UNDERSTANDING BIG DATA: A SURVEY

[1]Dimple Chehal & [2]Surabhi Lingwal
[1,2] Research Scholar
[1,2] Department of Computer Engineering
[1,2] J.C. Bose University of Science and Technology, YMCA, Sector 6, Faridabad, 121006, Haryana,India

Received: January 29, 2019 Accepted: March 12, 2019

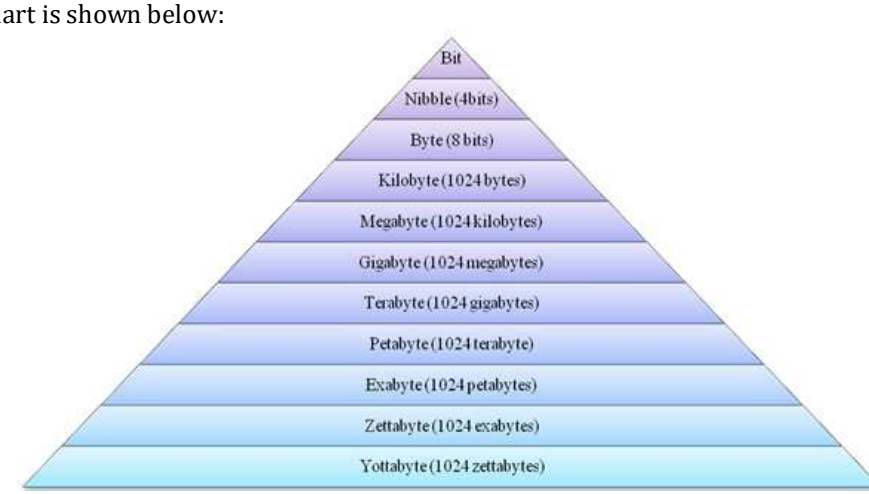
ABSTRACT *This paper aims to study basic terminologies associated with big data. Big Data's seven V characterizations have been overviewed with value being the ultimate desired output for any business entity. Various challenges that crop up due to big data have been explained here. Through this paper, understanding of big data platforms and application areas of big data can be achieved.*

Key Words: *Big Data, Characteristics, Challenges, Hadoop, Spark, Application*

I. Introduction

Big Data is a term coined for data which is beyond the processing capability of traditional systems. According to IBM, 90% of all the data present has been created in the last two years and 2.5 quintillion bytes of data is generated everyday [3]. As the analysis of data by traditional relational databases was not possible, NoSQL Databases were introduced to overcome their challenges. NoSQL databases have flexible schema without any downtime. Big data is generated from clickstreams, log history, sensors, scientific experiments, social media etc. and data is in semi structured or unstructured format. Analysis of such data will help both the consumer as well as the business. For instance, fraud detection can help secure a customer's bank deposit and save the associated bank from risks involved.

With the advent of big data, systems with scaling hardware/software have become mandatory [2]. Several frameworks supporting big data have cropped up. Choosing these platforms requires knowledge of programming paradigms usage. The various parameters to evaluate a framework are, time to compute result, data size to process, number of iterations, scalability, data transfer speed, fault tolerance Data measurement chart is shown below:



II. Characteristics of big data

Big data was characterized by 3Vs by Gartner in 2012 who referred to big data as 'Big data is high-volume, high velocity and high variety information aspect that requires new format of processing to enable enhanced decision making insight discovery and process optimization [4]'. Volume refers to the size of data generated from various sources like sensors, transaction logs, satellite data, history logs, experimental, scientific data and social data. Velocity refers to the speed of generation of big data. Not only should this velocity be handled by big data ecosystems but also the data streaming rate should be controlled to transfer the data to storage places for later retrieval and analysis. Variety refers to the different types of data which

makes it all the more complex. Unlike traditional systems where the data was primarily textual, data in big data era is made up of videos, audios, text, images etc. 4th V was given by IBM – Veracity indicates data integrity. It means how certain are we for the data received through all the sources as such data contains uncleansed data as well. Validity refers to correctness or accuracy of data with respect to the concerned usage. Data's validity depends on the application in which it is to be used. Volatility refers to the retention period of big data. Retaining of big data is difficult due to other Vs associated with it. Value is the result of big data processing.

II. BIG DATA ANALYSIS STEPS

To analyse big data the steps are as mentioned above. Initially, data needs to be acquired from various sources and should be cleaned before storing it into data storehouses. Then, in order to analyse the data, relevant data must be retrieved and finally visualized to understand the processed results.

