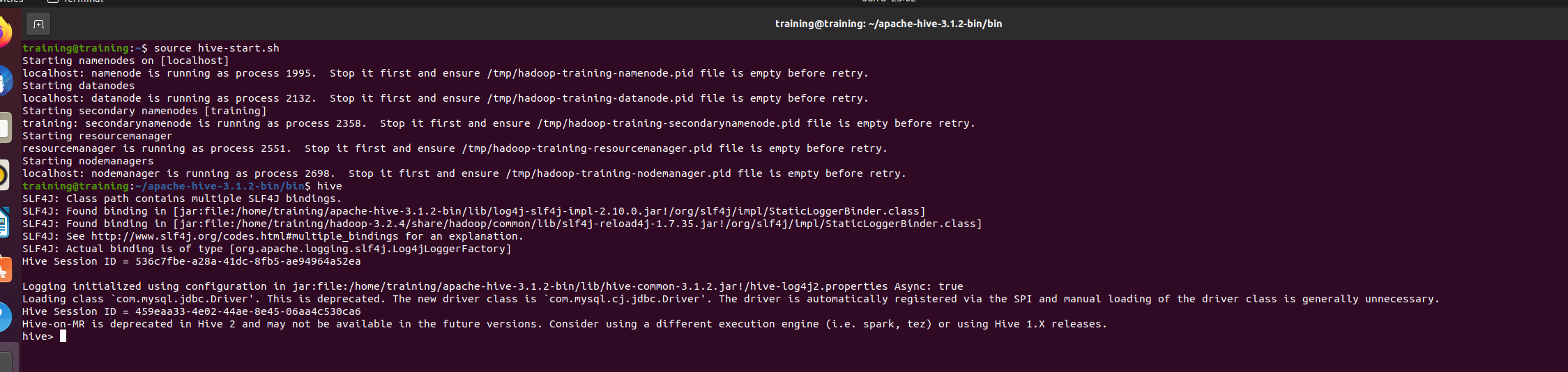
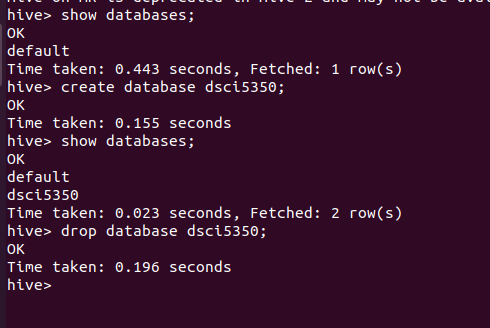
DSCI5350 Hive Activity 1&2 Budda satyasriharsha

Hive Hands-on Activity 1

Starting the Hive Shell  
Creating Database:  
show databases;

Create a database named DSCI5350  
create database dsci5350;  
Drop the database DSCI5350  
drop database dsci5350

