Task- To Load data into HBASE using PIG scripts

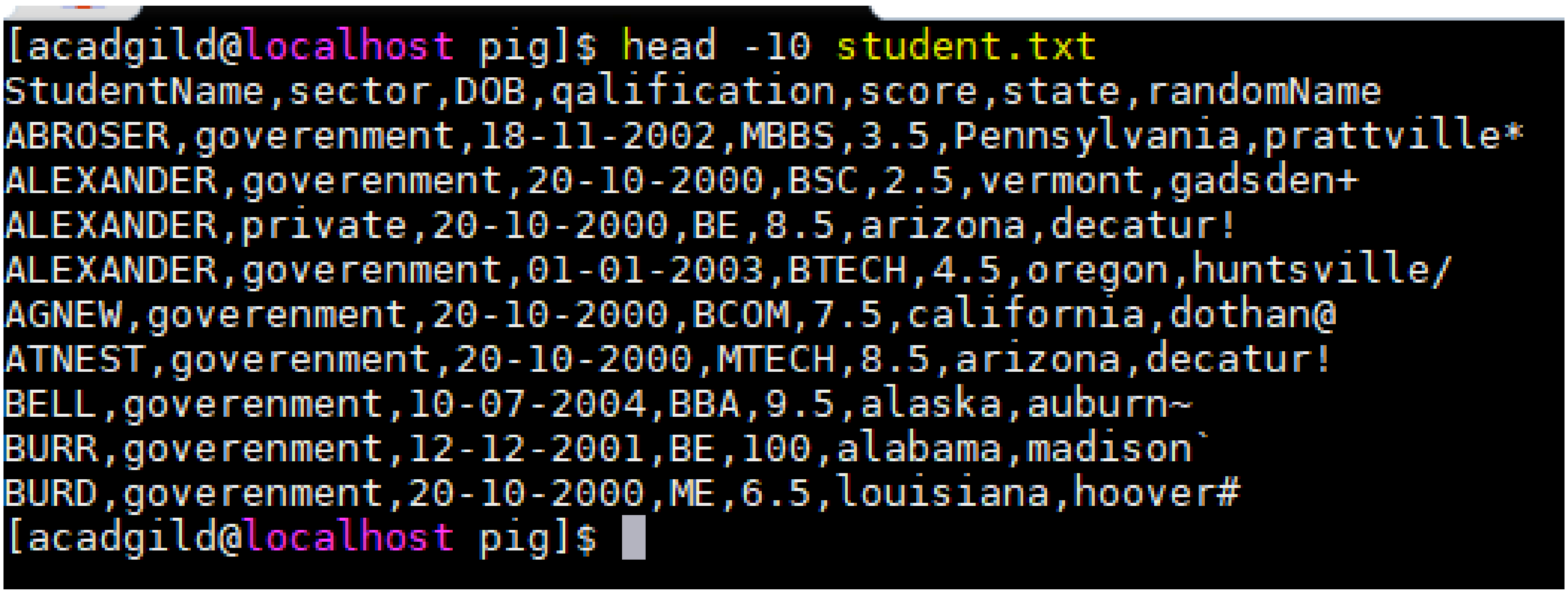
In this use case we have to load data into HBase table using Pig Script

We have implmented this via following below steps-

Step1. We have one input file named student.txt. We will load this file in local directory-

