ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

An Overview of Database management System, Data warehousing and Data Mining

Ramandeep Kaur[1], Amanpreet Kaur[2], Sarabjeet Kaur[3], Amandeep Kaur[4], Ranbir Kaur[5]

Assistant Prof., Deptt. Of Computer Science, Baba Farid College, Bathinda, India[1]

Assistant Prof., Deptt. Of Computer Science, CGC Landran, Mohali, India[2]

Assistant Prof., Deptt. Of Computer Science, Baba Farid College, Bathinda, India[3]

Assistant Prof., Deptt. Of Computer Science, Baba Farid College, Bathinda, India[4]

Assistant Prof., Deptt. Of Computer Science, Baba Farid College, Bathinda, India[5]

ABSTRACT: DBMS, Data Ware House and Data mining which basically focus on the management of data. Data retrieval and Data security are basic concept for decision support in data management. Various commercial Services are available in DBMS. So Database plays a very important role in our real life. This paper provides and overview of database, database warehousing and mining the data in database, with an emphasis on their new requirement. We discuss here bank end tools for managing the data extracting, cleaning and loading data into a data warehouse and front end client tools for querying and data analysis server extensions for processing the efficient query, tools for metadata management and for managing the warehouse. This overview gives us the basic knowledge of various database tools and techniques.

KEYWORDS: Architecture, Database, Datawarehouse, Data Mining, management

1. INTRODUCTION

A DBMS is a collection of programs which provides the management of database. It has proper control access to the data and query language to retrieve the information .The database contains proper data up to the needs and occupies minimum storage space.The database contains no unnecessary data and we can added and upadate data .

1.1 ADVANTAGES OF DATBASE APPROACH

1.1.1 Control on Data Redundancy: It controls the redundancy in data storage and in and in development and maintenance efforts. In non database system traditional computer file processing each application program has its own files. In this case duplicated copies of the same files will be created at many places. But in DBMS all the data of the organization is itegrated into the single database. The data is recorded as one place and it's not duplicated.

1.1.2 Data Sharing: A database allows the sharing of data under its control by any number of application programs or users. In other words we can say that accessing the data by multiple users at same time.

1.1.3 Security**:** Restricting unauthorized access of data. When multiple users sharing the data in large database, it is likely that most of users will not be authorized to access all information in the database.

1.1.4 Providing persistent storage: A complex object in C++ can be stored permanently in an Object Oriented DBMS. Such an object is said to be persistence, since it survives the termination of the program execution and can later be directly retrieved by another C++ program.

1.1.5 Integrity: Integrity means to validity and consistency of the stored data. For the Integrity we need to apply various types of constraints like Primary Key, Foreign Key, Unique, NOT NULL, Check.

1.1.6 Remove Inconsistencies: The main advantage of avoiding duplication is the elimination of inconsistencies that tend to be present in redundant data files. Any redundancies that exist in the DBMS are controlled and the system ensures that these multiple copies are consistent.

1.1.7 Backup and recovery: It provides recovery and backup from the failures like disk crash, power failure etc which help to recover the database from inconsistent state.

1.2 ARCHITECTURE OF DBMS

Three level Architecture suggested by ANSI/SPARC.It produced an interim report in 1972 followed by a final report in 1977.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

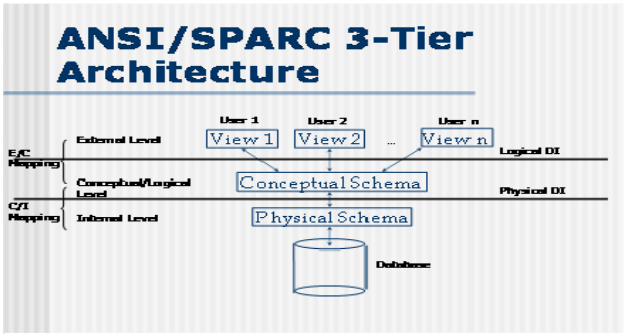

Figure 1 Levels in 3-tier architecture

1.2.1 External level: It is the highest level the architecture. It defines the various types of user's views.

1.2.2 Conceptual level: It is the middle level in the architecture. It defines the logical structure of whole database for a community of users. It also defined by the conceptual schema which describe all the database entities, attributes and relationships with integrity.

