A Review on Data Warehouse and Data Mining Technique

Chiranjiw Kumar Sahu[1], Rahul Kumar Chawda[2]
[1]*Student of MCA,* [2]*Assistant Professor*
Department of Computer Science, Kalinga University, Raipur.

Abstract **-** In this era of globalization and tough competition, every enterprise has an indispensable role to play proficiently and productively in order to maintain its existence in the market and increase its profitability ratios. This challenge becomes more complicated with advancement in information technology along with increasing volume and complexity of information. Presently, success of an enterprise is not just the outcome of efforts or utilization of resources but also depends upon its ability to mine the knowledge from the stored information. Emerging technologies and automation of systems, now allow collection and storage of large amount of data in a well-organized and economical way within a central repository called 'data warehouse'. Data warehousing is a collection of decision making techniques that has emerged to integrate and manage the large multivariate data effectively and systematically. Thus, the problem arises here is not a collection and storage of data but the way data is extracted to enhance the knowledge. Data mining or knowledge discovery shores up organizations, scrutinize their data more effectively and proficiently to attain valuable information, that can award a competitive edge through intelligent and strategic decision making. Data mining comprises several techniques and mathematical algorithms that can be used to mine the value aided information to augment the business performance and decision-making even in the case of incomplete data.

Keyword **-** Data Warehouse, Data Mining Technique

I. INTRODUCTION

Presently, the technological revolution in data capture, processing power, data transmission and storage capabilities, are available to organizations to assimilate their databases into a central repository so that data can be managed effectively and systematically. Before the advent of data warehouses, operational databases were used to satisfy their functional requirements, like data processing, analysis and reporting, however informational needs were the secondary considerations. But with the advent of information technology and increased complexity of data, business houses started demanding an information tool to improve their decision making capabilities. Now the data becomes heterogeneous (mixture of text, symbolic, numeric, texture, image), huge (both in dimension and size), scattered and growing at a phenomenal rate. As a result, traditional ad-hoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of data. Data warehousing has evolved to meet these challenges without disrupting operational processing. Warehouses optimize database query and reporting tools because of their ability to analyze data, often from disparate databases and in remarkable ways. Data warehousing technology helps the managers and decision makers to extract information quickly and easily to retrieve the patterns in data, hidden but useful facts and relationships between the data items. Data mining is a powerful technique for the knowledge discovery and extraction of predictive information from data warehouse to help an enterprise to catch and focus on the vital information while making decisions. There are huge varieties of data mining methods and algorithms for information extraction and prediction. These different technological aspects are discussed in the following sections.

A. Data Warehouse and Data Warehousing - Data warehouse is a repository of an organization's electronically stored data. A data ware-house is the consistent store of data which is made available to end users, so that they can understand and use in a business context. It is not just an individual repository product, rather an overall strategy or process for building decision support systems and a knowledge-based architecture & environment that supports everyday tactical decision making as well as long-term business strategy. Data warehouse is used in the businesses to convert data into business intelligence and making management decisions, based on the facts and not on intuition. The data warehouse environment positions a business to utilize an enterprise-wide data store to link information from diverse sources and make the information accessible for a variety of user purposes, most notably data warehouses are designed to facilitate reporting and strategic analysis. Business analysts must be able to use the warehouse for such strategic purposes as trend identification, forecasting, competitive analysis, and targeted market research. Data warehouse is one of the steps on the long road towards the ultimate goal of accomplishing the objectives of a concern.

B. Applications of Data Warehouse - The fundamental reason for building a data warehouse is to improve the quality of information in the organization. Some of the applications of data warehousing are as given below.

- Credit cards chum analysis
- Insurance fraud analysis
- Call record analysis
- Logistics management
- Data warehousing technologies have been successfully deployed in many industries
- [Chaudhuri and Dayal, 1997] such as:
- Manufacturing (for order shipment and customer support)
- Retail (for user profiling and inventory management)

- Financial services (for claims analysis, risk analysis, credit card analysis, and
- fraud detection)
- Transportation (for fleet management)
- Telecommunications (for call analysis and fraud detection)
- Utilities (for power usage analysis)
- Healthcare (for outcomes analysis)

C. Data Mining - Data Mining may be viewed as automated search procedures for discovering credible and actionable insights from large volumes of high dimensional data. Data mining or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems . Data mining tools can answer business questions that traditionally were too time-consuming to resolve. They scour databases for the hidden patterns, finding predictive information that experts can miss because it lies outside their expectations. Data mining techniques can be applied on a wide variety of data types including databases, text, spatial data, temporal data, images, and other complex data .Data mining is a technology to enable data exploration, data analysis as well as data visualization of large databases at a high level of abstraction, without a specific hypothesis in mind. It can be defined as the non-trivial extraction of novel, implicit, and actionable knowledge from large datasets that includes

- Extremely large datasets.
- Discovery of the non-obvious i.e. hidden and unknown facts.
- Useful knowledge that can improve processes.
- Cannot be done manually.

Data Mining employs techniques from statistics, pattern recognition, and machine learning, high performance computers, parallel algorithms, visualizations, database, etc. Many of these methods are also frequently used in vision, speech recognition, image processing, handwriting recognition, and natural language understanding. However, the issues of scalability and automated business intelligence solutions differentiate data mining from the other applications of machine learning and statistical modeling

II. LITERATURE SURVEY

Inmon, W.H., and Wiley, John (1992), affirmed that data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions. The past three years have seen explosive growth, both in the number of products and services offered and in the adoption of these technologies by industry.

