In this project we are performing the data parsing and data analysis using Pig and export the results into MYSQL using SQOOP.

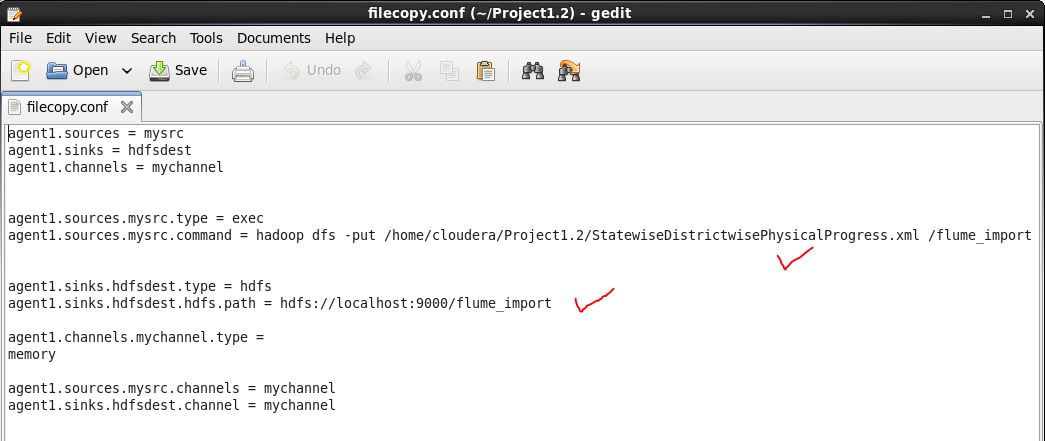
*Step 1*

Copy dataset from local file system to HDFS using flume.

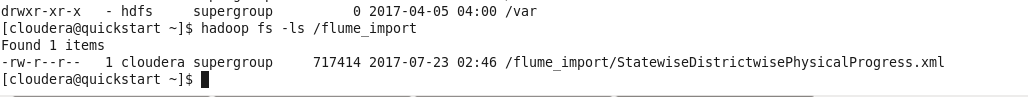
**Command:**

$ *flume-ng agent –n agent1 –c conf –f /home/cloudera/Project1.1/filecopy.conf*

***filecopy.conf file:***



***The input file is stored in /flume\_import folder in HDFS:***



Problem statement

1. Find out the districts who achieved 100 percent objective in BPL cards. Export the results to mysql using sqoop

First we have to convert XML data into csv using XMLLoader

* Register piggybank.jar

Then, pig commands as follows:



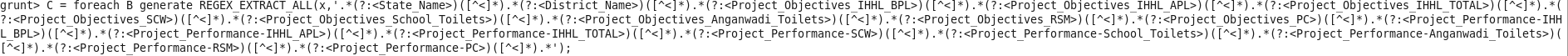
Here we will load the xml file using the default XML loader available in pig, inside the XML loader we are specifying that our root element is row and we are storing the whole thing with an alias name x as chararray.



