**This document contains the Solution Set for all the requirements and steps executed for processing the result along with Screenshots**

**Requirements**

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

**- Pig job for parsing the XML data.**

**- Create Pig scripts jobs to analyze the data**

**- Create the Sqoop job to store the data in database**

**First we will create the Hbase table where we will store our processing results**

create 'Statewise\_Hundred\_percent\_bpl\_obj','state\_data';

create 'Statewise\_Eighty\_percent\_bpl\_obj','state\_data';

**Create Mysql table which will be loaded from hdfs location using sqoop job**

create database stateDevelopment;

use stateDevelopment;

// **HunderedPercentBPLObjective**

create table **HunderedPercentBPLObjective**(

StateName VARCHAR(40),

DistrictName VARCHAR(40),

ObjectivesIHHLBPL INT,

ObjectivesIHHLAPL INT,

ObjectivesIHHLTOTAL INT,

ObjectivesSCW INT,

ObjectivesSchoolToilets INT,

ObjectivesAnganwadiToilets INT,

ObjectivesRSM INT,

ObjectivesPC INT,

PerformanceIHHLBPL INT,

PerformanceIHHLAPL INT,

PerformanceIHHLTOTAL INT,

PerformanceSCW INT,

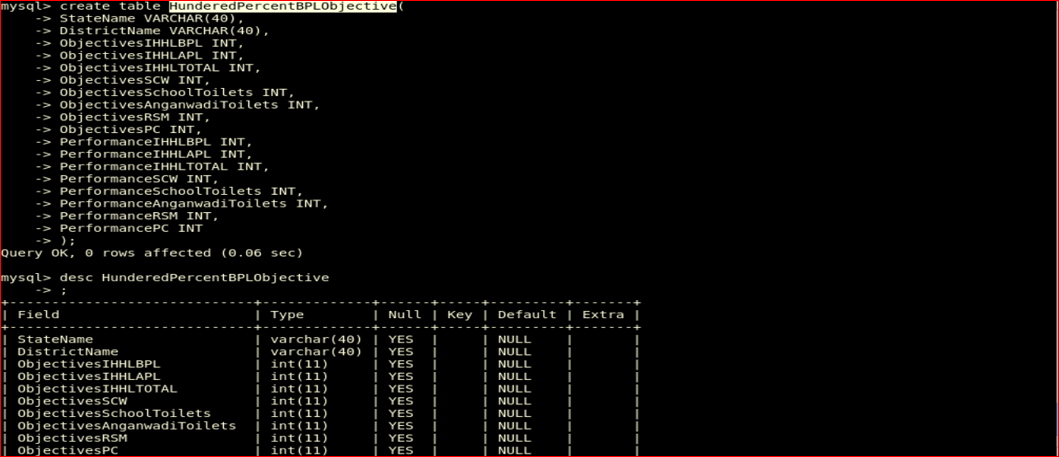
PerformanceSchoolToilets INT,

PerformanceAnganwadiToilets INT,

PerformanceRSM INT,

PerformancePC INT

);



// **EightyPercentBPLObjective**

create table **EightyPercentBPLObjective**(

StateName VARCHAR(40),

DistrictName VARCHAR(40),

ObjectivesIHHLBPL INT,

ObjectivesIHHLAPL INT,

ObjectivesIHHLTOTAL INT,

ObjectivesSCW INT,

ObjectivesSchoolToilets INT,

ObjectivesAnganwadiToilets INT,

ObjectivesRSM INT,

ObjectivesPC INT,

PerformanceIHHLBPL INT,

PerformanceIHHLAPL INT,

PerformanceIHHLTOTAL INT,

PerformanceSCW INT,

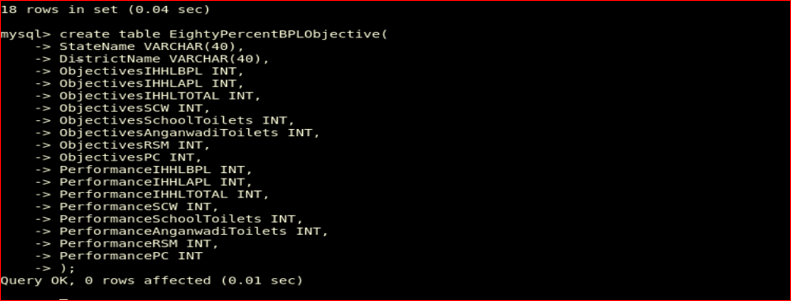
PerformanceSchoolToilets INT,

PerformanceAnganwadiToilets INT,

PerformanceRSM INT,

PerformancePC INT

);



