**1. Create the necessary structure in a MySQL database using the steps mentioned below:**

**a. Create a new database in MySQL with the name midproject**

**b. Create a table in this database with the name census\_adult to store the input dataset**

**c. Load the dataset into the table**

**d. Verify whether data is loaded properly**

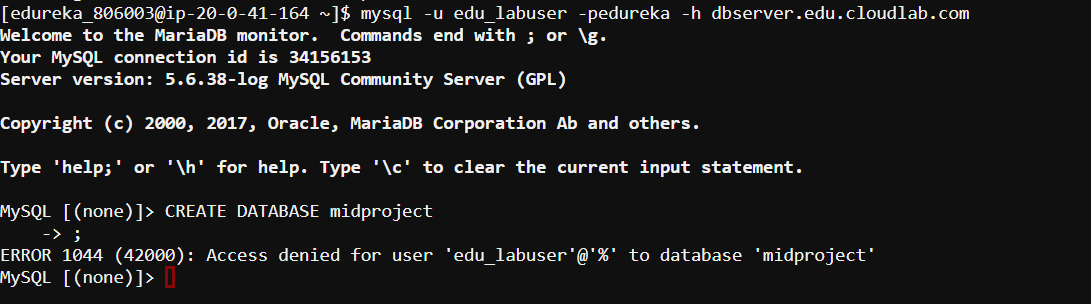
**e. Verify the table for unwanted data such as ‘?’,’Nan’ and ‘Null’**

**f. Get the counts for the columns which contain unwanted data**

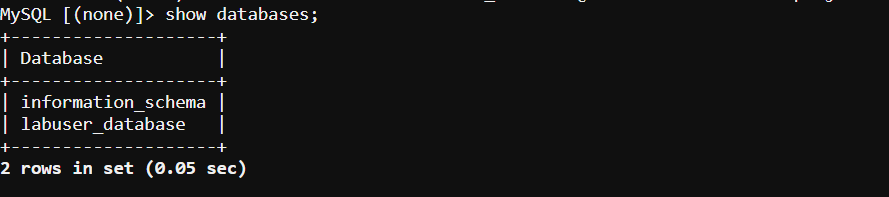
**g. Clean the data by replacing the unwanted data with others**

1. mysql -u edu\_labuser -pedureka -h dbserver.edu.cloudlab.com

CREATE DATABASE midproject;( do not have permissions for this)

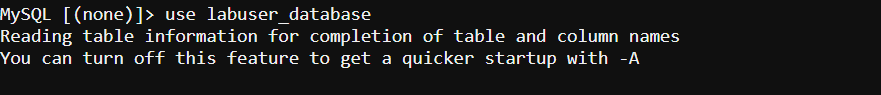


Show databases;



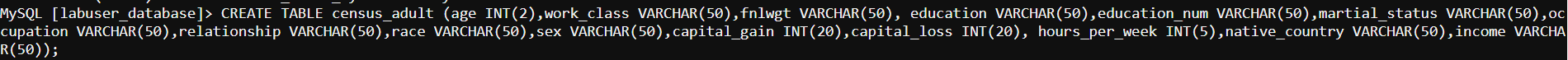
Use midproject

(using labuser\_database since we do not have permissions to create a database)



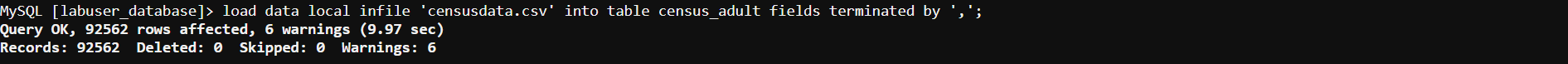
1. CREATE TABLE census\_adult(age INT(2),work\_class VARCHAR(50),fnlwgt VARCHAR(50), education VARCHAR(50),education\_num VARCHAR(50),martial\_status VARCHAR(50),occupation VARCHAR(50),relationship VARCHAR(50),race VARCHAR(50),sex VARCHAR(50),capital\_gain INT(20),capital\_loss INT(20),

hours\_per\_week INT(5),native\_country VARCHAR(50),income VARCHAR(50));

