Volume 8, No. 5, May-June 2017

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Big Data: A Survey Paper on Big Data Innovation and its Technology

Tasleem Nizam
Dept. of Computer Science & Engineering
JamiaHamdard(Hamdard University)
New Delhi, India

Syed Imtiyaz Hassan
Dept of Computer Science& Engineering
JamiaHamdard(Hamdard University)
New Delhi, India

Abstract:Any kind of datasets which are so large and complex which becomes difficult to process them using traditional data processing applications is considered as Big Data. While handling huge dataset different challenges may be faced by the user.One can get additional large data from analysis of single large set of related data as compared to separate smaller dataset with the same amount of data. For example correlations to be found to "prevent diseases, spot business trends, combat crime and so on." It is difficult to work with Big Data using traditional database management systems and visualization packages and desktop statistics requiring instead "massively parallel software running on hundreds, or even thousands of servers". Data sets with sizes beyond the capability of usually used software tools to capture, manage, and process data within a tolerable elapsed time, are included in Big Data. Big Data "size" is a constantly increasing, as of its ranging from a few dozen terabytes to many petabytes of data. So, Big Data is a collection of techniques and technologies that need new forms of integration to uncover large hidden values from large datasets that are complex, diverse and of a massive scale. Big Data environment is used to arrange and analyse the various types of data. Big Data is data which is so large in volume, so various in variety or moving with high velocity is called Big Data. Acquiring and analysing Big Data is a challenging task as it includes large distributed file systems which should be flexible, fault tolerant and scalable. Different technologies used by big data application to handle the huge amount of data are Hadoop, Map Reduce, etc. In this paper, firstly the definition of big data is presented.. Following that, the architecture of different technologies which are used for handling Big Data is defined. Finally, applications of Big Data system is represented.

Keywords:Big data, Hadoop, MapReduce, HDFS, YARN

I. INTRODUCTION

In today environment, data is generating from various sources.
These data is of different varieties. Capturing, retrieving, extracting, analyzing, manipulating and storing of these data is bit critical. This amount of massive data is considered as Big Data.
Big data is defined as data which is not only very huge, but also high in velocity and variety, which makes them difficult to handle using traditional tools and techniques. These data is generated from different social media like Twitter, Facebook etc., from different transactions done in company's databases, from supply chain scenarios which supplies tons of data, for example given number of scanners etc.
What Enables Big Data: There are some reasons which causes for generating Big Data. These are as follows-

- ***Commodity Hardware Compatibility:*** There are many small organizations which have cheap servers and cheap devices that makes them unable to access data which is very large.
- ***Reduction in Storage Costs:*** The down fall in the costs of storage devices makes cheaper to store data in large scale.
- ***Open Source Ecosystem:*** Big Data lies on open source software which makes organizations enable to use big data effectively comparatively to buy own proprietary or personal software for data handling.
- ***The web Economy:*** Today, most of the organizations or customers are based on web services. Maximum work of organizations are based on web services. It provides high access of data [1].

So; there is the great demand of Big Data Analytics. Big Data Analytics process the complex and massive datasets. This dataset is different from structured data in terms of five parameters, represented through Fig. 1. –volume, variety, velocity, value, veracity (5V's). These five V's (volume, variety, velocity, value, veracity) are
the basic challenges for big data management. These are as follows:

- ***Volume:***
  Data is generating day by day of all types ever KB, MB, TB, PB, YB, ZB of information. The data becomes into huge sized files. Too much volume of data is important issue of storage. This issue is resolved by deducting storage cost.
- ***Variety:***
  Data sources are too much heterogeneous. The files generates in various formats and of any type. That file may be unstructured or structured such as text, log files, audio, videos and more. These varieties are endless, and the data enters into the network without having been qualified in any way.
- ***Velocity:***
  The generated data comes at very high speed. This data velocity is also a main challenge. For example, credit card transactions and the social media messages done in millisecond and data generated from these are kept in to databases.
- ***Value:***
  This is a most important V in big data. Value is main property for big data because it is important for IT infrastructure system, businesses to keep large amount of values in database.
- ***Veracity:***

In today scenario, we are dealing with high velocity, variety and volume of data. These generated data are not going to be 100% correct, there may be some dirty data. Big data and analytics technologies work with these types of dirty data also [2].

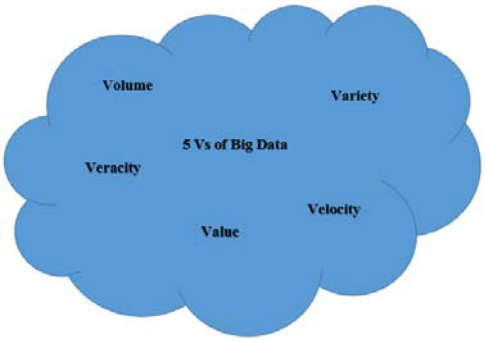

Figure 1: Parameters of Big Data

So big data is a buzz word which is giving a new path to IT world for handling datasets. In this paper, a review on Big Data is represented which includes definition of big data, different technologies and the application of big data. Section 2 gives the literature survey which is based on Big Data. Section 3 introduces different technologies used for handling Big Data. Section 4 includes application of Big Data in different fields.

II. LITERATURE SURVEY

The author [3] stated that Hadoop Map Reduce is an open source, large scale software framework which is dedicated to distributed, scalable, data-intensive computing. This framework first breaks up large data into smaller chunks which are parallel and handles scheduling. It maps each piece to an intermediate value using Map function and then reduces intermediate values to a solution using Reduce function. So MapReduce is a good way to solve those problems in which large dataset can be broken into smaller pieces parallel in a distributed environment.

The author [4] elaborated the importance of some of the technologies that can be used for handling Big Data like Hadoop, HDFS and Map Reduce. The author stated about various schedulers which can be used in Hadoop and about the different technical aspects of Hadoop also. The author also emphases on the importance of YARN which overpowers the restrictions of Map Reduce.

The author [5] have surveyed different technologies to handle the big data and its architecture. Under this paper, the author has also discussed the challenges of Big Data (volume, variety, velocity, value, veracity) and different advantages and disadvantages of these technologies. In this paper author discussed an architecture of Big Data using real-time NoSQL databases, Hadoop HDFS distributed data storage and MapReduce distributed data processing over a cluster of commodity servers.

The author [6] stated about Big Data definition and enhanced the definition by giving the 5V Big Data properties: Volume, Variety, Velocity, Value, Veracity, and also suggested other dimensions for Big Data analysis and taxonomy, in particular contrasting and comparing Big Data technologies in industry, e-Science, social media, , business, healthcare. In today scenario, working with constantly increasing volume of data, the scientific analysis methods can be offered by modern e-Science to industry, while industry can take advanced and fast evolving Big Data technologies and tools to wider public and science.

The author [7] detailed that data is growing rapidly day by day. Now these data is not only limited within gigabyte, instead it is more than this measure. The data generated is not only too huge in size, it is heterogeneous also. Big Data analysis tools like Map Reduce over Hadoop and HDFS, promises to support organizations better recognize their customers and the marketplace, which hopefully leads to competitive advantages and to better business decisions.

