8th Biennial International Conference
& Exposition on Petroleum Geophysics

"HYDERABAD 2010"

P-280

# Modeling Multidimensional Australian Resources Data for an effective Business Knowledge Management

***Shastri L NIMMAGADDA* and Heinz DREHER***
*School of Information Systems, Curtin University of Technology, Perth, Western Australia*

***Summary***

*Historical Australian resources (exploration and production) data are stored in data warehouse environment in the form of relational and hierarchical data structures in multiple dimensions. Significantly, these resources databases consist of periodic dimension, characterizing the role of period and its relation among other data dimensions, their attributes and fact tables. Data mining of periodic data instances in resources industry is an emerging discipline that can map business knowledge from variety of very large databases. Several materialized data views are accessed from the resources data warehouse using various data mining procedures for discovering data, links, associations and patterns; interpretation of these patterns (such as periodicity, seasonality, or cycles) that led to predictions for future business forecast. Mining models generated among multiple dimensions, will facilitate managers of decision support personnel for making future predictions. This present study extracts business intelligence from historical data, which is presented in terms of data visualization, an approach of business knowledge representation and interpretation.*

## 1 Introduction

Australian resources data are archived in different hardware platforms and or even in hard copies and at times difficult to retrieve, managing them in industry operations. Enormous resources data, when ontologically ([10] [11] and [12]) structured and warehoused them, have immense value in information extraction and further online analytical processing (OLAP) capabilities [8]. Typically, *surveys*, *wells* and *permits* data in different basins are handled by operating companies. All these datasets are heterogeneous, when logically organized and intelligently stored, so that data mining and data visualization can be more effective for interpretation of resources business data. Australia has vast mineral and petroleum resources untapped in different basins. These technologies are available, but not fully exploited in the operating companies so far. Authors attempt to design and develop these technologies in Australian resources industry situations.

## 2 Problem Statement

With the widespread use of databases and the explosive growth in their sizes, producing and servicing companies in Australia are faced with the problem of information overload and maintaining volumes of database repositories. Data integration and interoperability are other critical issues. A unified, standard interface to manage and share all business and technical information of integrated exploration and production model is needed and data warehousing and mining are problem solutions. All the *surveys* (including *seismic* and other *geological* data), *wells* and *permit* data are ontologically conceptualized and organized, developing multidimensional data structures [8]. These models are integrated and stored in a warehouse environment intelligently, for data mining. Fine grain structuring [14] is significant criteria for effective data mining. Real issue is, understanding the data dimensions, conceptualizing and integrating them for mining associations and links among data instances. Building knowledge and interpreting it for meaningful business information (as desired by explorers and operational managers), are other issues. Data mining and visualization, which are major challenges, are addressed in the current study. OLAP engine within a data warehouse provides an adequate interface for querying summarized and aggregated information across different dimension hierarchies.

*Presently with the Schlumberger Oilfield Eastern Ltd, P O Box 9056, East Ahmadi, 61001, Kuwait**
*Email: snimmagadda@kuwait.oilfield.slb.com/heinz.dreher@cbs.curtin.edu.au*

## 3 Methodologies and Solutions

The data mining research provides several procedures, which can explore volume of heterogeneous data and reveal different forms of interesting data patterns, including unknown data trends and associations. Data warehousing approach when combined with data mining has been extremely powerful in understanding data patterns from huge amounts of data. In the present study, heterogeneous data are structured and organized in the form multidimensional data schemas. Ontology addresses features of conceptualization, contextualization, specification and semantics. Ontology based multidimensional heterogeneous data structures that can accommodate in a warehouse environment ([3] and [10]), have succeeded in providing standard modeling approach upon which reusable analytical tools and other data mining can be designed and implemented. *Surveys*, *wells* and *permits* data are considered for generating multidimensional models. As narrated in Figs 1-4, all the "entities" that described in ER modeling [9] are analogously conceptualized and contextualized into multiple "dimensions".

