RESEARCH ARTICLE OPEN ACCESS

A Study on Some of Data Warehouses and Data Mining (Case Study of Data Mining for Environmental Problems)

Pooja Shrivastava[1], Dr. Manoj Shukla[2]

[1]Ph.D. Research Scholar,
Computer Science and Engineering,
Jayoti Vidyapeeth Women's University, Jaipur-India
[2]Associate Professor, Computer Science and Engineering
Sunder Deep Group of Institution, Ghaziabad U.P- India

ABSTRACT
We live in a world where vast amounts of data are collected daily. Data warehouses and Data mining given the important concepts analyzing such data. In this section we will gain a multidimensional view of data mining and data warehousing. In this paper we introduce a Data mining and Data warehousing some concepts and present a some few methods about this area and discuss case study of data mining for Environmental problems. In research field data mining has made significant progress and Data mining used in a vast array of areas include Biological and Environmental. Only a fewer research have focused on this area but analyzing such data is an important need. So we present how data mining can meet and gives an introduction to data warehouses.

Keyword's - Data warehouses concepts, Data mining concepts, methods, and spatial data mining.

I. INTRODUCTION

In recent year data warehouses and data mining techniques used in any industry and functional area. The construction of data include data cleaning, data integration, and data transformation and we can viewed as an important tool and step for data mining. Data mining is the (KDD) knowledge discovery for data base. Its concept used for discover for knowledge from data and any historical data also. Data warehouses given OLAP tool for analysis data and its effective data mining. Now days data mining used any functional area like biological, environmental any industries, hospital, management. In computing, a **data warehouse** (**DW**, **DWH**), or an **enterprise data warehouse** (**EDW**), is a database used for reporting (1) and data analyzing (2). Integrating data from one or more sources creates a central repository of data, a data warehouse (DW). Data warehouses collect current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. Data warehouse explore on data storage. The main source of the data is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support (Markkas & O'Brien 2009). However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an

expanded definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata. Data mart present small data ware houses system. Data warehouses maintain data history, improve data quality, making decision–support queries easier to write, and provide a single common data model.

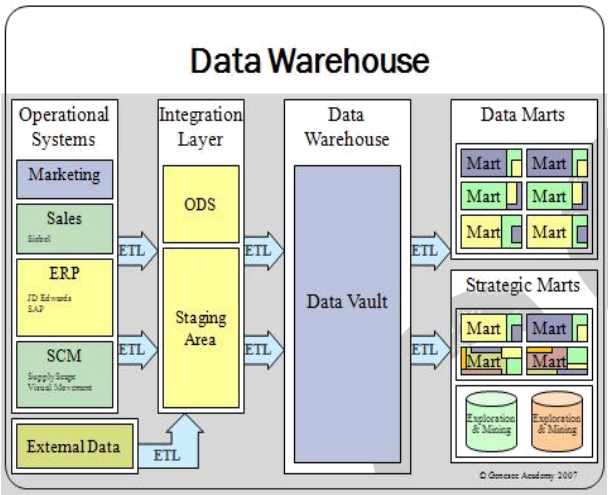

Fig: 1- processing of data warehouses

Above figure we can see that how data ware houses can be proceed. Data warehouses and data mining is related to each other. Data mining can be viewed as a result of the natural evolution of information technology. Many types of tools are provided WEKA, MATLAB, etc. Data mining used in biological and environmental functional area. Its used in identification and classification genome sequence and protein analysis, provided non negative matrix tool box and cancer prediction also. Its useful for predict the spatial data mining also. In earlier days we can see that many type of environmental problems occurred so data mining useful in this area also. In below figure we see how data mining work in basic way.

