IST769 Hive and Pig

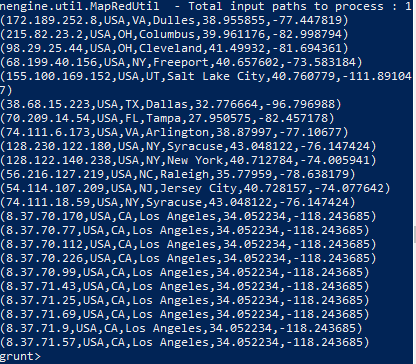
Your Name: Tajudeen Abdulazeez   
Your SUID: 69687-7373-0   
Your Email: toabdula@syr.edu   
Date Due:   
Homework #: 6

* 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.

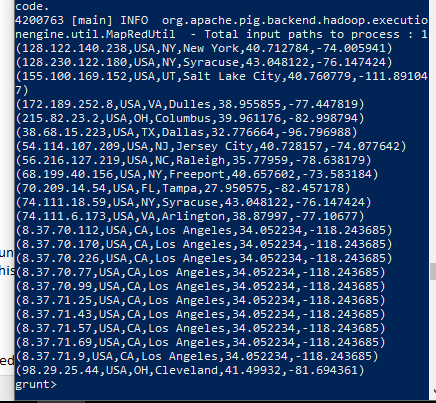
f1 = LOAD 'clickstream/iplookup/\*' USING PigStorage(',') AS (ip:chararray, country:chararray,state:chararray,city:chararray, lat:double, lng:double);

f1: {ip: chararray,country: chararray,state: chararray,city: chararray,lat: double,lng: double}

f2 = filter f1 by ip!='IP';

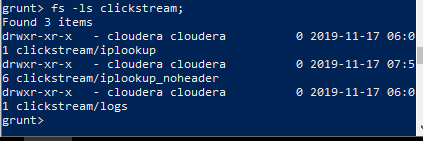


f3 = order f2 by ip asc;



store f3 into '/user/cloudera/clickstream/iplookup\_noheader' using PigStorage(',');

fs -ls clickstream;



* 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.

f4 = group f3 by state;

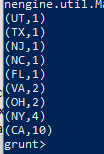
grunt> describe f4;

f4: {group: chararray,f3: {(ip: chararray,country: chararray,state: chararray,city: chararray,lat: double,lng: double)}}

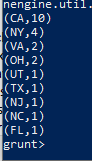
f5 = FOREACH f4 GENERATE group, COUNT(f3.ip) as counts;

f6 = order f5 by counts ASC;

dump f6;



f6 = order f5 by counts DESC;



* 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.

t1 = LOAD 'clickstream/logs/\*' USING PigStorage(' ') AS (reqdate:chararray, reqtime:chararray,x1:int, method:chararray, uri:chararray,x2:int, x3:int,x4:int,ipaddress:chararray, useragent:chararray);

grunt> describe t1;

t1: {reqdate: chararray,reqtime: chararray,x1: int,method: chararray,uri: chararray,x2: int,x3: int,x4: int,ipaddress: chararray,useragent: chararray}

t2 = FILTER t1 BY reqdate!='#';

grunt> t3 = FOREACH t2 GENERATE reqdate, reqtime,method,uri,ipaddress,useragent;

grunt> describe t3;

t3: {reqdate: chararray,reqtime: chararray,method: chararray,uri: chararray,ipaddress: chararray,useragent: chararray}

grunt> STORE t3 INTO '/user/cloudera/clickstream/logs\_noheader' USING PigStorage('\t');

fs -ls clickstream;

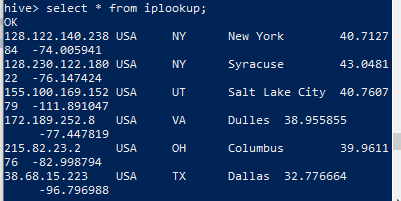
* Use hive to create two external tables for the **clickstream/logs\_noheader** and **clickstream/iplookup\_noheader** files you created in the previous steps. These tables should be named **weblogs** and **iplookup** respectively and should be placed in the **clickstream** database. Be sure to record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

create database clickstream;

show databases;

use clickstream;

CREATE EXTERNAL TABLE iplookup ( ip string, country string, state string, city string, lat double, lng double)ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/cloudera/clickstream/iplookup\_noheader/';



CREATE EXTERNAL TABLE weblogs (reqdate date, reqtime timestamp,method string, uri string, ipaddress string, useragent string)ROW FORMAT DELIMITED FIELDS TERMINATED BY "\t" LOCATION '/user/cloudera/clickstream/logs\_noheader/';

* 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.

select i.city, count(w.method) as num\_request from iplookup as i join weblogs w on i.ip=w.ipaddress group by i.city order by num\_request desc;

