A Review paper on basics of big data and Hadoop

Dr.Megha Gupta1st
Computer Engineering
Poornima Institute of Engineering and Technology
Jaipur, India
E-mail: megha.gupta@poornima.org

Laveena Chaturvedi2nd
Computer Engineering
Poornima Institute of Engineering and Technology
Jaipur, India
E-mail: 2015pietcslaveena@poornima.org

Abstract: In today's growing world very large amount of data are available in hands of decision maker. Big data is difficult to handle using traditional tools and techniques. Because they refer to datasets that are high in variety and velocity. Various solutions should be provided to handle getting values from this datasets because of increasing growth of data. We need to provide solution in order to handle these datasets. This paper provides basic information about the big data and its advantages, dimensions and its scope for the future research. This paper also gives an introduction to Hadoop and its components.

Keywords: big data, analytics, datasets, Hadoop, decision making.

I. INTRODUCTION

We can define big data as the datasets or the combinations of data sets which has basically 5V's which are volume, variety and velocity. Veracity and value are recently added to the list. In today's world we cannot imagine a world without big data as it plays an important role. It basically helps us in finding a diamond among a huge data. It even use in saving of large amount of data which are being lost after use. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data. The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed and pertaining information should be extracted. BIGDATA (Beyond storage capacity and processing) is a vague topic and there is no exact definition which is followed by everyone [1]. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks.

The question arises how we can store such huge amount of data and how to store such data within specific time?

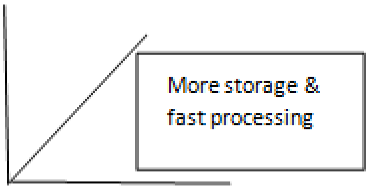
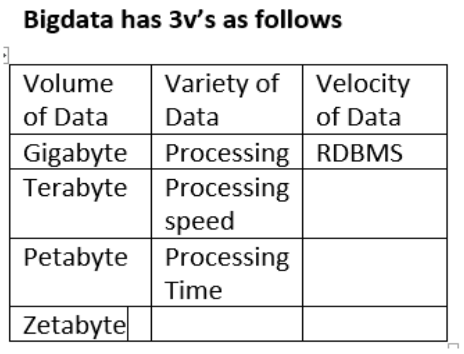


Fig. 1 Example of a graph showing more storage & fast processing

II. V'S OR DIMENSIONS OF BIG DATA


Bigdata has 3v's as follows

Volume of Data	Variety of Data	Velocity of Data
Gigabyte	Processing	RDBMS
Terabyte	Processing speed	
Petabyte	Processing Time	
Zetabyte		

Fig.2 3 v's of Big Data

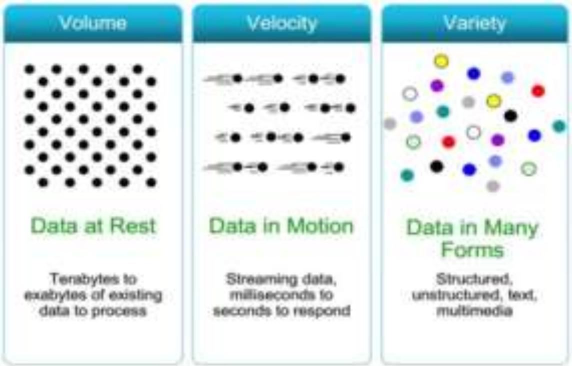

Volume
Velocity
Variety
Data at Rest
Terabytes to exabytes of existing data to process
Data in Motion
Streaming data, milliseconds to seconds to respond
Data in Many Forms
Structured, unstructured, text, multimedia

Fig. 1 Data in terms of Volume, Velocity and Variety

Volume: -We can see that the data is increasing more and more in the data storage, even we can also see that the data is more than the text data. We can find data in the format of videos, music's and large images on our social media channels. It is very common to have Terabytes and Peta bytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be revaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents Big Data.

