**State-Wise Development Analysis In India**

**Problem Statement:**

* 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

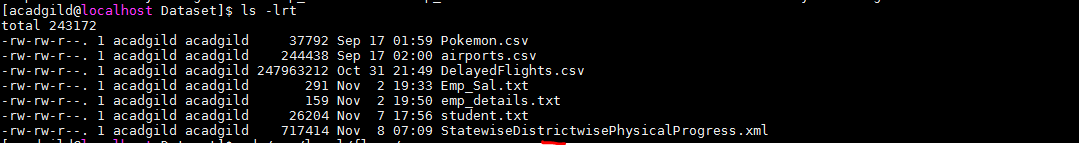
**Priority Definitions**

The following definitions are intended as a guideline to prioritize requirements.

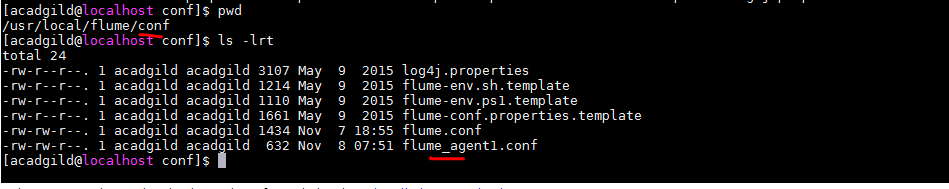
* Priority 1 – Create FLUME job for fetching log files from spool directory the data
* Priority 2 – MapReduce/pig job to preprocess

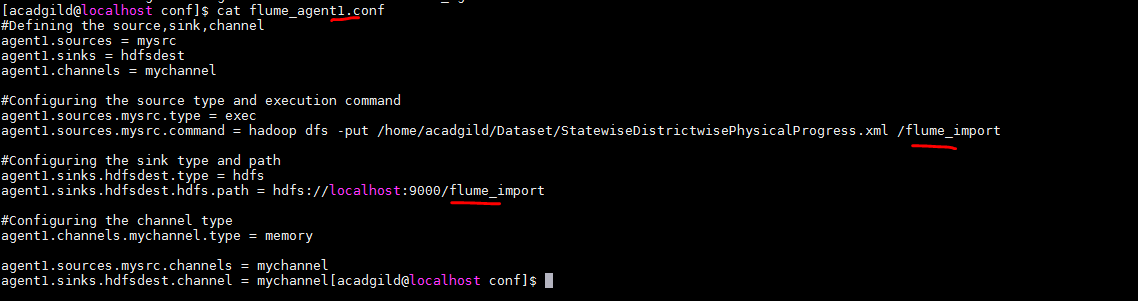
**Flume job:**

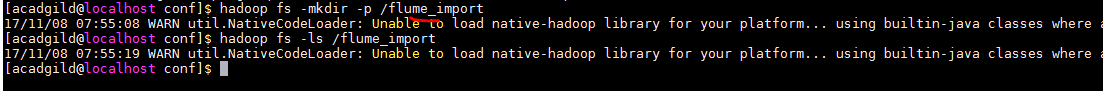
**Input Dataset in local:**

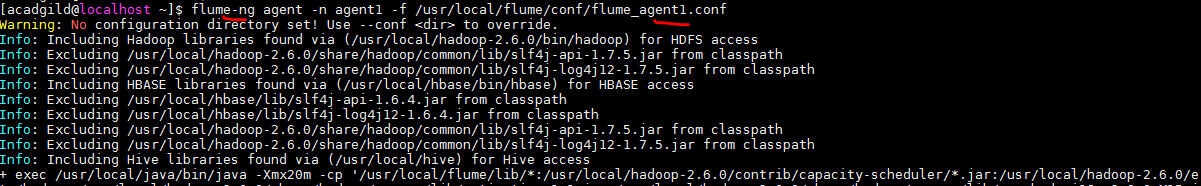
Input is in local and needs to be transferred to HDFS.