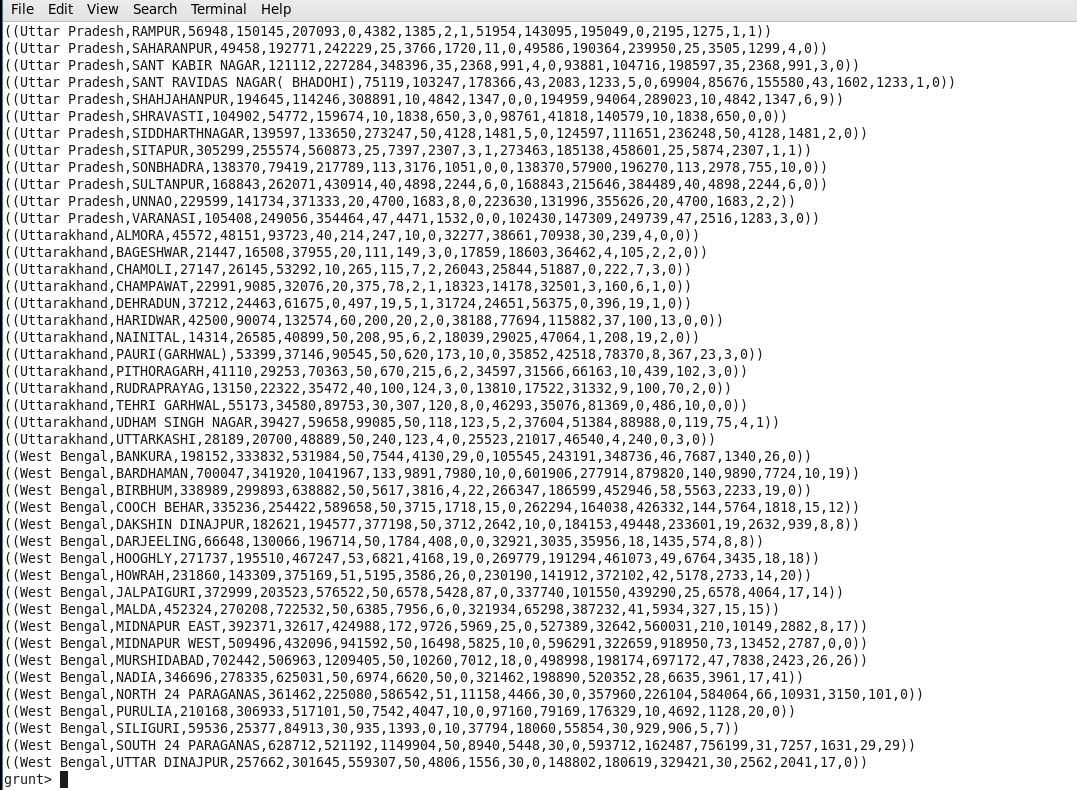
Here we are bringing the contents between the property tag in one line.

***C = foreach B generate REGEX\_EXTRACT\_ALL(x,'.\*(?:<State\_Name>)([^<]\*).\*(?:<District\_Name>)([^<]\*).\*(?:<Project\_Objectives\_IHHL\_BPL>)([^<]\*).\*(?:<Project\_Objectives\_IHHL\_APL>)([^<]\*).\*(?:<Project\_Objectives\_IHHL\_TOTAL>)([^<]\*).\*(?:<Project\_Objectives\_SCW>)([^<]\*).\*(?:<Project\_Objectives\_School\_Toilets>)([^<]\*).\*(?:<Project\_Objectives\_Anganwadi\_Toilets>)([^<]\*).\*(?:<Project\_Objectives\_RSM>)([^<]\*).\*(?:<Project\_Objectives\_PC>)([^<]\*).\*(?:<Project\_Performance-IHHL\_BPL>)([^<]\*).\*(?:<Project\_Performance-IHHL\_APL>)([^<]\*).\*(?:<Project\_Performance-IHHL\_TOTAL>)([^<]\*).\*(?:<Project\_Performance-SCW>)([^<]\*).\*(?:<Project\_Performance-School\_Toilets>)([^<]\*).\*(?:<Project\_Performance-Anganwadi\_Toilets>)([^<]\*).\*(?:<Project\_Performance-RSM>)([^<]\*).\*(?:<Project\_Performance-PC>)([^<]\*).\*');***

Now we are removing the brackets by using the above mentioned regular expression



The output will be in the following format:

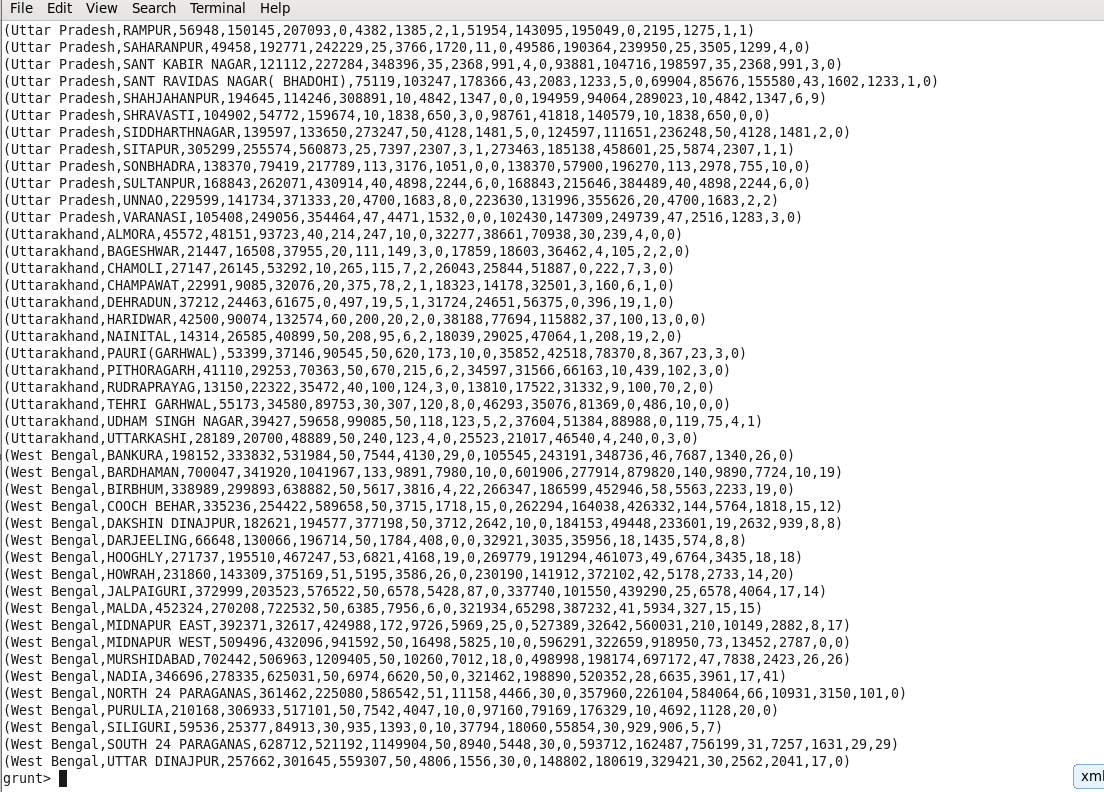


***D = FOREACH C generate FLATTEN (($0));***



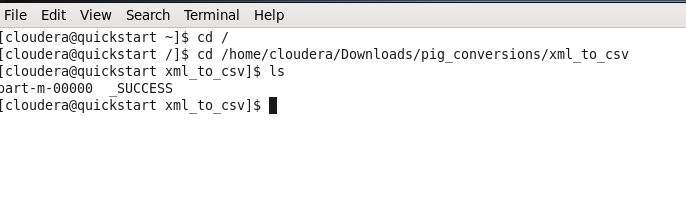
Here by using flatten it will remove the remaining brackets

