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. 6, Issue. 7, July 2017, pg.64 – 68

BIG DATA & HADOOP: A Survey

Rehana Hassan[1], Rifat Manzoor[2], Mir Shahnawaz Ahmad[3]

[1, 2]Student, Department of Computer Science Engineering, SSM College of Engg. & Tech., Kashmir, India

[3]Assistant Professor, Department of Computer Science Engineering, SSM College of Engg. & Tech., Kashmir, India

[1] sofi.rehana28@gmail.com, [2] sabajan509@gmail.com, [3] mirshahnawaz888@gmail.com

Abstract— In today's modern world, the availability of large number of online products like web-sites, web-portals, shopping sites, social networking sites, etc., give rise to a collection of extremely large and complex data sets known as Big Data. The analysis and processing of such an enormous volume of data in an efficient manner is a prime concern. Also, due the complex nature of Big Data it becomes very difficult to process using on-hand databases management tools. So, a specialized tool is required for processing Big Data, and one such tool is Hadoop. This paper presents a detailed overview of Big Data and finally describes an easy-to-use and highly scalable software framework, known as Map-Reduce framework, which is used to process Big Data.

Keywords— Big Data, Hadoop, Map-Reduce.

I. INTRODUCTION

Big data describes the large volume of data, it is a combination of huge datasets that can be handled using new techniques. Big data is not only containing data, it also contains various tools, techniques and frameworks. Data that has extra-large Volume, comes from Variety of sources like text, audio, video, xml files etc, Variety of formats and comes at us with a great Velocity is normally referred to as Big Data. Big data can be structured, unstructured or semi-structured. Big data that hold the data generated by various equipment and applications like Black box.

The term BIG DATA is simply used to describe the collection of complex and huge data sets such that it is difficult to analyse, store and process this kind of data using conventional database management tools and traditional databases management systems [1]. Earlier RDBMS tried to handle the unstructured and semi-structured large chunk of data but couldn't handle the same because the data is very huge. So hadoop came into existence. Using hadoop we can handle big data [2]. In this paper we present Hadoop and its core concepts which are HDFS and MapReduce. Hadoop is one of the few frameworks that support storing of big unstructured data like video files, xml files, audio files, image files etc. along with storing the normal structured data. The Hadoop is a distributed file system which is designed to operate on commodity hardware. A reliable and robust framework in Hadoop known as Map-Reduce is used to create applications that are capable of processing large amount of structured as well as unstructured data in parallel over a large group of machines. The detailed view of this framework is given in the upcoming sections of this paper.

II. BIG DATA & ITS PARAMETERS

As the data is bigger from different sources in different form, it is represented by the 4 V's [2]:

- *Volume:* Ratio of data or huge amount of data develops in every second. Machine develop data are examples for these components. Nowadays data volume is increasing very quickly from gigabytes to peta-bytes.
- *Velocity:* Velocity is the speed at which data is developing and processed. For example facebook, google, twitter etc
- *Variety:* Variety is important characteristic of big data. It is a type of data. Data can be in different styles such as Text, numerical, xml files, application programs, images, audio, video data. On facebook more than 2 billion people are sharing files, photos, txt, videos, audio,
- *Veracity:* Veracity means accuracy or anxiety of data. Data is uncertain due to the inconsistency and in completeness.

