Project 2.1- State-Wise Development Analysis In India

- 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

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.

Dataset - https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view -

Solution -

1st Phase

- First Checked the folders in hdfs , then created a folder 'flume_import' in it.

hadoop fs -ls /
hadoop fs -mkdir /flume_import

```
                        acadgild@localhost:~/RITESH/STATE
File  Edit  View  Search  Terminal  Help
[acadgild@localhost STATE]$ hadoop fs -ls /
18/06/21 01:16:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 5 items
drwxr-xr-x   - acadgild supergroup          0 2018-06-19 10:14 /hbase
drwxr-xr-x   - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
-rw-r--r--   1 acadgild supergroup      26204 2018-06-19 10:07 /student.txt
drwxrwx---   - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-02-09 14:50 /user
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost STATE]$ hadoop fs -mkdir /flume_import
18/06/21 01:17:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost STATE]$ hadoop fs -ls /
18/06/21 01:18:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 6 items
drwxr-xr-x   - acadgild supergroup          0 2018-06-21 01:17 /flume_import
drwxr-xr-x   - acadgild supergroup          0 2018-06-19 10:14 /hbase
drwxr-xr-x   - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
-rw-r--r--   1 acadgild supergroup      26204 2018-06-19 10:07 /student.txt
drwxrwx---   - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-02-09 14:50 /user
[acadgild@localhost STATE]$
```

- Then using flume we transfered the data from the local file system that is spool directory to HDFS.