**Conf file – for flume agent:**

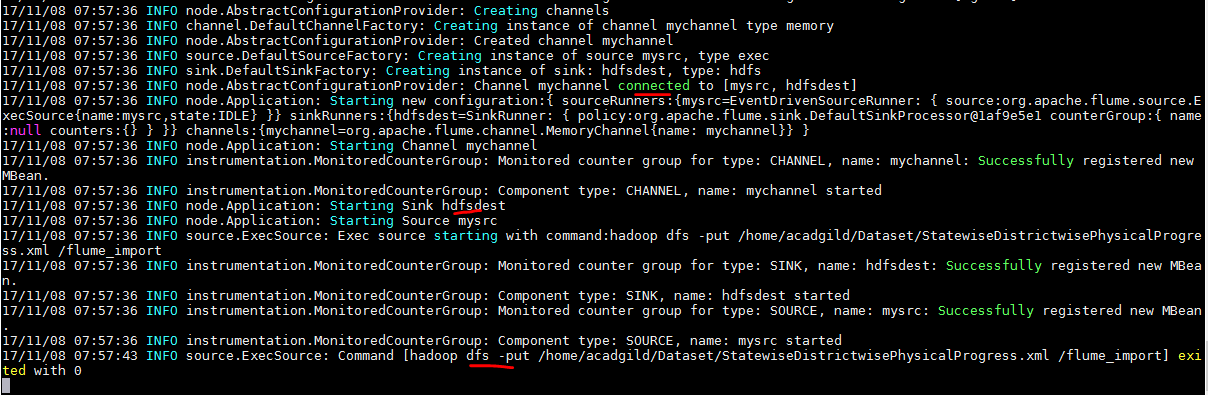
Below conf file is to define the properties of source(exec type),sink(target),channel.

Created a new directory in HDFS as a sink for flume(from where we can use the data for further analysis)

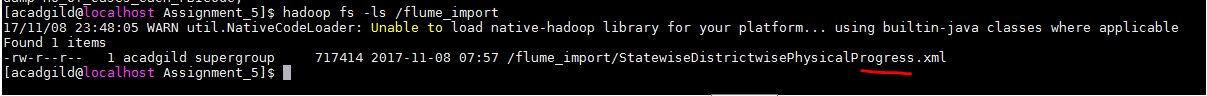


**Starting the flume agent – agent1:**

**Flume log:**



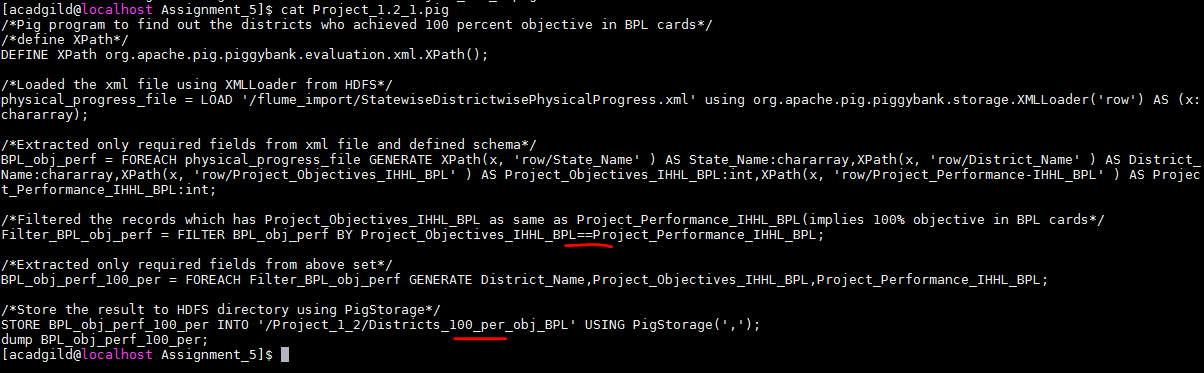
**HDFS sink after flume import:**

Input xml dataset is now transferred to HDFS by Flume agent.

