ACADGILD

Big Data & Hadoop Training

Project 1.2- State-Wise Development Analysis In India

Ankita Jaiswal

11/29/2017

Table of Contents

1. Executive Summary

✓ **Project Overview**

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

✓ **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:

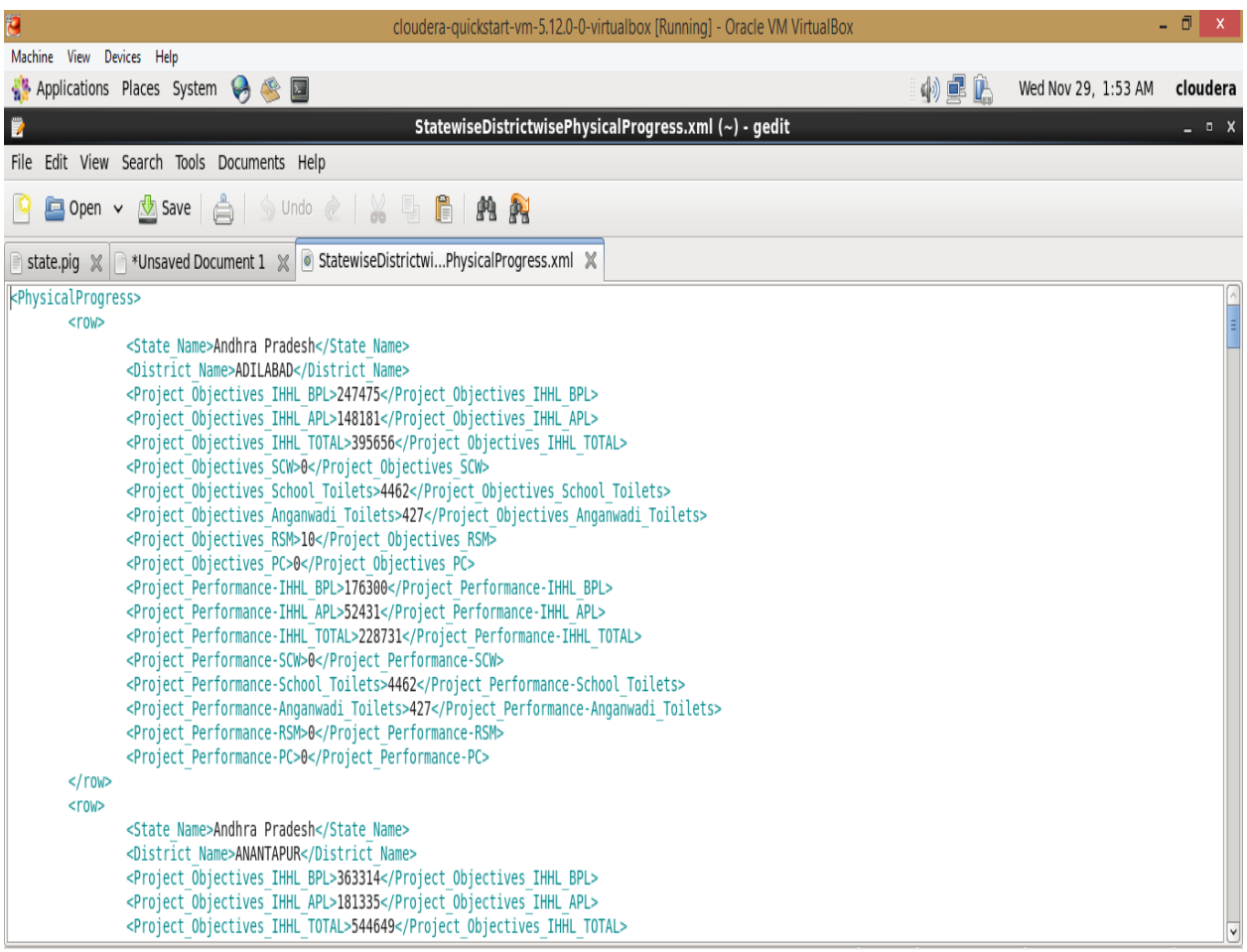
- Exporting the Data from the Local File System to the HDFS using Flume
- Performing Analysis on the data (in xml form) using PIG to get results for the below problem statements:
    - Find out the districts who achieved 100 percent objective in BPL cards
      Export the results to MySQL using Sqoop
    - 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.




```
<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>
```

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

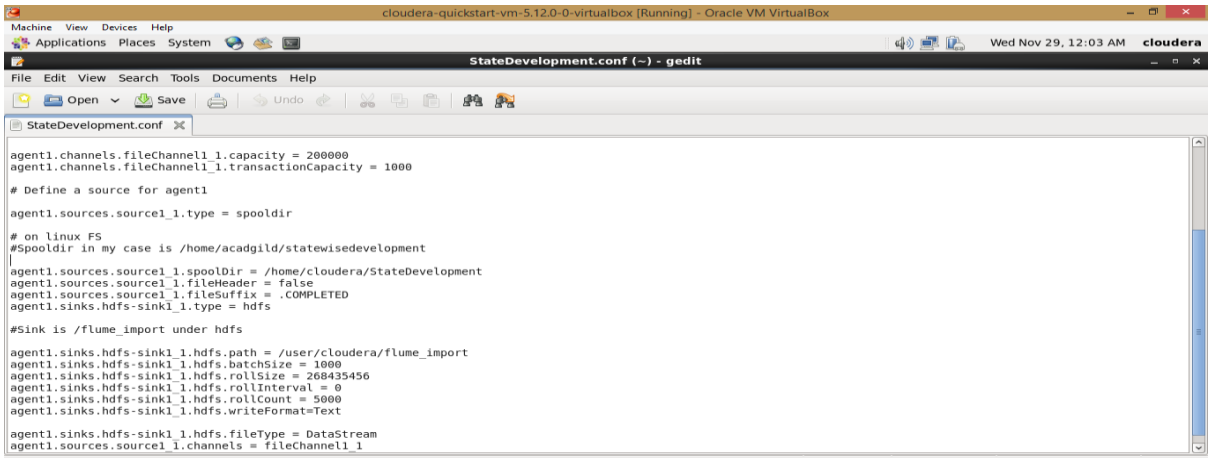
To perform this task we have to execute the following steps:

- Download Apache Flume for the cloudera VM and extract it Update the location of Apache Flume in the .bashrc file

- Create the configuration document for the flume job. This will contain the necessary information for **fetching** and **storing flume_import files in the HDFS.**

- I named the flume configuration file as **'StateDevelopment.conf'**

```
#Specify source,channel and sink
agent1.sinks =  hdfs-sink1_1
agent1.sources = source1_1
agent1.channels = fileChannel1_1
#Flume Configuration Starts
# Define a file channel called fileChannel on agent1
agent1.channels.fileChannel1_1.type = file
# on linux FS
agent1.channels.fileChannel1_1.capacity = 200000
agent1.channels.fileChannel1_1.transactionCapacity = 1000
# Define a source for agent1
agent1.sources.source1_1.type = spooldir
# on linux FS
#Spooldir in my case is /home/cloudera/StateDevelopment
agent1.sources.source1_1.spoolDir = /home/cloudera/StateDevelopment
agent1.sources.source1_1.fileHeader = false
agent1.sources.source1_1.fileSuffix = .COMPLETED
agent1.sinks.hdfs-sink1_1.type = hdfs
#Sink is /flume_import under hdfs
agent1.sinks.hdfs-sink1_1.hdfs.path = /user/cloudera/flume_import
agent1.sinks.hdfs-sink1_1.hdfs.batchSize = 1000
agent1.sinks.hdfs-sink1_1.hdfs.rollSize = 268435456
agent1.sinks.hdfs-sink1_1.hdfs.rollInterval = 0
agent1.sinks.hdfs-sink1_1.hdfs.rollCount = 5000
agent1.sinks.hdfs-sink1_1.hdfs.writeFormat=Text
agent1.sinks.hdfs-sink1_1.hdfs.fileType = DataStream
agent1.sources.source1_1.channels = fileChannel1_1
agent1.sinks.hdfs-sink1_1.channel = fileChannel1_1
```



