ACADGILD

BIG DATA HADOOP & SPARK TRAINING

SESSION – 12

PROJECT - **2**

State-Wise Development Analysis in India

TABLE OF CONTENTS

1. Executive Summary

1.1 Project Overview

To develop the System to analyze 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 analyzing 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)
- Analyze 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 preprocess

Problem Statement:

- Exporting the Data from the Local File System to the HDFS using Flume
- Performing Analysis on the data (in xml form) using PIG to get results for the below problem statements:
  - Find out the districts who achieved 100 percent objective in BPL cards
    Export the results to MySQL using Sqoop
  - Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.
    Export the results to MySQL using Sqoop.

Dataset:

The dataset is an xml file that contains the State-Wise Development data for India

Google Drive Link:
https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing

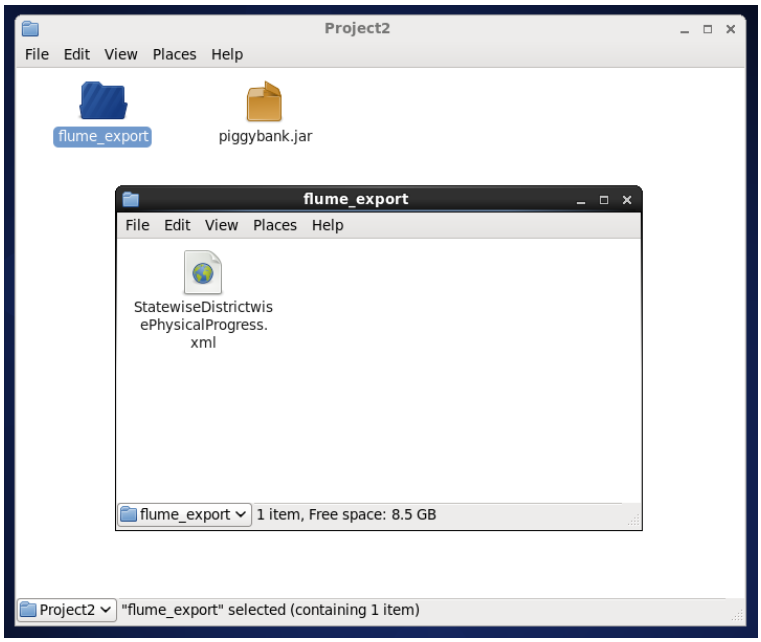
Screenshot:
A sample view of the data in the xml file.


```xml
<?xml version="1.0"?>
- <PhysicalProgress>
  - <row>
      <State_Name>Andhra Pradesh</State_Name>
      <District_Name>ADILABAD</District_Name>
      <Project_Objectives_IHHL_BPL>247475</Project_Objectives_IHHL_BPL>
      <Project_Objectives_IHHL_APL>148181</Project_Objectives_IHHL_APL>
      <Project_Objectives_IHHL_TOTAL>395656</Project_Objectives_IHHL_TOTAL>
      <Project_Objectives_SCW>0</Project_Objectives_SCW>
      <Project_Objectives_School_Toilets>4462</Project_Objectives_School_Toilets>
      <Project_Objectives_Anganwadi_Toilets>427</Project_Objectives_Anganwadi_Toilets>
      <Project_Objectives_RSM>10</Project_Objectives_RSM>
      <Project_Objectives_PC>0</Project_Objectives_PC>
      <Project_Performance-IHHL_BPL>176300</Project_Performance-IHHL_BPL>
      <Project_Performance-IHHL_APL>52431</Project_Performance-IHHL_APL>
      <Project_Performance-IHHL_TOTAL>228731</Project_Performance-IHHL_TOTAL>
      <Project_Performance-SCW>0</Project_Performance-SCW>
      <Project_Performance-School_Toilets>4462</Project_Performance-School_Toilets>
      <Project_Performance-Anganwadi_Toilets>427</Project_Performance-Anganwadi_Toilets>
      <Project_Performance-RSM>0</Project_Performance-RSM>
      <Project_Performance-PC>0</Project_Performance-PC>
    </row>
  - <row>
      <State_Name>Andhra Pradesh</State_Name>
      <District_Name>ANANTAPUR</District_Name>
      <Project_Objectives_IHHL_BPL>363314</Project_Objectives_IHHL_BPL>
      <Project_Objectives_IHHL_APL>181335</Project_Objectives_IHHL_APL>
      <Project_Objectives_IHHL_TOTAL>544649</Project_Objectives_IHHL_TOTAL>
      <Project_Objectives_SCW>0</Project_Objectives_SCW>
      <Project_Objectives_School_Toilets>3421</Project_Objectives_School_Toilets>
      <Project_Objectives_Anganwadi_Toilets>284</Project_Objectives_Anganwadi_Toilets>
      <Project_Objectives_RSM>10</Project_Objectives_RSM>
      <Project_Objectives_PC>0</Project_Objectives_PC>
      <Project_Performance-IHHL_BPL>366557</Project_Performance-IHHL_BPL>
      <Project_Performance-IHHL_APL>42000</Project_Performance-IHHL_APL>
      <Project_Performance-IHHL_TOTAL>408557</Project_Performance-IHHL_TOTAL>
      <Project_Performance-SCW>0</Project_Performance-SCW>
      <Project_Performance-School_Toilets>4258</Project_Performance-School_Toilets>
      <Project_Performance-Anganwadi_Toilets>302</Project_Performance-Anganwadi_Toilets>
      <Project_Performance-RSM>0</Project_Performance-RSM>
      <Project_Performance-PC>0</Project_Performance-PC>
    </row>
  - <row>
      <State_Name>Andhra Pradesh</State_Name>
      <District_Name>CHITTOOR</District_Name>
      <Project_Objectives_IHHL_BPL>296465</Project_Objectives_IHHL_BPL>
      <Project_Objectives_IHHL_APL>236986</Project_Objectives_IHHL_APL>
      <Project_Objectives_IHHL_TOTAL>533451</Project_Objectives_IHHL_TOTAL>
      <Project_Objectives_SCW>0</Project_Objectives_SCW>
      <Project_Objectives_School_Toilets>8171</Project_Objectives_School_Toilets>
      <Project_Objectives_Anganwadi_Toilets>375</Project_Objectives_Anganwadi_Toilets>
      <Project_Objectives_RSM>10</Project_Objectives_RSM>
      <Project_Objectives_PC>0</Project_Objectives_PC>
      <Project_Performance-IHHL_BPL>269750</Project_Performance-IHHL_BPL>
      <Project_Performance-IHHL_APL>190905</Project_Performance-IHHL_APL>
      <Project_Performance-IHHL_TOTAL>460655</Project_Performance-IHHL_TOTAL>
      <Project_Performance-SCW>0</Project_Performance-SCW>
      <Project_Performance-School_Toilets>8171</Project_Performance-School_Toilets>
      <Project_Performance-Anganwadi_Toilets>375</Project_Performance-Anganwadi_Toilets>
      <Project_Performance-RSM>11</Project_Performance-RSM>
      <Project_Performance-PC>0</Project_Performance-PC>
    </row>
```

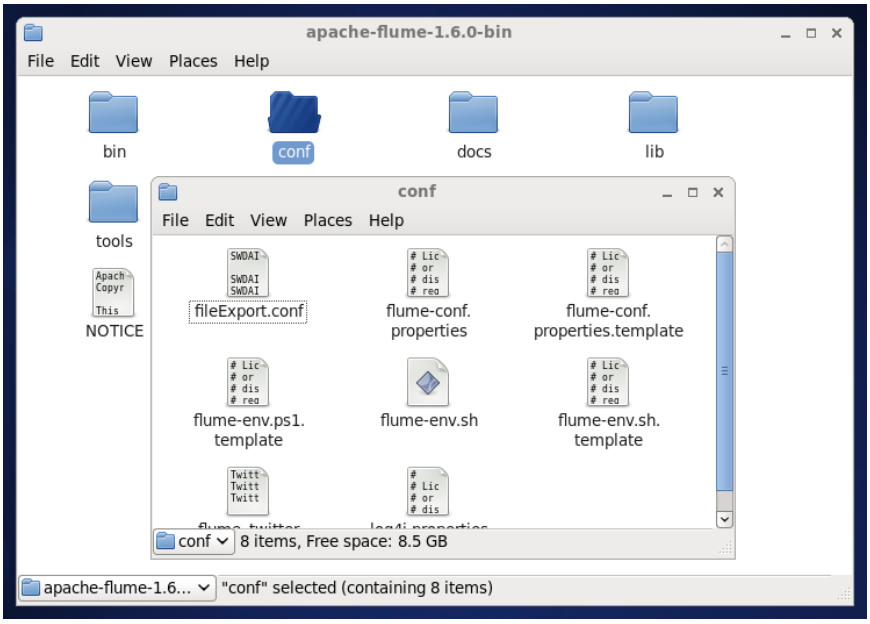
Exporting the Data from the Local File System to the HDFS using Flume

To perform this task we have to execute the following steps:

- Download Apache Flume for the Acadgild VM and extract it
  Update the location of Apache Flume in the .bashrc file
- Create the spool directory from where Flume will retrieve the data to be stored in the HDFS
  Here my spool directory is flume_export and my State and District Progression Log File for India is StatewiseDistrictwisePhysicalProgress.xml



- Create the configuration document for the flume job. This will contain the necessary information for **fetching log files from spool directory** and **storing these files in the HDFS**



My configuration file fileExport.conf is stored in the **conf** directory of Apache Flume directory

Below is the configuration file fileExport.conf, some important configurations are:

- Specifying the type of structure the file is coming in the channel: **file**
- Specifying the capacity of the transmission channel
- Specifying the type of source: **spool directory source**
- Specifying the path of the spool directory
- Specifying the suffix to be added to the name of the file in the spool directory
- Specifying the path in the HDFS to store the data

fileExport.conf (~/BDHST/Session12/apache-flume-1.6.0-bin/conf) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

fileExport.conf

```
SWDAIAgent.channels.fchannel.type = file

SWDAIAgent.channels.fchannel.capacity = 200000
SWDAIAgent.channels.fchannel.transactionCapacity = 1000
SWDAIAgent.sources.fsource.type = spooldir

SWDAIAgent.sources.fsource.spoolDir = /home/acadgild/BDHST/Session12/Projects/Project2/flume_export
SWDAIAgent.sources.fsource.fileHeader = false
SWDAIAgent.sources.fsource.fileSuffix = .COMPLETED
SWDAIAgent.sinks.hdfssink.type = hdfs

SWDAIAgent.sinks.hdfssink.hdfs.path = hdfs://localhost.localdomain:9000/flume_import
SWDAIAgent.sinks.hdfssink.hdfs.batchSize = 1000
SWDAIAgent.sinks.hdfssink.hdfs.rollSize = 268435456
SWDAIAgent.sinks.hdfssink.hdfs.rollInterval = 0
SWDAIAgent.sinks.hdfssink.hdfs.rollCount = 50000000
SWDAIAgent.sinks.hdfssink.hdfs.writeFormat=Text

SWDAIAgent.sinks.hdfssink.hdfs.fileType = DataStream
SWDAIAgent.sources.fsource.channels = fchannel
SWDAIAgent.sinks.hdfssink.channel = fchannel

SWDAIAgent.sinks =  hdfssink
SWDAIAgent.sources = fsource
SWDAIAgent.channels = fchannel
```

- Create the folder flume_import in the HDFS that will hold the data from Flume Agent/Job

acadgild@localhost:~

File Edit View Search Terminal Help

```
[acadgild@localhost ~]$ hdfs dfs -ls /
17/09/09 10:43:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 4 items
drwxr-xr-x   - acadgild supergroup          0 2015-11-09 19:21 /hbasestorage
drwxrwxr-x   - acadgild supergroup          0 2017-09-09 00:16 /tmp
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 01:56 /user
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 12:56 /zookeeper
[acadgild@localhost ~]$ hdfs dfs -mkdir /flume_import
17/09/09 10:43:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost ~]$ hdfs dfs -ls /
17/09/09 10:44:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 5 items
drwxr-xr-x   - acadgild supergroup          0 2017-09-09 10:43 /flume_import
drwxr-xr-x   - acadgild supergroup          0 2015-11-09 19:21 /hbasestorage
drwxrwxr-x   - acadgild supergroup          0 2017-09-09 00:16 /tmp
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 01:56 /user
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 12:56 /zookeeper
[acadgild@localhost ~]$ hdfs dfs -ls /flume_import
17/09/09 10:44:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost ~]$
```

- Execute the flume command that will create the flume job fetching data from the Local File System to the HDFS:

```
flume-ng agent -n <agentName> -f <path to fileExport.conf>
```

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ flume-ng agent -n SWDAIAgent -f /home/acadgild/BDHST/Session12/apache-flume-1.6.0-bin/conf/fileExport
.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/usr/local/hadoop-2.6.0/bin/hadoop) for HDFS access
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
eInfo: Including HBASE libraries found via (/usr/local/hbase/bin/hbase) for HBASE access
Info: Excluding /usr/local/hbase/lib/slf4j-api-1.6.4.jar from classpath
Info: Excluding /usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/usr/local/hive) for Hive access
+ exec /usr/local/java/bin/java -Xmx20m -cp '/home/acadgild/BDHST/Session12/apache-flume-1.6.0-bin/lib/*:/usr/local/hadoop-2.
6.0/contrib/capacity-scheduler/*.jar:/usr/local/hadoop-2.6.0/etc/hadoop:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/activ
ation-1.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/usr/local/hadoop-2.6.0/share/hadoo
p/common/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar
:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/asm-3
.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/avro-1.7.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons
-beanutils-1.7.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop-2.6.0
/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-codec-1.4.jar:/usr/local
/hadoop-2.6.0/share/hadoop/common/lib/commons-collections-3.2.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-c
ompress-1.4.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-configuration-1.6.jar:/usr/local/hadoop-2.6.0/share
/hadoop/common/lib/commons-digester-1.8.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-el-1.0.jar:/usr/local/had
oop-2.6.0/share/hadoop/common/lib/commons-httpclient-3.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-io-2.4.j
ar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-lang-2.6.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commo
ns-logging-1.1.3.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop-2.6.0/share/ha
doop/common/lib/commons-net-3.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/curator-client-2.6.0.jar:/usr/local/hadoo
17/09/09 10:45:19 INFO file.Log: Roll start /home/acadgild/.flume/file-channel/data
17/09/09 10:45:19 INFO file.LogFile: Opened /home/acadgild/.flume/file-channel/data/log-2
17/09/09 10:45:19 INFO file.Log: Roll end
17/09/09 10:45:19 INFO file.EventQueueBackingStoreFile: Start checkpoint for /home/acadgild/.flume/file-channel/checkpoint/ch
eckpoint, elements to sync = 0
17/09/09 10:45:19 INFO file.EventQueueBackingStoreFile: Updating checkpoint metadata: logWriteOrderID: 1504934119346, queueSi
ze: 0, queueHead: 12140
17/09/09 10:45:19 INFO file.Log: Updated checkpoint for file: /home/acadgild/.flume/file-channel/data/log-2 position: 0 logWr
iteOrderID: 1504934119346
17/09/09 10:45:19 INFO file.FileChannel: Queue Size after replay: 0 [channel=fchannel]
17/09/09 10:45:20 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: fchannel: Succ
essfully registered new MBean.
17/09/09 10:45:20 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: fchannel started
17/09/09 10:45:20 INFO node.Application: Starting Sink hdfssink
17/09/09 10:45:20 INFO node.Application: Starting Source fsource
17/09/09 10:45:20 INFO source.SpoolDirectorySource: SpoolDirectorySource source starting with directory: /home/acadgild/BDHST
/Session12/Projects/Project2/flume_export
17/09/09 10:45:20 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfssink: Success
fully registered new MBean.
17/09/09 10:45:20 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfssink started
17/09/09 10:45:20 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: fsource: Succes
sfully registered new MBean.
17/09/09 10:45:20 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: fsource started
17/09/09 10:45:21 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
17/09/09 10:45:22 INFO hdfs.BucketWriter: Creating hdfs://localhost.localdomain:9000/flume_import/FlumeData.1504934121099.tmp
17/09/09 10:45:22 INFO avro.ReliableSpoolingFileEventReader: Last read took us just up to a file boundary. Rolling to the nex
t file, if there is one.
17/09/09 10:45:22 INFO avro.ReliableSpoolingFileEventReader: Preparing to move file /home/acadgild/BDHST/Session12/Projects/P
roject2/flume_export/StatewiseDistrictwisePhysicalProgress.xml to /home/acadgild/BDHST/Session12/Projects/Project2/flume_expo
rt/StatewiseDistrictwisePhysicalProgress.xml.COMPLETED
17/09/09 10:45:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/09/09 10:45:49 INFO file.EventQueueBackingStoreFile: Start checkpoint for /home/acadgild/.flume/file-channel/checkpoint/ch
eckpoint, elements to sync = 12142
17/09/09 10:45:49 INFO file.EventQueueBackingStoreFile: Updating checkpoint metadata: logWriteOrderID: 1504934143766, queueSi
ze: 0, queueHead: 24280
17/09/09 10:45:49 INFO file.Log: Updated checkpoint for file: /home/acadgild/.flume/file-channel/data/log-2 position: 1742203
 logWriteOrderID: 1504934143766
```

This will now check the spool directory flume_export for the log file to export and then export/store it in the HDFS directory flume_import as given in the configuration file.

- Checking the HDFS import directory flume_import to see if the data has been exported successfully

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ hdfs dfs -ls /flume_import
17/09/09 10:49:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup     717415 2017-09-09 10:46 /flume_import/FlumeData.1504934121099
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -cat /flume_import/FlumeData.1504934121099
                <Project_Objectives_IHHL_BPL>628712</Project_Objectives_IHHL_BPL>
                <Project_Objectives_IHHL_APL>521192</Project_Objectives_IHHL_APL>
                <Project_Objectives_IHHL_TOTAL>1149904</Project_Objectives_IHHL_TOTAL>
                <Project_Objectives_SCW>50</Project_Objectives_SCW>
                <Project_Objectives_School_Toilets>8940</Project_Objectives_School_Toilets>
                <Project_Objectives_Anganwadi_Toilets>5448</Project_Objectives_Anganwadi_Toilets>
                <Project_Objectives_RSM>30</Project_Objectives_RSM>
                <Project_Objectives_PC>0</Project_Objectives_PC>
                <Project_Performance-IHHL_BPL>593712</Project_Performance-IHHL_BPL>
                <Project_Performance-IHHL_APL>162487</Project_Performance-IHHL_APL>
                <Project_Performance-IHHL_TOTAL>756199</Project_Performance-IHHL_TOTAL>
                <Project_Performance-SCW>31</Project_Performance-SCW>
                <Project_Performance-School_Toilets>7257</Project_Performance-School_Toilets>
                <Project_Performance-Anganwadi_Toilets>1631</Project_Performance-Anganwadi_Toilets>
                <Project_Performance-RSM>29</Project_Performance-RSM>
                <Project_Performance-PC>29</Project_Performance-PC>
        </row>
        <row>
                <State_Name>West Bengal</State_Name>
                <District_Name>UTTAR DINAJPUR</District_Name>
                <Project_Objectives_IHHL_BPL>257662</Project_Objectives_IHHL_BPL>
                <Project_Objectives_IHHL_APL>301645</Project_Objectives_IHHL_APL>
                <Project_Objectives_IHHL_TOTAL>559307</Project_Objectives_IHHL_TOTAL>
                <Project_Objectives_SCW>50</Project_Objectives_SCW>
                <Project_Objectives_School_Toilets>4806</Project_Objectives_School_Toilets>
                <Project_Objectives_Anganwadi_Toilets>1556</Project_Objectives_Anganwadi_Toilets>
                <Project_Objectives_RSM>30</Project_Objectives_RSM>
                <Project_Objectives_PC>0</Project_Objectives_PC>
                <Project_Performance-IHHL_BPL>148802</Project_Performance-IHHL_BPL>
                <Project_Performance-IHHL_APL>180619</Project_Performance-IHHL_APL>
                <Project_Performance-IHHL_TOTAL>329421</Project_Performance-IHHL_TOTAL>
                <Project_Performance-SCW>30</Project_Performance-SCW>
                <Project_Performance-School_Toilets>2562</Project_Performance-School_Toilets>
                <Project_Performance-Anganwadi_Toilets>2041</Project_Performance-Anganwadi_Toilets>
                <Project_Performance-RSM>17</Project_Performance-RSM>
                <Project_Performance-PC>0</Project_Performance-PC>
        </row>
</PhysicalProgress>
[acadgild@localhost ~]$
```