III. TECHNOLOGIES USED FOR BIG DATA

With the growth of technology and the improved multitudes of data flowing in and out of organizations daily, there has become a need for more efficient and faster ways of evaluating such data. Having heaps of data on hand is no longer sufficient to make efficient decisions at the right time. Such data sets can no longer be effortlessly evaluated with traditional analysis techniques and data management and infrastructures. Therefore, there is a need for new methods and tools specialized for big data analytics, as well as the required architectures for managing and storing such data. Some Big Data tools are as follows-

A. ***Hadoop:*** Hadoop is an open source software which is based on Java programming framework. Hadoop is a part of Apache project which is given by Apache Software Foundation. Hadoop framework supports storage and processing of extremely large dataset in a distributed environment.Hadoop makes it possible to handle thousands of terabytes of data and to run applications on systems having thousands of commodity hardware nodes. Its distributed filesystem enables speedy data transfer rates among nodes and also permits the system to continue functioning in case of a node failure. Because of this approach, it lowers the risk of unexpected data loss, and catastrophic system failure even if a major number of nodes become defective or inoperative. So, Apache Hadoop is emerged as a foundation for big data processing tasks, such as business and sales planning, scientific analytics and processing huge volumes of sensor data, including from internet of things sensors [7] [8].

B. ***Hadoop Architecture view:*** Hadoop has many similarities with existing distributed file systems. It follows a Master Slave architecture for the analysis and transformation of large datasets using HadoopMapReduce paradigm. There are generally five building blocks inside this runtime environment. Fig. 2 represent them. These are as follows-

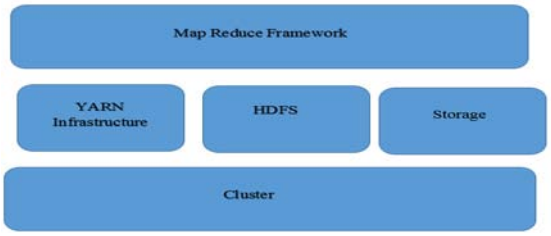

Figure 2: Structure of Hadoop

- ***Cluster:***Generally any set of tightly connected or loosely connected computers that work together as a single

system is considered as a Cluster. In other words, a computer cluster which is used for Hadoop is called Hadoop Cluster. The Hadoop cluster is the set of nodes/host machines.
For analyzing and storing vast amount of unstructured data in a distributed computing environment Hadoop cluster is used. It is a computational cluster that run on low cost commodity computers. These nodes may be divided in racks. Hadoop Cluster is the hardware part of the infrastructure.

- ***YARN:*** Apache Hadoop YARN (Yet Another Resource Negotiator) is a technology used for cluster management. YARN is considered as one of the key features in the second-generation Hadoop 2 version of the Apache Software Foundation's open source distributed processing framework. YARN is now considered as a large-scale, distributed operating system for big data applications. It is originally labelled by Apache as a redesigned resource manager. The YARN Infrastructure is responsible for providing the computational resources for example, CPUs, memory, etc, needed for application executions. YARN has two important elements [9] [10].

First is the Resource Manager, shown by Fig. 3. It is the master and one per cluster. It basically knows where the slaves are located which is called Rack Awareness and also knows how many resources they have. How to assign the resources is the most important task of Resource Manager.

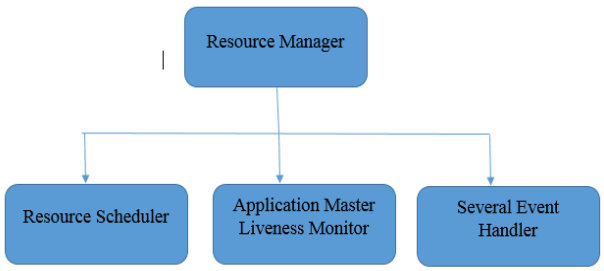

Figure 3: Resource Manager

Second is Node Manager, shown by Fig. 4. It is the slave of the infrastructure. It is many per cluster. When Node Manager starts, it publicizes himself to the Resource Manager. Periodically, a heartbeat is sent to the Resource Manager by node manager. Some resources are offered to the cluster by each Node Manager. The resource capacity is measured as the amount of memory and the number of v-cores. How to use this capacity is decided by Resource Scheduler at run-time. There is a Container which is a fraction of the Node Manager capacity and it is used by the client for running a program [11] [12].