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

```
[cloudera@quickstart ~]$ hadoop dfs -mkdir /user/cloudera/flume_import
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

- 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>

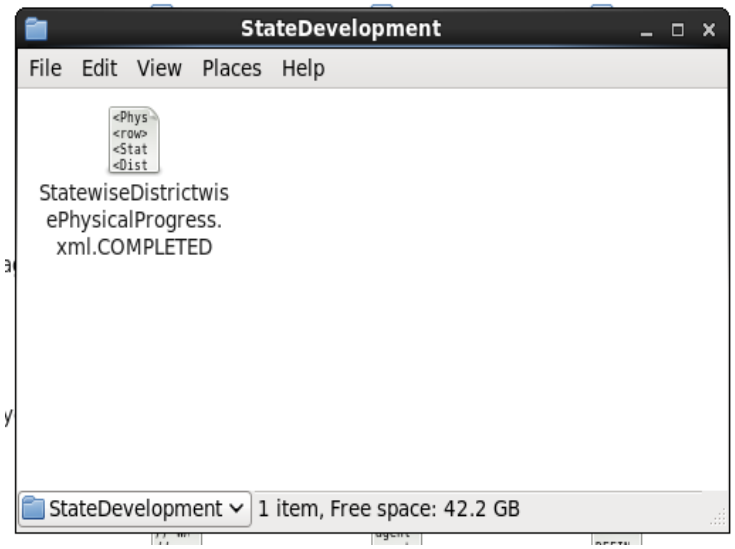
```
[cloudera@quickstart ~]$ flume-ng agent --conf-file StateDevelopment.conf --name agent1 --conf $FLUME_HOME/conf -Dflume.root.logger=INFO,console
Info: Including Hadoop libraries found via (/usr/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/usr/bin/hbase) for HBASE access
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java/jdk1.7.0_67-cloudera/bin/java -Xmx20m -Dflume.root.logger=INFO,console -cp '/home/cloudera/Downloads/apache-flume-1.6.0-bin/conf:/home/cloudera/Downloa
ds/apache-flume-1.6.0-bin/lib/*:/etc/hadoop/conf:/usr/lib/hadoop/lib/*:/usr/lib/hadoop/.//*:/usr/lib/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/.//*
:/usr/lib/hadoop-yarn/lib/*:/usr/lib/hadoop-yarn/.//*:/usr/lib/hadoop-mapreduce/lib/*:/usr/lib/hadoop-mapreduce/.//*:/usr/lib/hbase/bin/../conf:/usr/java/jdk1.7.0_67-cl
oudera/lib/tools.jar:/usr/lib/hbase/bin/..:/usr/lib/hbase/bin/../lib/activation-1.1.jar:/usr/lib/hbase/bin/../lib/apacheds-i18n-2.0.0-M15.jar:/usr/lib/hbase/bin/../lib/
apacheds-kerberos-codec-2.0.0-M15.jar:/usr/lib/hbase/bin/../lib/api-asn1-api-1.0.0-M20.jar:/usr/lib/hbase/bin/../lib/api-util-1.0.0-M20.jar:/usr/lib/hbase/bin/../lib/as
m-3.2.jar:/usr/lib/hbase/bin/../lib/avro.jar:/usr/lib/hbase/bin/../lib/aws-java-sdk-bundle-1.11.134.jar:/usr/lib/hbase/bin/../lib/commons-beanutils-1.9.2.jar:/usr/lib/h
base/bin/../lib/commons-beanutils-core-1.8.0.jar:/usr/lib/hbase/bin/../lib/commons-cli-1.2.jar:/usr/lib/hbase/bin/../lib/commons-codec-1.9.jar:/usr/lib/hbase/bin/../lib
/commons-collections-3.2.2.jar:/usr/lib/hbase/bin/../lib/commons-compress-1.4.1.jar:/usr/lib/hbase/bin/../lib/commons-configuration-1.6.jar:/usr/lib/hbase/bin/../lib/co
mmons-daemon-1.0.13.jar:/usr/lib/hbase/bin/../lib/commons-digester-1.8.jar:/usr/lib/hbase/bin/../lib/commons-el-1.0.jar:/usr/lib/hbase/bin/../lib/commons-httpclient-3.1
.jar:/usr/lib/hbase/bin/../lib/commons-io-2.4.jar:/usr/lib/hbase/bin/../lib/commons-lang-2.6.jar:/usr/lib/hbase/bin/../lib/commons-logging-1.2.jar:/usr/lib/hbase/bin/..
/lib/commons-math-2.1.jar:/usr/lib/hbase/bin/../lib/commons-math3-3.1.1.jar:/usr/lib/hbase/bin/../lib/commons-net-3.1.jar:/usr/lib/hbase/bin/../lib/core-3.1.1.jar:/usr/
lib/hbase/bin/../lib/curator-client-2.7.1.jar:/usr/lib/hbase/bin/../lib/curator-framework-2.7.1.jar:/usr/lib/hbase/bin/../lib/curator-recipes-2.7.1.jar:/usr/lib/hbase/b
in/../lib/disruptor-3.3.0.jar:/usr/lib/hbase/bin/../lib/findbugs-annotations-1.3.9-1.jar:/usr/lib/hbase/bin/../lib/gson-2.2.4.jar:/usr/lib/hbase/bin/../lib/guava-12.0.1
.jar:/usr/lib/hbase/bin/../lib/hamcrest-core-1.3.jar:/usr/lib/hbase/bin/../lib/hbase-annotations-1.2.0-cdh5.12.0.jar:/usr/lib/hbase/bin/../lib/hbase-annotations-1.2.0-c
dh5.12.0-tests.jar:/usr/lib/hbase/bin/../lib/hbase-client-1.2.0-cdh5.12.0.jar:/usr/lib/hbase/bin/../lib/hbase-common-1.2.0-cdh5.12.0.jar:/usr/lib/hbase/bin/../lib/hbase
2017-11-28 22:43:15,194 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO - org.apache.flume.sink.hdfs.BucketWriter.open(BucketWriter.java:234)] Creating /user/clou
dera/flume_import/FlumeData.1511937762898.tmp
2017-11-28 22:43:25,666 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO - org.apache.flume.sink.hdfs.BucketWriter.close(BucketWriter.java:363)] Closing /user/clou
dera/flume_import/FlumeData.1511937762898.tmp
2017-11-28 22:43:25,882 (hdfs-hdfs-sink1_1-call-runner-1) [INFO - org.apache.flume.sink.hdfs.BucketWriter$8.call(BucketWriter.java:629)] Renaming /user/cloudera/flume_i
mport/FlumeData.1511937762898.tmp to /user/cloudera/flume_import/FlumeData.1511937762898
2017-11-28 22:43:26,329 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO - org.apache.flume.sink.hdfs.BucketWriter.open(BucketWriter.java:234)] Creating /user/clou
dera/flume_import/FlumeData.1511937762899.tmp
2017-11-28 22:43:29,936 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO - org.apache.flume.sink.hdfs.BucketWriter.close(BucketWriter.java:363)] Closing /user/clou
dera/flume_import/FlumeData.1511937762899.tmp
2017-11-28 22:43:29,975 (hdfs-hdfs-sink1_1-call-runner-7) [INFO - org.apache.flume.sink.hdfs.BucketWriter$8.call(BucketWriter.java:629)] Renaming /user/cloudera/flume_i
mport/FlumeData.1511937762899.tmp to /user/cloudera/flume_import/FlumeData.1511937762899
2017-11-28 22:43:30,416 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO - org.apache.flume.sink.hdfs.BucketWriter.open(BucketWriter.java:234)] Creating /user/clou
dera/flume_import/FlumeData.1511937762900.tmp
2017-11-28 22:43:37,957 (Log-BackgroundWorker-fileChannel1_1) [INFO - org.apache.flume.channel.file.EventQueueBackingStoreFile.beginCheckpoint(EventQueueBackingStoreFil
e.java:230)] Start checkpoint for /home/cloudera/.flume/file-channel/checkpoint/checkpoint, elements to sync = 12142
2017-11-28 22:43:37,979 (Log-BackgroundWorker-fileChannel1_1) [INFO - org.apache.flume.channel.file.EventQueueBackingStoreFile.checkpoint(EventQueueBackingStoreFile.jav
a:255)] Updating checkpoint metadata: logWriteOrderID: 1511937782936, queueSize: 0, queueHead: 12140
2017-11-28 22:43:38,082 (Log-BackgroundWorker-fileChannel1_1) [INFO - org.apache.flume.channel.file.Log.writeCheckpoint(Log.java:1034)] Updated checkpoint for file: /ho
me/cloudera/.flume/file-channel/data/log-1 position: 1742347 logWriteOrderID: 1511937782936
```

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

