ABSTRACT

With the fast growth of networks now-a-days organizations has filled with the collection of millions of data with large number of combinations. This big data challenges over business problems. It requires more analysis for the high-performance process. The new methods of hadoop and MapReduce methods are discussed from the data mining perspective.

Keywords: Data Mining, Big data, BI, Big Data analytics, OLAP, EDA, Neural Networks, Hadoop and MapReduce technique, Advantages, Disadvantages.

1. INTRODUCTION

Data mining [4, 5] is designed to inquire the data by concerning methodical relationships among the variables and then by applying the detected patterns to the new subset of data, then findings can be validated. The mechanism of data mining is to apply regular algorithms to the large sets of data to extract the information from it. This large set of data that is related to business or market is known as “Big Data [1]”. The theme of business intelligence (BI) is to exchange the data that tends increased value to the enterprise. Rather than collecting the information on what organizations are really doing, it is better to understand how organizations view big data and to what extent they are currently using it to benefit their business. Now organizations are begun to understand and explore how to process and analyze this big data. Big data and data mining are spreading not only in BI but also in science fields such as meteorology, petroleum exploration and bio-informatics. This sequence of data needs support of software, hardware and sophisticated algorithms for the new level of analysis.

2. TRADITIONAL DATA MINING TECHNIQUES

In general data mining is categorized into three phase process (1) the elementary exploration, (2) Identifying the pattern with validation/verification and (3) deployment. (In this process the pattern is applied to the new model to generate predictions).

2.1 On-Line Analytic Processing (OLAP)
This is also known as fast analysis of shared multidimensional information – FASM. It allows the end users of multidimensional databases to generate the summaries accessible by the computer which contains the information of data and other systematic queries. Even though the name existed as on-line there is no compulsory of analyses of OLAP in real-time only. It just refers to analyze the multidimensional databases. OLAP [10] facilities can be unified into enterprise-wide database systems that allow analysts and managers to guide the performance of the business aspects like manufacturing process and various kinds of completed transactions at different locations. In this sense Data Mining can be a kind of analytic development of OLAP.

2.2 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis involves the both general approaches ad specific techniques. There is a need to understand the purpose of Data Mining and traditional Exploratory Data Analysis. Data Mining is more focused towards the general nature of the elemental phenomena and it concerns less about identifying the specific relations among the variables. Instead of it is focus to produce a solution that can generate useful predictions. Thus Data Mining [11] is a approach to explore the data which uses not only the traditional Exploratory Data analysis (EDA) techniques but also the Neural Networks that can generate valid predictions but are unable to identify the relations between the variables on which the predictions are based.

2.3 Neural Networks

Neural Networks [12] is one of the Data Mining techniques. Neural Networks are analytical techniques which are inspired by the biological nervous system such as brain to process the information. In this system based on the functions of brain it can predict new observations from other observations (i.e., from one variable to another variable) by using a process called learning from existed data.

The network architecture is designed such as it includes a definite number of layers each consisting of certain number of neurons. Each stage involves “trail and errors”. After each training, the resulting “network” developed in this learning process describes a pattern detected in the data. Thus this networking functioning similar to the traditional model approach. By using Neural Networks techniques, the analysis of explanatory models can be build. The major advantage of this model is they can approximate any continuous function and thus no need to have compulsory of any hypotheses about the underlying variable. A drawback of this model is that the final solution depends on the initial conditions of the network which is impossible.

3. HADOOP AND MAPREDUCE TECHNIQUE

Big data [2] plays an important role in Business applications. The coverage of big data includes Data acquisition, cleaning, distribution, and best practices. Big Data contains the risk of threat analysis, predicting failures of the network data and trade control. Big data analytics found that Apache Hadoop is preferred as solution to the problems in the traditional Data Mining. It acts as extensible for recovering the failures of the data storage and processing in the distributed system. Apache Hadoop is an Open-source software framework [13] for storing and processing of large data-sets on clusters of hardware. Here the hadoop is designed with the assumptions of hardware failures that are automatically handled by the software framework. The main components of Hadoop are Hadoop distributed file system (HDFS) which is useful for large files and MapReduce which acts as heart of Hadoop. HDFS is high bandwidth clustered storage. MapReduce performs two different tasks in Hadoop programs. the First job is to map, in which it takes a collection of data where it is transformed into another set of data .After transformation the data is broken in to tuples (with key/value pairs).The job of reduce is to take the output from the map job which acts as its input and these data tuples are combined into smaller sets of tuples. In this manner the reduce job is always achieved after the map job.

