International Journal of Engineering Research and General Science Volume 4, Issue 1, January-February, 2016
ISSN 2091-2730

DATA WAREHOUSE AND DATA MINING TECHNIQUES

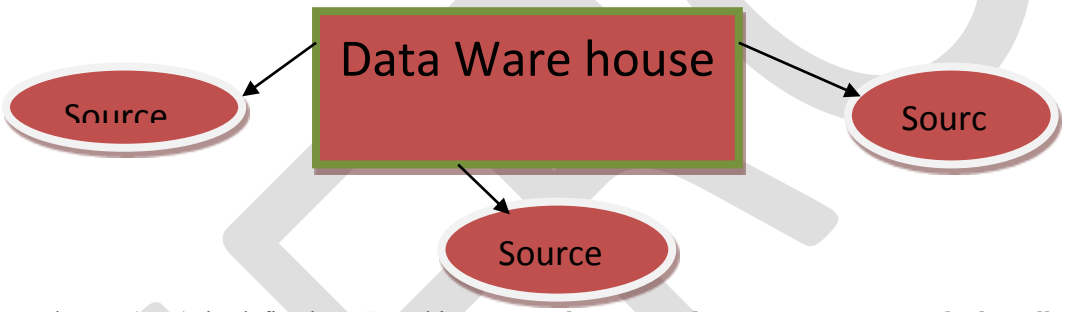
Karamjeet Kaur[1], Kiran Bala[2],
Assistant Professor, Mata Sahib Kaur Girls CollegeTalwandi Sabo, Bathinda, Punjab, India,
Karmjeet39@gmail.com,kiranmour@gmail.com

Abstract—The data mining techniques and applications are marked as important field in the data mining which we used in this paper. These are very helpful to understand the concept of data mining. Even beginners' are understand it very easily The concept of data mining was summarized and its significance towards its methodologies was illustrated. With the wide application of business intelligence in corporate, the demand for data mining software increases daily. To improve the efficiency and quality of the reusing data mining software and reduce the period and cost of developing data mining application system .Data Warehouse is today's biggest need so it's very important to learn about this, and we are trying to give our best to make people understand about data ware house and data mining by this paper.

Keywords- Data mining, DWH, Data mining application, data mining techniques

INTRODUCTION

Data Warehouse is a central managed and integrated database containing data from the operational source in an organization. A data warehousing is not just the data responstries to create a new ,central database but also the architecture and tools to collect, query and analyze the data. Data warehouse is a subject oriented ,integrated time varying, non-volatile collection of data that is used for organization decision making.



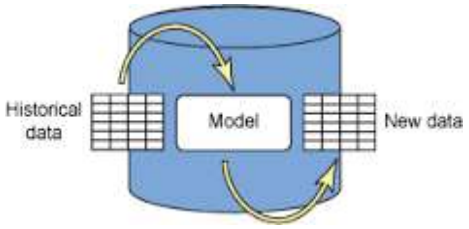
Data Warehouse (DW) is defined as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process". Data warehouses store huge amount of information from multiple data sources which is used for query and analysis. Therefore, the data is stored in the multidimensional (M D) structure. [4]

A data warehouse is a kind of management technique that collect business data from different stations of the enterprise network, so that it can provide efficient data analysis to decision makers. There are some architectural requirements which would govern development of architecture, some of them are: identifying potential users, defining security requirements, skill requirements etc.[2]

Some important points about data warehouse-

- Historical Data
- Large volume of data
- Update in frequently
- It is subject oriented

A multidimensional model stores information into facts and dimensions. A fact contains the interesting concepts or measures (fact attributes) of a business process (sales, deliveries, etc.), whereas a dimension represents the perspective or view for analyzing a fact (product, customer, time, etc.) using hierarchically organized dimension attributes.



Database Design Modal

The database design consists of following five phases. The first phase is Analysis of operational systems whose aim is to collect the information concerning the pre existing operational system. It involves the designer, along with the people involved in managing the information system and produces in output the (conceptual or logical) schemes of either the whole or part of the information system. The next phase consists in gathering and filtering the user requirements [4]

Database Design Modal

Step	Input	Output
Analysis of operational system	Information regarding the operational system	Database schema
Requirement elicitation	Database schema	Specification for DWH
Conceptual design	Database schema and specification	Conceptual schema
Logical design	Conceptual schema	Logical schema
Physical design	Logical schema	Physical schema

After gathering interest information they used different OLAP tools to provide different dimension of data and based on these different dimensions, decisions can be made. Distinguish between OLTP and OLAP: Online Transactional Process (OLTP) is day to day operation of an organization primary business. Such as ATM of banks, flight ticket of travel agency etc. Whereas Online Analytical Processing (OLAP) is capable of handling huge amount of integrated data to process ad-hoc queries, e.g. which books customers likely to buy together.[2]

Views regarding a data warehouse design must be considered: the top-down view, the data source view, the data warehouse view, of the information system.

