**HIVE PROJECT**

**All India Pincode Directory**

Description:

Postal Index Number (PIN) or PIN Code is a 6 digit code of Post Office numbering used by India Post. The PIN was introduced on August 15, 1972. There are 9 PIN regions in the country. The first 8 are geographical regions and the digit 9 is reserved for the Army Postal Service. The first digit indicates one of the regions. The first 2 digits together indicate the sub region or one of the postal circles. The first 3 digits together indicate a sorting / revenue district. The last 3 digits refer to the delivery Post Office.

DATA-SET:

http://data.gov.in/dataset/all-india-pincode-directory

Analyzing the data:

I will analyze the following from the data provided:

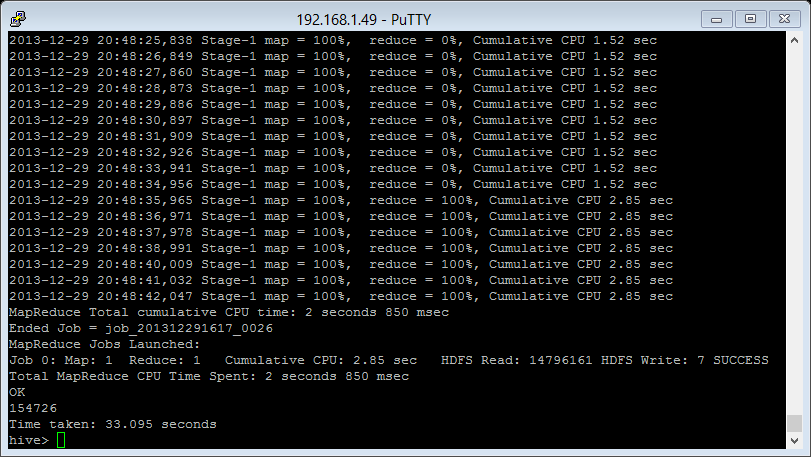
1. Select distinct state form the dataset.

select DISTINCT state from indiapincodes;



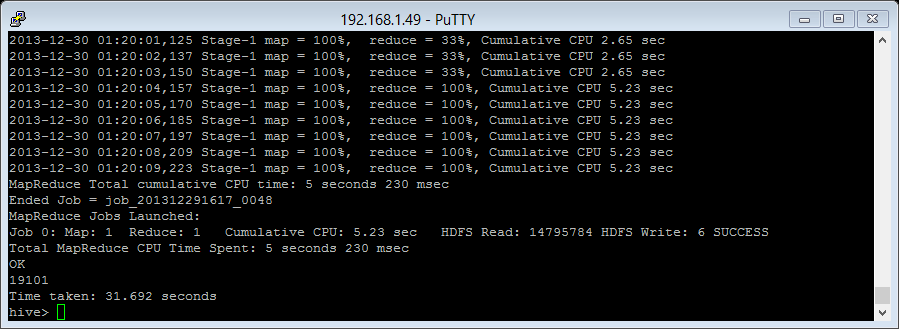
1. List the total number of records in the table.

select count(\*) from indiapincodes;



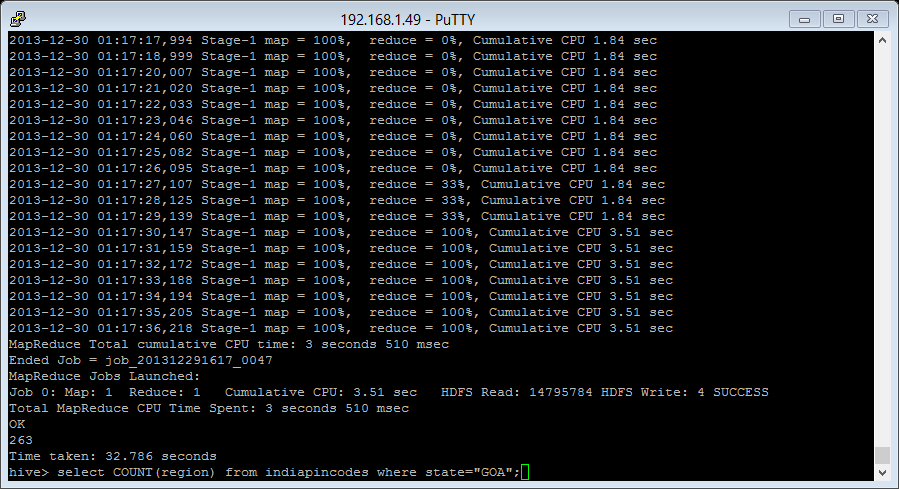
1. Count the number of distinct pincode in the table.

