ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

BIG DATA Analytics: A Comprehensive Study on Present and Future Trends

Asoke Nath[1], Anwesha Chakraborty[2], Trina Dutta[3], Sushmita Mondal[4]

Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, West Bengal, India[1]

Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, West Bengal, India[2]

Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, West Bengal, India[3]

Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, West Bengal, India[4]

ABSTRACT: We need data in various aspects of our lives. Storing data in databases and then retrieving, modifying or inserting new data is a huge aspect in every industry. The concept of Big Data is something that normal database cannot contain oranalyze; it is measured in Petabyte and Exabyte. With the increased online activity, Big Data is eventually going to grow even more. It is going to play an important role in the future era of technology. It is going to solve many problems related to huge amount of data in corporate sectors, governmentorganizations, education, business etc. along with bringing in some new challenges with itself. Analyzing and handling Big Data is more complex and finding redundant data is even tougher in it. Since, Big Data is stored in distributed file system architecture and accessed from multiple domains, its security is another major concern for such an enormous amount of data. In this paper, the authors have made a systematic study on various characteristics ,features and challenges with Big Data and also made a study on future trends.

KEYWORDS: Database, Big Data, Petabyte, Exabyte, Security.

I.INTRODUCTION

Big Data is a very common term in today's world of technology and business analysis. The term 'Big Data' can be described as huge collection of data of wide ranges both structured and unstructured – that includes a business on a day to day basis. It is so vast and complex that it cannot be processed using traditional data manipulating methods. But it's not the amount of data that is important. It's what organizations do with the data that matters. Big Data can be analyzed for insights that lead better decisions and strategic business moves. This term has been in used since the 1990s as it was coined by John Mashey. Big data "size" is constantly increasing, as of 2012 the range was from a few dozen terabytes to many Petabytes. Big Data can be applied to real time fraud detection, complex competitive analysis, call centre optimization, consumer sentiment analysis, intelligent traffic management, and to manage smart power grids, to name a few applications. [1]

From Big Data comes the term 'Big Data Analytics'.Big Data Analytics examine large amount of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions. [1]

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Types of Big Data:

- *Structured*– Any data that can be stored, accessed and processed in the form of fixed format is termed as structured data.
  e.g.: Here's an example-

ID	STUDENT NAME	GENDER
123	JOHN	MALE
345	MARY	FEMALE
678	STEVE	MALE

- *Unstructured*– Any data with the unknown form or structure is known as unstructured data.
  e.g.: Output returned by 'Google Search'
- *Semi- structured*– This type of data can contain both of the formats of structured and unstructured data.
  e.g.: a table definition in relational dbms [2]

Types of data analytics:

Based on complexity of the analysis, data analytics can be classified into four types. As it happens, the more sophisticated an analysis is, the more value it brings.

- *Descriptive Analytics*: Descriptive analytics describes the 'what happened'. It deals with raw data from multiple data sources to give valuable insights into the past. These findings signals whether something is wrong or right but does not answer why.
- *Diagnostic Analytics:*This type of analytics has the strength to drill down, to find out the dependencies and to identify patterns. With the help of this method historical data can be measured against other data to answer the question of why something happened.
- *Predictive Analytics*: This generally tells what is about to happen. It uses the findings of descriptive and diagnostic analytics to detect tendencies, clusters and exceptions, and to predict future trends, which makes it a valuable tool for forecasting.
- *Prescriptive Analytics*: As one can get to know about the future trends from predictive analytics, those future problems can eliminated by the prescriptive analytics. Hence, this type of analytics prescribes the remedies of the future problems that may arise. [3]

