Research Article **Impact Factor: 0.621** **ISSN: 2319-507X**
Aparna Kalaskar, IJPRET, 2014; Volume 2 (9): 186-193 **IJPRET**

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

GOLDEN DATA MINING AND WAREHOUSING

APARNA S. KALASKAR[1], DR. H. R. DESHMUKH[2], Z. I. KHAN[3], S. S. THANGAN[3]

1. Dept. of COE, IBSS College of Engineering, Amravati, Maharashtra, India.
2. Prof. & Head, IBSS College of Engineering, Amravati, Maharashtra, India.
3. Assistant Professor, IBSS College of Engineering, Amravati, Maharashtra, India.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: We are in an age often referred to as the information age. In this information age, Because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became Overwhelming. This initial chaos has led to the creation of structured databases and Database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. For that data mining concept has arise. Data mining *is* the extraction of hidden predictive information from large databases. A data warehouse as a storehouse, is a repository of data Collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof.

Keywords: Database Management System.

Corresponding Author: MS. APARNA S. KALASKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Aparna Kalaskar, IJPRET, 2014; Volume 2 (9): 186-193

PAPER-QR CODE

INTRODUCTION

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment can include an extraction, transportation, transformation, and loading (ETL) solution, online analytical processing (OLAP) and data mining capabilities, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

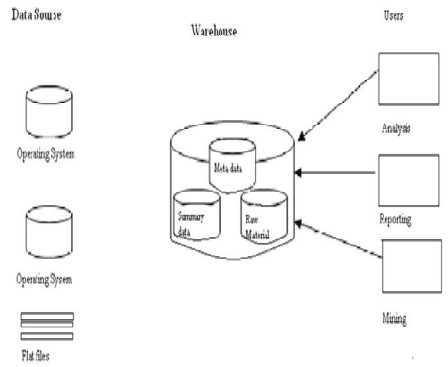
Data Mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

What is Data Warehousing:-

Data Warehousing is defined in the following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process".

The data warehouse is an environment not a product. It is an architectural construct of an information system that provides users with current and historical decision support information that is hard to access and present in traditional operational data stores. inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Data Warehouse Architecture (Basic):-



The components of Data warehousing are:-

1. Metadata management:- Metadata is data about data that describe the data warehouse. It is used for building, maintaining and using the data warehouse. Metadata can be classified into

Technical metadata includes

1. Information about data sources

2. Trasformation description

3. Warehouse object and data structure definitions for data targets .

4. The rules used to perform data cleanup and data enhancement.

Business metadata includes

Subject areas and information object type, including queries, reports, images, video and audio clips.

2. OLAP(On-line Analytical Processing) tools :

These tools are based on concept of multidimensional databases and allow a sophisticated user to analyze the data using elaborate, multidimensional , complex views. Typical business applications for these tools include product performance and profitability , effectiveness of sales program, sales forecasting and capacity planning.

3. Data mining:-

Data mining is the process of discovering meaningful new correlations patterns and trends by digging into large amounts of data stored in warehouses , using artificial intelligence and statistical and mathematical techniques. In these areas , data mining can reach beyond capabilities of the OLAP , since the major attraction of data mining is its ability to build predictive rather than retrospective models.

4. Data visualization:-

Data visualization goes far beyond simple bar and pie charts. It is collection of complex techniques that currently represent an area of intense research and development focusing on determining how to best display complex relationship and patterns on a two dimensional computer monitor.

5. Data Marts:-

The concept of the data mart is causing a lot of excitement and attracts much attention in the data warehouse industry. Mostly data marts are presented as an inexpensive alternative to a data warehouse that takes significantly less time and money to build.

6. Query tools:-

This category can be further divided into two groups

1. Reporting tools
2. Managed query tools

Features and Benefits:-

1) Complete data warehousing application tailored to your business
2) Platform independent architecture for seamless integration with your existing infrastructure Facilities to integrate data from a wide variety of sources.
3) Virtually unlimited scalability.
4) Options to update/synchronize data from multiple sources via automatic schedulers or on demand.
5) Completely Web-enabled system architecture to provide ultimate enterprise functionality for all company locations around the world (e.g., access via Web browsers from any location).
6) Advanced security model and authentication of users.
7) Complete document management options to optimize management of documents of any types and satisfy regulatory requirements.
8) Comprehensive OLAP functionality for online data exploration and reporting.
9) Advanced analytic components to clean/verify data and to integrate automated data mining, artificial intelligence, and real-time process monitoring.
10) Options to automatically run and post on Knowledge Portals (or broadcast) highly customized reports, including interactive (i.e., drillable, sliceable, and user-customizable) reports and results of advanced analytics
11) Backup and archiving options.

