School of Information Systems
Curtin Business School

Ontology Based Data Warehousing for Mining of Heterogeneous and Multidimensional Data Sources

Shastri Lakshman Nimmagadda

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

February 2015

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature:

Date: 26th February 2015

ACKNOWLEDGEMENT

The author acknowledges his deepest gratitude to Prof Dr Heinz Dreher, School of Information Systems, Curtin Business School (CBS), Curtin University, for his invaluable supervision throughout my PhD work. The critical discussions and comments made by Prof Dr Heinz Dreher during preparation of the thesis and research papers are commendable. The author and Prof Dr Heinz Dreher attended together few conferences and his keen interest in the current areas of the research work is well appreciated. The author is grateful to Dr Douglas Atkinson, Chairman of Research Committee for his encouragement and his continued support during the course of the research at Curtin University. The author wishes to accord his thankfulness to Dr Ananda Jeeva, School of Information Systems, for review of manuscript, fruitful suggestions and necessary facilities provided during preparation of research thesis at the School of Information Systems, CBS, Curtin University. Discussions made with the staff of School of Information Systems, especially with Dr Amit Rudra, CBS, Curtin University are useful at various stages of the work done at School of Information Systems premises, without which, current research could not have been completed. Prof Laurie Dickie's review and feedback helped the author in finalizing the manuscript. Dr Janine Rutledge, CBS Communication skills Centre, provided fruitful suggestions on the grammar and semantics of the text in the thesis.

The author wishes to acknowledge the feedback and critical comments made by the juries, Professor Tommi Karhela and Associate Professor Mustafa Jarrar that greatly helped to improvise the structure and contents of the thesis. The author has utilized this opportunity to update the contents in the thesis based on latest work done in the areas of “ontology modelling”, “big-data systems” and “business data analytics”.

The support and the immense patience from all the family members are well respected. Without their inspiration and encouragement, this work would not have been successfully completed.

….In memory of my beloved late parents

ABSTRACT

With an increased use of databases and explosive growth in their sizes, such as oil and gas exploration and production businesses face problems of data and information overload. Data search becomes tedious when specific queries are made to support crucial technical and financial decisions. Due to accumulated volumes of heterogeneous data and information in multiple primary sources, websites and company servers, the oil and gas industry needs a robust and flexible database organization. In order to address these major issues, Design Science (DS) research approach is adopted, conforming its guidelines through which, systematic data mapping workflows and procedures are proposed, integrating data structures from different sources and in different knowledge domains. Data warehousing framework is aimed at mapping and modelling comprehensive multidimensional domain ontologies that support various warehouse architectures. Building ontology based data warehouse (a digital oil field solution, for an oil and gas exploration and production industry, for example), resolving the semantics, schematic, syntactic and system heterogeneities attributed by entities and dimensions, classifying and interpreting them in several application domains, are key initial objectives. Emphasis is given to spatial-temporal dimensions, including multidimensional modelling and building of data relationships in large geographic regions (characteristically in a *basin* scale, as an example) and long historical periods, which are typical in the oil and gas business environment. A concept of a digital ecosystem is developed using these integrated frameworks. An ontology based data warehouse is expected to provide a mechanism for delivering comprehensive, consistent and flexible data structures. Other objectives are to develop mining models, aimed at providing infrastructure for extracting vital information from warehouses within short periods of time and interpret it for knowledge discovery in any application domain. Clustering of classified data, mining of association rules and construction of decision trees, are techniques proposed in the current research studies. 2D and 3D data visualization cross plots, map views, histograms, bubble plots, scalar plots, scatterograms and plots representing statistical multi-variates and regression analysis, are used for presenting the explored data views. The explored data graphically (data visualization), can uncover properties of the data quickly, easily and detect any patterns and or deviations from expected results. These ideas are extended to big-data systems for business data analytics. The proposed methodology is expected to deliver accurate and precise digital oil field information solution in multi-billion dollar resource projects, crucial for data interpretation, knowledge discovery and operational decision-making especially in an upstream and integrated business environment.

CONTENTS

FIGURES

Tables

LIST OF ACRONYMS

AIDS:	Acquired Immunodeficiency Syndrome
AI:	Artificial Intelligence
AOI:	Area of Interest
ASCII:	The American Standard Code for Information Exchange
API:	American Petroleum Institute
ASP:	Application Service Providers
BCNF:	Boyce-Codd Normal Form
B2B:	Business to Business
CF:	Clustering Feature (tree structure)
CDP:	Common Depth Point (a dimension representing depth domain)
COP:	Common Offset Point (a type dimension to represent the seismic instance)
CRP:	Common Receiver Point (a dimension representing the seismic instance)
CSP:	Common Source Point (a dimension representing seismic data instance)
CEOS:	Chief Executive Officers
CMP:	Common Midpoint (a dimension leveraging distance)
CBM:	Coal Bed Methane
DS:	Design Science
DB:	Database
DTD:	Document Type Definition
DW:	Data Warehouse
DDL:	Data Definition Language
DM:	Data Mining
DBA:	Database Administration
E & P:	Exploration and Production
ETL:	Extract, Transform and Load
ER:	Entity Relationship
ERM:	Entity Relationship Model
E^{2}R (EER):	Extended Entity Relationship
EDW:	Enterprise Data Warehouse
EDA:	Exploration Data Analysis
EHS:	Environment, Health and Safety
FP:	Frequent Pattern (tree structure)
F & A:	Finance and Administration
G & G:	Geology and Geophysics
GIS:	Geographic Information System
GGS:	Group Gathering Station
GNP:	Gross National Product
GP:	Gross Porosity
GA:	Geoscience Australia
HTML:	Hypertext Markup Language
I/O:	Input/Output
IS:	Information Systems
IOC:	International Oil Companies
IPA:	Indonesian Petroleum Association
KDD:	Knowledge Discovery in Databases
LASLOG:	Log Data Representation in LAS Format
LPG:	Liquefied Petroleum Gas
M^{2}R (MMR)/EMR:	Extended Multidimensional

MR:	Multidimensional Relationships
MOLAP:	Multidimensional Online Analytical Processing
MM:	Material Management
MM:	Multidimensional Modelling
NWS:	North West Shelf
NF:	Normal Form
NWSV:	North West Shelf Venture
NSW:	New South Wales (one of the states of Australia)
NP:	Net Porosity
NASA:	The National Aeronautics and Space Administration
NOC:	National Oil Companies
NT:	Northern Territory (one of the states of Australia)
OLAP:	Online Analytical Processing
OGIT:	Oil and Gas Information Technology
OO:	Object Oriented
OWL:	Web Ontology Language
PDE:	Petroleum Digital Ecosystems
PIS:	Petroleum Information Systems
P & A:	Personnel and Administration
PO:	Petroleum ontology
PROLOG:	Programming Logic (Language)
PC:	Personal Computers
PVT:	Pressure, Volume, Temperature
QC:	Quality Control
QLD:	Queensland (one of the states of Australia)
RDBMS:	Relational Database Management Systems
RDF:	Resources Description Framework
RMS:	Root Mean Square
SQL:	Structured Query Language
SP:	Shot Point (a location where seismic energy applied)
SST:	Sea Surface Temperature
SA:	South Australia (one of the states of Australia)
SHVF:	Shale Volume Fraction
TPS:	Total Petroleum System
TVDSS:	True Vertical Depth Subsea (ft)
UML:	Unified Modelling Language
VSP:	Vertical Seismic Profiling (data integrating surface and sub-surface domains)
WWW:	World Wide Web
WA:	Western Australia (one of the states of Australia)
WWAP:	The World Water Assessment Programme
XML:	Extensible Markup Language
XSD:	XML Schema Definition
ZMAP:	Depth Mapping Software

PUBLICATIONS GENERATED

The research approach is from a research framework, comprising of constructs, models, methods and implementation, validating the guidelines of Design Science (DS) perspective, providing proposed technology as a decision support and knowledge base system for operational businesses and industries. The outcome of research is year-wise appeared in the proceedings of different international conferences and journals.

2004

1. Nimmagadda, S.L. and Rudra, A (2004a) Applicability of data warehousing and data mining technologies in the Australian resources industry, *Published in the proceedings in the 7th international conference* on IT, held in Hyderabad, **India**.
2. Nimmagadda, S.L. and Rudra, A. (2004b) Data sources and requirement analysis for multidimensional database modeling – an Australian Resources Industry scenario, *published in the proceedings in the 7th international conference* on IT, held in Hyderabad, **India**.

2005

1. Rudra, A. and Nimmagadda, S.L (2005) Roles of multidimensionality and granularity in data mining of warehoused Australian resources data, *Proceedings of the 38th Hawaii International Conference on Information System Sciences*, Hawaii, **USA**
2. Nimmagadda, S.L. and Dreher, H. (2005) Ontology of Western Australian petroleum exploration data for effective data warehouse design and data mining, a paper presented and published in the *proceedings of the 3rd international IEEE conference on Industrial Informatics*, held in Perth, **Australia**, August, 2005.
3. Nimmagadda, S.L. and Dreher, H. (2005) Data warehouse structuring methodologies for efficient mining of Western Australian petroleum data sources, a paper presented and published in the *Proceedings of 3rd international IEEE conference on Industrial Informatics*, held in Perth, **Australia**, August, 2005.
4. Nimmagadda, S.L, Dreher, H. and Rudra, A. (2005) Warehousing of object oriented petroleum data for knowledge mapping, a paper presented and published in the *Proceedings of the 5th International Conference of IBIMA*, Cairo, **Egypt**.
5. Nimmagadda, S.L. and Rudra, A. (2005) Data Mapping Approaches for Integrating Petroleum Exploration and Production Business Data Entities for Effective Data Mining, a paper presented and published in the *proceedings of 3rd Kuwait International Petroleum Conference and Exhibition (KIPCE2005)*, **Kuwait City**.

2006

1. Nimmagadda, S.L, and Dreher, H. (2006a) Mapping and modelling of Oil and Gas Relational Data Objects for Warehouse Development and Efficient Data

Mining, a paper presented and published in the *proceedings of the 4th International Conference of IEEE Industry Informatics*, held in **Singapore**, August.

2. Nimmagadda, S.L, and Dreher, H. (2006b) Mapping of Oil and Gas Business Data Entities for Effective Operational Management, a paper presented and published in the *proceedings of the 4th International Conference of IEEE Industry Informatics*, held in **Singapore**, August.
3. Nimmagadda, S.L. and Dreher, H. (2006) Ontology-Base Data warehousing and Mining Approaches in Petroleum Industries: in Negro, H.O., Cisaro, S.G., and Xodo, D., (Eds.), Data Mining with Ontologies: Implementation, Findings and Framework, a book chapter published in 2007 by Idea Group Inc. http://www.exa.unicen.edu.au/dmontolo/
4. Nimmagadda, S.L, Dreher, H. Chang, E. and Rajab, M.R (2006) New technologies in mature gulf basins – multidimensional modelling of ontologically derived historical petroleum exploration data properties for effective basin knowledge mapping, a poster paper presented and published in the *AAPG international conference and exhibition*, 5-8 November, Perth, **Australia**

2007

1. Nimmagadda, S.L. Dreher, H. and Rajab, M.R (2007), Ontology-based Warehouse Time-Depth Data Modelling Framework for Improved Seismic Interpretation in Onshore Producing Basins, a paper presented in *the International Petroleum Technology Conference (IPTC)*, held in Dubai, **UAE**, Dec 2007
2. Nimmagadda, S. L. and Dreher, H (2007) DESIGN OF PETROLEUM COMPANY'S METADATA AND AN EFFECTIVE KNOWLEDGE MAPPING METHODOLOGY, a paper presented in the *IASTED* conference, held in Cambridge in **USA**, November 2007.
3. Nimmagadda, S.L., and Dreher, H. (2007) Ontology based data warehouse modelling and mining of earthquake data: prediction analysis along Eurasian-Australian continental plates, a paper published in the proceedings of *International Conference of IEEE in Industry Informatics Forum*, Vienna, **Austria**.

2008

1. Nimmagadda, S. L., and H. V. Dreher. 2008. "Ontology-based data warehousing and mining approaches in petroleum industries." In *Data warehousing and mining: concepts, methodologies, tools and applications*, ed. John Wang, 1901-1925. Hershey, New York and London, UK: Information Science Reference.
2. Nimmagadda, S.L and Dreher, H. (2008). Ontology Based Data Warehouse Modelling – a Methodology for Managing Petroleum Field Ecosystems, a paper presented in the *International conference of IEEE-DEST*, held in Bangkok, **Thailand**, Feb 2008.
3. Nimmagadda, S.L, Nimmagadda, S. K. and Dreher, H. (2008). Ontology based data warehouse modeling and managing ecology of human body for disease and drug prescription management, a paper presented and published in the proceedings of an *International conference of IEEE-DEST*, held in Bangkok, **Thailand**, Feb 2008.
4. Nimmagadda, S. L and Dreher, H. (2008). Petroleum Ontology: an effective data integration and mining methodology, aiding exploration of commercial

petroleum plays, a paper presented and published in the proceedings of an *International Conference of IEEE (INDIN'08)*, held in Daejeon, **South Korea**.

2009

1. Nimmagadda, SL. and Dreher, H. (2009). Petro-data-clustering – knowledge building analysis of complex petroleum systems, a technical paper accepted for presentation in the *international IEEE-ICIT conference*, held in Melbourne**, Australia**, Feb 2009.
2. Nimmagadda, SL. and Dreher, H. (2009). On designing Multidimensional Oil and Gas Business Data structures for effective data warehousing and mining, a technical paper presented *in an international conference of IEEE-DEST*, held in Istanbul, **Turkey,** June, 2009
3. Nimmagadda, S. L. and Dreher, H. (2009). On issues of Data Warehouse Architectures – Managing Australian Resources Data, a technical paper presented in an *international conference of IEEE-DEST*, held in Istanbul, **Turkey**, June, 2009
4. Nimmagadda, S. L. and Dreher, H. (2009). Ontology based data warehouse modelling and mining approaches for managing carbon emission ecosystems, a paper submitted to *20th Australasian Conference on Information Systems,* held in Melbourne, **Australia** 2009.
5. Nimmagadda, S.L, and Dreher, H. (2009) Technologies for adaptability in turbulent resources business environments, a book chapter published under a title: Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains, http://www.igi-global.com/, 2009, **USA**
6. Nimmagadda, S.L, and Dreher, H (2009) Ontology based data warehouse modeling for managing carbon emissions in safe and secure geological storages, a paper presented in the international *SEGJ symposium* – Imaging and Interpretation, in a forum "science and technology for sustainable development", held in Sapparo, **Japan**, Oct 2009.

2010

1. Nimmagadda, S. L. and Dreher, H. (2010) Modelling Multidimensional Australian Resources Data for an effective Business Knowledge Management, a technical paper presented and published in the 8th International Conference and Exposition on Petroleum Geophysics, organized by the *Society of Petroleum Geophysicists* (SPG) and sponsored by SEG and EAGE, held in Feb 2010, Hyderabad, **India**.
2. Nimmagadda, S. L and Nimmagadda, S. K. and Dreher, H (2010) "Multidimensional Ontology modelling of Human Digital Ecosystems affected by Social Behavioural Patterns", presented and published in the proceedings of an *IEEE-DEST-2010*, held in Dubai, **UAE**, April, 2010
3. Nimmagadda, S. L and Dreher, H (2010) "Ontology based warehouse modelling of fractured reservoir data – for an effective borehole and petroleum production management"*,* presented and published in the proceedings of an *IEEE-DEST-2010*, held in Dubai, **UAE**, April, 2010
4. Nimmagadda, S. L and Dreher, H (2010) "On new emerging concepts of Modelling Petroleum Digital Ecosystems by Multidimensional Data Warehousing and Mining Approaches" presented and published in the proceedings of an International Conference of *IEEE-DEST-2010*, held in Dubai, **UAE**, April, 2010
5. Nimmagadda, S. L and Dreher, H (2010) "On Data Integration Workflows for an effective Management of Multidimensional Petroleum Digital Ecosystems,

Arabian Gulf Basins" presented and published in the proceedings of an *IEEE-DEST-2010*, held in Dubai, **UAE**, April, 2010

2011

1. Shastri L Nimmagadda, Heinz Dreher, Fred Kabanda, Andrew Ochan, Philips Obita, Lyodah Kiconco and Proscavia Nabbanja*,* On Data Integration Workflows for an effective Management of Multidimensional Petroleum Digital Ecosystem in the Albertine Graben*, a technical paper submitted to the East African Petroleum Conference, held during February 2011, Kampala,* **Uganda**

2. Shastri L Nimmagadda, Heinz Dreher, Fred Kabanda, Andrew Ochan, Lyodah Kiconco, Proscavia Nabbanja, On new emerging concepts of Modeling Albertine Graben Petroleum Digital Ecosystem by Multidimensional Data Warehousing and Mining Approaches*, a technical paper submitted to the East African Petroleum Conference, held during February 2011, Kampala,* **Uganda**

3. Shastri L Nimmagadda and Heinz Dreher, (2011) Data warehousing and mining technologies for adaptability in turbulent resources business environments, *Int. J. Business Intelligence and Data Mining,* Vol. 6, No. 2, 2011, p 113-153.

4. Nimmagadda, S. L and Dreher, H (2011) "Shale-Gas Ontology, a robust data modeling methodology for integrating and connecting fractured reservoir petroleum ecosystems that affect production complexities" presented and published in the proceedings of an *IEEE-INDIN-2011* held in Lisbon, **Portugal**, July 2011.

5. Nimmagadda, S. L, Nimmagadda, S. K and Dreher, H (2011) "Multidimensional Data Warehousing and Mining of Diabetes & Food-domain ontologies for e-Health Management*"* presented and published in the proceedings of an *IEEE-INDIN-2011* held in Lisbon, **Portugal,** July 2011.

2012

1. Nimmagadda, S. L. and Dreher, H (2012) "On new emerging concepts of Petroleum Digital Ecosystem (PDE)", Journal *WIREs Data Mining Knowledge Discovery,* 2012, 2: 457–475 doi: 10.1002/widm.1070.

2. Nimmagadda, S.L, Dreher, H, Noventianto. A, Mustofa. A and Fiume. G. (2012) Enhancing the process of knowledge discovery from integrated geophysical databases using geo-ontologies, a paper presented and published in the Indonesian Petroleum Association (IPA) conference, held in Jakarta, **Indonesia**.

3. Nimmagadda, S.L, Dreher, H, Noventianto. A, Mustofa. A and Fiume. G. (2012) On new emerging concepts of Tarakan Sedimentary Basin – a Petroleum Digital Ecosystem (PDE), a paper published in the proceedings of an *International Geological Congress* (*IGC*) held in Brisbane, **Australia**, August 2012.

4. Nimmagadda, S.L and Dreher, H. Data Integration Methodologies for Effective Management and Data Mining of Petroleum Systems of South East Asian Sedimentary Basin, a technical paper communicated to *IEAT*, held in Kolkata, **India** 2012.

2013

1. Nimmagadda, S.L and Dreher, H., 2014 Ontology based Multidimensional Data Warehousing and Mining of Heterogeneous Data Sources for managing

Carbon Ecosystems, accepted for publication in the international journal of business analytics (*IJBAN*), USA, 1(2), Apr. 2014.

2. Nimmagadda, S.L, Dreher, H, Noventianto, A, Mustoffa, A and Parapaty, H. 2013, Sedimentary Basin - a Petroleum Digital Ecosystem*,* presented and published in the proceedings of an international conference of *IEEE, INDIN (Industrial Informatics),* held in Bochum, **Germany**.
3. Nimmagadda, S.L, Dreher, H, Shtukert, O and Zolotoi, N 2013, Multidimensional Data Warehousing and Mining - an Approach for Managing Multiple Reservoir Ecosystems*, presented and published in the proceedings of IEEE, INDIN (Industrial Informatics)*, held in Bochum, **Germany**.
4. Nimmagadda, S.L and Dreher, H 2013, Data Integration Methodologies for Effective Management and Data Mining of Petroleum Provinces*,* presented and published in the proceedings of an international conference of *IEEE-DEST (Digital Ecosystem Technologies)*, held in Stanford University, **USA**.
5. Nimmagadda, S.L, Dreher, H, Cardona Mora, A, P, Lobo, A, 2013, Ontology based Multidimensional Data Warehousing and Mining of Heterogeneous Unconventional Reservoir Ecosystems*,* presented and published in the proceedings of an international conference of *IEEE, INDIN (Industrial Informatics)*, held in Bochum, **Germany**.
6. Nimmagadda, S.L, Dreher, H, Noventianto, A, Mustoffa, A and Parapaty, H. 2013, On Heterogeneous, Multidimensional Unconventional Reservoir Ecosystems, presented and published *in the proceedings of an international conference of EAGE*, held in Kiev, **Ukraine** in May 2013.
7. Nimmagadda, S.L, Nimmagadda, S.K and Dreher, H 2013, On Robust Methodologies for Managing Public Health Care Systems*, International Journal of Environment Research and Public Health*, 2014, 11, 1106-1140; doi:10.3390/ijerph110101106.

2014

1. Nimmagadda S.L. and Dreher, V.H. (2014), Multidimensional Ontology Modelling – a Robust Methodology for Managing Complex and Heterogeneous Petroleum Digital Ecosystems, presented and published in the proceedings of an international conference of IEEE and IEEEExplore, held in Porto Alegre, **Brazil.**
2. Nimmagadda, S.L. Rudra, A. and Dreher, H.V (2014), Integration and Effective Management of Heterogeneous Petroleum Digital Ecosystems Using Big Data Paradigm, a technical paper presented in a symposium of Professional Petroleum Data Management (PPDM), held in **Perth**, Australia http://www.ppdm.org/event/view/archived/144.
3. Nimmagadda, S L. Rudra, A. and Dreher, H. V. (2014), Managing petroleum digital ecosystems using big-data paradigm, an invited lecture, School of Information Systems, Curtin Business School, Curtin University, WA, Perth, **Australia**.
4. Nimmagadda, S.L. Rudra, A. and Dreher, H.V. (2014), Big Data Information Systems for Managing Embedded Digital Ecosystems (EDE), accepted as a book chapter proposal for publication in a book entitled "Big Data and Learning Analytics: Current Theory and Practice in Higher Education" Springer, **New Zealand**.

2015

1. Nimmagadda, S L. and Rudra, A. (2015), On a Holistic Modelling Approach for Managing Carbon Ecosystems, communicated to the Journal of Greenhouse Gases – Science & Technology, Wiley Online Library, USA**.**
2. Nimmagadda, S L. and Rudra, A. (2015), Big-data Paradigm and its role in Design Science Research Framework communicated to an International

conference on Design Science Research in Information Systems and Technology (DESRIST 2015), Dublin, Ireland.

3. Nimmagadda, S L. and Rudra, A. (2015), On Managing Energy Data Sources using Big-data Paradigm, communicated to the Americas Conference on Information Systems (AMCIS 2015), Puerto Rico**.**

Chapter 1: Problem Statement, Research Questions, Significance and Motivation

1.0 Introduction

Handling multidimensional and heterogeneous data sources is major challenge in the oil and gas industries. This is because of sheer volume, size and complexity of data sources accumulated in many organizations. Although several domain ontologies are described for mining large volumes of heterogeneous and multidimensional business data, they are underutilized for mining and exploration of new knowledge. Jasper and Uschold (1999), Meersman (2004), Hadzic and Chang (2005) and Flahive et al. (2004) describe a number of ontologies and their relevance in different domain applications. Damiani (2008) describes digital ecosystems in service industry domains. A large number of heterogeneous and multidimensional oil and gas business data sources available on companies' archives, but are not yet used for mining and knowledge interpretation. As demonstrated in Nimmagadda et al. (2005c) and Nimmagadda and Dreher (2006c, 2008b), in the oil and gas domain, petroleum digital ecosystems (PDE), digital oil field solutions and petroleum information systems (PIS) are innovative ideas that address issues associated with multidimensional spatial-temporal data modelling, mining and interpretation for new knowledge discovery.

In this chapter, the author identifies issues, challenges and objectives of the research, and the significance of the proposed research including the motivations for the research. Research methodologies and the data sources, used in the modelling process are also discussed. In addition, domain ontologies, ontology based data modelling, schema architectures, data mining and visualisation including interpretation and knowledge discovery are considered relevant to upstream oil and gas upstream businesses.

1.1 Issues and Challenges with the Existing Data Sources and their Organization

If the *reservoir*, *structure* (geological), *seal*, *migration*, *source* and *timing* factors coexist in a system and contribute to sustained oil or gas production, it is said to be an oilplay. It has key role for building a constructive (productive) petroleum system in any sedimentary basin. In spite of major breakthroughs and advances in resources technologies, the identification and precise description of petroleum systems, their connectivity and the limitations that narrate oilplay factors (on a commercial scale), remain unresolved. This is because of poorly understood data sources and lack of integration in different domains. Furthermore, the phenomenon of petroleum ecosystems has not been readily descriptive. Heterogeneity and multidimensionality are additional issues, because of which data sources complicate the concept identification and interpretation in knowledge different domains.

Highly specialized data semantics make it infeasible to incorporate ideas within a consistent repository. The semantics of petroleum data are often hard to precisely define (Meersman 1999, 2000, 2001 and 2004, Cardenas and McLeod 1990) because they are not explicitly stated and are implicitly included in database designs. However, petroleum ontology (PO, much less all of geology or petroleum geology science) is not a single, consistent scientific domain; it is composed of dozens of smaller, focused research communities. This would not be a significant issue if researchers were able to access data from within a single domain, but that is not usually the case. Typically, researchers require integrated access to data from multiple domains, which requires resolving terms that have slightly different meanings across communities. This is further complicated because the specific community where the terminology is used, is usually not explicitly identified and because terminology evolves over time. For many of the larger, community data sources, the domain is obvious—the PO handles petroleum structure information, the petro- database provides petroleum sequence information and useful annotation and so forth—but the terminology used may not be current and can reflect a combination of definitions from multiple domains. *Geology* also demonstrates these challenges for data integration because they are in evolving scientific domains, not typically found elsewhere.

These issues exist with both conventional and unconventional oil and gas data sources and systems associated with big data. The major challenges are the quality within the existing data sources, also in published sources (Nimmagadda and Rudra 2004, 2005) and the enormous amount of time to reconcile this. Numerous findings on exploration are published in journals and technical reports (Nimmagadda and Dreher 2006c). These published reports (*AAPG/SEG/SPE*, all oil and gas reports from professional societies) cover descriptions of different oil and gas systems unfolding oilplays and so represent a huge range of exploration and development findings and their production potential (in the worldwide basins).

Many operating companies possess decades of oil and gas data sources, which are in awful state, because they are not well organized or appropriately used. The concepts of data warehousing and mining, data modelling procedures for data warehouse design, and their applications in the oil and gas industry, are relatively new and will be critically examined in this study. Petroleum systems, are analogous to other information systems, where connectivity is poorly understood (Nimmagadda and Dreher 2012a and Nimmagadda et al. 2006d). So, issues of multidimensionality, heterogeneity and granularity which are crucial for resolving issues associated with data integration and sharing of an interpretable knowledge among multiple domains, are examined in this current study.

Another issue is that the volumes of web data available through search engines are still raw and unstructured. This means, it is difficult to extract knowledge as the information can be shallow, multidimensional, heterogeneous, hidden or deep nature. In addition data types intricate the existing data mining schemes, user queries and knowledge extraction. Data types are also significant in ascertaining the type of knowledge to be extracted. Therefore, reducing the burden on data mining hardware and software resources by improving the logic of warehouse schemas and user response times are other key goals. As a consequence when designing data warehouse schemas, reduction of I/O operations and efficient computing operations are among other impacts to be considered. For example, improved data structuring design and development can drastically improve handling of massive and volumes of oil and gas data sources to the order of thousands of giga or tera/peta bytes. When response times are recovered, it reduces the burden on I/O system and minimizes the transaction response time by running efficient mining algorithms under lower workloads (that may comply with daily functioning of the database server). This will avoid network congestion and enable effective use of computing resources. Thus software and hardware resources are crucial, to the adoption of data warehouse and data mining applications in the oil and gas industry.

1.2 Research Problem Statement

The author has involved with exploration and field development projects in upstream oil & gas for many years. Therefore, the motivation for this research is based on current knowledge and previous work experience with operating and service companies. It is from these work experiences that deficiencies in the current data organizations are identified. Hundreds of dimensions and their associated attributes, including their conceptualized dimensions have ambiguities in their semantics. Description of dimensions and their relationships among multiple domains are challenging (Rudra and Nimmagadda 2005) in the upstream business. Models built based on conceptualized dimensions or constructs can deliver new knowledge, is another issue. The author explores whether conceptualized dimensions and their models can build new knowledge. The author by virtue his experiences, describes entities and or dimensions associated with geology, geophysics, geochemistry, oil & gas exploration, drilling, production, technical and logistics, with several of their data attributes. Prior knowledge from major commercial companies (in the oil and gas industry) is a prerequisite for fast tracking and developing infrastructure. The following major issues are identified, as in A, B, C and D sections:

A Business Data management

Business data management is a critical issue because of explosion of sheer volumes of data and their sizes. In addition, they are heterogeneous and multidimensional. In a competitive and integrated business environment, data integration is another major challenge.

- At present, massive storage devices store volumes of heterogeneous data (hundreds of GB for each basin) that describe numerous technical and commercial data entities. Large upstream integrated companies (such as oil and gas companies) are unable to manage multifaceted data, comprising of heterogeneous data structures (relational, hierarchical and networking) with complex data types, such as spatial-temporal data.
- Handling numerous entities and or dimensions (of the order of 500) and attributes (of the order of 1000), mapping and modelling thousands of tables is a tedious process.
- Data integration (sometimes among 10-15 operational centers) of enormous amount of multi-disciplinary data, is a serious business, in large (oil and gas business) commercial companies. Sharing of data in a multi-client environment must be a prerequisite for carrying out the successful (such as oil and gas exploration and production) business research.

B Business data mining and visualization; knowledge building

Mining and visualization of oil and gas data is another major challenge in building new knowledge and its interpretation.

- Knowledge building from these massive data structures is an intricate issue. With an increasing volume of periodic data, there is difficulty in understanding or retrieving knowledge from historical data. At times, many applications are forced to integrate and share volumes of multi-disciplinary data without prior knowledge of the business (such as basin or oil and gas-play entities).
- Knowledge of past business data exploration and exploitation data allows future business system improvements. Unknown relationships among different business data entities and or dimensions, also influence the economics involved in running the business. Improved data management may reduce running of future risky business through data analysis in time domain.
- The data from heterogeneous sources, at times are difficult to uncover and interpret data patterns because of its volume and complexity. The visualization techniques proposed in the present research, facilitate to uncover the hidden patterns of the data for presentation and interpretation.
- Finally, applicability and feasibility of data warehouse, supported by ontology, if successful, combined with application of data mining and visualization has tremendous impact on business data system knowledge discovery that can change the economics of commercial business (such as oil and gas business).

C Problems with the existing data structuring methods (Hoffer et al. 2005, Nimmagadda and Dreher 2006c and Nimmagadda and Dreher 2007b)

- Lack of proper domain knowledge and limited sharing of domain knowledge
- Semantic, schematic and syntactic issues during ontology descriptions
- Multidimensionality, granularity and heterogeneity of data cannot be resolved with the existing data structuring approaches
- Poor data integration and poor understanding of operational knowledge, leading to ambiguity in knowledge interpretation
- Data represented in relational structures, possess several constraints in applying business rules, with the result, inconsistency, less flexibility and inability to adapt to fast changing business situations

D Anticipated solutions for existing data structuring problems

- To share common understanding of the structure of information among entities and dimensions (ontologically represented)
- To enable reuse of domain knowledge among different applications
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge and its implementation
- To highlight the scope of extending these solutions in extracting new knowledge and use in other domain applications

The current research in the domain of interest is because of complexity in representing geological knowledge and its dynamics in the conceptualization and contextualization. The author considers the ways and means, in which semantic, syntactic, schematic and system heterogeneities impact the structuring process of data sources:

Unfortunately, the semantics of oil and gas data sources are hard to define (Meersman 1999, 2000 and 2001 and Cardenas and McLeod 1990) because they are not explicitly stated, but are implicitly included in the database design. The reason is simple: at a given time, within a single *sedimentary basin (*a geological key term, defined in glossary in Appendix-1) scenario, common definitions of various terms are often well understood and have precise meaning. As a result, those within that community usually understand the semantics of the data source without needing to be explicitly defined. However, petroleum ontology (much less all of geology or petroleum geology science) is not a single, consistent scientific domain; it is composed of dozens of smaller, focused research communities. This would not be a significant issue if researchers only accessed data from within a single domain, but that is not usually the case. Typically, researchers need integrated access to data from multiple domains, which requires resolving terms that have slightly different meanings across the communities. This is further complicated by observations that a specific community whose terminology is being used by the data source is usually not explicitly identified and that the terminology evolves over time. For many large community data sources, the domain is obvious—the petroleum ontology (PO) handles petroleum structure/reservoir information, the petro- database provides petroleum sequence information and useful annotation and so forth—but the terminology used may not be current and can reflect a combination of definitions from multiple domains.

Geology demonstrates these challenges for data integration that are common in evolving scientific domains, but not typically found elsewhere. The first is the sheer number of available data sources and an inherent heterogeneity of their contents. Some of these sources contain data from a single lab or project, whereas others are definitive repositories for very specific types of information (e.g., for a specific reservoir class). Not only do these sources complicate the concept identification issue because they use highly specialized data semantics, but they make it infeasible to incorporate all of them into a consistent repository. Second, the data formats and data access methods change regularly. These changes are an attempt (Wand 2000) to keep up with the scientific evolution occurring in the community at large. However, a change in a data source representation can have a dramatic impact on systems that integrate that source, causing the integration to fail on the new format. Third, the data and their related analysis are becoming increasingly complex. As the nature of petroleum ontology research evolves from a predominantly wet-lab activity into knowledge-based analysis, the scientists' need to access the wide variety of available information, increases dramatically.

Big-data that emerged from major *sedimentary basins,* is another key focus. This is because of the fact that the operational data are not appropriately integrated and the information shared by various operational units, may compromise crucial operational decisions. A systematic shared ontology, supported by data warehouse modelling and data mining research is necessary. Most of the published results, available on the Web through journal databases (given in Appendix-2); oil and gas databases in several formats and on software platforms are amenable to this approach. Historical data are available in hard copies. For example, historical exploration and production data in the Canning, Carnarvon, Bonaparte, Browse, Officer, Eucla and Perth basins are either available in hard copies (given in Appendix-2 and Nimmagadda and Dreher 2005a and 2005b) or on different software and hardware media and, at times, these valuable data are not simple to retrieve for documentation and computer analysis.

Understanding the *prospect* of a *sedimentary basin* is a significant problem (Beaumont and Foster 1999). Data integration and sharing of information among different fields or prospects of different *basins* are key issues of the present problem definition. According to (Mattison 1996), little attention is paid in integrating and organizing the historical data sources. To date, there has been no systematic investigation of these data sources using focused technologies. Meticulous analysis of *oilplays* (a key term defined in the glossary in Appendix-1) of different fields of petroleum systems, associated with different basins in their hierarchies, is a much-needed research. Even without additional surveys or exploratory drilling, many more prospects can be explored or discovered by data mining of existing exploration and production warehoused metadata. Unorganized volumes of massive stores of oil and gas business data hide undiscovered (unknown knowledge or intelligence) data patterns. Interpreting patterns, correlations and trends among exploration, drilling and production data as well as their *oilplay* factors into meaningful scientific geological and oil and gas business information, is one of the goals of the current research. The major challenge is identification of a new locale for an exploratory well and its delivery, for which the cost of drilling varies from five to hundreds of million dollars, depending upon the objectives of oil and gas investigations. These major challenges and issues demand for new robust methodologies, ontologies based data warehousing and mining, visualization and data interpretation solutions. The author describes them in details in the subsequent chapters.

1.3 Research Questions and Objectives of the Research

Research questions, framed in this research work are aimed at addressing the major upstream oil and gas and service providers worldwide. Ontology is a “shared understanding of certain domain represented in definitive logic (Gruber 1993 and 1995)”. In the present context, data are inherited from heterogeneous and multidimensional sources. ER and Multidimensional data models of petroleum systems’ *elements*, *processes* and *chains*, are used as logical articulations. While building these models, semantics, structure, syntactic and system heterogeneities too affect the process of systems’ integration. For this purpose, domain ontologies are articulated in the form of either ER and or Multidimensional Models for each system (in my context, system could be for a specific oil & gas field’s *elements*, *processes* and *chains* that share other with other fields and systems) and their integration. These are all simulated within an articulated integrated framework, in which, domain, data, data warehouse, data mining, data visualization and data interpretation play significant roles. The author uses and reuses these constructs, models and methods in the present domain to further validate in multiple domains and design science research guidelines. In order to interoperate, author applies these constructs and models in multiple oil & gas field situations.

The author focuses on current research on how best heterogeneous data are organized and new knowledge extracted from them. Based on the existing data sources, constructs, models and methodologies and their deficiencies and demerits, the author designs research questions in line with the objectives of the study.

1.3.1 Research questions (RQ)

1 RQ1 - Large amount of heterogeneous data sources are locked up in many industrial applications and knowledge domains. How are the heterogeneous data sources organized in oil and gas industries, keeping in view variety of entities, objects and dimensions of spatial-temporal nature of complex petroleum systems? Are petroleum systems analogous to information systems? How are the data mapping and structuring done, considering the variety of data sources in different petroleum systems' scenarios? The author proposes a more comprehensive work.

2 RQ2 - How are semantic, schematic, syntactic and system heterogeneities handled, keeping in view the current status of data sources in oil and gas industries?

3 RQ3 - How data integration is done, keeping in view the heterogeneities and multidimensionality?

4 RQ4 - How are the heterogeneous and multidimensional data sources accessed and explored for connections from metadata? Do the current data mining methods can explore such connections from data views of metadata?

5 RQ5 - How the correlations, trends and patterns of data views, extracted from metadata structures can be visualized and presented for interpretation?

6 RQ6 - Whether the visualized data views extracted from metadata structures, can be interpreted using the existing interpretation procedures?

7 RQ7 - There are number of sedimentary basins worldwide with hundreds of inherent petroleum systems with thousands of oil and gas fields geographically distributed. Are the producing companies equipped with better data organizing and analysing methodologies for interpreting the limits of petroleum systems?

8 RQ8 - Exploration business in oil and gas industry is risky business, keeping in view the economics involved in drilling onshore and offshore wells. Can the existing data structuring, integration, data mining and interpretation methodologies handle the risk of oil and gas exploration business?

It is evident that semantic, schematic, syntactic and system inconsistencies need attention during modelling and mapping of multiple knowledge domains. Therefore, the overall objective is to explore connections among variety of data sources and improve quality of information for quality interpretation. The author proposes a robust design and development of ontology-based data warehouse solution with data mining, data visualization and data interpretation articulations. Ultimately, this information will inform the existing decision support and knowledge discovery systems. More focussed research objectives are given in the following sections.

1.3.2 Research objectives (RO)

In the context of current research, the author designs artifacts with development of new innovative constructs, models and methodologies. These artifacts are intended to be part of an integrated methodological framework.

1) RO1: Ontology objectives – design and develop artifacts that describe ontologies for complex heterogeneous data dimensions, for creating knowledge-based structures including semantic information and rules/axioms and new knowledge – This objective is focused to address the RQ1 and RQ2.

2) RO2: Data warehouse objective – accommodate and integrate the ontologically structured logical data schemas in a warehouse for obtaining a metadata – this objective is focused to address the RQ3.

3) RO3: Data mining objective – investigate and develop data mining models for exploring warehoused metadata – this objective is focused to address the RQ4.

4) RO4: Data visualization objective – present the explored data views for better visualization and knowledge interpretation – the RQ5 addresses this objective.

5) RO5: Data interpretation objective – interpret the presented data views for extracting new knowledge and adding more value to the existing information (RQ6).

6) RO6: Ontology based data warehousing for mining of heterogeneous and multidimensional data sources (this research question is designed based on RQ 1 to RQ6) – an integrated framework for interpretation and new domain knowledge.

7) RO7*: Petroleum digital ecosystems and digital oil field solutions* - RO 7 addresses the RQ7. The above specific research objectives (that address RQ 1 to RQ 6) are taken advantage of designing and developing petroleum digital ecosystems and digital oil field solutions in oil and gas exploration industry scenarios. More details on methodologies and applications are given in Chapters 4 and 5.

8) RO8: *Risk minimize exploration* - RO8 addresses the RQ8.. Interpretation of multidimensional data views from warehoused metadata is expected to add value to knowledge domains of PDE and Digital Oil Field Solutions. Interpretative solutions can add new knowledge that can optimize the economics of exploratory drilling campaigns including expensive and risky oil and gas field development projects. More details on economic implications and analysis are provided in Chapters 4 and 5.

1.4 Significance, Motivation, Contributions and Goals of the Research

Petroleum information systems (*PIS*) and *petroleum digital ecosystems* (*PDE*) are new ideas, meant for better understanding of the connectivity and integration process among multiple information systems. The current research is a demonstration to understand how best they are embedded in nature. For example, *petroleum - ecological – geo-morphological systems* are inherently interrelated and interconnected, apparently embedded in nature and how best the new knowledge can be extracted from ontologically warehoused repository. Though previous researchers have investigated domain and data modelling, data warehousing and mining, visualization and interpretation in the business environment, it has been done in isolation. The approach of the current study is to model domains, data sources and integrate them in different applications, so that the knowledge extracted is effective for interpretation. Moreover, data modelling research, involving spatial-temporal data types in the oil and gas industries has not been the focus so far for explorers, data managers and data analysts.

Data warehouse schemas that depict multiple dimensions (combined with spatial-temporal) from multiple domains provide enhancements and scope for mining heterogeneous data sources. Ontology, data warehousing, data mining, data visualization and data interpretation, all in a single canvas, have been a focus in big data systems' design and development (Cleary et al. 2012), especially with spatial-temporal dimensions. Multiple oil and gas bearing *sedimentary basins* (Appendix – 1, see also Chapter 2) are targeted through the phenomena of connectivity and the process of integration. Oil and gas data integration and building knowledge from integrated models are significant research outcomes (Nimmagadda and Dreher 2008a and Shastri and Dreher 2011a). They are significant at further stages of data mining and analysis.

Warehoused multidimensional data cubes are effective for data mining, visualization and interpretation. But, the imposition of flexible business rules and constraints, is significant for the integration process. It is likely that there will be significant scope of analysing the heterogeneity, multidimensionality and granularity. Semantic, schematic, syntactic and system ambiguities that arise during structuring and integration process are expected to resolve during the description of domain ontologies and their integration. In addition to, faster operational and user responses that minimize the operational costs, reusability, search, reliability, flexibility, maintainability, integrity and security, will be significant in oil and gas domain research. Other significant operational criteria include, risk minimize exploration, appraisal and field development costs.

1.4.1 Significance

The author considers that the current research will be significant for academicians, researchers, oil and gas explorers and data management personnel. Research problem solutions have significance in implementing them in oil and gas companies, so that the risk involved in exploration and development can be minimized. Integrated methodological framework is expected to risk minimize the economics involved in exploratory drilling campaigns. There is huge demand for structured data and quality information in large industrial organizations (Winter and Strauch 2003). Petroleum exploration and production companies are such organizations, where there is immense scope of adapting new technologies and solving problems associated with geological and geophysical complexity, heterogeneity and multidimensionality.

1.4.2 Motivation

The *sedimentary basins* that occupy huge geographic regions worldwide, possess numerous *petroleum information systems* (*PIS*); and so understanding their connectivity is crucial in making multimillion dollar business decisions. This is what has motivated the author, to develop new ideas on ontology based data warehousing and mining of multidimensional and heterogeneous data sources. The use of an integrated framework has led to further development of *petroleum digital ecosystem* (*PDE*) ideas, and this has evolved in the form of *digital oil field solutions*. Because of the complexity and fast changing oil and gas business situations, there is growing demand for *digital oil field solutions* in both upstream and integrated business environments. Data integration and understanding the connectivity among multiple information systems (synonymous to petroleum systems and total petroleum systems) are paramount and are key motivating factors for the current research.

Several data sources available in multiple domains, with a large amount of historical data for various geographic regions have motivated the author to undertake the current research and design the research questions. The author has had consultations with variety of oil and gas producing companies. Based on the experiences with oil and gas producing and service companies, author experiences with pitfalls at various stages and chains of the operational units of oil and gas industries, especially keeping in view the heterogeneity and multidimensionality of data sources in the oil and gas industries.

1.4.3 Contributions

1 A research framework is in line with outcomes of *constructs*, *models*, *methods* and *instantiation* with combined research activities such as *build*, *evaluate*, *theorize* and *justify*, with an intention to apply this research framework in any domain. For this purpose, concept of design science and its guidelines are to be validated (Chapter 3).

2 A novel approach ontological structuring approach and integration of domain ontologies in a warehouse environment (Chapter 3).

3 A methodological framework consisting of domain, data, schema, warehouse, data mining, data visualization and data interpretation articulations (Chapter 3).

4 An exploration project, narrating the entities, objects and dimensions in different ontological constructs and their integration in a warehousing and mining is presented in Chapter 4.

5 Digital ecosystems in petroleum domain are new contributions, which provide improved understanding of the petroleum systems' new knowledge and its interpretation to risk minimize the exploration and drilling plans (Chapters 4 and 5).

6 Implementation of models and integrated framework and their evaluations, as demonstrated in Chapters 5 and 6. The commercial values of methodologies are emphasized.

1.4.4 Research Goals

An ontology defines a common vocabulary for researchers and users, in general who need to share information in different domains. In the context of current research work, ontology implies exploring relationships among various data sources. In addition, data integration is a key issue in developing a shared ontology in oil and gas domain, especially when integrating *exploration*, *drilling* and *production* domain ontologies. It includes machine-interpretable definitions of basic classes (concepts) in different domains and their respective relationships among them. Key goals of ontology based domain knowledge representation are:

1) To share common understanding of the structure of the information among entities: It is one of the common goals in developing ontologies. Several websites contain oil and gas information or provide oil and gas e-commerce services. If these websites share and publish the same ontological descriptions of the entities that are used, then computer agents can extract and aggregate information from these different sites. The agents can use this aggregated information to answer user queries or as input data to other applications.

2) To enable reuse of domain knowledge
Models from several domains need to represent the notion of space and time. This representation includes the notions of time-intervals, points in time, relative measures of time, and so on. If one group of researchers develops such ontology in detail, others can reuse it for their domains. As an example, the domain knowledge acquired from a particular model of a particular conventional oil and gas field, may be reused in the same ecosystem for an unconventional field. Additionally, if a large ontology needs to be built, several existing ontologies can be integrated describing portions of the large domain.

3) To make domain assumptions explicit, make necessary changes as per interpretation and implementation and in case the knowledge about the domain changes. Explicit specifications of domain knowledge are useful for researchers and investors who must learn what terms or entities/dimensions in the domain mean to earth science systems.

4) To separate domain knowledge from the operational knowledge
Existing known operational knowledge among entities or dimensions is separated from the undiscovered knowledge among emerging conceptualized entities or dimensions.

5) To analyse the domain knowledge (through data mining and visualization)
Data views extracted from warehoused oil and gas metadata are visualized for interpreting domain knowledge.

1.5 Thesis Outline

The present thesis embodies the results of an ontologically described data warehousing for mining of domain knowledge and its interpretation. Originally, the thesis contains 8 chapters. In Chapter 1, the author describes issues with the existing data sources, research problem statement, major challenges, operational issues, objectives of proposed research and motivations. An exhaustive literature survey is provided in Chapter 2. The research methodologies and workflows associated with data modelling, data warehousing, data mining, visualization and interpretation including a research framework are given in Chapter 3. In Chapter 4, an oil and gas exploration project is detailed with new insights on how entities, objects and dimensions are used in the modelling process. Modelling domain ontologies and applications, such as, design of company metadata, analyse systems that deal with big data, and conventional and unconventional petroleum digital ecosystems including shale gas ontologies are also detailed. In Chapter 5, results and discussions of models and methodologies including an outcome of oil and gas exploration project are provided. Evaluation and implementation of research outcomes are discussed in Chapter 6. The author summarizes the research work in Chapter 7. The final conclusions and recommendations are made in Chapter 8. The future scope, constraints, and limitations including documentation of other data sources in other domains are given. An exhaustive list of references is furnished at the end of the thesis. The description of key technical terms and the data sources used in the current research are given in Appendix-2.

1.6 Summary

The author describes the existing data sources, their organization in the oil and gas domain, defining issues and challenges associated with them. The author frames research questions in collaboration with objectives of research as summarized in the Figure 1.1. Figure 1.1 describes research questions referred to the chapters more than once or more. Similarly the research objectives are connected to the chapters multiple times. Each research question has at least one corresponding objective. Significance and motivation of the current research are narrated. The author surveys the literature that describes existing constructs, models and methodologies (see Chapter 2).

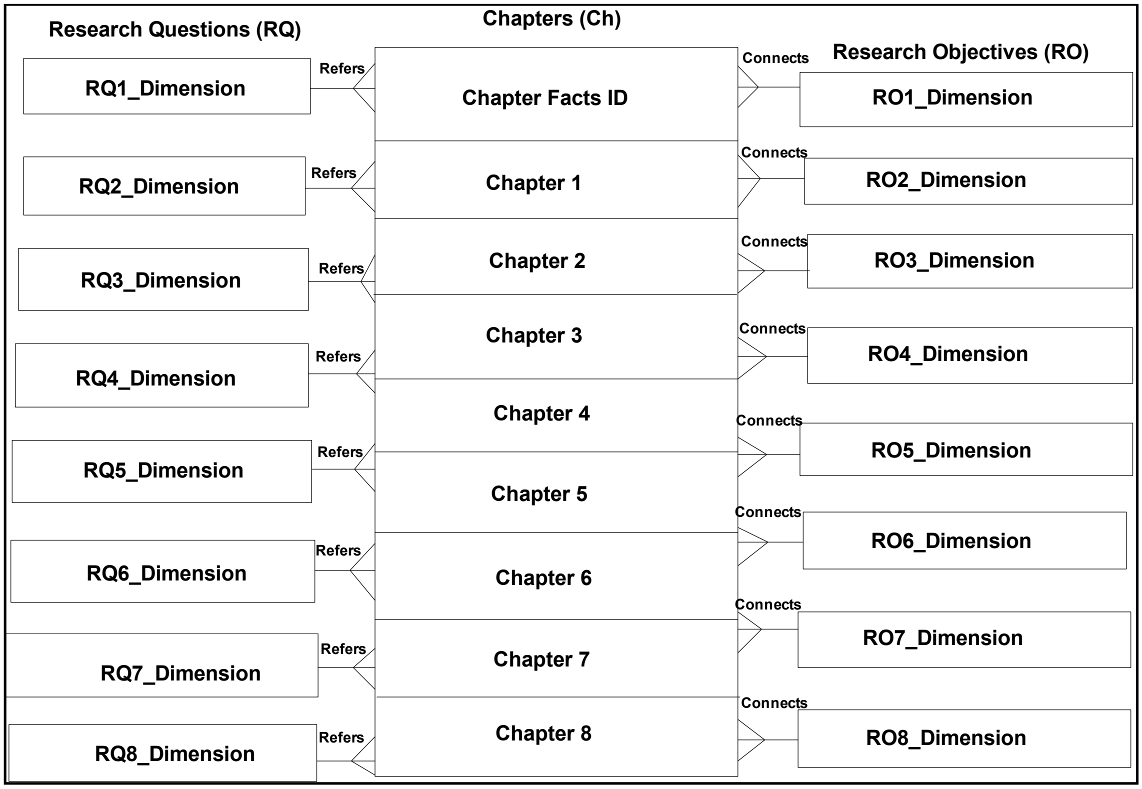


Figure 1.1: Establishing connections between research questions and research objectives

Chapter 2: The Literature Review

2.0 Introduction

The author refers to ISO15926 Oil & Gas Ontology focused in the plant/chemical/refinery specification in West (2006). West (2006) introduces 4 dimensionalism and ISO 15926, highlighting the ISO 15926-2 specification as an integration model that uses well defined metaphysics based on spatial-temporal extents, and is highly regarded as a fine ontological work. In Chapter 2, the author describes existing high level (or upper) ontologies, ontology mapping frameworks and procedures for standardising ontology models. In the current research work (Oil & Gas Exploration) ontology modelling ideas are brought from other domains where extensive ontologies are built in gene-technology, bio-informatics and software engineering applications. The author further reviews the literature, refereeing these contributions to contest and claim the research problems and objectives. The current literature, published and unpublished reports, volumes of data, both commercial and non-commercial nature, required for testing data models and evaluating systems' design and development of methodologies are used. The author examines ontological modelling in different domains, data models derived using conceptual models, data warehousing, data mining, data visualization and knowledge interpretation articulations. Existing data sources in the domain of interest are described in Section 2.1, describing their characterization in Section 2.2. Section 2.3 provides a review on ontology based data organization. The author gives a brief review on integration of multiple data sources in multiple domains in Section 2.4. Different data structuring methodologies are given in Section 2.5, describing data mining schemes in Section 2.6. The author describes data visualization and fusion in Section 2.7 with description of data interpretation and domain knowledge discovery aspects respectively in Sections 2.8 and 2.9. Domain knowledge of petroleum systems and digital ecosystems are described in Section 2.10 and Section 2.11. Systems' design and development in the domain of interest are described in Section 2.12. The author provides an anticipated research outcome and evaluation at the end of the chapter in Section 2.13.

Keeping in view the research problem and objectives (discussed in Chapter 1), the author carries out literature review on existing research methodologies. Neuman (2000) and Hevner et al. (2004) provide design science research methodologies, illustrating several quantitative and qualitative approaches. Vaishnavi and Kuechler

(2004, 2007) propose a specific design research that is distinguished from design by the production of (to a community) new domain knowledge. In a typical industry scenario, a new product or an integrated framework (artifact) is created, but in most cases, the more successful, the project is considered to be, the less is learned. It is generally desirable to produce a new product using state-of-practice application of state-of-practice techniques and readily available components. Engineering research or design science research as proposed by Hevner et al. (2004) is meant to collaborate with industry base artifact product designs, to fill the gap between the design research and system design. Hevner et al. (2004) provide an insight on the design science, substantiating the significance of industry base information systems' research. Venable et al. (2014) provide a framework for evaluation in design science (FEDS) providing various stages of design science and examples. Reinecke and Bernstein (2013) describes a design science approach to interfaces that automatically adopt to culture. The author finds these framework and design science constructs useful and extends their scope in the current research. Considering the nature of the research problem and its objectives, the author adopts the design science approach along with the guidelines published in the literature (Vaishnavi and Kuechler 2004, 2007).

Various research articles are explored on ontology modelling, data warehousing and mining for their feasibility and applicability in the business organisations. The author explores similar applications and scopes in the current research. Khatri et al. 2004 provide conceptual modelling using geo-spatial events. Guan and Zhu (2004), Uschold and Gruinger (1996) and O'Leary (2000) discuss concepts of ontology, principles of conceptual modelling and acquisition of ontology in various domains. Shanks et al. (2003, 2004), Hadzic and Chang (2005), Jasper and Uschold (1999) demonstrate ontology modelling, its validation, and semantic conceptualization in various industry applications. The author examines the existing concepts of data warehousing, data modelling procedures for data warehouse design and their applications in the oil and gas industry scenarios in Gornik (2002), Hoffer et al. (2005), Nimmagadda and Rudra (2005a), Nimmagadda and Rudra (2005b) and Nimmagadda and Rudra (2005c). Rudra and Nimmagadda (2005) investigate issues of multidimensionality and granularity of data structures in the oil and gas exploration business applications that are surrounded by heterogeneities. Nimmagadda and Rudra (2005a) examine the object oriented modelling approach in mapping the complex oil and gas business data entities described at various operational units. Pujari (2002) and Dunham (2003) provide data mining technologies, different computing algorithms and applications in various industry situations. They describe how data views are extracted through mining algorithms for an effective interpretation. Gornik (2002) discusses an industrial application of an airline reservation and control systems in a warehouse environment. Biswas et al. (1995), Cheung et al. (2000), Guha et al. (1998), Huang (1997), Matsuzawa and Fukuda (2000), Ng and Han (1994), Pei et al. (2000), Ramkumar and Swami (1998), Zhong et al. (1996), Yao and Zhong (2000) and Yun and Chen (2000) describe several data mining techniques such as clustering, associative rule mining and construction classifiers using the decision trees. Often oil and gas data are represented in the spatial-temporal forms and geographic information systems. Miller and Han (2001), Ott and Swiaczny (2001) and Zhou et al. (1996) illustrate the use of spatial-temporal datasets, organize them in a warehouse environment and explore these data using data mining procedures. Marakas (2003) provides data visualization concepts and practical applications, how the visualization can explore a hidden knowledge in multiple knowledge domains. Mattison (1996) discusses a case study of an oil and gas exploration, with an application to data visualization technique. Thomas et al. (2006) describe a warehouse design with bioinformatics focus.

Several published constructs, models and methods discussed so far are helpful in building the new constructs and models in the oil & gas domain. Literature on each component of the research questions, objectives and the problem statement is highlighted (in Sections 1.3.1 and 1.3.2 in Chapter 1) addressing issues with the existing constructs and models.

2.1 Description of Data Sources

The author describes data from multiple sources and application domains (see Appendix - 2). Oil and gas sources comprise of largely functional entities, such as *exploration*, *drilling*, *production*, *technical* and *marketing* including *human resources*, *finance* and other *project services* (Nimmagadda and Rudra 2004 and 2005). The operational, updated operational, archived, external either structured or unstructured, published and unpublished reports are other data sources. The operational entities include, strategic planning, management control and operational controls. The entities or objects described in the data sources are largely considered as dimensions. Though data characteristics are different in different domains, the author considers oil and gas data from heterogeneous and multidimensional sources, for testing and validating the constructs, models and methodologies in multiple domains. The objectives of the current research, discussed in Chapter 1, hold good in any domain.

Heterogeneous data sources: As described in the Appendix-2, the author acquires the existing data sources from different oil and gas and mining companies. *Heterogeneous data* are data from any number of sources, largely unknown and unlimited, and in many varying formats. As characterized in Figure 2.1 and Figure 2.2, the current data sources are widespread in different geographic regions, though the systems associated with these data sources are inherently interconnected, but largely unknown. Data sources, shown in pictorial views in Figure 2.1 are identified in digital form. In the oil and gas domain, sources of heterogeneous and multidimensional data are largely drawn from areas of *sedimentary basins* (defined in a glossary, see Appendix-1), as displayed in Figure 2.1.

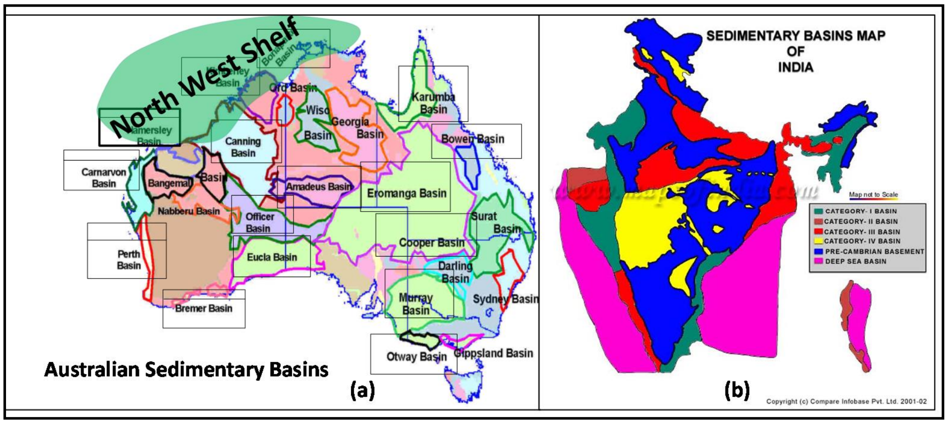


Figure 2.1a: Heterogeneous and Multidimensional data sources: (a) Australia; (b) India

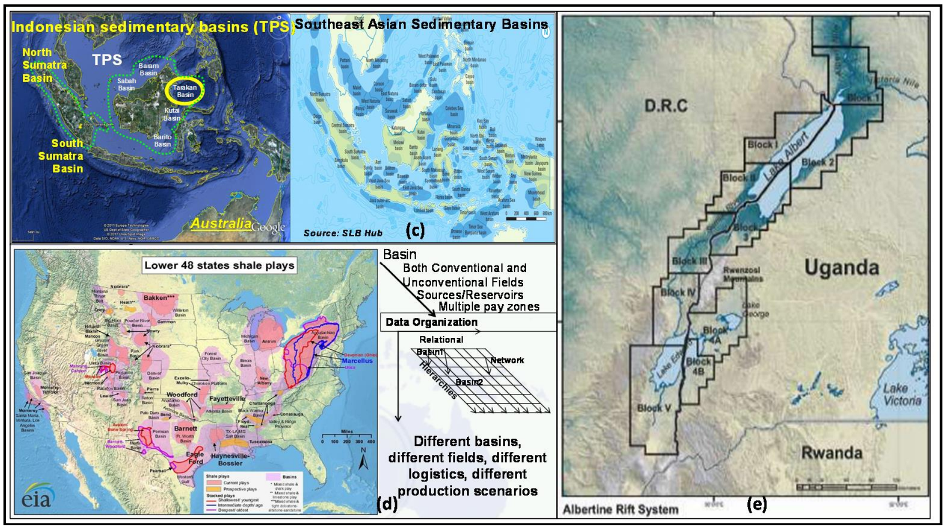


Figure 2.1b: Data from: (c) Indonesia; (d) USA; (e) Uganda

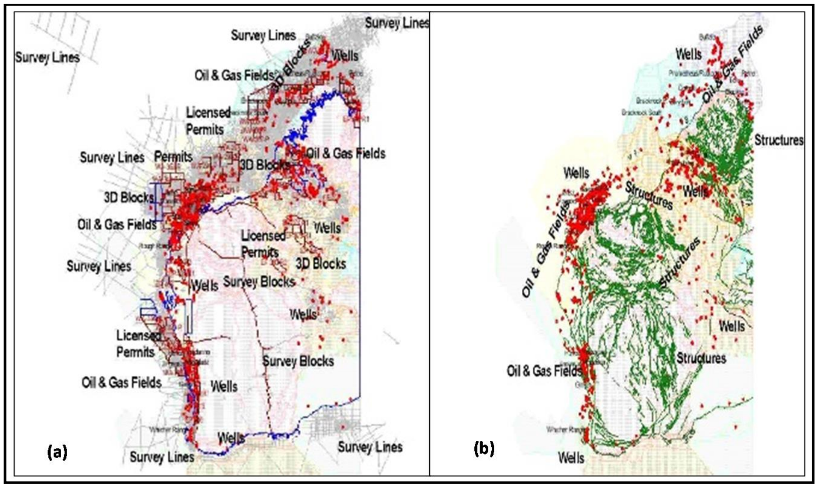


Figure 2.1c: Western Australian data: (a) *surveys, wells*, *permits* and *oil fields* and (b): oil & gas fields associated with geological *structures*

Multidimensional data sources: The term *multidimensional* tends to be applied only to datasets with three or more dimensions. The multidimensional data that originate from external sources are either from geographically or periodically varying multiple dimensions of unknown boundary conditions (similar to unknown extents of a petroleum system) and formats. They are said to be modelled as multidimensional. Here multidimensionality is viewed as addressing the complexity. The word dimension forms the root of multidimensional. For example, "how good the petroleum system that can produce commercial quantities of oil & gas", the author interprets the complexity of a system, how its' *elements*, *processes* and *chains* are in nature geographically distributed.

The exploration data are collected from Australian Bureau of Statistics from the mineral and oil and gas exploration quarterly censuses' sources. These publications contain actual, expected oil and gas exploration costs, onshore, offshore drilling costs including production lease and other costs data instances for analyzing production trends. As shown in Figure 2.2, the data sources generate volumes of oil and gas data. Often these data are in spatial-temporal form. The spatial data represent X, Y, Z coordinates (Cartesian coordinates represented in distance, meters) and the historical data are in time periods (years, months and days).

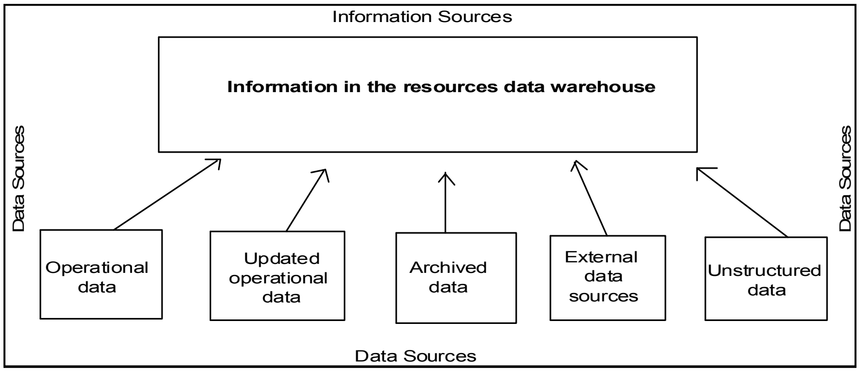


Figure 2.2: Oil and gas data sources

All the historical exploration data are from *geological*, *geophysical* and *geochemical surveys.* They are carried out in different prospective areas of different basins at different periods. The surface and sub-surface geologically mapped data, are key geological data. The *structure*, *stratigraphy* and other *geological* (defined in glossary in Appendix-1) data have significance in modelling the oil and gas data sources. The *biostratigraphy*, *coring*, *sedimentological*, *geo-chemical* and reservoir data are important data in ascertaining the oil and gas potentiality of a basin. The *reservoir, source, seal, migration, structure* are other critical elements of a *petroleum system (*as defined in glossary in Appendix-1*)*. The *navigational*, seismic or geological depth structure, well-logging, vertical seismic profiling (*VSP, vertical seismic profiling,* connecting composited dimension for the surface and sub-surface exploration data) and the reservoir are key data for interpreting the hydrocarbon prospects and evaluating them in every basin. Historical oil and gas data include several decades of oil and gas exploration data, drilling data, well data, original exploratory and development oil/gas discoveries, permits of oil and gas licenses and the production data of different wells, for different oil fields and basins and their descriptions. These data signify both the longitudinal and lateral dimensions and hierarchies. The additional dimensional attributes (which are deduced in various conceptual knowledge domains) have immense future scope of data mining, which adds the value in knowledge domains. The conceptual models, which form the basis of any logical design of multidimensional data model, make-up an integrated warehouse metadata model. The data that describe different sources, as illustrated in Figures 2.1 and 2.2, are:

- The data in the operational units of the resources industry are usually scattered throughout the enterprise. They are captured and stored on disparate systems using incompatible file formats and accessed only by using different application software. Mostly, the resources data in operational systems is not integrated.

Little attention paid to systems' interoperation, when individual operational sub-systems were created. Consequently the same information may be coded differently in different systems. The navigational data in the oil and gas industry may have been represented in the Cartesian coordinates in one database while the polar-radian coordinates in another.

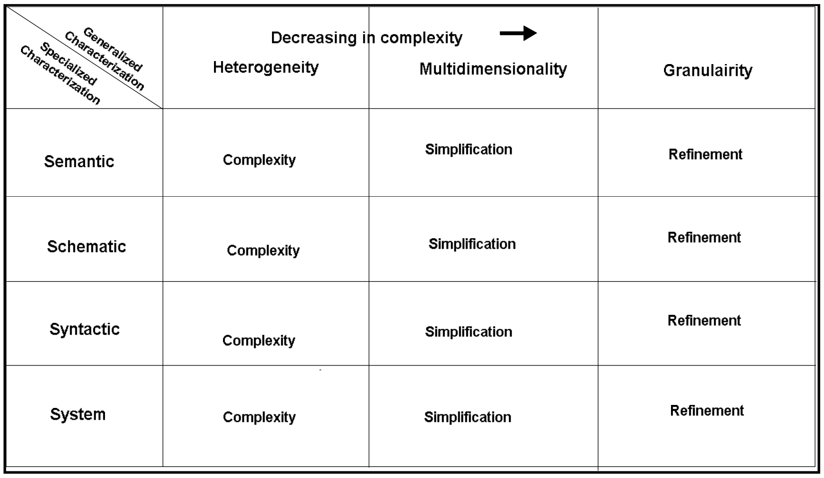
- The second data source for the data warehouse is from oil and gas data updates; e.g. operational data of drilling information is loaded into the data warehouse. The data warehouse is created, while a well in a particular basin is under drilling and the status of the well is not known. When the drilled well is completed and declared as a gas well, it should be updated periodically. Data updates often pose serious ongoing challenges for data warehouse architects and administrators, who devise ways to manipulate changes in the operational systems and data files without disrupting the operational systems. Moreover, the data entered must have time label to allow the users to ascertain the data's timeliness for analysis purposes. Dates of start and end of seismic data acquisition or start and end of well drilling etc. are examples of such operations.
- Third source of data for the data warehouse is archived oil and gas data, which is often generated by operational systems and moved offline. Companies often archive data for a specified period of time. Archived data could present data integration problems as variations may have occurred subsequently in the oil and gas data, when archived data may have been captured over time. For example, exploration operations may have been out-sourced over a period of time as a part of re-structuring process in a company, in which different lines of information system must have been designed for an exploration business.
- The fourth data source for the data warehouse is external information. Many exploration service companies purchase lot of exploration data and other information from the third party vendors. Integrating third party data with operational and archived data forms a more complete picture of the organization's exploration data. Although Internet makes an enormous amount of oil and gas resources information accessible to analysts and decision makers, the information is even more valuable if it is combined with proprietary, internal resources data.
- Enormous volumes of unstructured oil and gas data that reside on hard disks or hard copies throughout the oil and gas industry represent another data source. With prevalence of word processing, most significant company documents, including decision memos, strategic short notes on drilling or

survey acquisition proposals and the miscellaneous correspondence, are stored in electronic formats.

2.2 Data Characterization

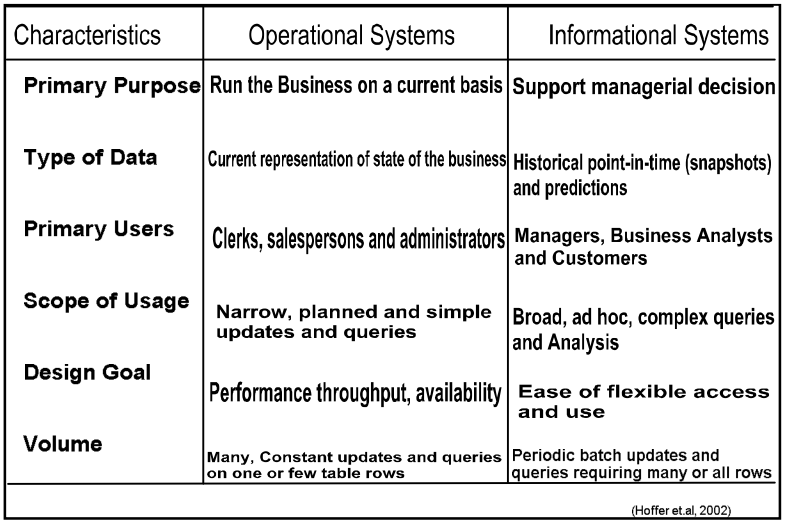
The author characterizes the data of existing data sources in this section. An entity is an identifiable existence, which can also be an abstract. Data records and their attributes describe their entities. An object is a visible and another tangible existence, but survives on the encapsulation and inheritance properties (Hoffer et al. 2005 and Coronel et al. 2011). Objects influence on the set of instructions or methods about what to do with data records. In the case of a dimension, it is a measure of identifiable existence or even an emerged conceptualized entity or dimension. The notion of a dimension provides lot of semantic information especially among hierarchies of elements. Here element is an individual existence of a composite entity. The entities, objects and dimensions, used in the ontology modelling process, are reusable in multiple domains. For example, a *reservoir* object (Nimmagadda et al. 2005c) that exists in a petroleum system domain, can be reused in multiple applications of petroleum production and reservoir engineering applications. Similar is the case with *reservoir* dimension, which quantifies or qualifies within a generic petroleum system domain, reusable in other domain applications.

In the context of digital ecosystems, entities, objects and dimensions and their attributes (that play significant roles in connecting multiple domains) are often associative (Coronel et al. 2011).The author describes a matrix of generalized and specialized data characterization (as per author's industries' interaction and experience), as tabulated in the Figure 2.3.



Generalized Characterization Specialized Characterization	Decreasing in complexity →		
	Heterogeneity	Multidimensionality	Granulairity
Semantic	Complexity	Simplification	Refinement
Schematic	Complexity	Simplification	Refinement
Syntactic	Complexity	Simplification	Refinement
System	Complexity	Simplification	Refinement

Figure 2.3a: Data characterization, describing the complexity



Characteristics	Operational Systems	Informational Systems
Primary Purpose	Run the Business on a current basis	Support managerial decision
Type of Data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary Users	Clerks, salespersons and administrators	Managers, Business Analysts and Customers
Scope of Usage	Narrow, planned and simple updates and queries	Broad, ad hoc, complex queries and Analysis
Design Goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, Constant updates and queries on one or few table rows	Periodic batch updates and queries requiring many or all rows

(Hoffer et.al, 2002)

Figure 2.3b: Data characterization in operational and informational systems

The data sources comprising of entities, objects and dimensions are used in building ontologies. Ontologies are described as per the characterization of data sources, their content and in which contexts they are used, as represented in Figure 2.3. Ontologies, as formal models of representation with explicitly defined concepts and named relationships linking them, are used to address the issue of *semantic heterogeneity* of data sources. In domains like *petro-informatics*, *earthquake prediction* (Nimmagadda and Dreher 2007c), *bioinformatics* (Sidhu et al. 2005 and Lee et al. 2006) and *nutrition-informatics, food-health* (Nimmagadda and Dreher 2011c), *human-anatomy* (Nimmagadda and Dreher 2008c), ecosystems (Nimmagadda and Dreher 2012a), the rapid development, adoption and public availability of ontologies (Shanks et al. 2003) have made it possible for the *data integration* community to leverage them for *semantic integration* of data and information in oil and gas domain.

2.3 Ontological Descriptions

In recent years, the research on ontologies has turned into an interdisciplinary subject (Jarrar 2005). It combines elements from different fields of Philosophy, Linguistics, Mathematics and Statistics. The author explores these opportunities and scope of research in domains of petroleum digital ecosystems, utilising the facts of capturing domain knowledge, existing independently in any domain application. An agreement on ontological content is the requirement in ontology modelling. The author focuses on the ontology requirement and development process in connecting groups of systems ultimately within a total petroleum ecosystem concept. The author takes advantage of facts of ontology reusability, task-independent ontology application, capturing semantics at domain level. Evolution of ontologies based on conceptual and epistemological changes (Jarrar 2005), scope extensions, large scale ontologies for heterogeneities and multidimensionality and quality improvements facilitate similar ontologies in the domain of research. Nimmagadda and Rudra (2004) and Nimmagadda and Dreher (2005) provide new insights on the ontological descriptions with respect to heterogeneity and multidimensionality.

Abdelali et al. (2003) discuss ontologies among cross disciplinary areas for the purpose of extracting new information retrieval. Similar cross disciplinary research is proposed, from which constructs and models are to be built from multiple domains. Aitken and Reid (2000) evaluate information obtained by ontological structures. They describe information retrieval tools. The author evaluates the data and information needed for building ontologies in the domain of interest. Bio-informatics is another domain in which ontologies are extensively investigated for quality of information and deriving new knowledge. Baker et al. (1998) describe ontologies applied in molecular biology. Bench-Capon and Malcolm (1999) formalise ontologies and their relationships for examining the interpretative information and knowledge. Bouquet et al. (2003) describe ontologies focusing on contextualised entities with OWL codes and their narrations. Similar OWL code are intended to be written for elements of petroleum system domain. For the purpose of acquiring new knowledge from the knowledge based systems, Bylander and Chandrasekaran (1988) attempt to provide knowledge base reasoning and abstraction. Clancey (1992) describes several model construction operators. Calvanese et al. (1998) investigate conceptual modelling and reasoning support systems for information integration and for quality information. The author explores options of developing conceptual models and integrate them through connectivity among attributes. Chandrasekaran et al. (1999) explore why at all ontologies needed in the data modelling and what exactly they contribute for quality information and with new knowledge, if any. The author intends to provide use of ontologies in the petroleum digital ecosystems' scenarios. Coronel et al. (2011) investigate database systems, their design, implementation and management and they provide several generic data models both in ER and Multidimensional diagrams. Relational database is intended to be used for documenting data instances from multiple domains. Ding and Fensel (2001) furnish ontology library systems for reuse supporting the concepts of interoperability. Degen et al. (2001) describe use of ontologies in information systems domain. They provide high level ontologies and their application in information systems. Demey et al. (2002) describe various business rules that support the interoperability and modelling systems.

Deridder and Wouters (2000) investigate ontologies support and its coupling between software models and their implementation. Embley (2005) examines the semantic web with an approach based systems for extracting ontologies. Elmasri and Navathe (1999) provide the basic fundamentals of database systems along with design, development and implementation. Franconi (2002) describes and provides tutorials on the logic based conceptual design for the purpose of information access and ontology integration. Frank (1997) demonstrates spatial ontologies with both spatial and temporal reasoning, which is geographical focused. Fonseca and Egenhofer (1999) provide the ontology-driven geographic information systems. The author considers this approach for analysing the geological and geophysical information systems. Gangemi et al. (2001a), Gangemi et al. (2001b) and Gangemi (2004) discuss ontology design aspects and ontology framework with an analysis for new information. These design aspects and framework in comparison with proposed framework, are followed up in the current domain of interest. Gilberg (1985) describes a schema methodology for large ER diagrams. For complex domains such as oil & gas attributes modelling, the author proposes similar larger ER and Multidimensional models and compare their use in complex petroleum ontologies.

Gomez-Perez and Benjamins (1999) insight knowledge sharing and reuse of components in problem solving methods. Similar reuse of ontologies in multiple domains of oil & gas exploration and prospecting is aimed at. Gruber (1995) describes principles and business rules in support of ontologies for knowledge sharing. Guarino (1994), Guarino and Giaretta (1995), Guarino (1997), Guarino (1998), Guarino and Welty (2000) and Guarino (2002) describe various aspects of ontology-driven conceptual modelling, connecting information systems and other knowledge based systems. Similar ontology driven conceptual modelling approach in the oil & gas domains is aimed at in the current research.

Guizzardi et al. (2002) use various ontological foundations for the UML conceptual modelling. UML modelling can map various objects, described in the oil & gas domain. This reference provides an analogy in the petroleum ontologies, finding similarities of gene-biology ontologies with respect to petroleum ontologies. Halpin (2001) describes the information modelling and relational databases. Van Heijst et al. (1997) provide explicitly knowledge based ontologies for information analysis. The author finds these ontologies, compatible to petroleum knowledge, access and share knowledge among many researchers. Jarrar (2005) and Jarrar and Meersman (2002) describe the methodological principles for an ontology engineering project, including scalability and knowledge reusability in ontology modelling and the author proposes to discuss the scalability and knowledge reusability in the current domain of interest. Keet (2004) describes various aspects of ontologies and their integration in exploiting new information and knowledge, to further follow up the integration of domain ontologies in the proposed integrated framework.

Meersman (1999), Meersman (2000), Meersman (2001) and Meersman (2004) describe semantic ontology tools in information systems' design with a special focus on databases. The author finds the Meersman's work inspirational and propose to apply similar ideas in the proposed domains of interest (with respect to achieving the goals of research objectives). Musen (1998) provides domain ontologies using Protégé with the EON Architecture. The author proposes to use a different architecture but supported by Protégé with various other graphical tools such as grapher and surfer solutions, used for drawing ER and Multidimensional diagrams. The grapher and surfer tools are intended to be used for mapping data views extracted from warehoused metadata. Richards (2000) uses and reuses the ontologies for knowledge and quality information and similar approach is intended for reusing the ontologies in various situations of the petroleum domain. Roberto et al. (1998) use ontology tools for representing geographic systems and other information systems. The author uses geographically based spatial-temporal dimensions in describing ontology models. Smith (2005) uses ontologies in information systems arena. Author intends to follow-up information systems' development using ontologies and their tools. Shoval (1985) provides several essential information structure diagrams narrating them with database schemas. Author proposes building several schemas that represent relational and hierarchical databases. Steels (1993) describes a computational framework and its reusability and author intends to develop an integrated framework in which domain ontologies are meant to be integrated to generate a metadata in the domain of interest. Siitola (1996) furnishes several large ER diagrams with illustrations. Similar large ER diagrams are intended in the current research, for supporting the ontologies and their integration. Suárez-Figueroa et al. (2005) identify standards for narrating a metadata built based on ontologies. The author intends to work and follow up standardization of ontologies in the oil & gas domain, especially when handling the heterogeneous and multidimensional data attributes. Sycara et al. (2006) describe ontologies in different domains narrating semantics. The author uses similar ontology modelling concepts to apply and evaluate them in the oil & gas domain, especially addressing heterogeneities and multidimensionality. Spaccapietra and Parent (1994) integrate the data views for understanding any structural conflicts and similar integration approach is aimed at for a company's metadata.

Uschold and Gruninger (1996) compile ontologies narrating principles, methods and applications. The author refers these ontology principles, applications and finds useful for narrating ontologies in oil & gas domain. Vermeir (1983) designs conceptual schemas and semantic hierarchies and author intends to build similar conceptual schemas in support of building ontologies in oil & gas domains. Ventrone and Heiler (1991) describe semantic heterogeneities as domain of interest becomes complex. Author too proposes to describe and address the heterogeneities in oil and gas domain. Wiederhold (1995) provides value added large scale information systems. The author finds this approach compatible to the domain of interest and author has chosen to analyse it in detail. Weinstein (1998) uses ontology based metadata for extracting information and new knowledge. Similar approach is intended to be used for generating metadata and extracting new knowledge in oil & gas domain. Welty and Ferrucci (1999a), Welty and Jessica (1999b) and Welty (2002) provide ontologies for reusing in software documents. The reuse of ontology models in various situations of oil & gas domain is the current focus. Wand et al. (1999) analyse the ontologies for building relationship constructs in a conceptual modelling and similar approach in building relationships among multidimensional oil & gas metadata is aimed at in the current research. The existing literature on various components of research methodologies is reviewed in the following sections.

Coronel et al. (2011) furnish interpreted data dimensions and structured in different domains. They cite various examples that represented in the ER modelling process. Pujari (2002) describes characteristics of different dimensions that go in the multidimensional modelling process. The author uses various algorithms adopted by Pujari (2002). Nimmagadda and Rudra (2004, 2005) describe several entities, objects and dimensions, used in oil and gas domain. Noy and McGuinness (2000) and O'Leary (2000) provide basic literature on ontology development and how different firms use different ontologies in different business environments. Guan and Zhu (2004) describe basic ontology acquisitions for developing agent tools. Hadzic and Chang (2005) provide ontologies in support of human diseases and healthcare domain. Sidhu et al. (2005) describe ontologies on protein data models. The author takes on ontology descriptions in other domains. Jasper and Uschold (1999) describe ontologies within an integrated framework. Nimmagadda and Rudra (2004) describe applicability of the data warehousing and data mining technologies in the Australian resources industry. They describe various conceptual models, feasible in the oil and gas industry.

Gilbert et al. (2004) provide data integration methodologies, for which scales of data sources for oil and gas reservoirs are different. Sasson and Blomgren (2011) highlight different dimensions that involved in the integration of oil and gas sources with a focus on the geo-statistics. The literature on the integration of domain ontologies for oil and gas business data sources though are limited, but the significance of data integration is emphasized in many oil and gas industries' technical manuals. The author intends to use the classical statistical mining tools and methods to analyse and interpret the data views of oil and gas metadata.

2.4 Data Structuring Methods

The data structures with various ER focus are explained in Gornik (2002) to intelligently integrate and store the data sources in a warehouse environment. He provides industry based solutions for aviation and transport services. Different data structures are given in Hoffer et al. (2005) and Pujari (2002) designed with a focus on effective data mining. Their studies are mere focus on academic rather than on any commercial interest. Hoffman (2003) designs the conventional databases for geoscience professionals. Ozakarahan (1990) provides different structuring models and methodologies, with interpretation of business rules for implementation. Ott and Swiaczny (2001) describe the geographic information systems using the spatial-temporal data. The author exploits the information given on the analysis of spatial-temporal data and their structures.

Ozkarahan (1990), Hoffer et al. (2005) and Pujari (2002) provide ER models, describing their entities and dimensions. They deduce in various application domains including educational institutions and industries, such as aviation, transport and software engineering domains. But there is no focus on spatial-temporal dimensions. Nimmagadda and Dreher (2004) and Nimmagadda and Dreher (2005) provide various entity relationship models for the resources industries. Chaudhri (1993) and Vadarparty (1996) use the data sources as objects in the modelling process. Atkinson et al. (2003) describe the golden rules for generating the object oriented database. They provide a step by step procedure for generating the databases. Opdahl et al. (2001) use the object oriented models in their data relationships analysis. It is challenging to find literature using oil and gas data sources for the object oriented modelling. Hoffer et al (2005) and Pujari (2002) provide several data models using entities, objects and dimensions in different domains. Nimmagadda et al. (2005c) have initiated the dimensional modelling in the oil and gas domain including mineral industries in Australia. The author takes advantage of the models deduced by previous researchers and the guidance for generating new and innovative data structures.

2.5 Data Warehousing and Mining Techniques

Jukic and Lang (2004) use offshore resources to develop and support data warehousing applications. Anahory and Murray (1997) describe data warehousing in real world situations. Gornic (2000), Hoffer et al. (2005) and Nimmagadda and Dreher (2005) describe the concepts of data warehousing, data modelling procedures for a data warehouse design, and their applications in different industry scenarios. They investigate issues of database structuring methodologies, multidimensionality and granularity of the data structures in various business applications. Berson and Smith (2004) provide data warehousing solutions with data mining and OLAP applications. Miller et al. (2002) review the medical data in a warehouse environment and a strategic information management.

Aldenderfer and Bashfield (1984) use the cluster analysis in quantitative applications. Biswas et al. (1995) have explained the cluster mining methodologies for domain knowledge in the petroleum domain. They use traditional databases for organizing the data sources. Cheung et al. (2000) describe density based data patterns for extracting the mining rules. Dunham (2003) provide various data mining tools including classical statistical mining. Frietas (2002) reviews various algorithms for data mining and knowledge discovery. Kaufman and Rousseeuw (1990) analyze different clusters for mining groups of data. Guha et al. (1998) provide cluster mining algorithms for large databases. Taniar et al. (2008) use associative rule mining for classifications and categorizations in the data sources. Tjioe and Taniar (2004) also use associative rule mining techniques for exploring the data warehouses and metadata. Huang (1997), Matsuzawa and Fukuda (2000), Ng and Han (1994), Pie et al. (2000), RamKumar and Swami (1998), Zhong et al. (1996), Yao and Zhong (2000) and Yun and Chen (2000) describe several data mining schemes such as clustering, associative rule mining and construction of classifiers using the decision tree structures. Often petroleum data sources are characterized in spatial-temporal formats and Miller and Han (2001), Ott and Swiaczny (2001) and Zhou et al. (1996) illustrate the use of spatial-temporal datasets and how they are organized in a warehousing environment. The author tracks spatial-temporal dimensions in the multidimensional models to them accommodate within an integrated warehouse environment.

2.6 Data Visualization and Fusion

Erdmann and Rudi (2001) provide data view structures with ontology focus. For transmission of data views geographically, they are structured in XML codes. Krishnamurthy (1999) describes various visualization techniques for representing data for interpretation. Cleveland (1994) describes various graphic tools for representing the data elements and their analysis. Han and Cercone (2000) use a visualization system for interpreting the association rules in the mining schemes.

2.7 Data Interpretation and Knowledge Mapping

The data acquisition, data processing and data interpretation are sequence of events described in any exploration project. For data analysis and new knowledge building in the oil & gas domain, the author uses a specialized data interpretation skills on windows and UNIX based workstations, whereas the formal interpretation is to explain and general intent of any subject matter. The data interpretation is routinely done in the exploration and field development stages of any oil & gas project. The current literature is sufficient enough to interpret domain knowledge and its use in the oil and gas industries. Though operating and service companies have their own proprietary interactive interpretation software, the interpretation methodologies are either developed based on user needs or with problem solutions. But still, qualitative and quantitative interpretation methodologies are popularly used in many domains including the oil and gas exploration domain. The data interpretation is a specialled skill either on windows or UNIX based workstations in the exploration project, whereas formal interpretation is to explain and intent of any subject matter. Maedche et al. (2002) provide ontology mapping procedures and a framework for analysing and obtaining new knowledge. The author intends to work on similar ontology mapping and framework for obtaining new knowledge. King (2000) explores warehoused data and implements data views for strategic knowledge management. In the oil and gas industries, especially exploration industries, “domain knowledge” is commonly used during interpretation of *elements*, *processes* and *chains* of petroleum systems and their ecosystems connectivity. Moving forward from the existing knowledge to new knowledge in new dimensions and domains, is a key criteria of the current research.

The author uses the basic ideas of data warehousing, mining, visualization and interpretation techniques.

2.8 Domain Applications

The term "domain knowledge" in the field of petroleum ecosystems and oil and gas industries is not commonly used, in spite of that several domain applications are involved in the oil and gas business environments. Though literature is available in this context in public domain, but it is highly commercialized.

Petroleum digital ecosystems (PDE)

The author examines literature on petroleum systems and their relevance in exploration and production environments from Magoom and Dow (1994). Longley et al. (2001) illustrate 21st century's frontier areas of petroleum prospects in Australian sedimentary basins. Telford et al. (1998) investigate critically the significance of geological, geophysical and geochemical exploration and prospecting for petroleum deposits in different geological environments. They examine types of oilplay and analysis. Huston et al. (2003) and Gilbert et al. (2004) discuss several issues of exploring and exploiting prospective reservoirs using data integration techniques. Several issues on the reservoir, structural and strati-structural plays under different geological settings are described. Dodds and Fletcher (2004) describe issues of risks involved in drilling the explored, under-explored and detailed exploration areas. Hori and Ohashi (2005) and Erdman and Rudi (2001) demonstrate the use of XML technologies for preparing and delivering the explored data through Internet. Weimer and Davis (1995) discuss different petroleum systems of USA, Middle-East, Russia and South America narrating stratigraphic and structural oilplays. Johnston (2004) discusses issues of time-lapse 4-D technology that help minimize drilling dry holes and the reservoir model uncertainty. Aside from brief discussions on ontology issues in oil companies Meersman (2004), there is no concrete literature available on ontology application in the upstream oil and gas industries as on today, especially in the contexts of digital ecosystems in petroleum domains.

Research objectives laid in Section 1.3.2 are followed up by the author to design and develop the digital ecosystems in the oil and gas domain. As demonstrated in Figure 2.1, several sedimentary basins described in Australia, India, Indonesia, Uganda, Middle East and USA are considered in the current research, keeping in view the intricacies and complexities of data sources in these regions. Matured oil and gas fields of Middle Eastern basins, Indian peninsula, Southeast Asian and Australian basins have a scope of analysing a huge volumes of heterogeneous and multidimensional data.

Australian sedimentary basins**:** The sedimentary basins possess heterogeneous and multidimensional data sources as displayed in Figure 2.1 and Figure 2.2 in digital form. Integrated framework and workflows explore the usage of exploration and production datasets and their integration, for risk minimizing the oil and gas business in Australia. *Super Westralian basin*, a total petroleum system (TPS), such as North West Shelf (NWS) possess *shelf*, *slope* and *deep* geological events, which appear to have a connectivity through phenomena of a digital ecosystem. In addition, this super basin has multitude of sub-basins, each basin is associated with multiple petroleum systems, and each system with unknown or limited areal extents or boundaries. Each petroleum system contains multiple oil and gas fields, with hierarchical structuring of dimensions and their associated attributes. The Western Australia (Figure 2.1) possesses big heterogeneous and multidimensional data systems with a scope of analysing their dimensions, attributes and instances in a warehousing environment and thus explore for a mining opportunity. There is immense scope of analysing and integrating many *sedimentary basins* of Australian onshore and offshore basins, especially in the Western Australia, which produces more oil and gas deposits than elsewhere in Australia. Multiple basins, petroleum systems and hundreds of oil and gas fields are located wide across many geographic regions of Australia. The author proposes various data mining, visualization and interpretation schemes for exploiting the untapped reservoirs in these basins.

Arab Gulf basins: The petroleum digital ecosystems and their embedded systems are described in the context of Arabian Gulf Basins (Middle Eastern onshore and offshore basins, discussed in the following sections), demonstrating the necessity of ontology modelling in the integrated workflows and their implementation in these *basins*. There is scope of specification of conceptualization and contextualization analysis in modelling and integrating multidimensional and heterogeneous data sources in Arab Gulf basins. Ecology, petroleum system and geomorphic systems cannot be isolated, which are otherwise inherently embedded, demonstrating within an ecosystem, with multiple systems' connections. For this purpose, an integrated methodology is proposed in gulf basins that enable to understand the ecosystem phenomena through interconnected multiple digital ecosystems.

Indonesian basins: Indonesia is an island country with more than 30000 islands, with scattered *sedimentary basins* (Figure 2.1b), with huge geographic regions and areal extents that provide volumes of multidimensional and heterogeneous data sources. Indonesian petroleum ontology (PO) descriptions are intended to be made good use of representing oil and gas data sources of sedimentary basins that facilitate integration of multiple systems in different knowledge and application domains. These descriptions can facilitate building digital oil field solutions in complex geological settings.

***East African Rift System*:** Several data sources exist within East African rifted basins. As a part of demonstrating the concept of petroleum digital ecosystem (PDE), the author identifies ontology based data warehousing associated with multiple petroleum systems in the East African Rift system, in the context of Albertine Graben (Figure 2.1b, located in the Western Part of Uganda). The Albertine Graben is considered to be a super basin, with sub-basins narrated within this super-basin. There are several sub-basins within this super-basin (super-type rifted graben dimension), each sub-basin consists of multiple petroleum systems and each system possesses multiple oil and gas fields. This whole basin concept is simulated to an ecosystem, in which all petroleum systems and their embedded oil and gas fields have a connectivity. Volumes of datasets acquired in these basins, facilitate the demonstration and representation of emerging petroleum ontologies (PO) in these *rift* systems. These emerging concepts have further scope of analysing the large number of productive basins in the entire east African rift system, starting from southern Sudan in the north to Malawi rift in the south-eastern parts of the Uganda.

Unconventional Energy Scenarios in USA: Several shale gas projects are developed (Figure 2.1b) and implemented in recent years, in the southern parts of USA to meet the demand and supply of energy resource in these regions. Because of growing demand of energy sources and the steady depletion of the current conventional oil and gas resources, there is an increasing demand of the unconventional oil and gas business worldwide. Shale gas is the future energy source, when the conventional gas energy resource gets depleted and or exhausted. The shale gas occurs within fractured shales. The data sources associated with the fractured shale (one of the unconventional sources) do exist in many company situations, but are still in awful state and an integrated framework is needed to organize and document these data and information. Problems associated with the drilling and production, especially in the fracture development areas, may be resolved, if fracture mechanism is well understood with associated sub-surface lithologies (geological sense) and their connectivity.

Author's Proposal: As per the design science guidelines laid in Hevner et al. (2004) and Weber (2010), the author proposes an integrated methodology that can connect and integrate multiple conventional and unconventional reservoir ecosystems. Using the concepts of ecosystem and embedded ecosystems, for example, the connectivity is explored among the fracture networks of reservoirs from multiple fractured shale systems. The well placement and oil and gas production around the fracture development areas are to be assessed using the fracture networks and their modelling. Within the petroleum digital ecosystem scenario, the author emphasises that all the energy resources exist within a single petroleum ecosystem, in which shale associated geological & geophysical data events are made good use of exploring connections through integrated workflows. The fractures and their networks are intended to be modelled through integrated frameworks and workflows. The nomenclature associated with the content and meaning of the dimensions and their associated attributes of fracture networks are handled by semantic based shale ontologies (Nimmagadda and Dreher, 2011b and 2012). The North America (Figure 2.1b) and few other countries within Europe, are already exploiting the unconventional oil and gas deposit. There is a wide scope of unconventional oil and gas opportunities in South America, Asia and Australia regions, though the science and technology behind shale gas are not yet popular in many countries because of environmental concerns. Huge heterogeneous and multidimensional shale data sources though unexplored, but oil and gas explorers believe, there is an opportunity to exploit this energy source worldwide in large scale.

Digital oil field solutions

Longley et al. (2001) provide useful sources of petroleum provinces of Australia. Dodds and Fletcher (2004) use probabilistic approaches while making drilling decisions and their analysis using traditional methods of data organizations. They have not provided any digital oil field solutions for extracting knowledge from oil and gas data sources using data modelling and data warehousing approaches. The author proposes a research framework and methodologies for risk minimising the oil and gas exploration in upstream industries through digital oil field solutions taking advantage of volumes of data, published in Guoyu (2011).

2.9 Systems Design and Development

The systems design, development and interpretation are well established aspects in the context of petroleum domain. Analogous to information systems, petroleum systems have not been the focus of interpretation and analysis of information quality. The author emphasizes petroleum systems and their analysis in information systems' view point focusing on the information qualities and their backing to decision support systems that enable to provide a new knowledge.

Systems that deal with big-data

The author has had an opportunity in presenting the ideas on big-data in the IEEE Industry Informatics forums in Porto Alegre, Brazil recently. The author has presented big-data ideas in a PPDM symposium, organized by energy industry data managers' forum in Perth, Australia. Similar ideas are presented to the researchers in the School of Information Systems, CBS, Curtin University on several occasions. Accordingly, more ideas are added on big-data in respective sections providing reasons, why and how big-data added as new dimensions in the current research work. Clearly et al. (2012) provide business analytics guide with big-data focus. Dhar et al. (2014) explain the features of big-data in industry situations. Debortoli et al. (2014) compare big-data business skills in alignment with the features. Schermann et al. (2014) conduct an interdisciplinary research using big-data exploring further opportunities in the information systems' arena. The author makes use of these ideas and extend them in oil & gas domain, since so far, big-data associated systems, and their applicability and feasibility have not been exploited in the oil and gas domain. The future scope of big data to be exploited is in upstream oil and gas industry. In this industry, the big-data idea is surrounded by six significant Vs, such as "volume, variability, velocity, veracity, visualization and value" features. All the features are incorporated within the systems built based on big-data. The author focuses on designing a system, within which these features are incorporated in an integrated framework that constitute set of components for collecting, storing, processing big-data and for communicating good quality information to the service providers.

For the purpose of handling the size and magnitude of data (including geographical and periodical extents) and their sources, a system is necessary to leverage volumes of various heterogeneous and multidimensional oil and gas data sources. For example, hundreds of square kilometres of *onshore* and *offshore* data sources are widely spread among many countries (for example, Middle East) comprise of hundreds of trillion bytes of data. Keeping in view the heterogeneity, multidimensionality and granularity of these data sources, an integrated framework is needed to handle big-data for an effective data mining, visualization and interpretation to extract a new knowledge and add value. An ontology based data warehousing approach thus proposed by the author, is intended to be generalized and extended in other domains. The big-data (systems) appear to have definite role in knowledge representations in multiple domains and explore multiple connections among heterogeneous data sources. The “ecosystem and embedded ecosystem” ideas are intended to be developed in the contexts of sedimentary basins’ in areas, where no geological boundaries exist.

Systems that deal with turbulent business environments

The author emphasizes that like any other industry, upstream oil and gas too undergoes many turbulent business situations (Shastri and Dreher 2011a). The primary focus is on designing and developing such systems that can forecast future resources, needed in the oil and gas and mineral exploration businesses, especially in the Australasian countries. Now falling energy prices, disappearance of mining boom and recession of global markets cause great anxiety and concern to other industries. These eventualities need to be considered while designing new systems and building their data models. More details on turbulent business systems and analysis are given in the forthcoming chapters.

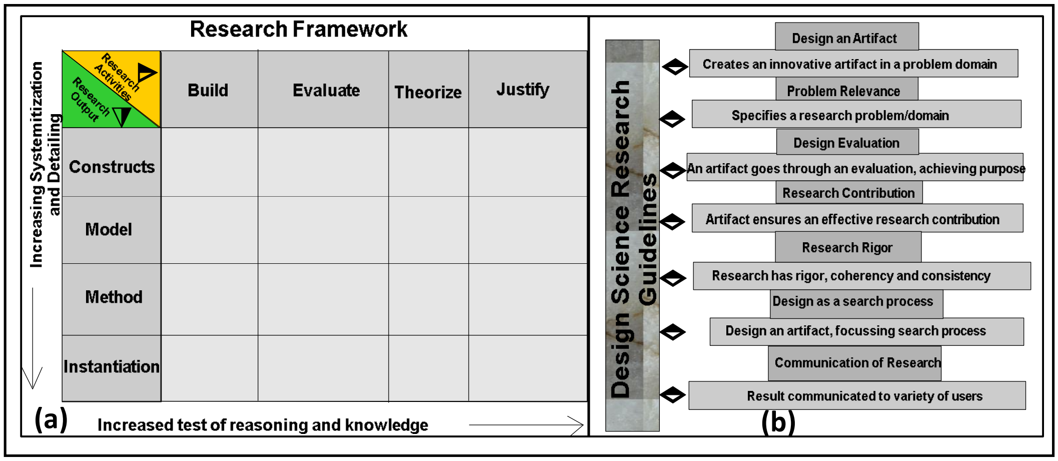
Conventional and Unconventional Digital ecosystems

Two types of reservoirs commonly hold commercial oil and gas deposits (Magoom and Dow 1994), such as *clastics* and *carbonates*. Both types of reservoirs exist within a single ecosystem. There are many giant oil and gas fields associated with the conventional petroleum systems worldwide. The volumes of data published (AAPG, SPE and Guoyu 2011) are good enough to do simulation of the reservoir systems using the concepts of petroleum digital ecosystems (PDE). Other types of reservoirs are “unconventional (mostly held by shales)” not uncommon in many sedimentary basins and among major producing companies. Most of the oil and gas production from these reservoirs, is from the North America. Because of want of huge logistics support for tapping this resource, there is already an existing market in the North America. This technology is slowly being transferred to other basins and other countries, where explorers intend the unconventional reservoirs). Currently the literature that piled up, is sufficient enough for simulating the both conventional and unconventional reservoirs, using the concepts of unconventional digital ecosystems (www.onepetro.org).

2.10 Anticipated Research Outcomes

The author anticipates that the constructs, models, methodologies and implementation representing integrated frameworks, as research deliverables in the current research work. In addition, several data, cross-plot and map views provide domain knowledge for interpretation.

The constructs or concepts (March and Smith 1995) are derived from different domains and contexts. They establish conceptualization and contextualization to describe problems within each domain and specify their solutions. The constructs are exceedingly formalized as in the semantic modelling formalism. The components of constructs are entities, attributes, relationships, identifiers and constraints. The conceptualization and contextualization are significant parts of constructs in the artifacts design in an integrated framework, which is driven by a design science research approach. The data warehouse is accommodated in an integrated framework with multidimensional data models, from which metadata are computed. The author ensures, these conceptualized constructs are well-formed logical and physical data structures, set by business rules and constraints. The author pursues the existing research framework described in March and Smith (1995) and Weber (2010) in the current study as described in Figure 2.4.


Research Framework

Research Activities / Research Output	Build	Evaluate	Theorize	Justify
Constructs				
Model				
Method				
Instantiation				

Increasing Systemitization and Detailing
(a)
Increased test of reasoning and knowledge
Design Science Research Guidelines
(b)

Design an Artifact
Creates an innovative artifact in a problem domain
Problem Relevance
Specifies a research problem/domain
Design Evaluation
An artifact goes through an evaluation, achieving purpose
Research Contribution
Artifact ensures an effective research contribution
Research Rigor
Research has rigor, coherency and consistency
Design as a search process
Design an artifact, focussing search process
Communication of Research
Result communicated to variety of users

Figure 2.4: (a) Existing research framework; (b) DS guidelines

Several data models represent different sources of knowledge domains. A model is a set of propositions or statements (March and Smith 1995) expressing relationships among constructs. In design activities, models represent situations as problem and solution statements. ER, EER, MR and MMR are set of constructs of data modelling formalism. A model is viewed simply as description and or graphical representation of how entities, objects and dimensions are presented and related each other in the modelling process. Semantic based data models are useful for designing information systems. The entities, objects and dimensions ensure their usefulness in the modelling process while representing and communicating with information system requirements.

A method is a set of steps (an algorithm or guideline) used to perform a task. The methods are based on a set of underlying constructs (language) and a representation (model) of the solution space (Vaishnavi and Kuechler 2004 and 2007). Although they may not be explicitly articulated, representations of tasks and results are intrinsic to methods. The methods can be tied to particular models, in that the steps taken as parts of the model are inputs. Further, the methods are often used to translate from one model or representation to another in the course of solving a problem. The data structures, for example, combine a representation of computer memory with algorithms to store and retrieve the data. The problem statement specifies the existing stored data and the data to be stored or retrieved. The method (algorithm) transforms this into a new specification of stored data (storage) or returns the requested data (retrieval). Many algorithms use tree-structured constructs to model the problem and its solution. The system development methods (Vaishnavi and Kuechler 2004 and 2007) facilitate the construction of a representation of user needs (expressed, for example, as problems, decisions, critical success factors, socio-technical and implementation factors, etc.). They further facilitate the transformation of user needs into system requirements (expressed in semantic data models, behaviour models, process flow models, etc.) and then into system specifications (expressed in database schemas, software modules, etc.), and finally into an implementation (expressed in physical data structures, programming language statements, etc.). These are further transformed into machine language instructions and bits stored on disks.

The design science creates the methodological tools that natural scientists use. The research methodologies prescribe appropriate ways to gather and analyse evidence to support (or refute) a posited theory (Neuman 2000, Vaishnavi and Kuechler 2004 and 2007). They are human created artifacts that have value in so far as they address this task. The design science validates the research framework, corroborating the guidelines as described in Figure 2.4b. An instantiation is a realization of an artifact in its environment (March and Smith 1995). Based on the secondary data, the models are tested and evaluated in commercial organizations. IT research instantiates both specific information systems and tools that address various aspect of designing information systems. Instantiations operationalize constructs, models, and methods. However, an instantiation actually precedes the complete articulation of its underlying constructs, models, and methods. That is, an IT system may be instantiated out of necessity, using intuition and experience. Only as it is studied and used, the author is able to formalize the constructs, models, and methods on which it is based.

At data modelling stage, the author limits the current research to description of high level ontologies in the oil & gas domain. The author agrees, the data integration came in to picture, because of heterogeneity and multidimensionality of data sources. As per the literature review done so far, the term "Ontologies" appears abstruse in the sense that its usage in different domains is different within different contexts. However, the concepts used by other domain experts in other domains seem convincing that their use in oil & gas domain is plausible. The author provides more details on "ontology descriptions in current domain of interest" in section 3.4 of Chapter 3. An attempt is made to analyse the concepts of ontologies in the context of oil & gas domains, since the data sources are heterogeneous and have numerous multiple dimensions and entities. The *exploration, drilling, production, technical, marketing* and *logistics* are subtype entities or dimensions of an upstream oil & gas company. At present, the connectivity is missing among these major entities and or dimensions. The author explores the application of ontologies (in the form of ER and Multidimensional articulates) in connecting these dimensions through known and unknown conceptualized attributes with relationship diagrams to bring out new domain knowledge.

2.11 Summary

The relevant literature review in this section covers on all topics of the research components and research objectives, as given in Sections of 1.3.1 and 1.3.2 within Chapter 1. The existing literature covers areas of domain, data modelling, schema, data warehouse, data mining, and visualization and interpretation components of an integrated framework. The author describes the design science research, research framework and methodologies in Chapter 3.

Chapter 3: The Research Framework and Methodologies

3.0 Introduction

The aim of this chapter is to describe a research framework, development of methodologies and their components, especially to address the research questions (RQ1 to RQ6) and achieve the objectives (RO1 to RO6), as narrated in Section 1.3.1 and Section 1.3.2 of Chapter 1. The author aims at validating the research objectives by design science guidelines (March and Smith 1995, Hevner et al. 2004 and Indulska and Recker 2008). Design Science (DS) is an evolving problem solving process. This chapter describes design science guidelines that support research methodologies. The author provides research paradigm and framework in Sections 3.1 and 3.2, discussing generic formulations needed for equipping integrated framework, connecting ecosystems in Sections 3.3 – 3.5 and describing components of integrated framework in Section 3.6. Simulation of an integrated framework is explained in the context of a digital ecosystem in the Section 3.7 (addressing RQ7 and RO7). Before articulating the constructs and models, the author reviews briefly the existing relevant literature on design science research with substantiative reasoning of adopting this research paradigm in the current study.

Churchill (1979) describes a research paradigm and research procedures for developing better measures for marketing constructs. This article outlines a procedure to develop better measures of marketing variables. The framework represents an attempt to unify and bring together in one place the scattered bits of information on how one goes about developing improved measures and how one assesses the quality of measures that have been advanced. Mintu et al. (1994) suggest Churchill's extended research paradigm and methodologies among cross-cultural constructs, addressing the issue of equivalence. With the growing cross-cultural research, findings of Mintu et al. (1994) are made valid, reliable ensuring new knowledge moved forward. The conceptual modelling, suggested by Brocke and Buddendick (2006) has significant role with advent of new software engineering principles such as model driven or service oriented architectures. They emphasize the application of reusable conceptual models as a promising approach to support model designers. They argue that the design science research paradigm delivers a framework to strengthen the theoretical foundations of research on conceptual models. In the end, they conclude, the principle of reusing artifacts is widely accepted and applied in software engineering.

Weber (2010) signifies with increasing evaluation of software prototypes, design science research (DSR) emerges as a new research direction ensuring rigor and relevance in prototyping research projects. Weber argues that though DSR has proven to deliver relevant research results, but failed to fully accept as a DSR approach, because of failure of developing theoretical contributions. Weber (2010) points out DSR is used as an approach in many different paradigms such as positivist, interpretative and it has more potential even in socio-technologist or developmental paradigms too. Weber (2010) concludes in his scientific reviews that DSR is on its way as an acceptable and pluralistic research approach. Ploesser (2012) develops design theory for context-awareness based information systems. March and Smith (1995) provide validity of design science research on information technology.

An evaluation of the design artifacts and theories (Venable et al. 2014) is a key activity in the Design Science Research (DSR), as it provides feedback for further development and assures the rigour on the research. The following steps motivated the author for ascertaining the DSR, so that the constructs, models and methods have rigor on their implementations in achieving the research objectives:

- Clarify the goals of evaluation
- Choose the evaluation strategy or strategies
- Determine the properties to evaluate
- Design the individual evaluation episodes
- Implementation of new knowledge from evaluations

3.1 The Design Science Research and Guidelines

The author makes use of the fundamental principle of design science research from which seven guidelines (Hevner et al. 2004) are derived. The guidelines framed by Hevner et al. (2004) on DS research, as explained in Figure 3.1a are:

1 Requires creation of an innovative and focused artifact.
2 Specifies problem domain and or domains.
3 The artifact designed has a purpose, yielding solution for a problem thorough implementation and evaluation.
4 Innovation is crucial, ensuring an effective and efficient problem solution.
5 The artifact defines rigorous, coherent and consistent problem solution.
6 The process, with which artifact created, is generic and effective, in constructing problem and solution spaces in any domain.
7 Results must be communicated effectively to variety of audience.

As described in Figure 3.1, each guideline is connected to various components of research methodology (as numbered in right hand side with circles) and what the research outcome can deliver for each research question. These DS guidelines (March and Smith 1995) are reproduced in the following sections, explaining each guideline to each component of research methodology and its activity in the current domain of interest.

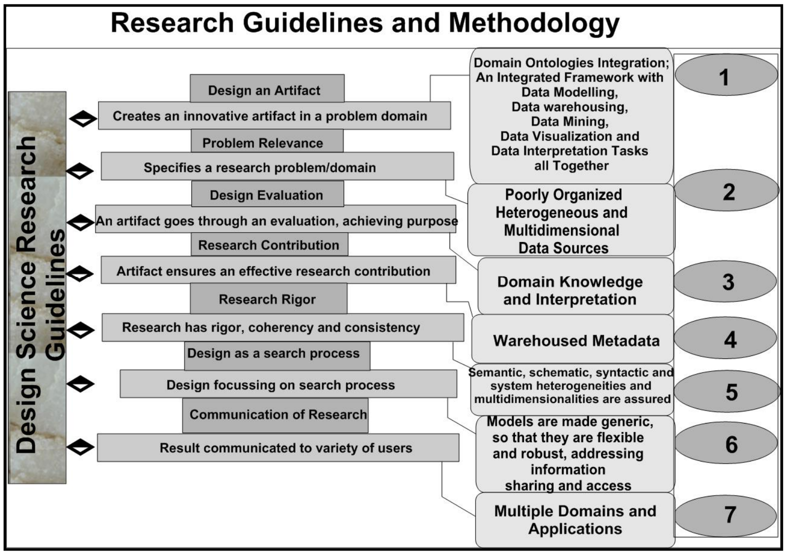

Research Guidelines and Methodology
Design Science Research Guidelines
Design an Artifact
Creates an innovative artifact in a problem domain
Problem Relevance
Specifies a research problem/domain
Design Evaluation
An artifact goes through an evaluation, achieving purpose
Research Contribution
Artifact ensures an effective research contribution
Research Rigor
Research has rigor, coherency and consistency
Design as a search process
Design focussing on search process
Communication of Research
Result communicated to variety of users
Domain Ontologies Integration; An Integrated Framework with Data Modelling, Data warehousing, Data Mining, Data Visualization and Data Interpretation Tasks all Together
1
Poorly Organized Heterogeneous and Multidimensional Data Sources
2
Domain Knowledge and Interpretation
3
Warehoused Metadata
4
Semantic, schematic, syntactic and system heterogeneities and multidimensionalities are assured
5
Models are made generic, so that they are flexible and robust, addressing information sharing and access
6
Multiple Domains and Applications
7

Figure 3.1a: DS research guidelines and methodologies

Guideline 1: Design as an Artifact

Artifacts, in the context of present research, are innovations, aiding the design of an integrated research framework and their components, which define various domain ontologies and their respective data structures. Data structures are designed by making use of the existing heterogeneous and multidimensional data sources in oil and gas business domains. Entity-relationship, object oriented and multidimensional data models are various constructs designed for different business scenarios and constraints. The design and development of domain, data modelling, and schema, data warehousing and mining, visualization including interpretation are parts of this integrated framework. These artifacts have relevance to the problem of research, described within the framework.

Guideline 2: Problem Relevance

The volumes of heterogeneous and multidimensional data sources are locked up in different media. The complexity of data sources precludes retrieving required information. An artifact that is designed has a relevance to the research problem and it is defined as the differences between methodologies or an integrated framework as a research outcome with existing methodologies. Poorly organized heterogeneous and multidimensional data sources, is one of such problems. Other current problems with existing methodologies are resisting to changes, poorly understood connectivity among ecosystems and their limits, access and sharing of domain knowledge among multiusers environment. The author aims at these artifacts addressing the research problems.

Guideline 3: Design Evaluation

The domain knowledge and its interpretation for decision support is key criteria. Data integration is crucial and artifact designs undergo the integration process. Successful integration of domain ontologies or structures infers design evaluation criteria. Another criteria is to ensure domain ontologies and their integration in a warehouse environment, with knowledge base solution. Based on variety of business and technical conditions, constraints and rules, the author evaluates the artifact designs, gaging IT artifacts in terms of their functionality, completeness, consistency, accuracy, performance, reliability, usability to fit with an ecosystem.

Guideline 4: Research Contributions

The author describes this guideline for ensuring quality research contributions based on the artifacts. The existing research contributions are based on design science and guidelines, as narrated in Chapter 2 and in section 3.0 of Chapter 3. The warehoused metadata derived from an integrated framework provides future scope and can generate research outcome keeping in view its robustness and flexibility. This guideline emphasizes the fact of scope of further research that can contribute and implement in other domain areas. Digital ecosystems as digital oil field solutions, ensure connectivity of systems with the designed artifacts. Modelling is ensured with domain ontologies with their feasibility and applicability in different knowledge based research problems. Research contributions also depend on creative development and evaluation of artifacts.

Guideline 5: Research Rigor

The rigor addresses the way in which research is conducted. The design science research requires the application of rigorous methods in both the construction and evaluation of the designed artifact. In particular, with respect to the construction activity, the rigor is assessed with respect to the applicability and generalizability of the artifact. Again, an overemphasis is on the rigor that can lessen the relevance. How well an artifact works, not to theorize about or prove anything about why the artifact works. It is imperative to understand why an artifact works or does not work to enable new artifacts to be constructed that exploit the former and avoid the latter. In addition, the author puts rigor on analysis of semantic, schematic, syntactic and system heterogeneities and multidimensionality, in particular in the oil and gas business domain.

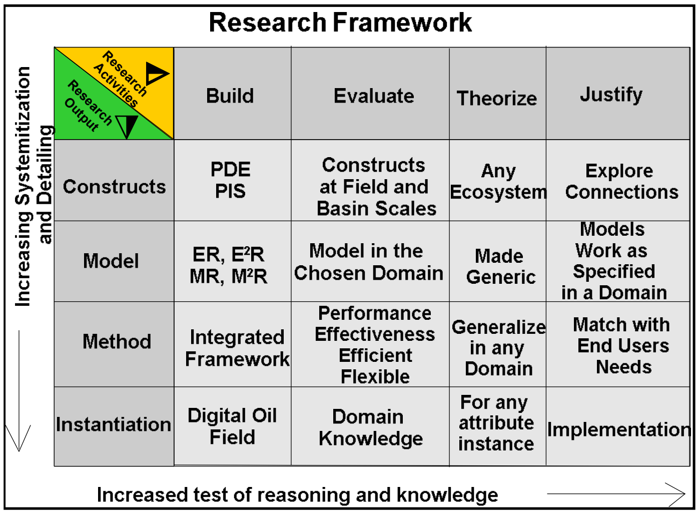
Guideline 6: Design as a Search Process

The author prepares models that are generic and flexible, so that they are robust in handling the information sharing and access under varied business and technical (geological) conditions. The design essentially focuses on a search process to discover an effective solution to a problem. The problem solving is viewed as utilization available means to reach the desired ends while satisfying laws existing in the environment (Simon 1992, 1996). Abstraction and representation of appropriate means, ends, and laws are crucial components of the design science research.

Guideline 7: Communication of Research

The author presents the design science research to technology oriented as well as management oriented audiences. Technology oriented audiences need sufficient details to enable the described artifact to be constructed (implemented) and used within an appropriate organizational context. This enables practitioners to take advantage of the benefits offered by the artifact and it enables researchers to build a cumulative knowledge base for further extension and evaluation. It is also important for such audiences to understand the processes by which the artifact was constructed and evaluated. This establishes repeatability of the research project and builds the knowledge base for further research extensions by design-science researchers in IS. That is, the emphasis is on the importance of the problem and the novelty and effectiveness of the solution approach realized in the artifact. In addition, knowledge facilitated from multiple domains and applications ensure the validity of the artifact and its implementation in other research domains.

These guidelines motivate the author to focus on current research objectives (RO1 to RO8 described in Section 1.3.1 in Chapter 1) and methodologies in multiple knowledge domains, using heterogeneous data sources. The entity relationship, multidimensional data relationship and object oriented data models are generated as sets of artifacts or constructs, representing semantics of data. Similar constructs are used to build and evaluate models in multiple domain applications within an oil and gas industry domain. As a follow-up of guidelines of design science information systems research, the author generalizes an integrated framework designed for its implementation in multiple domain applications. These innovations define the ideas, practices, technical capabilities and products through which the analysis, design, implementation and the use of information systems (IS) are expected to be effectively and efficiently accomplished. As identified in March and Smith (1995) and Ploesser (2012), the design processes and artifacts are evaluated analysing the IS problems and solutions. The author aims at the research framework as described in Figure 3.1b, focusing on research activities and expected research outcomes of the undergoing research.


Research Framework
Increasing Systemitization and Detailing
Increased test of reasoning and knowledge

Research Activities / Research Output	Build	Evaluate	Theorize	Justify
Constructs	PDE PIS	Constructs at Field and Basin Scales	Any Ecosystem	Explore Connections
Model	ER, E^2R MR, M^2R	Model in the Chosen Domain	Made Generic	Models Work as Specified in a Domain
Method	Integrated Framework	Performance Effectiveness Efficient Flexible	Generalize in any Domain	Match with End Users Needs
Instantiation	Digital Oil Field	Domain Knowledge	For any attribute instance	Implementation

Figure 3.1b: Updated research framework adopted in the current study

This research framework provides the author a guidance in preparing the constructs, models and methodologies in the oil and gas business domain. Models that use constructs represent the real world situations (for example, domains) along with a design problem and its associated solution space. Models aid the problem and solution, understanding of domain applications and frequently exploring the connections between problem and solution, enabling exploitation of effects on design decisions and changes in the real world. The current research method is more scientific and empirical in nature, thus based on certain basic postulates which are stated as under:

1. It relies on empirical evidence, such as testing models using data and or facts;
2. It utilizes relevant concepts, such as ontology based conceptualizations and contextualization;
3. It is committed to only objective considerations;
4. It presupposes ethical neutrality, i.e., it aims at nothing but making only adequate and correct statements about entities/dimensions/objects;
5. It results into probabilistic predictions and forecasts;
6. Its methodology is made known to all concerned for critical scrutiny for use in testing the conclusions through replication;
7. It aims at formulating most general axioms or what can be termed as scientific theories.

The data considered are from oil and gas industry worldwide, keeping in view the nature and types of data. Data are heterogeneous, multidimensional, multivariate and with hundreds of dependent and independent variables. In addition to oil and gas industries, medical ontologies, ecosystems research, earthquake prediction, embedded systems research are also targeted so that the proposed methodologies assessed to fit into other domain applications. Multidimensional data structures built are generic, so that their implementations can be extended in any application domain. The warehousing of multidimensional heterogeneous data, from a single repository system, can generate either data or plot, map and other graphic views that can effectively explore the data.

The author highlights that design science framework and guidelines (Henver et al. 2004, Vaishnavi and Kuechler 2004, 2007, Indulska and Recker 2008 and Neuman 2000) that provide a huge opportunity and scope for an exploratory, descriptive and explanatory research in information systems (IS) research domain including application domain research. In an application domain, IS research adds value to practice in an industry. As a part of invention, developing a technology solution to the petroleum industry in an application domain, is the subject of present industrial research. Data in major commercial industries (such as petroleum industry) are complex in nature and often, poorly organized, duplicated and exist in different formats. Business, in these companies, is operated both in space and time. Due to diverse nature of business products and operations in different geographic locations, these industries demand more accurate and precise information and data. Businesses operating in multi-client or multi-user environments with redundant data are prone to carry information with several ambiguities and anomalies.

In many cases, the data are summarized, according to some criteria and stored in a data warehouse. The "science" of data mining (Mattison 1996) involves exploration of data in an attempt to uncover patterns or detect anomalies. "Striking it rich", often means finding an unexpected result that requires users to think creatively. Tools for the data "miner", therefore, should allow for heuristic, iterative analysis where the user is looking for something but does not know quite what. Typically, however, the techniques available in data mining are restricted to spreadsheets and reports or simple bar charts or pie charts. Why not extend the capabilities of traditional business graphics to support the visualization of multidimensional data? Why not use the graphical methods available with visualization (Krishnamurthy 1999, Cleveland 1994 and Han and Cercone 2000) to transform multidimensional databases into images that would allow patterns inherent in the data to reveal themselves?

Keeping in view the research objectives and design science guidelines, the author constructs different large entity-relationship, multidimensional and object oriented data models. Author uses secondary data collection method, gathering the existing statistics (secondary analysis research) to carry out data modelling and do exploratory, explanatory and descriptive research (Neuman 2000). The author discusses a methodological framework, designed within purview of DS research in the forth-coming sections.

The ontological descriptions are proposed for simplifying the complexity of heterogeneous data. This results in design and development of conceptual models for translation into logical data models by a multidimensional data mapping approach compatible to any warehouse environment. The author translates the logical data models into implementation models, using a contemporary DBMS (for example, Oracle) and to make specific requests through SQL queries for locating a specific piece of data or information from massive storage warehouses. The author develops simple mining algorithms for extracting patterns, correlations and trends from heterogeneous data and statistical techniques are deployed for computing future forecast data models and also test measurement validities (Neuman 2000 and Gacenga 2013). Zaima and Kashner (2003) examine the data mining scopes in business industries.

Ontologies are generated using the concepts of knowledge representations (Cardenas and McLeod 1990) surrounding entity relationships, multiple objects and dimensions. The author finds storage for these heterogeneous data and thus for an integrated warehouse environment for business analytics purposes. The author does classification and specialization/generalization, facilitating semantics of data structuring and knowledge representation, aiming at knowledge base ontologies that target an effective data mining, at recognition of patterns, trends and correlations for large volumes of spatial-temporal data, stored in data warehouses and other information repositories. *Descriptive* and *predictive* data mining are used, in which general properties of data are characterized in the former one and predictions are made, based on inferences on the current data, in the later mining tasks. The entities, objects and dimensions that go into the conceptual data modelling represent different classes or categories as per the knowledge of conceptualization and specialization of datasets. For example, the spatial-temporal data are heterogeneous, their organization, structuring, modelling, mining, visualization and interpretation are critical for knowledge discovery and its management.

3.2 An Overview of the Research Framework

Research Activities

As described in Figure 3.1, the design science research framework and guidelines are followed up, the following research activities are pursued:

Build models: the author builds artifacts to perform a specific task. The basic question is, does it work? Building an artifact demonstrates feasibility. These artifacts then become the object of study. Constructs, models, methods and instantiations (implementation with live cases) are aimed at to build an integrated framework in oil and gas industry domain. Each is a technology that, once built, is evaluated scientifically.

*Evaluate models***:** the author evaluates artifacts to determine if any progress is made. The basic question is, how well does it work? Recall that progress is achieved when a technology is replaced by a more effective one. Requirements are evaluated for the development of metrics and the measurement of artifacts according to those metrics. Metrics here defines accomplishment and assesses the performance of an artifact.

Metrics is a quantitative indicator of attributes for software reuse or reusability. It is a measure of oil and gas organization's activities and performance.

Theorize models: Given an artifact whose performance is evaluated, it is significant to determine why and how the artifact worked or did not work within its environment. Such research applies natural science methods to IT artifacts. The author theorizes and then justifies theories about those artifacts. Theories clarify the characteristics of the artifact and its interaction with the environment that result in the observed performance. The natural laws governing the artifact and those governing the environment in which it gets operated are understood. Furthermore, the interaction of the artifact with its environment leads to theorizing about the internal workings of the artifact itself or about the environment.

Justify models: Once the data models are theorized, then the semantic data models correspond more to an end-user's conceptualization of a multidimensional database than does the relational model. Given a generalization or theory, an explanation is justified. That is, an evidence is gathered to test the theory.

The generic modelling articulations, needed for designing an integrated framework, are described in the following sections.

3.3 The Generic Modelling Concepts and Articulations

The domain, data modelling, data warehouse, data mining, data visualization are sequence of artifacts used and author describes the theory behind these artifacts or concepts before they are applied them in oil & gas domain in all the subsequent chapters of 4 and 5. All the generic structural models are presented in this section.

Keeping in view the research questions and objectives, RQ1 – RQ8 and RO1 – RO8, in Sections 1.3.1 and 1.3.2, generic formulations are described, as artifacts for the proposed research methodologies. Research methodologies in the current research work embody an *analytical research*, in which the author uses facts or information readily available for research, for modelling, analysis and critical evaluation of the information. This is an *applied research,* aimed at finding solutions for problems facing an industrial/business organization. This is a quantitative research, based on the measurement of quantity or amount. Initial conceptual research, such as ontology based modelling digs abstract idea(s) or theory. Author intends to develop new concepts or reinterpret the existing ontologies in multiple knowledge domains. Besides conceptual research, author does empirical research based on observations, utilizing systems and theories. It is a data-driven research, where large heterogeneous and multidimensional data sources involved, are used to draw conclusions, which are capable of being verified by an observation or an experiment. In such a research, it is necessary to get facts firsthand, at their source and actively to go about doing certain things to stimulate the production of a desired information. In such scenario, the author provides a working hypothesis or guesses so as to get the probable results. The researcher then works to get enough facts (data) to prove or disprove hypothesis. An empirical research is appropriate, when proof is sought that certain variables affect other variables in some way, as demonstrated in various application domains of oil and gas business in Chapter 3 and Chapter 4.

As per DS research guidelines in Section 3.1 and research questions and objectives (RQ1 to RQ8 and RO1 to RO8) placed in Sections 1.3.1 and 1.3.2 of Chapter1, a methodological framework is designed for building an integrated warehouse environment for accommodating the heterogeneous and multidimensional data sources. The integrated framework consists of data acquisition from multiple sources and domains, cleaning the data, organizing the data, identifying relationships among entities and or multidimensional data, structure the data in hierarchical, relational and network structuring architectures. Several data attributes in different domain applications are identified in order to build schemas. Different schemas are chosen to structure the data. The secondary data, published in the public domain are used for populating data instances in *fact* and *dimension* tables and test the data models. Ontologies are used to inter-relate and inter-connect these multiple entities and dimensions that have multiple meanings. In other words, ontologies search for multiple connections among multiple domains. The multidimensional data are varying in horizontal, vertical and lateral dimensions, which are logically and physically stored in one repository, so that they representable in different volumes or cubes for mining. The conceptualization and contextualization are used for explaining the models and their structuring process in the following sections.

Conceptualization and contextualization in the knowledge building process

Several data sources as displayed in Figures 2.1a, 2.1b and 2.1c are used to identify entities and dimensions for conceptual modelling. A declarative specification of the entities and attributes are made available. Formal analysis of these entities is extremely valuable when both attempting to reuse existing ontologies and extending them to other applications. The conceptual models are intended to be built for assembling and assimilating knowledge. The criteria for representing knowledge base models are mere understanding of domain knowledge. As shown in Figure 3.1, conceptualization and specification of conceptualization are attributed to evolution and emergence of both known and unknown dimensions and their associated attributes.

Each basin, has number of prospects which are derived based on the knowledge of exploration, drilling and production entities in a basin. Knowledge are two types, one, which is existing and the other, knowledge is built. As an example, prospect 2 interpreted based on the existing exploration and production data sources, is not known, until and unless the prospect is drilled. Similarly, as shown in hashed and solid lines in Figure 3.2, the author identifies the knowledge that exists and the knowledge that is to be built.

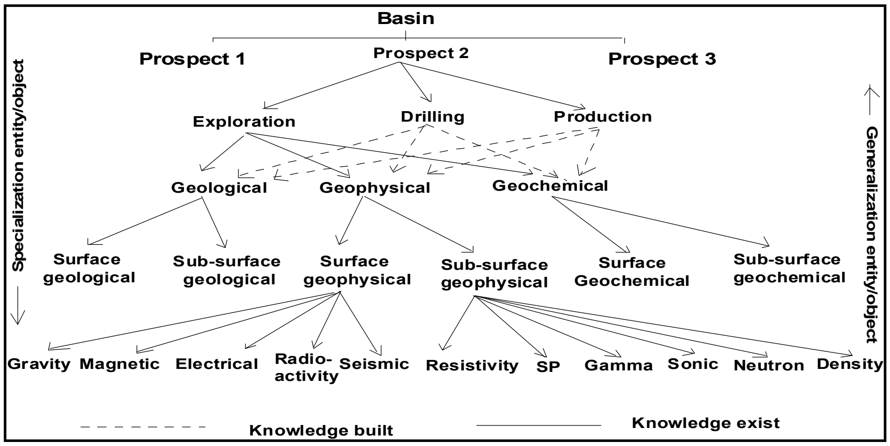


Figure 3.2a: Conceptual modelling displaying *basin,* with generalization and specialization dimensions

For example, in the exploration and prospecting of oil and gas scenario, the author describes a *sedimentary bearing basin* in a broad scale, as generalized super-type dimension from which different associated subtype dimensions are narrated at *prospect* level. The knowledge of the *basin prospectivity* is explored (a key term defined in a glossary in Appendix-1) among known and unknown attributed dimensions, as illustrated in Figure 3.2b. Knowledge is known from the existing measurements of surface- and subsurface geophysical methods and tools. The knowledge that is undiscovered, is built or explored among associated attributed dimensions (as highlighted hash lines) and their instances. The author understands the data relationships among these super and sub-type dimensions in a way, to explore connections and exploit the knowledge for implementation. The explicit in implementation and interpretation can lead to discovery of an oil and gas prospect. In another example, an ontology modelling explores connections among conceptualized dimensions, evolved from emerged data relationships as described in Figure 3.1b. Here the unknown data relationships are yet to be uncovered among the entities through conceptualization and contextualization.

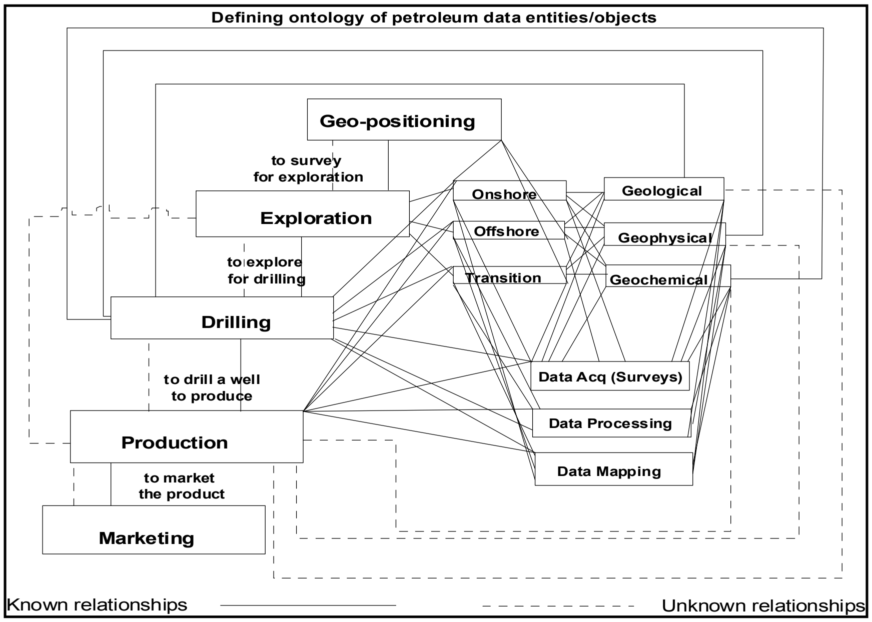


Figure 3.2b: Data relationships and their connectivity representing knowledge

The connections are explored among entities, described in different knowledge domains, linking the Figures 3.2a and 3.2b. For example, *exploration*, *drilling* and *production* are common entities or dimensions are connected through their common attributes. Many more such models built for exploring connections among data sources are described in Chapters 4 and 5. Key definitions that needed in modelling the generic ontologies in petroleum domain and sedimentary *basin* are:

A petroleum system is a natural system that encompasses a pod of active source rock and all related oil and gas geological elements and processes that are essential, if hydrocarbon accumulation is to exist (Magoon and Dow 1994). Elements are: structure (trap), reservoir, source, seal and processes are: generation, migration, timing of formation/migration and accumulations. *Petroleum Ontologies* (PO): The author uses ontologies for the purpose of making connections among petroleum systems' *elements*, *processes* and *chains*. Here *chains* are evolved from conceptualized attribute connections of *elements* and *processes*. Author uses all the dimensions and attributes associated with *elements, processes* and *chains* in the ontology modelling.

Sedimentary basins are regions of the earth of long-term subsidence creating accommodation space for infilling by sediments, in which most of oil and gas deposits get trapped. The subsidence results from the thinning of underlying crust, sedimentary, volcanic, and tectonic loading, and changes in the thickness or density of adjacent lithosphere. Sedimentary basins occur in diverse geological settings usually associated with plate tectonic activity. *Basin Ontologies* (BO): The author uses all the dimensions and attributes associated with the Basin in the ontology modelling for the purpose of making connections among subsidence, lithologies and tectonics. It is significant to have a clear understanding and explicit knowledge of geological settings, with which the oil & gas fields are accumulated. Other components of the integrated framework and its development are given in the following sections.

Multidimensional data, information and knowledge

For the purpose of designing an artifact within purview of design science guidelines, author explains knowledge domain and its representation explicitly. In the context of information technology and IS interpretative research, the author interprets knowledge distinctly from data and information (Rainer and Turban 2009). Data are collection of facts, measurements, and statistics; information is organized or processed data that are timely and accurate. Knowledge is information that is contextual, relevant and actionable. *Data* are elementary description of things, events, activities and transactions that are recorded, classified, and stored but are not organized to convey any specific meaning. Heterogeneous data refer to data from number of sources largely unknown, unlimited and in varying formats. Typically, such data are geology and geophysics and exploration & production (E & P) of oil & gas companies, data relevant to environment and disaster management, such as occurrence of natural calamities - earthquakes, tsunamis, and landslides. Organization of such data from number of sources is a huge challenge. The ontology is broadly a specification of conceptualization (Gruber 1993 and Meersman 2004) and many definitions of ontology exist in various contexts. In the present context, an ontology is a description of the concepts and relationships that can exist within an entity or a dimension or group of dimensions. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more in general. *Information* refers to these types of data that are organized, so that they have meaning in that context and value to the user. The *knowledge* refers to data and or information that are organized and processed from many sources to convey understanding, experience, accumulated learning and expertise as they apply to the current business problem. *Databases* collect related data files or tables containing the data. *Information system* (IS) is a process that collects, processes, stores, analyses and disseminates information for a specific purpose. Multidimensional databases refer to special types of databases that are optimized for a data warehouse and an online analytical processing (OLAP, a key term defined in the glossary, see Appendix - 1).

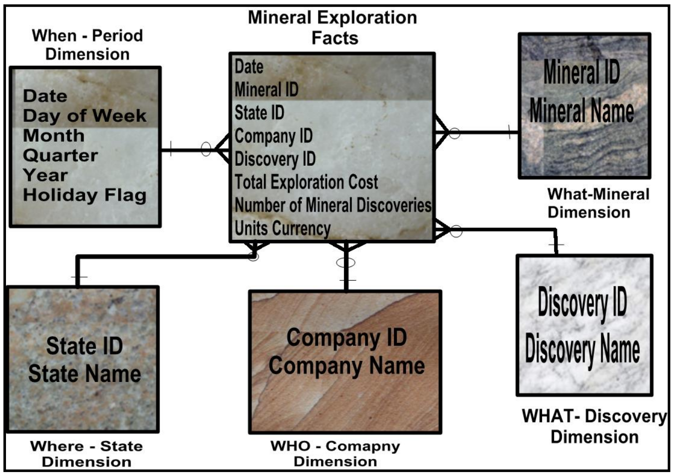
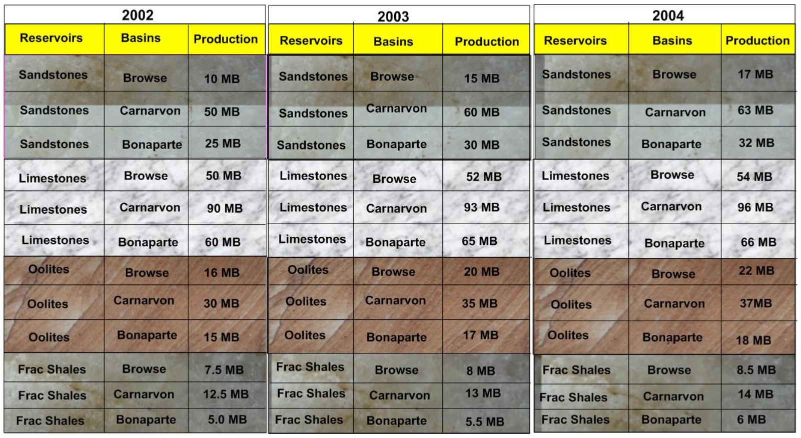


Figure 3.3a: A star schema model for mineral exploration, showing dimensions and their relationships with factual data

Numerous entities and dimensions are used in the mineral exploration industry for multidimensional modelling, as shown in Figure 3.3a. Data associated with petroleum (surface or sub-surface domain) or ecological and geomorphic systems (surface-domain), are from multiple systems, either from operational or functional units or their associated petroleum systems. ETL (extract, transform and load) process identifies the system to process data and then to accumulate in a warehouse environment. Such metadata and their connected data marts are explored for connections through *chains*, which are conceptualized dimensions evolved or emerged through modelling process. The author chooses to organize the data within the warehouse environment, either in relational or multidimensional format that are easy for end users to access. The differences between relational and multidimensional databases are analysed as shown in Figure 3.3b, as an example among several basins, having multiple *reservoirs* produce different petroleum products in different *years*. This matrix represents *production* dimensioned by *reservoir* and *basins* and *year* attributes.



2002			2003			2004		
Reservoirs	Basins	Production	Reservoirs	Basins	Production	Reservoirs	Basins	Production
Sandstones	Browse	10 MB	Sandstones	Browse	15 MB	Sandstones	Browse	17 MB
Sandstones	Carnarvon	50 MB	Sandstones	Carnarvon	60 MB	Sandstones	Carnarvon	63 MB
Sandstones	Bonaparte	25 MB	Sandstones	Bonaparte	30 MB	Sandstones	Bonaparte	32 MB
Limestones	Browse	50 MB	Limestones	Browse	52 MB	Limestones	Browse	54 MB
Limestones	Carnarvon	90 MB	Limestones	Carnarvon	93 MB	Limestones	Carnarvon	96 MB
Limestones	Bonaparte	60 MB	Limestones	Bonaparte	65 MB	Limestones	Bonaparte	66 MB
Oolites	Browse	16 MB	Oolites	Browse	20 MB	Oolites	Browse	22 MB
Oolites	Carnarvon	30 MB	Oolites	Carnarvon	35 MB	Oolites	Carnarvon	37MB
Oolites	Bonaparte	15 MB	Oolites	Bonaparte	17 MB	Oolites	Bonaparte	18 MB
Frac Shales	Browse	7.5 MB	Frac Shales	Browse	8 MB	Frac Shales	Browse	8.5 MB
Frac Shales	Carnarvon	12.5 MB	Frac Shales	Carnarvon	13 MB	Frac Shales	Carnarvon	14 MB
Frac Shales	Bonaparte	5.0 MB	Frac Shales	Bonaparte	5.5 MB	Frac Shales	Bonaparte	6 MB

Figure 3.3b: Relational database, showing production instances, from closely connected reservoirs of oil and gas bearing basins

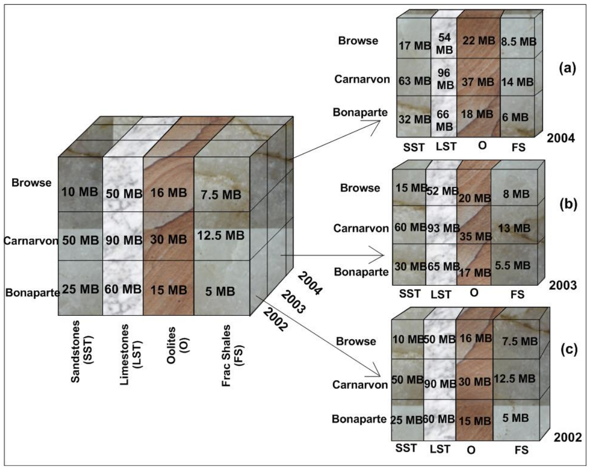


Figure 3.3c: Multidimensional database, showing a case as depicted in Figure 3.3b

The author transcribes ontologies for handling semantics, schematics and syntactic inconsistencies and heterogeneities that occur while designing multidimensional data structures from multiple oil and gas fields and basins. Multidimensional modelling is a database design technique specifically for designing the data warehouses and associated applications. The conceptual ER models are initially translated into dimensional models, classifying dimensions, designing high-level star schemas, using fine-grained dimension and fact tables. There are two ways of viewing fine-grained models, one is multidimensional cube, in which each cell contains one or more attributes and the dimensions are ways to categorize the raw data. The users (geoscientists, geo-engineers and reservoir modellers) of the data warehouse summarize the data (for example in terms of time periods, surveys conducted, contractors, wells drilled etc.), using these categories or dimensions. In a dimensional cube (Pujari 2002), each cell holds data relevant to the intersection of all its dimension values. For example, a cell might contain number of wells drilled in a period of time and basin with specific number of oil and gas producing wells. The equivalent view is also called the star schema. The author investigates each database as shown in Figure 3.3c, with dimension tables with centrally located fact table. The fact table is equivalent to cell in a multidimensional view. This table contains all the raw attributes and primary keys of all the surrounding dimension tables and other data column values. As an example, the author explains multidimensionality in the Figure 3.4, showing how different dimensions are connected to each other from different entities in a Petroleum Permits database. Survey_facts, well_facts and permit_facts are examples of fact data tables.

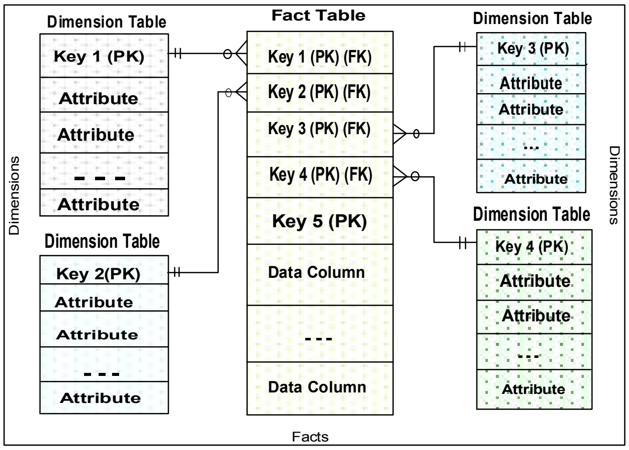


Figure 3.4: A star schema model depicting multiple dimensions and their relationships

Data attributes used in the modelling

In order to construct the data models, author tabulates data in multiple columns, in which each column is defined by a dimension, as shown in Table 1. The columns of a relation are named by attributes. For example, as shown in Table 1, the attributes are *Survey Name*, *Year*, *Length of Survey*, *and Type of Survey.* Here the domain is *Survey*. Similarly, drilled-well, year drilled, Depth Drilled, and Type of Well Drilled are attributes in another domain, called *drilled-well domain*. The geological-structure, reservoir, source, seal, timing, migration are attributes in a *petroleum-system* domain. The attributes appear at the tops of the columns. Often, an attribute describes the meaning of data entries in the column in a table. For instance, the column with attribute length holds the *Length of Survey*, in kilometres, of each survey.