**Analysis:**

**1. Find out the districts who achieved 100 percent objective in BPL cards**

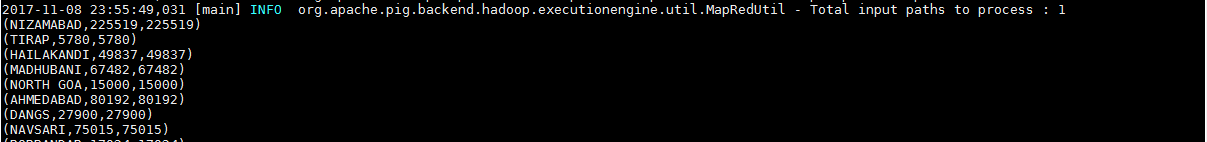
**Export the results to mysql using sqoop**

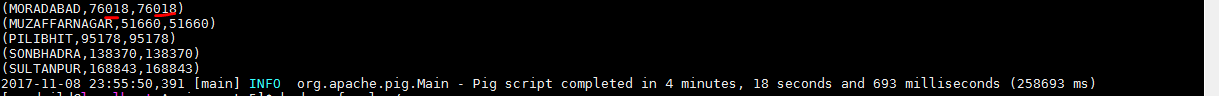
**Pig program:**

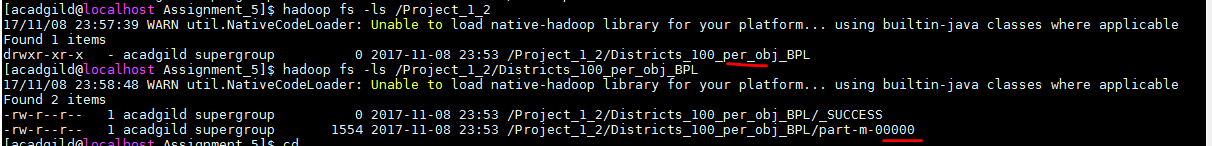
**Program invocation:**



**Output:**



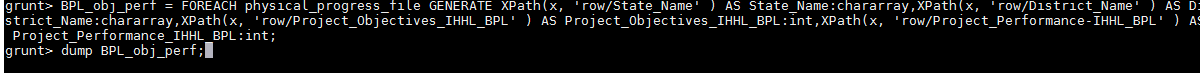


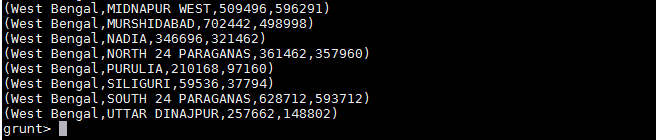


**Commands:**

**i) Loaded the xml file using XMLLoader(as input file is xml file) from HDFS**

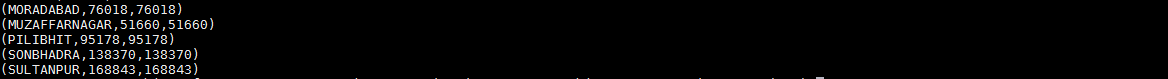


**ii) Extracted only required fields from xml file using XPath and defines a schema**



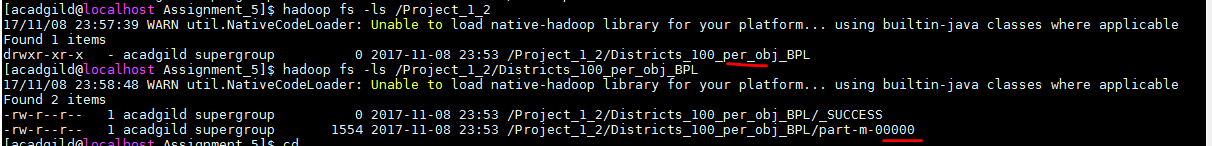
**iii) Filtered the records which has Project\_Objectives\_IHHL\_BPL as same as Project\_Performance\_IHHL\_BPL(implies 100% objective in BPL cards – implies who achieved 100 percent objective in BPL cards)**

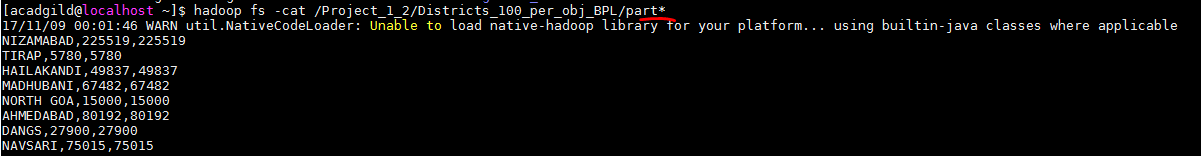
**iv) Extracted only required fields from above set**