The xml file has been successfully exported

Also, below is the xml data file in the spool directory. As mentioned in the configuration file, the flume job has added the **COMPLETED** as suffix addition to the file name. This shows us that the file has been successfully read from the spool directory.

Performing Analysis on the data (in xml form) using PIG

Find out the districts who achieved 100 percent objective in BPL cards. Export the results to MySQL using Sqoop

This is a summary of the commands used to execute the above problem statement

StateDistrictAnalysisIndia100.pig (~/BDHST/Session12/Projects/Project2) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

StateDistrictAnalysisIndia100.pig

```
--Find the districts which have achieved 100% of objectives of BPL Cards

REGISTER /home/acadgild/BDHST/Session12/Projects/Project2/piggybank.jar;

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


A =  LOAD 'hdfs://localhost:9000/flume_import/FlumeData.1504934121099' using org.apache.pig.piggybank.storage.XMLLoader
('row') as (sdaIndia:chararray);

B = FOREACH A GENERATE XPath(sdaIndia, 'row/State_Name') AS State_Name, XPath(sdaIndia, 'row/District_Name') AS
District_Name, XPath(sdaIndia, 'row/Project_Objectives_IHHL_BPL') AS PO_IHHL_BPL, XPath(sdaIndia, 'row/
Project_Objectives_IHHL_APL') AS PO_IHHL_APL, XPath(sdaIndia, 'row/Project_Objectives_IHHL_TOTAL') AS PO_IHHL_TOTAL, XPath
(sdaIndia, 'row/Project_Objectives_SCW') AS PO_SCW, XPath(sdaIndia, 'row/Project_Objectives_School_Toilets') AS
PO_School_Toilets, XPath(sdaIndia, 'row/Project_Objectives_Anganwadi_Toilets') AS PO_Anganwadi_Toilets, XPath(sdaIndia, 'row/
Project_Objectives_RSM') AS PO_RSM, XPath(sdaIndia, 'row/Project_Objectives_PC') AS PO_PC, XPath(sdaIndia, 'row/
Project_Performance-IHHL_BPL') AS PP_IHHL_BPL, XPath(sdaIndia, 'row/Project_Performance-IHHL_APL') AS PP_IHHL_APL, XPath
(sdaIndia, 'row/Project_Performance-IHHL_TOTAL') AS PP_IHHL_TOTAL, XPath(sdaIndia, 'row/Project_Performance-SCW') AS PP_SCW,
XPath(sdaIndia, 'row/Project_Performance-School_Toilets') AS PP_School_Toilets, XPath(sdaIndia, 'row/Project_Performance-
Anganwadi_Toilets') AS PP_Anganwadi_Toilets, XPath(sdaIndia, 'row/Project_Performance-RSM') AS PP_RSM, XPath(sdaIndia, 'row/
Project_Performance-PC') AS PP_PC;

C = FOREACH B GENERATE District_Name,PO_IHHL_BPL,PP_IHHL_BPL, ROUND_TO(((double)PP_IHHL_BPL/(double)PO_IHHL_BPL) * 100,2) AS
BPL_Percentage;

D = FILTER C BY BPL_Percentage == 100;

STORE D INTO 'Project2/statedistrictanalysis100' USING PigStorage('\t');
```

- Starting the Pig Shell using the command **pig** (not local so we can access the HDFS)
- Registering the **piggybank** jar that contains the executables for various pig functions. Ex: Parse XML (Used in this assignment)

acadgild@localhost:~

```
[acadgild@localhost ~]$ pig
2017-09-09 10:50:18,924 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-09-09 10:50:18,930 INFO  [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-09-09 10:50:18,930 INFO  [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-09-09 10:50:19,085 [main] INFO  org.apache.pig.Main - Apache Pig version 0.14.0 (r1640057) compiled Nov 16 2014, 18:02:0
5
2017-09-09 10:50:19,085 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/pig_1504934419084.log
2017-09-09 10:50:19,166 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2017-09-09 10:50:19,896 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2017-09-09 10:50:19,896 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-09-09 10:50:19,896 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: hdfs://localhost:9000
2017-09-09 10:50:19,909 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.used.genericoptionsparser is d
eprecated. Instead, use mapreduce.client.genericoptionsparser.used
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-09-09 10:50:20,442 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
2017-09-09 10:50:21,239 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> REGISTER /home/acadgild/BDHST/Session12/Projects/Project2/piggybank.jar;
2017-09-09 10:50:32,098 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.
hours is deprecated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
```

- Defining the XML Parse function as **XPath** (name used to call the function)
- Loading the data in the HDFS (that was exported using Flume) and using the XML Loader function to load the data into the relation **A** with every starting tag 'row' as one line of type: chararray with the name **sdaIndia**
- Generating the rows (sdaIndia) in relation A by using the XML Parser **XPath.** Every tag under the main tag **row** will be separated by the tag name and given a pseudo name in the relation.

```
grunt> DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();
2017-09-09 11:17:47,857 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2
time(s).
grunt>
grunt> A =  LOAD 'hdfs://localhost:9000/flume_import/FlumeData.1504934121099' using org.apache.pig.piggybank.storage.XMLLoade
r('row') as (sdaIndia:chararray);
2017-09-09 11:18:03,403 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-09-09 11:18:03,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-09-09 11:18:03,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-09-09 11:18:03,415 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2
time(s).
grunt> B = FOREACH A GENERATE XPath(sdaIndia, 'row/State_Name') AS State_Name, XPath(sdaIndia, 'row/District_Name') AS Distri
ct_Name, XPath(sdaIndia, 'row/Project_Objectives_IHHL_BPL') AS PO_IHHL_BPL, XPath(sdaIndia, 'row/Project_Objectives_IHHL_APL'
) AS PO_IHHL_APL, XPath(sdaIndia, 'row/Project_Objectives_IHHL_TOTAL') AS PO_IHHL_TOTAL, XPath(sdaIndia, 'row/Project_Objecti
ves_SCW') AS PO_SCW, XPath(sdaIndia, 'row/Project_Objectives_School_Toilets') AS PO_School_Toilets, XPath(sdaIndia, 'row/Proj
ect_Objectives_Anganwadi_Toilets') AS PO_Anganwadi_Toilets, XPath(sdaIndia, 'row/Project_Objectives_RSM') AS PO_RSM, XPath(sd
aIndia, 'row/Project_Objectives_PC') AS PO_PC, XPath(sdaIndia, 'row/Project_Performance-IHHL_BPL') AS PP_IHHL_BPL, XPath(sdaI
ndia, 'row/Project_Performance-IHHL_APL') AS PP_IHHL_APL, XPath(sdaIndia, 'row/Project_Performance-IHHL_TOTAL') AS PP_IHHL_TO
TAL, XPath(sdaIndia, 'row/Project_Performance-SCW') AS PP_SCW, XPath(sdaIndia, 'row/Project_Performance-School_Toilets') AS P
P_School_Toilets, XPath(sdaIndia, 'row/Project_Performance-Anganwadi_Toilets') AS PP_Anganwadi_Toilets, XPath(sdaIndia, 'row/
Project_Performance-RSM') AS PP_RSM, XPath(sdaIndia, 'row/Project_Performance-PC') AS PP_PC;
2017-09-09 11:18:13,797 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2
time(s).
```

