International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

A Review Paper on Big Data Hadoop Frame Work

Sindhu Daniel

Assistant Professor, MCA Department, Mount Zion College of Engineering, Kadammanitta, Kerala, India
sindhudaniel[at]yahoo.com

Abstract: ***Big data is dataset that having the ability to capture, manage & process the data in elapsed time .Managing the data is the big issue. And now days the huge amount of data is produced in the origination so the big data concept is in picture. It is data set that can manage and process the data. For managing the data the big data there are many technique are used .One of this technique is Hadoop. it can handle the huge amount of data, it is very cost effective, and it can handle huge amount of data so processing speed is very fast, and also it can create a duplicate copy of data in case of system failure or to prevent the loss of data. This paper contains the Introduction of big data and Hadoop, characteristics of big data, problem associated with big data, latest tools and components of hadoop.***

Keywords: Bigdata, Hadoop, Mapreduce, Hive, pig, HBAS, Spark

1. Introduction

Bigdata is a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it has become a complete subject, which involves various tools, techniques and frameworks. Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types. Structured data − Relational data, Semi Structured data − XML data, and unstructured data.

Traditional Approach

In this approach, an enterprise will have a computer to store and process big data. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc. In this approach, the user interacts with the application, which in turn handles the part of data storage and analysis.

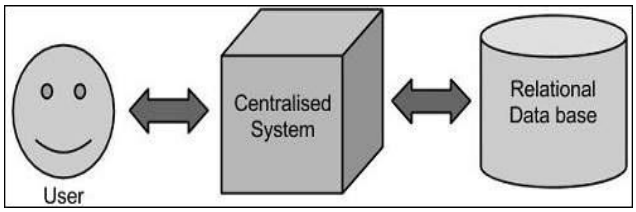

Figure 1: Traditional components

2. Limitation

This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data. But when it comes to dealing with huge amounts of scalable data, it is a hectic task to process such data through a single database bottleneck

Hadoop Components

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts Hadoop has two major:

(a) Hadoop Distributed File System (HDFS)
(b) Processing layer (MapReduce)

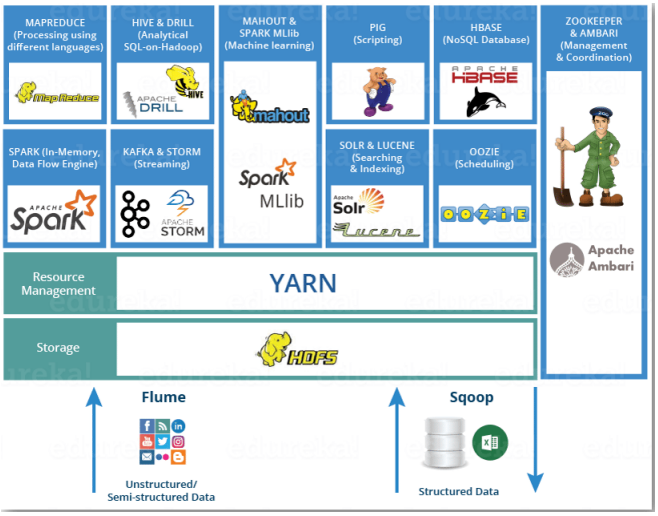

Figure 2: Hadoop Ecosystem

HDFS

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system. HDFS has master slave architecture. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. HDFS holds very large amount of data and provides easier access. The main components of HDFS are: Name Node and Data Node. The NameNode manages metadata. The DataNodes store the data.

