**Major Project**

**State-Wise Development Analysis in India**

**Problem statement:**

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

**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

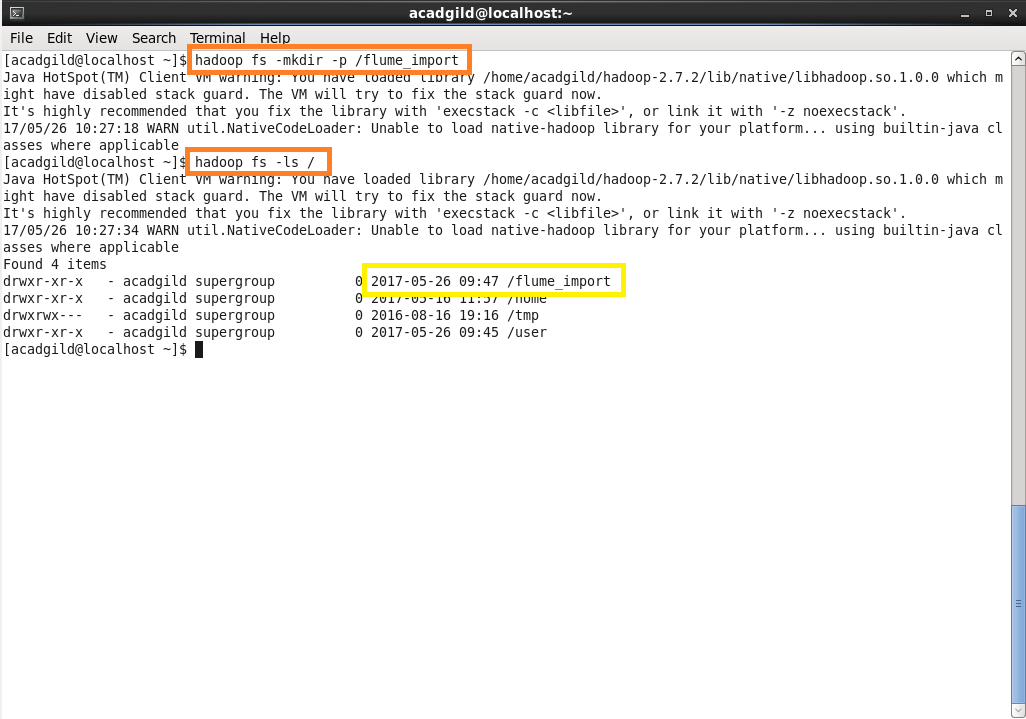
**Part 1:** Steps to be done to make the project run:

**Step 1:**

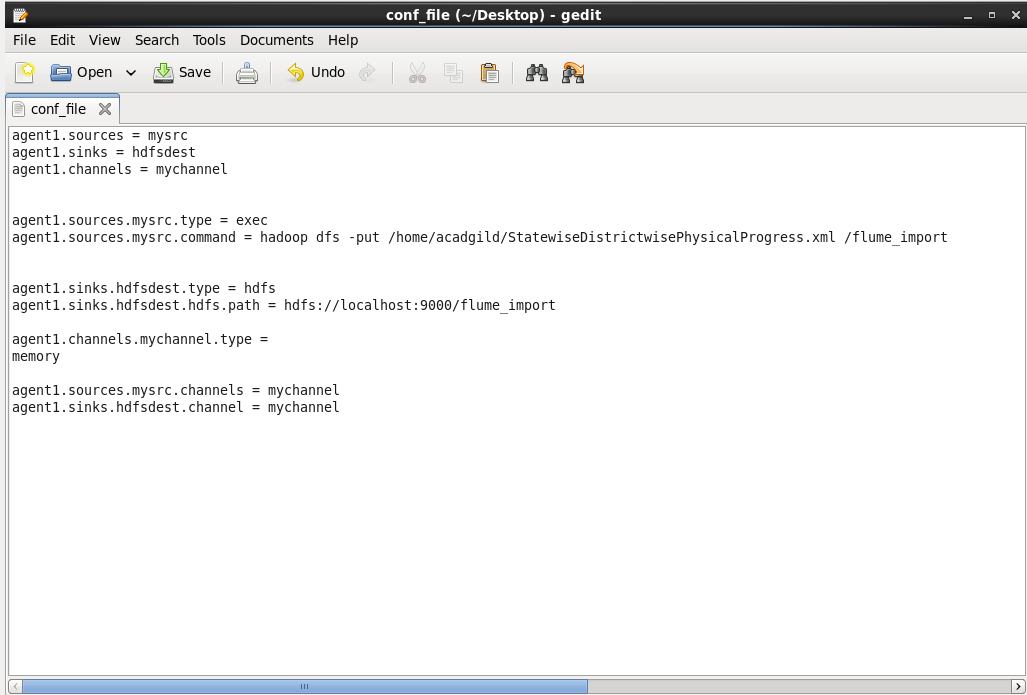
Copy dataset from local file system to HDFS using flume.

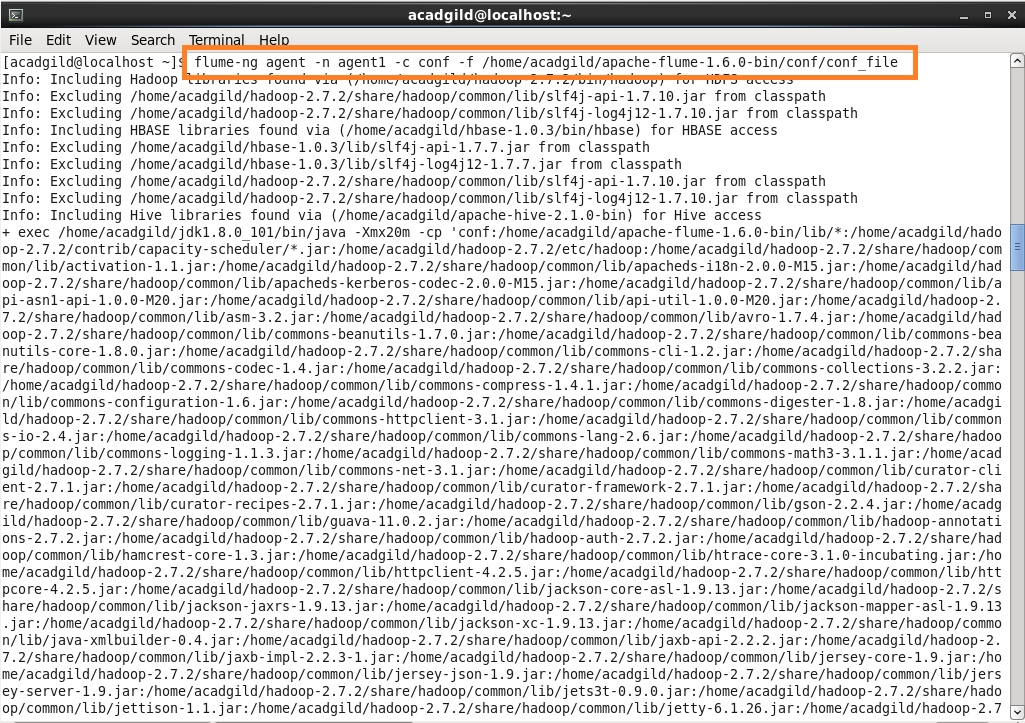
Command: flume-agent agent –n agent1 –c conf –f

