**STATE-WISE DEVELOPMENT ANALYSIS IN INDIA\**

This project is to give a glimpse over the overall content of Hadoop ecosystem. Here it starts with storing the dataset into hdfs through flume. Then the dataset is in xml format , so we can do it in two ways we can directly take it as input through pig latin or giving input format as xml in driver class in mapreduce program

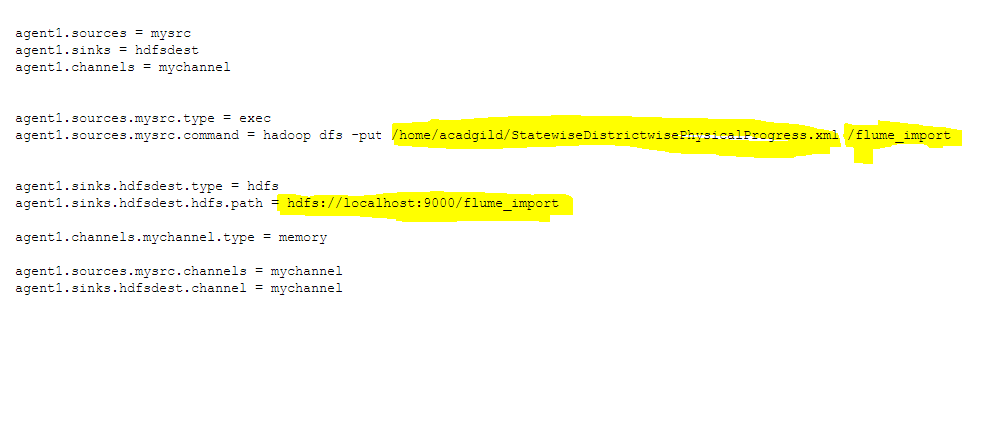
*Steps*

1:-

The flume configuration is downloaded and placed in /apacheflume/conf folder

Here this configuration file gives us the source ,sink,channel,sink command to be performed and where the content to be stored

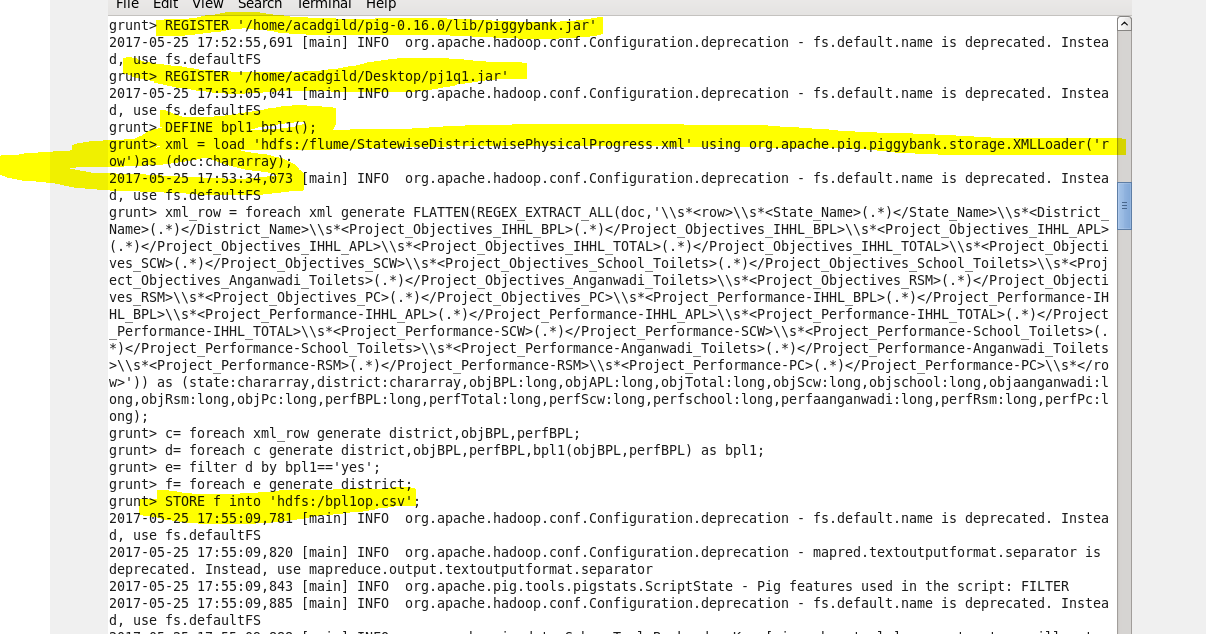
Here it is in local folder /home/acadgild/StatewiseDistrictwisePhysicalProgress.xml which is to be transferred to the hdfs folder /flume



*2):*

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

The question is to find the districts which achieved 100 % of its objective in bpl cards



Now we are creating the object from the java class

Then we are loading the xml file from hdfs system that consider each row as a value stored in a x of data type –chararray.

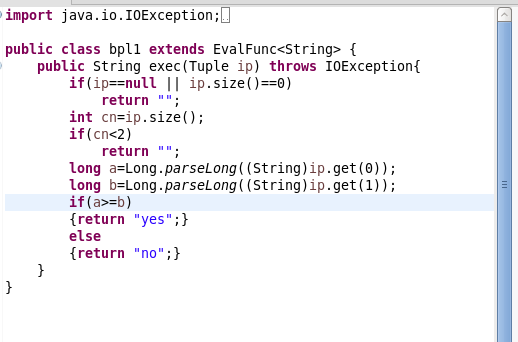
This would help in obtaining the each row value

Next we are using regular expression to obtain the value from each row removing the column names and we are giving each column’s data type

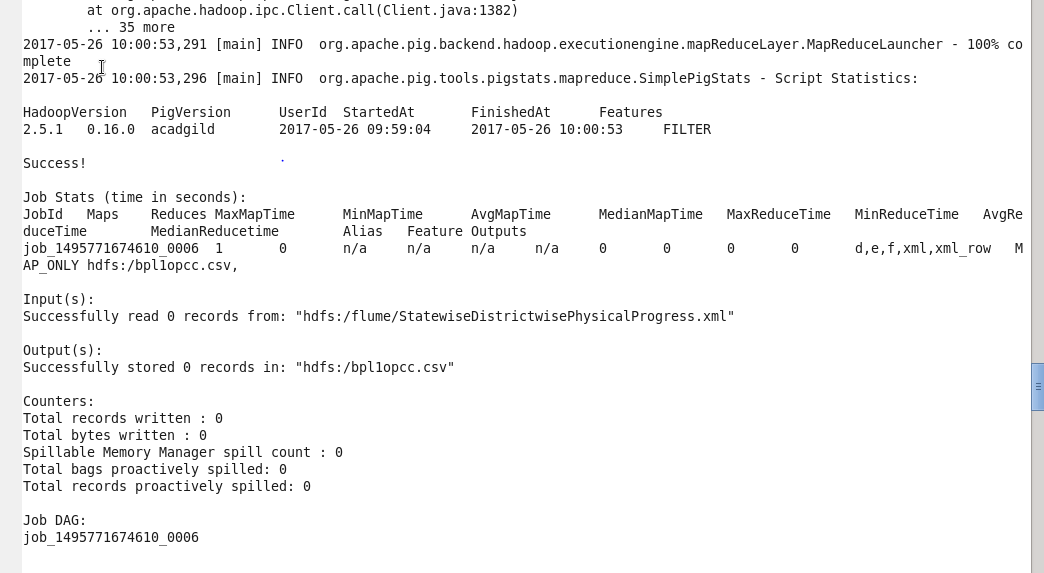
Then from this we are generating only the district name,objective of bpl and performance of bpl

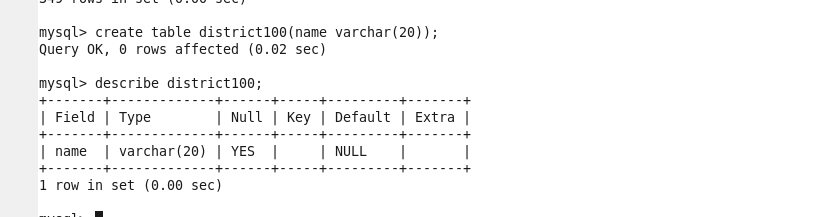
The filtered content is given as input for the jar class

