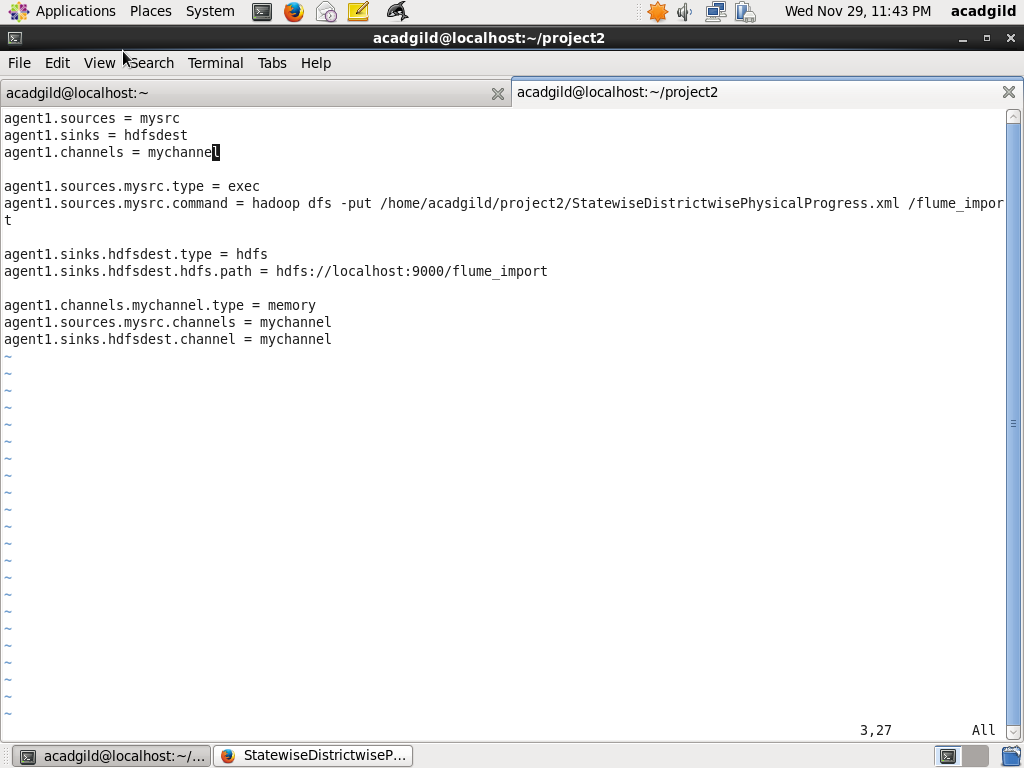
Step1:

Create Flume Config file filecopy.conf as below and copy to folder /home/acadgild/project2

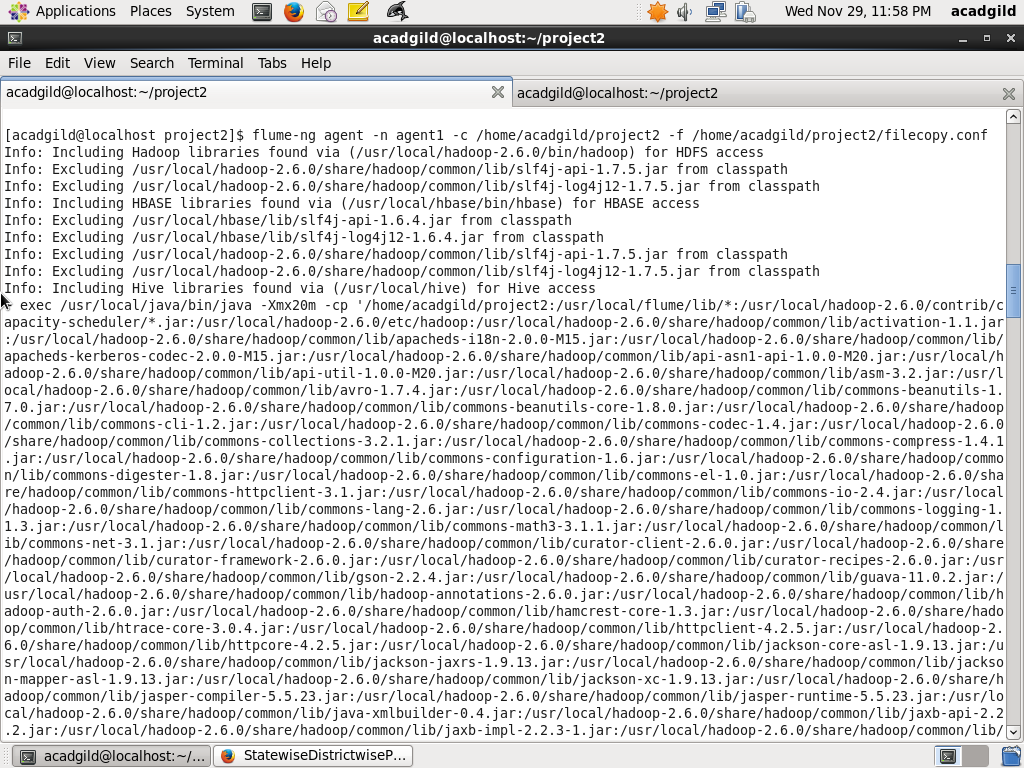
Screenshot is as below:

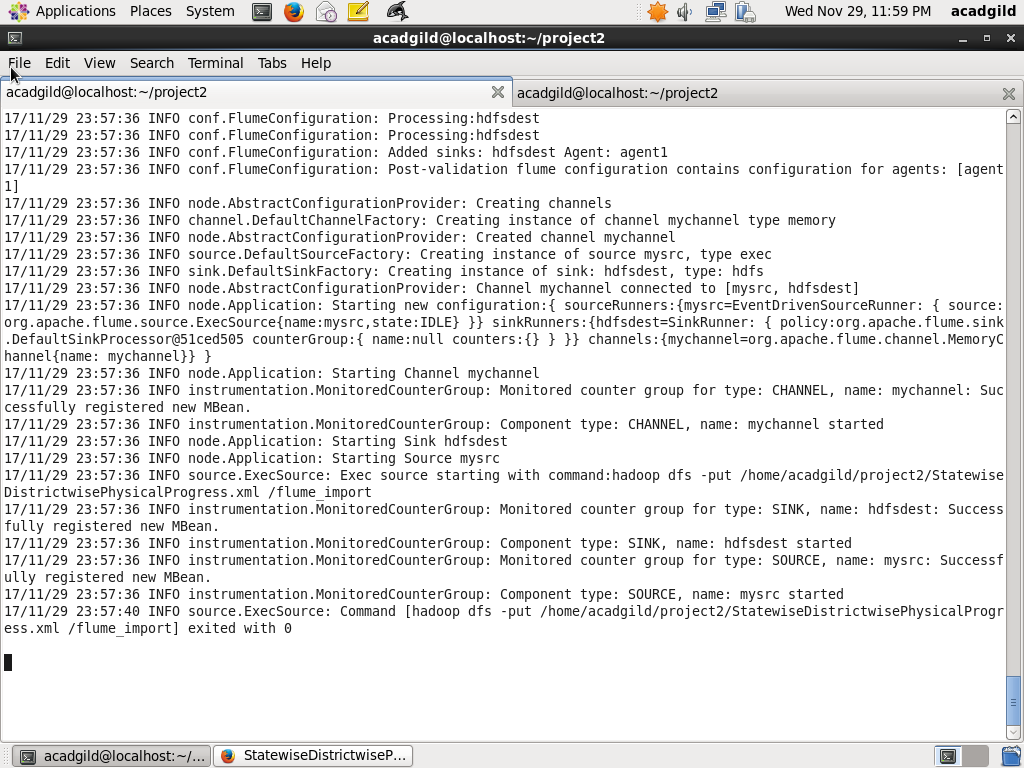


Step 2: Copy dataset from local file system to HDFS using flume.

flume-agent agent –n agent1 –c /home/acadgild –f /home/acadgild/project2/filecopy.conf

Screenshot is as below:

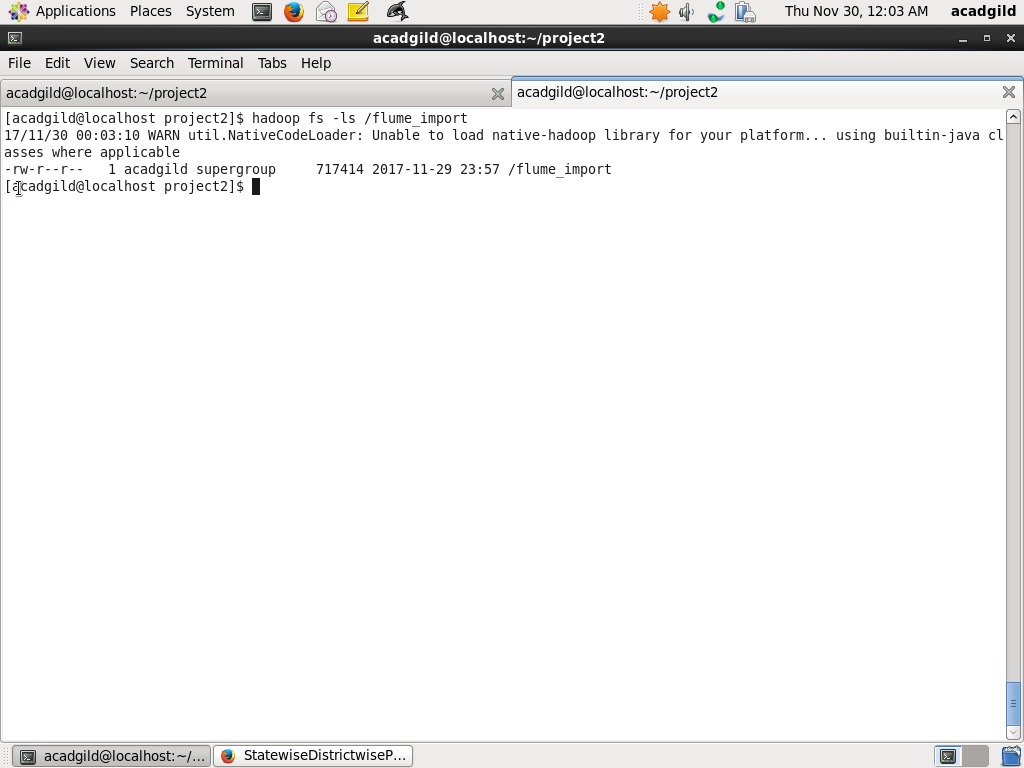




Step 3: Verify that file is copied using the HDFS commands below

The command below confirms the directory is created

hadoop fs –ls /flume\_import



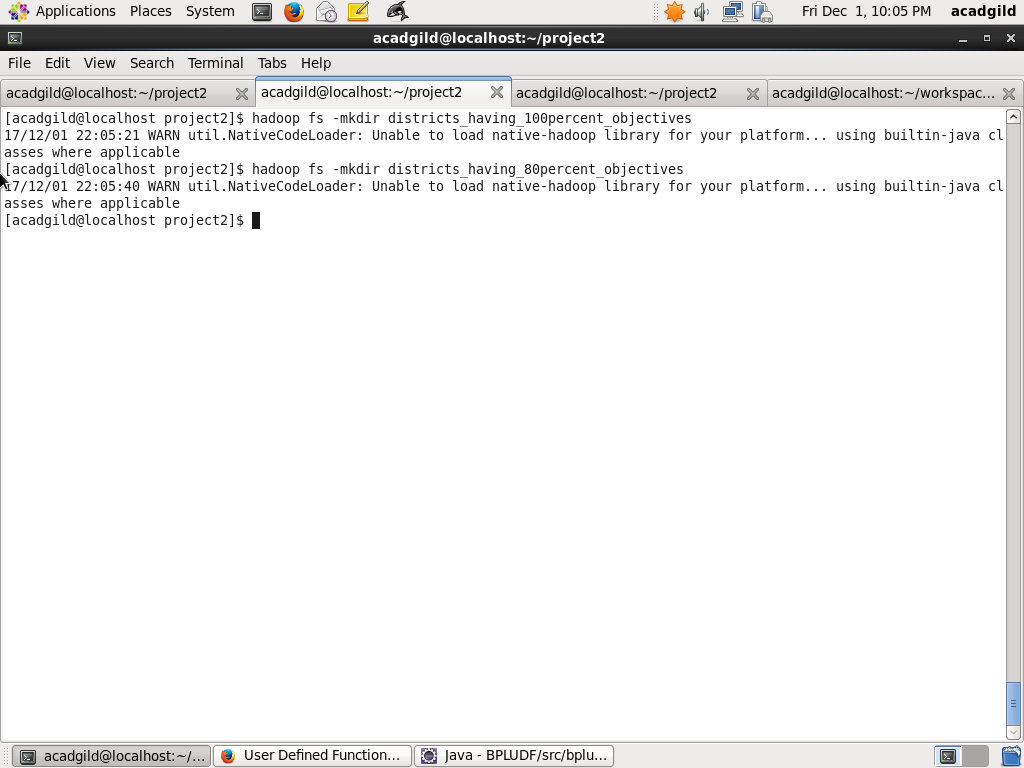
Step4: Create folders in HDFS to store query outputs

Create two folders districts\_having\_100percent\_objectives and districts\_having\_80percent\_objectives in HDFS.

hadoop fs –mkdir folders districts\_having\_100percent\_objectives

hadoop fs –mkdir folders districts\_having\_80percent\_objectives

Screenshots are as below:



Step5: Create mysql table to store the results of query

Start mysql using command

sudo service mysqld start

Login to mysql using

mysql –u root

Create Database bpl\_results

create database bpl\_results

use bpl\_results

Create tables districts\_having\_100percent\_objectives and districts\_having\_80percent\_objectives as below:

create table districts\_having\_100percent\_objectives

(

name varchar(40)

);

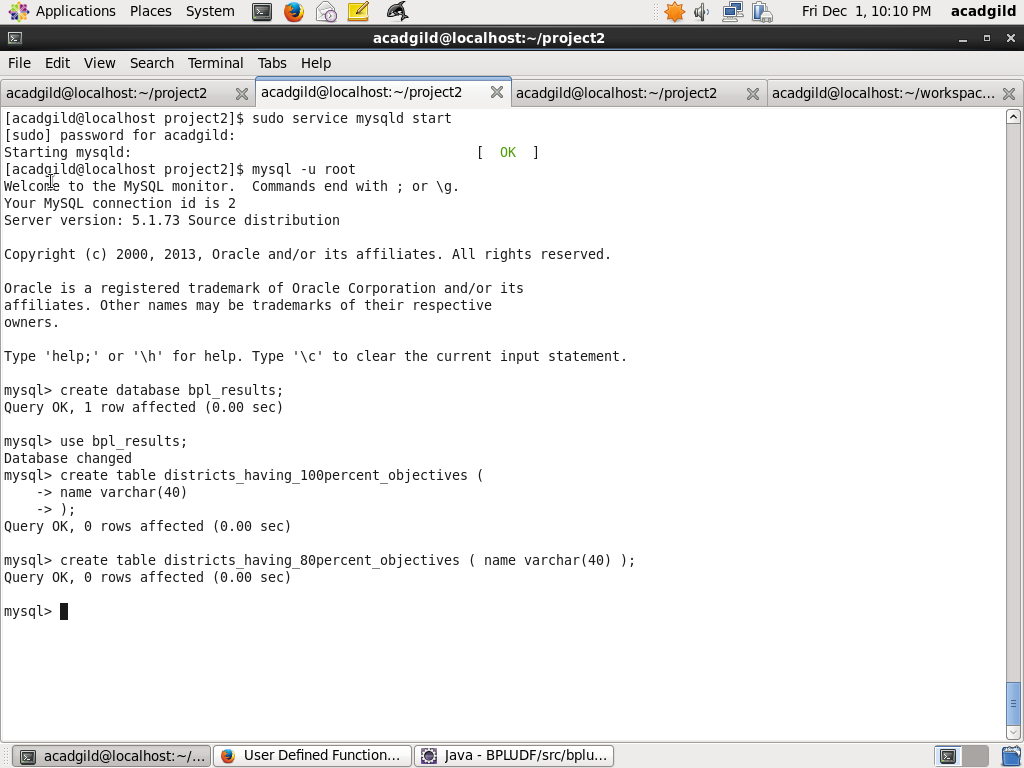
create table districts\_having\_80percent\_objectives

(

name varchar(40)

);

Screenshot is as below:



Step6: PIG query to process XML and store into PIG table

From HDFS LOAD to PIG alias row\_physical\_progress using XMLLoader as below

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

row\_physical\_progress = LOAD 'hdfs://localhost:9000/flume\_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (row:chararray);

Next, iterate over each row and load into alias physical\_progress which has schema fields same as XML schema hyphen(-) are replaced with underscore (\_)

physical\_progress = FOREACH row\_physical\_progress GENERATE XPath(row, 'row/State\_Name') AS State\_Name,

XPath(row, 'row/District\_Name') AS District\_Name,

XPath(row, 'row/Project\_Objectives\_IHHL\_BPL') AS Project\_Objectives\_IHHL\_BPL,