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

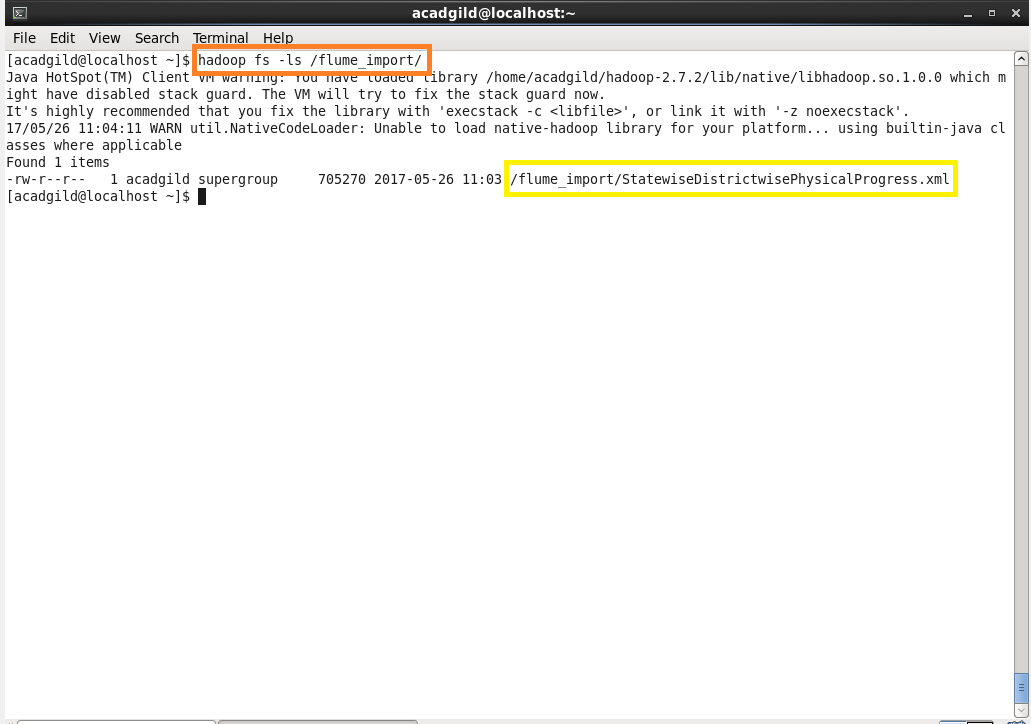


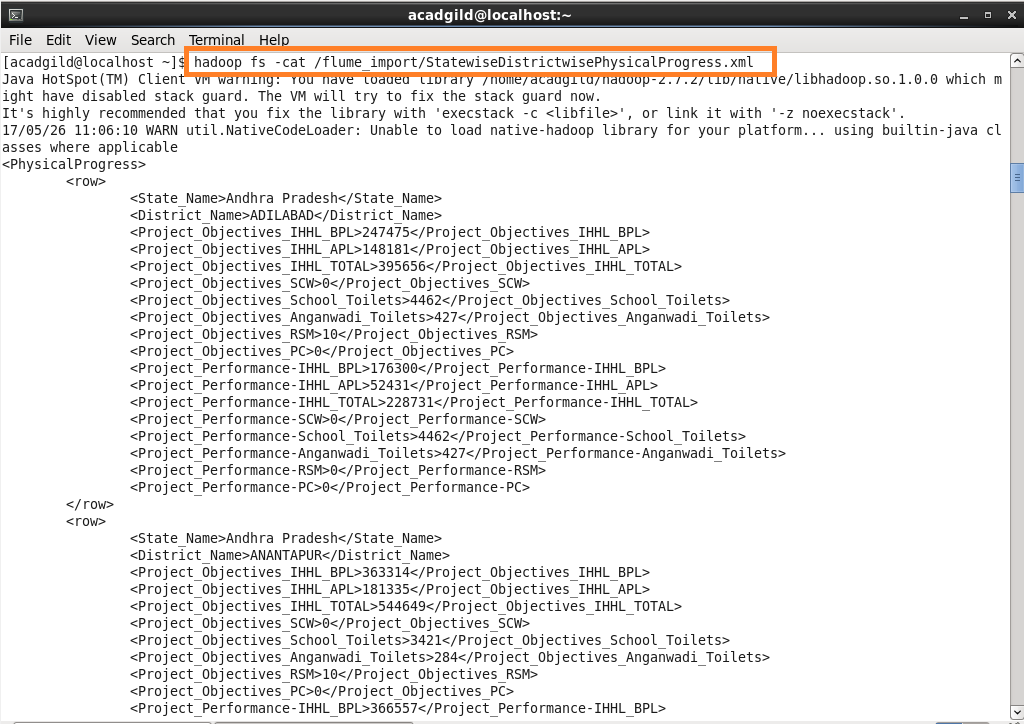
* Copying dataset from local file system to HDFS using flume by using the below conf\_file.

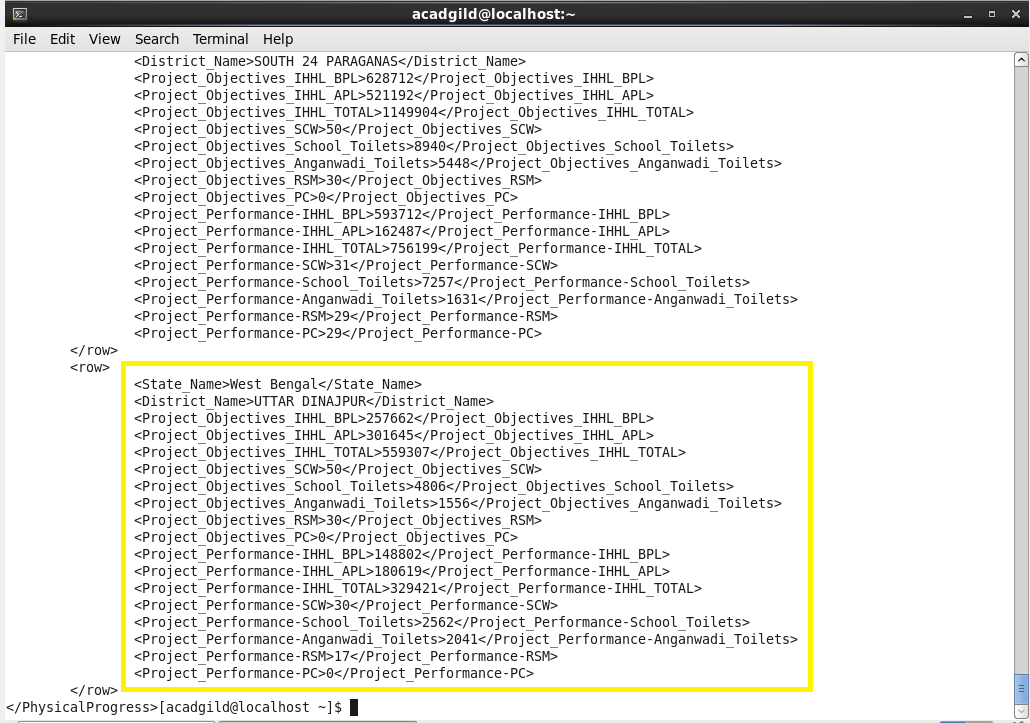




* Listing the created directory and performing basic cat operation.



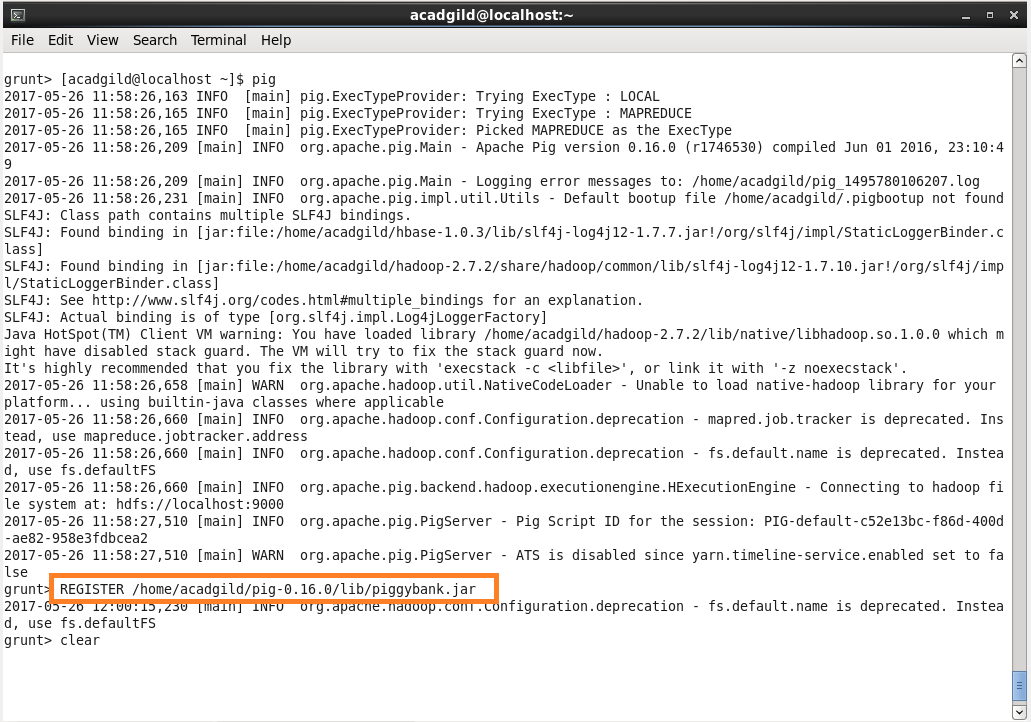


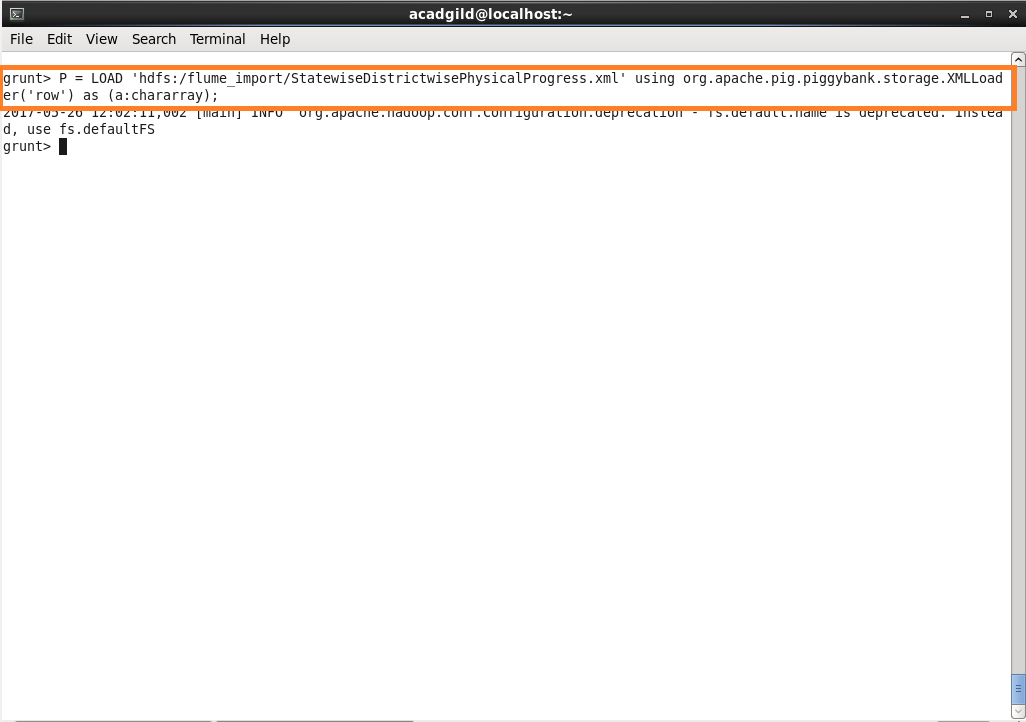


Here data is loaded from /flume project location which contains a XML file and in order to parse the xml file so that Pig can understand it.

