**Project 1.2**

**Project 1.2 - State-Wise Development Analysis in India**

Student Name: Abarajithan SA

Course: Big Data Hadoop & Spark Training

Contents

[1. Project Overview 2](#_Toc501362638)

[2. Product/Service Description 2](#_Toc501362639)

[2.1 Assumptions 2](#_Toc501362640)

[2.2 Constraints 2](#_Toc501362641)

[3. Requirements 2](#_Toc501362642)

[4. Dataset 3](#_Toc501362643)

[5. Problem statement 3](#_Toc501362644)

[Problem Statement1 - Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop 4](#_Toc501362645)

[Task 1 – Place Dataset in the target using flume, 4](#_Toc501362646)

[Task2 – Create folders in the HDFS to store the outputs, 6](#_Toc501362647)

[Task3 – Create Database and the Tables in the mysql 6](#_Toc501362648)

[Task4 - PIG query to process XML and store into PIG table 7](#_Toc501362649)

[Task5 – Find the districts who achieved 100 percent objective in BPL cards 8](#_Toc501362650)

[Task6 – Verifying the stored results in the HDFS 9](#_Toc501362651)

[Task7 – Export the results into mysql using sqoop 11](#_Toc501362652)

[Task8 – verify the data exported to mysql 11](#_Toc501362653)

[Problem statemet2 - Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop. 14](#_Toc501362654)

[Task1 – Create a PIG UDF using Java 14](#_Toc501362655)

[Task2 - Write PIG query to find out the districts who achieved 80 percent objective in BPL cards 15](#_Toc501362656)

[Task2 – verify the result stored in the HDFS 16](#_Toc501362657)

[Task4 – Export the results into mysql table using sqoop command, 18](#_Toc501362658)

[Task5 – Verify the result in the mysql 18](#_Toc501362659)

# Project Overview

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

* 1. Purpose and Scope of this Specification

The following requirement will be addressed in phase 1 of Project:

* Developing system to handle the incoming log feed and store the information in HadoopCluster (Flume)
* Analyze the data and understand the progress
* Store the results in Hbase/RDBMS

Out of scope

We can use this data and visualization and get more insights

# Product/Service Description

## 2.1 Assumptions

Log will be generated in XML format and stored in a server.

## 2.2 Constraints

Describe any item that will constrain the design options, including

* This system may not be used for searching for now. But it will be used for analysis and saving the relevant information as of now.
* System will be using mySql as a database

# 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

Priority Definitions

The following definitions are intended as a guideline to prioritize requirements.

* **Priority 1** – Create FLUME job for fetching log files from spool directory the data
* **Priority 2** – MapReduce/pig job to preprocess

# Dataset

Download the dataset using the below link:

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

Refer the below steps to understand the actual steps to create the above project.

**Step 1:**

Copy dataset from local file system to HDFS using flume.

Note: use the conf file by downloading from below link.



**Command:**

***flume-ng agent –n agent1 –c conf –f <path to filecopy.conf>***

**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.

# Problem statement

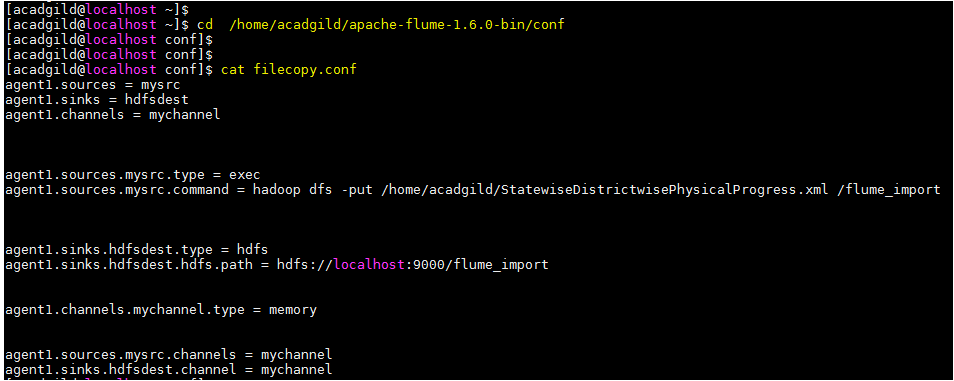
1. Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop
2. Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop.

