International Journal of Research in Engineering Science
ISSN: 2815-0511 Volume: 1 Issue: 2
Article Received: *10 October 2021* ***Revised:*** *15 November 2021* ***Accepted:*** *20 November 2021* ***Publication:*** *31 December 2021*

Analysis of the Hadoop Architecture, Modules, and Components, As Well as Knowledge of Its Potential and Restrictions

Dr. Kamaldeep Garg

Assistant Professor

Chitkara University, India

kamaldeep.garg@chitkara.edu.in

Abstract: -Big data sets can range in size from gigabytes to petabytes, and Hadoop is an open source framework used to store them. Unlike Hadoop architecture, which makes it possible to cluster data sets so that multiple datasets can be evaluated concurrently without interfering with one another, traditional methods of data storage will store the data in one large computer as a result of which only one data set can be analysed at a given point in time. In comparison to other data storage techniques, Hadoop processes data more quickly and provides quick responses. Additionally, the Hadoop architecture offers building blocks on which other services and applications may be developed. Spark, Hive, Presto, Hbase, and other popular Hadoop applications are listed here. In Hadoop.

Keywords: -Hadoop introduction, Hadoop's operational mechanism, Benefits and difficulties of the Hadoop architecture, Module and components of Hadoop.

Introduction: - [1]

We all know that in an organisation, large volumes of data is present and keeps on adding on a daily basis. The business uses advanced frameworks and tools and databases to store these set of data and information. They are so advanced that anybody who has authorised access to these databases can use them at any point of time as per their requirements. But in traditional methods of data storage, it has been found that the data processing to evaluate the data and perform data analysis is not possible. Only one data set can be executed or accessed at a time. This was time consuming process and use to take hours and great number of efforts by data analysts to perform on various data sets at the same time. This also use to give rise to the errors. In order to overcome these issues, data scientists have innovated latest framework and architecture known as Hadoop framework. It is Apacheopen-source framework which is used to store huge and large volumes of datasets which has big data sizes. This is done by using clustering technique where, clusters of datasets are developed which have same pattern and belongs to one particular category. Then these datasets can be processed simultaneously on distributed systems at the same time. This type of data analysis procedure is much better than traditional methods and gives faster and quick responses to the data analysts. Hadoop framework is the platform which operates in the environment that helps to facilitate distributed storage and computation across various clusters of the computers. Hadoop also ensures the safety of the data and information stored using this framework. They have various security groups implemented in the platform which helps to control the inbound and outbound network traffic to the cluster nodes. It also uses various identity and access management rules and regulations to grant or discard the permissions.

Working mechanism of Hadoop Framework: - [2]

It is very easy to store the data and information using Hadoop framework as it uses the clustering of the data sets. It will use clustering to make small clusters of datasets which shares common characteristics and then provides the facility of performing data analysis on these clusters via distributed systems simultaneously. Following tasks are performed which explains the working mechanism of Hadoop architecture: -

- It has been observed that it is very expensive to control and maintain high configuration severs to store the huge sizes of datasets. Instead of this number of clusters can be formed and the machines can read these simultaneously at a higher speed and gives quick responses and are low-cost initiatives for the data analysis.

- Hadoop executes the code on clusters in the computer systems. Following steps are performed: -

a. First of all, the data and information available is converted into files and folders and directories. The files are developed of same sizes which are known as blocks.

b. To process further, the clusters formed in the first steps are distributed across the number of clusters present in the network.

c. HDFS will monitor and guide the whole processing procedure as it is located on top of the local file system.
d. To avoid the hardware failure scenarios, the blocks are duplicated so that if one block is damaged the duplicate block can be used to avoid the system failure.

e. It is mandatory to evaluate whether the code is executed successfully or not. If the code execution is failed then all the vulnerabilities are evaluated to find the issue and fix it.

f. Playing out the sort that happens between the guide and diminish stages.

g. Once the data is filtered then it will be sent and stored in one specific computer.

h. The documentation of debugging logs for each job performed is necessary to maintain.

Components of Hadoop Architecture: - [3]

Hadoop is open-source architecture which is written using java language which is used to utilise and store big size data sets. Hadoop uses map reduce programming algorithm which was discovered by google. The main contribution of Hadoop is solving big data issues due to which many brand companies like Facebook, Netflix etc are using this technology. There are following four main components of Hadoop framework which will be discussed in detail: -

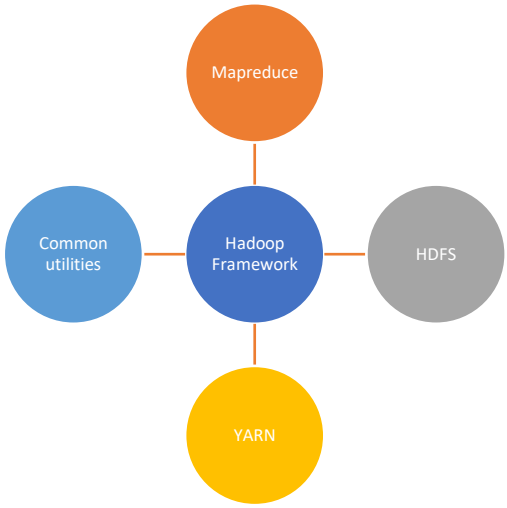

Figure 1 Components of Hadoop Framework.

1. Map reduce: -

- It is based on the YARN or is a data structure which is used to execute distributed processing parallel to each other in the form of data clusters. In MAP reduce the first phase is used to map and the second phase is to reduce.

