

©2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

# On Data Integration Workflows for an effective Management of Multi-dimensional Petroleum Digital Ecosystems in Arabian Gulf Basins

Shastri L. Nimmagadda<sup>1,2</sup>, Heinz Dreher<sup>1</sup>, Muhammad Nawaz<sup>2</sup>, and Kamran Laiq<sup>2</sup>, Ahmed Sabry<sup>2</sup>, and Mukarram Ahmed<sup>2</sup>

<sup>1</sup>Curtin Business School – Information Systems, Curtin University, Perth, Western Australia

<sup>2</sup>Schlumberger, East Ahmadi, Kuwait

e-mail: snimmagadda@slb.com; h.dreher@curtin.edu.au

**Abstract**— Data integration of multiple heterogeneous datasets from multidimensional petroleum digital ecosystems is an effective way, for extracting information and adding value to knowledge domain from multiple producing onshore and offshore basins. At present, data from multiple basins are scattered and unusable for data integration, because of scale and format differences. Ontology based warehousing and mining modeling are recommended for resolving the issues of scaling and formatting of multidimensional datasets, in which case, seismic and well-domain datasets are described. Issues, such as semantics among different data dimensions and their associated attributes are also addressed by Ontology modeling. Intelligent relationships are built among several petroleum system domains (structure, reservoir, source and seal, for example) at global scale and facilitated the integration process among multiple dimensions in a data warehouse environment. For this purpose, integrated workflows are designed for capturing and modeling unknown relationships among petroleum system data attributes in interpretable knowledge domains. This study is an effective approach in mining and interpreting data views drawn from warehoused exploration and production metadata, with special reference to Arabian onshore and offshore basins

**Index Terms**— data integration, multidimensional data, petroleum digital ecosystems, Arabian Gulf Basin

## I. INTRODUCTION

In many petroleum industries, data are exponentially growing; with the results data integration is increasingly tedious and complex. These heterogeneous data are handled both technologically and methodologically ([1] and [2]), in which solutions are offered in terms of storage capacity, processing power and access speed. In methodological front, ontology, indexing, views, data mining and temporal database organizations. Combination of both technological and methodological approaches is led to data warehousing approach, in which complex heterogeneous databases are handled, including data integration, data structuring and data mining. Ontology based multidimensional data warehousing approach is apparently new concept for petroleum industries and they have an interesting application and commission. Authors concentrate more on methodological solutions, which definitely facilitate future technological drives and investments in petroleum industries. Follow-up studies, future scopes and implementations, are recommended.

### A. Ontology Conceptualization

Ontology is a specification of a conceptualization [1] and [6]. Ontology is set of definitions that obey the conceptualizations and relations, derived among set of dimensions. Practically, an ontological commitment is an agreement to

use a vocabulary in a way consistent with respect to the entities or dimensions described within integrated workflows and knowledge from different known or unknown domains, still under use. We have described different dimensions that commit to ontologies. Ontologies are designed that can share knowledge within and among these dimensions, described in different workflows, which undergo integration process. In the context of knowledge sharing also, the term of ontology implies a specification of conceptualization. In our case, ontology is description of the concepts and relationships that may exist within dimensions or community of dimensions. This definition is consistent with the usage of ontology as a set of concepts or definitions, more general and at places with more specifications. Ontology has significant properties for knowledge sharing among intelligent dimensions, for example, semantics independent of reader and context. Authors attempt to use this concept within integrated workflow scenarios of a sedimentary basin.