Dorsey, Paul (1996) from site http:// www.dulcian.com/ papers/IOUG/1996/Logical _Design_Data_Warehouse.htm, illustrated: A data warehouse is a term that is used to mean many different things. The most natural definition for a data warehouse is a dedicated machine running its own DBMS with its own database, usually (but not exclusively) for use with an EIS. The central data warehouse is just that, a warehouse. All the enterprise's data are stored in that, "normalized", in order to minimize redundancy and so that each may be found easily. This is accomplished by organizing it according to the enterprise's corporate data model.

Jarke, M., and Vassiliou, Y. (1997), in "Data Warehouse Quality: A Review of the D WQ Project", stated that a data warehouse (DW) is a collection of technologies aimed at enabling the knowledge worker to make better and faster decisions. It is expected to present the right information in the right place at the right time with the right cost in order to support the right decision.

Surajit Chaudhuri Umeshwar Dayal (1997), in "An Overview of Data Warehousing and OLAP Technology" affirmed that data warehousing is a collection of decision support technologies, aimed at enabling the knowledge workers (Executive, manager, and analyst), to make better and faster decisions. Data warehousing technologies have been successfully deployed in many industries like manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). A data warehouse is a "subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making. Typically, the data warehouse is maintained separately from the organization's operational databases.

From Wikipedia, the free encyclopedia, under the title "Data warehouse" and from site: http://en.wikipedia.org/wiki/Data_warehouse stated "A data warehouse is a repository of an organization's electronically stored data. Data warehouses are designed to facilitate reporting and analysis. This definition of the data warehouse focuses on data storage. However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an expanded definition for data warehousing includes business intelligence tools, tools to extract, transform, and load data into the repository, and tools to manage and retrieve metadata". Further Chaudhuri and Dayal (1997), acknowledged that data warehousing technologies have

been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). A data warehouse is a "subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making."

Frawley, Piatetsky-Shapiro, & Matheus (1991), alleged that data mining techniques can be performed on a wide variety of data types including databases, text, spatial data, temporal data, images, and other complex data.

Palace, Bill (1996), in the article named "Data Mining: What is Data Mining?" from site: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/ data mining.htm; Palace enlightened that dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. From site: http://databases.about.eom/cs/datamining/g/dmining.html.

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items.

Howard Hamilton, Ergun Gurak, and Leah Findlate, and Wayne Olive, in article titled "Computer Science 831: Knowledge Discovery in Databases", retrieved from: http://www2.cs.uregina.ca/~hamilton/courses/831/index.html, had explained that the term Knowledge Discovery in Databases or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

III. CONCLUSION

The data warehouse, data mining and their methods, are study elaborated different definitions, methods articulated so far. It is concluded that data warehousing is a subject oriented, integrated, time varying, nonvolatile collection of data and not just a central repository. It is, rather, a strategy or a technique for building knowledge based decision support system to provide business intelligence by its proficient implementation and maintenance.

IV. REFERENCES

[1]. Dorsey, P. (1996). Logical Design of a Data Warehouse to Support Reporting, Ad Hoc Query, Executive Information Systems, and Decision Support Systems. Dulcian, Inc. Retrieved from: http://www.duIcian.com/papers/IOUG/1996/Logical_Design_ Data_ Warehouse.html.

[2]. Inmon, W.H., & Wiley, J. (1992). Building the Data Warehouse. Retrieved from site: http://www.google. co. in/url.

[3]. Jarke, M., & Vassiliou, Y. (1997). Data Warehouse Quality: A Review of the DWQ Project, DWQ : ESPRIT Long Term Research Project, No 22469 Invited Paper, Proc. 2nd Conference on Information Quality. Massachusetts Institute of Technology, Cambridge, 1997.

[4]. Palace, B. (1996). Data Mining. Spring 1996, Retrieved from: http://www.anderson.uda. edu/ faculty/ jason. frand/teacher/technologies/palace/datamining.html

[5]. Hamilton, H., Gurak, E., Findlater, L., & Wayne O. (2003, July 4), Computer Science 831: Knowledge Discovery in Databases, Copyright © 2000-2. Retrieved from: http://www2.cs.uregina.ca/~hamilton/courses/831/index.html.

[6]. Chaudhuri, S., & Dayal, U. (1997, March). An Overview of Data Warehousing and OLAP Technology, ACM SIGMOD, 26 (1), 65 - 74, Retrived from: http://research.microsoft.com/pubs/76058/sigrecord.pdf, http://portal.acm.org/ citation, cfm? id = 248616.

[7]. Fayyad, U.M., Piatetsky-Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 Panel - Data mining: The next 10 years. SIGKDD Explorations, 5(2). Retrieved March 22,2006 from ACM Digital Library database.

[8]. Yeha, I. C., & Lienb, C. (2007). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Department of Information Management, Chung-Hua University, Hsin Chu 30067, Taiwan, ROC. Department of Management, Thompson Rivers University, Kamloops, BC, Canada doi:10.1016/j.eswa.2007.12.020. Copyright © 2007 Elsevier Ltd.

[9]. Yen, J., & Langari, R. (1999). Fuzzy Logic-Intelligence, Control, and Information. Pearson Education. New jersey: Prentice Hell. 125-170.

[10]. Zhang, Y., & Xu, G. (2009). Web Community Mining and Analysis. Retrieved from: http:// www.culturegrid.net/ SKG2007/ keynote/ web CommunityMining.ppt 2009/04/09.