International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 1, pp 13-16, January-2020

An Analytical Study on Big Data and Hadoop

[1]Mr. Sagar Sureshrao Kalbande, [2]Mr. Shubham Anil Pardakhe, [3]Prof. Anurag Dwivedi

[1,2,3]MCA-III, Department of Research and PG Studies in Science & Management, Vidyabharati Mahavidyalaya, Amravati, India

***Abstract* - The term "big data" describes innovative techniques and technologies for capturing, storing, distributing, managing and analyzing quantities of data in peta bytes or large format at high speed and with different structures. Big data can be structured, unstructured or semi-structured, which leads to the inability of conventional methods of organizing data. Big data is data whose scope, variety and complexity require new architectures, techniques, algorithms and analyzes in order to process it and to extract hidden values and knowledge. Hadoop is the basic platform for structuring big data and solves the problem of making it useful for analytical purposes.**

Keywords: Big Data, Hadoop, Map Reduce, HDFS, Hadoop Components.

I. INTRODUCTION

1.1 Big Data: Definition

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability) and growth rate (speed) make it difficult to capture, manage, processing or using conventional technologies and tools to analyze B. relational databases and office statistics or visualization packages, in the time required for this. Although the size used to determine whether a particular record is considered big data is undefined and continues to change over time, most analysts and practitioners currently refer to records of 30 to 50 terabytes. (10 to 12 or 1000 gigabytes per terabyte) to several petabytes (1015 or 1000 terabytes per petabyte) as big data. Figure 1 shows the layered architecture of the Big Data system. It can be divided into three levels from top to bottom, including infrastructure, computer and application levels.

1.2 3 Vs of Big Data

Volume of data: The amount of data refers to the amount of data. The amount of data stored in corporate repositories has gone from megabytes and gigabytes to petabytes.

Variety of data: Different types of data and data sources. The variety of data ranges from structured and old data stored in corporate repositories to unstructured, semi-structured, audio, video, XML, etc.

Speed of data: Speed refers to the speed of data processing. Urgent processes such as contagious fraud must use big data during transfer to your business to maximize their value.

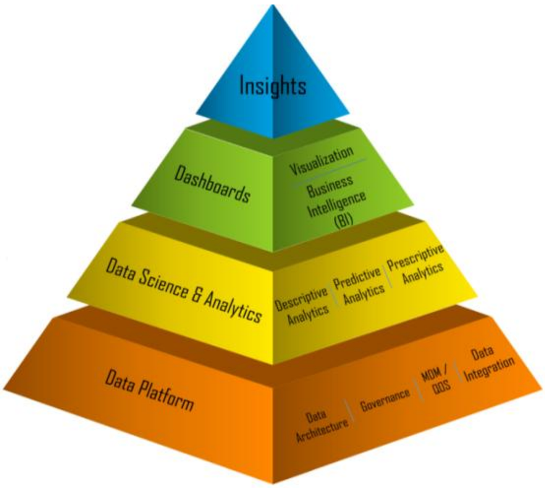

Figure 1: Layered Architecture of Big Data System

1.3 Problem with Big Data Processing

1.3.1 Heterogeneity and Incompleteness

When people consume information, much of the heterogeneity is pleasantly tolerated. Indeed, the nuances and the richness of the natural language can bring a precious depth. However, machine analysis algorithms expect consistent data and cannot understand the nuances. Therefore, the data must be carefully structured during the first step (or before) of data analysis. Computer systems operate more efficiently when they can store multiple items, all of the same size and structure. Efficient display, access and analysis of semi-structured data requires additional work.

1.3.2 Scale

Of course, the first thing you think about in big data is size. Finally, the word "fat" is in the name. Managing large amounts of rapidly growing data has been a challenge for many decades. In the past, this challenge has been mitigated by the fact that Moore processors have become faster and faster to provide us with the resources necessary to handle increasing amounts of data. However, a fundamental change is currently underway: the volume of data is changing faster than computing resources and the processor speeds are static.

1.3.3 Timeliness

The downside to size is speed. The larger the data set to be processed, the longer the analysis takes. Designing a system that effectively manages size will likely result in a system that can process a certain amount of data faster. However, it is not only this speed that we hear when we talk about speed in connection with big data. Rather, there is a challenge in terms of employment rate

1.3.4 Privacy

Data privacy is another major concern that grows with big data. Electronic patient records have strict laws that govern what can and cannot be done. Other data have less stringent regulations, especially in the United States. However, the general public is very afraid of the inappropriate use of personal data, in particular by combining data from multiple sources. Privacy management is both a technical and sociological target problem that must be tackled from both angles in order to realize the promise of big data.

1.3.5 Human Collaboration

Despite the huge advances in computer analysis, there are still many models that humans can easily recognize, but computer algorithms are struggling to find them. Ideally, big data analysis is not fully computerized, but has been specially designed to keep someone up to date. The new field of visual analysis attempts to do this at least with regard to the modeling and analysis phase in the pipeline. In today's complex world, several experts from different fields are often needed to really understand what is going on. A Big Data analysis system must support the contribution of several human experts and the joint exploration of the results. These multiple experts can be spatially and temporally separated if it is too costly to assemble an entire team in one room. The data system must accept this contribution from distributed experts and support its cooperation.

II. HADOOP: SOLUTION FOR BIG DATA PROCESSING

Hadoop is a programming framework that supports the processing of large amounts of data in a distributed computing environment. Hadoop was developed by Google Map Reduce. It is a software framework in which the applications are divided into different parts. The current Apache Hadoop ecosystem includes the Hadoop kernel, Map Reduce, HDFS and a number of different components such as Apache Hive, Base and Zookeeper. HDFS and Map Reduce are explained in the following points.

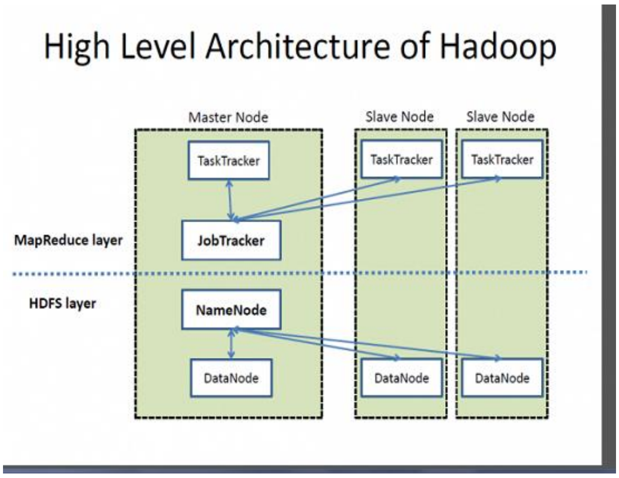

Figure 2: Hadoop Architecture

2.1 HDFS Architecture

Hadoop includes an error-free storage system called Hadoop Dispersed File System (HDFS). HDFS is capable of storing large amounts of information, gradually evolving it and surviving the failure of essential parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work between them. Clusters can be created with inexpensive computers. If an error occurs, Hadoop continues to run the cluster without losing data or interrupting the job by moving the job to the other computers in the cluster. HDFS manages storage in the cluster by dividing incoming files into blocks called "blocks" and storing each block redundantly in the server pool. HDFS typically stores three full copies of each file by copying each part to three different servers.

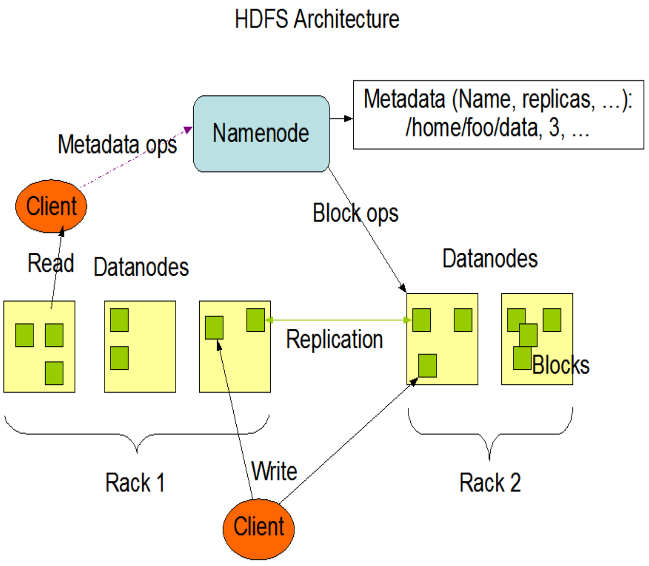

Figure 3: HDFS Architecture

ISSN (online): 2581-3048
Volume 4, Issue 1, pp 13-16, January-2020

2.2 Map Reduce Architecture

The processing pillar of the Hadoop ecosystem is the Map Reduce Framework. With the framework, the specification of an operation can be applied to a large data set, the problem and the data can be divided and executed in parallel. From an analyst's point of view, this can be done in several dimensions. For example, a very large amount of data can be reduced to a smaller subset in which analyzes can be applied. In a traditional data warehousing scenario, this may mean that an ETL operation is applied to the data to create something that can be used by the analyst. In Hadoop, these processes are written in Java because the map reduces work. There are a number of superior languages such as Hive and Pig that make writing these programs easier. The output of these jobs can be rewritten on HDFS or stored in a conventional data warehouse. There are two functions in Map Reduce as follows:

Map - The function takes key/value pairs as input and generates an intermediate set of key/value pairs.

Reduce - The function which merges all the intermediate values associated with the same intermediate key.

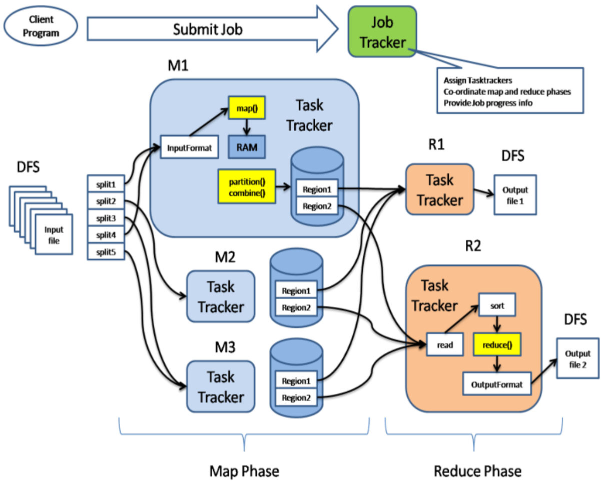

Figure 4: Map Reduce Architecture

2.3 Improving Science and Research

Science and research are being transformed by the new possibilities of big data. Take, for example, CERN, the Swiss laboratory for nuclear physics with the Large Hadrons Collider, the world's largest and most powerful particle accelerator. Experiments to decipher the secrets of our universe - how it started and how it works - generate huge amounts of data. The CERN data center has 65,000 processors to analyze its 30 petabytes of data. However, it uses the computing power of thousands of computers in 150 data centers around the world to analyze the data. Such computing power can be used to transform so many other areas of science and research.

2.4 Financial Trading

High frequency trading (HFT) is an area in which big data today plays a major role. Big Data algorithms are used here to make business decisions. Today, the majority of stock market transactions are carried out via data algorithms which increasingly take into account the signals of social media networks and news sites to make buying and selling decisions in a fraction.

III. CONCLUSION

We have entered an era of big data. The paper describes the concept of big data with 3 Vs, volume, speed and variety of big data. The document also discusses big data processing issues. These technical challenges must be met for efficient and rapid processing of big data. These technical challenges often occur in a large number of application areas and can therefore only be addressed cost-effectively in one area. The article describes Hadoop, open source software for processing big data.

REFERENCES

[1] S.Vikram Phaneendra, E.Madhusudhan Reddy,“Big Datasolutions for RDBMS problems- A survey”, *In 12thIEEE/IFIP Network Operations & Management Symposium (NOMS 2010)* (Osaka, Japan, Apr 19{23 2013).

[2] Aveksa Inc. (2013). Ensuring “Big Data” Security with Identity and Access Management. Waltham, *MA: Aveksa.*

[3] Hewlett-Packard Development Company. (2012). Big Security for Big Data. L.P.: *Hewlett-Packard Development Company.*

[4] Kaisler, S., Armour, F., Espinosa, J. A., Money, W. (2013). Big Data: Issues and Challenges Moving Forward. International Confrence on System Sciences (pp. 995-1004). *Hawaii: IEEE Computer Soceity*.

[5] Katal, A., Wazid, M., Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. *IEEE*, 404-409.

[6] Marr, B. (2013, November 13). The Awesome Ways Big Data is used Today to Change Our World. Retrieved November 14, 2013, from LinkedIn: https://www.linkedin.com/today /post/article/20131113065157-64875646-the-awesome-ways-bigdata-is-used-today-tochange-our-world

Citation of this Article:

Mr. Sagar Sureshrao Kalbande, Mr. Shubham Anil Pardakhe, Prof. Anurag Dwivedi, “An Analytical Study on Big Data and Hadoop” Published in International Research Journal of Innovations in Engineering and Technology (IRJIET), Volume 4, Issue 1, pp 13-16, January 2020.

****