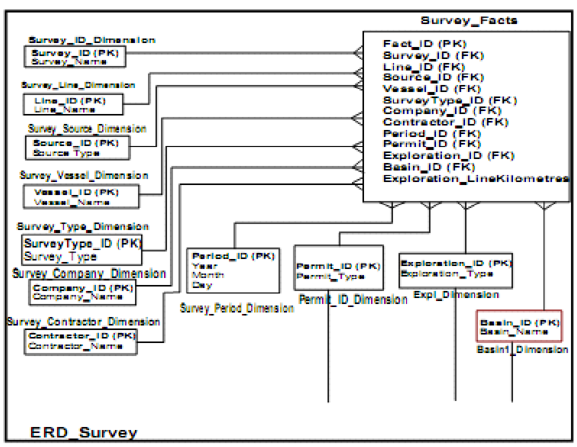


Fig. 1: Star Schema of "Survey" dimension

Seismic survey data, including *company*, *contractor* and *permits* data, are categorized in multiple dimensions, based on business rules and constraints [8]. As described in Figs. 1 and 4, *surveys* and *wells* data facts and instances are structured in a star schema model, in which dimensional data possess one-to-many relationships among with fact data tables. Similar structure models are designed for mineral discoveries and mineral exploration cost dimensions for Australian situations, as narrated in Figs. 2 and 3.

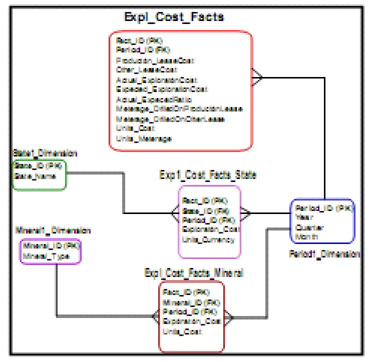


Figure 2: Star schema "Exploration Cost" dimension

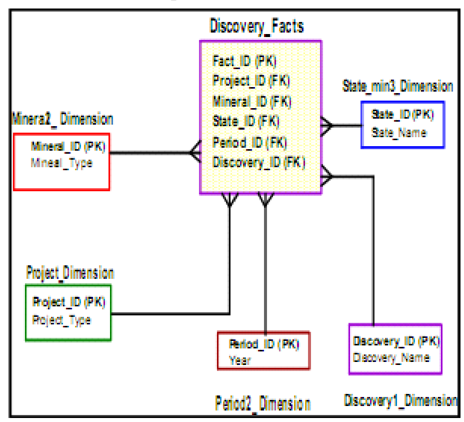


Figure 3: Star schema of "mineral discovery" dimension – fact

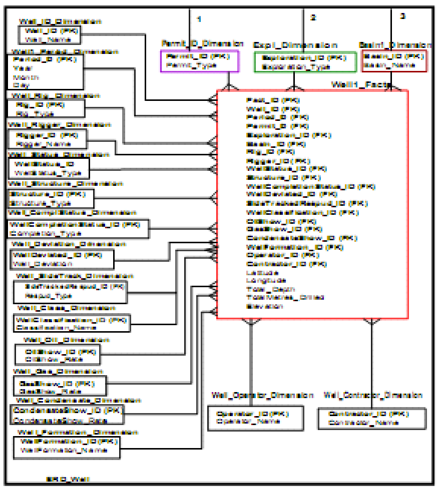

Figure 4: Star schema – "wells" dimension facts

### 3.1 What can Data Mining do for the Exploration Industry?

It is a non trivial process of identifying, valid, novel, potentially useful and ultimately understandable patterns in the exploration and production data. Once all *surveys*, *wells*, *permits* data are warehoused in the form of multidimensional structures, data views can be extracted by data mining algorithms that have been discussed in [1] and [2]. Ontologically derived data are used exploring data patterns and correlations among several petroleum data attributes. Besides ontology modeling, the mining scheme, as discussed in the Fig. 6, is effective in analyzing data views (extracted from a warehouse). Line and bubble plots, representing resources data instances of attributes of multiple dimensions with multiple associations, are used for building and interpreting knowledge. Data mining scheme discussed in the Fig. 6, is classic and more traditional, even though ontology based warehouse enables a more effective data mining and brings up links and associations among data instances. Correlations, trends and patterns that are brought out from these data instances, are replicate of links and associations.