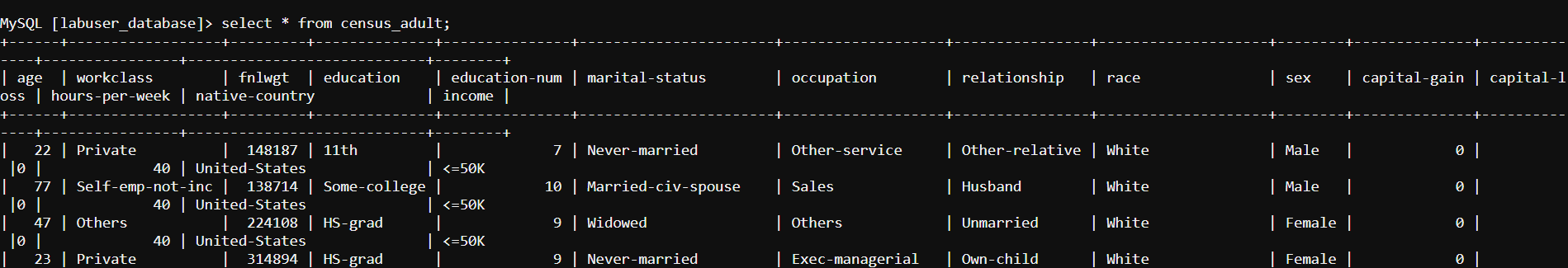




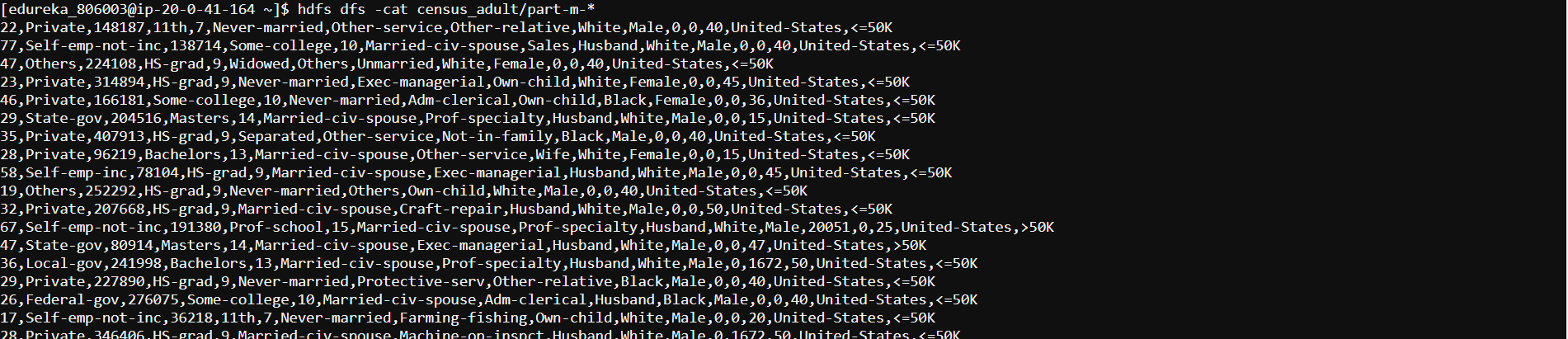
1. load data local infile '/user/edureka\_806003/data/ censusdata.csv ' into table census\_adult fields terminated by ',';



select \* from census\_adult



1. hdfs dfs -cat census\_adult/part-m-\*



1. and f.
2. select count(\*) from census\_adult where age is null or age='Nan' or age='?';

select count(\*) from census\_adult where workclass is null or workclass='Nan' or workclass='?';

select count(\*) from census\_adult where fnlwgt is null or fnlwgt ='Nan' or fnlwgt ='?';

select count(\*) from census\_adult where education is null or education='Nan' or education='?';

select count(\*) from census\_adult where education-num is null or education-num='Nan' or education-num='?';

select count(\*) from census\_adult where marital-status is null or marital-status ='Nan' or marital-status ='?';

select count(\*) from census\_adult where occupation is null or occupation='Nan' or occupation='?';

select count(\*) from census\_adult where relationship is null or relationship='Nan' or relationship='?';

select count(\*) from census\_adult where race is null or race ='Nan' or race ='?';

select count(\*) from census\_adult where sex is null or sex='Nan' or sex='?';

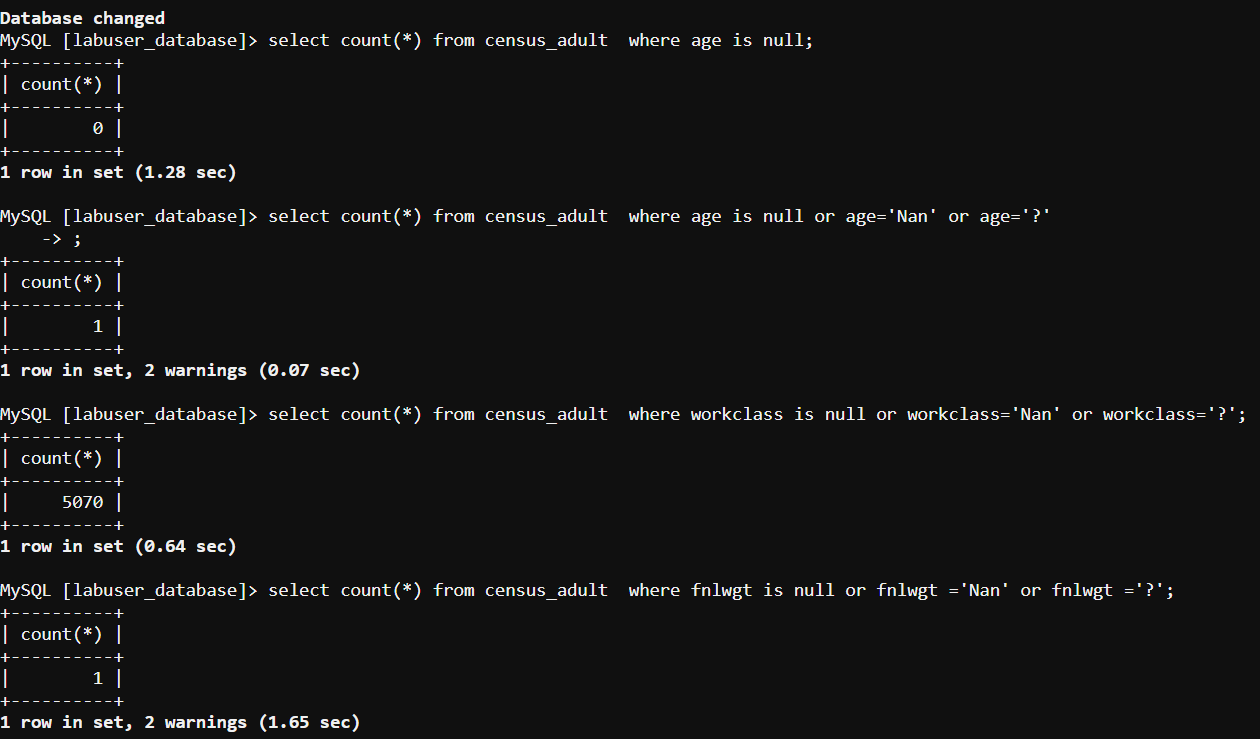
select count(\*) from census\_adult where capital-gain is null or capital-gain='Nan' or capital-gain='?';

