ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 8, February 2018

A Review-Big data mining issues and challenges

[1]Dr Attili Venkata Ramana, Sreenidhi Institute of science and Technology
[2]Dr C.Sunil Kumar, Professor in ECM, SNIST

Abstract*— Data has become an indispensable part of every economy, industry, organization, business function and individual. Big Data Mining is a term used to identify the datasets that whose size is beyond the ability of typical database software tools to store, manage and analyze. The Big Data introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation and measurement errors. These challenges are distinguished and require new computational and statistical paradigm. This paper presents the literature review about the big data mining issues and challenges with emphasis on the distinguished features of Big Data. It also discusses some methods to deal with big data. The hurdles of securing the data and democratizing it have been elaborated amongst several others such as inability in finding sound data professionals in required amounts and software that possess ability to process data at a high velocity.

***Index Terms*— Big data mining, Security, Hadoop, MapReduce.**

I. INTRODUCTION

Data is the collection of values and variables related in some sense and differing in some other sense. In recent years the sizes of databases have increased rapidly. This has lead to a growing interest in the development of tools capable in the automatic extraction of knowledge from data [1].Data are collected and analyzed to create information suitable for making decisions. Hence data provide a rich resource for knowledge discovery and decision support. A database is an organized collection of data so that it can easily be accessed, managed, and updated. Data mining is the process discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories. A widely accepted formal definition of data mining is given subsequently. According to this definition, data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data **[2]**. Data mining uncovers interesting patterns and relationships hidden in a large volume of raw data. Big Data is a new term used to identify the datasets that are of large size and have grater complexity **[3]**. So we cannot store, manage and analyze them with our current methodologies or data mining software tools. Big data is a heterogeneous collection of both structured and unstructured data. Businesses are mainly concerned with managing unstructured data. Big Data mining is the capability of extracting useful information from these large datasets or streams of data which were not possible before due to its volume, variety, and velocity. The extracted knowledge is very useful and the mined knowledge is the representation of different types of patterns and each pattern corresponds to knowledge. Data Mining is analyzing the data from different perspectives and summarizing it into useful information that can be used for business solutions and predicting the future trends. Mining the information helps organizations to make knowledge driven decisions. Data mining (DM), also called Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process of searching large volumes of data automatically for patterns such as association rules **[4].** It applies many computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining extract only required patterns from the database in a short time span. Based on the type of patterns to be mined, data mining tasks can be classified into summarization, classification, clustering, association and trends analysis.

Enormous amount of data are generated every minute. A recent study estimated that every minute, Google receives over 4 million queries, e-mail users send over 200 million messages, YouTube users upload 72 hours of video, Face book users share over 2 million pieces of content, and Twitter users generate 277,000 tweets **[5].** With the amount of data growing exponentially, improved analysis is required to extract information that best matches user interests. Big data refers to rapidly growing datasets with sizes beyond the capability of traditional data base tools to store, manage and analyze them. Big data is a heterogeneous collection of both structured and unstructured data. Increase of storage capacities, Increase of processing power and availability of data are the main reason for the appearance and growth of 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, and Variety **[6].**

Volume -the size of data now is larger than terabytes and peta bytes. The large scale and rise of size makes it difficult to store and analyze using traditional tools.

Velocity – big data should be used to mine large amount of data within a pre defined period of time. The traditional methods of mining may take huge time to mine such a volume of data.

Variety – Big data comes from a variety of sources which includes both structured and unstructured data. Traditional database systems were designed to address smaller volumes

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 8, February 2018

of structured and consistent data whereas Big Data is geospatial data, 3D data, audio and video, and unstructured text, including log files and social media. This heterogeneity of unstructured data creates problems for storage, mining and analyzing the data.

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. 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 **[8].**

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. This paper is organized as follows. Section 1 gives introduction and Section 2 presents literature review. Section 3 presents the issues and challenges of big data mining. Section 4 provides an overview of security and privacy challenges of big data and Section 5 describes some technologies to deal with big data analysis. Section 6 concludes this paper with summaries.

