# Big Data and Hadoop

Simmi Bagga[1] Satinder Kaur[2]

[1]Assistant Professor, Sant Hira Dass Kanya MahaVidyalaya, Kala Sanghian, Distt Kpt. INDIA
E-mail: *simmibagga12@gmail.com*

[2]Assistant Professor, GNDU, RC, Sathiala, INDIA

***Abstract:*** In this era data are continuously acquired for a variety of purposes. Data are generated from large-scale simulations, astronomical observatories, high-throughput experiments, or high-resolution sensors. As they all are used lead to new discoveries if scientists have adequate tools to extract knowledge from them. Hadoop is a Software platform that process vast amount of data. It helps to manage Big data. Hadoop can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email etc. In this paper, we discuss the concept of Big data, its sources and its characteristics. Some additional characteristics of Big data further more we describe the platform Hadoop which provides solution for Big data problem.

***Keywords:*** Big Data; Era Data; Hadoop; Large-scale Simulations; Astronomical Observatories

## I. INTRODUCTION

Big Data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big data is a voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity the term is often used when speaking about petabytes and exabytes of data etc. But today Big data can describe with 3Vs: the extreme Volume of data, the wide Variety of types of data and the Velocity at which data is traversing. As Big data takes too much time and costs too much money to load into a traditional relational database for analysis. So, new approaches to storing and analyzing data have been emerged which rely less on data schema and data quality.

The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. Corporations, government agencies and other organizations employ big data management strategies to help them compete with fast-growing pools of data, typically involving many terabytes or even petabytes of information which have been saved in a variety of file formats. Effective big data management helps companies to locate and represent valuable information in large sets of unstructured data and semi-structured data from a variety of sources, including call detail records, system logs and social media sites.

There are huge volumes of data in the world. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data. In 2011, the same amount was created every two days. In 2013, the same amount of data is created in every 10 minutes. 90% of the data in the world today has been created in the last two years alone.

There are two major architectural changes that are needed in the traditional integration platforms in order to fulfill future requirements. First, the ability and needs for organizations to store

  

and use big data. Most of the big data should always be available for a longtime and there should be tools and techniques available to process be same for the business benefits. Then second, the need for predictive analytical model which must be based on the history or patterns of past or hypothetical data driven models. While the business intelligence deals with what has happened, business analytics deal with what is expected to happen that is to forecast the situation.

## II. SOURCES OF BIG DATAT

Big data comes from different sources like sensors which are used to gather climate information from social media sites while posting digital pictures and videos from transaction records while purchasing goods, and from cell phone GPS signals etc. as shown in fig1.

Using social media like Facebook we produce lots of big data when we use to look it as when we look at another person's timeline, send or receive a message, when we post things like photos or videos on Facebook etc. We receive data from or about the computer, mobile phone, or other devices you use to install Facebook apps or to access Facebook .We receive data whenever we visit a game, application, or website that uses Facebook Platform. Sometimes we get data from some advertising partners, customers and other third parties that help us to deliver ads.

