International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-6, November 2021
https://doi.org/10.55524/ijircst.2021.9.6.22
Article ID IRPS12032, Pages 96-99
www.ijircst.org

A Review Paper on Hadoop Architecture

Pankaj Saraswat[1], and Swapnil Raj[2]

[1, 2] SOEIT, Sanskriti University, Mathura, Uttar Pradesh, India

Correspondence should be addressed to Pankaj Saraswat; pankajsaraswat.cse@sanskriti.edu.in

ABSTRACT- Hadoop is an open-source software-programming platform for storing and processing huge amounts of data. Its framework is built on Java programming, with some native C code and shell scripts thrown in for good measure. HDFS (Hadoop Distributed File System) is a global; highly failure file system intended to operate on low-cost commodity hardware. Big Data is the term used to describe this massive amount of data. In a world where data is generated at such a rapid pace, it must be preserved, evaluated, and dealt with. It is a subproject of the Apache Hadoop project. HDFS is designed for operations with big data volumes and offers high accessibility to application data. The main features of HDFS are discussed in this article, as well as a high-level overview of the HDFS structure. Hadoop is a technology that will be used in the future, particularly by big businesses. The quantity of data being generated is only going to grow, and the need for this software is only going to grow.

KEYWORDS- Architecture, Big Data, Framework, Hadoop, HDFS.

I. INTRODUCTION

Hadoop is an open source framework for a collection of tools released under the Apache License. In today's society, the concept of a strong individual has shifted. One who has access to data is powerful. This is because data is growing at an exponential pace. Assume we live in a world where everything is data. The data is then generated in the past two to four years, accounting for 90 percent of the total. This is because today, when a kid is born, she is exposed to the light of the camera before her mother. All of these images and movies are just data. Similarly, there is email data, data from different smartphone apps, statistics data, and so on. All of this information has tremendous potential to influence different events and trends. This information is utilized not just by businesses to influence their customers, but also by politicians to influence elections. Big Data is the term used to describe this massive amount of data. In a world where data is generated at such a rapid pace, it must be preserved, evaluated, and dealt with. This is where Hadoop enters the picture [1,2].

A. Evolution of Hadoop

Doug Cutting and Michael Cafarella created Hadoop in 2005. Hadoop's design is based on Google's. Hadoop uses the HDFS to store and handle massive amounts of data, and the Map Reduce technology to process it. The Google File System (GFS) and Map Reduce were used to develop HDFS and Map Reduce. Google surpassed all other search engines and became the most popular and lucrative search engine in the year 2000. Google's success has been ascribed to its proprietary Google File System and Map Reduce. Until that moment, no one but Google knew about it [3,4]. Therefore, in the year 2003 Google published several papers about GFS. However, comprehending Google's entire operation was insufficient. As a result, Google re-released the remaining documents in 2004. In the year 2005, two enthusiasts, Doug Cutting and Michael Cafarella, examined those articles and created Hadoop. Doug's kid had a toy elephant named Hadoop; therefore, Doug and Michael named their new invention "Hadoop" and used the symbol "toy elephant" to represent it. Hadoop has progressed in this manner. Thus, although Doug Cutting and Michael Cafarella developed HDFS and Map Reduced, Google initially motivated them [5].

Hadoop is an open source framework for a collection of tools released under the Apache License. It manages, stores, and processes data for a variety of big data applications that operate on clustered computers. Big Data was formerly characterized by the "3Vs," but now there are "5Vs" of Big Data, which are also known as

B. Big Data characteristics

1) Volume

Data is being produced in enormous quantities as our reliance on technology grows. Data generated by different social networking sites, sensors, scanners, airlines, and other businesses are common examples.

2) Velocity

The pace at which data is produced is enormous. Every person will generate 3mb data every second by the end of

2020, according to estimates. This massive amount of data is being produced at a rapid pace.

3) Variety

There are three kinds of data generated by various methods:

- ***Structured Data***

Relational data that is kept in the form of rows and columns is referred to as structured data.

- ***Unstructured Data***

Texts, images, videos, and other unstructured data that cannot be stored in rows and columns are examples of unstructured data.