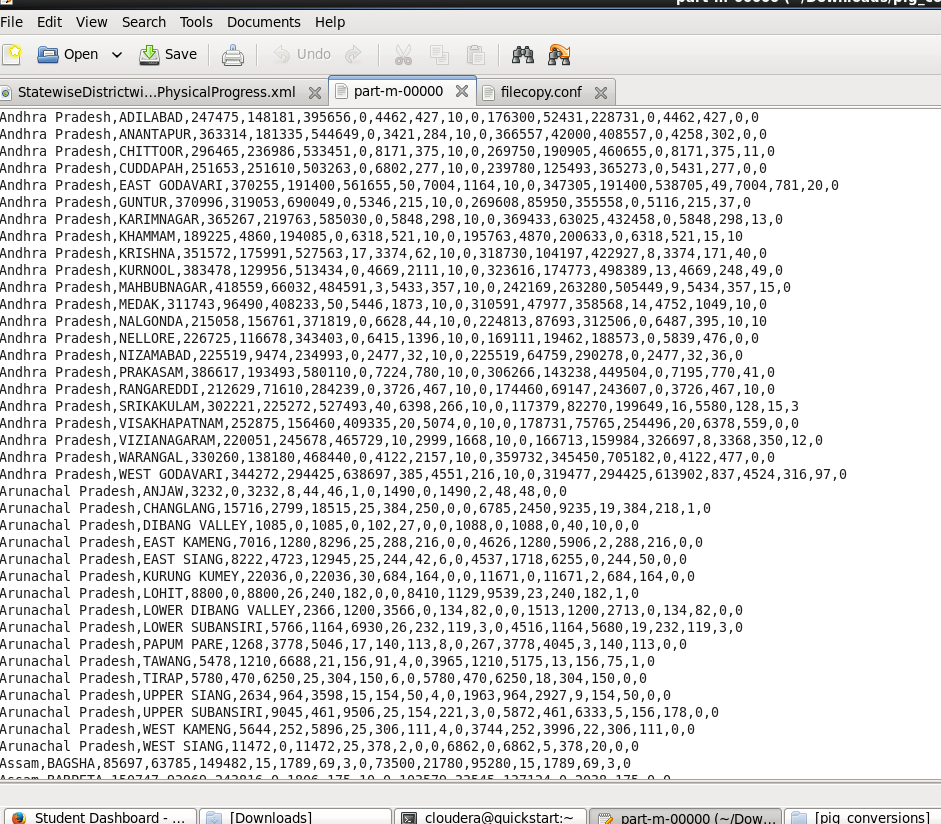
***Output:***



We have to store the result into local file system as CSV format

C:\Users\w10\Desktop\project\2-4.PNG



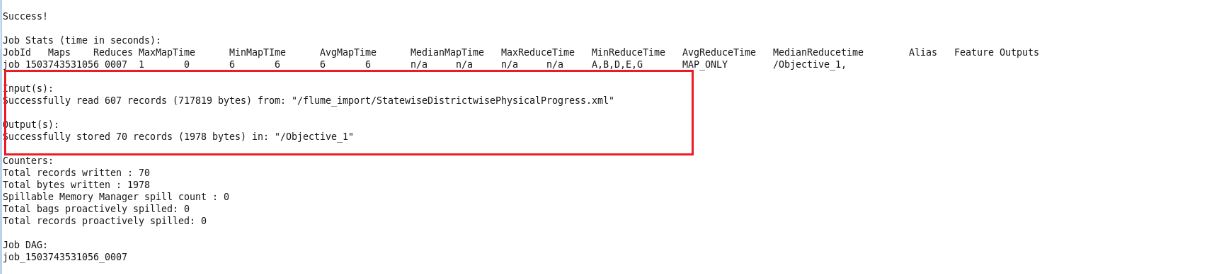


Now we need to get BPL Cards who have achieved 100 percent using below command

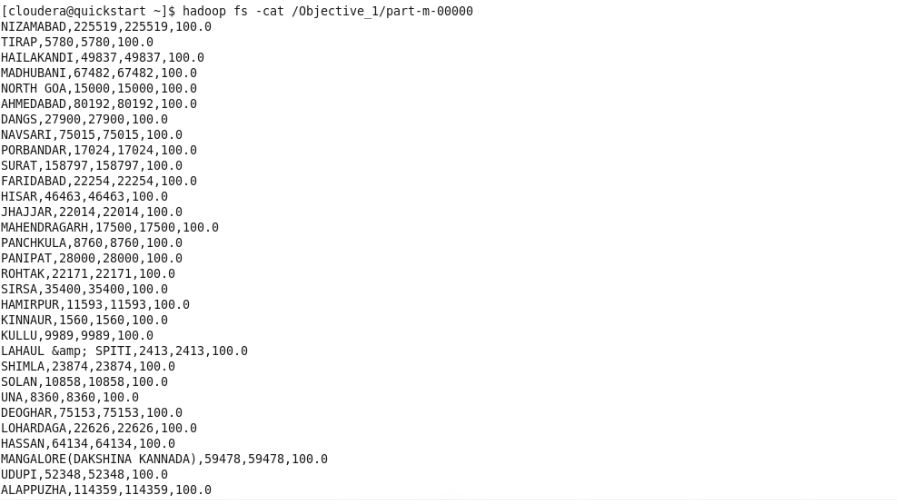
***E = foreach D generate $1,$2,$10,(($2/$10)\*100) as BPL\_Index ;***

***G = Filter E by BPL\_Index == 100;***

***STORE G into '/Objective\_1' using org.apache.pig.piggybank.storage.CSVExcelStorage();***

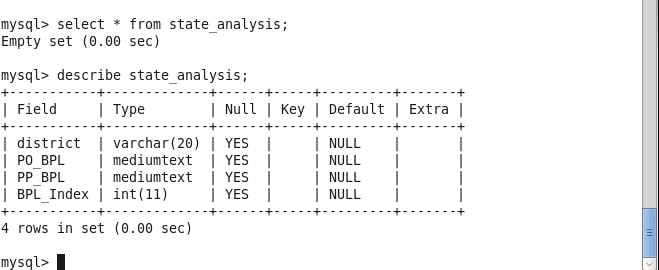


Now we can see the results stored in the /objective\_1 folder.

