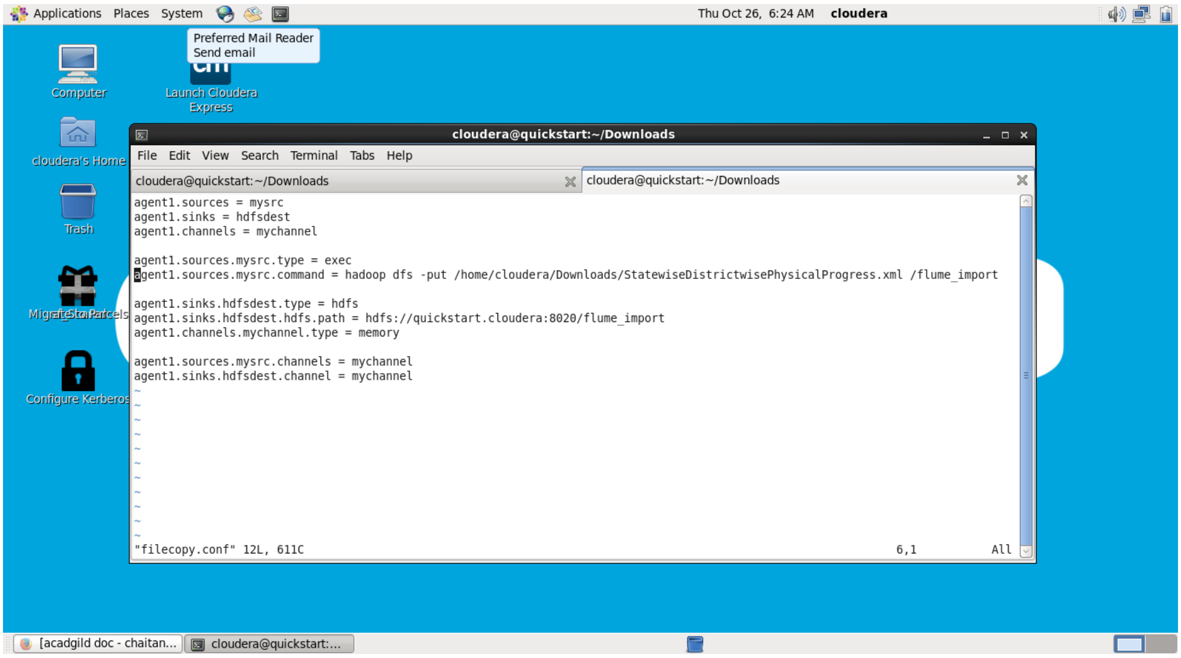
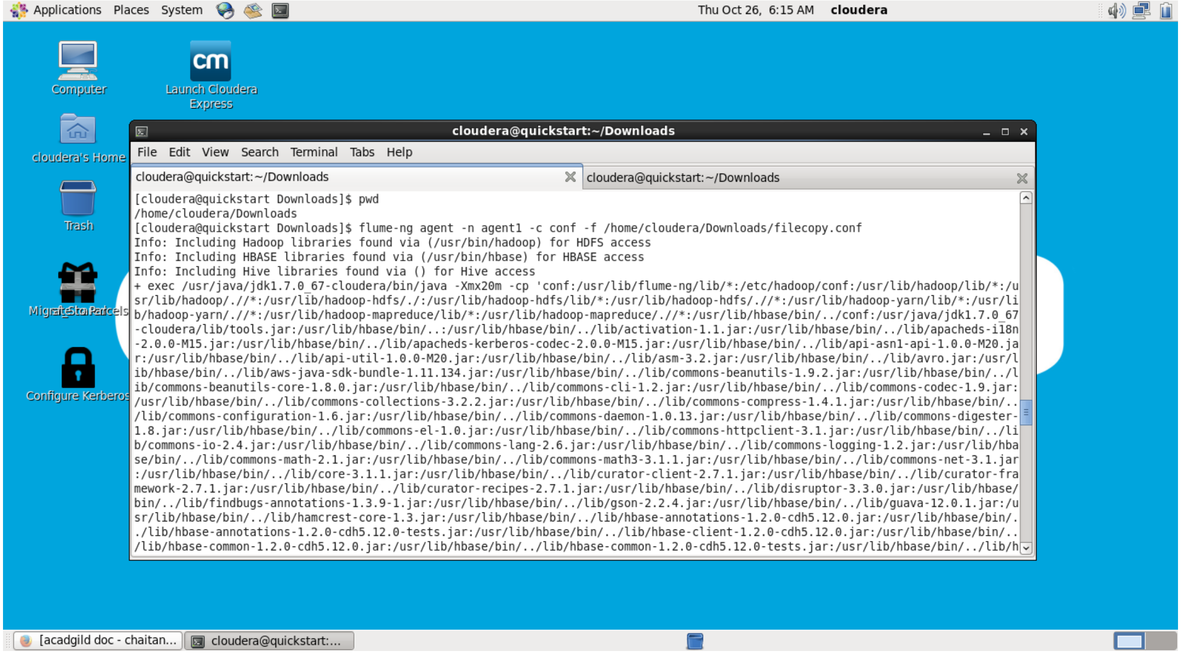
Task 1:
The FLUME job which will format the data and place the data to HDFS

1. Conf file to download dataset from local file system to HDFS flume:

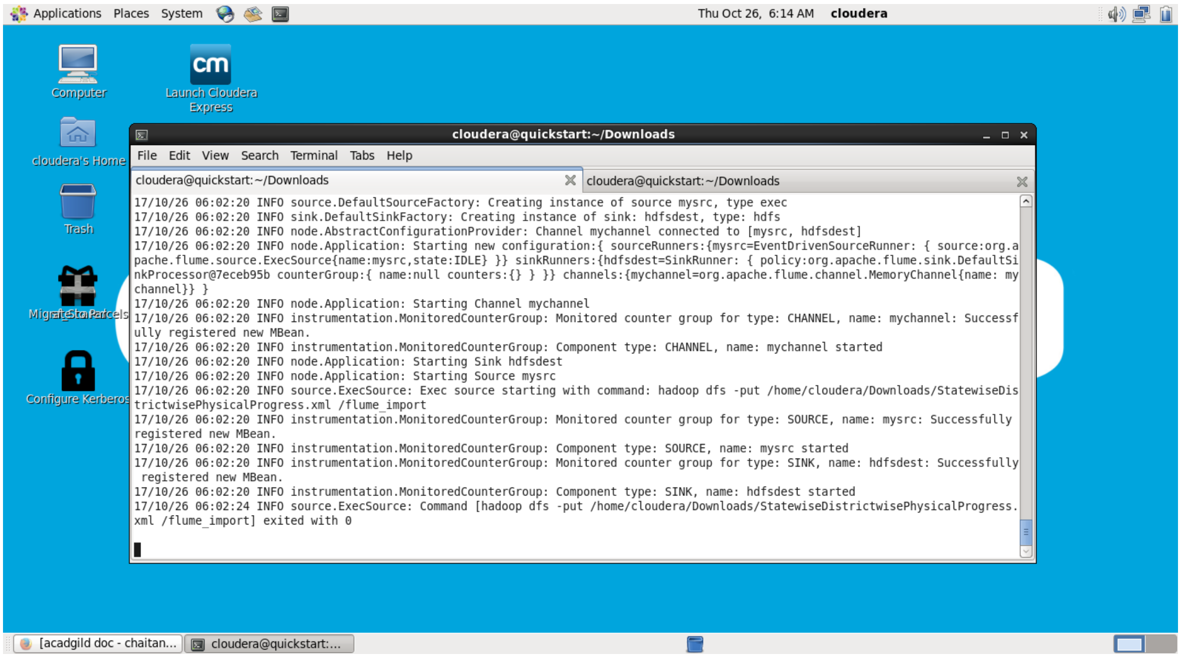

Applications Places System
Thu Oct 26, 6:24 AM cloudera
Preferred Mail Reader
Send email
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

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

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

agent1.sinks.hdfsdest.type = hdfs
agent1.sinks.hdfsdest.hdfs.path = hdfs://quickstart.cloudera:8020/flume_import
agent1.channels.mychannel.type = memory

agent1.sources.mysrc.channels = mychannel
agent1.sinks.hdfsdest.channel = mychannel
~
~
~
~
~
~
~
~
~
~
~
"filecopy.conf" 12L, 611C                                                             6,1           All
```

2. Run flume-ng command to copy the file to HDFS.

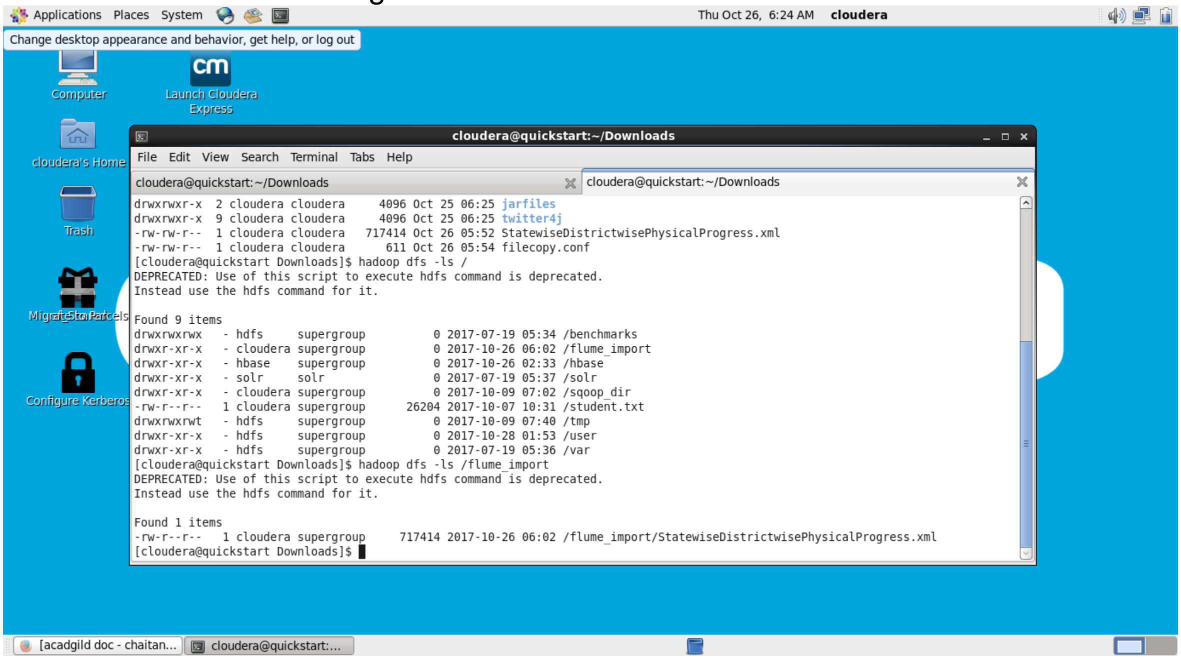

Applications Places System
Thu Oct 26, 6:15 AM cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

