# **PROJECT 1.2**

# **State-Wise Development Analysis In India**

**Project Overview :**

To develop the System to analyze the log data (In XML format) of government progress of various development activities.

**Requirements :**

The FLUME job which will format the data and place the data to HDFS.

Pig/MapReduce job for parsing the XML data.

Create Pig scripts/MapReduce jobs to analyze the data.

Create the Sqoop job to store the data in database.

**Dataset :**

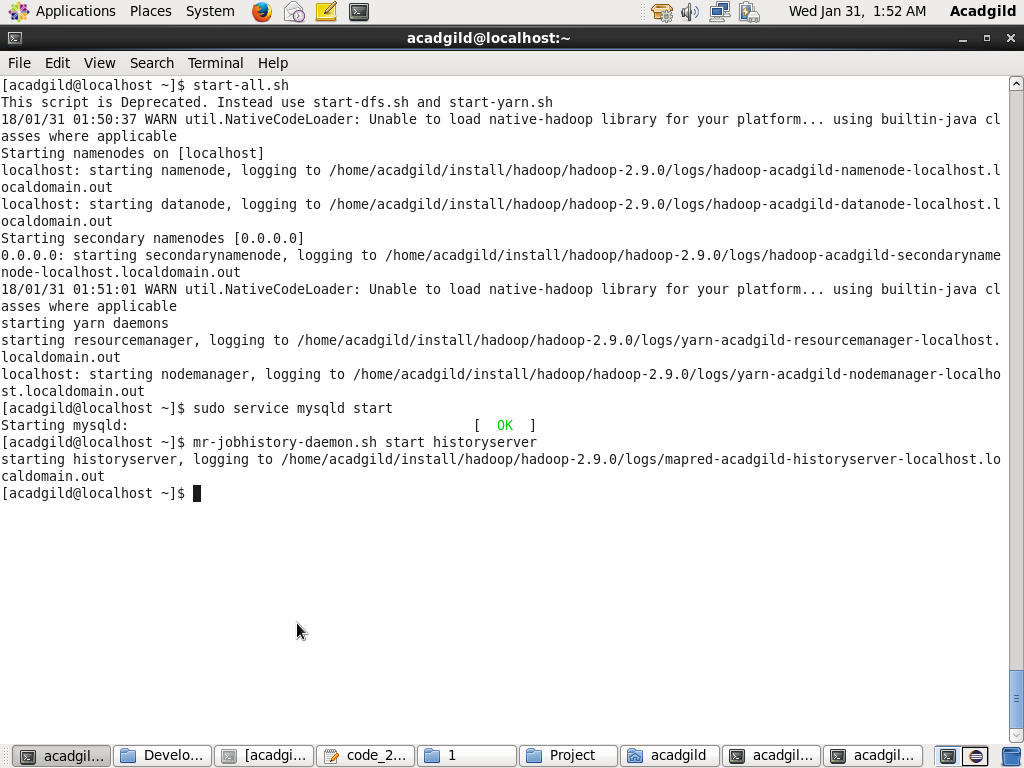
<https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing>

**Starting required daemons:**

Start-all.sh

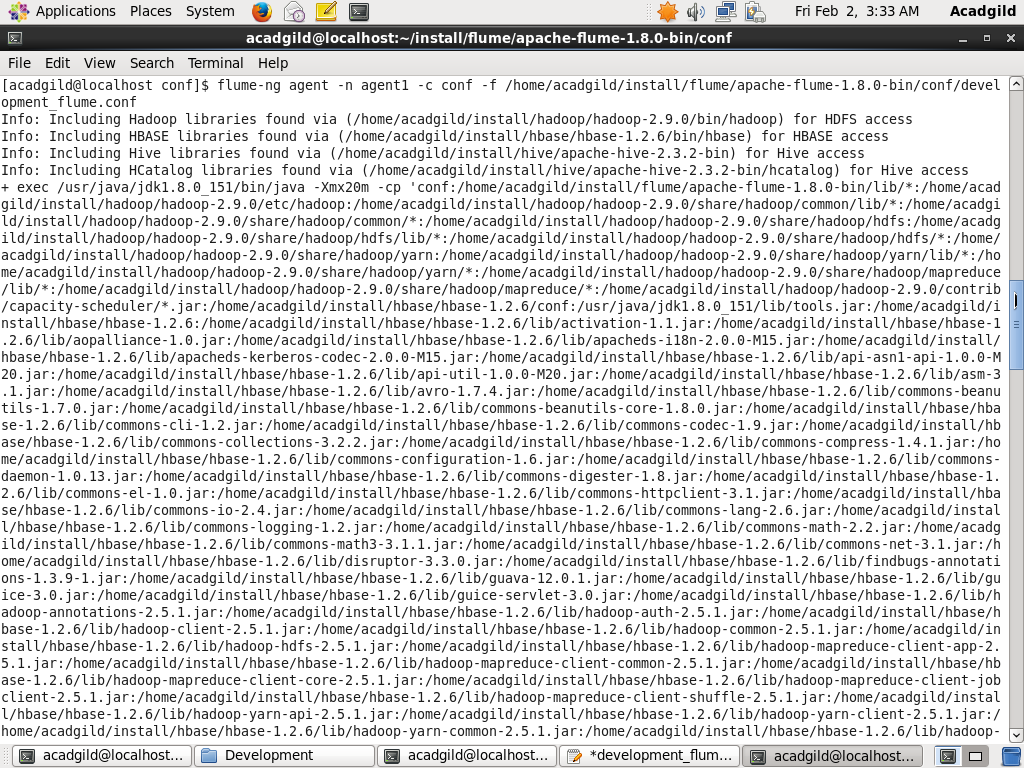
Sudo service mysqld start

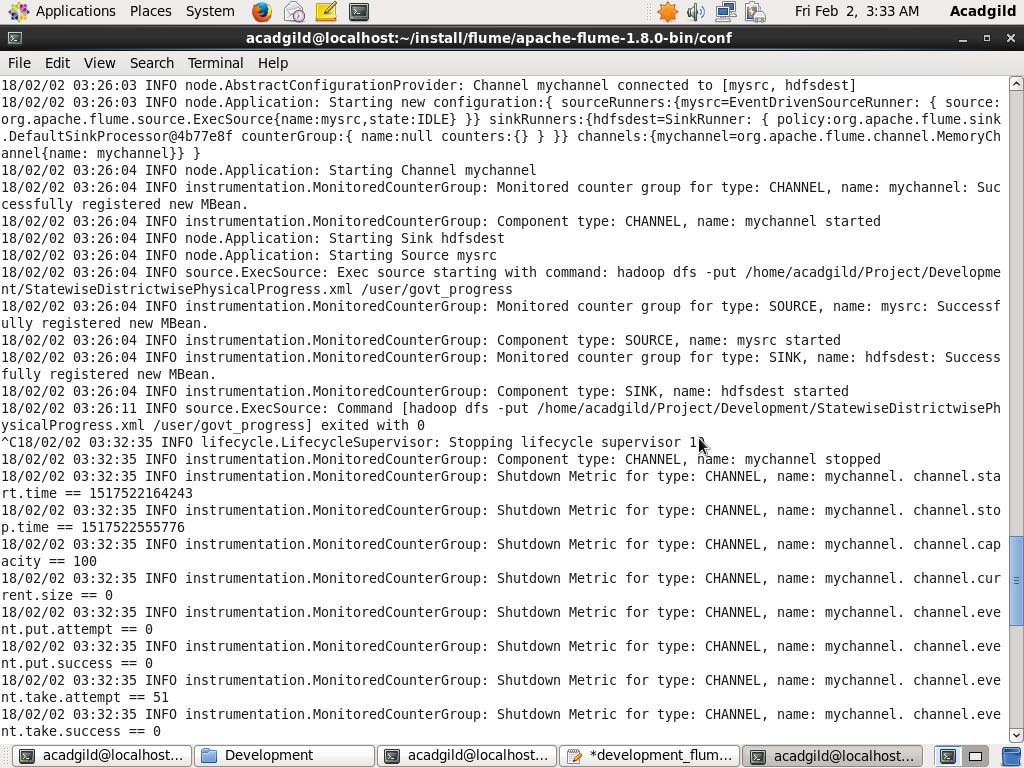
Mr-jobhistory-daemon.sh start historyserver



**Importing dataset from local file system to HDFS using flume :**

flume-ng agent –n agent1 –c conf –f /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/development\_flume.conf





The file which we have put into hdfs is XML format and for analysis of data in hive, we will first have to convert it into CSV format. For that purpose, we will use class present in piggybank.jar.

