

Group B Assignment 2

AIM: Design a distributed application using Map-Reduce which processes a log file of a system.

THEORY:

What is MapReduce?

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). Secondly, 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.

1. Steps for Compilation & Execution of Program:

```
#sudo mkdir analyzelogs
ls
#sudo chmod -R 777 analyzelogs/
cd
ls
cd
..
pwd ls cd
pwd
#sudo chown -R hduser analyzelogs/
cd ls
#cd analyzelogs/
ls cd ..
```

2. Copy the Files (Mapper.java,Reduce.java,Driver.java to Analyzelogs Folder)

```
#sudo cp /home/isha/Desktop/count_logged_users/* -/analyzelogs/
```

Start HADOOP

```
#start-dfs.sh
```

```
#start-yarn.sh
```

```
#jps
```

3. Compile Java Files

```
# javac -d .
```

```
DsbdaMapper.java DsbdaReducer.java DsbdaDriver.java ls
```

```
#cd Dsbda/
```

```
ls cd .. #sudo gedit Manifest.txt
```

```
#jar -cfm analyzelogs.jar Manifest.txt Dsbda/*.class
```

```
ls cd jps
```

```
#cd analyzelogs/
```

4. Create Directory on Hadoop #sudo mkdir ~/input2000

```
ls pwd #sudo cp access_log_short.csv ~/input2000/
```

```
# $HADOOP_HOME/bin/hdfs dfs -put ~/input2000 /
```

```
# $HADOOP_HOME/bin/hadoop jar analyzelogs.jar /input2000 /output2000
```

```
# $HADOOP_HOME/bin/hdfs dfs -cat /output2000/part-00000
```

```
# stop-all.sh
```

```
# jps
```

Input:

Output:

CONCLUSION: Thus, we have learnt how to design a distributed application using MapReduce and process a log file of a system