Project Execution

# Problem Statement1 - Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop

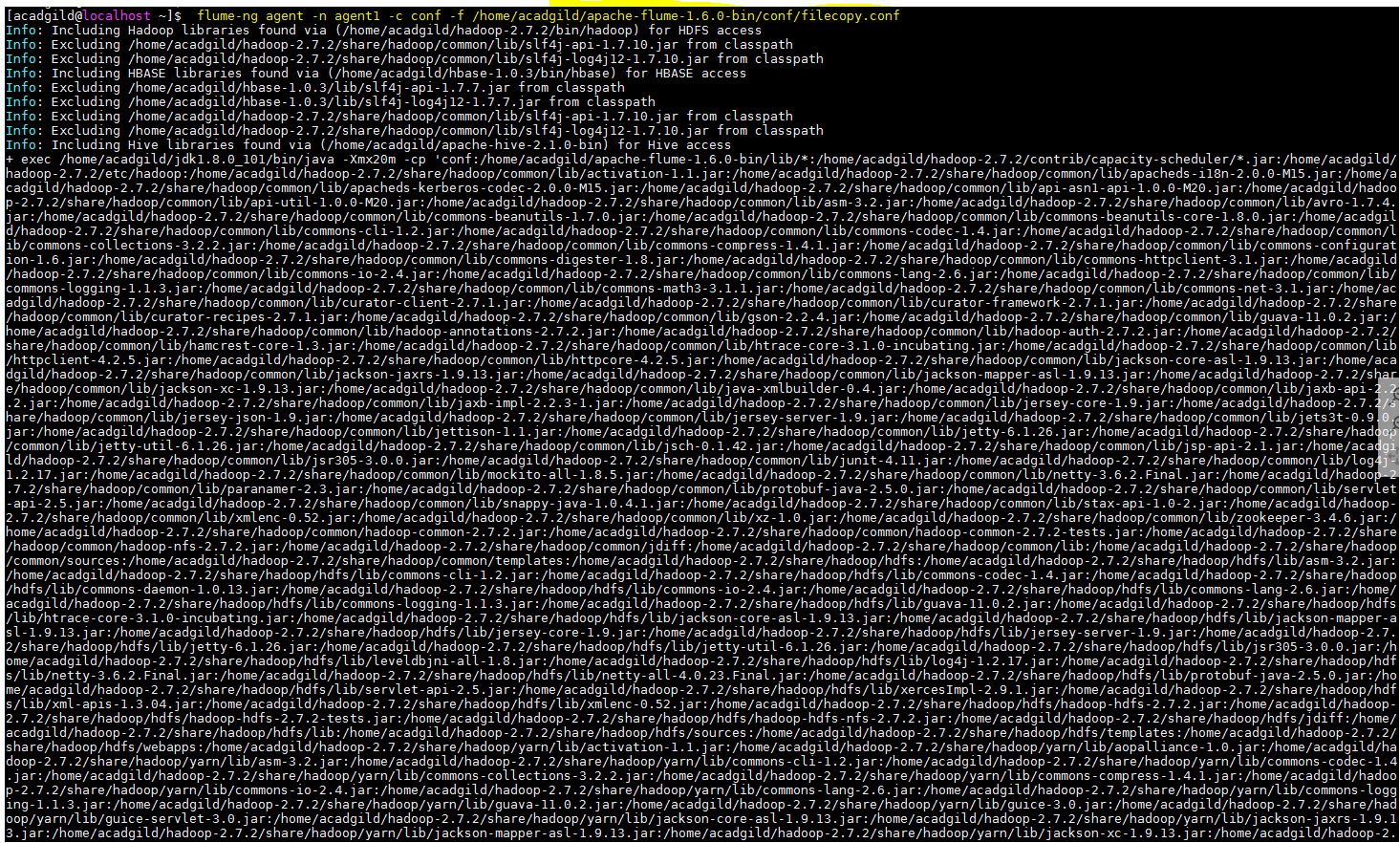
## Task 1 – Place Dataset in the target using flume,

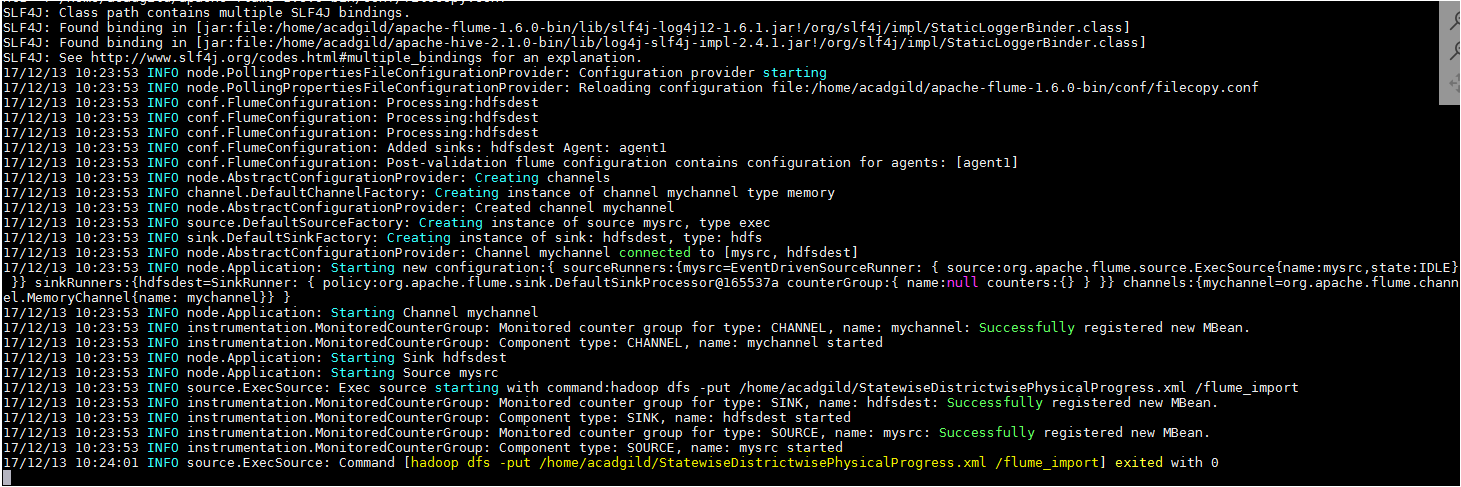
Place the flume config file provided at the location, **/home/acadgild/apache-flume-1.6.0-bin/conf**



Copy the dataset downloaded from the link from local file system to HDFS using flume using the below command,

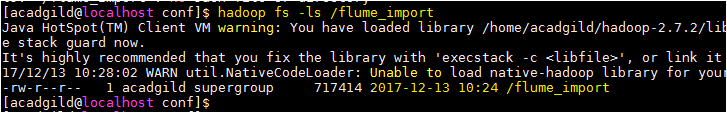
***flume-ng agent -n agent1 -c conf -f /home/acadgild/apache-flume-1.6.0-bin/conf/filecopy.conf***