**Final output:**

**v) Store the result to HDFS directory using PigStorage for further transferring it to MySql.**

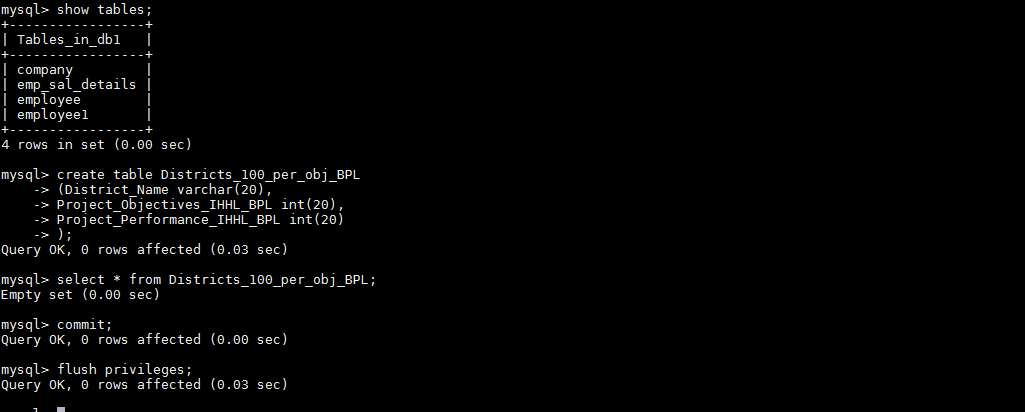


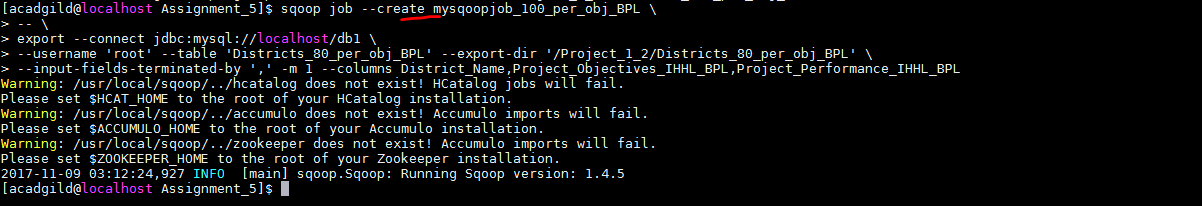




**Sqoop export from HDFS to Mysql:**

**For transferring data from HDFS to Mysql table should exist.Created a new table “Districts\_100\_per\_obj\_BPL” with columns as from above result inside ‘db1’ database.**

