Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 2, February-2019, ISSN 2230-7540

Feature, Planning and Uses of Big Data

Ankit Garg*

Assistant Professor in Computer Science, R.K.S.D. (P.G.) College, Kaithal

Abstract – In the era of Computing Technology, data plays a role of lifeblood in various fields to make decisions. Without consistent and reliable data, the organizations cannot make decisions at right time. To design the data in an effective way, traditional methods were not proved to be effective because data is increasing from Kilo Bytes to Giga Bytes and then to Petta Bytes. Because of large diversity of data that is changing very rapidly and massively, big data is an emerging technology that takes an initiative to overcome these problems. In addition to the large volume of data, big data also handles varieties, velocity, veracity and value of data. Various tools being used to analyze the big data like Hadoop, MapReduce are also discussed in this paper. Architecture of Hadoop which is used to convert diverse and unstructured data into uniform and structured form is also described in this paper. Big data plays a significant role in various areas like banking, industries, E-Governance etc.

Keywords — Big Data, 5 V's, Architecture, Hadoop, MapReduce, Applications of Big Data.

---------------------------------X---------------------------------

I. INTRODUCTION

Here, in the era of emerging technology each field whether it is banking, hospitals, school, government organization, social media and lot more produces a lake of data in a structured, semi-structured and unstructured format called **BIG DATA.** Big data describes collection of data sets so large and complex that it becomes difficult to process, manage using relational database system and traditional database system. Its tool like data mining, text analytics, predictive analytics, statistics, cloud computing are so large that are impractical to manage with traditional software tools.

Big data can be collected through two sources: internal and external sources. Its efforts are currently focus on analyzing internal 0 data to extract result and whereas some organization look outside to get such as social media. Internal data are enterprise data transaction details (which occur in e-shopping, banking, etc.) log files (consider example of railway ticket booking at IRCTC and we make payment in reserving our seats. The amount of respective seats is deducted from account but the seat is not reserved which leads to inconsistency then this log file generate unique Id. With the help of this Id within 24 hours our deducted amount will be cash back in an account. External data sources are social media data which is increasing day by day and platform promoted this is Facebook, Instagram Google Duo, Pinterest, Twitter and a lot more. In short, big data is a combination of social data and enterprise data. The 5V's of Big Data: Volume, Velocity, Value, Variety and Veracity as shown in Figure 1.

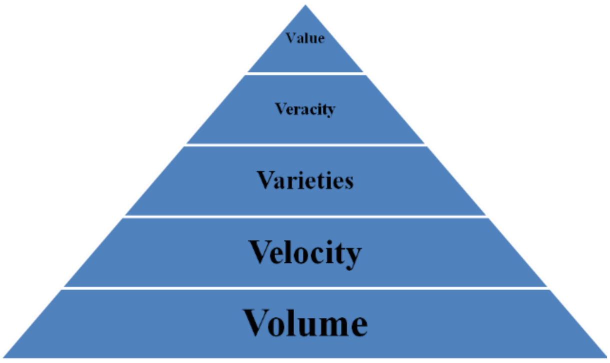

Figure 1 Representation of 5 V's of Big Data

VOLUME:-

Big Data is like a pyramid where volume is the base. Today, Facebook gives 600 TB of data everyday consisting of 20 billion messages, 5 billion times, that the 'like' is pressed and photos uploaded and a lot more. Data is produced at an astronomical rates from gigabytes to Petabytes and now Zettabytes because sources of data has increased from enterprises, social media sites, cell phones , cars , M2M sensors etc. in each and every second.

VELOCITY: -

As we know data is being generated at an alarming rate. Velocity refers to the increasing speed 0 at which data is created so the increasing speed at which data can be analysed, processed and stored in a relational database. After every 60seconds there are more than 1 Lakh tweets, 695000 status update on Facebook, 11,000,000+ instant messages, 6,900,445 google searches, 217+ new mobile users, 1820 TB data created and a lot more. It is generally in tabular form in RDBMS, flat files, multi-dimensional DBS, legacy system.

VARIETIES:-

Variety is defined as the different types of data we can now use. We have different kinds of data. Structured data is that data in 0 which the schema is previously defined and collected from various enterprise. But today big data generally covers data in which schema is not defined generally covers audio, video, images, social media updates, etc.