```
[cloudera@quickstart ~]$ hadoop dfs -ls /user/cloudera/flume_import
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 4 items
-rw-r--r--   1 cloudera cloudera          0 2017-11-28 22:43 /user/cloudera/flum
e_import/FlumeData.1511937762897.tmp
-rw-r--r--   1 cloudera cloudera     295196 2017-11-28 22:43 /user/cloudera/flum
e_import/FlumeData.1511937762898
-rw-r--r--   1 cloudera cloudera     295013 2017-11-28 22:43 /user/cloudera/flum
e_import/FlumeData.1511937762899
-rw-r--r--   1 cloudera cloudera      59456 2017-11-28 22:43 /user/cloudera/flum
e_import/FlumeData.1511937762900.tmp
[cloudera@quickstart ~]$
```

Once the export is complete the folder where the dataset is kept will show as completed and the xml file has been successfully exported



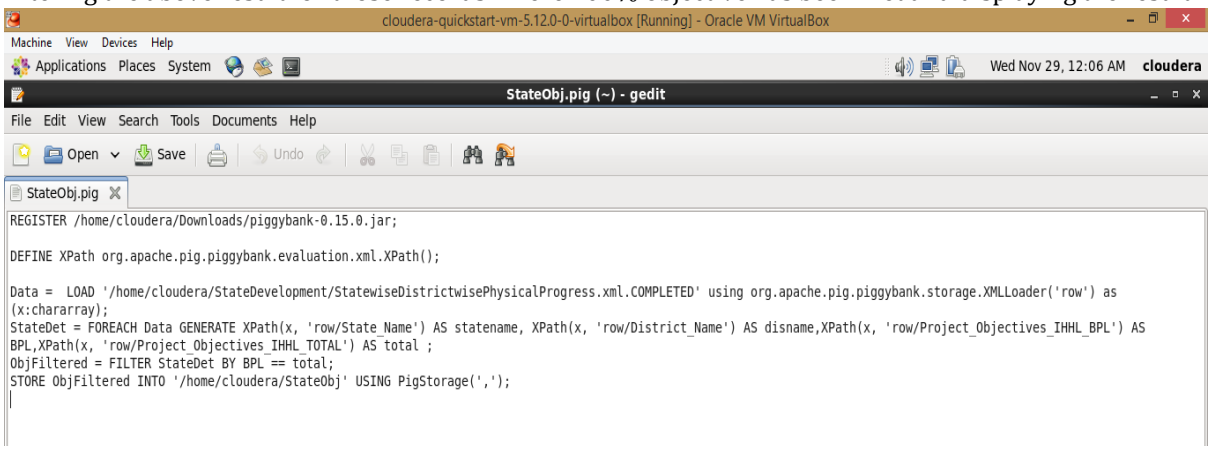
Performing Analysis on the data (in xml form) using PIG

Problem 1-----

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

Here I created a Pig File with filename "**StateObj.pig**" and below steps explain how the file works.

- Starting the Pig Shell using the command **pig**
- Registering the **piggybank** jar that contains the executables for various pig functions. Ex: Parse XML
- Defining the XML Parse function as **XPath** (name used to call the function)
- Loading the data in the LFS 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 **x** Generating the rows (x) in relation Data 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 column names pertaining to **statename,disname,BPL and total** and finding the **Percentage(bpl == total)** 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.



- The result of the above procedure:

```
[cloudera@quickstart ~]$ pig -x local StateObj.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2017-11-28 23:15:54,096 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.12.0 (rexported) compiled Jun 29 2017, 04:34:31
2017-11-28 23:15:54,098 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1511939753992.log
2017-11-28 23:15:54,191 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2017-11-28 23:15:59,650 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/cloudera/.pigbootup not found
2017-11-28 23:16:00,554 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-28 23:16:00,560 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-28 23:16:00,615 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2017-11-28 23:16:01,858 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-28 23:16:01,910 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-28 23:17:18,101 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - map
2017-11-28 23:17:18,102 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task 'attempt_local721053020_0001_m_000000_0' done.
2017-11-28 23:17:18,105 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local721053020_0001_m_000000_0
2017-11-28 23:17:18,105 [Thread-7] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2017-11-28 23:17:19,641 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2017-11-28 23:17:19,642 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2017-11-28 23:17:25,645 [main] WARN  org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local721053020_0001
2017-11-28 23:17:25,679 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2017-11-28 23:17:25,681 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2017-11-28 23:17:25,712 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2017-11-28 23:16:11     2017-11-28 23:17:25     FILTER

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local721053020_0001 Data,ObjFiltered,StateDet       MAP_ONLY        /home/cloudera/StateObj,

Input(s):
Successfully read records from: "/home/cloudera/StateDevelopment/StatewiseDistrictwisePhysicalProgress.xml.COMPLETED"