```
[cloudera@quickstart Downloads]$ pwd
/home/cloudera/Downloads
[cloudera@quickstart Downloads]$ flume-ng agent -n agent1 -c conf -f /home/cloudera/Downloads/filecopy.conf
Info: Including Hadoop libraries found via (/usr/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/usr/bin/hbase) for HBASE access
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java/jdk1.7.0_67-cloudera/bin/java -Xmx20m -cp 'conf:/usr/lib/flume-ng/lib/*:/etc/hadoop/conf:/usr/lib/hadoop/lib/*:/u
sr/lib/hadoop/.//*:/usr/lib/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/.//*:/usr/lib/hadoop-yarn/lib/*:/usr/li
b/hadoop-yarn/.//*:/usr/lib/hadoop-mapreduce/lib/*:/usr/lib/hadoop-mapreduce/.//*:/usr/lib/hbase/bin/../conf:/usr/java/jdk1.7.0_67
-cloudera/lib/tools.jar:/usr/lib/hbase/bin/..:/usr/lib/hbase/bin/../lib/activation-1.1.jar:/usr/lib/hbase/bin/../lib/apacheds-i18n
-2.0.0-M15.jar:/usr/lib/hbase/bin/../lib/apacheds-kerberos-codec-2.0.0-M15.jar:/usr/lib/hbase/bin/../lib/api-asn1-api-1.0.0-M20.ja
r:/usr/lib/hbase/bin/../lib/api-util-1.0.0-M20.jar:/usr/lib/hbase/bin/../lib/asm-3.2.jar:/usr/lib/hbase/bin/../lib/avro.jar:/usr/l
ib/hbase/bin/../lib/aws-java-sdk-bundle-1.11.134.jar:/usr/lib/hbase/bin/../lib/commons-beanutils-1.9.2.jar:/usr/lib/hbase/bin/../l
ib/commons-beanutils-core-1.8.0.jar:/usr/lib/hbase/bin/../lib/commons-cli-1.2.jar:/usr/lib/hbase/bin/../lib/commons-codec-1.9.jar:
/usr/lib/hbase/bin/../lib/commons-collections-3.2.2.jar:/usr/lib/hbase/bin/../lib/commons-compress-1.4.1.jar:/usr/lib/hbase/bin/..
/lib/commons-configuration-1.6.jar:/usr/lib/hbase/bin/../lib/commons-daemon-1.0.13.jar:/usr/lib/hbase/bin/../lib/commons-digester-
1.8.jar:/usr/lib/hbase/bin/../lib/commons-el-1.0.jar:/usr/lib/hbase/bin/../lib/commons-httpclient-3.1.jar:/usr/lib/hbase/bin/../li
b/commons-io-2.4.jar:/usr/lib/hbase/bin/../lib/commons-lang-2.6.jar:/usr/lib/hbase/bin/../lib/commons-logging-1.2.jar:/usr/lib/hba
se/bin/../lib/commons-math-2.1.jar:/usr/lib/hbase/bin/../lib/commons-math3-3.1.1.jar:/usr/lib/hbase/bin/../lib/commons-net-3.1.jar
:/usr/lib/hbase/bin/../lib/core-3.1.1.jar:/usr/lib/hbase/bin/../lib/curator-client-2.7.1.jar:/usr/lib/hbase/bin/../lib/curator-fra
mework-2.7.1.jar:/usr/lib/hbase/bin/../lib/curator-recipes-2.7.1.jar:/usr/lib/hbase/bin/../lib/disruptor-3.3.0.jar:/usr/lib/hbase/
bin/../lib/findbugs-annotations-1.3.9-1.jar:/usr/lib/hbase/bin/../lib/gson-2.2.4.jar:/usr/lib/hbase/bin/../lib/guava-12.0.1.jar:/u
sr/lib/hbase/bin/../lib/hamcrest-core-1.3.jar:/usr/lib/hbase/bin/../lib/hbase-annotations-1.2.0-cdh5.12.0.jar:/usr/lib/hbase/bin/.
./lib/hbase-annotations-1.2.0-cdh5.12.0-tests.jar:/usr/lib/hbase/bin/../lib/hbase-client-1.2.0-cdh5.12.0.jar:/usr/lib/hbase/bin/..
/lib/hbase-common-1.2.0-cdh5.12.0.jar:/usr/lib/hbase/bin/../lib/hbase-common-1.2.0-cdh5.12.0-tests.jar:/usr/lib/hbase/bin/../lib/h
```


Applications Places System
Thu Oct 26, 6:14 AM cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

```
17/10/26 06:02:20 INFO source.DefaultSourceFactory: Creating instance of source mysrc, type exec
17/10/26 06:02:20 INFO sink.DefaultSinkFactory: Creating instance of sink: hdfsdest, type: hdfs
17/10/26 06:02:20 INFO node.AbstractConfigurationProvider: Channel mychannel connected to [mysrc, hdfsdest]
17/10/26 06:02:20 INFO node.Application: Starting new configuration:{ sourceRunners:{mysrc=EventDrivenSourceRunner: { source:org.a
pache.flume.source.ExecSource{name:mysrc,state:IDLE} }} sinkRunners:{hdfsdest=SinkRunner: { policy:org.apache.flume.sink.DefaultSi
nkProcessor@7eceb95b counterGroup:{ name:null counters:{} } }} channels:{mychannel=org.apache.flume.channel.MemoryChannel{name: my
channel}} }
17/10/26 06:02:20 INFO node.Application: Starting Channel mychannel
17/10/26 06:02:20 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: mychannel: Successf
ully registered new MBean.
17/10/26 06:02:20 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: mychannel started
17/10/26 06:02:20 INFO node.Application: Starting Sink hdfsdest
17/10/26 06:02:20 INFO node.Application: Starting Source mysrc
17/10/26 06:02:20 INFO source.ExecSource: Exec source starting with command: hadoop dfs -put /home/cloudera/Downloads/StatewiseDis
trictwisePhysicalProgress.xml /flume_import
17/10/26 06:02:20 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: mysrc: Successfully
registered new MBean.
17/10/26 06:02:20 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: mysrc started
17/10/26 06:02:20 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfsdest: Successfully
 registered new MBean.
17/10/26 06:02:20 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfsdest started
17/10/26 06:02:24 INFO source.ExecSource: Command [hadoop dfs -put /home/cloudera/Downloads/StatewiseDistrictwisePhysicalProgress.
xml /flume_import] exited with 0
```

4. Check file in HDFS using ls command :


Applications Places System
Thu Oct 26, 6:24 AM cloudera
Change desktop appearance and behavior, get help, or log out
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

```
drwxrwxr-x  2 cloudera cloudera     4096 Oct 25 06:25 jarfiles
drwxrwxr-x  9 cloudera cloudera     4096 Oct 25 06:25 twitter4j
-rw-rw-r--  1 cloudera cloudera   717414 Oct 26 05:52 StatewiseDistrictwisePhysicalProgress.xml
-rw-rw-r--  1 cloudera cloudera      611 Oct 26 05:54 filecopy.conf
[cloudera@quickstart Downloads]$ hadoop dfs -ls /
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 9 items
drwxrwxrwx   - hdfs     supergroup          0 2017-07-19 05:34 /benchmarks
drwxr-xr-x   - cloudera supergroup          0 2017-10-26 06:02 /flume_import
drwxr-xr-x   - hbase    supergroup          0 2017-10-26 02:33 /hbase
drwxr-xr-x   - solr     solr                0 2017-07-19 05:37 /solr
drwxr-xr-x   - cloudera supergroup          0 2017-10-09 07:02 /sqoop_dir
-rw-r--r--   1 cloudera supergroup      26204 2017-10-07 10:31 /student.txt
drwxrwxrwt   - hdfs     supergroup          0 2017-10-09 07:40 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-10-28 01:53 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-07-19 05:36 /var
[cloudera@quickstart Downloads]$ hadoop dfs -ls /flume_import
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 1 items
-rw-r--r--   1 cloudera supergroup     717414 2017-10-26 06:02 /flume_import/StatewiseDistrictwisePhysicalProgress.xml
[cloudera@quickstart Downloads]$
```

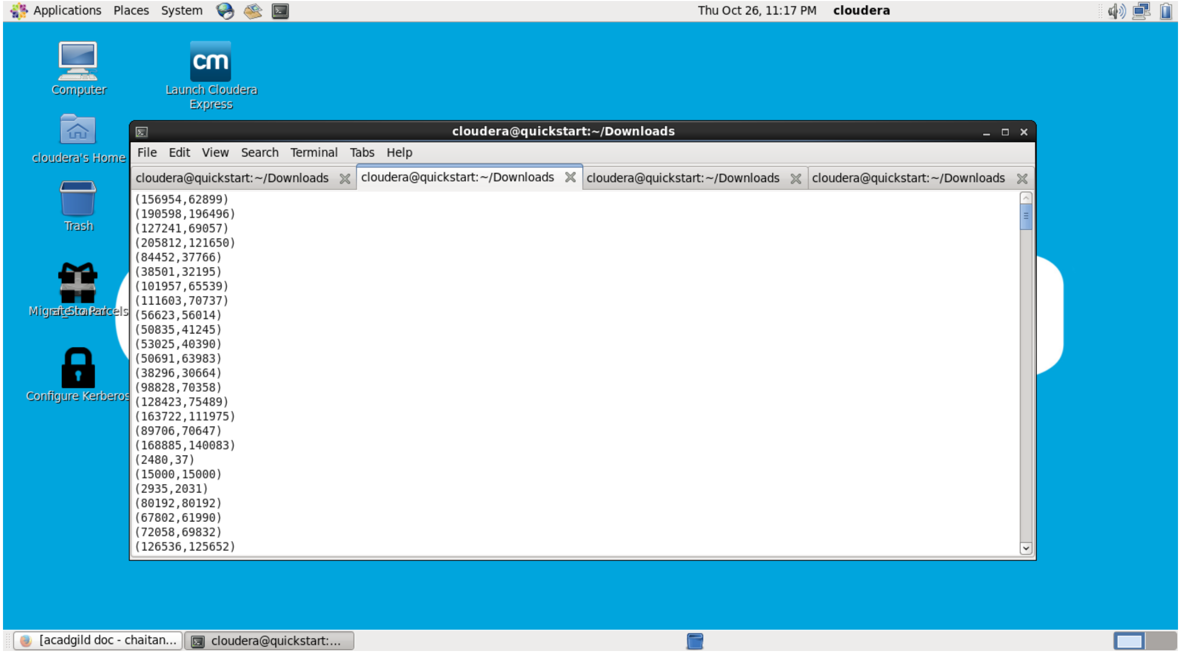
Task 2:
Pig/MapReduce job for parsing the XML data.

Pig Script:

```
REGISTER '/home/cloudera/Downloads/jarfiles/piggybank-0.17.0.jar'
DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();
A = LOAD '/flume_import/StatewiseDistrictwisePhysicalProgress.xml' using
org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
B = FOREACH A GENERATE XPath(x,'row/Project_Objectives_IHHL_BPL'),
XPath(x,'row/Project_Performance-IHHL_BPL');
dump B;
```

Execution:
Pig <pig_file_name>

Output:



Applications Places System
Thu Oct 26, 11:17 PM cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

```
(156954,62899)
(190598,196496)
(127241,69057)
(205812,121650)
(84452,37766)
(38501,32195)
(101957,65539)
(111603,70737)
(56623,56014)
(50835,41245)
(53025,40390)
(50691,63983)
(38296,30664)
(98828,70358)
(128423,75489)
(163722,111975)
(89706,70647)
(168885,140083)
(2480,37)
(15000,15000)
(2935,2031)
(80192,80192)
(67802,61990)
(72058,69832)
(126536,125652)
```

Task 3:
Create Pig scripts/MapReduce jobs to analyze the data

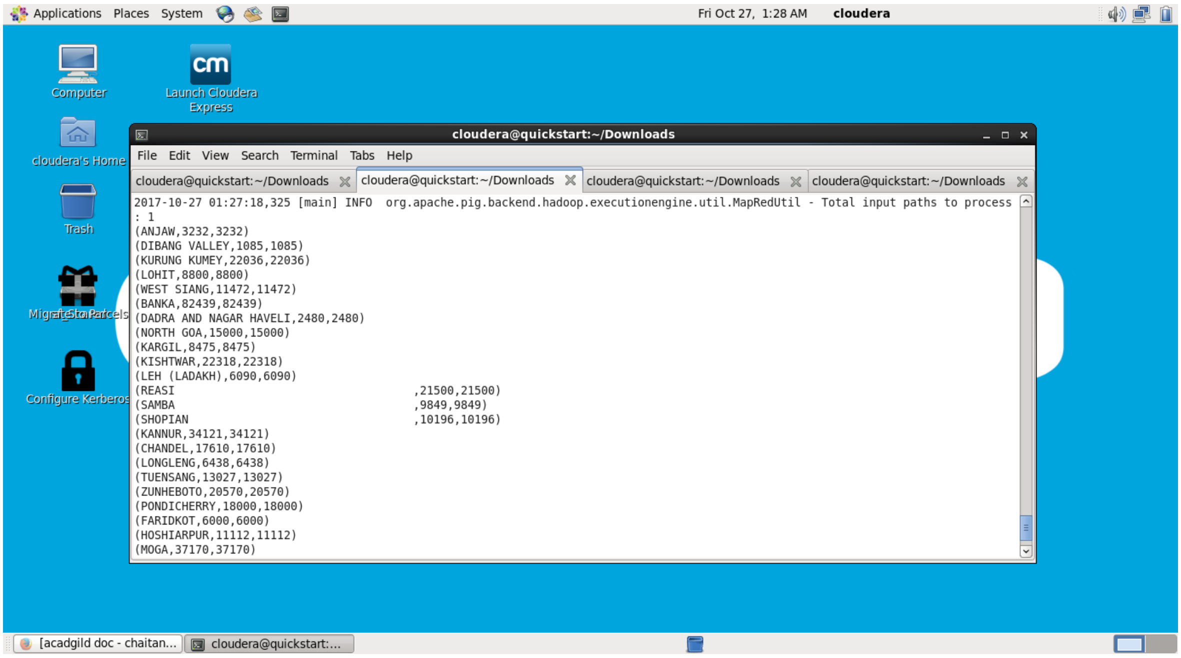
Find out the districts who achieved 100 percent objective in BPL cards

Pig script:

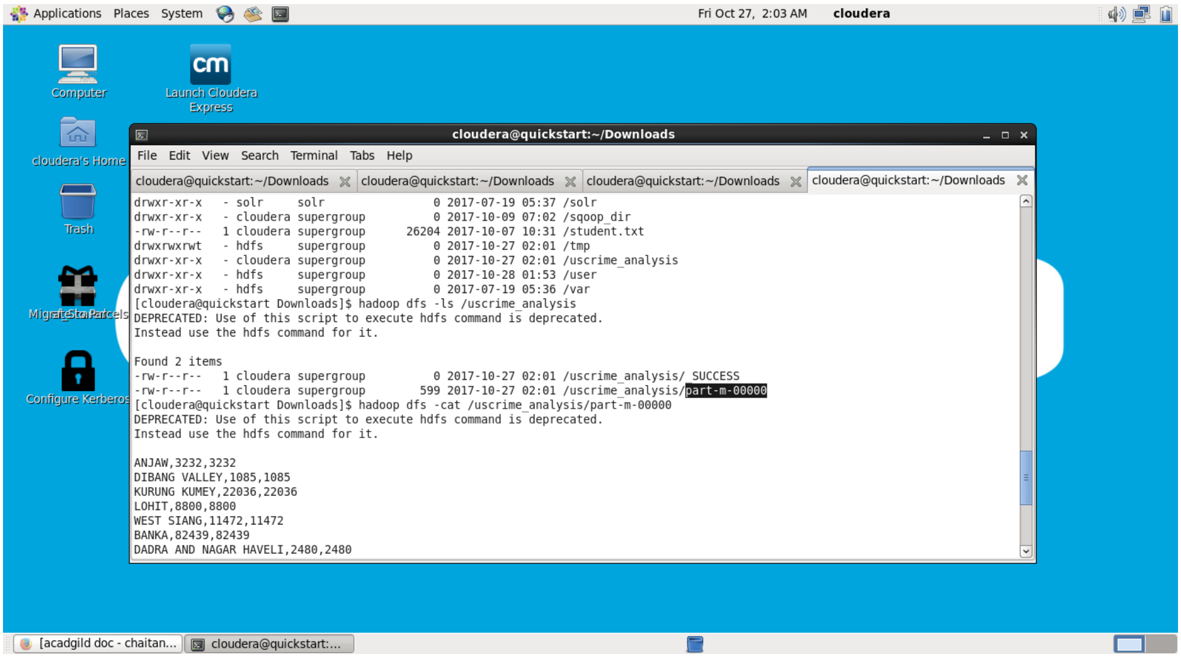
```
REGISTER '/home/cloudera/Downloads/jarfiles/piggybank-0.17.0.jar'
DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();
A = LOAD '/flume_import/StatewiseDistrictwisePhysicalProgress.xml' using
org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
B = FOREACH A GENERATE XPath(x,'row/District_Name') as district ,XPath(x,'row/
Project_Objectives_IHHL_BPL') as BPL_Objective, XPath(x,'row/
Project_Objectives_IHHL_TOTAL') as BPL_Objective_total;
C = filter B by (((int)BPL_Objective * 100)/(int)BPL_Objective_total) == 100;
STORE C INTO 'hdfs://quickstart.cloudera:8020/uscrime_analysis' USING
PigStorage (',');
dump C;
```

Execution:
Pig <file_name_path>

Output:


Applications Places System
Fri Oct 27, 1:28 AM
cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

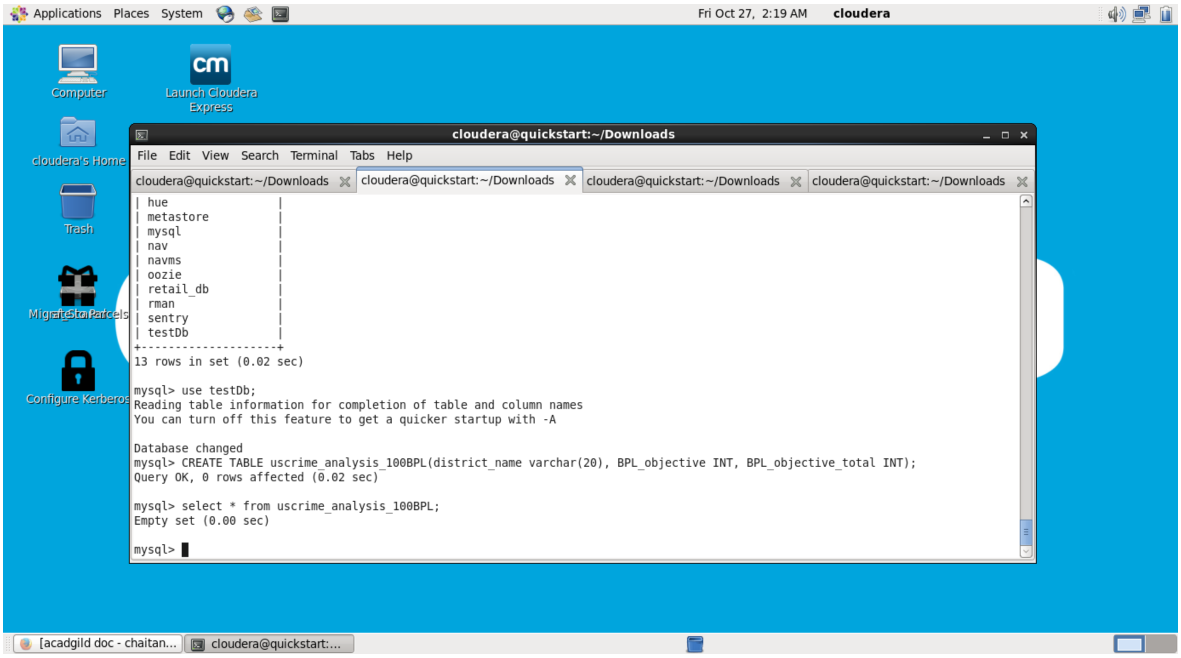
```
2017-10-27 01:27:18,325 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
: 1
(ANJAW,3232,3232)
(DIBANG VALLEY,1085,1085)
(KURUNG KUMEY,22036,22036)
(LOHIT,8800,8800)
(WEST SIANG,11472,11472)
(BANKA,82439,82439)
(DADRA AND NAGAR HAVELI,2480,2480)
(NORTH GOA,15000,15000)
(KARGIL,8475,8475)
(KISHTWAR,22318,22318)
(LEH (LADAKH),6090,6090)
(REASI                                  ,21500,21500)
(SAMBA                                  ,9849,9849)
(SHOPIAN                                ,10196,10196)
(KANNUR,34121,34121)
(CHANDEL,17610,17610)
(LONGLENG,6438,6438)
(TUENSANG,13027,13027)
(ZUNHEBOTO,20570,20570)
(PONDICHERRY,18000,18000)
(FARIDKOT,6000,6000)
(HOSHIARPUR,11112,11112)
(MOGA,37170,37170)
```


Applications Places System
Fri Oct 27, 2:03 AM
cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate to Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

