© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Big Data with Data Mining

P. Nandhini[1], M. Pavithra[2], R. Suganya[2]
[1]PG Scholar, Department of C.S.E, Jansons Institute of Technology, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of C.S.E, Jansons Institute of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Big Data relates large-volume, complex, increasing data sets with multiple independent sources. With the rapid evolution of data, data storage and the networking collection capability, Big Data are now speedily expanding in all science and engineering domains. Big Data mining is the ability of extracting constructive information from huge streams of data or datasets, that due to its variability, volume, and velocity [4]. Data mining includes exploring and analyzing big quantity of data to locate different molds for big data. Artificial intelligence (AI) and statistics are the fields which develop these techniques, This paper discusses a characterizes applications of Big Data processing model and Big Data revolution, from the data mining outlook [2]. The analysis of big data can be troublesome because it often involves the collection and storage of mixed data based on different patterns or rules (heterogeneous mixture data). This has made the heterogeneous mixture property of data a very important issue [5]. We are in a world of increasing data; the day-by-day generation of data becomes very vast. The true of it lies in collecting these data, analyze and perform some computations to obtain some meaningful information. This meaningful information can be derived using some data mining tasks [8]. In short we can call big data as a "asset" and data mining is a "handler" that is used to provide beneficial results. To perform these analysis data mining algorithms can be used and also the big data methods [7]. The research dealt with exploration methods and tools for big data. It identifies the challenges encountered in the analysis of big data. Such huge data is now referred to as Big Data. Big data mining leads to the discovery of the useful information from huge data repositories [9]. However, this huge amount of data hinders existing data mining tools and thus creates new research challenges that open the door for new research opportunities [6]. In this paper, we provide an overview of the research challenges and opportunities of big data mining. We present the technologies and platforms that are required for mining big data. A number of applications that can benefit from mining big data are also discussed. We discuss the status of big data mining, current efforts and future research directions in the UAE [1].

Keywords: Big Data, Data Mining, HACE Theorem, Structured and Unstructured

I. INTRODUCTION

Big data refers to the enormous amount of structured and unstructured data that overflow the organization. If the overflowed data is used in a proper way it leads to meaningful information [5]. When big data is compared to traditional databases it includes a large number of data which requires more processing in real time [2]. It also provides opportunities to discover new values, to understand an in-depth knowledge from hidden values and also provides space to manage those data effectively. Big Data concern large-volume, complex, growing datasets with multiple data sources [3]. With the fast development of networking, data storage and data collection capacity, big data are now expanding in all science and engineering domains,

including physical, biological and biomedical sciences [1]. Data Mining is a task of identifying relevant and significant information from large data set.

The amount of data is growing exponentially as it is doubling in size every two years. By 2020, as indicated by Fig. 1, the size of the world's digital data will reach 44 zettabytes (44,000 petabytes), according to a recent IDC study, (IDC, 2014). This dramatic increase in the amount of data is paralleled with an increase in the various data-generating devices, such as sensors, mobile devices and cameras, and data generating applications such as social media, location-based services and the Internet [5]. While Google currently stores about 15 zettabytes of data in its warehouses and processes data at a daily rate of 100 petabytes. The massive data increases are beyond the capability of current technologies to store analyze and extract useful information from big data [3].

Researchers refer to the huge datasets that are unmanageable by current technologies and software tools as Big Data. Big data may come in different forms such as text, images, videos, sounds, or their combinations [6]. In addition to the large size and variety of big data, its incoming rate is often very high. The author of argued that the main characteristics of big data are the three V's (Volume, Velocity, and Variety). These characteristics pose serious challenges to big data management and mining [4].

In this paper, we review the big data concepts, mining and required technologies. We also discuss the characteristics of big data, its issues and challenges, benefits to the industry and community at large. We also explore some future research directions [7]. The main applications in the UAE that could benefit from big data are also presented. The aim of our effort is to shed light on some of the benefits and challenges of big data research in the UAE and help researchers in data mining to explore new research directions to solve emerging challenges brought about by big data [8].

In the modern world it would be difficult to find areas of human activity, in which data is not collected and analyzed. Data is rapidly increasing by development of new technologies, while increasing the need to analyze the available data [1]. Recently huge amounts of data are constantly generated and stored in the vaults in the various research fields. It is often not only the amount of data is great, but these data are constantly updated and supplemented with the new ones. In addition, there is a very wide variety of data types and sources. Such data is called big data [10]. Processing and analysis of big data are encountered with difficulties in various fields, such as medicine, finance, economics, engineering, etc. Different methods are being developed for big data investigation tasks, such as clustering, classification, statistical and visual analysis. Study of big data is one of the biggest challenges faced by data [11].

