LOAD DATA INTO HDFS

-put /home/cloudera/StateWiseDistrictWisePhysicalProgress.xml /user/flume/statewise

EDIT CONF FILE IN FLUME

agent1.sources = mysource

agent1.sinks = HDFS

agent1.channels = mychannel

agent1.sources.mysource.type = statewise

agent1.sources.source1.StateWise = /home/cloudera/StatewiseDistrictwisePhysicalProgress.xml

agent1.sinks.HDFS.type = hdfs

agent1.sinks.HDFS.hdfs.path = /user/flume/statewise

agent1.sinks.HDFS.hdfs.fileType=DataStream

agent1.channels.mychannel.type=file

agent1.sources.mysource.channels = mychannel

agent1.sinks.HDFS.channel = mychannel

START FLUME

flume-ng agent --conf-file /usr/lib/flume-ng/conf/flume.conf --name agent1 --conf $FLUME\_HOME/conf -Dflume.root.logger=INFO,console

PIG grunt shell

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

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>'));

dump B; (it showz in tuple)

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

C = store B into '/user/flume/xml\_to\_csv' using org.apache.pig.piggybank.storage.CSVExcelStorage();

D = load '/user/flume/xml\_to\_csv' 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\_Obj

ectives\_SCW:int,Project\_Objectives\_School\_Toilets:int,Project\_Objectives\_Anga

nwadi\_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);

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

F = GROUP E by District\_Name;

G = foreach F generate group;

store E into '/user/flume/xml\_to\_csv/output' using PigStorage(',');

CREATING TABLE IN MYSQL

create table State\_wise(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 IN SQL USING SQOOP

sqoop export --connect jdbc:mysql://localhost/statewise --username 'root' --password 'cloudera' --table 'State\_wise' --export-dir /user/flume/xml\_to\_csv/part-m-00000 --input-fields-terminated-by ',' --m 4

CHECK OUTPUT IN SQL

select \* from Statewise;

**Question2. 80% objective of BPL**

**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 {**

**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)**

**{**

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

**}**

**}**

**}**

REGISTER /home/cloudera/Fil\_pigUDF.jar

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

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

J = foreach I generate FLATTEN($0);

STORE J INTO '/user/flume/xml\_to\_csv1/pigudf\_output' using PigStorage(',');

CREATING TABLE IN MYSQL

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

EXPORTING IN SQL USING SQOOP

sqoop export --connect jdbc:mysql://localhost/statewise --username 'root' --password 'cloudera' --table 'Statewise\_udf' --export-dir /user/flume/xml\_to\_csv/part-m-00000 --input-fields-terminated-by ',' --m 4

CHECK OUTPUT IN SQL

select \* from Statewise\_udf;