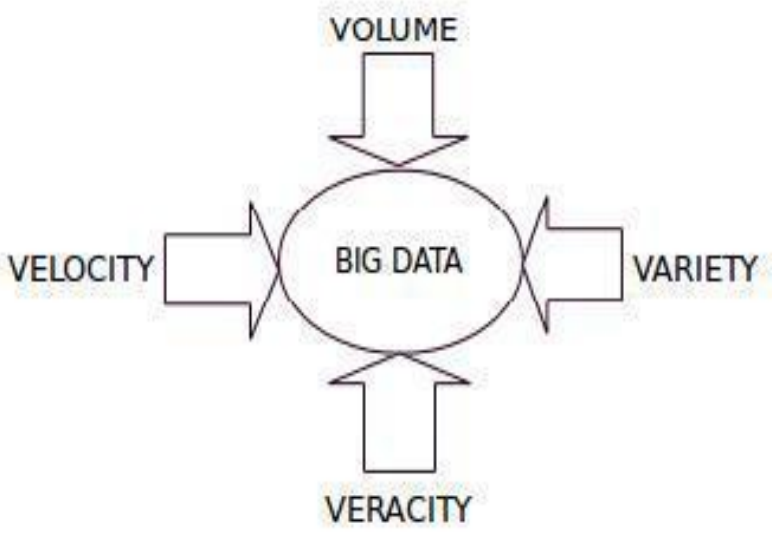

Fig. 1: Four – V's of BIG DATA

A. *The Challenges with Big Data*

1. *Heterogeneity and Incompleteness:* If we want to evaluate the data, it should be structured data but when we deal with the Big Data, then data may be structured or unstructured as well. Heterogeneity is the big challenge in data analysts and Analysis need to cope with it [3]. Suppose an example of patient in Hospital. We will create each record for each medical test. And we will also generate a record for hospital stay. This will be not equal for all patients. This design is not well structured. So managing with the Heterogeneous and incomplete is required. A good data analysis can be applied to this.
2. *Privacy:* Privacy of data is another huge concern that increases in the context of big data. In some countries there are strict laws about the data privacy, for example in USA there are strict law for health records, but for others it is less forceful. For example in social media we cannot get the private posts of users for sentiment analysis.
3. *Scale:* Big Data is having very huge size of data sets. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In past years, this problem has been solved by the processors getting faster but now data quantity is becoming large and processors are static. World is moving towards the Cloud technology, due to this shift data is generated in a very high rate [3]. This high rate of increasing data is becoming a challenging problem to the data analysts. Hard disks are used for storage of data. They are slower I/O performance. But now Hard Disks are replaced by the solid state drives and other technologies.
4. *Human Collaborations:* In spite of the advanced computational models, there are many patterns that a computer cannot recognize. A big data analysis system must support input from multiple human experts. Wikipedia is the perfect example. We are reliable on the data given by the strangers, however most of the time they are correct. But there can be other people with other motives as well as like providing false data. We need technological model to handle with this. As humans, we can look the review of book and find that few are positive and few are negative and come up with a decision to whether buy or not.

*B. **Opportunities to Big Data***

1. *Media:* Media is using big data for the boost and sale of products by focus the interest of the user on internet. For example social media posts, data analysts get the number of posts and then evaluate the interest of user. It can also be done by taking the positive or negative reviews on the social media.

2. *Government:* Big data can be used to handle the issues faced by the government. Obama government declared big data research and development initiative in 2012 [4]. Big data analysis played an important role of BJP winning the elections in 2014 and Indian government is implement big data analysis in Indian electorate.

3. *Technology:* Almost each top organization like Facebook and yahoo has adopted Big Data and are spending on big data. Facebook holds 50 Billion photos of users. Every month Google holds 100 billion searches. From this one can say that there are a lot of opportunities on internet and social media.

4. *Science and Research:* Big data is an up-to-the-minute topic of research. Large number of Researchers is working on big data [4]. There are so many papers being published on big data.

5. *Healthcare:* According to IBM Big data for Healthcare, 80% of medical data is unstructured. Healthcare organizations are adapting big data technology to grab the complete data about a patient. To boost the healthcare and low down the cost big data analysis are needed and certain technology should be adapted.

III. TECHNIQUES AND TECHNOLOGIES FOR BIG DATA

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data. There are many solutions to handle the Big Data, but the Hadoop [5] is one of the most widely used technologies.