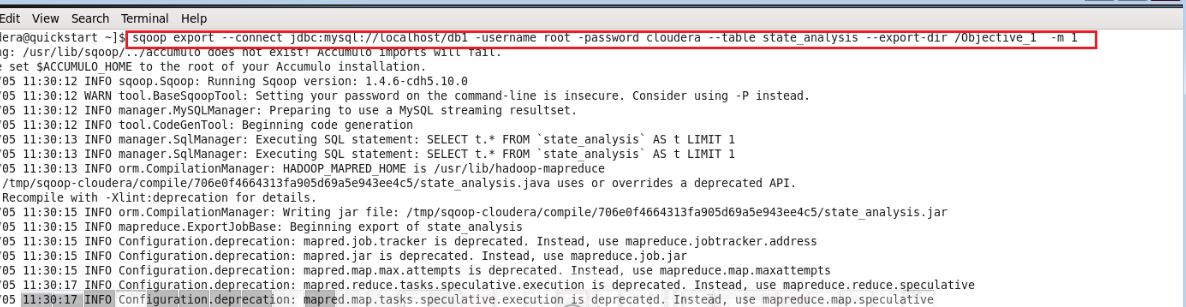


Now we have to use SQOOP command to export data into the state\_analysis table in MYSQL

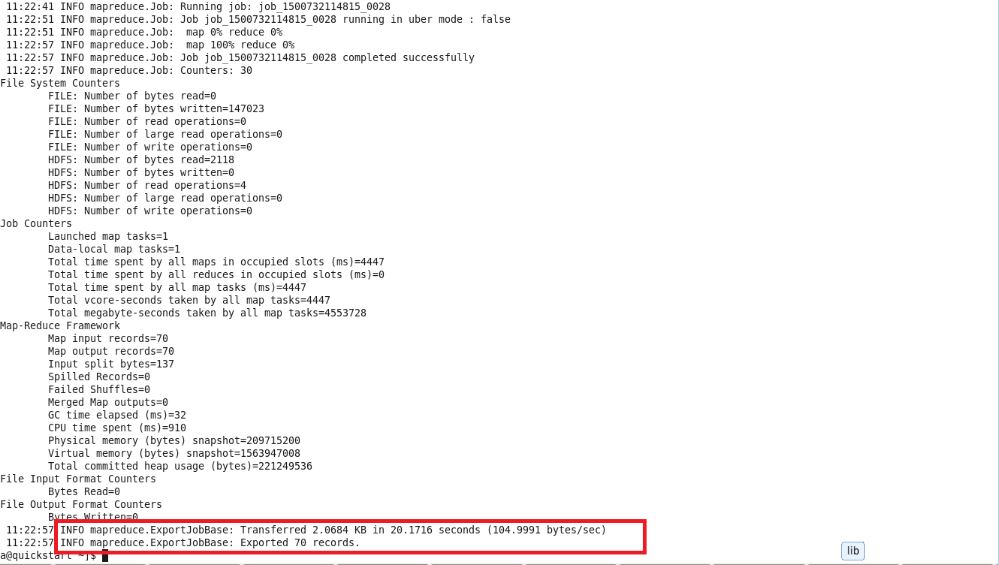
Here we have the mysql table ready to get data from HDFS using SQOOP.

