**ACADGILD**

**MAJOR PROJECT REPORT**

**Index**

|  |  |  |
| --- | --- | --- |
| **Sr No.** | **Title** | **Page No.** |
| **1.** | **Project Requirements** | **3** |
| **2.** | **Flow of Solution** | **5** |
| **4.** | **Using Pig Script** | **6** |
| **5.** | **Conclusion** | **24** |

**Project Requirements**

1. Executive Summary

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

1.2 Purpose and Scope of this Specification

The purpose of this project is to capture the data for analyzing the progress of various activities.

In scope

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 Hbase 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

 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.

Click here to download Command: flume-agent agent –n agent1 –c conf –f

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.

4. 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.

**Flow of the Solution(Generalized)**

**FIRST STEP-**

-Download the file from source into the local system.

-Using flume get that file from local system to the HDFS(Hadoop Distributed file system.)

**SECOND STEP-**

-Convert the given XML file to CSV (Comma Separated Values).

This can be done in both ways

* Using Mapreduce i.e creating a custom input format.
* Using Pig i.e.Using the piggybank methods.

In this project I have demonstrated via Pig script of converting the file from XML to CSV.

**THIRD STEP-**

-To get all the districts which have achieved the 100% in BPL cards.

The above can be demonstrated with help of pig.

**FOURTH STEP –**

-To get all the districts which have achieved the 80% in BPL cards.

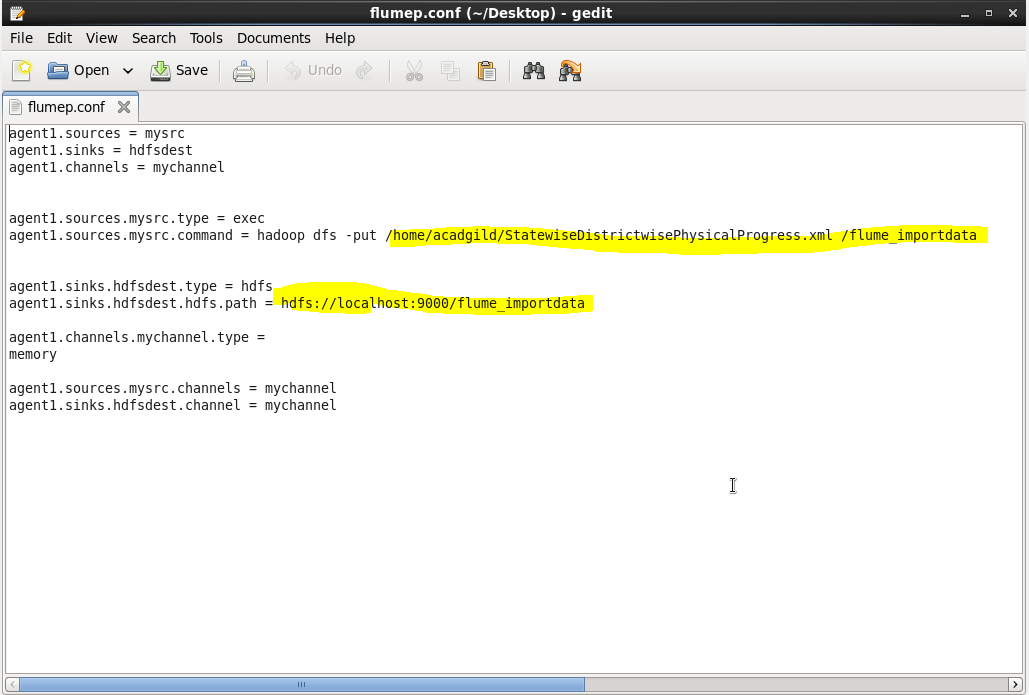
As in the project requirements pig udf has to be implemented hence this is done using the pig implementation.