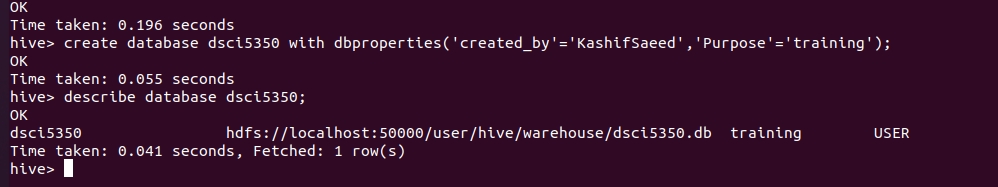


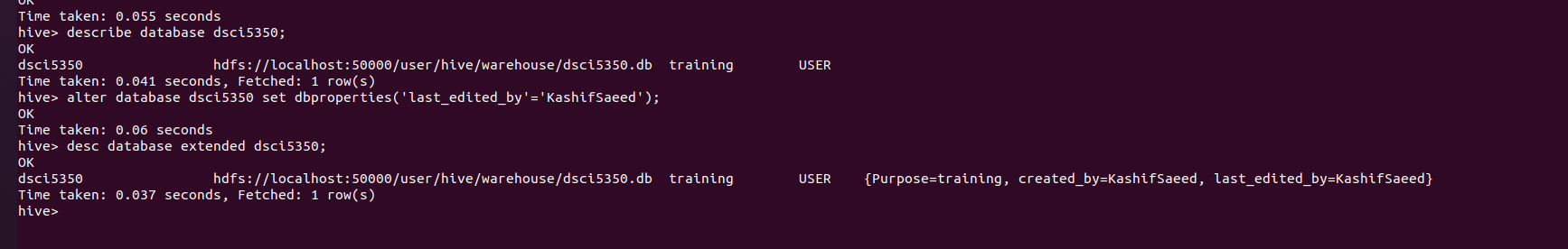


create database dsci5350 with  
dbproperties(‘created\_by’=’KashifSaeed’,’Purpose’=’Training’);

describe database dsci5350;

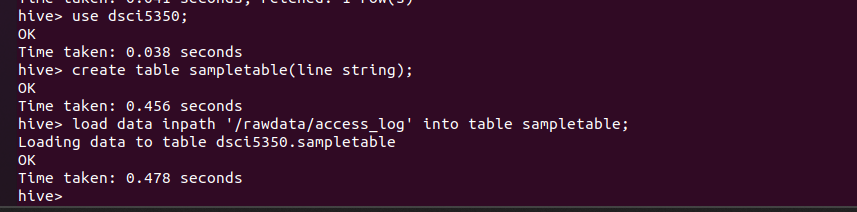
alter database dsci5350 set dbproperties(‘last\_edited\_by’=’KashifSaeed’);  
desc database extended dsci5350;



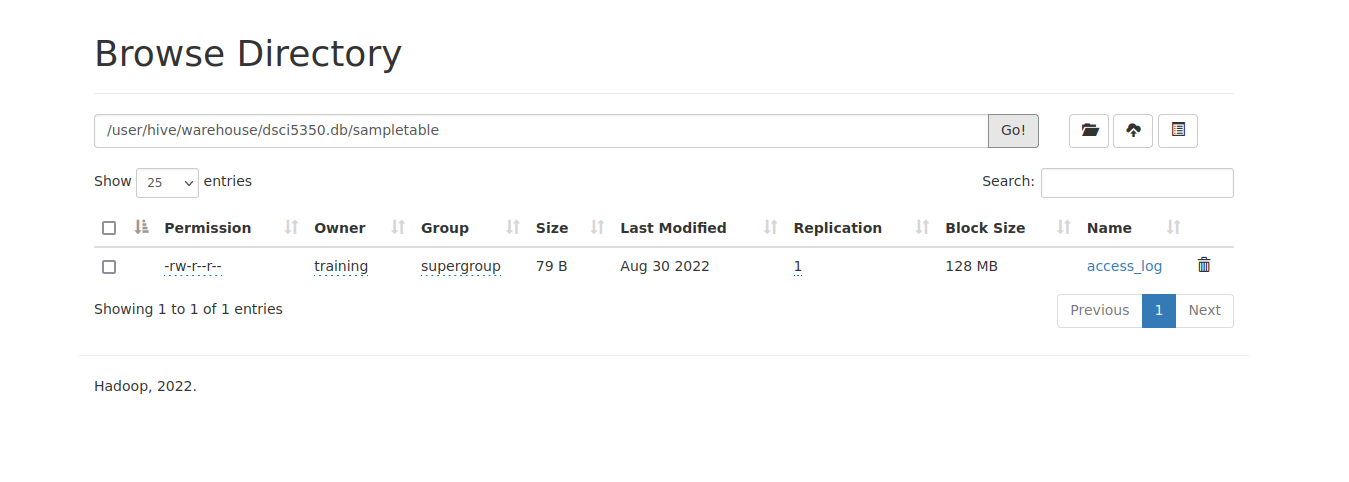


Exercise 1: Create a database with your EUID and set some properties: Created by you, for  
the purpose of performing activities



