**Project 1.2 State Wise Development Analysis In India**

(014,228)

* **Loaded configure and xml data Files in Local**

[acadgild@localhost ~]$ pwd

/home/acadgild

[acadgild@localhost ~]$ ls

derby.log filecopy.conf Music Templates

Desktop files Pictures Videos

Documents hadoop pig\_1507458541475.log workspace

Downloads hdfs: pig\_1520436558337.log

eclipse hive-site.xml Public

eclipse-jee-neon-M3-linux-gtk-x86\_64.tar.gz metastore\_db StatewiseDistrictwisePhysicalProgress.xml

[acadgild@localhost ~]$



* **XML Data file is copied from Local directory to HDFS using conf file**

[acadgild@localhost sbin]$ flume-ng agent -n agent1 -c conf -f /home/acadgild/filecopy.conf

Info: Including Hadoop libraries found via (/usr/local/hadoop-2.6.0/bin/hadoop) for HDFS access

Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath

Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath

Info: Including HBASE libraries found via (/usr/local/hbase/bin/hbase) for HBASE access

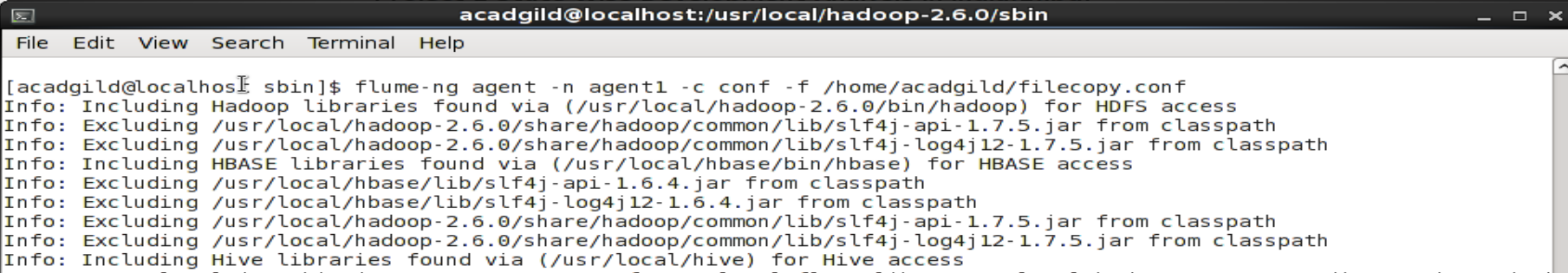
Info: Excluding /usr/local/hbase/lib/slf4j-api-1.6.4.jar from classpath

Info: Excluding /usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar from classpath

Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath

Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath

Info: Including Hive libraries found via (/usr/local/hive) for Hive access



* **Displaying the files successfully copied to HDFS**

[acadgild@localhost sbin]$ hadoop fs -ls /

18/03/08 20:57:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 7 items

-rw-r--r-- 1 acadgild supergroup 28 2017-10-08 13:24 /HelloWorld.txt

-rw-r--r-- 1 acadgild supergroup 212 2017-11-19 19:21 /employee.txt

-rw-r--r-- 1 acadgild supergroup 717414 2018-03-08 20:51 /flume\_import

drwxr-xr-x - acadgild supergroup 0 2017-10-08 16:09 /hbasestorage

drwxrwxr-x - acadgild supergroup 0 2017-11-19 19:26 /tmp

drwxr-xr-x - acadgild supergroup 0 2015-11-17 01:56 /user

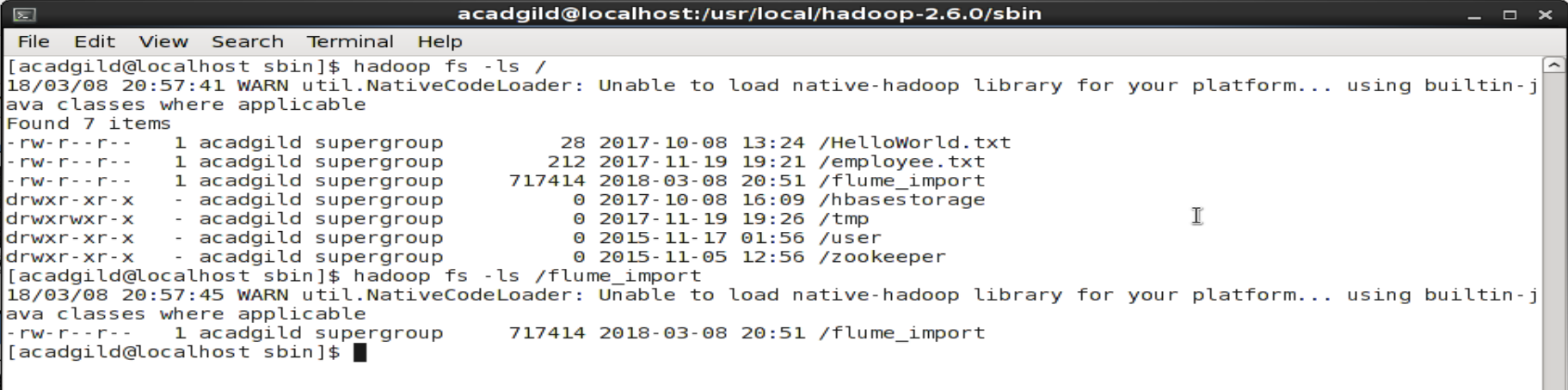
drwxr-xr-x - acadgild supergroup 0 2015-11-05 12:56 /zookeeper

[acadgild@localhost sbin]$ hadoop fs -ls /flume\_import

18/03/08 20:57:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

-rw-r--r-- 1 acadgild supergroup 717414 2018-03-08 20:51 /flume\_import

[acadgild@localhost sbin]$



* **Creating Output Directories For Problem 1 (100% BPL Objective) and Problem 2 (80% BPL Objective)**

[acadgild@localhost sbin]$ hadoop fs -mkdir /districts\_100P\_objectives

18/03/08 21:11:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

[acadgild@localhost sbin]$ hadoop fs -mkdir /districts\_80P\_objectives

18/03/08 21:11:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

^[[A[acadgild@localhost sbin]$ hadoop fs -ls /

18/03/08 21:11:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 9 items

-rw-r--r-- 1 acadgild supergroup 28 2017-10-08 13:24 /HelloWorld.txt

drwxr-xr-x - acadgild supergroup 0 2018-03-08 21:11 /districts\_100P\_objectives

drwxr-xr-x - acadgild supergroup 0 2018-03-08 21:11 /districts\_80P\_objectives

-rw-r--r-- 1 acadgild supergroup 212 2017-11-19 19:21 /employee.txt

-rw-r--r-- 1 acadgild supergroup 717414 2018-03-08 20:51 /flume\_import

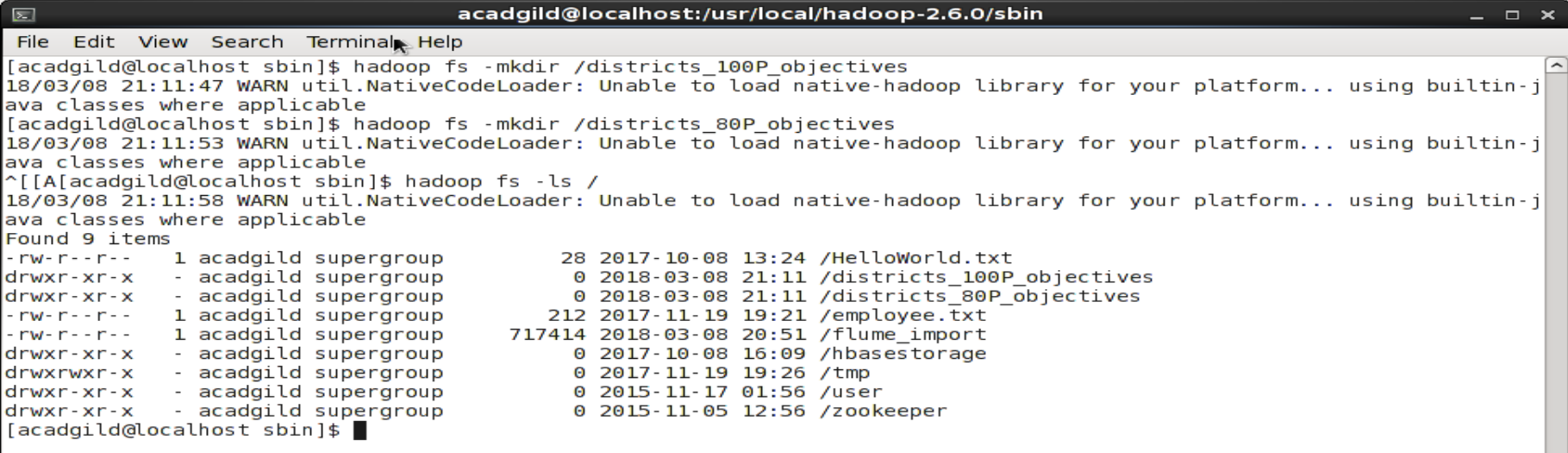
drwxr-xr-x - acadgild supergroup 0 2017-10-08 16:09 /hbasestorage

drwxrwxr-x - acadgild supergroup 0 2017-11-19 19:26 /tmp

drwxr-xr-x - acadgild supergroup 0 2015-11-17 01:56 /user

drwxr-xr-x - acadgild supergroup 0 2015-11-05 12:56 /zookeeper

[acadgild@localhost sbin]$



* **Creating Database and Tables For Problem 1 (100% BPL Objective) and Problem 2 (80% BPL Objective)**

mysql> create database BPLResults;

Query OK, 1 row affected (0.00 sec)

mysql> use BPLResults;

Database changed

mysql> create table districts\_100P\_objectives

-> (

-> name varchar(40)

-> );

Query OK, 0 rows affected (0.01 sec)

mysql> create table districts\_80P\_objectives

-> (

-> name varchar(40)

-> );

Query OK, 0 rows affected (0.01 sec)

mysql> show tables;

