Reprinted from:

2005 3rd IEEE International Conference on Industrial Informatics (INDIN) Perth, Australia 10-12 August 2005

IEEE Catalog Number ISBN 05EX1057
ISBN 0-7803-9094-6

Data Warehouse Structuring Methodologies for Efficient Mining of Western Australian Petroleum Data Sources

Shastri L Nimmagadda, Heinz Dreher and Amit Rudra
School of Information Systems, Curtin Business School
Curtin University of Technology, Perth, WA, Australia

Abstract – **Representing the knowledge domain of a petroleum system is a complex problem. In the present study, logical modelling of shared attributes of resources industry entities (dimensions or objects) has been used for construction of a dynamic and time-variant metadata model. This work demonstrates effectiveness of multidimensional data modelling for petroleum industry, which will be further investigated for fine-grain data presentation and interpretation for quality knowledge discovery.**

Key words: **Multidimensional data, schemas, petroleum data, data mining**

I. INTRODUCTION

More often in large oil and gas companies, there are several operational databases catering to the data management needs of different departments. Top managers of these companies at corporate level may extract enterprise wide summarized information to support their decision-making activities. Operational data such as day-to-day survey and drilling data are not summarized. However, informational transactions such as yearly petroleum production data are often represented in aggregated or summarized form for future decision support and auditing purposes. Surveys, wells and permits petroleum exploration and production data are typical resources data illustrated in the present paper. Star, snowflake and fact constellation schemas are used for designing warehouse data structures. Resources metadata can be created integrating these schemas logically, but are not necessarily restricted to individual schemas, and may be combination of them, because of the very nature of complex oil and gas business data. In other words, the complexity of the resources metadata structuring can be minimized by judicious adoption of these schemas and their combination, and normalizing/denormalizing attribute relationships. This process reduces the complexity of the warehoused metadata and also facilitates faster accessing of data during data mining stages.

A. Definition of multidimensional modeling in the petroleum industry scenario

The notion of dimension provides a lot of semantic information in a multidimensaional modelling domain, especially about the hierarchical relationship among its elements. It is important to note that dimension modeling is a special technique for structuring the petroleum exploration and production data around business concepts. Unlike ER modeling, which describes entities and relationships; dimensional modeling structures the numeric measures and hundreds of petroleum data dimensions. The dimension schema can represent the details of dimensional modeling and factual data along with their units.

Complex business data identified as entities are converted into dimensions (and or objects, if dimensions interpreted as objects) and organized them logically into a metadata structure model. Data integration and sharing of data are key objectives. The purpose of mapping and modelling petroleum business data is to simplify the very nature of complex representation and description into multidimensional data cubes. There are many basins all over in Australia, each atleast 5000 sq. Km in size with several oil and gas fields. For each of these basins, data is best gathered and stored as dimensions and facts forming a huge data warehouse. As all these basins share certain common characteristics, the authors believe that all the models built based on certain logic in different (petroleum system) knowledge-domains, can be translated into physical or implementation models as is done in warehouse modelling. Description of petroleum business logical data models that facilitate integrating into a warehouse environment is the subject of present study.

II. LITERATURE REVIEW

Issues of applicability of data warehousing and data mining in the resources industries have been discussed in detail in [7], [8], [9], [10] and [12]. They demonstrate various warehouse schemas deployed in the oil and gas industries. Different data sources, types of data and data requirement analysis for petroleum industries have been described in [8], [9], and [16]. They also describe ontology and conceptual models of petroleum exploration and production data. Petroleum exploration and production concepts and description of various operational entities involved in exploring economic deposits hidden underneath great depths of the earth's crust have been given in [1] and [14].