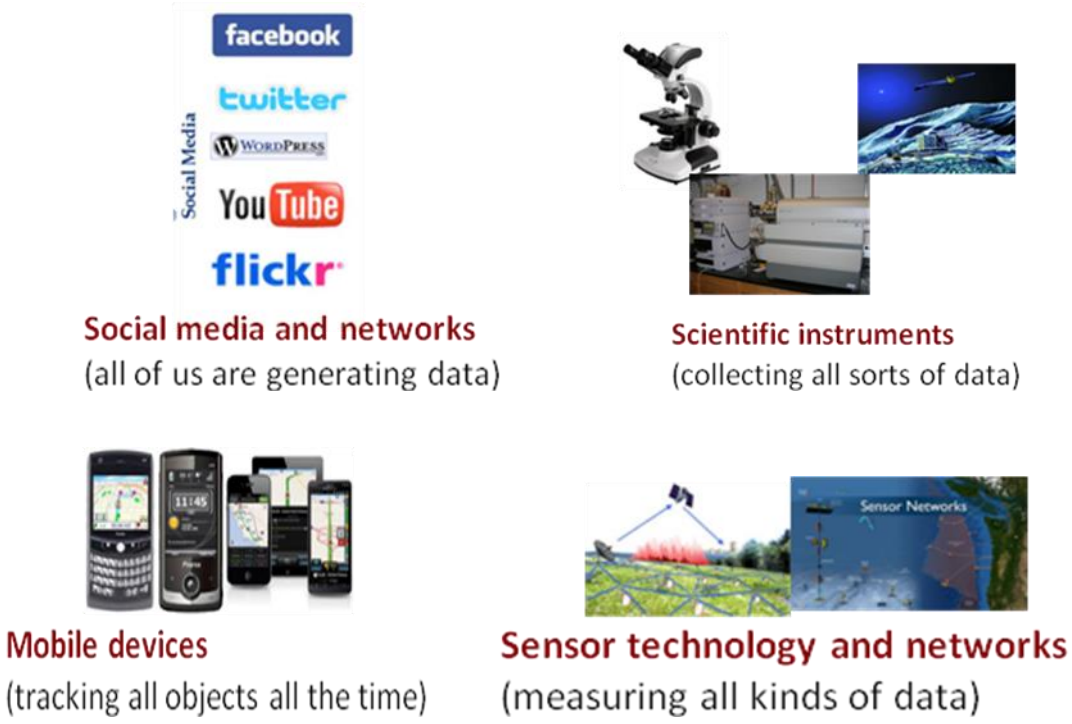


**Fig1. Sources of Big Data**

Different Physical Sensor data also produces high velocity, volume, variety of data. Sensors are used for geolocation, temperature, noise, attention, engagement, biometrics purposes to collect data in a variety of ways. When we use this large data to user context and to predict behavior we have to process this huge data that is a big challenge.

The progress and innovation is no longer hindered by the ability to collect data. But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion is necessary.

## III. CHARACTERISTICS OF BIG DATA

"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" Gartner 2012.Big data can be described by the following characteristics:

### *3(a). Main Charactristics i.e. 3 V's*

#### I. Volume

Big data management software enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The quantity of data that is generated is very important in this context. Which data is generated by machines, networks and human interaction on systems like social media is very massive. The name 'Big Data' itself contains a term which is related to characterized size that describe it.

#### II. Variety

Variety describes different formats of data. These include a long list of data such as documents, emails, social media text messages, video, still images, audio, graphs, and the output from all types of machine-generated data from sensors, cell phone GPS signals, DNA analysis devices, and more. This type of data is characterized as unstructured or semi-structured. This variety of unstructured data creates problems for storage, mining and analyzing it. Unstructured data is growing much more rapidly than structured data. It is estimated that unstructured data doubles every three months and offers the example that there are seven million web pages added each day.

There are two primary challenges regarding variety of data. First, storing and retrieving these data types quickly and cost efficiently. Second, during analysis, blending or aligning data types from different sources so that all types of data describing a single event can be extracted and analyzed together.

#### III. Velocity

Today Data is generated too fast and also need to be processed fast. Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers only if you are able to handle the velocity. Velocity is applied to data in motion. There are various information streams and the increase in sensor network deployment has led to a constant flow of data at a pace that has made it impossible for traditional systems to handle. Initially analysis of data is done by using a batch process. With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