- Displaying the results of the Load statement

```
grunt> DUMP A;

(<row>  <State_Name>West Bengal</State_Name>    <District_Name>SILIGURI</District_Name> <Project_Objectives_IHHL_BPL>59536</P
roject_Objectives_IHHL_BPL>     <Project_Objectives_IHHL_APL>25377</Project_Objectives_IHHL_APL>        <Project_Objectives_I
HHL_TOTAL>84913</Project_Objectives_IHHL_TOTAL> <Project_Objectives_SCW>30</Project_Objectives_SCW>     <Project_Objectives_S
chool_Toilets>935</Project_Objectives_School_Toilets>   <Project_Objectives_Anganwadi_Toilets>1393</Project_Objectives_Anganw
adi_Toilets>    <Project_Objectives_RSM>0</Project_Objectives_RSM>      <Project_Objectives_PC>10</Project_Objectives_PC>   <
Project_Performance-IHHL_BPL>37794</Project_Performance-IHHL_BPL>       <Project_Performance-IHHL_APL>18060</Project_Performa
nce-IHHL_APL>   <Project_Performance-IHHL_TOTAL>55854</Project_Performance-IHHL_TOTAL>  <Project_Performance-SCW>30</Project_
Performance-SCW>        <Project_Performance-School_Toilets>929</Project_Performance-School_Toilets>    <Project_Performance-
Anganwadi_Toilets>906</Project_Performance-Anganwadi_Toilets>   <Project_Performance-RSM>5</Project_Performance-RSM>    <Proj
ect_Performance-PC>7</Project_Performance-PC></row>)
(<row>  <State_Name>West Bengal</State_Name>    <District_Name>SOUTH 24 PARAGANAS</District_Name>       <Project_Objectives_I
HHL_BPL>628712</Project_Objectives_IHHL_BPL>    <Project_Objectives_IHHL_APL>521192</Project_Objectives_IHHL_APL>       <Proj
ect_Objectives_IHHL_TOTAL>1149904</Project_Objectives_IHHL_TOTAL>       <Project_Objectives_SCW>50</Project_Objectives_SCW> <
Project_Objectives_School_Toilets>8940</Project_Objectives_School_Toilets>      <Project_Objectives_Anganwadi_Toilets>5448</P
roject_Objectives_Anganwadi_Toilets>    <Project_Objectives_RSM>30</Project_Objectives_RSM>     <Project_Objectives_PC>0</Pro
ject_Objectives_PC>     <Project_Performance-IHHL_BPL>593712</Project_Performance-IHHL_BPL>     <Project_Performance-IHHL_APL
>162487</Project_Performance-IHHL_APL>  <Project_Performance-IHHL_TOTAL>756199</Project_Performance-IHHL_TOTAL> <Project_Perf
ormance-SCW>31</Project_Performance-SCW>        <Project_Performance-School_Toilets>7257</Project_Performance-School_Toilets>
<Project_Performance-Anganwadi_Toilets>1631</Project_Performance-Anganwadi_Toilets>     <Project_Performance-RSM>29</Project_
Performance-RSM>        <Project_Performance-PC>29</Project_Performance-PC></row>)
(<row>  <State_Name>West Bengal</State_Name>    <District_Name>UTTAR DINAJPUR</District_Name>   <Project_Objectives_IHHL_BPL>
257662</Project_Objectives_IHHL_BPL>    <Project_Objectives_IHHL_APL>301645</Project_Objectives_IHHL_APL>       <Project_Obje
ctives_IHHL_TOTAL>559307</Project_Objectives_IHHL_TOTAL>        <Project_Objectives_SCW>50</Project_Objectives_SCW>     <Proj
ect_Objectives_School_Toilets>4806</Project_Objectives_School_Toilets>  <Project_Objectives_Anganwadi_Toilets>1556</Project_O
bjectives_Anganwadi_Toilets>    <Project_Objectives_RSM>30</Project_Objectives_RSM>     <Project_Objectives_PC>0</Project_Obj
ectives_PC>     <Project_Performance-IHHL_BPL>148802</Project_Performance-IHHL_BPL>     <Project_Performance-IHHL_APL>180619<
/Project_Performance-IHHL_APL>  <Project_Performance-IHHL_TOTAL>329421</Project_Performance-IHHL_TOTAL> <Project_Performance-
SCW>30</Project_Performance-SCW>        <Project_Performance-School_Toilets>2562</Project_Performance-School_Toilets>   <Proj
ect_Performance-Anganwadi_Toilets>2041</Project_Performance-Anganwadi_Toilets>  <Project_Performance-RSM>17</Project_Performa
nce-RSM>        <Project_Performance-PC>0</Project_Performance-PC></row>)
grunt>
```

- Displaying the result of the Row Generating statement. All the data has been separated by tag name and formatted into a tuple of multiple fields.

```
grunt> DUMP B;
2017-09-09 11:20:25,694 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2017-09-09 11:20:25,781 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
(Uttar Pradesh,SHRAVASTI,104902,54772,159674,10,1838,650,3,0,98761,41818,140579,10,1838,650,0,0)
(Uttar Pradesh,SIDDHARTHNAGAR,139597,133650,273247,50,4128,1481,5,0,124597,111651,236248,50,4128,1481,2,0)
(Uttar Pradesh,SITAPUR,305299,255574,560873,25,7397,2307,3,1,273463,185138,458601,25,5874,2307,1,1)
(Uttar Pradesh,SONBHADRA,138370,79419,217789,113,3176,1051,0,0,138370,57900,196270,113,2978,755,10,0)
(Uttar Pradesh,SULTANPUR,168843,262071,430914,40,4898,2244,6,0,168843,215646,384489,40,4898,2244,6,0)
(Uttar Pradesh,UNNAO,229599,141734,371333,20,4700,1683,8,0,223630,131996,355626,20,4700,1683,2,2)
(Uttar Pradesh,VARANASI,105408,249056,354464,47,4471,1532,0,0,102430,147309,249739,47,2516,1283,3,0)
(Uttarakhand,ALMORA,45572,48151,93723,40,214,247,10,0,32277,38661,70938,30,239,4,0,0)
(Uttarakhand,BAGESHWAR,21447,16508,37955,20,111,149,3,0,17859,18603,36462,4,105,2,2,0)
(Uttarakhand,CHAMOLI,27147,26145,53292,10,265,115,7,2,26043,25844,51887,0,222,7,3,0)
(Uttarakhand,CHAMPAWAT,22991,9085,32076,20,375,78,2,1,18323,14178,32501,3,160,6,1,0)
(Uttarakhand,DEHRADUN,37212,24463,61675,0,497,19,5,1,31724,24651,56375,0,396,19,1,0)
(Uttarakhand,HARIDWAR,42500,90074,132574,60,200,20,2,0,38188,77694,115882,37,100,13,0,0)
(Uttarakhand,NAINITAL,14314,26585,40899,50,208,95,6,2,18039,29025,47064,1,208,19,2,0)
(Uttarakhand,PAURI(GARHWAL),53399,37146,90545,50,620,173,10,0,35852,42518,78370,8,367,23,3,0)
(Uttarakhand,PITHORAGARH,41110,29253,70363,50,670,215,6,2,34597,31566,66163,10,439,102,3,0)
(Uttarakhand,RUDRAPRAYAG,13150,22322,35472,40,100,124,3,0,13810,17522,31332,9,100,70,2,0)
(Uttarakhand,TEHRI GARHWAL,55173,34580,89753,30,307,120,8,0,46293,35076,81369,0,486,10,0,0)
(Uttarakhand,UDHAM SINGH NAGAR,39427,59658,99085,50,118,123,5,2,37604,51384,88988,0,119,75,4,1)
(Uttarakhand,UTTARKASHI,28189,20700,48889,50,240,123,4,0,25523,21017,46540,4,240,0,3,0)
(West Bengal,BANKURA,198152,333832,531984,50,7544,4130,29,0,105545,243191,348736,46,7687,1340,26,0)
(West Bengal,BARDHAMAN,700047,341920,1041967,133,9891,7980,10,0,601906,277914,879820,140,9890,7724,10,19)
(West Bengal,BIRBHUM,338989,299893,638882,50,5617,3816,4,22,266347,186599,452946,58,5563,2233,19,0)
(West Bengal,COOCH BEHAR,335236,254422,589658,50,3715,1718,15,0,262294,164038,426332,144,5764,1818,15,12)
(West Bengal,DAKSHIN DINAJPUR,182621,194577,377198,50,3712,2642,10,0,184153,49448,233601,19,2632,939,8,8)
(West Bengal,DARJEELING,66648,130066,196714,50,1784,408,0,0,32921,3035,35956,18,1435,574,8,8)
(West Bengal,HOOGHLY,271737,195510,467247,53,6821,4168,19,0,269779,191294,461073,49,6764,3435,18,18)
(West Bengal,HOWRAH,231860,143309,375169,51,5195,3586,26,0,230190,141912,372102,42,5178,2733,14,20)
(West Bengal,JALPAIGURI,372999,203523,576522,50,6578,5428,87,0,337740,101550,439290,25,6578,4064,17,14)
(West Bengal,MALDA,452324,270208,722532,50,6385,7956,6,0,321934,65298,387232,41,5934,327,15,15)
(West Bengal,MIDNAPUR EAST,392371,32617,424988,172,9726,5969,25,0,527389,32642,560031,210,10149,2882,8,17)
(West Bengal,MIDNAPUR WEST,509496,432096,941592,50,16498,5825,10,0,596291,322659,918950,73,13452,2787,0,0)
(West Bengal,MURSHIDABAD,702442,506963,1209405,50,10260,7012,18,0,498998,198174,697172,47,7838,2423,26,26)
(West Bengal,NADIA,346696,278335,625031,50,6974,6620,50,0,321462,198890,520352,28,6635,3961,17,41)
(West Bengal,NORTH 24 PARAGANAS,361462,225080,586542,51,11158,4466,30,0,357960,226104,584064,66,10931,3150,101,0)
(West Bengal,PURULIA,210168,306933,517101,50,7542,4047,10,0,97160,79169,176329,10,4692,1128,20,0)
(West Bengal,SILIGURI,59536,25377,84913,30,935,1393,0,10,37794,18060,55854,30,929,906,5,7)
(West Bengal,SOUTH 24 PARAGANAS,628712,521192,1149904,50,8940,5448,30,0,593712,162487,756199,31,7257,1631,29,29)
(West Bengal,UTTAR DINAJPUR,257662,301645,559307,50,4806,1556,30,0,148802,180619,329421,30,2562,2041,17,0)
grunt>
```