III. PROBLEM DEFINITION

Data warehousing is essentially a process that facilitates the data-driven approach. It is a new technology that provides tools to store summarized information from multiple, heterogeneous databases in a single repository. It is a process of integrating enterprise-wide corporate data into a single repository, from which end-users can run queries, make reports and perform multidimensional data analysis. There is huge demand and requirement [10], [16] and quality information in a distributed environments [3]. In our case, operational data, such as from exploration, drilling, production and marketing departments possess hundreds of dimensions with similar number of fact tables.

3rd International Conference on Industrial Informatics(INDIN 2005)

Petroleum bearing sedimentary basins of Western Australia are highly prospective compared to other Australian provinces. Historical exploration and production data in Canning, Carnarvon, Bonaparte, Browse and Perth basins (Fig. 1) 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 analyse them for knowledge building purposes through data-mining.

As stated earlier, resources industries handle complex and large volumes of data with numerous dimensions, (see Fig. 1) and attributes with multiple associations. It is intended to simplify 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 operational data of oil and gas industry. Various dimensions and attributes for resources industries have to be identified and analysed for data modelling purposes.

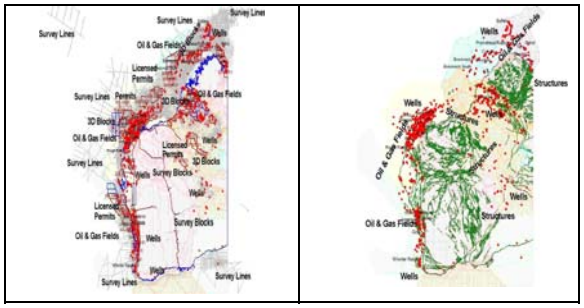

Fig. 1: Maps of Western Australia showing areas of surveys, oil & gas producing fields, wells and permits (modified version, originally reproduced from: www.doir.wa.gov.au)

Super-type entities such as exploration, drilling and production have been used for multidimensional data modelling. Various other associated sub-type entities, attributes and their key indices have been identified. Logical ER (entity-relationship) and multidimensional models are basis of the data warehouse design and development, which will be discussed in the forth coming sections.

IV. Description of Petroleum Data Sources

As shown in Fig. 2, **Fig.**there are several data sources to generate volumes of petroleum data. 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 spatio-temporal form. Spatial data represented in the form of X, Y, Z coordinates and historical data are in periods.

All the historical exploration data are from geological, geophysical and geochemical surveys carried out in different prospective areas of different basins in different periods. Surface and sub-surface geological mapping data are key geological data. Structure, stratigraphy and other geological data have significance in exploring for oil and gas. Biostratigraphy, coring, sedimentological, geochemical and reservoir data are important data in ascertaining the petroleum potentiality of basin. Reservoir, source, seal, migration, structure are critical factors of a petroleum system. Navigational, seismic or geological depth structure, well logging, vertical seismic profiling (VSP, that connects the surface and sub-surface exploration data) and reservoir are also key data for interpreting the hydrocarbon prospects and evaluating them for every basin.

Historical resources data include several decades of oil and gas exploration supporting data, drilling data, well data, original exploratory and development oil/gas discoveries, permits of petroleum licenses and production data of different wells, for different oil fields and basins and their descriptions. These data signify both longitudinal and lateral dimensional attributes. The additional dimension attributes have immense future scope of data mining, which adds value in the application domain. Three types of data structuring models have been considered, keeping in view the complexity of petroleum data, as discussed in the following sections.

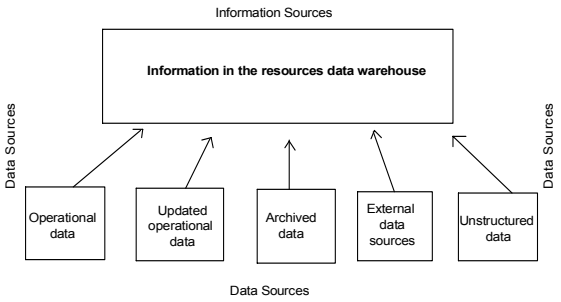

Fig. 2: Resources data sources

V. Data Structuring Models

Logical ER, multidimensional and object oriented data structuring models used in the present study have been investigated in the following sections for their application in petroleum industry.