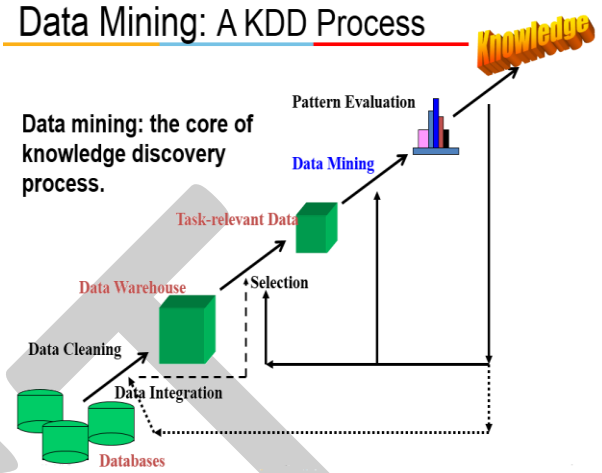

Fig: 2- Data Mining process

In this figure explore the how data mining proceed and how relate to data ware houses. In data mining many types of concepts, algorithms and tools are found that. In this paper we present that such type of concepts, tools and algorithms. Data mining is so important to help companies, biological area and environmental area focus on the most important information in the data they can collect information about the behavior of their customers form big databases and reveal it effectively. Data warehousing provides tools for business executives to system apically organize, understand and use their data to make decision. There are many types' data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and

Comprehensive any functional area and any industry are involve. Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods.

II. CONCEPTS AND TECHNIQUES OF DATAWAREHOUSES AND DATA MINING

Many researcher given lot of point of view for data warehouses and data mining. In data warehouses many type of concept is found that. In source system and Execution Systems CRM, ERP, Legacy, e-Commerce is include. Sample Technologies is very fast of industry and with the help of these technology many types of problems is solved. The **Sample Technologies** are PeopleSoft, SAP, Siebel Oracle, Applications, Manugistics, Custom Systems and these technology is very useful in research field.

ETL Tools: Informatics Power art **,**ETI**,**Oracle Warehouse, Builder**,** Custom programs**,** SQL scripts**,** Oracle**,** SQL Server**,** Tera data **,** DB2 tis is the ETL tool it is also very use full in this field.

In this types of tools provided fast and clean data set. OLAP tools are based on a multidimensional data model. OLAP operation shown that: star schema, snowflake schema, and fact constellation. Dimensional data modeling provide **E-R model** and E-R model divided(1) Symmetric E-R model (2) Divides data into many entities and many types paradigm (3)Describes entities and relationships and Seeks to eliminate data redundancy (4) Good for high transaction performance. And second is **Dimensional model** and its divide (1) Asymmetric dimensional model (2) Divides data into dimensions and facts (3) Describes dimensions and measures (4) Encourages data redundancy (5) Good for high query performance.

2.1 Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models:

It's the most popular data model for data warehouses in **Star Schema** two table contain (1) Fact table (2) dimension table

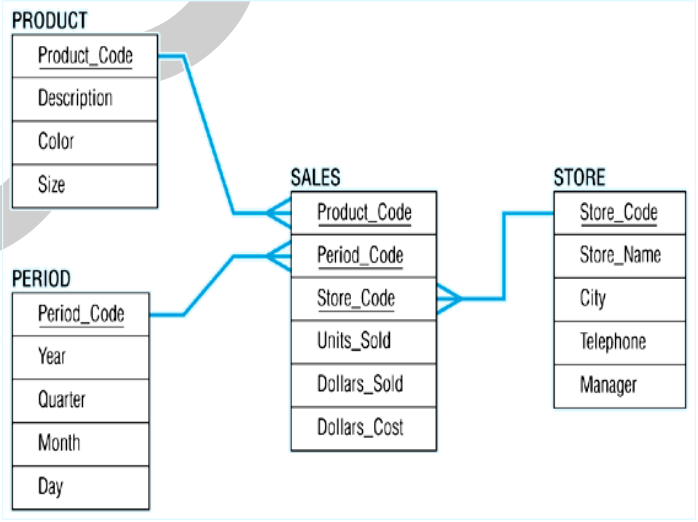

Fig 3:- Star schema of data warehouses

In this figure we can see that processing of star schema. Star schema, each dimension is represented by only one table, and each table contains a set of attributes. Second is Snowflake

schema the snow flake schema is a most variant of the star schema model. The difference between stare flake and snowflake is that the dimension table. Snowflake is reduce the redundancies and maintain the table in easy form. A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.In below figure we can see that how snowflake schema can be proceed in data warehouses.

