Impact of Data Warehousing and Data Mining in Decision Making

Monika Pathak
Department of Computer Science
Multani Mal Modi College, Patiala
Patiala, Punjab, India

Sukhdev Singh
Department of Computer Science
Multani Mal Modi College, Patiala
Patiala, Punjab, India

Sukhwinder Singh Oberoi
Department of Computer Science
Guru Hargobind Sahib Khalsa
Girls College, Karhali Sahib
Patiala, Punjab, India

Abstract-Today's reporting environment give users to access their data, but it does not solve all the problems of user. The users have privilege to access the data but do not guarantee the integrity of the data and adequacy of response time. Data warehousing solve the above problems and provide technology which enables the user or decision maker to process the huge amount of data in a short amount of time. With the help of data warehousing, user extract the knowledge in a real time and its help the user in the decision making. Many companies want to use that data for other purposes. So data mining techniques are evolved for extracting new knowledge from data warehouse. Data warehousing and data mining provide the right foundation for building decision support and executive information system tools which help to measure the progressing speed of organization toward its goal. Data warehousing and data mining provide a technology that enables the user or decision-maker in the corporate sector/govt. to process the huge amount of data and make decisions which are useful for whole organisation. This paper tries to explore the overview, advantages and disadvantages of data warehousing and data mining with suitable diagrams. In this paper, roles and responsibilities of organizational members of data warehousing are also discussed. As a concluding point, we are trying to show as how "Date Warehouses & Data Mining" can be used in organizations, how their data help in decision making and allow the manager to perform more accurate, substantive and consistent analysis.

Keywords-Data warehousing, data mining, decision support system, Staging Layer, Data Marts, Operational Data Store, Knowledge Discovery.

I. INTRODUCTION

The concept of Data Warehousing and Data Mining is becoming increasingly popular as a business information management tool where it is expected to disclose knowledge structures that can guide decisions in conditions of limited certainty. A data warehouse supports [1] business analysis and decision-making by creating an enterprise-wide integrated database of summarized, historical information. It integrates data from multiple, incompatible sources. By transforming data into meaningful information, and a data warehouse allows the manager to perform more substantive, accurate and consistent analysis.
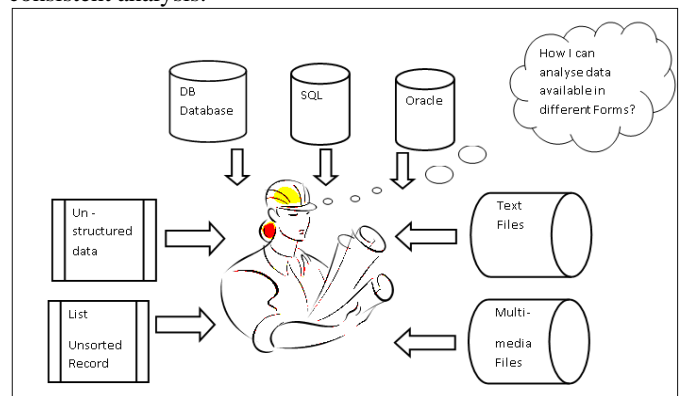

Figure: Problem in Decision Making

The data warehouse is not the normal database, as we understand the term "database". The main difference is that the traditional databases hold operational-type most often, transactional type data and that many of the decision-support type applications put too much strain on the databases intervening into the day-to-day operation (operational database). A data warehouse is of course a database, but it contains summarized information. Data warehouse refers to database that is maintained separately from an organizations operational databases. A warehouse holds read-only-data. Data mining, also called Knowledge-Discovery in Databases or Knowledge-Discovery. Data mining, the extraction of

hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

II. DATA WAREHOUSING

A data warehouse is a collection of integrated databases designed to support a DSS. It is a collection of integrated, subject-oriented databases designed to support the DSS function, where each unit of data is non-volatile and relevant to some moment in time. Numerous roles and responsibilities will need to be acceded to in order to make data warehouse efforts successful and generate return on investment. For the technical [6] personnel (application programmer, system administrator, database administrator, data administrator), it is recommended that the following roles be performed full-time by dedicated personnel as much as possible and that each responsible person receive specific Data Warehouse training. The data warehouse team needs to lead the organization into assuming their roles and thereby bringing about a partnership with the business. Management also needs to make actionable plans out of these directives and make sure the staff executes on them. Following are the team, team members and their responsibilities to make data warehouse make effective and helpful to user and organization [2-4]:

Member	Role
Manager/Director	The data warehouse manager or director ensures support for the data warehouse program at the highest levels of the organization and understand high level requirements of the business. Manager staff the team and ensure adherence to a set of guiding principles for data warehousing.
Project Manager	Project managers delivers commitments on time. Project managers maintains highly detailed plan and caring about progress on it. Project manager matching team member's skills and issue list of tasks to them.
Chief Architect	The Manager/Director of data warehouse will need to rely on a Chief Architect position, as one of his/her direct reports, to work on complex issues of architecture, modeling, and tools. Chief Architect would have significant interface with the internal clients and increase their confidence in the data warehouse organization. Chief Architect should have great knowledge of business.
End User	Data warehouse is made to meet end users requirements. Data warehousing is used to answer the end users queries and generate reporting. End user receive ID and password on the data warehouse system and provide feedback to the data warehouse team like performance, functionality, data quality, metadata quality and completeness.
Database Administrator	Data warehouse group is the placement of the database administration function and the division of roles and responsibilities between the support group and the user community. Database administrator has many responsibilities like database maintenance, backup and recovery, data replication, Performance Monitoring and Summary table creation.
Application Programmer Specialist	The Data Warehouse Application Programmer is responsible for applying transformation rules as necessary to keep the data clean and consistent. Application responsibilities has many responsibilities like sourcing the data from operational systems, applying the business transformation rules.
System Administrator	Responsibilities of system administrator are: Installing and maintaining the Database Management, monitoring the performance, architecting the data warehouse architecture. The Data Warehouse System Administrator is responsible for the performance of data transfers, either in response to a query or as part of a data replication or synchronization the effort

Data are organized based on how the users refer to them.

Integrated: All inconsistencies regarding naming convention and value representations are removed.

Nonvolatile: Data are stored in read-only format and do not change over time.

Time Variant: Data are not current but normally time series.

Summarized: Operational data are mapped into a decision-usable format

Large Volume: Time series data sets are normally quite large.

Not Normalized: DW data can be, and often are, redundant.

Metadata: Data about data are stored.

Data Sources: Data come from internal and external unintegrated operational systems.

III. ARCHITECT AND WORKING OF DATA WAREHOUSING

Data warehouse is a database used for reporting and analysis. It is a place where data is stored by integrating different Data bases. It can be used for storing current and historical data. With the help of historical and current database new prediction can be drawn. The following diagram [2-3] shows different compounds of data warehouse.
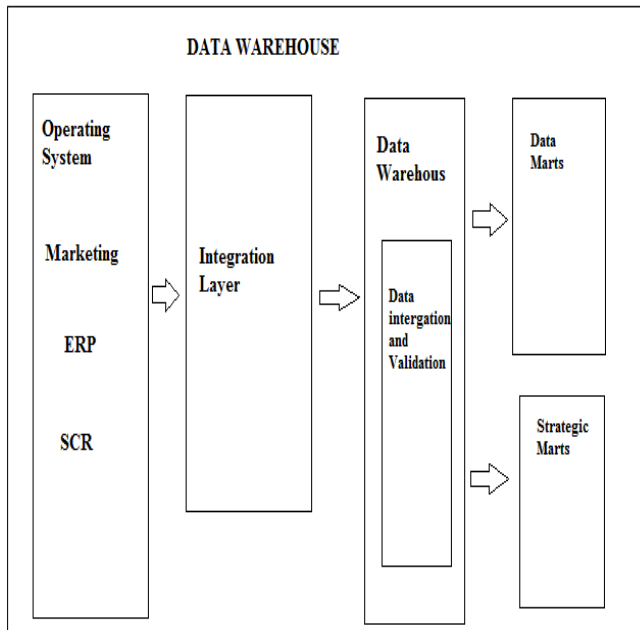

Figure: Components of Data warehouse

Data warehouse is a database used for reporting and data analysis. It is a central source of data which is created by integrating data from one or more different sources. The data stored in the warehouse are received from the operational systems.