III. Challenges of Big data

Many problems arise due to inherent nature of big data. The following pose a challenge to big data analysis [1, 4]:

1. *Privacy*: This is one of the major concern of the users involved in big data analytics as there is fear of data compromise or data misusage by unintended recipients.
2. *Security:* Data procured and stored in servers should be secure from hackers. No confidential data of the customer should be accessed by unapproved groups.
3. *Data storage*: Initial data dump consisting of data in its actual form needs to be stored for future processing. This data long with its meta data needs storage houses for real time processing. The storage should not be as expensive as cloud and reliable enough for future retrieval
4. *Heterogeneity of data*: Data being unstructured is not of homogeneous nature as expected by processing algorithms. Also, not all unstructured data can be converted into structured data due to underlying overheads involved. This data diversity is cumbersome to deal with.
5. *Inconsistent, Missing and low quality data*: Data received from several sources can be full of errors, miss key values and be of quality not useful for processing.
6. *Scalability*: Processing or accuracy of results should not be affected with changing data size.
7. *Visualization*: Results of big data analytics should be understandable with the help of plots and graphs and one should not feel lost in the pool of data calculations
8. *Fault tolerance:* Systems handling big data should be tolerant to damage if any and the damage done should be in acceptable limits.

IV. Apache Hadoop and Spark and NoSQL databases

Scaling is the ability to handle demands of growing data. Scaling of data can be handled with two approaches namely, scaling out and scaling up or a combination of both. While the former requires addition of commodity machines to existing system so as to improve the assigned work, the later requires upgrading of the existing system machines by increasing memory, processors and hardware. Platforms that make use of scaling out approach are Apache Spark and Apache Hadoop[5] and platforms which make use of scaling up approach are multi-core processors, high performance computing and general purpose graphic processing unit.

1. Apache Hadoop is Apache's open source framework for distributed computing written in Java. It is based on scaling out approach to process growing data by addition or removal of commodity hardware as per the data need. Instead of transferring the data to a set of slave nodes, computation is transferred to them and portion of data set is processed as decided by the master node. Major components of Hadoop include:
   a. *HDFS (Hadoop Distributed File System)*: As the name suggests, it is file system used in Hadoop framework in distributed manner. The file content is broken into blocks of 128Mb and by default three replicas of a file are maintained to make the system tolerant to faults.
   b. *YARN (Yet another resource negotiator)*: Manages resource and schedules jobs across the nodes in clusters.
   c. *MapReduce:* helps to process and analyze large datasets in parallel.
   d. *Other components:* Pig, Hive, HBase, Mahout, Zookeeper, Cassandra, Tez etc.
2. Apache Spark performs in-memory data computation on large clusters. Unlike, Hadoop, Spark is capable of doing iterative analysis of data. It is much faster than Hadoop due to its capability to store data in memory
3. Types of NoSQL Database are categorized into [1]:

1. *Key value database*: employs hash table usage where there is unique key and a pointer to set of values.
2. *Column oriented database*: row key, each row can have multiple columns Eg: HBase, Cassandra, big table used in google earth
3. *Document database*: a document stores a record and related data. These database do not have a schema. Eg: MongoDB, Couchbase
4. *Graph database:* Attribute-Data is represented as key value pairs in nodes and nodes are related to each other via edges.

CAP theorem states that such databases must have two properties among partition tolerance, consistency and availability

V. Applications of Big data

Big data plays an important role in many fields to benefit both the organization and its consumers. Some of its applications are listed below [4]:

1. *Health care:* Through analysis of patient's reports, medical history, health patterns can be identified to take preventive care, detect diseases at early stages and cure current ailments.
2. *Education:* Analysing the current job opportunities in market, student preferences for courses and number of admissions done in years, enrolment prediction for following years can be done easily. Based on student's profile, courses can be recommended to the students.
3. *Social network analysis*: an application of graph theory can help classify social relations and identify social groups sharing a common interest.

VI. Conclusion

Big data has immense potential for development of future systems. Trend analysis and prediction rules can help in solving real world problems. Big data analysis can help increase in business productivity and thus make the decision making capabilities stronger.

References

1. S. Srivastava, S. Agarwal, A. Srivastava, and A. K. Pandey, "Big data - An emerging and innovative technology: Survey," Proc. - 2016 2nd Int. Conf. Comput. Intell. Commun. Technol. CICT 2016, pp. 180–183, 2016.
2. M. Merrouchi, M. Skittou, and T. Gadi, "Popular platforms for big data analytics: A survey," 2018 Int. Conf. Electron. Control. Optim. Comput. Sci., pp. 1–6, 2019.

3. M. A. U. D. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," Proc. 2014 Zo. 1 Conf. Am. Soc. Eng. Educ. - "Engineering Educ. Ind. Involv. Interdiscip. Trends", ASEE Zo. 1 2014, 2014.
4. P. V. Desai, "A survey on big data applications and challenges," Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, no. Icicct, pp. 737–740, 2018.
5. Apache Hadoop http://hadoop.apache.org/
6. Apache Spark-Unified Analytics Engine for Big Data http://spark.apache.org/