



*Journal of Advances in
Science and Technology*

*Vol. IV, No. VII, November-
2012, ISSN 2230-9659*

AN ANALYSIS OF EMERGING TECHNIQUES FOR DATA WAREHOUSING

An Analysis of Emerging Techniques for Data Warehousing

S. K. Humayun

Research Scholar (Computer Science), CMJ University, Shillong, Meghalaya

Abstract— The data warehouse is a decision support system with integrated and cleaned data for the knowledge discovery and data mining systems. Data warehousing is the phenomenon and it have a capacity to store huge amount of electronic data. Data warehousing is acts as the cornerstone of the organization's information infrastructure. Data warehouse have many methods and techniques in order to improve the decision making process. This study discuss about data warehousing and its effective decision making and also about the strategies that reflect changes in the external and internal business environment.

Index Terms— Data warehouse (DW), Database, Applications and Techniques

I. INTRODUCTION TO DATA WAREHOUSING

Data warehouse is the collection of data or information which supports the decision making processes. Data warehousing is the collection of techniques, tools and methods and it is used to support the knowledge workers of directors, senior managers, analysts and managers in order to conduct the data analyses which helps in improving the information resources and performing the decision making processes. Data warehouse is the computer database which is responsible for both collecting and storing the huge amount of information or data. The main goal of data warehouse is to manage and analyze the data effectively. According to Cohen (2006), effective data warehousing will help to create the meaningful relationship between business and information technology by facilitating the enterprise or organization level strategic planning and growth.

Data warehouse is a decision support system which provides the multidimensional view and it has huge amounts of current and historical data from the operational sources and it supplies the useful information to the users when they needed and allows the decision makers to improve their business processes in organizations ((Zikmund and William, 2003). There are many differences in operational databases and data warehouse. Operational database usually covers the short period of time and here most of the transactions will involve the latest data. The data warehouse enable analyses about the historical data that is it covers for last few years.

The following table illustrates the difference between the features of operational database and data warehouse.

Feature	Operational Databases	Data Warehouses
Users	Thousands	Hundreds
Workload	Preset transactions	Specific analysis queries
Access	To hundreds of records, write and read mode	To millions of records, mainly read only mode
Goal	Depends on applications	Decision-making support
Data	Detailed, both numeric and alphanumeric	Summed up, mainly numeric
Data integration	Application-based	Subject-based
Quality	In terms of integrity	In terms of consistency
Time coverage	Current data only	Current and historical data
Updates	Continuous	Periodical
Model	Normalized	De-normalized, multidimensional
Optimization	For OLTP access to a database part	For OLAP access to most of the database

Table: Difference between the Features of Operational Database and Data Warehouse.

Source: Kelly and Sean (1997): Data warehousing in action, Wiley, New York

According to Kimball and Ross (2002), the data warehouse has 4 main components and they are:

- operational systems of record,
- the data staging area,
- the data presentation area and
- data access tools

Each and every component of data warehouse will serve unique functions in preparing the data for examination and manipulation. The following figure illustrates the basic components of data warehouse.

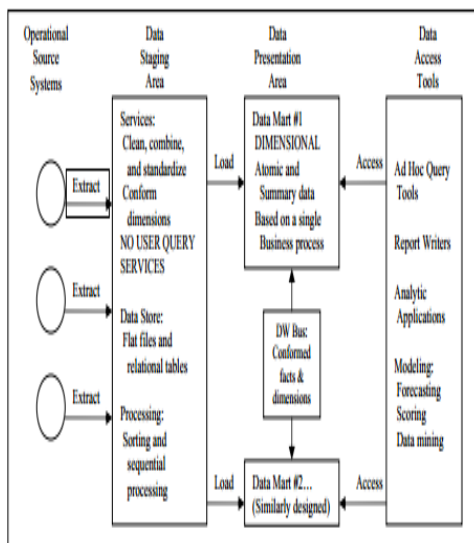


Figure: Basic elements of the Data Warehouse

Source: Kimball, Ross, Margy (2002): The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, John Wiley and Sons, Inc., Chichester, 2002.

1. OPERATIONAL SYSTEMS OF RECORD:

The operational source systems or operational systems of records will capture and process the day to day transactions of the organization. The operational source systems will concentrate effectively on processing performance and they deal with the high volume of transactions. The data or information that acquired in these systems will be uploaded into data staging area.

2. DATA STAGING AREA:

The data staging area acts both as storage area for captured data and platform for set of processes and this process is also called as the ETL (extract, transformation and load). This process occurs in order to standardize the raw information (data) and incorporate that data into the data warehouse. In

extraction process, the data will be extracted from the source systems and then it will be copied into the data staging area. Then, the data will be combined, cleansed and transformed to form a new standard format and structure. In addition to these, all the missing elements, misspellings, duplicate data, incorrect labels, and other errors are corrected and manipulated in this phase. When the data is standardized, it will be loaded into data presentation area, and it is final place where it will be accessible to users.

3. DATA PRESENTATION AREA:

Once the data is formatted, then it is organized, located and available for the user in data presentation area. The data presentation area will be considered as the set of integrated data marts. According to Inmon, (1999), data mart is the subset of data warehouse and it represents data are selected according to specific business function. The data that contained in data presentation area will be detailed and also logically organized.

4. DATA ACCESS TOOLS:

Once the formatted data are updated in the data presentation area, then the users are allowed to utilize the data access tools in order to perform the queries. Some of the data access tools may include sophisticated forecasting tools, data mining applications and ad hoc query tools. Users may use these tools for customizing queries to search the specific data in the data presentation area.