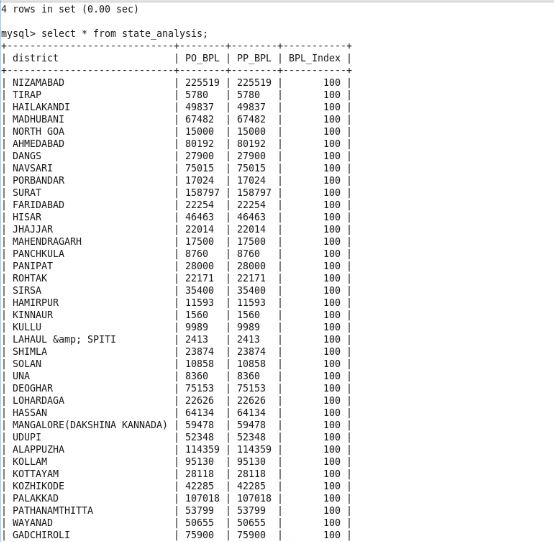


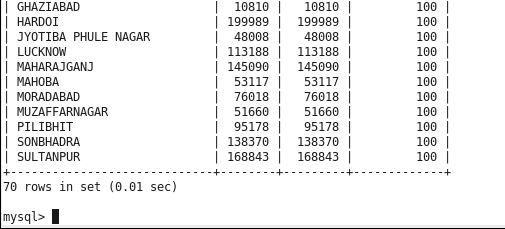
SQOOP command:



Exported records are stored in MYSQL state\_analysis table.





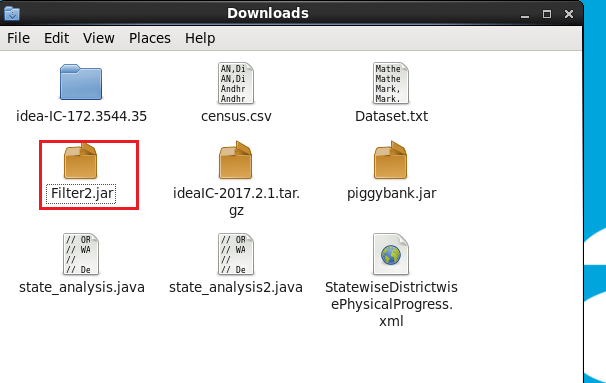
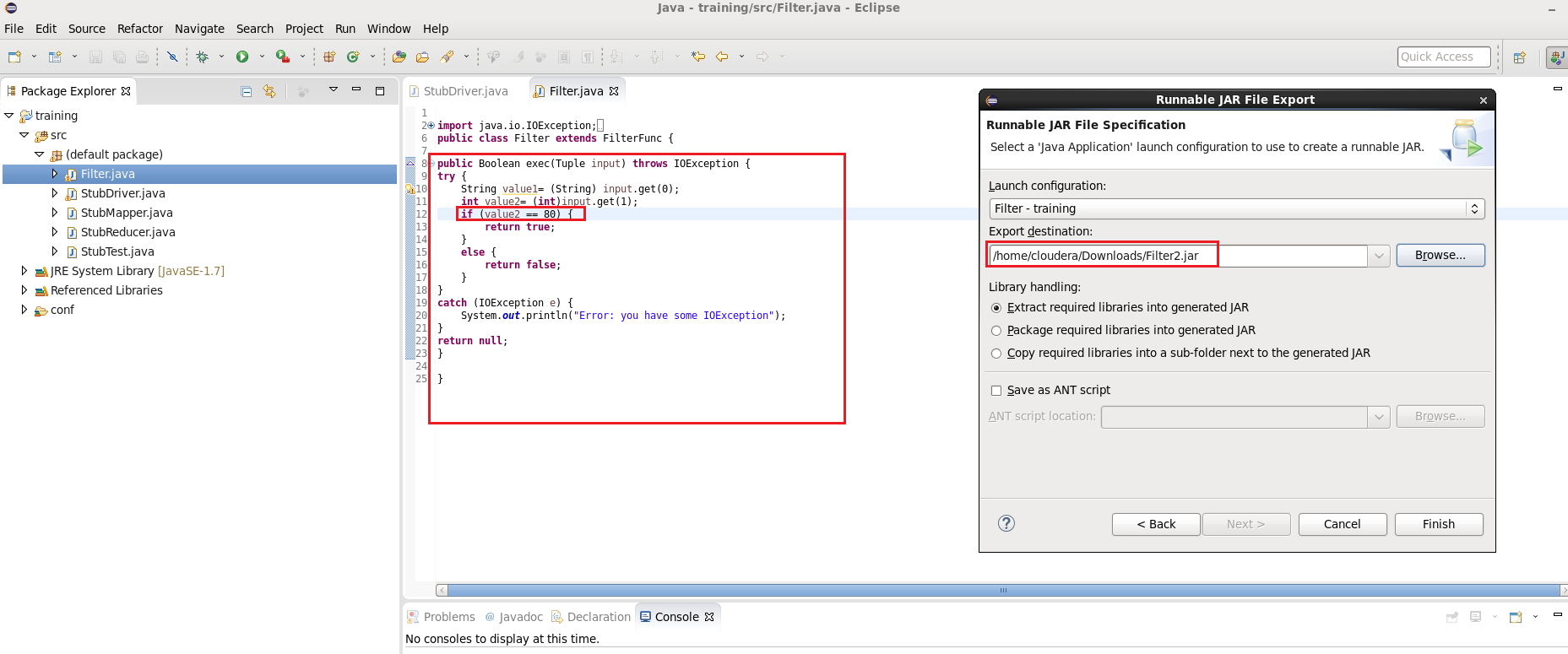


