International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 22 (2018) pp. 16021-16026

A Review on Big Data Mining

A. Raja
Assistant Professor, Kavitha's College of Arts & Science,
Tiruchengodu, Vaiyappamalai Rd, Chinnamanali, Tamil Nadu, India.

Dr. S. Prema
Assistant Professor, K.S.R. College of Arts & Science,
Tiruchengodu, Namakkal, Tamil Nadu, India

Abstract

With the fast development of technologies, social media generates huge amount of data rapidly every minuteand this accentuate the need of big data that can hold fast transactions and analytics synchronously. Big data examines large amounts of data to uncover hidden patterns, correlations and other insights to make extract value. Big data analytics eventually makes creating smarter and learner organizations to retrieve and analyze faster and more efficiently compared to the traditional analytics. These papers propound the nature, characteristics and concept of big data and how organizations can advance their systems with big data technologies. With today's technology, it's possible to analyze any massive data to reveal accurate retrieval with a very short span of time.

Keyword: Big data, Five V's, Hadoop, MapReduce.

I. INTRODUCTION

According to Gartner IT Glossary, big data is defined as: "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

History and evolution of big data analytics: Since, big data is a recent upcoming technology but the concept has been around for years. In most organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get significant value from it. But even in the 1950s, decades before anyone uttered the term "big data" businesses were using traditional analytics examined using a spreadsheet.

Most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information and to develop techniques that automatically discover new, hidden or unsuspected data from the large text collection. The nature of user's interest can be changed dynamically which reflects to the demand of services.

Big data generates variety of data includes both structured and unstructured data. While, structured data include numbers, words, letters and alpha numeric that can be easily categorized and analyzed. Whereas unstructured data include more complex information, such as multimedia information consists of image, audio and video file. These data cannot easily be separated into categories or analyzed numerically.

The large volume of data exploded to unimaginable levels and the challenge is to make sense of this large pool of data. The process of converting large amounts of unstructured raw data, retrieved from diverse sources to process as useful data for organizations forms the core of Big Data Analytics.

The following are implementation of big data in various sectors:

- ✓ Machine learning implementation could be a classification algorithm, a regression model or a segmentation model.
- ✓ Recommender system recommends choices based on user behavior.
- ✓ Dashboard is a graphical mechanism to make visualize aggregate data accessible.
- ✓ Ad-Hoc analysis performs hypotheses or myths that can be answered with data.

II. LITERATURE SURVEY

Xindong Wu, et.al [1] proposes a HACE theorem that characterizes the features of the big data revolution, and processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. It analyzes the challenging issues in the data-driven model and also in the big data revolution.

S.Castro, et.al [2] enhancesa security and privacy in mobile data centers is challengeable with efficient security key management. The proper security standards of algorithms can be activated using quantum cryptographic key management scheme and which has been interfaced with the data node. Each type of data has been further classified to provide adequate security and enhance the overhead of the security system.

Bhagyshri Adhau, et.al [3] suggests the conceptof pattern based filtering in which it automatically discovers new, hidden or unsuspected data from the large text collection. The propose model consist of topic distribution describing topic preference of each collection of document and Pattern-based topic representation.

Puneet Goswami, et.al [4] illustrates security and privacy is the important concerns with data. However, there exists incongruity between the big data security and privacy and the extensive use of big data. These insights on overview of big

data, associated challenges, privacy and security concerns and differentiates between privacy & security requirements in big data.

Ptiček, et.al [5] provides an overview on research and attempts to incorporate map reduce with data warehouse in order to empower it for handling of big data. Big data analytics has become a very active research area in the last few years, as well as the research of underlying data organization that would enhance it, which could be addressed as big data warehousing. Research direction is enhancing data warehouse with new paradigms that have proven to be successful at handling big data using map reduce paradigm.

Sushmita, et.al [6] focused on big data refers to datasets high in variety and velocity, so that difficult to handle using traditional tools and techniques. The process of research into massive data to reveal secret correlations named as big data analytics. Big Data is a data whose complexityrequires new techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Hadoop as the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes.

Neha Begam, et.al [7] improves on potential fortune of non-traditional data got from various sources, for instance, web based systems administration, messages, online diagrams, web based shopping that can be borrowed for accommodating data. With the decreasing in the cost of capacity limit and figuring, it has wound up workable for dares to use this data to benefit. This leads learning into the boundless perspective of Big Data.