VERACITY: -

Data being stored in database comes from different sources may have anomalies, bugs, noise and unfiltered. This is the most crucial task of Big Data to maintain the worthiness, accuracy and preciseness of the data.

VALUE: -

This is at the top of the pyramid of big data. Value includes a large volume and variety of data that is easy to access and delivers quality analytics that enables informed decisions.

II. ARCHITECTURE OF HADOOP

Big data architecture is outlined to manage the processing and then analyzing data that is very big and complicated for conventional database. Enterprise worry about managing big data is increasing day by day. In order to respond to changing conditions of market and needs of customers, big data imparts advantage to collect process and examine the huge amount of data. Hadoop is very important big data tool. For implementing infrastructure of big data, Hadoop and MapReduce are good choices 0.

Hadoop is very similar to existing distributed file system. For data storage, master slave architecture is used and data processing for transforming very large dataset, Hadoop MapReduce is used. Basically 5 blocks are used as shown in Figure 2.

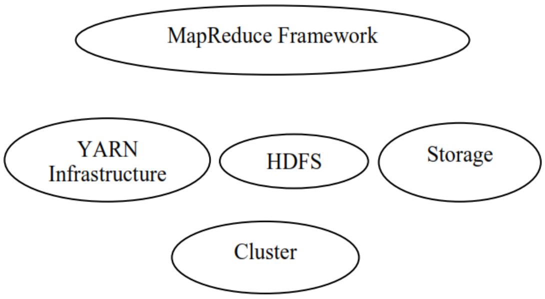

Figure 2 Architecture of Hadoop

- **Cluster: -** Clusters are designed for accumulating and analyzing big amount of disorganized data in an environment which is distributed. Hadoop clusters are used in increasing speed of analyzing data. Scalability of clusters 0 is very high. Hadoop clusters are very reliable which is invulnerable to failure because data can be copied into other cluster nodes.

- **YARN Infrastructure: -** YARN (Yet Another Resource Negotiator) is the resource management and arranging jobs. It is considered as a big scale, scattered operating system. YARN 0 is in control for delivering computational resources like memory, CPU etc. Resource management is provided by YARN.

- **HDFS**: - Hadoop distributed file system acts as a main system for storing data. HDFS works on hardware having very low cost and HDFS 0 is fault tolerant. It provides consistent method for big data management. Master/Slave architecture is used by HDFS. Files are stored redundantly so that if one of the file is lost, we can prevent system from complete loss of file.

- **Storage**: - Except HDFS, different companies provide different solutions to store data. For ex, Amazon provides simple storage services.

- **MapReduce Framework**: - Large amount of data can be processed by MapReduce framework. Various languages support MapReduce programs like Python, Java, C++ etc. Two tasks are included which are Map and Reduce. The function of Map is used to convert one dataset into another, where separate elements are split to get two tuples which contain pairs of key/value. In Reduce, data tuples are combined into very small tuples set. Basically 3 stages are involved. MapReduce is efficient in providing 0 distributed processing to huge set of data.

1. **Mapping: -** Input data in form of input file is passed in mapping stage and then data is processed to create very small portion of data.

2. **Shuffling: -** The task of this phase is to merge the records from output of mapping phase.

3. **Reducing:** - This phase is used to collect the output from shuffling phase and then

only single output is returned, which is stored in HDFS.

III. APPLICATIONS OF BIG DATA

Big data has totally changed and revolutionized the way businesses and organization work. In this era where every aspect of our day to day life has been revolutionalized there is a huge volume of data has been generating from various digital sources. Various tools are used to analyze the big data like Hadoop which contributes in various fields.

- **Big Data contributes to Education field:-**

Education is filled with enormous and tremendous amount of data related to students, faculties, courses, resources, curriculum, results, achievements and lot more. Proper analysis of the data provides insights which can be useful and fruitful to improve the working of educational institutes. These are the following aspects where big data brings great change like refraining course material 0 , grade system, career prediction, student portfolio, online portal sharing digital course material, digitalized record of every student , decrease dropout. Big data also contributes to online learning.

- **Big Data contributes to Healthcare:-**

Healthcare system 0 is one of the complex system where patient care, large amount of data driven by record keeping, compliance regulatory requirement is required. A lot of data is collected from various sources such as surveys, electronic financial transactions for health insurance claims, computer-based patient records (CPRs), and disease registries. So it leads to lot of data which must be managed properly. Big Data helps a lot to manage the data and it has some positive and life saving outcomes. Big Data refer to the vast quantities of information created by digitalization of health data from various health resources.