### B. Ontologies as a specification mechanism supporting integrated workflows

A sedimentary basin of formally represented knowledge, at times is based on a *conceptualization*. Several workflows from multi disciplines, such as geology, geophysics, geochemistry, rock physics, and reservoir engineering [8] possess different domains of knowledge contributing to the process of integration. These dimensions are integrated for extracting knowledge based on conceptualizations and conceptualizations. Similar analogy is made for extracting knowledge from integrated petroleum systems through conceptualizations and contextualization. These entities or dimensions 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 world that wishes to be represented for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. A dimension in an integrated workflow commits to ontology, if observable actions are consistent with the definitions of the ontology. The idea of ontological commitment is based on knowledge level perspective, which is again description of knowledge of a dimension that is independent of the symbol level representation used internally by the dimension. Knowledge is attributed to dimensions in integrated workflows by observing their actions, a dimension in the workflow “knows” something if it acts as if it has had information and is acting rationally to achieve the purpose of integrated workflows. In client/server based systems, knowledge of a dimension needs to be understood, in which client interacts with dimensions through logical assertions and queries. This has overall impact on the integrated work-

flow, handled by server based systems. A commitment to common ontology in an integrated workflow is a guarantee of consistency; including its completeness (no model or integrated workflow is complete or full-proof, with respect to queries and assertions using vocabulary defined in the ontology).

Ontologies in our present application scenarios are equated with taxonomic hierarchies of classes, class definitions, 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.

### C. Concept of an ecosystem supporting integrated workflows

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.

Each dimension in an integrated work flow, works in a unique knowledge domain, but when it starts connecting with other domains, several other conceptualized dimensions may be evolved and emerged, depending on the context or user knowledge and perception. In the context of knowledge sharing among several dimensions, in our case, surface seismic dimension shares knowledge with sub-surface dimension, similarly, the petrophysics dimension shares knowledge with surface seismic dimensions. Each dimension not only lives with it but also survives with other neighboring dimensions. The knowledge that is hidden within dimensions is in turn inter-related with dimensions of broader (global) petroleum systems and again sedimentary basins in nearby basins that are described previously with ontology descriptions. The relationships among these dimensions are not only symbiotic, but also they benefit from each other's participation. This process continues till all the knowledge extracted from all the dimensions reaches a stage that the life cycle of the designed workflow.

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 among several petroleum systems in a basin, can become mutually beneficial, self-sustaining and (somewhat) closed. This is clearly the case for Silicon Valley with the entrepreneurial industry, the venture capital industry needed to fund the entrepreneurial industry, and Stanford University, supplying the human capital needed to develop innovative/creative ideas and technologies. The goal of this ecosystem is to generate entrepreneurial ventures. Once an ecosystem is established, and is able to take first-mover advantage, it becomes very difficult for other regions to emu-

late. The region exhibits network effects and is able to establish lock-in since the switching costs associated with moving to another region are prohibitive. The collective costs of many moving out of the region (i.e. if another region tried to incentivize a large move) would be prohibitive. Thus current members have a clear incentive to remain, and new would-be entrepreneurs, venture capitalists and students interested in this industry have a significant incentive to relocate to this region.

An array of ideas from computer science, information technology, statistics and management science is being applied for organizing these data. Petroleum explorers and analysts hypothesize petroleum systems on geographic and magnitude basis. With the pursuit of acquiring more knowledge on oil/gas producing basins, authors propose existing data mining techniques for exploring data patterns. Several data mining tools are available to explore data patterns and extract knowledge

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 ([3], [7] and [9]) are used in geophysical exploration and prospecting. Broad Issues relevant to computer applications in exploration industry have been discussed, demonstrating their applicability and feasibility.

Data warehouse is more popularly designed, using multidimensional metadata conceptualization, with hierarchical and relational data structuring. Each composite data dimension refers to a conceptual schema, such as seismic-time (acquired and defined), geological-depth (already existing, inherent dimension) and velocity (conceptualized dimension) are composite data dimensions. The conceptual schema communicates with an integrated system encompassing a multilevel data warehouse system.

Sub-schemas may also be views that can be integrated in an oil and gas company to achieve a complete multidimensional conceptual schema (intelligent enough for extracting knowledge of petroleum systems). This type of fine-grained data structuring is more effective for knowledge mining from users' perspective.