- ***Semi-Structured Data***

Semi-structured data, such as log files, is one example of this kind of information.

4) Veracity

The word veracity refers to data that is inconsistent or incomplete, resulting in dubious or uncertain information. The number or quantity of data often causes data inconsistency, for example, data in bulk may generate confusion, while data in little amounts can only communicate half or partial information.

5) Value

Unless information is transformed and becomes useful, a big volume of information with no worth is pointless to the organization. Knowledge has no worth or meaning in and of themselves; in addition to collect content, it must be turned into anything usable. Resulting, you may say that Value is the most significant of the five Vs.

C. Components of Hadoo

1) HDFS

HDFS is a specialized file system for storing large amounts of data on commodity or low-cost hardware using a streaming access pattern. It allows data to be stored across many cluster nodes, ensuring data security and fault tolerance.

2) Map Reduce

Once data is saved in HDFS, it must be processed. Assume a query is made to the HDFS to process a data set. Hadoop now determines where this data is stored, a process known as mapping. This is referred to as the reduction process. As a result, Map Reduce is used to process the data while HDFS is utilized to store it [6,7]. Figure 1 shows the Map reduce architecture.

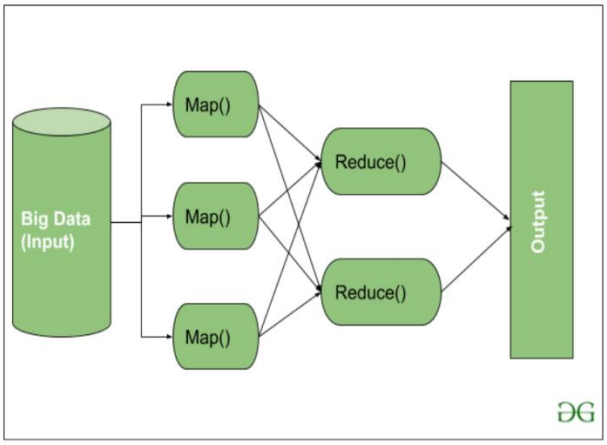

Figure 1: The above figure shows the Map Reduce architecture [geeksforgeeks]

3) YARN

YARN is an acronym for Yet another Resource Negotiator. A Hadoop-specific operating system controls the cluster's resources and serves as a foundation for Hadoop task scheduling. First Come, First Serve, Fair Share Scheduler, and Capacity Scheduler are some of the many kinds of scheduling. YARN's default scheduling is set to First Come, First Serve [8].

D. Versions of Hadoop

1) Hadoop1

This is Hadoop's earliest and most basic version. Hadoop Standard HDFS, and Map Reduce are all included. Figure 2 shows the version of Hadoop i.e. Hadoop1.

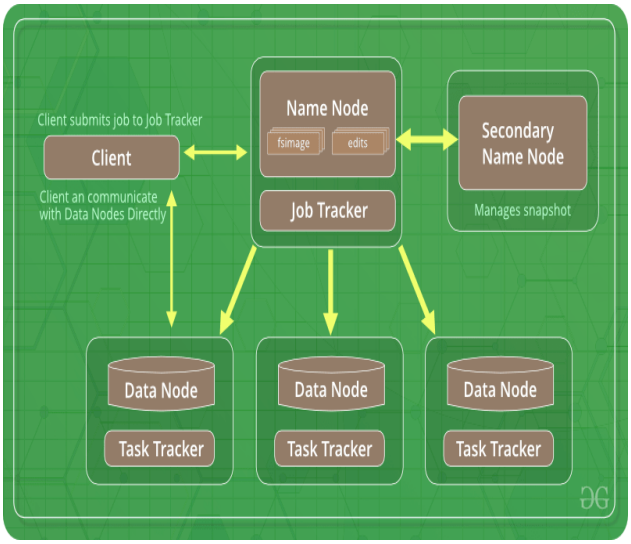

Figure 2: The above figure shows the version of Hadoop i.e. Hadoop1 [geeksforgeeks]

2) Hadoop2

Hadoop 2 includes YARN, which is the sole change between Hadoop 1 and Hadoop 2. (Yet another Resource Negotiator). YARN's two daemons, job tracking and progress monitoring, aid in resource management and

task scheduling. Figure 3 shows the version of Hadoop i.e. Hadoop2.

