Roles of Multidimensionality and Granularity in Warehousing Australian Resources Data

Amit Rudra and Shastri L Nimmagadda
School of Information Systems
Curtin University of Technology, P O Box U1987
Perth, WA, Australia
Amit.Rudra@cbs.curtin.edu.au

Abstract

Granularity of data modeled in multidimensional data structures is an important factor for a data warehouse. Grain sizes and number of dimensions participating in the model are critical in ascertaining the quality of analytical queries that are run on such data warehouses. In this paper, exploration and production data of Australian resources industry, pertinent to oil and gas, over the past five decades have been examined for multidimensionality and grain size. This research shows how using an ER approach combined with multidimensional data modeling helps in considerable reduction in the size of the data warehouse, making it more effective and efficient.

1. Introduction

Mapping of multidimensional data is often carried out in the data warehousing applications. This could be achieved in a couple of ways. First, one can view a multidimensional cube in which each cell stores one or more values of simple attributes and dimensions. Users may be interested in aggregating these attributes or dimensions in the form of summary or segmenting the data such as time periods, costs of exploration, functions of resources industry or people. A second and equivalent way is to use a star schema. In this case, several dimensional tables surround a fact table. A fact table's primary key is made up of primary keys of all the related dimensional tables. The surrounding dimensional data should have already been categorized (such as description of the petroleum company or contractor involved with exploration of natural resources). A key design issue is identifying the lowest level of each dimension participating in data model by which the users desire to aggregate certain measures.

Research involving oil and gas exploration data warehousing has been very rare [6]. However, the details of Li et al's research [6] are published in a language other than English. While for some of basic issues one can consult a number of work and studies, there are some peculiarities that are typical of oil and gas industries. For example, Jarke et al [4] describe data warehouse design solutions and present several case studies in the industrial domain, while Dunham [2] discusses computing algorithms used in the data mining with several case studies. Fayyad et al. [3] also furnish computing algorithms for analyzing the time series data accessed from large databases and extracting knowledge or business intelligence from the periodic data. Others like, Mattison [7] and Moody and Kortink [9] discuss designs of data warehouses and their implementation in the industrial environments. However, one must note that several studies have been conducted in the area of warehousing geographic data [1, 5, 8, 9, 10]. Saarvenvirta [10] demonstrates design and analysis of multidimensional geographic data. Miller and Han [8] provide an excellent insight into complex spatial data mining, extracting knowledge from volume of databases. This research is an attempt to study some of the special data warehousing needs of the oil and gas industry. In particular, it attempts to study the design aspects of granularity of fact tables and related dimension tables based on real-life data obtained from the Australian resources industry.

The rest of the paper is organized as follows. The following section looks at the methodology used in the study. Section 3 presents an analysis and discussions on the granularity of fact and dimension tables of resources data obtained from various states (NSW, QLD, SA and WA) of Australia. It discovers and determines how the units of measurement used during the various phases of oil and gas exploration, viz. survey, well drilling and permit allocation can affect the design of a warehouse. Finally, section 4 provides the conclusion and some recommendations of the study.

2. Methodological views

Resources data, due to its very nature, involve a large number of dimensions, fact tables and different data types. Storing historical information of such businesses could result in massive data warehouses. Needless to say, this may be very useful to extract business intelligence for day-to-day operational decision-making purposes for the business to stay competitive in the market. For example, if an exploration manager needs analyzing the number of surveys conducted in a period of time, he or she must access survey data by state, by basin (within that state) and by exploration mode. This may entail drilling down resources data by state in order to compare surveys conducted. In contrast, rolling-up the data would provide a view aggregated to a higher level.

As stated earlier, multidimensionality and granularity are two critical issues for data organization (both logical and physical) in the resources industry. These issues have to be addressed effectively in order to accomplish improved analysis of warehoused resources data. The following sub-sections discuss some methodological views in the context of resources industry.

