BIG DATA
DEVELOPMENT

ACADGILD

Session 10: Oozie and Sqoop

Assignment 2 Question

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.

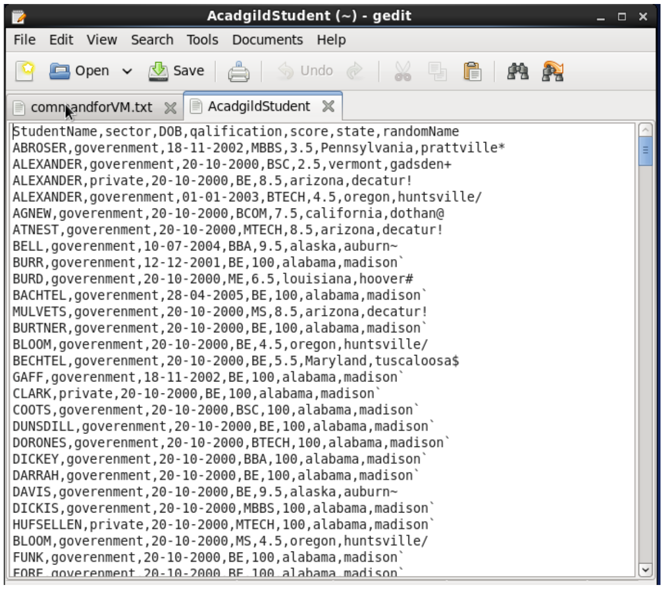
Note: You need to download the following versions of Hadoop, HBase and Pig to implement the steps discussed to load the data into HBase using Pig.

- **Hadoop version: hadoop-2.6.0**
- **Hbase version: hbase-0.98.4-hadoop2-bin**
- **Pig version: pig-0.14.0**

DataSet
https://drive.google.com/file/d/0B1QaXx7tpw3SVUlWTUQyNTMzdG8/view

SampleData(AcadgildStudent)

```
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#
BACHTEL,goverenment,28-04-2005,BE,100,alabama,madison`
MULVETS,goverenment,20-10-2000,MS,8.5,arizona,decatur!
BURTNER,goverenment,20-10-2000,BE,100,alabama,madison`
BLOOM,goverenment,20-10-2000,BE,4.5,oregon,huntsville/
```


AcadgildStudent (~) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
commandforVM.txt
AcadgildStudent

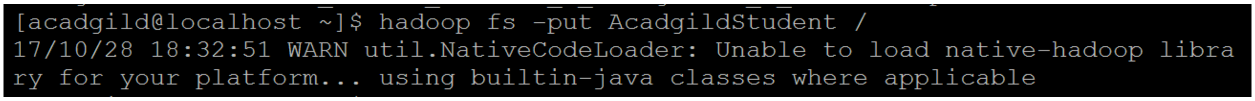
```
StudentName,sector,DOB,qalification,score,state,randomName
ABROSER,goverenment,18-11-2002,MBBS,3.5,Pennsylvania,prattville*
ALEXANDER,goverenment,20-10-2000,BSC,2.5,vermont,gadsden+
ALEXANDER,private,20-10-2000,BE,8.5,arizona,decatur!
ALEXANDER,goverenment,01-01-2003,BTECH,4.5,oregon,huntsville/
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#
BACHTEL,goverenment,28-04-2005,BE,100,alabama,madison`
MULVETS,goverenment,20-10-2000,MS,8.5,arizona,decatur!
BURTNER,goverenment,20-10-2000,BE,100,alabama,madison`
BLOOM,goverenment,20-10-2000,BE,4.5,oregon,huntsville/
BECHTEL,goverenment,20-10-2000,BE,5.5,Maryland,tuscaloosa$
GAFF,goverenment,18-11-2002,BE,100,alabama,madison`
CLARK,private,20-10-2000,BE,100,alabama,madison`
COOTS,goverenment,20-10-2000,BSC,100,alabama,madison`
DUNSDILL,goverenment,20-10-2000,BE,100,alabama,madison`
DORONES,goverenment,20-10-2000,BTECH,100,alabama,madison`
DICKEY,goverenment,20-10-2000,BBA,100,alabama,madison`
DARRAH,goverenment,20-10-2000,BE,100,alabama,madison`
DAVIS,goverenment,20-10-2000,BE,9.5,alaska,auburn~
DICKIS,goverenment,20-10-2000,MBBS,100,alabama,madison`
HUFSELLEN,private,20-10-2000,MTECH,100,alabama,madison`
BLOOM,goverenment,20-10-2000,MS,4.5,oregon,huntsville/
FUNK,goverenment,20-10-2000,BE,100,alabama,madison`
FORE,goverenment,20-10-2000,BE,100,alabama,madison`
```

the description for the above data set containing seven columns named as:
StudentName, sector, DOB, qualification, score, state, randomName.

