# IST769 Homework Submission #6

## Basic Information

Your Name: Thulasi Ram Ruppa Krishnan   
Your SUID: 460746269  
Your Email: truppakr@syr.edu  
Date Due: 5-06-2020   
Homework #: HW6

## Instructions

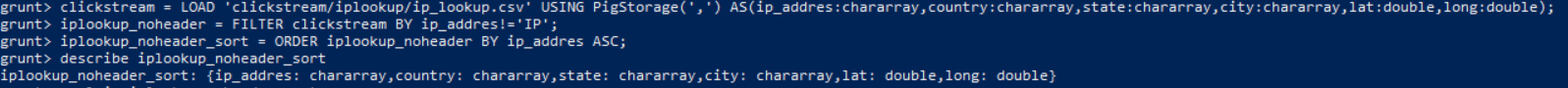
For each answer, please include your answer as text, and any screenshot(s) which demonstrate your answer was executed. Most importantly, make sure to include evidence your answer is correct. This will most likely be a screenshot. If you had issues, problems, or had to make assumptions include them in your answer.

## Answers:

commands required to complete the task.

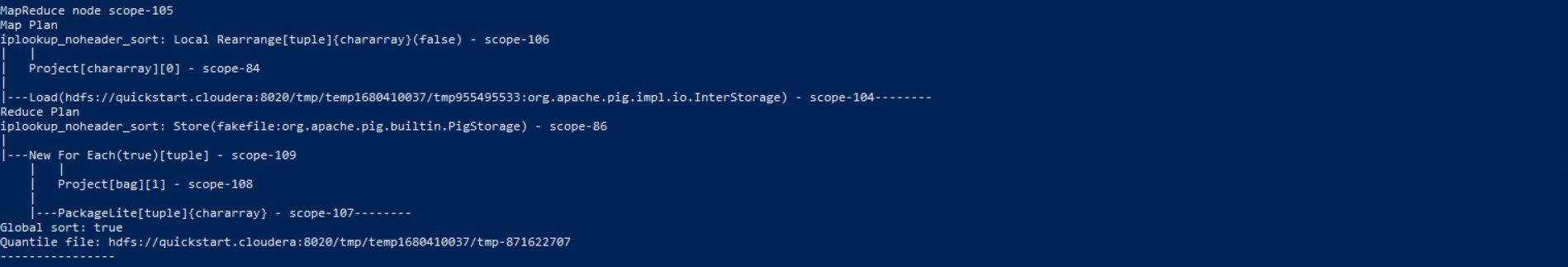
1. Load the comma delimited HDFS dataset at **clickstream/iplookup** into a relation with an explicit schema. Use filter logic to remove the first row (which contains a header) then sort the output by IP and dump a comma-delimited data set to **clickstream/iplookup\_noheader.** Record all of your Pig commands required to complete your transformation.

