Project 1- State-Wise Development Analysis In India

Priority 1 – Create FLUME job for fetching log files from spool directory the data

**Step 1: Copy dataset from local file system to HDFS using flume**

🡪Create file.conf

**File.conf:**

agent1.sources = source1

agent1.sinks = sink1

agent1.channels = channel1

agent1.sources.source1.type = exec

agent1.sources.source1.command = hadoop dfs -put /home/acadgild/flume/StatewiseDistrictwisePhysicalProgress.xml /user/acadgild/hadoop/

agent1.sinks.sink1.type = hdfs

agent1.sinks.sink1.hdfs.path = hdfs://localhost:9000/user/acadgild/hadoop/

agent1.channels.channel1.type = memory

agent1.sources.source1.channels = channel1

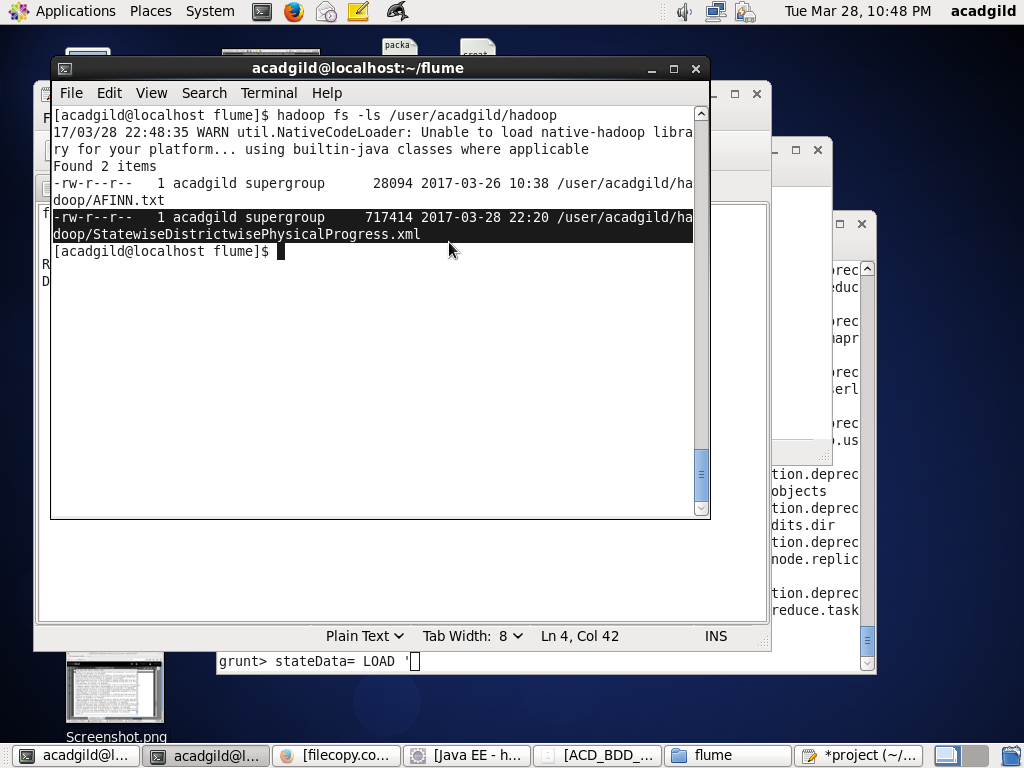
agent1.sinks.sink1.channel = channel1

**Run file.conf:**

flume-ng agent --name agent1 --conf-file /home/acadgild/flume/file.conf

**Checking HDFS location:**

hadoop fs -ls /user/acadgild/hadoop/



**Step 2: Input file is in the XML format use Map reduce or pig to parse the data and get the results for the below problem statements.**