II. APPLICATIONS OF DATA WAREHOUSE

The data warehouse technologies are successfully used in some fields of application:

Trade: Data warehouse technologies is used in trade for shipment and inventory control, sales and claims analyses, public relations and customer care.

Financial services: Data warehouse technologies are used in financial services for credit cards, fraud detection and risk analysis.

Craftsmanship: Data warehouse technologies are used in craftsmanship for supplier and order support and also for production cost control.

Telecommunication services: Data warehouse technologies are used in telecommunication services for customer profile analysis and call flow analysis.

Transport industry: Data warehouse technologies are used in transport industry for vehicle management

Health care service: Data warehouse technologies are used in health care service for bookkeeping in

the accounts departments and also for patient admission and discharge analysis.

III. ONTOLOGY FOR DATA WAREHOUSE DESIGN:

Ontologies will help to designers in development of the data warehouses (Romero and Abelló, 2007; Klein, and Noy 2003). Some of the processes in which the ontologies help for designers are: requirement analysis in the multidimensional design; incompleteness in the multidimensional models; conformed dimensions; reconciling requirements and data sources; data types of measures; semantically-traceable models; security constraint validation; reasoning on OLAP queries; and asserting suitable visualizations.

Here, the ontology integrated data warehouse for multidimensional association mining is discussed as a technique of data warehouse. The concept or idea of ontology has emerged into four effective methods for the domain knowledge sharing and representation (van Elst & Abecker, 2002; Uschold & Gruninger, 1996).

The following figure illustrates the model for ontology integrated data warehouse for the multidimensional association rule mining.

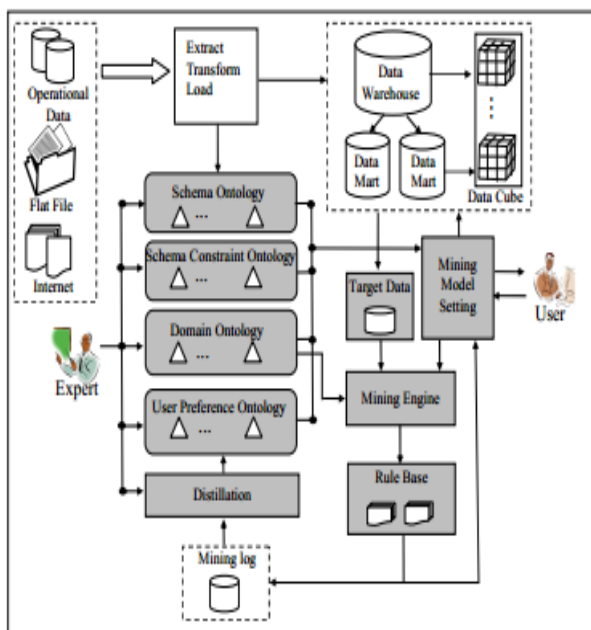


Figure: Ontology-Integrated Data Warehouse for the Multidimensional Association Rule Mining

Source: Wu, C.A., Lin, W.Y., Tseng, M.C., & Wu, C.C. (2007): Ontology-incorporated mining of association rules in data warehouse. *Journal of Internet Technology* 8(4), 477–485.

Schema ontology: It describes about the schema relationship and structure between the attributes of data warehouse;

Schema constraint ontology: It describes about the constraints between the attributes;

Domain ontology: It collects helps to collect domain and expert knowledge; and

User preference ontology: It integrates some derived common mining models.

The mining process begins in setting of mining model by user. Here, the target data is prepared and then mining engine will be launched (Tjioe, & Taniar, 2005). Then it is possible for user to tunes the mining model repetitively until their results are found. This will be saved in mining log and it will be provided for further analysis in order to gather the related model patterns. Here, analyzed results will be utilized to build the user preference ontology.

IV. CONCLUSION

This study concludes that, data warehouses are non-redundant, stable and consistent and also flexible in the terms of data usage. It is concluded that data warehouses will be highly risky and it have poor information quality, long development cycles, complex architectures, and incapability to adapt quickly as the changes that occurs in the business conditions. Although data warehouses have some disadvantages but they provide access and control to users in centralized and formatted data in order to choose the best action and also to support the business decisions.

REFERENCES

1. Kimball, Ross, Margy (2002): *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd Edition, John Wiley and Sons, Inc., Chichester, 2002
2. Inmon, William H. (1999). Data Mart Does Not Equal Data Warehouse. *DM Review*. Retrieved January 2, 2007 from http://www.dmreview.com/article_sub.cfm?articleId=1675
3. Cohen, Rich (2006): *Business Intelligence Strategy: Seven Principles for Enterprise Data Warehouse Design*. *DM Review*, Retrieved December 18, 2006, from http://www.dmreview.com/article_sub.cfm?articleId=1045818.

4. Zikmund, William G, (2003): Business Research Methods, 7th Edition, Dryden Press, New York, 2003.
5. Van Elst, L., & Abecker, A. (2002). Ontologies for information management: balancing formality, stability, and sharing scope. *Expert Systems with Applications* 23(4), 357–366.
6. Uschold, M., & Gruninger, M. (1996): Ontologies: principles, methods and applications, *Knowledge Engineering Review* 11(2), 93–155.
7. Tjioe, H.C., & Taniar, D. (2005): Mining Association Rules in Data Warehouses, *International Journal of Data Warehousing and Mining* 1(3), 28–62.
8. O. Romero, A. Abelló, (2007): Automating multidimensional design from ontologies, in: *DOLAP, 2007*, pp. 1–8.
9. Klein, M. and N. F. Noy (2003): A component-based framework for ontology evolution. *Proceedings of eighteenth International Joint Conference on Artificial Intelligence*.