- MAP Tasks: -

a. Record reader: - The objective of this is to break the records. It is use to provide key value pairs in the map function. The key associates with the location and the value is the data linked to it.
b. Map: - It is used to process tuples which are obtained form record reader. The map function is responsible for generating no key value pair or is responsible for creating number of key-value pairs.
c.
ombiner: - It is use to group data in the workflow of the map. It is optional and it is used to combine the key pair values that are generated in middle.

- Reduce Tasks: -

A. Shuffle and sort: - When the key value generated by the mapper is transferred to reducer then it is known as shuffling process.

B. It is use to sorting the data using the key values.

C. Reduce: - The principal capability or errand of the Reduce is to assemble the Tuple created from Map and afterward play out some arranging and conglomeration kind of interaction on those key-esteem contingent upon its key component.

D. Output format: - Once all the tasks are completed then using record writer the key value pairs are written into the file. Each record will be maintained in the new line and the key and value in space.

2.
DFS (Hadoop distributed file system): -

- HDFS is the Hadoop distributed file system which is used to store the large chunks of data files instead of storing small data blocks.
- It also serves the purpose of fault tolerance and high availability to the storage layer. There are following two nodes in HDFS: -

A. Name node
B. Data node

- Name node: -

1.
ame node works as a master node whose role is to guide the slave which is the data node.
2.
he goal of name node is to store the metadata which is another data regarding the existing data.
3.
eta data is used to record the activities of the client in one particular Hadoop cluster.
4.
nother function of name node is to guide the data node to perform various functions like to delete, create, replicate etc.

- It also gives information about the data node which is closer to that particular data node to provide better communication.
- B. Data Node: -

1. Data node is the slave node in the Hadoop architecture whose job is to follow guidelines given by the Name node.
2. It is used to store data in Hadoop cluster where the number of data nodes can range from 1-500 or more.
3. If the number of data nodes are more then it will be used to store more data, therefore the data nodes must be capable of storing large amount of data.

4. YARN (Yet another resource negotiator): -

- It is the framework on which Map reduce performs its functions. Following are the two main functions performed by the YARN: -

- Job scheduling: - The purpose of job scheduler is to divide the big tasks into small tasks so that the processing time for each job can be reduced. It is also use to identify which job has more priority and which job has low priority. It also identifies dependencies of jobs and job timing information etc.

- Resource management: - Its sole responsibility is to manage all the resources that are available for the execution of Hadoop cluster.

1. Common Utilities: -

- Common utilities are the java files and java library as well as java script which are required for all other components in the Hadoop cluster.
- These utilities are used by YARN, HDFS, Mapreduce to perform their tasks.
- Hadoop Common confirm that Hardware disappointment in a Hadoop bunch is normal so it should be settled consequently in programming by Hadoop Framework.

Advantages of Hadoop Framework: - [4]
Following are few advantages of Hadoop framework: -

1. Scalability: - Hadoop is scalable framework which is used to store big size data which can be processed parallely. Based upon the requirements the number of machines used in the framework can be increased or decreased

2. Flexibility: - Hadoop is flexible framework which can be used with any kind of data set like structured data, unstructured data etc. Hence it can process any format of data as it is flexible and independent of the data type.

3. Speed: - Since Hadoop uses HDFS so it divides the large size data sets into small blocks of data hence the processing speed to analyse these small datasets in the cluster is faster and gives quick responses.

4. Fault tolerance: - Hadoop also takes care of the failure of the hardware components. Hence it keep the replica of the data stored in it which can be useful in case of fault or can be used to control the damage caused.

5. Less network traffic: - Since the task is divided into small jobs and given to nodes and hence the network traffic is low.

Disadvantages of Hadoop: -
Hadoop also has some challenges and limitations as described below: -

1. Vulnerability: - Hadoop is open source frame work which is written using java language which has many vulnerabilities as it can easily be exploited or attacked by the intruder.

2. Low performance with less data volumes: - The efficiency of Hadoop decreases if the amount and size of data stored is less.

3. Security issues: - Hadoop uses kerberos to implement security feature but fails to do so as it does not have storage and network encryption.

4. Supports only batch processing: - The cluster cycle is only the cycles that are running behind the scenes and has no sort of collaboration with the client. The motors utilized for these cycles inside the Hadoop center isn't unreasonably much proficient. Creating the result with low inactivity is unimaginable with it.

Conclusion: - - Hadoop is an edge work which is open source and it is use to store bigdata sets whose sizes can differ from gigabytes to petabytes. Conventional strategies for information capacity will store the information in one enormous PC because of which only one informational collection can be broke down at given mark of time while in Hadoop engineering, grouping of informational indexes is conceivable because of which numerous datasets can be assessed parallelly without impeding one another. The information handling speed utilizing Hadoop is quicker and gives fast reaction when contrasted with different information stockpiling strategies. Hadoop engineering likewise gives building blocks on which many administrations and applications can be assembled. The normal uses of Hadoop are Spark, Hive, Presto, Hbase and so on. Hadoop can be executed on stages like AWS where executing it is simple. The charges for the stage depends on the use of the stage which implies the expense can be diminished as it relies on the quantity of hours it is being utilized by the associations to accomplish their business objectives.

Reference:-

1.https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm#:~:tex
2.https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/
3.https://www.geeksforgeeks.org/hadoop-architecture/
4.https://www.geeksforgeeks.org/hadoop-pros-and-cons/