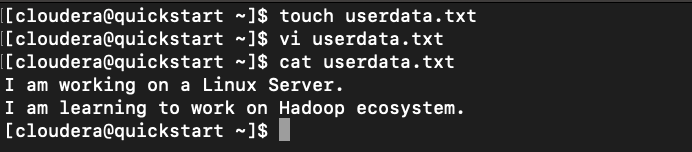
**HDFS Assignments**

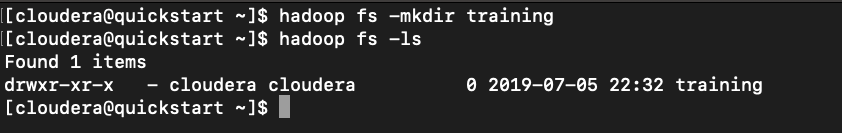
a. Create a text file **userdata.txt** in your home directory having following lines using a text editor:

**I am working on a Linux Server.**

**I am learning to work on Hadoop ecosystem**

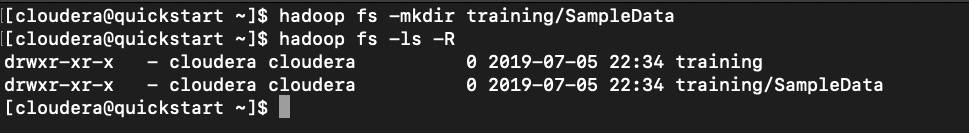
b. Create a folder called **training** under your home directory in HDFS

Write the command used here:

**hadoop fs -mkdir training**

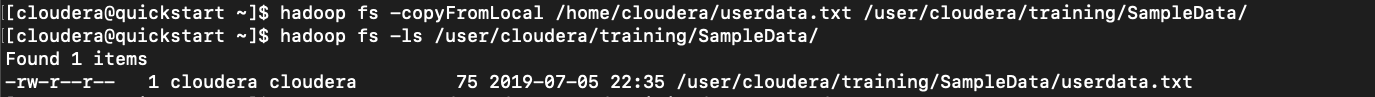
c. Create a folder called **SampleData** under **training**

Write the command used here:

**hadoop fs -mkdir training/SampleData**

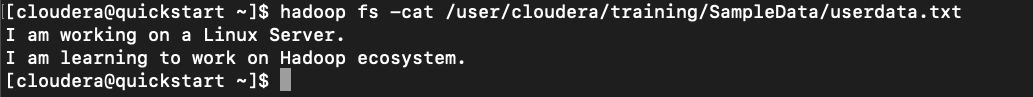
d. Copy file **userdata.txt** to **SampleData** folder in HDFS

Write the command used here:

**hadoop fs -copyFromLocal /home/cloudera/userdata.txt /user/cloudera/training/SampleData/**

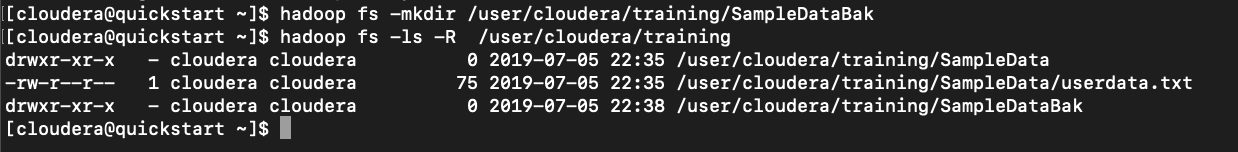
e. Display the content of **userdata.txt** in **hdfs** using **cat** command

Write the command used here:

**hadoop fs -cat /user/cloudera/training/SampleData/userdata.txt**

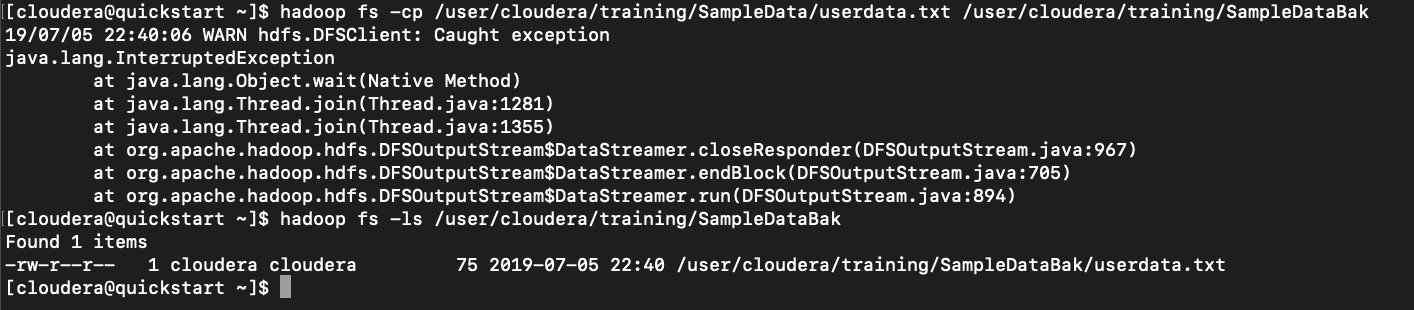
f. Create another directory called **SampleDataBak** under training directory in HDFS

Write the command used here:

**hadoop fs -p -mkdir /user/cloudera/training/SampleDataBak**

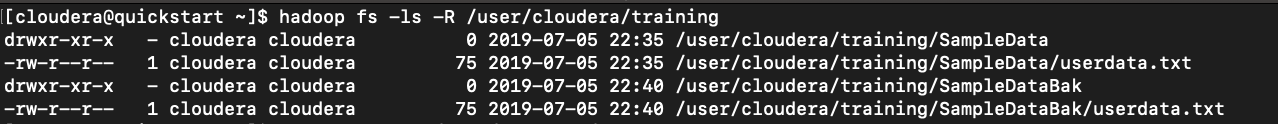
g. Copy the file **userdata.txt** from **SampleData** folder to **SampleDataBak**:

Write the command used here:

**hadoop fs -cp /user/cloudera/training/SampleData/userdata.txt /user/cloudera/training/SampleDataBak**

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

Write the command used here:

h**adoop fs -ls -R /user/cloudera/training**

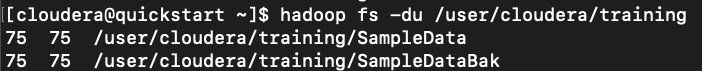
i. Display the total disk space used by **training** directory

Write the command used here:

**hadoop fs -du /user/cloudera/**

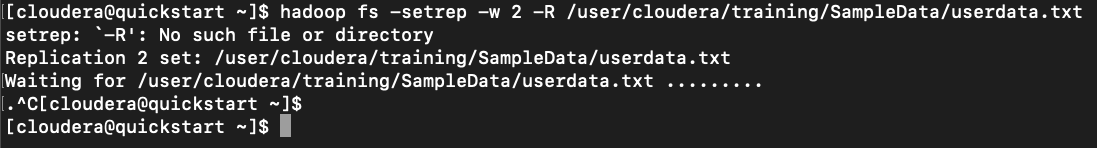
j. Display the disk space used by each folder under **training** directory

Write the command used here:

**hadoop fs -du /user/cloudera/training**

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

Write the command used here:

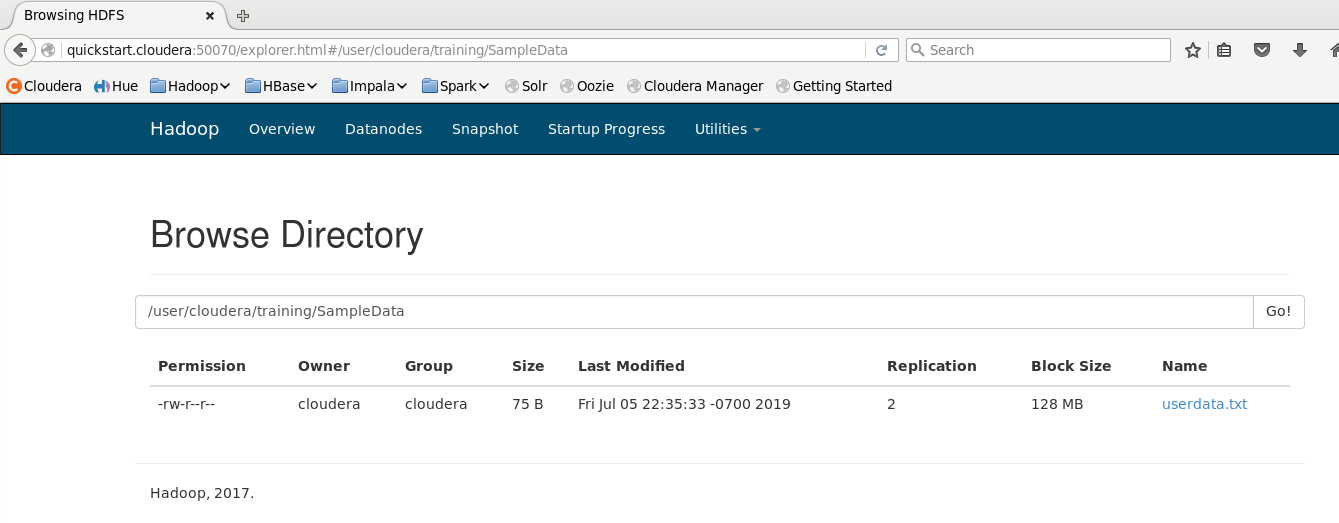
**hadoop fs -setrep -w 2 -R /user/cloudera/training/SampleData/userdata.txt**

