International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 1
ISSN: 2454-4248
322 – 325

Overview of Big Data

Kavita D. Zinjurde
Asst. Professor: Department of MCA
Jawaharlal Nehru Engineering College
Aurangabad, Maharashtra, India
kavita.zinjurde2011@gmail.com

Sujata S. Magare
Asst. Professor: Department of MCA
Jawaharlal Nehru Engineering College
Aurangabad, Maharashtra, India
magaresujata@rediffmail.com

Sandeep S. Parwe
Asst. Professor: Department of MCA
Jawaharlal Nehru Engineering College
Aurangabad, Maharashtra, India
sandeep.s.parwe@gmail.com

***Abstract*—** In this paper we have focused on basics of Data and Big Data, characteristics and yet challenges of Big data. Big data can store massive amount of data, we have described approaches and the architecture that are used to store data. Nowadays Big Data is used in various subject area like pattern recognition, medical, commercial industries and social networking.

Keywords- *BigData;Hadoop; MapReduce;Veracity; Veriability*

******

I. INTRODUCTION

Big data can be used in every application where large amount of massive data is needed and which demands innovative computational methodologies and more powerful tools. With rapid growth of data analysis of knowledgeable data have become more challenging [1].

Big data demands special infrastructure and trained workforce. Trained workforce is needed to take benefits of Big data analysis. Security and privacy are challengeable issues. Data is physically stored in a different place and can be access through a network [2]. Big data generally includes Traditional Enterprise data, Machine generated or sensor generated data and Social data [3].

II. DATA

This Data is a special kind of information which has been derived/generated from our/someone's experiences, with sense, with intellectuals, with knowledge& with understanding. Data has always been required for proper Decision Making. On the basis of available data one can take a decision for his further task.

Data has always been everywhere and there has every time been a need for storage, processing, and management of data, meanwhile the beginning of human evolution. Still, the amount and type of data captured, stored, processed, and managed depended then and even now on various aspects including the requirement felt by humans, available tools/technologies for storage, processing, management, effort/cost, and ability to increase understandings into the data, make decisions, and so on.

To capture, store, and process the data has enabled human beings to propagate knowledge and research from one generation to the next, so that the next generation does not have to re-invent the pulley.

The capacity of data storage has been increasing dramatically, and today with the availability of the cloud infrastructure, possibly one can store unlimited amounts of data. Today Petabytes and Exabytes of data is being generated, captured, processed, stored, and managed[4].

If the data is storing is live i.e. continuously running data example, Facebook data where user are recurrently sending likes, comments, uploading/sharing photos, videos etc. these all things are get stored on the server. At the backend of the server there is storage which is nothing but the place where all data is get stored. And Size of these storage is depends on the company from where you are taking/hiring the storage [5].

A. Data Measurement

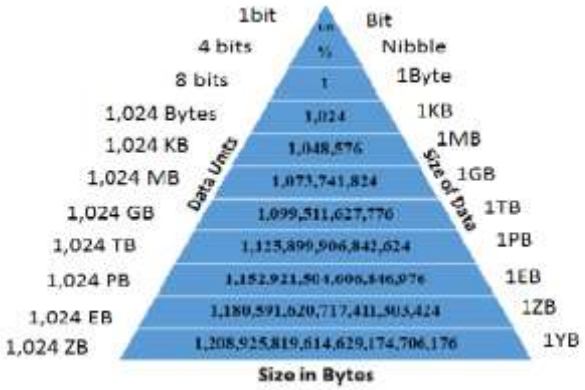

Figure 1. Pyramid of Data Measurement

Byte= A single letter, like "A."

Kilobyte=A 14-line e-mail. A pretty lengthy paragraph of text.

Megabyte= A good sized novel. Shelley's "Frankenstein" is only about four-fifths of a megabyte.

Gigabyte=About 300 MP3s. About 40 minutes of video at DVD quality (this varies, depending on maker). A CD holds about three-fourths of a gigabyte.

Terrabyte=statistically, the average person has spoken about this much by age 25.

Petabyte= the amount of data available on the web in the year 2000 is thought to occupy 8 petabytes.

Exabyte= in a world with a population of 3 billion, all information generated annually in any form would occupy a single exabyte.

 ISSN: 2454-4248
322 – 325

Zettabyte= Three hundred trillion MP3s; Two hundred billion DVDs.

Yottabyte=???????? [6]

III. BIG DATA

As the name signifies it consist of huge amount/volume of heterogeneous data which is being generated at high speed. This high speed generating data cannot be managed and processed using traditional data management tools and application at hand. A Big Data is an emerging technology which can deal with high speed generating huge volume of complex data with the use of a new set of tools, applications and frameworks to process and manage the data [7].

For example, if we construct one house with single labor in 20 days and if we construct same house with 20 labors so it will take 1 day to construct it. Same thing is there in big data by adding or using parallel resources. You can reduce the processing time of big data to analysis for decision making.