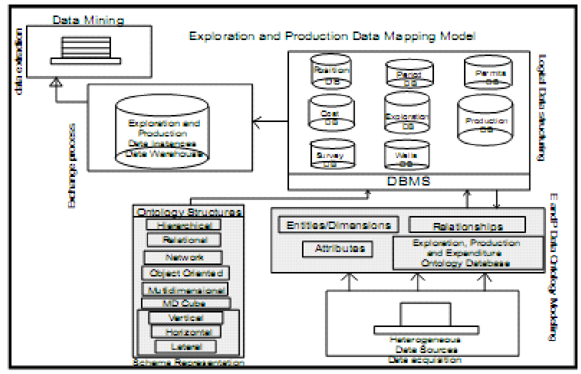


Figure 5a: An integrated framework for ontology modeling and mining

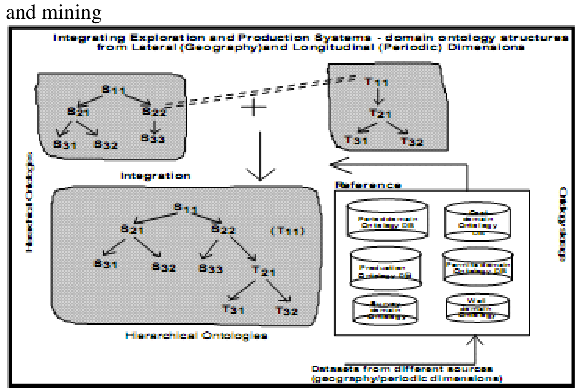


Figure 5b: Illustration of hierarchical ontology

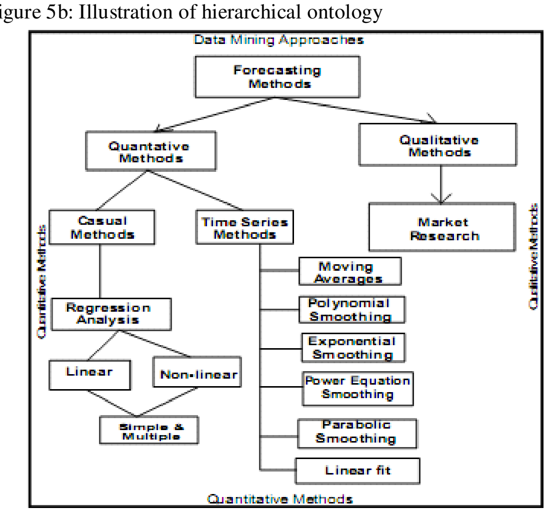


Figure 6: Data mining scheme

Oil and gas and other mineral resources exploration and production data that have been analyzed ([10], [11], [12], and [13]), are briefly discussed in the forthcoming sections. Today, resources business analysts have relied on reports and ad hoc query tools to bring together useful information from the exploration, drilling and production data. Since exploration data volumes continue to pile up, finding valuable information from these volumes is an overwhelming task. Data mining approaches mentioned in Fig. 6, though traditionally used for analyzing trends and aggregations in the specialized data [14] in the form of links and associations, these associations can further be analyzed more precisely using decision-trees, clusters and mining rules [13].

As stated before, exploration, drilling, production and marketing data are acquired from heterogeneous sources and these data are ontologically structured using logical relational, hierarchical (Fig. 5b), lateral, horizontal and vertical structuring in a Oracle driven database [3]. These structured data are further relationally organized and stored with multiple data links and associations. Once conceptualized data relationships are integrated and stored in a warehouse, it is ready for mining data views. This is an integrated framework (Fig. 5a) used in the present study.

