Volume 1, Issue I, Feb 2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Review of Data Mining & Data Warehousing

Anju Saini

A.P. in PIET, Samlakha, Panipat,India
anju.saini.20@gmail.com

Abstract- *Data mining and Data warehousing process involves multiple stages. Data warehousing is a method of bringing together all of a company's data from various computer systems, including those relating to customers, employees, vendors, products, inventory and financials. Data warehouse connects different databases together in order to offer a more comprehensive data set for making decisions. Data mining is a part of a process called KDD-knowledge discovery in databases. This process consists basically of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation. Association rule techniques are used for data mining if the goal is to detect relationships or associations between specific values of categorical variables in large data sets. In the past user would repeat the whole procedure to solve the basic problem, which is time-consuming in addition to its lack of efficiency. From this, the importance of data mining and data warehousing process appears and for this reason this problem is going to be the main topic of this paper. Therefore the purpose of this study is to give the basic knowledge about the data mining and data warehousing that will help to solve the basic problems of the organization.*

Keyword: *Data warehousing, data mining, KDD*

1. INTRODUCTION

Data Mining is a process of looking for unknown relationships and patterns and extracting useful information volumes of data in data warehouse. The rapid progress of computers and databases has enable companies to store data about customers and transactions for future use. The sheer amounts of data to be analyzed in order to make better decisions require dramatically improved new automated data modeling technologies then the concept of data mining is developed.

Data Mining Process: KDD refers to the broad process of finding knowledge in data, and emphasizes the high-level application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. There are various steps that are involved in data mining as shown below.

Data Integration: First of all the data are collected and integrated from all the different sources.

Data Selection: Data will have collect in the first step. So in this step select only those data which author thinks useful for data mining.
Data Cleaning: The data which author collected are not clean and may contain errors, missing values, noisy or inconsistent data. So author needs to apply different techniques to get rid of such anomalies.

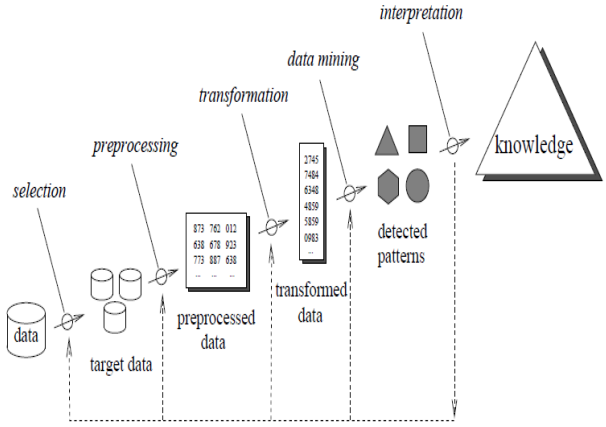

Figure 1: Data Mining Process [2]

Data Transformation: The data even after cleaning are not ready for mining as author need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
Data Mining: Now author is ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.
Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated. [1]

Volume 1, Issue I, Feb 2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Decisions / Use of Discovered Knowledge: This step helps user to make use of the knowledge acquired to take better decisions.

Data Warehouse: Data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. It is typically a combination of detailed demographic data on a customer, combined with a historical transactional history, which may include not only the purchases that were made by the customer, but also include contact or interaction data such as what type of promotions were made to each customer, which ones did they respond to, have they called on their own with support related questions, or inquire about a certain product.

Data Warehouse Process

Subject-oriented: The data in the data warehouse is organized so that all the data elements relating to the same real-world event or object are linked together.

Time-variant: The changes to the data in the data warehouse are tracked and recorded so that reports can be produced showing changes over time.

Non-volatile: Data in the data warehouse is never over-written or deleted - once committed, the data is static, read-only, and retained for future reporting.

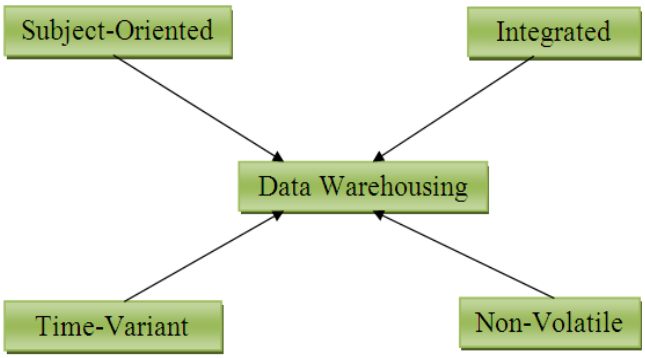

Figure 2: Data Warehousing Process

Integrated: The data warehouse contains data from most or all of an organization's operational systems and this data are made consistent.