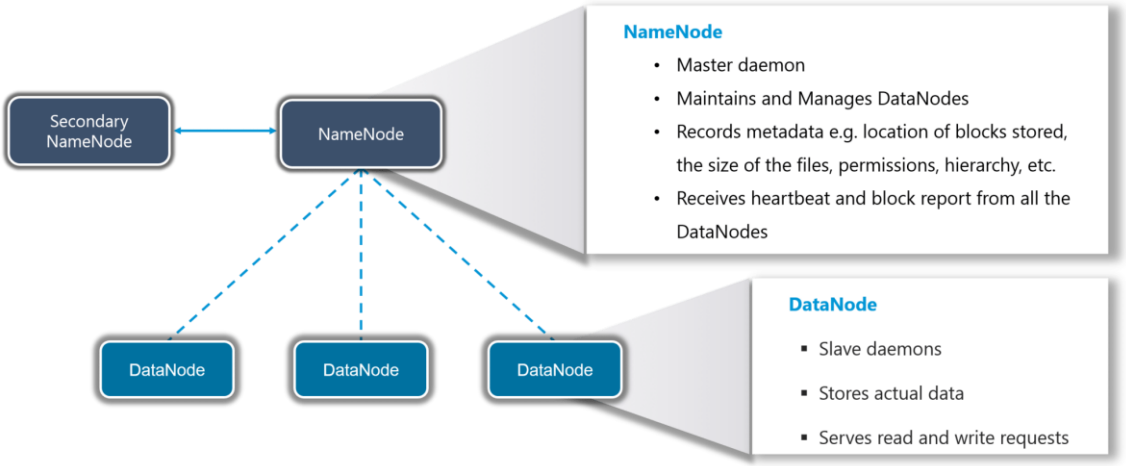

Figure 3: HDFS Nodes

MapReduce

MapReduce framework is the processing pillar of hadoop. The framework is applied on the huge amount of data divided in part and run parallel. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

- Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
- Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

Working of MapReduce

The mapReduce programming model also works on an algorithm to executes the map and reduce operations. The algorithm can be depicted as follows:

- Take a large dataset or set of records.
- Perform iteration over the data.
- Extract some interesting patterns to prepare an output list by using the map function.
- Arrange the output list properly to enable optimization for further processing.
- Compute a set of results by using the reduce function.
- Provide the final output.

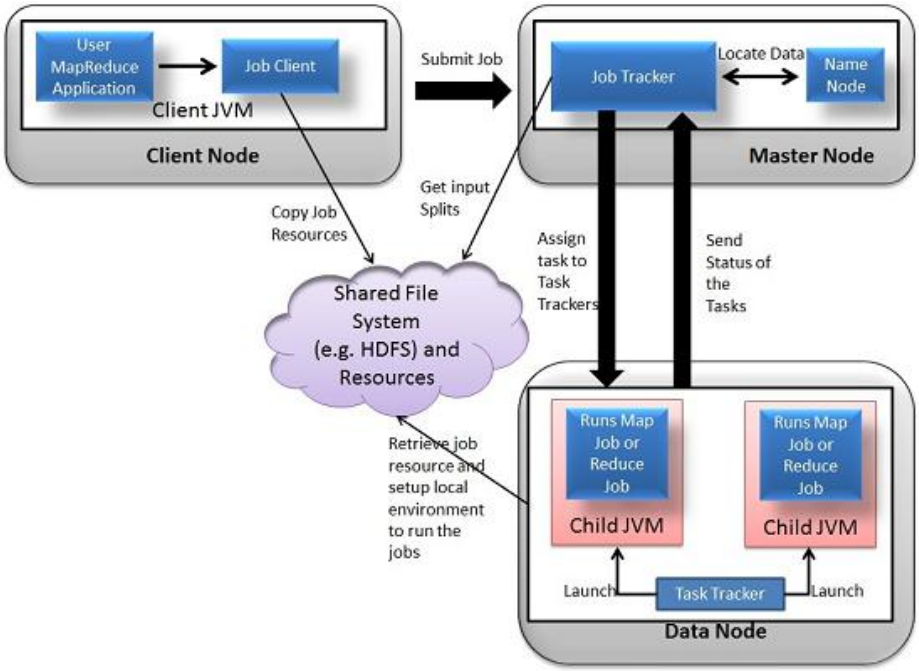

Figure 4: Working of MapReduce

YARN
Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System. Consists of three major components i.e.

