PROJECT-1.2

1.1 Project Overview-

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

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

- Developing system to handle the incoming log feed and stores the information in Hadoop Cluster (Flume).
- Analyze the data and understand the progress.
- Store the results in Hbase/RDBMS.

Step 1: Copy dataset from local file system to HDFS using flume.

In order to proceed we have to develop as system which can store the incoming log file in XML format to HDFS.
To do this we are making a directory /home/acadgild/Project-1.2 in local as shown below and currently storing the file there as shown below-

```
Quick connect...        3. 192.168.113.130 (acadgild)
[acadgild@localhost ~]$ cd Project-1.2
[acadgild@localhost Project-1.2]$ pwd
/home/acadgild/Project-1.2
[acadgild@localhost Project-1.2]$ ls -l
total 704
-rw-rw-r--. 1 acadgild acadgild 717414 Dec 12 23:28 StatewiseDistrictwisePhysicalProgress.xml
[acadgild@localhost Project-1.2]$
```

Now we have to configure the flume configuration to take the above file as input and store it in HDFS. Below is the configuration defined to do so-

```
[acadgild@localhost conf]$ ls -l
total 28
-rw-rw-r--. 1 acadgild acadgild    0 Dec 12 23:38 filecopy.conf
-rw-rw-r--. 1 acadgild acadgild 1457 Dec  5 15:33 flume.conf
-rw-r--r--. 1 acadgild acadgild 1661 May  9  2015 flume-conf.properties.template
```

filecopy.conf (~/apache-flume-1.6.0-bin/conf) - gedit@localhost.localdomain

File Edit View Search Tools Documents Help

New Open Save Print Undo Redo Cut Copy Paste Find Replace

filecopy.conf

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


agent1.sources.mysrc.type = exec
agent1.sources.mysrc.command = hadoop dfs -put /home/acadgild/Project-1.2/
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
```

Now we will run the flume agent to store the file in HDFS-

Flume-ng agent –n agent1 –f /home/acadgild/apache-flume-1.6.0-bin/conf/filecopy.conf

```
[acadgild@localhost ~]$ flume-ng agent -n agent1 -f /home/acadgild/apache-flume-1.6.0-bin/conf/filecopy.conf
Warning: No configuration directory set! Use --conf <dir> to override.
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
```

Below screenshot shows that the job ran to success-

```
{mychannel=org.apache.flume.channel.MemoryChannel{name: mychannel}} }
17/12/12 23:49:43 INFO node.Application: Starting Channel mychannel
17/12/12 23:49:43 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel: Successfully registered new MBean.
17/12/12 23:49:43 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
17/12/12 23:49:43 INFO node.Application: Starting Sink hdfsdest
17/12/12 23:49:43 INFO node.Application: Starting Source mysrc
17/12/12 23:49:43 INFO source.ExecSource: Exec source starting with command:hadoop dfs -put /home/acadgild/Project-1.2/StatewiseDistrictwisePhysicalProgress.xml /flu
me_import
17/12/12 23:49:43 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfsdest: Successfully registered new MBean.
17/12/12 23:49:43 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfsdest started
17/12/12 23:49:43 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: mysrc: Successfully registered new MBean.
17/12/12 23:49:43 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc started
17/12/12 23:49:48 INFO source.ExecSource: Command [hadoop dfs -put /home/acadgild/Project-1.2/StatewiseDistrictwisePhysicalProgress.xml /flume_import] exited with 0
^C17/12/12 23:56:44 INFO lifecycle.LifecycleSupervisor: Stopping lifecycle supervisor 11
17/12/12 23:56:44 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider stopping
17/12/12 23:56:44 INFO source.ExecSource: Stopping exec source with command:hadoop dfs -put /home/acadgild/Project-1.2/StatewiseDistrictwisePhysicalProgress.xml /flu
me_import
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc stopped
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. source.start.time == 1513102783748
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. source.stop.time == 1513103204923
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.append-batch.accepted == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.append-batch.received == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.append.accepted == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.append.received == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.events.accepted == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.events.received == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: mysrc. src.open-connection.count == 0
17/12/12 23:56:44 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel stopped
```

The same result can be seen in the HDFS location. The XML file is present at below location-

<u>hdfs://localhost:9000/flume_import</u>

```
[acadgild@localhost conf]$ hadoop fs -cat hdfs://localhost:9000/flume_import
Java HotSpot(TM) Server VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/l
ill 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
17/12/12 23:59:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform
<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>
```

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.

Problem statement

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

Solution-

Now we will proceed with starting pig in mapreduce mode as shown below-