A. Why Big Data

To perform deep Analytics it is necessary to integrate Big data with the organizational data [3]. Following are some reasons that fascinate industries towards big data shown in fig. 2. Table I shows reasons and the benefits of using big data

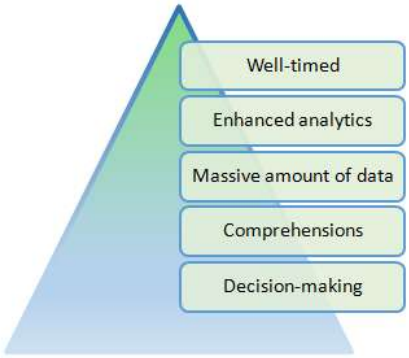

Figure 2. Motives for Big Data

TABLE I. REASONS AND BENEFITS OF BIG DATA

Reasons	Big Data Benefits
Well-timed	Increase immediate insights from different data sources
Enhanced Analytics	Improvement of business performance through real-time analytics
Massive amount of data	Huge amounts of data can be managed
Comprehensions	Unstructured and semi-structured data helps to provide better Perception
Supervisory Decision-Making	Risk analysis helps to moderate risk

B. Big Data Sources

Big data has many sources that includes A large stock exchange, Video sharing portal (like YouTube, Video, Daily motion etc.), Social networks (like Instagram, LinkedIn, Facebook, and Twitter etc.), Network sensors, Web pages, text and documents, Logs (such as Web logs, System logs, Search index data) etc.

Each mouse click on a *web site* can be captured in Web log files and analyzed in order to better understand shoppers' buying activities and to encourage their shopping by vigorously endorsing products. *Social media* sources generate tremendous amounts of comments, likes and tweets. There is remarkable amounts of *geospatial* (e.g., GPS) data, such as that produced by cell phones, that can be used by applications like Four Square to assist you know the locations of friends and to receive offers from nearby stores and restaurants. Applications such as recognition systems in security systems can be analyzed data like *Image, voice, and audio*.

C. Big Data Usages

- To understand pattern like analyze pattern for a analytics
- To understand behavior of people for example, In medical industry for patients under observation for unknown diseases.
- Share market
- In bank transaction to analyze suspicious behavior
- Many more notifications on social sites like Facebook, twitter etc. & many more [8]

D. Storage Component and Processing

The Data acquisition, data organization and data analysis are stages that are required to carry out infrastructural platform. Data Acquisition refers to collection data from different sources. Data may be in any format and size; it has to be organizing in such manner so that it could be easy to extract knowledgeable and useful information from such diverse data. Data organization and Analysis does this work. Data is uniformly stored and can be extractable for analysis purpose. Statistical method and Data mining techniques is used to achieve this task.

- A big data is stored with special kind of technique known as HDFS i.e. *Hadoop Distributed File System*, which is storage component of Hadoop. HDDS provides high performance across data clusters[9].
- Normally you stored your data in txt.file, .xml files, .ldb, .mdf.
- Processing part of Big Data is done by MapReduce (Business Logic)

1) Architecture of Storing Data

Fig.3 shows Traditional architecture and fig.4 shows Master-Slave Architecture

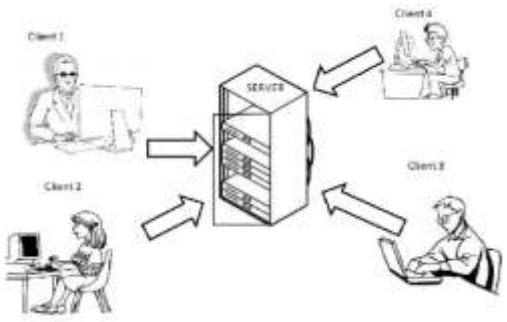

Figure 3. Traditional Client Server Architecture

The Master-Slave Architecture consist of Name Node and Data Node

 ISSN: 2454-4248
322 – 325

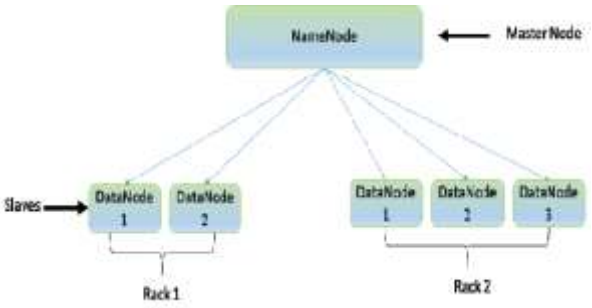

Figure 4. Modern Master Slaves Architecture

- NameNode (Master):- Name node works on Master system or in other word we can say that Name node work as a Master node. Name node is responsible for managing Meta data which consist of controls access to file system by the clients, file system namespace, records of the DataNodes, replication factor and confirms that it is continuously maintained.
- DataNodes (Slave):- Data node works on slave system. The main task of data node is to follow the instruction coming from Master or name node. You can read/write data from/to data node. Actually whatever data receives from user is stored on datanode in the form of blocks. DataNode is unaware about the location on which the block it is stored. As a outcome of this, if the NameNode crashes then the data in HDFS is unusable as only the NameNode tells which blocks belong to which file, where each block located exactly etc. [10]

IV. Challenges of Big Data

As data is increasing day by day, the management of data and Integration of data from the different sources have has become a biggest challenges [11]. Big data infrastructure is required to support statistical method and data mining techniques. Fig.5 shows characteristics and yet challenges to Big Data.