select count(DISTINCT pincode) from indiapincodes;



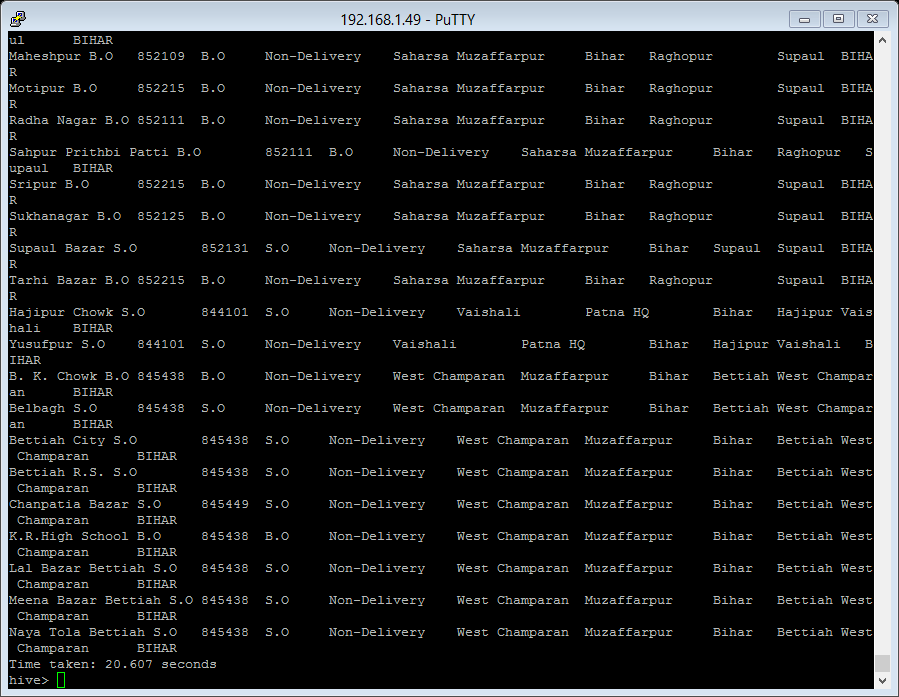
1. List the total number of regions in the GOA state.

select COUNT(region) from indiapincodes where state="GOA";



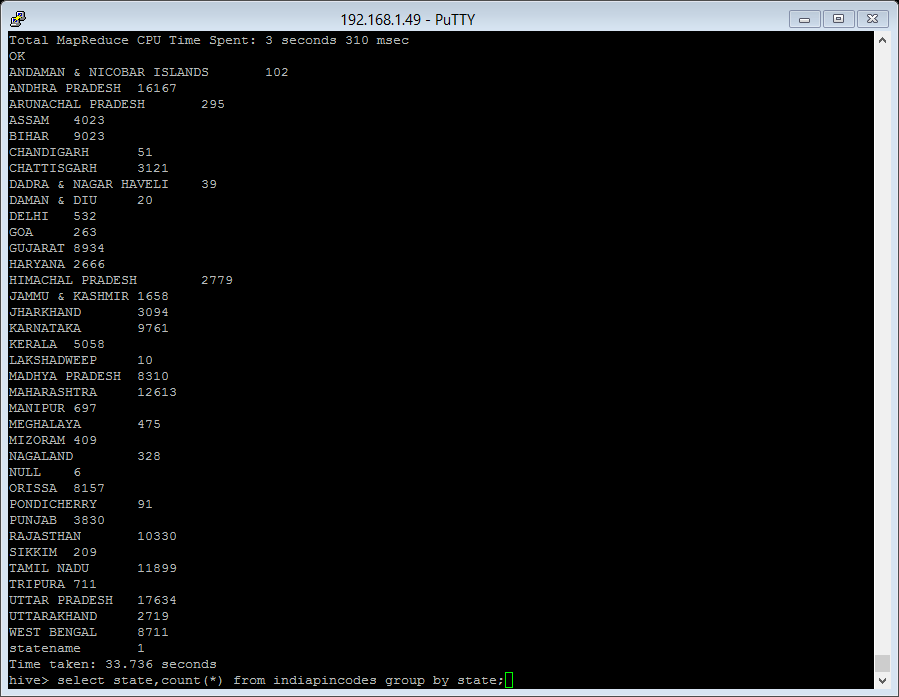
1. List the number of non-delivery areas in Bihar.