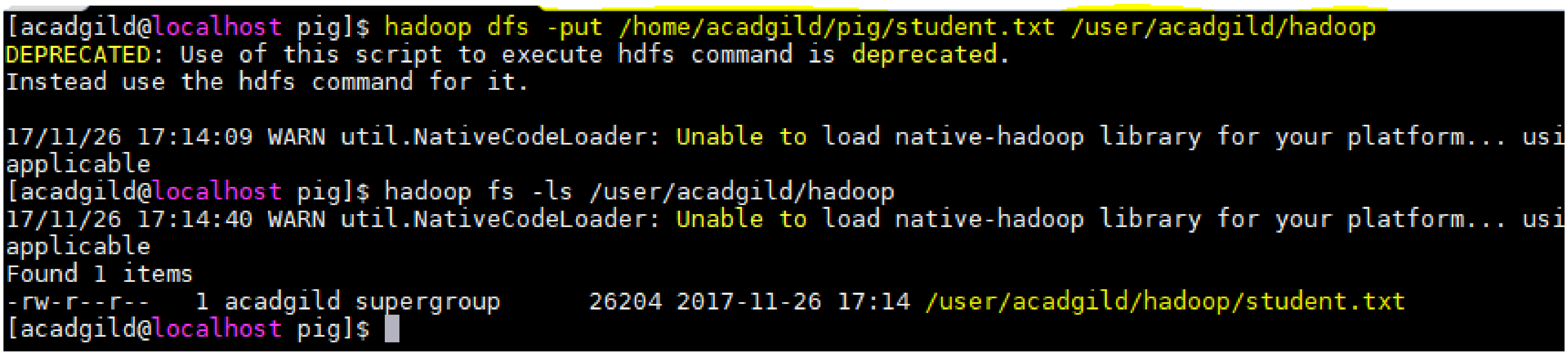
The structure of this file is- **StudentName, sector, DOB, qualification, score, state, randomName**.

Some contents of this file are shown below-



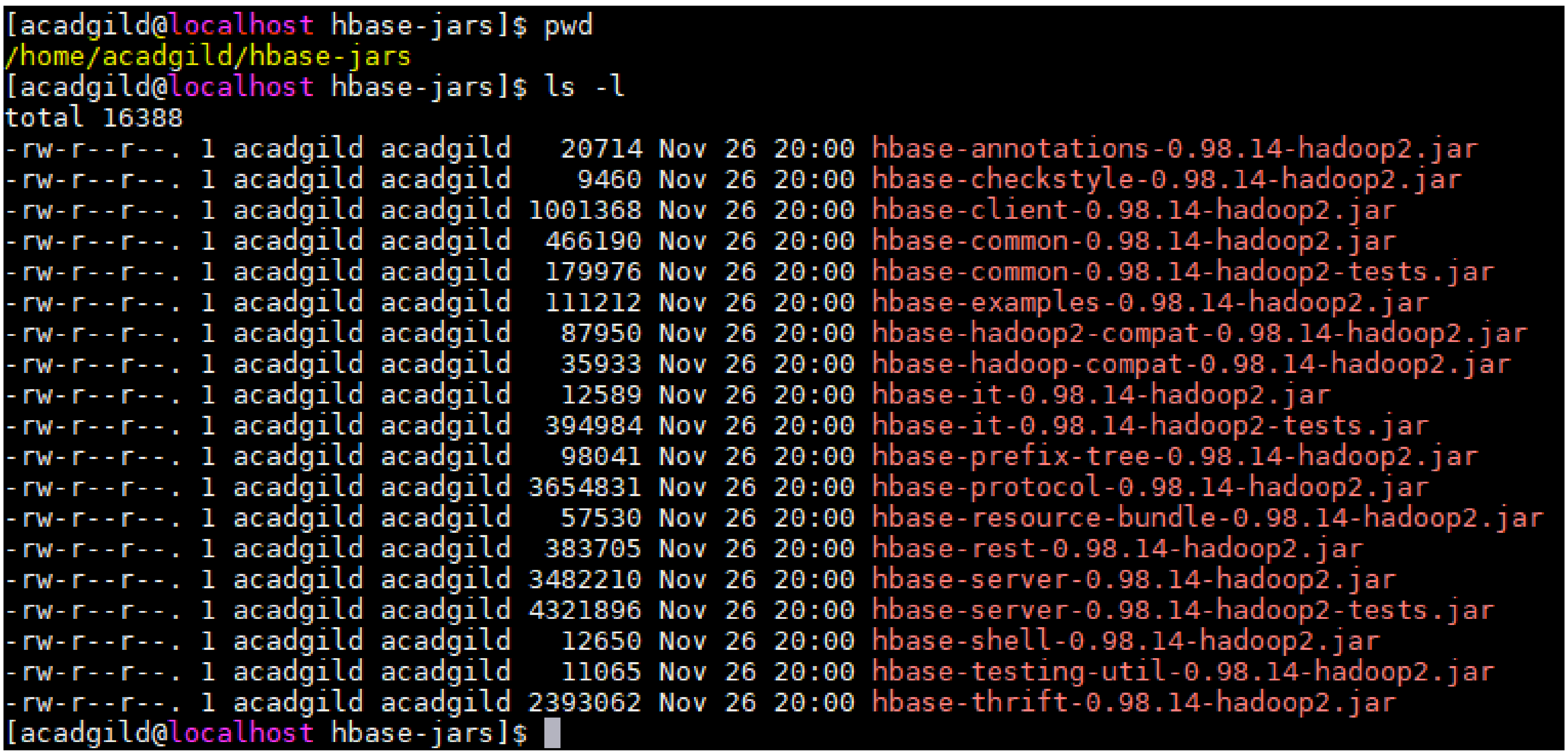
```
[acadgild@localhost pig]$ head -10 student.txt
StudentName,sector,DOB,qalification,score,state,randomName
ABROSER,goverenment,18-11-2002,MBBS,3.5,Pennsylvania,prattville*
ALEXANDER,goverenment,20-10-2000,BSC,2.5,vermont,gadsden+
ALEXANDER,private,20-10-2000,BE,8.5,arizona,decatur!
ALEXANDER,goverenment,01-01-2003,BTECH,4.5,oregon,huntsville/
AGNEW,goverenment,20-10-2000,BCOM,7.5,california,dothan@
ATNEST,goverenment,20-10-2000,MTECH,8.5,arizona,decatur!
BELL,goverenment,10-07-2004,BBA,9.5,alaska,auburn~
BURR,goverenment,12-12-2001,BE,100,alabama,madison`
BURD,goverenment,20-10-2000,ME,6.5,louisiana,hoover#
[acadgild@localhost pig]$
```