II. BIG DATA EVOLUTION

The introduction of the Internet and World Wide Web in 1990s have increase the number of web applications and web services, which in turn increased the amount of generated data tremendously [5]. Companies and governmental agencies are now dealing with large databases whose sizes are in petabytes and in the near future the sizes of these databases will reach Exabyte's [7].

2.1 Characteristics of Big Data

The size of big data is beyond the capability of traditional database algorithms to store, manage and analyze it. However, it is not just the volume of data but also the variety and velocity of the data. These three attributes form the three Vs of Big Data [1]. Organizations are now storing massive amount of data in the form of textual, numerical, or multimedia data. The size of such data in each individual organization will soon reach Exabyte's. That is big data may come from different sources such as sensors, mobile devices, social networks, etc. [3]. Moreover, big data may consist of various types and different levels of

complexities [5]. Big data could be structured such as relational databases, unstructured such as multimedia data, and/or semi-structures such as xml and social media data. These different types of data often generated and shared at high velocities. Therefore, the three vs. have made it difficult for current technologies to handle big data [2].

2.2 Emergence of Big Data

The wide spread of Web 2.0 with various popular applications including the various forums, newsgroups and social media contributed to the increase of content of data. Advances in digital sensors, mobile devices, communications, computers, and storage devices have contributed to the massive growth of data. Some organizations are trying to extract valuable information from these massive collections of data [5]. Google and other search engines have created profitable businesses by collecting the information posted on the Internet and presenting it to people in a useful way [6]. The use of big data by search engines has transformed the way people access information. Other big organizations and companies are now collecting huge amounts of their daily transactions and are trying (or hoping) to process it and extract valuable information to give them a competitive edge in the market [10].

2.3 Motivation of Big Data

As the size of collected and stored data is increasing tremendously, the chances of extracting useful information to gain business advantage is also increasing. Furthermore, the cost of new technologies to store this huge amount of data has fallen dramatically allowing even average-sized companies to collect and manage big data [6]. The peer pressure of competitor organizations forces individual organizations to collect and process big data to extract valuable information to allow them to remain competitive in the market [7]. Decision making will learn from big data to find patterns, relationships between data, correlations between different types of data, groupings of data, etc. [5].

III. TECHNIQUES FOR BIG DATA MINING

Traditional data mining discovers useful information, interesting groupings, valuable patterns and relationships hidden in the data [4]. The discovered results help make valuable predictions and help formulate decisions in the real world. Various applications have benefited from data mining such as medicine, business, and science [5]. Recently, data mining algorithms have been facing many challenges when applied to huge amount of data due to the limitations of these algorithms in handling the characteristics of big data [6]. Despite of these limitations, big data brings new opportunities for extracting valuable insights from the complex and heterogeneous contents. It is believed that big data will play a critical role in the future and affect the way businesses and services are conducted [8].

For example, governments may make use of big data in the form of social media and other sources of online information to gauge public satisfaction about governmental services, identify the need for new government facilities, detect suspicious criminal groups, or predict future threats [4]. However, the capabilities of existing database technologies cannot extend to the requirements of managing big data. To meet the requirements of big data, Google introduced MapReduce, which is a new programming model [5]. This programming model requires a distributed platform; therefore, a distributed file system, called GFS (Google File System), was used by Google to divide the huge the datasets among thousands of computers that are referred to as a cluster. On another front, Yahoo created Hadoop MapReduce, which uses the Hadoop Distributed File System (HDFS). Both of Yahoo's technologies are open source version of the Google's technologies [2].

3.1 MapReduce

With MapReduce, the input data is partitioned into large sets of key-value pairs. These sets are then processed by map () functions in parallel. Each map ()

function processes the local data, and writes the output to a temporary storage [5]. This mapping task is then followed by applying a reduce () function to the result of the map () functions. The reduce () functions merges each group of output data based on common keys, in parallel, to obtain the final result [6]. MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Hadoop MapReduce 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 [1].

The MapReduce 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 [7]. 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 [5]. The MapReduce runtime system groups together all intermediate pairs based on the intermediate keys and passes them to reduce () function for producing the final results [8].