4. HADOOP WITH OTHER TECHNOLOGIES

4.1 Using Hadoop techniques with parallel Databases [14]

In the earlier days of Hadoop and MapReduce there are several problems. But now the current versions have been used with various data management techniques to reduce the performance gap. In this context the Hadoop...
is studied in the sense of similarities and differences with parallel databases. The parallel databases techniques like job optimization, data layouts and indexes are focused in this discussion. Even though the usage of Hadoop requires a little knowledge of databases background, it can’t compete with the parallel databases with their efficiency of query processing. Many researchers found that it is efficient to use parallel databases with the combination of Hadoop MapReduce. In the previous versions of Hadoop, most of the problems are found in the physical organization of data including data layouts and indexes. In general Hadoop and MapReduce affected from row-oriented layout. Hence other data layout techniques are proposed for Hadoop MapReduce. Similarly Hadoop has deficit of appropriate indexes. A good source of indexing techniques has been proposed for Hadoop MapReduce [15, 16, 17, 18].

4.2 Hadoop and Data Warehouse [19]
As the advertising of Hadoop is unrestrained, the practitioners are easily affected by diversity of opinions like Hadoop is becoming the new data warehouse. But it is not really as it seems. There are a lot of differences between Hadoop and data warehouse. This context explores when to use hadoop and when to switch data warehouse. Let us consider an example of firm uses Hadoop to preprocess raw click streams generated by customers using their website. These click streams are passed to the data warehouse as its processing provides the vision of customers’ preferences. The data warehouse sets these customer preferences with marketing campaigns and recommendation engines to offer investment suggestions and analysis to customers. So the Data Warehouse is used as a source in complex Hadoop jobs. This will brings the advantages of these two systems in parallel. Finally choosing hadoop and data warehouse depends on the requirements of the organization. In most of the cases Hadoop and data warehouse work together as a group in the information supply.

5. HADOOP IN OTHER SCIENCES

5.1 Biological data analysis
As the biological databases grow larger, dealing with increasing information of genomic data is spreading tremendously. In order to handle this big data, hadoop is used mostly in next generation sequencing. High-throughput expression data for genes, proteins, and metabolites can be benefit from Big Data Analysis of Hadoop. Advances in Bioinformatics [21] bring challenges in processing, storing and analyzing the data. This can be achieved by using Hadoop with the combination of Cloudburst algorithm. It is also examining in protein sequence clustering [22] which finds MapReduce algorithm could be useful for handling more complex clusters with a high degree of parallelism.

5.2 Geographical analysis
This category of data includes location and is enhanced with geographic information in a structured form which is nothing but spatial data. This data requires an understanding of geometry and operations that can be performed on it. The GIS tools are used for spatial Framework of Hadoop [23]. It allows users for the analysis of spatial data. Some geographical networks [24] problems like Distributed Denial of Service (DDoS) which are common attempts in security hacking can be analyzed by using Hadoop and MapReduce.

6. HADOOP MAPREDUCE ADVANTAGES
The main advantage of Hadoop MapReduce it allows the users (even though if they are not experts) to easily handle analytical risk over Big data. It gives complete control on processing the input datasets. MapReduce can be easily used by the developers without having much knowledge of databases but with a little knowledge of java is needed. It gives satisfied performance in scaling large clusters.

- It supports distributed data and computation
- The computation is performed local to data and thus it prevents the network overload.
- The tasks are independent hence, it can easily handle partial failures such as when the nodes fail, and it can automatically restart.
- It is a Simple programming model. The end-user programmer only writes MapReduce tasks.
- HDFS stores vast amount of information.
- HDFS is simple and robust coherence model thus it stores data reliably.
- HDFS provide streaming read performance.
- Flat scalability [20]
- It has the ability to process the large amount of data in parallel.
- HDFS has capability for replicating the files which can easily handle situations like software and hardware failure.
- In HDFS the data can be written only once and it can be read for many times.
- It is more economic way as the data and processing are distributed across the clusters of personal computers.
It can be offered as on-demand service, for example as part of Amazon’s EC2 cluster computing service. Ability to write MapReduce programs in Java, a language which even many noncomputer scientists can learn with sufficient capability to meet powerful data-processing needs.

7. DISADVANTAGES OR LIMITATIONS OF HADOOP
These are major common areas where the Hadoop framework is found uncertain.

- As the both the Hadoop HDFS and MapReduce software are under active development, they are found to be uneven.
- Possibility of preventing central data leads to restrictive programming model.
- HDFS is weak in handling small files, and inadequacy of transparent compression. The design of HDFS is such that it doesn’t work with random reads on small files because of its optimization for sustained throughput.
- There is a necessary of managing job flow is when there is intermediate data.
- Managing the cluster is hard in operations like debugging, distributing software, collection logs etc.
- Because of single-master model, it requires more care and may limit scaling.
- Hadoop offers high security model, but because of its complexity it is hard to implement it.
- MapReduce is a batch-based architecture which means it doesn’t allow itself to use cases that needs real-time data access.

CONCLUSION
To keep track of current state of business, advanced analytical technique of big data such as predictive analysis, data mining, statistics and natural language processing are to be examined. New techniques of big data such as Hadoop and MapReduce create alternatives to traditional data warehousing. Traditional Hadoop with combination of new technologies explores a new scope of study in various fields of science and technologies.

REFERENCES