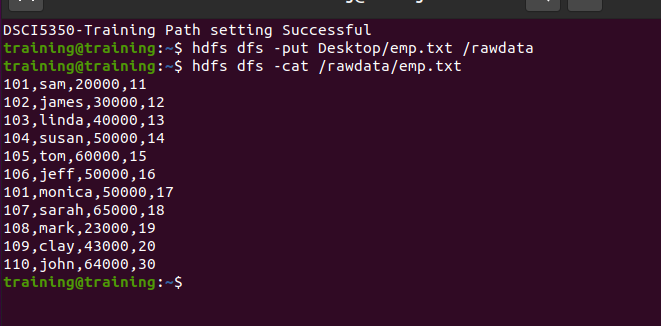
Creating Table  
reate table sampletable(line string);  
Load data inpath ‘/rawdata/access\_log’ into table sampletable  


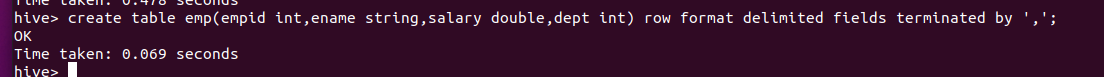
Now, go to the HDFS portal to validate that the database and the table is stored in Hive as  
directories and files.

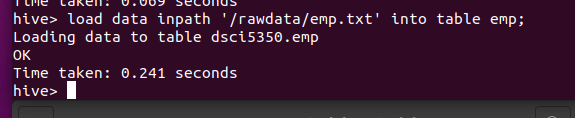


Creating a table with multiple columns- Loading Data from HDFS to Hive

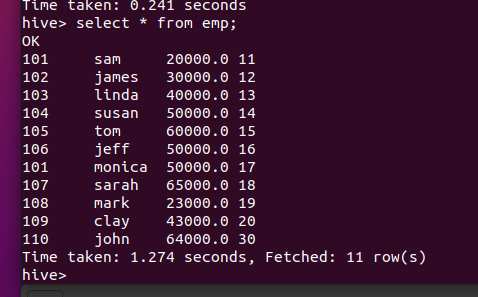
1. Copy emp.txt file to your Desktop VM (if you are unable to copy it, try logging to  
Canvas via the VM)  
2. Load the file into HDFS rawdata directory  
3. View the content of the file using the cat command

  
4. Using the information furnished in the previous step about the content of the file,  
create a table in Hive:

  
5. Next, load data from the /rawdata/emp.txt into the emp table using the command:

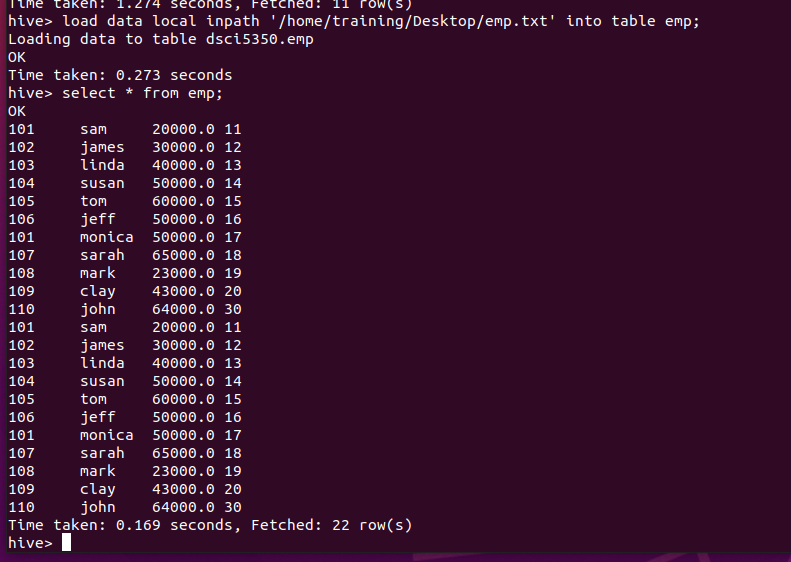


6.Run a query against the emp table to validate if the data got loaded



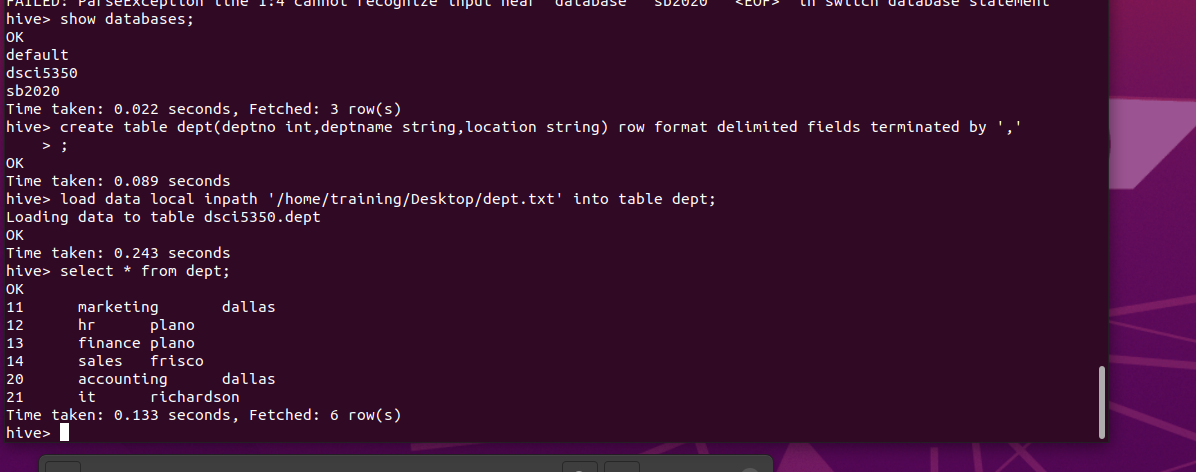
Creating a Table with multiple columns – loading data from local filesystem to Hive

1. Load emp.txt file from your local filesystem Desktop into emp table. Make sure you  
are using the hive shell and using dsci5350 database. The command for loading from  
the local filesystem is load data local inpath  
2. View the content of the table using a query.

  
3. Next, instead of appending the records as done in the previous step, overwrite the  
records using the following command:  
4. Validate the data in the table to ensure that it was overwritten.

