Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 5, Issue. 12, December 2016, pg.160 – 164

Big Data – A Growing Technology

Kusum Munde[1], Nusrat Jahan[2]

[1]Department of Computer Engineering, Savitribai Phule Pune University, India

[2]Department of Computer Engineering, Savitribai Phule Pune University, India

[1] kusum15.it@gmail.com; [2] nusratkota@gmail.com

Abstract— In recent years, Big data is increasingly becoming popular in business. Every communication is generating huge data. Data may be a analogue or digital data. Data may be present in various formats. Unstructured data storage is key feature of big data. Big data deals with large amount of data. Big data analytics is very important to get business solutions & to increase productivity of business. Further, Big data analytics is used to make decisions.

Keywords— Types of data, Characteristics, Architecture, Challenges, Technologies, Applications

I. INTRODUCTION

Now a days, big data is being generated rapidly. Big data describes the large volume of structured and unstructured data which is difficult to process using traditional database and software techniques. Data are generated from different sectors like Education, healthcare, retail, transportation, banking, insurance, communication Medias, social networks, Mobiles devices, industries and many more. This data generation may be in the form of texts, numbers, images, audio, video etc. Examples Of 'Big Data' includes:

1) The New York Stock Exchange generates about one terabyte of new trade data per day.
2) Statistic shows that 500+terabytes of new data gets ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
3) Single Jet engine can generate 10+terabytes of data in 30 minutes of a flight time. With many thousand flights per day, generation of data reaches up to many Petabytes [1].

Big Data Analytics plays an important role. It is used to find out information and knowledge from big data. Big data analytics is the process of collecting, organizing and analyzing large sets of data to discover hidden patterns, finding correlations, extracting useful business information [2].

II. TYPES OF DATA

Big Data includes voluminous, high velocity, and extensible variety of data. It consists of three types of data.

1) ***Structured data:***

- It contains data which is represented in the form of rows and columns.
- 5-10% data is structured data.
- e.g. Relational data. An 'Employee' table in a database.

2) Semi Structured data:

- These types of data have some organizational properties.
- 5-10% data is semi structured data.
- e.g. json, NoSQL, XML data. We can store Personal data in a XML file.

3) Unstructured data:

- Almost 80% of data generated is unstructured.
- This data is generated from individuals, organizations or machines.
- e.g. Word document, videos, E-mails, PDF, Text, data generated from satellites Output returned by 'Google Search' [3, 4].

III. CHARACTERISTICS OF BIG DATA

Five V's of big data are as follows. Following fig1 shows the characteristics of Big data: [4][5].

