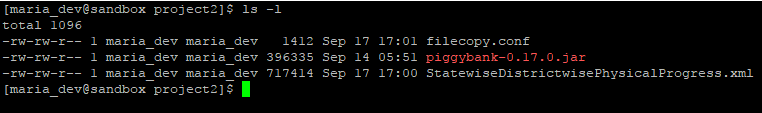
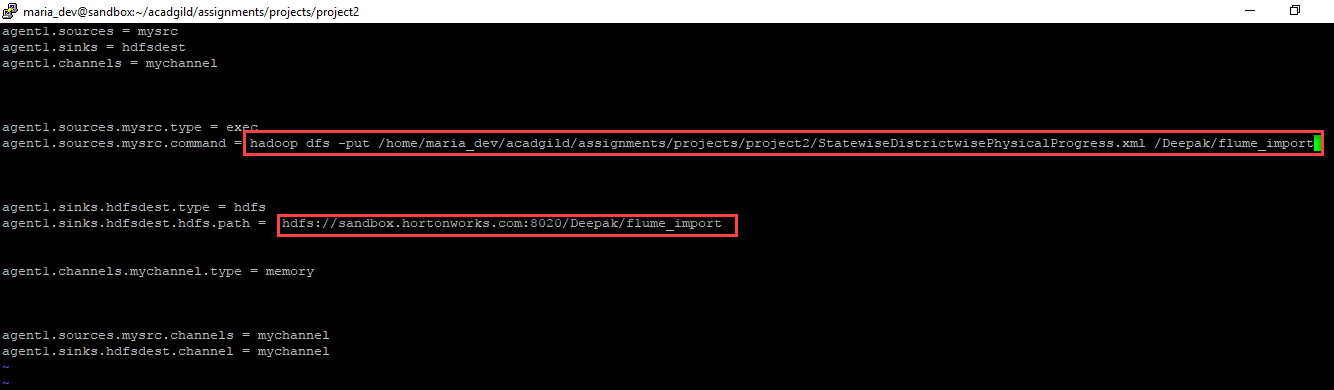
# Project 2 – State wise development analysis

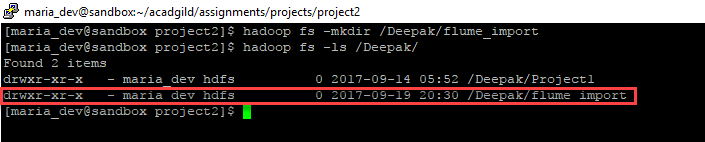
Place the files filecopy.conf and the xml file in one directory



Now, edit the required paths

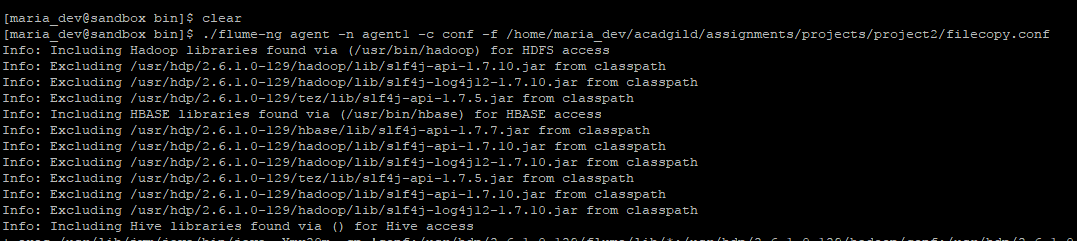


Create a directory flume\_import in HDFS

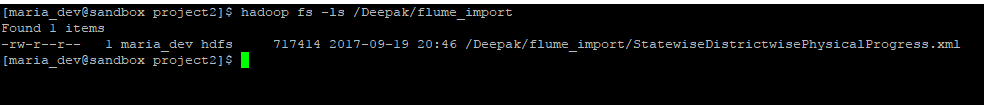


Now, navigate to FLUME\_HOME and execute the following command

./flume-ng agent –n agent1 –c conf –f /home/deepakraybigdata/acadgild/projects/filecopy.conf



Now, check the directory flume\_import in HDFS



We can see that the xml file has been imported from local file system.

**Task 1**

Find out the districts who achieved 100 percent objective in BPL cards

Export the results to mysql using sqoop

Step 1:

Create following directory structure in hdfs

hdfs dfs -mkdir /Deepak/Project2

hdfs dfs -mkdir /Deepak/Project2/Task1

hdfs dfs -mkdir /Deepak/Project2/output

Step 2:

Place the following pig script and piggybank.jar in /Deepak/Project2/Task1

--############################################################

-- Developed By: Deepak Ray

-- Date : 19/09/2017

-- Descr : Pig Script to list down all the districts

-- who achieved 100% objective in BPL cards

--###########################################################

REGISTER piggybank-0.17.0.jar

DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();

-- Load XML data

xmldata = LOAD 'hdfs://sandbox.hortonworks.com:8020/Deepak/flume\_import/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

-- Fetch the required columns using Xpath

xmldata\_getcolumns = FOREACH xmldata GENERATE XPath(x, 'row/District\_Name'),(int) XPath(x, 'row/Project\_Objectives\_IHHL\_BPL'),(int) XPath(x, 'row/Project\_Performance-IHHL\_BPL');

-- Filter the rows where performed number is greater than objective number

xmldata\_filter = FILTER xmldata\_getcolumns by $2 >= $1 ;

-- Store the output into a csv file

STORE xmldata\_filter INTO 'hdfs://sandbox.hortonworks.com:8020/Deepak/Project2/Task1/output' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','WINDOWS');



Step 4:

Execute the pig script



exec project2\_1.pig

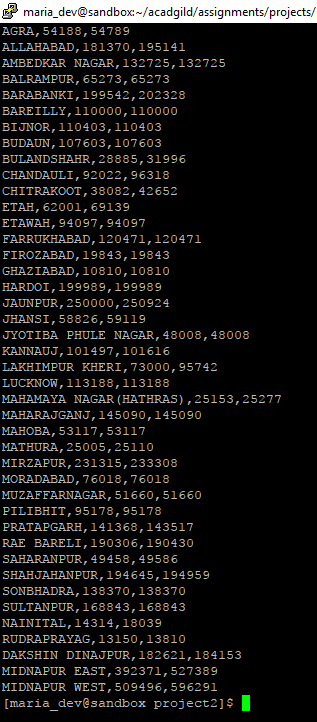


Step 5:

Check the output

hdfs dfs -cat /Deepak/Project2/Task1/output/part-v000-o000-r-00000

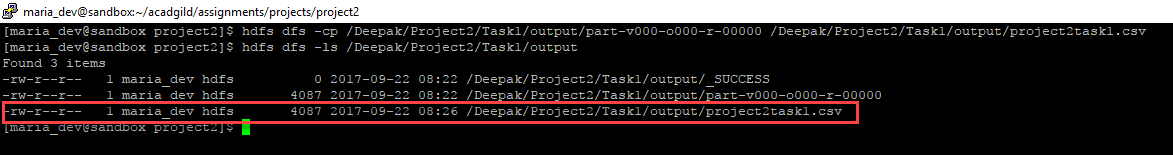




Step 6:

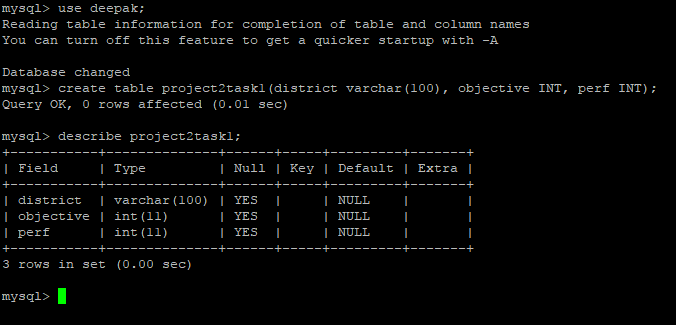
Create output file project2task1.csv from the original output file

hdfs dfs -cp /Deepak/Project2/Task1/output/part-v000-o000-r-00000 /Deepak/Project2/Task1/output/project2task1.csv



Step 7:

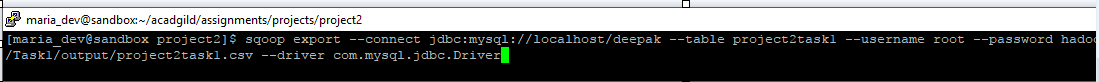
Connect to mysql and create a table to hold the output data from file project2task1.csv

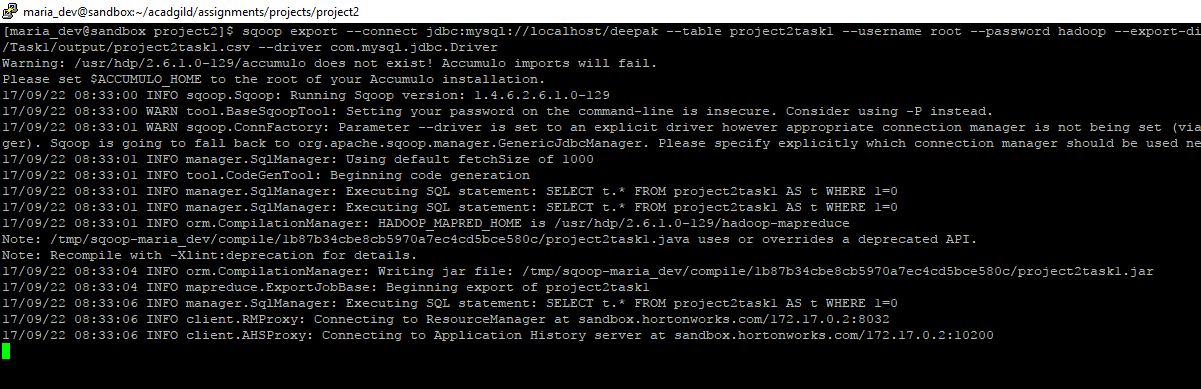


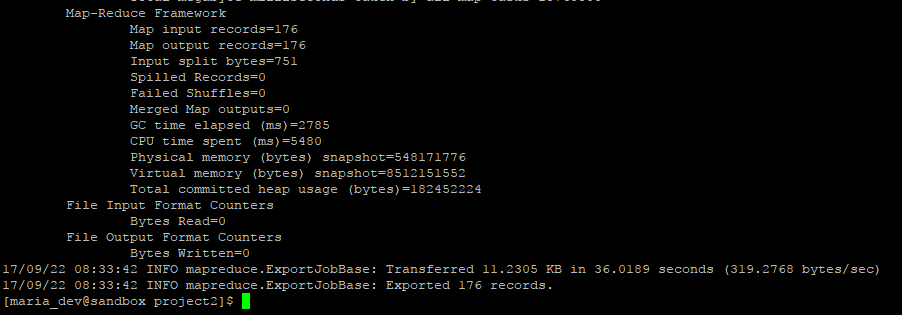
Step 8:

Export the data from /Deepak/Project2/Task1/project2task1.csv to the mysql table created above

sqoop export --connect jdbc:mysql://localhost/deepak --table project2task1 --username root --password hadoop --export-dir /Deepak/Project2/Task1/output/project2task1.csv --driver com.mysql.jdbc.Driver







Now, check the data in the table in mysql.