### 3.2 Time Domain Analysis of Historical Data

Data extracted laterally and longitudinally ([1], [4] and [5]) from databases at different historical times and places, have been input into data mining procedures to generate interpretable information. It is planned to explore the fluctuations among the component factors of the time series data annually, quarterly or monthly since database contents vary over time. For example, in a database that contains oil and gas exploration, drilling and production information, the expenditure for each product (entity) may be varying as material, labour costs as per market conditions change. These changes could be due to change in technology innovations in periods of time. If only a current cost is required, then the cost of these items can be modelled as a single valued attribute. However for future financial decisions and other documentation purposes, there is need to preserve history of these costs and time period over which, each cost was in effect. This can conceptualize a requirement as a series of costs and the effective date for each cost. This result is a (composite) multi-valued attribute named "Actual and Expected Expenditure in Exploration_Expenditure" entity and their corresponding periods. An important characteristic of such a composite, multi-valued attribute is that the component attributes go together. Each value of the attribute cost is time stamped with its effective date. A time stamp is simply a time value (date) that is associated with a data value. A time stamp may be associated with any data value that changes over time when there is need to maintain a history of data values. Time stamps may be recorded to indicate the date at which the value was entered (quarterly expenditure, for example), the date the value becomes valid or stops being valid, or the date when critical actions were performed (such as updates, corrections, or audits). Data transformations [8] can be performed on this data for user queries. OLAP and Data Visualization can be used for learning and discovering data. As a matter of fact, time domain modeling may generate complexities, because expenditure of each item that is analyzed could be due many other factors, which cannot be explained. There might be some data missing and could be inconsistent at different periods of time. In another example, database that contains exploration cost of mineral and petroleum, production rate of oil producing well and number of surveys and wells drilled on a prospect, may flexibly be changed as costs increase and prospect conditions change. If only current values of these attributes are considered, then these valued can be modeled as a single valued attribute. However, for informational modeling purposes, history of these attribute values with period, is needed. In our case, it is conceptualized as a requirement with series of exploration expenditure effecting each period, drilling or production rate effecting the date as well. Multi-valued attributes are derived in each case by the process of normalization into finer dimensional attribute values. For example, daily exploration cost, quarterly, monthly and yearly costs. In all these multidimensional time data, each value of attribute is time stamped with a specified period. Time values associated with other multidimensional data values have significant role in the data visualization and interpretation, thus finally implementation of warehouse application with time series data.

In our case, a time stamp [5] is associated with periodic data value or instance of exploration and production data that change with time while maintaining the historical data values. Time stamps may be recorded to indicate the time the value was entered; the time the value becomes valid or stops being valid, or when the critical actions (update, corrections, or audits) were performed. It may so happen in the middle of the year, due to reorganization of the exploration function, or outsourcing this function, expenditure data may change. It could be assigned to a new exploration line. Based on the present data, and discussions made with resources company managers, current data models are inadequate in handling the time-dependent data and some organizations ignore these problems. Due to complexities in time data modeling, data warehousing applications are designed to remove uncertainties by providing explicit designs for time dependent data. One has to be alert to these complexities posed by time-dependent data as data models are designed with time stamps and constraints in the resources organization.

## 4 Results and Discussions

Grapher, Surfer and Sigma are used to present and interpret the data views for exploring data patterns and correlations among datasets. These views are interpreted for business knowledge. Mineral exploration costs patterns have been examined for Queensland (QLD), Western Australia (WA), Rest of Australia and Northern Territory (NT). These patterns possess similarity (see Fig. 7) in their responses particularly for the years 1970, 1983 and 1990. Three peaks observed in these periods, indicating more exploration activity in these states.

As narrated in [6], data patterns are computed from the extracted data views of the warehouse, in order to understand relationships and links among exploration and production data attributes and instances as demonstrated in Figs. 7-17.

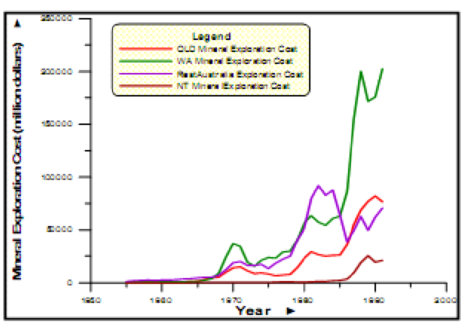


Figure 7: State wise mineral exploration cost data analysis

As shown in Fig. 8, in general, response patterns of base metal discoveries are highly irregular compared to the Nickel and Gold discoveries. More exploration activity of Nickel discovery has taken place (because of rising prices and more demand) between years 1965 and 1983 with decreasing trend leaving maximum number of discoveries around the year 1970. The situation with Gold Exploration is different and more discoveries have been made in between 1978 and 1990 as interpreted by increasing trends (see Fig. 8).

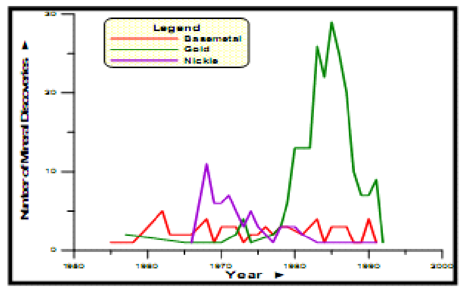


Figure 8: Analysis of base-metal, gold, and nickel mineral discoveries data

Fig. 9 narrates the state wise mineral discoveries in Australia. Western Australian state has dominated the mineral discovery status compared to other states. NT, QLD, NSW (New South Wales) and Victoria states follow the other mineral discovery status. There are prominent peaks in the discovery status interpreted in the years 1963, 1968 and 1985, since these periods provided good economic boom in the Western Australia and Australia, with more mineral exploration activity, demand and competitiveness between companies. Major upsetting periods are 1975 and 1997, as interpreted by the presentation of response curves (see Fig. 9). As shown in Fig. 10, the original mineral discoveries made in different periods have been interpreted and how the development of these mines has taken place during the process of original discovery. One can carefully examine the patterns of various mineral discoveries, especially during 1963, 1969 and 1985 with peaks in number of mineral discoveries.