Table 1: Attributes and Tuples (key terms defined in a glossary, see Appendix-1)

Survey Name	**Year**	**Length of Survey**	**Type of Survey**
ABC_1980	**1908**	**100**	**Explosive**
XYZ_1200	**1900**	**1000**	**Vibroseis**
MNQ_150	**1800**	**1500**	**Hydrophone**
KALI_1	**1980**	**7000**	**Exploratory**
KUTAI_2	**2000**	**10000**	**Development**
TARA_1	**2010**	**9000**	**Wildcat**

The name of a relation and the set of attributes for a relation are called the schema for that relation. The author illustrates schemas for data relations with the relation name followed by a parenthesized list of its attributes. For example, the schemas for the relations in different domains can be given here:

- Survey (*Survey Name*, *Year*, *Length of Survey*, *Type of Survey*)
- Drilled-Well (*Well-Name*, *Yea*r, *Depth*, *Type of Well*)
- Petroleum-System (Elements: (*Structure*, *Reservoir*, *Source*, *Seal*); Processes: (*Timing*, *Migration*, *Accumulations*))

The generic forms are: Dimension (attribute 1, attribute 2, attribute 3…) and in more complex generic term: Multidimensional (Dimension: (attribute 1, attribute 2, attribute 3…); Dimension: (attribute 1, attribute 2, attribute 3…)).

The attributes in a relation schema are a set, not a list. However, in order to talk about relations, a “standard” order is specified for the attributes. Wherever a relation schema with list of attributes is introduced, author takes this specific order as standard whenever the relation or any of its rows are displayed. In a relational model, a database may consist of one or more relations. The set of schemas for the relations of a database is called a *relational database schema*, or simply a *database schema*. The author references data instances associated with multiple dimensions with factual data tables, to make a connectivity among various other schemas. How the factual data identified with reference to the dimensions are as described in the Figure 3.5.

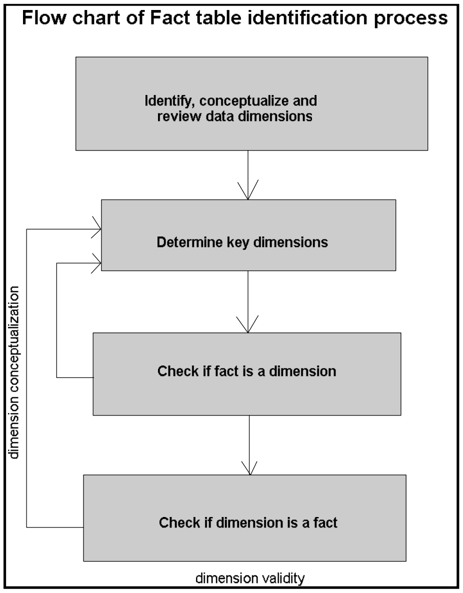

Flow chart of Fact table identification process
Identify, conceptualize and review data dimensions
Determine key dimensions
Check if fact is a dimension
Check if dimension is a fact
dimension conceptualization
dimension validity

Figure 3.5: Facts identification - workflow

Spatial – temporal domains and dimensions

The author identifies and describes hundreds of dimensions, their associated attributes and their instances with measurable magnitudes and strengths. In the current research (even in other domains as explained in appendices), the author focuses on spatial-temporal dimensions in oil and gas domain, which represent geographic (lateral dimension) and periodic (longitudinal) dimensions. In relational and hierarchical modelling approaches, the author uses multiple dimensions. The relational model requires that each component of each tuple to be atomic; that is it must be of some elementary type such as integer or string. It is not permitted for a value to be a record structure, set, list, array or any other type that reasonably can have its values broken into smaller components.

The author assumes each attribute of a relation in a *domain*, that is, a particular elementary type. The components of any tuple of the relation must have, in each component, a value that belongs to the domain of the corresponding column. It is possible to include the domain, or data type, for each attribute in a relation schema. Appending by a colon and a type after attributes, is used. The schema represented for either survey or drilled-well relation as:

Survey (Survey_Name: string, year: integer, length: integer/floating, type: string)

Here, a *survey* implies an investigation by any type of survey, such as *seismic, gravity, magnetic* or any other domain survey.

Multidimensional data structures

The heterogeneous data are unstructured, and therefore do not lend itself to analysis; in a data warehouse one can apply (hopefully dynamically) a variety of structures to suit a particular purpose. These data structures need to be designed and thus a methodology is proposed, connecting different knowledge domains (as shown via three case studies in the forthcoming chapters), which is an innovative element of current research. If one can analyse what was previous not able to be analysed, and thereby gain advantage, the discovery of the methodology and its application is significant in design science research.

The author structures multiple dimensions of data in various special logical schemas. It is variation of relational models that uses multidimensional structures and expresses the relationships among multidimensional data from multiple sources. The structure is broken into cubes and cubes can store and access data within confines of each cube. Each cell within a multidimensional cuboid structure contains aggregated data related to elements along each of its dimension. Even when data are manipulated, structuring allows an easy access from a compacted database format. The data still remains unaltered and continue to be interrelated. Data are viewed in multiple angles and dimensions, which describes a broader perspective of problem solution (O’Brien and Marakas 2009).

It is structure of a database management system, more commonly in a relational database management system (Hoffer et al. 2005 and Coronel 2011), the structure describes the tables, the fields in each table and the relationships between fields and tables. In a typical multidimensional database schema, author organizes and structures several dimensions in different dimension and fact tables and the relationships among tables of multiple dimensions and fact tables as demonstrated graphically in the Figure 3.6. In this model, dimensions and fact data dimensions and their instances are modelled from multiple fields, in one of the matured field areas of the Middle Eastern region. In this case, the author uses other structures such as star-schema, snow-flake schema and fact constellation schemas (Hoffer et al. 2005 and Rainer and Turban 2009) to accommodate multiple dimensions in an integrated metadata model. Several fact tables are constructed along with their associated dimension tables as demonstrated in the Figure 3.6.

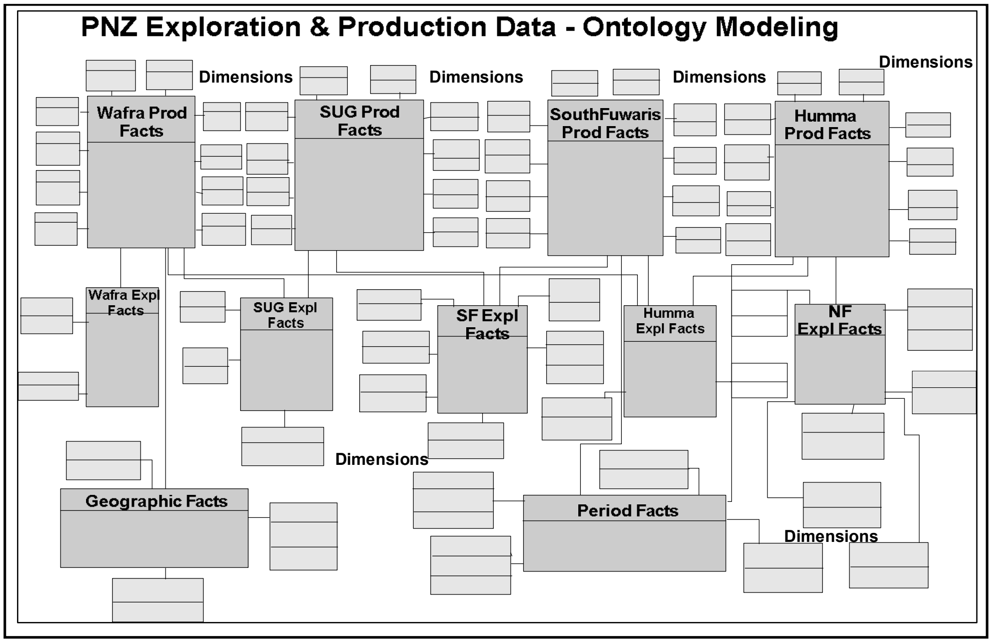


Figure 3.6: A conceptualized view: multidimensional schemas (for onshore producing fields and demonstrating their connectivity, as an example – Middle Eastern Matured Fields)

Heterogeneous and multidimensional data integration

The heterogeneous and multidimensional data are characteristically suitable for integration (in a warehouse environment) in an upstream. For the purpose of integration, the author explores for a connectivity among data sources associated with geological and geophysical (G & G) and oil & gas exploration and production (E & P) activities. Digital petroleum ecosystems and digital oil & gas field solutions are thus said to have emerged from these integrated systems. As described in Figure 3.6 and Figure 3.7, E & P and petroleum systems get interconnected through schemas, generated using super-type and sub-type generalization and specializations. The ontology based multidimensional data schemas are integrated for conceptualizing the undiscovered knowledge and its mapping, as illustrated in the Figure 3.7 and Figure 3.8. As shown in Figures 3.7 and 3.8, petroleum systems get interconnected through entities and dimensions of elements (*structure, reservoir, source, seal*) and processes (*migration, timing, critical movement and accumulations*), using E & P systems. The author considers navigational, surveys, wells, permits, production and other basins data to connect various E & P systems, examining the conceptualization and contextualization to explore unexplored dimensions and their attributes among petroleum systems and E & P (exploration & production) systems.

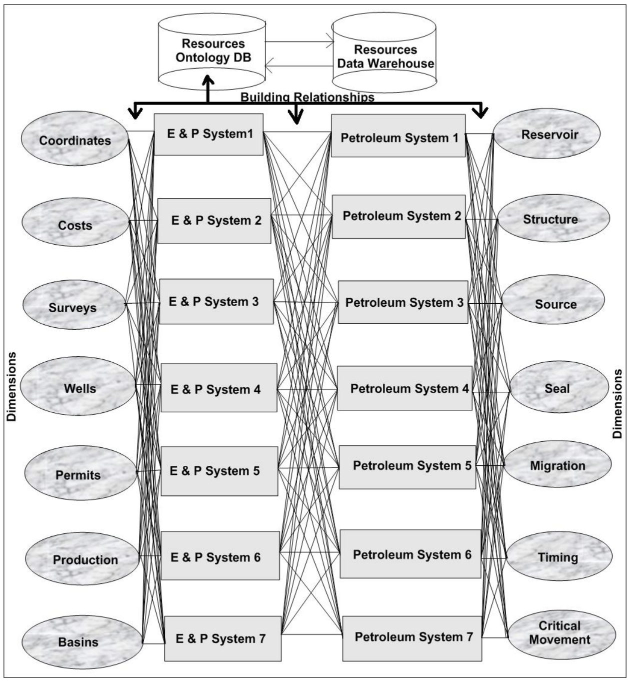


Figure 3.7: Building relationships among E & P and petroleum systems for integration

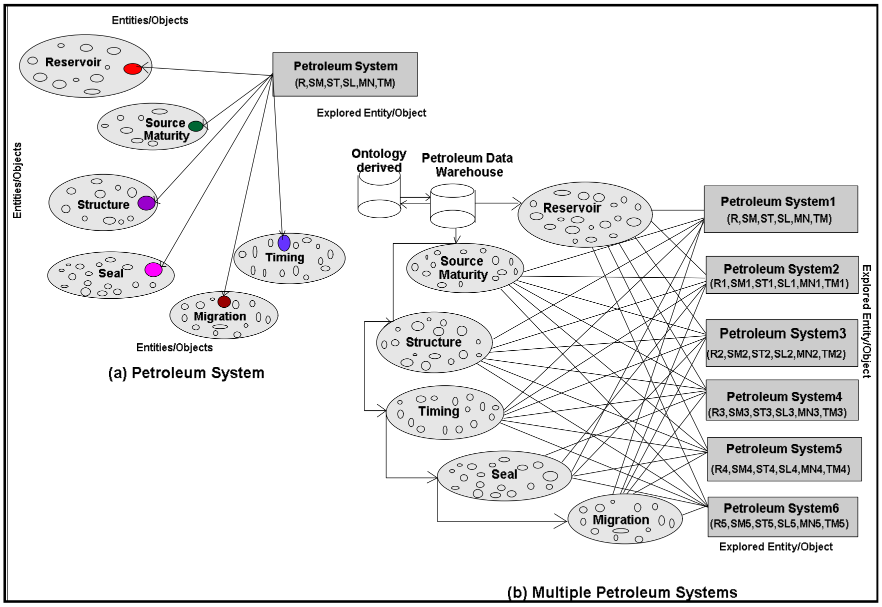


Figure 3.8: Designing a petroleum system – conceptualization of an ecosystem

The petroleum system defined here, is a super-type, composited in a generalized dimension in which *elements, processes and chains* are defined. From each set of *element* and *process* dimensions, a conceptualized *chain* dimension is emerged. As demonstrated in the Figure 3.8, the author uses ontologically derived metadata to explore connections among different petroleum systems through their *elements-processes-chains* structure.

Constraints and design of business rules

In the design science research, a data model is an artifact and a notation, for describing data or information. Garcia et al. (2008) describe data model into three parts, such as structure of the data, operations on the data and constraints in the data. Database models in a way are limited and constrained by description of operations and constraints on what the data can be, providing a strength in the implementation of operations efficiently.

While designing an artifact or construct of a model, the author has identified and documented business rules and constraints. While designing ER and multidimensional data models, author examines business rules rigorously, keeping in view the heterogeneity of data structures. Data models designed in each domain ensure their data relationships, addressing business rules and constraints of the framework. Keeping in view the nature, role and scope of the data, business rules are characterized, controlling the business processes. Entities, dimensions and objects are appropriately classified with business rules and constraints. Both data model designers and users are responsible for these business rules and constraints. Typical business rules are “at least one petroleum field must exist within a sedimentary basin”, “each petroleum system must have at least one or more oil and gas fields”. Other rules are “each field at least has either a survey or a drilled well”. “Each survey or drilled well has a producing horizon with oilplay attribute”. Similarly in the model design view point, one of the “one-to-one, one-to-many or many-to—many relationships” must exist. The author discusses more business rules and constraints in Chapters 4 and 5, where oil and gas exploration details are given.

Multidimensional data cubes and online analytical processing

The data cubes (Pujari 2002) are generated representing multidimensional 2D tables’ extensions; geometrically, each cube is a three-dimensional extension of a square. 3D data cube is a set of similarly structured 2-D tables stacked on top of one another. Data cubes, today, are built with many more dimensions (more than three-dimensions) allowing 64 dimensions. 4-D data cubes (Weimer and Davis 1995) consist of series of 3D cubes, through visualizing multidimensional spatial, periodic and geometry-base attributes. The data cube interpretation is popular in G & G (geology & geophysics) and E & P applications, in which multidimensionality of structured cubes is representative of multidimensional arrays. The MOLAP (multidimensional online analytic processing) analyses structured sub-sets of arrayed data. Another type of OLAP uses relational databases, called ROLAP that implements relational tables (up to twice as many as number of dimensions) instead of as a multidimensional array. Each of the tables, called a cuboid, represents a particular view; multiple cuboids are conventional database tables, which are processed and queried using indices and joins. This ensures inclusion of tables in a format efficient for large datasets, since the tables include only data cube that actually contain the data instances, but lack built-in implementation of indexing of MOLAP.

The multidimensional data analysis tools provide multiple data views that can facilitate interpretation of "what is happening or what has happened". For this purpose, multidimensional analysis tools allow users to "slice and dice" the data in any direction. In the data warehouse, relational tables are linked, forming multidimensional data structures, or cubes (Figure 3.9 and Figure 3.10). Statistical tools provide users with mathematical models that can be applied to the data to gain answers to their queries. For example, in *a sedimentary basin* scenario, the author is able to connect multiple petroleum systems (information systems) through cuboid structuring. These systems possess several oil and gas fields and each field with multiple producing horizons. The data organized in *cubes*, are sliced and diced to see certain facts or instances of producing horizons, for example, "how much oil/gas held in a particular geological structure or reservoir".

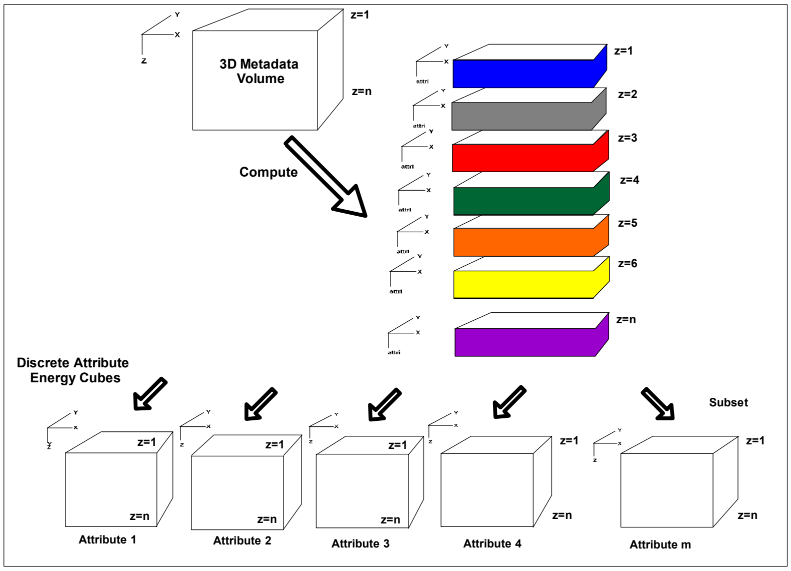


Figure 3.9: A workflow - attribute extraction and representation from a metadata volume

The geologists, geophysicists and production engineers might like to see *structural entrapment* or *reservoir potential* in each petroleum system of the sedimentary basin, so that they could evaluate the total potentiality (of oil and gas) of the basin. In this case, oil and gas business organization is reflected in the multidimensional structure. It can be said for all logical reasons, inherently sedimentary basin is reflected as a multidimensional data structure or a *digital ecosystem*. The power of multidimensional analysis lies in its ability to analyse the data in such a way as to allow users to quickly answer either business or scientific questions through mining and visualization. Figure 3.9 and Figure 3.10 are good examples of multidimensional representations for visualization and interpretation.

Several entities and or dimensions are described and for each dimension an attribute is described. As shown in Figures 3.8 and 3.9, for each attribute, the author computes a cube for mining, visualization and interpretation. For integrated interpretation, the author merges cubes of all attributes, so that an interpretative integrated cube is generated representing all the dimensions. Number of slices are cut from each cube, each slice is a view that can interpret attribute of conceptualized knowledge, such as "unexplored connectivity among two reservoirs in a single petroleum ecosystem". These slices are presented in different colours, each colour attribute representing a conceptualized interpretable reservoir property.

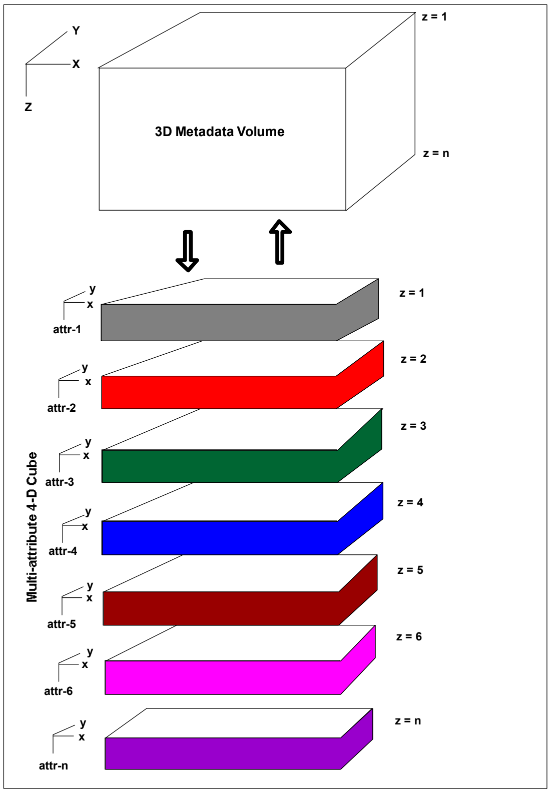


Figure 3.10: A workflow - volume attributes taken from a metadata volume

Figures 3.9 and 3.10 are parts of an integrated framework as more details emerge in the forthcoming sections of Chapters 4 and 5. Interpretation of multidimensional data views in different geographic regions interactively is mere a challenge for visualization and interpretation. For this purpose, the author constructs XML and XML schemas (Heather 2004) in oil and gas domain, as described in the following sections.

Description of XML data views

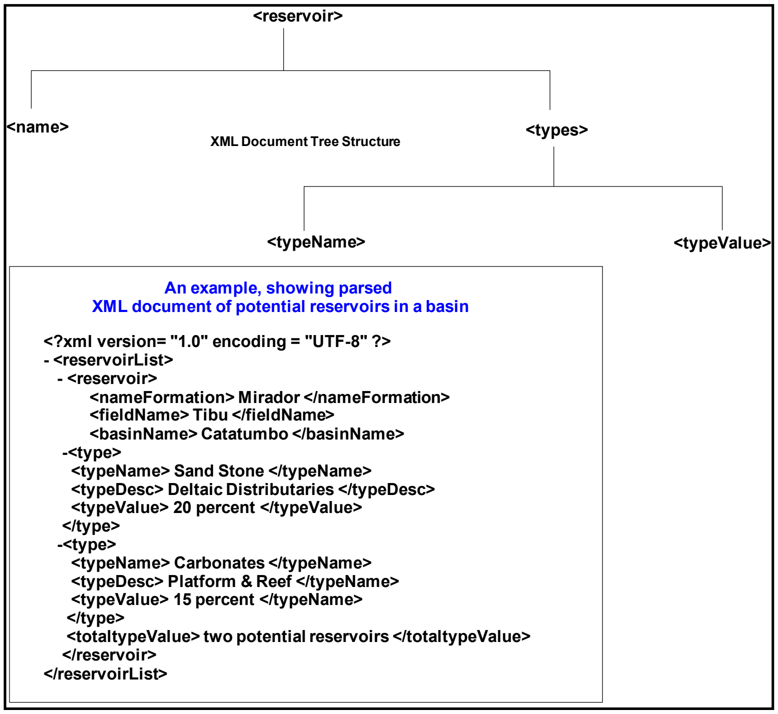
The author describes data views that need to be transmitted in different geographic locations, in a way they are transmittable by XML documents. XML stands for eXtensible Markup Language (Heather 2004), meant to design, transport and store data in any geographic region. In the present scope of multidimensional and heterogeneous data, XML is designed to transport among multiple geographical regions. Documents are prepared with this markup language. An XML Schema describes the structure of an XML document. A typical XML schema looks like:

```
<?xml version = “1.0”?>
<xs: schema xmlns:xs = http://www.w3.org/2001/XMLSchema>
<xs:element name = “note”>
  <xs:complexType>
    <xs:sequence>
      <xs:element name = “to” type = “xs:string”/>
      < xs:element name = “from” type = “xs:string”/>
      <xs:element name = “heading” type = “xs:string”/>
      <xs:element name = “body” type = “xs:string”/>
    </xs:sequence>
   </xs:complexType>
  </xs:element>
</xs:schema>
```

The purpose of a DTD (Document Type Definition) is to define the legal building blocks of an XML document. A DTD defines the document structure with a list of legal elements and attributes. XML Schema is an XML-based alternative to DTD and describes the structure of an XML document. XML Schema language is also referred to as XML Schema Definition (XSD). The purpose of an XML Schema is to define the legal building blocks of an XML document, just like a DTD. An XML Schema (Heather 2004):

- defines elements that can appear in a document
- defines attributes that can appear in a document
- defines which elements are child elements
- defines the order of child elements
- defines the number of child elements

- defines whether an element is empty or can include text
- defines data types for elements and attributes
- defines default and fixed values for elements and attributes



An example, showing parsed
XML document of potential reservoirs in a basin

```
<?xml version= "1.0" encoding = "UTF-8" ?>
- <reservoirList>
  - <reservoir>
       <nameFormation> Mirador </nameFormation>
       <fieldName> Tibu </fieldName>
       <basinName> Catatumbo </basinName>
   -<type>
     <typeName> Sand Stone </typeName>
     <typeDesc> Deltaic Distributaries </typeDesc>
     <typeValue> 20 percent </typeValue>
   </type>
  -<type>
     <typeName> Carbonates </typeName>
     <typeDesc> Platform & Reef </typeName>
     <typeValue> 15 percent </typeName>
    </type>
    <totaltypeValue> two potential reservoirs </totaltypeValue>
   </reservoir>
</reservoirList>
```

Figure 3.11: A data view description, carrying *reservoir* information

A typical XML document as shown in Figure 3.11, carries *reservoir* dimension and its information. The conceptualized dimension is either derived in between the known attributes or undiscovered attribute dimensions or even in between two conceptualized attribute dimensions. Here, the author describes conceptualization as a data relationship evolved during ontological data descriptions and modelling processes. An *oilplay* or *prospect* are examples of these conceptualized dimensions. In the case of denormalized data relationship scenarios, multidimensional data structures with conceptualized dimensions get emerged to fine-grained data structures, representing undiscovered finer knowledge.

3.4 Ontologies in Heterogeneity and Multidimensional Scenarios

Ontologies are explicitly explained in this section, adding references in Chapters 1 and 2. The author presents high-level ontologies all in one place in Chapter 4, described in an “Exploration Project”, where the context arose to present ontologies in an exploration project. To this extent, only generic and high level ontologies are presented in Chapters 3 and 4. Connotation of "ontologies" (Jarrar 2005 and Meersman 1999) is compared in various domains. The use of ontologies is emphasized with their definitions though they have some inconsistencies in different contexts and domain applications. But looking at applications in gene technologies, bio-informatics, and software engineering, the author finds ontology, which has role to play in petroleum engineering and systems analysis as well. The author in the context of oil & gas domain, takes definitions not so different from others. The current research is limited to description of high level ontology constructs in oil & gas domain. The constructs are the basis for building models as per research objectives narrated in Sections 1.3.1 and 1.3.2. In order to address RQ2 and RO1 and RO2, data from multiple sources are characterized by multiple types and multiple dimensions (Figures 2.1 and 2.2 in Chapter 2). The heterogeneities are:

- ***Syntactic Heterogeneity***: is a result of differences in representation format of data
- ***Schematic or Structural Heterogeneity***: the native model or structure to store data differs in data sources leading to structural heterogeneity. Schematic heterogeneity that particularly appears in structured databases is also an aspect of structural heterogeneity.
- ***Semantic Heterogeneity:*** differences in interpretation of the 'meaning' of data are source of semantic heterogeneity
- ***System Heterogeneity:*** use of different operating system, hardware platforms lead to system heterogeneity

Many researchers in the literature have different opinions on ontologies. Because of the heterogeneities, while describing data relationships and models, dimensions and their associated attributes pose semantic inconsistencies for vocabularies and different terminologies used. In order to address these issues, the author describes semantic ontologies (Meersman 1999, 2000, 2001 and 2004 and **http://lingo.stanford.edu/sag/L221a/gs-ch3.pdf**). The author uses ontologies, as formal models of representation with explicitly defined concepts and named relationships linking them, to address the issue of *semantic heterogeneity* in data sources. The author is of the opinion that semantic ontologies are described keeping in view the heterogeneities in the meanings of vocabularies and terminologies that arise in the modelling process. Jarrar (2005) describes domain axiomatization, focusing on the characterization of the intended meaning (i.e. intended models) of a vocabulary at the domain level, application axiomatizations mainly focusing on the usability of this vocabulary according to certain application/usability perspectives. Jarrar (2005) further generalizes an ontology in general to an agreed understanding (i.e. semantics) of a certain domain, axiomatized and represented formally as logical theory in a computer resource. By sharing an ontology, autonomous and distributed applications can meaningfully communicate to exchange data and make transactions interoperate independently of their internal technologies. Meersman (1999, 2000 and 2001) describe semantic ontology tools in information systems' design with a special focus on databases. The author reiterates that use of ontologies and their definitions though have some inconsistencies in different contexts and domain applications, but looking at its application in gene technologies, bio-informatics, software engineering, and the author finds ontology has much role to play in petroleum engineering and systems analysis as well. In the context of oil & gas domain, the author takes definitions not so different from Mustafa (2005) and Meersman (2004), believing that these are articulations expressed in different data relationships in hierarchies in specializations and generalizations levels. All the data relationships are expressed in ER and Multidimensional modelling articulates and the inconsistencies in terms of semantics, schematic, syntactic and systems are resolved while building, connecting and integrating the data models from different domains. Other tools that drive ontologies, such as domain, data modelling, schemas, data warehouse, data mining, visualization and interpretation that make up an integrated framework, is the basis of major articulation and contribution in this thesis. The author uses these articulations in building and connecting petroleum systems' *elements* and *processes*, for designing total petroleum information systems. These are termed as "petroleum digital ecosystems and digital oil field solutions". More details on ontology terminologies are updated in Chapters 1 and 2.

Keeping in view the flexibility and versatility of modelling, the rapid development, adoption and public availability of ontologies has made it possible for the *data integration* community to leverage them for *semantic integration* of data and information. Specifically, ontologies play the following key roles:

- Content Explication

The ontology enables accurate interpretation of data, especially when they derived from multiple sources through the explicit definition of terms and relationships in the ontology.

- Information annotation

Writing information annotation facilitates annotation ontologies. This helps capturing and reusing of existing ontologies, such as semantic and textual tags in many domain applications.

- Query Model

Queries are formulated using the ontology as a global query schema.

- Verification

An ontology verifies the mappings used to integrate data from multiple sources. These mappings may either be user specified or generated by a system. There are three main architectures that are implemented in ontology-based data integration applications, namely:

Single ontology approach

A single ontology is used as a global reference model in the system. This is the simplest approach as it can be simulated by other approaches. This approach handles global schemas, but is complex and difficult to understand by DB design, development and implementation.

Multiple ontologies

Multiple ontologies, each modelling for an individual data source, are used in combination for integration. Though, this approach is more flexible than the single ontology approach, it requires creation of mappings and modelling between multiple ontologies. Today, the ontology mapping is a challenging issue and is a focus of large number of research efforts in the computer science. Since data integration is a prerequisite in any business situation, data sources from multiple domains need description of multiple domain ontologies.

Hybrid approaches

An ISO 15926 Part 4 aims at exactly to this approach (West 2006). The hybrid approach involves the use of multiple ontologies that subscribe to a common, top-level vocabulary and content interpretation. The top-level vocabulary defines the basic terms of the domain. Thus, the hybrid approach makes it easier to use multiple ontologies for integration in presence of a common vocabulary.

The author uses these approaches in different domains of data/information structuring and knowledge interpretation. In the current approach, an ontology based multidimensional data warehousing is used for integrating heterogeneous data sources and their domain ontologies. The author describes these heterogeneities within contexts of petroleum systems.

Syntactic heterogeneity:

The author supports the idea of ontology based fine-grained multidimensional data structuring approach (Rudra and Nimmagadda 2005) appears to be feasible and applicable, when data are syntactically to be addressed from multiple sources. Syntactic ambiguity, in the context of oil and gas domain, focuses on multidimensional attributes, such as *petroleum systems their associated elements and processes*, which are differing in their formatting and storage requirements, such as interpreting attributes in more than one context or concept in multiple domains. Differences in syntactic ambiguity arise here not from the range of meanings of single words or vocabularies from multiple dimensions and domains, but from the relationships among the words of *geology*, *geophysics*, *reservoir* and *production* domains, including properties of these dimensions and attributes. In the star-schema multidimensional data structures references are made to connect all the dimensions in spite of subtle differences in their syntactical representations. *Geological structure* and or *reservoir* are interpreted with similar structure in multiple knowledge domains, implying syntactic ambiguity. Two or more meanings are possible among attributes described in the multidimensional data structures. The syntactic heterogeneity addresses issue with consistent syntaxes among attributes of the dimensions involved in the multidimensional data structure mapping and modeling process. During integration of multiple star and or snow-flake schemas and building relationships, including rules and constraints associated with data structures, syntax makes consistent in formatting the data structure. Variations or conflicts that arise during formatting, the data sources bring up syntactic ontologies. While formatting and integrating multiple domain ontologies, syntaxes maintain the consistency among attribute relationships. When the relationships have an ambiguity in interpreting the attributes and their associated dimensions, syntactically described ontologies facilitate fixing these ambiguities. For example, mapping of multiple dimensions while designing data structures, based on petroleum systems' elements, syntactic ontologies make the structures more consistent including domain knowledge described. Joining of several dimensions together makes relationships conceptualized and contextualized more effectively without any ambiguity. Both local and broader contexts, if appropriately described, syntactic ambiguities can be resolved. Structural and analytic ambiguities are fixed by syntactic ontologies. The composited dimensions, joined dimensions, including splitting at times, create structural ambiguities for which knowledge based graphics; visualization and interpretation reduce intricacy of an overall system. Finally, the purpose of syntax heterogeneity is to standardize the formatting styles and their vocabularies.

Schematic or structural heterogeneity

During logical organization and storage of oil and gas data, structural heterogeneities may have occurred. Whenever and wherever, schemas needed merging, restructuring and joining, contextualized and conceptualized data representations do emerge. Heterogeneities that arise during this process are handled by schematic ontologies and ambiguities are removed. For example, the composition of petroleum system elements during generalization and specialization processes including representation of dimensions in map and graphic views, value or instances may not have been distorted.

Semantic heterogeneity

Ontologies are described as building relationships among different entities and or dimensions for the oil & gas data sources for fine-grain structures (through denormalized ER and Multidimensional articulates). The author uses semantics as the study of meaning that is evolved, while modelling relationships, built among multiple dimensions and to better understand them with no ambiguity. Meersman (1999, 2000, 2001 and 2004 and http://lingo.stanford.edu/sag/L221a/gs-ch3.pdf) describes semantic ontologies in designing and developing information systems. Semantic ontologies are described keeping in view heterogeneities in the meanings of vocabularies and terminologies. While describing data relationships and models, the dimensions and their associated attributes pose semantic inconsistencies for vocabularies and different terminologies used in the modelling. Semantic heterogeneity has an ambiguity, in which differences in interpretation of *meaning* of data from multiple sources may have occurred during structuring domain ontologies and mapping domain knowledge. Using the description of ontologies of multiple dimensions, meaning of geological *structure* is ensured not to get confused with data structure. Meaning may have changed in multiple domains while building the concepts and contexts. For example, meaning of *data structure* is same as it is either envisaged during mapping/modelling or conceptualization of ontological descriptions of other domains, such as *geological structure,* an element or a dimension of a petroleum system. While populating the data in fact tables, uniqueness of the meaning of multiple dimensions that connected to fact tables is ensured. Meaning or the content of the information may not have been distorted during multidimensional structural design and development. Semantic ontologies make the meaning of multiple dimensions more consistent as it is required in metadata descriptions later. Meaning of concepts and contexts are made more consistent and standardized.

System heterogeneity

System heterogeneities occur, when the spatial-temporal data are documented on different hardware and software platforms. Different operating systems may need to convert at times the data in different formats, addressing the syntactic and semantic variations and their conflicts. For example, workstations, nowadays accept all the standard formatted data and are easily convertible to other multiple formats, without changing the meaning and syntax of the data. Typically, all the spatial-temporal data are made available in much hardware and driven on different software systems. Ambiguities that arise due to system heterogeneities, are resolved using standardized data sources, maintaining standardized syntactic, schematic and semantic ontologies, especially during exchange of information at geographically and periodically varying dimensions. The denormalization of data structures also facilitates standardized data characterizations, addressing intensive system heterogeneities.

3.5 Ontological Descriptions in Ecosystems' Modelling

For addressing RQ1 - RQ7, the author describes how ontologies narrate ecosystems in this section. While describing ecosystems in oil and gas domain, author emphasizes the fact that characterization and description of data sources are crucial and critical. Millions of records from thousands of attributes are in one repository, termed as a digital ecosystem. In an analogy, productive or non-productive petroleum bearing *sedimentary basin* consists of several petroleum systems, each system here is simulated as an information system, with information of several oil and gas and producing fields and their horizons. Each *pay* or a producing horizon is characterized and categorized by several drilled wells. Each drilled well has oil/gas *pay* horizon. Here a hierarchy of generalization, on a broad sedimentary scale (with multiple domains) to specialization, to a drilled-well (into a single dimension) scale, is interpreted. This is inherently an ecosystem, a system whose members are hierarchically connected and communicating each other. Similar participation of relationships and or positively summed relationships may have been benefited each other, which may be referred to as self-sustaining system or overall system participation with other neighbouring (petroleum) systems.

As it applies to any business, an ecosystem is viewed as a system where the relationships established across different *basins,* become mutually beneficial, self-sustaining and (somewhat) closed (Damiani 2008). This is clearly the case for a broader sedimentary basin, where a total petroleum system (TPS) is extended and described with multiple systems among several countries. A digital ecosystem is an emerging phenomena, in which all the existing attributes, conceptualized and contextualized attributes are ontologically described and interconnected within a warehouse environment. This approach leads to further development of innovative/creative ideas and technologies in oil and gas domain. The goal of an ecosystem is to generate “local to global” opportunities, leads and prospects of oil and gas. Once an ecosystem is established, the region exhibits network of areal extents. This approach makes an attractive venture to explorers and prospect generators, which is significant in terms of knowledge enhancement of the TPS.

3.6 An Integrated Methodological Framework

As per the research objectives laid in Sections 1.3.1 and 1.3.2, in Chapter 1, the author aims at designing an integrated framework (addressing RQ1, RQ2 and RO1 and RO2). As discussed in Moody and Kortink (2003), Hoffer et al. 2005, Coronel (2011) and Pujari (2002), the author uses star, snowflake and fact constellation schemas for constructing multidimensional logical data models. As suggested by Mattison (1996) and Marakas (2003), warehoused data are hierarchically structured in different knowledge domains. Figure 3.11 describes one of the initial hierarchical structural views. For integration, ontologically structured data are warehoused through multidimensional structuring. Hierarchically designed structures are intended to a goal of fine-grain multidimensional data structural design. Process of integrating domain knowledge from fine grained metadata is significant part of methodological framework.

The author equates ontologies in the present application scenarios with taxonomic hierarchies of classes, class definitions, and class conceptualizations of relationships described among multiple dimensions. To specify conceptualizations, business rules and axiom constraints need to be committed during contextual interpretations of the conceptualizations. In the context of an integrated workflow, the concept of an ecosystem is benefited with several multi-disciplinary entities or dimensions participation in the integration process through conceptualized relationships (in other words through symbiotic relations, positive sum relationships). It is a term that originated from *biology*, and refers to self-sustaining systems. It is complex community of large size multiple dimensions and its environment functioning as an ecological unit. More realistically, it is a term of volume of attributes gathered from multiple sources (both geographical and periodic) all in one place. Similar analogy is applied in a broader sense of large size *sedimentary basin* (local scale), with multiple petroleum systems with several hundreds of attributes, is connected to other large size *basins* elsewhere, on a global scale, where geology has no boundaries. As it applies to any business, an ecosystem, in case of petroleum system, can be viewed as a system where the relationships established across different *structures* and *reservoirs* (dimensions representing geological elements) among several petroleum systems in a basin, can become mutually beneficial, self-sustaining and (somewhat) closed. The goal of this ecosystem design is to generate commercial values.

Design of an integrated information system in the oil and gas industry depends on individual design of conceptual schemas of oil and gas industry's operational units "entities/objects/dimensions". This is an ontology based data conceptualization. Integration of schemas belonging to various operational sub-systems is a requirement for an oil and gas industry to accomplish the legality and validity of data. Intelligent and expert data systems (Hoffer 2005, Plastria et al. 2008 and Wand 2000) are used in geophysical exploration and prospecting. Broad Issues relevant to computer applications in exploration industry are discussed, demonstrating their applicability and feasibility.

3.6.1 Components of an integrated workflow

Several artifacts generated, are parts of the components of an integrated workflow. The author discusses workflows associated with domain, schema and metadata modelling, data mining, visualization and interpretation/knowledge based models, as done in the following sections. These models are critical in addressing the heterogeneity, multidimensionality and granularity of systems' design and development.

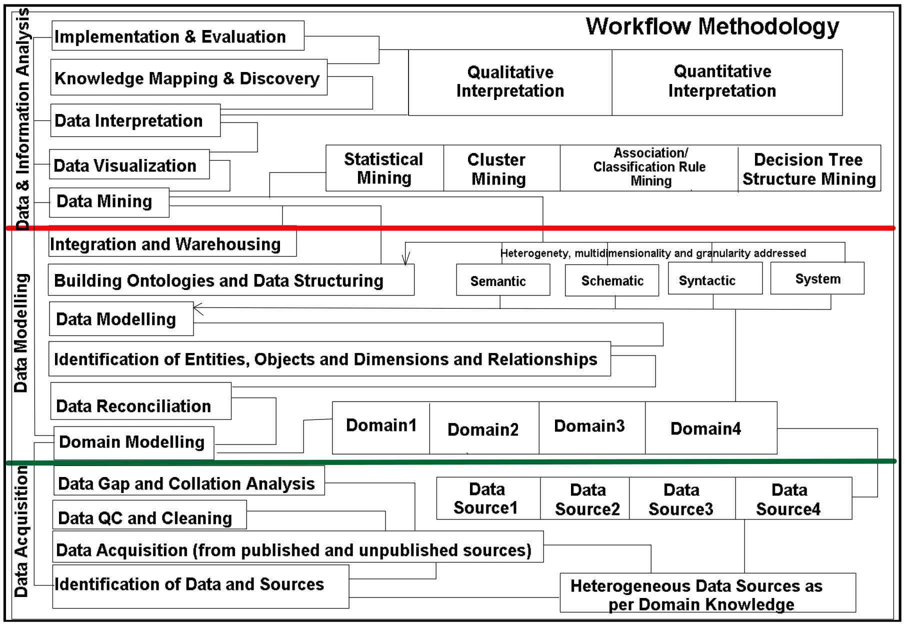

Workflow Methodology
Data & Information Analysis
Implementation & Evaluation
Knowledge Mapping & Discovery
Data Interpretation
Data Visualization
Data Mining
Qualitative Interpretation
Quantitative Interpretation
Statistical Mining
Cluster Mining
Association/ Classification Rule Mining
Decision Tree Structure Mining
Data Modelling
Integration and Warehousing
Building Ontologies and Data Structuring
Heterogenety, multidimensionality and granularity addressed
Semantic
Schematic
Syntactic
System
Data Modelling
Identification of Entities, Objects and Dimensions and Relationships
Data Reconciliation
Domain Modelling
Domain1
Domain2
Domain3
Domain4
Data Acquisition
Data Gap and Collation Analysis
Data QC and Cleaning
Data Acquisition (from published and unpublished sources)
Identification of Data and Sources
Data Source1
Data Source2
Data Source3
Data Source4
Heterogeneous Data Sources as per Domain Knowledge

Figure 3.12: An integrated workflow – a methodological framework

Keeping in view the research questions, objectives and components of the research methodology, the author designs an integrated workflow as described in Figure 3.12. As shown in Figure 3.12, data acquisition, data modelling and information analysis largely describe a framework. In each description, how the acquired data are quality controlled, how modelled data are organized in a way the information produced, are processed and interpreted for further evaluation and implementation at later stages. It is worth mentioning that all the events, such as data acquisition, data modelling, and information analysis, interconnect in a way an integrated process does. Detailed description of components of research methodology are given in the following sections.

3.6.1.1 Domain modelling methodology

As per DS guideline-2 (Hevner et al. 2004) and research questions and objectives, described in Sections 1.3.1 and 1.3.2 (RQ1 and RO1) of Chapter 1, the author intends to model multiple domains. For example, within a petroleum ecosystem, different domains do exist such as conventional, unconventional and shale-gas. Domain modelling, in general, describes a model of real world entities or objects or dimensions and the relationships built among them. Domain is again representation of multidimensional, either in space, time or depth domains. Identification of kind of entities, objects and or dimensions in each domain, provides an effective basis for understanding maintainability, testability and incremental development of a system. The domain modelling facilitates filling the gap between understanding of the problem domain and an ultimate interpretation and or knowledge domain description. In case of an object modelling, the domain modelling envisages as a solution and set of domain objects that collaborate and fulfil system-level scenarios. The domain modelling enables understanding the flexibility, efficiency and granularity of system design and development. Whenever a system needs to be changed its design and methodology, at par with domain applications and requirements with which the agile organizations need to be driven, domain models have a scope to refactor, update and maintain to control inherent complexity of the system. In order to address a common language and fundamental structure, features of domain models need to be analysed. A drastic reduction in cost of maintenance and increase in system enhancement effort are big advantages of domain modelling. As an example, the Figure 3.13 demonstrates the real world G & G (geology and geophysics) problem domain that can be modelled from generalization to specialization addressing the granularity and flexibility of data models.

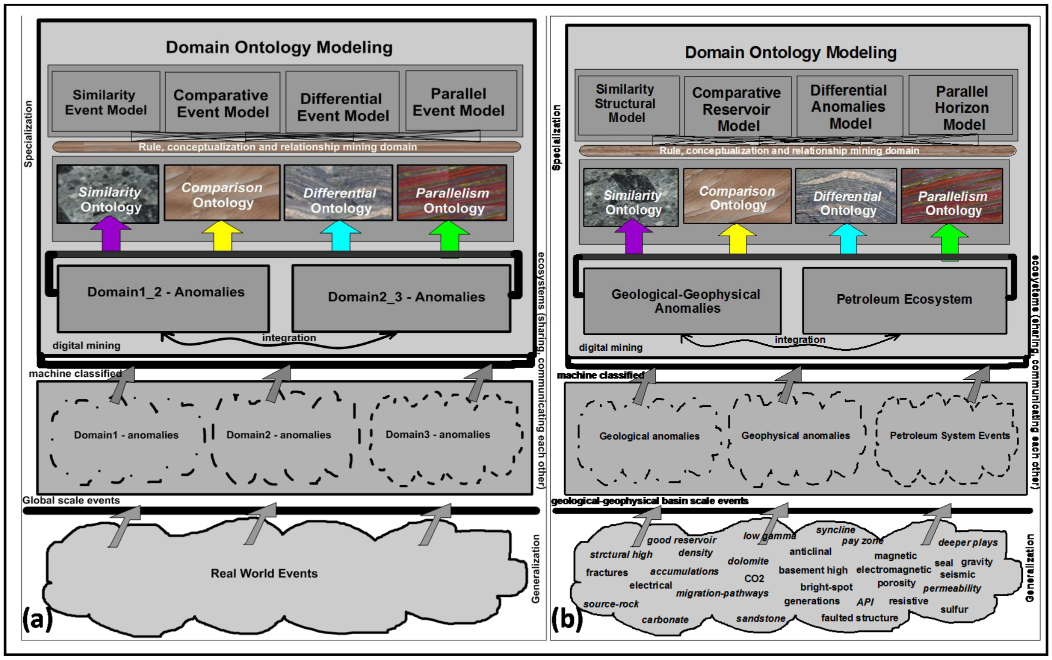


Figure 3.13: Domain modelling (a) a general process model (b) G &G domain modelling

The domain models, as described in Figure 3.13 and Figure 3.14, represent imposition of vocabulary and key concepts of the problem domain in any exploration industry. The author designs and uses the domain model (Figure 3.14), identifying the relationships among all the dimensions or entities, existing within the scope of the problem and application domain. This procedure commonly categorizes domains based on data types and attributes of data sources. A domain model is designed with business constraints and within scope of the problem definition. Several components of domain modelling are described from selection of domain to its analysis, which domain is appropriate for domain modelling and its connectivity with other. Once a domain is decided, then data associated with that domain are focused for acquisition.

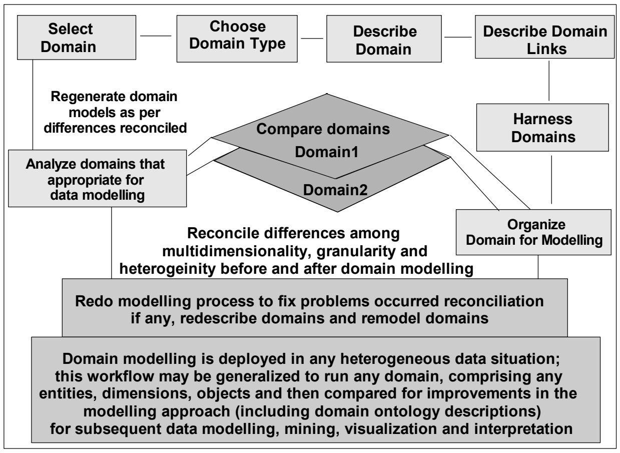

Select Domain
Choose Domain Type
Describe Domain
Describe Domain Links
Regenerate domain models as per differences reconciled
Compare domains
Domain1
Domain2
Harness Domains
Analyze domains that appropriate for data modelling
Organize Domain for Modelling
Reconcile differences among multidimensionality, granularity and heterogeinity before and after domain modelling
Redo modelling process to fix problems occurred reconciliation if any, redescribe domains and remodel domains
Domain modelling is deployed in any heterogeneous data situation; this workflow may be generalized to run any domain, comprising any entities, dimensions, objects and then compared for improvements in the modelling approach (including domain ontology descriptions) for subsequent data modelling, mining, visualization and interpretation

Figure 3.14: Domain modelling - workflow

3.6.1.2 Data modelling methodology

Different types of data exist nature in multiple domains and dimensions. As per research objectives, described in Sections 1.3.1 and 1.3.2 (RQ1 – RQ3 and RO1, RO2) in Chapter 1, different entities and dimensions are identified for logical and physical modelling. User is responsible to narrate, which type of data to consider in the data modelling approach. Three levels of data modelling are proposed, such as conceptual, logical and physical levels. A conceptual model investigates or examines the highest level of data relationships, either among entities, objects or dimensions. In this analysis, more focus is on dimensions, for organizing and modelling multiple dimensions of heterogeneous datasets. No attributes and keys are described at this stage. In the logical data modelling, the author describes the data in detail and finer relationships, without any concern on physical organization. In the logical data model, the following features are included:

- All entities, objects and dimensions; their relationships among them.
- All attributes for each entity/object/dimension, are specified.

- The primary and foreign keys are specified, identifying relationships among entities, objects and dimensions.
- Normalization occurs at this level.

The steps for designing the logical data model are as follows:

1. Specify primary keys for all entities/objects/dimensions.
2. Find the relationships between different entities/objects/dimensions.
3. Find all attributes for each entity/object/dimension.
4. Resolve many-to-many (or one-to-many where ever exists) relationships.
5. Normalization/Denormalization.

On comparing the logical data model shown above with the conceptual data model diagram, the main differences found between the two are:

- In a logical data model, primary keys are present, whereas in a conceptual data model, no primary keys are present.
- In a logical data model, all attributes are specified within an entity/object/dimension. No attributes are specified in a conceptual data model.
- Relationships between entities/objects/dimensions are specified using primary keys and foreign keys in a logical data model. In a conceptual data model, the relationships are simply stated, not specified, ensuring that two entities/objects/dimensions are known to be related, without specifying what attributes and their relationships exist in the model.

The conceptual and logical models are different, as narrated in Figure 3.15 and Figure 3.16:

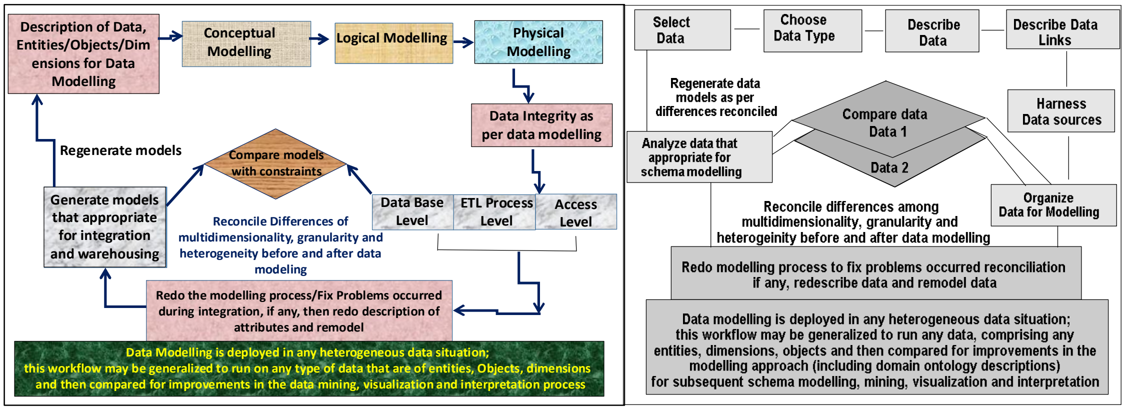

Description of Data, Entities/Objects/Dimensions for Data Modelling
Conceptual Modelling
Logical Modelling
Physical Modelling
Data Integrity as per data modelling
Regenerate models
Compare models with constraints
Generate models that appropriate for integration and warehousing
Reconcile Differences of multidimensionality, granularity and heterogeneity before and after data modeling
Data Base Level
ETL Process Level
Access Level
Redo the modelling process/Fix Problems occurred during integration, if any, then redo description of attributes and remodel
Data Modelling is deployed in any heterogeneous data situation; this workflow may be generalized to run on any type of data that are of entities, Objects, dimensions and then compared for improvements in the data mining, visualization and interpretation process
Select Data
Choose Data Type
Describe Data
Describe Data Links
Regenerate data models as per differences reconciled
Compare data
Data 1
Data 2
Harness Data sources
Analyze data that appropriate for schema modelling
Organize Data for Modelling
Reconcile differences among multidimensionality, granularity and heterogeinity before and after data modelling
Redo modelling process to fix problems occurred reconciliation if any, redescribe data and remodel data
Data modelling is deployed in any heterogeneous data situation; this workflow may be generalized to run any data, comprising any entities, dimensions, objects and then compared for improvements in the modelling approach (including domain ontology descriptions) for subsequent schema modelling, mining, visualization and interpretation

Figure 3.15 (a): Data modelling scheme – workflow, (b): an updated workflow

The author describes primary keys in a logical schema, but no keys are described in conceptual schema. Attributes existing within an entity or dimension are described in a logical schema, whereas no attributes are specified in the conceptual schema. While describing relationships among either entities or dimensions in a logical schema, several primary and foreign keys are used. In conceptual schema, only relationships are stated, but not specified. Two entities or dimensions related, are known, but not specified what attributes are used to relate or connect the entities or dimensions in a conceptual schema.

A physical data model describes how the data and model can physically be fitted within a database. Physically organized databases show all table structures, including column name, column data type, column constraints, primary and foreign keys and relationships among tables. While building physical data models, the following features are incorporated:

- Specification of all tables and columns.
- Foreign keys are used to identify relationships among tables.
- Denormalization may occur based on nature and type of industry and user requirements.
- Physical considerations may cause the physical data model to be quite different from the logical data model.
- Physical data models are different for different RDBMS.

The steps followed, for physical data model design, are as follows:

1. Convert entities or dimensions into tables.
2. Convert relationships into foreign keys.
3. Convert attributes into columns.
4. Modify the physical data model based on physical constraints / requirements.

The complexity increases from conceptual to logical and physical. It is important to understand the high level entities, objects and dimensions of the data and how they relate among each other. More details are added in the logical models without any implementation details.

For the purpose of finetuning the workflow, the author compares data models for achieving the objectives as demonstrated in Figure 3.15b. Ultimately, the implementation of data model is done by physical modelling, which has been detailed in the Figure 3.16.

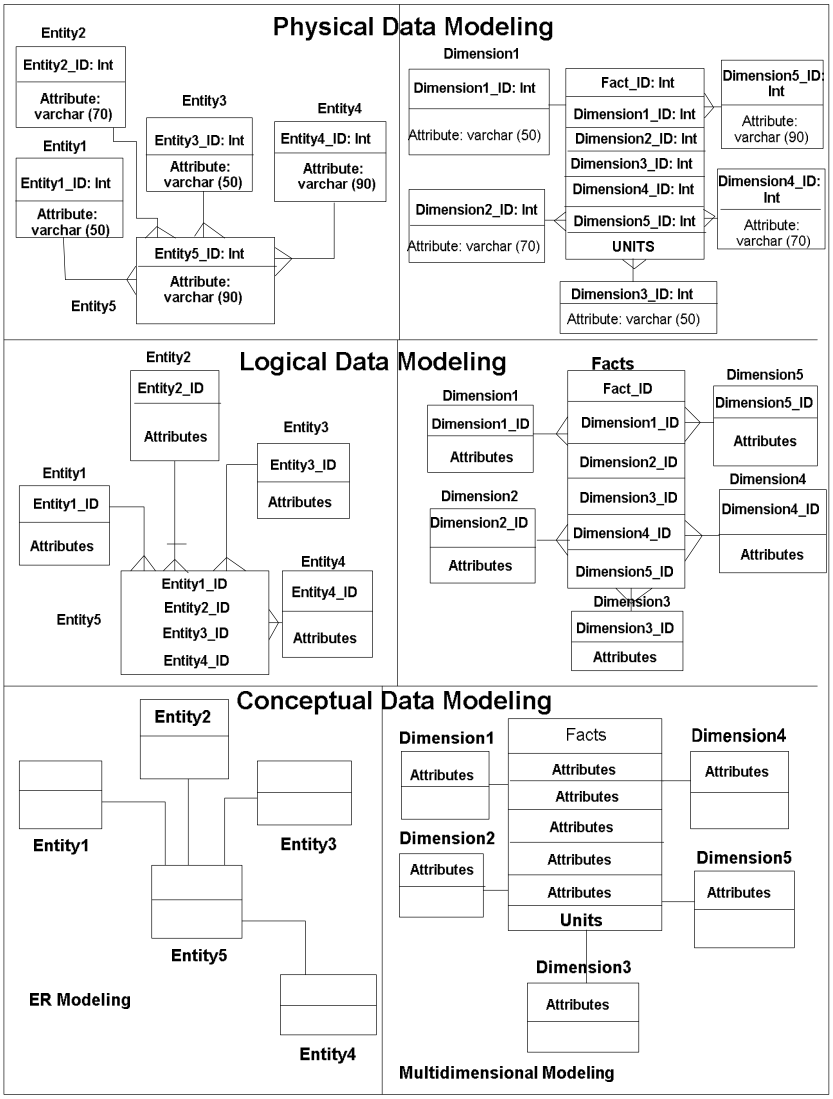


Figure 3.16: Comparison between data modelling approaches

3.6.1.3 Normalization and denormalization

The normalization is done with a specific purpose to maintain the data integrity (Hoffer et al. 2005 and Coronel 2011). But in real life situations, where multiple dimensions and domains are involved in data warehouse projects, typically data redundancy needs to be considered for performance and maintaining the history. A comprehensive normalized schema is designed, showing current state of the database. For example, *drilledWell-survey* application, a normalized design is intended to be developed, keeping the geography and period in the *operator* table. Suppose, if the *operator* attribute moves, geographic attribute is updated with new one. During the transition period, *operator* misses some details of *survey* information, requesting for the lost information. Sometimes, it is difficult to retrieve the exact information, if the *survey* data are not appropriately time-stamped or tagged with history of *operator*'s address. Issue may have been resolved, if the copy of the operator's information is documented as the *surveydate* in the *survey* table. Similarly, *operators* change the names, at times, copy of *operator* names need to be maintained as a part of *survey* acquisition/documentation in the *survey* tables as well.

The tables are decomposed into many tables during normalization process. More tables imply more joins are performed. More joins at times, have negative impact on the performances including maintenance of periodic volumes. A foreign key is replicated from the first child table to the other one. This process is continued for all the joins. In the case of *contractor* parent table, the *contractorID* is added to the *operator* table to serve as the foreign key, because every *operator* has information about the surveys and typical queries must find surveys together with operator's information, requiring joining three tables' *contractor*, *operator* and *surveys*. However, if the *contractorID* column in the invoices table replicated, same goal is achieved, satisfying the same query by joining, only the *surveys* and *contractor* tables.

Using derived data in a normalized system, is the best way to increase performance. For example, the need of *survey* is assessed, whether proposed well needs to be drilled in an area. All the data views needed for analysing the events are aggregate of all the events of area under investigation, just to tell the operator whether the drilled well, if available in the area that data need to be purchased, if no additional well drilling needed. The problem is solved by maintaining all the drilled-well information in the study area and by maintaining a separate DrilledWell – Warehouses table that hold the number of wells_drilled and available information in the warehouses (if multiple warehouses are made available). Finally there may be many queries that aggregate production data from operators. In a fully normalized system, all *survey* details are aggregated for a single operator multiple times. Performance can greatly be improved by maintaining *year-to-date* production summary in a column of the *operator* table. Figure 3.17 shows ER diagram of *surveys* database before normalization and after denormalization.

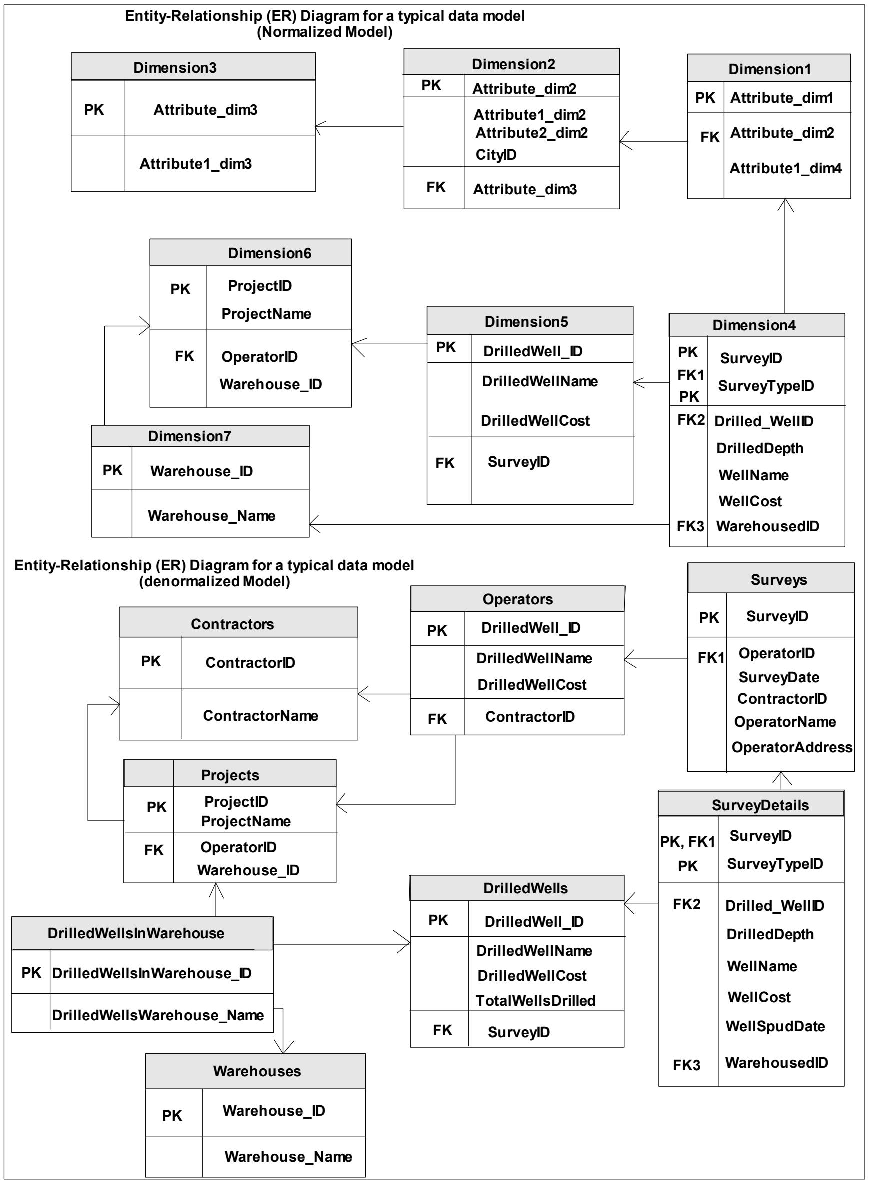


Figure 3.17: An Entity-Relationship Model (ERM), showing normalization and denormalization events

The denormalization brings out the risk of updating anomalies back to the database. Wherever necessary, it is done deliberately. Denormalization is systematically illustrated and assessed; ensuring that application correctly maintains the denormalized data and transactions are thus used appropriately. Here transaction is the smallest unit of work that must either fully complete or not at all. For example, in the *surveys* system, a transaction implies an insertion to a table such as SurveyDetails, which must be followed immediately by an update of the derived column TotalDrilledWells in the *Wells* table or a DrilledWellStatus column in the WellsInWarehouse table depending on where the drilled-well information is maintained. If one of the actions is failed, the entire transaction is rolled back to ensure the data consistency. DML (data manipulation language) triggers are used for maintaining the data. RDBMS fires a trigger automatically as a part of transaction. In the *surveys* example, a DML trigger for insert, update and delete on the SurveysDetails table can maintain the TotalDrilledWell table and WellsInWarehouses table. In addition, procedures are in place to rebuild the derived data if necessary from scratch. Data are rebuilt from events table in case of inconsistency between sum of events and states. After denormalization, denormalized data of transactions are maintained; for example, update of DrilledWell table whenever the status of a particular well is known.

1 Denormalization of databases

For the purpose of improving the performance, a data model is denormalized.

The current design is:

Projects (ProjectID, ProjectName, OperatorID)
ProjectDetails (ProjectID, ItemID, ActivityID, ContractorID, WorkDate, TimeSpent)
Operators (OperatorID, OperatorName)
Activities (ActivityID, ActivityName)
Contractor (ContractorID, ContractorName)

2 Denormalization to maintain history

The author performs denormalization to maintain history. For example, system needs to maintain an operator's name from the time; a project is started for an operator. An operator changes the name later. How do we achieve this requirement? In the *operators* table, the existing name is maintained. An operator name is added to the *Projects* Table. In addition, it is always a good practice to have information on the *date* when the operator name is valid. The start date is added in the *project* table as well. After denormalizing the attributes, *StartDate* and *OperatorName*, the improved design looks like:

Projects (ProjectID, ProjectName, OperatorID, StartDate, OperatorName)

Here StartDate could be SurveyDate or WellSpudDate. Denormalization improves performance. For example, a report calculates total time spent on an exploration project. When multiple projects are tracked, multiple times are reported in a day. Improved design of the data model, after denormalizing attributes TotalTimeSpent, TotalTimeExploration, TotalTimeDrilling and TotalTimeTakenProduction:

Projects (ProjectID, ProjectName, OperatorID, StartDate, OperatorName, TotalTimeSpent)

Attributes, TotalTimeExploration, TotalTimeDrilling and TotalTimeTakenProduction are further denormalized from composited attribute such as TotalTimeSpent.

Projects (ProjectID, ProjectName, OperatorID, TotalTimeExploration, TotalTimeDrilling and TotalTimeTakenProduction)

3.6.1.4 Schema modelling methodology

Several schemas exist in the literature (Hoffer et al. 2005). The author proposes the star-, snow-flake and constellation schemas, since they are compatible accommodating them in multidimensional heterogeneous data structuring process, in a warehouse environment. Several such schemes are provided in the Figure 3.18. Oil and gas exploration and production datasets are characteristically multidimensional and heterogeneous. For example, spatial-temporal dimensions are characteristic in nature, as they are geographically or periodically varying, especially among many oil and gas industry situations. The author uses schema architectures in petroleum domain as documented in the Figure 3.18.

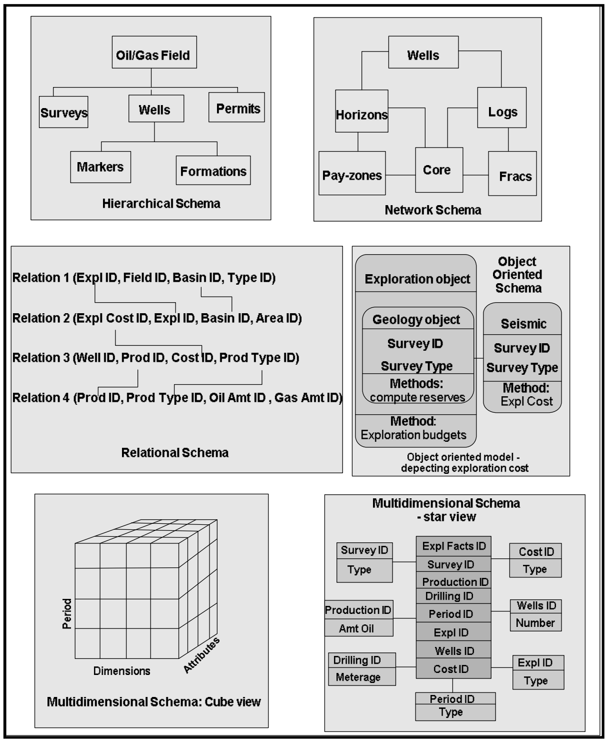


Figure 3.18: Different architectural schemas used in the current study

In a warehouse modelling approach, the author keeps key criteria as integration of different data structures, described from multiple domains. Data structures are multidimensional, representative of different multidimensional schemas. Figure 3.19 provides a workflow, in which data structure schemes used for modelling multidimensional datasets. These multidimensional data structures modelled from multiple domains are in the form of metadata, representing description of hundreds of multiple dimensions, simulated as if they are embedded in nature. All the entities, objects and dimensions are identified to select the structuring methodology, normalize/denormalize data relationships and construct schemas through either star, snow-flake and or constellation with their combinations.

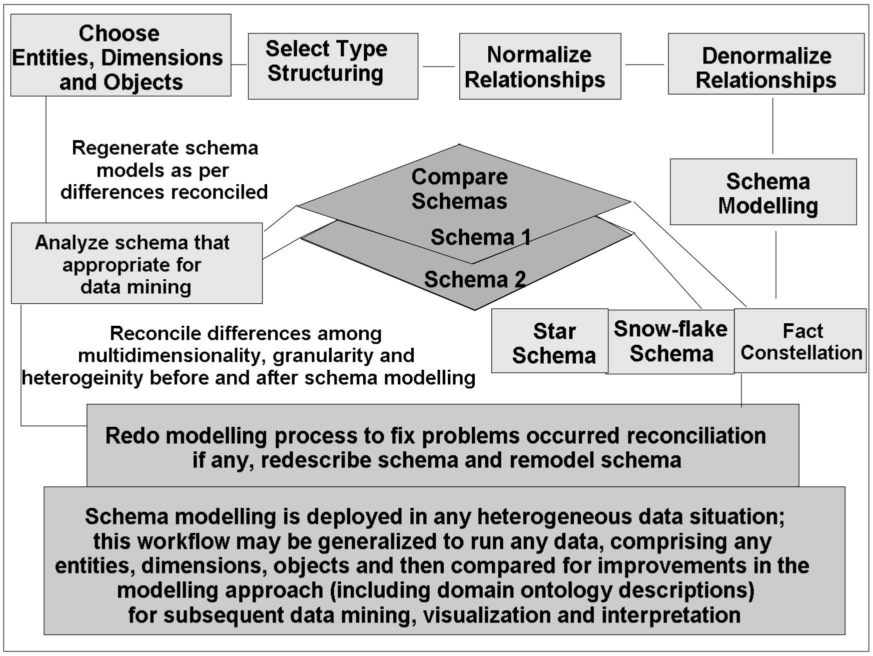

Choose Entities, Dimensions and Objects
Select Type Structuring
Normalize Relationships
Denormalize Relationships
Regenerate schema models as per differences reconciled
Compare Schemas
Schema 1
Schema 2
Schema Modelling
Analyze schema that appropriate for data mining
Reconcile differences among multidimensionality, granularity and heterogeinity before and after schema modelling
Star Schema
Snow-flake Schema
Fact Constellation
Redo modelling process to fix problems occurred reconciliation if any, redescribe schema and remodel schema
Schema modelling is deployed in any heterogeneous data situation; this workflow may be generalized to run any data, comprising any entities, dimensions, objects and then compared for improvements in the modelling approach (including domain ontology descriptions) for subsequent data mining, visualization and interpretation

Figure 3.19: Schema modelling – workflow

The modelling process is reconciled till the objectives of multidimensionality, heterogeneity and granularity are achieved. A multidimensional data model identifies the dimensions, their hierarchies, the measures, units, functions, for the design of the data cube. After having data relationships normalization done, author captures all the data and necessary information needed to build the multidimensional schemas that represent data cubes. The author uses multidimensional arrays to define data cubes and the well-known arithmetic matrix formulations (O'Brien and Marakas 2009) for OLAP operations. The schemas are modelling paradigms, in which the data warehouse contains a large, single, central fact tables and set of smaller dimension tables, one for each dimension, data relationships denormalized with fine grained schemas and sub-schemas. Here fact tables contain detailed summary of data or their conceptualized relationships. Each dimension has a single denormalized table. Every tuple in the fact table consists of the fact or subject of interest, and the dimensions. Several such data schemas used in the study are provided in Figures 3.20 and 3.21.

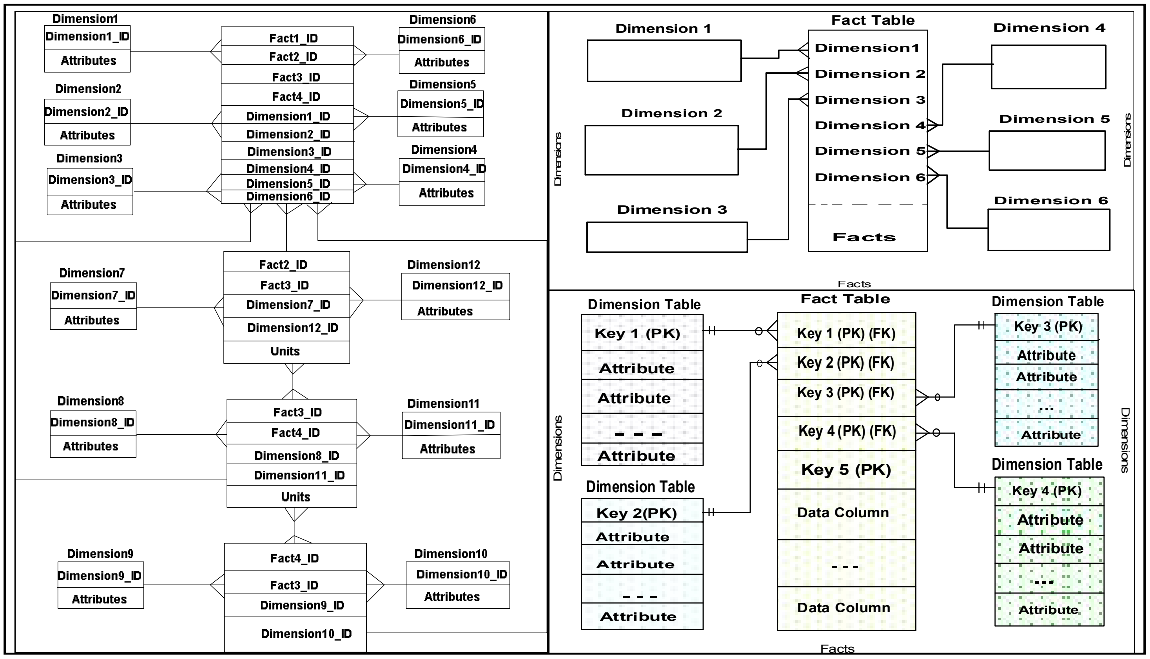


Figure 3.20: Multidimensional schemas

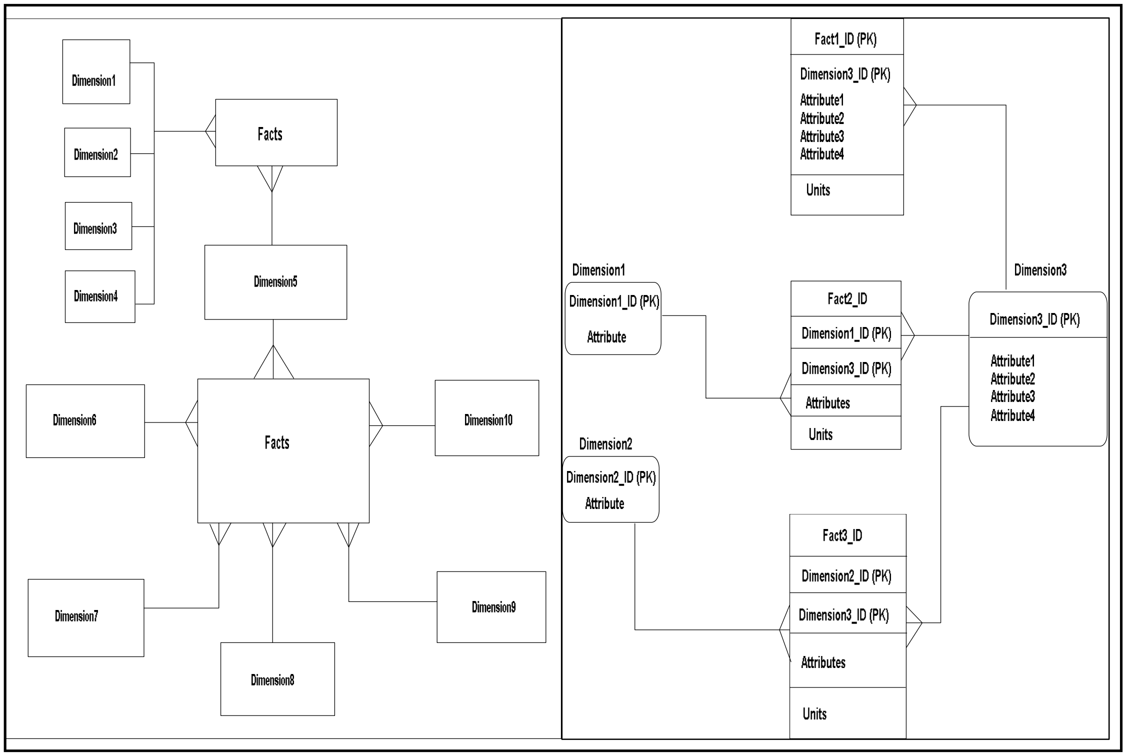


Figure 3.21 Fact constellation schemas

3.6.1.5 Data mining schemes

In this section, the author addresses RQ4, RQ5 and RO3. Mattison (1996) and Marakas (2003) investigate the effectiveness of data mining from warehoused data structures. Structured associated patterns (Matsuzawa and Fukuda 2000) are explored from warehoused metadata. The scientific goals (Neuman 2000) of data mining are:

1) Explanatory - explain observed events or conditions (such as why exploration expenditure has been increased in a particular period of time);
2) Confirmatory - confirm a hypothesis, (such as whether a particular field can produce similar quantity of oil and gas, compared with other fields);
3) Exploratory - to analyze data for new or unexpected relationships (such as what oil and gas exploration expenditure patterns are likely to appear in a particular period of time).

As examined in Figure 3.22, the author focuses on building several known and undiscovered relationships among several oil and gas *surveys*, *wells* and *permits* data. These relationships are incorporated in the modelling and integration process, which have an impact on refinement of data structuring process.

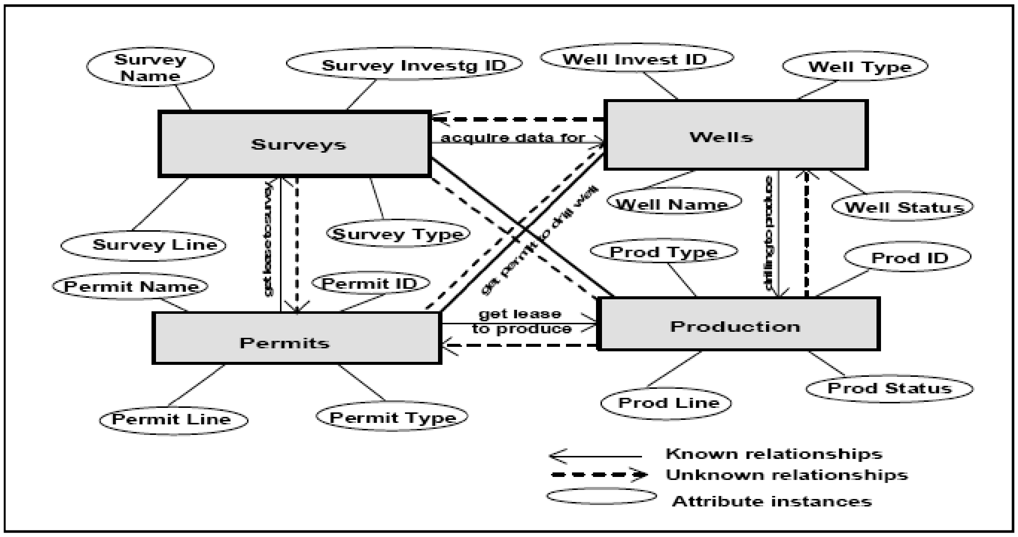


Figure 3.22a: Building ontology relationships among *surveys*, *wells*, *permits* and *production* data

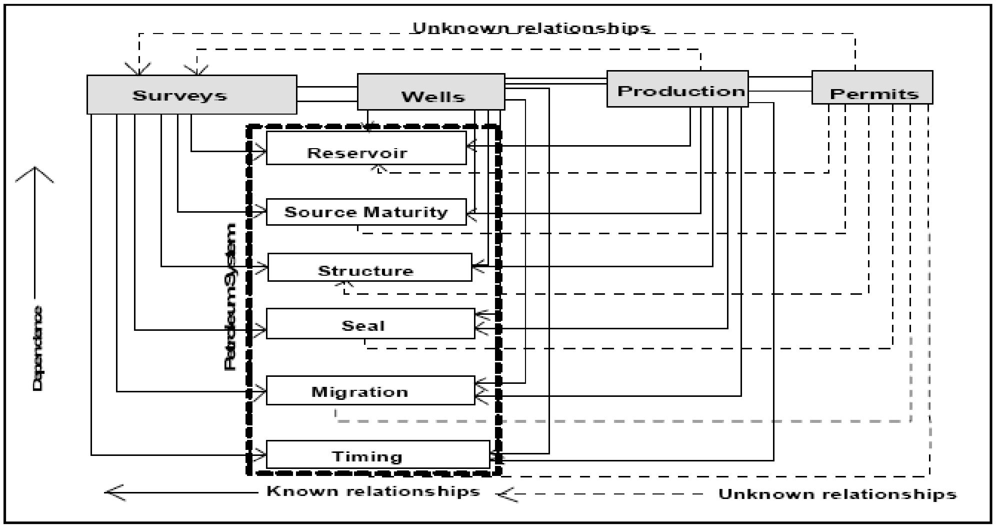


Figure 3.22b: Building ontology relationships among petroleum system elements

The author uses data mining schemes for extracting user specific data for interpretation purposes. Data mining processes (Pujari 2002) include individual or combined tasks of identifying data associations, classifying, regression mapping, time series analysis, prediction, clustering, summarization, associative rules, and sequence patterns discovery from warehouse compromising of heterogeneous data sources. If these processes are combined, data mining can become effective, since a single process may generate anomalies or ambiguities in its interpretation. Data mining is productive, when data extracted from warehouses or ontology repositories adhere with the following mining tools:

1. Run SQL queries or other data mining logics – for accessing required data from warehoused metadata; temporal data views shown in Figure 3.23, are such examples;
2. Interpreting the data views – for extracting shallow, multidimensional, hidden and deep knowledge – by means of correlations, trends and patterns perceived from data mining; build knowledge base models using mining algorithms;
3. Provide data mining solutions for detailed exploration or knowledge of heterogeneous data cluster (Figure 3.24): partitioning; hierarchical clustering; categorical clustering (Pujari 2002); (for example, CF-Tree structure algorithm);
4. Association rule mining (for example, FP-Tree growth algorithm), (Zhang 2000);
5. Build decision trees (for example, splitting indices, splitting criteria, pruning);
6. Extract and interpret hidden knowledge in support of business decision making analysis (for example, computing gain and entropy).

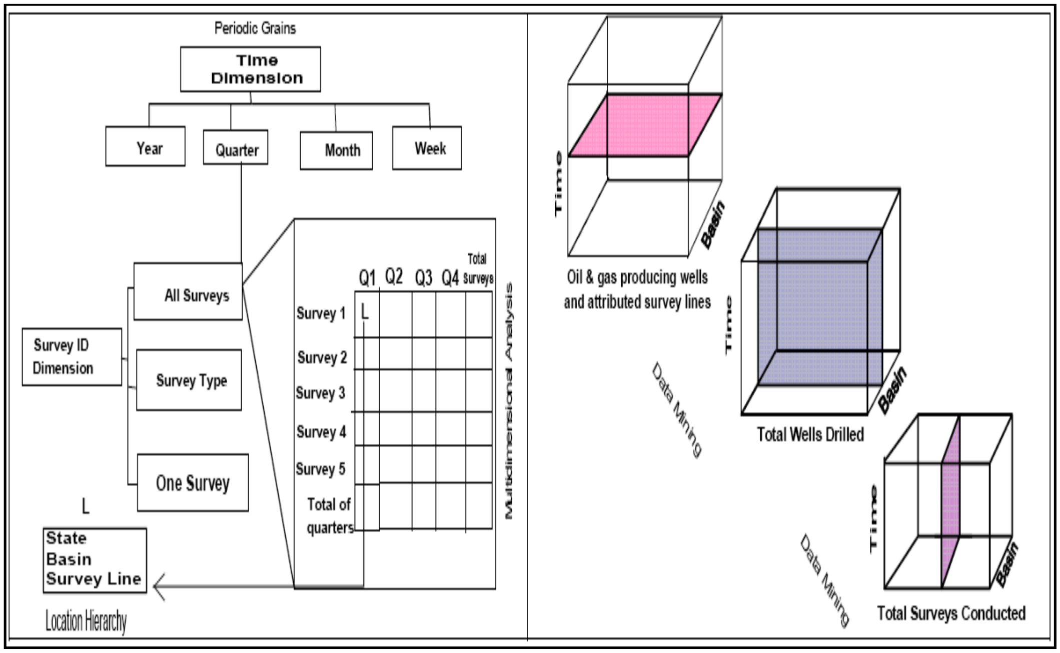


Figure 3.23: Periodic data views from temporal data

P*eriodic* dimensions are included *among surveys* data sources and integration process. Data views extracted with reference to the *period* dimension are representative, for carrying out an interpretable information as explained in Figure 3.23. In other words, data mining of *surveys'* volumes that characterize with multidimensional visualization, perceives an interpretable new knowledge in such multiple domains.

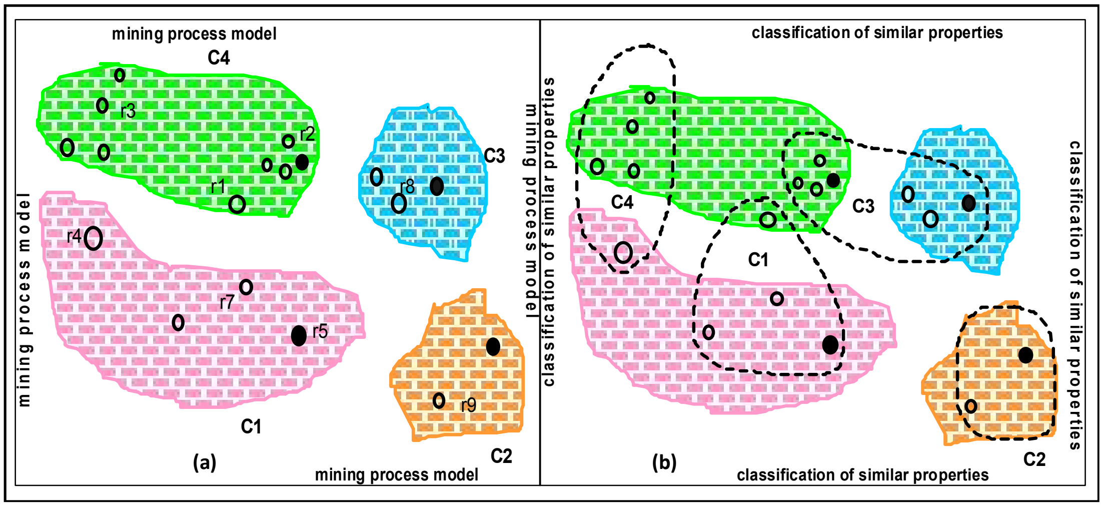

mining process model
C4
r3
r2
r1
C3
r8
r4
r7
r5
C1
r9
C2
(a)
mining process model
mining process model
mining process model
classification of similar properties
classification of similar properties
C4
C3
C1
C2
(b)
classification of similar properties
classification of similar properties

Figure 3.24a: Data mining (a) Partitioning before swapping; (b) Clustering of business data properties after swapping

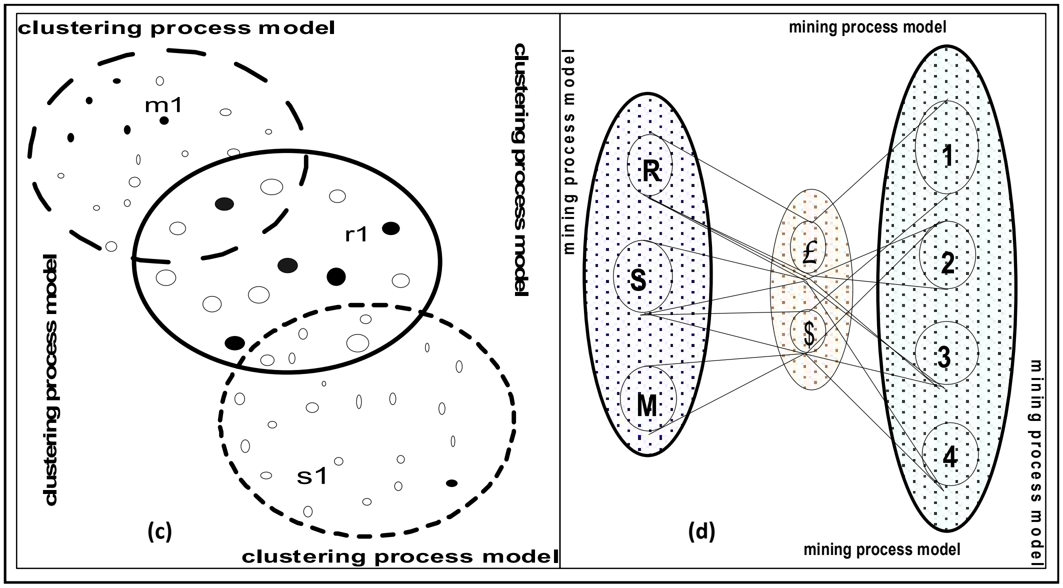


Figure 3.24b: Data mining (c) Density based spatial clustering of arbitrary shapes; (d) Weight based clustering

The entities (or dimensions) and their associated attributes identified from all the data sources are given weighting counters. Based on number of counters and strength of counters, clusters are formed. The author considers the cluster of the weighted counters as the measure and strength of a composite attribute in which all the low level attributes responsible for making the composite attribute are better clustered. The strength of associativity is interpreted among attributes and also among clusters for building new knowledge. The author carries out grouping of heterogeneous data attributes using factor, cluster and discriminate analysis and for this purpose simple mining algorithms are written in C++ as programming modules. This methodology makes the data analyst easy to prioritize the business (such as oil and gas exploration and development) strategies at viable economics. The author aims at this analysis to facilitate and ascertain which data structure is playing a definite role in what data mining process data model.

SQL Mining: The author uses MS Access and Oracle DB programs to load the data models, populating their data instances. SQL queries are run to extract structured information from Oracle driven data warehouses. Data views extracted through SQL mining, describe expenditure analysis on mineral and oil and gas exploration for the past 100 years, attempts to summarize major influences on that expenditure. Expenditure patterns of both mineral and oil and gas exploration are examined for extracting any intelligent information pertained to economic performance of the resources industry. Production patterns with respect to the expenditure in mining and oil and gas sectors are to be examined. Data pertained to the exploration, lease and state wise expenditures and basin wise production rates in their respective states are considered to explore patterns, trends, for any correlations and clusters between the data tables. Data mining performs two basic operations: predicting trends and behaviours, correlations and identifying previously unknown patterns. Multidimensional analysis provides a scope to users with a view “what is happening”, “why it is happening” including predictions of “what will happen in the future”. Other operations include:

- Data mining automates the process of finding predictive information in large databases.
- Data mining identifies previously hidden patterns in a single step.

Spatial data mining

In the current research, spatial data sources contain geographic dimensions and their instances are used in the modelling and integrating them with other attribute dimensions of oil and gas business data sources. Different countries and petroleum systems of geographic extents, possessing X, Y, Z coordinates, are typically incorporated within dimension and fact tables of data models. Several data structures, such as hierarchical, relational and networking approaches are used. A top-down approach can answer spatial data mining queries. With the hierarchical structure of grid cells on hand, a top-down approach is used to answer spatial data mining queries. For each query, cells on a high level layer are examined. Note that it is not necessary to start with the root; it may begin from an intermediate layer.

Mining modelling

Discovering interesting patterns, trends and correlations among large volume of datasets and interpreting them for knowledge discovery are objectives of data mining. The data instances, which are stored in multidimensional data warehouses, or other information repositories are used for searching patterns. Data mining tasks are classified into two main categories: *descriptive* and *predictive.* Descriptive tasks characterize, the general properties of the data in the database. Predictive tasks are performed for inferring the data predictions. All these descriptions and predictions are presented in the form of plots, graphs and correlograms in the forthcoming Chapters 4 and 5. The author provides different mining schemes used in the current studies and a robust workflow designed for mining data views in the Figure 3.25.

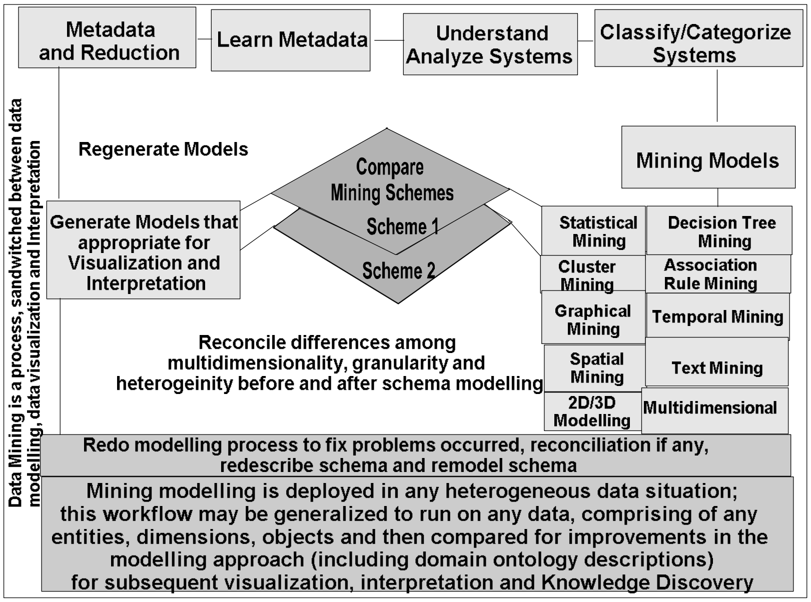

Metadata and Reduction
Learn Metadata
Understand Analyze Systems
Classify/Categorize Systems
Data Mining is a process, sandwitched between data modelling, data visualization and Interpretation
Regenerate Models
Compare Mining Schemes
Scheme 1
Scheme 2
Mining Models
Generate Models that appropriate for Visualization and Interpretation
Statistical Mining
Decision Tree Mining
Cluster Mining
Association Rule Mining
Graphical Mining
Temporal Mining
Spatial Mining
Text Mining
2D/3D Modelling
Multidimensional
Reconcile differences among multidimensionality, granularity and heterogeinity before and after schema modelling
Redo modelling process to fix problems occurred, reconciliation if any, redescribe schema and remodel schema
Mining modelling is deployed in any heterogeneous data situation; this workflow may be generalized to run on any data, comprising of any entities, dimensions, objects and then compared for improvements in the modelling approach (including domain ontology descriptions) for subsequent visualization, interpretation and Knowledge Discovery

Figure 3.25: Mining modelling – workflow

Data mining objectives are

- *Classification*

  Conceptualized attribute dimensions among several classes of datasets are in relationships. Capturing summaries and aggregations of the data and essential properties of the data models.

- *Knowledge Reduction/Knowledge Expansion*

  Knowledge is implicitly declared when two classes are in relationship. Classes in relationship, can be expanded with explicit knowledge.

- *Knowledge Derivation*

  A notion is proposed to characterize the maps from conceptual classes to logical classes. Each map is based on conceptual schema decompositions preserving the model equivalence, and can be interpreted as a link between conceptual and logical database schemas. These links are able to explicate classes implicitly declaring through graph of classes supported by semantic data models.

The Knowledge incorporation solution offers a tool for searching, in which the use of ontologies corresponding to the content and domain of the database is emphasize. Structured associated patterns (Matsuzawa and Fukuda 2000) are explored from warehoused data for relationships, such as what reservoir patterns are likely to appear in a given seismic and drilled-well data integration scenario. For achieving these mining goals, the following are performed:

1. Run SQL queries for accessing required data from warehoused metadata

2. Interpret the data views for extracting multidimensional knowledge – by means of correlations, trends and patterns perceived from data mining; build knowledge base models (Figure 3.26)

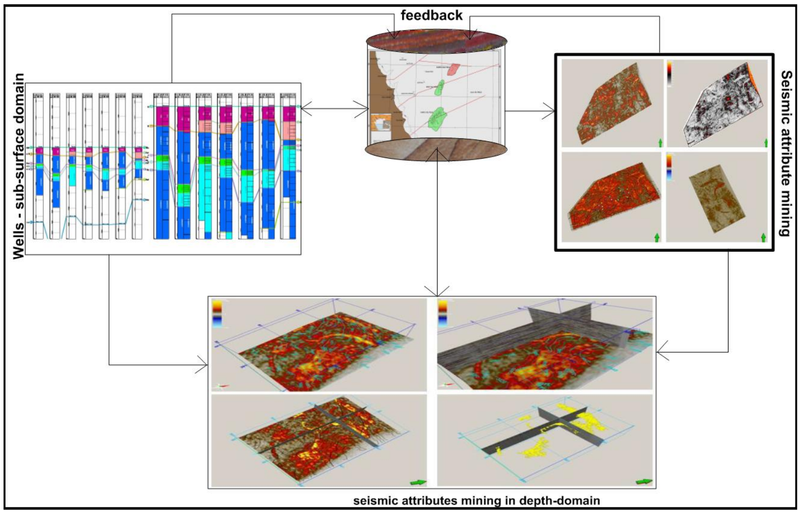

feedback
Wells - sub-surface domain
Seismic attribute mining
seismic attributes mining in depth-domain

Figure 3.26: Attribute data mining, views from integrated *surface* and *sub-surface* domains

As an example, the connectivity attained between surfaces (earth's) - to – subsurface domains in a warehoused repository, is shown with different data visualization views, providing new knowledge, which is inherent and unexplored. The author uses data mining rules to obtain the visualization views for interpretation. In the current research, associate rule, cluster and decision tree mining are designed for qualitatively analysing the patterns, correlations and trends of data views in different knowledge domains.

Associative rule mining

The problem of deriving associations among big data systems, has received great attention in recent years. The problem is to analyse *elements*, *processes* and *chains* of petroleum systems, occupied in large geographic regions (called *sedimentary basins*) by finding their associations among *elements, processes* and *chains and their attributes*. In the case studies, all these *elements, processes* and *chains* are described as dimensions. Basins, in terms of their geographic positions and periodic dimensions, produce enormous association rules that help analyse productive areas, with insights which system's *elements*, *processes* or *chains* can frequently occur, in which geographic regions and or periodic times. This can help explorers or researchers focus on these geographic regions for investments. Discovery of association rules depends on detection of frequent sets in data. Number of association rules make use of occurrences of presence of similar and or dissimilar systems and their elements. Constraints set by users and limits of the elements interpreted by users, are typical characteristics of mining rules. Grapher and surfer solutions are used in identifying the associativity among attributes of multiple dimensions in different application domains.

Cluster mining

Discovery and interpretation of dense and sparse patterns of regions of the data are goals of the cluster mining. Instances of attributes of large number of dimensions have different magnitudes as shown in Figure 3.27, thus creating different metrics or measurements, especially with respect to geographic dimensions. Clusters have multiple bubbles plotted in differing sizes and magnitudes that can measure the density of cluster. Number of attributes in such large number of databases is very large in spite of the size of individual existence and participation in the cluster is meagre. For example, similar and or dissimilar reservoirs of good qualities (possessing good porosities and permeability) that belong to a cluster though small, but larger quantities make cluster denser, compared to the dissimilar values narrating sparse patterns of regions. In the current research, author identifies several geographic and periodic patterns that describe clusters and for visualization and qualitative interpretation. Detailed studies on clustering paradigms are beyond the scope of current research.

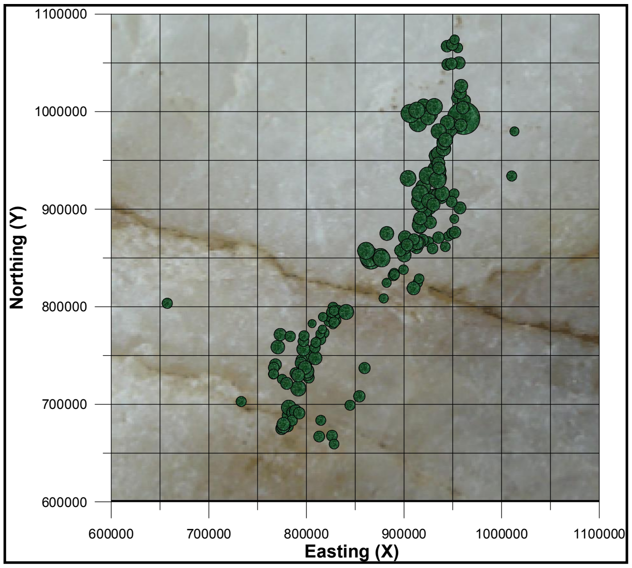
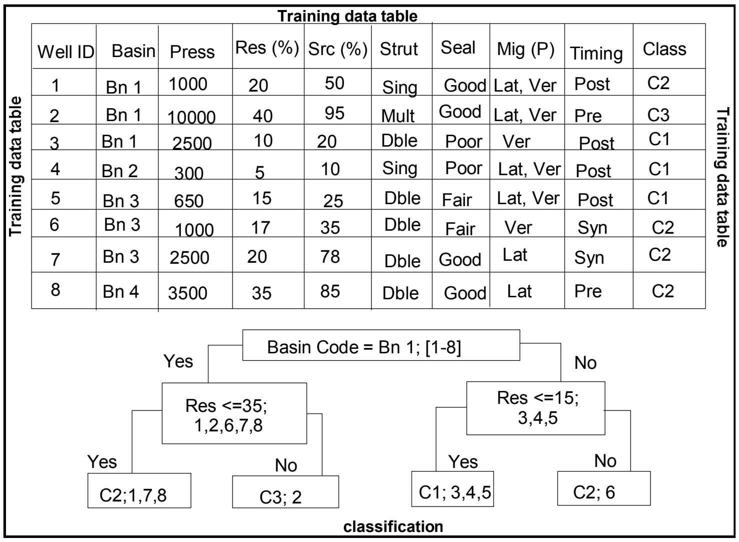


Figure 3.27: Clusters of *structure* anomalies drawn in a plot view, showing geographic trends for interpretation (sizes and densities of bubbles suggest attribute strengths)

Decision tree mining

Identifying classifications in large datasets, is an important problem in the data mining. Databases have number of records and sets of classes, each record belonging to a given class, problem of classification is to decide the class with which a given record belongs. The classification problem is concerned with generating description or a model for each class from the given dataset. For example, similar production data instances, in specific reservoir regimes can depict the classification. These regimes classified, are supervised by training the datasets. Using the training sets, the classifications attempt to generate the descriptions of the classes. These descriptions help classify the unknown records and sets of the databases. In addition to training, the author tests the datasets which may have been used to determine the effectiveness of a classification. Decision tree mining is a classification scheme which generates a tree and sets of rules, representing the model of different classes, from a given dataset. The sets of records available, for developing the classification, is generally divided into two disjoint subsets – a *training set* and *test set* (Figure 3.28). The former is used to describe the classifier and the latter is used to measure the accuracy of classifier. The accuracy of classifier determines the percentage test examples that are correctively classified.



Training data table

Well ID	Basin	Press	Res (%)	Src (%)	Strut	Seal	Mig (P)	Timing	Class
1	Bn 1	1000	20	50	Sing	Good	Lat, Ver	Post	C2
2	Bn 1	10000	40	95	Mult	Good	Lat, Ver	Pre	C3
3	Bn 1	2500	10	20	Dble	Poor	Ver	Post	C1
4	Bn 2	300	5	10	Sing	Poor	Lat, Ver	Post	C1
5	Bn 3	650	15	25	Dble	Fair	Lat, Ver	Post	C1
6	Bn 3	1000	17	35	Dble	Fair	Ver	Syn	C2
7	Bn 3	2500	20	78	Dble	Good	Lat	Syn	C2
8	Bn 4	3500	35	85	Dble	Good	Lat	Pre	C2

Figure 3.28: Decision tree mining

As described in Figure 3.28, decision tree structures generate understandable rules, especially among petroleum systems' elements. These tree structures handle numerical and categorical attributes and provide a clear indication of which element or process of a system is important for prediction/forecast or classification. More decision trees are described in Chapters 4 and 5, explaining the domain knowledge and interpretation of *reservoir* properties in different sedimentary basins' scenarios.

3.6.1.6 Data visualization methodology

In this section, the author addresses RQ6 and RO4. For this purpose, several graphic tools provide with one of most effective means of communication, because of highly developed 2D and 3D pattern-recognition capabilities. This allows perceiving and processing the pictorial and digital data more rapidly and efficiently. Data views are extracted from warehoused metadata and presents for visualization and interpretation. Mattison (1996) discusses a case study on oil and gas exploration, with an application to data visualization technique. By using visualization, data are summarized and the trends are highlighted. Unknown phenomena are uncovered through various kinds of graphical representation. The author uses several other visualization techniques to analyze spatial-temporal multidimensional data views. The visualization is one of the key features of big-data.

The methods that focus multidimensional data visualization are:

1) One method of visualizing the results of data search is through scatter plot display in a three-dimensional grid. (A scatter plot is a visualization technique that displays each data point as a color sphere, or bubble. Scatter plots are referred to as bubble viz.). The size, shape and color of the bubble each can be used to represent a variable in the data (Figure 3.27, more bubble plots are presented in Chapters 4 and 5). The three axes in this visualization ideally should be able to represent any dimension within the data and should be able to be randomly selected and modified by the user. The user can choose to explore more fully those data points using visual methods or perhaps click on the data and see a traditional numerical display in a spreadsheet. The volume of data points can also be rotated to observe clusters in different areas. By preparing and presenting the data graphically, the user can uncover properties of the data and quickly and easily detect any patterns or deviations from expected results.

2) Bubble charts are other ways of presenting data, because they convert pages of *hard-to-understand* numerical and textual data into something that is easily comprehensible to analyze. Bubble plot is very simple example of the use of graphics to quickly convey information about the data that they present. This bubble plot, representing bubble sizes, densities and trends, suggests several inferences such as *structure*, *reservoir* and *production* attributes, their strengths and magnitudes that can interpret qualitative and quantitative properties (of reservoirs, for example).

3) To visualize the data, it is typically read in as a 3D block of data (examples are given in Chapters 4 and 5), called a *volume*. Because the data sets are so large, the data is not read into the computer all at once but is converted in *chunks*. Each point of data in the selected area represents a physical location (i.e., an X, Y, Z location) in the 3D

space represented by that particular area. The value (or values, depending on the data acquisition methods used) at each data point represents many attributes or properties (such as amplitude, phase, frequency, velocity etc.). A collection of surrounding values in a given area, is in turn identified (to a certain degree of probability) as the type of entity (or object) or a dimension at each geographic location, as a lateral dimension.

4) Another method that assigns a value to each data point (or more typically to a range of data points), corresponds to color to display for that range of values or instances. All the points that fall within the selected range of values, have the same color. In a 3D representation of the object, the colors smooth out to form layers.

5) The author uses a visualization system (Han and Cercone 2000) for discovering numeric association rules among the structured resources business data.

The data visualization is a focus, keeping in view its increasing demand (even in big-data) for representing data views extracted from different domains for interpretation. Visualization explicitly facilitates domain interpretation from mining of fine grained metadata volumes. Visualization is a display of multiple views, such as map, plot, chart, bubble plot views and how multiple dimensions can be visualized in a single plot in order to make effective to visualize data patterns, trends and correlations more explicit. OLAP visualization is a presentation of metadata views of warehoused G & G and E & P datasets. As shown in the Figure 3.29, both periodic and geographic dimensional views are displayed for visual interpretation.

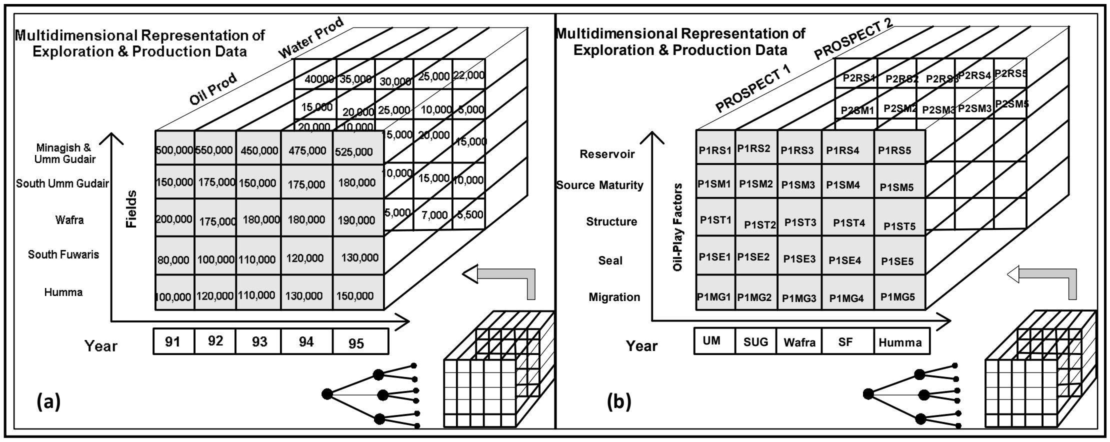


Figure 3.29: 3D Cube multidimensional representation of matured fields with reference to the elements of a petroleum system

The data visualization (Post et al. 2002) is the study of the visual representation and graphical depiction of data, meaning information that have been mined and processed in different schematic forms, including variables for different units of information. Data fusion is a sort of visualization, in which data instances are described in different graphic visuals in a way meaningful knowledge extracted from the interpreted data. The main goal of data visualization is to communicate information clearly and effectively through graphical means. To convey ideas effectively, both visual form and functionality need to go hand in hand, providing insights into a rather sparse and complex dataset by communicating its key-aspects in a more intuitive way. Yet designers often fail (Post et al. 2002) to achieve a balance between form and function, creating gorgeous data visualizations which fail to serve their main purpose — to communicate “information". Data fusion in recent years is on its wide usage in oil and gas industry situations in understanding domain knowledge. Without losing the clarity and perception, knowledge is achieved through visual representations and graphic displays of G & G and E & D data views, which has been demonstrated in Chapters 4 and 5.

Organizations associated with big-data, data distribution and cloud computing systems (synonymous to networking ecosystems in the present context), all involve with delivery of products of knowledge and information management including visualization of information (Figure 3.30). The author uses the visualization workflow as given in Figure 3.30. This process includes business analytics, delivering quality and interpretable information to variety of users. Business data analysis and presentation of processed data are criteria for interpreting the knowledge and information, extracted from warehoused metadata.

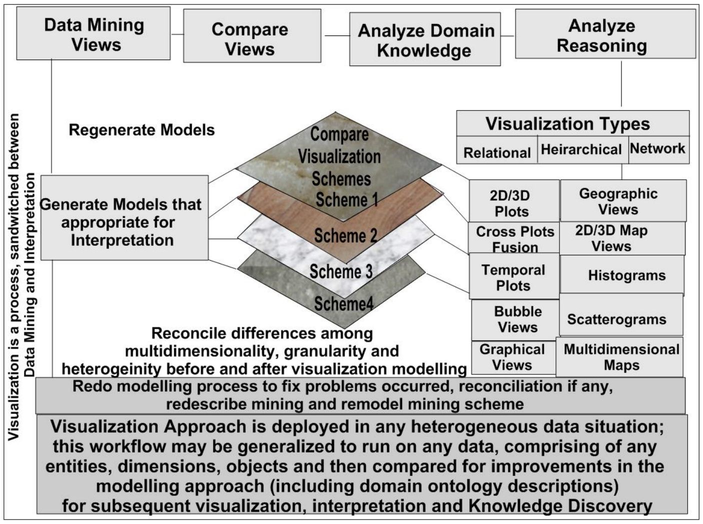

Data Mining Views
Compare Views
Analyze Domain Knowledge
Analyze Reasoning
Visualization is a process, sandwitched between Data Mining and Interpretation
Regenerate Models
Compare Visualization Schemes
Scheme 1
Scheme 2
Scheme 3
Scheme4
Visualization Types
Relational
Heirarchical
Network
Generate Models that appropriate for Interpretation
2D/3D Plots
Geographic Views
Cross Plots Fusion
2D/3D Map Views
Temporal Plots
Histograms
Bubble Views
Scatterograms
Graphical Views
Multidimensional Maps
Reconcile differences among multidimensionality, granularity and heterogeinity before and after visualization modelling
Redo modelling process to fix problems occurred, reconciliation if any, redescribe mining and remodel mining scheme
Visualization Approach is deployed in any heterogeneous data situation; this workflow may be generalized to run on any data, comprising of any entities, dimensions, objects and then compared for improvements in the modelling approach (including domain ontology descriptions) for subsequent visualization, interpretation and Knowledge Discovery

Figure 3.30: Visualization modelling – workflow

The information represented in the form of graphics, audio and video have immense use in interpreting spatial-temporal data and other heterogeneous datasets.

Visualization modelling has an impact in any application domain (Marakas 2003, Nimmagadda and Dreher 2012a). Visualization is designed for creative thinking, product ideation, and advanced business analytics. Questions such as "what" to explain the "why" of engineering graphics, are incorporated within the design of visualization models.

Design and development of data visualization modelling

The author uses various visualisation techniques for uncovering the patterns in the data. During data visualization process (Hevner et al. 2004, p.83, design science guideline 7), the author ensures that data views extracted from the warehoused metadata are at par with users' requirements so as to interpret them for new knowledge discovery.

3.6.1.7 Data interpretation methodology

The author addresses research question, RQ 6 and research objective, RO5 in this section. As a part of implementation and evaluation of the methodologies (one of the guidelines of DS research) used in the present study, author proposes several interpretation methodologies. The extracted data-views are interpreted for knowledge, thus for evaluating the effectiveness of integrated framework and data models designed in different knowledge domains. Data interpretation is crucial, which can test the validity of the data models, data warehousing and mining including effectiveness of visualization. The trends, patterns and correlations observed among data events are qualitatively interpreted, for understanding knowledge enhancements and interpretations. In addition, relevance, effectiveness, efficiency, impacts and sustainability criteria are described. The extent and duration of usage of data models and integrated framework including implementation of contextual, short- and long term research outcomes among latitude and longitude dimensions are interpretation objectives.

Data analysis and interpretation are meant for transforming the data collected or processed into meaningful knowledge and its interpretation. Interpretation outcomes ensure effective evaluation of data organization and descriptive analysis. Data interpretation is expected to confirm the measure, consistency and effectiveness of multidimensional and heterogeneous data organization, modeling, mapping, data mining including effectiveness of data visualization. Interpretation may be qualitative and quantitative and the data patterns, trends and correlations interpreted that lead to discovery of knowledge and thus implementing the knowledge. Interpretation is done for evaluating the data models based on the following criteria:

Relevance: methodologies, models, data mining, visualization are relevant to support the interpretation in different application domains.

Effectiveness: achieved the research objectives

Efficiency: within the available resources, maximum objectives and goals achieved
Impacts: there is an immense impact in the implementation of data models in various application domains and including an updated integrated framework

Sustainability: models, methodologies and implementation will continue to be extended in other domains

Impact based evaluations are refined based on the use of the models in multiple domains and applications with criteria:

1. ***Extent of use*** – how many stakeholders identified this approach and what degree outcome of research findings used
2. ***Duration and extent of usage*** – will the models, methodologies and implementations continue to be in multiple dimensions, such as geographical and periodic; to use for multiple countries and historical periods

Significant issues are, interpretation of research findings and understanding the research outcomes for evaluation and implementation purposes. Associativity and consistency among findings and research outcomes is another significant parameter that can leverage the methodologies and integrated frameworks used in the study. For example, an association or consistency exists when an event, such as implementation, happens in one domain, may not happen in another domain. In other words, associativity and consistency exist when one event is more likely to occur because another event taken place. Although the implementations may be associated, one does not necessarily cause the other or the second event may still occur independently of the first. For example, the current research on implementation of integrated frameworks in multiple domains supports associativity and consistency patterns, such as use and implementation of methodologies not only in petroleum industries, but in other domains. Patterns of implementations in different application domains support the idea of organization of heterogeneous and multidimensional datasets from multiple sources within integrated frameworks. Cause and effect of implementations in multiple domains can also determine the validity, integrity and consistency of methodologies. Data models described in one application domain, schematic representations hold good in other domains, in spite of semantics and syntactic inconsistencies. Semantic and syntactic inconsistencies exist during conceptualization of contextualization of specifications or ontological descriptions, even though data relationships mapped in different application domains are similar. Additional information or measures on fine tuning of data models are needed to fix these issues. Interestingly, denormalized multidimensional data models possess more flexibility and fine-grain of data relationships in building conceptualization and contextualization, especially among heterogeneous data relationships. This is demonstrated in Chapters 4 and 5.

Interpretation of cross-sectional and longitudinal dimensions

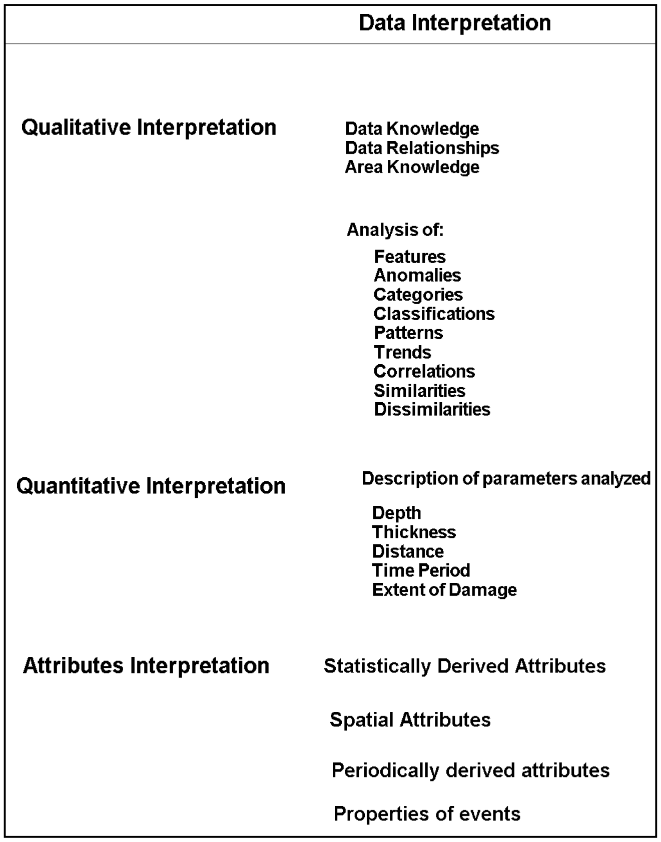
The interpretative research is part of design science IS research. Data views represented in multidimensional views provide insights of interpretation and anticipated domain knowledge. Multidimensional metadata and their data views are interpretative for systems analysis and development scenarios. Big-data from geographically spread countries and historical periods are significant in testing the current data models. Knowledge is built based on both short and long-term outcomes for interpretations. It is good idea evaluating the implementations of short and long term outcomes separately, so that a fair assessment and examination of time-frame and resources needed for the projects and sustained-impacts are understood at different stages. A comparative study of long- and short-term impacts are made, from cross-sectional and longitudinal analysis.

Contextual implementation

Interpretation of results or implementation outcomes are possible in proper contexts, which include what outcomes are expected from current implementations, based on similar implementations that have been made in previous studies or years.

Interpretation and knowledge modelling

Depending upon the domain application and knowledge, the author chooses interpretation objectives. But in the present context, author narrates methodologies as explained in Figures 3.31 and 3.32. Knowledge is built based on the analysis and interpretations. These are qualitative, quantitative and interpretations of attributes computed by mining methods. Anomalies are deviations from the common rules or standardized or expected values. Interpretation and qualitative analysis of anomalies are the basis of building knowledge, such as attitude attributes of petroleum systems' elements and or processes. The author does perform a quantitative analysis, by measuring thicknesses of reservoirs or their areal extents.



Data Interpretation

Qualitative Interpretation

Data Knowledge
Data Relationships
Area Knowledge

Analysis of:
Features
Anomalies
Categories
Classifications
Patterns
Trends
Correlations
Similarities
Dissimilarities

Quantitative Interpretation

Description of parameters analyzed
Depth
Thickness
Distance
Time Period
Extent of Damage

Attributes Interpretation

Statistically Derived Attributes

Spatial Attributes

Periodically derived attributes

Properties of events

Figure 3.31: Interpretation methodologies

Different data views are examined (extracted for visualization and interpretation) for anomalies and their evaluation and thus corroborating a specific model that achieves the knowledge interpretation objective (RO5, Section 1.3.1 of Chapter 1). For this purpose, different interpretation approaches are adopted for evaluating the qualitative and quantitative anomalies for knowledge discovery. The author attempts several interpretation schemes for achieving the set of objectives of interpretation as narrated in Figure 3.31 and Figure 3.32, with a focus on number and type of property attributes as mentioned in Figure 3.32 in the present study.

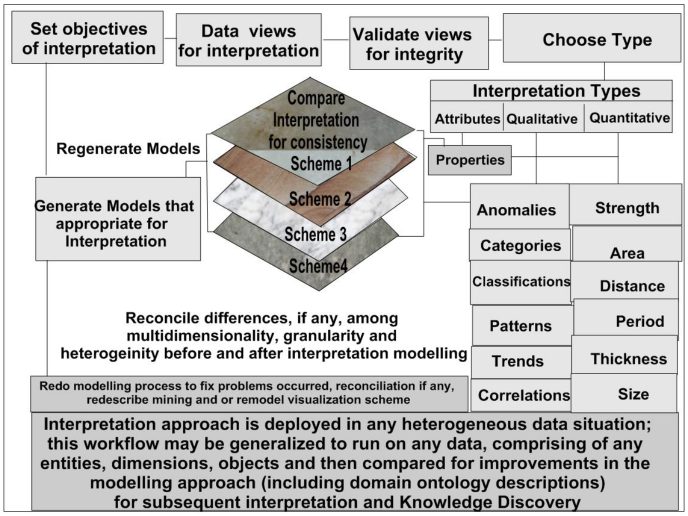

Set objectives of interpretation
Data views for interpretation
Validate views for integrity
Choose Type
Interpretation Types
Attributes
Qualitative
Quantitative
Compare Interpretation for consistency
Scheme 1
Scheme 2
Scheme 3
Scheme4
Regenerate Models
Properties
Generate Models that appropriate for Interpretation
Anomalies
Strength
Categories
Area
Classifications
Distance
Patterns
Period
Trends
Thickness
Correlations
Size
Reconcile differences, if any, among multidimensionality, granularity and heterogeinity before and after interpretation modelling
Redo modelling process to fix problems occurred, reconciliation if any, redescribe mining and or remodel visualization scheme
Interpretation approach is deployed in any heterogeneous data situation; this workflow may be generalized to run on any data, comprising of any entities, dimensions, objects and then compared for improvements in the modelling approach (including domain ontology descriptions) for subsequent interpretation and Knowledge Discovery

Figure 3.32: Interpretation modelling - workflow

The author takes advantage of ontologies for representing the knowledge and modelling the domain knowledge. Ontology supports storage and manipulation of knowledge, including drawing inferences and making decisions. Mechanism of generalization and specialization including classification facilitate semantics and fine tuning of knowledge representation. Selected data views consist of interesting patterns and trends, which may be descriptive and predictive. Properties of data are either qualitatively or quantitatively interpreted. Author uses attributes that depict spatial and periodically varying properties for interpreting data inferences.

Interpretation of data views

The author validates data views, taken for interpretation and type of interpretation, chosen for its consistency and knowledge discovery. Data are either qualitatively or quantitatively interpreted based on the objectives of interpretation and project goals. Models computed from statistical mining are used for interpreting their properties, more often for qualitative interpretation. Metadata that represent demographic, geographic and periodic data instances, is interpreted to have a meaningful information for decision support systems that assist in understanding system’s behaviour, further analyse for improvements. Measures, strengths and anomalies of the properties of the attributed dimensions are interpretable parameters. Limitations and uncertainties of systems are interpretable.

3.6.1.8 Knowledge discovery methodology

Data views are used for extracting domain knowledge for interpretation and supporting knowledge based systems. These data views are derived from metadata. The knowledge obtained in all case studies, is ensured with meaningful interpretations and implementation of metadata in different application domains. For example, an *element* within a petroleum system, is found to be more productive and its areal extents discovered is large enough, so that similar strength of attributes is predicted in other fields of associated systems. Creation and discovery of knowledge play a decisive role on increased the availability of knowledge from a system and effectiveness of its associated systems. Knowledge acquired in a system, has an impact in perceiving knowledge of other related (or associated) domains, for which data models are described for mining, visualization and interpretation.

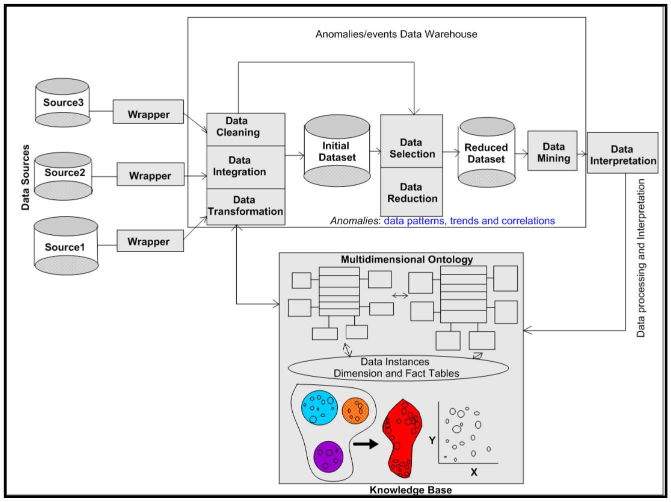


Figure 3.33: Knowledge building process model

A generalized knowledge process model as given in Figure 3.33, depicts a workflow from modelling of data sources to interpretation and then implementation of knowledge. Data exploration, prospecting, appraisal and development stages, produce enormous amount of new knowledge at multiple levels of systems' investigation and analysis, each level adding information for knowledge, interpretation and its analysis.

3.6.2 Integration of domain ontologies in a warehouse environment

Data integration is one of the key objectives of the present study. The author addresses RQ3 and RO6 in this section. The author tests several schemas in different domains, describing ontologies that make sense semantically and schematically. Ontologies designed and developed are accommodated in a framework facilitating understanding, mapping and modelling (Hevner et al. 2004, p.83, guideline 1) of complex heterogeneous data entities, creating a knowledge-structure including semantic information and rules/axioms to inform the design of data warehouse. As described by

Khatri and Ram (2004), Rocha et al. (2004) and Jasper and Uschold (2000), the author follows the following steps to design ontologies:

1. Acquire heterogeneous data (data sources: Figure 2.1 and Figure 2.2);
2. Identify and annotate entities and attributes;
3. Build relationships among entities with their common attributes;
4. Decontextualize data semantics
5. Structure and de-structure complex relationships among data entities;
6. Represent entities either in relational, hierarchical and or network forms; generate conceptual models using ontology (Khatri and Ram 2004)

Various types of operational data sources considered in the present study are shown in Figure 3.34. Data considered in the proposed research study are from the secondary data sources that contain volumes of historical exploration and production data of the petroleum industry from different operational units. Mostly data are in numerical form and at places in categorical form, on different software formats, such as Excel, Access and Websites and even in hard copies. Data are unstructured and archived in different domains.

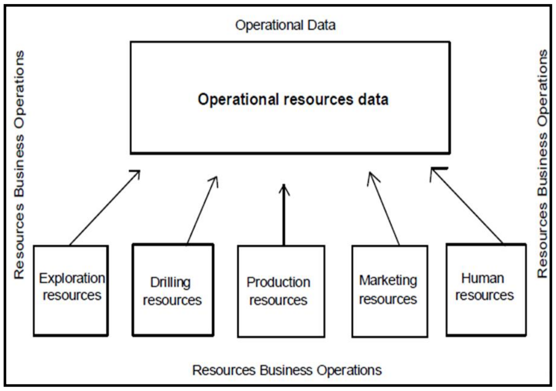

Operational Data
Operational resources data
Resources Business Operations
Resources Business Operations
Exploration resources
Drilling resources
Production resources
Marketing resources
Human resources
Resources Business Operations

Figure 3.34: Operational data sources of oil and gas industry

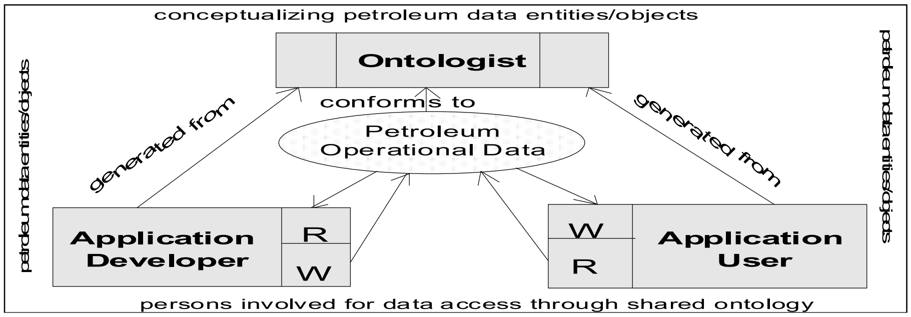

conceptualizing petroleum data entities/objects
Ontologist
conforms to
Petroleum Operational Data
generated from
generated from
petroleum data entities/objects
petroleum data entities/objects
Application Developer
R
W
W
R
Application User
persons involved for data access through shared ontology

Figure 3.35a: Conceptualization of ontology application in petroleum domain (W = Write and R= Read)

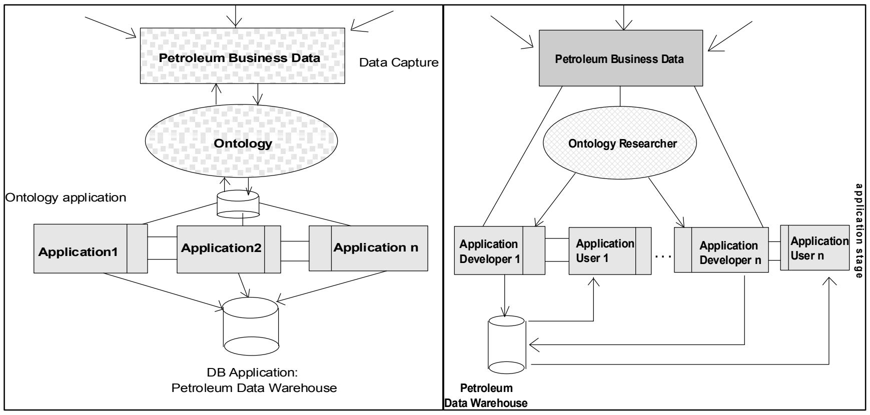


Figure 3.35b: Ontology application framework

The author demonstrates a mechanism (in Figure 3.35) that an ontologist, data warehouse developer and data miner continuously share information, thus integrating the whole process of ontology, data-warehouse and mining within an integrated framework. The proposed ontology application framework discussed in Figure 3.5, is an incorporated version for petroleum systems' research domain. As described in Figure 3.35, the author addresses an ontological view of the heterogeneous data, such as semantic information, rules/axioms. They further undergo rigorous mapping and modelling in the form of warehouse, which contains new structured data consisting of subject-oriented, integrated, time-variant, non-volatile collection of data used in support of management decision-making process and business intelligence. Ontology supports domain analysis research (Neuman 2000), which has tremendous impact on classifications of resources data in different domains.

Valued information is extracted from an application domain that holds several associated domains via a data mining process. This proposed mechanism provides an integration of heterogeneous data captured from various operational units for ontological structuring, including semantics and any business rules applied during ontology mapping.

Figure 3.36, Figure 3.37 and Figure 3.38a narrate generalized and proposed architectural methodological views. Figure 3.36 describes the process of building an integrated framework, in which ontologies from different systems made, are integrated. Data mining and visualization including interpretation on interactive workstations are attached to the warehoused repositories. Ontologically structured multidimensional data are warehoused through multidimensional data structuring (Henver et al 2004, p.83, guideline 4), a design methodology put forth in Figure 3.36. The ontological knowledge-structure is translated into logical multi-dimensional data schemas that are accommodative in a warehouse environment. As discussed in Moody and Kortink (2003), Hoffer et al. (2005) and Pujari (2002), star, snowflake and fact constellation schemas are used for constructing multidimensional logical data models. As suggested by Mattison (1996) and Marakas (2003), warehoused data are mined for exploring data patterns.

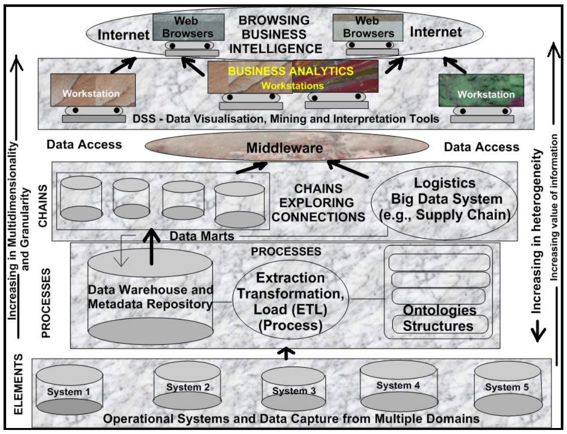

Internet
Web Browsers
BROWSING BUSINESS INTELLIGENCE
Web Browsers
Internet
Workstation
BUSINESS ANALYTICS
Workstations
Workstation
DSS - Data Visualisation, Mining and Interpretation Tools
Data Access
Middleware
Data Access
CHAINS
CHAINS EXPLORING CONNECTIONS
Logistics
Big Data System
(e.g., Supply Chain)
Data Marts
PROCESSES
PROCESSES
Data Warehouse and Metadata Repository
Extraction Transformation, Load (ETL) (Process)
Ontologies Structures
ELEMENTS
System 1
System 2
System 3
System 4
System 5
Operational Systems and Data Capture from Multiple Domains
Increasing in Multidimensionality and Granularity
Increasing in heterogeneity
Increasing value of information

Figure 3.36: Process of building warehouse framework

As shown in the Figure 3.37, the author designs workflows, in which, schemas are connected to multidimensional database management systems. The entities and or dimensions and their relationships are used in modelling various schemas, appropriate to the mapping and modelling process as detailed in Figure 3.36. Several *chains* explore connections among data sources including data marts. The author uses several tools available with data mining, visualization and interpretation. Increasing in multidimensionality is noticed from generalisation to specialization. Hierarchies are used in ontological descriptions as shown in Figure 3.37c (more hierarchies in the petroleum systems' perspective are given in Chapter 4.

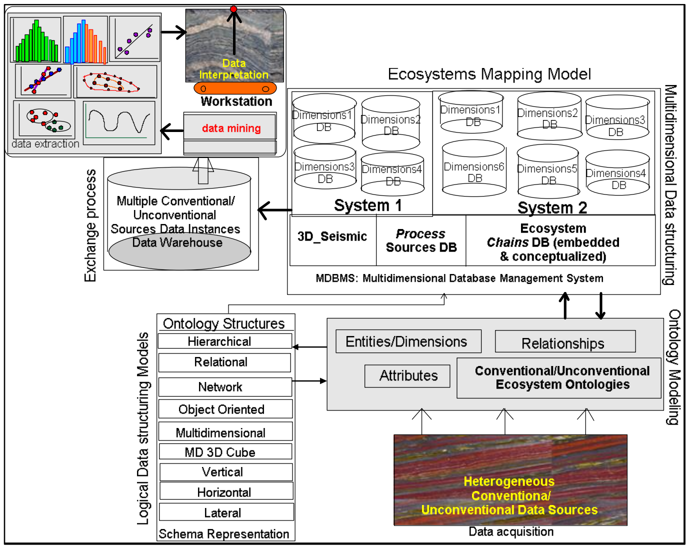


Figure 3.37a: Integrated framework, a generalized framework

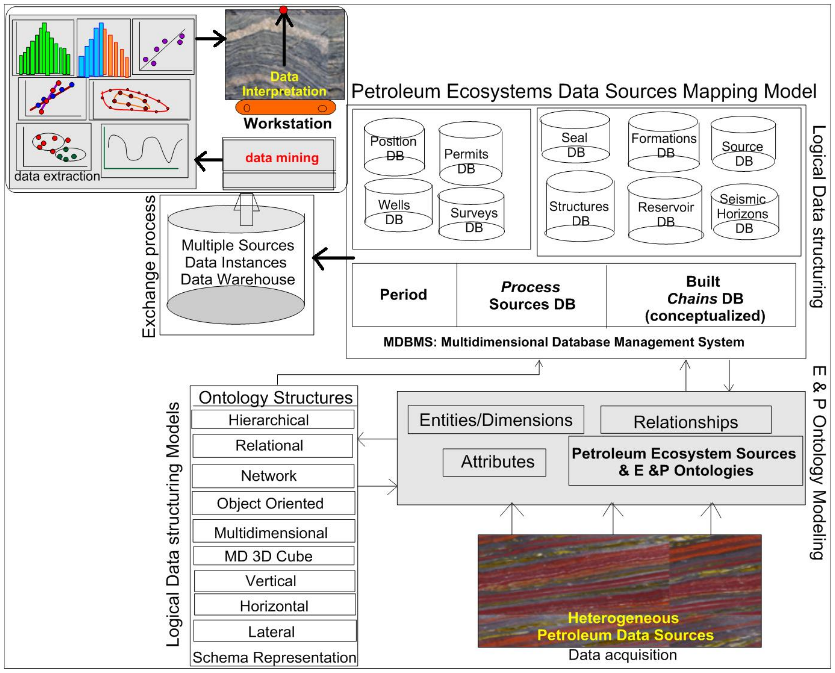


Figure 3.37b: Integrated frameworks, an integrated framework for petroleum domain (updated from Figure 3.37a)

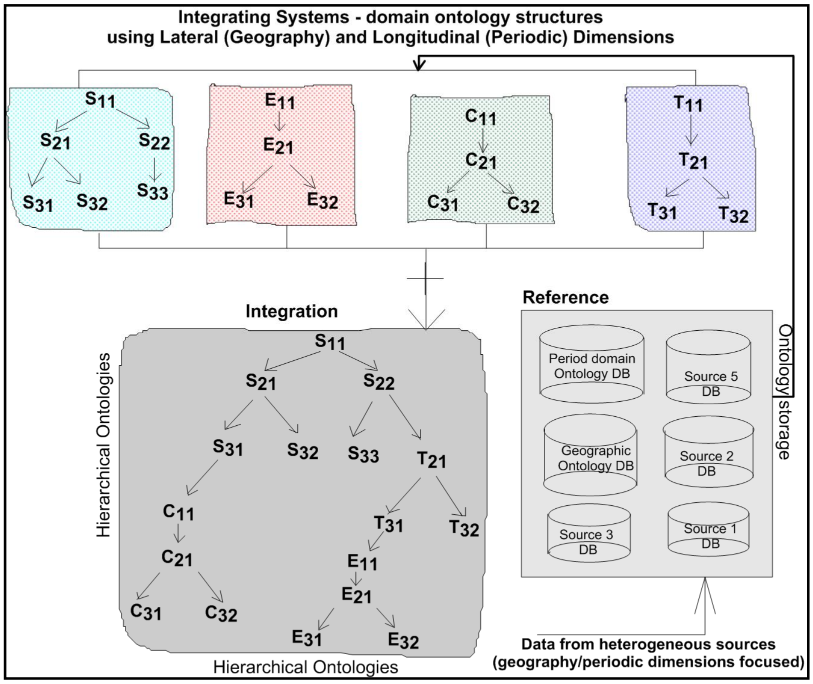

Integrating Systems - domain ontology structures
using Lateral (Geography) and Longitudinal (Periodic) Dimensions
S11
S21
S22
S31
S32
S33
E11
E21
E31
E32
C11
C21
C31
C32
T11
T21
T31
T32
Integration
Reference
Period domain Ontology DB
Source 5 DB
Geographic Ontology DB
Source 2 DB
Source 3 DB
Source 1 DB
Ontology storage
Hierarchical Ontologies
Data from heterogeneous sources
(geography/periodic dimensions focused)

Figure 3.37c Generalized hierarchical ontology

The author has an updated generalized hierarchical data structure used in the current case studies as shown in Figure 3.37c. This enables an effective means of data mining, which is elaborated in Chapters 4 and 5. The data mining procedures proposed to study various data patterns, are given in the following sections.

3.7 Simulating an Integrated Framework with Digital Ecosystems

The author explores the research objective RO7, addressing the research question RQ7. The ontology connotations are based on examining domains, data modelling, data warehousing, mining, visualization and data interpretation, all combination in a single canvas within an integrated framework. This integrated framework is simulated to a digital ecosystem, within petroleum system scenario. Ontological structuring explores the connections among multiple fields or systems that use several artifacts (of an integrated framework) including interpretation and evaluation of oil and gas data sources existing in a *sedimentary basin*. Investigation and interpretation of digital data of a petroleum system or number of petroleum systems, existing within a *sedimentary basin* that can lead to either a new opportunity or a prospect, is called a digital oil field solution. This concept risk minimizes the exploration and field development of drilling campaigns (addressing RQ8 and RO8). More details on design, development and implementation of digital ecosystems are given in Chapters 4 and 5.

3.8 Summary

The design science research guidelines are highlighted. Research methodologies are described addressing the research objectives. The domain, data modelling, data warehouse, data mining, visualization and interpretation modelling components are all integrated in an integrated framework (Figure 3.38). Under data modelling approach, ontologies are designed for heterogeneous and multidimensional oil and gas domain data sources. Then domain ontologies are integrated in a warehouse environment. Oracle driven, warehoused metadata is used for data mining, extracting data views for visualization and thus for interpretation of new domain knowledge. All the artifacts and their constructs are envisaged as per the guidelines of design science research discussed in Sections 3.1 and 3.2. Figure 3.38 is a composite image of all artifacts discussed in the Section 3.6.1, including an integrated framework discussed in Figures 3.36 and 3.37.

Elements and processes or group of elements of interconnecting petroleum system or groups of petroleum systems (as in the case of total petroleum system) constantly interact and communicate through their multidimensional attributes. A pictorial view is shown in Figure 3.38. To this extent, an exploration project discussed in Chapter 4 provides a comprehensive view of these ideas along with their results and discussions made in Chapter 5.

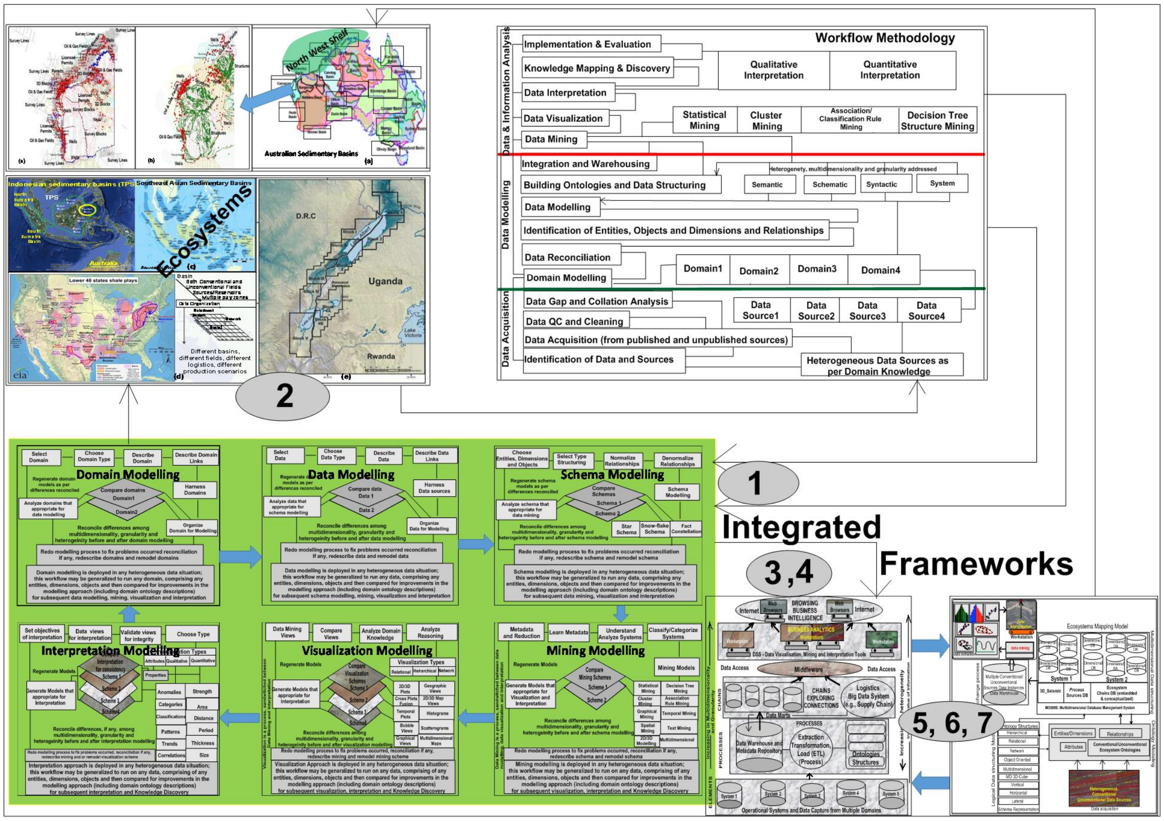

Workflow Methodology
Data & Information Analysis
Implementation & Evaluation
Knowledge Mapping & Discovery
Qualitative Interpretation
Quantitative Interpretation
Data Interpretation
Data Visualization
Statistical Mining
Cluster Mining
Association/ Classification Rule Mining
Decision Tree Structure Mining
Data Mining
Data Modelling
Integration and Warehousing
Heterogeneity, multidimensionality and granularity addressed
Building Ontologies and Data Structuring
Semantic
Schematic
Syntactic
System
Data Modelling
Identification of Entities, Objects and Dimensions and Relationships
Data Reconciliation
Domain Modelling
Domain1
Domain2
Domain3
Domain4
Data Acquisition
Data Gap and Collation Analysis
Data QC and Cleaning
Data Source1
Data Source2
Data Source3
Data Source4
Data Acquisition (from published and unpublished sources)
Identification of Data and Sources
Heterogeneous Data Sources as per Domain Knowledge
North West Shelf
Australian Sedimentary Basins
Indonesian sedimentary basins (TPS)
Southeast Asian Sedimentary Basins
Ecosystems
D.R.C
Uganda
Rwanda
Lower 48 states shale plays
Different basins, different fields, different logistics, different production scenarios
2
1
Integrated
Frameworks
3 ,4
5, 6, 7
Domain Modelling
Data Modelling
Schema Modelling
Interpretation Modelling
Visualization Modelling
Mining Modelling
Ecosystems Mapping Model
Operational Systems and Data Capture from Multiple Domains

Fig. 3.38: An integrated workflow used in the current research methodology (revised version)

An exploration project done using the integrated framework that comprises of conceptual modelling to its implementation, is demonstrated in Chapter 4.