flume-ng agent -n agent1 -g /home/aadgild/RITESH/STATE/filecopy.conf


```
                        acadgild@localhost:~/RITESH/STATE
File  Edit  View  Search  Terminal  Help
[acadgild@localhost STATE]$ flume-ng agent -n  agent1 -f /home/acadgild/RITESH/STATE/filecopy.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/home/acadgild/install/hbase/hbase-1.2.6/bin/hbase) for HBASE access
Info: Including Hive libraries found via (/home/acadgild/install/hive/apache-hive-2.3.2-bin) for Hive access
+ exec /usr/java/jdk1.8.0_151/bin/java -Xmx20m -cp '/home/acadgild/install/flume/apache-flume-1.8.0-bin/lib/*:/home/acadgild/
install/hadoop/hadoop-2.6.5/etc/hadoop/:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/*:/home/acadgild/i
nstall/hadoop/hadoop-2.6.5/share/hadoop/common/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs:/home/acadgild/
install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/*:/home/acad
gild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/*:/home
/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/ma
preduce/*:/home/acadgild/install/hadoop/hadoop-2.6.5/contrib/capacity-scheduler/*.jar:/home/acadgild/install/hbase/hbase-1.2.
6/conf:/usr/java/jdk1.8.0_151/lib/tools.jar:/home/acadgild/install/hbase/hbase-1.2.6:/home/acadgild/install/hbase/hbase-1.2.6
/lib/activation-1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/aopalliance-1.0.jar:/home/acadgild/install/hbase/hbase-1
.2.6/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home
/acadgild/install/hbase/hbase-1.2.6/lib/api-asn1-api-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-util-1.0.
0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/asm-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/avro-1.7.4.ja
r:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-1.7.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commo
ns-beanutils-core-1.8.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-cli-1.2.jar:/home/acadgild/install/hbase/hba
se-1.2.6/lib/commons-codec-1.9.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-collections-3.2.2.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/commons-compress-1.4.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-configuration-1
.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-daemon-1.0.13.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/co
mmons-digester-1.8.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-el-1.0.jar:/home/acadgild/install/hbase/hbase-1.2
.6/lib/commons-httpclient-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-io-2.4.jar:/home/acadgild/install/hbas
e/hbase-1.2.6/lib/commons-lang-2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-logging-1.2.jar:/home/acadgild/in
stall/hbase/hbase-1.2.6/lib/commons-math-2.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math3-3.1.1.jar:/home/a
cadgild/install/hbase/hbase-1.2.6/lib/commons-net-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/disruptor-3.3.0.jar:/h
ome/acadgild/install/hbase/hbase-1.2.6/lib/findbugs-annotations-1.3.9-1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guav
a-12.0.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-se
rvlet-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-annotations-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.
2.6/lib/hadoop-auth-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-client-2.5.1.jar:/home/acadgild/install/hba
se/hbase-1.2.6/lib/hadoop-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-hdfs-2.5.1.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-app-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapre
duce-client-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-core-2.5.1.jar:/home/acadgi
ld/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-jobclient-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/had
oop-mapreduce-client-shuffle-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-api-2.5.1.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/hadoop-yarn-client-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-common-2.
5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-server-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1
.2.6/lib/hbase-annotations-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-annotations-1.2.6-tests.jar:/home/aca
```

```
                          acadgild@localhost:~/RITESH/STATE                         _ □ ×
File  Edit  View  Search  Terminal  Help
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
18/06/21 01:22:49 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting
18/06/21 01:22:49 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/acadgild/RITESH/ST
ATE/filecopy.conf
18/06/21 01:22:49 INFO conf.FlumeConfiguration: Processing:hdfsdest
18/06/21 01:22:49 INFO conf.FlumeConfiguration: Processing:hdfsdest
18/06/21 01:22:49 INFO conf.FlumeConfiguration: Processing:hdfsdest
18/06/21 01:22:49 INFO conf.FlumeConfiguration: Added sinks: hdfsdest Agent: agent1
18/06/21 01:22:49 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [agent
1]
18/06/21 01:22:49 INFO node.AbstractConfigurationProvider: Creating channels
18/06/21 01:22:49 INFO channel.DefaultChannelFactory: Creating instance of channel mychannel type memory
18/06/21 01:22:49 INFO node.AbstractConfigurationProvider: Created channel mychannel
18/06/21 01:22:49 INFO source.DefaultSourceFactory: Creating instance of source mysrc, type exec
18/06/21 01:22:49 INFO sink.DefaultSinkFactory: Creating instance of sink: hdfsdest, type: hdfs
18/06/21 01:22:49 INFO node.AbstractConfigurationProvider: Channel mychannel connected to [mysrc, hdfsdest]
18/06/21 01:22:49 INFO node.Application: Starting new configuration:{ sourceRunners:{mysrc=EventDrivenSourceRunner: { source:
org.apache.flume.source.ExecSource{name:mysrc,state:IDLE} }} sinkRunners:{hdfsdest=SinkRunner: { policy:org.apache.flume.sink
.DefaultSinkProcessor@3c95d288 counterGroup:{ name:null counters:{} } }} channels:{mychannel=org.apache.flume.channel.MemoryC
hannel{name: mychannel}} }
18/06/21 01:22:49 INFO node.Application: Starting Channel mychannel
18/06/21 01:22:49 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel: Suc
cessfully registered new MBean.
18/06/21 01:22:49 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
18/06/21 01:22:49 INFO node.Application: Starting Sink hdfsdest
18/06/21 01:22:49 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfsdest: Success
fully registered new MBean.
18/06/21 01:22:49 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfsdest started
18/06/21 01:22:49 INFO node.Application: Starting Source mysrc
18/06/21 01:22:49 INFO source.ExecSource: Exec source starting with command: hadoop dfs -put /home/acadgild/RITESH/STATE/spoo
l/StatewiseDistrictwisePhysicalProgress.xml /flume_import
18/06/21 01:22:49 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: mysrc: Successf
ully registered new MBean.
18/06/21 01:22:49 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc started
18/06/21 01:22:52 INFO source.ExecSource: Command [hadoop dfs -put /home/acadgild/RITESH/STATE/spool/StatewiseDistrictwisePhy
sicalProgress.xml /flume_import] exited with 0
```

This flume process goes on continuosly to fetch data from local file system to HDFS.

The flume configuration file **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/RITESH/STATE/spool/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

- Then the flume import folder is checked wheteher data got imported.
  hadoop fs -ls /flume_import

```
                              acadgild@localhost:~/RITESH/STATE              _  □  ×
File  Edit  View  Search  Terminal  Help
[acadgild@localhost STATE]$ hadoop fs -ls /flume_import
18/06/21 01:25:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup      717414 2018-06-21 01:22 /flume_import/StatewiseDistrictwisePhysicalProgress.xml
[acadgild@localhost STATE]$ ▊
```

- Startedthe pig shell
  pig -x local

```
                              acadgild@localhost:~/RITESH/STATE              _  □  ×
File  Edit  View  Search  Terminal  Help
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost STATE]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/06/21 01:41:39 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/06/21 01:41:39 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2018-06-21 01:41:39,693 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:4
9
2018-06-21 01:41:39,693 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/RITESH/STATE/pig_1529525
499665.log
2018-06-21 01:41:39,760 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/acadgild/.pigbootup not found
2018-06-21 01:41:40,084 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2018-06-21 01:41:40,087 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-21 01:41:40,088 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: file:///
2018-06-21 01:41:40,273 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-21 01:41:40,310 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-821cb433-23fc-46f0
-b835-069840796eba
2018-06-21 01:41:40,310 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to fa
lse
grunt> ▊
```

1.
   - Registering the **piggybank** jar that contains the executables for various pig functions. to parse XML.