```
[acadgild@localhost ~]$ pig -x mapreduce
2017-12-13 00:33:05,777 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOCAL
2017-12-13 00:33:05,780 INFO  [main] pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2017-12-13 00:33:05,802 INFO  [main] pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-12-13 00:33:05,934 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compi
2017-12-13 00:33:05,934 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/
2017-12-13 00:33:06,113 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acad
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/i
SLF4J: Found binding in [jar:file:/home/acadgild/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1
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) Server VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/li
ill 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 n
2017-12-13 00:33:07,374 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable to load native-
here applicable
2017-12-13 00:33:07,380 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
2017-12-13 00:33:07,380 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
2017-12-13 00:33:07,381 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine -
2017-12-13 00:33:08,993 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-de
2017-12-13 00:33:08,993 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-s
grunt>
```

We will also REGISTER piggybank.jar-

```
grunt> REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar'
2017-12-13 01:02:40,969 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

Below is the PIG script we will run to find out the districts who achieved 100 percent objective in BPL cards-

- A = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
- B = FOREACH A GENERATE FLATTEN(REGEX_EXTRACT_ALL(x,'<row>\\s*<State_Name>(.*)</State_Name>\\s*<District_Name>(.*)</District_Name>\\s*<Project_Objectives_IHHL_BPL>(.*)</Project_Objectives_IHHL_BPL>\\s*<Project_Objectives_IHHL_APL>(.*)</Project_Objectives_IHHL_APL>\\s*<Project_Objectives_IHHL_TOTAL>(.*)</Project_Objectives_IHHL_TOTAL>\\s*<Project_Objectives_SCW>(.*)</Project_Objectives_SCW>\\s*<Project_Objectives_School_Toilets>(.*)</Project_Objectives_School_Toilets>\\s*<Project_Objectives_Anganwadi_Toilets>(.*)</Project_Objectives_Anganwadi_Toilets>\\s*<Project_Objectives_RSM>(.*)</Project_Objectives_RSM>\\s*<Project_Objectives_PC>(.*)</Project_Objectives_PC>\\s*<Project_Performance-IHHL_BPL>(.*)</Project_Performance-IHHL_BPL>\\s*<Project_Performance-IHHL_APL>(.*)</Project_Performance-IHHL_APL>\\s*<Project_Performance-IHHL_TOTAL>(.*)</Project_Performance-IHHL_TOTAL>\\s*<Project_Performance-SCW>(.*)</Project_Performance-SCW>\\s*<Project_Performance-School_Toilets>(.*)</Project_Performance-School_Toilets>\\s*<Project_Performance-Anganwadi_Toilets>(.*)</Project_Performance-

Anganwadi_Toilets>\\s*<Project_Performance-RSM>(.*)</Project_Performance-RSM>\\s*<Project_Performance-PC>(.*)</Project_Performance-PC>\\s*</row>'));

- C = FOREACH B GENERATE (chararray)$0 as State, (chararray)$1 as District, (int)$2 as BPL_Objective, (int) $10 as BPL_Performance;
- D = FILTER C BY (BPL_Objective == BPL_Performance);
- E = FOREACH D GENERATE District;
- STORE E INTO /home/acadgild/Project-1.2/BPL100P.txt

