## **Assignment No. 1**

Title: Single node/Multiple node Hadoop Installation.

Execution: <https://drive.google.com/file/d/1qGymlkdCyQ07bOPEZk_ETb_GYn8Qq6jb/view?usp=share_link>

## 

## **Assignment No. 2**

Title: Design a distributed application using MapReduce(Using Java) which processes a log file of a system. List out the users who have logged for a maximum period on the system. Use a simple log file from the Internet and process it using a pseudo distribution mode on the Hadoop platform.

**Implementation Steps:**

1.Create folder Data on Desktop and copy all given files (i. access\_log\_short.txt – Data set to analyze

ii.SalesCountryDriver.java : Manages input and output from Mapper and Reducer

iii.SalesMapper.java : Reads input file

iv.SalesCountryReducer.java : Reads the input from mapper file)

2.Open file named access\_log\_short.txt using libreoffice/wps and use separator as “-” (Uncheck remaining options), save the file as text CSV format

3.su – hduser (Change the user)

4.Sudo mkdir analyzelogs (Create folder analyzelogs)

and copy the contents of Data folder into analyzelogs folder using command

5.sudo cp /home/ll01/Desktop/Data/\* ~/analyzelogs

6.Give permissions using

sudo chmod -R 777 analyzelogs /

7.Change the owner to hduser using following sudo chown -R hduser analyzelogs /

8.cd analyzelogs

9.sudo chmod +r \*.\*

10.export CLASSPATH="$HADOOP\_HOME/share/hadoop/mapreduce/hadoop- mapreduce-clientcore-2.9.0.jar:$HADOOP\_HOME/share/hadoop/mapreduce/hadoop- mapreduce-client-common2.9.0.jar:$HADOOP\_HOME/share/hadoop/common/hadoop- common2.9.0.jar:~/analyzelogs/SalesCountry/\*:$HADOOP\_HOME/lib/\*"

(above command is available in file: hadoop-mapreduce-example-file.txt)

11.sudo gedit Manifest.txt

12.Add following line and press enter and save the file

Main-Class: SalesCountry.SalesCountryDriver

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

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

15.start-dfs.sh

16.Start-yarn.sh

17.jps

18.cd analyzelogs/

19.sudo mkdir ~/input

20.sudo cp access\_log\_short.csv ~/input/

## 

## **Assignment No. 3**

**Title:** Write an application using HiveQL for flight information system which will include

a. Creating, Dropping, and altering Database tables.

b. Creating an external Hive table.

c. Load table with data, insert new values and field in the table, Join tables with Hive

d. Create index on Flight Information Table

e. Find the average departure delay per day in 2008.

**Execution:**

**Steps for HIVE Installation:**

##Download Apache Hive from this link

##https://dlcdn.apache.org/hive/

##Download version 2.x only, as version 3.x has problems with java

##I have downloaded apache-hive-2.3.9-bin.tar.gz

##After downloading extract the file, right click on extracted folder , go to properties and copy the path

##Use the below command to move the folder into home of hduser

sudo mv /home/pranaypawar/Downloads/apache-hive-2.3.9-bin /home/hduser

##Start hadoop using

start-dfs.sh

start-yarn.sh

##check all nodes are running with jps

jps

##if all ok continue with next step

##Create Hive directories within HDFS. The directory ‘warehouse’ is the location to store the table or data related to hive.

hdfs dfs -mkdir -p /user/hive/warehouse

hdfs dfs -mkdir /tmp

##Set read/write permissions for table.

hdfs dfs -chmod g+w /user/hive/warehouse

hdfs dfs -chmod g+w /tmp

##Give all permissions to apache-hive-2.3.9-bin folder

sudo chmod a+rwx apache-hive-2.3.9-bin/

sudo chmod -R 777 apache-hive-2.3.9-bin/

sudo chown -R hduser apache-hive-2.3.9-bin/

##add hive path in bashrc using below command

sudo nano ~/.bashrc

##go to the last line and add these lines in bashrc file

export HIVE\_HOME=/home/hduser/apache-hive-2.3.9-bin

export PATH=$PATH:$HIVE\_HOME/bin

##Do ctrl+x, hit Y and then Enter. this will save the file

source ~/.bashrc

cd $HIVE\_HOME

cd conf/

sudo nano hive-env.sh.template

##add the following lines at the end of the file

export HADOOP\_HOME=/usr/local/hadoop

HIVE\_CONF\_DIR="${HIVE\_CONF\_DIR:-$HIVE\_HOME/conf}"