     REGISTER /home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar;

```
grunt> REGISTER /home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar;
2018-06-21 01:42:25,552 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-21 01:42:25,564 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt> ▊
```

   - Defining the XML Parse function as **Xpath**

     **DEFINE XPath
     org.apache.pig.piggybank.evaluation.xml.XPath;**

```
grunt> DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath;
grunt> ▮
```

- Loading the data in from HDFS and then using the XML Loader function to load the data into the relation with every starting tag 'row' as one line of type: chararray with the name **x.**

  data = load 'hdfs://localhost:8020/flume_import/StatewiseDistrictwisePhysicalProgress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

```
grunt> data = load 'hdfs://localhost:8020/flume_import/StatewiseDistrictwisePhysicalProgress.xml'
>> using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
```

- Generating the rows in relation by using the XML Parser **XPath.** Every required tag for the anlysis under the main tag **row** will be separated by the tag name and given a pseudo name in the relation.

req_data = foreach data generate XPath(x,'row/State_Name') as state, XPath(x,'row/District_Name') as DistrictName, XPath(x,'row/Project_Objectives_IHHL_BPL') as POIHHLBPL, XPath(x,'row/Project_Performance-IHHL_BPL') as PPIHHLBPL;

```
grunt> req_data = foreach data generate XPath(x,'/row/State_Name') as state, XPath(x,'/row/District_Name') as DistrictName, X
Path(x,'/row/Project_Objectives_IHHL_BPL') as POIHHLBPL, XPath(x,'/row/Project_Performance-IHHL_BPL') as PPIHHLBPL;
```

- required fields are filtered if null is present.
  filter_null = filter req_data by (PPIHHLBPL is not null) AND (POIHHLBPL is not null);

```
grunt> filter_null = filter req_data by (PPIHHLBPL is not null) AND (POIHHLBPL is not null);
```

- Filter the data where performance objective is equal to performance achiedved.

centPercentObjective = filter filter_null by((int)PPIHHLBPL==(int)POIHHLBPL);

```
grunt> centPercentObjective = filter filter_null by ((int)PPIHHLBPL==(int)POIHHLBPL);
grunt> █
```

- List is generated containg state , districts , performance objective and performance achieved of BPL cards.
  districts = foreach centPercentObjective generate state,DistrictName,POIHHLBPL,PPIHHLBPL;

```
grunt> districts = foreach centPercentObjective generate state,DistrictName,POIHHLBPL,PPIHHLBPL;
grunt> █
```

- Stored the final results into HDFS
  store districts into 'hdfs://localhost:8020/pig_output/districtwiseachieved' using PigStorage(',');

```
grunt> store districts into 'hdfs://localhost:8020/pig_output/districtwiseachieved' using PigStorage(',');█
```

```
                                  acadgild@localhost:~/RITESH/STATE                                    _  □
 File  Edit  View  Search  Terminal  Help
[acadgild@localhost STATE]$ hadoop fs -ls /pig_output/districtwiseachieved
18/06/21 10:18:43 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--   3 acadgild supergroup          0 2018-06-21 10:12 /pig_output/districtwiseachieved/_SUCCESS
-rw-r--r--   3 acadgild supergroup       2330 2018-06-21 10:12 /pig_output/districtwiseachieved/part-m-00000
[acadgild@localhost STATE]$ hadoop fs -cat /pig_output/districtwiseachieved/part-m-00000
18/06/21 10:19:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Andhra Pradesh  NIZAMABAD        225519  225519
Arunachal Pradesh       TIRAP    5780    5780
Assam   HAILAKANDI      49837    49837
Bihar   MADHUBANI       67482    67482
Goa     NORTH GOA       15000    15000
Gujarat AHMEDABAD       80192    80192
Gujarat DANGS   27900   27900
Gujarat NAVSARI 75015   75015
Gujarat PORBANDAR       17024    17024
Gujarat SURAT   158797  158797
Haryana FARIDABAD       22254    22254
Haryana HISAR   46463   46463
Haryana JHAJJAR 22014   22014
Haryana MAHENDRAGARH    17500    17500
Haryana PANCHKULA       8760     8760
Haryana PANIPAT 28000   28000
Haryana ROHTAK  22171   22171
Haryana SIRSA   35400   35400
Himachal Pradesh        HAMIRPUR        11593   11593
Himachal Pradesh        KINNAUR 1560    1560
Himachal Pradesh        KULLU   9989    9989
Himachal Pradesh        LAHAUL & SPITI  2413    2413
Himachal Pradesh        SHIMLA  23874   23874
Himachal Pradesh        SOLAN   10858   10858
Himachal Pradesh        UNA     8360    8360
Jharkhand       DEOGHAR 75153   75153
Jharkhand       LOHARDAGA       22626   22626
Karnataka       HASSAN  64134   64134
Karnataka       MANGALORE(DAKSHINA KANNADA)     59478   59478
Karnataka       UDUPI   52348   52348
```