Now we will try to understand each script in details-

- A = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
  Since the file is in XML format we are loading it using XMLloader by taking the tag <row> as parent field and processing it as chararray-

```
grunt> A = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
2017-12-13 01:22:39,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> lim1 = LIMIT A 2;
grunt> dump lim1;
2017-12-13 01:23:32,549 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-12-13 01:23:32,583 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-12-13 01:23:32,583 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-12-13 01:23:32,583 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculato
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEach
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-12-13 01:23:32,620 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-12-13 01:23:32,628 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-13 01:23:32,630 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 01:23:32,630 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-12-13 01:23:32,642 [main] INFO  org.apache.hadoop.mapreduce.lib.input.LineRecordReader - Found UTF-8 BOM and skipped it
2017-12-13 01:23:32,684 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to hdfs://localhost
:9000/tmp/temp-1676464451/tmp-663366910/_temporary/0/task__0001_m_000001
2017-12-13 01:23:32,713 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-13 01:23:32,718 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 01:23:32,718 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(<row>          <State_Name>Andhra Pradesh</State_Name>         <District_Name>ADILABAD</District_Name>         <Project_Objectives_IHHL_BPL>247475</Project_Objectiv
es_IHHL_BPL>            <Project_Objectives_IHHL_APL>148181</Project_Objectives_IHHL_APL>               <Project_Objectives_IHHL_TOTAL>395656</Project_Objectives_IHH
L_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_Obje
ctives_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_Performan
ce-SCW>         <Project_Performance-School_Toilets>4462</Project_Performance-School_Toilets>           <Project_Performance-Anganwadi_Toilets>427</Project_Performan
ce-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_Object
ives_IHHL_TOTAL>                <Project_Objectives_SCW>0</Project_Objectives_SCW>              <Project_Objectives_School_Toilets>3421</Project_Objectives_School_To
ilets>          <Project_Objectives_Anganwadi_Toilets>284</Project_Objectives_Anganwadi_Toilets>               <Project_Objectives_RSM>10</Project_Objectives_RSM> <
Project_Objectives_PC>0</Project_Objectives_PC>         <Project_Performance-IHHL_BPL>366557</Project_Performance-IHHL_BPL>             <Project_Performance-IHHL_APL
>42000</Project_Performance-IHHL_APL>           <Project_Performance-IHHL_TOTAL>408557</Project_Performance-IHHL_TOTAL>         <Project_Performance-SCW>0</Project_P
erformance-SCW>         <Project_Performance-School_Toilets>4258</Project_Performance-School_Toilets>           <Project_Performance-Anganwadi_Toilets>302</Project_P
erformance-Anganwadi_Toilets>           <Project_Performance-RSM>0</Project_Performance-RSM>            <Project_Performance-PC>0</Project_Performance-PC>      </row
>)
```

- 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>'));

Above command is parsing the XML fields of the input fields and extracting the data out of it

```
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_Objec
tives_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> lim2 = LIMIT B 2;
grunt> dump lim2;
```

If we dump above relation we can see below data extracted out of the XML file-

```
2017-12-13 01:28:18,808 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,814 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:18,887 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,890 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:18,935 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,939 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:18,982 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,988 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:19,047 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:19,055 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:19,103 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:19,110 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:19,189 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-13 01:28:19,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
2017-12-13 01:28:19,190 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate co
2017-12-13 01:28:19,195 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 01:28:19,195 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(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)
grunt>
```

- C = FOREACH B GENERATE (chararray)$0 as State, (chararray)$1 as District, (int)$2 as BPL_Objective, (int) $10 as BPL_Performance;

Now here we are extracting only required fields like State, district, BPL Objective and BPL Performance-

```
grunt> C = FOREACH B GENERATE (chararray)$0 as State, (chararray)$1 as District, (int)$2 as BPL_Objective, (int) $10 as BPL_Performance;
grunt> lim3 = LIMIT C 5;
grunt> dump lim3;
```

If we dump C we can see below sample result-

```
2017-12-13 02:04:44,180 [main] INFO  org.apache
(Andhra Pradesh,ADILABAD,247475,176300)
(Andhra Pradesh,CHITTOOR,296465,269750)
(Andhra Pradesh,CUDDAPAH,251653,239780)
(Andhra Pradesh,ANANTAPUR,363314,366557)
(Andhra Pradesh,EAST GODAVARI,370255,347305)
grunt>
```

- D = FILTER C BY (BPL_Objective == BPL_Performance);

Now we will run above relation which will filter C based on above condition to find out which district achieved 100 percent objective in BPL cards

```
2017-12-13 02:10:42,156 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-13 02:10:42,156 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaul
2017-12-13 02:10:42,158 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-12-13 02:10:42,164 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 02:10:42,164 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Goa,NORTH GOA,15000,15000)
(Assam,HAILAKANDI,49837,49837)
(Bihar,MADHUBANI,67482,67482)
(Andhra Pradesh,NIZAMABAD,225519,225519)
(Arunachal Pradesh,TIRAP,5780,5780)
```

- E = FOREACH D GENERATE District;

Since we are interested in finding only District name we will extract District name from D relation-

```
grunt> E = FOREACH D GENERATE District;
grunt> lim4 = LIMIT E 10;
grunt> dump lim4;
```

Below shows a sample dump of it-

```
2017-12-13 02:16:48,810 [main]
2017-12-13 02:16:48,811 [main]
2017-12-13 02:16:48,812 [main]
2017-12-13 02:16:48,818 [main]
(NORTH GOA)
(HAILAKANDI)
(MADHUBANI)
(DANGS)
(SURAT)
(NAVSARI)
(AHMEDABAD)
(PORBANDAR)
(NIZAMABAD)
(TIRAP)
```