Jayesh Surana, et.al [8] propounds privacy and security concerns and privacy in big data is divided into three stages - data generation, data storage and data processing. It covers some traditional methods adopted for privacy in big data, the challenges faced by these techniques. The goal is to study the recent techniques adopted for privacy and draw their comparison in order to declare the most efficient technique among all of them.

Gayathri Ravichandran, et.al [9] describes a formal definition of Big Data and look into its industrial applications. Further, understands how traditional mechanisms prove inadequate for data processing due to the sheer volume, velocity and variety of big data. We will then look into the Hadoop Architecture and its underlying functionalities. This will include delineations on the HDFS and MapReduce Framework.

M.Chalapathi Rao, et.al [10] enhances the massive volume of big data sets are toocomplicated to be managed and processed by conventional relational databases the term “Big Data” was coined to address this massive volume of data storage and processing. The quality of captured data can vary greatly, affecting accurate analysis. Protecting privacy is mechanism for data processing and producing right information to favor corporate sectors, business managers, stake holders and other users make highly informed business decisions.

III. CHARACTERISTICS OF BIG DATA

Big data features are generally construe by 5Vs namely Volume, velocity, variety, veracity and value.

Volume: Due to the massive amount of data in the form of text, videos, music and image files is now stored in terms of exabyte in different enterprises. The sheer volume of data being generated makes the issue of data processing a complicated task. Big data stores huge volume of data which dealt with efficiently.

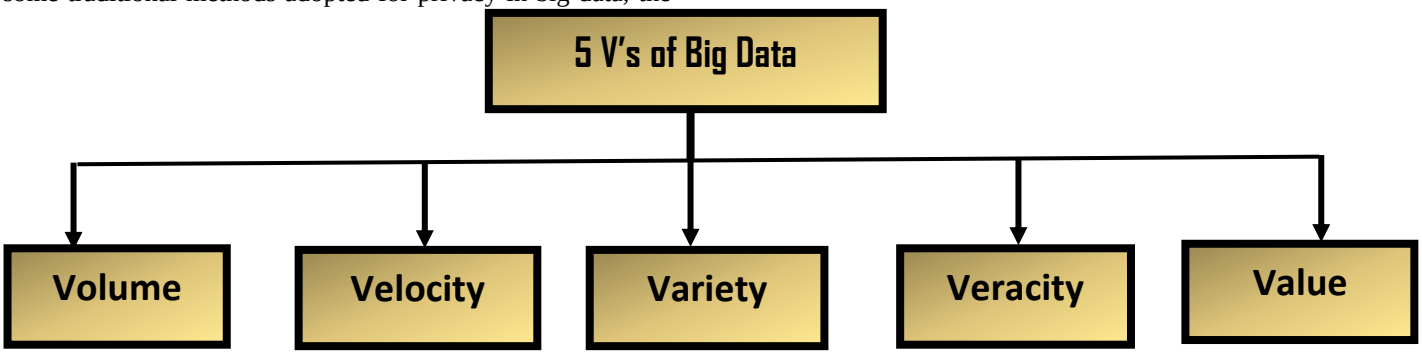

Figure 1: Five V’s of big data

elocity: Velocity refers to the rate at which data is being processed. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

Variety : Data is being generated from various sources includes social media, stock exchange and black box. Most of the data generated today is in unstructured manner. Furthermore, the data can assume various forms numerals, text, media files, etc. Using Big Data tools clusters and manages all structured, semi structured and unstructured data.

Veracity: Veracity refers to cleanliness or trustworthiness of data. Many of the data lacks quality and accuracy. For example twitter posts with hash tags, abbreviations and so on. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks.

Value: Value gives importance to the profit gained by organizations who investin big data technologies. It refers to the ability of companies to analyze data and to provide a

better understanding of the various key areas that include customer behavior, personalized services, and so on. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

Big data characteristics: HACE theorem

Big data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data. The most important characteristics of big data according to HACE theorem are listed below,

Huge Data with Heterogeneous and Diverse Dimensionality: One of the fundamental characteristics of the Big Data is the huge volume of data and heterogeneous diverse data from any number of sources, largely unknown and infinite, and in multiple formats. This heterogeneous huge volume of data comes from various social media like twitter, facebook, orkut and LinkedIn etc.,

Autonomous Sources with Distributed and Decentralized Control: Autonomous data sources with distributed and decentralized controls are one of the important characteristic of big data applications. It collects information without involving any centralized control. For e.g., In World Wide Web setting, each web server gives a certain amount of information and function individually without depending on others.