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

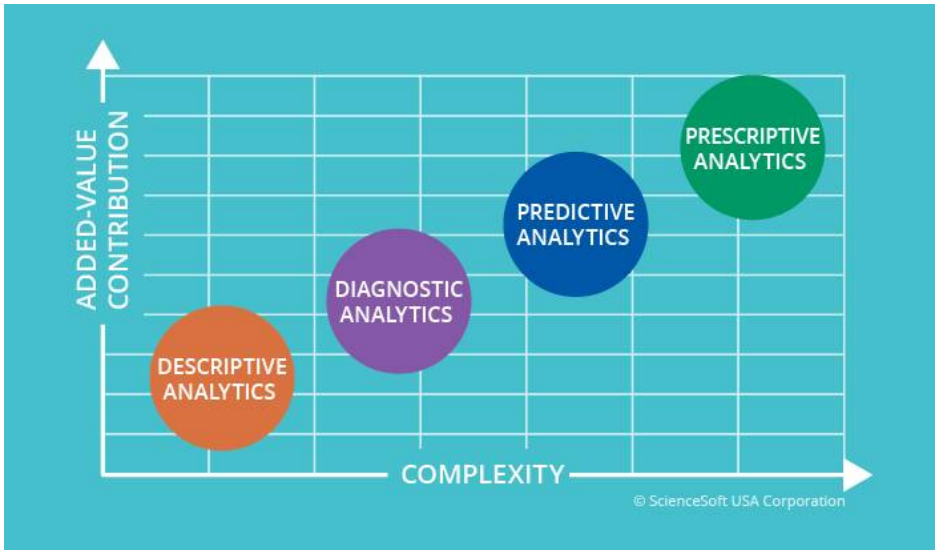

Fig: Graph showing types of data analytics

II.MAPREDUCE ALGORITHM TO HANDLE BIG DATA

MapReduce is a concept that is applied to handle big data. It can be described as a programming model for processing large data sets with a parallel distributed algorithm on a cluster. For big data processing MapReduce is coupled with HDFS which gives rise to the concept of Hadoop. It was introduced by Google.

The basic unit of information used in MapReduce is a (key, value) pair. All types of data whether it is structured or unstructured need to be translated to this basic unit before feeding it into the MapReduce model. This MapReduce model consists of two separate routines, viz. – Map function and Reduce function. The computation of MapReduce on the set of information (the pair) occurs in three stages.

Step 1: The map stage

Step 2: The shuffle stage

Step 3: The Reduce stage

Here, the data is distributed in the map and shuffle stage and the reduce phase performs the computation.

In the **map stage**, the mapper takes a single pair as input and produces any number of pairs as output. Most of the time, the map phase is simply used to specify the desired location of the input value by changing its key.

The **shuffle stage** is automatically handled by the MapReduce framework. The underlying system implementing MapReduce routes all of the values that are associated with an individual key to the same reducer.

In **reduce stage**, the reducer takes all of the values associated with a single key and outputs any number any number of pairs. This highlights one of the sequential aspects of MapReduce computation: all of the maps need to finish before the reduce stage can begin. Since the reducer has access to all the values with the same key, it can perform sequential computations on these values. For this stage, a function is designed that takes in input a list of values associated with a single key and outputs any number of pairs.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Overall, a program in MapReduce paradigm can consist of many rounds (usually called jobs) of different map and reduce functions, performed sequentially one after another. [11]

III.BIG DATA ANALYTICS AND HADOOP

The concept of Hadoop was introduced by Doug Cutting, Mike Cafarella and their team. They used the MapReduce Algorithm developed by Google to build this architecture. Hadoop is an Apache Open source framework, which is written in Java. In Hadoop, various sets of data run in parallel with each other on various CPU nodes. Hadoop has the capability to run applications on clusters of nodes and possesses the capability to provide statistical analysis on large amount of data like Big Data. Needless to mention, it has enormous storage space and has powerful processing capability.

The Hadoop architecture can be broken down into four major components:

1. *Hadoop Common*: These are Java libraries and utilities required by all the components of the Hadoop architecture. It contains the required Java files and scripts to start the Hadoop.
2. *Hadoop YARN*: This is a kind of a framework which is used for job scheduling and cluster resource management due to large amount of clusters of nodes working in parallel.
3. *Hadoop Distributed File System*: This distributed file system provides high-throughput access to application data.
4. *Hadoop MapReduce*: This is a YARN based system which helps large chunks of data process in parallel with each other.

What are the advantages of using Hadoop architecture for Big Data Analytics?