* We will use XMLoader class to load the xml file
* And, CSVExcelStorage class for storing it into CSV format.

**Content of Code.pig file:**

A = load '/user/govt\_progress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

B = foreach A generate REPLACE(x,'[\\n]','') as x;

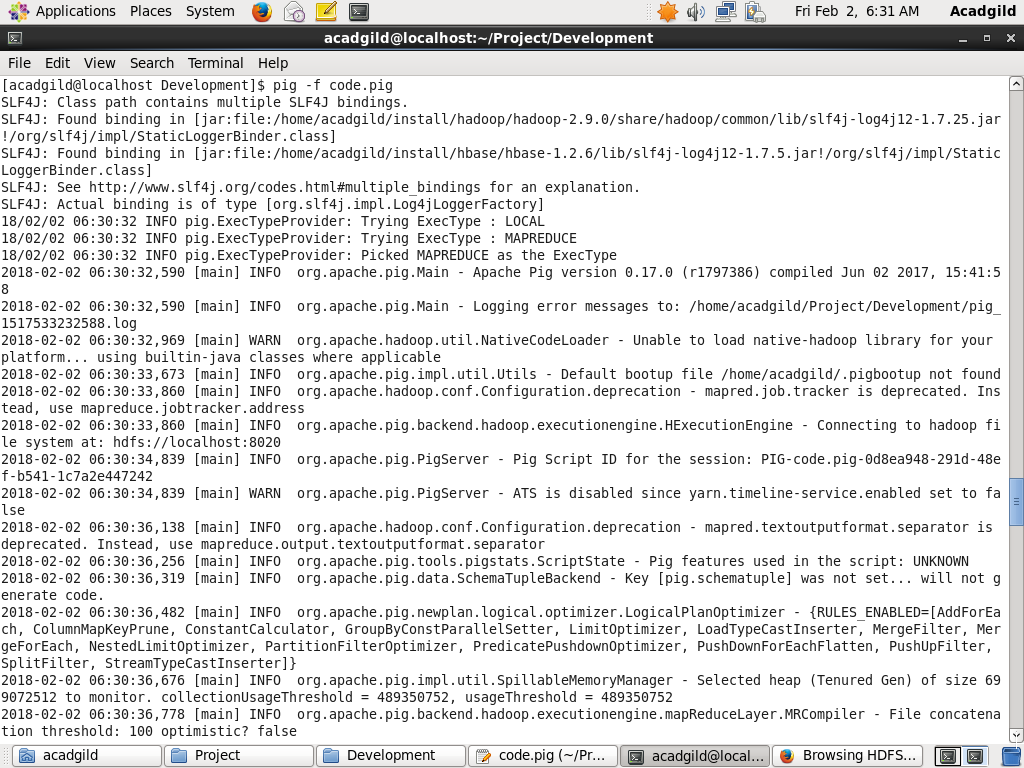
C = foreach B generate REGEX\_EXTRACT\_ALL(x,'.\*(?:<State\_Name>)([^<]\*).\*(?:<District\_Name>)([^<]\*).\*(?:<Project\_Objectives\_IHHL\_BPL>)([^<]\*).\*(?:<Project\_Objectives\_IHHL\_APL>)([^<]\*).\*(?:<Project\_Objectives\_IHHL\_TOTAL>)([^<]\*).\*(?:<Project\_Objectives\_SCW>)([^<]\*).\*(?:<Project\_Objectives\_School\_Toilets>)([^<]\*).\*(?:<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>)([^<]\*).\*');

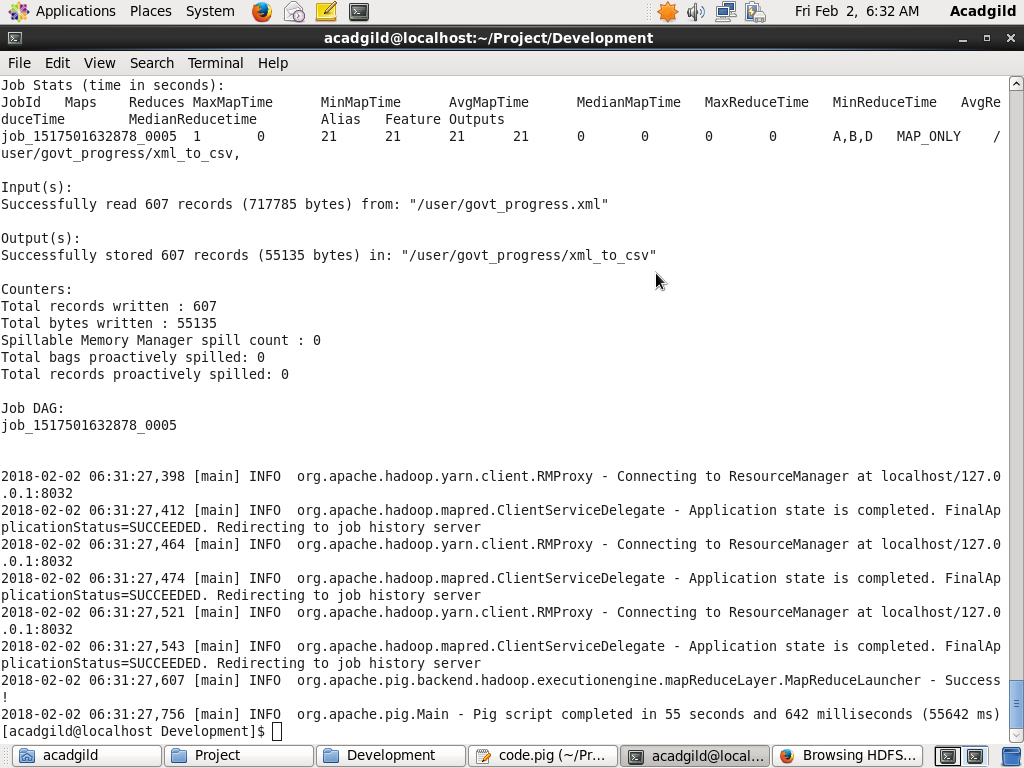
D =FOREACH C GENERATE FLATTEN (($0));

STORE D INTO '/user/govt\_progress/xml\_to\_csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage();