Step2- We will be copying the data set in to HDFS which will be further loaded into HBase using put command as shown below-



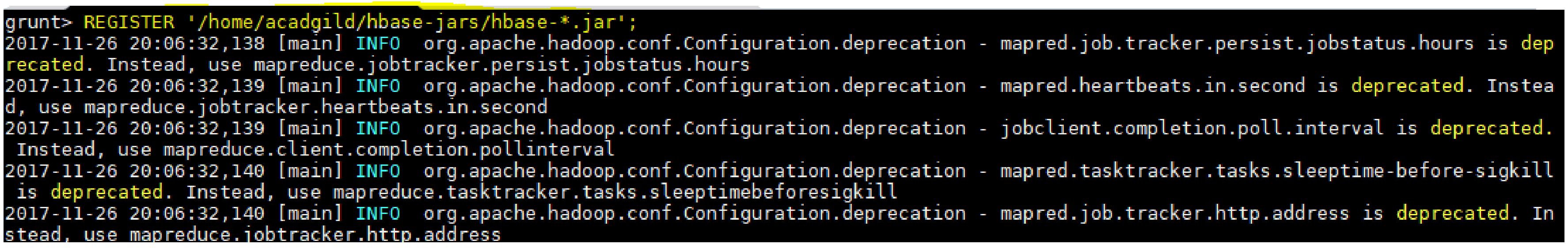
```
[acadgild@localhost pig]$ hadoop dfs -put /home/acadgild/pig/student.txt /user/acadgild/hadoop
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

17/11/26 17:14:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usi
applicable
[acadgild@localhost pig]$ hadoop fs -ls /user/acadgild/hadoop
17/11/26 17:14:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usi
applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup      26204 2017-11-26 17:14 /user/acadgild/hadoop/student.txt
[acadgild@localhost pig]$
```

Step3-Now in order to make pig communicate with HBase we have to register some jars. So we will include below jars shown-