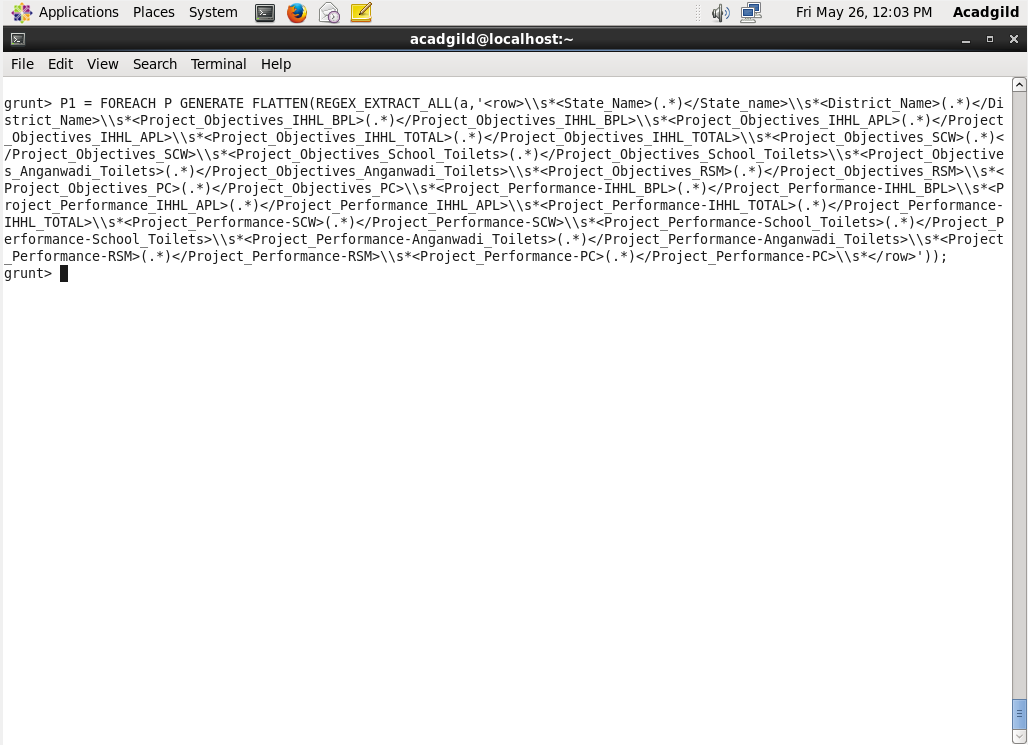
**Step 2:** Input file is in the XML format use Map reduce or pig to parse the data and get the results for the below problem statements.

* XMLLoader () function which is present in the jar Piggybank of Apache is used.

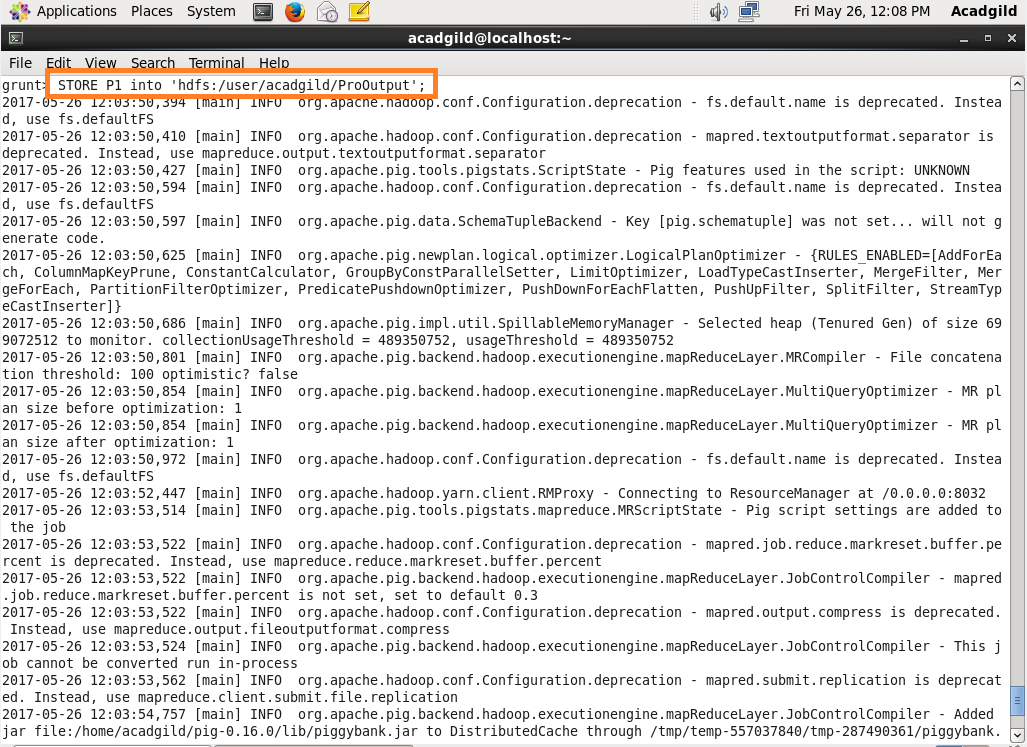




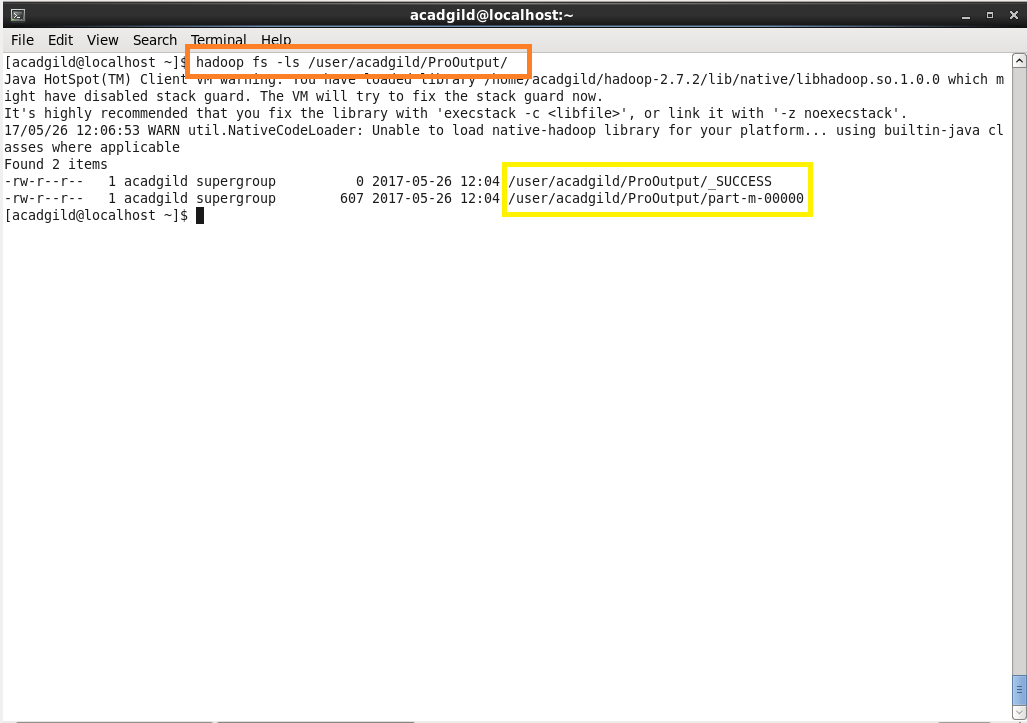
* Using regex-function for capturing the values within tag and FLATTEN is used flattening the TUPLE.



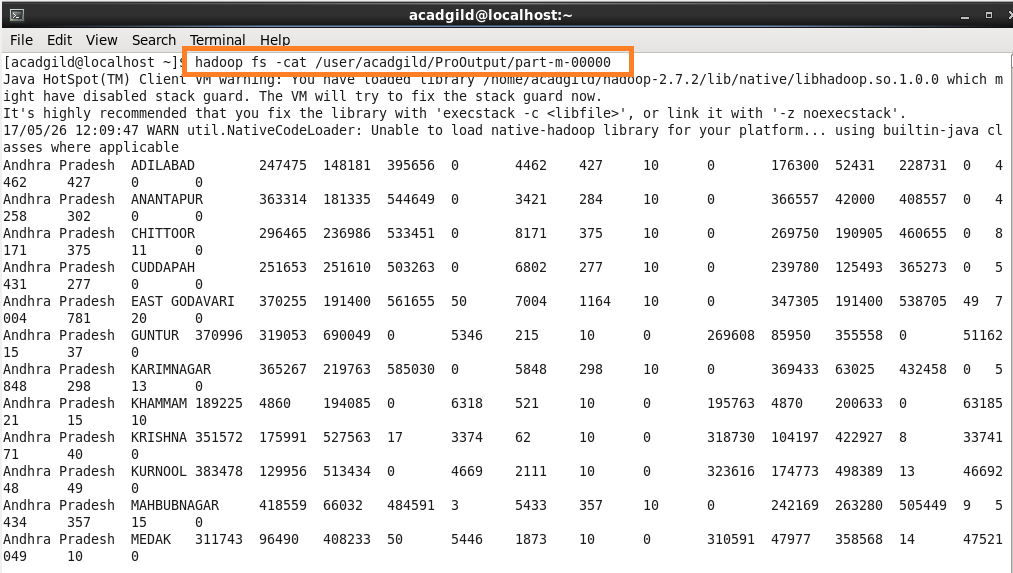
* Storing the above results into HDFS for further analysis.

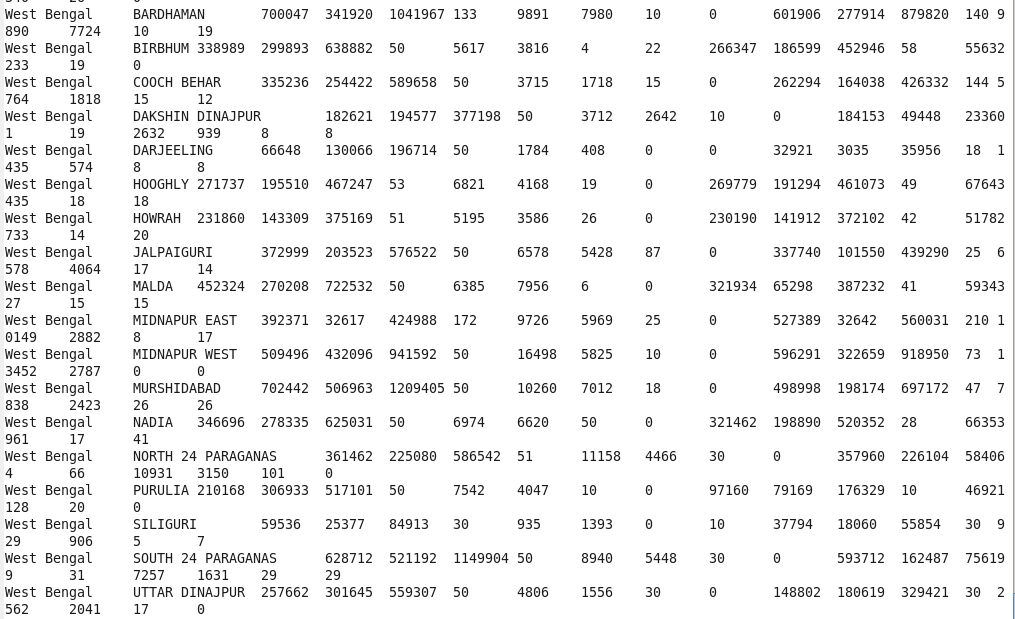


* Here, we can see that data is successfully stored.



* Using cat command we can see the stored data.





**Problem 1:**

Find out the districts who achieved 100 percent objective in BPL cards Export the results to MySQL using Sqoop.

**Logic:**

**Step 1:** Loading the dataset

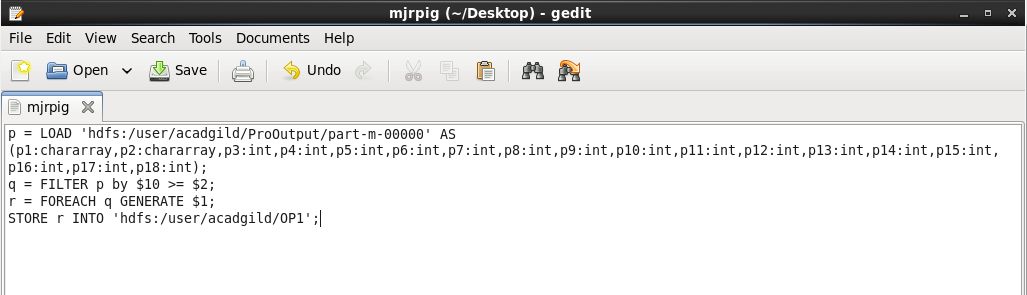
**Step 2:** Filtering the loaded data so as to get the districts who achieved 100 percent objective in BPL cards.

* $10 is the project objective in BPL cards column
* $2 is the project performance in BPL column

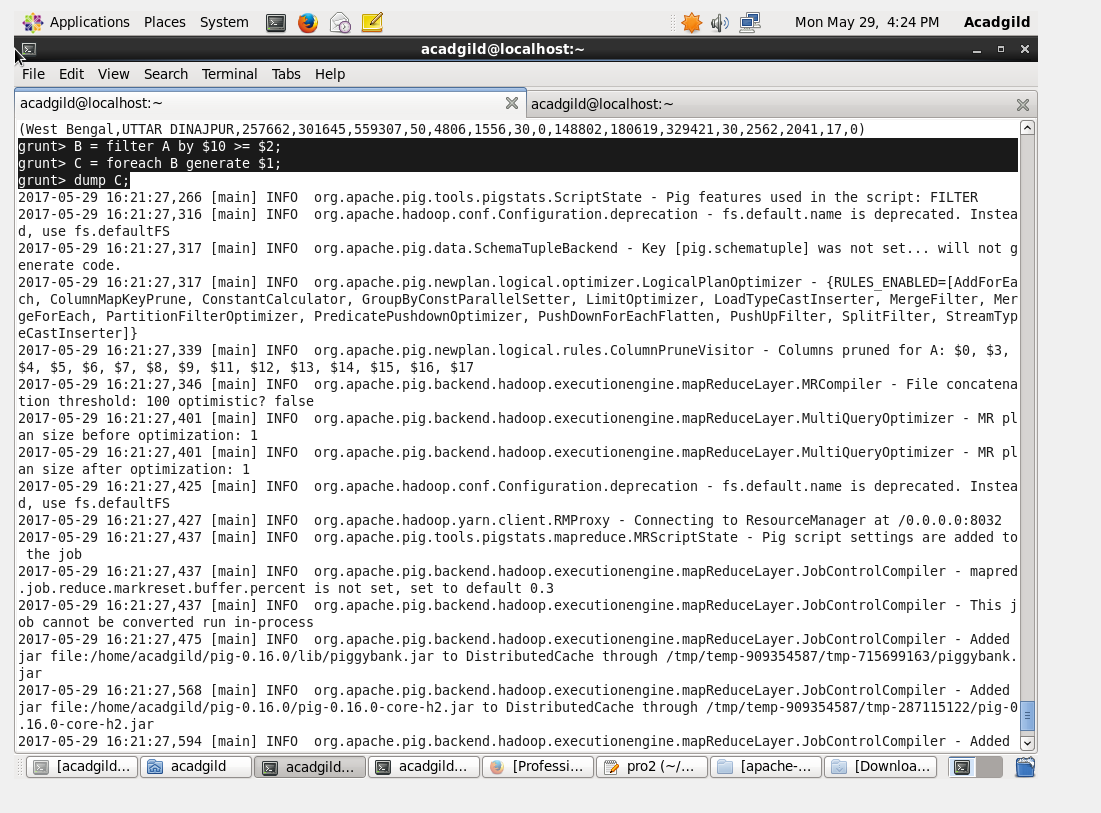
Comparing whether $10 i.e., project objectives in BPL have achieved 100 percent or not.

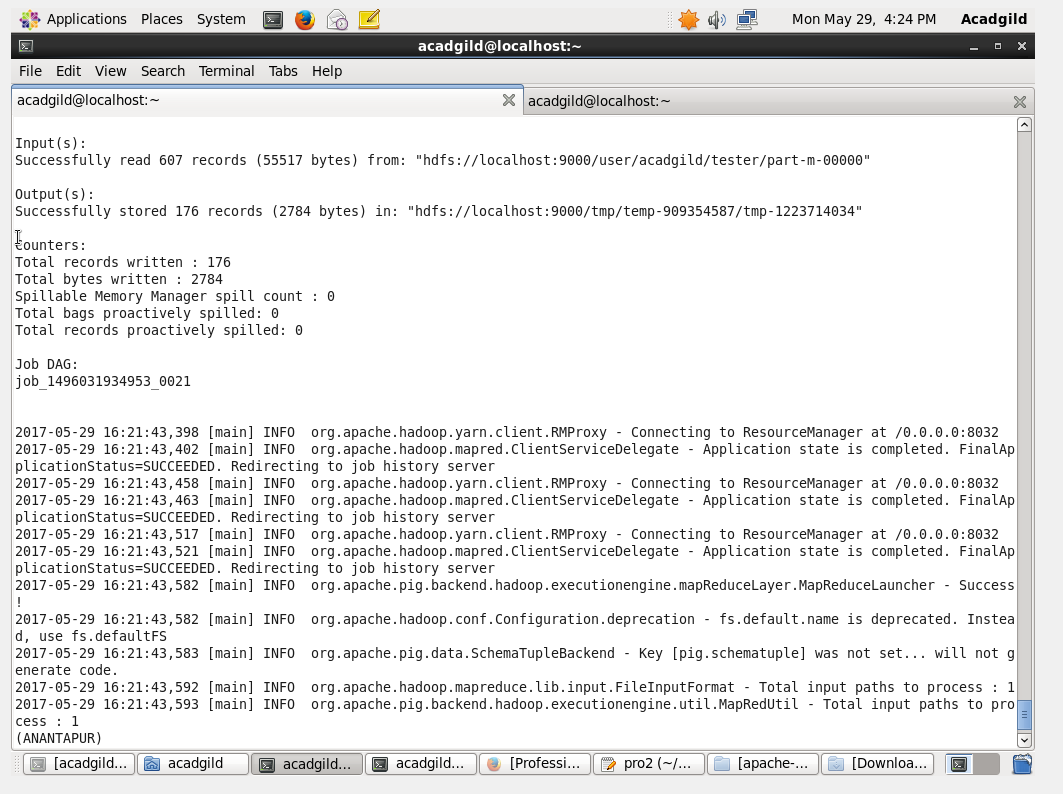
**Step 3:** If achieved then generating $1 which is the district column