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

**The following flume configuration file will be used to load data from local system to hdfs file system**

# flumeCopy.conf

agent1.sources = mysrc

agent1.sinks = hdfsdest

agent1.channels = mychannel

agent1.sources.mysrc.type = exec

agent1.sources.mysrc.command = hadoop dfs -put /home/acadgild/HadoopProject\_2/StatewiseDistrictwisePhysicalProgress.xml /flume\_import

agent1.sinks.hdfsdest.type = hdfs

agent1.sinks.hdfsdest.hdfs.path = hdfs://localhost:9000/flume\_import

agent1.channels.mychannel.type = memory

agent1.sources.mysrc.channels = mychannel

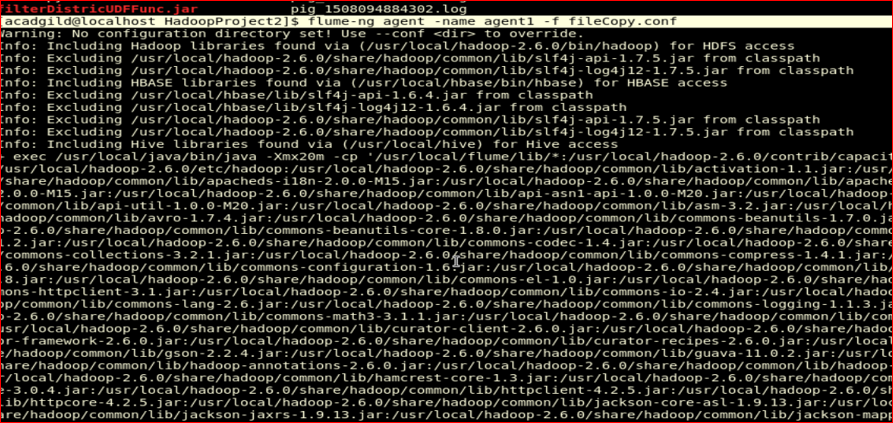
agent1.sinks.hdfsdest.channel = mychannel

# Run the following command to start the flume job

flume-ng agent –n agent1 –f /home/acadgild/HadoopProject\_2filecopy.conf

This Job will load the file from source destination home/acadgild/HadoopProject\_2/ to hdfs target destination /flume\_import

The type of channel used for this job is memory, we can also choose it to be file depending on the performance requirement of the project



**Pig/MapReduce job for parsing the XML data.**

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

# District\_Hundred\_Percent\_BPL.pig

**#Register Jar required for using XML Loader**

REGISTER /home/acadgild/HadoopProject2/piggybank.jar;

#Register Jar required to load data into Hbase Table

REGISTER /usr/local/hbase/lib/hbase-common-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/hbase-client-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/hbase-server-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/hbase-protocol-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/htrace-core-2.04.jar;

REGISTER /usr/local/hbase/lib/zookeeper-3.4.6.jar;

REGISTER /usr/local/hbase/lib/guava-12.0.1.jar;

**#Deine XPath function to use for parsing xml files**

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

**#Load the data from xml using XML loader**

row = LOAD '/home/acadgild/HadoopProject2/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (rowElement:chararray);

**#Get all the xml elements into columns which would get stored into relation #elements**

elements = FOREACH row GENERATE

XPath(rowElement, 'row/State\_Name') AS State\_Name,

XPath(rowElement, 'row/District\_Name') AS District\_Name,

XPath(rowElement,'row/Project\_Objectives\_IHHL\_BPL') AS Project\_Objectives\_IHHL\_BPL,

XPath(rowElement,'row/Project\_Objectives\_IHHL\_APL') AS Project\_Objectives\_IHHL\_APL,

XPath(rowElement,'row/Project\_Objectives\_IHHL\_TOTAL') AS Project\_Objectives\_IHHL\_TOTAL,

XPath(rowElement,'row/Project\_Objectives\_SCW') AS Project\_Objectives\_SCW,

XPath(rowElement,'row/Project\_Objectives\_School\_Toilets') AS Project\_Objectives\_School\_Toilets,

XPath(rowElement,'row/Project\_Objectives\_Anganwadi\_Toilets') AS Project\_Objectives\_Anganwadi\_Toilets,