Chapter 4: Oil and Gas Exploration Project

4.0 Introduction

In this chapter, the author discusses several ontologies in different application domains within oil and gas exploration business scenarios. The aim of this chapter is to describe an oil and gas exploration project that uses various artifacts of an integrated framework (Figure 3.38). These artifacts depict ontology models, deduced in different knowledge domains, as per guidelines of investigated design science research, discussed in Chapter 3. As per research questions and objectives (RQ1 – RQ6 and RO1-RO6) of Sections 1.3.1 and 1.3.2 in Chapter 1, the author takes an advantage of the methodological framework, integrating ontology models, articulating in different knowledge domains that can accomplish in generating a metadata for data mining, visualization and interpretation purposes.

Keeping in view the research questions RQ1 – RQ8 and research objectives RO1 – RO8, the author highlights the ontological descriptions for data sources in oil and gas exploration business in Section 4.1. Various business entities, objects and dimensions used in the ontological representation of petroleum ecosystems in Section 4.2. The author emphasizes exploration data integration and metadata integration in Section 4.3 and investigates design aspects of big-data systems in exploration business arena in Section 4.4 and systems' design and development facets in turbulent exploration business in Section 4.5. Design and development aspects of conventional and unconventional digital reservoir ecosystems in the exploration business are described in Sections 4.6 and 4.7. Petroleum digital ecosystems, part of digital oil field solutions, narrated for different sedimentary basins in Australia, Indonesia, Uganda and Arab-Gulf are given in Section 4.8.

4.1 Ontology Descriptions in Oil and Gas Exploration Domain

An upstream business system discussed here, is a simulation of one of the national oil companies in a real situation. The author describes more details in Figures 4.1a to 4.1c and also in Figure 4.2. The author has taken this as basis for building blocks (sub-systems) and workflows in the initial part of this chapter, ensuring that a producing national oil company (NOC) has well established assets comprising of Exploration, Drilling, Production, Marketing and Logistics as major entities and or dimensions. This simulation is the basis and foundation to represent an integrated systems' metadata and to this extent, the author provides a comprehensive information in various sections of Chapter 4.

The author emphasizes that ontology provides basic knowledge on semantic information while designing various data models and thus an integrated metadata in oil and gas exploration business domain. In the petroleum exploration and production, oilplay is key factor and criteria for ascertaining new knowledge domains. The author uses conceptual modelling to design hundreds of fact and dimension tables adopting relational and hierarchical structures. A commitment to ontology helps ensuring the consistency among data relationships and properties of *geological*, *geophysical* and *geo-chemical* entities (Figure 3.2, in Chapter 3). The key entities are considered for understanding the relationships among *surveys*, *wells*, *permits* and *production* operational activities (Figures 3.37 and 3.38 in Chapter 3). Based on these operational activities, the author describes the hierarchical structuring. *Exploration* is a key, subject-oriented and a generalized super-type entity, from which, other specialized dimensions such as *onshore*, *offshore* and *geological*, *geophysical* and *geo-chemical* dimensions are derived for building associative conceptualized data relationships. The present study embodies the scenarios for applying ontologies and their integration for achieving the objective of an articulated design. In addition, as part of digital oil field solutions development, the author designs petroleum digital ecosystems in different contexts of sedimentary basins of Australia, Indonesia, Uganda and Arabian Gulf perspectives, after analysing volumes of heterogeneous and multidimensional data sources, a requirement for generating "petro_2" database in this project.

The author uses volumes of heterogeneous data from oil and gas industries (Figures 2.1 and 2.2 and 4.1a, 4.1b and 4.1c), from which various entities, objects and dimensions are identified and organized them in different conceptual, logical and physical models. Types of structuring are chosen depending on the need of information that support the knowledge base decision support systems. Crucial decisions are taken for budgeting on risky exploration and drilling operations, based on the supply of accurate, timely structured data and information. The requirement of an integrated data supporting system is considered, which allows smooth flow of processed information among different operational units. An oil and gas company is described as an integrated system, consisting of several operational units, identified as sub-systems and again each sub-system into many smaller units as shown in Figure 4.1a.

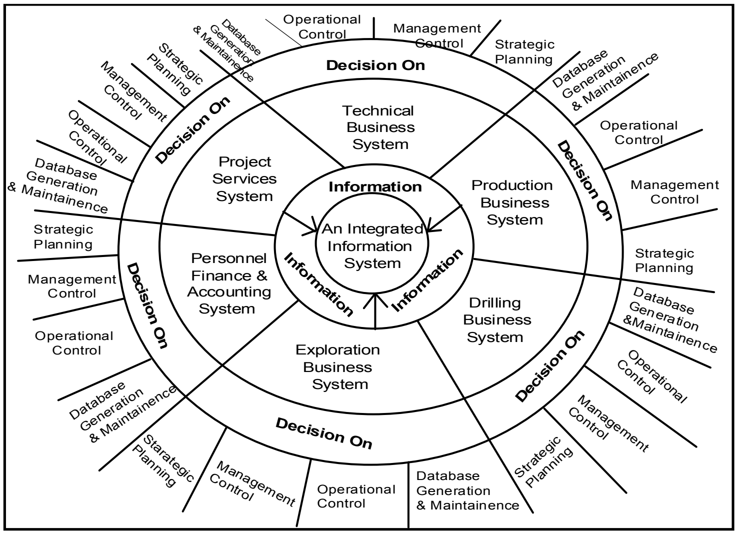


Figure 4.1a: Information – decision-making roles in an oil & gas company metadata structure

Longley et al. (2001), Gilbert et al. (2004) and Weimer and Davis (1995) provide literature on the exploration and production data entities and volume of attributes used in oil and gas industries. *Exploration*, *drilling*, *production operations*, *marketing*, *technical*, *finance* and *accounting*, *personnel* and *administration*, *project services* and *research* and *development* are key areas of an integrated oil company (business system) segregated into subsystems and each of these subsystems is divided into smaller units for the purpose of mapping into manageable units. The manager of each unit handles several supporting staff consisting of geologists, geophysicists, reservoir engineers, financial controllers, personnel administrators and even economists. Each sub-system carrying the unique operational and functional data, is conceptualized initially within an entity-relationship model. All these sub-systems (referred as business entities) that linked with their associativity, sub-type entities are integrated. Various activities of the business system are identified with all sub-type entities. Attributes in each of these entities are responsible for identifying associated entities.

The organizational data represented in matrix form (Figure 4.1b) designate that all the functions and operational activities of the oil and gas company that are systematically organized and needed for mapping and modelling purposes. Various conceptual models are prepared using all the business entities. The relationships are normalized wherever necessary, so that the entities involved in the mapping process are conforming to the relationships of other associated entities. As an example, Figure 4.1b narrates the process of identifying and analyzing company's business entities of both specialized and generalized types. As shown in Figures 4.1a – 4.1c, different business entities are involved in building an integrated oil and gas company, from which hierarchical, relational and networking nature of structural relationships are construed in the modelling process.

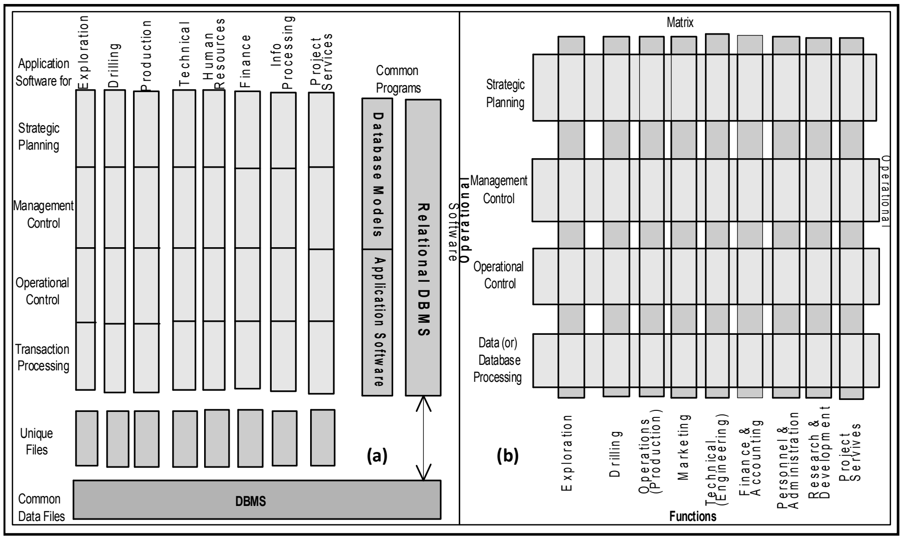


Figure 4.1b: Functions and activities of the oil and gas company, identifying business entities

As examined in Figures 4.1a - 4.1c, various business entities are building blocks of an integrated business system scenario, showing how complex view of business entities of Oil and Gas Company are simply segregated into various generalized and specialized entities.

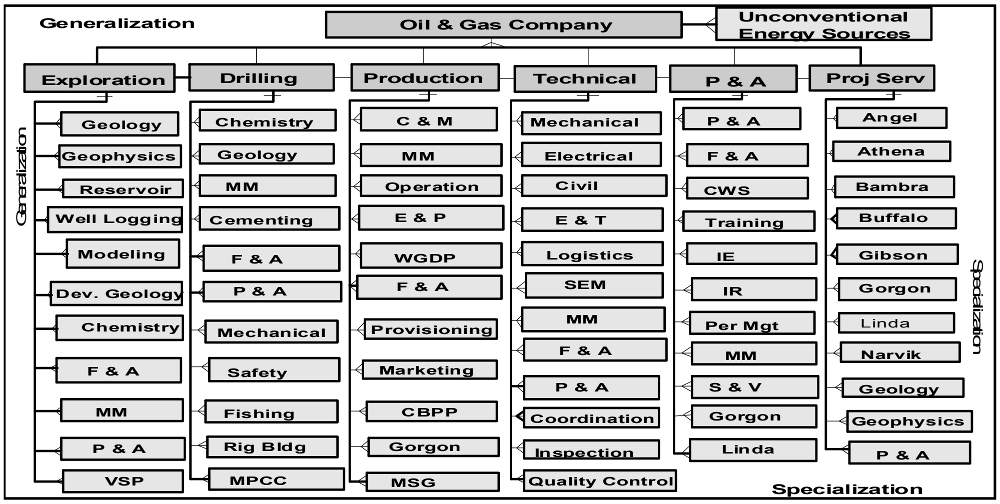


Figure 4.1c: Metadata model presentation showing super type and sub type business entities

Methodologies as discussed in Chapter 3, are used for designing and developing data models in oil and gas industry domain. As per domain knowledge and its application, the author considers entities, objects and dimensions in ontology modelling, but the real focus is on multidimensional modelling and mapping of dimensions, keeping in view multidimensionality in oil and gas industry domain. Detailed workflows and methodologies of data modelling, data warehousing and data mining, data visualization and interpretation are described in the following application domains.

The author identifies an upstream oil and gas industry, a primary focus on the businesses of exploration, drilling/mining and production entities, as shown in Figure 4.2. Upstream oil and gas data and information flow within their information systems complicate the organizational databases, especially of if they are spatial-temporal nature. The author draws a simple conceptual model as shown in in Figure 4.2a, based on which different super-type and sub-type systems are drawn. The information flow within these systems is shown in Figure 4.2a. Models are conceptualized in such a way that they are logically representative for implementation and evaluation at later stages.

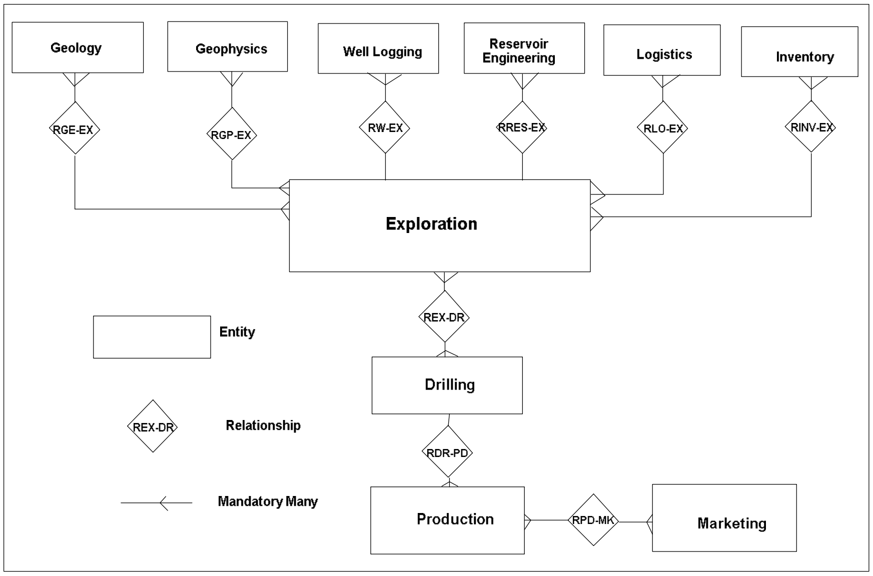


Figure 4.2a: A typical ER model for an oil and gas exploration company

There are various activities and functions undertaken in every project of Oil & Gas Company. The activities are various measurable tasks performed in each and every exploration, drilling, production, marketing, logistics functions. The author acquires all the available data associated with each activity and function to store and use for modelling within an integrated framework, which is termed as a sub-system. This is illustrated schematically in Figure 4.2b.

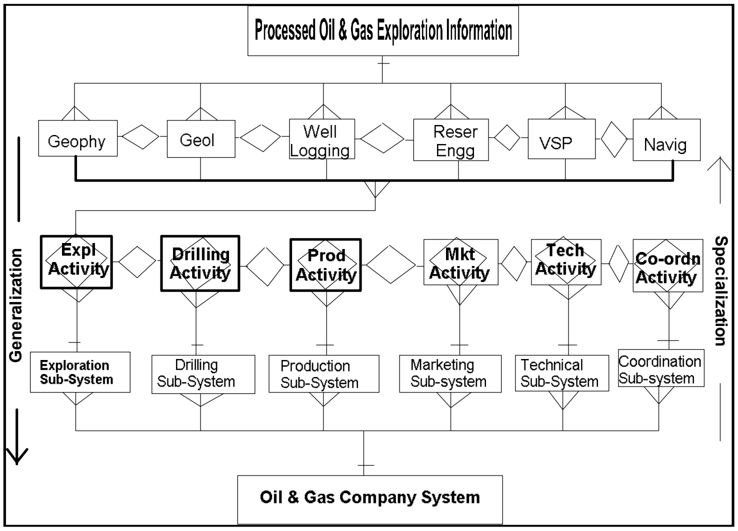


Figure 4.2b: A typical petroleum exploration company design, narrating various systems

Before developing the actual data schemas, the author has identified the sources of data and requirements necessary to analyse and organize them in different modelling approaches. As an example, information flow among subsystems, is shown in Figure 4.3. Keeping in view the research objectives and guidelines of design science, historical oil and gas data sources are analysed, such as attributes, facts and measures for building constructs and models. Subsequently, the author prepares conceptual models that are feasible for construction of logical star schemas, for physically connecting and integrating them with other oil and gas (exploration and production) data attributes.

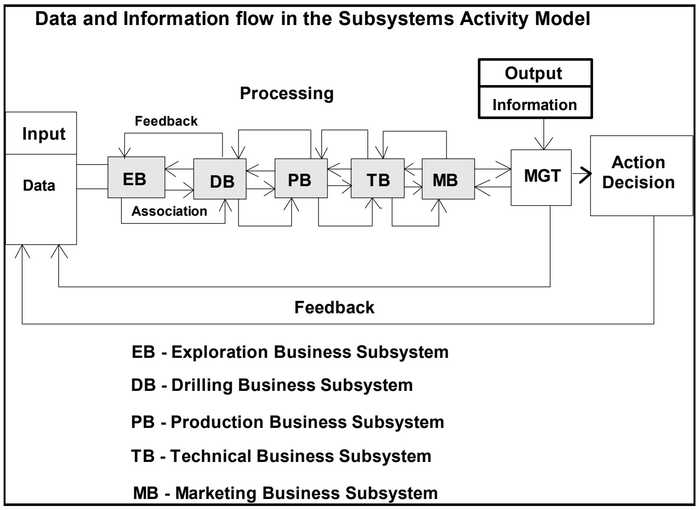

Data and Information flow in the Subsystems Activity Model
Processing
Output
Information
Input
Data
Feedback
EB
DB
PB
TB
MB
MGT
Action Decision
Association
Feedback
EB - Exploration Business Subsystem
DB - Drilling Business Subsystem
PB - Production Business Subsystem
TB - Technical Business Subsystem
MB - Marketing Business Subsystem

Figure 4.3: Data and information flow among sub-systems

The data acquisition is a part of an ongoing design and development of an integrated framework. Data are organized in different logical data structures representing logical data organization. In case of multidimensional modelling approach too, hierarchies play crucial role in describing dimensions and their relationships. The author describes few such multidimensional relationships, characterizing oil and gas data attributes in the following sections.

Explicit hierarchies in dimensions: The hierarchies of the dimensions that are captured by the schema should be explicit, so the user has the clear understanding of relations between the different levels in the hierarchy. In the present study, the hierarchies Basin_Surveys_Monthly or ExplorationCost_State_Mineral_Quarterly are captured.

Symmetry of dimensions and measures: In an industry situation, oil & gas data attributes are numerous and voluminous. Geology, geophysics and geochemistry have variety of dimensions and attributes which are spatial-temporal in general. These types of dimensions and their attribute instances at places pose symmetrical properties by virtue of spatial-temporal data. The data model should allow measures to be treated as dimensions and vice versa. The attribute quarterly or annual exploration costs or monthly surveys or wells drilled are some examples. The exploration costs would typically be treated as a measure, to allow for computations such as average, etc. The period dimension that is defined, allows to group the number of surveys conducted or wells drilled in quarterly or monthly or yearly period.

Multiple hierarchies in each dimension: In each dimension, there can be more than one path along which to aggregate data. As an example, let us assume that a time dimension on the date of survey conducted or well drilled. Days roll up to weeks and to months, but weeks do not roll up to months. To model this, multiple hierarchies in each dimension are needed.

Support for correct aggregation: The data model should support, getting results that are “correct,” i.e., meaningful to the user, when aggregating data. One aspect of this is, is to avoid double counting of data. Petroleum production of a well in a basin implies that it could be oil or gas or condensate or combination of all three oil and gas. In this case study (of petro-2 database), when asking for the number of petroleum producing wells in different basin or states groups, the same oil or gas discovery once per group should be counted, even though the oil or gas discoveries are made in several basins or states in a group.

Many-to-many relationships between facts and dimensions: The relationship between fact and dimension is not always the classical many-to-one mapping. In our case study (petro-2), the same contractor may carry out several surveys, even at the same point of time, widening their contracts with more *surveys* in adjacent or same *basins*, with increasing resources.

***Handling change and time*:** Data change over time, but meaningful analysis results should be dug across changes. Exploration costs, production flow rate, number of wells or surveys conducted over a period of time are few examples of the present problem situations. In the example, one survey can supersede with another survey in the same basin or with a bit of overlap. It may so happen that new surveys are conducted or new wells are drilled in the areas, where old survey data are already accumulated. It should be possible to easily combine data across changes. The problem typically referred to as handling *slowly changing dimensions* is part of this problem.

***Handling uncertainty in the data*:** The author uses different vintages in the data modelling process. The vintages with missing or uncertainty data too have different versions. In this case study (building a warehouse for oil and gas data sources), old *surveys* or *permit* (key terms defined in a glossary in Appendix-1) facts of historical data supersede with new ones. Ninety percent of the data consist of old surveys or permits history; new data histories are added to the old one. Thus, when requesting data grouped by surveys or permits history for a period that spans the change, the old survey or permit history needs to be counted together with the first new surveys or permits data history. The data model should allow expressing some kind of indications in the query results, signifying how many of these “converted” surveys or permits history are counted in the result with both new and old surveyed data.

***Handling different levels of granularity*:** Fact data might be registered at different granularities. In this example, *survey* data may be registered in the *survey* facts, *well* data in *well* facts and *permit* data in the *permit* facts. The registration could be different from different surveyors or contractors. Some users use a very specific survey such as “well based surveys or surveys based wells,” while others use the more imprecise “survey or well data,” which covers several lower level survey data. It should still be possible to get correct analysis results when data are registered at different granularities.

Design and development of petroleum ontologies

The following steps are considered in the oil and gas domain:

1. Acquire data attributed to the development of petroleum system;
2. Identify entities, attributes of exploration and production data;
3. Build relationships among entities with their common attributes;
4. Structure and de-structure complex relationships among data entities;
5. Acquire *surveys*, *wells, permits* and *production* data (from Figures 2.1 and 2.2);
6. Represent all the entities in to relational, hierarchical and networked data structures;
7. Collect *reservoir*, *structure*, *seal*, *source-maturity*, *migration* and *timing of occurrence* data instances and generate conceptual models using ontology (Figures 3.36 3.37 and Figure 3.38 including Figure 3.2 ;
8. Integrate exploration and production costs data and developing relationships with petroleum systems data.

Nimmagadda et al. (2005c) demonstrate an ontology application in petroleum industry that evaluates the inventory of the petroleum reserves in the Western Australian producing basins. Hundreds of dimensions and their attributed elements of associated petroleum systems are organized and documented in strict ontological and semantic sense in a warehousing approach, improvising the basin concept, petroleum system knowledge and thus managing petroleum reserve inventories. Similar to multidimensional schemas, object-oriented data schemas are constructed using petroleum ontology. In general, in multidimensional schemas, the relationships among the common data attributes are de-normalized, so that the final data views become finer for effective data mining. Here, data relevant to activities such as *exploration*, *drilling* and *production* are organized in their corresponding data structures for the purpose of integrating them to build a metadata model as shown in Figures 3.36 -3.38 (in Chapter 3). At conceptual modelling stage, the author conceptualizes oilplay, performance indicators (which are conceptualized dimensions) and various other dimensions of petroleum systems within a basin configuration. The specifications of concepts are used for expressing the knowledge, including types and classes, the kinds of attributes and properties they can have, the relationships and functions they can perform and constraints that they can hold.

As shown in Figure 4.4, the author describes a basin through hierarchical structuring, connecting through data location attributes. Each basin consists of several *surveys*, *wells*, *permits* and other geographical data such as *river-channels*, *geomorphology* in land and coastal areas, lineation features, hard and soft rock areas, which are all used during integration of petroleum systems. In the context of petroleum provinces, instances associated with *surveys*, *wells* and *permits* as shown the Figure 3.8, are needed. As mentioned earlier, each petroleum province (region) consists of a sedimentary basin, *survey-lines* falling in *dip* and *strike* (attributes) directions of a region. Each survey consists of hundreds of *survey points* such as *CDP (common depth point)* or *CMP (common mid-point)*. These points are in 2D or 3D dimensions, as depicted in Figure 4.4.

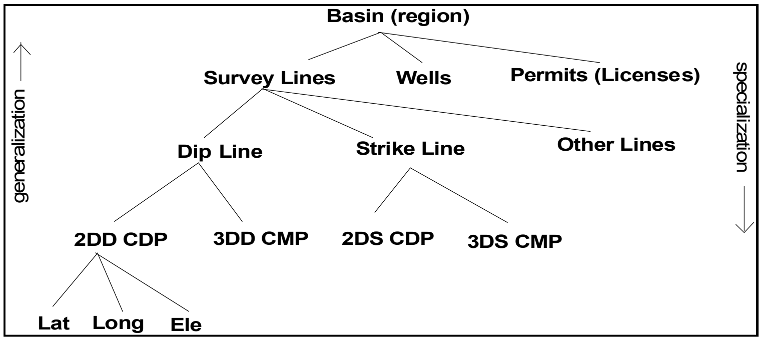


Figure 4.4: Demonstrating the *basin* location hierarchy in *CDP* domain

Other Lines associated with seismic profiles have connectivity and all the dimensions are made to make connectivity among seismic profiles. *Point, line, regions* are connected through geographic coordinate system. A geographic coordinate system is typically connected to a petroleum system and basin under investigation. The author incorporates *strike* and *dip* geological and geographical attributes associated with the seismic CDP/CMP dimensions within an exploration super type dimension. An ontology carries sufficient and useful information of relationships in the structuring. The ontology definition provider is responsible for defining a suitable ontology representation, which may save time and cost in capturing the data and information content from other multiple entities and dimensions in other domains. The incoming ontologies are integrated in an application framework (Figures 3.36, 3.37 and 3.38) with other ontologies. Firstly, it checks the petroleum ontology database for similarities to establish the relationships that can map between two ontologies. If some dimensions or facts of the relationships contradict each other, the part in contradiction are rejected or placed into other future ontology representations in the integration processes. Hierarchical ontologies as described in Figure 3.37c, shows similar S22 and T11 elements that are common in their properties or characteristics, so T11 is merged with S22 in the resultant integrated hierarchical ontology. During the ontology mapping process, if a parent ontology finds a common child ontology, then they are merged and integrated; accordingly the petroleum database ontology is updated after each and every integration and merging process. For example, for the purpose of data integration, ontology structures of *navigational*, *surveys* and *well-data* dimensions need to be merged into one single metadata structure. Other issues of ontology structuring involved in connecting different specialized entities of super-type entities are:

Ontology as specification: The author creates ontologies in domains such as exploration, drilling or production, and uses as a basis for specification and development of data warehouse schemas or UML models for surveys (within the generalized entity of exploration) entity. Benefits of this approach include documentation, maintenance, reliability and knowledge (re) use.

Common access to information: An information is required by one or more persons (geologists, geophysicists and reservoir engineers in our application) or computing applications (seismic workstation utilities such as ZMAP, LASLOG, GIS or remote sensing), but is expressed using unfamiliar vocabulary, or in an inaccessible format. Ontology facilitates rendering of information intelligible by providing a shared understanding of the terms, or by mapping between sets of terms. Benefits of this approach include inter-operability, and more effective use and reuse of knowledge of oil and gas in other applications and domains.

Ontology-based search: The author uses ontology for searching an information repository for desired oil and gas data source. The chief benefit of this approach is faster access to important information of oil and gas domain, including more effective use and reuse of knowledge.

Ontology application framework

Jasper and Uschold (1999) discuss a framework for understanding and classifying ontology applications in software engineering. Figures 3.36 – 3.38 show methodological views of the application, narrating ontology at application design and development levels. The author represents petroleum data in either relational or hierarchical structures. Constructing the ontology structure is a bonus for relational database application, since all the relationships and properties among entities described in ontology come to play a similar role as in relational data structure design.

The ontology framework (as illustrated in Figures 3.37 and 3.38) provides a scope of integration of heterogeneous data, captured from various operational units (of oil and gas industry) for ontological structuring, including semantics and business rules applied during ontological mapping. The actual mechanism involving “the ontologist, application developer and user” is described. An ontologist develops several semantic base database structures that are in agreement with petroleum data application developers’ needs. As demonstrated in Figures 2.1, 2.2 and 3. 38 (in Chapters 2 and 3), operational petroleum data have to be understood in terms of the true meaning of entities and their relationships participating either in the relational and or hierarchical database mapping and modelling process. An ontologist and application developer conform in reading/writing the data. Ontologically processed petroleum data are fed to the applications involved in data warehousing and data mining of oil and gas databases. For the sake of simplicity, key ontological structures are described in a standard syntax.

Knowledge-base structural model: The author attempts to build knowledge-base structural models based on unions and joins of set theory (Bartle 1976, Halmos 1974 and West 2006) Here the structure or hierarchy shown in an example of oil & gas company metadata is explicitly described here. A model structure Z for L (logic) is a 5-tuple <P, E, D, P, M>. Here S = U {(P, EI) R1} and U {(E, D) R2} are domains of Z structure, and consists of the union of two mutually disjoint sets (P, E), R1 and (E, D), R2. (P, E) is a set of individual entities of S and R is a set of relationships between (P, E) and (E, D) entities. R is partitioned in different ways as designer wanted it as R1 and R2, since the prior entity combinations are logically related. Here P = position; E = exploration; D = drilling; P=production; and M=marketing entities.

Another structural model basin_line_point representing a region (R) composite (at generalization level) of all basins is about locating a survey point (specialization level) in a basin:

$R = (B_1, B_2, B_3 \ldots B_i)$; i = basin numbers; B = basins
$B_m = U \{(L_{1D}, L_{2D} \ldots L_{mD}) + (L_{1S}, L_{2S} \ldots L_{mS})\}$;
$L_{m.n} = U \{(P_{m,\ (n,n+1,n+2)}\ P_{m,\ (n+3,n+4,n+5)}\ P_{m,\ (n+6,n+7,n+8)}\ . . .\ )\}$;
U= union; L = location; P = point;
m = number of survey lines;
n = number of survey points on each survey line

Similar semantic models are made for other production (P) and marketing (M) entities and their relationships.

4.2 Ontology Representations in Petroleum Ecosystems

The author draws ER diagrams based on real data dimensions, attributes and instances, taken from a role-model company situations as given in sections 4.2.1, 4.2.2 and 4.2.3. The author uses data instances from published literature in the modelling process as described in various sections of this chapter and Chapter 5. The beauty of the current research work is working with the constructs and models, based on factual data. Dimension and facts are derived from the real oil & gas data situations taken from sedimentary basins from Australia, Indonesia, Uganda and Arabian Gulf countries. To this extent, the author added notes in Chapters 1 and 2, describing several data schema representations in the literature, such as simple star and snow-flake schemas (Hoffer et al. 2005 and Nimmagadda et al. 2010). Several such data models representing both logical and physical data organizations in petroleum industry situations are given in Shastri and Dreher (2011). For more complex and typical applications such as in petroleum ecosystems, star schema and or snowflake schema, in combination are used to construct a fact constellation schema. This schema is more complex than star or snowflake architectures, because of the fact, it contains multiple fact tables, but necessitated for petroleum systems' representation. This allows dimension tables to be shared amongst many fact tables, as in the case of petroleum sub-systems design and analysis. In addition, the domain ontologies integrated in warehouse environment, are extended their simulations in ecosystems scenarios as described in the following sections A, B, C, D and E:

A. Petroleum Ecosystems

An ecosystem is a system whose members benefit from each other's participation via symbiotic relationships (Thomas et al. 2006) and (Karp 1995) (positive sum relationships). In the context of sedimentary basin research, modelling and petroleum systems analysis, narration of several petroleum systems and different oil and gas fields, in each petroleum system, is a complex community, but its environment functioning as a single ecological unit. More realistically, it is term of millions of data attributes and properties from volumes of DBs of multiple basins and their respective petroleum systems all that store in one place.

B. Ontology as a specification mechanism describing a sedimentary basin with one or more petroleum systems

A sedimentary basin of formally represented knowledge, is based on a *conceptualization* (Thomas et al. 2006) of petroleum systems' elements: the *structure, reservoir, source* and other *source maturity* and *migration-path* process dimensions that are assumed to exist in some area of interest and the relationships that hold among them. A conceptualization is an abstract, simplified view of the basin that wish to be represented for some purpose. Every knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. In the case of an investigating *sedimentary basin*, its associated petroleum systems, interpreted oilplays and producing prospects have complex known and undiscovered relationships and many undiscovered are known to have hidden relationships (may be conceptualized). Multiple dimensions or entities are interpreted among these petroleum systems and each dimension has multiple data attributes and or characteristics.

The domain knowledge is key in narrating the ontology and when domain knowledge is represented in a declarative formalism, the set of dimensions that can be represented is called, in a broad sense *digital basin* or with sets of digital petroleum systems. Ontology of a petroleum system is developed and represented based on systematic existence, logic and intelligent design of data structures. Relationships are built among *structure, reservoir, source rock, seal rock elements* and *processes* such as timing of formation of elements and migration paths with several logical conceptualizations. Prior knowledge of petroleum system domain, among elements and processes, classes, relationships, functions and dimensions (super-type and sub-type) associated with sedimentary basin may be a necessity, if not essential.

The connectivity and intelligent communication among these dimensions are well described during course of ontological commitments. A dimension commits to ontology if its observable actions are consistent with the definitions and designs of ontology. The idea of ontological commitment is based on the knowledge-level perspective. The knowledge level is a level of description of the knowledge of a dimension that is independent of symbol-level representation used internally by the dimension. Knowledge is attributed to dimensions by observing their actions, a dimension 'knows" something if it acts as if it had information and is acting rationally to achieve its goals. The actions of dimensions – including knowledge base servers and knowledge based systems – can be seen through a "tell and ask" functional interface (Thomas et al.

2006, Davidson et al. 1995 and Karp 1995), where a client interacts with a dimension by making logical assertions (tell), and posing queries (ask).

Logically, a common ontology defines the vocabulary within which queries and assertions are exchanged among dimensions. Ontological commitments are also agreements to use the shared vocabulary (logic) in a coherent and consistent manner. The dimensions sharing a vocabulary need not share a knowledge base; each knows things the other does not, and a dimension that commits to ontology is not required to answer all queries that can be formulated in a shared vocabulary (logic). A commitment to a common ontology is an agreement to consistency, but not completeness, with respect to queries and assertions using the vocabulary (logic) defined in the ontology.

The warehouse modelling is a logical data integration process through multi-dimensional structuring, from different sources of petroleum ecosystems, such as environmental ecology and geomorphology. Environment and geomorphic systems (Pujari 2002, Shanks et al. 2003 and Sheriff 2002) cannot be separated and their coexistence is interpreted to be ecologically constructive. This implies that entities involved in these systems are so eco-friendly embedded in nature, but apparently unknown because of their poor understanding. An entity of a system becomes affected for any reason, can cause a ripple effect throughout the other systems. For example, geomorphology and its structural pattern are constantly changing on this Earth with a definite impact on its surrounding environment and ecology.

The data from these pertinent systems are inherent in nature. Having known a particular system digitally, the knowledge of its associated system can be predicted. For this purpose, these systems, in terms of their digital datasets, are composed intelligently in a data warehouse (Jukic and Lang 2004, Nimmagadda and Dreher 2005a, Nimmagadda and Dreher 2005b and Nimmagadda and Dreher 2007) environment through an ontology based multidimensional modelling approach. Ontology depicts the semantics and the association among these systems, in which entities are represented in multiple dimensions. Hierarchical and relational ontologies (Jasper and Uschold 1999, Lowrie 1997, Nimmagadda and Dreher 2006 and Nimmagadda and Dreher 2007) have been described, in which dimensions and their attributes vary in several directions. Ontology builds relationships from fine- to coarse-grained dimensions (Rudra and Nimmagadda 2005) through bottom-up or up-down hierarchical dimensions. Data attributes of several related entities can also be relationally linked through one-to-many or many-to-many or even one-to-one relationships. Seismic data instances of these attributes are logically organized in vertical, horizontal or lateral directions in relational and hierarchical structuring methodologies. Besides, data instances of dimensions, such as point-to-point, line-to-line, zig-zag and data- looping, can also be structured to integrate different domain ontologies, such as seismic (surface) and well (sub-surface) domains.

C. Exploration domain ontology development for ecosystems management

The author further examines several data (geological) events during analysis of conceptual models. For example, sinkholes and the geological formations with which sinkholes investigated are based on integrated interpretation of geological and geophysical data and their domain knowledge outcomes. In principle, *seismic* data distort their amplitudes and frequencies with possible loss of *seismic* energy around the sinkholes where severe absorption effects take place. Seismic data in surface domain are integrated with drilled-well data, from sub-surface domain and then interpret the effects of sinkholes and their occurrences at geological ages. An ontology based approach narrated to conceptualize the entities involved in seismic and drilled-well domains, organizes multidimensional data in a warehouse environment. The author computes geomorphic structural patterns for different geological *ages* (stratigraphic events), using ontologically derived metadata structures for interpreting the connections among elements and processes of multiple petroleum ecosystems. The following mapping procedures are considered for connecting multiple petroleum systems:

D. Ontology-based conceptual modelling

Volumes of dimensions are conceptualized to generate ontology-based multidimensional data models that led building relationships among instances of dimension and fact data tables of seismic and drilled-well entities. Where ever one-to-one, one-to-many and many-to-many relationships exist, they are mapped more logically. Business rules, constraining the data mapping process, are imposed while building relationships among data structures.

E. Functional mapping in seismic domain ontology

In the multidimensional ontology modelling (Hoffer et al. 2005 and Nimmagadda and Dreher, 2006) approach, the author deals with multidimensional data involving several entities or dimensions and attributes, to build ontology relationships. In discrete theory of datasets and relations (Nimmagadda and Dreher 2010a), a relationship between two domain ontologies defines a correspondence, or connection, or mapping among elements of datasets. These datasets can consist of numerals and symbols as in mathematics or entities in databases. A relation (Nimmagadda and Dreher 2010a) is an ordered pair of elements taken, from the related domain ontologies. In other words, one element from one dataset of particular domain ontology can be combined with one from the other dataset of a related domain. In our seismic data instances case, the relationships are often of functional type, in which the mapping of structure of a relationship (between seismic and well-base domains) determines the type of a relationship (as shown in Figure 4.5 and Figure 4.6) and also obeys the mathematical definition of a function.

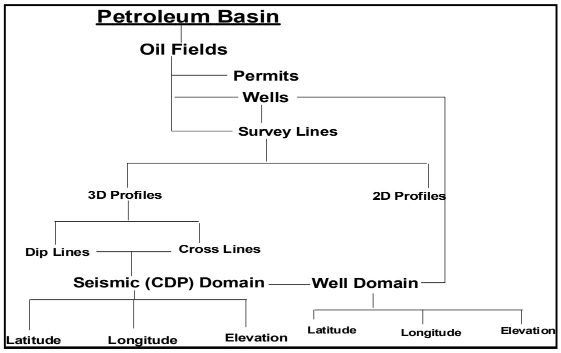


Figure 4.5: Hierarchy of seismic *Common Depth Point (CDP)* domain ontology, showing different attribute varying data dimensions

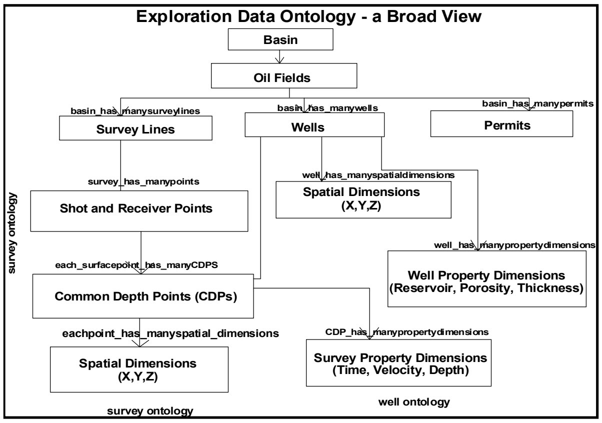

Exploration Data Ontology - a Broad View
Basin
Oil Fields
basin_has_manysurveylines
basin_has_manywells
basin_has_manypermits
Survey Lines
Wells
Permits
survey_has_manypoints
well_has_manyspatialdimensions
Spatial Dimensions
(X,Y,Z)
Shot and Receiver Points
survey ontology
well_has_manypropertydimensions
each_surfacepoint_has_manyCDPS
Well Property Dimensions
(Reservoir, Porosity, Thickness)
Common Depth Points (CDPs)
eachpoint_has_manyspatial_dimensions
CDP_has_manypropertydimensions
Spatial Dimensions
(X,Y,Z)
Survey Property Dimensions
(Time, Velocity, Depth)
well ontology
survey ontology

Figure 4.6: Relational representation of seismic (*CDP*) domain ontology, showing different attribute varying data dimensions

A special type of mapping (Nimmagadda and Dreher 2010) that is of great importance in ontology is a mapping that obeys the mathematical definition of a function. In a mathematical function, a domain D, a range, R, and a functional mapping or function f such that f: D→ R, which means that for all values in each domain ontology, D map to the values of other related ontology R, such that for each value in D, there corresponds a unique value in R. If every value in D maps to R, then it is totally “functional”; otherwise the mapping is called partially functional. In our case, seismic data instances, all the domains, specified, such as *time*, *velocity* and *depth*, every value in each domain, maps every value in the other domains. So these are totally functional mappings.

Hierarchical databases have their idiosyncrasies (Ozkarahan 1990). Some are useful for applications, such as functional mapping between parent and children in a hierarchical structure. Hierarchies pose restrictions too. To be more specific, in a hierarchy or tree data structure, there is a many–to-one mapping (which includes one-to-one) from child nodes (data record instances) to the root instance. This must be a totally functional mapping. The hierarchy must obey the tree structure in the sense that no node can exist without a parent except the root. For example, “petroleum basin” and “exploration” entities have been their related sub-type entities or dimensions (as discussed in the forthcoming sections). Further, either one or more attributes must have been related among these dimensions in a deduced hierarchy.

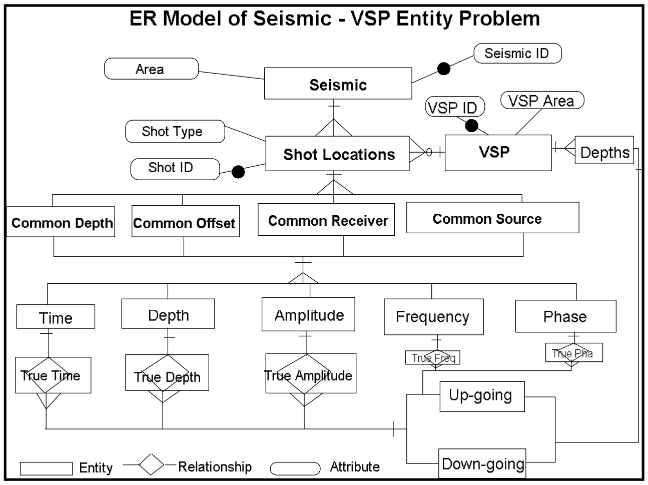


Figure 4.7: An ER model showing *Common Depth Point (CDP)* dimensions and their relationships with *common receiver, common offset* and *common source* dimensions

The hierarchy may have several applications too, where for example, children need to be unassigned to any parent; handle the deletion of a parent differently; and build relationships of a more complex nature such as many-to-many relationships or the multidimensional relationships ending up in a lattice or network structure, as the case of CDP dimensions are organized in Figure 4.8 rather than a tree structure.

The exploration entity or dimension has a complex hierarchical data structure, represented in a multidimensional conceptual data model. In the case of seismic and well-data structuring, for example, 3D seismic data are distributed in multiple dimensions, each dimension characterizing a property. In the case of hierarchical structure, such dimensions are either mapped vertically or horizontally or laterally and in combinations. These ontologically derived conceptual entities are further translated into different multidimensional data properties for intelligently storing in a warehouse environment, and discussed in the following sections.

Data structuring – hierarchical ontology

In the hierarchical data structuring (Nimmagadda and Rudra 2004 and Nimmagadda et al. 2006d) process, the author identifies the hierarchies using geomorphic structural entities. A hierarchy is a tree structure, in which a geomorphic surface, as represented by *seismic* dimension has several nodes and each node is connected by *lines, points and contours* dimensions and all these nodes are interconnected. Each geomorphic surface consists of number of *contours*, each contour has equal data instance values (such as time, depth and velocity values) that are distributed on several survey grids and on each grid with several survey lines and on each line with several points. Each point has latitude, longitude and elevation as spatial dimensions along with property instance. These are typical hierarchical data structural features. In the case of hierarchical data organization, conceptual ontologies are built among several geomorphic surfaces surrounded by petroleum fields. A profile layout describing generalized hierarchical and relational seismic domain ontology is shown in Figure 4.8. CDP data instances from reflecting surfaces, how they are acquired from field geometries are provided in the Figure 4.8.

Vertically varying dimensions

Vertically varying dimensions possess vertical dimensionality in which *time*, *velocity* and *depth* are key dimensions, varying their instances hierarchically in the vertical direction. When their instance values vary in this direction, the data structure changes into different relationship or other conceptual domains, such as slopes (steepness dimension) of horizons.

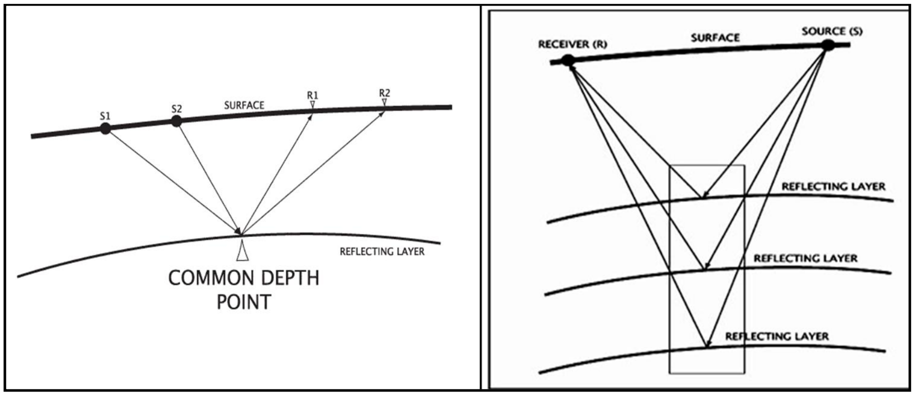


Figure 4.8: Modelling data instances from *Common Depth Point (CDP)* domain ontologies

The *dip* and *strike* attributes (their respective data instances) of horizon dimensions thus come into play in these ontology descriptions. Entities participating in this type of data structuring, may be top-down from coarse grain to fine grain (until it is mapped to its atomic level) or down-top from fine to coarse (from denormalization to normalization) structuring approach.

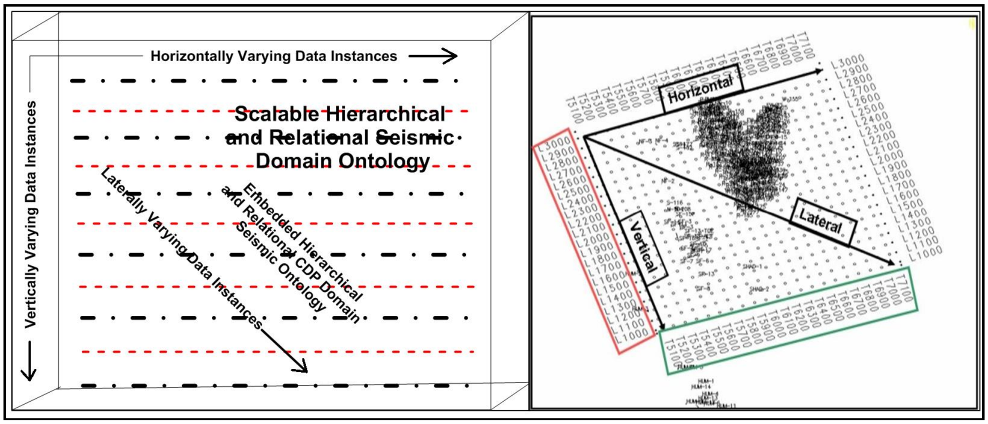

Horizontally Varying Data Instances
Vertically Varying Data Instances
Scalable Hierarchical and Relational Seismic Domain Ontology
Laterally Varying Data Instances
Embedded Hierarchical and Relational CDP Domain Seismic Ontology
Horizontal
Vertical
Lateral

Figure 4.9: Map views of 3D seismic *survey* areas with spatial dimensions – analysing their hierarchical and relational ontologies for geomorphic representation

As shown in Figure 4.9, the author identifies horizontal, vertical and lateral dimensions to describe them in relation to the hierarchical data structuring. Horizontal dimension is typically a *space*; vertical dimension is either time or depth; and lateral may be a composite dimension, comprising both space and time or depth dimensions. Another composited dimension such as *velocity* is derived from space and time or depth dimensions. *Line* dimension is a composite dimension (geographic nature), described in the horizontal, vertical and lateral directions.

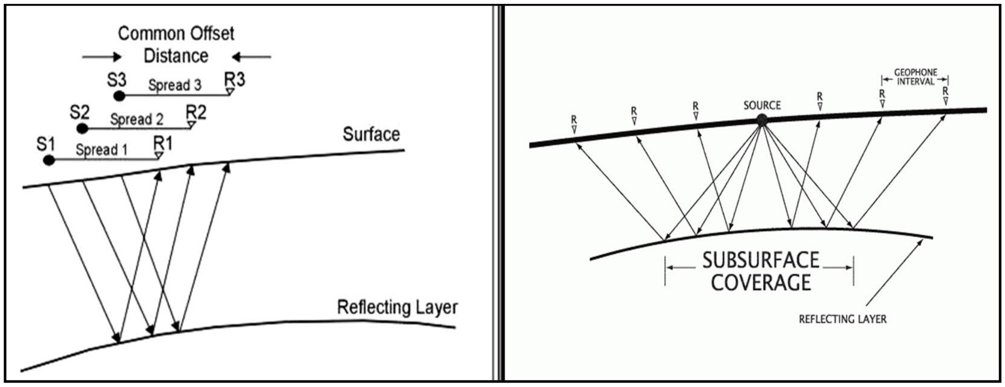


Figure 4.10a: Modelling data instances of *common offset point* (COP) domain ontologies

Another domain ontology description is associated with seismic reflecting surfaces and their data instances of Common Offset Point (COP) dimensions, as shown in Figure 4.10, which is an organization of COP data instances with laterally varying source and receiver-points (geophones) laid on the surface of the earth.

Horizontally varying dimensions

Each layer discussed in hierarchical data structuring, has distinct dimension attributing to a certain property. *Points* and *survey lines* described in seismic domains are represented each in different data layers in horizontally changing hierarchies. As shown in Figure 4.11, horizontally and vertically varying hierarchies and their associated data instances are gathered from the stacked data volumes. These values are interconnected to navigational systems, in which the coordinate data instances are described for all *point* dimensions of seismic amplitude values, representing *trough* and *peak* dimensions and their data instances as shown in Figure 4.10b.

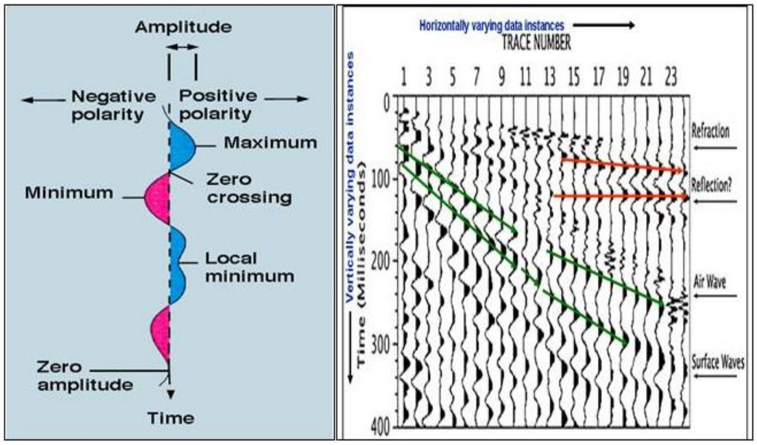


Figure 4.10b: Data instances captured from seismic records for describing and modelling *horizon* domain ontologies

Laterally varying dimensions

In case of laterally varying hierarchies, data layers are conceptualized laterally. For example, a geomorphic structured surface may have several *point* dimensions from several data grids. For example, 2D and 3D survey lines have multiple data instances gathered from laterally varying dimensions. As shown in Figures 4.8 – 4.10, laterally (hierarchical) varying dimensions and their instances are organized using ontologically described multidimensional data structuring procedures.

Relational data structuring-relational ontology

These are new ideas in the presentation of seismic data acquisition and seismic data processing arrays in which sensors' (geophones) data are vital for analysis and interpretation. The author explains all the dimensions involved in arrays' layouts (Lowrie 1997 and Sheriff 2002) in these models, calling them as "ontologies" because there are definite relationships existing among sensors' dimensional and factual data (for example, geophone, shot, offset distances are various dimensions). In fact, these are simulated arrays represented in terms of ontologies. Because of flexibility and robustness in presentation, the author designs different ontology models in these domains.

Relational data structuring is based on the relational model (Coronel 2011), in which data sources may have internal conceptual relationships, describing relations among several attributes, called relational-base ontology. Different datasets in a domain can make up a relation. A significant feature of relational data structuring is that attribute values come from a homogeneous pool of values called a domain, which represents all possible values in a finite dataset. However, more than one attribute can assume values from the same domain and a single attribute can take several values from a domain. In the seismic data structuring case, single "seismic" domain ontology contains *time* dimension and their instances contributing to describe a *time-surface*. In case of computed *depth* surface, besides *time* dimension, other composite dimension, *velocity* anomalies are involved in the integration process that can describe a shallow near-surface or deeper surface characterizing sink-hole features. These surfaces communicate and interact with representation of a geomorphic surface, narrating several structural features. Relationally organized data structures and their attribute dimensions (along with their instances) are integrated and intelligently stored in a warehouse environment. *Point* and *line* dimensions are made related in the navigational and geometrical data that support the seismic data and their patterns. These data are conceptualized from different *geophone* and *source* attribute dimensions. These are especially targeted to "seismic data acquisition techniques, in which relationships among signals/patterns coming from number of source (shots) positions and sensors (geophones placed on earth in an investigating area) locations and how they can be described in ontological sense. The aim of this presentation is to understand the seismic data patterns and enhancement of signal/noise ratios in the seismic data processing centres. The author describes different *point* dimensions (including their instances) that vertically, horizontally and laterally in *CDP, COP, CRP and CSP* domains (Figures 4.11-4.14) articulated, modelled and their instances stored in a warehouse environment.

***CDP Ontology*:** The common depth point (CDP) is a *point* dimension, derived from a configuration of *geophone* (a sensor for acquiring instances of seismic energies) and *shot point* (energizing dimension) dimensions (along a survey profile, as shown in X axis in Figure 4.11).The data instances are acquired from different 2D/3D survey profiles, laid on the surface of the ground (Figure 4.11a). The CDP ontologies are based on the relationships build based on common depth point dimensions extracted from survey profiles.

***COP Ontology*:** The common offset point (COP) dimensions, derived from a type of profile of *geophone* and *shot* locations, laid on the surface of ground. Only offset *point* dimensions and their fact instances are acquired for modelling. The purpose of this type of ontology structuring (Figure 4.11b) is to analyse the domain knowledge, affected by different offset distances between *sensors* and *shot* positions on the ground.

***CRP Ontology**:* These are sets of common receiver point (CRP) dimensions acquired from CRP profiles for modelling CRP data instances. Geological features affecting these types of survey profiles are investigated from CRP ontology structured data (Figure 4.12a).

***CSP Ontology**:* Similar to common receiver structuring, common shot point (CSP) instances acquired from CSP profiles (Figure 4.12b) are structured and analysed in its knowledge domain. CSP dimensions and their instances are deduced from survey profiles data.

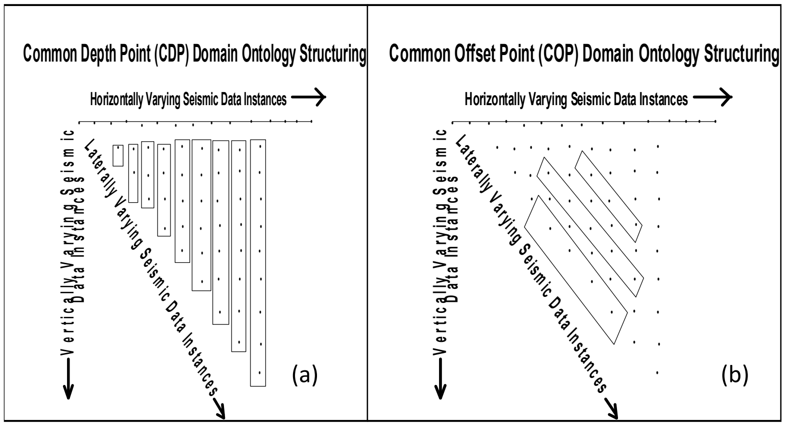

Common Depth Point (CDP) Domain Ontology Structuring
Horizontally Varying Seismic Data Instances
Vertically Varying Seismic Data Instances
Laterally Varying Seismic Data Instances
(a)
Common Offset Point (COP) Domain Ontology Structuring
Horizontally Varying Seismic Data Instances
Vertically Varying Seismic Data Instances
Laterally Varying Seismic Data Instances
(b)

Figure 4.11: (a) CDP Domain and (b) COP Domain ontology structuring

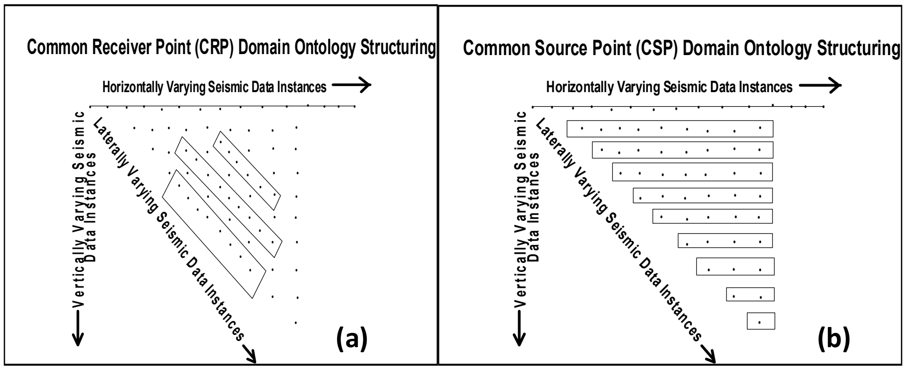

Common Receiver Point (CRP) Domain Ontology Structuring
Horizontally Varying Seismic Data Instances
Vertically Varying Seismic Data Instances
Laterally Varying Seismic Data Instances
(a)
Common Source Point (CSP) Domain Ontology Structuring
Horizontally Varying Seismic Data Instances
Vertically Varying Seismic Data Instances
Laterally Varying Seismic Data Instances
(b)

Figure 4.12: (a) *Common receiver* point (CRP) and (b) Common *source* point (CSP) domain ontology structuring

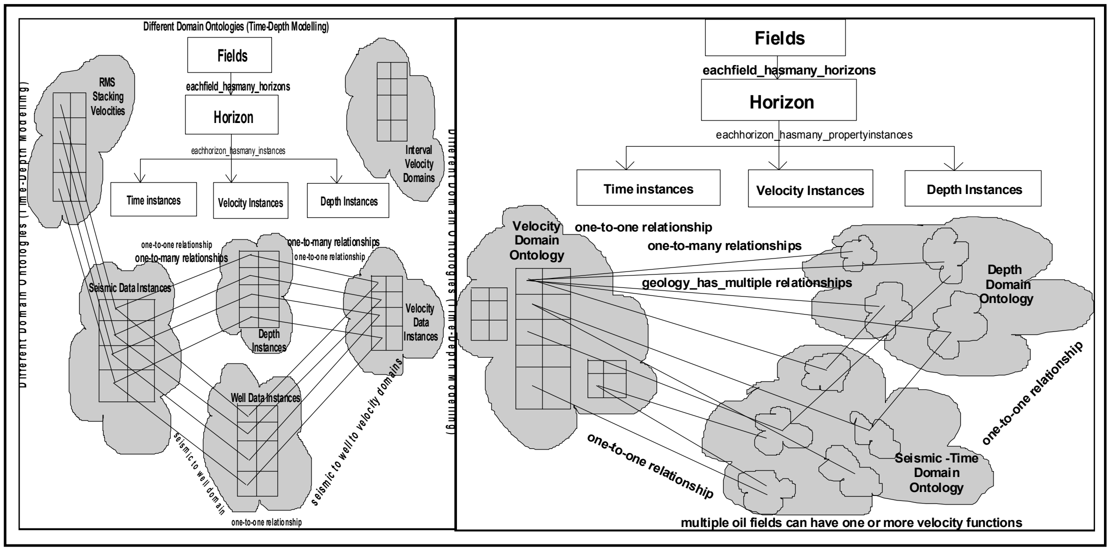

Different Domain Ontologies (Time-Depth Modelling)
Fields
eachfield_hasmany_horizons
Horizon
eachhorizon_hasmany_instances
Time instances
Velocity Instances
Depth Instances
RMS Stacking Velocities
Interval Velocity Domains
Different Domain Ontologies (Time-Depth Modelling)
one-to-one relationship
one-to-many relationships
one-to-many relationships
one-to-one relationship
Seismic Data Instances
Depth Instances
Velocity Data Instances
Well Data Instances
seismic to well domain
seismic to well to velocity domains
one-to-one relationship
Different Domain Ontologies (Time-Depth Modelling)
Fields
eachfield_hasmany_horizons
Horizon
eachhorizon_hasmany_propertyinstances
Time instances
Velocity Instances
Depth Instances
Velocity Domain Ontology
one-to-one relationship
one-to-many relationships
geology_has_multiple relationships
Depth Domain Ontology
one-to-one relationship
one-to-one relationship
Seismic -Time Domain Ontology
multiple oil fields can have one or more velocity functions

Figure 4.13: *Horizon* based relational-domain ontology structuring

The CDP, COP, CRP and CSP ontologies are used to build horizons. As described in Figures 4.13 and 4.14, different horizon attributes are gathered to accommodate them in multidimensional structuring. *Peak* and *trough* dimensions (Figure 4.14) of the seismic attributes acquired from surveys, make connections with horizon attribute dimensions.

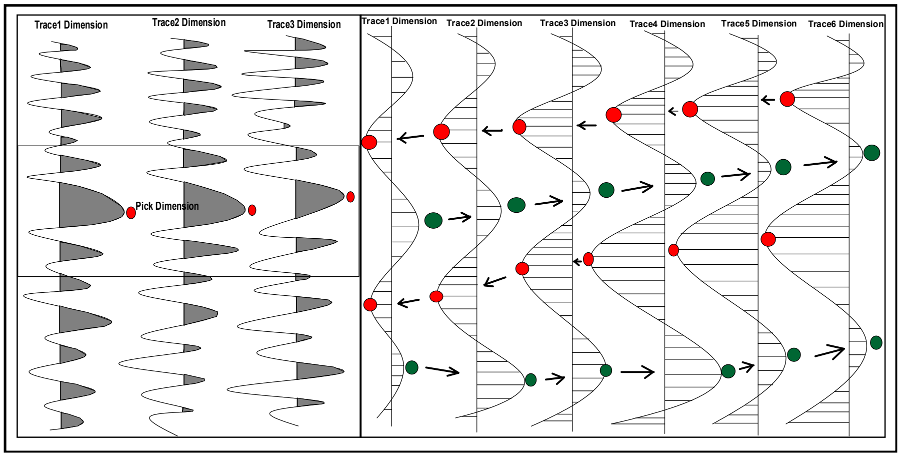


Figure 4.14: *Seismic-pick* dimensions, showing (arrowed) *horizon* relational data instances

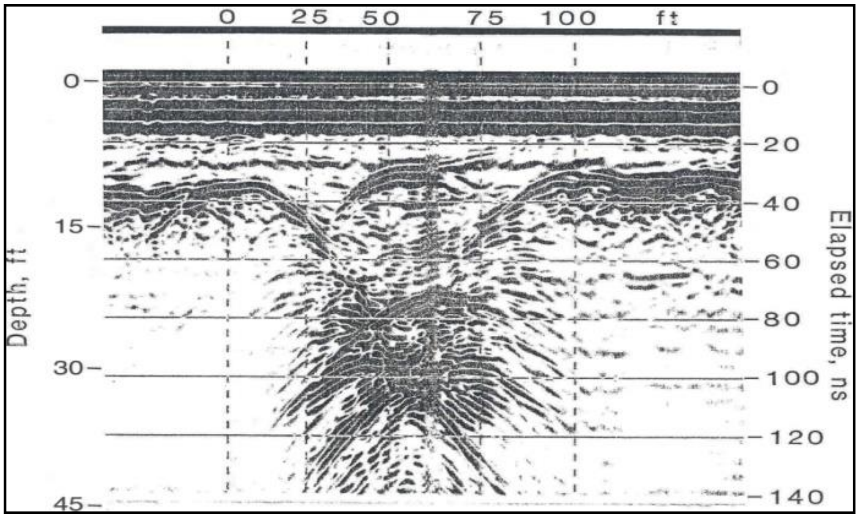


Figure 4.15: Seismic view showing *geomorphic* deformation, around a sink hole

As shown in Figures 4.15 and 4.16, in typical seismic sections, impressions or images of a typical sink hole are visualized, in which multiple dimensions such as *depth, time* and *space* are described. Instances of lateral dimensions are extracted from the digitally recorded data. The contoured surface has an alignment of equal time, or depth or velocity data values, taken from several in-lines and cross-lines, with *points*; each point has a property value or an instance. As stated earlier, each *point* consists of several dimensions and their instances.

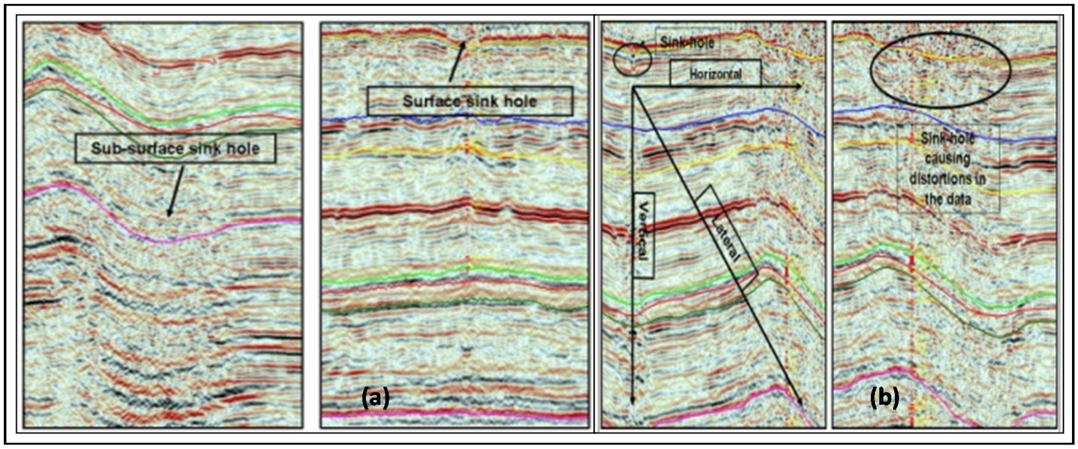


Figure 4.16: Data views of seismic data instances in space and time dimensions – depicting signatures of sink-holes, causing distortions in the seismic amplitudes and frequencies

The data associated with sinkholes, are considered for interpreting shallow surfaces as shown in Figures 4.15 and 4.16. The data instances of deeper sections of the seismic data are affected by shallow data instances. These conceptualized attribute dimensions are used in the modelling process. The data instances from these dimensions are extracted from the digitally recorded seismic data from both shallow and deeper sink-hole events.

Vertically varying data relationships

The most common dimension, varying its instance in vertical direction is either time or depth. Either time or depth has definite relationship with geomorphic structure attributes. From seismic times, depth relationship is built using another relationship, called velocity. Velocity is a key composited dimension and derived from instances attributing to depth variations in vertical direction. Utmost care is necessary to interpret seismic data instances along computed survey grids. Having generated contoured surfaces at chosen time or depth, one can interpret for geomorphic data features. Other dimensions changing in vertical direction are seismic amplitude, frequency and phase attributes. The knowledge of geomorphic surface structure depends on interpretation of the vertically varying data attributes.

Horizontally varying data relationships

Most common dimension, attributing to changes in the horizontal direction and connecting horizontally varying data relationships, is spatial dimension, with a distance property, separating two *points* on a surface. Distance between points has a definite relationship with a geomorphic structure property. Horizontal variations in seismic times, depth or velocity dimensions and their attributes are interpreted to have associated with conceptual structural variations of geomorphic surface features.

Laterally varying data relationships

The relationships of data dimensions, built laterally are conceptual. Ontologically, all these data instances are conceptually described and interpreted along laterally changing seismic data structural features. As shown in Figure 4.9, the survey location map, representing positional or coordinate data, narrates different domains of ontology in which *surveys* and *wells* are located. There data instances are common in different domains. Relationships are conceptualized wherever they need to be built for the purpose of extracting knowledge of hazards on geomorphic structured surfaces.

Warehouse modelling of ontologically structured data instances

The changes in geomorphology are causative to impacts on environmental ecology because of systems' interconnectivity in nature. This phenomenon is well understood in a data warehouse modelling situation (framework as described in Figures 3.36, 3.37 and 3.38 in Chapter 3), when ontology modelling and integrating the datasets of geomorphic structures and other associated geological outcrop features. The author has gathered large amount of data to organize intelligently from different levels of hierarchies in both horizontal and vertical dimensional structuring process. The author documents and stores all the dimension tables (with rows and columns) with their corresponding fact tables intelligently using normalization process. For fine grained data structuring purposes, at places, relationships are denormalized among several dimension and fact tables.

In addition to *time*, *depth, location, horizon* and *positional* dimensions (Figures 4.9 – 4.16), *dip* and *strike* dimensions and their corresponding instances are significant in interpreting the structural attitudes of a horizon, from which all the associated relationships are correlated and mapped respectively in horizontal and vertical *point-to-point* and *dip* tracking dimensions. One can take advantage of the fact that multiple dimensions existing in *exploration* super-type dimension (Nimmagadda and Dreher 2006 and Nimmagadda and Dreher 2007) are effectively used in building conceptual data models and their integration into metadata for extracting new knowledge, especially in between seismic and well-data domains. Basic framework used in Figures

3.36 and 3.37, is the basis to construct map views from ontology based multidimensional metadata (discussed more details in Chapter 5) for interpretation. Logically organizing the seismic and well data instances in the form of metadata is an art of data representation, in which metadata serves to identify the contents and location of these datasets in a warehouse.

The author considers seismic, well-data and other available geological inputs for ontology-based warehouse modelling. Besides seismic time, velocity (geophysical dimensions) and formation depth (geological dimension) dimensions, structure dimension with "structure-high" and "structure-low" attributes (discussed in OLAP views in the forthcoming sections) are also interpreted. Type of karsts, volume of karsts, which subside the producing geological formations and liquefied media, which leach the karsts, have definite impacts on amplitude and frequency attributes of seismic data and their instances. This has close relationship with conceptualized "structure" attributes. Karst terrains represent unique characteristics of seismic data instances along *dip* dimension and they are different when data instances are represented along *strike* dimension. The attribute properties change as per geological *attitude* dimensions.

4.2.1 Business data entities

The Relational Database Management System (RDBMS) technology has become an accepted methodology (D'Orazio and Happel 1996 and Hoffer et al. 2005) for primary data storage and access platform for knowledge-tone applications. Data structures in the form of logical and physical designs, data manipulation procedures like SQL query, data integrity and security-stored procedures, are important components of RDBMS. RDBMS manages data as a collection of tables in which all data relationships are represented by common values in related tables. ER analysis and normalization are performed, before designing database structures, for oil and gas data warehouses. Issues related to indexing like primary and foreign keys have also been analyzed during normalization process. Traditionally, architecture of data warehousing, is optimized by an approved RDBMS, especially for decision support.

The author identifies entities (dimensions), attributes and relationships to create data tables for various attributes relevant to navigation/satellite, exploration, drilling, and production and support service functions. To begin with, the author carries out conceptual modeling and data mapping for building a RDBMS for oil and gas business environment. Cross-reference keys link the tables, representing the relationships between entities. The primary and foreign keys provide an easy access to the databases. Similar to the petroleum oil and gas data structuring, mineral oil and gas data also possess several high-level and low-level business operations. Accordingly, data items of mineral databases possess definite relationships among their entities or dimensions and attribute values. A sample of Entity-Relation (ER) is presented in Figure 4.17.

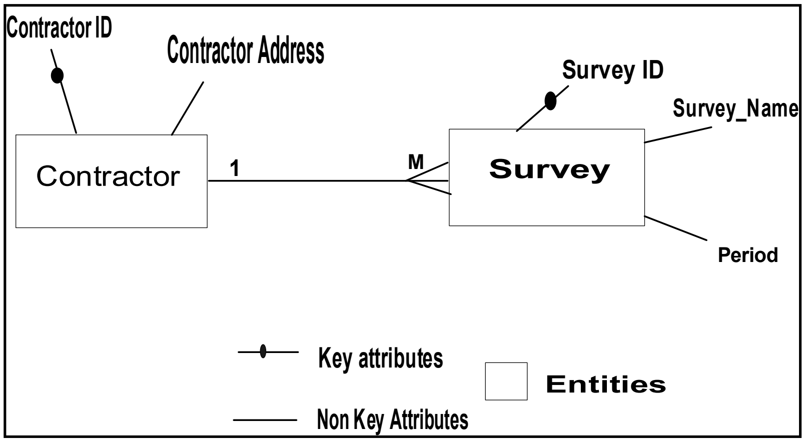


Figure 4.17: A simple ER diagram for petro2 (database) -*surveys* showing one-to-many relationships

Another conceptual model as described in Figure 4.18 represents a database structure enabling the oil and gas user to model the spatial-temporal changes of the reservoir region units. The data structure combines an ER model with an arc-node structure (Ott and Swiaczny 2001) on the logical level. In another example, the basic problem derived from the reservoir units from certain geological regions, may only be valid for a certain period of time, since all the world oil or gas reservoirs have definite life span. Reservoir is crucial in oilplay analysis. The reservoir pay aerial or region objects vary with time and space as well. When they originally produced petroleum, they possess definite boundaries, as time passes, they have differing boundaries. A spatial unit may evolve from either one or the combination of several spatial units. It may lose or gain an area from several units. The author characterizes the spatial reservoir units by spatial and temporal object attributes or dimensions as well as their relations to the neighboring reservoir unit. Here each object is related to other object categories: an oil producing basin is a part of state owned oil and gas company. Each entity is described by attribute data, which may carry a period dimension within the lifespan of the object. The author translates ER model into a star schema multidimensional model, to finally accommodate into the data warehouse that comprise of spatial-temporal dimensional oil and gas data.

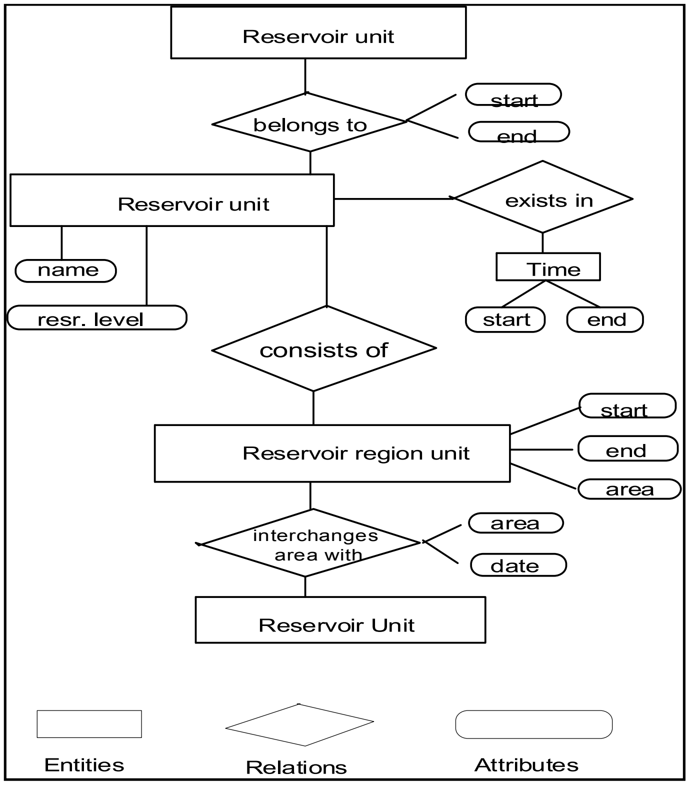


Figure 4.18: Ontology base "reservoir" conceptual model

Methodologies

D'Orazio and Happel (1996) and Rob and Coronel (2004) discuss several database designs demonstrating different types of data models in industrial scenarios. The author draws an entity- relationship model among the identified business sub-type entities. As depicted in a conceptualized model in Figure 4.19, one can go for bottom-up as a specialized entity and top-down as a generalized entity. Exploration as a generalized type business entity is further divided into smaller units as specialized entities. The super type is interrelated with sub-type entities as shown in Figure 4.19 through sharing of common attributes of all entity types. R1 – R10 are relationships constructed among business entities, are represented as associated entities. While mapping entity-relationships, the author extends ER models incrementally further to build associative relations as entities through an extended entity-relationship (E^2/R) mapping approach. The precise definitions of generalized and specialized entities, ER and E^2/R are discussed in the forthcoming sections.

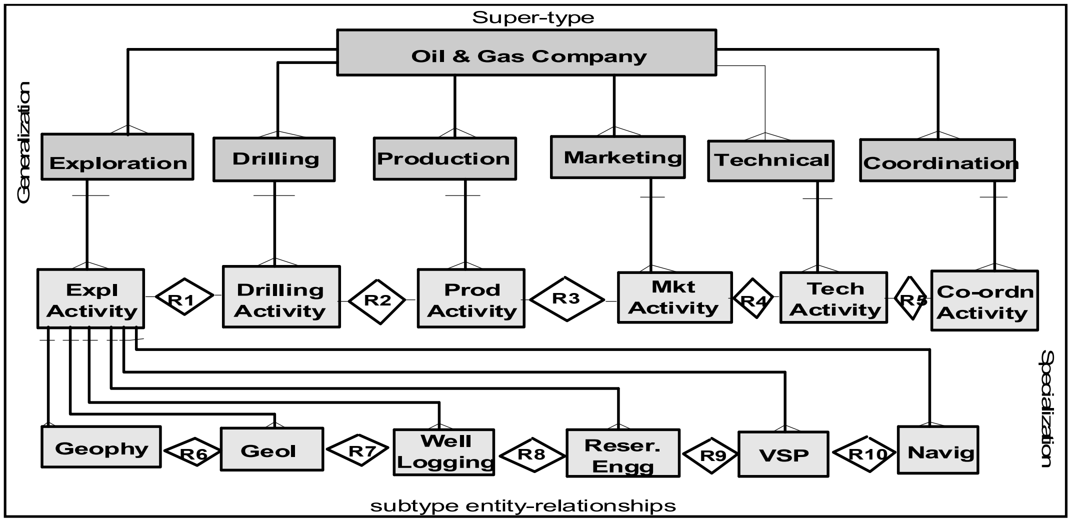


Figure 4.19: An entity relationship (ER) model showing specialization and generalization

The author constructs a conceptual model using the ER data mapping constructs, such as, entities, attributes and their relationships. As shown Figure 4.20, a logical model constructed and presented, consist of *navigational, seismic*, *VSP* and *well-logging* as subclasses of *exploration* class. The author identifies key attributes associated with each and every sub-type entities and represents relationships that mapped between sub-type entities with diamond boxes. These relationships are evolved due to the business scenarios, attributed as associative entities.

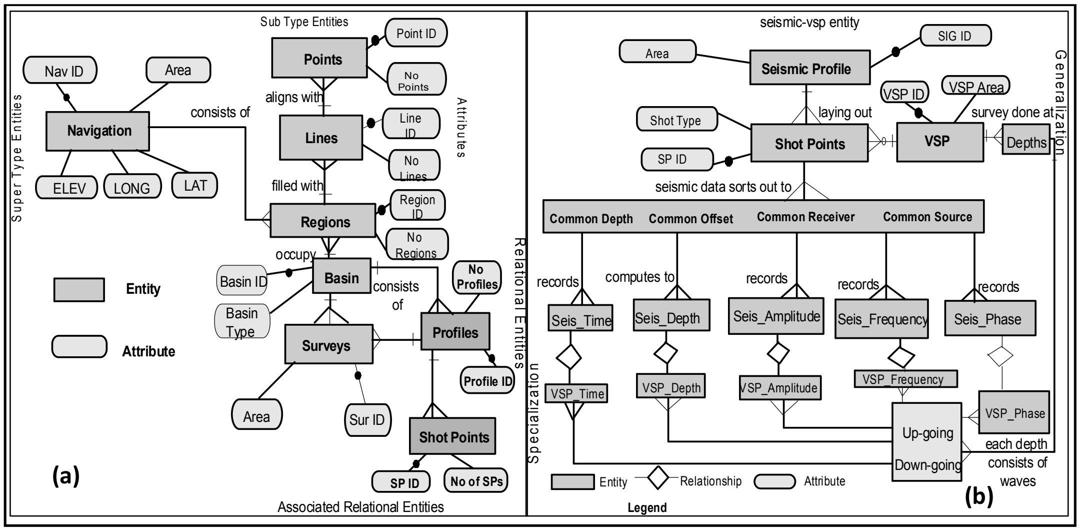


Figure 4.20: ER models for (a) *navigational-surveys* and (b) *seismic-VSP* entities

Only key entities are displayed in these ER models. Similarly, ER models constructed for *VSP* (Vertical Seismic Profiling) and *well-logging* entities are in Figure 4.21. The author maps one-to-one, one-to-many and many-to-many relationships wherever they exist. How business rules imposed, relevant to exploration data structures, is explained in the following sections.

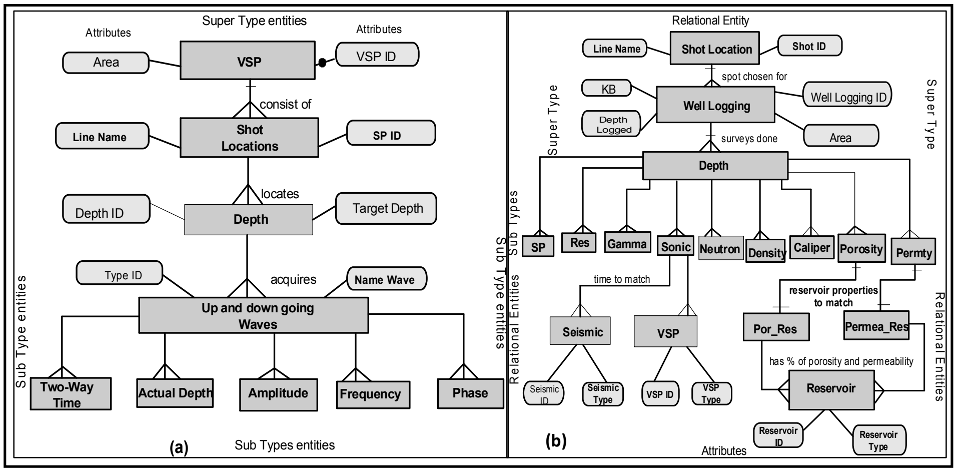


Figure 4.21: ER models for (a) *VSP* and (b) *well-logging* (sub-surface survey) data entities

Enforcing business rules in the modelling process: ER and E^2/R (extended entity-relationship) diagrams are a useful means of setting business rules (Nimmagadda and Rudra 2005a). Participation and dis-jointness associated with super-type and sub-type entities or objects are expressions of business rules associated with specialized sub-type relationships. The cardinality, optionality, derivation, integrity, action and structural assertions, encapsulation of attributes and operations of oil and gas industry data are key business rules incorporated while mapping complex object data structures. The polymorphism and inheritance are important other concepts of object oriented modelling, applicable to heterogeneous petroleum data structures as well. In many companies, the business rules represented, are in the form of derivations and algorithms, which enforce organizations and make the object data structures and business models work as per the enforced rules. Logical modelling depends on how logical object classes are identified, mapped and tested as per enforced rules.

Description of relational exploration data structures: As an example, the author demonstrates how relational relationships among several entities are presented in a logical sense using the data structure. The text description of structure of a relation by a shorthand notation in which the relationships followed by the names of attributes in that relation is represented as:

Navigation (Nav ID, area, longitude, latitude, elevation, Region ID)

Region (Region ID, No_of _Regions, Area_of _Region, Line ID)
Line (Line ID, No_of _Lines, Length, Point ID)
Point (Point ID, No_of _Points, Distance, Shot ID)
Basin (BasinID, No_of_Basins, BasinType, ProfileName, Profile ID)
Profile (Profile ID, No_of_Profiles, Shot ID)
Shot (Shot ID, Point ID, No_of_Shots, Shot_Distance, ShotLineNumber)

In the above notation, solid lines represent primary key attributes and the dashed lines to foreign key attributes through which the entities are made interrelated. “Shot” has a composite key consisting of the attribute keys *Shot ID* and *Point ID*. Similar relations can be represented with other data structures. Author discusses exploration, drilling, oil and gas production and coordination identified, as key entities in the following sections. Since exploration is a key business entity in any oil and gas company’s overall business system, how it is described as a business entity in association with other sub-type entities.

Exploration as a business entity

The entities and attributes identified from exploration business are accountable for making up connections within the super type entity and also share common attributes with other sub type entities of other super-type business entities. *Geophysics*, *geology*, *reservoir*, *well-logging*, *VSP* and other research services are key functions, interpreted as inherent sub-type entities of exploration business entity. They share common attributes among sub-type entities. Further, “geophysics” sub-type entity possesses inherent sub-type entities interpreted as *data acquisition*, *data processing* and *data interpretation*, based on their activities. The author classifies functions and their activities within several operational units to identify as inherent associativity entities. Again strategic planning, management control, operational control and data reporting, control each entity or operational unit of exploration business. Manager, who is playing a key role in the assembly of entities, also performs planning, organizing, staffing, directing and controlling tasks.

As shown in Figure 4.22, R1 – R3 relationships constructed between business entities appear to possess associativity entities. These associated entities are based on presence of common attributes between business entities. Shot location, latitude, longitude and elevation are key attributes that are associated with other business entities. Exploration and drilling business entities share shot location (a primary key) attribute. Many other sub-type entities of exploration business super-type entity share common attributes though they are not shown in the ER models. Logical models that have been generated showing entity relationships (sharing common attribute properties) of exploration and drilling business entities are shown in Figure 4.22. The author takes these entities in the ER mapping process and apples business rules of one-to-one, one-to-many and many-to-many relationships wherever they exist.

Geophysics, geology, VSP, reservoir are key operational activities having one-to-many relationships among exploration activities. There are attributes derived for each and every specialized entity as well as in the generalized exploration entity. As an example, exploration entity is a part of petroleum system, again inherent with associative sub-type entities.

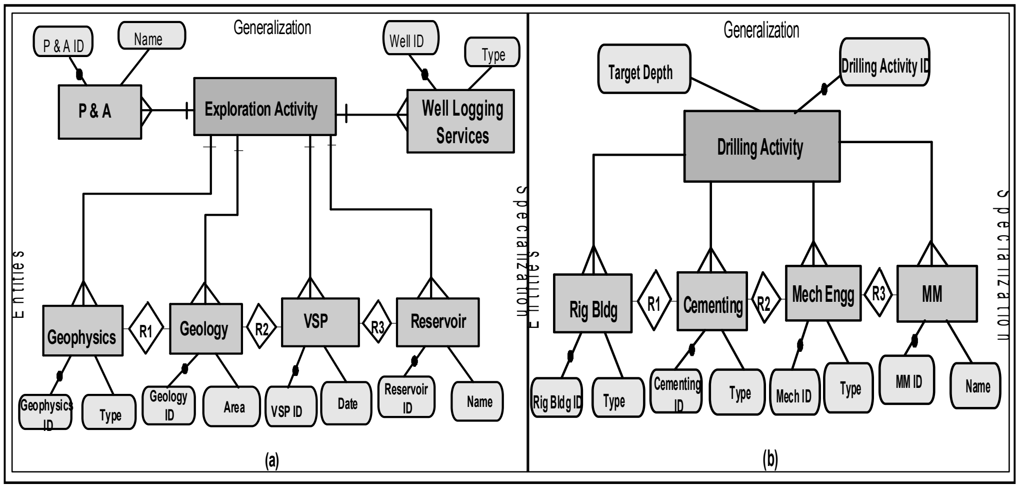


Figure 4.22: Entity relationships (a) entity relationship (ER) models of exploration activity and (b) drilling activity

A petroleum system, as a generalized business entity, consists of several specialized entities, such as *reservoir*, *source*, *structure (trap)*, *migration* paths and timing of deposition of structures and occurrence of sedimentation. Petroleum systems of onshore and offshore gulf basins need integration of petroleum system elements or specialized entities, so that explorationist can understand and build new knowledge of each petroleum system of each and every sub-basin.

Drilling as a business entity

The next business entity immediately after exploration is drilling. Once a geotechnical order or a detailed exploratory proposal plan with financial commitments is approved by the company's technical management, drilling infrastructure and associated logistics are mobilized to an approved site either for detailed exploration, development, step-out or delineation well, from which data analyst starts identifying again all the associativity entities. Author identifies entities associated with strategic planning, management control, operational control and data reporting. The number of wells, type of well, and categories of well that will be decided depending upon the size of prospect, are key attributes. Type of rig, time involved while drilling a well, targeted depth, meterage to be drilled, cycle speed and other budgetary constraints are also identified.

A logical model is built using the entities and their relationship with associated attributes of drilling business entity as shown in Figure 4.22. There may be common attributes with other business entities, such as exploration and production business, which are reported and included by the author in the conceptual model. Author analyses with rigor, all the data pertained to engineering, cementing, chemistry and variety of chemicals used, safety, rig-building, personnel and administration, finance and accounting. Various other engineering activities associated with drilling business such as mechanical, civil, electronics and telecommunications and electrical engineering are made involved in the conceptual data models by the author. These entities and their associated attributes are also present in the technical business subsystem.

When once, the target drilled depth is reached, the contractor conducts the logging surveys within the drilled-well including the vertical seismic profiling (VSP) surveys. Again author analyses systematically all the planning, management and operational control activities associated with the conduct of the valuable surveys in the drilling business subsystem. *Shot location*, *latitude, longitude* and *elevation* of the well site are key attributes that are associated with exploration business subsystem as well. The author is involved with all entities and attributes with conceptual ER modeling (Figure 4.22) and the mapping is done through one-to-one and one-to-many and many-to-many relationships. Business rules set from all constraints are incorporated in the data mapping.

Production as a business entity

After having explored and drilled the prospect (from previous roles), if the drilled well proves to be oil bearing, the next step is production operation that is carried out for extracting oil and or gas from targeted depths. Production engineering, exploitation of oil and gas from drilled well, oil dispatch, refinery supply, liquefied petroleum gas (LPG) production, wells work over, the flow lines laid to transfer oil and gas to group gathering stations (GGS) are key entities of the production business. The author identifies all the attributes associated with these entities, for not only associating with other sub-types of production business entity, but also with its associated super-type such as, exploration and drilling business entities, discussed earlier.

Wells and geological formations tested, type of activation of sick well, secondary recovery type are key attributes, besides, *shot location, latitude, longitude* and *elevation* of the current producing well data. The author identifies *men*, *machinery money* dimensions and other constraints, as entities with their corresponding attributes. *Shot location*, *latitude*, *longitude* and *elevation* are other data attributes that are associated with each other business subsystems. A logical model involving the entities and relationships is generated as shown in Figure 4.23. These data are mapped through one-to-one and one-to-many relationships.

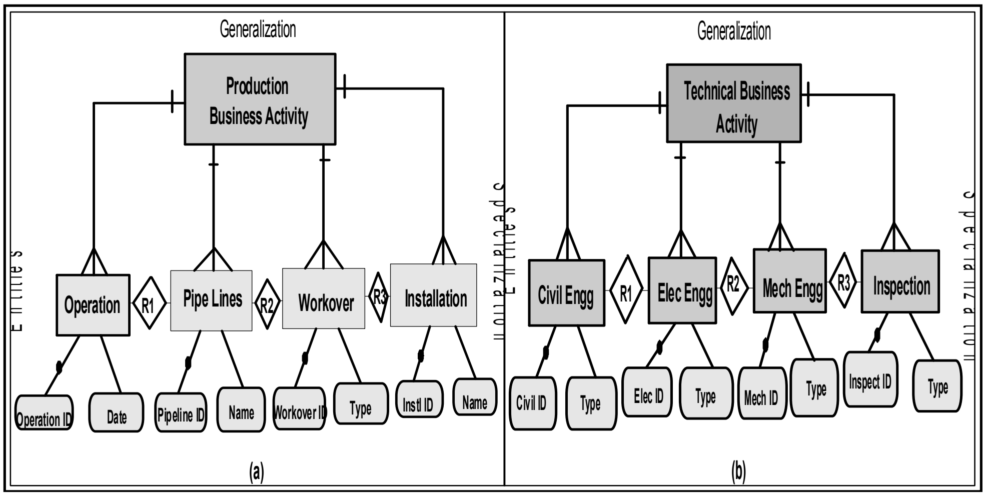


Figure 4.23: Entity relationships (a) model for production business activity and (b) technical business activity

Technical as a business entity

The technical business manager generally provides logistics and materials that needed to the exploration, drilling and production units. This entity concentrates on civil, electrical, mechanical, transport, inspection, quality control, safety and environment, establishment of task force disposals and other combined activities pertained to other subsystems such as exploration, drilling and production businesses. Data relevant to mechanical, electrical, logistics, transport, tele-communications, civil, safety and environment, inspection, energy conservation, technology up gradation, import substitution, cost reduction engineering and inventory are key entities of this business subsystem. The author creates a logical ER model with mapping of one-to-one, one-to-many and many-to-many relationships among different sub-type entities of this business subsystem as shown in the Figure 4.23.

Coordination and collaboration as business entity

In general, the manager of the coordination business deals with personnel and administration; finance and accounts aspects of the oil and gas company. Manpower planning, selection and recruitment, training and development of staff, welfare, security, industrial relations, sports, and public relations are key activities of this business subsystem. In addition, financial performances and sales turn over, plan and non-plan expenditures, material-in-transit, disposal and distribution of materials and material procurement are other entities of coordination business. Similar ER models can be drawn for this business. As shown in the Figures 4.17 – 4.24, all the specialized entities represented within a generalized business entity are associated either with one-to-one, one-to-many, and or many-to-many relations. The graphical representation of entity-relationship mapping is called ER modeling. It is not necessary all the relationships exist and explainable in all data structures that undergo mapping. Data analyst is responsible for identifying these relationships and mapping them appropriately, so that the logical models generated can strictly be implemented. Entities, attributes and relationships are key constructs of these ER models. The author discusses how the specialized and generalized entity concepts applied in oil and gas ER modelling scenario, in the next section.

Representing specialization and generalization in ER mapping

As discussed in Coronel (2011) and Hoffer et al. (2005), author uses super-type/subtype relations relationships to represent the entities of oil and gas company's data models. Generalization and specialization are key processes for examining the super-type and subtype relationships.

Generalization entity

In data modeling, the generalization is the process of defining a more general entity type from a set of more specialized entity types. It is a bottom up process. Examples of such generalized entities are shown in Figures 4.26 – 4.28. The *exploration*, *drilling*, *production*, *technical and coordination* business entities are sub-type entities of super-type Oil and Gas Company generalized entity system.

Specialization entity

Unlike generalization, the specialization is top-down process in which, the author examines one or more subtypes of the super-type and formation of super-type/subtype relationships. As shown in Figure 4.24, an entity type named, 'drilled well' has several attributes. An attribute called, *formation* is multi-valued, because there may be more than one formation with an associated *formation name* with its *ID*. *Drilled Well* is a generalized with specialized exploratory or development subtypes. ER logical model suggests that *formation* and its *ID* are also associated with specialized subtype entities. A new relationship may be evolved because of these associated relationships between entities.

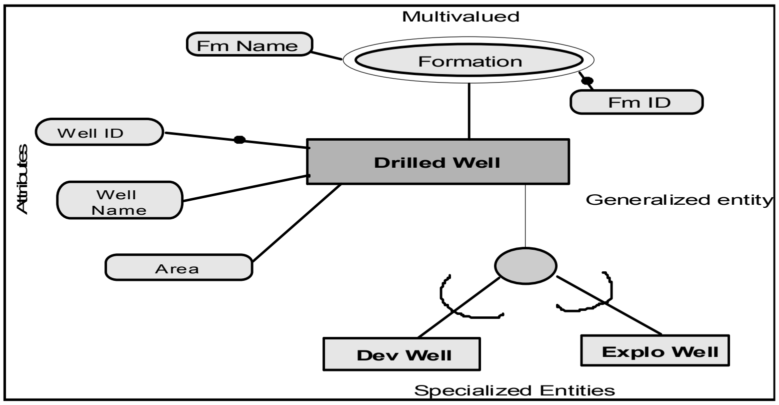


Figure 4.24: Specialization of a "drilled well" entity

These two processes are key techniques to describe relationships between super type and sub type entities. Data analyst perceives the combined usage of these processes in the development of conceptual and logical data models. Hoffer et al. (2005) also discuss several constraints of specifying super type and sub type entities. Extended entity relationships (E^2/R) ultimately lead to incremental building of complex conceptual data models, which are discussed in the following sections.

Extended ER (E^2/R) model

With the advent of database modelling using relational database technology, author introduces more advanced and new forms of abstractions in conceptual modelling. More sophisticated abstract data types are emerging in the complex oil and gas company situations. At this point, the author exploits certain extensions introduced to the ER model, which are grouped under the following sections.

Total mapping constraints

As shown in Figure 4.25a, total relationships are denoted by putting a dot (.) on the side of the relationship facing the constrained entity type. It is likely to have relationships that are total on both sides. This extension not only provides more expressiveness in database design, but also provides mechanism to enforce updates. If the constrained entity is updated, the relationships are also updated and vice versa. If an *employee* is hired to work in an oil and gas company*, employee's name* must be entered in the database with the work to which he or she is associated. If the *driller* is deleted, some other *driller* must be associated with the assignment; otherwise entity "well" must be deleted from the database. Similar situation can be narrated for a petroleum system, reservoir qualities and quantities are entered in the database periodically.

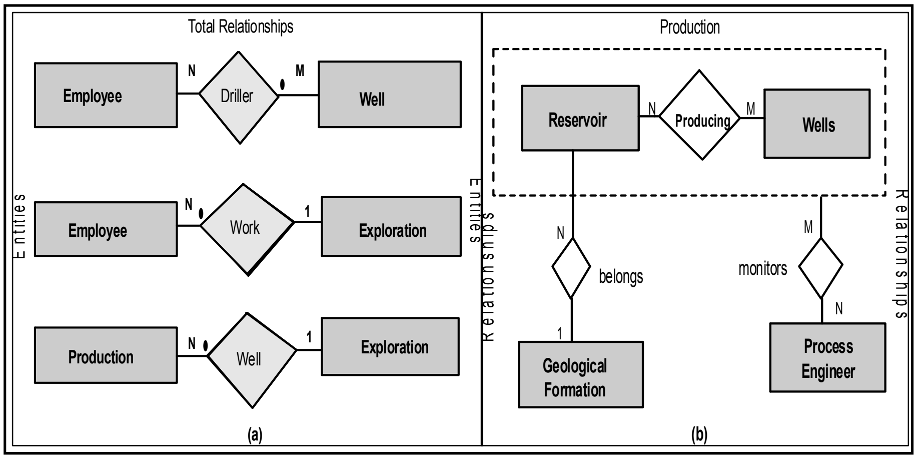


Figure 4.25: Entity relationships (a) oil and gas company business entities showing total relationships and constraints and (b) aggregations at type levels

Aggregation at data type level

This helps to build entity types of higher levels. The aggregation serves (see Figure 4.25b) as a higher-level entity, hence further relationships can be entered. In the original ER model, relationship types can only connect the entity types, implying that there cannot be relationship with another relationship. The aggregation overcomes this situation and is constructed in a recursive manner to build entities of higher and higher levels. The author builds a relationship between an aggregate and an entity type (or another aggregate) and calls it yet another aggregate, and so on.

Generalization and classification in E^2/R

As shown in Figure 4.26, a generalization is represented with an 'IS-A' designator and adapt to E^2/R modelling. Author converts conceptual schema to external data schemas (see Figure 4.27) as discussed in Ozkaharan (1990). The author demonstrates E^2/R models for implementation. If inconsistencies arise when two specialized models are integrated into a generalized model, extended entity models make them consistent. In the extended-entity-relationship (E^2/R) model, the author extends the basic ER model to include the super type/sub-type relationships and remove the inconsistencies that arise while linking the specialized and generalized data structures.

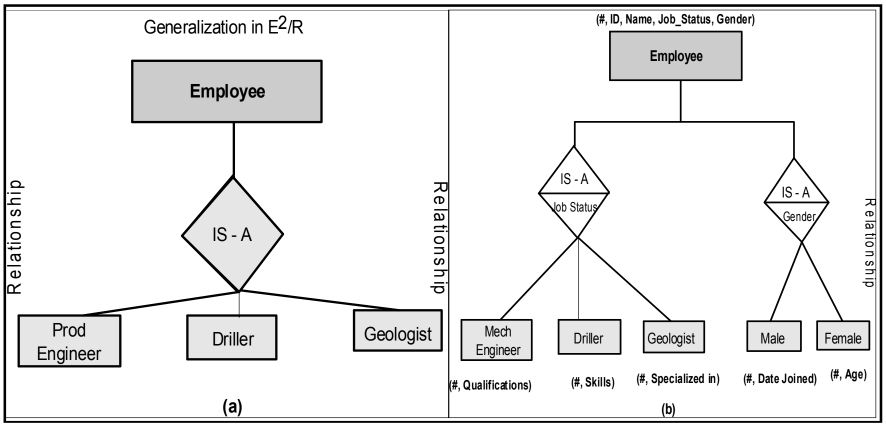


Figure 4.26: Entities (a) business entities showing generalization entity and (b) its properties with multiple classifications in E^2/R

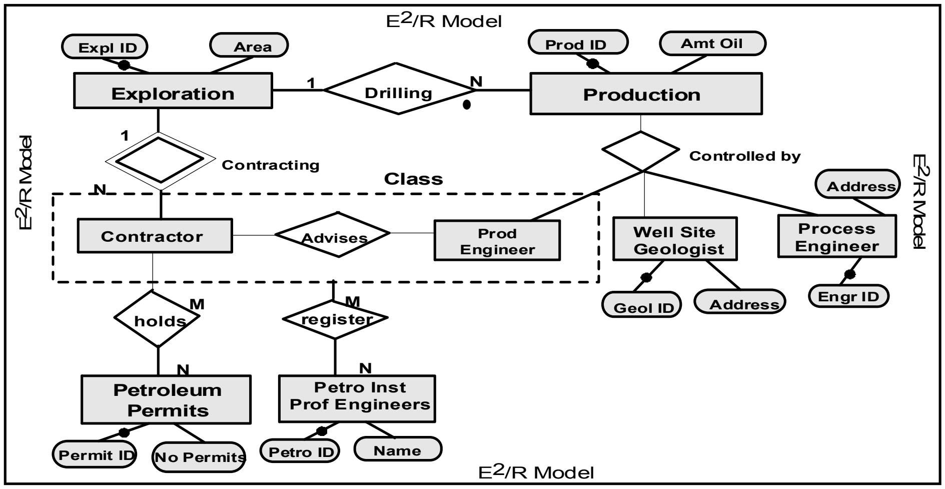


Figure 4.27: A conceptual schema in E^2/R (for E&P Oil & Gas Company)

The extended multidimensional (M^2/R)

The database modelling has become more sophisticated with advent of new tools and ideas database technology (Ozkarahan 1990). Keeping view variety of applications, the author introduces more advanced and new forms of abstractions in conceptual modelling. More sophisticated abstract data types are emerging in the case of complex oil and gas company situation. At this point, certain extensions are introduced to the existing MR model, which are grouped in the following sections.

Aggregation at type level

This facilitates to build dimension types of higher levels. The aggregation serves as a higher-level dimension; hence further relationships can be entered. In the original MR model, relationship types can only connect the dimension types, implying there cannot be relationship with a relationship. The aggregation overcomes this situation and the author constructs it in a recursive manner to build dimensions of higher and higher levels. Such relationships are between aggregate and a dimension type (or another aggregate) and call it yet another aggregate, and so on.

Generalization and classification in M^2/R

As shown in Figure 4.28, it is a generalization model to represent with an ‘IS-A’ designator and adapt to M^2/R modelling. Different designators can be used within the same hierarchy as shown in Figure 4.28. Conceptual schema is converted to external data schemas (see Figure 4.29) as discussed in Ozkarahan (1990). The author demonstrates EMR models for their implementation. If inconsistencies arise when two specialized models are integrated in a generalized model, extended dimension models make them consistent. In the extended-multidimensional-relationship (EMR) model, the author extends the basic MR model to include the super type/sub-type relationships and remove the inconsistencies that arise while linking the specialized and generalized data structures.

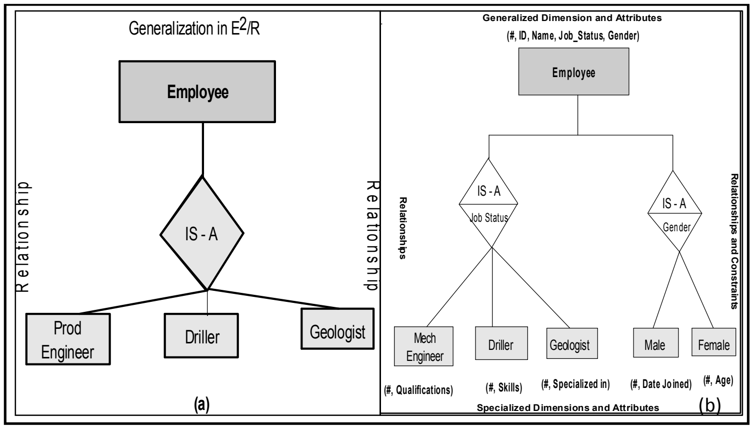


Figure 4.28: Dimensions (a) business data dimensions showing generalization and (b) its properties with multiple classifications in M^2/R

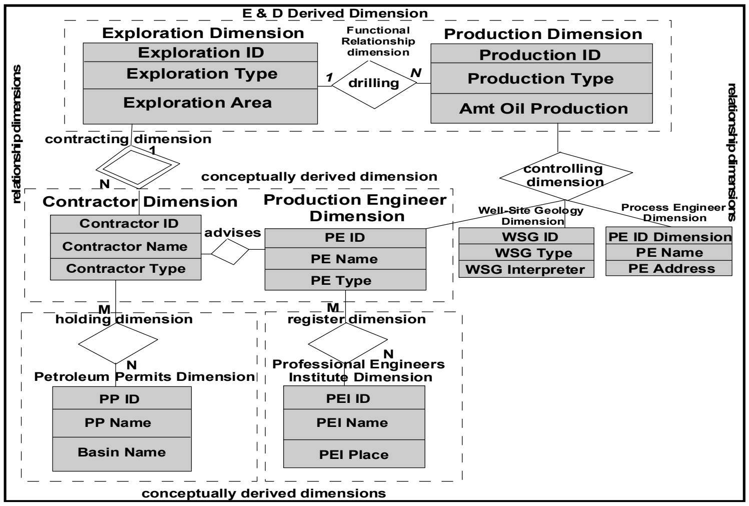


Figure 4.29: A conceptual multidimensional schema in M^2/R (Exploration & Production)

So far, the author conceptualizes company's data ontologically to describe all in the ER diagrams. How objects accommodated within the ontological modelling are described in the following sections.

4.2.2 Business data objects

Applying warehousing and business intelligence concepts to oil and gas data sources yield interesting results, especially when oil and gas information technology (OGIT) is coupled with non-oil and gas information technologies (e.g. coupling OGIT with a data warehouse). When data warehousing technology linked with oil and gas database technologies, issues such as slow response times, poor image or map navigation capabilities, data fusion bottlenecks have to be resolved. As such OGIT does mostly transactional processing with less analysis and hardly supports decision-making applications. However, user point of view, functionalities and response times of both spatial (such as OGIT) and non-spatial technologies can be tuned.

In oil and gas databases, the author handles different types of data, such as exploration, drilling and production used to be typically stored and processed by separate tools. Typically these data are in the form of numeric and spatial. Semantic data pertained to the OGIT can as well be stored separately. In which case, the author models data as a set of object instances. Objects have a set of attributes, each of which can take one or several values. Objects are linked with other objects by associations. A class represents a set of similar objects and summarizes the related attributes and associations. With this model, author views data as a labelled directed graph. Specifically, each node in the graphical representation corresponds to either an object or a literal. Edges from an object to a literal are attributes, and edges from one object to another are associations. However, in the case of ontology based multidimensional representation, the author derives metadata from an integrated framework, in which the entire data structuring organization and process in oil and gas domain, including effective use of data mining, visualization and interpretation are facilitated.

The author presents geometric elements in different objects. The conceptualization of the real world spatial objects is interpreted through object oriented data modeling (Ott and Swiaczny 2001) reducing the complexity of object data structures. For example, geological field samples are represented in pointed objects. “Seismic survey lines” is another spatial-temporal view showing occurrence of spatial events in different basins represented in line objects. The navigational data of oil and gas in different polygons and 2D geophysical contours, representing maps as region objects, are other views of geometric elements as shown in Figure 4.30.

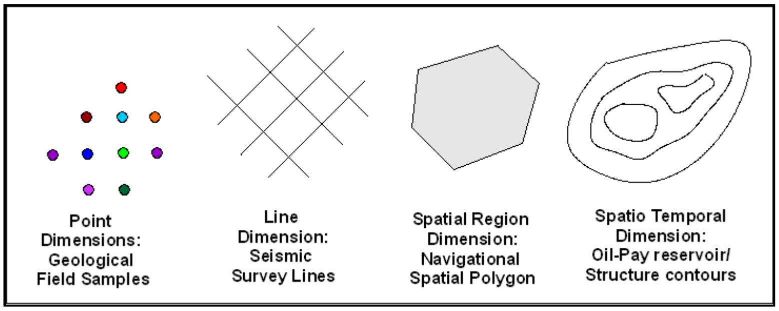

Point Dimensions: Geological Field Samples
Line Dimension: Seismic Survey Lines
Spatial Region Dimension: Navigational Spatial Polygon
Spatio Temporal Dimension: Oil-Pay reservoir/ Structure contours

Figure 4.30: Process of representing spatial views of geometrical elements of oil and gas data

The last two types of spatial dimensions of geo-scientific data indicate geometric nature that have more than one way of being generalized to high-level concepts. The generalized concepts are geometric, such as maps representing larger regions, or non-geometric, such as named areas or general descriptions of the regions. These are used as alternate ways to go from fine granularity to coarser granularity, even within the same spatial dimension. Author presents the logical entity-relationship models for building multidimensional logical models and converting them into object oriented data models.

After having constructed the logical ER models, DB users check for performance and programming of database structures (Ambler 2001). But *point*, *line* and *area* objects narrated in the *navigational* sub class are difficult to handle by traditional ER modelling. *Structure region* from *seismic* objects, *oil-water/gas-oil contact region* from *well-logging* object and *net oil-pay region* (reservoir thickness) from *reservoir* objects are difficult to handle by ER traditional modelling. Accessing these *exploration* data, depends on appropriate storing and manipulating these types of data objects.

Universal Modified Language (UML), as described in Shanks et al. (2004), the author interprets all the previously defined entities or dimensions as objects or class objects in object oriented modeling domain. Exploration is a super- class object. Figure 4.31 demonstrates *seismic* and *wells* sub- class objects in exploration super class, how they are modelled describing the relationships during logical and implementation stages. Similar sub-class objects can be identified and interpreted from drilling and production super class objects and modelled logically to implement them in an object data warehouse environment.

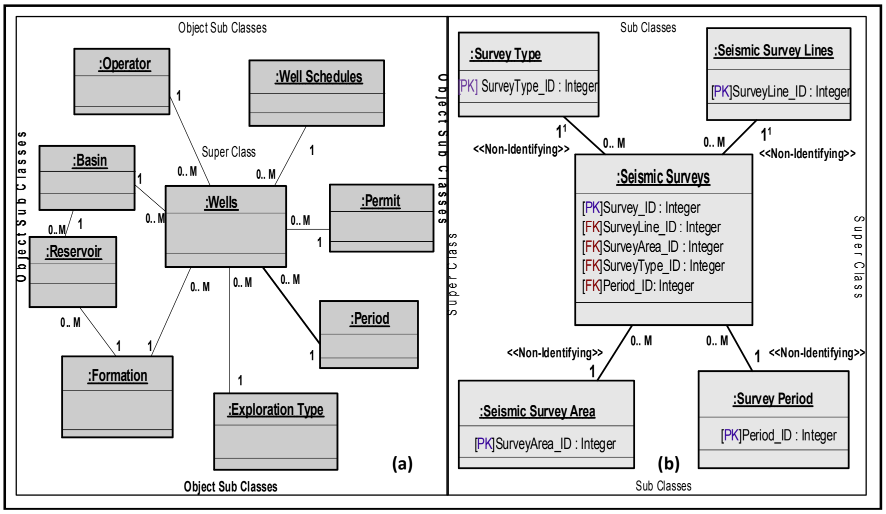


Figure 4.31: Wells data – an object schema model

In this project work, the author imports data from various programming applications such as MS Excel and Access. Data cleaning and loading of data are key operations done before data reach storage devices. When once data are loaded in the form of relational data structures in an Oracle environment, SQL queries are run for accessing data views for interpretation and analysis. All the necessary ontology descriptions used in modelling the heterogeneous data sources may be connected to the semantic web environment, through which data are allowed to share and reuse among multiple domains and semantic web applications. These ontologies are in align with agreement on common vocabularies and terminologies. As a matter of fact, for complex applications such as petroleum digital ecosystems, complex ontologies with complex reasoning are needed. The current application warrantees critical requirement analysis and goals of application. All the schemas generated in the current research in the oil & gas domain support the data interchange on web through the Resource Description Framework (RDF). The structured data can be mixed, exposed and shared among different domains and applications. The high level ontologies expressed in various graphical schematic views in Chapters 3, 4 and 5, are in conformation with RDF, which are easy –to-understand through visual explanations. All the queries (SPARQL) used in extracting the data views from a relational database of an integrated framework (metadata) can be linked with the RDF store. The author is of the opinion that the data views are based on certain focused ontologies and or on sets of rules. The schemas and artifacts built within the purview of the design science framework and guidelines are used and reused in various domains and applications for evaluations.

The models discussed so far, include modules for *well*, *lease*, *seismic*, *culture*, *asset* and *expenditure* data for petroleum exploration. As an example, the comprehensive well data module includes tables for *formation-tops*, *cores*, *lab* and *field tests*, *production*, *geophysical logs*, *deviation surveys* and *well locations*. The seismic module uses records of acquisition and processing history of seismic lines along with the location data. Database is populated with existing data quickly using text, ASCII and SQL and control files (for data loading).

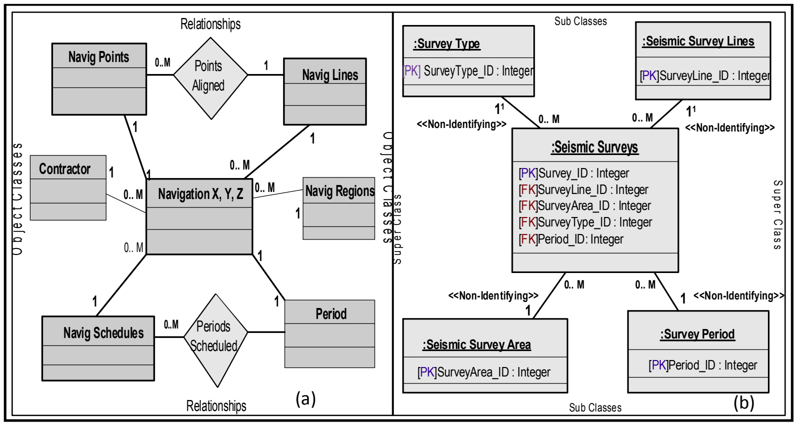


Figure 4.32: Class diagrams for (a) *navigation* and (b) *seismic* objects

As shown in Figures 4.32 and 4.33, the diamond at one end of the relationship between *basin* and *survey lines* is not hollow, but solid. A solid diamond represents a stronger form of aggregation called composite aggregation. In composition, a part object belongs to only one whole object; for example, a *survey line* is part of only that particular basin that is under study. Therefore multiplicity on the aggregate end may not exceed one. The author generates parts after creation of the whole object; for example, more *survey lines* may be added to the existing basin for investigation. However, once a part of the composition is created, it lives and dies with the whole. Deletion of the aggregate object cascades to its components. However it is possible to delete a survey line before it's aggregate in the basin dies.

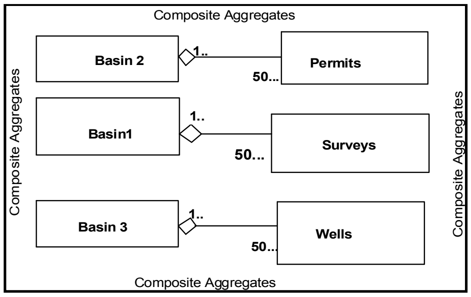


Figure 4.33: Representation of the composite aggregates of petro2 -*surveys* database

Representing an aggregated exploration entity as spatial object

So far in the ER modeling, the author discusses types of data and measures involved in a typical petroleum industry. As stated earlier, in the oil and gas industry, the data types are in the form of numeric and or geometric measures. An object in the spatial data modeling is equivalent to an entity or dimension that has well-defined role in the application domain as well as state, behavior and identity. An object (Hoffer et al. 2005) could be a concept or abstract, or thing that makes sense in an application context. Entities in the ER modeling are represented as objects in the object model, if one wishes to model spatial objects such as *survey lines* or *well locations* or *gravity contours* etc… In addition to storing information, an object also exhibits its behavior, through operations that examine or affects its state.

As an example, a petroleum company is an aggregate object consisting of several component objects. *Navigation, geology, geophysics, reservoir engineering, well logging* are other component objects of the aggregate exploration object. Similar objects can be identified in other aggregated objects such as *drilling, production and marketing* of the oil and gas industry.

An exploration object is an aggregation of several component objects. Note that aggregation involves a set of distinct object instances, one of which contains or is composed of the others. It is a stronger form of association relationships and is represented with a hollow diamond at the aggregate end. The component objects can be added and even if one component is missing, still the aggregated object still survives.

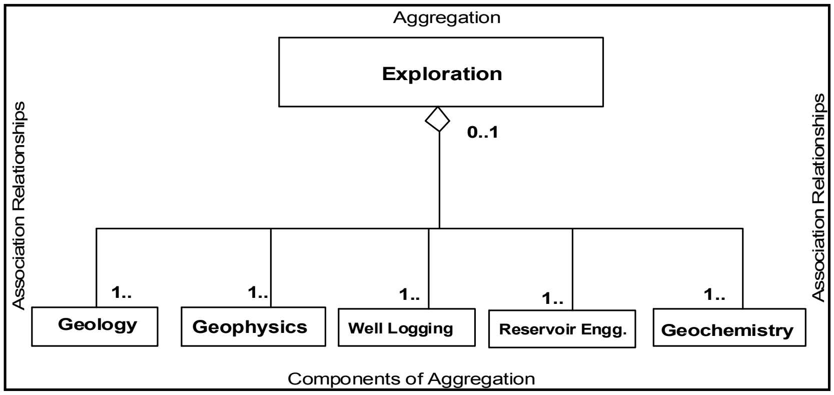


Figure 4.34: Representation of aggregated *exploration* object

Each of the component objects of exploration aggregate object has measures of numeric and spatial data. As discussed earlier, author processes data from each of the object component and stores in the oil and gas data warehouse. These data must have been pre- or post-computed and stored in the storage components of the warehouse. As shown in Figure 4.34, each component of the aggregated object has multiple sub-components, such as *geology*, *geophysics*, *well logging*, *reservoir* and *geochemistry* possess multiple entities or dimensions indicating the types of operations and their corresponding data types.

4.2.3 Business data dimensions

As demonstrated in (Marakas 2003), the author makes a comparison between entity-relational modelling (ERM) and multidimensional modelling (MM). The goal of ERM is to remove all redundancy in the data and create a set of stable relationships that can easily be translated into a physical set of tables or databases. MM is a logical data design technique in which several data dimensions from various domains explore connections, provides an access to data warehouse (DW) and its associated data-marts. The MM approach differs significantly from the ERM. Every dimensional model consists of three elements: fact table, dimensional tables and star joins between fact tables surrounded by dimensional tables. The key difference between ER models and dimensional models lies in their relationships. This relationship is understood with single ERM that is broken down into multiple fact tables with multiple dimensional tables, as already narrated in Chapter 3. The ERM appears to be so complex and unmanageable for data query purposes and often represents modelling with several processes and business rules, which sometimes become more complex and expensive to operate. In contrast, the dimensional model characterizes with a single business process. As such, in a data warehouse environment where the data are organized into a set of star joins via dimensional modelling; a single fact table is queried independently from all others with a simple query. The author draws a dimensional model (Figure 4.35) for a *basin* with number of dimensions, depicted around a basin *fact* table.

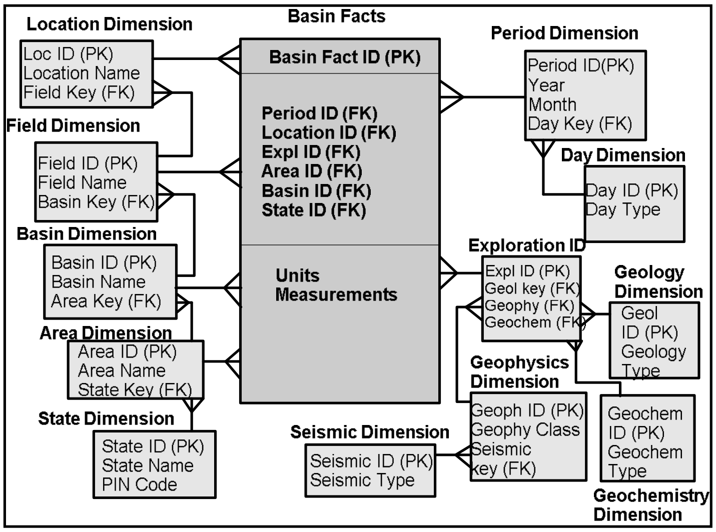


Figure 4.35: A typical multidimensional star schema drawn for a *basin*

A data warehouse stores facts, such as unit of production, production rate, mineral discoveries, exploration costs, crude exports etc. In addition, the author creates calculated measures by specifying mathematical formulas such as Ratio_ofActual and ExpenditureCosts, differences between exports and imports of petroleum products, NPV and least square curve fitting, second order parabola and regression statistical formulae. Author creates measures from stored facts as well as from other measures. For example, the measure ‘volume change’ is calculated by subtracting a prior period volume from the current period volume, yields a second measure, ‘percentage change’. Facts and measures change over time. Multidimensionality is demonstrated that includes a period key, as shown in the Figure 4.36a, linking with other multiple dimensions of Petro2-permits database. Time or period, is the one of the database dimensions, common to every data warehouse and is the most difficult for many analysis tools to handle because of the infinite number of possible time period combinations. The time periods considered in the present study are months, quarterly and yearly accounting periods. At places half-yearly measures have been reported and analysed. In the present scenario, there are multiple time periods in different databases organized. The data warehouse considers in each situation, as a single database. A single dimension analysis allows users to retrieve facts and calculate measures for one or more time periods, including totalling a group of periods.

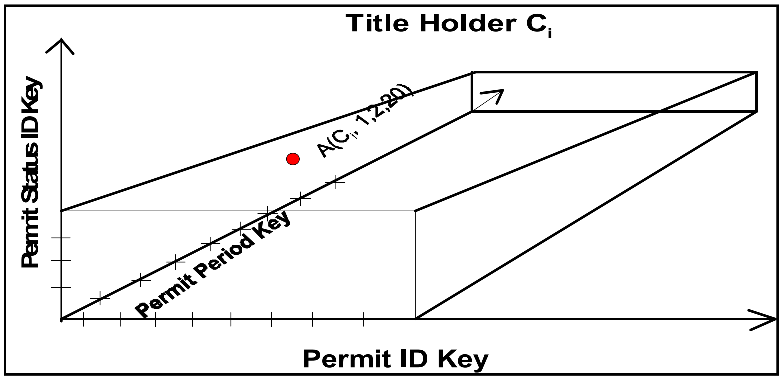


Figure 4.36a: Multidimensionality of the star schema for petro2-*permits* database

In an exploration aggregation entity, spatial or coordinate data dimension has significance in the oil and gas data computations, since various types of geometrical and non-geometrical data are generally referred to earthly positions and thus make up geo-scientific databases. When building a geo-scientific data warehouse with a multidimensional approach, one may have to consider three types of dimensions (in multidimensional sense), according to the theory of scale measurement (nominal, ordinal, interval and ratio scales where each scale allows for richer analysis than its precedent one (Miller and Han 2001). Each type of dimension is considered that deals with a geometric spatial reference such as X, Y coordinate systems (i.e. quantitative data of the interval and ratio scales), with a semantic spatial reference such as place names (i.e. qualitative data of the nominal and ordinal scales) or with combination of both, such as survey coordinates, survey name and line, basin name; permit number and name in a particular basin (which is a combination of quantitative and qualitative data where the quantitative data can be located precisely or interpolated along the linear axis identified by the qualitative data). The type of dimension involving the geo-scientific database system supported by the warehousing/decision-support technology influences the type of spatial dimension one may use, or in other words, the type of hierarchy of a dimension:

Non-geometric spatial dimension: This dimension contains only non-geometric data. For example, “survey names, survey line numbers, well numbers, survey IDs and well IDs and permit IDs and numbers”, are used in the construct of geo-scientific warehouse. These dimensions contain only nominal data instances to locate a phenomenon in space. Such dimensions could start with the names of exploration permit names, survey line numbers and basin names, which are non-geometric such as state and country names. Such a solution is implemented as long as navigation representation is not required.

Geometric-to-non-geometric spatial dimension: An exploration aggregation component, which has several dimensions or entities, is further represented by several object elements characterizing with several patterns. This is a dimension whose prehistoric level data is geometric but whose generalization, starting at a certain high level, becomes non-geometric. For example, a survey line represented by a polygon in the Canning basin map, that is geometric data, is the finest granularity level of this spatial dimension. However, each survey line can be generalized to some value which is solely nominal, such as *Canning survey line* 1, *Canning survey line* 2 etc… and its further generalization remains nominal, thus playing a similar role to a non-geometric dimension at coarser granularities of this spatial dimension. Using such design techniques, oil and gas personnel are allowed to benefit from the generalization.

Fully geometric spatial dimension: This is a dimension whose primitive level and all of its high-level generalizations are geometric elements. For example, polygons of equi-type or value onshore survey regions data (such as equal property of gravity, magnetic or seismic survey data) or for offshore data are geometric data. It could also be polygons of equal elevations or altitude regions are geometric data, and every generalization, such as elevations covering 0-700m, 700-1000m and so on… are also geometric.

Additional database dimensions that may represent the business characteristics are unique to each enterprise. Representative dimensions and facts are identified from mining, oil and gas and mineral industry domains (Figure 4.36b). The unique requirements of multidimensional analysis begin to emerge when a user requires sub-totalling of facts and measures along multiple dimensions, including the time dimension. Many facts and most calculated measures are 'non-additive'. In one-way multidimensional analysis, software shares a similar concept with the spreadsheet. Rather than specifying cell locations in a spreadsheet to define a calculated measure, the location of data in the database is used in specifying the formula to compute a measure. In the multidimensional database, the formula requires the locations of two facts defined by the database dimensions. The power of multidimensional software is that the formula does not change regardless of what product, markets, or periods are specified. The users are able to navigate any combination of dimensions while the logic to calculate the required measure is maintained.

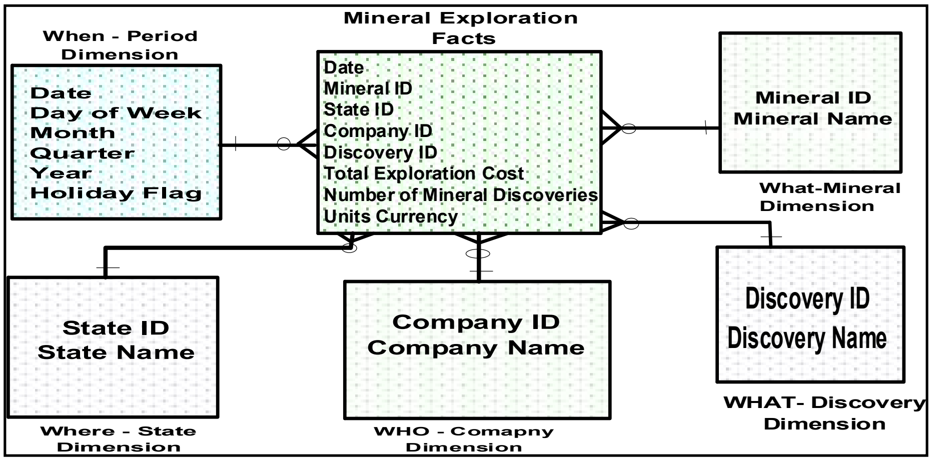


Figure 4.36b: Multidimensional star schema model for *mineral* exploration company

As narrated in the Figure 4.36b, a dimension is a collection of logically related attributes and is viewed as an axis for modelling the data. Attributes identified in the oil and gas industries are analysed. *Time* dimension is divided into many different grains, such as monthly, quarterly and yearly. The specific data stored are called facts and usually numeric data. At places facts consists of measures and contextual data, such as *million dollars cost* and *million litres of petroleum products*. Explorationists involved in the Western Australian oil and gas exploration industry provide a broad and in-depth dimensional view of the data warehouse that shields them from the complexities of the underlying data structures, query algorithms and allows them to pose questions from their business perspective. Each dimension is a part of standard hierarchies, including multiple hierarchies with a single dimension at places.

As demonstrated earlier, various business data dimensions are building blocks of an integrated business system scenario showing how complex view of business dimensions of Oil and Gas Company can simply be segregated into generalized and specialized data dimensions. Several database designs have been narrated in D'Orazio and Happel (1996) and Rob and Coronel (2004), demonstrating different types of data models in industrial scenarios. The author draws a multi-dimensional data model among the identified business sub-type dimensions. Dimensions among several oil & gas dimensions are associated with several sub-type dimensions that represent the specialized and generalized type dimensions. The "exploration" as a generalized type business dimension is further divided into smaller units as specialized dimensions. The author attempts to inter-relate the super type with sub-type dimensions as shown in Figures 4.37 and 4.38 through sharing common attributes of all dimension types. The author identifies and builds relationships between business dimensions representing their associativity. While mapping multi-dimensional-relationships, multi-dimensional models are incrementally extended further building associated relations as dimensions (MR) through extended multi-dimensional-relationship (EMR) mapping approach. The precise definitions of generalized and specialized dimensions, MR and EMR are discussed in the forthcoming sections.

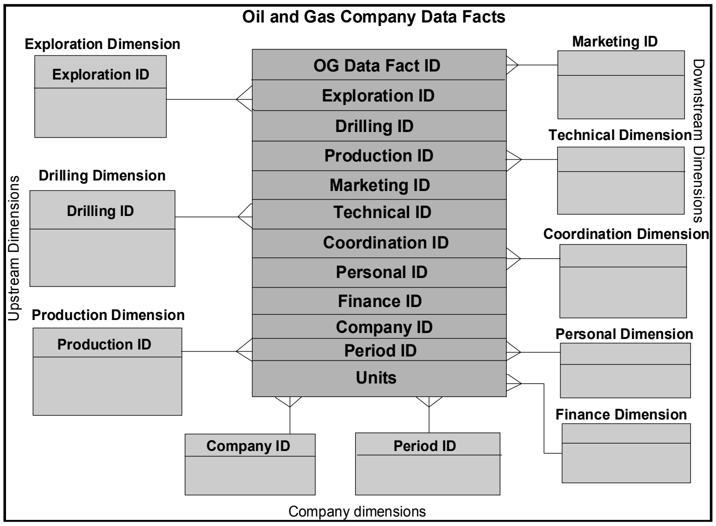


Figure 4.37: Multidimensional oil and gas business data schema representation

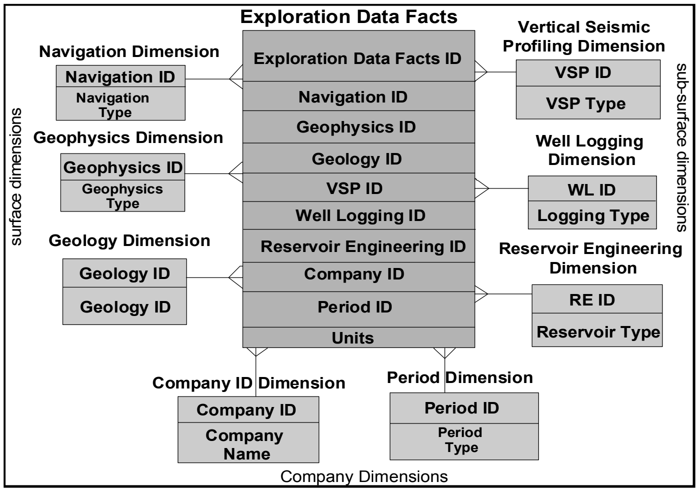


Figure 4.38: A Multidimensional relationship (MR) model for exploration

Exploration, drilling, oil and gas production and coordination businesses identified by the author, as key dimensions are briefly discussed in the following sections. The conceptual models developed designed and developed for exploration dimension; provide a scope and flexibility of extending them in an integrated company environment (Nimmagadda et al. 2005c and Nimmagadda and Rudra 2004). Since exploration is a key business super-type dimension in any oil and gas company's overall business system, how it is described as a business dimension in association with other sub-type dimensions is discussed in the following sections.

Exploration as a business dimension

The dimensions and attributes identified from exploration business are responsible for making up connections within this super-type dimension and also share common attributes with other sub type dimensions of other super-type business dimensions. The geophysics, geology, reservoir, well-logging, VSP and other research services are key functions, interpreted as inherent sub-type dimensions of an exploration business. They also share common attributes among these sub-type dimensions. Further, "geophysics" sub-type dimension possesses inherent sub-type dimensions interpreted as data acquisition, data processing and data interpretation, based on their activities. Functions and their activities within several operational units again classify and identify inherent dimensions. Again strategic planning, management control, operational control and data reporting, control each dimension or operational unit of exploration business. A manager, who is playing a key role in each of these entities, performs planning, organizing, staffing, directing and other controlling tasks.

The author rationally uses multidimensional modelling to logically map and model typical data objects. Dimensional data mapping approach (Hoffer et al. 2005, Coronel 2011 and Pratt and Adamski 2000) is one of the logical data model development techniques considered in the present study, since these are used to design exploration data warehouse, accommodating volumes of object oriented and multidimensional data structures. The author designates all sub-type classes interpreted from *navigational* and *seismic* operational data object classes as dimensions, as shown in the Figure 4.39, on top of each and every rectangular box.

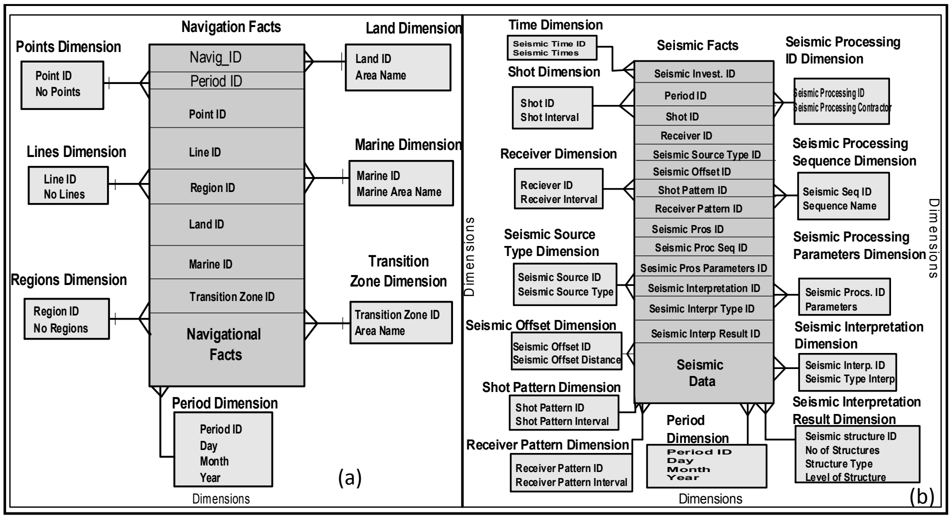


Figure 4.39: Multidimensional star schema models for (a) navigation and (b) seismic data

Primary key attributes of dimensional tables, that are common, are shared and linked with the fact tables through their corresponding foreign key attributes. This procedure is followed for all data structures, one may notice that each fact table is surrounded by related dimensional tables, linking them with foreign keys and the corresponding primary keys located in the dimension tables. One-to-many relationships are mapped in all the tables. Similar logical data models are developed using *surveys* and *navigational* dimensions. Exploration data such as *navigational, seismic*, *well-logging*, *VSP* and *reservoir* are represented dimensions with one-to-many relationships (Figure 4.40).

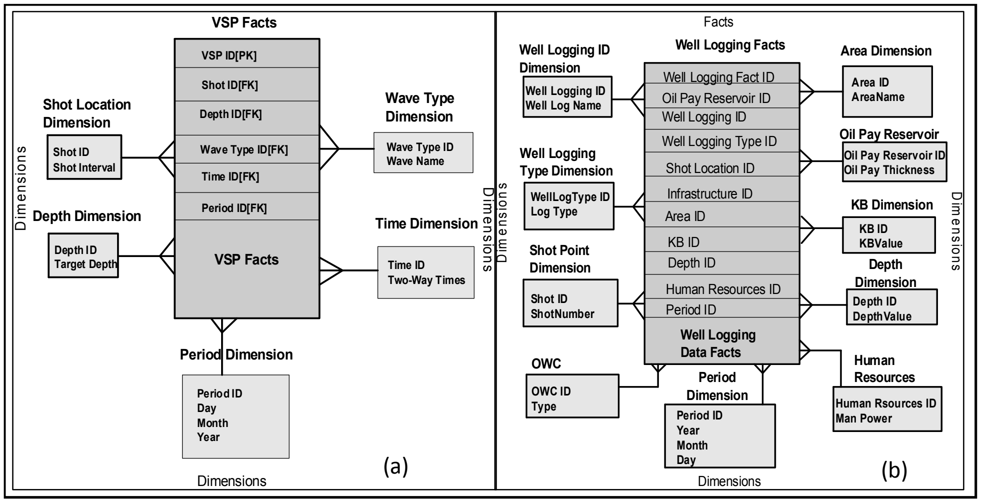


Figure 4.40: Multidimensional star schema models for (a) *VSP* and (b) *well-logging* dimensions

Drilling as a business data dimension

The next business immediately after exploration is drilling. Once a geotechnical order or an exploratory proposal plan is approved by the company's technical management, drilling infrastructure and associated logistics are mobilized to an approved site either for exploratory, development, step-out or delineation well type from which data analyst identifies all the associated dimensions. Other dimensions associated with strategic planning, management control, operational control and data reporting are also identified. *The number of wells*, *type of well*, and *categories of well* that are decided depending upon the size of prospect, are key attributes. Type of rig, time involved in drilling a well, targeted depth, meterage to be drilled, cycle speed and other budgetary constraints are identified.

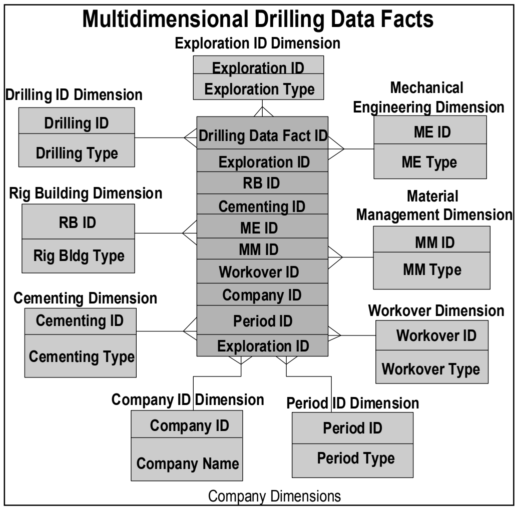


Figure 4.41: Multidimensional relationship (MR) model for *drilling* business dimension

A conceptual model is generated using dimensions and their relationship with associated attributes of drilling business entity as shown in Figure 4.41. There are common attributes with other business data dimensions' attributes, such as exploration and production business, which need to be included in the conceptual model. Data pertained to engineering, cementing, chemistry and variety of chemicals used, safety, rig building, personnel and administration, finance and accounting have to be rigorously be analyzed. Various other engineering activities associated with drilling business such as mechanical, civil, electronics and telecommunications and electrical engineering may have to be involved in the conceptual data models. These dimensions and their associated attributes are also present in the technical business subsystem. When once, the drilled target depth is reached, the contractor can conduct the logging surveys within the drilled-well including the vertical seismic profiling (VSP) surveys. Again all the planning, management and operational control activities associated with the conduct of these valuable surveys will have to be systematically analyzed in the drilling business subsystem. Shot location, latitude, longitude and elevation of the well site are key attributes that will be associated with exploration business subsystem as well. All dimensions and attributes have to be involved with conceptual MR modelling (Figures 4.39 – 4.43) and mapping is done through one-to-one and one-to-many and many-to-many relationships. Business rules are imposed wherever necessary as per constraints analyzed.

***Production* as a business dimension**

After having explored and drilled the prospect, if the drilled well proves to be oil bearing, the next step is production operation to be carried out for extracting oil and or gas from target depths. Production engineering, exploitation of oil and gas from *drilled well*, *oil dispatch*, *refinery supply, LPG production*, *wells worked over*, *the flow lines* laid to transfer oil and gas to group gathering stations are key dimensions of the production business. The attributes associated with these dimensions are identified, for not only associating among sub-types of production business dimension, but also with other super-type such as, exploration and drilling business dimensions.

Using the dimensions approach, the author identifies and examines wells and geological formations tested, type of activation of sick well, secondary recovery procedure method attributes, besides, shot location, latitude, longitude and elevation of the current producing well data. Men, machinery money and other constraints, are identified as other dimensions with corresponding attributes. The shot location, latitude, longitude and elevation are key data attributes that can be associated with other business subsystems. The author creates a conceptual model involving these dimensions and relationships as shown in the Figure 4.42. Data mapping is done through one-to-one and one-to-many relationships among these dimensions.

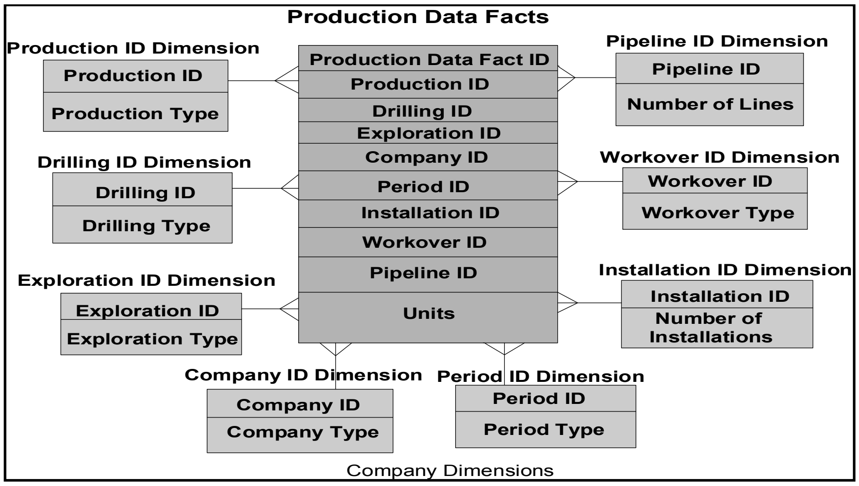


Figure 4.42: Multidimensional relationship model of *production* business data dimensions

Technical as a business data dimension

The technical business manager generally focuses on civil, electrical, mechanical, transport, inspection, quality control, safety and environment, establishment task force disposals and other combined activities pertained to other subsystems such as exploration, drilling and production businesses.

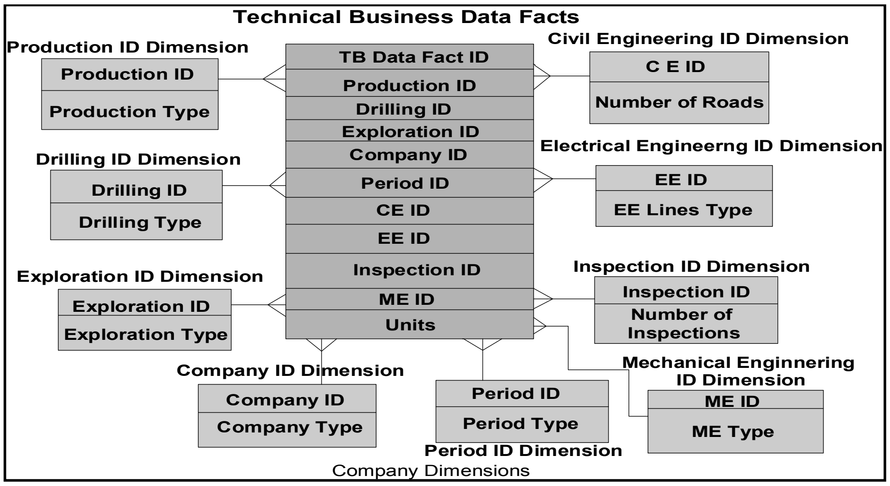


Figure 4.43: Multidimensional relationship (MR) model for *technical* business dimension

The data relevant to mechanical, electrical, logistics, transport, tele-communications, civil, safety and environment, inspection, energy conservation, technology up gradation, import substitution, cost reduction engineering and inventory are key dimensions of this business subsystem. The author draws a conceptual MR model with data mapping of one-to-one, one-to-many and many-to-many relationships among different dimensions of this business subsystem as shown in Figure 4.43. MR model stands for multidimensional relationship model.

Coordination and collaboration as business data dimension

In general, the manager of the coordination business deals with personnel and administration and finance and accounts aspects of the oil and gas company. Manpower planning, selection and recruitment, training and development of staff, welfare, security, industrial relations, sports, and public relations are key activities and also dimensions of this business subsystem. In addition, financial performances and sales turn over, plan and non-plan expenditures, material-in-transit, disposal of and distribution of materials and material procurement are other dimensions of the coordination business. Similar MR models can be drawn for this business. In addition, the author discusses MR mapping concepts and their practice in the following sections.

Multidimensional data mapping and modelling

The graphical representation of multidimensional-relationship mapping is called MR modelling. As shown in the Figures 4.38 – 4.43, all the dimensions represented within a business dimension are either associated with one-to-one, one-to-many, and or many-to-many relations. It is not necessary all the relationships exist in all data structures that are mapped. The data analyst is responsible for identifying these relationships and mapping the dimensions appropriately, so that conceptual models generated, truly represent the logical data models for future implementation. Dimensions, attributes and relationships are key constructs of these MR models. The author further discusses how the specialized and generalized dimensions concepts applied in oil & gas conceptual MR modelling in the following sections.

Representing specialization and generalization in multidimensional data mapping

As discussed in Hoffer et al. (2005) and Coronel (2011), the author uses super-type/subtype relationships to represent the dimensions of oil and gas company's data models. The generalization and specialization are key descriptions to use in examining the super-type and sub-type relationships.

Generalization dimensionality

In the data modelling, the generalization is the process of defining a more general dimension type from a set of more specialized dimension types. It is a bottom up process. Examples of such generalized dimensions are shown in Figures 4.38 – 4.43. The exploration, drilling, production and technical business dimensions are sub-type dimensions of super-type oil and gas company generalized system.