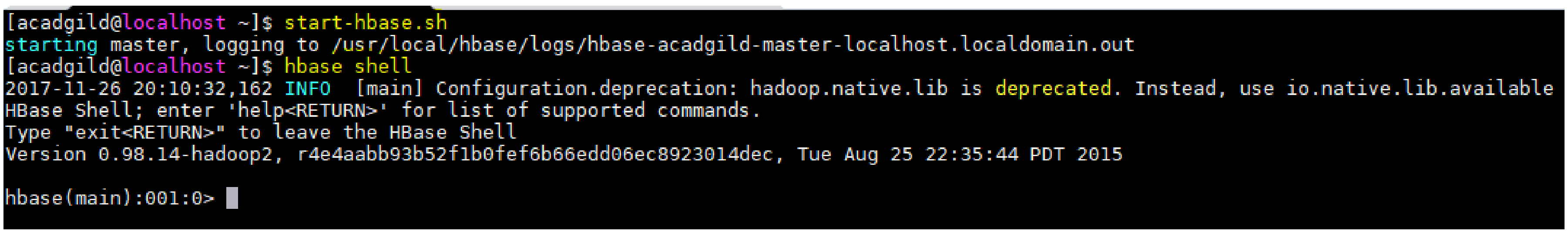
```
[acadgild@localhost hbase-jars]$ pwd
/home/acadgild/hbase-jars
[acadgild@localhost hbase-jars]$ ls -l
total 16388
-rw-r--r--. 1 acadgild acadgild   20714 Nov 26 20:00 hbase-annotations-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild    9460 Nov 26 20:00 hbase-checkstyle-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild 1001368 Nov 26 20:00 hbase-client-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild  466190 Nov 26 20:00 hbase-common-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild  179976 Nov 26 20:00 hbase-common-0.98.14-hadoop2-tests.jar
-rw-r--r--. 1 acadgild acadgild  111212 Nov 26 20:00 hbase-examples-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild   87950 Nov 26 20:00 hbase-hadoop2-compat-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild   35933 Nov 26 20:00 hbase-hadoop-compat-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild   12589 Nov 26 20:00 hbase-it-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild  394984 Nov 26 20:00 hbase-it-0.98.14-hadoop2-tests.jar
-rw-r--r--. 1 acadgild acadgild   98041 Nov 26 20:00 hbase-prefix-tree-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild 3654831 Nov 26 20:00 hbase-protocol-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild   57530 Nov 26 20:00 hbase-resource-bundle-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild  383705 Nov 26 20:00 hbase-rest-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild 3482210 Nov 26 20:00 hbase-server-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild 4321896 Nov 26 20:00 hbase-server-0.98.14-hadoop2-tests.jar
-rw-r--r--. 1 acadgild acadgild   12650 Nov 26 20:00 hbase-shell-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild   11065 Nov 26 20:00 hbase-testing-util-0.98.14-hadoop2.jar
-rw-r--r--. 1 acadgild acadgild 2393062 Nov 26 20:00 hbase-thrift-0.98.14-hadoop2.jar
[acadgild@localhost hbase-jars]$
```

We will REGISTER all jars shown in above diagram using below command-



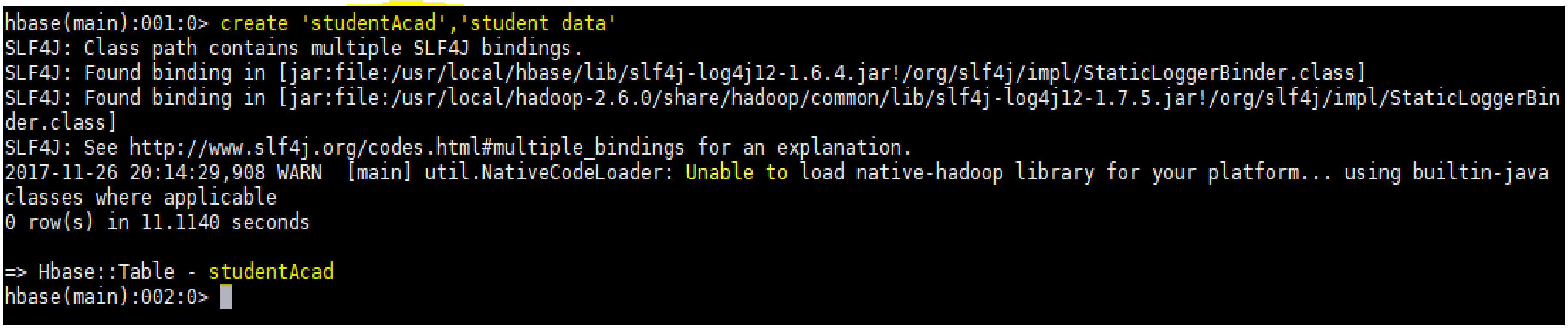
```
grunt> REGISTER '/home/acadgild/hbase-jars/hbase-*.jar';
2017-11-26 20:06:32,138 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.persist.jobstatus.hours is dep
recated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2017-11-26 20:06:32,139 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in.second is deprecated. Instea
d, use mapreduce.jobtracker.heartbeats.in.second
2017-11-26 20:06:32,139 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion.poll.interval is deprecated.
 Instead, use mapreduce.client.completion.pollinterval
2017-11-26 20:06:32,140 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.tasks.sleeptime-before-sigkill
 is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebeforesigkill
2017-11-26 20:06:32,140 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. In
stead, use mapreduce.jobtracker.http.address
```

