

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/>

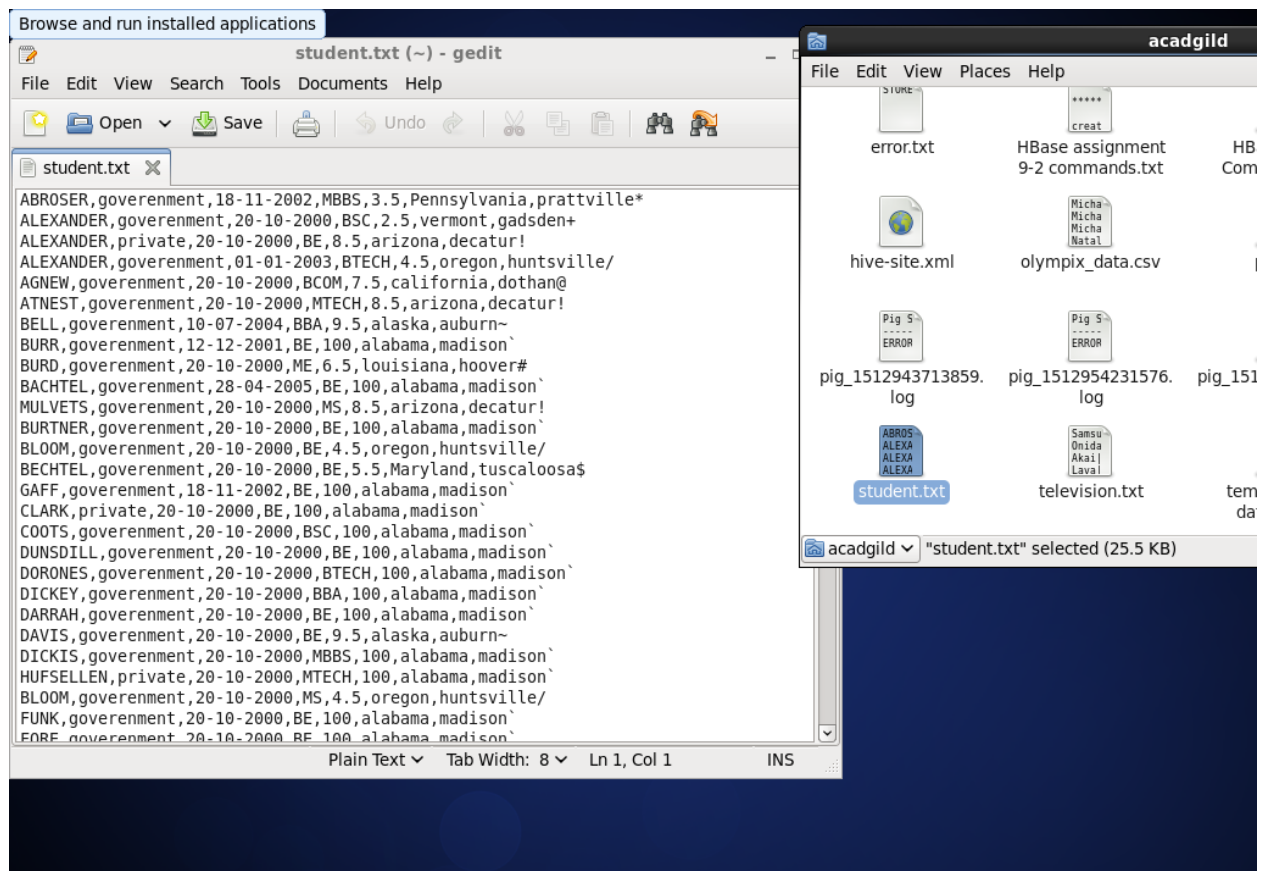
Solution:

Loading Data into HBase Using PIG Scripts:

We are taking sample data set of student which will be loaded into HBase. We have attached snapshot with every step for better understanding.

Please refer the description for the above data set containing seven columns named as:

StudentName, sector, DOB, qualification, score, state, randomName.



