Significance of Data Warehousing and Data Mining in Business Applications

Madhuri V. Joseph

***Abstract*-Information technology is now required in every aspect of our lives which help business and enterprise to make use of applications like decision support system, query and reporting online analytical processing, predictive analysis and business performance management. In this aspect this paper focuses on the significance and role of Data Warehousing and Data Mining technology in business. A Data Warehouse is a central repository of relational database designed for query and analysis. If helps the business organization to consolidate data from different varying sources. These warehouses are analyzed by the latest technique known as Data Mining. In Data Mining data sets will be explored to yield hidden and unknown predictions which can be used in future for the efficient decision making. Now companies use techniques of Data Mining that involves pattern recognition, mathematical and statistical techniques to search Data Warehouses and help the analyst in recognizing significant trends, facts relationships and anomalies.**

***Index Terms*: Data Warehousing, Data Mining, OLAP, OLTP, CART & CHAID.**

I. INTRODUCTION

Since Data Warehouses are gaining enormous ground in Business Intelligence (BI), every organization gives highest priority to maintain a corporate Data Warehouse. Most business applications like online analytical processing, statistical/predictive analysis, complex query processing and critical business decisions are based on the data available in the Data Warehouse.

Data Warehouse (DW) is a system that extracts, cleans, confirms and source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making. Sophisticated OLAP and Data Mining tools are used to facilitate multinational analysis and complex business models. Inmon W.H defines the Data Warehouse as a subject oriented, integrated, time variant andnon-volatile collection of data in support of management's decision making process [1]. BI applications in enterprises provide reports for the strategic management of business by collaborating the business data and electronic data interchange. This ensures competitive intelligence and thereby helps in good decision making [2]. According to B de Ville, BI refers to the technologies and application for collecting, storing and analyzing business data that helps the enterprise to make better decisions [3].

Data Marts were used to analyze the data and it's a complex task that is time consuming. Thus for the improved analysis if data, Data Mining methodologies is used. The Data Mining process involves computer assisted analysis and extraction of large volume of business data. Frawley, Piatetsky and Mathues defined Data Mining as a nontrivial extraction of implicit, previously unknown and potentially useful information from data [4].

The combination of data warehousing and Data Mining technology has become an innovative idea in many business areas through the automation of routine tasks and simplification of administrative procedures.

II. DATA WAREHOUSE: DEFINITION

Data Warehouse is a repository of enterprise or business databases which provides a clear picture of current and historical operations of organizations [5]. Since it provides a coherent picture of the business conditions at a particular point of time, it is used for the efficient decision making process. it involves the development of system that helps the extraction of data in flexible ways. Data Mining describes the process of designing how the data is stored in order to improve the reporting and analysis. Data Warehouse experts consider that the various stores of data are connected and related to each other conceptually as well as physically. A business's data is usually stored across a number of Databases. However, to be able to analyze the broadest range of data, each of these databases needs to be connected in some way. This means that the data within them need a way of being related to other relevant data and that the physical databases themselves have a connection so their data can be looked at together for reporting purposes..

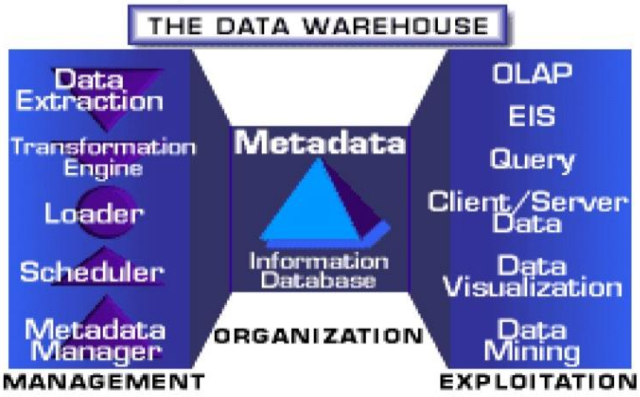

Fig 1: Data Warehouse [6]

Multiple data stores are integrated by the Data Warehouses and this information is used by the mangers for better decision making. Data warehousing environment includes the Extraction of relational database, Transformation, Loading (ETL process), Online Analytical Processing (OLAP) engine and client analysis tools.

Manuscript received March, 2013.

Madhuri V. Joseph, Sr.Lecturer, Dept. of Computing, Muscat College, Muscat, Sultanate of Oman.

