# Project 2.1

**Step 1:**

Copy dataset from local file system to HDFS using flume.

***Flume.conf***

*agent.sources = localdir*

*agent.channels = channel\_f*

*agent.sinks = HDFS*

*# For each one of the sources, the type is defined*

*agent.sources.localdir.type = spooldir*

*# The channel can be defined as follows.*

*agent.sources.localdir.channels = channel\_f*

*agent.sources.localdir.spoolDir = /home/flumedata/flumespool*

*agent.sources.localdir.fileHeader = false*

*agent.sources.localdir.fileSuffix = .COMPLETED*

*# Each sink's type must be defined*

*agent.sinks.HDFS.type = hdfs*

*#Specify the channel the sink should use*

*agent.sinks.HDFS.channel = channel\_f*

*agent.sinks.HDFS.hdfs.path = hdfs://localhost/user/ec2-user/flume/logdata/*

*agent.sinks.HDFS.hdfs.fileType = DataStream*

*agent.sinks.HDFS.hdfs.writeFormat = Text*

*agent.sinks.HDFS.hdfs.batchSize = 1000*

*agent.sinks.HDFS.hdfs.rollSize = 268435456*

*agent.sinks.HDFS.hdfs.rollInterval = 0*

*agent.sinks.HDFS.hdfs.rollCount = 50000000*

*# Each channel's type is defined.*

*agent.channels.channel\_f.type = file*

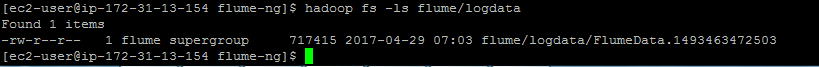
*# Other config values specific to each type of channel(sink or source)*

*# can be defined as well*

*# In this case, it specifies the capacity of the memory channel*

*agent.channels.memoryChannel.capacity = 10000*

*agent.channels.memoryChannel.transactionCapacity = 100*



**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

**Register piggybank to use xml loader**

REGISTER '/home/ec2-user/Installations/pig/piggybank-0.15.0.jar';

**Load the xml file to pig using xml loader from piggybank**

B = LOAD '/user/ec2-user/flume/logdata/FlumeData.1493463472503' using org.apache.pig.piggybank.storage.XMLLoader('row') as (xmlrows:chararray);

**Make data suitable for analysis**

data = FOREACH B GENERATE FLATTEN(REGEX\_EXTRACT\_ALL(xmlrows,'<row>\\s\*<State\_Name>(.\*)</State\_Name>\\s\*<District\_Name>(.\*)</District\_Name>\\s\*<Project\_Objectives\_IHHL\_BPL>(.\*)</Project\_Objectives\_IHHL\_BPL>\\s\*<Project\_Objectives\_IHHL\_APL>(.\*)</Project\_Objectives\_IHHL\_APL>\\s\*<Project\_Objectives\_IHHL\_TOTAL>(.\*)</Project\_Objectives\_IHHL\_TOTAL>\\s\*<Project\_Objectives\_SCW>(.\*)</Project\_Objectives\_SCW>\\s\*<Project\_Objectives\_School\_Toilets>(.\*)</Project\_Objectives\_School\_Toilets>\\s\*<Project\_Objectives\_Anganwadi\_Toilets>(.\*)</Project\_Objectives\_Anganwadi\_Toilets>\\s\*<Project\_Objectives\_RSM>(.\*)</Project\_Objectives\_RSM>\\s\*<Project\_Objectives\_PC>(.\*)</Project\_Objectives\_PC>\\s\*<Project\_Performance-IHHL\_BPL>(.\*)</Project\_Performance-IHHL\_BPL>\\s\*<Project\_Performance-IHHL\_APL>(.\*)</Project\_Performance-IHHL\_APL>\\s\*<Project\_Performance-IHHL\_TOTAL>(.\*)</Project\_Performance-IHHL\_TOTAL>\\s\*<Project\_Performance-SCW>(.\*)</Project\_Performance-SCW>\\s\*<Project\_Performance-School\_Toilets>(.\*)</Project\_Performance-School\_Toilets>\\s\*<Project\_Performance-Anganwadi\_Toilets>(.\*)</Project\_Performance-Anganwadi\_Toilets>\\s\*<Project\_Performance-RSM>(.\*)</Project\_Performance-RSM>\\s\*<Project\_Performance-PC>(.\*)</Project\_Performance-PC>\\s\*</row>')) AS (state\_name:chararray, district:chararray,po\_ihhl\_bpl:INT, po\_ihhl\_apl:INT, po\_ihhl\_total:INT,po\_scw:INT,po\_schl\_toilets:INT,po\_anganwadi\_toilets:INT,po\_rsm:INT,po\_pc:INT,pp\_ihhl\_bpl:INT,pp\_ihhl\_apl:INT,pp\_ihhl\_total:INT,pp\_scw:INT,pp\_schl\_toilets:INT,pp\_anganwadi\_toilets:INT,pp\_rsm:INT,pp\_pc:INT);