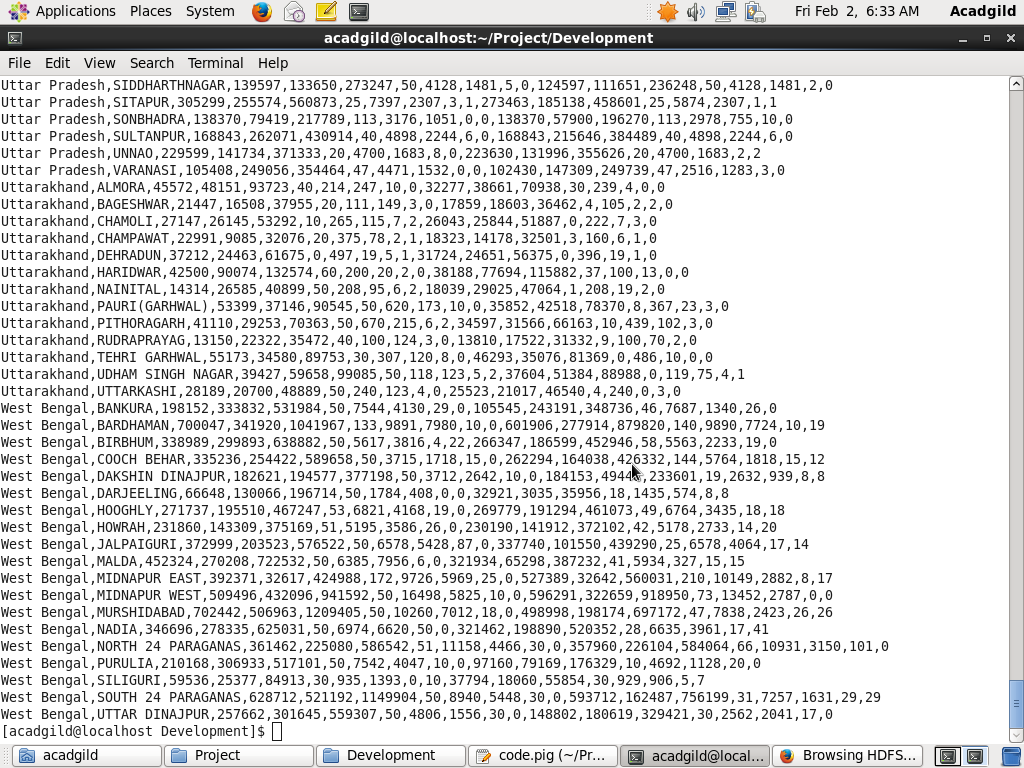
**Executing the file:**

Pig -f code.pig

****



**Data After Conversion (CSV):**

****

Now we do have data in CSV format, we can analyse the data.

**Requirement 1:**

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

**Hive Code:**

A = load '/user/govt\_progress/xml\_to\_csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX');

B = FOREACH A GENERATE (chararray)$0 as state, (chararray)$1 as district, (int)$2 as Project\_Objectives\_IHHL\_BPL, (int)$10 as Project\_Performance\_IHHL\_BPL;

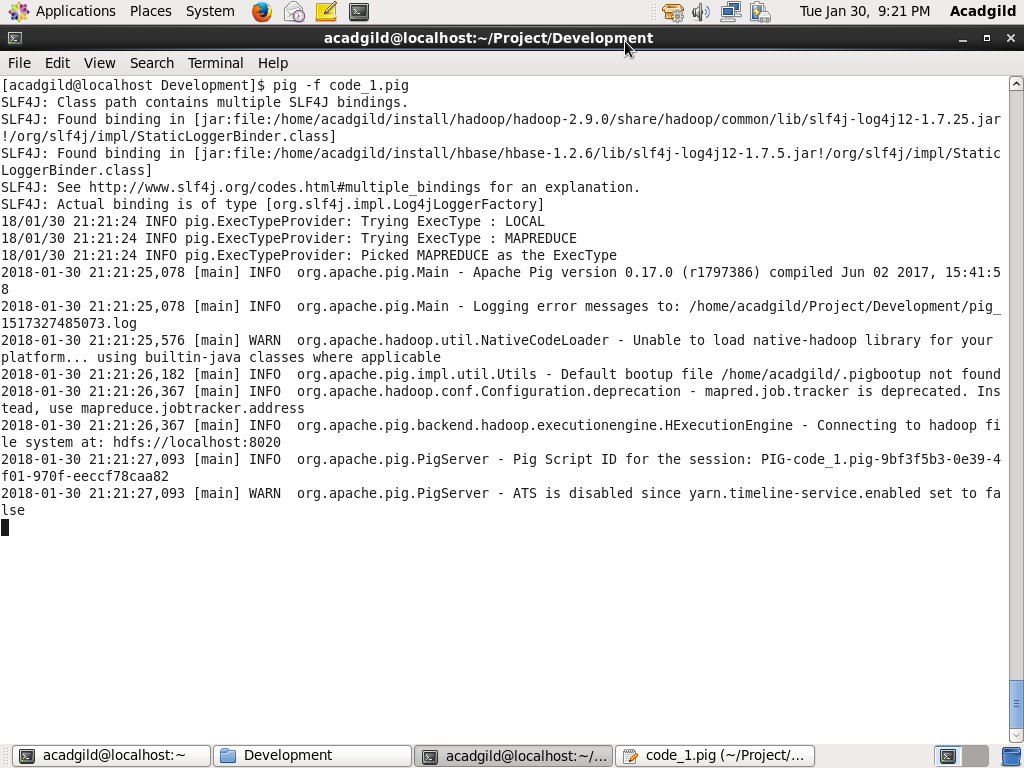
C = FILTER B BY (Project\_Performance\_IHHL\_BPL == Project\_Objectives\_IHHL\_BPL);

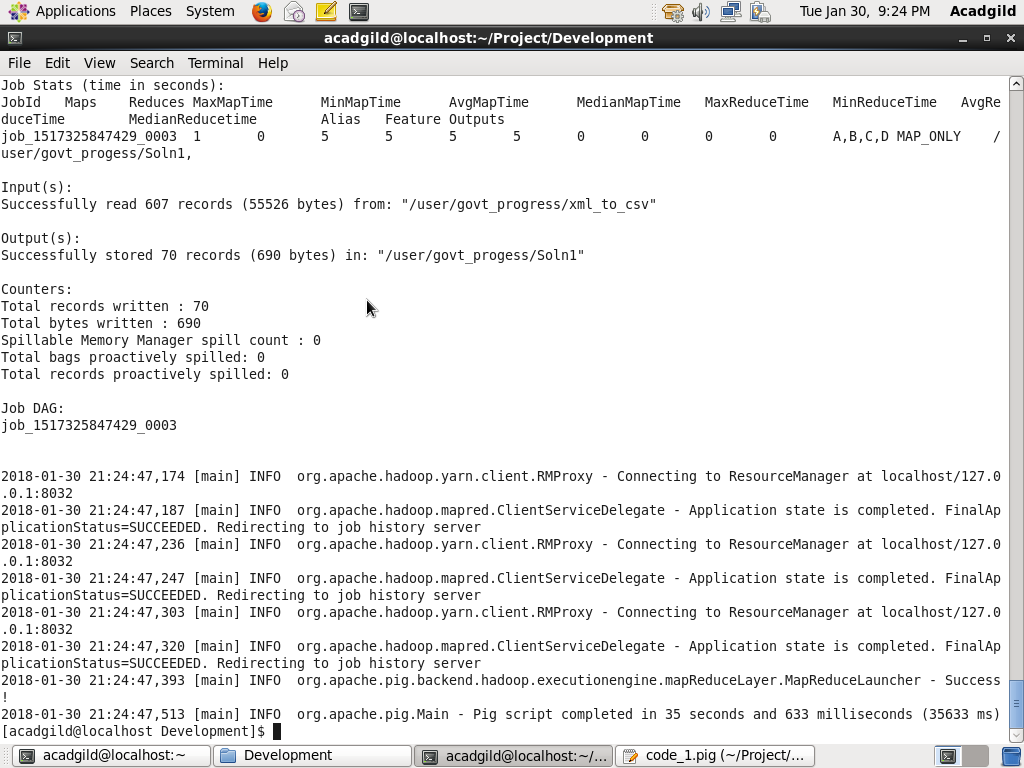
D = FOREACH C GENERATE district;