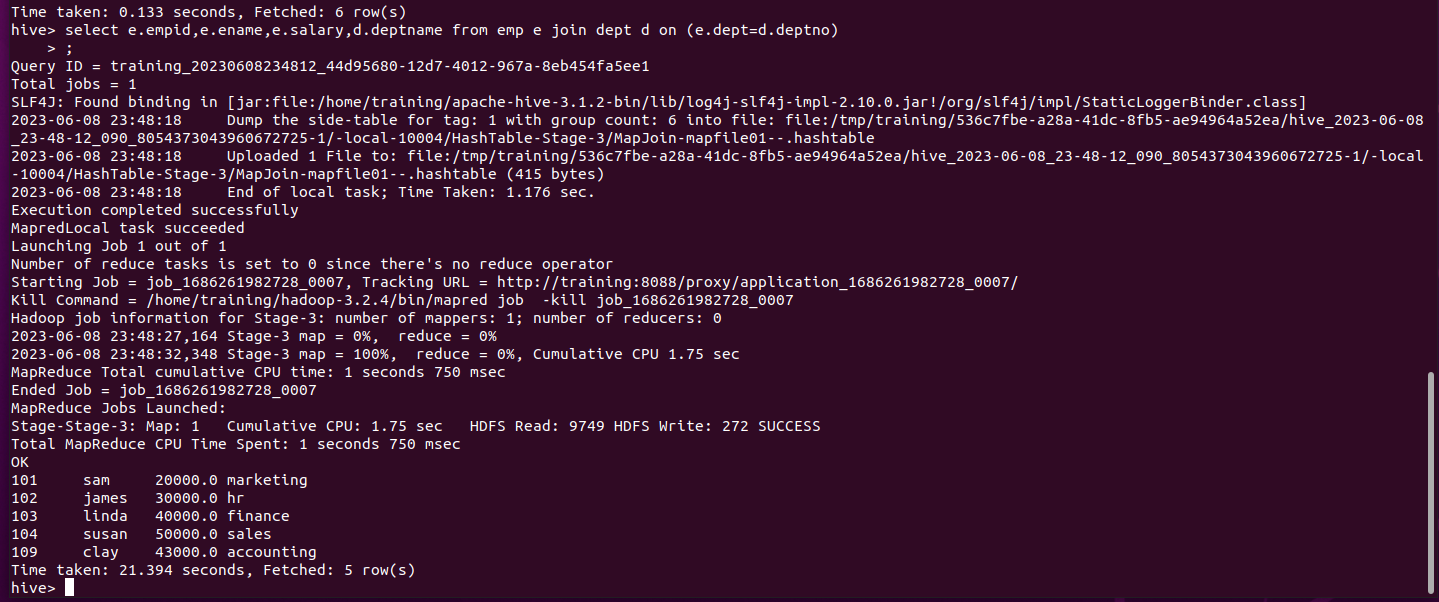


Exercise 2: Create a table in dsci5350 database using the dept.txt file provided on Canvas.  
Load the data into the table.

