**Project 2: Overview**

To develop the System to analyse the log data (In XML format) of government progress of various development activities.

1.2 **Purpose and Scope of this Specification**

The purpose of this project is to capture the data for analysing the progress of various activities.

**In scope**

The following requirement will be addressed in phase 1 of Project:

**Developing system to handle the incoming log feed and store the information in Hadoop Cluster (Flume)**

Analyse the data and understand the progress

**Store the results in Hbase/RDBMS**

*~~Out of scope~~*

*~~We can use this data and visualization and get more insights~~*

*~~2. Product/Service Description~~*

*~~2.1 Assumptions~~*

*~~Log will be generated in XML format and stored in a server~~*

*~~2.2 Constraints~~*

*~~Describe any item that will constrain the design options, including~~*

*~~ This system may not be used for searching for now. But it will be used for analysis and~~*

*~~saving the relevant information as of now~~*

*~~ System will be using Hbase as a database~~*

3. **Requirements**

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

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

**Create Pig scripts/MapReduce jobs to analyze the data**

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

Priority Definitions

The following definitions are intended as a guideline to prioritize requirements.

Priority 1 – Create FLUME job for fetching log files from spool directory the data

Priority 2 – MapReduce/pig job to pre-process

**Download the dataset using the below link:**

Link:

<https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing>

Refer the below steps to understand the actual steps to create the above project.

**Step 1:**

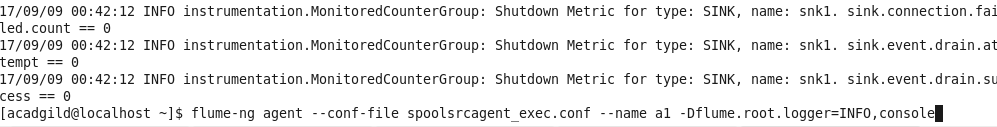
Copy dataset from local file system to HDFS using flume.

Note: use the conf file by downloading from below link.

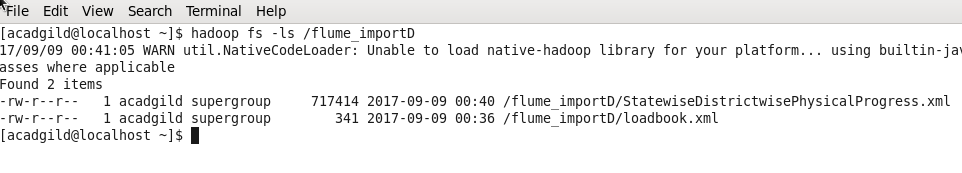
Click here to download

Solution:

* flume-ng agent --conf-file spoolsrcagent\_exec.conf --name a1 -Dflume.root.logger=INFO,console



* Hadoop fs –ls /flume\_importD



**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.

**4. Problem statement**

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

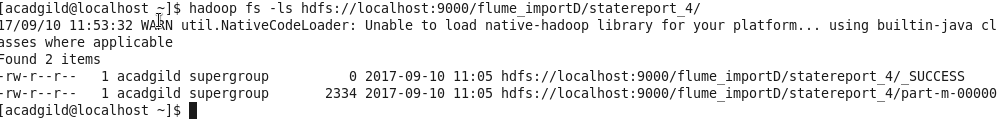
A = LOAD 'hdfs://localhost:9000/flume\_importD/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

B = foreach A 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>'));

D = foreach B generate $0 as state, $1 as dist,(int)$2 as obj\_bpl,(int)$10 as obj\_prf;

Ans1 = filter D by (obj\_bpl==obj\_prf);

STORE Ans1 into 'hdfs://localhost:9000/flume\_importD/statereport\_4' using PigStorage (',') ;



Export the results to mysql using sqoop:

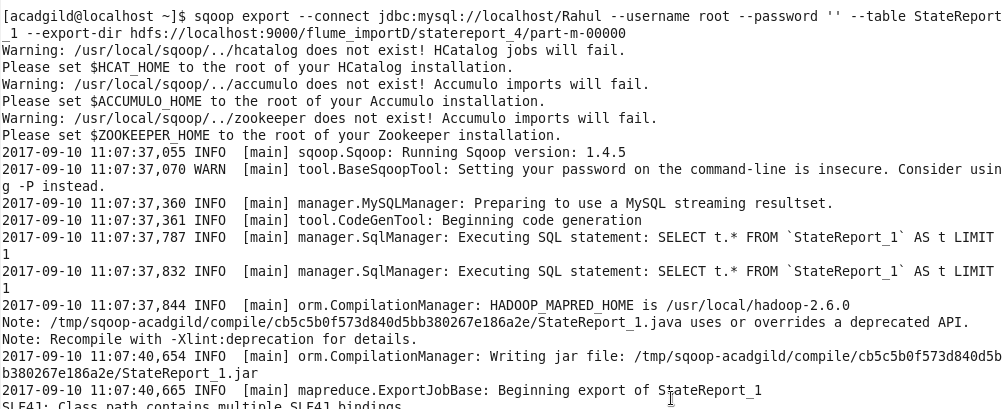
Start mysql in Acadgild:

sudo service mysqld start

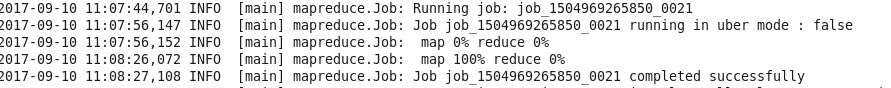
mysql -u root

sqoop export --connect jdbc:mysql://localhost/Rahul --username root --password '' --table StateReport\_1 --export-dir hdfs://localhost:9000/flume\_importD/statereport\_4/part-m-00000

