ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 05, pp.51124-51127, May, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

BIG DATA: TOOLS, CHALLENGES AND FUTURE TRENDS

***Kumar Harsh**

Adityam Digitech, GSS, Varanasi, India

ARTICLE INFO

Article History:
Received 24th February, 2017
Received in revised form
21st March, 2017
Accepted 04th April, 2017
Published online 31st May, 2017

Key words:
Big Data,
Hadoop, Map Reduce,
Hive,
NoSQL and Cloud Computing.

ABSTRACT

Today data is everywhere, increasing itself in a very fast pace every second, minute, hour and day. A huge amount of data is generated by social media, broadcasting videos and audios, business transactions, sensor data, digital technologies like internet of things and cloud computing etc. This large amount of data generated is called Big Data and it has become a part of our economy, business, organization and individual. This paper presents the analysis of what the big data actually is and what are the tools to handle the big data. It also focuses on future scope and trends of Big Data.

Citation: Kumar Harsh. 2017. Big data: tools, challenges and future trends", *International Journal of Current Research*, 9, (05), 51124-51127.

INTRODUCTION

Today one cannot skip the most hyped topic in the computer world Big Data. Just like the big bang explosion in the universe there is a explosion of data coming from various sectors today. Data is increasing day by day with a great speed and it is evident that the amount of data produced in the last two years is equal to the total data produced in the entire history of mankind (https://www.forbes.com/sites/bernardmarr /2015/09/30/big-data-20-mind-boggling-facts-everyone- must-read/#58987f7b17b1). Big data does not have a specific definition it can be defined in a variety of ways. It cannot be defined that how much amount of data is big data i.e. is 500 GB is a big data or 100000 GB is a big data or some other quantity. Big data can be defined in one way as extremely large data sets coming from different sources and in different forms which can be analyzed and processed with the help of different tools to reveal patterns, trends, insights about human behavior and interaction which will lead to better decision making and planning strategic moves in various sectors. Big data is very relative and continuously evolving. A big data five years ago is not a big data today. Now, what size of data is big data? if there is a capacity of a laptop to store 500 GB of data but the incoming data exceeds the storage and manipulation capacity then the data is big data for the laptop, if there is a large organization with large storage spaces but if data cannot fit in it then it is a big data so it means if the data goes beyond the storage and manipulation capacity of the system then it is big data (Soumya Shukla, 2015; http://www.guru99.com/what-is-big-data. html; Why is BIG Data Important, 2012). Enormous amount of data is generated every minute, recent studies shows that We are seeing a massive growth in video and photo data, where every minute up to 300 hours of video are uploaded to YouTube alone, everyday about 2.5 quintillion bytes of data is produced which is nearly equal to 10 million blue ray discs in a year (https://bigdatauniversity.com) and by the year 2020, about 1.7 megabytes of new information will be created every second for every human being on the planet (https://www. forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#58987f7b17b1).

Now days data is getting recognition and importance and decision making based on big data analysis is now being recognized broadly and there is a growing enthusiasm for BIG data. Big data analysis now drives nearly in every sector of our society.

Characteristics of Big Data

Characteristics of Big data helps to identify if a problem requires a big data solution or not. The different characteristics of big data are Volume, Variety, Velocity, Veracity and Value

Corresponding author: Kumar Harsh,
Adityam Digitech, GSS, Varanasi, India.

Volume

It refers to the huge and enormous size of the data. This data is generated by machines, sensors, social media, cloud computing, organizations, transaction etc. The data to be analyzed is massive. Due to the large volume of data, traditional relational database management systems fail to handle the big data hence different tools are used to store and analyze the data (6,8,9).

Velocity:

Data can be divided in two types i.e. Data at rest and Data in motion. Velocity refers to Data in motion for example the continues data taken from a sensor or a email being sent to someone. Velocity is the frequency of data coming from different sources that needs to be processed and analyzed.

Variety

With the increase in volume of data it is obvious that there will be a increase in variety of data coming from different sources. Data can be categorized in three types i.e. structured data, unstructured data and semi structured data. Structured data always follow a set of rules or have a predefined model e.g. relational databases, tables, spreadsheet, dates which follow specific patterns, names will have only text etc. Unstructured data on the other hand has no set of rules like pictures, videos, tweets etc (6,8,9). All these data are different but represents some behavioral aspect of human beings. Semi structured data are the mixture of both structured and unstructured data.