A. ***Hadoop***

Hadoop is a programming framework user to support the processing large data sets in a disturbed computing environment. It is an open source project hosted by Apache Software Foundation. It consists of various small sub projects which belong to the category of infrastructure for distributed computing. Hadoop was the name of a yellow toy elephant owned by the son of one of its inventors. Hadoop mainly consists of:

1. File System (The Hadoop File System) .
2. Programming Paradigm (Map Reduce) [6].

There exist many problems in dealing with storage of huge amount of data. The storage capacities of the drives have expanded massively but the rate of reading data from them has not shown that considerable progress. The reading process takes huge amount of time and the process of writing is also slower. The time can be decreased by reading from multiple disks at once. Only using one hundredth of a disk may seem wasteful. But if there are one hundred datasets, each of which is one terabyte and providing shared access to them is also a solution. There exist many problems also with using several pieces of hardware as it increment the chances of failure. This can be deflecting by Replication i.e. creating redundant copies of the similar data at distinct devices so that in case of failure the copy of the data is possible. The main problem is of combinative the data being read from different machines. There are so many methods are able in distributed computing to handle this problem but still it is quite challenging. All the complication discussed is easily managed by Hadoop. The problem of failure is directed by the Hadoop Distributed File System and problem of combining data is directed by Map reduce programming Paradigm. Map Reduce basically diminish the complication of disk reads and writes by giving a programming model dealing in computation with keys and values.

B. ***Hadoop Distributed File System***

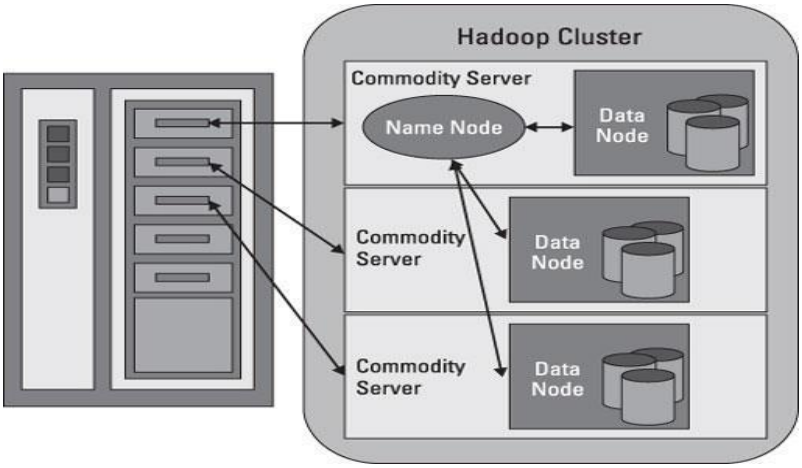

Fig. 2: HDFS Architecture

HDFS stands for Hadoop Distributed File System. The Hadoop Distributed File System is a versatile, clustered way to handling files in a big data environment. It includes a fault-tolerant system HDFS is able to store huge amounts of information. HDFS, it is a kind of data service that offers a different set of capabilities required when data volumes and velocity are high.

HDFS manages storage on the cluster by breaking incoming files into pieces called blocks. The blocks are stored on data nodes to notice what blocks on which data nodes make up the complete file.

The Name Node also performs as a "traffic cop," handling all access to the files. The entire collection of all the files in the cluster is sometimes referred to as the file system namespace. Even though a strong relationship occurs between the Name Node and the data nodes, they run in a "loosely coupled" fashion. This allows the cluster elements to act dynamically. The data nodes communicate among themselves so that they can cooperate during normal file system operations. This is mandatory because blocks for one file are likely to be stored on multiple data nodes.

C. ***Map Reduce***

Map-The Function takes key or value pairs as input and generates an intermediate set of key or value paris. Reduce- The Function which merges all the intermediate values associated with the same intermediate key [7]. Map-Reduce is an implementation of the algorithm developed and managed by the Apache Hadoop project. Hadoop Map Reduce consists of many stages, each with a meaningful set of operations helping to get the answers you need from big data. The development starts with a user request to run a Map Reduce program and go on until the results are written back to the HDFS.