Specialization dimensionality

Unlike generalization, the specialization is top-down process in which, the author examines one or more subtypes of the super-type dimensions and formation of super-type/subtype dimension relationships. As shown in the Figure 4.44, an entity type named, 'drilled well' has several attributes. An attribute called, formation is multi-valued, because there may be more than one formation with an associated *formation-name* with its ID. The drilled-well is generalized with specialized exploratory or development subtype dimensions. The MR conceptual model suggests that *formation* (geological) and its ID are also associated with specialized subtype dimensions. A new relationship is emerged because of these associated relationships between dimensions.

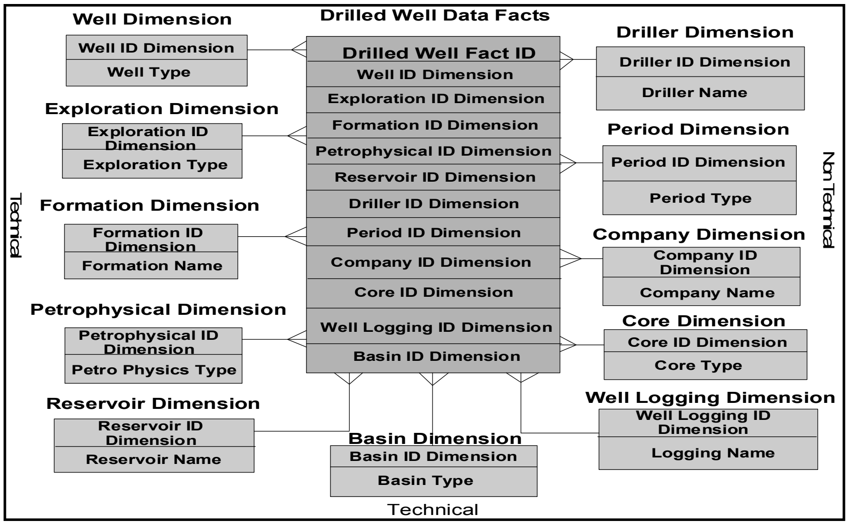


Figure 4.44: Multidimensional structure of *drilled-well* dimension

These two processes are key to describe relationships between super-type and sub-type dimensions. The data analyst perceives the combined usage of these processes in the development of conceptual dimensional data models. The author ultimately interprets several constraints of specifying super type and sub type dimensions (Hoffer et al. 2005) for logically implementing them from generalization to specialization or vice versa approach. Schemas and sub-schemas are described, how they are integrated in a company situation, are given in the following sections.

4.3 Exploration Data Integration and Exploration Metadata

The schemas and sub-schemas are integrated after their validation. The process of integration and critical factors associated with development of integration process are narrated in the following sections.

4.3.1 Schema integration

A conceptual schema that describes an enterprise of an oil and gas company is the result of an association, bridging various functional sub-systems or divisions of exploration business. It is conceivable to have each functional sub-system (*exploration* as an example) developing its own schema. When schemas of different oil and gas data entities or objects are disjointed, a union operation is invoked, constructing the overall conceptual schema. However, schemas overlap at several points and it is quite a challenge to draw semantic boundaries between subsystems of a complex nature such as found in the oil and gas industry, with system-wide functional interactions. Components of centralized information systems are described and each part, referred to as a schema, corresponds to a part of the overall conceptual schema which is called a view. Only when individual views are put together, or integrated, an overall conceptual schema is obtained. Diversity in semantics causes conflicts and variations in modeling, which have to be taken care of systematically. Distributed information systems add a further dimension to the problem. Integration of views is different from database integration where databases of all individual schemas of oil and gas data items are merged centrally, or combined in a distributed database by constructing schemas of schemas, i.e. global schemas. View integration is at a lower level; views are combined into a conceptual schema that represents a database.

4.3.2 Reasons for integration in the exploration industry

The data found in the oil and gas industry are often very diverse, multidimensional and heterogeneous. A methodology for data view integration is necessary (Ozkarahan

1990) because it is impossible for people working independently (with individual operations) on data modeling to consistently arrive at the same representation with similar concepts. People have different viewpoints in perceiving data semantics. The richer the abstract model in providing alternatives in representation, the more diverse, is the views explaining the same concepts. A given concept with different names, possibly causing naming conflicts as well as other problems, can be represented with different types. Moreover, incomplete information in the conflicting parts gives rise to inconsistencies, such as differing cardinalities for the same entity or relationship. This often happens in the case of heterogeneous data, where similar data attributes and corresponding values have a common influence in the different operational environments. In addition, the semantics and naming conventions of data attributes, integration of these attributes of different company entities (with semantics applied), can facilitate the logical and physical data organization of heterogeneous petroleum oil and gas data as a metadata, ultimately permitting the building of knowledge of petroleum systems previously hidden among the data.

4.3.3 Relating the data views from multiple domains

The data views contain anything from totally disjoint events to closely related events. When two views are brought together, enlarging the semantic context, the following possible interactions are discovered:

- Inter-schema connections or inter-schema relationships. A geologist object class in one schema and an exploration object in the other, when merged together result in “explorationist” relationship between the geologist and exploration objects (or entities);
- Common parts of the views may be found to be identical; in which case, the merged schema contains a single copy of the identical representation;
- Common parts modeled by two views are not identical but equivalent to each other. In other words, one view can be mapped to the other by some algorithmic transformation, such as when the same concept is an entity in one view and an attribute in the other. Depending on the direction of transformation, the views are merged after one representation is converted to the other.

4.3.4 Integration of the data relationships

In this process, the author combines N views by pair-wise merging at one extreme or by an all-at-once n-ary merge at the other extreme. In either case, author examines the schemas (or views) in a pre-integration process to determine the degree of conformity, and for conflicts among the views. When schemas are compared, the following are considered:

- *Naming*: Problems will be discovered due to homonyms and synonyms in the industrial data. An example of a homonym is the use of the same name for different concepts. Such as the use of "*exploration*" as an entity in the "*geologist*-<qualified>-*exploration*" schema and also as an entity in the schema "driller-<belongs>-*exploration* (angle brackets denoting relationships). An example of a synonym is that between the schemas "*employee*-<*assignedto*>-*exploration* and "*surveyor*-<*conducts*>-*survey*," the same concept being described by two different names, "*exploration*" and "*survey*";
- *Attribute Correspondences:* Industrial data such as exploration, drilling, mining or production, possess different attributes and their correspondences. The same attributes between schemas may have different data-types (integer vs. real), units etc., which must be converted to a common form. If the keys and underlying domains of attributes are identical, they can be unified with a union operation. Another possibility is containment between sets of attributes. For example, the domain of "*engineers* (or *geologist*)" is a subset of "technical personnel" or "surveyor". Integrating the schemas containing these domains, one per schema, yields a generalization hierarchy between the two entity-sets in the integrated schema. If the underlying domains of the attributes are different, however, the decision rests with the data analyst. The data analyst may choose to create a generalization hierarchy under "*employee*" for *Secretary* (*Eno, Name, Sal, Type-speed*) and *Engineer* (*Eno, Name, Sal, Degree*) even though the domains of engineer and secretary do not intersect. The case *explorationist* (*Dept, Name, Age*) and *secretary* (*Dept, Name, Age*) are not integrable, despite these entities having common attributes;
- *Structural Correspondence:* These may involve conflicts in types, keys, types of relationships, etc. Type conflicts arise when different types are chosen for the same concept in different schemas. For example, *Geologist* (*ID, Ssn, Name*)-<worksin>-*Dept* is used in one schema, yet *Geologist* (*Ssn, ID, Name, Dept*) is in another schema. *Dept* is an entity in the first schema, and the *Geologist*'s connection to it is represented by the <working> relationship. In the other more compact representation, *Dept* is expressed as an attribute of the *Geologist* entity. In key conflicts, the same concept may be represented with different primary keys in different schemas, such as the case of *Geologist*

having the primary key ID in one schema yet *Ssn* in the other, as shown. Where relationships conflict, for example, a relationship "*Geologist*-(n)<worksin>(1)-*Dept*" can be represented as a weak relationship in one schema (where n, 1 are mapping constraints), making the existence of *employee* dependent on the existence of *Geologist* dependent on the existence of *Department*. Yet, in another schema, the relationship can be a (strong) relationship focused on the department side, meaning that there cannot be any department without any geologists assigned to it.

The objects (Hoffer et al. 2005, Coronel 2011 and Huynh et al. 2000) under different schemas are similar due to their key and domain similarities. This similarity ranges from identical to a commonality in some domain containment relationships, and to totally dissimilar with disjoint domains (object instances are not common despite being conceptually alike). In the case of identical objects, a single copy of schema and union of the attributes (some may differ) and instances are kept. For example, two different functional units of an enterprise may keep the object "*employee*" between the schemas. Even though the two representations mostly share the same domains, there may be attributes unique to individual schemas such as between *Driller* (*Ssn, Name, Sal, Commission*) and *Driver* (*Ssn, Name, Sal, Overtime*), where commission and overtime are schema-specific. Yet the newly integrated schema contains the concept *Employee* over the entire enterprise as *Employee* (*Ssn, Name, Sal, Commission, Overtime*), which may introduce the use of 'nulls'.

In all the other cases of object integration, the major contribution of integration is identifying generalization hierarchies among similar and dissimilar object classes. As mentioned in the attribute correspondences, the examples for these are as follows: for similar objects, the integration between *surveyor*, and *geologist*; for dissimilar objects, *explorationist* and *secretary* in one case and *driller* and *driver* in the other.

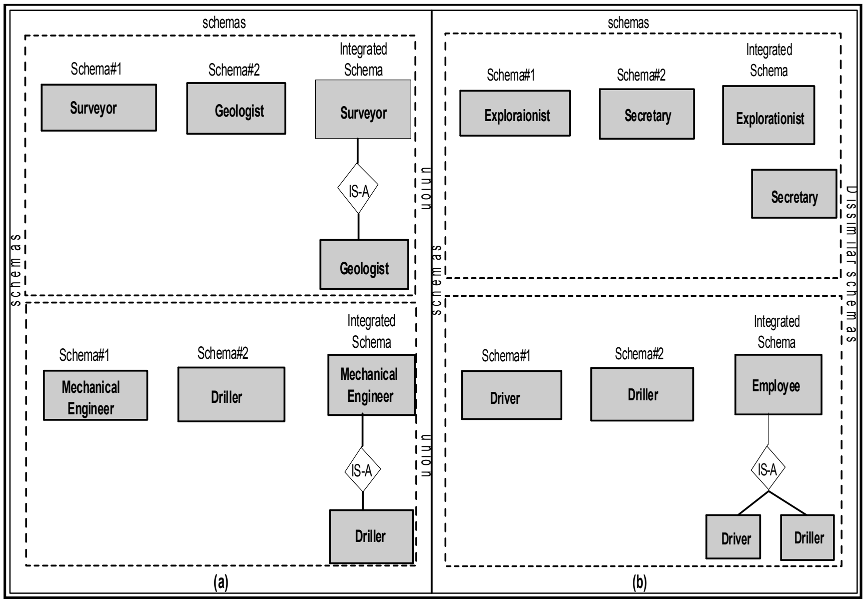


Figure 4.45: Objects integration; (a) similar objects and (b) dissimilar objects

As mentioned earlier, dissimilarity here is in the un-commonality of instances, i.e. an *explorationist* is not a *secretary* even though they may not be described similarly. A *driller* is not a *driver*, but in this case there is an opportunity for them to be integrated into a generalization hierarchy as shown in Figure 4.45. However, there is a difference between the IS-A hierarchies in the sense of domain containment. In *Explorationist*, an inclusive union has to be made (of generalization), whereas in *Employee*, one can have an exclusive union due to the common and disjoint domains of the former and the latter, respectively.

Integration of relationships is the next most difficult type of process. Various aspects must be considered, such as the degree of a relationship, roles of entity sets or objects participating in the relationship, and structural features such as type of relationship and mapping constraints. The degree of the relationship refers to the number of entities a relationship involves. For example, a *Driller-WellSite-Rig* relationship involves the entity sets *driller*, *rig*, and *well-site* and therefore has a degree of 3. When two relationships are compared for conformity, there may be many combinations among the variables considered. Two relationships can be exactly identical in degree, roles of participating objects, and structural features, or there may exist large differences. In between, there may be several partial matches.

Two relationships with respect to two views (schemas) are highlighted in Figure 4.45(a). Both of these relationships share the same degree, roles, and structural features; hence they are identical. Therefore, in the integrated schema, only one relationship is placed, as shown in Figure 4.45(b). The domains of entity-types *technical personnel* and *Engineer* are related by containment. The integrated schema is combined in a generalization hierarchy, under *driller* which is more generic. However, this generalization hierarchy would overlap in its specializations. When relationships differ in degree, these may be merge-able if they correspond to each other conceptually and one is a more detailed view of the other. Putting it differently, the relationship of lower degree must be derivable from the relationship of the higher degree by projecting or creating a subset of the latter.

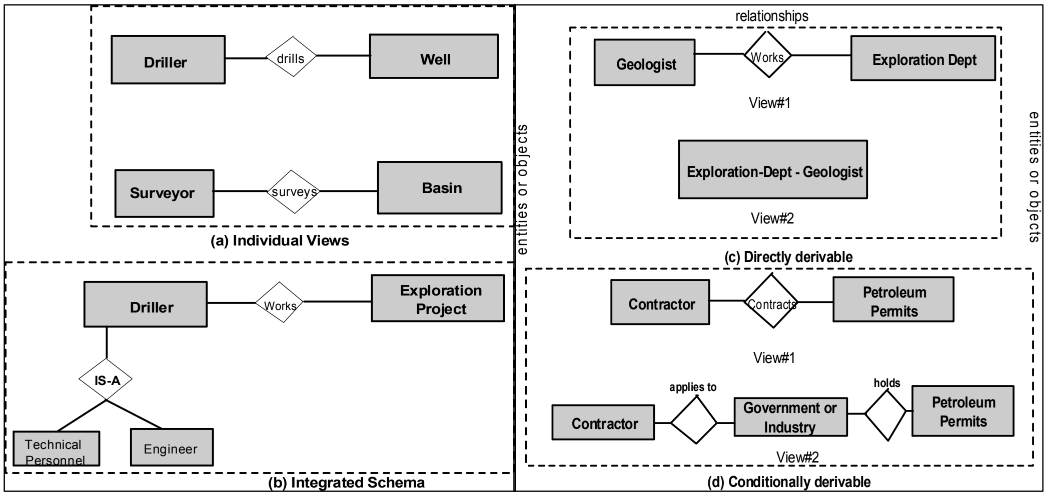


Figure 4.46: Identical relationships (a) Individual views (b) integrated schema; relationships of different degrees (c) directly derivable (d) conditionally derivable

In integrating such schemas, the one belonging to higher degree and representing both schemas, is retained. Figure 4.46 shows two views that correspond to relationships of differing degrees again. Figure 4.46 shows directly derivable relationship.

View#1 in Figure 4.46c is a relationship of degree 2, whereas View#2 represents the whole entity (i.e., in a way, a relationship of degree 1) and is unrestricted in mapping, as opposed to the 1:n mapping constraint of View#2. Therefore, these views are equivalent. View#2 is derived from View#1 directly. The integrated schema is represented by View#1, i.e., the relationship is of higher degree. In Figure 4.46(d), View#2 represents the integrated schema if View#1 is derived from it. For this to happen, all the attributes of the *contracts* relationship must be contained in *applies_to*, *government/industry* and *holds*. This case represents a conditional derivability such that if the condition holds, then the schemas are integrable. Figure 4.47 shows an extreme case where even though the entity sets are common to the relationships, their roles, semantics, and degrees are different; hence their views are un-integrable. Notice the roles explicitly added for clarity on the connections to the ternary relationship.

Figure 4.47 depicts the integration process among these concepts that are placed between the two schemas.

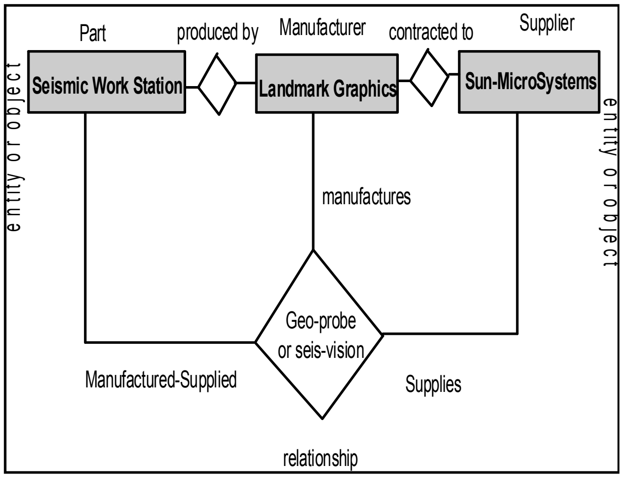


Figure 4.47: Un-integrable relationships

So far, different data schemas and their integration are discussed. The author attempts to implement them in one of petroleum exploration cases.

4.3.5 An exploration ontology implementation case study (RO1 – RO7 focus)

As per the research questions and objectives (RQ1 – RQ5 and RO1 – RO5 in Sections 1.3.1 and 1.3.2 in Chapter 1), the conceptual models and integrated methodological framework, described in Figures 3.2 and 3.36 - 3.38, the author discusses a case study justifying the models and methodologies. There are numerous data entities and attributes used in the modelling and integrated methodologies in the exploration case study. ER models constructed for surveys, wells and permits along with other exploration data such as seismic, reservoir parameters and interpretation inputs are used for implementation. They are briefly described in the following sections.

Data type descriptions

The author explains the need for warehousing and data mining technologies in the petroleum companies in (Nimmagadda and Dreher 2006b and Nimmagadda and Rudra 2004b). Relational and hierarchical data structures are popularly used, conceptualizing all the data entities and their relationships. Petroleum data from several heterogeneous sources (Hoffer et al. 2005) are conceptualized including their relationships. These are intelligently stored in a warehousing environment. Entities are used in conceptual ER modeling and objects in object oriented data modeling approaches. Analogous to entities and objects, the author uses dimensions conveniently in multidimensional data structuring approach. In the past, petroleum companies have typically stored data consisting of only text and numbers, but today, graphics, drawings, photographs, video, sound, voice mails, spreadsheets and other complex objects are stored. Relational database management systems store these data objects and types with certain limitations. The concept of object, which is the core of all OO systems, is some unit of data along with actions, affecting its behavior (Hoffer et al. 2005). A *reservoir* object, for example, could consist of the data relevant to *wells* object (*reservoir* name, type, quality and *production rate*) together with the actions that can take place on *reservoir* object (multi-reservoirs, predicting qualities, *reservoir* extents and thickness for computing the *geological* and *recoverable* reserves from petroleum prospects).

Figure 4.48 shows two schemas that convey semantics about contractors and their petroleum permits. A problem is discovered with critical examination of two views. They refer to the same concept and *contractor*, *government* or *upstream_company* is selected as the common name. Further, a structural conflict in *contractor* is observed; while it is an entity in View#2, it is used as an attribute in View#1. Because integration retains the higher degree relationship and View#1 is derivable from View#2, a change can be made in View#1 for conformity, by adding an entity set for contractor and tying it to the *licensed area for exploration* with the applicable relationship while deleting the attribute contractor. The result is shown in Figure 4.48(b). Now the schemas conform with each other conceptually, they are merged as shown in Figure 4.49. As can be seen, identical objects *contractor* and *government/upstream_company* are copied only once and connected to the other entities with their appropriate relationships.

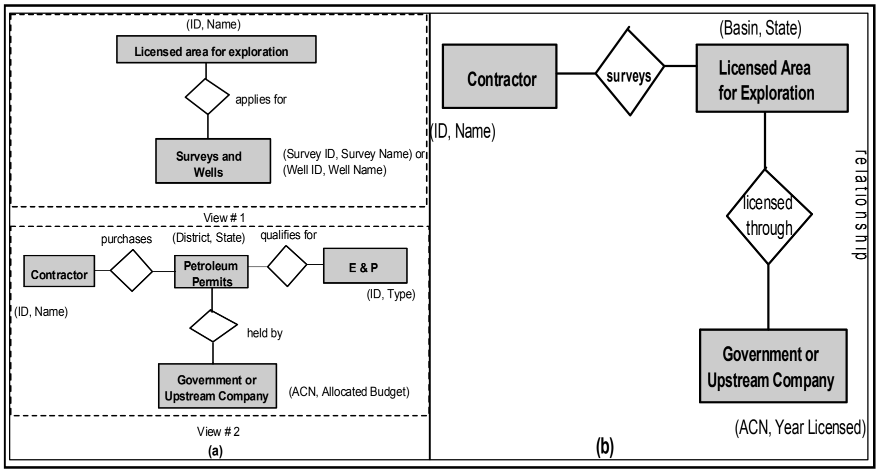


Figure 4.48: Schemas (a) original schemas (b) conforming views

These relationships however are duplicates, and their simplification can only be possible with a transformation on the merged schemas. Such a transformation is possible since *licensed area for exploration* and *petroleum permits* have common domains and are related to each other through containment. That is, *petroleum permit* is a subset of a licensed area for exploration. This is shown by an IS-A hierarchy. The *permit* inherits all the relationships to its superset, *licensed area for exploration*. Figure 4.49(b) shows the transformation and the resulting elimination of the relationships in the integrated schema. Fine-grained refinement of data schemas, combined with data integration process is effective in knowledge mapping (Nimmagadda and Rudra 2004b). Briefly, the refinement of schemas is discussed in the following sections.

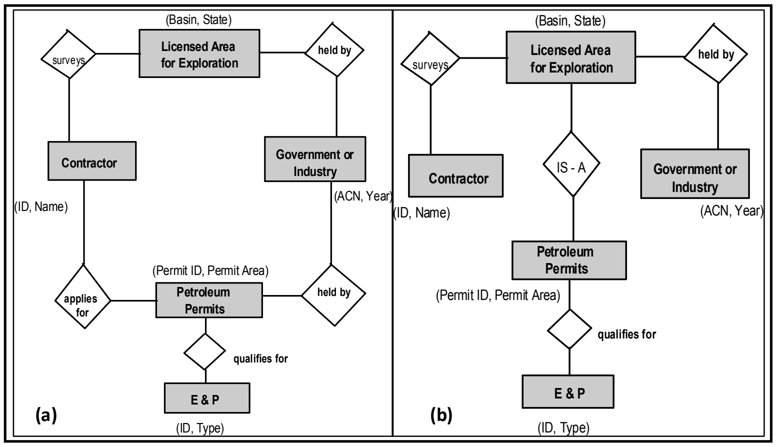


Figure 4.49: Schemas (a) merged schemas (b) totally integrable schemas

Conceptual schema refinement

As shown in Figure 3.2, the author specifies conceptual schemas in an enterprise description, and refines until it meets the criteria of a legal conceptual schema. The relational schema, which is developed based on a relational theory, is an integral target of the conceptual schema. Here, a manual refinement procedure is given citing an example from the oil and gas industry. For deriving a legal schema, the author demonstrates possible refinements to the conceptual schema. This refinement process involves design, information requirements analysis, enterprise description, transaction analysis, schema analysis, normalization and lossless joins.

A Manual Refinement Process: Manual refinement process is discussed in this section.

Design Topic

The design topic is a surveyor-activity information system. A subset of activities is encompassed which occurs between surveyors and members of the oil and gas company. The project can be divided into the following steps:

1. *Requirements Analysis:* Acquire data in order to describe what is needed and desired by the user;
2. *Enterprise Description:* Begin a rough conceptualization of the model;
3. *Schema Analysis:* Manually refine the system;
4. *Database Description:* Populate the database, apply security controls, and execute queries.

Requirements Analysis: In the Oil and Gas Exploration Data case, data available from several survey documents in multimedia format are gathered for modeling purposes (Tables 2 and 3). Author considers the enterprise to be the *surveyor* community with the following sources of information:

Table 2: Documents narrating survey data

Documents in the survey enterprise	*Activities in the survey enterprise*
1. All previous technical and financial reports 2. Maps and geological cross sections and their descriptions	1. Recruitment 2. Training and development 3. Procurement of raw materials 4. Stores & purchases 5. Attendance 6. Book keeping 7. Records & transcripts processing

Table 3: Survey data attributes and description

Functional description of enterprise functions	*Enterprise description:* Entities and their attributes have explicitly been identified:
1. Surveyors are employed by companies 2. Surveyors are qualified for survey work	1. Surveyor: 2. Company: 3. Incentives: 4. Department:

3. Surveyors possess skills 4. Surveyors are allotted to field work 5. Surveyors are given responsibility of handling equipment 6. Surveyors are instructed with terms and conditions of working 7. Surveyors acquire exploration data 8. Surveyors discuss the well positions with drillers 9. Surveyors are offered incentives for acquiring quality and quantity exploration data	5. Survey Activity: 6. Survey Type: 7. Survey Name: 8. Survey Line: 9. Survey Coordinates: 10. Survey Documents: 11. Survey Budgets:

Figure 4.50 shows the functional-dependency diagram corresponding to the semantics of the application. Most FDs result from the dependencies on keys. Keeping in view the result, the author designs a preliminary E^2/R conceptual schema for surveyor-activity information system.

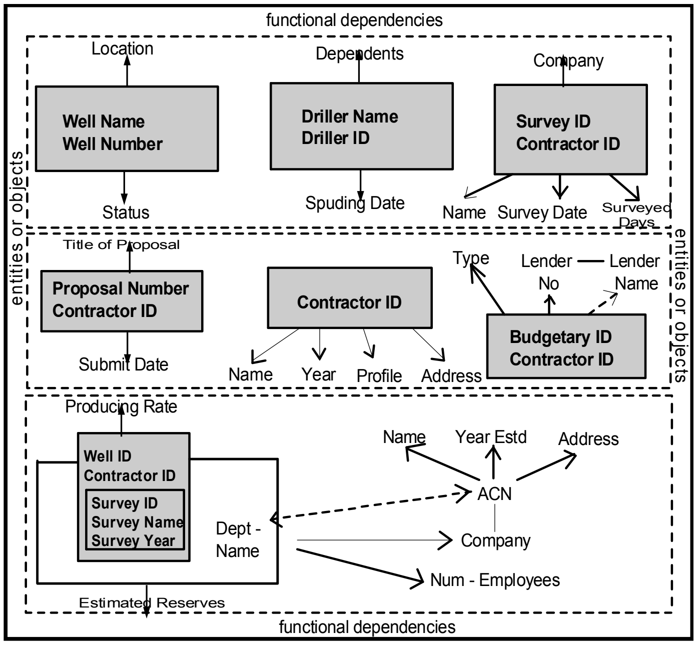


Figure 4.50: Functional dependencies in exploration data object

Transaction Analysis

This process identifies transactions needed in the system. Two examples given are as follows:

Transaction: List the surveys and the production rate for a contractor:

Entity types: surveyor, geology

Relationship types: survey or exploration

1. Retrieve the surveyor entity
2. Retrieve geology related to the surveyor entity via survey relationship
3. Retrieve geology with seismic-drilled well inputs and connect relationships

Transaction: List all surveyors assigned to a particular basin

Entity types: department, surveyor

Relationships: Approved for survey in a basin

1. Retrieve department entity
2. Retrieve a surveyor entity related to the department entity via approved relationship

Schema analysis

The analysis is carried out in the following areas:

Normalization: No MVDs are present; therefore, achieving BCNF will automatically provide 4NF.

Dependency-preserving decomposition: Decompositions to improve normalization must preserve all functional dependencies.

Lossless joins: All decompositions to improve normalization must be lossless join decompositions.

Normalization

The functional dependency (FD) observes the property of functional mapping in which semantics and integrities among attributes are described. Analysis of relations (NF as described in Table 4) and FDs shows that all relations are in BCNF except *inspection* and *expert*, which are in 2NF. After applying the membership algorithms and the algorithm for dependency-preserving 3NF decompositions, the following are obtained:

Table 4: Normalized Relations

QC-ID-no: Expert-name is redundant so it is removed	The three relations are in BCNF: Geologist ID-no; well-drilling is also redundant so it is removed
Quality control: ID-no, type Expert: ID-no, name Inspects: Inspect-no, Expert-ID	*Geologist* becomes: Geologist: ID-no, name, sex, salary Approved: ID-no, Well name Well-drilling: name, basin

Since a relation department (Well-name, basin, num-surveys) already exists, it is used instead of the relation "well-drilling" above. All relations are in BCNF. 2nd normal form (2NF) is one that embodies two disjoint facts together. 2NF is converted into 3NF relational schema, by separating disjoint facts by decomposing "well-drilling" relations. This 3NF is further converted into Boyce-Codd normal form (BCNF) without having dependency of primary attributes (well-name, well-ID) on non-prime attributes.

Lossless joins

Because all entities and relationships are linked via their primary keys, all join paths yield lossless joins. In this example decomposition of entities and relationships is not performed and hence no further lossless-ness check is required. As a result of schema analysis, the FD diagram and E^2/R conceptual schema can be refined. As shown in Figure 4.50, FDs connected to the dashed lines are ignored, and the required FD diagram is obtained.

Rule based refinement

The design of a good (or refined) conceptual schema that yields a legal relational schema is an iterative process and cannot be easily accomplished in an ad hoc manner. A legal relational schema is one that contains relations that are lossless, dependency-preserving, normalized, and free of unnecessary redundancies and anomalies. With the advent of knowledge-based systems and rule-based programming, it is feasible to encapsulate a substantial amount of the database administrator's knowledge into a system capable of performing normalization operations and making database-design decisions.

The methodologies used in the knowledge based systems are discussed with the conceptual schema conversion (Bayle and Ozkarahan 1988). The system is programmed in PROLOG and it is an iterative design session. Design methodologies that use the concepts in relational theory are discussed here. The algorithms can easily be programmed using recursion. The refinement flowchart is shown in Figure 4.51, and the process consists of E^2/R DDL information. The main ingredients of DDL are entities, relationships, aggregates and generalizations. The knowledge based systems convert these objects into their relational equivalent before applying the refinement procedure. The input may also contain functional and multi-valued dependencies. Normalization is the main highlight of this process. The output of the refined relational schema is synthesized back into E^2/R DDL to complete the cycle. The feedback for refinement process as seen in Figure 4.51(b), in reality takes place during the normalization phase.

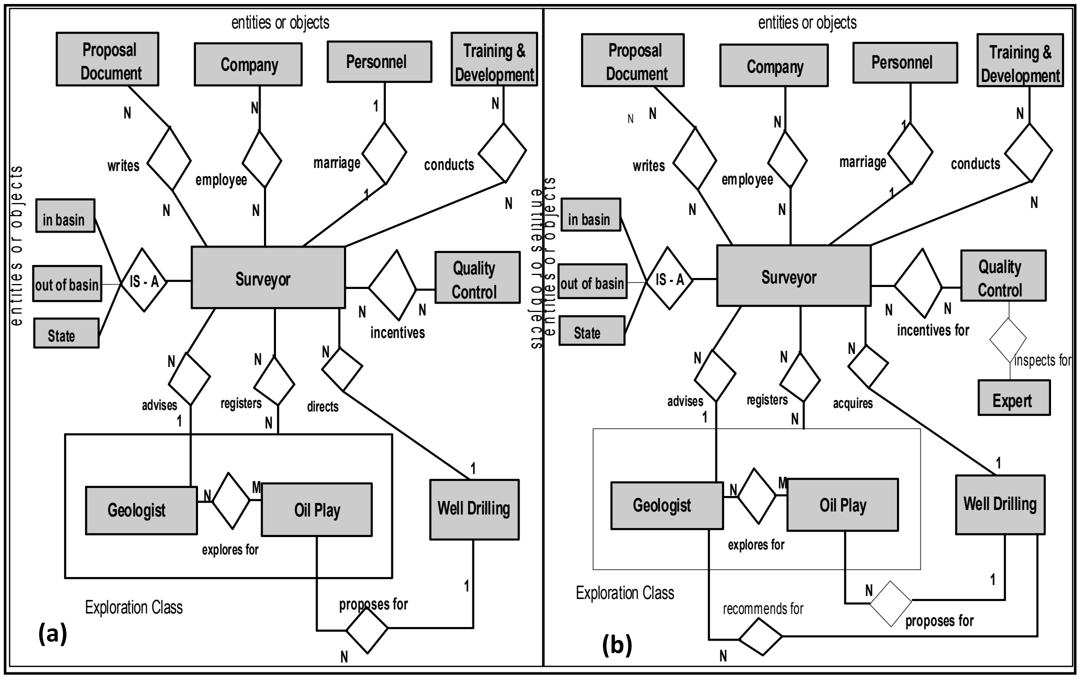


Figure 4.51: Conceptual schemas (a) preliminary E^2/R conceptual schema (b) refined E^2/R conceptual schema

The main body of refinement process is therefore in the normalization phase, which may consist of 3NF, BCNF as discussed in (Hoffer et al. 2005). It is not a simple normalization filtering process, but application of the complete know-how to produce lossless, dependency preserving, and normalized schema. The know-how can be provided with algorithms developed for normalizing relations in the different stages. Briefly, the algorithms which can easily be programmable (Deitel and Deitel 2001), are discussed in (Ozkarahan 1990). The author uses all these conceptual models for translating them in logical and physical models. Further, the process and purpose of writing syntax for heterogeneous data structures are:

Logic of integrating data models: The intelligent integration of several logical models is based on different ontologies of different concepts. The concepts simply constitute the properties and relationships that are constituents of the propositions of that ontology. Fine-grained structural properties, relations and propositions facilitate analysing the concept of ontology and its integration.

Process of integration: The main part of the framework represents a set of ontologies for applications, building an integrated metadata model. The application implies the process of translating ontological model to implementation data model. As stated earlier, data warehouses and data marts are used in a wide range of applications (Hoffer et al. 2005). Business executives in oil and gas industry use the data warehouses and data marts, to perform data analysis and make strategic decisions, such as well planning and budgetary proposals, including dedicated exploration and production activities. Generating reports and answering predefined queries are key uses of current data warehouse. In order to perform multidimensional data analysis, such as OLAP and slice and dice operations, the author addresses merits of warehouse architecture designs affecting data integration in Figure 4.52.

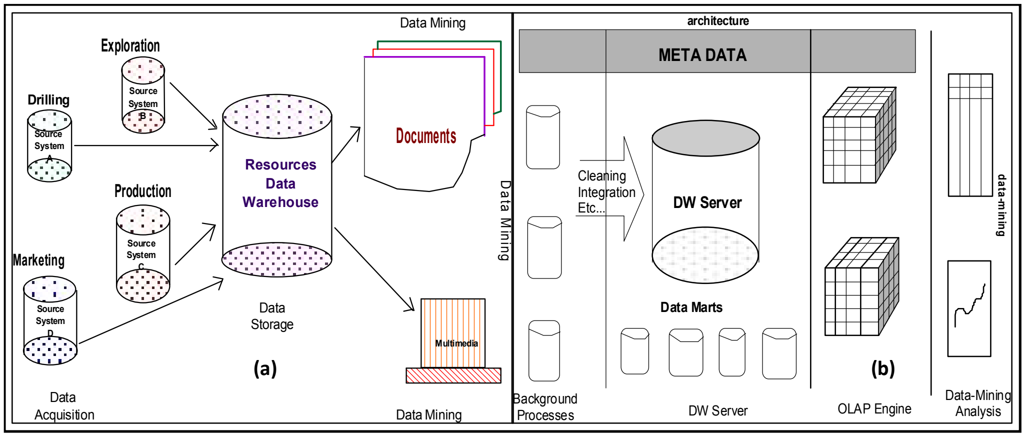


Figure 4.52: Data integration process and warehousing of petroleum data

As an example, as shown in Figure 4.53, the data instances of *seismic structure* and *reservoir* attributes (in the form of contours), interpreted in an integrated map view, suggest more probable locales of prospects or delineation of oil and gas deposits within a periphery of a matured field area. *Reservoir* and *structure contour* values (contours shown in Figure 4.53) are drawn to show equal values or instances of *reservoir* and *structure* attribute dimensions on the map views. Several drilled wells are represented, encircled in respectively as *oil* and *gas* in *green* and *red* colour attributes. This map is a demonstration of addressing research question RQ 8 and research objective RO 8, as described in Sections 1.3.1 and 1.3.2 in Chapter 1.

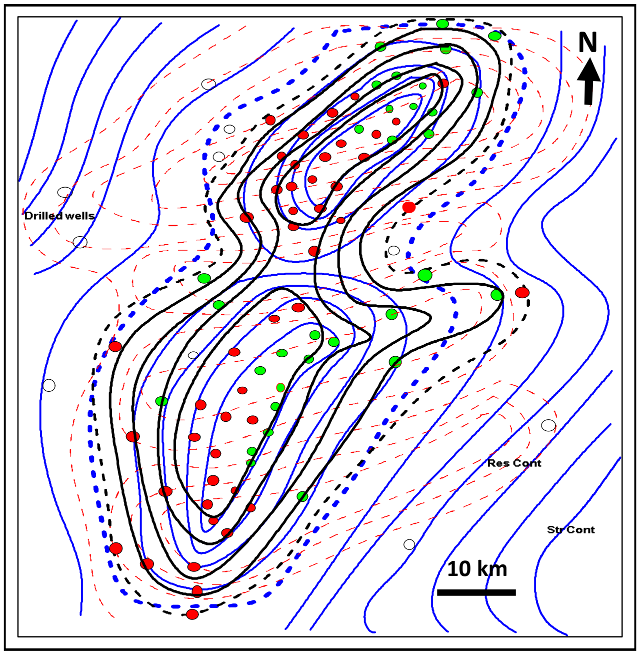


Figure 4.53: Domain ontologies integration, a schematic map view of *seismic* and *well-data* instances for risk minimizing various phases of exploration/appraisal/field development (circled symbols are drilled wells, green: oil; red: gas)

Ontology model implementation (addressing RQ 8 and RO 8 of 1.3.1 Chapter 1)

The conceptualized and logically designed petroleum data are stored physically in a modern database program (such as Oracle) for implementation. Extracting user defined business data views from warehouses and interpreting them for business intelligence are key challenges for successful implementation of this methodology. Implementation of ontology-based data warehouse approach for mining of petroleum data is vital for knowledge building from petroleum systems and effective reservoir management. Measuring the reservoir and production property attributes through periodic dimension has immense impact on economies of oil and gas business while carrying out expensive drilling operations.

The implementation of ontology approach validates petroleum oil and gas data structuring and ultimately for finer search of petroleum data and information from large online or offline warehouses and or information repositories. The ontology approach instigates improving data mining precision as well as reducing the overall amount of time spent searching for data. Conceptual or ontological models also support data delivery technologies that include agents for searching the data, data delivery agents (Erdmann and Rudi 2001) using meta-data languages and knowledge representation tools. Users trigger warehouse access to a piece of petroleum data of a drilled well, ontology identifies the description of that data view and search-engine acts to locate the data from that data warehouse as demonstrated in the next section.

This research attempts to implement a key data structure model as described in (Nimmagadda et al. 2006). The author uses metadata of the current data structure for extracting data views. One of such database view processed and generated through mapping, is represented as shown in Figure 4.54. As demonstrated in Figure 4.54, the author extracts multidimensional object views from a warehouse and maps them in an integrated data object structure (OLAP model). This is a combined view of several sub-classes and their associated attributes such as *reservoir* thickness, *oil-water contact* (OWC, an attribute derived from composite data structure, such as *Navig_Seismic_VSP_WellLog_Reservoir,* (Nimmagadda et al. 2006a), and *net oil pay* thickness. In other words, all the sub-class objects such as *points*, *seismic lines* and *regions* are uploaded and integrated to a single object data model, linking all their respective attributes from multiple object classes. The author interprets this model with an inference that all the ingredients of *oilplay*, such as, favourable *structure* and *reservoir* are present in a hydrocarbon (oil and gas) province. An explorer can conclude the model implementation, interpreting "H" as a locale (Figure 4.54) for drillable exploratory location, which minimizes the risk of *exploration* and also optimize the economics involved in future well-drilling plans around the vicinity of "H" location.

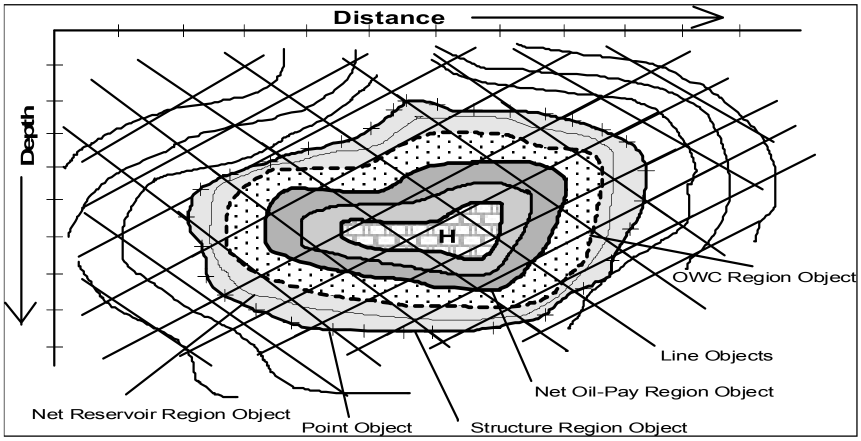


Figure 4.54: OLAP multidimensional objects data implementation model

As per research questions and objectives RQ8 and RO8 in Section 1.3.1, the knowledge base models deduced from metadata (Figures 4.53 - 4.54), are interpreted to be risk minimizing the exploratory drilling campaigns.

At this stage, the concept of a petroleum system is introduced, from which several elements and processes, described as entities and or dimensions. In ecosystem situations, all these elements and processes continuously interact and communicate each other. In the context of broader notion of a sedimentary basin, integration of these entities or dimensions, fit with notion of data warehousing, with true concept of metadata representation amongst multiple petroleum systems. In this study, the author acquires data from published sources from multiple petroleum systems from different sedimentary basins (Courteney et al. 1991 and Guoyu 2011) and integrate their domain ontologies, through data warehousing and for metadata representation (as spelt in research objectives (RQ 2- RQ 5), described in Section 1.3.1 of Chapter 1). The models deduced in Figures 4.53 and 4.54 are based on ontology models' and their integration in warehouse environment for implementation. Domain ontologies are integrated to arrive at a metadata and extract OLAP views (Figures 4.53 and 4.54) from metadata for visualization and interpretation of new knowledge. References and open access sources added in the Appendix-2 are taken advantage of data sources for modelling.

4.4 Systems dealing with the Big-data

The author discusses various features of big-data in respect of oil & gas industry in Section 2.9 in Chapter 2. Features of big-data are "volume, variability, velocity, visualization, veracity and value". All these features are incorporated in big-data systems design. As discussed in an exploration project in Chapter 4, it is mere an articulation of a system that deals with big-data. Systems dealing with big-data in oil and gas exploration industries, in particular, in the field of geo-informatics play an increasing role in the study of fundamental geological problems owing to the exponential explosion of sequence and structural information with time. There are two major challenging areas in geo-informatics: data management and knowledge discovery. In order to address these challenges, multiple heterogeneous data are integrated into logical database structures and thus derived, metadata range into several scales presenting huge analysis, for data mining and visualization opportunities. A challenge to data management involves managing and integrating the existing G & G databases. However, in some situations, a single database cannot provide answers to the complex problems of geologists and geophysicists. Integrating or assembling information from several databases to solve problems and discovering new knowledge are other major challenges in geo-informatics. The transformation of voluminous exploration data into valuable geological knowledge is the challenge of knowledge discovery process. Data mining and interpretation of interesting patterns hidden in trillions of seismic and other exploration data are critical goals of geo-informatics. The goals cover identification of useful reservoir bearing *structures* from petroleum ecosystems (Nimmagadda and Dreher 2012) that deal with big-data. The author discusses analysis of big-data systems in Chapter 5.

4.4.1 Petroleum ontology (PO), a general framework

This is integrated framework in which ontology models are mere articulates. Ontologies connect different articulates in the integrated framework as shown in Figure 4.55. The author builds and narrates ontology models that accommodating in the integration process. Author changes the title of Section 4.4.2, petroleum ontology- in an ecosystem scenario, keeping in view the content of big-data under title of section 4.4 in Chapter 4**.** Our common conceptual model for the internal representation of PO is based on the work done in Sidhu et al. (2009). Dimensions and attributes acquired with big-data focus are used in the ontology modelling process. These models are incorporated in an ontology framework, providing a common vocabulary for structured and unstructured information, to geologists, geophysicists and oil & gas (G- & G- and E- & P) explorers; it is a medium to share and access knowledge of petroleum/reservoir-geology domain.

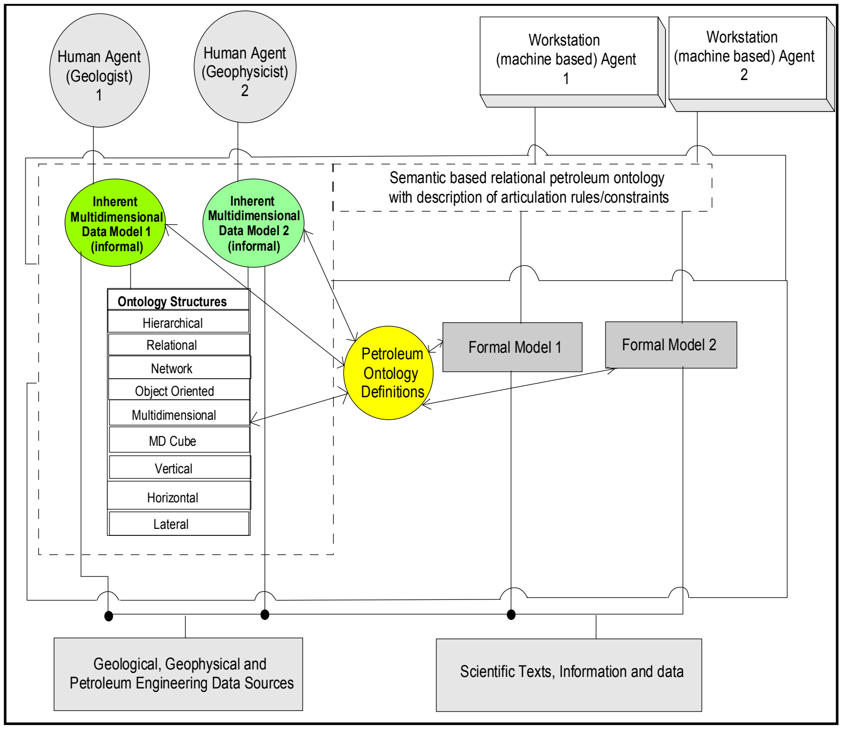

Human Agent (Geologist) 1
Human Agent (Geophysicist) 2
Workstation (machine based) Agent 1
Workstation (machine based) Agent 2
Semantic based relational petroleum ontology with description of articulation rules/constraints
Inherent Multidimensional Data Model 1 (informal)
Inherent Multidimensional Data Model 2 (informal)
Ontology Structures
Hierarchical
Relational
Network
Object Oriented
Multidimensional
MD Cube
Vertical
Horizontal
Lateral
Petroleum Ontology Definitions
Formal Model 1
Formal Model 2
Geological, Geophysical and Petroleum Engineering Data Sources
Scientific Texts, Information and data

Figure 4.55: PO general framework for modeling heterogeneous data sources

Specifically, the PO describes a series of events (Figure 4.55) in a big-data focus, especially in the context of the petroleum exploration domain that is used to depict reservoirs in any field or to depict in the drillable wells. PO describes:

1. Reservoir or/ seismic sequence and structure information
2. Structure or /reservoir integration process
3. G & G and geochemical processes of structures, reservoirs and sources
4. Reservoir/ or structural internal and external data associations to petroleum accumulations;
5. Constraints affecting the final petroleum trap/s or seal conformation.

The PO removes the constraints of potential conflicts arising during conceptualization and interpretations of terms in various data sources and provides a structured vocabulary that integrates all data and knowledge sources in a unified domain. There are eight concepts of PO, called generic concepts, which are used (and reused in different domains and in a broad scale of big-data, *basins*) to define complex concepts: *{{structure, reservoir, source, seal}, {chain}, {generation, migration, timing, accumulations}},* linked by *"chains" an evolved* concept. These conceptualized dimensions are typically narrated in hierarchical structure, as demonstrated in Figure 4.56. The PO specifies different topological relationships among *structure*, *reservoir* and *source*, in order to illustrate how mandatory semantic constraints are represented.

Source and *structure* have a mandatory constraint with *reservoir* because every *structure* (trap) and every *source* must topologically *touch* (analogues to *chain* as derived in Figure 4.56) one or more instances of a *reservoir.* Producing oil or gas field must *contain* at least one *reservoir*, while *reservoirs* have a mandatory relationship with *trap* in a producing field.

The elements of a petroleum system, their constraints and limits are explored on the internet through public knowledge domains of the Web. Notice in the Web Ontology Language (OWL) representation that minimum cardinality “1” is explicitly represented and is retrieved. Instances of *chains* of former *elements* of PO are defined in the *chains* concept (Figure 4.56). The multidimensional data structure of PO elements is represented with instances of the *elemen*ts, *chains* and *processes* concepts. Defining *elements*, *chains* and *processes* as individual concepts, has the benefit that any special properties or changes affecting a particular *element*, *chain* and *process*, can easily be added to PO.

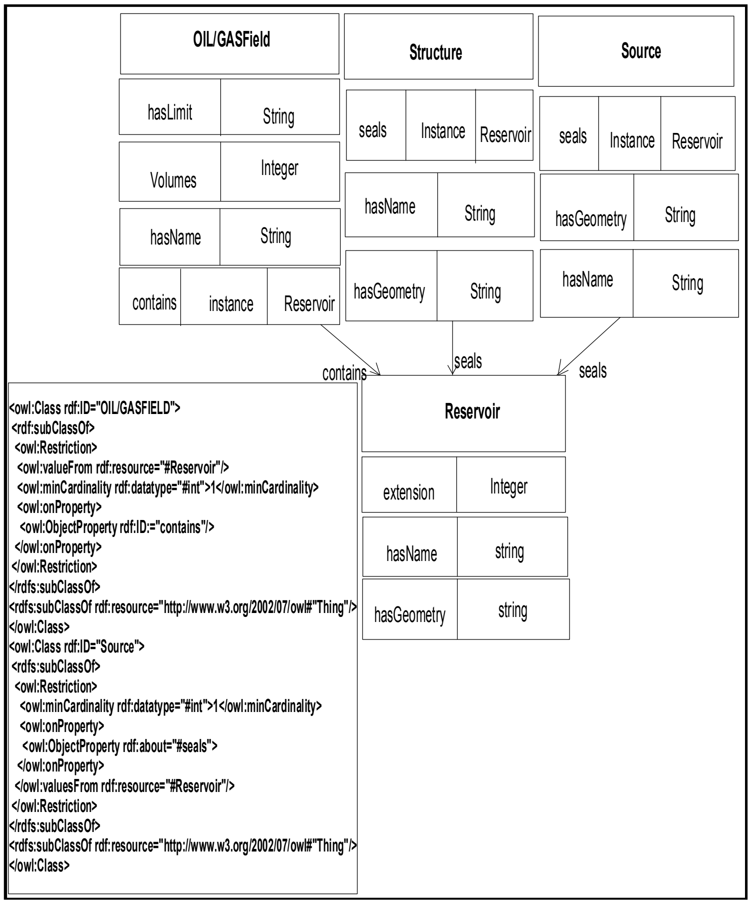
It is inevitable to associate the big-data with a total petroleum system (TPS). If the PO concept is tagged to TPS, all the details of petroleum system *elements*, *processes* and *chains* can be described within that contextualization, conceptualization and specification. Data about binding elements and processes, each in chains, are again entered into ontology modelling with their instances. Data among elements are chained by the elements-chains concept. Similarly, data among processes are chained by the process-chains concept. Similarly, data among elements and processes are chained by *elements-processes-chains* concepts. *Chains* are conceptually described by their instances.

- ● **MultidimensionalHierarchicalPetroleumOntology**
  - ○ **TotalPetroleumSystem(TPS)**
  - ○ **SedimentaryBasinDomain**
  - ○ **PetroleumSystemDomain**
  - ○ **ChainsEcosystemDomain**
  - ○ **PetroleumSystemsElements**
  - ○ **PetroleumField**
  - ○ **FieldRulesConstraints**
    - □ **StructureTrapDomains (Rule)**
      - ● **4-WayStructuralClosure**
      - ● **FaultGeneratedClosure**
    - □ **Constraints**
      - ● **PoorlyUnderstood**
      - ● **SurfaceExposed**
      - ● **NonSealing**
    - □ **ReservoirDomains (Rule)**
      - ● **Carbonates**
      - ● **Clastics**
      - ● **FracturedShales**
    - □ **Constraints**
      - ● **PoorQuality**
      - ● **ShalingDomain**
    - □ **SourceDomains (Rule)**
      - ● **Shales**
      - ● **Limestones**
      - ● Coals
    - □ **Constraints**
      - ● ThermalMaturity
      - ● **Quality**
      - ● **Quantity**
    - □ **SealDomains (Rule)**
      - ● **Shales**
      - ● **SiltStones**
      - ● **ClayStones**
      - ● **MudStones**
    - □ **Constraints**
      - ● **LocalSeal**
      - ● **PoorSeal**
    - □ **ProcessDomainRulesConstraints**
    - □ **MigrationDomains (Rule)**
      - ● **FarDistance**
      - ● **CloseToKitchen**
    - □ **TimingDomain (Rule)**
      - ● **NoTiming**
      - ● **BeyondTime**
      - ● **PoorlyUnderstood**
    - □ **GenerationDomain**
      - ● **OilWindow**
      - ● **FavorableTemperature**
    - □ **StratigraphicTrapDomain (Rule)**
      - ● **Channel**
      - ● **OnLap, TopLap, DownLap**
    - □ **Constraints**
      - ● **PoorlyUnderstood**
      - ● **PoorTrap**
      - ● **NoTrap**

(Elements, processes and chains of Petroleum System: Multiple Dimensions with hierarchies)

Figure 4.56: A multidimensional hierarchical structure view - describing dimensions of each system's element

The structure of PO provides concepts necessary to describe individual *elements, processes* and *chains*. The author stores instances of these dimensions either in excel files, ASCII, or in databases in Web ontology language (OWL) format. The complete multidimensional hierarchy of PO is shown in Figures 4.56 and 4.57. Another significant feature of representation of the structures is refining the boundaries or limits of system elements or processes. As narrated in Figure 4.57, the author draws the relationships among *elements* and *processes* of petroleum systems. Several logical rules and constraints applicable for scalability of PO models and multiple levels of information stored in PO models for query, information retrieval and presentation, are illustrated in Figure 4.57. Data views retrieved from metadata cubes are analyzed for data correlations and patterns. For this purpose, data visualization, data fusion and interpretation for new knowledge discovery are adopted. It is small piece of OWL code for elements of petroleum system. Representing *structure*, *reservoir* and *source* dimensions in a schema through OWL code (multiple dimensions of *elements*, *chains* and *processes,* are presented in a hierarchy in the Figure 4.56).



```
<owl:Class rdf:ID="OIL/GASFIELD">
 <rdf:subClassOf>
  <owl:Restriction>
   <owl:valueFrom rdf:resource="#Reservoir"/>
   <owl:minCardinality rdf:datatype="#int">1</owl:minCardinality>
   <owl:onProperty>
    <owl:ObjectProperty rdf:ID:="contains"/>
  </owl:onProperty>
 </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#"Thing"/>
</owl:Class>
<owl:Class rdf:ID="Source">
 <rdfs:subClassOf>
  <owl:Restriction>
   <owl:minCardinality rdf:datatype="#int">1</owl:minCardinality>
   <owl:onProperty>
    <owl:ObjectProperty rdf:about="#seals">
   </owl:onProperty>
  </owl:valuesFrom rdf:resource="#Reservoir"/>
 </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#"Thing"/>
</owl:Class>
```

Figure 4.57: Representing *structure*, *reservoir* and *source* dimensions in a schema through OWL code (multiple dimensions of *elements*, *chains* and *processes,* are presented in a hierarchy in the Figure 4.56*)*

4.4.2 Petroleum ontology – in an ecosystem scenario

The author discusses a comprehensive framework as given in Nimmagadda and Dreher (2011) and Nimmagadda and Dreher (2012) for different petroleum ecosystems' scenarios. Big-data and its features have definite roles in connecting multiple petroleum systems in a basin. A conceptual framework of PO, as demonstrated in Figure 4.55, keeps criteria (1) a hierarchical classification of concepts represented as classes, from general to specific; (2) a list of attributes related to each concept, for each class; (3) a set of relations between classes to link concepts in ontology in more information-rich ways than implied by the hierarchy, thus promoting reuse of concepts in the ontology; and (4) a set of algebraic operators for querying petroleum ontology instances. The contexts of PO domains, (Figures 4.55 – 4.57), depict the overall layered approach for generic PO structure. Three layers are distinguished here:

1. The author considers real-world data and information from—various proprietary G- & G- and petroleum engineering data sources, including scientific reports, company and project documents, and published literature.
2. Data elements and models (both of inherent data descriptors resulting from patterns in data, and formal data models defined by scientists for specialized applications) possess different formats and the anomalies are addressed by PO syntactic structure. These data models represent post analysis information exchange. Both human and machine agents that access and analyze the ontology are considered.
3. Human agents, such as geoscientists interpret and analyze ontology directly using their expertise. For machine agents to interpret and analyze ontology, at the same level as human agents, the expertise represents using semantic relationships and articulation rules in an ontology framework, as shown in the Figure 4.55.

4.5 Systems in a Turbulent Business Environment

Here the author refers “turbulent” as unstable because of falling in energy prices, global currencies and geo-politics. As described in RQ8 and RO8 in Sections 1.3.1 and 1.3.2 in Chapter 1, risk minimization of exploration and its economic implications are explored with the turbulent situations. The systems designed and developed in the current study have business and economic focus. Several operational units are considered in the oil and gas industry domain. As described in Flahive et al. (2004) and Bhatt et al. (2004), the author uses ontologies in distributed business environments. Oil and gas data belonging to multiple operational units are typically *exploration*, *drilling*, *production* and *marketing* even in distributed business environment and these business units operate both locally and or globally. The author discusses data sources and requirements for conceptualizing and building data models in Nimmagadda and Dreher (2006). For integration purposes, the data from multiple sources are cleaned, reformatted, logically and physically organized in an intelligent storage environment.

The exploration in an oil and gas business is a key generalized dimension, in which hundreds of dimensions and attributes associated, as are described in (Nimmagadda and Dreher 2007 and 2008). Several other factors, based on processes in the exploration business from which these data dimensions conceptualized, are narrated. During data integration process, data from several operational dimensions are gathered and inter-connected through clustering. The periodic dimension, which has been the subject of current multidimensional data modeling approach, is addressed for several basins. Oil and gas data are organized in different data schema approaches, as narrated in Nimmagadda et al. (2006) and discuss a data warehousing environment approach using the multidimensional modelling.

A petroleum system is an information system, in which inherent natural data entities and their associated attributes are conceptualized and contextualized, by connecting data relationships of multiple systems' entities (dimensions) and attributes. Analogous to entity, dimensions are used in the present data modeling schemas. For this purpose, the author addresses an integrated framework. But dimension names and their contexts are semantically interpreted, which are not inherent, such as the ones' in any project management, *surveys*, *wells*, *permits*. Similar many other entities associated with *drilling*, *production* and *marketing* business units are interpreted. Major composite dimensions , such as oil and gas fields (mineral provinces in case of mining farms) in which *wells*, *boreholes*, *surveys*, *geological markers*, *seismic horizons* and multiple *log-profiles* (including *vertical seismic profiling) sub-type* dimensions are described, interact (communicate) logically between surface (known seismic data instances) and sub-surface unknown geology (drilled well data instances). There are many users in different operational centers, share these multi-disciplinary datasets in different contexts.

In both collaborative and individual projects, several functions and activities of oil and gas exploration and field development are performed. AOI (areas of interest) sub-projects are also created to handle a specialized data, such as *well* and *seismic* data accessed by geologists and geophysicists working in a shared project under *exploration* entity and or dimension. Other important functions of *exploration* operational unit are procurement of necessary financial budgets for conducting exploration operations. Author describes them as external entities, in which each entity is narrated by a different data dimension and its corresponding fact data tables. Various concepts, forecasting methods and their practical applications in industries are discussed in Chapter 5.

Surveys, *wells* and *permits* are key entities or dimensions from which several instances, used in the study for generating conceptual data models, described in star schemas, as demonstrated in Figures 4.58 and 4.59. The integrated framework for warehousing oil and gas data given in Figures 3.36 - 3.38 is used, in which all the workflows for integration including data modeling, warehouse and other analytic procedures needed to compute and interpret data views, are described. Besides technical data, the author interprets financial data with mineral and petroleum exploration costs that affect industry turbulent business environments. Few models used in the integration process are given in Figures 4.58 and 4.59.

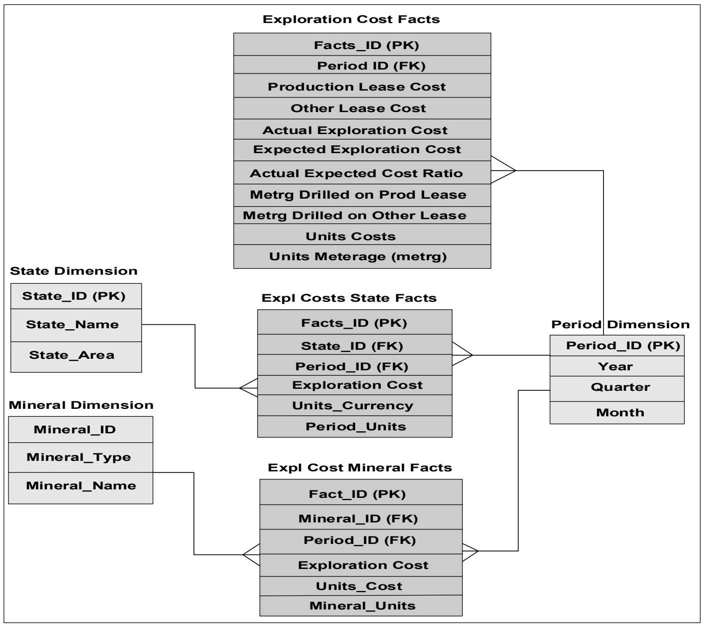


Figure 4.58: Multidimensional star schema data model for *exploration costs*

Daily, volumes of dimension and fact tables with multiple rows and columns are handled in petroleum business operations. Data modeling makes use of the concepts of similarity and associativity to inter-relate several data attributes among several dimension and fact tables. As shown in Figures 4.58 and 4.59, data models offer links to multiple fact tables of a dimension named, *petro2-surveys* database.

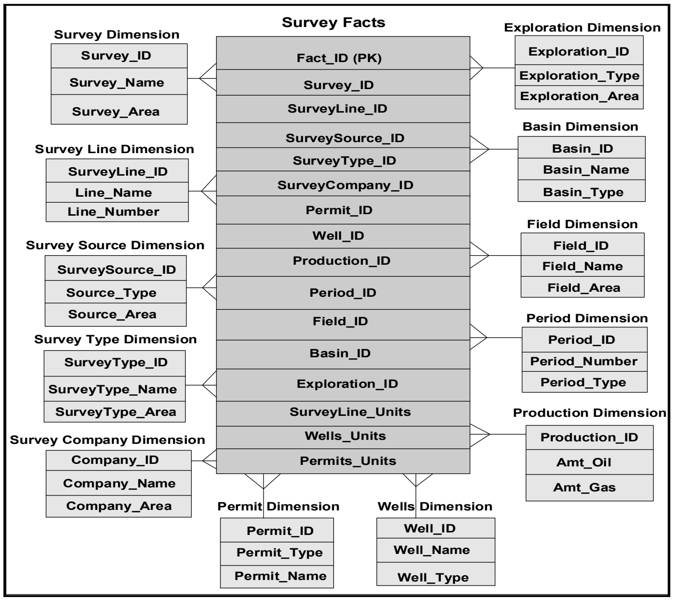


Figure 4.59: Multidimensional star schema data model for *surveys* (database)

Petro2-surveys database (built for the current study and analysis) consists of multiple fact and dimension tables, from which only representative dimensions are denoted in Figures 4.58 and 4.59 for demonstration purposes. The author uses an overall integrated framework for integrating warehoused multidimensional data structures with data mining and other forecast procedures. The author discusses more focused economic objectives (for minimizing exploration risk) of the study with benefits including forecasting and decision-making in the oil and gas industry in Chapter 5.

4.6 Conventional Digital Ecosystems

The author addresses the research question RQ7 and the research objective RO7 in this section. The conventional digital ecosystems refer to reservoirs of variety of petroleum systems. Exploration of conventional reservoirs is one of the key objectives of any oil and gas exploration project. For the purpose of identifying data relationships among several conventional reservoirs and integrating them in a warehouse environment, an integrated framework (Nimmagadda and Dreher 2012) is adopted using ontology-based multidimensional and heterogeneous warehousing for mining multi-disciplinary petroleum oil and gas data. For example, seismic and drilled-well data are integrated, during which, multiple horizons are correlated and interpreted (or seismic, geological, or well markers). Each individual horizon is a *sedimentary* layer that has several multiple dimensions. Here, properties of horizons are considered as multiple dimensions, at places conceptualizing them, for the purpose of mapping and documenting them in a data warehouse environment. The author organizes all the seismic horizons and their associated property dimensions for integrating in a warehouse environment. Typically, such dimensions are *seismic (surface-domain)*, drilled-*well (subsurface-domain)*, *point*, *line* and *navigational* (surface-domain) dimensions. Each *point* on a line has a unique coordinate data instance and it is tagged with other multiple property dimensions. Such properties are - *time* (seismic time), *depth*, *velocity*, *density*, *porosity*, and *fluid* type. Here, the petroleum ecosystem concept is introduced, in which, each point has a description of *structure*, *reservoir*, *source*, *seal* attributes and their instances including other data instances related to processes, existing within an ecosystem.

These processes are typically migration pathways and *timing* of occurrence of each element within the system. In fact, each layer or horizon interpreted in a sedimentary basin has numerous dimensions. For the purpose of denormalizing data relationships in multidimensional data structures, their attributes, for each and individual dimension (element or process) is, again segregated or categorized in a way that enables a more effective integration at later stages may facilitate effective data mining process as well. In the case of petroleum digital ecosystems, the author categorizes all the elements participating in an ecosystem process in a way so that, the ontology description understands each element's terminologies, including semantics, schematic and syntactic heterogeneities. The author describes ontologies for horizons (oil and gas producing layers) and elements of petroleum system, affecting these producing horizons are described in the following sections *1, 2, 3, 4, 5 and 6*:

1 Description of horizon ontology

An interface that might be represented by a seismic reflection, such as the contact between two bodies of rock, having different physical properties, seismic velocity, density, porosity and fluid content, is associated with sedimentary layer properties. Here *seismic* is a super-type dimension, based on which other multiple horizons are interpreted keeping in view these properties. Seismic horizons are a significant part of inputs that contribute to prospective depth models and then to a description of an ecosystem both at field and basin scales.

A horizon is an informal term, used to denote a surface of rock, or a distinctive layer of rock that might be represented by a reflection in seismic data. The term is often used incorrectly to describe a zone from which hydrocarbons are produced. Generally, seismic data possess multiple horizons that are used in modeling time or depth structures. The author believes that a horizon (either seismic or geological) should never be viewed as a single entity or dimension or isolated from multiple horizons at any stages of processing and or interpretation. Multiple horizons are ontologically interlinked or interconnected in an ecosystem scenario. Horizons in a sedimentary basin form a collective ecosystem, in which all the horizons interact and communicate (during several geological periods) among themselves through shared properties. If there is any change in a single horizon, there is a corresponding change in the other horizons (reasons could be structural connection and reservoir extensions in multiple horizons in 2D/3D datasets). For example, multiple horizons are interlinked with the *Recent, Pleistocene, Pliocene, Miocene, Oligocene, Eocene, and Paleocene*, multiple geological ages. Each horizon is characterized by the attributes of *structure*, *reservoir*, *seal* and *source* elements.

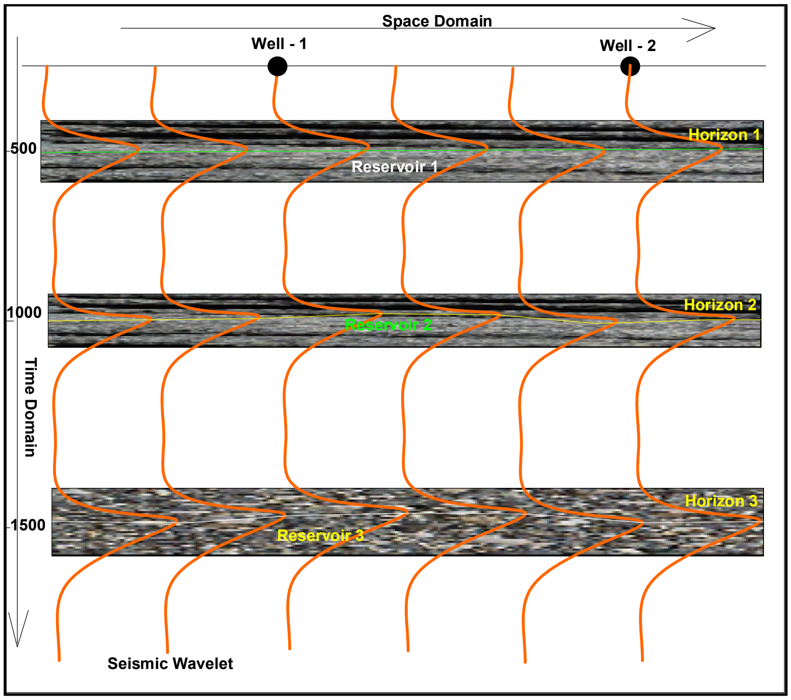


Figure 4.60: Integration of *seismic* and *well-data* dimensions and their instances – representation of *horizons* in time-domain ontology

Keeping in view the concept of an ecosystem, interpretation of a single horizon in the seismic data has no meaning for knowledge building of reservoir connectivity. As per principle of an *ecosystem*, horizons, which are either in time or depth domain, constantly interact and communicate with each other at field and basin scales. If there is any effect on a single horizon, a similar effect is interpreted on other horizons so that multiple horizons are collectively targeted and interpreted (Figure 4.60) to understand the overall effect of horizon ecosystem in a sedimentary basin. For example, the author tags horizons, interpreted in an exploration project with several keys, such as navigation or positional data (X, Y, Z Cartesian coordinates) representing space and certain property (Z) dimension. Each horizon may have several *peaks* and *troughs* with pointed seismic data, each representing seismic time data instances (Z) and corresponding easting and northing (X, Y) instances. Multiple horizons can have multiple X, Y, Z positional data and volumes of *peak*s and *trough* data instances.

2 Basin and petroleum system ontology modeling

Sedimentary basins, or simply basins, vary from bowl-shaped to elongated troughs. If rich hydrocarbon source rocks occur in combination with appropriate depth and duration of burial, hydrocarbon generation can occur within the basin. In this context, the author models a sedimentary basin or groups of sub-basins, using robust multidimensional modeling (Shastri and Dreher 2011a and Nimmagadda and Dreher 2011b) and set theory (Bartle 1976, Halmos 1974 and West 2006) principles. The author uses set theory in order to make unions and joins among categorized groups of sets, identified among dimensions of elements of petroleum system. Even dimensions associated with horizons and horizon ontology descriptions and connectivity of ecosystems, set theory principles are applicable. But for now, all the horizon ontologies described with multiple dimensions are integrated within a warehouse or computer environment in the form a petroleum digital basin. Each digital basin may have multiple digital fields (either oil or gas or both) and each field has number of surveys and wells.

A sedimentary basin is a depression in the crust of the Earth formed by plate tectonic activity in which sediments accumulate. The continued deposition can cause further depression or subsidence. Petroleum systems are developed based on the tectonics and depositional environments and these establish the geometries and strengths of elements and processes. As described in Figures 3.36 – 3.38 and Figure 4.61, the basic business constraint is that each basin has one or more petroleum systems. Each petroleum system has one or more oil/gas fields or it may not have any. Each proven or productive petroleum system has all the essential *elements* and *processes* (Magoom and Dow 1994).

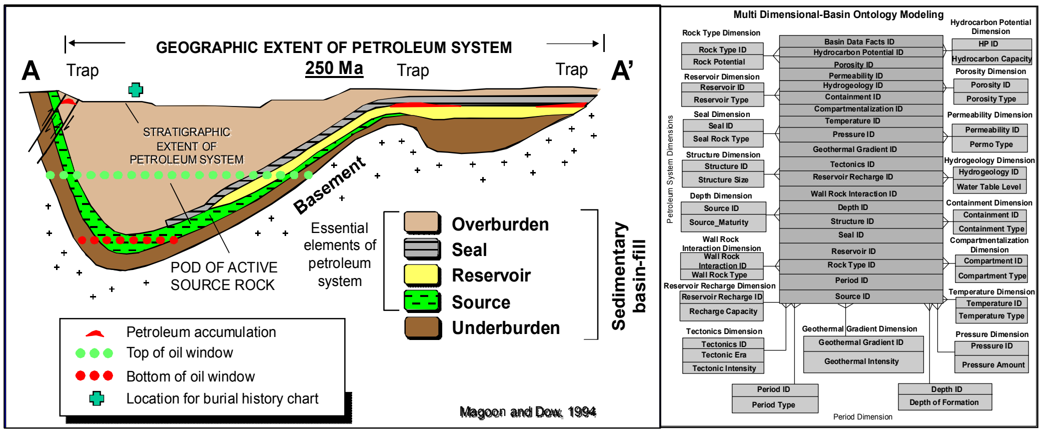


Figure 4.61: Depiction of multidimensional petroleum system and its ontology

It is an established fact (Magoom and Dow 1994) that all the producing giant fields and or matured fields are from these basins, where sediments form structures due to tectonic activity and the hydrocarbons generated in source rocks migrate long distances (to the order of several kilometers) and get entrapped within geological structures. Typically structures embed with reservoirs of differing qualities that hold the hydrocarbon fluids. Here the author introduces the concepts of set theory.

In order to ease the representation of complexity that exist in the reservoir ecosystems, author uses set theory concepts. Intention is to represent conveniently the conceptualized dimensions and attributes evolved from elements and processes of a petroleum system. The author tries to link different elements of petroleum systems in different oil & gas fields of different basins. For example, certain elements or processes show favourable conditions, other very poor conditions for petroleum accumulations. Unions and joins are made among these elements and processes. This process facilitates categorization of elements and processes and thus make cluster, based on correlations, trends and patterns. *Source*, *structure*, *reservoir* and *seal* are primary elements and *migration*, *timing* and *maturity* are interconnected processes (Figure 4.61). The connectivity is represented as:

{{structure, reservoir, source, seal} ↔ {chains} ↔ {source maturity, generation, migration, timing, accumulations}} – a representation, appears to be exploring connections in a petroleum ecosystem. The relationships are defined and explored using set theory (Bartle 1976, Halmos 1974 and West 2006).

If X = {*A, B, C, D*} and Y = {structure, reservoir, source, seal}, then $|X| = |Y|$ because {(*A*, structure), (*B*, reservoir), (*C*, source), (D, seal)} is a bijection between the sets X and Y. The cardinality of each of X and Y is 4.

- If $|X| < |Y|$, then there exists Z such that $|X| = |Z|$ and $Z \subseteq Y$.
- If $|X| \leq |Y|$ and $|Y| \leq |X|$, then $|X| = |Y|$. This holds even for infinite cardinals.

If A, B, C, D are sets of sedimentary basins or/ sub-basins in the context of a global sedimentary basin (global ontology), several permutations and combinations of sets of elements (of these basins/or sub-basins) are possibly relevant to each petroleum ecosystem domain that is organized through unions and intersections of sets as shown here:

Case 1: $|X| = |Y|$

Two sets, X and Y, have the same cardinality, and, if it exists, a bijection (that is, an injective and surjective function) from A to B.

For example, the set *reservoir* = {0, 2, 4, 6 ...} of non-negative even numbers has the same cardinality as the set N = {0, 1, 2, 3 ...} of natural numbers that represent the total number of sets existing within an ecosystem scenario, since the function $f(n) = 2n$ is a bijection from *reservoir* to N. Here, *reservoir* represents the qualities of *reservoirs* among all petroleum ecosystem reservoir elements, depicting multiple reservoirs of a particular basin or group of sub-basins (B). Bijection, injection and surjection may be different permutations. In the database relationship terminologies, these permutations may represent one-to-one or one-to-many data relationships.

Case 2: $|X| \geq |Y|$

A has cardinality greater than or equal to that of B if there exists an injective function from B into A.

Case 3: $|X| > |Y|$

A has cardinality strictly greater than that of B if there is an injective function, but no bijective function, from B to A. Similarly, sets of elements can be represented as unions or intersections based on the classifications of attributes or properties of elements of a petroleum ecosystem. If A, B, C, D are basins/sub-basins of a broad sedimentary basin (or in a Total Petroleum System, the TPS scenario) within a particular set (to a category to which elements belong), if *A* and *B* are *disjoint* sets, then

$$|A \cup B| = |A| + |B| \ .$$

From this, one can show that in general the cardinalities of unions and intersections are related by

$$|C \cup D| + |C \cap D| = |C| + |D| \ .$$

Alike elements in an ecosystem and existing processes are also affected. The elements and processes are explored for connections (through permutations and combinations) to a new conceptualized element derived, from {chains}. *Chains* within an ecosystem are more conceptualized and based on the interaction among elements and processes. A *chain* can be either between individual elements of basins or processes and or elements and processes of individual petroleum systems. The seismic data, making up these *chains*, consist of numerous *peak* and *trough* dimensions (Shastri and Dreher 2011, Nimmagadda and Dreher 2011 and Nimmagadda and Rudra 2004) and *chains* are used for connecting horizons and their associated system elements *structure*, *reservoir*, *source* and *seal in a* petroleum ecosystem. The author initially uses all the petroleum ecosystem elements' in the modelling, as sets described in set theory (West 2006); cardinalities of elements are thus described in several permutations and combinations as well. Then each element is narrated with an ontological description to formulate the connectivity. Ontologies are described for each individual element here. Primarily, more focus is on narrating a reservoir ontology.

3 Source ontology

A source rock is the main element in the ontology description. A rock rich in organic matter, if heated sufficiently, will generate oil or gas. Typical source rocks, usually shales or limestones, contain about 1% organic matter and at least 0.5% total organic carbon (TOC), although a rich source rock might have as much as 10% organic matter. Preservation of organic matter without degradation is critical to creating a good source rock, and necessary for a complete petroleum system. Under the right conditions, source rocks can also be reservoir rocks, as in the case of shale gas reservoirs. Among several dimensions involved within source ontology, ontology description of maturity of source rock is a characteristic property of any petroleum ecosystem.

4 Structure (geological) ontology

The structure and entrapment of hydrocarbons are key elements of any petroleum ecosystem. All the necessary inputs needed to explore and describe ontology connections through trapping mechanism or structural connectivity, are given in Magoom and Dow (1994).

5 Seal ontology

A seal rock is another characteristic property of an entire petroleum ecosystem habitat. Unless an appropriate seal rock is interpreted on a regional scale, oil and gas field existence cannot be well explained in its totality within an ecosystem setting. Accordingly, all the descriptions of ontologies associated with seals are narrated (Magoom and Dow 1994) to make seal connectivity possible.

6 Reservoir ontology

A reservoir is one of the principal components of a petroleum ecosystem, being the accumulations of hydrocarbons. The primary focus is modelling and exploring reservoir connections, based on multidimensional warehouse modelling and mining that supported by an integrated ontology framework, (Figures 3.36 - 3.38). Case-1, Case-2 and Case-3 are reviewed in accordance with permutations, for further refining and establishing fine-grained data and thus to derive a metadata, including reservoir attribute plot and map views (Figure 4.62).

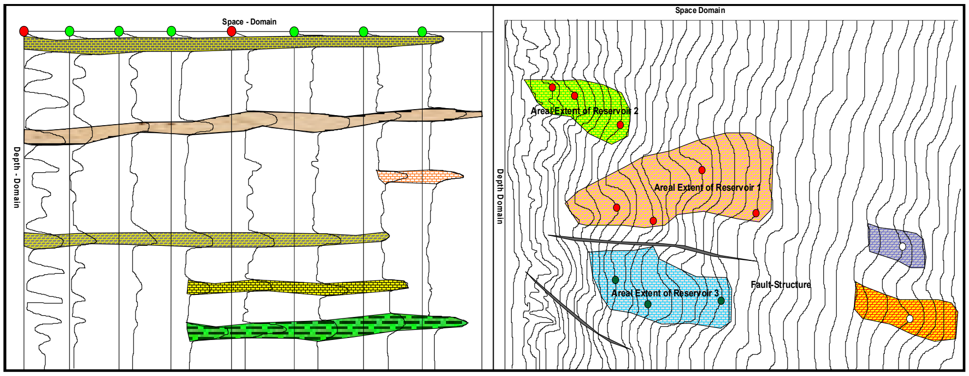

Space - Domain
Depth - Domain
Space Domain
Depth Domain
Areal Extent of Reservoir 2
Areal Extent of Reservoir 1
Areal Extent of Reservoir 3
Fault-Structure

Figure 4.62: Exploring reservoir connections in space and depth dimensions on plot and map graphic views (seismic and well-log views integrated)

Exploring connections among reservoirs through their domain ontologies are effective, when the multidimensional and heterogeneous structures are fine-grained.

4.7 Unconventional Digital Ecosystems

The unconventional digital ecosystems refer to unconventional reservoirs of variety of petroleum systems. The author builds couple of models in time and depth domain that can connect to a different composited dimension “velocity”. Time, depth, velocity dimensions, their attributes and instances facilitate in building connectivity process. When once it is done for a field, ontologies establish its connectivity to nearby fields in a basin. The author believes, the whole process is a development process. In this section, author addresses RQ7 and RO7. It is meant for designing and connecting ecosystems of “unconventional” nature. In recent years, exploration of unconventional reservoirs has been increasing momentum worldwide, especially in the areas of shale gas and fractured shales. It is vital to put forth all facts and figures of gas shale and organize the data instances in a way that users can perceive and extract knowledge for interpreting the associated geological structures and reservoirs. The author uses different data warehousing and mining methodologies for oil and gas industries (Gornik 2002 and Khatri and Ram 2004) to address data integration and interoperability issues. Keeping in view the data management and application issues in the current research, we propose the data schemas and integrated frameworks as a part of systems’ development for unconventional reservoirs. For this purpose domain ontologies are intended to be integrated in a warehouse environment. Data dimensions are used in the current modeling studies as given in Nimmagadda and Dreher (2008d) and Nimmagadda and Dreher (2012), in which detailed description of ontologies are given. Often, *point*, *line*, *spatial* and spatio-temporal datasets vary with space and time and hydrocarbon-producing reservoirs are associated with high density fractured networks and their orientations. Hundreds of dimensions and their attributes are involved in this resource business including periodic dimension. The geographic (space) and periodic (life span of resource or reserve) dimensions and heterogeneous data from seismic and well-domains are other dimensions that were included in the current data modeling.

A shale reservoir by definition is a hydrocarbon source, reservoir trap and a seal in a single system, called an unconventional petroleum ecosystem. No two shales are similar, though they present as complex and heterogeneous with extremely low permeability instances. Stress anisotropy is common among shales, which is addressed by 3D seismic surveys. Integration of geology, geophysics, reservoir engineering and geo-mechanics is recommended to reduce the risk and uncertainty involved in mapping and interpretation of shale- and tight-gas reservoirs including

CBM. Anisotropy, Poisson's ratio and Young's modulus properties facilitate interpretation of stress images from 3D acoustic characterization studies. Reservoir and completion qualities are crucial for successful development of these reservoirs. Key properties and characteristics of unconventional reservoirs, addressed in any ecosystem scenario are described here:

Tight Gas Reservoirs

- Sandstone or carbonate, single porosity
- Low permeability (< 0.1md)
- Stimulation with Hydraulic fracturing or acidizing needed for production

Coal Bed Methane (CBM) Reservoirs

- Naturally fractured system; Matrix (coal) controls gas storage
- Fracture (cleat) controls fluid flow
- Gas adsorbed onto the coal matrix (<1~>25m3/t)
- Fractures (usually) initially water-filled
- Primary production: initially by dewatering natural fractures
- Reduced pressure in fracture system, which allows gas desorption from coal surface to fracture
- Secondary Production (enhanced CBM recovery) : CO_2 and N_2 injection

Shale Gas Reservoirs

- Gas adsorbed on the internal surface of the organic matter (<10 m^3/tonne)
- Free gas in the micropores and fractures
- Very heterogeneous with nano-darcy matrix permeability
- Produced from matrix, through fractures, or from more permeable sands
- Horizontal wells + Hydraulic Fracturing
- Coals and shales interbedded in a single reservoir

The recent worldwide economic downturn and steep fluctuations of hydrocarbon pricing have created a momentum for shale gas exploration and domestic consumption. New ideas are explored in shale gas data management as explained in the following sections A, B, and C:

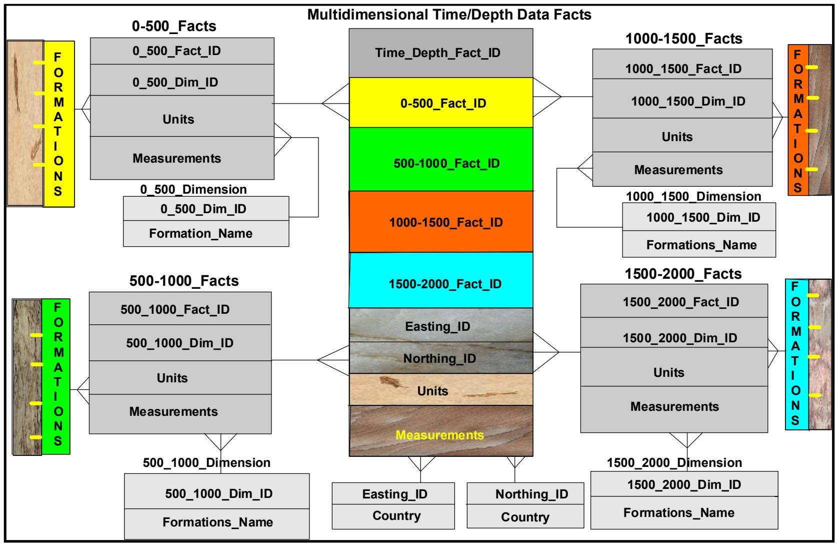


Figure 4.63: Multidimensional modelling of *time-depth* data instances

A Time-domain ontologies

The seismic data are represented in different space and time dimensions and author writes separate ontologies for multiple seismic vertical time ranges: 0-500ms, 500-1000ms, 1000-1500ms, 1500-2000ms and 2000-2500ms. Each (vertical) time range has different, geologically interpretable knowledge (Figure 4.63), such as horizons, structures, reservoirs, source and seal rocks, and different processes. In a typical 2D seismic section (Figure 4.64), several features, such as *structure*, *facies changes* in seismic reflection character that attribute to reservoir distributions are explained.

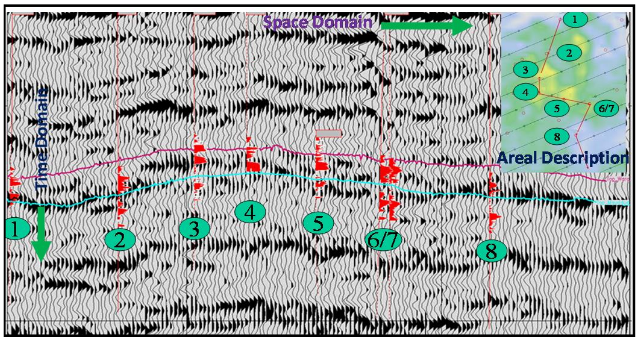


Figure 4.64: Multidimensional data descriptions in *time*, *space* and *areal* domains (seismic and well-log data instances used in integration)

B Depth-domain ontologies

The author writes similar knowledge based ontologies in depth-domain, in which depths ranging 0-500m, 500-1000m, 1000-1500m and 1500-2000m are described (Figure 4.60, Figure 4.63 and Figure 4.65). Each range has different knowledge levels;

for example, CBM occurs in the range of 0-500m, conventional non-associated gas from 500-1000m, conventional associated gas from 1000-1500m and tight gas sands beyond 2000m. Hydrocarbons associated within shale can be deeper (more than 3000m), where in general, they experience matured temperatures for hydrocarbon generation. Extraction and interpretation of knowledge for deeper reserves of shale gas requires more effort and is more expensive.

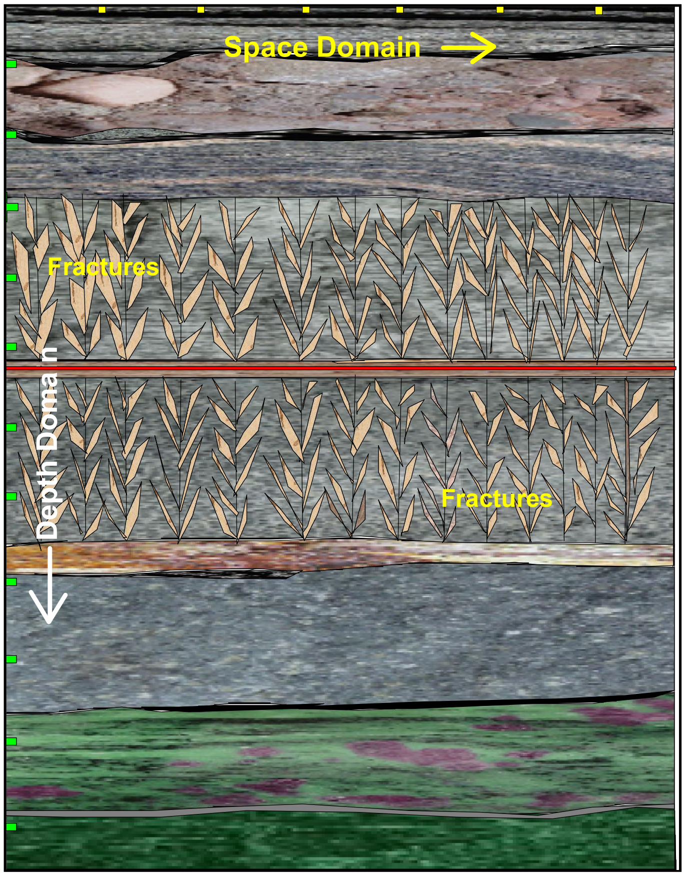


Figure 4.65: Multidimensional data, depicting *space* and *depth* dimensions in fractured formations (geological containers/reservoirs)

C Description of ontologies for multidimensional unconventional data sources

For the purpose of bringing multidimensional heterogeneous data for integration (Castañeda et al. 2012) through integrated frameworks, the author gathers several data instances from unconventional oil and gas provinces. Hundreds of dimension- and -fact tables including attributes and their corresponding data instances are documented for modeling purposes. One of the key multidimensional models representing the *shale-gas* exploration and development (Nimmagadda and Dreher 2011b) is presented in the Figure 4.66.

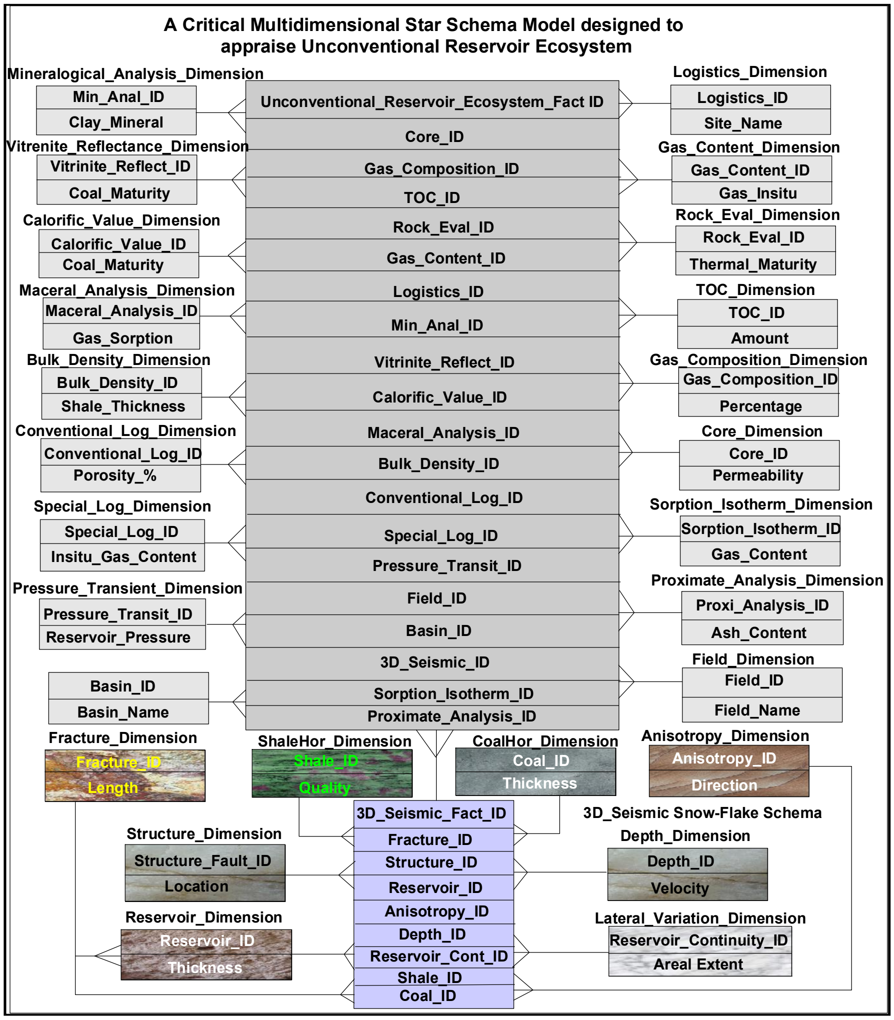


Figure 4.66: A multidimensional star schema describing an unconventional reservoir ecosystem

Several interconnected dimensions which represent one-to-one, one-to-many and many-to-many relationships to *exploration* domain, are logically (ontologically described) connected to their corresponding fact tables within a warehouse environment. All these dimensions and associated attributes are either hierarchical or relational or both, and are used for modeling data instances. The data or map views mined from these metadata structures must be interpretable in terms of meaningful fractured (structure or reservoir) geology. In Figure 4.63-4.66, several *point*, *line*, *spatial* (areal) dimensions are ontologically described, so that data structures connected to seismic- and well-domain datasets, are spatially mappable, for interpreting the fractured networks. These multidimensional data models are in both time- and depth-domains, narrating ontologies for every time- and - depth ranges, which enabled developing the logistics needed for development of unconventional resource. Figure 4.66 describes the ontology, based on the logistics descriptions given in Figure 4.67.

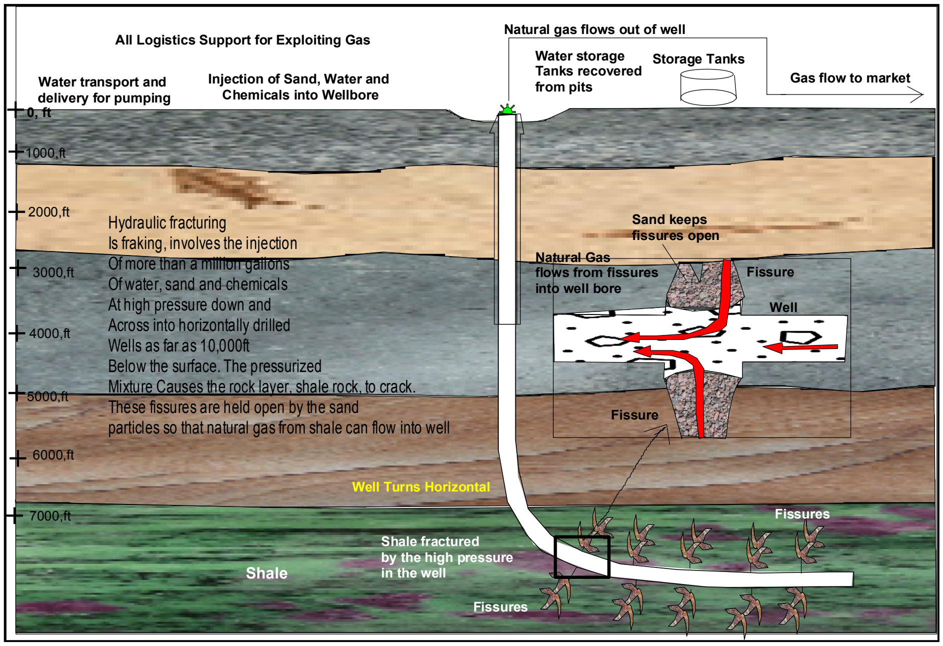

All Logistics Support for Exploiting Gas
Water transport and delivery for pumping
Injection of Sand, Water and Chemicals into Wellbore
Natural gas flows out of well
Water storage Tanks recovered from pits
Storage Tanks
Gas flow to market
0, ft
1000,ft
2000,ft
3000,ft
4000,ft
5000,ft
6000,ft
7000,ft
Hydraulic fracturing
Is fraking, involves the injection
Of more than a million gallons
Of water, sand and chemicals
At high pressure down and
Across into horizontally drilled
Wells as far as 10,000ft
Below the surface. The pressurized
Mixture Causes the rock layer, shale rock, to crack.
These fissures are held open by the sand
particles so that natural gas from shale can flow into well
Sand keeps fissures open
Natural Gas flows from fissures into well bore
Fissure
Well
Fissure
Well Turns Horizontal
Fissures
Shale fractured by the high pressure in the well
Shale
Fissures

Figure 4.67: Description of hydraulic fracturing and horizontal drilling dimensions

4.8 Petroleum Digital Ecosystem (PDE), a Digital Oil Field Solution

The research questions RQ1 – RQ7 and research objectives RO1 – RO7 describe events associated with data modelling and digital ecosystems' design and their use in the exploration project. The description of ontologies and integration of domain ontologies are lead to design and development of petroleum digital ecosystems to solve the problems (RQ7 – RQ8) of exploration and field development opportunities in oil and gas industries. The data sources from North West Shelf (NWS) in the Western Australia, such as *shelf, slope* and *deep* events are used for ecosystems modelling. The Romanian Offshore continental basin margins too narrate shelf, slope and deep (geological) events. Though ontology design ideas are originated from the literature on Entity-Relationship (ER) modeling and Object-Oriented (OO) designs, ontology development is different from designing entities or classes and relationships in ER and OO modeling. The author emphasizes involvement of programmer and ontology designer. The entities or classes are basic elements used in programming, which center primarily on entities or classes – programmer makes design decisions on the operational properties of entities or classes, whereas ontology designer makes these decisions based on structural properties of an entity or a class. As a result, an entity structure (or class structure) and relationships among entities or classes in an ontology, are different from the structure for a similar domain in the ER and OO model designs.

4.8.1 Integration of data sources of onshore and offshore basins

RQ1 – RQ3 and RO1 – RO2 are addressed in this section for integrating data sources from onshore and offshore regions. As demonstrated in the Figure 4.68, the continental basin shelf, slope and deep (geological) events are typical in any typical petroleum system scenario, where onshore and offshore data sources are connected and integrated through ontological descriptions and warehoused metadata. Examples from Romanian offshore and North West Shelf basins of the Western Australia are discussed with the following sections:

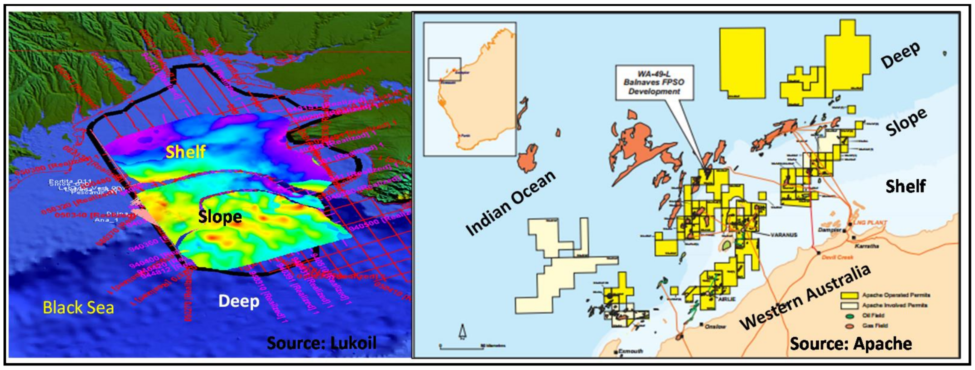


Figure 4.68: *Shelf*, *slope* and *deep* (geological) events showing the need of their connectivity (Romanian Offshore Basin and North West Shelf, Australia)

What are petroleum system and class representation?

An occurrence of a petroleum in any basin is a combined presence of *reservoir*, *structure*, *source rock maturity*, *seal*, *migration of petroleum* and *timing* of petroleum accumulated in reservoir rocks, which are elements and processes, and is termed as a total package of a petroleum system. Petroleum explorers must ensure the presence of these elements in any basin and if any of these elements is missing, explorers must invest in understanding the petroleum system and it is important that all geoscientists and petroleum investors make sure a complete understanding of each and every element of petroleum system. Author attempts to view the petroleum system in a different perspective for exploring its effectiveness using different ontology based data warehousing and data mining technologies. Petroleum system's potentiality depends on its effectiveness. In the present study, all the petroleum elements, such as *reservoir*, *structure*, *seal*, *source maturity*, *migration* and *timing* are considered as classes of super class *petroleum system*. Other classes, such as *surveys*, *wells* and *permits* are represented in *exploration* super class. *Production* is a key sub-class represented under E&P class. For prospect analysis in the NWS, integration of multiple dimensions in different domains is a prerequisite, as shown in the Figure 4.69.

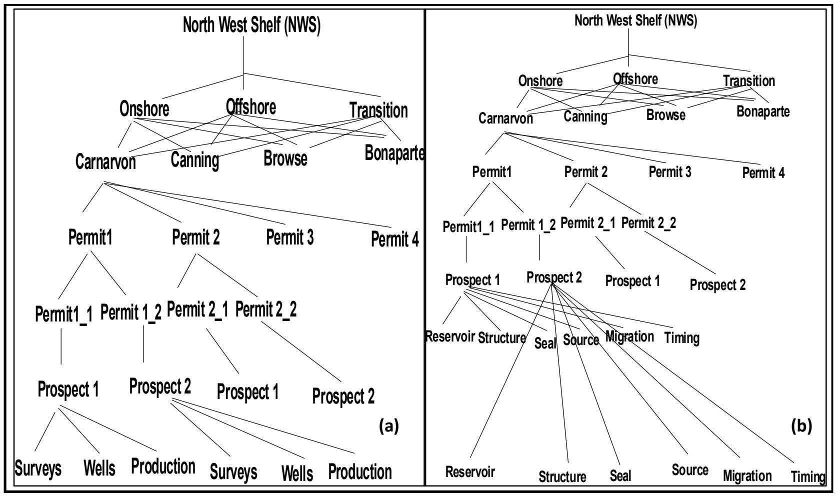


Figure 4.69: Conceptualized models showing basin – field (oil/gas) connectivity in North West Shelf (NWS) system

What is North West Shelf (NWS) and Why to model these events?

As described in Figure 4.68, NWS is a Mesozoic intracratonic basin, developed into a rift/drift margin of approximately 2400Kms long and 140Kms wide in the northwest of Australia. It is divided into different segments (sub-basins), Northern Carnarvon, Offshore Canning, Browse and Bonaparte basins. Much of the undeveloped oil and gas along NWS are large gas fields (of the order of 88 tcf reserves), which are the focus of exploration and development at early 21st century. The largest nine undeveloped fields in these basins form nuclei for further exploration. They are Gorgon (North Carnarvon), Scott Reef (Browse), Sunrise (Bonaparte), Troubadour (Bonaparte), Brecknock (Browse), Scarborough (North Carnarvon), Bayu-Undan (Bonaparte), Chrysaor (North Carnarvon) and Dionysis (North Carnarvon) fields. There are many undiscovered reserves, risk free in the NWS particularly from deep-water prospects. Many large structures are still under explored. Petrel sub-basin of the Bonaparte basin is under explored, where a proven Paleozoic oil-charge system exists beyond the limits of current oil discoveries.

The primary geological risk in these frontier areas is a combination of presence of *reservoir*, *source maturity* and *seal* effectiveness (elements of a system). In spite that areas along western flanks of Barrow-Dampier sub-basin (in the Western Australia), are recently proven discoveries. However, the under-explored deep-water area is huge and the entire untested basins are located in the Northern Beagle sub-basin of the Carnarvon basin, offshore canning and Browse basins (Figures 2.1 and 2.2, narrated in Chapter2).

Why develop an Ontology for NWS Petroleum Systems (as per research question and objective RQ1 and RO1 described in Section 1.3.1 and Section 1.3.2 in Chapter 1)

In recent years, the development of ontologies, which is explicit formal specification of the classes of petroleum systems of NWS in different application domains and relationships among them, has been moving from the realm of artificial intelligence (AI) to desktops of domain experts. Ontologies have become very common on World Wide Web (WWW). The present work has a scope to develop Petroleum Oil and gas Description Framework, a language for encoding knowledge on web pages to make it understandable to electronic agents searching for information. It is a process model for understanding the petroleum systems that can help the knowledge builder to easily extract knowledge of petroleum system. The ontologies of the petroleum systems range from large taxonomies categorizing petroleum systems to categorizations of basins, each field in each basin, drilled-wells in each field and their hierarchies and relational properties. Many areas now develop standardized ontologies that domain experts can use to share and annotate information in their fields. Presence of *structure*, *reservoir*, *source-maturity* and *seal* with petroleum system process elements, such as *timing*, *migration* and *accumulation* are better understood if these are represented as classes and studied there relationships through conceptual models (ontology perspective) that translate into knowledge-domain petroleum system models.

Role of ontology in the NWS

The *shelf*, *slope* and *deep* are specific geological events in the continental basin areas. The author is of the opinion, these events are interconnected and can be not isolated. These events are modelled to understand the role of connectivity, their integration for basin history and prospectivity in the basin margin areas (Figure 4.68). There are multiple domains in NWS, which are needed for integration. An ontology defines a common vocabulary for researchers who need to share information in different domains. Data integration is a challenge in developing a shared ontology. It includes machine-interpretable definitions of basic classes (concepts) in different domains and their respective relationships among them. Key research goals described in Section

1.4.4 as given in Chapter 1 are extended to NWS, Australia. To analyze the domain knowledge in NWS similar data mining, visualization and interpretation methodologies are considered addressing research objectives, RO 3 and RO 4 of Section 1.3.1 in Chapter 1. A declarative specification of the entities and attributes must be made available. Formal analysis of these entities is extremely valuable when both attempting to reuse of existing ontologies and extending them to other applications in NWS basins.

Conceptual modelling

The author uses conceptual models and framework discussed in Figures 3.36, 3.37, 3.38 and 4.69 in the NWS modelling and generating a metadata from integrated frameworks.

Significance and scope of ontology modelling

Data integration and shared logical data structures have proved to be successful in many financial and health science applications. Integration of shared ontologies can simplify the Western Australian heterogeneous petroleum data sources for building the knowledge from unknown geological systems. The Canning offshore and onshore petroleum systems, where exploration success is poor (possibly because of poor understanding of certain domain knowledge), are in the vicinity of high quality and proven discoveries. The ontology approach can facilitate to connect these petroleum systems, connecting their classes, sub-classes and add more information to the already existing pool of knowledge. Explorers or industries involved in investing in these basins are direct in contact with petroleum fields and are using all significant information to explore the petroleum systems and invest in the commercial ventures. The present study is an attempt to integrate all the petroleum ontologies and facilitate for shared ontologies for multi-users. This can categorize priority areas of exploration and production to explorers and optimizing investments in promising ontologically derived petroleum domains.

If ontology is successfully implemented in NWS, returns on investments are huge. The North West Shelf Venture (NWSV) is Australia's the largest resource development project, with investment totaling more than AUD$12 billion since 1984. Based on huge gas and condensate fields on the North West Shelf of Western Australia, the Venture supplies natural gas to the domestic market in WA, liquefied natural gas (LNG) to Japan and condensate, crude oil and liquefied petroleum gas (LPG) to international markets, such as China and USA.

NWS ontology framework

After deciding what the author is going to use the ontology for, how detailed or general the ontology going to be, directs the author with many of the modeling decisions to the development of warehousing. Among several viable alternatives, which one would work better for projected problem solution, be more intuitive, more extensible, and more maintainable, needs to be worked out. The ontology is a model of represented in world of reality and the concepts in the ontology must reflect this reality. After defining the initial version of ontology, it is evaluated and debug by using it in applications or problem solving methods or by discussing it with experts in the field. As a result certainly the initial ontology must be revised. This process of iterative design does likely to continue through the entire lifecycle of the ontology. Author follows up for developing the petroleum ontology for NWS petroleum systems:

Determine the domain and scope of ontology

Defining a domain and its scope are key criteria in the development of ontology, while answering the following basic questions:

- What is the domain that the ontology will cover? - oil and gas domain
- For what, we are going to use the ontology? – domain descriptions and structuring
- For what type of questions, the information in the ontology should provide answers?
    - explore connections among multiple reservoirs (for example) of oil and gas fields
- Who will use it and maintain ontology? – data analysts and oil and gas explorers

The answers to these questions may change during the ontology-design process, but any given time; they may facilitate limiting the scope of the model. Considering the case of petroleum system of NWS, it is the domain of ontology. It is planned to use this ontology for the applications that suggest petroleum system elements are good combinations in different geological settings in different basins.

Naturally, the concepts (classes) describing different types of petroleum elements, major sub-basins, the notion of good combination of presence of these petroleum elements and also bad combination are sorted out in ontology. At the same time, it is likely that the ontology will include concepts (classes) for managing the inventory of petroleum system elements, such as *surveys*, *wells*, *permits* and *production,* since these classes are related to the classes of petroleum system elements. If the ontology designed is used to assist in natural – language processing of petroleum systems analysis in the PRDF, it is important to include synonyms and parts of speech information for concepts (classes) in ontology. If the ontology is used to help explorers or geoscientists to decide which petroleum system element is dominant, production rate of a particular well (in a derived permit area) representing this element must be included. If this concept or class is used for exploration and production purposes, basin acreage and its potentiality may be necessary. If the geoscientists or petroleum explorers maintain the ontology describe the domain in a model (language) that is different from the model (language) of ontology users, it is necessary to provide the mapping between these models (languages).

Interoperability

It is always worth considering what someone else has done and checking, if it can be refined and extended to the existing sources for the particular domain and task. Reusing existing ontologies may be a requirement if one system needs to interact with other applications of other petroleum systems defined in adjacent basins that have already been committed to particular ontologies or with controlled semantics. Many ontologies may be available in electronic form and can be imported into the PRDF that is being developed. Many knowledge base systems can import or export the ontology, in whichever the formalism it is expressed. Since there are no existing ontologies for petroleum systems, development of ontology is started right from scratch.

Enumerate important classes in the ontology

It is useful to write down a list of all terms (concepts or classes) to develop ontology. What are the classes and what properties do they have? Name of basin, names of petroleum system elements, surveys: survey id, survey area, survey type, type of exploration; well: well id, well name, type of well; permits: permit id, name of permit, type of permit; production: production id, type of production, prod rate and pressure, if sub-classes or concepts are available. Initially, it is important to get a comprehensive list of all super-class, its classes and sub-classes without worrying about overlap between classes or concepts they represent, relationships among classes or any properties that the concepts may have, or whether the concepts are classes or slots.

The next two steps are – developing the class hierarchy and defining properties of concepts (slots) – are closely related. It is hard to do one of them first and then do the other. Typically, a few definitions of concepts (classes) in the hierarchy are created and then continued by describing properties of these concepts (classes). These two steps are key steps of ontology-design process. The more complicated issues of design will be discussed in the following sections.

Class definition and the class hierarchy

There are several approaches in developing class hierarchy (Uschold and Gruninger 1996):

1. A *top-down* development process starts with the definition of the most general classes (concepts) in the domain and subsequent specialization of the classes (concepts). For example, for general concepts of petroleum system or basin, their corresponding classes can be created. Then the *basin* class can be specialized in to its subclasses, such as onshore, offshore or transition. Or another example is permit super-class is categorized into its subclasses such as exploration permit, well approval permit, production permit or pipeline permit. In the case of NWS petroleum system, a super-class, can be categorized into Carnarvon, Canning, Browse and Bonaparte petroleum systems as subclass. Further Carnarvon petroleum system can be categorized into prospect level subsystems as sub-classes.
2. A *bottom-up* development process starts with the definition of more specific classes in a hierarchy, with subsequent grouping of these classes into more general classes (concepts). For example, Gorgon or Scott Reef is a prospect representing a sub-class petroleum system within North Carnarvon basin as a general class petroleum system. Another representation is *oil well*, *gas well*, *condensate well* are subclasses of general class *wells.* Oil well from a particular geological formation may be another subclass of *oil well* class.
3. A *combination* development process is a combination of top-down and bottom-down approaches. First more salient concepts (classes) are defined, then generalized and specialized them appropriately. A few top-level concepts may be included with a few specific concepts, which can finally be related to a middle level concept. We may also generate all of the regional classes, thereby linking them with a number of middle-level concepts.

Though none of these methods is inherently applicable in practical situations (Ozkarahan 1999), the approach depends on the individual requirements and needs of classes to be identified in each domain. If a developer has a systematic top-down view of the domain, then it may be easier to use the top-down approach. The combination approach is often the easiest for many ontology developers, since the concepts "in the middle" tend to be the more descriptive concepts in the domain. For distinguishing the most general classification, the top-down approach may be feasible. If one starts with specific examples, the bottom-up approach may be more appropriate. Whichever the approach is chosen, firstly, concepts are defined as classes in the ontology and will become anchors of class hierarchy. The classes are organized in a hierarchical taxonomy and instance of one class may necessarily be an instance of some other classes. If a class of A is super class of B, then every instance of B is also an instance of A.

Define the properties of classes – slots

Once the classes are defined and the internal structure of concepts must be described. Each and every class, a property is described, such as quality of reservoir, type of structure and production rate in their respective classes. For each property in the list, we must determine which class it describes. These properties become slots attached to classes. There are numerous class properties in the heterogeneous petroleum data that can become slots in an ontology.

- Intrinsic properties, such as quality of reservoir, texture and color of geological unit
- Extrinsic properties such as well name, survey id, date of production
- Parts, if the class is structured; these can be both physical and abstract "parts" (such as ingredients of petroleum system, the connectivity between systems and basins.

In addition to these properties, all sub-classes of a class inherit the slot of that class and their associated properties must be defined. All the slots of the class will be inherited to all sub-classes. All the additional slots representing the classes and subclasses must be defined. A slot should be attached at the most general class that can have that property. For instance, petroleum type, its composition and production rate, its *PVT values* should be attached to the *production* class, since it is the most general class whose instances have *PVT* and *production rates.*

Define the facets of the slots

The slots have different facets describing the value type, allowed values, the number of values (cardinality), and other features of the values the slot can take. For example, the value of a *name* slot (as in "the name of reservoir") is one string. That is, *name* is a slot with value type String. A slot *produces* (as in "a basin produces petroleum from these reservoirs") can have multiple values and the values are instances of the class *Basin*. That is, produces is a slot with value type: Instance with: Basin as allowed class.

The author describes several other common facets in the following sections:

The slot cardinality defines how many values a slot can have. Some systems distinguish only between single cardinality (allowing at most one value) and multiple cardinality (allowing any number of values). A tectonics of a basin will be a single cardinality slot (a basin can have only one tectonics). Basins produced by a particular petroleum system, in a multiple-cardinality slot *produce* from a *petroleum system* class.

Some systems allow specification of a minimum and maximum cardinality to describe the number of slot values more precisely. Minimum cardinality of N means that a slot must have at least N values. For example, the reservoir slot of a basin has a minimum cardinality of 1: each basin is made of at least one variety of reservoir. Maximum cardinality of M means that a slot can have at most M values. The maximum cardinality for the reservoir slot for single variety of basin is 1:these basins are made from only one variety of reservoirs. Sometimes it may be useful to set the maximum cardinality to 0. This setting would indicate that the slot cannot have any values for a particular subclass.

Slot-value type

A value type facet describes what type of values can fill in the slot. Here is the list of the more common value types:

String: is the smallest value type which is used for slots such as name: the value is a simple string.
Number: describes (both floating and integer) slots with numeric values. For example, basin produces oil at the rate of, can have value type Float

Boolean slots: are simple yes-no flags. For example, if petroleum system exists (yes) or not (no) are used. Or source rocks if matured, say yes or immature, say no.

Enumerated slots: specify a list of specific allowed values for the slot. For example, petroleum production can take on three possible values: oil, gas or condensate.

Instance-type slots: allow definition of relationships between classes. Slots with value type Instance must also define a list of allowed classes from which the instances can come. For example, a slot *produces* for the class reservoir may have instances of the class Basin as its values.

Domain and range of slot

Allowed classes for slots of type instance are range of slots. In our case, the class basin is the range of the produces slot. Some systems allow restricting the range of a slot when the slot is attached for a particular class. The class to which the slot is attached or classes which property a slot describes, are called the domain of the slot. The reservoir class is the domain of the produces slot. In the systems where the slots are attached to classes, the classes to which the slot is attached usually constitute the domain of that slot. There is no need to specify the domain separately. The basic rules for determining a domain and range of a slot are:

Find most general classes or class that can be respectively the domain or the range for the slots. On the other hand, do not define a domain and range that is overly general: all the classes in the domain of a slot should be described by the slot and instances of all the classes in the range of a slot should be potential fillers for the slot. An overlay of a general class is observed, but one would want to choose a class that covers all fillers.

Instead of listing all possible subclasses of the basin class for the range of the produces slot, just list basin. At the same time, the range of the slot cannot be specified as THING – the most general class in an ontology. In more specific terms, *if a list of classes defining a range or a domain of a slot includes a class and its subclass, remove the subclass.*

If the range of the slot contains both basin and the reservoir class, since they are separate hierarchical classes, the reservoir class becomes a subclass of basin and therefore the slot range already implicitly includes it as well as all other subclasses of the basin class. *If a list of classes defining a range or a domain of a slot contains all subclasses of a class A, but not the class A itself, the range should contain only the class A and not the subclasses.* Instead of defining the range of the slot to include good reservoir, poor reservoir and no reservoir, the range can be limited to the class reservoir itself. *If a list of classes defining a range or a domain of a slot contains all but a few subclasses of a class A, consider if the class A would make a more appropriate range definition.*

In systems where attaching a slot to a class is the same as adding the class to the domain of the slot, the same rules apply to the slot attachment. In other words, it is made more generalized and ensured that each class, to which the class is attached, can indeed have the property that the slot represents. Fine-grained property slot can be attached to each of the reservoir classes representing poor reservoirs. However, since all poor reservoirs have the fine-grained property, the slot can be attached to a more general class of poor reservoirs. Generalizing the fine-grain property slot further would not be correct since fine-grain property is not described for good reservoirs.

Create instances

This is last step in creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires (1) choosing a class, (2) creating an individual instance of that class, and (3) filling in the slot values. For example, an individual instance *Birdrong Sandstone-Barrow Sub-basin- Permian - North Perth Basin* to represent a specific type of reservoir in that basin. *Birdrong Sandstone-Barrow Sub-basin- Permian - North Perth Basin* is an instance of the class *North Perth Basin* representing all North Perth Basin's reservoirs types. As an example, this instance has the following slot values:

• Basin/Field:	Carnarvon/North Rankin
• Field_Size:	100 sq km
• Reservoir_Quality:	High
• Type_Reservoir:	Sandstone
• Porosity (%):	30%
• Production:	100 million barrels
• NWS_Basin_Name:	Carnarvon
• Production_LifeSpan:	20 Years (existing known reserves)

Defining classes and a class hierarchy

The author describes a class of hierarchy in a chosen domain in this section. There could be many ways of representing classes for any given domain. The hierarchy depends on the possible uses of the ontology, the level of detail that is necessary for the application, personal preferences, and sometimes requirements for compatibility with other models or when integrating it with other classes in other domains. Certain basic guidelines must be followed in developing a class hierarchy. After defining a considerable number of new classes, it is helpful to stand back and check if the emerging hierarchy conforms to these guidelines.

Ensuring that the class hierarchy is correct

An “is-a” relationship

The class hierarchy represents an “is-a” relationship: a class A is a subclass of B if every instance of B is also an instance of A. For example, Barrow is a sub-basin of Carnarvon basin or coarse-grain reservoir is a sub-class of reservoir. Another way to think of taxonomic relationship is a “kind of relation”.

Author does Integration of Shared Petroleum Ontologies through integrated framework and hierarchical ontologies as described in Figure 3.37. Researchers add more information to the already existing pool of knowledge regarding a particular basin or field. More details on results and discussions of the modelling methodologies in NWS are given in Chapter 5.

4.8.2 Modelling petroleum digital ecosystems (RO1-RO7 focus)

To demonstrate RQ1 to RQ7 and RO1 to RO7 and improvise exploration and field development activities of oil and gas businesses, petroleum digital ecosystems’ opportunities (in the contexts of multiple sedimentary basins) are explored. An ecosystem refers to an interdependent group of natural elements and entities that exist in a particular environment and the habitat with which these elements interact. Ecosystems are important because they sustain the natural world, providing humans with the oil and gas that are required in order to live and thrive. An ecosystem is a unit of the biosphere that has the functional components necessary to sustain itself, though there is sometimes significant interchange between ecosystems that exist next to each other (Davidson et al. 1995 and Karp 1995). When adjacent ecosystems interact, they share material and energy. If one ecosystem collapses, it could take the surrounding ecosystems with it. This is especially the case when man-made ecosystems are involved (such as urban ecosystems, croplands and farms), in which case the natural balance has been altered by humans. Similar to freshwater ecosystems, ocean ecosystems and terrestrial ecosystems, petroleum system is an ecosystem, a collective entity, in which several elements and processes interact and communicate both geographically and also in periodical dimensions. In our current case study, the author analyses several elements and processes (of different petroleum ecosystems) involved in geographic and periodic interaction among sub-systems used in building relationships using ontology conceptualization, specifications and contextualization.

In general, the narration of petroleum system of any basin is complex and several elements, making up the system, possess variety of attributes and each attribute is characteristic in its representation, classification and conceptualization. Investigation of a sedimentary basin is made and often interpreted categorizing and classifying groups of structures and reservoirs of petroleum systems. The contextualization is significant, which has been ignored in several knowledge domains of a petroleum system. Hierarchy is similar to any information system's representation. Each petroleum system comprises of group of hydrocarbon fields making up the petroleum system. Nimmagadda et al. 2010 narrate relationships among different entities and attributes are drawn to build ontology conceptual models. Each hydrocarbon field is again composed of groups of structural and or stratigraphic plays making up the hydrocarbon field and then finally interpreting the prospect in a particular knowledge domain describing its economic leverage. Each petroleum system, within a broader Sedimentary Basin (on a global scale, Total Petroleum System, TPS) is an information system. All the elements of the local petroleum system share their attributes and strengths with elements of other petroleum systems (extendable to the TPS).

Number of sedimentary basins exist in Southeast Asia, similar to Westralian Super basin in Australia. Indonesian basins are highly productive and hundreds of such basins exist with wealth of exploration, drilling, production and marketing data and their instances. Several data models are deduced that represent ontology, narrating relationships among petroleum system elements of different basins. Data structures or schemas are modelled in different star-schemas. In these data schemas, there are multiple dimensions narrated and interpreted conceptually with relationships among several dimensions and fact data tables, physically. It is important to describe intelligently and design multiple dimensions in a way to understand or increase domain knowledge that is unknown. For example, the knowledge built among fields, API, reservoir quality, production rate, is significant in understanding knowledge of particular petroleum systems, producing for long period of time. The author constructs ontology models for *shallow marine* and *deltaic* petroleum systems, in which several dimensions are described, connecting the factual data with one-to-many relationships in star-schemas.

Data modelling approaches

The author designs data schemas simple and flexible enough to modify as per the geological situations and users requirements. If the data schemas are too large, complexity is a significant barrier to widespread adoption of the warehouse technology, because users will find the schema so difficult to understand that they will be unable to write queries and application programs. Schemas must be evolved to grow and support new data types. Limiting the scalability to more data sources though has definite advantage, but in petroleum exploration industry, data structures are very large size. The warehouse schema should facilitate scalability and coercion of different sources of the same type of data into a common semantic framework. If there is no common information of attribute or data property among data dimensions, then different DBs exist side by side within a warehouse environment, without having merged. Once common data property information explores and connects, DBs get merged. Several star-schema models are merged as shown in Figures 4.70 and 4.71, with common logical key attributes among the participating dimensions in the structuring process.

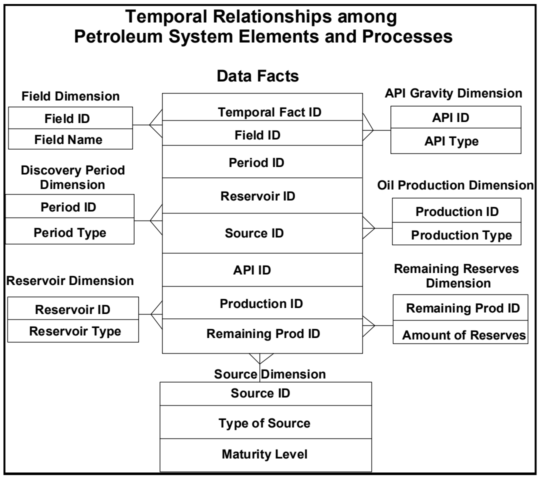


Figure 4.70: A star schema model representing relationships among *source rock, API, production, period* dimensions and their data facts

Data attributes and instances are used in the models as shown in Figures 4.70 and 4.71, for an integrated framework. For this purpose, data instances are documented and tabulated in different rows and columns of tables. These dimensional and factual data are loaded in an Oracle database program as explained in the following sections A and B:

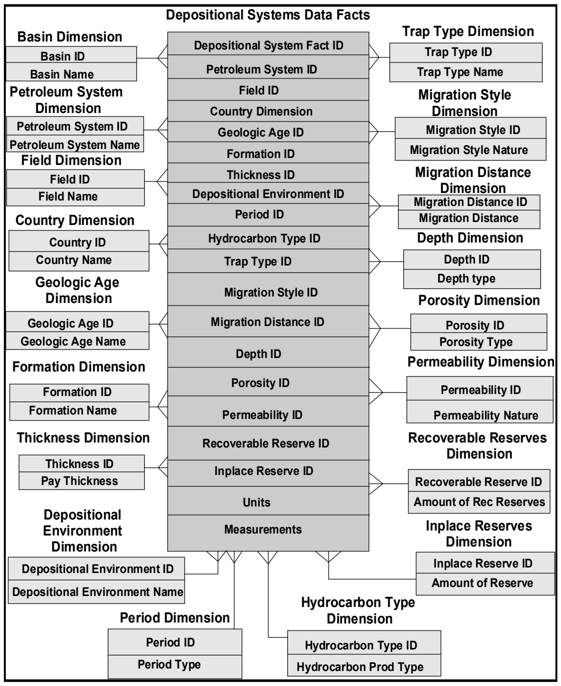


Figure 4.71: A star schema model representing relationships and their connectivity among various attributes of *depositional* systems of *basins*

A. Loader requirements

Because basins and petroleum systems are in large size and their associated data are very large in general. Because of poorly defined syntax and load failures are frequent, could result crashing of loaders. For this reason, DB loaders DB loaders should be able to recover gracefully from errors encountered during parsing of their input files. The loaders are designed to keep loading even in the presence of an error. If partial data has been inserted into the warehouse, a load error maintained on the related dimensions is updated to indicate that an error occurred while parsing the dimension and that the warehouse entry for the dimension may therefore be incomplete or contain errors.

B. Petroleum ecosystems warehouse design

A warehouse schema is designed by first analysing the schemas of each DB or super-type dimension to be integrated (Nimmagadda et al. 2005c and Rudra and Nimmagadda 2005), and connected with the schemas of other DBs that use the same data type and semantics. The development of warehouse is guided by several principles DBs schemas for sub-sets, each having similar data type or property. Since DBs typically conceptualize ingredients of petroleum systems in different ways, any kind of cross-DB operation faces the problem of semantic heterogeneity, whereby, information partitioned in different fields use different definitions (such as different units of measure). One possible approach to supporting petroleum system DBs within warehouse would be to create different schema definitions for each of the conceptualization (domain ontology) of petroleum systems used by the source DBs. Though this approach may achieve the semantic heterogeneity among these DBs, but may be designed not complicating the overall resulting integrated schema of the petroleum systems warehouse. As narrated in Figures 3.36 and 3.37, the author designs an integrated framework, explaining data structuring and integration process.

A typical hierarchical structuring may be useful in deducing relationships not only connecting different data dimensions within a petroleum system, but also among different petroleum systems. By and large, this warehouse may typically contain a different petroleum system for every petroleum system DB that it loads. With the result, the warehouse schema could be larger and more complex, and users would have more difficulty understanding the schema, making it more difficult for a user to query the DB, in which case, queries are made for all petroleum systems, accommodated in the warehouse, and essentially write a separate sub-query for each sub-schema.

In this approach, a single set of schema definitions covers a given data type, even if that data type is ontologically conceptualized differently in different DBs. For example, the author has created a single set of schema definitions to span all attributes for petroleum systems that possess similar structural and reservoir characteristics of producing horizons. Another example may be a single set of schema definitions for petroleum systems that have similar geo-chemical compositions of hydrocarbons. The DB loaders are responsible for translating from the ontology conceptualization (specifications), used in each DB within the family to the domain conceptualization used by the warehouse. This approach eliminates the semantic heterogeneity of these multiple DBs, allowing users to query all petroleum systems DBs using the same schema. This approach ensures the encoding of the dataset from which each data dimension within the warehouse is derived. Since entries from any petroleum system DBs are loaded into the same set of tables (multi-dimensional schema definitions), it is critical for user queries (data views for interpretation) to be able to distinguish different petroleum bearing basins in the Middle East, or basins in the western hemisphere or eastern hemisphere. Thus queries are made to extract data views for all warehouse shallow marine and or deltaic depositional system DBs that were loaded. Systems are simulated and developed for the following different sedimentary basins.

4.8.2.1 The Australian case study

The author has analysed four different case studies validating my research ideas and appropriately quoting the research objectives. Ultimately, author addresses the research goals, how author achieves them in narrating and implementing the constructs, models and methodologies in multiple sedimentary basins under varied geological and geophysical conditions. I have reviewed these case studies and reworded sentences. Author has quoted and referred the research objectives (RO1-RO7) in each case study from pages 222-297.

The constructs and models are built and used in this case study. As demonstrated in Figures 2.1a and 2.1b in Chapter 2, the author depicts several sedimentary basins and volumes of heterogeneous and multidimensional data sources available in these basins are used in constructing data structures (following up the research objectives, RO1 – RO3, explained in Section 1.3.1 of Chapter 1). For the purpose of connecting different petroleum systems within multiple basins, multidimensional integrated framework with warehousing approach is used. Periodic dimension is focused in the modeling and data warehouse design, analyzing time-series historical data of oil and gas data sources.

Data acquisition methodologies

Data sources collected among large sedimentary basins of Australia (geographic) and over 70 years of time (longitudinal), is a requirement of data design collection. The author collects secondary data on expenditure involved in the exploration of mineral and petroleum oil and gas in different basins in the Australian situations. The Australian Bureau of Statistics, Western Australian Department of Resource and Industry, South Australian Department of Primary Industry and Resource and Library of Curtin University have provided these data and information on sales and purchases, performance indicators, products and services of different oil and gas companies. Data collected from different geographic locations in Australia and over considerable time periods, are used to construct a metadata for qualitative and quantitative interpretation views. In addition to the Western Australian mineral and petroleum oil and gas business sources, data sources are from Western Uganda (Albertine-Graben), Middle Eastern basins, Indian Peninsula, Indonesia, USA, Romanian offshore basins.

Description of Australian petroleum data sources

As shown in Figures 2.1 and 2.2 in Chapter 2, there are several data sources to capture volumes of petroleum data in digital form. After having acquired these data volumes, warehouse designer identifies entities, dimensions and objects in their respective domains with hundreds of tables. Often these data are in spatial-temporal form. Spatial data represented in the form of X, Y, Z coordinates and historical data are in periods.

Logical storage of oil and gas data sources

Data warehouse storage area is a collection of relational database tables. Each of these tables will hold a different subset of the information that people want to access to. There are certain characteristics that data warehouse tables share that are not common to other types of system. The warehouse usually consists of a specially identified hardware platform that runs a specific database software product. Into this environment all the information and data are loaded and stored. Often all the data stored in the warehouses are in the form of multidimensional structures. There is huge demand and requirement (Nimmagadda and Rudra 2004 and Winter and Strauch 2003) for quality of information in a distributed environments (Jukic and Lang 2004). In Australian case study, operational data, such as from exploration, drilling, production and marketing departments possess hundreds of dimensions with similar number of fact tables.

The petroleum bearing sedimentary basins of Western Australia are highly prospective and productive compared to other Australian petroleum provinces. Historical exploration and production data in Canning, Carnarvon, Bonaparte, Browse and Perth basins (Figures 2.1 and 4.72) are available in different formats, often in duplicate. Understanding the prospect and petroleum system of a basin are significant issues. Data integration and sharing of data among different fields or prospects of different basins are other key issues of the present problem. Solution of this technical problem has great impact on the health and economics of the drilled well. Data-warehousing and data mining call for addressing these issues and analyze them for knowledge building purposes through data-mining.

In the Western Australia, oil and gas industries handle complex and large volumes of data with numerous dimensions and attributes with multiple associations. The author simplifies these data structures into logical and physical schemas, so that volume of data views generated in response to queries in a short period of time, are fast and precise for quality business decisions. Multidimensional logical models can also be designed for petroleum exploration and production data through ontological process. In order to achieve this, one needs to effectively and logically design data warehouse utilizing different basinal data (Figure 4.72) for oil and gas industry's exploration. Various dimensions and attributes for oil and gas industries have to be identified and analyzed for data modelling purposes.

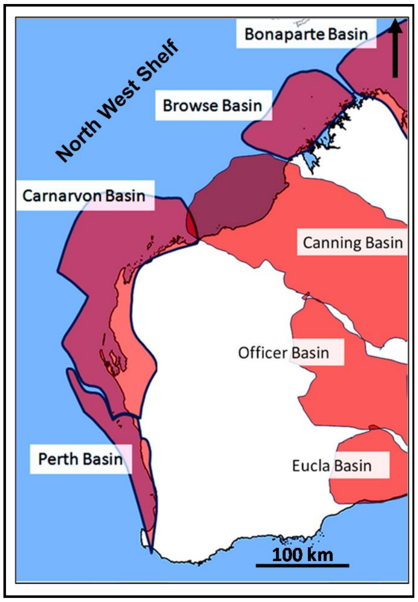


Figure 4.72: Map of Western Australia showing basinal data sources for oil and gas producing fields, wells and permits (www.doir.wa.gov.au)

The author identifies and uses super-type entities such as exploration, drilling and production for multidimensional data modelling and various other associated sub-type entities, attributes and their key indices. Logical ER (entity-relationship) and multidimensional models are basis of the data warehouse design and development, which are discussed in the fourth coming sections.

Data structuring methodologies

The logical ER, multidimensional and object oriented data structuring models used in the present study are investigated in the following sections for their application in petroleum industry.

Exploration, drilling, production and marketing are considered as super type entities, participating in the conceptual models (as shown in Figures 4.1, 4.2 and 4.5 and 4.6). Exploration, as a super entity has several sub-type entities, such as geology, geophysics, well logging, reservoir, logistics and inventory are few to mention. Similar sub-type entities can be derived from drilling, production and marketing operational data entities. An ontological framework can be derived using overall petroleum business data to maintain and represent consistent semantic information among these data entities. Various data entities and attributes identified for petroleum exploration and production have been conceptualized in graphical form so that all the associated data entities and their relationships are explicitly understood.

Ontologies are used for building relationships and design petroleum data structures in different knowledge domains (Meersman 2004) and it is a basis for incremental design of logical data schemas such as ER, EER and MR and MMR in Australian situations. Schemas and components of schemas have been structured using entity-relationship (ER) constructs and relational data theory. Good understanding of conceptual data structuring (as narrated in Figures 4.1 and 4.2) is required for designing quality logical ER models. Two of such conceptual models have been narrated, based on which the logical schemas are generated. The author shows a conceptual model and a sample of an ER diagram for petroleum exploration and production industry in Figures 4.1 and 4.2, narrating data model constructs**.**

Multidimensional logical warehouse schemas

A star schema expresses data as collection of facts and dimensions. Unlike relational schema, which is flat, a star schema is a graphical view of the data (Hoffer et al. 2005 and Coronel 2011). It is a simple database design in which the dimensional data (describing how data are commonly aggregated) are separated from fact or event data (that describe individual business transactions). Another name that is often used is the dimensional model. This schema is suited to ad-hoc queries (and other forms of informational processing); it is not suited to on-line transaction processing and therefore is not generally used in operational systems, operational data stores, or an EDW. Australian multidimensional oil and gas data use all facts, dimensions and measures. Facts are business data items, transactions, or events used to analyze the business process. Facts are stored in core tables that hold numerical information. Dimensions determine the contextual information for the facts. Each data point in the fact table is associated with one dimension. Each dimension comprises members of unique data points. For example, location dimension (in oil and gas industry data) may consist of the basin under study, its identity, location in state and country. Figures 3.36 – 3.38 demonstrate how dimensions can be organized into hierarchical structures. The location dimension can be constructed from basin data, dealt with in a state. Measures are numeric value of facts. Each is a variable and represents of the business in relation to the dimension.

Visualizing dimensional models through star schemas is increasingly common in the oil and gas warehouse. Star schemas provide a query-centric view of the data. They are dimensional models that rely on classifying information as either facts or dimensions. The query searches for specific fact qualified by a set of dimensions. In this way analysts avoid navigating a maze of interrelated tables to access the desired information. The author emphasizes that for many oil and gas users, star schemas, are more intuitive ways to visualize the data structures.

Fact tables and dimension tables: A star schema consists of two types of tables: fact tables and dimension tables. Fact tables contain factual or quantitative data about a business such as units sold, orders booked and so on. The dimension tables hold descriptive data about the subject of the business. The dimension tables are usually the source of attributes used to qualify, categorize or summarize the facts in queries, reports, or graphs. The simplest star schema consists of one fact table, surrounded by several dimension tables. Figure 4.73 demonstrates a typical star schema models describing exploration costs, with four dimensions and one fact table. Typical business dimensions for mineral exploration are *state*, *mineral*, *discovery* and *period* as shown in Figure 4.74. The *period* is always one of the dimensions, enabling the historical data presented for extracting trends in the data and interpreting the trends in meaningful form.

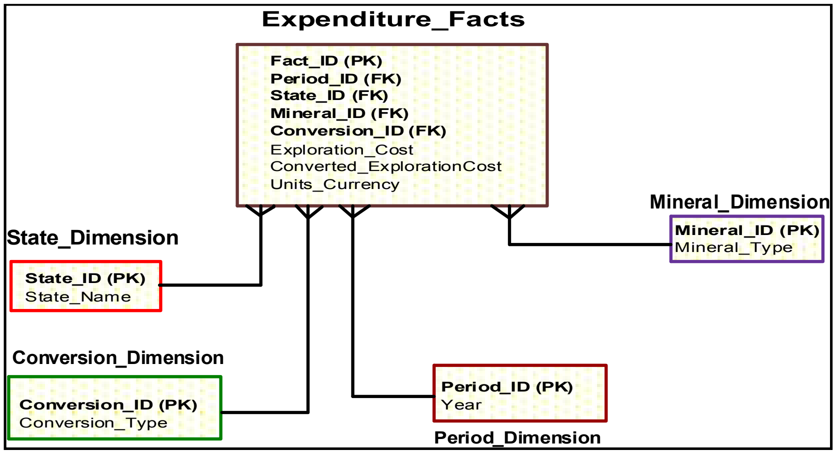


Figure 4.73: Multidimensional star schema for mineral *exploration costs* dimension

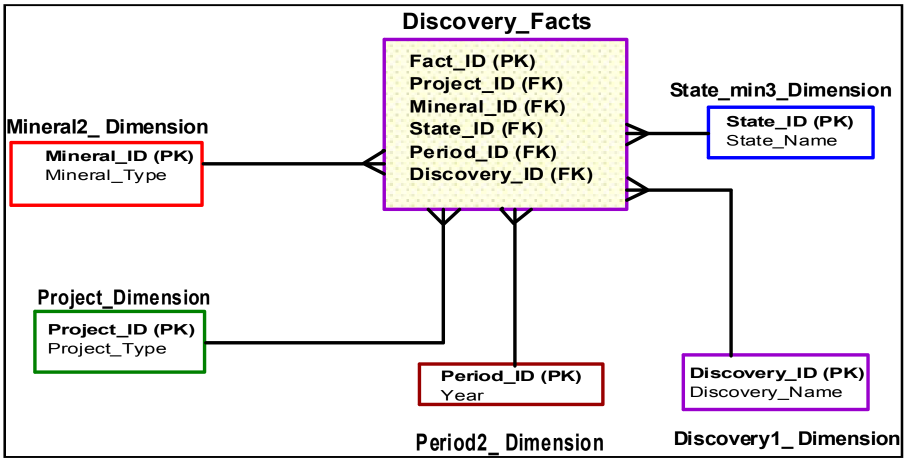


Figure 4.74: Multidimensional star schema for mineral *discovery* dimension

Each dimension table has a one-to-many relationship to the central fact table. Each dimension table generally has a simple primary key, as well as several non-key attributes. The primary key in turn is a foreign key in the fact table. The primary key of the fact table is a composite key that consists of the concatenation of all the foreign keys, plus possibly other components that do not correspond to the dimensions. The relationship between each dimension table and the fact table provides a join path that allows the users to query the database easily, using SQL statements for either predefined or ad-hoc queries. Non-key attributes are generally called data columns. The fact table plays the role of an n-array associative entity that links the instances of the various dimensions as shown in Figures 4.73 - 4.75. Expl_Expenditure_Facts, Lease_Facts, Petro_Production_Facts and Statewise_Expl_Cost_Facts are fact tables surrounded by dimensions tables linked by primary and foreign keys as narrated in Figure 4.75.

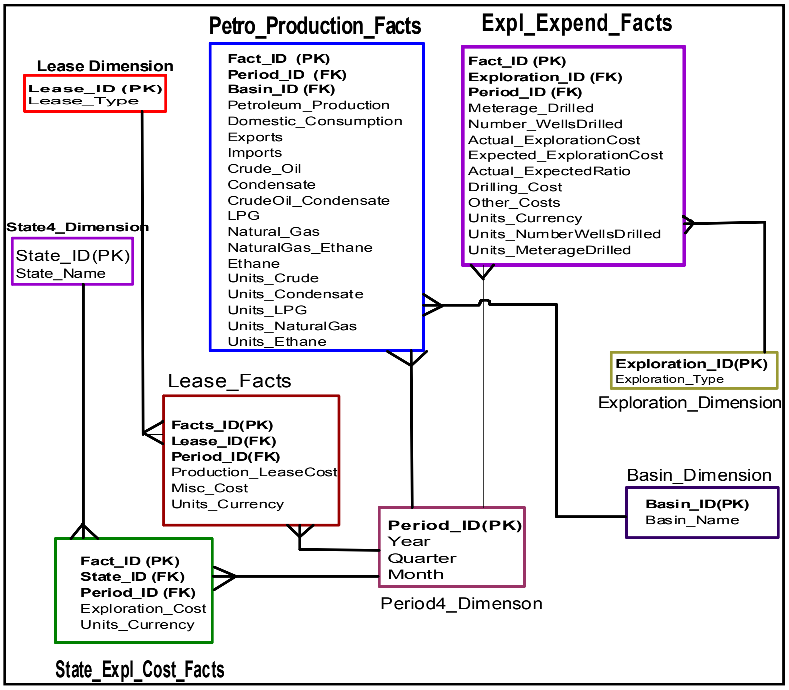


Figure 4.75: Multidimensional star schema for petro1 (database) - with *exploration* and *production* dimensions

Indexing: With multidimensional data, indices help to reduce the overhead of scanning the extremely large tables. Exploration and production data of a oil and gas company consist of volume of data with hundreds of dimension and fact tables. When used for multidimensional data, a common approach creates a join index between a dimension table and a fact table.

Surrogate key: This is best-used key in the present database structures as shown in Figures 4.73 - 4.75. Surrogates are system generated identifiers used to remove ambiguities, occurring while referring conflicting entities or objects for implementation including database maintenance. They are permanent identifiers and never change during any stage of modelling. For example, a reservoir, called *Ratawi Limestone*, a unique reservoir occurring in multiple petroleum provinces of Middle Eastern region, which never changes its identification. Every key used to join the fact table with a dimension table should be a surrogate (non-intelligent or system assigned) key, not a key that uses a business value (sometimes called a smart key or a production key).

Multiple fact tables: It is often desirable for performance or other reasons to define more than one fact table in a given schema as shown in Figure 4.75. For example, various users require different levels of aggregation, such as exploration expenditure involved in a particular state for a particular mineral or oil or gas resource between certain periods. The performance can be improved by defining a different fact table for each level of aggregation. The obvious trade-off is that storage requirements may increase dramatically with each new fact table. More commonly, multiple fact tables are needed to store facts for different combinations of dimension.

A conformed dimension means the same thing with each fact table, and hence, uses the same surrogate primary keys. Even when two star schemas are stored in separate physical data marts, if dimensions are conformed, there is a potential for asking questions across the data marts. In general, conformed dimensions allow users to:

- Share non-key dimension data
- Query across fact tables with consistency
- Work on facts and business subjects for which all users have the same meaning.

Fact-less Fact Tables: There are applications for fact tables without any non-key data, only the foreign keys for the associated dimensions. In case of Min4 Database, there are all key attributes without facts as shown in Figure 4.76.

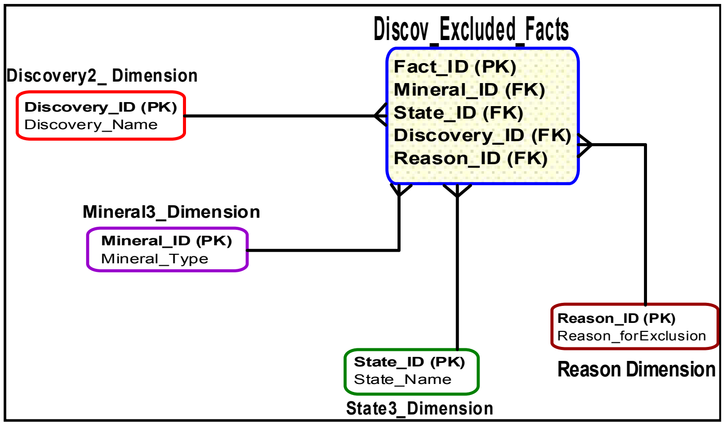


Figure 4.76: Multidimensional star schema model for attribute dimensions of Min4 database

Normalizing and denormalizing dimension tables:

Fact tables are normalized because each fact depends on the whole composite primary key and nothing but the composite key. On the other hand, dimension tables may not be normalized. Most data warehouse experts find this acceptable for a data mart optimized and simplified for a given user group, so that all of the dimension data are only one join away from associated facts. Sometimes, as with any relational database, the anomalies of a denormalized dimension table cause add, update and delete problems.