- **The Top** - Down view allows the selection of the relevant information necessary for the data warehouse. This information matches current and future business needs.
- **The Data source view-** exposes the information being captured, stored, and managed by operational system. This information may be documented at various levels of detail and accuracy, from individual data source tables to integrate at various levels of detail and accuracy, form individual data source tables to integrated data source tables. Data sources are often modeled by traditional data modeling techniques, such as the E-R model or DASE tools.
- **The Data warehouse view**- includes fact tables and dimension tables. It represents the information that is stored inside the data ware house, including pre calculated totals and counts, as well as information regarding the source, date and time of origin added to provide historical context.
- **The Business Query View** -is the data perspective in the data warehouse form the end-user's view point
  So, building and using a data warehouse is a complex task[6]

DATA MODELING TECHNIQUES

Two data modeling techniques that are relevant in a data warehousing environment are ER modeling and Multidimensional modeling.

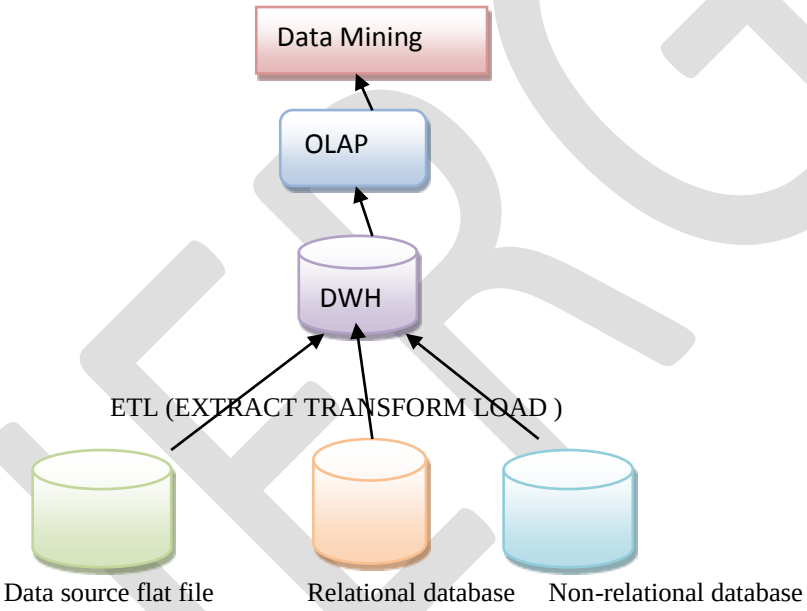
ER modeling produces a data model of the specific area of interest, using two basic concepts: entities and the relationships between those entities. The ER model is an abstraction tool because it can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments.

Multidimensional modeling uses three basic concepts: measures, facts, and dimensions. Multidimensional modeling is powerful in representing the requirements of the business user in the context of database tables. Both ER and Multidimensional modeling can be used to create an abstract model of a specific subject.[3]

Business metadata is content from the data warehouse described in more user-friendly terms. The business metadata tells you, what data you have, where it comes from, what it means and what its relationship is to other data in the data warehouse. We are solving the modern business problems like market analysis and financial forecasting requires query- centric databases schemas that are array-oriented and multidimensional. These business problems are specified by the need to manipulate large numbers of records from very large data sets.[6]

Data Mining-

The Data Mining is the collection of the data, when the data is in the current state then it is called data mining. The inaccurate information is not persisting in data mining. It is the process of discovering hidden data. Purpose of data mining is to extract useful information from different sources for future planning. Data mining is also called KDD (Knowledge Discovery in database) We are live in a world where vast amount of data are collected daily. Analyzing such data is an important need. Data mining used for analyzing data .Given diagram help us to illustrate the Data mining.



DATA MINING LIFE CYCLE:

The life cycle of a data mining project consists of six parts. It depends on the outcome of each part. The main parts are:

- **Business Understanding:**
  This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
- **Data Understanding:**
  It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation:**
  In this stage, it collects all the different data sets and constructs the varieties of the activities basing on the initial raw data
- **Modeling:**
  In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- **Evaluation:**

In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

- **Deployment:**
  The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

DATA MINING APPLICATIONS

Data mining in Healthcare: Data mining applications in health can have tremendous potential and usefulness .However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications [10]

Data Mining in Banking and Finance: Data mining has been used extensively in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction and so on.

Data Mining in Market Basket Analysis: These methodologies based on shopping database. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together. The stores can use this information by putting these products in close proximity of each other and making them more visible and accessible for customers at the time of shopping.

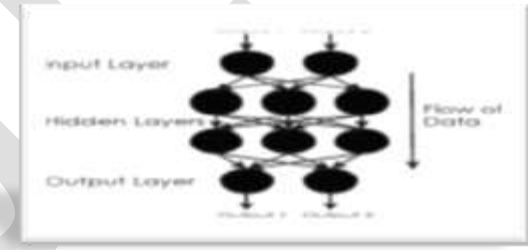
Data Mining in Earthquake Prediction: Predict the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance).

Data Mining in Bioinformatics: Bioinformatics generated a large amount of biological data. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data [11]

DATA MINING TECHNIQUES

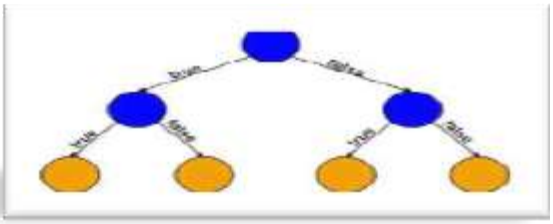
(1). Artificial neural networks:

Non-linear predictive models that learn through training and resemble biological neural networks in structure.[12]



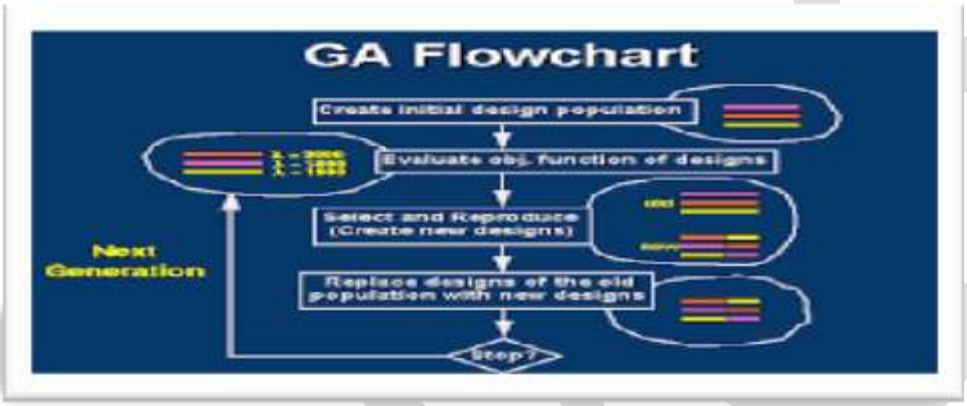
(2). Decision trees:

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).



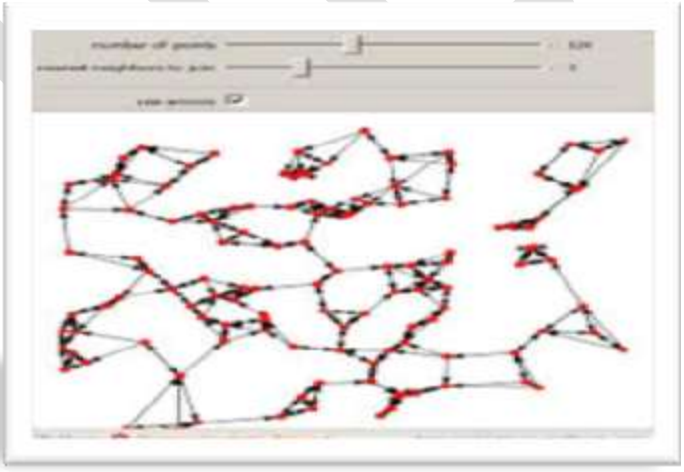
(3). Genetic algorithms:

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.



