

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

Download the dataset using the below link:

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

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

Step 1: Copy dataset from local file system to HDFS using flume. Note: use the conf file by downloading from below link. Click here to download Command: flume-agent agent –n agent1 –c conf –f

Step 2: Input file is in the XML format use Map reduce or pig to parse the data and get the results for the below problem statements.

4. Problem statement

1. Find out the districts who achieved 100 percent objective in BPL cards Export the results to mysql using sqoop
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.

Solution

Scripting the flume conf file

The screenshot shows a terminal window titled "flumeprj.conf" containing Flink configuration code. The code defines an agent (agent1) with a source (mysrc), sink (hdfsdest), and channel (mychannel). The source is an exec type that runs a hadoop dfs command to read from a local XML file. The sink is an hdfs type pointing to a local host at port 9000. The channel is a memory type.

```
agent1.sources = mysrc
agent1.sinks = hdfsdest
agent1.channels = mychannel

agent1.sources.mysrc.type = exec
agent1.sources.mysrc.command = hadoop dfs -put /home/acadgild/Desktop/StatewiseDistrictwisePhysicalProgress.xml /user/acadgild/flumeimp

agent1.sinks.hdfsdest.type = hdfs
agent1.sinks.hdfsdest.hdfs.path = hdfs://localhost:9000/flumeimp

agent1.channels.mychannel.type = memory

agent1.sources.mysrc.channels = mychannel
agent1.sinks.hdfsdest.channel = mychannel
```

Executing the flume conf file in terminal

```
[acadgild@localhost Desktop]$ flume-ng agent -n agent1 -c conf -f /home/acadgild/Desktop/flumeprj.conf
Info: Including Hadoop libraries found via /home/acadgild/hadoop-2.7.2/bin/hadoop for HDFS access
Info: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath
Info: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath
Info: Including HBASE libraries found via /home/acadgild/hbase-1.0.3/bin/hbase) for HBASE access
Info: Excluding /home/acadgild/hbase-1.0.3/lib/slf4j-api-1.7.7.jar from classpath
Info: Excluding /home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar from classpath
Info: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath
Info: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath
Info: Including Hive libraries found via /home/acadgild/apache-hive-2.1.0-bin) for Hive access
+ exec /home/acadgild/jdk1.8.0_101/bin/java -Xmx20m -cp "conf:/home/acadgild/apache-flume-1.6.0-bin/lib/*:/home/acadgild/hadoop-2.7.2/contrib/capacity-scheduler/.jar:/home/acadgild/hadoop-2.7.2/etc/hadoop:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/apachesn-kerberos-codec-2.0.0-M15.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/asn-3.2.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/avro-1.7.4.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-client-1.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-compress-1.4.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-configuration-1.6.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-digester-1.8.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-discovery-0.2.4.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-lang-2.6.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-math3-3.1.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-net-3.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/curator-client-2.7.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/curator-framework-2.7.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/curator-recipes-2.7.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/gson-2.2.4.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/guava-11.0.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-annotations-2.7.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-auth-2.7.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hmcrcrest-core-1.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/htrace-core-3.1.0-incubating.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/httpclient-4.2.5.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jaxb-impl-2.2.3-1.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jersey-core-1.9.jar:/home/acadgild/Mozilla Firefox 7.2/share/hadoop/common/lib/jersey-json-1.9.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jers
```

Flume conf file execution success status

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
        at org.apache.flume.node.PropertiesFileConfigurationProvider.getFlumeConfiguration(PropertiesFileConfigurationProvide
r.java:189)
        at org.apache.flume.node.AbstractConfigurationProvider.getConfiguration(AbstractConfigurationProvider.java:89)
        at org.apache.flume.node.PollingPropertiesFileConfigurationProviders$FileWatcherRunnable.run(PollingPropertiesFileConf
igurationProvider.java:140)
        at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
        at java.util.concurrent.FutureTask.runAndReset(FutureTask.java:308)
        at java.util.concurrent.ScheduledThreadPoolExecutor$ScheduledFutureTask.access$301(ScheduledThreadPoolExecutor.java:1
80)
        at java.util.concurrent.ScheduledThreadPoolExecutor$ScheduledFutureTask.run(ScheduledThreadPoolExecutor.java:294)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)
17/05/25 10:31:57 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agent
1]
17/05/25 10:31:57 INFO node.AbstractConfigurationProvider: Creating channels
17/05/25 10:31:57 INFO channel.DefaultChannelFactory: Creating instance of channel mychannel type memory
17/05/25 10:31:57 INFO node.AbstractConfigurationProvider: Created channel mychannel
17/05/25 10:31:57 INFO source.DefaultSourceFactory: Creating instance of source mysrc, type exec
17/05/25 10:31:57 INFO node.AbstractConfigurationProvider: Channel mychannel connected to [mysrc]
17/05/25 10:31:57 INFO node.Application: Starting new configuration:{ sourceRunners:{mysrc=EventDrivenSourceRunner: { source:
org.apache.flume.source.ExecSource{name:mysrc,state:IDLE} } } sinkRunners:{} channels:{mychannel=org.apache.flume.channel.Memo
ryChannel{name: mychannel} }
17/05/25 10:31:57 INFO node.Application: Starting Channel mychannel
17/05/25 10:31:57 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel: Suc
cessfully registered new MBean.
17/05/25 10:31:57 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
17/05/25 10:31:57 INFO source.ExecSource: Exec source starting with command:hadoop dfs -put /home/acadgild/Desktop/StatewiseD
istrictwisePhysicalProgress.xml /user/acadgild/flumeimp
17/05/25 10:31:57 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: mysrc: Successf
ully registered new MBean.
17/05/25 10:31:57 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc started
17/05/25 10:32:01 INFO source.ExecSource: Command [hadoop dfs -put /home/acadgild/Desktop/StatewiseDistrictwisePhysicalProgr
ess.xml /user/acadgild/flumeimp] exited with 0
Click to switch to "Workspace 2"

```

[Mozilla Firefox] StatewiseDistrictwise... acadgild@localhost:~/...

Displaying the extracted dataset using ls command in terminal

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
[acadgild@localhost Desktop]$ hadoop fs -ls /user/acadgild
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ight have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 11:58:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 62 items
drwxr-xr-x - acadgild supergroup          0 2017-05-08 14:42 /user/acadgild/16
drwxr-xr-x - acadgild supergroup          0 2017-05-08 14:34 /user/acadgild/16-2c
drwxr-xr-x - acadgild supergroup          0 2017-05-02 14:19 /user/acadgild/16-2output
drwxr-xr-x - acadgild supergroup          0 2017-05-08 14:42 /user/acadgild/16-4c
drwxr-xr-x - acadgild supergroup          0 2017-05-08 14:45 /user/acadgild/16-5c
drwxr-xr-x - acadgild supergroup          0 2017-05-08 14:48 /user/acadgild/16-6c
drwxr-xr-x - acadgild supergroup          0 2017-04-28 09:59 /user/acadgild/16_2o
drwxr-xr-x - acadgild supergroup          0 2017-04-22 12:41 /user/acadgild/16_2oP1
drwxr-xr-x - acadgild supergroup          0 2017-04-28 10:29 /user/acadgild/16_4o
drwxr-xr-x - acadgild supergroup          0 2017-04-22 12:57 /user/acadgild/16_4oP1
drwxr-xr-x - acadgild supergroup          0 2017-05-02 18:04 /user/acadgild/17
drwxr-xr-x - acadgild supergroup          0 2017-04-22 06:59 /user/acadgild/18_3
drwxr-xr-x - acadgild supergroup          0 2017-03-23 15:12 /user/acadgild/Data
-rw-r--r-- 1 acadgild supergroup        406 2017-02-24 15:57 /user/acadgild/Data.txt
drwxr-xr-x - acadgild supergroup          0 2017-03-23 16:00 /user/acadgild/Data5
drwxr-xr-x - acadgild supergroup          0 2017-03-27 15:42 /user/acadgild/Data6
drwxr-xr-x - acadgild supergroup          0 2017-03-28 17:36 /user/acadgild/Data7
drwxr-xr-x - acadgild supergroup          0 2017-03-28 15:11 /user/acadgild/Data8
drwxr-xr-x - acadgild supergroup          0 2017-02-24 15:57 /user/acadgild/Desktop
-rw-r--r-- 1 acadgild supergroup        869 2017-02-24 23:02 /user/acadgild/Flight.txt
-rw-r--r-- 1 acadgild supergroup         71 2017-02-24 17:16 /user/acadgild/Mov.txt
-rw-r--r-- 1 acadgild supergroup        102 2017-02-24 18:49 /user/acadgild/Movie.txt
drwxr-xr-x - acadgild supergroup          0 2017-02-24 17:27 /user/acadgild/OP1
drwxr-xr-x - acadgild supergroup          0 2017-02-24 17:31 /user/acadgild/OP2
drwxr-xr-x - acadgild supergroup          0 2017-02-24 17:42 /user/acadgild/OP3
-rw-r--r-- 1 acadgild supergroup        212 2017-02-24 15:08 /user/acadgild/Sam1.txt
-rw-r--r-- 1 acadgild supergroup        129 2017-02-24 14:54 /user/acadgild/Sample.txt
-rw-r--r-- 1 acadgild supergroup        678 2017-03-22 14:12 /user/acadgild/TVfinal
-rw-r--r-- 1 acadgild supergroup        501 2017-02-24 19:09 /user/acadgild/Titanic.txt
-rw-r--r-- 1 acadgild supergroup        405 2017-02-24 19:14 /user/acadgild/Titanic1.txt

```

[Mozilla Firefox] StatewiseDistrictwise... acadgild@localhost:~/...

Dataset extracted as xml file

```

File Edit View Search Terminal Help
-rw-r--r-- 1 acadgild supergroup 678 2017-03-22 14:12 /user/acadgild/TVfinal
-rw-r--r-- 1 acadgild supergroup 501 2017-02-24 19:09 /user/acadgild/Titanic.txt
-rw-r--r-- 1 acadgild supergroup 405 2017-02-24 19:14 /user/acadgild/Titanic1.txt
-rw-r--r-- 1 acadgild supergroup 25489723 2017-04-22 08:48 /user/acadgild/cc.csv
          0 2016-08-18 09:34 /user/acadgild/employee
          717414 2017-05-25 10:32 /user/acadgild/flumeimp
          0 2017-02-24 15:02 /user/acadgild/out1
drwxr-xr-x - acadgild supergroup 0 2017-02-24 16:23 /user/acadgild/out12
drwxr-xr-x - acadgild supergroup 0 2017-02-24 16:25 /user/acadgild/out13
drwxr-xr-x - acadgild supergroup 0 2017-02-24 15:07 /user/acadgild/out2
drwxr-xr-x - acadgild supergroup 0 2017-02-24 15:10 /user/acadgild/out5
drwxr-xr-x - acadgild supergroup 0 2017-02-24 15:29 /user/acadgild/out6
drwxr-xr-x - acadgild supergroup 0 2017-02-24 15:37 /user/acadgild/out7
drwxr-xr-x - acadgild supergroup 0 2017-02-24 15:58 /user/acadgild/out8
drwxr-xr-x - acadgild supergroup 0 2017-02-24 18:03 /user/acadgild/output2
drwxr-xr-x - acadgild supergroup 0 2017-02-24 18:43 /user/acadgild/output3
drwxr-xr-x - acadgild supergroup 0 2017-02-24 19:11 /user/acadgild/output4
drwxr-xr-x - acadgild supergroup 0 2017-02-24 19:15 /user/acadgild/output5
drwxr-xr-x - acadgild supergroup 0 2017-02-24 19:23 /user/acadgild/output6
drwxr-xr-x - acadgild supergroup 0 2017-02-24 23:03 /user/acadgild/output6_1a
drwxr-xr-x - acadgild supergroup 0 2017-02-24 23:12 /user/acadgild/output6_1aa
drwxr-xr-x - acadgild supergroup 0 2017-02-24 23:14 /user/acadgild/output6_laaa
drwxr-xr-x - acadgild supergroup 0 2017-02-24 23:09 /user/acadgild/output6_lb
drwxr-xr-x - acadgild supergroup 0 2017-02-24 23:18 /user/acadgild/output6_lbb
drwxr-xr-x - acadgild supergroup 0 2017-02-24 23:27 /user/acadgild/output6_1c
drwxr-xr-x - acadgild supergroup 0 2017-02-25 00:26 /user/acadgild/output6_2a
drwxr-xr-x - acadgild supergroup 0 2017-02-24 20:10 /user/acadgild/output7
drwxr-xr-x - acadgild supergroup 0 2017-04-24 06:54 /user/acadgild/p
-rw-r--r-- 1 acadgild supergroup 24 2017-03-24 13:54 /user/acadgild/pigdemo.txt
drwxr-xr-x - acadgild supergroup 0 2017-04-05 11:59 /user/acadgild/project
drwxr-xr-x - acadgild supergroup 0 2017-04-22 08:50 /user/acadgild/project2aout
drwxr-xr-x - acadgild supergroup 0 2017-04-05 13:57 /user/acadgild/projectcb
-rw-r--r-- 1 acadgild supergroup 159 2017-05-05 14:44 /user/acadgild/wc.txt
drwxr-xr-x - acadgild supergroup 0 2017-05-25 11:55 /user/acadgild/xmlcsv
-rw-r--r-- 1 acadgild supergroup 2497496 2017-05-06 14:42 /user/acadgild/zipcode.csv

```

Dataset as xml file in gedit

```

<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>

```

Conversion of dataset from xml file format to csv file format using pig script

```

2017-05-25 11:44:24,019 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/Desktop/pig_149569286401.log
2017-05-25 11:44:24,078 [main] INFO org.apache.pig.impl.Utils - Default bootup file /home/acadgild/.pigbootup not found
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/:org/slf4jimpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/:org/slf4jimpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
2017-05-25 11:44:24,912 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2017-05-25 11:44:24,914 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-05-25 11:44:24,914 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 11:44:24,914 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2017-05-25 11:44:25,996 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-fed00471-3de7-4d93-876b-7bf4ac0030cf
2017-05-25 11:44:25,996 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
    ...
grunt> A = LOAD '/user/acadgild/flumeimp' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
2017-05-25 11:49:28,809 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = FOREACH A generate REPLACE(x,'[\n]', '') as x;
grunt> C = FOREACH B generate REGEX_EXTRACT_ALL(x, '.*(?:<State Name>)(([^<]*).*?:<District Name>)(([^<]*).*?:<Project Objectives_IHML_BPL>)(([^<]*).*?:<Project Objectives_IHML_TOTAL>)(([^<]*).*?:<Project Objectives_SCW>)(([^<]*).*?:<Project Objectives_School_Toilets>)(([^<]*).*?:<Project Objectives_Anganwadi_Toilets>)(([^<]*).*?:<Project Objectives_RSM>)(([^<]*).*?:<Project Objectives_PC>)(([^<]*).*?:<Project_Performance-IHML_BPL>)(([^<]*).*?:<Project_Performance-IHML_AP>)(([^<]*).*?:<Project_Performance-IHML_TOTAL>)(([^<]*).*?:<Project_Performance-SCW>)(([^<]*).*?:<Project_Performance-SCH>)(([^<]*).*?:<Project_Performance-PC>)(([^<]*).*?);
grunt> 

```

```

    ...
2017-05-25 11:44:24,914 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2017-05-25 11:44:25,996 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-fed00471-3de7-4d93-876b-7bf4ac0030cf
2017-05-25 11:44:25,996 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
    ...
grunts> A = LOAD '/user/acadgild/flumeimp' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
2017-05-25 11:49:28,809 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = FOREACH A generate REPLACE(x,'[\n]', '') as x;
grunt> C = FOREACH B generate REGEX_EXTRACT_ALL(x, '.*(?:<State Name>)(([^<]*).*?:<District Name>)(([^<]*).*?:<Project Objectives_IHML_BPL>)(([^<]*).*?:<Project Objectives_IHML_TOTAL>)(([^<]*).*?:<Project Objectives_SCW>)(([^<]*).*?:<Project Objectives_School_Toilets>)(([^<]*).*?:<Project Objectives_Anganwadi_Toilets>)(([^<]*).*?:<Project Objectives_RSM>)(([^<]*).*?:<Project Objectives_PC>)(([^<]*).*?:<Project_Performance-IHML_BPL>)(([^<]*).*?:<Project_Performance-IHML_AP>)(([^<]*).*?:<Project_Performance-IHML_TOTAL>)(([^<]*).*?:<Project_Performance-SCW>)(([^<]*).*?:<Project_Performance-SCH>)(([^<]*).*?:<Project_Performance-PC>)(([^<]*).*?);
grunt> D = FOREACH C generate FLATTEN (($0));
grunt> STORE D INTO '/user/acadgild/xmlcsv' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
2017-05-25 11:55:25,987 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 11:55:26,013 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator

```

Dataset converted into csv file format from xml file format

```
[Acadgild_32_Bit [Running] - Oracle VM VirtualBox]
File Machine View Input Devices Help
[acadgild@localhost Desktop]$ hadoop fs -ls /user/acadgild
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 11:58:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 62 items
drwxr-xr-x 1 acadgild supergroup 0 2017-05-08 14:42 /user/acadgild/16
drwxr-xr-x 1 acadgild supergroup 0 2017-05-08 14:34 /user/acadgild/16-2c
drwxr-xr-x 1 acadgild supergroup 0 2017-05-02 14:19 /user/acadgild/16-2output
drwxr-xr-x 1 acadgild supergroup 0 2017-05-08 14:42 /user/acadgild/16-4c
drwxr-xr-x 1 acadgild supergroup 0 2017-05-08 14:45 /user/acadgild/16-5c
drwxr-xr-x 1 acadgild supergroup 0 2017-05-08 14:46 /user/acadgild/16-6c
drwxr-xr-x 1 acadgild supergroup 0 2017-04-28 09:59 /user/acadgild/16_2o
drwxr-xr-x 1 acadgild supergroup 0 2017-04-22 12:41 /user/acadgild/16_2opl
drwxr-xr-x 1 acadgild supergroup 0 2017-04-28 10:29 /user/acadgild/16_4o
drwxr-xr-x 1 acadgild supergroup 0 2017-04-22 12:57 /user/acadgild/16_4opl
drwxr-xr-x 1 acadgild supergroup 0 2017-05-02 18:04 /user/acadgild/17
drwxr-xr-x 1 acadgild supergroup 0 2017-04-22 06:59 /user/acadgild/18_3
drwxr-xr-x 1 acadgild supergroup 0 2017-03-22 15:12 /user/acadgild/Data
-rw-r--r-- 1 acadgild supergroup 406 2017-02-24 15:57 /user/acadgild/Data.txt
drwxr-xr-x 1 acadgild supergroup 0 2017-03-23 16:08 /user/acadgild/Data5
drwxr-xr-x 1 acadgild supergroup 0 2017-03-27 15:42 /user/acadgild/Data6
drwxr-xr-x 1 acadgild supergroup 0 2017-03-28 17:36 /user/acadgild/Data7
drwxr-xr-x 1 acadgild supergroup 0 2017-03-28 15:11 /user/acadgild/Data8
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:57 /user/acadgild/Desktop
drwxr-xr-x 1 acadgild supergroup 869 2017-02-24 23:02 /user/acadgild/Flight.txt
-rw-r--r-- 1 acadgild supergroup 71 2017-02-24 17:16 /user/acadgild/Mov.txt
-rw-r--r-- 1 acadgild supergroup 102 2017-02-24 18:46 /user/acadgild/Movie.txt
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 17:27 /user/acadgild/OP1
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 17:30 /user/acadgild/OP2
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 17:42 /user/acadgild/OP3
-rw-r--r-- 1 acadgild supergroup 212 2017-02-24 15:00 /user/acadgild/Sam1.txt
-rw-r--r-- 1 acadgild supergroup 129 2017-02-24 14:54 /user/acadgild/Sample.txt
-rw-r--r-- 1 acadgild supergroup 678 2017-03-22 14:12 /user/acadgild/TVfinal
-rw-r--r-- 1 acadgild supergroup 501 2017-02-24 19:09 /user/acadgild/Titanic.txt
-rw-r--r-- 1 acadgild supergroup 405 2017-02-24 19:14 /user/acadgild/Titanic1.txt
[Mozilla Firefox] StatewiseDistrictwise... acadgild@localhost:~... Right Ctrl
```

```
[Acadgild_32_Bit [Running] - Oracle VM VirtualBox]
File Machine View Input Applications Places System acadgild@localhost:~/Desktop
File Edit View Terminal Help
-rw-r--r-- 1 acadgild supergroup 678 2017-03-22 14:12 /user/acadgild/TVfinal
-rw-r--r-- 1 acadgild supergroup 501 2017-02-24 19:09 /user/acadgild/Titanic.txt
-rw-r--r-- 1 acadgild supergroup 405 2017-02-24 19:14 /user/acadgild/Titanic1.txt
25489723 2017-04-22 08:48 /user/acadgild/cc.csv
0 2016-08-18 09:34 /user/acadgild/employee
717414 2017-05-25 10:32 /user/acadgild/flumeimp
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:02 /user/acadgild/out1
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 16:23 /user/acadgild/out12
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 16:25 /user/acadgild/out13
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:07 /user/acadgild/out2
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:10 /user/acadgild/out5
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:29 /user/acadgild/out6
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:37 /user/acadgild/out7
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:58 /user/acadgild/out8
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 18:03 /user/acadgild/output2
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 18:43 /user/acadgild/output3
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 19:11 /user/acadgild/output4
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 19:15 /user/acadgild/output5
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 19:23 /user/acadgild/output6
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:03 /user/acadgild/output6_1a
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:12 /user/acadgild/output6_1aa
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:09 /user/acadgild/output6_1b
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:17 /user/acadgild/output6_1bb
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:27 /user/acadgild/output6_1c
drwxr-xr-x 1 acadgild supergroup 0 2017-02-25 00:20 /user/acadgild/output6_2a
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 20:11 /user/acadgild/output7
drwxr-xr-x 1 acadgild supergroup 0 2017-04-22 06:56 /user/acadgild/p
-rw-r--r-- 1 acadgild supergroup 24 2017-03-24 13:54 /user/acadgild/pigdemo.txt
drwxr-xr-x 1 acadgild supergroup 0 2017-04-05 11:59 /user/acadgild/project
drwxr-xr-x 1 acadgild supergroup 0 2017-04-22 08:50 /user/acadgild/project2aout
drwxr-xr-x 1 acadgild supergroup 0 2017-04-05 13:57 /user/acadgild/projectb
drwxr-xr-x 1 acadgild supergroup 159 2017-05-05 14:44 /user/acadgild/wc.txt
drwxr-xr-x 1 acadgild supergroup 0 2017-05-25 11:55 /user/acadgild/xmlcsv
-rw-r--r-- 1 acadgild supergroup 2497496 2017-05-08 14:42 /user/acadgild/zipcode.csv
Right Ctrl
```

Dataset extracted from xml file format

```
[acadgild@localhost Desktop]$ hadoop fs -ls /user/acadgild/xmlcsv
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 12:00:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 acadgild supergroup 0 2017-05-25 11:55 /user/acadgild/xmlcsv/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 55135 2017-05-25 11:55 /user/acadgild/xmlcsv/part-m-00000
[acadgild@localhost Desktop]$ hadoop fs -cat /user/acadgild/xmlcsv/part-m-00000
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 12:00:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Andhra Pradesh,ADILABAD,247475,148181,395656,0,4462,427,10,0,176300,52431,228731,0,4462,427,0,0
Andhra Pradesh,ANANTAPUR,363314,181335,544649,0,3421,284,10,0,366557,42000,408557,0,4258,302,0,0
Andhra Pradesh,CHITTOOR,296465,236986,533451,0,8171,375,10,0,269750,199905,466655,0,8171,375,11,0
Andhra Pradesh,CUDDAPAH,251653,251610,503263,0,6802,277,10,0,239789,125493,365273,0,5431,277,0,0
Andhra Pradesh,EAST GODAVARI,370255,191409,561655,50,7004,1164,10,0,347305,191408,538705,49,7004,781,20,0
Andhra Pradesh,GUNTUR,370996,319053,690049,0,5346,215,10,0,269608,85950,355558,0,5116,215,37,0
Andhra Pradesh,KARIMNAGAR,365267,219763,585030,0,5848,298,10,0,369433,63025,432458,0,5848,298,13,0
Andhra Pradesh,KHAMMAM,189225,4860,194085,0,6318,521,10,0,195763,4870,200633,0,6318,521,15,10
Andhra Pradesh,KRISHNA,351572,175991,527563,17,3374,62,10,0,318730,104197,422927,8,3374,171,40,0
Andhra Pradesh,KURNool,383478,129956,513434,0,4669,2111,10,0,323616,174773,498389,13,4669,248,49,0
Andhra Pradesh,MAHBUBNAGAR,418559,66032,484591,3,5433,357,10,0,242169,263280,505449,9,5434,357,15,0
Andhra Pradesh,MEDAK,311743,96499,408233,50,5446,1873,10,0,310591,47977,358568,14,4752,1049,10,0
Andhra Pradesh,NALGONDA,215058,156761,371819,0,6628,44,10,0,224813,87693,312560,0,6487,395,10,10
Andhra Pradesh,NELLORE,226725,116678,343483,0,6415,1396,10,0,169111,19462,188573,0,5839,476,0,0
Andhra Pradesh,NIZAMABAD,225519,9474,234993,0,2477,32,10,0,225519,64759,290278,0,2477,32,36,0
Andhra Pradesh,PRAKASAM,386617,193493,580110,0,7224,788,10,0,306266,143238,449564,0,7195,770,41,0
Andhra Pradesh,RANGAREDDI,212629,71610,284239,0,3726,467,10,0,174460,69147,243667,0,3726,467,10,0
Andhra Pradesh,SRIKAKULAM,302221,225272,527493,46,6398,266,10,0,117379,82270,199649,16,5580,128,15,3
Andhra Pradesh,VISAKHAPATNAM,252875,156460,499335,20,5074,0,10,0,178731,75765,254496,20,6378,559,0,0
Andhra Pradesh,VIZIANAGARAM,230851,245678,465729,10,2999,1668,10,0,166713,159984,326697,8,3368,350,12,0
Andhra Pradesh,WARANGAL,330260,138180,468440,0,4122,2157,10,0,359732,345450,705182,0,4122,477,0,0
Andhra Pradesh,WEST GODAVARI,344272,294425,638697,385,4551,216,10,0,319477,294425,613982,837,4524,316,97,0
Arunachal Pradesh,ANJW,3232,0,3232,8,44,46,1,0,1490,0,1490,2,48,48,0,0
```

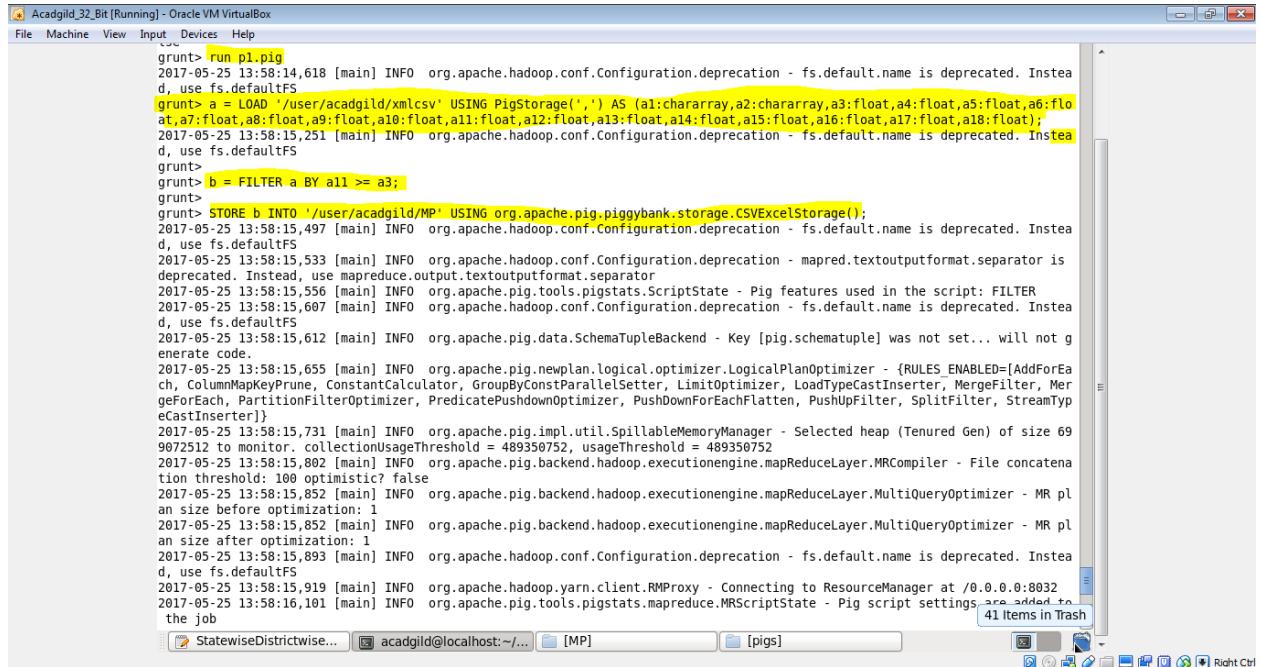
Pig script for the 1st question

Q1: Find out the districts who achieved 100 percent objective in BPL cards?

```
p1.pig (~/Desktop/MP... acadgild@localhost:~/... [MP] pigs Right Ctrl
```

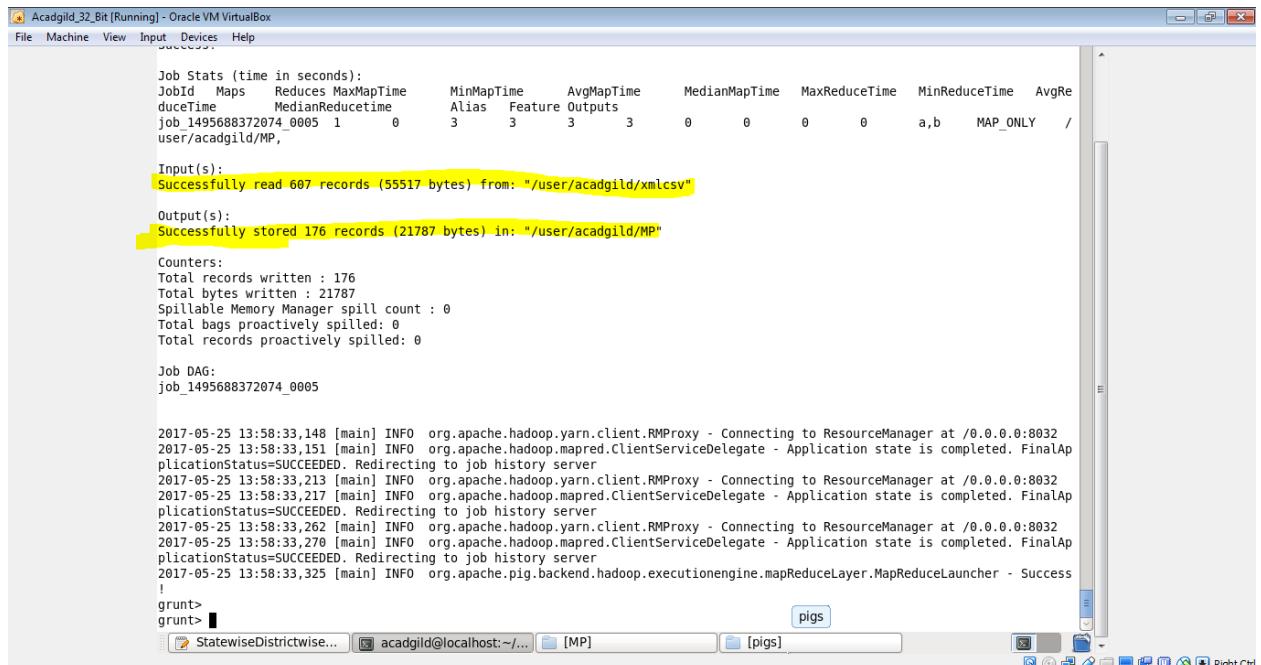
```
a = LOAD '/user/acadgild/xmlcsv' USING PigStorage(',') AS
(a1:chararray,a2:chararray,a3:float,a4:float,a5:float,a6:float,a7:float,a8:float,a9:float,a10:float,a11:float,a12:float,a13:float);
b = FILTER a BY a11 >= a3;
STORE b INTO '/user/acadgild/MP' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
```

Executing the pig script



```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
grunts> run pi.pig
2017-05-25 13:58:14,618 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts> a = LOAD '/user/acadgild/xmlcsv' USING PigStorage(',') AS (a1:chararray,a2:chararray,a3:float,a4:float,a5:float,a6:float,a7:float,a8:float,a9:float,a10:float,a11:float,a12:float,a13:float,a14:float,a15:float,a16:float,a17:float,a18:float);
2017-05-25 13:58:15,251 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
grunts> b = FILTER a BY a11 >= a3;
grunts>
grunts> STORE b INTO '/user/acadgild/MP' USING org.apache.pig.piggystash.storage.CSVExcelStorage();
2017-05-25 13:58:15,497 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 13:58:15,533 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2017-05-25 13:58:15,556 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2017-05-25 13:58:15,607 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 13:58:15,612 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-05-25 13:58:15,655 [main] INFO org.apache.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-05-25 13:58:15,731 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699672512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2017-05-25 13:58:15,802 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2017-05-25 13:58:15,852 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2017-05-25 13:58:15,852 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2017-05-25 13:58:15,893 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 13:58:15,919 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 13:58:16,101 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
41 items in Trash
[ StatewiseDistrictwise... ] [ acadgild@localhost:~/... ] [ MP ] [ pigs ]
```

Success status of the executed pig script



```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Success:
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime
duceTime MedianReducetime Alias Feature Outputs
job_1495688372074_0005 1 0 3 3 3 3 0 0 0 a,b MAP_ONLY /
user/acadgild/MP,
Input(s):
Successfully read 607 records (55517 bytes) from: "/user/acadgild/xmlcsv"
Output(s):
Successfully stored 176 records (21787 bytes) in: "/user/acadgild/MP"
Counters:
Total records written : 176
Total bytes written : 21787
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1495688372074_0005

2017-05-25 13:58:33,148 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 13:58:33,151 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-25 13:58:33,213 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 13:58:33,217 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-25 13:58:33,262 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 13:58:33,270 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-25 13:58:33,325 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
!
grunts> [ StatewiseDistrictwise... ] [ acadgild@localhost:~/... ] [ MP ] [ pigs ]
```

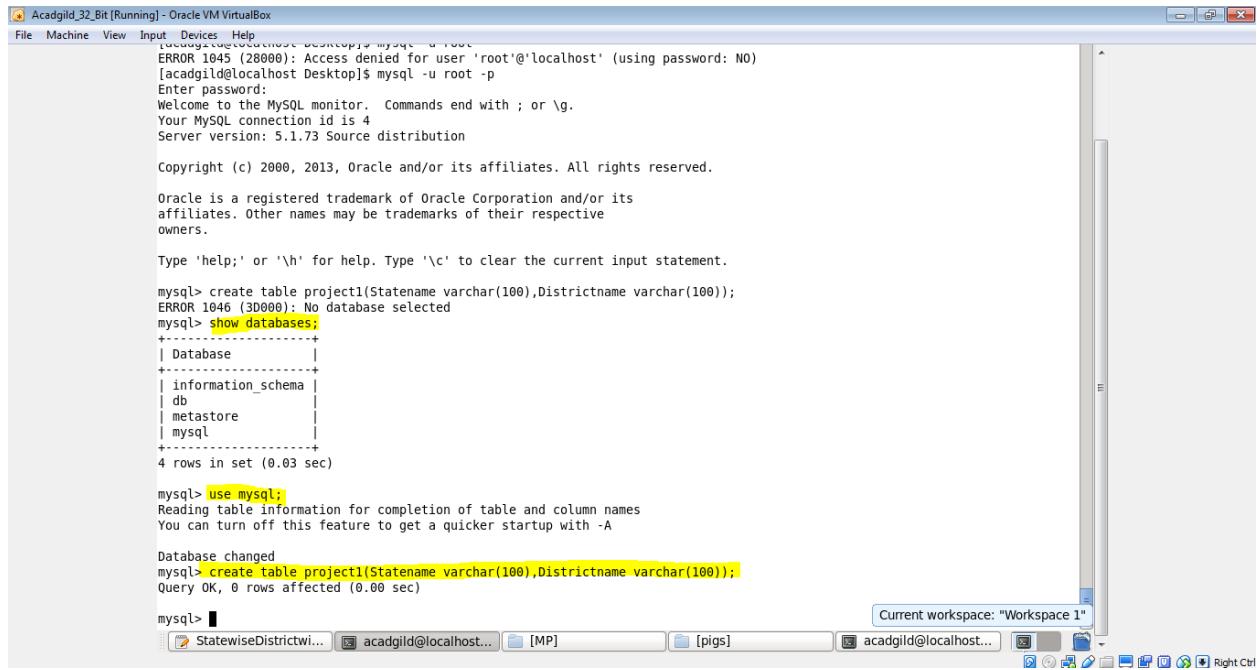
Displaying the stored output of the pig script