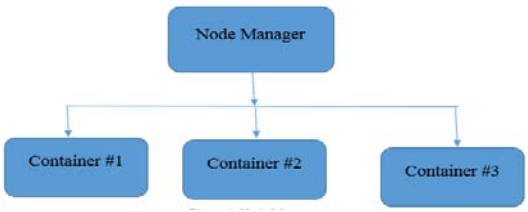

Figure 4: Node Manager

- ***HDFS:***The Hadoop Distributed File System (HDFS) is based on distributed file system which is designed to run on commodity hardware. HDFS has many similarities with existing distributed file systems but there are some differences also from other distributed file systems which are significant. HDFS is designed to be deployed on low-cost hardware and is highly fault-tolerant. HDFS is highly suitable for those applications which have large data sets and delivers high throughput access to application data. At first, HDFS was implemented as infrastructure for the Apache Nutch web search engine project. Now HDFS is an Apache Hadoop subproject. HDFS stores huge data and for storing such huge data, the files are stored across multiple machines. These files are stored in redundant manner so that it can prevent the system from possible data losses in case of failure. HDFS provides parallel processing also.

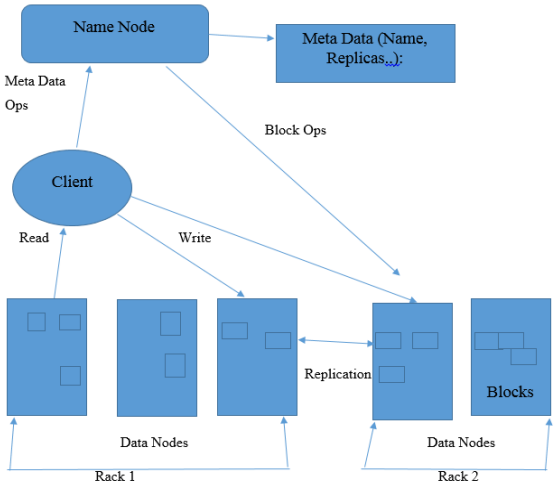

Figure 5: HDFS Architecture

HDFS contains a master/slave architecture. Fig. 5 shows its architecture. This architecture consists of a master server and a single Name-Node that handles the file system namespace and regulates access to files by clients. In this architecture, there is one Data Node in each cluster.This manages storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. In HDFS internally, a file is split into one or more blocks. These blocks are stored in a set of Data Nodes. The Name-Node performs file system namespace operations like closing, opening and renaming directories and files. It also controls the mapping of blocks to Data Nodes. The Data Nodes are responsible for serving write and read requests from the file system's clients. The Data Nodes also perform block deletion, creation and replication upon instruction from the Name Node [13].

- ***Storage:*** Except of the HDFS storage, there are other alternative also for storage solutions. These storage services are provided by different companies. For example, Amazon provides the Simple Storage Service (S3). This Simple Storage Service is storage for the Internet which is designed to make web-scale computing easier for developers.

Amazon S3 provides simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It offers any developer access to the same highly reliable, scalable, inexpensive and fast data storage infrastructure that Amazon uses to run its own global network of web sites.
As the Amazon provides storage services, same is provided by different companies also, like Microsoft Azure, IBM etc.

- ***MapReduce Framework:*** MapReduce is a program model for distributed computing based on java. It is a processing technique. The MapReduce algorithm includes two important tasks, namely Map and Reduce.

Map takes a dataset and then converts it into another dataset, where individual elements are broken down into two tuples (key/value pairs). Secondly, in reduce task, the output of Map is inputted in Reduce and then combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce infers, the reduce task is always done after the map job.

In MapReduce, it is easy to scale data processing over multiple computing nodes. Under this model, the data processing primitives are called mappers and reducers. Writing the application in the MapReduce form is not so easy but once it is written it allows scaling the application to run over thousands or even tens of thousands of machines in a cluster. This simple scalability has attracted many programmers to use the MapReduce model [14].