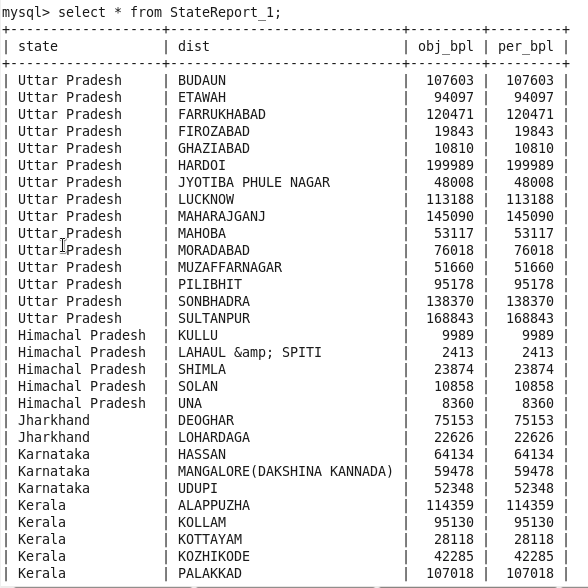
Screen:1

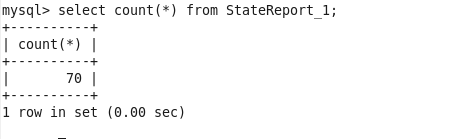


Screen: 2



Screen 3:





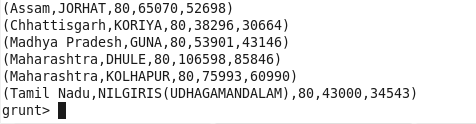
**2. Filter the districts which have reached 80% of objectives of BPL cards.**



**F =** foreach D generate $0 as state, $1 as dist, (obj\_prf \* 100)/obj\_bpl as gain\_per , obj\_bpl, obj\_prf;

G = Filter F by (gain\_per == 80);

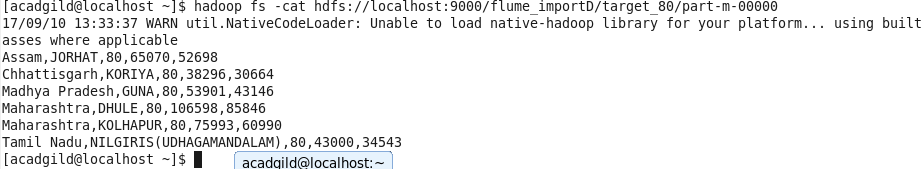
**Ans : 6 records found**



--store data @HDFS location

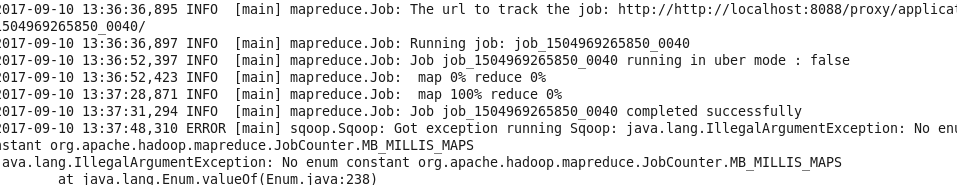
STORE G into 'hdfs://localhost:9000/flume\_importD/target\_80' using PigStorage (',');

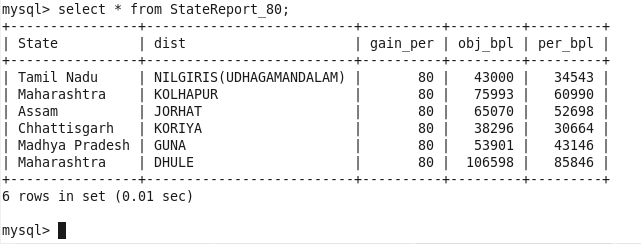




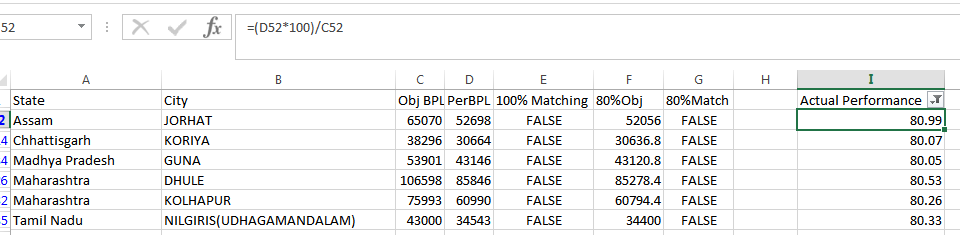
Export to MySQL from HDFS:

sqoop export --connect jdbc:mysql://localhost/Rahul --username root --password '' --table StateReport\_80 --export-dir hdfs://localhost:9000/flume\_importD/target\_80/part-m-00000





**Data Validation @ Excel**



80% Match:



Attached excel for clear understanding: