IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

Comprehensive Study on Big Data Analytics

[1]Sarala N R, [2]Gagana R P, [3]Manisha R, [4]Monisha P V, [5]Roja L

[1,2,3,4,5]Department of Computer Science & Engineering SJBT, Bangalore, India

DOI: https://doi.org/10.26438/ijcse/v7si15.265269 | Available online at: www.ijcseonline.org

Abstract- Big Data is termed has any type of datasets which are so vast and compound which becomes difficult to process them using traditional data processing applications. While handling vast dataset different challenges may be faced by the user. In recent times, the internet application and communication have observed a lot of growth and reputation in the field of Information Technology. These internet applications and communication are frequently generating the large size, different variety and with some authentic difficult multifaceted structure data called big data. As a result, we are now in the era of enormous automatic data collection. For example, E-commerce transactions include activities such as online buying, selling or investing. Thus they generate the data which are high in dimensional and complex in structure. The traditional data storage techniques are not adequate to store and analyses those huge volume of data. Many researchers are doing their research in dimensionality reduction of the big data for effective and better analytics report and data visualization. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, and Apache Hive. Hence, the aim of the survey paper is to provide the overview of the big data analytics, issues, challenges and various technologies related with Big Data.

***Keywords*:** Big Data, Big Data Analytics, Hadoop, Map Reduce.

I. INTRODUCTION

Big data is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis. Big data is defined as data which is not only very large, but also high in velocity and variety, which makes them difficult to handle using traditional tools and techniques [3]. These data is generated from different social media like Twitter, Facebook etc. When handling with huge datasets, Organizations come across some difficulties in being able to create, manipulate, and manage big data. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyse huge datasets. In recent days the field of Information Technology (IT) is improving more, this leads in enormous data generation, for every instance, approximately 72 hours of video files are uploaded to YouTube by the people. This data growth challenges the main problems of acquisition and integrating massive volume of data from widely distributed data sources such as social media applications [4].

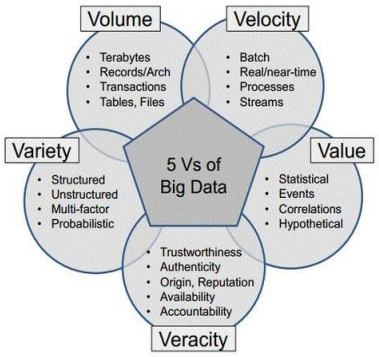

Fig. 1: 5v's of Big Data

Fig.1 illustrates a general big data network model with Map Reduce. The five V's (volume, variety, velocity, value, veracity) are the challenges of big data management are:

1. Volume:
Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. The data becomes into large sized files which causes storage issue. This issue is resolved by deducting storage cost.

2. Velocity:
Data sources are extremely varied. The files comes in many formats and type, it may be structured or unstructured such as text, audio, videos, log files and more. Data streams come in unparalleled speed and should be allocated with in

an suitable manner. Different kind of RFID(Radio Frequency Identification)tags, IoT(Internet of Things) sensors and smart metering are driving the necessity to deal with data flows in real time.

3. Variety:

Data originates in different categories of formats such as structured, numeric data in traditional databases to unstructured text documents: email, video, audio, stock and financial transactions. But these 3V's are extended as 5V's. In addition, two more V's such as variability and veracity. They are as follows:

4. Variability:

The growing velocities and varieties of data, data flows can be highly unreliable with periodic peaks. When we dealing with high volume, velocity andvariety of data, the all of data are not going 100% accurate, there will be dirty data. Big data works on these kind of data.

5. Value:

It is a most significant V in big data. Value is main buzz for big data because it is important for businesses, it is widely used in IT setup to store enormous amount of values in database.

II. LITERATURE SURVEY

The author [1] quoted that Hadoop MapReduce is used for storing massive data and processing the data. As it is a large scale and also an open source software framework that is dedicated to data-intensive computing, scalable and distribution. Initially this frame work splits up the huge data into smaller segments known as chunks which can handle scheduling in parallel. This maps every piece of chunks into an intermediate values by using map function in the map phase and by reducing this intermediate values into a solution using reduce function done in reduce phase.

