Assignment 2

→ **Problem Statement :**

Design a distributed application using MapReduce which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet & process it using a pseudo distribution mode on Hadoop.

→ **Theory :**

A] **What is MapReduce ?**

- MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains 2 important tasks, namely Map & Reduce.
- Map takes a set of data & converts it into another set of data, where individual elements are broken down into tuples.
- The reduce task, takes the output from Map as input & combines those data tuples into smaller set of tuples.

- It is hence easier to scale data processing over multiple computing nodes.
- Scaling an application over hundreds of machines in a cluster is just a matter of configuration change.

B] The Algorithm:

- MapReduce executes in 3 stages.

a. Map Stage: Maps the data. The input file is passed line by line and the mapper processes the data & creates several small chunks of data.

b. Reduce Stage: Combination of Shuffle Stage & Reduce Stage. Processes data mapped by mapper & produces new output to be stored in HDFS.

- During MapReduce job, Hadoop sends Map & Reduce Tasks to appropriate servers in the cluster.

- The framework manages all the details of data-passing such as issuing tasks, verifying task(s) completion & copying data around the clusters.

- After Completion of task, cluster collects and reduces data to form an appropriate result, and sends it back to Hadoop Server.

c] Terminology :

- Payload : Applications implement Map & Reduce functions and form the core of the job.
- Mapper : Maps input key/value pair to a set of key/value pair.
- NameNode : Node that manages Hadoop Distributed File System. (HDFS)
- DataNode : Node where data is presented in advance before any processing takes place.
- MasterNode : Node where JobTracker runs & which accepts job requests from clients.
- SlaveNode : Node where Map & Reduce program runs.
- TaskTracker : Tracks task & reports status to Job Tracker.
- Job Tracker : Schedules Jobs & tracks the the assigned Jobs to task tracker.
- Job : A program in execution of Mapper & Reducer across dataset

- Task: An execution of a Mapper or a Reducer on a slice of data.
- Task Attempt: A particular instance of an attempt to execute a task on SlaveNode.

→ Conclusion :

Understand the uses of distributed data processing using MapReduce.