As a business grows globally, the parameters and complexities involved in analysis and decision making become more complex. Data access portion which is available in the form of products is the most visible part of a Data Warehouse project. Data warehousing process involves transformation of data from original format to a dimensional data store which consumes a greater percentage of effort, time and expenses. Since e implementation of a data warehousing is expensive and critical, there are a number of data extraction and data cleaning tools and load and fresh utilities are available for the same. One of the most important characteristic of the Data Warehouse is data integration.

A. Example of Data warehousing – Facebook

A great example of data warehousing is what Facebook does. Facebook gathers al your data such as your friends, your likes, your groups etc. All these data are stored into one central repository. Although Facebook is storing all these information into separate databases, they store the most relevant and significant information into one central aggregated database. This is because of many reasons like to make sure that you see the most relevant ads that you are most likely to click on or the friends that they suggest are the most relevant to you.

B. Relevance of Data Warehouse.

Data Warehouse is a subject oriented, time variant, integrated and non-volatilecollection of data. Data cleansing, data integration and Online Analytical Processing (OLAP) are a part of the data warehousing technology. It provides a complete and consistent data store from multiple sources which can be easily understood and used in business applications. Some of the application areas include:

- Integration of data across the enterprise.
- Quick decisions on current & historical data
- Provide ad-hoc information for loosely-defined system
- Manage & control businesses
- Solving what-if analysis.

C. Data Warehousing: Process

Data warehousing is the process of centralizing or aggregating data from multiple sources into one common repository. Data warehousing occurs before Data Mining takes place. Data warehousing involves a strict engineering phase, where no business users are involved. In data warehousing, data stored in different databases are combined into one comprehensive and easily accessible database. This is available to business professionals or managers who use the data for Data Mining and to create forecasts. Data is fed from a variety of disparate sources into the Data Warehouse which is again converted, reformatted, summarized and used for managerial decision making.

The process of data warehousing acts as a guideline to identify the business requirements, develop the business plan and create Data Warehouse also includes project management, startup and wrap-up activities.

D. Data Warehouse: Architecture

Data Warehouse architecture is based on the various business processes associated with an organization. Some other considerations while going for the architecture of a Data Warehouse include data modeling, adequate security, metadata management, extent of query requirement and utilization of full technology. Metadata is data about data which is stored either as a unstructured or semi-structured form. These summary data are very useful in Data Warehouses. For example simple Data Warehouse query can be used to retrieve January sales. Data Warehouse architecture can be shown with the materialized view in Oracle 9i as below.

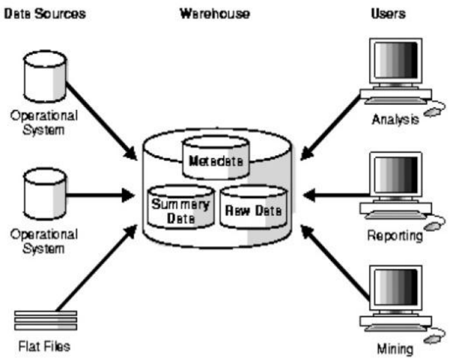

Fig 2: Data Warehouse Architecture[7]

III. FROM DATA WAREHOUSE TO DATA MINING

It is necessary to choose adequate Data Mining algorithms for making Data Warehouse more useful. Data Mining algorithms are used for transforming data into business information and thereby improving decision making process. Data Mining is a set of methods used for data analysis, created with the aim to find out specific dependence, relations and rules related to data and making them out in the new higher level quality information [8]. Data Mining gives results that show the interdependence and relations of data. These dependences are mainly based on various mathematical and statistical relations [9].

Data are collected from internal database and converted into various documents, reports, list etc. which can be further used in decision making processes. After selecting the data for analysis, Data Mining is applied to the appropriate rules of behavior and patterns. That is the reasons why Data Mining is also known as “extraction of knowledge”, “data archeology” or “pattern analysis”.

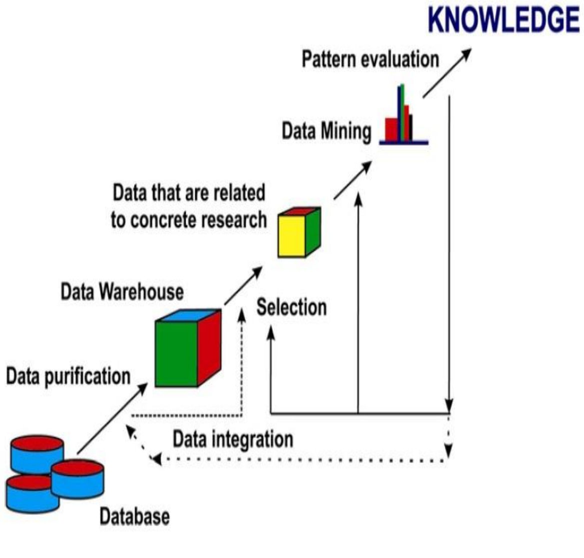

Fig 3: The process of knowledge recovery from database by using Data Warehousing and Data Mining technologies. [10].

