**State-Wise Development Analysis In India**

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

Queries:

**Flume:**

flume-ng agent --conf-file /usr/local/flume/conf/flumeproject.conf --name agent -Dflume.root.logger=INFO,console

**Inside .conf file:**

agent1.sources = mysrc

agent1.sinks = hdfsdest

agent1.channels = mychannel

agent1.sources.mysrc.type = exec

agent1.sources.mysrc.command = hadoop dfs -put /home/gv/Desktop/Execute/flume/pro/State /user/flume/mainproject1

agent1.sinks.hdfsdest.type = hdfs

agent1.sinks.hdfsdest.hdfs.path = hdfs://localhost:9000/user/flume/mainproject1

agent1.channels.mychannel.type = memory

agent1.sources.mysrc.channels = mychannel

agent1.sinks.hdfsdest.channel = mychannel

**PIG:**

A = load '/user/flume/mainproject1/State' using org.apache.pig.piggybank.storage.XMLLoader('row') as (state:chararray);

**Parsing xml file:**

B = foreach A generate FLATTEN(REGEX\_EXTRACT\_ALL(state,'<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>'));

**Storing B:**

C = STORE B into '/user/hdusr/mainproject' using org.apache.pig.piggybank.storage.CSVExelStorage();

**Loading data to pig:**

D = load '/user/hdusr/mainproject' using PigStorage(',') as (State\_Name:chararray,District\_Name:chararray,Project\_Objectives\_IHHL\_BPL:int,Project\_Objectives\_IHHL\_APL:int,Project\_Objectives\_IHHL\_TOTAL:int,Project\_Objectives\_SCW:int,Project\_Objectives\_School\_Toilets:int,Project\_Objectives\_Anganwadi\_Toilets:int,Project\_Objectives\_RSM:int,Project\_Objectives\_PC:int,Project\_Performance\_IHHL\_BPL:int,Project\_Performance\_IHHL\_APL:int,Project\_Performance\_IHHL\_TOTAL:int,Project\_Performance\_SCW:int,Project\_Performance\_School\_Toilets:int,Project\_Performance\_Anganwadi\_Toilets:int,Project\_Performance\_RSM:int,Project\_Performance\_PC:int);

**Filtering:**

E = filter D by Project\_Objectives\_IHHL\_BPL == Project\_Performance\_IHHL\_BPL;

we can get output here it self but if you need district name then below is code

**Grouping:**

F = group E by District\_Name;

G = foreach F generate group; (this is to display only district names)

**Storing:**

store E into '/user/hdusr/mainproject/results2' using PigStorage(',');

**Creating table in SQL:**

create table mainproject(State\_Name varchar(50),District\_Name varchar(50),Project\_Objectives\_IHHL\_BPL int(20),Project\_Objectives\_IHHL\_APL int(20),Project\_Objectives\_IHHL\_TOTAL int(20),Project\_Objectives\_SCW int(20),Project\_Objectives\_School\_Toilets int(20),Project\_Objectives\_Anganwadi\_Toilets int(20),Project\_Objectives\_RSM int(20),Project\_Objectives\_PC int(20),Project\_Performance\_IHHL\_BPL int(20),Project\_Performance\_IHHL\_APL int(20),Project\_Performance\_IHHL\_TOTAL int(20),Project\_Performance\_SCW int(20),Project\_Performance\_School\_Toilets int(20),Project\_Performance\_Anganwadi\_Toilets int(20),Project\_Performance\_RSM int(20),Project\_Performance\_PC int(20));

**Exporting to SQL using sqoop:**

sqoop export --connect jdbc:mysql://localhost/gova --table mainproject --export-dir /user/hdusr/mainproject/results2/part-m-00000 --fields-terminated-by ' '--username root -P;

**checking output in SQL:**

select \* from mainproject;(hence the results are too many so cant take screen shot for whole data(total 70 records))

**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.**

Queries:

**UDF Program:**

**package** pig\_UDF;

**import** java.io.IOException;

**import** org.apache.pig.FilterFunc;

**import** org.apache.pig.data.Tuple;

**public** **class** Fil **extends** FilterFunc {

@Override

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

// **TODO** Auto-generated method stub

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

{

**return** **null**;

}

**try**{

//String str = (String) input.get(0);

**int** completed = (**int**) input.get(0);

**int** total = (**int**) input.get(1);

**int** percentage = (**int**) (completed/total \* 100.0f) ;

**if**(percentage > 80.0f)

{

**return** **true**;

}**else**{

**return** **false**;

}

}

**catch** (Exception e)

{

// **TODO**: handle exception

**throw** **new** IOException("caught excepion in fil uDf", e);

}

}

}

**Registering in pig:**

register /home/hdusr/Desktop/Exefile/filpertrywithelse.jar

**Filtering:**

H = filter D by pig\_UDF.Fil(Project\_Performance\_IHHL\_BPL,Project\_Objectives\_IHHL\_BPL);

**Foreach:**

I = foreach H generate (State\_Name,District\_Name,Project\_Objectives\_IHHL\_BPL,Project\_Performance\_IHHL\_BPL);

Dump I(here values are storing with () so remove them by below command)

**Flatten:**

J = foreach I generate FLATTEN($0);

**Storing:**

store J into '/user/hdusr/mainproject/udf\_result' using PigStorage(',');

**Creating table in SQL:**

Create table mainproject\_udf(State\_Name varchar(40),District\_Name varchar(40),Project\_Objectives\_IHHL\_BPL int(20),Project\_Performance\_IHHL\_BPL int(20));

**Exporting to SQL using sqoop:**

sqoop export --connect jdbc:mysql://localhost/gova --table mainproject\_udf --export-dir /user/hdusr/mianproject/udf\_result/part-m-00000 --fields-terminated-by ',' --username root -P

**checking output in SQL:**

select \* from mainproject\_udf;(hence the results are too many so cant take screen shot for whole data(total 176 records))