Your Name: STEPHEN OMONDI  
Your SUID: 946934043  
Your Email: SOOMONDI@SYR.EDU  
Date Due: 2019-11-10  
Homework #: LABWORK 6

# 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.

**Silencing log-in feedback:**

Echo “log4j.rootloger=fatal” > nolog.conf

**Loading clickstream/iplookup**

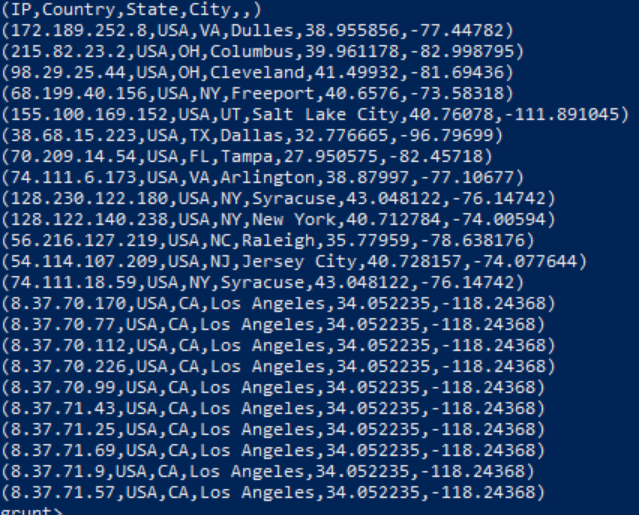
iplookupc = LOAD ‘clickstream/iplookup/\*’

USING PigStorage(‘,’)

AS(IP:chararray, Country:chararray, State:chararray, City:chararray, ApproxLat:float, ApproxLng:float);

**Verify output**

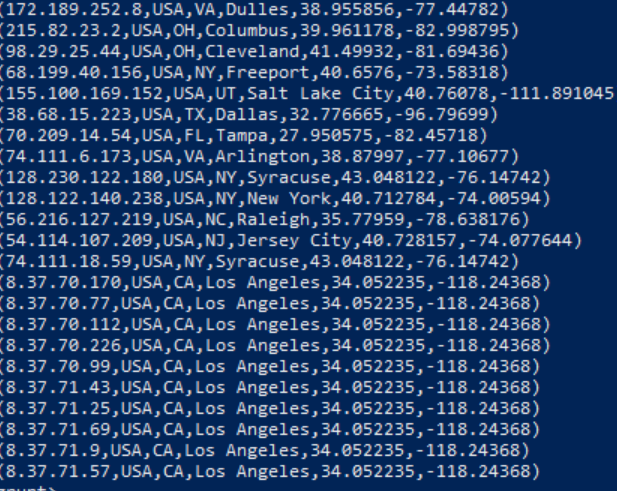
Dump iplookupc



**Filtering output**

iplookupcNoHeader = FILTER iplookupc BY IP !=’IP’;

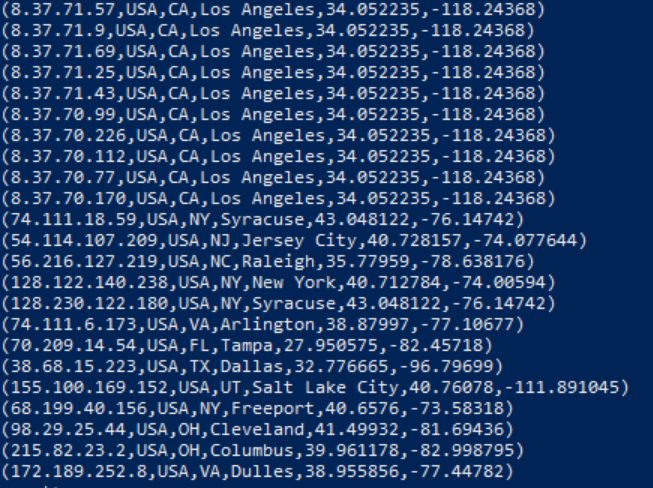
**Verify filtered output**



**Sort the output**

sort\_iplookupcNoHeader = ORDER iplookupcNoHeader BY $1 ASC;

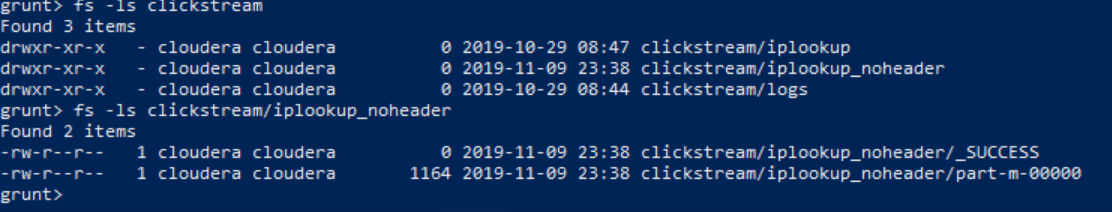
**Verify the sort**



**Dumping a csv into clickstream/iplookup\_noheader**

STORE sort\_iplookupcNoHeader INTO ‘clicksteram/iplookup\_noheader’ USING PigStorage(‘,’);