**SOLUTION**

**Using Pig Scripts**

**FIRST STEP-**

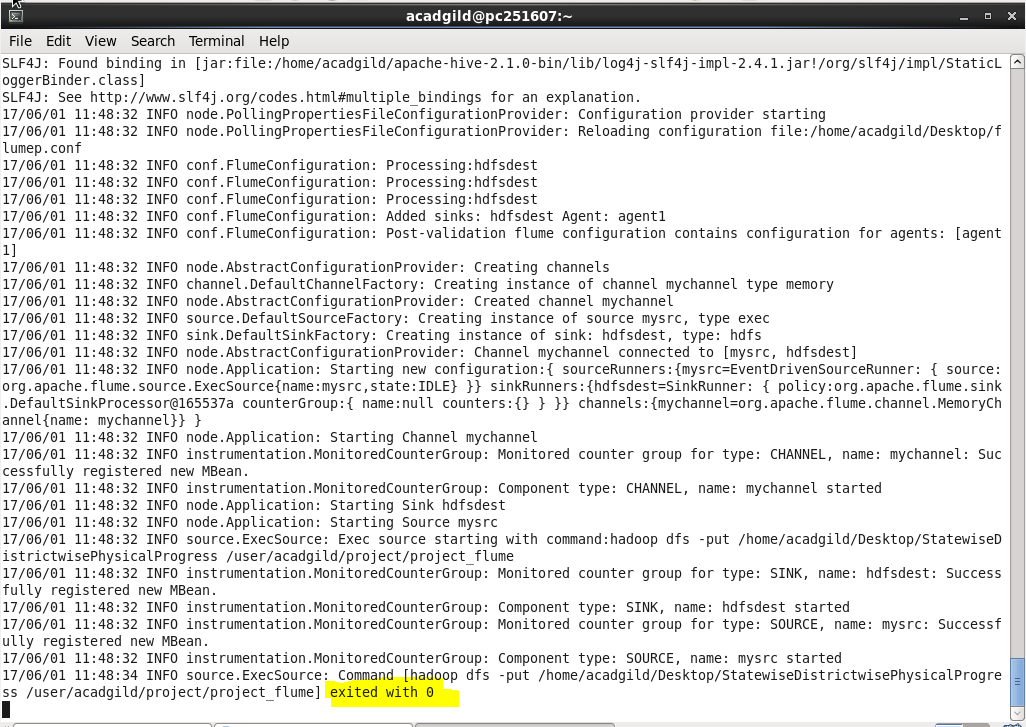
Extracting data using flume to HDFS



Above shows the flume agent that was used to extract the data from local file system to HDFS.

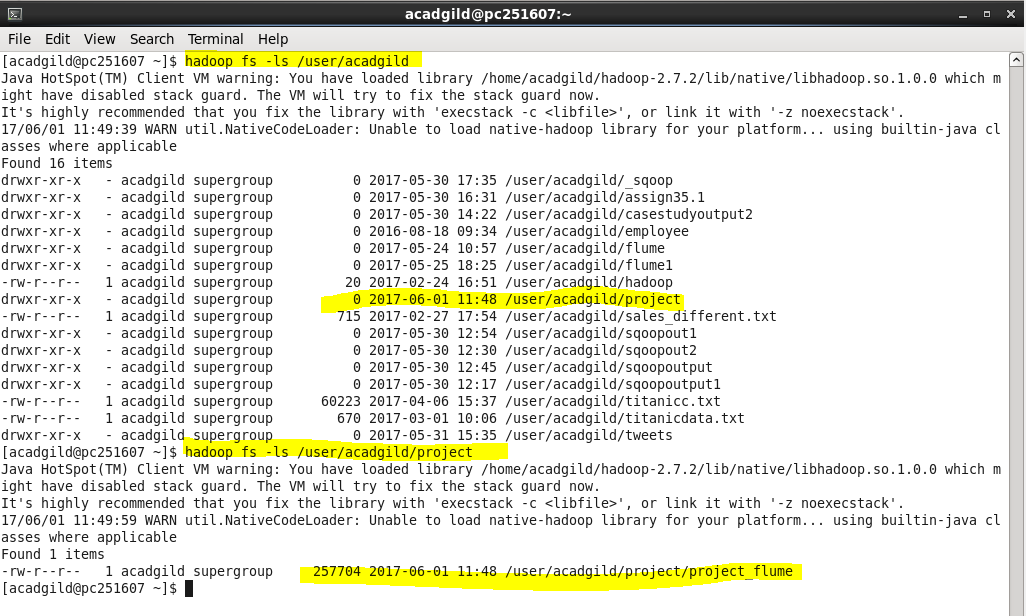


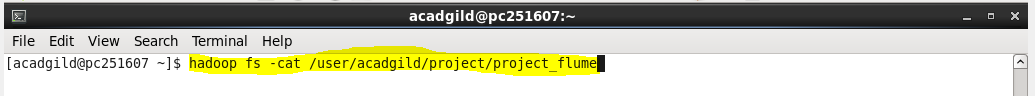
The command used to extract the file to HDFS . Above we have mentioned the location of the configuration file used.



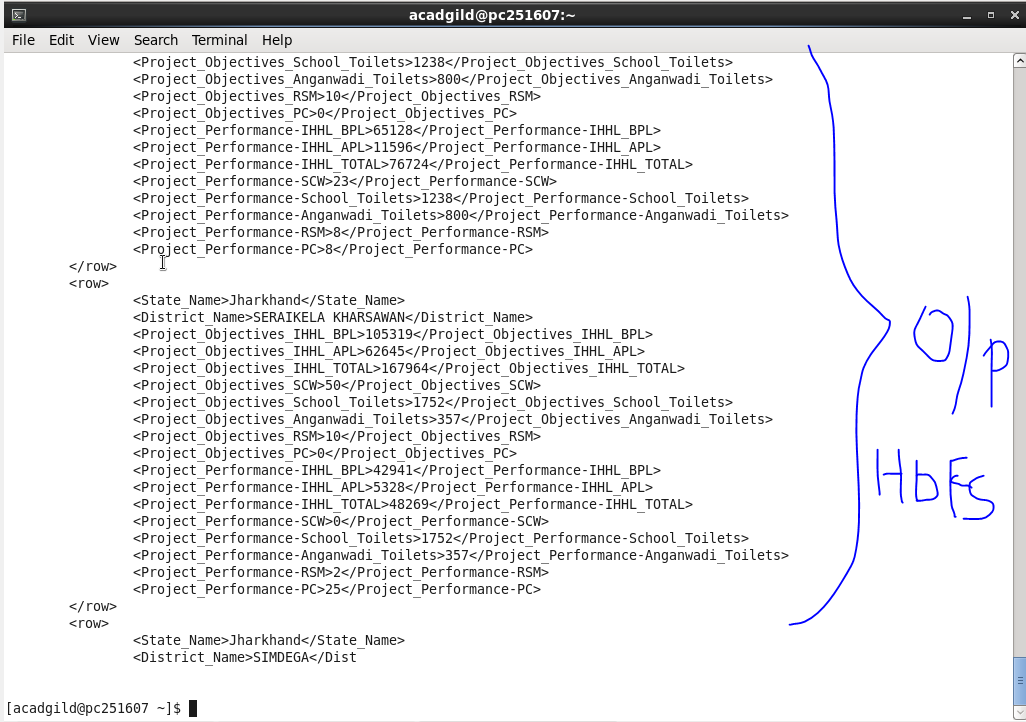
Here we can see that transfer of the file from local files system to HDFS is over. As flume is used for getting the streaming data hence it keeps the channels connected expecting the new input. Hence in order to stop it from listening on the channel Crlt-c has to be pressed.