2.1 Multidimensionality

In order to understand and visualize the nature of geophysical data a dimensional model provides an increasingly common view to data warehousing applications [7]. Star schemas that gained popularity provide a query centric view of the resources data [4]. They rely on classifying information as either facts or dimensions. A data warehouse, handling numerous dimensions and several fact tables, also distinguishes informational and operational resources data. Current transactions include daily well site reports, day-to-day survey operational data and daily oil and gas volume and pressure status of the well-built reservoirs. Operational management, such as survey unit level technicians or desk operators in the field parties, is concerned with day-to-day transactions. While highly skilled personnel responsible for making periodic business decisions handle analytical data. Several dimensions have been recognized from all these activities.

Multidimensional nature of some of the measurable resource data has some peculiarities that need to be considered during the design phase of a warehouse. For example, for petroleum wells drilled by a company during a certain time span one can consider how many wells are oil producers. If oil producing, then what have been their temperatures and pressures during the past quarter (or month). For example, a company may be interested in finding out how a newly drilled oil-producing well compares to an existing one in the same basin over the past six month period. In this case, both time and space are of interest to the data warehouse. In other words, they tend to include many dimensions, thus providing a multidimensional view of the data.

2.2 Granularity

A crucial decision to be made while designing a star schema should be the level of detail (termed the grain or granularity), at which measures will be recorded. It is critical that every row in the surveys, wells and permits fact tables be recorded at exactly the same level of detail. Similarly, in order to relate to other fact tables, grain levels followed in them have to be recognized. As the level of success in making effective future business forecasts for exploration and development will depend on the level of knowledge extracted from the historical resources data.

2.3 Star schema for multidimensional data mapping

A star schema created from the resources industry's database is best explained using a figure. The theme of the star schema (Fig. 1) is exploration and production of minerals and petroleum in the Western Australian basins. The four dimensions of the schema are petroleum exploration permit holder; permit status; permit number and permit period. The dimension keys determine the coordinates of the multidimensional structure represented by the star schema.

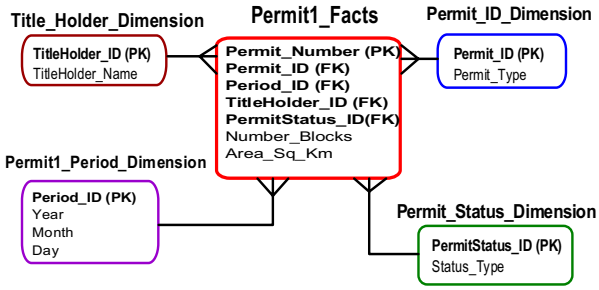

Fig. 1 Star schema for Petro2-permits data mart

The permit period dimension table stores period allowing temporal information in terms of days, months and years. While the fact table, in addition to the dimension keys, stores measures (attributes with metrical or observable data) such as number of blocks and permits area in square kilometers. Further, a petroleum permit holder with a permit number, title and a permit status has specific number of blocks and certain amount of area (ocean bed or land) under its possession for a particular period.

2.4 Handling multiple dimensions using a star schema

The fact table in Fig. 2 describes a four dimensional space. It maps of three of the four dimensions viz. permit number, permit status and period that exist for each permit holder (the fourth dimension) within the star schema. The figure indicates the general case where *i*th permit holder is shown as C_i.

Every row contained in the fact table is represented by exactly one point in a four-dimensional space. The cell A (C_i, 1, 2, 30) has the dimensional key values, with fact values of number of blocks and area in square kilometers.

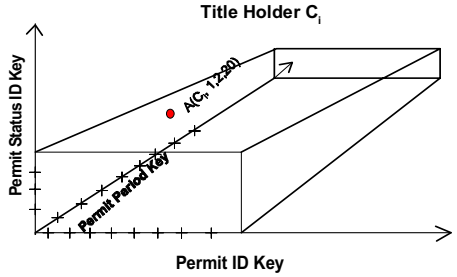

Fig. 2 Dimensional attributes of Petro2-permits fact table

The finest (lowest) level of granularity of the fact table is at the level of a transaction. That is, all petroleum permits having number of blocks and area in square kilometers have been individually recoded in the fact table. In order to increase the granularity, summary of all permits of a single status type by each permit holder for a single day can be recorded. A higher level of granularity can be achieved, if petroleum permits by status for each permit holder is represented on a monthly basis. In this situation, period dimension value is in months rather than in days. To determine the level of granularity, one would need to have a clear understanding of the level of detail required by individuals involved in using the data warehouse. For any analytical problem, a higher or coarser level of granularity will benefit the data warehouse performance and ultimately allow the users to have less number of rows in the fact table. However, it would lack detailed analysis of data.