The jar class compares the performance with objective and check whether it is more than that or not .If it is achieved it return ‘yes’ or else ‘no’

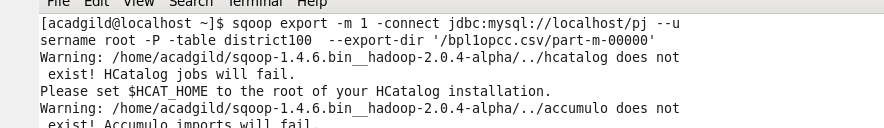


* First the input tuple is checked whether it is null or 0,if yes it returns null value
* Then it checks whether it has two column values also and then we are converting these values into long
* No we are checking whether the performance is more than the objective,if it is it returns ‘yes’.
* Then from the obtained output ‘yes’ content alone is filtered
* Then for each ‘yes’ corresponding district name is obtained
* This is then stored into a hdfs file

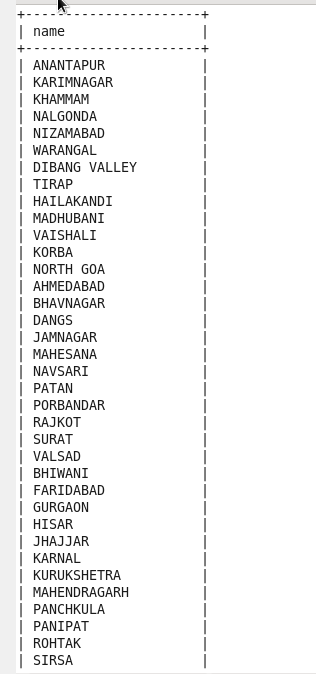


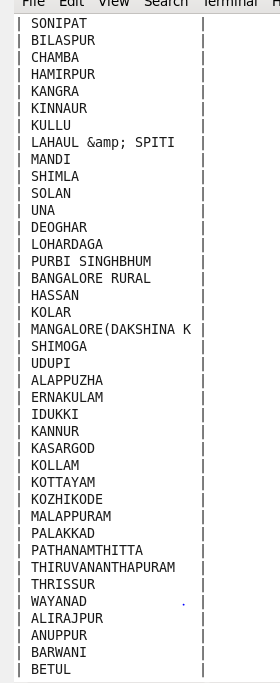


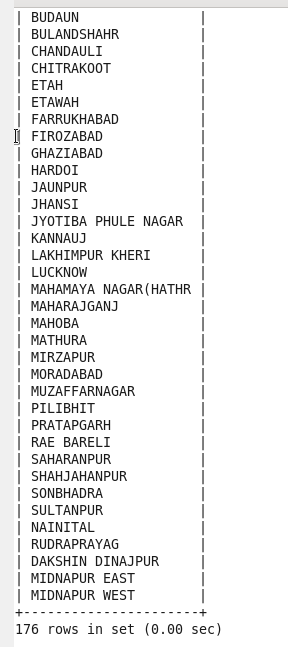
**Now the hdfs file is exported to mysql through sqoop which connects mysql and store the content in pj database ,district100 table from the output hdfs file ‘/bpl1opcc.csv/part-m-00000’**



