

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

### Contents

1. Project Overview .....	2
2. Product/Service Description.....	2
2.1 Assumptions.....	2
2.2 Constraints.....	2
3. Requirements.....	2
4. Dataset .....	2
5. Problem statement.....	2
Problem Statement1 - Find out the districts who achieved 100 percent objective in BPL cards Export the results to MySQL using sqoop .....	3
Task 1 – Place Dataset in the target using flume.....	4
Task2 – Create folders in the HDFS to store the outputs .....	4
Task3 – Create Database and the Tables in the MySQL.....	5
Task4 - PIG query to process XML and store into PIG table.....	6
Task5 – Find the districts who achieved 100 percent objective in BPL cards.....	7
Task6 – Verifying the stored results in the HDFS.....	7
Task7 – Export the results into MySQL using sqoop.....	10
Task8 – verify the data exported to MySQL.....	11
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.....	13
Task1 – Create a PIG UDF using Java .....	13
Task2 - Write PIG query to find out the districts who achieved 80 percent objective in BPL cards.....	14
Task3 – verify the result stored in the HDFS .....	14
Task4 – Export the results into MySQL table using sqoop command.....	23
Task5 – Verify the result in the MySQL .....	24

## **1. Project Overview**

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

### **1.1 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 Hadoop Cluster (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

## **2. 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

## **3. 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

## 4. 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.

filecopy.conf

### **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.

## 5. 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***

```
[acadgild@localhost ~]$ hadoop fs -ls /
18/02/22 10:37:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 6 items
drwxr-xr-x - acadgild supergroup          0 2018-02-20 14:58 /AcadgildStudent
drwxr-xr-x - acadgild supergroup          0 2018-02-20 14:59 /AcadgildStudent
drwxr-xr-x - acadgild supergroup          0 2018-02-21 19:52 /hbase
drwxr-xr-x - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
drwxrwx--- - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x - acadgild supergroup          0 2018-02-21 11:42 /user
[acadgild@localhost ~]$ hadoop fs -mkdir /flume_import
18/02/22 10:38:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost ~]$ hadoop fs -ls /
18/02/22 10:38:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 7 items
drwxr-xr-x - acadgild supergroup          0 2018-02-20 14:58 /AcadgildStudent
drwxr-xr-x - acadgild supergroup          0 2018-02-20 14:59 /AcadgildStudent
drwxr-xr-x - acadgild supergroup          0 2018-02-22 10:38 /flume_import
drwxr-xr-x - acadgild supergroup          0 2018-02-21 19:52 /hbase
drwxr-xr-x - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
drwxrwx--- - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x - acadgild supergroup          0 2018-02-21 11:42 /user
[acadgild@localhost ~]$ █

18/02/22 10:40:49 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel. Suc
cessfully registered new MBean.
18/02/22 10:46:49 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
18/02/22 10:46:49 INFO node.Application: Starting Sink hdfsdest
18/02/22 10:46:49 INFO node.Application: Starting Source mysrc
18/02/22 10:46:49 INFO source.ExecSource: Exec source starting with command: hadoop dfs -put /home/acadgild/StatewiseDistrict
wisePhysicalProgress.xml /flume_import
18/02/22 10:46:49 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: mysrc: Successf
ully registered new MBean.
18/02/22 10:46:49 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc started
18/02/22 10:46:49 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfsdest: Success
fully registered new MBean.
18/02/22 10:46:49 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfsdest started
18/02/22 10:46:54 INFO source.ExecSource: Command [hadoop dfs -put /home/acadgild/StatewiseDistrictwisePhysicalProgress.xml /
flume_import] exited with 1
[acadgild@localhost:~] █
```

Verify whether the file is copied in the target,

***Hadoopfs -ls /flume\_import***

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ hadoop fs -ls /flume_import  
18/02/22 11:10:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl  
asses where applicable  
Found 1 items  
-rw-r--r-- 1 acadgild supergroup 717414 2018-02-22 10:43 /flume_import/StatewiseDistrictwisePhysicalProgress.xml  
[acadgild@localhost ~]$
```

## **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,

***hadoopfs -mkdir districts\_100per\_objectives***

***hadoopfs -mkdir districts\_80per\_objectives***