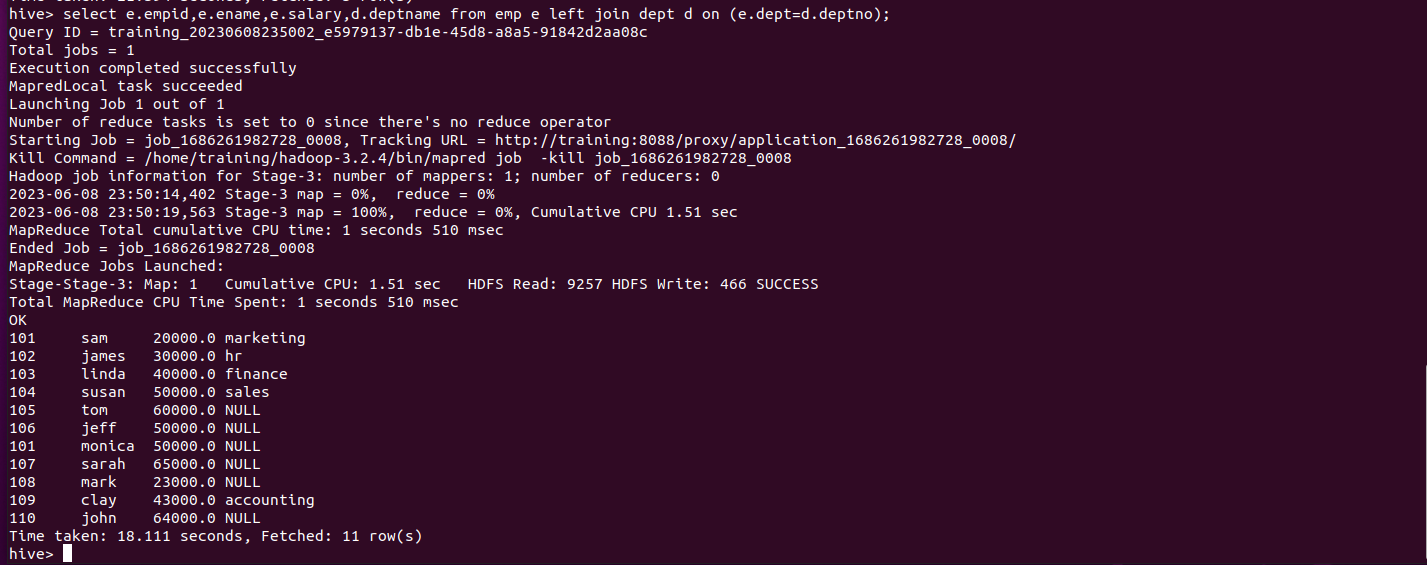


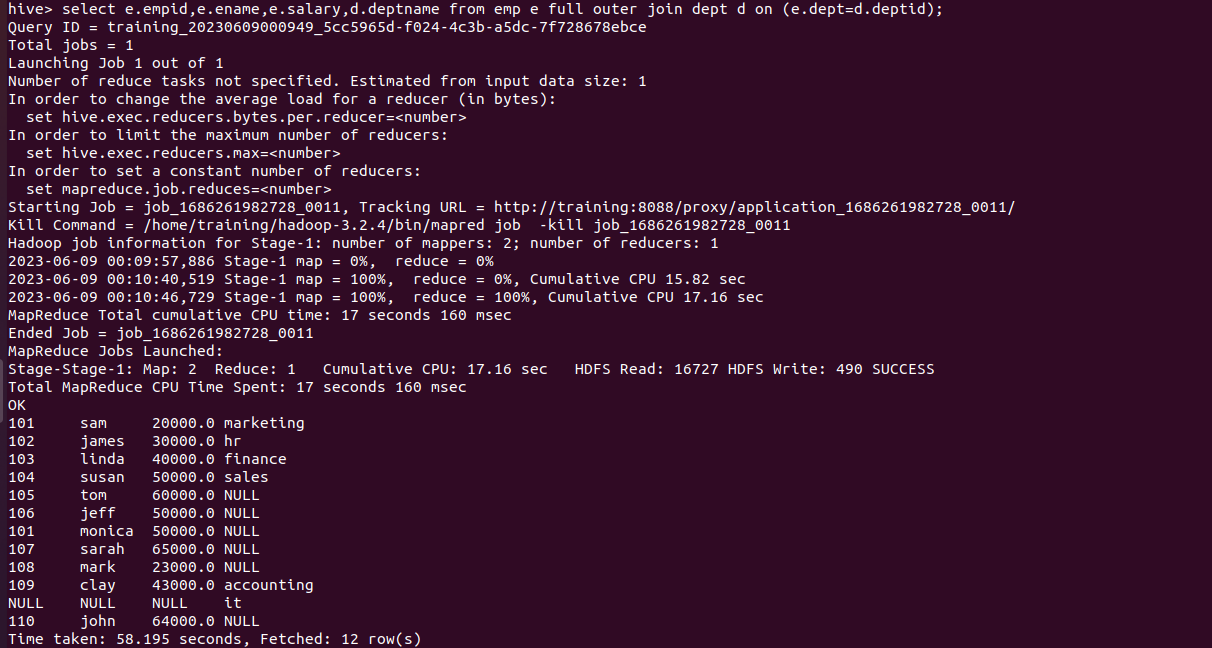
Joining Tables