A. Conceptual and ER models

Exploration, drilling, production and marketing are considered as supertype entities participating in an entity-relationship model (see Fig. 5). 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.

Ontology is a building block of petroleum data structure in a knowledge-domain [4] and is a basis for incremental design of logical data schemas such as ER. 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 Fig. 3 and 4) is required for designing quality logical ER models. Two of such conceptual models have been narrated, based on which the logical schemas are generated. A conceptual model and a sample of an ER diagram for petroleum exploration and production industry is shown in Fig. 5 narrating data model constructs**.**

In the present study, data relevant to activities such as exploration, drilling, production and marketing have been organized in their corresponding data structures and domains, thus an integrated metadata model evolved from different operational business activities. For each exploration, drilling, production and marketing entities, there are several associatied entities, for which the relationships mapped during metadata design process. As shown in Fig. 5, *exploration* entity (supertype) is sub divided into geology, geophysics, well logging, and reservoir engineering, logistics and inventory of sub-type entities. Diamond shape relationship is built based on the knowledge between entities. Similar such knowledge base entities can be mapped for drilling, production and marketing super type entities.

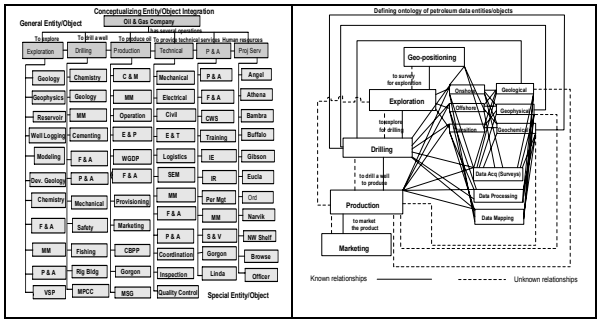

Fig. 3: Conceptual models of hierarchies in petroleum resources data

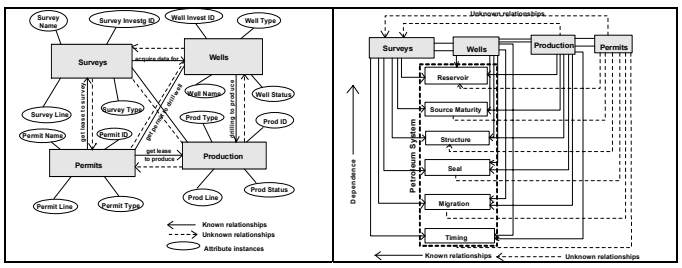

Fig. 4: Conceptual models for building logical schemas

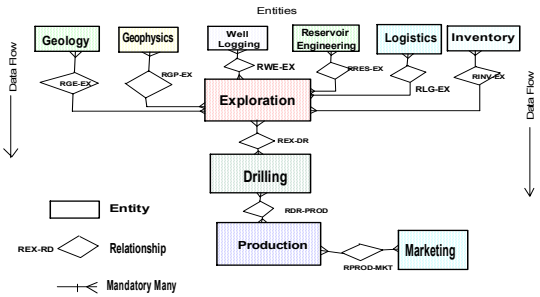

Fig. 5: ER model of petroleum industry describing the entities and relationships

B. Multidimensional logical warehouse schemas

Multidimensional data schemas are constructed for petroleum business data in a logical domain, which will further 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. All the common attributes appearing in all these tables have been used for building relationships among several dimensions and fact tables. In our case, relationships among the common attributes are denormalized, so that the final data become fine-grained for effective warehousing and mining purposes.

Similar to entities in ER modeling, dimensions are used in dimensional modeling. Fig.s 6-8 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 Fig. 6 and Fig. 7, 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 as shown in Fig. 6 and 7.