1.2.3 Physical Level: It is the third level of 3-tier architecture. It defines the physical storage of data.

II. DATA WAREHOUSE

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.

2.1 CHARACTERISTICS OF DATA WAREHOUSE

2.1.1 Subject oriented: Data are organized based on how the users refer to them.
2.1.2 Integrated: All inconsistencies regarding naming convention and value representations are removed.
2.1.3 Nonvolatile: Data are stored in read-only format and do not change over time.
2.1.4 Time variant: Data are not current but normally time series.
2.1.5 Summarized: Operational data are mapped into a decision-usable format.
2.1.6 Large volume: Time series data sets are normally quite large.
2.1.7 Not normalized: DW data can be, and often are, redundant.
2.1.8 Metadata: Data about data are stored.
2.1.9 Data sources: Data come from internal and external unintegrated operational systems.

2.2 DATA WAREHOUSE ARCHITECTURE

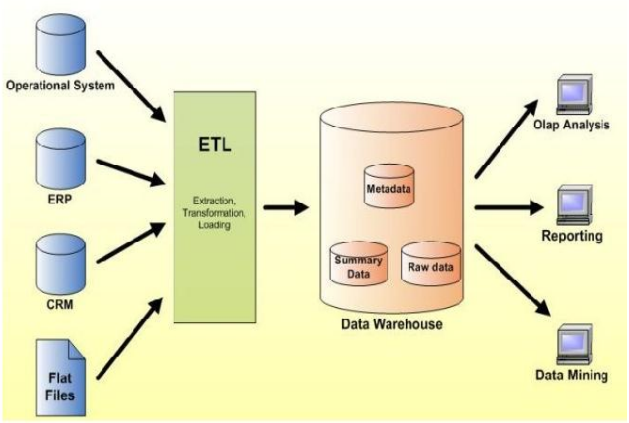

Figure 2 Data warehouse Architecture

2.2.1Datawarehouse Architecture: It includes tools for extracting data from multiple operational databases and external sources for cleaning, transforming and integrating this data for loading data into the data warehouse and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, for slower archival storage. In addition to the main warehouse, there may be several departmental data marts. Data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of frontend tools: query tools, report writers, analysis tools, and determining tools.
2.3 Data Mining:Data mining (knowledge discovery from data) Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

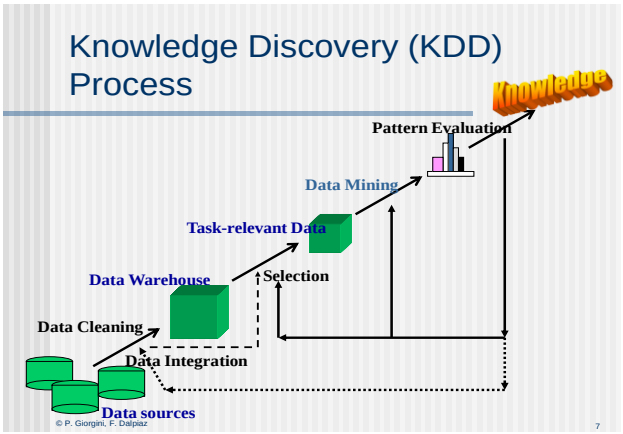

Figure 3 Knowledge Discovery Process

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

2.3.1Data Mining Operations and Associated Techniques

Operations	Data mining techniques
Predictive modeling	Classification
	Value prediction
Database segmentation	Demographic clustering
	Neural clustering
Link analysis	Association discovery
	Sequential pattern discovery
	Similar time sequence discovery
Deviation detection	Statistics
	Visualization

2.3.2 Predictive Modelling: Similar to the human learning experience, It uses observations to form a model of the important characteristics of some phenomenon. Generalizations of 'real world' and ability to fit new data into a general framework. It can analyze a database to determine an essential characteristic (model) about the data set.Model is developed using a supervised learning approach, which has two phases: training and testing.

Training builds a model using a large sample of historical data called a training set. Testing involves trying out the model on new, previously unseen data to determine its accuracy and physical performance characteristics. Applications of predictive modelling include customer retention management, credit approval, cross selling, and direct marketing.

2.3.3.1 There are two techniques associated with predictive modelling: classification and value prediction, which are distinguished by the nature of the variable being predicted.

Predictive Modeling – Classification-Used to establish a specific predetermined class for each record in a database from a finite set of possible, class values. Two specializations of classification: tree induction and neural induction.

