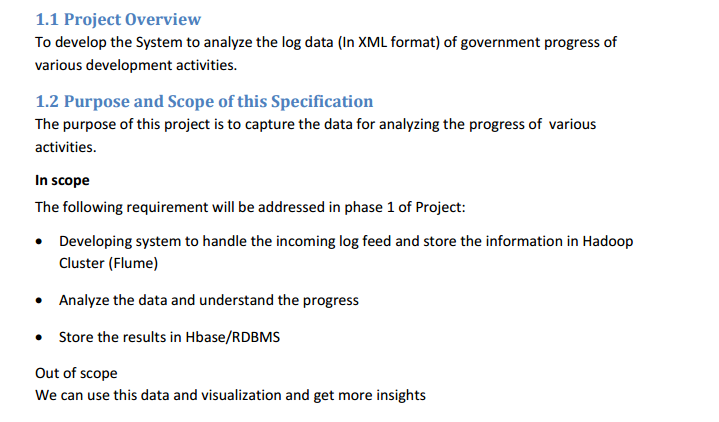
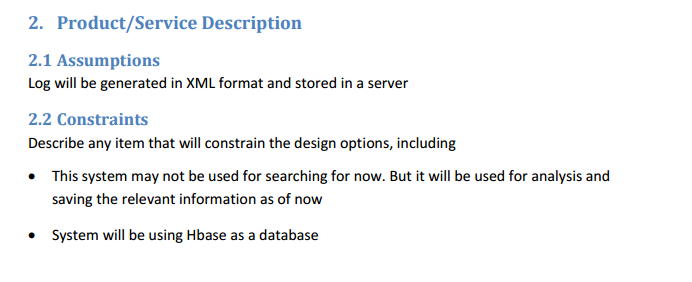
**ACADGILD**

**MAJOR PROJECT REPORT**

**STATE-WISE DEVELOPMENT ANALYSIS IN INDIA**

**BY- AZRA PARVEEN**

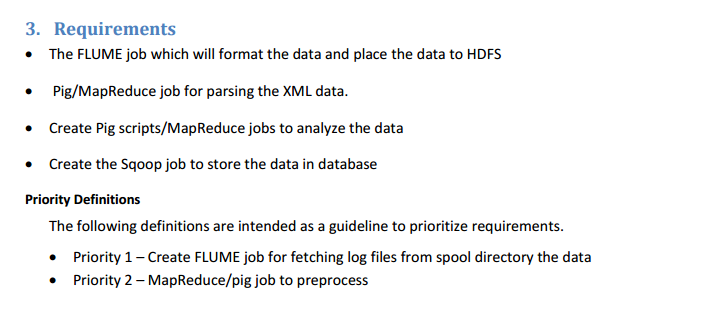




**Now we will download the content from the following link,**



The downloaded content will be in local system, then we will transfer the dataset into HDFS to perform data analytics to get a valuable information over the dataset



Now we have two options to perform analysis

* **mapreduce**
* **PIG**

So what should we choose -

**- Mapreduce:** Mapreduce can handle extreme level of complexity, but considering the coding complexity and further use of the obtained result I prefer pig latin over map reduce

**- Piglatin:** In Piglatin the commands are easy to execute and in case if the obtained results need a join with other dataset ,then there would be a complexity in mapreduce.So considering the future performance, I chose piglatin.

**Flow of the Solution**

**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.**(We will use this)**

In this project I have demonstrated both ways of converting the file from XML to CSV.

**THIRD STEP-**

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

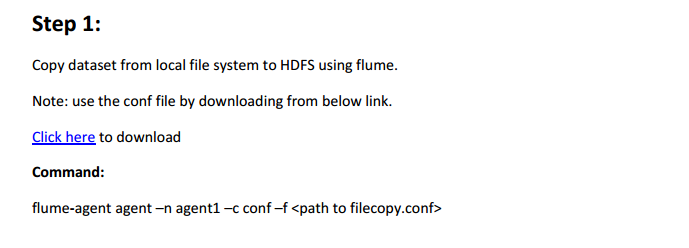
The above can be demonstrated with mapreduce as well as with help of pig. Pig method has been implemented below.

**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 only using the pig implementation.

**Now we will download the flume configuration file**

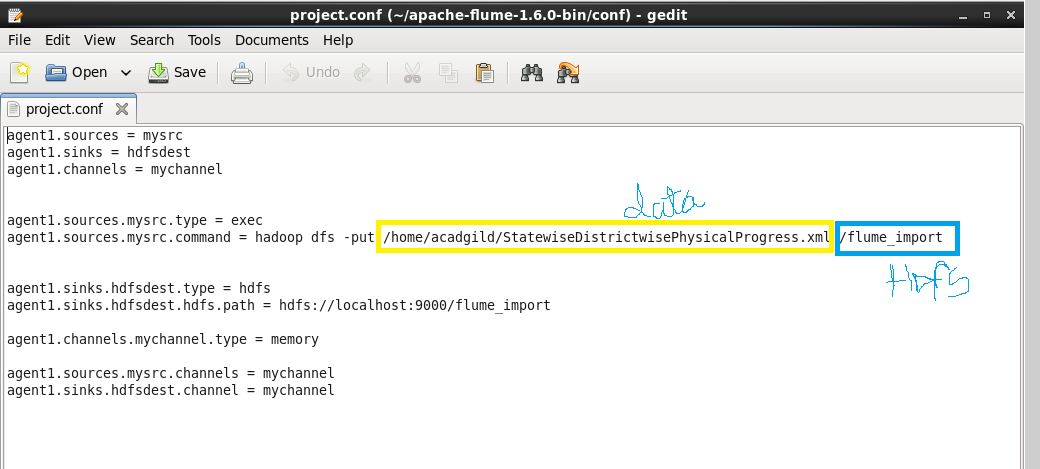


The flume configuration is downloaded and placed in following folder

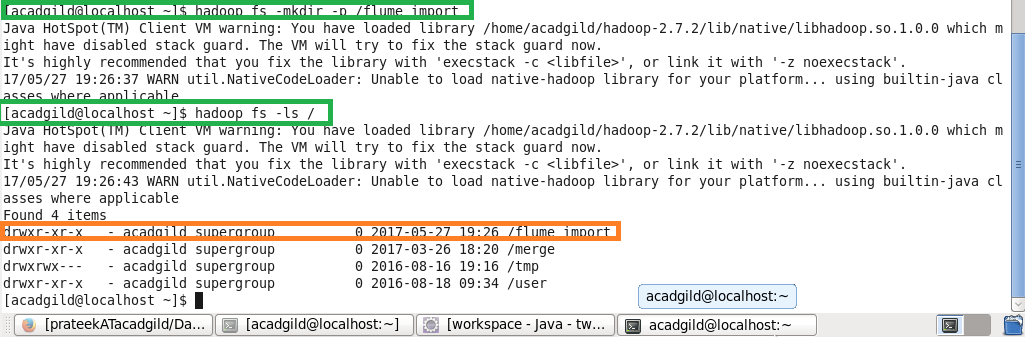
**/home/acadgild/apache-flume-1.6.0-bin/conf**

