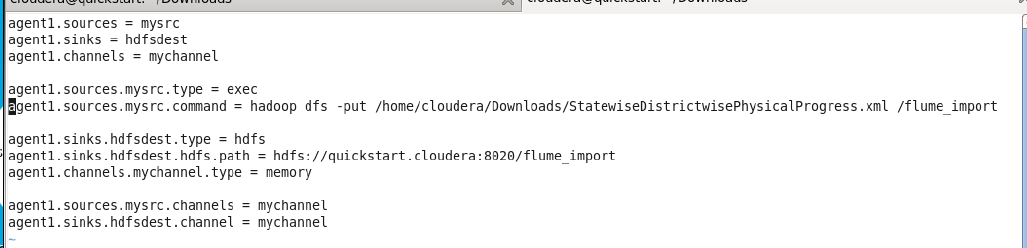
Task 1:

The FLUME job which will format the data and place the data to HDFS

1. Conf file to download dataset from local file system to HDFS flume:



1. Run flume-ng command to copy the file to HDFS.



Task 2:

Pig/MapReduce job for parsing the XML data.

Pig Script:

REGISTER '/home/cloudera/Downloads/jarfiles/piggybank-0.17.0.jar'

DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();

A = LOAD '/flume\_import/StatewiseDistrictwisePhysicalProgress.xml' using

org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

B = FOREACH A GENERATE XPath(x,'row/Project\_Objectives\_IHHL\_BPL'), XPath(x,'row/Project\_Performance-IHHL\_BPL');

dump B;

Execution:

Pig <pig\_file\_name>

OUTPUT:



Task 3:

Create Pig scripts/MapReduce jobs to analyze the data

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

Pig script:

REGISTER '/home/cloudera/Downloads/jarfiles/piggybank-0.17.0.jar'

DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();

A = LOAD '/flume\_import/StatewiseDistrictwisePhysicalProgress.xml' using

org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

B = FOREACH A GENERATE XPath(x,'row/District\_Name') as district ,XPath(x,'row/

Project\_Objectives\_IHHL\_BPL') as BPL\_Objective, XPath(x,'row/

Project\_Objectives\_IHHL\_TOTAL') as BPL\_Objective\_total;

C = filter B by (((int)BPL\_Objective \* 100)/(int)BPL\_Objective\_total) == 100;

STORE C INTO 'hdfs://quickstart.cloudera:8020/uscrime\_analysis' USING

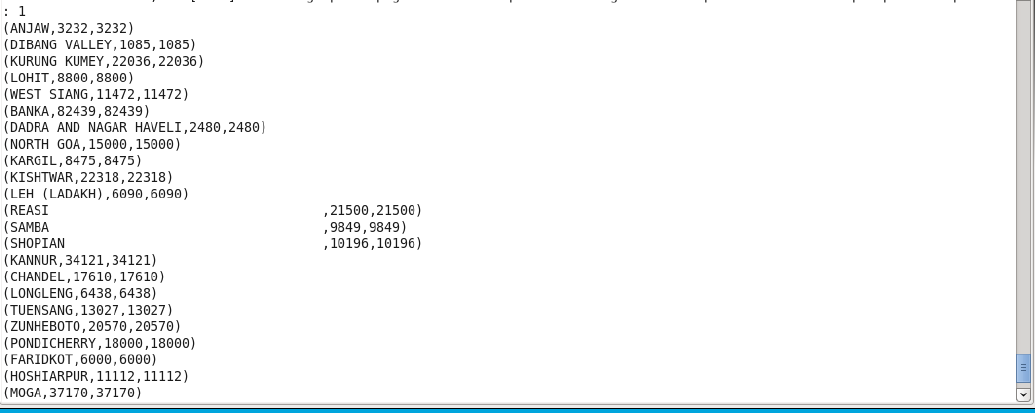
PigStorage (',');

