ANALYSIS OF MINING DATA IN DATA WAREHOUSE AND BIG DATA

C.Radha[1], S.Nithyananth[2], M.Menakapriya[3], S.Senthamaraiselvi[4]
Assistant Professor/ MCA,
Muthayammal Engineering College Namakkal, Tamilnadu

Abstract

Data Mining (DM) is known as the process of extracting useful patterns or information from huge amount of data. Data mining task is the automatic or semi-automatic analysis of large quantities of data to extract unknown interesting patterns such as groups of data records (cluster analysis), unusual records and dependencies (association rule mining). The aim of this paper is to show the analysis of data mining in Data Warehouse and Big Data. Data warehouses can provide the information required by the decision makers. The data mining from data warehouse can be a ready and effective system for the decision makers. Big data are the large amount of data being processed by the data mining environment. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This paper shows how to extract data in data warehouse and Big data.

Keywords: Data Mining, Data Warehousing, Big Data, Analysis, Knowledge Discovery

I. INTRODUCTION

Data Mining is the new technology to extract the knowledge from the data. It is used to explore and analyze the same. The data to be mined varies from a small data set to a large data set i.e. big data. The data mining environment produces a large volume of the data. The information retrieved in the data mining step is transformed into the structure that is easily understood by its user. Data mining involves various methods such as frequent pattern tree algorithm, association rule and cluster analysis, to disclose the hidden patterns inside the large data set.

DW is a very important repository especially for the historical data and non-every-day transaction which is useful. A data warehouse is a subject oriented integrated, non-volatile, and time variant collection of data in support of management decisions .Data warehouse obtains the data from a number of operational data base systems which can be based on RDBMS (Relational Database Management System)/ERP (Enterprise Resource Planning) package, etc. Practically, data warehousing and data mining is really useful for any organization which has huge amount of data. Data warehousing and data mining help regular (operational) databases to perform faster. Data Mining (DM) is a combination of Database and Artificial Intelligent used to provide useful information to both technical and non-technical users which will help them to make better decisions. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information-information that can be used to increase revenue, cuts costs, or both.

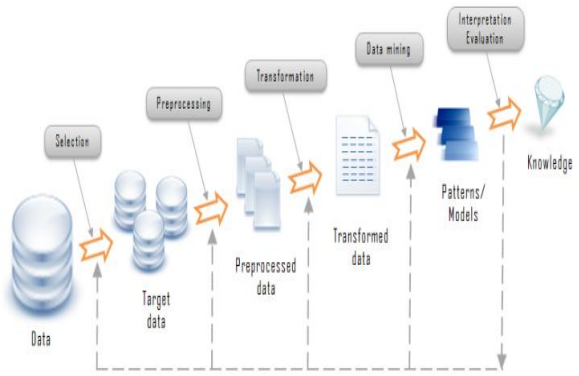

Fig. 1. Knowledge Discovery Process

Now the term Data Mining, Finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining. So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information.DM is a set of methods for data analysis, created with the aim to find out specific dependence, relations and rules related to data and making them out in the new, higher-level quality information.

As distinguished from the data warehouse, which has unique data approach, DM gives results that show relations and interdependence of data. Mentioned dependences are mostly based on various mathematical and statistic relations represents the process of knowledge data discovery. Process of knowledge data discovery based.

The Big Data is nothing but a data , available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. For example, the data stored at the server of Facebook, we upload various types of information, upload photos. All the data get stored at the data warehouses at the server of Facebook. This data is nothing but the big data, which is so called due to its complexity. Also another example is storage of photos at Flicker. These are the good real-time examples of the Big Data. Another best example of Big data would be, the readings taken from an electronic microscope of the universe. It' s huge in nature because, there is the collection of data from various sources together. If we consider the example of Facebook, lots of numbers of people are uploading their data in various types such as text, images or videos. The people also keep their data changing continuously. This tremendous and instantaneously, time to time changing stock of the data is stored in a warehouse.

This large storage of data requires large area for actual implementation. As the size is too large, no one is capable to cont rol it oneself. The Big Data needs to be controlled by dividing it in groups. Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time. For major Big Data -related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets.

II. RELATED WORK

A. *Distributed Data Mining*

The rapid development of computers together with communication networks led to the generation of huge datasets by software applications developed for various services such as supermarket, banking, stock market and mobile service. Distributed Data Mining (DDM) can be implemented using one of the four approaches, mentioned below.