Multivalued Dimensions: There may be a need for facts to be qualified by a set of values for the same business subject. In the present study, Permit ID or Number is one of the dimensional attributes used commonly for three different fact tables. As indicated in Figure 4.77, “1” “2” and “3” represent the link to the surveys and wells fact tables (with Figure 4.78) associated with petro2 database.

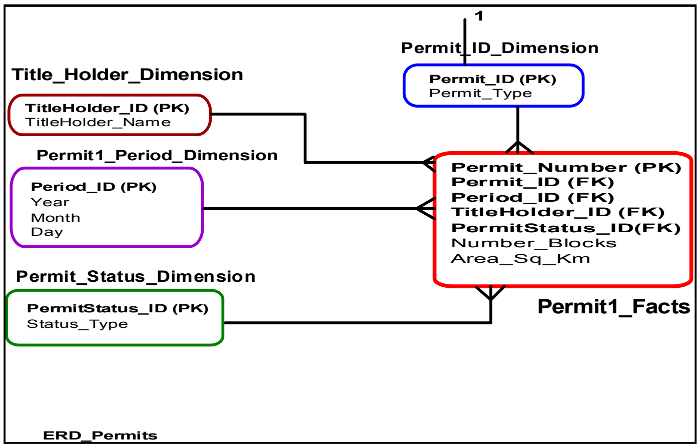


Figure 4.77: Multidimensional star schema model for petro2-*permits* dimension

Slow changing dimensions: The data warehouses and data marts track business activities over time. The business does not remain static over time – exploration costs change by exploration objectives, contractor relocation, basin exploration priority change, and exploration staff are assigned to different locations. Most systems of record keep only the current values for business subjects. But in a data warehouse or data mart, one needs to know the history of values to match the history of facts with correct dimensional descriptions at the time the facts happened. For example, a survey ID needs to be associated with the description of the associated company or contractor during the time period of the survey facts, which may not be the description of that company today. Business subjects’ change slowly compared to most operational transactions. Thus, the dimensional data change, but change slowly.

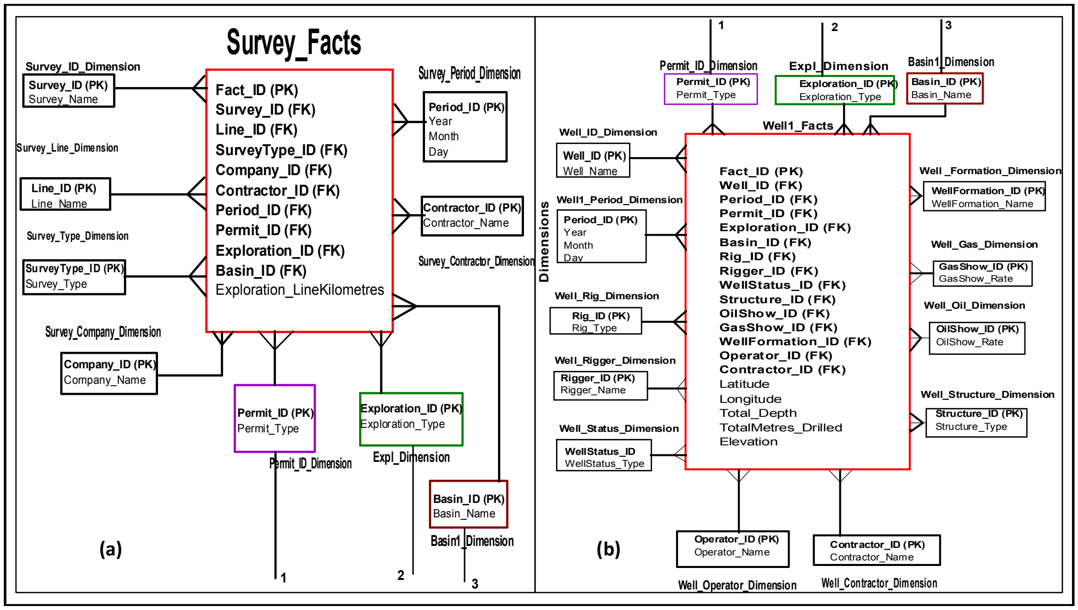


Figure 4.78: Star schema models for *surveys* and *wells* dimensions and their facts data instances

Multidimensional data schemas constructed for petroleum business data in a logical domain, lead to development of physical data models in implementation domain. At the core of the design of the data warehouse, lies a multidimensional view of the data model. Many statistical data models represented in tables (columns and rows) can be straightway used for building the multidimensional data models. These tables must have been already in relational or hierarchical or network representation, which are described with the existing knowledge. Author uses all the common attributes appearing in all these tables for building relationships among several dimensions and fact tables. In this case, relationships among the common attributes are denormalized, so that the final data become fine-grained for effective warehousing and mining purposes.

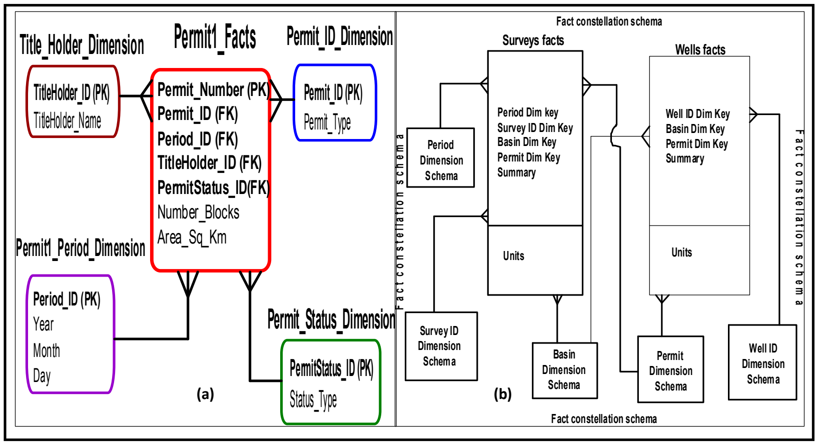


Figure 4.79: Schema styles of representing exploration data sources (a) star schema model (b) fact constellation schema

Similar to the entities in the ER modeling, the author uses dimensions in dimensional modeling. Figures 4.77-4.80 represent a petroleum exploration (surveys) and wells database, but chopped into three diagrams, due to presentation convenience. This multidimensional model consists of three major facts tables (surveys, wells and permit facts) surrounded by several dimensions. Being a fact constellation schema, high-level granularity has been maintained in order to derive fine-grained queries. Each dimension table has one-to-many relationship with a central fact table as shown in Figures 4.77 – 4.80, in a petroleum exploration and production situation. Each dimension table generally has a simple primary key, as well as several non-key attributes. The primary key in turn is a foreign key in the fact table. The primary key of the fact table is a composite key that consists of the concatenation of all the foreign keys, plus possibly other components that do not correspond to the dimensions. The relationship between each dimension table and the fact table provides a join path that allows the users to query the database easily, using SQL statements for either predefined or ad-hoc queries. Non-key attributes are generally called data columns. The fact table plays the role of an n-array associative entity that links the instances of the various dimensions.

The above *surveys, wells and permits* schemas are represented as star schemas. In fact, they contribute to the design and development of fact constellation schema, since 1, 2, 3 (as represented in Figures 4.78 – 4.79) share common dimensions with surveys and wells fact tables. These two are just examples. There could be many other fact tables that can share many other associated dimension tables. The primary keys (PK) represented in the dimension tables become foreign keys in the fact tables and fact table itself has a primary key attribute represented as surrogate key to maintain uniqueness in the data. As shown in Figure 134, petroleum database has been generated using fact constellation schema, containing two fact tables of surveys and wells and surrounded by period, basin, permits, survey ID and basin dimensions.

The petroleum production data dimensions and facts are modelled in the snowflake star schemas, in which relationships between common attributes of each entity are normalized as demonstrated in Figure 4.80. In the present context, to support attribute hierarchies, the dimension tables are normalized to create snowflake schemas. This consists of a single fact table and multiple dimension tables again. Like star schema, each tuple of the fact table consists of an attribute (foreign key) pointing to each of the dimension table that provides its multidimensional coordinates. It also stores numerical values (non-dimensional attributes) for these coordinates. An advantage of this normalized schema is easy in its usage and maintenance; normalizing also saves storage space, though navigation of petroleum data across multiple tables may not be effective due to large number of join operations.

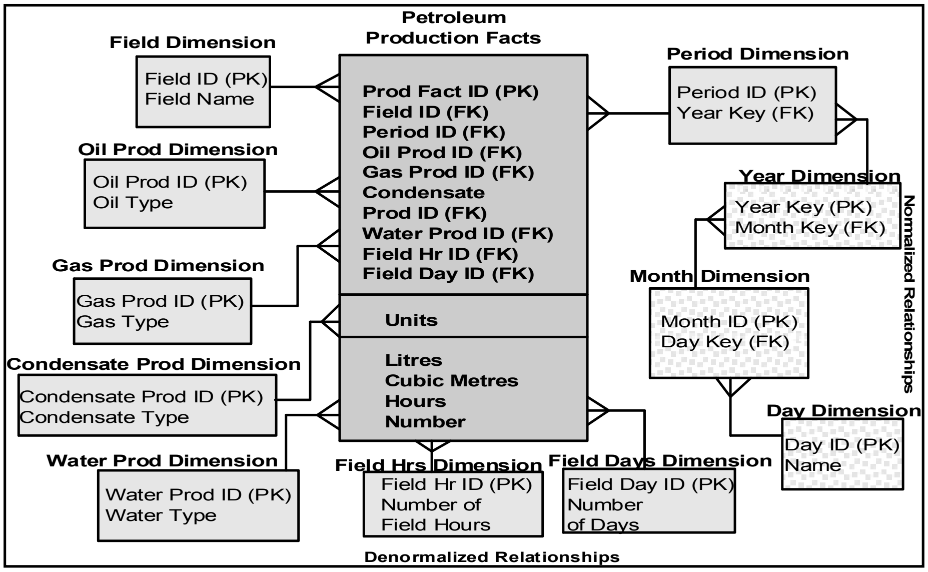


Figure 4.80: Petroleum production business data – snowflake schema model

Oil and gas data warehouse architectures

The data warehouses are electronic repositories of summarized exploration, drilling, production, personnel and administration data, often extracted from disparate departmental or project sites databases. It is interesting to observe that this model has three important components. Data acquisition, data storage and data mining contribute to the data warehouse design and development. Get all the company data working together, so explorationists, engineers, project managers can see more, learn more, and make the organization work better. Today majority of fortune oil and mineral companies are busy in constructing their data warehouses to serve as “networked oil & gas information service hub” for optimizing their business operations.

Grain of the fact table: The author keeps the raw data of a star schema in the fact table. Determining the lowest level of detailed fact data stored is arguably the most important data mart design step. The level of detail of this data is determined by the intersection of all the components of the primary key of the fact table. Determining the grain is critical and must be determined from business decision-making needs. There is always a way to summarize fact data by aggregating using dimension attributes, but there is no way in the data mart to understand business activity at a level of detail finer than the fact table grain. In the present study, data warehouse design considered, involves with three significant components, as discussed in the next section.

Components of data warehouse

The data warehouse is made up of three major components:

- The acquisition component – data may be acquired through any compatible software or hardware means.
- The storage component – the warehouse, stores databases from different operational business units.
- The access component – data mining tools used to access data and information and extract business intelligence through statistical or mathematical means.

Data warehouse design with periodic dimension

1. Distinguish between instantaneous and historical problems and derive an appropriate data model
2. Describe how temporal aspects affect the cardinality of relationships.
3. Recognize how changes in business rules affect instantaneous or historical views.
4. Apply appropriate mechanisms for storing historical data.

The cost of exploration, drilling and production depend on the cost per meter for exploration type, casing type used in the borehole, borehole depth, the daily hire rate charged by the contractor and the time taken to explore and drill the borehole. Sometimes more than one casing type is used on the same borehole. Same contractor may have several leases or permits in the same or different basins. Part of the contractual arrangements with the operators require oil and gas companies to pay various bonuses or loadings to the rig operators for each man day per borehole with no lost-time injuries. Each day, for which the drilling or mining of a borehole or mine is finished ahead of schedule. Each truck or rig used in the drilling or mining operations. Each lease is given a permit number. To achieve a more efficient operation, drilling rigs are moved between adjoining permits, where possible, when moving to drill a new borehole. Thus, there are innumerable data entities, items and values that can be identified from the oil and gas companies in the Western Australia. User’s involvement is necessary for achieving the satisfying results.

Why to design and develop data warehouse in space and time dimensions

Analyzing the petroleum and mineral exploration data dimensions with time dimension is a significant criterion for extracting useful information from past historical datasets. The author uses different data structure schemas in the current study to build relationships between period and other associated oil and gas data dimensions of Australian mineral industry situations. As shown in Figures 4.81-4.82, time dimension is made relevant to both exploration data and mineral discovery data facts, implying a period dimension is linked to different operations and activities of the oil and gas industry.

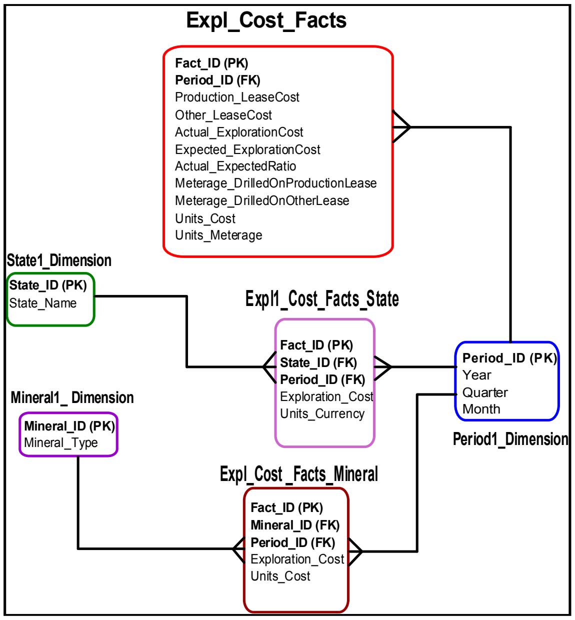


Figure 4.81: Star schema model – *exploration cost facts* with *periodic* dimension

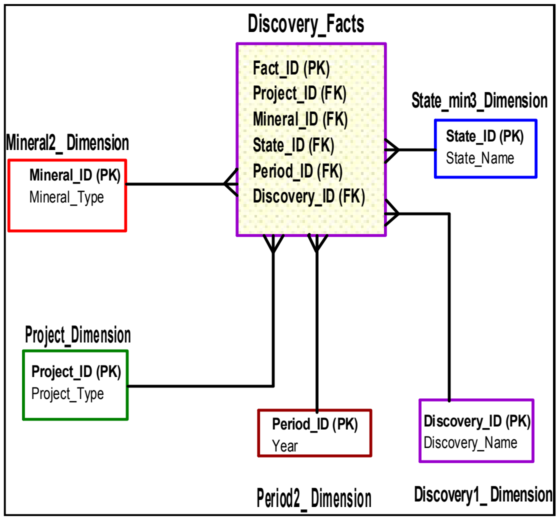


Figure 4.82: Star schema model with *discovery* facts showing the *space* and *time* dimensions

Periodic dimension of the multidimensional data sources

In this case study, the author uses data relevant to activities such as exploration, drilling and production in the Western Australian situations, for modelling data structures, thus a data warehouse containing various data structures of different business activities is evolved. These business functions and activities are transformed into many data dimensions in a star like schema structures. Time or period dimension is one of such dimensions, can be stored either in the form of period or detailed version of snowflake schema, representing, *day*, *month*, *quarter* and *year*. It is interesting to note that presenting the period dimension into multiple dimensions, particularly when an oil and gas company has to handle the day-to-day transactions. Data warehouse, thus presented in the present study is special multidimensional database, again disseminated into data marts, each aimed at accessing informational data and extracting periodic information in the form of business intelligence. Entity- relationship diagrams have been drawn for four Mineral Exploration and Production Databases and for two Petroleum Exploration and Production Databases. The present system includes modules for *well*, *lease*, *seismic*, *culture*, *assets* and *expenditure* data for petroleum exploration. The comprehensive well data module includes tables for tops, cores, tests, production, logs run, deviation surveys and well location. The seismic module uses records of the acquisition and processing history of seismic lines along with the location data.

Time domain data modeling

West (2006) describes a space-time map view within a 3D space. Time domain analysis and conceptual modelling of time-varying data have been discussed by (Gregersen and Jensen 1998). Author in the present context discusses the significance of time domain mapping of data within a data warehousing and multidimensional analysis (Coronel 2011) perspective. Gupta (1990) uses several statistical techniques to process the time domain data and identify the trends. The concepts and applications in the business statistical analysis are described in Berenson and Levine (1992). They provide literature on regression analysis and time domain modeling techniques. These techniques are briefly discussed in the forthcoming chapters.

A time series is a set of quantitative data that are obtained at regular periods over time. Monthly or quarterly expenditure are some of these examples. In time series analysis, the value of an attribute is examined as it varies over time. The values are obtained as evenly spaced time points (monthly, quarterly, yearly). A time series plot is used to visualize the time series.

There are two main goals of time-domain mapping:

a) Identifying the nature of the phenomenon represented by the sequence of observations in space and time and
b) Forecasting (predicting future values of the time series variable).

Both these goals require patterns of the observed time series data, identified and formally described. Factors influencing the patterns of activity in the past and present are expected to continue in the similar manner in the future. Once the pattern is established, it can be interpreted and integrated with other data. The identified pattern is extrapolated to predict the future events (forecasts). Previous studies suggest two general aspects of time series patterns and they are trend and seasonality. Trend represents a linear or non-linear component that changes over time and does not repeat or does not repeat within a time range captured by the data. Seasonality may have similar nature and repeats itself in systematic intervals over time. For example, expenditure and sales of a company and or their sister-companies can grow rapidly over years, but they still follow consistent seasonal patterns.

Currently, the author analyzes the data that contain a time dimension. Predictions are made with one or more time-dependent attributes using time series data. Typical applications include: tracking individual wells that produced both oil and gas over a period of time, predicting the neighborhood wells' status or predicting the opencast mining extension of gold vein deposit in a particular direction. The ability to succeed in these predictions depends on the availability stronger correlation, trends and patterns of relevant attributes and their instances. As discussed in Dunham (2003), a time series is a set of attribute values over a period of time. This may be discrete or continuous. Typical data mining applications for time series data include determining the similarity between two different time series and predicting future values for an attribute, given a time series of known values. The prediction is a type of classification, while similarity can be thought of as either clustering or classification. Given several time series samples, which time series are like each other in terms of clustering and or classification. A special type of similarity analysis is that of identifying patterns within time series. As described before, time series is viewed as finding patterns in the data and predicting future values. Detected patterns may include:

Oil and gas Data Trends: it is viewed as systematic non-repetitive changes to the attribute values over time. For example, investment or cost of exploration may increase when oil prices rise. Or with increase in oil production, there could be corresponding increase in exports.

Cycles: Response may be cyclic. Quarterly mineral exploration cost may be cyclic.

Seasonal: The response behavior within a month or quarter or year. In general, with the increase of surveys, and wells drilled in an area, there could be increase in the number of oil or gas producing wells.

Outliers: Techniques improve signal and noise ratios in the data. For example, surveyed data typically possess different sub-surface patterns, which are detected by data processing techniques. Neural networks can also predict the well plan prediction based on the supervised learning of the neighborhood oil and gas producing wells. Similarly reservoir quality can also be predicted with time domain petroleum production data.

Time (periodic) series analysis

In the time series analysis, it is assumed that data consist of systematic patterns and random noise, which usually makes the pattern difficult to identify and interpret. Some form of filtering is required to smooth the data and make the patterns appear in the data. There are no proven automatic techniques to identify trend components in the time domain data. As long as trend is monotonous (consistently increasing or decreasing), data analysis is not difficult. If the time series contains some errors, the first step in the process of trend identification is smoothing or weighted averaging.

Many monotonous time series data can be adequately approximated by a linear function; if there is a clear monotonous nonlinear component, the data first need to be transformed to remove the nonlinearity. Usually a logarithmic, exponential, or (less often) polynomial function can be used. The data analysis techniques and the functions used in the present study will be discussed in the subsequent chapters. Gregersen and Jensen (1998) discuss time domain analysis and conceptual modeling of time-varying data. Database contents vary over time. For example, in a database that contains product information, the unit price for each product may be changed as material and exploration costs and market conditions change. If only a current price is required, then *price* can be modeled as a single valued attribute. However for accounting, billing, and other purposes, there is need to preserve a history of all expenditure and the time period over which each was in effect. This can conceptualize this requirement as series of exploration costs and the effective date for each cost. This result is a (composite) multi-valued attribute named ExplorationCost_History. The components of ExplorationCost_History are ExplorationCost and Effective_Date (Figure 4.83). An important characteristic of such a composite, multivalued attribute is that the component attributes go together. Each value of the attribute ExplorationCost is time stamped with its effective date. A time stamp is simply a time value (such as date and time) that is associated with a data value. A time stamp (West 2006) may be associated with any data value that changes over time when we need to maintain a history of those data values. The author documents and reports time stamps to indicate the time the value was entered (transaction time), the time the value becomes valid or stops being valid, or the time when critical actions were performed (such as updates, corrections, or audits).

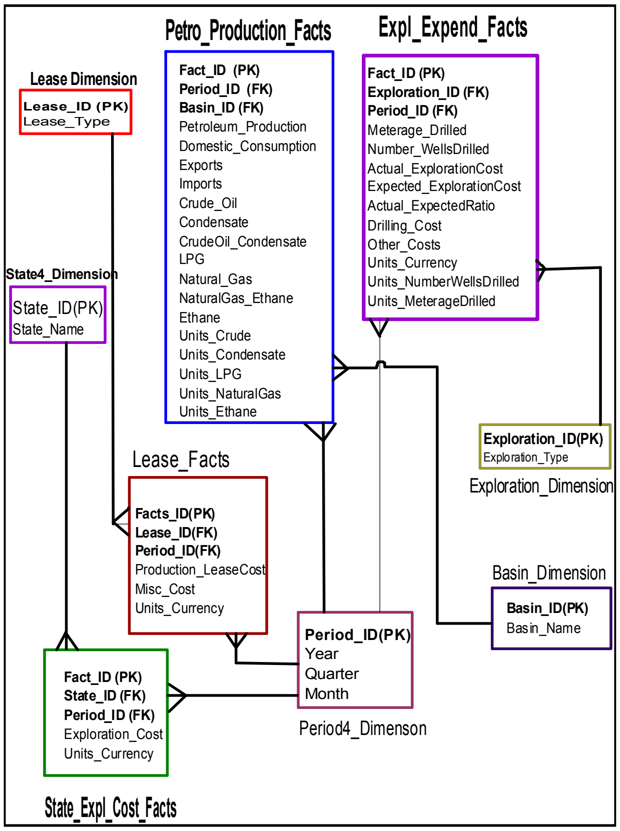


Figure 4.83: Multidimensional star modeling of *exploration-production* facts

The use of simple time stamping (West 2006) is often adequate for modeling the time-dependent data. However, time often introduces subtler complexities in data modeling. For example customer orders are processed throughout the year, and monthly summaries are reported by product line and by product within product line. The current data models are generally inadequate in handling time-dependent data, and that organizations often ignore this problem and hope that the resulting inaccuracies balance out. However, data warehousing applications are designed to remove many of these uncertainties by providing explicit designs for time-dependent data. One needs to be alert to the complexities posed by the time-dependent data as one develops data models in the organization.

There are three basic functions performed in the time series analysis. The “distance” measure is used to determine the similarity between different time series. Secondly, the structure of the line is examined to determine its behavior. The third application is use of historical time series plot to predict future values. Since it is an ongoing process of acquiring good quality data, the actual original data have been checked for missing data, irrelevant data and noisy data. At all stages, good quality data is ensured to maintain quality data mining or data exploration. Some burning issues of data design criteria, exporting and importing of data and data validity checks have been discussed in the following sections.

Modeling Date and Time: Because data marts record facts about dimensions over time, date and time is always a dimension table and a date surrogate key is always one of the components of the primary key of any fact table. Because a user may want to aggregate facts on many different aspects of date, a date dimension may have many non-key attributes. Also, because some characteristics of dates are country or event specific (e.g., whether the date is a holiday or there is some standard event on a given day, such as a festival or football game), modeling, the date dimension can be more complex than illustrated so far. As shown in Figures 4.84 and 4.85, a typical multidimensional date has been designed. A date surrogate key appears to be a primary key in the fact table and is the primary key of the date dimension table. The non-key attributes of the date dimension table include all the characteristics of dates that users use to categorize, summarize, and group facts that do not vary by country or event. For an oil and gas company doing business at several locations in Australia and in several countries, possess different characteristics of dates. The author adds a country calendar table to hold the characteristics of each date in each country. Thus the date key is a foreign key in the country calendar table, and each row of the country calendar table is unique by the combination of date key and country, which form the composite primary key for this table. A special event may occur on a given date (for simplicity, no more than one special event may occur on a given date), event data have been normalized by creating an event table, so each descriptive data on each event (e.g., Christmas, Good Friday etc…) are stored only once. In case of oil and gas industry, however, the historical data are either in date or monthly/quarterly or yearly. This is much-needed situation for data warehousing. Another snowflake star model with normalized period dimension is shown in Figure 4.84 associated with survey facts.

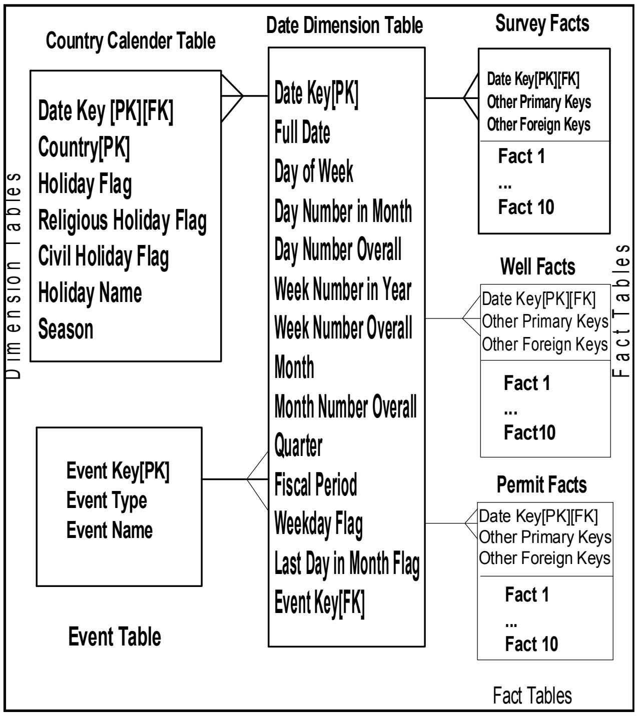


Figure 4.84: Modeling date attributes with *survey*, *well* and *permit* facts

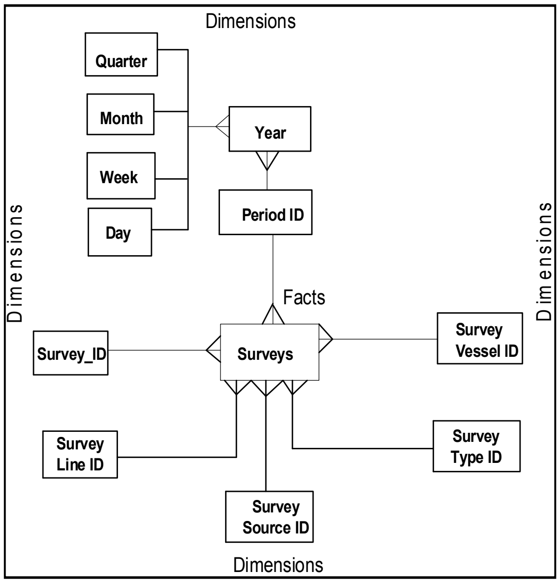
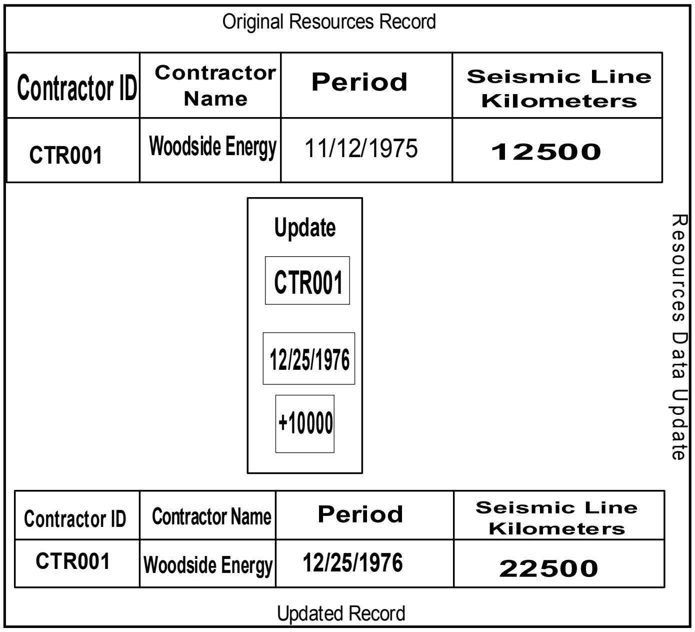


Figure 4.85: Snowflake schema with the normalized *period* dimension and *survey* facts

Transient versus periodic data

In data warehouses, it is often necessary to maintain a record of when events occurred in the past. This is necessary, for example, to compare costs or production or discovery levels on a particular date or during a particular period with the previous year's costs on the same date or during the same period. Most operational systems are based on the use of transient data. Transient data are data in which changes to existing records are written over previous records, thus destroying the previous data content. Records are deleted without preserving the previous contents of those records. However, because of database logs, both images are preserved. Periodic data are never physically altered or deleted once added to the store. Before images and after images, are periodic data (see Figures 4.86 and 4.87). Each record contains a timestamp that indicates the date when the most recent update event occurred.



Original Resources Record

Contractor ID	Contractor Name	Period	Seismic Line Kilometers
CTR001	Woodside Energy	11/12/1975	12500

Contractor ID	Contractor Name	Period	Seismic Line Kilometers
CTR001	Woodside Energy	12/25/1976	22500

Updated Record

Figure 4.86: A typical DBMS log entry

The characteristics of time varying warehoused data

In general, if a data warehouse is built, the data will meet the following criteria:

- The tables are extremely large.
- The data in those tables will have high degree interdependency with the data in other tables
- The principal means of accessing these tables will be *ad hoc* access
- Not only will the tables within the warehouse be large, but there will be a large number of tables available to access.
- The data is accessed in a read-only mode from the user's perspective

- The data will need to be refreshed periodically from multiple sources.
- Much of the data collected will be historical (time-dependent)

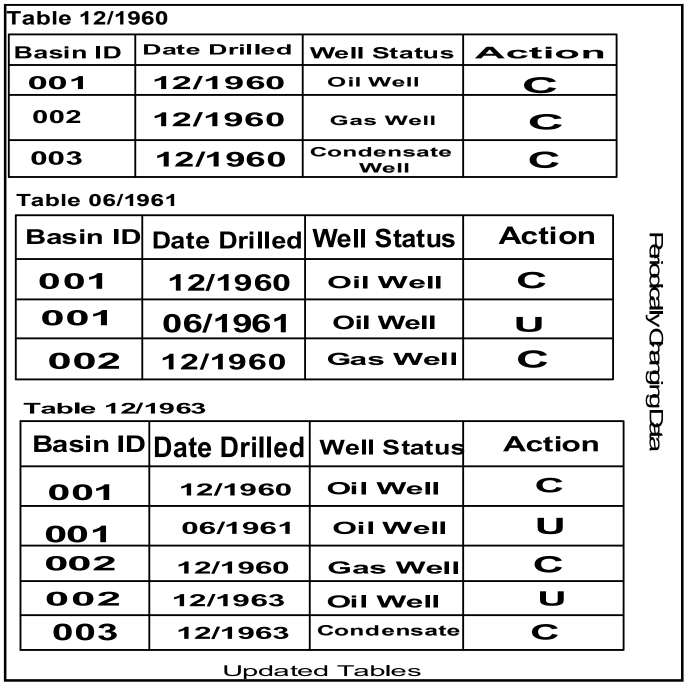


Table 12/1960

Basin ID	Date Drilled	Well Status	Action
001	12/1960	Oil Well	C
002	12/1960	Gas Well	C
003	12/1960	Condensate Well	C

Table 06/1961

Basin ID	Date Drilled	Well Status	Action
001	12/1960	Oil Well	C
001	06/1961	Oil Well	U
002	12/1960	Gas Well	C

Table 12/1963

Basin ID	Date Drilled	Well Status	Action
001	12/1960	Oil Well	C
001	06/1961	Oil Well	U
002	12/1960	Gas Well	C
002	12/1963	Oil Well	U
003	12/1963	Condensate	C

Figure 4.87: An example of *periodic* warehoused data

These characteristics can be grouped into three categories, each of which requires considering some specific organisational assumptions as one proceeds with the construction process:

- High volume/ad hoc access
- Complexity of the environment
- Time sensitivity

The sheer number of tables: A full-scale data warehouse involves hundreds of data tables. As the number of tables grows, it becomes more and more difficult for people to know what each one contains. Therefore, a catalogue of tables must be developed that is not simply a list of contents. It must be organized in a way that makes it extremely easy for people to zero in on what they need to find.

Table interdependencies: Besides the problems of simple table inventory management, one has to find a way to allow users to understand what the relationships between those tables are. This can raise the level of sophistication required from the catalogue many times over.

Timing: Additional sets of complexities arise around the issues of timing. The data warehouse is not time stagnant by any means. Hundreds of tables have to be tracked that are being tracked at different times. The time frame that applies to a given population of data is critical to the users' analysis and the time at which each table is refreshed is also a key. One must develop the means to track and report on these time and synchronization complexities, for the sake of the users and for the administrators of the warehouse.

Features of warehoused periodic oil & gas data

Typical objective for a data warehouse is to maintain a historical record of key events or to create a time series for particular variables such as exploration costs. This often requires storing periodic data, rather than transient data. In Figure 4.87, the first table shows a well drilled during 12/1960 in a Basin (identified as 001) and another well added in the same Basin (under same identification of Basin) during 06/1961, entered as update (U). Similarly another well added in another Basin (identified as 002) under different date (as update) as shown in Figure 4.87. One can see clearly how data warehouses tend to grow rapidly. Storing periodic information can impose large storage requirements. Users of data warehouse must choose very carefully the key data that require this form of processing.

Problem of contractor-surveyor in exploration case

When the problem is historically oriented, it usually results in one or more of the following business rules:

- One-to-one relationships become one-to-many or many-to-many
- One-to-many relationships become many-to-many

The oil and gas company explores and produces (mines) minerals and petroleum products for local and overseas consumers. This company intends to record data about the items that are explored and exploited. The data to be stored includes exploration costs, petroleum production data, and mineral discoveries. More specifically, well data acquired from the drill sites, data on well site installations and well drilling expenditure in different periods of time. To start with, an ER diagram with exploration entity or dimension, describes (Figure 4.88) attributes of exploration ID, exploration name and start date of exploration. An exploration manager decides that it would be useful to record the end date of exploration and costs of exploration during the last 10 years of exploration work along with petroleum production. This allows a more detailed production analysis report to be created. The present ER diagram cannot deliver this data; the present model needs to be adapted to meet the new requirements. The new entity or dimension exploration history is a dependent entity that keeps track of different exploration costs and production over time. Note the attribute start date of exploration with a single occurrence per exploration has turned into an attribute with multiple occurrences in the exploration history.

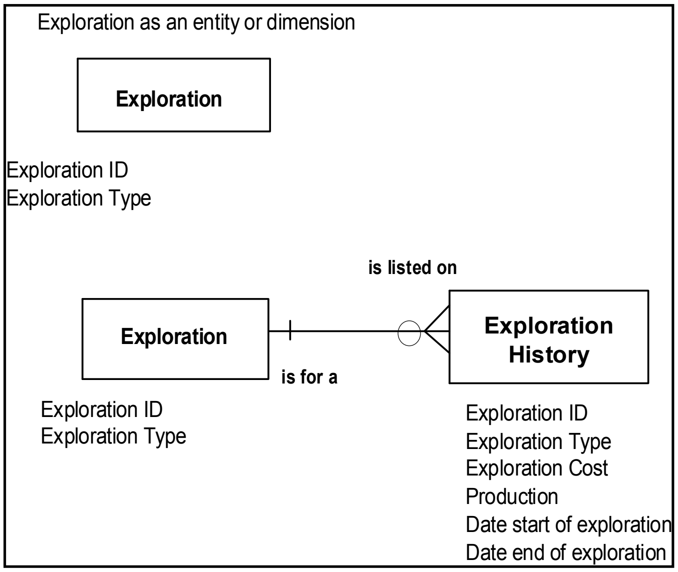


Figure 4.88: ER diagram for an *exploration* entity

The attribute start date is the date of a particular exploration work came into effect. One needs to make decisions at this stage about recording of exploration costs and production rates and pressure of well (if available). There may be two possibilities:

- Every time the exploration cost and production change
- At regular intervals, for example, each day or each week

The choice depends on the frequency of change and importance of recording of every change. In case of production changes, it might be appropriate to record every change, if an average production changed every week. However, in case of stock exchange, it would not be feasible to record every single change to affect in the shares price. One may settle for recording the change on that day, in mid-session or the end of the day. For exchange rates, it is usually the daily rates, which are relevant, so these may be recorded on on-going basis. It is up to the client who decides the recording frequency.

Changes in relationship cardinality

One-to-one and one-to-many relationships can turn into many-to-many relationships. This is always not true for some situations. The cardinality change can be caused by two reasons:

- Business rules change
- Historic data need to be kept (usually for analysis purposes)

Changes in business rules

The changes in business rules often affect the data model, particularly when the business rules change to require historical data to be stored. Oil and gas Company explores and exploits vast oil and gas in different basins. It wants to store a list of surveys and the contractor details responsible for exploration work. The current business rules are:

- Each contractor may be responsible for one or more surveys
- Each survey is looked after by one contractor

ER diagram for this problem (with some sample attributes) is shown in Figure 4.89.

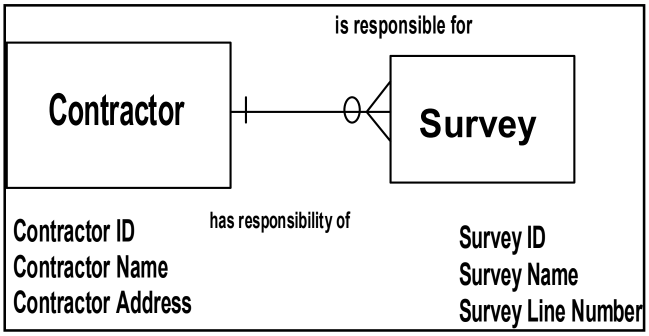


Figure 4.89: ER diagram for oil and gas company's *contractor-survey* problem

Awarding licenses or permits to different contractors is an on-going business of the oil and gas company. Two weeks after successfully implementing the tables in the relational databases, another contractor with the approval of management of the oil and gas company decides that more surveys will be conducted, so surveys have to be looked after by more than one contractor or the employees associated with the contractor. Though the database manager got upset and wished someone had told him these changes before database is implemented. These business situations are unpredictable and unavoidable. This results change in business rules, implying that the relationship becomes many-to-many. Additional attributes need to be stored such as dates and times when a contractor looks after a particular survey. This means that many-to-many relationships need to be resolved with an associative entity. The ER diagram for the modified is shown in Figure 4.90. Changing business rules so that many contractors or employees associated with employees can look after a survey is conceptually simple change to the problem, but it results in several changes to the database structure. One must appreciate to avoid frequent changes in the databases. Database changes are time-consuming and costly. The author examines and investigates business rules and assesses the probability of changes in the short or medium term. If a change is likely, incorporate it into the database structure sooner and earlier.

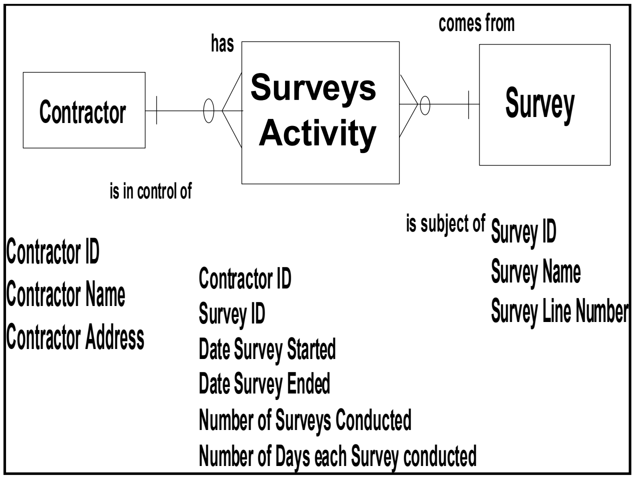


Figure 4.90: Modified ER *contractor-survey* problem

Let us illustrate two simple examples that behave differently when temporal aspects into the problem:

An oil and gas company wants to store all positions within the company and the contractors who are associated with these companies. The business rules are:

1. Each exploration holds production
2. Each production may be held by exploration

An oil and gas company wishes to store petroleum permit information and the contractors who hold these permits. The business rules are:

The ER diagrams for both above examples are shown in Figures 4.91-4.92. ER diagrams look fairly similar. The cardinality is same; but their optionality is different. Let us introduce the time aspect, for example, the situation, which demands that a five-year history be kept for exploration productions and contractor permits. The analysis reveals that in the first example, the relationship changes into many-to-many because over time production can be held responsible by several exploration techniques. An associative entity is required to hold the dates when the exploration reports the production event.

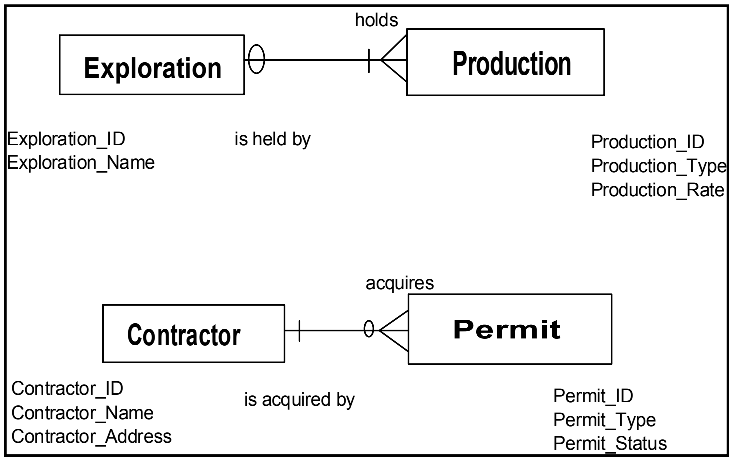


Figure 4.91: Data model depicting, historical data representation

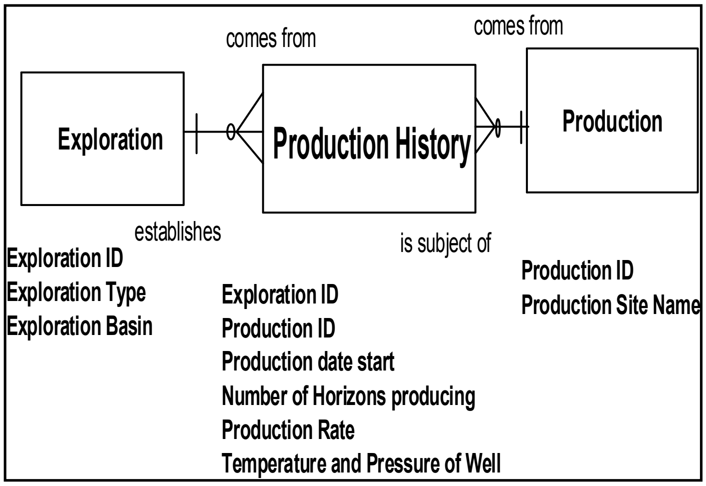


Figure 4.92: *Exploration-production* problem with historical data

In *contractor-permit* problem – the relationship does not change over time. Each permit is always placed by only one contractor. What could be the reason for this different behavior? In the exploration/production example, the relationship is said to be transferable, that is production held by exploration can be transferred to another type of exploration over time. In other words, when prospect is explored by one variety of exploration techniques, can now be exploited by another variety of exploration methods, when a development plan is proposed over a period of time. This causes the relationship to be many-to-many when time aspect is considered. It is up to the user to determine if it is sufficient to only keep track of the incumbent in a particular position or whether a detailed history is required.

In the second example, the relationship is not transferable, that is, permits always belong to one and only contractor. If the contractor with existing permits wants to place some other permits, a new permit order is required to be placed. This is a fairly simple criterion to determine when deciding if one-to-one or one-to-many relationships develop into many-to-many relationships over time. If they are transferable, which is more common and turns into many-to-many relationship, and if they are not transferable, they do not.

Referring back to the ER diagram in Figure 4.92, the question may arise: how one distinguishes between the current survey history and the past survey history. The ER diagram in Figure 4.92 does not separate their survey histories. The current survey history from those occurrences in the associative relationship survey activity is null, where start date and end dates of surveys. Alternatively, the author keeps two relationships between contractor and survey, one for the current history and the other for history. An ER diagram for this solution is shown in Figure 4.93. In effect, Figure 4.93 is combination of Figures 4.88 -4.93.

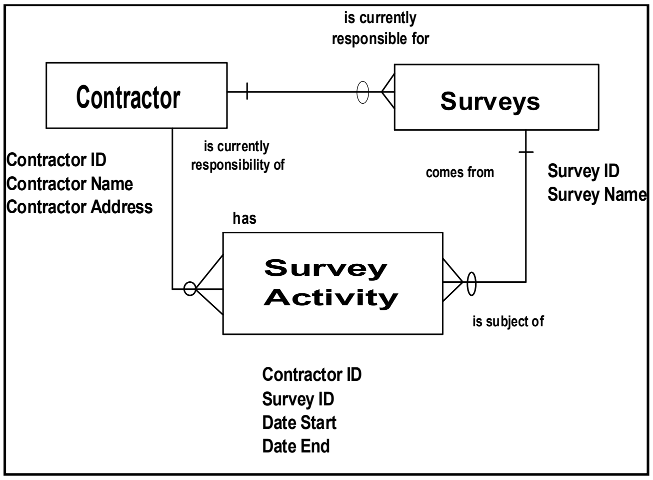


Figure 4.93: *Contractor-survey* problem with current and historical data

Data mining and delivery

By its basic definition (Marakas 2003), data mining (DM) is a set of actions or procedures used to find new, hidden, or unexpected patterns in the data. In the present study, information processed and contained within the data warehouse, DM provides unanswered questions about the oil and gas industry that a corporate manager had previously not thought to raise issues like:

1. Which particular quarter or month the actual exploration cost incurred in a particular state for a particular mineral?
2. What is the meterage drilled on a production lease for a particular mineral in that period?
3. How many original minerals discovered in that particular period and exploration cost and later developed the fields with different costs?
4. How many companies involved in a particular period of time in the same basin for petroleum exploration?
5. How many permits released to contractors for exploratory drilling?
6. What is the total petroleum production in that period?

Computational considerations

The author documents and populates spatial and non-spatial data, of geometric and non-geometric dimensions, generally in the databases of oil and gas data warehouses. Issues regarding choices of computing spatial measures, in particular, in spatial data cube construction are:

1. Collect and store the corresponding spatial object pointers but do not perform pre-computation of spatial measures in a spatial data cube. The polygons of survey data or analog contour grids of different properties of oil and gas data may be collected and stored in the data cubes. In other words, this can be accomplished by storing, in the corresponding cube cell, a pointer to a collection of spatial objects. This choice indicates a merge of a group of spatial objects. Other numeric data collected, are stored, as is done by multidimensional modeling.
2. Pre-compute and store some rough approximation/estimation of the spatial measures in a spatial data cube. Initial processing of survey polygons and contour analog data may be carried out and stored in the data cubes. At this stage, data that have been stored could be of different grains. The choice of computing or processing the data and storing in data cubes, expects coarse estimation of merged data, which generally takes little storage space for the estimated merged result.

3. Selectively pre-compute some spatial measures in a spatial data cube. The question is how to select a set of spatial measures for pre-computation. The selection are performed at the cuboids level, that is, either pre-compute and

store each of merge- able spatial regions for each cell of a selected cuboid or pre-compute none if the cuboid is not selected.

Since a cuboid usually consists of a larger number of spatial objects, it may involve pre-computation and storage of a large number of merge-able spatial objects but some of them could be rarely be used. Therefore, it is recommended that the selection is performed at a finer granularity level by examining each group of merge-able polygons or contours in a cuboid to determine whether such a merge should be pre-computed. The best choice is to select pre-computed some aggregated survey polygons or other oil and gas data grid contours and then perform efficient online polygon amalgamation operations. An application example of such a oil and gas data warehouse construction and online analytical processing is to do multidimensional analysis of oil and gas data and draw the correlations, patterns and trends from datasets.

4.8.2.2 The Indonesian case study

To address RQ1 - RQ7 and follow up RO1-RO7, another case study is considered in this section for validating the research objectives. The author provides the constructs and models used in this case study. Heterogeneous data sources exist in several basins of Indonesia. Ontology based data warehouse approach is used to handle these data sources. Tools and concepts used for designing and developing data-warehouse for bio-informatics (Thomas et al. 2006 and Shastri and Dreher 2011) support our ideas of data integration process of system elements. All the elements and processes of a sedimentary basin (basins illustrated in Figure 2.1b) are visualized via multi-dimensional metadata, representing several attributes and their characteristics. The data warehouse approach brings together petroleum systems' data from different oil and gas fields from sedimentary basins' different depositional and geological regimes. In the Petroleum Digital Ecosystem (PDE) scenario, data from the geological, geophysical and geochemical domains are integrated in a data warehouses environment. The data warehouse approach (Hoffer et al. 2005 and Shastri and Dreher 2011) is used to benchmark and track the effectiveness of petroleum system productivity over space and time dimensions. It also allows processed (knowledge based) data shared among professionals and geographically distributed worldwide. The need to integrate petroleum systems data from multiple systems and sources is well known (Magoom and Dow 1994). It is important that data warehouse designers define the scope, depth, comparability and accuracy of data entering the warehouse. The scope of data refers petroleum systems data, sedimentary basins data, and geological, geophysical and geo-chemical data from multiple periods (time-dimension), and geographic locations (space-dimension). Depth of data refers the level of details obtained. To be comparable, data from multiple dimensions and different sites should adopt the same classifications, as much as possible. No matter how differently data are collected across sites, they are significantly altered for integration before moving into the data warehouse environment. To reduce the burden of alteration, it is important for petroleum systems analysts and geo-modellers to use compatible software systems to acquire and send data to the repositories. It is also important to standardize the data collection processes. Accuracy of data is desired for all types of data in any given situation and this is a fundamental requirement for reliable use of data. A sedimentary basin is typical example; in which varied data dimensions are embedded, such as *seismic*, *drilled-well*, *petro-physical and production datasets*.

Data structures or schemas are drawn in different star-schemas for different petroleum systems' elements and processes. For this purpose, author narrates multiple dimensions for interpreting them conceptually, with multiple relationships among several other related dimensions and physically, with fact data tables. Ontology models are constructed for shallow marine and deltaic petroleum systems, in which several dimensions are described, connecting the factual data with one-to-one, one-to-many and many-to-many data relationships of dimensions. These models are represented in star-schemas, as described in the following sections.

Data modelling in Indonesian basins

In a large-scale sedimentary basin, the data schemas are simple and flexible enough to modify as per the geological situations and users requirements, despite complex geological situations. If the data schemas are too large, complexity is a significant barrier to widespread adoption of the warehouse technology, because users will find the schema so difficult to understand that they will be unable to write queries and application programs. Schemas are evolved to grow and support new data types. Limiting the scalability to more data sources has definite advantage, but in petroleum exploration industry, data structures are typically very large in size and multidimensional.

More logical data structuring defines single common tables for information that is common to many warehouse data types, to decrease the schema complexity. For example, for many basins and petroleum systems, there may be common *reservoirs*, *structure* styles, *source* and *seal* types (Magoom and Dow 1994). Each is implemented through a single common table. To this extent, an associative dimension or entity needs to be defined among common *structure, reservoir, source, seal* and depositional systems. Dimensions, representing spaces, such as *unique ID* and its *type* are required, for constructing the schema. Their associations are uniquely identified in the warehouse schema. Different dimension identifiers may have the same Fact ID. Thus a simpler approach is used in which all data warehouse dimensions share a single space of dimension identifiers within a warehouse instance. The identifiers are known as warehouse identifiers and are integers that are assigned at database load time. The schemas designed for a warehouse have a support with concurrent presence, accessibility and addressability of multiple datasets, multiple versions of a given database (Nimmagadda and Dreher 2006a, Rudra and Nimmagadda 2005 and Shastri and Dreher 2011), within a petroleum systems warehouse perspective.

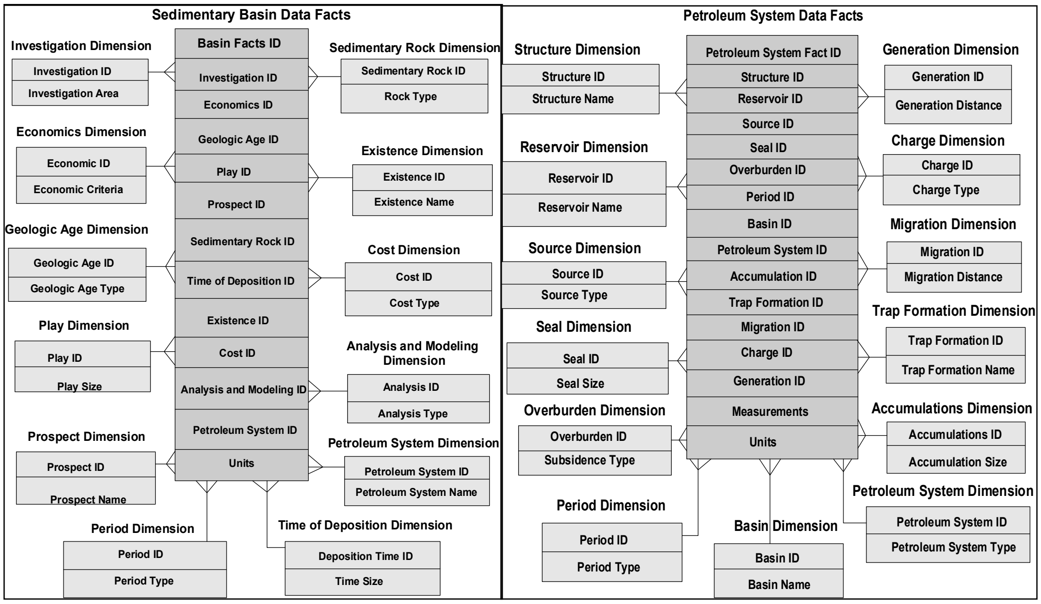


Figure 4.94: Star-schema models representing relationships and their connectivity among various attributes of petroleum and depositional systems of a sedimentary basin

The warehouse schema facilitates coercion of different sources of the same type of data into a common semantic framework. If there is no common information of attribute or data property among data dimensions, then different databases exist side by side within a warehouse environment, without having merged. Once common data property information exists, databases get merged. The author merges star-schema models derived in Figure 4.94 with logical key attributes among the other participating dimensions in the structuring process.

Data loader requirements

Because basins and petroleum systems are large in size (to the order of 19,300 mi^2 [50,000 –km^2] or terabytes in storage) and their associated data are often large. Poorly defined syntax and frequent load failures could result crashing of loaders. For this reason, database loaders should be able to recover gracefully from errors encountered during parsing of their input files. The loaders are designed to keep loading even in the presence of an error. If partial data are inserted into the warehouse, a load error maintained on the related dimension is updated to indicate that an error occurred while parsing the dimension and that the warehouse entry for the dimension may therefore be incomplete or contain errors.

Petroleum systems warehouse design

In the case of Indonesian basins too, the author follows similar procedure for schema and warehouse designs as narrated for Australian case study discussed in Section 4.8.2 but the models keep changing because of changing geological situations.

Knowledge discovery from integrated geo-ontologies

Several geological events occur in each and every sedimentary basin. For example, *pre-rift, syn-rift and post-rift events* (Nigel et al. 2009 and Al-Fares et al. 1998) taken place in a basin at global scale, cannot be isolated, but always have an interconnectivity, because of the echoing effect, and all these events are connected to each other. In addition, in a total petroleum system (TPS) scenario as well, all the elements of system share each of the petroleum system's elements and processes in TPS. As illustrated in Figure 4.95, surface and subsurface information are integrated using survey and drilled-well dimensions are interconnected and *survey*- and *well*-ontologies that make possible integrating/connecting the petroleum systems. Specialized geo-ontologies that describe the conceptualized attributes are used in any domain ontologies. Data models are constructed representing the Geophy_Anom and Total Petroleum System (TPS) concepts. Different geophysical criteria are discussed based on anomalies observed in these datasets.

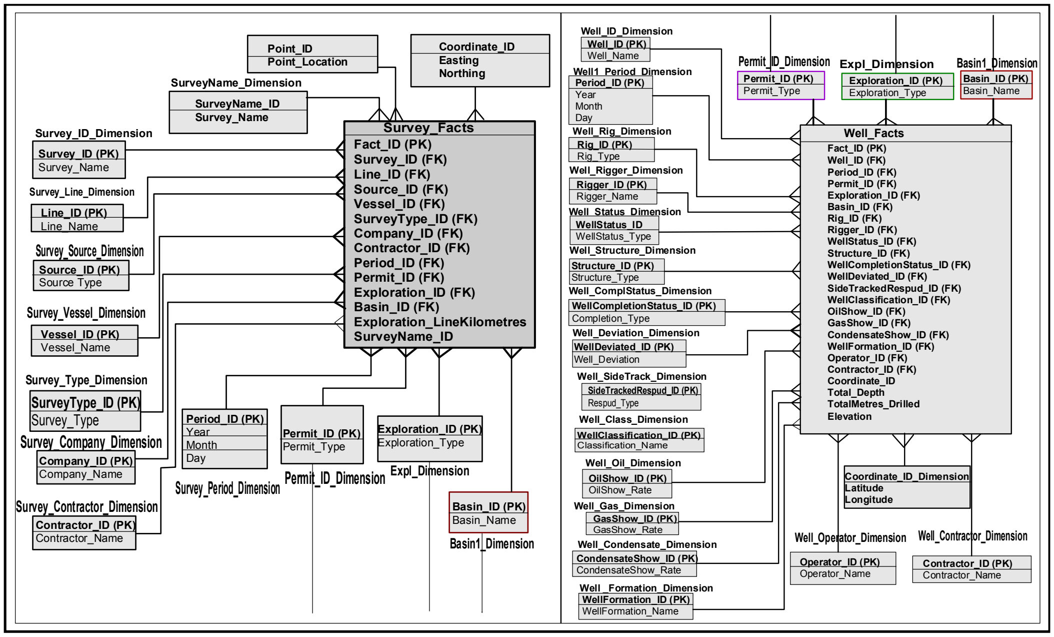


Figure 4.95: Mapping and modeling of *surface-subsurface* data dimensions and facts

Geophysical anomalies criteria

Most of the technical jargon used here may be found in Parasnis (1997). Important criteria for applicability of any geophysical method or methods are to delineate the existing physical property contrasts between the geological objects and the surrounding host rocks (Green 1991). For example, sedimentary rocks are feebly magnetic, while some of metamorphic and igneous are more magnetic. The magnetic susceptibility increases from acidic to basic and ultra-basic rocks. More specifically, it is a case for a geological event, such as a salt-dome. Salt domes are emplaced when a buried salt layer, because of its low density and ability to flow, rises through overlying denser strata in a series of approximately cylindrical bodies. The rising columns of salt pierce the overlying strata or arch them into a domed form. A salt dome has physical properties that are different from the surrounding sediments and which enable its detection by geophysical methods. These properties are: (1) a relatively low density; (2) a negative magnetic susceptibility; (3) a relatively high propagation velocity for seismic waves; and (4) a high electrical resistivity (specific resistance).

The Bouguer gravity (Parasnis 1997) anomalies interpreted on either side of a fault, where density contrasts exist; there is an increasing in gravity value or instance from the lower density region to the higher density region. The rate of change of gravity depends on the depth of the faulted bed, its dip, the thickness of the bed and the differential density (density contrast at the fault plane). In case of light sedimentary fill over a granitic basement structure, gravity anomaly decreases towards the center of the basin, the amplitude and rate of decrease indicates the slope of the basin, depth and mass difference. For bodies extending over large distances, the anomaly maps show contour lines more or less parallel to the strike of the body. Bodies confined in their areal extents, have closed contour forming circles, polygons or ellipses. Residual are computed by removing the regional anomalies from the gravity data. Residual gravity anomalies depict regular geometric bodies. Though gravity data yield information about depth, nature and extent of anomalous source, it has limited scope because of ambiguous and non-unique solutions obtained by gravity method. Anyway, the gravity data play a vital role in delineating tectonic zones, mapping sedimentary structures associated with oil and gas deposits. Gravity data, if are used in conjunction with magnetic and seismic data, may minimize the ambiguity in the interpretation of complex geological events. Lateral variations of the Bouguer gravity anomalies can infer the variations of sub-surface lithology. Gravity anomaly is the departure of the observed gravity from that of the expected or theoretical gravity.

Magnetics and micro-magnetics (Parasnis 1997) detect hydrocarbon induced mineralization at shallow depths in sediments above oil and gas accumulations, applicable to onshore and offshore basins. Magnetic prospecting is based on the measurement of, analysis and interpretation of the cause-effect relation between the spatial variations in the normal magnetic field (magnetic anomaly) and the geometry, attitude and geology of the anomaly due to the causing body. This method uses natural and spontaneous field of force, superimposed over which are the fields due to the geological inhomogeneity. Magnetic character of the rock depends on the type and amount of magnetic mineral present, grain size, the geological and structural history of the rock, and the magnetic effects of a buried mass is a function of magnetic susceptibility, magnetic and geological history, size, shape, attitude and depth of burial. As a result there is no unique solution in the interpretation and understanding of the data. The geological mapping of contacts between different rock formations with varying magnetic properties, demarcation of intrusive, study of composition and relief of crystalline basement and if the topography of basement and if basement is associated with magnetic and non-magnetic materials. Sediment to basement contacts and sediment thickness are other investigating features from the magnetic data, by means of their susceptibility variations between oil bearing sediments and igneous rocks. Combining magnetic and gravity data allows us to resolve ambiguities in subsurface structural interpretation that would be unresolvable if only one or the other dataset was used by itself. Several "blind" intrusive and plutons have been recognized, and two of these have been modeled. Known mineral deposits and prospects are now placed in the context of the subsurface geology influencing and moderating their emplacement.

Induced Polarization method (IP, Parasnis 1997) method attempts to detect alteration zones or pyrite chimneys caused by micro-seepages from hydrocarbon reservoirs into iron rich sediments near the surface. Hydrocarbon induced changes in sediments above hydrocarbon accumulations or directly resistive-hydrocarbon bearing formations are detected. Radiometric or gamma radiation surveys detect generally low radiation values at the surface above the oil and gas accumulations. Controlled source audiomagnetotellurics method measures electrical and magnetic fields and their strengths based on geometry of GMT profiles. The surface expression of hydrocarbon seepage and hydrocarbon-induced alteration of soils and sediments can take many forms including (1) anomalous hydrocarbon concentrations in soils, sediments, and waters; (2) microbiological anomalies and the formation of “paraffin dirt”; (3) mineralogical changes such as formation of calcite, pyrite, uranium, elemental sulfur, and certain magnetic iron oxides and sulfides; (4) bleaching of red-beds; (5) clay mineral alteration; (6) electrochemical changes; (7) electromagnetic and telluric changes, (8) radiation anomalies; and (9) biogeochemical and geo-botanical anomalies. These different manifestations have led to development of an equally varied number of geochemical and non-seismic geophysical exploration techniques. These include direct and indirect geochemical methods, magnetic and electrical methods, radioactivity-based methods, and remote sensing methods. Survey scales are variable, depending upon the objectives of geological investigations, such as regional, reconnaissance and detailed. All the anomalies observed and documented are incorporated in the warehouse design through multidimensional geo-ontology models.

Modelling data anomalies

Hundreds of geophysical anomalies are neither documented nor at times noticed from the geophysical data. These anomalous features could either be attributed and or distinguished from producing and non-producing field areas. Different multiple dimensions and their associated fact data instances are logically connected and stored with relationships, in which case, one-to-many (one – to - many→) are described. *Structure*, *reservoir, source* and *seal* are elements and their anomalies interpreted from different exploration data instances, derived from seismic, drilled-well and all other geological data, which are again multiple domains, are required for logically storing exploration data both for integration and data mining purposes.

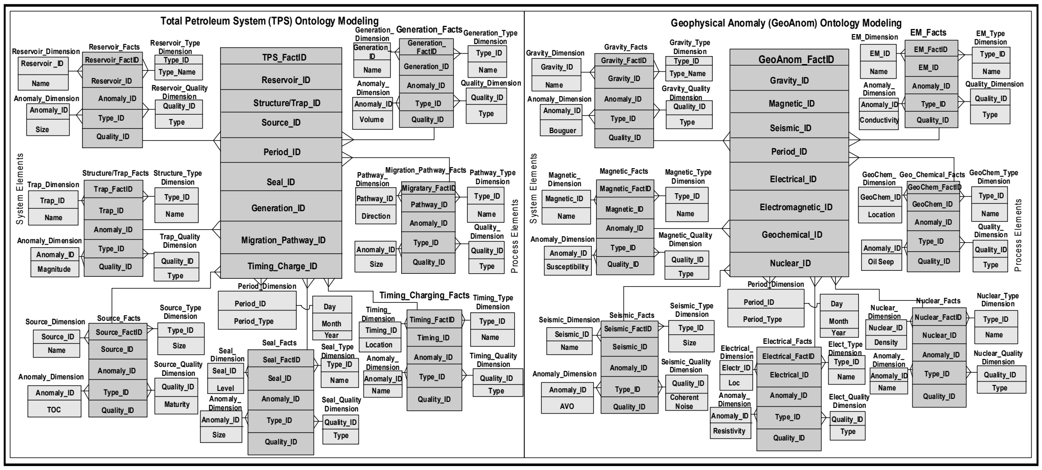


Figure 4.96: Petroleum systems and geophysical *anomaly* ontology modelling

Several contexts and conceptual attributes are deduced as shown in Figure 4.96 for star-schema models showing one-to-many relationships among multiple dimensions described from survey, well and other attributes of petroleum systems' elements. The author conceptualizes these dimensions based on the knowledge among known attributes, such as seismic horizons, structures and their associated anomalous features, such as *structural highs*, *shaling-out* or *wedging out* features. When these data or anomalous features are integrated with other attributed anomalies from other domains, such as sub-surface (borehole data), gravity, magnetic, electrical and seismic, an increased understanding and perception of the geological and geophysical features are explored. When the data are communicating among petroleum systems, as illustrated in total petroleum system domain, the perception of geophysical features (anomalies/attributes) is explicit in terms of geological interpretation of anomalies interpreted from either individual geophysical data or combined methods of geophysical exploration and prospecting.

Integrated framework

As illustrated in Figures 3.2 and 3.36-3.38 in Chapter 3, author uses and reuses domain ontologies with hierarchical structures. Several geophysical anomalies, including elements of petroleum systems are accommodated from a Total Petroleum System in the global scale, in multiple dimensions. These dimensions are structured in several ontology hierarchical structuring methodologies. Logically, dimension and fact data tables (and their instances) accommodated in multiple dimensions in a warehouse environment (as demonstrated in an integrated framework in Figures 3.36 and 3.37), are formally defined as a triple <T, S, M>, where *T* is the Total Schema (Basin or sets of Basins, Total Petroleum Systems (TPS) Scale), *S* is the heterogeneous set of Petroleum Systems, and *M* is the mapping that maps queries between the Petroleum Systems and the Total Petroleum Systems (Magoon and Dow 1994). Both *G* and *S* are expressed in languages over alphabets composed of symbols for each of their respective relations. The mapping *M* consists of assertions between queries over *G* and queries over *S*. When users pose queries over the data integration system, they pose queries over *G* and the mapping then asserts connections between the elements in the global schema and the source schemas. Author apprehends processing of several anomalies to logically store them in a metadata structure, to enable data mining algorithms capture these data views.

Logically, a database over a schema is defined as a set of sets, one relating to each other in a relational database schema. The database corresponding to the source schema *S* would comprise the set of sets of tuples for each of the heterogeneous data sources and is called the *source database*. Note that this single source database may actually represent a collection of disconnected databases. The database corresponding to the virtual mediated schema *G* is called the *global database*. The global database must satisfy the mapping *M* with respect to the source database. The legality of this mapping depends on the nature of the correspondence between *G* and *S*. Two popular ways to model this correspondence exist: *Global data view and Local data view*. Global systems model the global database as a set of views over *S*. In this case *M* associates to each element of *G* as a query over *S*. Query processing becomes a straightforward operation due to the well-defined associations between *G* and *S*. The burden of complexity falls on implementing mediator code, instructing the data integration system exactly how to retrieve elements from the source databases. If any new sources join the system, considerable effort may be necessary to update the mediator, thus the global approach appears preferable when the sources seem unlikely to change. In a global approach, the system designer would first develop mediators for each of the *survey* or *drilled-well* information sources and then design the global schema around these mediators. For example, consider if one of the sources serve a particular petroleum system. The designer would likely then add a corresponding element of that system to the global schema. Then the bulk of effort concentrates on writing the proper mediator code that will transform predicates on a petroleum habitat or province into a query over the Total Petroleum System described at a basin scale (for example from large basins in the South East Asia). This effort can become complex if some other source also relates nearby producing fields that associate with petroleum systems with far-off distances, because the designer may need to write code to properly combine the results from the two sources. While searching for conceptualized relationships among petroleum systems, semantics are used to contextualize the content in relationships.

Semantic integration is the process of interrelating information on meaning of vocabularies and terminologies from domains of different geological and geophysical domains in geographic and periodic dimensions, for example *structure, reservoir*, *source*, *seal*, and processing agents, such as *migratory path-ways* and *timing* of occurrence of structure/trap, when connected to *survey*, *well* and *permits* information. In this regard, semantics focuses on the organization of and action upon information by acting as a mediatory between heterogeneous data sources which may conflict not only by structure but also context or value. In Enterprise Application Integration, semantic integration facilitates or potentially automates the communication between computer systems using metadata publication. Metadata publication potentially offers the ability to automatically link ontologies. One approach to automated ontology mapping requires the definition of a semantic distance or its inverse, semantic similarity and appropriate rules. Other approaches include so-called *lexical methods,* as well as methodologies that rely on exploiting the structures of the ontologies. For explicitly stating similarity/equality or other relational ontologies, there exist special properties or relationships in most web-base ontology languages. OWL, for example has “sameIndividualAs” or “same-ClassAs”. Eventually systems design may see the advent of composable architectures where published semantic-based interfaces are joined together in new and meaningful capabilities. These are predominately described through design-time declarative specifications that could ultimately be rendered and executed at run-time. As shown in Figure 4.97, the author computes multiple volumes in the form of cubes from warehoused metadata for visualization and interpretation of data views.

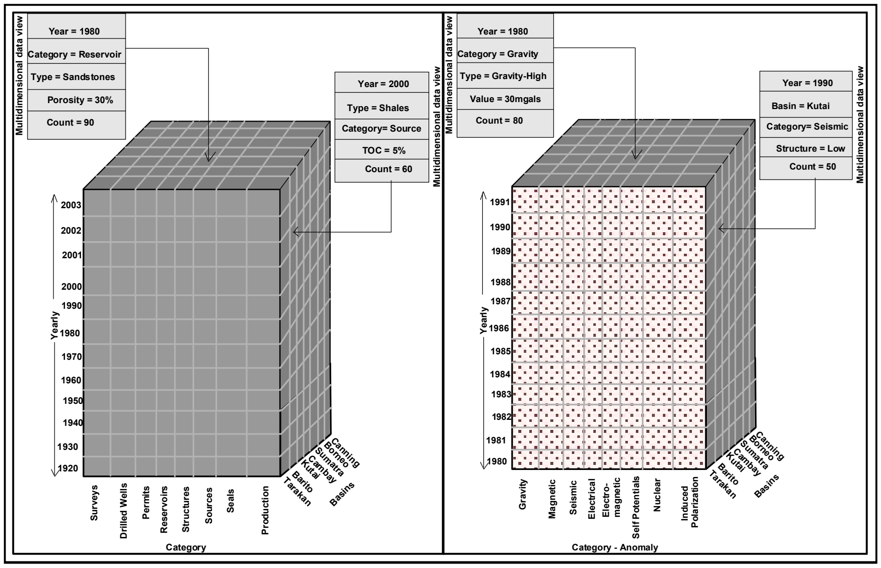


Figure 4.97: Multidimensional data cubes

The data views extracted for user queries are represented in the form of OLAP visualizations, subsequently used for interpretations and knowledge mapping. Semantic integration uses facilitating design-time activities of interface design and mapping. In this model, semantics are only explicitly applied to design and the run-time systems work at the syntax level. This "early semantic binding" approach can improve overall system performance while retaining the benefits of semantic driven design. Further, author extracts data views for different scalable multiple dimensions maps, such as seismic structure, reservoir and production data, which are superimposed by the author for understanding the concept of integration process and also relationships among geophysical anomalies, in both exploration and field development project scenarios.

4.8.2.3 The Ugandan case study

The author presents another case study in this section addressing RQ1 – RQ7 to follow up RO1 – RO7, as given in Sections 1.3.1 and 1.3.2 in Chapter 1. The constructs and models used in this case study are highlighted. Multidimensional data from sub-basins are handled in integrated workflows in the case of Ugandan Albertine Graben (Figure 2.1b) basin. Keeping in view, the volumes and sizes of horizon-structure-reservoir-production datasets in petroleum bearing sedimentary basin in unmanageable way, it is imperative to use warehouse and mining approaches in the Albertine Graben basin, in which data are ontologically conceptualized in multiple dimensions. These valuable data are more intelligently stored and so that accessibility of domain knowledge is easy during data mining stage at later stages. The data integration, sharing of knowledge and interoperability are significant issues that are addressed more robustly by data warehousing and mining approaches in these basins. Several ontology models, deduced in Figures 4.98 and 4.99, are made of data structures in multiple domains and for their integration purposes.

The author describes ontology models deduced for different data dimensions of Albertine graben are narrated in Figures 4.98 and 4.99. Different dimensions such as, *geologic age*, *hydrocarbon play*, *time of deposition* of *sediments*, period of investigation, economic factors, overall petroleum system are chosen in the current data model and these are all interconnected through attribute dimensions and their instances (this is a generalized ontology model for broader scope of Albertine Graben). Similarly, as shown in Figure 4.99, data dimensions involved in the data integration process have interconnectivity and this connectivity facilitates the systems through which interoperability needs to be addressed, especially during implementation stage of these models. Acoustic impedance, porosity, rock physics calibration, seismic structure and other attributes are integrated through their fact data instances. In addition, the author uses seismic time, depth and velocity dimensions for integrating composited structure data attributes.

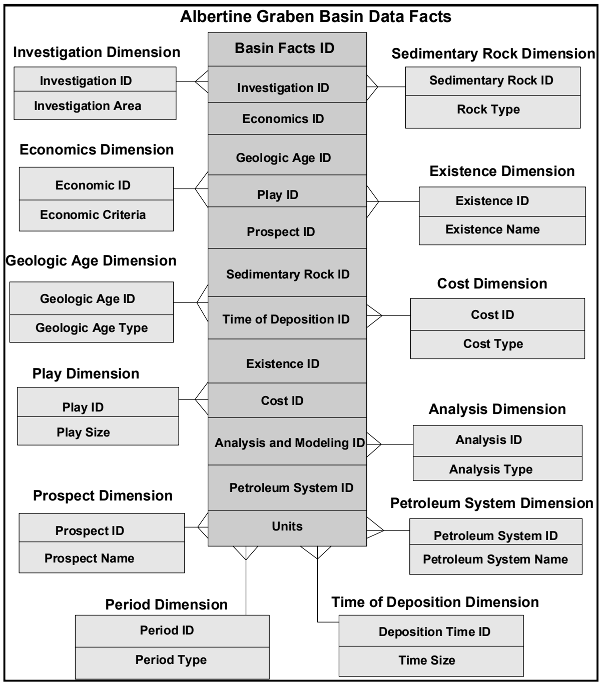


Figure 4.98: Ontology model for designing and mapping the Albertine Graben data attributes and their instances

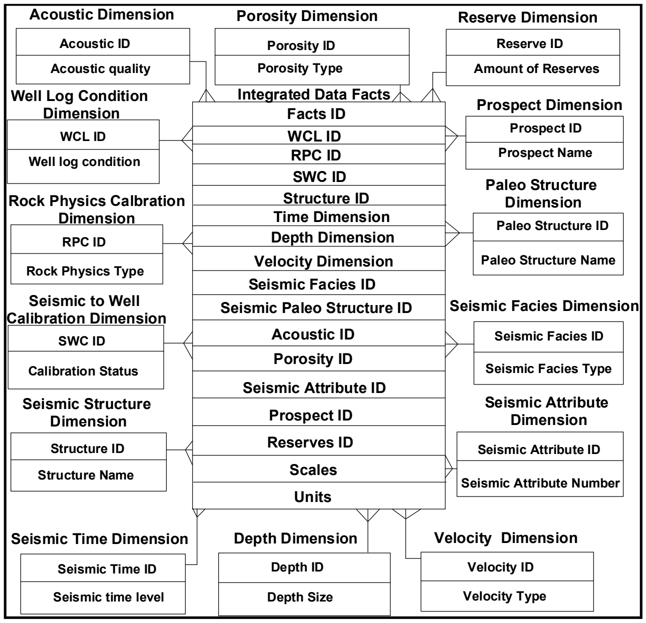
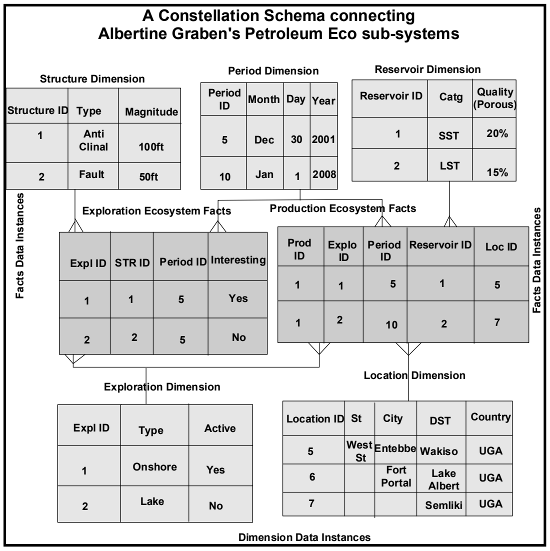


Figure 4.99: A multidimensional ontology model, integration of reservoir parameters

The constellation schema is used in the warehouse structuring model as represented in the Figure 4.100, demonstrating the connectivity of sub-basins (systems) through their common data properties and their instances. Peak and trough attributes of seismic horizon dimension are used in integrating with log and petrophysical data attributes. As shown in Figure 4.101, *peak* and *trough* attributes are interrelated and are integrated through calibration of petrophysical data attributes. Hierarchical and relational data structuring methods are used for this purpose.


A Constellation Schema connecting
Albertine Graben's Petroleum Eco sub-systems

Structure Dimension

Structure ID	Type	Magnitude
1	Anti Clinal	100ft
2	Fault	50ft

Period Dimension

Period ID	Month	Day	Year
5	Dec	30	2001
10	Jan	1	2008

Reservoir Dimension

Reservoir ID	Catg	Quality (Porous)
1	SST	20%
2	LST	15%

Exploration Ecosystem Facts

Expl ID	STR ID	Period ID	Interesting
1	1	5	Yes
2	2	5	No

Production Ecosystem Facts

Prod ID	Explo ID	Period ID	Reservoir ID	Loc ID
1	1	5	1	5
1	2	10	2	7

Exploration Dimension

Expl ID	Type	Active
1	Onshore	Yes
2	Lake	No

Location Dimension

Location ID	St	City	DST	Country
5	West St	Entebbe	Wakiso	UGA
6		Fort Portal	Lake Albert	UGA
7			Semliki	UGA

Figure 4.100: An ontology model depicting the connectivity among data instances

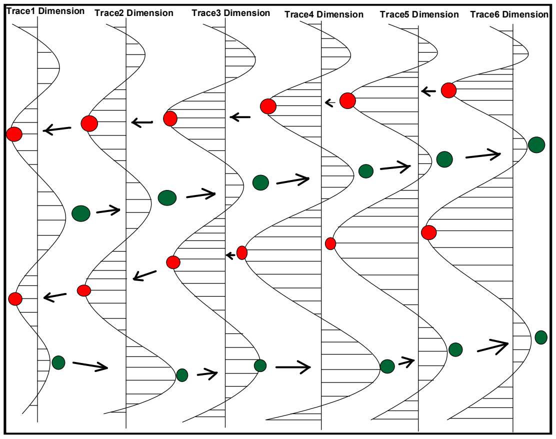


Figure 4.101: Seismic horizon attributes and their trace attribute instances, used in the data modelling and integration process of Albertine – Graben sub-systems

As explained in Figures 3.36 and 3.37, data facts are modelled in a warehouse model, in order to establish the connectivity among multiple horizons including elements' of petroleum systems. The author uses the integrated framework, incorporating the data warehouse and performing mining tasks (as demonstrated in Figure 3.38). Entities or dimensions deduced in the models go through data structuring procedures, as per the type of data and whichever logic data accept. Ontology supports easy organization of these logical data structuring procedures. Other criteria that support compatibility and design of a warehouse are:

- For sharing *structure-reservoir-production* data among several fields in a producing basin and among multiple petroleum systems
- Data models are flexible to fast changing geological and geophysical data situations in a basin
- Reusability of composite data models among several petroleum fields and systems for knowledge discoveries
- To model rapidly changing seismic data (because of velocities), because of geological situations, changing data structures and models in a warehousing environment, is more flexible
- Conceptualization and contextual designs are significant in understanding the integration process. Warehouse design addresses the issue of data integration process.

Need of PDE and ontologies in the Albertine Graben basin

There are several tools and methodologies narrated in the literature (Nimmagadda et al. 2006d) that explain the petroleum systems and their integration. The author characterizes these petroleum systems as ecosystems in which each sub-system is illustrated as an information system, in which all the components are assembled and interconnected (for communication and interaction in an eco-system scenario). These elements could be several parts, correlatable horizons, members, fields, factors, procedures and processes that make up of petroleum system. Information system is shaped by interaction of a community of networks or structures (either in the form of data schema or a geological structure) and their organization including their existence with their environments. Petroleum system is an information system, in which all these elements, processes and environments are interactive and communicative, so that they exist in the Albertine Graben basin in the form of entities and or dimensions. The Albertine Graben – a complex Petroleum System, is beneficially modelled as an information system, and for emphasising the interactivity and simplicity of natural complexity of models, a Petroleum Digital Ecosystem (PDE). For substantiating PDE, the concept of ecosystem is described as an emerging concept in a petroleum analysis scenario, in sustainable economic and environment perspective.

This is another example to extract and build relationships among sub-basins of the Albertine Graben, petroleum systems, oil/gas fields, hydrocarbon plays, leads and prospects. These representations and categorizations are ontologically described in different contextualization (or conceptualizations) and domains. The Albertine Graben petroleum system is an information system, similar to any other information systems described in different domains of applications. Petroleum system requires timely convergence of certain geological elements, processes and events essential to the formation of the hydrocarbon deposits in this graben. A petroleum basin (or province) is a geologic entity (in its generalization representation) containing at least one or more petroleum systems. The concept of petroleum system is in a contextualization domain, in which continuous portions of sedimentary reservoirs may contain hydrocarbon pools, describing:

1. Reservoirs of similar or dissimilar productive geologic sequence (or seismic sequences)
2. Similar or dissimilar chemical compositions
3. Similar or dissimilar trap types (structural and or stratigraphic)
4. Similar or dissimilar reservoir types

Integration of domain ontologies within the PDE scenario

The author construes several data models that represent ontology, narrating relationships among petroleum systems and their elements of different sub-basins of the Albertine Graben. As described in Figure 4.102, the author interprets several such domains and genetic relationships, built between a particular pod of generating source rock and the resulting petroleum resource that can ontologically be expressed. A lead or prospect, in either case is conceptual initially, when a successful prospect turns into an oil/gas field after drilling or the prospect may be unsuccessful, if structural prospects are devoid of hydrocarbons. Domain ontologies organize the needs of information modelling and PDE development needs.

As described in Figure 4.102, the author classifies each petroleum system genetically (Nimmagadda and Dreher 2010d) in terms of several processing factors, such as charge, migration drainage style, and entrapment style. Each charge factor is again critically categorized in terms of *super-charged*, *normally-charged* and *under-charged*. Under *migration* process category, hydrocarbons are either vertically and or laterally charged. Under entrapment category, it could be associated with high or low impedance reservoirs, again vertically, horizontally and or laterally.

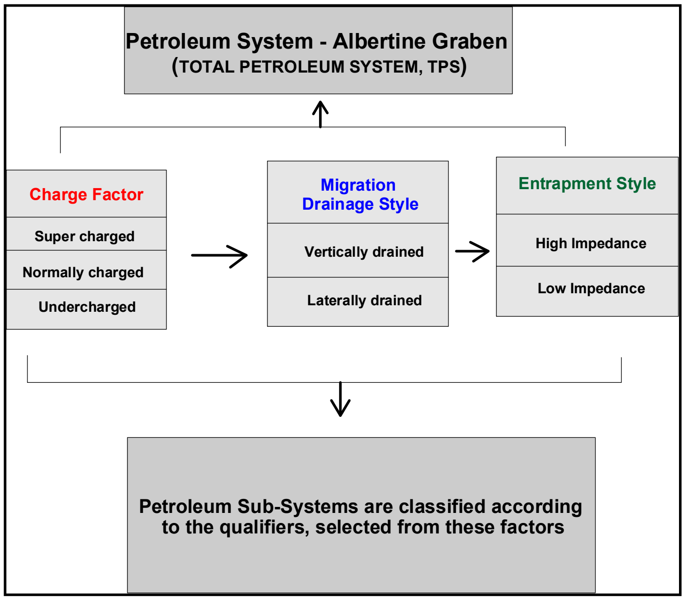

Petroleum System - Albertine Graben
(TOTAL PETROLEUM SYSTEM, TPS)
Charge Factor
Super charged
Normally charged
Undercharged
Migration Drainage Style
Vertically drained
Laterally drained
Entrapment Style
High Impedance
Low Impedance
Petroleum Sub-Systems are classified according to the qualifiers, selected from these factors

Figure 4.102: Genetic classification of petroleum system of Albertine Graben, Western Uganda

As a matter fact, the relationships are conceptualized among these classifications for an effective integration (Davidson et al. 1995) process. Total petroleum system (analogues to concept of PDE) concept states that all the petroleum system elements (including processes) are shared among sub-basins of the Albertine Graben (Nimmagadda and Dreher 2012a). An effective way of interpreting economically viable petroleum system is to integrate the tectonic framework, sequence stratigraphy, geologic history, thermal history along with sedimentary basin analysis and modelling.

The basin analysis is an investigation of a sedimentary basin, in which different domains of data are described for each sub-basin. Basin modelling is how different system elements are put together to formulate a hypothesis. The difference between analysis and modelling is that in analysis, an existing item is dissected to determine how it functions, where as in modelling; a hypothetical item is dissected to determine how it should function. Prospect modelling is used on a prospect to justify drilling, where as a prospect analysis is carried out to investigate enhancement of production in an existing prospect or why prospect lacked hydrocarbons, even after post drilling. Several factors such as investigation, economics, geologic time, existence, cost and analysis and modelling are investigated in each sedimentary basin, petroleum system, hydrocarbon-play and prospect. Hierarchical, relational and networking relationships can be constructed among these factors and levels of investigations in the Albertine Graben. Similarly, elements and processes of petroleum sub-systems of the Albertine Graben are hierarchically and relationally linked among attributes and characteristics of each of these attributes of sub-systems. The evaluation of the effect, which has accrued from the application of methodological views (that addressed data organizations, data integration, data mining and data interpretation from knowledge models) as producing results are described in the following sections.

Methodology – petroleum digital ecosystem (PDE) framework

The data warehouse approach brings together petroleum systems data of the Albertine Graben from different sedimentary sub-basin settings of different depositional and geological regimes, in addition to bringing data from other geophysical and geochemical data warehouses. The author uses the data warehouse approach to benchmark and track the effectiveness of petroleum system productivity with respect to both time-period (periodic dimension) and space dimension. It also allows processed (knowledge-domain models) data shared among professionals and geographically distributed worldwide. The need to integrate petroleum systems data from multiple systems and sources is well known (Magoom and Dow 1994 and Coronel et al. 2011). It is important for data warehouse designers to define the scope, depth, comparability and accuracy of data entering the warehouse, besides describing the data dimensions, attributes and their types. The scope of data refers petroleum systems data, sedimentary basins data and geological, geophysical and geo-chemical data in multiple periods, geographic locations (space dimension) within the Albertine Graben, such as Lake Albert, Pakwach and Reino Camp sub-basins in the north and Lake Edward and George sub-basins in the south. The depth of data refers the level of details of these sub-basins. For comparison, data from multiple dimensions and different sites (sub-basins of Albertine Graben) should adopt to the same classifications, as much as possible. No matter how differently data are collected across sites, they are significantly altered dynamically (based on the nature of data types, exploration and production datasets are always temporal and space varying, for example) for integration before moving into the data warehouse. To reduce the burden of alteration, petroleum systems analysts and geo-modellers are made vital to use compatible software systems to acquire and send data to the repositories. It is also imperative to standardize the data collection processes. Accuracy of data is desired for all types of data (that need to be intelligently stored in PDE) in any given situation and this is fundamental requirement for reliable use and documentation of data. Data facts are documented from the Albertine Graben basin that are needed for mapping their data instances.

The facts and dimensions are significant tables used to organize, build and construct relationships among multidimensional models through their instances that engage in interaction and communication among multiple dimensions. In case of petroleum systems and sub-systems analysis, there are volumes of facts and dimension tables, it is important to identify these facts and dimensions and review them for multidimensional modelling and mapping purposes. Key facts and dimensions are identified and interpreted. Facts may be dimensions and dimensions may also be facts. The relationships identified based on logical contexts and concepts may also be facts and or dimensions. In petroleum systems scenarios, there are several tools, procedures and processes used to integrate structure and organize these facts, so that the entity/dimension and its existence and survival are described in the form of an ecosystem. The author characterizes this fact, as a Total Petroleum System (TPS as global to PDE) concept, in which several elements and processes are interacted, communicated and shared among each other. *Source rock, reservoir, structure, trap* and *seal* including processes such as migration, maturity and timing of deposition in each and every oil and gas field in each and every sub- basin of Albertine Graben are interacted and shared each other, so that if one element is missing in one sub-system, that element is shared with other sub-systems. Several facts and dimensions are organized and structured in a way to get knowledge on interconnectivity among sub-systems.

The author does map the data dimensions in different star-schemas in different sub-basins of Albertine Graben. In these data schemas, there are multiple dimensions narrated and data relationships interpreted conceptually (logical data organization) among several dimensions (Nimmagadda et al. 2010b) and fact data tables, physically organized. It is significant to design intelligently/logically among these multiple dimensions in a way to understand or increase domain knowledge that is unknown. The author uses Oracle database procedures for documenting, organizing and mapping petroleum ecosystem models. For example, the knowledge built among fields, within a specific classification of attributes, such as API (A specific gravity scale developed by the American Petroleum Institute (API) for measuring the relative density of various petroleum liquids, expressed in degrees), reservoir quality, production rate, characterized by their properties, is significant for understanding domain knowledge of petroleum systems of sub-basins that produce for long periods of time in different geographic locations. Ontology models are constructed for shallow marine and deltaic petroleum systems, in which several dimensions are described (Nimmagadda et al. 2007a), connecting the factual data with one-to-one and one-to-many relationships (Hoffer et al. 2005). These models are represented in star-schemas, designed for Lake Petroleum systems (interpreted for Lake-Albert basin in Western Uganda).

The production data facts and their associated dimensions, for example play roles in ecosystems scenarios, integrating different exploration data from different petroleum systems and their sub-systems of Albertine-Graben. This fact is even conceptualized based on seismic sequences analysis and mapping, in which several producing horizons (seismic sequence) interact, communicate and share their production, including other facts and dimensions of petroleum system elements and processes. These facts can better be corroborated with factual data acquired from different producing horizons, in both *seismic* and *drilled-well* domains (Nimmagadda and Dreher 2007). Production data facts also depend on other facts such as reservoir type, API, source rock maturity and trap type. As narrated in Figure 4.103, an ontology model is described for building relationship of these facts and dimensions. Models built based on these relationships-facts, are used for extracting and mining data views for geological interpretations at later stages. For example, if a reservoir is fractured, the interconnectivity of fractures among different producing horizons is imminent. The phenomena of interconnectivity are well established through ontological conceptualization and contextualization.

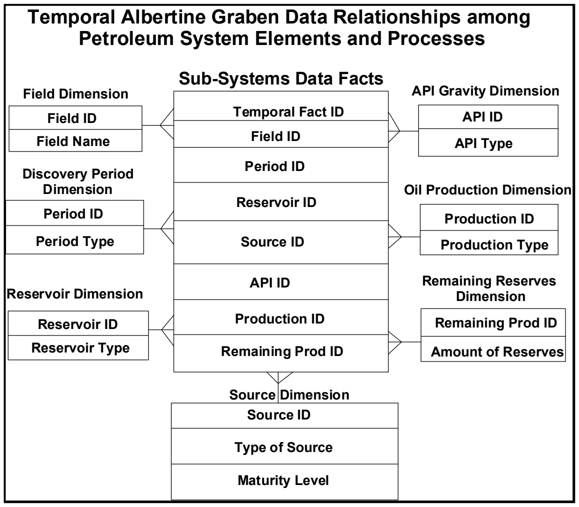


Figure 4.103: Temporal data attribute relationships and building an ontology model

The ecosystems are dynamic because of geographic and periodic nature. Facts and dimension instances relevant to location and period are acquired for modelling and accommodating them in a warehouse. These facts and dimensions are relevant to our petroleum ecosystem scenarios. A multidimensional model drawn depicting location dimension with hierarchies can ontologically interact with other models that have relations (Figure 4.104) to other petroleum eco sub-systems.

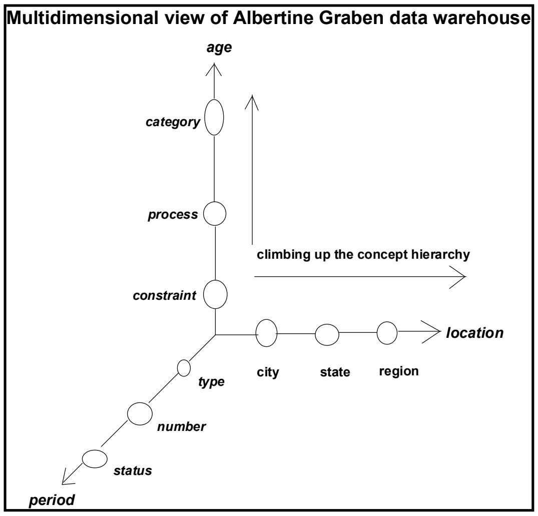

Multidimensional view of Albertine Graben data warehouse
age
category
process
constraint
climbing up the concept hierarchy
location
city
state
region
type
number
status
period

Figure 4.104: A multidimensional model, depicting the hierarchy of data attributes

Several hierarchical models are accommodated in a PDE warehouse. The warehouse schema should also facilitate coercion of different sources of the same type of data into a common semantic framework. If there is no common information of attribute or data property among data dimensions, then different DBs exist side by side within a warehouse environment, without having merged. Based on common data property information that exists in data structures, DBs get merged.

4.8.2.4 The Arabian-gulf case study

The Arab-gulf is another case study for addressing research questions and following up the research objectives (RO1 – RO7) mentioned in Chapter 1. The constructs and models used in this case study are analyzed in the Arabian Peninsula. There are number of prolific petroleum bearing sedimentary basins of Kuwait, Saudi Arabia, Iran and Iraq countries. The data sources in these regions are heterogeneous and multidimensional. There are common reservoirs and structures, producing commercial quantities of oil and gas for many decades are associated each other though geology has no boundary among these countries. In Arabian Gulf regions, large areal extents of gulf basins occupying onshore, offshore and transition zones, are sources of multidimensional data. Keeping in view, the volumes and sizes of *horizon-structure-reservoir-production* data structure in petroleum bearing sedimentary basins in unmanageable way, it is imperative to use robust warehouse and mining approaches, in which multiple dimensions are ontologically conceptualized, so that valuable data are more intelligently stored. Accessibility of domain knowledge is easy during data mining stage. Data integration, sharing of knowledge and interoperability are significant issues that are addressed in these basins. An ontology model, deduced in Figure 4.105, is an initial data structure, representing exploration data dimensions and facts.

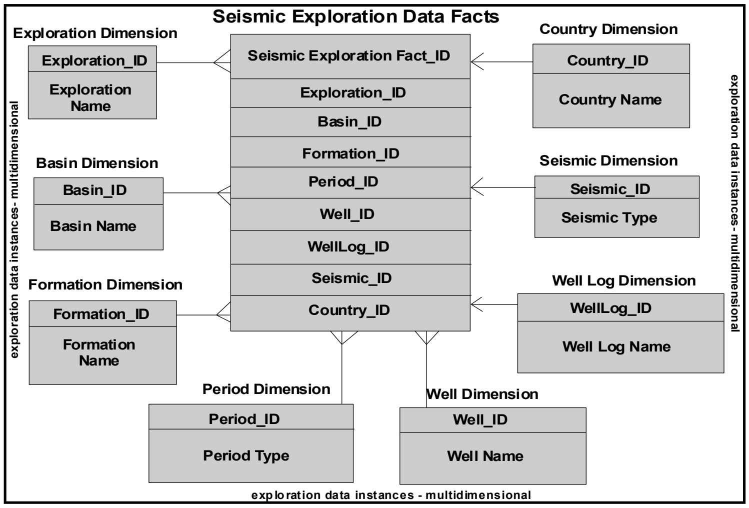


Figure 4.105: Seismic exploration data attribute dimensions and their facts

In addition, incorporating domain knowledge of seismic horizons in a warehouse modelling process, led to integration of multiple dimensions in multiple domains.

Seismic data instances play key roles connecting multiple dimensions in multiple domains of exploration and field development. As narrated in Figure 4.106, there are several *peaks* and *troughs* data instances of the seismic wavelets for each and every horizon, are analysed and interpreted without any ambiguity.

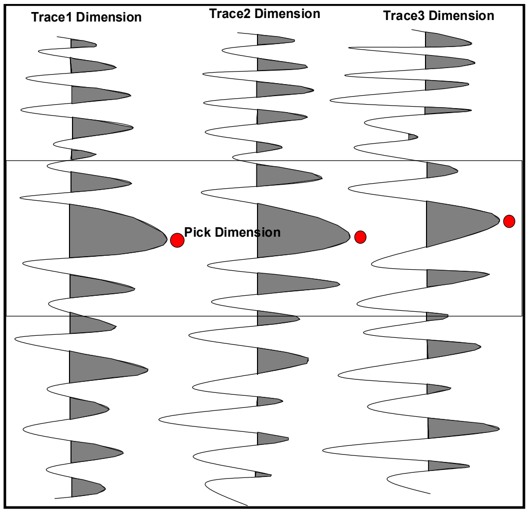


Figure 4.106: Flattened seismic wavelets (for different trace attributes) and their polarity pick dimensions, an example showing peak (dark) and trough (white) dimension relationships, for describing horizon ontology

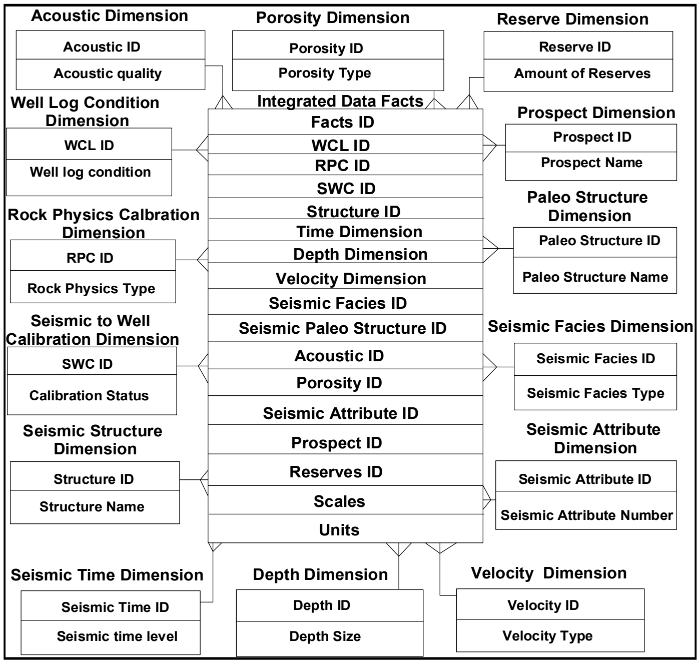


Figure 4.107: An integrated data schema, narrating relationships among several sub-sets and their respective sub-schemas

The magnitude, size and dimension (both time and space varying dimensions) of these peaks and troughs of the seismic wavelets are controlling the integrity of the relationship *structure-reservoir-petrophysical* structure and their properties (Figure 4.107). When there is change in shapes and dimensions of these peaks' and troughs' (Figure 4.106) instances, in a set of seismic wavelets, there is corresponding change in the properties of these dimensions. The author makes an attempt to analyse the connectivity of local level seismic wavelet to broader group of seismic wavelets of a set of horizons, interpreted in a petroleum system. Picking of peak and or trough for a horizon and maintaining database for these horizon data is an interpreter's task. As shown in Figures 3.36 – 3.38, data facts are collected and stored in a warehouse model, so that the connectivity among multiple horizons is established. Author designs an integrated framework incorporating the data warehouse and mining tasks. Entities or dimensions deduced in the models go through data structuring procedures, as per the type of data and whichever logic the data accept. Ontology descriptions support the logical data structuring procedures. One of such structuring approaches described in Figures 3.36 – 3.37, is an example of organizing the data in hierarchical ontologies.

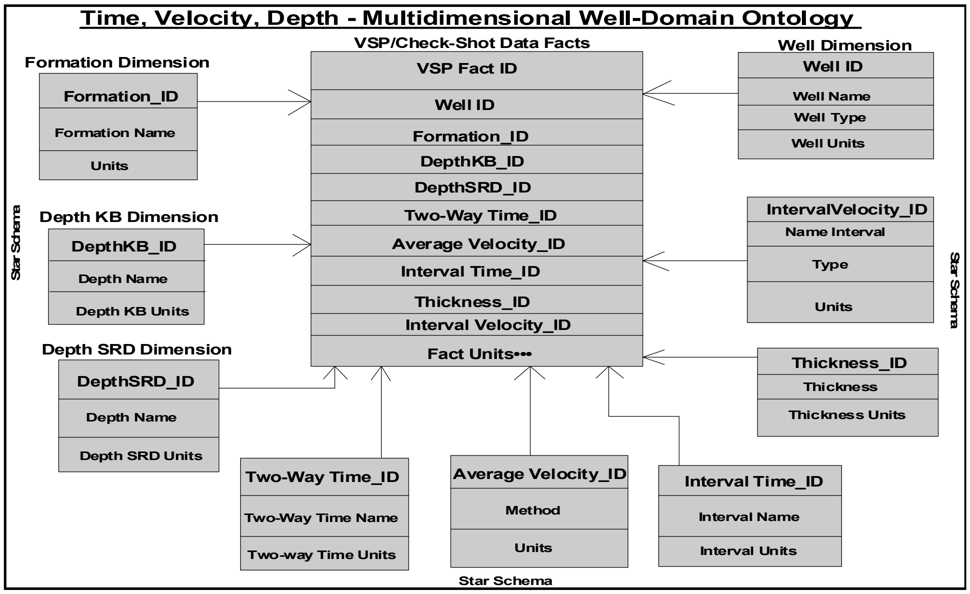


Figure 4.108: Domain ontology model connecting *geology-seismic* dimensions

What do these methodologies do in making structure models in Gulf basins?

For processing and interpreting the data, prior to integration, a core framework captures the processed data from several domain applications. Thus data are integrated from multiple fields and petroleum systems. During conceptual (ontology) development stage, an integrated metadata, can save enormous computing and time of oil and gas exploration during data-mining and knowledge extraction stage, other merits include:

1. This approach has flexibility to update scalable data attributes, depending upon the size of the oil-field and basin; thus meeting local and global geological situations (i.e., field to basin level hierarchies)
2. Our models consist of a package, in which petroleum data from different petroleum fields and basins are captured and intelligently structured through logical data organization.
3. Enormous amount of knowledge is extracted from producing petroleum fields in terms of varied attribute dimensions and thus assessing an implementation of an effective data mining approach, such as cluster mining, rule mining and decision-trees
4. Sink holes investigation, geological storages of CO_2, areas of H_2S, reservoirs polluted by CO_2 and guiding smart wells and also multi-lateral wells are few typical applications. Methodologies described so far address these domain applications in oil and gas industries. The following are typical stages of development of methodologies and procedures for generating an integrated metadata, considered for the gulf regions.

Significance

1. Integration and data sharing
2. Resolving complexity of heterogeneous time-depth/velocity data presentation – in several multidimensional views
3. Flexibility in changing data structures based on geology
4. Exploration and development risk minimization
5. Better understanding of data/views with least effort
6. Flexibility in changing and reusing the time-depth data structures quickly, (depending upon the fast changing geology situations, including drilling and production scenarios; addressing interoperability issues.

Stages of development (as per research objectives RO1-RO6, described in Sections 1.3.1 and 1.3.2 of Chapter 1)

1. Ontology – Specification, conceptualization and contextualization
2. Data warehousing – data integration
3. Data mining – classifications, clusters, design of rule mining and decision trees; extraction of knowledge
4. Data visualization – presentation of data views
5. Data Interpretation: interpretation of presented knowledge

The following procedure is followed for designing the data warehouse and data mining:

1. Acquire oil and gas data
2. Identify dimensions, entities and objects
3. attributes of petroleum exploration and production data
4. build ontology models using petroleum data entities/dimensions
5. structuring and de-structuring of complex relationships among data entities/objects
6. Identify all the dimensions/entities in relational, hierarchical and networked environments; star, snowflake and fact constellation schemas
7. Petroleum ontology; Build ER, UML, Logical multi-dimensional & implementation data models
8. Develop star, snowflake and fact constellation schemas constructing multidimensional logical and implementation data models; load data into Oracle database program for storing integrated petroleum- metadata in a warehouse environment

The conceptual schema design process is typically an iterative process, refinement and integration of views that involve:

1. Decomposition and/or synthesis of dimensions
2. Redefinition of relationships & relationship types
3. Redefinition of mapping constraints
4. Redefinition of high-level abstractions (e.g. conceptualization, generalization or specialization) based on semantics and contexts
5. Rearrangement of attributes among multiple dimensions (e.g. structure-reservoir-production dimensions and sets of their relationships)

Load the data into Oracle and cluster the data using cluster algorithms

1. Classify each cluster to qualify understanding of each dimension
2. Run SQL queries or other data mining logic – for accessing required data from warehoused metadata
3. Interpreting the data views – for extracting shallow, multidimensional, hidden and deep knowledge – by means of correlations, trends and patterns perceived from data mining; build statistical data models for future prediction of oil and gas
4. Extract hidden knowledge for interpreting petroleum geology

5. Knowledge discovery on petroleum systems, especially with fast changing geological situations (affecting velocity datasets, exploration phase) saving millions of dollars for drilling (drilling phase) hazards and during mining stages
6. IT/Communication technology opportunities must be utilized to understand these systems and datasets
7. To develop forecast models – for predicting knowledge, geologically interpretable, but maximise efforts in understanding petroleum fields and systems, extract knowledge from these Metadata systems.

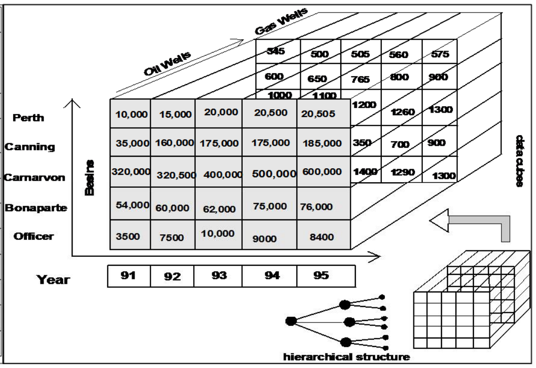
Modelling digital ecosystems of Middle Eastern Gulf basins

The author focuses on a petroleum data warehouse for integrating the multiple, heterogeneous data sources such as relational databases and OLTP files from gulf-basins. Since petroleum data reside in many operational units and in several geographic locations, a consistent encoding of data is necessary. Data integration is carried out to maintain consistency in the naming convention, measures of variables, encoding structure and physical attributes. There are hundreds of dimensions, narrated by petroleum systems and their elements, which are represented in hierarchies. These dimensions and numeric measures are carried out to facilitate in the multidimensional modelling of petroleum data. The author computes cuboids, OLAP and multidimensional data views for analysing the ontology base multidimensional petro-data as discussed in the next section.

Multidimensional petro-data model

The core design of the data warehouse (Pujari 2002) lies with a multidimensional view (Figure 4.109) of the petroleum data model. In order to understand this concept, the datasets shown as petroleum in-place in the gulf onshore and offshore basins (Al-Fares et al. 1998) are characterized by fields, wells, reservoir quality and by year. The rows and columns represent more than one dimension are observed, if the dataset contains more than 2 dimensions. The rows in the typical inventory, representing in two dimensions; *field* and *year* (Figure 4.109)*,* which are described as *field performance* first and then *year* (the order is arbitrary). The columns however do not really represent 2 distinct dimensions, but they represent some sort of taxonomy of a dimension. The *field performance* represents a hierarchical relationship between instances of *field class* and the instances of the *fluid type, such as oil production and* *water production*. We also examine the summary information, which is the main theme of the table. In our case, the summary function is sum.

Type	Field Attribute	Field Class					
Field Perf		Year	Umm Gudair (ltr)	SUG (ltr)	Wafra (ltr)	SF (ltr)	Humma (ltr)
	Oil Prod	91	5,000,000	5,500,000	4,500,000	4,750,000	5,250,000
		92	1,500,000	1,750,000	1,500,000	1,750,000	1,800,000
		93	2,000,000	1,750,000	1,800,000	1,800,000	1,900,000
		94	800,000	1,000,000	1,100,000	1,200,000	1,300,000
		95	1,000,000	1,200,000	1,100,000	1,300,000	1,500,000
	Water Satu	91	400,000	350,000	300,000	250,000	220,000
		92	150,000	200,000	250,000	100,000	150,000
		93	200,000	100,000	150,000	200,000	150,000
		94	90,000	80,000	100,000	150,000	100,000
		95	40,500	30,000	50,000	70,000	50,500



Basins					
Perth	10,000	15,000	20,000	20,500	20,505
Canning	35,000	160,000	175,000	175,000	185,000
Carnarvon	320,000	320,500	400,000	500,000	600,000
Bonaparte	54,000	60,000	62,000	75,000	76,000
Officer	3500	7500	10,000	9000	8400
Year	91	92	93	94	95

Figure 4.109: Oil and gas production profile representing it in a cube

Petro-data cubes

A popular conceptual model that influences data warehouse architectures is the multidimensional view of the petroleum data. A multidimensional cube-slice view of the information, representing *year* wise *oil/gas field performance*, is generated with structure-cube corresponding to a table narrated in Figure 4.109. This model views petroleum data in the form of petro data-cube. It has three dimensions, namely field performance, fluid type and year. Again each dimension is divided into sub-dimensions. In a multidimensional petro-data model, there is a set of numeric measures that are the main subject of the analysis. In our study, the numeric measure is field performance. There may be more than one numeric measure. Some examples of numeric measures are reserve inventory, well plans, production performance, reservoir quality etc. Each numeric measure depends on a set of dimensions, which provide context to the measure. All the dimensions together are assumed to uniquely determine the measure. Thus the multidimensional data represented as measures, are values placed in individual cells in the multidimensional space. Each dimension, in turn, is described by a set of attributes. The attributes of a dimension may be related or linked to a hierarchy of relationships or by a lattice of cube.

Petro-dimensional modeling

The notion of a dimension provides a lot of semantic information, especially about the hierarchical relationship between elements. It is important to note that dimensional modeling is a special technique for structuring data around business concepts. Unlike ER modeling, which describes entities and relationships, dimension modeling structures numeric measures and the dimensions. The dimension schema represents details of dimensional modeling. The dimension hierarchy facilitates the view of multidimensional data in several data cube representations. Conceptually, multidimensional data is viewed as a lattice of cuboids.

Data mining (OLAP) and implementation of integrated data schemas

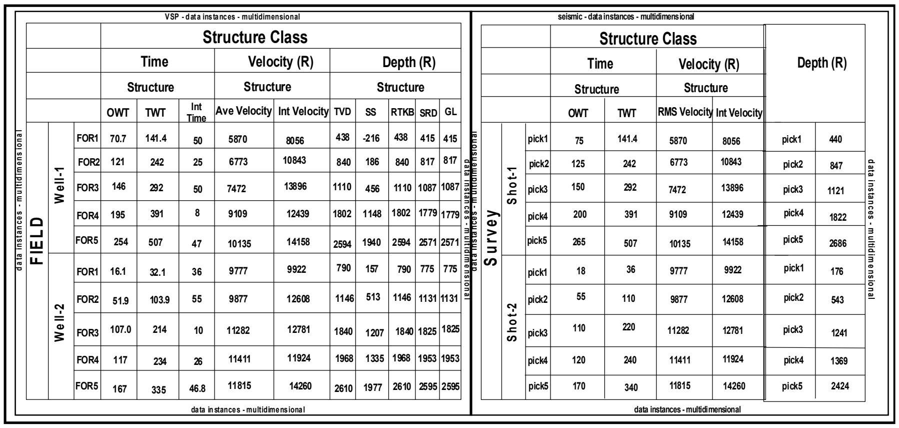
Once an ontology base data warehouse is ready with representation of multidimensional data cubes, it is necessary to explore the exploration data using different OLAP tools, for building and interpreting the knowledge of OLAP views. The survey and oil-field base ontologies are defined for each petroleum system of a basin for knowledge mapping purposes. Slicing, dicing, drilling, drill-up, drill-down, drill-within, drill-across and pivot are some of the operations on petro multidimensional data cubes, which further allow interactive querying, analysis and interpretation of these exploration data. According to the underlying multidimensional view, classification hierarchies are defined for each dimension. These tools are used to access the live data on line and analyse them for interpretation. In the multidimensional model, petro-data are organized into multiple dimensions and each dimension contains multiple levels of abstraction. The author uses multidimensional views, taken from ontologically derived petroleum data for geological interpretations and detailed modelling studies. This model is used for well planning and economic evaluation as well.

Time-depth modelling – Arabian –Gulf perspective

An ontology is a modelling procedure of specifications of concepts (Meersman 2004, Noy and McGuinness 2000 and Nimmagadda and Dreher 2007) and contexts that are organized logically with certain business constraints. Similar conceptually described entities or dimensions in different hierarchies and relationships are representative in gulf basins. Modelling multiple dimensions is a core concept in the data warehouse structuring (Pujari 2002) in the time-depth modelling analysis of gulf petroleum provinces. Various issues involved in presenting time-depth data in a data warehouse and how ontologies can come to the rescue of these issues, are discussed in the following sections.

Time-depth-velocity datasets

The author transforms scalable seismic (in time) datasets into corresponding depths needed for structural data interpretation. The depth-domain dataset is scalable and but independent of velocity. The seismic times and RMS velocities derived after NMO (normal move-out) processing (Sheriff 2002) are used to create an initial stacked dataset in time-domain and these datasets are further improved qualitatively, keeping in view data interpretation objectives by refining velocity datasets (Figure 4.110). Velocities derived in the seismic-domain are obtained from "normal moveout" (Liner 1998) corrected (seismic) events. These are basic RMS (root mean square) velocity datasets. In a warehouse environment, author integrates velocity datasets at drilled wells and seismic survey grids for building time-depth conversion. The author uses VSP and check-shot data to calibrate seismic horizons with well-top datasets (formation tops). The datasets involved in these calibrations are briefly described for integration of seismic domain (with time dimension) and well-base domain ontologies (with depth dimension) in the following sections.



VSP - data instances - multidimensional

			Structure Class										
			Time			Velocity (R)		Depth (R)					
			Structure			Structure		Structure					
			OWT	TWT	Int Time	Ave Velocity	Int Velocity	TVD	SS	RTKB	SRD	GL
FIELD	Well-1	FOR1	70.7	141.4	50	5870	8056	438	-216	438	415	415
		FOR2	121	242	25	6773	10843	840	186	840	817	817
		FOR3	146	292	50	7472	13896	1110	456	1110	1087	1087
		FOR4	195	391	8	9109	12439	1802	1148	1802	1779	1779
		FOR5	254	507	47	10135	14158	2594	1940	2594	2571	2571
	Well-2	FOR1	16.1	32.1	36	9777	9922	790	157	790	775	775
		FOR2	51.9	103.9	55	9877	12608	1146	513	1146	1131	1131
		FOR3	107.0	214	10	11282	12781	1840	1207	1840	1825	1825
		FOR4	117	234	26	11411	11924	1968	1335	1968	1953	1953
		FOR5	167	335	46.8	11815	14260	2610	1977	2610	2595	2595

data instances - multidimensional

seismic - data instances - multidimensional

			Structure Class				Depth (R)	
			Time		Velocity (R)			
			Structure		Structure			
			OWT	TWT	RMS Velocity	Int Velocity		
Survey	Shot-1	pick1	75	141.4	5870	8056	pick1	440
		pick2	125	242	6773	10843	pick2	847
		pick3	150	292	7472	13896	pick3	1121
		pick4	200	391	9109	12439	pick4	1822
		pick5	265	507	10135	14158	pick5	2686
	Shot-2	pick1	18	36	9777	9922	pick1	176
		pick2	55	110	9877	12608	pick2	543
		pick3	110	220	11282	12781	pick3	1241
		pick4	120	240	11411	11924	pick4	1369
		pick5	170	340	11815	14260	pick5	2424

data instances - multidimensional

data instances - multidimensional

Figure 4.110a: A typical data instances list, showing structure classes

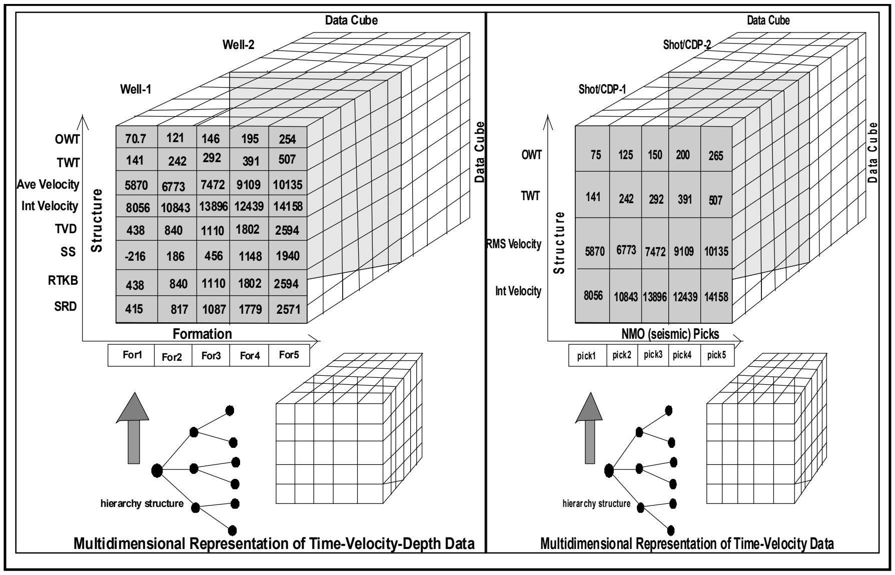


Figure 4.110b: Data instances list, representing the structures in cuboid slices

***Check-shot surveys**:* These data instances connect surface to subsurface data. In a type of borehole seismic survey, seismic travel times are measured from the surface to a known depth. P-wave velocity of the formations encountered in a well-bore is measured directly by lowering a geophone to each formation of interest, sending out a source of energy from the surface of the Earth, and recording the resultant signal. The data are correlated to surface seismic data by correcting the sonic log and generating a synthetic seismogram to confirm or modify seismic interpretations. This differs from a vertical seismic profile in the number and density of receiver depths recorded. The geophone positions may be widely and irregularly located in the well bore, whereas a vertical seismic profile (VSP) usually has numerous geophones positioned at closely and regularly spaced intervals in the well bore (Lowrie 1997).

***Vertical Seismic Profiling (VSP) surveys**:* It is a method of measuring borehole seismic travel times for correlation with surface seismic data (Sheriff and Oliveira 2002) to obtain images of higher resolution than surface seismic images - also called Vertical Seismic Profiling (VSP). The VSP refers to measurements made in a vertical well bore using geophones inside the well bore and a source at the surface near the well. VSPs vary with well configuration, the number and location of sources and geophones, and how they are deployed. Most VSPs use a surface seismic source which is commonly a vibrator on land and an air gun in offshore or marine environments. It is a much more detailed survey than a check-shot survey because, geophones are more closely spaced, typically on the order of 25 m [82 ft], whereas a check-shot survey might include (Lowrie 1997) measurements of intervals hundreds of meters apart. VSP survey uses the reflected energy contained in the recorded trace at each receiver position as well as the first direct path from source to receiver. The check-shot survey uses only the direct path travel time data. In addition to matching well data to seismic data, the vertical seismic profile also enables converting seismic data to zero-phase data and distinguishing primary reflections from multiples.

Initially, the author describes conceptualized relationships ontologically well-wise and field-wise among different dimensions. Relationships among check-shot and VSP datasets with seismic and well datasets are further ontologically conceptualized. These data structures are integrated in a warehouse for mining different data views for future seismic and geological interpretations. The author discusses an overview of ontology and how it works in modelling time-depth datasets in the following sections.

Need for ontologies in the gulf basins

The gulf regions occupy multiple basins, petroleum systems and multiple oil and gas fields in hundreds of square kilometers of areal extents. The data sources in these regions are heterogeneous and multidimensional. The connectivity and understanding the phenomena of mutual interaction among attributed elements, processes and chains of petroleum systems are needed for enhanced understanding of exploration in the sedimentary basins of gulf regions. Exploration datasets are characterized at the atomic level, isolated, and in non-hierarchical structures. Different databases may describe different aspects of the same geological formation and the relationships among them must have been established by links that are not intrinsically part of the data archives themselves. Development of individual databases has generated a large variety of formats in their implementations. Intercommunication between databases (Ozkarahan 1990 and Wand et al. 1999) of different structures and formats implies the need for common semantic standards. The problem is especially acute in comparing entities and their relationships among several petroleum fields and sedimentary basins. The author addresses these problems by using ontologies for integrating and sharing knowledge in petroleum geology or petroleum exploration. The author creates a repository of terms and concepts relevant to petroleum exploration, hierarchically organized by means of “is a subset of”, or “is member of” operators. In this classical case, velocity data are constructed through relating time-depth entities (or dimensions) in both seismic-well ontology domains. Similarly, depth-data are constructed by interrelating the time and velocity; time data are built interrelating the depth and velocity datasets. As an ontology-based warehouse modeler, this interpretation needs to be translated into common goals and interoperability since structures built based on these relationships can be reused in other oil fields of other basins.

In the case of Arabian Gulf situations, ontology is a formal explicit description of concepts of petroleum systems (Meersman 2004) in a domain of discourse (called *classes* or *concepts*), properties of each concept describing various features and attributes of the concept, and restrictions on roles or properties. It is a set of individual instances of classes that constitute a *knowledge base* in the Arabian Gulf exploration domains. In basic exploration ontology, the author describes several domains with different relationships and then classified with certain constraints applied; time-depth, time-velocity, depth-velocity relationships are ontologically conceptualized (Figure 4.111). A brief outline is discussed here.

Decomposition

- Bottom-up dimension modelling

- Commitment modelling (making connections with constraints)

Stepwise

- Extract (data acquisition scheme)
- Abstract (conceptualization of data entities)
- Organize (logical organization of entities and relationships)
- Specify commitments (instantiations, constraints, connection of dimensions)
- Configure commitments (introduce layers, templates)

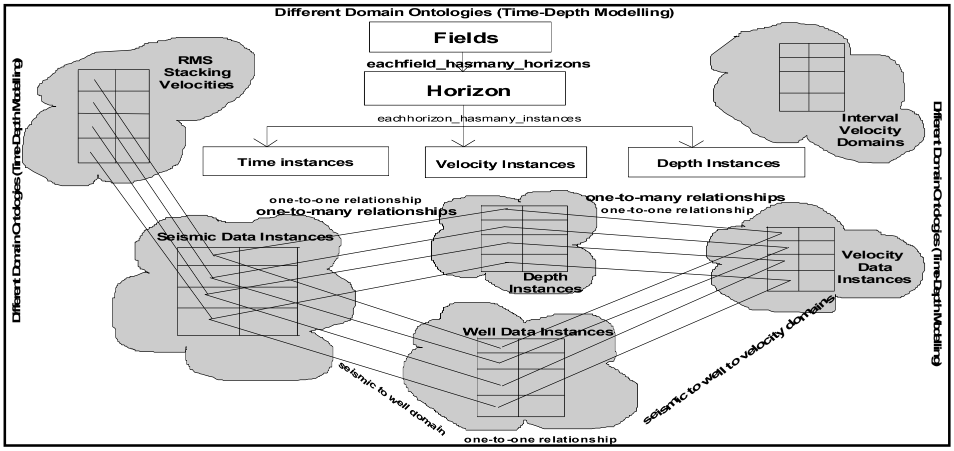

Different Domain Ontologies (Time-Depth Modelling)
Fields
eachfield_hasmany_horizons
Horizon
eachhorizon_hasmany_instances
Time instances
Velocity Instances
Depth Instances
RMS Stacking Velocities
Interval Velocity Domains
Different Domain Ontologies (Time-Depth Modelling)
Different Domain Ontologies (Time-Depth Modelling)
one-to-one relationship
one-to-many relationships
one-to-many relationships
one-to-one relationship
Seismic Data Instances
Depth Instances
Velocity Data Instances
Well Data Instances
seismic to well domain
seismic to well to velocity domains
one-to-one relationship

Figure 4.111: Relational ontology – connecting time, depth and velocity entities (or dimensions) and building relationships

Understanding data warehouse models

A data warehouse is a subject-oriented, integrated, time-variant, non-updatable collection of data (Coronel et al. 2011) used in support of managing decision-making process and business intelligence. The data warehouse is organized with high-level entities, as an example, *exploration*, *drilling*, *production* and *marketing* (Nimmagadda and Rudra 2004a and Nimmagadda and Rudra 2004b) in petroleum industries. More specifically, in *exploration* entity (Figures 4.5 and 4.6), basins, oil fields, horizons, seismic times, velocities and depths are described as sub-type entities or dimensions in multidimensional logical data representations. These data are stored in a warehouse environment using consistent naming conventions, formats, encoding structures and related characteristics gathered from several internal systems. Here the author conceptualizes ontologically, the relationships among entities and attributes to build and extract knowledge of *basin* and *exploration* entities (Figures 4.5 and 4.6). The time-depth conversion and deriving appropriate velocities for each field (survey grid and well locations) are key aspects of data interpretation. In the ontology modelling, the author derives time, depth and velocities, conceptualized entities and or dimensions from an *exploration* entity hierarchy (Figure 4.5).

Two different categories of data, *velocity* and *time* are used in the modelling process. Warehousing approach organizes seismic data (of time domain) as per *coordinates* and *points*, *surfaces* and even *contours* in several fine-grain structures (to their lowest atomic structure), regrouping into multiple dimensions and fact-tables of actual seismic data. An ontology illuminates seismic domain attributes (Nimmagadda and Rudra 2004b), their instances and locates the drilled-well point in depth domain within the vicinity of seismic survey grid. Warehouse framework is responsible for calibrating the time data into depth domain through velocity attributes. Wherever factual velocity data are found, a procedure multiplying with seismic time values to get a depth point, is followed for all multiple horizons (of varying depths). Thus, the author makes use of these attributes available for modelling multidimensional data in a warehouse environment, in which time, depth and velocity dimension tables are constructed for multiple horizons (Figure 4.112). The respective fact data tables are reorganized for documenting all the horizons, interpreted in a given seismic volume.

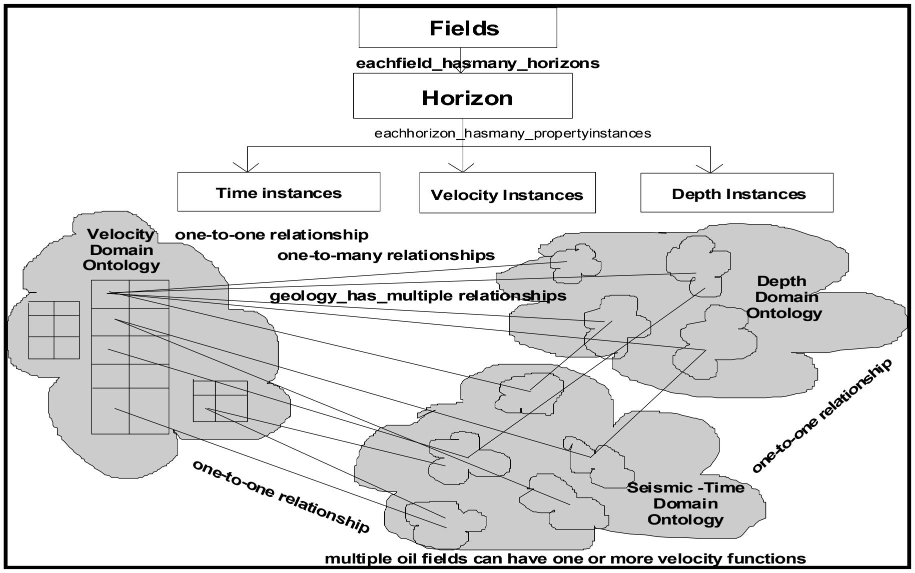

Fields
eachfield_hasmany_horizons
Horizon
eachhorizon_hasmany_propertyinstances
Time instances
Velocity Instances
Depth Instances
Velocity Domain Ontology
one-to-one relationship
one-to-many relationships
geology_has_multiple relationships
Depth Domain Ontology
one-to-one relationship
one-to-one relationship
Seismic -Time Domain Ontology
multiple oil fields can have one or more velocity functions

Figure 4.112: Data integration process model - relational ontology

Velocity models

A velocity model (Parasnis 1997) is a list of low-level data types that can be grouped together in a geological sense, including:

- Velocity functions (both processed and user defined)
- Time-depth tables (both processed and user defined)

- Geologic/seismic horizons (interpreted and user defined)

“Exploration” is a broad entity or dimension (an exploration ontology describing all data dimensions is given in Figures 4.5 and 4.6) consisting of all sub-dimensions such as described in the seismic and well-domains. Oil fields are connected in a warehouse model by several surveys and wells. Geologists, geophysicists and reservoir engineers acquire lot of data from oil field areas in the form of time-depth tables that characterize velocities. These data are also computed using different software modules (such as Geographix, OpenWorks, Petrel workstations). Initially, broad view of velocity model is constructed using well-base time-depth data tables. The author updates these models, modifying and refining using RMS velocity datasets in between wells and away from wells, by adding or removing entities (or dimensions), or editing individual data dimensions within volumes stored in a warehouse framework. The warehouse approach allows us to revise the existing models and build new velocity models as well, which are capable of further integrating with other data dimensions of wells located adjacent to the current datasets. The time-depth volume is simulation for all points of seismic survey layouts and the procedure is continued for each individual field in a total basin. Thus a metadata volume is generated.

Time-depth datasets

A time-depth volume is a set of X-Y-Z points within a volume that can be interpolated using linear and triangulation methods. This volume is linked through velocity model that generated – usually by sharing a common name using shared ontologies of seismic and well-data domains. Using time-depth volumes and grids, ontology-base warehouse performs time-depth conversions of points, surfaces, seismic time traces and other common data objects types. Warehouse uses 2D or 3D field layouts for organizing these datasets. Warehouse generated velocity models, consist of several vertical time-depth functions. Each time-depth function is a series of time-depth pairs at a single surface location depicting each point object or dimension.

Constructing multidimensional data structuring for warehouse modelling

The author interprets each entity discussed in the ontology modelling as a dimension in the multidimensional data mapping (Pujari 2002 and Nimmagadda et al. 2006). In a multidimensional structuring approach, author denormailzes all the entities and or dimensions and their relationships and including conceptualization and contextualization stages. This form of representing multidimensional tables is popular in statistical data analysis. The rows represent two dimensions; *field* and *formation,* which are ordered as field first, then formation (the order may not be arbitrary, since formation is represented in the order of age from bottom to top). The columns, however, do not really represent two distinct dimensions, but they do represent some sort of taxonomy of a dimension. The structure class and structure represent a hierarchical relationship (Figure 4.113) between instances of structure class and the instances of the structure.

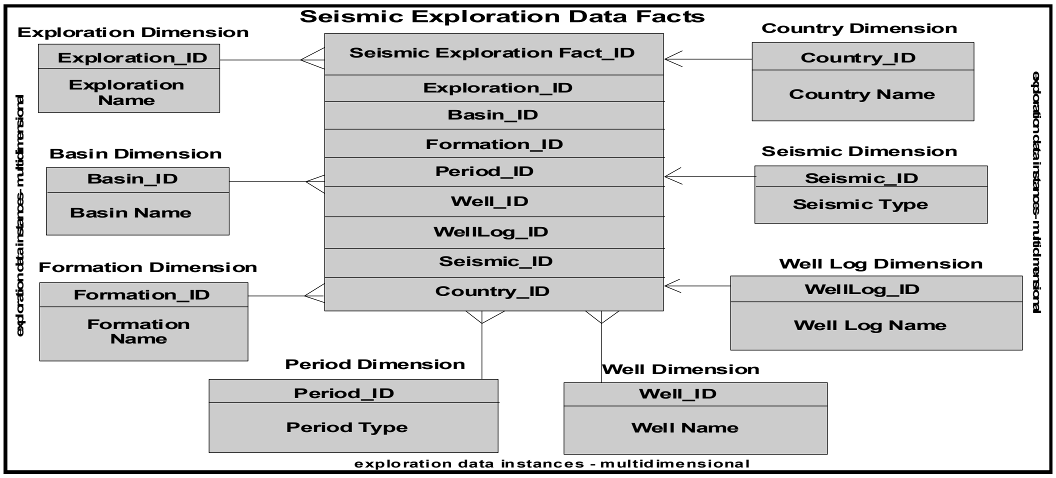


Figure 4.113: Multidimensional data structuring of "exploration"– a star schema model

Data descriptions in multidimensional petro data cubes

A popular conceptual model that influences the data warehouse architecture (Pujari 2002, Hoffer et al. 2005 and Jukic and Lang 2004) is a multidimensional view of petro data. This model views data in the form of a data cube, which has three dimensions namely, *field*, *structure* and *formation*. Each dimension is subdivided into sub-dimensions. In a multidimensional data model there is a set of numeric measures which are used as the focus of, or context for analysis. In this example, the numeric measure is *seismic time.* Each numeric measure (e.g. depth, velocity) depends on set of dimensions which provide the context for the measure. The contexts and semantics are referred from their respective ontologies. All the dimensions structured in Metadata model uniquely determine measures of attribute strengths and magnitudes. Thus, multidimensional data view is a measure and a value placed in a cell in the multidimensional space (Nimmagadda et al. 2005c and Nimmagadda and Rudra 2004b). Each dimension, in turn is described by a set of attributes. In general terms, dimensions are the perspectives or entities or concepts with respect to a field, which are described by well, seismic and formations. "Geological formation" is again described by its "structure" and "structure" is described by "seismic time" and or "geological depth". "seismic-times" make link with velocity or geological-depth with velocity. That is how, relationships are constructed among time, depth and velocity dimensions, further implying that on a broad view, conceptualized relationships are from hierarchy of relationships and through a lattice. In our case, each "drilled well" possesses several "geological formations" in a hierarchy relationship. In a conceptual model, a constraint is made, such as; each "drilled well" must have one or more "geological formations". Each "geological formation" represents a unique set of "seismic time" events at several CDP/SP (in a spatial domain). Each set of seismic data instances characterizes with structural "highs" and "lows" anomalies. More such anomalies are modelled, as given in Chapter 5.

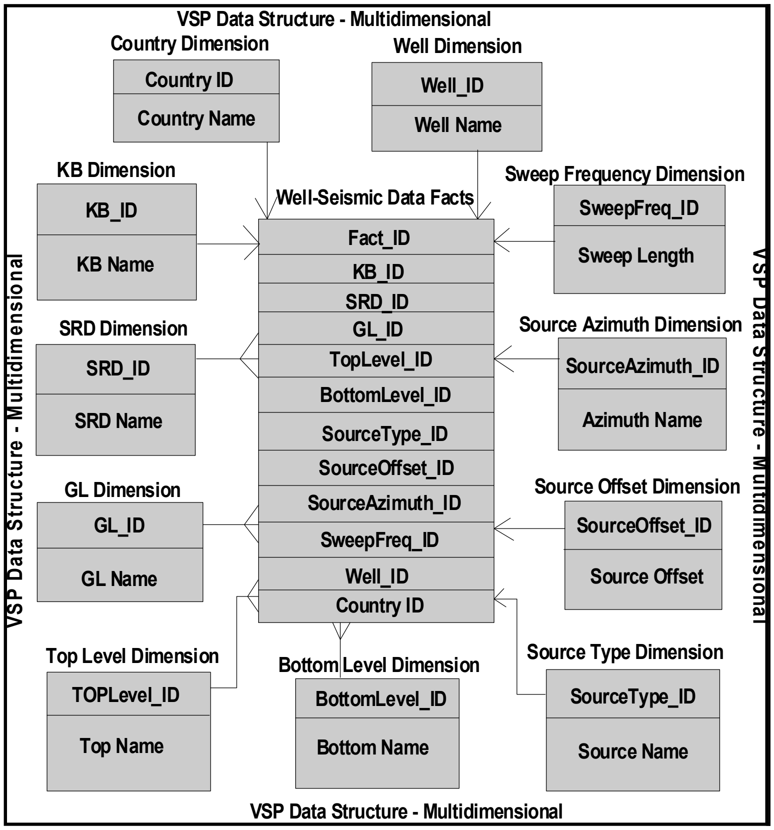
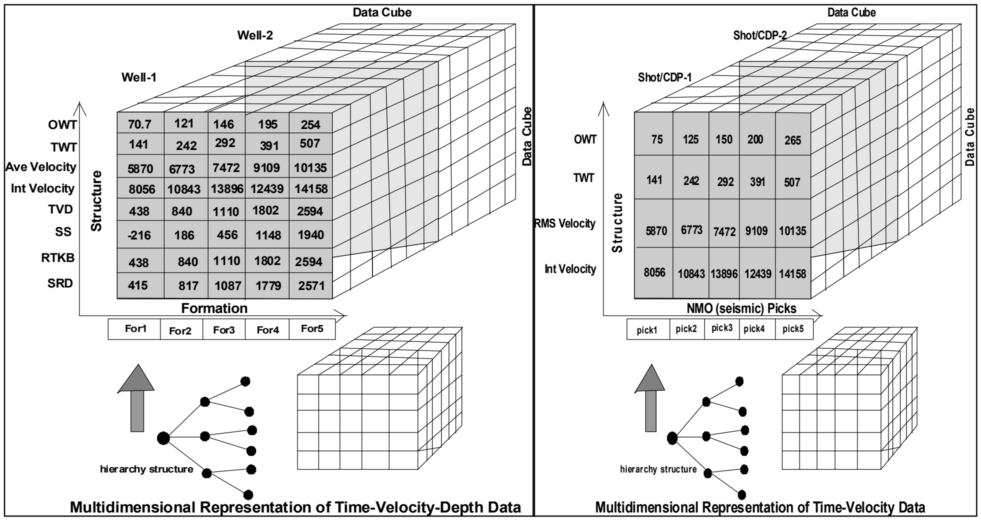


Figure 4.114: VSP Ontology – data descriptions

In a prospect area, the author interprets dimensions in two key domains, seismic and well-domain. Wells are spatially distributed in a prospect area. Datasets describing these two domains are different, as one is based on the information at well locations (time-depth tables derived from sonic logs, VSP- Vertical Seismic Profiling, check-shots and synthetics, Figure 4.114) and the other is based on information away from the wells (derived from the seismic data processing, velocities and regional geology). The author uses the datasets in the integration process from two different domains using standard calibration procedure (Lowrie 1997). The calibration is a part of data interpretation procedure, in which all geological formation tops (in well domain) are scaled to the seismic horizon (in seismic domain) using time-velocity, time-depth and depth-velocity data tables (Figure 4.115).



	For1	For2	For3	For4	For5
OWT	70.7	121	146	195	254
TWT	141	242	292	391	507
Ave Velocity	5870	6773	7472	9109	10135
Int Velocity	8056	10843	13896	12439	14158
TVD	438	840	1110	1802	2594
SS	-216	186	456	1148	1940
RTKB	438	840	1110	1802	2594
SRD	415	817	1087	1779	2571

	pick1	pick2	pick3	pick4	pick5
OWT	75	125	150	200	265
TWT	141	242	292	391	507
RMS Velocity	5870	6773	7472	9109	10135
Int Velocity	8056	10843	13896	12439	14158

Figure 4.115: Multidimensional representations of “velocity” – data cubes

Alternatively, the author loads datasets independently and then relationships are built based on the constraints extracted from ontology. Several frameworks, their applicability and feasibility are discussed in (Jasper and Uschold 1999). A framework of ontology based warehousing, proposed in the current study as presented in Figures 3.36 and 3.37 is used. To this extent, the author discusses multidimensional metadata with fine and coarse grained structuring as in (Rudra and Nimmagadda 2005). While designing the data structure, grain size is taken into consideration since it has great impact on data-mining. In the present case, time, depth and velocity attribute events discussed in different seismic and well-based ontologies are hierarchically structured and their instances are linked through a warehouse integration process until finer grain size is achieved (denormalization). Datasets acquired in the present study are organized in different hierarchies and relational structures. Relationships built among these hierarchies are further conceptualized into several dimensions. An ontology framework (Jasper and Uschold 1999) used in the present study characterizes seismic and well domain ontologies. These are further structured and integrated in multidimensional form (Pujari 2002).

Data needed for warehouse framework

- Time-depth tables stored multidimensional structures

- Time-depth functions stored in ASCII files
- RMS, Interval and Velocities stored in ASCII files

When input data are RMS (root-mean-square) or average velocities, the warehousing model converts them into interval velocities, which are further used to construct time-depth functions to be stored and integrated into a metadata. The models derived from time-depth tables at well locations are used as reference for calibrating models derived from seismic velocities. Well based velocity models are integrated with seismic derived velocity models. Map grids, seismic horizons, fault polygons and seismic data may be used in the ontology-based data warehouse modelling.

Types of data converted into *time-depth* pairs

At every survey *point* dimension, pair of time and velocity datasets is interpreted with interval velocity instances tabulated. Survey lines, 2D/3D profiles, user grids stored in the warehouses, horizons data, polygons derived from structures (such as faults, folds, joints and unconformities)**,** pointed seismic times and well data (including deviated well polygons) are typical datasets used**.** Time-velocity pairs are derived in the seismic domain from normal moveout (Liner 1998) corrected (seismic) events.

How ontology-base data warehouse computes time-depth models

Once all available time-velocity-depth data tables are set up in a warehouse environment, the modelling approach enables one to convert time-to-depth or depth-to-time data with reference to seismic and well domain ontologies. Data points are taken along user chosen survey grids. More specifically, for any surface horizon (x, y, t), it is possible to extract the corresponding depth from the corresponding time or velocity volumes. For any depth horizon surface (X, Y, Z), it is possible to compute corresponding time. For each input data point, our warehouse approach computes depth equivalent by linear interpolation of time-depth volume.

Points, lines, contour surfaces, and *seismic times* are typical data dimensions. Time or depth (vertical) and shot-to-shot or well-to-well distance (lateral) are user specified dimensions. A time-depth volume is a set of x-y-z points within a volume that interpolated using linear and triangulation methods (Lowrie 1997, Al-Fares et al. 1998 and Castañeda Gonzalez et al. 2012). The time-depth volume does not retain information as to the source or origin of its points, nor can it be meaningfully edited. It is however linked to the velocity model that was used to generate it usually by sharing a common name. Time-depth volumes are the raw materials that our warehouse model uses to perform time-depth conversions of points, surfaces, seismic traces, and other common data object types.

Relational and hierarchical dimensional data structuring

For the purpose of computing time-depth-velocity and connecting pairs of these dimensions in different hierarchies and relationships from both seismic and base-base domains, it is convenient to interpolate and/or extrapolate among survey grids. The procedure for gridding these datasets is discussed here.

Computation of data grids

For each point of observed seismic time, the warehousing approach computes depth point by connecting neighboring points, extracted as per survey grids, either by hierarchical or relational structuring methods. By warehouse procedure, velocities from all the interpreted horizons are gathered from different dimensions, all vertically, horizontally and laterally and integrated to generate a Metadata velocity model. Velocities, close to the drilled-well are computed using depths of geological formation tops. Velocities close to the drilled point is more accurate; however velocities are modeled away from well points. Using integrated velocity Metadata structure, a depth volume is computed. Structuring is carried out in a hierarchical manner, if horizontal or lateral dimensions need to be addressed. For each CDP/SP (common depth point or shot point) distribution on survey grid, a depth-point is computed. Depth is computed by linear interpolation of two time-depth pairs, a function derived, close to interpreted CDP/CMP trace.

The warehousing framework gathers all the data points together based on survey grids and permits computation of all depth points from the collected velocity datasets. Warehouse modelling can also permit lateral triangulation connecting the x-y locations of the time-depth functions. For example, to convert a data point x,y,t to depth, warehouse process locates and determines in which schema, the point lies. At each schema, the depth at “t” is computed by linear interpolation of the two time-depth pairs in the function that are close to t. The linear interpolation across time-depth functions (along the triangle) is done to determine the depth at x,y,t.

Gridded data volume

Data points extracted in the required grids are stored and all these points are interpolated and extrapolated within known domains, such as seismic and or well-base domains. Warehousing also reads gridded time-depth datasets generated by other applications from different kinds of models. An interpolation is either carried out in time domain or depth domain with a fair matching of time-depth values away from well data points. At well locations, a good match between seismic and well tops is maintained. The velocity model is composed of vertical time-depth functions each of which is a series of time-depth pairs at a single surface location. Warehouse constructed model can have as little as one time-depth function, but more functions result in greater detail and reliability.