2.5 Attribute hierarchies in multidimensional analysis

Attributes of a dimension can be ordered in a well-defined attribute hierarchy. This provides top-down data organization that is used for two purposes: aggregation and drill-down or roll-up operations on the data. For example, period dimension can be organized by day, week, month, quarter and year. Another example is that mineral discovery can be organized by state, basin and discovery type. As illustrated in Fig. 3**,** a scenario in which the analyst studies survey facts using the basin, time, and location dimensions, the basin dimension is set to all basins, implying that the analyst will see all the basins on Y-axis. The time dimension is set to month, meaning that the data are aggregated by quarters (for example, total of surveys in basins 1, 2, 3, 4 and 5 during Q1, Q2, Q3, Q4). Finally the location dimension is initially set to state. Thus ensuring each cell containing the total surveys carried out in the state in a given quarter. Data analysis illustrated in Fig. 3 provides the resources analyst with three different information paths. On the basin dimension (Y axis), the analyst can request to see all basins grouped by type, or just one basin. On the item dimension (X axis), the analyst can request time-variant data at different levels of aggregation: year, quarter, month or week.

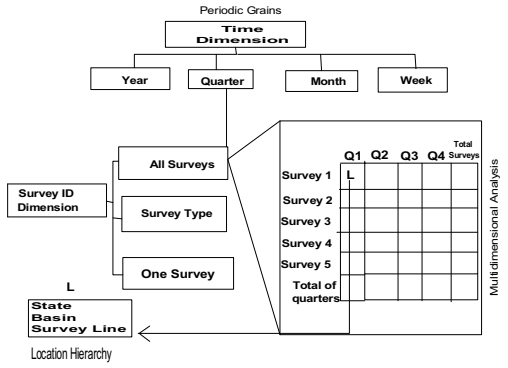

Fig. 3 Granularity of period dimension for Petro2-surveys database

2.6 Using OLAP for exploration permits

OLAP is a query-based methodology that supports data analysis, which can be used to logically structure resources data in the form of a cube as shown in Fig. 4. While a three-dimensional cube is easy to visualize, it is difficult to illustrate more than three dimensions. However, dimensions of more than three can be visualized by thinking of *n*-three-dimensional cubes where *n* is the total number of possible attribute values for the fourth dimension. This process could be repeated to conceptualize higher-level dimensions.

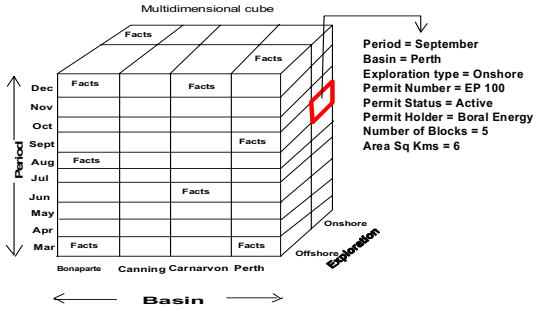

Fig. 4 Data cube from Petro2-permits

Fig. 5 shows another fine-grained data view that has been extracted from the resources data warehouse in the present research. The figure is similar to the previous one except that it stores information regarding oil wells as compared to permits in the previous figure (the legend on the right hand side of the figure indicates the difference). Likewise, a similar cube could be visualized for surveys' data mart.