1) Volume:

Volume can be *described as data at rest*. Currently the data gathering is in petabytes and is assume to be increase to zettabytes in immediate future. The social networking sites, E-Commerce sites, banking sites etc. are themselves generating data in order of terabytes every day and this amount of data is no doubt difficult to be handled using the present traditional systems. Machine generated data are larger in terms of traditional data [3].

2) Velocity:

Velocity is *all about data in motion* (Stream data within millisecond to respond). Velocity talks about the rate at which data are produced and the speed at which it should be analyzed and acted upon.

*3) Variety***:**

Continuous flow of heterogeneous data is called as Variety, Which consist of structured (tabular data), unstructured (Text, images, audio, and video) and semi-structured (XML) data. Technological advances permits organizations to use variety structured, semi-structured, and unstructured data.

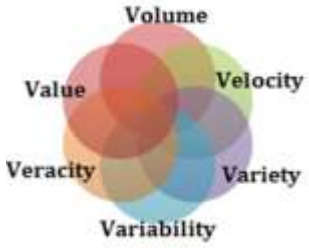

Figure 5. Six V's in Big Data

4) Veracity:

Veracity signifies the unreliability characteristic in some sources of data. For example, data generating from social media are uncertain in nature. However they contain valuable information. Thus the requirement to deal with uncertain data is another aspect of big data, which is handled using various tools developed for management and mining of uncertain data.

5) Variability (and complexity):

Variation in the rate of flow of data is referring as Variability. Complexity is all about to the fact that big data are spawned through a numerous of sources. This enforces a critical challenge: the need to link, match, cleanse and transform data received from various sources.

6) Value:

Value definition, big data are characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analyzing large volumes of such data [12].

Acknowledgment

The Authors are grateful to Department of MCA, Jawaharlal Nehru Engineering College, Aurangabad. The authors would like to thank all the faculty members and staff and also to the Institute Authorities for providing the infrastructure to carry out the research.

References

[1] Feng Xia,Wei Wang, Teshome Megersa Bekele and Huan Liu, "Big scholarly data: A survey", IEEE Transactions on Big Data, vol. 3, Issue 1, pp.18-35, March 2017

[2] D. R. Luna, J.C Mayan,M.J. García A.A. Almerares and M. Househ,"Challenges and Potential Solutions for Big Data Implementations in Developing Countries", NCBI, Yearbook Med Inform. 2014, 9(1), pp.36–41. Published online 2014 Aug 15, doi: 10.15265/IY-2014-0012

[3] Oracle White Paper, "Big Data for the Enterprise", pp.1-16, June 2013

[4] Wikibon,"Big Data Statistics",http://wikibon.org/blog/big-data-statistics, in Analytics and Bigdata [Accessed on Jan 03, 2018]

[5] Dattatrey Sindol, "Introduction to big data", https://www.mssqltips.com/sqlservertip/3132/big-data-basics--part-1--introduction-to-big-data, [Accessed on Jan 02, 2018].

[6] http://www.hjo3.net/bytes.html, [Accessed on Jan 02,2018]

[7] Ibrahim Abaker et al,"The rise of big data on cloud computing: Review and open research issues", ELSEVIER, Information Systems 47, pp. 98–115, 2015

 ISSN: 2454-4248
322 – 325

[8] "Application of Big Data in Real Life", Article in Intellipaat, https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life, [Accessed on Jan. 8 ,2018]

[9] Fazlur Rahman, "Beginners Guide- Introduction of Big Data Hadoop", Article in Codeproject, Jan 2017

[10] Mohd Rehan Ghazia, Durgaprasad Gangodkara , "Hadoop, MapReduce and HDFS: A Developers Perspective" , Published by Elsevier B.V., International Conference on Intelligent Computing, Communication & Convergence(ICCC-2015), 1877-0509, pp.45-50,2015

[11] Hammond WE, Bailey C, Boucher P, Spohr M, Whitaker P., "Connecting Information To Improve Health", Health Aff (Millwood) , 29(2), pp. 284–288, Feb 2010

[12] Amir Gandomi, Murtaza Haider," Beyond the hype: Big data concepts, methods, and analytics",ELSEVIER,International Journal of Information Management 35 , pp. 137–144, 2015.