**HDFS Assignments**

1. Create a text file **userdata.txt** in your home directory having following lines using a text editor:

I am working on a Linux Server.

I am learning to work on Hadoop ecosystem.

1. Create a folder called **training** under your home directory in HDFS

Write the command used here:

1. Create a folder called **SampleData** under **training**

Write the command used here:

1. Copy file **userdata.txt** to **SampleData** folder in HDFS

Write the command used here:

1. Display the content of **userdata.txt** in **hdfs** using **cat** command

Write the command used here:

1. Create another directory called **SampleDataBak** under training directory in HDFS

Write the command used here:

1. Copy the file **userdata.txt** from **SampleData** folder to **SampleDataBak**:

Write the command used here:

1. List all sub-folders and files under **training** folder in **HDFS**:

Write the command used here:

1. Display the total disk space used by **training** directory

Write the command used here:

1. Display the disk space used by each folder under **training** directory

Write the command used here:

1. Change the Replication factor of **training/SampleData/userdata.txt** to 2

Write the command used here:

1. Browse the **training** folder in HDFS using Web UI (localhost:50070) and note down the following:

**File / Folder Name Size Replication Block Size**

training

userdata.txt (under SampleData)

userdata.txt (under SampleDataBak)

1. Delete training folder along with all sub folders and files:

Write the command used here:

**Hive Assignments**

**Input Data**

**Employee.txt** - has the following columns - ***EmpID, Name, Band, DepartmentID, Salary***

**Employee.txt**

A1001,Ramesh,B1,IT,40000

A1002,Ganesh,B2,HR,35000

A1003,Latha,B1,HR,30000

A1008,Shirish,B2, IT,55000

A1009,Shibu,B2, MKTG,48000

**EmpProj.txt** - has following columns – ***EmpID, projectID, year-week, EffortHrs***

**EmpProj.txt**

A1001,IDW,201601,40

A1002,IDW,201601,45

A1003,GDW,201601,25

A1008,IDW,201601,35

A1009,GDW,201601,50

A1001,IDW,201602,45

A1002,IDW,201602,48

A1003,GDW,201602,45

A1008,IDW,201603,50

A1009,GDW,201602,40

A1001,IDW,201602,45

A1002,IDW,201602,49

A1003,GDW,201602,46

A1009,GDW,201602,45

A1002,CDW,201603,50

A1002,CDW,201604,50

**Department.txt** – has following columns – ***DepartmentID, Department Name***

**Department.txt**

IT,Information Technology

HR,Human Resources

MKTG, Marketing

**Assignment 1: Create a Managed internal table**

1. Connect to Hive CLI
2. Create a database called **projectdb**
3. Create a table **employee** to store data in **employee.txt** under **projectdb**
4. List the structure of **employee** table

Write the command used here:

1. List all storage parameters of **employee** table using describe command.

Write the command used here:

1. List data in **employee** table using select statement (make sure that the output has column header s)

Write the commands used here:

**Assignment 2: Create External Table and execute Join query**

1. Copy file **Department.txt** to folder **dept** under your home directory in HDFS
2. Create an external Hive table **department** to read data from **dept** folder in hdfs
3. Write a join query to join **department** and **employee** tables and get the following output:

Department Name, Total Salary

1. How many mappers and reducers are executed in the map reduce job executed by Hive?

Number of Mappers:

Number of reducers:

1. Display the explain plan for the join query using **explain** statement

**Assignment 3: Create a Partitioned Table and load data**

1. Create a table **project\_details** partitioned by **project ID** and having following columns:

*EmpID, year-week, EffortHrs*

1. Load data from file empProj.txt into table **project\_details**
2. Write a select query to get total effort spent by empID = A1002 by project name.

Required Output columns: *projectID, empID, empName, totalEffort*