Loading Data Into HBase Using PIG Scripts

Step 1: Starting hadoop daemons and job history server using below commands

```
[acadgild@localhost ~]$ $HADOOP_HOME/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/08/21 14:26:51 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-aca
dgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-aca
dgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/had
oop-acadgild-secondarynamenode-localhost.localdomain.out
17/08/21 14:27:50 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadgild-
resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-ac
adgild-nodemanager-localhost.localdomain.out
[acadgild@localhost ~]$ mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /usr/local/hadoop-2.6.0/logs/mapred-acadgild-
historyserver-localhost.localdomain.out
[acadgild@localhost ~]$
```

Step 2: Starting hbase daemons using below command, and checking using jps whether all required daemons have started or not

```
[acadgild@localhost ~]$ $HBASE_HOME/bin/start-hbase.sh
starting master, logging to /usr/local/hbase/logs/hbase-acadgild-master-localhos
t.localdomain.out
[acadgild@localhost ~]$ jps
3363 JobHistoryServer
3300 NodeManager
3846 Jps
2761 NameNode
3785 HMaster
2858 DataNode
3036 SecondaryNameNode
3197 ResourceManager
[acadgild@localhost ~]$
```

Required daemons have started

Step 3: Copying the data set "student.txt" into HDFS which will further be loaded into HBase

```
[acadgild@localhost ~]$ hadoop fs -put /home/acadgild/Documents/SharedDatafromWi
ndows/Session10/student.txt /user/
17/08/21 13:34:39 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop fs -ls /user/
17/08/21 13:34:58 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 6 items
drwxr-xr-x   - acadgild supergroup          0 2017-08-20 14:16 /user/acadgild
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 12:52 /user/hive
drwxr-xr-x   - acadgild supergroup          0 2017-08-13 00:11 /user/my_pig_stuf
f
drwxr-xr-x   - acadgild supergroup          0 2017-08-20 23:29 /user/oozie
drwxr-xr-x   - acadgild supergroup          0 2015-11-08 17:35 /user/prateek
-rw-r--r--   1 acadgild supergroup      26204 2017-08-21 13:34 /user/student.txt
[acadgild@localhost ~]$
```

Below screenshot shows that data has been loaded successfully in hdfs:

```
[acadgild@localhost ~]$ hadoop fs -cat /user/*.txt
17/08/21 13:36:06 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
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`
```

Step 4: Including few jar files of HBase to the Pig classpath

```
[acadgild@localhost ~]$ export PIG_CLASSPATH=/usr/local/hbase/lib:/usr/loca
se/lib/*.jar;
[acadgild@localhost ~]$
```

Step 5: Starting HBase shell and creating a table "studentAcad"

We only need this table as skeleton so PIG can store data inside this by referring the table name.

```
[acadgild@localhost ~]$ hbase shell
2017-08-21 13:42:03,060 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

                                                    studentAcad - Table
hbase(main):001:0> create 'studentAcad','student_data'   student_data - column_family
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/li
b/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-08-21 13:42:47,655 WARN  [main] util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes where applicabl
e
0 row(s) in 9.8420 seconds

=> Hbase::Table - studentAcad   <-- studentAcad table is created successfully
hbase(main):002:0>
```

Using describe command, we can check schema of "studentAcad" table

```
hbase(main):004:0> describe 'studentAcad'
Table studentAcad is ENABLED
studentAcad
COLUMN FAMILIES DESCRIPTION
{NAME => 'student_data', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'fa
lse', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FORE
VER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZ
E => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.2540 seconds

hbase(main):005:0>
```

Step 6: Starting PIG in mapreduce mode

```
[acadgild@localhost Session10]$ pig -x mapreduce
2017-08-21 13:46:50,601 INFO  [main] pig.ExecTypeProvider: Trying ExecType : LOC
AL
2017-08-21 13:46:50,607 INFO  [main] pig.ExecTypeProvider: Trying ExecType : MAP
REDUCE
2017-08-21 13:46:50,607 INFO  [main] pig.ExecTypeProvider: Picked MAPREDUCE as t
he ExecType
2017-08-21 13:46:51,064 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
14.0 (r1640057) compiled Nov 16 2014, 18:02:05
2017-08-21 13:46:51,065 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/acadgild/Documents/SharedDatafromWindows/Session10/pig_1503303411058
.log
```

```
org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/li
b/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-08-21 13:46:54,946 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - U
nable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
2017-08-21 13:46:56,916 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

Step 7: As we are inside PIG grunt shell, so loading data from HDFS to Alias relation

```
grunt> rawD = LOAD '/user/student.txt' USING PigStorage(',') AS (StudentName:cha
rarray,sector:chararray,DOB:chararray,qualification:chararray,score:int,state:ch
ararray,randomName:chararray);
2017-08-21 13:51:00,776 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.c
ounters.max
2017-08-21 13:51:00,779 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-08-21 13:51:00,780 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

Step 8: Now transferring 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:D
OB,student_data:qualification,student_data:score,student_data:state,student_data
:randomName');
```

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

```
Job DAG:
job_1503302416799_0001


2017-08-21 14:10:42,416 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2017-08-21 14:10:42,470 [main] INFO  org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2017-08-21 14:10:42,740 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2017-08-21 14:10:42,759 [main] INFO  org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2017-08-21 14:10:42,930 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2017-08-21 14:10:42,971 [main] INFO  org.apache.hadoop.mapred.ClientServiceDeleg
ate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirect
ing to job history server
2017-08-21 14:10:43,209 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Unable to retrieve job to compute warning
aggregation.
2017-08-21 14:10:43,209 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Step 9: The result can be displayed through scan command followed by table name inside quotes (‘ ’)

```
hbase(main):001:0> scan 'studentAcad';
```

```
WORK                 column=student_data:StudentName, timestamp=1503304806722,
                     value=goverenment
WORK                 column=student_data:qualification, timestamp=1503304806722
                     , value=100
WORK                 column=student_data:score, timestamp=1503304806722, value=
                     alabama
WORK                 column=student_data:sector, timestamp=1503304806722, value
                     =28-04-2005
WORK                 column=student_data:state, timestamp=1503304806722, value=
                     madison`
YOUNG                column=student_data:DOB, timestamp=1503304806726, value=MS
YOUNG                column=student_data:StudentName, timestamp=1503304806726,
                     value=goverenment
YOUNG                column=student_data:qualification, timestamp=1503304806726
                     , value=9
YOUNG                column=student_data:score, timestamp=1503304806726, value=
                     alaska
YOUNG                column=student_data:sector, timestamp=1503304806726, value
                     =20-10-2000
YOUNG                column=student_data:state, timestamp=1503304806726, value=
                     auburn~
230 row(s) in 12.8980 seconds

hbase(main):002:0>
```