- STORE E INTO /home/acadgild/Project-1.2/BPL100P.txt

Now we will store the name of districts in HDFS

```
Quick connect...    3. 192.168.113.130 (acadgild)
grunt> STORE E INTO '/home/acadgild/Project-1.2/BPL100P.txt';
```

```
2017-12-13 02:20:32,247 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 02:20:32,250 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
o job history server
2017-12-13 02:20:32,289 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 02:20:32,294 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
o job history server
2017-12-13 02:20:32,333 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 02:20:32,336 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAppli
o job history server
2017-12-13 02:20:32,398 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Below screenshot shows the file content-

```
grunt> cat hdfs://localhost:9000/home/acadgild/Project-1.2/BPL100P.txt
NIZAMABAD
TIRAP
HAILAKANDI
MADHUBANI
NORTH GOA
AHMEDABAD
DANGS
NAVSARI
PORBANDAR
SURAT
FARIDABAD
HISAR
JHAJJAR
MAHENDRAGARH
PANCHKULA
PANIPAT
ROHTAK
SIRSA
HAMIRPUR
KINNAUR
KULLU
LAHAUL &amp; SPITI
SHIMLA
SOLAN
UNA
DEOGHAR
LOHARDAGA
HASSAN
MANGALORE(DAKSHINA KANNADA)
UDUPI
ALAPPUZHA
KOLLAM
KOTTAYAM
KOZHIKODE
PALAKKAD
PATHANAMTHITTA
WAYANAD
```

Export the results to MySQL using sqoop-

Now for the second part of this requirement we will create a table in MySQL in database **db** as shown below-

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| db                 |
| metastore          |
| mysql              |
+--------------------+
4 rows in set (0.03 sec)

mysql> use db;
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>
```

The table name is BPL100P with one column District_name-

```
mysql> CREATE TABLE BPL100P
    -> (
    -> District_name varchar(50)
    -> );
Query OK, 0 rows affected (0.05 sec)

mysql>
```

Now we will run Sqoop to export the data present in HDFS to MySQL using below command-

- sqoop export --connect jdbc:mysql://localhost/db \
- --username 'acadgild' -P --table 'BPL100P' --export-dir '/hdfs://localhost:9000/home/acadgild/Project-1.2/BPL100P.txt' \
- --input-fields-terminated-by ',' \
- -m 1 --columns District_name

```
[acadgild@localhost ~]$ sqoop export --connect jdbc:mysql://localhost/db \
> --username 'acadgild' -P --table 'BPL100P' --export-dir '/hdfs://localhost:9000/home/acadgild/Project-1.2/BPL100P.txt' \
> --input-fields-terminated-by ',' \
> -m 1 --columns District_name
```

The result can be seen in the table BPL100P if we select records from it-

```
mysql> select * from BPL100P;
+-----------------------------+
| District_name               |
+-----------------------------+
| NIZAMABAD                   |
| TIRAP                       |
| HAILAKANDI                  |
| MADHUBANI                   |
| NORTH GOA                   |
| AHMEDABAD                   |
| DANGS                       |
| NAVSARI                     |
| PORBANDAR                   |
| SURAT                       |
| FARIDABAD                   |
| HISAR                       |
| JHAJJAR                     |
| MAHENDRAGARH                |
| PANCHKULA                   |
| PANIPAT                     |
| ROHTAK                      |
| SIRSA                       |
| HAMIRPUR                    |
| KINNAUR                     |
| KULLU                       |
| LAHAUL &amp; SPITI          |
| SHIMLA                      |
| SOLAN                       |
| UNA                         |
| DEOGHAR                     |
| LOHARDAGA                   |
| HASSAN                      |
| MANGALORE(DAKSHINA KANNADA) |
| UDUPI                       |
| ALAPPUZHA                   |
| KOLLAM                      |
| KOTTAYAM                    |
| KOZHIKODE                   |
```

```
| PALAKKAD                    |
| WAYANAD                     |
| GADCHIROLI                  |
| SINDHUDURG                  |
| WEST GARO HILLS             |
| CHAMPHAI                    |
| LAWNGTLAI                   |
| HANUMANGARH                 |
| ERODE                       |
| KARUR                       |
| NAMAKKAL                    |
| TIRUCHIRAPPALLI             |
| TIRUVANNAMALAI              |
| DHALAI                      |
| SOUTH TRIPURA               |
| WEST TRIPURA                |
| AMBEDKAR NAGAR              |
| BALRAMPUR                   |
| BAREILLY                    |
| BIJNOR                      |
| BUDAUN                      |
| ETAWAH                      |
| FARRUKHABAD                 |
| FIROZABAD                   |
| GHAZIABAD                   |
| HARDOI                      |
| JYOTIBA PHULE NAGAR         |
| LUCKNOW                     |
| MAHARAJGANJ                 |
| MAHOBA                      |
| MORADABAD                   |
| MUZAFFARNAGAR               |
| PILIBHIT                    |
| SONBHADRA                   |
| SULTANPUR                   |
+-----------------------------+
71 rows in set (0.00 sec)
```

