# IST769 Homework Submission Template

## Basic Information

Your Name: Wanyue XiaoYour

SUID: 720633297

Your Email: xwanyue@syr.edu

Date Due: March 11, 2020

Homework #: HW06

## Instructions

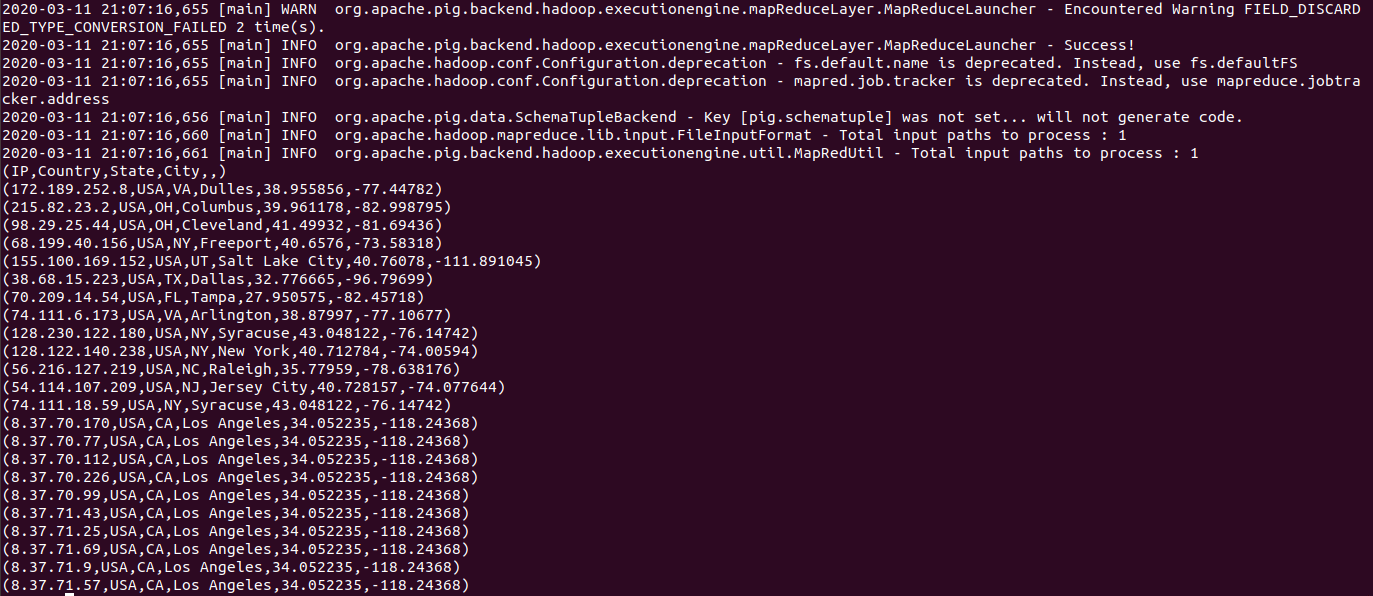
For each answer, please include your answer as text, and any screenshot(s) which demonstrate your answer was executed. Most importantly, make sure to include evidence your answer is correct. This will most likely be a screenshot. If you had issues, problems, or had to make assumptions include them in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

## Your Answers:

1.

Iplookup = LOAD 'clickstream/iplookup/\*' USING PigStorage(',') AS (IP:chararray, Country:chararray, State:chararray, City:chararray, ApproxLat: chararray, ApproxLng: chararray);

Dump Iplookup;



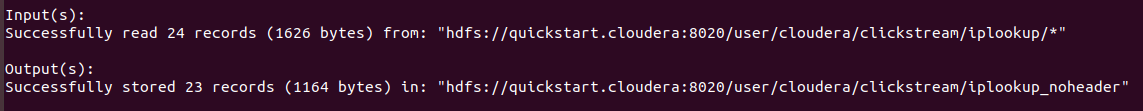
remove = FILTER Iplookup BY IP != ‘IP’;

dump remove;

byIP = **ORDER** remove **BY** IP **ASC;**

**DUMP byIP;**

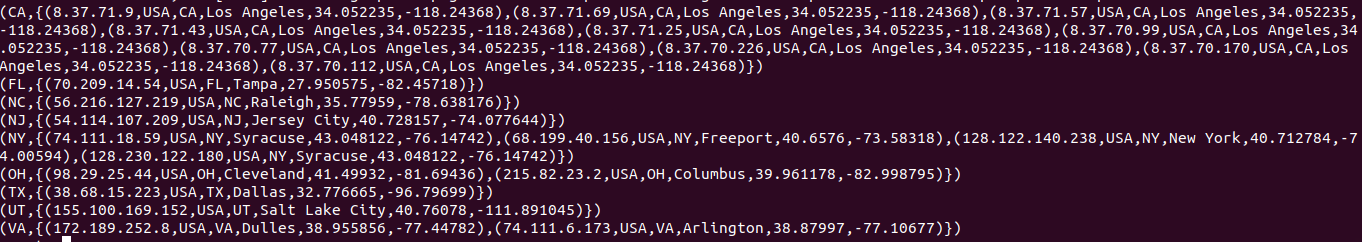
**STORE** byIP INTO ‘**clickstream/iplookup\_noheader**’ USING PigStorage(‘,’);



2.