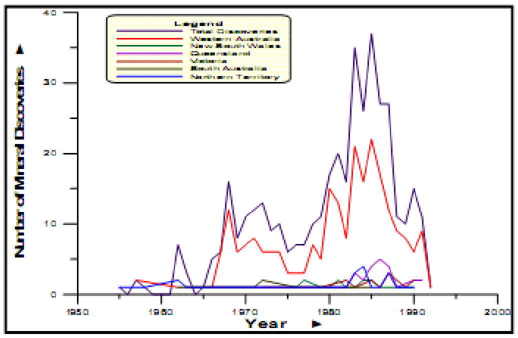


Figure 9: State wise mineral discovery data of Australia

There are random fluctuations in discovery status in all their troughs and peaks, due to seasonal, and or external economic pressures. Original discoveries have been made between periods 1965 and 1974 and again in between 1977 and 1988 years. Development of mines has taken place more during 1980, when there are no original discoveries made, two significant one in 1982 and the other 1990 (see Fig. 10 for more details).

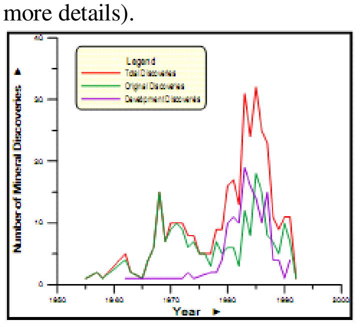


Figure 10: Analysis of mineral discoveries data

Offshore and onshore petroleum Exploration Costs data have been plotted as illustrated in Fig. 11, to compare and establish any correlation between these two independent variables. Offshore exploration is more expensive than the onshore exploration costs, because of its infrastructure and operational complexities (see Fig. 11).

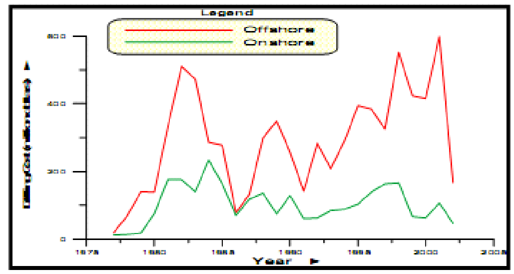


Figure 11: Correlation between offshore and onshore drilling cost data

Correlation analysis done between *surveys* and *hydrocarbon producing wells* provides a good correlation particularly in the years 1963-1975, 1980-1985 and 19862000 as demonstrated in Fig. 12. Similar good correlation is observed between the attributes *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 Fig. 13.

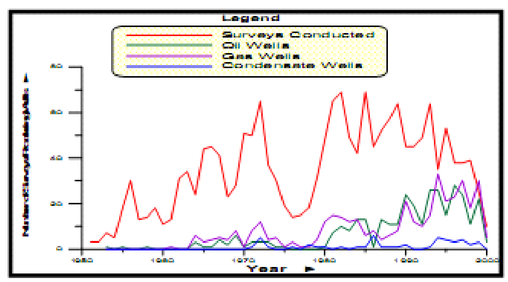


Figure 12: Correlation between number of surveys and number of oil producing wells data

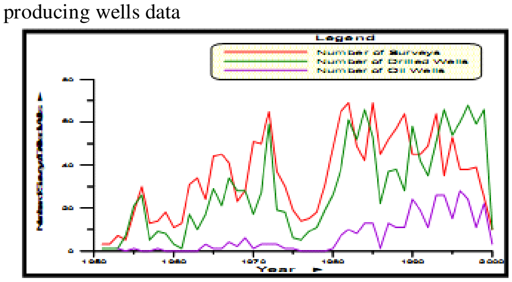


Figure 13: Multidimensional data correlation analysis among number of surveys, number of wells drilled and number of oil producing wells data

*The number of surveys* carried out has been plotted with period as shown in Fig. 14. There are three peaks one between 1954 and 1960, 1961-1976 and 1978-2000. In general, there is general increase in the trend of *number of surveys* from 1953 to 1985. Since 1990, there is decrease in the trend in *number of surveys*. Similarly Carnarvon, Canning and Perth basins have dominated with increase of *number of surveys* (see Fig. 14 for details). Similar analysis is made for the number of wells drilled with period as narrated in Fig. 15.