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 Fig. 6) 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. 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 Fig. 7, 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.

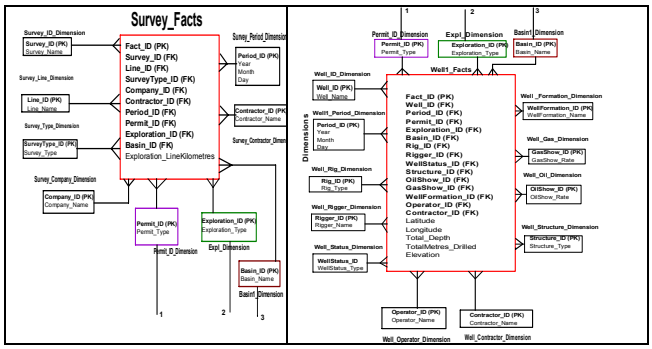

Fig. 6: Star schema models for surveys and wells dimensions and facts data

Petroleum production data dimensions and facts are also represented in the snowflake star schemas, in which relationships between common attributes of each entity are normalized as demonstrated in Fig. 8.

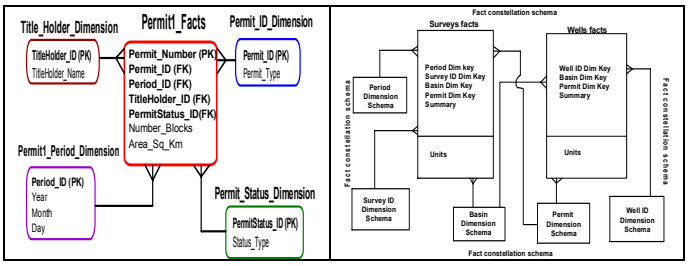

Fig. 7 (a) Star Schema Model (b) Fact Constellation schema

In our 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. Advantage of this normalized schema is easy in its usage and maintenance; normailizing also saves storage space, though navigation of petroleum data across multiple tables may not be effective due to large number of join operations.

C. Object oriented modeling

Another means of representing the petroleum exploration and production data is by Universal Modified Language (UML), [13], in which all the previously defined entities or dimensions can be interpreted as objects or class objects in object oriented modeling domain. Exploration is a superclass object. Fig. 9 demonstrates *seismic* and *wells* subclass 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.

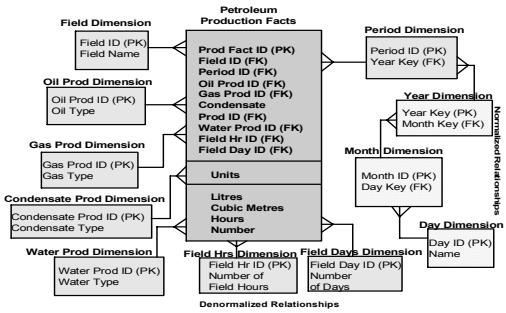

Fig. 8: Petroleum Production Business Data – Snowflake Schema Model

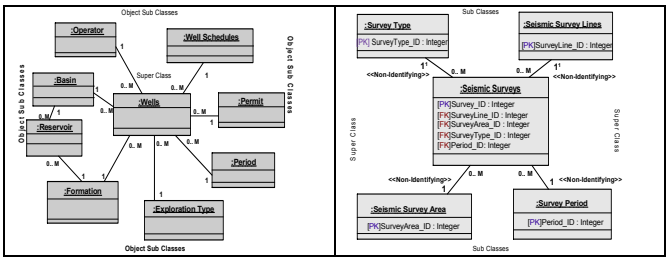

Fig. 9: Wells data – an object schema model

In the present work, data have been imported from various programming applications such as MS Excel and Access. Data cleaning and loading of data are key operations before data reach storage devices. When once data have been loaded in the form of relational data structures in an Oracle environment, SQL queries have been run and used to access data views.