Velocity: - velocity generally refers to the speed of data processing. The vision of data has been changed along with the changing of time. When we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today's youth reply on social media to update them with the latest happening. On social media sometimes, a few seconds old messages are even old to read so basically the speed of sharing data is increasing so fast that they don't even need a second to have a new update's in media. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represents ***Big** Data*. Velocity is the speed at which data is generated and processed. For example, social media posts [2].

Variety:-In Variety we can store the data in various formats. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world has data in many different formats and that is the challenge we need to overcome with the *Big Data*. This variety of the data represents big data**.** There are two more dimensions or 2 more v's which are recently added to the list Veracity & Value. When we talk about value, we're referring to the worth of the data being extracted. Veracity is the quality or trustworthiness of the data [3]. On twitter 400 million tweets are sent per day and there are 200 million active users on it.

There are many things through which big data are made of: -

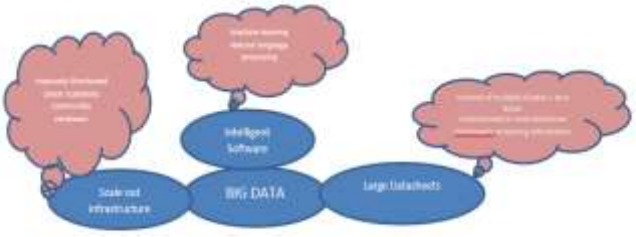


Fig.4 Architecture of Big Data

III. III BENEFITS OF BIGDATA

- Big Data is timely: -big data is timely as it doesn't require much time to manage it and find it. The high speed of tools like Hadoop and in-memory analytics can be easily identify new sources of data which helps Businesses analyzing data immediately and make quick decisions based on the learning.
- Cost-Reduction: - When there is a need of storing large amount of data there are few tools of big data like Hadoop and Cloud-based Analytics which bring cost-based advantages to the business and the best way of doing business with maximum profit.
- Understand the market conditions: by using big data we can find out what is there in the market in trend. And after that we can work according to the need of market. For example, by noticing what is the behavior of the customer's while purchasing, companies basically try to find out which is most sold product in the market which product is in demand and try to produce products according to their need[4].
- Big data is Holistic: - information is kept in Silos within an organization. The data is very safe in this place.

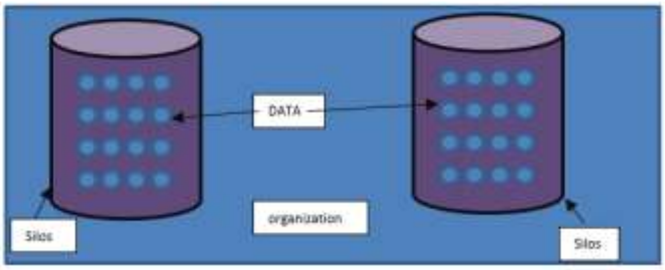


Fig.5 Data stored in Silos

We required a marketing data which might be develop sin web analytics, mobile analytics, social analytics, CRM3 (customer relationship management) and even email marketing system. Every one of them has its own silos [9].

- Faster and better decision making: -It is also a benefit of big data that it is faster and has better decision making suppose we have large organization those who are seeking to replace that they require they make better decision it is driven by the speed of Hadoop and in-memory analytics. Hadoop is using cluster of machines together. So basically, companies require Hadoop and in-memory analytics for speed and faster work.