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 [3]. 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 [1].

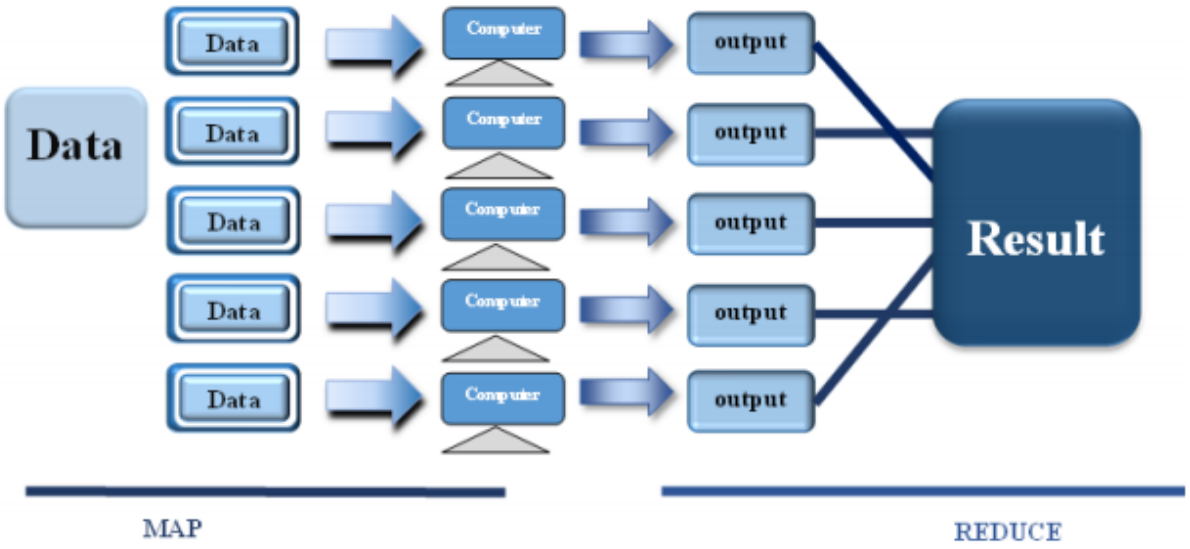

Figure 1

3.2 Hadoop

Hadoop is a distributed and scalable file-system written in Java. It is an open-source framework for distributed storage and distributed processing of Big Data on clusters of computers. Hadoop stores huge amounts of data and processes them efficiently via distributed processing [3]. Hadoop also replicates the sets of data on several computer nodes to achieve reliability [9]. 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 [5]. It runs MapReduce for distributed data processing and is works with structured and unstructured data [11].

For handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are

parts of Hadoop and HDFS framework [10]. 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 [6].

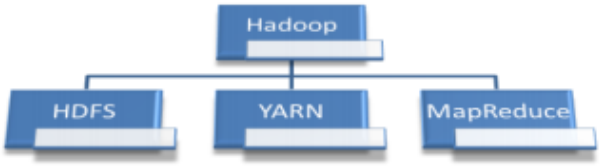

Figure 2

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 [5]. The present Hadoop ecosystem, as shown in Figure, consists of the Hadoop kernel, MapReduce, 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 [6]:

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

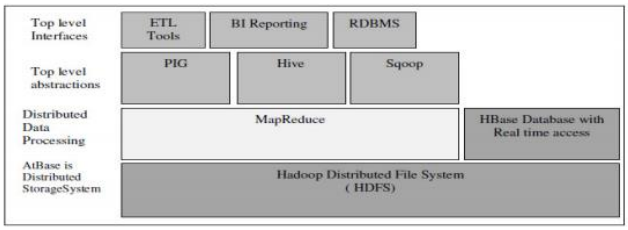

Figure 3

BIG DATA WITH DATA MINING

Generally big data refers to a collection of large volumes of data and these data are generated from various sources such as internet, social media, business organizations etc., with these data some useful information can be extracted with the help of data mining [6]. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data [2].