STORE D INTO '/user/govt\_progess/Soln1';

**Code Workflow:**

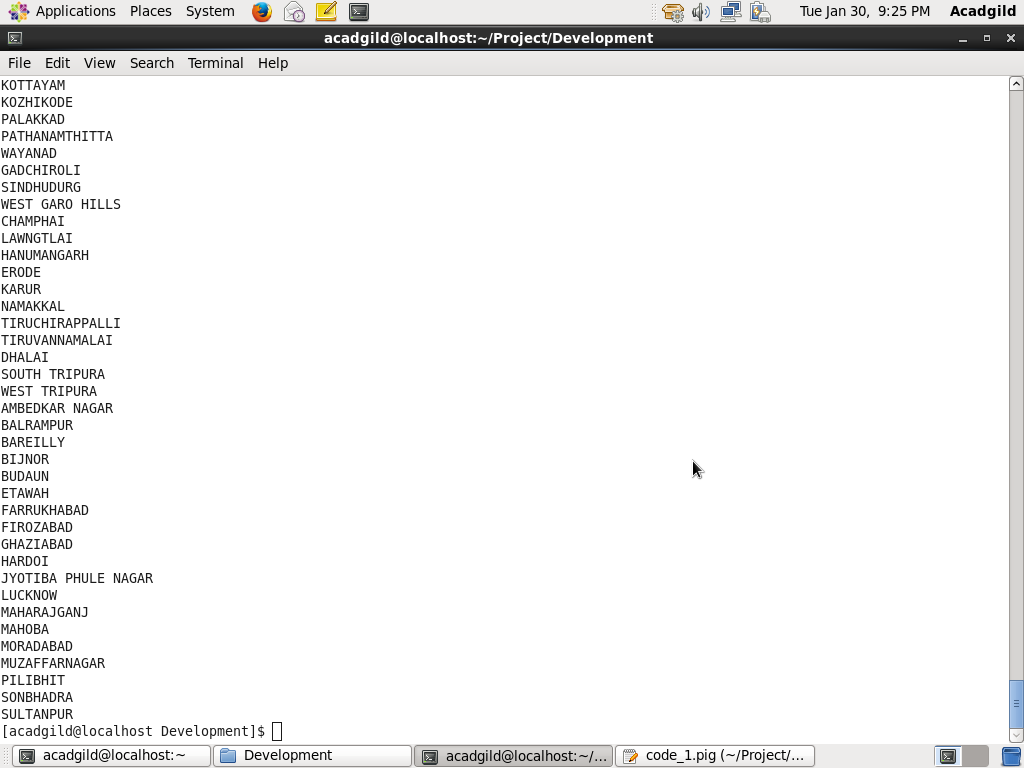
We will load CSV file and define schema name for the columns we require. According to the requirement, we want the districts with 100% objective completion. So, we will filter those records which has Project\_Performance\_IHHL\_BPL equal to Project\_Objectives\_IHHL\_BPL i.e. the objectives were met perfectly. Then, those districts will be stored into HDFS.





**OUTPUT:**

**Total Districts - 70**

****

Before exporting the result to MySQL we will first create a mysql table.