The districts that achieved 100 percent BPL score are exported from HDFS to MYSQL successfully.

2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using SQOOP.

Since some custom functionalities are not built in pig. So we need to create some UDFs for our own purpose.

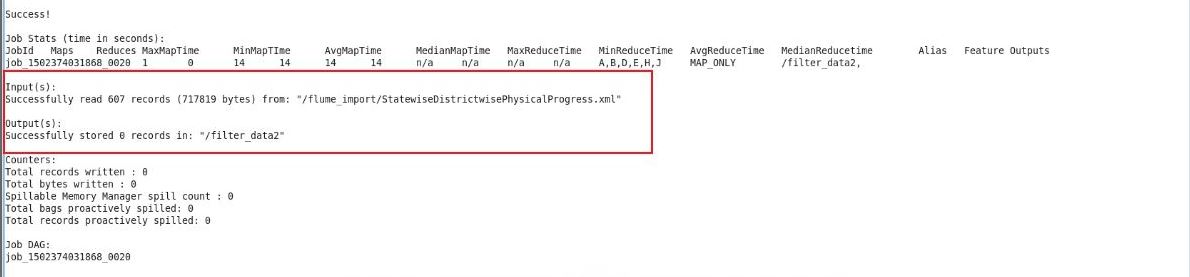
***Districts with BPL\_score = 80%***



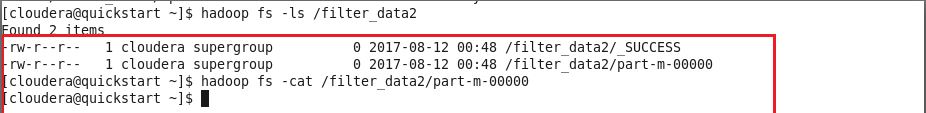
***Commands:***

*grunt> Register ‘/home/cloudera/Downloads/Filter2.jar’;  
  
grunt> J = FOREACH G FILTER($0, $1);*

*grunt> STORE J into ‘/filter\_data2’ using org.apache.pig.piggybank.storage.CSVExcelStorage();*

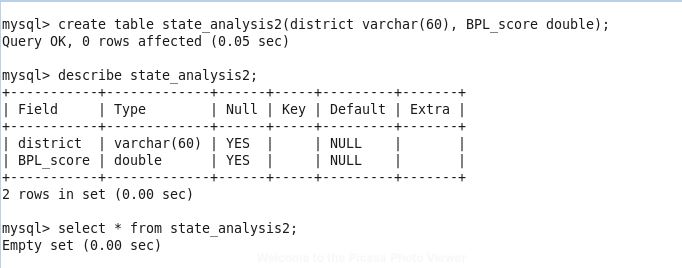


Now we have the exported result within HDFS location.