- **Big Data contributes to Governance:-**

From local level to national level and then to international level, US census to local municipalities there are bulk of data in an unstructured way. Big data can be very well used for E-Governance. It helps to capture details of complete activity of online purchase and to classify them as regular tax 0 payer or fraudulent etc. It can also help government to track tax payers and to retrieve the data i.e. missing person / case history , identify the criminals, terrorists etc. can be done very easily Various projects sponsored by the government , their details , schemes can be easily available. So, Big Data play a significant role in managing data of Government agencies in collecting, capturing and creating huge amount of data in multi lingual forms. It also helps in making faster decision and implement the decision with various technology like MAPREDUCE. Big data can be very well used for e governance.

- **Big Data contributes to Transportation:-**

Traffic is increasing day by day 0 and everyday number of nuances to driver safety, from roads to police officers, handling the rise of ride hailing apps, add more public transit option, how to minimize the construction for expanding public transit and lot more. To solve these problems, there is a need of Big Data. Big Data ensures better roads, better routes, safer roadways and new routes.

- **Big Data contributes to Agriculture:-**

“The world doesn’t need any more engineers we didn’t run out of planes and mobiles…. We ran out of food. “Population is increasing day by day and to satisfy the population nothing is more important than food supply. There are various problems which are facing to meet the proper food supply like uncertainty about weather, pests, consumer demand and water use, inefficiencies in planting, harvesting leads to intangible loss. On the consumer end packaging and labelling can lead to wastage of food and illness due to pathogens.

- **Big Data contributes to Banking**: - Banking sector has undergone a transformation in their database day by day. Due to this many banks failed to utilize the information within their own database. Banks has to face various challenges to satisfy customers 0 like privacy, Data quality, data integrity, legal challenges, fraud detection, service delivery, data visualization, general ledger transformation etc. can be resolved by Big data. It helps to generate the data in a manner which meets the bank vision and mission. It solves all the above challenges, make the service faster, provide platform for new application etc.

IV. CONCLUSION

This paper presents the fundamental concepts of big data. Big Data tracks the large amount of data and it discovers hidden patterns and gives the results immediately. Big data provides consistent, integral, reliable, confidential, available and normalized data which comes from diverse sources and different formats (Unstructured, Semi-Structured and unstructured). The variation in data can be handled by the various tools like Hadoop, MapReduce and its architecture of Hadoop is also discussed in this paper. Big data has the potential to transform the various fields like banking, industries, E-governance and agriculture etc.

REFERENCES

1. Al-kahtani, M. S. (2016). Big Data Networking: Requirements, Architecture and Issues. *International journal of wireless and mobile networks (IJWMN), 8* (6).

2. Chan, J. O. (2013). An Architecture of big data analytics. *Communications of IIMA, 13* (2).

3. Charles, P. J., Bharathi, S. T., & Sushmita, V. (2018). Big data- concepts, analytics, architectures-overview. *International research journal of engineering and technology (IRJET), 05* (02).

4. Mukherjee, S., & Shaw, R. (2016). Big Data- Concepts, Application, Challenges and future scope. *International journal of advanced research in computer and communication engineering , 5* (2), pp. 66-74.

5. Nizam, T., & Hassan, S. I. (2017). Big Data: A Survey paper on big data innovation and its technology. *International journal of advanced research in computer science , 8* (5).

6. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare:promise and potential. *Health Information Science and Systems , 2* (3).

7. Siddiqui, A. A., & Qureshi, R. (2017, February). Big data in banking: opportunities and challenges post demonstration in India. *IOSR Journal of computer engineering (IOSR-JCE)*, pp. 33-39.

8. Tawade, S. S. (2018). Applications of big data: Review paper. *International research journal of engineering and technology (IRJET) , 05* (02), pp. 1631-1633.

9. Thomas, M. (2015). A review paper on big data. International research journal of engineering and technology (IRJET) , 02 (09), pp. 1030-1034.

10. Woodside, J., & Amiri, S. (2015, January). The impact of ICT and big data on e-government. *Int'l Conf. on advances in big data analytics*, pp. 116-118.

Corresponding Author

Ankit Garg*

Assistant Professor in Computer Science, R.K.S.D. (P.G.) College, Kaithal