**Load Data and Describe Explicit schema**



**explain iplookup\_noheader\_sort**



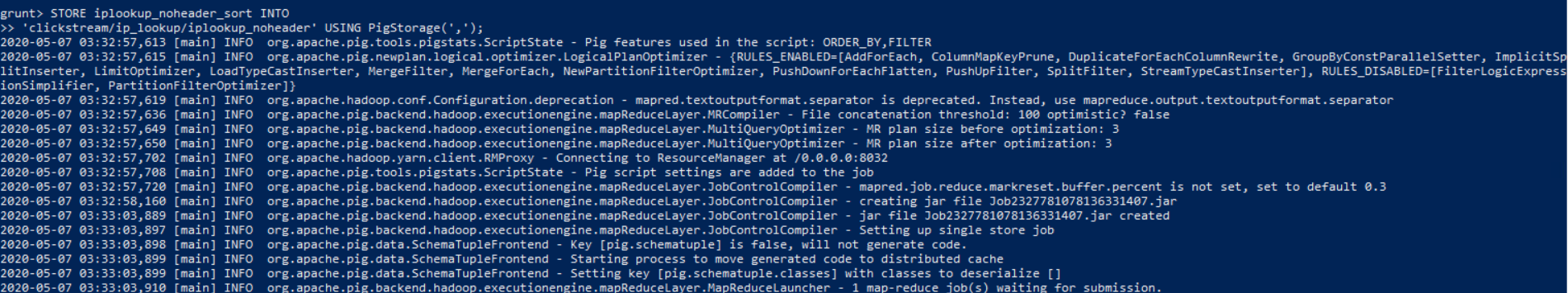


**dump ip\_lookup\_noheader\_sort**

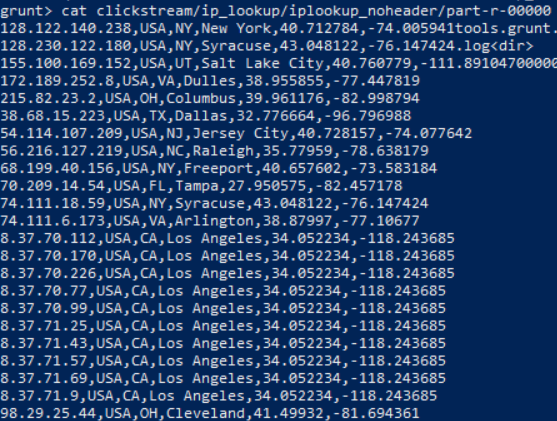


**STORE iplookup\_noheader\_sort INTO**

**‘clickstream/ip\_lookup/iplookup\_noheader’ USING PigStorage(‘,’);**

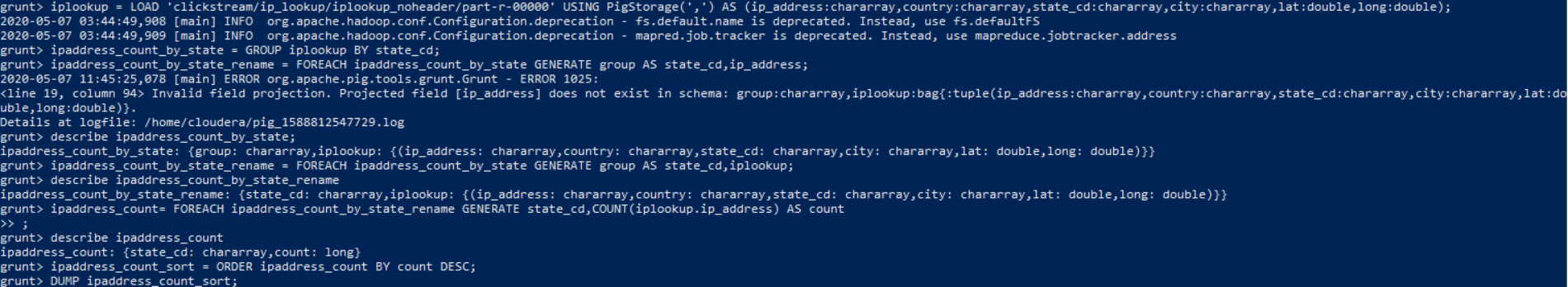




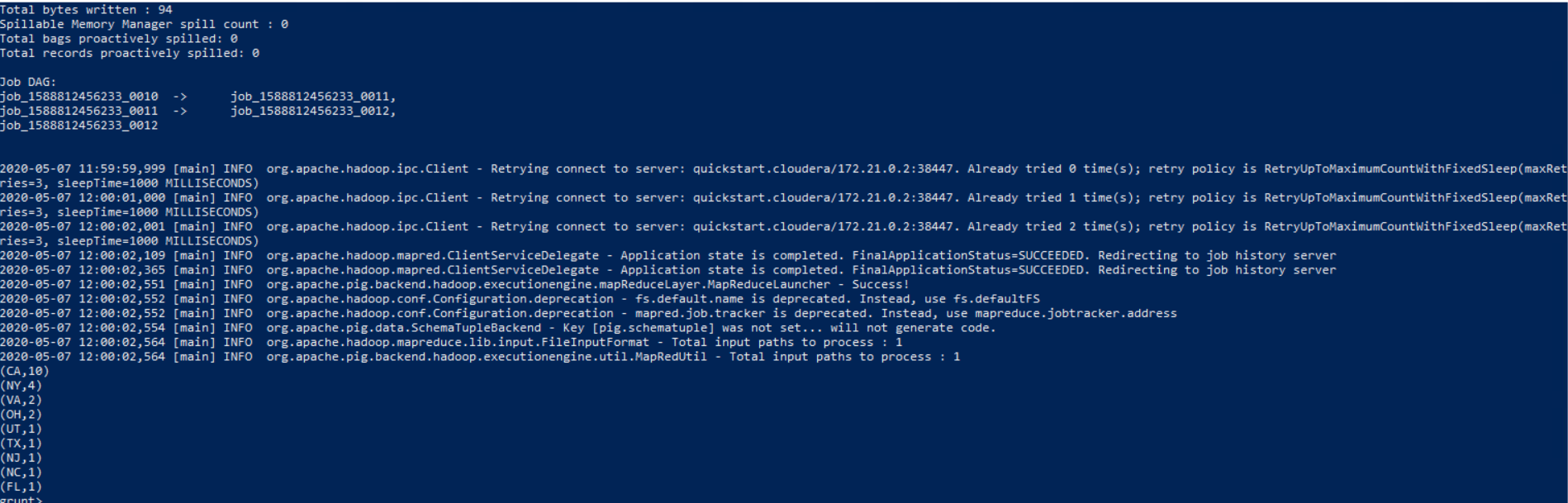


1. Write Pig commands to produce a count of IP Addresses by state codes, sorted by the count with highest values first, like this:  
   (CA, 10)  
   (NY, 4)  
   (VA, 2)  
   Etc…  