The author [2] emphasis on the importance to handle Big Data such as Hadoop, MapReduce and Hadoop Distributed File System (HDFS) by using some of the methodologies. This author also quoted that, Hadoop can used different schedulers as well as various technical aspects of Hadoop. This author also elaborates on importance of YARN which overcomes the limitations of MapReduce.

The author [3] has been focused on various procedures in order to handle the Big Data and its architecture. In this paper , the author has discussed the drawbacks of Big Data such as Volume, Variety, Velocity, Value, Veracity and also several benefits of these technologies and also the author has discussed on the architecture of Big Data using Hadoop distributed data storage HDFS, MapReduce distributed data processing over bunch of produces servers and real-time NoSQL databases.

The author [4] stated about Big Data definition and expanded the definition by providing the 5V Big Data properties: Volume, Variety, Velocity, Value, and Veracity. This author also measured other dimensions for Big Data analytics and taxonomy, in specific relating and contrasting the Big Data tools in social media, industry, business, e-Science, healthcare and so on. At present tradition, working with consistently enlarging the volume of data, the scientific analysis methods can be provided by modern e-Science to industry, though industry can take advanced quick evolving Big Data machineries and tools to broader a public and science.

The author [5] detailed that data is developing enormously day by day. Currently, these data are not only restricted within Gigabyte, instead it is more than this measure that is Terabyte, Petabyte, Exabyte and so on. The data that is generated are not only too gigantic quantity in size and also heterogeneity and ability to gather data from dizzing array of sources. Big data analysis tools like HDFS, MapReduce over Hadoop assurances to support organizations better recognizing their market place and customers, which hopefully leads to computation welfares and for better business decisions. The main aim of our paper is to focus on numerous Big Data handling techniques that can handle a gaint amount of data from different sources by combination of advanced methods, tools, functions, techniques in order to progress overall performance of the system.

III. BIG DATA TECHNOLOGIES AND METHOD

Due to the growth of the technologies, huge data flows in and out of organization on the daily basis, this has led to more efficient and quick way of evaluating this large amount of data. Having heaps of data on hand is no longer sufficient to make proper decision at exact time. Big data is a new method for handling enormous data through diverse tools. Some of the technologies used in big data are:

A) HADOOP

Development started on the Apache Nutch project, but was moved to the new Hadoop sub project in January 2006. Dong cutting found hadoop along with Mike Cafarella in 2005. Dong was working at Yahoo! At the interval, named it after his son's toy elephant. Apache Hadoop is an open source software built on two technologies LINUX operating system and java programming language. Java is used for storing, analysis and processing large data sets**.** The benefits of Hadoop are distributed storage and computational capabilities, extremely scalable, optimized for high throughput, large block sizes, tolerant of software and hardware. Hadoop enables reliable, scalable, distributed computing on clusters.

Distributed: Handle replication, Offers massively parallel programming model, Map Reduce.
Scalable: Designed for massive scale of processors, memory and local attached storage.
Reliable: The software is fault tolerant, it expects and handles hardware and software failures. Hadoop is mostly useful when:
a) Complex information processing is needed:

- Converting unstructured data into structured form and vice versa.
- Complex but parallelizable algorithms needed, such as genome sequencing or geo-spatial analysis.

b) Machine learning:

- Data sets are too huge to fit into record discs, RAM or require too many Hubs (10's of TB up to PB) .
- Fault tolerance is critical.
- Job scheduling is handled.

B)MapReduce

MapReduce is a popular framework for bigdata processing because of its simple programming model and automatic parallel execution. It was introduced by Google in order to process and store vast datasets on product hardware. It is a programming environment that licenses large jobs enactment scalability against group of server. MapReduce is a popular software framework for distributed processing of large datasets on computer clusters. MapReduce is based on master-slave architecture where a number of slave node are handled by one master node

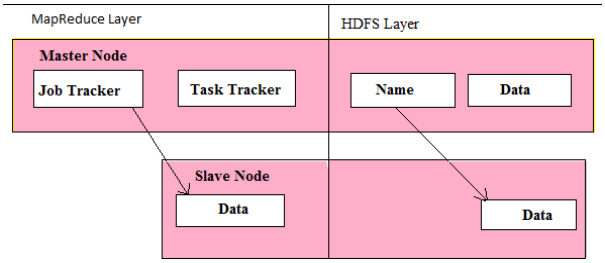

Fig 2: Master slave architecture.