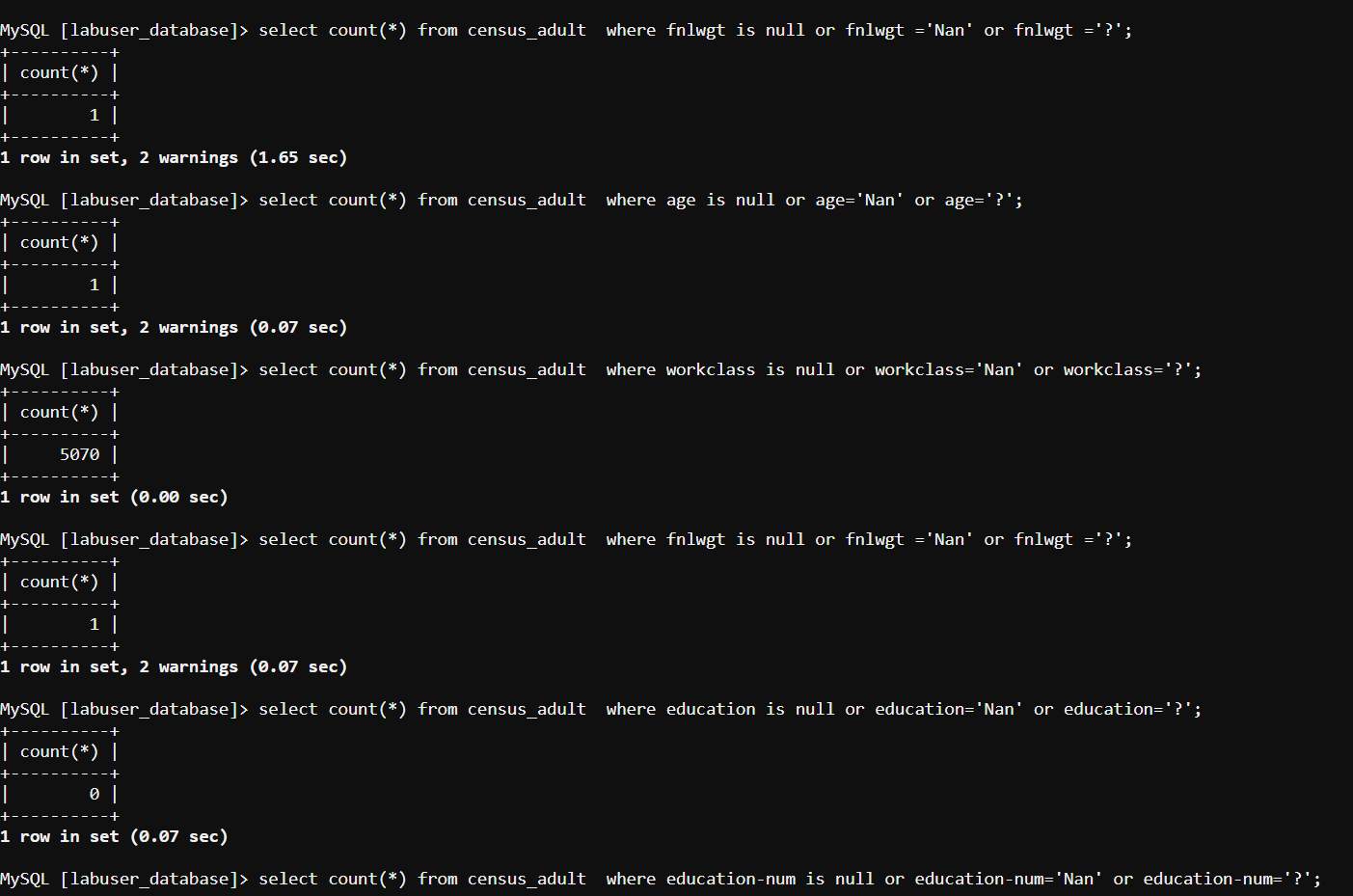
select count(\*) from census\_adult where capital-loss is null or capital-loss ='Nan' or capital-loss ='?';