**Creation of MySQL table:**

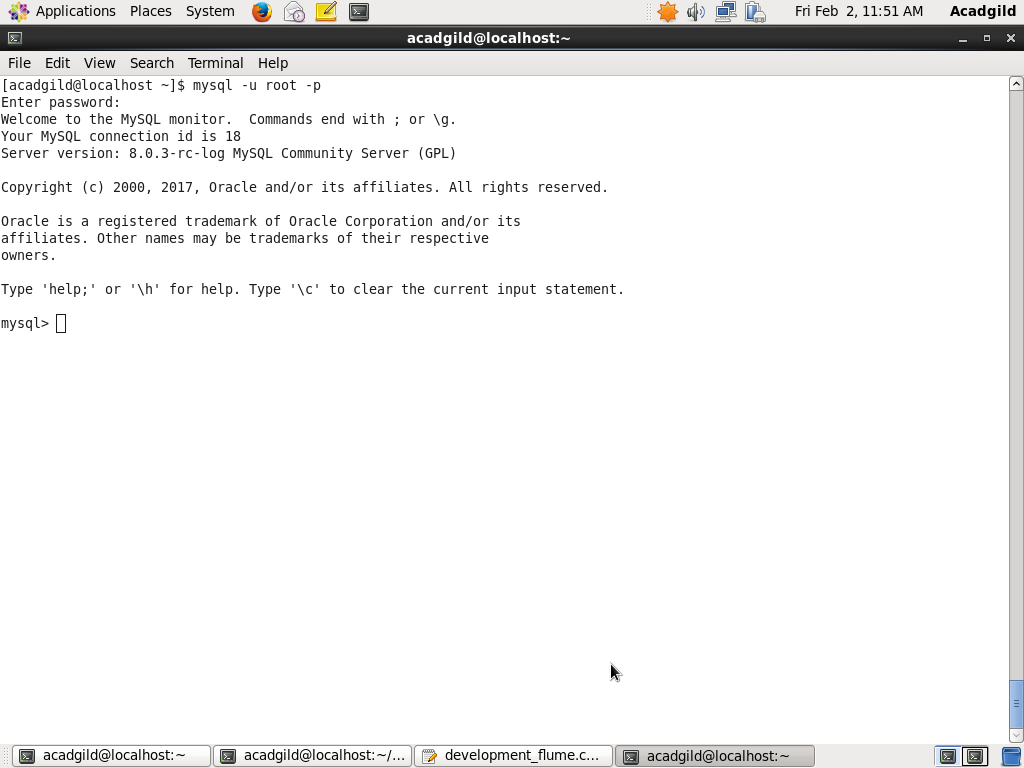
Mysql -u root -p

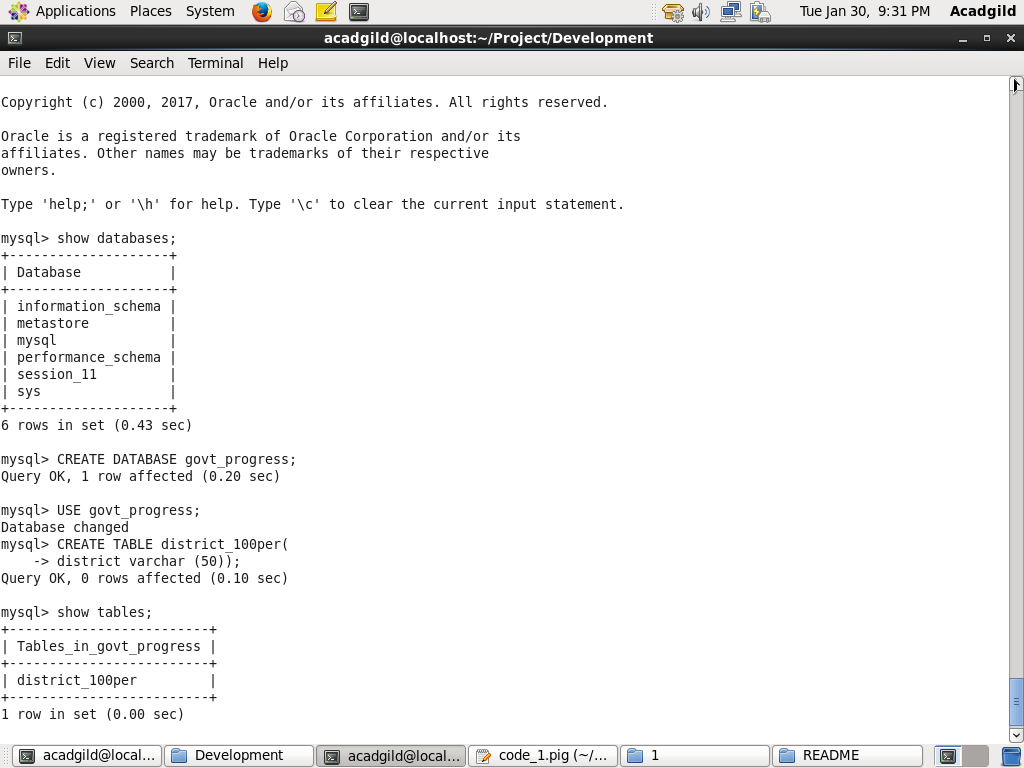
CREATE DATABASE govt\_progress;

USE govt\_progress;

CREATE TABLE district\_100per(

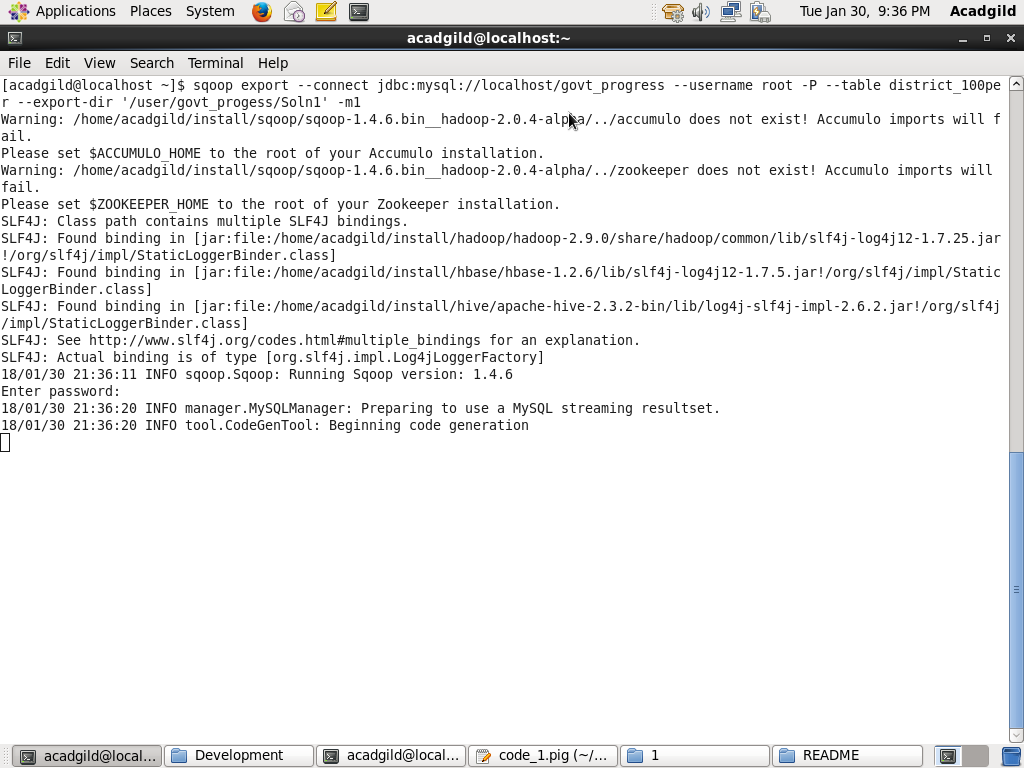
district varchar(50));