Ontology modelling and building time-depth models

The author stores time, depth tables, time-depth functions, RMS, average and interval velocities in the databases. Interval velocities are computed using time and depth data, tabulated in a warehouse environment. For this purpose all the data are modelled using ontology-base data warehousing and mining approaches, representing alternate approaches to solving problems associated with time-depth conversions. All naming and semantic conflicts involved in time to depth conversion are handled by the ontology. Ontologically derived data instances are warehoused for integrating seismic and well datasets and used for mining different data views for the purpose of seismic and geological interpretation. The author interprets some of the data views taken from these metadata in the next section.

Petro-data cluster mining – Arabian – Gulf perspective

The basic idea of clustering petroleum business data is to integrate and merge similar properties of oil-play factors which otherwise remain undiscovered and hence unintelligible. Benefits of this approach include inter-operability and knowledge reuse. As stated earlier, clustering, in the present context is broadly described as identification of similar property or characteristics petroleum business data and investigating their relationships. More specifically, oil and gas data entities, possess in general spatio-temporal dimension. Spatial data are in the form of X, Y, Z coordinates. Historical data are periodic in nature. The author has described details of data types in (NImmagadda and Dreher 2007). Exploration, drilling, production and marketing are key activities of any petroleum producing company. Exploration, as a super-entity has several sub-type entities or objects, such as geology, geophysics, well logging, reservoir, logistics and inventory. Similar sub-type entities or objects can be derived from drilling, production and marketing operational data. An ontological framework (Nimmagadda and Dreher 2007) is derived keeping in view an overall petroleum exploration business data. Different data visualization software is used to present the extracted data views for interpretation. The author discusses a mining methodology in the current study in the following sections.

Cluster mining

The entities, attributes and instances of petroleum database clustered in multiple dimensions (Pujari 2002) are used for mining various patterns. Multidimensional data modeling is carried out for warehousing large size petroleum data in gulf-basins. Data cubes and their views further facilitates clustering algorithms to group and categorize *similarity* property attributes. Data structures are denormalized (Coronel et al. 2012) to arrive at finer groups that may provide more knowledgeable and intelligent clusters. However, the inclusion of irrelevant attributes can be futile to a successful clustering outcome because they negatively affect proximity measures and eliminate clustering tendency. A sound Exploratory Data Analysis (EDA) is required as given in (Becher et al. 2000). EDA eliminates inappropriate attributes and reduces cardinality of the retained categorical attributes. EDA provides attribute selection, but cluster-specific attributes are yet to be invented - attribute scaling is viewed as the continuation of attribute selection. Applications that derive their data from measurements may have an associated amount of noise, which can be viewed as legitimate records having abnormal behavior even if clustering techniques do not distinguish between noise and the abnormalities that fit into clusters. Pre-processing of data, such as partitioning or data summarization, may help in minimizing the outliers.

Clustering of petroleum data is a method by which large sets of exploration and production data are grouped into clusters of smaller sets of similar data. The examples shown in the following sections demonstrate the clustering into balls/bubbles of similar and dissimilar colors. Author is interested to group of different balls of similar colors into particular set of groups. Clustering implies grouping of exploration and production data or dividing a large set into smaller data sets with some similarity. The similarity property of attributes of petroleum business data entities or objects is the criterion used in the present studies by plotting similar attribute values from different oil fields and basins in the same scale. Exploration and production characteristics of different wells in different fields of different basins can be revealed if plotted in the same scale to permit their visualization along with their densities (distributions) and magnitudes The relative shift or position of clusters shows which cluster characterizes and contributes more to oilplay of a petroleum system and thus informs the exploration and development of an oil field.

Warehouse schemas contributing to finer data clustering in Gulf basins

Multidimensional and object oriented data schemas are deployed keeping in view what is expected for interpreting finer clusters of petroleum business data. In general, relationships among the common attributes of petroleum-plays data are denormalized, so that the final data clusters become finer. In the present study, data relevant to performances of petroleum systems have been organized in their appropriate data structures, for the purpose of fine-grained data mining processes (Keogh et al. 2001 and Rudra and Nimmagadda 2005). The multidimensional model consists of several fact tables (surveys, wells and permit facts or oil-play factors) surrounded by hundreds of dimension tables (Nimmagadda and Dreher 2007) in several star-schemas. Being a fact constellation schema, high-level granularity (Pujari 2002) has been maintained in order to derive fine-grained queries. The combination of star, snowflake and fact constellation schemas definitely with normalized and denormalized relationships can expect to generate fine-grained clusters from large volume of data warehouse. The above *surveys, wells and permits* schemas, though represented as star schemas, are also used for designing fact constellation schemas.

The relationships among sub-type data dimensions are conceptualized through ontology (Meersman 2004 and Nimmagadda and Dreher 2007) descriptions. Surveys, basins, oil and gas fields, reservoirs, structures and production data instances are mapped through a warehouse modelling as given in an integrated framework in Figure 3.38. Several data views are extracted from the warehoused exploration and production data, for mining and interpreting them in a knowledge domain. All data instances, visualized as data views, are represented into certain groups or categories, narrating the clusters. The basic idea of this approach is to use data warehouse for investigating and computing finer clusters of petroleum data and accessing desired clusters (e.g. clustering of similar reservoir characteristics from several fields, cluster of similar structural properties from several fields and basins, cluster of similar surveys attributed to drillable exploratory well) for interpretation purposes. The motivation is to improve precision and/or recall as well as reduce the overall amount of time spent searching for data clustering. Supporting technologies include agents for searching the data and group by similar property data clusters, data delivery agents using meta-data languages (e.g., XML and HTML), and other knowledge representation tools.

Data mining views

Data mining of gulf-basin data deals with large databases that impose on clustering analysis severe computational requirements. Multidimensional Scaling (MS) (Hair et al. 1984) is intended to identify and examine underlying dimensions from a series of similarities provided by the correlation, patterns and trend analysis plots of the data taken out from databases in the form of views. MS provides an idea what, how many and how the dimensions are used in the analysis of situations. The main aim of MS approach in the present study is to examine and understand the similarities and differences in similarities between plots of data, spatial representation of data that clarify relationships, determine the number of dimensions to represent the data and finally interpret the correlations, trends and patterns so that dimensions that have undergone clustering, have been interpreted. MS also identifies and also ranks relationships described in the data more clearly. From the parabolic, power, linear and polynomial equations constructed among different attribute dimensions of petroleum exploration and production data, similarity is inferred among properties. There are many similarities in between these equations and thus dimensions in the data. Similarly, for petroleum exploration data, several statistical fits are constructed among different dimensions providing several coherencies and similarities in the characteristics of the equations as given in the forthcoming Chapter 5. In view of these coherencies and similarities in these data, several inherent relationships are interpreted.

Classification by clusters in bubble plots

The grouping is accomplished (Pujari 2002) by finding similarities between data according to characteristics found in the actual data. The groups are called clusters. In other words, clustering is a class of modelling used to place items into groups, having similar characteristics of their attributes. In the present study, bubbles plotted with similar sizes have been clustered into groups as shown in Figure 4.116. Bubble plot displays two variables on a scatter-type plot. In a bubble plot, the diameter of each bubble can vary in size, providing a way to represent an additional dimension of data.

For example, consider a traditional scatter plot that shows the number of surveys conducted in the Canning basin over a period of time. Using a bubble plot can also display a third dimension of data that shows the average petroleum production over the same time span. This bubble plot has been used to study the following features of the data mining. The bubble line and fill color are set in the bubble plot properties of the grapher program. On the bubble plot, min radius and max radius fields are set the range of the bubble's radius. The smallest value in the column value is displayed as a bubble with the min radius value. The largest value in the column value is displayed as a bubble with the max radius value.

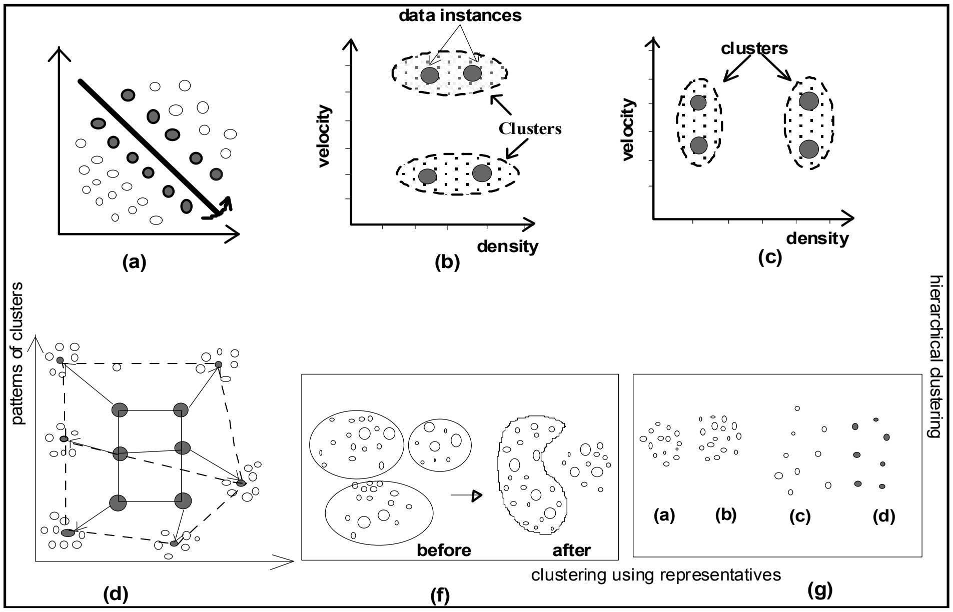


Figure 4.116: Oil and gas data mining schemes

Clustering has many applications in biology, medicine, anthropology, marketing and financial institutions (Fraley and Raftery 1998). It has ability to process images, recognizing patterns, in oil & gas exploration applications, in particular with geochemical and geo-statistics prospecting. In the present study, author makes an attempt to create data views in the form of clusters using the oil and gas data.

The author has examined these plots for classification of group of data from different clusters. This sort of cluster analysis identifies and classifies variables so that each variable is very similar (bubble size) to others in its cluster with respect to the predetermined selection criteria. This is presented in a graphic form as bubble plots for exploration data. Partitioning of individual or group of similar clusters or similarity in variables or objects, interpretation of these individually grouped variables and profiling of these clusters to detail the characteristics, can be performed in the process of cluster analysis. One has to carefully observe and analyze the size of the bubble and the direction, their sizes are varying to.

A computing logic or algorithm is necessitated to instruct the program to control the petroleum datasets to identify the similar property attribute data. Clustering algorithms, which attempt to find natural groups of components (or data) based on some similarity uncover and examine the centroid of a group of datasets. To determine clusters, most algorithms evaluate the distance between a point and the cluster centroids. The output from a cluster algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster. The centroid of a cluster is a point whose attribute values are the mean of the attribute values of all the points in the clusters. The common metric computed between two points will assess the similarity among the attributes of petroleum business data. The most commonly used distance measure (Gupta 1990) is the Euclidean metric which defines the distance between two points s = ($s_1, s_2, s_3, \ldots$) and t = ($t_1, t_2, t_3, \ldots$) as:

$D = \sqrt{}\ \Sigma\ (s_i - t_i)^2$ i = 1 to k; k = number of points

Interpretation of clusters (Fraley and Raftery1998**) for building knowledge of a petroleum system** (research objective, RO 5, as explained in section, 1.3.1 of Chapter 1)

The results of clustering must be appropriately interpreted in the wider context of the application. Clustering of different petroleum data attributes may be considered as an initial data exploration tool before the design of further decision-tree or classification system or constructing an associativity rule, which entertain fixed number of classes. In all these cases, the clusters must suitably be transformed or interpreted. The most common representation of a clustering is by centroids of each cluster, or a prototype data instance nearest to the centroid. This is effective for compact and isotropic clusters, but not for elongated or anisotropic clusters. In certain applications, the extremities of a cluster are used to form the conjunctive expressions in rule sets.

Clustering is division of petroleum data into smaller groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses finer details, but achieves simplification. Hidden patterns from clusters are interpreted in terms of variations in distances or concepts of oil-play data. Clustering in the present context, is said to be unsupervised learning of hidden petroleum system data. More details on results and discussions of clusters in oil and gas domain are provided in Chapter 4.

Comparison among modelling methodologies in oil and gas domain

Multidimensional modelling methods appear to have more compatibility with data warehousing, data mining, visualization including interpretation, compared with conventional ER and EER mapping methods. Keeping in view spatial-temporal data types and their heterogeneities, fine-grain data schemas are necessary for accommodating multidimensional structures within a warehouse environment. In addition, data types and their characteristics are significant in choosing a particular type of modelling approach. To substantiate these merits and comparisons, more reasoning is provided in the Chapter 5.

4.9 Summary

The author has described ontologies in different knowledge domains and applications of oil and gas business scenario. For this purpose entities, objects and dimensions are identified in an upstream project and described in modelling ontologies and their accommodation in a warehouse development. Data integration and metadata design aspects are described, providing schema integration, integration process, and integration of relationships among entities, objects and dimensions. The author has described big-data systems, conventional, unconventional reservoir systems and systems that describe shale gas and their ontologies in each domain.

Chapter 5: Results and Discussions (RQ1 – RQ8 and RO1 – RO8 Focus)

5.0 Introduction

The author has described the computation of data views and analysis of their views for multiple dimensions and in different knowledge domains. The overall aim of the chapter is interpretation of various data views extracted from warehoused metadata structures, described in Chapter 4. As per RQ1-RQ8 and RO1-RO8, laid in Sections 1.3.1 and 1.3.2 of Chapter 1 and the guidelines of DS, the author implements data models and the warehoused metadata in the form of domain knowledge extraction and interpretation of new knowledge. This chapter reiterates that the domain knowledge interpreted can further provide leads to the evaluation of the research objectives to be achieved.

The methodologies and applications mentioned in Chapters 3 and 4 are the artifacts of the current research. The outcome of the current research embodies validation of the design science research guidelines and interpretation of domain knowledge from metadata structures. The results and discussions made in this chapter are new insights, based on the analysis of heterogeneous and multidimensional data sources in different application domains. In this chapter, the author analyzes models deduced from integrated workflows as described in Chapter 4, for interpretation. Oil and gas data sources possessing, longitudinal and lateral dimensions, respectively as periodical and geographic dimensions, provide valuable information for interpretation and new knowledge discovery. From the conceptual modelling to its implementation, these dimensions play significant roles in revealing new knowledge through interpretation of several data views, extracted from the warehoused metadata, generated in multiple domains.

The author analyses warehoused multidimensional exploration metadata in Section 5.1, providing metadata analysis for ER, OO and MR cases in Section 5.2. Interpretation views and analysis of domain knowledge obtained from multiple digital ecosystems (as discussed in Chapter 4) are provided in Section 5.3. Analysis of petroleum digital ecosystems and digital oil field solutions (RQ7 and RO7) are given in Section 5.4.

5.1 Analysis of Multidimensional Exploration Data Warehousing

The author accommodates the multidimensional data in an integrated warehouse environment with variety of dimensions involved in an oil and gas exploration project. To make use of the potential information available in different operational units of oil and gas exploration, these multidimensional data are stored in a warehouse as a part of every-day's work and they are selected, consolidated and aggregated. Not every single information, is necessary to deal with day to day requests that need to be stored in the data warehouse. Other data are processed or aggregated first, before it is of use for analytical or prognostic purposes. Nevertheless, all the operational data stored are accompanied by metadata which allows accessing the content or information from datasets. In the end, the oil and gas data warehouse provides the user, with pre-structured data and tools to access and explore the information content.

The aggregation of oil and gas data, especially when it involves with space-time components and continual update of such spatial-temporal data from the operational processes is an essential task of the oil and gas data warehouse. Even under changing circumstances like new spatial divisions, the data warehouse is able to provide with the spatial-temporal oil and gas data that the user needs. It is important the spatial-temporal changes are frequently updated and integrated in the data warehouse, both as means of data collection, processing and storage as well as a tool for the exploration of spatial-temporal oil and gas data trends, correlations and patterns.

5.2 Metadata Analysis

The key purpose of the metadata is to facilitate and improve the retrieval of information. The Information with which the author deals with large upstream businesses, its data sources are heterogeneous and multidimensional. The oil and gas data are complex in nature with several entities, dimensions, objects and attributes. The data integration is crucial for industry managers to make technical, financial and human resources procurement decisions. Methodologies discussed in this thesis narrate all the conceptual schemas simplifying the logical data models and facilitate implementation for oil and gas companies. Sub-schemas interpreted as views are added to the existing schemas, so that data integrated are current and allow metadata warehouses to extract user defined views more precisely. Implementation of conceptual and logical data structures for knowledge mapping has been discussed. Logical and physical data are organized for many sedimentary basins. This demonstrates the data integration procedure (as described in Figure 3.38) which is a prerequisite to explore and exploit interesting geological features attributable to the petroleum prospects from metadata of sedimentary basins, especially from the Middle Eastern Gulf regions.

Analysis of a metadata in terms of data or map views is useful representation for interpretation and knowledge discovery. In a typical exploration and production petroleum company, geologists, geophysicists, petrophysicists, reservoir engineers and production engineering professionals who get involved in a team environment utilize the metadata. The metadata are analyzed with classifications in surface and sub-surface based, seismic, well, reservoir and production data attributes represented in different views. The metadata derived (though surface-based, but interpreted as sub-surface) in the producing areas (Figures 2.1, 2.2 and 3.38) are represented in different attributes, such as *number of wells* contributed in the mapping process of a producing geological horizon. Several attribute maps are generated for making crucial technical decisions on drillable exploratory and development locations in the offshore basin. As an example, structure attributes of a geological horizon are integrated with formation tops derived from wells. Good *porosity* attributes are extracted from a geological horizon from many wells. The author takes advantage of the development of good seismic signatures of formation in the oil field producing area for better well data integration.

There is scope of discovering more petroleum oil and gas resources if the exploration and production datasets and their structures are logically connected, modeled and imaged for detailed visualization and interpretation. Thus, the knowledge built from metadata is representative from map views (Figures 5.1 and 5.2) that characterize the power of data visualization and data mining, assisting interpretation for drillable exploratory and development locations in the Arabian Gulf basins (addressing RQ8 and RO8).

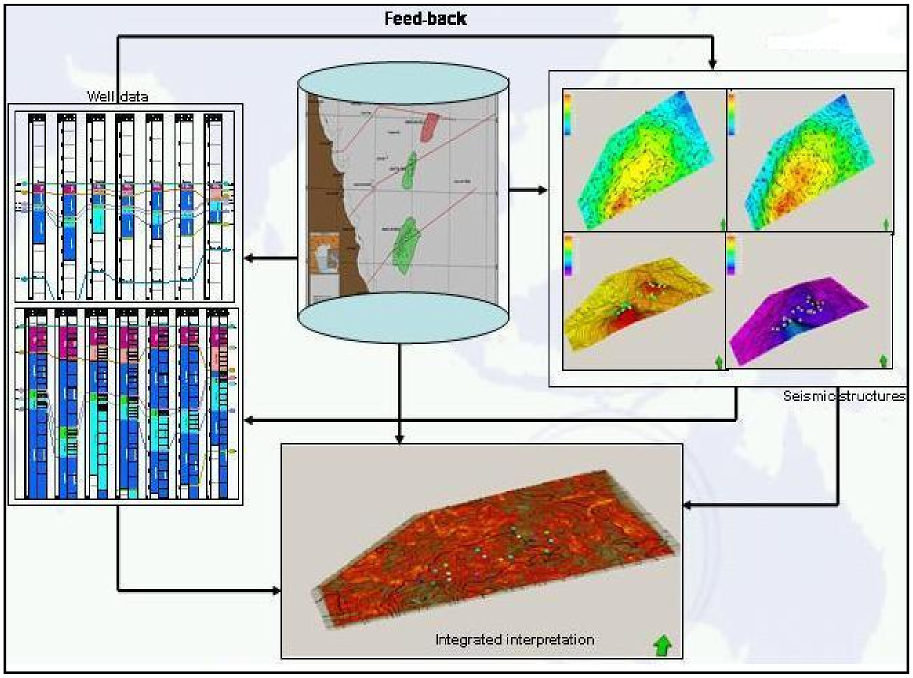


Figure 5.1: Mapping domain knowledge features for *geological* interpretation

Similar data integration and knowledge mapping is under study among several other database entities in the producing basins in the middle-east region. There is further scope of connecting onshore and offshore data entities, dimensions and their possible *areal-extents* by the data integration methodologies.

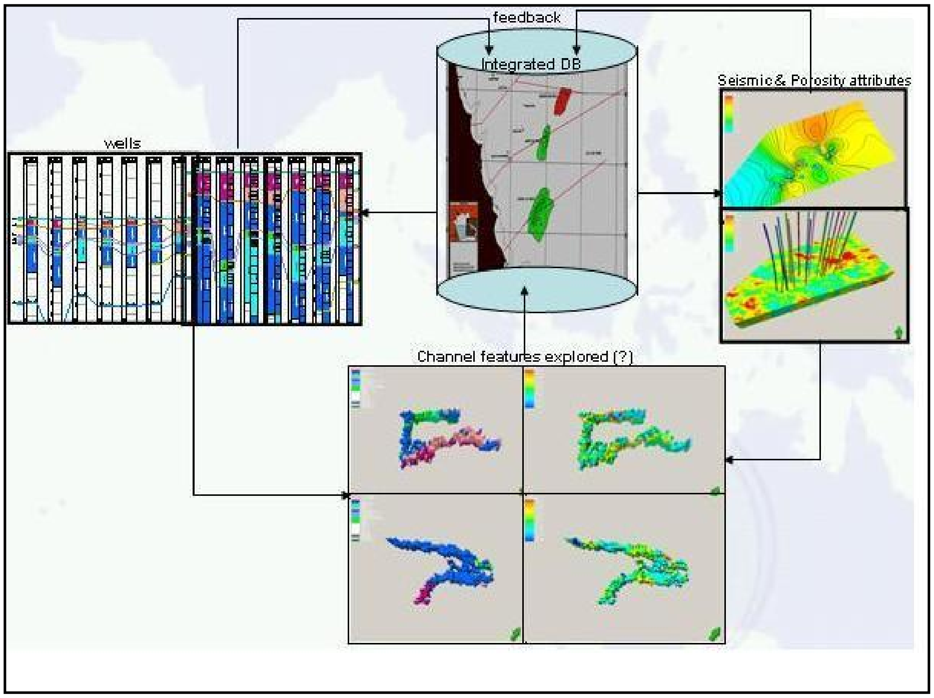


Figure 5.2: Interpreting of knowledgeable features, identified (in Figure 5.1) for drillable locations

5.2.1 ER and E^{2}R metadata analysis

A generalized structure of the integrated exploration and production oil and gas industry (E&P) is made up of several specialized structural data models. These structures are flexible and interoperable, in the sense, more entities and their relationships can be added into the existing ER models to build an updated and integrated metadata model. When new entities start emerging into the ER models, because of the change in the system or organization situation in their sizes or new oil and gas fields are explored for further exploitation, extended entity modeling (E^2/R) approach appears to be logically feasible. Logical E^2/R models can also be converted into relational, hierarchical and network models (Ozkaharan 1990) to further simplify the existing logical models and implement them in the integrated E & P petroleum industry. Every entity that is added is made linked to other entities. Every relation that is converted to another relation bears a unique name of the associated type of entity.

Any warehoused metadata that is designed has inclusion of combination of the data structures based on the several logics and business constraints. As shown in Figure 5.3, a sub-system process model is conceptualized and made logical based on sub-type entities identified from generalized entities of exploration-drilling-production-technical. Based on decision rules, oil and gas company’s data files are extracted from a centralized warehoused repository. The accessed files from database are further processed to extract business intelligence and future forecast of organizational resources. The data structuring methodologies facilitate data analysis, understand better data integration and make the models more explicit in designing and implementing logical models in a data warehousing environment. Metadata pertained to petroleum system elements in complex geological settings of different sedimentary basins of the Middle East, East and West Africa and Asia-Pacific regions are result of logically and physically organized structures accommodated in a warehouse. This contribution facilitates the data mining users for effective data mining of warehoused exploration and production metadata.

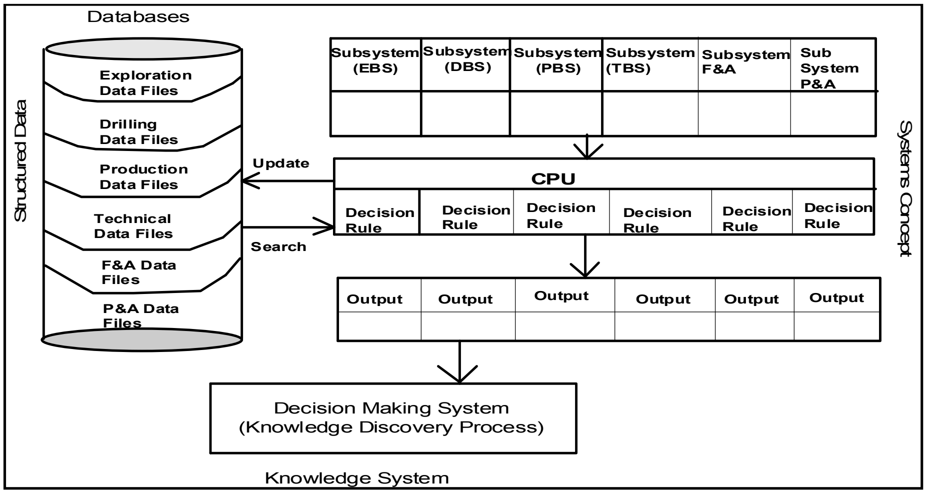


Figure 5.3: A knowledge based subsystems concept

5.2.2 Object oriented metadata analysis

The basic star schema creates a multidimensional space (often called dice), using the basic capabilities of a relational database utilities (Nimmagadda and Rudra 2004). One needs to understand a multidimensional space. A multidimensional analysis space is depicted in the Figure 5.4. A geometrical dice is an example of three dimensional space with all three dimensions of the same size. Imaging a cube with each object class dimension of three units, we get 44 = 256 cells of equal structure. The multidimensional analysis space (or a data warehouse dice) differs just in details from a geometrical space.

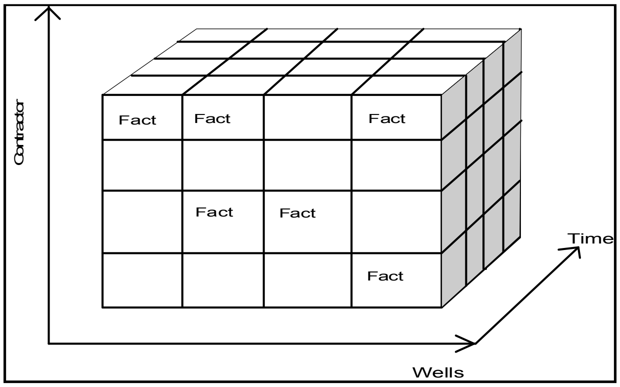


Figure 5.4**:** A dice with dimensions *wells, contractor, time*

However, the dimensions are not just limited to just three. It is not easy to handle a cube with several object class dimensions, which often result in most of the implementations, limited to six or seven object dimensions. However, one should never expect a good graphical representation of more than three object dimensions. All object dimensions are not of the same size and unit. The size differs from few units to several millions of units. The units can be period (day, quarter, month or year), contractor, wells or an exploration department, with several cuboid cells, describing fact tables (see Figure 5.4). The data cube needs much memory to store all the facts. For this reason, we design multidimensional schemas in a warehouse environment, optimizing storage capacity and preserving the flexibility of data structuring. The purpose of mapping objects of oil and gas company data is to design a warehoused metadata and keep the operational managers active and sharp in their managerial decision support to act promptly. Million dollar decisions are taken based on the information obtained from warehoused metadata with accountability and it provides an added value in return, tens of millions of dollars saved. Data objects from different warehouse marts, such as exploration, drilling and production, when integrated, furnish added values.

An online analytical processing (Nimmagadda and Dreher 2005) tool has capability to look into oil company's data object models in depth domain, map and present them in interpretable ways. Company's oil-field base data schemas, either multidimensional or object oriented models are reused among multiple petroleum fields (Figure 5.5), so that knowledge built from data integration is interpreted in terms of drillable exploratory or development location as given in a couple of OLAP models (Figures 4.53 and 4.54).

Exploration managers provide valued processed exploration information, so that critical decisions made in planning new exploratory or development wells in the frontier oil bearing sedimentary basins are accurate and precise. Figure 5.5 exhibits the interoperability of petroleum data objects among basins (Australian basins), when data objects are conceptualized using petroleum ontology (Nimmagadda and Dreher 2006).

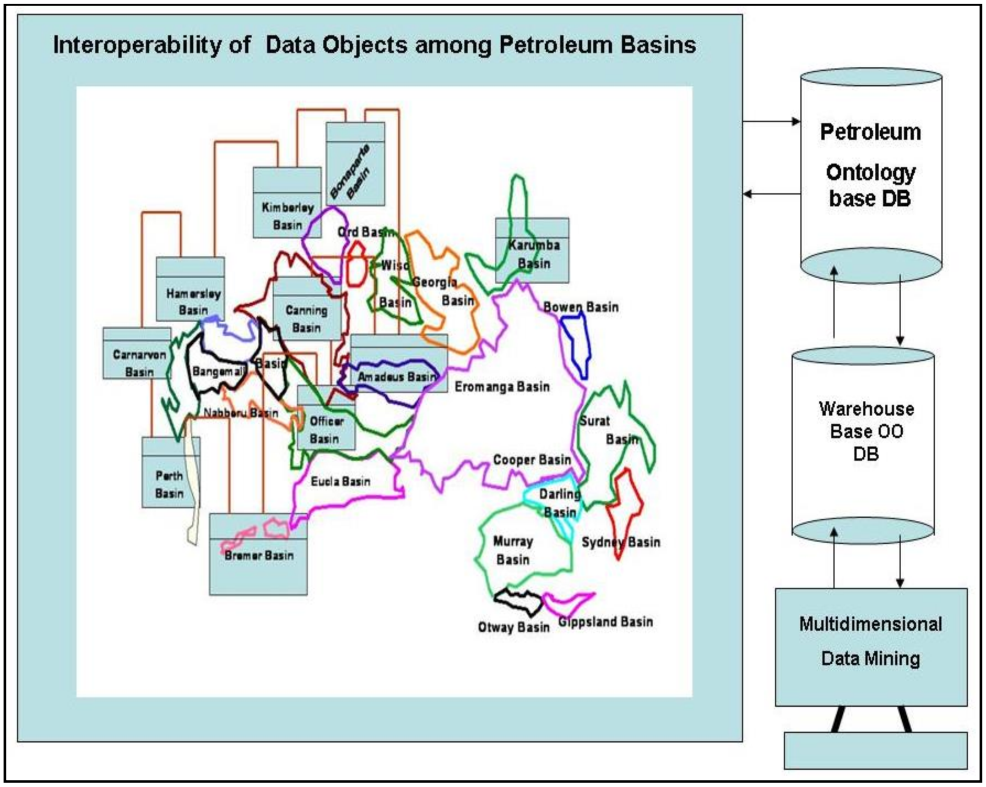


Figure 5.5: Interoperability of petroleum data object attributes among petroleum bearing sedimentary basins of Australia

All the survey information processed by OLAP is presented by different combination of aggregated data views as shown in Figure 5.6. Well data facts stored in the oil and gas data warehouse or data marts, are processed by OLAP procedure (Moody and Kortink 2003) and presented in aggregated data views convenient to interpret and extract knowledge from the past historical oil and gas business data.

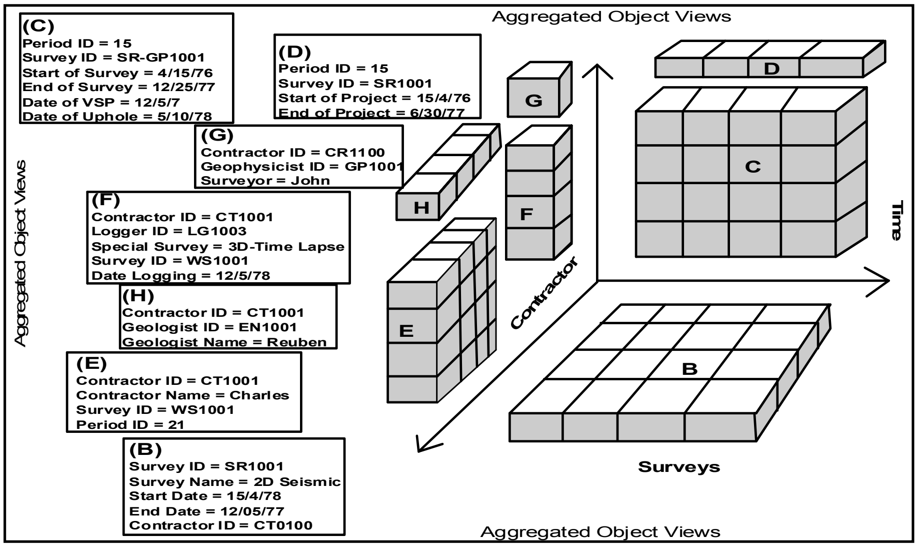


Figure 5.6: Aggregated computed views of “survey” object

The multidimensional object data views for super class object “wells” are shown in Figures 5.7a and 5.7b. All the available data and information are now integrated and available in a centrally located enterprise data warehouse (EDW), making it easy for all managers to make timely decisions. Different data warehouse architectures are tried creating several data marts that can map and process individual operational units’ object classes and make them available to exploration managers. Similar data marts may be initiated for drilling, production, marketing, human resources and other support engineering class objects. The Big-data systems too can ease the complexity of the data organization.

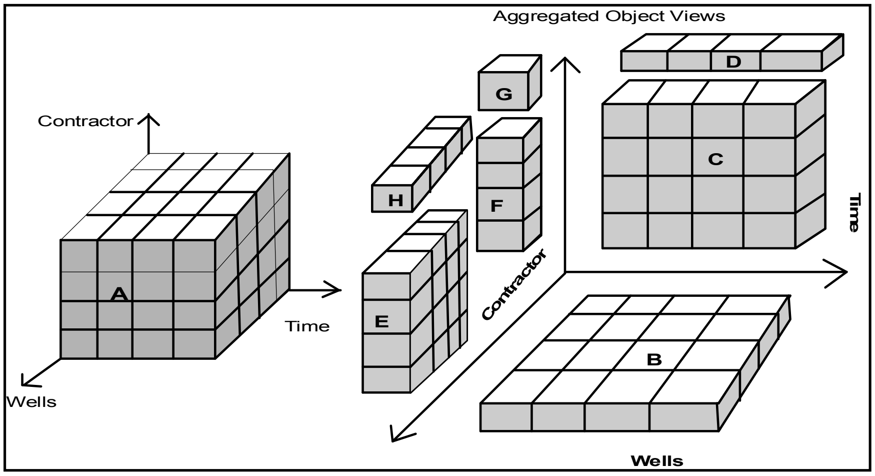
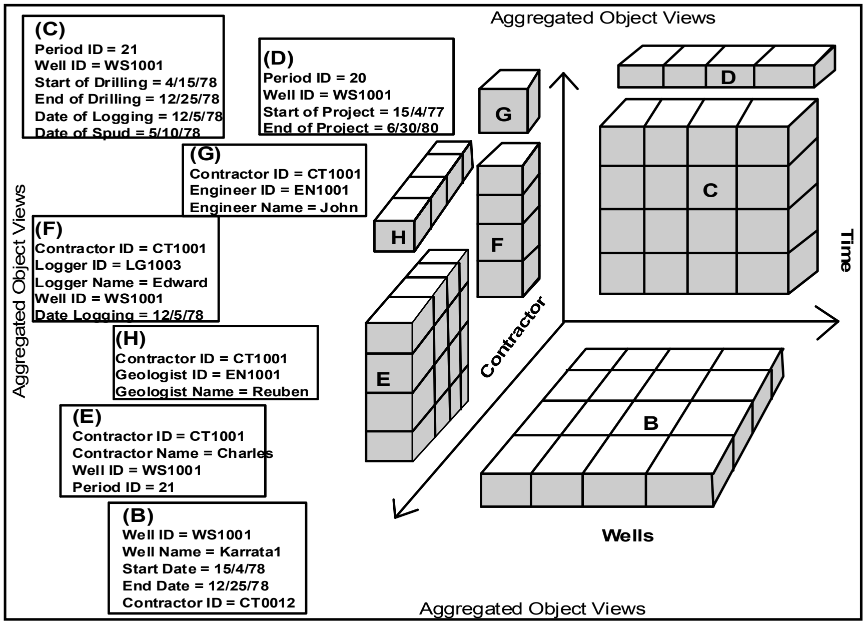


Figure 5.7a: Example of 3D data cube showing three dimensions *wells*, *contractor* and *time*



(C)
Period ID = 21
Well ID = WS1001
Start of Drilling = 4/15/78
End of Drilling = 12/25/78
Date of Logging = 12/5/78
Date of Spud = 5/10/78

(D)
Period ID = 20
Well ID = WS1001
Start of Project = 15/4/77
End of Project = 6/30/80

(G)
Contractor ID = CT1001
Engineer ID = EN1001
Engineer Name = John

(F)
Contractor ID = CT1001
Logger ID = LG1003
Logger Name = Edward
Well ID = WS1001
Date Logging = 12/5/78

(H)
Contractor ID = CT1001
Geologist ID = EN1001
Geologist Name = Reuben

(E)
Contractor ID = CT1001
Contractor Name = Charles
Well ID = WS1001
Period ID = 21

(B)
Well ID = WS1001
Well Name = Karrata1
Start Date = 15/4/78
End Date = 12/25/78
Contractor ID = CT0012

Figure 5.7b: Computed aggregate views of “wells” object

5.2.3 Multidimensional metadata analysis

Information retrieval is one of the measures of multidimensional metadata and its analysis. The author implements these logically organized models successfully in an integrated petroleum company for effective data mining and visualization by generating interpretable data views and interpreting them with new knowledge. Decomposition/denormalization of more specialization entities of the generalized dimension can be effective (Nimmagadda and Dreher 2006) in structuring these data models for warehousing and data mining purposes.

5.3 Systems Analysis and Interpretation of Domain Knowledge

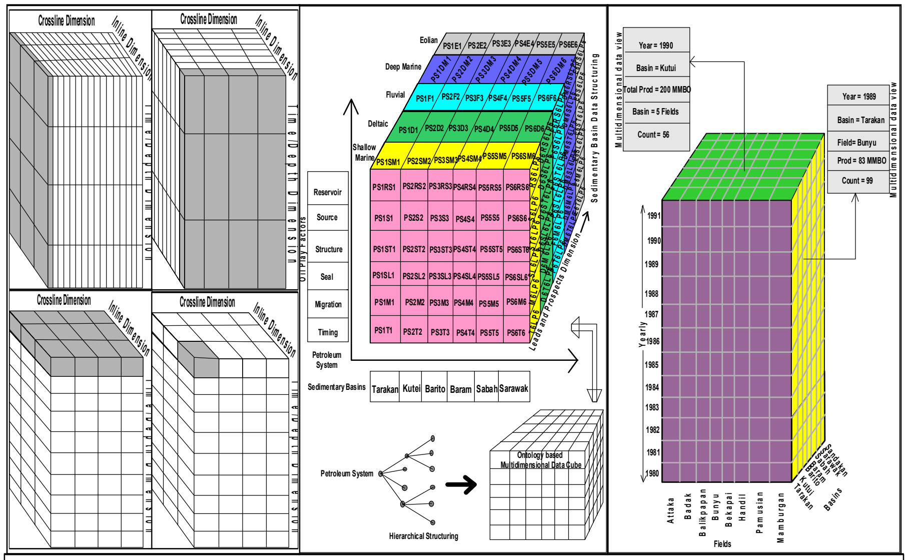
The author analyses exploration data sources, modelled in an integrated framework (that refer RO5 and RO6) in cases such as big data systems, systems in which turbulent business conditions exist, conventional and unconventional reservoir digital ecosystems, as discussed in the following sections. The results presented in section 5.3 are scientifically significant (309-313), especially petroleum exploration and field development point of view. Figures 5.8 to 5.11 are real, obtained as plot and map views from the warehoused metadata. These are obtained from an integrated warehouse environment. This methodology is described in details in Chapters 3 and 4. Fine-grained multidimensional real data are used for modelling, domain ontologies and generating metadata. The author reiterates the sequence of events in the methodology:

- Domain modelling (3.6.1.1)
- Data modelling (used ontologies) (3.6.1.2)
- Schemas schemes (3.6.1.4)
- Data Warehouse (for integration) (3.6.2)
- Data mining (3.6.1.5)
- Data visualization (3.6.1.6)
- Data interpretation and discovery of new knowledge (3.6.1.7 and 3.6.1.8)

This sequence is accommodated within an integrated framework, which also includes data acquisition and other ETL utilities. Author has mentioned various references and the data sources in the Appendix-2. Author does not disclose the raw-data sources that have confidentiality issues.

5.3.1 Big data systems analysis

The systems dealing with big-data take advantage of the evaluation of petroleum digital ecosystems (PDE), complex data models and implementation of their interpretative analysis of domain knowledge in business environments. “Volume, variability, velocity, veracity, visualization and value” are features through which petroleum data sources are better managed. One of the primary goals of examining big data (Cleary et al. 2012) systems is to discover repeatable technical and business-data patterns and their intelligence. It is generally an accepted notion that unstructured data, most of it located in text files, accounts for at least 80% of an organization’s data. If left unmanaged, the sheer volume of unstructured data that’s generated each year within an enterprise or company can be costly in terms of storage. Big data analytics (Cleary et al. 2012) is often associated with cloud computing because the analysis of large data sets in real-time requires a framework like *MapReduce* to distribute the work among tens, hundreds or even thousands of computers. In big-data analytics, the author examines large amounts of data of a variety of types of heterogeneous and multidimensional nature to uncover hidden patterns, unknown correlations and other useful information. Such information provides competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue.



1. SELECT operator is defined as: *OS = σ (NS, ES, RS) where NS = Nodes (condition = true) and ES = Edges (N ∈ NS).*

2***. INTERSECTION****: OI (1, 2) = O1 $\cap_{SR}$ O2 = (NI, EI, RI), where: NI = Nodes (SR (O1, O2)), EI = Edges (E1, NI ∩ N1) + Edges (E2, NI ∩ N2) + Edges (SR (O1, O2)), and RI = Relationships (O1, NI ∩ N1) + Relationships (O2, NI ∩ N2) + SR (O1, O2) – Edges (SR (O1, O2)).*

3. ***UNION****: OI (1, 2) = O1 $\cup_{SR}$ O2 = (NU, EU, RU), where, NU = N1 ∪ N2 ∪ NI (1, 2), EU = E1 ∪ E2 ∪ EI (1, 2), and RU = R1 ∪ R2 ∪ RI (1, 2), where, OI (1, 2) = O1 $\cap_{SR}$ O2 = (NI (1, 2), EI (1, 2), RI (1, 2))*

DIFFERENCE*: O1 – (O1 $\cap_{SR}$ O2); O1, O2 are two different domain ontologies; SR: Semantic Relationships; N: set of Nodes; E: set of edges;*

Figure 5.8: Multidimensional data views from a seismic cube from ontologically described metadata for interpretation and knowledge discovery (mining algebra used for extracting data views)

As shown in Figure 5.8, several data cube views extracted from mining algebra are representative to interpretation and new knowledge discovery. The crucial goal of big data analytics is to help operating companies make better technical and business decisions by enabling data scientists and other users to analyze huge volumes of technical and commercial data as well as other data sources that are left untapped by conventional business intelligence (BI) programs. The other data sources may include Web server logs and Internet clickstream data, social media activity reports, mobile-phone call detail records and information captured by sensors. Some people exclusively associate with big-data and big data analytics (Cleary et al. 2012) with unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid forms of big-data.

The author adopts the big-data analytics (Cleary et al. 2012) with software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data mining. But the unstructured data sources used for big data analytics may not fit in traditional data warehouses. Potential pitfalls that can rise with organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals, plus challenges in integrating *Hadoop* systems and data warehouses, although vendors are starting to offer software connectors between those technologies and integrated platforms. The author attempts to use various integrated frameworks and as designed and developed in Shastri and Dreher (2011) and Nimmagadda and Dreher (2012), for visualization and mining of the G & G data views for interpretation in the upstream oil and gas.

A G- & G- data mining and fusion

As narrated in an integrated framework (Nimmagadda and Dreher 2012), the ontology based multidimensional modeling is performed to generate a metadata. As described in Figures 3.36 and 3.37, the author collaborates integrated frameworks (Nimmagadda and Dreher 2012) to facilitate the data visualization, mining and interpretation processes. Data mining in essence, explores data correlations, trends and patterns among petroleum heterogeneous metadata. Several algorithms are available in the literature (Pujari 2002); one of the popular ones is representation of multidimensional data cube and its *slicing* and *dicing* (Pujari 2002). Data fusion is the process of integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and interpretable representation. Fusion of the data sources (multiple dimensions) yields a classifier superior to any other classifiers based on dimensions and their fusion representations. Data fusion processes are often categorized as low, intermediate or high, depending on the processing stage at which fusion takes place. Low-level data fusion combines several sources of raw data to produce a new arrangement of raw data. The expectation is that fused data are more informative and synthetic than the original inputs. In a much more complicated application, petroleum exploration researchers use "data fusion" to review and display integrated products of exploration and production data with other associated systems (Wight et al. 1992). Other ecosystems include data acquired from a marine system such as bathymetric; in an environment system, such as meteorological, sea surface temperature (SST) and bore-hole temperature are examined to understand petroleum system integration and its ecosystem behavior in reaction to other associated ecosystems. Through the use of data fusion, all data and attributes are brought together into a single view in which, a more complete picture of the system is created and interpreted for the case of Southeast Asian larger sedimentary basin setting (examples of data fusion as narrated in Figures 5.9 – 5.11). This enables geoscientists to gain new insights on interactions between different petroleum ecosystems. The data fusion is significant for reservoir-trap presentation, which provided an increased understanding of petroleum ecosystem visualization process.

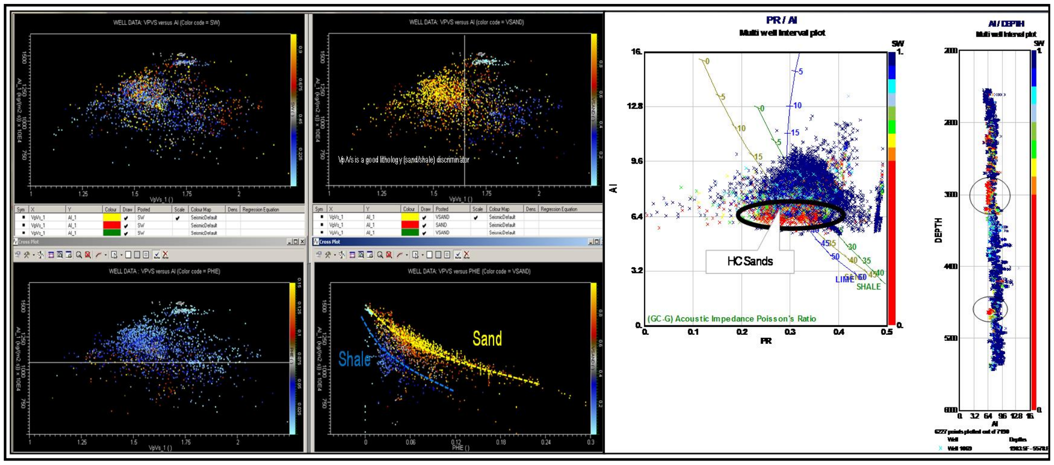


Figure 5.9: Multidimensional data views for extracting knowledge of a *clastic reservoir* (sandstone) from integrated metadata

In the petroleum systems analysis and knowledge domains, the data fusion is often synonymous to representation of integrated interpreted data or map-views of multidimensional data views. In these applications, there is often a need to combine systems' elements and their datasets into a unified (fused) dataset, in which all the data instances from domains, (such as *seismic*, *drilled-well*, *reservoir*, *drilling* and *production)* are interlinked and integrated into a single conceptual domain, which could be *chains*. The fused dataset is different from a combined superset. The data points in the fused data set can contain attribute strengths and magnitudes (computed or derived from hypotheses). In fact, fused metadata may not represent the contextual or conceptual views as that described in the original data. Data instances built with conceptual *chains* are preserved in the original fact tables. More commonly, data warehousing supports the data fusion concept. In this context, the data warehousing consists of collecting functions providing access to, and management of, data fusion databases, including data retrieval, storage, archiving, compression, relational queries, and protection.

Knowledge based data fusion: Interpretation of the fused data for geology and geophysics (G & G) interpretations, with valid calibrations and assumptions, is used to obtain data trends and patterns, ensuring good agreement with the original data. These calibrations are in the form of two-dimensional graphic or map views drawn for Southeast Asian sedimentary basins (Figures 5.9 – 5.11). As articulated an integrated framework and described in Nimmagadda and Dreher (2012), the author extracts data views through OLAP as described in Pujari (2002) and Nimmagadda and Dreher (2012) from integrated metadata structures. In particular, these are extended-form of knowledge-based or expert systems developed to interpret the results of processing, analyzing issues such as contexts in which the data, the relationships among multiple dimensions, hierarchical groupings of targets or objects and predictions of future actions of targets perceived. Such reasoning is normally performed by analytical solutions, but it is approximated by hypotheses, wherever necessary.

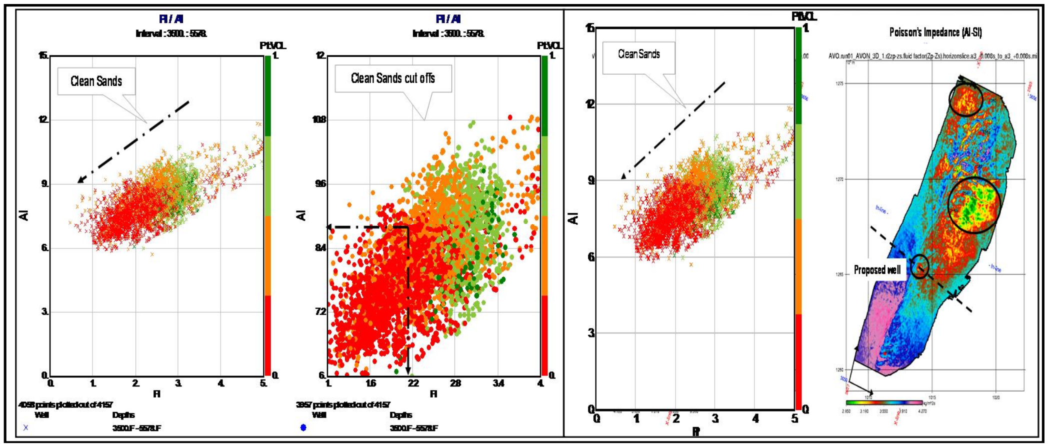


Figure 5.10: Geology and geophysics fusion data mining and visualization

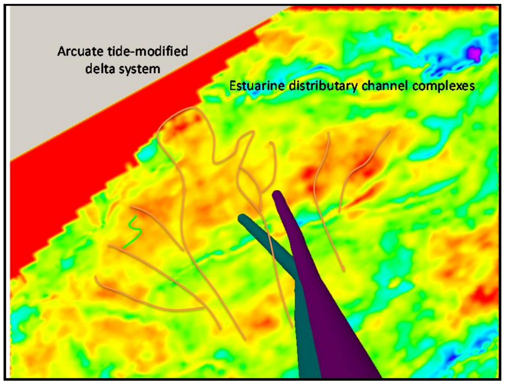


Figure 5.11: Knowledge discovery from one of the attributed metadata models

The big-data systems too, at times come to the rescue of oil and gas business projects in managing their turbulent business environments. The author analyses systems that deal with managing turbulent resources in the following sections.

5.3.2 Turbulent business systems analysis

As a part of research objectives, RO-5 and RO-6, as cited in Section 1.3.1 of Chapter 1, several queries are interpreted in the form of data views for implementation of metadata. The author uses many queries and sub-queries, generated from the grapher and surfer solutions, intelligent data analysers. Statistical and graphic techniques are used for visualization of the data. Several data views drawn from 3D cubes (Figure 5.12) are visually represented and interpreted in terms of predictions and forecasts.

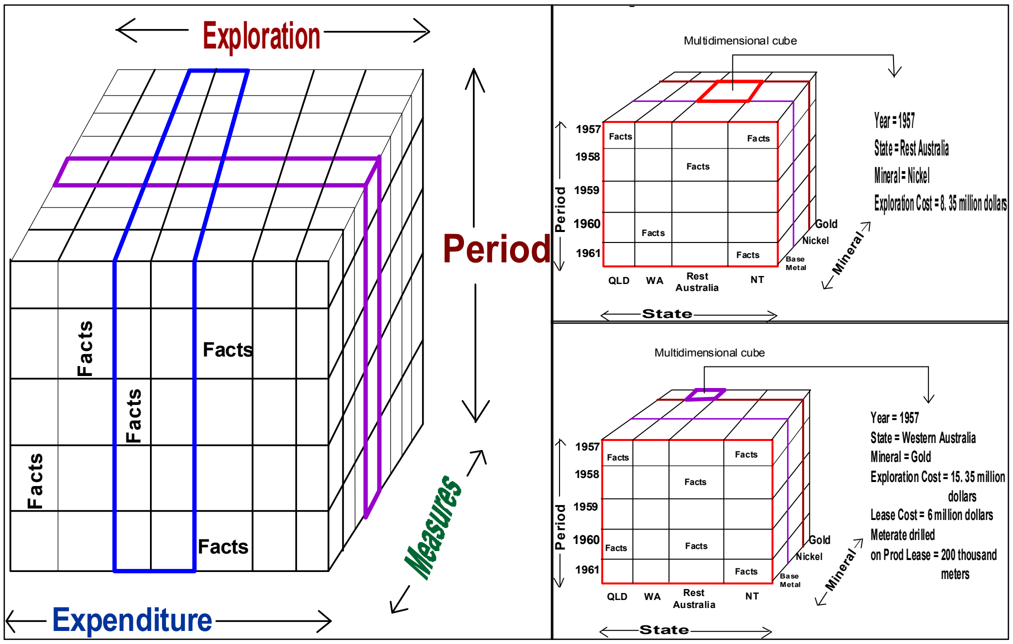


Figure 5.12: Data views from multidimensional data cubes

It is observed visually from the characteristics of time series data plots, the construction expenditures have shown a tendency to increase in a curvilinear over a period of 30-50 years of time. In the present studies, data have been considered from 1953 to 2000 years. This overall long-term tendency or impression is known as trend. For particular periods of time, the observed values are dipping below the trend curve. They represent the peaks of their respective business cycles. Any observed data that do not follow the smooth fitted trend curve modified by the aforementioned cyclical movements are indicative of the irregular or random factors of influence. When data are recorded monthly rather than annually, an additional factor has an effect on the time series data. It could be due to seasonal component. At certain periods of time, the trends appear to be irregular and or random and seasonal at other periods of time.

As described earlier, in addition to various time series components, Fayyad et al. (1996) discuss several other patterns of time series stock market data and detect temporal variations from which author discovers stock market knowledge. They use dynamic programming approach for interpreting time series patterns. The author observes similar or dissimilar shaped patterns in the historical oil and gas data, which have been interpreted for forecast new knowledge. Rounding top oil and gas data reversal indicates more demand of oil and gas at the period indicated “1”, where economic boom or growth may be interpreted. In the rounding bottom reversal case, probably, at a period indicated at “1”, there may be economic recession. In case of panic reversals, high trading of resources may be observed. There are two top reversals, which may be cyclic, indicating inflation pressures. All these patterns measure expansion and contraction of periodic dimension as envisaged in the oil and gas database of oil and gas industry, which forecast future economic growth.

Forecasting vs. decision making

For decision-making, *exploration* data and information analysis, personal judgments, evaluating alternative actions in terms of probable exploration costs and pay offs are used. Impacts on entrepreneurship, while making company’s operational decisions have been addressed. Forecasting is used to simply predict future – it is not a decision making tool, but a key input to a decision model, therefore, forecast is a guiding tool. Data refer to any number of facts that may be available in the oil and gas industry. Information refers to that portion of data that is relevant to what happens in the future. One of the most difficult tasks in forecasting is separating informational content of data and conversely identifying these data, which has particular informational content in the mining and or petroleum industry.

Forecasting Techniques

There is wide diversity in presenting forecasting techniques; all have three basic elements in common: (1) which deal with situations in the future, that is, every forecast must be made for a specific point of time, referred to as the time frame of the forecast. (2) all forecasts deal with some level of uncertainty; therefore it is necessary to make assumptions, judgments, or hypothesis about relevant conditions and interactions. Some error in forecasting must be expected. (3) all forecasts must to some extent, rely on information that is contained in historical data. The forecasting methods discussed in the following sections also refer to various statistical techniques (as classic) data mining methodologies. Data that have been accessed in the form of data views from warehouse have been analyzed and generated statistical models with varied attributes of the oil and gas data. Time series analysis is used for quantitative data interpretation.

The classification of the forecasting methods is discussed in Figure 5.13.

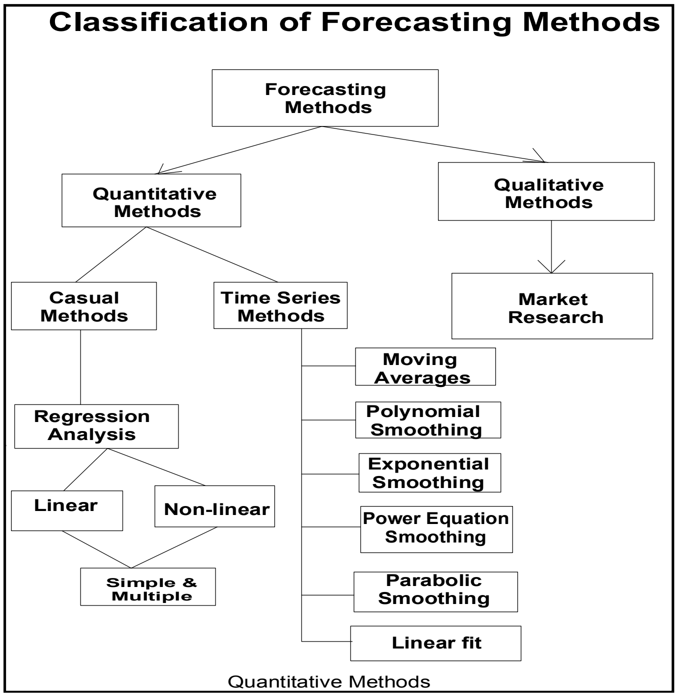


Figure 5.13: Classification of forecasting methods (Gupta 1990)

Irrespective of any forecasting method used, the development of a particular forecast technique requires to the completion of the following steps:

1. Determine the objective of forecast – what is it that we are trying to predict and for what purpose.
2. Determine the variables to be forecast – such as *exploration* costs, *production* performance, other oil and gas industry's performance indicators
3. Determine the time horizon (or period dimension) for the forecast – weekly, monthly, quarterly, half-yearly, annually
4. Decide upon which forecasting method best suits – such as moving average, regression
5. Collect the required data – past data or records
6. Make the forecast
7. Implement the results and review the forecasting process through an effective feedback system.

Time series forecasting

In the time series modelling, variables that do the predictions are analyzed with respect to different historical periods and data patterns are obtained for analysis. Such models assume the criteria that the time sequential patterns (Graham and Desmond 1992) that have occurred in the past also occur in the future. As such, historical quantitative data provide the basis of the time series analysis. For example, exploration costs or production data acquired are over a period of 70 years. These historical time series data interpret patterns and trends. These patterns are used to assist in developing forecasts as basis for future decision-making in the oil and gas industry. Time series analysis uses guidelines that are set for the future, but are designed around events that have taken place in the past. More observations or data instances can make the data mining procedures more effective and further realistic forecasts.

For time series analysis and forecasting methods, the following criteria (Graham and Desmond 1992) are taken into account:

1. Identification of the underlying trend line.
2. Measurement of past data patterns and the assumption that these patterns will be repeated in future time periods.
3. Forecast of the future trends

Four main components of a time series

The scrutiny of historical set of quantitative data series identifies 4 key movements that are appropriate to the oil and gas business data:

1. Secular trend
2. Cyclical movement
3. Seasonal movement
4. Irregular movement

Secular trend

A secular trend identifies the underlying trend (direction) of the data – increasing, decreasing or remaining constant. It is a long-term direction of the data, usually described by the "line of best fit". As an example, the increase in petroleum production is linearly proportional to its consumption. Increase in *exploration* costs or dollar fluctuations and market values of essential commodities will affect the discovery index of mineral and or petroleum deposits.

Cyclical movement

This reflects the level of business activity and economic movement over time by fluctuating patterns. These variations measure periods of expansion and contraction in industry and the economy, in general. Their regularity and intensity are not predictable, however certain economic indicators contribute to their existence – level of investment, confidence, and confidence in the economy, GNP, trade indexes and government policy. Cyclical movements in the economy – for example, inflationary pressures affect the time series trends. Any regular pattern of sequences of observations above and below the trend line, which lasts more than one year is a result of the cyclical component of the time series. For meeting demand and supply of oil and gas, economic measures taken in the oil and gas industries that boost the mineral and petroleum production, are interpreted to have associated with the cyclical movements of the periodic historical data.

Seasonal movement

Seasonal movement refers to regular periodic fluctuations that occur in each time period – usually yearly, monthly, and daily variations in the oil and gas data. Some examples are special periods during discoveries of petroleum oil and gas, monthly petroleum loads, and production decline and oil-field sickness. At times, the seasonal variations can also be documented in the quarterly exploration costs data. Seasonal variations greatly impact on the outcomes of recorded data and often belie the underlying trend. Businesses need to identify the following seasonal impacts:

1. So that a measurable (index) can be used to adjust the expected outcome.
2. In order to recognize the direction of the underlying trend

Irregular movements

These patterns refer to the uncontrollable, random variations that impact greatly on the level of oil and gas business activity. Some examples are extreme weather patterns (flood, fire, cyclone), extreme business variation (stock market crash, drop in Aus $$), political climate (sudden elections, wars, death of a leader) and industry changes (pilot strikes, waterside strikes). The resulting patterns will exert a great pressure on the predicted underlying trends and such eventualities must be accounted for, when planning for the future. All these data patterns, described so far, have been identified and well documented in the oil and gas data. Interpretation of these patterns is carried out for business intelligence analysis in the up-coming sections. The forecasting methodologies are briefly given in the next section.

Forecasting – general methodology

In general, a forecasting focuses on the decomposition of historical data into each of the above components, estimating each pattern separately, and then combining the projected impact of each component in the future, to produce the final forecast. For instance, the determination of whether a drop in sales is due to seasonal, random, or trend variations (or how much can be attributed to each) can be vitally important to any level of management in evaluating current policies and indicating the corrective action required. The methodologies considered in the present data computations are more classic and derived from basic statistical techniques. Before interpreting the quantitative models, a brief discussion of computational methods is given in the next section.

The computational considerations

The author considers moving average, multivariate regression, polynomial regression, construction of exponential, power and linear equations using the actual oil and gas data. The author presents them in the following sections.

Forecasting using smoothing methods

The smoothing techniques (Gupta 1990) are appropriate for forecasting purposes in those situations where the time series is fairly stable, in which there is no significant trend, cyclical or seasonal effects. In these situations, the objective of the forecasting method is to “smooth” out the irregular component of the time series through some type of averaging process. The methods to be covered here include:

I. Moving Average
II. Weighted Moving Average
III. Exponential Smoothing

Moving average

The moving averages method consists of computing an average of the most recent n data values in the time series. Average is then used as the forecast for the next period. Mathematically, the moving average calculation is made from:

Moving Average = Σ (most recent n data values)/ N

Weighted moving average

The moving average method provides equal weights to actual data observed in each period. This is one of the reasons for its slow reaction to variations in the most recent observations in the time series of oil and gas data. A modified version, using the weighted moving averages is used to assign a greater weight to the more current data.

For example weightings may be applied as follows:

3 to the most recent observations t_{-1}

2 to the second most recent observations t_{-2}

1 to the third most recent observations t_{-3}

[the sum of the weights is equal to 6 i.e. (3+2+1)], and the forecast may be calculated as follows:

Weighted Moving Average = $(\Sigma\ (3t_{-1} + 2t_{-2} + 1t_{-3}))/6$

A major problem, in this case is the determination of the weightings to be assigned. If greater weight is assigned to most recent observations, the weighted average may over-react to an irregular movement. However, if the weight assigned to the most recent observations is not too much greater than that is assigned to earlier observations, the meaning of weighted average may be lost.

Exponential smoothing

An exponential smoothing forecast method attempts to predict the time series in the next period based on the moving average of the current method. This method also weights on most recent observations, more heavily than older data. Consequently, the most recent changes are strongly reflected in the forecast. The exponential smoothing uses a single weighting factor called *alpha* symbolized as α. The exponential smoothing formula is as follows:

$E_t = \alpha.\ A_{t-1} + (1- \alpha.)\ E_{t-1}$

or

$E_t = E_{t-1} + \alpha.\ (A_{t-1} – E_{t-1})$

Both formulae provide the same result. Where,

E_t = forecast of the time series

A_{t-1} = actual or observed time series value for the most recent period t-1

E_{t-1} = forecast of the time series for the most recent period t-1 (old forecast)

α. = smoothing factor which has a positive value between 0 and 1

1- α. = damping factor (if α. = 0.3, then damping factor = 0.7)

When starting with the exponential smoothing calculations, the first actual result becomes the forecast for the second period, and from then on, the exponential smoothing formula will provide the forecast for next data event. These forecasting methods are more basic, but have been used in the present computations to test the validity of these techniques. In this case, the basic and more advanced data mining techniques are combined to confirm the forecast. Each one of these mining techniques has been described in the following sections.

Regression analysis

The statistical regression (Graham and Desmond 1992) is a supervised learning technique that generalizes a set of numeric data by creating a mathematical equation relating one or more input attributes to a single numeric output attribute. The author uses regression analysis for the purpose of prediction and in the present study, author develops a statistical model through regression analysis and to predict the values of a dependent or response variable based upon the values of at least one independent variable. Regression analysis or curve fitting is a procedure (Gupta 1990) for estimation of average of a variable (Y) corresponding to a given value of X. This is called regression of Y on X. If the average value of X corresponding to a given value of Y is estimated then it is known of X on Y. Depending upon the fancy of the work, several regressions are straight lines to a given set of points. However, regardless of the type of curve, fitted, there exists a relationship between two variables, which can be defined with the help of correlation coefficient.

The correlation coefficient cannot exceed one and can be less than -1. A value of 1 denotes perfect functional relationship between Y and X, an increasing X being associated with an increasing in Y. A value of -1 indicates perfect functional relationship, though now an increasing X is associated with a decreasing Y. The graphs show different configurations of forecast at plotted points. Configurations of plotted points may or may not indicate trends or relationships in the data. But quite often such inherent trends or data relationships are measured when they actually exist. Curve fitting is the solution for a given problem, in which period; a set of points is fitted along a particular trend.

For example:

$Y = a + b X$, is a linear equation in which points (X_i, Y_i) lie on a line.

From this equation, it is significant to derive and analyze the values of a and b from the constructed line. For every unit of increase in X, there is corresponding change in Y, thus b, measuring the steepness of the line and it is termed as regression coefficient. When the value of b is positive, the line ascends from left to right. When b is negative, the line descends from left to right.

Fitting straight line by method of lease squares

This implies that finding values of the parameters a and b of straight line that fits with actual observed data. The "least squares" method assumes the best fitting line in which the sum of the squares of the vertical distances of the points (Xi, Y_i) from the line is minimal. The vertical distance E_i from the line of any point P_i with coordinates (X_i, Y_i) is

$E_i = Y_i – (a + b X_i)$

The best fitting line is that line for which the sum of squares, ΣE_i^2 is minimum.

Method of computing multi variable regression

Simplest kind of relation between two statistical variables X and Y, a least square line, is made up of a line called regression line. For a change in the value of variables X, there is corresponding variation in the variable Y. A trend or data relationship is defined by constructing equations between two variables X and Y. The strength of regression between two different attribute variables is interpreted by means of correlation coefficient and association between variables.

Computation of correlation coefficients

It is the degree of similarity, both in direction and magnitude, of variations in corresponding pairs of observations of two variables. Simple correlation implies finding out degree of association between pairs of observations. In the present study, author uses the following Karl Pearson's Method of finding correlation coefficient:

$R = \Sigma XY/N\sigma_x\sigma_y$

Alternative formula is

$R = \Sigma XY/ (SQRT (\Sigma X^{2}* \Sigma Y^{2}))$

Where X = difference between the actual period value and the average of its values, and Y = the difference between the actual cost and its average values.

When the actual means are taken in fraction, the use of the above formula becomes time consuming. The following assumed method is used for computing the correlation coefficient when the cost values are in fractions:

$R = (\Sigma dxdy – (\Sigma dx* \Sigma dy)/N)/ ((SQRT (\Sigma dx^2 – (\Sigma dy)^2/N))(SQRT(\Sigma dy^2 – (\Sigma dy)^2/N)))$

At places, probable and standard errors of Correlation Coefficients have been computed using the following formulae:

Probable error of R = 0.6745 ((1- $(R)^2$)/N)   where N = Total observations

Standard error of R = ((1-$(R)^2$)/N)

The following pairs of attribute instances are considered for computing correlation coefficients (for different mineral exploration costs):

Trends computed for mineral exploration costs

1. 1st Quarter/2nd Quarter of Coal Exploration Cost: R = 0.753
2. 1st Quarter/Yearly of Coal Exploration Cost: R = 0.848
3. 2nd Quarter/Yearly of Coal Exploration Cost: R = 0.969
4. 3rd Quarter/Yearly of Coal Exploration Cost: R = 0.96
5. 4th Quarter/Yearly of Coal Exploration Cost: R = 0.965
6. 1st Quarter/Yearly of Diamond Exploration Cost: R = 0.80
7. 2nd Quarter/Yearly of Diamond Exploration Cost: R = 0.869
8. 3rd Quarter/Yearly of Diamond Exploration Cost: R = 0.78
9. 4th Quarter/Yearly of Diamond Exploration Cost: R = 0.80
10. Nickel Actual Exploration Cost/Computed Cost: R = 0.68
11. Gold Actual Exploration Cost/Computed Cost: R = 0.90
12. Base metal Actual Exploration Cost/Computed Cost: R = 0.870
13. WA Actual Mineral Exploration Cost/Computed Cost: R = 0.94
14. QLD Actual Mineral Exploration Cost/Computed Cost: R = 0.94
15. NT Actual Mineral Exploration Cost/Computed: R = 0.850
16. Rest Australia Actual Mineral Exploration Cost/Computed Cost: R = 0.86

Estimated errors in correlation

1. Probable Error of R for Base Metal Exploration Cost/Computed Cost = 0.026
2. Standard Error of R for Base Metal Exploration Cost/Computed Cost = 0.039
3. Probable Error of R for Gold Exploration Cost/Computed Cost = 0.021
4. Standard Error of R for Gold Exploration Cost/Computed Cost = 0.031
5. Probable Error of R for Nickel Exploration Cost/Computed Cost = 0.059
6. Standard Error of R for Nickel Exploration Cost/Computed Cost = 0.088

Trends computed for petroleum exploration costs

1. Actual /Expected Offshore Petroleum Exploration Cost: R = 0.92
2. Actual Offshore Exploration Cost/Computed Cost: R = 0.91
3. Onshore Actual Petroleum Exploration/Expected Cost: R = 0.658

Basin wise petroleum production trends

1. Gippsland/Total: R = 0.22
2. Gippsland/Eromanga: R = 0.70
3. Gippsland/Carnarvon Barrow: R = 0.899
4. Gippsland/Perth: R = -0.306
5. Eromanga/Carnarvon Barrow: R = 0.410
6. Carnarvon Barrow/Perth: R = -0.360

Other petroleum industry performance indicators

1. Number of Surveys/Number of Wells Drilled: R = 0.680
2. Number of days surveyed/Number of days wells drilled: R = 0.26
3. Total Survey Line Kilometers/Total Depth Meters Drilled: R = 0.615
4. Number of Surveys/Number of Hydrocarbon Producing Wells: R = 0.470
5. Number of Surveys Conducted/Number of Structures Interpreted: R = 0.655
6. Number of Structures/Number of Hydrocarbon Producing Wells: R = 0.932
7. Number of Wells Drilled/Number of Hydrocarbon Producing Wells: R = 0.88

Essential difference between correlation and regression

The correlation analysis is used, in contrast to regression, to measure the strength of an association between quantitative variables. For example, *number of surveys* conducted may have a correlation to *number of wells drilled* (attributes) in an oil field area. The *cost of exploration* may be proportional to *number of discoveries* made (a dimension, in our multidimensional modeling analysis) and the production forecast. In a pure correlation problem, a sample of pairs of observations is chosen from a bivariate. A trend is defined by correlating these two pairs. Here the functional relationship that exists is reversible from the statistical standpoint. In a pure regression problem, there is an independent or casual variable X and a dependent variable Y. The values of X assume to be selected in advance and held fixed, and then the corresponding values of Y are monitored.

Two lines of regression

In a correlation problem, it is sometimes useful to consider two lines of regression, that of Y on X and that of X on Y. In the former case, a least squares line that is minimized, sum-up the squares of the vertical or Y distances of the points from the line that is used. In the latter case, the sum of the squares of the horizontal or X distances of the points turns the line. After having evaluated the plots between actual exploration cost values and the period, parabola type of curve has been computed. Regression analysis has been carried out to further evaluate this parabola curve.

The following equation is used for fitting the second order parabola:

$Y = a + bX + cX^2$

Calculating the values of a, b, c, is crucial for fitting the parabola curve. Since the deviations of X and Y series can be taken from their means, the normal equations are reduced to:

$\Sigma Y = na + c\Sigma X^2$

$\Sigma XY = b\Sigma X^2$

$\Sigma X^2Y = b\Sigma X^2 + c\Sigma X^4$

Using these equations, the parabolic trend values have been computed for the following actual mineral exploration cost values (for Australian situations):

1. Queensland Actual Mineral Exploration Cost
   $Y = 6590.87 + 1816.58X + 104.05X^2$
2. Western Australian Actual Mineral Exploration Cost
   $Y = 15634.35 + 4557.2X + 258.23X^2$
3. Rest of Australia Actual Mineral Exploration Cost
   $Y = 24240.14 + 2268.08X + 38.32X^2$
4. Northern Territory Actual Mineral Exploration Cost
   $Y = -1143.14 + 393.53X + 36.34X^2$

Where X = period and Y = exploration cost

Similar parabolic trend values have also been computed for the following exploration cost values of

1. Base metal Minerals
   $Y = 44607.94 + 3694.43X + 33.36X^2$ (Fig. 12)

2. Gold Mineral
   $Y = -5343.57 + 4859.74X + 394.23X^2$ (Fig. 13)

3. Nickel
   $Y = 6057.9 + 481.22X + 9.35X^2$

Where X = period; Y = Exploration Cost;

A parabola curve fitting method has also been used for evaluating the offshore petroleum exploration costs and trend values have been computed for the following:

1. Offshore petroleum other exploration cost
   $Y = 124.5 + 10.3X + 0.28X^2$

2. Other Petroleum Lease Costs
   $Y = 555.03 + 12.62X + 0.24X^2$

3. Offshore Actual Exploration Cost
   $Y = 519.77 + 17.79X + 0.106X^2$

Where Y = Exploration Cost; X = Period

Types of regressive models and computation of regression equations

The nature of the relationship can be in many forms, ranging from simple mathematical functions to extremely complicated ones. Simplest one is a linear or straight-line relationship in which for each increasing in one value, there is corresponding increasing in the other. But some relations may be curvilinear, in which with increasing in value, there is corresponding decrease in the other value. The following regressive equations are used in computing trends:

X on Y: $x – X = r\sigma_x/\sigma_y (y- Y)$

Y on X: $y – Y = r\sigma_y/\sigma_x (x- X)$

$r\sigma_x/\sigma_y = (\Sigma d_x d_y – \Sigma d_x \, \Sigma d_y) / (\Sigma d_x^2 – ((\Sigma d_x)^2/N))$

Where X = Y = average of the observed or actual values; d_x and d_y are the difference of the average and actual values. Using these formulae regression equations have been computed for the following pairs of Industry Performance indicators:

1. Number of Surveys Conducted (x)/Number of Wells Drilled (y):
   $x = 0.605y + 17.22$; $y = 0.773x + 2.11$

2. Number of Surveys (x)/Petroleum Producing Wells (y)
   $x = 0.506y + 26.99$; $y = 0.437x + 0.11$

3. Number of Wells Drilled (x)/Number of Petroleum Producing Wells (y):
   $x = 1.0675y + 12.63$; $y = 0.722x + 5.61$

4. Number of Surveys (x)/Number of Structures interpreted (y):
   $x = 0.591y + 20.06$; $y = 0.726x – 0.4$

5. Number of Structures (x)/Number of Petroleum Producing Wells (y):
   $x = 1.11y + 7.84$; $y = 0.78x – 4.06$

The author represents some of the regression models in the form of polynomial regressions and regression with exponential and power equations. The author provides a brief description of the different regressions here:

Polynomial regression:

Statistical analysis carried out in the business organizations with current and past data (also in time domain), has been well demonstrated in Aczel (1993) and Berenson and Levine (1992). Often, the relationship between dependent variable, Y and one or more of the independent X variable is not a straight-line relationship but, rather, has some curvature to fit. In our present analysis, in each of the situations shown, a straight line provides a poor fit to the data. Instead, polynomials of the order higher than 1, that is, functions of higher powers of X, such as X^2, X^3, provide much better fit to the data. Such polynomials in the X variable, or several X_i variables are still considered linear regression models. The multiple linear regression models thus cover situations of fitting data to the polynomial functions. The general form of a polynomial regression model in one variable X:

$$Y = \beta_0 + \beta_1X + \beta_2X^2 + \beta_3X^3 + \ldots + \beta_mX^m + \varepsilon$$

M is the degree of polynomial. The degree of the polynomial is the order of the model. The polynomial regression fit displays a curve based on the equation above. The polynomial degree can be set from 0 to 10. A polynomial degree of zero is the average Y value, degree one is a linear fit, degree two is a quadratic fit, degree three is cubic fit and degree four is a quadric fit. Polynomial equations have been constructed for some of the situations of the data explored from *petro-2 surveys* database, which are interpreted in details in Chapter 5.

Exponential regression:

One method more often useful in forecasting time series (Aczel 1993) is exponential regression. One of such methods is a simple exponential smoothing, a useful method for forecasting time series that have no pronounced trend or seasonality. The concept is an extension of a moving average. In the exponential smoothing, more recent values of the time series are allowed to have greater influence on the forecasts of future values than the more distant observations. Exponential smoothing is based on a weighted average of current and past series values. The largest weight is given to the present observation, less weight to the immediately preceding observation, even less weight to the observation before that, and so on. The weights decline geometrically as one goes back in time. The exponential smoothing model:

$Z_{t+1} = w(Z_t) + (1\text{-}w)\ (Z_t);$

Where, Z_t is the actual known series value at time t, and Z_t is the forecast value for time t.

In the forecast studies, the following exponential equation fits with actual data:

$\ln Y = bX + a$ or $Y = a\ e^{bX}$

Using these concepts and formulae, exponential equations are constructed using the queried data. Power fit has also been used in the present studies for some situations. The following power fit displays a power fit through the data:

$\ln Y = b(\ln X) + a$ or $Y = a\ X^b$

Using this relation, power equations are computed for situations of queries generated from mineral and petroleum exploration data sources.

Interpretation of the explored data

As cited in research objectives, RO 5 in Section 1.3.2 of Chapter 1, author interprets the data views extracted from metadata for new patterns and trends. For this, mineral exploration and discovery and petroleum exploration and production data dimensions are analyzed, considering the attributes and attribute instances that vary with time and space. The trends in the data in particular, when the data vary with period, have significance while forecasting for future oil and gas resources in Australia. Four main types of trends are distinguished from the mineral and petroleum data views. First one is secular trend, when attribute values are interpreted in increasing or decreasing direction. Second trend, called cyclical movement or variation is due to expansion or contraction of economy. At times, inflationary pressures affect these cyclical movements. Third one is seasonal, interpreted in regular periodic fluctuations, for example monthly or quarterly. The last one is irregular variation or trend that refers to uncontrollable and random variations, impacting greatly on the level of business activity. Some examples of this type are: extreme weather patterns (cyclone, floods, fires, for example), extreme business variations (oil price down, stock market crash, drop in Australian dollar value) or extreme political situations.

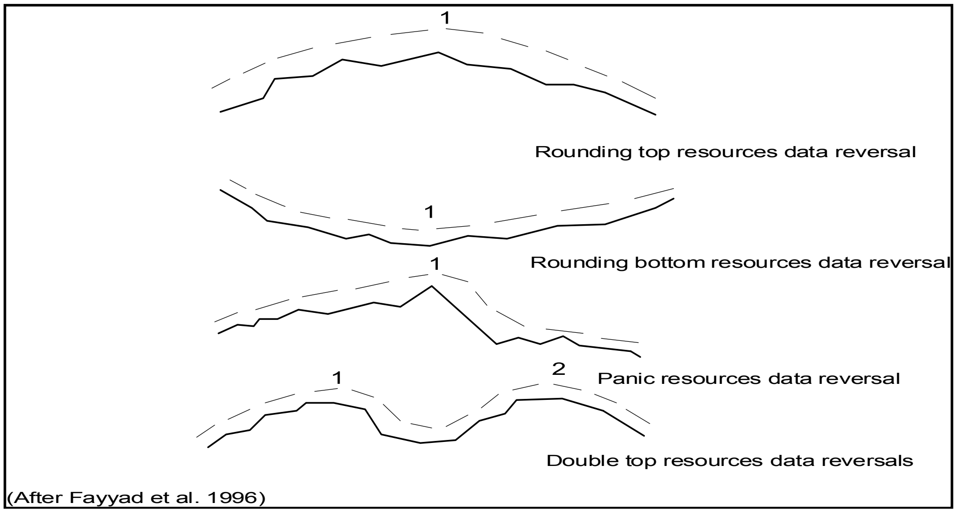

1
Rounding top resources data reversal
1
Rounding bottom resources data reversal
1
Panic resources data reversal
1
2
Double top resources data reversals
(After Fayyad et al. 1996)

Figure 5.14: Patterns analysis used in the oil and gas data domain

As shown in Figure 5.14 (Fayyad et al. 1996), the data reversal implies similarities and coherencies at early and later periods. The author examines quarterly data for investigating similarity trends in the exploration costs data except in the second quarter during 1994. The author interprets an economic peak for all quarters, indicating a rigorous mineral exploration and mining activity during this period. A polynomial equation is fit with actual cost data, with a good correlation. Rounding top oil and gas data reversals (Figure 5.14) are observed. The variations observed in the actual data during years 1992 and 1997 are cyclic or seasonal (see Figure 5.15). There is more investment on mineral exploration, thereby indicating more demand of minerals, in particular with state of New South Wales (NSW).

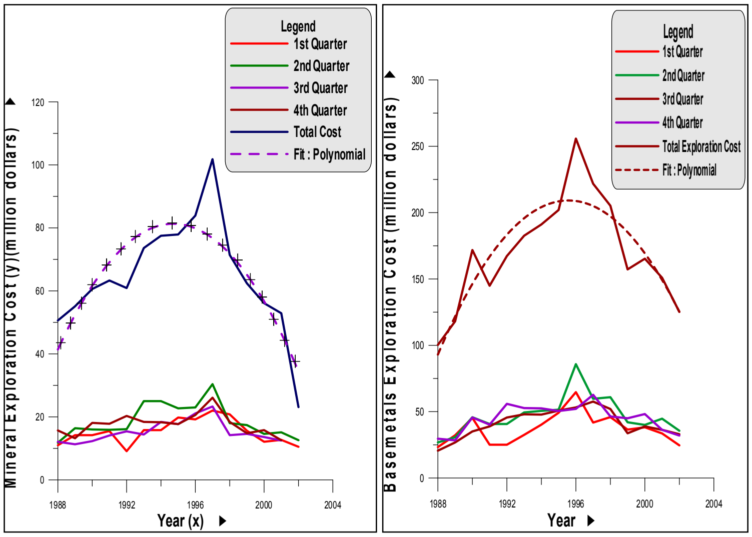


Figure 5.15: Quarterly presentation of NSW state's mineral (base metals) exploration cost data attribute

The author plots quarterly exploration costs data (Figure 5.15) for base metal minerals. There are periodic fluctuations especially during years 1990, 1996 and 1999, which could be due to seasonal as illustrated in Figure 5.16. Maximum exploration cost is interpreted during 1996, with a decreasing trend and increasing period, which appears to be rounding top resources reversal trends (as patterns of models envisaged in Figure 5.14). A polynomial equation is constructed and thus to use for future exploration cost predictions. The author plots quarterly exploration cost for gold mineral (see Figure 5.16) for detecting any trends in the cost data. Maximum and minimum exploration costs have respectively been interpreted during years 1997 and 1991. In all these quarters, the author interprets similar exploration cost patterns. But the computed trend, as detailed in Figure 5.16, does not fit well with the actual cost data; because of actual data contain unexpected exploration costs, and indirectly affecting rise of prices and market demand. The panic reversals of data patterns are observed (as patterns of models envisaged in Figure 5.14), indicating high volume of gold mineral trading during 1997. As shown in Figure 5.16, a bar chart is another way of presenting the gold exploration cost data.

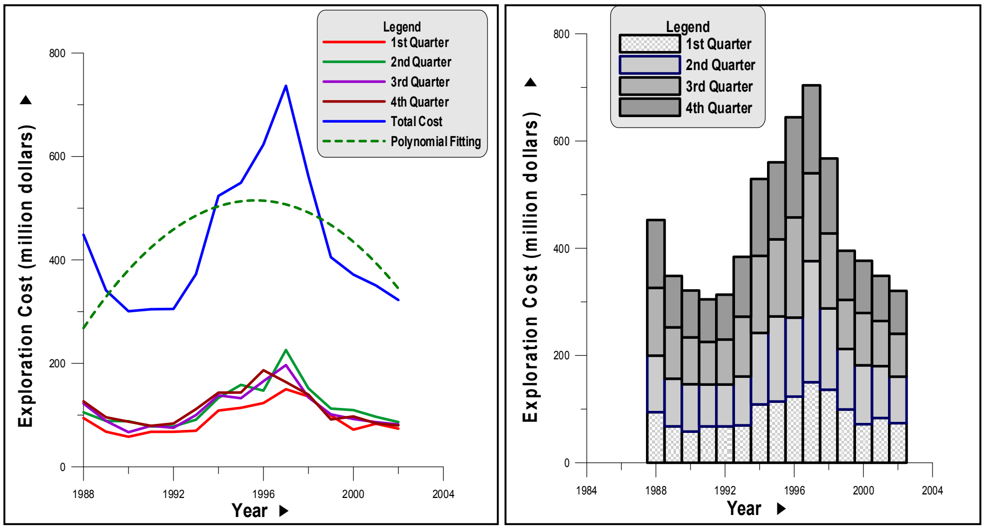


Figure 5.16: Quarterly presentations of gold exploration cost data attribute

In general, as narrated in Figure 5.17, the quarterly exploration expenditure patterns are similar to all quarters in the Western Australian state. The author reports maximum expenditure in 1997 and minimum in 1990. The computed trend provides a poor match with actual exploration cost, with lot of fluctuations in exploration costs data. Here panicked reversal patterns (Figure 5.14) indicate high volume of minerals trading in the WA state during the year 1997.

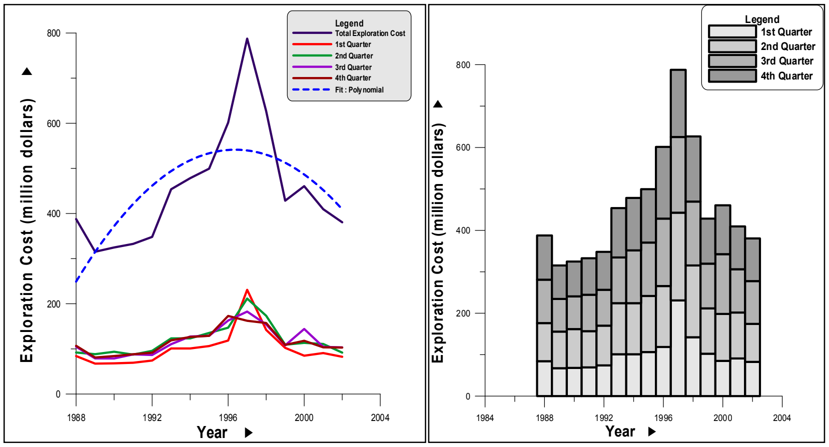


Figure 5.17: Quarterly presentation of WA state mineral exploration cost data attribute

Upward and downward periodic movements of actual exploration cost around the computed trend are cyclical and irregular at some periods as illustrated in Figure 5.18. Seasonal variations affect these trends in the actual data. In general, base metal exploration cost is in increasing trend with period. A parabolic curve is fitted with the actual data and thus an equation is constructed. Large periodic fluctuations observed between 1980 and 1990 are worth mentioning for base metals exploration. The variations are smaller in between 1955 and 1980. The author interprets peaks and troughs reported during 1965-1975 in the business point of view (see Figure 5.18). Any observed data that do not follow the smooth fitted curve without cyclical movements are indicative of random factors of influence. Secular and regular periodic fluctuation trends are again due to seasonal effects.

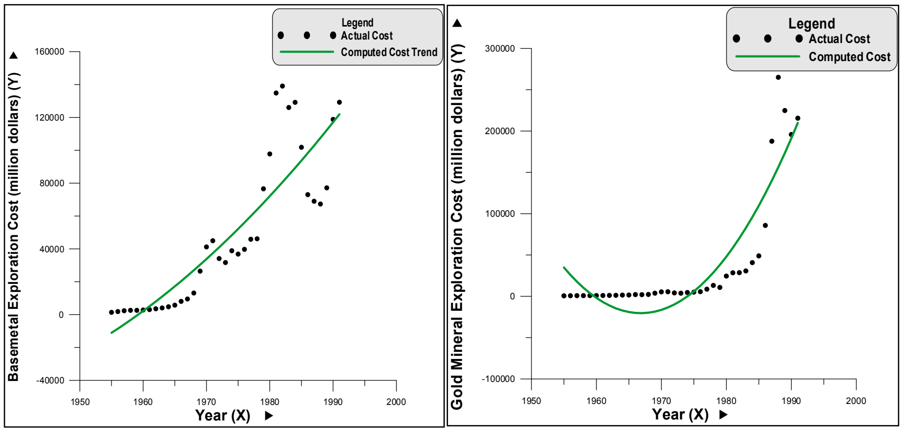


Figure 5.18: Construction of parabolic equations for base metal and gold actual exploration cost attribute

As illustrated in Figure 5.18, author fits a parabolic curve with the data of actual exploration cost of gold, which matches well. This equation may be used for predicting the future exploration costs of gold mineral in Australia. Again, the author computes and interprets another parabolic curve match-fit (Figure 5.19) with the actual exploration cost data of Northern Territory (NT). Fairly a good match is observed and may be used to compute future exploration costs in northern Australia.

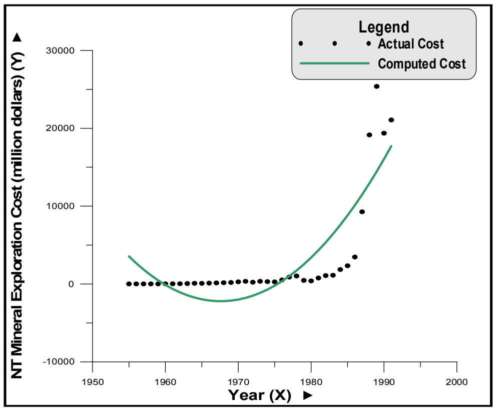


Figure 5.19: Construction of parabolic equation for NT (Northern Territory) mineral exploration cost attribute

The author plots total mineral exploration cost and number of mineral discoveries (both viewed as dimensions in the database) made in Australia (Figure 5.20) for establishing any trends and or correlations. In general, the actual data are not user friendly with the computed trend. There are many random fluctuations in the actual data and the constructed polynomial equation interprets an increase in exploration cost trend with corresponding decreasing trend in the number of mineral discoveries. As shown in Figure 5.20, computed trend indicates that even after increase in exploration cost, the number of mineral discoveries substantially has fallen down.

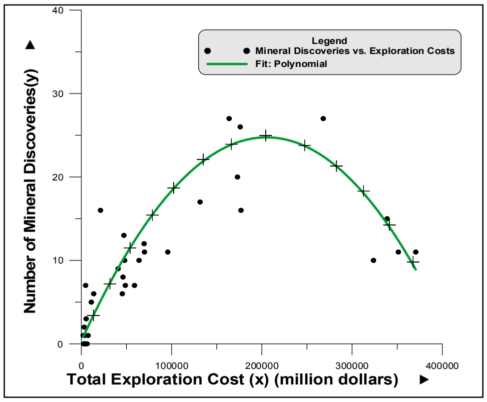


Figure 5.20: Construction of polynomial equation between total exploration cost and number of mineral discoveries attributes made in Australia

The author makes similar analysis in the case of Western Australian mineral discoveries vs. mineral exploration cost data (Figure 5.21). Number of mineral discoveries has fallen down in WA even with increase of mineral exploration costs. Future trends of "mineral discoveries against exploration costs" can be predicted from the polynomial curve fit, as provided in Figure 5.21.

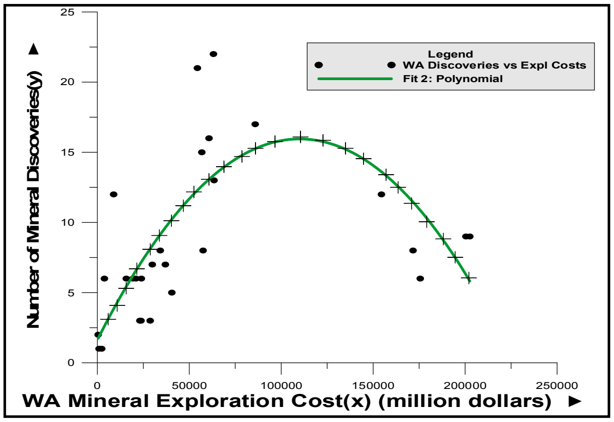


Figure 5.21: Construction of polynomial equation between (Western Australia) WA mineral exploration cost and mineral discoveries' attributes made

A similar analysis is done using Queensland (QLD) State's mineral exploration cost data (Figure 5.22) and the number of mineral discoveries made. In general, the parabolic curve fits with the actual data. There are periodic fluctuations in the exploration costs particularly in the years 1970, 1980 and 1985 with peaks and troughs, which are interpreted as seasonal. The computed trend again, can guide the future predictions in the QLD state. A minimum exploration cost is reported for the year 1965, as shown in Figure 5.22.

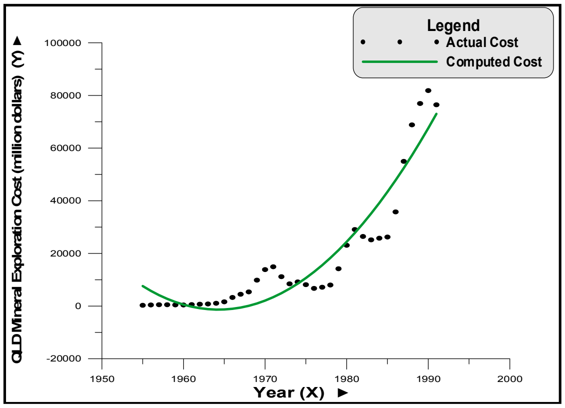


Figure 5.22: Construction of parabolic equation for Queensland (QLD) state actual mineral exploration cost data attribute

As illustrated in Figure 5.23, mineral exploration costs in the rest of Australia (excluding WA, NT and QLD states) seem to be in increasing trend with period (secular variation), but could have affected due to seasonal variations especially during 1970 and 1988 years. The parabolic computed trend matches with actual data, but with minor fluctuations in the exploration costs around the computed trend. However, the computed trend is used to predict the future exploration costs.

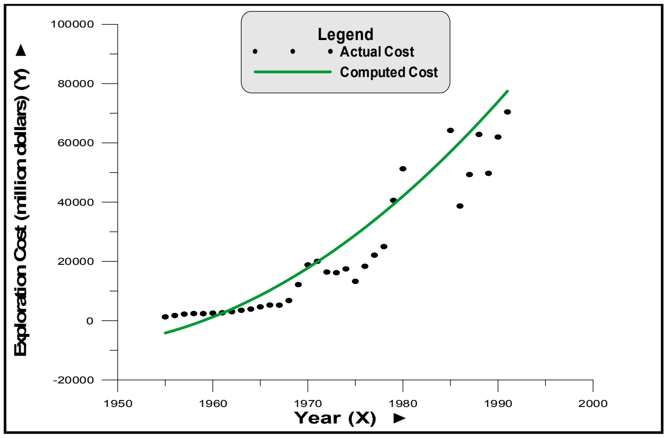


Figure 5.23: Construction of parabolic equation for mineral exploration costs data attribute

The Western Australia's actual exploration costs (as shown in Figure 5.24) data are fit well with the parabolic trend, with minor fluctuations (irregular trends) around the computed trend. The correlation coefficient between these costs and the period appears to be good.

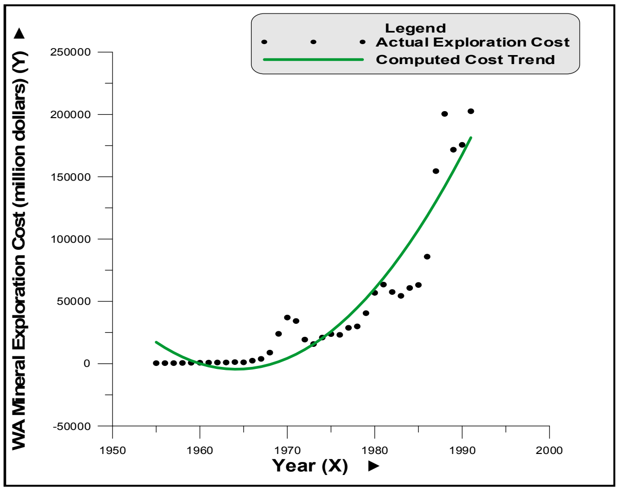


Figure 5.24: Construction of parabolic equation for WA mineral exploration cost data attribute

So far, the author interprets mineral exploration costs data. Exploration of petroleum resources is also active in Australia, especially in the state of WA. Quarterly costs of petroleum lease exploration areas do not match each other as shown in Figure 5.25. But there is general increase in the trend (secular variation) of other lease costs with period. Fluctuations in the actual lease costs are irregular, but minor. The coefficient of determination is fairly good as shown in the Figure 5.25.

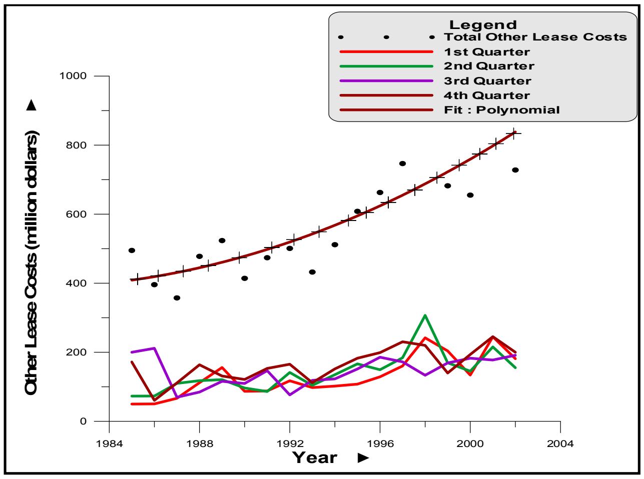


Figure 5.25: Correlation among quarterly data of other petroleum lease costs data attribute

The combined quarterly offshore meterage data, in general, do match data patterns (see Figure 5.26) with computed trend. But data are not correlatable to each quarter. Maximum meterage drilled in the year 1983 have two peaks, interpreted one in 1983 and the other in 1979, separated by a trough in 1981. This is another top reversal resources data model pattern (as envisaged in Figure 5.14).

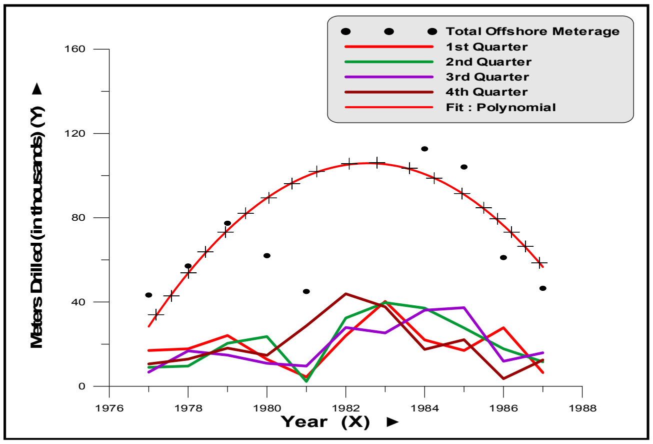


Figure 5.26: Analysis of offshore meterage drilled data attributes

In general, the patterns (Figure 5.27) of quarterly onshore petroleum meterage drilled data responses do match. Maximum meterage, drilled is in the year 1984. Peaks are in 1982 and 1984 and troughs are in 1983 and 1986. In general, there is periodic increase in the data. This is another panicked top reversal data pattern shape as narrated in Figure 5.14. This explicitly indicates more wells are drilled in recent years, with a pursuit of more oil production and meet demand. Global oil pricing and more demand of crude are also contributing factors for this type of data response patterns. The current global fall of oil prices have created panic in other industries.

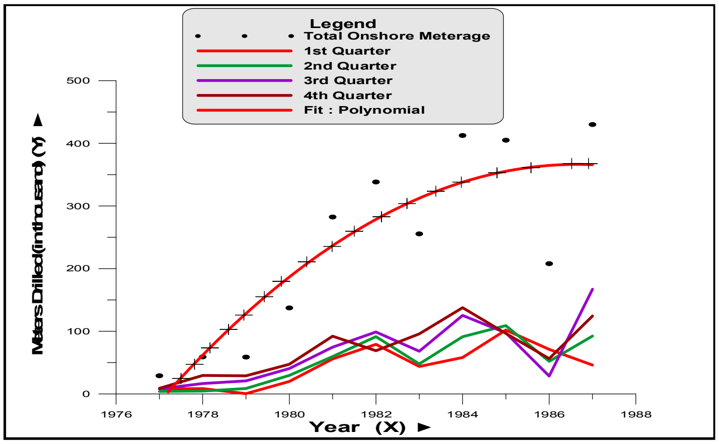


Figure 5.27: Analysis of onshore meterage drilled data attributes

An exponential equation (as drawn in Figure 5.28) drawn between two attribute variables seismic (surface) line kilometers and meters drilled (sub-surface), suggests a poor correlation and is not user friendly. But one is independent of the other variable. But the data trend indicates an exponential relation between these two variables as illustrated in Figure 5.28. However, these computed trends may help the seismic field parties to have an advanced fact of the meters to be drilled in an under investigation.

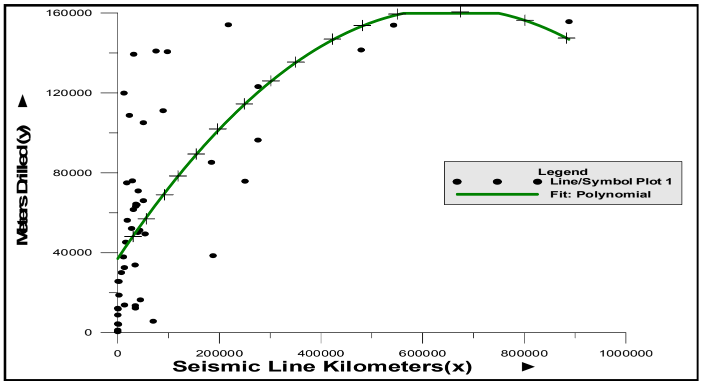


Figure 5.28: Construction of polynomial equation between seismic line kilometers and meterage drilled data attributes

A linear equation (Figure 5.29) is established between the petroleum production and petroleum consumption. The actual data perfectly matches with the computed trend. With increase in production there is corresponding effect on its consumption as demonstrated in Figure 5.29. The author clarifies that this is a plot between public consumption of petroleum products and petroleum production, narrating a linear relationship. Though time has a role in petroleum products' consumption, but in Figure 5.29, it has no role to play.

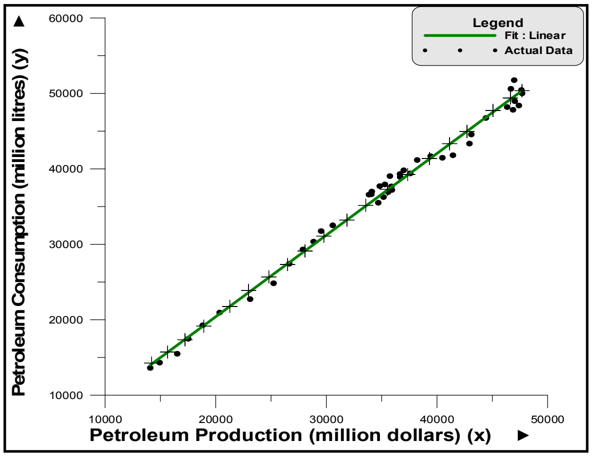


Figure 5.29: Construction of linear equation between petroleum production and petroleum consumption data attributes

The actual offshore exploration costs data is fitted with the computed parabolic trend as shown in Figure 5.30. There is good fit, because of good correlation coefficient. The fluctuations in the offshore exploration costs data are due to spending time.

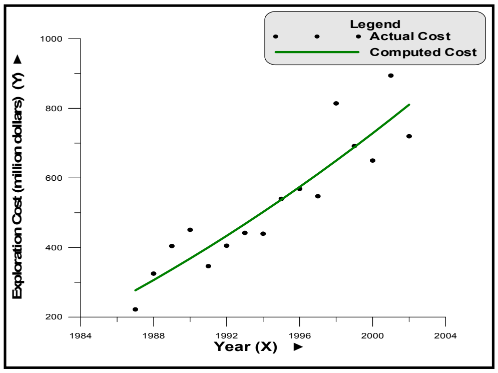


Figure 5.30: Construction of parabolic equation for actual offshore exploration costs

Irregular fluctuations of the actual data around the computed trend may be due to the irregular or random variations with various external factors. They may be political or drop in Australian dollar value or any other natural calamities during these periods. Author plots petroleum actual exploration cost and petroleum production to explore for any correlation (see Figure 5.30). Original exploration costs data are not user friendly. A polynomial equation constructed between these two dependent (if they are linearly proportional) variables, fairly matches with the actual data as illustrated in Figure 5.31. One has to be careful in using this computed trend, though coefficient of determination is fairly good. The actual data appears to have irregular trends. These irregular variations in the actual exploration costs data could be due to unforeseen and external situations. But the computed trend appears to follow the actual data trend and so one can make use of this model to predict the future petroleum production, having known the petroleum exploration costs.

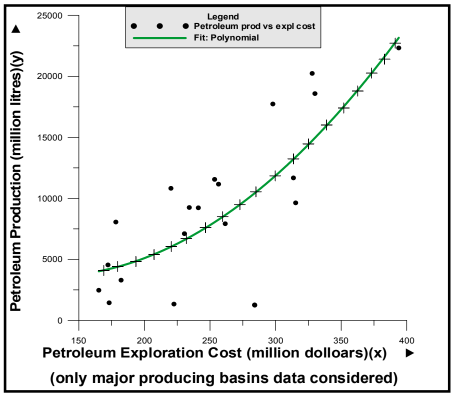


Figure 5.31: Construction of polynomial equation between petroleum exploration cost and petroleum production data attributes

A trend analysis is done between petroleum production and petroleum export attributes (see Figure 5.32). The constructed computed trend matches with the actual data, though there are irregular fluctuations. The constructed curve provides future predictions of petroleum production and exports.

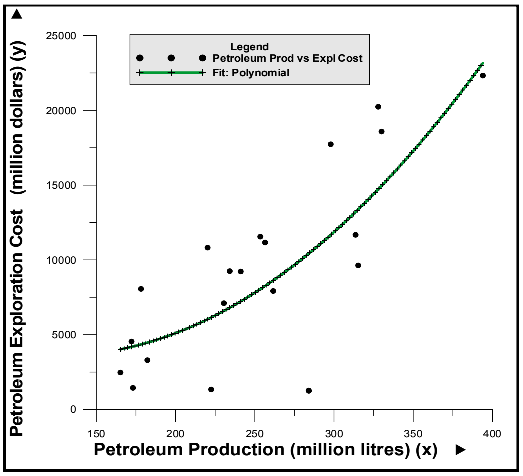


Figure 5.32: Construction of polynomial equation between petroleum production and petroleum exports data attributes

The regression analysis done between two dependent variables, such as number of surveys conducted and number of hydrocarbon producing wells suggests that the correlation coefficient (r = 0.47, see Figure 5.33) is poor. Regression analysis done between number of surveys and number of structures provided a fair (r = 0.655, see Figure 5.34) correlation. Similar analysis done between wells drilled and number of hydrocarbon producing wells provides a good correlation coefficient (r= 0.878, Figure 5.35). Regression analysis carried out between number of structures interpreted v. number of hydrocarbon producing wells has provided a strong correlation coefficient (r=0.932, Figure 5.36). Regression equations may be used for future prediction of the attributes. The number of surveys v. number of wells drilled has been used for regression analysis and a fair (r=0.68, see Figure 5.37a) correlation has been established. These equations (X on Y and Y on X) may be used to predict the two variables. Figure 5.37b shows relationship and construction of a trend between *number of surveys* and *number of wells drilled* attributes.

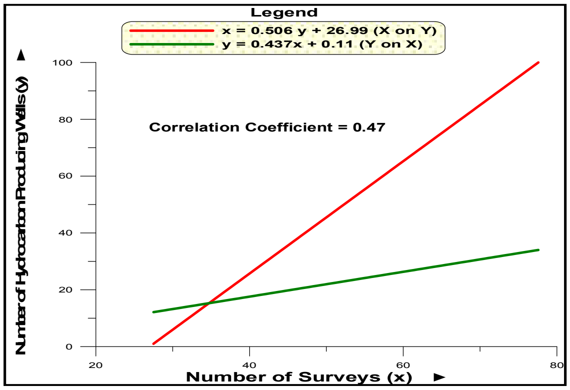


Figure 5.33: Regression analysis between *number of surveys* and *number of hydrocarbon producing wells* data attributes

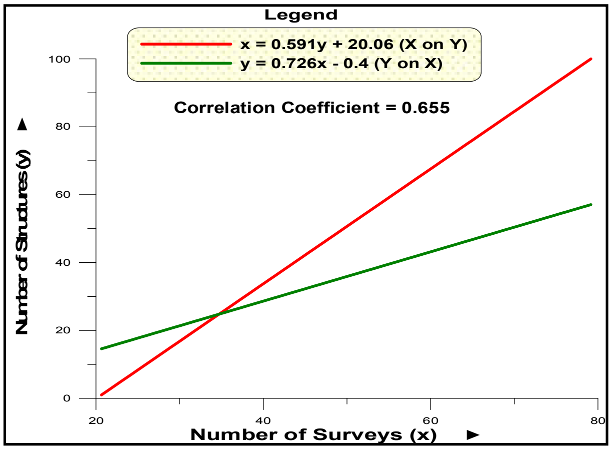


Figure 5.34: Regression analysis between *number of surveys* and *number of structures* interpreted data attributes

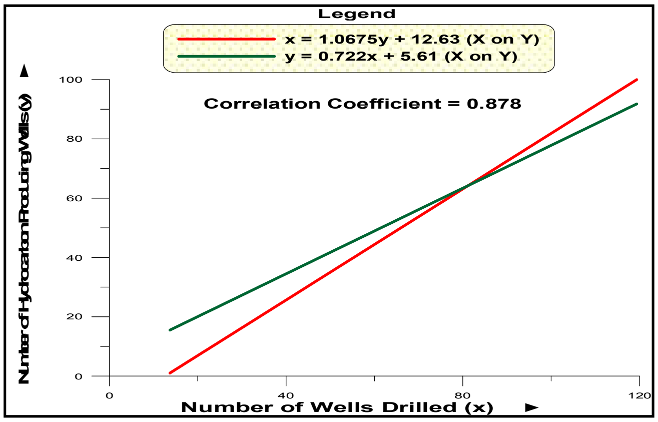


Figure 5.35: Regression analysis between *number of wells* drilled and *number of hydrocarbon producing wells* data attributes

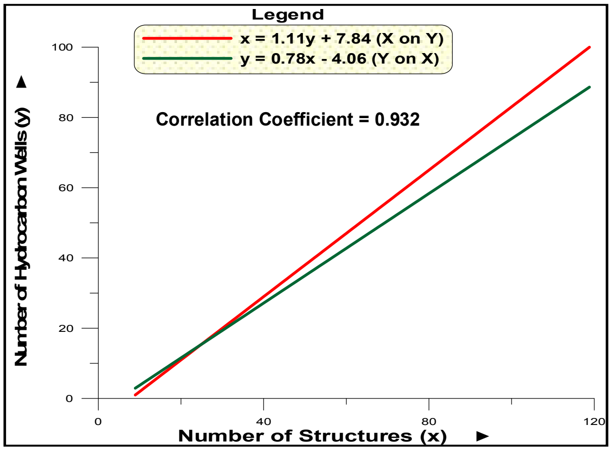


Figure 5.36: Regression analysis between *number of structures* and *number of hydrocarbon producing wells* data attributes

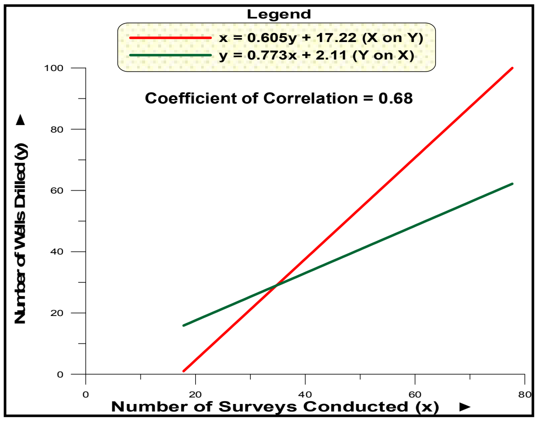


Figure 5.37a: Regression analysis between *number of surveys* and *number of wells* drilled attributes

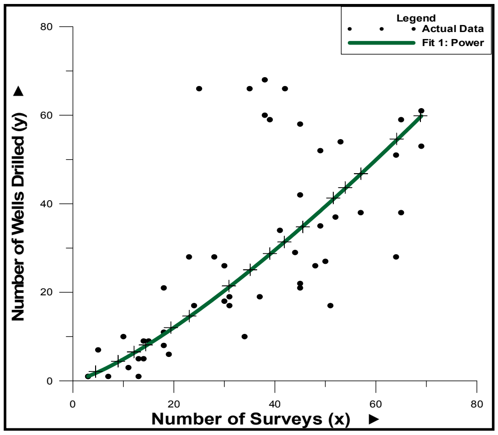


Figure 5.37b: Polynomial fit between *number of surveys* vs. *number of wells drilled* attributes

The author clarifies that the dimensions involved in the petroleum exploration are considered for demonstrating their connectivity and establishing associations. Necessary analysis is done for each correlation, establishing the implementation of metadata through analysis of data views.

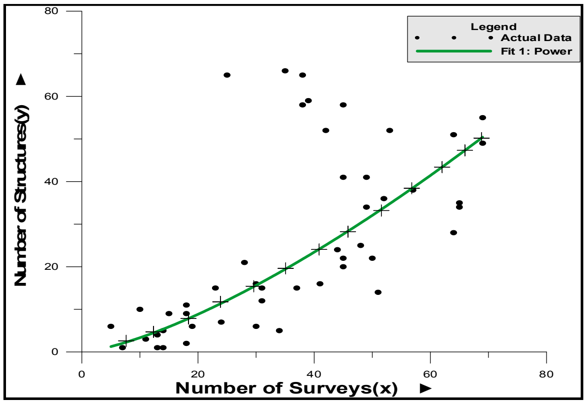


Figure 5.37c: Polynomial fit between *number of surveys* vs. *number of structures* attributes

A polynomial fit generated (Figure 5.37c) between number of surveys vs. number of structures, appears good with initial number of surveys, but increasing number of surveys, the fit between number of structures is just fair. A correlation is searched (as shown in Figure 5.38) between the number of wells drilled and the number of hydrocarbon producing wells (attribute instances). One variable depends on the other. The computer trend is very gentle initially and much steeper with increase in the number of wells drilled. This suggests that with increasing number of wells drilled, there is much more chance of getting more number of hydrocarbon wells (with more petroleum production).

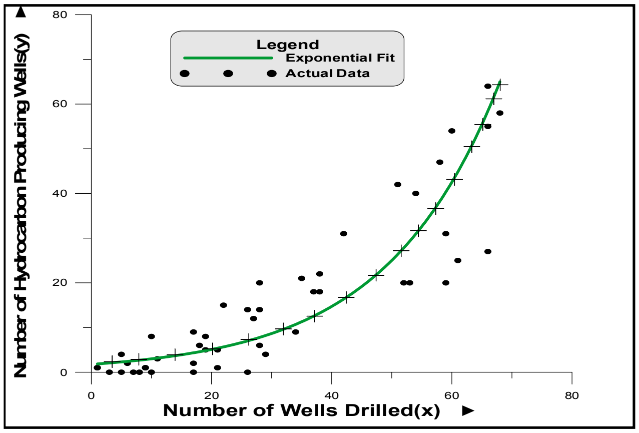


Figure 5.38: Construction of exponential equation between *number of wells* drilled and *number of hydrocarbon producing wells* data attributes

Similar trends in the data of petroleum products such as crude, LPG, condensate and a general periodic increase in petroleum products with regular fluctuations in crude and condensate after 1985 as demonstrated in Figure 5.39 are due to seasonal. Coefficients of determinations are strong as shown in the Figure 5.39. This panicked shaped data patterns (Figure 5.14) indicate more demand for crude and strongly suggests global oil pricing.

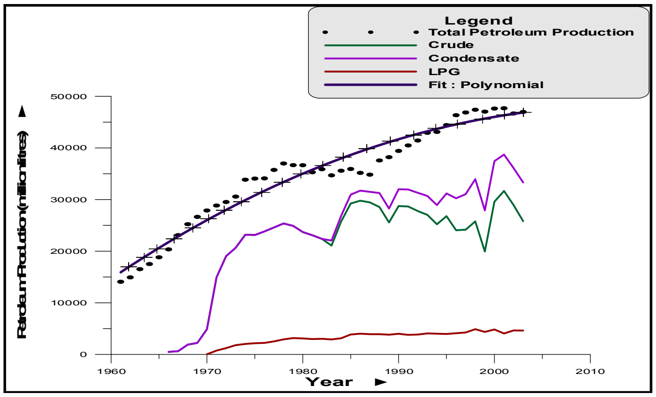


Figure 5.39: Correlation analysis of Australian petroleum products data attributes

The number of structures interpreted vs. *the number of gas producing wells* (attributes and their instances) are correlated (Figure 5.40). The computer trend interprets a very gentle rise initially and the trend rises sharply at higher number of structures. This implies that with more number of petroleum structures interpreted, there is an increase in number of gas producing wells (see Figure 5.40 for details) in Australia.

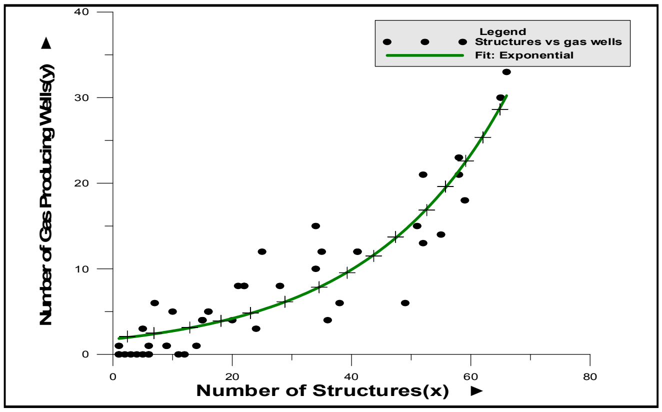


Figure 5.40: Construction of exponential equation between *number of structures* and *number of gas producing wells* data attributes

The author correlates attribute instances of, *the number of structures* (and surveys) interpreted vs. *the number of hydrocarbon producing wells* (Figure 5.41a). The computer trend interprets a very gentle rise initially slowly and the trend rises steeply at higher number of structures. This implies that with more number of petroleum structures interpreted, there is an increase in number of hydrocarbon producing wells. In case of surveys vs. number of producing wells, there is noise present in the data, as envisaged in Figure 5.41b. However, for increase in number of surveys, there is corresponding increase in number of hydrocarbon producing wells.

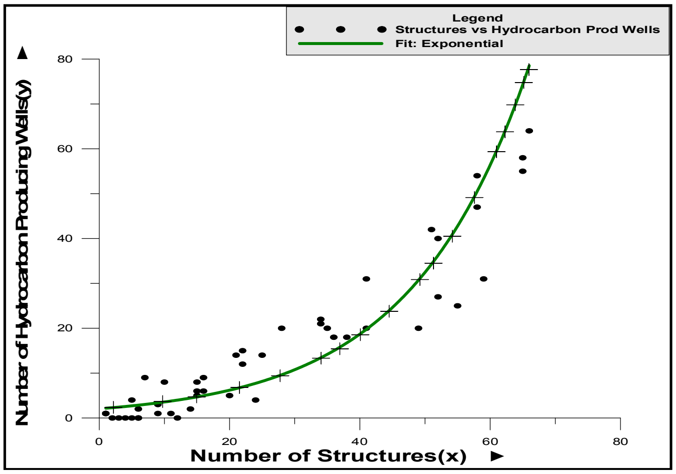


Figure 5.41a: Construction of exponential equation between *number of structures* and *number of hydrocarbon producing wells* data attributes

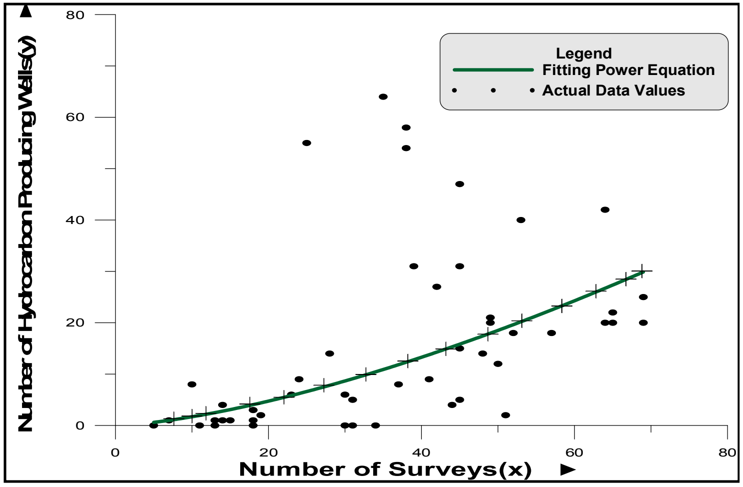


Figure 5.41b: Construction of power equation between *number of surveys* and *number of hydrocarbon producing wells* attributes

An attribute instance such as, *the number of petroleum structures* is again correlated with *the number of oil producing wells* as shown in Figure 5.42. In this case, the computer trend is much sharper even for smaller number of structures, which indicates that in spite of less number of petroleum structures (geological new knowledge) interpreted, with increasing number of oil producing wells.

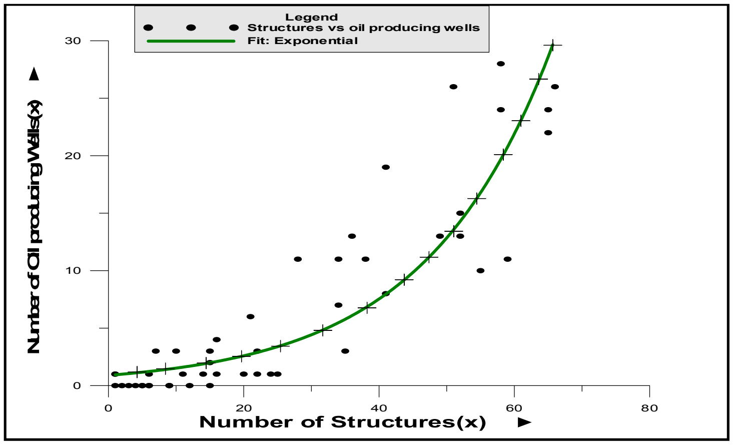


Figure 5.42: Construction of exponential equation between number of *structures* and *number of oil producing wells* data attributes

Number of wells and number of condensate wells plotted has provided a good correlation in the years 1971-72, 1993-2000, but fair correlation in between years 1985-1990 (Figure 5.43). There is quite good correlation drawn between wells drilled and number of gas producing wells in the years 1963-1970, 1971-1990 and 1991-2000 (Figure 5.44).

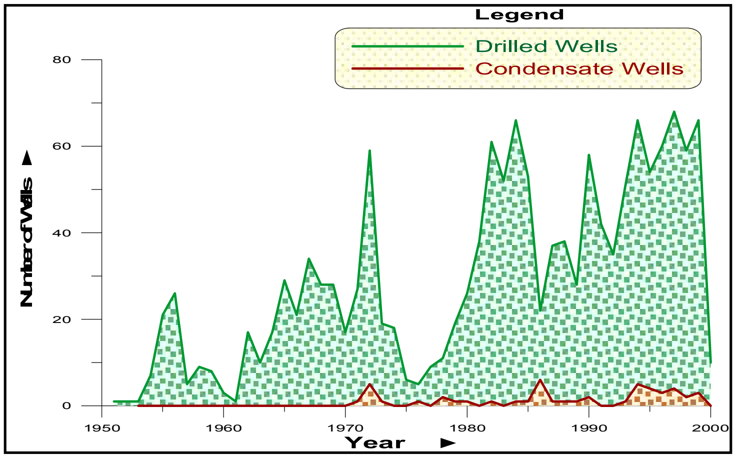


Figure 5.43: Correlation analysis between number of *wells drilled* and *condensate producing wells* data attributes

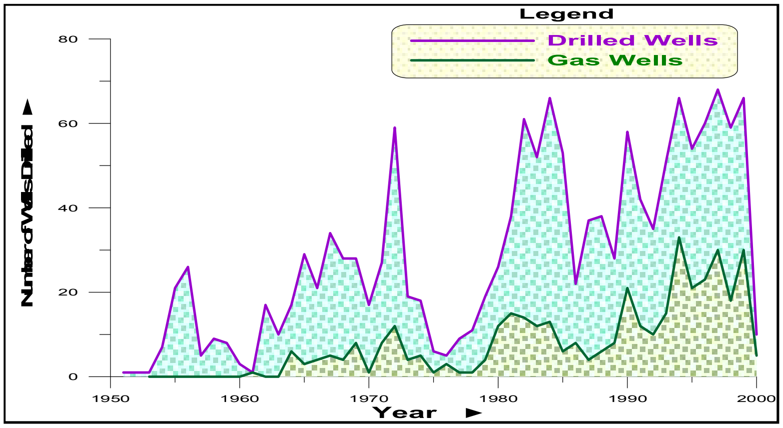


Figure 5.44: Correlation analysis between number of *wells drilled* and number of *gas producing wells* attributes

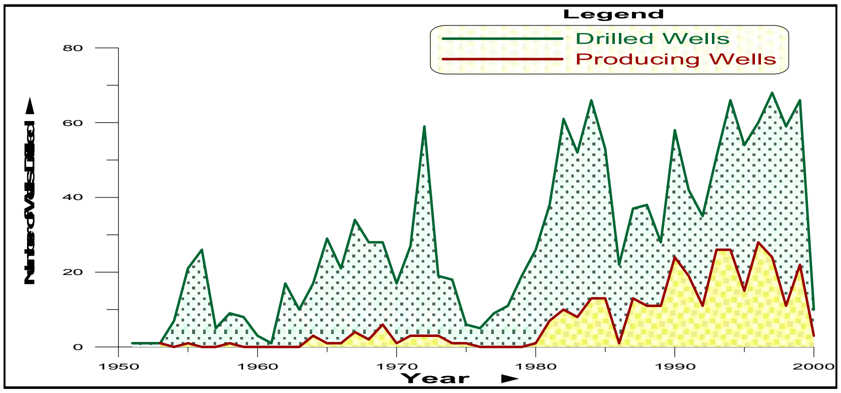


Figure 5.45: Correlation analysis between *wells drilled* and *oil producing wells* attributes

The author correlates *number of drilled wells* vs. *number of oil producing wells* instances. These two variable attributes have dependence each other and has quite good correlation as illustrated in Figure 5.45. In general, the number of surveys conducted and wells drilled in Australia do match, except in the years 1972 and 1997. Similar good correlation is observed between the number of surveys, number of wells drilled and number of oil producing wells particularly in the years 1955, 1970 and 1980-2000 as narrated in Figure 5.46. As narrated in Figure 5.46, number of wells drilled matches with number of surveys conducted. In between years 1985 and 1990, less number of wells drilled in spite of increase in number of surveys, in which period, number of oil producing wells has also gone up. On the contrary, as demonstrated in Figure 5.46, in between 1995 and 2000, more number of wells drilled is drilled with similar increase of oil producing wells. In the same period, surprisingly, oil business was in bad shape in the sense; the value of barrel went drastically down during 1997.

Though an economic slowdown is perceived during this period, the oil and gas industry has been still active.

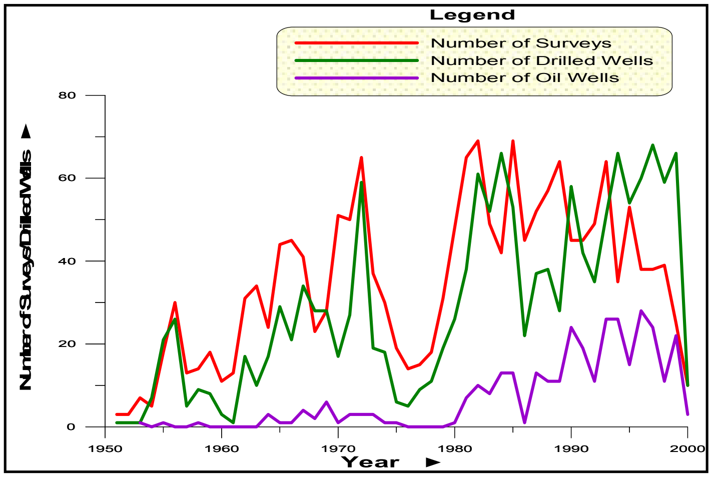


Figure 5.46: Trend analyses among number of *surveys*, *number of wells drilled* and *producing wells* for different *period* attributes

Data warehousing and mining methods facilitate so far the interpretation of predictions done during oil and gas business operations and their exploration costs controls. Keeping in view the current (in the year 2015) global fall of oil prices, other industries have created panic. There are potential opportunities and scope for further studies, especially in the context of development of data mining procedures and translating them for new knowledge interpretation.

5.3.3 Conventional digital ecosystems analysis

The author uses ontologies narrated for different elements of petroleum ecosystem for integrating volumes of data instances acquired from *structure*, *reservoir*, *source* and *seal* rock dimensions and their attributes. Several permutations and combinations are made in order to make sense of each classification of elements and processes. For example, similar characteristic reservoir instances documented for an interpreted horizon in a basin are initially warehoused for data integration after reservoir dimension ontology descriptions. As narrated in the following sections, A, B and C, the structure of dimensions are further fine-grained (Rudra and Nimmagadda 2005) using set theory (West 2006). Metadata volumes for each and every element attribute are computed for slicing and dicing cube operations, as demonstrated in Figures 5.48 and 5.49. Author plots data and map views representing group of reservoirs that characterize similarity or dissimilarity properties for visualization of the reservoir qualities and their interpretation. Another feature is combination attributes, in which multiple sets and their elements including attributes among ecosystems may have been interconnected. For example, the potential reservoirs interpreted within structural traps (in a geological sense, a trapping mechanism) and their framework or structural compartments appear to have a connections. Several tools, narrated in sections A, B and C are components of the methodological framework (objectives, RO2 to RO4) and big-data perspective, for analyzing the geological new knowledge.

A Data mining and visualization

Mining and visualization of data views extracted from warehouses are the next stages of any E- & -P data management project. Various mining algorithms are available in the literature and many tools are in the public domain (Pujari 2002 and Yao and Zhong 2000). We use classical statistical mining (Shastri and Dreher 2011 and Rudra and Nimmagadda 2005) and *Grapher* and *Surfer* solutions for data visualization purposes. Figures 5.47 and 5.48 show examples, in which the volume of seismic data instances, after converting to an attribute property (that characterizes the reservoir and its areal extents), has brought out limits of reservoir extents.

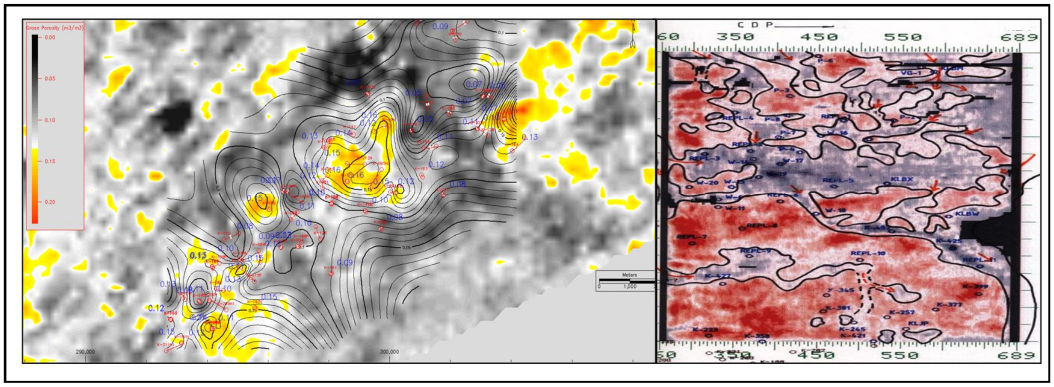


Figure 5.47: Map views showing connectivity among reservoir development areas at field scale

B Data Interpretation

The data interpretation and knowledge discovery are the final stages of reservoir management project (from G & G perspective). They facilitate the interpretation of reservoir-knowledge in different domains of oil and gas fields that can connect petroleum ecosystems and sedimentary basins. Seismic attributes at volume and surface levels are plotted and map views suggest multiple reservoirs and their areal extents are explicit both at field and basin levels. Through geological ages, seismic-driven reservoir attributes and their extents are known and interpreted by space and depth dimensions.

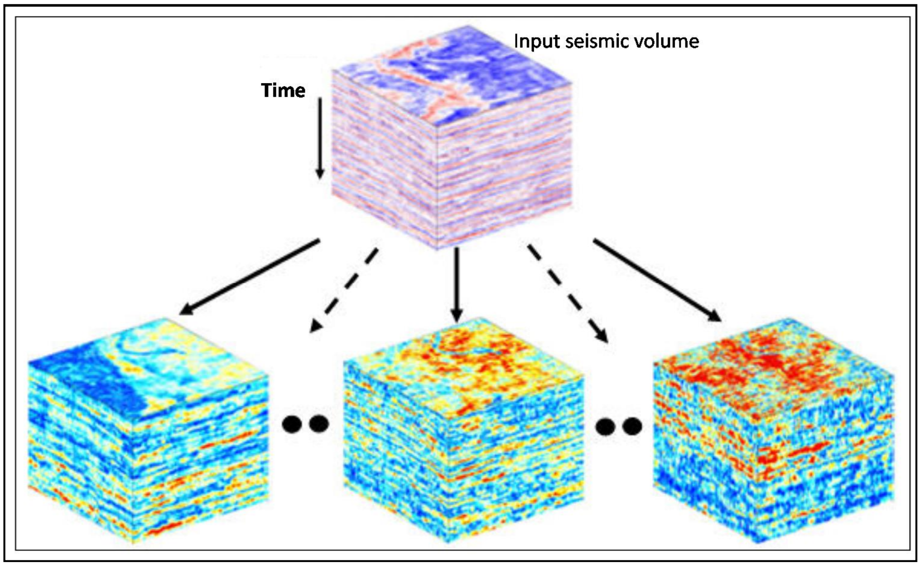


Figure 5.48: Exploring reservoir connections among data cubes

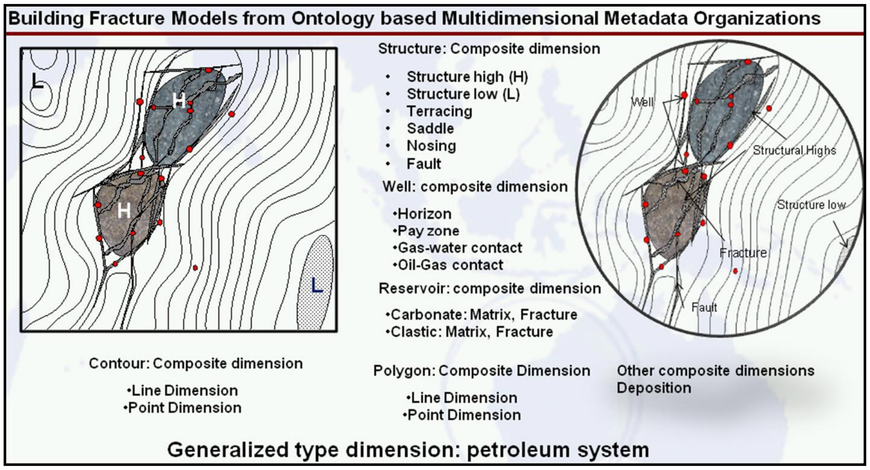
C Results and discussions

Determining the limits of reservoir extent is often an interpreter's challenging task and its scope is enormous in reservoir estimations. Multiple dimensions involved in organizing oil and gas data, for constructing domain ontologies and for integrating them, author uses data warehousing and mining approaches. Obtaining reliable multidimensional and heterogeneous data from the operational sources is crucial. Often, either data are not available for analysis or existing data are poorly organized, and not suitable for modelling and implementation of metadata models. Metadata are built using integrated frameworks. The author implements data views extracted from metadata, and integrated database for interpreting knowledge from multiple reservoirs occurring at different depths, ensuring that reservoir models explore for multiple connections. These models are then ready for reservoir simulation and management. In the case of fractured reservoirs, establishing connectivity among multiple fracture systems is significant in terms of production and productivity. The author reuses the data structures, deduced for each and every oil and gas field in other nearby fields within the same sedimentary basin setup. This facilitates an easy way to design a data warehouse on a sedimentary basin scale, which is as big as 10000 km^2 requiring several terabyte of database storage. Validity of data schemas designed for a particular field is ensured and the making of global schemas is facilitated.

5.3.4 Unconventional digital ecosystems analysis

Lancaster (1996) demonstrates commercial gas production from fractured shales. Shales are part of the source-cum-reservoir element in unconventional reservoirs. Keeping in the importance of fractures, their connectivity is explored through ecosystem phenomena and an integrated framework. Implementation of the data models and framework (Figures 3.36 - 3.38 in Chapter 3) is a major challenge for any producing (upstream) company. Several major challenges and implementation issues are discussed in Rudra and Nimmagadda (2005); since multidimensional data are fine-grained and all the heterogeneous data are in denormalized form. Knowledge building analysis, including interpretation of data views from massive data structures is a challenging task. With the increased volumes of periodic data, perceiving and retrieving knowledge from historical heterogeneous data is now relatively easy. At times, many applications were forced to integrate and share volumes of multi-disciplinary data without prior knowledge of the business (such as unconventional petroleum-play entities that broadly associate petroleum ecosystems' knowledge domains). The knowledge from past business data is explored and exploited for future business system improvements.

Unknown relationships among different business data entities also influence the economics, involved in running the business. Improved data management in time – and depth – domains has become effective in data mining multidimensional metadata. The data created from heterogeneous sources at times are difficult to interpret because of the volume and complexity of information and bounded (embedded) patterns. The visualization techniques facilitate uncovering of the hidden patterns in the data, we presented and interpreted. Different data views were extracted from an integrated ontology framework (Figures 3.36-3.38) for interpretation and knowledge extraction. In all the shale-gas projects, the reservoir ecosystems data, integrated in a warehouse modeling environment underwent domain ontology modeling process. Finally, the applicability and feasibility of data warehousing, supported by ontology, combined with application of data mining and visualization, have tremendous impact on business data system knowledge discovery systems that can change the economics of exploiting unconventional oil and gas in massive geological structures. As described in Figures 5.49 – 5.52, online analytic processing (OLAP) models in the form of map views are extracted from explored metadata (as described in Figures 5.49 – 5.52), in which seismic and interpreted well-domain data dimensions are representative of an integrated metadata structural model. This structural model is the basis for working on an unconventional resources project and it is further refined, based on geological and geophysical inputs provided in the framework (Figures 3.36-3.38, in Chapter 3).



Building Fracture Models from Ontology based Multidimensional Metadata Organizations

Structure: Composite dimension

- Structure high (H)
- Structure low (L)
- Terracing
- Saddle
- Nosing
- Fault

Well: composite dimension

- Horizon
- Pay zone
- Gas-water contact
- Oil-Gas contact

Reservoir: composite dimension

- Carbonate: Matrix, Fracture
- Clastic: Matrix, Fracture

Contour: Composite dimension

- Line Dimension
- Point Dimension

Polygon: Composite Dimension

- Line Dimension
- Point Dimension

Other composite dimensions
Deposition

Generalized type dimension: petroleum system

Figure 5.49: Several data attributes and their dimensions described in fractured reservoir ecosystems

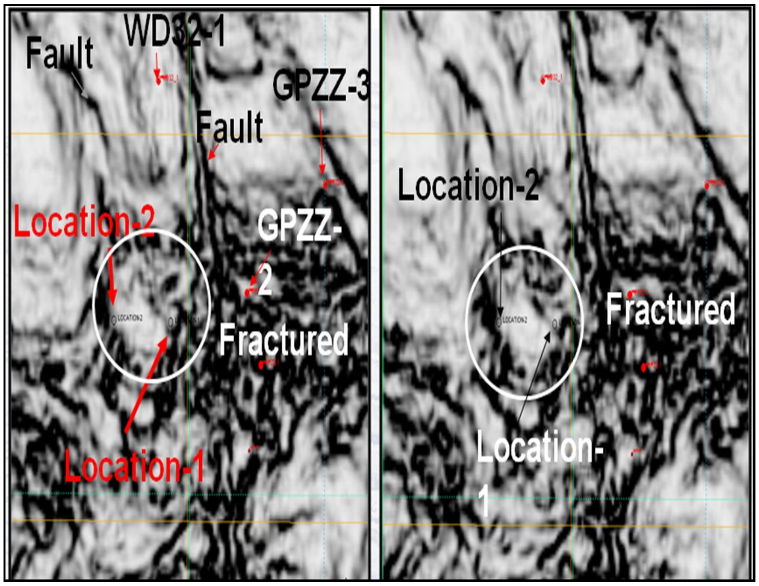


Figure 5.50: Fractures interpreted (fracture signatures), based on the connections interpreted from multidimensional metadata

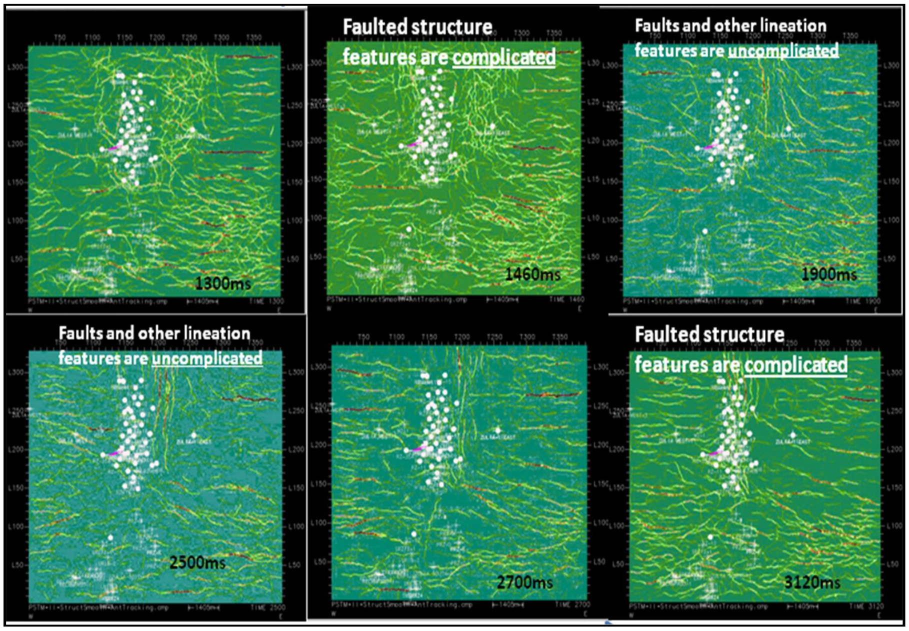


Figure 5.51: Fractured reservoir networks integrated and represented in map views

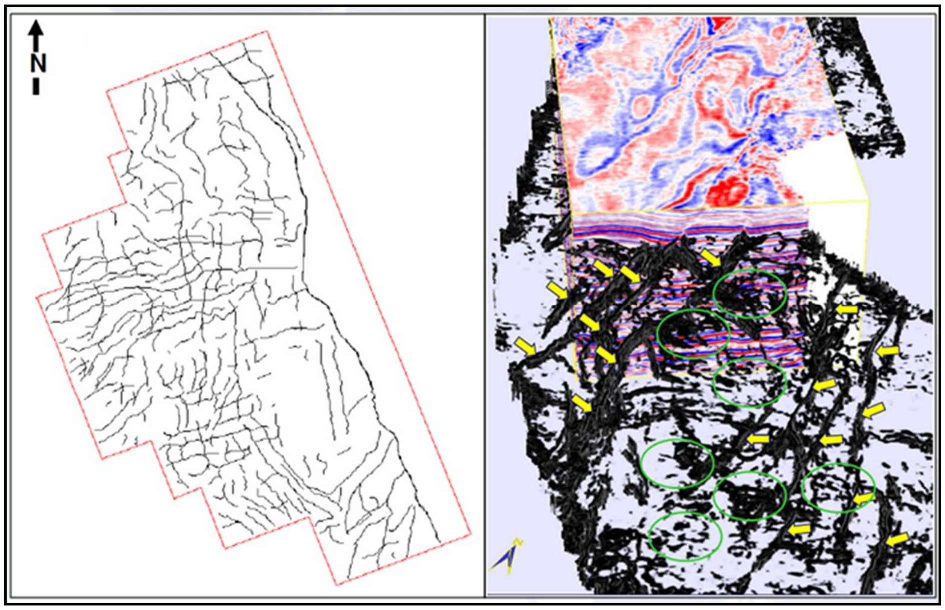


Figure 5.52: Map views of fracture ecosystems systems, built based on integrated metadata model (with arrowed fracture impressions)

Construction of a decision-tree mining model

A decision tree is a classification scheme (Castañeda et al. 2012 and Pujari 2002) that generates a tree (Figure 5.53) style mining and with set of mining rules or business constraints, representing the model of different classes from a given dataset. Two types of attributes are characterized - numerical and categorical. Training datasets and test datasets are two disjoint subsets, from which classifiers are derived, and the accuracy of the classifier measured. As an example, in the unconventional reservoir plays, based on *porosities* and *kerogen* content, several numerical and categorical attributes are deduced for evaluating fractured reservoir plays. *Plays* and *non-plays* are typical categorical attributes. Several leaf nodes are described from the decision-tree; each leaf node represents a rule.