Output(s):
Successfully stored records in: "/home/cloudera/StateObj"

Job DAG:
job_local721053020_0001


2017-11-28 23:17:31,719 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
[cloudera@quickstart ~]$
```

- Checking the data was successfully stored at **/home/cloudera/StateObj**

```
[cloudera@quickstart ~]$ ls /home/cloudera/StateObj
part-m-00000  _SUCCESS
[cloudera@quickstart ~]$ cat /home/cloudera/StateObj/p*
Arunachal Pradesh,ANJAW,3232,3232
Arunachal Pradesh,DIBANG VALLEY,1085,1085
Arunachal Pradesh,KURUNG KUMEY,22036,22036
Arunachal Pradesh,LOHIT,8800,8800
Arunachal Pradesh,WEST SIANG,11472,11472
Bihar,BANKA,82439,82439
D & N Haveli,DADRA AND NAGAR HAVELI,2480,2480
Goa,NORTH GOA,15000,15000
Jammu & Kashmir,KARGIL,8475,8475
Jammu & Kashmir,KISHTWAR,22318,22318
Jammu & Kashmir,LEH (LADAKH),6090,6090
Jammu & Kashmir,REASI                                    ,21500,21500
Jammu & Kashmir,SAMBA                                     ,9849,9849
Jammu & Kashmir,SHOPIAN                                   ,10196,10196
Kerala,KANNUR,34121,34121
Manipur,CHANDEL,17610,17610
Nagaland,LONGLENG,6438,6438
Nagaland,TUENSANG,13027,13027
Nagaland,ZUNHEBOTO,20570,20570
Puducherry,PONDICHERRY,18000,18000
Punjab,FARIDKOT,6000,6000
Punjab,HOSHIARPUR,11112,11112
Punjab,MOGA,37170,37170
Punjab,MUKTSAR,33148,33148
[cloudera@quickstart ~]$
```

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

```
[cloudera@quickstart ~]$ hadoop dfs -put StateObj /user/cloudera
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

- 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

```
[cloudera@quickstart ~]$ hadoop dfs -ls /user/cloudera/StateObj
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2017-11-28 23:21 /user/cloudera/StateObj/_SUCCESS
-rw-r--r--   1 cloudera cloudera        889 2017-11-28 23:21 /user/cloudera/StateObj/part-m-00000
[cloudera@quickstart ~]$
```

- 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 "state" and table to hold the data. Here my table is named **BPLObjectivesMet**

```
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 54
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 state;
Query OK, 1 row affected (0.00 sec)

mysql> use state;
Database changed
mysql> create table BPLObjectivesMet (State varchar(20), district varchar(50), BPL int, total int);
Query OK, 0 rows affected (0.09 sec)

mysql> show tables;
+------------------+
| Tables_in_state  |
+------------------+
| BPLObjectivesMet |
+------------------+
1 row in set (0.00 sec)
```

- Using the Sqoop command given below:
  - ✓ Specifying the name of the database to hold the data
  - ✓ Specifying the username 'root' and password is entered while executing sqoop command
  - ✓ 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

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/state --username 'root' -P --table BPLObjectivesMet --export-dir '/user/cloudera/StateObj/part-m-
00000' --input-fields-terminated-by ',' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/28 23:31:39 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/11/28 23:31:53 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/11/28 23:31:53 INFO tool.CodeGenTool: Beginning code generation
17/11/28 23:31:57 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `BPLObjectivesMet` AS t LIMIT 1
17/11/28 23:31:57 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `BPLObjectivesMet` AS t LIMIT 1
17/11/28 23:31:57 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/8aac73da32a82c1cc98b8aee6bf480f1/BPLObjectivesMet.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/11/28 23:32:16 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/8aac73da32a82c1cc98b8aee6bf480f1/BPLObjectivesMet.jar
17/11/28 23:32:16 INFO mapreduce.ExportJobBase: Beginning export of BPLObjectivesMet
17/11/28 23:32:16 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
17/11/28 23:32:19 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
17/11/28 23:32:28 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
17/11/28 23:32:28 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
17/11/28 23:32:28 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
17/11/28 23:32:29 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/11/28 23:32:36 WARN hdfs.DFSClient: Caught exception
```