•Move the disperse data to a central site, and then perform data mining on the entire data. This involves heavy communication cost to pool the entire data to a central site. Another problem with this approach is the preservation of data privacy which can't be guaranteed.

•Second approach is to perform mining locally at each site to build a local model. All the local models built so are moved to central site to combine them in to a global model. This approach suffers from scalability problem with increase in number of sites.

•The third approach is using samples. A small set of representative data is carefully sampled at each site to form one global representative data set. Data mining can then be performed on the global representative data set.

•Unlike all the previous three approaches where a central site to facilitate distributed data mining is involved, here in the fourth approach data mining is performed on P2P networks, where sites communicate directly with each other without involving a central server.

Based on the use of underlying network infrastructure, distributed data mining systems can be broadly categorized into Client-Server DDM systems, and Mobile agent-based DDM systems.

B. *Distributed data warehousing*

As in distributed databases, in distributed data warehousing a new phase needs to be added to the design method: the one for designing the distribution, from both the architectural and the physical points of view. During architectural design, general decisions will be taken about which distribution paradigm (P2P, federation, grid) better suits the requirements, how to deploy the DW on the infrastructure, which communication protocols to use, etc. The physical point of view mainly addresses how to fragment the DW and how to allocate fragments on the different sites in order to maximize local references to data and to take advantage of the intrinsic parallelism arising from distribution, thus optimizing the overall performance. Though some approaches to fragmentation of DWs have been tempted, they are mainly aimed at exploiting local parallelism or at designing ad hoc view fragments for a given workload. Indeed, distribution is particularly useful in contexts where new data marts are often added, typically because of company mergers or acquisitions. In this case, the most relevant issue is related to integration of heterogeneous data marts.

C. *Data Mining Tasks*

Data Mining is the new technology to extract the knowledge from the data. Data mining involves various methods such as frequent pattern tree algorithm, association rule and cluster analysis, to disclose the hidden patterns inside the large data set. The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

Exploratory Data Analysis-In the repositories vast amount of information's are available .This data mining task will serve the two purposes

(i).Without the knowledge for what the customer is searching, then

(ii) It analyze the data

Descriptive Modeling-It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

Predictive Modeling-This model permits the value of one variable to be predicted from the known values of other variables.

Discovering Patterns and Rules-This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and many more techniques in the data mining algorithm like (K-Means /K-Medoids). These are called the clustering algorithm.

Retrieval by Content-The primary objective of this task is to find the data sets of frequently used in the for audio/video as well as images It is finding pattern similar to the pattern of interest in the data set.

III. EXISTING SYSTEM

A. *Data Warehousing*

A data warehouse is defined as a "subject-oriented, integrated, time variant, non-volatile collection of data that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. In data warehouses historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be much larger than operational databases. Most queries on data warehouses are ad hoc and are complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Due to the complexity query throughput and response times are more important than transaction throughput. Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions.

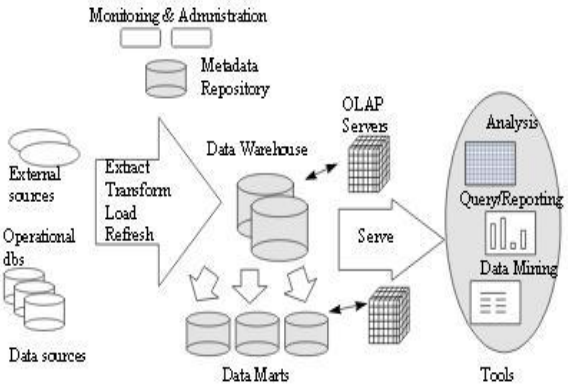

Fig. 2. Data Warehousing Architecture

A Decision Support System (DSS) is any tool used to improve the process of decision making in complex Systems. A DSS can range from a system that answer simple queries and allows a subsequent decision to be made, to a system that employ artificial intelligence and provides related datasets. Amongst the most important application areas of DSS are those complicated systems that directly "answer" questions, in particular high level "what-if" Scenario modeling. Over the last decade there was a transition to decision Support using data warehouses. The data warehouse environment is more controlled and therefore more reliable for decision support than the previous Methods. The data warehouse environment supports the entire decision support requirements by providing high-quality information, made available by accurate and effective cleaning routines and using consistent and valid data transformation rules and documented pre-summarization of data values. It contains one single source of accurate, reliable information that can be used for Analysis. Data Warehouses (DW) integrate data from multiple heterogeneous information sources and transform them into a multidimensional representation for decision support applications.