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 tops, cores, tests, production, logs run, deviation surveys and well locations. The seismic module uses records of acquisition and processing history of seismic lines along with the location data. Database has been populated with existing data quickly using text, ASCII and SQL and control files (for data loading).

VI. Analysis and Discussions

Composite syntax, composite modelling, presenting its content and meaning play key role in multi-dimensional modelling of petroleum industry's data. Composite attribute integration [13] is another issue addressed in the present study, a gateway of building the logical warehouse schemas. Exploring semantics of an up-stream and integrated oil and gas company is often more complex and tedious. Ontology [4] approach facilitates integrating various facets of exploration and production data from which finer data can be 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 [5] [6] may be interpreted for business knowledge.

OLAP system provides specialized analysis tools as narrated in Fig. 10 and Fig. 11.

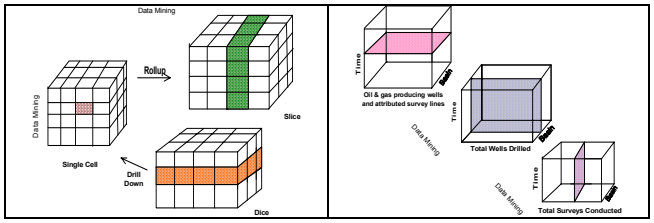

Fig. 10: Multidimensional data operations

In multidimensional petroleum data models, there could be several numeric measures for data analysis purposes. These models view the data in several cubes, more precisely named as hyper-cubes [11]. Each cube has three dimensions and each is further divided into several sub-dimensions.

A. Validating multidimensional schemas in petroleum industry

Fundamentally, metadata in the context of petroleum business data is an integrated model that communicates through different dimensions such as wells, surveys, petroleum permits and oil play factors (and or objects) intelligently. This process is practiced to achieve inter-operability or to further improve finer exploration of data carried out at later stages. The following criteria are adapted [15] for customizing the multidimensional modeling approach for petroleum business industry:

Communication: Ambiguity is minimized by intelligent communication between dimensions (or entities/objects) and their attributes and relationships (properties) and linking them by robust logic.

Inter-operability: Models (data cubes) built based on multidimensional OLAP [11], can be used on different software platforms of oil and gas industry's 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. OLAP or SQL operations can be carried out to extract piece of information without any distortion on other computing platforms (see Fig. 11).

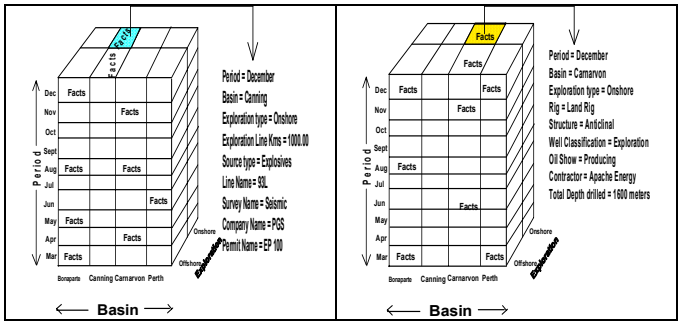

Fig. 11: Typical data views extracted from multidimensional data cubes

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.

B. Warehoused multidimensional data analysis

Once the data warehouse is modelled in the form of a multidimensional data cube, it is necessary to explore different analytic tools with which one can perform complex analysis of petroleum data. These analysis tools are called OLAP (On line analytical processing). 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.

A resources 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 [12]. In order to preserve data relationships with finer granularity, designer must ensure that data relationships are denormalized and the dimension is at its atomic level. The data warehouse project initiated for oil and gas business industry will fail 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 resources data in a warehouse will be invalid and may loose credibility. Creating the resources data for a data warehouse involves several issues. First, the data warehouse will draw data 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 resources organization's operational systems and also informational systems.