Transfer into hdfs

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

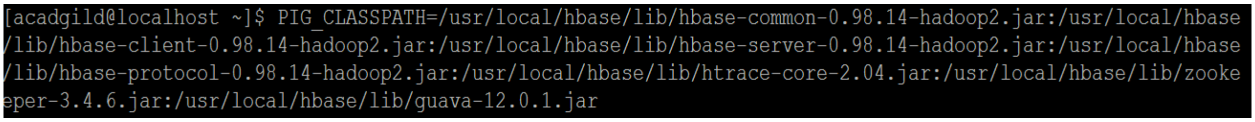
hadoop fs -put AcadgildStudent /



```
[acadgild@localhost ~]$ hadoop fs -put AcadgildStudent /
17/10/28 18:32:51 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
```

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

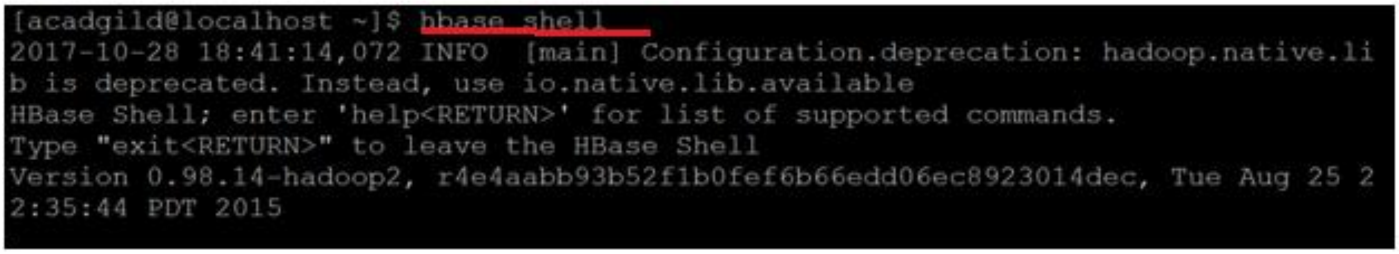
```
PIG_CLASSPATH=/usr/local/hbase/lib/hbase-common-0.98.14-hadoop2.jar:/usr/local/hbase/lib/hbase-client-0.98.14-hadoop2.jar:/usr/local/hbase/lib/hbase-server-0.98.14-hadoop2.jar:/usr/local/hbase/lib/hbase-protocol-0.98.14-hadoop2.jar:/usr/local/hbase/lib/htrace-core-2.04.jar:/usr/local/hbase/lib/zookeeper-3.4.6.jar:/usr/local/hbase/lib/guava-12.0.1.jar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



```
[acadgild@localhost ~]$ 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/zooke
eper-3.4.6.jar:/usr/local/hbase/lib/guava-12.0.1.jar
```

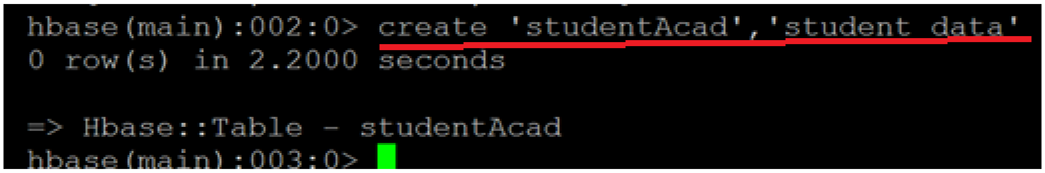
We will now start HBase shell and create a table. We only need this table as skeleton so PIG can Store data inside this by referring the table name.

```
hbase shell
```



```
[acadgild@localhost ~]$ hbase shell
2017-10-28 18:41:14,072 INFO  [main] Configuration.deprecation: hadoop.native.li
b 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 2
2:35:44 PDT 2015
```