- Can store huge amount of data.
- Has fast data processing capability.
- Since it is a distributed system, its hardware fault tolerance capability is high. Even if one node goes down, it task is redirected to some other node.
- Hadoop architecture provides flexibility to process and store data as much as you want.
- In spite of all these capabilities, the Hadoop framework is not all expensive, because this open-source framework is free and uses commodity hardware.
- Hadoop architecture also provides scalability to increase the distributed system. [10]

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

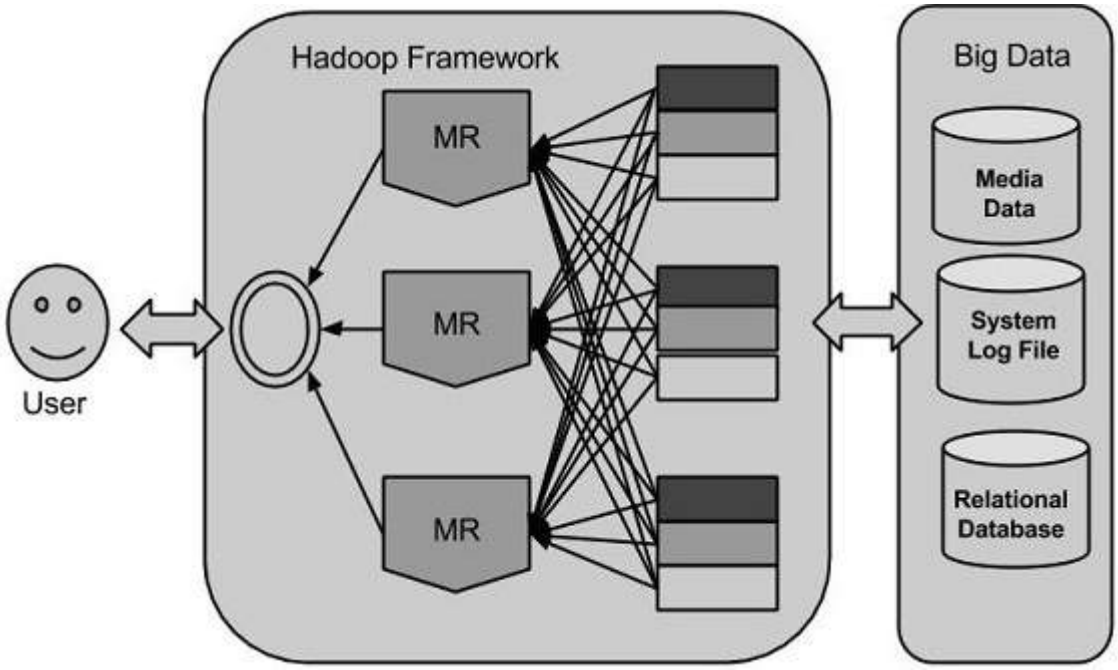

Fig: Hadoop Framework

IV.CHARACTERISTICS OF BIG DATA

Big Data can be characterized basically by 3 V dimensions that were first introduced in 2001. Later a few other characteristics have been added. The Big Data characteristics were compiled from several sources including IBM, Paxata, Dataflow, Data Science Central and National Institute of Standards and Technology (NIST). [5]

Here is the description of the 3 V's:

1. *VOLUME*: This refers to the vast amounts of data generated every second all over the world. Size of data is increasing in exponential growth. In order to deal with such huge amount of data, traditional databases can be used to handle or process them. Nowadays data is not only stored in Terabytes but also in Petabytes, Zettabytes.
2. *VELOCITY*: Velocity is the speed at which data is created, stored,analyzed, and visualized. The speed at which data is created currently is almost unimaginable. It is a challenge to deal with such enormous speed with which torrents of data is streamed in real time.
3. *VARIETY*: Different types of data are not new in today's world of technology. Every day one sends, receives, creates, process and manipulates various kinds of data viz. image, text, sound, video etc. Previously, the focus was mostly on structured data e.g.: financial data. But now, about 80% of the data dealt with is unstructured. So, with the help of big data analytics, we can analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, and video or voice recordings. [4]

Later, some other features have been added to the 3 V dimensions to describe Big Data. The other characteristics are:

- *Veracity:* It refers to the noise and abnormalities in data.
- *Validity:* It is important toanalyze whether the stored data is correct and accurate for the intended use.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

- *Volatility*: It defines how long a data is valid and how long it should be stored.