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

```
[acadgild@localhost ~]$ hadoop dfs -put student.txt /
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

17/12/26 12:25:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop dfs -ls /
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

17/12/26 12:25:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 6 items
drwxr-xr-x - acadgild supergroup          0 2017-12-26 10:48 /hbasestorage
drwxr-xr-x - acadgild supergroup          0 2017-12-09 17:01 /nig data
-rw-r--r-- 1 acadgild supergroup    26204 2017-12-26 12:25 /student.txt
drwxrwxr-x - acadgild supergroup          0 2015-11-05 13:46 /tmp
drwxr-xr-x - acadgild supergroup          0 2015-11-17 01:56 /user
drwxr-xr-x - acadgild supergroup          0 2015-11-05 12:56 /zookeeper
[acadgild@localhost ~]$
```

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

PIG_CLASSPATH=/home/hadoop/HADOOP/hbase-0.98.4-hadoop2/lib/hbase-server-0.98.4-hadoop2:/home/hadoop/HADOOP/hbase-0.98.4-hadoop2/lib/hbase-*.jar;

```
[acadgild@localhost lib]$ PIG_CLASSPATH=/usr/local/hbase/lib/hbase-server-0.98.14-hadoop2.jar:/usr/local/hbase/lib/hbase-*.jar;
[acadgild@localhost lib]$ echo $PIG_CLASSPATH
/usr/local/hbase/lib/hbase-server-0.98.14-hadoop2.jar:/usr/local/hbase/lib/hbase-*.jar
[acadgild@localhost lib]$
```

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.

```
/usr/local/hbase/lib/hbase-server-0.98.14-hadoop2:/usr/local/hbase/lib/hbase-*.jar
[acadgild@localhost lib]$ hbase shell
2018-01-22 21:55:31,792 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

hbase(main):001:0> create 'studentAcad','student data'
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2018-01-22 21:56:46,323 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
0 row(s) in 4.9220 seconds

=> Hbase::Table - studentAcad
hbase(main):002:0>
```

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.

```
2018-01-22 21:58:46,336 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> rawD = LOAD '/student.txt' USING PigStorage(',') AS (StudentName:chararray, sector:chararray, DOB:chararray, qualification:chararray, score:int, state:chararray, randomName:chararray);
2018-01-22 22:04:46,625 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.hours is deprecated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2018-01-22 22:04:46,625 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is deprecated. Instead, use mapreduce.jobtracker.heartbeats.in.second
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion.poll.interval is deprecated. Instead, use mapreduce.client.completion.pollinterval
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.tasks.sleeptimebefore-sigkill is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebefore-sigkill
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.address
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.replication.considerLoad is deprecated. Instead, use dfs.namenode.replication.considerLoad
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.block.size is deprecated. Instead, use dfs.blocksize
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.permissions is deprecated. Instead, use dfs.permissions.enabled
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - topology.node.switch.mapping.impl is deprecated. Instead, use net.topology.node.switch.mapping.impl
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.access.time.precision is deprecated. Instead, use dfs.namenode.access.time.precision
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.skip.map.max.skip.records is deprecated. Instead, use mapreduce.map.skip.maxrecords
2018-01-22 22:04:46,629 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.submit.replication is deprecated. Instead, use mapreduce.client.submit.file.replication
```

```
2018-01-22 22:04:46,652 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.userlog.retain.hours is deprecated. Instead, use mapreduce.job.userlog.retain.hours
2018-01-22 22:04:46,652 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.max.objects is deprecated. Instead, use dfs.namenode.max.objects
2018-01-22 22:04:46,652 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.name.edits.dir is deprecated. Instead, use dfs.namenode.edits.dir
2018-01-22 22:04:46,652 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - dfs.replication.interval is deprecated. Instead, use dfs.namenode.replication.interval
2018-01-22 22:04:46,652 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.local.dir.minspacekill is deprecated. Instead, use mapreduce.tasktracker.local.dir.minspacekill
grunt> DESCRIBE rawD;
rawD: {StudentName: chararray,sector: chararray,DOB: chararray,qualification: chararray,score: int,state: chararray,randomName: chararray}
grunt>
```

Now we can transfer the data inside HBase by STORE command.