```
                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)=35704
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=35704
                Total vcore-milliseconds taken by all map tasks=35704
                Total megabyte-milliseconds taken by all map tasks=36560896
        Map-Reduce Framework
                Map input records=24
                Map output records=24
                Input split bytes=148
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=677
                CPU time spent (ms)=3440
                Physical memory (bytes) snapshot=127770624
                Virtual memory (bytes) snapshot=1508028416
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/11/28 23:34:16 INFO mapreduce.ExportJobBase: Transferred 1.0156 KB in 107.9205 seconds (9.6367 bytes/sec)
17/11/28 23:34:16 INFO mapreduce.ExportJobBase: Exported 24 records.
[cloudera@quickstart ~]$
```

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

OUTPUT:

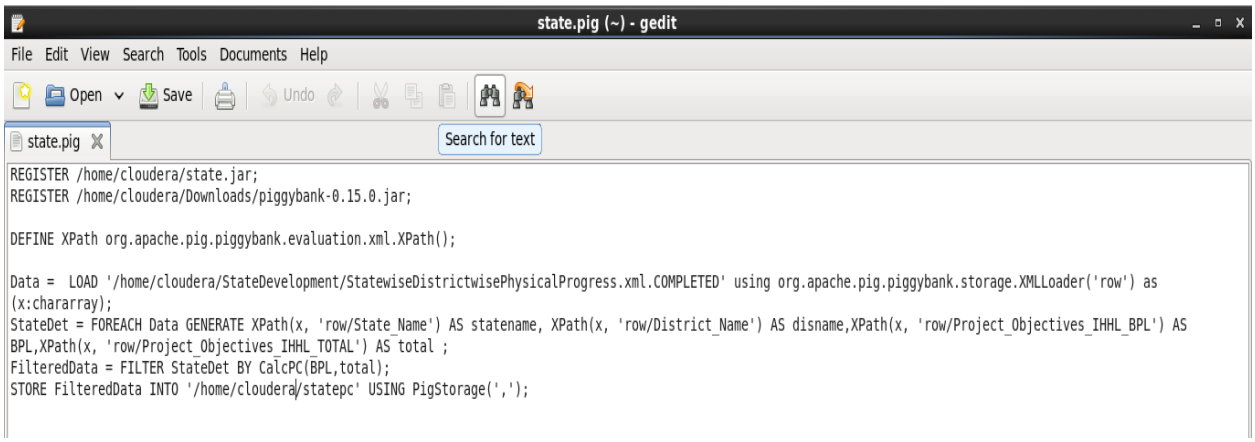
- To check the contents of the MySQL table **BPLObjectivesMet** use the **SELECT *** command

```
mysql> select * from BPLObjectivesMet;
+-------------------+-----------------------------------------+-------+-------+
| State             | district                                | BPL   | total |
+-------------------+-----------------------------------------+-------+-------+
| Arunachal Pradesh | ANJAW                                   |  3232 |  3232 |
| Arunachal Pradesh | DIBANG VALLEY                           |  1085 |  1085 |
| Arunachal Pradesh | KURUNG KUMEY                            | 22036 | 22036 |
| Arunachal Pradesh | LOHIT                                   |  8800 |  8800 |
| Arunachal Pradesh | WEST SIANG                              | 11472 | 11472 |
| Bihar             | BANKA                                   | 82439 | 82439 |
| D & N Haveli      | DADRA AND NAGAR HAVELI                  |  2480 |  2480 |
| Goa               | NORTH GOA                               | 15000 | 15000 |
| Jammu & Kashmir   | KARGIL                                  |  8475 |  8475 |
| Jammu & Kashmir   | KISHTWAR                                | 22318 | 22318 |
| Jammu & Kashmir   | LEH (LADAKH)                            |  6090 |  6090 |
| Jammu & Kashmir   | REASI                                   | 21500 | 21500 |
| Jammu & Kashmir   | SAMBA                                   |  9849 |  9849 |
| Jammu & Kashmir   | SHOPIAN                                 | 10196 | 10196 |
| Kerala            | KANNUR                                  | 34121 | 34121 |
| Manipur           | CHANDEL                                 | 17610 | 17610 |
| Nagaland          | LONGLENG                                |  6438 |  6438 |
| Nagaland          | TUENSANG                                | 13027 | 13027 |
| Nagaland          | ZUNHEBOTO                               | 20570 | 20570 |
| Puducherry        | PONDICHERRY                             | 18000 | 18000 |
| Punjab            | FARIDKOT                                |  6000 |  6000 |
| Punjab            | HOSHIARPUR                              | 11112 | 11112 |
| Punjab            | MOGA                                    | 37170 | 37170 |
| Punjab            | MUKTSAR                                 | 33148 | 33148 |
+-------------------+-----------------------------------------+-------+-------+
24 rows in set (0.00 sec)
```

TASK 2----

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) **"state.pig"** and executed it via the pig shell



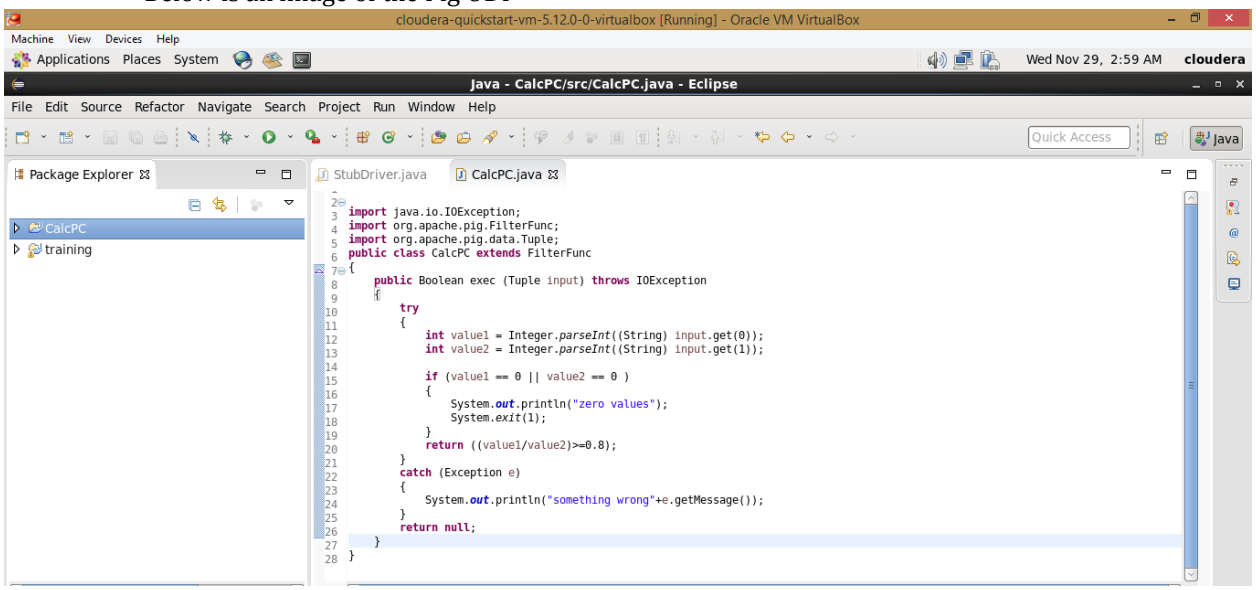
The steps followed are explained as below:

- Registering the piggybank jar that contains the executables for various pig functions. Ex: Parse XML
- Defining the XML Parse function as **XPath** (name used to call the function)
- Registering the Pig UDF "state.jar" created to filter the districts which have reached 80% of objectives of BPL cards. (Written in Java)
- Defining exec as the function to be used to execute the UDF class **CalcPC**

- Loading the data in the HDFS (that was exported using Flume) and using the XML Loader function to load the data into the relation **Data** with every starting tag 'row' as one line of type: chararray with the name **x**
- Generating the rows (x) in relation Data 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 80% and more** 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



- The UDF exported as a jar **"/home/cloudera/state.jar"**
- Filtering the above result for those records that have received 80% and above in BPL cards
- 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