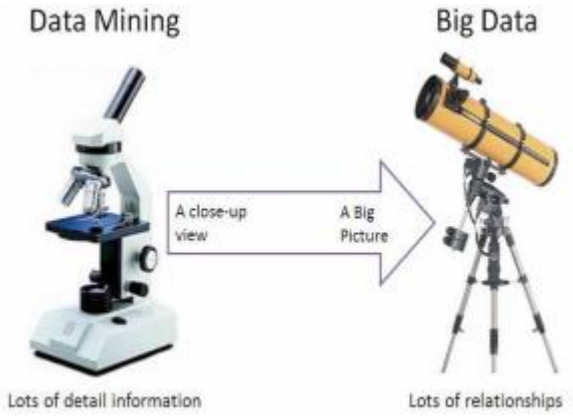

Figure 4. Data Mining with Big Data

The figure given above portraits the relationship of big data with data mining. From the figure it is observed that big data gives lots of relationships and data mining gives lots of information [3].

BIG DATA CHARACTERISTICS -HACE THEOREM

Big Data starts with large volume, heterogeneous autonomous sources with distributed and decentralized control and seeks to explore complex and evolving relationships among data [1].These characteristics makes it an extreme challenge for discovering useful information from big data. In connection with this scenario, let us imagine a scenario where blind people are asked to draw the picture of an elephant. The information collected by

each blind people will be such that they may think the trunk as a „wall", leg as a „tree", body as a „wall" and tail as a „rope". In this case one blind man can exchange information with other which may be biased [9].

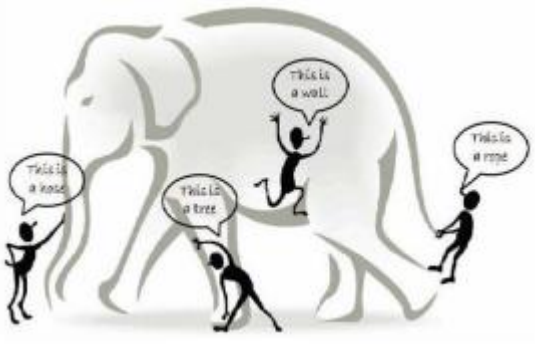

Figure 5

i. Vast data with heterogeneous and diverse sources one of the fundamental characteristics of big data is the large volume of data represented by heterogeneous and diverse dimensions [4]. For example in the biomedical world, a single human being is represented as name, age, gender, family history etc., For X-ray and CT scan images and videos are used. Taking the example heterogeneity refers to the different types of representations of same individual and diverse refers to the variety of features to represent single information [1].

ii. ii. Autonomous with distributed and de-centralized control These are the main characteristics of big data. Since the sources are autonomous, i.e., automatically generated, it generates information without any centralized control. We can compare it with World Wide Web (WWW) where each server provides a certain amount of information without depending on other servers [5].

DATA MINING FOR BIG DATA

Generally data mining also known as data or knowledge discovery is the process which analyses data from different perspectives and discover useful information from it [8]. Data mining contains several algorithms which fall into four categories.

They are

1) Association Rule.
2) Clustering.
3) Classification.
4) Regression.

(1) Association is used to search relationship between variables. It is applied in searching for frequently visited items. In short it establishes relationship among objects [3].

(2) Clustering discovers groups and structures in the data. i.e., it classifies the data belongs to which group [9]. Classification deals with associating an unknown structure to a known structure.

(3) Regression finds a function to model the data.

These data mining algorithms can be converted into big data map reduce algorithm which is based on parallel computing basis [7]. As data clustering has attracted a significant amount of research attention in past decades, many clustering algorithms has been proposed. However the enlarging data in applications makes clustering of very large scale of data a challenging task [6].

DATA MINING FOR BIG DATA

Data mining includes extracting and analyzing bulky amounts of data to discover models for big data. The methods came out of the grounds of artificial intelligence (AI) and statistics with a tad of database management [8].

Searching information from data takes two major forms: prediction and description. it is tough to know what the data shows?. Data mining is used to summarize and simplify the data in a way that we can recognize and then permit us to gather things about specific cases based on the patterns normally; the objective of the data mining is either prediction or classification [10]. In classification, the thought is to arrange data into sets. For example, a seller might be attracted in the features of those who answered versus who didn't answered to an advertising. There are two divisions [5]. In prediction, the plan is to predict the

rate of a continuous variable. For example, a seller might be involved in predicting those who will reply to a promotion [11].

Distinctive algorithms used in data mining are as follows:

A. Classification trees:

A famous data-mining system that is used to categorize a needy categorical variable based on size of one or many predictor variables [4]. The outcome is a tree with links and nodes between the nodes that can be interpret to form if-then rules.

