International Journal of Scientific Research & Engineering Trends
Volume 8, Issue 2, Mar-Apr-2022, ISSN (Online): 2395-566X

The Basic HADOOP: HDFS, MapReduce, YARN

Ayushi Ankita Rakshit
Dept. of Electronics and Communication Engineering,
Asansol Engineering College,
Asansol, West Bengal.

Abstract- In our day-to-day life, we encounter with a huge amount of data. As a student, the name of the subjects, syllabus, terminologies, etc should be kept in mind, as a teacher, the name of the students, subjects, syllabus, etc should be kept in mind. Every small thing here represented are considered as a data. For a computer, files like images, videos, documents, etc, are considered as a data. We as a human deal with fewer amounts of data as compared to that of the computers. Working with millions and billions of data is not possible by humans. So, to overcome the problem Big Data was introduced and works with Hadoop as an ecosystem. This paper is a technical brief on how Hadoop ecosystem works, its principles and basic algorithms.

Keywords- Big Data, Hadoop, HDFS, MapReduce, YARN, Python.

I. INTRODUCTION

Companies like Google, Microsoft, Amazon deals with millions of data in a day. The basic principle they use is Hadoop. It was originated by Google File System paper published by Google. Hadoop became the fastest system to deal with 1 terabyte of data on a 900-node cluster within 209 seconds in 2008. After many modifications, Hadoop can deal with petabytes of data.

As,

1 petabyte = 1000 terabyte

1 terabyte = 1000 gigabyte

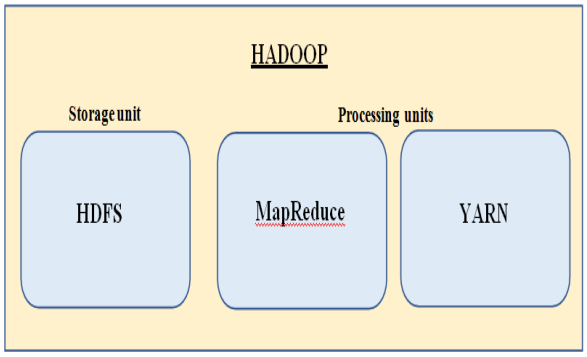

Fig 1. Three major parts of HADOOP.

So, we can just imagine the amount of data that can be handled by Hadoop ecosystem. The data are being processed and stored in HDFS i.e., Hadoop Distributed File System. The basic operation to be performed is decided by MapReduce (major component of Hadoop) as it has two parts: Mapper and Reducer. YARN (Yet Another Resource Negotiator) is that portion of Hadoop that performs the operation or a set of operations.

So, the work is divided between MapReduce and YARN. There are many more parts of Hadoop but we focused on only these three parts.

In Hadoop, the huge amount of data is parallelised, which means divided in parts and stored in nodes. Mainly, the division is done in 128mb. So, data are distributed in many nodes and the nodes together are known as Cluster.

This makes processing faster than other techniques. The metadata of each node is being stored in a parent node which is named as Name node. In case of corruption of any of the node Name node is responsible for backup. This is maintained by Hadoop itself.

II. FINDING THE FREQUENCY OF WORDS IN CERTAIN TEXT FILE

Let us consider a book which contains more than 100 pages, and the given task is to find the frequency of each word present in the book and time limit is set for 30 minutes. So, for a human being it is not possible to do the task in the stipulated time frame. Here comes the concept of MapReduce,

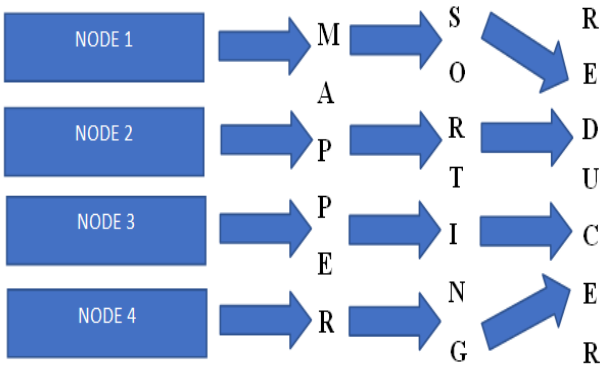

Fig 2. Basic MapReduce operation.

Here, the data is being divided into several nodes means the lines of the books or paragraphs are divided into several nodes and the metadata of each node is stored in name node. Then Hadoop decides how many mappers to be introduced for processing, sorting is done by Hadoop itself and finally the reducer gives the output. There are three class to be specified: Mapper class, Reducer class and Job class. Mapper class will perform the map operation, reducer class will perform the reduce operation and objects for each class will be specified in job class which will point to the specified classes.

After the operations to be selected, YARN (Yet another Resource Negotiator) comes into action. The specified tasks, map and reduce operations are performed by YARN. It is responsible for the operations and processing part.

Finally, the processed data is being stored in HDFS. Hadoop always takes text as an input and produces text as an output. It works on key-value pair. As Hadoop is written in JAVA language, preference is given to the programmers to write the code in JAVA but using API, a programmer can code in Python, ruby or any other object-oriented language.

III. CONCLUSION

In this paper, the main idea of Big Data, Hadoop, YARN and HDFS has been described briefly. Also, the example of calculating the word frequency in a text has been explained. The basic operations of Hadoop: MapReduce, YARN and HDFS are taken into consideration for making this paper more effective.

REFERENCES

[1] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google, Inc.
[2] Nada Elgendy Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", August 2014 Lecture Notes in Computer Science 8557:214-22 DOI: 10.1 007/978-3-319-08976-8_16 Conference: Industrial Conference on Data Mining.
[3] D. P. Acharjya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.
[4] Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur," Research Paper on Big Data and Hadoop", IJCST Vol. 7, Issue 4, Oct - Dec 2016.
[5] Yusuf Perwej, "The Hadoop Security in Big Data: A Technological Viewpoint and Analysis", Isroset-Journal Vol.7, Issue.3, pp.10-14, Jun-2019.
[6] Wulun Du; Depei Qian; Ming Xie; Wei Chen, "Research and Implementation of MapReduce Programming Oriented Graphical Modeling System", 2013 IEEE 16th International Conference on Computational Science and Engineering.