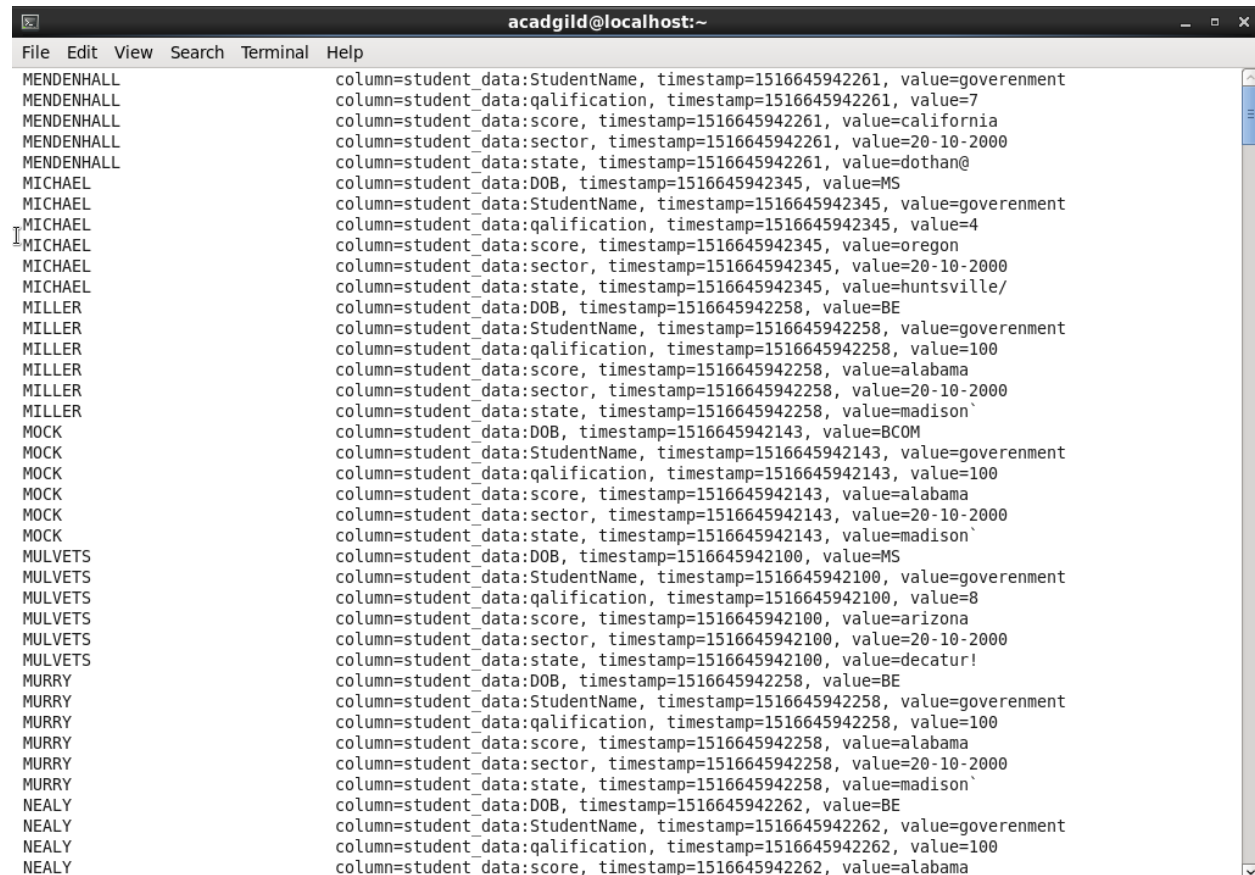
We need to ensure that we give the correct name for table name created inside HBase. Also the parameters should be kept in mind to avoid mistake.

```
grunt> 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');
2018-01-22 22:30:48,538 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2018-01-22 22:30:48,538 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-01-22 22:30:48,538 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-01-22 22:30:49,100 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-01-22 22:30:49,101 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2018-01-22 22:30:49,101 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-01-22 22:30:49,501 [main] INFO org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper - Process identifier=hconnection-0x56913163 connecting to ZooKeeper ensemble=localhost:2181
2018-01-22 22:30:49,534 [main] INFO org.apache.zookeeper.ZooKeeper - Client environment:zookeeper.version=3.4.5-1392090, build.version=3.4.5-1392090
```