1) Resource Manager
2) Nodes Manager
3) Application Manager

Resource manager has the privilege of allocating resources for the applications in a system whereas Node managers work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager. Application manager works as an interface between the resource manager and node manager and performs negotiations as per the requirement of the two.

PIG:
Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL. It is a platform for structuring the data flow, processing and analyzing huge data sets. Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS. Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the JVM. Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

HIVE:
With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL data types are supported by Hive thus, making the query processing easier. Similar to the Query Processing frameworks, HIVE too comes with two components: *JDBC Drivers* and *HIVE Command Line*. JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.

Mahout:
Mahout, allows Machine Learn the ability to a system or application. Machine learning as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms. It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning. It allows invoking algorithms as per our need with the help of its own libraries.

Apache Spark:
It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc. It consumes in memory resources hence, thus being faster than the prior in terms of optimization. Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing; hence both are used in most of the companies interchangeably.

Apache HBase:
It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively. At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data.

Other Components: Apart from all of these, there are some other components too that carry out a huge task in order to make Hadoop capable of processing large datasets. They are as follows:

- **Solr, Lucene:** These are the two services that perform the task of searching and indexing with the help of some java libraries, especially Lucene is based on Java which allows spell check mechanism, as well. However, Lucene is driven by Solr.
- **Zookeeper:** There was a huge issue of management of coordination and synchronization among the resources or the components of Hadoop which resulted in inconsistency, often. Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.
- **Oozie**: Oozie is an open source Hadoop service used to manage and process submitted jobs.

Advantages of Hadoop

- Varied Data Sources: Hadoop accepts a variety of data. Hadoop can accept data in a text file, XML file, images, CSV files etc.
- Cost-effective: This requires fewer machines to store data as the redundant data decreased significantly.
- Performance: Hadoop with its distributed processing and distributed storage architecture processes huge amounts of data with high speed. .
- Fault-Tolerant: In Hadoop 3.0 fault tolerance is provided by erasure coding.
- High Throughput: Throughput means job done per unit time. Hadoop stores data in a distributed fashion which allows using distributed processing with ease.
- Open Source: Hadoop is an open source technology i.e. its source code is freely available. We can modify the source code to suit a specific requirement.
- Scalable: Hadoop works on the principle of horizontal scalability i.e. we need to add the entire machine to the cluster of nodes

3. Conclusion

Managing the data is the big issue. And now days the huge amount of data is produced in the origination so the big data concept is in picture. It is data set that can manage and process the data. For managing the data the big data

technique is used i.e. hadoop. Hadoop can handle the huge amount of data, it is very cost effective, and it can handle huge amount of data so processing speed is very fast, and also it can create a duplicate copy of data in case of system failure or to prevent the loss of data. In this paper, we have seen an overview of big data, characteristic of big data(volume, variety, velocity), its importance in business, advantages, some issues and challenges regarding to big data and also latest tools and components which is used to implement big data concept. This paper specifies the Hadoop environment, its architecture and its components.

References

[1] R. H. Katz, E. D. Lazowska and R. Bryant, "Big-data computing: creating revolutionary breakthroughs in commerce", 2008.
[2] B.M. Purcell, "Big Data using cloud computing", 2013
[3] Apache Software Foundation, http://hadoop., apache.org /" (2017) [Online].
[4] Rodero-Merino, Luis, and Gilles Fedak. "MapReduce and Hadoop." Open Source Cloud Computing Systems: Practices and Paradigms (2012): 197.
[5] "A Review Paper on Big data & Hadoop" Rupali Jagadale, Pratibha Adkar.
[6] A.Ting,https://wiki.apache.org.(2017,Dec.)[Online].https://wiki.apache.org/hadoop/PoweredBy
[7] J. Manyika, "Big data: The next frontier for innovation, competition, and productivity", San Francisco, 2011.
[8] N. Guruprasad, "Hadoop Vendor Distributions"