MES College of Engineering Pune-01

Department of Computer Engineering

Name of Student:	**Class:**
Semester/Year:	**Roll No:**
Date of Performance:	**Date of Submission:**
Examined By:	**Experiment No: Part B-02**

PART: B) ASSIGNMENT NO: 02

AIM: Design a distributed application using MapReduce which processes a log file of a system.

OBJECTIVES:

- Students will be able to understand and use MapReduce framework.
- To learn the concept of how to Input data to HDFS and view resultant file in the output folder.

APPARATUS:

- Operating System recommended: 64-bit Open source Linux or its derivative.
- Front End: Java,Hadoop
- Dataset: Logfile.txt or Logfile.csv

PREREQUISITE:

Fundamentals of Hadoop commands and concept of MapReduce framework,HDFS.

THEORY:

- MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.
- MapReduce is a processing technique and a program model for distributed computing based on java
- 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) and reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.
- The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.
- The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable- Comparable interface to facilitate sorting by the framework.

- Input and Output types of a MapReduce job:

  (Input) <k1,v1> -> map -> <k2, v2>-> reduce -> <k3, v3> (Output).

CONCLUSION:

QUESTIONS:

1. Explain data structure of big data
2. Explain applications of Hadoop
3. Explain Hadoop Ecosystem