+---------------------------+

| Tables\_in\_BPLResults |

+---------------------------+

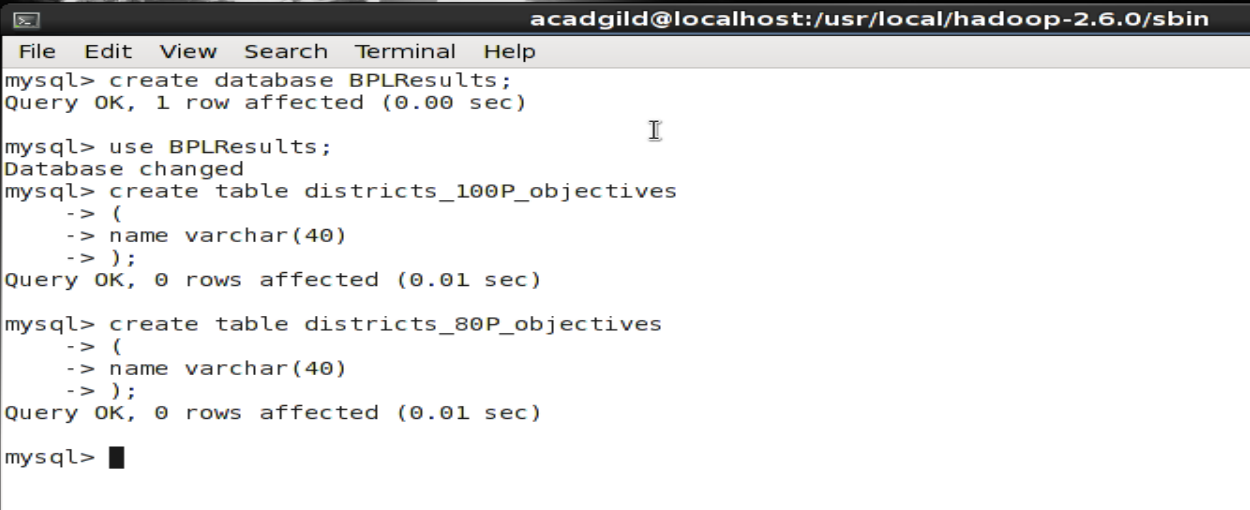
| districts\_100P\_objectives |

| districts\_80P\_objectives |

+---------------------------+

2 rows in set (0.00 sec)

mysql>



* **Loading data from HDFS into Pig**

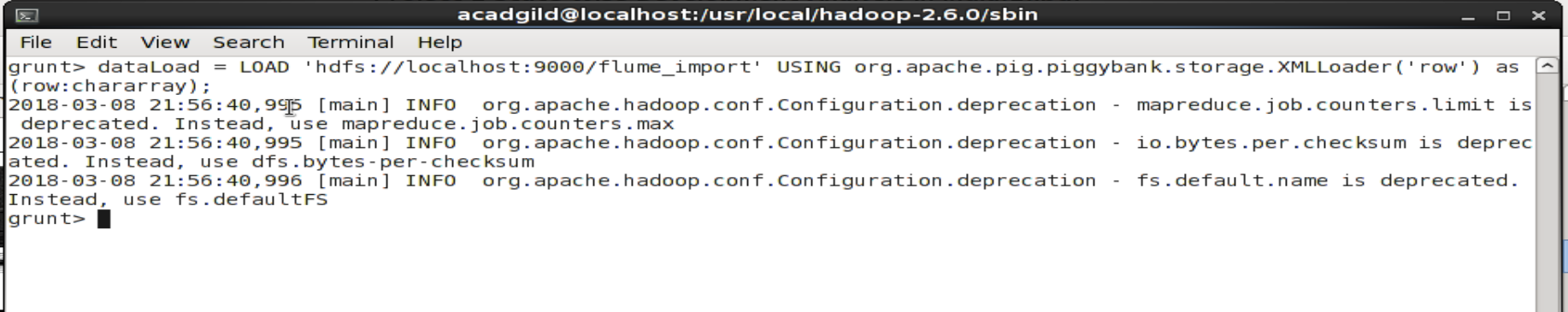
grunt> dataLoad = LOAD 'hdfs://localhost:9000/flume\_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (row:chararray);

2018-03-08 21:56:40,995 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2018-03-08 21:56:40,995 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2018-03-08 21:56:40,996 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

grunt>



* **Generating a Data Structure for Loaded XML Data**

grunt> generatingDataStruct = FOREACH dataLoad GENERATE org.apache.pig.piggybank.evaluation.xml.XPath(row,

>> 'row/State\_Name') AS State\_Name,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/District\_Name') AS District\_Name,org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_IHHL\_BPL') AS Project\_Objectives\_IHHL\_BPL,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_IHHL\_APL') AS Project\_Objectives\_IHHL\_APL,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_IHHL\_TOTAL') AS Project\_Objectives\_IHHL\_TOTAL,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_SCW') AS Project\_Objectives\_SCW,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_Anganwadi\_Toilets') AS Project\_Objectives\_Anganwadi\_Toilets,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_RSM') AS Project\_Objectives\_RSM,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Objectives\_PC') AS Project\_Objectives\_PC,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-IHHL\_BPL') AS Project\_Performance\_IHHL\_BPL,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-IHHL\_APL') AS Project\_Performance\_IHHL\_APL,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-IHHL\_TOTAL') AS Project\_Performance\_IHHL\_TOTAL,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-SCW') AS Project\_Performance\_SCW,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-School\_Toilets') AS

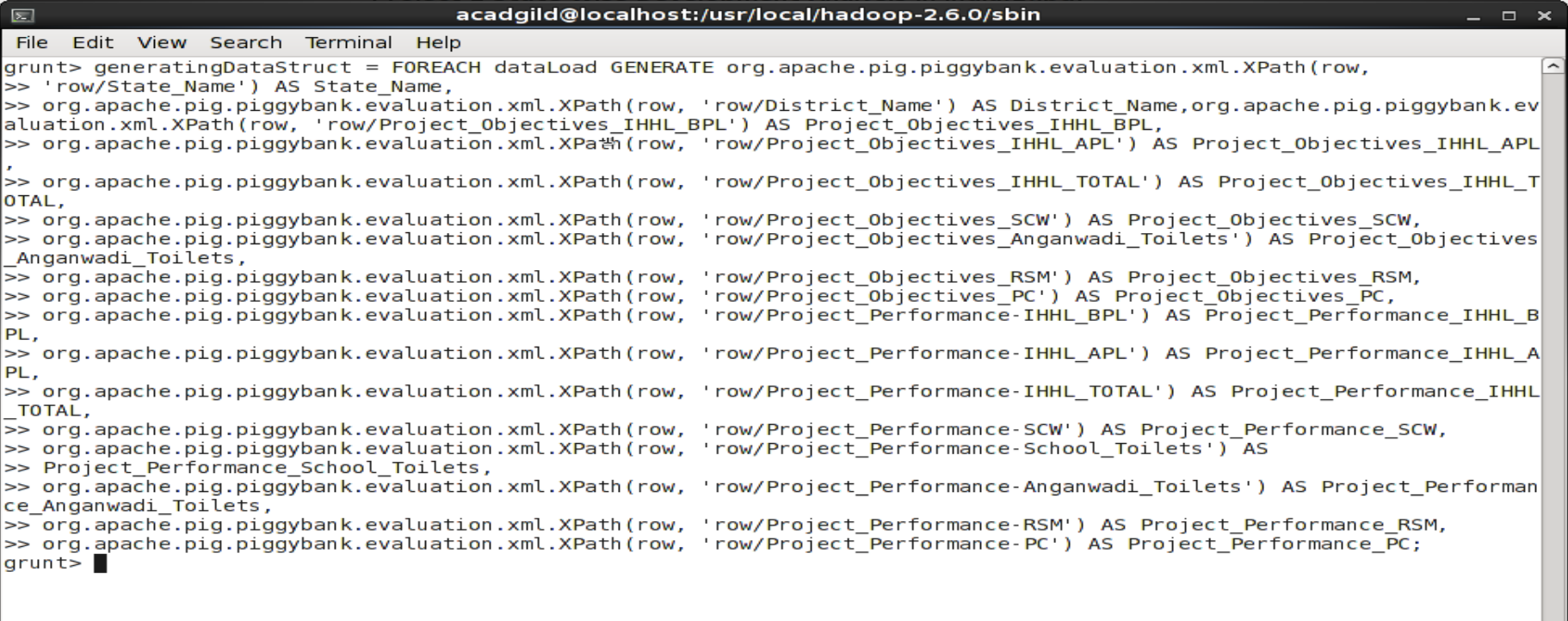
>> Project\_Performance\_School\_Toilets,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-Anganwadi\_Toilets') AS Project\_Performance\_Anganwadi\_Toilets,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-RSM') AS Project\_Performance\_RSM,

>> org.apache.pig.piggybank.evaluation.xml.XPath(row, 'row/Project\_Performance-PC') AS Project\_Performance\_PC;

grunt>



* **Problem-1 Output**

**// Filtering the structured data to have Project\_Performance\_IHHL\_BPL**

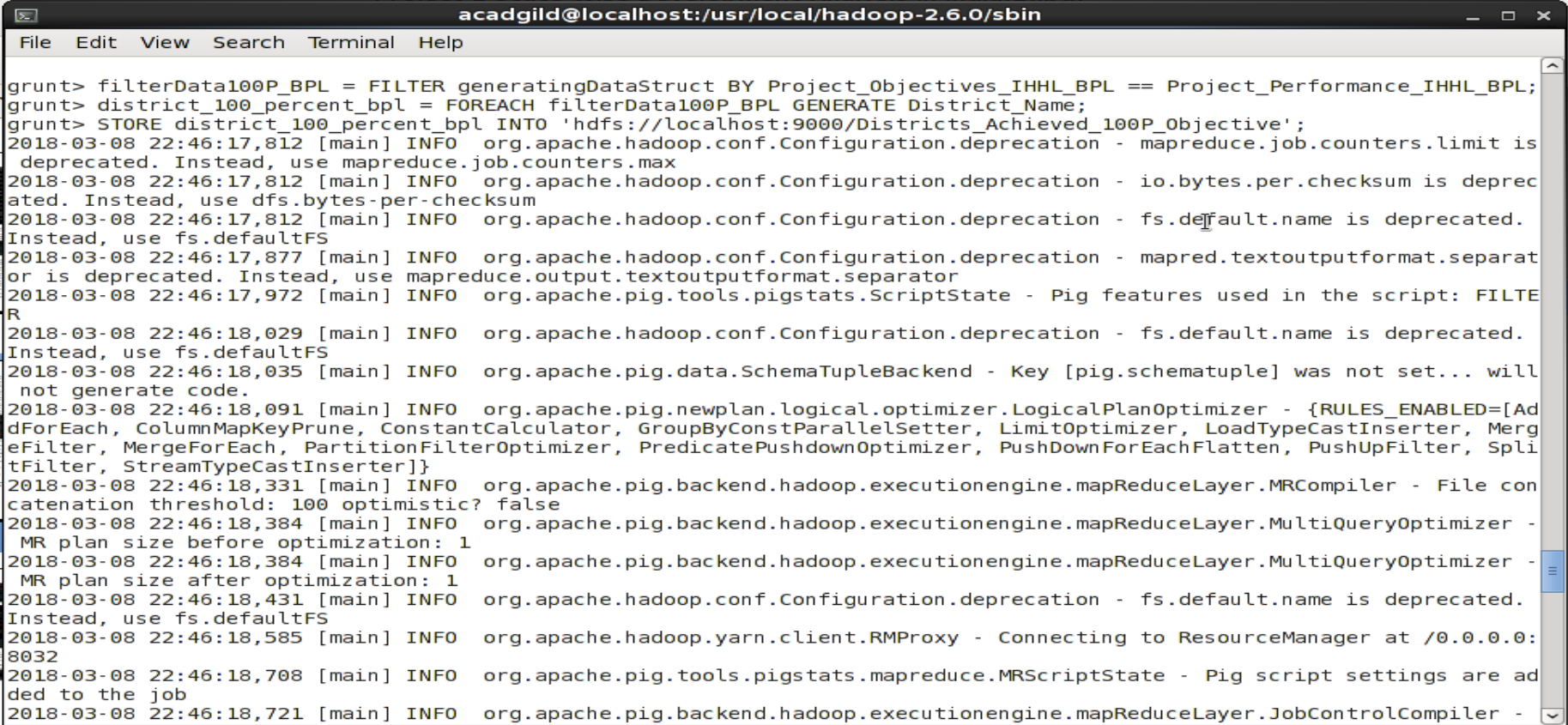
grunt> filterData100P\_BPL = FILTER generatingDataStruct BY Project\_Objectives\_IHHL\_BPL == Project\_Performance\_IHHL\_BPL;

**// Generating District Name for above filtered data**

grunt> district\_100\_percent\_bpl = FOREACH filterData100P\_BPL GENERATE District\_Name;

**// Exporting the data from PIG into a HDFS file**

grunt> STORE district\_100\_percent\_bpl INTO 'hdfs://localhost:9000/Districts\_Achieved\_100P\_Objective';



**// Displaying the file created**

[acadgild@localhost ~]$ hadoop fs -ls /

18/03/08 22:58:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 10 items

drwxr-xr-x - acadgild supergroup 0 2018-03-08 22:46 /Districts\_Achieved\_100P\_Objective

-rw-r--r-- 1 acadgild supergroup 28 2017-10-08 13:24 /HelloWorld.txt

drwxr-xr-x - acadgild supergroup 0 2018-03-08 21:11 /districts\_100P\_objectives

drwxr-xr-x - acadgild supergroup 0 2018-03-08 21:11 /districts\_80P\_objectives

-rw-r--r-- 1 acadgild supergroup 212 2017-11-19 19:21 /employee.txt

-rwxrwxrwx 1 acadgild supergroup 717414 2018-03-08 20:51 /flume\_import

drwxr-xr-x - acadgild supergroup 0 2017-10-08 16:09 /hbasestorage

drwxrwxr-x - acadgild supergroup 0 2018-03-08 22:46 /tmp

drwxr-xr-x - acadgild supergroup 0 2015-11-17 01:56 /user

drwxr-xr-x - acadgild supergroup 0 2015-11-05 12:56 /zookeeper

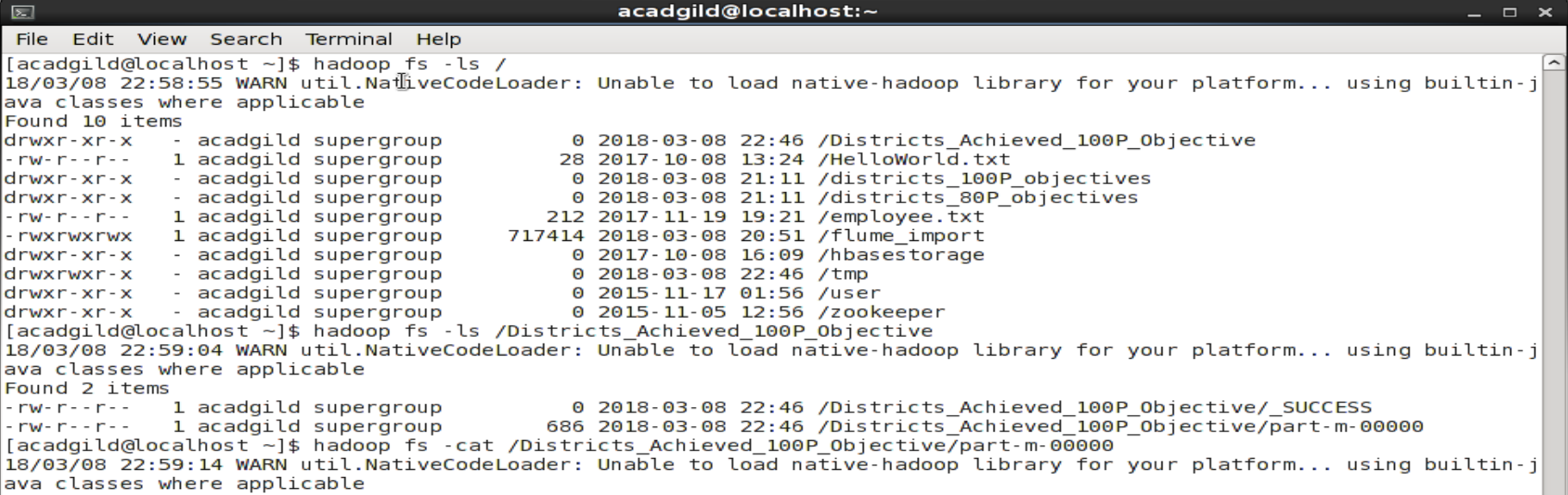
[acadgild@localhost ~]$ hadoop fs -ls /Districts\_Achieved\_100P\_Objective

18/03/08 22:59:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Found 2 items

-rw-r--r-- 1 acadgild supergroup 0 2018-03-08 22:46 /Districts\_Achieved\_100P\_Objective/\_SUCCESS

-rw-r--r-- 1 acadgild supergroup 686 2018-03-08 22:46 /Districts\_Achieved\_100P\_Objective/part-m-00000



**// Displaying the output data inside the file in HDFS**

[acadgild@localhost ~]$ hadoop fs -cat /Districts\_Achieved\_100P\_Objective/part-m-00000

18/03/08 22:59:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

NIZAMABAD

TIRAP

HAILAKANDI

MADHUBANI

NORTH GOA

AHMEDABAD

DANGS

NAVSARI

PORBANDAR

SURAT

FARIDABAD

HISAR

JHAJJAR

MAHENDRAGARH

PANCHKULA

PANIPAT

ROHTAK

SIRSA

HAMIRPUR

KINNAUR

KULLU

LAHAUL & SPITI

SHIMLA

SOLAN

UNA

DEOGHAR

LOHARDAGA

HASSAN

MANGALORE(DAKSHINA KANNADA)

UDUPI

ALAPPUZHA

KOLLAM

KOTTAYAM

KOZHIKODE

PALAKKAD

PATHANAMTHITTA

WAYANAD

GADCHIROLI

SINDHUDURG

WEST GARO HILLS

CHAMPHAI

LAWNGTLAI

HANUMANGARH

ERODE

KARUR

NAMAKKAL

TIRUCHIRAPPALLI

TIRUVANNAMALAI

DHALAI

SOUTH TRIPURA

WEST TRIPURA

AMBEDKAR NAGAR

BALRAMPUR

BAREILLY

BIJNOR

BUDAUN

ETAWAH

FARRUKHABAD

FIROZABAD

GHAZIABAD

HARDOI

JYOTIBA PHULE NAGAR

LUCKNOW

MAHARAJGANJ

MAHOBA

MORADABAD

MUZAFFARNAGAR

PILIBHIT

SONBHADRA

SULTANPUR

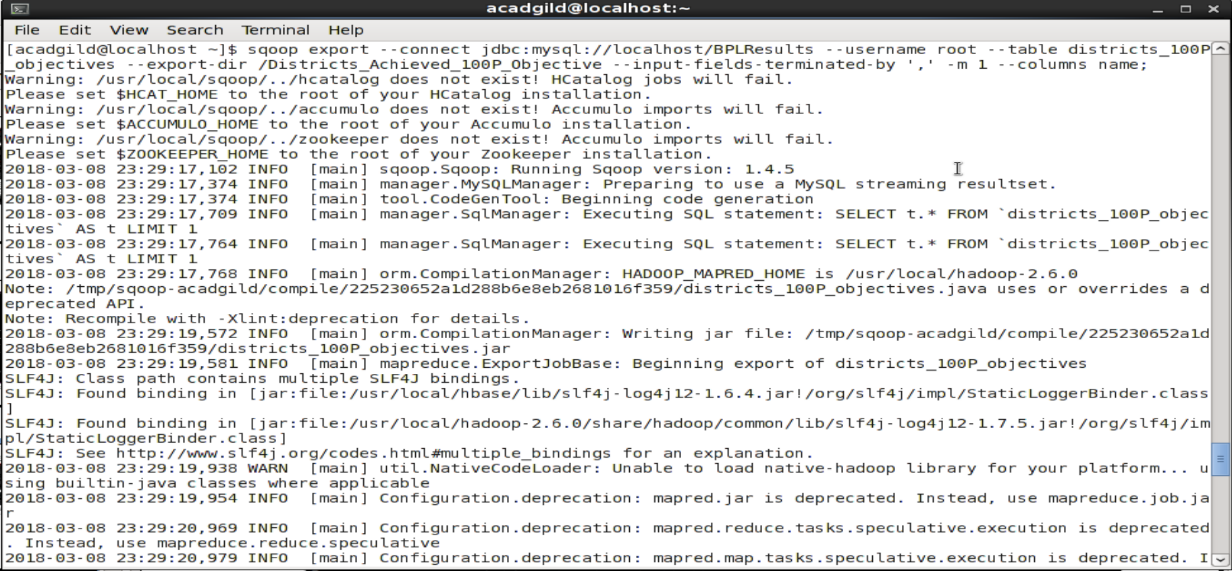
[acadgild@localhost ~]$



* **Exporting data from HDFS file into Mysql table**

**// Exporting the data from HDFS to Mysql**

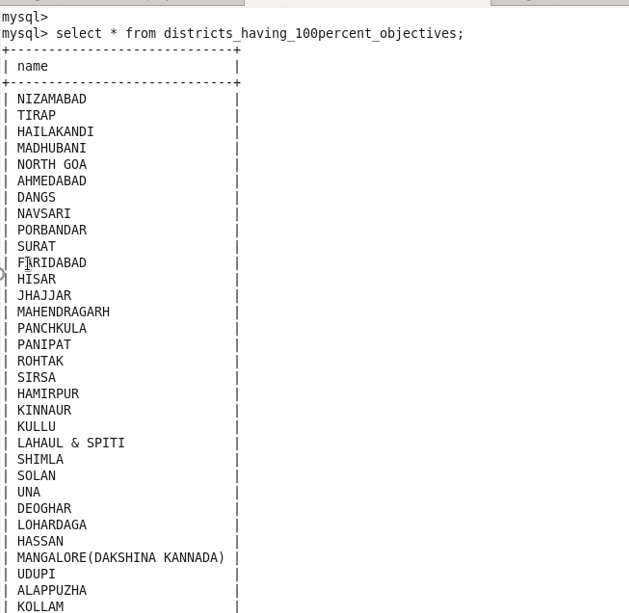
[acadgild@localhost ~]$ sqoop export –connect jdbc:mysql://localhost/bpl\_results --username 'root' --table 'districts\_having\_100percent\_objectives' --export-dir '/districts\_having\_100percent\_objectives' --input-fields-terminated-by ',' -m 1 --columns name;



* **Run the below query to verify the data export in Mysql table**

**// Displaying the data exported to Mysql table**

Mysql> select \* from districts\_having\_100P\_objectives;





* **Problem-2**

**// Write a Java code and Create a jar file for UDF**

package bpludf;

import org.apache.pig.FilterFunc;

import org.apache.pig.backend.executionengine.ExecException;

import org.apache.pig.data.Tuple;

import java.io.IOException;

public class FilterDistrictsHavingEightyPercentBPL

extends FilterFunc{

@Override

public Boolean exec(Tuple input) throws IOException{

try{

if (input == null || input.size()==0) {

return false;

}

Object valueTuple = input.get(0);

if (valueTuple instanceof Tuple) {

Object value1 = ((Tuple) valueTuple).get(0);

Object value2 = ((Tuple) valueTuple).get(1);

long objective\_value = Long.valueOf((String) value1);

long performance\_value = Long.valueOf((String) value2);

if (performance\_value > objective\_value \* 80 /100){

return true;

}

}

}

catch(ExecException ee){

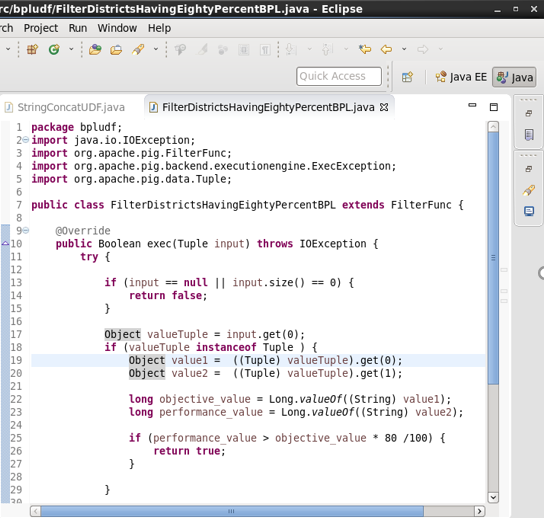
throw ee;

}

return false;

}

}



* **Registering UDF and Filtering the data**

**// Registering the created UDF jar file**

grunt> REGISTER '/home/acadgild/Downloads/bpludf.jar';

2018-03-10 11:30:16,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2018-03-10 11:30:16,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2018-03-10 11:30:16,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

grunt>

**//Using UDF to filter those tuple for which Project\_Performance\_IHHL\_BPL is equal or more than 80% of Project\_Objectives\_IHHL\_BPL**

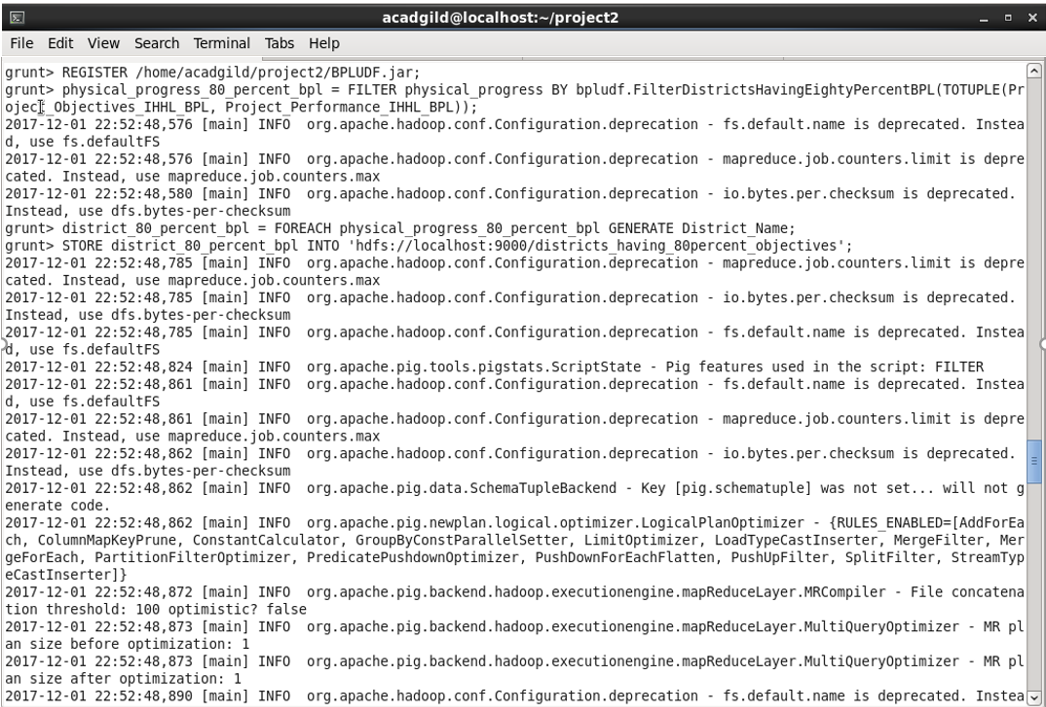
grunt>physical\_progress\_80\_percent\_bpl = FILTER physical\_progress BY bpludf.FilterDistrictsHavingEightyPercentBPL(TOTUPLE(Project\_Objectives\_IHHL\_BPL, Project\_Performance\_IHHL\_BPL));