The districts who achieved 100 percent objective in BPL cards are

```
[Acadgild_32_Bit [Running] - Oracle VM VirtualBox]
File Machine View Input Devices Help
Applications Places System Thu May 25, 2:19 PM Acadgild
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
[acadgild@localhost Desktop]$ hadoop fs -ls /user/acadgild/MP
Java HotSpot(TM) Client VM warning: You have loaded library '/home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0' which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 14:17:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 acadgild supergroup 0 2017-05-25 13:58 /user/acadgild/MP/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 21787 2017-05-25 13:58 /user/acadgild/MP/part-m-00000
[acadgild@localhost Desktop]$ hadoop fs -cat /user/acadgild/MP/part-m-00000
Java HotSpot(TM) Client VM warning: You have loaded library '/home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0' which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 14:18:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Andhra Pradesh,ANANTAPUR,363314.0,181335.0,544649.0,0.0,3421.0,284.0,10.0,0.0,366557.0,42000.0,408557.0,0.0,4258.0,302.0,0.0,0.0
Andhra Pradesh,KARIMNAGAR,365267.0,219763.0,585030.0,0.0,5848.0,298.0,10.0,0.0,369433.0,63025.0,432458.0,0.0,5848.0,298.0,13.0,0.0
Andhra Pradesh,KHAMMAM,189225.0,4960.0,194085.0,0.0,6318.0,521.0,10.0,0.0,195763.0,4870.0,200633.0,0.0,6318.0,521.0,15.0,10.0
Andhra Pradesh,NALGONDA,215058.0,156761.0,371819.0,0.0,6628.0,44.0,10.0,0.0,224813.0,87693.0,312506.0,0.0,6487.0,395.0,10.0,1.0
Andhra Pradesh,NIZAMABAD,225519.0,9474.0,234993.0,0.0,2477.0,32.0,10.0,0.0,225519.0,64759.0,290278.0,0.0,2477.0,32.0,36.0,0.0
Andhra Pradesh,WARANGAL,330260.0,138180.0,468440.0,0.0,4122.0,2157.0,10.0,0.0,359732.0,345450.0,705182.0,0.0,4122.0,477.0,0.0,0.0
Arunachal Pradesh,DIBANG VALLEY,1085.0,0.0,1885.0,0.0,102.0,27.0,0.0,0.0,1088.0,0.0,40.0,10.0,0.0,0.0,0.0
Arunachal Pradesh,TIRAP,5789.0,470.0,6250.0,25.0,304.0,150.0,6.0,0.0,5789.0,479.0,6250.0,18.0,304.0,150.0,0.0,0.0,0.0
Assam,HAJIBAKANDI,49837.0,12445.0,62828.0,0.0,1140.0,35.0,4.0,0.0,49837.0,10570.0,60497.0,0.0,1140.0,35.0,4.0,0.0
Bihar,MAUDHUBANI,67482.0,69746.0,137228.0,63.0,4572.0,1127.0,10.0,0.0,67482.0,32211.0,99693.0,18.0,2704.0,452.0,5.0,5.0
Bihar,VAISHALI,190598.0,195325.0,385923.0,50.0,3224.0,518.0,10.0,0.0,196496.0,91065.0,287561.0,12.0,2908.0,455.0,1.0,95.0
Chhattisgarh,GORAKSHI,50691.0,74471.0,125162.0,50.0,3618.0,1223.0,0.0,0.0,639839.0,49162.0,113145.0,17.0,3618.0,1223.0,5.0,5.0
Goa,NORTH GOA,15000.0,0.0,15000.0,150.0,283.0,116.0,0.0,0.0,15000.0,0.0,208.0,8.0,0.0,0.0,0.0
Gujarat,AHMEDABAD,80192.0,169360.0,249525.0,35.0,920.0,678.0,31.0,0.0,80192.0,119352.0,199544.0,26.0,917.0,473.0,31.0,0.0
Gujarat,BHAVNAGAR,31305.0,87900.0,119205.0,54.0,600.0,1145.0,0.0,0.0,31563.0,72870.0,104433.0,109.0,628.0,839.0,42.0,0.0
Gujarat,DANGS,77000.0,16142.0,11042.0,0.0,216.0,200.0,0.0,0.0,27000.0,13392.0,41292.0,10.0,216.0,700.0,1.0,1.0
```

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
0.0,10.0
Uttar Pradesh,MAHAMAYA NAGAR(HATHRAS),25153.0,129568.0,154721.0,10.0,2423.0,1130.0,3.0,1.0,25277.0,114597.0,139874.0,10.0,214
6.0,1130.0,1.0,0.0
Uttar Pradesh,MAHARAJGANJ,145090.0,111256.0,256346.0,55.0,2956.0,1461.0,5.0,1.0,145090.0,90842.0,235932.0,55.0,2956.0,1461.0,
4.0,0.0
Uttar Pradesh,MAHOBIA,53117.0,70666.0,123783.0,0.0,0.2153.0,464.0,1.0,1.0,53117.0,65145.0,118262.0,0.0,1756.0,464.0,2.0,0.0
Uttar Pradesh,MATHURA,25005.0,182330.0,207335.0,10.0,3228.0,1475.0,3.0,1.0,25110.0,129553.0,154663.0,10.0,2936.0,979.0,1.0,0.
0
Uttar Pradesh,MIRZAPUR,231315.0,195266.0,426581.0,133.0,3941.0,1582.0,10.0,0.0,233308.0,182883.0,416191.0,133.0,3941.0,1582.0
,10.0,10.0
Uttar Pradesh,MORADABAD,76018.0,354539.0,430557.0,0.0,0.8504.0,2641.0,4.0,1.0,76018.0,273367.0,349385.0,0.0,0.4870.0,2095.0,2.0,1
.0
Uttar Pradesh,MUZAFFARNAGAR,51660.0,238443.0,290103.0,14.0,3700.0,2074.0,2.0,0.0,51660.0,169008.0,220668.0,14.0,3478.0,2074.0
,2.0,0.0
Uttar Pradesh,PILIBHIT,95178.0,119893.0,215071.0,21.0,2786.0,889.0,2.0,1.0,95178.0,103271.0,198449.0,21.0,2786.0,889.0,1.0,1.
0
Uttar Pradesh,PRATAPGARH,141368.0,264341.0,405709.0,50.0,2525.0,1955.0,10.0,0.0,143517.0,167106.0,310623.0,79.0,2525.0,2030.0
,3.0,0.0
Uttar Pradesh,RAE BARELI,198306.0,135862.0,326168.0,25.0,5889.0,1860.0,5.0,0.0,198438.0,124312.0,314742.0,25.0,4188.0,1860.0
,1.0,0.0
Uttar Pradesh,SAHARANPUR,49458.0,192771.0,242229.0,25.0,3766.0,1720.0,11.0,0.0,49586.0,190364.0,239950.0,25.0,3505.0,1299.0,4
.0,0.0
Uttar Pradesh,SHAHJAHANPUR,194645.0,114246.0,308891.0,10.0,4842.0,1347.0,0.0,0.0,194959.0,94064.0,289023.0,10.0,4842.0,1347.0
,6.0,9.0
Uttar Pradesh,SONBHADRA,138370.0,79419.0,217789.0,113.0,3176.0,1051.0,0.0,0.0,138370.0,57900.0,196270.0,113.0,2978.0,755.0,10
.0,0.0
Uttar Pradesh,SULTANPUR,168843.0,262071.0,430914.0,40.0,4898.0,2244.0,6.0,0.0,168843.0,215646.0,384489.0,40.0,4898.0,2244.0,6
,0.0,0
Uttarakhand,NAINITAL,14314.0,26585.0,40899.0,50.0,208.0,95.0,6.0,2.0,18039.0,29025.0,47064.0,1.0,208.0,19.0,2.0,0.0
Uttarakhand,RUDRAPRAYAG,13150.0,22322.0,35472.0,40.0,109.0,124.0,3.0,0.0,13810.0,17522.0,31332.0,9.0,109.0,70.0,2.0,0.0
West Bengal,DAKSHIN DINAJPUR,182621.0,194577.0,377198.0,50.50.0,3712.0,2642.0,10.0,0.0,184153.0,49448.0,233601.0,19.0,2632.0,0.939
.0,8.0,8.0
West Bengal,MIDNAPUR EAST,392371.0,32617.0,424988.0,172.0,9726.0,5969.0,25.0,0.0,527398.0,32642.0,560031.0,210.0,10149.0,2862
.0,8.0,17.0
West Bengal,MIDNAPUR WEST,509496.0,432096.0,941592.0,50.0,16498.0,5825.0,10.0,0.0,596291.0,322659.0,918950.0,73.0,13452.0,278
7.0,0.0,0.0
[acadgild@localhost Desktop]$
```

Creating table in mysql



```
[acadgild@localhost Desktop]$ mysql -u root
ERROR 1045 (28000): Access denied for user 'root'@'localhost' (using password: NO)
[acadgild@localhost Desktop]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 4
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.

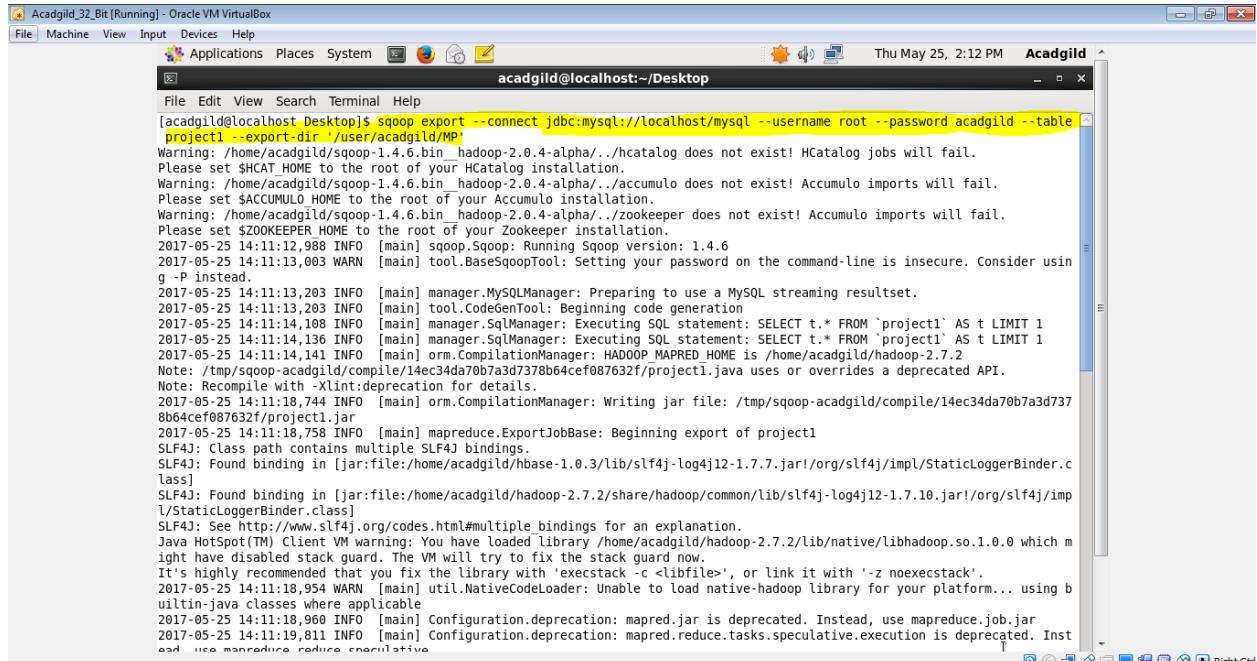
mysql> create table project1(Statename varchar(100),Districtname varchar(100));
ERROR 1046 (3D000): No database selected
mysql> show databases;
+--------------------+
| Database          |
+--------------------+
| information_schema |
| db                |
| metastore         |
| mysql             |
+--------------------+
4 rows in set (0.03 sec)

mysql> use mysql;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> create table project1(Statename varchar(100),Districtname varchar(100));
Query OK, 0 rows affected (0.00 sec)

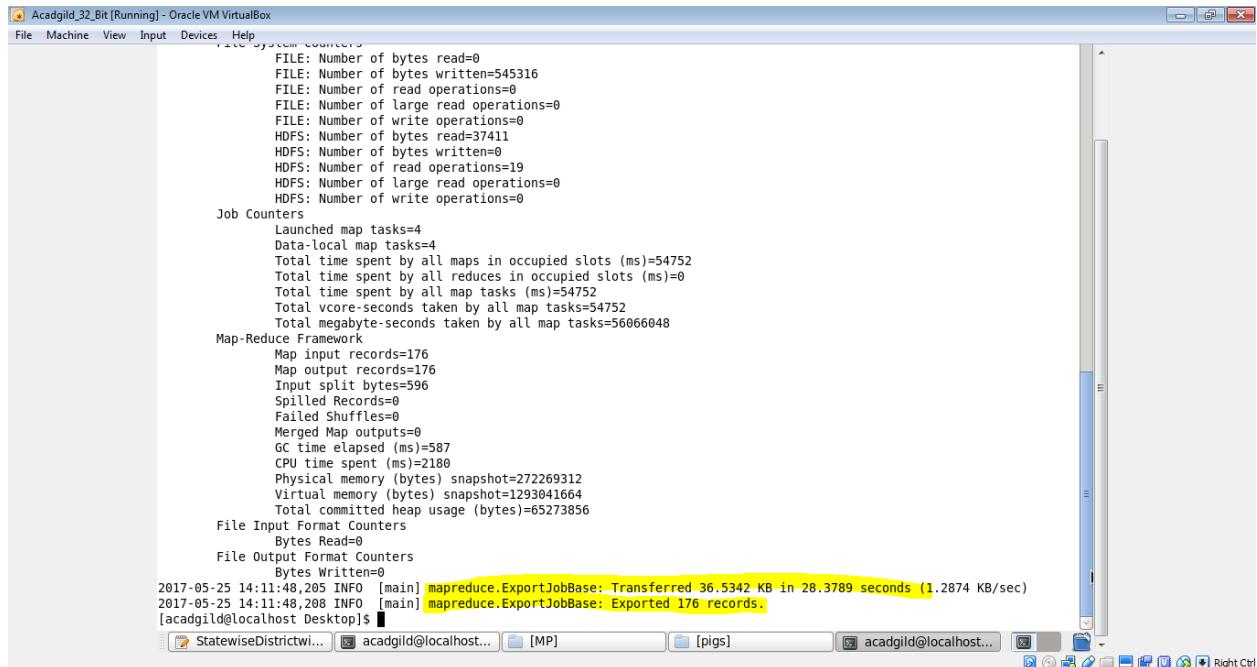
mysql>
```

Exporting the output of pig script to mysql using sqoop command



```
[acadgild@localhost Desktop]$ sqoop export --connect jdbc:mysql://localhost/mysql --username root --password acadgild --table project1 --export-dir '/user/acadgild/MP'
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop2-0.4-alpha/../.hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop2-0.4-alpha/../.accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop2-0.4-alpha/../.zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2017-05-25 14:11:12,988 INFO  [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
2017-05-25 14:11:13,003 WARN  [main] tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -g instead.
2017-05-25 14:11:13,203 INFO  [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-05-25 14:11:13,203 INFO  [main] tool.CodeGenTool: Beginning code generation
2017-05-25 14:11:14,108 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project1` AS t LIMIT 1
2017-05-25 14:11:14,136 INFO  [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project1` AS t LIMIT 1
2017-05-25 14:11:14,141 INFO  [main] orm.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/hadoop-2.7.2
Note: /tmp/sqoop-acadgild/compile/14ec34da70b7a3d7378b64cef087632f/project1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2017-05-25 14:11:18,744 INFO  [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/14ec34da70b7a3d7378b64cef087632f/project1.jar
2017-05-25 14:11:18,750 INFO  [main] mapreduce.ExportJobBase: Beginning export of project1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
2017-05-25 14:11:18,954 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable
2017-05-25 14:11:18,960 INFO  [main] Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2017-05-25 14:11:19,811 INFO  [main] Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
```

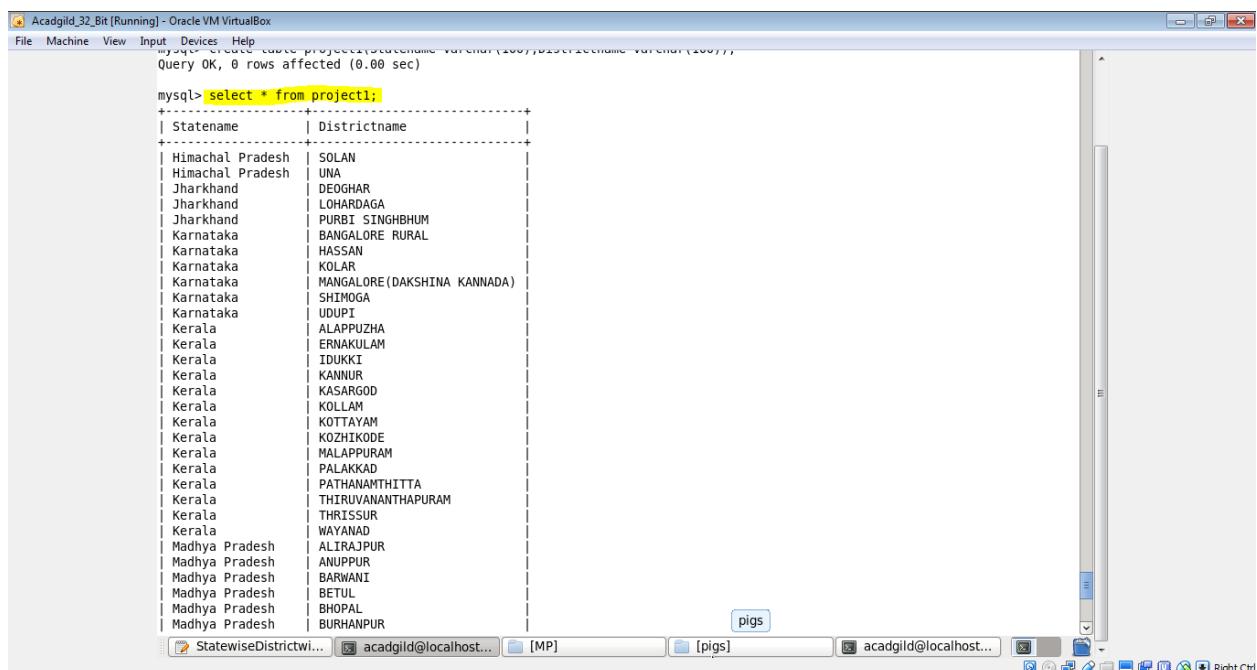
Status of the executed sqoop command



A screenshot of a Windows desktop showing a terminal window titled "Acadgild_32_Bit [Running] - Oracle VM VirtualBox". The terminal displays the output of a Sqoop export command. The output includes various counters and metrics such as file and HDFS operations, job counters, map-reduce framework details, and file input/output counters. A key line in the output is highlighted in yellow: "2017-05-25 14:11:48,208 INFO [main] mapreduce.ExportJobBase: Exported 176 records." This indicates that the data has been successfully transferred from HDFS to MySQL.