- Checked the content of the pig output in hdfs.
  hadoop fs -ls /pig_output/districtwiseachieved

- Created a database project and a table 'districtCentPercent'.
  create database project;

use project;
create table districtCentPercent(state varchar(25), district_name
varchar(30), POIHHLBPL int , PPIHHLBPL int);

- Using sqoop we exported the data from HDFS to mysql.
  sqoop export –connect jdbc:mysql://localhost/project –username
  'root' -password 'Root@123' –table 'districtCentPercent' –
  export-dir '/pig_output/districtwiseachieved/part-m-00000' –
  input-fields-terminated-by ',' -m 1 –columns
  state,district,name,POIHHLBPL,PPIHHLBPL

```
                File System Counters
                        FILE: Number of bytes read=0
                        FILE: Number of bytes written=127623
                        FILE: Number of read operations=0
                        FILE: Number of large read operations=0
                        FILE: Number of write operations=0
                        HDFS: Number of bytes read=2480
                        HDFS: Number of bytes written=0
                        HDFS: Number of read operations=4
                        HDFS: Number of large read operations=0
                        HDFS: Number of write operations=0
                Job Counters
                        Launched map tasks=1
                        Data-local map tasks=1
                        Total time spent by all maps in occupied slots (ms)=4938
                        Total time spent by all reduces in occupied slots (ms)=0
                        Total time spent by all map tasks (ms)=4938
                        Total vcore-milliseconds taken by all map tasks=4938
                        Total megabyte-milliseconds taken by all map tasks=5056512
                Map-Reduce Framework
                        Map input records=70
                        Map output records=70
                        Input split bytes=147
                        Spilled Records=0
                        Failed Shuffles=0
                        Merged Map outputs=0
                        GC time elapsed (ms)=62
                        CPU time spent (ms)=1180
                        Physical memory (bytes) snapshot=117760000
                        Virtual memory (bytes) snapshot=2063929344
                        Total committed heap usage (bytes)=62980096
                File Input Format Counters
                        Bytes Read=0
                File Output Format Counters
                        Bytes Written=0
18/06/21 12:06:01 INFO mapreduce.ExportJobBase: Transferred 2.4219 KB in 22.1694 seconds (111.8661 bytes/sec)
18/06/21 12:06:01 INFO mapreduce.ExportJobBase: Exported 70 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost STATE]$
```

- We checked the content of the table in mysql.
  select * from districtCentPercent;

```
mysql> select * from districtCentPercent;
```

