- su - hduser
- sudo mkdir analyzelogs
- Right click on data folder open properties and copy the path

sudo cp /home/pranaypawar/Desktop/Data/* ~/analyzelogs

(change the path according to your path)

sudo chmod -R 777 analyzelogs

sudo chown -R hduser analyzelogs

cd analyzelogs

sudo chmod +r *.*

- paste the below line in terminal

export CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.9.0.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-common-2.9.0.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-2.9.0.jar:~/analyzelogs/SalesCountry/*:$HADOOP_HOME/lib/*"

javac -d . SalesMapper.java SalesCountryReducer.java SalesCountryDriver.java

sudo nano Manifest.txt

- copy paste this line

Main-Class: SalesCountry.SalesCountryDriver

- after copy pasting do ctrl + x, Y , Enter

jar -cfm analyzelogs.jar Manifest.txt SalesCountry/*.class

cd

start-dfs.sh

start-yarn.sh

jps

- check hadoop running properly)

cd analyzelogs

sudo mkdir ~/input

sudo cp access_log_short.csv ~/input/

$HADOOP_HOME/bin/hdfs dfs -put ~/input /

$HADOOP_HOME/bin/hadoop jar analyzelogs.jar /input /output

$HADOOP_HOME/bin/hdfs dfs -cat /output/part-00000

- this gives mapreduce outputjps

stop-dfs.sh

Stop-yarn.sh

What is map reduce ?
MapReduce performs the processing of large data sets in a distributed and parallel manner.
MapReduce consists of two distinct tasks-Map and Reduce.
Two essential daemons of MapReduce: Job Tracker & Task Tracker.

Map reduce performs the processing of large data sets in a distributed and parallel manner

Mappper is used divided and conquer split data into key pair value

The reduce :- this job takes the output of previously executed map task as input and combines those data tuples into a smaller set of tuples

advantage

Scalable

Simple coding model

Support unstructured data

Simple to scale data processing over multiple computing model