The designer must also deal with data warehouse administrative processes, which are complex in structure, large in number and hard to code; deadlines must be met for the population of the data warehouse and contingency actions taken in the case of errors. Finally, the evolution phase involves a combination of design and administration tasks: as time passes, the business rules of an organization change, new data are requested by the end users, new sources of information become available, and the data warehouse architecture must evolve to efficiently support the decision-making process within the organization that owns the data warehouse.

Fig.3. Different viewpoints for the metadata repository of a data warehouse

B. *Data Mining*

Data Mining is the extraction or "Mining" of knowledge from a large amount of data or data warehouse. To do this extraction data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge form stored data. Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. The spreadsheet is still the most compiling front-end Application for Online Analytical Processing (OLAP). The challenges in supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet Operation effectively over large multi-gigabytes databases.

Data Mining Life Cycle Phases are,

Business Understanding-This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding-It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation-It collects all the different datasets and constructs the varieties of the activities basing on the initial raw data

Modeling-Various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation-The model is thoroughly evaluated and reviewed. The steps executed to

construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment-The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

c. Data Mining Algorithms

1) The K-Means Algorithm

The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters, k. Gray and Neuh off provide a nice historical background for k-means placed in the larger context of hill-climbing algorithms. The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1, . . . , N}, where xi ∈_d denotes the ith data point. The algorithm is initialized by picking k points in _d as the initial k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence:

- Data Assignment

Each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data.

- Relocation of "means"

Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

The algorithm converges when the assignments (and hence the cj values) no longer change. Each iteration needs N × k comparisons, which determines the time complexity of one iteration. The number of iterations required for convergence varies and may depend on N, but as a first cut, this algorithm can be considered linear in the dataset size. The greedy-descent nature of k-means on a non-convex cost also implies that the convergence is only to a local optimum, and indeed the algorithm is typically quite sensitive to the initial centroid locations. The local minima problem can be countered to some extent by running the algorithm multiple times with different initial centroids, or by doing limited local search about the converged solution. k-means is a limiting case of fitting data by a mixture of k Gaussians with identical, isotropic covariance matrices (_ = σ2I), when the soft assignments of data points to mixture components are hardened to allocate each data point solely to the most likely component.

2) The Apriori Algorithm

Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of item sets, "if an item set is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or item set are sorted in lexicographic order. Let the set of frequent item sets of size k be Fk and their candidates be Ck. Apriori first scans the database and searches for frequent item sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent item sets.

1. Generate Ck+1, candidates of frequent item sets of size k +1, from the frequent item sets of size k.

2. Scan the database and calculate the support of each candidate of frequent item sets.

3. Add those item sets that satisfies the minimum support requirement to Fk+1.

Function apriori-gen in line 3 generates Ck+1from Fk in the following two step process:

1. Join step

2. Prune step

The Apriori achieves good performance by reducing the size of candidate sets. However, in situations with very many frequent item sets, large item sets, or very low minimum support, it still suffers from the cost of generating a huge number of candidate sets and scanning the database repeatedly to check a large set of candidate item sets. The most outstanding improvement over Apriori would be a method called FP-growth(frequent pattern growth) that succeeded in eliminating candidate generation. It adopts a divide and conquer strategy by compressing the database representing frequent

items into a structure called FP-tree (frequent pattern tree) that retains all the essential information and dividing the compressed database into a set of conditional databases, each associated with one frequent item set and mining each one separately.

IV. PROPOSED SYSTEM

A. *Big Data*

Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making. The data may be enterprise specific or general and private or public. Big data are characterized by 3 V's: Volume, Velocity, Variety. Big Data mining refers to the activity of going through big data sets to look for relevant information. Big data samples are available in astronomy, atmospheric science, social networking sites, life sciences, medical science, government data, natural disaster and resource management, web logs, mobile phones, sensor networks, scientific research, telecommunications.

