Lab Assignment - 2
(Grp - B)

Sinhgad Institutes

Date

Title :-
Design a distributed application using map-reduce which processes a log file of a system.

Objectives :-
Students should be able to perform the mapreduce operatn in Hadoop.

Prerequisite :-
1) Hadoop installatn
2) Basics of hadoop : HDFS, map reduce.
3) java programming.

Theory :-
MapReduce :-
MapReduce is a processing technique & a program model for distributed computing based on java. The mapReduce algorithm contains two important tasks, namely Map & Reduce. Map takes a set of data & converts it into another set of data. The major advantage of mapReduce is that it easy to scale data processing over multiple computing nodes. This simple scalability is what has attracted many programmers to use to mapReduce model.

Terminology:-

1) Payload:- Applications implement the map & reduce functions & form the core of job.

2) Mapper:- maps the input key/value pair.

3) Named Node:- manages Hadoop distributed file system.

4) DataNode:- data is represented in advanced.

5) SlaveNode:- where maps & reduce program runs.

6) Job Tracker:- schedules jobs & tracks the assign jobs to task tracker.

7) Job - A program is an execution of mapper & reducer across a dataset.

8) Task - execution of mapper or reducer on slice of data.

9) Task attempt:- particular instance of an attempt to execute a task on slavenode.

= Important commands:-

- Name-node format
- secondary name node
- datanode
- fsck.

- balancer
- FS
- Oiv
- job
- queue
- version
- jar <jar>
- classpath
- <parent path> <src>* <dest>

=conclusion:- In this manner we have successfuly designed a distributed applicatn using Map-Reduce which processes a log file of a system.