Procedure for Paper Submission

II. ISSUES AND CHALLENGES

Big data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and Interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining.

A. Heterogeneity and Incompleteness

The difficulties of big data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data. In the case of complicated heterogeneous mixture data, the data has several patterns and rules and the properties of the patterns vary greatly. Data can be both structured and unstructured. 80% of the data generated by organizations are unstructured. They are highly dynamic and does not have particular format. It may exists in the form of email attachments, images, pdf documents, medical records, X rays, voice mails, graphics, video, audio etc. and they cannot be stored in row/ column format as structured data. Transforming this data to structured format for later analysis is a major challenge in big data mining. So new technologies have to be adopted for dealing with such data.

Incomplete data creates uncertainties during data analysis and it must be managed during data analysis. Doing this correctly is also a challenge. Incomplete data refers to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values. While most modern data mining algorithms have inbuilt solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field which seeks to impute missing values in order to produce improved models (compared to the ones built from the original data). Many imputation methods exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance.

B. Scale and complexity

Managing large and rapidly increasing volumes of data is a challenging issue. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analyzed.

C. Timeliness

As the size of the data sets to be processed increases, it will take more time to analyze. In some situations results of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed by preventing the transaction from taking place at all. Obviously a full analysis of a user's purchase history is not likely to be feasible in real time. So we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 8, February 2018

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. In such cases Index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria.

III. 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. 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. 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. Information security is becoming a big data analytics problem where massive amount of data will be correlated, analyzed and mined for meaningful patterns. 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. The following are some of the security threats:

- An unauthorized user may access files and could execute arbitrary code or carry out further attacks.
- An unauthorized user may eavesdrop/sniff to data packets being sent to client.
- An unauthorized client may read/write a data block of a file.
- An unauthorized client may gain access privileges and may submit a job to a queue or delete or change priority of the job.

Security of big data can be enhanced by using the techniques of authentication, authorization, encryption and audit trails. There is always a possibility of occurrence of security violations by unintended, unauthorized access or inappropriate access by privileged users. The following are some of the methods used for protecting big data:

A. Using authentication methods: - Authentication is the process verifying user or system identity before accessing the system. Authentication methods such as Kerberos can be employed for this.

B. Use file encryption: Encryption ensures confidentiality and privacy of user information, and it secures the sensitive data. Encryption protects data if malicious users or administrators gain access to data and directly inspect files, and renders stolen files or copied disk images unreadable. File layer encryption provides consistent protection across different platforms regardless of OS/platform type. Encryption meets our requirements for big data security. Open source products are available for most Linux systems, commercial products additionally offer external key management, and full support. This is a cost effective way to deal with several data security threats.

C. Implementing access controls: Authorization is a process of specifying access control privileges for user or system to enhance security.

D. Use key management: File layer encryption is not effective if an attacker can access encryption keys. Many big data cluster administrators store keys on local disk drives because it's quick and easy, but it's also insecure as keys can be collected by the platform administrator or an attacker. Use key management service to distribute keys and certificates and manage different keys for each group, application, and user.

E. Logging: To detect attacks, diagnose failures, or investigate unusual behavior, we need a record of activity. Unlike less scalable data management platforms, big data is a natural fit for collecting and managing event data. Many web companies start with big data particularly to manage log files. It gives us a place to look when something fails, or if someone thinks you might have been hacked. So to meet the security requirements, we need to audit the entire system on a periodic basis.

F. Use secure communication: Implement secure communication between nodes and between nodes and applications. This requires an SSL/TLS implementation that actually protects all network communications rather than just a subset. Thus the privacy of data is a huge concern in the context of Big Data. There is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. So, unauthorized use of private data needs to be protected.

To protect privacy, two common approaches used are the following. One is to restrict access to the data by adding certification or access control to the data entries so sensitive information is accessible to a limited group of users only. The other approach is to anonymize data fields such that sensitive information cannot be pinpointed to an individual record. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 8, February 2018

sensitive information can be misconduct by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals **[10].**

IV. 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 MapReduce offer new and exciting ways to process and transform big data, defined as complex, unstructured, or large amounts of data, into meaningful knowledge.