2. ARCHITECTURE

The classical architecture of data mining systems is one-tier architecture. Such a system is completely client based. Basically all data mining systems of the first generation are based on this architecture. The user has to select a small subset of data warehouse data and load it on the client in order to make it accessible to the data mining tool. This tool may offer several data mining techniques. The most obvious drawback of the one-tier approach is the size of the data set that can be mined and the speed of the mining process. [5] This is often overcome by selecting a random sample from the data.

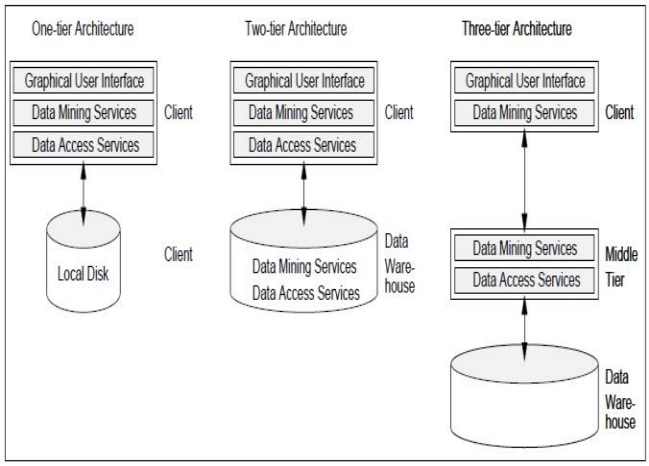

Figure 3: Data Mining Architecture

A truly random (unbiased) sample is needed to ensure the accuracy of mined patterns, and even then patterns relating to small segments of the data can be lost. The data resides in raw files of the client's file system. Another disadvantage is the absence of a multi-user functionality. Each User has to define his own subset of the data warehouse and load it separately onto the client machine. Since each user runs his own client-based data mining software, there is no way for data mining specific access control and control of system resources. Optimization of the data mining process is restricted to choosing more efficient implementations of the data mining techniques. [9]

Data Warehousing Architecture

An organization's data life cycle management's policy will dictate the data warehousing design and methodology. The goal of Data Warehousing is to generate front-end analytics that will support business executives and operational managers.

Pre-Data Warehouse: The pre-Data Warehouse zone provides the data for data warehousing. Data Warehouse designers determine which data contains business value for insertion.

Data Cleansing: Before data enters the data warehouse, the extraction, transformation and cleaning (ETL) process ensures that the data passes the data quality threshold. ETLs are also

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

responsible for running scheduled tasks that extract data from OLTPs.

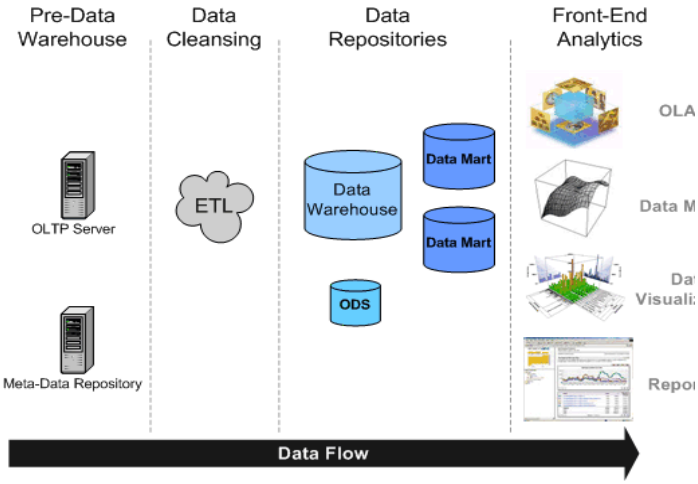

Figure 4: Data Warehousing Architecture [3]

Data Repositories: The Data Warehouse repository is the database that stores active data of business value for an organization. The Data Warehouse modeling design is optimized for data analysis.
Data Warehouses collects data and is the repository for historical data. Hence it is not always efficient for providing up-to-date analysis. This is where ODS, Operational Data Stores, come in. ODS are used to hold recent data before migration to the Data Warehouse.
Front-End Analysis: The last and most critical potion of the Data Warehouse overview are the front-end applications that business users will use to interact with data stored in the repositories.