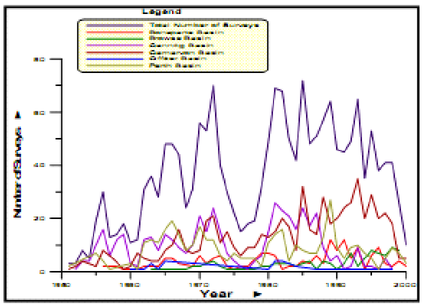


Figure 14: Correlation analysis of number of surveys conducted for petroleum exploration in WA State

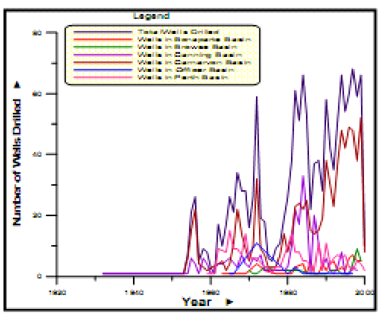


Figure 15: Correlation analysis of data on number of wells drilled in different Western Australian basins

*Number of wells* and *number of condensate wells* plotted has provided a good correlation in the years 1971-72, 1993-2000, but fait correlation in between years 1985-1990 (Fig. 16). 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 (Fig. 17). Number of drilled wells vs. number of oil producing wells has been correlated. These two variable attributes have dependence each other and has quite good correlation as illustrated in Fig. 18. In general, the number of surveys conducted and wells drilled in Australia do match, except in the years 1972 and 1997.

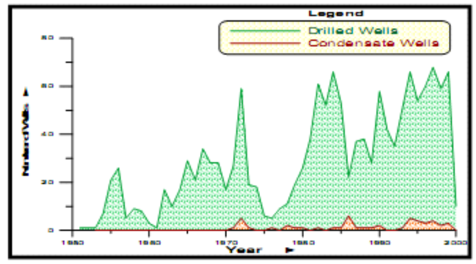


Figure 16: Correlation analysis between number of wells drilled and condensate producing wells data

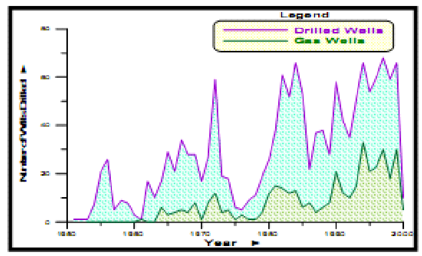


Figure 17: Correlation analysis between number of wells drilled and number of gas producing wells data

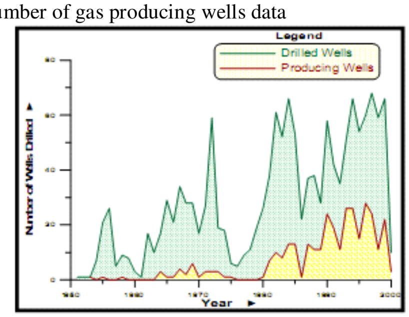


Figure 18: Correlation analysis between *number of wells drilled* and *oil producing wells* data

### 4.1 Multidimensional scaling for similarities of dimensions

Multidimensional Scaling (MS) [7] is intended to identify and examine the underlying exploration data 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 have been appropriately understood. MS identifies and also ranks relationships described in the data more explicitly. As discussed in [7], parabolic and polynomial equations have been constructed between different dimensions and their attribute values of Mineral Exploration data. There are many similarities (as well as dissimilarities) in between these equations and thus dimensions in the data. Similarly, for petroleum exploration data, several statistical fits have been constructed between different dimensions providing several coherencies and similarities in the characteristics of the equations as illustrated [7] and [9]. In view of these coherencies and similarities in these data, several inherent relationships are interpreted [2] and [5].

