Loading Data Into HBase Using PIG Scripts.

In this assignment we are going to transfer data into HBase using Pig. We are taking sample data set of student which will be loaded into HBase.

The columns in the dataset are : StudentName,sector,DOB,qualification,score,state,randomName

Loading the data in HDFS:

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

hadoop fs -put /home/acadgild/hadoop/student.txt /

Screenshot of Mobaxterm for loading the student.txt in HDFS:

```
[acadgild@localhost ~]$ hadoop fs -put /home/acadgild/hadoop/student.txt /

[acadgild@localhost ~]$ hadoop fs -ls /
Found 17 items
-rw-r--r--   1 acadgild supergroup        416 2015-11-16 23:43 /Wc.txt
drwxr-xr-x   - acadgild supergroup          0 2017-11-11 16:16 /hbasestorage
-rw-r--r--   1 acadgild supergroup         28 2015-11-17 01:49 /inp
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 02:03 /out23
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 01:47 /sample-mr
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 13:46 /sqoopout
-rw-r--r--   1 acadgild supergroup      26204 2017-11-11 16:58 /student.txt
drwxrwxr-x   - acadgild supergroup          0 2017-11-10 20:22 /tmp
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 01:56 /user
drwxr-xr-x   - acadgild supergroup          0 2015-11-16 23:57 /wc_out
drwxr-xr-x   - acadgild supergroup          0 2015-11-16 23:45 /wc_out1
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 00:01 /wc_out2
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 00:04 /wc_out3
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 00:07 /wc_out4
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 00:11 /wc_out5
drwxr-xr-x   - acadgild supergroup          0 2015-11-19 15:17 /wc_out6
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 12:56 /zookeeper
```

Including HBase jar files in pig classpath:

Once the data is loaded into HDFS, Using the below command we are registering the hbase.har in pig class path:
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;

Screenshot of Mobaxterm for registering the hbase.jar in pig_classpath

```
[acadgild@localhost ~]$ PIG_CLASSPATH=/home/hadoop/HADOOP/hbase-0.98.4-hado.98.4-hadoop2/lib/hbase-*.jar;
```

Creating the table in HBase:

Once the pig_classpath is registered, We will now start HBase shell and create a table named studentAcad_Tab using the below command. We only need this table as skeleton so PIG can Store data inside this by referring the table name.

create 'studentAcad_Tab','student_data'

Screenshot of Mobaxterm for creating the table in HBase:

```
hbase(main):009:0> create 'studentAcad_Tab','student_data'
0 row(s) in 0.4180 seconds

=> Hbase::Table - studentAcad_Tab
```

Pig shell to upload the data to HBase table:

1) Once the table is created in HBase, We will now start the pig grunt shell and load the student.txt using PigStorage
2) store the contents of students.txt to Alias relation 'rawD'.

```
rawD = LOAD 'student.txt' using PigStorage(',') AS
(StudentName:chararray,sector:chararray,DOB:chararray,qualification:chararray,score:chararray,state:chararray,randomName:chararr
ay);
```

Screenshot of Mobaxterm for loading student.txt using pigstorage:

```
grunt> rawD = LOAD 'student.txt' using PigStorage(',') AS  (StudentName:chararray,sector:chararray,DOB:chararray,qualification:chararray,score:c
hararray,state:chararray,randomName:chararray);
```

3) Once the student data is loaded in rawD, we are storing the contents of rawD into the hbase table 'hbase://studentAcad_Tab' , using store command as below. 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_Tab' 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');
```

Screenshot of Mobaxterm for storing the rawD into hbase table : studentAcad_Tab:

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

Viewing the table in HBase:

Once the data is stored to HBase, we can we the contents of HBase table using Scan command.

```
scan 'studentAcad_Tab'
```

Screenshot of Mobaxterm for creating the table in HBase:

```
hbase(main):010:0> scan 'studentAcad_Tab'
ROW                                   COLUMN+CELL
 ABEDNIGO                             column=student_data:DOB, timestamp=1510398242641, value=BBA
 ABEDNIGO                             column=student_data:StudentName, timestamp=1510398242641, value=goverenment
 ABEDNIGO                             column=student_data:qualification, timestamp=1510398242641, value=100
 ABEDNIGO                             column=student_data:score, timestamp=1510398242641, value=alabama
 ABEDNIGO                             column=student_data:sector, timestamp=1510398242641, value=20-10-2000
 ABEDNIGO                             column=student_data:state, timestamp=1510398242641, value=madison`
 ABROSER                              column=student_data:DOB, timestamp=1510398242470, value=MBBS
 ABROSER                              column=student_data:StudentName, timestamp=1510398242470, value=goverenment
 ABROSER                              column=student_data:qualification, timestamp=1510398242470, value=3.5
 ABROSER                              column=student_data:score, timestamp=1510398242470, value=Pennsylvania
 ABROSER                              column=student_data:sector, timestamp=1510398242470, value=18-11-2002
 ABROSER                              column=student_data:state, timestamp=1510398242470, value=prattville*
 AGNES                                column=student_data:DOB, timestamp=1510398242641, value=BE
 AGNES                                column=student_data:StudentName, timestamp=1510398242641, value=goverenment
 AGNES                                column=student_data:qualification, timestamp=1510398242641, value=100
 AGNES                                column=student_data:score, timestamp=1510398242641, value=alabama
 AGNES                                column=student_data:sector, timestamp=1510398242641, value=20-10-2000
 AGNES                                column=student_data:state, timestamp=1510398242641, value=madison`
 AGNEW                                column=student_data:DOB, timestamp=1510398242471, value=BCOM
 AGNEW                                column=student_data:StudentName, timestamp=1510398242471, value=goverenment
 AGNEW                                column=student_data:qualification, timestamp=1510398242471, value=7.5
 AGNEW                                column=student_data:score, timestamp=1510398242471, value=california
 AGNEW                                column=student_data:sector, timestamp=1510398242471, value=20-10-2000
 AGNEW                                column=student_data:state, timestamp=1510398242471, value=dothan@
 ALEXANDER                            column=student_data:DOB, timestamp=1510398242642, value=MBBS
 ALEXANDER                            column=student_data:StudentName, timestamp=1510398242642, value=goverenment
 ALEXANDER                            column=student_data:qualification, timestamp=1510398242642, value=100
 ALEXANDER                            column=student_data:score, timestamp=1510398242642, value=alabama
 ALEXANDER                            column=student_data:sector, timestamp=1510398242642, value=20-10-2000
 ALEXANDER                            column=student_data:state, timestamp=1510398242642, value=madison`
```

Scan command shows the student data is loaded into the HBase table.