- Generating column names pertaining to **District** and **BPL information** and finding the **Percentage** of performance achieved for the objective that was set for BPL Cards in India.
- Filtering the above result for those records where 100% objective has been met and displaying the result.

```
grunt> C = FOREACH B GENERATE District_Name,PO_IHHL_BPL,PP_IHHL_BPL, ROUND_TO(((double)PP_IHHL_BPL/(double)PO_IHHL_BPL) * 100
,2) AS BPL_Percentage;
2017-09-09 11:22:05,872 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 3
time(s).
grunt> D = FILTER C BY BPL_Percentage == 100;
2017-09-09 11:22:18,879 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 4
time(s).
grunt> DUMP D;
2017-09-09 11:22:27,962 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2
time(s).
2017-09-09 11:22:27,975 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2017-09-09 11:22:28,040 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-09-09 11:22:28,040 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-09-09 11:22:28,043 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-09-09 11:22:28,043 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2017-09-09 11:22:28,044 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Mer
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTyp
eCastInserter]}
2017-09-09 11:22:28,089 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
```

- The result of the above procedure:

```
acadgild@localhost:~
(NIZAMABAD,225519,225519,100.0)
(TIRAP,5780,5780,100.0)
(HAILAKANDI,49837,49837,100.0)
(MADHUBANI,67482,67482,100.0)
(NORTH GOA,15000,15000,100.0)
(AHMEDABAD,80192,80192,100.0)
(DANGS,27900,27900,100.0)
(NAVSARI,75015,75015,100.0)
(PORBANDAR,17024,17024,100.0)
(SURAT,158797,158797,100.0)
(FARIDABAD,22254,22254,100.0)
(HISAR,46463,46463,100.0)
(JHAJJAR,22014,22014,100.0)
(MAHENDRAGARH,17500,17500,100.0)
(PANCHKULA,8760,8760,100.0)
(PANIPAT,28000,28000,100.0)
(ROHTAK,22171,22171,100.0)
(SIRSA,35400,35400,100.0)
(HAMIRPUR,11593,11593,100.0)
(KINNAUR,1560,1560,100.0)
(KULLU,9989,9989,100.0)
(LAHAUL & SPITI,2413,2413,100.0)
(MANDI,34407,34408,100.0)
(SHIMLA,23874,23874,100.0)
(SOLAN,10858,10858,100.0)
(UNA,8360,8360,100.0)
(DEOGHAR,75153,75153,100.0)
(LOHARDAGA,22626,22626,100.0)
(HASSAN,64134,64134,100.0)
(MANGALORE(DAKSHINA KANNADA),59478,59478,100.0)
(UDUPI,52348,52348,100.0)
(ALAPPUZHA,114359,114359,100.0)
(KOLLAM,95130,95130,100.0)
(KOTTAYAM,28118,28118,100.0)
(KOZHIKODE,42285,42285,100.0)
(PALAKKAD,107018,107018,100.0)
(PATHANAMTHITTA,53799,53799,100.0)
(WAYANAD,50655,50655,100.0)
(GADCHIROLI,75900,75900,100.0)
(SINDHUDURG,43874,43874,100.0)
(WEST GARO HILLS,44385,44385,100.0)
(CHAMPHAI,11077,11077,100.0)
(LAWNGTLAI,16544,16544,100.0)
(HANUMANGARH,31621,31621,100.0)
(ERODE,165306,165306,100.0)
(KARUR,105280,105280,100.0)
(NAMAKKAL,117538,117538,100.0)
(TIRUCHIRAPPALLI,77747,77747,100.0)
(TIRUVANNAMALAI,209116,209116,100.0)
(DHALAI,53507,53507,100.0)
(SOUTH TRIPURA,139456,139456,100.0)
(WEST TRIPURA,183405,183405,100.0)
(AMBEDKAR NAGAR,132725,132725,100.0)
(BALRAMPUR,65273,65273,100.0)
(BAREILLY,110000,110000,100.0)
(BIJNOR,110403,110403,100.0)
(BUDAUN,107603,107603,100.0)
(ETAWAH,94097,94097,100.0)
(FARRUKHABAD,120471,120471,100.0)
(FIROZABAD,19843,19843,100.0)
(GHAZIABAD,10810,10810,100.0)
(HARDOI,199989,199989,100.0)
(JYOTIBA PHULE NAGAR,48008,48008,100.0)
(LUCKNOW,113188,113188,100.0)
(MAHARAJGANJ,145090,145090,100.0)
(MAHOBA,53117,53117,100.0)
(MORADABAD,76018,76018,100.0)
(MUZAFFARNAGAR,51660,51660,100.0)
(PILIBHIT,95178,95178,100.0)
(SONBHADRA,138370,138370,100.0)
(SULTANPUR,168843,168843,100.0)
grunt>
```

- Now we store the result in the HDFS for the **Sqoop job to export the data to a MySQL database**

- Storing the data in the HDFS under the path given below and separating the fields by tab space

```
grunt> STORE D INTO 'Project2/statedistrictanalysis100' USING PigStorage('\t');
2017-09-09 11:26:45,269 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-09-09 11:26:45,269 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-09-09 11:26:45,269 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-09-09 11:26:45,283 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 4
time(s).
2017-09-09 11:26:45,328 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2
time(s).
2017-09-09 11:26:45,334 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is
```

- To check if the file has been successfully stored in the HDFS, we check the output folder of its contents.
  The data has been stored successfully as seen by the file named **part-m-00000** that hold the output of the MapReduce job

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ hdfs dfs -ls Project2
17/09/09 15:11:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
drwxr-xr-x   - acadgild supergroup          0 2017-09-09 15:11 Project2/statedistrictanalysis100
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -ls Project2/statedistrictanalysis100
17/09/09 15:12:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2017-09-09 15:11 Project2/statedistrictanalysis100/_SUCCESS
-rw-r--r--   1 acadgild supergroup       1998 2017-09-09 15:11 Project2/statedistrictanalysis100/part-m-00000
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -cat Project2/statedistrictanalysis100/part-m-00000
17/09/09 15:12:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
NIZAMABAD       225519  225519  100.0
TIRAP   5780    5780    100.0
HAILAKANDI      49837   49837   100.0
MADHUBANI       67482   67482   100.0
NORTH GOA       15000   15000   100.0
AHMEDABAD       80192   80192   100.0
DANGS   27900   27900   100.0
NAVSARI 75015   75015   100.0
PORBANDAR       17024   17024   100.0
SURAT   158797  158797  100.0
FARIDABAD       22254   22254   100.0
HISAR   46463   46463   100.0
JHAJJAR 22014   22014   100.0
MAHENDRAGARH    17500   17500   100.0
PANCHKULA       8760    8760    100.0
PANIPAT 28000   28000   100.0
ROHTAK  22171   22171   100.0
SIRSA   35400   35400   100.0
HAMIRPUR        11593   11593   100.0
KINNAUR 1560    1560    100.0
KULLU   9989    9989    100.0
```

- Now we export the data in the HDFS to a Table in MySQL by the following steps:
  - Start the MySQL service and terminal and create the database and table to hold the data
    Here my table is named **SD_Analysis_100**

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ sudo service mysqld start
Starting mysqld:                                           [  OK  ]
[cloudera@quickstart ~]$ mysql -uroot -pcloudera
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 22
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> CREATE DATABASE MiniProject;
Query OK, 1 row affected (0.00 sec)

mysql> USE MiniProject;
Database changed
mysql> CREATE TABLE SD_Analysis_100
    -> (
    -> District_Name varchar(30),
    -> PO_IHHL_BPL int,
    -> PP_IHHL_BPL int,
    -> BPL_Percentage double
    -> );
Query OK, 0 rows affected (0.03 sec)

mysql> SELECT * FROM SD_Analysis_100;
Empty set (0.00 sec)

mysql>
```