Verify whether the file is copied in the target,

***Hadoop fs –ls /flume\_import***

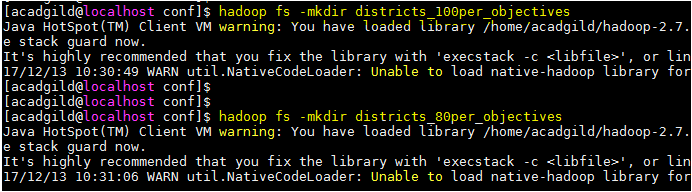


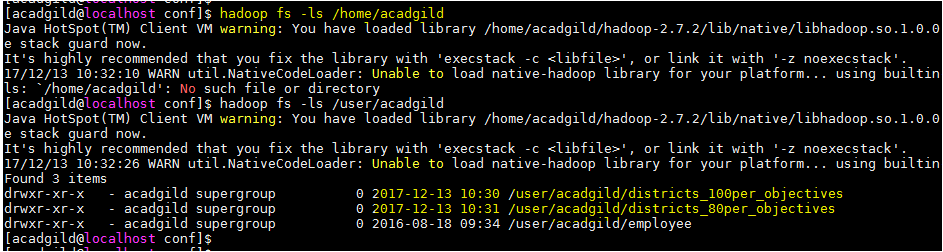
## Task2 – Create folders in the HDFS to store the outputs,

Create 2 folders in the HDFS where we are going to store the output from PIG execution,

***hadoop fs -mkdir districts\_100per\_objectives***

***hadoop fs -mkdir districts\_80per\_objectives***





## Task3 – Create Database and the Tables in the mysql

Start mysql> sudo service mysqld start

Login as root user,

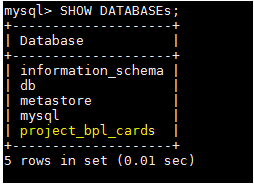
***create database project\_bpl\_cards;***

***use project\_bpl\_cards;***

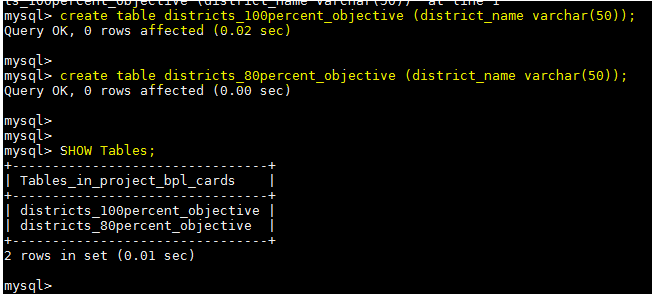
***create table districts\_100percent\_objective (district\_name varchar(50));***

***create table districts\_80percent\_objective (district\_name varchar(50));***









## Task4 - PIG query to process XML and store into PIG table

In this section we are going to Load data from HDFS to PIG alias ***StatewiseDistrictwisePhysicalProgress*** using below query:

PIG Queries,

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

***StatewiseDistrictwisePhysicalProgress = 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 ***StatewiseDistrictwisePhysicalProgress*** which has schema fields same as XML schema hyphen (-) are replaced with underscore (\_)

***PhysicalProgress = FOREACH StatewiseDistrictwisePhysicalProgress 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,***

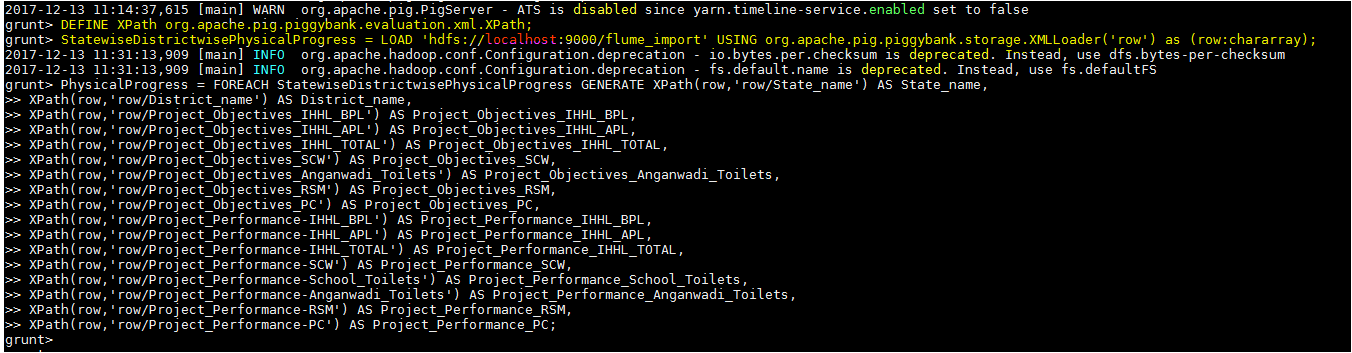
***XPath(row, 'row/Project\_Performance-SCW') AS Project\_Performance\_SCW,***

***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;***



## Task5 – Find the districts who achieved 100 percent objective in BPL cards

Filter the records by ***Project\_Objectives\_IHHL\_BPL*** is equal to ***Project\_Performance\_IHHL\_BPL***

***PhysicalProgress\_100\_percentage\_bpl = FILTER PhysicalProgress BY Project\_Objectives\_IHHL\_BPL == Project\_Performance\_IHHL\_BPL;***