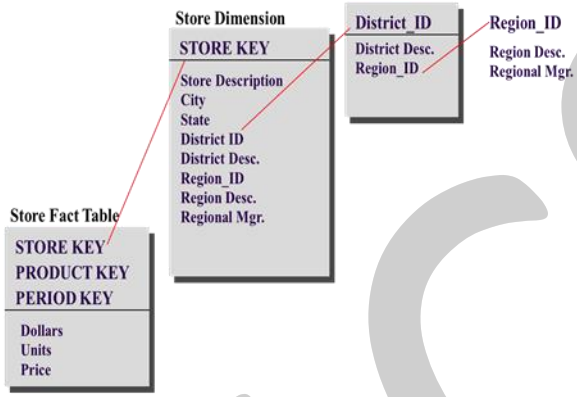

Fig 4: Snowflake schema of data warehouses

Fact constellation is sophisticated application in this tables fact tables to share dimension tables. Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation.

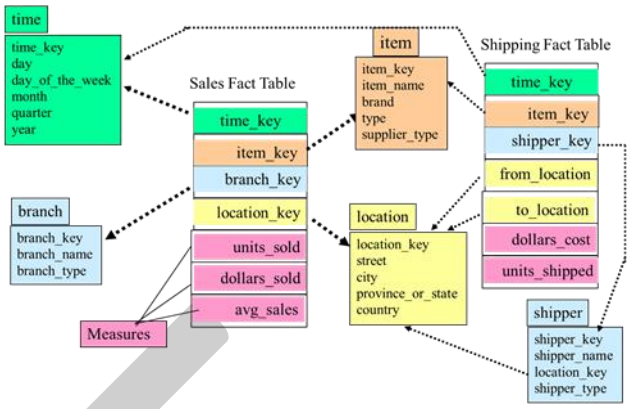

Fig 5:-fact constellation schema of data ware house

2.2 OLAP Operations:

OLAP provides a user friendly environment for data analysis. Ist of all **Roll up (drill-up)** summarize data by climbing up hierarchy or by dimension reduction its reduced the data. Roll-up may be performed by removing say the time dimension. **Drill down (roll down)** reverse of roll-up from higher level summary to lower level summary or detailed data, or introducing new dimensions. Its navigates from less detailed data to more detailed data. **Slice and dice:** project and select the slice operation performs a selection on one dimension of the given cube, resulting in a sub cube**. Pivot (rotate):** reorient the cube, visualization, 3D to series of 2D planes, it is also called rotates

Other operations:

Drill across: involving (across) more than one fact table

Drill through: it s through the bottom level of the cube to its back-end relational tables with the help of SQL.

2.3 Data mart:

A Data mart contains a subset of corporate-wide data that is of value to a specific group of users. Data mart implemented on low-cost departmental servers that are windows based UNIX and LINUX also. In figure we can see that dependent data mart.

Fig6:- Independent Data mart

2.4 Clustering:

Cluster a collection of data objects. And these objects similar to one another with in the same cluster dissimilar to the objects in other clusters. With the help of cluster we finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Unsupervised learning and no predefined classes in clusters. Researcher worked on in this field. It is used in UCI repository iris data sets and created the simple work on iris data sets. And many research clustering is also used in biological problem DNA/RNA analysis. It have rich application for example

Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

Land use: with the help of land we can identification of areas of similar in an earth observation database.

Insurance: it is identifying groups of motor insurance policy holders with a high average claim cost etc.

City-planning: it is identifying groups of houses according to their house type, value, and geographical location. Etc.

Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults.

Basic concepts of clustering:

Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters, s.t., and min sum of squared distance. It is the simplest and most fundamental version of clustering. Given the database with the help of cluster find the dissimilarity function based on distance. Cluster are similar to one another and dissimilar to objects in other cluster in terms of the data set attributes. Given a k, find a partition of k clusters that optimizes the chosen partitioning criterion. Global optimal: exhaustively enumerate

all partitions. Heuristic methods: k-means and k-medoids algorithms**.**

Hierarchical Method: Use distance matrix as clustering criteria. This method does not require the number of clusters *k* as an input, but needs a termination condition. A hierarchical clustering method works by grouping data objects into a hierarchy or tree of cluster. In this method many types of application algorithms are used ROCK algorithm, two phase algorithms etc and many types of method are also found that Density-Based Methods ,Grid-Based Methods, Model-Based Methods it is clustering methods.