The staging layer stores raw data collected from each of the different source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store database.

Raw Data→ Integrated Data Source→ Data Warehousing→Report→ Decision Making.

A data mart is a small data warehouse concentrated on a specific area of interest. Data warehouses can be subdivided into data marts for improved performance in use. The company can have one or more data marts towards a larger and more complex enterprise data warehouse.

A Data Warehouse saves time of business user and helps to generate the reports quickly. Business users can quickly use these reports on one place and can take decisions quickly. Business users won't waste their precious time in collecting data from multiple sources. With the help of data warehousing, business can query the data themselves and saves money and time

IV. DATA MINING

The data mining applications are available on all size systems for mainframe, client/server, and PC platforms. Data base mining or Data mining is a process that aims to use existing data to invent new facts and to uncover new relationships.

Data mining includes several steps: problem analysis, data extraction, data cleansing, rules development, output analysis and review. Data mining sources are typically flat files extracted from on-line sets of files, from data warehouses or other data source. Data may however be derived from almost any source.

Whatever the source of data, data mining will often be an iterative process involving these steps. Following are the steps[3-8] of data mining are:-.

1. Uniqueness Identification of the Objective -- Before you begin, be clear on what you hope to accomplish with your analysis. Know in advance the business goal of the data mining. Establish whether or not the goal is measurable.

2. Choice of the Data -- Once you have defined your goal, your next step is to select the data to meet this goal. This may be a subset of your data warehouse or a data mart that contains specific product information. It may be your customer information file. Segment the data as much as possible the scope of the data to be mined. Here are some key issues like 1.How current and relevant are the data to the business goal? 2. Are the data stable—will the mined attributes be the same after the analysis?

3. Compilation of the Data -- Once you've assembled the data, you must decide which attributes to convert into usable formats. Consider the input of domain experts/creators and users of the data. Establish strategies for handling missing data, extraneous noise, and outliers. Decide on a log or square transformation, if necessary. Determine the distribution frequencies of the data?

4. Evaluate the Data -- Evaluate the structure of your data. What is the nature and structure of the database? What is the overall condition and distribution of the dataset?

5. Choice of Appropriate Tools -- Two important factors for the selection of the appropriate data-mining tool business objectives and data structure. Both should guide you to the same tool. No single tool is preferred to answer the queries

6. Prepare indented the Solution – Find out the answers of some questions like: What are the available format options? What is the goal of the solution? What do the end-users need graphs, reports, code?

7. Prepare the desired Model -- Now the data mining process begins. User split data into sets, construct and evaluate the model. The generation of classification rules, decision trees, clustering sub-groups, scores, code, weights and evaluation data/error rates takes place at this stage.

8. Check and Validate the Findings -- Share and discuss the results of the analysis with the business client or domain expert. Ensure that the findings are correct and appropriate to the business objectives. Find out the answers of many queries like-Do the findings make sense?

9. Reporting the Findings -- Prepare a final report for the business unit or client. The report should document the entire data mining process including data preparation, tools used, test results, source code, and rules. This report helps in decision making and plays important role in the growth of organization.

10. Combine components to integrate the solution -- Share the findings with all interested end-users. You might wind up incorporating the results of the analysis into the company's business procedures. Although data mining tools automate database analysis, they can lead to faulty findings and erroneous conclusions if you're not careful.

