International Journal of Business
Management & Research (IJBMR)
ISSN(P): 2249-6920; ISSN(E): 2249-8036
Special Edition, Jan 2015, 21-26
© TJPRC Pvt. Ltd..

DATA MINING AND DATA WAREHOUSING OF BIG DATA

S. POORNIMA[1] & B. DEEPHA[2]

[1]Associate Professor, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India

[2] Research Scholars, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India

ABSTRACT

Big data is everywhere today from lighting houses t finding soul mates online. Big data is the term used to describe the exponential growth and availability of data both structured and unstructured. Big data is important to businesses and the society as internet has become a key player for everything. Big data leads to availability of more data and more data means more accurate analyses. More accurate analyses may lead to more confident decision making. And better decisions can mean greater operational efficiencies, cost reductions and reduced risk. To gather this Data Mining becomes inevitable which is designed to explore data and also in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Once the data is gathered Data Warehousing comes into play where in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. This study focuses on the advantages of big data, and its implications in business. It also throws light on the steps involved in data warehousing and data mining its advantages and its implications.

KEYWORDS: Data Mining, Data Warehousing

INTRODUCTION

Need for Study

The sheer volume of data generated, stored, and mined for insights has become economically relevant to businesses, government, and consumers. Data are now woven into every sector and function in the global economy, and, like other essential factors of production such as hard assets and human capital, much of modern economic activity simply could not take place without them. The use of Big Data — large pools of data that can be brought together and analyzed to discern patterns and make better decisions — will become the basis of competition and growth for individual firms, enhancing productivity and creating significant value for the world economy by reducing waste and increasing the quality of products and services. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data warehouse helps increase the effectiveness and efficiency of a business operation and equip managers with information to make intelligent, informed decisions that will translate into a competitive advantage. With such importance to the Big Data, Data Mining and Data Warehousing this study will help understand its role and effectiveness in business.

Objectives of Study

- Advantages of big data,

- Application of Big Data in business.
- Steps involved in Data Warehousing and Data Mining
- Advantages and its implications.

INTRODUCTION

Big Data

Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions. While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes tools and processes) that an organization requires to handle the large amounts of data and storage facilities. The term big data is believed to have originated with Web search companies who needed to query very large distributed aggregations of loosely-structured data.

Data Mining

Data mining of big data involves going through big data sets for relevant or pertinent information. Businesses collect massive sets of data that may be homogeneous or automatically collected but decision-makers need access to smaller, more specific pieces of data from those large sets. They use data mining to uncover the pieces of information that will inform leadership and help chart the course for a business. Data mining can involve the use of different kinds of software packages such as analytics tools. It can be automated, or it can be largely labor-intensive, where individual workers send specific queries for information to an archive or database. Generally, data mining refers to operations that involve relatively sophisticated search operations that return targeted and specific results. For example, a data mining tool may look through dozens of years of accounting information to find a specific column of expenses or accounts receivable for a specific operating year.

The following is the process in Data Mining:

- Selection
- Pre-processing
- Transformation
- Data Mining
- Interpretation/Evaluation.

Pre-Processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned.

Data cleaning removes the observations containing noise and those with missing data.

Data Mining

Data mining involves six common classes of tasks:

- **Anomaly Detection (Outlier/Change/Deviation Detection)**: The identification of unusual data records, that might be interesting or data errors that require further investigation.
- **Association Rule Learning (Dependency Modeling)**: Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- **Clustering**: Is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Classification**: Is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- **Regression**: Attempts to find a function which models the data with the least error.
- **Summarization**: Providing a more compact representation of the data set, including visualization and report generation.

Evaluation

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called overfitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had *not* been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be used to evaluate the algorithm, such as ROC curves. If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

Data Warehousing of Big Data

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 it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users. Data warehouses can be classified further as viz:

- Subject Oriented
- Integrated
- Non Volatile
- Time Variant

Subject Oriented

Data warehouses are designed to help analyze data. For example, to learn more about company's sales data, you can build a warehouse that concentrates on sales. This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

Integrated

Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

Nonvolatile

Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

Time Variant

In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to **online transaction processing (OLTP)** systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

Data Warehouses has the Following Features

- **Workload**

Data warehouses are designed to accommodate *ad hoc* queries. It is hard to know the workload of your data warehouse in advance, so a data warehouse should be optimized to perform well for a wide variety of possible query operations.

- **Data Modifications**

A data warehouse is updated on a regular basis by the ETL process (run nightly or weekly) using bulk data modification techniques. The end users of a data warehouse do not directly update the data warehouse.

- **Schema Design**

Data warehouses often use denormalized or partially denormalized schemas (such as a star schema) to optimize query performance.

- **Typical Operations**

A typical data warehouse query scans thousands or millions of rows. For example, "Find the total sales for all customers last month."

- **Historical Data**

Data warehouses usually store many months or years of data. This is to support historical analysis.

Data Warehouse Architectures

Data warehouses and their architectures vary depending upon the specifics of an organization's situation. Three common architectures are:

- Data Warehouse Architecture (Basic)
- Data Warehouse Architecture (with a Staging Area)
- Data Warehouse Architecture (with a Staging Area and Data Marts)

FORECAST

It is being forecasted that as a Data Warehouse becomes a mature part of an organization, it is likely that it will become as “anonymous” as any other part of the IS. But one challenge to face is coming up with a workable set of rules that ensure privacy as well as facilitating the use of large data sets. There will also arise a need to store unstructured data such as multimedia, maps and sounds. The growth of the Internet allows integration of external data into a Data Warehouse, but its varying quality is likely to lead to the evolution of third-party intermediaries whose purpose is to rate data quality. The quality of data from mining needs to be improved which can break or make the data mining efforts. It is also predicted that companies first have to integrate, transform and cleanse the data mined to obtain value from the process.

CONCLUSIONS

It is thus being aptly concluded from this study that both data mining and data warehousing procedures play an indispensible role in the smooth functioning of the businesses in order to make the right choices and decisions. Big Data being an important tool today it is important to understand the challenges in gathering and storing the data and overcome them to reap the best benefits of the same.