**Register piggybank.jar**

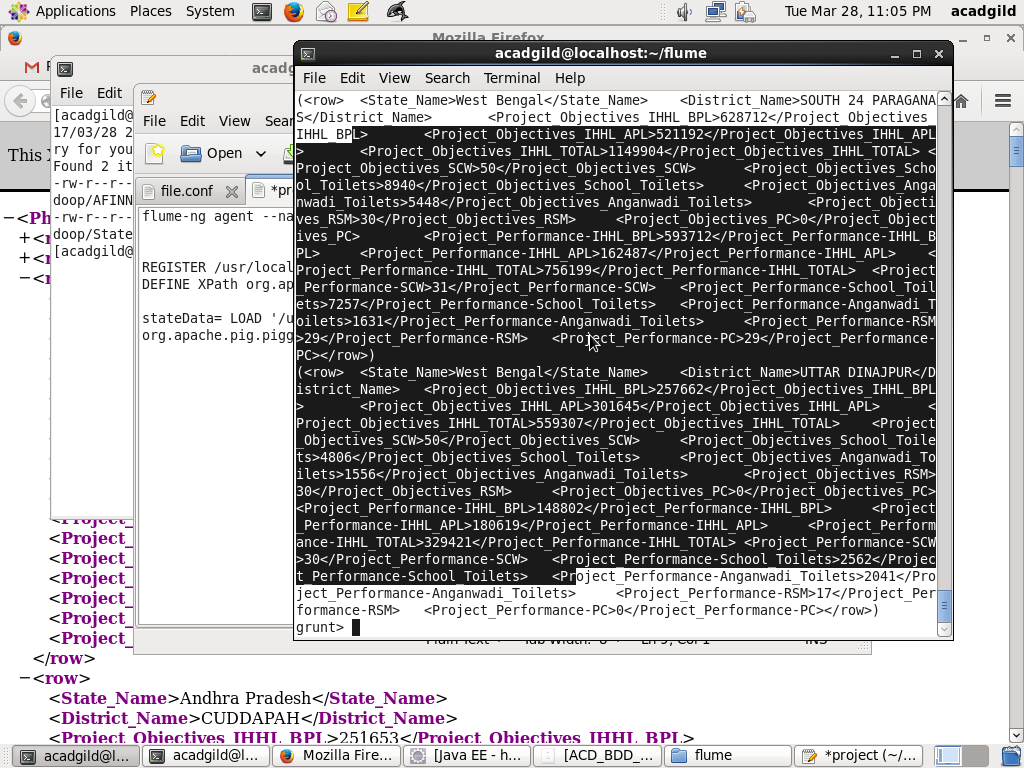
REGISTER /usr/local/pig/lib/piggybank.jar;

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

**Loading XML data**

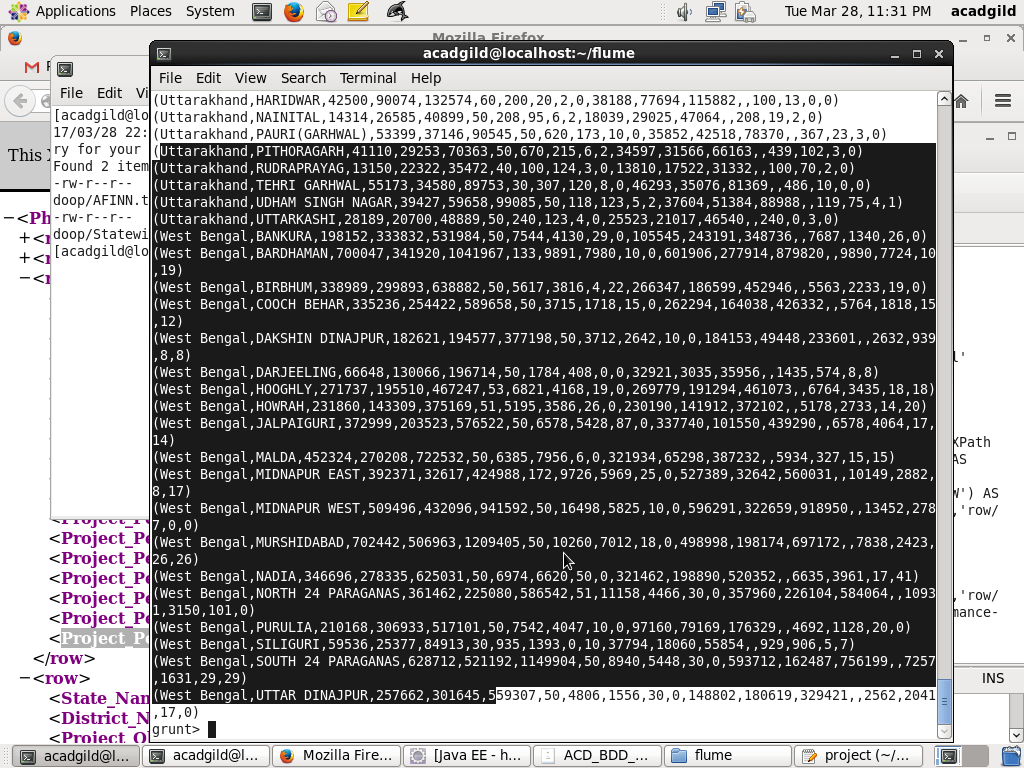
**stateData**= LOAD '/user/acadgild/hadoop/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray)