Master node comprises -

- Name node (HFDS layer)
- Data node (HFDS layer)
- Task tracker node (MapReduce layer)
- Job tracker node (MapReduce

layer) Multiple slave nodes comprise -

- Single JobTracker per master is responsible for scheduling.
- MapReduce layer has job and task tracker nodes
- HFDS layer has name and data nodes
- Data node (HFDS layer)
- Task tracker node (MapReduce layer)

Single Job Tracker per master is answerable for scheduling the jobs' portion errands on the slaves. It monitors slave progress. It also re-executing unsuccessful tasks As well as single Task Tracker per slave performs the tasks as engaged by the master. Map reduce core functionality is based on the Map phase and reduce phase. Code usually written in Java though it can be transcribed in other languages with the Hadoop Streaming Aplication programming interface(API).

MapReduce Algorithm:
In MapReduce algorithm three stages are used for the program to run namely map stage, shuffle stage, and Reduce stage. Map stage: Generally, map tasks are launched in parallel to convert the original input splits into intermediate data in the form tuple i.e, key/value pairs. These key/value pairs are stored on local machine and organized into multiple data partition, one per reduce task

- Shuffle stage: this step is between map and reduce phase. In this stage , data produced by map reduce are ordered, partitioned and transferred to appropriate machines executing the reduce phase.
- Reduce stage: each reduce task fetches its own share of data partitions from all map tasks generate the final result.

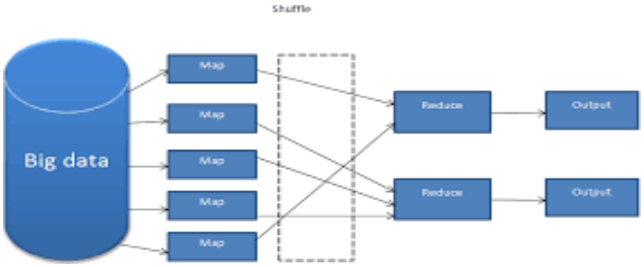

Fig 3: MapReduce architecture.

IV. LAYERED VIEW ARCHITECTURE OF BIG DATA SYSTEM

The layered structure, of the Big data system is as shown in the following figure 4. The layered structure is divided into three layers i.e, the application layer, computing layer and infrastructure layer from top to bottom. Conceptual hierarchy is provided by the layered view to reduce the complexity of the system. The operation performed by the layers is as follows

Application layer: It acts as the interface given by the programming models to implement various data analysis function, that includes classification, statistical analysis, clustering and querying. The basic analytical method is combined to develop various related application. Some of the application domains are retail, global manufacture, public sector administration, healthcare etc.

Computational layer: It is the middle layer that runs over raw ICT resources which encapsulates various data tools. Data integration, the programming model and data management are the tools included in this layer.

Infrastructure layer: it contains pool of Information and communication technology(ICT) resources. The cloud computing infrastructure organizes this resources and virtualization technology is enabled. These resources are efficiently used in the upper-layer in ne-grained way with a specific service level agreement (SLA). In this model, we must allocate the resources to meet the demands of big

.