Figure: Working of Data Mining with Data warehouse

Data mining can be applied to operational databases with individual transactions. Both private and public sectors such as banking, insurance, pharmaceutical manufacturers, health care providers, and retailing are using data mining for a variety of purpose to reduce costs, enhance research, predict the effectiveness of a procedure or medicine, and increase sales.

Data mining is used to predict future trends, customer purchase habits and help in decision making. Data mining

improve company revenue and lower costs. Data mining is also used in analyzing the market and find out the frauds.

But data mining has many limitations too. Data mining has privacy or security issues. Data mining sometimes is costly at implementation stage. Data mining has privacy issues and misuse of information. Data mining cannot promise perfect results, cannot explain why an outcome occurs, and cannot correct problems in your data.
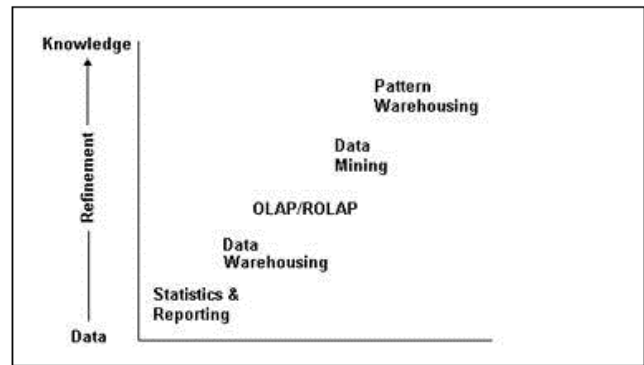

Figure: Refinement of Knowledge with Data representation

V. ROLE OF DATA WAREHOUSING AND DATA MINING IN DECISION

The goal of a data warehouse is to support decision making with data. Data mining [9-10] can be used in conjunction with a data warehouse to help with certain types of decisions. To be successful, data warehousing and data mining needs a skilled user who will supply the correct data and a specialist who can make objective conclusions out of the output that is created. If the user supplies incorrect or minimal amount of information, output will be affected and forecast will not be credible.

Data warehousing and data mining plays an important role in decision making of the organization. Data warehousing provide answers of many queries to the organization and the user and helps in decision making. There are many types of queries of the organization like tactical query, strategic query, and update query.

A tactical query [8] is a database operation that attempts to determine the best course of action right now. Whereas the strategic query provides the information necessary to make long term business decision, a tactical query provides information to rank and file elements in the field that need to respond quickly to a set of unfolding events. Tactical queries tend to produce a very small result set. It is not uncommon for the result set to be less than a dozen rows. Usually the result set is designed to fit into a single window on a display screen.

A strategic query is a database operation that attempts to determine what has happened, why it happened, and/or what will happen next. It typically accesses vast amounts of detailed data from the warehouse and ranges in complexity from simple table scans to multi-way joins and sub queries. Applications that generate strategic queries include; report generation, OLAP, decision support, ad-hoc, data mining, etc.
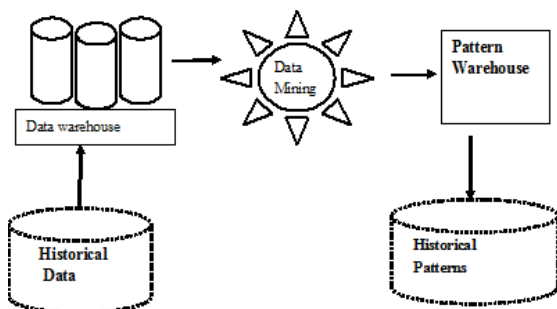
An update query is a database operation that modifies the state of a database. Teradata provides a set of bulk load utilities used to load large quantities of data into the database in an efficient fashion.