Once the success message comes as shown below, it is confirmed our data is loaded inside HBase.

```
2018-01-23 00:31:28,852 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2018-01-23 00:31:28,859 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-01-23 00:31:29,350 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2018-01-23 00:31:29,592 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-01-23 00:31:29,798 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2018-01-23 00:31:29,822 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalAp
plicationStatus=SUCCEEDED. Redirecting to job history server
2018-01-23 00:31:30,264 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable
to retrieve job to compute warning aggregation.
2018-01-23 00:31:30,330 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-01-23 00:31:30,726 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-01-23 00:31:30,730 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
```

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



```
acadgild@localhost:~
File Edit View Search Terminal Help
MENDENHALL column=student_data:StudentName, timestamp=1516645942261, value=governement
MENDENHALL column=student_data:qalification, timestamp=1516645942261, value=7
MENDENHALL column=student_data:score, timestamp=1516645942261, value=california
MENDENHALL column=student_data:sector, timestamp=1516645942261, value=20-10-2000
MENDENHALL column=student_data:state, timestamp=1516645942261, value=dothan@
MICHAEL column=student_data:DOB, timestamp=1516645942345, value=MS
MICHAEL column=student_data:StudentName, timestamp=1516645942345, value=governement
MICHAEL column=student_data:qalification, timestamp=1516645942345, value=4
MICHAEL column=student_data:score, timestamp=1516645942345, value=oregon
MICHAEL column=student_data:sector, timestamp=1516645942345, value=20-10-2000
MICHAEL column=student_data:state, timestamp=1516645942345, value=huntsville/
MILLER column=student_data:DOB, timestamp=1516645942258, value=BE
MILLER column=student_data:StudentName, timestamp=1516645942258, value=governement
MILLER column=student_data:qalification, timestamp=1516645942258, value=100
MILLER column=student_data:score, timestamp=1516645942258, value=alabama
MILLER column=student_data:sector, timestamp=1516645942258, value=20-10-2000
MILLER column=student_data:state, timestamp=1516645942258, value=madison`
MOCK column=student_data:DOB, timestamp=1516645942143, value=BCOM
MOCK column=student_data:StudentName, timestamp=1516645942143, value=governement
MOCK column=student_data:qalification, timestamp=1516645942143, value=100
MOCK column=student_data:score, timestamp=1516645942143, value=alabama
MOCK column=student_data:sector, timestamp=1516645942143, value=20-10-2000
MOCK column=student_data:state, timestamp=1516645942143, value=madison`
MULVETS column=student_data:DOB, timestamp=1516645942100, value=MS
MULVETS column=student_data:StudentName, timestamp=1516645942100, value=governement
MULVETS column=student_data:qalification, timestamp=1516645942100, value=8
MULVETS column=student_data:score, timestamp=1516645942100, value=arizona
MULVETS column=student_data:sector, timestamp=1516645942100, value=20-10-2000
MULVETS column=student_data:state, timestamp=1516645942100, value=decatur!
MURRY column=student_data:DOB, timestamp=1516645942258, value=BE
MURRY column=student_data:StudentName, timestamp=1516645942258, value=governement
MURRY column=student_data:qalification, timestamp=1516645942258, value=100
MURRY column=student_data:score, timestamp=1516645942258, value=alabama
MURRY column=student_data:sector, timestamp=1516645942258, value=20-10-2000
MURRY column=student_data:state, timestamp=1516645942258, value=madison`
NEALY column=student_data:DOB, timestamp=1516645942262, value=BE
NEALY column=student_data:StudentName, timestamp=1516645942262, value=governement
NEALY column=student_data:qalification, timestamp=1516645942262, value=100
NEALY column=student_data:score, timestamp=1516645942262, value=alabama
```