2.5 Classification:

Classification consists of assigning a class label to a set of unclassified cases. It is advanced techniques for data classification. Classification are two types **Supervised Classification** The set of possible classes is known in advance. **Unsupervised Classification** Set of possible classes is not known. After classification we can try to assign a name to that class. Unsupervised classification is called clustering. In **Bayesian Classifier** the attributes are independent given the class. Called "Naïve" classifier because of these assumptions. Empirically proven to be useful. Scales very well. **Robust Bayesian Classifier** makes no assumption about the nature of the data. It provides probability intervals that contain estimates learned from all possible completions of the database. Classification Techniques are very wide but in this paper we gives only idea about the concept**. Regression, Distance, Decision Trees Rules, Neural** Networks it is the classification methods.

III. CASE STUDY OF DATA MINING FOR ENVIRONMENTAL STUDIES

Now days in research field environental studies is very important. Environmental studies as a subject has a wide scope in every field. Data mining consists of evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data, spatial data, forest fire, landslide etc. Some researcher focused on this area because its very challenging task. We present case study of this matter and how researcher provided ideas in this area. **J-S. Lai ,F.[4]** Tsai focused on validation and risk assessment of landslide. In this research used some tools and methods of data mining. In this research induced by heavy torrential rains in the Shimen reservoir watershed of Taiwan using spatial analysis and data mining algorithms. In this research we can see that, the spatial analysis and data mining algorithms combining the mechanism of filtering uncertainty data can perform verification and forecast of landslides with more reliable and most accuracy results in the study site.


The most common hazard in forests is forests fire. Forests fires are as old as the forests themselves. They pose a threat not only to the forest wealth but also to the entire regime to fauna and flora seriously disturbing the bio-diversity and the ecology and environment of a region. During summer, when there is no rain for months, the forests become littered with dry senescent leaves and twinges, which could burst into flames ignited by the slightest spark. The Himalayan forests, particularly, Garhwal Himalayas have been burning regularly during the last few summers, with colossal loss of vegetation cover of that region. **Paulo Cortez and Anibal [5]Morais** worked on Predict Forest Fires using Meteorological Data and in this work used data mining concepts such as Support Vector Machines (SVM) and Random Forests, and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes), were tested on recent real-world data collected from the northeast region of Portugal. And after some time researcher give the valuable feedback from this work.**Yong Poh Yu, Rosli Omar, Rhett D. Harrison, Mohan Kumar Sammathuria and Abdul Rahim Nik[6] given** some idea related to this area. In this work two hybrid approaches to investigate the nonlinear relationship between size of a forest fire and meteorological variables (temperature, relative humidity, wind speed and rainfall) and its key word is Forest fire, self-organizing map, back-propagation neural network, rule-based system. The experimental result is accuracy of burnt area prediction. Many researcher gives the different ideas and different approaches on this area. Its difficult task but more popular concept used in this area. Many algorithms used in spatial data mining. Generalized Density-Based Clustering is used for spatial data mining. Data mining techniques are used in every field green media is most popular field for the data mining. like television, radio, newspaper, magazines, hoardings, advertisements etc.

IV. CONCLUSION

Data mining and data warehouses is the very big area in the world. Now days many research is found up in this field. In this paper we present only some few concepts and case study of environmental problems. How data mining support any type of environmental problem we can see that in this paper. Data mining has seen great successes in many application and we discuss highly successful and popular application example of data mining and data warehousing.

V. REFERENCE

1-Data mining concepts and techniques(Jiawei han , Micheline Kamber,Jian Pei).

2 -Shreyas Sen, Seetharam Narasimhan, and Amit Konar(Engineering Letters, 14:2, EL_14_2_8 (Advance online publication: 16 May 2007).

3- www.google.com as url link.

4- J-S. Lai , F. Tsai(International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B2, 2012 XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia).

5- Paulo Cortez and An´ıbal Morais.

6- Yong Poh Yu, Rosli Omar, Rhett D. Harrison, Mohan Kumar Sammathuria and Abdul Rahim Nik. Journal of Computational Biology and Bioinformatics Research Vol. 3(4), pp. 47-52, July 2011 Available online http://www.academicjournals.org/jcbbr ISSN-2141-2227 ©2011 Academic Journals.

7 - Chandrasekhar Jakkampudi ,MR BRAHMAM.