B. Logistic regression:

A algebraic technique that is a modification of standard regression but enlarges the idea to deal with sorting. It constructs a formula that predicts the possibility of the occurrence as a role of the independent variables [5].

C. Neural networks:

A software algorithm that is molded after the matching architecture of animal minds. The network includes of output nodes, hidden layers and input nodes. Each unit is allocated a weight. Data is specified to the input node, and by a method of trial and error, the algorithm correct the weights until it reaches a definite stopping criteria [7]. Some groups have likened this to a black–box system.

D. Clustering techniques like K-nearest neighbors:

A procedure that identifies class of related records. The K-nearest neighbor technique evaluates the distances between the points and record in the historical data [9]. It then allocates this record to the set of its nearest neighbor in a data group.

THE DATA MINING IN ERA OF BIG DATA

Big data mean those data sets, which become quite difficult or impossible to handle by using simple data-processing programs and tools because of their size and complex structure [2], [8]. Big data term is used quite often, but it is not always defined correctly [6]. Sometimes big data is characterized only by their volume. Although the word "significant" means the volume and size but big data is described by more than one characteristic [4]:

- Volume - one of the big characteristics, illustrating the size of the data.
- Variety - the characteristic that defines difference of the data types.
- Velocity - the term is understood as the satisfaction of a requirement for updating of data, growth and processing.
- Variability - characteristic used to describe the constant change of data and renewal.
- Veracity - the term is associated with the correctness and accuracy of the data and their analysis.
- Complexity or Value - associated with the ever-growing amounts of data, their variety and the problems arising from the analysis of these data. Data mining is the main task of big data.

The main steps of data processing are shown in the picture at the top. The data features which makes processing of data complex and complicated are presented in the bottom [7]. The scheme demonstrates that the data processing is composed of many steps, each of which is encountering some challenges, requiring appropriate solutions [9]. Storage and processing of big data is different from traditional data analysis methods. When individual computer resources are not enough, it is recommended to use distributed and parallel systems or a distributed data mining [10]. If the analyzed data are divided in some ways, the data research task addresses in parallel computer clusters or Grid. Computer cluster - computers combined in a single network which are able to carry out distributed computing. Grid - this is like a cluster, it is freely available, combined infrastructure, but it consists of separate computing clusters [2].

DIFFERENCES BETWEEN BIG DATA AND DATA MINING.

We cannot say data mining as „big data“ and big data as „data mining“. There are some differences between these two and they are shown below [3].

Table 1

Data Mining	Big Data
Data mining is the old big data	Big data is everything in the world now
Data size is smaller	Data size is Larger
Finding interesting patterns	Involves large scale storage and processing of large data sets.
All data mining tasks are not big data	All big data tasks are data mining

CHALLENGING ISSUES WITH BIG DATA

Big data has been one of the current and future research problems. In the year 2014, Gartner listed "Top ten Strategic Technologies trends for 2013" and „Top ten critical Technology Trends for the next five years" and big data is listed in both two [5]. Challenges in big data are very large. On one hand big data had many opportunities and on the other hand it is facing lot of challenges too. When handling big data challenges occurs in the following areas.

- Data Capture and Storage.
- Data Transmission.
- Data Curation.
- Data Analysis.

i. Data Visualization. According to [1] challenges of big data mining is generally divided into three tiers.

Figure 6

The first tier includes the setup of data mining platforms. The second one includes

1. Information sharing and Data privacy.

2. Domain and Application Knowledge. The third one includes Local Learning and model fusion for multiple information sources.

3. Mining from sparse, uncertain and incomplete data.

4. Mining complex and dynamic data.

Generally mining of data from different data sources is tedious one as the data size is larger. And also big data is stored at different places collecting those data [9].

SECURITY AND PRIVACY CHALLENGES FOR BIG DATA

Big data refers to collections of data sets with sizes outside the ability of commonly used software tools such as database management tools or traditional data processing applications to capture, manage, and analyze within an acceptable elapsed time. Big data sizes are constantly increasing, ranging from a few dozen terabytes in 2012 to today many petabytes of data in a single data set [7]. Big data creates tremendous opportunity for the world economy both in the field of national security and also in areas ranging from marketing and credit risk analysis to medical research and urban planning. The extraordinary benefits of big data are lessened by concerns over privacy and data protection [4]. As big data expands the sources of data it can use, the trust worthiness of each data source needs to be verified and techniques should be explored in order to identify maliciously inserted data [5]. Information security is becoming a big data analytics problem where massive amount of data will be correlated, analyzed and mined for meaningful patterns [8].