Rule1: If the shale has, more than 5% porosity, it is play.

Rule 2: If the shale kerogen content is more than 7%, the shale is play.

Rule 3: If the kerogen and porosity are each less than 5% and 7%, the bulk volume is less than 3%, the shale is not a play (relatively poor fractured reservoir)

Rule: 4: Based on Rule 1 and Rule 2, if the shales are good fractured reservoirs, then this rule 4 holds good.

Rule: 5: If the attributes are not favorable as narrated in Rule 3, then the shale play does not hold good.

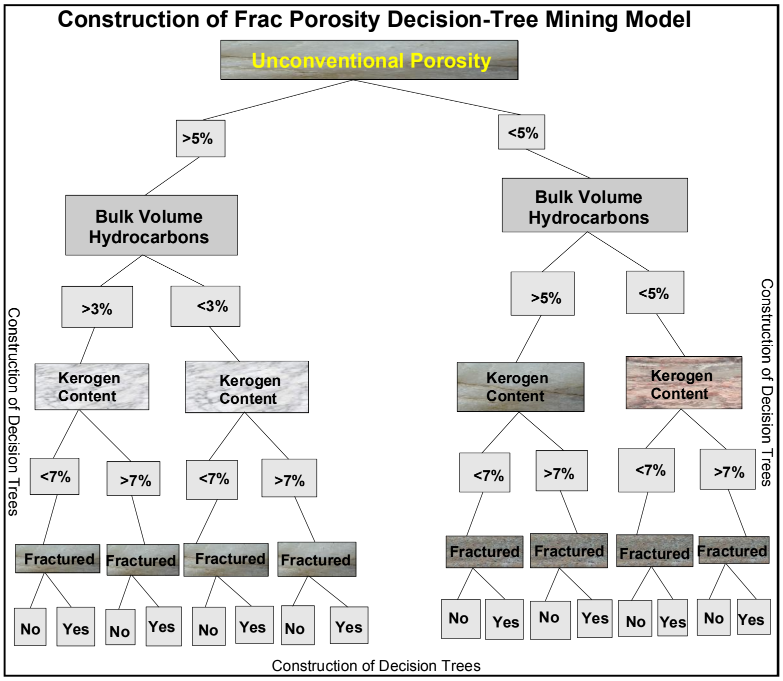


Figure 5.53: Decision-tree test-data mining model describing categorical and numerical attributes with rules

The data acquired from the public domain, are used to test accuracy of the classifier. Accuracies of rules are calculated from the training and test datasets and rule 4 appears to be 70% accuracy, compared to the accuracy of the other rules. Several multidimensional shale reservoirs (reservoir plays from different fields) and their properties have been analyzed. Each has unique characteristics and different properties. Though their individual attributes match for particular shale-play-dimensions, other attribute dimensions of the shale system (interpreting multidimensional and heterogeneity) are different as demonstrated in Figure 5.54.

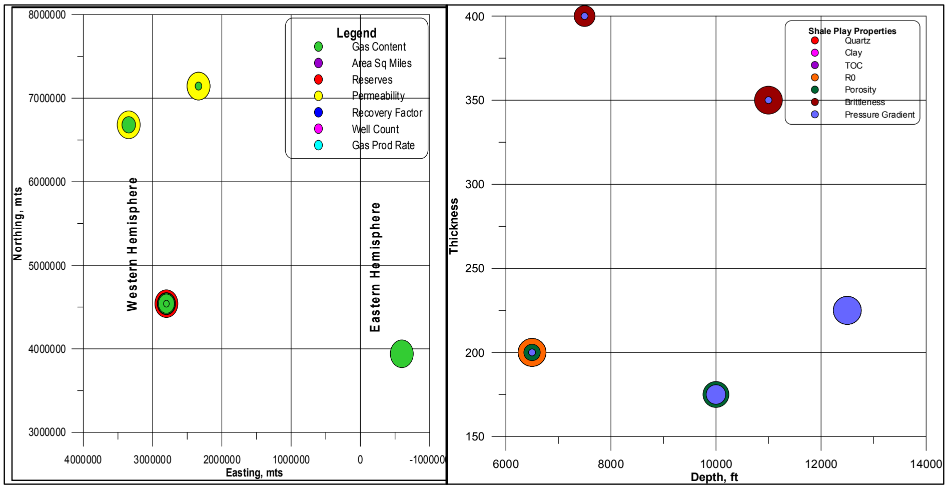


Figure 5.54: Plot views of attributes of unconventional shale gas plays extracted, from multidimensional metadata

Analysis of shale-gas ontology

The fractures and existing fracture networks among shales, including shales to be fracked or defracked have linking reservoir connections. The author uses ontology descriptions to make interconnections and explore effective networks of fractured shales. As interpreted in Figures 5.49 – 5.52, several fractured networked signatures are used for modelling reservoir capabilities, responsible for holding massive hydrocarbon accumulations and thus for reserve calculations. Classifications, decision trees and other mining rules are described for interpreting such fracture anomalies, in the following sections, A, B, C, and D:

A. Classifying multiple dimensions for rule mining

The discovery of association mining rules (Brown 2013, Creties et al. 2008, Durham 2013 and Gornik 2002) is solely dependent on discovery of frequent occurrence of multidimensional data attributes. Oil/gas businesses are often interested in *Yes* or *No* response, such as reservoir engineers wanting to know whether a reservoir is productive or not; and explorers seeking to establish if the petroleum system is productive or not with existing secondary porosity fracture system. These are classification issues, in which data attributes and their instances with finite number of classes, are explicitly described.

The classifying attributes are related to many other attributes, which may have been conceptualized among other classifications. For example, structure and reservoir attributes among several horizons have similarity and scalable property instances. Among reservoir subsets, there may be fracture attributes, both relationally and hierarchically interconnected among multiple horizons. Classifying the intensity and frequency of fractures in different orientations is an interesting data mining issue. Each fracture is characterized by its physical properties such as surface area and shape, and each has specific fluid flow properties—of permeability, compressibility, and aperture. It integrates the information from a wide range of sources including 2D and 3D seismic, maps, outcrops, reservoir geo-mechanics, well logs, well tests, and flow logs, as well as structural or depositional conceptual models. Several data views are extracted from the fractured reservoir data warehouse. Several dimensions are chosen for mining the data views and interpreting them for significant fractured reservoirs from integrated data warehouse model. Some such data views are deduced and interpreted in terms of multidimensional decision tree mining model, views from data cubes and cluster mining through bubble plot analysis among multiple dimensions and are given in the following sections.

B. Design of multidimensional decision trees

Which horizon (a production layer) has greater number of fractures, indicating the strength of porosity? The answer can assist in planning for borehole placement. A decision tree is a classification scheme which generates a tree type model and with a set of rules, representing the model of different classes from a given data set. Two disjoint subsets are made, which are ‘training set” and “test set”. The former is used for delivering the classifier while the latter is used to measure the classifier accuracy.

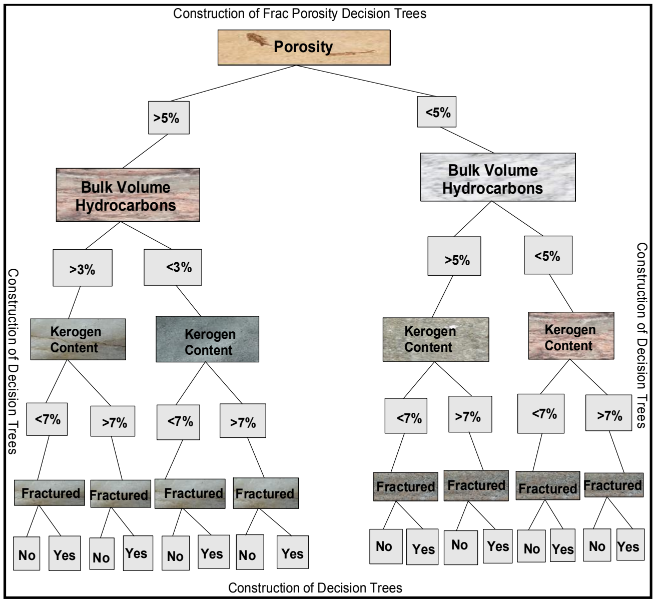


Figure 5.55: Multidimensional decision tree structure

The accuracy of a classifier is determined by the percentage of the test examples correctly classified. In our case, attributes are two different types - one is porosity, and the other is kerogen content. Attributes whose domain is numerical are called numerical attributes and non-numerical attributes are called categorical. Figure 5.55 shows a construct of decision tree mining model, in which various rules associated with porosity and kerogen instances and their cut-offs are described. Favorable data instances of porosity and kerogen content of rocks contribute to various fractures and their categorizations. Interpreters make use of this information model, as a decision making tool for ascertaining which type of rocks and kerogen content contribute to the hydrocarbon accumulations with favorable porosities. Major strengths of decision tree are, generating more logical and understandable mining rules, handling of both numerical and categorical attributes, and also providing clear clues of which fields are significant for prediction and classification.

C. Dimension modelling and data cube

The dimension modelling provides much semantic information (Matsuzawa and Fukuda 2000) especially about the hierarchical relationships between its elements. It is important to note that dimension modelling is a special technique for structuring data around fracture systems. The dimension modelling structures the numeric measures and the dimensions. The dimension schemas here, represent the details of the dimension modelling; in which period is key dimension that enables analysis of historical datasets. The dimension hierarchy helps viewing multidimensional fractured data in several data cube representations. The data views are finer to access, since they are derived from the fine-grained structuring (Pujari 2002 and Rudra and Nimmagadda 2005) and their domain ontologies.

A popular conceptual model that influences data warehouse architecture is a multidimensional view of the data, as shown in Fig 5.56. This model views data in the form of a data cube (more precisely, hypercube). It has multiple dimensions, each dimension again is subdivided. In this multidimensional model, there are sets of numerical measures that are the main theme or subject of the analysis. Each fracture type, such as open fracture, has different dimensional attributes, such as dip, azimuth, and density. There is more than one numeric measure. Each numeric measure depends on a set of dimensions, which provide the context for the measure. All the dimensions together are assumed to uniquely determine the measure; the multidimensional data views a measure as a value placed in a cell in the multidimensional space. Each dimension, in turn, is described by set of attributes. The attributes of a dimension may be related via a hierarchy of relationships (Figure 3.37) or by a lattice (Pujari 2002 and Hoffer et al. 2005). As an example, in Figure 5.57, drawn from the "frac data cube" is during 1998, under depth category, open fracture system has 5% fracture density (porosity) with a specific count of 76 and during 2002, under structure category, fracture density is 8% (porosity) with counting rate 100.

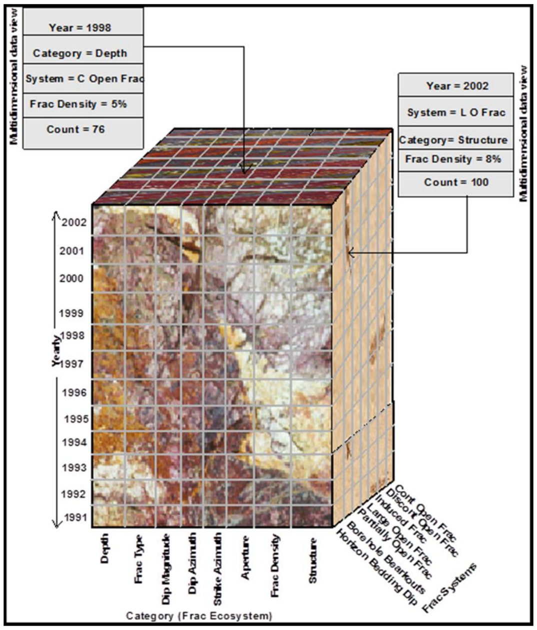


Figure 5.56: Multidimensional frac data cube – data views for interpretation

D. Multidimensional cluster mining

The cluster mining discovers data patterns (Figures 5.57 – 5.59) and distributions from large number of multidimensional attributes, organized in a warehouse environment. Identifying the dense and sparse regions of datasets, is significant and it is the real goal of multidimensional clustering. Number of attributes is multidimensional, in large size datasets. Horizons having more and similar type of fracture patterns belong to a particular cluster. Data instances that consist of large numerical data may be categorized into two groups, in which partition and hierarchical types are popular. Most of the algorithms existing today can handle multidimensional data, but they differ in their ability to handle different types of attributes, numerical, categorical, and accuracy of clustering.

Measuring the distances or similarity metrics among partitioned or hierarchical clusters is also a significant concept. Knowledge of which horizon has a greater number of fractures occurring in particular groups or types of fracture patterns, is helpful in planning for new borehole placement. As described in Figures 5.57, 5.58 and 5.59, based on strike and dip attributes and their magnitudes, different bubble sizes, densities and orientations are interpreted suggesting dip attribute magnitudes play roles on fracture orientations. In addition, geomechanical attributes, such as *stress* and *strain* attributes and their relationships on rock properties and around the drilling wellbore describe the orientations of fractures as interpreted in *dip* and *strike* attributes in Figures 5.57 – 5.60.

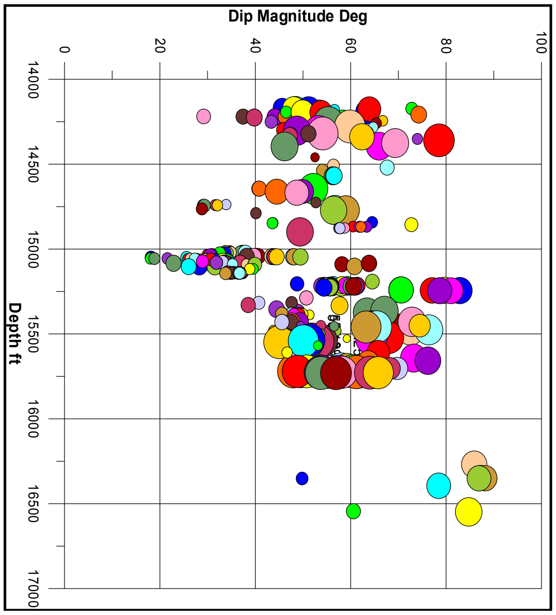


Figure 5.57: Depth vs. dip magnitude deg

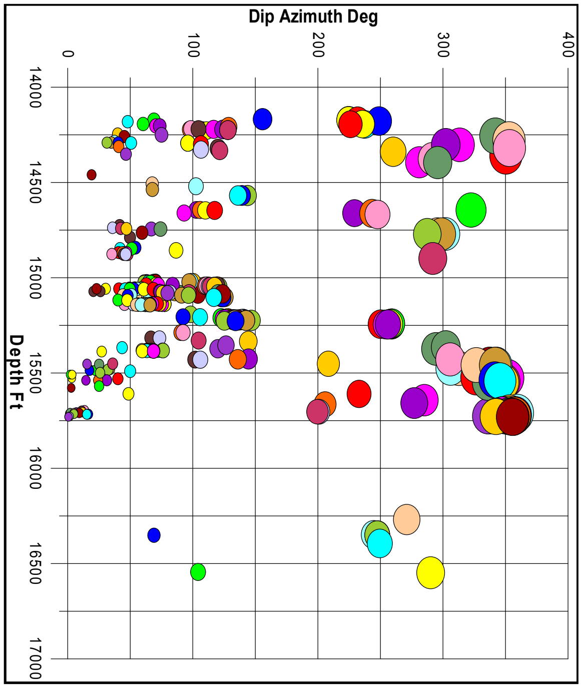


Figure 5.58: Depth vs. dip azimuth deg

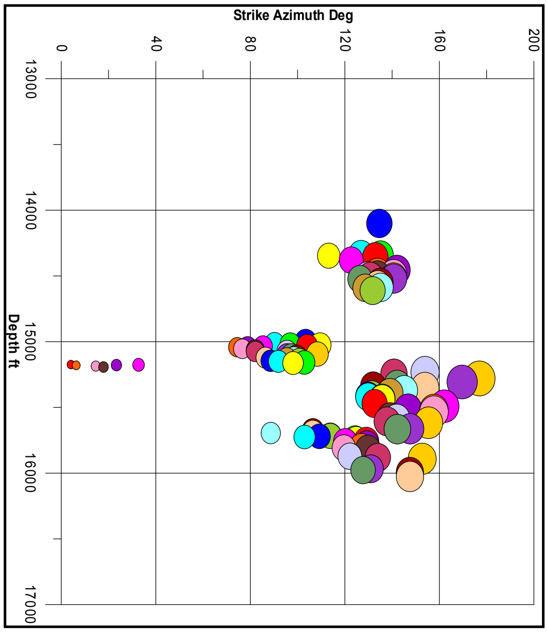


Figure 5.59: Depth vs. strike azimuth deg (borehole breakouts)

Results and discussions

The fractured reservoirs, especially carbonates hold significant oil and gas reserves, besides it is challenging to predict these reservoirs under complex anisotropic and heterogeneous conditions. Most carbonate reservoirs are naturally fractured and due to brittleness and size of fracture may vary from isolated microscopic fissures to kilometres-wide. At geological times and places, these fractures create complex paths for fluid movement, which impact reservoir characterization and ultimately production performance and total recovery.

Various operating companies in the Middle East have been exploring and developing the fractured reservoirs (Lancaster 1996), especially when the reservoirs are associated with particular *geological-age* attributes. Low porosity carbonates with high kerogen (geochemical property) contents of the horizons also act *source rock* attribute. Certain reservoirs are entirely dependent on natural fractures for their productivities. Hydrocarbon pore volumes in reservoirs cannot be produced commercially unless there is connectivity among natural fracture systems (Figure 5.60), especially dense systems around the drilled wellbore. In order to plan and select drillable exploratory and development targets, it is necessary to optimally design the trajectories, completions and develop sustainable field development plans, with improved understanding of the natural fracture systems. The scope of the current study is to assess the application of directional/horizontal drilling (in new wells or side track of existing wells) and hydraulic fracturing, develop better understanding of reservoir fracture/matrix architectures (fracture storativity, connectivity, replenishment, flow capacity, intensity), and finally develop a fracture network model with predictive capabilities.

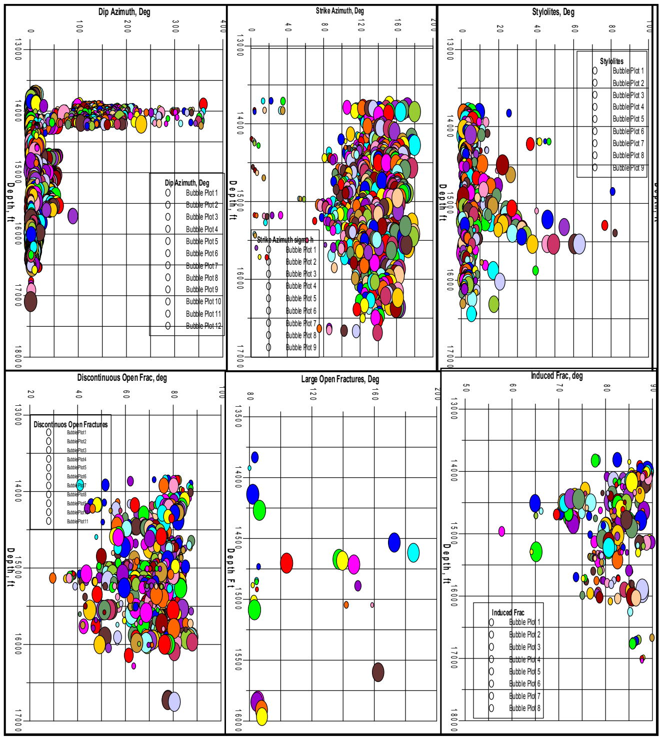


Figure 5.60: Comparison of different domain fracture systems

The author uses spectral amplitude and velocity anisotropy from 3D seismic datasets with known drilled-well information in assessing the fracture reservoirs. For this purpose, fracture image logs and core data are integrated with interpreted fracture systems, obtained from 3D seismic data cubes. Borehole breakouts are also considered in terms of their measured depths and orientations. Logs and 3D seismic data suggest wide-spread fracture porosity and permeability distributions in the study areas. Oil and gas production rates are dependent on the quality and distribution of fractures and their densities, which also significantly provide decline rates (because of reduced porosities and permeability of interpreted lithologies). Depth surfaces, gridded with faulted structures are attributable to interpret the compressional and extensional structure dimensions of the fracture reservoir systems. Reactivation attributes interpreted based on the geological age, are integrated with structure attributes, which ultimately made through an ontology connectivity to fracture reservoir systems. For example, Late Jurassic, Late Cretaceous and Tertiary aged *structures* and *reservoirs* are well connected through ecosystems and thus ontology based data warehousing and data mining approaches.

Fracture (reservoir) analysis

Fractures identified in the borehole wells are classified as natural and induced fractures. Natural fractures cut across the entire borehole, are traced as sine waves on borehole images. These fractures appear as darker than the surrounding rocks and contain drilling mud. Drilling induced fractures appear as dark (low amplitude) thin vertical lines and 180 degrees apart on images and as echelon chatter fractures at places. These fractures are produced during drilling.

Low amplitude fractures: these appear as dark sine waves on the image since they absorb more acoustic energy than the surrounding rock matrix. When the filling material is drilling mud, these fractures are open, but fractures sealed with clay can have the same signatures if the acoustic contrast between clay/formation is sufficient. Small size bubble clusters are noticed.

High amplitude fractures: these attributes appear as bright sine-waves since they absorb less acoustic energy than the surrounding rocks. In this case, the filling material is necessarily a material which is tighter than the matrix, usually quartz and or carbonate cements. Depending upon their density, sealed fractures can act as strong permeable barriers in the direction perpendicular to their strike. So based on the filling material, fractures appear to represent separate bubble clusters and their sizes.

5.4 Analysis of Petroleum Digital Ecosystems and Digital Oil Field Solutions

The author simulates a petroleum ecosystem to a data warehouse environment (RO2 and RO7), a system designed for archiving and analysing an organization's historical data, such as oil and gas exploration and production data, drilling data, including day-to-day operations. Normally, an organization summarizes and copies information from its operational systems to the data warehouse (Uschold 1998, Uschold and Gruninger 1996 and Wand et al. 1999) on a regular schedule, such as every night or every weekend; after that, management performs complex queries and analysis on the information without slowing down the operational systems. Data integration is key issue; combining data residing at different sources and providing the user with a unified view of these data. The data integration is an emerging process in a variety of situations both commercial (when two similar companies need to merge their databases) and scientific. Data integration process uses instances of different dimensions of exploration, drilling, production including navigational data. Relationships constructed in the conceptual modelling are ontologically analysed (Uschold and Gruninger 1996 and Wand et al. 1999). Integrated framework discussed in the Figures 3.36 – 3.38, is the basis for generating metadata and extracting data views for visualization and interpretation. *Horizon*, *structure*, *seismic time*, *velocity* and *depth* are used pointing to the navigational data. Besides, *frequency* and *amplitude* dimensions, as they are needed for logically connecting the other data dimensions for describing the sink-hole geometries. Preparation of a data warehouse to data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data (Ozkaharan 1990, Pujari 2002 and Uschold 1998) for patterns and deriving association rules. Though this is fairly recent a topic in computer science, but applies to many computational techniques from statistics, information retrieval, machine learning and pattern recognition.

The relationships constructed among conceptual models are semantically analyzed (Uschold and Gruninger 1996 and Wand et al. 1999). An integrated framework discussed in the Figures 3.36-3.38, is the basis for generating metadata and extracting data views for visualization and interpretation. *Horizon*, *structure*, *seismic time*, *velocity* and *depth* are used pointing to the navigational data. Besides, *frequency* and *amplitude* dimensions, as they are needed for logically connecting the other data dimensions for describing the sink-hole geometries, for example. Preparation of a data warehouse to Data Mining (DM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data (Ozkarahan 1990 and Uschold 1998) for patterns and deriving association rules. Though this is fairly a recent topic in computer science, but applies to many computational techniques from statistics, information retrieval, machine learning and pattern recognition. These recent advances enable development of digital ecosystems in petroleum industries. In an example of the Westralian Super Basin, multiple basins, multiple oil and gas fields exist within the knowledge domains of multiple petroleum ecosystems. Similar is the case with connecting onshore – offshore – transition zones of Romanian continental basin margins and Indonesian sedimentary basins through petroleum digital ecosystems.

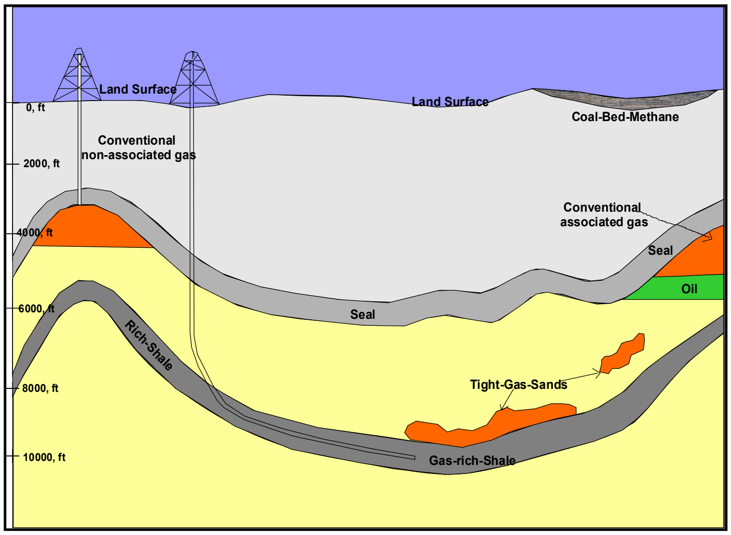


Figure 5.61: A multidimensional digital ecosystem view, showing conventional and unconventional resources

As an example, as demonstrated in Figure 5.61, the author is of the view both conventional and unconventional petroleum system elements and their connectivity including their processes, modelled and implemented, exist within in a single digital ecosystem. This concept equips the robust ontology based data warehousing approach, in which multidimensional data sources are accommodated. Integrated warehouse approach facilitates the connectivity among these systems. Fine-grained schemas for both conventional and unconventional systems constructed, simplifies the complexity of the digital ecosystem process.

Ontologically described and warehoused seismic data volumes

The author explores warehoused metadata for interpreting data views in terms of karstfication geology. Karst terrains are developed (Telford et al. 1990) in the areas underlain by carbonate rocks such as limestone rocks. They often have drainage systems that are reflected on the surface as sinkholes, springs, disappearing streams or even caves. The term "karst", therefore, refers to the terrain and the term sinkhole is one of the types of drainage features reflected by that type of terrain.

In the data warehouse modelling approach, the seismic data instances from hierarchical and relational structuring in both horizontal and vertical dimensions are gathered and intelligently stored in multiple dimensions, described in the domain ontology structuring methodologies. Several map views are drilled down from the warehouse metadata, representing seismic, drilled-well, known domains, structure and sink-hole interpretation, unknown, but built knowledge from the mined data. From different data mining representations of seismic data instances (in time and depth domain) as shown in Figures 5.62 and 5.63, demonstrate several geomorphic features in which sink holes are represented by structural lows. These structural lows, unlike structural anomalies, have unusual shapes and patterns. Amplitude maps and formation thickness maps (subtraction from formation top with formation base), narrate shallow drainage patterns and also strength of karstification. Karstification can visually be seen as described in the Figure 5.62, in which several mine-fields reported, are incorporated in the seismic grids (left image of Figure 5.63).

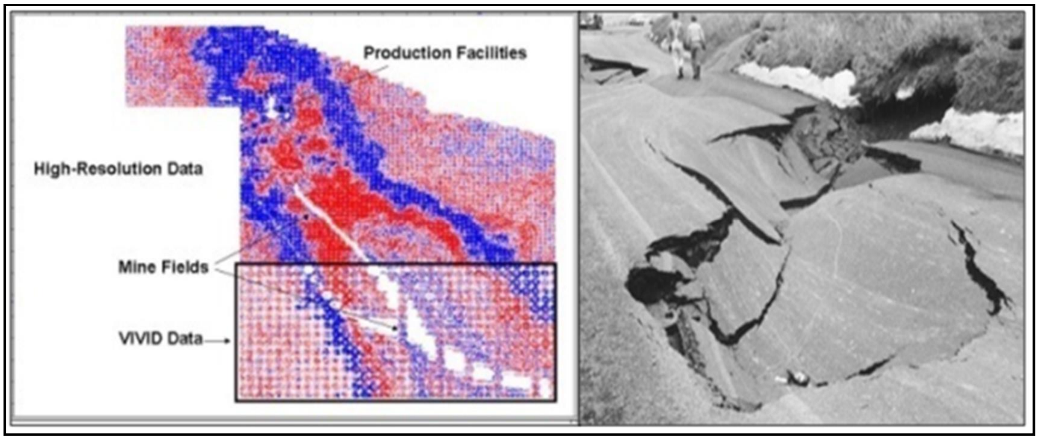


Figure 5.62: Seismic-time dimensions, showing data variations across mine-field survey areas – a scenario associated with an affected geomorphic system (right image, a sunken surface)

Near surface *geomorphic* features such as data acquisition layout *foot-prints*, *mine-fields*, production facilities are explicitly visualized in high resolution and sparse data grid areas as shown in Figure 5.63. To this extent, the near-surface that affected by sink-hole is distinct in the Figure 5.63.

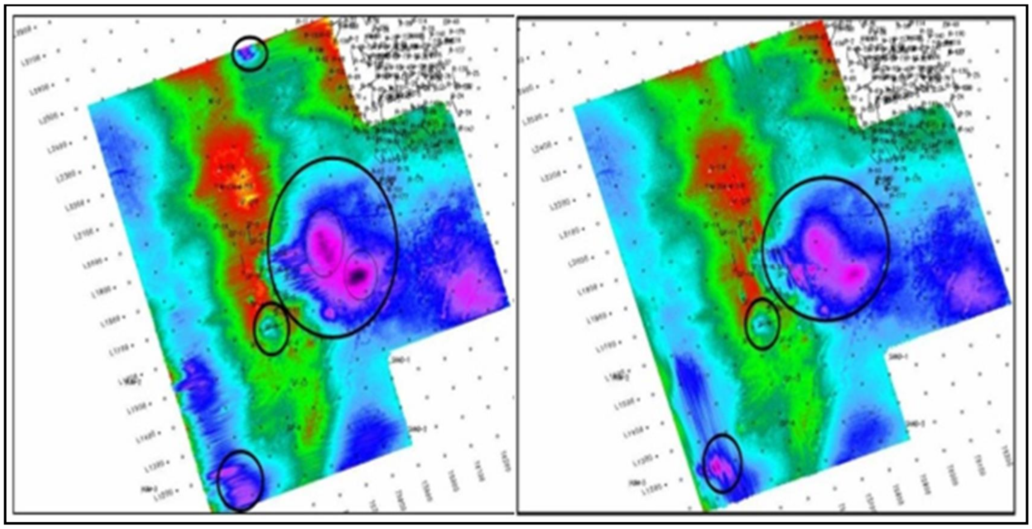


Figure 5.63: Map views of hierarchically organized data (in vertical, lateral horizontal directions) – sinkhole signatures (encircled are sunken areas) of geomorphic representation

There are many oil producing wells in the areas from carbonate reservoirs in the study area. As stated earlier, the carbonate rocks when associated with liquefied media cause chemical reactions to take place thus creating cavities, voids and even fractures within carbonate rocks. These anomalies must have definitely been deformed the rock properties which are represented as sinkholes as shown pinkish circular features in Figure 5.63. Sharp signatures due to sinkholes (encircled blue bodies) are representative of loss of seismic energy and dissipation of seismic amplitudes and frequencies seen in the form of circular features as shown in Figures 5.63 and 5.64. The author uses the framework described in Figure 3.37 (in Chapter 3), for constructing integrated metadata structures and interpreting them into multidimensional data structure map views as narrated in Figures 5.63 and 5.64.

The structural relationship is established based on the integrated interpretation of data dimensions and their instances extracted from seismic and well-base data sources. Structure attributes, represented as “structure-highs” (red color notation) and “structure-lows” (blue colored notation) as shown in Figures 5.63 – 5.64 are anomalous features, are interpreted as geomorphologic structure patterns. These sinkholes, which are seen in the seismic data as structural low anomalies manifest environmentally as geomorphic structure features shown in Figure 5.65.

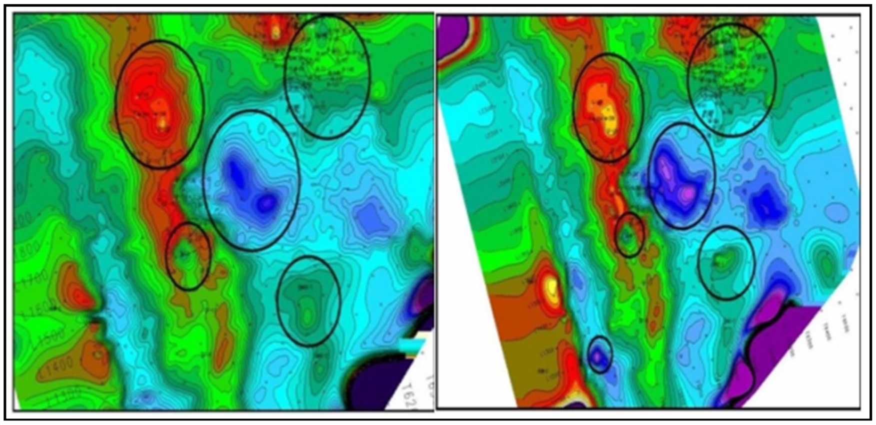


Figure 5.64: Map views of structure dimensions with "seismic time" attribute instances contoured over a survey of geomorphic representation (encircled) of sub-surface sink-hole patterns and trends

The real difference between actual depositional structural lows and sinkhole is that depositional structural low is broader and gently dipping towards depo-centre, whereas sinkhole structural low has more sharp edges with steep dips in the sinkholes. As shown in Figures 5.64 and 5.65, where one can physically interpret sharp edge with steep cut. Though the sinkholes described in the seismic images do not correspond to the sinkholes demonstrated in Figure 5.64, but ultimately, geomorphic features discussed in Figure 5.66, take shape of the image as shown in Figure 5.66. As examined in Figure 5.66, the sink-hole dimensions possess sharp edges, which perfectly show sharp seismic-structure dimension as interpreted in Figures 4.63 – 4.65.

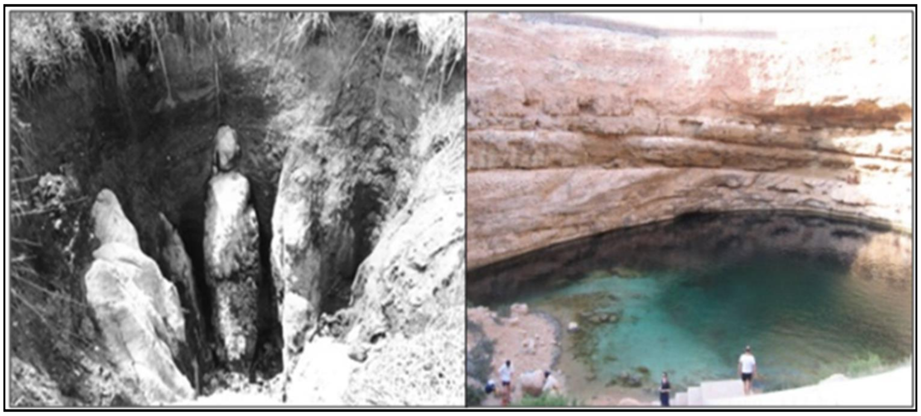


Figure 5.65: Surface sink holes affecting geomorphic system

As demonstrated in Figures 5.62 – 5.65, keeping in view possible hazards and environmental ramifications (Rolf 2005) for both exploration and exploitation of oil and gas around both surface and sub-surface sink holes; careful and judicious well planning is needed. Having established the connectivity among petroleum ecosystem, the geomorphic and ecological systems including sedimentary basin knowledge, future surveys and well campaigns are planned. Several sedimentary basins as described in Figure 2.1 (in Chapter 2) need careful understanding of the connectivity or embedded ecosystems for managing exploration and development. The conclusions and recommendations made, are useful for managers, well planners and explorers.

5.4.1 The Australian exploration business perspective

The author uses composite syntax and modelling, which play key roles in multidimensional modelling of petroleum industry's heterogeneous data in the Australian basins. Identification of composite attribute and its integration (Nimmagadda and Rudra 2004) are other issues addressed in the present study, a gateway for building the logical warehouse schemas. Exploring semantics from data sources of an up-stream and integrated oil and gas company is often more complex and tedious. An ontology (Jasper and Uschold 1999) approach facilitates integrating various facets of exploration and production data from which finer data are explored for patterns analysis. In the dimensional data structuring, multidimensional views of oil and gas exploration data extracted using OLAP server engine with OLAP operations (Jukic and Lang 2004 and Lowrie 1997) are interpreted for business knowledge. OLAP system provides specialized analysis tools as narrated in Figure 5.66 and 5.67.

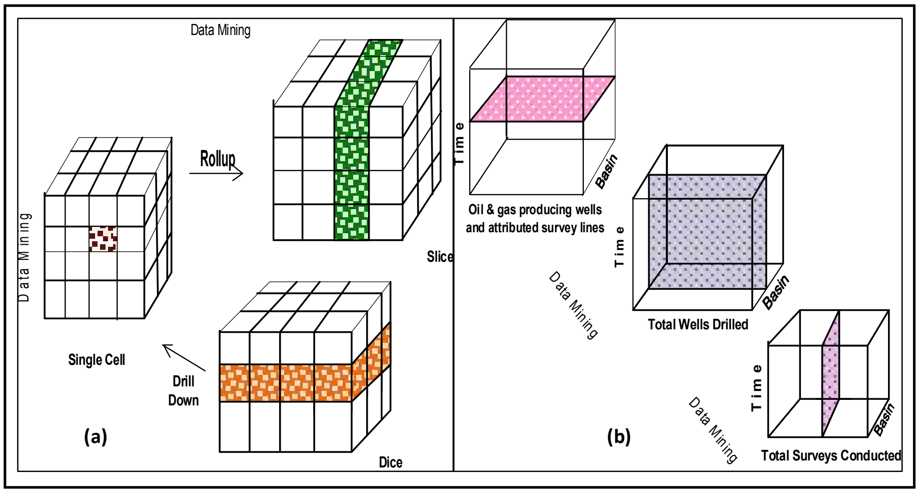


Figure 5.66: Multidimensional data operations

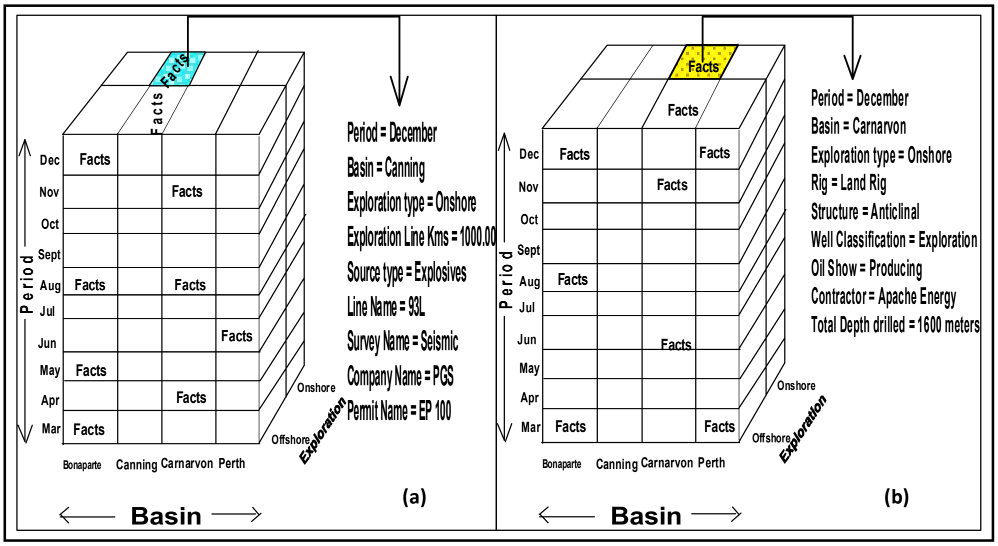
In multidimensional petroleum data models, there could be several numeric measures, for data analysis purposes. The data are viewed in several cubes, more precisely named as hyper-cubes (Nimmagadda and Dreher 2007). Each cube has three dimensions and each is further divided into several sub-dimensions.

Validating multidimensional schemas in the upstream oil and gas

Fundamentally, the metadata in the context of petroleum business data, is an integrated model that communicates through different dimensions such as wells, surveys, petroleum permits and oilplay factors (and or objects) intelligently. This process is practiced to achieve inter-operability or improved understanding of finer exploration of data carried out at later stages. The following criteria are adopted (Nimmagadda et al. 2005) for customizing the multidimensional modeling approach in the upstream:

Communication: An ambiguity is minimized by intelligent communication between dimensions (or entities/objects) and their attributes and relationships (properties) and linking them by robust logic.

Interoperability: The author uses models (data cubes), built based on multidimensional OLAP (Hoffer et al. 2005 and Nimmagadda and Dreher 2008). They are on different software platforms of an oil and gas industry, for data integration process and help in implementing them in a warehouse environment. It is achieved by linking and integrating different logical models built in different domains so that this process simplifies the complexity of petroleum business industry situations. The OLAP or SQL operations are carried out to extract piece of information without any distortion on other computing platforms.



Period = December
Basin = Canning
Exploration type = Onshore
Exploration Line Kms = 1000.00
Source type = Explosives
Line Name = 93L
Survey Name = Seismic
Company Name = PGS
Permit Name = EP 100

Period = December
Basin = Carnarvon
Exploration type = Onshore
Rig = Land Rig
Structure = Anticlinal
Well Classification = Exploration
Oil Show = Producing
Contractor = Apache Energy
Total Depth drilled = 1600 meters

Figure 5.67: Typical data views extracted from multidimensional data cubes

Multidimensional schema design benefits in a warehousing environment:

Reusability: Multidimensional logical structures built in different domain applications and or modeling languages can be reused. Even data structures used in one basin (in a basin domain) can be used in other basins elsewhere by simply changing the numerical data. Dimensions used in the process model remain unchanged.

Search: Ease of searching or viewing a piece of data from the warehoused petroleum metadata through data mining.

Reliability: Logical data structures built in knowledge domain makes the petroleum business data more reliable, because of intelligent communication (logical) between data structures. Resolving data conflicts (through ontology approach) prior to logical data design make the logical structural and implementation designs more reliable.

Specification: The multidimensional data mapping can assist the process of identifying future requirements and defining specifications for data warehousing.

Maintenance: Documentation of logical data model design is of immense help to the manager to reduce the operation and maintenance costs. This facilitates addressing the future needs of data structuring.

Knowledge acquisition: Speed and reliability for which logic has been developed for designing the data warehouse will facilitate the data-mining task much easier and faster in extracting business intelligence from petroleum business data.

Warehoused multidimensional data analysis

Once the data warehouse is ready with multidimensional data cube, then author explores the cube using different analytic tools with which complex analysis of petroleum data is performed. These analysis tools are called OLAP (a key term defined in the glossary in appendix). These OLAP tools are designed to accomplish such analysis on very large databases (VLDB) such as the current one.

In the multidimensional model, data are organized into multiple dimensions and each dimension contains multiple levels of abstraction. Such an organization provides the users with a flexibility to view data from different perspectives. Number of OLAP operations are carried out on data cubes, which allow interactive querying and analysis of data. According to the underlying multidimensional view with classification hierarchies defined for multiple dimensions, OLAP provides slicing, dicing, drilling (drilling-up, drilling-down, drill-within and drill-across) operations on multidimensional data cubes. An oil and gas data warehouse can only be effective when the data stored address issue of multidimensionality. Roles of multidimensionality and granularity in maintaining the data warehouse design and development have been emphasized in (Rudra and Nimmagadda 2005). In order to preserve data relationships with finer granularity, author ensures that data relationships are denormalized and the dimension is at its atomic level. The data warehouse project initiated for oil and gas business industry fails if the data structures in the data warehouse are poorly organized or with an inappropriate structure. If strict standards are not adopted, decisions based on this kind of oil and gas data in a warehouse are invalid and may lose credibility. Creating the oil and gas data for a data warehouse involves several issues. First, the data warehouse draws correlations from several data structures: operational data structures, updated operational data structures, already built-in data archives, data structures constructed from external data sources, and unstructured data. The primary source of data for the data warehouse is the oil and gas organization's operational systems and also informational systems.

Multidimensional schemas vs. business data-mining

The basic idea of this approach is to refine search of petroleum data from a warehouse and or information repository for desired data or information (e.g. well data of a particular field and a particular basin, oilplay analysis of a field, oil and gas data of other basins worldwide viewed as web documents, names of surveys conducted for spudding and drilling a well). OLAP engine presents the user a multidimensional view of the data warehouse as well as tools for operations. If the warehouse server organizes the data warehouse multidimensional arrays, then the implementation considerations of the OLAP engine are different from those when the server keeps the warehouse in a relational form.

The data mining of the oil and gas data is an important part of any data warehouse process model. The data from oil and gas data sources are mined externally using various statistical or other mathematical approaches. The author processes the warehoused oil and gas data for extracting correlations, trends and patterns from the oil and gas data. Author presents some of the explored data in Figures 5.68 – 5.72 and for interpreting any meaningful correlations and trends. The author examines mineral exploration cost patterns for QLD, WA, Rest of Australia and Northern Territory. These patterns possess similarity (see Figure 5.68) in their responses particularly in the years 1970, 1983 and 1990.

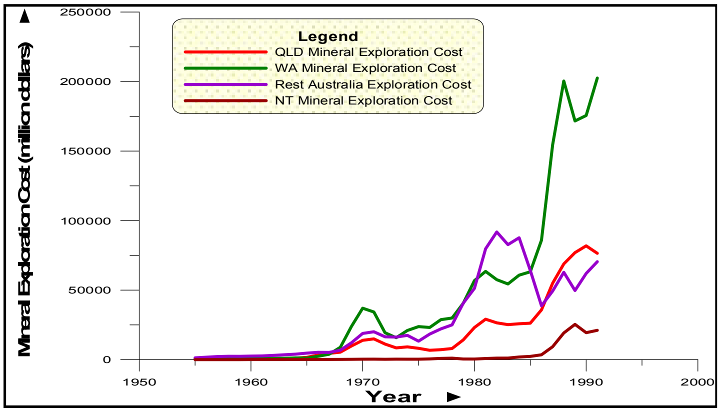


Figure 5.68: State wise mineral exploration cost data attributes (Australia)

Three peaks observed in these periods, indicate again more exploration activity in these states. Figure 5.69 is another means of presenting the quarterly mineral exploration costs data, which indicates more expenditure involved in the WA state's during 1996, because of more demand of the minerals and also higher prices of the precious mineralization.

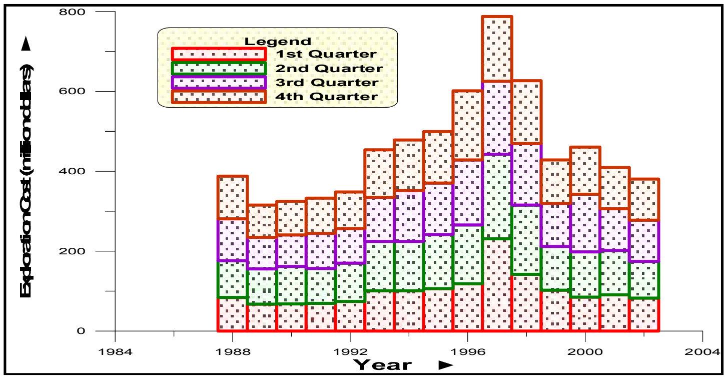


Figure 5.69: Bar chart analysis of WA State mineral exploration costs data attributes

The data acquisition, storage and access are key components of the data warehouse. The data capturing, exporting and importing from different sources have been smooth through the study, since the data are acquired from reliable sources and exported through standard different hardware and software platforms supporting the data warehousing. Oil and gas data warehouse is implemented as a global one to work across associated business units/subsidiaries and other functional areas such as survey field parties, drill sites, production platforms, assisting service providers (with limited access) and systems services. The oil and gas data, diverse in nature, provide synchronized move through different compatible hardware platforms, operating systems and database structures.

***Clustering of multiple oil and gas fields in the Western Australian basins**:* Volumes of data are available from different Westralian basins as described in Figure 281. *Period*, *number of surveys conducted*, and *number of drilled-wells* dimensions plotted on one scale, narrate interesting trends of clusters in circles and ellipsoids in the Canning and Carnarvon basins. The number of wells or surveys conducted in these basins is steady up to 1980 and shows a cluster of surveys conducted not to be related to the number of wells drilled. Clusters of surveys and wells drilled are highly variable during 1990; maximum wells drilled with corresponding number of surveys, suggesting that they relate each other and is interpreted to be that *wells-drilled* is dependent upon the *surveys* conducted. Size of bubbles of these clusters particularly in later years is decreasing. In the case of Carnarvon basin (Figure 2.1), the situation is quite different and the clusters of *number of surveys* conducted and *wells drilled* reveal an exponentially increasing trend over the period. In the case of Perth basin, as shown in Figure 2.1, surveys and number of wells drilled are decreasing over the period, but these clusters are in increasing order with decreasing periods. Figures 5.71-5.73 show that bigger sized bubbles are interpreted in a big cluster in the increasing periods inferring the number of surveys and wells drilled are correlatable in the sense that wells drilled in the Western Australian basins (Figures 2.1 and 5.70) depend on the number of surveys conducted. Encircled clustered bubbles suggest close proximity among attribute strengths. For example, with increasing *number of surveys* attribute there is corresponding increase in *number of drilled wells*. Similar conclusions are made among other bubble plots drawn among other attributed dimensions as shown in Figures 5.72 and 5.73. Another interesting interpretation is that where the red and green bubbles are clustered together, suggesting more wells drilled based on surveys data.

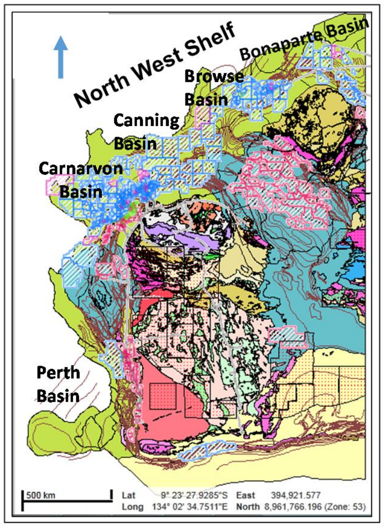

North West Shelf
Bonaparte Basin
Browse Basin
Canning Basin
Carnarvon Basin
Perth Basin
500 km
Lat 9° 23' 27.9285"S East 394,921.577
Long 134° 02' 34.7511"E North 8,961,766.196 (Zone: 53)

Figure 5.70: Western Australian map showing volumes of permits, surveys and other geological data sources

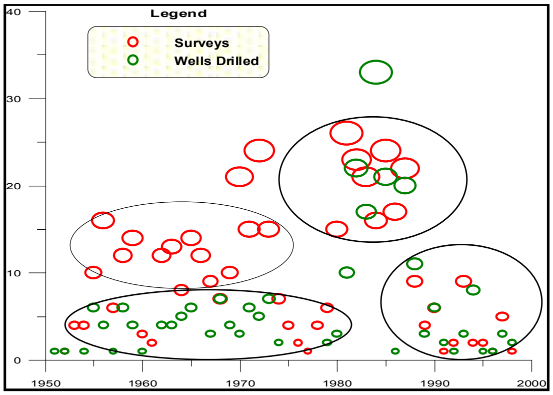


Figure 5.71: Data views extracted from warehouse, showing clusters of bubbles attributed by several *petroleum surveys* and *wells* (plot drawn between *period* vs *number of surveys* in Canning basin, WA)

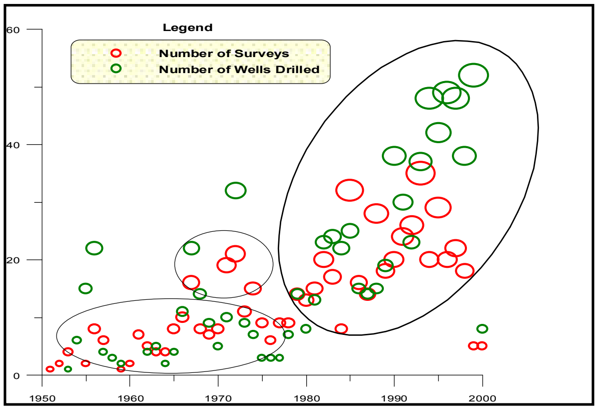


Figure 5.72: Data views extracted from warehouse, showing clusters of bubbles attributed by several petroleum surveys and wells (plot drawn between *period* vs *number of surveys*, Carnarvon basin, WA)

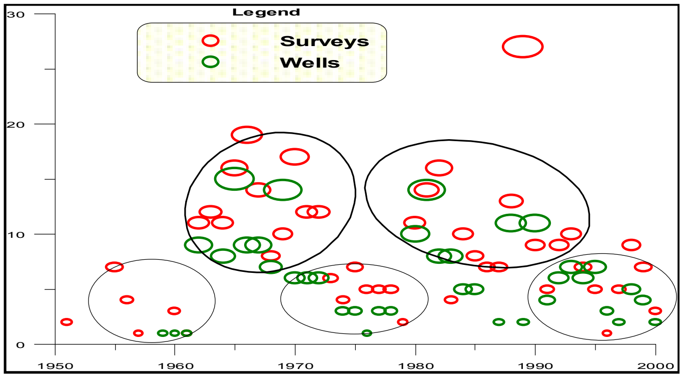


Figure 5.73a: Data views extracted from warehouse, showing clusters of bubbles attributed by several petroleum surveys and wells (plot drawn between *period* vs *number of surveys* in Perth basin, WA)

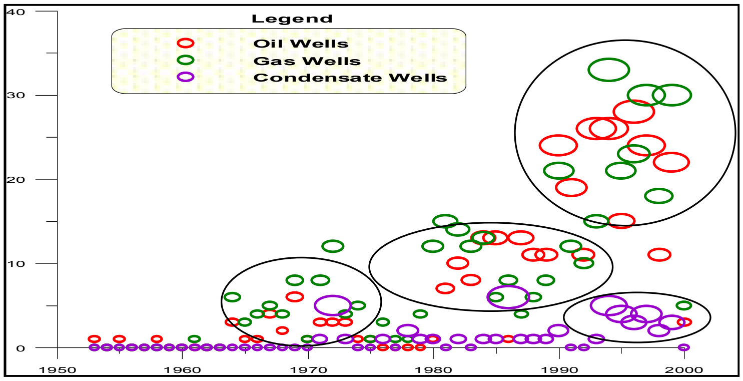


Figure 5.73b: Data views extracted from warehouse, showing clusters of bubbles attributed by several petroleum surveys and wells (plot drawn between *period* vs *number of surveys* in Perth basin, WA)

As shown in Figure 5.72a, clusters gathered among number of surveys and wells attributes appear to be matching each other in the increasing periods, interpreted to be dependent on each other. As clusters bounded by circles are exponentially increasing with increasing bubble sizes and with period as well. This is interpreted as all the producing wells perfectly match all the drilled wells in all key basins of Western Australia. Wherever oil wells are drilled quite a few gas wells are also discovered, showing multiple reservoirs (Figure 5.72b), providing good relationship between oil and gas producing wells. A similar situation is observed (Figure 5.73a) between *number of structures* interpreted and *number of hydrocarbon producing wells* as encircled with bubbles and their densities. As shown in Figure 5.73b, when clusters associated with *number of oil, gas and condensate wells* attributes plotted together (Figure 5.73c and Figure 5.73d) with *total number of wells drilled*, it is interesting to observe that cluster associated with number of drilled wells attribute is separated, thereby leaving an impression of cluster associated with *number of drilled wells* attribute has least relationship with producing wells data attribute clusters.

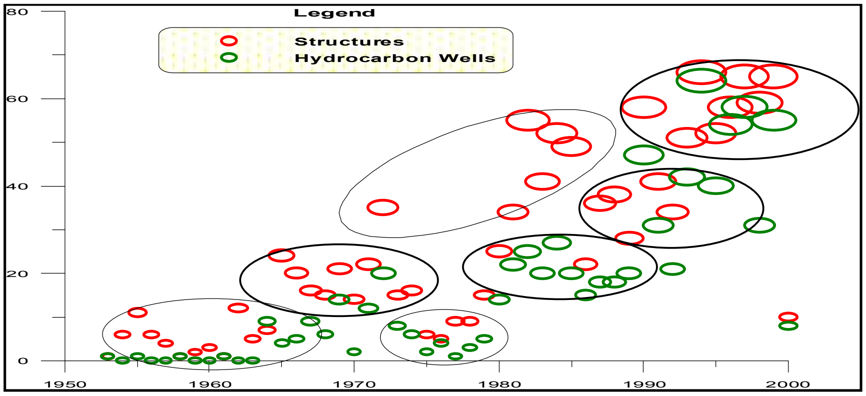


Figure 5.73c: Data views extracted from warehouse, showing clusters of bubbles attributed by several structures (geological) and hydrocarbon wells (plot drawn between *period* vs *number of wells,* Perth basin, WA)

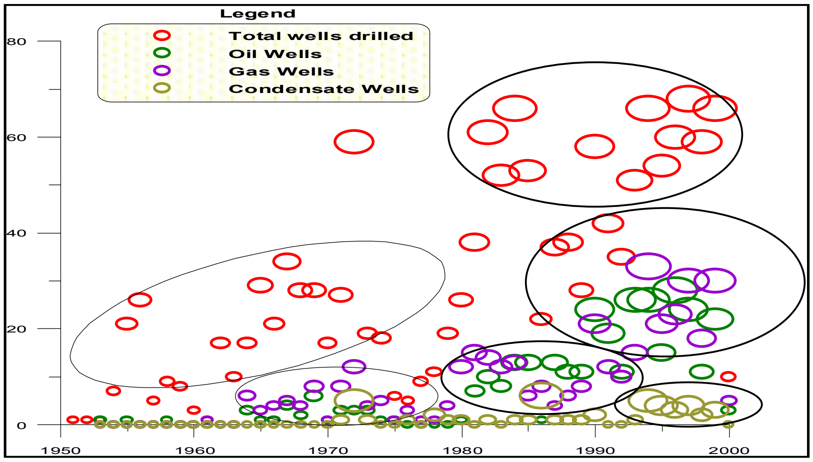


Figure 5.73d: Data views extracted from warehouse, showing clusters of bubbles attributed by several drilled wells (plot drawn between *period* vs *number of wells* in Perth basin, WA)

Interpretation of similar and dissimilar attributes (as per research objective, RO 5 in Section 1.3.1 in Chapter 1)

Multiple dimensions of exploration are typically, basin, reservoir, structure, drilled well, surveys conducted and oil/gas produced in each field and basin. Clustering algorithms attempt to find natural groups from exploration and production data instances based on certain similarity. These algorithms find the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distances between a point and the cluster centroids. The output from clustering algorithm is a statistical description of the cluster centroids with number of data in each cluster. The distance between two points is taken as a common metric to assess the similarity among datasets of volume of exploration data. In the true representation of clusters for interpretation, author considered several data instances from the warehoused metadata, populated in several spreadsheets. These data instances are plotted in the form of bubbles, representing clusters in 2D plots (Figures 5.74a – 5.74e). Bubbles of clusters display two variables on a scatter-type plot. In the present study, author uses bubbles for narrating and investigating clusters, for varying data dimensions. In a 2D bubble plot, the diameter of each bubble can vary in size, providing a way to represent an additional dimension of data. For example, consider a traditional scatter plot that shows the number horizons/reservoirs producing in the Perth, Carnarvon, Cambay basins or Middle-eastern basins over a period of time. Using a 2D bubble plot, a third dimension of data is displayed, showing average production or reservoir type over the same time span or entire basin. Different worksheets are used for changing the data to plot different bubbles. X, Y, Z coordinates are changed to get appearance of different bubbles with differing sizes and locations. X, Y, Z and size data columns of worksheets and their corresponding attributes are plotted in the same scale and axes so explore similarity or dissimilarity properties. Encircled bubble plots show close proximity among several attribute dimensions as shown in Figure 5.74. Change of axes is also performed on different plots, to check the data validity and integrity. Point dimensions attributed in the data trends are also counted, so that strength of attribute property is explored.

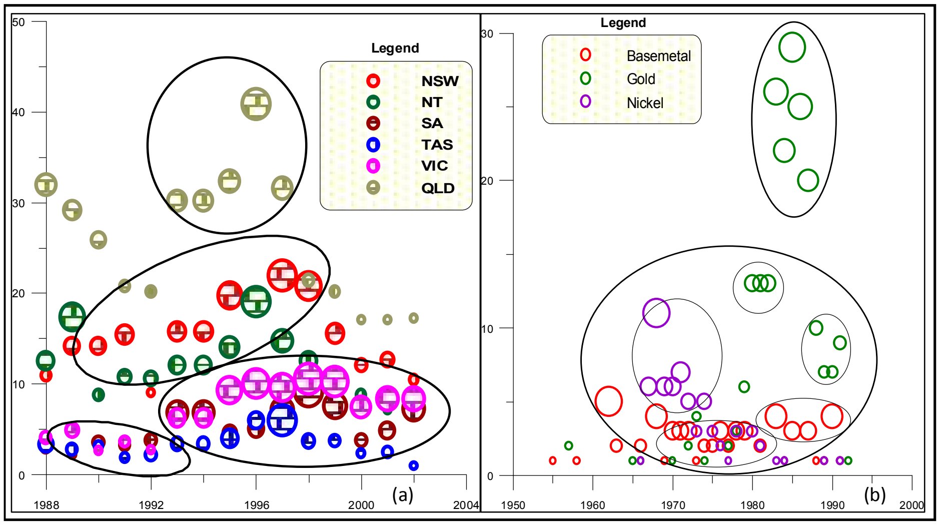


Figure 5.74a: Cluster of similar values (based on the direction and color) of states of Australia (plot drawn between *period* and *number of reservoirs*)

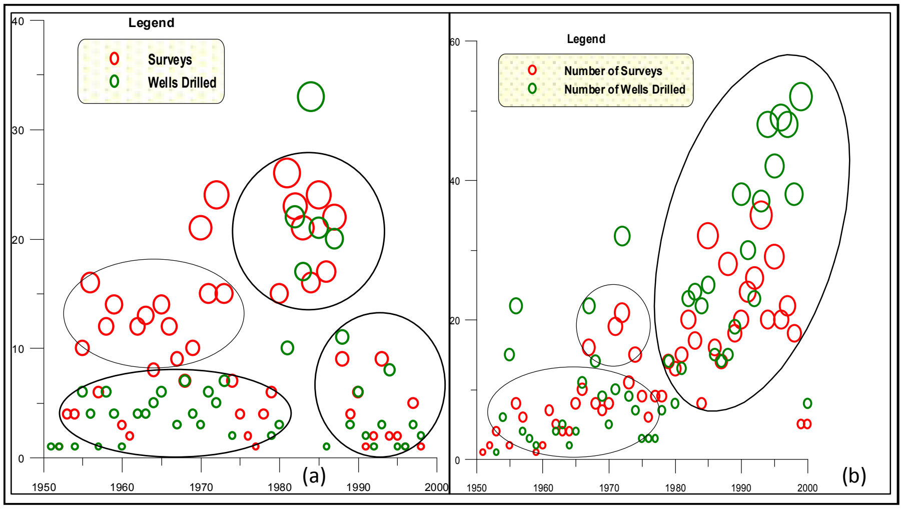


Figure 5.74b: Cluster of similar values (based on the direction and color) of Australian oil and gas (plot drawn between *period* and *number of surveys*)

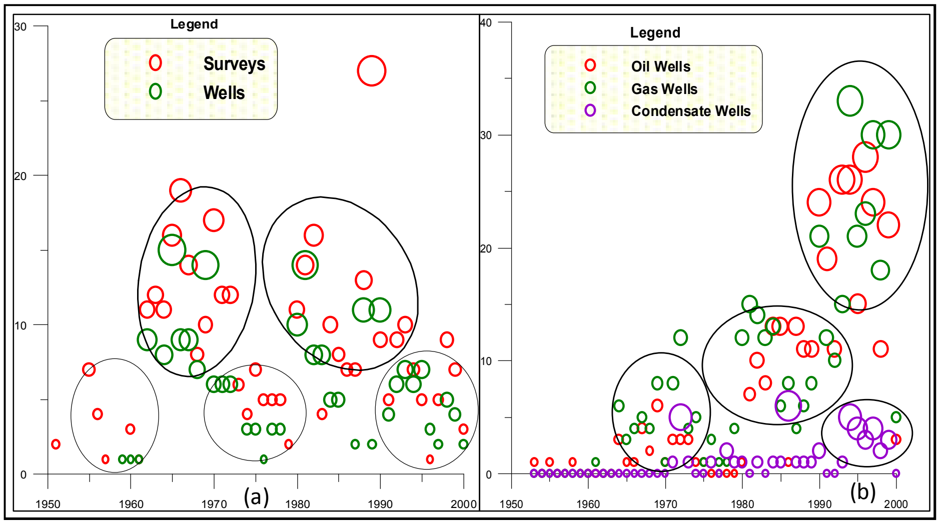


Figure 5.74c: Cluster of similar values (based on the direction and color) of Australian oil and gas (plot drawn between *period* vs *number of wells)*

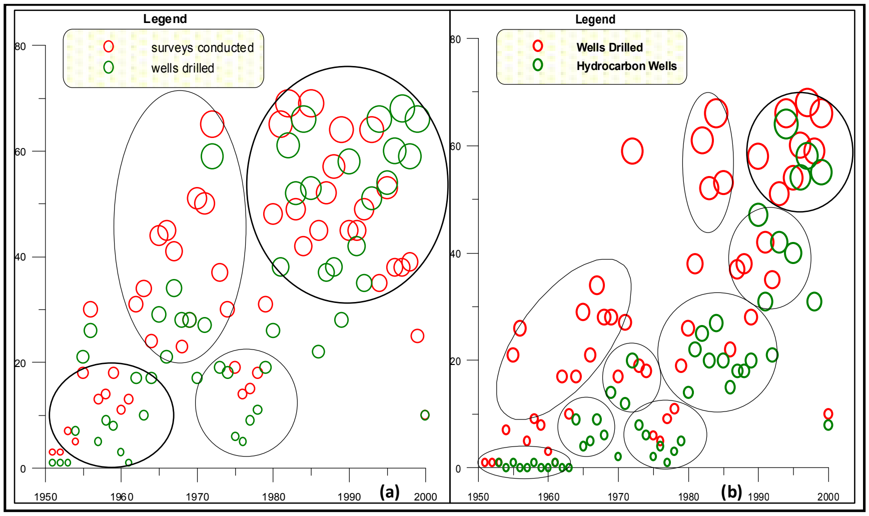


Figure 5.74d: Cluster of similar values (based on the direction and color) of Australian oil and gas (plot drawn between *period* vs *number of wells)*

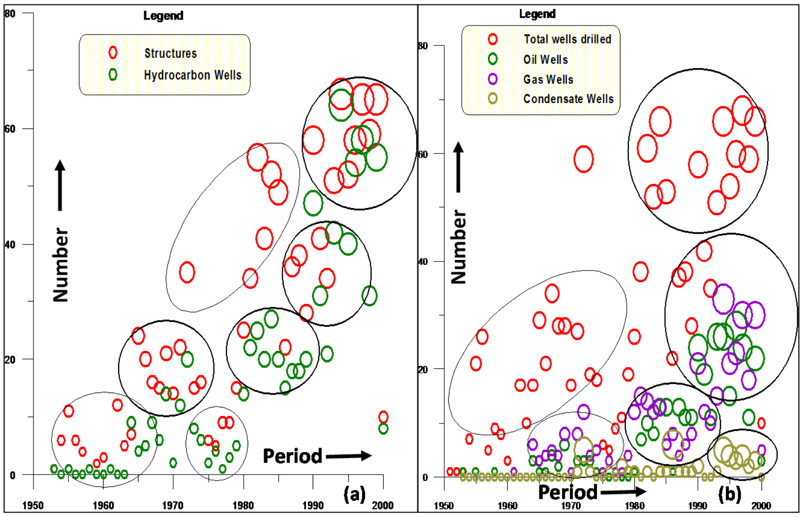


Figure 5.74e: Cluster of similar values (based on the direction and color attributes) of Australian oil and gas data

5.4.2 The Indonesian exploration business perspective

Big data analysis

The secondary data (Wight et al. 1992, Nimmagadda and Dreher 2011, Nimmagadda and Dreher 2012a and Nimmagadda et al. 2012b) sources published by the Indonesian Petroleum Association (IPA) are used for testing the data models and analyse them in the current study. These secondary data sources possess big data features “volume, variability, velocity, veracity, visualization and value”, meant to be analyzed in Indonesian sedimentary basins. As described in Courteney et al. (1991), Noon et al.

(2003) and Biantoro et al. (1996), volumes of big-data published, made use of modelling and analysis. Petroleum Multidimensional star schema definitions are integrated in a single metadata schema of Tarakan (name of a basin in Indonesia) petroleum system warehouse. This type of multidimensional data warehouse respects and maintains the integrity of individual datasets, preserving information about the source of warehouse entries, so that users know and determine exactly, which depositional systems are part of particular set of petroleum systems. Different data views are provided about different sub-basins of Tarakan basin in similar and dissimilar depositional environments of different petroleum systems (as views can be extracted from different data cubes, Figure 5.75). Ontology based multidimensional approach removes the semantic heterogeneity among petroleum systems and basins, removing redundancy among different datasets. This approach also allows detecting the non-redundant petroleum system DBs, for which different algorithms are written for mining purposes. Non-redundant petroleum system DB may be created as a separate schema and DB within warehouse, which can apply to a specific mining algorithmic application or a specific user query. This could satisfy the users, who want to study large size non-redundant petroleum systems' data sources. Similar big-data are available as described in Figures 2.1 and 2.2 in the Western Australian sedimentary basins.

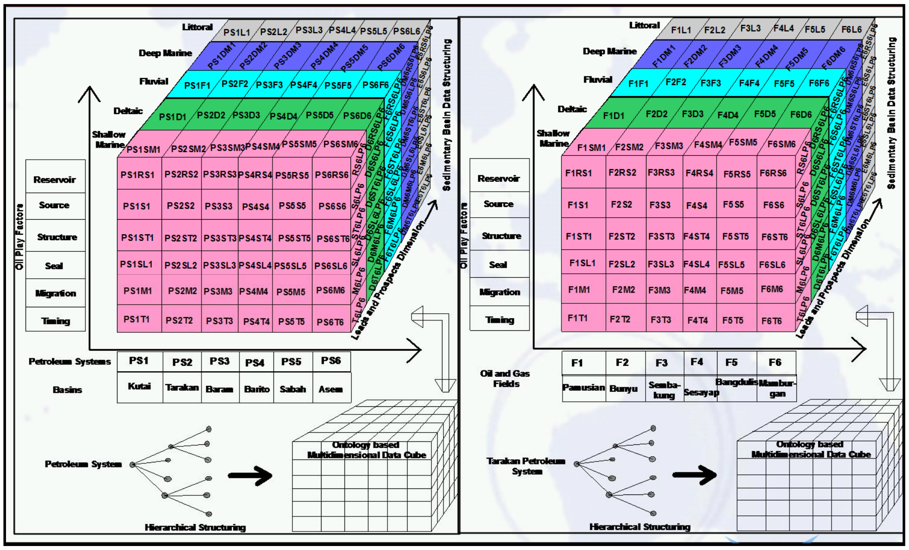

Littoral
Deep Marine
Fluvial
Deltaic
Shallow Marine
PS1L1 PS2L2 PS3L3 PS4L4 PS5L5 PS6L6
PS1DM1 PS2DM2 PS3DM3 PS4DM4 PS5DM5 PS6DM6
PS1F1 PS2F2 PS3F3 PS4F4 PS5F5 PS6F6
PS1D1 PS2D2 PS3D3 PS4D4 PS5D5 PS6D6
PS1SM1 PS2SM2 PS3SM3 PS4SM4 PS5SM5 PS6SM6
Reservoir
Source
Structure
Seal
Migration
Timing
Oil Play Factors
PS1RS1 PS2RS2 PS3RS3 PS4RS4 PS5RS5 PS6RS6
PS1S1 PS2S2 PS3S3 PS4S4 PS5S5 PS6S6
PS1ST1 PS2ST2 PS3ST3 PS4ST4 PS5ST5 PS6ST6
PS1SL1 PS2SL2 PS3SL3 PS4SL4 PS5SL5 PS6SL6
PS1M1 PS2M2 PS3M3 PS4M4 PS5M5 PS6M6
PS1T1 PS2T2 PS3T3 PS4T4 PS5T5 PS6T6
Sedimentary Basin Data Structuring
Leads and Prospects Dimension
Petroleum Systems
PS1 PS2 PS3 PS4 PS5 PS6
Basins
Kutai Tarakan Baram Barito Sabah Asem
Petroleum System
Hierarchical Structuring
Ontology based Multidimensional Data Cube
Littoral
Deep Marine
Fluvial
Deltaic
Shallow Marine
F1L1 F2L2 F3L3 F4L4 F5L5 F6L6
F1DM1 F2DM2 F3DM3 F4DM4 F5DM5 F6DM6
F1F1 F2F2 F3F3 F4F4 F5F5 F6F6
F1D1 F2D2 F3D3 F4D4 F5D5 F6D6
F1SM1 F2SM2 F3SM3 F4SM4 F5SM5 F6SM6
Reservoir
Source
Structure
Seal
Migration
Timing
Oil Play Factors
F1RS1 F2RS2 F3RS3 F4RS4 F5RS5 F6RS6
F1S1 F2S2 F3S3 F4S4 F5S5 F6S6
F1ST1 F2ST2 F3ST3 F4ST4 F5ST5 F6ST6
F1SL1 F2SL2 F3SL3 F4SL4 F5SL5 F6SL6
F1M1 F2M2 F3M3 F4M4 F5M5 F6M6
F1T1 F2T2 F3T3 F4T4 F5T5 F6T6
Sedimentary Basin Data Structuring
Leads and Prospects Dimension
Oil and Gas Fields
F1 F2 F3 F4 F5 F6
Pamusian Bunyu Semba-kung Sesayap Bangdulis Mambur-gan
Tarakan Petroleum System
Hierarchical Structuring
Ontology based Multidimensional Data Cube

Figure 5.75: Multidimensional representation of data cubes for multiple oil and gas fields and sedimentary basins of Indonesia

Schema implementation and data mining in Indonesian basin scenarios

The petroleum systems' warehouse supports several petroleum ecosystem informatics for different data types, each of which is implemented as one or more tables in the schemas. Dimensional and factual data tables are not shown in this work, for privacy and protecting the intellectual property rights. Every dimension in the warehouse has multiple rows in the entry table that defines the metadata such as the time it is inserted in the warehouse, and its time of last update. Every warehouse entity is also associated with the dataset and DB from which domain, it was derived. Different data views are drawn from a metadata of a warehouse for interpreting trends and correlations (Shastri and Dreher 2011a and Wand 2000) that make up the connectivity and communication among relevant data attributes of multiple dimensions. Using grapher and surfer solutions, bubble plots and surface map views are drawn and plotted as shown in Figures 5.76 – 5.80. The data relationships plotted between period and other conceptualized (logical) data attributes are deduced with interesting connectivity among data properties. *Structure*, *reservoir* (in geological domain) and production data attributes among different petroleum systems and among different depositional regimes, provide valuable data correlations and trends among periodically and geographically distributed petroleum systems. Indonesia has geographically distributed basins with productive petroleum provinces. Big data features of these data sources support correlatable trends of system's elements. As visualized in Figures 5.76 – 5.80, several data views drawn between oil and gas data wells, interesting trends are interpreted with new knowledge.

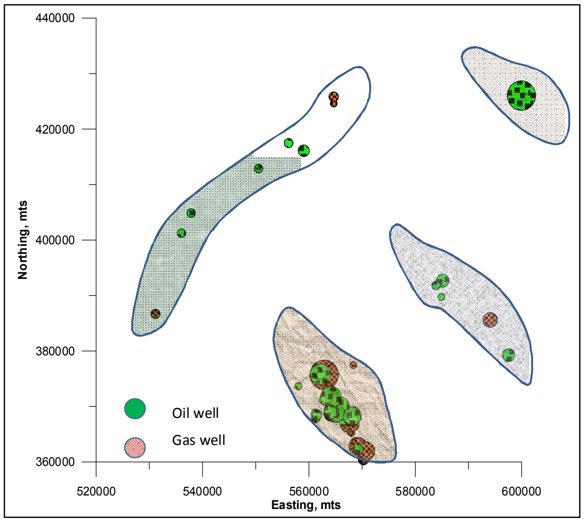


Figure 5.76: Data view showing production attribute trends of oil and gas drilled wells in a Tarakan (name of basin) ecosystem

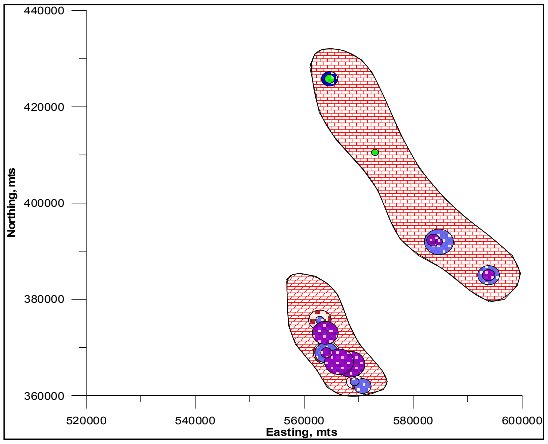


Figure 5.77: Data view showing multidimensional trends of *porosity*, *permeability*, *cumulative production*, *API* and *methane* (%) attribute strength trends and orientations

In a 2D bubble plot, as shown in Figure 5.78, the author draws several multidimensional data views to understand varying instances of multidimensional attributes and their strengths. In the bubble plot, the diameter of each bubble varies in size, providing a way to represent an additional dimension of the data. For example, as demonstrated in Figure 5.78, the scatter plot of *drilled-depths* with *salinity of water formations* at different stratigraphic intervals and geographic locations (inside basin boundaries) suggests correlations and trends among formations within an ecosystem scenario.

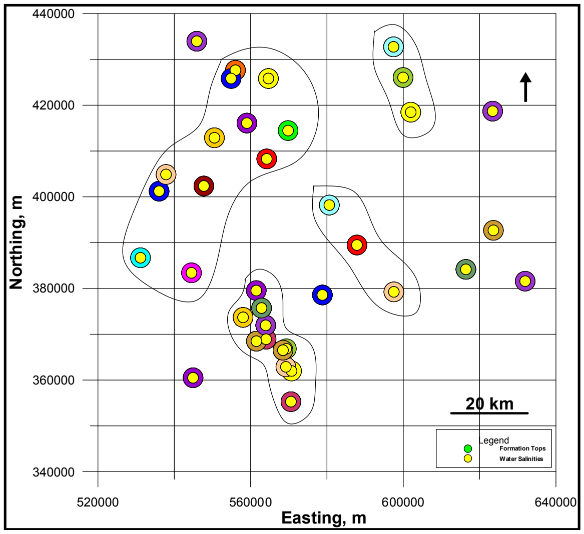


Figure 5.78: Data relationships, correlations and trends of water salinities within a Tarakan digital ecosystem

As stated earlier, for the Tarakan basin, several multidimensional data attributes and their strengths given in Courteney et al. (1991), Noon et al. (2003) and Biantoro et al. (1996), are ontologically described for designing and developing Petroleum Digital Ecosystem (PDE). As an example, in the current paper, PDE of the Tarakan basin is demonstrated. Similar PDE may be designed for other sedimentary basins in the Indonesian basins. These studies can further be extended in the South East Asia (Figure 2.1, Chapter 2). In Southeast Asia, Total Petroleum Systems (TPS) scenarios, have hierarchies in multiple sedimentary basins, "each basin has multiple oil and gas fields, each field has multiple oil/gas drilled wells, each well has multiple oil/gas net-pays". Each well has encountered different geological formations and each formation has multiple reservoir characteristics. Hierarchically, in geological domain, all these multidimensional data are ecologically (inherent) interconnected and being interacted (evolution) continuously in multiple geological ages. In global scenario, hierarchically organized multidimensional data (ontologically structured) in geological domain (in sub-surface, i.e., depth, for example) are connected to surface-geophysical data (user defined) through process of data integration, named data warehousing. The author extracts data views from warehoused systems for interpreting geological knowledge. Geological *ages* and navigational data coordinates are significant dimensions that are responsible connecting geological and geophysical domains. These phenomena are demonstrated in Figures 5.79 and 5.80 in 3D plot views.

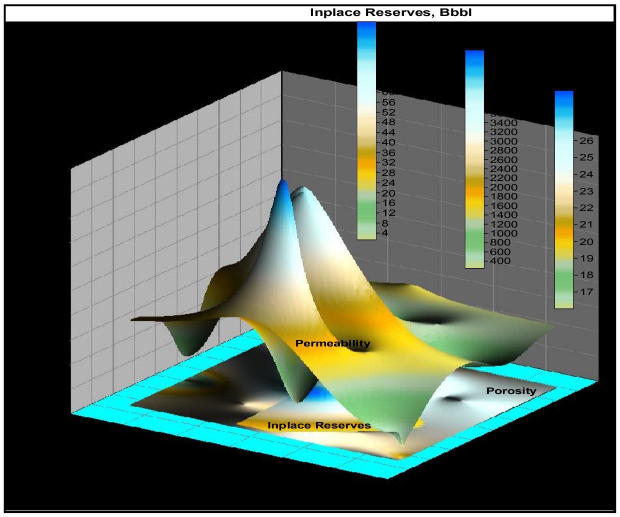


Figure 5.79: 3D surface map view, showing global production from deltaic reservoir ecosystem

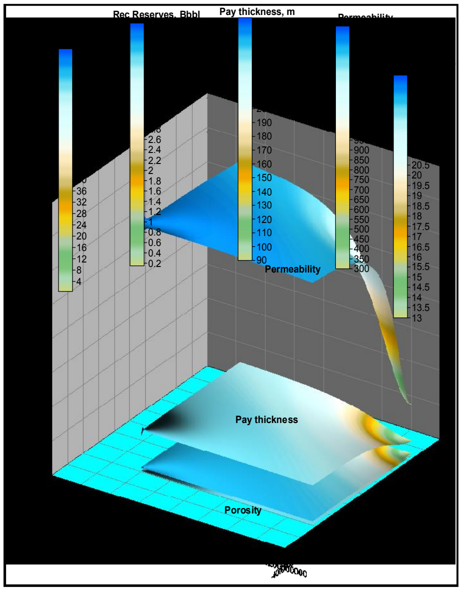


Figure 5.80: 3D Surface map view of global production within a constrained reservoir ecosystem

Petroleum data management and knowledge discovery

The field of geo-ontologies plays an increasing role in the study of fundamental geological problems owing to the exponential explosion of sequence and structural information with time (Nimmagadda and Dreher 2010). There are two major challenging areas in geo-informatics and exploration studies: (1) data management and (2) knowledge discovery. After the completion of the geo-ontologies project (Nimmagadda and Dreher 2011), a new, pre and post-exploration era, is beginning to analyze and interpret the huge amount of geo-scientific information. The geo-ontologies would have not been completed in advance without collaborative efforts carried out at research centers through hubs and networks. After completion of geo-ontologies, professionals face different problems related to topics such as discovering petroleum systems knowledge in terms of interactions and processes (such as migration pathways and timing of charging the reservoirs along geological structures) through elements. These processes need intensive computations and new informatics approaches. The greatest challenge facing the geo-ontologies, understanding petroleum systems at basin to prospect scale (Magoon and Dow 1994), and indeed all of geology, is on the capture, management, and analysis of the torrent of new data and information flooding over us in the next few decades. Many different types of data must be integrated into databases and metadata structures that will range well into hundreds of terabyte scale—presenting huge analysis, data mining, and visualization challenges. According to a survey, there were at least hundreds of data sources in the last century and the number climbed to several thousands in the current century. The most popular sources are from the AAPG, SEG, SPE, and National oil companies (NOC) and bigger producing private companies, such as Exxon-Mobile, Shell. Good sources are from the service companies, such as Schlumberger and Halliburton. The increasing volume and diversity of digital information related to geo-informatics and exploration have led to a growing problem that computational data management systems do not have, namely finding which information sources out of many candidate choices is more relevant and most accessible to answer a given user query. Moreover, there is rich domain information on how results for queries should be obtained and strong user preference for sources (for instance, one geoscientist may prefer to keep structure or reservoir information due to their connectivity to petroleum systems at basin scale (global scale). The completeness (but not correctness) of results is negotiable in favor performance and timeliness.

A real challenge to data management is managing and integrating the existing exploration databases. However, in some situations, a single database cannot provide answers to the complex problems of petroleum systems that bother the geoscientists and explorers. Integrating or assembling information from several databases to solve problems and discovering new knowledge, are other major challenges in geo-informatics. The transformation of voluminous exploration and geo-scientific data into useful information and valuable knowledge is the challenge of knowledge discovery. Identification and interpretation of interesting patterns hidden in trillions of bytes of seismic and well data is a critical goal of geo-informatics. This goal covers identification of useful reservoir and geological structures from sequences, derivation of diagnostic knowledge from experimental data and extraction of scientific information from the literature. The existing research in geo-informatics is related to knowledge discovery, event analysis, and structure analysis. Event analysis is the discovery of functional and structural similarities and differences between multiple geological sequences. This can be done by comparing the new (unknown) sequence with well-studied and annotated (known) sequences. Geo-scientists find two similar sequences, possessing the petroleum system's characteristics including migratory pathway, and geological structure. If two similar sequences are from different fields or two different petroleum systems, they are said to be homologous sequences. Finding homologous sequences is important in predicting the nature or type of reservoir. Structure analysis is the study of individual or two different geological formations (or corroborative seismic sequences) and their interactions/communications. Reservoirs are complex in their distributions depending upon the structural setting. Each reservoir and its deposition are unique at unique geological settings. The structures of reservoir facies are hierarchical and consist of primary, secondary, and tertiary structures. In other words, at molecular level, reservoirs may be viewed as 3D structures. The understanding of reservoir geometries leads to new understanding for diagnosis and health of exploration or even producing matured fields. The current research analyzes reservoir characterizations using relational domain ontologies such as, comparison and prediction of reservoir geometries and their associated structures/trap anomaly attributes. Due to the complexity and gigantic volume of geological and geophysical data, the traditional computer science techniques and algorithms fail to solve the complex geological problems and discover new knowledge discover knowledge from multiple petroleum systems or total petroleum system. As narrated in Figure 5.81, multiple volumes are created in the form of cubes from warehoused metadata for further analysis and interpretation.

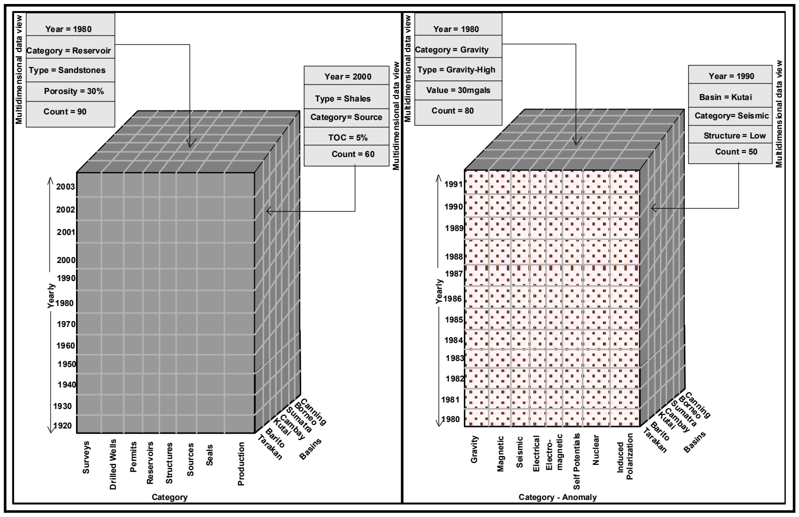


Figure 5.81: Multidimensional data cubes of *anomalies* attributes

Further, data views are extracted for different scalable maps of attributes, *seismic structure*, *and reservoir* and *production* data. The map views of these different attributes are superimposed (Figure 4.53) for understanding the concept of an integration and also relationships among geophysical property anomalies, both in both exploration and field development scenarios.

Computation of data views and discussion of results

Faulted structures are *displacement* attributes and they are in the order of several feet and hundreds of feet in steep dipping salt domes, usually interpreted from seismic, gravity and magnetic datasets (as envisaged from integrated seismic information, Figure 5.82). *Weathering* and *erosion* events associated with these faulted structures, are prominent, in which case attributes such as, *the vertical extent* and *dip of fault* at depths, are unknown or known that create an ambiguity usually with a single geophysical property and anomaly.

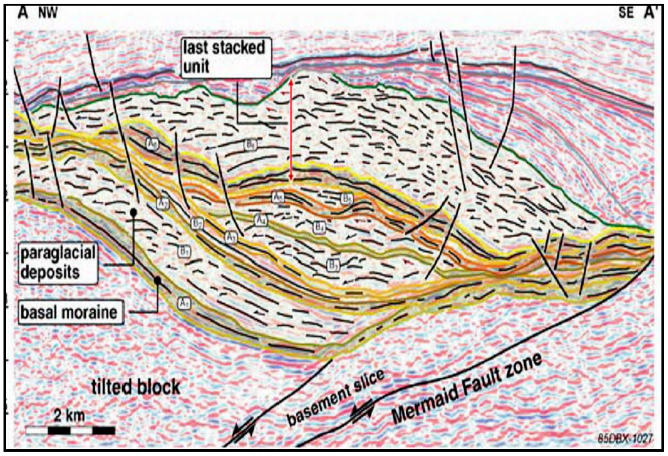


Figure 5.82: An interpreted section view showing *seismic structure* anomalies

Integrated geophysical anomalies may better reveal the nature of the fault at that depth. Electrical resistivity, ground penetrating radar, seismic refraction and reflection, gravity and magnetic surveys conducted across complex geological anomaly attributes may bring out the attitudes of the structures. Data views extracted and illustrated in Figure 5.83, suggest several interesting trends among multiple attributes and anomalies, with envisaged reservoir models. The production data when integrated with structure and isopach (thickness patterns) anomalies (Figures 4.53, 4.54 and 5.83) provide characteristic correlations that can describe prospective oilplays.

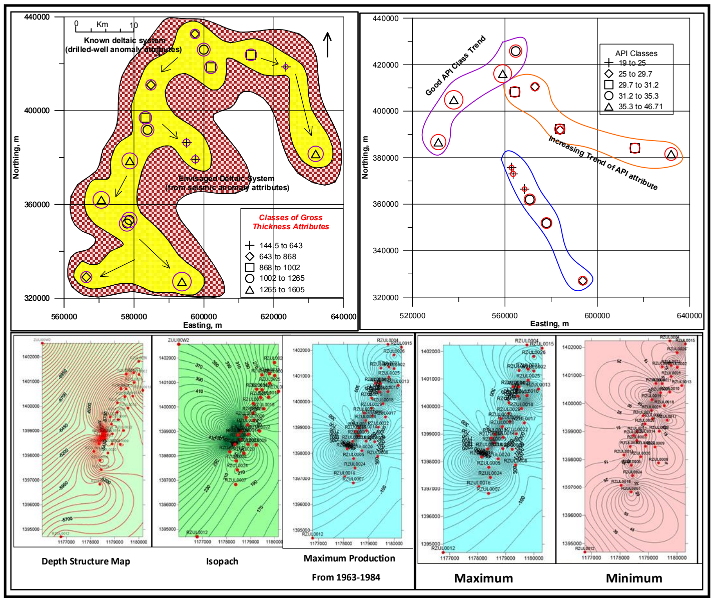
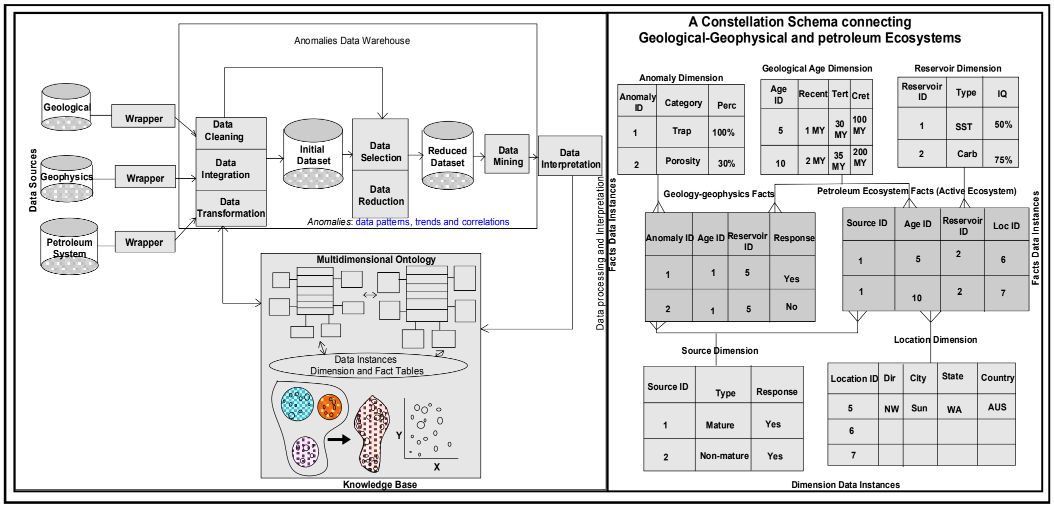


Figure 5.83: Data views extracted from warehouse and data mining models

Steep gradients of gravity and magnetic anomalies often suggest the orientation of a fault including fault zones associated with salt domes. The seismic data alone may not provide solutions on high resolution anomaly attributes such as, dip and strike of sedimentary beds. There is currently no official classification of structures by a single attribute. The use of multiple geophysical methods is aimed at the reduction of ambiguity within the entire data set for describing structures. As shown in Figure 5.84a, a knowledge-base model is built in which all the geophysical anomalies including anomalies of elements of petroleum systems and integrated to compute data views for knowledge mapping purposes. The author retrieves structure anomalies such as *normal* and *reverse* fault documented as *location* attributes (Figure 5.82) along with other anomalies, such as *Bouguer gravity* (or residual) and *magnetic* intensities (Parasnis 1997).



Anomaly Dimension

Anomaly ID	Category	Perc
1	Trap	100%
2	Porosity	30%

Geological Age Dimension

Age ID	Recent	Tert	Cret
5	1 MY	30 MY	100 MY
10	2 MY	35 MY	200 MY

Reservoir Dimension

Reservoir ID	Type	IQ
1	SST	50%
2	Carb	75%

Geology-geophysics Facts

Anomaly ID	Age ID	Reservoir ID	Response
1	1	5	Yes
2	1	5	No

Petroleum Ecosystem Facts (Active Ecosystem)

Source ID	Age ID	Reservoir ID	Loc ID
1	5	2	6
1	10	2	7

Source Dimension

Source ID	Type	Response
1	Mature	Yes
2	Non-mature	Yes

Location Dimension

Location ID	Dir	City	State	Country
5	NW	Sun	WA	AUS
6				
7				

Figure 5.84a: Knowledge discovery process model, showing views of data for mapping and analyzing for knowledge

Each display has different physical parameters of the fault (geological structure) and surrounding host rocks. The use of multiple methods may be able to isolate some physical aspect of the fault system and provide geoscientists with more detailed information that could be used for future classification of structures. In one of the matured basins of Southeast Asia, the datasets are integrated to get a fair idea of sedimentary patterns, which otherwise not possible by seismic method, because of poor quality of the seismic data. From the residual gravity data, sediment patterns are clear, which are further verified the presence of sediments and their orientations (depositional features, for example) by seismic structure maps. The magnetic data provide definite clues on the type of basement and basement geometries.

Data mining and interpretation

Several data views are extracted from metadata through data mining algorithms. Simple algorithm used by Han et al. (2001) classifies the categories based on contexts. To fix the context and to clarify prolific terminology, a dataset X consisting of data points is considered (or synonymously, objects, instances, cases, patterns, trends, correlations, tuples, transactions) $x_i = (x_{i1}, \ldots, x_{id})$ Є A in attribute (anomaly) space A, where i = 1: N, and each component $x_{il,}$ Є A_l is a numerical or nominal categorical attribute (or synonymously, feature, variable, dimension, component, field, fact). Han et al. (2001) discuss various data types and their attributes in which point-by-attribute data format conceptually corresponds to a matrix and is used by majority of algorithms in cluster mapping. However, data of other formats, such as variable length sequences and heterogeneous geological-geophysical data, is becoming more and more popular. The simplest attribute space subset is a direct Cartesian product of sub-ranges C = Π C_l Є A, C_l Є A_l, l = 1: d, called a segment (also cube, cell, region). A unit is an elementary segment whose sub-ranges consist of a single category value, or of a small numerical bin. Describing the numbers of data points per every unit represents an extreme case of clustering, a histogram (that represent a data view), where no actual clustering takes place. This is a very expensive representation, and not a very revealing one. User driven segmentation is another commonly used practice in data exploration that utilizes expert knowledge regarding the importance of certain sub-domains. The author distinguishes clustering from segmentation to emphasize the importance of the automatic learning process. The ultimate goal of anomaly/attribute clustering is to assign points to a finite system of k-subsets, clusters. Usually subsets do not intersect and their union is equal to a full dataset with possible exception of outliers. Data views are extracted from cubes in the form of slices that represent bubble clusters with visualization features to further analyze for interpretation and mapping for a geological knowledge.

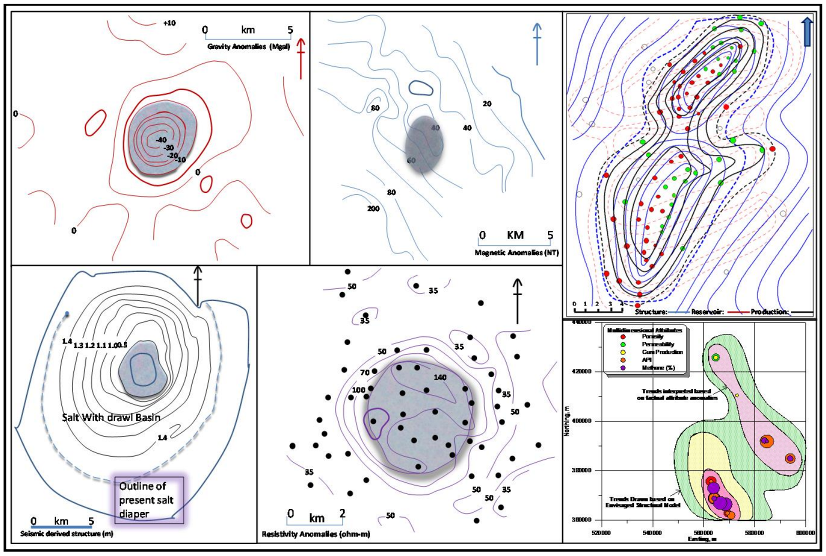


Figure 5.84b: Data views of different geophysical anomaly attributes of geological features

The contour values represent quantities describing the extent (Figure 5.84b) to which the geophysical anomaly attributes are scattered and distorted by subsurface phenomena and their configuration reflects the shape of subsurface *salt dome* with some accuracy (in this case, for example). The author integrates *structure*, *reservoir* and *production* (Figure 5.84b) anomalies to plan for oil and gas field development including exploration portfolios (to address RQ8 and RO8).

5.4.3 The Ugandan exploration business perspective

The analysis of multiple petroleum systems in a Total Petroleum System (TPS) scenario has been more challenging, in terms of organizing TPS data, information and representing them in different geological knowledge domains. The Geology and Geophysics (G & G) information in the case of Albertine Graben is described in a petroleum digital ecosystem, in which all the elements of petroleum systems of sub-systems and their attributes (through instances of their strengths) are shared through inherent communications and interactions.

Having analysed the concepts, tools and methodologies of organizing the Albertine Graben as PDE, the author takes on the concept of PDE to its implementation stage. All the elements and processes of all sub-systems are now intelligently stored in digits in a warehouse and now ready to extract different data views, for example, for an interpretable hydrocarbon (oil) play. An oilplay is a geological trend based on the *structure, reservoir* and *stratigraphy* that may contain a series of prospects. In the context of the Albertine-Graben, to evaluate an oilplay and begin generating leads and prospects, a significant amount of structural, stratigraphic frameworks and seismic interpretation tasks are completed. The current analysis suggests five elements are necessary for a viable prospect, identified from a system (*source rock*, *reservoir rock*, *seal-rock*, *trap-rock*, hydrocarbon *migration* pathway), including *porosity*, *permeability*, *API* attributes and characterized from ontologically structured multidimensional warehousing and mining that led to metadata representations. Multidimensional star schema definitions and descriptions are integrated in a single schema in the Albertine Graben's petroleum systems' warehouse. Data views extracted from warehouses are used, for visualisation and interpretation purposes.

Data visualization tools, methods and analysis

Visualization in information science point of view is to present the concepts hidden under vast amount of datasets, termed as *persuasive computing*, which uses technology to make people aware of complex concepts, in the ultimate goal to display processed data and information. Visualization of heterogeneous data, such as geological and geophysical data has significance in its representation and interpretation. Petroleum data presentation is the lowest level of abstraction, information is the next level, and finally the geological knowledge is the highest level among all three. Data on its own carries no meaning. For data to become information, it must be interpreted and take on a meaning. Main goal of petroleum data visualization is to interpret *exploration* and *production* information explicitly and effectively through graphical data views. To convey ideas of data integration more effectively, mined petroleum systems' data from warehouses are represented in both petro-info-graphics and statistical graphics. Petroleum exploration data are presented in elegant and descriptive form of tables, graphs, histograms, cross-plots, map views and bubble plots. Further, interpretation of maps, cross-plots and bubble plots among multidimensional attributes, convey knowledge on geographic and historical information, especially the exploration and production histories in terms of geography, geological ages including forecasts of exploration and drilling plans. Several tools and visualization methodologies are described in the following sections.

The visualization of various geological and geophysical (G & G) events in multiple dimensions is an objective and is analogous in representing G & G events in different knowledge domains. Several tools are available for representing multidimensional and heterogeneous petroleum exploration data for visualizing correlations, trends and patterns in multiple dimensions. Rock-Ware, Petrel, Database (DB) Studio offer popular visualization solutions in representing the spatio-temporal data. These visualizations offer multiple data views in multiple dimensions for geological interpretations and also assess the geological uncertainties. Visualization of data is done for facilitating interpretations in different knowledge domains, up-scaling of multidimensional data, such as drilled-well logs, uncertainty analysis, analysis of facies map views, structural and petrophysical data models. The visualization is in fact, maximizing information usage to optimize exploration and production targets and for analyzing reservoir properties that have serious production complexities. Histograms, variograms, correlations, trends and patterns are various other methods of analyzing property or attribute variations in the petroleum system models.

Figure 5.85 is another demonstration of a multidimensional view of the information corresponding to graben data represented in cubes. It has multiple dimensions, such as yearly exploration and production data of the Albertine Graben. Using the grapher and surfer solutions, bubble plots and surface map views are plotted. A typical data cube built for the Albertine Graben (Figure 5.85) and different map views computed by surfer (Rockware Mapping Solutions) for interpretation are key highlights of the data mining and visualization. This cube is a representation of logical organization of data instances of sub-basins of the Albertine Graben and their associated attributes of geological, geophysical and depositional systems.

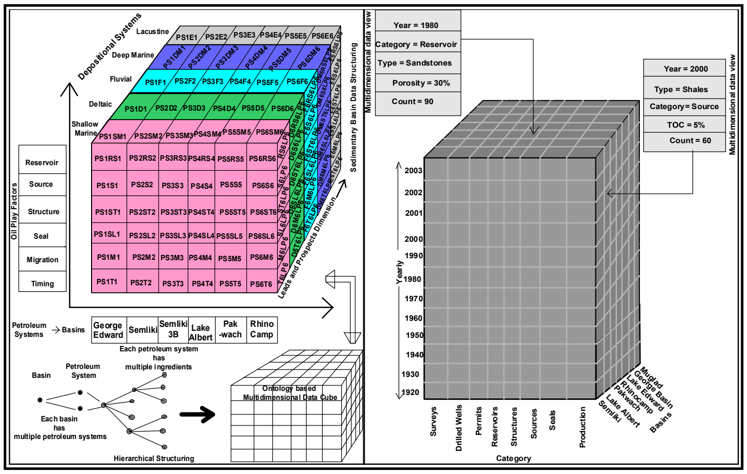

Depositional Systems
Lacustine
PS1E1 PS2E2 PS3E3 PS4E4 PS5E5 PS6E6
Deep Marine
PS1DM1 PS2DM2 PS3DM3 PS4DM4 PS5DM5 PS6DM6
Fluvial
PS1F1 PS2F2 PS3F3 PS4F4 PS5F5 PS6F6
Deltaic
PS1D1 PS2D2 PS3D3 PS4D4 PS5D5 PS6D6
Shallow Marine
PS1SM1 PS2SM2 PS3SM3 PS4SM4 PS5SM5 PS6SM6
Oil Play Factors
Reservoir
PS1RS1 PS2RS2 PS3RS3 PS4RS4 PS5RS5 PS6RS6
Source
PS1S1 PS2S2 PS3S3 PS4S4 PS5S5 PS6S6
Structure
PS1ST1 PS2ST2 PS3ST3 PS4ST4 PS5ST5 PS6ST6
Seal
PS1SL1 PS2SL2 PS3SL3 PS4SL4 PS5SL5 PS6SL6
Migration
PS1M1 PS2M2 PS3M3 PS4M4 PS5M5 PS6M6
Timing
PS1T1 PS2T2 PS3T3 PS4T4 PS5T5 PS6T6
Leads and Prospects Dimension
Sedimentary Basin Data Structuring
Petroleum Systems → Basins
George Edward
Semliki
Semliki 3B
Lake Albert
Pak -wach
Rhino Camp
Each petroleum system has multiple ingredients
Basin
Petroleum System
Each basin has multiple petroleum systems
Hierarchical Structuring
Ontology based Multidimensional Data Cube
Multidimensional data view
Year = 1980
Category = Reservoir
Type = Sandstones
Porosity = 30%
Count = 90
Multidimensional data view
Year = 2000
Type = Shales
Category= Source
TOC = 5%
Count = 60
Yearly
2003
2002
2001
2000
1990
1980
1970
1960
1950
1940
1930
1920
Surveys
Drilled Wells
Permits
Reservoirs
Structures
Sources
Seals
Production
Category
Muglad
George Basin
Lake Edward
Rhinocamp
Pakwach
Lake Albert
Semliki
Basins

Figure 5.85: Representation of petroleum sub-systems data organization of Albertine Graben (an ecosystem) within a data-cube

Data analysis

Analysis of petroleum exploration and production data, like any other heterogeneous nature, is a process of inspecting, cleaning, transforming, and modeling them with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Heterogeneous data analysis has multiple facets and approaches, encompassing diverse techniques under variety of names, and in different business situations. For example, geological and geophysical (G&G) data have different dimensions and these multiple dimensions are integrated, structured and represented in data cubes for analyzing different geological knowledge domains. Analysis of variables in multiple dimensions and their correlations, description of reservoir heterogeneities and establishing uncertainties are other implementations, as described in Parasnis (1997). Data and map views depict property variations and for example, each property is *structure*, *reservoir*, *seal* and *source* rock strengths. Exploration and production forecasts and predictions are the result of visualized data views and their interpretations. Besides, probability is measurement of likelihood of an event (%), variance, and a measurement of how different dimensions are correlated from each other, and a way to measure relationships between two separate dimensions. Anisotropy is another property, to measure whether variance within a collection of data is determined by a direction property. Histogram, normal probability distributions, variance, cumulative distributions, cross-plots, correlation coefficients are different methods used for analyzing petroleum exploration data at different historical periods. 2D/3D graphs, 2D/3D contour and surface maps are further methods used to represent multidimensional petroleum data.

2D Graphs*:* Exploration data are displayed as a line, as symbols, or as a combination of a line and symbols. Petroleum heterogeneous spatial-temporal data are plotted in an order in which they appear in the data file and they may be multiple line/scatter plots in a graph. In addition, these plots can contain fit curves, error bars, labels and color fills. Histograms display data in groups or bins. Each bin represents a range of values on the X axis. The height of a bin represents the number of data points that fall within that bin's range. If there is one Y value for each X value in the data set, use the bar chart rather than the histogram. Polar graph displays degree, radian, or grad data versus a radial distance. Bubble plot displays two variables on a scatter-type plot. In a 2D bubble plot, the diameter of each bubble can vary in size, providing a way to represent an additional dimension of data. For example, consider a traditional scatter plot that shows the number of new hydrocarbon plays in a basin over a period of time. Using a 2D bubble plot, a third dimension is displayed from data that shows the average crude production from a basin (under petroleum systems' constraints) and under the same time span. Bubble plot displays four variables on a scatter-type plot.

The diameter of each bubble can vary in size and Z position (attribute instance of an element of petroleum system, for example), providing a way to represent more data on a single plot.

2D/3D Contour Maps: XY Data (or XZ Data) creates a contour map from a data file (such as a [.DAT] or [.XLS] file). Data are gridded using the inverse distance algorithm. A contour map is a two-dimensional representation of three-dimensional data. Two types of contour data maps are created: XY data maps and XZ data maps. The first two dimensions of the XY data map are the XY coordinates; the third dimension (Z) is represented by lines of equal value (for example, exploration/production data instances). The relative spacing of the contour lines indicates the relative slope of the surface. The area between two contour lines contains only grid nodes having Z values within the limits defined by the two enclosing contours. The difference between two contour lines is defined as the contour interval. Gridded data are used in contouring the data trends and patterns. All the data represented in graph and maps views, are interpreted in geological knowledge domains. Implementation of the data mining is described in the next section.

Schema implementation and data mining in the Albertine Graben

The Albertine Graben petroleum system's PDE warehouse supports several petroleum ecosystem's data types, each of which is implemented as one or more tables in the schemas. Dimensional and factual data tables of the Albertine Graben are not shown here because of intellectual property rights. Like previous cases, every dimension in the warehouse of the basin has multiple rows in the entry table that defines the metadata, such as the *seismic* time instance or *geological-age* instance is inserted in the warehouse, and its time of last update. Every warehouse entity is also associated with the dataset DB from which domain it was derived. However, different data views are drawn from a metadata (of a warehoused mode) for interpreting trends and correlations (Rudra and Nimmagadda 2005 and Wand 2000) that make up the connectivity and communication among relevant data attributes of multiple dimensions. Data relationships plotted between period and other conceptualized (logical) data dimensions have been deduced with interesting connectivity among data properties of sub-basins of the Albertine Graben. *Structure*, *reservoir*, *production* data attributes, among different petroleum sub-systems and different depositional environments, provide valuable data correlations and trends among periodically and geographically distributed petroleum systems of rift basins. As shown in Figure 5.86, an integrated framework is used for building and analysing domain knowledge from the Albertine-Graben petroleum digital ecosystem (PDE).

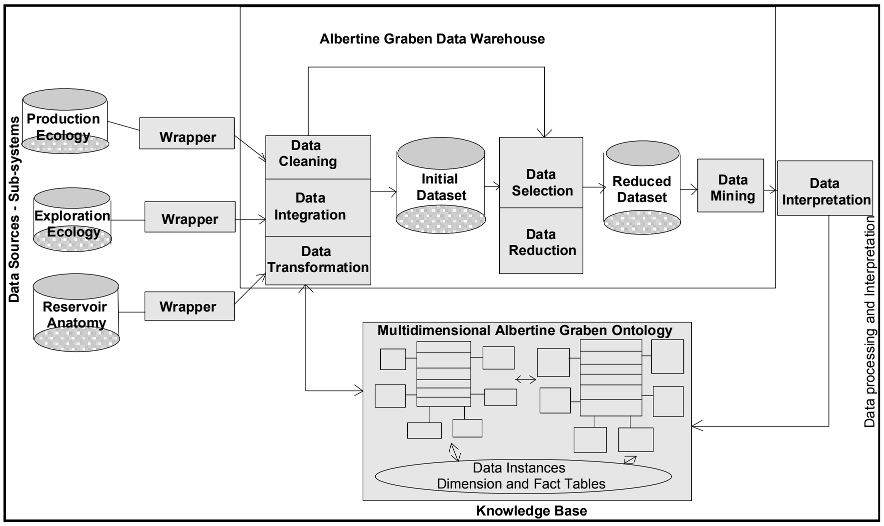
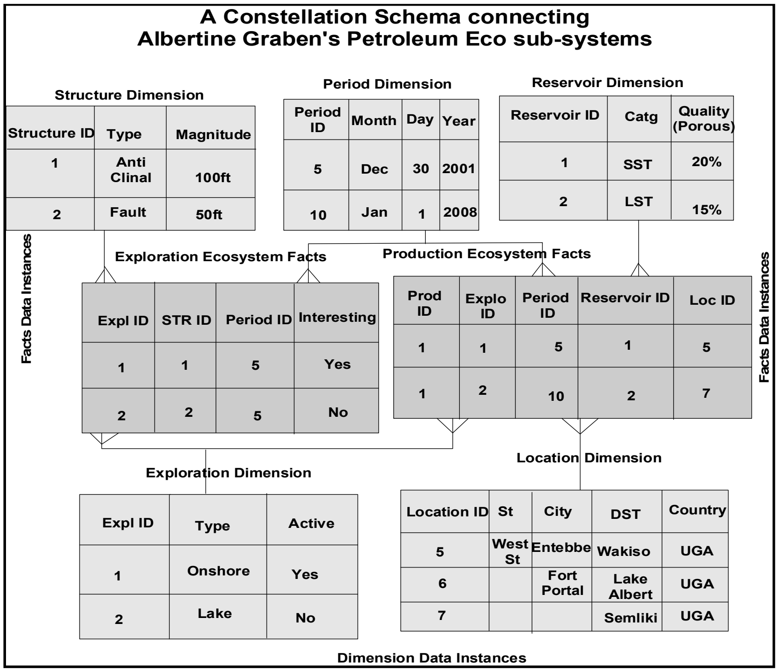


Figure 5.86: Knowledge building process model (modified version)

The schema design and implementation solutions are very flexible, but may be more complicated because of many variants of aggregation. In a fact constellation schema, different fact tables are explicitly assigned to the dimensions, which are for given facts relevant. This may be useful in cases when some facts are associated with a given dimension level and other facts with a deeper dimension level. In that case using two or more different fact tables on a different level of grouping is realized through a fact constellation model. Sophisticated applications may require multiple fact tables to *share* dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation. As shown in Figure 5.87, a constellation schema with facts from multiple dimensions and their tables are interconnected. This ontology representation is responsible for narrating a generalized ecosystem and specified sub-systems. The geological events inherited within petroleum ecosystems are narrated in different visualization views as described in the next section.


A Constellation Schema connecting
Albertine Graben's Petroleum Eco sub-systems
Facts Data Instances
Facts Data Instances
Dimension Data Instances

Structure Dimension

Structure ID	Type	Magnitude
1	Anti Clinal	100ft
2	Fault	50ft

Period Dimension

Period ID	Month	Day	Year
5	Dec	30	2001
10	Jan	1	2008

Reservoir Dimension

Reservoir ID	Catg	Quality (Porous)
1	SST	20%
2	LST	15%

Exploration Ecosystem Facts

Expl ID	STR ID	Period ID	Interesting
1	1	5	Yes
2	2	5	No

Production Ecosystem Facts

Prod ID	Explo ID	Period ID	Reservoir ID	Loc ID
1	1	5	1	5
1	2	10	2	7

Exploration Dimension

Expl ID	Type	Active
1	Onshore	Yes
2	Lake	No

Location Dimension

Location ID	St	City	DST	Country
5	West St	Entebbe	Wakiso	UGA
6		Fort Portal	Lake Albert	UGA
7			Semliki	UGA

Figure 5.87: An ontology model depicting the connectivity among data instances, for mapping and modelling multiple system elements' dimensions

G & G data visualization

Visualization of the exploration data, such as seismic, well-log data and other petrophysical data are represented in 2D/3D windows (base maps), 3D windows, interpretation windows, including displaying and manipulating dip- and strike lines (in case of 2D), in-lines and cross-lines time-slices and random lines and drilled-well-cross sections (Parasnis 1997 and Magoom and Dow 1994). Exploration and production data of producing oil and gas fields in sedimentary basins, in different domains are integrated in a warehouse environment. Integrated metadata are mined for understanding anisotropy, heterogeneity and similarity of properties of attributes. These could be through variograms, correlograms, histograms and krigging. A krigging is a special statistical procedure used with variograms, or two-point statistical functions that describe the increasing difference or decreasing correlation between sample values as separation between them increases, to determine the value of a point in a heterogeneous grid from known values nearby. In the multidimensional heterogeneous data analysis view point, a two-point statistical function describes the increasing difference or decreasing correlation, or continuity, between sample values as separation between them increases. The term variogram is sometimes used incorrectly in place of semi-variogram. The two differ only in that the semi-variogram uses each pair of data elements only once, whereas the variogram uses all possible data pairs. Semi-variograms are usually used instead of variograms, but opposite vector directions (for example, north and south) are recognized as representing the same thing and having identical ranges, sills, nugget points and the like. The author makes similar analogy between dip and strike attributes, the attitudes of sedimentary beds either may be increasing or decreasing in multiple directions, including similar multiple trends and correlations interpreted among petrophysical properties and drilled-well data attributes. These are graphical representations in which author brings out trends and correlations for geological knowledge representation and interpretation. Handling large amount of data, provide consistent analysis in multiple dimensions, visualize multiple geological interpretations and assess uncertainty are other purposes of visualization tools. Different visualization tools, used in the petroleum industries, characterize geological and geophysical data processed and interpretation results.

Visualization for data calibration and correlations: The exploration data in multiple domains are connected and the connectivity is made through their common data attributes. As shown in Figure 5.88, visualization of data views is representative for correlation among multidimensional attributes, magnitudes and their strengths.

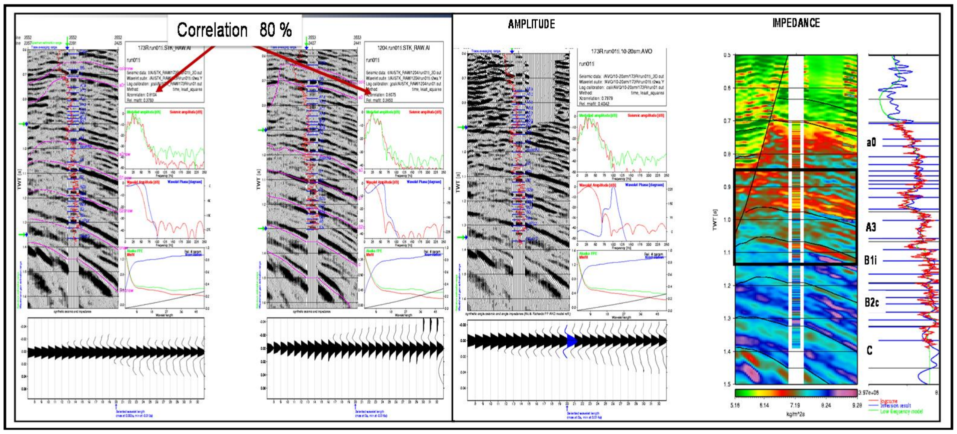


Figure 5.88: Calibration, correlation and integration of multidimensional data attributes

Visualization for structure interpretation: An accurate characterization of fractured reservoirs requires an ability to integrate multi-scale measurements (seismic, image logs, well tests) and build geologic models, incorporating fractures. As demonstrated in Figure 5.89, structure impressions such as faults, folds and anticlinal closures are visualized as on map views for easy interpretation. The color and scale bars can quickly identify visually the attribute strengths.

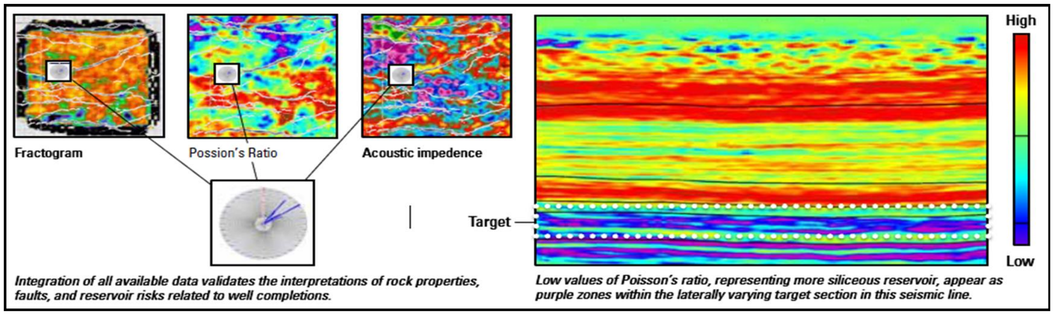


Figure 5.89: Visualization of fault and fracture attributes of a petroleum system

***Visualization for multidimensional attribute representation**:* Multidimensional exploration data, such as 2D/3D seismic-lines, point, polygons and their corresponding coordinates, seismic, drilled-well and geological outcropping data are populated in an integrated data warehouse environment. Warehoused metadata in a volume are sliced using mining algorithms available in Grapher and Surfer solutions, in which each data slice is represented as a data view and in a dimension (Figure 5.90). These data views are either in the form of cross-plots (Figure 5.91), map views or cross-sections, extracted from warehoused seismic volumes. The cross-plots are usually represented in single, two-dimensional and multidimensional domains. For example, cross plots drawn among porosity and permeability of productive reservoirs can visually narrate the correlation of the attributes and trends in the form of strengths and magnitudes of reservoirs among associated lithologies (Figure 5.90).

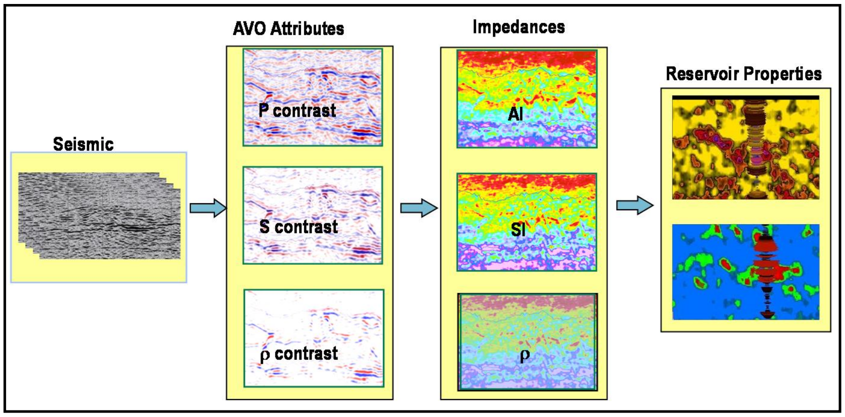


Figure 5.90: Multidimensional data views of *exploration* metadata

***Visualization for addressing data issues**:* Data quality is an important issue, dealt with at many stages of exploration and production project design and implementations. For example, during review of seismic reflection continuities, seismic resolution issues, for the purpose of preserving geological events during acquisition and processing stages of exploration, depth matching, cycle skip and data mis-ties in the log- and seismic data interpretations (Figure 5.91), are some of the data issues, easily addressed through visualization and representation of the data views in multiple dimensions.

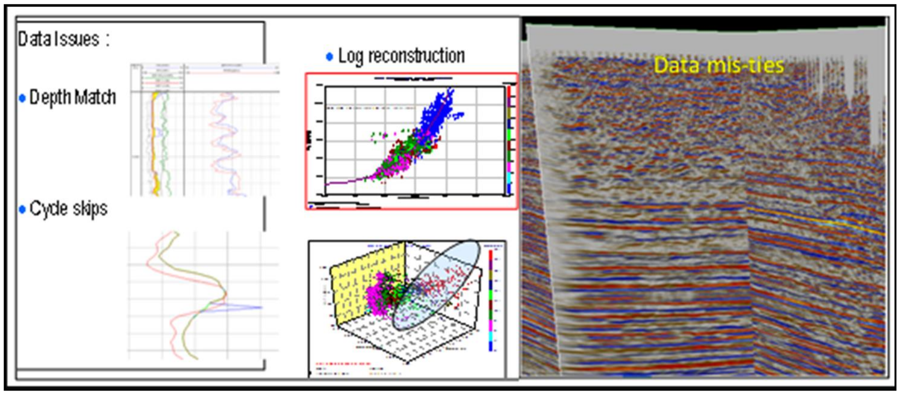


Figure 5.91: Visualization addressing data issues

The author uses data views extracted from warehouses, for discovering new knowledge and interpretation of knowledge in true geological information. As shown in Figures 5.89 -5.90 and 5.92, plot views are drawn in which, multidimensional data are represented in knowledge domains on fracture reservoir ecosystems. The Albertine-Graben sedimentary basin, as a petroleum digital ecosystem, is analyzed using multidimensional data mining and visualization approaches. Interpretation of these map views are discussed in the next section.

Data mining and interpretation

The data mining in the context of petroleum exploration and development of the Albertine Graben, is a search for the relationships and global patterns that are hidden among petroleum systems of sub systems of sub-basins of the Albertine Graben. The relationships are in between production data and the petroleum system analysis of all sub-basins. These relationships, if mapped and visually plotted in multiple dimensions, make known valuable knowledge of petroleum systems and elements of systems. Hydrocarbon play elements, such as *structure*, *reservoir*, *source, seal* and play processes, such as *migration-pathways* and *timing* of depositions are represented in multiple dimensions. Several data views extracted from warehoused metadata of petroleum ecosystem, in which data property relationships hidden in metadata are brought out with correlations, patterns and trends of reservoir properties including hydrocarbon reserve forecasts. These views are used by a petroleum geologist for interpreting geological knowledge and forecast of exploration and development of oil and gas fields. These data views are also used for drillable exploration and development new target areas. The color bars show the magnitude and strengths of the attributes of producing horizons. The maps with multiple dimensions and comparable attribute strengths suggest productivity of each horizon and forecast of similar quality reservoirs elsewhere in the other sub-basins.

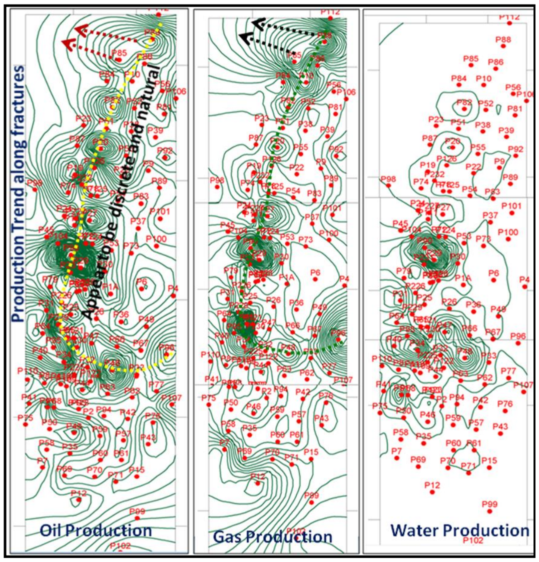

Production Trend along fractures
Appear to be discrete and natural
Oil Production
Gas Production
Water Production

Figure 5.92: Data mining views extracted from relationships built among production and other spatial data attributes (crowded contours because of dense fractures)

As narrated in Figure 5.92, it is a demonstration (RQ 8 and RO 8) of representing interconnectivity among fracture systems, which hold most of oil and gas deposits in several sub-basin ecosystems, risk minimizing exploration. As shown in Figure 5.92, data views are interpreted for fracture connectivity from warehoused petroleum ecosystems, in which relationships built are representative of *structure* data attributes. As interpreted from Figure 5.92, contours are either narrowed or clustered at one place, suggesting fractures developed at different zones. These clustered contours at different zones appear to be due to interconnectivity among fractures that held commercial quantities of hydrocarbons at multiple levels (horizons). The data and maps views are effective for data interpretation and geological knowledge extraction. The fracture density and orientations are other key attributes narrated from map views, as demonstrated in Figure 5.92.

5.4.4 The Arabian-gulf exploration business perspective

TD (Time-Depth Data) analysis

The author has analyzed various data views from warehoused metadata. For example, *TVD* (ft) and *interval velocity* dimensional views are extracted and they are thus plotted for different wells and formations. As shown in Figures 5.93 – 5.99, there are definite data trends among the time-depth-velocity datasets and the author uses these trends for gridding separately to achieve a better structural definition.

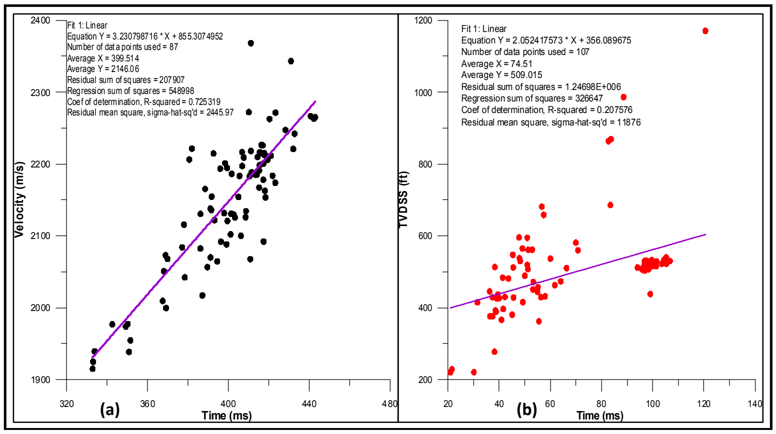


Figure 5.93: Constructing time-depth and time-velocity relationships, used for exploring multiple connections among oil and gas fields

Datasets relevant to certain closely associated wells and their corresponding fields have been extracted from warehoused Metadata. As shown in Figure 5.93, the author plots seismic time (ms) and TVDSS (ft) values.

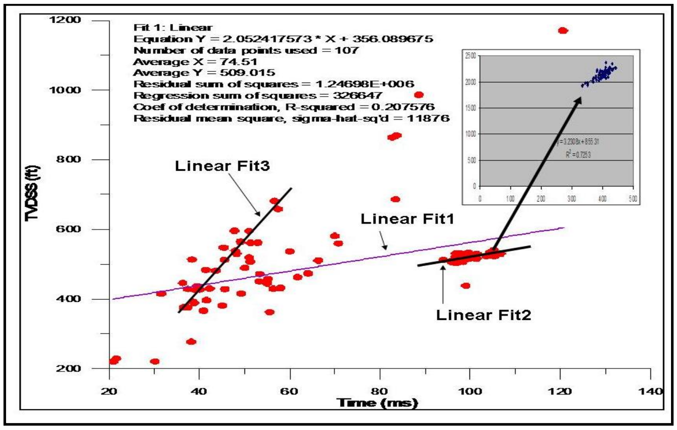


Figure 5.94: Constructing relationships among seismic time and depth datasets extracted from multiple oil and gas fields

It is interesting to observe from a scatter-plot of all data for key drilled-wells, where a broad trend-line exists. However, within this broader view, there appear to be smaller and finer data patterns (as indicated with Linear Fit 1, 2, and 3, Figure 5.94). The ontological warehousing extracts these finer data patterns that will reveal representative trends. If an equation is computed for each smaller dataset and applied for gridding the volume for that particular formation, it can give a better structural refinement in the contoured data.

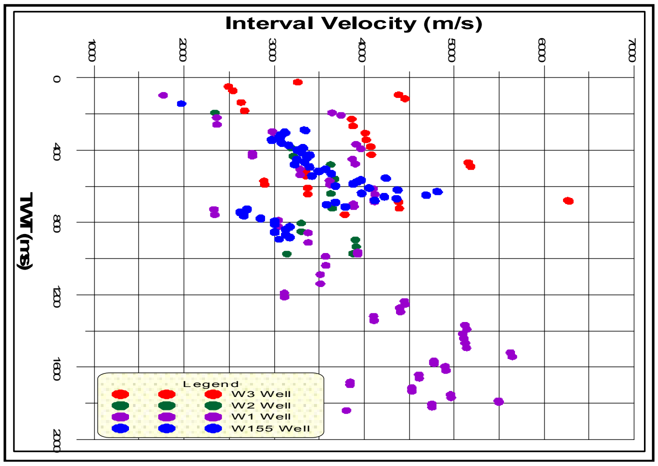


Figure 5.95: Constructing relationships between two-way time and interval velocity datasets, extracted for different wells

Similarly, the author plots several data views extracted from warehoused data volumes separately permitting finer data patterns. As shown in Figures 5.93 – 5.99, it is interesting to observe data patterns among combination of time-depth and velocity plots. The warehousing gathers all the finer data patterns and constructs relationships - fast, slow and steady changing velocity data views gathered and interpreted as represented in Figures 5.93 – 5.99, in which data relationships are used further in the grid computations and contouring of gridded data. It is always good to have firsthand geological knowledge of the study area, where velocities have definite roles on structural data anomalies. This facilitates an ontologist and the data warehouse designer, the type of grids and GI to be chosen for time-depth computations.

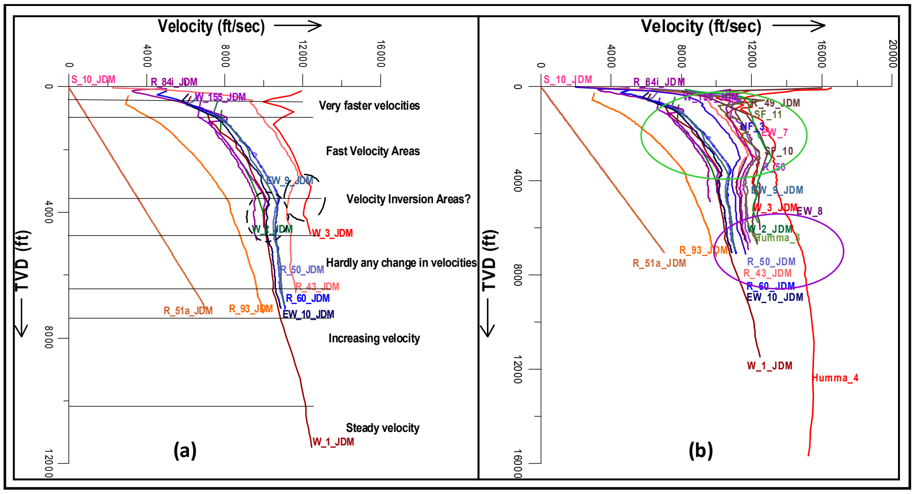


Figure 5.96: Mining of data views from warehoused metadata – construction of relationships among true vertical depths and average velocities

The author plots TWT and velocities for different wells of several fields in the matured basin. As shown in Figures 5.96 – 5.98, different velocity data patterns are observed because of anomalous geological situations associated with low-velocity oil/gas anomalies. The author interprets fast changing velocities with near surface geology.

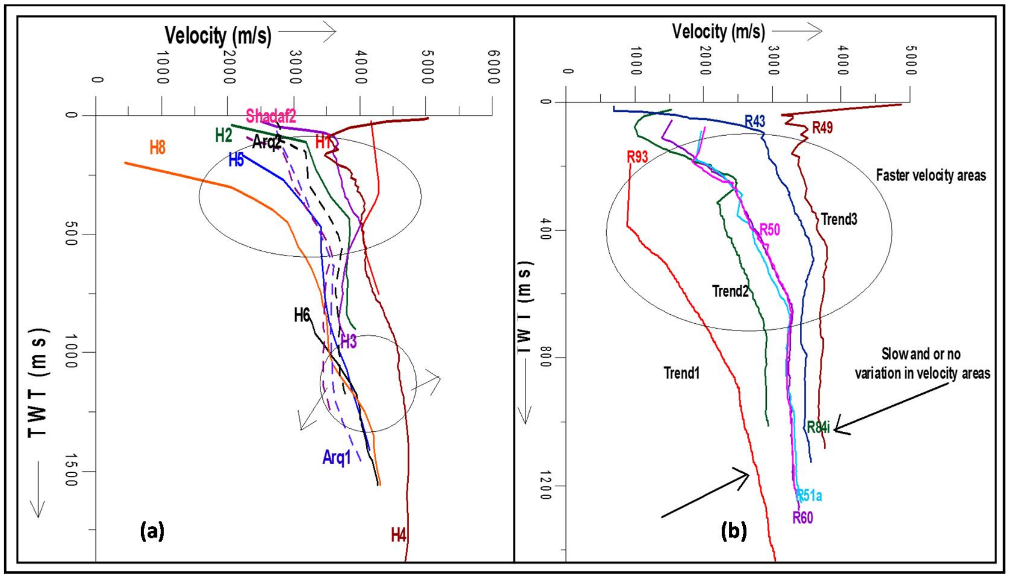


Figure 5.97: Constructing relationships between time-velocities in different oil fields

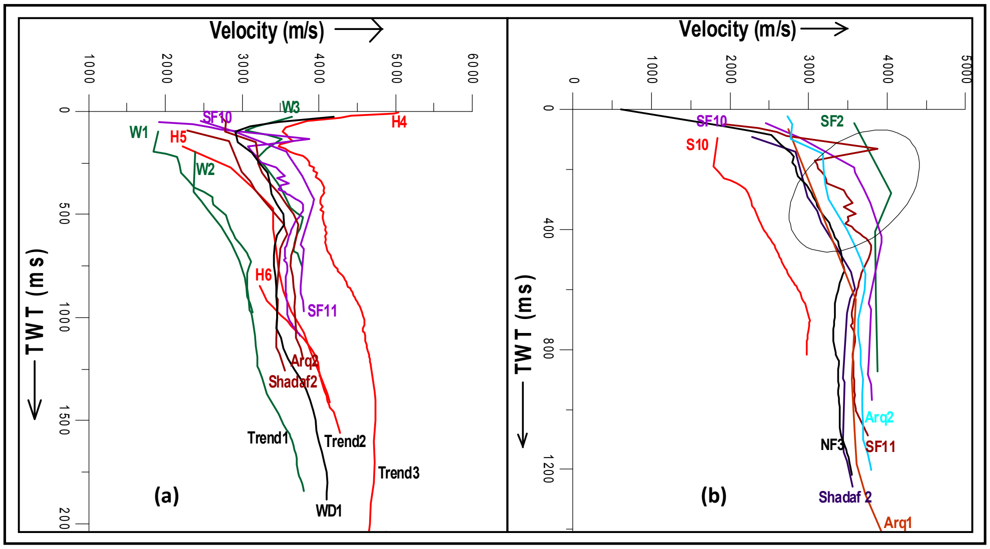


Figure 5.98: Constructing graphic relationships between time-velocities attributes in different oil fields

The computed grid data are contoured for understanding the data patterns for different conceptually derived relationships. The new knowledge is extracted from an interpretable structure at different well locations. It is interesting to observe, for different constructed relationships among time-depth-velocity pairs, the structural patterns are different. For Linear Fit 1 (as shown in Figure 5.94), it is interesting to notice a closed structure data pattern in the northern part of the study area. When more and more wells and their formation tops are involved in the mapping process in well- and seismic-domains, interpretation has definitely enhanced the knowledge of the structure. In the matured basins, when ontologically described historical data are warehoused (integrated into a metadata) and various geological views are interpreted, there is definite improvement in the structural resolution. The spatial-temporal dimensions have added new knowledge in the study area. From Figure 5.99, better structural resolution is observed and definitions of VSP (vertical seismic profiling), check-shot data, and stacking-velocity datasets (key terms defined in glossary in Appendix-1) are applied. These are conceptually construed relationships between time-depth data sets. It can be inferred that structural resolution is improved when the data are integrated in both seismic and well-base domain ontologies.

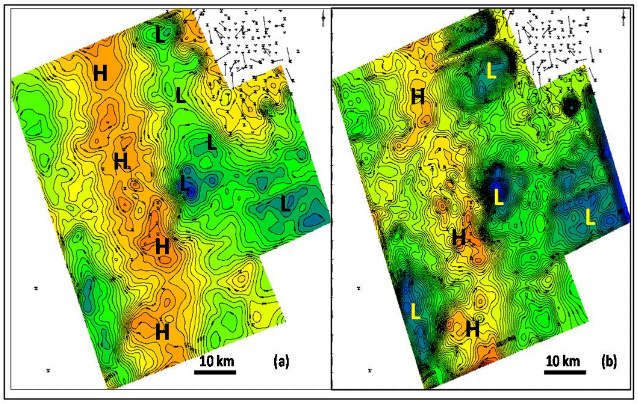


Figure 5.99: Time to depth structure map views after applying VSP and stacking velocity instances to seismic data instances (H: Structural High; L: Structural Low)

It is interesting to observe that structure contours are getting closed, when a linear equation is applied in the grid computations. The equation extracted from the cross plots (Figure 5.94), is applied to compute structural anomalies (H and L) as shown in Figure 5.99. Another form of structural pattern is observed when velocity data are computed and modelled at well locations using well tops (in depth-domain) and seismic times in seismic domain (from ontology structural views). The structure is better resolved (as represented in Figure 5.100) when VSP data with check shots and RMS stack velocity data from all SP/CDPs are applied in the grid computations. Grid interval (GI) which has definite impact on structural resolution, with smaller GI, obviously structural anomalies are much finer, which may not necessarily be attributable to actual structural patterns as shown on the right hand image of Figure 5.100. After careful analysis of structures with different grid interval (GI) and velocity models, an interpretable structural map has been arrived at, as shown (right hand image) in Figure 5.100.

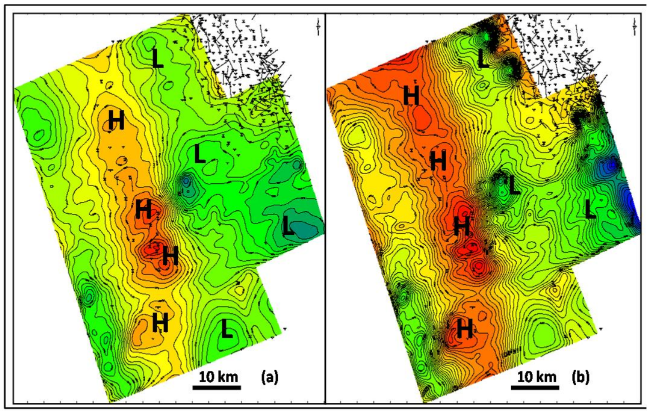


Figure 5.100: Construction of structure map views using VSP and linear relationships between time-depth datasets (H and L are structure anomalies)

Interpretation of clusters for knowledge building and discovery

The author interprets the results of clustering appropriately in the wider context of the application and implementation. Clustering of different petroleum data attributes may be considered as an initial data exploration tool before the design of further decision-tree or classification system or constructing an association rule involving a fixed number of classes. In all these cases, the clusters must be suitably transformed or interpreted. The most common representation of clustering is by centroids. This is effective for compact and isotropic clusters but not for elongated or anisotropic clusters. In certain applications, the extremities of a cluster are used to form the conjunctive expressions in rule sets. *Clustering* in our present study is division of petroleum data into smaller groups of similar objects. Each group consists of objects possessing similarities and dissimilar to objects of other groups. Representing data by fewer clusters necessarily diminishes fine-grained details, but achieves simplification. Author interprets hidden patterns from clusters in terms of variations in distances or concepts of oil-play data. The clustering in the present context, is said to be an unsupervised learning of a hidden petroleum system data.

The data mining process starts with the assessment of whether any cluster trend has meaning and interpretation, and correspondingly includes appropriate selection, and in many cases feature construction. Validation and evaluation of the resulting clusters must be assessed with their interpretability and visualization. Interpretability depends on the size of data (mostly numerical) and method used for building cluster. Dense areas around centroids score well. Cluster validation and their visualization issues are addressed in Jain and Murthy (1999) and Kandogan (2001). When two partitions (Pujari 2002, Jain and Murthy 1999 and Rudra and Nimmagadda 2005) occur as visualized and interpreted as different sub-sets, each is implied as separable granularity and an interpretable cluster. The distance between any two centroids normalized by corresponding clusters' radii (standard deviation) and averaged (with cluster weights) is a reasonable choice of coefficient of separation.

Clustering of onshore oil and gas field data: In one of the medium sized onshore fields in the South-east Asia, there are multi-stacked reservoirs, producing from this field creating surprises. More than 500 wells have been drilled on this single field. Geological structures and reservoirs and their areal extents are uncertain and unknown in the petroleum system. Many wells are wet or abandoned thus costing the industry millions of dollars. The author attempts to use data warehousing and mining methodologies to understand data trends in the known oil accumulations and predict the currently unknown. Figure 5.101 is a drilled-well geographic location map view, showing an existing well placement. The metadata is created using a warehousing modeling approach. Hierarchical ontology is used for building relationships among geological structures and reservoirs with their production data instances. These data are integrated in a warehouse environment and mined to create interpretable data views.

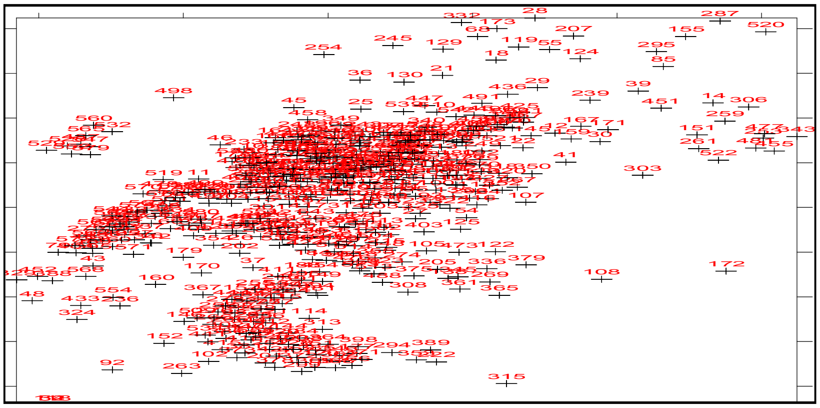


Figure 5.101: Base map showing well data associated with one of the onshore producing basins in South-East Asia, a single medium sized field (+, well position)

The author extracts several data views for interpreting data trends among several attributes of structure, reservoirs and production entities and these trends are shown in Figures 5.102a – 5.102c. Reservoir, geological structure, and production data attributes have been plotted with several bubbles on a single scale to explore data trends. Distribution of bubbles, size of bubbles, and direction of bubbles are key criteria in understanding and exploring data entities and their attributes.