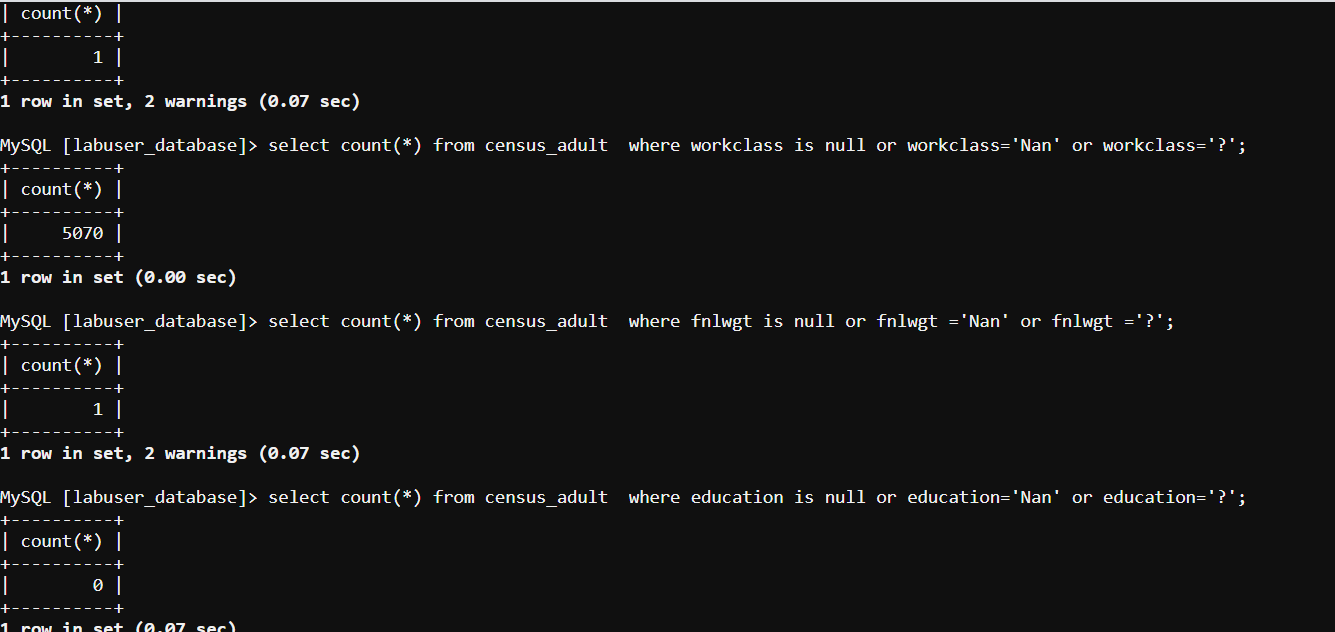
select count(\*) from census\_adult where hours-per-week is null or hours-per-week='Nan' or hours-per-week='?';

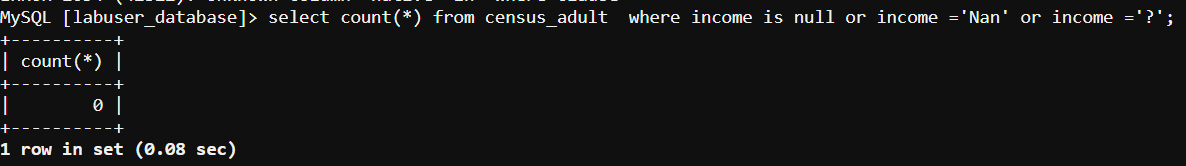
select count(\*) from census\_adult where native-country is null or native-country='Nan' or native-country='?';

select count(\*) from census\_adult where income is null or income ='Nan' or income ='?';









1. UPDATE census\_adult SET age = 'others' WHERE age is null or age='Nan' or age='?';

UPDATE census\_adult SET workclass = 'others' WHERE workclass is null or workclass='Nan' or workclass='?';

UPDATE census\_adult SET fnlwgt = 'others' WHERE fnlwgt is null or fnlwgt ='Nan' or fnlwgt ='?';

UPDATE census\_adult SET education = 'others' WHERE education is null or education='Nan' or education='?';

UPDATE census\_adult SET education-num = 'others' WHERE education-num is null or education-num='Nan' or education-num='?';

UPDATE census\_adult SET marital-status = 'others' WHERE marital-status is null or marital-status ='Nan' or marital-status ='?';

UPDATE census\_adult SET occupation = 'others' WHERE occupation is null or occupation='Nan' or occupation='?';

UPDATE census\_adult SET relationship = 'others' WHERE relationship is null or relationship='Nan' or relationship='?';

UPDATE census\_adult SET race = 'others' WHERE race is null or race ='Nan' or race ='?';

UPDATE census\_adult SET sex = 'others' WHERE sex is null or sex='Nan' or sex='?';

