Data Warehousing and Data Mining in Business Applications

Eesha Goel

CSE Deptt. GZS-PTU Campus, Bathinda.

Abstract—*Information technology is now required in all aspect of our lives that helps in business and enterprise for the usage of applications like decision support system, query and reporting online analytical processing, predictive analysis and business performance management. This paper focuses on the significance and role of Data Warehousing and Data Mining technology in business. Data Warehouse is a central repository of relational database designed for query and analysis. Business organization uses this technique to consolidate data from different varying sources. The latest technique used for analyzing these warehouses is known as Data Mining. In Data Mining data sets will be explored to yield hidden and unknown predictions that can be used in future for the efficient decision making. In global data now companies use techniques of Data Mining that includes pattern recognition, mathematical and statistical techniques for searching Data Warehouses and to help the analyst in recognizing significant trend, facts relationships and anomalies.*

Keywords: - *Data Warehousing, Data Mining, OLAP, OLTP.*

I. INTRODUCTION

Data Warehouse are emotionally gaining in Business Intelligence (BI), so every organization gives highest priority to maintain a corporate Data Warehouse. Popular business applications like online analytical processing, statistical/ predictive analysis, complex query processing and critical business decisions all are based on the data i.e. 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 for decision making by querying and analyzing. Sophisticated OLAP and Data Mining tools are used to facilitate multinational analysis and complex business models. BI applications in enterprises provide reports for the strategic management of business by collaborating the business data and electronic data interchange in order to ensure competitive intelligence and further helps in good decision making.

To analyze the data Data Marts are used which is a complex task. Due to its complexity it is a time consuming process. In order to improve the analysis of data, Data Mining methodologies are used. The Data Mining process includes computer assisted analysis and extraction of large volume of business data. The combination of data warehousing and Data Mining technology has become a creative idea in many business areas through the automation of routine tasks and simplication of administrative procedures.

II. DATAWAREHOUSE DEFINITION:

Data Warehouse is a repository of business databases that provides a clear view of current and historical operations of organizations. As it provides a clear picture of the business conditions at a particular point of time, it is used in making efficient decision, which includes the development of system that further helps in the extraction of data flexibly. Data Mining describes the process for designing how should the data be stored in order to improve the reporting and analysis. Data Warehouse experts consider that the various stores of data are converted and related to each other conceptually as well as physically. Data Warehousing environment includes the Extraction of relational database, Transformation, Loading (ETL process), online analytical processing (OLAP) engine and client analysis tools. Since business is growing globally, the parameters and complexities that are involved in analysis and decision making become more complex. The most visible part of a Data Warehouse project is the Data Access portion which is available in the form of products. Data Warehousing process involves the transformation of data from original format to a dimensional data store that consumes a greater percentage of effort, time and expenses. Data Integration is one of the most important characteristic of the Data Warehouse. The features of Data Warehouse are described in figure given below:

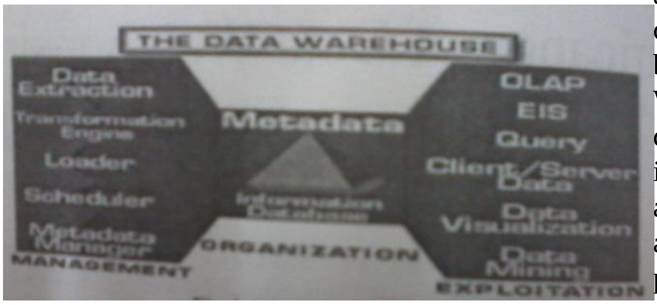

Figure1: Data Warehouse

A. Example of Data Warehousing

A good example of data warehousing is the management of data by Facebook. Facebook gathers all your data such as your friends, your likes, your groups, etc and stored into one central repository. Facebook stores all the information into separate databases but the most relevant and significant information is being stored into one central aggregated database. The reason for using this approach is to make sure that you see the most relevant adds 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-volatile collection of data. The various part of the data warehousing technology are Data Cleaning, data integration and Online Analytical Processing (OLAP). This helps in providing a complete and consistent data store from multiple sources that can be easily understood and used in business applications. Some of the application areas include:

1. Integration of data across the enterprise.
2. Quick decisions on current and historical data.
3. Provide ad-hoc information for loosely defined system.
4. Manage and control business.
5. Solving what-if analysis.