- Using the Sqoop command given below:
  - ✓ Specifying the name of the database to hold the data
  - ✓ Specifying the password of the VM (Can also be manually entered or got from a password file)
  - ✓ Specifying the name of the table to hold the data
  - ✓ Specifying the directory in the HDFS that holds the data
  - ✓ Specifying how the fields are terminated
  - ✓ Specifying the number of MapReduce jobs :1
  - ✓ Specifying the column names to import to the MySQL table

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/MiniProject --username 'root' -password cloudera
--table 'SD_Analysis_100' --export-dir 'statedistrictanalysis100' --input-fields-terminated-by '\t' -m 1 --columns Distr
ict_Name,PO_IHHL_BPL,PP_IHHL_BPL,BPL_Percentage
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/09/10 20:51:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
17/09/10 20:51:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P inst
ead.
17/09/10 20:51:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/09/10 20:51:47 INFO tool.CodeGenTool: Beginning code generation
17/09/08 23:30:22 INFO mapreduce.Job: Job job_1504933809120_0005 completed successfully
17/09/08 23:30:22 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=151469
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2152
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
```

```
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=10112
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=10112
                Total vcore-milliseconds taken by all map tasks=10112
                Total megabyte-milliseconds taken by all map tasks=10354688
        Map-Reduce Framework
                Map input records=71
                Map output records=71
                Input split bytes=151
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=145
                CPU time spent (ms)=1260
                Physical memory (bytes) snapshot=136859648
                Virtual memory (bytes) snapshot=1508085760
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/09/08 23:30:22 INFO mapreduce.ExportJobBase: Transferred 2.1016 KB in 35.6129 seconds (60.4275 bytes/sec)
17/09/08 23:30:22 INFO mapreduce.ExportJobBase: Exported 71 records.
[cloudera@quickstart ~]$
```

The file has been successfully written to the MySQL table **SD_Analysis_100**

OUTPUT:

- To check the contents of the MySQL table **SD_Analysis_100** use the **SELECT *** command

cloudera@quickstart:~

File Edit View Search Terminal Help

```
mysql> SELECT * FROM SD_Analysis_100;
+-----------------------------+-------------+-------------+----------------+
| District_Name               | PO_IHHL_BPL | PP_IHHL_BPL | BPL_Percentage |
+-----------------------------+-------------+-------------+----------------+
| NIZAMABAD                   |      225519 |      225519 |            100 |
| TIRAP                       |        5780 |        5780 |            100 |
| HAILAKANDI                  |       49837 |       49837 |            100 |
| MADHUBANI                   |       67482 |       67482 |            100 |
| NORTH GOA                   |       15000 |       15000 |            100 |
| AHMEDABAD                   |       80192 |       80192 |            100 |
| DANGS                       |       27900 |       27900 |            100 |
| NAVSARI                     |       75015 |       75015 |            100 |
| PORBANDAR                   |       17024 |       17024 |            100 |
| SURAT                       |      158797 |      158797 |            100 |
| FARIDABAD                   |       22254 |       22254 |            100 |
| HISAR                       |       46463 |       46463 |            100 |
| JHAJJAR                     |       22014 |       22014 |            100 |
| MAHENDRAGARH                |       17500 |       17500 |            100 |
| PANCHKULA                   |        8760 |        8760 |            100 |
| PANIPAT                     |       28000 |       28000 |            100 |
| ROHTAK                      |       22171 |       22171 |            100 |
| SIRSA                       |       35400 |       35400 |            100 |
| HAMIRPUR                    |       11593 |       11593 |            100 |
| KINNAUR                     |        1560 |        1560 |            100 |
| KULLU                       |        9989 |        9989 |            100 |
| LAHAUL & SPITI              |        2413 |        2413 |            100 |
| MANDI                       |       34407 |       34408 |            100 |
| SHIMLA                      |       23874 |       23874 |            100 |
| SOLAN                       |       10858 |       10858 |            100 |
| UNA                         |        8360 |        8360 |            100 |
| DEOGHAR                     |       75153 |       75153 |            100 |
| LOHARDAGA                   |       22626 |       22626 |            100 |
| HASSAN                      |       64134 |       64134 |            100 |
| MANGALORE(DAKSHINA KANNADA) |       59478 |       59478 |            100 |
| UDUPI                       |       52348 |       52348 |            100 |
| ALAPPUZHA                   |      114359 |      114359 |            100 |
| KOLLAM                      |       95130 |       95130 |            100 |
| KOTTAYAM                    |       28118 |       28118 |            100 |
| KOZHIKODE                   |       42285 |       42285 |            100 |
| PALAKKAD                    |      107018 |      107018 |            100 |
```

PATHANAMTHITTA	53799	53799	100
WAYANAD	50655	50655	100
GADCHIROLI	75900	75900	100
SINDHUDURG	43874	43874	100
WEST GARO HILLS	44385	44385	100
CHAMPHAI	11077	11077	100
LAWNGTLAI	16544	16544	100
HANUMANGARH	31621	31621	100
ERODE	165306	165306	100
KARUR	105280	105280	100
NAMAKKAL	117538	117538	100
TIRUCHIRAPPALLI	77747	77747	100
TIRUVANNAMALAI	209116	209116	100
DHALAI	53507	53507	100
SOUTH TRIPURA	139456	139456	100
WEST TRIPURA	183405	183405	100
AMBEDKAR NAGAR	132725	132725	100
BALRAMPUR	65273	65273	100
BAREILLY	110000	110000	100
BIJNOR	110403	110403	100
BUDAUN	107603	107603	100
ETAWAH	94097	94097	100
FARRUKHABAD	120471	120471	100
FIROZABAD	19843	19843	100
GHAZIABAD	10810	10810	100
HARDOI	199989	199989	100
JYOTIBA PHULE NAGAR	48008	48008	100
LUCKNOW	113188	113188	100
MAHARAJGANJ	145090	145090	100
MAHOBA	53117	53117	100
MORADABAD	76018	76018	100
MUZAFFARNAGAR	51660	51660	100
PILIBHIT	95178	95178	100
SONBHADRA	138370	138370	100
SULTANPUR	168843	168843	100

71 rows in set (0.03 sec)

mysql>

Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards. Export the results to MySQL using Sqoop.

- To filter the districts that have reached 80% of their objectives in BPL Cards, I have created a Pig Script(with commands similar to the problem before) and executed it via the pig MapReduce shell

StateDistrictAnalysisIndia80.pig (~/BDHST/Session12/Projects/Project2) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

StateDistrictAnalysisIndia80.pig

```
--Pig UDF to filter the districts which have reached 80% of objectives of BPL Cards

REGISTER /home/acadgild/BDHST/Session12/Projects/Project2/piggybank.jar;

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

REGISTER '/home/acadgild/BDHST/Session12/Projects/Project2/miniprojectudf.jar';

DEFINE getPercentage filterBPL80.Filter80;


A =  LOAD 'hdfs://localhost:9000/flume_import/FlumeData.1504934121099' using org.apache.pig.piggybank.storage.XMLLoader
('row') as (sdaIndia:chararray);

B = FOREACH A GENERATE XPath(sdaIndia, 'row/State_Name') AS State_Name, XPath(sdaIndia, 'row/District_Name') AS
District_Name, XPath(sdaIndia, 'row/Project_Objectives_IHHL_BPL') AS PO_IHHL_BPL, XPath(sdaIndia, 'row/
Project_Objectives_IHHL_APL') AS PO_IHHL_APL, XPath(sdaIndia, 'row/Project_Objectives_IHHL_TOTAL') AS PO_IHHL_TOTAL, XPath
(sdaIndia, 'row/Project_Objectives_SCW') AS PO_SCW, XPath(sdaIndia, 'row/Project_Objectives_School_Toilets') AS
PO_School_Toilets, XPath(sdaIndia, 'row/Project_Objectives_Anganwadi_Toilets') AS PO_Anganwadi_Toilets, XPath(sdaIndia, 'row/
Project_Objectives_RSM') AS PO_RSM, XPath(sdaIndia, 'row/Project_Objectives_PC') AS PO_PC, XPath(sdaIndia, 'row/
Project_Performance-IHHL_BPL') AS PP_IHHL_BPL, XPath(sdaIndia, 'row/Project_Performance-IHHL_APL') AS PP_IHHL_APL, XPath
(sdaIndia, 'row/Project_Performance-IHHL_TOTAL') AS PP_IHHL_TOTAL, XPath(sdaIndia, 'row/Project_Performance-SCW') AS PP_SCW,
XPath(sdaIndia, 'row/Project_Performance-School_Toilets') AS PP_School_Toilets, XPath(sdaIndia, 'row/Project_Performance-
Anganwadi_Toilets') AS PP_Anganwadi_Toilets, XPath(sdaIndia, 'row/Project_Performance-RSM') AS PP_RSM, XPath(sdaIndia, 'row/
Project_Performance-PC') AS PP_PC;