Veracity

It refers to the validity, reliability or accuracy of data. It also tells about the completeness of the data .It tells that the data being analyzed and mined is related to the problem being solved or not. Veracity is an important characteristic which tells about the truthfulness of the data.

Value

Ultimately the main purpose of doing any work is to derive value from it. We are making all the efforts of storing and analyzing big data to mine some value from it. At the end of the day all the efforts made must give some insights which must help the organization to grow. All other characteristic must be considered to finally obtain some value from the big data (Fig 1).

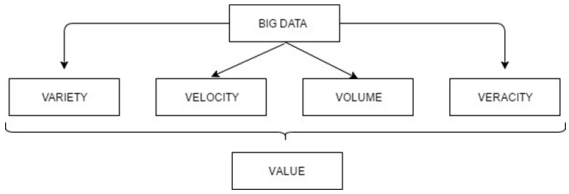

Fig. 1. Characteristics of Big Data

Tools for handling big data

The amount of data produced is large enough for the traditional methods to handle, store and analyze hence we need to use different tools to handle and store big data and do the analytics. There are thousands of tools present today in the market which are capable of doing different unique tasks.

The choice of tool may differ from person to person or the type of analytics one is going to do. Some of the major big data tools which are helping users to manage and utilize big data in a cost effective manner are explained below:

Map Reduce

Map reduce is a programming model using which we can generate and process Big data. It is the heart of apache hadoop framework. Map reduce is divided into two parts viz. map and reduce (Acharjya, 2016). The map stage is used for processing the input data line by line and produces several small chunks of data. The reduce stage is divided into two parts i.e. shuffle and reduce .In this stage data coming from map stage is processed to produce new list of reduced outputs (http://searchcloud computing.techtarget.com/definition/MapReduce).

HADOOP

It is a open source framework which is used to store, analyze and process big data based on commodity hardware .Hadoop consist of two main parts viz. Hadoop distributed file system (HDFS) and map reduce programming framework. HDFS is a file system used for storing big data and map reduce framework is used for processing it. Hadoop has major advantages like reliability, scalability, flexibility and low cost.

HIVE

Hadoop and map reduce framework were a major breakthrough in storing and processing big data but writing code in map reduce was very difficult and top quality skills are required for it. Professionals with such skills were very less and dependency on them was increasing day by day. This is where hive comes in .It provides a SQL like interface which is familiar to more people to extract data from hadoop systems hence increasing the audience. Hive also known as HiveQL converts SQL like quarries to java map reduce code to extract information from hadoop systems (http://searchdatamanagement. techtarget.com/ definition/Apache-Hive).

NoSQL

Nosql data-based are those databases which are not in tabular relation but are represented in some other means. Various data-based that fits into this category are column store, document store, key-value store, graph, multi model etc. Nosql databases are completely different type of databases as compared to traditional databases which provide high performance and quick as well as easy processing of data as a massive scale hence this database is satisfying heave demands of big data.

CHALLENGES TO BIG DATA

Size, Storage and Quality

As the name suggest 'Big', managing and storing large and increasing day by day data has always been as issue for many decades .today data in an organization is increasing exponentially and to store that data effectively and process it is becoming the real challenge. Companies are often using data lakes to store their vast amount of data (Agrawal et al., 2012; https://www.qubole.com/resources/solution/big-data-challenges/. Another main issue is the quality of the data

weather the data is incomplete or inaccurate or duplicate etc. The data which consist of mistakes incomplete values or duplicacy are often referred as dirty data. According to The Data Warehouse Institute (TDWI),dirty data causes a loss of 600$ to U.S. companies every year. Businesses and organizations must come up with ways to relies the cause of dirty data and fix it.

Technical Manpower to Combat Big Data

Not anyone can just hop in and start doing big data analysis. It requires unique and specific set of skills to become a data scientist. As big data is a new technology there is a shortage of data professionals and data scientist (Acharjya, Kauser Ahmed, 2016). There is a shortage of data scientist as well as carrying a data project to success require a full team with significant domain knowledge to find valuable insights.