## II. ISSUES AND OBJECTIVES OF CURRENT RESEARCH

1. Improve: knowledge extraction from conceptualized data (an example, patterns), having known *seismic times* and *geological depths* in a total petroleum system and faster data access and better user (geologists and geophysicists) queries
2. Establish: better connectivity among different petroleum fields and systems, through intelligent and logical conceptualization (example, time-depth-velocity relationships and their attributes)
3. Data mapping and modeling: Ontology representation and in hierarchical and relational structures.
4. Present Scenarios: our approach is new and cost effective; these are 21st century IT applications in petroleum industries that can help analyse and address issues of data management and also EHS (Environment, Health

and Safety management)

5. Future scopes: These technologies can be adapted analyzing on any heterogeneous data organizations. Present data mining studies have scope in implementing petroleum eco systems
6. To develop Ontology-base warehouse modelling- facilitating an intelligent storage and easy understanding of heterogeneous data; building basin knowledge and effective data access using data-mining algorithms
7. To develop a domain-specific data schemas (to improve knowledge of a structure/reservoir attributes that narrate classified petroleum systems in a basin)
8. To design and develop simple logical data schemas warehousing and storing large volumes of data
9. To access user specified data and attend user queries in a shortest response time based on data mining logic
10. Having known knowledge of intelligent petroleum systems, data management issues are aimed at understanding relationships among attributes of multiple dimensions, such as structure, reservoir, production data patterns that facilitated from conceptualized interpretations of workflows and petroleum systems. More popularly, dimensions are conceptualized in place of entities in a warehouse design scenario.
11. Understanding the connectivity among multiple dimensions through cluster mining techniques
12. To interpret time-depth data (translating seismic domain knowledge into depth domain) for building knowledge of seismic data patterns and thus an interpretable geological (structural as well as reservoir) data model
13. Development of large size conceptual and intelligent seismic data models for implementation in petroleum exploration industries.

Other basic research objectives broadly are:

**To develop ontology** – a framework facilitating easy understanding, mapping and modelling of complex petroleum data dimensions and build initial knowledge-structure that leads to development of data warehouse models

**Data warehouse objectives** - to design and develop logical data schemas for storing large volume of time-depth-velocity data for large number of seismic horizons interpreted for modelling Metadata to link up with data mining stage

**Data mining objectives** - to access and perceive user specific data in a shortest response time based on data mining logic and computational or data-mining algorithms including statistical techniques

**Knowledge extraction** - Understanding geological-geophysical-petrophysical connectivity (relationships) surrounding petroleum fields

#### A. Petroleum data analysis objectives

1. To interpret explored data for knowledge discovery from petroleum fields and systems, building knowledge among other several associated systems; to measure property attributes through logical data design or structuring
2. To distribute and deliver the data to several petroleum industries through interfaces/agents for easy retrieval and in required formats (XML technology proposed)
3. Carry on interpretation from field (drilled-well level, more fine-grained structure) level to basin level, using

*structure, reservoir, production* contour data patterns

4. Analyse similar property data dimensions such that complexity of information hidden under large size petroleum systems and its complexity eased.

#### B. Definitions

1. Ontology: specification of conceptualization, a formal, computer-based linguistic approximation of a shared conceptualization ([5] and [6]), building relationships among several dimensions of petroleum and geomorphic systems, geographically located
2. Data Warehouse: subject oriented, integrated, time-variant, non-updatable collection of data used in support of management decision-making processes and business intelligence [3].
3. Data Mining: in a broad sense is a process of knowledge discovery from data warehouse [9].
4. Petroleum System – an information system
5. Ontology based Multidimensional Metadata: data that describe about data. Metadata describe how and when and by whom, a particular set of data (such as, seismic time, geological depth and velocity, geological, geophysical, geochemical, reservoir engineering and production engineering) was collected, and how these data are logically structured. Metadata is essential for understanding information stored in a data warehouse and has become increasingly important in XML-based Web applications.