Any security control used for big data must meet the following requirements:

- ✓ It must not compromise the basic functionality of the cluster.
- ✓ It should scale in the same manner as the cluster.
- ✓ It should not compromise essential big data characteristics.
- ✓ It should address a security threat to big data environments or data stored within the cluster.

Unauthorized release of information, unauthorized modification of information and denial of resources are the three categories of security violation [6].

IV. CONCLUSION

Big data is directed to continue rising during the next year and every data scientist will have to handle a large amount of data every year .This data will be more miscellaneous, bigger and faster. We discussed in this paper several insights about the subjects and what we think are the major concern and the core challenges for the future. Big Data is becoming the latest final border for precise data research and for business applications [5]. Data mining with big data will assist us to discover facts that nobody has discovered before. The heterogeneous mixture learning technology is an advanced technology used in big data analysis. In the above, we introduced difficulties that are inherent in heterogeneous mixture data analysis, the basic concept of heterogeneous mixture learning and the results of a demonstration experiment that deal with electricity demand predictions [4]. As the big data analysis increases its importance, heterogeneous mixture data mining technology is also expected to play a significant role in the market [7]. The range of application of heterogeneous mixture learning will be expanded broader than ever in the future [8]. To investigate Big Data, we have examined a number of challenges at the system levels, data and model. To hold Big Data mining, high performance computing platforms are necessary, which enforce organized designs to set free .the complete power of the Big Data [9]. By the data level, the independent information sources and the range of the data gathering environments, habitually result in data with complex conditions, such as missing unsure values. The vital challenge is that a Big Data mining structure needs to consider complicated interaction between data sources, samples and models along with their developing changes with time and additional probable factors [11].

The data mining techniques can be applied on big data to acquire some useful information from large datasets. Thus these two terms are not different instead they are coupled together to acquire some useful picture from the data [4]. Thus we conclude that big data will become an excellent opportunity in the fourth coming years. We discussed some of the useful information about big data and data mining and have identified the research gaps and open research areas [10]. A system wants to be cautiously designed so that unstructured data can be connected through their composite relationships to form valuable patterns, and the development of data volumes and relationships should help patterns to guess the tendency and future [11].

V. FUTURE ENHANCEMENT

The amounts of data is growing exponentially worldwide due to the explosion of social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. Big Data is becoming the new area for scientific data research and for business applications [5]. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules [6]. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities [7]. In this paper we discussed about the issues and challenges related to big data mining and also 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 [5]. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data [8]. That will help the research scholars to choose the best mining tool for their work.

VI. REFERENCES

[1]. Xindong Wu, Xingquan Zhu, Gong Qing Wu, Wei Ding, "Data mining with Big data", IEEE, Volume 26, Issue 1, January 2014.

[2]. Bharti Thakur, Manish Mann , " Data Mining for Big Data- A Review", IJARCSSE, Volume 4, Issue 5, May 2014.

[3]. Rohit Pitre, Vijay Kolekar, "A Survey Paper on Data Mining With Big Data", IJIRAE, Volume 1, Issue 1, April 2014.

[4]. Dr. A.N. Nandhakumar, Nandita Yambem , "A Survey of Data Mining Algorithms on Apache Hadoop Platforms", IJETAC, Volume 4, Issue 1, January 2014.

[5]. C.L. Philip Chen, C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Inform. Sci, http://dx.doi.org/10.1016/j.ins.2014.01.015,2014.

[6]. Che, D., Safran, M., Peng, Z, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities", DASFAA Workshops, LNCS 7827, pp. 1–15, 2013.

[7]. H.V. Jagadish, "Big data and science: myths and reality", Big Data Research, vol. 2(2), pp. 49–52, 2015.

[8]. X. Jin, B.V. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research", Big Data Research, vol. 2(2), pp. 59–64, 2015.

[9]. C. Sherman, "What's the big deal about big data?" Online Searcher 38.2. ProQuest Central, pp.10–17, 2014.

[10]. Albert Bifet, "Mining Big data in Real time", Informatica 37, pp15-20, 2013.

[11]. Richa Gupta, "Journey from data mining to Web Mining to Big Data", IJCTT, 10(1), pp18-20, 2014.

[12]. Priya P. Sharma, Chandrakant P. Navdeti, "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131, 2014.