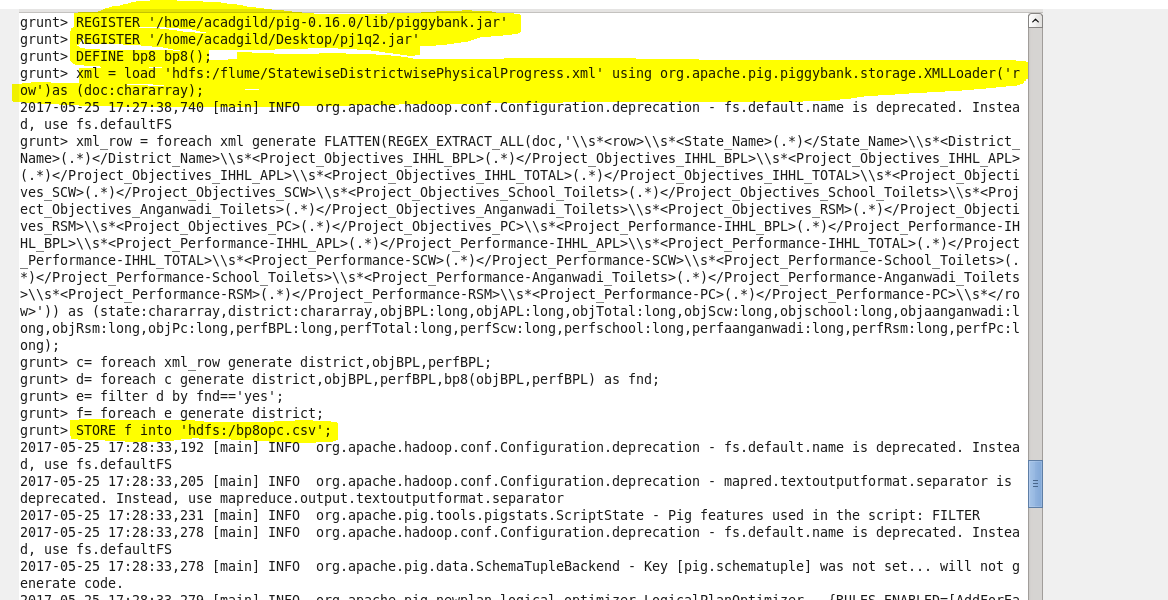
*The following districts has achieved 100% of objective in bpl*







*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.*

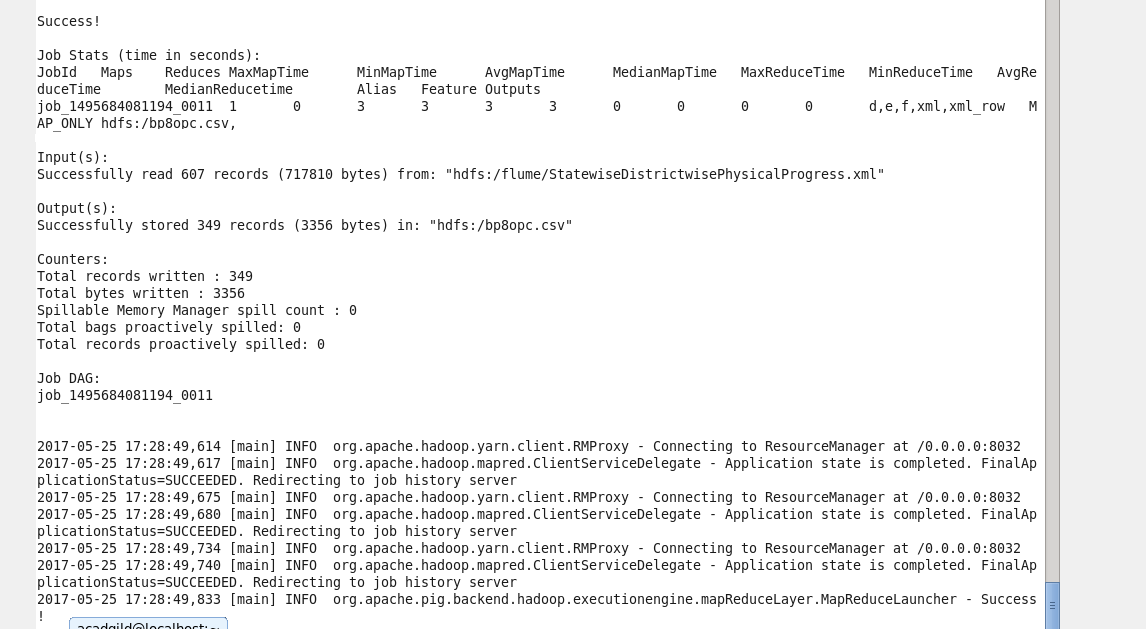


Next we are using regular expression to obtain the value from each row removing the column names and we are giving each column’s data type