#### C. Data Types

1. Hierarchical, relational and networking
2. Temporal and space-varying data types
3. Non-geometric, geometric to non-geometric and fully geometric spatial dimensions (entities/dimensions/objects)

#### D. Data Warehousing

1. Data management – because of volume of data,
2. Data integration – because of multi-dimensions having multiple attributes, properties and scales
3. Data communication and connectivity – because of poorly perceived knowledge between entities/dimensions
4. So many oil/gas wells, so many oil/gas fields, so many systems and so many basins – because of so many hierarchies

#### E. Data Mining and Knowledge Interpretation

1. Attributes classified from different heterogeneous – multiple data dimensions, are explored from multiple petroleum systems – each petroleum system is an information system
2. All data classifications interpreted from multiple dimensions are made interconnected and interrelated through data integration and interoperability
3. Data classification as identified and interpreted as per data characteristics and property gets affected, if there is any affect on other classification
4. Current research investigates this phenomenon

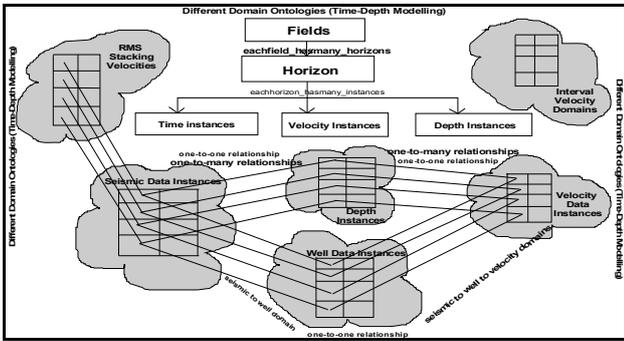


Fig. 1: Time-depth relationships for interpreted horizons, building conceptualized velocity dimension attributes

Volumes of petroleum exploration and production data, geological, geophysical and geochemical datasets are routinely acquired. Based on the evaluation of these exploration data, drilling and production are other data, subsequently built based on the exploration data. Authors concentrate more specifically, the exploration datasets such as, seismic, well-log, petrophysical, and the connectivity among these entities or dimensions through integrated workflow strategies. Interestingly, Seismic horizons (Figs.1-2) interpreted comprise of seismic data, consisting of several wavelets with crests and troughs with positive and negative peak data. These data are critically analyzed in the seismic, well-log and also petrophysical data and integrated using typical workflows. This workflow is broadly written in ontology model, in which several dimensions are described building relationships through fact data tables. These facts are data instances from several dimensions, either generic or conceptualized.

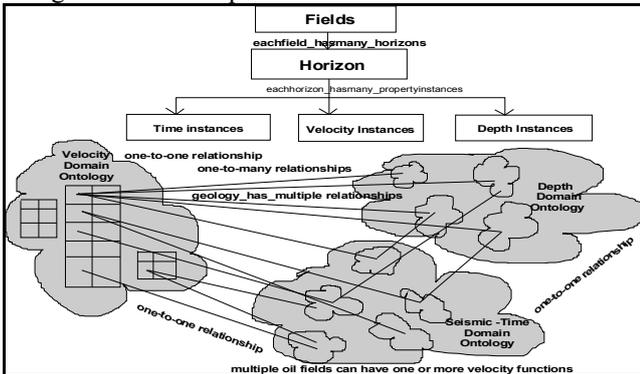


Fig. 2: Horizon domain ontologies with relationships among time, depth and velocity dimensions

### III. METHODOLOGY

#### A. Why use this methodology

Most of the data used and handled in integrated workflows are heterogeneous. 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 which these data are ontologically conceptualized in multiple dimensions, so that these valuable data are more intelligently stored and accessibility of domain knowledge is easy during data mining stage. Data integration, sharing of knowledge and interoperability are significant issues that can be addressed more robustly by data warehousing and mining approaches. Several ontology models, deduced in Figs 3, 5-6, do make up of data structuring and also data integration process.

