*International Journal on Emerging Technologies* (Special Issue NCETST-2017) **8**(1): 48-50(2017)

*(Published by Research Trend, Website: www.researchtrend.net)*

**ISSN No. (Print) : 0975-8364**
**ISSN No. (Online) : 2249-3255**

# Big Data-Challenges and opportunities: A Review Paper

***Ms. Preeti Pandey, Dr. Sadhana Rana and Ms. Garima Joshi***
*Assistant Professor-CSE, Amrapali Institute- Haldwani, (Uttarakhand), India*

**ABSTRACT: As we are living in the world of technology, everyday there is a new innovation to enhance the drawbacks of the older technology. With the advancement of technology and devices, the rate at which data is growing day by day is hard to manage. The term 'Big Data' in itself describes the flow of information in the real world. This king size of data comes from everywhere, every day world widely. Big data can be in structured, unstructured, distributed form. To manage big data we need certain tools and techniques. This paper presents the overview of the concept big data, Issues, challenges in this field, tools & techniques.**

**Keywords:** Big data, Hadoop framework, challenges, opportunities in big data

## I. INTRODUCTION

Big data is the large and complex volume of data sets that are difficult to manage using traditional tools and techniques. "Big Data" the huge amount of data collected from everywhere in this universe comes with new opportunities and challenges for researchers and technical people. Figure1 shows the different areas from where huge amount of data is producing every second. Data may be in any format such as images, text, audio, video, numbers etc.

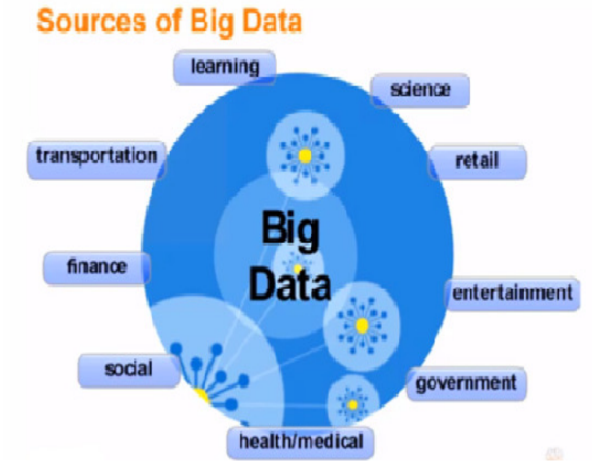
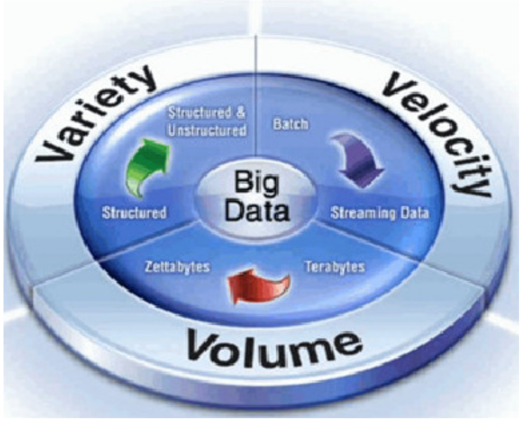


**Fig.1.** Sources of Big Data.

Big data has three main characteristics:

- Volume of data means the amount of data stored in different databases/servers. This data is growing from megabytes to petabytes and from petabytes to exabytes.
- Velocity of data means the data generation rate. As this is the era of digitization, the data generation rate is going high and increasing the volume of data.
- Variety of data means data may be in any format due to variety of data generation sources.



Big Data involves the study of technologies and mathematical approaches which focuses on the acquisition, condition, analysis and evaluation of large volume of data.

## II. CHALLENGES WITH BIG DATA

Some of the facts available to explain the situation of data are as follows. In 1998 when Google came into picture it had the figure to process near about 10,000 search query per day but now it has reached the figure of more than

40,000 search queries per second. Moreover more than 42,000 GB of data flow in a single second and figures doesn't stop here. Users of Instagram share near about 40 million photos in a single day. Facebook has 30 Petabytes of data generated by its users to process and manage. IBM indicates that 2.5 Exabyte data is created everyday which is very difficult to analyze. These figures are enough to explain the amount of Big Data circulating everywhere. To store, retrieve, manage this amount of data is not an easy task. Some very challenging question arises regarding big data that when we have plenty of data what data should be analyzed, how to find out the relevant data out of big data, how to achieve best data to to get advantages and so on. There are certain challenges while managing Big Data. Some are discussed below:

I. **Size of data:** Big data is a collection of very large size of data sets. It is really problematic to handle such huge amount of data available. Technologists worked on processor's speed so that processing can be improved. But with the usage of internet data is becoming more and more large in size.

II. **Processing speed:** As data volume is very high, processing speed is very important. For example in social media analysts have to analyze the interest of users after examining various posts. In any situation when we need results immediately, processing speed matters.

III. **Data Privacy and Security:** Data privacy becomes very important in terms of securing user's data. Cloud computing is becoming so popular now a days. Data has been stored on virtual servers and thus becomes important to provide protection to data.

IV. **Timeliness:** Timeliness is related to size of data. If the volume of data is high (as in the case of Big Data), it will take more time to process the data. Moreover it will take more time to analyze data.

V. **Heterogeneity of data:** As data comes from different, known and unknown resources, sometimes it is structured and sometimes unstructured. Structured data is easy to manage but unstructured data itself comes with lot of difficulty in accessing , storing and managing.

## III. HADOOP -A TECHNIQUE TO HANDLE BIG DATA

To overcome all the challenges faced to manage the Big data sets, a programming framework came into picture named as Hadoop. Hadoop was started by two Yahoo employee, Doug Cutting and Mike Cafarella, in 2006.It is developed by Google's MapReduce as a software platform and also it is an open source framework. In the past few years, Hadoop has taken its position as the most powerful framework for data storage and processing.

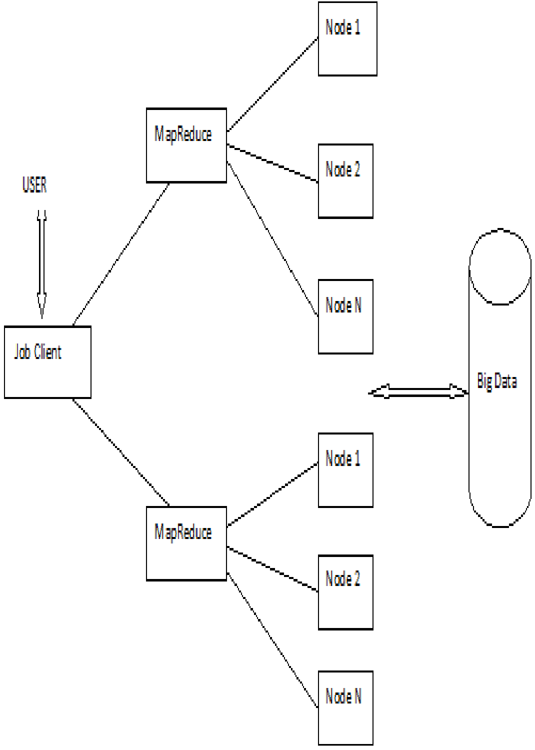


**Fig. 2.** Architecture of Hadoop.

In Hadoop large data sets are processed in a distributed computing environment. An Apache Hadoop system consists of the Hadoop Kernel, Map Reduce, HDFS.
There are two main components in Hadoop:

(i) *Hadoop Distributed File System:* HDFS is a distributed storage that is best suited for the applications that contains large data sets. HDFS follows master-slave architecture and can store data sets coming across the servers from various locations.

(ii) Map Reduce: Map Reduce is a programming model. It works on large data sets, fragments the problems into smaller fragments(lager data sets into smaller data sets) and runs them parallel. Map Reduce is a mix of two functions. First one is map function which is used to filter and transform the data sets. It divides the input into small parts and assigns those parts to many computers connected over the network. Second function is reduce function, is used to collects the results generated from servers/computers and reduce it to final result.

## IV. OPPORTUNITIES TO BIG DATA

Big data has gain lots of importance in last few years. It has open the door of opportunities
for researchers as well as many business organizations so that they can expand their business and can manage huge data also. Big data is playing important role in many fields for example medical, banking, e-commerce, government sectors, private sectors etc.

1. **Social media:** Social media like Facebook, twitter, LinkedIn have to deal with a huge amount of structured and unstructured data. Facebook handles more than 60 Billion photos of users. Social media collects lots of data every minute and have lot of opportunities.
2. **Medical:** Healthcare field is also having plenty of data to analyze. Patient's everyday records, online offline consultation, patients history etc are used to get the desired results. This area has gained lots of research importance.
3. **Research and Development:** Big data is a latest topic of research. Many researchers are working on big data. There are so many articles and research papers being published on big data.
4. **Science and Technology:** Almost all top organizations are investing on Big data.
5. **Education :** Can improve students performance and learning abilities making the lessons more personal. The courses can be adjusted from the teachers with the help of analytics.

Big data is really working as a helping tool for the decision makers. It is analyzing data coming across mobile devices, web, online services, social media and hence improving the profit of an organization.

## V. CONCLUSION

Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data. In this review paper, we focused on the basic concept of big data, characteristics of big data, working of Hadoop and its applications. Hadoop has an important role in the processing and managing of big data. This open source software platform managed by the Apache Software Foundation has proven to be very helpful in storing and managing vast amounts of data cheaply and efficiently. Lot of research and experiments are going on this area so that we can store and manage the current data as well as future data efficiently.

## REFERENCES

[1]. Rahul Beakta, "Big Data And Hadoop: A Review Paper", **2015.**

[2]. Bijesh Dhyani, Anurag Barthwal, " Big Data Analytics using Hadoop" International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014.

[3]. www.researchgate.net.

[4]. Nirali Honest and Atul Patel,"A survey of big data analytics", *International Journal of Information Sciences and Techniques* (IJIST) Vol.**6**, No.1/2, March 2016.

[5]. Shilpa* Manjit Kaur, " BIG Data and Methodology-A review" Volume **3**, Issue 10, October 2013

[6]. Harshawardhan S. Bhosale1, Prof. Devendra P. Gadekar," A Review Paper on Big Data and Hadoop", *International Journal of Scientific and Research Publications,* Volume **4**, Issue 10, October 2014.

[7]. Athanasios S. Drigas and Panagiotis Leliopoulos, "The Use of Big Data in Education", *IJCSI International Journal of Computer Science Issues,* Vol. **11**, Issue 5, No 1, September 2014.

[8]. Prity Vijay, Bright Keshwani,"Emergence of Big Data with Hadoop: A Review", *IOSR Journal of Engineering(IOSRJEN)*, March 2016.

[9]. Nisha Bhardwaj, Dr Balkishan, Dr Anubhav Kumar, "Big Data and Hadoop: A Review", *IJIRSET*, June 2015.