select \* from pincode where delivery="Non-Delivery" and state="BIHAR";



1. List the number of different states in the table

select state,count(\*) from indiapincodes group by state;



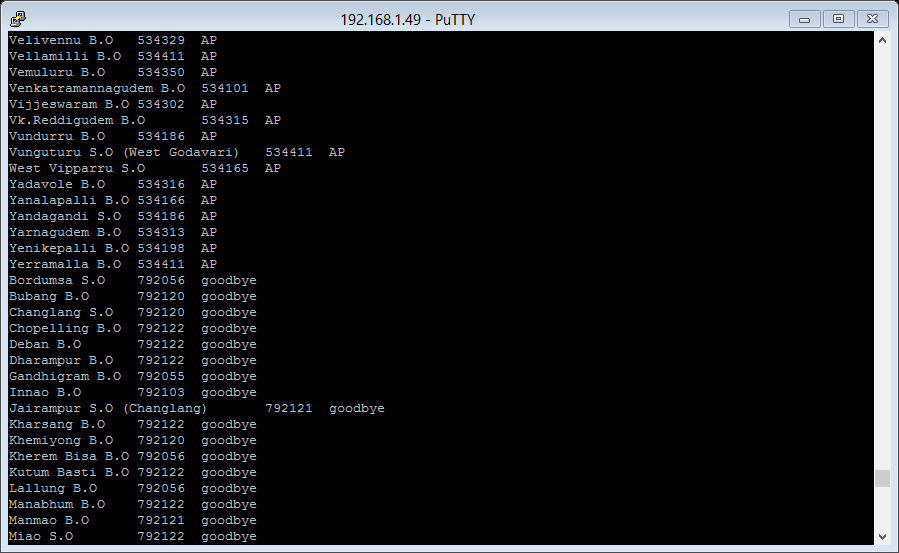
1. Analyze the data with a condition (using when and then).

SELECT officename,pin, WHEN state = 'MADHYA PRADESH' then 'MP'

WHEN state = 'ANDHRA PRADESH' then 'AP'

else 'goodbye'

end as bracket from indiapincodes;



**Analyze Data Using Pig:**

**The data will be analyzed from**

[**https://nycopendata.socrata.com/Education/SAT-Results/f9bf-2cp4**](https://nycopendata.socrata.com/Education/SAT-Results/f9bf-2cp4)**?**

Following analysis will be done to the dataset

1. Load the table in Pig

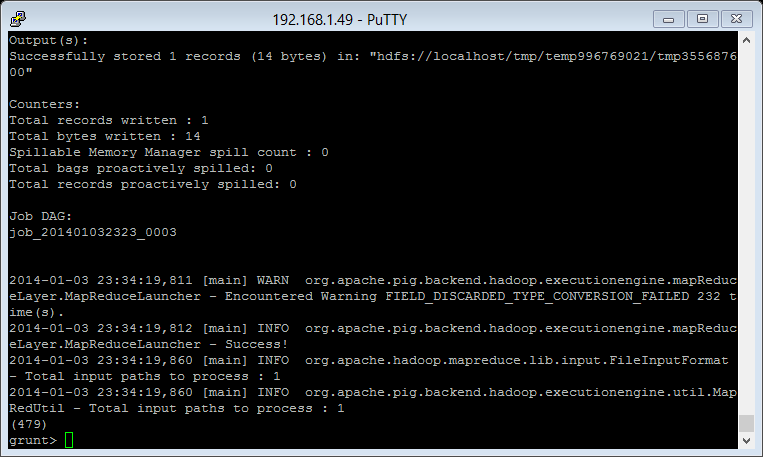
satresults = LOAD 'pig/satresults' using PigStorage(',') AS(DBN:chararray,schoolname:chararray,sattaken:int,criticalreading:int,mathscore:int,writingscore:int);