IV. CONCEPT OF BIGDATA HADOOP

Hadoop is open source software used to process the Big data. It is being noticed that Hadoop is one of the popular used software of big data in the field of organizations/researchers. There were two people who decided to work on Hadoop, Hadoop was the solution provided by Goggle [8]. The two who decided to work on this open source project Hadoop were Doug cutting, Mike Cafarella in 2005. Now Apache Hadoop is registered trademark of the Apache Software Foundation. Hadoop use applications using the Map Reduce algorithm, where the data is processed in parallel on different CPU nodes. In short Hadoop framework is capable of running on clusters of computers and they could perform complete statistical analysis for a huge amount of data.An Apache Hadoop ecosystem consists of the Hadoop Kernel, Map Reduce, HDFS and other components like Apache Hive, Base and Zookeeper [5]

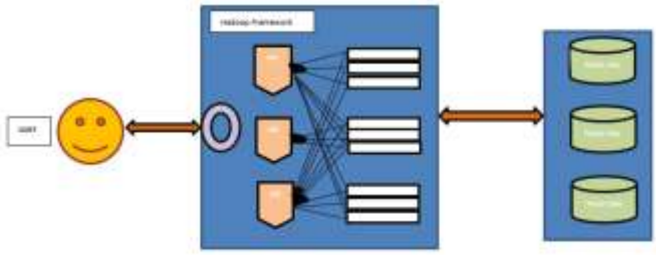


Fig.6 Concept of Big Data

A. How Does Hadoop Work?

Stage 1

A user/application can submit a job to the Hadoop (a Hadoop job client) for required process by specifying the following items:

1. The location of the input and output files in the distributed file system.
2. The java classes in the form of jar file containing the implementation of map and reduce functions.
3. The job configuration by setting different parameters specific to the job.

Stage 2

In the second stage job is been submitted by the Hadoop job client (jar/executable etc.) and configuration to the Job Tracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client[6].

Stage 3

The Task Trackers on different nodes execute the task as per Map Reduce implementation and output of the reduce function is stored into the output files on the file system.

V. ADVANTAGES OF HADOOP

Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores [7].

- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.

- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.

- Last but not the least advantage of big data is that as it is java based it is compatible in all the platforms. It is also an open source [10].

VI. CONCLUSION

In this paper we have enlighten the light on the topic of basics of big data in which we have covered introduction, benefits and its dimensions as 5V's of big data and we have also shown from which the big data is made of. We have even given the brief introduction about the concept of Hadoop along with its diagram of framework and advantages in the big data and also how does it work. So, basically this paper shows how important is Hadoop in big data and how big data is important in today's world.

ACKNOWLEDGMENT

First and foremost, I would like to thank my loving parents and respected faculty Dr. Megha Gupta for standing beside me throughout my career as well as while writing this review paper. Without her support it would have been totally impossible to achieve this knowledge sharing. As I started writing this review paper. Also, thanks to my friends for encouraging me to initiate writing for big technologies such as Hadoop.

REFERENCES

[1]. Harshawardhan S. Bhosale, Proof Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[2]. SMITHA T, V. Suresh Kumar "Application of Big Data in DataMining" in International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 7, July 2013).

[3]. IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data

[4]. Mrigank Mridul, Akashdeep Khajuria, Snohomish Dutta, Kumar N Analysis of Big data using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[5]. Big Data, Wikipedia, http://en.wikipedia.org/wiki/Big_dataWebster,Phil. "Supercomputing the Climate: NASA's Big Data Mission". SC World. Computer Sciences Corporation. Retrieved 2013-01-18

[6]. www.udemy online course

[7]. Poonam S. Patel et al. "Survey Paper on Big Data Processing and Hadoop Components", International Journal of Science and Research, Volume 3, Issue 10, October 2014.

[8]. Apache HBase. Available at http://hbase.apache.org

[9]. Apache Hive. Available at http://hive.apache.org

[10]. Abhishek S, "Big Data and Hadoop", White Paper

AUTHOR'S BIOGRAPHIES

Dr.Megha Gupta completed her BTech (CS) from Mody Institute of Engineering and Technology, Luxmangarh in year 2004, MTech from Banasthali Vidyapith in 2006 and PhD from Jagannath University in 2016.She has published a number of papers in various national and international journals and conferences. Her field of interest includes networking, neural network and IOT..

Ms.Laveena Chaturvedi pursuing her Btech final year from Poornima Institute of Engineering and Technology.