

ASSIGNMENT 2

Aim: Design a distributed application using Map Reduce

Problem Statement: 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 of the system. Use simple log file from the internet and process it using a pseudo distribution mode on hadoop platform.

Objectives: • To understand concept of MapReduce

- · To understand the details of Hadoop File System
- · To understand the technique for log file processing
- · Analyze the properformance of hadoop file system
- · To understand use of distributed processing

Theory: What is Mapkeduce?

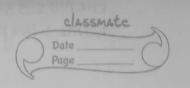
Mapkeduce is a processing technique and a program model for distributed computing based on Java. The Mapkeduce algo contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value) pairs). Secondly, a reduce task, which takes output from map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Mapkeduce implies, the reduce task is always performed after the map job.

The Algorithm ?

1) Crenerally MapReduce paradigm is based on the sending the computer to where the data resides.

2) MapReduce program executes in three stages, namely map stage and reduce stage

i) Map Stage - The map or mapper's job is to process the input data is in the form of file or



directory and is stored in HDFs. The input file is pasced to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

- Fi) 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 HDFS.
- 3) During a Map Reduce job, Hadoop sends the map and reduce tasks to the appropriate servers in the cluster
- The framework manages all the details of data-passing such as issuing tasks, verifying tasks of completion and copying data around the cluster between the nodes.
- 5) most of the computing takes places on nodes which with the data on local dicks that reduces network traffic.
- 6) After completion of the given task, the cluster collects and reduces the data to form an appropriate result and sends it back to the hadoop server.

Terminology:

- Degload Applications implement the map and reduce
- 2) Mapper It maips the input key / value pairs to a set of intermediate key / value pair
- 3) Named Node Node that marages the HDFS
- 4) Data Nøde Nøde where data is presented in advance before any processing takes place
- 5) MasterNode Node where Job Tracker runs and which accepts job requests from clients
- 6) staveNode Node where map and reduce program runs
- 7) Job Tracker Schedulu jobs and tracks the assigned jobs to Task Scheduler

8)	Task Tracker - Tracks the task and reports status to
9)	Job - A program & an execution of a mapper or a reducer on a slice of data
10)	Task Attempt - A particular instance of an attempt to execute a task on a Slave Node.
	conclusion: I understood the uses of distributed data processing using MapReduce.