3. DATA CUBE

The data cube, also known in the OLAP community as the multi-dimensional database. A data cube is constructed from a subset of attributes in the database.
Certain attributes are chosen to be measure attributes, i.e., the attributes whose values are of interest. Other attributes are selected as dimensions or functional attributes. The measure attributes are aggregated according to the dimensions. The Figure 1 below depicts a small, practical data cube example; consider a Hypothetical database of sales information maintained by a company. This particular data cube has three feature attributes - store, product, and time - and a single measure attribute – product sales for a large chain of stores (sales is computed with the sum function).

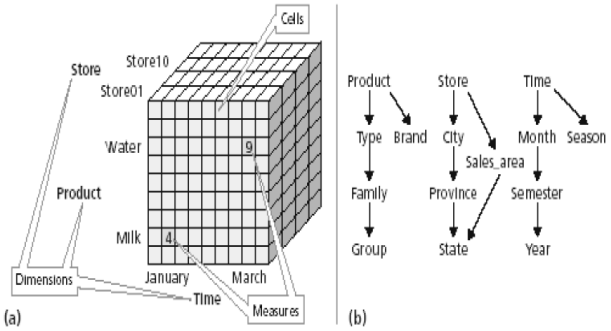

Figure 5: Multidimensional Model Data Cube

(a) The cube itself is composed of cells that define fact attributes, while (b) the classification hierarchies display the dimensions that define the cube – product, store and time. By selecting cells, planes, or subcubes from the base cuboid, we can analyze sales figures at varying granularities. Such queries form the basis of OLAP functions like roll-up and drill-down. In total, a d-dimensional base cube is associated with 2d cuboids. Each cuboid represents a unique view of the data at a given level of granularity. Not all these cuboids need actually be present, however, since any cuboid can be computed by aggregating across one or more dimensions in the base cuboid. Nevertheless, for anything but the smallest data warehouses, some or all of these cuboids may be computed so that users may have rapid query responses at run time. [15]

4. APPLICATION OF DATA WAREHOUSING

There are three kinds of data warehouse applications: information processing, analytical processing, and data mining:
Information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts or graphs. A current trend in data warehouse information processing is to construct low cost Web-based accessing tools which are then integrated with Web browsers. [13]
Analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historical data in both summarized and detailed forms. The major strength of on-line analytical processing over information processing is the multidimensional data analysis of data warehouse data.
Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

5. FROM ON-LINE ANALYTICAL PROCESSING TO ON-LINE ANALYTICAL MINING

In the field of data mining, substantial research has been performed for data mining at various platforms, including transaction databases, relational databases, spatial databases, text databases, time-series databases, data warehouses, etc. Among many different paradigms and architectures of data mining systems, OLAM integrates OLAP with data mining and mining knowledge in multidimensional databases is particularly important for the following reasons.

High quality of data in data warehouses. Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation and data integration as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of high quality data for OLAP as well as for data mining. Notice that data mining may also serve as a valuable tool for data cleaning and data integration as well. [6]

Available information processing infrastructure surrounding data warehouses. Comprehensive information processing and data analysis infrastructures have been or will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple, heterogeneous databases, ODBC/OLE DB connections, Web-accessing and service facilities, reporting and OLAP analysis tools. It is prudent to make the best use of the available infrastructures rather than constructing everything from scratch [8]. OLAP-based exploratory data analysis. Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms. On-line analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction, by drilling, pivoting, filtering, dicing and slicing on a data cube and on some intermediate data mining results. This, together with data/knowledge visualization tools, will greatly enhance the power and flexibility of exploratory data mining. [14]

On-line selection of data mining functions. Often a user may not know what kinds of knowledge that she wants to mine. By integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically. [10]

6. INTEGRATED ARCHITECTURE

An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing. An integrated OLAM and OLAP architecture is shown in Figure below, where the OLAM and OLAP engines both accept users' on-line queries (or commands) via a User GUI API and work with the data cube in the data analysis via a Cube API. A Meta data directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases and/or by filtering a data warehouse via a Database API which may support OLE DB or ODBC connections. Since an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, etc., [12] it usually consists of multiple, integrated data mining modules and is more sophisticated than an OLAP engine. Data warehousing provides users with large amounts of clean, organized, and summarized data, which greatly facilitates data mining. For example, rather than storing the details of each sales transaction, a data warehouse may store a summary of the transactions per item type for each branch, or, summarized to a higher level, for each country. The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining. Moreover, it also believes that data mining should be a human-centered process. Rather than asking a data mining system to generate patterns and knowledge automatically, a user will often need to interact with the system[7]