```
[acadgild@localhost ~]$ hadoop fs -mkdir /districts_100per_objectives  
18/02/22 11:12:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl  
asses where applicable  
[acadgild@localhost ~]$ hadoop fs -mkdir /districts_80per_objectives  
18/02/22 11:12:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl  
asses where applicable  
[acadgild@localhost ~]$ hadoop fs -ls /  
18/02/22 11:12:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl  
asses where applicable  
Found 9 items  
drwxr-xr-x - acadgild supergroup 0 2018-02-20 14:58 /AcadgildStudent  
drwxr-xr-x - acadgild supergroup 0 2018-02-20 14:59 /AcadgildStudent  
drwxr-xr-x - acadgild supergroup 0 2018-02-22 11:12 /districts_100per_objectives  
drwxr-xr-x - acadgild supergroup 0 2018-02-22 11:12 /districts_80per_objectives  
drwxr-xr-x - acadgild supergroup 0 2018-02-22 10:43 /flume_import  
drwxr-xr-x - acadgild supergroup 0 2018-02-21 19:52 /hbase  
drwxr-xr-x - acadgild supergroup 0 2018-02-02 12:49 /sqoopout111  
drwxrwx--- - acadgild supergroup 0 2018-02-09 11:35 /tmp  
drwxr-xr-x - acadgild supergroup 0 2018-02-21 11:42 /user  
[acadgild@localhost ~]$
```

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

Start mysql>sudo service mysqld start

Login as root user,

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

```

useproject_bpl_cards;
create table districts_100percent_objective (district_namevarchar(50));
create table districts_80percent_objective (district_namevarchar(50));

```

```
mysql> create database project_bpl_cards;
Query OK, 1 row affected (0.67 sec)
```

```
mysql> use project_bpl_cards
Database changed
mysql> create table districts_100percent_objective(district_name varchar(50));
Query OK, 0 rows affected (0.47 sec)
```

```
mysql> create table districts_80percent_objective(district_name varchar(50));
Query OK, 0 rows affected (0.13 sec)
```

```
mysql> █
```

```
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| metastore |
| mysql |
| oozie |
| performance_schema |
| project_bpl_cards |
| simplidb |
| sys |
+-----+
8 rows in set (0.00 sec)
```

```
mysql> show tables;
+-----+
| Tables_in_project_bpl_cards |
+-----+
| districts_100percent_objective |
| districts_80percent_objective |
+-----+
2 rows in set (0.00 sec)
```

```
mysql> █
```

```

grunt> StatewiseDistrictwisePhysicalProgress = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row')as(row:chararray);
2018-02-22 11:32:21,808 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-02-22 11:32:21,808 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> █

```

## **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;*

```
grunt> 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;
grunt> █
```

### **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 location where we created at the Task2

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



```
acadgild@localhost:~  
File Edit View Search Terminal Help  
  
grunt> PhysicalProgress_100_percentage_bpl = FILTER PhysicalProgress BY Project_Objectives_IHHL_BPL == Project_Performance_IH  
HL_BPL;  
grunt> districts_100_percentage_bpl = FOREACH PhysicalProgress_100_percentage_bpl GENERATE District_name;  
grunt> █
```

## **Task6 – Verifying the stored results in the HDFS**

*hadoopfs -ls /districts\_100per\_objectives*

```
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ hadoop fs -cat /distr  
18/02/23 17:42:53 WARN util.NativeCodeLoader:  
asses 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
```

```
File Edit View Search Terminal Help
KOZHIKODE
PALAKKAD
PATHANAMTHITTA
WAYANAD
GADCHIROLI
SINDHUDURG
WEST GARO HILLS
CHAMPHAI
LAWNGTLAI
HANUMANGARH
ERODE
KARUR
NAMAKKAL
TIRUCHIRAPPALLI
TIRUVANNAMALAI
DHAI
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
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

## **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
```



```

mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| metastore |
| mysql |
| oozie |
| performance_schema |
| project_bpl_cards |
| simplidb |
| sys |
+-----+
8 rows in set (0.45 sec)

mysql> use project_bpl_cards
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+-----+
| Tables_in_project_bpl_cards |
+-----+
| districts_100percent_objective |
| districts_80percent_objective |
+-----+
2 rows in set (0.00 sec)

mysql> select count(district_name) from districts_100percent_objective;
+-----+
| count(district_name) |
+-----+
| 70 |
+-----+
1 row in set (0.01 sec)

mysql>

```

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

```

mysql> select * from districts_100percent_objective;
+-----+
| district_name |
+-----+
| 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 |
+-----+

```

Recent - Google Drive - Mozilla Firefox



```

{
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 e)
{
throw e;
}
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 *Statewise.jar*.