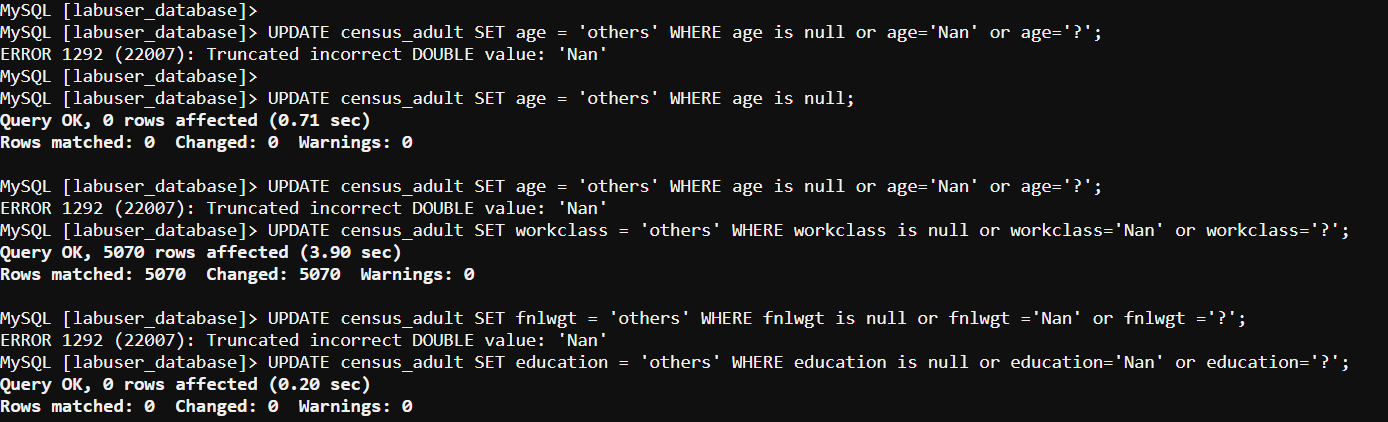
UPDATE census\_adult SET capital\_gain = 'others' WHERE capital\_gain is null or capital\_gain='Nan' or capital\_gain='?';

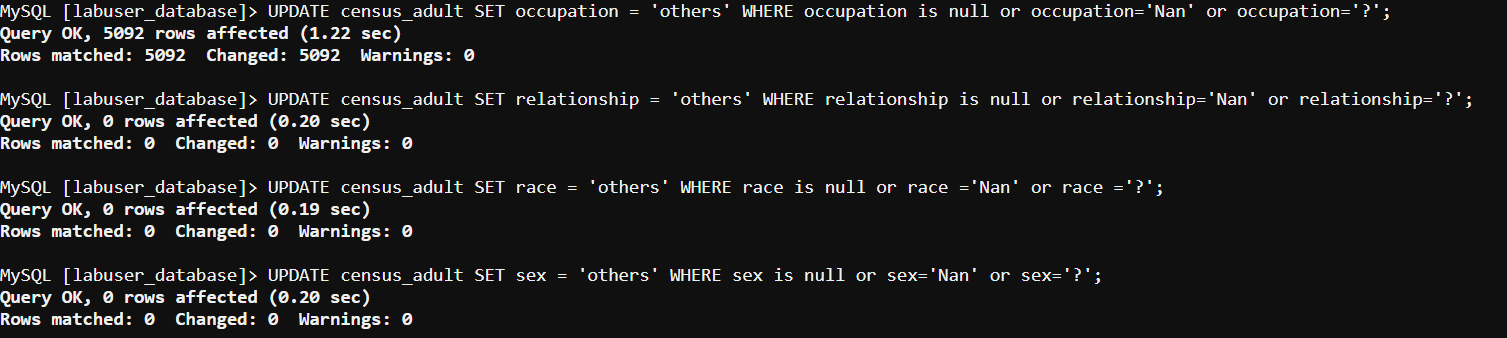
UPDATE census\_adult SET capital-loss = 'others' WHERE capital-loss is null or capital-loss ='Nan' or capital-loss ='?';

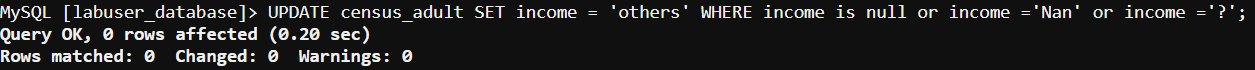
UPDATE census\_adult SET hours-per-week = 'others' WHERE hours-per-week is null or hours-per-week='Nan' or hours-per-week='?';

UPDATE census\_adult SET native-country = 'others' WHERE native-country is null or native-country='Nan' or native-country='?';

UPDATE census\_adult SET income = 'others' WHERE income is null or income ='Nan' or income ='?';