Here this configuration file gives us the sources ,sinks and channels command to be performed and where the content to be stored

Here our input data xml file is stored in local folder **/home/acadgild/StatewiseDistrictwisePhysicalProgress.xml** which is to be transferred to the **hdfs** folder **/flume\_impot**

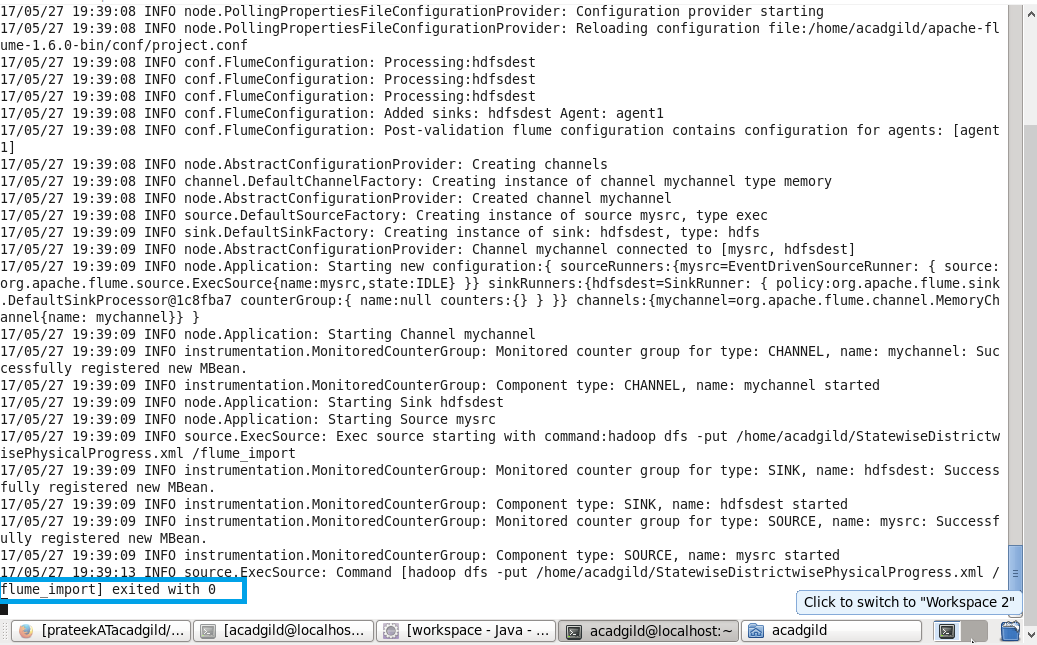
****

**First, creating a directory named as flume\_import. In the below screenshot we can see that directory has been created.**

****

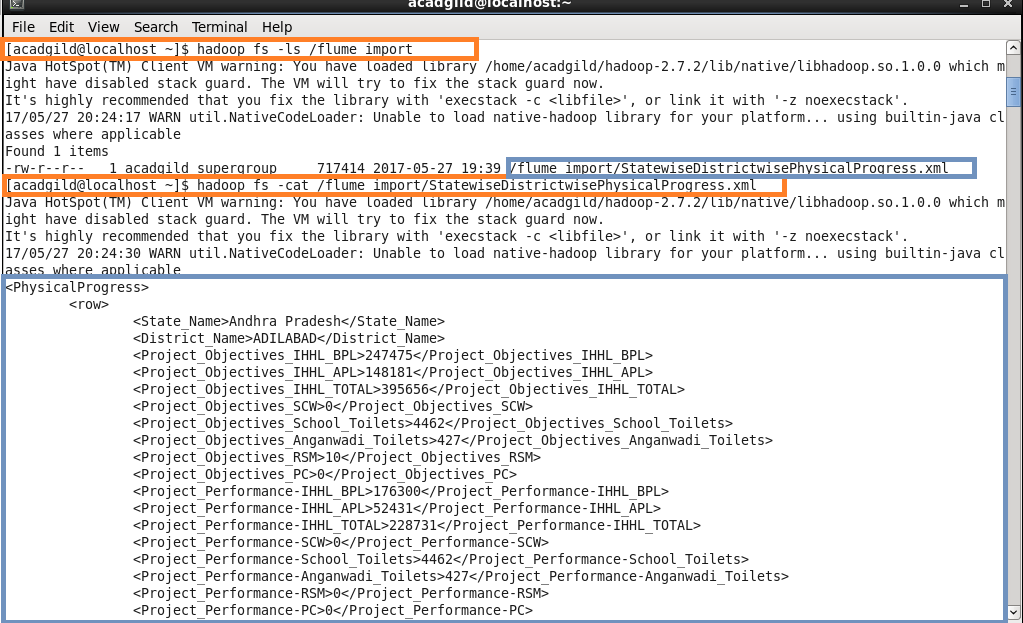
****

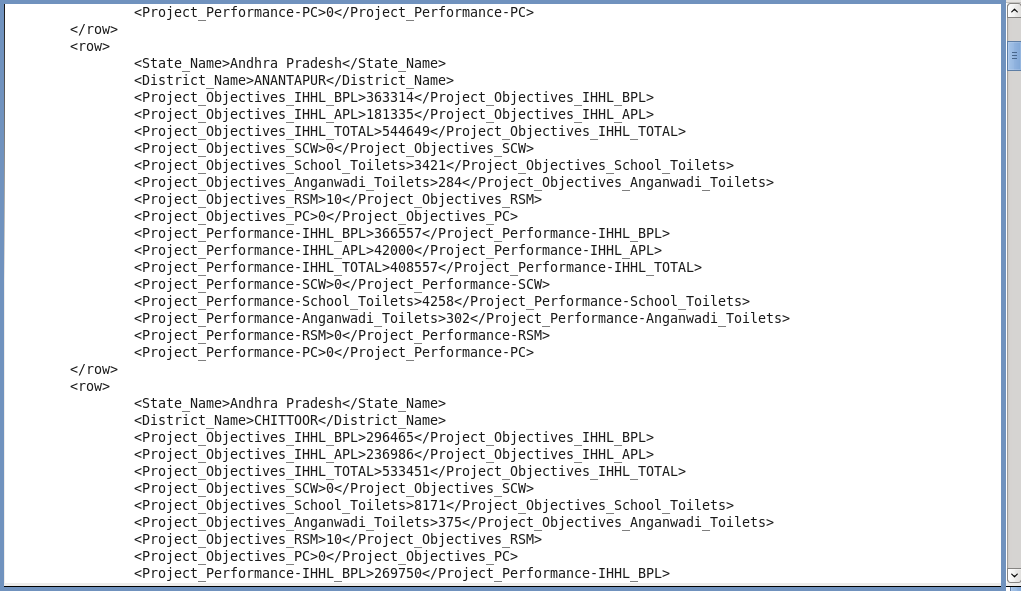
The command used to extract the file to HDFS . Above we have mentioned the location of the configuration file used.



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.