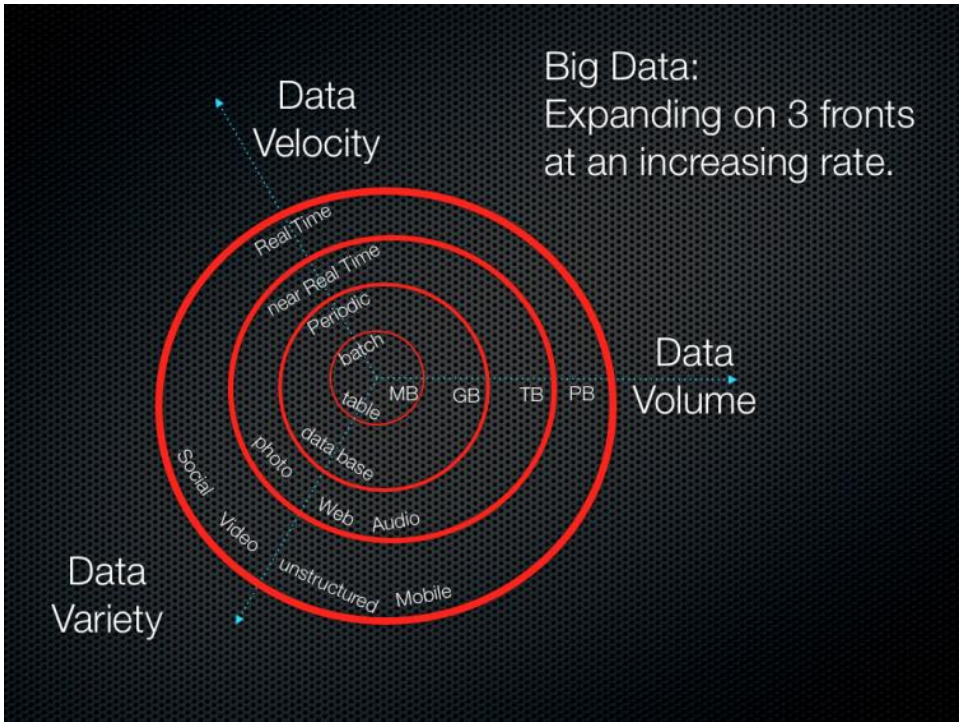


Figure: 2. How 3 V's exploring

**3(b). Additional Characteristics**

**I. Veracity**

Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Veracity in data analysis is the biggest challenge as compared to volume and velocity. In scoping your big data strategy ones need to have his own team and partners together to work to clean the data in the system to avoid 'dirty data' from accumulating in his systems.

II. **Validity**

Like big data veracity, the issue of validity deals with the accuracy of data. So that it can be properly use from decision making and fir future conclusions.

III. **Volatility**

Big data volatility refers to how long is data valid and how long should it be stored. In this world of real time data one need to determine at what point is data no longer relevant to the current analysis.

Big data management clearly deals with issues beyond volume, variety and velocity to other concerns like veracity, validity and volatility to hear about other big data trends and presentation and uses.

## IV. HADOOP : A BIG DATA MANAGEMENT PLATFORM

Hadoop is a processing engine that is designed to handle extremely high volumes of data in any structure. Hadoop provides distributed storage and distributed processing for very large data sets. Hadoop has two main components:

- The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between. HDFS is a reliable distributed file system that provides high through put access to data.

- The MapReduce programing paradigm which is meant for managing applications on multiple distributed servers. It is a framework for performing high performance distributed data processing and is based on divide and aggregate paradigm.

HDFS, or the Hadoop Distributed File System, gives the programmer unlimited storage. The advantages of HDFS are:

Horizontal scalability: Thousands of servers holding petabytes of data. When you need even more storage,we don't switch to more expensive solutions, but add servers instead.

Commodity hardware: HDFS is designed with relatively cheap commodity hardware in mind. HDFS is self-healing and replicating.

Fault tolerance: Hadoop also deal with hardware failures. If one have 10 thousand servers, then he will see one server fail every day, on average. HDFS foresees that by replicating the data, by default three times, on different data node servers. Thus, if one data node fails, the other two can be used to restore the third one in a different place.

MapReduce takes care of distributed computing. It reads the data, usually from its storage, the Hadoop Distributed File System (HDFS), in an optimal way. However, it can read the data from other places too; including mounted local file systems, the web, and databases. It divides the computations between different computers (servers, or nodes). It is also fault-tolerant. If some of nodes fail, Hadoop knows how to continue with the computation, by re-assigning the incomplete work to another node and cleaning up after the node that could not complete its task. It also knows how to combine the results of the computation in one place

A set of machines running HDFS and MapReduce is known as a Hadoop Cluster. Individual machines are known as nodes. A cluster can have as few as one node, as many as several thousands. More nodes in the cluster mean better is the performance.

These both the components are Hadoop provide ultimate solution to handle the problem of Big data. Hadoop meets all the requirements related to reliability, scalability etc which are the major challenges with the big data.

## V. CONCLUSION

In this paper, we explained big data and its sources. We also explained the characteristics of Big data and also explained how Hadoop can be the solution of the problem of Big Data. Due to the scale, diversity, and complexity of Big data a new architecture, techniques, algorithms, and analytics are required to manage it and extract value and hidden knowledge from it. There are various

architectural changes that are needed in the traditional platforms in order to overcome future challenges. Hadoop has become a valuable business intelligence tool. Hadoop is a processing engine that is designed to handle extremely high volumes of data in any structure.

## VI. REFERENCES

[1] Beyer, M.A, Laney, D.: The Importance of 'big data': a Definition. Gartner (2012).
[2] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013.
[3] VinayakBorkar, Michael J. Carey, Chen Li, Inside "Big Data Management":Ogres, Onions, or Parfaits?, EDBT/ICDT 2012 Joint Conference Berlin, Germany,2012 ACM 2012,
[4] Hadoop,"PoweredbyHadoop,"http://wiki.apache.org/hadoop/PoweredBy.