C = FOREACH B GENERATE State_Name .. PP_PC, ROUND_TO(getPercentage((double)PP_IHHL_BPL,(double)PO_IHHL_BPL),2) AS BPL_Percent;

D = FILTER C BY BPL_Percent!=0.0;

E = FOREACH (GROUP D ALL) GENERATE COUNT_STAR(D);

STORE D INTO 'Project2/statedistrictanalysis80' USING PigStorage('\t');
```

The steps followed are explained as below:

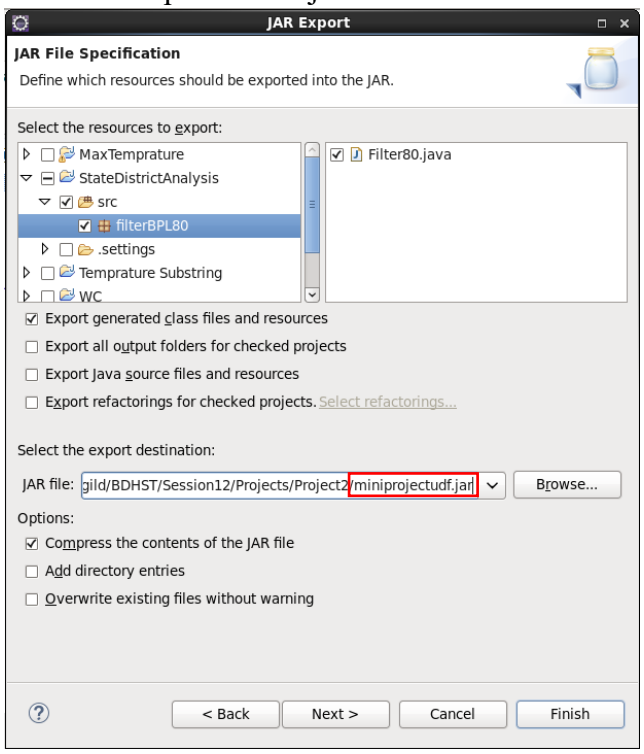
- Registering the piggybank jar that contains the executables for various pig functions. Ex: Parse XML (Used in this assignment)
- Defining the XML Parse function as **XPath** (name used to call the function)
- Registering the Pig UDF miniprojectudf created to filter the districts which have reached 80% of objectives of BPL cards. (Written in Java)
- Defining getPercentage as the function to be used to execute the UDF in package **filterBPL80** and class **Filter80**
- Loading the data in the HDFS (that was exported using Flume) and using the XML Loader function to load the data into the relation **A** with every starting tag 'row' as one line of type: chararray with the name **sdaIndia**
- Generating the rows (sdaIndia) in relation A by using the XML Parser **XPath.** Every tag under the main tag **row** will be separated by the tag name and given a pseudo name in the relation.

- Generating all column and finding the **Percentage** of performance achieved, for the objective that was set for BPL Cards in India, by using a Pig UDF written in java and exported as a jar as below:

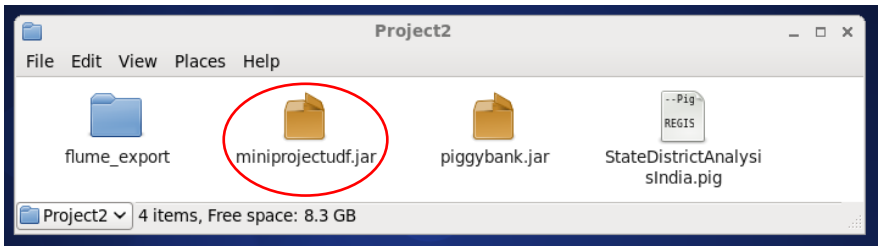
  - Below is an image of the Pig UDF

```java
package filterBPL80;

import java.io.IOException;
import org.apache.pig.data.Tuple;
import org.apache.pig.EvalFunc;

public class Filter80 extends EvalFunc<Double> {

    Double percent = null;

    public Double exec(Tuple input) throws IOException {

        //get the project performance and objective for BPL cards
        //from tuple sent as parameter to func call
        double p_performance = (double) input.get(0);
        double p_objective = (double)input.get(1);

        //get percentage of performance achieved for objective set
        double p_percentage = (p_performance/p_objective) * 100;

        //check whether percentage greater than or equal to 80
        //if yes, return percentage else 0
        if(p_percentage >= 80.00)
            return p_percentage;
        else
            return 0.00;
    }
}
```

  - The UDF exported as a jar



- The UDF in the directory from where it is accessed



- Filtering the above result for those records where percentage is 0.0% (The records that do not meet the 80% objective). Therefore giving us the records that have received 80% and above in BPL cards
- Getting the count of the filtered records
- Storing the results, i.e. the filter records into a directory in the HDFS and separating the fields by tab space

- Executing the Pig Script in MapReduce mode (can access HDFS) as below:

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ pig -x mapreduce /home/acadgild/BDHST/Session12/Projects/Project2/StateDistrictAnalysisIndia80.pig

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgRe
duceTime        MedianReducetime        Alias   Feature Outputs
job_1504933943113_0015  1       0       36      36      36      36      0       0       0       0       A,B,C,D MAP_ONLY    h
dfs://localhost:9000/user/acadgild/Project2/statedistrictanalysis80,

Input(s):
Successfully read 0 records from: "hdfs://localhost:9000/flume_import/FlumeData.1504934121099"

Output(s):
Successfully stored 0 records in: "hdfs://localhost:9000/user/acadgild/Project2/statedistrictanalysis80"

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1504933943113_0015


2017-09-09 15:16:30,743 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-09-09 15:16:30,757 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-09 15:16:30,939 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-09-09 15:16:30,996 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-09 15:16:31,160 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-09-09 15:16:31,192 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-09-09 15:16:31,246 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable
to retrieve job to compute warning aggregation.
2017-09-09 15:16:31,246 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-09-09 15:16:31,312 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 7 seconds and 934 milliseconds (
67934 ms)
[acadgild@localhost ~]$
```

The execution is successful.

- Checking the contents of the folder statedistrictanalysis80 in HDFS that contains the filtered data

  The data has been stored successfully as seen by the file named **part-m-00000** that hold the output of the MapReduce job.

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ hdfs dfs -ls Project2
17/09/09 15:17:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
drwxr-xr-x   - acadgild supergroup          0 2017-09-09 15:11 Project2/statedistrictanalysis100
drwxr-xr-x   - acadgild supergroup          0 2017-09-09 15:16 Project2/statedistrictanalysis80
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -ls Project2/statedistrictanalysis80
17/09/09 15:17:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2017-09-09 15:16 Project2/statedistrictanalysis80/_SUCCESS
-rw-r--r--   1 acadgild supergroup      34158 2017-09-09 15:16 Project2/statedistrictanalysis80/part-m-00000
[acadgild@localhost ~]$
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hdfs dfs -cat Project2/statedistrictanalysis80/part-m-00000
Uttarakhand     CHAMOLI 27147   26145   53292   10      265     115     7       2       26043   25844   51887   0       222 7
3       0       95.93
Uttarakhand     DEHRADUN        37212   24463   61675   0       497     19      5       1       31724   24651   56375   0   3
96      19      1       0       85.25
Uttarakhand     HARIDWAR        42500   90074   132574  60      200     20      2       0       38188   77694   115882  37  1
00      13      0       0       89.85
Uttarakhand     NAINITAL        14314   26585   40899   50      208     95      6       2       18039   29025   47064   1   2
08      19      2       0       126.02
Uttarakhand     PITHORAGARH     41110   29253   70363   50      670     215     6       2       34597   31566   66163   10  4
39      102     3       0       84.16
Uttarakhand     RUDRAPRAYAG     13150   22322   35472   40      100     124     3       0       13810   17522   31332   9   1
00      70      2       0       105.02
Uttarakhand     TEHRI GARHWAL   55173   34580   89753   30      307     120     8       0       46293   35076   81369   0   4
86      10      0       0       83.91
Uttarakhand     UDHAM SINGH NAGAR       39427   59658   99085   50      118     123     5       2       37604   51384   88988
0       119     75      4       1       95.38
Uttarakhand     UTTARKASHI      28189   20700   48889   50      240     123     4       0       25523   21017   46540   4   2
40      0       3       0       90.54
West Bengal     BARDHAMAN       700047  341920  1041967 133     9891    7980    10      0       601906  277914  879820  140 9
890     7724    10      19      85.98
West Bengal     DAKSHIN DINAJPUR        182621  194577  377198  50      3712    2642    10      0       184153  49448   23360
1       19      2632    939     8       8       100.84
West Bengal     HOOGHLY 271737  195510  467247  53      6821    4168    19      0       269779  191294  461073  49      67643
435     18      18      99.28
West Bengal     HOWRAH  231860  143309  375169  51      5195    3586    26      0       230190  141912  372102  42      51782
733     14      20      99.28
West Bengal     JALPAIGURI      372999  203523  576522  50      6578    5428    87      0       337740  101550  439290  25  6
578     4064    17      14      90.55
West Bengal     MIDNAPUR EAST   392371  32617   424988  172     9726    5969    25      0       527389  32642   560031  210 1
0149    2882    8       17      134.41
West Bengal     MIDNAPUR WEST   509496  432096  941592  50      16498   5825    10      0       596291  322659  918950  73  1
3452    2787    0       0       117.04
West Bengal     NADIA   346696  278335  625031  50      6974    6620    50      0       321462  198890  520352  28      66353
961     17      41      92.72
West Bengal     NORTH 24 PARAGANAS      361462  225080  586542  51      11158   4466    30      0       357960  226104  58406
4       66      10931   3150    101     0       99.03
West Bengal     SOUTH 24 PARAGANAS      628712  521192  1149904 50      8940    5448    30      0       593712  162487  75619
9       31      7257    1631    29      29      94.43
[acadgild@localhost ~]$
```

- Now we export the data in the HDFS to a Table in MySQL by the following steps:
  - Start the MySQL service and terminal and create the database and table to hold the data Here my table is named **SD_Analysis_80**

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
mysql> USE MiniProject;
Database changed
mysql> CREATE TABLE SD_Analysis_80
    -> (State_Name varchar(30),
    -> District_Name varchar(30),
    -> PO_IHHL_BPL int,
    -> PO_IHHL_APL int,
    -> PO_IHHL_TOTAL int,
    -> PO_SCW int,
    -> PO_School_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_School_Toilets int,
    -> PP_Anganwadi_Toilets int,
    -> PP_RSM int,
    -> PP_PC int,
    -> BPL_Percentage double);
Query OK, 0 rows affected (0.00 sec)

mysql> SELECT * FROM SD_Analysis_80;
Empty set (0.00 sec)
```