```
create 'studentAcad','student_data'
```



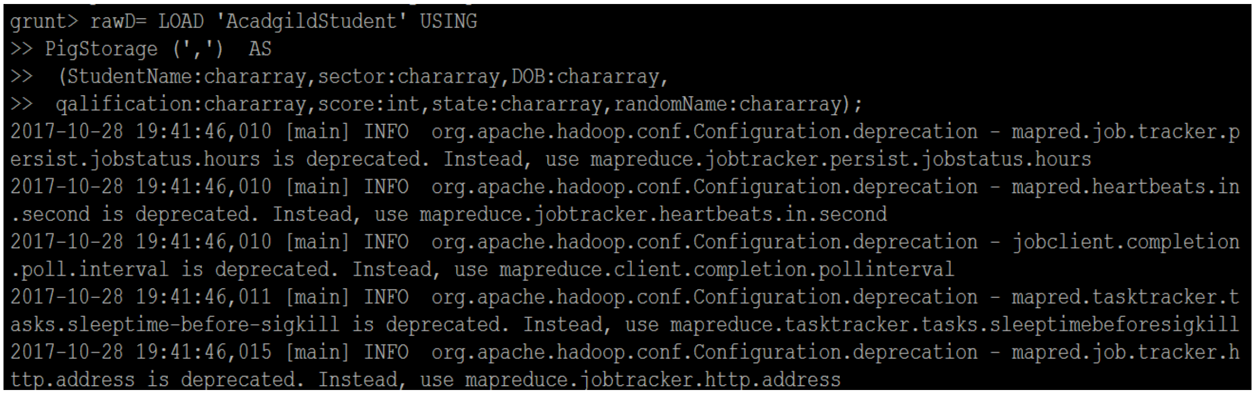
```
hbase(main):002:0> create 'studentAcad','student data'
0 row(s) in 2.2000 seconds

=> Hbase::Table - studentAcad
hbase(main):003:0>
```

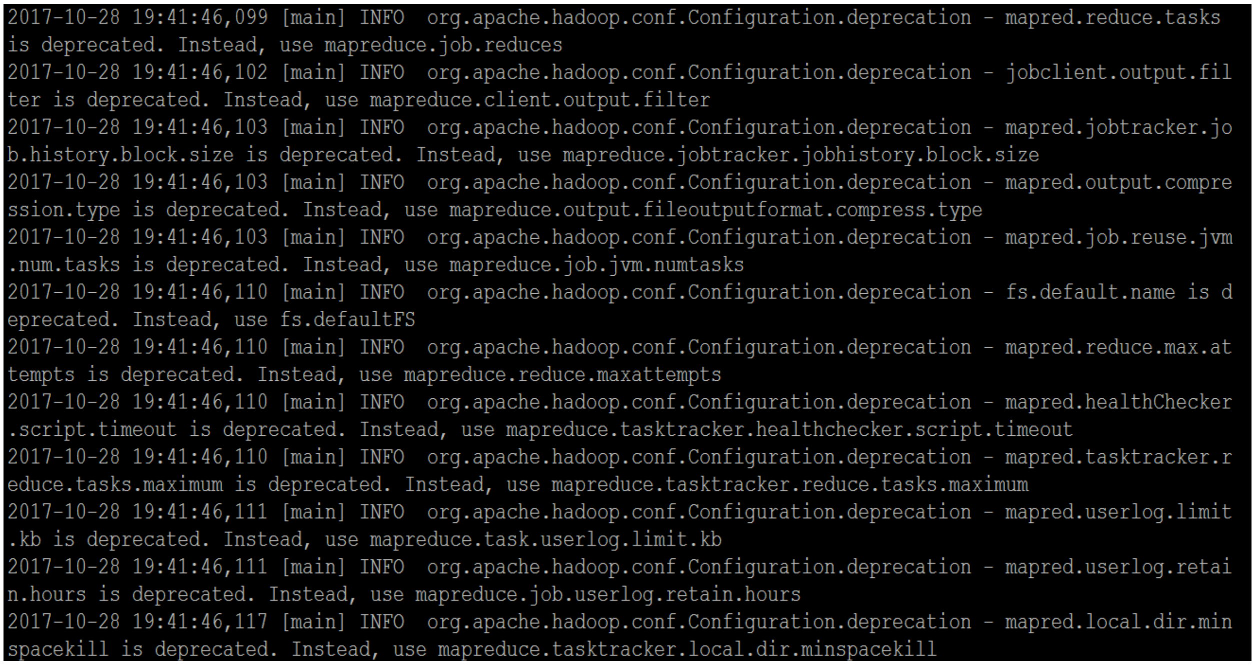
We can come out from HBase by typing exit and switch to PIG grunt shell.Once we are inside PIG mode we can load data from HDFS to Alias relation.

