Overview of Big Data And Its Issues

Himani Maheshwari[1], Luxmi Verma[2], Umesh Chandra[3]

[1]*M.Tech Scholar, Dev Bhoomi Institute of Engineering and Technology, Dehradun, Uttarakhand*
[2]*Associate Professor, Dev Bhoomi Institute of Engineering and Technology, Dehradun, Uttarakhand*
[3]*Assistant Professor, Banda University of Agriculture And Technology, Banda, Uttar Pradesh*
(E-mail: Himani_bahmah@yahoo.com, luxmi.verma@gmail.com, uck.iitr@gmail.com)

***Abstract*—**Data is growing in the universe at a very fast pace due to internal and external sources including social media, emails, log files, internet, stock data, transport data and many more. Big data is describes in five V's: volume, variety, velocity, value and veracity. It also insights the technologies behind big data: operational and analytical capabilities. This paper mainly discusses the issues associated with big data.

Keywords*—big data, hadoop, fiveV's, HDFS, MapReduce.*

I. INTRODUCTION

Today, we are living in an information age where data is being generated at an alarming rate and this huge amount of data is termed as Big Data. It consists of large datasets that cannot be managed efficiently by the common database management systems. These datasets range from terabytes to exabytes [1.11].

New technologies, devices, communications are growing day by day as a result amount of data produced by mankind is growing rapidly every year. Approximately, ninety percent of the world's data was generated in the last few years [9]. Data that comes from multiple sources such as database, ERP, systems, weblogs, chat history, maps varies in its format.

Data is obtained primarily from the following type of sources:

Machine Data: It refers to the information generated from sensors, barcode scanners, Global Positioning System (GPS), Radio Frequency Identification (RFID) chip readings.

Social Media Data: It refers to the information collected from various social networking sites like facebook, twitter, LinkedIn, Flickr etc. and online portals.

Black box Data: It refers to the black box of airplanes, helicopters; jets are used to store microphone voices, performance information of airplane, etc.

Stock Exchange Data: It refers to the information about buy and sell of shares, etc.

Broadly, sources of data are categorized as internal and external sources as given in Table 1.

TABLE I. SOURCES OF BIG DATA

Internal Sources (Organizational data)	External Sources (Social data)
Structured or organized data.	Unstructured or unorganized data.
Sources: ERP, customer relationship management, product and sales data, etc.	*Sources:* government, internet, business partners, market research organization, etc.
Used to support daily business needs.	Analyze to understand the entities external to the organization.

On the basis of data received from varied sources, the big data are categorized into three different types: structured, semi-structured and unstructured (Figure 1). 20-30% of existing data are structured and semi- structured data and rest 70-80% is unstructured data [4, 5].

Structured data is organized data in a predefined format and stored in tabular form (Relational Database Management System). Machine generated structured data are generated from sensors and weblogs and human generated structured data are taken as information from human like their names, addresses, age, nationality, etc. Sources of structured data are relational database, flat files (CSV files, tab separated files), legacy database and multidimensional database.

Unstructured data has no clear format in storage. It is a set of data that might or might not have any logical or repeating patterns. Machine generated unstructured data are satellite generated images, scientific data and human generated unstructured data are images, videos, social media, etc. Sources of unstructured data are social media, mobile data, text both internal and external to an organization [7,10].

Semi-structured data is very difficult to categorize sometimes they look like structured or sometimes unstructured. XML or JSON documents, NoSQL database data items are semi-structured data. Sources of semi-structured data are web data in the form of cookies and data exchange formats such as JSON data [6].

II. ELEMENTS OF BIG DATA

According to Gartner, data is growing at the rate of 59% every year; this growth is depicted in terms of the five V's: volume, variety, velocity, value and veracity in Figure 2. Volume is the amounts of data generated by organizations or individuals in term of exabytes and predict to reach zettabytes in the coming years [12]. Veracity describes the rate at which data is generated, captured and shared in real time. Data is being generated at a very fast pace and in varied formats: structured, semi-structured and unstructured. Veracity refers to the uncertainty of data, i.e., where obtained data is correct or consistent[2]. Out of the huge amount of data, only the correct and consistent data can be used for further analysis.

Value is the mechanism to extract useful and meaningful information from huge data.