1) Incomplete:-

Missing records:

This means a record that should be in a source system is not there. Usually this is caused by a programmer who diddled with a file and did not clean up completely. You may not spot this type of error unless you have another system or old reports to tie to.

Missing fields:- These are fields that should be there but are not. There is often a mistaken belief thought a source system requires entry of a field.

2) Incorrect:-

Sometimes wrong, Sometimes right codes: - This usually occurs when an old transaction processing system is assigning a code that the transaction processing system users do not care about.

Now if the code is not valid, you are going to catch it. The "gotcha" comes when the code is wrong but it is still a valid code. For example, you may have to extract data from an ancient repair parts ordering system that was programmed in 1968 to assign a product code of 100 to all transactions. Now, however, product code 100 stands for something other than repair parts.

Wrong calculations, agreements:-

This situation refers to when you decide to or have to load data that have already been calculated or aggregated outside the data warehouse environment. You will have to make a judgment call on whether to check the data. You may find it necessary to bring data into the warehouse environment solely to allow you to check the calculation.

Duplicate records:- There usually are two situations to be dealt with. First, there are duplicate records within one system whose data are feeding the warehouse. Second, there is information that is duplicated in multiple systems that feed in the same type of information.

3) Incomprehensibility:-

Multiple fields within one field:-This is the situation where a source system has one field which contains information that the data warehouse will carry in multiple fields.

Unknown codes:- Many times you can figure out what 99% of what codes mean. However, you usually find that there will be a handful of records with unknown codes and usually these records contain huge or minuscule dollar amounts and are several years old.

Spreadsheets and word processing files:-

Often in order to perform the initial load of a data warehouse it is necessary to extract critical data being held in spreadsheet files and/or "merge list" files.

4) Inconsistency:-

Inconsistent use of different codes:-

Much of the data warehousing literature gives the example of one system that uses "M" and "F" and another system that uses "1" or "2" to distinguish gender. May I suggest that you wish that this is the toughest data cleaning problem you will face.

Overlapping codes:- This is a situation where one source system records, say, all its sales to Customer A with three customer numbers and another source system records its sales to customer A with two different customer numbers. Now, the obvious solution is to use one customer number here. The problem is that there is usually some good business reason why there are five customer numbers.

What is Data Mining:-

“Data mining is the extraction of hidden predictive information from large databases.” It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout.

An Architecture for Data Mining:-

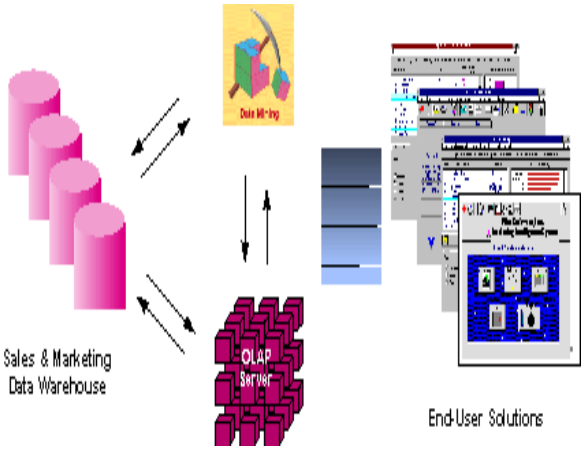


Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the

data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database, for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore.

Data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors.**
- **Automated discovery of previously unknown patterns.**
- **Databases can be larger in both depth and breadth.**

CONCLUSION:

Data warehouse is platform with integrated data of improved qualities. Data warehousing is blend of technologies aimed at effective integration of operational databases into an environment that enables strategic use of data.

Data mining is useful to extract useful information from large database. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

Data warehousing and mining are very powerful tools for storing large amount of information and extracting useful information from it respectively.

REFERENCES:-

1. Data warehousing, Data Mining & OLAP. By Alex Berson & Stephen J. Smith.

2. Introduction to Data Mining-José Hernández-Orallo

3. Data warehousing in the real world. By Sam Anahory & Dennis Murray.

4. Data Mining on the Web: There's Gold in that Mountain of Data by Dan Greening (Web Techniques Magazine)**.**

5. Cipolla,EmilT. Data Mining: Techniques to Gain Insight into Your Data Enterprise Systems Journal.