**DUMP stateData;**



**physicalProgress**= FOREACH stateData GENERATE XPath(x,'row/State\_Name') AS name,XPath(x,'row/District\_Name') AS district,XPath(x,'row/Project\_Objectives\_IHHL\_BPL') AS obj\_bpl,XPath(x,'row/Project\_Objectives\_IHHL\_APL') AS obj\_apl,XPath(x,'row/Project\_Objectives\_IHHL\_TOTAL') AS obj\_total,XPath(x,'row/Project\_Objectives\_SCW') AS obj\_scw,XPath(x,'row/Project\_Objectives\_School\_Toilets') AS obj\_toilets,XPath(x,'row/Project\_Objectives\_Anganwadi\_Toilets') AS obj\_anganwadi,XPath(x,'row/Project\_Objectives\_RSM') AS obj\_rsm,XPath(x,'row/Project\_Objectives\_PC') AS obj\_pc,XPath(x,'row/Project\_Performance-IHHL\_BPL') AS per\_bpl,XPath(x,'row/Project\_Performance-IHHL\_APL') AS per\_apl,XPath(x,'row/Project\_Performance-IHHL\_TOTAL') AS per\_total,XPath(x,'Project\_Performance-SCW') AS per\_scw,XPath(x,'row/Project\_Performance-School\_Toilets') AS per\_toilets,XPath(x,'row/Project\_Performance-Anganwadi\_Toilets') AS per\_anganwadi,XPath(x,'row/Project\_Performance-RSM') AS per\_rsm,XPath(x,'row/Project\_Performance-PC') AS per\_pc;

**DUMP physicalProgress;**



Store XML loaded data into temporary file , so that in future(**Problem statement-2**) if we want to work on the data..no need to do all the above steps again and again.

STORE physicalProgress INTO '/home/acadgild/pig/data';

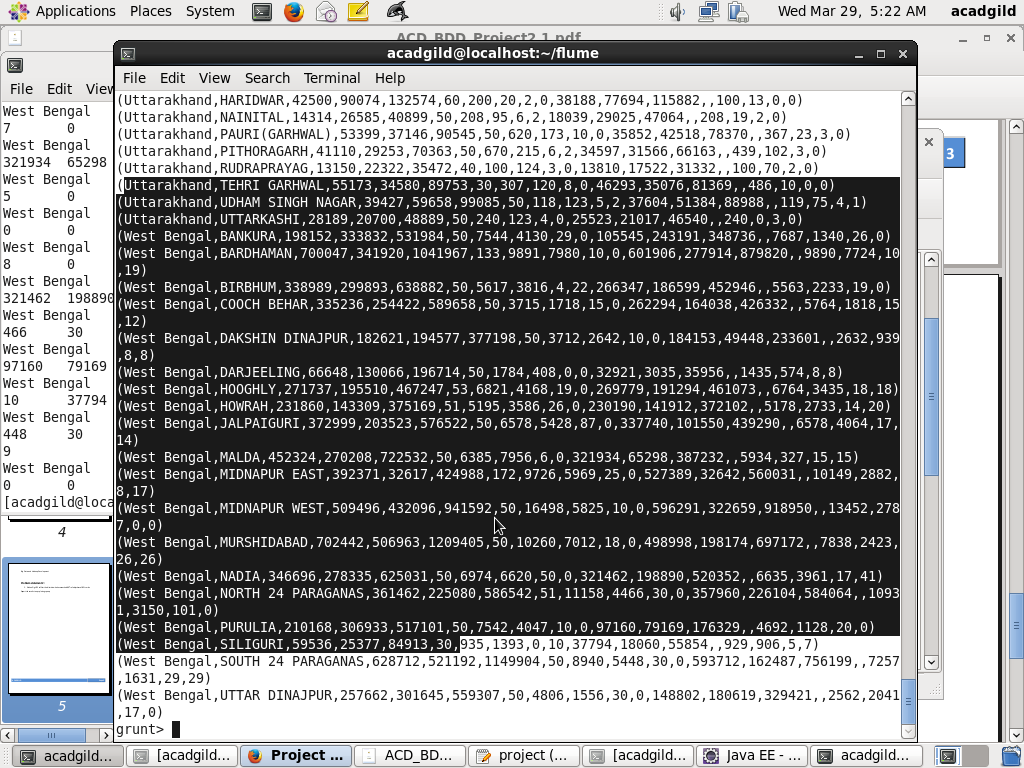
hadoop fs -cat /home/acadgild/pig/data/part-m-00000

**data**= LOAD '/home/acadgild/pig/data/part-m-00000' AS (name:chararray,district:chararray,obj\_bpl:int,obj\_apl:int,obj\_total:int,obj\_scw:int,

obj\_toilets:int,obj\_anganwadi:int,obj\_rsm:int,obj\_pc:int,per\_bpl:int,per\_apl:int,

per\_total:int,per\_scw:int,per\_toilets:int,per\_anganwadi:int,per\_rsm:int,per\_pc:int);

**DUMP data;**



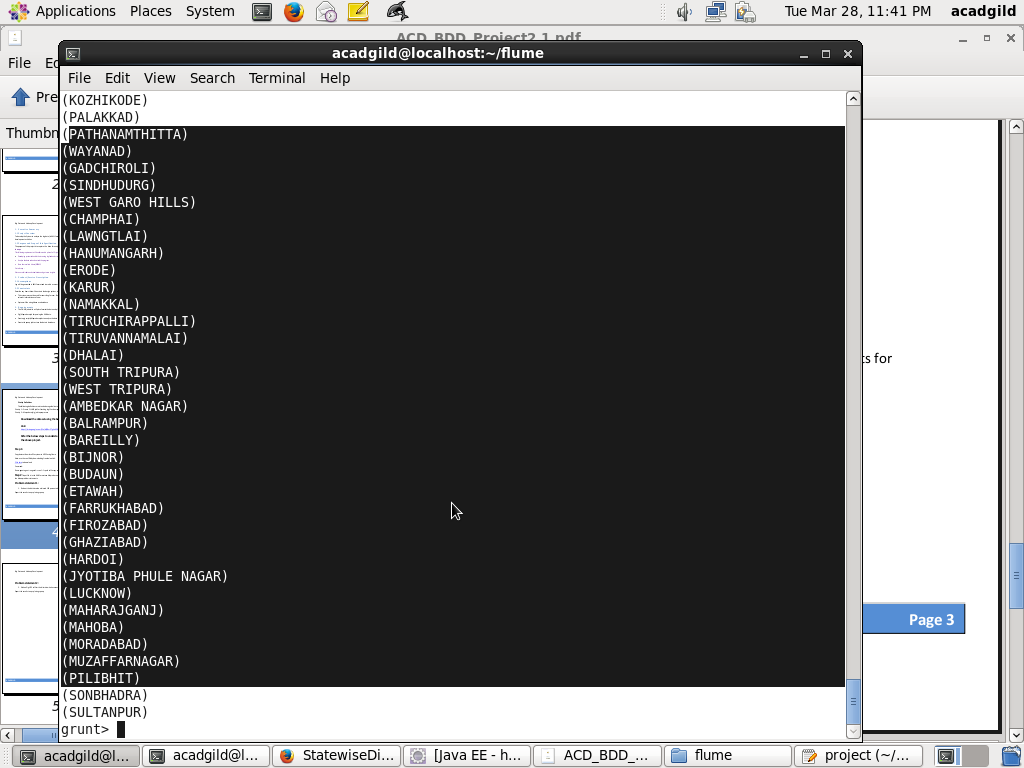
**Problem statement 1:**

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

**bpl**= FILTER physicalProgress BY obj\_bpl==per\_bpl;

**districts**= FOREACH bpl GENERATE district;

**DUMP districts;**

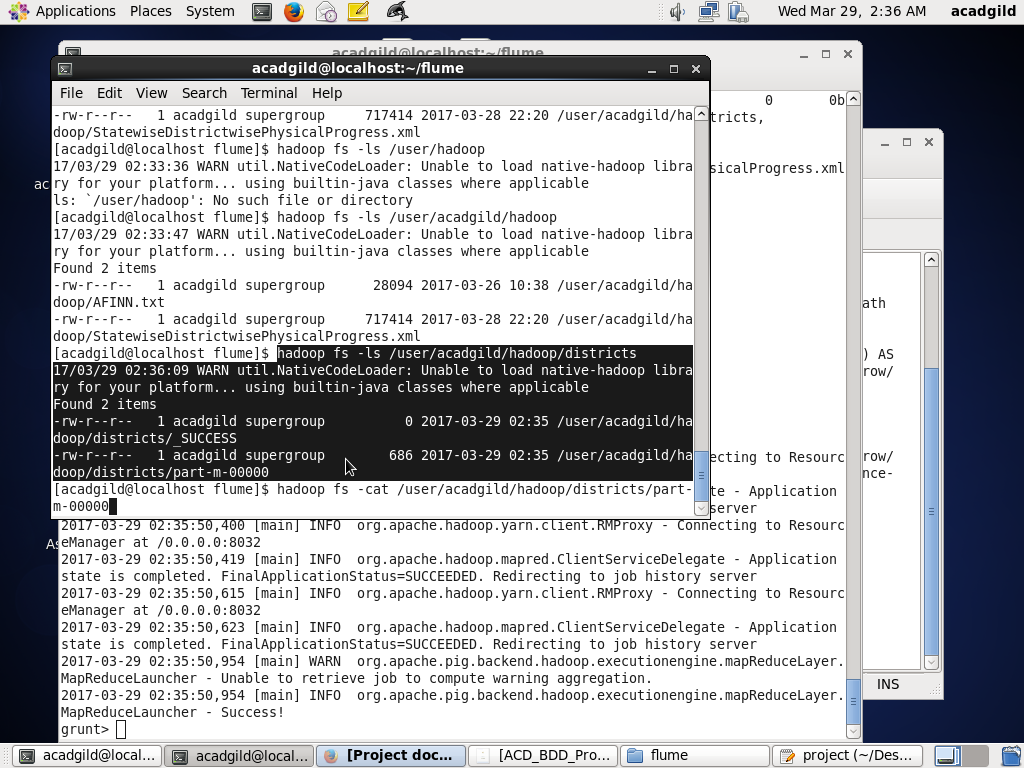


**Storing the result into HDFS:**

STORE districts INTO '/user/acadgild/hadoop/districts';

**Checking in HDFS location:**

hadoop fs -ls /user/acadgild/hadoop/districts



**Export the results to mysql using sqoop**

**Creating table in mysql:**

mysql>**create database project;**

Query OK, 1 row affected (0.00 sec)

mysql> **use project;**

Database changed

mysql> **create table districts**

**-> (**

**-> district\_name varchar(20)**

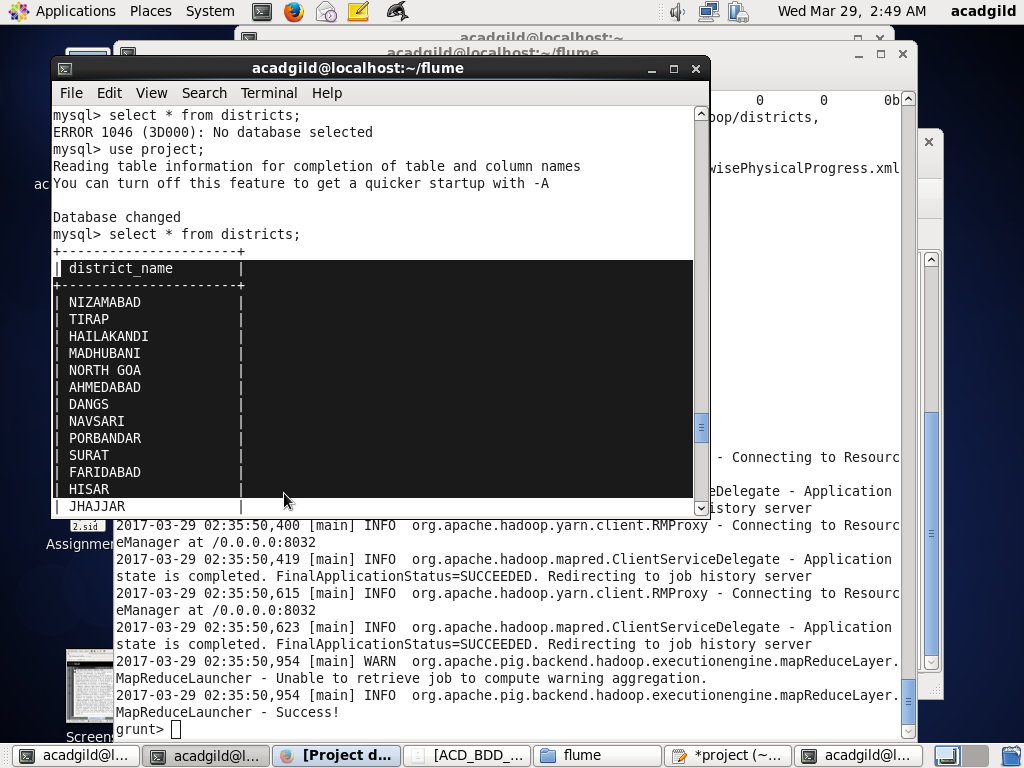
**-> );**

**Exporting to mysql using sqoop:**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table 'districts' --export-dir '/user/acadgild/hadoop/districts' -m 1 --columns district\_name

**Checking in ‘districts’ mysql table:**

select \* from districts;



**Problem statement 2:**

**2. Write a Pig UDF to filter the districts who have reached 80% of objectives of BPL cards.**

Write Java class for UDF

**FilterUDF.class:**

package pig\_udf;

import java.io.IOException;

import org.apache.pig.EvalFunc;

import org.apache.pig.data.Tuple;

public class FilterUDF extends EvalFunc<String>{

public String exec(Tuple arg0) throws IOException {