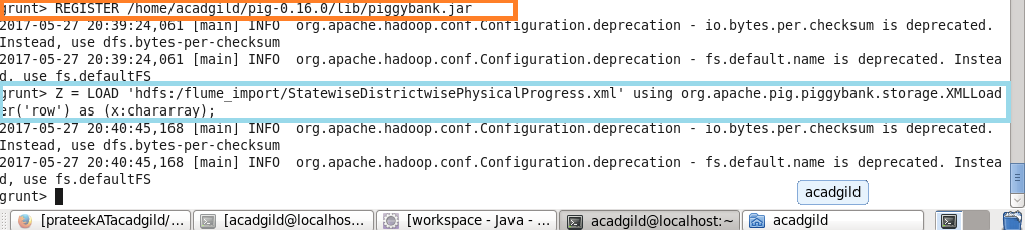
**Listing the created directory and performing basic cat operation.**



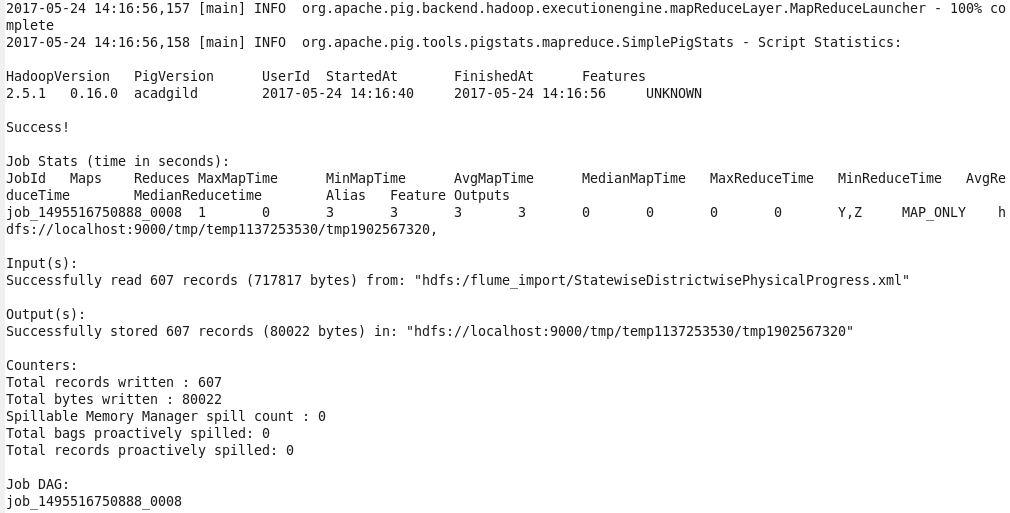
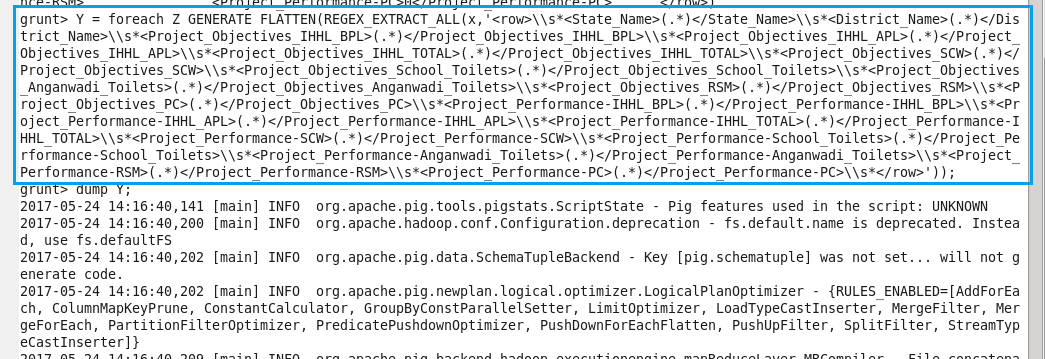


Here data is loaded from /flume project location which contains a XML file and in order to parse the xml file so that Pig can understand it.

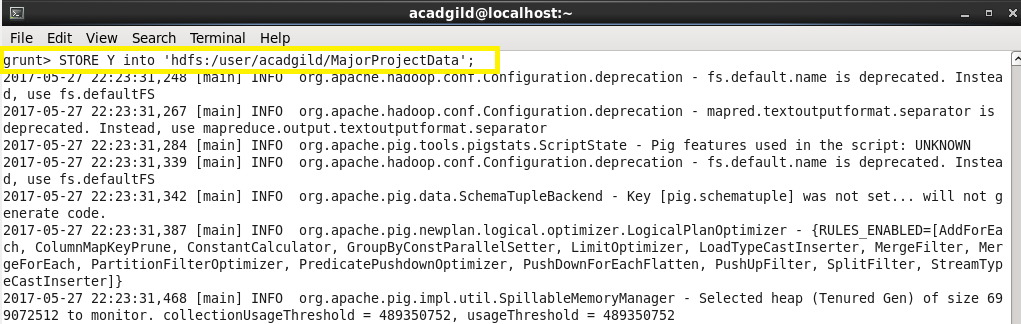
**XMLLoader() function which is present in PiggyBank of Apache is used.**

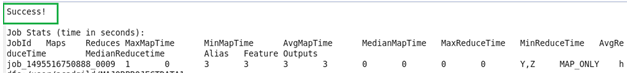


* **Using regex-function for capturing the values within tag and FLATTEN is used for flattening the TUPLE.**

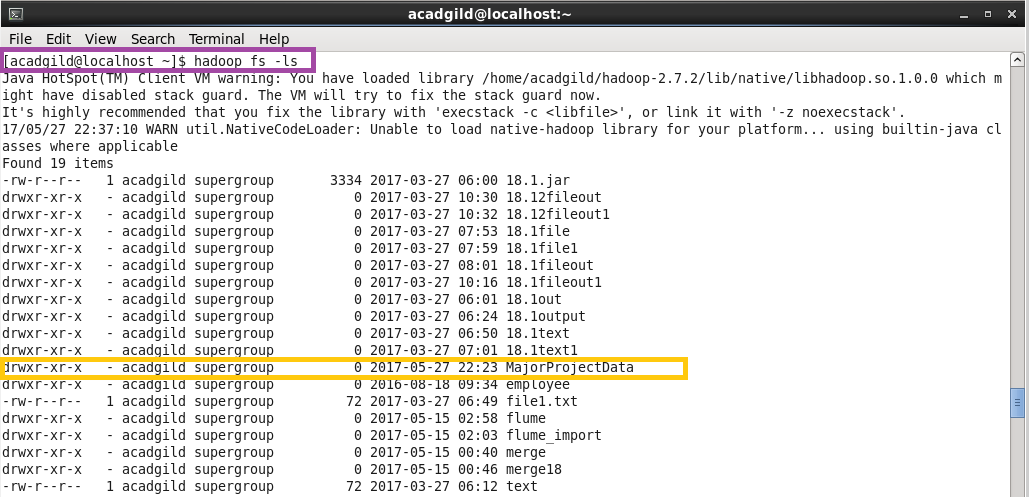


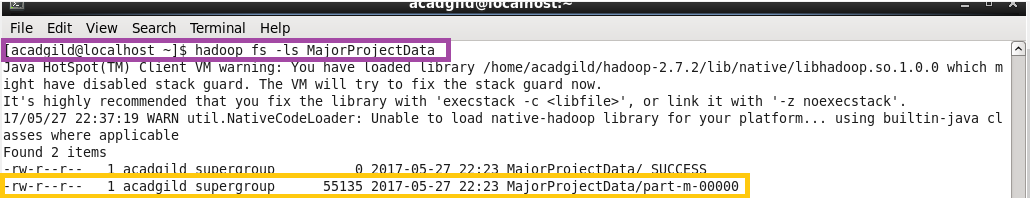
**Storing the above results into HDFS for further analysis.**



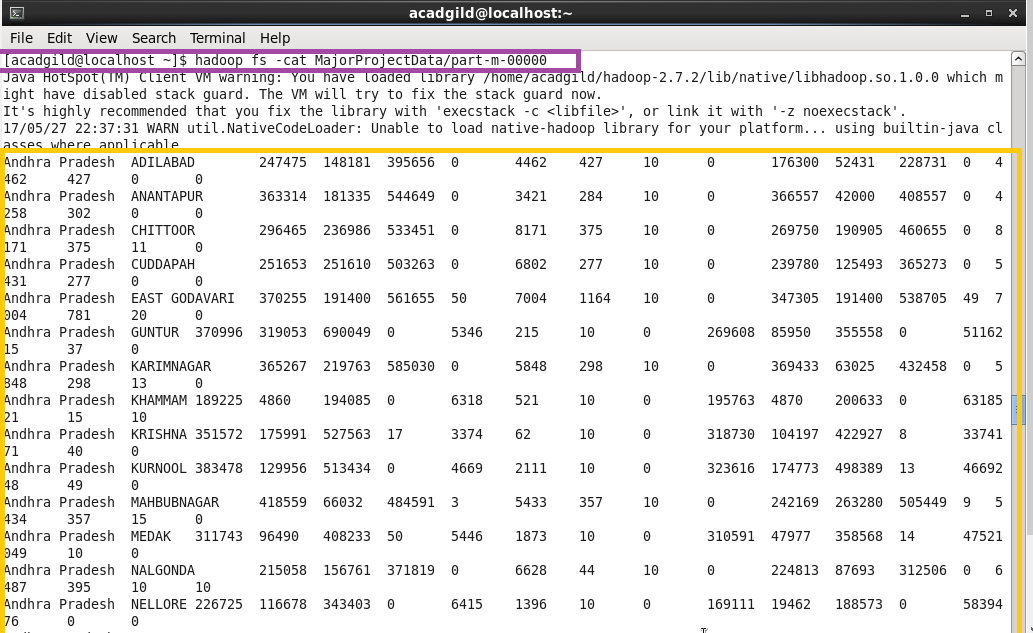


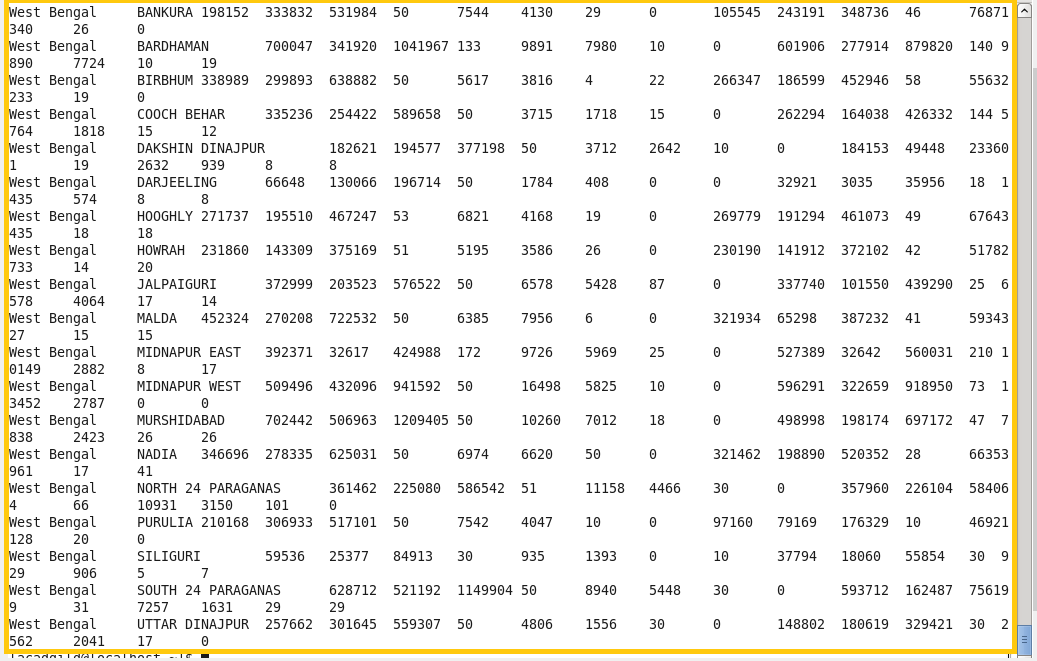
Here, we can see that data is successfully stored.

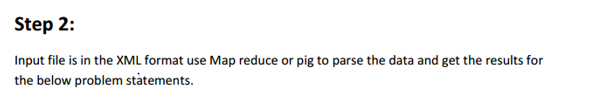




Using cat command we can see the stored data.







* **Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop .**

**Step1:**  loading the dataset

**Step2:**  filtering the loaded data so as to get the districts who achieved 100 percent objective in BPL cards.

$10 is the project objective in BPL cards column

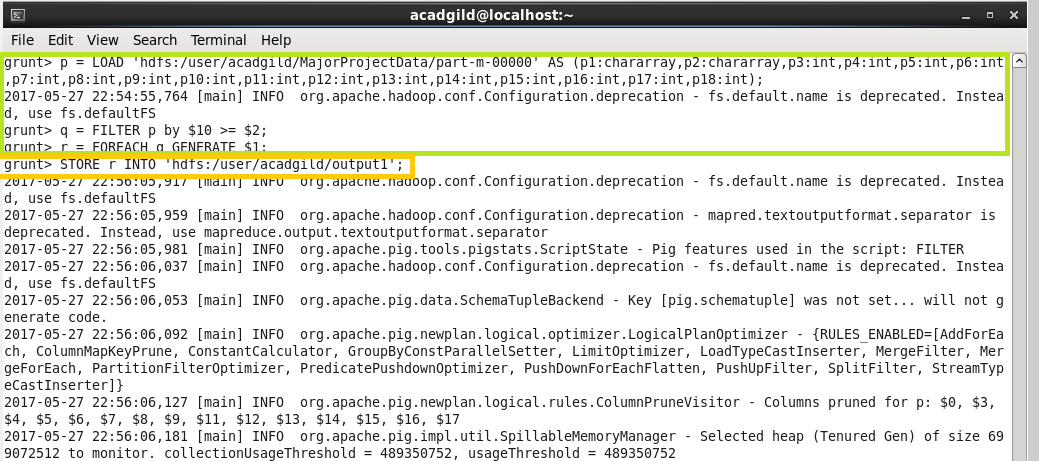
$2 is the project performance in BPL column

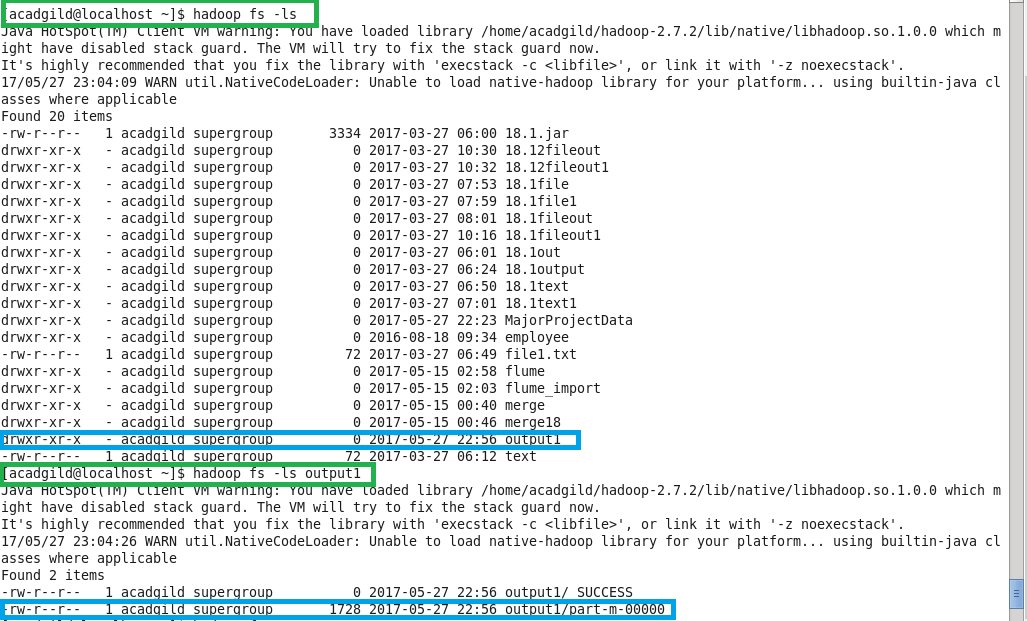
Comparing whether $10 i.e, project objectives in BPL have achieved 100 percent or not.

**Step3:** If achieved then generating $1 which is the district column

**Step4:** finally storing the results as output1.

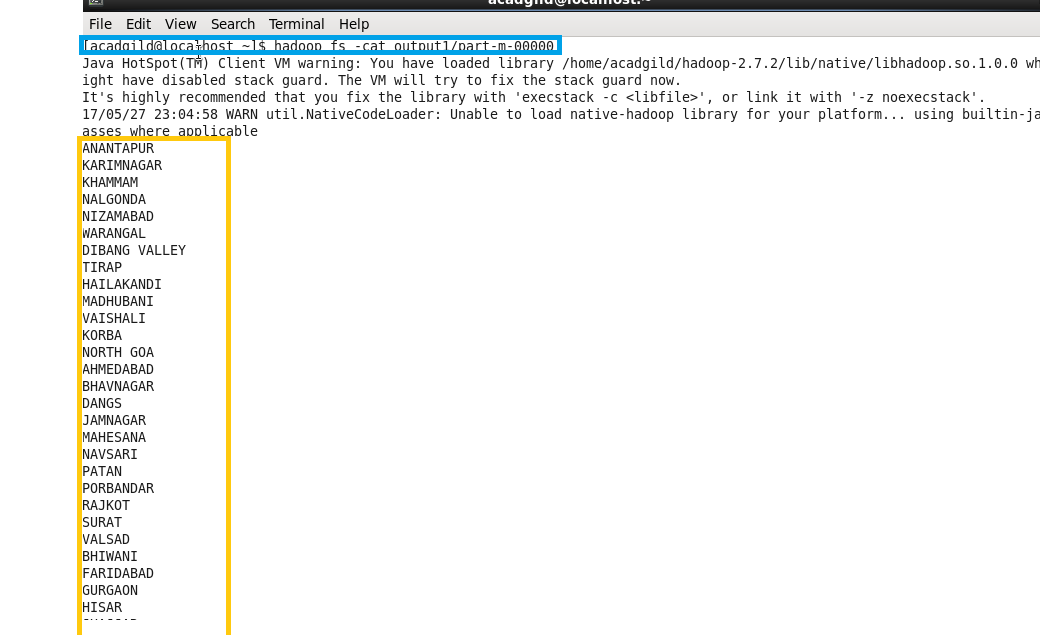
Running pigscript :



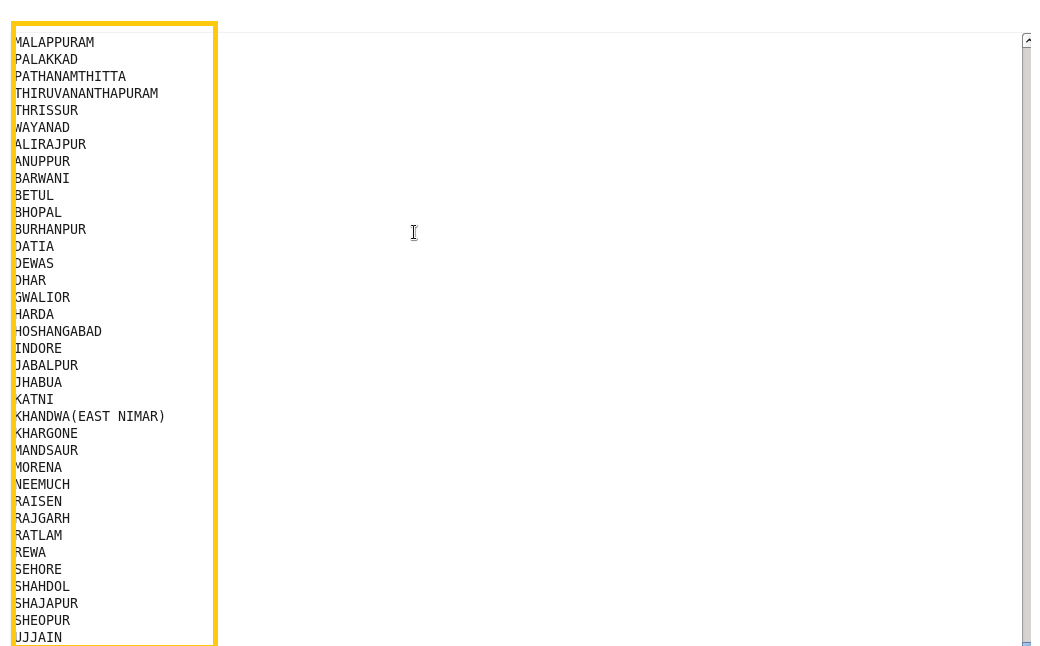


Here, we can see that job ran successfully.

**Now, using cat command we can display the results.**

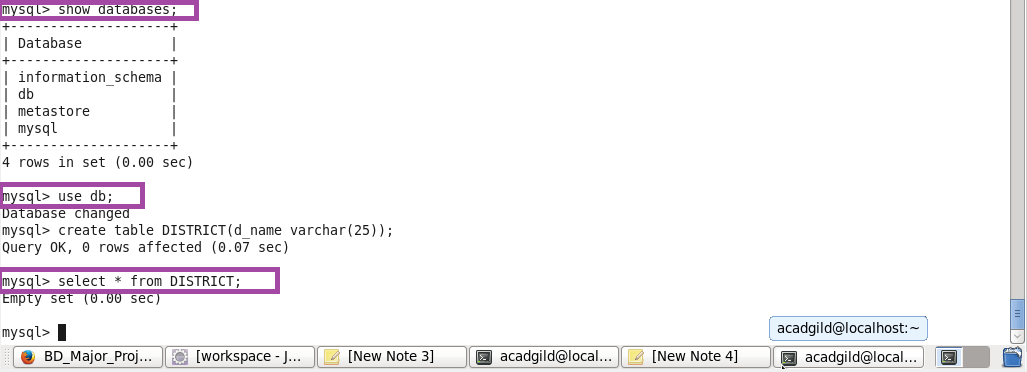
****



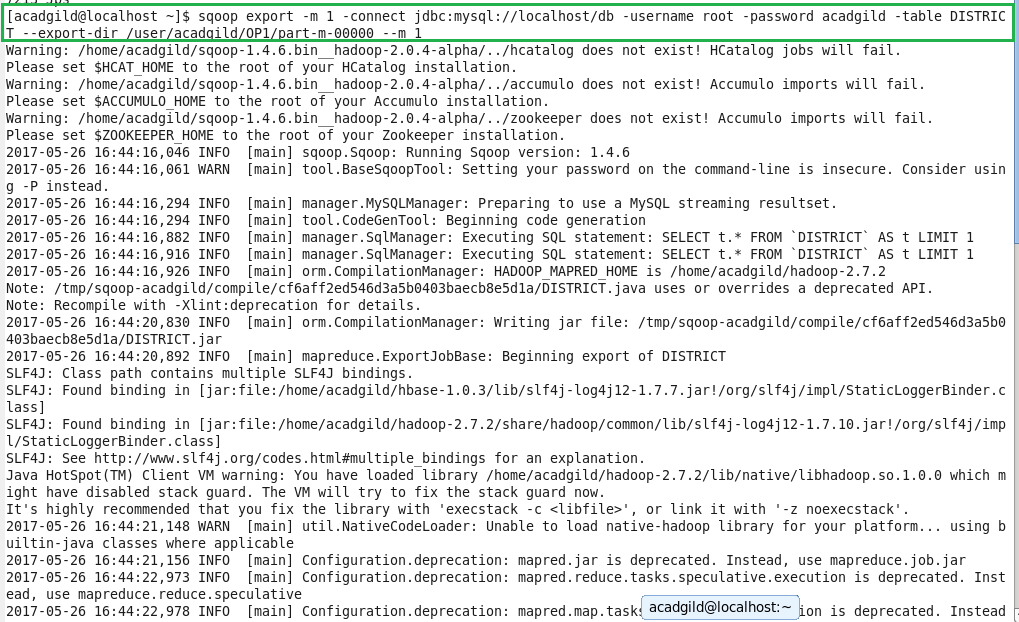


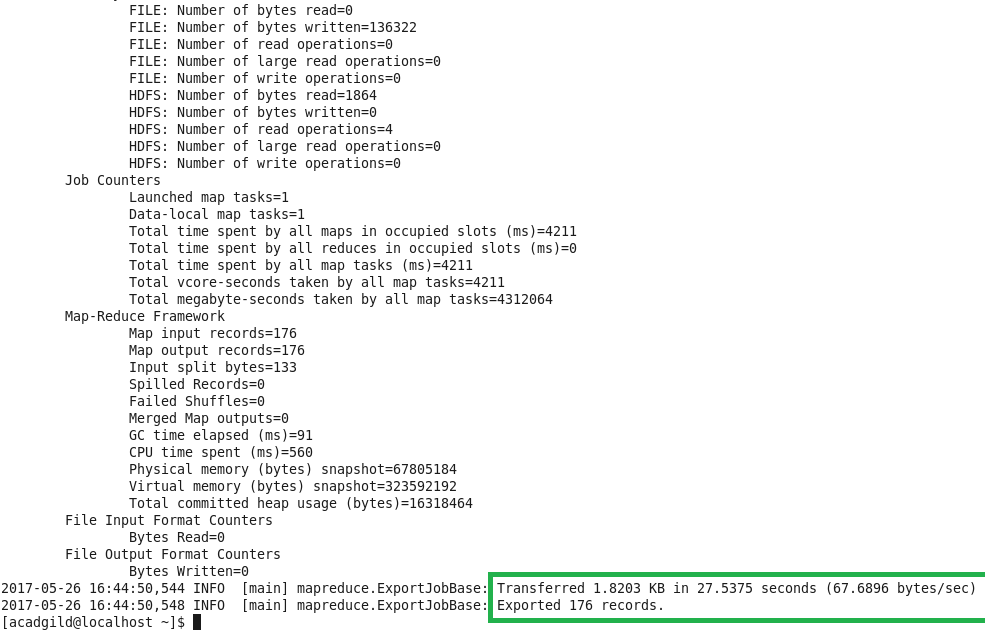


To export the result from pig using sqoop creating a table in mysql and inserting no values in the table so as to perform the export.



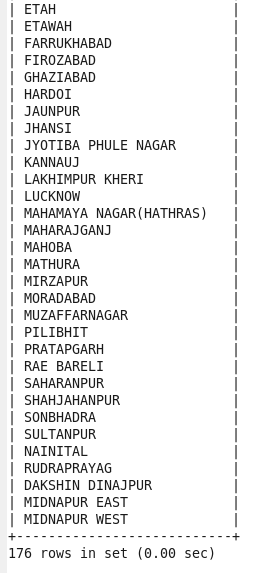
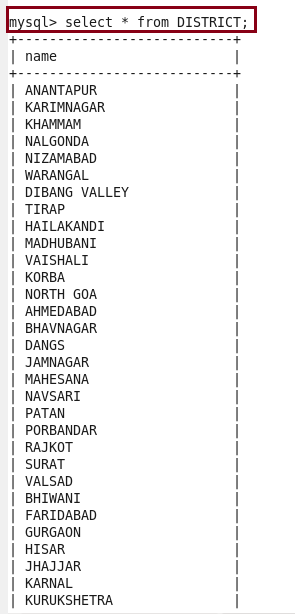
Below is the export command to export the records from pig to mysql