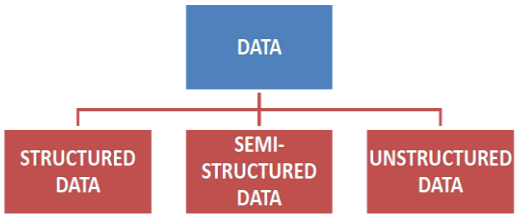

Figure 1: Types of Big Data

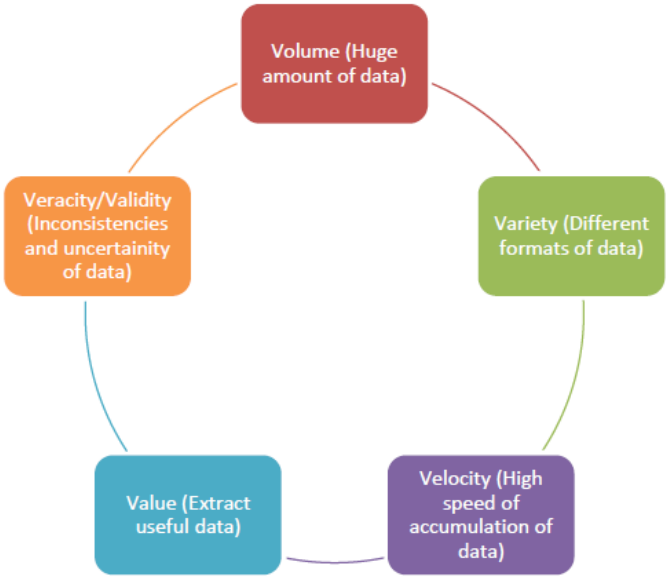

Figure2: Elements of Big Data

III. BIG DATA TECHNOLOGIES

Two technologies used in Big Data are: Operational and Analytical (Table 2). Operational system includes capturing and storing data in real time while analytical system includes complex analysis of all the data. Operational capabilities include NoSql database which deals with real time interactive databases [3]. Analytical capabilities focus on complex queries using Hadoop which handle almost all the data.

TABLE II. TECHNOLOGIES OF BIG DATA

Operational	**Analytical**
Real time interactive databases	Batch Oriented analytic database
NoSQL	*Hadoop*
Operational, Velocity	Analytical, Volume
Online	Offline
Web/Mobile/IoT Apps	Analytical Apps
Millions of customers/consumers	Hundreds of business analysts

you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:An excellent style manual for science writers is [7].

IV. ISSUES WITH BIG DATA

There are three problems associated with big data:

Issue 1: Storing exponentially growing huge datasets:

Data generated in the past two years is more than the previous history in total. By 2021, it is predicted that total digital data will grow to forty four zetta bytes approximately or more and about one and half mega byte of new information will be created every second for each individual and to store all this data is tedious task.

Issue 2: Processing data having complex structure:

Data is not only huge but present in different format-structured, semi-structured and unstructured given in Table 3.

Issue 3: Processing data transfer:

The data is growing at much faster rate than compared to disk read/write speed and process this data is hilarious.

V. SOLUTIONS OF THE ISSUES ASSOCIATED WITH BIG DATA

Hadoop comes as a solution of all problems of big data. Hadoop is a framework that allows us to store and process large datasets in parallel and distributed fashion. It can easily handle vast amount of data economically with commodity hardware cluster. It is scalable and fault tolerant framework. It is not only used as a storage system rather it can processed data also. The Apache Hadoop is an open source framework which is written in Java by Yahoo, IBM and Cloudera. The

Hadoop ecosystem can be defined as a comprehensive collection of tools and technologies that can be successfully implemented and deployed to provide big data solutions in a cost-effective manner. Hadoop Distributed File System (HDFS) and Map reduce are the two core components of the hadoop ecosystem as depicted in Figure 3 and other elements are given in Table 3.

TABLE III. DIFFERENT FORM OF BIG DATA

Structured	Semi-structured	Unstructured
Organized data format	Partial organized data	Un-organized data
Data schema is fixed	Lacks formal structure of a data model	Unknown schema
e.g. RDBMS data	e.g. XML and JSON files	e.g. multimedia files

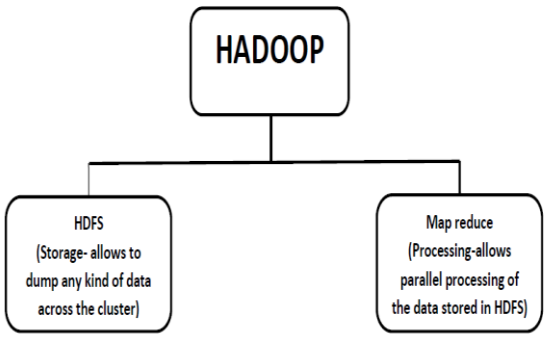

Figure 3: Main Components of Hadoop System.

TABLE IV. OTHER COMPONENTS OF HADOOP ECOSYSTEM

Component Name	Working
Sqoop	Data exchange
Flume	Log control
Zookeeper	Coordination
Pig	Scripting
Hive	SQL Query
Oozie	Workflow
Mahout	Machine Learning
Hbase	Columnar data store
R Connection	Statistics

MapReduce and HDFS provide the necessary services and basic structure to deal with the core requirements of big data solutions. Other services and tools of the ecosystem provide the environment and components required to build and manage purpose driven big data applications.