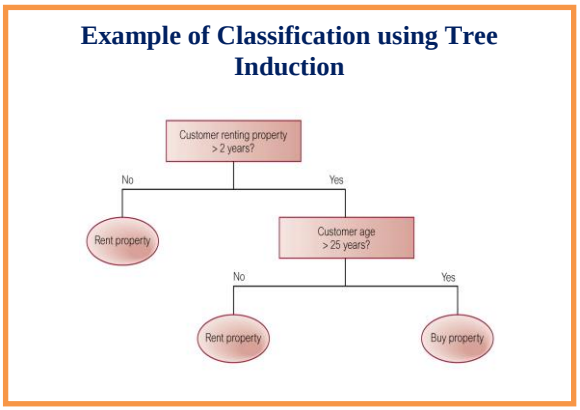

Figure 4 Predictive Modelling- Classification

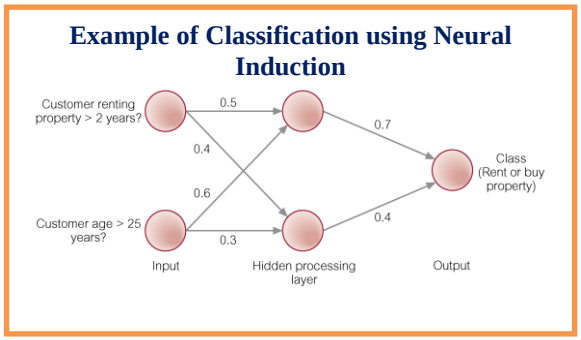

Figure 5 Predictive Modelling-Neural Induction

Predictive Modelling - Value Prediction – It used to estimate a continuous numeric value that is associated with a database record and the traditional statistical techniques of linear regression and nonlinear regression. It is relatively easy-to-use and understands. Linear regression attempts to fit a straight line through a plot of the data, such that the line is the best representation of the average of all observations at that point in the plot.

Problem is that the technique only works well with linear data and is sensitive to the presence of outliers (that is, data values which do not conform to the expected).Although nonlinear regression avoids the main problems of linear regression, it is still not flexible enough to handle all possible shapes of the data plot. Statistical measurements are fine for building linear models that describe predictable data points; however, most data is not linear in nature. Data mining requires statistical methods that can accommodate non-linearity, outliers, and non-numeric data. Applications of value prediction include credit card fraud detection or target mailing list identification.

2.3.4 Database Segmentation- Aim is to partition a database into an unknown number of segments, or clusters, of similar records. It uses unsupervised learning to discover homogeneous sub-populations in a database to improve the accuracy of the profiles.it is less precise than other operations thus less sensitive to redundant and irrelevant features. Sensitivity can be reduced by ignoring a subset of the attributes that describe each instance or by assigning a weighting factor to each variable. Applications of database segmentation include customer profiling, direct marketing, and cross selling.

2.3.5 Example of Database Segmentation using a Scatter plot

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

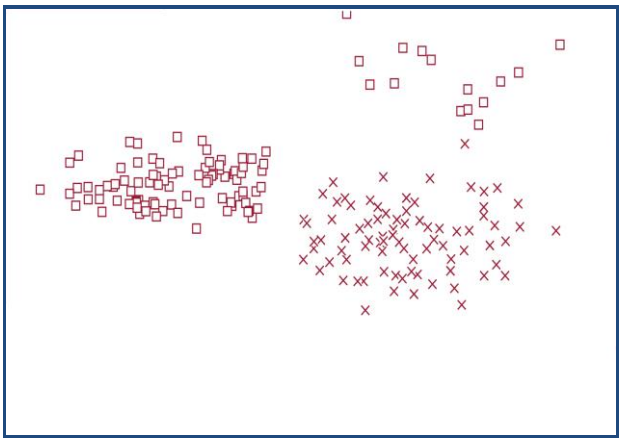

Figure 6 Database Segmentation

2.3.6 Associated with demographic or neural clustering techniques, which are distinguished by Allowable data inputs, methods used to calculate the distance between records, presentation of the resulting segments for analysis

2.3.7 Link Analysis-Aims to establish links (associations) between records, or sets of records, database.Applicationsinclude product affinity analysis, direct marketing, and stock price movement.