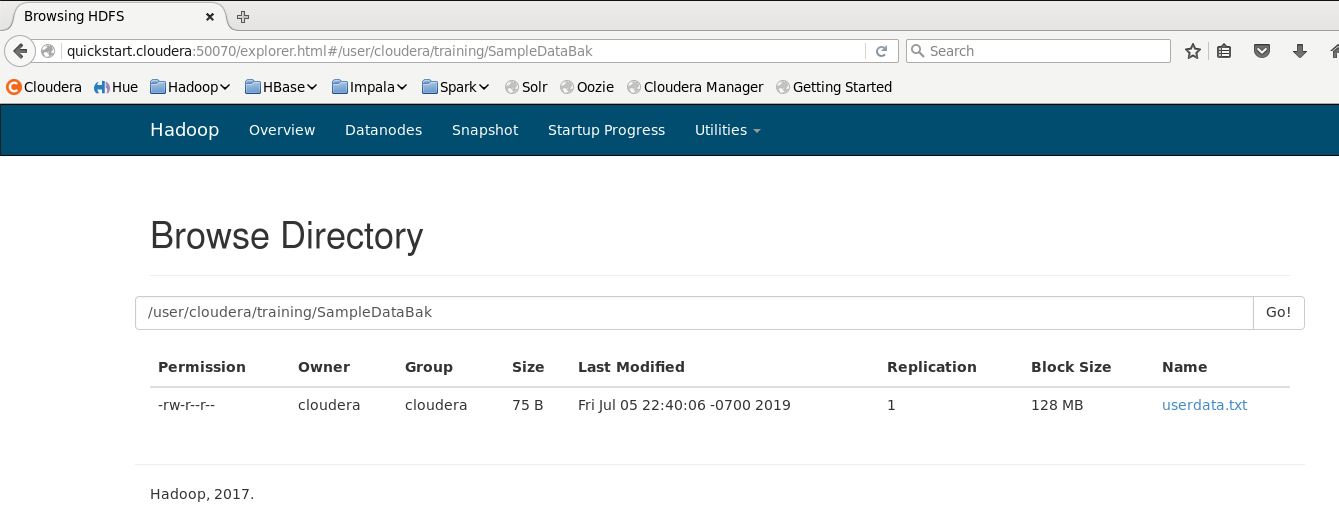
l. Browse the **training** folder in HDFS using Web UI (localhost:50070) and note down the following:

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

training 0 B. 0. 0

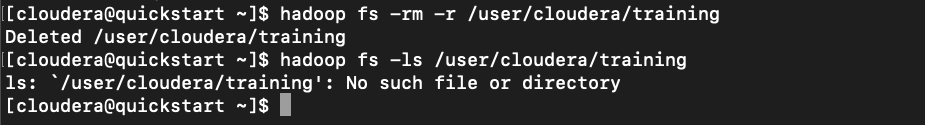
userdata.txt (under SampleData) 75B. 2. 128MB

userdata.txt (under SampleDataBak) 75B. 1. 128MB



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

Write the command used here:

h**adoop fs -rm -r /user/cloudera/training**

**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**

a. Connect to Hive CLI

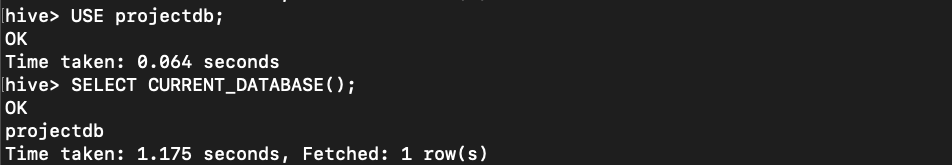
**hive**

b. Create a database called **projectdb**

**CREATE DATABASE projectdb;**



c. Create a table **employee** to store data in **employee.txt** under **projectdb**



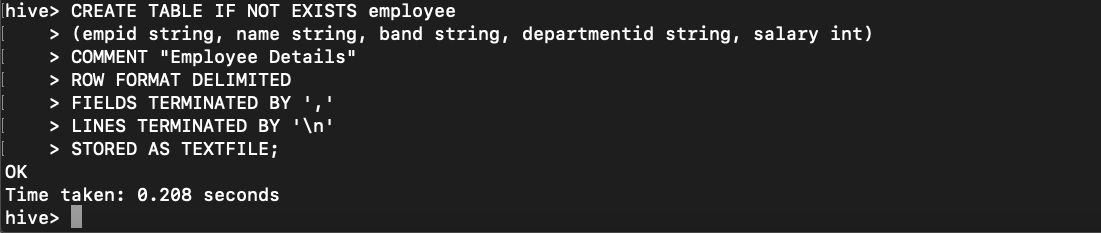
**CREATE TABLE IF NOT EXISTS employee (EmpID string, Name string, Band string, DepartmentID string, Salary int)**

**COMMENT “Employee Details”**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY “,”**

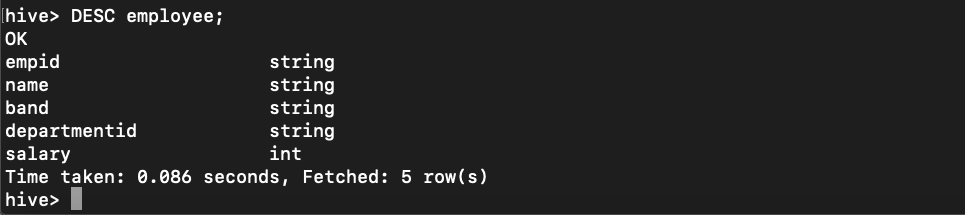
**LINES TERMINATED BY “\n”**

**STORED AS TEXTFILE;**

d. List the structure of **employee** table

Write the command used here:

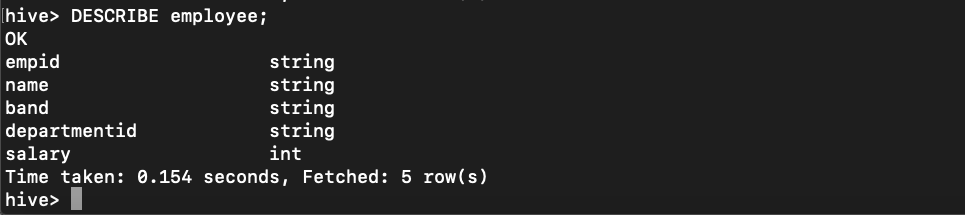
**DESC employee;**

**DESCRIBE employee;**

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

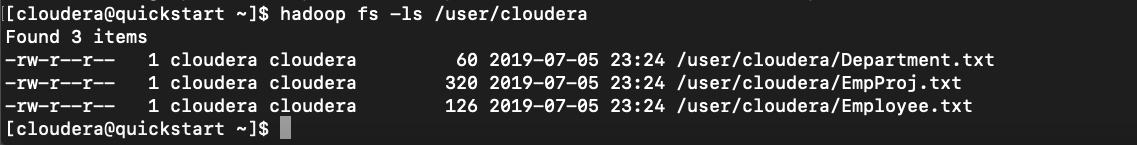
Write the command used here:

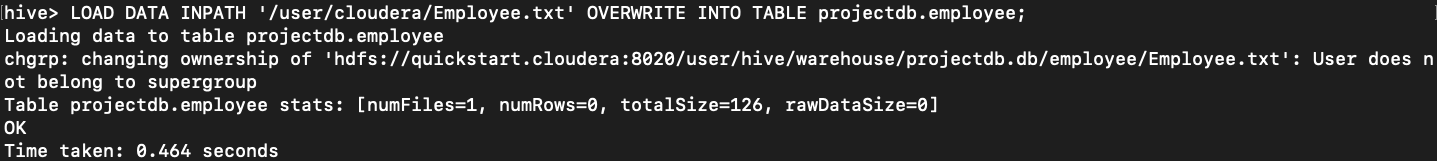
**DESCRIBE employee;**



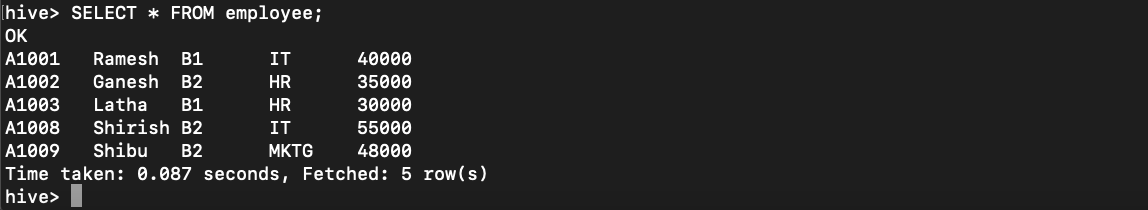
f. List data in **employee** table using select statement (make sure that the output has column header s)

Write the commands used here:



**LOAD DATA INPATH '/user/cloudera/Employee.txt' OVERWRITE INTO TABLE projectdb.employee;**

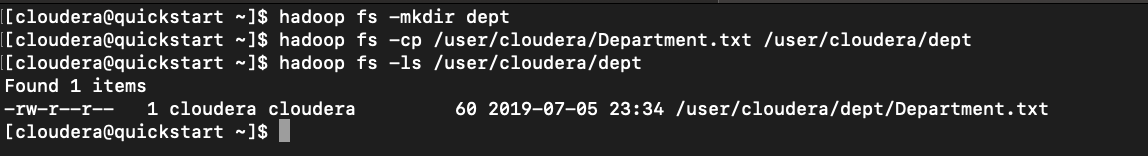
**set hive.cli.print.headers=true**

**SELECT \* FROM employee;**

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

a. Copy file **Department.txt** to folder **dept** under your home directory in HDFS

**hadoop fs -mkdir dept**

**hadoop fs -cp /user/cloudera/Department.txt /user/cloudera/dept**

b. Create an external Hive table **department** to read data from **dept** folder in hdfs

hive> CREATE EXTERNAL TABLE department

> (departmentid string, departmentname string)

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY ","

> LINES TERMINATED BY "\n"

> STORED AS TEXTFILE;

LOAD DATA INPATH '/user/cloudera/dept/Department.txt' OVERWRITE INTO TABLE projectdb.department;

c. Write a join query to join **department** and **employee** tables and get the following output:

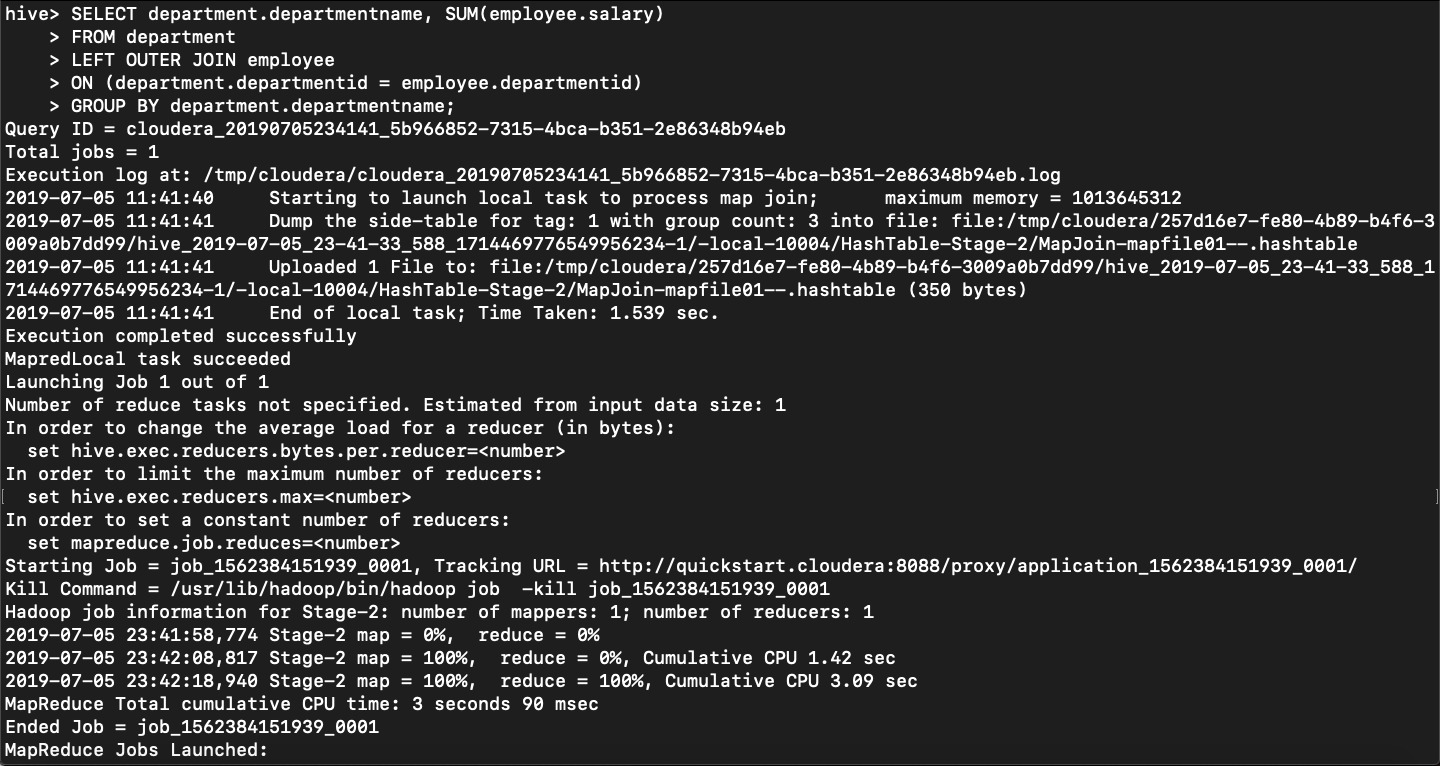
Department Name, Total Salary

hive> SELECT department.departmentname, SUM(employee.salary)

> FROM department

> LEFT OUTER JOIN employee

> ON (department.departmentID = employee.departmentid)

> GROUP BY department.departmentname;



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

Number of Mappers: 1

Number of reducers: 1

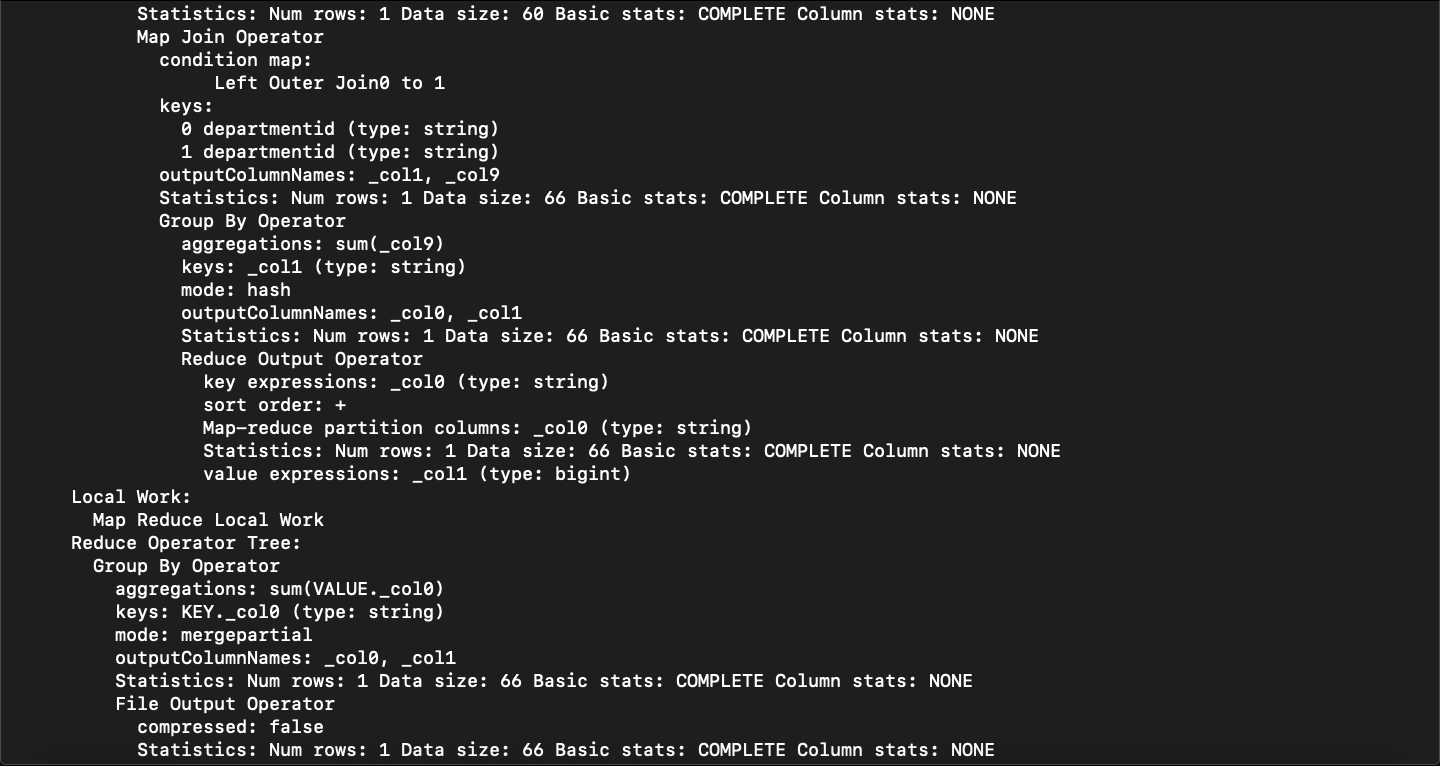
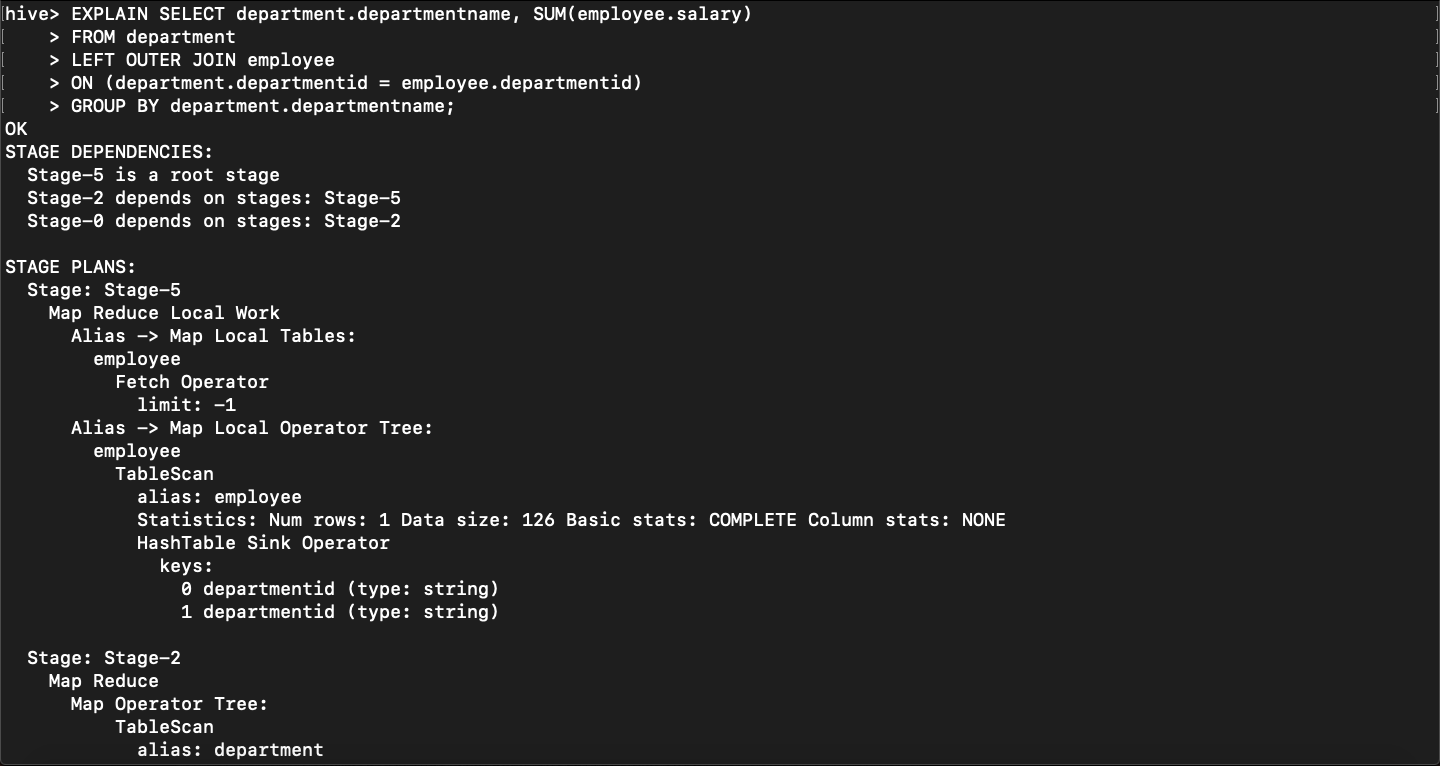
e. Display the explain plan for the join query using **explain** statement

hive> EXPLAIN SELECT department.departmentname, SUM(employee.salary)

> FROM department

> LEFT OUTER JOIN employee

> ON (department.departmentID = employee.departmentid)

> GROUP BY department.departmentname;