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-

Below is the Pig Script used to find the result-

- A = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
- B = FOREACH A GENERATE FLATTEN(REGEX_EXTRACT_ALL(x,'<row>\\s*<State_Name>(.*)</State_Name>\\s*<District_Name>(.*)</District_Name>\\s*<Project_Objectives_IHHL_BPL>(.*)</Project_Objectives_IHHL_BPL>\\s*<Project_Objectives_IHHL_APL>(.*)</Project_Objectives_IHHL_APL>\\s*<Project_Objectives_IHHL_TOTAL>(.*)</Project_Objectives_IHHL_TOTAL>\\s*<Project_Objectives_SCW>(.*)</Project_Objectives_SCW>\\s*<Project_Objectives_School_Toilets>(.*)</Project_Objectives_School_Toilets>\\s*<Project_Objectives_Anganwadi_Toilets>(.*)</Project_Objectives_Anganwadi_Toilets>\\s*<Project_Objectives_RSM>(.*)</Project_Objectives_RSM>\\s*<Project_Objectives_PC>(.*)</Project_Objectives_PC>\\s*<Project_Performance-IHHL_BPL>(.*)</Project_Performance-IHHL_BPL>\\s*<Project_Performance-IHHL_APL>(.*)</Project_Performance-IHHL_APL>\\s*<Project_Performance-IHHL_TOTAL>(.*)</Project_Performance-IHHL_TOTAL>\\s*<Project_Performance-SCW>(.*)</Project_Performance-SCW>\\s*<Project_Performance-School_Toilets>(.*)</Project_Performance-School_Toilets>\\s*<Project_Performance-Anganwadi_Toilets>(.*)</Project_Performance-Anganwadi_Toilets>\\s*<Project_Performance-RSM>(.*)</Project_Performance-RSM>\\s*<Project_Performance-PC>(.*)</Project_Performance-PC>\\s*</row>'));
- C = FOREACH B GENERATE (chararray)$0 as State, (chararray)$1 as District, (int)$2 as BPL_Objective, (int)$10 as BPL_Performance;
- D = FOREACH C GENERATE District, Pig_UDF.Pig_UDF.PigUdf(BPL_Objective,BPL_Performance) as Result;
- E = FILTER D BY Result == true;
- F = FOREACH E GENERATE District;
- STORE F INTO /home/acadgild/Project-1.2/BPL80P.txt