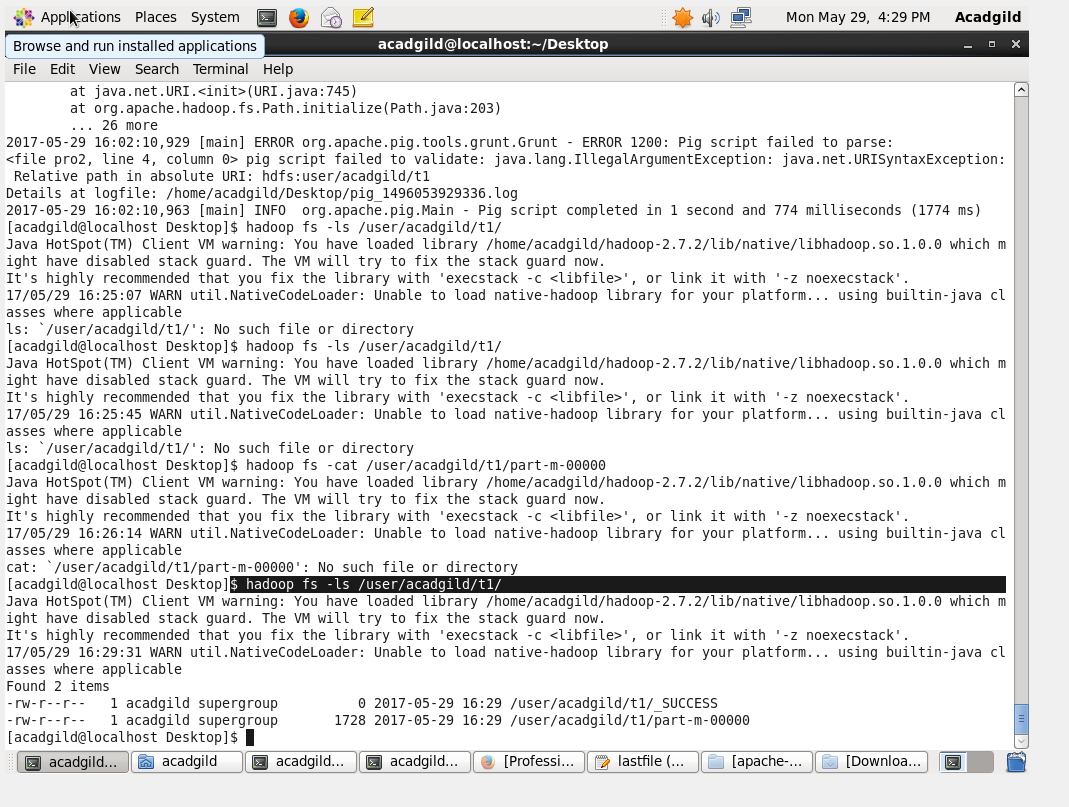
**Step 4:** Finally storing the results as OP1.



* Running the above Pig Script:

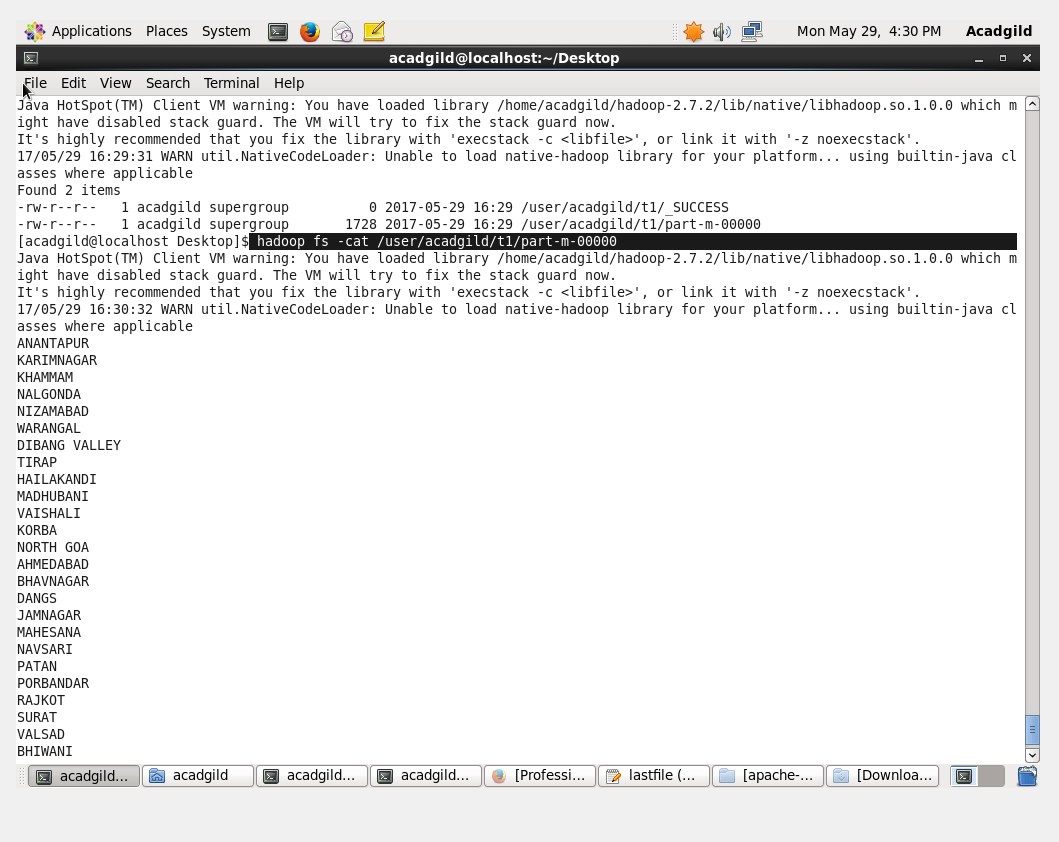


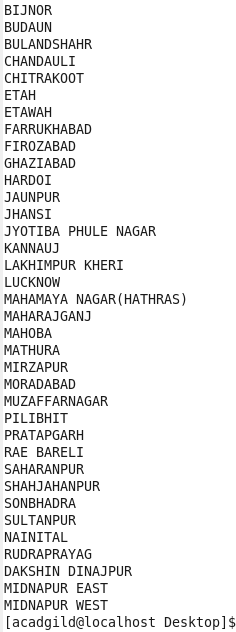




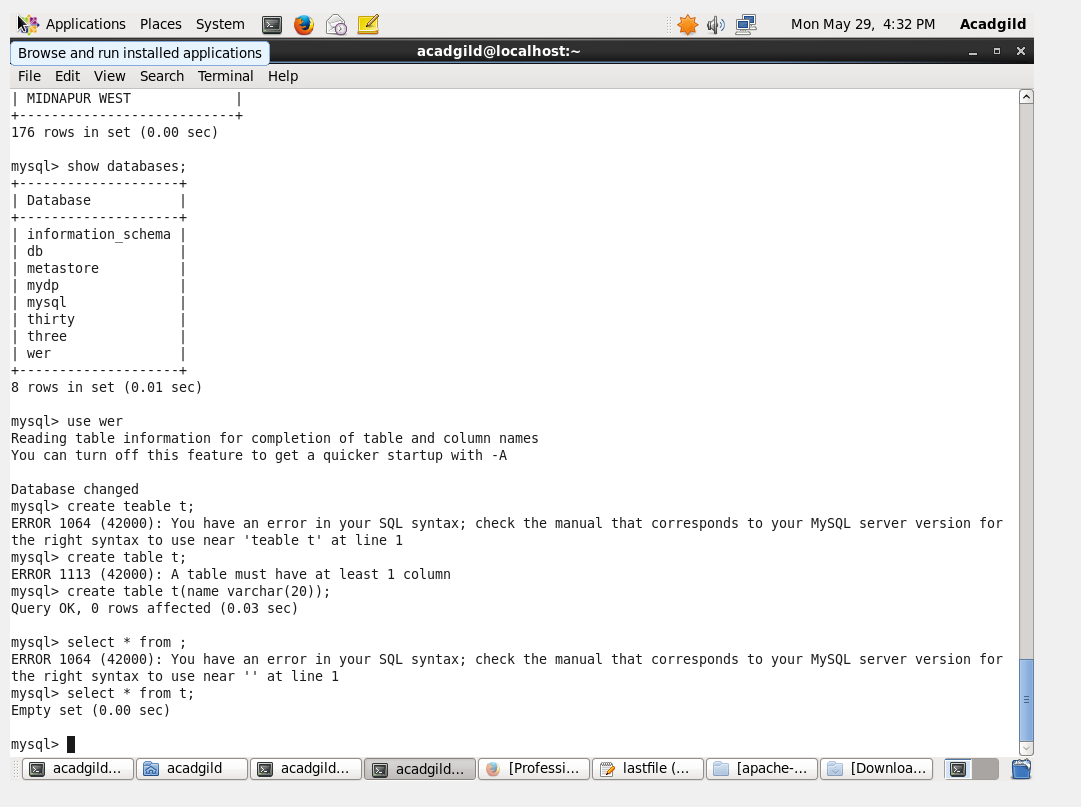
Here, we can see that job ran successfully.

Then, using cat command displaying the results.