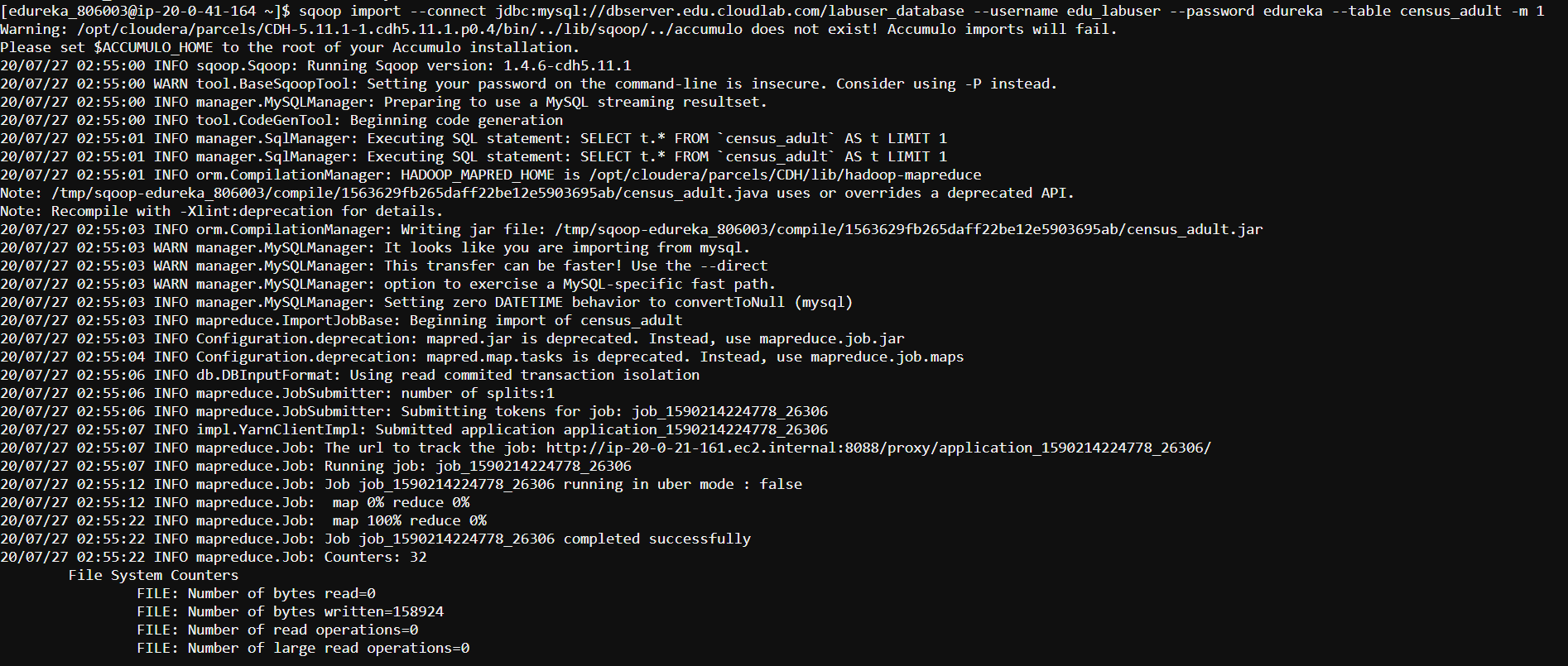


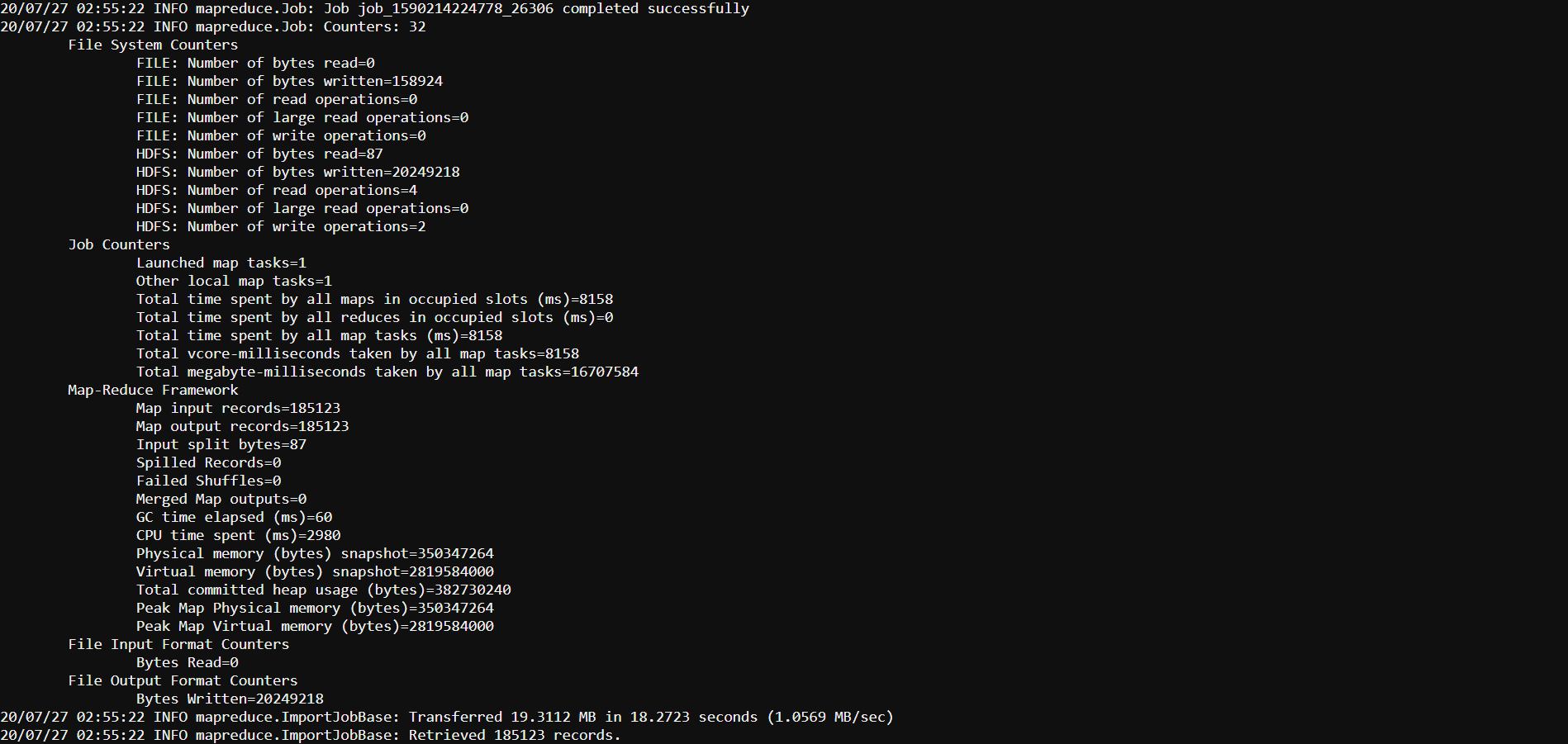




**2. Import the above data from MySQL into a Hive table using Sqoop**

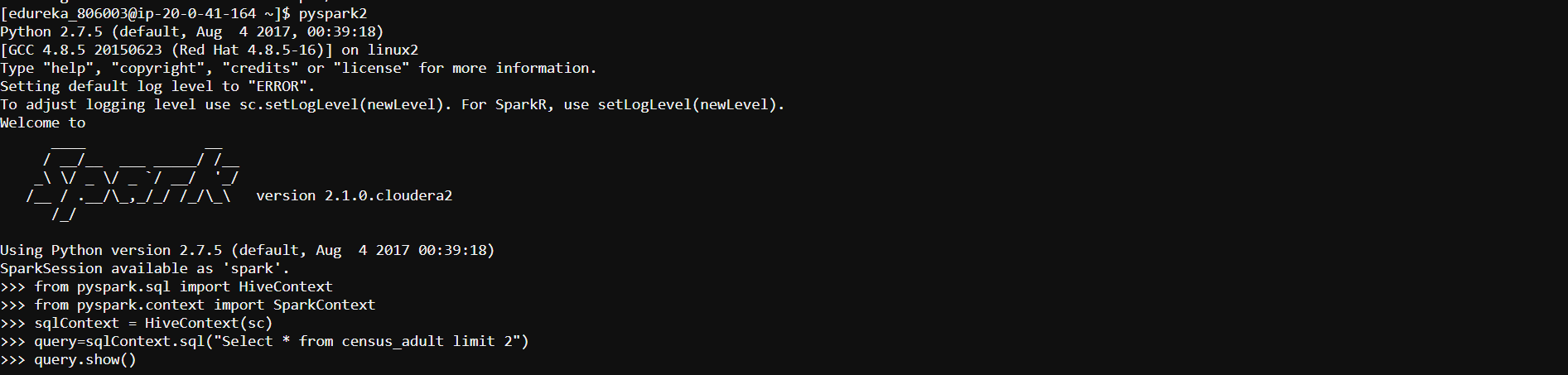
sqoop import --connect jdbc:mysql://dbserver.edu.cloudlab.com/ labuser\_database --username edu\_labuser --password edureka --table census\_adult-m 1

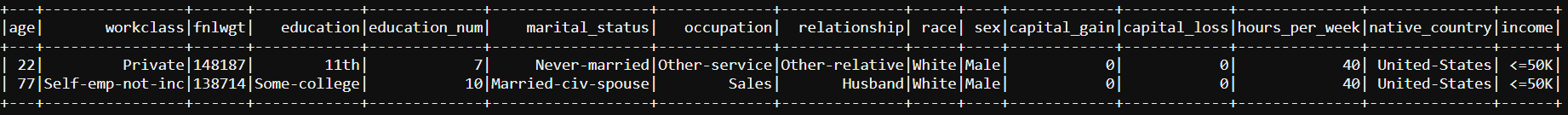