Now we will try to understand each script in details-

- A = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
  Since the file is in XML format we are loading it using XMLloader by taking the tag <row> as parent field and processing it as chararray-

```
grunt> A = LOAD 'hdfs://localhost:9000/flume_import' USING org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
2017-12-13 01:22:39,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> lim1 = LIMIT A 2;
grunt> dump lim1;
2017-12-13 01:23:32,549 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2017-12-13 01:23:32,583 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-12-13 01:23:32,583 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-12-13 01:23:32,583 [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]}
2017-12-13 01:23:32,620 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-12-13 01:23:32,628 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-13 01:23:32,630 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 01:23:32,630 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-12-13 01:23:32,642 [main] INFO  org.apache.hadoop.mapreduce.lib.input.LineRecordReader - Found UTF-8 BOM and skipped it
2017-12-13 01:23:32,684 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to hdfs://localhost:9000/tmp/temp-1676464451/tmp-663366910/_temporary/0/task__0001_m_000001
2017-12-13 01:23:32,713 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-12-13 01:23:32,718 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 01:23:32,718 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(<row>          <State_Name>Andhra Pradesh</State_Name>         <District_Name>ADILABAD</District_Name>         <Project_Objectives_IHHL_BPL>247475</Project_Objectives_IHHL_BPL>            <Project_Objectives_IHHL_APL>148181</Project_Objectives_IHHL_APL>               <Project_Objectives_IHHL_TOTAL>395656</Project_Objectives_IHHL_TOTAL>               <Project_Objectives_SCW>0</Project_Objectives_SCW>              <Project_Objectives_School_Toilets>4462</Project_Objectives_School_Toilets> <Project_Objectives_Anganwadi_Toilets>427</Project_Objectives_Anganwadi_Toilets>         <Project_Objectives_RSM>10</Project_Objectives_RSM>             <Project_Objectives_PC>0</Project_Objectives_PC>            <Project_Performance-IHHL_BPL>176300</Project_Performance-IHHL_BPL>            <Project_Performance-IHHL_APL>52431</Project_Performance-IHHL_APL>          <Project_Performance-IHHL_TOTAL>228731</Project_Performance-IHHL_TOTAL>         <Project_Performance-SCW>0</Project_Performance-SCW>         <Project_Performance-School_Toilets>4462</Project_Performance-School_Toilets>           <Project_Performance-Anganwadi_Toilets>427</Project_Performance-Anganwadi_Toilets>           <Project_Performance-RSM>0</Project_Performance-RSM>            <Project_Performance-PC>0</Project_Performance-PC>      </row>)
(<row>          <State_Name>Andhra Pradesh</State_Name>         <District_Name>ANANTAPUR</District_Name>                <Project_Objectives_IHHL_BPL>363314</Project_Objectives_IHHL_BPL>            <Project_Objectives_IHHL_APL>181335</Project_Objectives_IHHL_APL>               <Project_Objectives_IHHL_TOTAL>544649</Project_Objectives_IHHL_TOTAL>               <Project_Objectives_SCW>0</Project_Objectives_SCW>              <Project_Objectives_School_Toilets>3421</Project_Objectives_School_Toilets>          <Project_Objectives_Anganwadi_Toilets>284</Project_Objectives_Anganwadi_Toilets>               <Project_Objectives_RSM>10</Project_Objectives_RSM> <Project_Objectives_PC>0</Project_Objectives_PC>         <Project_Performance-IHHL_BPL>366557</Project_Performance-IHHL_BPL>             <Project_Performance-IHHL_APL>42000</Project_Performance-IHHL_APL>          <Project_Performance-IHHL_TOTAL>408557</Project_Performance-IHHL_TOTAL>         <Project_Performance-SCW>0</Project_Performance-SCW>         <Project_Performance-School_Toilets>4258</Project_Performance-School_Toilets>           <Project_Performance-Anganwadi_Toilets>302</Project_Performance-Anganwadi_Toilets>           <Project_Performance-RSM>0</Project_Performance-RSM>            <Project_Performance-PC>0</Project_Performance-PC>      </row>)
```

- 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>'));

Above command is parsing the XML fields of the input fields and extracting the data out of it

```
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_Objec
tives_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> lim2 = LIMIT B 2;
grunt> dump lim2;
```

If we dump above relation we can see below data extracted out of the XML file-

```
2017-12-13 01:28:18,808 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,814 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:18,887 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,890 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:18,935 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,939 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:18,982 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:18,988 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:19,047 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:19,055 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:19,103 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-13 01:28:19,110 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationS
o job history server
2017-12-13 01:28:19,189 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-13 01:28:19,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
2017-12-13 01:28:19,190 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate co
2017-12-13 01:28:19,195 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 01:28:19,195 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(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)
grunt>
```

- C = FOREACH B GENERATE (chararray)$0 as State, (chararray)$1 as District, (int)$2 as BPL_Objective, (int) $10 as BPL_Performance;

Now here we are extracting only required fields like State, district, BPL Objective and BPL Performance-

```
grunt> C = FOREACH B GENERATE (chararray)$0 as State, (chararray)$1 as District, (int)$2 as BPL_Objective, (int) $10 as BPL_Performance;
grunt> lim3 = LIMIT C 5;
grunt> dump lim3;
```