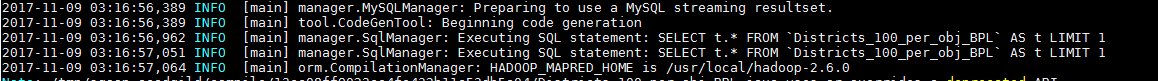


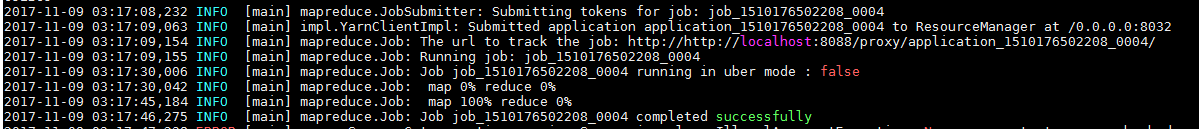
**Sqoopjob:**

**Sqoop job execution:**



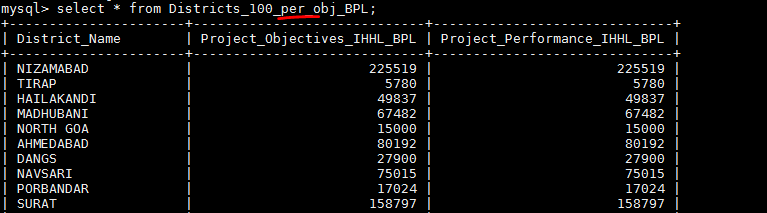
**Sqoop log:**

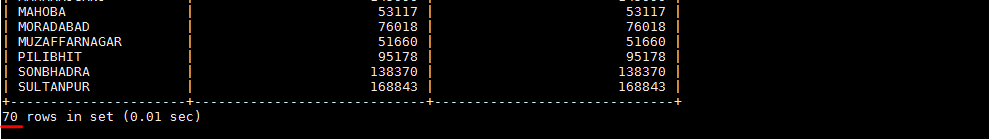




**Mysql table:**

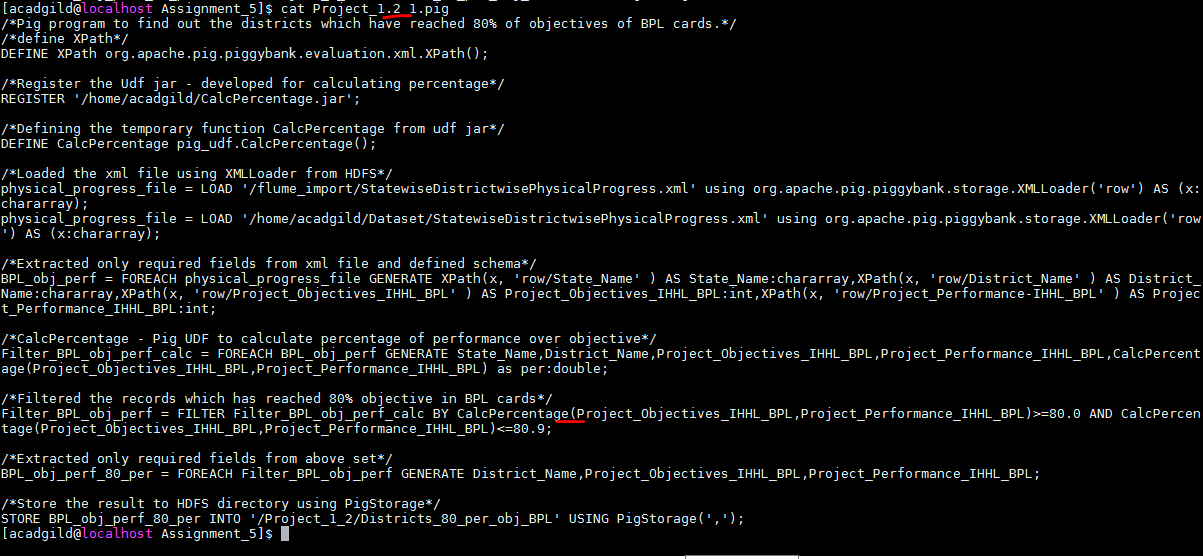
Data got loaded from HDFS to Mysql table.