Complex Relationships: While the volume of data increases, so do the complexity and the relationship of the data also increases. In an early stage, it focuses on finding best feature values to represent each observation. It characterize each individuals of data fields, such as name, age, gender, income, etc.,

Evolving Relationships: The complex data which is nonlinear and has many to many data relationships evolve. For instance, customer comments on a website. This type of data must be gathered over significant periods of time in order to find out patterns and trends.

IV. BIG DATA LIFE CYCLE

A big data life cycle can be classified in the following stages:

Business Problem Definition:

It defines the problem and evaluate correctly by how much potential gain it may have for an organization.

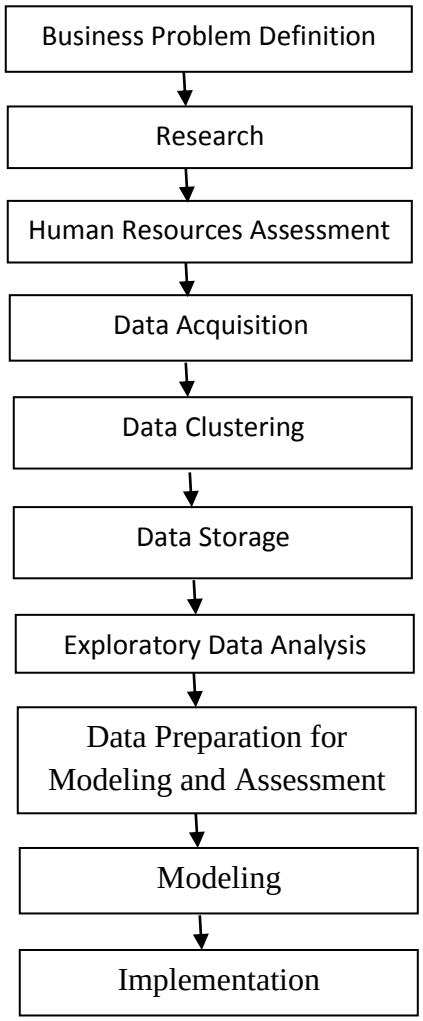

Figure 2: Big Data Life Cycle

Research:

Analyze for solutions that are reasonable for organization, though it involves adapting other solutions to the resources and requirements. In this stage, a methodology for the future stages should be defined.

Human Resources Assessment:

Once the problem is defined, it's reasonable to continue analyzing in order to complete the project successfully. It provides an optimal solution to all the stages, before starting the project if there is a need to outsource a part of the project.

Data Acquisition:

Data Acquisition stage is key in a big data life cycle; it defines which type of profiles should be needed. It also involves gathering unstructured data from different sources and generally requires a significant amount of time to be completed.

Data Clustering:

Suppose one data source gives reviews in terms of rating in stars, therefore it is possible to read response variable $x \in \{1, 2, 3, 4, 5\}$. Another data source gives reviews using two arrows system, one for up voting and the other for down

voting. This would imply a response variable of the form x ∈ {positive, negative}.

Data Storage:

Once the data is processed, it needs to be stored in a database. Big data technology uses the Hadoop File System for storage that provides users a limited version of SQL, known as HIVE Query Language.

Exploratory Data Analysis:

The data exploratory is mandatory once the cluster and storage of data in a way that insight retrieve from it. Exploratory data analysis is to understand the data which normally done with statistical techniques and plotting the data.

Data Preparation for Modeling and Assessment:

It involves reshaping cleanse data retrieved previously and using statistical preprocessing for missing values imputation, outlier detection, normalization, feature extraction and selection.

Modeling:

It involves different models and looking forward to solving the business problem. Finally, the best model or combination of models is selected evaluating its performance on a dataset.

Implementation:

In this stage, the data product has developed and also involves setting up a validation scheme while the data product is working, in order to track its performance.