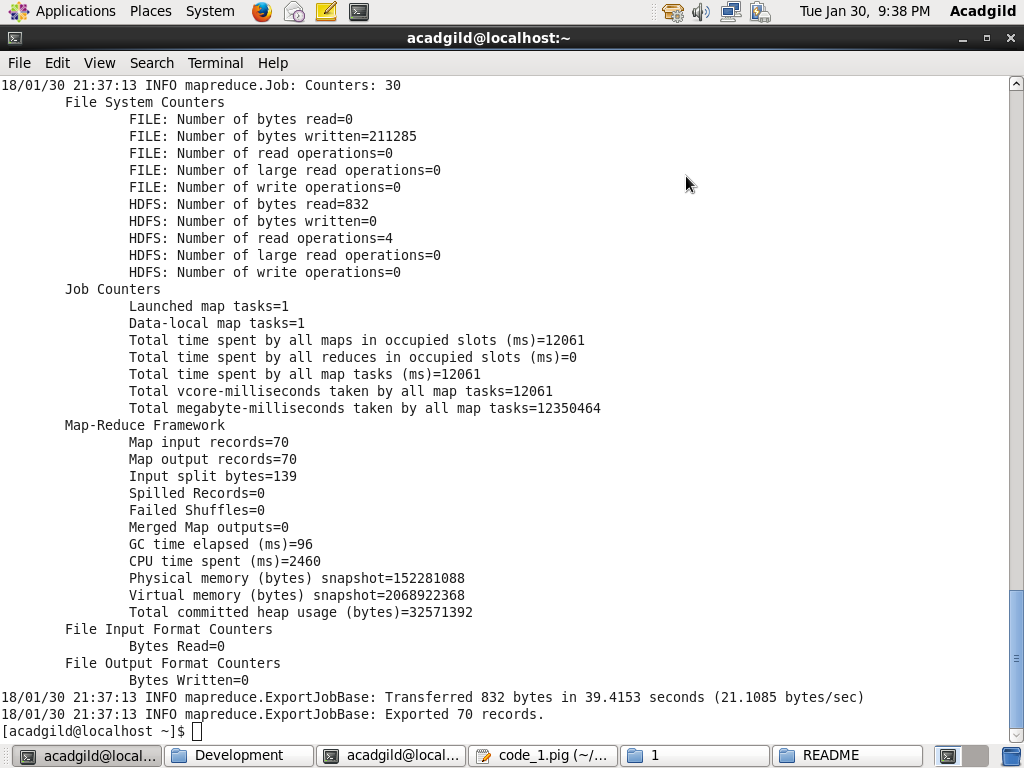




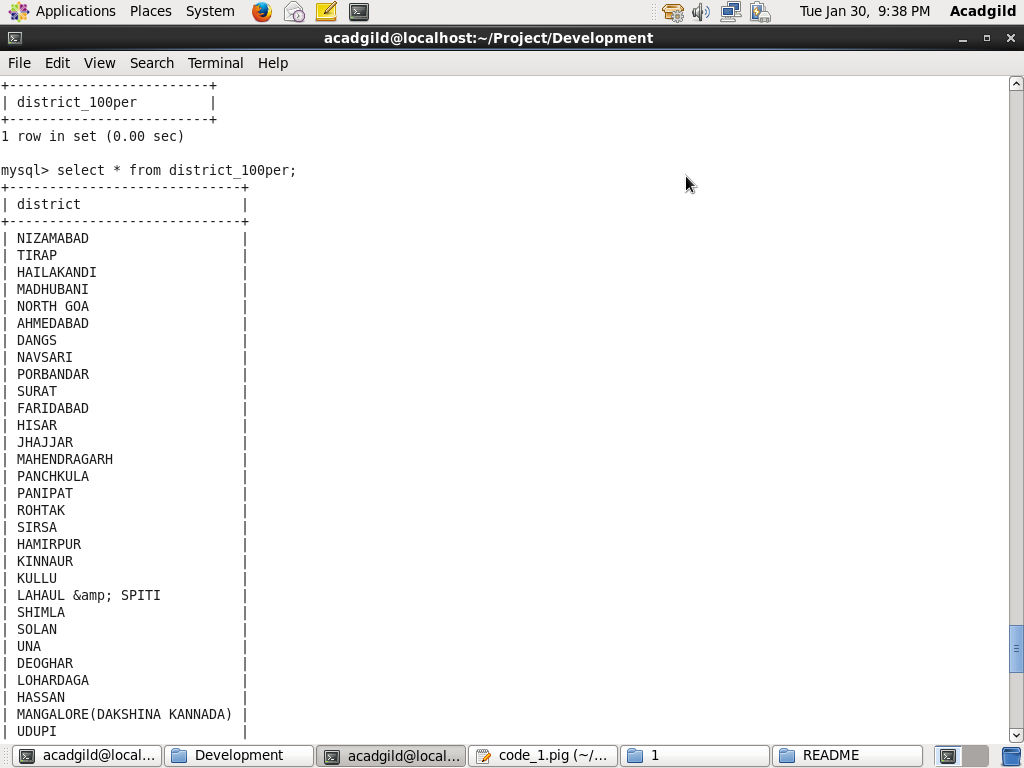
**Sqoop Export Command:**

sqoop export --connect jdbc:mysql://localhost/govt\_progress --username root -P --table district\_100per --export-dir ‘/user/govt\_progress/Soln1’ -m1





After successful Sqoop job, we will check result in SQL..



**Requirement 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.

We will create a Pig UDF, which will take a tuple as parameter. It will calculate what percent of objective were completed when we will pass Project\_Performance\_IHHL\_BPL and Project\_Objectives\_IHHL\_BPL as parameter.

**govt\_80perc.java**

package govt\_udf;

import java.io.IOException;

import org.apache.pig.EvalFunc;

import org.apache.pig.data.Tuple;

import org.apache.pig.impl.util.WrappedIOException;

public class govt\_80perc extends EvalFunc<Float> {

@SuppressWarnings("deprecation")

public Float exec(Tuple arg0) throws IOException {

// Checking if parameters are null or not

if(arg0 == null || arg0.size() == 0)

return null;

try {

Integer num1 = (Integer)arg0.get(0);

Integer num2 = (Integer)arg0.get(1);

//Finding Percentage

Long per = Math.round(num2\*100.0/num1);

Float val = per.floatValue();

return val;

} catch (Exception e) {

// TODO: handle exception

throw WrappedIOException.wrap("999.99",e);

}

}

}

**Code Workflow:**

We will now export this java file as jar. To use this UDF we will first register the jar and define an alias for the UDF and use it to get the percentage of completion of objective. Then, we will filter records with 80% as per the requirement.

REGISTER '/home/acadgild/Project/Development/2/govt\_80perc.jar';

DEFINE perc govt\_udf.govt\_80perc();

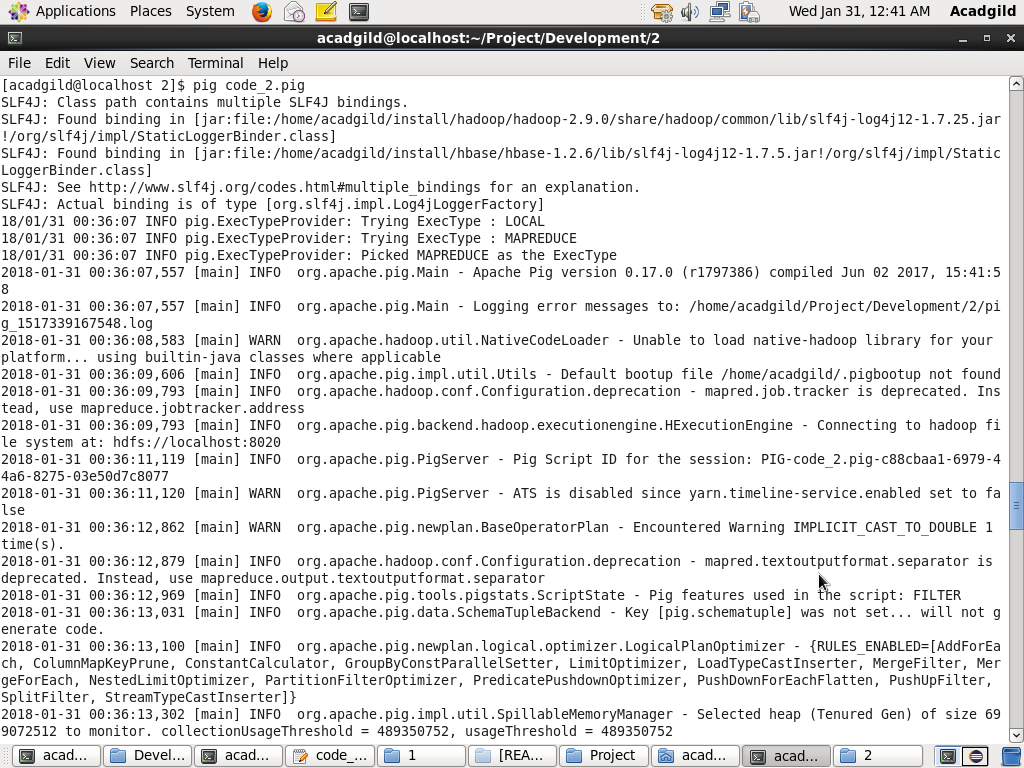
A = load '/user/govt\_progress/xml\_to\_csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX');