XPath(row, 'row/Project\_Objectives\_IHHL\_APL') AS Project\_Objectives\_IHHL\_APL,

XPath(row, 'row/Project\_Objectives\_IHHL\_TOTAL') AS Project\_Objectives\_IHHL\_TOTAL,

XPath(row, 'row/Project\_Objectives\_SCW') AS Project\_Objectives\_SCW,

XPath(row, 'row/Project\_Objectives\_Anganwadi\_Toilets') AS Project\_Objectives\_Anganwadi\_Toilets,

XPath(row, 'row/Project\_Objectives\_RSM') AS Project\_Objectives\_RSM,

XPath(row, 'row/Project\_Objectives\_PC') AS Project\_Objectives\_PC,

XPath(row, 'row/Project\_Performance-IHHL\_BPL') AS Project\_Performance\_IHHL\_BPL,

XPath(row, 'row/Project\_Performance-IHHL\_APL') AS Project\_Performance\_IHHL\_APL,

XPath(row, 'row/Project\_Performance-IHHL\_TOTAL') AS Project\_Performance\_IHHL\_TOTAL,

XPath(row, 'row/Project\_Performance-SCW') AS Project\_Performance\_SCW,

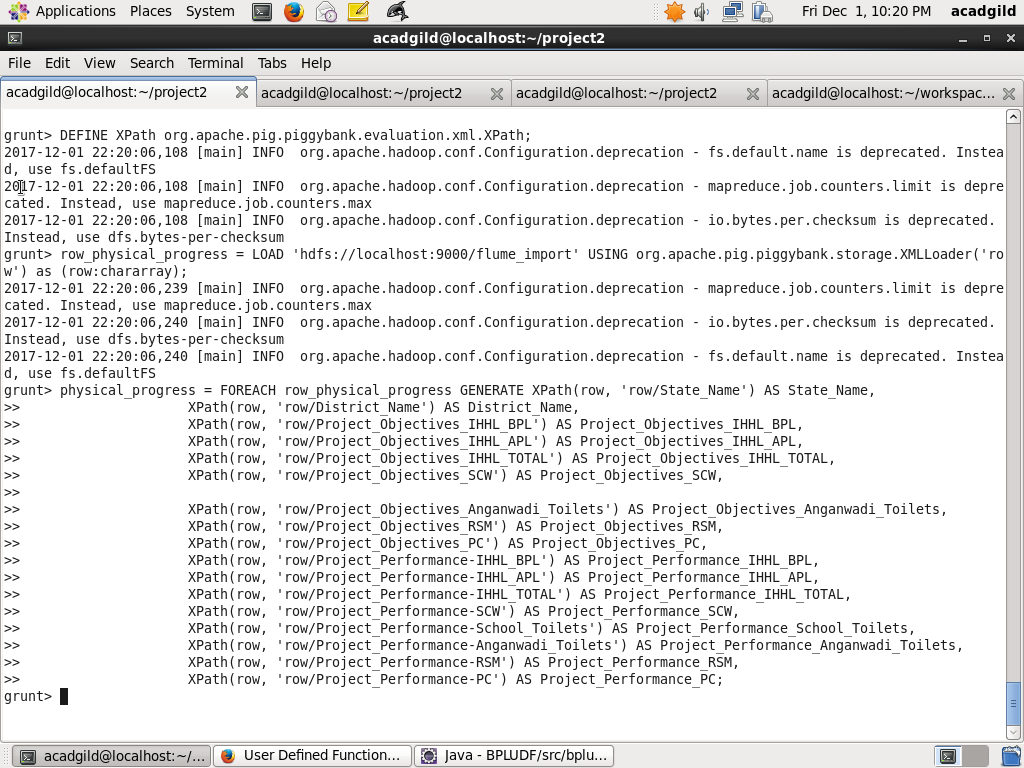
XPath(row, 'row/Project\_Performance-School\_Toilets') AS Project\_Performance\_School\_Toilets,

XPath(row, 'row/Project\_Performance-Anganwadi\_Toilets') AS Project\_Performance\_Anganwadi\_Toilets,

XPath(row, 'row/Project\_Performance-RSM') AS Project\_Performance\_RSM,

XPath(row, 'row/Project\_Performance-PC') AS Project\_Performance\_PC;

Screenshot is as below:



Step7: PIG Query to find out the districts who achieved 100 percent objective in BPL cards

Here first filter the records where Project\_Objectives\_IHHL\_BPL is equal to Project\_Performance\_IHHL\_BPL

physical\_progress\_100\_percent\_bpl = FILTER physical\_progress BY Project\_Objectives\_IHHL\_BPL == Project\_Performance\_IHHL\_BPL;

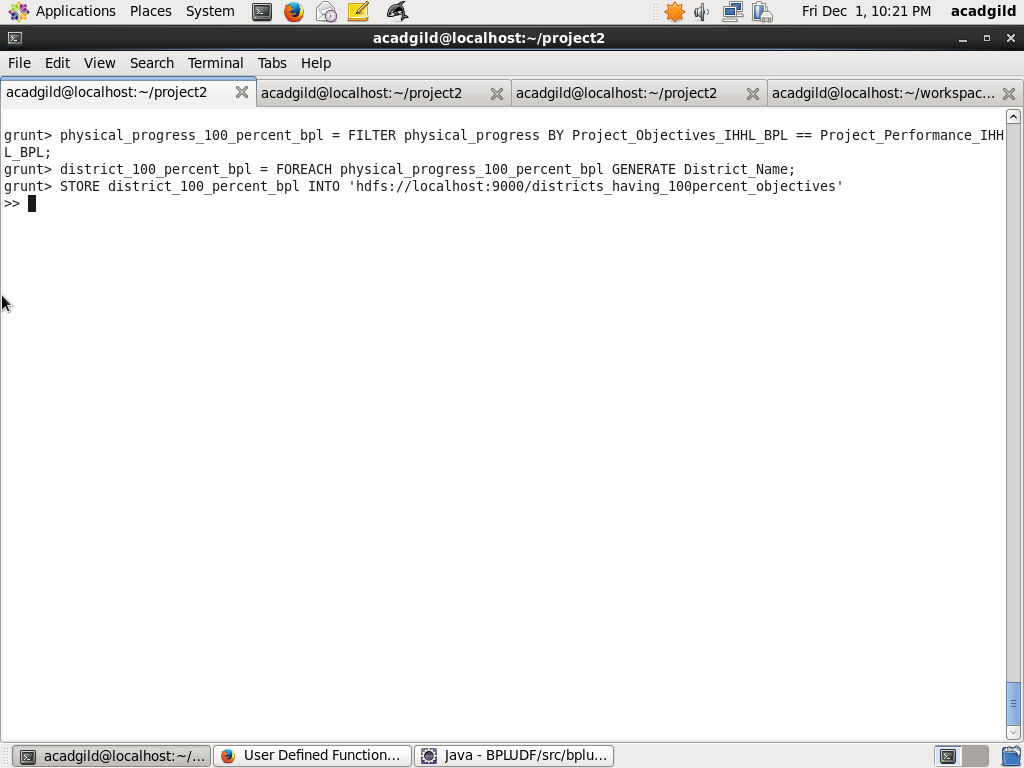
Next, Select only District\_Name field using command below:

district\_100\_percent\_bpl = FOREACH physical\_progress\_100\_percent\_bpl GENERATE District\_Name;

Next, Store into HDFS directory districts\_having\_100percent\_objectives using command below:

STORE district\_100\_percent\_bpl INTO 'hdfs://localhost:9000/districts\_having\_100percent\_objectives'

Screenshot is as below:



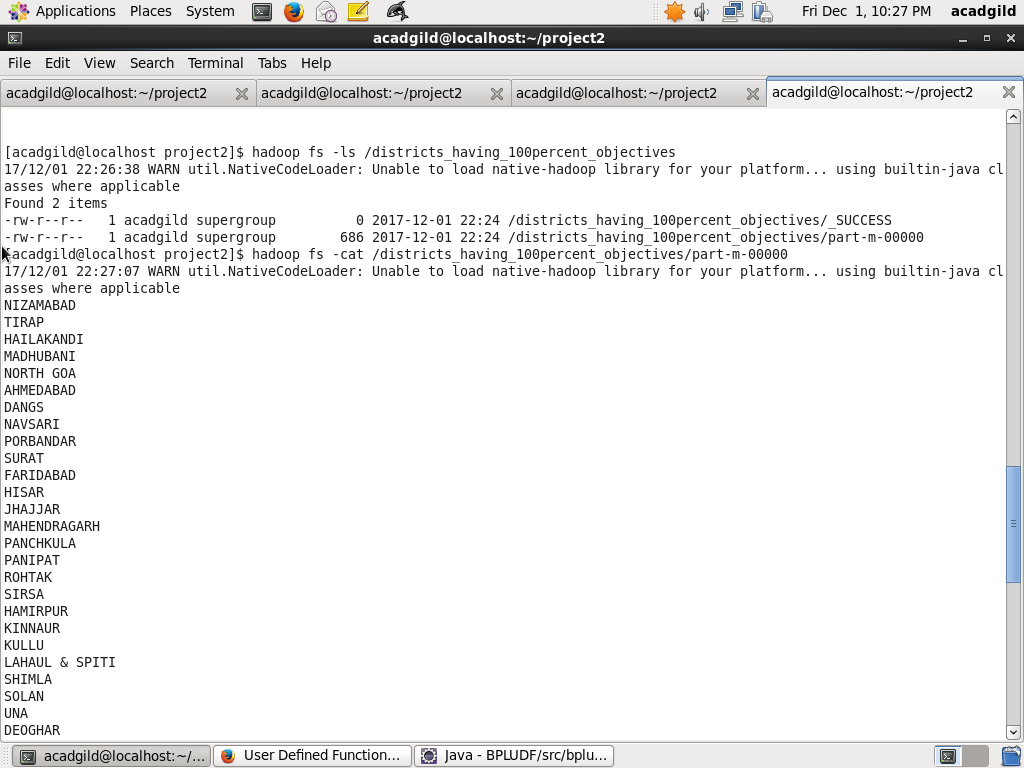
Step8: Verify that results are stored in HDFS

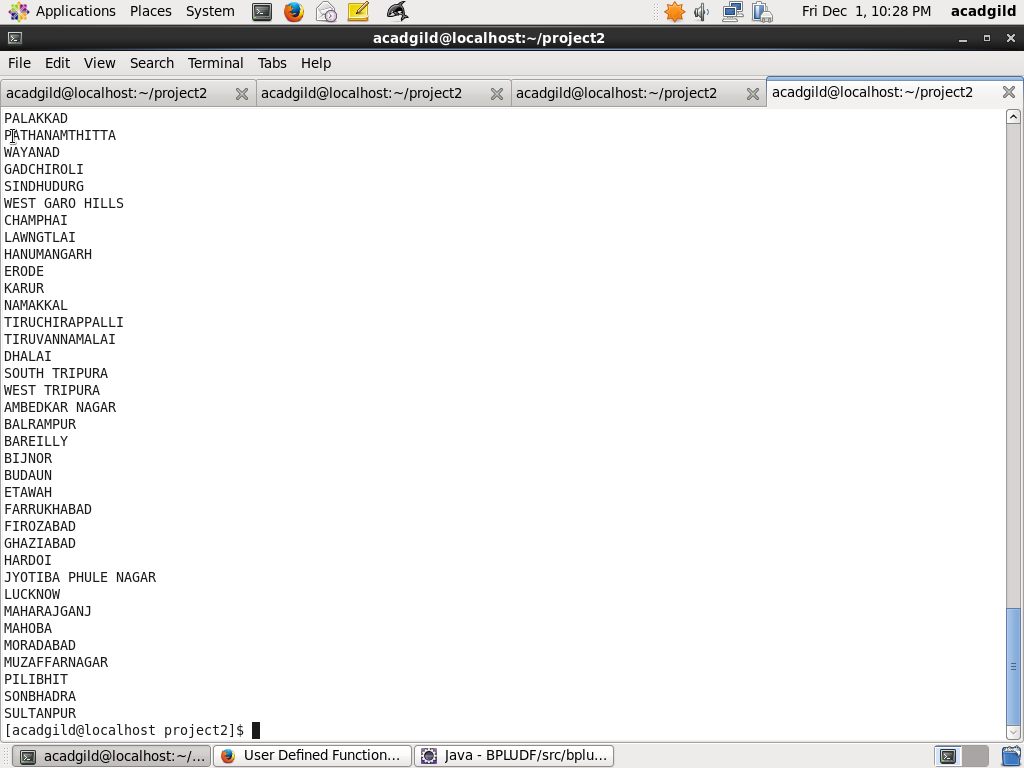
The following command shows that folders are created under districts\_having\_100percent\_objectives

hadoop fs –ls /districts\_having\_100percent\_objectives

Next, use the following HDFS command to show the results

hadoop fs –ls /districts\_having\_100percent\_objectives/part-m-00000



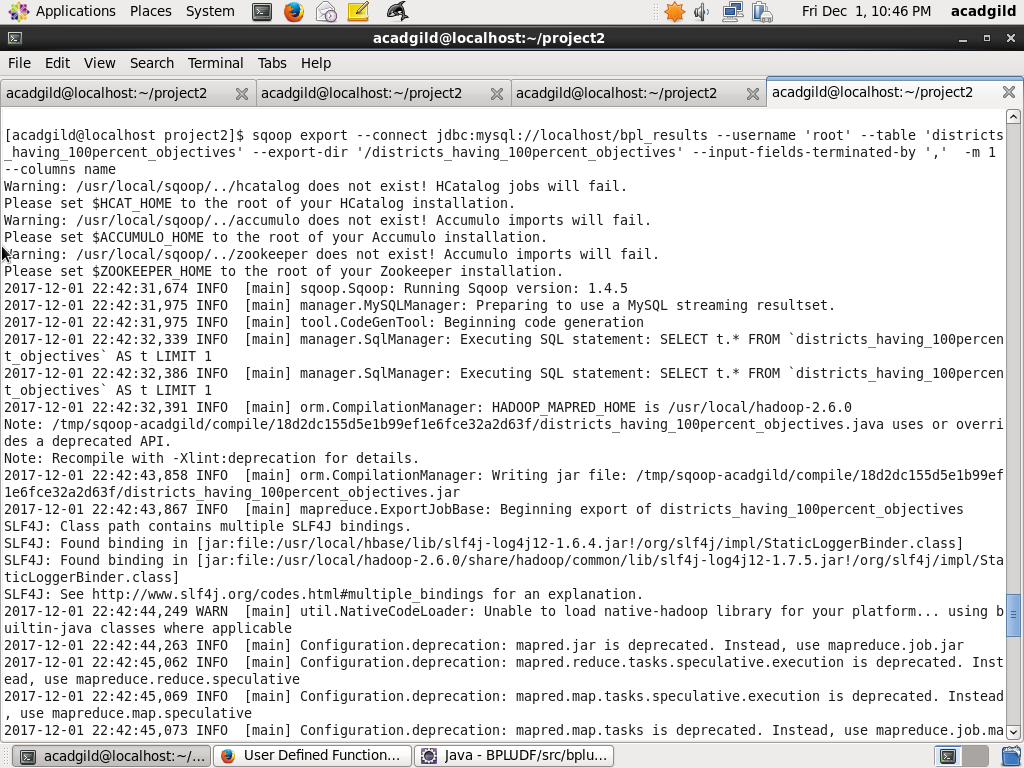


Step9 : Use sqoop command to export data from HDFS into mysql table districts\_having\_100percent\_objectives in database bpl\_results

The following sqoop command is used to export data from HDFS folder districts\_having\_100percent\_objectives into already created mysql table 'districts\_having\_100percent\_objectives

sqoop export --connect jdbc:mysql://localhost/bpl\_results --username 'root' --table 'districts\_having\_100percent\_objectives' --export-dir '/districts\_having\_100percent\_objectives' --input-fields-terminated-by ',' -m 1 --columns name

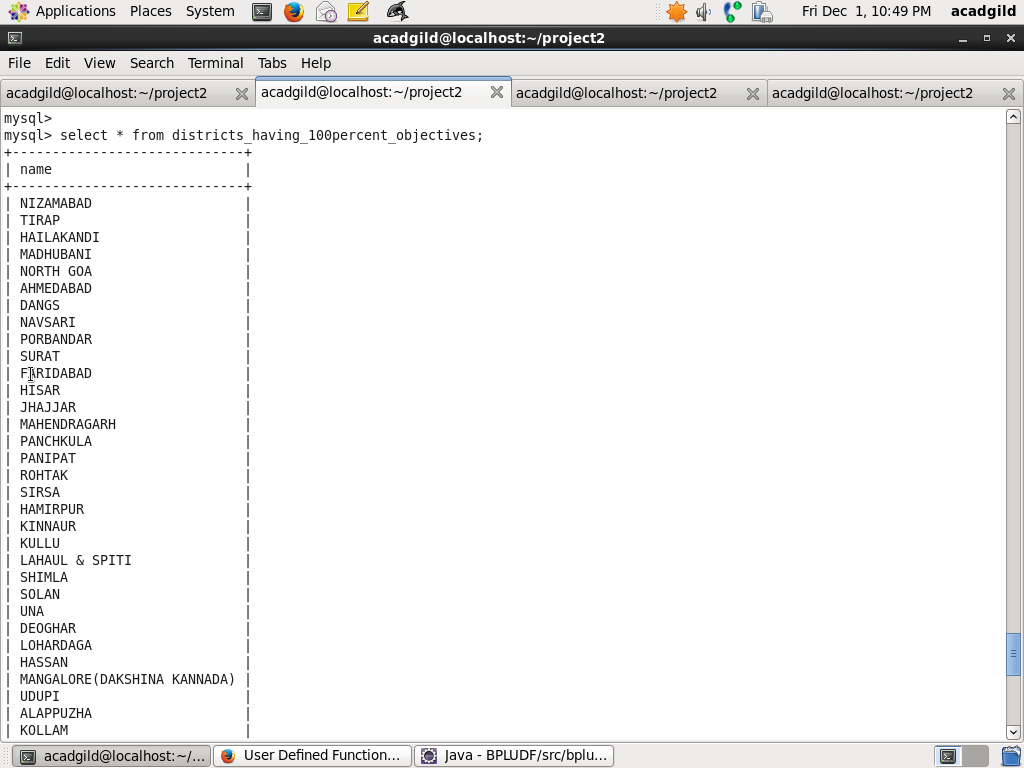
Screenshots are as below:

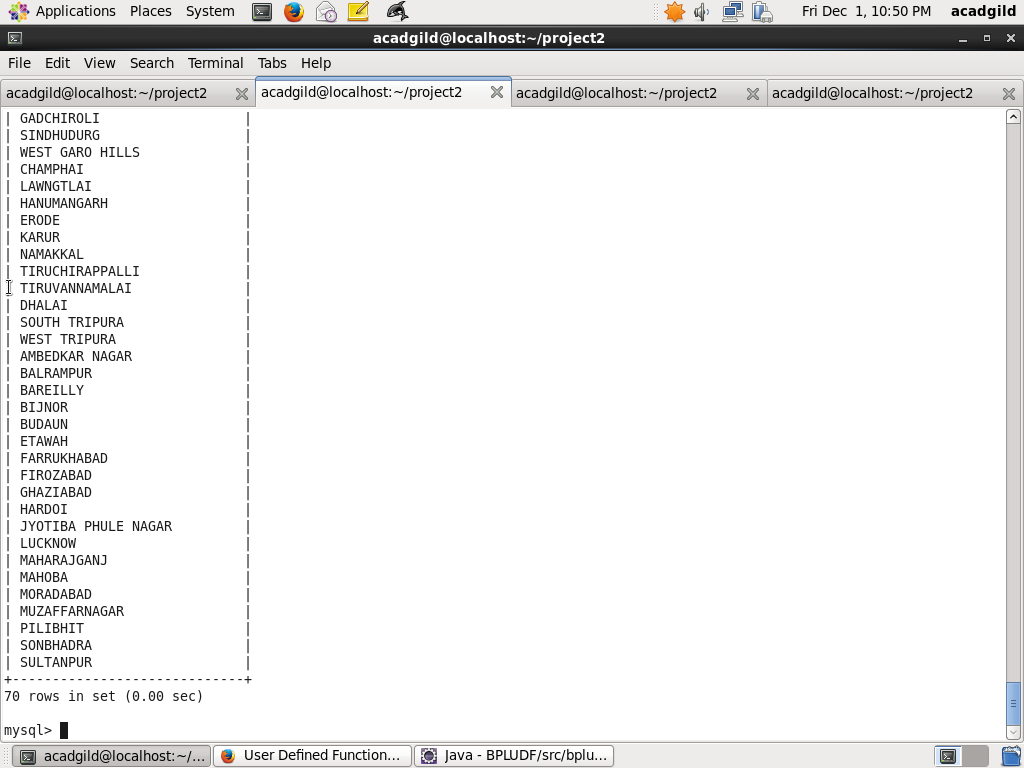


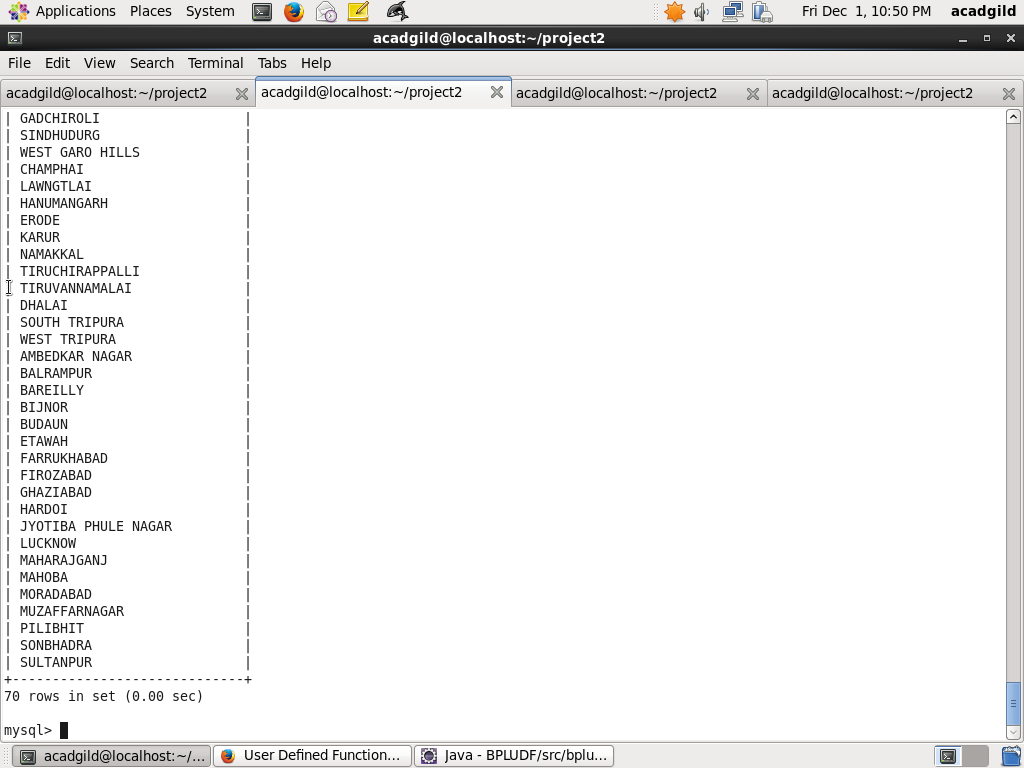
Step10: Verify Result in Mysql

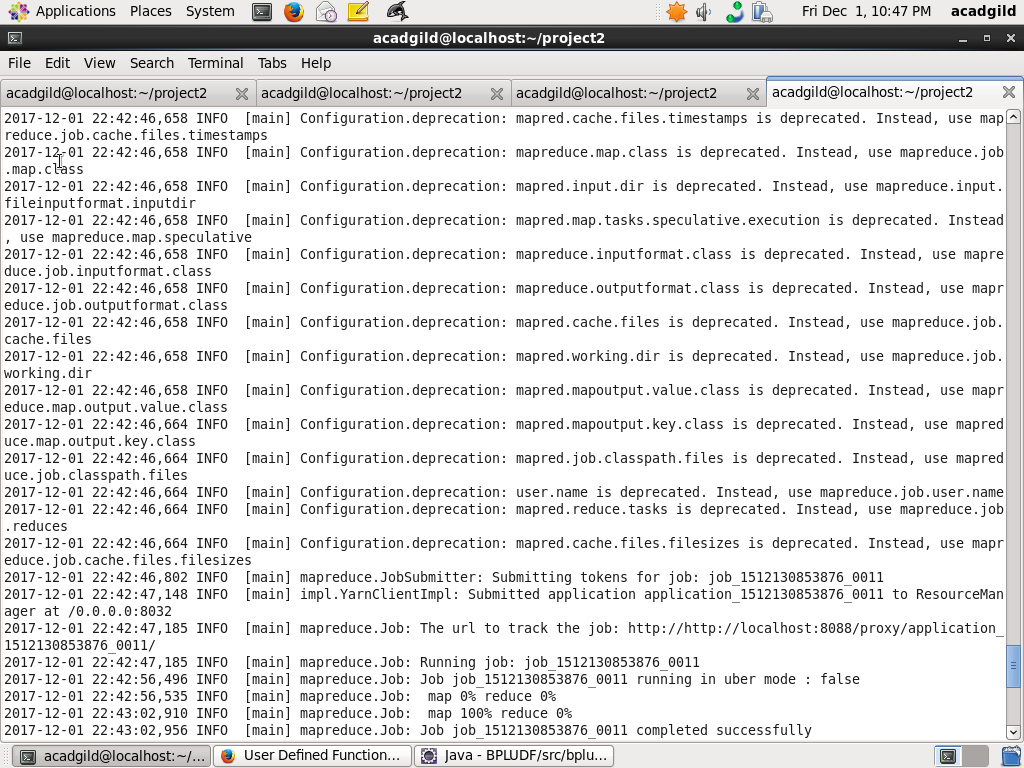
Use the following command in mysql to verify results in mysql