V. BIG DATA FRAMEWORK: A PARALLEL PROGRAMMING WITH HADOOP AND MAP REDUCE

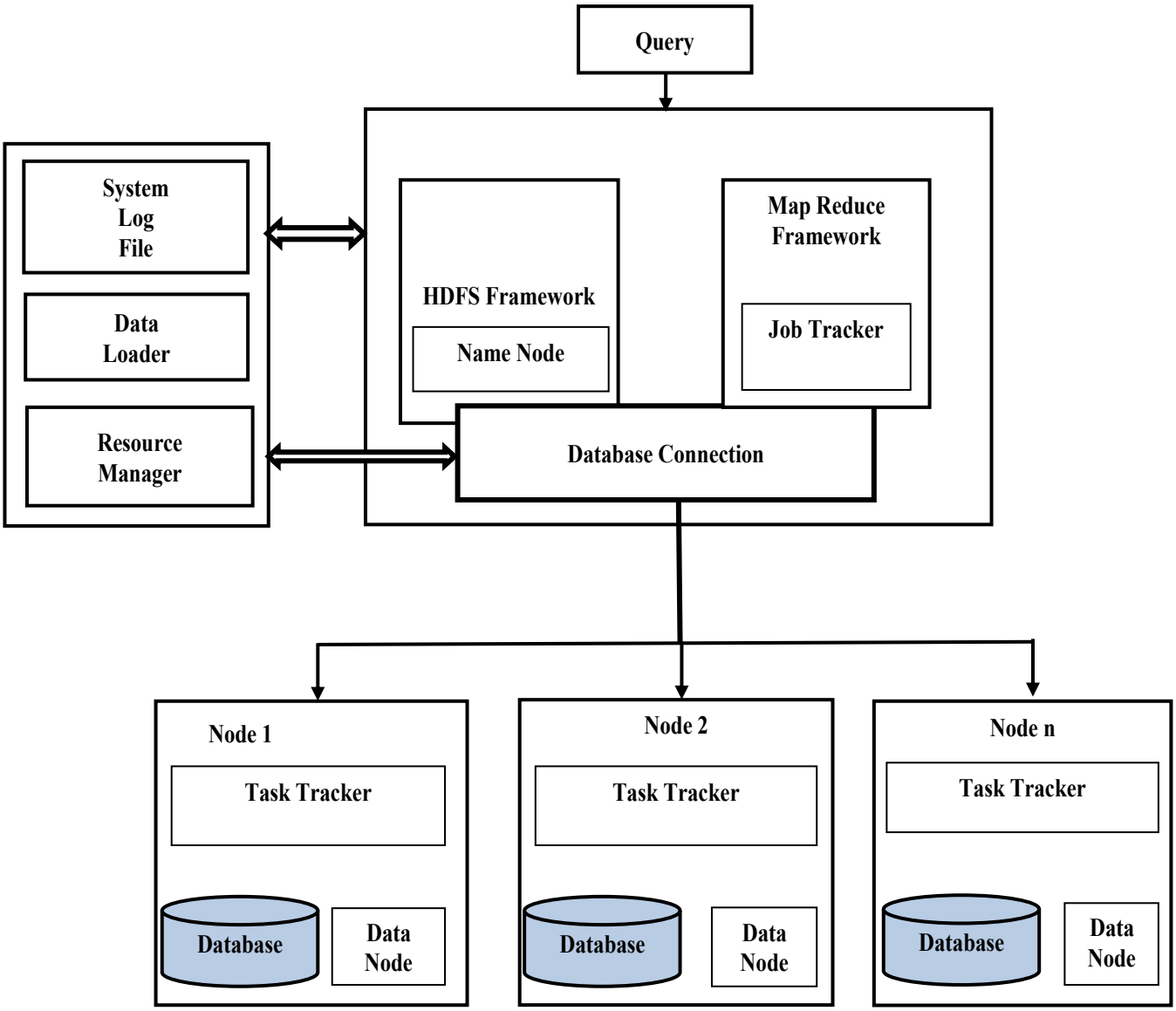

Figure 3: Big Data Hadoop Architecture

The Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It consists of two layers: a data storage layer Hadoop Distributed file system called (HDFS) and a layer called Map Reduce for data processing. HDFS performs storage function, while Map Reduce performs processing and sorting function.

Hadoop Distributed File System (HDFS)

Hadoop uses the Hadoop Distributed File System is an open-source software framework which is written in Java for distributed storage. It is designed to run parallel processing of a large number of records distributed across clusters of computers with simple programming models. It is used for storing large files with streaming data access patterns and provides file permissions, authentication, and streaming access to system data. There are two types of nodes in HDFS cluster, namely name node and data node.