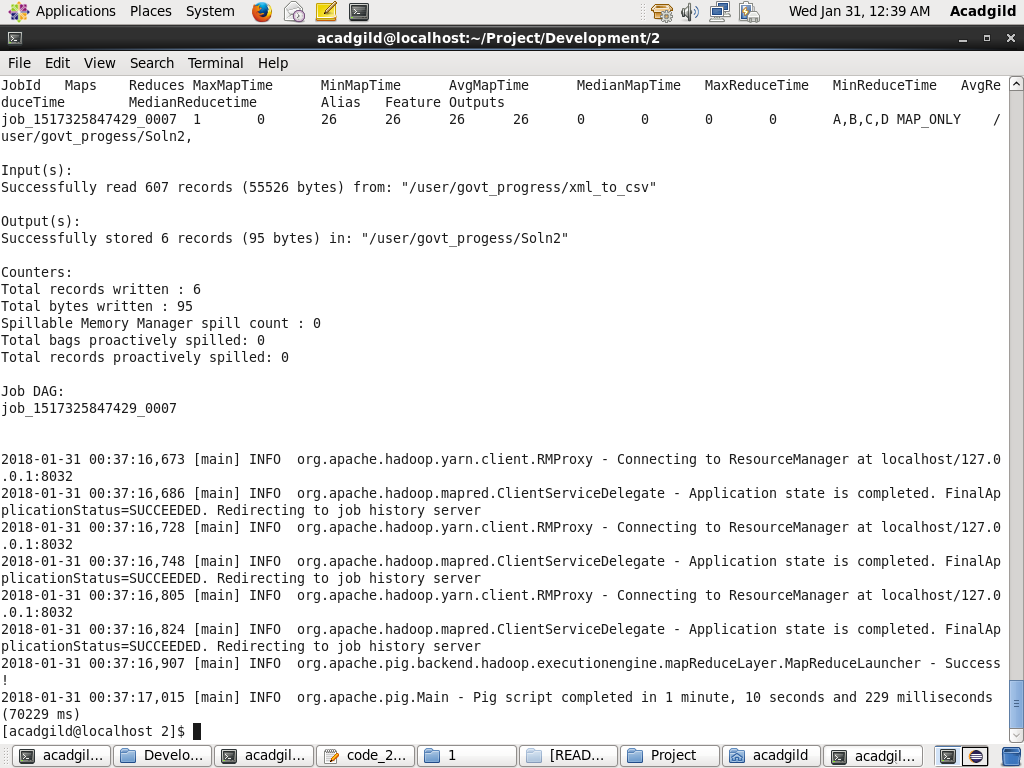
B = FOREACH A GENERATE (chararray)$0 as state, (chararray)$1 as district, (int)$2 as Project\_Objectives\_IHHL\_BPL, (int)$10 as Project\_Performance\_IHHL\_BPL;

C = FOREACH B GENERATE district, perc(Project\_Objectives\_IHHL\_BPL, Project\_Performance\_IHHL\_BPL) as percentage;

D = FILTER C BY percentage == 80.0;

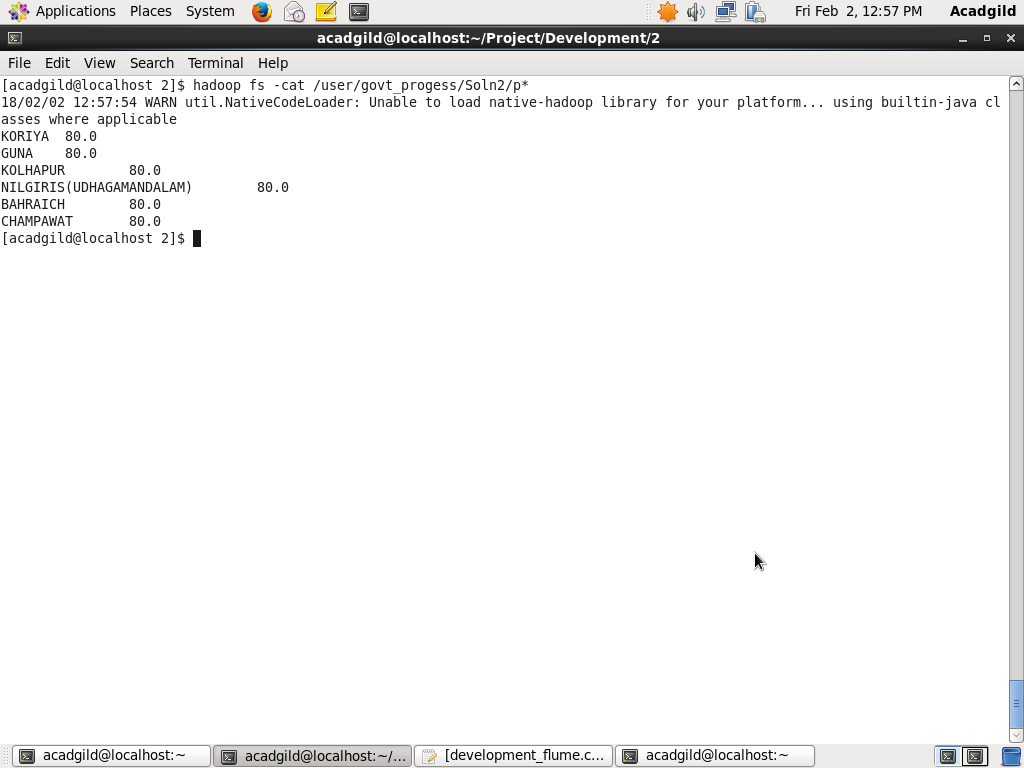
STORE D INTO '/user/govt\_progess/Soln2';