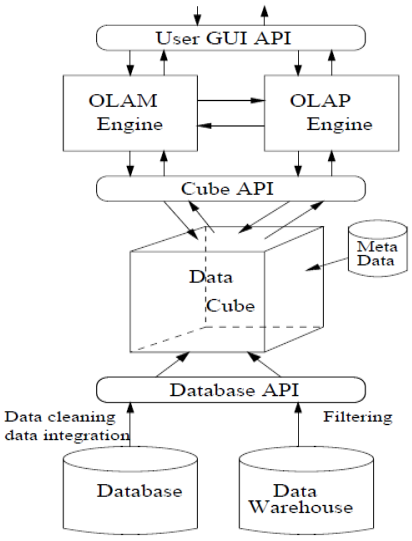

Figure 6: Integrated Architecture

to perform exploratory data analysis. OLAP sets a good example for interactive data analysis, and provides the necessary preparations for exploratory data mining. Consider the discovery of association patterns, for example. Instead of mining associations at a primitive (i.e., low) data level among

transactions, users should be allowed to specify roll-up operations along any dimension. [11]
For example, a user may like to roll-up on the item dimension to go from viewing the data for particular TV sets that were purchased to viewing the brands of these TVs, such as SONY or Panasonic. Users may also navigate from the transaction level to the customer level or customer-type level in the search for interesting associations. Such an OLAP-style of data mining is characteristic of OLAP mining.

7. CONCLUSION

It is clear from all that have been said that data mining is still in its infancy, or at the beginning of the road as there are many aspects of data mining that have not been tested. Up-to-date most of the data mining projects have been dealing with verifying the actual data mining concepts. Same as, data warehouses are still an expensive solution and typically found in large firms. The development of a central warehouse is a huge undertaking and capital intensive with large, potentially unmanageable risks. Since this has now been established most researchers will move into solving some of the problems that stand in the way of data mining, this research will deal with such a problem, in this case the research is to concentrate on solving the basic problem of the organization.

ACKNOWLEDGEMENT

Author is grateful to the organizers of PIET, for giving me this invaluable opportunity of presenting my research paper at the Conference. Author is grateful to all the Data Mining and Data Warehousing Entrepreneurs for their cooperation in giving us first hand information to complete my research survey on time. Author is also thankful to Research Institutes for the unlimited access to various National and International journals and various books pertaining to my research work for providing me with right data of data mining and warehousing. Most important, author is grateful and thankful to my Almighty, Miraculous Lord God who gave me the strength and wisdom to complete this research paper on time.

References

[1] Maria Halkidi, "Quality assessment and Uncertainty Handling in Data Mining Process" http://www.edbt2000.unikonstanz.de/phd-workshop/papers/Halkidi.pdf.
[2]Fayyad, U. M., G. P. Shapiro, P. Smyth. "From Data Mining to Knowledge Discovery in Databases", 0738-4602-1996,AI Magazine (Fall 1996): 37–53.
[3]Jiawei Han, Micheline Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Champaign: CS497JH, fall 2001, www.cs.sfu.ca/~han/DM_Book.html.
[4]David Hand, Heikki Mannila, Padhraic Smyth. "Principles of Data Mining", ISBN: 026208290 MIT Press, Cambridge, MA, 2001.
[5] Fernando Crespoa, Richard Weberb. "A methodology for dynamic data mining based on fuzzy clustering", Fuzzy Sets and Systems 150 (2005) 267–284.
[6] Papadimitriou, J.Sun, C.Faloutsos, "Streaming Pattern Discovery in Multiple Time-Series", Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005,p697-708.
[7] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy. "Mining Data Streams: A Review", VIC3145, Australia, ACM SIGMOD Record Vol. 34, No. 2; June 2005.
[8] Two Crows Corporation. "Introduction to Data Mining and knowledge Discovery", ISBN: 1-892095-02-5,
[9] Ruoming Jin and Gagan Ag awal. "A Middleware for Developing Parallel Data Mining Applications", Proc. of the 1-stSIAM Conference on Data Mining, 2000 - cs.ubc.ca.
[10] W.H. Inmon and C. Kelley. Rdb/VMS: Developing the Data Warehouse. QED Publishing Group, Boston, Massachussetts, 1993.
[11] Codd, E.F., S.B. Codd, C.T. Salley, "Providing OLAP (On-Line Analytical Processing) to User Analyst: An IT Mandate.
A. Gupta and I.S. Mumick. Maintenance of materialized views: Problems, techniques, and applications. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Warehousing, 18(2):3{18, June 1995}.
[12] Sen, A. and Sinha, A. P. (2005): A Comparison of Datawarehousing Methodologies, Communication *of the ACM*, **48**(3), 79-84.