**store the data file in hdfs**

STORE data INTO '/user/ec2-user/india\_state\_wise\_progress/' USING PigStorage(',');

**Problem statement 1:**

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

Export the results to mysql using sqoop

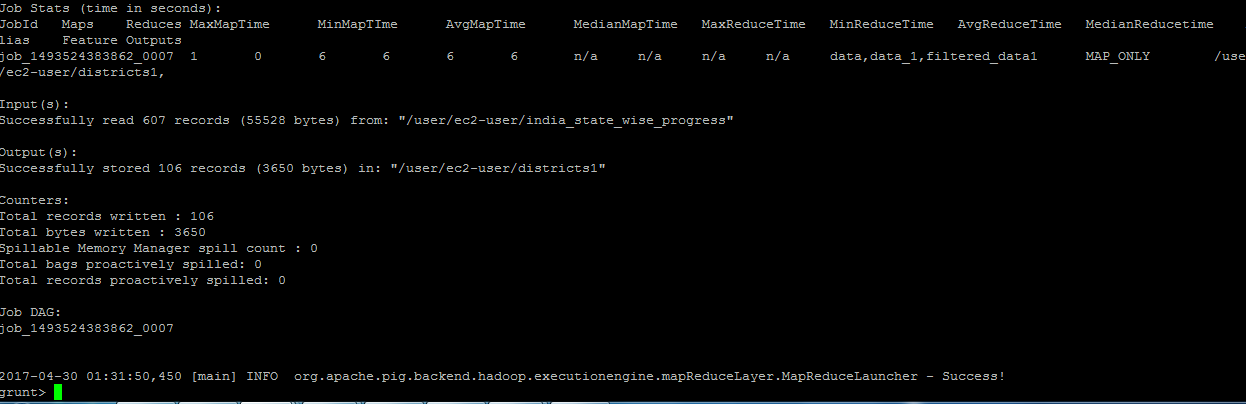
**Calculating performance indicator greater than objectives indicator (Project\_Performance-IHHL\_BPL > Project\_Objectives\_IHHL\_BPL)**

data\_1 = foreach data generate state\_name,district,po\_ihhl\_bpl,pp\_ihhl\_bpl;

filtered\_data1 = FILTER data\_1 by pp\_ihhl\_bpl>po\_ihhl\_bpl;

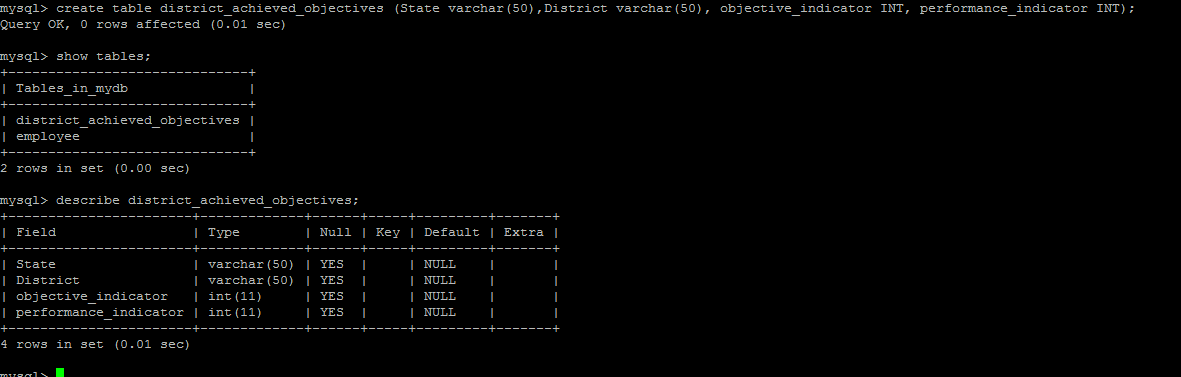
**storing the result to hdfs to use in sqoop export**

store filtered\_data1 into '/user/ec2-user/districts1/' USING PigStorage(',');



