DATA MINING ON HETEROGENEOUS DATABASES SYSTEM

Sagar Varma *

Revakant Chaudhari *

Anuprita Nagpure *

ABSTRACT

Distributed system facilitates the Storage, bandwidth and CPUs utilization. It allows the number of people and devices connected to the internet to grow continually. Data storage requirements increase as data accumulation from all sources grows as does the number of sources. Distributed system supports two types of architectures namely homogeneous and Heterogeneous. Today all the industries use the homogeneous approach. As the today's need is to use the different data modeling schemas at different databases, this cannot be done in the homogeneous system. This problem can be easily solves by using the heterogeneous distributed system.

Evaluating the performance of any organization is an essential part for overcoming their weaknesses. The main purpose of this paper is how different data mining techniques can extract respectable knowledge from the larger database and analyze user behavior to improve the business performance of an organization. Heterogeneous distributed data mining techniques have become necessary for large and multi-scenario data sets requiring resources, which are heterogeneous and distributed.

Keywords: *Data mining, Heterogeneous database system, Data cube, Homogeneous database, ETL process.*

* PVPIT, University of Pune, India.

1. INTRODUTION

This system deals with the data mining on heterogeneous distributed databases. Generally, **distributed database** is a database that is under the control of a central database management system (DBMS) in which storage devices are not all attached to a common processor. It may be stored in multiple computers located in the same physical location, or may be dispersed over a network of interconnected computers.

A **Heterogeneous Database System** is an automated or semi-automated system for the integration of heterogeneous, disparate database management systems to present a user with a single, unified query interface. For example, different data modeling schemas such as Relational, Object-Oriented, Object-Relational etc.

There are many difficulties in distributed heterogeneous databases in retrieving or storing the information. This system will solve the problem of dealing with different schemas simultaneously performing the operations on data in distributed environment. It uses the mapping algorithm to deal with data on database of various schemas. As in the large organization the data is stored in the huge amount so it requires the data warehouse. It is a process of centralized data management and retrieval. The huge amount of data is difficult to handle and gives less response time to carry out the different operations. So the concept of data mining is essential. **Data mining** is 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 warehouse.

It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

2. SYSTEM ARCHITECTURE

The functionalities of different subsystem are explained below:

2.1. Heterogeneous Distributed Databases

This architecture consists of two sites where datasets are stored heterogeneously in two different schemas of Relational and Object Relational. The data at two sites are distributed by horizontal fragmentation fashion. In this system half data records are stored on one site whereas rest half on another site.

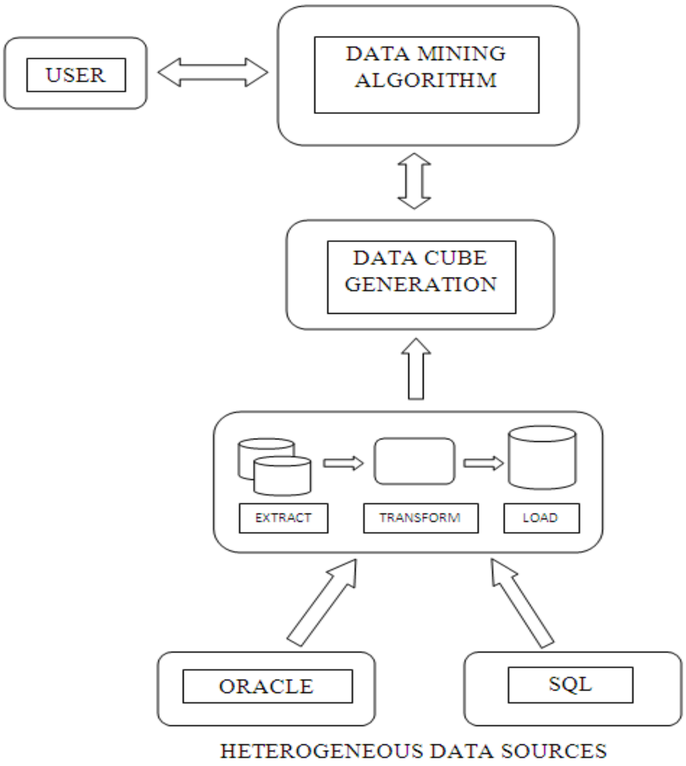

Fig. no.1 Architecture of a system

2.2. ETL process

Extract, transform, and load (**ETL**) is a process in database usage that involves:

- Extracting data from heterogeneously distributed sources
- Transforming it to fit operational needs (which can include quality levels)
- Loading it into the end target.

In ETL process data preprocessing and data cleaning will be carried out.

These both processes are used to clean the data which is incomplete or missing values, a kind of noisy data. These all are removed and quality level data is passed to data cube.

2.3. Data Cube

A Data Cube is a data structure that allows fast analysis of data. It is a multi-dimensional data model which views data in the form of a cube. It can also be defined as the capability of manipulating and analyzing data from multiple data sources. At the very start, the mapped data is fetched. In this step, we execute the cube query. The data obtained is classified as per the required facts and dimensions. The attributes of the data is considered as the dimensions and the operations performed on these are termed as facts.

2.4. Data Mining

Data Mining (sometimes called data or knowledge discovery) is a non-trivial process of identifying valid, novel, useful and ultimately understandable patterns in data. In this the data mining algorithm are applied on the data which is generated in data cube and finally the results generated are displayed to the user.

3. CONCLUSION:

The goal of this system is to use fragmented databases and data cubes in order to extract data by applying data mining. So the system is able to detect future prediction with the help of efficient data mining.

4. ACKNOWLEDGEMENT

We would like to express our deep gratitude to director of ICCIEMI for approving the project research work. We would also like to thank our supervisor Prof. S.U.Kadam whose help, simulating suggestions and encouragement helped us in all the time of research work.

Especially, we would like to give our special thanks to our parents for their encouragement and support.

5. REFERENCES

1) Research on Distributed Data Mining Technology Based on Grid Ning Zhang, Hong Bao,
2) Distributed multi-relational data mining based on genetic algorithm, Wenxiang Dou Jiznglu Hu Hirasawa, K. Gengfeng WuGrad, 808-0135, Japan.
3) Distributed/Heterogeneous Query Processing in Microsoft SQL Server, José A. Blakeley Conor Cunningham Nigel Ellis Balaji Rathakrishnan Ming- Chuan Wu.
4) Data mining in telecommunications, Gary M. Weiss, Department of Computer and Information Science Fordham University.
5) Efficient implementation of data mining : improve customer's behavior, Abbullah –Al-Mudimigh , Farrukh Salim , Zahid Ullah,
6) Query Optimization in multidatabase system: Solutions and open Issues, Tadeusz Morzy,Zbyszko Krolikowski, Poznan University of Technology.
7) The Future of heterogeneous databases, Witold Litwin, INRIA 78153, Le Chesnay, France.
8) An Application of Data mining to identify Data quality problems, Esherf Januzaj, Visar Januzaj Technische universitat Darmstadt Formal Methods in System Engineering Darmstadt, Germany.