DSDBA

Sinhgad Institutes | Date | 1

Assignment-1 Part A

Title of Assignment :- Hadoop installation on a) Single Node b) Multiple Node

Objective :- To learn the installation steps of Hadoop on a Ubuntu platform

Prerequisite :- An Ubuntu 16.04 server with a non-root user with sudo privilege & knowledge of basic command in Ubuntu.

Theory Concept :-

1] What is Hadoop?

→ Hadoop is an opensource software framework that provides for processing of large data sets across clusters of computer using simple programming model.

2] Which are different layers of Hadoop?

→ Hadoop has following 4 layers :-

a) HDFS (Hadoop distributed file system)
b) YARN (Yet Another Resource Negotiator)
c) MAP REDUCE
d) HADOOP common

3] Explain the each layer of Hadoop in details.

→ Hadoop has following 4 layers :-

a) HDFS :- HDFS stands for Hadoop Distributed file system. It states that the file will be broken into blocks and stored in nodes over the distributed arch. It provides high-throughput access to appⁿ data.

(b) YARN :- YARN stands for Yet Another Resource Negotiator. It is used for Job Scheduling and managing the cluster.

(c) MAP REDUCE :- This is YARN based system for parallel processing of large data set using key value pair the map task takes input data & convert it into a dataset which can be computed in key value pair.

(d) Hadoop Common :- These Java libraries & utilities are used to start Hadoop. These are used by other Hadoop modules. These libraries provide file system & OS level abstraction.

4] Conclusion :-

In this assignment we learned how to install Hadoop on single node & multiple node on ubuntu platform.

Sinhgad Institutes | Date

Assignment 2 : PART A

Title of the Assignment :- Design a distributed application using Map Reduce which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log system from the Internet and process it using a pseudo distribution mode on Hadoop platform.

Objective of the Assignment :- To learn the concept of Map Reduce.

Prerequisite :- Java Programming.

Theory Content :-

1] What is Map Reduce paradigm?

→ MapReduce paradigm was created in 2003 to enable processing of large data sets in a massively parallel manner. The goal of MapReduce model is to simplify the approach to transformation and analysis of large datasets, as well as to allow developers focus on algorithms instead of data management. The MapReduce model consists of two phases: the map phase and the reduce phase, expressed by the map function and reduce function, respectively.

Map :- The map function, also referred to as a map task, processes a single key/value input pair and produces a set of intermediate key/value pairs.

Reduce :- The reduce function, also referred to as the reduce task, consists of taking all key/value pairs produced in the map phase that share the same intermediate key and producing zero, one, or more data items.

2] Discuss the main components of Map Reduce job.

→ The three main components of the MapReduce Job are the Jobtracker and Tasktracker.

Date

JobTracker :- It is the master that creates and runs the job in the MapReduce. It runs the job in the MapReduce. It runs on the name node and allocates the job to the task trackers.

TaskTracker :- It is the slave that runs on the data node. TaskTracker runs the job sent by the JobTracker and report the status of the task back to it. The TaskTracker will be assigned with the Mapper and Reducer tasks to execute the job.

3) Explain working of mapper, reducer, and driver.

→ The Hadoop java programs are consist of Mapper class and Reducer class along with the driver class.

Mapper class :- The first stage in data processing using MapReduce is the Mapper class. Here, RecordReader processes each input record and generates the respective key-value pair. Hadoop's mapper store saves this intermediate data into the local disk.

Reducer class :- The intermediate output generated from the mapper is fed to the reducer which processes it and generates the final output which is then saved in the HDFS.

Driver class :- The major component in a MapReduce job is a Driver class. It is responsible for setting up a MapReduce Job to run in Hadoop. We specify the names of Mapper and Reducer classes along with data types and their respective job names.

4) Explain the basic parameters of mapper and reducer function.

→ The four basic parameters of a mapper are as follows :-

a) LongWritable

b) text

c) text

d) IntWritable.

The first two represent input parameters and the second two represent intermediate output parameters.

Sinhgad Institutes | Date

The four basic parameters of a reducer are as follows:-

a) Text

b) IntWritable

c) Text

d) IntWritable.

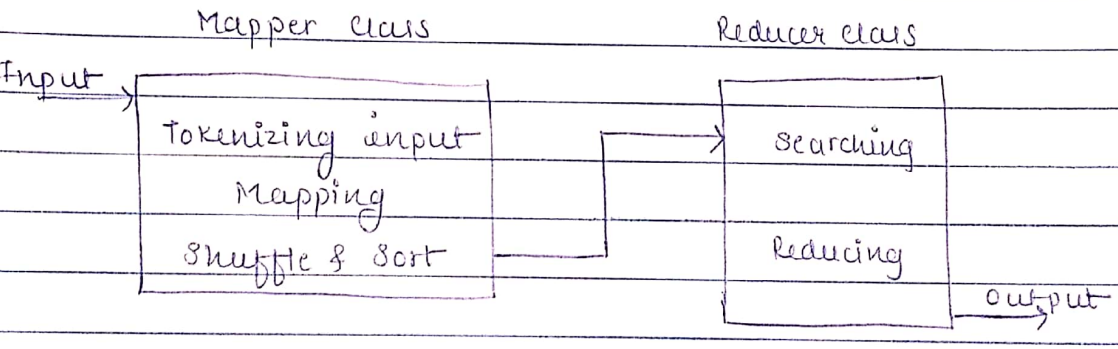
The first two represent intermediate output key, value parameters and the second two represent final output key, value parameters.

5] Algorithm:-

The MapReduce Algorithm contains two important tasks, namely Map and Reduce.

- The map task is done by means of Mapper class.
- The reduce task is done by means of the Reducer class.

Mapper class takes the input, tokenizes it, maps and sorts it. The output of Mapper class is used as input by Reducer class, which in turn searches matching pairs and reduces them.



MapReduce implements various mathematical algorithms to divide a task into small parts and assign them to multiple systems. In technical terms, MapReduce algorithm helps in sending the Map & Reduce tasks to appropriate servers in a cluster.

Conclusion:- Hence we conclude that by the end of this assignment we have studied the concept of Map Reduce.