```
drwxr-xr-x   - solr     solr              0 2017-07-19 05:37 /solr
drwxr-xr-x   - cloudera supergroup        0 2017-10-09 07:02 /sqoop_dir
-rw-r--r--   1 cloudera supergroup    26204 2017-10-07 10:31 /student.txt
drwxrwxrwt   - hdfs     supergroup        0 2017-10-27 02:01 /tmp
drwxr-xr-x   - cloudera supergroup        0 2017-10-27 02:01 /uscrime_analysis
drwxr-xr-x   - hdfs     supergroup        0 2017-10-28 01:53 /user
drwxr-xr-x   - hdfs     supergroup        0 2017-07-19 05:36 /var
[cloudera@quickstart Downloads]$ hadoop dfs -ls /uscrime_analysis
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 cloudera supergroup        0 2017-10-27 02:01 /uscrime_analysis/_SUCCESS
-rw-r--r--   1 cloudera supergroup      599 2017-10-27 02:01 /uscrime_analysis/part-m-00000
[cloudera@quickstart Downloads]$ hadoop dfs -cat /uscrime_analysis/part-m-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

ANJAW,3232,3232
DIBANG VALLEY,1085,1085
KURUNG KUMEY,22036,22036
LOHIT,8800,8800
WEST SIANG,11472,11472
BANKA,82439,82439
DADRA AND NAGAR HAVELI,2480,2480
```

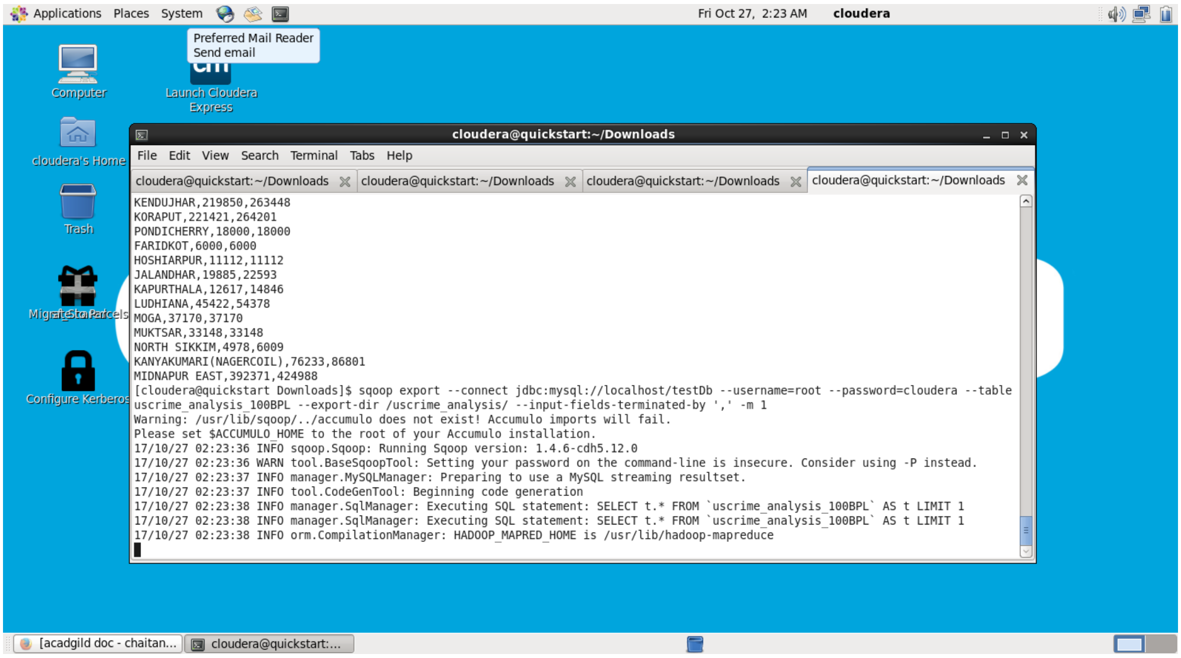
Export the results to mysql using sqoop:

1. Create table in mysql



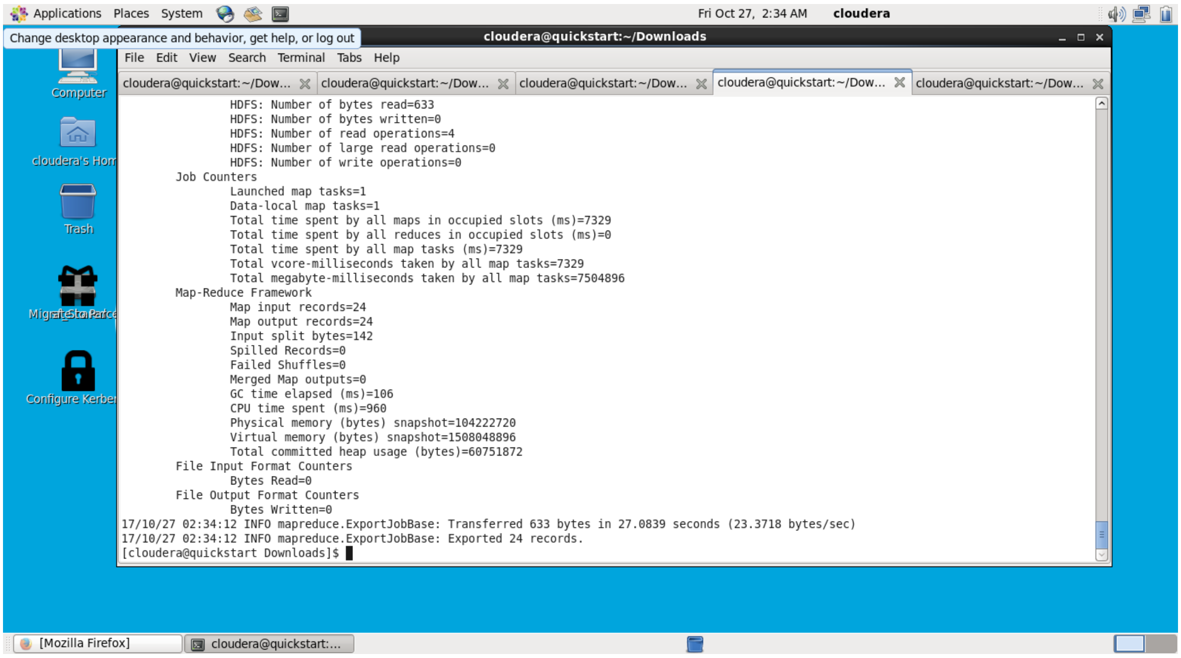
```
| hue                |
| metastore          |
| mysql              |
| nav                |
| navms              |
| oozie              |
| retail_db          |
| rman               |
| sentry             |
| testDb             |
+--------------------+
13 rows in set (0.02 sec)

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

Database changed
mysql> CREATE TABLE uscrime_analysis_100BPL(district_name varchar(20), BPL_objective INT, BPL_objective_total INT);
Query OK, 0 rows affected (0.02 sec)

mysql> select * from uscrime_analysis_100BPL;
Empty set (0.00 sec)