VI. CONCLUSION

Data Mining and Warehousing talks about the change in business trends these days. All the small and big industries are collecting and using data from various sources to identify their own business trends. Organizations understand the strengths and the weaknesses of their competitor improve their progressing speed towards the goal and expand their business empire. A data warehouse is a solution to a business problem not a technical problem. The data warehousing and data mining need to constantly overcome obstacles that are yet undefined and help the organization in decision making and improves the goodwill of organization. Data mining helps in securing and processing the data into understandable chunks, where warehousing helps in analyzing the data and put it in such a way as to facilitate comparison between trends, analyzing the data for the business predictions and accelerate decision making. In short, a data warehousing and data mining implementation includes the conversion of data from various source systems into a common format with accuracy, help the organization in the strong business decision and help to expand the business empire. A Data Warehouse Enhances Consistency and Data Quality each data from the various departments is standardized, each department will produce results that are in line with all the other departments. It is relevant and organized in an efficient manner. One powerful feature of data warehouses is that data from different locations can be combined in one location.

VII. FUTURE SCOPE

Data mining offers an important approach to achieving values from the data ware house for use in decision support. Data warehousing becomes a standard part of an organization, there will be efforts to find new ways to use the data. Data warehousing and data mining will bring several new challenges in future like 1. Regulatory constraints may limit the ability to combine sources of disparate data. 2. These disparate sources are likely to contain unstructured data which is hard to store. 3. The internet makes it possible to access data from virtually "anywhere". This just increases the disparity. Today the challenge is to design data warehousing and data mining applications that are reliable, easy to use and supports effective decision making. As the amount of data increases in the future, data mining and data warehousing will become a valuable tool in industries/business. Data mining [13-16] will be helpful in finding new quality products, predict the benefits from that quality data, and can help optimize use of sales resources like manpower and marketing.

REFERENCES

[1]. Edwin M Knorr and Raymond T. N.(1998), "Algorithms for Mining Distance-Based Outliers in Large Datasets", Proceedings of 24rd International Conference on Very Large Data Bases, York, USA.

[2]. Brachman R. J. and Anand. T.(1996), "The process of knowledge discovery in databases: A human centered approach", chapter 2, 37-57. AAAI/MIT Press.

[3]. Wirth, R. and Hipp. J.(2000)," CRISP-DM: Towards a standard process model for data mining", The 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39, Manchester, UK.

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

[5]. Frawley W., Piatetsky Shapiro G. and Matheus C.(1992), "Knowledge Discovery in Databases: An Overview", AI Magazine, Fall, 213-228,.

[6]. Daskalaki, S., Kopanas, I., Goudara, M., and Avouris, N.(, 2003) "Data mining for decision support on customer insolvency in telecommunications business", Europeon Journal of Operational Research, Vol. 145, Issue. 2, 239-255.

[7]. Chen, C. and Lewis, B(2002) , "A basic primer on data mining, Information Systems Management", 56-60.

[8]. M.J.A.Berry,G. Linoff, "Data Mining Techniques: For Marketing, Sales and Customer Relationship Management", second ed., Wiley, New York,2004.

[9].C.X. Ling, C. Li(1998), "Data mining for direct marketing: Problems and solutions", in Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.

[10]. Berry, Michael J.A., and Gordon Linoff (1997)," Data mining techniques: for marketing, sales, and customer support", New York: Wiley.

[11]. S.C. Hui, G. Jha(2000), "Data Mining for Customer Service Support", Information & Management, Elsevier.

[12]. Jiawei Han, Micheline Kamber, Jian Pei(2005), "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann.

[13]. J. M. Zytkow and W. Klösgen(2002), Handbook of Data Mining and Knowledge Discovery. New York: Oxford.

[14]. Barry, D.(1997), "Data Warehouse from Architecture to Implementation", Addison-Wesley.

[15]. Fayyad U., Piatetsky-Shapiro G., and Smyth P.(1996), "Knowledge Discovery and Data Mining: Towards a Unifying Framework", proceeding of 2nd International Conference on Knowledge Discovery and Data Mining, Portland, pp.82-88.

[16]. N. R. T., Han J.(1994), "Efficient and Effective Clustering Methods for Data Mining", International conference on Very Large Data Bases, Santiago, Chile, pp.144-15