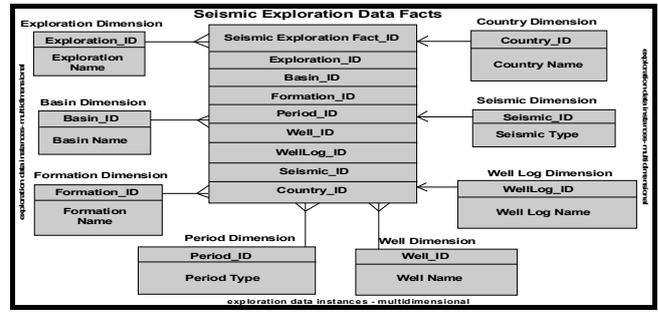


Fig. 3: Seismic Exploration Data Dimensions and their facts

Another interesting phenomenon is to design and develop domain knowledge of seismic horizons (in a warehouse modeling) that undergo data mining and interpretation stage. As narrated in Fig. 4, there are several peaks and troughs in the seismic wavelets for each and every horizon that needs to be analysed and interpreted.

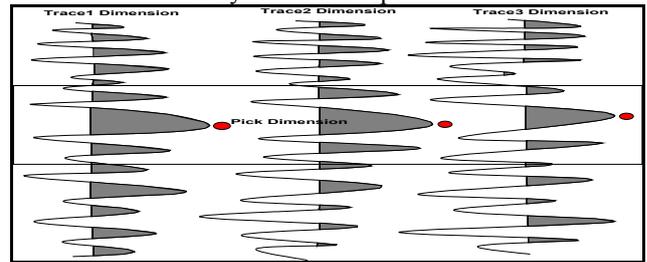


Fig. 4a: Seismic wavelets (for different trace attributes) and their polarity pick dimensions, an example showing peak (dark) and trough (white) dimension relationships, for describing horizon ontology

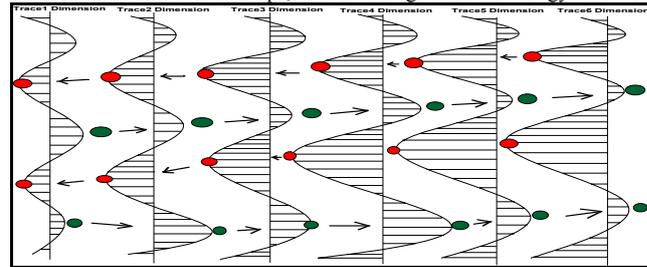


Fig. 4b: Peak and trough dimension relationships that describe horizons (for different geological structures, reservoirs and ages) and also establish horizon relationships

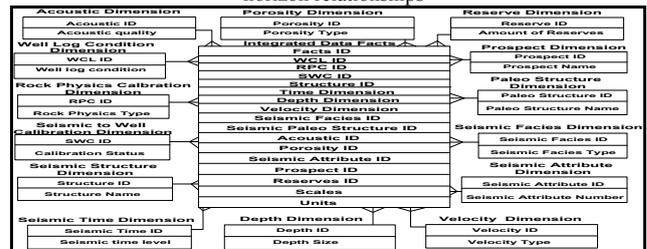


Fig. 5: An integrated data schema, narrating relationships among several sub-sets and their respective sub-schemas

Magnitude, size and dimension (both time and space varying dimensions) of these peaks and troughs of these seismic wavelets are controlling the integrity of the relationship *structure-reservoir-petrophysical* structure and their properties. When there is change in shapes and dimensions of these peaks and troughs, in a set of seismic wavelets, there is corresponding change in the properties of these dimensions. Authors make 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 Fig.7, data facts are collected and stored in a warehouse model, so that the connectivity

among multiple horizons is established. An integrated framework is designed 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 data accept. Ontology supports easy organization of these logical data structuring procedures. One of such structuring approaches described in Fig.8, is an example of organizing the data in several hierarchies. 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.