Displaying data from hadoop





Hence after displaying the data it appears like



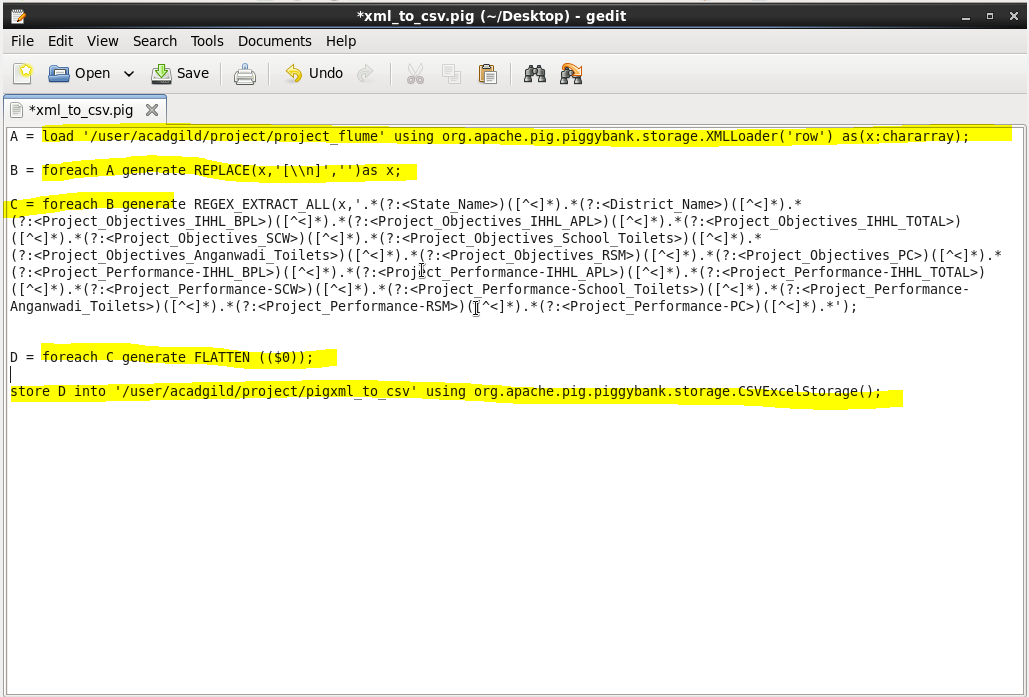
Successful extraction of data was possible.

**STEP TWO-**

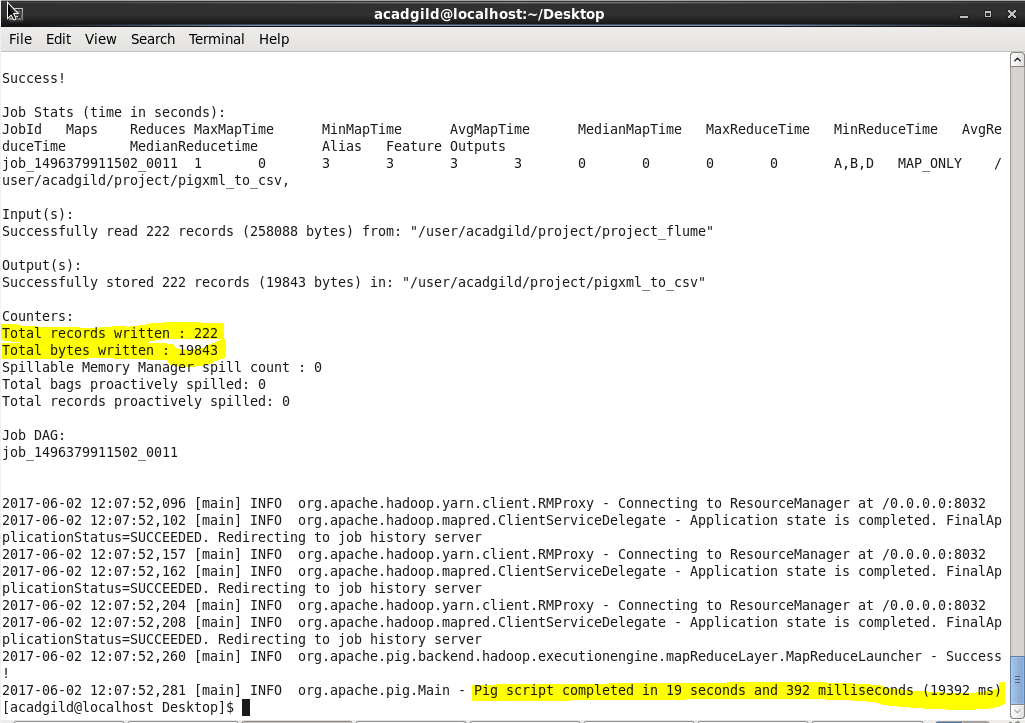
Conversion of the XML to CSV file format.

In this part we are accomplishing the below part by using the **Pig** .

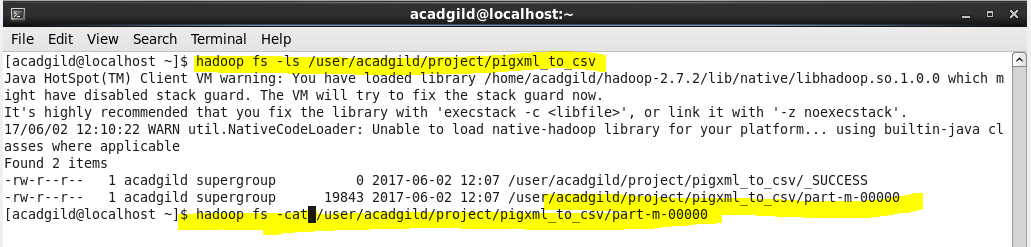
Following is the **pig script** .



Process completed successfully



On running the above script we get the following csv output data on HDFS.

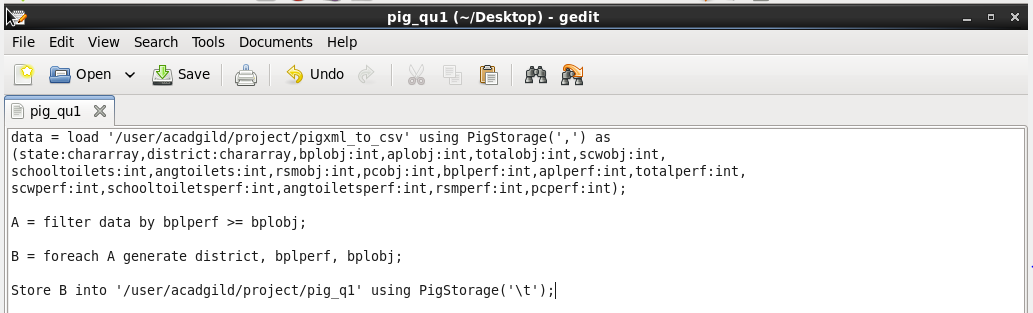


Thus execution of the pig script converts the file from XML to CSV format.

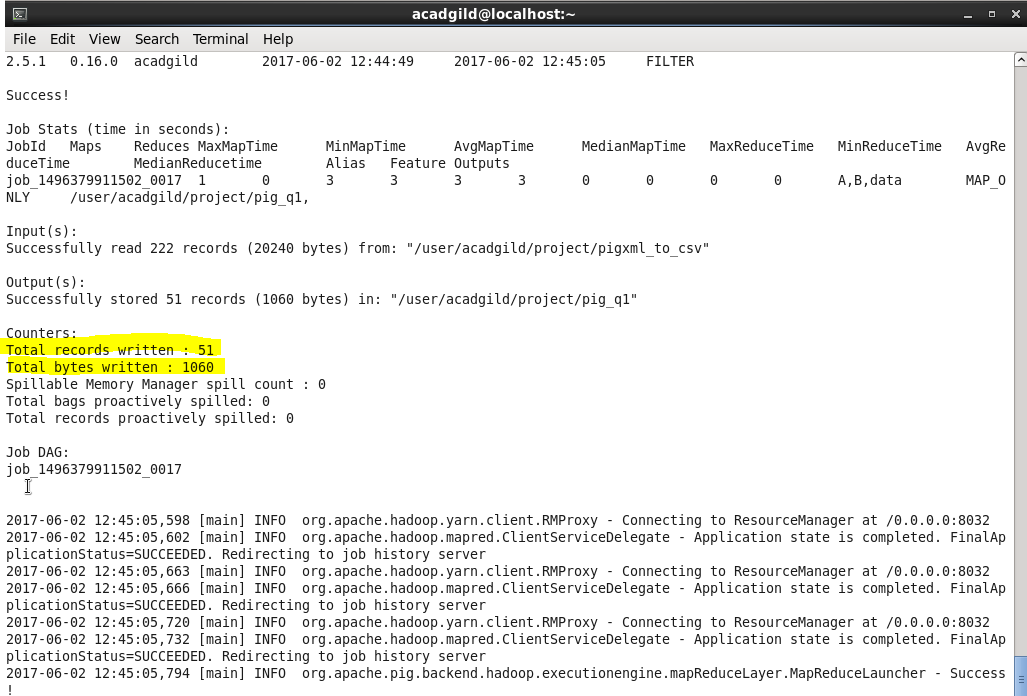
**STEP THREE-**

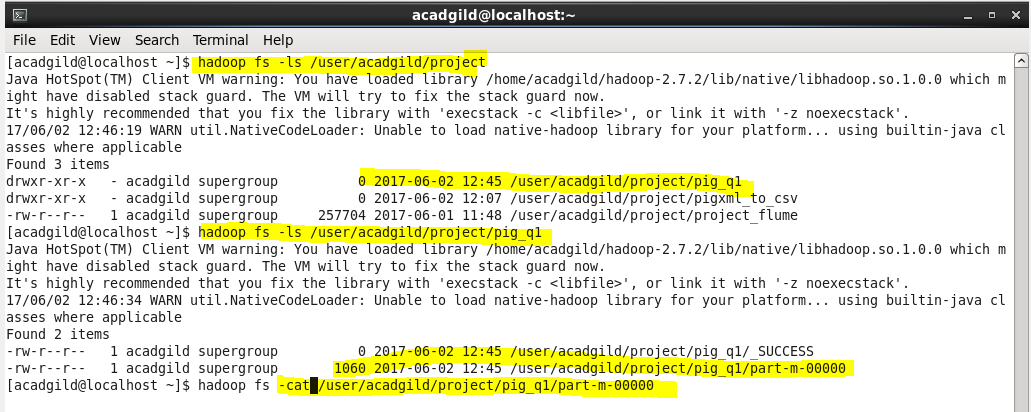
Getting the districts that are achieved the 100% in bpl cards

Again created a Pig Script to achieve the intended output.

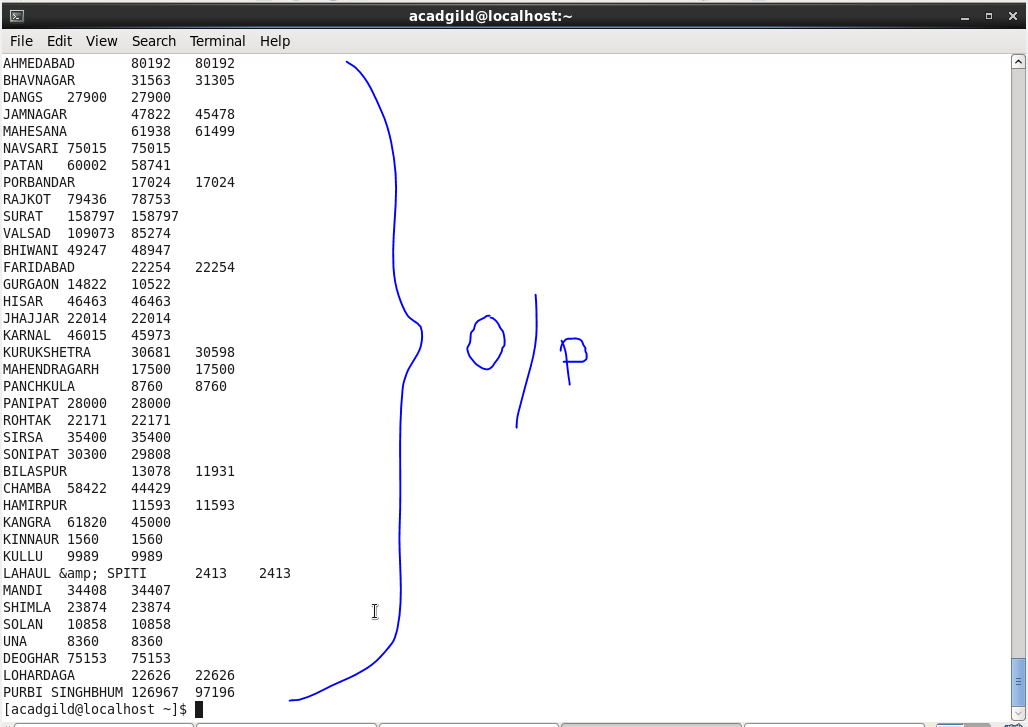


After Execution of the Script we get the following output -

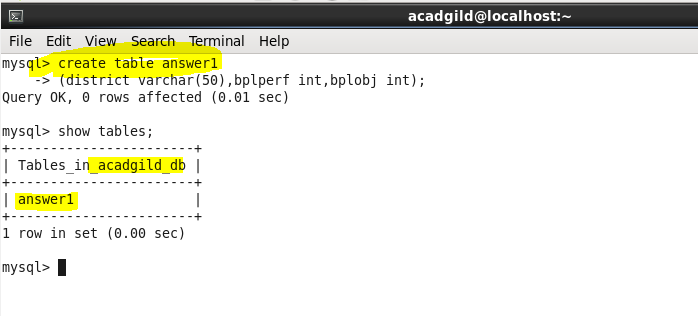




Sample output containing records which have achieved 100% mark or more.

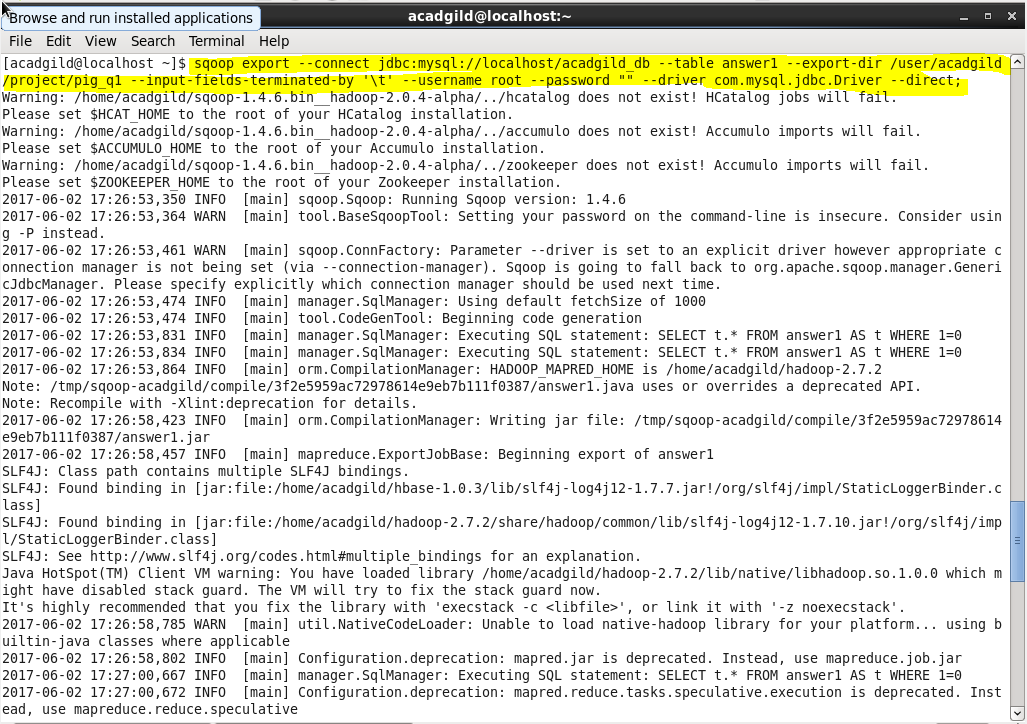


Creating the required empty tables in mysql database for data to be exported.

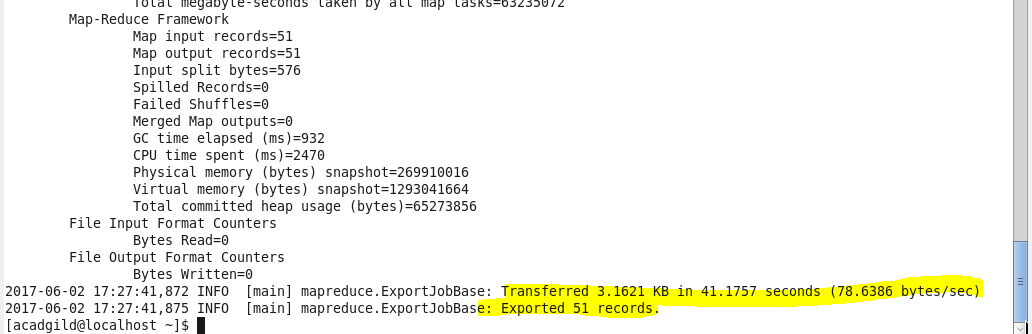


Now importing the output file into mysql db using Sqoop job.

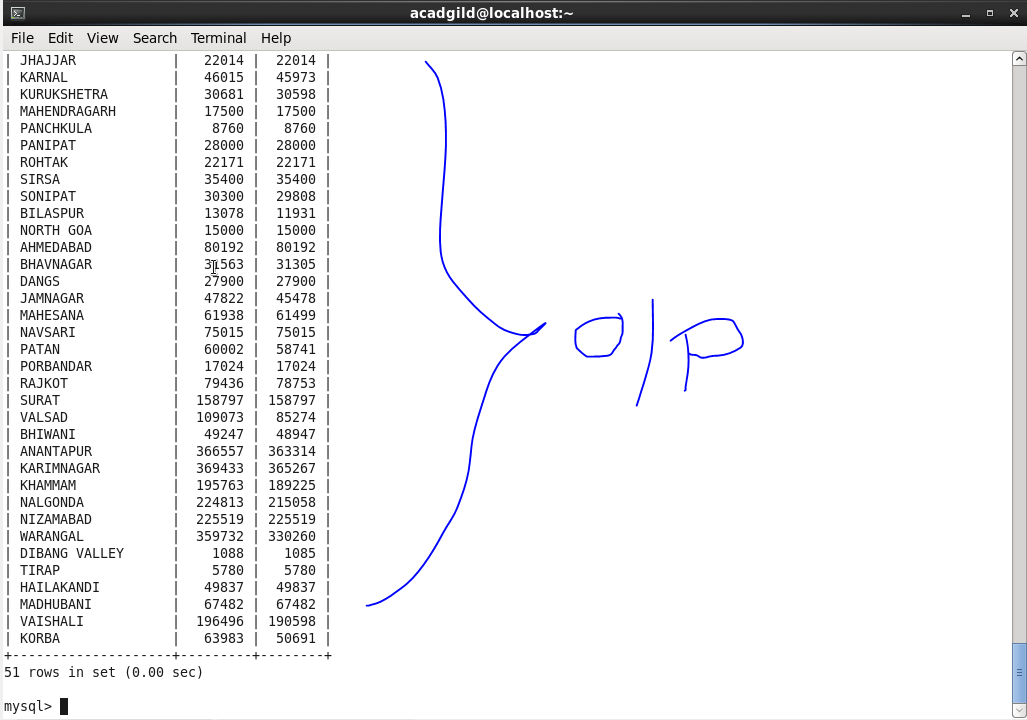
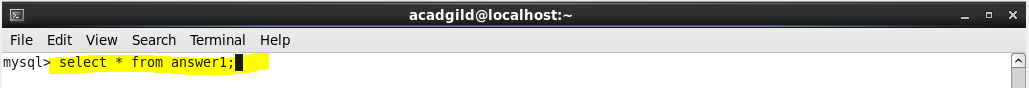
**Sqoop job implementation**



**Completion of the Sqoop job**



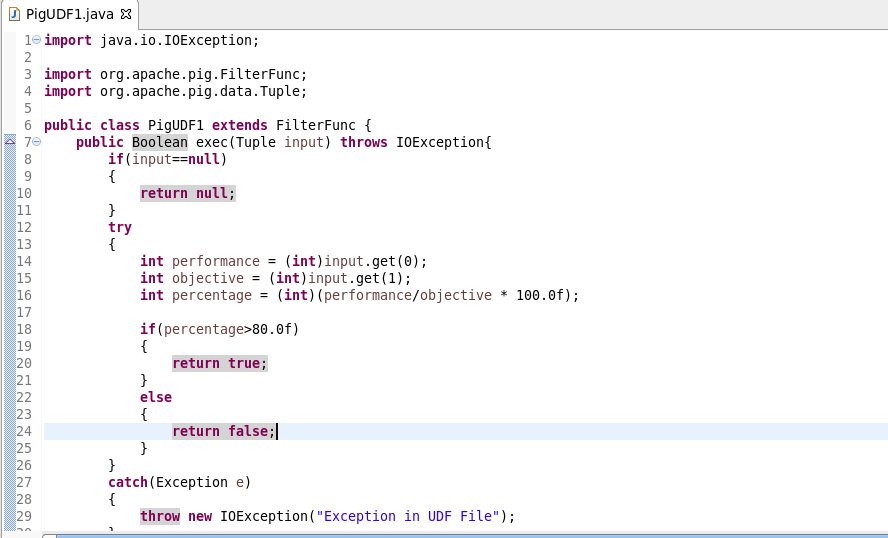
**Sample Output from mysql db.**



From above results we can see that all the percentages are equal or above 100%

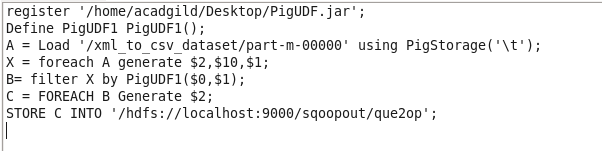
**STEP FOUR-**

Getting the districts that are achieved the 80% in bpl cards. Here as per the instructions **Pig UDF** has been implemented**.**



Filter function of pig is used in order to filter out the all the records who have achieved the 80% mark or above in bpl cards.

The function returns true if and only if the percentage is equal to 80 or above 80 percent, else it returns false.



Here is the **Pig Script** written in order to get the records that have achieved the 80% or more in bpl cards.

-I have first registered the udf jar.

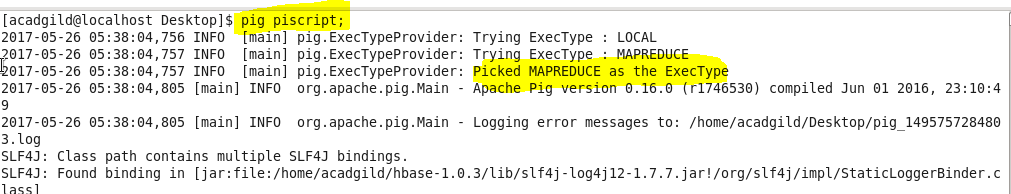
-Then loaded the data from the converted file with the above mentioned schema.

-Used my udf to filter out the records as needed.

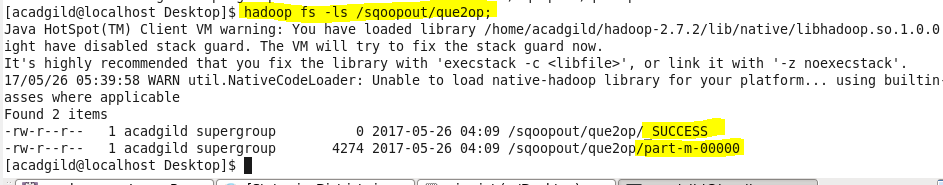
-Then displayed only the state name, accomplished value, target value and percentage value.

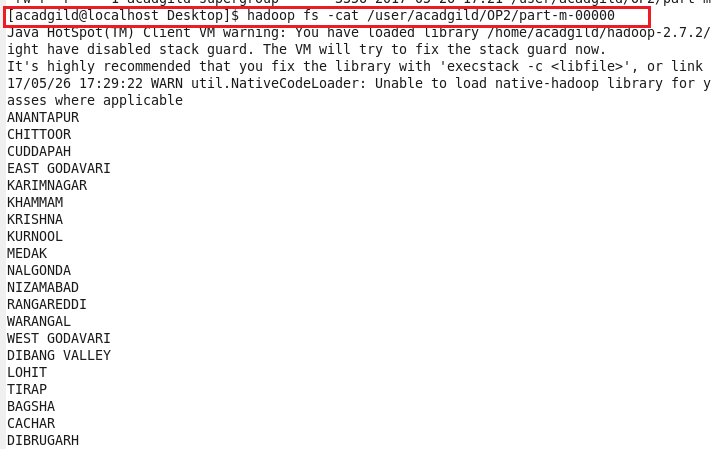
Note – As mentioned in the requirements the task was to complete using the pig udf hence executed with pig Script.

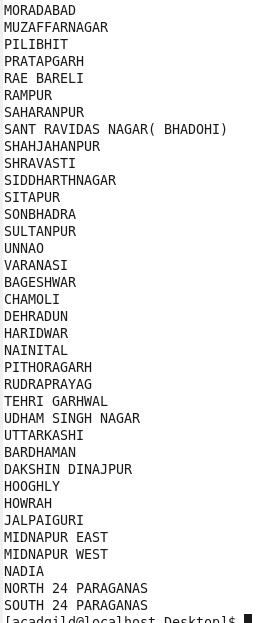
Below are the screen shots of the implementation of the pig script.



This is the final result of the pig script.

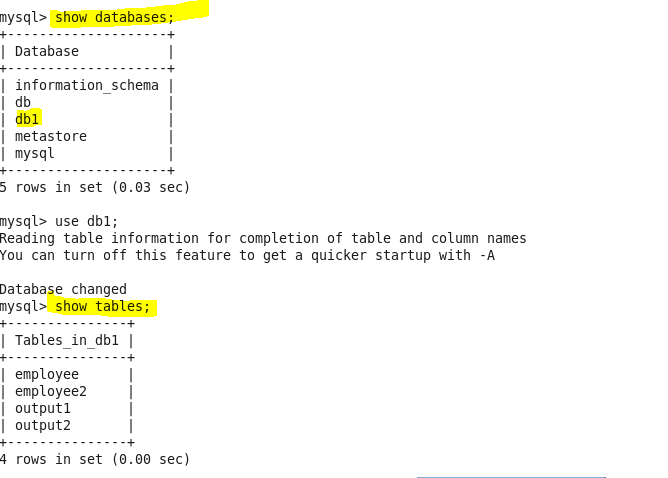


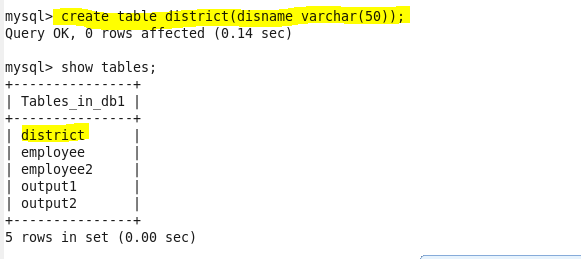




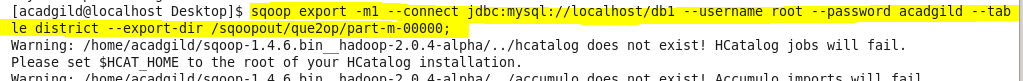
This value is put in mysql db using sqoop .

First created a table in the mysql bd in order to store the values into mysql db.



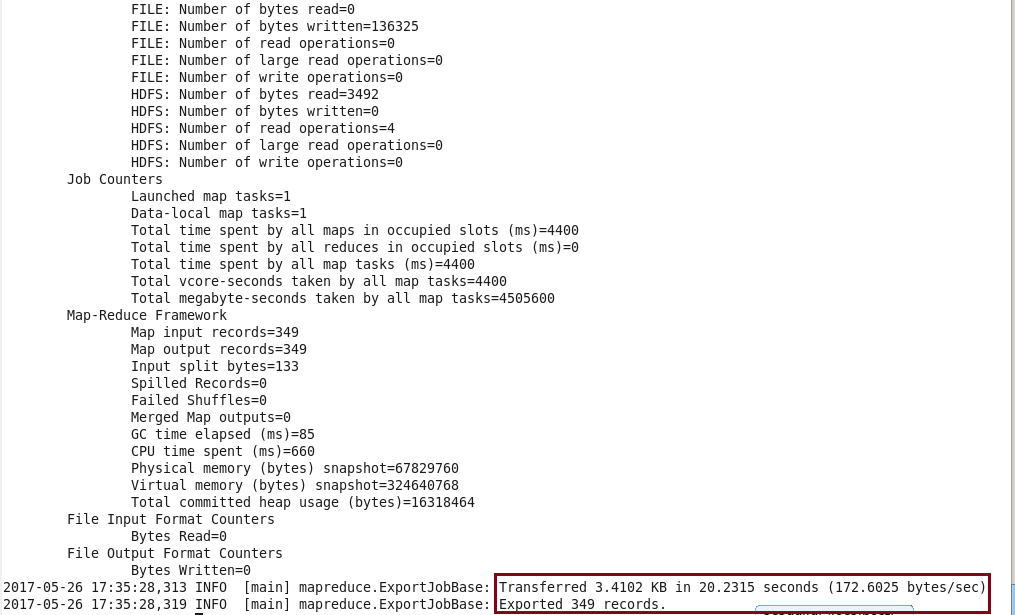


**Using sqoop job**

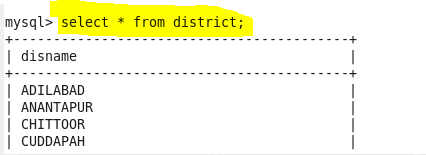


In the sqoop command we have connected to database where we created the database. Also mentioned the table name and the directed from which data is to be extracted.

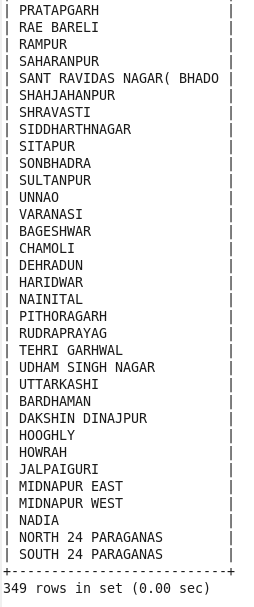
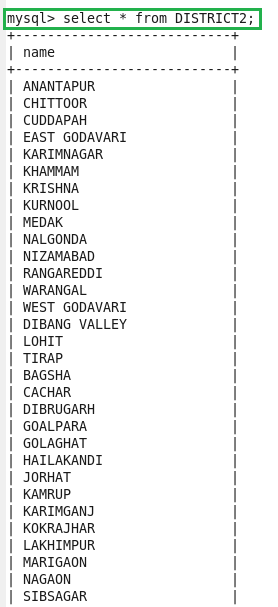
**Completion of the sqoop job.**



**This is the output what we get in the mysqldb after the sqoop job.**



**Below is the full output , copied from the terminal**



**From above results we can see that all the percentages are equal or above 80%.**

**Conclusion**

Thus we have meet the requirements mentioned above also with it learnt the implementation of the

1. Parsing XML File using map reduce program
2. Parsing XML file using Pig Script.
3. Ingestion of the datafrom hdfs into MySql Database using Sqoop.