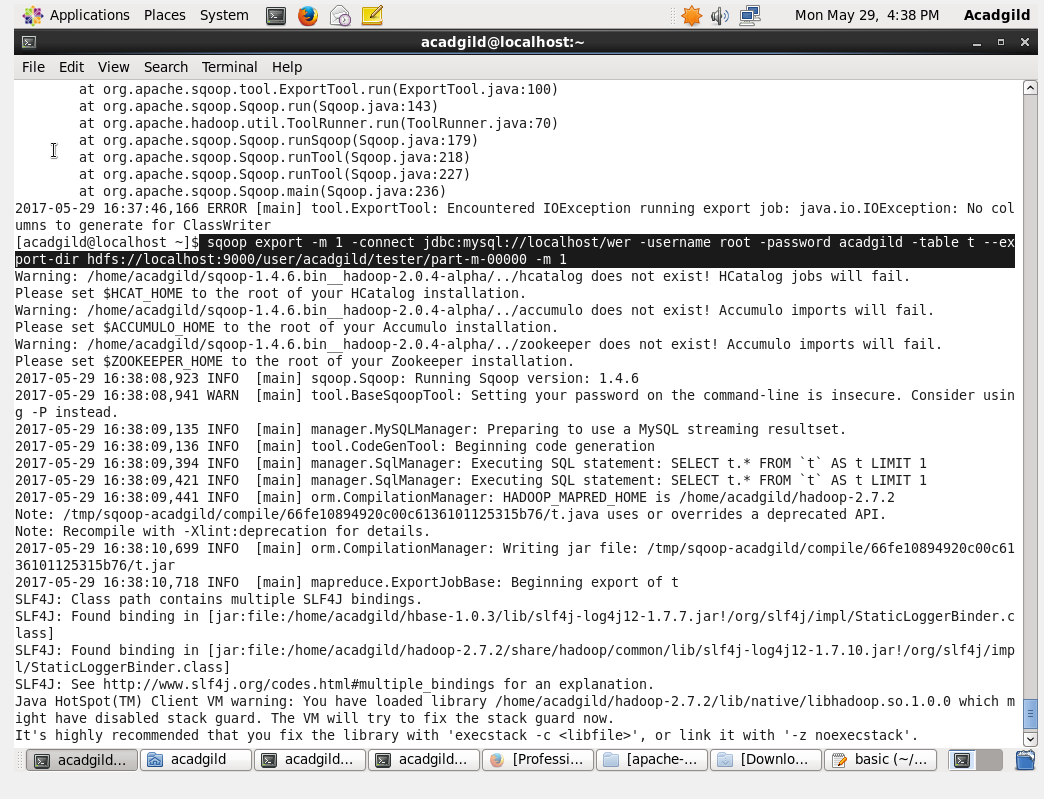


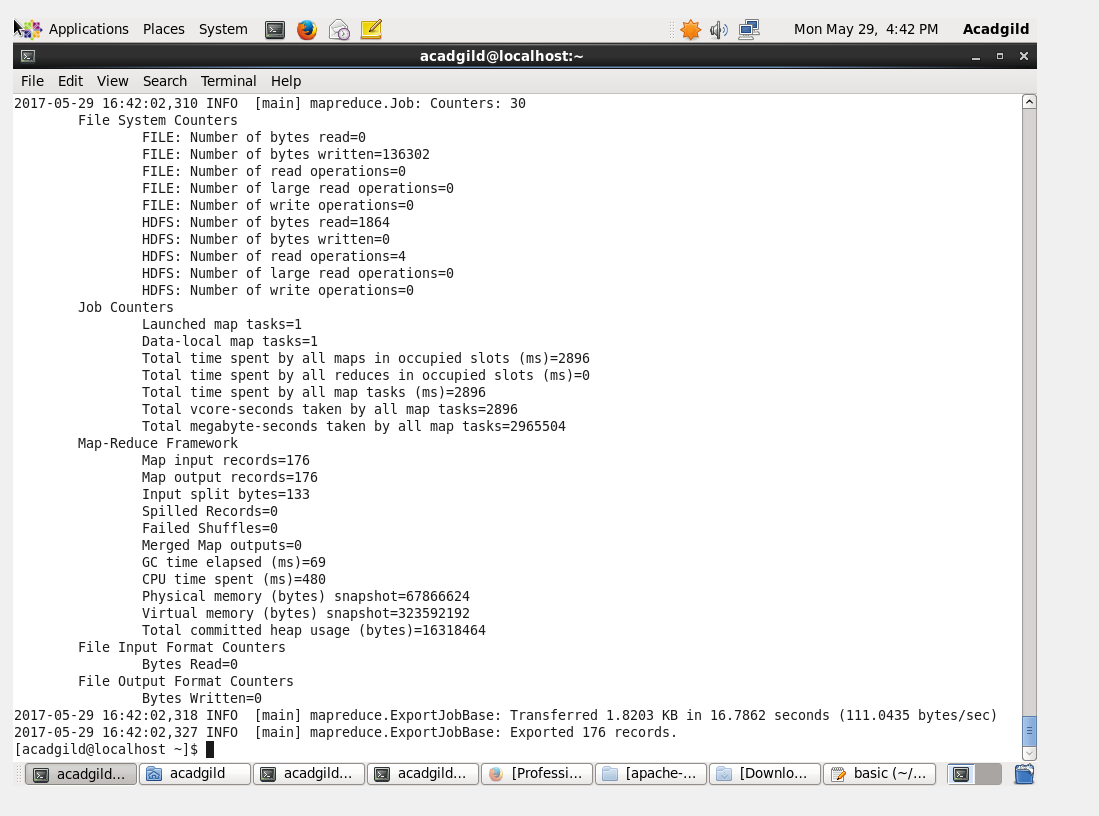


* To export the result from pig using Sqoop creating a table in MySQL and inserting no values in the table so as to perform the export.



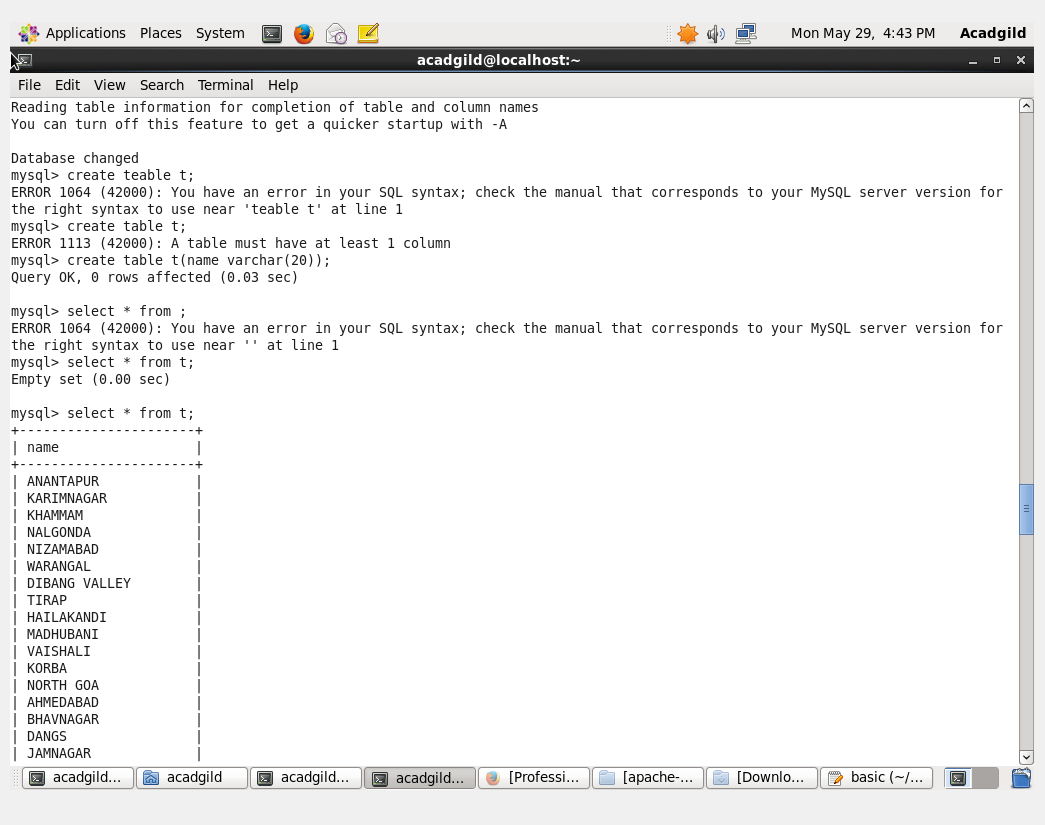
* Below is the export command to export the records from pig to mysql

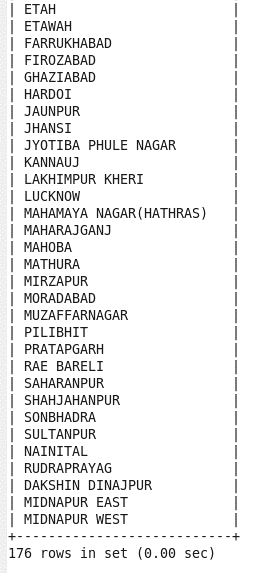




Here, we can see that the data is successfully exported.