```
[cloudera@quickstart ~]$ pig -x local state.pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2017-11-29 01:08:49,109 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.12.0 (rexported) compiled Jun 29 2017, 04:34:31
2017-11-29 01:08:49,110 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1511946528955.log
2017-11-29 01:08:49,223 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - user.name is deprecated. Instead, use mapreduce.job.user.name
2017-11-29 01:08:54,669 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/cloudera/.pigbootup not found
2017-11-29 01:08:55,516 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-29 01:08:55,523 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-29 01:08:55,559 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2017-11-29 01:08:56,722 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-29 01:08:56,760 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-29 01:08:57,877 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-29 01:08:57,951 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-29 01:08:58,748 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-29 01:08:58,764 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-29 01:08:59,226 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-29 01:08:59,240 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-11-29 01:08:59,576 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
2017-11-29 01:09:59,750 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - map
2017-11-29 01:09:59,750 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task 'attempt_local1570722355_0001_m_000000_0' done.
2017-11-29 01:09:59,750 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner - Finishing task: attempt_local1570722355_0001_m_000000_0
2017-11-29 01:09:59,750 [Thread-7] INFO  org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2017-11-29 01:09:59,974 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2017-11-29 01:09:59,974 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2017-11-29 01:10:05,978 [main] WARN  org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local1570722355_0001
2017-11-29 01:10:05,998 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2017-11-29 01:10:05,998 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2017-11-29 01:10:06,023 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2017-11-29 01:09:08     2017-11-29 01:10:05     FILTER

Success!

Job Stats (time in seconds):
JobId   Alias   Feature Outputs
job_local1570722355_0001        Data,FilteredData,StateDet      MAP_ONLY        /home/cloudera/statepc,

Input(s):
Successfully read records from: "/home/cloudera/StateDevelopment/StatewiseDistrictwisePhysicalProgress.xml.COMPLETED"

Output(s):
Successfully stored records in: "/home/cloudera/statepc"

Job DAG:
job_local1570722355_0001


2017-11-29 01:10:12,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
[cloudera@quickstart ~]$
```

The execution is successful.

- Checking the records is successfully stored in /home/cloudera/statepc folder

```
[cloudera@quickstart ~]$ ls /home/cloudera/statepc
part-m-00000  _SUCCESS
[cloudera@quickstart ~]$ cat /home/cloudera/statepc/p*
Arunachal Pradesh,ANJAW,3232,3232
Arunachal Pradesh,DIBANG VALLEY,1085,1085
Arunachal Pradesh,KURUNG KUMEY,22036,22036
Arunachal Pradesh,LOHIT,8800,8800
Arunachal Pradesh,WEST SIANG,11472,11472
Bihar,BANKA,82439,82439
D & N Haveli,DADRA AND NAGAR HAVELI,2480,2480
Goa,NORTH GOA,15000,15000
Jammu & Kashmir,KARGIL,8475,8475
Jammu & Kashmir,KISHTWAR,22318,22318
Jammu & Kashmir,LEH (LADAKH),6090,6090
Jammu & Kashmir,REASI                                    ,21500,21500
Jammu & Kashmir,SAMBA                                     ,9849,9849
Jammu & Kashmir,SHOPIAN                                   ,10196,10196
Kerala,KANNUR,34121,34121
Manipur,CHANDEL,17610,17610
Nagaland,LONGLENG,6438,6438
Nagaland,TUENSANG,13027,13027
Nagaland,ZUNHEBOTO,20570,20570
Puducherry,PONDICHERRY,18000,18000
Punjab,FARIDKOT,6000,6000
Punjab,HOSHIARPUR,11112,11112
Punjab,MOGA,37170,37170
Punjab,MUKTSAR,33148,33148
[cloudera@quickstart ~]$
```

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

```
[cloudera@quickstart ~]$ hadoop dfs -put statepc /user/cloudera
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