### 4.2 Classification by cluster of bubble plots

Grouping process is accomplished 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 modeling used to place items into groups, having similar characteristics of their attributes. In the present study, bubble plots with similar sizes have been clustered into groups as shown in Figs. 19-25. Bubble plot displays two variables on a scattertype 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 data mining features. The bubble line and fill colour are set in the bubble plot properties of the grapher program. On the bubble plot, min radius and max radius fields are set to 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. Intermediate values are displayed proportionally between these two values as shown in Figs. 19-25. Clustering has many applications in biology, medicine, anthropology, marketing and financial institutions. It has ability to process images, recognising data patterns, in oil and gas exploration applications, in particular with geochemical and geo-statistics prospecting. In the present study, an attempt has been made to create data views of clusters using the present resources data.

These plots have been examined for certain classifications in groups of data from different clusters. This sort of cluster analysis identifies and classifies variables, in a way each variable is very similar (in bubble size) to others in its cluster with respect to the predetermined selection criteria.

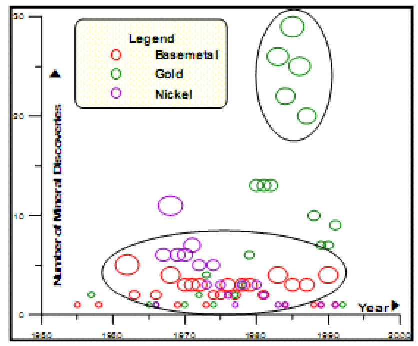


Figure 19: Bubble plot among basemetal, gold, and nickel mineral discoveries data

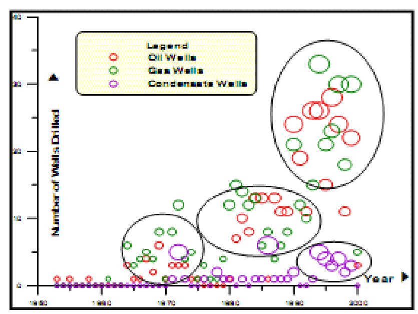


Figure 20: Bubble plot among number of oil wells, number of gas wells and number of condensate wells data

This is presented in a graphic form as a bubble plot in Fig. 19 for mineral exploration data. Partitioning of individual or group of similar clusters that may represent to similarity in time-varying variables further signifies properties or characteristics of these time-dependent variables.

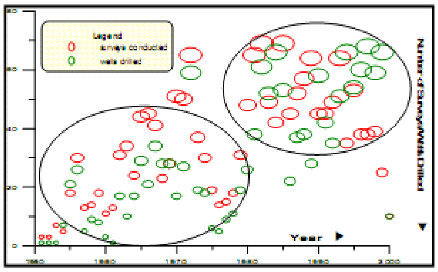


Figure 21: Bubble plot between number of surveys and number of wells drilled data

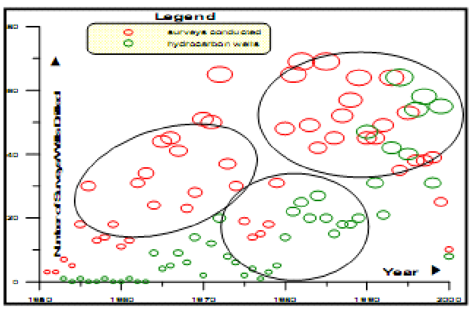


Figure 22: Bubble plot between number of surveys and number of hydrocarbon producing wells data

One can critically examine and interpret the size of the bubble and the direction at which their sizes are varying, so that data relationships, for example, between *number of drilled wells* and *producing wells* and *number of surveys* with varying *number of drilled wells*, are precisely known and understood. Similar characteristics have been observed for petroleum exploration data as shown in Figs. 20-25. As demonstrated in Fig. 21, based on the surveys conducted in later periods of 1990, many wells have been drilled. It is interesting to notice bubbles of different data attributes and their instances, very close together as shown in Figs. 19-25, which suggest most of the wells drilled in these basins are hydrocarbon bearing. As shown in Fig. 24, it is significant to deduce that the geological structures interpreted, have proven to be hydrocarbon bearing, especially during periods between 1990 and 2000. During 1995, number of wells drilled in the Western Australian basins, produced different types of hydrocarbons, such as oil, gas and condensate.