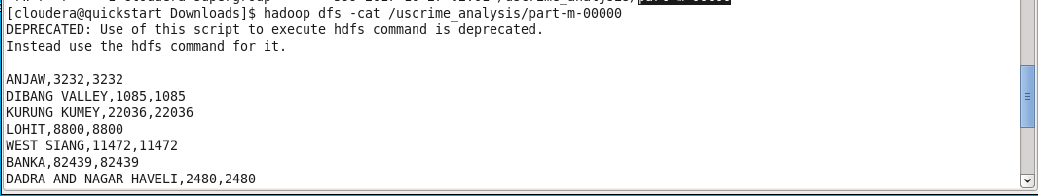
dump C;

Execution:

Pig <file\_name\_path>

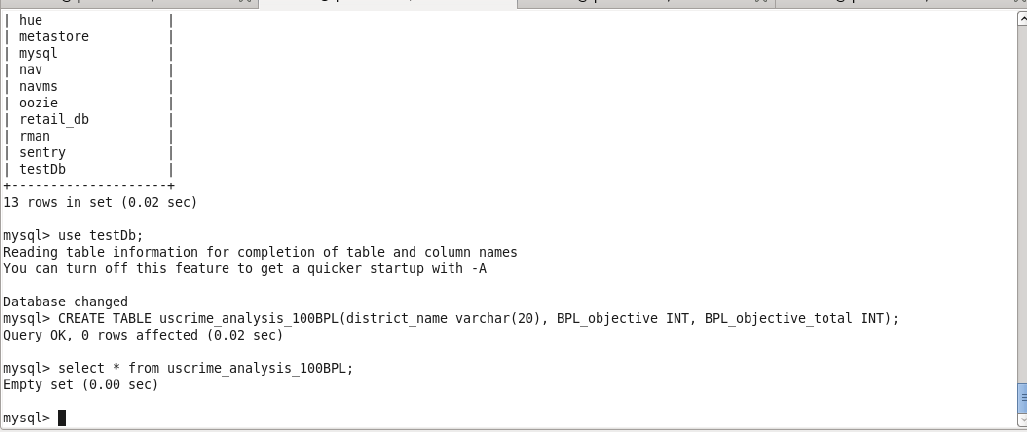
Output:





Export the results to mysql using sqoop:

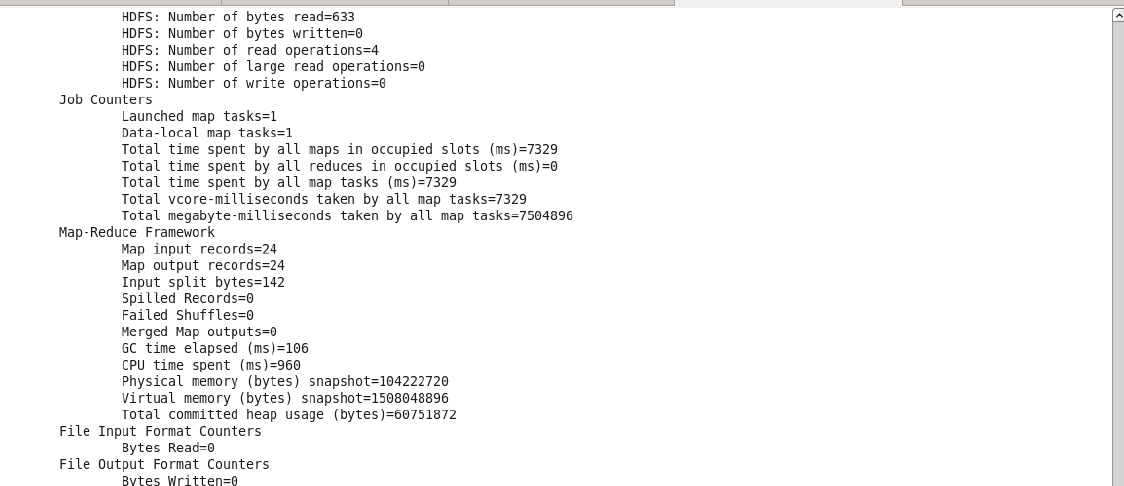
1. Create table in mysql



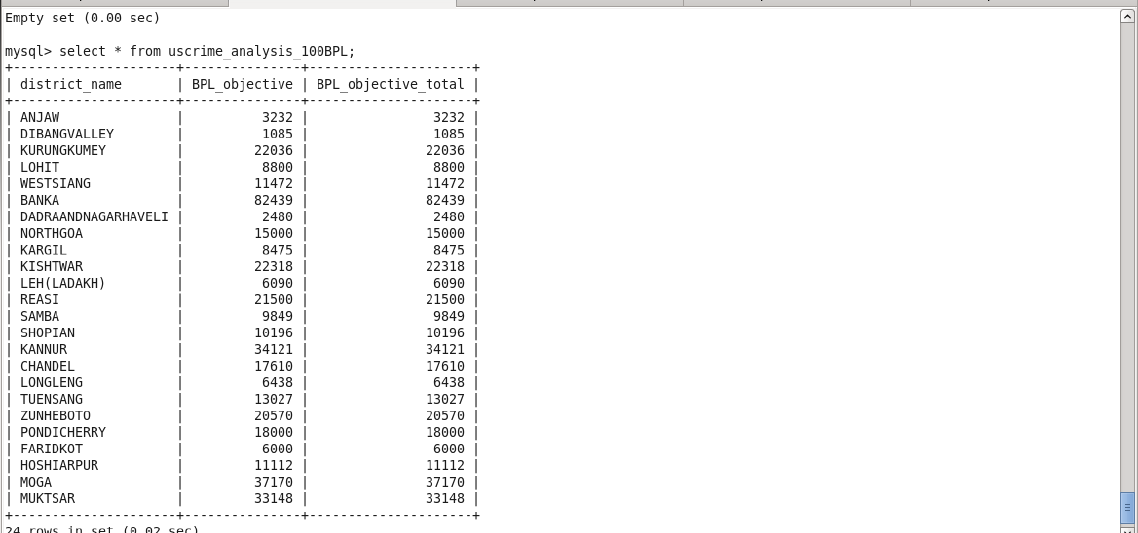
1. Scoop command to export the data from HDFS to mysql



1. once the job completed, check for the success command:



1. Check the exported data in mysql select statement:



2.Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL

cards.

Pig script:

REGISTER '/home/cloudera/Downloads/jarfiles/piggybank-0.17.0.jar'

DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();

A = LOAD '/flume\_import/StatewiseDistrictwisePhysicalProgress.xml' using

org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

B = FOREACH A GENERATE XPath(x,'row/District\_Name') as district ,XPath(x,'row/

Project\_Objectives\_IHHL\_BPL') as BPL\_Objective, XPath(x,'row/

Project\_Objectives\_IHHL\_TOTAL') as BPL\_Objective\_total;

C = filter B by (((int)BPL\_Objective \* 100)/(int)BPL\_Objective\_total) >= 80;