mysql> 
```

2. Scoop command to export the data from HDFS to mysql



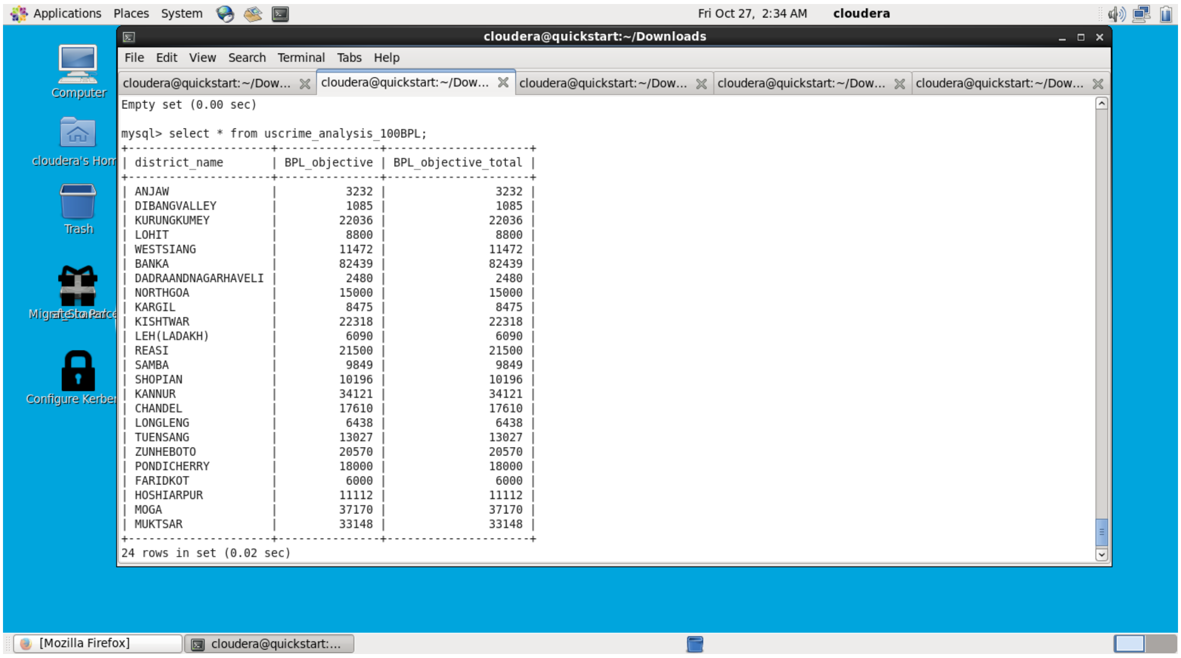
```
KENDUJHAR,219850,263448
KORAPUT,221421,264201
PONDICHERRY,18000,18000
FARIDKOT,6000,6000
HOSHIARPUR,11112,11112
JALANDHAR,19885,22593
KAPURTHALA,12617,14846
LUDHIANA,45422,54378
MOGA,37170,37170
MUKTSAR,33148,33148
NORTH SIKKIM,4978,6009
KANYAKUMARI(NAGERCOIL),76233,86801
MIDNAPUR EAST,392371,424988
[cloudera@quickstart Downloads]$ sqoop export --connect jdbc:mysql://localhost/testDb --username=root --password=cloudera --table
uscrime_analysis_100BPL --export-dir /uscrime_analysis/ --input-fields-terminated-by ',' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/10/27 02:23:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
17/10/27 02:23:36 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
17/10/27 02:23:37 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/10/27 02:23:37 INFO tool.CodeGenTool: Beginning code generation
17/10/27 02:23:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `uscrime_analysis_100BPL` AS t LIMIT 1
17/10/27 02:23:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `uscrime_analysis_100BPL` AS t LIMIT 1
17/10/27 02:23:38 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
```

3. once the job completed, check for the success command:


Applications Places System
Fri Oct 27, 2:34 AM cloudera
Change desktop appearance and behavior, get help, or log out
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
Computer
cloudera's Home
Trash
Configure Kerberos
[Mozilla Firefox]
cloudera@quickstart:...

```
                HDFS: Number of bytes read=633
                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)=7329
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=7329
                Total vcore-milliseconds taken by all map tasks=7329
                Total megabyte-milliseconds taken by all map tasks=7504896
        Map-Reduce Framework
                Map input records=24
                Map output records=24
                Input split bytes=142
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=106
                CPU time spent (ms)=960
                Physical memory (bytes) snapshot=104222720
                Virtual memory (bytes) snapshot=1508048896
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
17/10/27 02:34:12 INFO mapreduce.ExportJobBase: Transferred 633 bytes in 27.0839 seconds (23.3718 bytes/sec)
17/10/27 02:34:12 INFO mapreduce.ExportJobBase: Exported 24 records.
[cloudera@quickstart Downloads]$ 
```

4. Check the exported data in mysql select statement:


Applications Places System
Fri Oct 27, 2:34 AM cloudera
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
Computer
cloudera's Home
Trash
Configure Kerberos
[Mozilla Firefox]
cloudera@quickstart:...

```
Empty set (0.00 sec)