The MapReduce Algorithm:

- Generally, MapReduce program runs in three stages, as shown in Fig. 6, namely map stage, shuffle stage, and reduce stage.

Map stage: The input data is stored in the form of files in the Hadoop file system (HDFS). This input file is then passed to the mapper function line by line. These data is then processed by the mapper and several small chunks of data is created.

Reduce stage: Reduce stage is the combination of the Shuffle stage and the Reduce stage. The Reducer takes the output of mapper and the process it and generates new set of output which will be stored in HDFS.

- Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster when MapReduce job is executing.
- The framework manages all the details of data-passing such as verifying task completion, issuing tasks and copying data around the cluster between the nodes.
- The cluster accumulates and cuts the data to form an appropriate result, and sends it back to the Hadoop server when the task is completed [15].

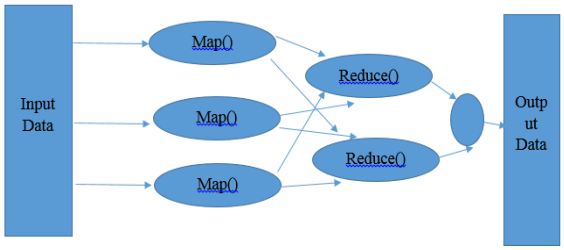

Figure 6: MapReduce Architecture

IV. APPLICATIONS

In the present time, due to large amount of heterogeneous data, there is great demand of Big Data. Big data is widely used in many applications. Some applications are as follows-

Big Data Contributions to Public Sector:Big Data is growing in the world by storm. Huge amount of data is generating from various sources with high speed. For handling these data, Big Data is used. Big Data provides better insights of unstructured and structured data. It provides smart decision by proper risk analysis. This is the reason why all the companies are inclined towards Big Data. Big Data provides a wide range of facilities to the government sectors including deceit recognition, fitness interconnected exploration, the power investigation, ecological fortification and economic promotion investigation.

FDA uses Big Data to examine the food based infections. Big data provides fast investigation of a huge volume of communal complaints big data analytics is used.

- ***Big Data Contribution to Healthcare:***The big data has wide use in the field of healthcare and medicine. With the raise of technology cost, the cost of health care is also increasing more and more. Big data is providing great helping hand in this field. It helps the physicians to keep the track of all patient's history which can be accessed by the only the patient or his particular physician. All the records related to patients are stored safely in database forever.

There are large number of medical devices which are big data oriented. Today data is used to such an extent that doctor recommends the medicines without even visiting the patient. They just notice the heartbeat and temperature through watch fitted on the patient's hand which monitor heart and temperature that stays in a remote place.

Big Data Contributions to Learning:Big data has great effect in the education world too. Now these days, almost every course of learning is present online. There are many more examples of the use of big data in the education industry. There is an application named as the Bubble Score which allows teachers to convey multiple-choice valuations through mobile devices and notch up paper tests through the cameras of the mobile phones. Along with this equipment, there are many more techniques which can be implemented using Big Data in education field.

Big Data Contributions to Industrial and Natural Resources:Big data provides solutions for natural resources also. The high volume as well as the velocity of big data is challenged by the high demand of the natural sources on this earth. Similarly, a great amount of data commencing the built-up industry is unexploited. The unused data avoids power competence, advanced eminence of merchandise, improved income boundaries and dependability. In the natural wealth industry, big data empowers for analytical modelling to sustain judgment creation that is used to incorporate and consume vast amounts of information from graphical information, geographical information, chronological and manuscript statistics. Big data has as well been worn to grow aggressive improvements in the middle of former settlements and in finding the solution to the development of confrontations.

Big Data Contributions to Banking Zones and Fraud Detection: In the banking sectors, Big Data is hugely used in the fraud detection. Big Data finds out all the mischief tasks done in the banking sectors. It detects the misuse of debit cards, misuse of credit cards, venture credit hazard treatment, archival of inspection tracks, customer statistics alteration, business clarity, IT action analytics, public analytics for business and IT strategy fulfillment analytics [16] [17] [18].