state	district_name	POIHHLBPL	PPIHHLBPL
Andhra Pradesh	NIZAMABAD	225519	225519
Arunachal Pradesh	TIRAP	5780	5780
Assam	HAILAKANDI	49837	49837
Bihar	MADHUBANI	67482	67482
Goa	NORTH GOA	15000	15000
Gujarat	AHMEDABAD	80192	80192
Gujarat	DANGS	27900	27900
Gujarat	NAVSARI	75015	75015
Gujarat	PORBANDAR	17024	17024
Gujarat	SURAT	158797	158797
Haryana	FARIDABAD	22254	22254
Haryana	HISAR	46463	46463
Haryana	JHAJJAR	22014	22014
Haryana	MAHENDRAGARH	17500	17500
Haryana	PANCHKULA	8760	8760
Haryana	PANIPAT	28000	28000
Haryana	ROHTAK	22171	22171
Haryana	SIRSA	35400	35400
Himachal Pradesh	HAMIRPUR	11593	11593
Himachal Pradesh	KINNAUR	1560	1560
Himachal Pradesh	KULLU	9989	9989
Himachal Pradesh	LAHAUL & SPITI	2413	2413
Himachal Pradesh	SHIMLA	23874	23874
Himachal Pradesh	SOLAN	10858	10858
Himachal Pradesh	UNA	8360	8360
Jharkhand	DEOGHAR	75153	75153
Jharkhand	LOHARDAGA	22626	22626
Karnataka	HASSAN	64134	64134
Karnataka	MANGALORE(DAKSHINA KANNADA)	59478	59478
Karnataka	UDUPI	52348	52348
Kerala	ALAPPUZHA	114359	114359
Kerala	KOLLAM	95130	95130
Kerala	KOTTAYAM	28118	28118

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

```
70 rows in set (0.00 sec)

mysql>
```

2.

- Registering the **piggybank** jar that contains the executables for various pig functions. to parse XML.

  REGISTER /home/acadgild/install/pig/pig0.16.0/lib/piggybank.jar;

```
grunt> REGISTER /home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar;
2018-06-21 01:42:25,552 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-21 01:42:25,564 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
grunt>
```

- Defining the XML Parse function as **Xpath**

  DEFINE XPath org.apache.pig.piggybank.evaluation.xml.Xpath;

```
grunt> DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath;
grunt>
```

- Registering pig UDF jar.
  REGISTER '/home/acadgild/RITESH/STATE/Filter80.jar';

```
grunt> REGISTER /home/acadgild/RITESH/STATE/Filter80.jar;
```

  <u>Filter80 code:</u>

```
package pigUDF;

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

public class Filter80 extends EvalFunc<Double>{
        public Double exec(Tuple input) throws IOException {
                double performance = (double)input.get(0);
                double objective = (double)input.get(1);

                double percentage = (performance/objective)*100;

                if(percentage>80.0)
```

```
                    return percentage;
              else
                    return 0.0;
        }
    }
```

- Loading the data in from HDFS and then using the XML Loader function to load the data into the relation with every starting tag 'row' as one line of type: chararray with the name **x.**

data = load 'hdfs://localhost:8020/flume_import/StatewiseDistrictwisePhysicalPro gress.xml' using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);

```
grunt> data = load 'hdfs://localhost:8020/flume_import/StatewiseDistrictwisePhysicalProgress.xml'
>> using org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
```

- Generating the rows in relation by using the XML Parser **XPath.** Every required tag for the anlysis under the main tag **row** will be separated by the tag name and given a pseudo name in the relation.

req_data = foreach data generate XPath(x,'row/State_Name') as state, XPath(x,'row/District_Name') as DistrictName, XPath(x,'row/Project_Objectives_IHHL_BPL') as POIHHLBPL, XPath(x,'row/Project_Performance-IHHL_BPL') as PPIHHLBPL;

```
grunt> req_data = foreach data generate XPath(x,'/row/State_Name') as state, XPath(x,'/row/District_Name') as DistrictName, X
Path(x,'/row/Project_Objectives_IHHL_BPL') as POIHHLBPL, XPath(x,'/row/Project_Performance-IHHL_BPL') as PPIHHLBPL;
```

- required fields are filtered if null is present.
  filter_null = filter req_data by (PPIHHLBPL is not null) AND (POIHHLBPL is not null);

```
grunt> filter_null = filter req_data by (PPIHHLBPL is not null) AND (POIHHLBPL is not null);
```

- List is generated where it contains sate,district name, percentage completion of BPL cards.

percentObjective = foreach filter_null generate state,DistrictName,ROUND_TO(pigUDF.Filter80((double)PPIH HLBPL,(double)POIHHLBPL),2) as BPL_Percent;

```
grunt> percentObjective = foreach filter_null  generate state,DistrictName,ROUND_TO(pigUDF.Filter80((double)PPIHHLBPL,(double
)POIHHLBPL),2) as BPL_Percent;
```

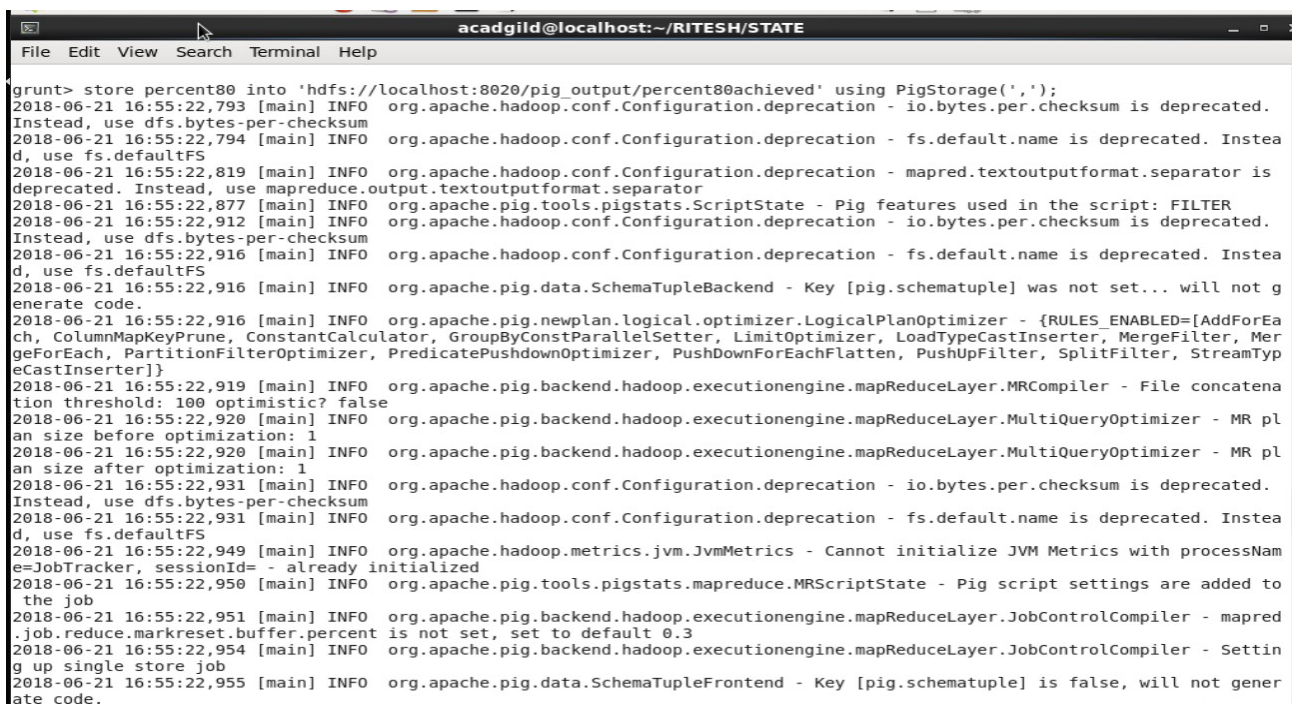- Filtered the data where BPL cards completion is greater than 80

percent80 = FILTER percentObjective BY BPL_Percent>=80.00;

```
grunt> percent80 = FILTER percentObjective BY BPL_Percent>=80.00;
```

- Store the data into hdfs

store percent80 into 'hdfs://localhost:8020/pig_output/percent80achieved' using PigStorage(',');

```
acadgild@localhost:~/RITESH/STATE
File  Edit  View  Search  Terminal  Help

grunt> store percent80 into 'hdfs://localhost:8020/pig_output/percent80achieved' using PigStorage(',');
2018-06-21 16:55:22,793 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-21 16:55:22,794 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-21 16:55:22,819 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is
deprecated. Instead, use mapreduce.output.textoutputformat.separator
2018-06-21 16:55:22,877 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2018-06-21 16:55:22,912 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-21 16:55:22,916 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-21 16:55:22,916 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not g
enerate code.
2018-06-21 16:55:22,916 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Mer
geForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTyp
eCastInserter]}
2018-06-21 16:55:22,919 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2018-06-21 16:55:22,920 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2018-06-21 16:55:22,920 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2018-06-21 16:55:22,931 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-06-21 16:55:22,931 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-06-21 16:55:22,949 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-06-21 16:55:22,950 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to
 the job
2018-06-21 16:55:22,951 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2018-06-21 16:55:22,954 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Settin
g up single store job
2018-06-21 16:55:22,955 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not gener
ate code.
```

- Created the table 'Percent80_district'
  create table districtCentPercent(state varchar(25), district_name varchar(30),BPL_percent double);

- Checked the wheteher the pig output was stored or not in HDFS.