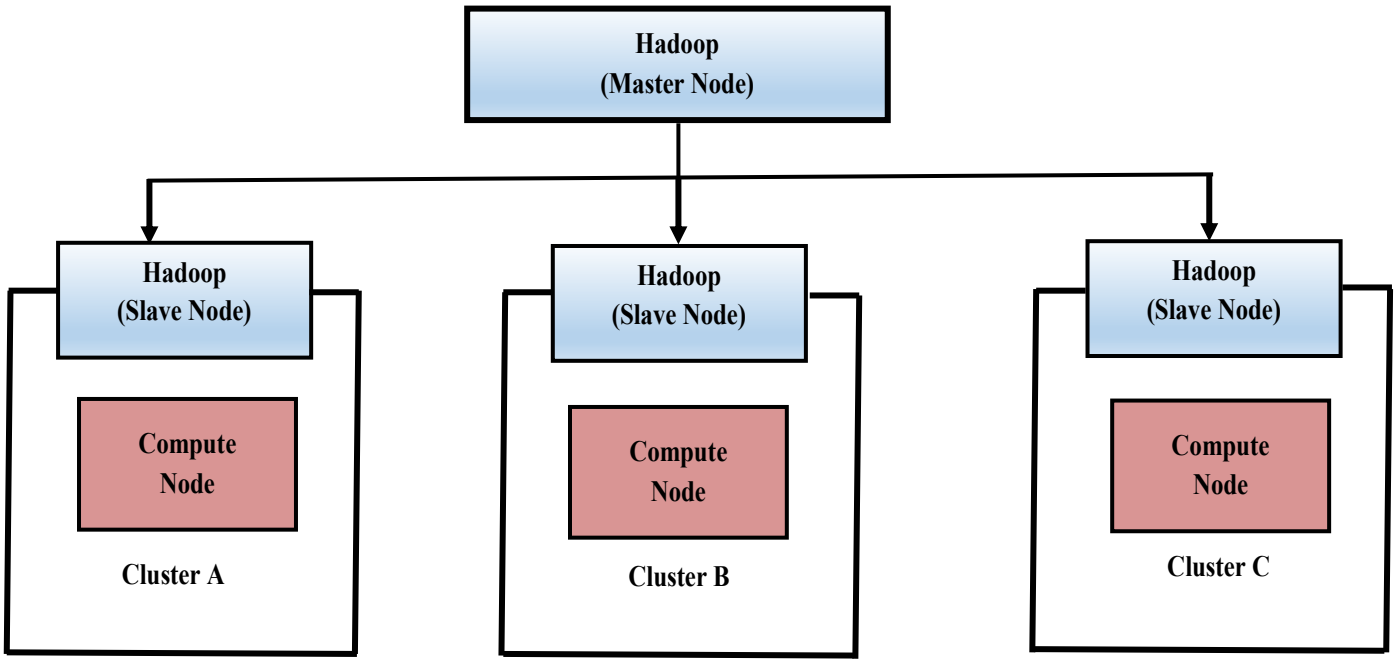

Figure 4: HDFS Framework

1. **Name node**

   The HDFS consists of a single name node, which acts as the master node. A file system namespace consists of a hierarchy of files and directories, where users can create, remove or move files based on their privilege. It controls and manages the file system namespace.

2. **Data node**

The HDFS consists of more than one data node. It is responsible for performing read and write operations from file systems as per client. The data node stores and retrieve blocks as per the instructions of clients or the name node.

Hadoop Map Reduce Framework

Hadoop uses the Map Reduce framework for distributed computing applications to process large amounts of data. It is the widely used parallel processing programming model and played a significant role in processing big data. It consists of two functions, map () and reduce () functions. Hadoop is also based on Master/Slave communication model.

1. **Map stage**

   The map function takes in a set of data as the input, and returns a key-value pair as the output. The input may be in the form of a file or directory. The output of the map stage serves as input to the reduce stage.

2. **Reduce stage**

   The reduce function will combine the data tuples into a smaller set. The map task always proceeds the reduce task. The output of reduce stage is stored in the HDFS.

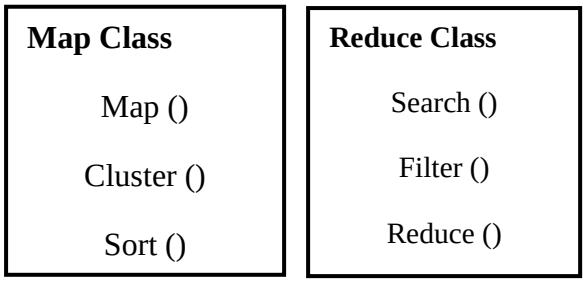

Figure 5: Map Reduce Function

The idea behind Map Reduce is that Hadoop can first map a large data set, and then perform a reduction on that content for specific results. A reduce function can be thought of as a kind of filter for raw data. Mapper class takes the input, tokenizes it, maps and sorts it while the reducer class which in turn searches matching pairs and reduces them.