[cloudera@quickstart ~]$ hadoop dfs -ls /user/cloudera/statepc
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2017-11-29 01:15 /user/cloudera/statepc/_SUCCESS
-rw-r--r--   1 cloudera cloudera        889 2017-11-29 01:15 /user/cloudera/statepc/part-m-00000
[cloudera@quickstart ~]$
```

- Now we export the data in the HDFS to a Table in MySQL by the following steps:

  - Start the MySQL service and terminal and use database 'state' and create table to hold the data
    Here my table is named "state80percent"

```
mysql> create table state80percent (State varchar(20), district varchar(50), BPL int, total int);
Query OK, 0 rows affected (0.05 sec)

mysql> show tables;
+------------------+
| Tables_in_state  |
+------------------+
| BPLObjectivesMet |
| state80percent   |
+------------------+
2 rows in set (0.01 sec)
```

  - Using the Sqoop command given below:
    - ✓ Specifying the name of the database to hold the data
    - ✓ Specifying the username 'root' and password is entered while executing sqoop command
    - ✓ 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

```
[cloudera@quickstart ~]$ sqoop export --connect jdbc:mysql://localhost/state --username 'root' -P --table state80percent --export-dir '/user/cloudera/statepc/part-m-000
00' --input-fields-terminated-by ',' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/11/29 01:19:52 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
Enter password:
17/11/29 01:20:04 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/11/29 01:20:04 INFO tool.CodeGenTool: Beginning code generation
17/11/29 01:20:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `state80percent` AS t LIMIT 1
17/11/29 01:20:08 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `state80percent` AS t LIMIT 1
17/11/29 01:20:08 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/01fc484f4dabfaab8947035e04737732/state80percent.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/11/29 01:20:22 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/01fc484f4dabfaab8947035e04737732/state80percent.jar
17/11/29 01:20:22 INFO mapreduce.ExportJobBase: Beginning export of state80percent
17/11/29 01:20:22 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
17/11/29 01:20:25 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
17/11/29 01:20:31 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
17/11/29 01:20:31 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
17/11/29 01:20:31 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
17/11/29 01:20:32 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/11/29 01:20:35 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
17/11/29 01:20:37 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
```

```
            HDFS: Number of bytes read=1039
            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)=25224
            Total time spent by all reduces in occupied slots (ms)=0
            Total time spent by all map tasks (ms)=25224
            Total vcore-milliseconds taken by all map tasks=25224
            Total megabyte-milliseconds taken by all map tasks=25829376
    Map-Reduce Framework
            Map input records=24
            Map output records=24
            Input split bytes=147
            Spilled Records=0
            Failed Shuffles=0
            Merged Map outputs=0
            GC time elapsed (ms)=205
            CPU time spent (ms)=2020
            Physical memory (bytes) snapshot=129044480
            Virtual memory (bytes) snapshot=1508319232
            Total committed heap usage (bytes)=60751872
    File Input Format Counters
            Bytes Read=0
    File Output Format Counters
            Bytes Written=0
17/11/29 01:21:56 INFO mapreduce.ExportJobBase: Transferred 1.0146 KB in 84.948 seconds (12.231 bytes/sec)
17/11/29 01:21:56 INFO mapreduce.ExportJobBase: Exported 24 records.
[cloudera@quickstart ~]$
```

The file has been successfully exported to the MySQL table **state80percent**

OUTPUT:

- To check the contents of the MySQL table **state80percent** use the **SELECT *** command

```
mysql> select * from state80percent;
+--------------------+-----------------------------------------+-------+-------+
| State              | district                                | BPL   | total |
+--------------------+-----------------------------------------+-------+-------+
| Arunachal Pradesh  | ANJAW                                   |  3232 |  3232 |
| Arunachal Pradesh  | DIBANG VALLEY                           |  1085 |  1085 |
| Arunachal Pradesh  | KURUNG KUMEY                            | 22036 | 22036 |
| Arunachal Pradesh  | LOHIT                                   |  8800 |  8800 |
| Arunachal Pradesh  | WEST SIANG                              | 11472 | 11472 |
| Bihar              | BANKA                                   | 82439 | 82439 |
| D & N Haveli       | DADRA AND NAGAR HAVELI                  |  2480 |  2480 |
| Goa                | NORTH GOA                               | 15000 | 15000 |
| Jammu & Kashmir    | KARGIL                                  |  8475 |  8475 |
| Jammu & Kashmir    | KISHTWAR                                | 22318 | 22318 |
| Jammu & Kashmir    | LEH (LADAKH)                            |  6090 |  6090 |
| Jammu & Kashmir    | REASI                                   | 21500 | 21500 |
| Jammu & Kashmir    | SAMBA                                   |  9849 |  9849 |
| Jammu & Kashmir    | SHOPIAN                                 | 10196 | 10196 |
| Kerala             | KANNUR                                  | 34121 | 34121 |
| Manipur            | CHANDEL                                 | 17610 | 17610 |
| Nagaland           | LONGLENG                                |  6438 |  6438 |
| Nagaland           | TUENSANG                                | 13027 | 13027 |
| Nagaland           | ZUNHEBOTO                               | 20570 | 20570 |
| Puducherry         | PONDICHERRY                             | 18000 | 18000 |
| Punjab             | FARIDKOT                                |  6000 |  6000 |
| Punjab             | HOSHIARPUR                              | 11112 | 11112 |
| Punjab             | MOGA                                    | 37170 | 37170 |
| Punjab             | MUKTSAR                                 | 33148 | 33148 |
+--------------------+-----------------------------------------+-------+-------+
24 rows in set (0.00 sec)
```