Homework 6: Hive and Pig

# Learning Objectives

In this exercise, you will learn basic Hadoop Commands:

* Pig for ETL and other data transformations
* Hive for SQL-Like Queries from Hadoop data

# Setup

To complete this homework or follow along with the instructor’s class demos from your coursework, you must first complete the steps in this setup section. For this exercise we require the Cloudera Hadoop instance in the Docker environment. For more details on the how’s and why’s of managing the database environments used in this course, consult Lab 1. Instructions for this assignment:

1. Open the PowerShell prompt on Windows or the Terminal on MacOS.
2. Type:  
   cd adv-db-labs  
   to change the working directory to the repository folder. If you are in the correct spot, your command prompt should have adv-db-labs in it, for instance: PS adv-db-labs>
3. Change into the hadoop folder, type:  
   cd hadoop  
   if you are in the correct folder, your command prompt should have hadoop in it.
4. To be safe, let’s bring down Cloudera Hadoop, type:  
   docker-compose down
5. Bring up the Cloudera Hadoop environment, type:  
   docker-compose up -d
6. Check to make sure the environment is running, type:  
   docker-compose ps

The state of the **cloudera** container should be **Up** on **ports 7180, 8080,** and **80**.

1. You are now ready to connect to the running instance, and login as the **cloudera** user:   
   docker-compose exec cloudera bash -c "su -l cloudera"
2. Your prompt should now say: [cloudera@quickstart ~] where you are ready to execute lab commands!
3. When you need to connect to pig:  
   **[cloudera@quickstart ~]** pig
4. When you need to connect to Hive:  
   **[cloudera@quickstart ~]** beeline -u jdbc:hive2://localhost:10000/default -n cloudera -p cloudera --silent=true

# Exercises

Complete each of the following exercises. If you are unsure how to accomplish the task, please consult the coursework videos where there are explanations and demos.

1. Load the comma-delimited HDFS dataset at **clickstream/iplookup** into a relation with an explicit schema. Use filter logic to remove the first row (which contains a header) then sort the output by IP and dump a comma-delimited data set to **clickstream/iplookup\_noheader.** Record all of your Pig commands required to complete your transformation.
2. Write Pig commands to produce a count of IP Addresses by state codes, sorted by the count with highest values first, like this:  
   (CA, 10)  
   (NY, 4)  
   (VA, 2)  
   Etc…  
   Record all your Pig commands required to complete your transformation.
3. Use pig to load the web log files from **clickstream/logs** using the following schema:   
   reqdate:chararray, reqtime:chararray, x1:int, method:chararray, uri:chararray, x2:int ,x3:int, x4:int ,ipaddress:chararray, useragent:chararray, filter any rows which begin with a “#” (these are header rows and should be removed, then writes out the reqdate, reqtime, method, uri, ipaddress and useragent columns to a tab-delimited data set in HDFS **clickstream/logs\_noheader**. HINT: The data is space delimited.
4. Use hive to create two external tables for the **clickstream/logs\_noheader** and **clickstream/iplookup\_noheader** files you created in the previous steps. These tables should be named **weblogs** and **iplookup** respectively and should be placed in the **clickstream** database. Be sure to record all HQL steps to complete the operations.
5. Write an HQL query to display the name of the city and the number of HTTP requests from that city (NOTE: each row in the web logs is an HTTP request). Order the output so cities with the most requests are at the top. If you complete the query correctly, you should see Syracuse has 272-page requests and Los Angeles has 24.

# Turning it in

Take your copy and paste each of the solutions to the exercises into the submission template file included with this assignment. Make sure your name and SU email are at the top and turn in your work through the course learning management system.

signment. Make sure your name and SU email are at the top and turn in your work through the course learning management system.

# Tear-Down / Troubleshooting Common Issues

**IMPORTANT NOTE:** Hadoop is a complex environment with many interworking parts. Since there are many services running in the container, it is suggested that you bring down the container completely whenever:

1. You are finished with the lab, or
2. Whenever your computer goes to sleep. When that happens, the running Hadoop container is going to lose connectivity among its services.

When you are **finished with the homework or stopping work for an extended period of time** you should bring down the environment like this:

1. From the terminal window where you typed docker-compose up -d type in the following:  
   docker-compose down

If your **Hadoop container seems to have broken services**. Tear down the container and bring it back up again:

1. From the terminal window where you typed docker-compose up -d type in the following:  
   docker-compose down and then docker-compose up -d