STORE C INTO 'hdfs://quickstart.cloudera:8020/uscrime\_analysis\_2' USING

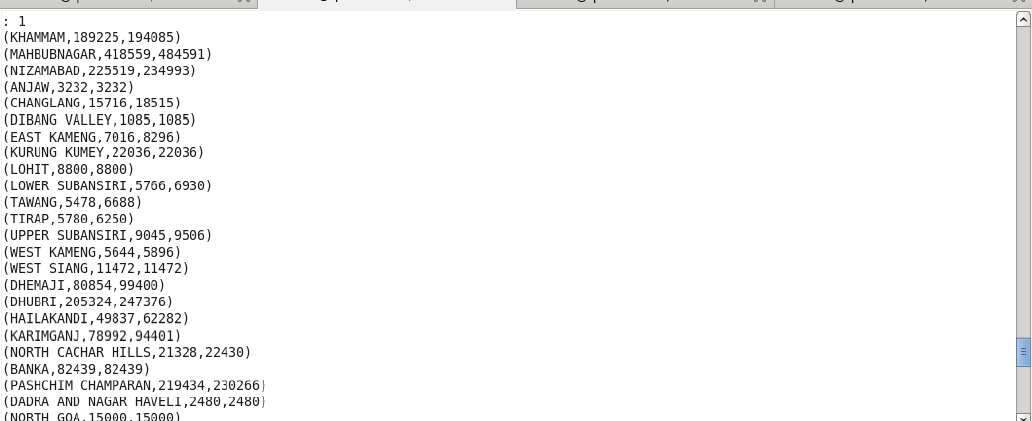
PigStorage (',');

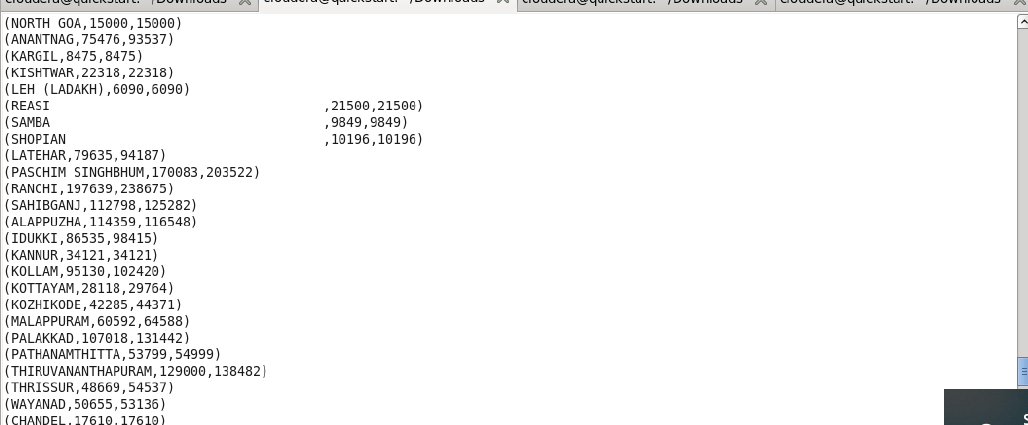
dump C;

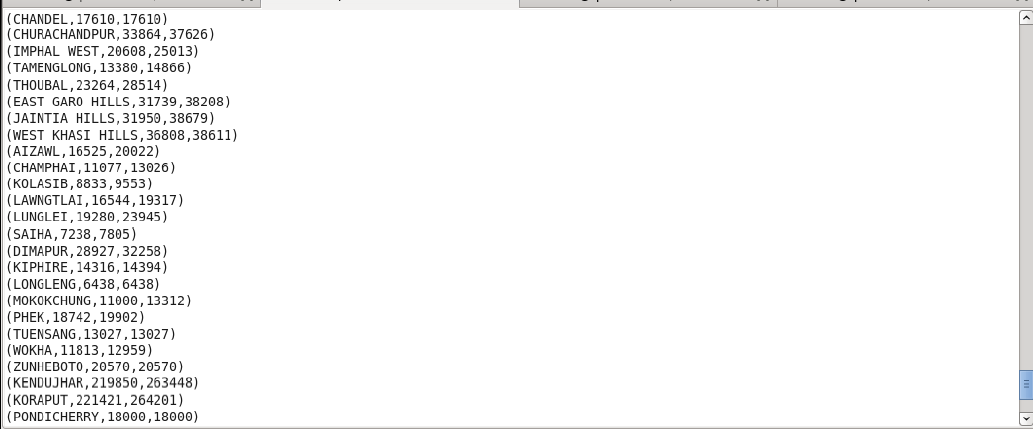
Execution:

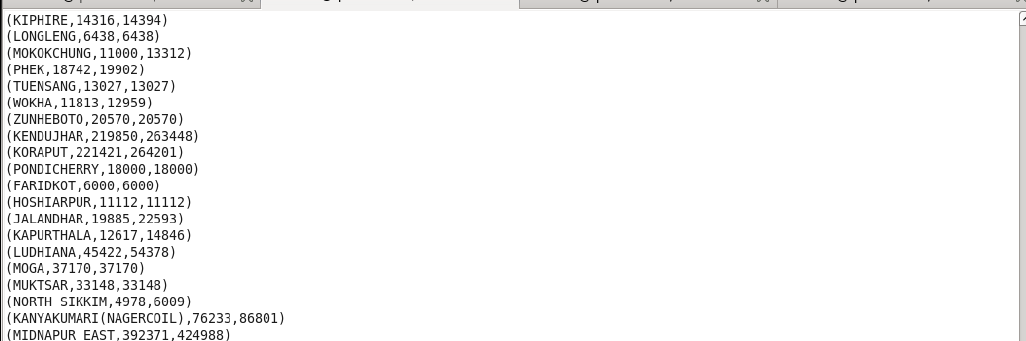
Pig <pig\_script\_filename>

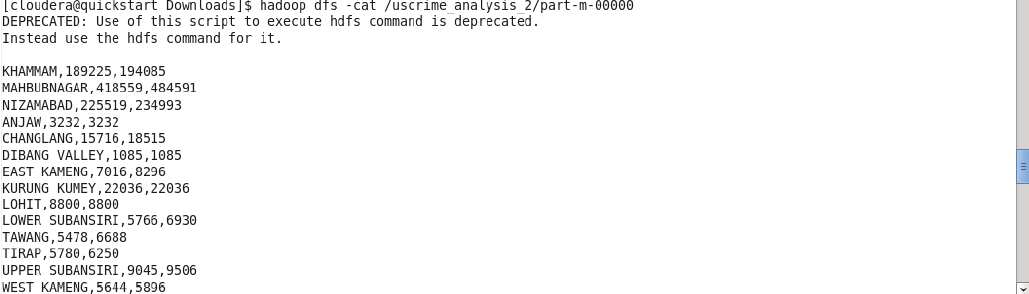
Output:





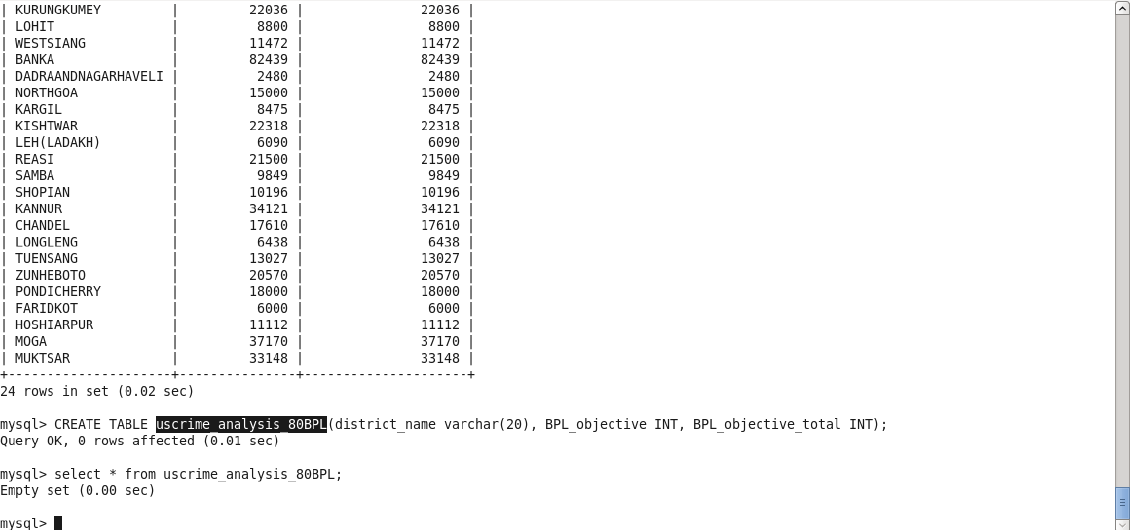




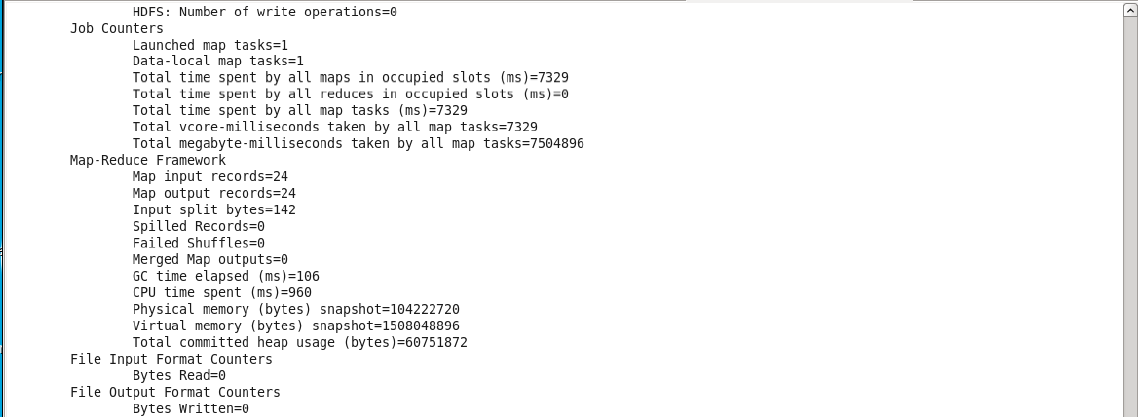


Export the results to mysql using sqoop:

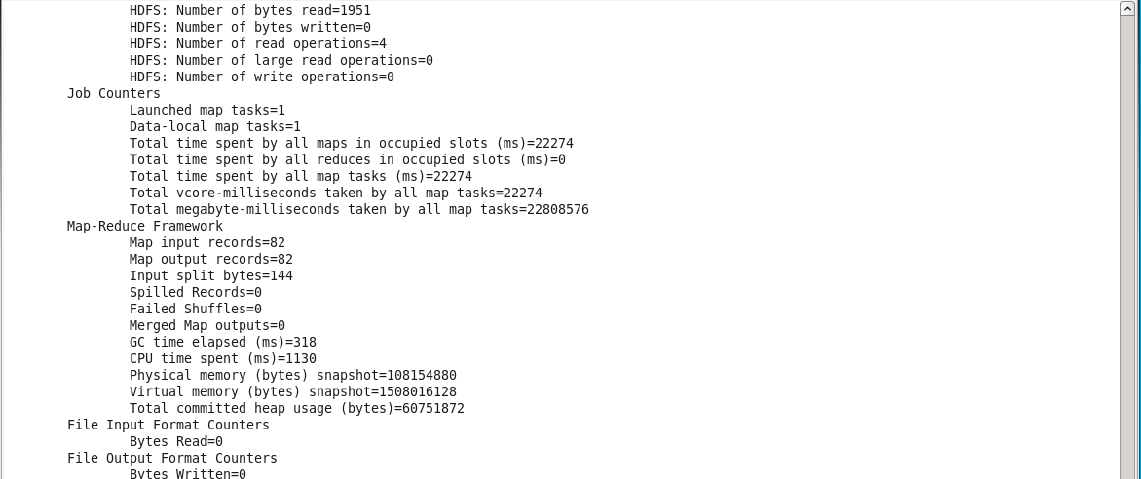
1. Create table in mysql to store the data:



1. Using sqoop export data from HDFS to mysql using command:



1. Check the success message for the job completion:



1. Check the data in mysql table using command; 82 rows copied successfully

