

MAIN PROJECT

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 <path to filecopy.conf>
```

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.

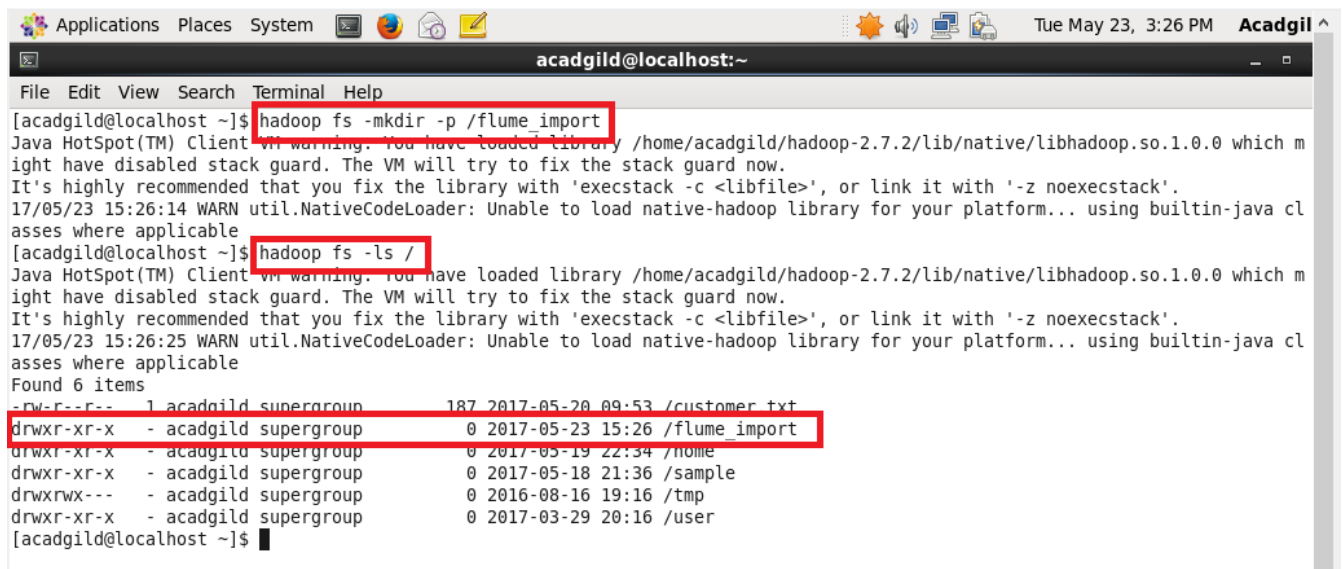
Export the results to MySQL using Sqoop.

DATA INGESTION INTO HDFS USING FLUME:

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.

Flume is used to ingest data into the hdfs at very high speed.

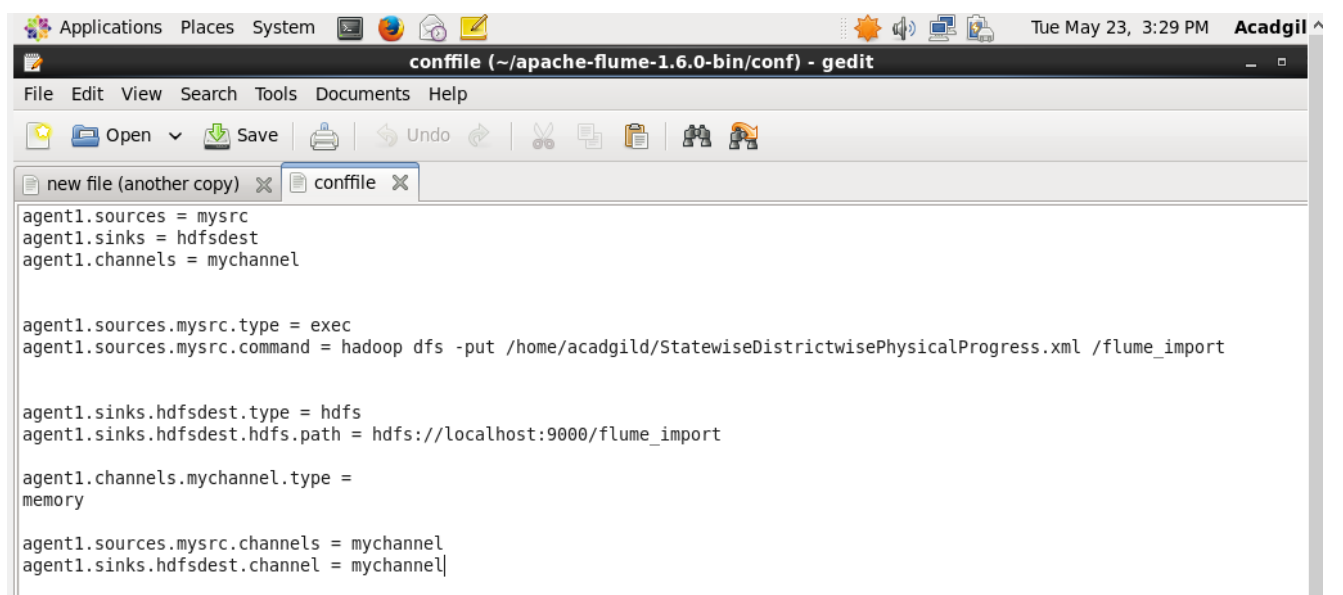
STEP 1: CREATING A DIRECTORY TO STORE THE DATASET IN HDFS



The screenshot shows a terminal window with the following commands and output:

```
[acadgild@localhost ~]$ hadoop fs -mkdir -p /flume_import
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'.
17/05/23 15:26:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop fs -ls /
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'.
17/05/23 15:26:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 6 items
-rw-r--r-- 1 acadgild supergroup 187 2017-05-20 09:53 /customer.txt
drwxr-xr-x - acadgild supergroup 0 2017-05-23 15:26 /flume_import
drwxr-xr-x - acadgild supergroup 0 2017-05-19 22:34 /home
drwxr-xr-x - acadgild supergroup 0 2017-05-18 21:36 /sample
drwxrwx--- - acadgild supergroup 0 2016-08-16 19:16 /tmp
drwxr-xr-x - acadgild supergroup 0 2017-03-29 20:16 /user
```

STEP 2: COPYING AND PASTING THE CONFIGURATION FILE INSIDE THE LIB FOLDER OF FLUME DIRECTORY.



The screenshot shows a Gedit editor window with the following configuration:

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

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

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

agent1.channels.mychannel.type =
memory

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

STEP 3: USING THE FLUME COMMAND TO INGEST THE DATASET(IN XML FORMAT) INTO THE HDFS.

A screenshot of a Linux terminal window titled "Applications Places System". The top bar shows system icons and the date/time "Tue May 23, 3:54 PM". The terminal prompt is "acadgild@localhost:~". The user has entered the command "PhysicalProgress>[acadgild@localhost ~]\$ flume-ng agent -n agent1 -c conf -f /home/acadgild/apache-flume-1.6.0-bin/conf/conffile". This command is highlighted with a red rectangular box. Below the command, the terminal displays several status messages: "fo: Including Hadoop libraries found via (/home/acadgild/hadoop-2.7.2/bin/hadoop) for HDFS access", "fo: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath", "fo: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath", "fo: Including HBASE libraries found via (/home/acadgild/hbase-1.0.3/bin/hbase) for HBASE access", "fo: Excluding /home/acadgild/hbase-1.0.3/lib/slf4j-api-1.7.7.jar from classpath", "fo: Excluding /home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar from classpath", "fo: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath", "fo: Excluding /home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath", "fo: Including Hive libraries found via (/home/acadgild/apache-hive-2.1.0-bin) for Hive access", and finally "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/activation-1.1.jar:/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/apacheds-kerberos-codec-2.0.0-M15.jar:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/asn1-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-1.7.0.jar:/home/acadgild/hadoop-2.7.2/share/hadoop-common/lib/commons..."

STEP 4: THE COMMAND HAS STARTED THE SINK, SOURCE AND THE CHANNEL

```

Acadgild_32_Bit [Running] - Oracle VM VirtualBox
t Devices Help
ion -n agent1 -f /home/acadgild/apache-flume-1.6.0-bin/conf/conffile
F4J: Class path contains multiple SLF4J bindings.
F4J: Found binding in [jar:file:/home/acadgild/apache-flume-1.6.0-bin/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLog
rBinder.class]
F4J: Found binding in [jar:file:/home/acadgild/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticL
gerBinder.class]
F4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
/05/23 15:53:36 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting
/05/23 15:53:36 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/acadgild/apache-fl
e-1.6.0-bin/conf/conffile
/05/23 15:53:36 INFO conf.FlumeConfiguration: Processing:hdfsdest
/05/23 15:53:36 INFO conf.FlumeConfiguration: Processing:hdfsdest
/05/23 15:53:36 INFO conf.FlumeConfiguration: Processing:hdfsdest
/05/23 15:53:36 INFO conf.FlumeConfiguration: Added sinks: hdfsdest Agent: agent1
/05/23 15:53:36 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agent

/05/23 15:53:36 INFO node.AbstractConfigurationProvider: Creating channels
/05/23 15:53:36 INFO channel.DefaultChannelFactory: Creating instance of channel mychannel type memory
/05/23 15:53:36 INFO node.AbstractConfigurationProvider: Created channel mychannel
/05/23 15:53:36 INFO source.DefaultSourceFactory: Creating instance of source mysrc, type exec
/05/23 15:53:36 INFO sink.DefaultSinkFactory: Creating instance of sink: hdfsdest, type: hdfs
/05/23 15:53:36 INFO node.AbstractConfigurationProvider: Channel mychannel connected to [mysrc, hdfsdest]
/05/23 15:53:36 INFO node.Application: Starting new configuration:{ sourceRunners:{mysrc=EventDrivenSourceRunner: { source:
g.apache.flume.source.ExecSource{name:mysrc,state:IDLE} }} sinkRunners:{hdfsdest=SinkRunner: { policy:org.apache.flume.sink
efaultSinkProcessor@165537a counterGroup: { name:null counters:{ } }} } channels:{mychannel=org.apache.flume.channel.MemoryCh
nel{name: mychannel}} }
/05/23 15:53:36 INFO node.Application: Starting Channel mychannel
/05/23 15:53:37 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel: Suc
ssfully registered new MBean.
/05/23 15:53:37 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
/05/23 15:53:37 INFO node.Application: Starting Sink hdfsdest
/05/23 15:53:37 INFO node.Application: Starting Source mysrc
/05/23 15:53:37 INFO source.ExecSource: Exec source starting with command:hadoop dfs -put /home/acadgild/StatewiseDistrictw
ePhysicalProgress.xml /flume import
/05/23 15:53:37 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfsdest: Success
lly registered new MBean.

acadgild@localhost:~ [Student Dashboard - ...

```

STEP 5: AFTER RUNNING THE FLUME COMMAND, THE DATA GETS STORED IN THE DESTINED PATH INSIDE THE HDFS

```
Applications Places System acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fs -ls /flume_import/
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'.
17/05/23 15:55:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 acadgild supergroup 717414 2017-05-23 15:39 /flume_import/StatewiseDistrictwisePhysicalProgress.xml
[acadgild@localhost ~]$
```

STEP 6: READING THE DATASET USING CAT COMMAND JUST TO CHECK WHETHER THE DATA HAS BEEN PROPERLY INGESTED.

```
acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fs -ls /flume_import/
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'.
17/05/23 15:55:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 acadgild supergroup 717414 2017-05-23 15:39 /flume_import/StatewiseDistrictwisePhysicalProgress.xml
[acadgild@localhost ~]$ hadoop fs -cat /flume_import/StatewiseDistrictwisePhysicalProgress.xml
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'.
17/05/23 15:56:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

STEP 7: THIS ENSURES THAT THE DATA HAS BEEN PROPERLY INGESTED INTO THE HDFS.

```
<Project_Performance-School_Toilets>125</Project_Performance-School_Toilets>
<Project_Performance-Anganwadi_Toilets>1631</Project_Performance-Anganwadi_Toilets>
<Project_Performance-RSM>29</Project_Performance-RSM>
<Project_Performance-PC>29</Project_Performance-PC>
</row>
<row>
  <State_Name>West Bengal</State_Name>
  <District_Name>UTTAR DINAJPUR</District_Name>
  <Project_Objectives_IHHL_BPL>257662</Project_Objectives_IHHL_BPL>
  <Project_Objectives_IHHL_APL>301645</Project_Objectives_IHHL_APL>
  <Project_Objectives_IHHL_TOTAL>559307</Project_Objectives_IHHL_TOTAL>
  <Project_Objectives_SCW>50</Project_Objectives_SCW>
  <Project_Objectives_School_Toilets>4806</Project_Objectives_School_Toilets>
  <Project_Objectives_Anganwadi_Toilets>1556</Project_Objectives_Anganwadi_Toilets>
  <Project_Objectives_RSM>30</Project_Objectives_RSM>
  <Project_Objectives_PC>0</Project_Objectives_PC>
  <Project_Performance-IHHL_BPL>148802</Project_Performance-IHHL_BPL>
  <Project_Performance-IHHL_APL>180619</Project_Performance-IHHL_APL>
  <Project_Performance-IHHL_TOTAL>329421</Project_Performance-IHHL_TOTAL>
  <Project_Performance-SCW>30</Project_Performance-SCW>
  <Project_Performance-School_Toilets>2562</Project_Performance-School_Toilets>
  <Project_Performance-Anganwadi_Toilets>2041</Project_Performance-Anganwadi_Toilets>
  <Project_Performance-RSM>17</Project_Performance-RSM>
  <Project_Performance-PC>0</Project_Performance-PC>
</row>
PhysicalProgress>[acadgild@localhost ~]$
```

PARSING THE DATA FROM XML FORMAT USING PIG:

I am running the pig with map reduce mode.

STEP 1: I AM USING THE XMLLOADER() IN PIGGY BANK UDF TO LOAD THE XML
HENCE REGISTERING THE PIGGYBANK UDF WITH GRUNT SHELL.

```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
grunt> REGISTER /home/acadgild/pig-0.16.0/lib/piggybank.jar
2017-05-24 15:45:30,560 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

STEP 2: LOADING THE DATA IN XML FORMAT INTO THE RELATION A

```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
grunt> A = LOAD 'hdfs://flume_import/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
2017-05-24 15:46:35,405 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

STEP 3: TO ELIMINATE THE TAGS, I AM USING REGEX_EXTRACT_ALL COMMAND.
THIS COMMAND FETCHES ONLY THE RECORDS PRESENT IN BETWEEN THE TAGS.
I'VE USED FLATTEN COMMAND TO UNNEST THE BAGS.

```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
grunt> B = foreach A GENERATE FLATTEN(REGEX_EXTRACT_ALL(x, '<row>\s*<State_Name>(.*?)</State_Name>\s*<District_Name>(.*?)</District_Name>\s*<Project_Objectives_IHHL_BPL>(.*?)</Project_Objectives_IHHL_BPL>\s*<Project_Objectives_IHHL_APL>(.*?)</Project_Objectives_IHHL_APL>\s*<Project_Objectives_IHHL_TOTAL>(.*?)</Project_Objectives_IHHL_TOTAL>\s*<Project_Objectives_SCW>(.*?)</Project_Objectives_SCW>\s*<Project_Objectives_School_Toilets>(.*?)</Project_Objectives_School_Toilets>\s*<Project_Objectives_Anganwadi_Toilets>(.*?)</Project_Objectives_Anganwadi_Toilets>\s*<Project_Objectives_RSM>(.*?)</Project_Objectives_RSM>\s*<Project_Objectives_PC>(.*?)</Project_Objectives_PC>\s*<Project_Performance_IHHL_BPL>(.*?)</Project_Performance_IHHL_BPL>\s*<Project_Performance_IHHL_APL>(.*?)</Project_Performance_IHHL_APL>\s*<Project_Performance_IHHL_TOTAL>(.*?)</Project_Performance_IHHL_TOTAL>\s*<Project_Performance_SCW>(.*?)</Project_Performance_SCW>\s*<Project_Performance_School_Toilets>(.*?)</Project_Performance_School_Toilets>\s*<Project_Performance_Anganwadi_Toilets>(.*?)</Project_Performance_Anganwadi_Toilets>\s*<Project_Performance_RSM>(.*?)</Project_Performance_RSM>\s*<Project_Performance_PC>(.*?)</Project_Performance_PC>\s*</row>');
grunt>
```

STEP 4: STORING THE FLATTENED FILE INTO HDFS USING STORE COMMAND.

```
acacgild@localhost:~/Desktop
File Edit View Search Terminal Help
grunt> STORE B into 'hdfs://user/acadgild/MAINPROJDATA';
2017-05-24 15:48:47,062 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-24 15:48:47,108 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2017-05-24 15:48:47,146 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2017-05-24 15:48:47,245 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-24 15:48:47,251 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-05-24 15:48:47,318 [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]}
```



```

acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/MAINPROJDATA/
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'.
17/05/24 15:56:58 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 2017-05-24 15:49 /user/acadgild/MAINPROJDATA/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 55135 2017-05-24 15:49 /user/acadgild/MAINPROJDATA/part-m-00000
[acadgild@localhost ~]$

```

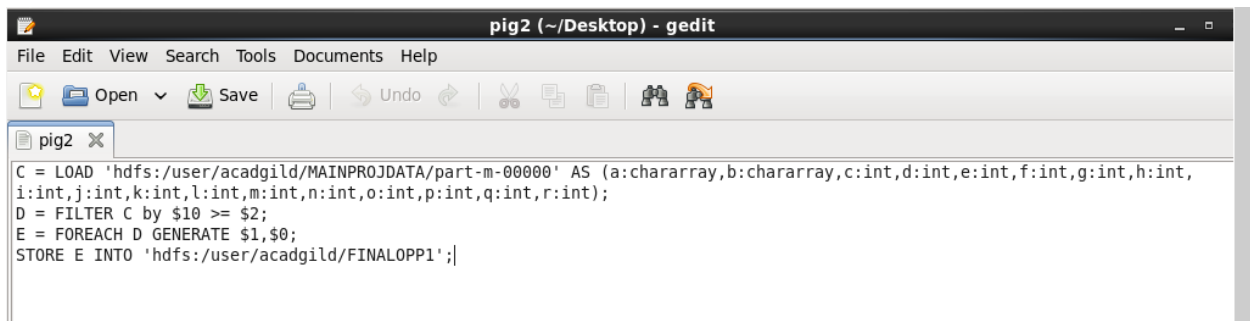
```
acagdild@localhost:~  
File Edit View Search Terminal Help  
[acagdild@localhost ~]$ hadoop fs -cat /user/acagdild/MAINPROJDATA/part-m-00000  
Java HotSpot(TM) Client VM warning: You have loaded library /home/acagdild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. 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'.  
17/05/24 15:57:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

acadgild@localhost:~																	
File	Edit	View	Search	Terminal	Help												
West Bengal 340 26			BANKURA 0	198152	333832	531984	50	7544	4130	29	0	105545	243191	348736	46	76871	
West Bengal 890 7724			BARDHAMAN 10 19		700047	341920	1041967	133	9891	7980	10	0	601906	277914	879820	140 9	
West Bengal 233 19			BIRBHUM 0	338989	299893	638882	50	5617	3816	4	22	266347	186599	452946	58	55632	
West Bengal 764 1818			COOCH BEHAR 15 12		335236	254422	589658	50	3715	1718	15	0	262294	164038	426332	144 5	
West Bengal 1 19			DAKSHIN DINAJPUR 2632 939 8			182621	194577	377198	50	3712	2642	10	0	184153	49448	23360	
West Bengal 435 574			DARJEELING 8 8		66648	130066	196714	50	1784	408	0	0	32921	3035	35956	18 1	
West Bengal 435 18			HOOGLHY 18	271737	195510	467247	53	6821	4168	19	0	269779	191294	461073	49	67643	
West Bengal 733 14			HOWRAH 20	231860	143309	375169	51	5195	3586	26	0	230190	141912	372102	42	51782	
West Bengal 578 4064			JALPAIGURI 17 14		372999	203523	576522	50	6578	5428	87	0	337740	101550	439290	25 6	
West Bengal 27 15			MALDA 15	452324	270208	722532	50	6385	7956	6	0	321934	65298	387232	41	59343	
West Bengal 0149 2882			MIDNAPUR EAST 8 17		392371	32617	424988	172	9726	5969	25	0	527389	32642	560031	210 1	
West Bengal 3452 2787			MIDNAPUR WEST 0 0		509496	432096	941592	50	16498	5825	10	0	596291	322659	918950	73 1	
West Bengal 838 2423			MURSHIDABAD 26 26		702442	506963	1209405	50	10260	7012	18	0	498998	198174	697172	47 7	
West Bengal 961 17			NADIA 41	346696	278335	625031	50	6974	6620	50	0	321462	198890	520352	28	66353	
West Bengal 4 66			NORTH 24 PARAGANAS 10931 3150 101 0			361462	225080	586542	51	11158	4466	30	0	357960	226104	58406	
West Bengal 128 20			PURULIA 0	210168	306933	517101	50	7542	4047	10	0	97160	79169	176329	10	46921	
West Bengal 20 006			SILIGURI 5 7		59536	25377	84913	30	935	1393	0	10	37794	18060	55854	30 9	

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

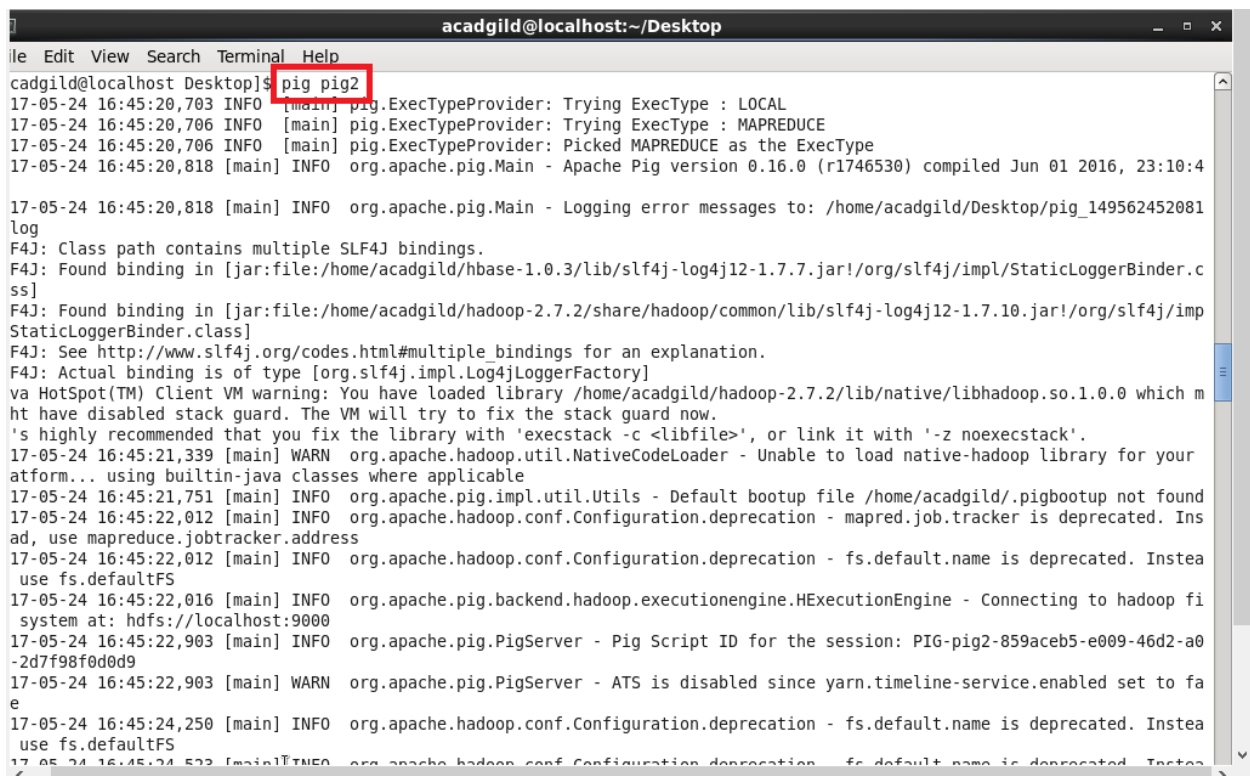
STEP 1:

- Loading the dataset into relation C by specifying datatype with the columns.
- Filtering the dataset by specifying condition (performance(\$10) >= objective (\$2))
- As we want only district, we are generating only districts and states by specifying the column \$1 and \$0.
- Storing the output in hdfs by specifying the path



```
pig2 (~/Desktop) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
pig2 x
C = LOAD 'hdfs:/user/acadgild/MAINPROJDATA/part-m-00000' AS (a:chararray,b:chararray,c:int,d:int,e:int,f:int,g:int,h:int,i:int,j:int,k:int,l:int,m:int,n:int,o:int,p:int,q:int,r:int);
D = FILTER C by $10 >= $2;
E = FOREACH D GENERATE $1,$0;
STORE E INTO 'hdfs:/user/acadgild/FINALOPP1';
```

STEP 2: RUNNING THE PIG SCRIPT IN MAPREDUCE MODE



```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
acadgild@localhost Desktop]$ pig pig2
17-05-24 16:45:20,703 INFO [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
17-05-24 16:45:20,706 INFO [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17-05-24 16:45:20,706 INFO [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
17-05-24 16:45:20,818 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
17-05-24 16:45:20,818 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/Desktop/pig_149562452081
log
F4J: Class path contains multiple SLF4J bindings.
F4J: Found binding in [jar:file:/home/acadgild/hbase-1.0.3/lib/slf4j-log4j12-1.7.7.jar!/org/slf4j/impl/StaticLoggerBinder.c
ss]
F4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/imp
StaticLoggerBinder.class]
F4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
F4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
va HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
ht have disabled stack guard. The VM will try to fix the stack guard now.
's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17-05-24 16:45:21,339 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your
atform... using builtin-java classes where applicable
17-05-24 16:45:21,751 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
17-05-24 16:45:22,012 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
ad, use mapreduce.job.tracker.address
17-05-24 16:45:22,012 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
use fs.defaultFS
17-05-24 16:45:22,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
system at: hdfs://localhost:9000
17-05-24 16:45:22,903 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-pig2-859aceb5-e009-46d2-a0
-2d7f98f0d0d9
17-05-24 16:45:22,903 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
e
17-05-24 16:45:24,250 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
use fs.defaultFS
17-05-24 16:45:24,532 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
```

STEP 3: READING THE OUTPUT USING CAT COMMAND. THIS IS JUST TO CHECK THE OUTPUT. EXPORTING STEP IS THE NEXT STEP.

```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
[acadgild@localhost Desktop]$ hadoop fs -cat /user/acadgild/FINALOPP1/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 16:10:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

STEP 4: SAMPLE OUTPUT

```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
BARABANKI      Uttar Pradesh
BAREILLY      Uttar Pradesh
BIJNOR      Uttar Pradesh
BUDAUN      Uttar Pradesh
BULANDSHAHR    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
MAHOBA      Uttar Pradesh
MATHURA      Uttar Pradesh
MIRZAPUR      Uttar Pradesh
```

STEP 5: USING **DB** DATABASE AND CREATING A TABLE '**DISTRICLIST**' IN **MYSQL**. USING SELECT COMMAND TO SHOW THAT THE TABLE IS EMPTY

```
acadgild@localhost:~
File Edit View Search Terminal Help
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| db |
| metastore |
| mysql |
+-----+
4 rows in set (0.00 sec)

mysql> use db;
Database changed
mysql> create table districlist (District varchar(50),State varchar(20));
Query OK, 0 rows affected (0.00 sec)

mysql> select * from districlist;
Empty set (0.00 sec)

mysql>
```

STEP 6: RUNNING SQOOP COMMAND BY SPECIFYING THE HDFS PATH , DATABASE NAME AND THE TABLE NAME.

```
2017-05-25 16:00:38,298 INFO [main] mapreduce.ExportJobBase: Exported 176 records.
[acadgild@localhost ~]$
[acadgild@localhost ~]$ sqoop export -m 1 -connect jdbc:mysql://localhost/db -username root -password acadgild -table district --export-dir /user/acadgild/FINALOPP1/part-m-00000 --input-fields-terminated-by '\t' --input-lines-terminated-by '\t' -m 1
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 16:00:38,261 INFO [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
2017-05-25 16:00:38,298 WARN [main] tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
```

Map reduce framework is running

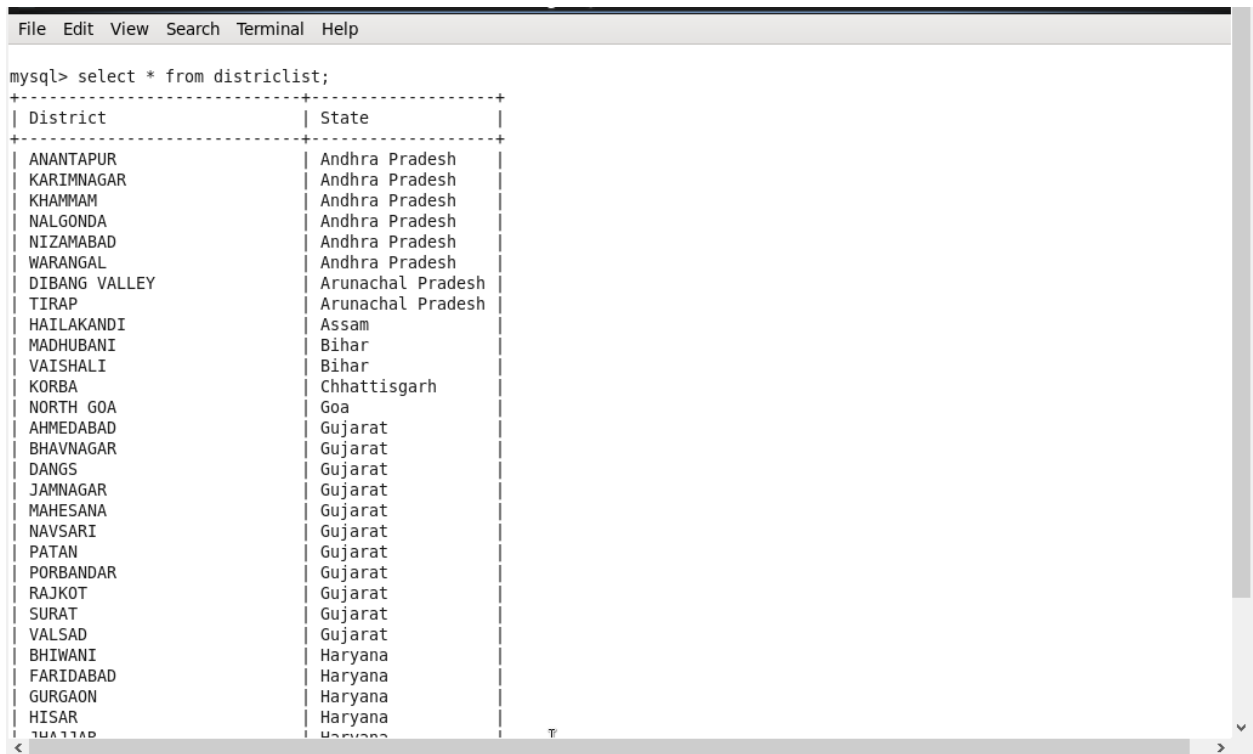
```
2017-05-25 16:00:45,423 INFO [main] Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.map.tasks
2017-05-25 16:00:45,620 INFO [main] client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 16:00:49,299 INFO [main] input.FileInputFormat: Total input paths to process : 1
2017-05-25 16:00:49,303 INFO [main] input.FileInputFormat: Total input paths to process : 1
2017-05-25 16:00:49,487 INFO [main] mapreduce.JobSubmitter: number of splits:1
2017-05-25 16:00:49,532 INFO [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
2017-05-25 16:00:49,803 INFO [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1495705447939_0007
2017-05-25 16:00:50,595 INFO [main] impl.YarnClientImpl: Submitted application application_1495705447939_0007
2017-05-25 16:00:50,866 INFO [main] mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1495705447939_0007/
2017-05-25 16:00:50,869 INFO [main] mapreduce.Job: Running job: job_1495705447939_0007
2017-05-25 16:01:07,885 INFO [main] mapreduce.Job: Job job_1495705447939_0007 running in uber mode : false
2017-05-25 16:01:07,889 INFO [main] mapreduce.Job: map 0% reduce 0%
2017-05-25 16:01:17,433 INFO [main] mapreduce.Job: map 100% reduce 0%
2017-05-25 16:01:18,513 INFO [main] mapreduce.Job: Job job_1495705447939_0007 completed successfully
2017-05-25 16:01:18,847 INFO [main] mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=136347
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=3939
```

Successfully exported the output from hdfs to mysql

```
Total megabyte-seconds taken by all map tasks=7599104
Map-Reduce Framework
  Map input records=176
  Map output records=176
  Input split bytes=139
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=141
  CPU time spent (ms)=980
  Physical memory (bytes) snapshot=67887104
  Virtual memory (bytes) snapshot=323592192
  Total committed heap usage (bytes)=16318464
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
2017-05-25 16:01:18,879 INFO [main] mapreduce.ExportJobBase: Transferred 3.8467 KB in 33.4046 seconds (117.918 bytes/sec)
2017-05-25 16:01:18,894 INFO [main] mapreduce.ExportJobBase: Exported 176 records.
[acadgild@localhost ~]$
```

acadgild@localhost: ~/... acadgild@localhost: ~ acadgild@localhost: ~ [pig2 (~/.Desktop) - ge...

SAMPLE OUTPUT:



```
mysql> select * from districlist;
```

District	State
ANANTAPUR	Andhra Pradesh
KARIMNAGAR	Andhra Pradesh
KHAMMAM	Andhra Pradesh
NALGONDA	Andhra Pradesh
NIZAMABAD	Andhra Pradesh
WARANGAL	Andhra Pradesh
DIBANG VALLEY	Arunachal Pradesh
TIRAP	Arunachal Pradesh
HAILAKANDI	Assam
MADHUBANI	Bihar
VAISHALI	Bihar
KORBA	Chhattisgarh
NORTH GOA	Goa
AHMEDABAD	Gujarat
BHAVNAGAR	Gujarat
DANGS	Gujarat
JAMNAGAR	Gujarat
MAHESANA	Gujarat
NAVSARI	Gujarat
PATAN	Gujarat
PORBANDAR	Gujarat
RAJKOT	Gujarat
SURAT	Gujarat
VALSAD	Gujarat
BHIWANI	Haryana
FARIDABAD	Haryana
GURGAON	Haryana
HISAR	Haryana
MEERUT	Haryana

I've attached the output with this document. Please check that.

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

STEP 1: WRITING USER DEFINED FUNCTION TO FILTER THE DATASET.

FILTERING THE DATASET BY SPECIFYING CONDITION (PERFORMANCE ≥ 0.8 * OBJECTIVE)

```
*check.java
1 import java.io.IOException;
2
3
4
5
6 public class check extends FilterFunc {
7
8     @Override
9     public Boolean exec(Tuple input) throws IOException {
10         // TODO Auto-generated method stub
11
12         if(input == null || input.size() == 0)
13         {
14             return null;
15         }
16
17         String val1 = input.get(0).toString();
18         String val2 = input.get(1).toString();
19
20         int a = Integer.parseInt(val1);
21         int b = Integer.parseInt(val2);
22
23         if(a>=(b*0.8))
24         {
25             return true;
26         }
27         return false;
28     }
29 }
```

STEP 2:

- Loading the dataset into relation C by specifying datatype with the columns.
- check function filters the dataset as we have already specified condition in the udf
- As we want only district and state, we are generating only districts by specifying the columns \$1 and \$0.
- Storing the output in hdfs by specifying the path.

```
pig3 (~/Desktop) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
pig3
REGISTER '/home/acadgild/Desktop/pig.jar';
C = LOAD 'hdfs://user/acadgild/MAINPROJDATA/part-m-000000' AS (a:chararray,b:chararray,c:int,d:int,e:int,f:int,g:int,h:int,i:int,j:int,k:int,l:int,m:int,n:int,o:int,p:int,q:int,r:int);
D = FILTER C BY check($10,$2);
E = FOREACH D GENERATE $1,$0;
STORE E INTO 'hdfs://user/acadgild/FINALOPP2';
```

STEP 3: RUNNING THE PIG SCRIPT IN MAPREDUCE MODE

```
[acadgild@localhost Desktop]$ pig pig3
2017-05-24 17:40:03,133 INFO [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-05-24 17:40:03,136 INFO [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-05-24 17:40:03,136 INFO [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-05-24 17:40:03,230 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2017-05-24 17:40:03,230 [main] INFO org.apache.pig.Main - Logging error messages to: /home/acadgild/Desktop/pig_1495627803227.log
SLF4J: Class path contains multiple SLF4J bindings.
```

STEP 4: READING THE CONTENTS OF THE OUTPUT JUST TO CHECK WHETHER THE CODE IS CORRECT

```
acadgild@localhost:~/Desktop
File Edit View Search Terminal Help
[acadgild@localhost Desktop]$ hadoop fs -cat /user/acadgild/FINALOPP2/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 16:15:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ANANTAPUR      Andhra Pradesh
CHITTOOR       Andhra Pradesh
CUDDAPAH       Andhra Pradesh
EAST GODAVARI  Andhra Pradesh
KARIMNAGAR     Andhra Pradesh
KHAMMAM        Andhra Pradesh
KRISHNA        Andhra Pradesh
KURNOOL        Andhra Pradesh
MEDAK          Andhra Pradesh
NALGONDA       Andhra Pradesh
NIZAMABAD      Andhra Pradesh
RANGAREDDI     Andhra Pradesh
WARANGAL       Andhra Pradesh
WEST GODAVARI  Andhra Pradesh
DIBANG VALLEY  Arunachal Pradesh
LOHIT          Arunachal Pradesh
TIRAP          Arunachal Pradesh
BAGSHA         Assam
```

STEP 5: USING DB DATABASE AND CREATING A TABLE 'DISTRICLIST2' IN MYSQL. USING SELECT COMMAND TO SHOW THAT THE TABLE IS EMPTY

```
acadgild@localhost:~
File Edit View Search Terminal Help
mysql> show databases;
+-----+
| Database |
+-----+
| information_schema |
| db |
| metastore |
| mysql |
+-----+
4 rows in set (0.00 sec)

mysql> use db;
Database changed
mysql> create table districlist2 (District varchar(50),State varchar(20));
Query OK, 0 rows affected (0.00 sec)

mysql> select * from districlist2;
Empty set (0.00 sec)

mysql>
```

STEP 6: RUNNING SQOOP COMMAND BY SPECIFYING THE HDFS PATH , DATABASE NAME AND THE TABLE NAME.

```
[acadgild@localhost ~]$ sqoop export -m 1 -connect jdbc:mysql://localhost/db -username root -password acadgild -table district2 --export-dir /user/acadgild/FINALOPP2/part-m-00000 --input-fields-terminated-by '\t' --input-lines-terminated-by '\t' -m 1
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 16:20:22,773 INFO [main] sqoop.Sqoop: Running Sqoop version: 1.4.6
2017-05-25 16:20:22,811 WARN [main] tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2017-05-25 16:20:23,292 INFO [main] manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2017-05-25 16:20:23,292 INFO [main] tool.CodeGenTool: Beginning code generation
2017-05-25 16:20:23,943 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `district2` AS t LIMIT 1
2017-05-25 16:20:24,004 INFO [main] manager.SqlManager: Executing SQL statement: SELECT t.* FROM `district2` AS t LIMIT 1
2017-05-25 16:20:24,025 INFO [main] orm.CompilationManager: HADOOP MAPRED_HOME is /home/acadgild/hadoop-2.7.2
Note: /tmp/sqoop-acadgild/compile/81521afe9fd868ea6211c0ed7de78348/district2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2017-05-25 16:20:27,519 INFO [main] orm.CompilationManager: Writing jar file: /tmp/sqoop-acadgild/compile/81521afe9fd868ea6211c0ed7de78348/district2.jar
2017-05-25 16:20:27,556 INFO [main] mapreduce.ExportJobBase: Beginning export of district2
SLF4J: Class path contains multiple SLF4J bindings.
```