```
rawD= LOAD 'AcadgildStudent' USING
PigStorage (',')  AS
 (StudentName:chararray,sector:chararray,DOB:chararray,

qalification:chararray,score:int,state:chararray,randomName:chararray);
```



```
grunt> rawD= LOAD 'AcadgildStudent' USING
>> PigStorage (',')  AS
>>  (StudentName:chararray,sector:chararray,DOB:chararray,
>>  qalification:chararray,score:int,state:chararray,randomName:chararray);
2017-10-28 19:41:46,010 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.p
ersist.jobstatus.hours is deprecated. Instead, use mapreduce.jobtracker.persist.jobstatus.hours
2017-10-28 19:41:46,010 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.heartbeats.in
.second is deprecated. Instead, use mapreduce.jobtracker.heartbeats.in.second
2017-10-28 19:41:46,010 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - jobclient.completion
.poll.interval is deprecated. Instead, use mapreduce.client.completion.pollinterval
2017-10-28 19:41:46,011 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.t
asks.sleeptime-before-sigkill is deprecated. Instead, use mapreduce.tasktracker.tasks.sleeptimebeforesigkill
2017-10-28 19:41:46,015 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.h
ttp.address is deprecated. Instead, use mapreduce.jobtracker.http.address
```

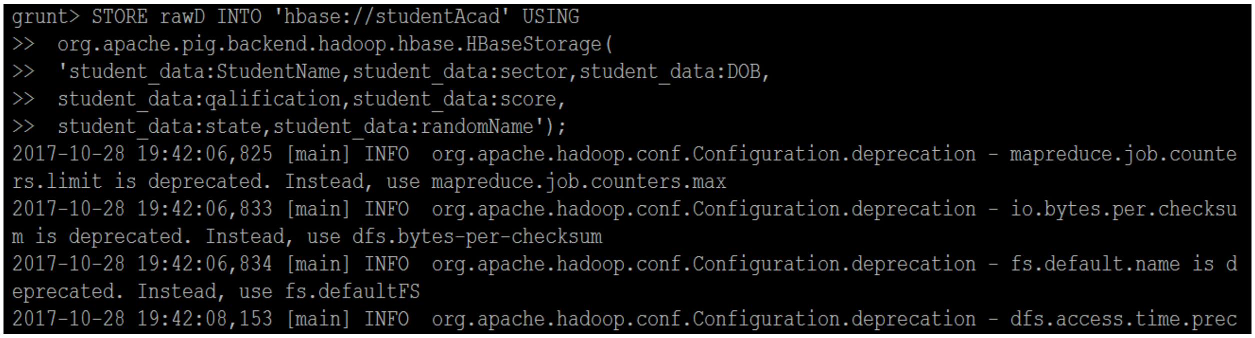