V. WHY DO WE NEED THE CONCEPT OF BIG DATA?

- As the world is advancing in network system huge number of data is being produced every day. So, it has been difficult to keep track of the data and has been difficult to maintain such large data. With the increase in data, complexity has also increased and it has become difficult to manage it using traditional data management tools[6]. Here comes the concept of Big Data which can maintain both structured and unstructured data. Big Data can be seen in finance and business where huge amount of data are exchanged on day to day basis.
- With the increase in companies competition between them have also increased. Companies have sought for decades to make best use of information to improve their business capabilities. They use the concept of Big Data for it. Big Data is special because it gives significant information of data and a way to analyze the data to open new doors.
- The thing about Big Data is that it can be interpreted in many ways. Big Data is defined by having Variety, Volume and Velocity of data however, processing this way always hard [6]. Still Big Data is a phrase that echoes across all corners of the business. It is coming into use eventually and many ways are being invented to use it in a more advance way.
- Big Data analytics are being used by various organizations to harness their data. It examines large amount data to reveal the hidden data, their patterns, and correlations. It reduces the cost since the Big Data technologies used such as Hadoop and cloud-based analytics can store large amount of data in significantly low cost.
- Big Data uses the technologies which are faster and uses better decision making techniques and so the concept of Big Data are used to analyze a large amount of data and give the good decision immediately on what they have learned.

VI.CHALLENGES WITH BIG DATA

- With the advancement in network technology data privacy has become the topmost priority for all the companies andorganization. [7] Security of these vast data is a big challenge for Big Data. Big Data contains various personal details and important information of many clients and customers of bigorganization and companies. It also stores crucial information about the companies. But since the data are so much vast and spreads over a wide range of variations authenticate them has become a challenge for Big Data. Authentication of the user for accessing data and access restriction is still a big challenge.[1] Data security is the most important issue for everyone and if it is not strong enough then data becomes vulnerable to many threats and can lead to an immense damage of data.
- The data which are stored using Big Data are mostly unstructured i.e. they contains data from various fields likes documents, videos, audios, photos etc. To store and analyze them is really hard since it may contain various unreliable data which should be discarded otherwise it may lead to wrong results. So data must be cleaned to avoid the dirty data otherwise it may lead to some serious problems in future.[8]
- The data which are stored and analyzed using Big Data helps in achieving some business goals. The data are analyzed to meet certain objectives. The Big Data provides insightsanalyzing data but the challenge is that not every time it provides the meaningful insights [8]. It must be improved in such a manner that it should generate proper insight time to time for guiding the business in correct direction.
- Using Big Data it is really necessary to scale up and down with demands. Big Data has the capability to evolve and drive a business but Big Data adoption should be nicely planned for effective output. Many companies

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

can't cope up with the pace of big data growth and evolution. Since the nature of Big Data workloads are bursty in nature so it is really hard to allocate proper resources.[8]

- Hadoop is an open-source; Java-based programming framework used to store large extensive data and is one of the famous technologies used in Big Data. But it is not easy to manage. It is really hard. Since it is a really a new technology and totally new concept some programmers aren't familiar with it and how to use it[8]. So to work with Big Data people should first get properly acquainted with Hadoop otherwise it would lead to many other problems.

VII.APPLICATIONS OFBIG DATA IN VARIOUS ASPECTS OF REAL LIFE

With the huge amount of data developing and emerging from various aspects of our day to day lives, there is no way in denying the fact that Big Data has become an important part of variousorganization and industries. That time is not far away when each and every MNCs and other private and governmentorganization will have to work with Big Data. Following are some industries in which Big Data Analysis is of utmost importance:

1. Public sectors.
2. Various medical and healthcareorganizations.
3. Educational services.
4. Various insurance companies.
5. Transportation services.
6. Banking systems and various scam detection.

*Contributions of Big Data in Public Sectors***:** Big Data has some notable contribution in various aspects of government sectors viz. investigation of power, recognition of fraudulent activities, fitness interconnected exploration, investigation of economic promotion, ecological enrichment etc.