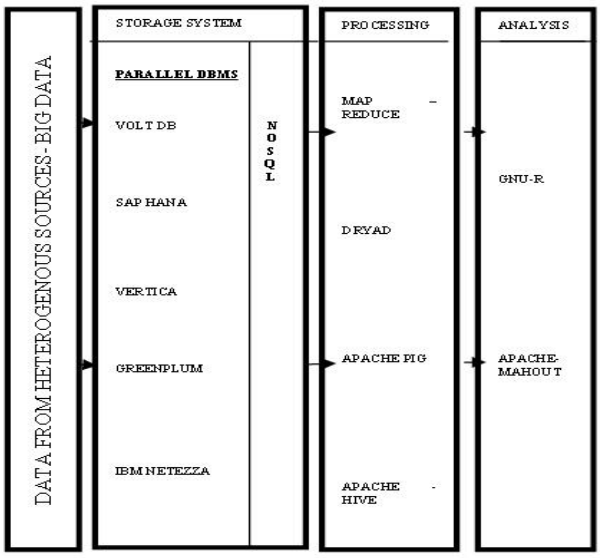

Fig.4. Big Data Architecture

Two main goals of high dimensional data analysis are to develop effective methods that can accurately predict the future observations and at the same time to gain insight into the relationship between the features and response for scientific purposes. Big data have applications in many fields such as Business, Technology, Health, Smart cities etc. These applications will allow people to have better services, better customer experiences, and also to prevent and detect illness much easier than before. The rapid development of Internet and mobile technologies has an important role in the growth of data creation and storage. Since the amount of data is growing exponentially, improved analysis of large data sets is required to extract information that best matches user interests. New technologies are required to store unstructured large data sets and processing methods such as Hadoop and Map Reduce have greater importance in big data analysis. To process large volumes of data from different sources quickly, Hadoop is used. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It allows running applications on systems with thousands of nodes with thousands of terabytes of data. Its distributed file system supports fast data transfer rates among nodes and allows the system to continue operating uninterrupted at times of node failure. It runs Map Reduce for distributed data processing and is works with structured and unstructured data.

B. *Big Data Mining*

The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analyzing fast and massive stream data may lead to new valuable insights and theoretical concepts. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights in the target domain. However, this brings a series of new challenges to the research community. Overcoming the challenges will reshape the future of the data mining technology, resulting in a spectrum of groundbreaking data and mining techniques and algorithms. One feasible approach is to improve existing techniques and algorithms by exploiting massively parallel computing architectures (cloud platforms in our mind). Big data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and interactiveness that existing mining techniques and algorithms are incapable of. The need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM) has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop Map Reduce. NIMBLE is a portable infrastructure that has been specifically designed to enable rapid implementation of parallel ML-DM algorithms, running on top of

Hadoop. Apache's Mahout is a library of machine learning and data mining implementations.

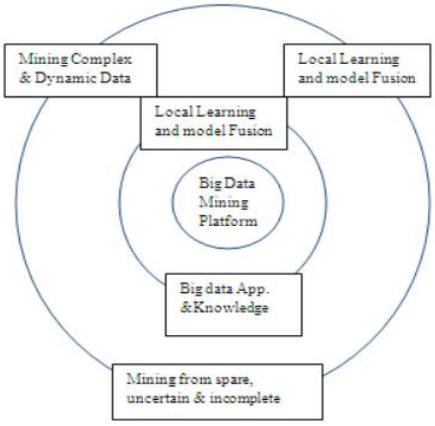

Fig. 5. Big Data processing Framework

Tier III contains three stages. In first stage sparse, heterogeneous, uncertain, incomplete and multi source data is preprocessed by data fusion technique.. In second stage after preprocessing stage complex and dynamic data are mined. Third stage is for local learning and model fusion, where the global knowledge is obtained by local learning and model fusion is tested and the relevant information is feedback to preprocessing stage. Big Data is carry out computing on the PB (Peta byte) or even on EB (Exabyte) data with complex computing process, so parallel computing infrastructure, programming language support and software model utilizing to efficiently analyze and mine distributed data. Map Reduce mechanism is suitable for large scale data mining task on clusters

c. Techniques for big data mining

Big data has great potential to produce useful information for companies which can benefit the way they manage their problems. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. These massive data sets are too large and complex for humans to effectively extract useful information without the aid of computational tools. Emerging technologies such as the Hadoop framework and Map Reduce offer new and exciting ways to process and transform big data, defined as complex, unstructured, or large amounts of data, into meaningful knowledge.

1) Hadoop

Hadoop is a scalable, open source, fault tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS which is fault-tolerant high bandwidth clustered storage architecture. It runs Map Reduce for distributed data processing and is works with structured and unstructured data . For handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. Hadoop and HDFS (Hadoop Distributed File System) by Apache is widely used for storing and managing big data. Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration . HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system. The present Hadoop ecosystem, as shown in Figure 1, consists of the Hadoop kernel, Map Reduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper and these components are explained as below :

•HDFS: A highly faults tolerant distributed file system that is responsible for storing data on the clusters.