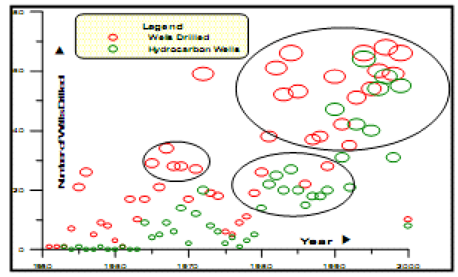


Figure 23: Bubble plot between number of wells and number of hydrocarbon producing wells data

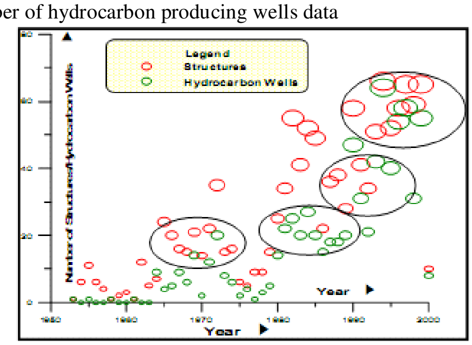


Figure 24: Bubble plot between number of structures and number of hydrocarbon producing wells data

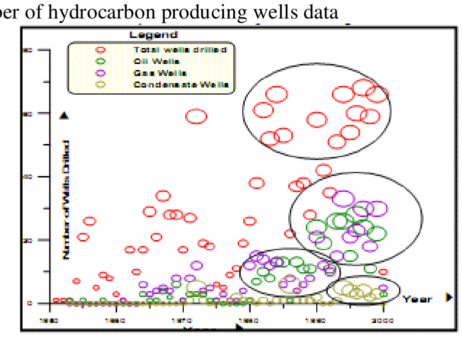


Figure 25: Bubble plot between number of wells drilled and number of oil, gas and condensate wells data

## 5 Conclusions and Recommendations

Data integration and issue of interoperability are addressed in this paper. Ontology based data warehouse approach is feasible and flexible with large size databases. Historical data are analyzed, providing interesting trends, correlations and patterns among multi dimensionally modeled data. Several interesting trends have been drawn from the resources data. These trends are useful in terms of extracting business knowledge and interpreting it particularly in predicting the future resources forecasts. It is recommended to utilize these studies in other basins else where in Middle East and Asia regions, where volume of data are available.

**References**

Berenson, M.L and Levine, D.M. (1992). Basic Business Statistics, Concepts and Applications, fifth edition, 1-953p.

Dunham, H. M. (2003). Data Mining, Introductory and Advanced Topics, Prentice Hall Publications, 10-200p.

Gornic, D (2000) Data Modeling for Data Warehouses, Rational Software White Paper, www.rational.com/worldwide.

Graham, P. and Desmond, J. K. (1992). The Use and Misuse of Statistical Methods in Information Systems Research, *Information Systems Research*, pp.208-229

Gregersen, H. and Jensen, C. S., (2002). Conceptual Modelling of Time-Varying Information, http://powerdb.net/database. (Date accessed: 15th March 2009)

Gupta, S.P., (1990) Practical Statistics, M/S Chan & Co Publishers, 1-563p

Hair, F.J, Anderson, R.E Tatham, R.L (1984). Multivariate Data Analysis, 2nd Edition, Maxwell Macmillan Publishers, 1-449p

Hoffer, J.A, Presscot, M.B and McFadden, F.R (2005). Modern Database Management, 7th Edition, Prentice Hall.

Nimmagadda, S.L. and Dreher, H. (2006) Ontologybase 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 published in 2007 by Idea Group Inc. http://www.exa.unicen.edu.au/dmontolo/

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 4 International Conference of IEEE Industry Informatics*, held in Singapore, August.

Nimmagadda, S.L. and Dreher, H., Rudra, A (2005a) Ontology of Western Australian petroleum exploration data for effective data warehouse design and data mining, *Proceedings of 3rd international IEEE conference on Industrial Informatics*, held in Perth, Australia.

Rudra, A and Nimmagadda, S.L. (2005) Roles of multidimensionality and granularity in data mining of warehoused Australian resources data, Proceedings of 38 Hawaii International Conference on Information Sciences, Hawaii, USA.

Pujari, A.K. (2002) Data Mining Techniques, Universities Press, p. 114-147 and p. 251-279.

Shawkat Ali, A. B. M. and Wasimi, S. A. (2007) Data Mining: Methods and Techniques, p. 196-219 and p. 25-267.