hadoop fs -ls /pig_output/percent80achieved
hadoop fs -cat /pig_output/percent80achieved/part-m-00000

```
                                    acadgild@localhost:~/RITESH/STATE                          _ □ ×
File  Edit  View  Search  Terminal  Help
[acadgild@localhost STATE]$ hadoop fs -ls /pig_output/percent80achieved
18/06/21 17:02:16 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--   3 acadgild supergroup          0 2018-06-21 16:55 /pig_output/percent80achieved/_SUCCESS
-rw-r--r--   3 acadgild supergroup       9478 2018-06-21 16:55 /pig_output/percent80achieved/part-m-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost STATE]$ hadoop fs -cat  /pig_output/percent80achieved/part-m-00000
18/06/21 17:02:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Andhra Pradesh,ANANTAPUR,100.89
Andhra Pradesh,CHITTOOR,90.99
Andhra Pradesh,CUDDAPAH,95.28
Andhra Pradesh,EAST GODAVARI,93.8
Andhra Pradesh,KARIMNAGAR,101.14
Andhra Pradesh,KHAMMAM,103.46
Andhra Pradesh,KRISHNA,90.66
Andhra Pradesh,KURNOOL,84.39
Andhra Pradesh,MEDAK,99.63
Andhra Pradesh,NALGONDA,104.54
Andhra Pradesh,NIZAMABAD,100.0
Andhra Pradesh,RANGAREDDI,82.05
Andhra Pradesh,WARANGAL,108.92
Andhra Pradesh,WEST GODAVARI,92.8
Arunachal Pradesh,DIBANG VALLEY,100.28
Arunachal Pradesh,LOHIT,95.57
Arunachal Pradesh,TIRAP,100.0
Assam,BAGSHA,85.77
Assam,CACHAR,84.28
Assam,DIBRUGARH,90.09
Assam,GOALPARA,85.67
Assam,GOLAGHAT,97.8
Assam,HAILAKANDI,100.0
Assam,JORHAT,80.99
Assam,KAMRUP,98.76
Assam,KARIMGANJ,95.96
Assam,KOKRAJHAR,89.46
Assam,LAKHIMPUR,83.06
Assam,MARIGAON,92.96
```

- Using sqoop we exported the data from HDFS to mysql.
  **sqoop export –connect jdbc:mysql://localhost/project –username
  'root' -password <u>'Root@123</u>' –table 'Percent80_district' –export-
  dir '/pig_output/percent80achieved/part-m-00000' –input-fields-
  terminated-by ',' -m 1 –columns state,district,name,BPL_Percent**


```
                                    acadgild@localhost:~/RITESH/STATE                          _ □ x
File  Edit  View  Search  Terminal  Help
[acadgild@localhost STATE]$ sqoop export --connect jdbc:mysql://localhost/project --username 'root' -password 'Root@123' --ta
ble 'Percent80_district' --export-dir '/pig_output/percent80achieved/part-m-00000' --input-fields-terminated-by ',' -m 1 --co
lumns state,district_name,BPL_Percent
Warning: /home/acadgild/install/sqoop/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/install/sqoop/sqoop-1.4.6.bin__hadoop-2.0.4-alpha/../accumulo does not exist! Accumulo imports will f
ail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/06/21 17:06:06 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
18/06/21 17:06:06 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/06/21 17:06:06 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
```

```
                                    acadgild@localhost:~/RITESH/STATE                          _ □
File  Edit  View  Search  Terminal  Help
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=127609
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9625
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6571
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=6571
                Total vcore-milliseconds taken by all map tasks=6571
                Total megabyte-milliseconds taken by all map tasks=6728704
        Map-Reduce Framework
                Map input records=349
                Map output records=349
                Input split bytes=144
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=59
                CPU time spent (ms)=1270
                Physical memory (bytes) snapshot=120860672
                Virtual memory (bytes) snapshot=2062385152
                Total committed heap usage (bytes)=62980096
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
18/06/21 17:06:37 INFO mapreduce.ExportJobBase: Transferred 9.3994 KB in 24.593 seconds (391.3714 bytes/sec)
18/06/21 17:06:37 INFO mapreduce.ExportJobBase: Exported 349 records.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost STATE]$
```

- Checked the content of the table in mysql

select * from Percent80_district;