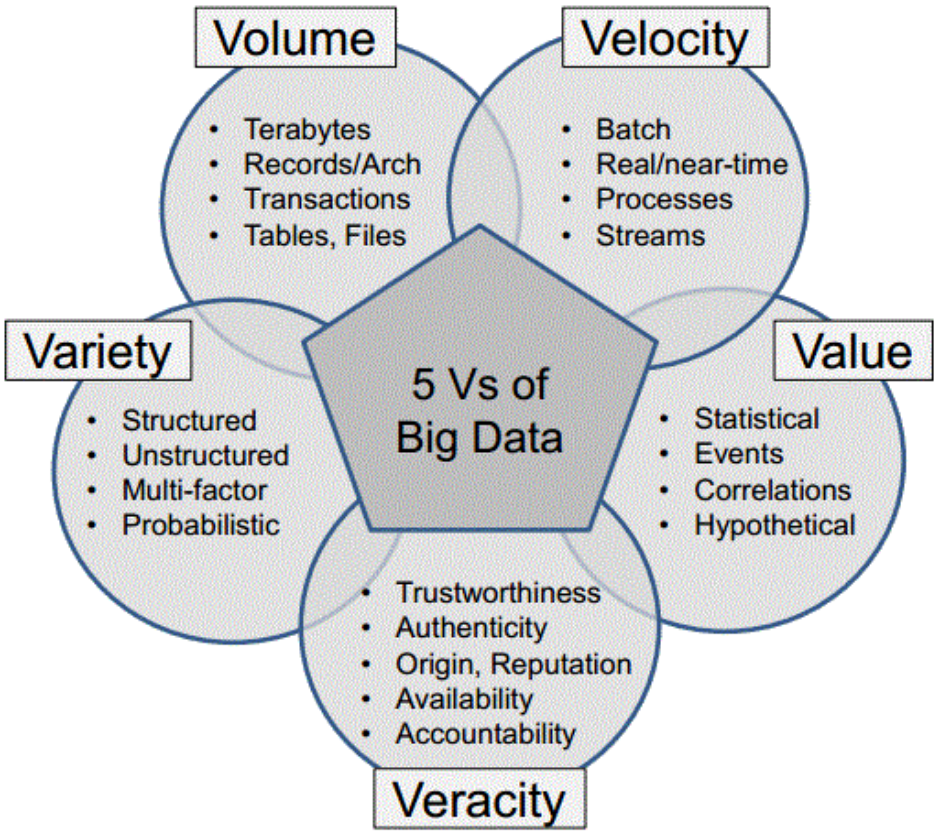

Fig. 1 Characteristics of big data

1. Velocity:

It refers to data in motion. It is the rate or speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

2. Volume:

he quantity of generated data is considered here. The size of the data determines the value and potential of the data under consideration. Large amount of data is generated continuously from various activities. Data which is in terabyte, petabyte and beyond is big data.

3. Variety:

It refers to the type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Data comes from different sources & data is present in different types.

4. Veracity:

The quality of captured data can vary greatly, affecting accurate analysis. Veracity defines that the data is being stored, and mined meaningful to the problem being analysed .It also refers to the noise and abnormality of data [6].

5. Value:

It is good to have access to big data but unless we can turn it into value it is useless. So 'value' is the most important V of Big Data. Value refers to our ability turn our data into value. It is important that businesses make a case for any attempt to collect and leverage big data. It is easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of the business value it will bring.

Enormous information can convey esteem in any zone of business or society:
It helps organizations to better comprehend and serve clients: Examples incorporate the proposals made by Amazon or Netflix. It permits organizations to advance their procedures: Uber can anticipate request, powerfully cost travels and send the nearest driver to the clients. It helps us to enhance security [7].

IV. BIG DATA ARCHITECTURE

Fig 2 gives Big Data Architecture. It shows various components are associated with each other. In Big Data, various different data sources are part of the architecture. The most essential layers of the architecture are extract, transform and integration. Data is stored in relational as well as non relational data marts and data warehousing solutions. As per the business need data is processed as well converted to proper reports and visualizations for end users. Software & hardware are the most important part of the Big Data Architecture. In the big data architecture hardware infrastructure is extremely important and failure over instances as well as redundant physical infrastructure is usually implemented [8].