**3. Connect to PySpark using web console to access the created Hive table. Perform the following queries and note the time taken for execution in each of the queries.**

**a. Query the table to get the number of adults based on income and gender b. Query the table to get the number of adults based on income and workclass.**

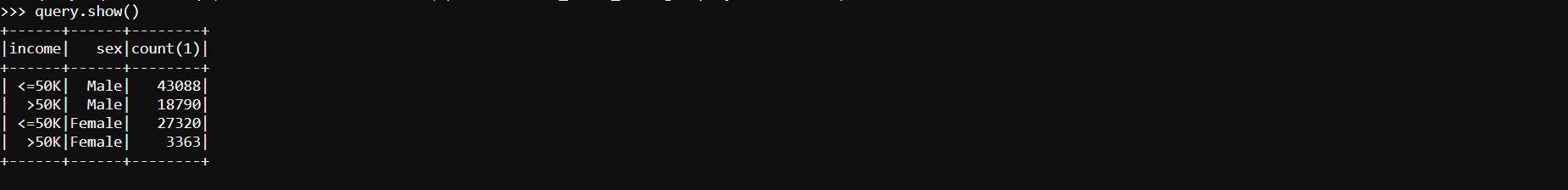




1. query=sqlContext.sql("select income,sex,count(\*) from census\_adult group by sex,income")