A. *Example of Data Mining: Fraud detection of credit card usage*

Credit card companies will alert you when they think your credit card is fraudulently used by someone other than you. Companies will have a history of the customer's purchases and know geographically where the purchases have been made. If a purchase is made in a city far away from where you live, the companies will put an alert to possible fraud since their Data Mining shows that you don't normally make purchases in that city. Companies can either disable the card for that transaction or put a flag for suspicious activity.

B. *Data Mining Process.*

The process of Data Mining provides ways to make best use of data through rapid computerization [11]. Data Mining software uses modeling techniques to make a model that is a set of examples or a mathematical relationship based on data from situations where the answer is known and then applying the same model to other situations where answers are hidden. [12]

The 3 main stages involved in the process of Data Mining are:

1) Exploration: data preparation, cleaning and transformations are involved in this stage. A subset of records will be selected to reduce the number of variables to a manageable range. This depends on the complexity of analysis of graphical and statistical data.
2) Model building and validation: in this stage the best model will be taken based on their predictive performance. Various techniques used for comparison of models include bagging, boosting, stacking and Meta learning.
3) Dependent: in this final stage the best model is selected and it is applied to the new data sets to generate predictions of the expected outcome. One simple example for this is the online shopping site doing e-commerce transactions through credit card deploys neural networks and Meta learner to identify fraud.

Data Mining process involves use of various techniques and methods. Most common techniques are:

1) Classification: stored data will be grouped into different classes. This allows locating data into pre-determined groups.
2) Clustering: Data items are grouped into clusters of similar groups. It may be of It may be of hierarchical or non-hierarchical.
3) Regression: this method uses a numerical data set to develop a best fit mathematical formula. This formula can be used to feed new data sets and get a better prediction. This is suitable for continuous quantitative data.
4) Association: it is a rule X->Y such that X and Y are data items sets.
5) Sequential pattern matching: it allows predicting behavior patterns and trends based on the sequential rule A->B which implies that event B will always be followed by A.

C. *Next Generation Data Mining Techniques*

Data Mining uses black box approach to explore data and discovered knowledge using Exploratory Data Analysis (EDA) techniques. The techniques used in Data Mining are a blend of statistics, database research and artificial intelligence [13]. Next generation Data Mining techniques include artificial neural networks, decision trees, induction rules and genetic algorithms.

1) Artificial neural networks: This technique uses non-linear predictive models to enable learning through training. Computers are trained to think, act and take decision similar to humans. These models are quite complex to use even by the experts because it is packed as a complete solution[14]. It determines relevant prediction for a model.
2) Rule induction: This technique enables knowledge discovery and unsupervised learning. It extracts useful patterns from database based on accuracy and statistical significance. Prediction will be more correct and has better logic by neural network. It creates a certain confusions to select the best rule from a pool of rules. Normally rule induction is used on databases with many columns of binary fields or fields with higher cardinality in order to collect the suitable patterns for making a better prediction, a bottom – to – top approach is chosen.
3) Decision trees: Decision tree is a Data Mining technique where tree shapes structures are representing the set if decision generating rules for a data set classifications. The starting node or the top node is known as the root. Depending results of test, the root partitioned into two or more nodes. It is a fast Data Mining technique since its required less or no pre-processing of business data. It is used for both exploration and prediction using Classification And Regression Trees (CART) and Chi Square Automatic Interactions detection (CHAID). CART generates two way splits from data set segmentation which needs less preparation of data than CHAID which generates a multi-way split. Rules are mutually exclusive and relatively exhaustive.
4) Genetic algorithms: This optimized technique of Data Mining is based on genetics and natural selections, combination and mutation [15]. Genetic algorithms are used in patterns recognition either as classifier or as an optimization tool. According to Chuck Kelly (2002), genetic algorithms support the survival of the fittest using heuristic functions even by posing the problems [16].

IV. INFRASTRUCTURE FOR IMPLEMENTING DATA WAREHOUSE AND DATA MINING

Data Warehouse and Data Mining application are quite divorced in size and storage capacities. Enterprise applications range from 10 GBs to higher. Data Warehouse is a very flexible solution that can explore database more efficiently than any other Online Transaction Processing (OLTP) environment. The major advantage of this is that the user does not have to possess knowledge of relational model and complex query languages.