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



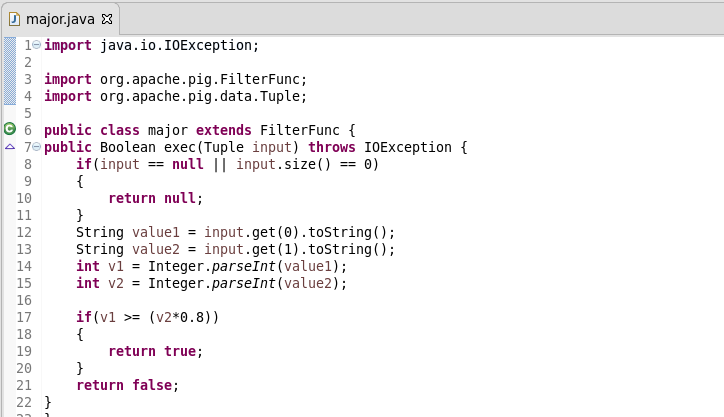


**Problem 2:**

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

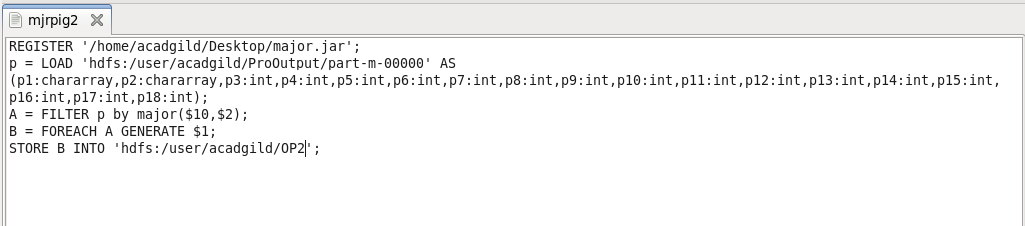
**Creating UDF:**

Adding the necessary Imports

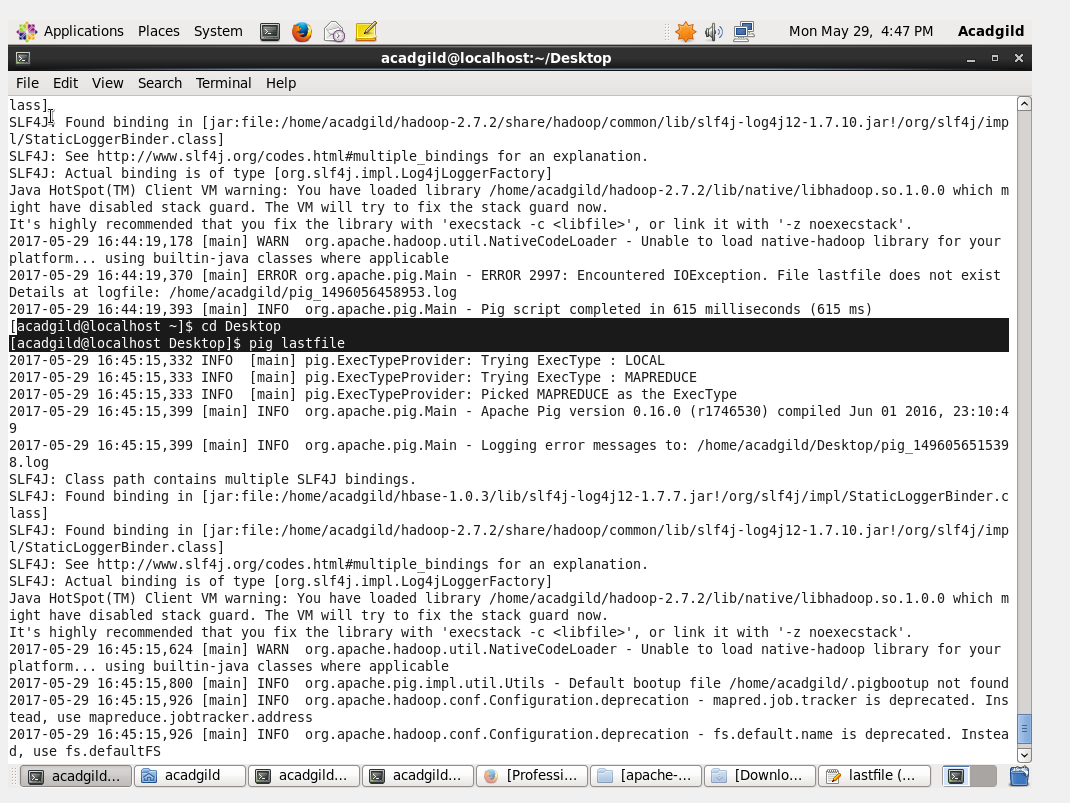


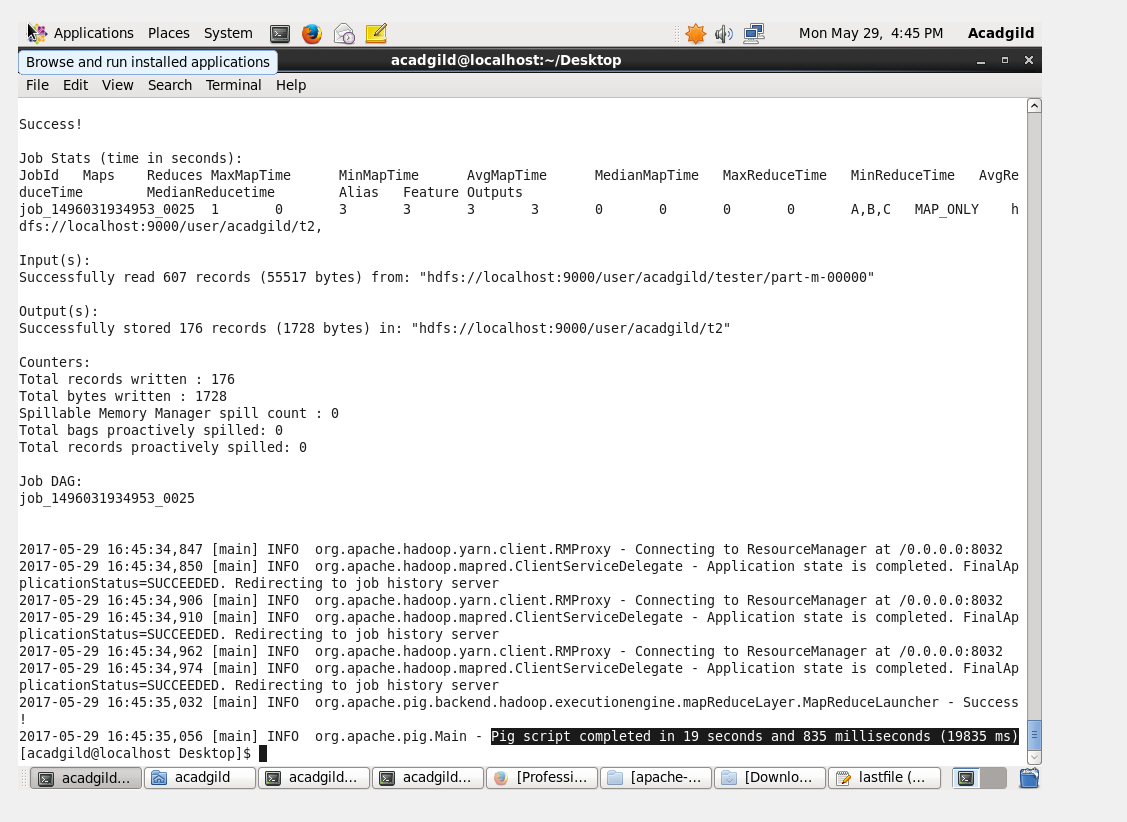
**Logic:**

* Firstly, registering the jar.
* Then, loading the data and the datatype and name of fields are defined.
* Filtering Districts that achieved 80%:
* Used a UDF function “major” which filter the districts which have achieved 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 “OP2” in HDFS from which it will be exported to MySQL.