```
FILE: Number of bytes read=0
FILE: Number of bytes written=545316
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=37411
HDFS: Number of bytes written=0
HDFS: Number of read operations=19
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
Launched map tasks=4
Data-local map tasks=4
Total time spent by all maps in occupied slots (ms)=54752
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=54752
Total vcore-seconds taken by all map tasks=54752
Total megabyte-seconds taken by all map tasks=56066048
Map-Reduce Framework
Map input records=176
Map output records=176
Input split bytes=396
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=587
CPU time spent (ms)=2180
Physical memory (bytes) snapshot=272269312
Virtual memory (bytes) snapshot=1293041664
Total committed heap usage (bytes)=65273856
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
2017-05-25 14:11:48,208 INFO [main] mapreduce.ExportJobBase: Transferred 36.5342 KB in 28.3789 seconds (1.2874 KB/sec)
2017-05-25 14:11:48,208 INFO [main] mapreduce.ExportJobBase: Exported 176 records.
[acadgild@localhost Desktop]$
```

Exported result in mysql table



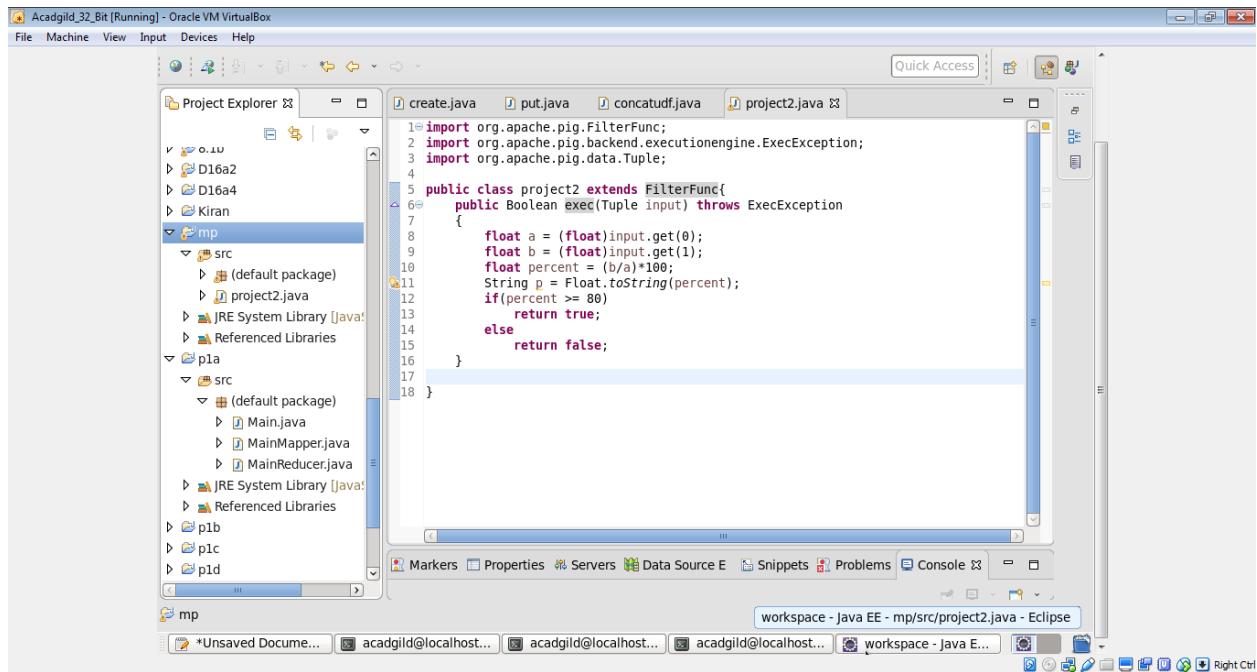
A screenshot of a Windows desktop showing a terminal window titled "Acadgild_32_Bit [Running] - Oracle VM VirtualBox". The terminal shows the creation of a MySQL table named "project" and then executes a SELECT query on it. The output of the SELECT query is a table listing state names and their corresponding district names. The districts listed are: SOLAN, UNA, DEOGHAR, LOHARDAGA, PURBT SINGHBHUM, BANGALORE RURAL, HASSAN, KOLAR, MANGALORE (DAKSHINA KANNADA), SHIMOGA, UDUPI, ALAPPUZHA, ERIAKULAM, IDUKKI, KANNUR, KASARGOD, KOLLAM, KOTTAYAM, KOZHIKODE, MALAPPURAM, PALAKKAD, PATHANAMTHITTA, THIRUVANANTHAPURAM, THRISSUR, WAYANAD, ALIRAJPUR, ANUPPUR, BARWANI, BETUL, BHOPAL, and BURHANPUR.

```
mysql> create table project(StateName varchar(100),DistrictName varchar(100));
Query OK, 0 rows affected (0.00 sec)

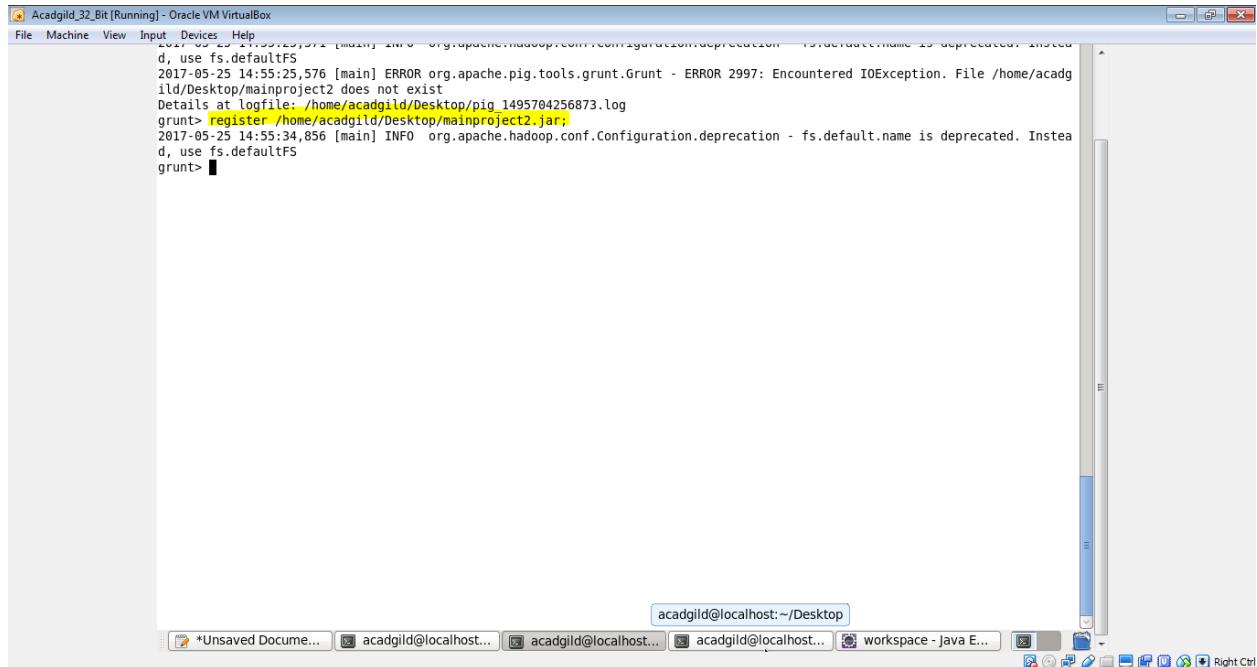
mysql> select * from project;
+-----+-----+
| StateName | DistrictName |
+-----+-----+
| Himachal Pradesh | SOLAN
| Himachal Pradesh | UNA
| Jharkhand | DEOGHAR
| Jharkhand | LOHARDAGA
| Jharkhand | PURBT SINGHBHUM
| Karnataka | BANGALORE RURAL
| Karnataka | HASSAN
| Karnataka | KOLAR
| Karnataka | MANGALORE (DAKSHINA KANNADA)
| Karnataka | SHIMOGA
| Karnataka | UDUPI
| Kerala | ALAPPUZHA
| Kerala | ERIAKULAM
| Kerala | IDUKKI
| Kerala | KANNUR
| Kerala | KASARGOD
| Kerala | KOLLAM
| Kerala | KOTTAYAM
| Kerala | KOZHIKODE
| Kerala | MALAPPURAM
| Kerala | PALAKKAD
| Kerala | PATHANAMTHITTA
| Kerala | THIRUVANANTHAPURAM
| Kerala | THRISSUR
| Kerala | WAYANAD
| Madhya Pradesh | ALIRAJPUR
| Madhya Pradesh | ANUPPUR
| Madhya Pradesh | BARWANI
| Madhya Pradesh | BETUL
| Madhya Pradesh | BHOPAL
| Madhya Pradesh | BURHANPUR
+-----+-----+
```

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox  
File Machine View Input Devices Help  
| Uttar Pradesh | CHANDAULI  
| Uttar Pradesh | CHITRAKOOT  
| Uttar Pradesh | ETAH  
| Uttar Pradesh | ETAWAH  
| Uttar Pradesh | FARRUKHABAD  
| Uttar Pradesh | FIROZABAD  
| Uttar Pradesh | GHAZIABAD  
| Uttar Pradesh | HARDOI  
| Uttar Pradesh | JAUNPUR  
| Uttar Pradesh | JHANSI  
| Uttar Pradesh | JYOTIBA PHULE NAGAR  
| Uttar Pradesh | KANNAUJ  
| Uttar Pradesh | LAKHIMPUR KHERI  
| Uttar Pradesh | LUCKNOW  
| Uttar Pradesh | MAHAMAYA NAGAR(HATHRAS)  
| Uttar Pradesh | MAHARAJGANJ  
| Uttar Pradesh | MAHOB  
| Uttar Pradesh | MATHURA  
| Uttar Pradesh | MIRzapur  
| Uttar Pradesh | MORADABAD  
| Uttar Pradesh | MUZAFFARNAGAR  
| Uttar Pradesh | PILIBHIT  
| Uttar Pradesh | PRATAPGARH  
| Uttar Pradesh | RAE BARELI  
| Uttar Pradesh | SAHARANPUR  
| Uttar Pradesh | SHAHJAHANPUR  
| Uttar Pradesh | SONBHADRA  
| Uttar Pradesh | SULTANPUR  
| Uttarakhand | NAINITAL  
| Uttarakhand | RUDRAPRAYAG  
| West Bengal | DAKSHIN DINAJPUR  
| West Bengal | MIDNAPUR EAST  
| West Bengal | MIDNAPUR WEST  
+-----+  
176 rows in set (0.00 sec)  
mysql> [MP]  
[StatewiseDistrictwi... acadgild@localhost... [MP] [pigs] acadgild@localhost... Right Ctrl
```

Pig udf in eclipse

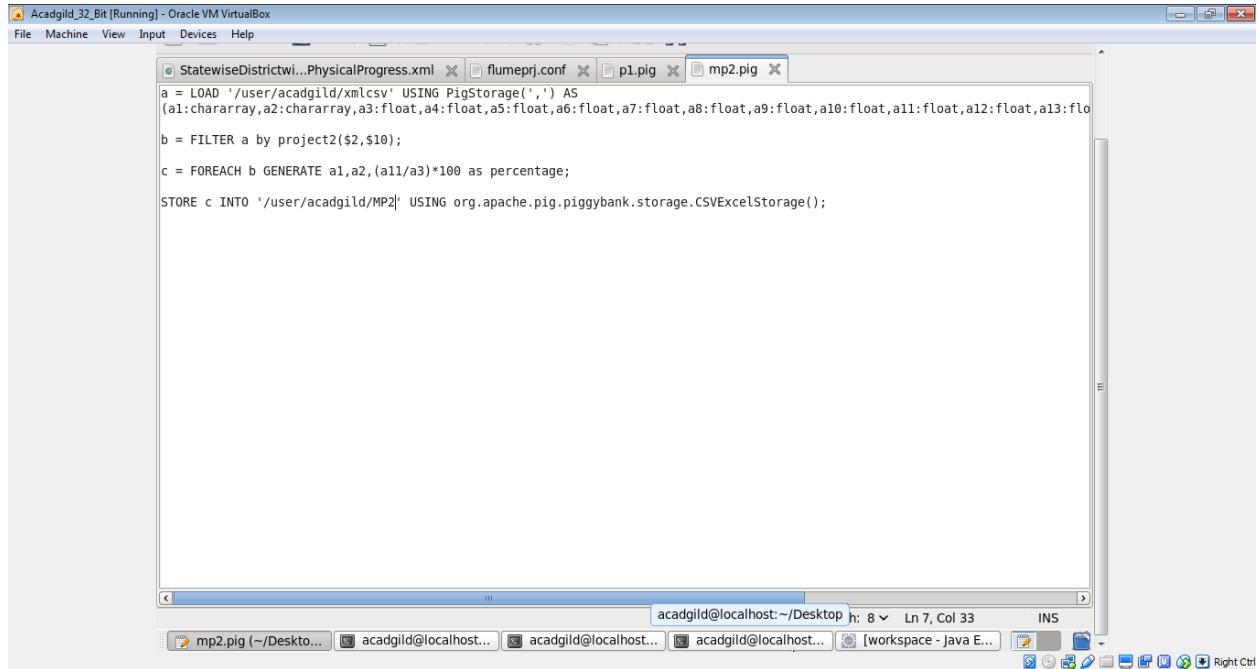


Registering the pig udf in grunt shell



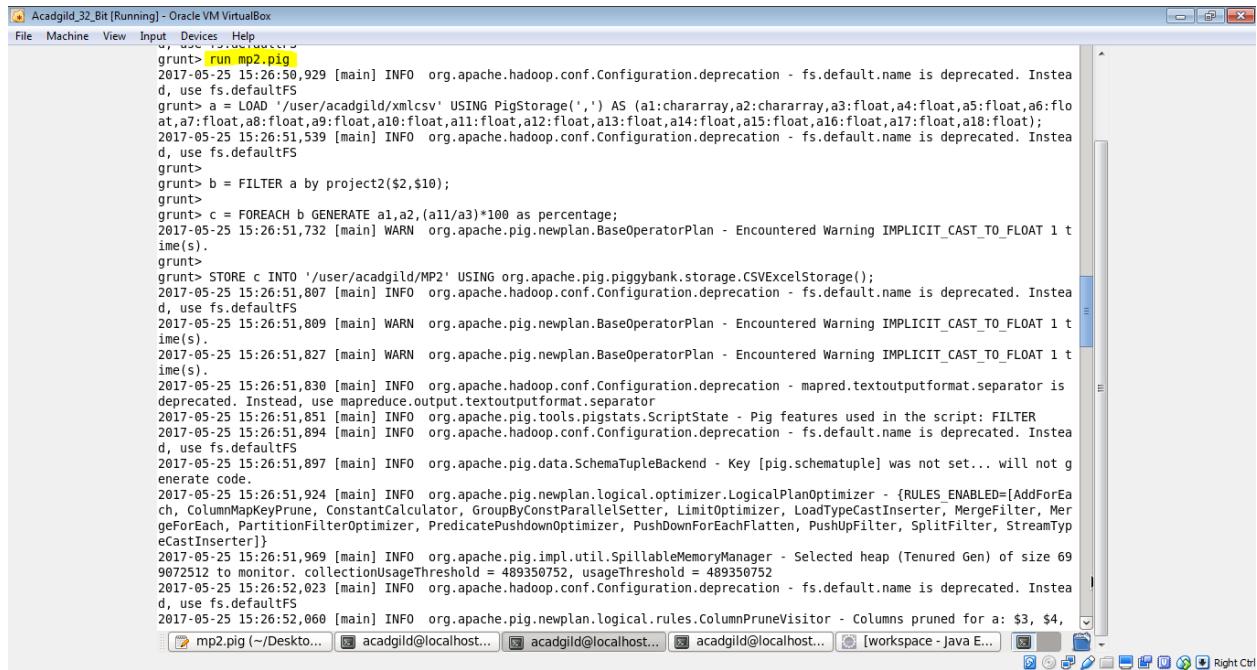
Pig script for the 2nd question

Q2: Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.



```
a = LOAD '/user/acadgild/xmlcsv' USING PigStorage(',') AS
(a1:chararray,a2:chararray,a3:float,a4:float,a5:float,a6:float,a7:float,a8:float,a9:float,a10:float,a11:float,a12:float,a13:float);
b = FILTER a by project2($2,$10);
c = FOREACH b GENERATE a1,a2,(a11/a3)*100 as percentage;
STORE c INTO '/user/acadgild/MP2' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
```

Executing the pig script for the 2nd question



```
grunts> run mp2.pig
2017-05-25 15:26:50,929 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts> a = LOAD '/user/acadgild/xmlcsv' USING PigStorage(',') AS
(a1:chararray,a2:chararray,a3:float,a4:float,a5:float,a6:float,a7:float,a8:float,a9:float,a10:float,a11:float,a12:float,a13:float);
2017-05-25 15:26:51,539 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts>
grunts> b = FILTER a by project2($2,$10);
grunts>
grunts> c = FOREACH b GENERATE a1,a2,(a11/a3)*100 as percentage;
2017-05-25 15:26:51,732 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 times(s).
grunts>
grunts> STORE c INTO '/user/acadgild/MP2' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
2017-05-25 15:26:51,807 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 15:26:51,809 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 times(s).
2017-05-25 15:26:51,827 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 times(s).
2017-05-25 15:26:51,830 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2017-05-25 15:26:51,851 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2017-05-25 15:26:51,894 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 15:26:51,897 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-05-25 15:26:51,924 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-05-25 15:26:51,969 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
2017-05-25 15:26:52,023 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-25 15:26:52,060 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for a: $3, $4,
```

Success status of the pig script corresponding to the 2nd question

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgRe
duceTime MedianReduceTime Alias Feature Outputs
job_1495688372074_0007 1 0 3 3 3 3 0 0 0 a,b,c MAP_ONLY /
user/acadgild/MP2,

Input(s):
Successfully read 607 records (55517 bytes) from: "/user/acadgild/xmlcsv"

Output(s):
Successfully stored 349 records (10382 bytes) in: "/user/acadgild/MP2"

Counters:
Total records written : 349
Total bytes written : 10382
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1495688372074_0007

2017-05-25 15:27:09,944 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 15:27:09,954 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
cationStatus=SUCCEEDED. Redirecting to job history server
2017-05-25 15:27:10,007 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 15:27:10,013 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
cationStatus=SUCCEEDED. Redirecting to job history server
2017-05-25 15:27:10,062 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 15:27:10,076 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
cationStatus=SUCCEEDED. Redirecting to job history server
2017-05-25 15:27:10,135 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
grunt>
grunt> ■

```

The screenshot shows the Oracle VM VirtualBox interface with a running VM named "Acadgild_32_Bit". The terminal window displays the execution of a Pig script. The output shows the job statistics, including the number of records read from the input file "xmlcsv" (55517 bytes) and successfully stored in the output file "MP2" (10382 bytes). It also shows the application's final state as "SUCCEEDED" and the logs of the MapReduce tasks.

Displaying the stored output file in terminal

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 16:25 /user/acadgild/out13
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:07 /user/acadgild/out2
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:10 /user/acadgild/out5
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:29 /user/acadgild/out6
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:37 /user/acadgild/out7
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 15:58 /user/acadgild/out8
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 18:03 /user/acadgild/output2
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 18:43 /user/acadgild/output3
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 19:11 /user/acadgild/output4
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 19:15 /user/acadgild/output5
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 19:23 /user/acadgild/output6
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:03 /user/acadgild/output6_1a
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:12 /user/acadgild/output6_1aa
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:14 /user/acadgild/output6_1aaa
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:00 /user/acadgild/output6_1bb
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:18 /user/acadgild/output6_1bb
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 23:27 /user/acadgild/output6_1c
drwxr-xr-x 1 acadgild supergroup 0 2017-02-25 00:26 /user/acadgild/output6_2a
drwxr-xr-x 1 acadgild supergroup 0 2017-02-24 20:18 /user/acadgild/output7
drwxr-xr-x 1 acadgild supergroup 0 2017-04-22 06:56 /user/acadgild/p
-rw-r--r-- 1 acadgild supergroup 24 2017-03-24 13:54 /user/acadgild/pigdemo.txt
drwxr-xr-x 1 acadgild supergroup 0 2017-04-03 11:59 /user/acadgild/project
drwxr-xr-x 1 acadgild supergroup 0 2017-04-23 08:51 /user/acadgild/project2aout
drwxr-xr-x 1 acadgild supergroup 0 2017-04-03 13:57 /user/acadgild/projectb
-rw-r--r-- 1 acadgild supergroup 159 2017-05-03 14:44 /user/acadgild/wc.txt
drwxr-xr-x 1 acadgild supergroup 0 2017-05-25 11:55 /user/acadgild/xmlcsv
-rw-r--r-- 1 acadgild supergroup 2497496 2017-05-08 14:42 /user/acadgild/zipcode.csv
[acadgild@localhost Desktop]$ hadoop fs -ls /user/acadgild/MP2
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ight have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 15:30:58 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 2017-05-25 15:27 /user/acadgild/MP2/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 10382 2017-05-25 15:27 /user/acadgild/MP2/part-m-00000
[acadgild@localhost Desktop]$ ■

```

The screenshot shows the Oracle VM VirtualBox interface with a running VM named "Acadgild_32_Bit". The terminal window displays the command "hadoop fs -ls /user/acadgild/MP2" being run. The output lists the contents of the directory, including two files: "_SUCCESS" (0 bytes) and "part-m-00000" (10382 bytes). The terminal prompt ends with a yellow █ symbol.

Displaying the contents of the output file

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
[acadgild@localhost Desktop]$ hadoop fs -cat /user/acadgild/MP2/part-m-00000
[rw-r--r-- 1 acadgild supergroup 10382 2017-05-25 15:27 /user/acadgild/MP2/part-m-00000
[acadgild@localhost Desktop]$ JavaHotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/25 15:32:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Andhra Pradesh,ANANTAPUR,100.89262
Andhra Pradesh,CHITTOOR,99.9882
Andhra Pradesh,CUDDAPAH,95.282
Andhra Pradesh,EAST GODAVARI,93.80157
Andhra Pradesh,KARTHNAGAR,101.14053
Andhra Pradesh,KHAMMAM,103.45515
Andhra Pradesh,KRISHNA,90.65853
Andhra Pradesh,KURNool,84.38972
Andhra Pradesh,MEDAK,99.63046
Andhra Pradesh,NALGONDA,104.53599
Andhra Pradesh,NIZAMABAD,100.0
Andhra Pradesh,RANGAREDDI,82.04901
Andhra Pradesh,WARANGAL,108.923874
Andhra Pradesh,WEST GODAVARI,92.797844
Arunachal Pradesh,DIBANG VALLEY,100.2765
Arunachal Pradesh,LOHIT,95.56818
Arunachal Pradesh,TIRAP,100.0
Assam,BAGSHA,85.767296
Assam,CACHAHA,84.277626
Assam,DIBRUGARH,90.088394
Assam,GOALPARA,85.672356
Assam,GOLAGHAT,97.79542
Assam,HILAKANDI,100.0
Assam,JORHAT,88.98663
Assam,KAMRUP,98.76435
Assam,KARIMGANJ,95.95908
Assam,KOKRAJHAR,89.4601
Assam,LAKHIMPUR,83.06156
Assam,MARIGAON,92.95705
Assam,NAGAON,94.00146
[acadgild@localhost Desktop]$ mp2.pig (~/Desktop... acadgild@localhost... acadgild@localhost... acadgild@localhost... workspace - Java E... Right Ctrl
```

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
[acadgild@localhost Desktop]$ hadoop fs -cat /user/acadgild/MP2/part-m-00000
Uttar Pradesh,MORADABAD,100.0
Uttar Pradesh,MUZAFFARNAGAR,100.0
Uttar Pradesh,PILIBHIT,100.0
Uttar Pradesh,PRATAPGARH,101.52014
Uttar Pradesh,RAE BARELI,100.06516
Uttar Pradesh,RAMPUR,91.2306
Uttar Pradesh,SAHARANPUR,100.258804
Uttar Pradesh,SANT RAVIDAS NAGAR( BHADOHI),93.05768
Uttar Pradesh,SHAHJAHNPUR,100.161316
Uttar Pradesh,SHRAVASTI,94.145966
Uttar Pradesh,SIDDHARTHNAGAR,89.25478
Uttar Pradesh,SITAPUR,89.57219
Uttar Pradesh,SONBHADRA,100.0
Uttar Pradesh,SULTANPUR,100.0
Uttar Pradesh,UNNAO,97.400246
Uttar Pradesh,VARANASI,97.17479
Uttarakhand,BAGESHWAR,83.270386
Uttarakhand,CHAMOLI,95.93325
Uttarakhand,DEHRADUN,85.25207
Uttarakhand,HARIDWAR,89.05412
Uttarakhand,NAINITAL,126.02347
Uttarakhand,PITHORAGARH,84.15714
Uttarakhand,RUDRAPRAYAG,105.01901
Uttarakhand,TEHRI GARHWAL,83.905174
Uttarakhand,UDHAM SINGH NAGAR,95.37627
Uttarakhand,UTTARKASHI,90.54241
West Bengal,BARDHAMAN,85.9808
West Bengal,DAKSHIN DINAJPUR,100.8389
West Bengal,HOOGHLY,99.27945
West Bengal,HOWRAH,99.27974
West Bengal,JALPAIGURI,90.547165
West Bengal,MIDNAPUR EAST,134.4108
West Bengal,MIDNAPUR WEST,117.03546
West Bengal,NADIA,92.72157
West Bengal,NORTH 24 PARAGANAS,99.03116
West Bengal,SOUTH 24 PARAGANAS,94.43306
[acadgild@localhost Desktop]$ mp2.pig (~/Desktop... acadgild@localhost... acadgild@localhost... acadgild@localhost... workspace - Java E... Right Ctrl
```

Exporting the output to mysql table using sqoop command

```
[acadgild@localhost Desktop]$ sqoop export --connect jdbc:mysql://localhost/mysql --username root --password acadgild --table project2 --export-dir '/user/acadgild/MP2'
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop-2.0.4-alpha/.hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop-2.0.4-alpha/.accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop-2.0.4-alpha/.zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2017-05-25 15:45:43,791 INFO [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
2017-05-25 15:45:43,814 WARN [main] tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2017-05-25 15:45:44,028 INFO [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-05-25 15:45:44,028 INFO [main] tool.CodeGenTool: Beginning code generation
2017-05-25 15:45:44,285 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project2` AS t LIMIT 1
2017-05-25 15:45:44,323 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project2` AS t LIMIT 1
2017-05-25 15:45:44,332 INFO [main] org.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/hadoop-2.7.2
Note: /tmp/sqoop-acadgild/compile/8cf4dcff8f89a5564fcbb2dea929c8a89/project2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2017-05-25 15:45:47,647 INFO [main] org.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/8cf4dcff8f89a5564fcbb2dea929c8a89/project2.jar
2017-05-25 15:45:47,670 INFO [main] mapreduce.ExportJobBase: Beginning export of project2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
2017-05-25 15:45:47,922 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2017-05-25 15:45:47,929 INFO [main] Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2017-05-25 15:45:48,058 INFO [main] Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
2017-05-25 15:45:48,662 INFO [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
acadgild@localhost:~/Desktop$
```

Completion status of the sqoop export command to mysql

```
[acadgild@localhost Desktop]$ sqoop export --connect jdbc:mysql://localhost/mysql --username root --password acadgild --table project2 --export-dir '/user/acadgild/MP2'
FILE: Number of bytes read=0
FILE: Number of bytes written=545364
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=23086
HDFS: Number of bytes written=0
HDFS: Number of read operations=19
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
Launched map tasks=4
Data-local map tasks=4
Total time spent by all maps in occupied slots (ms)=60243
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=60243
Total vcore-seconds taken by all map tasks=60243
Total megabyte-seconds taken by all map tasks=61688832
Map-Reduce Framework
Map input records=349
Map output records=349
Input split bytes=601
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=758
CPU time spent (ms)=2420
Physical memory (bytes) snapshot=272084992
Virtual memory (bytes) snapshot=1293869056
Total committed heap usage (bytes)=65273856
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
2017-05-25 15:46:18,884 INFO [main] mapreduce.ExportJobBase: Transferred 22.5449 KB in 30.1668 seconds. (765.2773 bytes/sec)
2017-05-25 15:46:18,887 INFO [main] mapreduce.ExportJobBase: Exported 349 records.
[acadgild@localhost Desktop]$
```

Output of the mysql table with exported data

Acadgild_32_Bit [Running] - Oracle VM VirtualBox

```

File Machine View Input Devices Help
+----+----+
| 176 rows in set (0.00 sec)

mysql> create table project2(Statename varchar(100),districtname varchar(100),percentage float);
Query OK, 0 rows affected (0.06 sec)

mysql> select * from project2;
+----+----+----+
| Statename | districtname | percentage |
+----+----+----+
| Uttar Pradesh | BANDA | 92.4382 |
| Uttar Pradesh | BARABANKI | 101.396 |
| Uttar Pradesh | BAREILLY | 100 |
| Uttar Pradesh | BASTI | 81.1466 |
| Uttar Pradesh | BIJNOR | 100 |
| Uttar Pradesh | BUDAUN | 100 |
| Uttar Pradesh | BULANDSHAHAR | 110.77 |
| Uttar Pradesh | CHANDAULI | 104.668 |
| Uttar Pradesh | CHITRAKOOT | 112 |
| Uttar Pradesh | DEORIA | 95.0321 |
| Uttar Pradesh | ETAH | 111.513 |
| Uttar Pradesh | ETAWAH | 100 |
| Uttar Pradesh | FAIZABAD | 98.3188 |
| Uttar Pradesh | FARRUKHABAD | 100 |
| Uttar Pradesh | FATEHPUR | 95.7352 |
| Uttar Pradesh | FIROZABAD | 100 |
| Uttar Pradesh | GAUTAM BUDDHA NAGAR | 98.1337 |
| Uttar Pradesh | GHAZIABAD | 100 |
| Uttar Pradesh | GHAZIPUR | 99.1362 |
| Uttar Pradesh | GONDA | 86.8813 |
| Uttar Pradesh | GORAKHPUR | 99.9891 |
| Uttar Pradesh | HAMIRPUR | 84.0198 |
| Uttar Pradesh | HARDOI | 100 |
| Uttar Pradesh | JALAUN | 89.106 |
| Uttar Pradesh | JAUNPUR | 100.37 |
| Uttar Pradesh | JHANSI | 100.498 |
| Uttar Pradesh | JYOTIBA PHULE NAGAR | 100 |
+----+----+----+
41 Items in Trash

```

mp2.pig (~/Deskt... acadgild@localhost... acadgild@localhost... acadgild@localhost... workspace - Java E... Right Ctrl

A total of 349 records were inserted into the table

Acadgild_32_Bit [Running] - Oracle VM VirtualBox

```

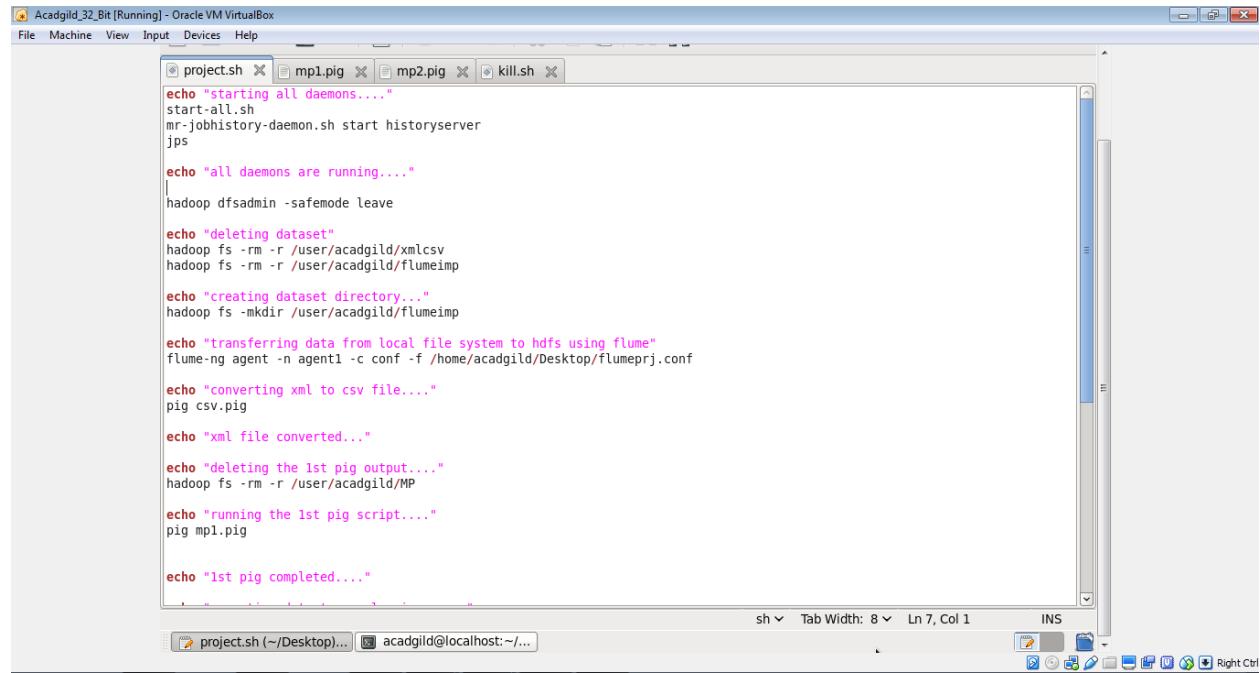
File Machine View Input Devices Help
+----+----+
| Gujarat | PORBANDAR | 100 |
| Gujarat | RAJKOT | 100.867 |
| Gujarat | SABAR KANTHA | 99.8712 |
| Gujarat | SURAT | 100 |
| Gujarat | SURENDRANAGAR | 83.9947 |
| Gujarat | VADODARA | 91.6543 |
| Gujarat | VALSAD | 127.909 |
| Haryana | AMBALA | 90.3726 |
| Haryana | BHILWANA | 100.613 |
| Haryana | FARIDABAD | 100 |
| Haryana | FATEHABAD | 98.9041 |
| Haryana | GURGAON | 140.867 |
| Haryana | HISAR | 100 |
| Haryana | JHAJJAR | 100 |
| Haryana | JIND | 98.3431 |
| Haryana | KAITHAL | 98.7043 |
| Haryana | KARNAL | 100.091 |
| Haryana | KURUKSHETRA | 100.271 |
| Haryana | MAHENDRAGARH | 100 |
| Haryana | MEWAT | 90.688 |
| Haryana | PANCHKULA | 100 |
| Haryana | PANIPAT | 100 |
| Haryana | REWARI | 88.4933 |
| Haryana | ROHTAK | 100 |
| Haryana | SIRSA | 100 |
| Haryana | SONIPAT | 101.651 |
| Haryana | YAMUNANAGAR | 98.8123 |
| Himachal Pradesh | BILASPUR | 109.614 |
| Himachal Pradesh | CHAMBA | 131.495 |
| Himachal Pradesh | HAMIRPUR | 100 |
| Himachal Pradesh | KANGRA | 137.378 |
| Himachal Pradesh | KINNAUR | 100 |
| Himachal Pradesh | KULLU | 100 |
+----+----+
349 rows in set (0.01 sec)
mysql> 
41 Items in Trash

```

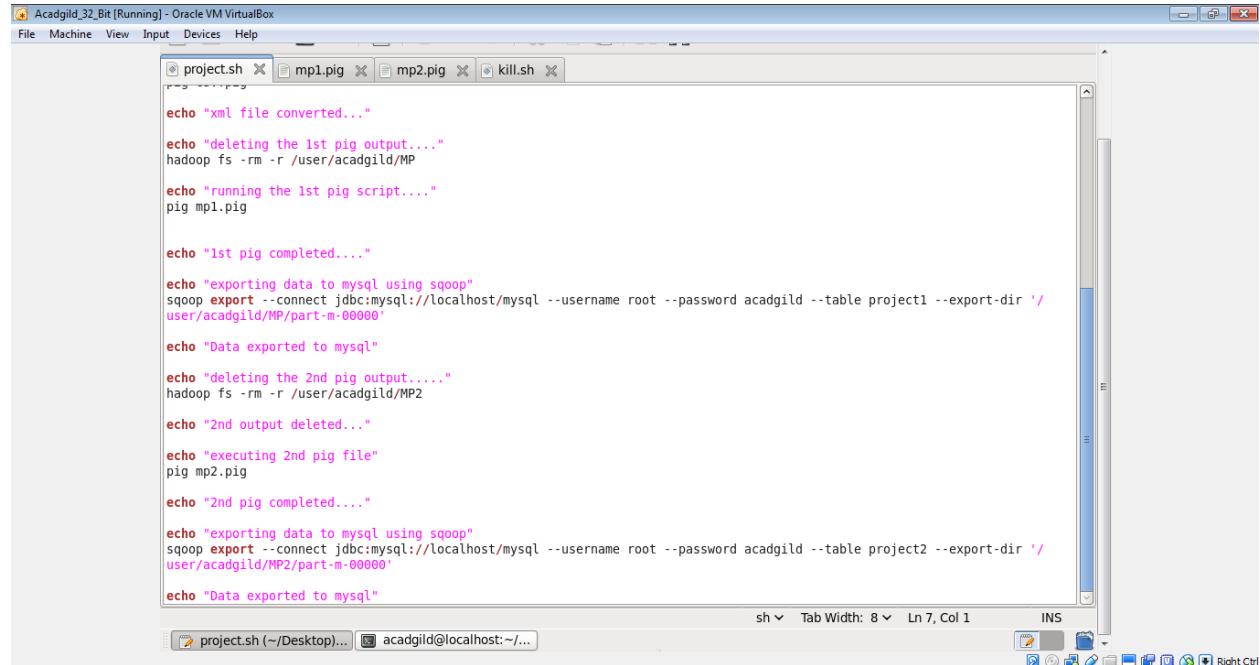
mp2.pig (~/Deskt... acadgild@localhost... acadgild@localhost... acadgild@localhost... workspace - Java E... Right Ctrl

Project using shell scripting

Shell script

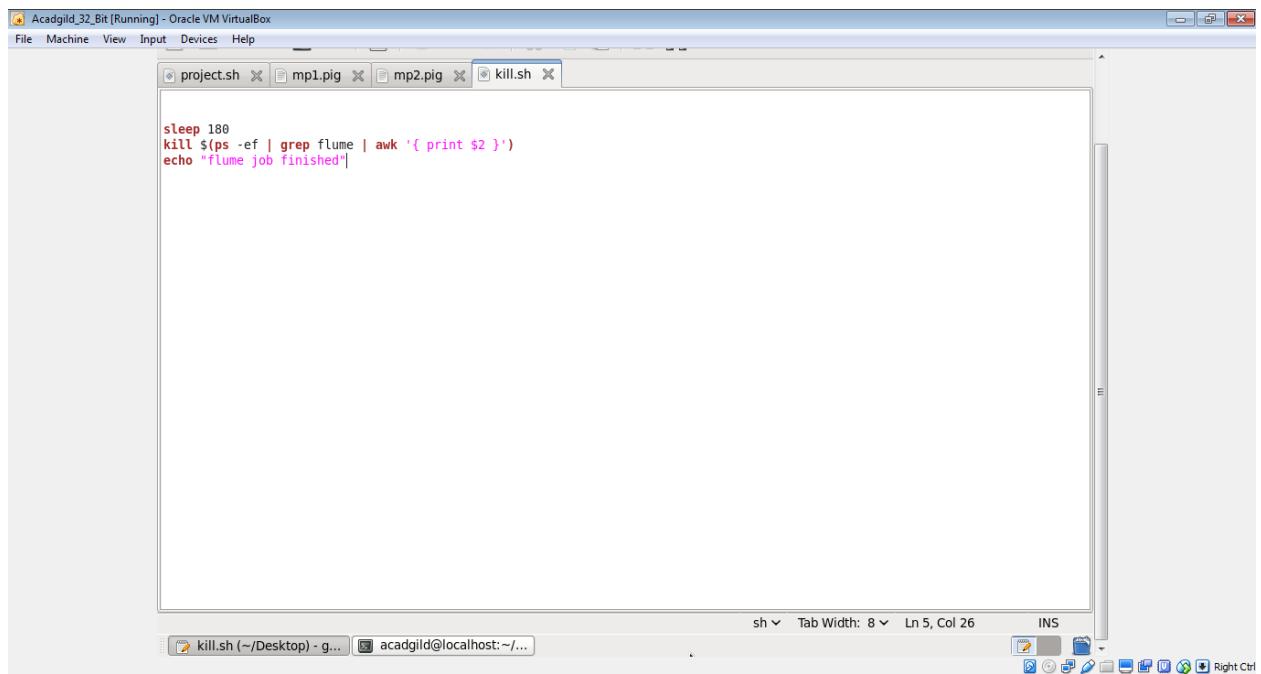


```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
project.sh mp1.pig mp2.pig kill.sh
echo "starting all daemons...."
start-all.sh
mr-jobhistory-daemon.sh start historyserver
jps
echo "all daemons are running...."
hadoop dfsadmin -safemode leave
echo "deleting dataset"
hadoop fs -rm -r /user/acadgild/xmlcsv
hadoop fs -rm -r /user/acadgild/flumeimp
echo "creating dataset directory..."
hadoop fs -mkdir /user/acadgild/flumeimp
echo "transferring data from local file system to hdfs using flume"
flume-ng agent -n agent1 -c conf -f /home/acadgild/Desktop/flumeprj.conf
echo "converting xml to csv file...."
pig csv.pig
echo "xml file converted..."
echo "deleting the 1st pig output...."
hadoop fs -rm -r /user/acadgild/MP
echo "running the 1st pig script...."
pig mp1.pig
echo "1st pig completed...."
```



```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
project.sh mp1.pig mp2.pig kill.sh
echo "xml file converted..."
echo "deleting the 1st pig output...."
hadoop fs -rm -r /user/acadgild/MP
echo "running the 1st pig script...."
pig mp1.pig
echo "1st pig completed...."
echo "exporting data to mysql using sqoop"
sqoop export --connect jdbc:mysql://localhost/mysql --username root --password acadgild --table project1 --export-dir '/user/acadgild/MP/part-m-00000'
echo "Data exported to mysql"
echo "deleting the 2nd pig output...."
hadoop fs -rm -r /user/acadgild/MP2
echo "2nd output deleted..."
echo "executing 2nd pig file"
pig mp2.pig
echo "2nd pig completed...."
echo "exporting data to mysql using sqoop"
sqoop export --connect jdbc:mysql://localhost/mysql --username root --password acadgild --table project2 --export-dir '/user/acadgild/MP2/part-m-00000'
echo "Data exported to mysql"
```

Script to kill the flume data extraction command



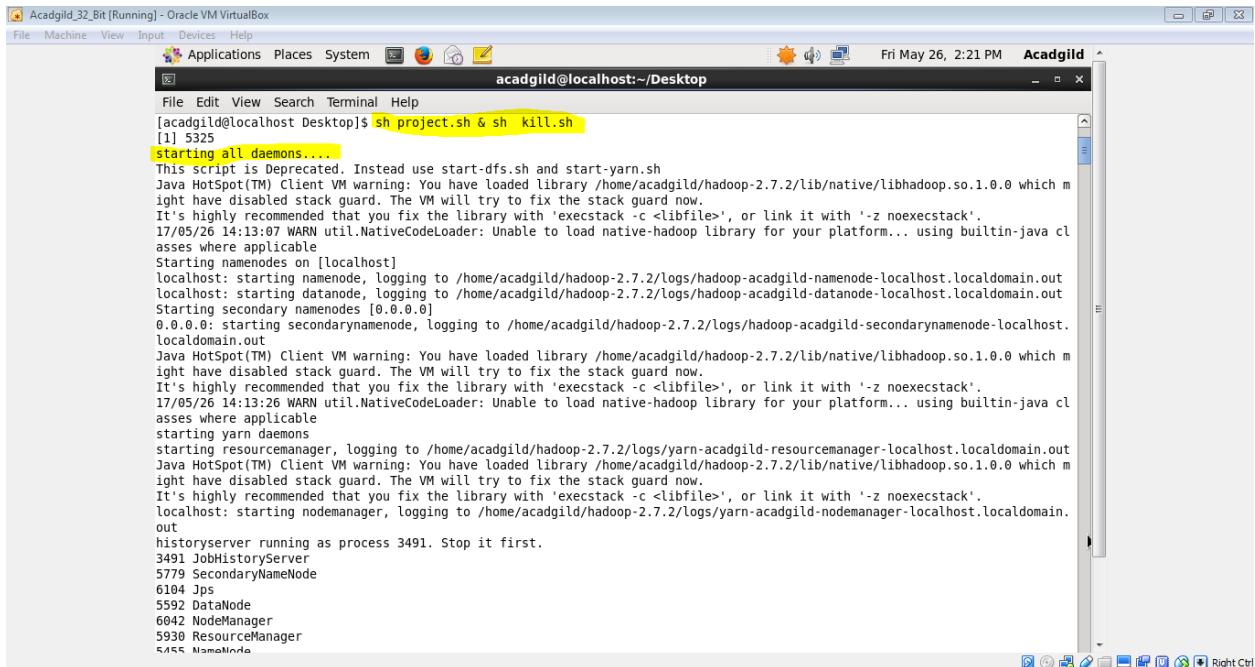
The screenshot shows a terminal window titled "Acadgild_32_Bit [Running] - Oracle VM VirtualBox". The window contains four tabs: "project.sh", "mp1.pig", "mp2.pig", and "kill.sh". The "kill.sh" tab is active and displays the following script:

```
sleep 180
kill $(ps -ef | grep flume | awk '{ print $2 }')
echo "flume job finished"
```

The terminal interface includes a status bar at the bottom showing "sh", "Tab Width: 8", "Ln 5, Col 26", and "INS". Below the status bar is a toolbar with various icons.

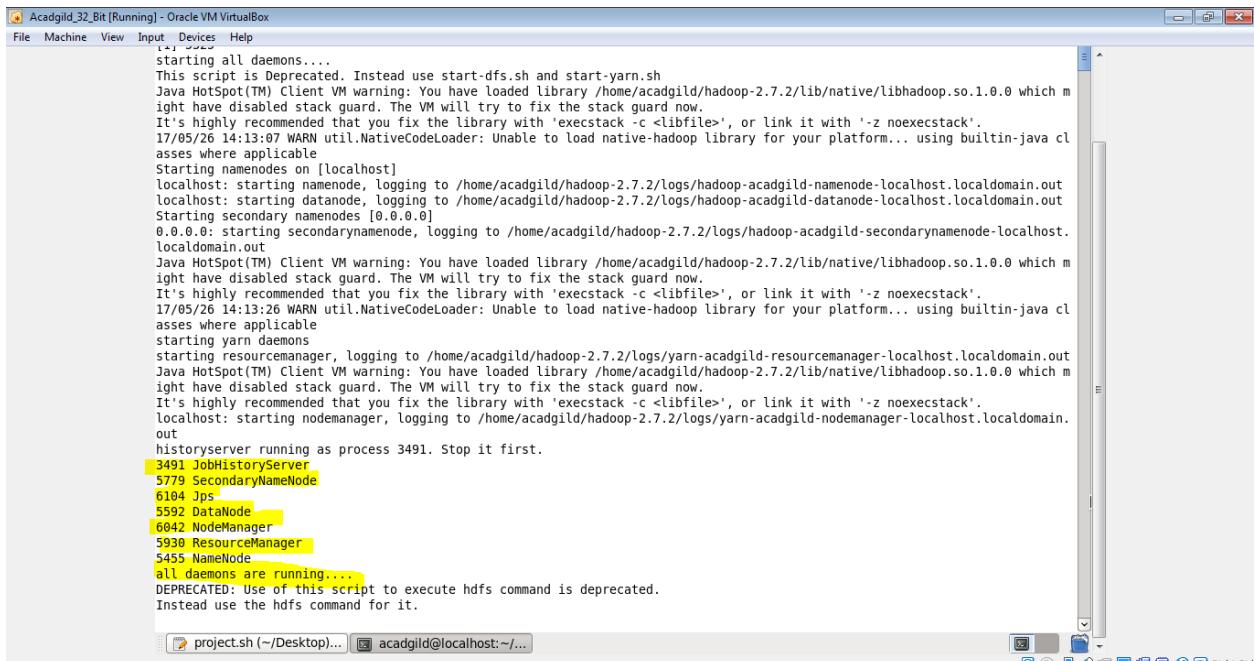
Stages of the project

Stage 1: String the shell script



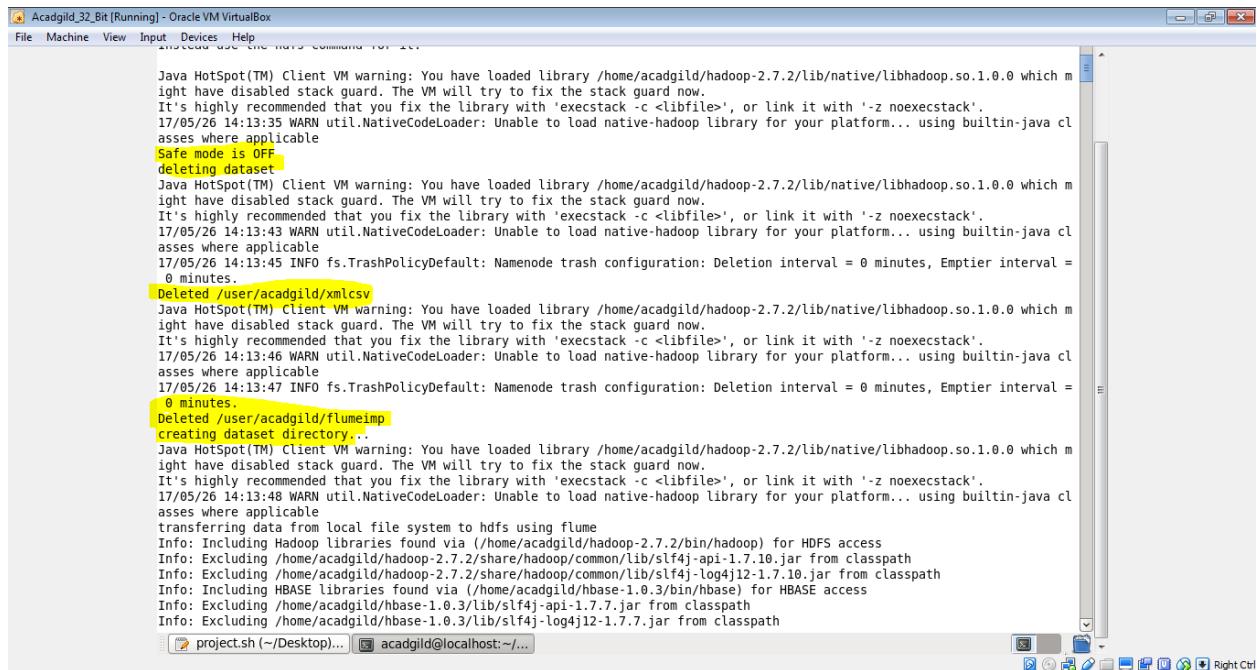
```
[acadgild@localhost Desktop]$ sh project.sh & sh kill.sh
[1] 5325
starting all daemons....
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/acadgild/hadoop-2.7.2/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /home/acadgild/hadoop-2.7.2/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/acadgild/hadoop-2.7.2/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/acadgild/hadoop-2.7.2/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
localhost: starting nodemanager, logging to /home/acadgild/hadoop-2.7.2/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
historyserver running as process 3491. Stop it first.
3491 JobHistoryServer
5779 SecondaryNameNode
6104 Jps
5592 DataNode
6042 NodeManager
5930 ResourceManager
5455 NameNode
```

Stage 2: Starting all daemons



```
starting all daemons....
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/acadgild/hadoop-2.7.2/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /home/acadgild/hadoop-2.7.2/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/acadgild/hadoop-2.7.2/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/acadgild/hadoop-2.7.2/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
localhost: starting nodemanager, logging to /home/acadgild/hadoop-2.7.2/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
historyserver running as process 3491. Stop it first.
3491 JobHistoryServer
5779 SecondaryNameNode
6104 Jps
5592 DataNode
6042 NodeManager
5930 ResourceManager
5455 NameNode
all daemons are running ...
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

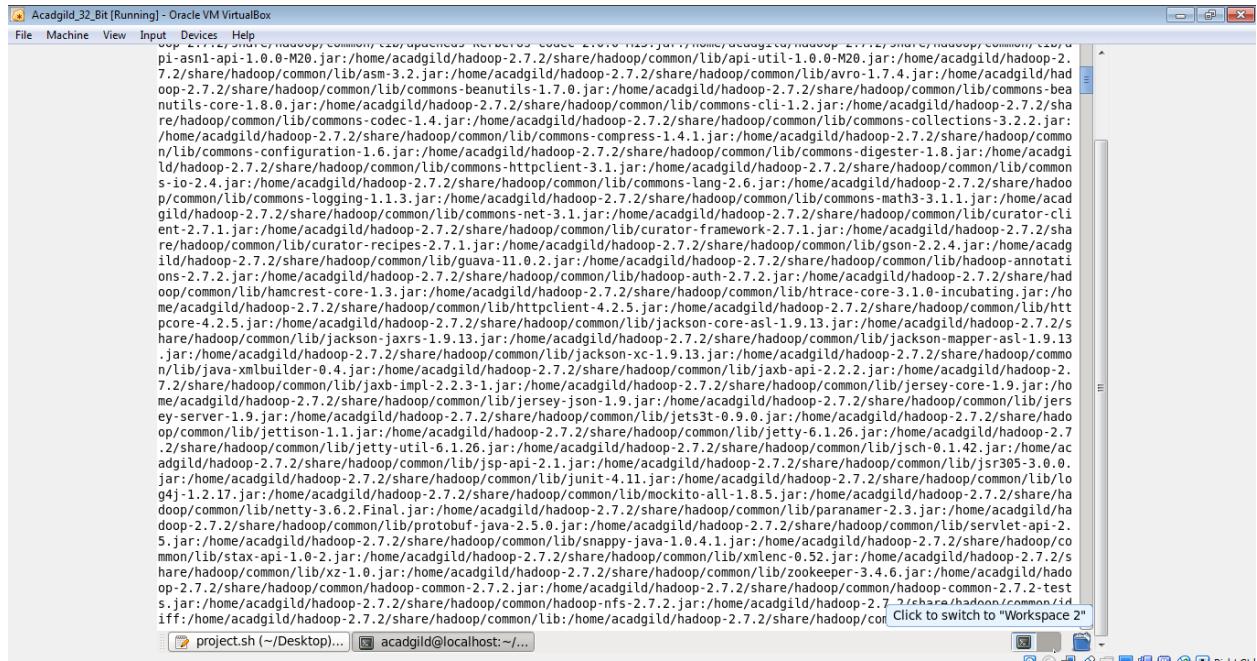
Stage 3: Deleting the flume output directory



```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
instantiated due to the 'java' command for it.

Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Safe mode is OFF
deleting dataset
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/05/26 14:13:45 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/acadgild/xmlcsv
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/05/26 14:13:47 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/acadgild/flumeimp
creating dataset directory...
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:13:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
transferring data from local file system to hdfs using flume
Info: Including Hadoop libraries found via (/home/acadgild/hadoop-2.7.2/bin/hadoop) for HDFS access
Info: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath
Info: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath
Info: Including HBASE libraries found via (/home/acadgild/hbase-1.0.3/bin/hbase) for HBASE access
Info: Excluding /home/acadgild/hbase-1.0.3/lib/slf4j-api-1.7.7.jar from classpath
Info: Excluding /home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar from classpath
[project.sh (~/Desktop)...] acadgild@localhost:~/...
```

Stage 4: Starting the flume conf file



```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
pl-ashn1-api-1.0.0-M20.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/asm-3.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/avro-1.7.4.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-cli-1.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-codec-1.4.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-compress-1.4.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-configuration-1.6.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-digester-1.8.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-httpclient-3.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-lang-2.6.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-logging-1.1.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-math3-3.1.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/commons-net-3.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/curator-client-2.7.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/curator-framework-2.7.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/curator-recipes-2.7.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/gson-2.4.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/guava-11.0.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-annotations-2.7.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-auth-2.7.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hmcrest-core-1.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/httppre-4.2.5.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-jaxrs-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jersey-core-1.9.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jersey-server-1.9.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jets3t-0.9.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jettison-1.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jetty-6.1.26.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jetty-util-6.1.26.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jscn-0.1.42.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/jsp-api-2.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/junit-4.11.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/loq4j-1.2.17.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/mockito-all-1.8.5.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/netty-3.6.2.Final.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/paranamer-2.3.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/protoBuf-java-2.5.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/servlet-api-2.5.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/snappy-java-1.0.4.1.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/stax-api-1.0-2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/xmlenc-0.52.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/xz-1.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/zookeeper-3.4.6.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-common-2.7.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-nfs-2.7.2.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-common-2.7.2-test5.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-common-2.7.2-test5.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/hadoop-conf Click to switch to "Workspace 2"
[project.sh (~/Desktop)...] acadgild@localhost:~/...
```

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
at org.apache.flume.conf.FlumeConfiguration.access$9000(FlumeConfiguration.java:129)
at org.apache.flume.conf.FlumeConfiguration.validateConfiguration(FlumeConfiguration.java:127)
at org.apache.flume.conf.FlumeConfiguration.<init>(FlumeConfiguration.java:109)
at org.apache.flume.node.PropertiesFileConfigurationProvider.getFlumeConfiguration(PropertiesFileConfigurationProvide
r.java:189)
at org.apache.flume.node.AbstractConfigurationProvider.getConfiguration(AbstractConfigurationProvider.java:89)
at org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConf
igurationProvider.java:140)
at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
at java.util.concurrent.FutureTask.runAndReset(FutureTask.java:308)
at java.util.concurrent.ScheduledThreadPoolExecutor$ScheduledFutureTask.access$301(ScheduledThreadPoolExecutor.java:1
80)
at java.util.concurrent.ScheduledThreadPoolExecutor$ScheduledFutureTask.run(ScheduledThreadPoolExecutor.java:294)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
17/05/26 14:13:53 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agent
1]
17/05/26 14:13:53 INFO node.AbstractConfigurationProvider: Creating channels
17/05/26 14:13:53 INFO channel.DefaultChannelFactory: Creating instance of channel mychannel type memory
17/05/26 14:13:53 INFO node.AbstractConfigurationProvider: Created channel mychannel
17/05/26 14:13:53 INFO source.DefaultSourceFactory: Creating instance of source mysrc, type exec
17/05/26 14:13:53 INFO node.AbstractConfigurationProvider: Channel mychannel connected to [mysrc]
17/05/26 14:13:53 INFO node.Application: Starting new configuration:{ sourceRunners:{mysrc=EventDrivenSourceRunner: { source:
org.apache.flume.source.ExecSource{name:mysrc,state:IDLE} }} sinkRunners:{} channels:{mychannel=org.apache.flume.channel.Memo
ryChannel{name:mymychannel} }
17/05/26 14:13:53 INFO node.Application: Starting Channel mychannel
17/05/26 14:13:54 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel: Suc
cessfully registered new MBean.
17/05/26 14:13:54 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
17/05/26 14:13:54 INFO node.Application: Starting Source mysrc
17/05/26 14:13:54 INFO source.ExecSource: Exec source starting with command:hadoop dfs -put /home/acadgild/Desktop/StatewiseD
istrictwisePhysicalProgress.xml /user/acadgild/flumeimp
17/05/26 14:13:54 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: mysrc: Successf
ully registered new MBean.
17/05/26 14:13:54 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc started
17/05/26 14:13:57 INFO source.ExecSource: Command [hadoop dfs -put /home/acadgild/Desktop/StatewiseDistrictwisePhysicalProgr
ss.xml /user/acadgild/flumeimp] exited with 0
project.sh (~/Desktop)... acadgild@localhost:~...

```

Stage 5: Converting dataset from xml file format to csv file format

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
rt.time == 1495788234015
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.sto
p.time == 1495788366070
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.cap
acity == 100
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.cur
rent.size == 0
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.eve
nt.put.attempt == 0
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.eve
nt.put.success == 0
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.eve
nt.take.attempt == 0
17/05/26 14:16:06 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: mychannel. channel.eve
nt.take.success == 0
converting xml to csv file...
2017-05-26 14:16:07,596 INFO [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-05-26 14:16:07,597 INFO [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-05-26 14:16:07,598 INFO [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-05-26 14:16:07,676 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2017-05-26 14:16:07,677 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/Desktop/pig_149578836767
5.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.c
lass]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/imp
l/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ight have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
2017-05-26 14:16:07,933 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
2017-05-26 14:16:08,131 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found

```

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
mpstats
2017-05-26 14:16:33,233 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.5.1 0.16.0 acadgild 2017-05-26 14:16:10 2017-05-26 14:16:33 UNKNOWN

Success!

Job Stats (time in seconds):
JobID Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgRe
duceTime MedianReduceTime Alias Feature Outputs
job_1495788209826_0001 1 0 4 4 4 0 0 0 A,B,D MAP_ONLY /
user/acadgild/xmlcsv,

Input(s):
Successfully read 607 records (717827 bytes) from: "/user/acadgild/flumeimp/statewiseDistrictwisePhysicalProgress.xml"

Output(s):
Successfully stored 607 records (55135 bytes) in: "/user/acadgild/xmlcsv"

Counters:
Total records written : 607
Total bytes written : 55135
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1495788209826_0001

2017-05-26 14:16:33,236 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:33,239 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:33,300 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:33,306 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:33,356 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:33,369 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:33,434 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-05-26 14:16:33,475 [main] INFO org.apache.pig.Main - Pig script completed in 26 seconds and 14 milliseconds (26014 ms)
xml file converted...
deleting the 1st pig output...
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ight have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:16:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/05/26 14:16:36 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval =
0 minutes.
Deleted /user/acadgild/MP
running the 1st pig script...
2017-05-26 14:16:37,964 INFO [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-05-26 14:16:37,966 INFO [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-05-26 14:16:37,967 INFO [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-05-26 14:16:38,035 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2017-05-26 14:16:38,035 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/Desktop/pig_149578839803
3.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.c
lass]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/] 41 Items in Trash

```

Stage 6: Deleting the output of 1st pig script and starting the pig script corresponding to 1st question

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
project.sh (~/Desktop)... acadgild@localhost:~...
job_1495788209826_0001

2017-05-26 14:16:33,236 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:33,239 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:33,300 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:33,306 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:33,356 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:33,369 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:33,434 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-05-26 14:16:33,475 [main] INFO org.apache.pig.Main - Pig script completed in 26 seconds and 14 milliseconds (26014 ms)
xml file converted...
deleting the 1st pig output...
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ight have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/05/26 14:16:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/05/26 14:16:36 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval =
0 minutes.
Deleted /user/acadgild/MP
running the 1st pig script...
2017-05-26 14:16:37,964 INFO [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-05-26 14:16:37,966 INFO [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-05-26 14:16:37,967 INFO [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-05-26 14:16:38,035 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2017-05-26 14:16:38,035 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/Desktop/pig_149578839803
3.log
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.c
lass]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/] 41 Items in Trash

```

Stage 7: Status of the 1st pig script execution

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
2017-05-26 14:16:57,685 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:57,685 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
cationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:57,765 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% co
mplete
2017-05-26 14:16:57,769 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.5.1 0.16.0 acadgild 2017-05-26 14:16:40 2017-05-26 14:16:57 FILTER

Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgRe
duceTime MedianReducetime Alias Feature Outputs
job_1495788209826_0002 1 0 4 4 4 4 0 0 0 a,b MAP_ONLY /
user/acadgild/MP,
Input(s):
Successfully read 607 records (55517 bytes) from: "/user/acadgild/xmlcsv/part-m-00000"
Output(s):
Successfully stored 176 records (21787 bytes) in: "/user/acadgild/MP"

Counters:
Total records written : 176
Total bytes written : 21787
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1495788209826_0002

2017-05-26 14:16:57,770 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:57,774 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
cationStatus=SUCCEEDED. Redirecting to job history server
[project.sh (~/Desktop)... acadgild@localhost:~...]

```

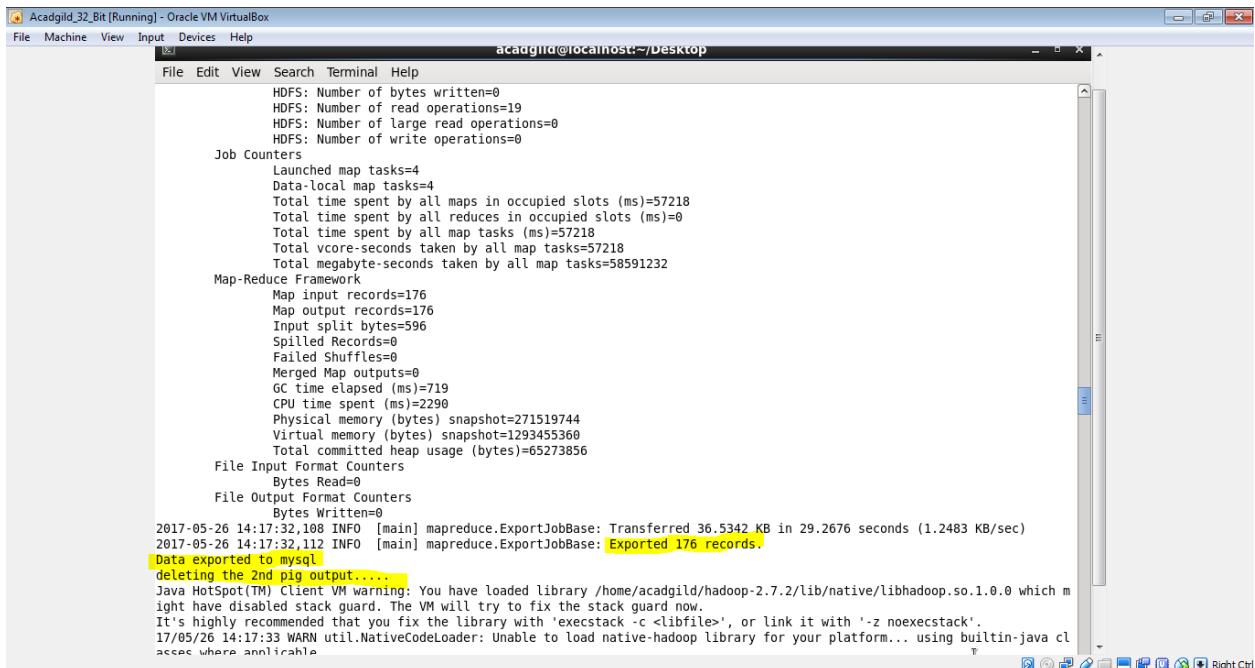
Stage 8: exporting the output to mysql table using sqoop command

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
project.sh (~/Desktop)... acadgild@localhost:~...
2017-05-26 14:16:57,887 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:16:57,892 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
cationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:16:57,955 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-05-26 14:16:57,979 [main] INFO org.apache.pig.Main - Pig script completed in 20 seconds and 146 milliseconds (20140 ms)
1st pig completed...
exporting data to mysql using sqoop ...
Warning: /home/acadgild/sqoop-1.4.6.bin_hadoop-2.0.4-alpha/..hcatalog does not exist! HCatalog jobs will fail.
Please set $CAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/sqoop-1.4.6.bin_hadoop-2.0.4-alpha/..accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/acadgild/sqoop-1.4.6.bin_hadoop-2.0.4-alpha/..zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2017-05-26 14:16:59,394 INFO [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
2017-05-26 14:16:59,415 WARN [main] tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider usin
g -P instead.
2017-05-26 14:16:59,639 INFO [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-05-26 14:16:59,639 INFO [main] tool.CodeGenTool: Beginning code generation
2017-05-26 14:16:59,966 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project1` AS t LIMIT 1
2017-05-26 14:17:00,022 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project1` AS t LIMIT 1
2017-05-26 14:17:00,034 INFO [main] org.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/hadoop-2.7.2
Note: /tmp/sqoop-acadgild/compile/9332855a524af7d6622d5d67310d1df5/project1.java uses or overrides deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2017-05-26 14:17:01,670 INFO [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/9332855a524af7d66a
2d5d67310d1df5/project1.jar
2017-05-26 14:17:01,701 INFO [main] mapreduce.ExportJobBase: Beginning export of project1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.c
lass]
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/imp
l/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ight have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
2017-05-26 14:17:01,933 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
[project.sh (~/Desktop)... acadgild@localhost:~...]

```

Stage 9: Deleting the output of 2nd pig script

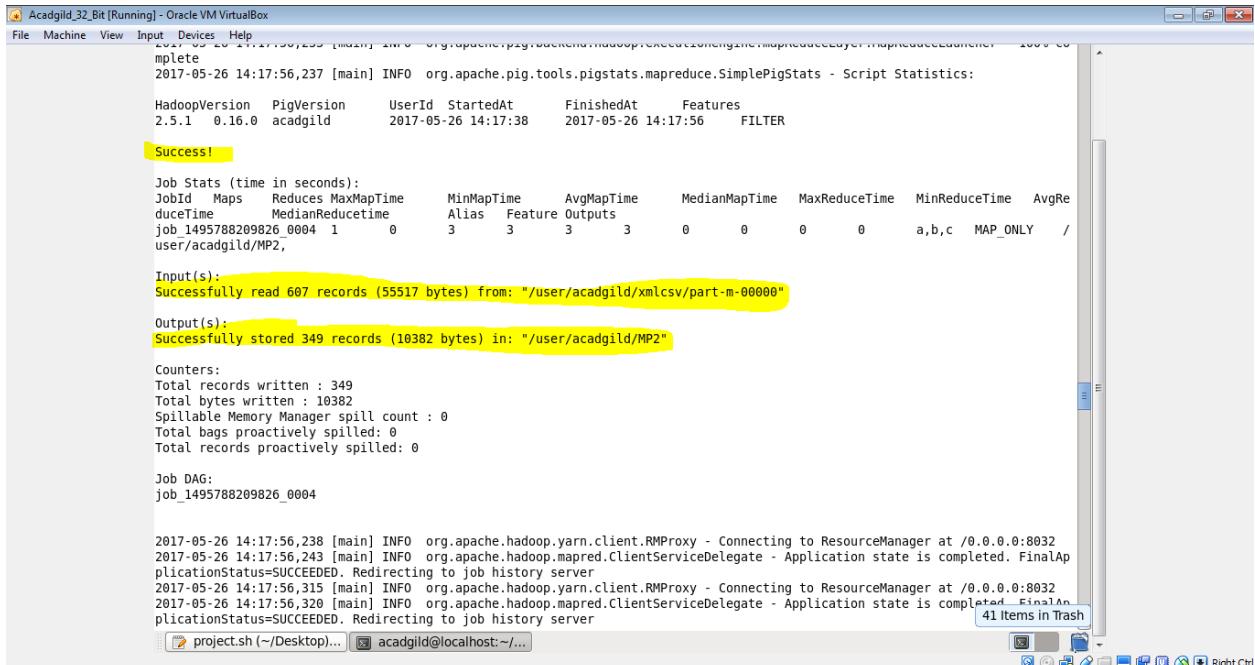


```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
HDFS: Number of bytes written=0
HDFS: Number of read operations=19
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
Launched map tasks=4
Data-local map tasks=4
Total time spent by all maps in occupied slots (ms)=57218
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=57218
Total vcore-seconds taken by all map tasks=57218
Total megabyte-seconds taken by all map tasks=58591232
Map-Reduce Framework
Map input records=176
Map output records=176
Input split bytes=596
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=719
CPU time spent (ms)=2290
Physical memory (bytes) snapshot=271519744
Virtual memory (bytes) snapshot=1293455360
Total committed heap usage (bytes)=65273856
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
2017-05-26 14:17:32,108 INFO [main] mapreduce.ExportJobBase: Transferred 36.5342 KB in 29.2676 seconds (1.2483 KB/sec)
2017-05-26 14:17:32,112 INFO [main] mapreduce.ExportJobBase: Exported 176 records.
Data exported to mysql
deleting the 2nd pig output.....
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <library>', or link it with '-z noexecstack'.
17/05/26 14:17:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

```

Stage 10: Success status of the 2nd pig script



```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
2017-05-26 14:17:56,233 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduce.MapReduceLauncher - 100% CO
mpleted
2017-05-26 14:17:56,237 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.5.1 0.16.0 acadgild 2017-05-26 14:17:38 2017-05-26 14:17:56 FILTER
Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgRe
duceTime MedianReducetime Alias Feature Outputs
job_1495788209826_0004 1 0 3 3 3 3 0 0 0 a,b,c MAP_ONLY /
user/acadgild/MP2,
Input(s):
Successfully read 607 records (55517 bytes) from: "/user/acadgild/xmlcsv/part-m-00000"
Output(s):
Successfully stored 349 records (10382 bytes) in: "/user/acadgild/MP2"
Counters:
Total records written : 349
Total bytes written : 10382
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1495788209826_0004

2017-05-26 14:17:56,238 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:17:56,243 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:17:56,315 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:17:56,320 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
41 Items in Trash
project.sh (~/Desktop)... acadgild@localhost:~/...

```

Stage 11: Exporting the output of pig script to mysql table using sqoop command

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Job DAG:
job_1495788209826_0004

2017-05-26 14:17:56,238 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:17:56,243 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:17:56,315 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:17:56,320 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:17:56,377 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-05-26 14:17:56,383 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-05-26 14:17:56,457 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success !
2017-05-26 14:17:56,480 [main] INFO org.apache.pig.Main - Pig script completed in 21 seconds and 60 milliseconds. (21060 ms)
2nd pig completed...
exporting data to mysql using sqoop
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop-2.0.4-alpha/../.hcatalog does not exist! HCatalog jobs will fail.
Please set $CAT_HOME to the root of your HCatalog installation.
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop-2.0.4-alpha/../.accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/acadgild/sqoop-1.4.6-bin_hadoop-2.0.4-alpha/../.zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2017-05-26 14:17:57,928 INFO [main] Sqoop.Sqoop: Running Sqoop version: 1.4.6
2017-05-26 14:17:57,951 WARN [main] tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -p instead.
2017-05-26 14:17:58,189 INFO [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-05-26 14:17:58,189 INFO [main] tool.CodeGenTool: Beginning code generation
2017-05-26 14:17:58,492 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project2` AS t LIMIT 1
2017-05-26 14:17:58,546 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `project2` AS t LIMIT 1
2017-05-26 14:17:58,561 INFO [main] org.CompilationManager: HADOOP_MAPRED_HOME is /home/acadgild/hadoop-2.7.2
Note: /tmp/sqoop-acadgild/compile/195f962047d75bb98ba08fe5ba9c27f3/project2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2017-05-26 14:18:00,334 INFO [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/195f962047d75bb98ba08fe5ba9c27f3/project2.jar
2017-05-26 14:18:00,357 INFO [main] mapreduce.ExportJobBase: Beginning export of project2
[ project.sh (~/Desktop)... acadgild@localhost:~... ] Click to switch to "Workspace 2"

```

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=23086
HDFS: Number of bytes written=0
HDFS: Number of read operations=19
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
Launched map tasks=4
Data-local map tasks=4
Total time spent by all maps in occupied slots (ms)=57794
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=57794
Total vcore-seconds taken by all map tasks=57794
Total megabyte-seconds taken by all map tasks=59181056
Map-Reduce Framework
Map input records=349
Map output records=349
Input split bytes=601
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=783
CPU time spent (ms)=2490
Physical memory (bytes) snapshot=272142336
Virtual memory (bytes) snapshot=1293869056
Total committed heap usage (bytes)=65273856
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
2017-05-26 14:18:30,923 INFO [main] mapreduce.ExportJobBase: Transferred 22.5449 KB in 29.4646 seconds. (783.5173 bytes/sec)
2017-05-26 14:18:30,927 INFO [main] mapreduce.ExportJobBase: Exported 349 records.
Data exported to mysql
^C
[1]+ Done sh project.sh
[acadgild@localhost Desktop]$ sudo service mysqld start
[sudo] password for acadgild:
[ project.sh (~/Desktop)... acadgild@localhost:~... ] Click to switch to "Workspace 2"

```

End of project using shell scripting

Displaying the output

Output of 1st question

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
[acadgild@localhost Desktop]$ sudo service mysql start
[sudo] password for acadgild:
Starting mysqld:                                         [ OK ]
[acadgild@localhost Desktop]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 49
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> use mysql;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from project;
+-----+-----+
| statename | districtname |
+-----+-----+
| Madhya Pradesh | NEEMUCH |
| Madhya Pradesh | RAISEN |
| Madhya Pradesh | RAJGARH |
| Madhya Pradesh | RATLAM |
| Madhya Pradesh | REWA |
| Madhya Pradesh | SEHORE |
| Madhya Pradesh | SHAHDOL |
| Madhya Pradesh | SHAJAPUR |
| Madhya Pradesh | SHEOPUR |
| Madhya Pradesh | UJJAIN |
| Madhya Pradesh | UMARIA |
+-----+-----+
project.sh (~/Desktop)...  acadgild@localhost:~/...
```

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Kerala
| Kerala | ALAPPUZHA
| Kerala | ERIAKULAM
| Kerala | IDUKKI
| Kerala | KANNUR
| Kerala | KASARGOD
| Kerala | KOLLAM
| Kerala | KOTTAYAM
| Kerala | KOZHIKODE
| Kerala | MALAPPURAM
| Kerala | PALAKKAD
| Kerala | PATHANAMTHITTA
| Kerala | THIRUVANANTHAPURAM
| Kerala | TIRISSUR
| Kerala | WAYANAD
Madhya Pradesh
| Madhya Pradesh | ALIRAJPUR
| Madhya Pradesh | ANUPPUR
| Madhya Pradesh | BARWANI
| Madhya Pradesh | BETUL
| Madhya Pradesh | BHOPAL
| Madhya Pradesh | BURHANPUR
| Madhya Pradesh | DATIA
| Madhya Pradesh | DEWAS
| Madhya Pradesh | DHAR
| Madhya Pradesh | GWALIOR
| Madhya Pradesh | HARDA
| Madhya Pradesh | HOSHANGABAD
| Madhya Pradesh | INDORE
| Madhya Pradesh | JABALPUR
| Madhya Pradesh | JHABUA
| Madhya Pradesh | KATNI
| Madhya Pradesh | KHANDWA(EAST NIMAR)
| Madhya Pradesh | KHARGONE
| Madhya Pradesh | MANDSAUR
| Madhya Pradesh | MORENA
+-----+
176 rows in set (0.00 sec)

project.sh (~/Desktop)...  acadgild@localhost:~/...
```

Output of 2nd question

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
+----+-----+-----+
| statename | districtname | percent |
+----+-----+-----+
| Madhya Pradesh | MANDSAUR | 100
| Madhya Pradesh | MORENA | 100
+----+-----+
176 rows in set (0.00 sec)

mysql> select * from project2;
+----+-----+-----+
| statename | districtname | percent |
+----+-----+-----+
| Maharashtra | GADCHIROLI | 100
| Maharashtra | GONDIA | 94.4182
| Maharashtra | HINGOLI | 91.2322
| Maharashtra | JALNA | 97.2189
| Maharashtra | KOLHAPUR | 86.2574
| Maharashtra | NAGPUR | 88.3415
| Maharashtra | OSMANABAD | 99.9779
| Maharashtra | PARBHANI | 87.1717
| Maharashtra | PUNE | 97.7041
| Maharashtra | RATNAGIRI | 106.744
| Maharashtra | SANGLI | 82.9894
| Maharashtra | SATARA | 85.8256
| Maharashtra | SINDHUDURG | 100
| Maharashtra | THANE | 95.5074
| Maharashtra | WARDHA | 92.7833
| Manipur | BISHNUPUR | 87.0639
| Manipur | IMPHAL EAST | 95.1551
| Manipur | TAMENGLONG | 88.5351
| Meghalaya | RI BHOI | 84.4249
| Meghalaya | SOUTH GARO HILLS | 90.5283
| Meghalaya | WEST GARO HILLS | 100
| Mizoram | CHAMPHAI | 100
| Mizoram | KOLASIB | 98.0867
| Mizoram | LAWNGT LAI | 100
| Mizoram | LUNGLEI | 89.362
| Mizoram | MAMIT | 97.7102
| Mizoram | SAIHA | 88.6571
| Mizoram | SERCHHIP | 88.9024
+----+-----+
project.sh (~/Desktop)... acadgild@localhost:~/...
```

```
Acadgild_32_Bit [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
+----+-----+-----+
| Gujarat | PORBANDAR | 100
| Gujarat | RAJKOT | 100.867
| Gujarat | SABAR KANTHA | 99.8712
| Gujarat | SURAT | 100
| Gujarat | SURENDRANAGAR | 83.9947
| Gujarat | VADODARA | 91.6543
| Gujarat | VALSAD | 127.909
| Haryana | AMBALA | 96.3726
| Haryana | BHIWANI | 100.613
| Haryana | FARIDABAD | 100
| Haryana | FATEHABAD | 98.9041
| Haryana | GURGAON | 140.867
| Haryana | HISAR | 100
| Haryana | JHAJJAR | 100
| Haryana | JIND | 98.3431
| Haryana | KAITHAL | 98.7043
| Haryana | KARNAL | 100.091
| Haryana | KURUKSHETRA | 100.271
| Haryana | MAHENDRAGARH | 100
| Haryana | MEWAT | 90.688
| Haryana | PANCHKULA | 100
| Haryana | PANIPAT | 100
| Haryana | REWARI | 88.4933
| Haryana | ROHTAK | 100
| Haryana | SIRSA | 100
| Haryana | SONIPAT | 101.651
| Haryana | YAMUNANAGAR | 98.8123
| Himachal Pradesh | BILASPUR | 109.614
| Himachal Pradesh | CHAMBA | 131.495
| Himachal Pradesh | HAMIRPUR | 100
| Himachal Pradesh | KANGRA | 137.378
| Himachal Pradesh | KINNAUR | 100
| Himachal Pradesh | KULLU | 100
+----+-----+
349 rows in set (0.00 sec)

mysql> 
project.sh (~/Desktop)... acadgild@localhost:~/...
```

Pig files

XML to CSV pig script

```
A = LOAD '/user/acadgild/flumeimp/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

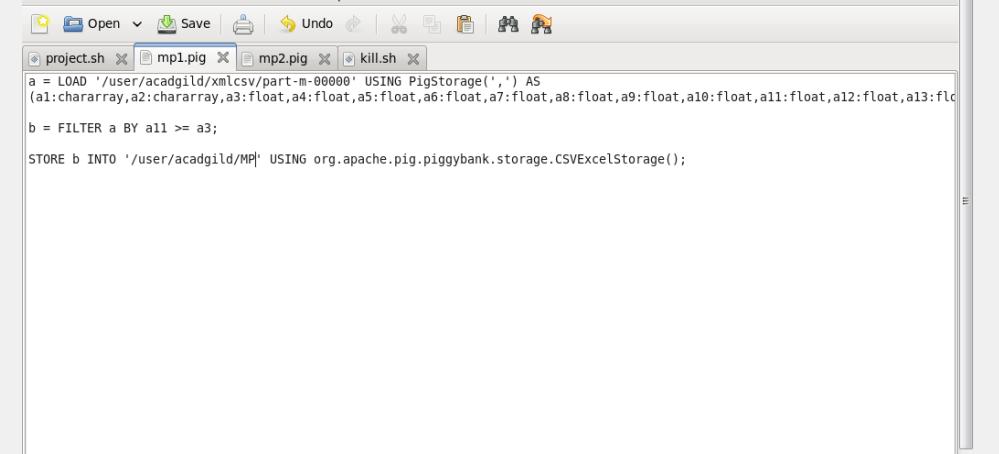
B = FOREACH A generate REPLACE(x,'[\n]','') as x;

C = FOREACH B generate REGEX_EXTRACT_ALL(x,'.*(?:<State Name>)({[^<]*}.*(?:<District Name>)({[^<]*}.*(?:<Project Objectives_IHHL_BPL>)({[^<]*}.*(?:<Project Objectives_IHHL_AP>)({[^<]*}.*(?:<Project Objectives_IHHL_TOTAL>)({[^<]*}.*(?:<Project Objectives_SCW>)({[^<]*}.*(?:<Project Objectives_School Toilets>)({[^<]*}.*(?:<Project Objectives_RSM>)({[^<]*}.*(?:<Project Objectives PC>)({[^<]*}.*(?:<Project Performance-IHHL_BPL>)({[^<]*}.*(?:<Project Performance-IHHL_AP>)({[^<]*}.*(?:<Project Performance-IHHL_TOTAL>)({[^<]*}.*(?:<Project Performance-SCW>)({[^<]*}.*(?:<Project Performance-School Toilets>)({[^<]*}.*(?:<Project Performance-Anganwadi Toilets>)({[^<]*}.*(?:<Project Performance-RSM>)({[^<]*}.*(?:<Project Performance-PC>)({[^<]*}.*');

D = FOREACH C generate FLATTEN (($0));

STORE D INTO '/user/acadgild/xmlcsv' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
```

1st question pig script



Acadgild_32_Bit [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System Fri May 26, 2:43 PM Acadgild

mp1.pig (~/Desktop) - gedit

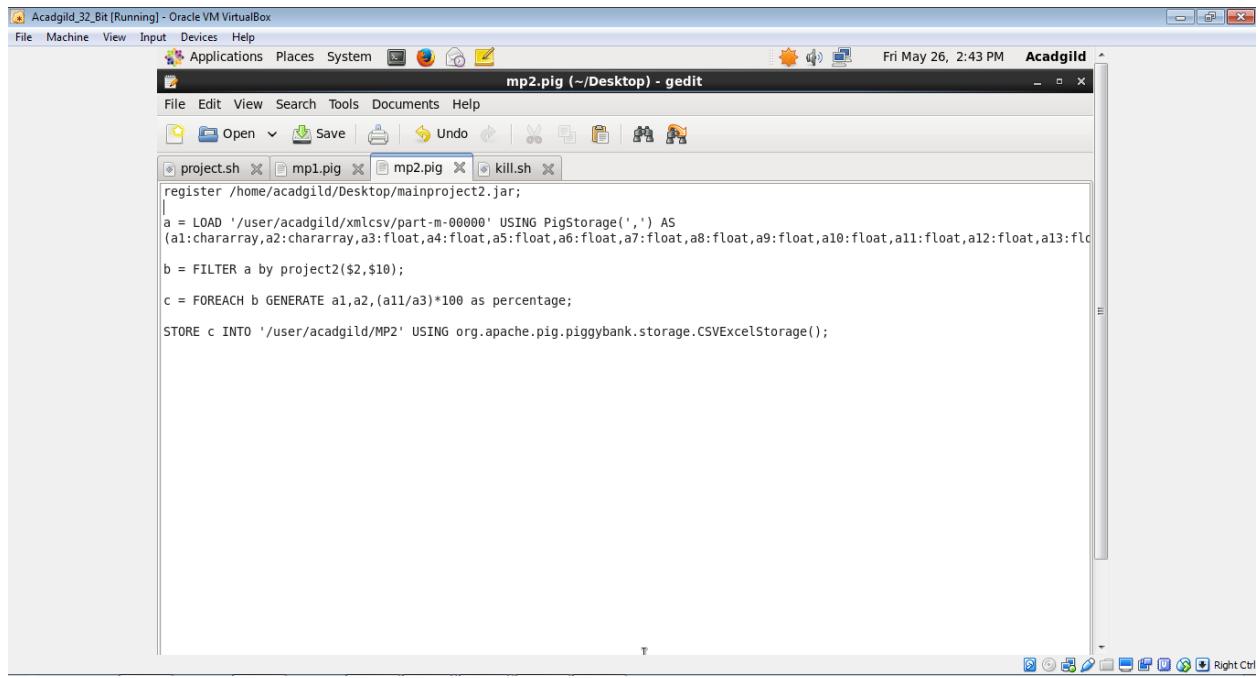
File Edit View Search Tools Documents Help

Open Save Undo

project.sh mp1.pig mp2.pig kill.sh

```
a = LOAD '/user/acadgild/xmlcsv/part-m-00000' USING PigStorage(',') AS  
(a1:chararray,a2:chararray,a3:float,a4:float,a5:float,a6:float,a7:float,a8:float,a9:float,a10:float,a11:float,a12:float,a13:float);  
  
b = FILTER a BY a11 >= a3;  
  
STORE b INTO '/user/acadgild/MP' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
```

2nd question pig script



The screenshot shows a Gedit window titled "mp2.pig (~/Desktop) - gedit". The window contains a Pig Latin script:

```
register /home/acadgild/Desktop/mainproject2.jar;
a = LOAD '/user/acadgild/xmlcsv/part-m-00000' USING PigStorage(',') AS
(a1:chararray,a2:chararray,a3:float,a4:float,a5:float,a6:float,a7:float,a8:float,a9:float,a10:float,a11:float,a12:float,a13:float);
b = FILTER a by project2($2,$10);
c = FOREACH b GENERATE a1,a2,(a1/a3)*100 as percentage;
STORE c INTO '/user/acadgild/MP2' USING org.apache.pig.piggybank.storage.CSVExcelStorage();
```