In banking sector, at present they are using natural speech processors and network analytics to grasp unlawful business activity in the economic marketplaces. Private and public actor banks, Retail traders, prevaricate funds and others in

the monetary marketplace make use of big data for business analytics used in reaction dimension, big businesses, prognostic Analytics etc. In businesses big data helps a lot in knowing CRM tactics of the competitors and the shopping patterns of customers so that they can put on them in their businesses in order to improve the sales.

V. CONCLUSION

In this paper concept of Big Data and various technologies hasbeen surveyed which are used handle the big data. This paper discussed an architecture of Big Data using Hadoop HDFS distributed data storage under which its different components are also explained. The main objective of this paper was to make a survey of various Big Data architecture, its handling techniques which handle a huge amount of data from different sources and improves overall performance of systems and its applications which shows its importance and uses in the present IT world.

VI. REFERENCES

[1] C. Lakshmi and V. V. Nagendra Kumar, "Survey paper on Big Data", 2016 International Journal of Advanced Research in Computer Science and Software Engineering.

[2] Ms. Vibhavari Chavan and Prof. Rajesh. N. Phursule, "Survey Paper on Big Data", 2014 International Journal of Computer Science and Information Technologies.

[3] Yuri Demchenko, "The Big Data Architecture Framework (BDAF)", Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.

[4] Amogh Pramod Kulkarni, Mahesh Khandewal, "Survey on Hadoop and Introduction to YARN", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[5] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, et al., "KNIME: The Konstanz Information Miner", in Data Analysis, Machine Learning and Applications (Studies in Classification, Data Analysis, and Knowledge Organization), Springer Berlin Heidelberg, pp. 319–326, 2008.

[6] Sagiroglu, S.Sinanc, D.,"Big Data: A Review",2013, 20-24.

[7] Ms. Vibhavari Chavan, Prof. Rajesh and N. Phursule, "Survey Paper On Big Data", International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[8] Kyuseok Shim, "MapReduce Algorithms for Big Data Analysis", DNIS 2013, LNCS 7813, pp. 44–48, 2013.

[9] S. G. Manikandan and S. Ravi, "Big Data Analysis Using Apache Hadoop," 2014 International Conference on IT Convergence and Security (ICITCS), Beijing, 2014, pp. 1-4.doi: 10.1109/ICITCS.2014.7021746

[10] Y. Demchenko, C. de Laat and P. Membrey, "Defining architecture components of the Big Data Ecosystem," 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, 2014, pp. 104-112. doi: 10.1109/CTS.2014.6867550

[11] Mata Gujri College Fatehgarh Sahib, "REVIEW PAPER ON BIG DATA USING HADOOP", 2015 International Journal of Computer Engineering & Technology.

[12] Poonam S. Patil1 and Rajesh. N. Phursule, "Survey Paper on Big Data Processing and Hadoop

[13] Components", 2012 International Journal of Science and Research.

[14] Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat,JaspreetKaur and Navjot Kaur, "Research Paper on Big Data and Hadoop", 2016 International Journal of Computer Science and Telecommunications.

[15] BijeshDhyaniand Anurag Barthwal, "Big Data Analytics using Hadoop", 2014International Journal of Computer Applications.

[16] Samiddha Mukherjee and Ravi Shaw, "Big Data – Concepts, Applications, Challenges and Future Scope" ,2016International Journal of Advanced Research in Computer and Communication Engineering.

[17] Hua Fang, Zhaoyang Zhang, ChanpaulJin Wang, Mahmoud Deshmand, Chonggang Wang, and HonggangWang, "A Survey of Big Data Research", 2015 IEEE Network.

[18] KuchipudiSravanthi and TatireddySubba Reddy, "Applications of Big Data in Various Fields", International Journal ofCmputer Science and Information Technology, 2015.