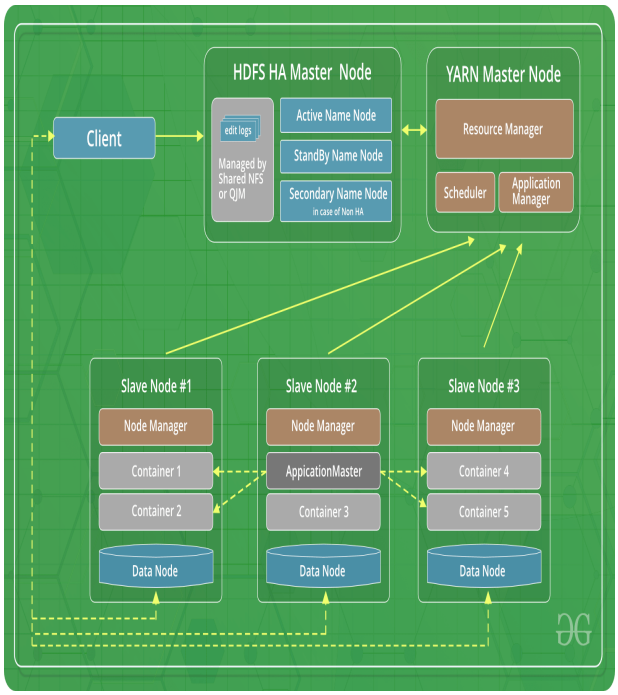

Figure 3: The above figure shows the version of hadoop i.e. Hadoop2 [geeksforgeeks]

3) Hadoop3

This is the most current Hadoop version. Hadoop 3 offers one major benefit in addition to the benefits of the previous two versions. By having many name nodes, it has addressed the problem of single point failure. Other features, like as erasure coding, GPU hardware, and Dockers, make it superior than previous Hadoop versions [9,10].

- ***Economically Feasible***

It is less expensive to store and analyze data than it was in the past. Because the computers that hold the data are just commonplace hardware.

- ***Easy to Use***

The Apache Hadoop projects or collections of tools are simple to work with when analyzing large data volumes.

- ***Open Source***

There is no requirement to buy for Hadoop because it is public sourced platform licenced underneath the Apache License. Simply install and use it.

- ***Fault Tolerance***

Because Hadoop maintains three copies of data, the data is secure even if one copy is lost due to a commodity hardware failure. Furthermore, since Hadoop version 3 includes many name nodes, Hadoop's single point of failure has been eliminated.

- ***Scalability***

Hadoop's scalability is unrivalled in the industry. If the cluster has to be scaled up or down, all that is required is to alter the quantity of commodity hardware in the cluster.

- ***Distributed Processing***

HDFS and Map Reduce guarantee that data is stored and processed in a distributed manner.

- ***Locality of Data***

Data locality is one of Hadoop's most attractive and promising characteristics. Instead of sending the data to the local computer to conduct a query over a data collection, Hadoop sends the query to the server and retrieves the result from there. This is referred to as data locality.

E. Advantages of Hadoop

- Possibility of storing a huge quantity of information.
- High degree of adaptability.
- It is economical.
- High-performance computing.
- Tasks are self-contained.
- Scaling that is linear.

F. Disadvantages of Hadoop

- For tiny amounts of data, this method is ineffective.
- Cluster administration is a difficult task.
- Has problems with stability.
- There are worries about security.

II. DISCUSSION

The Big Data Hadoop ecosystem evolves at a rapid pace, and some studies may become outdated in as little as two to three years. According to a 2015 article, NoSQL systems "only offer basic query interfaces, such as key-based access or single-table queries, forcing application developers to perform more sophisticated processes such as joins themselves". As mentioned in this article, SQL enquiries can now be incorporated in NoSQL datasets, perhaps with the assistance of request motors that can incorporate information from diverse source materials using a block of statements, such as having joined a table from one NoSQL directory with that other from another NoSQL directory or even from a distributed information inventory remedy like Hive.