HDFS is a storage unit of hadoop which divides input data into smaller chunks and stores it across the cluster. It is scalable as per requirement. It is a specially designed file system for storing huge datasets with clusters of commodity hardware with streaming access patterns. It is based on the concept of WORA (Write Once Run Anywhere). There are five basic components of HDFS are: Name Node, Secondary Name Node, Job Tracker, Data Node and Task Tracker. The first three are master services and others are slave services. Every master service can communicate to each other and each slave service can also communicate to each other. For master Name Node slave node is Data Node and for master Job Tracker node slave node is Task Tracker node. NameNode (the master) is the main node that deals with the file system and stores the metadata for all the documents and indexes. DataNodes (slaves) store and recover blocks when they are asked by NameNode and report back to it. DataNodes are collection of commodity hardware in the distributed environment and data is replicated in three DataNodes. When one data node is down, the data can be accessed through any other data node which contains the replication. It is very cost effective as it is used very simple commodity hardware to process and store data blocks [8].

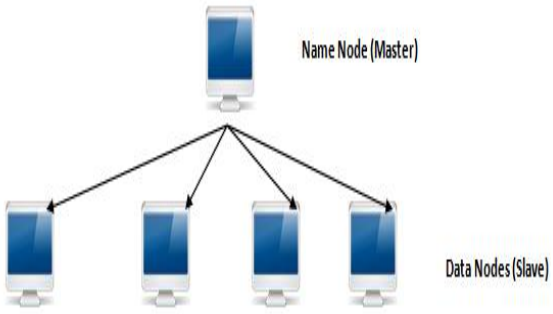

Figure 4: Hadoop Cluster.

MapReduce enables computational processing of data stored in a file system without the requirement of loading the data initially into a database. It primarily supports two operations: nap and reduce. These operations execute in parallel on a set of worker nodes. MapReduce works on a master slave approach in which the master process controls and directs the entire activity, such as collecting, segregating and delegating the data among different workers.

VI. CONCLUSION

Handling huge volumes of data generating from billions of online activities and transactions requires continuous up gradation and evolution of Big Data.

This paper highlights the problems associated with big data i.e. storing huge datasets, processing complex data and processing data transfer. One such technology to solve these problems is Hadoop, which is an integral part in almost all big data processes. MapReduce and Hadoop distributed file system are the two core components of the Hadoop ecosystem that helps to manage big data along with other components.

References

[1] Chen, J.; Liang, Q.; Wang, J. Secure transmission for big data based on nested sampling and coprime sampling with spectrum efficiency. Secur. Commun. Netw. 8, 2447–2456, 2015.

[2] Colombo, P.; Ferrari, E. Privacy Aware Access Control for Big Data: A Research Roadmap. Big Data Res. 2, 145–154, 2015.

[3] Dharminder Yadav, Himani Maheshwari and Umesh Chandra, "Big Data Hadoop: Security and Privacy", Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering Applications, 2019.

[4] Dharminder Yadav, Umesh Chandra , " Modern Technologies of Big Data Analytics: Case study on Hadoop Platform " , International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 6, Issue 4, 2018, pp. 044-050.

[5] E. S. A. Ahmed and R. A. Saeed, A Survey of Big Data Cloud Computing Security, International Journal of Computer Science and Software Engineering, vol. 03, Issue 01, pp. 78-85, 2014.

[6] J. Shafer, S. Rixner, and A. L. Cox. The Hadoop Distributed File system: Balancing Portability and Performance. Proc. of 2010 IEEE Int. Symposium on Performance Analysis of Systems & Software (ISPASS), March 2010, White Plain, NY, pp. 122-133.

[7] Shagufta Praveen , Umesh Chandra, "NoSQL: IT Giant Prespectives", International Journal of Computational Intelligence Research, Volume 13, Number 8 ,pp. 2125-2133, 2017.

[8] Shagufta Praveen, Umesh Chandra, "Influence of Structured, Semi Structured, Unstructured data on various data models", International Journal of Scientific & Engineering Research, Volume 8, Issue 12, 2017.

[9] Stephen, J.J.; Savvides, S.; Seidel, R.; Eugster, P. Program analysis for secure big data processing. In Proceedings of the 29th ACM/IEEE international conference on Automated software engineering, Vasteras, Sweden, pp. 277–288, 15–19 September 2014.

[10] Thilakanathan, D.; Calvo, R.; Chen, S.; Nepal, S. Secure and controlled sharing of data in distributed computing. In Proceedings of the 16th IEEE International Conference on Computational Science and Engineering (CSE 2013), Sydney, Australia, pp. 825–832, 2013.

[11] V.N. Inukollu, S. Arsi, and S.R. Ravuri, Security issue Associated with Big data in Cloud Computing, Int. J. of Network Security and its Application, vol. 6, Issue 3, 2014.

[12] Weber, A.S. Suggested legal framework for student data privacy in the age of big data and smart devices. In Smart Digital Futures; IOS Press: Washington, DC, USA, vol. 262, 2014.

Dr. Himani Maheshwari is a M.Tech scholar in the Department of Computer Science and Engineering, DBIT, Uttarakhand Technical University, Dehradun. She received her Master of Computer Applications degree with honors from Uttar Pradesh Technical University and PhD degree from Indian Institute of Technology (IIT) Roorkee, in 2011, and 2017 respectively.

Dr. Luxmi Sapra is an Associate