Step-4 Start HBase and go to hbase shell



```
[acadgild@localhost ~]$ start-hbase.sh
starting master, logging to /usr/local/hbase/logs/hbase-acadgild-master-localhost.localdomain.out
[acadgild@localhost ~]$ hbase shell
2017-11-26 20:10:32,162 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> 
```

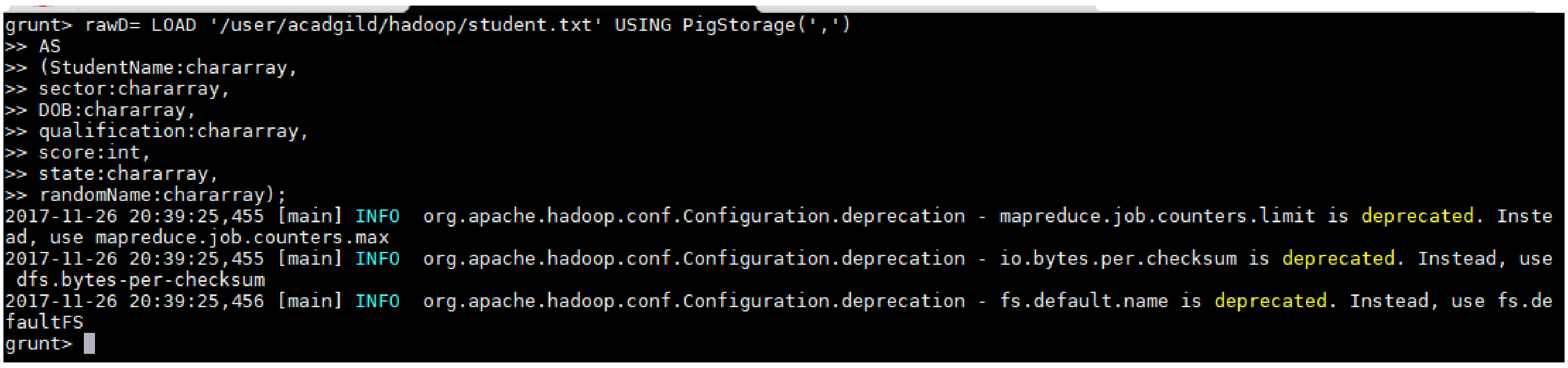
Step5- Inside hbase shell we will create a table named- '**studentAcad'** as shown below-



```
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/StaticLoggerBin
der.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-11-26 20:14:29,908 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
0 row(s) in 11.1140 seconds

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

Step6- Now we will go inside grunt shell and create a relation named rawD to load the student.txt file-

- rawD = LOAD '/home/acadgild/pig/student.txt' USING PigStorage(',')
- AS
- (
- StudentName:chararray,
- sector:chararray,
- DOB:chararray,
- qualification:chararray,
- score:int,
- state:chararray,
- randomName:chararray
-);

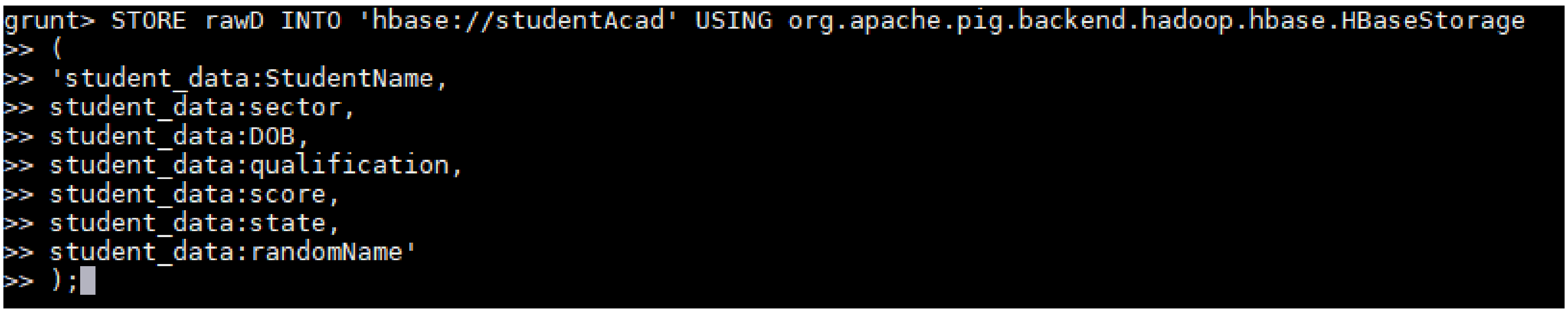


