International Journal of Scientific and Innovative Research 2018; 6(1) : 23 - 26
P-ISSN 2347-2189, E- ISSN 2347-4971

BIG DATA: ISSUES AND CHALLENGES

*** Rajeev Srivastava**

Department of Decision Sciences, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India

* Address for correspondence: Dr. Rajeev Srivastava,Head, Department of Decision Sciences,University of Petroleum and Energy Studies, Dehradun,Uttarakhand,India;
Email ID: rajeevspn21@yahoo.com

ABSTRACT

Data is growing with tremendous rate not only in the form of volume but also in different formats mainly semi-structured or unstructured. With the growing use of social networking websites, use of smart phones this growth will increase day by day. Companies are showing their interest not only to store that data but also to analyze that data to extract knowledge from that data. This growing data is in the form of audio, video, images, text and transactional data. So, we can say thatBig data is collection of huge amount of data which is complex in nature. To analyze that huge amount of data requires huge storage, the use of high speed parallel and distributed computing and high ended analytical tools. There are various issues and challenges to handle the volume, velocity, distributed nature of data, and its analysis poses challengesin data analytics. This paper covers various challenges and issues to handle that Big Data for various organizations.

Keywords: *Big Data; Data Analytics; Social Media; UnstructuredData*

INTRODUCTION

First time the term "Big Data" was used by John Mashey in 1998 in his slide during presentation. It is used in the form of the data sets, so large and cannot be processed with existing traditional data,management tools. [1]Many authors have described "Big Data" in their own words. The summary for definition given by different authors is given in Table 1.

Table 1: Various definitions of Big Data

Author	Definition
S. Kaisler et al.	"Big Data" is the amount of data which is beyond technology's capability to store, manage and process efficiently.
J. M. David et al.	"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
Gartner	Gartner's describes Big Data as having three dimensions: volume, variety, and velocity
IDC	IDC defined "Big data technologies" as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery, and/or analysis."
IBM	Although IBM defines Big Data with four characteristics by adding veracity to the above 3 characteristics. Other Big Data Researchers defined it with 5 Vs by adding Value.

Big data analytics can be defined as the process of collection, organizing collected data and examining that huge amount of the data to know the hidden pattern or other useful information, so that better decision making can be taken. With the big data analytics techniques, data scientists can analyze huge volume of data that cannot be possible with business intelligence or conventional analytics tools. Jangalaet al. [1]have categorized the big data analytics in two categories: Stream processing and batch processing. In stream processing, data comes in streams and it is processed as soon as possible to generate results. Storm and Apache kafka are popular stream processing models. In batch processing, first data is stored and then processed. MapReduce is popular batch processing model.

Google, eBay and LinkedIn were among the first to experiment with big data. Initially they used big data analytics to check if analytical model can be improved and they indeed got positive results. According to studies reported by Julie M. DavidandKannanBalakrishnan[2], we are creating about 2.5 quintillion bytes of data and 90% of the data in the world today has been created in the last two years alone and it is going to double in every two years [3]. Most of the data is created by individual users via social media and other sources of data are sensors data used to gather climate transaction records, digital pictures and videos, transactions records and cell phone GPS signals etc. This data is big data.

Now a day around 75% of data is generated by social media and most of that data is unstructured. As per the forecast of the IDC, the market of Big data technology and its services has multibillion dollar worldwide opportunity and is expected to grow at the growth rate of around 23.1 percent compounded annuallyand will reach $48.6 billion in 2019. This growth will be approximately 6 times the growth of the IT market. According to the recent IDC forecast, Big Data technology and services market represents a fast-growing multibilliondollar worldwide opportunity and it will grow at a 23.1% compound annual growth rate to $48.6 billion in 2019, or about six times the growth rate of the overall information technology market. IDC also believes that by 2020, real- time intelligence growth will be in double digits and exploration/discovery of the unstructured worlds [10]. Most of the companies are using online Big data tools to handle big data for reducing their IT cost, which on one side reducing the cost but on the other side it effects the security and privacy of databecause the data is hosted on third party infrastructure[11].

ISSUES OF BIG DATA

According to JASON [8] still there is no universally accepted method to store and reduce Big data. There is needto make robust, open source and platform independent solution for this problem. Finally, we can say that there is no perfect Big data management solution yet. This is one of the important gap in the literature that still needs to be filled [4].

Storage of Data

Keeping in view the increase in the use of mobile phones especially smart phones and social networking websites the huge amount of unstructured data is accumulating at very high speed. However, the invention of new storage medium is going on but speed at which the accumulation of new data is going on is very high as compared to new storage invention [4].

Transmission of Data

There are two basic methods of transmitting data. In first method, collected data transfer to the processing point. In second method, the collected data remain in place while code to process that data transfer to collected data and then the final result transferred. However, in both the cases integrity and provenance need to be transferred along with the actual data [4].

Security and Privacy

In the process of collecting, analyzing and mining for the meaningful pattern information security is the major problem. Because an unauthorized client may get access to that data, modify it or change the priority of the job [3]

Data validation

Data quantification or finding missing data and outliers from a huge amount of unstructured data is almost impractical, so there is urgent need for new data validation and quantification techniques.

Processing Issues

Processing of huge amount of unstructured data is also one of the major issue. There is need to develop new analytical algorithms to effectively process that huge amount of data in order to

perform quick decision making [4].

CHALLENGES OF BIG DATA

The problem related to storage of Big data has been resolved, but tools required for analytics of that data are still missing. So development of these tools for extraction of knowledge from that data is big challenge [2,7]. Challenge in terms of incompatible formats of data, non-aligned structure of data and inconsistency in data semantics also need to be considered [4].

Huge data generated examples

As per the recent study conducted it has been found that every minute huge amount of data generated by different sources like Google receives around 4 million queries, around 200 million messages send through e-mail, 72 hours of video uploaded in YouTube, users of Facebook share 2 million structured and unstructured content and around 277,000 tweets generated by Twitter[56].

Unstructured data

Most of the data collected now a day is unstructured and the major problem with that unstructured data is that it's not easy to categorize and analyzed [1].

Timeliness

Increase in the amount of data also demands faster processing of that data otherwise it will take too much time.So there is need to develop a system which can process huge amount of data timely [9].

Heterogeneity and Incompleteness

The data consumed by human beings are normally heterogeneous because it's easily understood by them. However,the datarequired by machine learning algorithm demands homogeneous data. So before doing analysis of heterogeneous data it requiresfurther work to make it homogeneous for meaningful results [9].

Inconsistencies

Big data cloud collect and analyze all domains data like Spatial, manufacturing, physical sciences etc. This combination of data from different domains produces inconsistency in data. So inconsistency at the level of data, information and knowledge level need to be addressing properly [12].

CONCLUSION

Handling huge amount of structured and semi-structured Big data using traditional RDBMS system is a big challenge. This paper presented the various issues and challenges related to Big data storage, processing and analysis. Some of these challenges can be easily overcome but few required great concern in terms of tools and techniques, cost, software and hardware. Therefore, there is need to support and encourage fundamental research towards addressing these technical challenges if we are to attain the assured benefits of Big Data.We have tried to discover the issues and challenges that Big Data is facing from data storage and analytics perspectives. Some of the challenges that we have mentioned can easily be overcome. These practical challenges are common across a large variety of application domains, and consequently not cost-effective to address in the context of one domain alone. Therefore, there is urgent need to focus on these technical challenges to fully utilize the potential of Big data.

REFERENCES

1. K. S. Dr. Jangala., M. Sravanthi, K. Preethi and M. Anusha, "Recent Issues and Challenges on Big Data in Cloud computing," *IJCST , ,* vol. Vol. 6, no. Issue 2, April - June 2015.
2. Julie M. David, KannanBalakrishnan, "Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets," *Int. J. of Computer and Electrical Engineering, Hong,* vol. 3(1), pp. pp163-169, 2011.
3. K. U. Jaseena and J. M. David, "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING," *Computer Science & Information Technology (CS & IT),* no. pp. 131–140, 2014.
4. S. Kaisler, F. Armour and J. A. Espinosa, "Big Data: Issues and Challenges Moving Forward," *Hawaii International Conference on System Sciences,* no. 46th, 2013.
5. J. K. U. and J. M. David, "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING," *Computer Science & Information Technology (CS & IT).*
6. "http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/," [Online].
7. A. A. TOLE, "Big Data Challenges," *Database Systems Journal,* vol. vol. IV, no. no. 3, pp. 31-40, 2013.

8. JASON, "“Data Analysis Challenges”," *Mitre Corporation, McLean, VA, JSR-08-142,* 2008.
9. "Challenges and Opportunities with Big Data," *A community white paper developed by leading researchers across the United States.*
10. "Australian Department of Immigration, “Fact sheet 70 - managing the border,” Internet, 2013. [Online].," [Online]. Available: Available: http://www.immi.gov.au/media/factsheets /70border.htm.
11. M. M. Gaber, A. Zaslavsky and S. Krishnaswamy, "“Mining data streams: A review,”," *ACM SIGMOD Record.,* Vols. vol. 34, no. 2, p. pp. 18–26, 2005.
12. S. J. Samuel, K. RVP, K. Sashidhar and C. R. Bharathi, "A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES," *ARPN Journal of Engineering and Applied Sciences,* Vols. VOL. 10, NO. 8, MAY 2015.