```
2017-10-28 19:41:46,099 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks
is deprecated. Instead, use mapreduce.job.reduces
2017-10-28 19:41:46,102 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - jobclient.output.fil
ter is deprecated. Instead, use mapreduce.client.output.filter
2017-10-28 19:41:46,103 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.jobtracker.jo
b.history.block.size is deprecated. Instead, use mapreduce.jobtracker.jobhistory.block.size
2017-10-28 19:41:46,103 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compre
ssion.type is deprecated. Instead, use mapreduce.output.fileoutputformat.compress.type
2017-10-28 19:41:46,103 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reuse.jvm
.num.tasks is deprecated. Instead, use mapreduce.job.jvm.numtasks
2017-10-28 19:41:46,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is d
eprecated. Instead, use fs.defaultFS
2017-10-28 19:41:46,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.max.at
tempts is deprecated. Instead, use mapreduce.reduce.maxattempts
2017-10-28 19:41:46,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.healthChecker
.script.timeout is deprecated. Instead, use mapreduce.tasktracker.healthchecker.script.timeout
2017-10-28 19:41:46,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.tasktracker.r
educe.tasks.maximum is deprecated. Instead, use mapreduce.tasktracker.reduce.tasks.maximum
2017-10-28 19:41:46,111 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.userlog.limit
.kb is deprecated. Instead, use mapreduce.task.userlog.limit.kb
2017-10-28 19:41:46,111 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.userlog.retai
n.hours is deprecated. Instead, use mapreduce.job.userlog.retain.hours
2017-10-28 19:41:46,117 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.local.dir.min
spacekill is deprecated. Instead, use mapreduce.tasktracker.local.dir.minspacekill
```

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

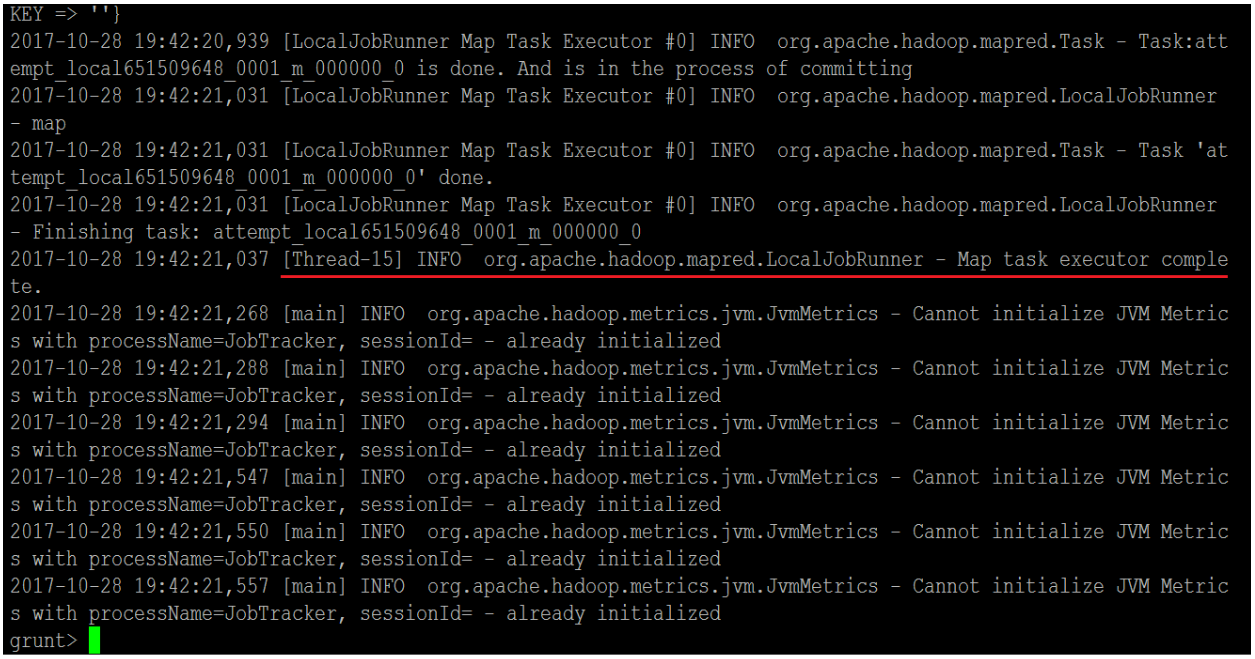
We need to ensure that we give the correct name for table name created inside HBase. Also the parameters should be kept in mind to avoid mistake.

```
 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> 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');
2017-10-28 19:42:06,825 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counte
rs.limit is deprecated. Instead, use mapreduce.job.counters.max
2017-10-28 19:42:06,833 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksu
m is deprecated. Instead, use dfs.bytes-per-checksum
2017-10-28 19:42:06,834 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is d
eprecated. Instead, use fs.defaultFS
2017-10-28 19:42:08,153 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - dfs.access.time.prec
```

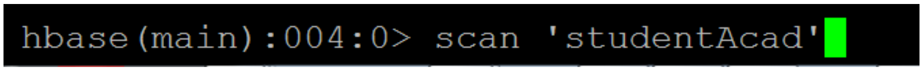
Once the success message comes as shown below , it is confirmed our data is loaded inside HBase.



```
KEY => ''}
2017-10-28 19:42:20,939 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task:att
empt_local651509648_0001_m_000000_0 is done. And is in the process of committing
2017-10-28 19:42:21,031 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner
- map
2017-10-28 19:42:21,031 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.Task - Task 'at
tempt_local651509648_0001_m_000000_0' done.
2017-10-28 19:42:21,031 [LocalJobRunner Map Task Executor #0] INFO  org.apache.hadoop.mapred.LocalJobRunner
- Finishing task: attempt_local651509648_0001_m_000000_0
2017-10-28 19:42:21,037 [Thread-15] INFO  org.apache.hadoop.mapred.LocalJobRunner - Map task executor comple
te.
2017-10-28 19:42:21,268 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metric
s with processName=JobTracker, sessionId= - already initialized
2017-10-28 19:42:21,288 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metric
s with processName=JobTracker, sessionId= - already initialized
2017-10-28 19:42:21,294 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metric
s with processName=JobTracker, sessionId= - already initialized
2017-10-28 19:42:21,547 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metric
s with processName=JobTracker, sessionId= - already initialized
2017-10-28 19:42:21,550 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metric
s with processName=JobTracker, sessionId= - already initialized
2017-10-28 19:42:21,557 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metric
s with processName=JobTracker, sessionId= - already initialized
grunt>
```

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

scan 'studentAcad'



```
hbase(main):004:0> scan 'studentAcad'
```