File   Edit   View   Search   Terminal   Help

```
mysql> select * from Percent80_district;
+--------------------+------------------------+-------------+
| state              | district_name          | BPL_Percent |
+--------------------+------------------------+-------------+
| Andhra Pradesh     | ANANTAPUR              |      100.89 |
| Andhra Pradesh     | CHITTOOR               |       90.99 |
| Andhra Pradesh     | CUDDAPAH               |       95.28 |
| Andhra Pradesh     | EAST GODAVARI          |        93.8 |
| Andhra Pradesh     | KARIMNAGAR             |      101.14 |
| Andhra Pradesh     | KHAMMAM                |      103.46 |
| Andhra Pradesh     | KRISHNA                |       90.66 |
| Andhra Pradesh     | KURNOOL                |       84.39 |
| Andhra Pradesh     | MEDAK                  |       99.63 |
| Andhra Pradesh     | NALGONDA               |      104.54 |
| Andhra Pradesh     | NIZAMABAD              |         100 |
| Andhra Pradesh     | RANGAREDDI             |       82.05 |
| Andhra Pradesh     | WARANGAL               |      108.92 |
| Andhra Pradesh     | WEST GODAVARI          |        92.8 |
| Arunachal Pradesh  | DIBANG VALLEY          |      100.28 |
| Arunachal Pradesh  | LOHIT                  |       95.57 |
| Arunachal Pradesh  | TIRAP                  |         100 |
| Assam              | BAGSHA                 |       85.77 |
| Assam              | CACHAR                 |       84.28 |
| Assam              | DIBRUGARH              |       90.09 |
| Assam              | GOALPARA               |       85.67 |
| Assam              | GOLAGHAT               |        97.8 |
| Assam              | HAILAKANDI             |         100 |
| Assam              | JORHAT                 |       80.99 |
| Assam              | KAMRUP                 |       98.76 |
| Assam              | KARIMGANJ              |       95.96 |
| Assam              | KOKRAJHAR              |       89.46 |
| Assam              | LAKHIMPUR              |       83.06 |
| Assam              | MARIGAON               |       92.96 |
| Assam              | NAGAON                 |          94 |
| Assam              | SIBSAGAR               |       85.39 |
| Assam              | SONITPUR               |       84.34 |
| Assam              | TINSUKIA               |       85.51 |
| Bihar              | BEGUSARAI              |       87.28 |
| Bihar              | MADHUBANI              |         100 |
```

File   Edit   View   Search   Terminal   Help

Uttar Pradesh	MUZAFFARNAGAR	100
Uttar Pradesh	PILIBHIT	100
Uttar Pradesh	PRATAPGARH	101.52
Uttar Pradesh	RAE BARELI	100.07
Uttar Pradesh	RAMPUR	91.23
Uttar Pradesh	SAHARANPUR	100.26
Uttar Pradesh	SANT RAVIDAS NAGAR(BHADOHI)	93.06
Uttar Pradesh	SHAHJAHANPUR	100.16
Uttar Pradesh	SHRAVASTI	94.15
Uttar Pradesh	SIDDHARTHNAGAR	89.25
Uttar Pradesh	SITAPUR	89.57
Uttar Pradesh	SONBHADRA	100
Uttar Pradesh	SULTANPUR	100
Uttar Pradesh	UNNAO	97.4
Uttar Pradesh	VARANASI	97.17
Uttarakhand	BAGESHWAR	83.27
Uttarakhand	CHAMOLI	95.93
Uttarakhand	DEHRADUN	85.25
Uttarakhand	HARIDWAR	89.85
Uttarakhand	NAINITAL	126.02
Uttarakhand	PITHORAGARH	84.16
Uttarakhand	RUDRAPRAYAG	105.02
Uttarakhand	TEHRI GARHWAL	83.91
Uttarakhand	UDHAM SINGH NAGAR	95.38
Uttarakhand	UTTARKASHI	90.54
West Bengal	BARDHAMAN	85.98
West Bengal	DAKSHIN DINAJPUR	100.84
West Bengal	HOOGHLY	99.28
West Bengal	HOWRAH	99.28
West Bengal	JALPAIGURI	90.55
West Bengal	MIDNAPUR EAST	134.41
West Bengal	MIDNAPUR WEST	117.04
West Bengal	NADIA	92.72
West Bengal	NORTH 24 PARAGANAS	99.03
West Bengal	SOUTH 24 PARAGANAS	94.43

349 rows in set (0.00 sec)

mysql>