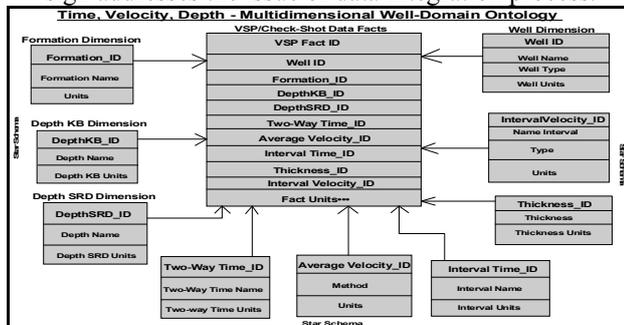


Fig. 6: Domain ontology model connecting geology-seismic dimensions

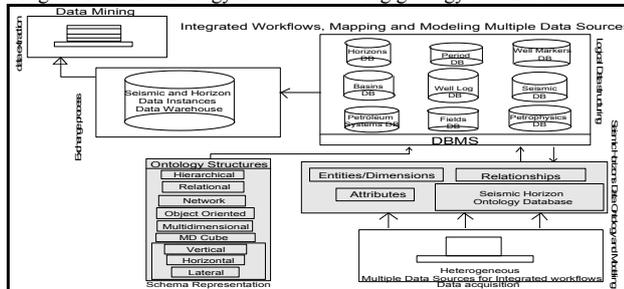


Fig. 7: An integrated framework for modeling warehouse and mining

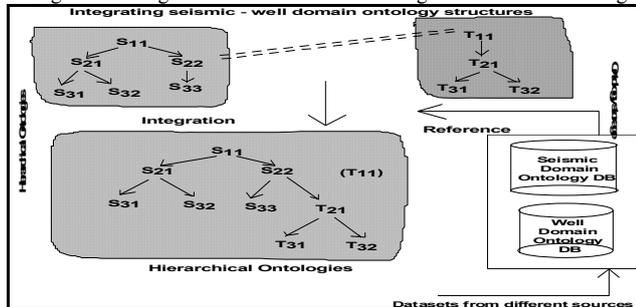


Fig. 8: A hierarchical data structure for mapping and modeling integrated (seismic-well domains integration) heterogeneous datasets

### B. What do these methodologies do:

For processing and interpreting the data, prior to integration, core program captures the processed data from several modules and then integrates data from multiple fields and

petroleum systems. During conceptual (ontology) development stage, an integrated Metadata, can save enormous computing and time resources 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<sub>2</sub>, areas of H<sub>2</sub>S, reservoirs polluted by CO<sub>2</sub> and guiding smart wells and also multi-lateral wells are few typical applications to our proposed methodologies.

### C. 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.

### D. Stages of Development

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 resources 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

- 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:

- Decomposition and/or synthesis of dimensions
- Redefinition of relationships & relationship types
- Redefinition of mapping constraints
- Redefinition of high-level abstractions (e.g. conceptualization, generalization or specialization) based on semantics and contexts
- Rearrangement of attributes among multiple dimensions (e.g. structure-reservoir-production dimensions and sets of their relationships)

#### E. Load the data into Oracle and cluster the data using cluster algorithms

- Classify each cluster to qualify understanding of each dimension
- Run SQL queries or other data mining logic – for accessing required data from warehoused metadata
- 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 resources
- Extract hidden knowledge for interpreting petroleum geology
- 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
- IT/Communication technology opportunities must be utilized to understand these systems and datasets
- To develop forecast models – for predicting knowledge, geologically interpretable, but maximise efforts in understanding petroleum fields and systems, extract knowledge from these Metadata systems.

#### F. Data Mining and Implementation of Integrated Data Schemas (workflows)

The dimension hierarchy facilitates viewing the multidimensional data in several data cube representations [3] and [4]. Conceptually, multidimensional data are viewed as lattice of cuboids. An n-dimensional data cube,  $C [A_1, A_2 \dots A_n]$ , is a database with n dimensions as  $A_1, A_2 \dots$  and  $A_n$ , each of which represents a theme and contains  $|A_i|$ , number of distinct elements in the dimension  $A_i$ . There are many data cells in the data cube, in which each cell,  $C[a_1, a_2, \dots, a_n]$  stores the numeric measures of the data for  $A_i=a_i$ . 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  $C [inline, crossline, time/depth]$  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 [8]. For

each CDP point,  $C [inline \text{ CDP}, crossline \text{ CDP}, time/depth]$ , there is unique value of  $[easting, northing, time/depth]$  value. In the case of horizon or seismic wavelet in horizon domain, a typical cell  $C [easting, northing peak/trough]$  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 Figs. 9a-9d, 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.

