ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Data Mining and Data Warehousing

Hitesh Naidu, Aishwarya Tiwari
Computer Science Department, Mahamaya Technical University, Noida, Uttar Pradesh-201309, India

Abstracts

Today in organizations, the developments in the transaction processing technology requires that, amount and rate of data capture should match the speed of processing of the data into information which can be utilized for decision making. A data warehouse is a subject- oriented, integrated, time-variant and non-volatile collection of data that is required for decision making process. Data mining involves the use of various data analysis tools to discover new facts, valid patterns and relationships in large data sets. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. OLTP is customer-oriented and is used for transaction and query processing by clerks, clients and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives and analysts. Data warehousing and OLAP have emerged as leading technologies that facilitate data storage, organization and then, significant retrieval. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications.

Keywords: Data mining, Data warehousing .

Introduction

Data warehousing

In computing, a data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more disparate sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. Large amount of operational data are routinely collected and stored away in the archives of many organizations. To take a simple example, the railway reservation system has been operational for over a decade and large amount of data is generated each day on train bookings. Much of this data is probably archived for audit purposes. This archived operational data can be effectively used for tactical strategic management of the railways. Data warehouse is a storage area for processed and integrated data across different sources which will be both operational data and external data. Data warehouses offer organizations the ability to gather and store enterprise information in a single conceptual enterprise repository. It allows its users to extract required data for business analysis and strategic decision making. One can also define a warehouse as a copy of transaction data specifically structured for query and analysis. It is a repository of information, integrated from several operational databases. Data warehouses store large amount of data which can be frequently used by decision support system. It is maintained separately from the organizations operational database. They are relatively static with only infrequent updates. The most effective advantages of data warehousing is high speed of data processing and summarized data.

Data warehouses have several distinguishing characteristics

These systems combine data from multiple sources. Operational systems such as ERP systems provide production data, financial systems supply revenue and expense data, and human resource systems present employee data.

The data copied into a data warehouse does not change (except to correct errors). The data warehouse is a historical record of the state of an organization. The frequent changes of the source OLTP systems are reflected in the data warehouse by adding new data, not by changing existing data.

Data warehouses are subject oriented, that is, they focus on measuring entities, such as sales, inventory, and

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

quality. OLTP systems, by contrast, are function oriented and focus on operations such as order fulfillment.

A data warehouse is always a physically separate store of data, which is transformed from the application data found in the appropriate environment. Due to this separation, data warehouses do not require transaction processing, recovery, concurrency control, etc. The data is not updated or changed in any way once they enter the data warehouse, but are only loaded, refreshed and accessed for queries.

Time variant: Data is stored in data warehouse to provide a historical perspective. Every key structure in the data warehouse contains, implicitly or explicitly, an element of time. The data warehouse contains a place for sorting data that are 5 to 10 years old, or older, to be used for comparisons, trends and forecasting.

Data mining

Data mining an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. It is a process of extracting hidden predictive information from large databases. It is a powerful new technology 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. For a commercial business, the discovery of previously unknown statistical patterns or trends can provide valuable insight into the function and environment of their organization. Data-mining techniques can generally be grouped into two categories: predictive method and descriptive method.

Characteristics of a data mining system

Large quantities of data
The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.

Noisy, incomplete data
Imprecise data is the characteristic of all data collection.

Complex data structure conventional statistical analysis not possible
Heterogeneous data stored in legacy systems

Tools of Data Mining

Traditional Data Mining Tools**.** Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

Dashboards. Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen - often in the form of a chart or table - enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

Text-mining Tools**.** The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text - from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

Techniques of Data Mining:

Regression modeling-This technique applies standard statistics to data to prove or disprove a hypothesis. One example of this is linear regression, in which variables are measured against a standard or target variable path over time. A second example is logistic regression, where the probability of an event is predicted based on known values in correlation with the occurrence of prior similar events.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Visualization-This technique builds multidimensional graphs to allow a data analyst to decipher trends, patterns, or relationships.

Correlation-This technique identifies relationships between two or more variables in a data group.

Variance analysis-This is a statistical technique to identify differences in mean values between a target or known variable and nondependent variables or variable groups.

Discriminate analysis-This is a classification technique used to identify or "discriminate" the factors leading to membership within a grouping.

Forecasting-Forecasting techniques predict variable outcomes based on the known outcomes of past events.

Cluster analysis-This technique reduces data instances to cluster groupings and then analyzes the attributes displayed by each group.

Decision trees-Decision trees separate data based on sets of rules that can be described in "if-then-else" language. Neural networks-Neural networks are data models that are meant to simulate cognitive functions. These techniques "learn" with each iteration through the data, allowing for greater flexibility in the discovery of patterns and trends.

What's the difference between data mining and data arehousing?

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 - in fraud detection, as an aid in marketing campaigns, and even supermarkets use it to study their consumers.
Data warehousing can be said to be the process of centralizing or aggregating data from multiple sources into one common repository.

Conclusion

Organizations today are under tremendous pressure to compete in an environment of tight deadlines and reduced profits. Business processes that require data to be extracted and manipulated prior to use will no longer be acceptable. Instead, enterprises need rapid decision support based on the analysis and forecasting of predictive behavior. Data- warehousing and data-mining techniques provide this capability.

References

1. Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, R., and Sarawagi. S. On the computation of multidimensional aggregates. Proc. VLDB, 1996.
2. Choi, W., Kwon, D., and Lee, S. Spatio-temporal data warehouses using an adaptive cell-based approach. DKE, 59, 1 (Oct. 2006), 189-207.
3. eCourier.co.uk dataset, http://api.ecourier.co.uk/. (URL valid on June 20, 2009).
4. Giannotti, F., Nanni, M., Pinelli, F., and Pedreschi, D. Trajectory pattern mining. Proc. KDD, 2007.
5. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. Data cube: A relational aggregation operator generalizing groub-by, cross- tab and sub-totals. DMKD, 1, 1 (Mar. 1997), 29-53.