Assignment 10.2

Problem Statement

Implement the concept given in below blog link and share the complete steps along with screenshots.

https://acadgild.com/blog/loading-data-into-hbase-using-pig-scripts/

To implement the concepts discussed, user is expected to have a Hadoop cluster with Pig and HBase running on it.

Transfer into hdfs

We will be copying the data set in to HDFS which will be further loaded into HBase.

hadoop fs -put AcadgildStudent /

We will be including few jar files of HBase to the Pig classpath.

PIG_CLASSPATH=/usr/local/hbase/lib/hbase-common-0.98.14-hadoop2.jar:/usr/local/hbase/lib/hbase-client-0.98.14-

hadoop2.jar:/usr/local/hbase/lib/hbase-server-0.98.14-
hadoop2.jar:/usr/local/hbase/lib/hbase-protocol-0.98.14-
hadoop2.jar:/usr/local/hbase/lib/htrace-core-2.04.jar:/usr/local/hbase/lib/zookeeper-
3.4.6.jar:/usr/local/hbase/lib/guava-12.0.1.jar

```
[acadgild@localhost Downloads]$ PIG_CLASSPATH=/home/hadoop/HADOOP/hbase-0.98.4-hadoop2/lib/hbase-server-0.98.4-hadoop2:/home/hadoop/HADOO
P/hbase-0.98.4-hadoop2/lib/hbase-*.jar;
[acadgild@localhost Downloads]$ hbase shell
2017-12-18 09:02:48,853 INFO  [main] Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 0.98.14-hadoop2, r4e4aabb93b52f1b0fef6b66edd06ec8923014dec, Tue Aug 25 22:35:44 PDT 2015
```

We will now start HBase shell and create a table. We only need this table as
skeleton so PIG can Store data inside this by referring the table name.

```
hbase(main):010:0> create 'studentAcad','student_data'
0 row(s) in 2.5910 seconds

=> Hbase::Table - studentAcad
hbase(main):011:0>
```

We can come out from HBase by typing exit and switch to PIG grunt shell.Once we
are inside PIG mode we can load data from HDFS to Alias relation and then we can
transfer the data inside HBase by STORE command.

We need to ensure that we give the correct name for table name created inside
HBase. Also the parameters should be kept in mind to avoid mistake.

```
STORE rawD INTO 'hbase://studentAcad' USING
org.apache.pig.backend.hadoop.hbase.HBaseStorage(
'student_data:StudentName,student_data:sector,student_data:DOB,
student_data:qalification,student_data:score,
student_data:state,student_data:randomName');
```

```
grunt>
grunt> rawD = load '/student' using PigStorage (',') as (StudentName:chararray,sector:chararray,DOB:chararray,qualification:chararray,sco
re:int,state:chararray,randomName:chararray);
2017-12-18 09:35:28,630 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
2017-12-18 09:35:28,630 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
ad, use mapreduce.job.counters.max
2017-12-18 09:35:28,632 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2017-12-18 09:35:28,916 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
ad, use mapreduce.job.counters.max
2017-12-18 09:35:28,917 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2017-12-18 09:35:28,917 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> store rawD into 'hbase://studentAcad' using org.apache.pig.backend.hadoop.hbase.HBaseStorage('student_data:StudentName,student_dat
a:sector,student_data:DOB,student_data:qualification,student_data:score,student_data:state,student_data:randomName');
2017-12-18 09:36:29,210 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
ad, use mapreduce.job.counters.max
```

```
2017-12-18 09:37:18,455 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-18 09:37:18,461 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2017-12-18 09:37:18,579 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-18 09:37:18,585 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2017-12-18 09:37:18,675 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2017-12-18 09:37:18,686 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationSta
tus=SUCCEEDED. Redirecting to job history server
2017-12-18 09:37:18,796 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve
job to compute warning aggregation.
2017-12-18 09:37:18,796 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

The result can be displayed through scan command followed by table name inside quotes(' ').
scan 'studentAcad'

```
hbase(main):011:0>
hbase(main):011:0> scan 'studentAcad'
ROW                           COLUMN+CELL
 ABEDNIGO                     column=student_data:DOB, timestamp=1513570021894, value=BBA
 ABEDNIGO                     column=student_data:StudentName, timestamp=1513570021894, value=goverenment
 ABEDNIGO                     column=student_data:qualification, timestamp=1513570021894, value=100
 ABEDNIGO                     column=student_data:score, timestamp=1513570021894, value=alabama
 ABEDNIGO                     column=student_data:sector, timestamp=1513570021894, value=20-10-2000
 ABEDNIGO                     column=student_data:state, timestamp=1513570021894, value=madison`
 ABROSER                      column=student_data:DOB, timestamp=1513570021763, value=MBBS
 ABROSER                      column=student_data:StudentName, timestamp=1513570021763, value=goverenment
 ABROSER                      column=student_data:qualification, timestamp=1513570021763, value=3
 ABROSER                      column=student_data:score, timestamp=1513570021763, value=Pennsylvania
 ABROSER                      column=student_data:sector, timestamp=1513570021763, value=18-11-2002
 ABROSER                      column=student_data:state, timestamp=1513570021763, value=prattville*
 AGNES                        column=student_data:DOB, timestamp=1513570021894, value=BE
 AGNES                        column=student_data:StudentName, timestamp=1513570021894, value=goverenment
 AGNES                        column=student_data:qualification, timestamp=1513570021894, value=100
 AGNES                        column=student_data:score, timestamp=1513570021894, value=alabama
 AGNES                        column=student_data:sector, timestamp=1513570021894, value=20-10-2000
 AGNES                        column=student_data:state, timestamp=1513570021894, value=madison`
 AGNEW                        column=student_data:DOB, timestamp=1513570021764, value=BCOM
 AGNEW                        column=student_data:StudentName, timestamp=1513570021764, value=goverenment
 AGNEW                        column=student_data:qualification, timestamp=1513570021764, value=7
 AGNEW                        column=student_data:score, timestamp=1513570021764, value=california
 AGNEW                        column=student_data:sector, timestamp=1513570021764, value=20-10-2000
 AGNEW                        column=student_data:state, timestamp=1513570021764, value=dothan@
 ALEXANDER                    column=student_data:DOB, timestamp=1513570021895, value=MBBS
 ALEXANDER                    column=student_data:StudentName, timestamp=1513570021895, value=goverenment
 ALEXANDER                    column=student_data:qualification, timestamp=1513570021895, value=100
 ALEXANDER                    column=student_data:score, timestamp=1513570021895, value=alabama
 ALEXANDER                    column=student_data:sector, timestamp=1513570021895, value=20-10-2000
 ALEXANDER                    column=student_data:state, timestamp=1513570021895, value=madison`
 ANDREW                       column=student_data:DOB, timestamp=1513570021895, value=BE
 ANDREW                       column=student_data:StudentName, timestamp=1513570021895, value=goverenment
 ANDREW                       column=student_data:qualification, timestamp=1513570021895, value=100
 ANDREW                       column=student_data:score, timestamp=1513570021895, value=alabama
 ANDREW                       column=student_data:sector, timestamp=1513570021895, value=20-10-2000
```