**Verify storage**



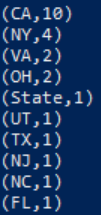
1. Write Pig commands to produce a count of IP Addresses by state codes, sorted by the count with highest values first, like this:  
   (CA, 10)  
   (NY, 4)  
   (VA, 2)  
   Etc…  
   Record all your Pig commands required to complete your transformation.

state\_group = GROUP iplookup BY State;

state\_count = FOREACH state\_group GENERATE group, COUNT(iplookup.IP);

sort\_state\_count = ORDER state\_count BY $1 DESC;

DUMP sort\_state\_count;



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.

log\_files = LOAD ‘clickstream/logs/\*’

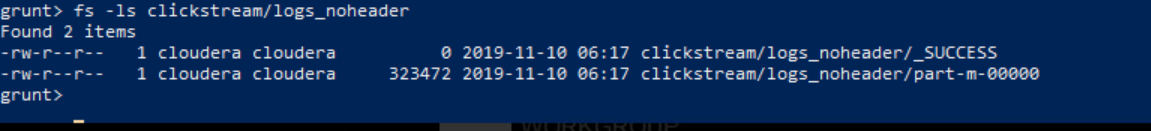
USING PigStorage(‘,’)

AS (reqdate:chararray, reqtime:chararray, x1:int, method:chararray, uri:chararray, x2:int ,x3:int, x4:int ,ipaddress:chararray, useragent:chararray);

filter\_header\_rows = FILTER log\_files BY reqdate != ‘#%’;

filter\_columns = FOREACH filter\_header\_rows GENERATE reqdate, reqtime, method, uri, ipaddress and useragent;

STORE filter\_columns INTO ‘clickstream/logs\_noheader’ USING PigStorage(‘\t’);



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.

CREATE EXTERNAL TABLE weblogs(

reqdate string,

reqtime string,

method string,

ipaddress string,

useragent string

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\t’

LOCATION ‘/user/cloudera/clickstream/logs\_noheader/’

CREATE EXTERNAL TABLE iplookup(

IP string,

Country string,

State string,

City string,

ApproxLat double,

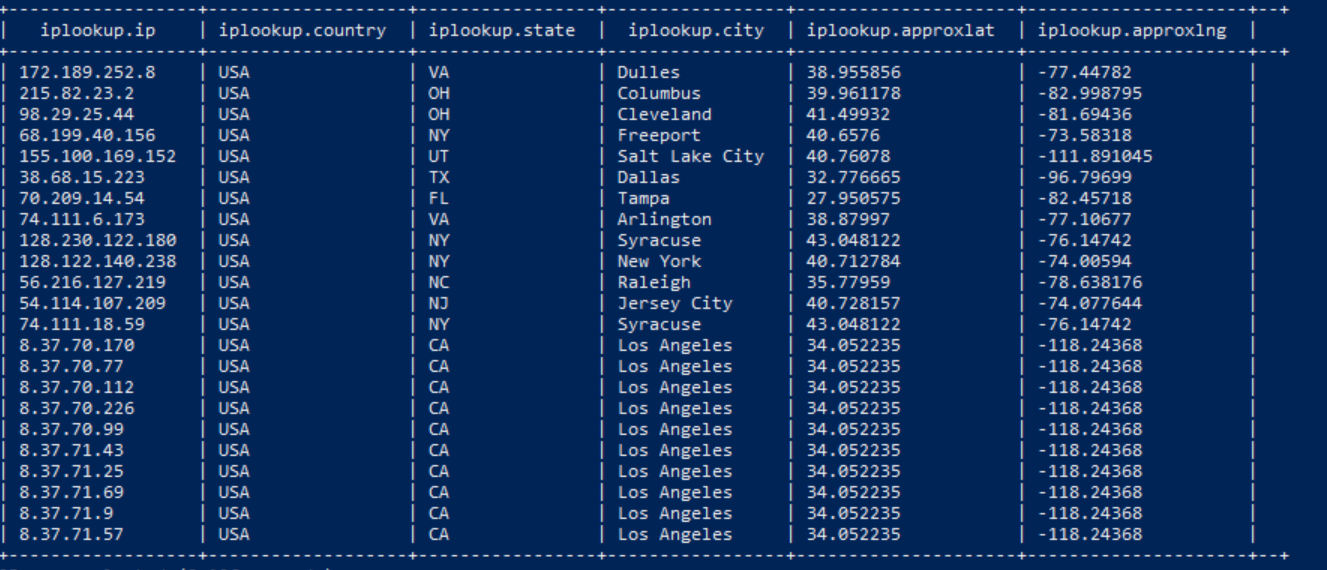
ApproxLng double

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\t’

LOCATION ‘/user/cloudera/clickstream/iplookup\_noheader/’;



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.

SELECT \* FROM (

SELECT city, COUNT(\*) AS NumRequests

FROM iplookup

INNER JOIN weblogs.ipaddress = iplookup.ip

GROUP BY city

) AS count\_a

SORT BY NumRequests DESC;