A. 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 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. 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, 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:

- 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 dataon 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 2 gives an overview of the Big Data analysis tools which are used for efficient and precise data analysis and management 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 CloudStore 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 [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. 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.

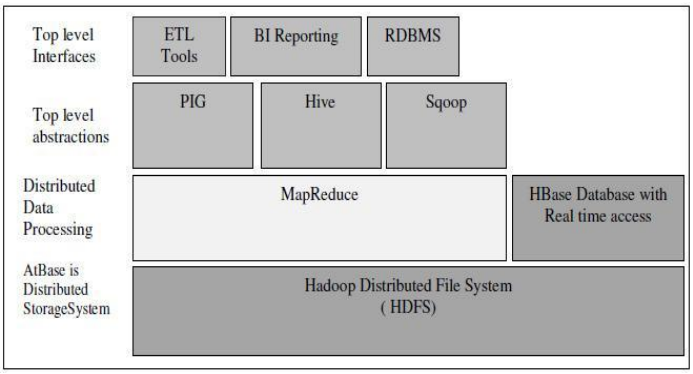

Fig 1: Hadoop Architecture Tools [6]

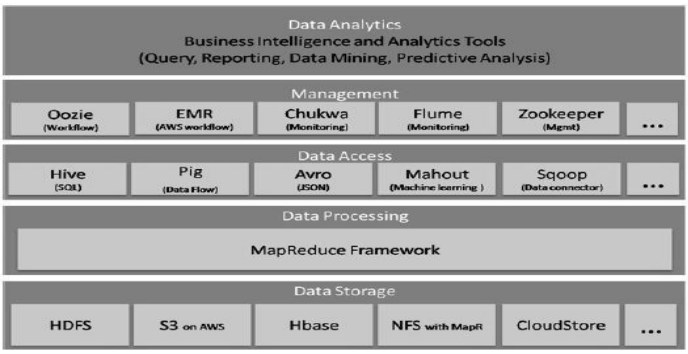

Fig 2: Big data analysis tools [11]

V. MAPREDUCE

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.

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. Users can implement their own processing logic

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 7, Issue 8, February 2018

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 MapReduce 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 [11].

VI. CONCLUSION AND FUTURE SCOPE

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. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. 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. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data. That will help the research scholars to choose the best mining tool for their work.

Today, Big Data is influencing IT industry like few technologies have done before. The massive data generated from sensor-enabled machines, mobile devices, cloud computing, social media, satellites help different organizations improve their decision making and take their business to another level. "Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives," - Susan Hauser, corporate vice president of Microsoft. Data is the biggest thing to hit the industry since PC was invented by Steve Jobs. As mentioned earlier in this paper, every day data is generated in such a rapid manner that, traditional database and other data storing system will gradually give up in storing, retrieving, and finding relationships among data.

REFERENCES

[1] Julie M. David, Kannan Balakrishnan, (2011), Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets, Int. J. of Computer and Electrical Engineering, Hong Kong, 3(1), pp163-169.

[2] Julie M. David, Kannan Balakrishnan, (2011), Prediction of Learning Disabilities in School-Age Children using SVM and Decision Tree, Int. J. of Computer Science and Information Technology, ISSN 0975-9646, 2(2), pp829-835.

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

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

[5] http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/.

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

[7] Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), "Big Data: Overview", IJCTT, 9 (5).

[8] Wei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, 14 (2), pp1-5.

[9] Chanchal Yadav, Shullang Wang, Manoj Kumar, (2013) "Algorithm and Approaches to handle large Data- A Survey", IJCSN, 2(3), ISSN: 2277-5420(online), pp2277-5420.

[10] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data".

[11] Puneet Singh Duggal, Sanchita Paul, (2013), "Big Data Analysis: Challenges and Solutions", Int. Conf. on Cloud, Big Data and Trust, RGPV.