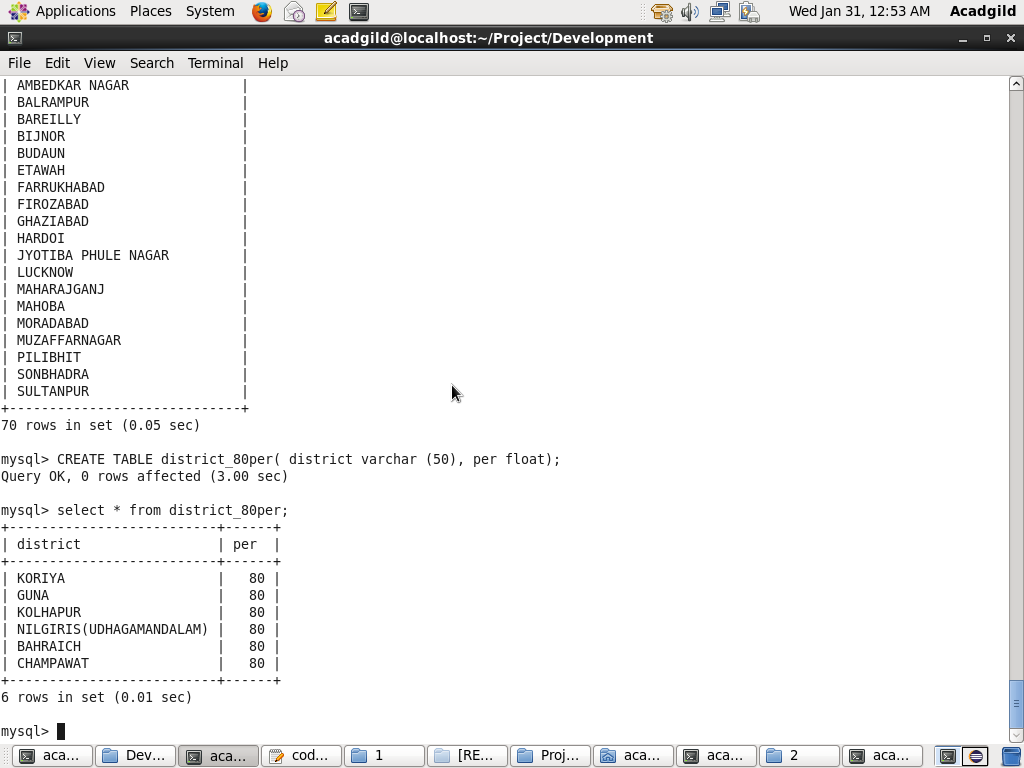
**OUTPUT:**

**Total Districts - 6**

****

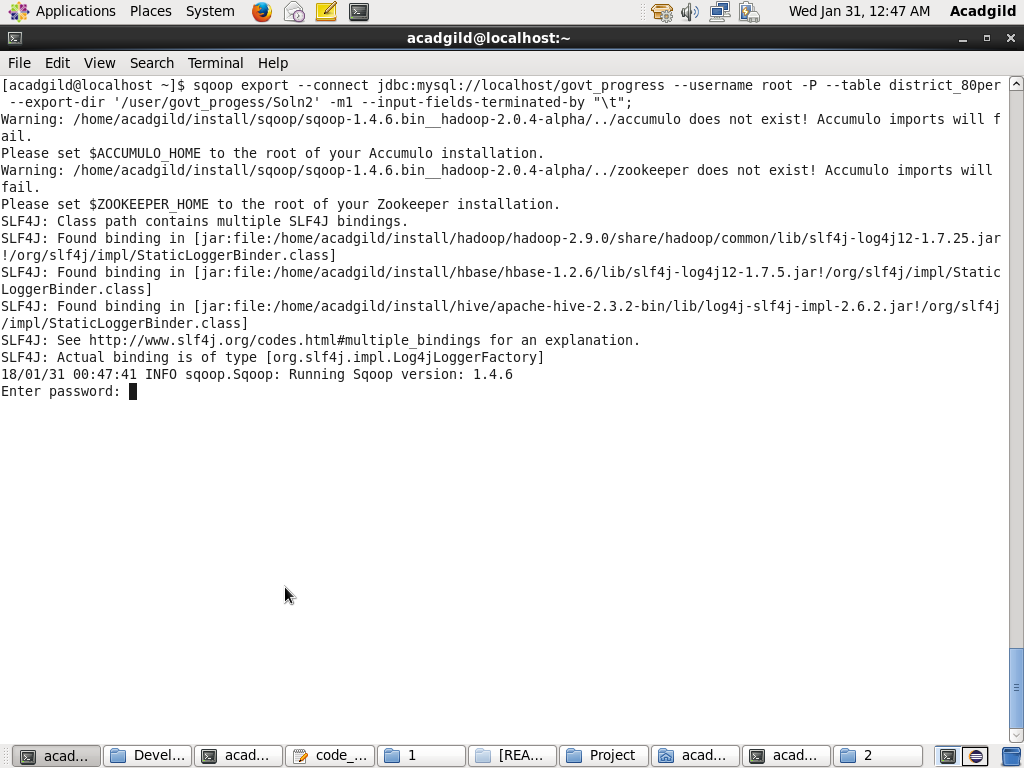
We will create a table for the data in sql:

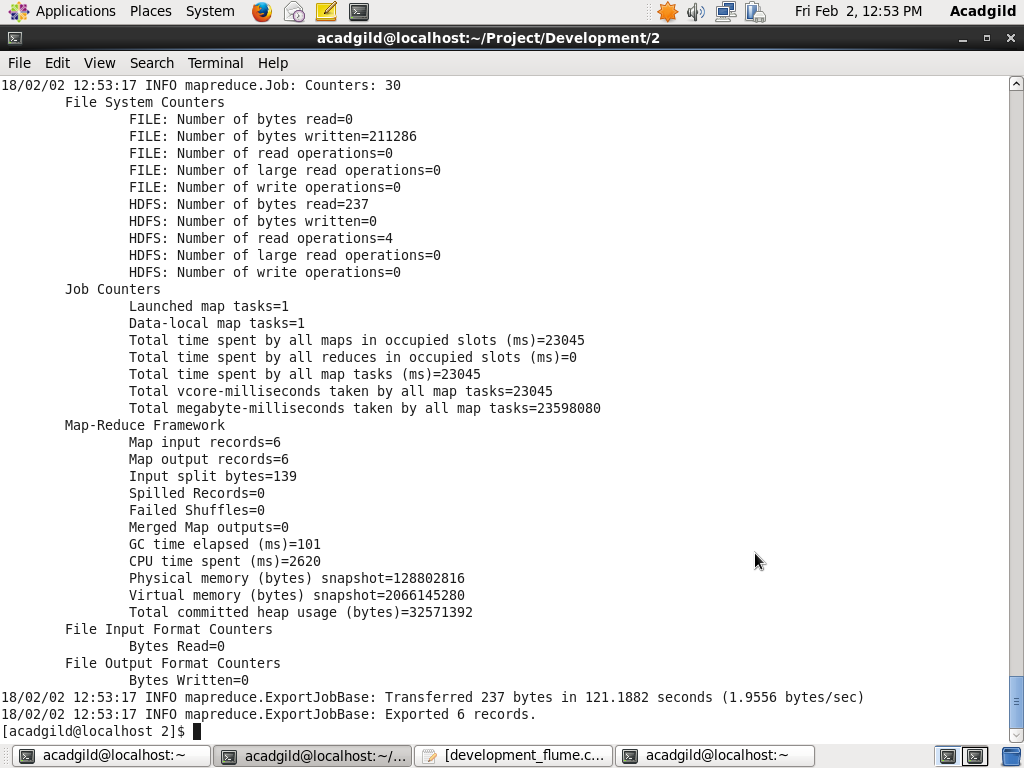
CREATE TABLE district\_80per( district varchar(50), per float);



**Sqoop Export Command:**

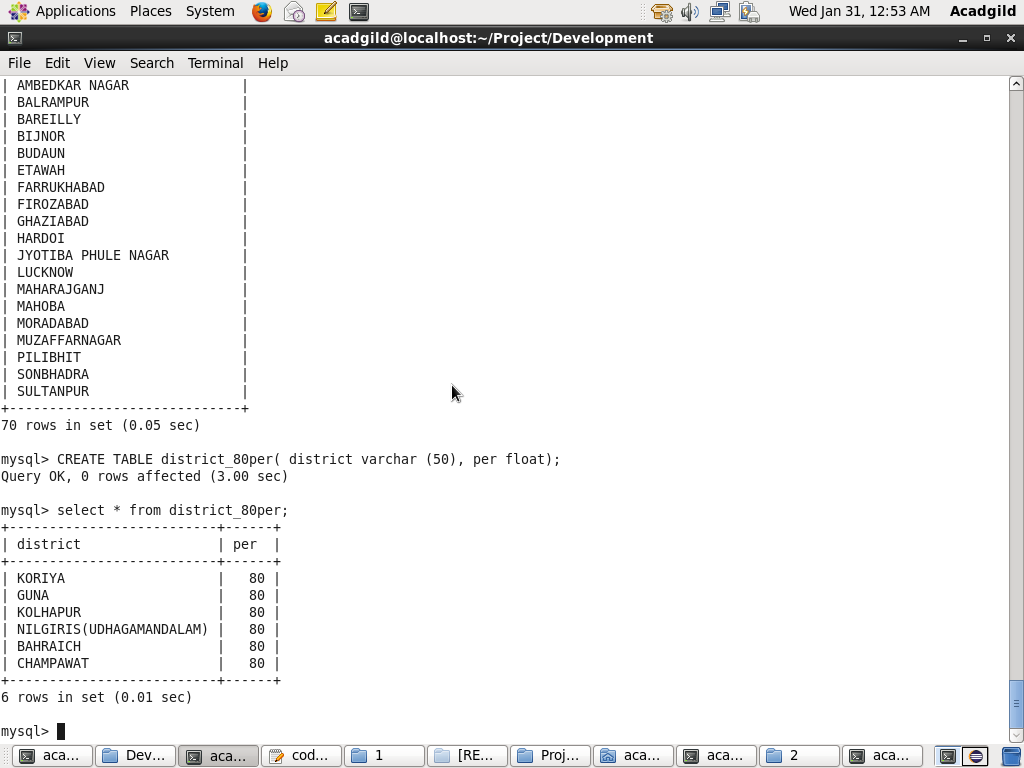
sqoop export --connect jdbc:mysql://localhost/govt\_progress --username root -P --table district\_80per --export-dir ‘/user/govt\_progress/Soln2’ -m1





**Checking data in MySQL:**

Select \* from district\_80per;



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*