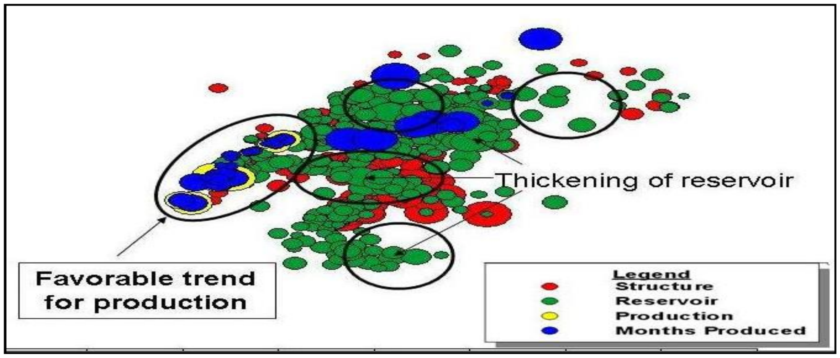


Figure 5.102a: *Structure*, *reservoir* and *production* data attributes at producing *horizon-1*

These bubbles are clustered in 2D X, Y plane. Bubbles are closely clustered because of the fact that data are from single field. However, size of bubble differentiates the attribute strength.

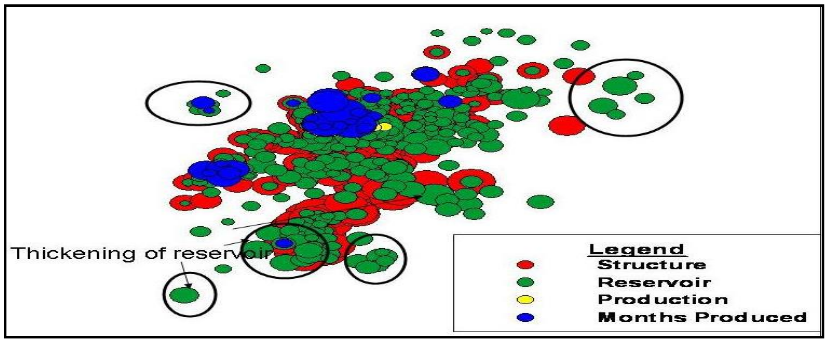


Figure 5.102b: *Structure, reservoir* and *production* attributes at a producing *horizon-2*

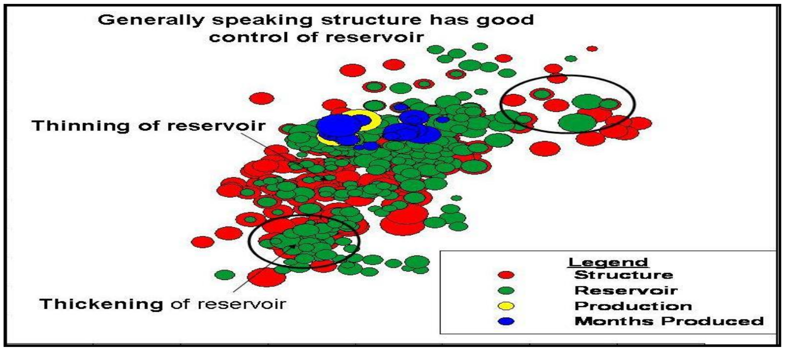


Figure 5.102c: *Structure, reservoir* and *production* attributes at a producing *horizon-3*

As shown in Figures 5.102a – 5.102c, the author plots clusters of bubbles for different attributes and instances of *structure, reservoir* and *production* properties to explore data patterns. Interesting trends are explored, with good fit among producing structure and reservoir attributes. Some of these data are fitted with linear trends, especially with reservoir thickness and porosity (reservoir properties), inferring relationships among porosity properties among multiple fields, resolving the issues related to structure and reservoir. This has significance in knowledge building and conceptualizing relationships among complex petroleum systems. A prerequisite for achieving this knowledge depends on effective data integration and denomalizing relationships during fine-grained multidimensional data structuring process (at warehouse modeling stage).

Clustering of multiple onshore oil and gas field data from Arabian Gulf Basins: There are multiple oil and gas fields within a producing basin (Figure 5.103). Each field has hundreds of producing wells, several survey lines and geological information. Though many fields, in general, have good structural bearing, there are reservoir uncertainties, resulting in the abandonment of some wells. The author carries out warehouse modeling to understand the data trends and patterns of attributes among these interconnected fields (Figure 5.103).

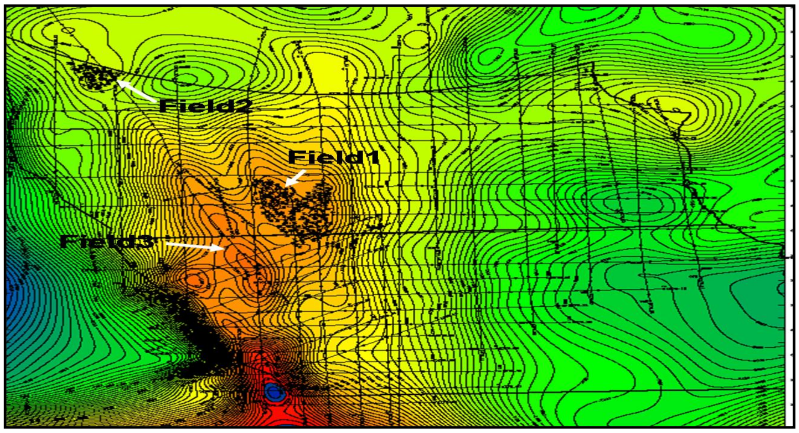


Figure 5.103: Multiple oil and gas fields in an onshore Arabian Gulf basin

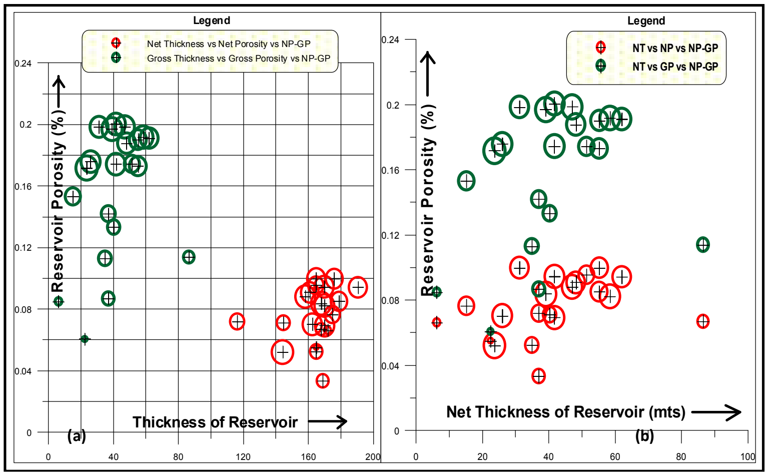


Figure 5.104a: Clusters of porosity values of Middle Eastern oil and gas reservoirs

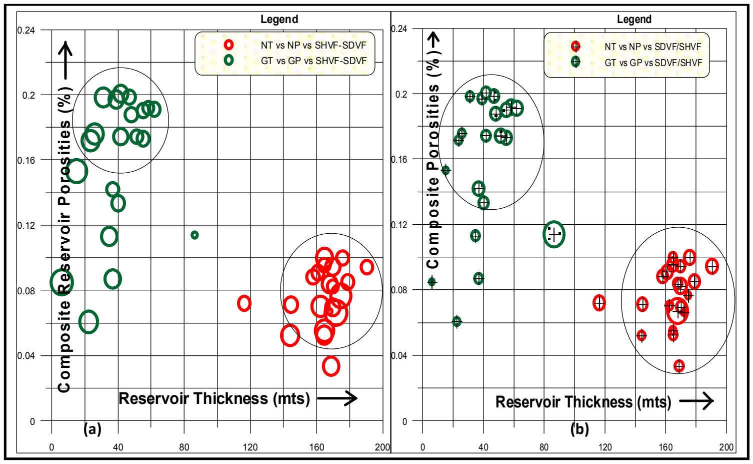


Figure 5.104b: Clusters of composite porosities of Middle-eastern oil and gas reservoirs

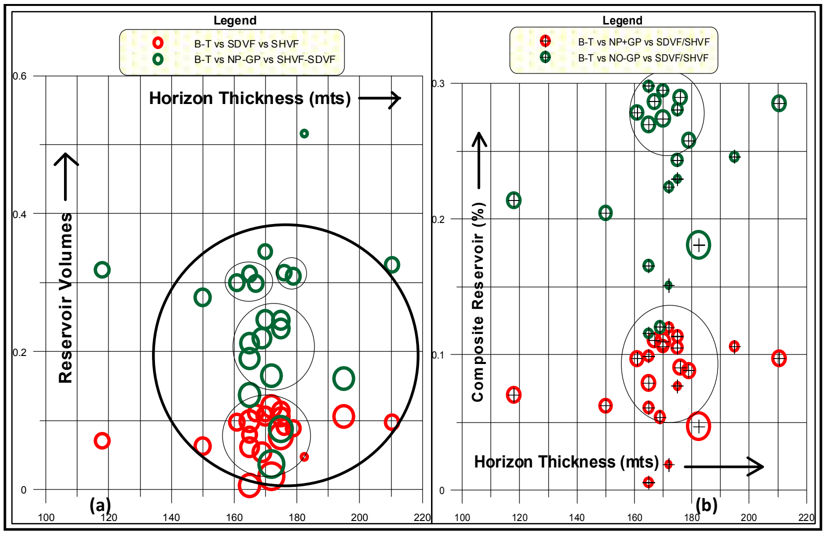


Figure 5.104c: Reservoir volume ratio attributes

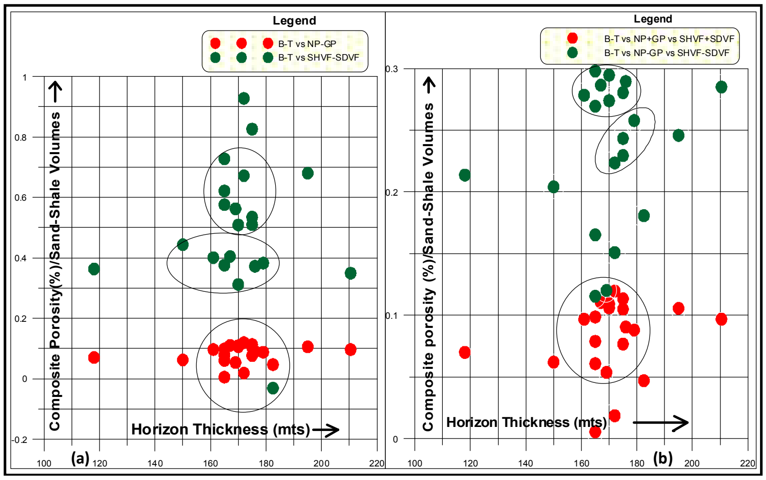


Figure 5.104d: Clusters of composite *porosities*

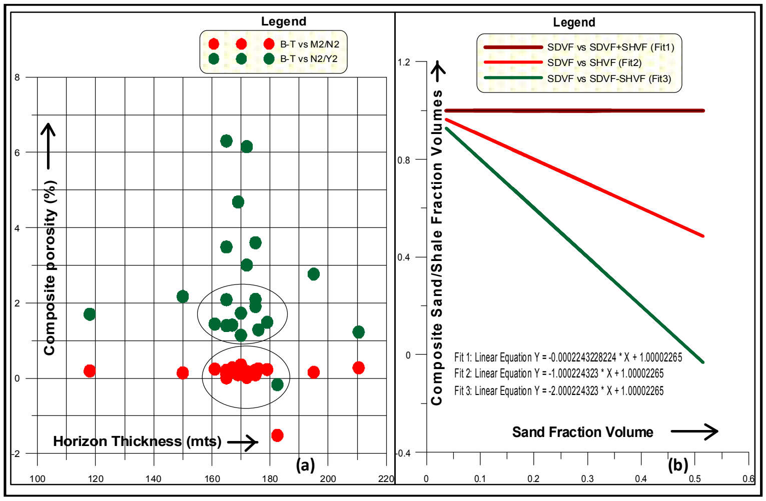


Figure 5.104e: Sand/Shale fraction volume attributes

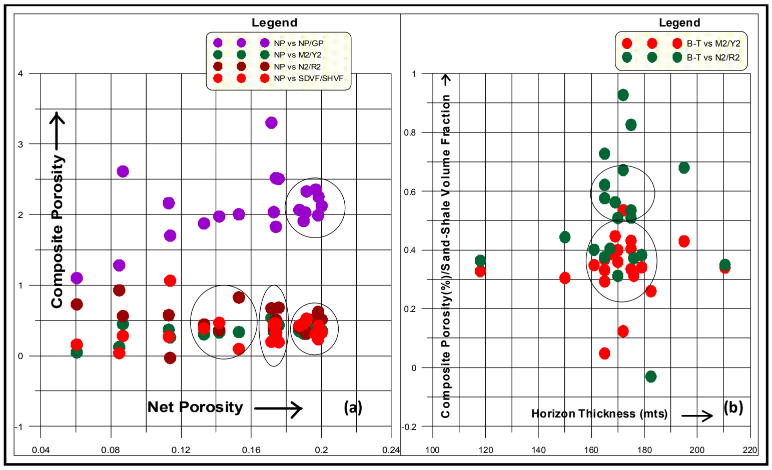


Figure 5.104f: Composite porosity clusters

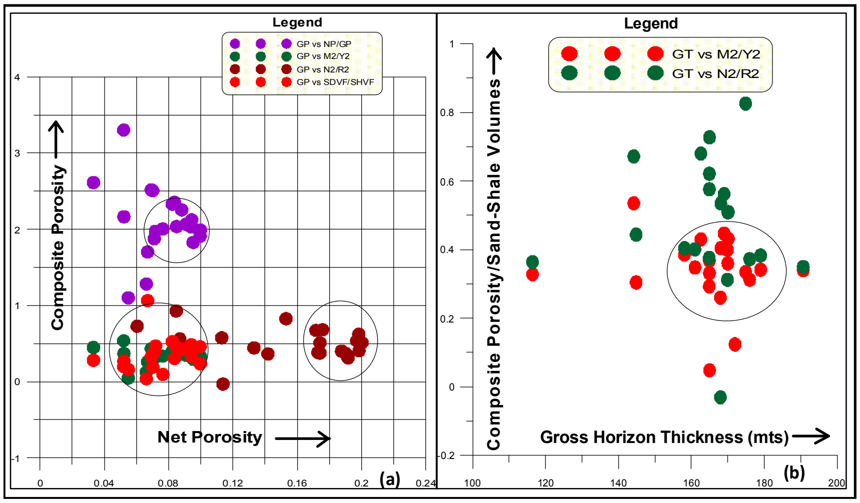


Figure 5.104g: Clusters of composite porosities

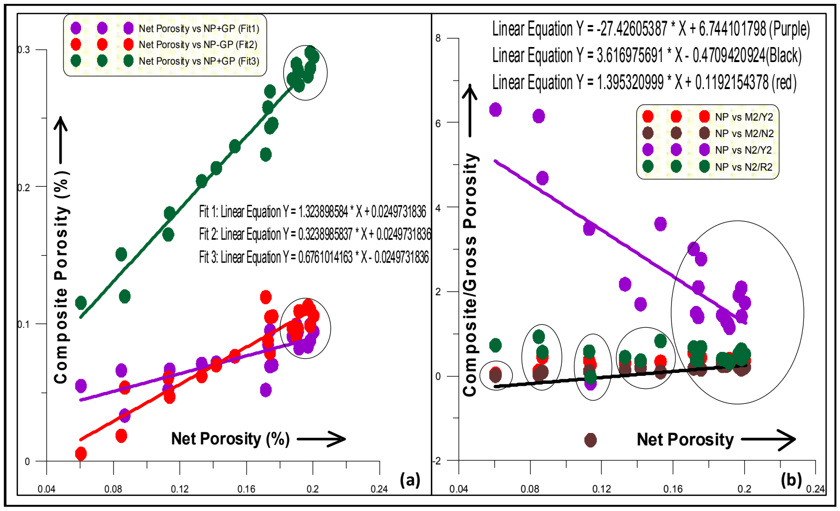


Figure 5.105a: Computed data showing, negative and positive relationships

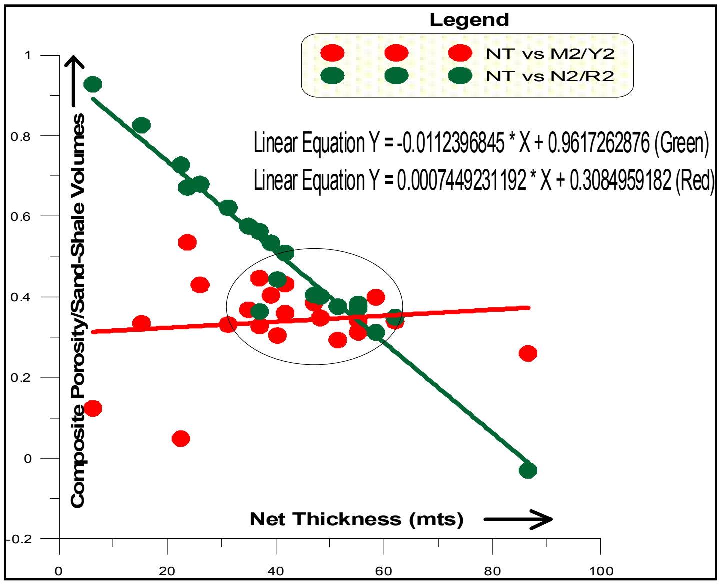


Figure 5.105b: Constructing the data models using clusters of composite porosities

Associativity, density and orientation of clusters, as shown in Figures 5.104a – 5.104h and 5.105a – 5.105b are green and red colored bubbles. They are represented in encircled views, interpreted to be in groups of clusters, each cluster firming up the following equation:

(NP – GP)/ (NP + GP) = (SHVF-SDVF)/ (SHVF + SDVF)

The derived equation from this graph is useful for reservoir engineers for computing better reservoir quality areas of the formation. The sand volumes and porosities appear to be forming a cluster within a net reservoir thickness ranging 35-65mts. This process demonstrates that once net-porosities (NP) and gross-porosities (GP) and shale volume fraction (SHVF) are known, the sand volume fractions can be computed from the above equation. The porosity attributes clustered in red color are corroborating with sand volume fractions, which appear to have better correlation within the interpreted net reservoir thickness range.

Description of clustered data

The basic idea of clustering petroleum business data in terms of building knowledge from petroleum systems is to integrate and merge similar property of oilplay factors, which is otherwise unintelligible. The benefits of this approach include interoperability and knowledge reuse. The clustering, in the present context is broadly described as identification of similar property or characteristics of petroleum business data instances and investigate their relationships.

Significance of cluster petroleum data mining

The pattern recognition, artificial intelligence and cybernetics are using the concepts of cluster analysis in the Engineering Sciences (Fraley and Raftery 1998 and Han et al. 2001).Typical examples include handwritten characters, samples of speech, fingerprints and pictures. In the life sciences, especially, biology, botany, zoology, entomology, cytology, microbiology, objects of analysis are life forms such as plants, animals and insects. The cluster analysis ranges from developing complete taxonomies to classification of species into subspecies. The cluster analysis is also used in information analysis, policy and decision sciences. The author attempts to use these ideas in the fields of petroleum engineering and petroleum systems. Nonhierarchical data structures are also suitable for representing various clusters. The centroids and minimum distances are key parameters and values, and thus criteria for exploring and exploiting clusters for maximum knowledge. Various hierarchical data types are used for doing cluster analysis. In the present context, several data views (taken from warehouse) represented as bubbles of clusters, are explored for finding similarity of patterns among petroleum systems data that will further explore the prospectivity of a basin.

The basic idea of this approach is to build conceptual models (Khatri et al. 2004) on a finer scale (in space-temporal domains) and store them intelligently in warehouse environment for investigating fine-grained clusters of petroleum data and accessing cluster views (e.g. clustering of similar reservoir and structural characteristics from several fields and basins, cluster of similar surveys attributed to drillable exploratory well) for interpretation purposes. Clustering algorithm attempts to find natural groupings among exploration and production data based on some similarity. This algorithm finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distances between a point and the cluster centroids. The output from the clustering algorithm is a statistical description of the cluster centroids with number of data in each cluster.

Another example, with clustered anomalies plotted in Figure 5.106, shows arrows indicating within encircled data clusters, suggesting complex structure trends and the petroleum occurrences within the entrapped structures (geological).

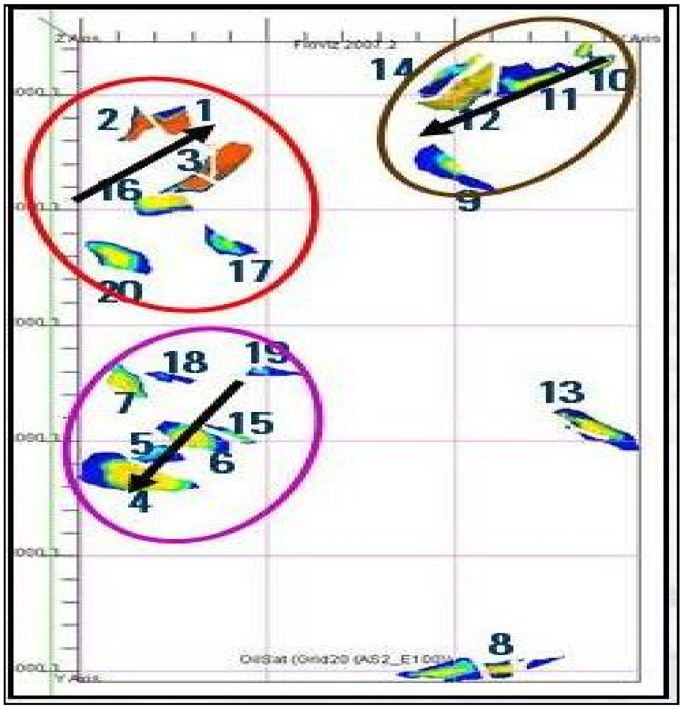


Figure 5.106: Data views drawn to show geological *structure* and *oil saturation anomaly* trends, narrating complex geology (number represents number of prospects and in the order their priority, to risk minimize exploration)

Data mining, visualization and implementation of integrated data schemas (in the cuboid structures)

The dimension hierarchy facilitates viewing the multidimensional data in several data cube representations (Coronel et al. 2011 and Pujari 2002). Conceptually, multidimensional data are viewed as lattice of cuboids. An n-dimensional data cube, C [A1, A2… An], is a database with n dimensions as A1 A2… and An, each of which represents a theme and contains │Ai│, number of distinct elements in the dimension

Ai. There are many data cells in the data cube, in which each cell, C [a1, a2, …, an] stores the numeric measures of the data for Ai=ai. Thus, a data cell corresponds to an instantiation of all dimensions. In the following, C [dip, strike, vertical] is the data cube, and a data cell is C [inline, cross line, time /depth] stores number of instances and associated measures. There are hundreds of in-lines and cross-lines (of a typical seismic acquisition campaign) and for each in-line and cross-line, could be hundreds of CDP points (Telford et al. 1990 and Parasnis 1997). For each CDP point, C [inline CDP, cross line CDP, time/depth], there is unique value of [easting, northing, time/depth] value. In the case of horizon or seismic wavelet in horizon domain, a typical cell C [easting, northing, and peak/trough] has instance, for a seismic time or depth cube, thousands of cells have data cell instances. Each dimension has unique set of rows and column instances, having connections among peak and trough data instances. Interestingly, in-line, cross-line dimensions have relationships and their properties (either peak or trough or seismic wavelets of horizons) are linked through instances of easting and northing. As shown in Figures 5.107a – 5.107d, several cube representations are demonstrated, to present the data mining views for in-line and cross-lines, in terms of surface map views for interpretation. In-line and cross-line data visualizations are extracted for interpreting various properties of horizons, such as structure (geological structure). Multidimensional data models possess summary measure, summary function, dimension and dimension hierarchy, the basic conceptual components. A measure value is computed for a given cell by aggregating the data corresponding to the respective dimension-value that sets and defines the cell. The measures are categorized into different groups based on the kind of aggregate function used.

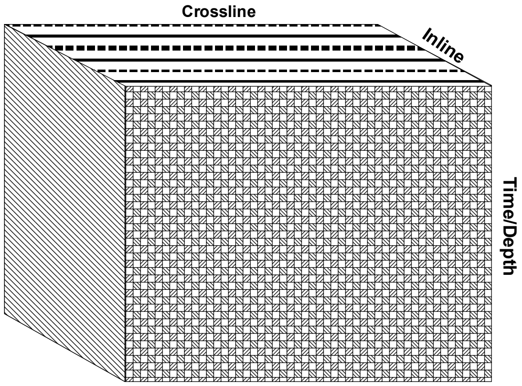


Figure 5.107a: Cross-line/Inline data dimensions their relationships with seismic time or geologic depth dimensions

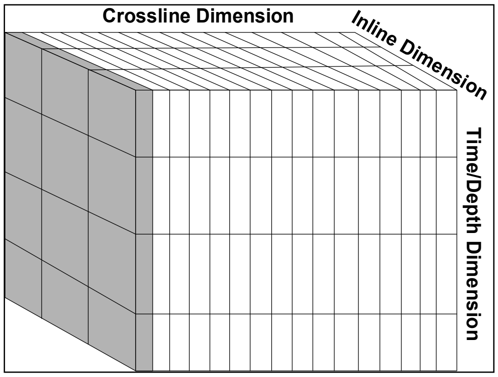


Figure 5.107b: An inline dimension with respect to time/depth dimension

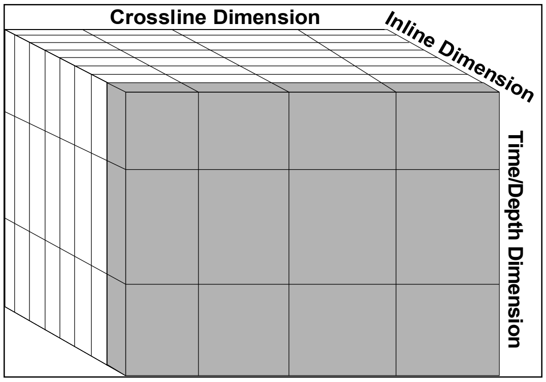


Figure 5.107c: A Cross line dimension with respect to time/depth dimension

The data visualization (Plastria et al. 2008 and Pujari 2002) is an important stage for creating and representing rich and compelling visualizations to explain the multiple dimensions and connections within broader (global) integrated workflows. Visualizations affecting the data integration process and each dimension (within an integrated workflow) contributing to this process, have a bearing on the strength and depth of interpretation. In the current scenarios, geological (structure and reservoir), geophysical dimensions (their attributes) and petrophysical dimensions and their attributes, used for interpretation, are presented with several visualizations.

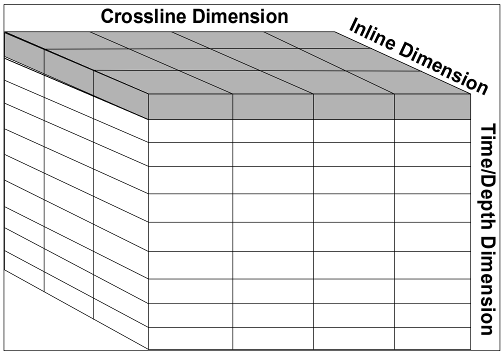


Figure 5.107d: Time slice from a cube

The author uses workstation graphics for generating visualizations and their interpretations. Warehoused data views led interpretation of geological structure attributes and their connectivity among different fields. Different horizons are interpreted from volumes of seismic data (connecting several wavelets, with peak and trough dimensions). The seismic in-lines are plotted with horizons posted on them, showing up several seismic peaks and troughs. Geo-spatial visualizations are often representative, when easting and northing of data properties are drawn, thus understanding the connections among data characteristics makes easy. The connectivity between seismic (horizons) and well data attributes are made by data integration process. *Structure* (geological) is a typical data characteristic feature, measured for different horizons as illustrated in visualizations shown in Figures 5.108a and 5.108b. The connectivity among structures is visualized as characterized in Figure 5.108c. The author plots several dimensions with *structure* property instances, extracted from warehoused metadata, showing connectivity among fields (seismic map views).

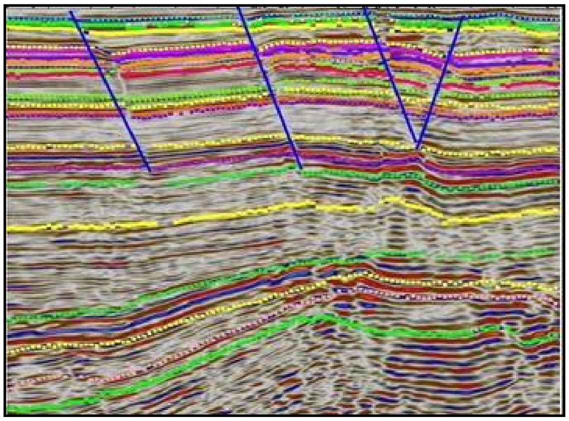


Figure 5.108a: An inline seismic section, showing *horizon* attributes

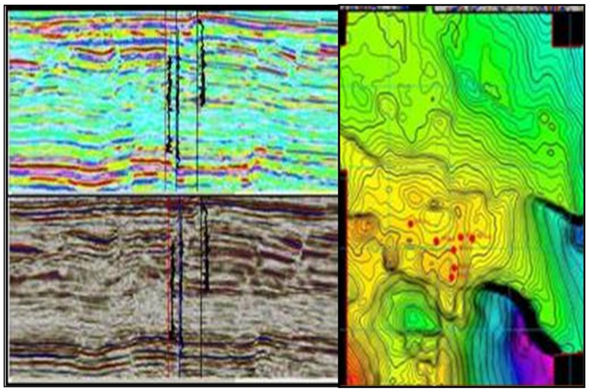


Figure 5.108b: Typical seismic map views (after data integration)

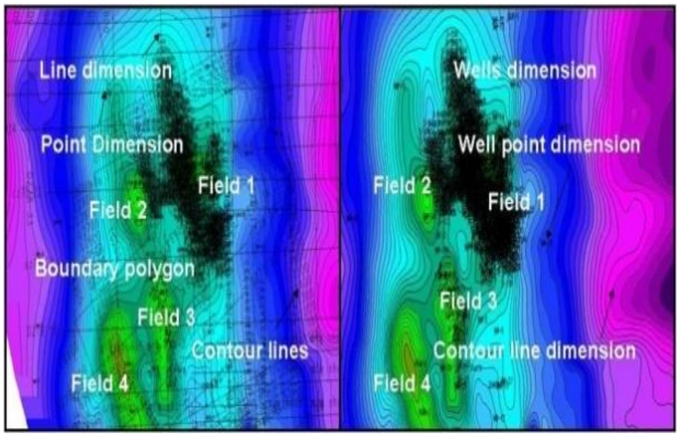


Figure 5.108c: Several dimensions plotted with structure property instances, extracted from warehoused metadata, showing connectivity among fields (seismic map views)

5.5 Summary

As per research objectives (RO 1 to RO 5) described in Section 1.3.2 in Chapter 1, the author interprets domain knowledge and its implementation in industry scenarios. Domain ontology models accommodated within a warehouse environment for integration, arrive at a metadata structure that enables to interpret domain knowledge in different applications. The domain knowledge and its interpretation have immense application, scope and implementation in oil and gas exploration industries. The economic implications are enormous, when different map views are superimposed and interpreted for precise exploratory drilling campaigns including expensive and risky field development projects.

Chapter 6: Evaluation of the Integrated Research Framework and Implementation

6.0 Introduction

As a part of implementation, the author presents implementation models in Chapter 4 (Figures 4.53 and 4.54), because of the context came up with “an exploration project” as a case study highlighting the implementation. In addition, in Chapter 5, all the map views (structure map views presented in Figures 5.99 and 5.100) extracted are ready for implementation. These maps are the basis for exploratory drilling campaigns. Similar models are computed for different basins in Australia, Indonesia, Uganda and Middle East. To this extent, constructs, models and methodologies along with the articulated integrated framework are ready for implementation in an industry arena. New constructs, models and methodologies used in an integrated framework (that address research objectives RO1 to RO8) are evaluated in the contexts of different sedimentary basins and their ecosystems scenarios. The author implements domain ontologies and their structures, designed and developed, in different sedimentary basins of Australia, Indonesia, Uganda and Middle East, as described in Chapters 4 and 5. For this purpose, map views drawn from metadata, are judiciously and successfully used for extracting new knowledge.

For the purpose of validating the design science IS research guidelines, the author evaluates the data warehousing design and data mining processes and procedures. Especially the context of warehouse design, using heterogeneous and multidimensional data, a bench mark is assessed, even though, there is no standard measurement for data warehouse performance that is examined. The benchmark is how successful the metadata structure and their data and map views yield required oil and gas domain knowledge interpretation that further facilitate exploration and drilling campaigns. As a design science guideline, the models deduced are analysed and evaluated in different domains, to establish the validity of framework. The performance varies according to the hardware and software environments used. Database size affects data warehouse performance. In the present study, ontology based heterogeneous and multidimensional data warehouse repository is designed and its performance is evaluated in terms of testing it, using live data from different domains. As described in Chapters 4 and 5, several data views interpreted, provide useful new knowledge for drillable new targets or wells.

Several schemas, relevant to oil and gas business environment considered, are not optimum and flexible in the other domains. For example, as per domain situations of exploration, drilling, production, the models are modified using the semantically derived dimensions and attributes. The design and development of a data warehouse and its application for testing, are evaluating factors, satisfying the research questions and objectives. Here is an outcome of the analysis:

- Integrated data warehouse and data mining implementation in industries.
- Standardisation of data warehousing and data mining technologies, saving enormous amount of time and money in the resources and energy Industries.
- Consistency, reliability, integrity, and quality of oil and gas data/information are ensured.
- These technologies help the CEOS as decision support systems for future planning and forecast of resources in energy and resources industries.

The author completes the initial data gathering, design and mapping, loading and testing, building and testing tasks of data warehousing. Rollout and feedback from end users still remain to be carried out, though quite good reception obtained from current research outcomes from producing and service companies worldwide.

6.1 Review of Research Questions with Respect to the Integrated Framework

Fast tracking and developing an infrastructure for accessing accurate and precise data from heterogeneous and multidimensional data sources in the oil and gas business domain are current focus and the author achieves the following solutions:

1. The research methodologies simplify the complexity of data models by representing in various schemas and categorizing them in geographic and periodic dimensions.
2. Ontologies make the structuring process easy and enable to represent the data models in multiple dimensions. Mapping multidimensional data in multiple domains and applications has become easy.
3. Ontology structures are made use of and reuse in different domains and applications.
4. Semantics, schematic, syntactic and system heterogeneities though addressed in many knowledge domains in oil and gas business environment, but still there

is an application and scope of analysing these heterogeneities in many other domains.

5. An integration process through warehousing environment has simplified the heterogeneous and multidimensional data integration, when the data are captured from multiple domains and applications.
6. Data and information sharing among users of multiple domains, has become easy without losing credibility, security and integrity of data and information
7. The knowledge built among volumes of historical data sources has become easy for interpretation, when the data comprise of multiple dimensions and domains. The data relationships that are hidden and undiscovered among these data sources are easy to explore connections through integrated methodological workflows.
8. The applicability and feasibility of data warehouse, supported by ontologies, combined with data mining, data visualization, data interpretation, have made huge impact in terms of the economics of exploration and production business of oil and gas industry. Figures 4.53 and 4.54 are good demonstration of addressing research question RQ8 and research objective RO8, in which a drillable exploratory well is risk minimized by pinpointing a new opportunity/prospect. There are many other data and map views used for domain knowledge and interpretation of new drillable targets in many basins.

Keeping in view these observations, the author evaluates the following research objectives that are achieved:

Ontology objectives (RO1) – developed ontologies for complex heterogeneous data dimensions, creating knowledge based structures addressing semantic information and rules/axioms.

Data warehouse objective (RO2) – domain ontology structures integrated within a warehouse environment. Metadata created in an Oracle environment is a basic input for computing mining, visualization and interpretation of data models.

Data mining objective (RO3) – computed different mining models and data views for the purpose of analyzing and interpreting domain knowledge. Ontologically structured data warehouse is the basic input for achieving this objective.

Data visualization objective (RO4) – data views computed through data mining represented in different visualizations that explored domain knowledge for interpretation. Data views presented in different plot, maps, cross-plots in 2D and 3D views, are the basic inputs for good quality data interpretation that provides new knowledge.

Data interpretation objective (RO5) – the data models presented in different visualizations, interpreted for qualitative and quantitative information for extracting knowledge and use of this new knowledge in different commercial and financial ventures of oil and gas business. The domain knowledge and models deduced, have economic implications especially in optimizing the economics of exploratory and field development drilling campaigns.

Ontology based data warehousing and mining (RO6), is the solution keeping in view the heterogeneity and multidimensionality of data sources in oil and gas business environments.

Petroleum digital ecosystems and digital oil field solutions

Petroleum Digital Ecosystems (PDE) and Digital Oil Field Solutions (RO7) – the author took the above specific research objectives advantage of designing and developing petroleum digital ecosystems and digital oil field solutions in oil and gas exploration industry scenarios. These ideas are new and have further application and scope of extending them in many sedimentary basins worldwide. These ideas create more exploration opportunities for many oil and gas explorers, in the form of new investments, making their industries more economical and financially viable.

Exploration risk minimization objectives in exploration and field development campaigns

Interpretation of multidimensional data views from warehoused metadata is expected to add value to knowledge domains of PDE and Digital Oil Field Solutions. Interpretative solutions can optimize the exploration risks (RO8) and their associated economics of exploratory drilling campaigns including expensive oil and gas field development projects.

Validation of DS research guidelines:

1. New and innovative artefacts are created as discussed in Chapters 1, 2 and 3.

2. The new artifacts address the research problem domain, as described in Chapter 2.
3. The artifacts yielded solutions for research problems, as demonstrated in Chapter 3.
4. The artifacts ensure their consistency and coherency among multiple domains and applications as demonstrated in Chapters 4 and 5.
5. The models presented are generic and they can be applied in any application domain. More details are given in Chapter 4 and Chapter 5.

The author describes merits and demerits of the methodologies in the following sections.

Issues of design and development of an integrated framework and workflows comprising of data models, schemas, data warehouse, data mining, visualization and interpretation and their solutions, transform the operational databases into decision support systems. The author uses historical data, comprising of periodic and geographic dimensions of operational systems of the resources businesses, for analysing the decision support data systems. Identified all the entities, dimensions, relationships, attributes and their synonyms. The conceptual models and entity- and dimension-attribute worksheets have been produced for resources companies and big data associated industries including health-care and environment management systems. Multidimensional ER models and worksheets have been generated and successfully used the listing of all entities, dimensions, objects and relationships among them, the relationship cardinalities and optionality. Data views representing Excel-produced datasets and ASCII files have been successfully used for drawing correlations, trends and patterns from multidimensional metadata. Classic data mining models have been generated for forecasting of future resources in the mineral, oil and gas business environments. The author successfully uses grapher and surfer solutions for drawing models, mining, visualization and interpretation for extracting the new knowledge.

6.2 Comparison of Data Models

The author analyses different modelling approaches to evaluate for multidimensionality and heterogeneity. The entity relationship (ER) modelling method is initially used to depict the entities of heterogeneous data in their graphical representation (narration of ontologies) and how best they are compatible and integrating them in a warehouse environment. Exploration and production (E & P) data sources in space and time domains are heterogeneous and multidimensional, posing semantic, schematic and system inconsistencies (as addressed in research problem and questions). ER (entity-relationship) and E^{2}R (extended entity-relationship) modelling approaches have been used for building structure models and integrating them in a warehousing environment. Because of semantics, schematic, syntactic and system heterogeneities that complicate the ER and E^{2}R structuring process of G & G and E & P data sources, the author used judiciously a multidimensional approach in the modelling. The author evaluated ER, E^{2}R and multidimensional approaches for an effective data mining, visualization and interpretation as demonstrated in Chapters 4 and 5, suggesting heterogeneous data sources when translated into multidimensional data structures, yielded better domain knowledge for interpretation and risk evaluation of petroleum prospects. The multidimensional approach is more accommodative and compatible to warehousing compared to ER mapping.

6.3 Implementation of the Integrated Framework

The integrated framework provides scopes and opportunities for revising and constructing new constructs, models for evaluations and justifications. During implementation of the data warehouse, the data warehouse is further evaluated, in terms of performances that check the reasonableness of the response times. This implies:

- Load up the databases in the warehouse with set of records to the volumes expected in a production environment
- Simulate the peak number of simultaneous users in a multi-user database environment
- Run a mix of the most common time-critical functions

Resultant efficiency considerations are:

- Data access from volume of databases loaded into warehouses in Oracle environment, response times are fast and easy to access. If the response times are too slow, a remedial action needs to be taken.
- Present studies suggest that appropriate indexing and denormalization procedures improve the efficiency of accessing the data from warehouses.

- Storing many different codes and code descriptions in one common reference table, if considered, the application could read this table as a start-up menu and use the data for as long as it continues to run, without having to continually access the table.
- When performing table joins, it is more efficient to filter out the required subset of data from each table, and then perform the joins than joining two large tables, then filter out the subset of data.
- Separating the current and historical transactions might help the efficiency of loading and accessing the wanted transactions more quickly and efficiently.
- Data presentation in the form of correlation, trends, and bubbles has been very efficient and effective, especially while interpreting the resources data. Clustering of multidimensional data and presenting in different graphic images help in interpreting and evaluating the data more explicitly.
- The benefits received from the data mining studies, are more than the cost of the data mining costs.
- The results of data mining approaches can confidently be used for interpretation of resources data. The available software and hardware used have successfully made possible to prepare data models and establish the utility of data warehouse in the resources organization.
- The information and data that have been extracted from the statistical models and interpreted datasets are very knowledgeable and intelligent enough to predict the future resources information in the resources industry. Bubble plots could improve the interpretation of multidimensional data in many basins.
- As stated earlier, missing values could create problems in the data organizations. Missing values have been intelligently trained and conditioned the data tables appropriately. Interpretation of the resources warehoused data has definitely been improved through intelligent conditioning of the data.
- Evaluation of the written documentation to ensure that the documentation and procedures are accurate and easy to follow.
- Observance of standards for naming, documenting and coding
- Data duplication conflicts with existing data and enforcement of all data validation rules.
- SQL queries, user and system files have to be regularly monitored and updated so that performance of use and usage of the resources databases can be improved.

The innovative artifacts, such as constructs, models and methods associated with petroleum digital ecosystems, digital oil field solutions, big data systems, integrated framework comprising of ontology based data warehousing of mining heterogeneous and multidimensional data sources, are solutions for the problem statement, discussed in Chapter 1. As discussed in Chapter 5, analysis and implementation innovations ensure an effective and efficient problem solution evaluation. The models and methodologies used in oil and gas domain, are made generic, so that ensuring in other domains, their coherency and consistency. Interpretation of domain knowledge and analysis presented in Chapters 4 and 5, describe innovativeness of artifacts, as designed in Chapter 3.

The data warehouse approach brings together petroleum systems data from different sedimentary basins of different depositional and geological regimes along with data from other geophysical and geochemical data warehouses. The data warehouse approach is used to benchmark and track the effectiveness of petroleum system productivity over time. It also allows processed (knowledge based) data shared among professionals, geographically distributed worldwide. The need to integrate petroleum systems data from multiple systems and sources is well known (Magoom and Dow 1994 and Hoffer et al. 2005). It is important for data warehouse designers to define the scope, depth, comparability and accuracy of data entering the warehouse. The scope of data refers petroleum systems data, sedimentary basins, data, geological, geophysical and geo-chemical data from multiple periods (time dimension), geographic locations (space dimension). Depth of data refers to the level of details. To be comparable, data from multiple dimensions and different sites should adopt the same classifications, as much as possible. No matter how differently data are collected across sites, they are significantly altered for integration before moving into the data warehouse. To reduce the burden of alteration, it is important for petroleum systems analysts and geo-modellers to use compatible software systems to acquire and send data to the repositories. It is also important to standardize the data collection processes. An accuracy is desired for all types of data in any given situation and this is fundamental requirement for reliable use of the data or even a construct/model.

6.4 Overall Research Outcomes

Building, evaluating, theorizing and justification are different research activities (RO1 to RO8) analysed for evaluating the research outcomes, such as concepts/constructs, models, methodologies and implementation within an integrated framework, in the oil and gas domain. The author summarizes them in the following sections.

6.4.1 Conceptual models

Several conceptual models built in Chapters 4 and 5 are inputs resultant to building logical structures. ER, EER, MR and MMR including object oriented models are built keeping in view the current domain knowledge. Business rules and constraints imposed on these models satisfy the business conditions.

6.4.2 Entity relationship (ER) and multidimensional relationships (MR) data models

The entity and multidimensional data models, used in Chapters 4 and 5, are relevant to the oil and gas industries for building fine grained physical data models for warehousing and mining. Various entities, attributes and relationships appeared in the volume of oil and gas data, utilize multidimensional designs. The multidimensionality and granularity have been maintained throughout this modelling process, while addressing heterogeneity. Missing values have been intelligently interpolated wherever necessary to make the data fit to the model. The information processed from the warehoused data has been evaluated for its utilization in the oil and gas industry. Analysis and interpretation of computed data infer that good quality data views generated support intelligent decisions made by top management of the resources organization. Multidimensional ER models with star schemas created for surveys, wells and permits data provide, good quality queries as described in the previous chapters. Multidimensionality and granularity of the Petro2 databases allow producing good quality and fine grained data views from the warehoused metadata. Oil and gas engineers, scientists, technicians and managers involved in the exploration and development of mineral and petroleum resources are contented with these analyses. If the data warehousing successfully applied and implemented, there is definite impact on exploration, drilling and production departments especially while sharing required information and data and also when making future forecast of inputs for field parties and oil drill site projects.

The present studies suggest that oil and gas industries have been very optimistic and cut over-head costs in the exploration and drilling activities. Interpretation of statistical models and their usage in the oil and gas industries, in general, provide good quality business intelligence to support exploration, drilling and production business operations.

6.4.3 Integrated methodological framework

Domain, data modelling, schema, data warehousing and mining with visualization and interpretation, all used in a single canvas in an integrated framework, is a good demonstration as given in Chapters 4 and 5. As described in Chapters 4 and 5, exploration of oil and gas is given more emphasis. As described in Chapter 3, the integrated framework and its utility are meant to be extended in different domains in other domains such as drilling, production and marketing.

6.4.4 Implementation of models

The models built based on the secondary data, deduce various statistical correlations, trends, and patterns among oil and gas data sources and their interpretation for new knowledge, is a demonstration of implementation. The data instances or values of attributes used in the interpretation and analysis vary depending upon the system and its connectivity with other ecosystems. Figures 4.53 and 4.54 discussed in Chapter 4 are good examples of implementation. The data structures made generic in many domains in spite of that the operational data vary in different geological situations, for example. An interoperability is achieved through integrated implementation process.

6.4.5 Petroleum digital ecosystems

The author has evaluated the research objectives, RO7 and RO8, as cited in Sections 1.3.1 and 1.3.2 in Chapter 1 and successfully used Australian, Indonesian, Ugandan and Arabian Gulf sedimentary basins and their data sources for understanding the phenomena of the ecosystem in petroleum domain. The integrated framework, which is simulated with a sedimentary basin, provides a metadata structure, for analysis and evaluation. The integrated framework successfully explores the connections among different elements, processes and chains of the different petroleum systems existing in a sedimentary basin. The petroleum digital ecosystem has more commercial value, since multiple oil and gas fields are driven by successful presence of elements and processes of different petroleum systems. Commercial values are exploited and from which, explorers and well planners know where and how much to invest in exploration and subsequent field development.

6.4.6 Digital oil field solutions

As explained in Figures 3.37 and 3.38 in Chapter 3, an integrated framework is a digital oil field solution. The metadata evolved from the process of integrated methodological framework is a research outcome, is the ultimate research deliverable for exploration and development investment. The author evaluated as a part of RQ 7 and RQ 8 and RO 7 and RO 8, mentioned in Sections 1.3.1 and 1.3.2 of Chapter 1, economic benefits (RQ 8 and RO 8) of the research work in the following sections.

Exploring connections among knowledge domains of oil and gas data sources, is one of primary goals of petroleum digital ecosystems. Integration of data sources associated with *onshore-transition-zone-offshore* areas explores connections among structures and reservoirs among multiple oil and gas fields in a sedimentary basin.

The connectivity and integration of geological events associated with onshore and offshore exploration is the most difficult and challenging part of any exploration and development project of oil and gas business. An understanding of the data relationships among these data sources need great attention, especially in the context of minimizing risks of expensive oil and gas exploration including field development campaigns in the *onshore*, *offshore* and *transition-zone* (overlapping areas between onshore and offshore) areas. Various aspects of data relationships, such as the degree of relationships, roles of entities or dimensions participating in the relationships, structural features such as the type of relationship and mapping constraints considered in the integration process enables data analysts, oil and gas explorers and energy researchers. As shown in Figure 6.1 and Figure 6.2, several oil and gas fields from *onshore-transition zones-offshore basins* are organized, addressing data integration and interoperability issues. The concept of a digital ecosystem in petroleum system domain, facilitates interconnectivity and integration in a warehouse environment, further enabling data mining, data visualization and interpretation more effectively and efficiently. Several data views drawn from ontologically structured metadata improvise interpretation and knowledge discovery (map views shown in Figure 6.1) among oil and gas data sources of undiscovered fields.

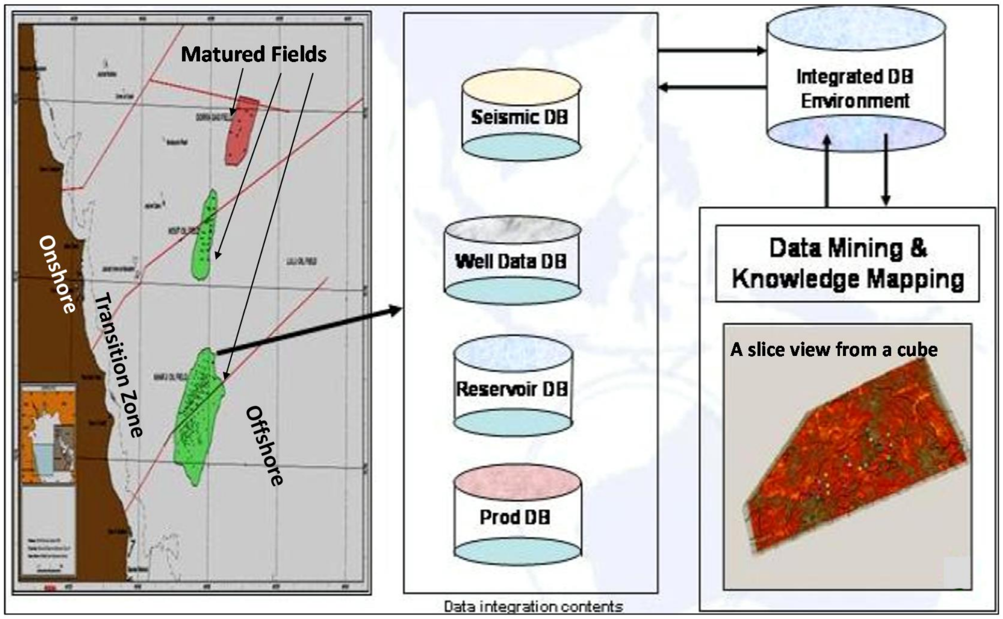


Figure 6.1: Integrated framework - connecting Middle Eastern offshore oil and gas data sources

As an example, in one of the Middle Eastern onshore and offshore productive areas, the connectivity among multiple petroleum systems is established through ontology based warehousing of mining heterogeneous and multidimensional data sources. These data sources are either from exploration or production or combined entities. Several seismic data sources considered in the modelling process as shown in Figure 6.2 are ontologically structured and connected among other nearby fields and their associated petroleum systems. The author has an opportunity to use and reuse ontology structures in different fields and their associated ecosystems, with minor changes of attribute naming conventions, instead of creating separate ontology structures for individual fields. As an example, several matured producing oil and gas fields, existing in the *Middle Eastern sedimentary basins* provide improved understanding of the prospectivity and production enhancements. How digital ecosystems phenomena facilitate connecting multiple matured oil and gas fields of onshore, offshore and transition zone areas, is another an innovative concept of the current research.

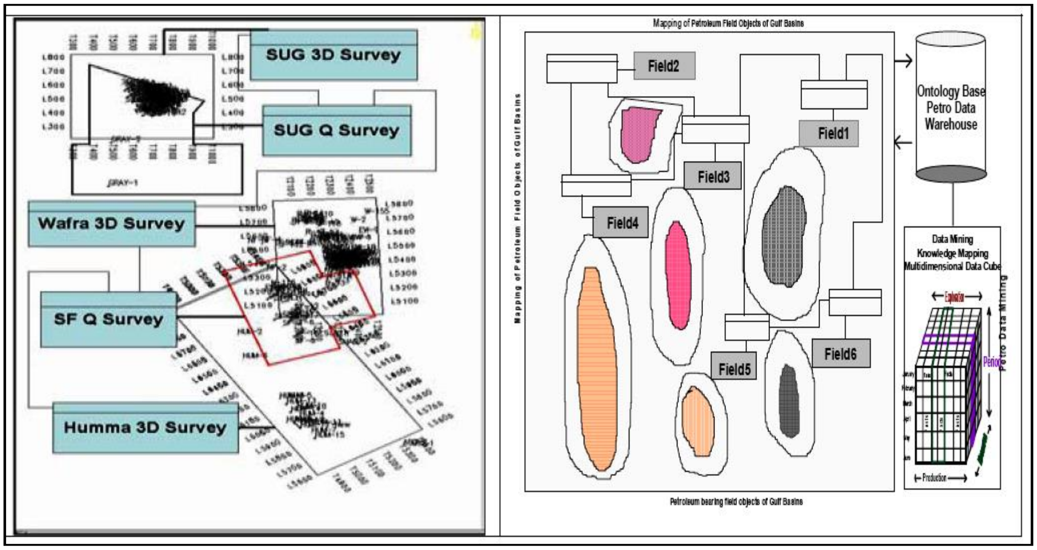


Figure 6.2: Interoperability in the oil and gas data integration process

In addition, in complex heterogeneous and multidimensional data sources environment, graphics and visualization provide effective means of communication because of highly developed 2D and 3D seismic pattern-recognition capabilities that allow processing and perceiving the pictorial data instantly and efficiently. The author successfully summarized data from different sources, for analyzing the trends and correlations for interpretation, as an example demonstrated in Figure 6.3. An unknown phenomena uncover through various kinds of graphical representations of multidimensional datasets.

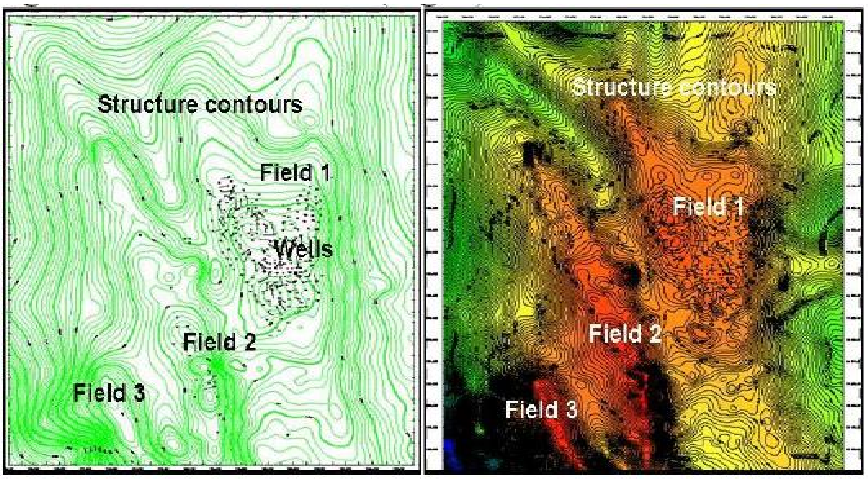


Figure 6.3: Establishing connectivity among multiple oil and gas fields

Bubble plots (Figure 6.4) convert pages of *hard-to-understand* numerical and textual data into something that is easily comprehensible to analyze. The bubble plot is a simple example of using graphics to quickly convey information about the data that they can present. Bubble plots, representing bubble sizes, densities and trends suggest several inferences such as *structure*, *reservoir* and *production* attributes and their strengths and magnitudes that interpret qualitative and quantitative characteristics of attributes.

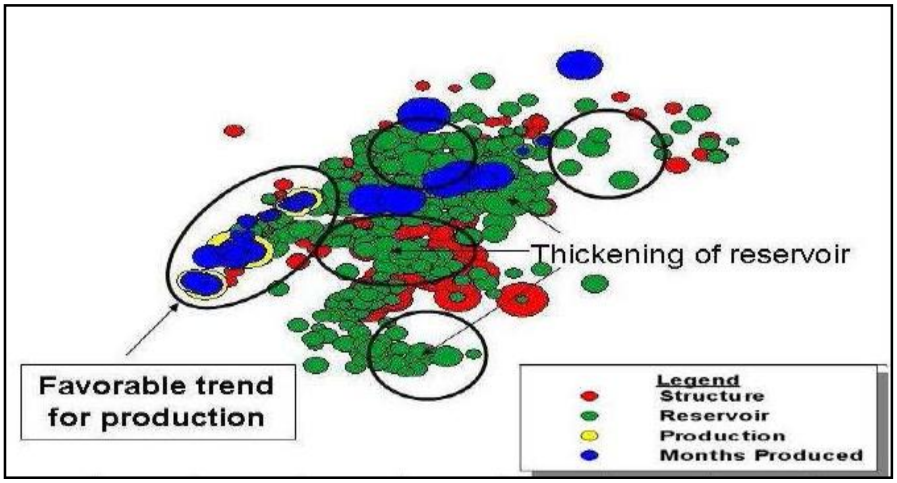


Figure 6.4: Representing multiple attribute dimensions on the same graph

As a part of implementation and evaluation of the methodologies used in the present study, author interprets several plot, map and other graphic views and extract knowledge, for evaluating the effectiveness of integrated research framework and the data models designed in different application domains (Chapter 4 and Chapter 5). Data interpretation tests the validity, effectiveness and flexibility of the data models, data warehousing and mining procedures including effectiveness of visualization. Qualitatively, the trends, patterns and correlations observed (as discussed in Chapters 4 and 5) among data events are interpreted for understanding knowledge enhancements and evaluations.

Ontology based warehouse modelling, mining, visualization and data interpretation, all in combination in a single canvas in an integrated framework, is advantageous in designing and developing digital ecosystems in petroleum domain. The digital data of a petroleum system or number of petroleum systems, existing in a *sedimentary basin*, all simulated within an integrated warehouse scenario, is considered as a *digital oil field solution*. How the *elements*, *processes* and *chains* (of an information system or a petroleum system) interacted and communicated each other among multiple systems, is another innovative concept in the design and development of a petroleum digital ecosystem (PDE). Petroleum digital ecosystem and digital oil field solutions' are research outcomes, can change the economics of the exploratory drilling campaigns.

6.4.7 Evaluation of research framework as per DS guidelines

As described in a research framework in Figure 3.1, the research activities and research outcomes are discussed in the context of a comprehensive research methodology. As highlighted in Venable et al. (2014), the evaluation of an integrated framework constituting the constructs, models and methods is performed by implementing them in the upstream oil & gas business strategies. The guidelines evaluated in the DSR, are summarized in the following sections.

Guideline 1: Design as an Artifact

Integrated research framework and methodology satisfy the design science guideline 1. For this purpose, new constructs, models and methodologies have been made in Chapters 3 and 4. This guideline has been analysed in the oil and gas business domain in Chapter 5.

Guideline 2: Problem Relevance

As defined in Sections 1.3.1 and 1.3.2 in Chapter 1, the author describes the existing issues with organization of heterogeneous and multidimensional data sources, research questions and objectives. Author discusses existing literature in Chapter 2. The author demonstrated how the new models and methodologies described in Chapters 3 address the research problems. The research questions and objectives (RQ 1 – RQ 8; RO 1 – RO 8 given in 1.3.1 and 1.3.2 in Chapter 1) evaluated in 6.1, highlights this guideline.

Guideline 3: Design Evaluation

Successful integration of domain ontologies and their structures in a warehouse environment is significant criteria achieved, ensuring a knowledge base solution to the research problem as described in Chapter 1. Various constructs, models and methods used in the study are evaluated.

Guideline 4: Research Contributions

This guideline highlights and emphasizes the scope of implementation and future research, keeping in view the research framework and methodology in the context of oil and gas business domain. There is an immense scope of using similar models and methodologies including implementations in many other domains of research.

Guideline 5: Research Rigor

The author achieves the rigor on research by evaluating the models and methodologies as demonstrated in Chapters 4 and 5. Interpretation and analysis of ambiguities, inconsistencies, heterogeneities and multidimensionality suggest much needed rigor is done in researching and evaluating the methodologies.

Guideline 6: Design as a Search Process

Integration, making and exploring connections among heterogeneous and multidimensional data structures facilitate sharing and accessing the fine-grained user and data views as demonstrated in Chapters 4 and 5.

Guideline 7: Communication of Research

The most significant part of the research is dissemination of new knowledge and communication to variety of users and domain experts. The current research framework and methodologies discussed in Chapter 3, though focuses in oil and gas business domain, but further generates research opportunities in other domains. Author documents well all the research ideas and scopes in a way, they are well perceptive and knowledge building to variety of researchers and oil & gas explorers. The presented integrated research methodology is comprehensive and expected to attract variety of researchers in many of domains of research.

6.5 Summary

System development projects produce results from multidisciplinary teams, such as users, designers, system analysts and programming specialists. Data used from multiple domains and for each domain, the systems development procedure is similar and also the data structures are analogous with flexibilities. Solving problems associated with datasets and fixing errors, missing and redundant data issues in the initial stages of project, save enormous amount of time and money, especially during implementation stage. Systems and solutions need to be examined carefully, ensuring that they are appropriate and optimum to business solutions. Feasibility and applicability analysis is must for safeguarding strategic interests of the organization, including economic, technical and behavioural impacts. Testing is the process of checking the data models within each domain, whether the models achieved desired results within the assigned business constraints. The domain, data modelling and type of schema chosen in the modelling process could ultimately impact the implementation.

Chapter 7: Summary of the Thesis

The design science research paradigm is used, based on which an integrated framework is designed and implemented in an oil and gas exploration project. Several components of the integrated framework are domain, data modelling, schema, data warehouse, data mining, visualization and interpretation modelling. The conceptual, logical and physical models are generated as artifacts. The author uses oil and gas application domain, for which oil and gas (including mineral exploration) data acquired from various sources are organized for integrating in a way acceptable to the database design, development principles and analysis requirements. Conceptual models and multidimensional ER Models have been drawn for Min1, Min2, Min3, Min4 and Petro1 and Petro2 databases in Oracle 9i environment. As a follow up of research objectives, author develops petroleum digital ecosystems and digital oil field solutions with the integrated warehouse framework articulation. Data views are extracted using SQL (.txt) files, control files (.ctl) and data files (.asc). The queries are created using SQLs and thus data views generated from these databases are interpreted for data trends, patterns and correlations through the principles of data mining and data visualization techniques. The classical statistical mining models are generated for predicting the future forecast of resources to be used in the oil and gas industry, analysing cost benefit analysis, exploration costs in mineral and petroleum industries in terms of cost of exploration, petroleum and mineral discoveries and production. The performance indicators are analysed for a resources industry, from the databases and for making the future predictions.

The present integrated business application is created with the real data, business, information and user focus. The data reliability and validity checks have to be made at various stages of implementation, so that information that is conveyed to the end users has achieved its objectives. Performance indicators of oil and mineral exploration and production business both in space and time have established significant quantitative and qualitative interpretation trends that help the managers to make technical decisions on future forecast and resources planning. It is recommended to analyse the application with live data and establish the efficacy of application. The feedback from industries and business organisations always help to improve the working condition of this integrated application and recommended to implement it in other resources industries. These technologies allow the oil business transactions done for cost-efficient, e-Commerce and seamless B2B integration.

The recent advancements in information technologies and telecommunications, data communications technologies drive the oil companies business electronically too far off places. The data warehouses are back end application of e-business activities. Developments in Internet, extranet, and Intranets are now able to provide better access and interfacing to data and information. XML technologies bridge the gap of exchange of documents. Data that have been mined from oil and gas data warehouses can be exchanged to different operational business areas through XML documents. The cross sectional data collection from different oil companies located in different parts of the world, processing and analysis of data may provide a better picture on trend of performance indicators analysis with respect to global economic and technological trends with time and space dimensions. The XML technology is poised to make integrated data connection a reality in the oil and gas business. A major step toward reaching the XML application in petroleum industry comes from the co-operation of several major company's software vendors that are participating in building the application for their commercial offerings. Standardised documents allow simplicity and cost effectiveness in moving data between disparate workflow systems. Registry services by oil and resource companies on net enable to come together, make business transactions and customer services. All the document exchange procedures are compatible with the data warehousing technologies. ASP (Application Service Providers) technologies are also enjoying rapid acceptance in many petroleum industries. According to International Data Corporation, the ASP market exceeds a turnover of $2 billion by 2003. An ASP leverages the power, speed and convenience that web-browser based solutions can provide. These web interfaces can be linked to data warehouses for accessing the data and information in remote and distant places. This amazing technology has an access to the powerful e-commerce framework that is now forming in the oil and gas industry.

The current research ideas can judicially be implemented in companies and or enterprises that deal with oil and gas business data sources (of heterogeneous and multidimensional scales). If these integrated approaches are wisely applied in oil and gas exploration industries, economic impacts and benefits are huge. Systems analysis especially, ecosystems research analysis and petroleum ecosystems development, if implemented in oil & gas companies, have an enormous impacts in terms of risk minimizing the exploration, appraisal and field development stages, saving billions of dollars.The author promoted the current research ideas, ontologies, data models, schemas, warehousing and mining, visualization and interpretation in many research, industry and educational forums worldwide. The author got good encouragement in terms of design, development and implementation of current research ideas in the industry perspectives.

Chapter 8: The Conclusions and Recommendations

The author has described a research paradigm and research methodologies for developing better procedures for oil and gas domain data constructs. The current study outlines a procedure to develop better methods of modelling heterogeneous and multidimensional oil and gas attribute variables and their instances. The framework represents an attempt to unify and bring together data sources in a single repository from the scattered bits of information on how one explores about developing improved procedures and how one assesses the quality of models. The author suggests the research paradigm and methodologies among various domain constructs. The research findings deduced among cross-sectional domains of oil and gas business are made valid, reliable and the new knowledge generated from the integrated framework is useful among variety of research communities and oil and gas explorers.

The conceptual models are reusable that support model designers. The design science research paradigm delivers a framework to fortify the theoretical foundations of research on conceptual models. The author emphasizes the principle of reusing artifacts, which is acceptable to data modellers and analysts. The current research ensures rigor and relevance in designs of integrated framework for research projects. This research approach appears promising in many other paradigms such as positivist and interpretive research areas. The integrated framework appears to be applicable and feasible for explorers, data management personnel and data analysts, especially when they are involved in other domains of research. The design science guidelines that guided the present research work, seem to be extendable in other domains of research such as health-care, environment and disaster management.

Based on the findings of the current research study and application of data warehousing and data mining studies in the oil and gas business, the author makes the following conclusions and recommendations:

1. Heterogeneous data sources are widely spread worldwide geographically and periodically. A rigor is needed to translate and put the heterogeneous data into multidimensional. It is an enormous challenging task initially, but it fetches in the entire life cycle of the system’s existence.
2. Multidimensional modelling with fine-grained dimension and fact table documentation, is a breakthrough application in the oil and gas industry. This

modelling approach appears to be ideal for designing powerful data warehouses that can handle volume of data and information. Since the resources industry operates both domestically and globally, data warehouse is demanding technology for all project personnel who operate the resources business geographically. It delivers accurate and precise information. Now it is possible to share the required information by all personnel.

3. Petroleum digital ecosystems (PDE) and petroleum information systems are future digital oil field solutions for major producing companies worldwide. These tools and methodologies offer huge economic benefits in the exploration and even at field development stages of oil & gas business.
4. In addition, data integration and integrated workflows can risk minimize the exploration and improved understanding of limits of elements and processes of a petroleum ecosystem.
5. Exploration business in the resources company has already been using various data mining procedures in some form or the other. Various software service providing companies have data mining utilities to meet the needs of the oil and gas companies. Some of the worth mentioning are statistical correlation techniques and neural networks. Data exploration approaches discussed in the present report, definitely add value to the existing data interpretation techniques.
6. The data warehouse can store and provide speedy access of comprehensive and aggregate datasets with guaranteed qualities.
7. Granularity of database design depends on the users' needs and requirements, since it is expensive to build fine-grained database structures and maintain them.
8. The data warehouse has uses other than data mining. However, the fullest use of a data warehouse must include the data mining. Data visualization is quite useful technique in the resources industry for processing the volume of numerical data in to meaningful images. This technique is already in place in oil industry. The ability to create multidimensional structural views and models from the raw data is one of the important innovations in decision support technologies. Data mining (DM) is a powerful technology that converts detail data into competitive intelligence and businesses can use these data knowledge to predict future trends and behaviours. In other words, data mining is one of the fastest business intelligence technologies because it pays off in quantitative values to the resources industry.

9. Volumes of historical data containing facts of the resources industry, business operations are analysed and used to predict what is going to happen in the future.
10. Join indexing facilitates the efficient identification of facts for a specific dimension level and or value. This is key issue for better data mining.
11. If the data warehousing is the hot topic in modern organizations, data mining is one of the hottest topics in the data warehousing. Virtually every organization has embraced the concept of data warehouse believing that data mining is part of its future investment.
12. In complex resource organizations, similar data warehousing strategies need to be implemented to suit even different environmental conditions.
13. Relational database management systems provide computer protocols, SQL, for writing data queries. SQL is a complex protocol, which by its nature makes it unusable by non-technical users seeking a simple means of data access. Data access, though SQL provides structured queries from volume databases, but they are very much constrained, which produce only simple lists of data.
14. Multidimensional analysis is used by a very large number of decision makers for rapid-fire management questioning sequence typical of day-to-day decision-making. Typical questions arise as a result of observed business event or operating performing issues. The number of questions will increase with increase in queries. Multidimensional analysis tools provide answers to unstructured and unpredictable questions that arise every day while managing business.
15. Multidimensional analysis provides a robust set of computational and navigational data capabilities that are far beyond the features of even the most advanced query tool. Multidimensional analysis provides a flexibility to answer questions that are stored in the data warehouse, but is derived from what is stored in the data warehouse. It is nearly impossible to predict the data queries and the analysis path from the user perspective. Therefore a multidimensional analysis tool allows the users to create reports that contain user-defined calculations and multiple layers of subtotaling on the fly. The user will navigate within the report (drill up/down/across) and change report parameters until an answer to particular question emerges. Then more questions will be raised. In summary, query tools simplify data retrieval of information stored in a data warehouse. Multidimensional analysis tools simplify the analytic requirements that underlie decision-making by allowing on-the-fly calculations and summarization of information in the data warehouse.

16. Data visualization is a very successful and widely used technology for viewing the resources hidden under great depths. Data mining is an iterative process, implying that each time it refines the resultant data.
17. Statistical models drawn in the present studies are useful for the resources industry in terms of predicting the future mineral and petroleum exploration costs. Mineral and petroleum production can also be estimated by studying the past performance indicators and expenditure.
18. It is recommended to use the data warehousing and data mining technologies together in the resources industry.
19. Exploration, drilling and production entities have been the focus in the present study. Marketing, human resources, other support services entities can also be included in the present data warehouse system.
20. Several map views provided in the text, are valuable tools for interpretation and implementation.
21. SQL mining and classical statistical mining are quite useful for interpreting the multidimensional data, including geographical and periodic dimensions
22. Heterogeneous data sources located in many company, government and private enterprises, national and multinational situations are useful translating them to multidimensional models.
23. Petroleum digital ecosystems (PDE) and petroleum information systems are digital oil field solutions for all the producing companies including multinational service companies.
24. There is immense scope of extending the current research in many other domains and applications.

Similar studies are recommended in other resources organisations, such as Woodside, Santos, SA Primary Industry and Resources and Victoria Primary Industry and Resources Departments in Australia. The author had an opportunity to present the research work in many operating and service companies, for which, great response received. These organizations handle various types of volume of spatial-temporal resources data that are analysed in the current research. In addition, heterogeneous and multidimensional data of other domains, such as health-care, food industries, CO2 emissions, earthquake prediction analysis, and bush-fire disaster management have immense scope implementing current research ideas.

Other advantages of multidimensional modelling of oil and gas data are:

- Standardized and predictable framework
- Data mining convenience
- Star schema or its extended version accommodates unexpected new data elements
- Easily maintaining the physical data organization
- None of the query or reporting tools need to be reprogrammed for accommodating the new data.

Now a warehoused seismic and well-base metadata repository consists of:

1. Description of structure of the data warehouse. This includes the warehouse schema, view, dimensions, hierarchies and derived data definitions, data locations, and contents;
2. Operational metadata, such as relationships, conceptualizing the dimensions, data units and monitoring information on seismic and well data qualities with updates;
3. The summarization processes which include dimension definition, data on granularity, partitions, summary measures, aggregation and summarizations;
4. Details of data sources, which may include source databases and their contents, gateway descriptions, data partitions, data extractions;
5. Data related to system performance, which include indices and profiles that improve data access and retrieval performances, in addition to rules for timing and scheduling of refresh, update and replication cycles;
6. Business metadata, which include business terms, rules, constraints and definitions, data ownership and changing policies on Environment, Health and Safety (EHS).
7. Ontologies have become very common on WWW. Future work is aimed to develop a WWW Consortium of Petroleum Resources Description Framework (PRDF), a language for encoding knowledge on web pages to make it understandable to electronic agents searching for information. Ontologies of the petroleum systems range from large taxonomies categorizing petroleum systems of NWS (North West Shelf) to categorizations of basins, fields and their features and properties. Many areas now develop standardized ontologies that domain experts can use to share and annotate information in their domain fields.

Having evaluated the data warehouse methodologies and their implementation in the oil and gas industry, future scope of the present study, limitations and constraints, are discussed in the following chapter.

The Limitations and Future Scope

The design science research paradigm used in the current study has future scopes. Integrated framework, which is driven by this research approach, can be generalized in other domain research areas. Warehouse modelling, which can integrate domain ontologies, can intelligently store big data systems such as petroleum ecosystems. Applicability of data warehousing in the oil and gas industry domain has been studied; warehoused metadata have been evaluated using the data mining and data visualization tools. The data correlations, patterns and trends have been interpreted for business intelligence and economic performances in the mining and oil and gas sectors of the resources industry. Investment decisions made by oil and gas companies are generally influenced by global and national economic trends, taxation, production sharing agreements between investors, and capital costs. Other specific issues include exploration mining decisions associated with native title requirements, cultural heritage protection, environment protection, political instabilities.

Implementation of the new research methodology is merely a challenge in its own domain of research. Several domains are considered in the current research in terms of identifying data issues, associated with modelling, warehousing, mining, visualization and interpretation tasks. Some of the issues related to applicability studies of data warehousing and data mining technologies in energy and resources industries are discussed. The company restructuring, takeover, merger and global economic crisis have an impact on the use of these technologies and make strategic decisions. Data warehouses need continuous repair and maintenance. Feedback and evaluation of these technologies are significant issues for a complete implementation. Data collection is covered here by both longitudinal (periodic) and lateral (geographic) designs. If the cross sectional design is combined with longitudinal and lateral designs, qualitative and quantitative interpretations of data views are more meaningful and significant in any domain of research, especially in the resources industry, since it is a billion dollar risky business.

The secondary data are available from various sources in haphazard manner in each domain. At times, data integration is difficult and it is necessary to organize these data sources systematically. The ER, E^{2}R, MR and other multidimensional (relationships) diagrams and models given, are representative of resources industry. The data warehousing is an ideal concept, to use in the resources industry, but at present it is rarely adopted. Training is needed to the personnel involved in the data warehousing design and development. Attitudes of personnel involved in various disciplines such as exploration, drilling, production, technical and administrative services, marketing may have to be tuned to the design and development of data warehousing overall process design. The data collected from the present resources industry are historical (from 1930 onwards), but in other organizations, these data are not available at all. So it is difficult to derive and develop effective statistical models unless they represent volumes of data from cross segmented, different organizations. The author had an opportunity to discuss these issues with the personnel involved in the data management in the Western Australian Department of Industry and Resources and Minerals and Petroleum Branch, Australia, which motivated to take up this research topic.

Multidimensional data and their connectivity from multiple domains and sources, including integration can significantly risk minimize crucial technical and financial decisions. These technologies provide immense future scopes in both private and public sectors of oil and gas industries in Australia and worldwide. There is an opportunity to extend and apply these proposed technologies, as *digital oil field solutions* in both operating and services oil and gas companies worldwide. What these industries have in common, is the need to analyse big datasets, produced and collected by a diverse range of methods, to understand particular phenomena and or hypotheses. Different mining rules, classifications, associations, clustering, spatial clustering and decision tree mining are made good use of for data mining of the complex data types, which are intended initially, to be studied for volume of resources and energy industry data and web sources, since these industrial data consist of complex data types. Data that have been mined can be converted into XML data documents. These XML documents can be exchanged when business-to business transactions are performed. There is a scope of further generating interfaces (both back office and front-end) and connecting to the data warehouses for suitably accessing and extracting knowledge from resources data and communicating it through XML documents in different operational business centres. Advanced statistical approaches, neural networks are quite useful technologies in the resources industries, which are already in place in some resources companies, such as Hampson and Russell Software Services providers. Golden software solutions and Matalab features are used where ever necessary. As a part of future scope and vision of current research, several other data sources are examined, as described in the following sections.

Indian Subcontinent: In the context of the Indian sub-continent (Figure 2.1b), there is immense scope of analysing different sedimentary basins and integrating them for risk minimizing the exploration & production tasks in onshore and offshore regions. The proposed methodologies are intended to be analysed for building knowledge and effectively managing petroleum systems with wide technical and economic objectives in the Indian sub-continent.

Data Sources from Healthcare Domain

The author uses the conceptual and logical models used in the current research domain successfully in the healthcare domain, especially the diabetic-food domains. Based on this research, author published research papers in one of the journals of public healthcare systems (Nimmagadda and Dreher 2009 and Nimmagadda et al. 2011c).

Data Sources from Environment Domain

The author has made an attempt to use the data models and methodologies including the integrated framework in the environment management, such as carbon emissions ecosystems domain (Nimmagadda and Dreher 2009). The earthquake data sources (Milne 1886, Abe and Kanamori 1980 and Trever et al. 2005) are used for analysing the domain knowledge and its implementation in the hazard management systems (Nimmagadda and Dreher 2007).

Other Data Sources on Hazard Management

Worldwide, because of extreme hot, humid, dry and windy conditions, several bushfires occur, causing damage to human and animal life and loss of billions of dollars of worth properties. Besides bushfires, cyclones, earthquakes, landslides, severe weather, tsunami, volcanoes and even nuclear monitoring are other domain hazards, can be managed by current system development methodologies. Geographic and periodic data sources may be found from the sites as described in the following Figures.

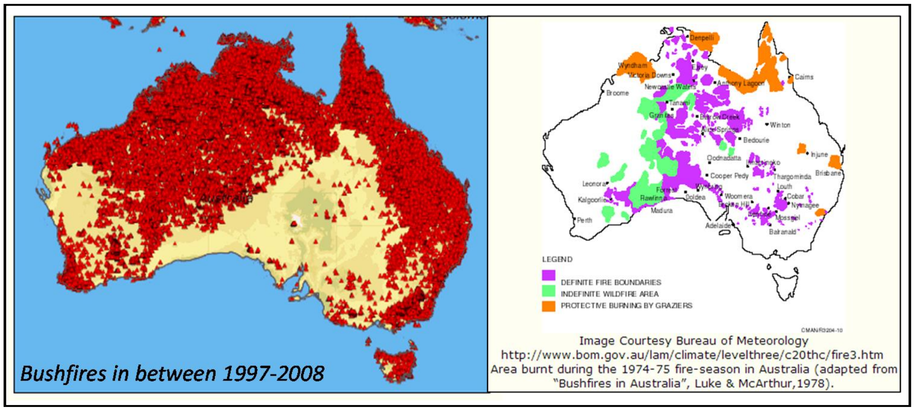
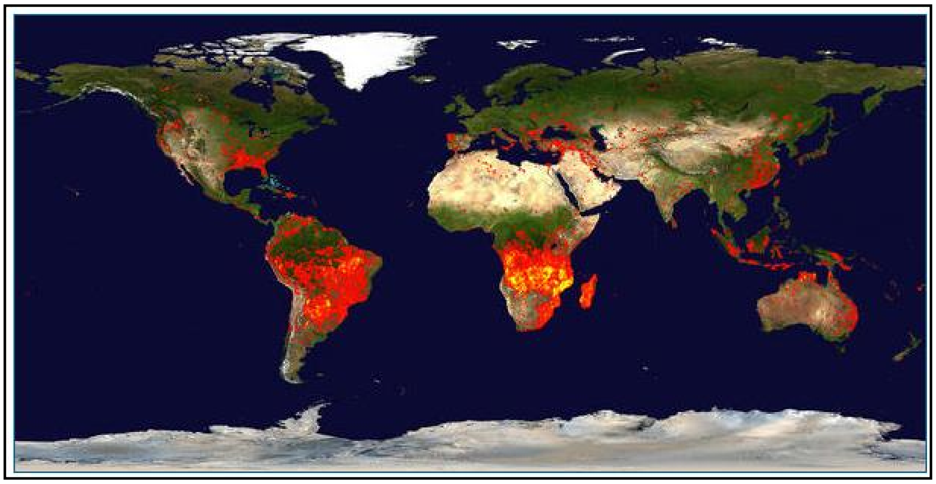


Image Courtesy Bureau of Meteorology
http://www.bom.gov.au/lam/climate/levelthree/c20thc/fire3.htm
Area burnt during the 1974-75 fire-season in Australia (adapted from "Bushfires in Australia", Luke & McArthur,1978).

Australian bushfires and their historical occurrences (Bureau of Meteorology, Australia)

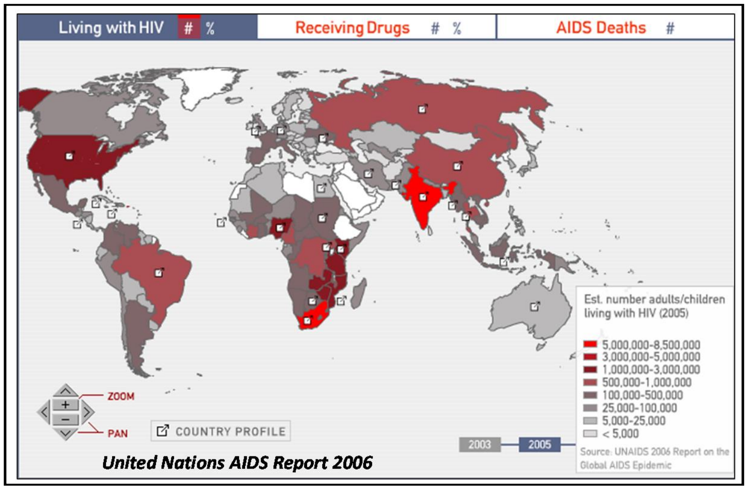
Brushfires' statistical data can be acquired from the Geoscience Australia (GA) sources. Different other domains, such as cyclone, tsunami warnings and earthquakes data sources for many decades are accumulated at these public domain sources. NASA (American National Space Agency) is another useful source of data, where large amount of information is available either for analysing and or documenting for future analysis.



Global fires, map view (http://lance-modis.eosdis.nasa.gov/cgi-bin/imagery/firemaps.cgi)

Data Sources on global pandemic AIDS disease management

In another domain, such as AIDS disease is spread worldwide and the author describes data sources as depicted here. In recent years, though this disease is curtailed, still it is a global pandemic disease.

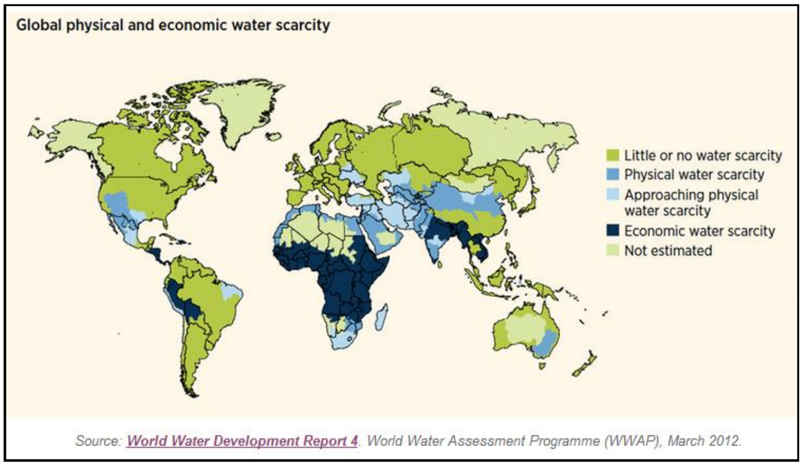


Source: UNAIDS 2006 Report on the Global AIDS Epidemic

World's AIDS disease statistics (Source: UNAIDS Report, 2006)

Data Sources on natural water resources management

Human and animal survival on this planet depend on water resources. Though 70% of our planet consists of water, but it is either unusable or not suitable for human or animal consumption. There are plenty of data sources, as described in the United Nations reports.


Global physical and economic water scarcity
Little or no water scarcity
Physical water scarcity
Approaching physical water scarcity
Economic water scarcity
Not estimated

Source: World Water Development Report 4. World Water Assessment Programme (WWAP), March 2012.

Global view of water scarcity (WWAP, 2012)

The author has documented several data sources for future studies. These are heterogeneous and multidimensional, based on the geographic and periodic dimensions. Documentation is an important asset of any business environment system. In the current research, methodologies are systematically organized and presented.

References

Abe, K. and Kanamori, H. (1980), “Magnitudes of great shallow earthquakes from 1953 to 1977”, *Tectonophysics*, Vol. 62, p. 196.

Abdelali, A., Cowie, J., Farwell, D., Ogden, B. and Helmreich, S. (2003), Cross-Language Information Retrieval using Ontology, Proceedings of TALN Batz-sur-Mer. France.

Aczel, D. A. (1993), “Complete Business Statistics”, 3rd Edition, New York: McGraw-Hill, 1-869p.

Aitken, S., and Reid, S. (2000), Evaluation of an Ontology-Based Information, Retrieval Tool, Proceedings of ECAI’00. Berlin, Germany.

Aldenderfer, M.S. and Bashfield, R.K. (1984), “Cluster Analysis”, Sage University paper series on quantitative applications in the social sciences No.07-44, Beverly Hills, CA: Sage Publications.

Al-Fares, A. A, Bouman, M. and Jeans, P. (1998), “A new look at the middle to lower Cretaceous stratigraphy”, Offshore Kuwait, *GeoArabia*, Vol. 3(4), p. 543-560.

Ambler, S.W. (2001), Mapping objects to relational databases: Agile Alliance, the Objects Primer, 2nd Edition; http://www.AmbySoft.com/mappingObjects.pdf.

Anahory, S. and Murray, D. (1997), “Data warehousing in the real world: a practical guide for building decision support systems”, Pearson Education, Addison-Wesley, UK, pp. 255-360.

Andritsos, P. (2002), Data clustering techniques, a qualifying oral examining paper, Department of Computer Science, University of Toronto, www.cs.toronto.edu/~periklis/pubs/depth.pdf.

Atkinson, M., DeWitt, D. Maier, D., Bancihon, F. Dittrich, K. and Zdonik, S. (2003), The object oriented database system manifesto; http://www.cl.cam.ac.uk/teaching/2003/DBaseThy/oo-manifesto.pdf.

Baker, G., Brass, A., Bechhofer, S., Goble, C., Paton, N. and Stevens, R. (1998), TAMBIS: Transparent Access to Multiple Bioinformatics, Information Sources. In: Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff D., Sensen, S. (eds.): 6th Int. Conf. on Intelligent, Systems for Molecular Biology. AAAI Press, Menlo Park. Montreal, Canada, pp 25–34.

Bartle, R. G. (1976), “The Elements of Real Analysis” (2nd ed.), New York: John Wiley & Sons, ISBN 978-0-471-05464-1, p. 17 *ff*.

Bayle, A. and Ozkarahan, E. (1988), Knowledge based system for relational normalization of GDBMS Conceptual schemas, “Proceedings of the fourth IEEE conference on artificial intelligence applications”, San Diego, CA, USA.

Beaumont, E.A. and Foster, N.H. (1999), Exploring for Oil & Gas Traps, *AAPG Treatise of Petroleum Geology*, Publications of Millennium Edition, Memoir 78, 2nd Edition, UK.

Becher, J. P. Berkhin, and E. Freeman, E. (2000), Automating exploratory data analysis for efficient data mining, Proceedings of the 6th *ACM SIGKDD*, pp. 424-429, New York, USA.

Bench-Capon T.J.M. and Malcolm G. (1999), Formalising Ontologies and Their Relations. Proceedings of DEXA'99, pp. 250–259.

Berenson, M.L. and Levine, D.M. (1992), "Basic Business Statistics, Concepts and Applications", sixth edition, Prentice Hall, New Jersey, USA, 1-953p.

Berkin, P. (2002), Survey of Clustering Data Mining Techniques, Accrue, Software Inc, San Jose, 2002. www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf.

Berson, A. and Smith, J. S. (2004), "Data warehousing, data mining & OLAP", Mc Graw – Hill Education (India) Pty Ltd, pp. 205-219, 221-513.

Bhatt, M. Flahive, A. Wouters, C. Rahayu, J. W., Taniar, D. and Dillon, T.S. (2004), A Distributed Approach to Sub-Ontology Extraction, Proceedings of the 18th International conference on advanced information networking and applications, IEEE computer society press, *AINA* (1): 636-641.

Biantoro, E., Kusuma, M.I., and Rotsinsulu, L. F. (1996), Tarakan Basin growth faults, North-East Kalimantan: Their roles in hydrocarbon entrapment. Indonesian Petroleum Association, published in the proceedings 25th annual convention, Jakarta, 1996, 25(1), 175-189.

Biswas, G. Weinberg, J. and Li, C. (1995), ITERATE: A conceptual clustering method for knowledge discovery in databases, *Artificial Intelligence in the Petroleum Industry*, Paris, France, pp.111-139.

Bouquet, P., van Harmelen, F., Giunchiglia, F., Serafini, L. and Stuckenschmidt H. (2003), C-OWL: Contextualizing ontologies. Proceedings of the second International Semantic Web Conference - ISWC'03, Sanibel Island, Florida. October.

Brocke, J. V. and Buddendick, C. (2006), Reusable conceptual models – requirements based on the design science research paradigm, DESRIST, CGU Publications, CA, USA.

Brown, D. (2013), Looking deeper into fracture impacts, *AAPG Explorer*, March Archive, AAPG Publications, USA.

Bylander, T., and Chandrasekaran, B. (1988), Generic tasks in knowledge based reasoning: The right level of abstraction for knowledge acquisition. In: Gaines B., Boose, J. (eds.): Knowledge Acquisition for Knowledge Based Systems. Vol. 1. Academic Press, London, pp. 65–77.

Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D. and Rosati, R. (1998), Information integration: Conceptual modelling and reasoning support. In Proceedings of the 6th International Conference on Cooperative Information Systems (CoopIS'98), pp. 280-291.

Cardenas, A. F., and McLeod. (1990), "Research Foundations in Object-Oriented and Semantic Database Systems". Prentice Hall, Englewood Cliffs, NJ 07632.

Castañeda Gonzalez, O. J., Nimmagadda, S.L, Cardona Mora, A. P, Lobo, A, and Darke, K. (2012), On Integrated Quantitative Interpretative Workflows for interpreting structural and combinational traps for risk minimizing the exploratory and field development plans, presented and published in the *Bolivarian Geophysical Symposium* proceedings, held in Cartagena, Colombia.

Chaudhri, A.B. (1993) Object Database Management Systems: An Overview in "BCS OOPS Newsletter", No.18 Summer '93, USA.

Chandrasekaran, B., Johnson, R. and Benjamins, R. (1999), Ontologies: what are they? why do we need, them?. IEEE Intelligent Systems and Their Applications, 14(1), Special Issue on Ontologies, pp. 20– 26.

Cheung, D.W. Wang, L. Yiu, S.M. and Zhou, B. (2000), Density based mining of quantitative association rules, *PAKDD, LNAI* 1805, pp. 257-268, Springer-Verlag, Berlin, Heidelberg.

Churchill, G.A. JR. (1979), A paradigm for developing better measures of marketing constructs, *Journal of Marketing Research*, AMA Publications, Vol. XVI, pp. 64-73, USA.

Clancey, W.J. (1992), “Model construction operators”, Artificial Intelligence, 53(1):1-115.

Cleary, L, Freed, B, and Elke, P. (2012), “Big Data Analytics Guide: 2012”, Published by SAP, CA 94607, USA.

Cleveland, W.S. (1994), The elements of graphing data, p. 297, Hobart Press, New Jersey, USA.

Coronel, C., Morris, S., and Rob, P. (2011), Database Systems, Design, Implementation and Management, Course Technology, Cengage Learning, USA.

Courteney, S. Cockcroft, R. Lorentz, R. Miller, R. Ott, H.L. Prijosoesilo, A.R, Suhendan, A.W.R, Wight and Biman, S.K. (1991), Indonesia – Oil & Gas Field Atlas Volume V: Kalimantan, *IPA* 032, p1-21, Jakarta, Indonesia.

Creties D J, De Golyer, Mac Naughton and Boyer (II). (2008), C. M, Coal-bed and shale gas reservoirs, *SPE*, 103514, Distingushed Lecture Series, USA.

Damiani, E. (2008), Key note address on ‘Digital Ecosystems: the next Generation of Service Oriented Internet”, IEEE-DEST, Phitsanulok, Thailand, Feb 2008.

Davidson, S.B, Overton C, Buneman P. (1995), Challenges in integrating biological data sources, *Journal of Computational Biology*, 2(4):557-572, Thompson Reuters, CA, USA.

Debortoli, S. Muller, O. and Brocke, J.V. (2014). Comparing Business Intelligence and Big Data Skills, BISE – RESEARCH PAPER, DOI 10.1007/s12599-014-0344-2, Springer Fachmedien Wiesbaden, 2014

Deitel, H.M. and Deitel, P.J. (2001), “C++ How to Program (Introducing Object-Oriented Design with the UML)”, Upper Saddle River Publishers, 3rd Edition, Prentice Hall Publishers, NJ, USA.

Degen, W., Heller, B., Herre, H. and Smith, B. (2001), GOL: Towards an Axiomatized Upper-Level Ontology. In: Welty, C., Smith B. (eds.): Formal Ontology in Information

Systems, proceedings of the Second International Conference (FOIS 2001), ACM Press, New York: October, pp. 34–46.

Demey, J., Jarrar, M. and Meersman, R. (2002), A Conceptual Markup Language that supports interoperability between Business Rule modeling systems, proceedings of the Tenth International Conference on Cooperative Information Systems (CoopIS 02), Springer Verlag LNCS 2519, pp. 19–35.

Deridder, D., and Wouters, B. (2000), The Use of an Ontology to Support a Coupling between Software Models and Implementation. European Conference on Object-Oriented Programming (ECOOP'00), International Workshop on Model Engineering.

Dhar, V. Jarke, M. Laartz, J. (2014). Big Data, WIRTSCHAFTSINFORMATIK, doi:10.1007/s11576-014-0428-0, Springer Fachmedien Wiesbaden 2014.

Ding, Y., and Fensel, D. (2001), Ontology library systems: the key for successful ontology reuse. Proceedings of the first Semantic Web Working Symposium, Stanford, CA, USA. August.

Dodds, K. and Fletcher, A. (2004) Interval probability process mapping as a tool for drilling decisions analysis – the R&D perspective, *The Leading Edge*, Vol. 23(6) (pp. 558-564).

D'Orazio, R., and Happel, G. (1996), "Practical Data Modelling for Database Design". The Information Technology Series, John Wiley & Sons Australia Ltd, Victoria.

Dunham, H. M. (2003), "Data Mining, Introductory and Advanced Topics", Prentice Hall Publications, 10-200p, USA.

Durham, S.L. (2013), An Unconventional Idea, Open to Interpretation, *AAPG Explorer*, March Series, AAPG Publications, USA.

Elmasri, R., and Navathe, S. (1999), Fundamentals of Database Systems. (3rd Edition). Addison-Wesley Publishing.

Erdmann, M. and Rudi, S. (2001), How to structure and access XML documents with ontology, *Data & Knowledge Engineering*, 36 (3), p. 317-335, Elsevier Science Publishers, B.V, Amsterdam, The Netherlands.

Fayyad, U.M., Shapiro, G.P., Smyth, P. and Urthurusamy, R. (1996), "Advances in Knowledge Discovery and Data Mining", MIT Press, Cambridge, MA, 229-247p.

Flahive, A., Rahayu, J. W, Taniar, D. and Apduhan, B.O. (2004), A Distributed Ontology Framework for the Grid, *PDCAT*: 3320, ISBN: 3-540-24013-6, pp.68-71, Parallel and Distributed Computing: Applications and Technologies, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.

Fraley, C. and Raftery, A. (1998), Which cluster method? Answers via model-based cluster analysis, *The Computer Journal*, 41 (8), pp. 578-588, Oxford Open Access Journal.

Franconi, E. (2002), Tutorial on Description Logics for Conceptual Design, Information Access, and Ontology Integration: Research Trends, Proceedings of the 1st International Semantic Web Conference.

Frank, A. (1997), Spatial Ontology: A Geographical Point of View. In: Stock, O. (eds.): Spatial and Temporal Reasoning, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 135–153.

Fonseca, F., and Egenhofer, M. (1999), Ontology-Driven Geographic Information Systems, in: the 7th ACM Symposium on Advances in Geographic Information Systems, Kansas City, MO: ACM Press, N.Y.

Frietas, A. A. (2002), "Data Mining and Knowledge Discovery with Evolutional Algorithms", 31-43, Springer – Verlag New York, Inc, NJ, USA.

Garcia M. H, Ullman, J. D., Widom. J. (2008), "Database Systems, Comprehensive Overview and Int", *Hardcover*, *1,152 Pages*, Published 2008 by Prentice Hall, USA.

Gacenga, F. N. (2013), A Performance Measurement Framework for IT service, PhD Thesis, School of Information Systems, Faculty of Business and Law, University of Southern Queensland, Australia.

Gangemi, A. (2004), Some design patterns for domain ontology building and analysis, an online presentation at (http://www.loacnr.it/Tutorials/OntologyDesignPatterns.zip April).

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2001a), Understanding top level ontological distinctions. Proceedings of IJCAI-01 Workshop on Ontologies and Information Sharing. AAAI Press. Seattle, USA, pp. 26–33.

Gangemi A., Pisanelli DM. and Steve G. (2001b), A formal Ontology Framework to represent Norm Dynamics. Proceedings of Second International Workshop on Legal Ontologies, Amsterdam, NL.

Gilberg, R. (1985), A Schema methodology for Large Entity-Relationship Diagrams, proceedings of the 4th International Conference on Entity Relationship Approach, Chicago, Illinois, ISBN O-13186-0645-2. October, pp. 320–327.

Gilbert, R. Liu, Y. and Abriel, W. (2004), Reservoir modeling: integrating various data at appropriate scales, *The Leading Edge*, Vol. 23(8) (pp. 784-788), EAGE, The Netherlands.

Gomez-Perez, A. and Benjamins, R. (1999), Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods, proceedings of the IJCAI-99, Workshop on Ontologies and Problem-Solving Methods (KRR5), Morgan Kaufmann.

Gornik, D. (2002), Data modeling for data warehouses, *IBM rational software white paper*, TP 161 05/02, Rational E-development Company, USA.

Graham, P. and Desmond, J.K. (1992), The Use and Misuse of Statistical Methods in Information Systems Research, *Information Systems Research*, No. 3, pp.208-229.

Green, W.R. (1991), "Exploration with a computer: Geoscience Data Analysis Applications", (*Computer Methods in Geosciences*, Vol. 9), 1st Edition, Pergamon Press, UK.

Gregersen, H. and Jensen, C. S. (1998), Conceptual Modelling of Time-Varying Information, Time center technical report TR-35, Springer, http://powerdb.net/database.

Gruber, T.R. (1993), A translation approach to portable ontologies. *Knowledge Acquisition*,
5(2):199-220; http://ksl web.stanford.edu/KSL_Abstracts/KSL-92-71.html

Gruber, T. (1995), Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 43(5/6).

Guan, S. and Zhu, F. (2004), Ontology acquisition and exchange of evolutionary product-brokering agents, *Journal of Research and Practice in Information Technology*, Vol. 36 (1), pp.35-46, IGI, USA.

Guarino, N. (1994), The Ontological Level. In R. Casati, B. Smith and G. White (eds.), Philosophy and the Cognitive Science, Hölder-Pichler Tempsky, Vienna: 443-456.

Guarino, N. and Giaretta, P. (1995), "Ontologies and Knowledge Bases: Towards a Terminological Clarification"` in: Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, N. Mars (ed.), pp 25-32, IOS Press, Amsterdam.

Guarino, N. (1997), Understanding, building, and using ontologies: A commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga." International Journal of Human and Computer Studies No. 46, pp. 293–310.

Guarino, N. (1998), Formal Ontology in Information System, proceedings of FOIS'98, IOS Press, Amsterdam, pp. 3–15.

Guarino, N. and Welty, C. (2000), A Formal Ontology of Properties. Proceedings of the ECAI-00 Workshop on Applications of Ontologies and Problem Solving Method. Berlin, Germany, pp. 12.1–12.8.

Guarino, N. (2002), Ontology-Driven Conceptual Modelling. Tutorial at 21st International Conference on Conceptual Modelling (ER'02), Tampere, Finland.

Guha, S. Rastogi, R and Shim, K. (1998), CURE: An efficient algorithm for clustering large databases, *Proceedings of ACM-SIGMOD International conference on management of data,* ACM, New York, USA.

Guizzardi, G., Herre, H. and Wagner G. (2002), Towards Ontological Foundations for UML Conceptual Models. proceedings of the 1st International Conference on Ontologies, Databases and Application of Semantics (ODBASE'02), Lecture Notes in Computer Science, Vol. 2519, Springer-Verlag, Berlin, pp. 1100–1117.

Gupta, S.P (1990), "Practical Statistics", M/S Chan & Co Publishers, New Delhi, 1-563p, India.

Guoyu, L. (2011), World Atlas of Oil and Gas Basins, Wiley-Blackwell, p. 1-474.

Hadzic, M. and Chang, E. (2005), Ontology-based support for human disease study, published in the *38th Hawaii international conference on system sciences*, Hawaii, USA.

Hair, J.F, Anderson, R.E., Tatham, R.L. (1984), "Multivariate Data Analysis", 2nd Edition, 1-449p, Maxwell Macmillan Publishers, New York, USA.

Halmos, Paul R. (1974), "Naive Set Theory", New York: Springer, ISBN 978-0-387-90092-6, p. 38 *ff*.

Halpin, T. (2001), Information Modelling and Relational Databases. 3rd edn. Morgan-Kaufmann.

Han, J. and Cercone, N. (2000), Aviz: A visualization system for discovering numeric association rules. In: Terano. T, Liu. H, Chen A.L.P. (eds.) *PAKDD* 2000, LNCS (LNAI) vol. 1805, pp. 269-280, Springer, Heidelberg.

Han, J., Kmber, M. and Tung, A.K.H. (2001), Spatial clustering methods in data mining: a survey, in Miller, H. and Han, J. (eds.) *Geographic Data Mining and Knowledge Discovery* (pp.33-50), London and New York: Taylor and Francis.

Heather, A.W. (2004), XML, “A quick reference of more than 300 XML tasks, Terms and Tricks, from A to Z”, Firewall Media, Laxmi Publications Pty Ltd, New Delhi, India.

Hevner, A.R., March, S.T., Park, J. Ram, S. (2004), Design science in information systems research, *MIS Quarterly*, Vol. 28 (1), pp.75-105, Society for Information Management and the Management Information Systems Research Center, MN, USA, ISSN: 0276 - 7783.

Hoffer, J.A, Presscot, M.B and McFadden, F.R. (2005), “Modern Database Management”, Sixth Edition, Prentice Hall, USA.

Hori, M. and Ohashi, M. (2005) Applying XML web services into health care management, *Proceedings of the 38th Hawaii international conference on system sciences*, Hawaii, USA.

Hoffman, D.R. (2003), “Effective Database Design for Geoscience Professionals”, PennWell Publishers, pp. 205-222, OK, USA.

Huang, Z. (1997), A fast clustering algorithm to cluster very large categorical data sets in data mining, *Proceedings of SIGMOD workshop* on research issues on data mining and knowledge discovery, pp. 311-322, Jones and Barlett Learning, LLC, UK and Canada.

Huston, D.C. Hunter, H. and Johnson, E. (2003), Geostatistical integration of velocity cube and log data to constrain 3D gravity modeling, Deepwater Gulf of Mexico, *The Leading Edge*, Vol. 23(4) (pp.842-846).

Huynh, T.N. Mangisengi, O. and Tjoa, A.M. (2000), Metadata for Object-Relational Data Warehouse, *Proceedings of the International Workshop on Design and Management of Data Warehouse,* Stockholm, Sweden, June 5-6.

Indulska, M. and Recker, J. C. (2008), Design Science in IS Research: A Literature Analysis. In Gregor, Shirely and Ho, Susanna, Eds. Proceedings 4th Biennial ANU Workshop on Information Systems Foundations, Canberra, Australia.

Jasper, R. and Uschold, M. (1999), A Framework for Understanding and Classifying Ontology Applications, p. 1-20, *published in the Proceedings of the IJCAI-99* workshop on ontologies and problem-solving method (KRR5), Stockholm, Sweden.

Jain, A.K. and Murthy, N.M. (1999), P.J. Flynn, Data clustering – a review, *ACM Computing Surveys*, Vol. 31(3), pp. 264-323, ACM, New York, USA.

Johnston, D.H. (2004) 4D-gives reservoir surveillance, AAPG Explorer, Vol. 25(12) (pp. 28-30).

Jarrar, M. (2005), Towards Methodological Principles for Ontology Engineering, PhD Dissertation, Vrije Universiteit Brussel, Faculty of science, Belgium.

Jarrar, M., Meersman, R. (2002), Scalability and Knowledge Reusability in Ontology Modelling. Proceedings of the International conference on Infrastructure for e-Business, e-Education, e-Science, and e-Medicine (SSGRR'2002s).

Jukic, N. and Lang, C. (2004), Using offshore resources to develop and support data warehousing applications, *Business Intelligence Journal*, p.6-14, Netmobius, USA.

Kandogan, E. (2001), Visualizing multi-dimensional clusters, trends and outliers using star co-ordinates, Proceedings of the 7th *ACM SIGKDD*, pp.107-116, ACM, NY, USA.

Karp, P. (1995), A strategy for database interoperation*, Journal of Computational Biology*, Thompson Reuters, 2(4): 573-586, NY, USA.

Kaufman, L. and Rousseeuw, P.J. (1990), "Finding Groups in Data: An Introduction to Cluster Analysis". Wiley, New York.

Keogh, E., Chakraborthy, K., Pazzani, M. and Mehrotra, S. (2001), Dimensionality reduction for fast similarity search in large time series databases, *Journal of Knowledge and Information Systems*, 3 (3), Springer, USA.

Keet, M. (2004), Aspects of ontology integration. Technical report. School of Computing, Napier University.

King, E. (2000), "Data Warehousing and Data Mining: Implementing Strategic Knowledge Management", 1st Ed, CTR Corporation, ISBN 1566070782, SC, USA.

Khatri, V. and Ram, S. and Snodgrass, R.T. (2004), Augmenting a conceptual model with geo-spatiotemporal annotations, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16(11), pp. 1324-1338, doi:10.1109/TKDE.2004.66.

Krishnamurthy, B. (1999), "Data visualization techniques", Wiley & Sons, Incorporated, John, USA.

Lancaster, D. E. (1996), Production from fractured shales, SPE Reprint Series, Production from Fractured Shale, No. 45, Published by *SPE* (Society of Petroleum Engineers), TX, USA.

Lee, T.J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D.W.J., Tenenbaum, J.D., and Karp, P.D. (2006), BioWarehouse: a bioinformatics database warehouse toolkit; *BMC Bioinformatics* 7:170, UK, http:///www.biomedcentral.com/1471-2105/7/170.

Liner C.L. (1998), Normal Moveout, an Effect, Process, *AAPG Explorer* 19, no. 4: 26-29, AAPG Publications,TX, USA.

Longley, I.M., Bradshaw, M.T., and Hebberger, J. (2001), *Australian petroleum provinces of the 21st century*. In M.W. Downey , J.C. Threet , and W.A. Morgan (Eds.), Petroleum provinces of the 21st century, AAPG Memoir, 74, 287-317, USA.

Lowrie W. (1997), "Fundamentals of Geophysics", New York, USA: Cambridge University Press.

Magoom B. L and Dow, D.W. (1994), “The Petroleum System – from Source to Trap”, AAPG Memoir 60, AAPG Publications, p. 1-625, USA.

March, S. T., and Smith, G. (1995), “Design and Natural Science Research on Information Technology,” Decision Support Systems 15 (4), December 1995, pp. 251-266, Elsevier Science B.V., USA.

Matsuzawa, H. and Fukuda, T. (2000), Mining structured association patterns from databases, *PAKDD* 2000, LNAI 1805, pp. 233-244, Springer-Verlag, Berlin, Heidelberg.

Marakas, M. G. (2003), “Modern Data Warehousing, Mining, and Visualization Core Concepts”, Prentice Hall Pub.

Maedche A., Motik B., Silva N., Volz R., (2002), MAFRA a Mapping FRAmework for distributed ontologies, Proc. 13th Int. Conf. Knowledge Engineering and Knowledge Management, Siguenza, Spain, pp. 235–250, (http://www.cs.ox.ac.uk/boris.motik/pubs/mmsv02mafra.pdf).

Matsuzawa, H. and Fukuda, T. (2000), Mining structured association patterns from databases, *PAKDD* 2000, LNAI 1805, pp. 233-244, Springer-Verlag Berlin Heidelberg.

Mattison, R. (1996), “Data Warehousing Strategies, Technologies and Techniques”, Mc-Graw Hill Publishers, 100-450p, NY, USA.

Meersman R. (1999), Semantic Ontology Tools in Information System Design. In, Ras, Z. & Zemankova, M., (eds.), Proceedings of the ISMIS 99 Conference, LNCS 1609, Springer Verlag, pp. 30–45.

Meersman R. (2000), Can Ontology Theory Learn from Database Semantics?. Proceedings of the Dagstuhl Seminar 0121 'Semantics on the Web.

Meersman R. (2001), New Frontiers in Modeling Technology: The Promise of Ontologies. Proceedings of the SISO ESM Conference on Simulation.

Meersman, R.A. (2004), Foundations, implementations and applications of web semantics, parts 1, 2, 3, lectures at School of Information Systems, CBS, Curtin University, Australia.

Miller, H.J and Han, J. (2001), Fundamentals of spatial data warehousing for geographic knowledge discovery, *Geographic data mining and knowledge discovery*, pp. 51-72, London: Taylor and Francis.

Miller, G.C, Dolan, C.T, Crowson, N., Stout, S.R. (2002), Data warehousing and information management strategies in the clinical immunology laboratory, Clinical and Applied Immunology Reviews 3, 73–86, Elsevier Science Inc, The Netherlands.

Milne, J. (1886), “Earthquakes and other Earth Movements”, The International Scientific Series, Kegan Paul, Trench and Company, Harvard University, p. 363, MA, USA.

Mintu, A.T., Calantone, R.J and Gassenheimer, J.B. (1994), Towards improving cross-cultural research: extending Churchill’s research paradigm, Journal of International Consumer Marketing, Vol. 7 (2), Haworth Press Inc., USA.

Moody, L. D and Kortink, M.A.R. (2003), From ER Models to Dimensional Models: Bridging the gap between OLTP and OLAP Design, Part1 and Part 2, *Journal of Business Intelligence*, Summer Fall editions, Vol. 8(3), http://www.tdwi.org.

Musen, M. (1998), Domain Ontologies in Software Engineering: Use of Protege with the EON Architecture. Methods of Information in Medicine, No. 37, pp. 540–550.

Nigel E. C, Cunningham, C, Cook, R. J, Taha, A. Esmaie, E and Swidan, N.E. (2009), Three-dimensional seismic geomorphology of deep-water slope-channel system: The Sequoia field, offshore west Nile Delta, Egypt, *AAPG Bulletin*, Vol. 93 (8) p. 1063-1086, USA.

Nimmagadda, S.L. and Rudra, A. (2004a), Applicability of data warehousing and data mining technologies in the Australian resources industry, *published in the proceedings of 7th international conference* on IT, held in Hyderabad, India.

Nimmagadda, S.L. and Rudra, A. (2004b), Data sources and requirement analysis for multidimensional database modeling – an Australian Resources Industry scenario, *published in the proceedings of 7th international conference on IT*, held in Hyderabad, India.

Nimmagadda, S.L. and Dreher, H. (2005a), Ontology of Western Australian petroleum exploration data for effective data warehouse design and data mining, a paper presented and published in the *proceedings of 3rd international IEEE conference on Industrial Informatics*, held in Perth, Australia.

Nimmagadda, S.L. and Dreher, H. (2005b), Data warehouse structuring methodologies for efficient mining of Western Australian petroleum data sources, a paper presented and published in the *proceedings of 3rd international IEEE conference on Industrial Informatics*, held in Perth, Australia.

Nimmagadda, S.L, Dreher, H. and Rudra, A. (2005c), Warehousing of object oriented petroleum data for knowledge mapping, a paper presented and published in the *proceedings of 5th International Conference of IBIMA*, Cairo, Egypt.

Nimmagadda, S.L. and Rudra, A. (2005d), Data Mapping Approaches for Integrating Petroleum Exploration and Production Business Data Entities for Effective Data Mining, a paper presented and published in the *proceedings of 3rd Kuwait International Petroleum Conference and Exhibition (KIPCE2005)*, Kuwait.

Nimmagadda, S.L, and Dreher, H. (2006a), Mapping and modelling of Oil and Gas Relational Data Objects for Warehouse Development and Efficient Data Mining, a paper presented and published in the *proceedings of 4th International Conference of IEEE Industry Informatics*, held in Singapore.

Nimmagadda, S.L, and Dreher, H. (2006b), Mapping of Oil and Gas Business Data Entities for Effective Operational Management, a paper presented and published in the *proceedings of the 4th International Conference of IEEE Industry Informatics*, held in Singapore.

Nimmagadda, S.L. and Dreher, H. (2006c), Ontology-Base Data warehousing and Mining Approaches in Petroleum Industries: in Negro, H.O., Cisaro, S.G., and Xodo, D., (Eds.), Data Mining with Ontologies: Implementation, Findings and Framework, a book chapter published in 2007 by Idea Group Inc. http://www.exa.unicen.edu.au/dmontolo/

Nimmagadda, S.L, Dreher, H. Chang, E. and Rajab, M.R. (2006d), New technologies in mature gulf basins – multidimensional modelling of ontologically derived historical petroleum exploration data properties for effective basin knowledge mapping, a poster paper presented and published in the *AAPG international conference and exhibition*, 5-8 November, Perth, Australia.

Nimmagadda, S.L. Dreher, H. and Rajab, M.R. (2007a), Ontology-based Warehouse Time-Depth Data Modelling Framework for Improved Seismic Interpretation in Onshore Producing Basins, a paper presented and published in the proceedings of an *International Petroleum Technology Conference (IPTC)*, held in Dubai, UAE.

Nimmagadda, S. L. and Dreher, H. (2007b), DESIGN OF PETROLEUM COMPANY'S METADATA AND AN EFFECTIVE KNOWLEDGE MAPPING METHODOLOGY, a paper presented and published in the proceedings of *IASTED* conference, held in Cambridge in USA.

Nimmagadda, S.L., and Dreher, H. (2007c), Ontology based data warehouse modelling and mining of earthquake data: prediction analysis along Eurasian-Australian continental plates, a paper published in the proceedings of an *International Conference of IEEE in Industry Informatics Forum*, Vienna, Austria.

Nimmagadda, S. L., and Dreher. H.V. (2008a), Ontology-based data warehousing and mining approaches in petroleum industries, In *Data warehousing and mining: concepts, methodologies, tools and applications*, ed. John Wang, 1901-1925. Hershey, New York and London, UK: Information Science Reference.

Nimmagadda, S.L and Dreher, H. (2008b), Ontology Based Data Warehouse Modelling – a Methodology for Managing Petroleum Field Ecosystems, a paper presented in the *International conference of IEEE-DEST*, held in Bangkok, Thailand.

Nimmagadda, S.L, Nimmagadda, S. K. and Dreher, H. (2008c), Ontology based data warehouse modeling and managing ecology of human body for disease and drug prescription management, a paper presented and published in the proceedings of an *International conference of IEEE-DEST*, held in Bangkok, Thailand.

Nimmagadda, S. L and Dreher, H. (2008d), Petroleum Ontology: an effective data integration and mining methodology, aiding exploration of commercial petroleum plays, a paper presented and published in the proceedings of an *International Conference of IEEE (INDIN'08)*, held in Daejeon, South Korea.

Nimmagadda, S. L. and Dreher, H. (2009a), Petro-data-clustering – knowledge building analysis of complex petroleum systems, a technical paper presented and published in the proceedings of an *international IEEE-ICIT conference*, held in Melbourne, Australia.

Nimmagadda, S. L. and Dreher, H. (2009b), On designing Multidimensional Oil and Gas Business Data structures for effective data warehousing and mining, a technical paper presented and published in the proceedings of *an international conference of IEEE-DEST*, held in Istanbul, Turkey.

.

Nimmagadda, S. L. and Dreher, H. (2009c), On issues of Data Warehouse Architectures – Managing Australian Resources Data, a technical paper presented and published in the proceedings of an *international conference of IEEE-DEST*, held in Istanbul, Turkey.

Nimmagadda, S.L, and Dreher, H. (2009e), Technologies for adaptability in turbulent resources business environments, a book chapter published under a title: Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains, http://www.igi-global.com/, 2009, USA.

Nimmagadda, S.L, and Dreher, H. (2009f), Ontology based data warehouse modeling for managing carbon emissions in safe and secure geological storages, a paper presented and published in the proceedings of an international *SEGJ symposium* – Imaging and Interpretation, in a forum "science and technology for sustainable development", held in Sapparo, Japan; published in the digital library of Society of Exploration Geophysicists (SEG), USA.

Nimmagadda, S. L. and Dreher, H. (2010a), Modelling Multidimensional Australian Resources Data for an effective Business Knowledge Management, a technical paper presented and published in the *proceedings of 8th International Conference and Exposition on Petroleum Geophysics*, organized by the *Society of Petroleum Geophysicists* (SPG) and sponsored by SEG and EAGE, held in Hyderabad, India.

Nimmagadda, S. L and Nimmagadda, S. K. and Dreher, H (2010b), Multidimensional Ontology modelling of Human Digital Ecosystems affected by Social Behavioral Patterns", presented and published in the proceedings of an *IEEE-DEST-2010 conference*, held in Dubai, UAE.

Nimmagadda, S. L and Dreher, H (2010c), Ontology based warehouse modelling of fractured reservoir data – for an effective borehole and petroleum production management, presented and published in the proceedings of an *IEEE-DEST-2010 conference*, held in Dubai, UAE.

Nimmagadda, S. L and Dreher, H (2010d), On new emerging concepts of modelling petroleum digital ecosystems (PDE) by Multidimensional Data Warehousing and Mining Approaches, presented and published in the proceedings of an International Conference of *IEEE-DEST-2010*, held in Dubai.

Nimmagadda, S. L and Dreher, H. (2010e), On Data Integration Workflows for an effective management of Multidimensional Petroleum Digital Ecosystems, Arabian Gulf Basins" presented and published in the proceedings of an *IEEE-DEST-2010 conference*, held in Dubai, UAE.

Nimmagadda, S. L and Dreher, H. (2011b), Shale-Gas Ontology, a robust data modeling methodology for integrating and connecting fractured reservoir petroleum ecosystems that affect production complexities" presented and published in the proceedings of an *IEEE-INDIN-2011* held in Lisbon, Portugal.

Nimmagadda, S. L, Nimmagadda, S. K and Dreher, H (2011c), Multidimensional Data Warehousing and Mining of Diabetes & Food-domain ontologies for e-Health Management, presented and published in the proceedings of an *IEEE-INDIN-2011* held in Lisbon, Portugal.

Nimmagadda, S. L. and Dreher, H. (2012a), On new emerging concepts of Petroleum Digital Ecosystem, *Journal Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2012, 2 (6): 457–475 doi: 10.1002/widm.1070.

Nimmagadda, S.L, Dreher, H, Noventianto. A, Mustofa. A and Fiume. G. (2012b), Enhancing the process of knowledge discovery from integrated geophysical databases using geo-ontologies, a paper presented and published *in the proceedings of Indonesian Petroleum Association (IPA)* conference, held in Jakarta, Indonesia.

Nimmagadda, S.L, Dreher, H, Noventianto. A, Mustofa. A and Fiume. G. (2012c), On new emerging concepts of Tarakan Sedimentary Basin – a Petroleum Digital Ecosystem (PDE), a paper published in the proceedings of an *International Geological Congress* (*IGC*) held in Brisbane, Australia.

Neuman, W.L. (2000), “Social research methods, qualitative and quantitative approaches”, 4th Edition, Allyn and Bacon Publishers, USA.

Ng, R.T and Han, J. (1994), Efficient and effective clustering methods for spatial data mining, *Proceedings of International conference on very large databases* (VLDB’94), Santiago, Chile, p. 144-155.

Noon, S. Harrington, J. and Darman, H. (2003), The Tarakan Basin, Eastern Kalimantan: Proven Neogene Fluvio-Deltaic, Prospective deep water and Paleogene Plays in a Regional Stratigraphic Context, *IPA Publication*, IPA03-G-136, Jakarta, Indonesia.

Noy, N. F and McGuinness, D.L. (2000), “Ontology Development 101: A Guide to Creating Your First Ontology”, Knowledge Systems Laboratory, Stanford University, USA.

O’Brien, J. A., and Marakas, G. M. (2009), “Management information systems”, (9th ed.), Boston, MA: McGraw-Hill/Irwin, USA.

O’Leary, D.E. (2000), Different firms, different ontologies, and no one best ontology, *IEEE Intelligent Systems*, EXPERT, Vol. 15 (5), pp. 72-78, USA.

Opdahl, A.L., Henderson-Sellers, B., and Barbier, F. (2001), Ontological analysis of whole-part relationships in OO-models, *Information and Software Technology*, 43 (6), pp. 387-399, Elsevier Science B.V, USA.

Ott, T. and Swiaczny, F. (2001), Time-integrative Geographic Information Systems; Management and analysis of spatio-temporal data, 1st Edition, pp. 1-234, Springer, Heidelberg, Berlin.

Ozkarahan, E. (1990), “Database Management, Concepts, Design and Practice”, Prentice Hall Publications, USA.

Parasnis, D.S. (1997), “Principles of Applied Geophysics”, 5th Edition, Chapman & Hall, UK.

Plastria, F. Bruyne, S. D. and Carrizosa, E. (2008), Dimensionality reduction for classification: Comparison of techniques and dimension choices, published *in the 4th International Conference, ADMA 2008,* Chengdu, China.

Ploesser, K. (2012), A design theory for context – aware information systems, PhD Thesis, Information Systems School, Science and Engineering Faculty Queensland University of Technology, Queensland, Australia.

Pei, J. Han, J. Mortazavi-asl, B. Zhu, H. (2000), Mining access patterns efficiently from web logs, In: Terano, T., Chen, A.L.P (eds), *PAKDD, LNCS*, 1805, pp. 396-407, Springer, Heidelberg, Berlin.

Post, H.F, Gregory, M. N. and Bonneau, G. (2002), Data Visualization: The State of the Art, *Research paper TU delft*, The Netherlands.

Pratt, J.P and Adamski, J.J. (2000), “Concepts of database management”, 3rd Edition, Excellence in Information Systems, p.253-275, Cambridge, Mass Course Technology.

Pujari, A.K. (2002), “Data mining techniques”, University Press (India) Pty Limited, Hyderabad, India.

Rainer, K. R. and Turban, E. (2009), “Introduction to Information Systems”, 2nd Edition, John Wiley & Sons, Inc, USA.

Ramkumar, G.D and Swami, A. (1998), Clustering data without distance functions, *Bulletin of IEEE Computer Society Technical Committee on Data Engineering,* Vol 21 No.1, Electronic edition, USA.

Reinecke, K., and Bernstein, A. (2013), What s user likes: A design science approach to interfaces that automatically adapt to culture, MIS Quarterly, Vol. 37, No 2, pp-427-453.

Richards. D. (2000), "The Reuse of Knowledge: A User-Centered Approach", International Journal of Human Computer Studies.

Rob, P. and Coronel, C. (2004), “Database Systems, Design, Implementation and Management”, 6th Edition, Thomson Course Technology, Boston, USA.

Roberto, C., Smith, B. and Varzi A. (1998), Ontological tools for geographic representation, in: N. Guarino (eds.): Formal Ontology in Information Systems, Proceedings of the First International Conference (FOIS’98). Amsterdam IOS Press. Trento, Italy, June, pp. 77–85.

Rocha, C., Schwabe, D. and de Aragao, M.P. (2004), A hybrid approach for searching in the semantic web, *WWW*, May 17-22, New York, USA, pp. 374-383.

Rolf K. E. (2005), Explosion Hazards in the Process Industry, *American Petroleum Institute Publication*, p. 441, USA.

Rudra, A. and Nimmagadda, S.L. (2005), Roles of multidimensionality and granularity in data mining of warehoused Australian resources data, *Proceedings of the 38th Hawaii International Conference on Information System Sciences*, Hawaii, USA.

Sasson, A. and Blomgren, A. (2011), Knowledge based oil and gas industry, a research report 3/2011, BI Norwegian Business School Department of Strategy and Logistics, BI Norwegian Business School N-0442 Oslo, www.bi.no/en/research-publications.

Schermann, M. Hemsen, H. Buchmüller, C. Bitter, T. Krcmar, H. Markl, V. and Hoeren, T. (2014). Big Data, An Interdisciplinary Opportunity for Information Systems Research, DOI 10.1007/s12599-014-0345-1, Springer Fachmedien Wiesbaden.

Shanks, G. Tansley, E. Weber, R. (2003), Using Ontology to validate conceptual models”, *Communications of the ACM*, 46(10), pp. 85-89, ACM, NY, USA.

Shanks, G., Tansley, E., and Weber, R. (2004), Representing composites in conceptual modeling, *Communications of the ACM*, Vol. 47(7), pp.77-80, ACM, NY, USA.

Shastri L Nimmagadda and Dreher, H. (2011a), Data warehousing and mining technologies for adaptability in turbulent resources business environments, *Int. J.*

Business Intelligence and Data Mining, Vol. 6, No. 2, 2011, p 113-153, Inderscience Publishers, Geneva, Switzerland.

Shastri L Nimmagadda, Dreher, H., Kabanda, F., Ochan, A., Obita, P., Kiconco, L. and Nabbanja, P*.* (2011b), On Data Integration Workflows for an effective Management of Multidimensional Petroleum Digital Ecosystem in the Albertine Graben*, a technical paper presented and published in the proceedings of East African Petroleum Conference, Kampala,* Uganda.

Sheriff, R.E, (2002), "Encyclopedic Dictionary of Applied Geophysics", 4th Edition, Reference Series 13, SEG Publishers, p. 1-429, USA.

Sheriff, R.E. and Oliveira, N. E. D. (2002), "Dictionary of Applied Geophysics", 4th Edition, SEG Publishers, p. 1-429.

Shoval, P. (1985), Essential information structure diagrams and database schema design. Information Systems, 10(4), pp. 417-423.

Sidhu, A.S., Dhillon, T.S. and Chang E. (2009), Data Integration through Protein Ontology, a book chapter published under a title: Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains, http://www.igi-global.com/ ,USA.

Sidhu, A.S. Dillon, T.S. and Chang, E. (2005), An ontology for protein data models, presented at the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC 2005), Shanghai, China.

Siirtola, (1996), Managing Large Entity-Relationship Diagrams. In: Thalheim, B., Yigitbasi, S., (eds.): Proceeding of the Workshop ER CASE Tools. Cottbus, Germany. October, pp. 29–42.

Simon, H. A. (1992), Scientific discovery as problem solving, *International Studies in the Philosophy of Science*, 6, 3-14.

Simon, H A. (1996), *The Sciences of the Artificial*, 3rd ed. MIT Press, USA.

Smith B. (2005), Ontology and information systems. Stanford Encyclopaedia of Philosophy (2002) http://ontology.buffalo.edu/ontology (PIC). Sure, Y.: Methodology, Tools and Case Studies for Ontology based Knowledge Management. PhD Thesis, University of Karlsruhe, Department of Economics and Business Engineering, 2003.

Spaccapietra, S. and Parent, C. (1994), View Integration: A Step Forward in Solving Structural Conflicts. IEEE Transactions on Data and Knowledge Engineering 6(2).

Steels, L. (1993), "The componential framework and its role in reusability", in: Second Generation Expert Systems, J.-M. David, J.-P. Krivine & R.Simmons, (eds.), pp. 273-298. Berlin: Springer-Verlag.

Suárez-Figueroa, M., García-Castro, R., Gómez-Pérez, A., Palma R., Nixon, L., Paslaru, L., Hartmann J. and Jarrar, J. (2005), Identification of standards on metadata for ontologies. Deliverable D1.3.2. EU-IST Network of Excellence (NoE) IST-2004-507482 (KWEB), Luxemburg.

Sycara, K. Chang, E. Damiani, E. Jarrar, M. and Dillon, T. (eds) (2006), Proceedings of the 2nd IFIP WG 2.12 and WG 12.4 International Workshop on Web Semantics

(SWWS'06). In OTM Workshops (2). Volume 4278 of LNCS, page (1723), Springer Berlin. ISBN: 9783540482734. Montpellier, France. November.

Taniar, D., Rahayu, J.W., Lee, V. and Daly, O. (2008), Exception rules in association rule mining, *Applied Mathematics and Computation*, Vol. 205, No. 2, Elsevier Publishers, pp.735–750, NY, USA.

Telford, W.M, Geldart, L.P. and Sheriff, R.E. (1990), “Applied Geophysics”, Cambridge University Press, 2nd Edition, p. 790, USA.

Thomas, J.L, Yannick, P. Valerie, W., Gupta, P. Stringer-Calvert, D.WJ, Tenenbaum, J.D and Karp, P.D Biowarehouse, P.D. (2006), a bioinformatics database warehouse toolkit; *BMC Bioinformatics*, 7:170, p.1-14, UK; http:///www.biomedcentral.com/1471-2105/7/170.

Tjioe, H. C and Taniar, D. (2004), A Framework for Mining Association Rules in Data Warehouses, IDEAL, LNCS 3177, pp. 159–165, Springer – Verlag Berlin, Heidelberg.

Trevor, J. Middelmann, M. and Corby, N. (2005), Cities Project Perth Report, in Sinadinovski et al. *Chapter – 5, Earthquakes Risk, GA Publications – Earthquakes Risk Assessment and Management in Australia.*

Uschold, M.E. (1998), Knowledge level modelling: concepts and terminology, *Knowledge Engineering Review*, 13(1), Cambridge University Press, USA.

Uschold, M. and Gruninger, M. (1996), Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, 11(2). Also available as AIAI-TR-191 from AIAI, The University of Edinburgh, UK.

Vaishnavi, V. and Kuechler, W. (2004), “Design Research in Information Systems” July 27. URL: http://www.isworld.org/Researchdesign/drisISworld.htm; vvaishna@gsu.edu and kuechler@unr.edu.

Vaishnavi, V. and Kuechler, W. Jr. (2007) Design Science Research Methods and Patterns: Innovating Information and Communication Technology, NY: Auerbach Publications, Boca Raton, FL, Taylor & Francis Group.

Vadarparty, K. (1996) Developing an ODBMS Application: Basic Steps in "Journal of Object Oriented Programming" January '96 pp 19-21, USA.

Van Heijst, G., Schreiber, A. and Wielinga, B. (1997), Using Explicit Ontologies in KBS Development. International Journal of Human Computer Studies, 46, pp. 183–292.

Vermeir D. (1983), Semantic Hierarchies and Abstraction in Conceptual Schemata, Journal of Information Systems. Vol. 8, No. 2, pp.117–124.

Venable, J., Pries-Heje, J., and Baskerville, R. (2014), FEDS: A framework for evaluation in design science research, European Journal of Information Systems, 1-13, doi:10.1057/ejis.2014.36.

Ventrone, V. and Heiler, S. (1991), Semantic Heterogeneity as a Result of Domain Evolution. SIGMOD Record 20(4), pp. 16–20.

Wand, Y. (2000), An ontological analysis of the relationship construct in conceptual modelling, ACM Transactions on Database Systems, Vol. 24 (4), pp. 494-528.

Wand, Y., Storey, V.C., and Weber, R. (1999), An ontological analysis of the relationship construct in conceptual modelling, *ACM Trans. On Database Systems* 24 (4), pp. 494-528, ACM, NY, USA.

Welty, C., Ferrucci, D. (1999), A Formal Ontology for Re-Use of Software Architecture Documents, proceedings of the 1999 International Conference on Automated Software Engineering, IEEE Computer Society Press, October, pp. 259–262.

Welty, C., Jessica, J. (1999), An Ontology for Subject, J. Data and Knowledge Engineering. 31(2). Elsevier, pp. 155–181.

Welty, C. (2002), Ontology-Driven Conceptual Modelling, invited talk at the Fourteenth International Conference on Advanced Information Systems Engineering (CAiSE), Toronto, Canada.

Weinstein, P.C. (1998), Ontology-Based Metadata: Transforming the MARC Legacy, ACM Digital Libraries, Pittsburgh, USA.

Weber, S. (2010), Design science research: paradigm or approach?, *AMCIS 2010 proceedings*, paper 214, Lima, Peru, http://aisel.aisnet.org/amcis2010/214.

Weimer, P. and Davis, T.L (1995), Applications of 3D-seismic data to exploration and production, *AAPG studies in geology, No.42, and SEG Geophysical Developments series, No.5, AAPG Publications, USA*.

West, M. (2006),Ontolog, http://ontolog.cim3.net/cgi-bin/wiki.pl?ConferenceCall_2006_02_23

Wiederhold, G. (1995), Value-added Mediation in Large-Scale Information Systems, DS-6, pp. 34–56.

Winter, R. and Strauch, B. (2003), A method for demand-driven information requirements analysis in data warehousing, *in the proceedings of the36th Hawaii International Conference on System Sciences, HICSS-03,* Hawaii, USA.

Wight, A.W.R, Hare, L.H, and Reynolds, J.R. (1992), A Sedimentary Basin, NE Kalimantan, Indonesia: a century of exploration and future potential, Geological Society of Malaysia, Circum – Pacific Council for Energy and Mineral Resources, Jakarta, Indonesia.

Yao, Y.Y and Zhong, N. (2000), On association, similarity and dependency attributes, *PAKDD, LNAI,* 1805, pp. 138-141, Springer-Verlag, Berlin, Heidelberg.

Yun, C.H and Chen, M.S. (2000), Mining web transaction patterns in an electronic commerce environment, *PAKDD, LNAI*, 1805, pp. 216-219, Springer-Verlag, Berlin, Heidelberg.

Zaima, A. and Kashner, J. (2003), A data mining premier for data warehouse professional, *Business Intelligence Journal,* Spring edition, 8 (2), pp. 44-54, USA.

Zhang, T. (2000), Association rules, *PAKDD* 2000, LNAI 1805, pp. 245-256, Springer-Verlag, Berlin, Heidelberg.

Zhong, T. Raghu, R. and Livny, M. (1996), An efficient data clustering method for very large databases, *Proceedings of ACM SIGMOD* International conference on management of data, ACM, NY, USA.

Zhou, J., Bruns, M.A., and Tiedje, J.M. (1996), DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol*. 62 (2): 316-322, ASM, USA.

Appendix –1: *Glossary of Key Terms*

***Actual Exploration Expenditure*:** refers to the actual expenditure incurred for carrying out mineral or petroleum exploration.
Audiomagnetotellurics: AMT is a higher-frequency *magneto telluric* technique for shallower investigations
Base metals are made up of the copper, silver, lead-zinc, nickel and cobalt.
Bouguer gravity: a gravity anomaly to measure gravitational acceleration
Deltaic: A river delta is a landform that is formed at the mouth of a river, where the river flows into an ocean, sea, estuary, lake, or reservoir. Deltaic environments are gradational to fluvial and coastal environments.
Development: This phase usually follows exploration phase where a prospective discovery (e.g., proven oil & gas field or concentrate of ore) has been made and brought into production. This includes the extension of the life of a current mine or well. Activities may include preparing the ground by the removal of overburden, constructing shafts, drives and winzes or by drilling and completing well. All activities are for the purposes of commencing extraction/mining or extending production.
Drilling expenditure: includes wages and salaries paid to employees; purchase, rental, hiring as well as operation and maintenance of drilling equipment together with activities associated with accessing the area where drilling is to occur (road creation, vessel/transport hiring, site preparation and restoration). Also includes expenditure on drilling done by contractors.
Drilling: where drilling expenditure includes the cost of access (roads, vessel hire etc.) to the drilling site and the site preparation, etc., and other expenditure includes costs of surveys, report writing, map preparation and all other activities attributable to exploration.
Exploration: this activity involves searching for concentration of naturally occurring solid, liquid or gaseous materials and includes new wildcat and stratigraphical and extension/appraisal wells and mineral appraisals intended to delineate or greatly extends the limits of known deposits by geological, geophysical, geochemical, drilling or other methods. This includes construction of shafts and adits primarily for exploration purposes but excludes mine development activities. Exploration for water is excluded. It is also an initial phase in petroleum operations that includes generation of a prospect or play or both, and drilling of an exploration well.
Exploration expenditure: covers all expenditure (capitalized and non-capitalized) during the exploration or evaluation stages in Australia, Australian waters. Costs include cost of exploration, determination of recoverable reserves, engineering and economic feasibility studies, procurement of finances, gaining access to reserves, construction of pilot plants and all technical and administrative overheads directly associated with these functions. Examples are costs of satellite imageries, airborne and seismic surveys, use of geophysical and other instruments, geochemical surveys and map preparation, license fees, land access and legal costs; geologic inspections, chemical analysis and payments to employees and contractors. Cash bids for offshore petroleum exploration permits are also included.
Exploration license/permit: is designed to cover the exploration phase of a project and confers exclusive rights to the exploration for and recovery of samples from the area designated. Relevant Commonwealth, State or Territory Governments grants these rights.
Expected Exploration Expenditure: refers to expected expenditure on mineral, Oil shale and petroleum exploration as reported by private and government organizations. Subsequent events (e.g., discoveries of new ore bodies, unexpected weather conditions, government policy changes, and unforeseen changes in the economic

conditions) may cause actual expenditure to differ from those previously expected. Consequently statistics of expected expenditure should be treated with caution.

Exploration on Production Leases: relates to exploration carried out on the production lease by miners currently producing or development for production of minerals. **Production lease:** is an area on which development to extract coal, minerals, liquids or gaseous materials is underway or where extraction/mining of these substances is already occurring. **Other leases** are those areas outside the production lease. These include areas under exploration license/permit or retention license, as well as non-licensed areas being assessed for exploration (e.g., through airborne surveys).

Facts: Facts are factual data instances. These are for example, *reservoir parameters*, such as *porosity* and *permeability* in their respective units and measurements, *production* data instances that may represent either periodical or historical data. In Australia, *oil and gas* deposits are termed as *resources*. The terms *oil and gas*, *petroleum* and *hydrocarbons* are synonymous each other, but used different contexts.

Geographic extent: The area over which the petroleum occurs.

Geomorphology: It is the scientific study of landforms and processes that shape the Earth

Induced Polarization method (IP): IP anomalies to measure electrical measurements due to base metals of sub-surface of earth

Karst: Karst is a landscape formed from the dissolution of soluble rocks including limestone, dolomite and gypsum. It is characterized by sinkholes, caves, and underground drainage systems

Kerogen: is a mixture of organic chemical compounds that make up a portion of the organic matter in sedimentary rocks.

Magnetics and micro-magnetics: magnetic anomaly, susceptibility measured by magnetic survey

Minerals: in the broad sense comprise metallic minerals, coal, construction materials, gemstones, other non-metallic minerals, oil shale and petroleum (oil or gas).

Migration: Migration is process, whereby hydrocarbons move from source rocks to traps.

Other expenditure: includes all other exploration costs, other than those associated with drilling expenditure.

Offshore/Onshore: where offshore includes all operations in a marine area under the Petroleum Act 1967 or any other acts administered by state governments. **Offshore** commences from the low water mark to three nautical miles out (referred to as coastal waters) under state ot Northern Territory legislation and extends to those areas beyond coastal waters governed by the Commonwealth under the Petroleum Act. **Onshore** includes all Australian territorial lands to the low water mark.

Oilplay: Reservoir, structure (geological), seal, migration, source and timing factors, playing a role for building a constructive (productive) petroleum system.

OLAP: Online analytical processing, is an approach to answering multi-dimensional analytical queries swiftly

Petroleum: is a naturally occurring hydrocarbon or mixture of hydrocarbons. As oil or gas in solution (LPG), associates with sedimentary rocks.

Periods: considered in the present study are historical periods covered are from 1953 to 2002. At places they are in the form of yearly, half-yearly, quarterly and monthly.

Production Lease/Other: is where a production leases an area on which production or development is actually taking place.

Seal: A shale or impervious rock that acts as a barrier to the passage of oil and gas, also called cap rock.

Sediments: Solid fragmented material, such as silt, sand, gravel, chemical precipitates, and fossil fragments, that are transported and deposited by water, ice, or wind or that accumulates through chemical precipitation or secretion by organisms, and that forms layers on the Earth's surface.

***Sedimentary Basin**:* A depression in the crust of the Earth, caused by plate tectonic activity and subsidence, in which sediments accumulate. Sedimentary basins vary from bowl-shaped to elongated troughs. *Oil and gas* are generated, migrated and trapped within the whole basin concept.

Shallow-marine: It is type of depositional environment under which conditions mayny organism deposited. It refers to the area in between the shore and the beginning of reef wall.

Sinkhole: is a depression or hole in the ground caused by some form of collapse of the surface layer

Source: It is rock unit containing sufficient organic matter (kerogen) of suitable chemical composition to biologically or thermally generate and expel oil and gas.

Stack velocity – velocity information derived from the seismic gathers data, for generating seismic stacks for interpretation.

Structure (geological): it is a trap, usually a result of powerful tectonic force in the form of fold or fault.

Surveys: investigation methods, either geophysical, geological or geochemical means; in geophysical surveys, seismic, gravity and magnetic methods are used for oil and gas exploration

System: A regularly interacting or interdependent group of items forming a unified whole whose organization forms a network for distributing something, such as oil and gas.

Tectonic setting: The tectonic setting is the alignment of the weak points on the earth's crust where regional blocks move horizontally, vertically, forward and backward.

Tuples: The rows of a relation, other than the header row containing the attribute names, are called *tuples*. A tuple has one component for each attribute of the relation. The row of each of the table of each dimension described for each domain is a tuple. For instance, the first of three tuples in Table 1 has four components for attributes.

***Permits**:* Oil and gas licensing areas.

Petroleum System: It is a natural system that encompasses a pod of active source rock and all related oil and gas geological elements and processes that are essential, if hydrocarbon accumulation is to exist (Magoon and Dow, 1994). **Elements** are: structure (trap), reservoir, source, seal and **processes** are: generation, migration, timing of formation/migration and accumulations.

Prospectivity: related to existence of number of oil or gas plays in a basin, significant potentiality of hydrocarbon occurrence in a basin, the success rate of oil and gas discovery in a productive basin, implying the ratio of number of producing wells to the number of wells drilled in that basin.

Radiometric or gamma radiation surveys: radiometric anomalies to measuring radiometric minerals

Reservoir: A subsurface volume of rock that has sufficient porosity and permeability to permit the migration and accumulation of oil and gas under adequate trap conditions

Resources: oil and gas and mineral deposits

Vertical seismic profiling, Check shot data – data that connect the surface domain seismic information with sub-surface drilled well information.

Wells: Wells drilled for exploration and production, which can be onshore or offshore

Appendix – 2: Data Sources Considered in the Present Research

1. Australian Bureau of Statistics, Australia (mineral & oil and gas exploration data).
2. Centre for Resource Studies, Queen's University, Canada (mineral exploration data).
3. Western Australian Department of Resource and Industry (oil and gas exploration and production data).
4. South Australian Department of Primary Industry and Resource (oil and gas exploration data).
5. Woodside Energy Company (useful discussions on oil and gas exploration data)
6. Published reports from American Association of Oil and gas Geologists (AAPG), Society of Exploration Geophysicists (SEG), Australian Society of Exploration Geophysicists (ASEG), Society of Oil and gas Engineers (SPE) and International Electronics and Electrical Engineering (IEEE).

Other Data Sources:

1. Eia: http://www.eia.gov/; http://www.statistics2013.org/?src=home-f3; http://www.eia.gov/countries/#rtabs
2. Petroleum Exploration and Production Department (PEPD, Brochure), Ministry of Energy & Mineral Development, Uganda, 2013. Ministry published annual reports.
3. Indonesian Petroleum Association (IPA, Brochure) and Google Indonesian Basin Maps. IPA published reports.
4. Australian Sedimentary basins, Department of Mines and Petroleum, WA, Australia
5. Geoscience Australia (GA); Australian Geological Survey Organization (AGSO): Earthquakes and their statistics
6. AAPG/SEG/SPE published reports.
7. World Health Organization (WHO) reports on health care systems.
8. World seismological map: http://geology.about.com/od/seishazardmaps/ss/World-Seismic-Hazard-Maps.htm
9. http://earthquake.usgs.gov/earthquakes/eqarchives/year/eqstats.php
10. Statistics of earthquakes: David R Brillinger,
11. http://www.stat.berkeley.edu/~brill/Papers/quakestat.pdf
12. http://databib.org/index.php: Databib
13. Topological modelling of Human Anatomy using Medical Data: MIRALab Copyright © Information 1998; US National Library of Medicine, National Institutes of Health: http://www.nlm.nih.gov/research/visible/
14. Australian national disasters: bushfires,
15. http://en.wikipedia.org/wiki/Bushfires_in_Australia

16. Australian bureau of statistics – bushfires, national disasters, natural calamities.
17. World Atlas: http://au.wiley.com/WileyCDA/WileyTitle/productCd-0470656611.html