1. List the total number of SAT tests

groupall = group satresults ALL;

countall = foreach groupall generate COUNT(satresults);

dump countall;

****

1. Display the maximum Math average score

maxmathscore = foreach groupall generate MAX(satresults.mathscore);

dump maxmathscore;

****

1. The minimum Writing average score

minwritingscore = foreach groupall generate MIN(satresults.writingscore);

dump minwritingscore;

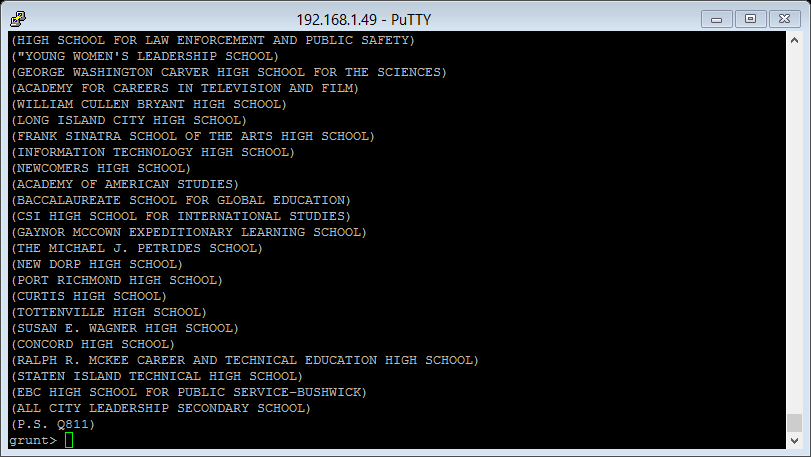
****

1. List the name of the schools that scored more than 400 in writing test

morethan400 = filter satresults by mathscore > 400;

listschoolname = foreach morethan400 generate schoolname;

dump listschoolname;

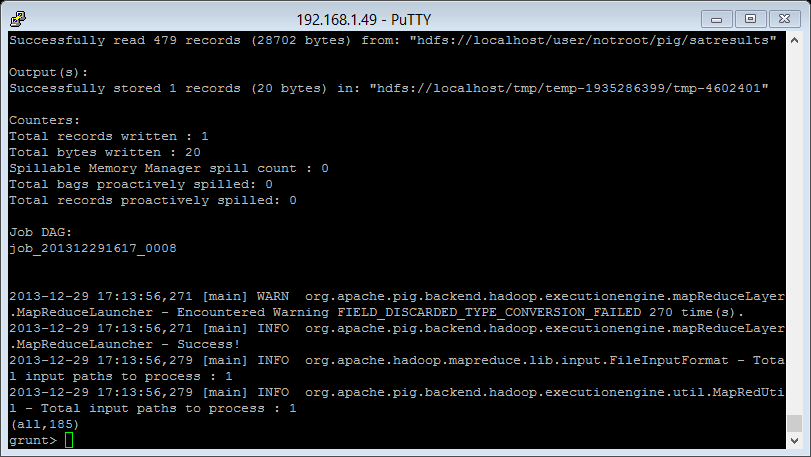
****

1. Count the total number of schools with mathscore more than 400

groupmorethan400 = GROUP morethan400 ALL;

countmorethan400 = foreach groupmorethan400 generate group, COUNT(morethan400);

