HU Extension School E-185 Big Data Analytics

# Assignment 04

Handed out: 03/02/2013 Due by 5:30PM on Friday, 03/08/2013

Files `all-bible.tar` and `all-shakespeare.tar` used during the lecture are attached. Files `kv1.txt, kv2.txt, kv3.txt` referenced in the lecture notes could be found on `ami-5abc2f33` and most probably any other instance created for a Hive interactive job flow by the Elastic Map Reduce service. The files reside in the directory: `/home/hadoop/hive/examples/files.`
Using AWS Elastic Map Reduce service open an interactive Hive job with a minimal cluster. Use installation of Hive on the master node of that cluster.

**Problem 1)** Add regular expression serializer/deserializer `RegexSerDe` by issuing the following command on the hive prompt:

```
hive> add jar hive/contrib/hive_contrib-0.8.1.jar
```

The version of your `hive-contrib.jar` file might be different. Now, you are ready to create hive table `apachelog` using create table command presented on slide 61 of the lecture notes.

```
CREATE TABLE apachelog (
host STRING,
identity STRING,
user STRING,
time STRING,
request STRING,
status STRING,
size STRING,
referer STRING,
agent STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
WITH SERDEPROPERTIES ( "input.regex" = "([^ ]*) ([^ ]*) ([^ ]*) (-
|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)(?: ([^
\"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\"))?", "output.format.string" =
"%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s" )
STORED AS TEXTFILE;
```

Verify that table apachelog accepts apache access logs by loading both apache log files in the directory `/home/hadoop/hive/examples/files.` Please make sure that you load all rows present in those files. Select all rows and present the result.

**Problem 2)** S3 bucket `s3n://elasticmapreduce/samples/pig-apache/input` contains `6 apache_access_logs.` Copy those logs to an HDFS directory on your cluster. Load all 6 access_logs into table apachelog. Query the number of rows you loaded.

**Problem 3)** You recall that we said that hive stores all values in its tables in a HDFS directory. Locate that directory and present a small portion of the file containing table apachelog.

**Problem 4)** Change the name of your table apachelog to apachelogold and create new table apachelog which will support partitioning. Load each of `apache_access_log_x` files into a separate partition. Find out how many rows you have in each partition.

**Problem 5)** Export data from one of above partitions to an OS file. Examine and display the first 10 rows of the new file.

**Problem 6)** You noticed how Hive provides you with the kill command for every job that you run. Our jobs are not very long, still demonstrate that you can kill one of them. You need to open another terminal window to do that.

Capture all steps of your implementation with comments indicating what is it you are accomplishing with every step in an MS Word document.
Please place all files you want to submit in a folder named: HW04. Compress that folder into an archive named E185_LastNameFirstNameHW04. ZIP. Upload the archive to the course drop box, i.e. the web site. Please send comments and questions to cscie185@fas.harvard.edu