   Record all your Pig commands required to complete your transformation.

**Pig Commands**

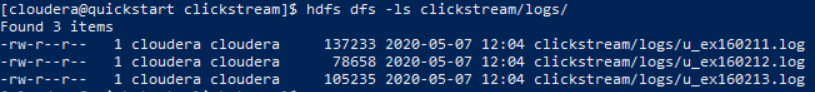


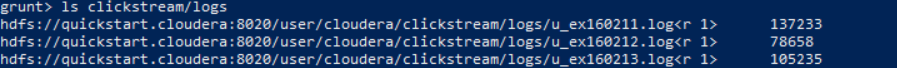
**Results**



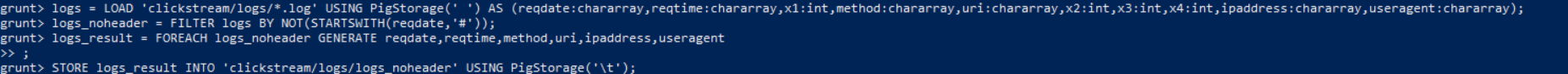
1. Use pig to load the web log files from **clickstream/logs** using the following schema:   
   reqdate:chararray, reqtime:chararray, x1:int, method:chararray, uri:chararray, x2:int ,x3:int, x4:int ,ipaddress:chararray, useragent:chararray, filter any rows which begin with a “#” (these are header rows 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.