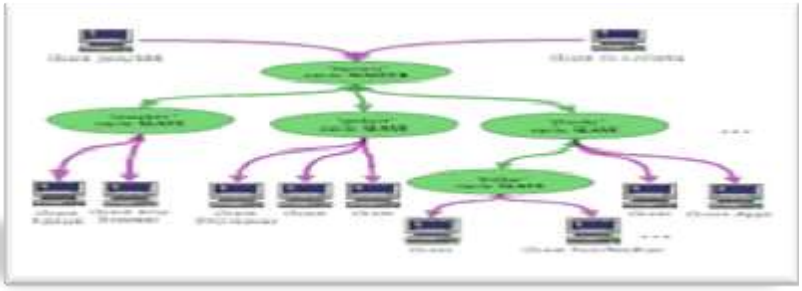
(4). Nearest Neighbor method:

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique. [10]



(5) Clustering:

Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning (i.e. desired output for a given input is not known). For example, image processing, pattern recognition, city planning. [11]



(5).Association:

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit. **Applications**: market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc.

Types of association rules: Different types of association rules based on

- Types of values handled
  - Boolean association rules
  - Quantitative association rules
- Levels of abstraction involved
  - Single-level association rules
  - Multilevel association rules
- Dimensions of data involved
  - Single-
-dimensional association rules
  - Multidimensional association rules [13]

CONCLUSION

A data warehouse is a kind of management technique that collect business data from different stations of the enterprise network, so that it can provide efficient data analysis to decision makers. There are some architectural requirements which would govern development of architecture, some of them are: identifying potential users, defining security requirements, skill requirements etc with the wide application of business intelligence in corporate, the demand for data mining software increases daily. To improve the efficiency and quality of the reusing data mining software and reduce the period and cost of developing data mining application system. In this paper we also include different views of DWH and also explain data ware house life cycle. We have included the detail of data mining techniques which are beneficial to understand the data mining. Everyone should aware about data mining and data warehouse so that we can take advantages of modern technology.DWH has become the necessity for the modern word so that our little

effort helps people to learn about this.

REFERENCES:

[1] Stefano Rizzi, Alberto Aiello's, "Research in Data Warehouse Modeling and Design: Dead or Alive?"

[2] Muhammad Arif, 1Ghulam Mujtaba," A Survey: Data Warehouse Architecture, International Journal of Hybrid Information Technology", Vol.8, No. 5 (2015), pp. 349-356 http://dx.doi.org/10.14257/ijhit.2015.8.5.37

[3] MS. Alpa R. Patel, Pro f. (DR.) Jayesh M. Patel**, "Data Modeling Techniques For Data Warehouse, International Journal Of Multidisciplinary Research", Vol.2 Issue 2, February 2012, ISSN 2231 5780

[4] Rajni Jindal1 and Shweta Taneja2, "Comparative Study Of Data Warehouse Design Approaches: A Survey",International Journal of Database Management Systems (IJDMS) Vol.4, No.1, February 2012

[5]Nirmal Sharma1 and S.K. Gupta2, "Design And Implementation Of Access The Contents In TheData Warehouse", International Journal of Information Technology and KnowledgeManagementDecember2012,Volume 6, No. 1, pp. 61-641 Aryan Institute of Technology, Ghaziabad (UP). E-mail:nirmal1709@rediffmail.com2 B.I.E.T. Jhansi (UP), E-mail: guptask_biet@rediffmail.com

[6] Mr. Dishek Mankad1, Mr. Preyash Dholakia2, " The Study on Data Warehouse Design and Usage, International Journal of Scientific and Research Publications", Volume 3, Issue 3, March 2013 1 ISSN 225315www

[7] Dr. Mohamed F. AlAjmi, Shakir Khan Dr. Arun Sharma "Studying Data Mining And Data Warehousing With Different E-Learning System", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.1, 2013

[8] David Sushil, 2B.Siva NagaRaju,3K.RaghavaRao, "Comparative Study Of Data Warehouse Design Approaches From Security Perspective", David et al. / IJEA Vol. 2 Issue 2 ISSN: 2320-0804

[9] P.Veeramuthu, Dr.R.Periasamy, "Application of Higher Education System for Predicting Student Using Data mining Techniques", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163Volume 1 Issue 5 (June 2014) http://ijirae.com

[10] Neelamadhab Padhy, Dr. Pragnyaban Mishra, and Rasmita Panigrah "The Survey of Data Mining Applications And Feature Scope"Vol.2, No.3, June 2012

[11] Smita1, Priti Sharma "Use Of Data Mining In Various Field: A Survey Paper ",e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 18-21 www.iosrjournals.org www.

[12] Nikita Jain1, Vishal Srivastava2 "Data Mining Techniques: A Survey Papere", ISSN: 2319-1163 | pISSN: 2321-7308

[13] Kalyani M Raval, "Data Mining Techniques", Volume 2, Issue 10, October 2012 ISSN: 2277 128X