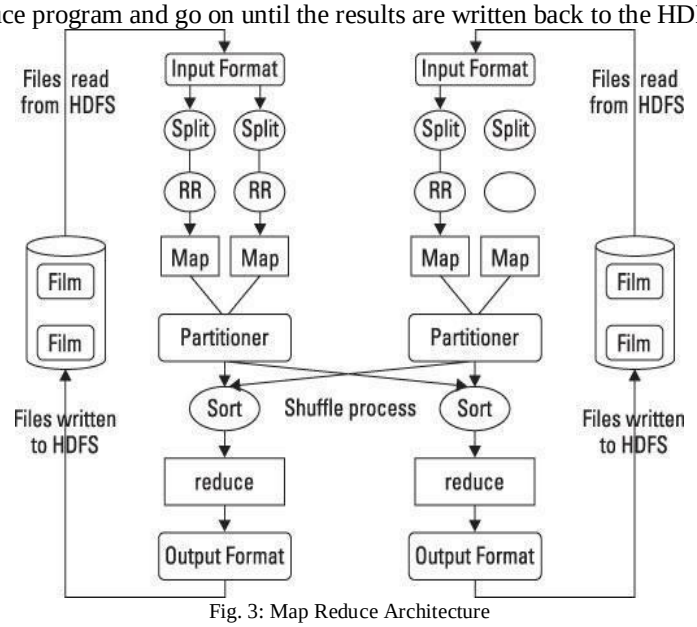

Fig. 3: Map Reduce Architecture

a. Get the Big Data Ready

When a client requests a Map Reduce program to run, the first step is to discover and study the input file. The file pattern is totally arbitrary, but the data must be changed to something the program can process [8]. This is the activity of Input Format and Record Reader. Input Format chooses how the file is going to be broken into smaller pieces for processing using a function called Input Split. It then assigns a Record Reader to transform the raw data for processing by the map. Different types of Record Readers are supplied with Hadoop, offering a wide variety of conversion options.

b. Let the Big Data Map Begin

Your data is now in a mode tolerable to map. For all input pair, a specific instance of map is called to process the data for each input pair. Map and reduce need to perform together to process your data, the program needs to gather the output from the separate map and pass it to the reducers. This task is done by an Output Collector. A Reporter function also present information collected from map tasks [9]. This entire task is being done on multiple nodes in the Hadoop cluster simultaneously. After all the map tasks are complete, the common results are collected in the partition and a shuffling occurs, sorting the output for excellent processing by reduce.

c. Reduce and combine for big data

For each output pair, reduce is called to do its task. In similar pattern to map, reduce collect its output while all the functions are processing. Reduce can't begin as far as all the mapping is done. The output of reduce is also a key and a value. Hadoop provides an Output Format feature, and it performs very much like Input Format. Output Format takes the key-value pair and formulates the output for writing to HDFS. The last task is to actually write the data to HDFS. This is done by Record Writer, and it performs similarly to Record Reader except in reverse. It takes the Output Format data and writes it to HDFS [10].

IV. CONCLUSION

The paper describes the concept of Big Data along with 4 V's, Volume, Velocity, variety veracity of big data. The paper also focuses on Big Data processing problems. These challenges must be addressed for good and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, , and visualization at all stages of the analysis pipeline from data acquisition to result interpretation.

The paper describes Hadoop which is an open source software used for processing of Big Da. That's where Hadoop has become one of the enabling data processing technologies for big data analytics. Hadoop allows companies to store and manage far larger volumes of structured and unstructured data than can be managed by relational database management systems

REFERENCES

[1] S.Vikram phaneedra & E.Madhusudhan Reddy, "Big data-solutions for RDBMS- A survey" in 12th IEEE/IFIP Network operations & management symposium (NOMS 2010) (Osaka,Japan,Apr 19 2013).
[2] Kiran kumara Reddi & Dnvsl Indra "Different technique to transfer Big Data; survey" IEEE Transations on 52(8) (Aug.2013).
[3] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.
[4] Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012.
[5] Jimmy Lin "Map Reduce Is Good Enough?" The control project. IEEE Computer 32 (2013).
[6] Kenn Slagter · Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013
[7] Ahmed Eldawy, Mohamed F. Mokbel "A Demonstration of Spatial Hadoop: An Efficient MapReduce Framework for Spatial Data", Proceedings of the VLDB Endowment, Vol. 6, No. 12 Copyright 2013 VLDB Endowment 21508097/13/10.
[8] Jeffrey Dean and sanjay Ghemawat "Map Reduce; Simplified Data Processing on Large Clusters" OSDI 2010.
[9] http://www.tutorialspoint.com/Hadoop
[10] http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.html