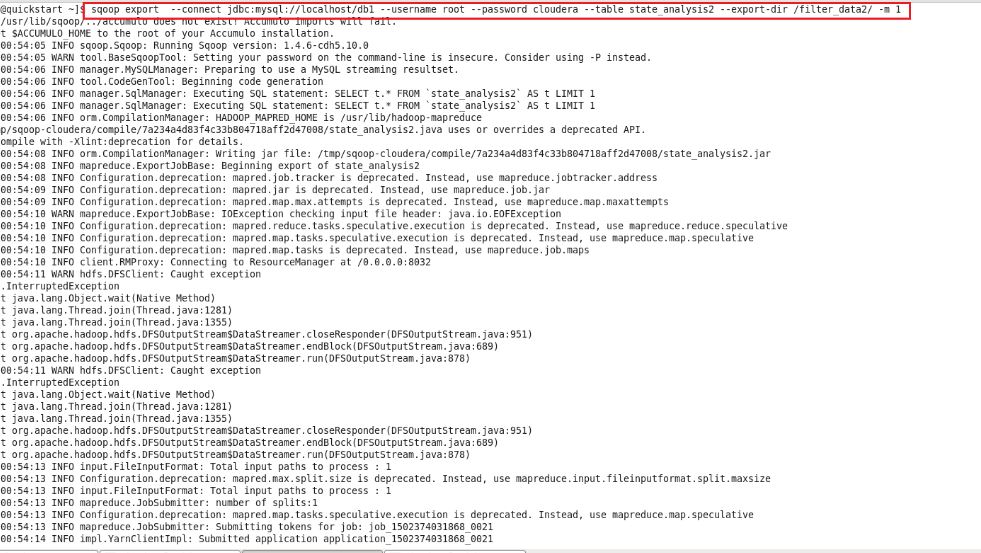


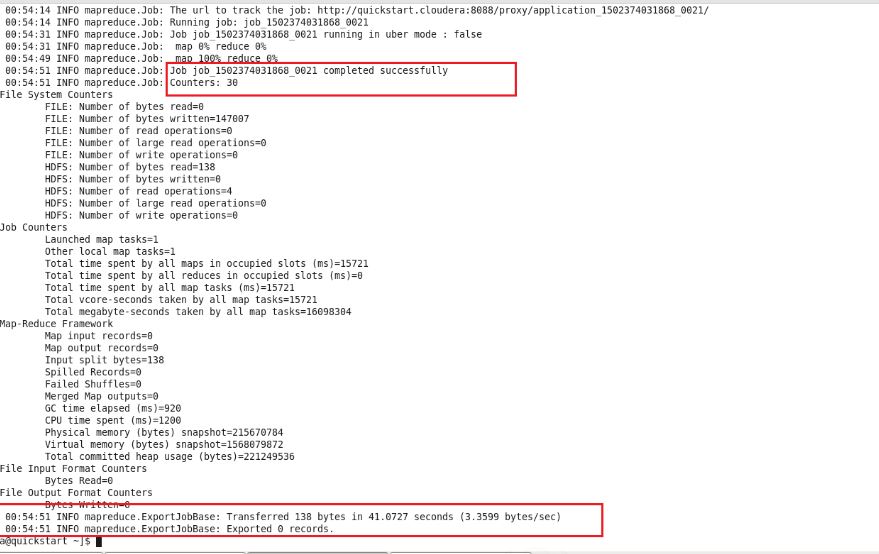
Here we can see there are zero records in folder ‘filter\_data2’ because there are zero districts that achieved 80% BPL score.

In MYSQL before SQOOP transfer:



Now Using SQOOP command we can export the results into MYSQL.

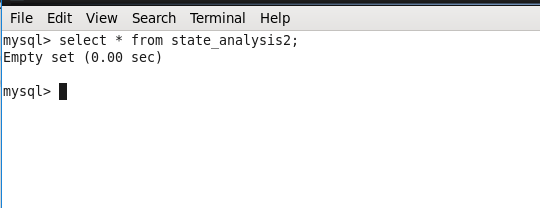




In MYSQL stored results are:

There are Zero districts that reached 80% BPL\_Score.

SQL Table after Sqoop import.



***Conclusion:***

***For data parsing and analyzing we used pig.***

***For data transfer between HDFS and MYSQL we used SQOOP.***