

Handed out: 04/05/2014

Due by 11:30PM on Saturday, 04/12/2014

Problem 1) From <http://www.cloudera.com/content/support/en/downloads.html> please download Cloudera's VM (4.4). You login to that VM as user cloudera with password cloudera. User root has the same password. User cloudera is a sudo user. On your VM, locate file `piggybank.jar` and copy it to user cloudera home directory. Open PIG's local mode and examine all commands in the attached file `script.pig`. Simply run those commands one by one and periodically create 10 line extracts of the most recent collections and DUMP those extracts to the screen. In order to save you from downloading file `access_log_1` from the `S3://elasticmapreduce` bucket, we are providing you with that file as well. Describe data structures you are observing and the effects of every command.

Problem 2) Once you are convinced that the above `script.pig` works create an EMR cluster in AWS Amazon Cloud, and navigate, i.e. `ssh`, to the master node of that cluster. From the command prompt of your PC or MAC transfer your script using `scp` command. Your command would look like the following:

```
$ scp -i e63.pem  
script.pig hadoop@ec2-54-35-13-40.compute-1.amazonaws.com:~hadoop
```

The last portion beyond the DNS name, "`:~hadoop`" is telling `scp` command to place file `script.pig` in the home directory of user `hadoop`.

Download `access_log_1` from `s3n://elasticmapreduce` bucket directly into `~hadoop` directory, using appropriate `hadoop fs` command. Subsequently run `script.pig` in `mapreduce` mode by invoking `pig` with option `-x mapreduce`. In that case, your input and output files have to reside in HDFS. Try running `script.pig` in the local mode as well by issuing a command similar to the following, all on one line:

```
$ pig -x local -p INPUT=file:///home/hadoop/access_log_1  
-p OUTPUT=file:///home/hadoop/results script.pig
```

As usual, please capture all the 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: `HW09`. Compress that folder into an archive named `E63_LastNameFirstNameHW09`. ZIP. Upload the archive to the course drop box on the class web site. Please send comments and questions to cscie63@fas.harvard.edu