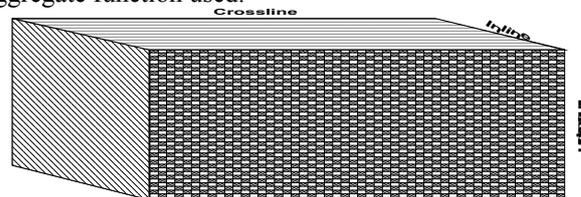


Fig. 9a: Crossline/Inline data dimensions their relationships with seismic time or geologic depth dimensions

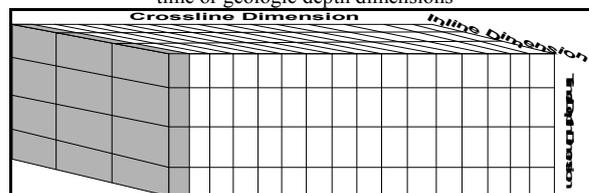


Fig. 9b: An inline dimension with respect to time/depth dimension

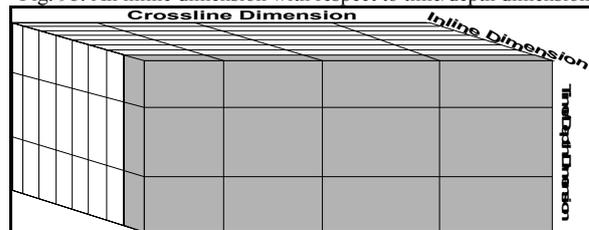


Fig. 9c: A Cross line dimension with respect to time/depth dimension

## IV. DATA VISUALIZATION AND INTERPRETATION

Data visualization ([7] and [9]) is 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 of visualizations. 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.

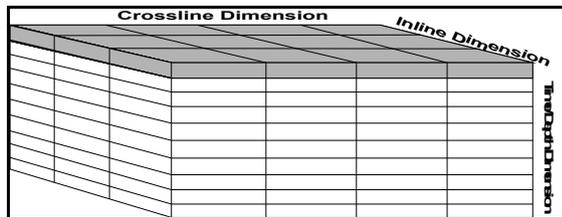


Fig. 9d: Time slice Cube

Workstation graphics are used for generating visualizations and their interpretations. As described in the Figs. 11-12, warehoused data views led to interpret in terms of geological structures their connectivity among different fields. Different horizons are interpreted from volumes of seismic data (several wavelets, with peak and trough dimensions). 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. As demonstrated in Fig. 10, all the peaks and troughs dimensions are plotted with horizon and structure attributes. The connectivity between seismic (horizons) and well data attributes has been made by data integration process. Structure is a typical data characteristics, measured for different horizons as illustrated in visualizations shown in Figs. 11-12.