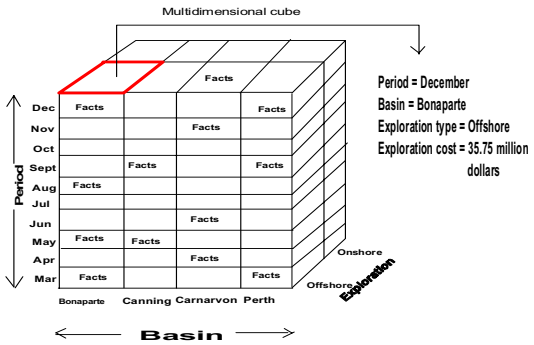

Fig. 5 Data Cube from Petro1 wells

3. Analysis and discussions

A key aspect of the resources companies business is the need to get a total picture of the operations over the years. As and when unexpected trends or variations are observed, they need to drill down the information to get more details in order to discover the reasons for these variations. For example, when the number of surveys or well drilled in a basin drops, or oil (or gas) flow goes down, one must find out if it is a general trend for all exploration entities within the described ones (e.g. for all basins, and for all states). If it is for a specific survey or well or permit, one may want to drill further down and find out if is for a certain type of exploration (such as onshore or offshore) or a 2D seismic or 3D seismic survey or oil or gas producing well. Thus, in order to provide fast answers to such multi-level questions, the warehouse must aggregate and summarize data by survey, basin, state, period of survey or well drilled etc. at different levels of generalization. One such hierarchy could be well-basin-state-country. Another hierarchy could be day-week-number-month-quarter-year (Fig. 6). A parallel hierarchy could be day-month-year.

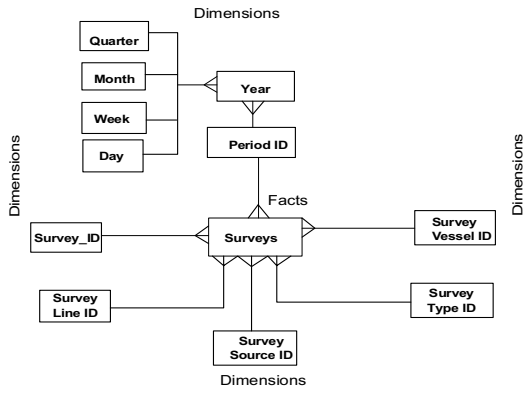

Fig. 6 Granularity of Period dimension (snowflake schema)

3.1 Variations in measurement intervals

The rationale for separating different fact tables such as surveys, wells and permits is that units of measures of interest in each fact table are different. For example, the periods at which surveys are conducted in a basin are different from the dates the wells are drilled in that basin. Seismic line kilometers acquired on a surface has different measure from the drilled wells' depth (represented either in meters or kilometers). Thus, even though, survey, wells and permits have data of similar nature they cannot be lumped into one table.

In order to get some idea what this all entails we list some of the characteristics of the various fact tables in the data warehouse built from the resources data in Table 1. Following are the brief descriptions of the data marts shown in the table:

Mineral Exploration Expenditure - Min1 and Min2
Mineral Exploration Discoveries - Min3 and Min4
Petroleum Exploration and Production - Petro1, Petro2.

Table 1. Number of attributes, rows, dimensions and cube sizes of fact tables in various data marts or databases

Petro1

No. of Attributes	Rows	Dimensions	Cube Size
13	200	2	1.6 KB
19	316	2	2.5 KB
6	66	2	0.5 KB
5	256	2	2 KB

Petro2

No. of Attributes	Rows	Dimensions	Cube Size
13	1736	11	76 KB
25	1473	19	112 KB
7	257	4	4 KB

Min1

No. of Attributes	Rows	Dimensions	Cube Size
8	444	4	3.5 KB

Min2

No. of Attributes	Rows	Dimensions	Cube Size
11	97	1	0.4 KB
5	406	2	3.2 KB
5	464	2	3.7 KB

Min3

No. of Attributes	Rows	Dimensions	Cube Size
6	347	5	3.5 KB

Min4

No. of Attributes	Rows	Dimensions	Cube Size
5	329	4	2.6 KB

The table provides some idea how the number of rows in a fact table and the number of dimensions affect the cube size. Granularity of some of the fact tables of the six resources databases or data marts has been discussed in the following sub-sections.