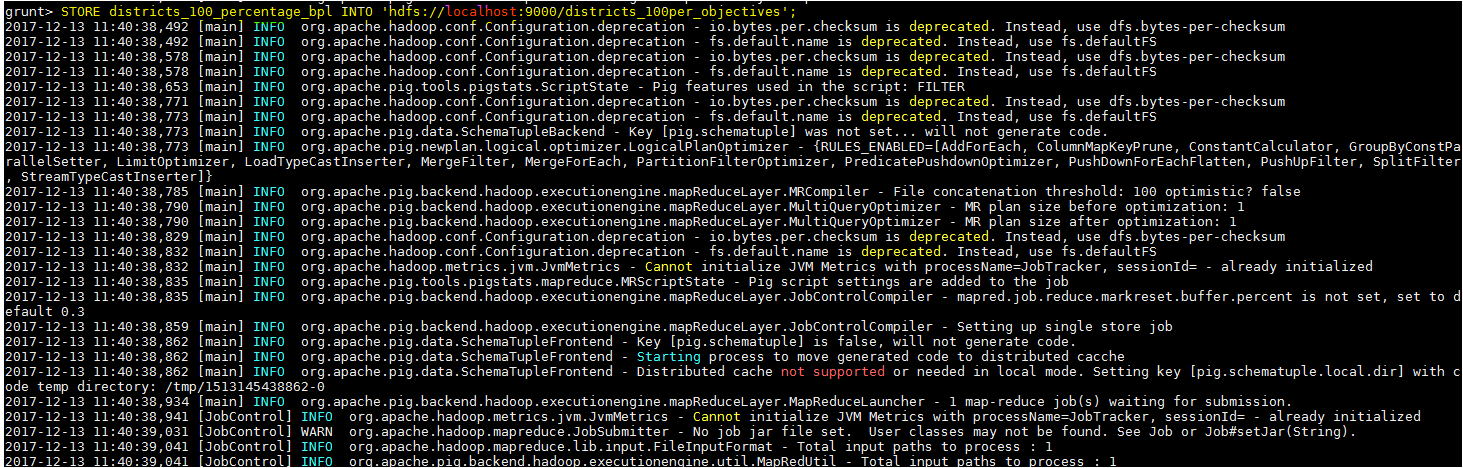
Select only District\_Name column,

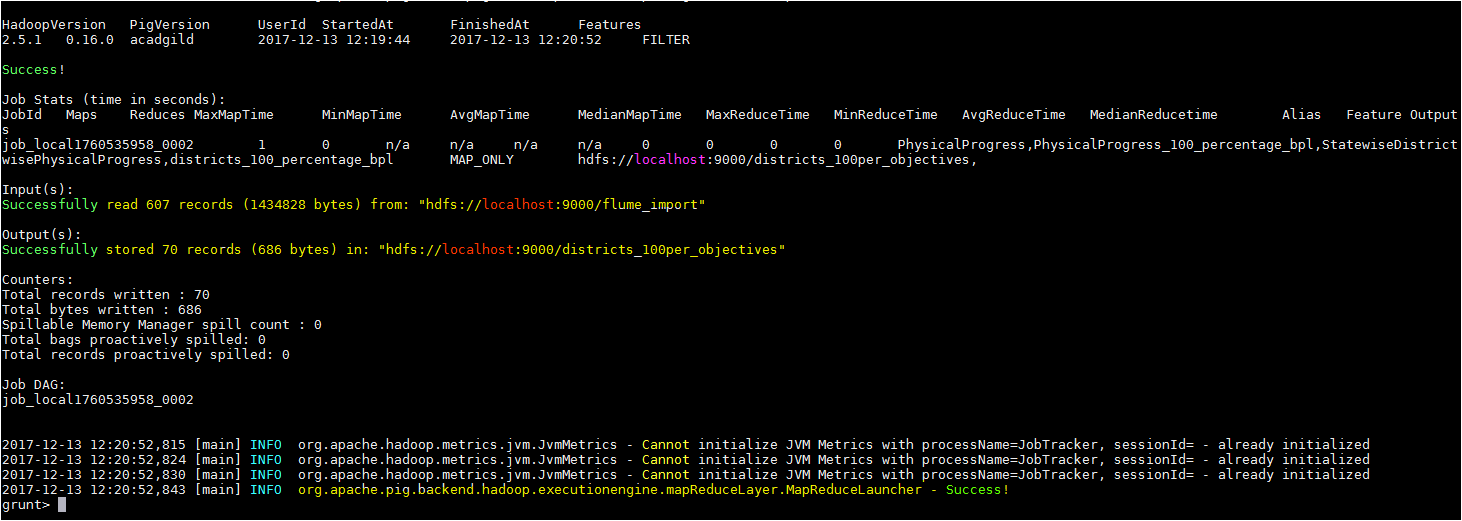
***districts\_100\_percentage\_bpl = FOREACH PhysicalProgress\_100\_percentage\_bpl GENERATE District\_name;***

Now store the data we received from the PIG alias ***districts\_100\_percentage\_bpl*** into the HDFS locationwhere we created at the [Task2](#_Task2_–_Create)

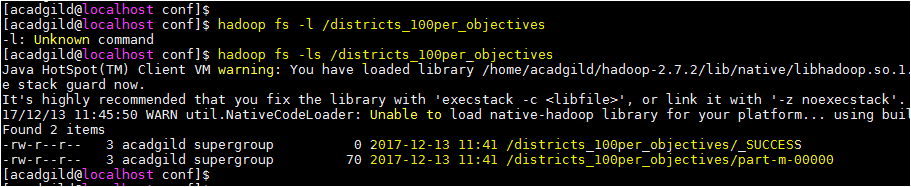
***STORE districts\_100\_percentage\_bpl INTO 'hdfs://localhost:9000/districts\_100per\_objectives';***



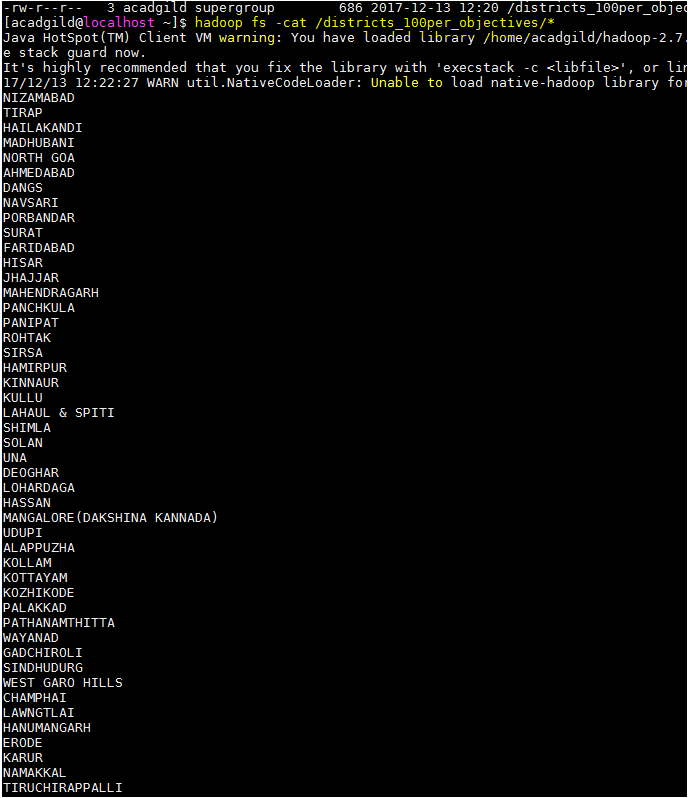




## Task6 – Verifying the stored results in the HDFS

***hadoop fs -ls /districts\_100per\_objectives***

***hadoop fs -cat /districts\_100per\_objectives/\****

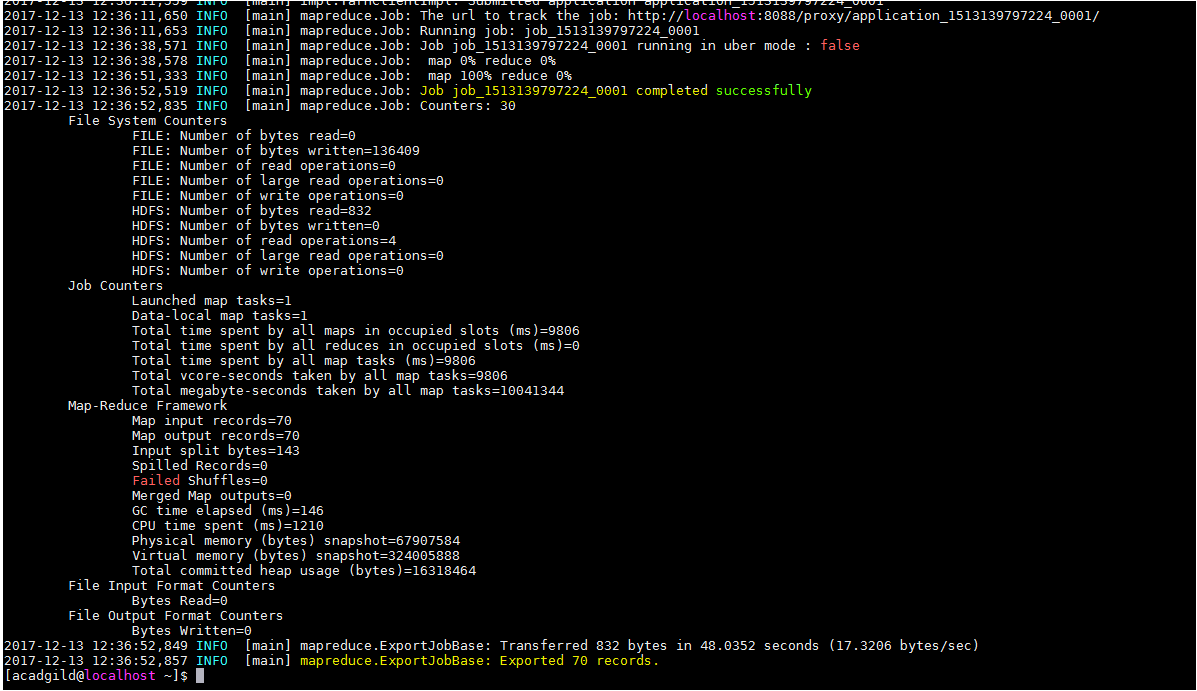
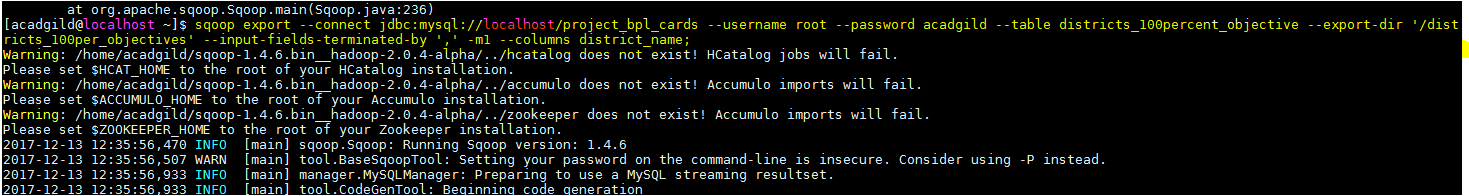




## Task7 – Export the results into mysql using sqoop

Sqoop command to export,

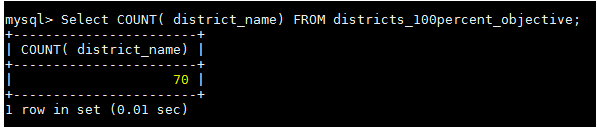
***sqoop export --connect jdbc:mysql://localhost/project\_bpl\_cards --username root --password acadgild --table districts\_100percent\_objective --export-dir '/districts\_100per\_objectives' --input-fields-terminated-by ',' -m1 --columns district\_name***



Task8 – verify the data exported to mysql

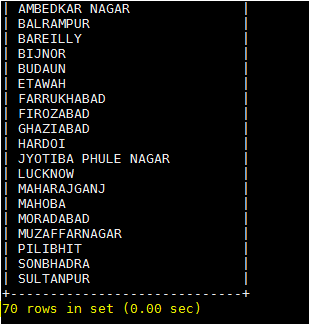
Use the following command in mysql to verify results in mysql

***Select COUNT( district\_name) FROM districts\_100percent\_objective;***



***select \* from districts\_100percent\_objective;***





Thus, as per the problem statement 1, we have successfully exported the result from HDFS to mysql database **project\_bpl\_cards** and into the table **districts\_100percent\_objective.**

# Problem statemet2 - Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop.

## Task1 – Create a PIG UDF using Java

Create a Maven project **StateAnalysis** and Write a Java class **StateAnalysis** 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.**

Java code

**package** StateAnalysis;

**import** java.io.IOException;

**import** org.apache.pig.FilterFunc;

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

**import** org.apache.pig.data.Tuple;

**public** **class** StateAnalysis **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;

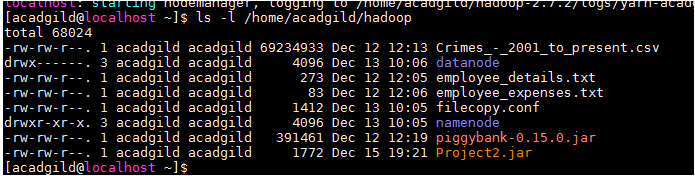
}

**return** **false**;

}

}

Compile this project and Export the project as .jar file to the acadgild local file system. Here we named the jar file as ***Project2.jar.***

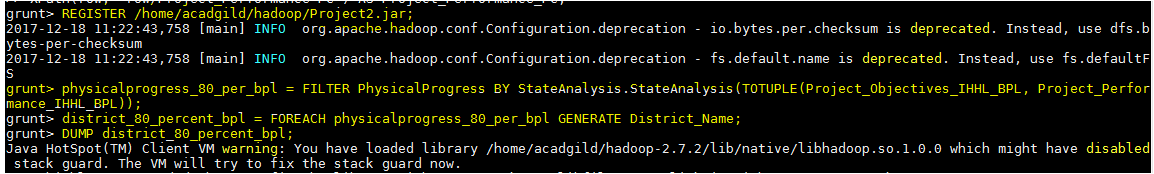


## Task2 - Write PIG query to find out the districts who achieved 80 percent objective in BPL cards

***REGISTER /home/acadgild/hadoop/Project2.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,**

***physicalprogress\_80\_per\_bpl = FILTER PhysicalProgress BY StateAnalysis.StateAnalysis(TOTUPLE(Project\_Objectives\_IHHL\_BPL, Project\_Performance\_IHHL\_BPL));***

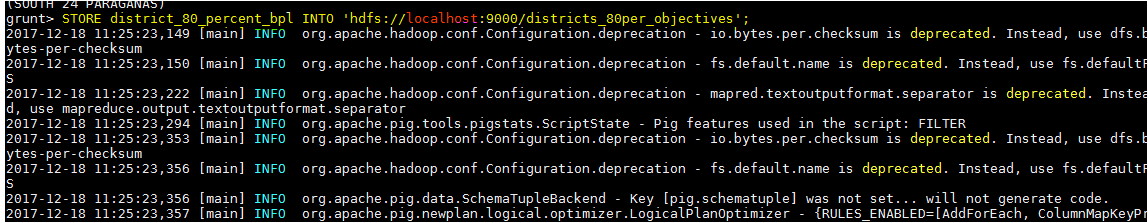


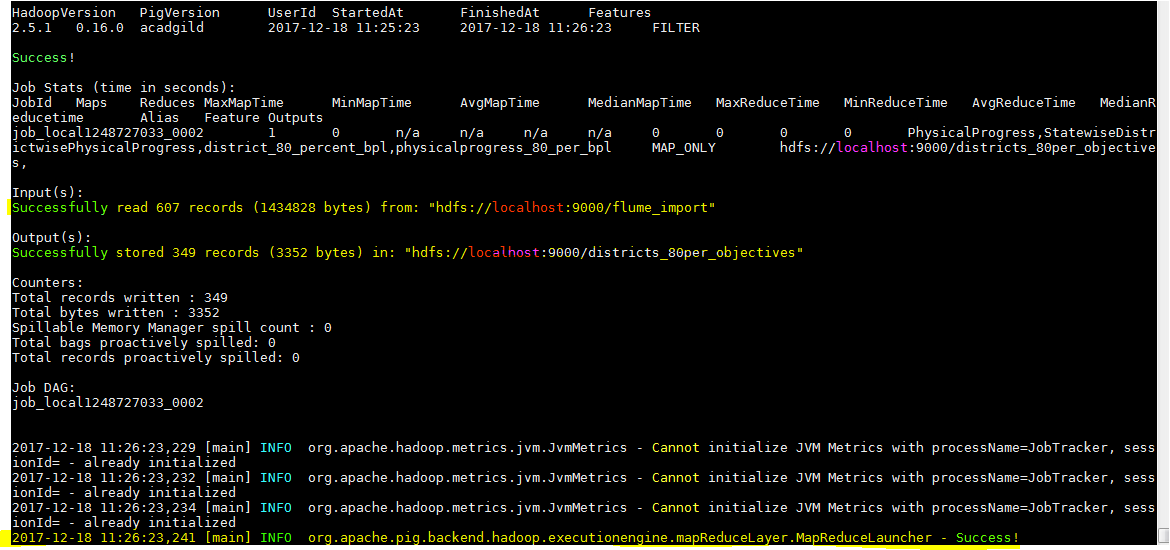
Next, select only **District\_Name** field using command below:

***district\_80\_percent\_bpl = FOREACH physicalprogress\_80\_per\_bpl GENERATE District\_Name;***

Now store the data we received from the PIG alias ***district\_80\_percent\_bpl*** into the HDFS locationwhere we created at the [Task2](#_Task2_–_Create)

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





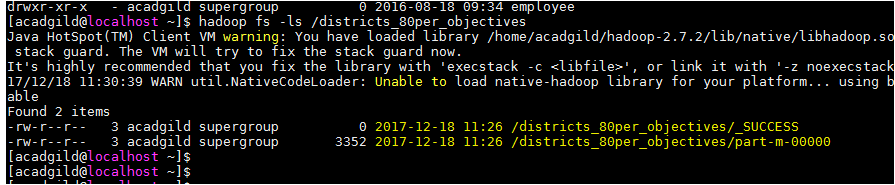
## Task2 – verify the result stored in the HDFS

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

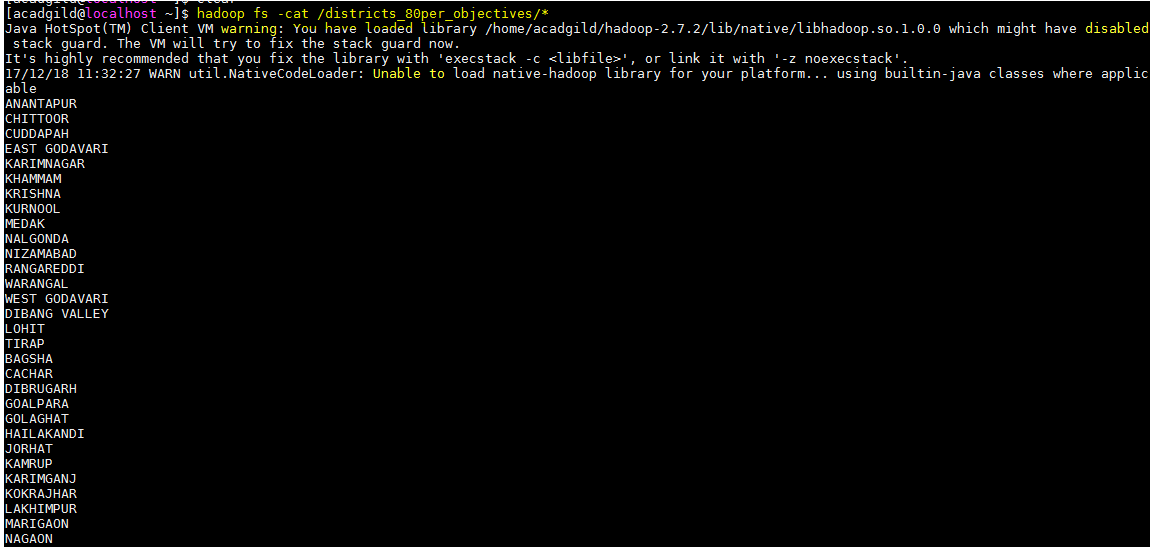
***hadoop fs -ls / districts\_80per\_objectives***

***hadoop fs –ls / districts\_80per\_objectives/part-m-00000***

The output file has been generated in the HDFS location,



***hadoop fs -cat /districts\_80per\_objectives/\****



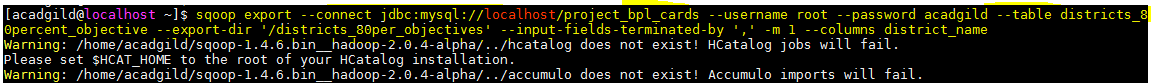


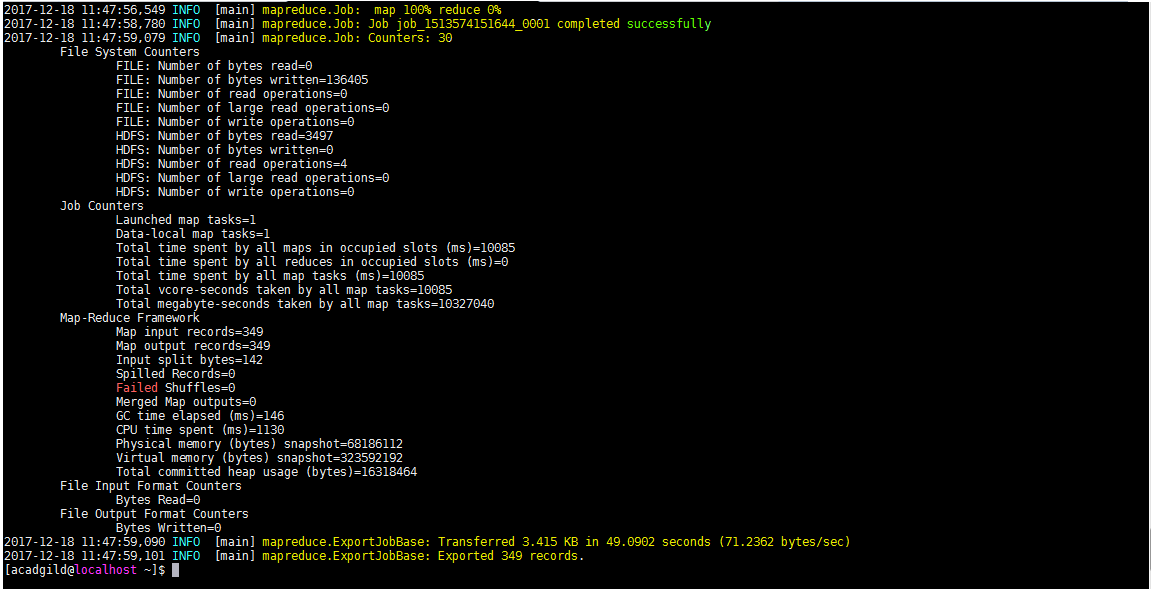
## Task4 – Export the results into mysql table using sqoop command,

In this task we are going use the sqoop to export the desired output stored in the HDFS location **hdfs://localhost:9000/districts\_having\_80percent\_objectives** to the mysql table **districts\_having\_80percent\_objectives** we created in the database **project\_bpl\_cards**

Sqoop command,

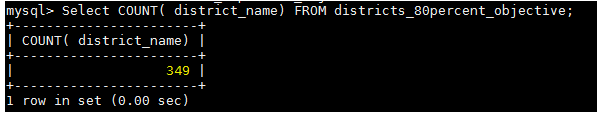
***sqoop export --connect jdbc:mysql://localhost/*project\_bpl\_cards *--username root --password acadgild --table* districts\_80percent\_objective *--export-dir '/districts\_80per\_objectives' --input-fields-terminated-by ',' -m 1 --columns district\_name***





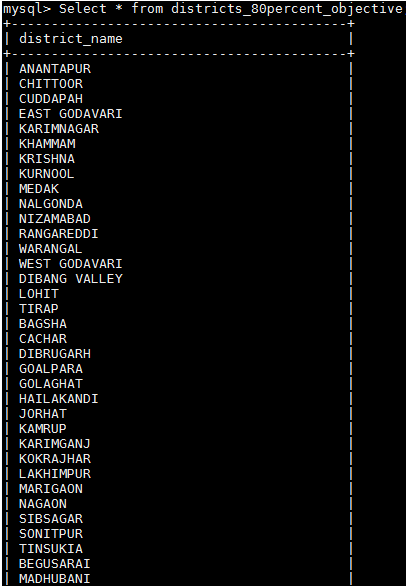
## Task5 – Verify the result in the mysql

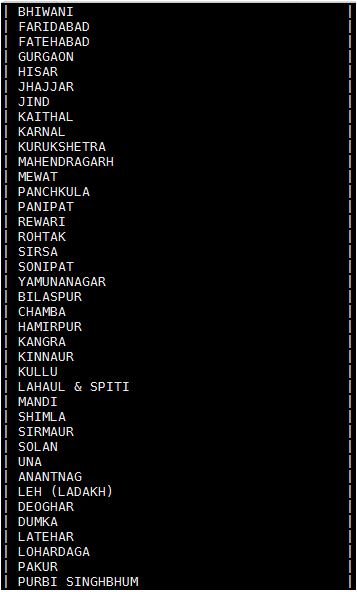
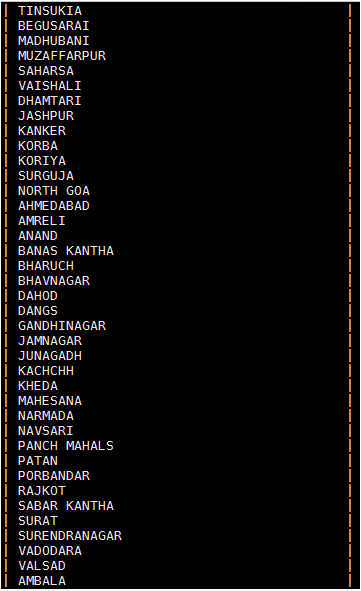
***Select COUNT( district\_name) FROM districts\_80percent\_objective;***

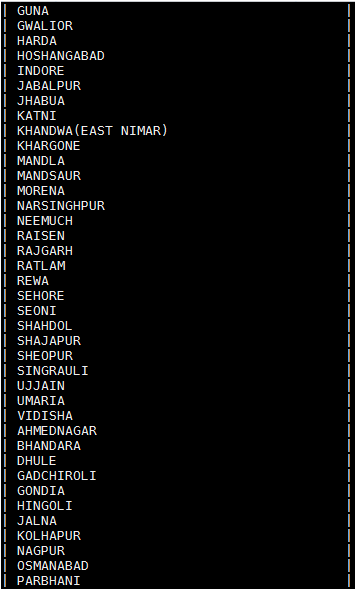
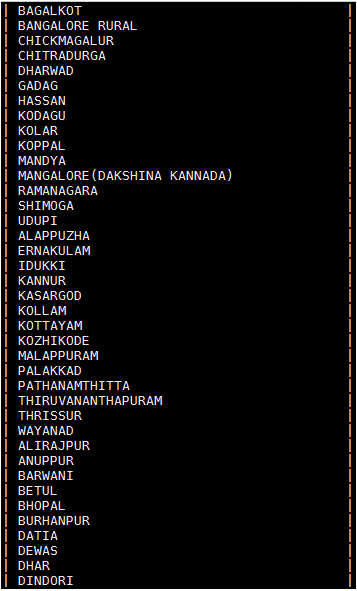


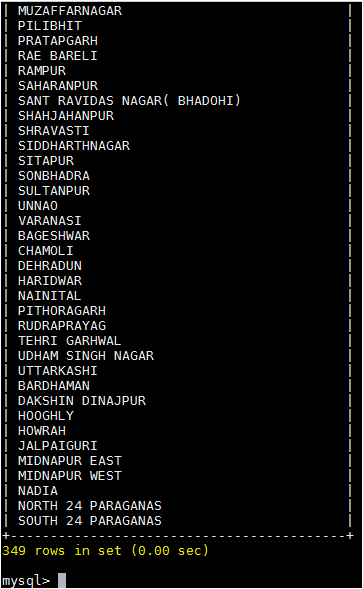
Now, verify the data present in the table,

***Select \* from districts\_80percent\_objective;***









Hence, using PIG UDF we have got the required result and stored into the **mysql** table using **sqoop** commands.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Document\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*