IST769 Hive and Pig

Your Name: Tajudeen Abdulazeez   
Your SUID: 69687-7373-0   
Your Email: toabdula@syr.edu   
Date Due:   
Homework #: 6

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.

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.

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.