export HADOOP\_HEAPSIZE=${HADOOP\_HEAPSIZE:-256}

export HIVE\_CONF\_DIR=$HIVE\_CONF\_DIR

export HIVE\_AUX\_JARS\_PATH=$HIVE\_AUX\_JARS\_PATH

##do ctrl+x, hit y , rename the file to hive-env.sh, press enter and then y

##this will save the file hive-env.sh.template to hive-env.sh

cd ..

cd bin

schematool -initSchema -dbType derby

##after using above command if you get something like this

//

Starting metastore schema initialization to 2.3.0

Initialization script hive-schema-2.3.0.derby.sql

Initialization script completed

schemaTool completed

//

##metastore\_db has been created succesfully then

hive

##the above command will start hive

**HIVE DATABASE COMMANDS:**

# Create Database

CREATE DATABASE databasename;

#Use database

use databasename;

# Create table

CREATE TABLE table\_name (

column1 datatype,

column2 datatype,

column3 datatype,

....

);

eg: CREATE TABLE Persons (

PersonID int,

LastName varchar(255),

FirstName varchar(255),

Address varchar(255),

City varchar(255)

);

#Alter table

ALTER TABLE table\_name

ADD column\_name datatype;

Eg: ALTER TABLE Customers

ADD Email varchar(255);

ALTER TABLE table\_name

DROP COLUMN column\_name;

Eg:ALTER TABLE Customers

DROP COLUMN Email;

ALTER TABLE table\_name

ALTER COLUMN column\_name datatype;

Eg: ALTER TABLE table\_name

MODIFY COLUMN column\_name datatype;

#index

CREATE INDEX index\_name

ON table\_name (column1, column2, ...);

Eg:

CREATE INDEX idx\_pname

ON Persons (LastName, FirstName);

<https://drive.google.com/file/d/1gnY9C6Up06AAV4DGhDK8WYuUJJIEOxpe/view?usp=share_link>

## 

## **Assignment No. 4**

**Title:** Perform the following operations using Python on the Facebook metrics data sets

a. Create data subsets

b. Merge Data

c. Sort Data

d. Transposing Data

e. Shape and reshape Data

**Execution:** <https://colab.research.google.com/drive/1ibPPVD8DiuXFn99N__i8h8-2VL5aUSUN?usp=sharing>

## **Assignment No. 5**

**Title:** Perform the following operations using Python on the Air quality and Heart Diseases data sets

a. Data cleaning

b. Data integration

c. Data transformation

d. Error correcting

e. Data model building

Execution: <https://colab.research.google.com/drive/10A42LPLQwdCh7v-cEt_anGOrhGwpY-kH?usp=sharing>

Execution: <https://colab.research.google.com/drive/1L-jHT6jGjS6bzzeCy7NHt7sBUK-wZqcs?usp=sharing>

## **Assignment No. 6**

Visualize the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment no. 4 and 5

Execution:

<https://colab.research.google.com/drive/1L-jHT6jGjS6bzzeCy7NHt7sBUK-wZqcs?usp=sharing>

## **Assignment No. 7**

Perform the following data visualization operations using Tableau on Adult and Iris datasets.

a. 1D (Linear) Data visualization b. 2D (Planar) Data Visualization

c. 3D (Volumetric) Data Visualization d. Temporal Data Visualization

e. Multidimensional Data Visualization f. Tree/ Hierarchical Data visualization

g. Network Data visualization

Execution: <https://drive.google.com/file/d/1gdlCp7KD7fjzrfPi7w9E0sZC1V9KL19C/view?usp=share_link>

## **Assignment No. 8**

Create a review scrapper for any ecommerce website to fetch real time comments, reviews, ratings, comment tags, customer name using Python

Execution:

<https://drive.google.com/file/d/10oNBRvKQk84tTdS1iNMsrRB24GdPMIOg/view?usp=share_link>