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

return null;

try{

int val=(Integer)arg0.get(0);

double value=val\*0.8;

return String.valueOf(value);

}catch(Exception e){

throw new IOException(e);

}

}

}

Make a jar of this project and include it in project\_udf.pig

JAR🡪**pig\_udf.jar**

**Project\_udf.pig:**

REGISTER '/home/acadgild/workspace/pig\_udf.jar';

DEFINE filterudf pig\_udf.FilterUDF;

**data**= LOAD '/home/acadgild/pig/data/part-m-00000' AS (name:chararray,district:chararray,obj\_bpl:int,obj\_apl:int,obj\_total:int,obj\_scw:int,

obj\_toilets:int,obj\_anganwadi:int,obj\_rsm:int,obj\_pc:int,per\_bpl:int,per\_apl:int,per\_total:int,per\_scw:int,per\_toilets:int,per\_anganwadi:int,per\_rsm:int,per\_pc:int);

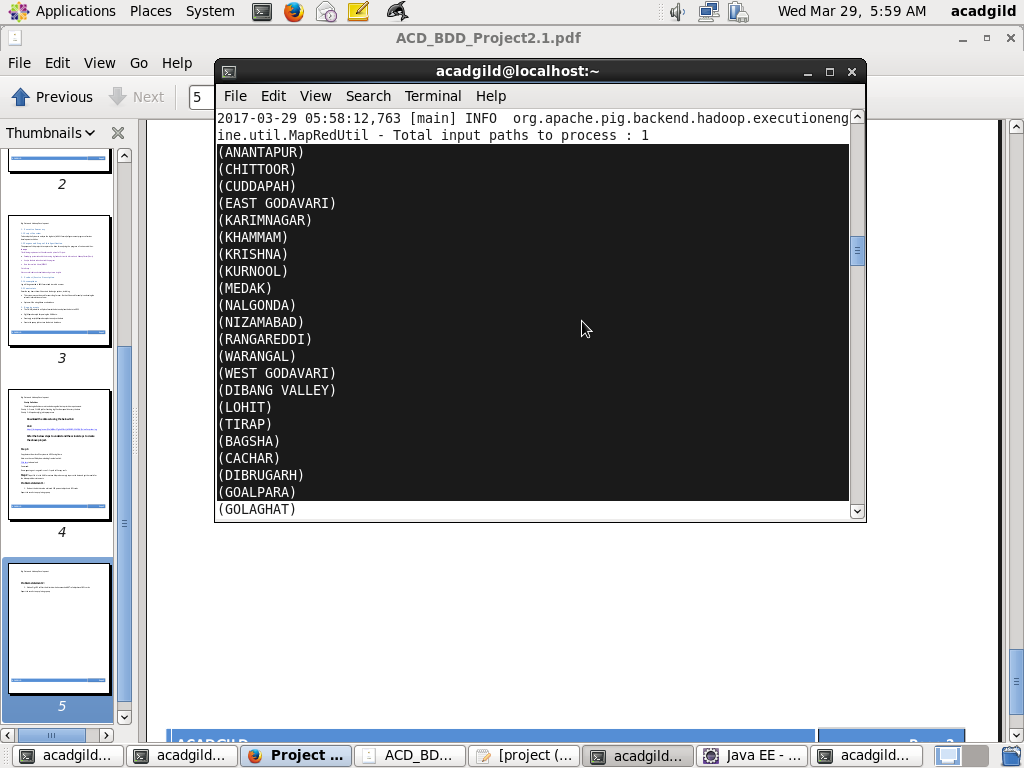
**percent**= FOREACH data GENERATE district,per\_bpl,filterudf(obj\_bpl) as obj\_per;