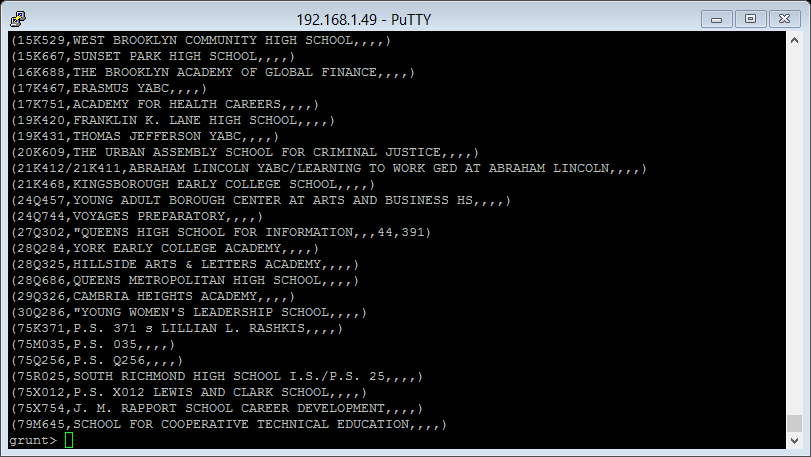
dump countmorethan400;

****

1. Pull out all of the invalid records in the dataset.

filternull = FILTER satresults by criticalreading is null;

dump filternull;

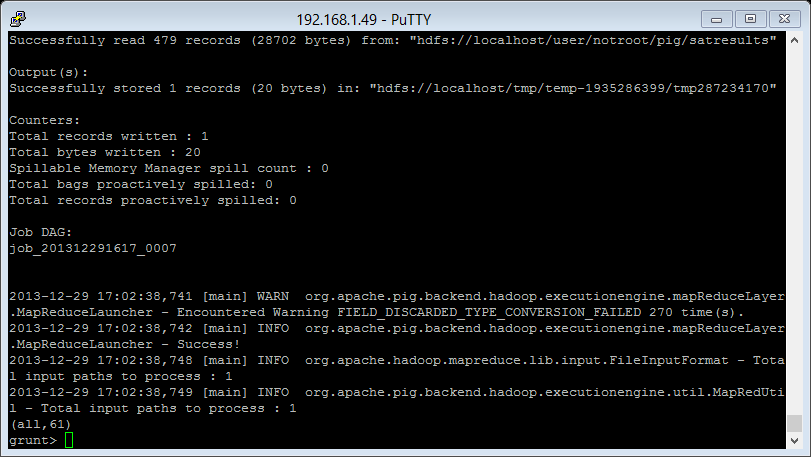
****

1. Count the total number of null values in the table

groupfilternull = GROUP filternull ALL;

countfilternull = foreach groupfilternull generate group, COUNT(filternull);

dump countfilternull;

****

1. Join the tables………………

There is one more table newsat1 that has the same columns as satresults but the only difference is the grade column.

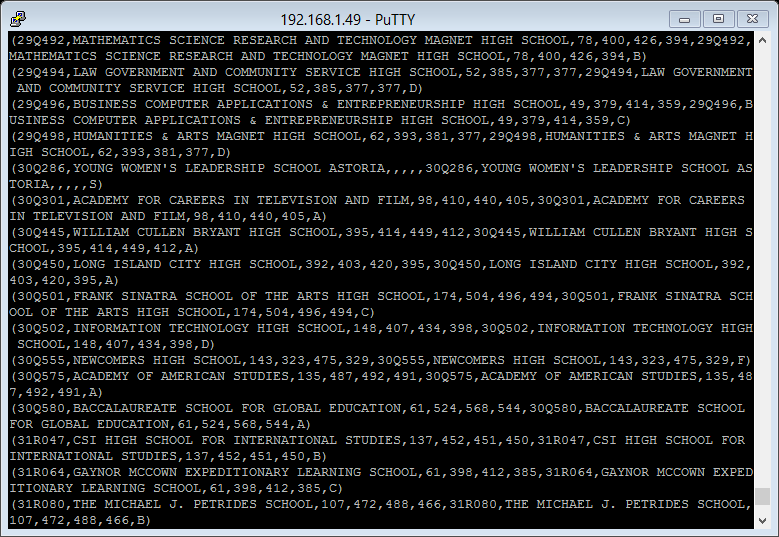
I would like to join these two tables and will do the same as follows:

The first job is to load the table in pig………….

newsat = LOAD 'newsat1' using PigStorage(',') AS(DBN1:chararray,schoolname:chararray,sattaken:int,criticalreading:int,mathscore:int,writingscore:int,grade:chararray);

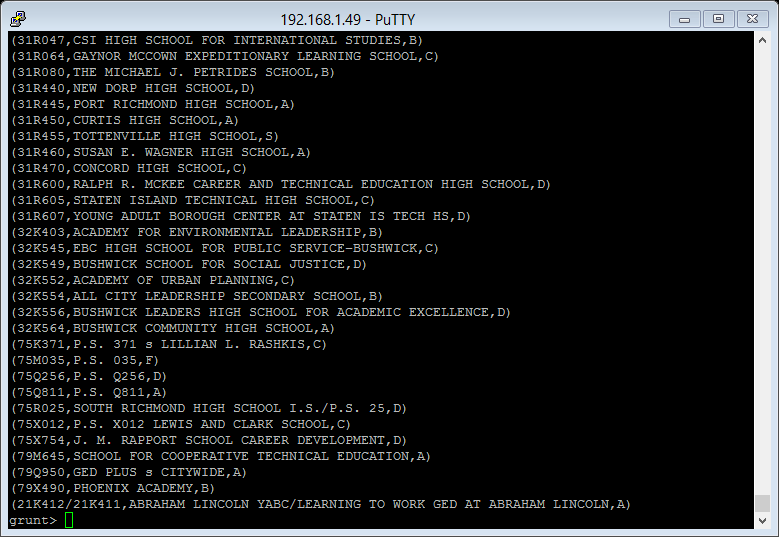
Now join the tables with the same column:

joinit = join satresults by DBN,newsat by DBN1;



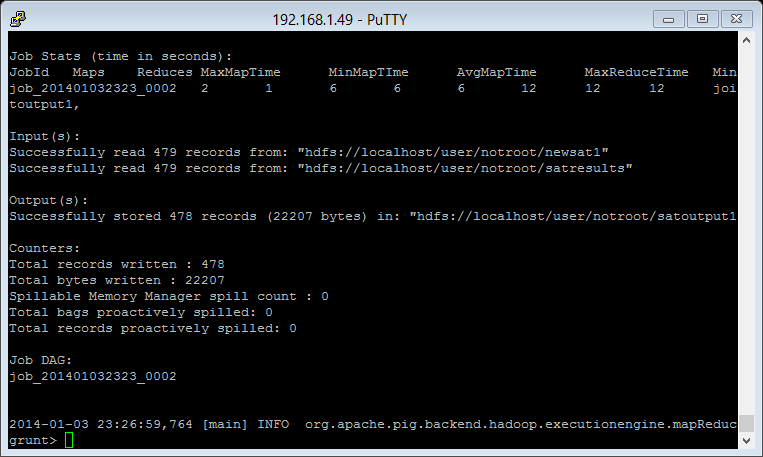
1. refinedjoin = foreach joinit generate $0,$1,$12;

dump refinedjoin;



11.The last step is to store the output in some file:

store refinedjoin into 'satoutput'