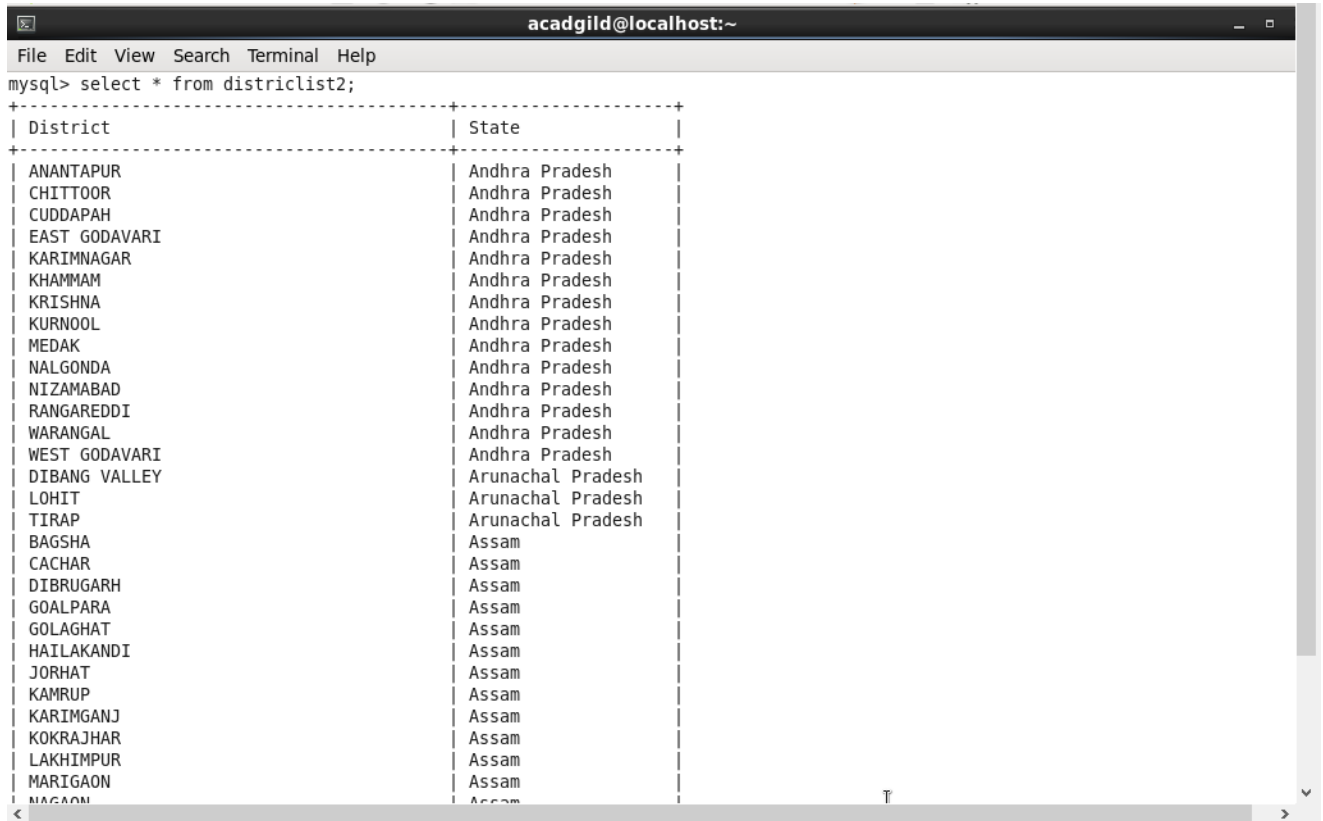
Map reduce framework is running

```
2017-05-25 16:20:30,128 INFO [main] client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2017-05-25 16:20:33,575 INFO [main] input.FileInputFormat: Total input paths to process : 1
2017-05-25 16:20:33,579 INFO [main] input.FileInputFormat: Total input paths to process : 1
2017-05-25 16:20:33,749 INFO [main] mapreduce.JobSubmitter: number of splits:1
2017-05-25 16:20:33,793 INFO [main] Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
2017-05-25 16:20:34,069 INFO [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1495705447939_0009
2017-05-25 16:20:34,801 INFO [main] impl.YarnClientImpl: Submitted application application_1495705447939_0009
2017-05-25 16:20:34,920 INFO [main] mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1495705447939_0009/
2017-05-25 16:20:34,922 INFO [main] mapreduce.Job: Running job: job_1495705447939_0009
2017-05-25 16:20:52,828 INFO [main] mapreduce.Job: Job job_1495705447939_0009 running in uber mode : false
2017-05-25 16:20:52,830 INFO [main] mapreduce.Job: map 0% reduce 0%
2017-05-25 16:21:04,104 INFO [main] mapreduce.Job: map 100% reduce 0%
2017-05-25 16:21:04,130 INFO [main] mapreduce.Job: Job job_1495705447939_0009 completed successfully
2017-05-25 16:21:04,451 INFO [main] mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=136350
```

Successfully exported the output from hdfs to mysql

```
Map-Reduce Framework
  Map input records=349
  Map output records=349
  Input split bytes=139
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=160
  CPU time spent (ms)=1330
  Physical memory (bytes) snapshot=68096000
  Virtual memory (bytes) snapshot=324005888
  Total committed heap usage (bytes)=16318464
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
2017-05-25 16:21:04,473 INFO [main] mapreduce.ExportJobBase: Transferred 7.2891 KB in 34.4904 seconds (216.4083 bytes/sec)
2017-05-25 16:21:04,491 INFO [main] mapreduce.ExportJobBase: Exported 349 records.
[acadgild@localhost ~]$
```


STEP 7: SAMPLE OUTPUT



```
mysql> select * from districlist2;
```

District	State
ANANTAPUR	Andhra Pradesh
CHITTOOR	Andhra Pradesh
CUDDAPAH	Andhra Pradesh
EAST GODAVARI	Andhra Pradesh
KARIMNAGAR	Andhra Pradesh
KHAMMAM	Andhra Pradesh
KRISHNA	Andhra Pradesh
KURNOOL	Andhra Pradesh
MEDAK	Andhra Pradesh
NALGONDA	Andhra Pradesh
NIZAMABAD	Andhra Pradesh
RANGAREDDI	Andhra Pradesh
WARANGAL	Andhra Pradesh
WEST GODAVARI	Andhra Pradesh
DIBANG VALLEY	Arunachal Pradesh
LOHIT	Arunachal Pradesh
TIRAP	Arunachal Pradesh
BAGSHA	Assam
CACHAR	Assam
DIBRUGARH	Assam
GOALPARA	Assam
GOLAGHAT	Assam
HAILAKANDI	Assam
JORHAT	Assam
KAMRUP	Assam
KARIMGANJ	Assam
KOKRAJHAR	Assam
LAKHIMPUR	Assam
MARIGAON	Assam
NAGAON	Assam

I've attached the output with this document. Please check that.