A. *Data ware house implementation phases.*

According to Barry D & Addison – Wesley, 1997 Data Warehouse implementation phases include [17]:

1) Analysis of current situations: this is a very important phase in the Data Warehouse design, since at this phase a possibility of realization and solution of the problems can be seen. Since the users will have a better knowledge about the problems than the designers, their opinion is very crucial for a good warehouse design.
2) Selecting the most appropriate data for analysis from the existing data: instead of using the entire OLTP database,

the data subset which includes all the interesting data related to the subject will be chosen.

3) Filtering data interesting for analysis: data analysis does not need all the data. Because of this the filtering of data will be done related to certain time period or some specific area.
4) Extracting data in staging database: after reducing and filtering of data, data are being extracted in staging database from which the Data Warehouse is being built. Data Transformation Services (DTS) package is written in SQL server 2000. Package writing is very significant in Data Warehouse implementation because packages can be arranged to function automatically so that the users can fresh and prompted data.
5) Selecting fact table, dimensional tables and appropriate schemas: entity-relationship model commonly used in the design of relational databases. This is suitable for OLTP. A Data Warehouse requires a concise, subject oriented schema that facilitates online data analysis. The simplest scheme is a single table scheme which consists of redundant fact table. Data Warehouse contains a large central fact table containing the bulk of data with no redundancy and a set of smaller dimension tables.
6) Selecting measurement, percentage of aggregations and warehouse models: the next step in designing Data Warehouse is selecting measurements. It needs calculated measurements that are attained from
7) Various arithmetic operations with other measurements. Data Warehouse solutions also use aggregations through which they solve the queries very fast. The increasing of the percentage of aggregated data speeds up the user defined queries.
8) Creating and using the cube: the cube is being created on either client or server computer. Basic factors for selecting the place for cube's storehouse are size of the cube, performance of the client's and server's computers, number of the cube users and throughput of the system. The cube created can be used by the support of various client tools.

B. Data Mining Implementations

Microsoft Decision Tree (MDT) algorithms are based on possibility of various attributes and it is when prediction is necessary. These algorithms also generate rules. MDT also enables the user to analyze a large number of Data Mining problems.

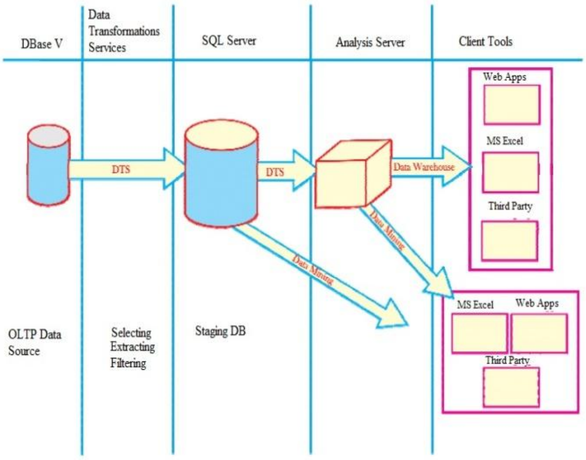

Fig 4: Scheme of Data Warehouse and Data Mining implementation [18].

Database size and query complexity are the 2 critical technological drivers for Data Mining. New hardware architectures like Massively Parallel Processors (MPP) are used which can link hundreds of speed processors to achieve better performance. Data Mining is now aggressively used in various industries [19]. All the major database vendors are using various Data Mining techniques in their platforms. Some of them are:

1) SQL server: it is Microsoft database platform that allows Data Mining through the use of clustering and classification algorithms.
2) SAS, SPSS and S-PLUS are advanced statistical packages for implementing Data Mining algorithms.
3) Darwin: is an Oracle Data Mining suite for implementing classification and decision trees, K-nearest neighbors, regression analysis, clustering algorithms and neural networks.

V. DATA WAREHOUSE AND DATA MINING: APPLICATION AREAS IN BUSINESS

Data warehousing and Data Mining has gained improved popularity in multiple areas of business to analyze the large databases quickly which would be too complex and time consuming. Some of these application areas are listed below.

1) Government: for searching terrorist profile and threat assessments.
2) Finance: analysis and forecasting of business performance, for stock and bond analysis.
3) Banking: to learn underwriting, mortgage approval etc.
4) Direct marketing: for identifying prospects that are included in mailing list so as to obtain highest response time.
5) Medicine: for drug analysis, diagnosis, quality control and epidemiological studies.
6) Manufacturing: for improved quality control and maintenance.
7) Churn analysis: to predict customers who are likely to quit the company and move to a competitor company.
8) Market segmentation: to identify customer's common characteristics and behavior that purchases the same products of a company [20].
9) Trend analysis: to analyze the difference between the customer's behavior over consecutive months.
10) Fraud detection: to identify the fraud users in telecommunication industry as well as credit card usage.
11) Web marketing: for advertisements and personalization opportunities.