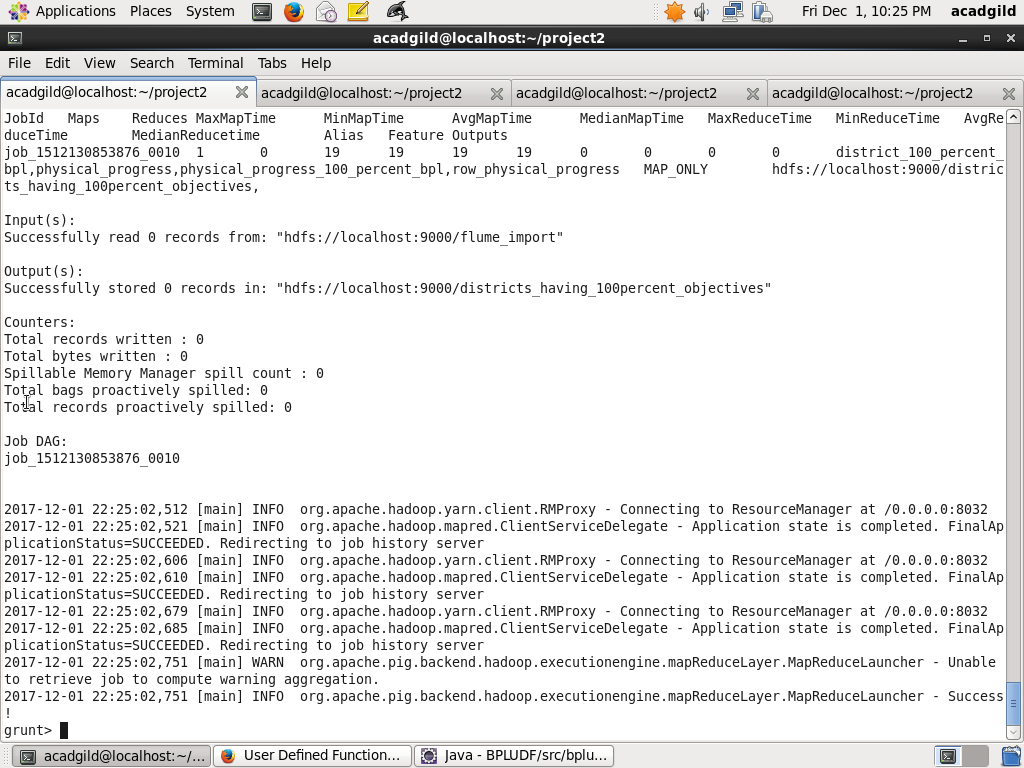
select \* from districts\_having\_100percent\_objectives









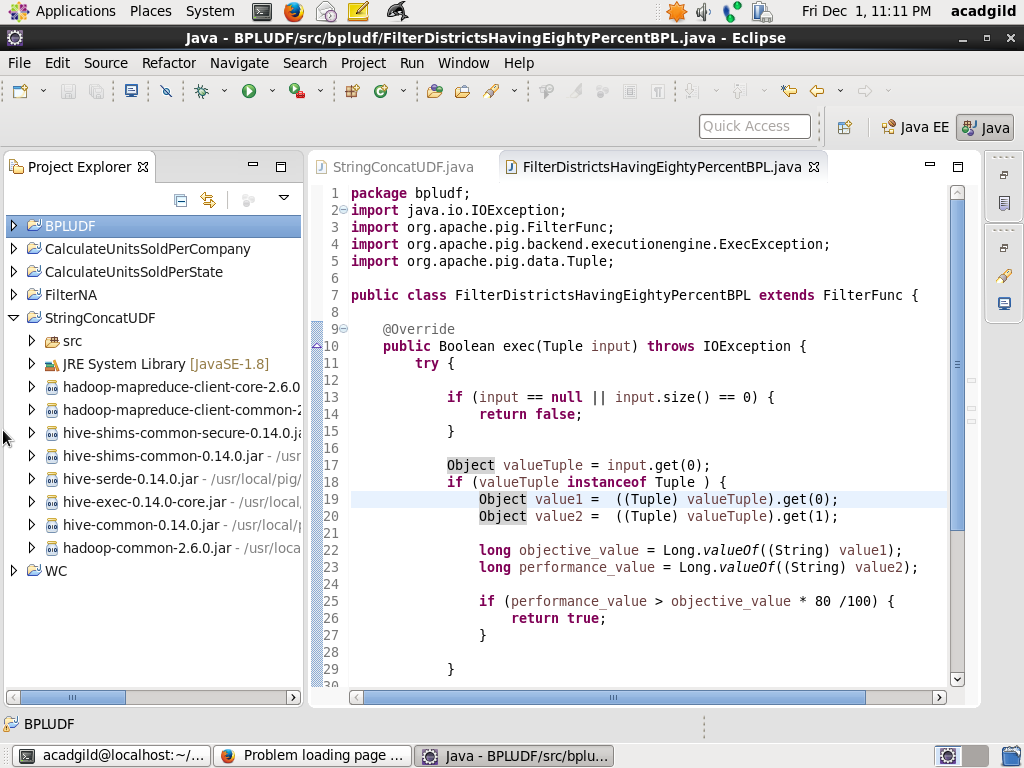


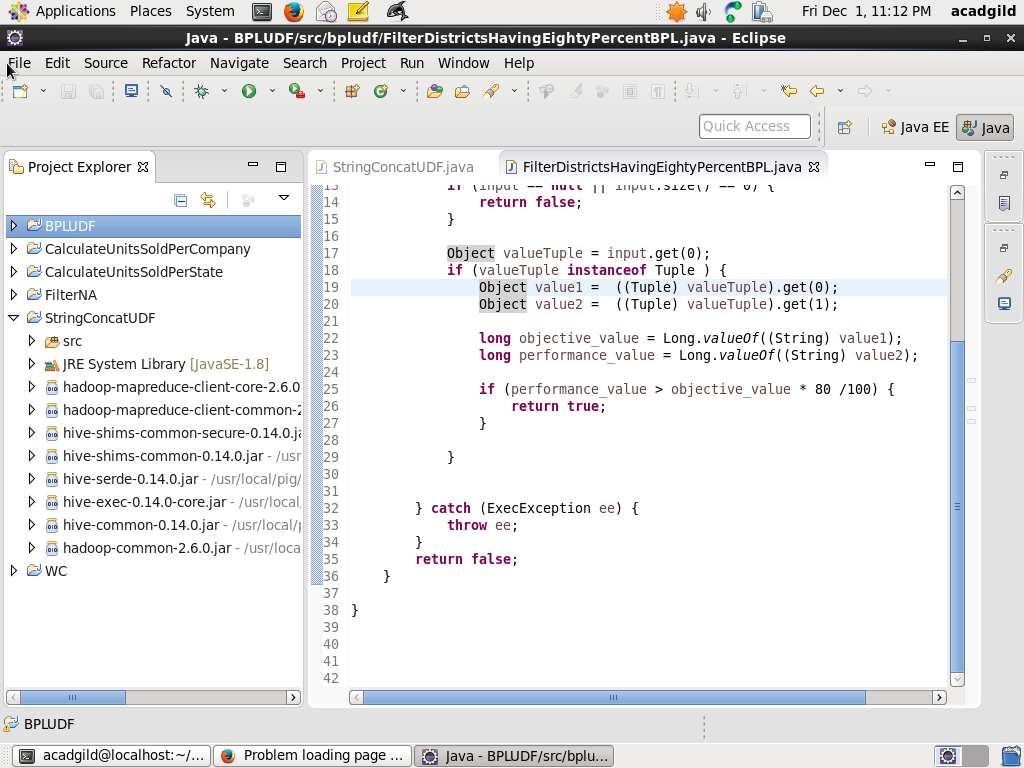
Step11: PIG UDF to find out the districts who achieved 80 percent objective in BPL cards

Create a Java project BPLUDF and Write a Java class FilterDistrictsHavingEightyPercentBPL in eclipse which will filter those tuples for which 80 percent objective in BPL cards are achieved. The logic put in exec method is value of Project\_Performance\_IHHL\_BPL is equal to more than 80% of Project\_Objectives\_IHHL\_BPL.