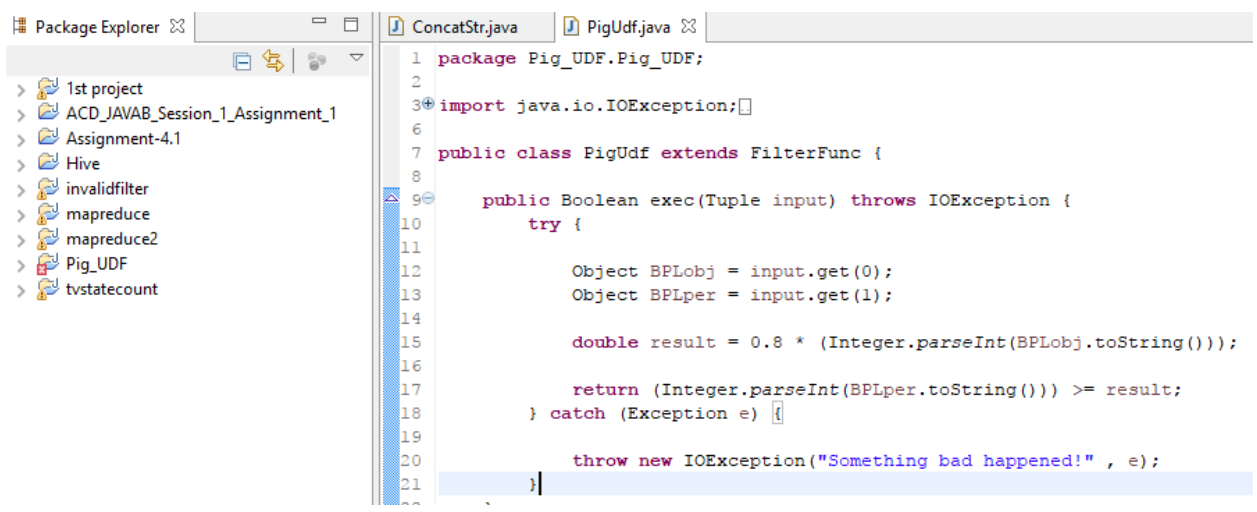
If we dump C we can see below sample result-

```
2017-12-13 02:04:44,180 [main] INFO  org.apache
(Andhra Pradesh,ADILABAD,247475,176300)
(Andhra Pradesh,CHITTOOR,296465,269750)
(Andhra Pradesh,CUDDAPAH,251653,239780)
(Andhra Pradesh,ANANTAPUR,363314,366557)
(Andhra Pradesh,EAST GODAVARI,370255,347305)
grunt>
```

PIG UDF-

Now we here we have made a PIG UDF named PigUdf under JAVA package Pig_UDF.Pig_UDF.

Below is screenshot for same-



In order to run the PIG UDF we will REGISTER the JAR first as shown below-

```
grunt> REGISTER '/home/acadgild/Project-1.2/pig-udf.jar';
2017-12-13 04:43:04,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

- D = FOREACH C GENERATE District, Pig_UDF.Pig_UDF.PigUdf(BPL_Objective,BPL_Performance) as Result;

Now as seen from the definition of PIG UDF it is taking 2 arguments as input and checking whether it has achieved 80% of second. Here we are taking BPL_Objective as first argument and BPL_Performance as second argument and if first argument is 80% of second we are making the result as true.

```
grunt> D = FOREACH C GENERATE District, Pig_UDF.Pig_UDF.PigUdf(BPL_Objective,BPL_Performance) as Result;
grunt> lim = LIMIT D 10;
grunt> dump lim;
```

- E = FILTER D BY Result == true;

Now we are filtering the result D by true

```
grunt> E = FILTER D BY Result == true;
grunt> lim2 = LIMIT E 10;
grunt> dump lim2;
```

Below screenshots shows the sample dump of same-

```
2017-12-13 04:52:31,637 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-13 04:52:31,637 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use f
2017-12-13 04:52:31,637 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
2017-12-13 04:52:31,643 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-12-13 04:52:31,643 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(MEDAK,true)
(KHAMMAM,true)
(KRISHNA,true)
(KURNOOL,true)
(CHITTOOR,true)
(CUDDAPAH,true)
(NALGONDA,true)
(ANANTAPUR,true)
(KARIMNAGAR,true)
(EAST GODAVARI,true)
grunt>
```

- F = FOREACH E GENERATE District;

Since we are interested only in name of districts we will extract District name from relation E as shown below

```
grunt> F = FOREACH E GENERATE District;
grunt> dump F;
```

Below screenshot shows the sample dump for same-

```
2017-12-13 04:55:58,432 [main] INFO  org.apache.pig.backend.hadoc
2017-12-13 04:55:58,432 [main] INFO  org.apache.hadoop.conf.Confi
2017-12-13 04:55:58,432 [main] INFO  org.apache.pig.data.SchemaTu
2017-12-13 04:55:58,439 [main] INFO  org.apache.hadoop.mapreduce.
2017-12-13 04:55:58,440 [main] INFO  org.apache.pig.backend.hadoc
(ANANTAPUR)
(CHITTOOR)
(CUDDAPAH)
(EAST GODAVARI)
(KARIMNAGAR)
(KHAMMAM)
(KRISHNA)
(KURNOOL)
(MEDAK)
(NALGONDA)
(NIZAMABAD)
(RANGAREDDI)
(WARANGAL)
(WEST GODAVARI)
(DIBANG VALLEY)
(LOHIT)
(TIRAP)
(BAGSHA)
(CACHAR)
(DIBRUGARH)
(GOALPARA)
(GOLAGHAT)
(HAILAKANDI)
(JORHAT)
(KAMRUP)
(KARIMGANJ)
(KOKRAJHAR)
(LAKHIMPUR)
(MARIGAON)
```

- STORE F INTO /home/acadgild/Project-1.2/BPL80P.txt

Now we will store above result in HDFS in above location-