Inner Join



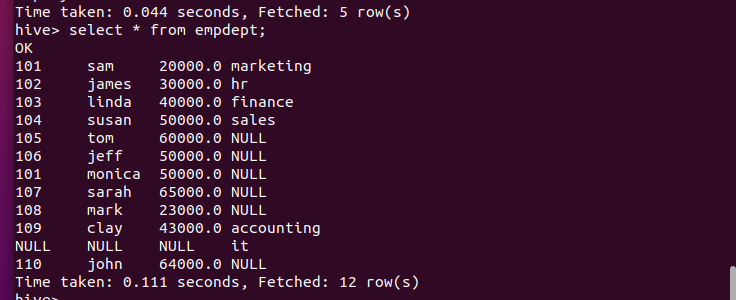
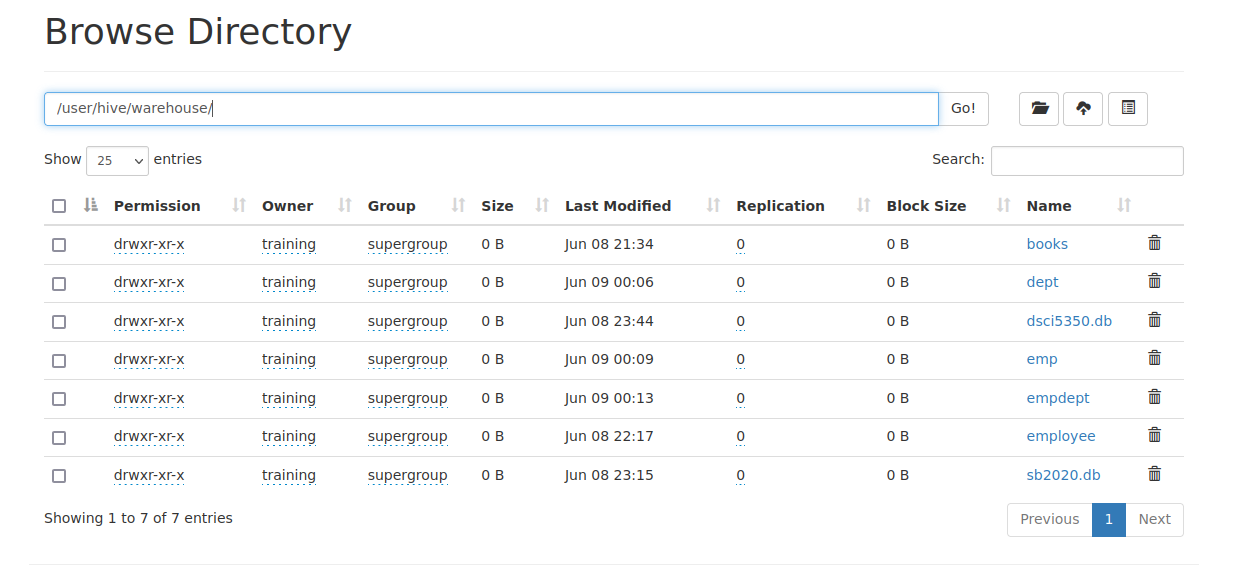
Left Join