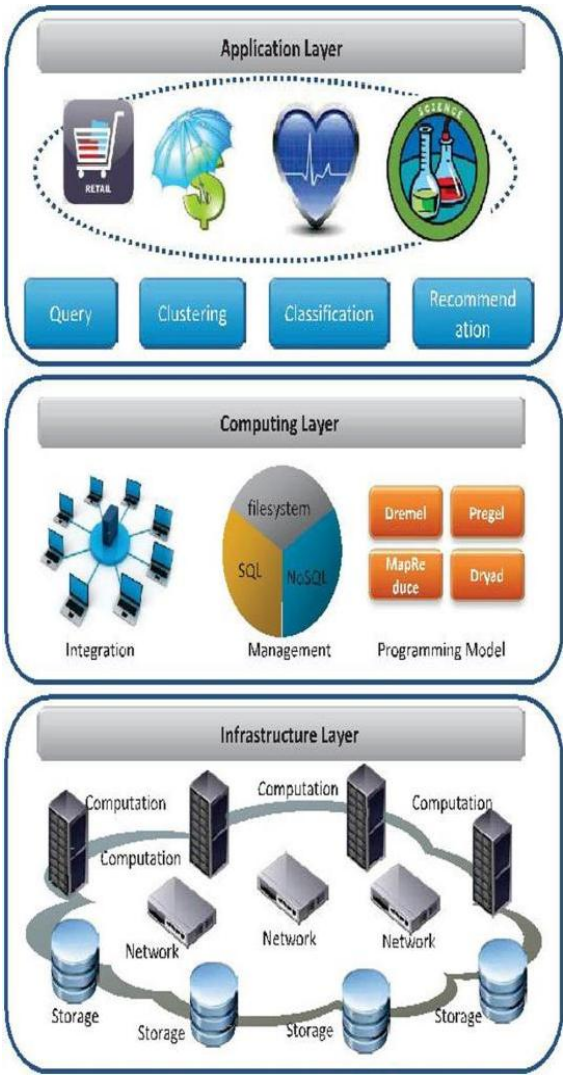

Fig 4: layered view architecture of Big Data

Data while resources can be efficiency achieved by maximizing system utilization, operational simplification, energy awareness etc.

APPLICATIONS

Fig 5: Applications of big data

1. Big data contribution to healthcare:
The sources of big data in healthcare are: Hospitals, Health insurance companies, Research institutions, etc. This improves the quality of healthcare by minimizing the costs, providing customized patient treatment, early prediction of diseases.

2. Big data contribution to public sector
It provides better view of structured data and unstructured data. Most companies are using Big data because it gives a smart decision by proper risk analysis. It also provides facilities to the government sectors.

3. Big data contribution to learning
Use of Big Data in the education field has increased. There are various course of learning online. Bubble score application allows the teacher to convey multiple choice valuation through mobile devices.

4. Big Data contribution to Banking Zone and fraud detection:
Big data is widely used in fraud detection by finding out the harmful task done in bank sectors. It detect the misuse of credit cards, debit cards and so on.

CHALLENGES

1. Data representation
Big Data is as large datasets are organization structure, accessibility, heterogeneity type etc. There are many software and hardware solutions available in the technological landscape that enable capturing, storing and subsequently analysis of Big Data.

2. Redundancy
Traditionally they are large number of redundant data in a row datasets. The big data removes the duplicate copies and reduces the memory space.

3. Data life cycle management
Data life cycle management process decides which data shall be stored and which data shall be discarded during the analytical process. There are challenges, one of which is

that the existing storage system could not support such huge amount of data. Therefore, a principle which makes the life cycle management system effective is needed.

4. Data confidentiality:
The service providers and owners of the data could not maintain and analyze such huge datasets effectively. They depend on professionals or third-party tools to analyze such data, which improve the potential safety risks. Hence, data confidentiality is important issue for the researchers.

5. Data privacy and security:
This is mainly concerned on analyzing and accessing personal information is rising. It provides privacy and security by allowing only the authorized person to access.

6. Connecting to social media:
Social media gives us a distinct parameter, such as statistical redundancy, availability of redundancy and vastness. Applications can achieve high levels of precision and distinct points of view by connecting interned data.

V. CONCLUSION

The main objective of this paper is to make a survey of numerous big data architecture, its handling techniques which handle a huge amount of data from different sources and improves overall performance of systems and also its applications which shows its importance and uses in the present IT world. It includes architecture using Hadoop HDFS, MapReduce distributed data processing over a cluster of commodity servers.

REFERENCES

[1] Yuri Demchenko, "The Big Data Architecture Framework (BDAF)", Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.
[2] 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).
[3] 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.
[4] Sagiroglu, S.Sinanc, D.,"Big Data: A Review",2013, 2024.
[5] 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.

