|  |
| --- |
| Introduction to Hadoop |
|  |
|  |
| **Name: Dipanwita Mullick**  **Associate ID:388069**  **Date: 12/22/2015** |
|  |

`

|  |
| --- |
|  |

**Table of Contents**

**INTRODUCTION** **3**

**ARCHITECTURE OF HADOOP 3**

**FLOW IN HADOOP 5**

**PROCESS OF MAPREDUCE** **5**

**ADVANTAGES OF HADOOP 6**

**REFERENCES 7**

**Introduction:**

Hadoop is basically an open source platform that is based on java and is used to store and process large amount of data among distributed computer systems. It is basically used to scale up multiple servers that do not share same memory. Hadoop can be considered as an ecosystem that contains several components that work together as a single platform. Hadoop is sponsored by Apache Software Foundation. Hadoop is inspired by Google’s MapReduce that is a software framework to break an application in several parts. The founder of Hadoop is Doug Cutting.

**Architecture of Hadoop:**

Basic components of Hadoop are:

* Cluster
* YARN infrastructure
* Hadoop Distributed File System(HDFS)
* Hadoop MapReduce

**Cluster:** Cluster is set of machines that is known as nodes.

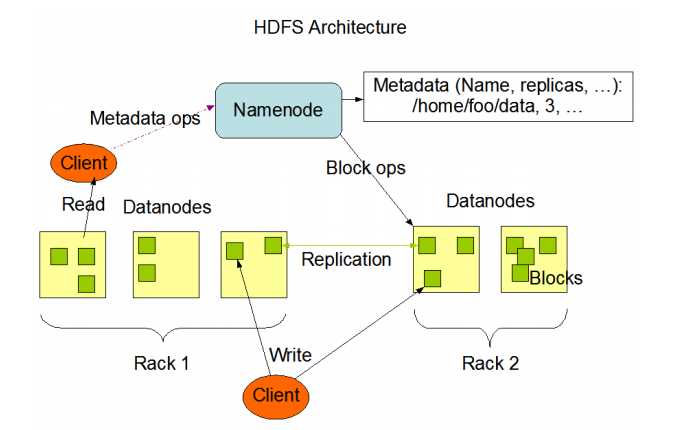
**YARN:** YARN is basically a framework that is used to provide the resources like CPU, memory to run the application.

**Hadoop Distributed File System(HDFS):** Hadoop Distributed file system is distributed, portable and scalable framework that is suitable for the large application. It is used to analyze the and transformation of data using MapReduce algorithm that contains large volume. HDFS stores the large volume of data across multiple machines by creating the replica of data. Usually the replication values is 3 that is the datas are stored in 3 nodes- one in different rack and other two in same rack. HDFS basically contains master slave architecture where there is one master node that is called NameNode and others are called DataNodes.

* **Namenode** is used to store the metadata of HDFS and contains information of location of the files and doesnot contain any actual data. When a file is entered into a server the corresponding location is stored in NameNode so that it can be retrieved easily in future whenever is required.

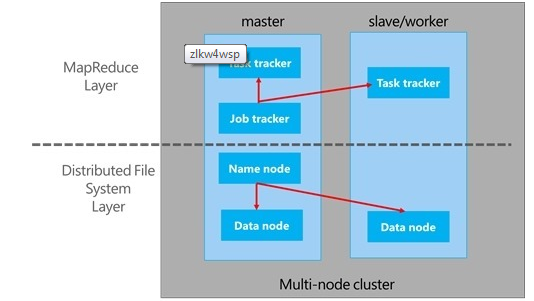
There is also a secondary NameNode that co ordinate with the Primary Namenode and keeps the snapshot of the metadata into local storage.