3.2 Granularity of mineral exploration expenditure (Min1 data mart) facts

A problem encountered in building the exploration expenditure-reporting schema is selecting a grain at transactional-level fact tables. Exploration expenditure (Min1) data mart contains rows of all transactions for each period, state and mineral type. Thus, a very detailed grain of fact table is created.

3.3 Dimensions recorded in Min1 data mart

The basic dimensions typically found for mineral exploration expenditure are:

Period: The effective period of transaction or record is the obvious first dimension, and in order to record individual records grain at the level of day is required for this dimension. Period dimension links the expenditure fact table by one-to-many relationship. Period has a primary key in this dimension and foreign key in the fact table. This dimension has only single attribute, year, since monthly or quarterly facts are not available. At times this dimension has composite attributes (such as day, week, quarter, month, and year) and constructed in the snowflake schema structure, can further be normalized, so that grain size at fact table becomes finer.

State: State, for which the exploration cost computed, is recorded in fact table corresponding to this dimension. It has primary key with its corresponding foreign key in fact table through again one-to-many relationship. The grain size at this dimensional level is coarse, because the fact is state level of the exploration cost at a particular period.

Mineral: The exploration cost for a particular mineral computed, is recorded in the fact table. This dimension has a type of mineral that associates with the exploration cost measure in the fact table.

Other dimensions can be added (such as conversion dimension) depending on the specific business objective and the depth of the exploration costs database system. The dimensions defined in this database have been shown in a three-dimensional view as shown in Fig. 8. The granularity of the exploration expenditure may typically include the contractor for whom this expenditure is incurred. In adding any additional dimensional attributes, the designer must avoid creating details that are not well understood.

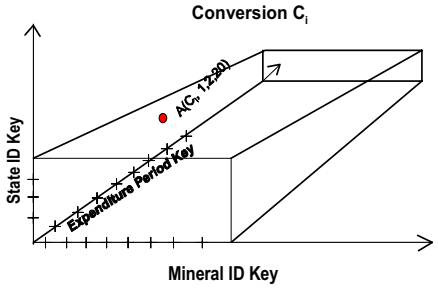

Fig. 7 Multidimensionality of Min1 data mart

3.4 Granularity of mineral exploration expenditure (Min2 data mart) facts

In this database, three fact tables are involved in detailing the grains separately by production lease cost, actual exploration cost, expected exploration cost, meterage drilled on production lease, meterage drilled on other leases. Ratio computed between actual and expected exploration costs minimizes the grain size in the fact table. Exploration costs by state have separate fact entity, as it has state facts and the third exploration cost facts by mineral has mineral type facts. However, all these fact tables involve period dimension.

3.5 Granularity of commercial mineral discoveries (Min3 data mart) facts

Around 350 total mineral discoveries made so far, have been considered for populating in the present database. Mineral, discovery and state are few important dimensions linked to this discovery fact table. The primary keys defined in all these dimensions are incorporated in the fact table in the form of foreign keys, making the grain of the fact data finer.

3.6 Granularity of unaccounted mineral discoveries (Min4 data mart) facts

Several unaccounted (total 329) mineral discoveries have been reported and considered in the database (Min4) construction. Again discovery, mineral, state dimensions surround the Disc_Excluded_Facts fact table. Another dimension added in this database is the reason for excluding these discoveries in the evaluation of the total discoveries. Fig. 10 indicates four dimensions with one-fact table makes the fact data fine grain.

3.7 Multidimensional view of the petro1 data mart

There are four fact tables with five dimension tables, making the grain of the facts tables comparatively coarse. Petro_Production_Facts, Expl_Expend_Facts, Lease_Facts, State_Expl_Cost_Facts have been surrounded by state, period, basin, lease and exploration dimensions. It is

interesting to observe that period dimension linked to all fact tables contains year, quarter and monthly-populated data enabling finer grain in the factual data. Fig 7 shows how dimensions have been connected in one-to-many relationships with fact tables.