C. Data Warehousing: Process

Data Warehousing is the process of aggregating data from multiple sources into one common repository. This process occurs before the Data Mining process. In Data Warehousing, data which is stored in different databases are combined into one central phase in order to access the database. This is further available to managers to use for Data Mining and for creating forecasts. Data warehousing process is used for conversion, reformatting, summarization and then for managerial decision making.

D. Data Warehouse: Architecture

Data Warehouse architecture is based on various business processes associated with an organization as shown in figure 2 below. The architecture of Data Warehouse includes data modeling, adequate security, meta data management, extent of query requirement and utilization of full technology. Meta data is data about data that can be stored either as a unstructured or in semi- structured form that are useful in Data Warehouses. For example simple Data Warehouse query can be used to retrieve January sales.

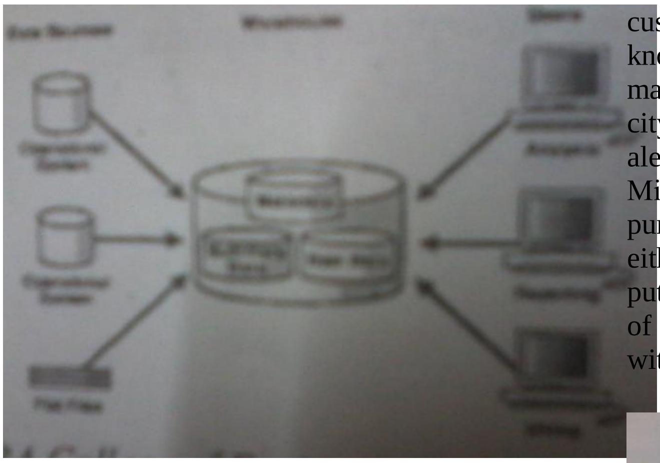

Figure2: Data Warehouse Architecture

III. FROM DATA WAREHOUSE TO DATA MINING

Choosing of adequate Data Mining algorithms is necessary for making Data Warehouse more useful. Data mining algorithms are used for transforming data into business information and further helps in improving the decision making process. Data Mining is a set of methods that are 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. Data Mining provides results that represents the interdependence and relations of data which are based on various mathematical and statistical relations.

Data are collected from internal database which is being converted into various documents etc that can be used in decision making. After the selection of data for analysis, Data Mining is applied to the appropriate rules of behavior and patterns. This is the reason why Data Mining is also known as "extraction of knowledge" or "pattern analysis".

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

When companies think that your credit card is being used fraudulently by some another person then Credit Card companies will alert you. These companies will have history of the customers purchases and they graphically know about where the purchases have been made. If a purchase is made far away from the city you alive, then the companies will put an alert as a possible fraud since their Data Mining shows that you don't normally make purchases in that city. For this companies can either disable the card for that transaction or put a flag for suspicious activity. The process of knowledge recovery has been described with figure given below:

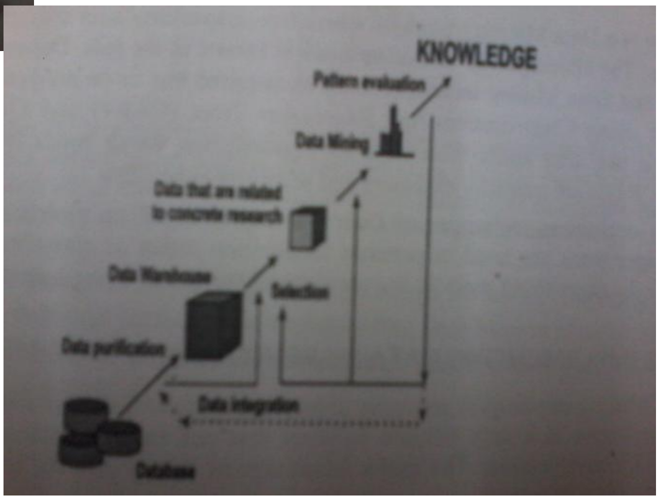

Figure3: The process of knowledge recovery from database by using Data Warehousing and Data Mining technologies.