The Hadoop MapReduce with master/slave architecture has JobTracker and TaskTracker. JobTracker is a master server in the MapReduce framework and responsible to manage the control flow of running MapReduce jobs.

It splits the job into tasks and assign those tasks to TaskTracker of slave. The TaskTracker accepts the job from JobTracker and report the health status to the JobTracker using heartbeat message. HDFS performs monitoring the HDFS file structure, locations, and the updated files. Similarly map reduce performs monitoring the list of applications, configuration of nodes, application status, etc.

VI CONCLUSION

Big data is not simply a data but it involves the large pool of data processed and generated by variety of gadgets, devices or applications. As a result, data are mined for deeper insights and requires advanced tools, techniques and frameworks rather traditional computing techniques. . The machine learning algorithm for big data needs to be more robust and easier to use. It overcomes the traditional ETL process into Extract-Load-Transform (ELT) process as big data and gives more efficient services. Big data challenge is to explore the large volumes of data and extract useful information or knowledge for future actions. This paper examined about importance of big data concept, characteristics and benefit of it. The availability of big data, low-cost commodity hardware, and analytic software has shaped a unique moment in the history of data analysis. The main objective of big data analytics in business intelligence is to enhance decision making proficiency, understanding of customer demands, innovative and developing plans, exploring and improving services. As a result the data are mined for deeper insights and improve various services in business functions and provides customer satisfaction, privacy and security. The benefits that big data analytics brings the ability to work faster and stay agile gives organizations a competitive edge they didn't have before. In future due to enormous information contains delicate and private data, so as to secure this huge volume of data is most challenging in big data.

REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding (2013), 'Data Mining with Big Data ', IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 1041-4347/13.

[2] S.Castro and R.Pushpalakshmi (2017), 'A Survey on Big Data Security and Related Techniques to Improve Security', Asian Journal of Applied Science and Technology (AJAST), Volume 1, Issue 5.

[3] Bhagyshri Adhau and Dr. V. T. Gaikwad (2017),'Pattern Based Filtering Approach for Big Data Application', IJSRSET 1732160.

[4] Dr. Puneet Goswami and Suman Madan (2017), 'A Survey on Big Data & Privacy Preserving Publishing Techniques', Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 3.

[5] M. Pticek and B. Vrdoljak (2017),' MapReduce Research on Warehousing of Big Data' MIPRO.

[6] Sushmita, Simranjeet Kaur, Simranjot Kaur (2017), 'A Review paper on BIG Data' ,International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT) Volume 2, Issue 3 , ISSN : 2456-3307.

[7] Neha Begam, Sandhya, Er.CK Raina (2017),' Big Data Challenges and Techniques', International Journal of Engineering Science and Computing (IJESC), Volume 7, issue no: 4.

[8] Jayesh Surana, Akshay Khandelwal, Avani Kothari,Himanshi Solanki and Meenal Sankhla (2017), 'Big Data Privacy Methods', International Journal of Engineering Development and Research IJEDR1702165.

[9] Gayathri Ravichandran (2017), 'Big Data Processing with Hadoop : A Review', International Research Journal of Engineering and Technology Volume: 04 Issue: 02.

[10] M.Chalapathi Rao and A.Kiran Kumar (2017), 'Challenges arise of Privacy Preserving Big Data Mining Techniques', International Research Journal of Engineering and Technology Volume: 04 Issue: 05 e-ISSN: 2395 -0056.

[11] Priya Chaudhari and Binita Patel (2017), 'Future of Big Data', International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 1, e-ISSN: 2395 -0056.

[12] Shrikant Rangrao Kadam and Vijaykumar Patil (2017), 'Review on Big Data Security in Hadoop', International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 01, e-ISSN: 2395 -0056.

[13] Neha D. Patil and Dr. D. S. Bhosale (2017),' Providing highly accurate service recommendation for semantic clustering over big data', International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 02, e-ISSN: 2395 -0056.

[14] Dipti Shikha Singh and Garima Singh (2017), 'Big data – A Review', International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 04, e-ISSN: 2395 -0056.

[15] Pwint Phyu Khine and Wang Zhao Shun (2017), 'Big Data for Organizations: A Review', Journal of Computer and Communications, 40-48.

[16] Asha Patel (2017), 'A Survey Paper on Security Issue with Big Data on Association Rule Mining', National Conference on Latest Trends in Networking and Cyber Security (IJIRST/Conf/NCLTNCS/2017/025).