C. 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, oil play analysis of a field, resources data of other basins worldwide viewed as web documents, names of surveys conducted for spudding and drilling a well). As stated earlier, 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 motivation is to improve precision and/or recall as well as reduce the overall amount of time spent searching for data. Supporting technologies include agents [2] for searching the data, data delivery agents using meta-data languages (e.g., UML, XML and HTML), and knowledge representation tools. User triggers the warehouse to access piece of petroleum data of a drilled well, warehouse identifies the description of that data view and search engine acts to locate the data from that data warehouse.

VII. CONCLUSIONS

Warehouse schemas described in this paper are relevant to the existing petroleum organizational heterogeneous (such as hierarchical) structures and easy to implement. Multidimensional modelling of petroleum business data entities opens more scope of understanding the data mining and its acceptance in interpreting business intelligence. In any petroleum industry, exploration, drilling, production and marketing are few key areas of operations from which volume of data is collected for exploring and exploiting economic resources. The data gathered must be made valid, in a way that their responses are suitably assessed to retrieve interpretable and meaningful information, so that workable quantitative multidimensional models can be predicted for future resources industry's forecasts.

REFERENCES

[1] Beaumont, E.A and Foster, N.H. Exploring for Oil & Gas Traps, AAPG Publications of Millennium Ed., 1999.

[2] Erdmann, M. Rudi, S. How to structure and access XML documents with ontology, *IEEE Data & Knowledge Engineering*, 36, 2001, p. 317-335.

[3] Jukic, N. and Lang, C. Using offshore resources to develop and support data warehousing applications, *Business Intelligence Journal*, 2004, p.6-14.

[4] Meersman, R.A. Foundations, implementations and applications of web semantics, parts 1, 2, 3, *lectures at School of Information Systems*, 2004.

[5] Moody, L.D. and Kortink, M.A.R. From ER models to dimensional models: bridging the gap between OLTP and OLAP design, part I, *Journal of data warehousing*, 2003, p. 7-24.

[6] Moody, L.D. and Kortink, M.A.R From ER models to dimensional models: bridging the gap between OLTP and OLAP design, part II, Advanced Design Issues, *Journal of data warehousing*, 2003, p. 7-24.

[7] Nimmagadda, S L. Rudra, A Object oriented modelling approach in mapping oil and gas exploration business data for effective operational management, a poster paper accepted for *presentation in the 7th international conference on enterprise information systems*, Miami, Florida, 2005.

[8] Nimmagadda, S L. Rudra, A Design methodologies for constructing petroleum company's metadata, a poster paper accepted for *presentation in the 7th international conference on enterprise information systems*, Miami, Florida, 2005.

[9] Nimmagadda, S.L. and Rudra, A 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 in Dec 2004.

[10] Nimmagadda, S.L. and Rudra, A. Data sources and requirement analysis for multidimensional database modelling – an Australian Resources Industry scenario, *published in the proceedings of 7th international conference on IT*, held in Hyderabad, India in Dec 2004.

[11] Pujari, A.K Data mining techniques, University Press (India) Pty Limited, 2002.

[12] Rudra, A. and Nimmagadda, S.L Roles of multidimensionality and granularity in data mining of warehoused Australian resources data, *published in the proceedings of the 38th Hawaii International Conference on Information System Sciences*, held in Hawaii, Jan 2005.

[13] Shanks, G. Tansley, E. Weber, R. Representing composites in conceptual modelling, *Communications of ACM* Vol. 47 (7), 2004, pp. 77-80.

[14] Telford, W.M. Geldart, L.P. and Sheriff, R.E. Applied Geophysics, Second Edition, 1998, p.100-350 and 600-750.

[15] Uschold, M.E. Knowledge level modeling: concepts and terminology, *Knowledge Engineering Review*, 1998, 13(1).

[16] Winter, R. and Strauch, B. Demand-driven information requirements analysis in data warehousing, *Journal of Data Warehousing*, v. 8(1), 2003, p. 38-46.