B. Data Mining Process

With the use of rapid computerization data mining process provides ways for best usage of data. Data Mining software uses modeling techniques for making a model which is composed of set of examples or a mathematical relationship based on data from situations where the answer is known and then applying some model to other situations where answers are hidden.

The main stages included in the process of Data Mining are:-

1. Exploration: This stage involves data preparation, cleaning and transformations. A subset of records will be selected to reduce the number of variables to manageable stages that depends on the complexity of analysis of graphical and statistical data.

2. Model Building and Validation: Based on the predictive performance of models, best model will be taken. For comparison of models the techniques that can be used are bagging, boosting, stacking and Meta learning.
3. Dependent: This is the final stage where the best model is being selected and applied to the new data sets to generate predictions of the expected outcome.

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

C. Next Generation Data Mining technique

Data Mining uses black box method for exploring data and discovered knowledge through Exploratory Data Analysis (EDA) techniques. Next Generation Data Mining techniques includes artificial neural network, decision trees, induction rules and genetic algorithms.

IV. INFRASTRUCTURE FOR IMPLEMENTING DATA WAREHOUSE AND DATA MINING

A. Data Warehouse implementation phase

Data Warehouse implementation phase includes:
1. Analysis of current situations.
2. Filtering data interesting for analysis.
3. Extracting data in staging database.
4. Selecting fact table, dimensional tables and appropriate schemas.
5. Selecting measurement, percentage of aggregations and warehouse models.
6. Various arithmetic operations with other measurements.
7. Creating and using the cube.

B. Data Mining Implementations

When prediction is necessary Microsoft Decision Tree (MDT) algorithms are used that are based on possibility of various attributes. These algorithms are used to generate rules and help the user to analyze the large number of Data Mining problems. The implementation has been described using figure shown below:

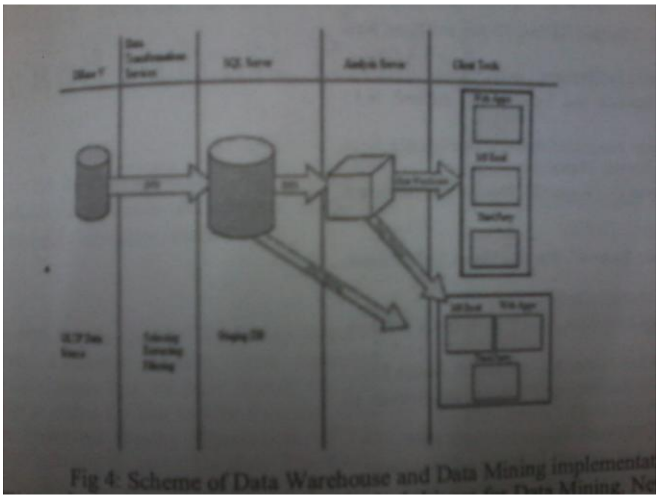

Figure4: Scheme of Data Warehouse and Data Mining implementation

The two critical technological drives for Data Mining are Database Size and Query Complexity.

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

Data Warehousing and Data Mining has gained improved popularity in multiple areas of business. Some of them are listed below:
1. Government: for searching terrorist profile and threat assessments.

2. Banking: analysis and forecasting of business performance for stock and bond analysis.
3. Direct marketing: for identifying prospects that are included in mailing list so as to obtain highest response time.
4. Medicine: for drug analysis, diagnosis, quality control and epidemiological studies.
5. Manufacturing: for improved quality control and maintenance.
6. Fraud detection: to identify the fraud users in telecommunication industry as well as credit and usage

VI. CONCLUSION

Data Warehouse and Data Mining technologies have bright future in business applications as it helps in generating 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 anomalies data that highlight errors generated during the data entry. Both techniques have become a hit in various industries. Both of them are Business Intelligence tools that are used to turn information and data into actionable language. Data Warehouse design storage system that connects relevant data in different data bases where as Data Minor run more meaningful and efficient queries to improve business.

REFERENCES

[1] Inmon W.H., " Building the Data Warehouse", Second Edition, J Wiley and Sons, New York, 1996.

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

[3] C.Date, (2003), " Introduction to Database Systems", * Edition

[4] Barry, D., Data Warehouse from Architecture to Implementation, Addison-Wisley,1997.

[5] Figures from Google Search Engine (Images).