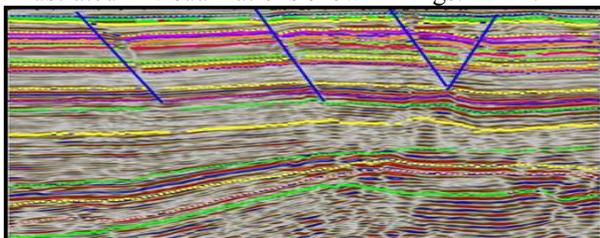


Fig. 10: An inline seismic section, showing horizon attributes

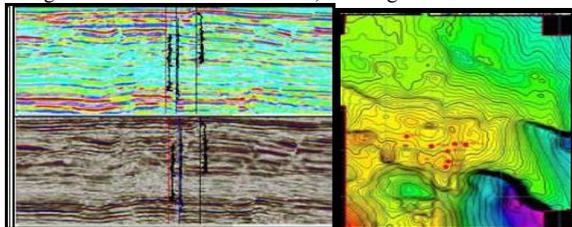


Fig. 11: Typical data integration process views

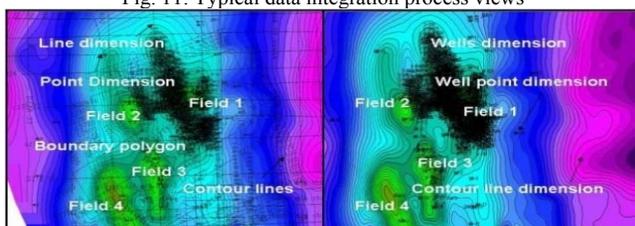


Fig. 12: Several dimensions plotted with structure property instances, extracted from warehoused metadata, showing connectivity among fields

## V. CONCLUSIONS AND RECOMMENDATIONS

1. 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
2. IT/Communication technology opportunities must be utilized to understand these systems and datasets
3. To develop forecast models – for predicting knowledge, geologically interpretable, but maximise efforts in un-

derstanding petroleum fields and systems, extract knowledge from these Metadata systems

4. Data Warehouse designed, embodies significant progress on the database integration problem for petroleum system ecosystems
5. A domain-specific (knowledge discovery: exploration and production domains; integrated framework development) Metadata (firm-up conceptualization)
6. Logical and implementation data models for warehousing the large volume of petroleum data
7. Data mining algorithms (for accessing data, interpret them for prediction knowledge)
8. Use of application and its evaluation in petroleum exploration companies

## VI. FUTURE SCOPE OF WORK

Implementation of an integrated petroleum metadata model consisting of different exploration and production data dimensions and attributes, is in progress

Application of association rule mining among petroleum system factors; to build the connectivity among systems of different geographic regions; maximise the predictability of petroleum production by favourable geology are in progress. Construction of decision trees, computing classifiers, determining the accuracy of classifiers among sets of exploration and production data items are planned. Petroleum data clustering – understanding and knowledge building among different petroleum systems, is under investigation

## VII. TESTING AND EVALUATION

Economic measures of ontology-base warehousing in the petroleum industries need to be implemented and assessed. OLAP Models are to be implemented in multi-disciplinary application scenarios.

## VIII. REFERENCES

- [1] Kalfus, O., Ronen, B. and Spiegler, I. A selective data retention approach in massive databases, *OMEGA*, 32 (2004), 87-95.
- [2] Thomas, J.L, Yannick, P. Valerie, W., Gupta, P. Stringer-Calvert, D.WJ, Tenenbaum, J.D and Karp, P.D. *Bioinformatics database warehouse toolkit*; BMC Bioinformatics 2006, 7:170, p.1-14; <http://www.biomedcentral.com/1471-2105/7/170>
- [3] Hoffer, J.A, Presscot, M.B and McFadden, F.R (2005). *Modern Database Management*, 7<sup>th</sup> Edition, Prentice Hall.
- [4] Nimmagadda, S.L. and Dreher, H. (2006) *Ontology-base data warehousing and mining approaches in petroleum industries*: in Negro, H.O, Cisar, S.G. and Xodo, D. (Eds.), *Data Mining with Ontologies: Implementation, Findings and Framework*, a book published in 2007 by Idea Group Inc. <http://www.exa.unicen.edu.au/dmontolo/>
- [5] Gornic, D (2000) *Data Modeling for Data Warehouses*, Rational Software White Paper, [www.rational.com/worldwide](http://www.rational.com/worldwide).
- [6] 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.
- [7] Plastria, F. Bruyne, S. D. and Carrizosa, E. (2008). *Dimensionality reduction for classification: Comparison of techniques and dimension choices*, published in the 4<sup>th</sup> International Conference, *ADMA 2008*, Chengdu, China, October, 2008.
- [8] Magoom B. L and Dow, D.W (1994), *the Petroleum System – from Source to Trap*, AAPG Memoir 60, p. 1-625.
- [9] Shawkat Ali, A. B. M. and Wasimi, S. A. (2007) *Data Mining: Methods and Techniques*, p. 196-219 and p. 25-267