**//For filtered data generating District Name**

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

**//Storing the above data into HDFS**

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



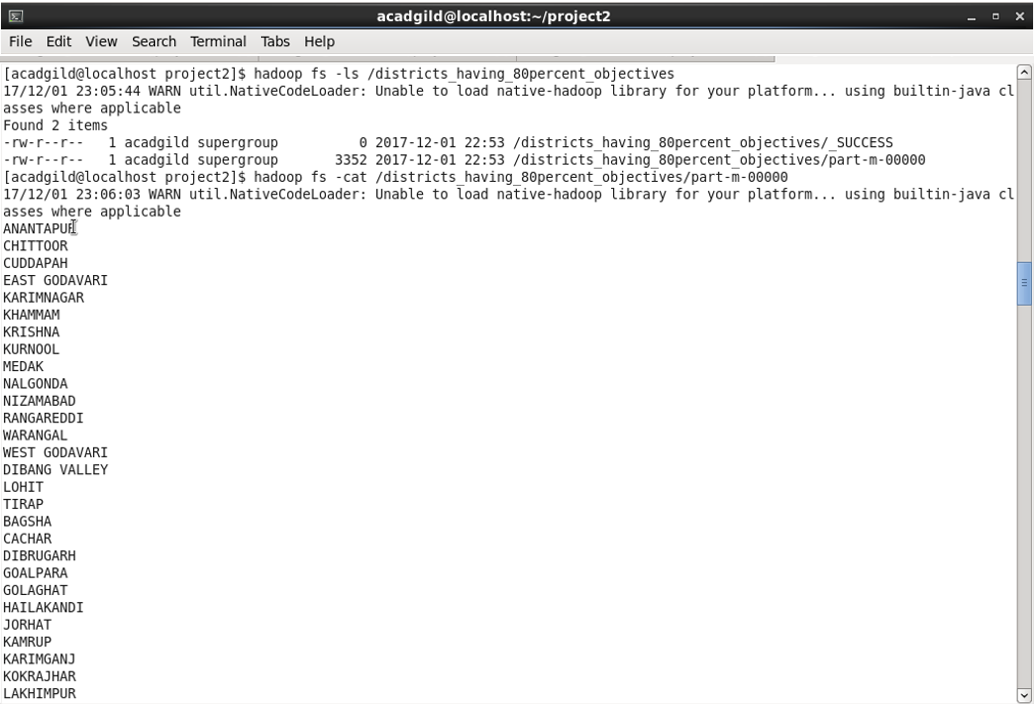
* **Verifying the data stored in the HDFS**

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

[acadgild@localhost ~]$ hadoop fs –ls /districts\_having\_80percent\_objectives

**//Next, use the following HDFS command to show the results**

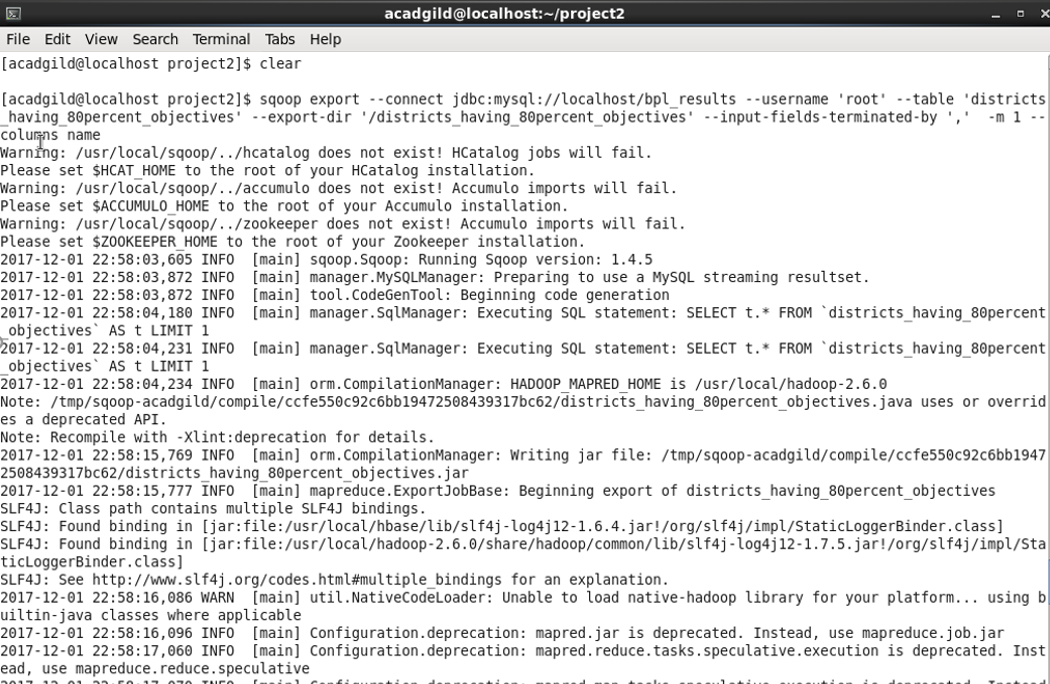
[acadgild@localhost ~]$ hadoop fs –ls /districts\_having\_80percent\_objectives/part-m-00000



* **Using Sqoop exporting data from HDFS to Mysql**

**// Sqoop command to export data from HDFS into mysql table**

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;



* **Run the below query to verify the data export in Mysql table**

**// Displaying the data exported to Mysql table**

Mysql> select \* from districts\_having\_80P\_objectives;