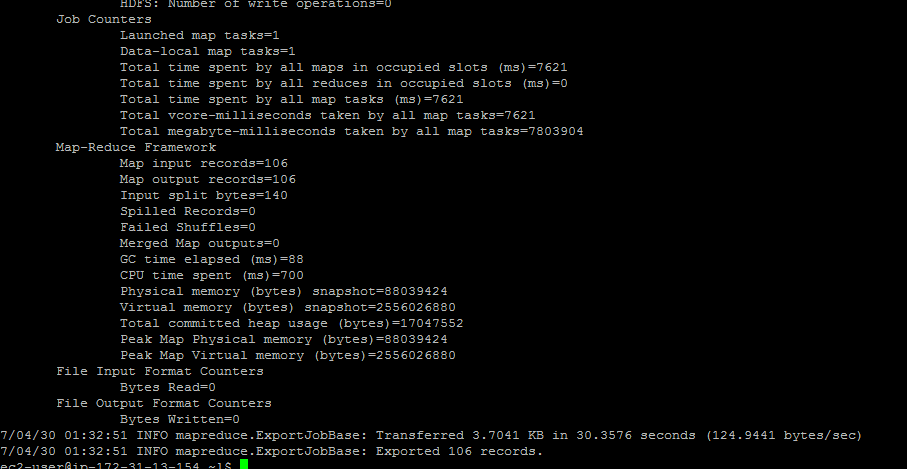
**Create table in mysql to export the result**

create table district\_achieved\_objectives (State varchar(50),District varchar(50), objective\_indicator INT, performance\_indicator INT);

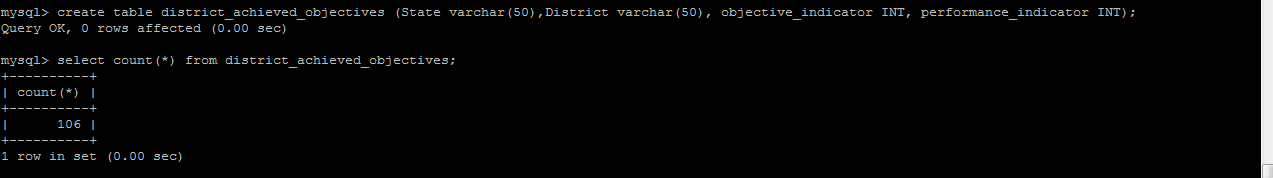


**Sqoop export to Mysql DB**

sqoop export --connect jdbc:mysql://54.233.145.80:3306/mydb --username sqoop --password Sqoop\_123 --table district\_achieved\_objectives --export-dir '/user/ec2-user/districts1' --input-fields-terminated-by ',' -m 1 --columns State,District,objective\_indicator,performance\_indicator



**Exported Data in mysql**

****

**Problem statement 2:**

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

Export the results to mysql using sqoop.

pigudf2.java is attached in the project folder

**Registering the udfjar**

REGISTER '/home/ec2-user/Installations/pig/pigudf2.jar';

**Defining alias**

DEFINE PigUdf2 hadoop.pig.PigUdf2();

**Taking previously loaded data from hdfs**

data = load '/user/ec2-user/india\_state\_wise\_progress/' USING PigStorage(',') as (state\_name:chararray,district:chararray,po\_ihhl\_bpl:INT,po\_ihhl\_apl:INT,po\_ihhl\_total:INT,po\_scw:INT,po\_schl\_toilets:INT,po\_anganwadi\_toilets:INT,po\_rsm:INT,po\_pc:INT,pp\_ihhl\_bpl:INT,pp\_ihhl\_apl:INT,pp\_ihhl\_total:INT,pp\_scw:INT,pp\_schl\_toilets:INT,pp\_anganwadi\_toilets:INT,pp\_rsm:INT,pp\_pc:INT);

**taking only the data required for calculation**

data\_1 = foreach data generate state\_name,district,po\_ihhl\_bpl,pp\_ihhl\_bpl;

using udf function to fetch data in which (**Project\_Performance-IHHL\_BPL /Project\_Objectives\_IHHL\_BPL) > 0.8 >>achieving more than 80 % of objective.**

data\_2 = foreach data\_1 generate PigUdf2(state\_name,district,po\_ihhl\_bpl,pp\_ihhl\_bpl) as datafiltered;

**code snippet**

@Override

**public** Tuple exec(Tuple input) **throws** IOException {

**if** (input == **null** || input.size() == 0)

**return** **null**;

Integer objective = (Integer)input.get(2);

Integer performance = (Integer)input.get(3);

// System.out.println("objective > "+objective+" performance > "+performance);

**int** o = objective.intValue();

**int** p = performance.intValue();

**if**(o>0){

**double** percentage = p/o;

**if**(percentage>0.8){

**return** input;

}

}

**return** **null**;

}

**Removing null values**

data\_3 = filter data\_2 by datafiltered is not null;

**Storing data to hdfs to sqoop out to mysql**

store data\_3 into '/user/ec2-user/district\_80/' USING PigStorage(',');

**Mysql table create**

create table district\_eighty (State varchar(50),District varchar(50), objective\_indicator INT, performance\_indicator INT);

**Sqoop Export**

sqoop export --connect jdbc:mysql://54.233.145.80:3306/mydb --username sqoop --password Sqoop\_123 --table district\_achieved\_objectives --export-dir '/user/ec2-user/district\_80' --input-fields-terminated-by ',' -m 1 --columns State,District,objective\_indicator,performance\_indicator