Export the project to BPLUDF.jar

Screenshot is given as below:





Step12: Write PIG query to find out the districts who achieved 80 percent objective in BPL cards

Register the Jar BPLUDF.jar for the UDF created in step11

REGISTER /home/acadgild/project2/BPLUDF.jar;

Next, using the UDF filter those tuple for which Project\_Performance\_IHHL\_BPL is equal to more than 80% of Project\_Objectives\_IHHL\_BPL

physical\_progress\_80\_percent\_bpl = FILTER physical\_progress BY bpludf.FilterDistrictsHavingEightyPercentBPL(TOTUPLE(Project\_Objectives\_IHHL\_BPL, Project\_Performance\_IHHL\_BPL));

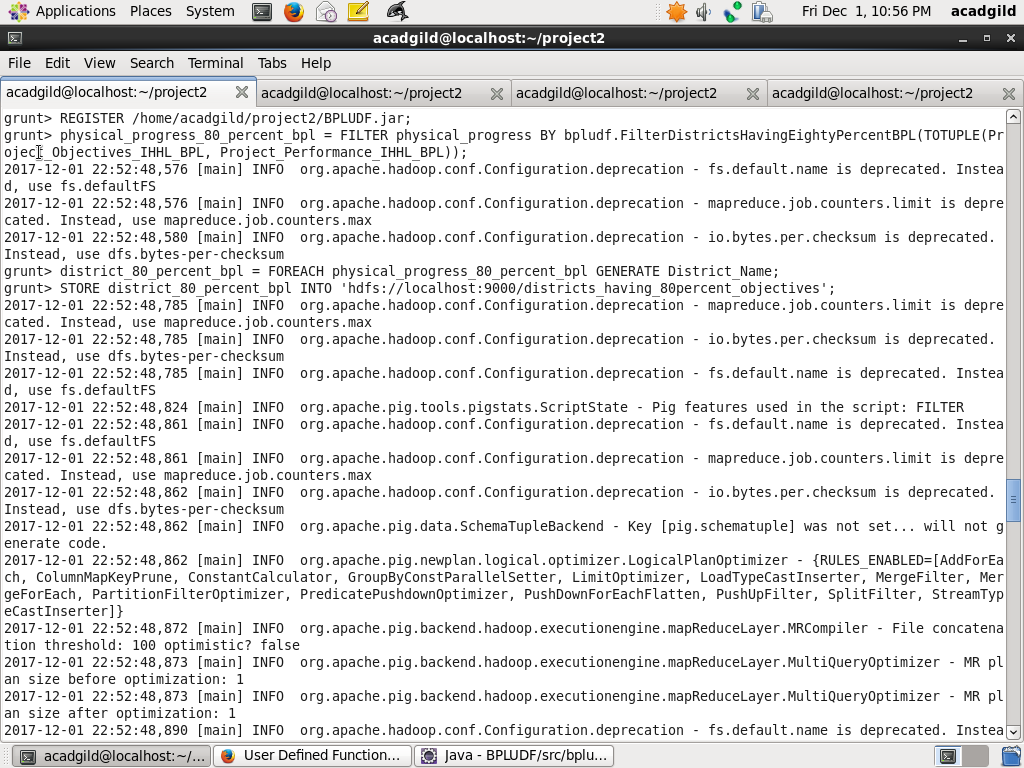
Next, Select only District\_Name field using command below:

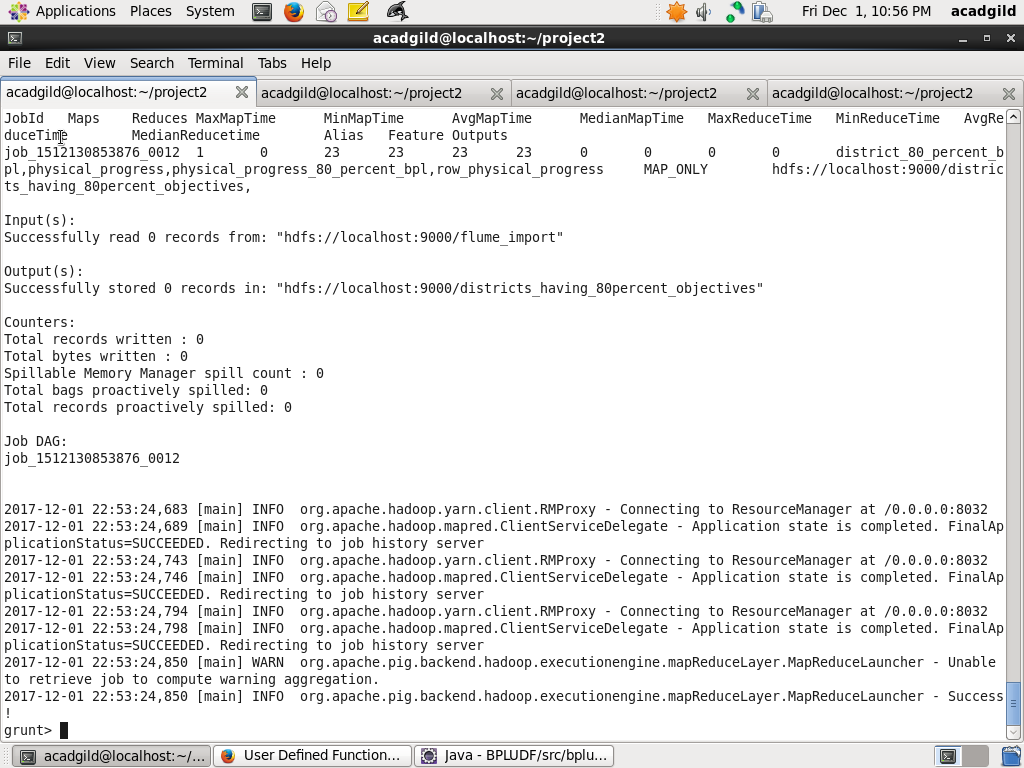
district\_80\_percent\_bpl = FOREACH physical\_progress\_80\_percent\_bpl GENERATE District\_Name;

Next, Store into HDFS directory districts\_having\_100percent\_objectives using command below:

STORE district\_80\_percent\_bpl INTO 'hdfs://localhost:9000/districts\_having\_80percent\_objectives';

Screenshot is as below:





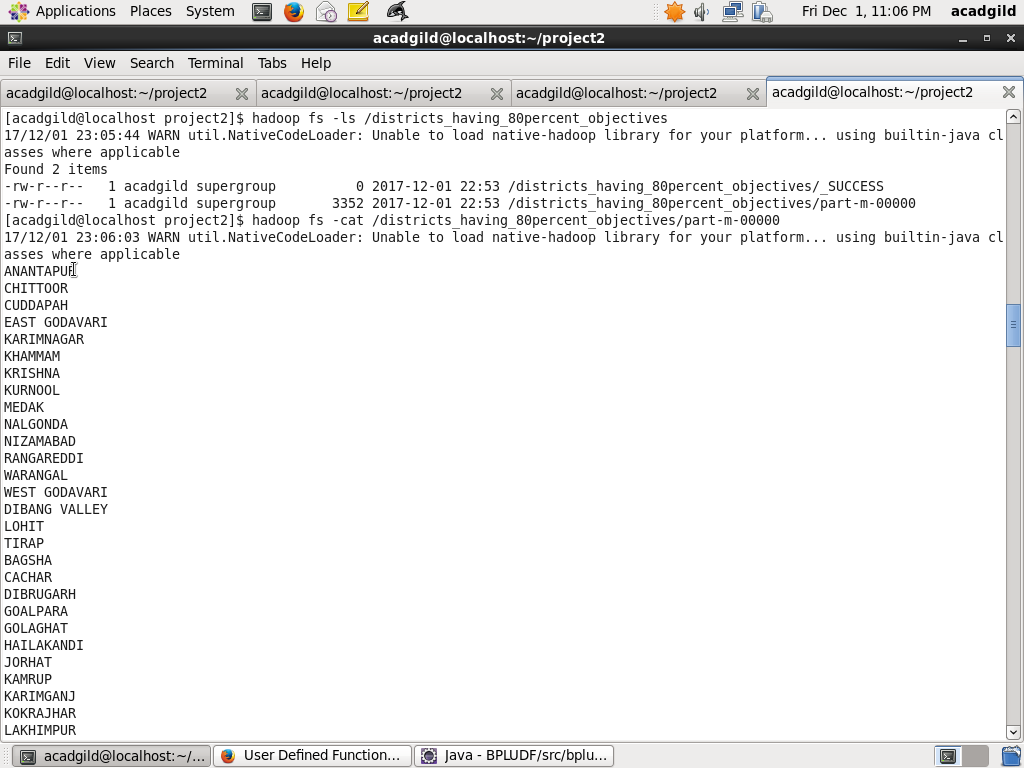
Step13: Verify that results are stored in HDFS