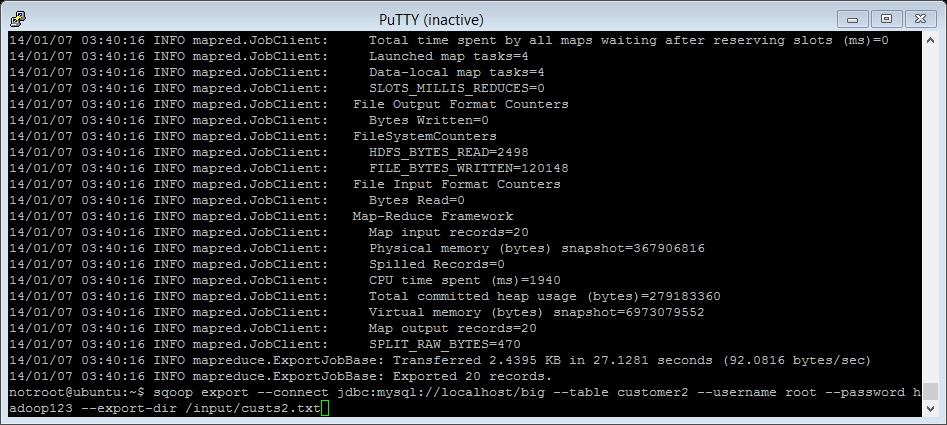
USING SQOOP

1. FROM HDFS TO MYSQL

sqoop export –connect jdbc:mysql://localhost/big

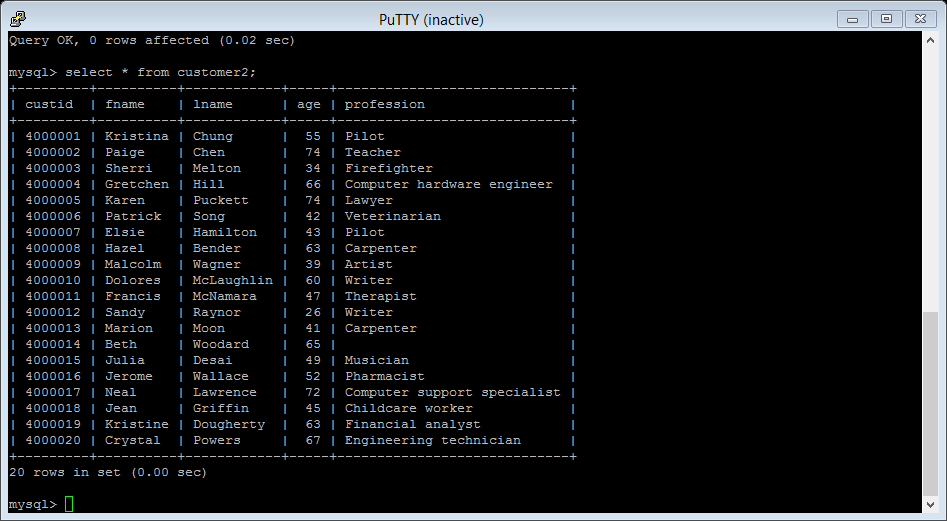
--table customer2 --username root

password hadoop123 --export-dir /input/custs2.txt



2.Check if the data is exported to mysql

Select \* from customer2;

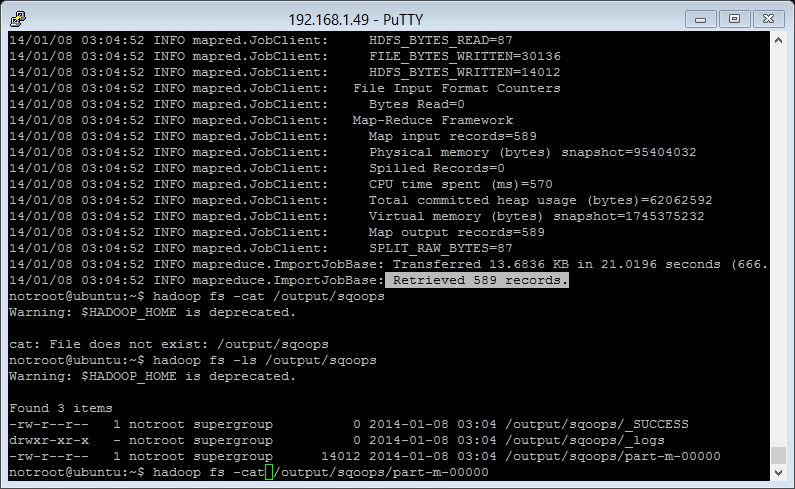


3.To import data from mysql to HDFS, we use import command. We wanted to import table name birthnames from mysql to the output directory in HDFS.

sqoop import --connect jdbc:mysql://localhost/big

--table birthnames --username root --password hadoop123

--target-dir /output/sqoops



4.To check if the data is being imported we use the following command.

hadoop fs -cat /output/sqoops/part-m-00000