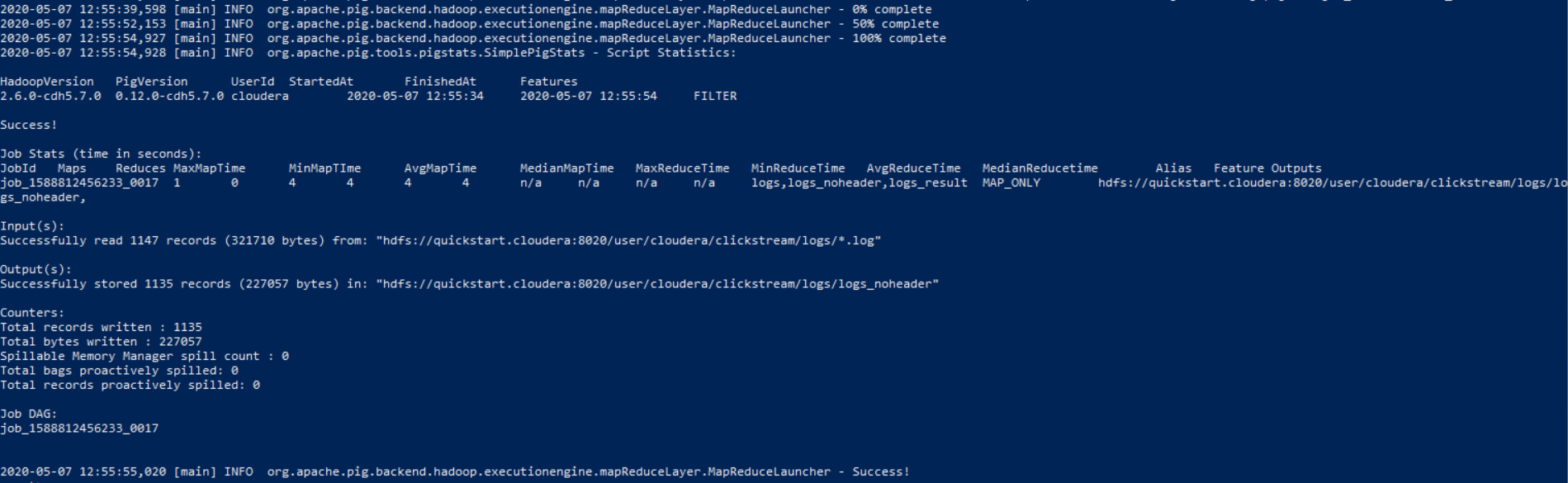




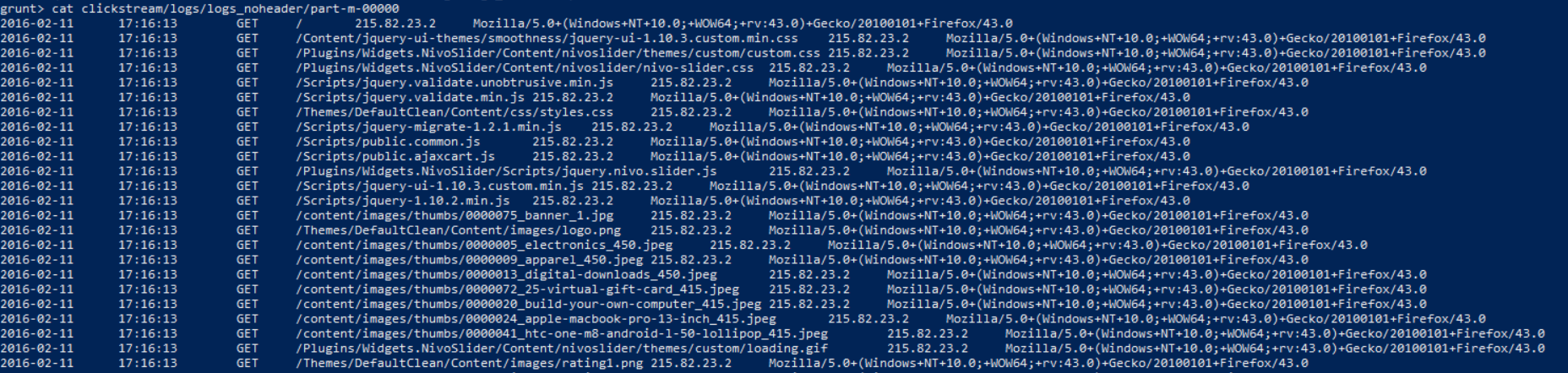




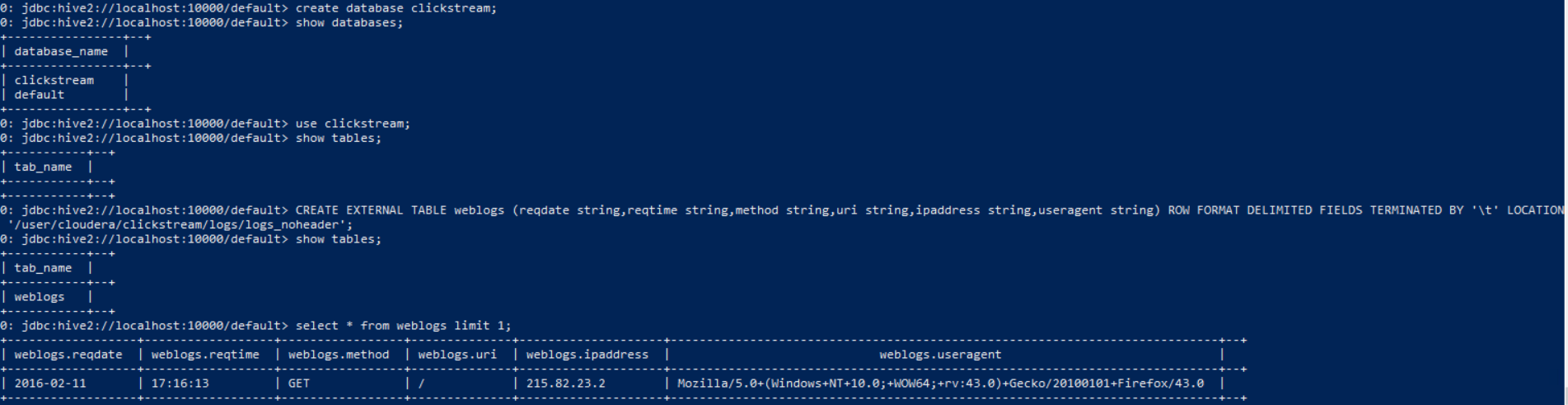


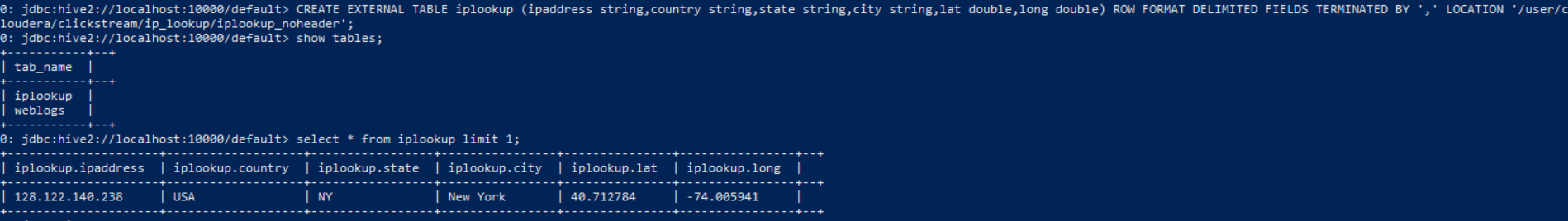






1. Use hive to create two external tables for the **clickstream/logs\_noheader** and **clickstream/iplookup\_noheader** files you created in the previous steps. These tables should be named **weblogs** and **iplookup** respectively and should be placed in the **clickstream** database. Be sure to record all HQL steps to complete the operations.





1. 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.

# 

# Tear-Down

When you are finished with the homework you should stop the environment:

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