```
grunt> rawD= LOAD '/user/acadgild/hadoop/student.txt' USING PigStorage(',')
>> AS
>> (StudentName:chararray,
>> sector:chararray,
>> DOB:chararray,
>> qualification:chararray,
>> score:int,
>> state:chararray,
>> randomName:chararray);
2017-11-26 20:39:25,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Inste
ad, use mapreduce.job.counters.max
2017-11-26 20:39:25,455 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use
 dfs.bytes-per-checksum
2017-11-26 20:39:25,456 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.de
faultFS
grunt> 
```

Step7- We will create another relation to store the relation rawD into the table studentAcad we created inside HBase as shown below-

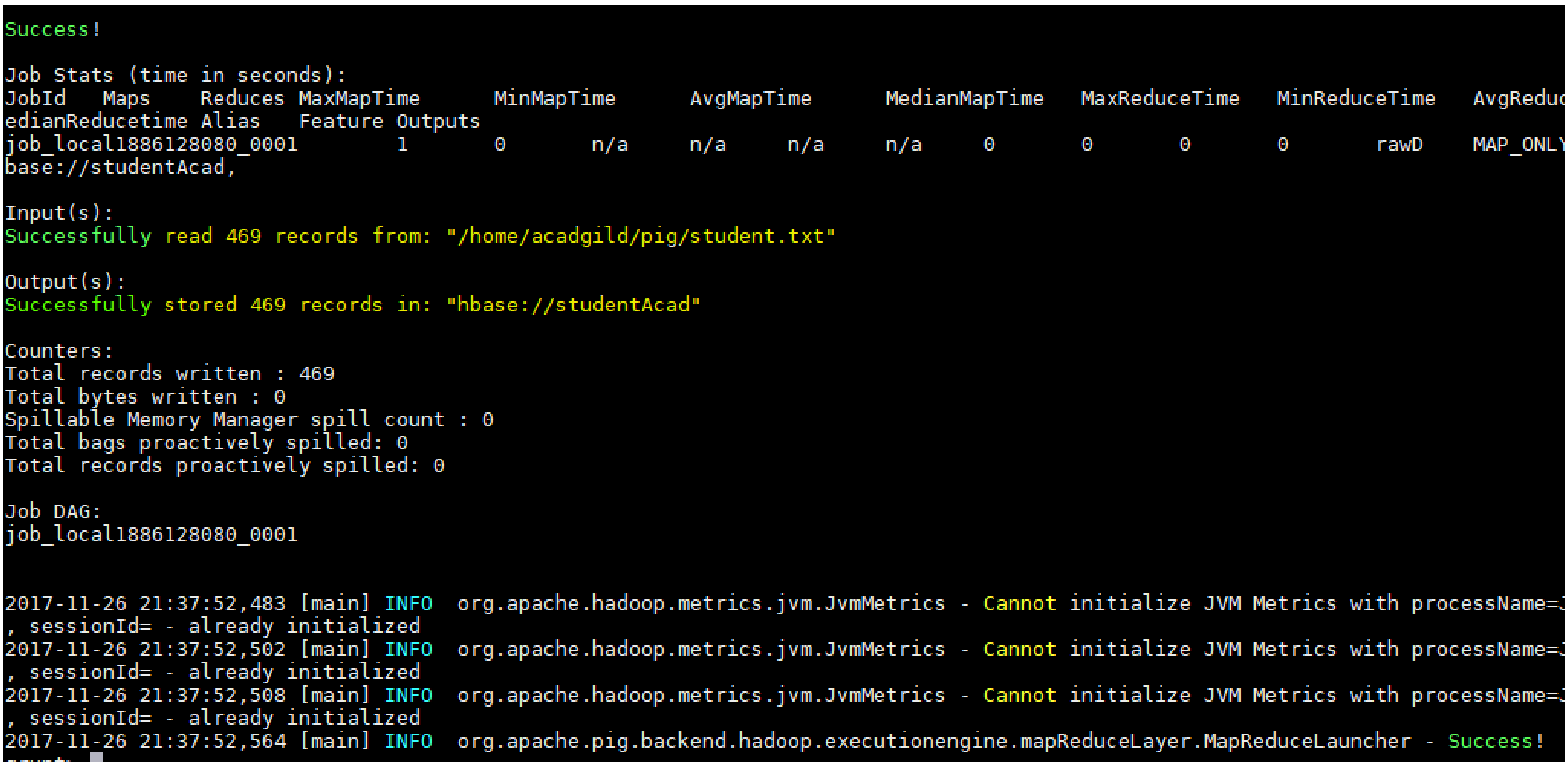
- STORE rawD INTO 'hbase://studentAcad' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage
- (
- 'student_data:StudentName,
- student_data:sector,
- student_data:DOB,
- student_data:qualification,

- student_data:score,
- student_data:state,
- student_data:randomName'
-);



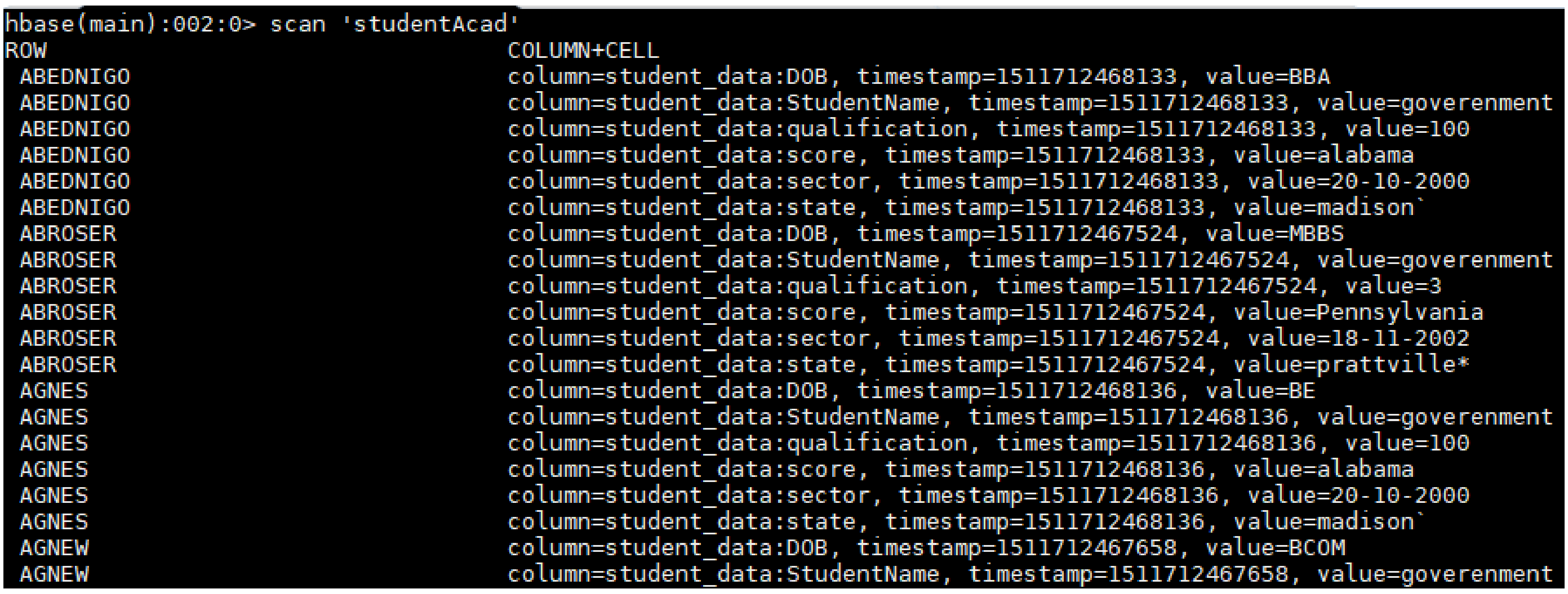
```
grunt> STORE rawD INTO 'hbase://studentAcad' USING org.apache.pig.backend.hadoop.hbase.HBaseStorage
>> (
>> 'student_data:StudentName,
>> student_data:sector,
>> student_data:DOB,
>> student_data:qualification,
>> student_data:score,
>> student_data:state,
>> student_data:randomName'
>> );
```

If the status shows as Success it mean our operation is successful-



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgRedu
edianReducetime Alias   Feature Outputs
job_local1886128080_0001        1       0       n/a     n/a     n/a     n/a     0       0       0       0       rawD    MAP_ONLY
base://studentAcad,

Input(s):
Successfully read 469 records from: "/home/acadgild/pig/student.txt"

Output(s):
Successfully stored 469 records in: "hbase://studentAcad"

Counters:
Total records written : 469
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1886128080_0001


2017-11-26 21:37:52,483 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=J
, sessionId= - already initialized
2017-11-26 21:37:52,502 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=J
, sessionId= - already initialized
2017-11-26 21:37:52,508 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=J
, sessionId= - already initialized
2017-11-26 21:37:52,564 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

<u>Step8-</u> The same will get reflected in Hbase table studentAcad which we can see using scan command-