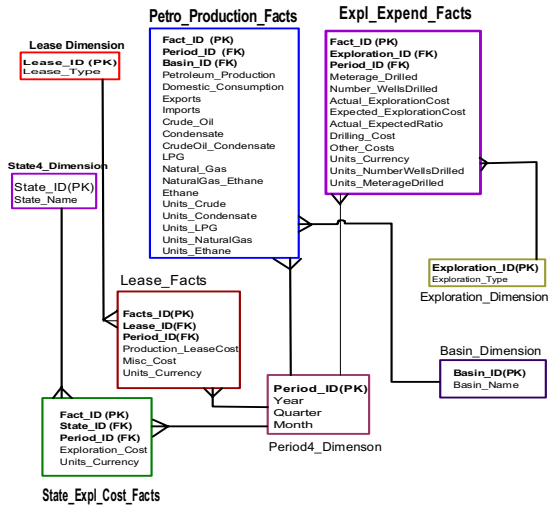

Fig. 7 Star schema of Petro1 data mart

3.8 Multidimensional views of Petro2 data mart

In Petor2 database, a total of thirty dimensions participate in the design and storage of measures in three different fact tables. Fact tables, survey_facts, well_facts and permit_facts, surrounded by dimensions appear to complicate the database structure. Each dimension linked to its respective fact table has one-to-many relationship. As an example, permits database has been represented in one-to-many relationships as shown earlier in Fig. 1.

3.9 Petroleum surveys, wells and permits data mart granularity

Database sizes have substantially been reduced through ER and multidimensional data mapping. These processes have improved the grain size. In all, 30 links have been made from 30 dimension tables to the three fact tables, indicating data with finer grains within fact tables.

3.10 Grains of the resources data fact tables

The most significant design step for a data mart is in determining the lowest level of detailed fact measure recorded in the data marts. The finest grain possible would be each business transaction, such as survey line or well name drilled in a basin in a particular period, or exploration cost incurred in a day, mineral discovery made on that day, petroleum production rate on that day etc. However, the finer the grain of the fact table, more the number of dimensions needed and increasingly more fact rows created. The choice of grain size thus, depends on users involved in the use of data marts and their requirements to ultimately drill-down for detailed data mining solutions and analysis.

3.11 Sizes of fact tables

The grain size of a measure has a direct impact on the size of the fact table. The number of rows in a fact table is estimated as follows:

1. Estimate the number (average) of possible values for each dimension associated with the fact table (i.e. the number of possible values for each foreign key in the fact table)
2. Multiply all the values obtained in step (1) above after making any necessary adjustments.

In this research six data marts representing Australian resources data were been generated for a variety of entities. These data were obtained from the resources industries of various Australian states. Table 2 and following steps show how the above computational method is applied to Petro2 data mart.

Table 2. Estimating the total size of Petro2 data mart

Total number of fact tables: 3
Total number of dimensions: 30

Survey facts:

1. Number of Surveys Conducted: 1736
2. Number of Survey Lines: 1114
3. Number of Survey Sources Used: 18
4. Number of Survey Vessels used: 400
5. Number of Survey Types: 13
6. Number of Survey Companies: 605
7. Number of Survey Permits: 1118
8. Number of Exploration Types: 3
9. Number of Basins: 30
10. Number of Periods: 1736 (in terms of dates)

Well facts:

1. Number of Wells considered: 1473
2. Number of Periods: 1459
3. Number of Rigs: 6
4. Number of Riggers: 144
5. Well Status: 23 (criteria number)
6. Number of Structures: 23
7. Well Completion Status: 10 (criteria number)
8. Well Deviation: 2 (criteria number)
9. Well Side Tracked: 2 (criteria number)
10. Well Classification Number:
11. Well Oil (Gas/Condensate) Producing Rate: 6x3 = 18
12. Well Formations: 41 (number of formations encountered or interpreted)

Permit facts:

1. Number of Permit Periods: 257
2. Permit Status: 4

Common dimensions for the three fact tables are:

1. Number of Permits
2. Exploration Type
3. Number of Basins
4. Number of Companies involved

ER resources data mapping carried out along with multidimensional mapping, thus separating surveys, wells and permits fact tables, has substantially reduced the number of rows in the fact tables and ultimately the size of the database though the grain size remains the same in all these tables.