XPath(rowElement,'row/Project\_Objectives\_RSM') AS Project\_Objectives\_RSM,

XPath(rowElement,'row/Project\_Objectives\_PC') AS Project\_Objectives\_PC,

XPath(rowElement,'row/Project\_Performance-IHHL\_BPL') AS Project\_Performance\_IHHL\_BPL,

XPath(rowElement,'row/Project\_Performance-IHHL\_APL') AS Project\_Performance\_IHHL\_APL,

XPath(rowElement,'row/Project\_Performance-IHHL\_TOTAL') AS Project\_Performance\_IHHL\_TOTAL,

XPath(rowElement,'row/Project\_Performance-SCW') AS Project\_Performance\_SCW,

XPath(rowElement,'row/Project\_Performance-School\_Toilets') AS Project\_Performance\_School\_Toilets,

XPath(rowElement,'row/Project\_Performance-Anganwadi\_Toilets') AS Project\_Performance\_Anganwadi\_Toilets,

XPath(rowElement,'row/Project\_Performance-RSM') AS Project\_Performance\_RSM,

XPath(rowElement,'row/Project\_Performance-PC') AS Project\_Performance\_PC ;

**#Filter the districts which have reached 100percent objective in BPL**

filterDistricts = FILTER elements BY TRIM(Project\_Objectives\_IHHL\_BPL) == TRIM(Project\_Performance\_IHHL\_BPL);

**#Store the result into hbase table Statewise\_Hundred\_percent\_bpl\_obj**

STORE filterDistricts INTO 'hbase://Statewise\_Hundred\_percent\_bpl\_obj'

USING org.apache.pig.backend.hadoop.hbase.HBaseStorage('state\_data:State\_Name,state\_data:District\_Name,state\_data:Project\_Objectives\_IHHL\_BPL,

state\_data:Project\_Objectives\_IHHL\_APL,state\_data:Project\_Objectives\_IHHL\_TOTAL,state\_data:Project\_Objectives\_SCW,

state\_data:Project\_Objectives\_School\_Toilets,state\_data:Project\_Objectives\_Anganwadi\_Toilets,state\_data:Project\_Objectives\_RSM,

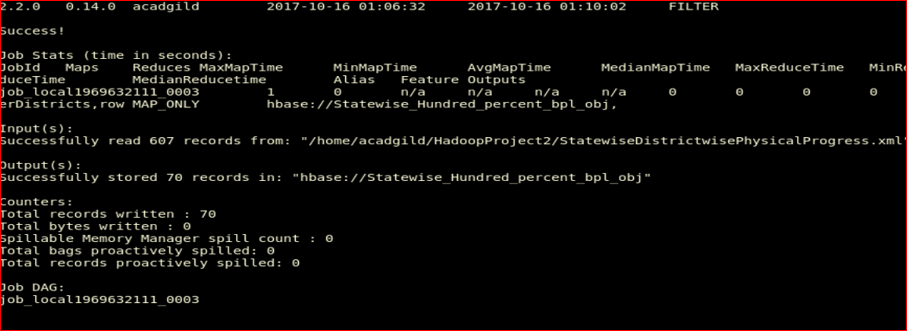
state\_data:Project\_Objectives\_PC,state\_data:Project\_Performance\_IHHL\_BPL,state\_data:Project\_Performance\_IHHL\_APL,

state\_data:Project\_Performance\_IHHL\_TOTAL,state\_data:Project\_Performance\_SCW,state\_data:Project\_Performance\_School\_Toilets,

state\_data:Project\_Performance\_Anganwadi\_Toilets,state\_data:Project\_Performance\_RSM,state\_data:Project\_Performance\_PC');

**# Store the result into hdfs location from where we can export the results # to mysql using sqoop job**

STORE filterDistricts into 'hdfs://localhost:9000/bpl100percentOutput' using PigStorage(',');