- Using the Sqoop command given below:
  - ✓ Specifying the name of the database to hold the data
  - ✓ Specifying the password of the VM (Can also be manually entered or got from a password file)
  - ✓ Specifying the name of the table to hold the data
  - ✓ Specifying the directory in the HDFS that holds the data
  - ✓ Specifying how the fields are terminated (tab separated)
  - ✓ Specifying the number of MapReduce jobs :1
  - ✓ Specifying the column names to import to the MySQL table (Only some of all the columns that are present in the HDFS are exported)

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/MiniProject --username 'root' -password cloudera
--table 'SD_Analysis_80' --export-dir 'statedistrictanalysis80' --input-fields-terminated-by '\t' -m 1
17/09/09 05:24:07 INFO mapreduce.Job: Job job_1504933809120_0006 completed successfully
17/09/09 05:24:07 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=151487
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=34324
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=10158
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=10158
                Total vcore-milliseconds taken by all map tasks=10158
                Total megabyte-milliseconds taken by all map tasks=10401792
        Map-Reduce Framework
                Map input records=349
                Map output records=349
                Input split bytes=163
                Spilled Records=0
```

```
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=162
                CPU time spent (ms)=1460
                Physical memory (bytes) snapshot=135442432
                Virtual memory (bytes) snapshot=1509138432
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/09/09 05:24:07 INFO mapreduce.ExportJobBase: Transferred 33.5195 KB in 32.3961 seconds (1.0347 KB/sec)
17/09/09 05:24:07 INFO mapreduce.ExportJobBase: Exported 349 records.
[cloudera@quickstart ~]$
```

The file has been successfully written to the MySQL table **SD_Analysis_80**

OUTPUT:

- To check the contents of the MySQL table **SD_Analysis_80** use the **SELECT *** command

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
mysql> SELECT * FROM SD_Analysis_80;
| West Bengal       | HOOGHLY                        |      271737 |      195510 |        467247 |     53 |
 6821 |                4168 |     19 |     0 |      269779 |      191294 |        461073 |     49 |              6764 |
                3435 |     18 |    18 |          99.28 |
| West Bengal       | HOWRAH                         |      231860 |      143309 |        375169 |     51 |
 5195 |                3586 |     26 |     0 |      230190 |      141912 |        372102 |     42 |              5178 |
                2733 |     14 |    20 |          99.28 |
| West Bengal       | JALPAIGURI                     |      372999 |      203523 |        576522 |     50 |
 6578 |                5428 |     87 |     0 |      337740 |      101550 |        439290 |     25 |              6578 |
                4064 |     17 |    14 |          90.55 |
| West Bengal       | MIDNAPUR EAST                  |      392371 |       32617 |        424988 |    172 |
 9726 |                5969 |     25 |     0 |      527389 |       32642 |        560031 |    210 |             10149 |
                2882 |      8 |    17 |         134.41 |
| West Bengal       | MIDNAPUR WEST                  |      509496 |      432096 |        941592 |     50 |
16498 |                5825 |     10 |     0 |      596291 |      322659 |        918950 |     73 |             13452 |
                2787 |      0 |     0 |         117.04 |
| West Bengal       | NADIA                          |      346696 |      278335 |        625031 |     50 |
 6974 |                6620 |     50 |     0 |      321462 |      198890 |        520352 |     28 |              6635 |
                3961 |     17 |    41 |          92.72 |
| West Bengal       | NORTH 24 PARAGANAS             |      361462 |      225080 |        586542 |     51 |
11158 |                4466 |     30 |     0 |      357960 |      226104 |        584064 |     66 |             10931 |
                3150 |    101 |     0 |          99.03 |
| West Bengal       | SOUTH 24 PARAGANAS             |      628712 |      521192 |       1149904 |     50 |
 8940 |                5448 |     30 |     0 |      593712 |      162487 |        756199 |     31 |              7257 |
                1631 |     29 |    29 |          94.43 |
+-------------------+--------------------------------+-------------+-------------+---------------+--------+-------------
------+----------------------+--------+-------+-------------+-------------+---------------+--------+-------------------+
----------------------+--------+-------+----------------+
349 rows in set (0.00 sec)

mysql>
```

- Using the below command you can check for specific columns in the table

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
mysql> SELECT State_Name,District_Name,PO_IHHL_BPL,PP_IHHL_BPL,BPL_Percentage FROM SD_Analysis_80;
```

Uttar Pradesh	MORADABAD	76018	76018	100
Uttar Pradesh	MUZAFFARNAGAR	51660	51660	100
Uttar Pradesh	PILIBHIT	95178	95178	100
Uttar Pradesh	PRATAPGARH	141368	143517	101.52
Uttar Pradesh	RAE BARELI	190306	190430	100.07
Uttar Pradesh	RAMPUR	56948	51954	91.23
Uttar Pradesh	SAHARANPUR	49458	49586	100.26
Uttar Pradesh	SANT RAVIDAS NAGAR(BHADOHI)	75119	69904	93.06
Uttar Pradesh	SHAHJAHANPUR	194645	194959	100.16
Uttar Pradesh	SHRAVASTI	104902	98761	94.15
Uttar Pradesh	SIDDHARTHNAGAR	139597	124597	89.25
Uttar Pradesh	SITAPUR	305299	273463	89.57
Uttar Pradesh	SONBHADRA	138370	138370	100
Uttar Pradesh	SULTANPUR	168843	168843	100
Uttar Pradesh	UNNAO	229599	223630	97.4
Uttar Pradesh	VARANASI	105408	102430	97.17
Uttarakhand	BAGESHWAR	21447	17859	83.27
Uttarakhand	CHAMOLI	27147	26043	95.93
Uttarakhand	DEHRADUN	37212	31724	85.25
Uttarakhand	HARIDWAR	42500	38188	89.85
Uttarakhand	NAINITAL	14314	18039	126.02
Uttarakhand	PITHORAGARH	41110	34597	84.16
Uttarakhand	RUDRAPRAYAG	13150	13810	105.02
Uttarakhand	TEHRI GARHWAL	55173	46293	83.91
Uttarakhand	UDHAM SINGH NAGAR	39427	37604	95.38
Uttarakhand	UTTARKASHI	28189	25523	90.54
West Bengal	BARDHAMAN	700047	601906	85.98
West Bengal	DAKSHIN DINAJPUR	182621	184153	100.84
West Bengal	HOOGHLY	271737	269779	99.28
West Bengal	HOWRAH	231860	230190	99.28
West Bengal	JALPAIGURI	372999	337740	90.55
West Bengal	MIDNAPUR EAST	392371	527389	134.41
West Bengal	MIDNAPUR WEST	509496	596291	117.04
West Bengal	NADIA	346696	321462	92.72
West Bengal	NORTH 24 PARAGANAS	361462	357960	99.03
West Bengal	SOUTH 24 PARAGANAS	628712	593712	94.43

349 rows in set (0.00 sec)

mysql>