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

**REGISTER /home/acadgild/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 location where we created at the Task2

```
STORE district_80_percent_bpl INTO  
'hdfs://localhost:9000/districts_having_80percent_objectives';
```

```
(SOUTH 24 PARAGANAS)  
grunt> STORE district_80_percent_bpl INTO 'hdfs://localhost:9000/districts_80per_objectives';
```

```
Counters:  
Total records written : 349  
Total bytes written : 0  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_local704783937_0005  
  
2018-02-27 12:36:10,243 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized  
2018-02-27 12:36:10,248 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized  
2018-02-27 12:36:10,249 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized  
2018-02-27 12:36:10,259 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success  
!
```

### Task3 – verify the result stored in the HDFS

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

```
hadoopfs -ls / districts_80per_objectives
```

```
hadoopfs -ls / districts_80per_objectives/part-m-00000
```

The output file has been generated in the HDFS location,

```

[acadgild@localhost ~]$ hadoop fs -ls /
18/02/27 12:38:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 14 items
-rw-r--r-- 1 acadgild supergroup 2274 2018-02-25 11:23 /A3-1
-rw-r--r-- 1 acadgild supergroup 2274 2018-02-25 11:24 /A3.jar
drwxr-xr-x - acadgild supergroup 0 2018-02-20 14:58 /AcadgildStudent
drwxr-xr-x - acadgild supergroup 0 2018-02-20 14:59 /AcadgildStudent
drwxr-xr-x - acadgild supergroup 0 2018-02-25 10:57 /Assignment3-1
drwxr-xr-x - acadgild supergroup 0 2018-02-23 17:41 /districts_100per_objective
drwxr-xr-x - acadgild supergroup 0 2018-02-23 17:29 /districts_100per_objectives
drwxr-xr-x - acadgild supergroup 0 2018-02-22 11:12 /districts_80per_objectives
drwxr-xr-x - acadgild supergroup 0 2018-02-27 12:37 /districts_having_80percent_objectives
drwxr-xr-x - acadgild supergroup 0 2018-02-22 10:43 /flume_import
drwxr-xr-x - acadgild supergroup 0 2018-02-24 04:11 /hbase
drwxr-xr-x - acadgild supergroup 0 2018-02-02 12:49 /sqoopout111
drwxrwx--- - acadgild supergroup 0 2018-02-23 16:33 /tmp
drwxr-xr-x - acadgild supergroup 0 2018-02-25 11:44 /user
[acadgild@localhost ~]$ hadoop fs -ls /districts_having_80percent_objectives
18/02/27 13:06:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r-- 1 acadgild supergroup 0 2018-02-27 12:37 /districts_having_80percent_objectives/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 3352 2018-02-27 12:37 /districts_having_80percent_objectives/part-m-000000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

***hadoopfs -cat /districts\_80per\_objectives/\****

```

[acadgild@localhost ~]$ hadoop fs -cat /districts_having_80pe
18/02/27 13:06:42 WARN util.NativeCodeLoader: Unable to load
asses where applicable
ANANTAPUR
CHITTOOR
CUDDAPAH
EAST GODAVARI
KARIMNAGAR
KHAMMAM
KRISHNA
KURNOOL
MEDAK
NALGONDA
NIZAMABAD
RANGAREDDI
WARANGAL
WEST GODAVARI
DIBANG VALLEY
LOHIT
TIRAP
BAGSHA
CACHAR
DIBRUGARH
GOALPARA
GOLAGHAT
HAILAKANDI
JORHAT
KAMRUP
KARIMGANJ
KOKRAJHAR
LAKHIMPUR
MARIGAON
NAGAON
SIBSAGAR
SONITPUR
TINSUKIA
BEGUSARAI
MADHUBANI
MUZAFFARPUR

```



MUZAFFARPUR  
SAHARSA  
VAISHALI  
DHAMTARI  
JASHPUR  
KANKER  
KORBA  
KORIYA  
SURGUJA  
NORTH GOA  
AHMEDABAD  
AMRELI  
ANAND  
BANAS KANTHA  
BHARUCH  
BHAYNAGAR  
DAHOD  
DANGS  
GANDHINAGAR  
JAMNAGAR  
JUNAGADH  
KACHCHH  
KHEDA  
MAHESANA  
NARMADA  
NAVSARI  
PANCH MAHALS  
PATAN  
PORBANDAR  
RAJKOT  
SABAR KANTHA  
SURAT  
SURENDRANAGAR  
VADODARA  
VALSAD  
AMBALA  
BHIWANI  
FARIDABAD  
FATEHABAD

File Edit View Search Terminal Help

FATEHABAD  
BURGAON  
HISAR  
HHAJJAR  
HIND  
KAITHAL  
KARNAL  
KURUKSHETRA  
MAHENDRAGARH  
MEWAT  
PANCHKULA  
PANIPAT  
REWARI  
ROHTAK  
SIRSA  
SONIPAT  
AMUNANAGAR  
BILASPUR  
CHAMBA  
HAMIRPUR  
ANGRA  
CINNAUR  
GULLU  
LAHAUL & SPITI  
MANDI  
SHIMLA  
SIRMAUR  
SOLAN  
JNA  
NANTNAG  
LEH (LADAKH)  
DEOGHAR  
DUMKA  
JATEHAR  
JOHARDAGA  
PAKUR  
PURBI SINGHBHUM  
BAGALKOT  
BANGALORE RURAL

---

File Edit View Search Terminal Help

BANGALORE RURAL  
CHICKMAGALUR  
CHITRADURGA  
DHARWAD  
GADAG  
HASSAN  
KODAGU  
KOLAR  
KOPPAL  
MANDYA  
MANGALORE (DAKSHINA KANNADA)  
RAMANAGARA  
SHIMOGA  
UDUPI  
ALAPPUZHA  
ERNAKULAM  
IDUKKI  
KANNUR  
KASARGOD  
KOLLAM  
KOTTAYAM  
KOZHIKODE  
MALAPPURAM  
PALAKKAD  
PATHANAMTHITTA  
THIRUVANANTHAPURAM  
THRISSUR  
WAYANAD  
ALIRAJPUR  
ANUPPUR  
BARWANI  
BETUL  
BHOPAL  
BURHANPUR  
DATIA  
DEWAS  
DHAR  
DINDORI  
GUNA

File Edit View Search Terminal Help

GUNA  
GWALIOR  
HARDA  
HOSHANGABAD  
INDORE  
JABALPUR  
JHABUA  
KATNI  
KHANDWA(EAST NIMAR)  
KHARGONE  
MANDLA  
MANDSAUR  
MORENA  
NARSINGHPUR  
NEEMUCH  
RAISEN  
RAJGARH  
RATLAM  
REWA  
SEHORE  
SEONI  
SHAHDOL  
SHAJAPUR  
SHEOPUR  
SINGRAULI  
UJJAIN  
UMARIA  
VIDISHA  
AHMEDNAGAR  
BHANDARA  
DHULE  
GADCHIROLI  
GONDIA  
HINGOLI  
JALNA  
KOLHAPUR  
NAGPUR  
OSMANABAD  
PARBHANI

File	Edit	View	Search	Terminal	Help
PARBHANI					
PUNE					
RATNAGIRI					
SANGLI					
SATARA					
SINDHUDURG					
THANE					
WARDHA					
BISHNUPUR					
IMPHAL EAST					
TAMENGLONG					
RI BHOI					
SOUTH GARO HILLS					
WEST GARO HILLS					
CHAMPHAI					
KOLASIB					
LAWNGTLAI					
LUNGLEI					
MAMIT					
SAIHA					
SERCHHIP					
KOHIMA					
MOKOKCHUNG					
PHEK					
BALESWAR					
JAGATSINGHAPUR					
BARNALA					
FATEHGARH SAHIB					
HOSHIARPUR					
JALANDHAR					
KAPURTHALA					
LUDHIANA					
MANSA					
NAWANSHAHR					
S.A.S Nagar					
AJMER					
CHURU					
DUNGARPUR					
GANGANAGAR					

File Edit View Search Terminal Help

GANGANAGAR  
HANUMANGARH  
JAISALMER  
NAGOUR  
SIKAR  
EAST SIKKIM  
NORTH SIKKIM  
SOUTH SIKKIM  
WEST SIKKIM  
COIMBATORE  
CUDDALORE  
DHARMAPURI  
DINDIGUL  
ERODE  
KANCHIPURAM  
KANYAKUMARI (NAGERCOIL)  
KARUR  
MADURAI  
NAMAKKAL  
NILGIRIS (UDHAGAMANDALAM)  
PERAMBALUR  
PUDUKKOTTAI  
RAMANATHAPURAM  
SALEM  
SIVAGANGA  
THENI  
THOOTHUKUDI  
TIRUCHIRAPPALLI  
TIRUNELVELI  
TIRUVANNAMALAI  
TIRUVARUR  
VELLORE  
VIRUDHUNAGAR  
DHALAI  
NORTH TRIPURA  
SOUTH TRIPURA  
WEST TRIPURA  
AGRA  
ALIGARH

File	Edit	View	Search	Terminal	Help
ALIGARH					
ALLAHABAD					
AMBEDKAR NAGAR					
AZAMGARH					
BAGPAT					
BALLIA					
BALRAMPUR					
BANDA					
BARABANKI					
BAREILLY					
BASTI					
BIJNOR					
BUDAUN					
BULANDSHAHR					
CHANDAULI					
CHITRAKOOT					
DEORIA					
ETAH					
ETAWAH					
FAIZABAD					
FARRUKHABAD					
FATEHPUR					
FIROZABAD					
GAUTAM BUDDHA NAGAR					
GHAZIABAD					
GHAZIPUR					
GONDA					
GORAKHPUR					
HAMIRPUR					
HARDOI					
JALAUN					
JAUNPUR					
JHANSI					
JYOTIBA PHULE NAGAR					
KANNAUJ					
KANPUR DEHAT					
KANPUR NAGAR					
KAUSHAMBI					
KUSHINAGAR					

```
File Edit View Search Terminal Help
KUSHINAGAR
LAKHIMPUR KHERI
LALITPUR
LUCKNOW
MAHAMAYA NAGAR(HATHRAS)
MAHARAJGANJ
MAHOBA
MAINPURI
MATHURA
MAU
MEERUT
MIRZAPUR
MORADABAD
MUZAFFARNAGAR
PILIBHIT
PRATAPGARH
RAE BARELI
RAMPUR
SAHARANPUR
SANT RAVIDAS NAGAR( BHADOHI)
SHAHJAHANPUR
SHRAVASTI
SIDDHARTHANAGAR
SITAPUR
SONBHADRA
SULTANPUR
UNNAO
VARANASI
BAGESHWAR
CHAMOLI
DEHRADUN
HARIDWAR
NAINITAL
PITHORAGARH
RUDRAPRAYAG
TEHRI GARHWAL
UDHAM SINGH NAGAR
UTTARKASHI
BARDHAMAN

BARDHAMAN
DAKSHIN DINAJPUR
HOOGHLY
HOWRAH
JALPAIGURI
MIDNAPUR EAST
MIDNAPUR WEST
NADIA
NORTH 24 PARAGANAS
SOUTH 24 PARAGANAS
[acadgild@localhost ~]$
```

#### **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 -m 1 --connect jdbc:mysql://localhost/project_bpl_cards--username root --password Root@123 --table districts_80percent_objective --export-dir
```



***'/districts\_having\_80per\_objectives'***

***;***

```
[acadgild@localhost ~]$ sqoop export -m 1 --connect jdbc:mysql://localhost/project_bpl_cards --username root --password Root@123 --table districts_80percent_objective --export-dir '/districts_having_80per_objectives';
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```

```
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=127630
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=3508
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=4
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=0
Job Counters
  Launched map tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=11202
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=11202
  Total vcore-milliseconds taken by all map tasks=11202
  Total megabyte-milliseconds taken by all map tasks=11470848
Map-Reduce Framework
  Map input records=349
  Map output records=349
  Input split bytes=153
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=113
  CPU time spent (ms)=2350
  Physical memory (bytes) snapshot=105934848
  Virtual memory (bytes) snapshot=2062385152
  Total committed heap usage (bytes)=32571392
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
18/02/27 13:27:15 INFO mapreduce.ExportJobBase: Transferred 3.4258 KB in 38.806 seconds (90.3983 bytes/sec)
18/02/27 13:27:15 INFO mapreduce.ExportJobBase: Exported 349 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Recent - Google Drive - Mozilla Firefox

## **Task5 – Verify the result in the mysql**

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

```
mysql> select count(district_name) from districts_80percent_objective;
+-----+
| count(district_name) |
+-----+
|                349 |
+-----+
1 row in set (0.05 sec)

mysql>
```

Now, verify the data present in the table

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

File Edit View Search Terminal Help

```
mysql> select * from districts_80percent_objective;
```

district_name
ANANTAPUR
CHITTOOR
CUDDAPAH
EAST GODAVARI
KARIMNAGAR
KHAMMAM
KRISHNA
KURNOOL
MEDAK
NALGONDA
NIZAMABAD
RANGAREDDI
WARANGAL
WEST GODAVARI
DIBANG VALLEY
LOHIT
TIRAP
BAGSHA
CACHAR
DIBRUGARH
GOALPARA
GOLAGHAT
HAILAKANDI
JORHAT
KAMRUP
KARIMGANJ
KOKRAJHAR
LAKHIMPUR
MARIGAON
NAGAON
SIBSAGAR
SONITPUR
TINSUKIA

File	Edit	View	Search	Terminal	Help
MADHUBANI					
MUZAFFARPUR					
SAHARSA					
VAISHALI					
DHMTARI					
JASHPUR					
KANKER					
KORBA					
KORIYA					
SURGUJA					
NORTH GOA					
AHMEDABAD					
AMRELI					
ANAND					
BANAS KANTHA					
BHARUCH					
BHAVNAGAR					
DAHOD					
DANGS					
GANDHINAGAR					
JAMNAGAR					
JUNAGADH					
KACHCHH					
KHEDA					
MAHESANA					
NARMADA					
NAVSARI					
PANCH MAHALS					
PATAN					
PORBANDAR					
RAJKOT					
SABAR KANTHA					
SURAT					
SURENDRANAGAR					
VADODARA					
VALSAD					
AMBALA					
BHIWANI					
FARIDABAD					

File Edit View Search Terminal Help

FARIDABAD  
FATEHABAD  
GURGAON  
HISAR  
JHAJJAR  
JIND  
KAITHAL  
KARNAL  
KURUKSHETRA  
MAHENDRAGARH  
MEWAT  
PANCHKULA  
PANIPAT  
REWARI  
ROHTAK  
SIRSA  
SONIPAT  
YAMUNANAGAR  
BILASPUR  
CHAMBA  
HAMIRPUR  
KANGRA  
KINNAUR  
KULLU  
LAHAUL & SPITI  
MANDI  
SHIMLA  
SIRMAUR  
SOLAN  
UNA  
ANANTNAG  
LEH (LADAKH)  
DEOGHAR  
DUMKA  
LATEHAR  
LOHARDAGA  
PAKUR  
PURBI SINGHBHUM  
BAGALKOT

\*S.txt (~) - gedit

BAGALKOT	
BANGALORE RURAL	
CHICKMAGALUR	
CHITRADURGA	
DHARWAD	
GADAG	
HASSAN	
KODAGU	
KOLAR	
KOPPAL	
MANDYA	
MANGALORE (DAKSHINA KANNADA)	
RAMANAGARA	
SHIMOGA	
UDUPI	
ALAPPUZHA	
ERNAKULAM	
IDUKKI	
KANNUR	
KASARGOD	
KOLLAM	
KOTTAYAM	
KOZHIKODE	
MALAPPURAM	
PALAKKAD	
PATHANAMTHITTA	
THIRUVANANTHAPURAM	
THRISSUR	
WAYANAD	
ALIRAJPUR	
ANUPPUR	
BARWANI	
BETUL	
BHOPAL	
BURHANPUR	
DATIA	
DEWAS	
DHAR	
DINDORI	

DINDORI
GUNA
GWALIOR
HARDA
HOSHANGABAD
INDORE
JABALPUR
JHABUA
KATNI
KHANDWA (EAST NIMAR)
KHARGONE
MANDLA
MANDSAUR
MORENA
NARSINGHPUR
NEEMUCH
RAISEN
RAJGARH
RATLAM
REWA
SEHORE
SEONI
SHAHDOL
SHAJAPUR
SHEOPUR
SINGRAULI
UJJAIN
UMARIA
VIDISHA
AHMEDNAGAR
BHANDARA
DHULE
GADCHIROLI
GONDIA
HINGOLI
JALNA
KOLHAPUR
NAGPUR
OSMANABAD

OSMANABAD
PARBHANI
PUNE
RATNAGIRI
SANGLI
SATARA
SINDHUDURG
THANE
WARDHA
BISHNUPUR
IMPHAL EAST
TAMENGLONG
RI BHOI
SOUTH GARO HILLS
WEST GARO HILLS
CHAMPHAI
KOLASIB
LAWNGTLAI
LUNGLEI
MAMIT
SAIHA
SERCHHIP
KOHIMA
MOKOKCHUNG
PHEK
BALESWAR
JAGATSINGHAPUR
BARNALA
FATEHGARH SAHIB
HOSHIARPUR
JALANDHAR
KAPURTHALA
LUDHIANA
MANSA
NAWANSHAHR
S.A.S Nagar
AJMER
CHURU
DUNGARPUR

DUNGARPUR
GANGANAGAR
HANUMANGARH
JAISALMER
NAGAUR
SIKAR
EAST SIKKIM
NORTH SIKKIM
SOUTH SIKKIM
WEST SIKKIM
COIMBATORE
CUDDALORE
DHARMAPURI
DINDIGUL
ERODE
KANCHIPURAM
KANYAKUMARI (NAGERCOIL)
KARUR
MADURAI
NAMAKKAL
NILGIRIS (UDHAGAMANDALAM)
PERAMBALUR
PUDUKKOTTAI
RAMANATHAPURAM
SALEM
SIVAGANGA
THENI
THOOTHUKUDI
TIRUCHIRAPPALLI
TIRUNELVELI
TIRUVANNAMALAI
TIRUVARUR
VELLORE
VIRUDHUNAGAR
DHALAI
NORTH TRIPURA
SOUTH TRIPURA
WEST TRIPURA
AGRA

AGRA
ALIGARH
ALLAHABAD
AMBEDKAR NAGAR
AZAMGARH
BAGPAT
BALLIA
BALRAMPUR
BANDA
BARABANKI
BAREILLY
BASTI
BIJNOR
BUDAUN
BULANDSHAHR
CHANDAULI
CHITRAKOOT
DEORIA
ETAH
ETAWAH
FAIZABAD
FARRUKHABAD
FATEHPUR
FIROZABAD
GAUTAM BUDDHA NAGAR
GHAZIABAD
GHAZIPUR
GONDA
GORAKHPUR
HAMIRPUR
HARDOI
JALAUN
JAUNPUR
JHANSI
JYOTIBA PHULE NAGAR
KANNAUJ
KANPUR DEHAT
KANPUR NAGAR
KAUSHAMBI



KAUSHAMBI
KUSHINAGAR
LAKHIMPUR KHERI
LALITPUR
LUCKNOW
MAHAMAYA NAGAR(HATHRAS)
MAHARAJGANJ
MAHOBA
MAINPURI
MATHURA
MAU
MEERUT
MIRZAPUR
MORADABAD
MUZAFFARNAGAR
PILIBHIT
PRATAPGARH
RAE BARELI
RAMPUR
SAHARANPUR
SANT RAVIDAS NAGAR( BHADOHI)
SHAHJAHANPUR
SHRAVASTI
SIDDHARTHANAGAR
SITAPUR
SONBHADRA
SULTANPUR
UNNAO
VARANASI
BAGESHWAR
CHAMOLI
DEHRADUN
HARIDWAR
NAINITAL
PITHORAGARH
RUDRAPRAYAG
TEHRI GARHWAL
UDHAM SINGH NAGAR
UTTARKASHI

```
| UTTARKASHI |
| BARDHAMAN |
| DAKSHIN DINAJPUR |
| HOOGHLY |
| HOWRAH |
| JALPAIGURI |
| MIDNAPUR EAST |
| MIDNAPUR WEST |
| NADIA |
| NORTH 24 PARAGANAS |
| SOUTH 24 PARAGANAS |
+-----+
349 rows in set (0.00 sec)

mysql> █
```

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