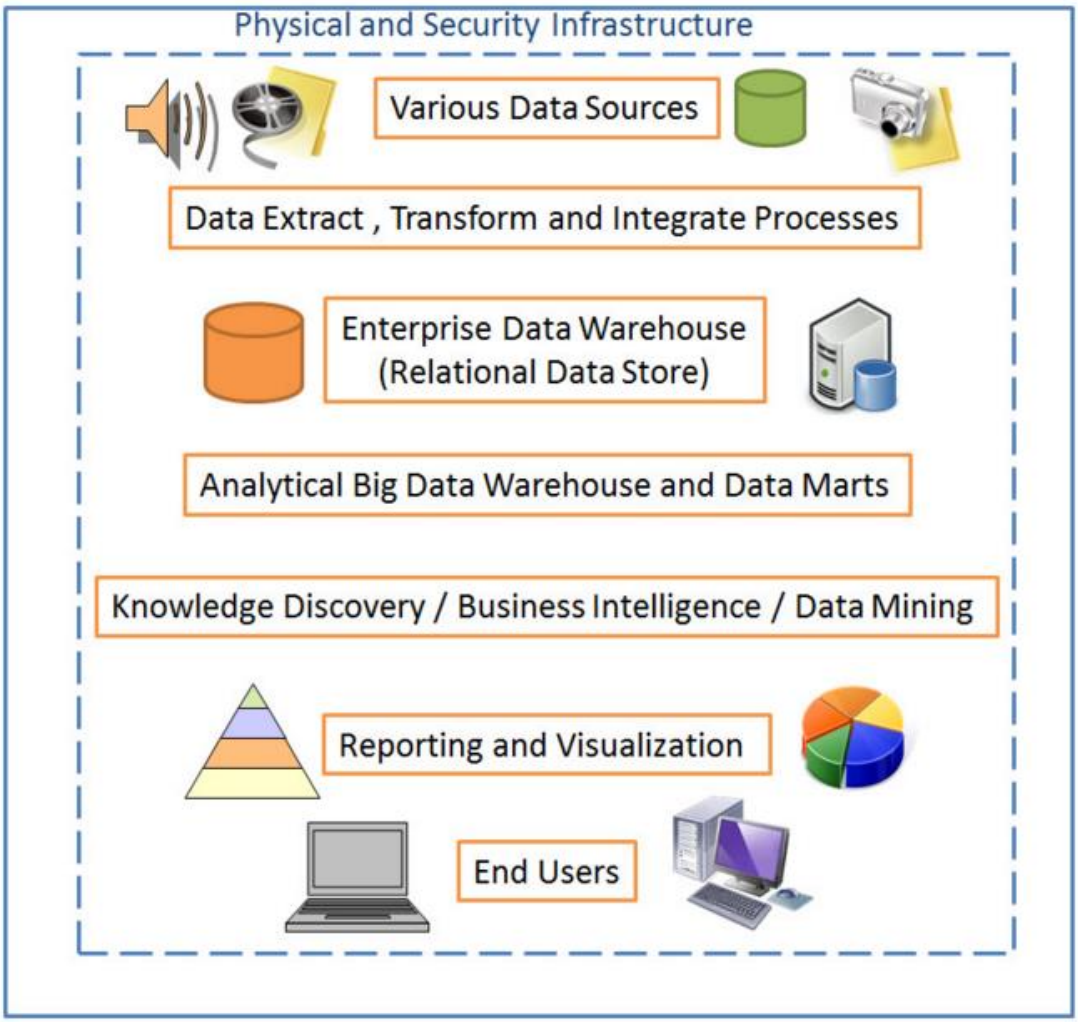

Fig 2. Big Data Architecture

V. CHALLENGES OF BIG DATA

1. Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time [9].
2. Data change very fast and the "timeliness" of data is very short, Due to the rapid changes in big data. If companies unable collect the required data in real time then they may obtain outdated and invalid information [9].
3. Data quality-decision-making will not be appropriate, if the data is not accurate. Only the data that conform to the relevant uses and meet requirements can be considered good quality data [9].
4. Data Acquisition: it is difficult to get the context into which data has been generated, filter non relevant data and to compress data [10].
5. Information Extraction and Cleaning: The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.
6. Data is transformed in order to extract information and present this information in a suitable format for analysis. Data may be uncertain. Data cleaning and data quality verification becomes complex [9, 10].
7. Data Integration, Aggregation and Representation: data can be heterogeneous and may have different metadata.
8. Data integration needs considerably more human efforts. Different data aggregation and representation strategies may be needed for different data analysis tasks [10].
9. Query Processing, and Analysis: Methods should deal with noisy, dynamic, heterogeneous, untrustworthy data and data characterized by complex relations. Sometimes noisy and uncertain data can be more valuable for identifying more reliable hidden patterns as compared samples of good data. Supporting query processing and data analysis requires scalable mining algorithms and powerful computing infrastructures [10].
10. Interpretation: Results should be interpreted by decision makers [10].

With these challenges, data visualization, visualizing results, security, dealing with uncertain data such as outliers is also a big challenge.

VI. TECHNOLOGIES USED TO HANDLE THE BIG DATA

A. NoSQL database:

NoSQL database are commonly used. Database like CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper etc, are also used.

B. Massively Parallel Processing and MapReduce:

MPP includes multiple processors, in this each processor works on different parts of the program and has its own operating system and memory. MPP processors communicate using a messaging interface. MapReduce is a computational approach. Generally used in distributed systems for processing large datasets.