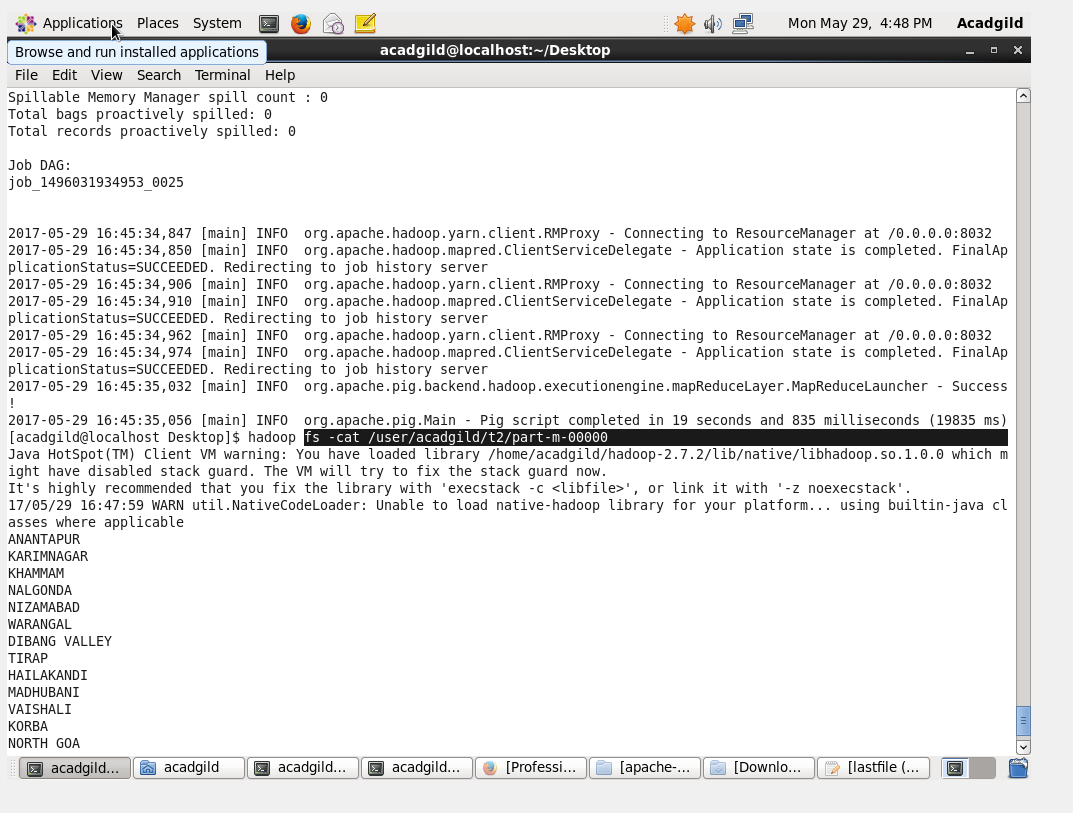


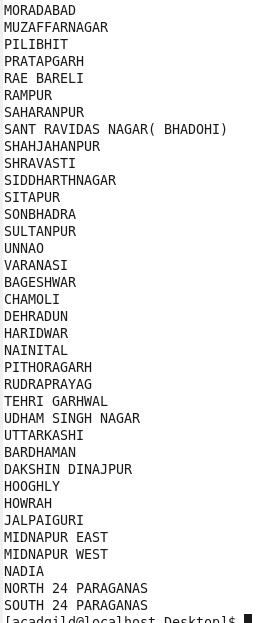
* Running above pigscript:



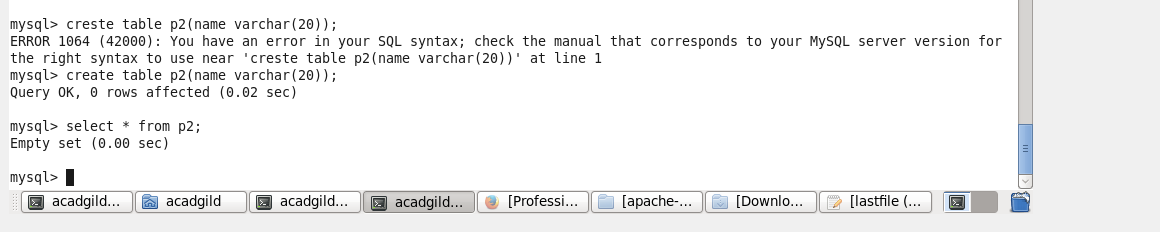


* Listing it and then performing cat operation on it.
* Cat command will display the stored output of the pig commands

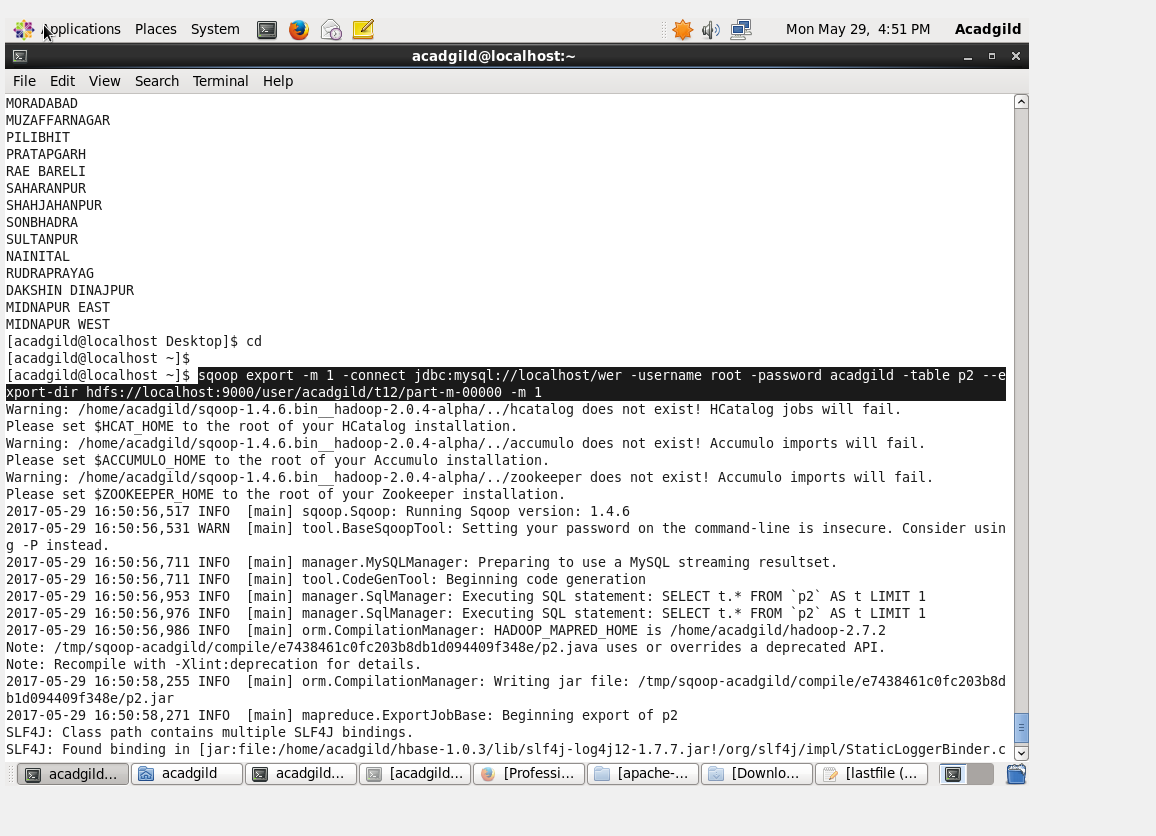




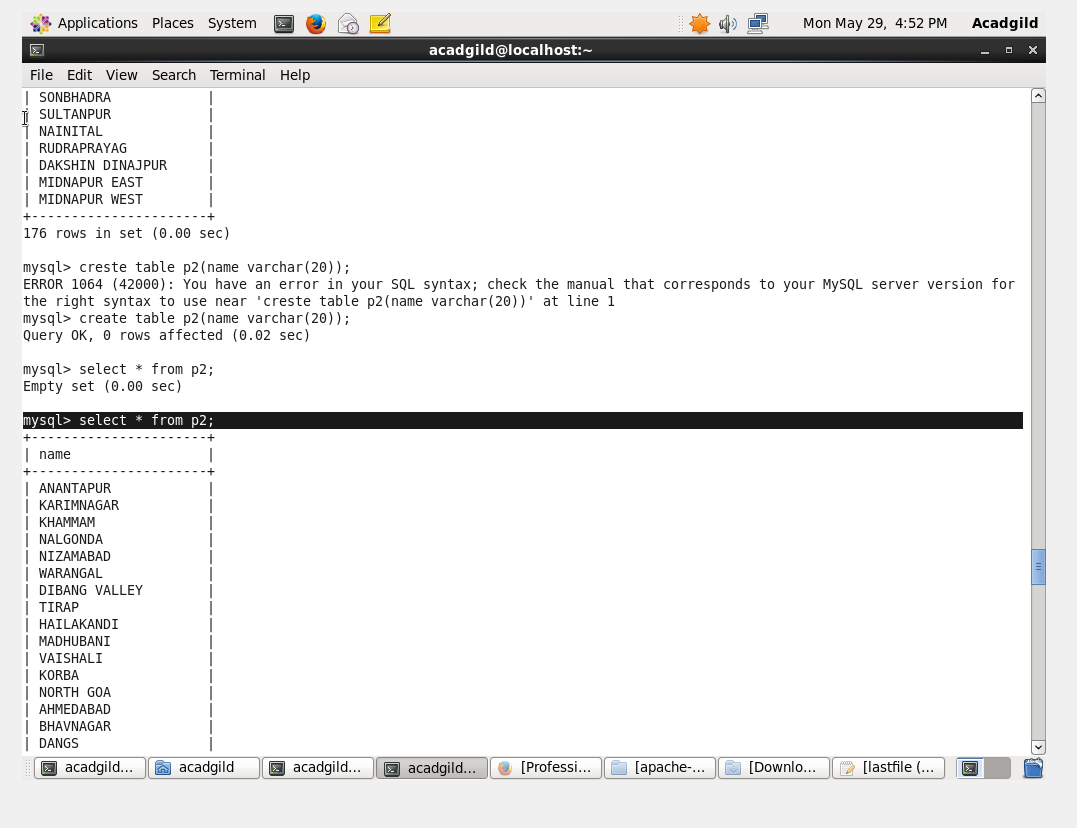
* To export the result from pig to MySQL using Sqoop creating a table in MySQL and inserting no values in the table so as to perform the export.

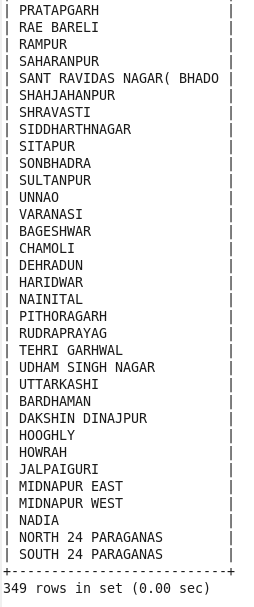


* Below is the export command to perform export operation



* Using select command displaying the results.





The full output was very long of 349 rows so only the first and last part of the outputs are attached above.