flume command to copy xml file "StatewiseDistrictwisePhysicalProgress.xml" in HDFS path /flume\_import/:

**flume-ng agent --conf-file /home/cloudera/Desktop/Shared/Project2/filecopy.conf --name agent1 -Dflume.root.logger=INFO,console**

Solution for problem 1:

-------------------------------------------------------------------------------------------------------------------

pig script to analyse and store output in /StateWiseAnalysis/output/problem1\_output:

**register '/usr/lib/pig/piggybank.jar'**

**xmldata = LOAD '/flume\_import/StatewiseDistrictwisePhysicalProgress.xml' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);**

**districtDetails = FOREACH xmldata GENERATE FLATTEN(REGEX\_EXTRACT\_ALL(x,'<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:chararray, district: chararray, p\_ihhl\_bpl: int, p\_ihhl\_apl:int, p\_ihhl\_total: int, p\_scw: int, p\_scl\_toilet: int, p\_angwadi\_toilet: int, p\_rsm:int , p\_pc:int, p\_perf\_ihhl\_bpl:int, p\_perf\_ihhl\_apl:int, p\_perf\_ihhl\_tot:int, p\_perf\_scw:int, p\_perf\_scl\_toilet:int, p\_perf\_angwadi\_toilet:int, p\_perf\_rsm:int, p\_perf\_pc:int);**

**districtBplPerf = foreach districtDetails generate district, p\_perf\_ihhl\_bpl, p\_ihhl\_bpl;**

**fullPerBPL = filter districtBplPerf by p\_perf\_ihhl\_bpl == p\_ihhl\_bpl;**

**districtsWith100PercPerf = foreach fullPerBPL generate district;**

**store districtsWith100PercPerf into '/StateWiseAnalysis/output/problem1\_output';**

Mysql table creation:

**create table districtsWithHundredPerBPL(district varchar(30));**

sqoop export to Mysql table "districtsWithHundredPerBPL" from HDFS path:

**sqoop export --connect jdbc:mysql://localhost/state\_performance --username root --password cloudera --table districtsWithHundredPerBPL -m 1 --export-dir /StateWiseAnalysis/output/problem1\_output/part-m-00000**

==================================================================================

Solution for problem 2:

-----------------------------------------------------------------------------------------------------------------

pig scripts to analyse and store output in "/StateWiseAnalysis/output/problem2\_output":

**register '/usr/lib/pig/piggybank.jar'**

**register '/home/cloudera/Desktop/Shared/Project2/check80Percentage.jar';**

**xmldata = LOAD '/flume\_import/StatewiseDistrictwisePhysicalProgress.xml' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);**

**districtDetails = FOREACH xmldata GENERATE FLATTEN(REGEX\_EXTRACT\_ALL(x,'<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:chararray, district: chararray, p\_ihhl\_bpl: int, p\_ihhl\_apl:int, p\_ihhl\_total: int, p\_scw: int, p\_scl\_toilet: int, p\_angwadi\_toilet: int, p\_rsm:int , p\_pc:int, p\_perf\_ihhl\_bpl:int, p\_perf\_ihhl\_apl:int, p\_perf\_ihhl\_tot:int, p\_perf\_scw:int, p\_perf\_scl\_toilet:int, p\_perf\_angwadi\_toilet:int, p\_perf\_rsm:int, p\_perf\_pc:int);**

**districtBplPerf = foreach districtDetails generate district, p\_ihhl\_bpl, p\_perf\_ihhl\_bpl;**

**find80PercBplPerformerDist = foreach districtBplPerf generate district, udf\_proj.CalculatePercentage(p\_ihhl\_bpl,p\_perf\_ihhl\_bpl) as achievedEighty;**

**filter80PercPerformer = filter find80PercBplPerformerDist by achievedEighty == true;**

**eightyPercPerfDistrict = foreach filter80PercPerformer generate district;**

**store eightyPercPerfDistrict into '/StateWiseAnalysis/output/problem2\_output'**

Mysql table creation:

**create table districtsWithEightyPerBPL(district varchar(30));**

sqoop export to Mysql table districtsWithEightyPerBPL from HDFS path:

**sqoop export --connect jdbc:mysql://localhost/state\_performance --username root --password cloudera --table districtsWithEightyPerBPL -m 1 --export-dir /StateWiseAnalysis/output/problem2\_output/part-m-00000**