**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.**

**Pig program:**

**Program invocation:**

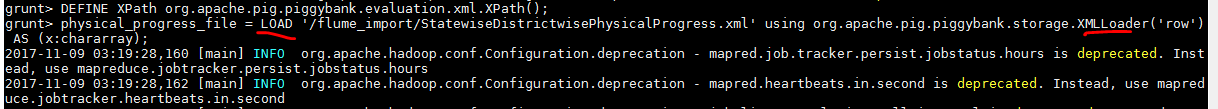


**Output:**

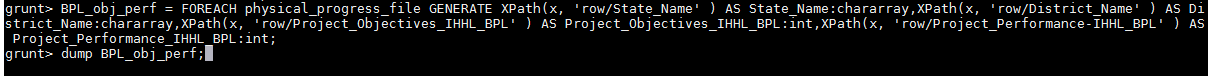
**Commands:**

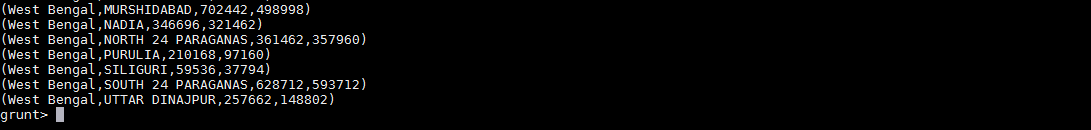
i) defineXPath

ii) Loaded the xml file using XMLLoader from HDFS



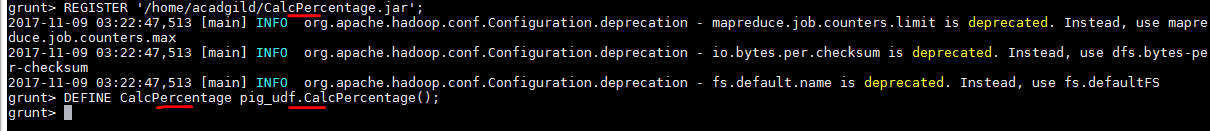
iii) Extracted only required fields from xml file and defined schema





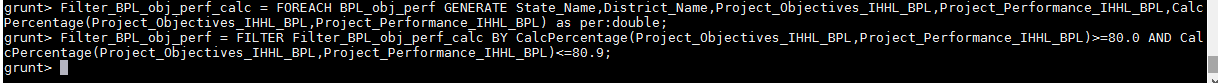
iv) Register the Udf jar - developed for calculating percentage

v) Defining the temporary function CalcPercentage from udf jar



vi) PerformCalcPercentage - Pig UDF to calculate percentage of performance over objective

vii) Filtered the records which has reached 80% objective in BPL cards



viii) Extracted only required fields from above set

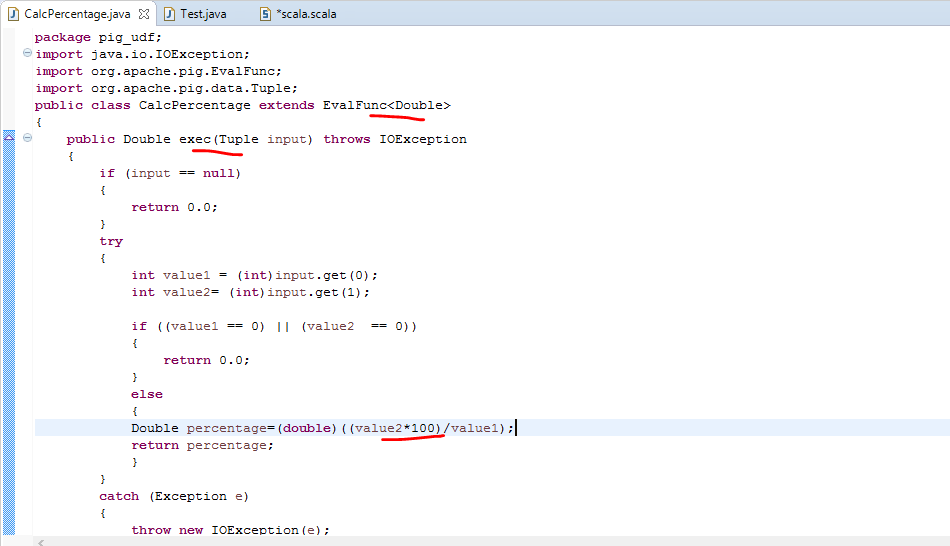


ix) Store the result to HDFS directory using PigStorage

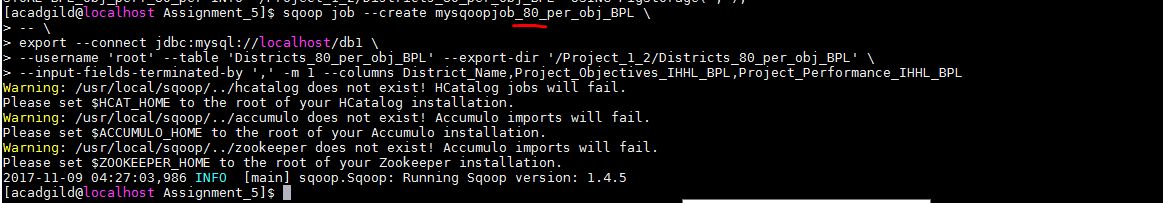


**Pig – UDF: to calculate percentage of objectives of BPL cards:**

Designed to calculate the percentage of two inputs and return percentage in double.



**Sqoop job:**



**Sqoop job execution:**