VI. CONCLUSION

Data Warehouse and Data Mining technologies have bright future in business applications as it helps to generate new possibilities by automated prediction of trends and behaviors in a large database. Data Mining techniques help to automatically discover the unknown patterns like identifying anomalous data that highlight errors generated during the data entry. Data Warehouse & Data Mining technologies have become a hit with various industries like sales & marketing, healthcare organization, financial institutions and many more. These technologies have a lot of benefits in varying fields. It can be said with pleasure that these technologies help the quick analysis of data and thereby improving the quality of decision making process. Both Data Mining and Data Warehousing are business intelligence tools that are used to turn information or data into actionable knowledge. Data Warehouse experts design data storage system that connects relevant data in different databases where as a Data miner run more meaningful and efficient queries to improve business. The immense data volumes and extremely complex knowledge discovery procedures associated with business organizations make the Data Warehouse with its OLAP and Data Mining tools a very significant technology supporting decision making. Thus Data Warehouse & Data Mining are very essential components in business operations to gain competitive intelligence. These technologies allow statistical multidimensional analysis of data to evaluate relationships, correlations and trends in business.

REFERENCES

[1] Inmon W.H., “*Building the Data Warehouse”*, Second Edition, JWiley and Sons, New York, 1996.

[2] P. Bergeron, C. A. Hiller, (2002), “Competitive intelligence”, in B. Cronin, Annual Review of Information Science and Technology, zedford, N.J.: Information Today, vol. 36, chapter 8

[3] B. de Ville, (2001), *“Microsoft Data Mining: Integrated Business Intelligence for e-Commerce and Knowledge Management”,* Boston: Digital press.

[4] Frawley W., Piatetsky – Shapiro G. and Matheus C., “Knowledge Discovery in Databases: An Overview”, Al Magazine, Fall 1992, pgs 213-228.

[5] C. Date, (2003), *“Introduction to Database Systems”,* 8th ed., Upper Saddle River, N.J.: Pearson Addison Wesley.

[6] Han Jiawei, Kamber Micheline, “*Data Mining: Concepts and Techniques”*, 2nd edition, Morgan Kaufman Publishers, March 2006. ISBN 1-55860-901-6.

[7] Oracle9i Data Warehousing Guide Release 2 (9.2), Part No. A96520-01, March 2002.

[8] Berry, M.J.A., and Linoff, G., *"Mastering data mining*", The Art and Science of Customer Relationship Management, 1999.

[9] Bhavani, T., *Data Mining: Technologies, Techniques, Tools and Trends,* 1999.

[10] Jiwei, H., and Micheline, K., *Data Mining: Concepts and Techniques*, Simon Fraser.

[11] D. Pyle, (2003), *“Business Modeling and Data Mining”,* Morgan Kaufmann, San Francisco, CA.

[12] M.H. Dunham, (2005), “Data *Mining – Introductory and Advanced Topics”*, Prentice Hall.

[13] Berson Alex, Smith J. Stephen, Thearling Kurt, (1999), *“Building Data Mining Applications for CRM”*, McGraw-Hill Companies.

[14] Gilman Michael, (2004), “*Nuggets and Data Mining*”, Data Mining Technologies Inc. Melville, NY 11714, (631) 692-4400

[15] Chen, S. H, (2002), “*Genetic Algorithms and Genetic Programming in Computational Finance*”, Boston, A: Kluwer.

[16] Kelly Chuck, (2002), “*What is the role of Genetic Algorithms in Data Mining*”, Information Management: How your Business Works,Electronic Newsletter, http://www.information-management.com/news/5755-1.html.

[17] Barry, D., *Data Warehouse from Architecture to Implementation*, Addison-Wesley, 1997.

[18] Krulj, D., "*Design and implementation of data warehouse systems*", M Sc. Thesis, Faculty ofOrganizational Sciences, Belgrade, 2003.

[19] Chapple Mike, “*Data Mining: An Introduction*”, (2011), http:/ /databases.about. com/od/datamining/a/datamining.htm.

[20] Alexander Doug, “*Data Mining*”, (2000), http://www.laits.utexas.edu/norman/BUS.FOR /course.mat/Alex/, electronic article.