* **DataNode** is responsible to store the files and actual data in the server by creating file blocks within the node. DataNodes are also known as Slave. It sends the location information of files to NameNode and always responds to the NameNode for any operation. DataNodes consume lot of memory space as it is used to store the actual data. When a DataNode is down, it does not affect the operation as NameNode creates a replica of that particular DataNode that is down.
* **File System NameSpace:** HDFS is basically a traditional file organization where an user can create directory and within that directory then can save the files. One can create or remove files, move the required files from one to another directory. This file system NameSpace is maintained by NameNode. Any change in this is tracked by NameNode.
* **Data Replication:** HDFS stores large files among several machines in a cluster. HDFS stores each file as sequence of blocks and except last block all blocks are of same size. All the blocks of file are replicated and maintained by NameNode. NameNode gets a response from DataNode at each point of time. When NameNode receives the response from DataNode it implies that DataNode is acting properly otherwise it implies that DataNode is not alive.



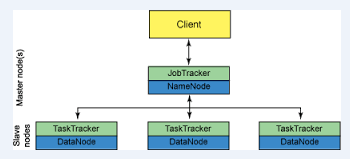
**Hadoop MapReduce:** Hadoop MapReduce is a framework based on java that consists of two tasks name Map and Reduce. In Map Stage the input data is divided to independent group that are processed in parallel and then in the Reduce Stage it sorts the output of the Map function and the result is stored in the filesystem. It also follows are master slave architecture where JobTracker acts as master and TaskTracker acts as slave.

* **JobTracker** is responsible to take the request from the client and assign them to the available tasktracker for execution. JobTracker is aware of file location, so it assigns the job to that tasktracker that is nearby the data. During job execution, JobTracker keeps the tracking of the progress of the job by co coordinating with the TaskTracker. If any job is failed by particular TaskTracker, then it will notify the JobTracker that the task has been failed and then JobTracker assigns the job to the another TaskTracker.
* **TaskTracker** is a component of HDFS that is responsible to receive tasks from jobtracker. The number of slots that is how many tasks can be taken by TaskTracker, it is preconfigured. TaskTracker always send the available slots to the JobTracker and also send an update of the task to JobTracker .



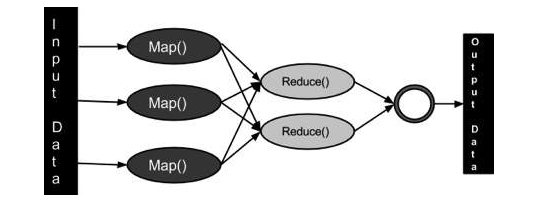
**Flow in Hadoop:**

* Client submits the job to JobTracker.
* JobTracker coordinates with NameNode to find the location of the data and assigns the job to the tasktracker.
* TaskTracker receives the request from JobTracker and execute the task as per MapReduce Function and output of Reduce function is stored.



**Process of MapReduce:**

MapReduce is a framework that is developed by Google to process large amount of data. The two main component of MapReduce are – Map and Reduce. During Map function, mapper retrieves input data that is stored in the files and divides the large data into small parts. After Map function is performed, data is redistributed based on the output of the map function so that data belonging to the same key is located on the same node. Then reduce method takes the output of the map method and applies shuffle and sort on the same. Main job of Reducer is to process the data that retrieves from Mapper. The final output of Reduce function is stored in HDFS. During MapReduce the Map and Reduce tasks are distributed to the appropriate servers. Each map function is independent of other map functions but they can be performed in parallel.



**Advantages of Hadoop:**

* Main advantage of Hadoop is it is used to store large amount of data in server that can work in parallel.
* As hadoop is based on distributed framework, so network overload can be prevented.
* The tasks can be performed independently.
* Replica is stored, so in case of failure of one datanode, another datanode can handle the operation.
* It is a simple programming model where developer needs to write simple mapreduce function.

References:

* <http://www.tutorialspoint.com/hadoop/>
* <http://www.plottingsuccess.com/hadoop-101-important-terms-explained-0314/>
* <http://www.aosabook.org/en/hdfs.html>
* <https://opensource.com/life/14/8/intro-apache-hadoop-big-data>
* <http://www.rohitmenon.com/index.php/introducing-hadoop-part-ii/>
* <https://wiki.apache.org/hadoop/JobTracker>
* <http://www.webopedia.com/TERM/H/hadoop_mapreduce.html>
* <http://www.vogella.com/tutorials/MapReduce/article.html>
* <http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm>
* <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html>
* <http://www.j2eebrain.com/java-J2ee-hadoop-advantages-and-disadvantages.html>