```
grunt> STORE F INTO '/home/acadgild/Project-1.2/BPL80P.txt';
```

Same can be seen in below screenshot-

```
grunt> cat hdfs://localhost:9000/home/acadgild/Project-1.2/BPL80P.txt
ANANTAPUR
CHITTOOR
CUDDAPAH
EAST GODAVARI
KARIMNAGAR
KHAMMAM
KRISHNA
KURNOOL
MEDAK
NALGONDA
NIZAMABAD
RANGAREDDI
WARANGAL
WEST GODAVARI
DIBANG VALLEY
LOHIT
TIRAP
BAGSHA
CACHAR
DIBRUGARH
GOALPARA
GOLAGHAT
HAILAKANDI
JORHAT
KAMRUP
KARIMGANJ
KOKRAJHAR
```

Now to store data in MySQL we will create a table BPL80P with column name District_name

```
mysql> CREATE TABLE BPL80P ( District_name varchar(50) );
Query OK, 0 rows affected (0.01 sec)

mysql> show tables;
+---------------+
| Tables_in_db |
+---------------+
| BPL100P       |
| BPL80P        |
| employee      |
+---------------+
3 rows in set (0.00 sec)
```

- sqoop export --connect jdbc:mysql://localhost/db \
- --username 'acadgild' -P --table 'BPL80P' --export-dir 'hdfs://localhost:9000/home/acadgild/Project-1.2/BPL80P.txt/part-m-00000' \
- --input-fields-terminated-by ',' \
- -m 1 --columns District_name

The above script will export data from HDFS location to the MySQL table BPL80P

```
[acadgild@localhost ~]$ sqoop export --connect jdbc:mysql://localhost/db \
> --username 'acadgild' -P --table 'BPL80P' --export-dir 'hdfs://localhost:9000/home/acadgild/Project-1.2/BPL80P.txt/part-m-00000' \
> --input-fields-terminated-by ',' \
> -m 1 --columns District_name
```

If we select from the table BPL80P the results can be seen-

```
Database changed
mysql> select * from BPL80P;
+----------------------------+
| District_name              |
+----------------------------+
| CUDDAPAH                   |
| EAST GODAVARI              |
| KARIMNAGAR                 |
| KHAMMAM                    |
| KRISHNA                    |
| KURNOOL                    |
| MEDAK                      |
| NALGONDA                   |
| NIZAMABAD                  |
| RANGAREDDI                 |
| WARANGAL                   |
| WEST GODAVARI              |
| DIBANG VALLEY              |
| LOHIT                      |
| TIRAP                      |
| BAGSHA                     |
| CACHAR                     |
| DIBRUGARH                  |
| GOALPARA                   |
| GOLAGHAT                   |
| HAILAKANDI                 |
| JORHAT                     |
| KAMRUP                     |
| KARIMGANJ                  |
| KOKRAJHAR                  |
| LAKHIMPUR                  |
| MARIGAON                   |
| NAGAON                     |
| SIBSAGAR                   |
| SONITPUR                   |
| TINSUKIA                   |
| BEGUSARAI                  |
| MADHUBANI                  |
| PARBHANI                   |
| PUNE                       |
| SATARA                     |
| SINDHUDURG                 |
| THANE                      |
| WARDHA                     |
| BISHNUPUR                  |
| MANSA                      |
| EAST SIKKIM                |
| MADURAI                    |
| WEST TRIPURA               |
| AGRA                       |
| ALIGARH                    |
| ALLAHABAD                  |
| AMBEDKAR NAGAR             |
| AZAMGARH                   |
| BAGPAT                     |
| BALLIA                     |
| BALRAMPUR                  |
| BANDA                      |
| BARABANKI                  |
| BAREILLY                   |
| BASTI                      |
| BIJNOR                     |
| BUDAUN                     |
| BULANDSHAHR                |
| CHANDAULI                  |
| CHITRAKOOT                 |
| DEORIA                     |
| ETAH                       |
| ETAWAH                     |
| HARDOI                     |
| HARIDWAR                   |
| NAINITAL                   |
+----------------------------+
```