**final**= FOREACH ( FILTER percent BY per\_bpl>=(double)obj\_per) GENERATE district;

DUMP final;

**Run project\_udf.pig:**

**>pig project\_udf.pig**



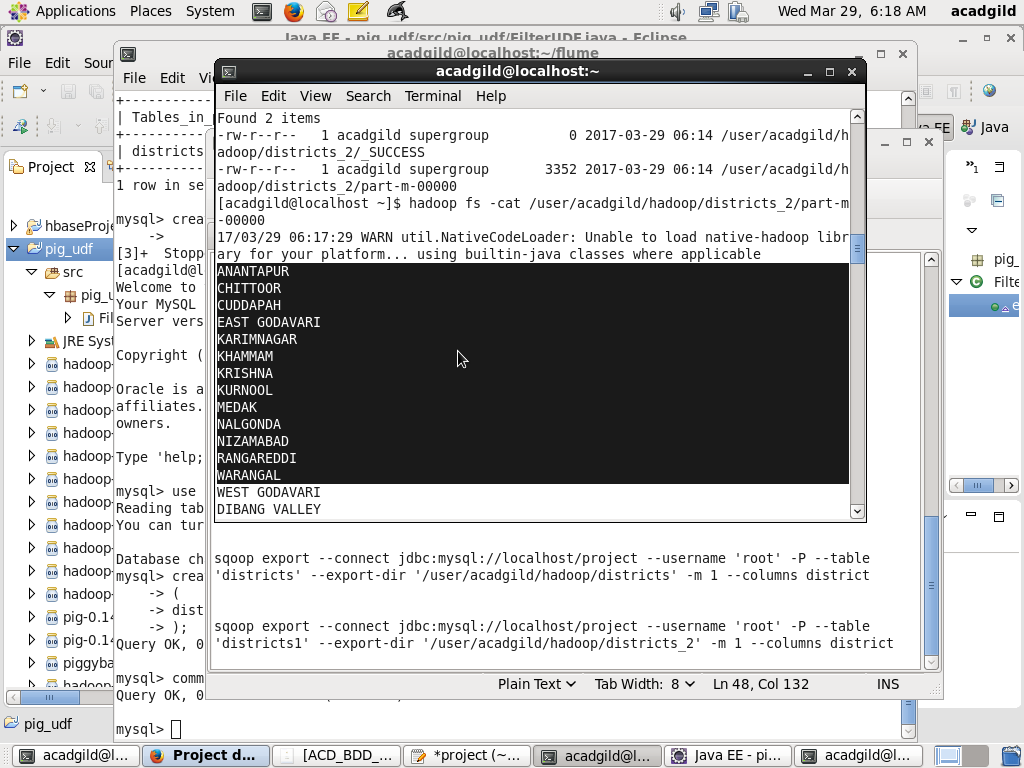
**Storing results in HDFS:**

STORE final INTO '/user/acadgild/hadoop/districts\_2';

**Checking in HDFS location:**

hadoop fs -ls /user/acadgild/hadoop/districts\_2;

hadoop fs -cat /user/acadgild/hadoop/districts\_2/part-m-00000;



**Creating table ‘districts1’ in mysql:**

create table districts1

-> (

-> district\_name varchar(20)

-> );

**Export the results to mysql using sqoop:**

sqoop export --connect jdbc:mysql://localhost/project --username 'root' -P --table 'districts1' --export-dir '/user/acadgild/hadoop/districts\_2' -m 1 --columns district\_name

**Checking ‘districts1’ table:**

Select \* from districts1;