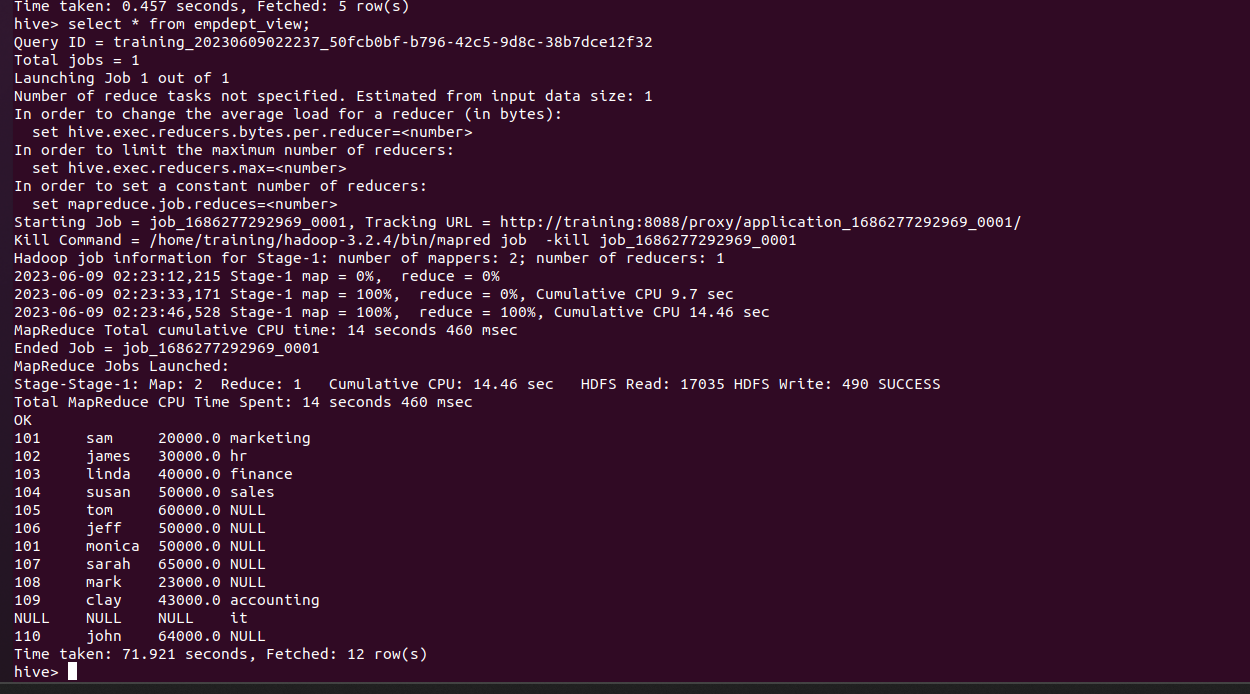
Full Outer Join  


Storing the results of a Query in a table

Tasks:  
- Validate the creation of this table in HDFS.  
- Validate the definition of this table.  
- Run a test query against it

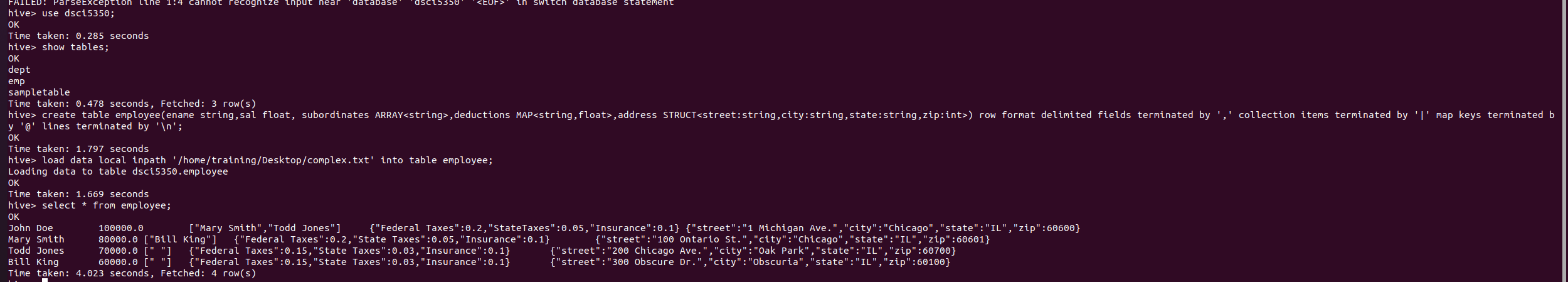
Creating a View  
Tasks:  
- Validate the creation of this view in HDFS.View can not be validated in HDFS as it is virtual.  
- Run a test query against it



Question: If data is appended in the emp table now (i.e. after the creation of the view and the  
empdept table), which one of the following will show the updated data?  
Empdept table  
Empdept\_view

Answer: Empdept\_view will give us dynamic and updated data.

Handling Complex Data Types in Hive  
1. Copy complex.txt file (provided in Canvas) to your VM desktop  
2. View the data to understand the format. We will need this information when creating  
a table in Hive. Use the following command to create the table:  
Ensure that the table is created in dsci5350 database.  
3. Next, load data from the /home/training/Desktop/complex.txt file into the employee  
table using the command:  
4. Validate that the table is loaded with data.

  
5. Next, we will write some DML commands to consume this data. To select an  
employee and all it the subordinates, write:

Now, select the first member of the subordinate’s array. For Arrays, we use the index  
of the array starting from 0 to n-1.  
Next, select all the deduction for John Doe.  
Next, select only the federal taxes for the employees. For MAPS, we must use the key  
to find the value. In this example, the key is “Federal Taxes”.  
Now, select the address for the employees.  
Next, select just the state column from the address STRUCT.