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

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

*EmpID, year-week, EffortHrs*

hive> CREATE TABLE project

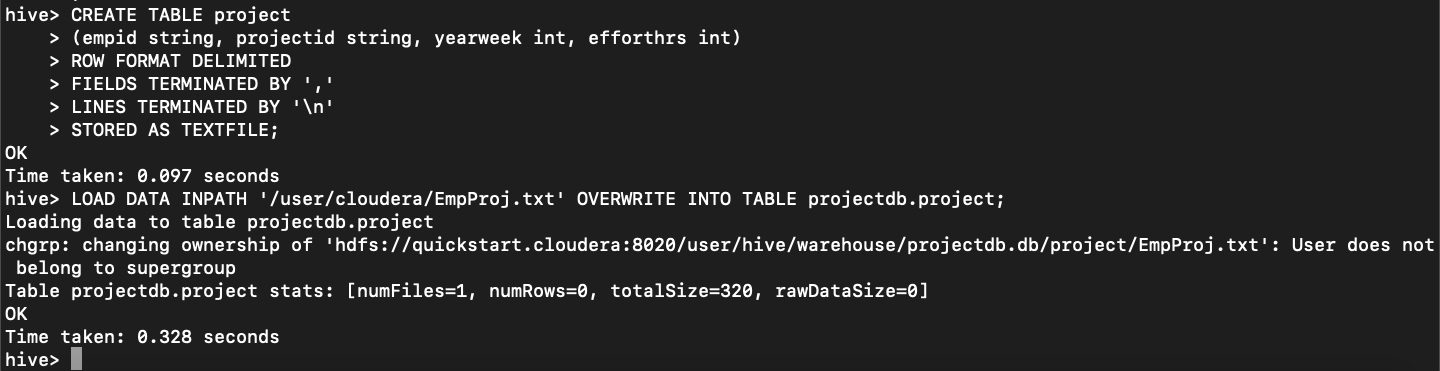
> (empid string,projectid string,yearweek int,efforthrs int)

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY ","

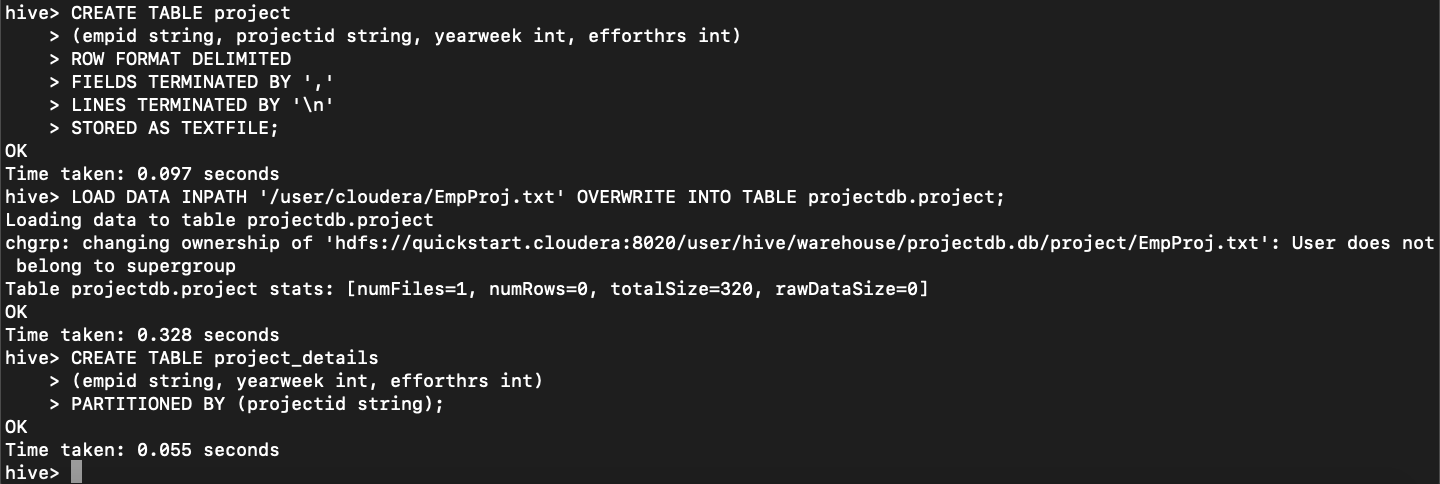
> LINES TERMINATED BY "\n"

> STORED AS TEXTFILE;

LOAD DATA INPATH '/user/cloudera/EmpProj.txt' OVERWRITE INTO TABLE projectdb.project;

hive> CREATE TABLE project\_details

> (empid string, yearweek int,efforthrs int)