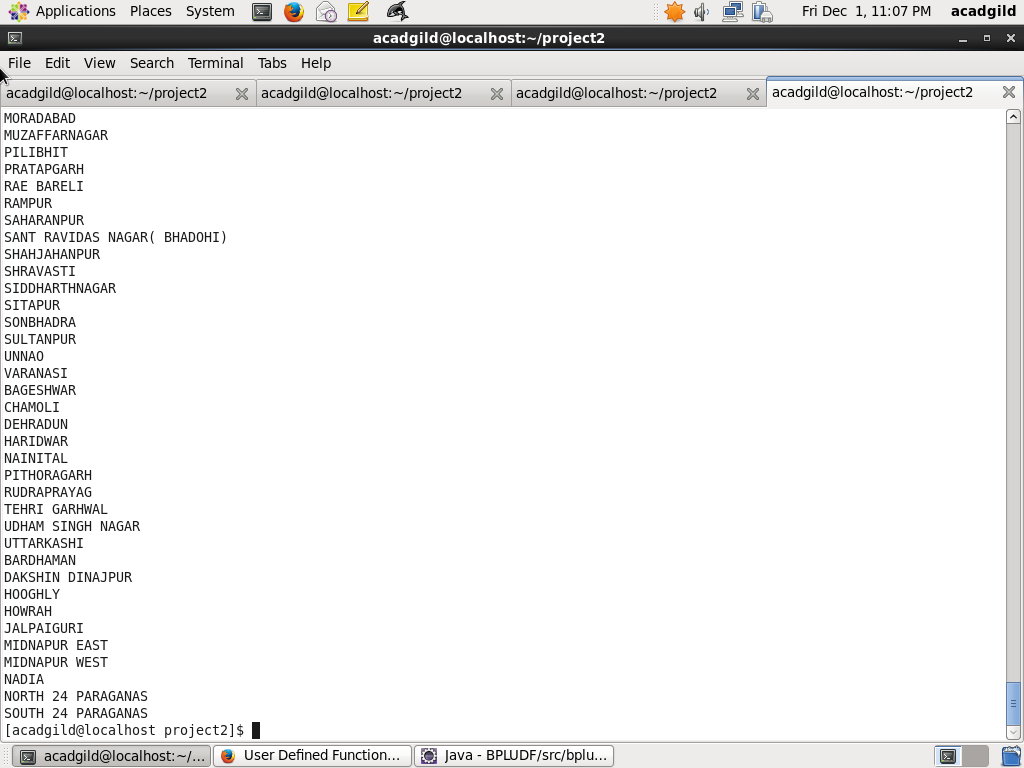
The following command shows that folders are created under districts\_having\_100percent\_objectives

hadoop fs –ls /districts\_having\_80percent\_objectives

Next, use the following HDFS command to show the results

hadoop fs –ls /districts\_having\_80percent\_objectives/part-m-00000



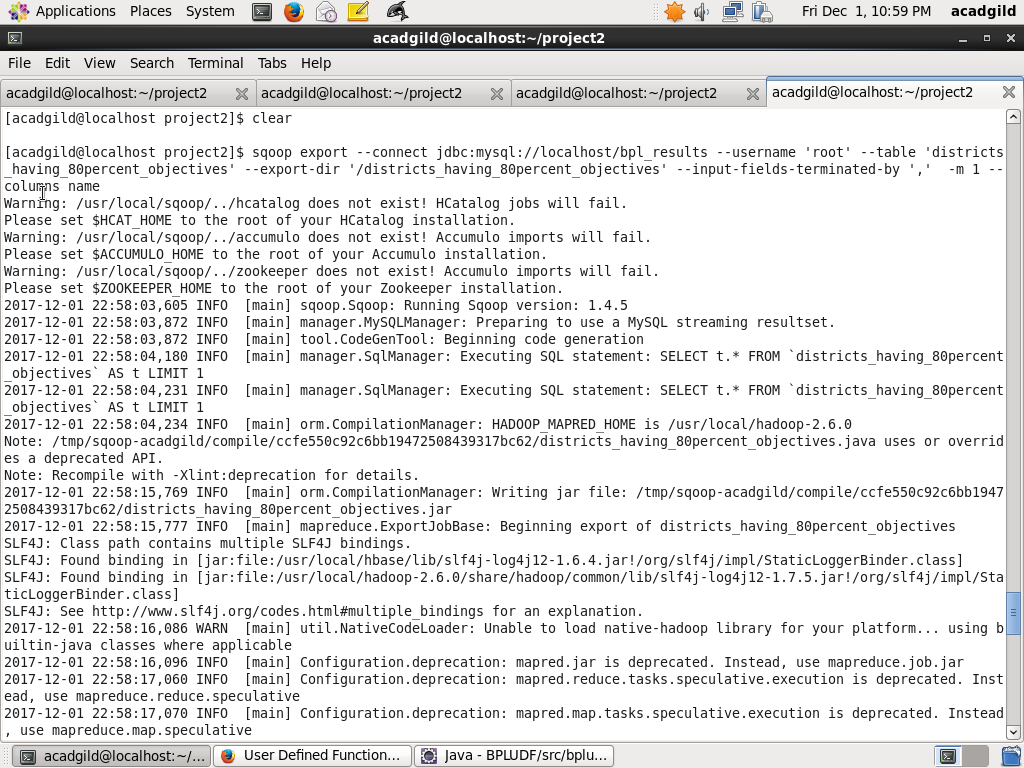


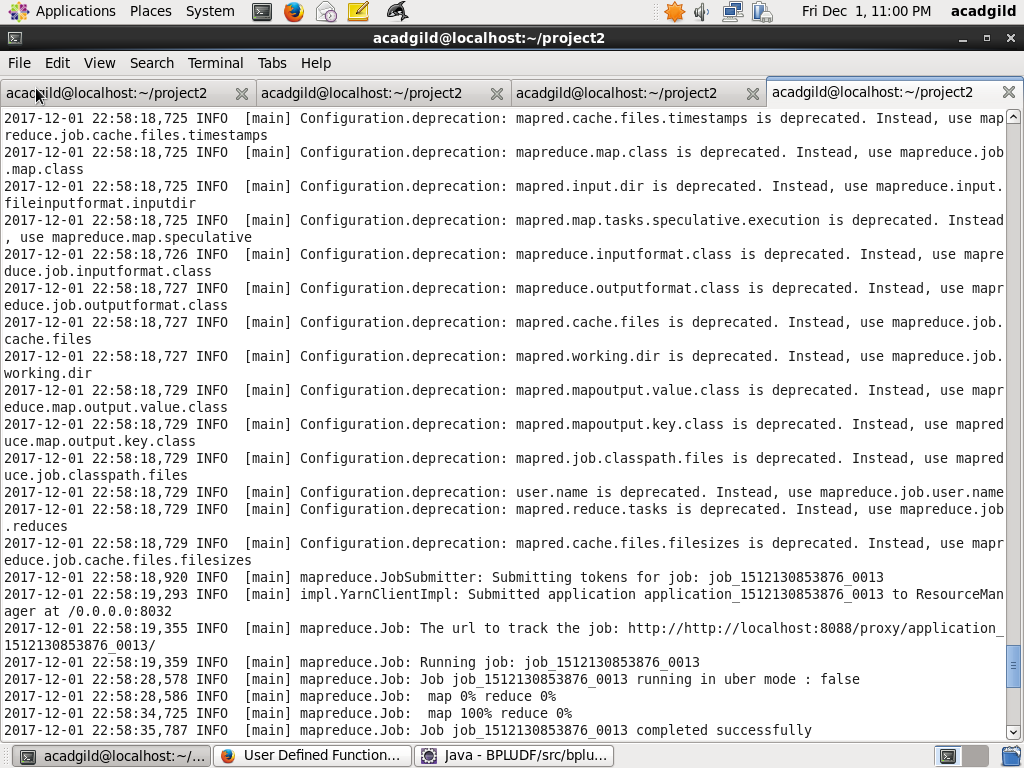
Step14 : Use sqoop command to export data from HDFS into mysql table districts\_having\_80percent\_objectives in database bpl\_results

The following sqoop command is used to export data from HDFS folder districts\_having\_80percent\_objectives into already created mysql table 'districts\_having\_80percent\_objectives

Screenshots are as below:

sqoop export --connect jdbc:mysql://localhost/bpl\_results --username 'root' --table 'districts\_having\_80percent\_objectives' --export-dir '/districts\_having\_80percent\_objectives' --input-fields-terminated-by ',' -m 1 --columns name





Ste15: Verify Result in Mysql

Use the following command in mysql to verify results in mysql

select \* from districts\_having\_80percent\_objectives