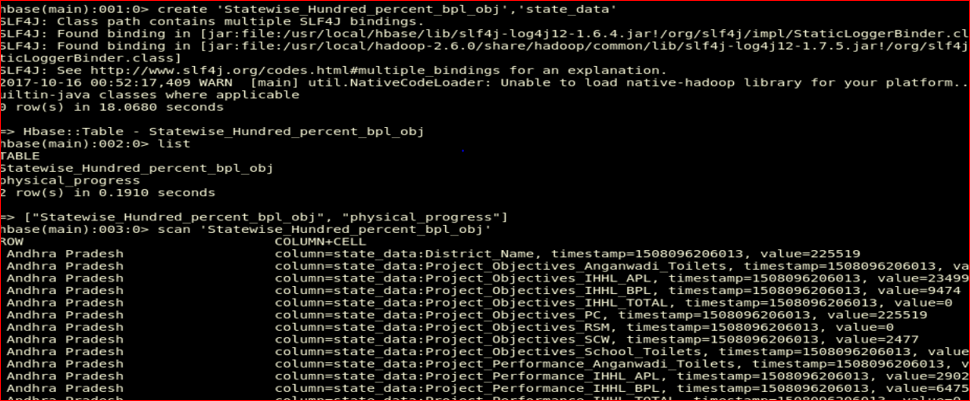
# Data in HDFS location:



# Data in Hbase Table

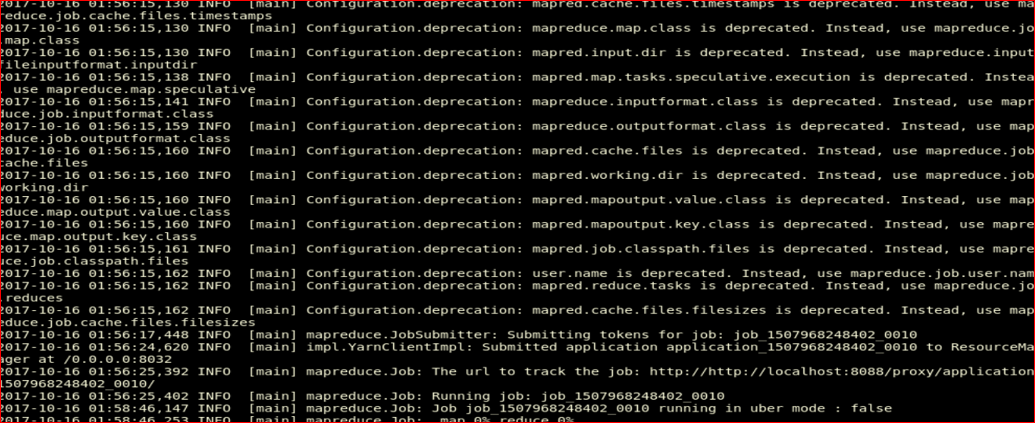
scan 'Statewise\_Hundred\_percent\_bpl\_obj'

**OutPut**



# Sqoop Job to export result to Mysql table HunderedPercentBPLObjective from hdfs location:

sqoop export --connect jdbc:mysql://localhost/stateDevelopment --username 'root' -P --table 'HunderedPercentBPLObjective' --export-dir 'hdfs://localhost:9000/bpl100percentOutput/' --input-fields-terminated-by ',' -m 1 --columns \ StateName,DistrictName,ObjectivesIHHLBPL,ObjectivesIHHLAPL,ObjectivesIHHLTOTAL,ObjectivesSCW,ObjectivesSchoolToilets,ObjectivesAnganwadiToilets,ObjectivesRSM,ObjectivesPC,PerformanceIHHLBPL,PerformanceIHHLAPL,PerformanceIHHLTOTAL,PerformanceSCW,PerformanceSchoolToilets,PerformanceAnganwadiToilets,PerformanceRSM,PerformancePC



## Exported data in mysql table HunderedPercentBPLObjective

select \* from HunderedPercentBPLObjective;



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

# District\_Eighty\_Percent\_BPL.pig

**#Register exported jar which contains UDF for filtering Eighty percent bpl #Objectives**

REGISTER /home/acadgild/HadoopProject2/filterUDF.jar;

**#Register Jar required for using XML Loader**

REGISTER /home/acadgild/HadoopProject2/piggybank.jar;

**#Register Jar required to load data into Hbase Table**

REGISTER /usr/local/hbase/lib/hbase-common-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/hbase-client-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/hbase-server-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/hbase-protocol-0.98.14-hadoop2.jar;

REGISTER /usr/local/hbase/lib/htrace-core-2.04.jar;

REGISTER /usr/local/hbase/lib/zookeeper-3.4.6.jar;

REGISTER /usr/local/hbase/lib/guava-12.0.1.jar;

**#Define XPath function to use for parsing xml files**

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

**#Load the data from xml using XML loader**

row = LOAD '/home/acadgild/HadoopProject2/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (rowElement:chararray);

**#Get all the xml elements into columns which would get stored into relation #elements**

elements = FOREACH row GENERATE

XPath(rowElement, 'row/State\_Name') AS State\_Name,

XPath(rowElement, 'row/District\_Name') AS District\_Name,

XPath(rowElement,'row/Project\_Objectives\_IHHL\_BPL') AS Project\_Objectives\_IHHL\_BPL,

XPath(rowElement,'row/Project\_Objectives\_IHHL\_APL') AS Project\_Objectives\_IHHL\_APL,

XPath(rowElement,'row/Project\_Objectives\_IHHL\_TOTAL') AS Project\_Objectives\_IHHL\_TOTAL,

XPath(rowElement,'row/Project\_Objectives\_SCW') AS Project\_Objectives\_SCW,

XPath(rowElement,'row/Project\_Objectives\_School\_Toilets') AS Project\_Objectives\_School\_Toilets,

XPath(rowElement,'row/Project\_Objectives\_Anganwadi\_Toilets') AS Project\_Objectives\_Anganwadi\_Toilets,

XPath(rowElement,'row/Project\_Objectives\_RSM') AS Project\_Objectives\_RSM,

XPath(rowElement,'row/Project\_Objectives\_PC') AS Project\_Objectives\_PC,

XPath(rowElement,'row/Project\_Performance-IHHL\_BPL') AS Project\_Performance\_IHHL\_BPL,

XPath(rowElement,'row/Project\_Performance-IHHL\_APL') AS Project\_Performance\_IHHL\_APL,

XPath(rowElement,'row/Project\_Performance-IHHL\_TOTAL') AS Project\_Performance\_IHHL\_TOTAL,

XPath(rowElement,'row/Project\_Performance-SCW') AS Project\_Performance\_SCW,

XPath(rowElement,'row/Project\_Performance-School\_Toilets') AS Project\_Performance\_School\_Toilets,

XPath(rowElement,'row/Project\_Performance-Anganwadi\_Toilets') AS Project\_Performance\_Anganwadi\_Toilets,

XPath(rowElement,'row/Project\_Performance-RSM') AS Project\_Performance\_RSM,

XPath(rowElement,'row/Project\_Performance-PC') AS Project\_Performance\_PC ;

**#Use UDF created to filter the districts which have reached 80% of objectives of BPL cards.**

filterEightyPercent = FILTER elements by FilterDistrictUdf.FilterDistrict(\*)

**#Store the result into hbase table Statewise\_Eighty\_percent\_bpl\_obj**

STORE filterEightyPercent INTO 'hbase://Statewise\_Eighty\_percent\_bpl\_obj'

USING org.apache.pig.backend.hadoop.hbase.HBaseStorage('state\_data:State\_Name,state\_data:District\_Name,state\_data:Project\_Objectives\_IHHL\_BPL,

state\_data:Project\_Objectives\_IHHL\_APL,state\_data:Project\_Objectives\_IHHL\_TOTAL,state\_data:Project\_Objectives\_SCW,

state\_data:Project\_Objectives\_School\_Toilets,state\_data:Project\_Objectives\_Anganwadi\_Toilets,state\_data:Project\_Objectives\_RSM,

state\_data:Project\_Objectives\_PC,state\_data:Project\_Performance\_IHHL\_BPL,state\_data:Project\_Performance\_IHHL\_APL,

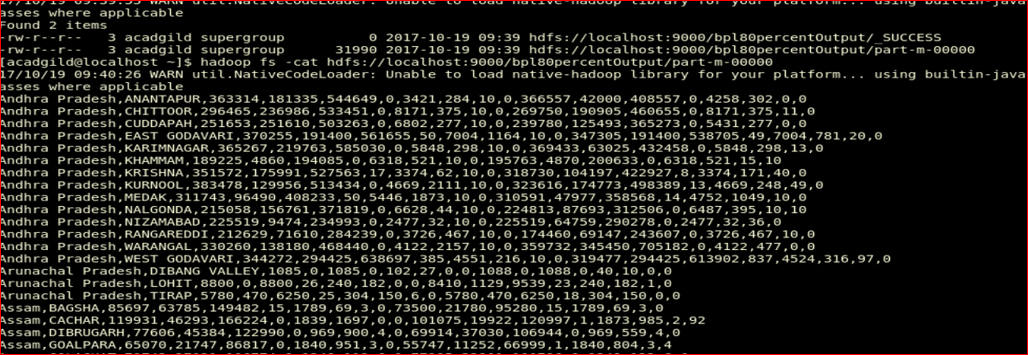
state\_data:Project\_Performance\_IHHL\_TOTAL,state\_data:Project\_Performance\_SCW,state\_data:Project\_Performance\_School\_Toilets,

state\_data:Project\_Performance\_Anganwadi\_Toilets,state\_data:Project\_Performance\_RSM,state\_data:Project\_Performance\_PC');

**# Store the result into hdfs location from where we can export the results to #mysql using sqoop job**

STORE filterEightyPercent into 'hdfs://localhost:9000/bpl80percentOutput' using PigStorage(',');

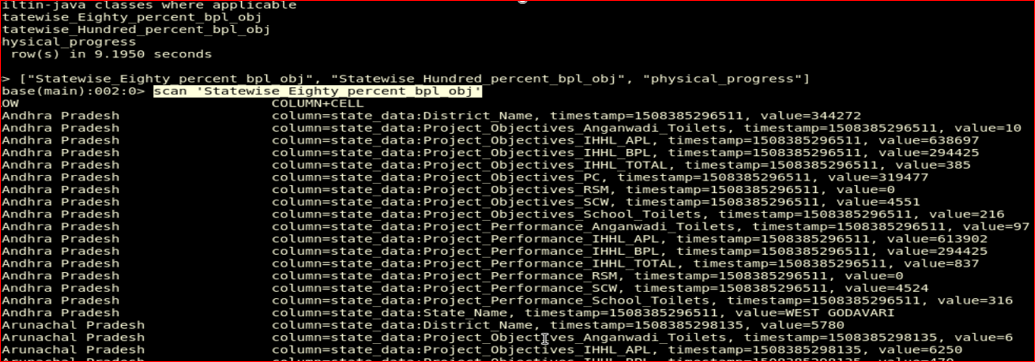
# Data in HDFS location:



# Data in Hbase Table

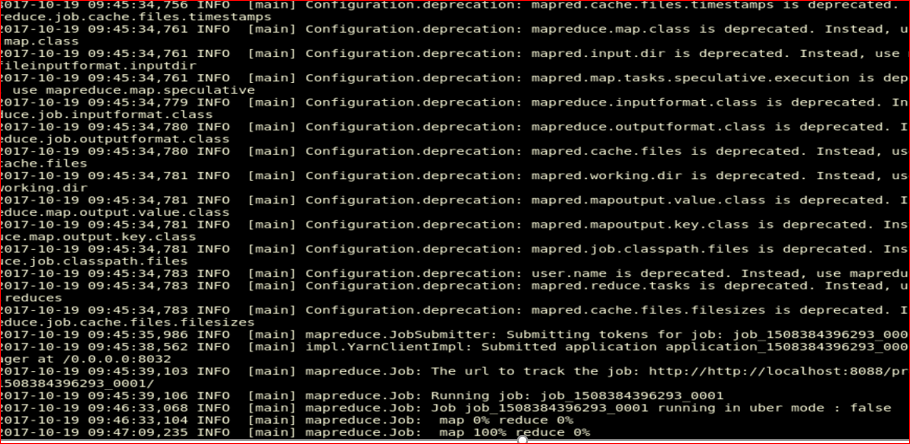
scan 'Statewise\_Hundred\_percent\_bpl\_obj'

**OutPut**



# Sqoop Job to export result to Mysql table EightyPercentBPLObjective from hdfs location:

sqoop export --connect jdbc:mysql://localhost/stateDevelopment --username 'root' -P --table 'EightyPercentBPLObjective' --export-dir 'hdfs://localhost:9000/bpl80percentOutput/' --input-fields-terminated-by ',' -m 1 --columns \ StateName,DistrictName,ObjectivesIHHLBPL,ObjectivesIHHLAPL,ObjectivesIHHLTOTAL,ObjectivesSCW,ObjectivesSchoolToilets,ObjectivesAnganwadiToilets,ObjectivesRSM,ObjectivesPC,PerformanceIHHLBPL,PerformanceIHHLAPL,PerformanceIHHLTOTAL,PerformanceSCW,PerformanceSchoolToilets,PerformanceAnganwadiToilets,PerformanceRSM,PerformancePC



# Exported data in mysql table EightyPercentBPLObjective

select \* from EightyHunderedPercentBPLObjective;