2.3.8 Deviation Detection- Relatively new operation in terms of commercially available data mining tools. Often a source of true discovery because it identifies outliers, which express deviation from some previously known expectation and norm. it can be performed using statistics and visualization techniques or as a by-product of data mining. Applications include fraud detection in the use of credit cards and insurance claims, quality control, and defects tracing.

Example of Database Segmentation using Visualization

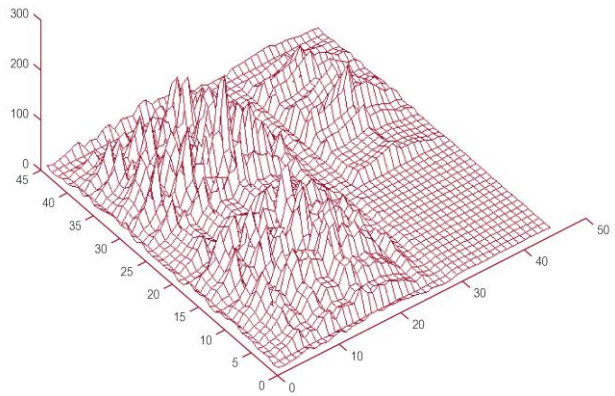

Figure 7 Database segmentation

2.3.10 Data Mining Architecture

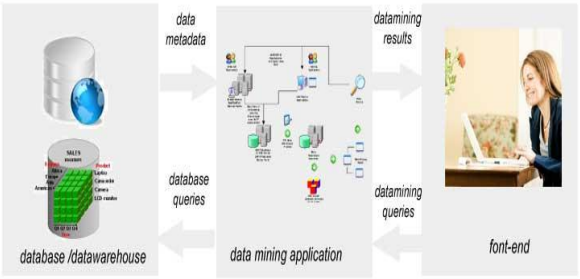

Figure 8 Data Mining Architecture

2.3.11 There are three tiers in the tight coupling data mining architecture.

Data Layer: Data layer can be database or Datawarehouse systems. This layer is an interface for all data sources. Data mining results are stored in the data layer so it can present to end users in form of reports or other kind of visualization.

Data mining application layer is used to retrieve data from database. Some transformation routine can be performed here to desired format. Then data is processed using various data mining algorithms.

Front-end layers provides intuitive and friendly user interface for end user to interact with data mining system. Data mining result presented in visualization form to user in the front-end layer.

III. COMPARISON BETWEEN DBMS, DATAWARE HOUSE, DATA MINING

DBMS is basically management of data and relational manner and organized that data. It's placing of data by removing duplicacy, inconsistency, and apply integrity rules so that our data be placed in secure form in data base management system. User can apply different query according to their needs.

Data mining is the process of finding patterns in a given data set. These patterns can often provide meaningful and insightful data to whoever is interested in that data. Data mining is used today in a wide variety of contexts- like in fraud detection, as an aid in marketing campaigns and even supermarkets use it to study their consumers. This is a perfect example of data mining – credit cards.

We can say that data warehousing is basically a process in which data from multiple sources /databases is combined into one comprehensive and easily accessible database. Then this data used by anyone who wants to mining the

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

some useful data relates to field. A great example of data warehousing is social website that is facebook that everyone can relate to is what facebook does.it gathers all of your data –yours friends , likes ,comments and your stalk etc- and then stores data into the central repository.

IV.CONCLUSION

In this paper, overview of database, Datawarehouse, Data Mining concept was introduced and further study about architecture. The idea and reasons behind the concept was to describe introduction, features, and comparison between them to make it easy for the learners.

REFERENCES

[1] http://www.bcanotes.com/Download/DBMS/Rdbms/database.pdf
[2] http://stackoverflow.com/questions/3419353/what-is-the-difference between -a-database-and-a-data-warehouse
[3] Inman, W.H., Data Warehouse. John Wiley, 1992.
[4] http://www.zentut.com/data-mining/what-is-data-mining/
[5] http://www.zentut.com/data-mining/data-mining-architecture/
[6] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective.
[7] Srinivasan Parthasarathy, Data Mining overview. srini@cse.ohio-state.edu
[8] http://navdeep19.blogspot.in/2012/04/advantages-and-disadvantages-of.html
[9] http://www.programmerinterview.com/index.php/database-sql/data-mining-vs-warehousing/