mysql> select * from uscrime_analysis_100BPL;
+-----------------------+---------------+---------------------+
| district_name         | BPL_objective | BPL_objective_total |
+-----------------------+---------------+---------------------+
| ANJAW                 |          3232 |                3232 |
| DIBANGVALLEY          |          1085 |                1085 |
| KURUNGKUMEY           |         22036 |               22036 |
| LOHIT                 |          8800 |                8800 |
| WESTSIANG             |         11472 |               11472 |
| BANKA                 |         82439 |               82439 |
| DADRAANDNAGARHAVELI   |          2480 |                2480 |
| NORTHGOA              |         15000 |               15000 |
| KARGIL                |          8475 |                8475 |
| KISHTWAR              |         22318 |               22318 |
| LEH(LADAKH)           |          6090 |                6090 |
| REASI                 |         21500 |               21500 |
| SAMBA                 |          9849 |                9849 |
| SHOPIAN               |         10196 |               10196 |
| KANNUR                |         34121 |               34121 |
| CHANDEL               |         17610 |               17610 |
| LONGLENG              |          6438 |                6438 |
| TUENSANG              |         13027 |               13027 |
| ZUNHEBOTO             |         20570 |               20570 |
| PONDICHERRY           |         18000 |               18000 |
| FARIDKOT              |          6000 |                6000 |
| HOSHIARPUR            |         11112 |               11112 |
| MOGA                  |         37170 |               37170 |
| MUKTSAR               |         33148 |               33148 |
+-----------------------+---------------+---------------------+
24 rows in set (0.02 sec)
```

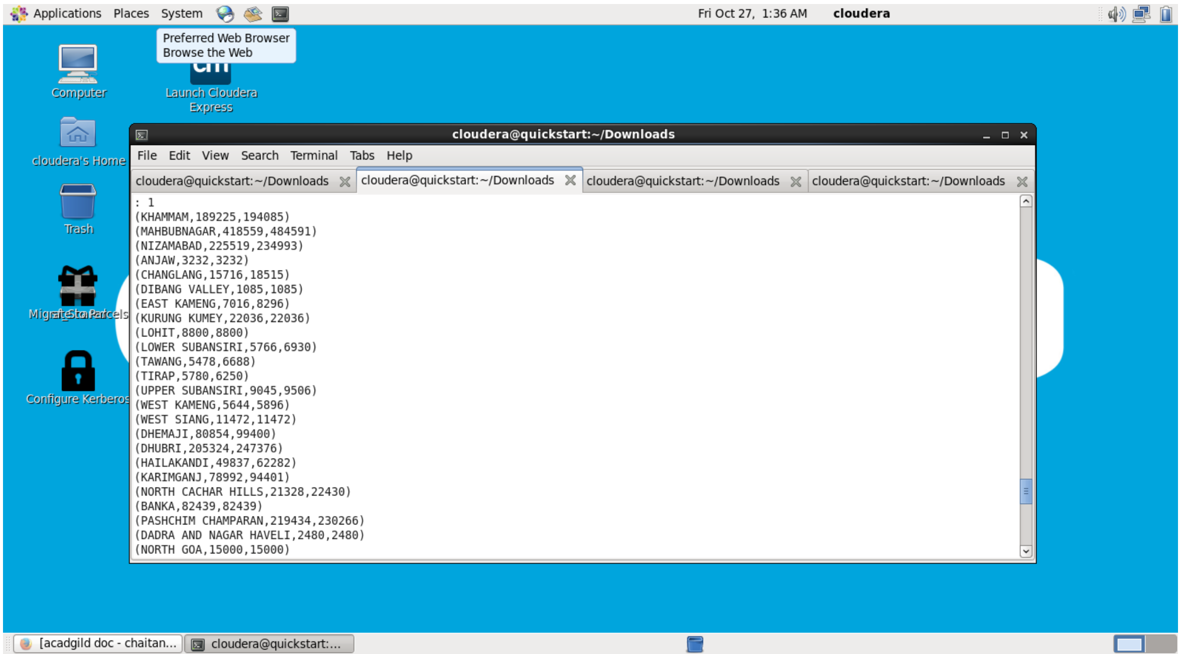
2.Write a Pig UDF to filter the districts which have reached 80% of objectives of BPL cards.

Pig script:

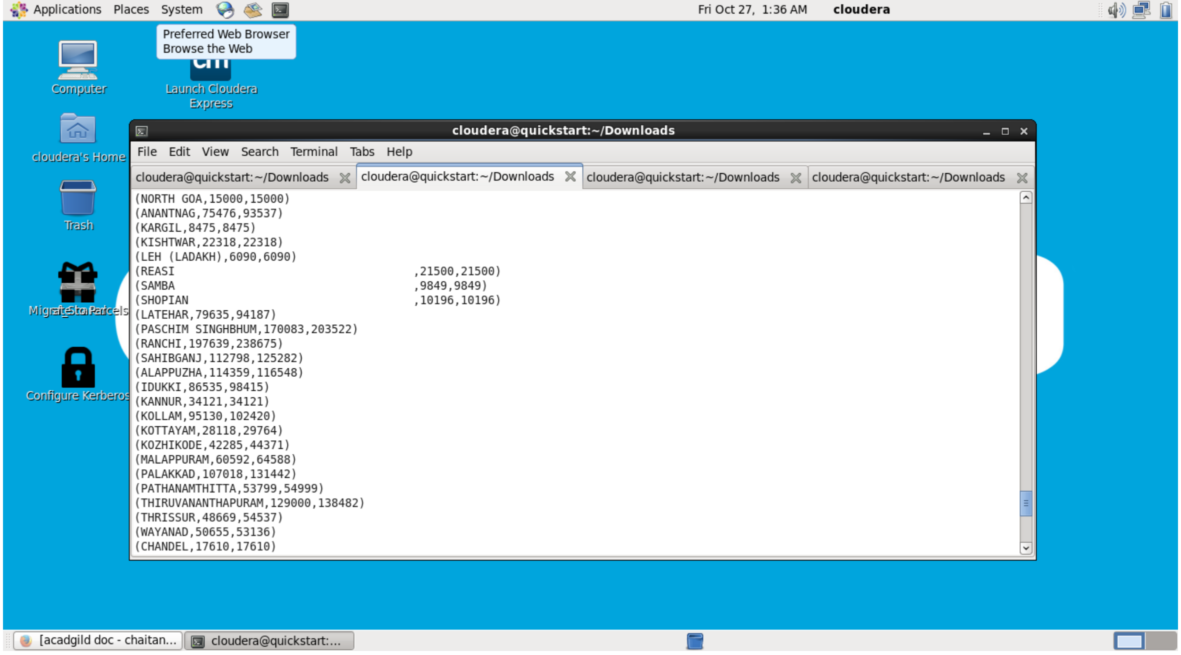
```
REGISTER '/home/cloudera/Downloads/jarfiles/piggybank-0.17.0.jar'
DEFINE XPath org.apache.pig.piggybank.evaluation.xml.XPath();
A = LOAD '/flume_import/StatewiseDistrictwisePhysicalProgress.xml' using
org.apache.pig.piggybank.storage.XMLLoader('row') as (x:chararray);
B = FOREACH A GENERATE XPath(x,'row/District_Name') as district ,XPath(x,'row/
Project_Objectives_IHHL_BPL') as BPL_Objective, XPath(x,'row/
Project_Objectives_IHHL_TOTAL') as BPL_Objective_total;
C = filter B by (((int)BPL_Objective * 100)/(int)BPL_Objective_total) >= 80;
STORE C INTO 'hdfs://quickstart.cloudera:8020/uscrime_analysis_2' USING
PigStorage (',');
dump C;
```

Execution:
Pig <pig_script_filename>

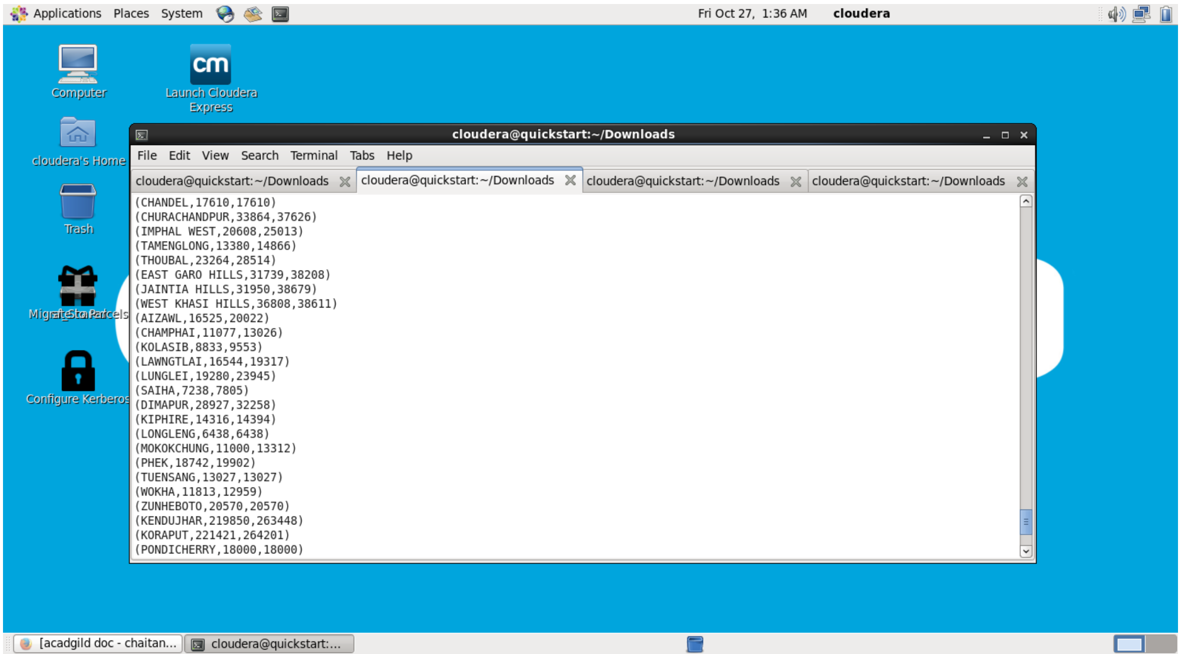
Output:



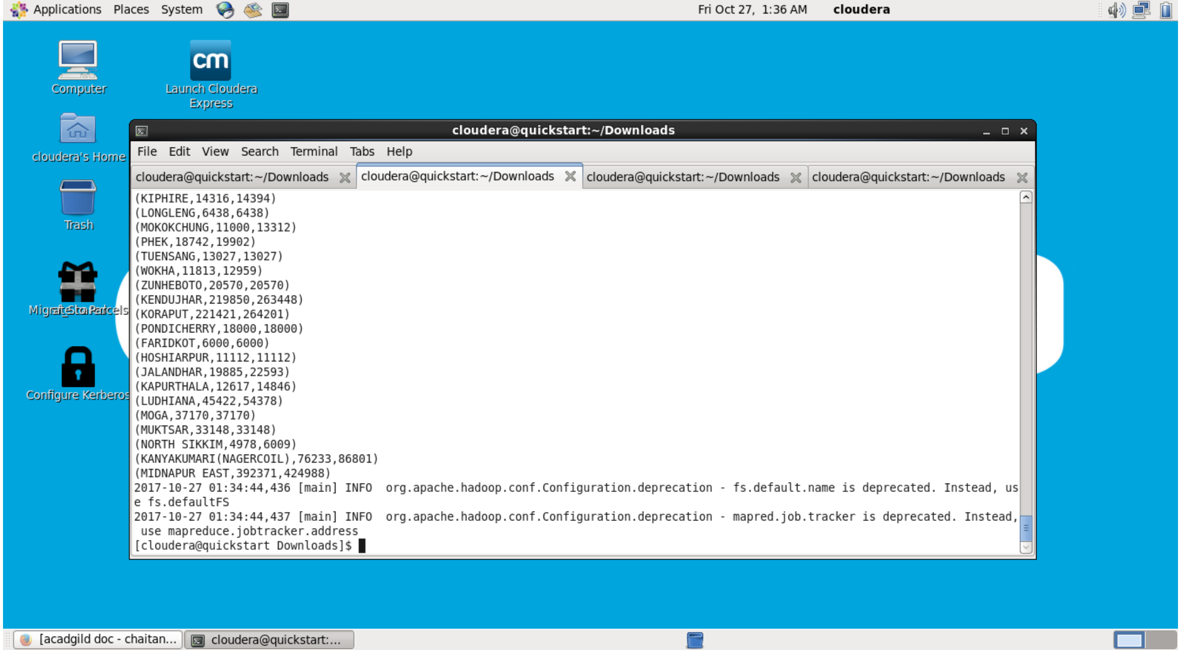
```
: 1
(KHAMMAM,189225,194085)
(MAHBUBNAGAR,418559,484591)
(NIZAMABAD,225519,234993)
(ANJAW,3232,3232)
(CHANGLANG,15716,18515)
(DIBANG VALLEY,1085,1085)
(EAST KAMENG,7016,8296)
(KURUNG KUMEY,22036,22036)
(LOHIT,8800,8800)
(LOWER SUBANSIRI,5766,6930)
(TAWANG,5478,6688)
(TIRAP,5780,6250)
(UPPER SUBANSIRI,9045,9506)
(WEST KAMENG,5644,5896)
(WEST SIANG,11472,11472)
(DHEMAJI,80854,99400)
(DHUBRI,205324,247376)
(HAILAKANDI,49837,62282)
(KARIMGANJ,78992,94401)
(NORTH CACHAR HILLS,21328,22430)
(BANKA,82439,82439)
(PASHCHIM CHAMPARAN,219434,230266)
(DADRA AND NAGAR HAVELI,2480,2480)
(NORTH GOA,15000,15000)
```



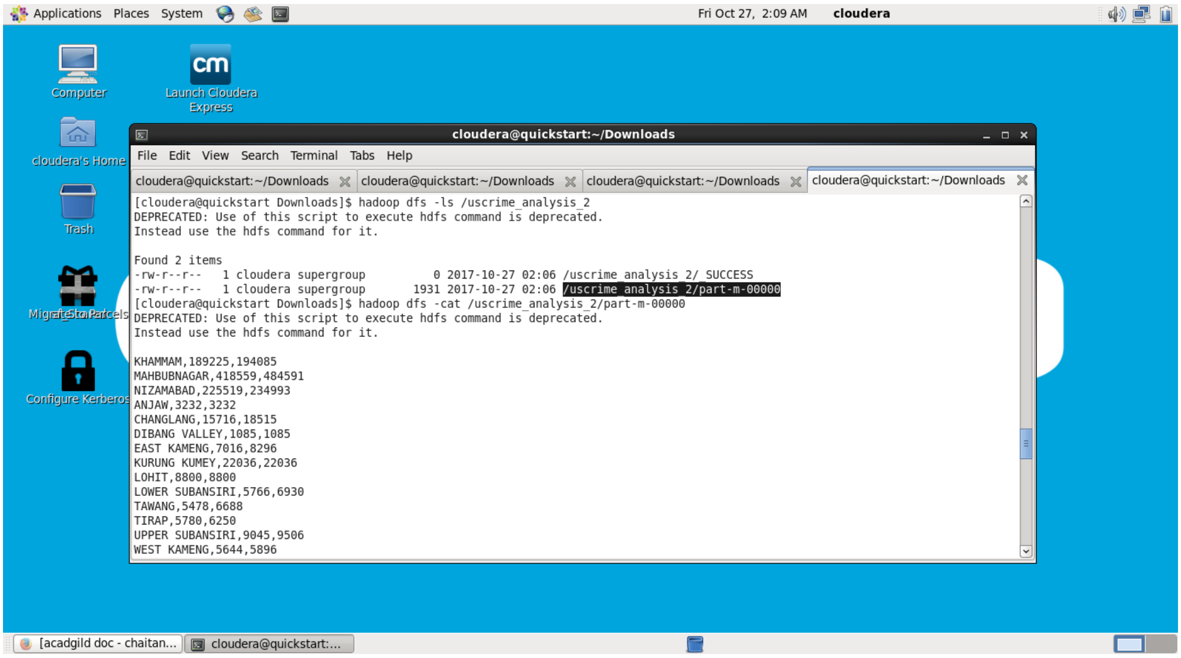
```
(NORTH GOA,15000,15000)
(ANANTNAG,75476,93537)
(KARGIL,8475,8475)
(KISHTWAR,22318,22318)
(LEH (LADAKH),6090,6090)
(REASI                                   ,21500,21500)
(SAMBA                                   ,9849,9849)
(SHOPIAN                                 ,10196,10196)
(LATEHAR,79635,94187)
(PASCHIM SINGHBHUM,170083,203522)
(RANCHI,197639,238675)
(SAHIBGANJ,112798,125282)
(ALAPPUZHA,114359,116548)
(IDUKKI,86535,98415)
(KANNUR,34121,34121)
(KOLLAM,95130,102420)
(KOTTAYAM,28118,29764)
(KOZHIKODE,42285,44371)
(MALAPPURAM,60592,64588)
(PALAKKAD,107018,131442)
(PATHANAMTHITTA,53799,54999)
(THIRUVANANTHAPURAM,129000,138482)
(THRISSUR,48669,54537)
(WAYANAD,50655,53136)
(CHANDEL,17610,17610)
```


Applications Places System
Fri Oct 27, 1:36 AM
cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate_to_Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