Time consuming and scalability

With size there always comes speed. The larger the size of data will be the longer it will take to analyze and process it. Sometime there is a requirement of immediate result therefore tools may be developed to analyze big data in time efficient way. One other major issue in big data is the scalability of the data. A big data project may grow and evolve very quickly and many companies cannot cope with this situation (Soumya Shukla, 2015; Agrawal et al., 2012). As the data will increase organization will have to upgrade their infrastructure to cope with it. Pausing a project to improve infrastructure will again consume time and cost.

Security

Maintaining and storing the vast amount of data is a challenge but keeping the vast data secure is another. Security is an important aspect in big data analytics. Big data may contain sensitive information's like credit card details personal details (name, address etc.), health care details, bank account numbers, corporate information etc. (https://www.qubole.com/ resources/solution/big-data-challenges/). This data is sensitive in nature and if fall in the wrong hands can cause considerable damage to an organization and society so it must be effectively protected from unauthorized access.

Future of Big Data

Today Big Data have taken the business world by storm and is still growing at a very fast pace. Data is growing exponentially and by the year 2020 it will grow from 4.4 zeta byte to 44 zeta byte (https://www.forbes.com/sites/ bernardmarr/2015/ 09/30/big-data-20-mind-boggling -facts-everyone-must-read/#58987f7b17b1). By the year 2020 one tenth of the data will be produced by machines and internet of things will contribute 10% data of the earth. The number of connected devices is predicted to rise by 285 percent says Juniper Research (https://www.forbes.com/sites/ber nardmarr/ 2016/ 03/15/17-predictions-about-the-future-of- big-data-everyone-should-read/#21bfd8111a32; http://www.com puterweekly. com/news/ 2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off). Big companies are already aware of the future spectrum of big data and already invested in the technology. Gartner survey shows that nearly 75% of organizations today have already invested or planning to invest in big data in next two years. In the upcoming years businesses whether small or large will use big data analytics at some extent to increase their business and meet more demands in a more effective way. Today big data is used in various sectors such as business, government, health care industry, travel and hospital, scientific research, retail and many more. Originally hadoop was designed to process the data cluster presents on physical machines but with the advent of cloud computing this has changed .Now many technologies are available which can process data on the cloud. Cloud computing and Big data are highly compatible with each other as cloud computing allows storing and processing big volumes of data with high accuracy hence Big data will be extensively used with cloud computing .This increase in Big data will affect large and small organizations as well as every individual in the future.In the upcoming future more company will appoint data analyst and data scientist and ways of processing and analyzing data will increase and improve. In the future machine learning will be used to reveal greater insights of big data. Big data will help organizations to make their existing product and services more profitable. It helps companies to understand customer preferences and use this data to influence their buying habits and preferences. Big data will also help the organizations to maintain their security and secrecy to greater extent.In the future big data will affect healthcare, marketing, scientific research, travel and transportation to a great extent.

Conclusion

In todays world we are living in the era of Big Data. By processing and analyzing data in more efficient way we can open new gates to further advancements to various sectors and make organizations grow in a profitable manner. New and more efficient tools to analyze data must be developed for this purpose and the challenges discussed in this paper must be overcome by further researches. In future big data has a great potential to change the world's economy as well as individual life.

REFERENCES

Soumya Shukla ,Vaishnavi Kukade , Sofiya Mujawar, 2015. "Big Data: Concept, Handling and Challenges: An Overview " International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 11, March.

Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M., Widom J. 2012. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc-led-whitepapers

FORBES,https://www.forbes.com/sites/bernardmarr/2016/03/15/17-predictions-about-the-future-of-big-data-everyone-should-read/#21bfd8111a32

QUOBOLE,https://www.qubole.com/resources/solution/big-data-challenges/

Acharjya, D. P. Kauser Ahmed, 2016. "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2

Dataintensity,http://www.dataintensity.com/characteristics-of-big-data-part-one/

Guru99,http://www.guru99.com/what-is-big-data.html

IBM,https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/

Datasciencecentral,http://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data
Forbes,https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#58987f7b17b1
Why is BIG Data Important? A Navint Partners White Paper May 2012
http://searchcloudcomputing.techtarget.com/definition/MapReduce
http://searchdatamanagement.techtarget.com/definition/Apache-Hive
Computerweekly,http://www.computerweekly.com/news/2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off
Big Data University, https://bigdatauniversity.com

****