If all three fact tables are combined, the total number of rows in the Petro2 data mart (or database) would be 1736 x 1473 x 1028 rows. After adjustment, that means only half of the transactions, if considered in the period dimension, the total database size could be in the order of

Total size: 1,778,564,000 x 6 fields x 4 bytes/field = 42,685,536,000 bytes (or around 42.7 gigabytes).

We can thus infer that ER modeling combined with star schema data mapping considerably reduces the database size. The size of a fact table depends on both the number of dimensions participating in the schema and the grain of the fact table. The size of most data warehouses of these resources companies is in terabyte range and growing rapidly as marketing department of the resources company continues to increase number of dimensions and finer grain of the fact tables. The ever demanding supply of information to various operational business units such as exploration, drilling, production, marketing and human resources with finer details is linearly proportional to size of the resources data warehouse.

4. Conclusions and recommendations

This research has investigated the enormous amount of data and information pertaining to the Australian resources industry; and the inherent relationships amongst the various entities involved therein. From the study of the relationships amongst this huge resource the following recommendations are made:

1. Granularity of a fact table is proportional to the number of dimensions involved in the multidimensional database design
2. Fine-grained multidimensional data modeling is best suited to manage, store, and analyze multidimensional resources data.
3. Multidimensional resources data modeling combined relational ERD substantially reduces the size of the resources data warehouse and enables data mining more effective.
4. Decision support data can be interpreted from multidimensional structured data using variety of processing techniques to produce each dimension. The ability to analyze, extract and present information in meaningful ways gains strength from high grained and multidimensional data.

5. Acknowledgement

Authors wish to acknowledge Jeff Haworth, Manager, Data Management Group, Western Australian Department of Industry and Resources, for providing the necessary data used in the present studies.

6. References

[1] Carr, K., Meyer, R., Duma, C., Bryant, K., Hartranft, R., Bergman, R. F., Fox, S. (2003). Storage and Retrieval of Spatially-qualified Data from NASA's EOSDIS Data Pool, *Geoscience and Remote Sensing Symposium, Proceedings of the IEEE International IGRASS '03*, Vol. 1, pp 657 - 659.

[2] Dunham, H. M. (2003). *Data Mining, Introductory and Advanced Topics*, Prentice Hall.

[3] Fayyad, U.M, Shapiro, G.P, Smyth, P and Urthurusamy, R (1996) *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA.

[4] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P. (2003). *Fundamentals of Data Warehouses*, Springer, Heidelberg, Germany.

[5] Liu, F. Shuai, Y., Xin, D., Zhu, Q., Liu, S., Tian, Z. (2003). Data Mining about Hyperspectral Data of Winter Wheat Based on Data Warehouse and Data Organization, *Geoscience and Remote Sensing Symposium, Proceedings of the IEEE International IGRASS '03*, Vol. 6, pp 3679 - 3681.

[6] Li, Q., Duan Y-C., Tang, J., Chen, J-P. (2002). System Development for the Data Warehouse of Oil Exploration, *Journal of the Chengdu Institute of Technology* Vol. 29 (3), pp 310-314.

[7] Mattison, R. (1996). *Data Warehousing – Strategies, Technologies and Techniques*, McGraw-Hill, pp. 425-450.

[8] Miller, J.H and Han, J (2001). Geographic Data Mining and Knowledge Discovery, *GIS Series*, pp.51-72.

[9] Moody, D. L., Kortink, M. A. R. (2003). Bridging the Gap between OLTP and OLAP Design – Part I, *Business Intelligence Journal*, The Data Warehousing Institute, Vol. 8(3), pp.7-24.

[10] Saarvenvirta, G. (2004). The Untapped Value of Geographic Information, *Business Intelligence Journal*, The Data Warehousing Institute,Vol. 9(1), pp.58-63.