C. Storage:

For storage most commonly Amazon Simple Storage Service (Amazon S3) and Hadoop Distributed File System (HDFS) are used. HDFS is a Java-based file system that provides scalable and reliable data storage [11].

VII. GOVERNMENT APPLICATIONS

- Big-data analysis was in parts, responsible for the BJP and its allies to win a highly successful Indian general election 2014.
- The Indian government utilizes numerous techniques to ascertain how the Indian electorate is responding to government actions, as well as ideas for policy augmentations.
- Joining up Data: The weather challenges in winter 2014 a local authority blended data about services, such as road gritting rotas, with services of people at risk, such as 'meals on wheels'. The connection of data allowed the local authority to avoid the any weather related delay[12].
- Google's DNAStack compiles and organizes DNA samples of genetic data from around the world to identify diseases and other medical defects. These fast and exact calculations eliminate any 'friction points,' or human errors that could be made by one of the numerous science and biology experts working with the DNA.

- DNAStack, a part of Google Genomics, allows scientists to use the vast sample of resources from Google's search server to scale social experiments that would usually take years, instantly [13, 14].

VIII. CONCLUSION

Big data is a current buzzword in business. Business efficiencies are increased in terms of reduced cost, reduced risks and increased performance. Big data is used in various fields such as Medical, Education, Retail, Banking, Transportation, Social media etc. As, large amount of data is already present. Now what kind of data need to be collected and where to store this rapidly growing data in future is seems to be a challenge. However, the utilization of huge information will turn into a key premise of rivalry and development for individual firms.

REFERENCES

[1] http://www.guru99.com/bigdata-tutorials.html.

[2] Dr. Siddaraju 1 , Sowmya C L 2 , Rashmi K 3 , Rahul M 4,"*Efficient Analysis of Big Data Using Map Reduce Framework",* International Journal of Recent Development in Engineering and Technology, ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014).

[3] Prof. A. R. Wasukar 1 ,Prof. P. A. Pawade 2," *A REVIEW ON WHAT IS BIG DATA AND HOW TO HANDLE THE ENORMOUS DATA THROUGH BIG DATA",*4th international conference on recent innovations in science , Engineering and Management, India International Centre, New Delhi.ISBN:978-81- 932074-6-8,march 2016.

[4] Hilbert, Martin. "*Big data for development: A review of Promises and challenges. Development Policy Review.*" Martinhilbert.net. Retrieved 2015-10-07.

[5] Hilbert, M (2015)." *Digital Technology and social change*" [Open Online Course at the University of California] (freely available). http://www.youtube.com/watch?v=XRVIh47sA&index=51&list=PLtjBSCvVVCU3rNm46D3r85efN0hrz juAlg Retrieved from https://canavas.instructure.cpm/course/949415.

[6] http://insidebigdata.com/2013/09/12/beyond-volume-velocity-issue-big-data-veracity.

[7] http://www.ibmbigdatahub.com/.

[8] http://blog.sqlauthority.com/2013/10/04/big-data-basics-of-big-data-architecture-day-4-of-21/.

[9] Cai, L. & Zhu, Y., (2015). "*The Challenges of Data Quality and Data Quality Assessment in the Big Data* Era". Data Science Journal. 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002.

[10] Elisa Bertino "*Big Data - Opportunities and Challenges*", Panel Position Paper 2013 IEEE 37th Annual Computer Software and Applications Conference.

[11] Nirali Honest 1 and Atul Patel 2,"*A S URVEY OF BIG D ATA A NALYTICS*", International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016.

[12] Siddharth Singh, 2Tuba Firdaus, 3 Dr. A.K. Sharma,"*Survey on Big Data Using Data Mining*", 2015 IJEDR | Volume 3, Issue 4 | ISSN: 2321-9939.

[13] "These six great neuroscience ideas could make the leap from lab to market". The Globe and Mail. 20 November 2014. Retrieved 1 October 2016.

[14] "DNAstack tackles massive, complex DNA datasets with Google Genomics". Google Cloud Platform. Retrieved 1 October 2016.