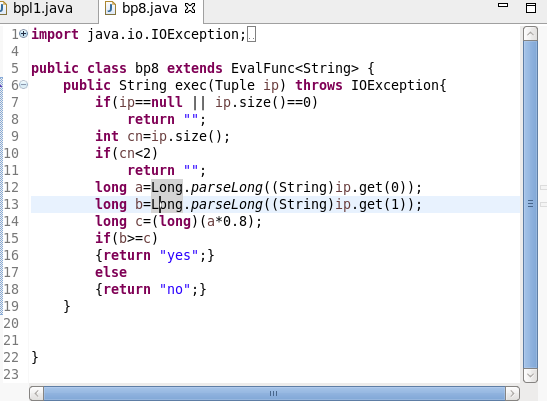
Then from this we are generating only the district name,objective of bpl and performance of bpl

The filtered content is given as input for the jar class

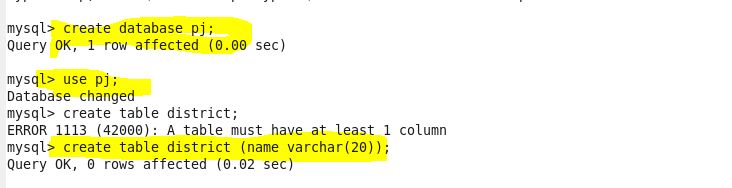
The jar class compares the performance with objective and check whether it is more than that or not .If it is achieved it return ‘yes’ or else ‘no’



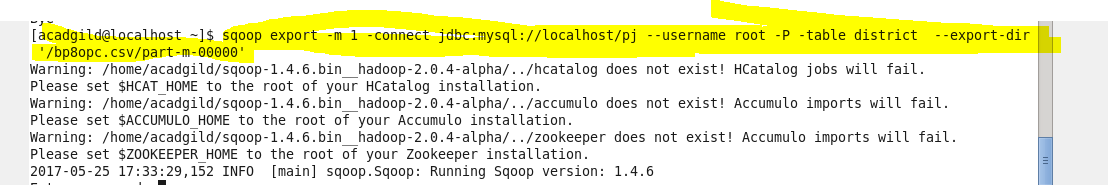
* First the input tuple is checked whether it is null or 0,if yes it returns null value
* Then it checks whether it has two column values also and then we are converting these values into long
* No we are checking whether the performance is 80% of the objective,if it is it returns ‘yes’.
* Then from the obtained output ‘yes’ content alone is filtered
* Then for each ‘yes’ corresponding district name is obtained
* This is then stored into a hdfs file



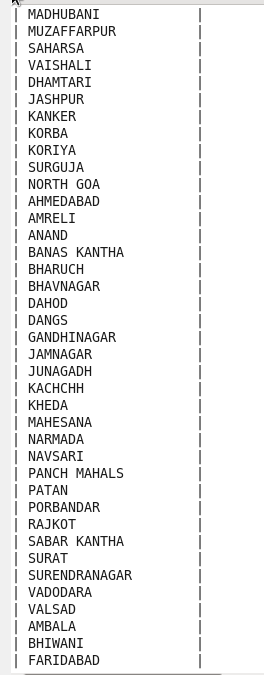
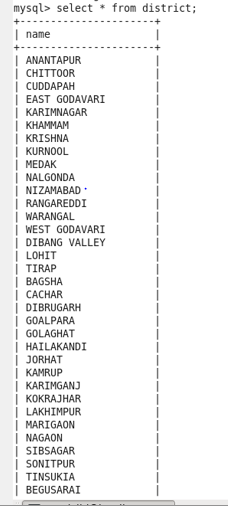
we are creating mysql table to store the districts that obtained 80% of the objective

****

Now the hdfs file is exported to mysql through sqoop which connects mysql and store the content in **pj database ,district table** from the output hdfs file **‘/bp8opc.csv/part-m-00000’**



*These districts have achieved 80% of objective in bpl*





*finish*