```
hbase(main):002:0> scan 'studentAcad'
ROW                                  COLUMN+CELL
 ABEDNIGO                            column=student_data:DOB, timestamp=1511712468133, value=BBA
 ABEDNIGO                            column=student_data:StudentName, timestamp=1511712468133, value=goverenment
 ABEDNIGO                            column=student_data:qualification, timestamp=1511712468133, value=100
 ABEDNIGO                            column=student_data:score, timestamp=1511712468133, value=alabama
 ABEDNIGO                            column=student_data:sector, timestamp=1511712468133, value=20-10-2000
 ABEDNIGO                            column=student_data:state, timestamp=1511712468133, value=madison`
 ABROSER                             column=student_data:DOB, timestamp=1511712467524, value=MBBS
 ABROSER                             column=student_data:StudentName, timestamp=1511712467524, value=goverenment
 ABROSER                             column=student_data:qualification, timestamp=1511712467524, value=3
 ABROSER                             column=student_data:score, timestamp=1511712467524, value=Pennsylvania
 ABROSER                             column=student_data:sector, timestamp=1511712467524, value=18-11-2002
 ABROSER                             column=student_data:state, timestamp=1511712467524, value=prattville*
 AGNES                               column=student_data:DOB, timestamp=1511712468136, value=BE
 AGNES                               column=student_data:StudentName, timestamp=1511712468136, value=goverenment
 AGNES                               column=student_data:qualification, timestamp=1511712468136, value=100
 AGNES                               column=student_data:score, timestamp=1511712468136, value=alabama
 AGNES                               column=student_data:sector, timestamp=1511712468136, value=20-10-2000
 AGNES                               column=student_data:state, timestamp=1511712468136, value=madison`
 AGNEW                               column=student_data:DOB, timestamp=1511712467658, value=BCOM
 AGNEW                               column=student_data:StudentName, timestamp=1511712467658, value=goverenment
```