```
(CHANDEL,17610,17610)
(CHURACHANDPUR,33864,37626)
(IMPHAL WEST,20608,25013)
(TAMENGLONG,13380,14866)
(THOUBAL,23264,28514)
(EAST GARO HILLS,31739,38208)
(JAINTIA HILLS,31950,38679)
(WEST KHASI HILLS,36808,38611)
(AIZAWL,16525,20022)
(CHAMPHAI,11077,13026)
(KOLASIB,8833,9553)
(LAWNGTLAI,16544,19317)
(LUNGLEI,19280,23945)
(SAIHA,7238,7805)
(DIMAPUR,28927,32258)
(KIPHIRE,14316,14394)
(LONGLENG,6438,6438)
(MOKOKCHUNG,11000,13312)
(PHEK,18742,19902)
(TUENSANG,13027,13027)
(WOKHA,11813,12959)
(ZUNHEBOTO,20570,20570)
(KENDUJHAR,219850,263448)
(KORAPUT,221421,264201)
(PONDICHERRY,18000,18000)
```


Applications Places System
Fri Oct 27, 1:36 AM
cloudera
Computer
Launch Cloudera Express
cloudera's Home
Trash
Migrate_to_Parcels
Configure Kerberos
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
[acadgild doc - chaitan...
cloudera@quickstart:...

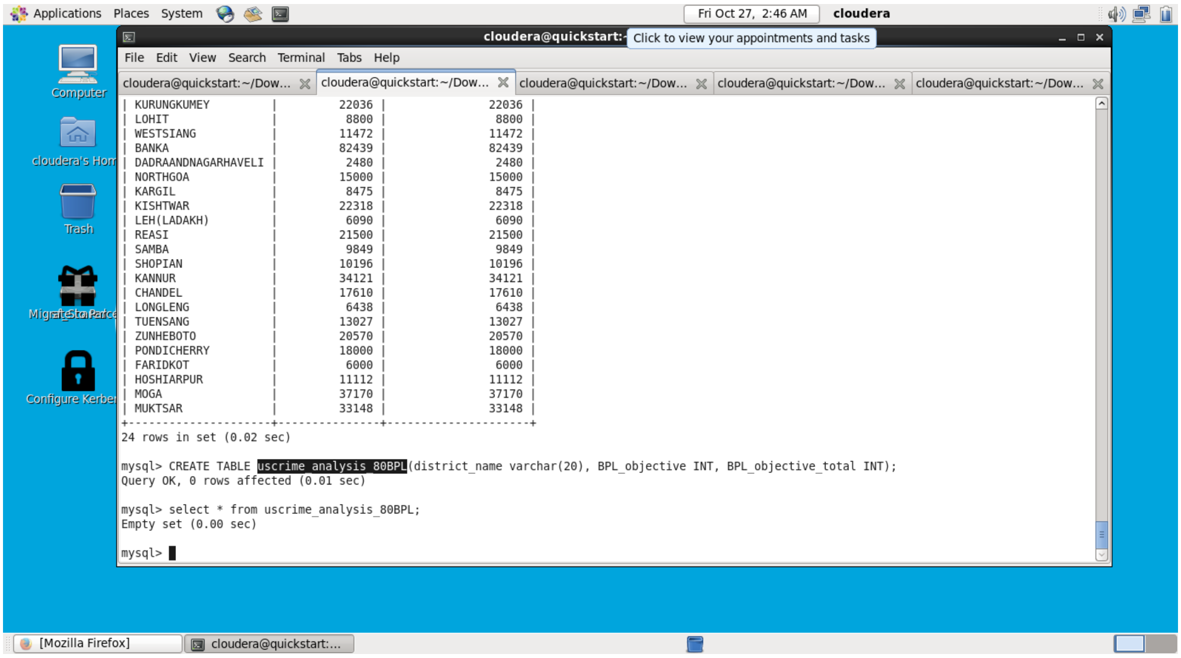
```
(KIPHIRE,14316,14394)
(LONGLENG,6438,6438)
(MOKOKCHUNG,11000,13312)
(PHEK,18742,19902)
(TUENSANG,13027,13027)
(WOKHA,11813,12959)
(ZUNHEBOTO,20570,20570)
(KENDUJHAR,219850,263448)
(KORAPUT,221421,264201)
(PONDICHERRY,18000,18000)
(FARIDKOT,6000,6000)
(HOSHIARPUR,11112,11112)
(JALANDHAR,19885,22593)
(KAPURTHALA,12617,14846)
(LUDHIANA,45422,54378)
(MOGA,37170,37170)
(MUKTSAR,33148,33148)
(NORTH SIKKIM,4978,6009)
(KANYAKUMARI(NAGERCOIL),76233,86801)
(MIDNAPUR EAST,392371,424988)
2017-10-27 01:34:44,436 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, us
e fs.defaultFS
2017-10-27 01:34:44,437 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead,
 use mapreduce.jobtracker.address
[cloudera@quickstart Downloads]$ 
```



```
[cloudera@quickstart Downloads]$ hadoop dfs -ls /uscrime_analysis_2
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2017-10-27 02:06 /uscrime_analysis_2/_SUCCESS
-rw-r--r--   1 cloudera supergroup       1931 2017-10-27 02:06 /uscrime_analysis_2/part-m-00000
[cloudera@quickstart Downloads]$ hadoop dfs -cat /uscrime_analysis_2/part-m-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

KHAMMAM,189225,194085
MAHBUBNAGAR,418559,484591
NIZAMABAD,225519,234993
ANJAW,3232,3232
CHANGLANG,15716,18515
DIBANG VALLEY,1085,1085
EAST KAMENG,7016,8296
KURUNG KUMEY,22036,22036
LOHIT,8800,8800
LOWER SUBANSIRI,5766,6930
TAWANG,5478,6688
TIRAP,5780,6250
UPPER SUBANSIRI,9045,9506
WEST KAMENG,5644,5896
```

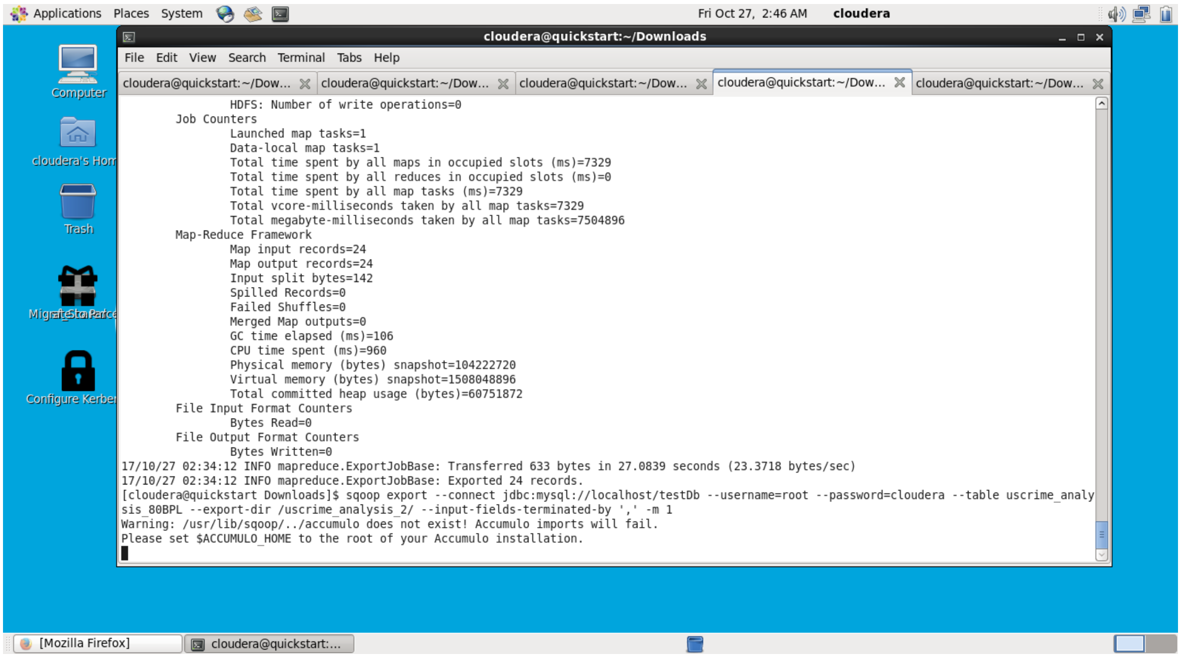
Export the results to mysql using sqoop:

1. Create table in mysql to store the data:



```
| KURUNGKUMEY         |        22036 |              22036 |
| LOHIT               |         8800 |               8800 |
| WESTSIANG           |        11472 |              11472 |
| BANKA               |        82439 |              82439 |
| DADRAANDNAGARHAVELI |         2480 |               2480 |
| NORTHGOA            |        15000 |              15000 |
| KARGIL              |         8475 |               8475 |
| KISHTWAR            |        22318 |              22318 |
| LEH(LADAKH)         |         6090 |               6090 |
| REASI               |        21500 |              21500 |
| SAMBA               |         9849 |               9849 |
| SHOPIAN             |        10196 |              10196 |
| KANNUR              |        34121 |              34121 |
| CHANDEL             |        17610 |              17610 |
| LONGLENG            |         6438 |               6438 |
| TUENSANG            |        13027 |              13027 |
| ZUNHEBOTO           |        20570 |              20570 |
| PONDICHERRY         |        18000 |              18000 |
| FARIDKOT            |         6000 |               6000 |
| HOSHIARPUR          |        11112 |              11112 |
| MOGA                |        37170 |              37170 |
| MUKTSAR             |        33148 |              33148 |
+---------------------+--------------+--------------------+
24 rows in set (0.02 sec)

mysql> CREATE TABLE uscrime_analysis_80BPL(district_name varchar(20), BPL_objective INT, BPL_objective_total INT);
Query OK, 0 rows affected (0.01 sec)

mysql> select * from uscrime_analysis_80BPL;
Empty set (0.00 sec)

mysql> 
```

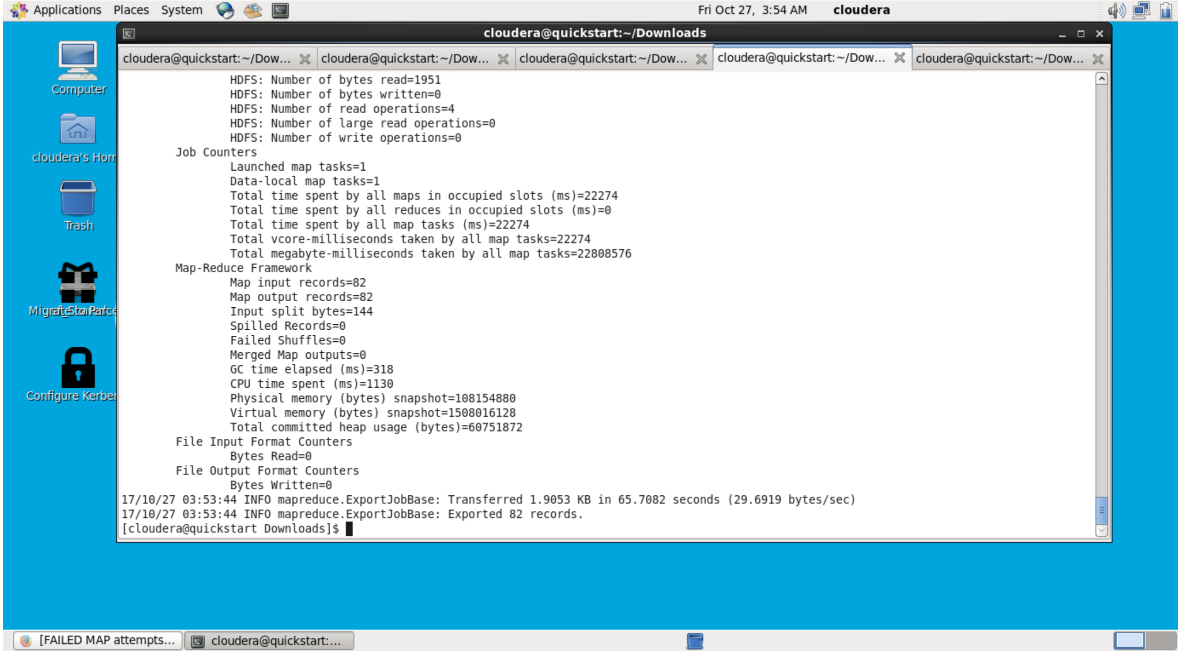
3. Using sqoop export data from HDFS to mysql using command:



Applications Places System
Fri Oct 27, 2:46 AM cloudera
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
Computer
cloudera's Hom
Trash
Migrate to Parce
Configure Kerber
[Mozilla Firefox]
cloudera@quickstart:...

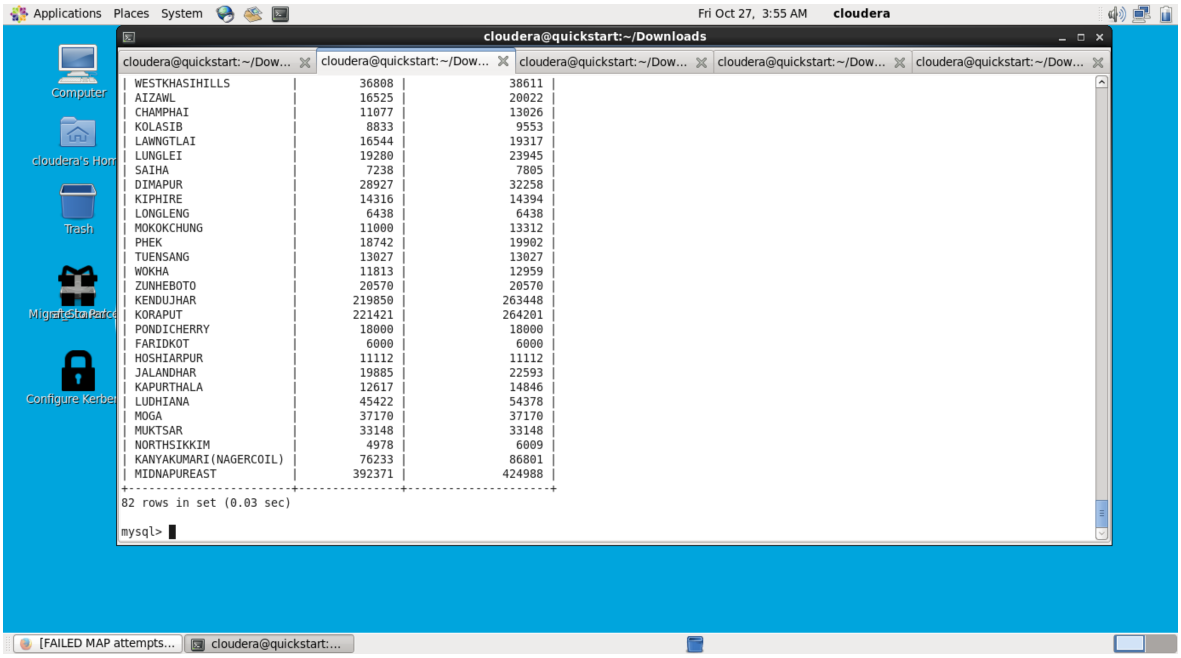
```
                HDFS: Number of write operations=0
        Job Counters 
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7329
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=7329
                Total vcore-milliseconds taken by all map tasks=7329
                Total megabyte-milliseconds taken by all map tasks=7504896
        Map-Reduce Framework
                Map input records=24
                Map output records=24
                Input split bytes=142
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=106
                CPU time spent (ms)=960
                Physical memory (bytes) snapshot=104222720
                Virtual memory (bytes) snapshot=1508048896
                Total committed heap usage (bytes)=60751872
        File Input Format Counters 
                Bytes Read=0
        File Output Format Counters 
                Bytes Written=0
17/10/27 02:34:12 INFO mapreduce.ExportJobBase: Transferred 633 bytes in 27.0839 seconds (23.3718 bytes/sec)
17/10/27 02:34:12 INFO mapreduce.ExportJobBase: Exported 24 records.
[cloudera@quickstart Downloads]$ sqoop export --connect jdbc:mysql://localhost/testDb --username=root --password=cloudera --table uscrime_analy
sis_80BPL --export-dir /uscrime_analysis_2/ --input-fields-terminated-by ',' -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```

3. Check the success message for the job completion:



```
                HDFS: Number of bytes read=1951
                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)=22274
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=22274
                Total vcore-milliseconds taken by all map tasks=22274
                Total megabyte-milliseconds taken by all map tasks=22808576
        Map-Reduce Framework
                Map input records=82
                Map output records=82
                Input split bytes=144
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=318
                CPU time spent (ms)=1130
                Physical memory (bytes) snapshot=108154880
                Virtual memory (bytes) snapshot=1508016128
                Total committed heap usage (bytes)=60751872
        File Input Format Counters 
                Bytes Read=0
        File Output Format Counters 
                Bytes Written=0
17/10/27 03:53:44 INFO mapreduce.ExportJobBase: Transferred 1.9053 KB in 65.7082 seconds (29.6919 bytes/sec)
17/10/27 03:53:44 INFO mapreduce.ExportJobBase: Exported 82 records.
[cloudera@quickstart Downloads]$ 
```

4. Check the data in mysql table using command; 82 rows copied successfully.



Applications
Places
System
Fri Oct 27, 3:55 AM
cloudera
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Dow...
cloudera@quickstart:~/Dow...
cloudera@quickstart:~/Dow...
cloudera@quickstart:~/Dow...
cloudera@quickstart:~/Dow...
Computer
cloudera's Hom
Trash
Migrate_to_Parce
Configure Kerber
[FAILED MAP attempts...
cloudera@quickstart:...

```
| WESTKHASIHILLS        |         36808 |               38611 |
| AIZAWL                |         16525 |               20022 |
| CHAMPHAI              |         11077 |               13026 |
| KOLASIB               |          8833 |                9553 |
| LAWNGTLAI             |         16544 |               19317 |
| LUNGLEI               |         19280 |               23945 |
| SAIHA                 |          7238 |                7805 |
| DIMAPUR               |         28927 |               32258 |
| KIPHIRE               |         14316 |               14394 |
| LONGLENG              |          6438 |                6438 |
| MOKOKCHUNG            |         11000 |               13312 |
| PHEK                  |         18742 |               19902 |
| TUENSANG              |         13027 |               13027 |
| WOKHA                 |         11813 |               12959 |
| ZUNHEBOTO             |         20570 |               20570 |
| KENDUJHAR             |        219850 |              263448 |
| KORAPUT               |        221421 |              264201 |
| PONDICHERRY           |         18000 |               18000 |
| FARIDKOT              |          6000 |                6000 |
| HOSHIARPUR            |         11112 |               11112 |
| JALANDHAR             |         19885 |               22593 |
| KAPURTHALA            |         12617 |               14846 |
| LUDHIANA              |         45422 |               54378 |
| MOGA                  |         37170 |               37170 |
| MUKTSAR               |         33148 |               33148 |
| NORTHSIKKIM           |          4978 |                6009 |
| KANYAKUMARI(NAGERCOIL) |         76233 |               86801 |
| MIDNAPUREAST          |        392371 |              424988 |
+-----------------------+---------------+---------------------+
82 rows in set (0.03 sec)

mysql> 
```