Set phoenix. Transactions. Enabled = true to allow ACID (Atomicity, Consistency, Isolation, Durability) compliant tables over NoSQL tables using query engines like Phoenix. Hive, on the other hand, is a data warehousing system that can offer complete ACID semantics at the row level right out of the box. The suggested architecture considers the needs for storing and processing huge amounts of data related with smart cities. Benchmarking every product, including desktop PCs, servers, data center infrastructures, and mobile devices, contains biases, fallacies, and dangers, in part because real-world situations are difficult to categorize and match. This may lead to architects identifying the incorrect bottlenecks and making incorrect trade-offs. Things become much more complex when it comes to benchmarking distributed systems since performance is affected by data

distribution, the route to data for a particular query, and the heterogeneous network that links the nodes.

III. CONCLUSION

There was no way to interact with anyone before the internet. The only method to go from one person to another is to write letters, which take a long time to arrive. Everyone now has access to the internet, which allows them to read whatever they want and do searches on their phones, computers, and other devices. Life has gotten easier because of the internet. Big data and Hadoop, as well as cloud computing, are two current technologies that the author has covered. Nowadays, everyone is considerably more familiar with these languages and some of these sites. Because everyone is using them to store their data, they are in short supply. The author has addressed cloud computing, big data, and Hadoop in this article. The four V's of big data, big data problems, Hadoop, and cloud computing When you have a lot of data, everyone prefers this since these platforms are safe if you put them to private mode, and your papers are stored. You can also have an overabundance of them whenever you desire. Hadoop is an open-source framework, and cloud computing is the internet that links the entire globe. Big data refers to enormous amounts of data, and hadoop is an open-source framework. We have also spoken about the benefits and varieties of cloud, as well as their service model. All of the technologies' characteristics.

The author has covered hadoop, big data, and cloud computing since they are extremely significant nowadays because it is quite difficult to upload data or research something about anything. All updates, whether organised or unstructured, are available there. Everything is connected to everything else. Nowadays, everyone has many data, and in order to keep it and avoid taking any risks, they all want to transfer it to the cloud. The author of this review article covered big data, which is defined as having a huge quantity of data saved on the cloud, hadoop, which is an open source framework, and cloud computing, which is defined as connecting to the rest of the world. What are the V's of big data, and what are the problems and benefits of cloud computing, big data, and Hadoop? Data volumes will continue to grow in the future, and more data will be transferred to the cloud. Data volumes will rise in the future and will be transferred to the cloud, as data quantities will grow in the future.

REFERENCES

[1] Hanson JJ. An introduction to the Hadoop Distributed File System. Dev Work. 2011;

[2] Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics. 2010;

[3] Arslan M, Riaz Z, Munawar S. Building Information Modeling (BIM) Enabled Facilities management using hadoop architecture. PICMET 2017 - Portl Int Conf Manag Eng Technol Technol Manag Interconnected World, Proc. 2017;2017-January:1–6.

[4] Diaconita V, Bologa AR, Bologa R. Hadoop oriented smart cities architecture. Sensors (Switzerland). 2018;18(4).

[5] Kumar DGVRRCHVVNS. Hadoop Distributed File System and Map Reduce Processing on Multi-Node Cluster. Int J Sci Res. 2015;

[6] Afrati FN, Ullman JD. Optimizing multiway joins in a map-reduce environment. IEEE Trans Knowl Data Eng. 2011;

[7] Shahabinejad M, Khabbazian M, Ardakani M. An efficient binary locally repairable code for hadoop distributed file system. IEEE Commun Lett. 2014;

[8] Mavani M. Comparative Analysis of Andrew Files System and Hadoop Distributed File System. Lect Notes Softw Eng. 2013;

[9] Borthakur D. The hadoop distributed file system: Architecture and design. Hadoop Proj Website. 2007;

[10] Shvachko K, Kuang H, Radia S, Chansler R. The Hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010. 2010.