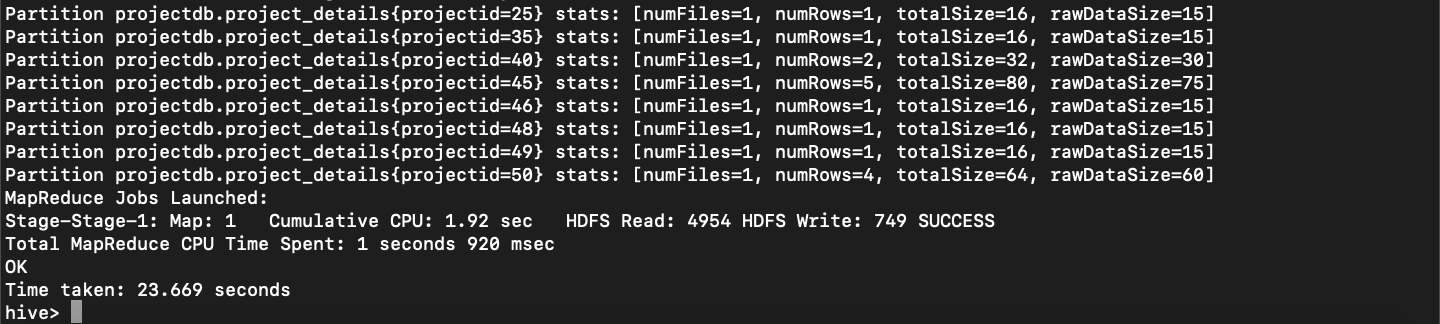
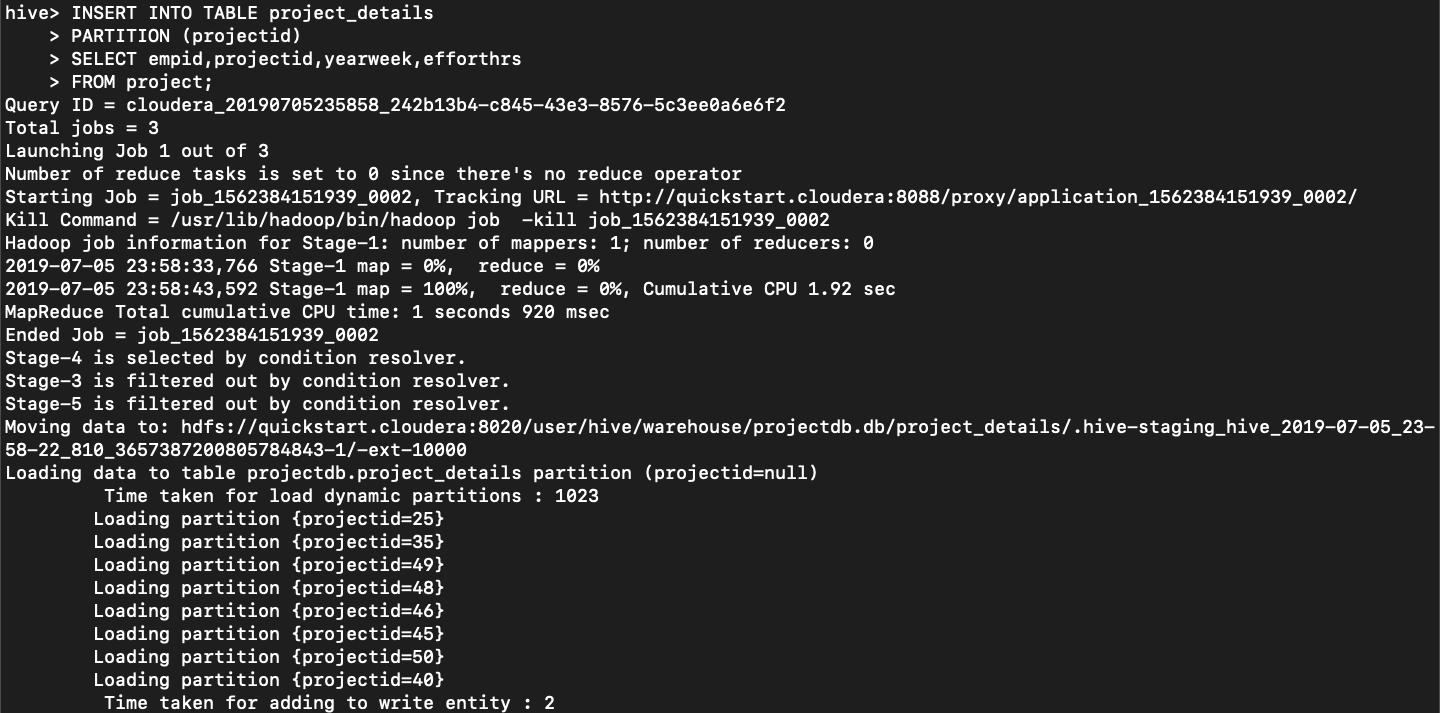
> PARTITIONED BY (projectid string);

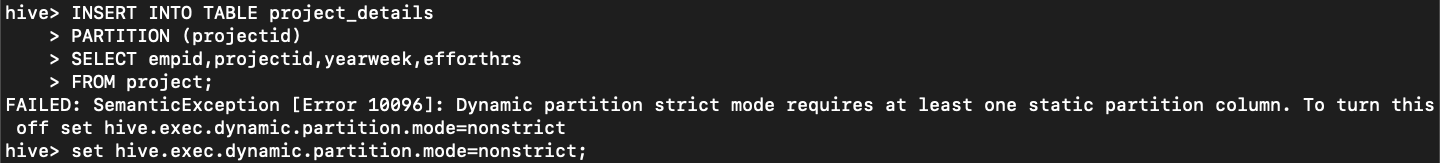
b. Load data from file empProj.txt into table **project\_details**

hive> INSERT INTO TABLE project\_details

> PARTITION (projectid)

> SELECT empid, projectid,yearweek,efforthrs

> FROM project;



c. Write a select query to get total effort spent by empID = A1002 by project name.

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

hive> SELECT project.projectid, employee.empid, SUM(project.efforthrs)

> FROM project

> LEFT OUTER JOIN employee

> ON project.empid = employee.empid

> WHERE employee.empid= 'A1002'

> GROUP BY project.projectid, employee.empid;