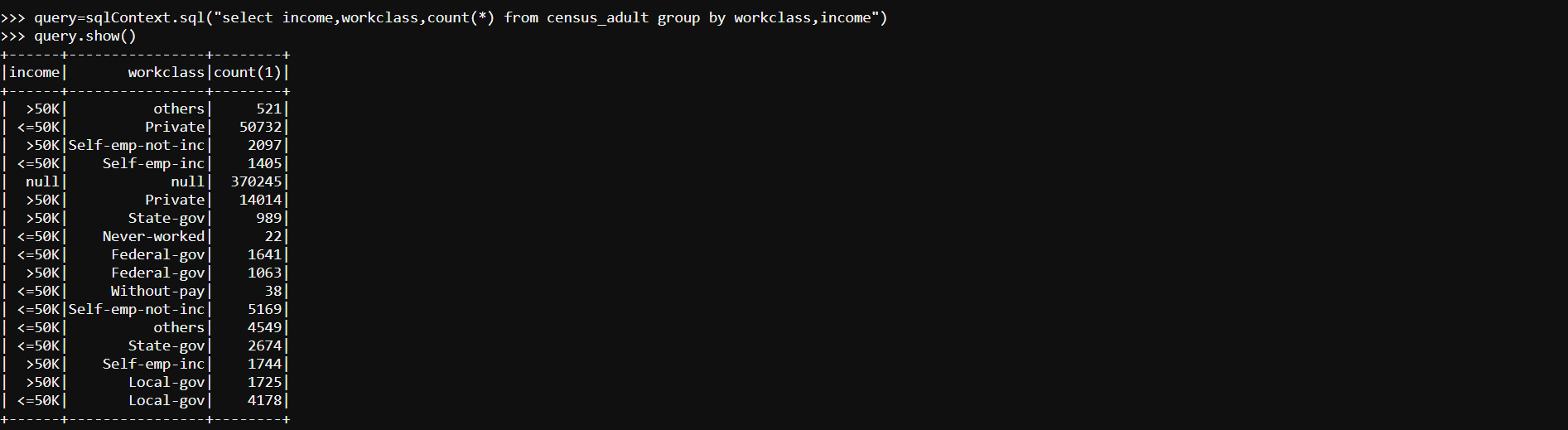
query.show()





1. query=sqlContext.sql("select income,sex,count(\*) from census\_adult group by workclass,income")

query.show()



**4. Access the following two tables created as part of Problem 1 (HDFS and Hive) and perform the steps as mentioned below:**

**a. Access Hive External Table with partition**

**i. Query the table to get the number of adults based on income and gender**

**ii. Query the table to get the number of adults based on income and workclass**

**b. Access Hive Internal Table with Partition**

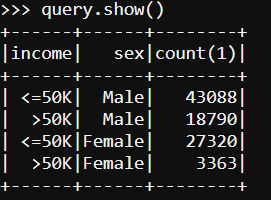
**i. Query the table to get the number of adults based on income and gender**

**ii. Query the table to get the number of adults based on income and workclass**

1. Hive External Tables
2. query=sqlContext.sql("select income,sex,count(\*) from census\_data\_external group by sex,income")

query.show()

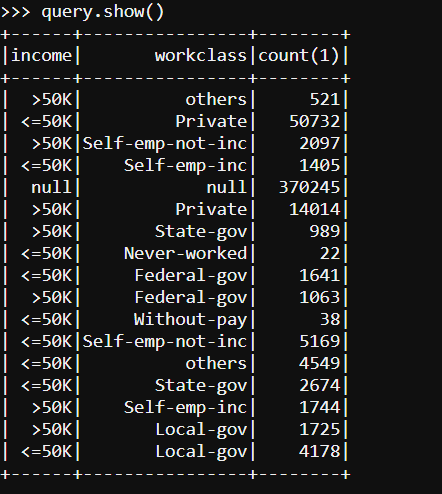




1. query=sqlContext.sql("select income, workclass,count(\*) from census\_data\_external group by workclass,income")

query.show()

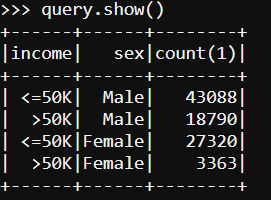




1. Hive Internal Tables
2. query=sqlContext.sql("select income,sex,count(\*) from census\_data group by sex,income")

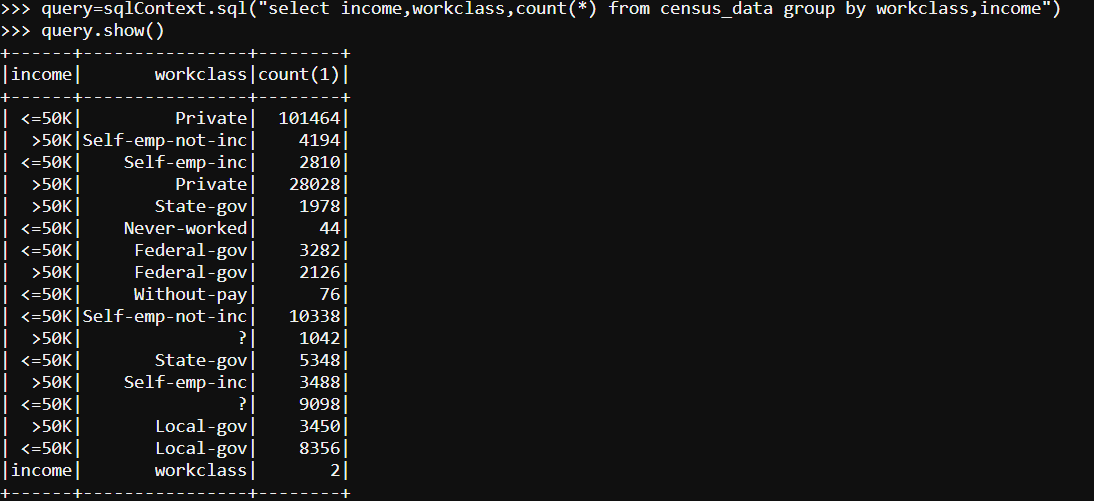
query.show()





1. query=sqlContext.sql("select income,workclass,count(\*) from census\_data group by workclass,income")

query.show()



**5. Comment on the time taken for executing these commands using Spark as compared to the time taken for execution in Hive (Problem Statement 1).**

Time taken to execute Query in spark is much less than what it takes in hive.