**Here, we can see that the data is successfully exported.**

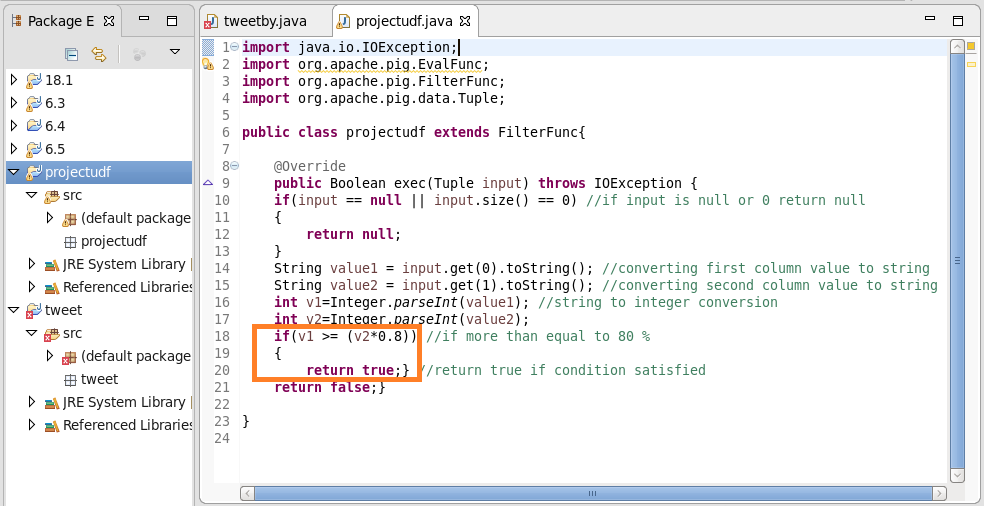
Now, if we perform the select operation on the empty created table it will show the exported data.



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

**Creating UDF:**

* Adding the necessary Imports



* Firstly, **registering the jar**
* Then, loading the data and the datatype and name of fields are defined.
* **Filtering Districts that acheived 80%:**

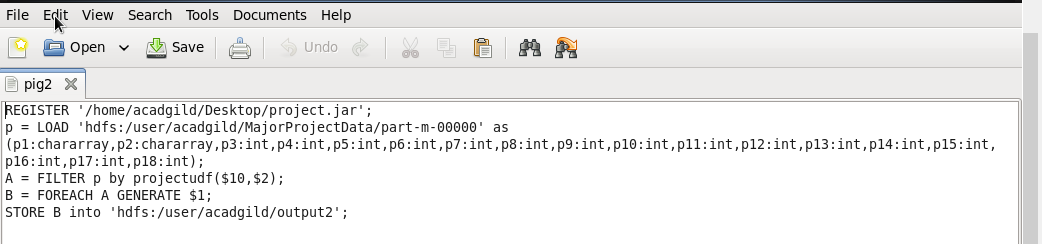
Used a udf function “projectudf” which filter the districts which have acheived less than 80%

**Generating District List:**

$1 is the district column. Here we are generating $1 so as to get the districts name which have reached 80% of objectives of BPL cards.

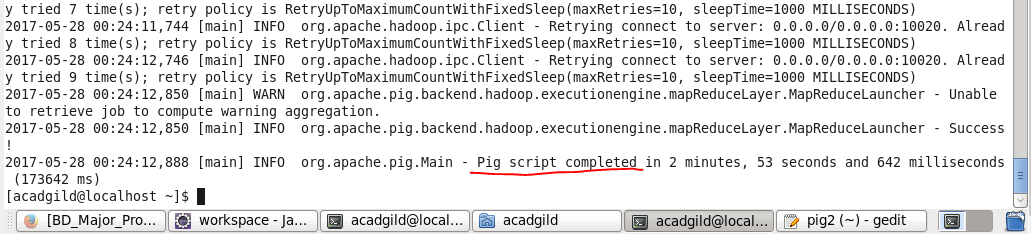
**Finally, Storing the result**

Data is stored into “output2” in Hdfs from which it will be exported to mysql.

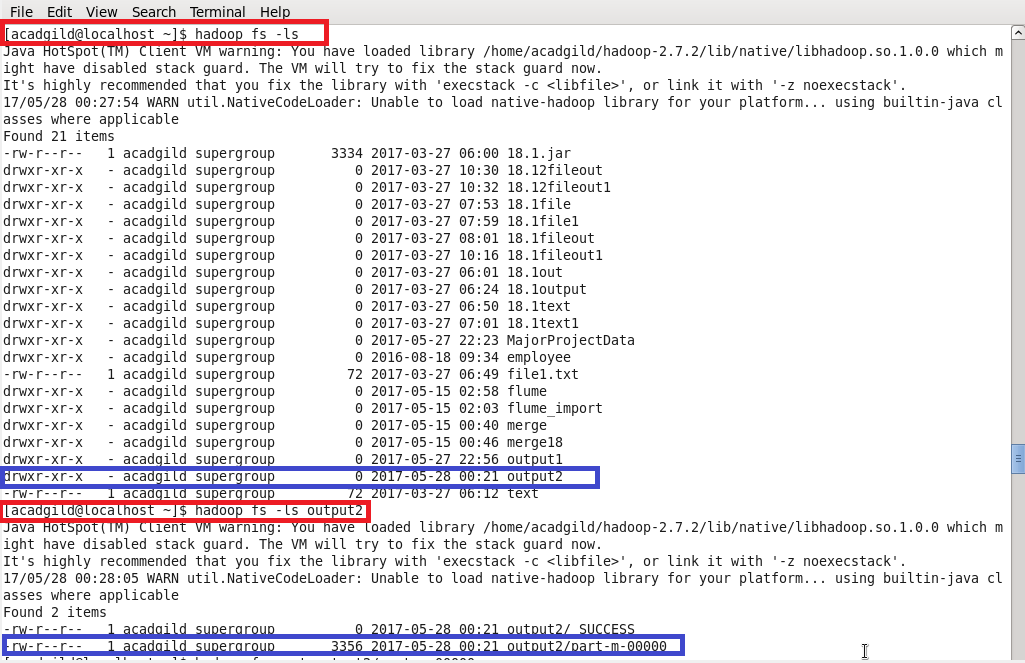


Running above pigscript:

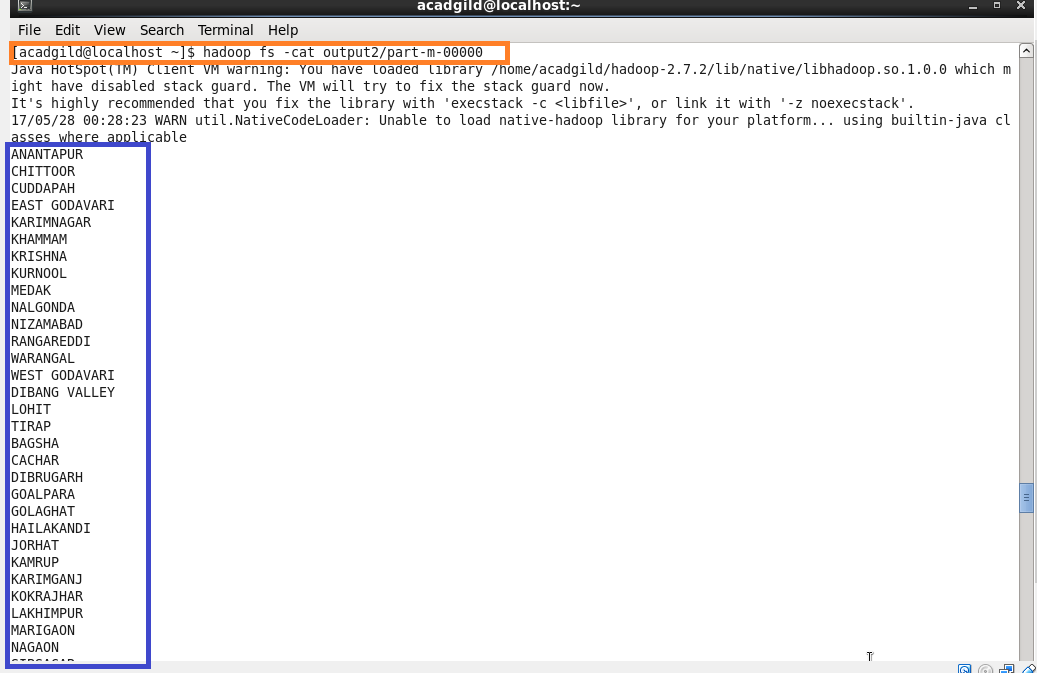


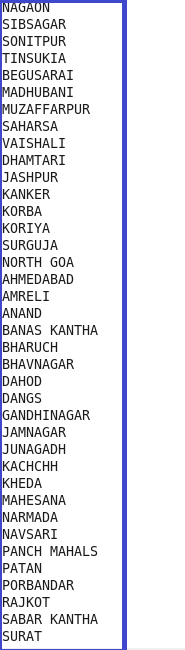


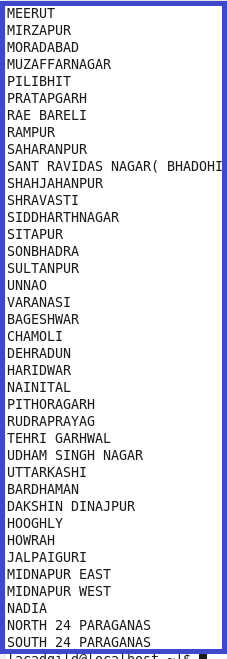
**Listing it and then performing cat operation on it.**



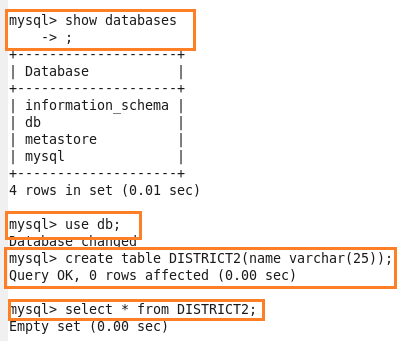
**Displaying the stored output of the pig commands using cat**



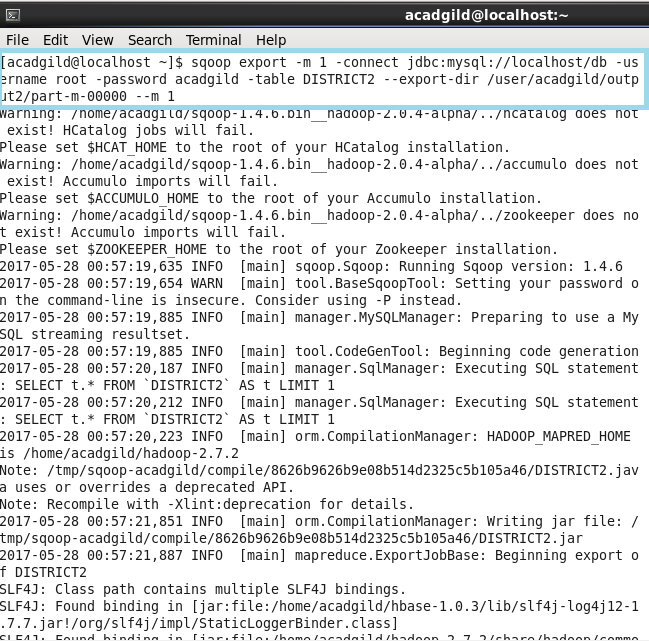


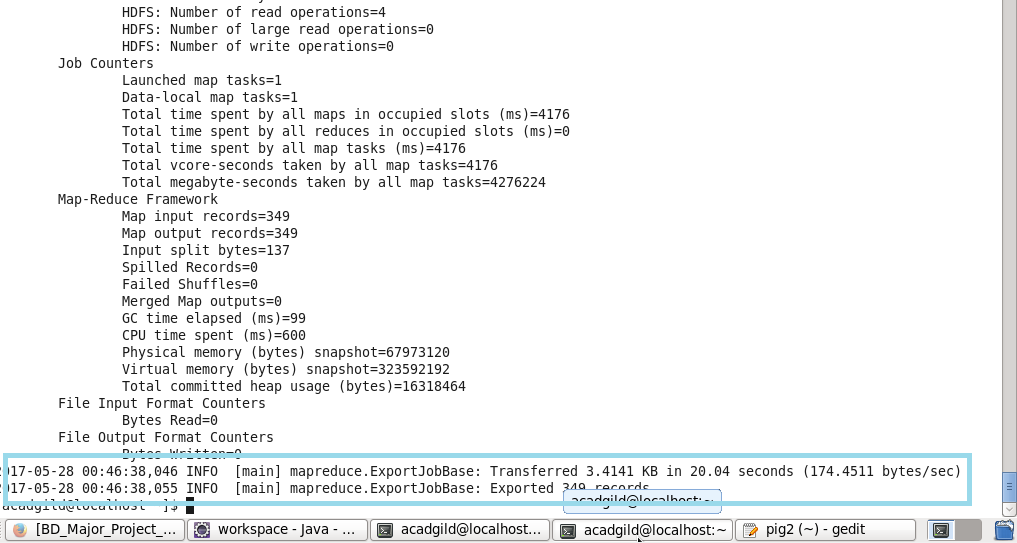


**To export the result from pig to mysql using sqoop creating a table in mysql and inserting no values in the table so as to perform the export.**



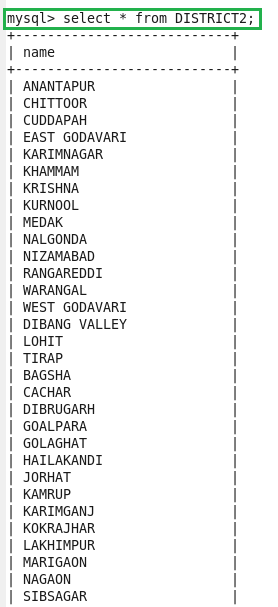
**Below is the export command to perform export operation**

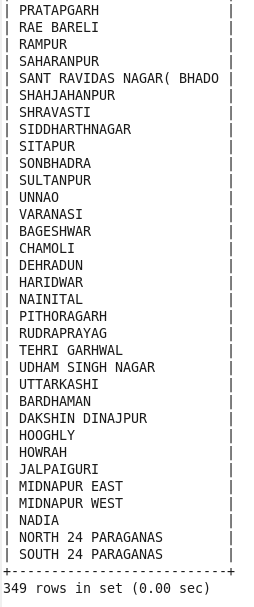




**Here, we can see that successfully exported the records.**

* **Using select command displaying the results in mysql.**





* Full output is in attached file output 2