•Map Reduce: A powerful parallel programming technique for distributed processing of vast amount of data on clusters.

•HBase: A column oriented distributed NoSQL database for random read/write access.

•Pig: A high level data programming language for analyzing data of Hadoop computation.

•Hive: A data warehousing application that provides a SQL like access and relational model.

•Sqoop: A project for transferring/importing data between relational databases and Hadoop.

•Oozie: An orchestration and workflow management for dependent Hadoop jobs.

The Big Data Analysis and management setup can be understood through the layered structured defined in the figure. The data storage part is dominated by the HDFS distributed file system architecture and other architectures available are Amazon Web Service, Hbase and Cloud Store etc. The data

processing tasks for all the tools is Map Reduce and it is the Data processing tool which effectively used in the Big Data Analysis. For handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. It is interesting to note that for all the tools used, Hadoop over HDFS is the underlying architecture. Oozie and EMR with Flume and Zookeeper are used for handling the volume and veracity of data, which are standard Big Data management tools.

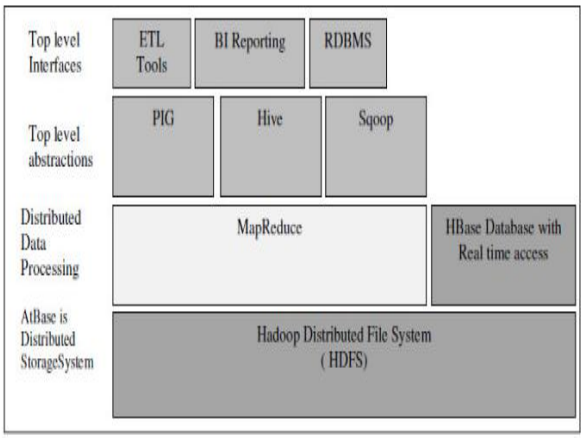

Fig. 6. Hadoop Architecture Tools

2) Map Reduce

Map Reduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Hadoop Map Reduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes . The Map Reduce consists of two functions, map() and reduce(). Mapper performs the tasks of filtering and sorting and reducer performs the tasks of summarizing the result. There may be multiple reducers to parallelize the aggregations . Users can implement their own processing logic by specifying a customized map() and reduce()function. The map() function takes an input key/value pair and produces a list of intermediate key/value pairs. The Map Reduce runtime system groups together all intermediate pairs based on the intermediate keys and passes them to reduce() function for producing the final results. Map Reduce is widely used for the Analysis of big data. Large scale data processing is a difficult task. Managing hundreds or thousands of processors and managing parallelization and distributed environments makes it more difficult. Map Reduce provides solution to the mentioned issues since it supports distributed and parallel I/O scheduling. It is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data.

v. CONCLUSION

Big Data mining will help us to discover knowledge that no one has discovered before. Big Data is a very challenging and recent trend research area in the IT industry. Big Data is an emerging trend and there is instant need of new machine learning and data mining techniques to analyze massive and complex amount of data in near future. To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. Hadoop and Map Reduce tool for big data is described in detail focusing on the areas where it needs to be improved so that in future Big data can have technology as well as skills to work with. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. Big Data analysis tools like Map Reduce over Hadoop and HDFS which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In real-world applications managing and mining Big Data is Challenging task, The Map Reduce framework is one of the most important parts of big data processing, and batch oriented parallel computing model. We regard Big data as an emerging trend and the need for Big data mining is rising in all science and engineering domains.

REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[3] Devlin, B. & Murphy, P. (1988) An Architecture for a Business and Information System, IBM Systems Journal, 27 (1), 60-80.

[4] Inmon, W.H. (1996) Building the Data Warehouse, Second Edition, New York: John Wiley & Sons.

[5] Kimball, R. (1996) The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, New York: John Wiley & Sons.

[6] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp.487–499.

[7] Ahmed S, Coenen F, Leng PH (2006) Tree-based partitioning of date for association rule mining. Knowl Inf Syst 10(3):315–331.

[8] Banerjee A, Merugu S, Dhillon I, Ghosh J (2005) Clustering with Bregman divergences. J Mach Learn Res 6:1705–1749.

[9] Frolick, M.N. & Lindsey, K. (2003) Critical Factors for Data Warehouse Failure, Journal of Data Warehousing, 8 (1).

[10] Hwang, H.-G., Kua, C.-Y., Yenb, D. C., & Chenga, C.-C. (2002) Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan, Decision Support Systems, In Press, Corrected Proof.

[11] "Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012.