The Food and Drugorganization (FDA) also uses Big Data to examine the various bacteria and other infection grown in food. Since Big Data Analysis provides quick results, it is proven to be very helpful.

*Contributions of Big Data in Healthcare***:** Big Data has also proven its worth in the field of medical science. With the advancement of technology and science, various lifesaving drugs have surfaced along with expensive equipments for the betterment of mankind. Due to these, the expenses in the field of healthcare have accelerated. The database contains the patient history of each and every patient and the physicians looking after them for the ease of doctors and hospitals staffs.

There aremany medical equipment that are linked to Big Data. Like heart and temperature monitoring devices, which the doctors checks remotely and prescribes medicines. Another link of Big Data with the healthcare are the miniatures robots called Nanobots which are designed to fight various germs and bacteria in the human body and increase immunity power.

*Contributions of Big Data in Insurance Companies***:** Insurance companies use Big Data as a tool to offer their customer various new ideas, schemes and insights for see-through and simpler commodities byanalyzing the behaviour and feedbacks from side by side information obtained from the internet, social media and CCTV footages. Scam detection has also improved greatly in the insurance companies after the introduction of Big Data.

*Contributions of Big Data in the field of Education***:** Big Data can have a huge influence in the field of education as well. In this era of technology, teaching and learning has become digital and is done electronically. The concept of virtual classroom is widespread nowadays. An application called Bubble Score allows teachers to provide MCQs

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

through mobile devices and notch up paper tests through the cameras of the mobile phones. This way many time and energy are saved, and accurate data is received.

*Contributions of Big Data in Transportation***:** Big Data plays an important role in traffic management, direction preparation, intellectual transportation arrangements and overcrowding administration. Direction forecasting to accumulate on petroleum is another personal usage of Big Data related to transportation.

*Contributions of Big Data in Banking Systems***:** Banking sectors uses Big Data to detect frauds and scams. Increased usage of technology and internet has brought many curses along with the blessings. Fortunately the introduction of Big Data has greatly reduced the cyber-crimes and other fraudulent cases related to banking. It detects the misuse of credit cards, misuse of debit cards, tracks various inspection being archived, analyzes credit hazard treatment, provides business clarity, customer statistics alteration, public analytics for business, IT action analytics, and IT strategy fulfilment analytics. [9]

VIII.CONCLUSION AND FUTURE SCOPE

Big Data is changing many viewpoints and the way we see things. It has become a very challenging and popular topic in the field of research. With the increment in Global Data, the amount of data will be four times by 2020 compared to the amount of data we have now. The presence and contributions of Big Data in various aspects of our lives are going to be inevitable and undeniable. Researchers believe that Big Data has a huge potential and gives faster output compared to any other systems. But there are many challenges with Big Data along with the advantages it is providing. Researchers are hopeful to overcome these challenges in the near future to discover more potential of Big Data in various industries. The future of Big Data seems quite promising and there are more scopes of research in this field that are yet to be discovered. [6]

REFERENCES

1.https://googleweblight.com/i?u=https://www.sas.com/en_us/insights/analytics/big-data-analytics.html&hl=en-IN
2.https://www.guru99.com/what-is-big-data.html
3.https://www.scnsoft.com/blog/4-types-of-data-analytics
4.http://googleweblight.com/i?u=http://www.dataintensity.com/characteristics-of-big-data-part-one/&hl=en-IN
5.https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
6. “Big data Security Issues and Challenges” by Raghav Toshniwal, Kanishka Ghosh Dastidar and Asoke Nath, International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN:2349-2163, Issue 2, Volume 2 (February 2015).
7. Security issues associated with Big Data in Cloud Computing, Venkata Narasimha Inukollu 1, Sailaja Arsi1 and SrinivasaRaoRavuri3 a. Department of Computer Engineering, Texas Tech University, USA
b. Department of Banking and Financial Services,Cognizant Technology Solutions, India
8. https://www.qubole.com/resources/big-data-challenges
9. https://intellipaat.com.
10.https://www.tutorialspoint.com
11.http://www.analyticsvidhya.com