IPbyState = GROUP byIP BY State;

DUMP IPbyState;



describe IPbyState;

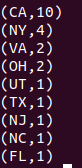


Result = **FOREACH** IPbyState **GENERATE** group, COUNT(byIP) as count;

DUMP Result;

Sort\_Result = ORDER Result BY count DESC;

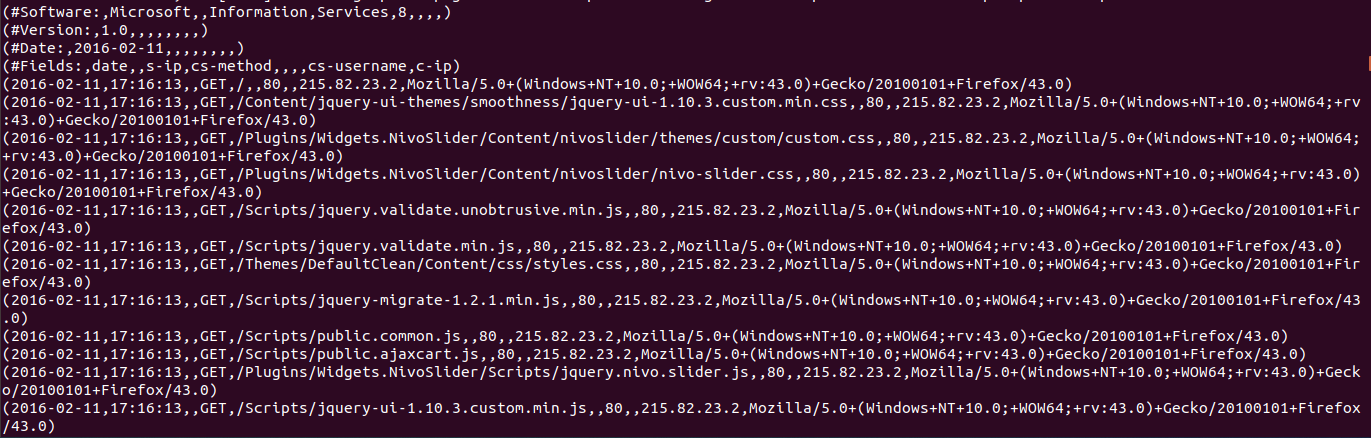
DUMP Sort\_Result;



3.

logs = 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);

dump logs;



remove = FILTER logs BY reqdate != '#Software:';

dump remove;

remove = FILTER remove BY reqdate != '#Version:';

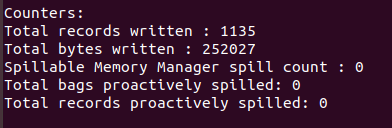
dump remove;

remove = FILTER remove BY reqdate != '#Date:';

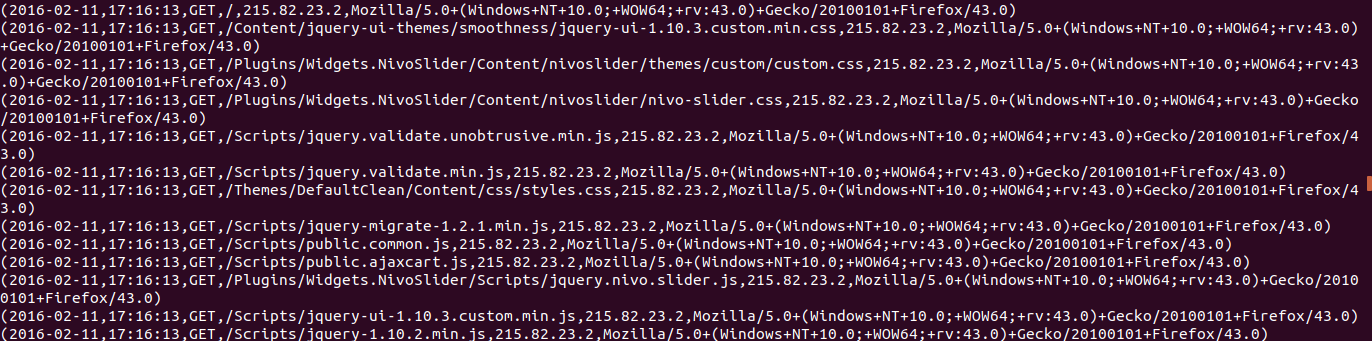
dump remove;

remove = FILTER remove BY reqdate != '#Fields:';

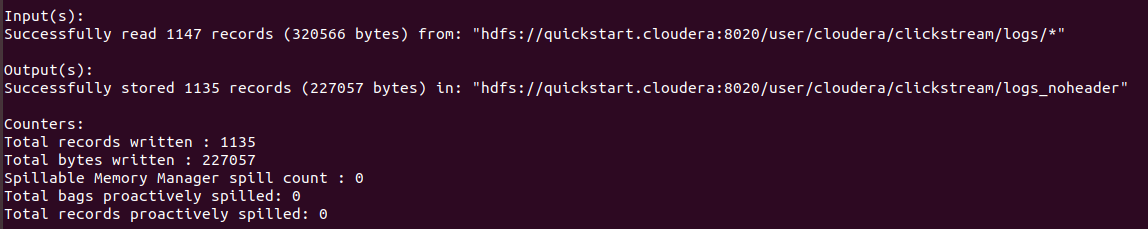
dump remove;



Result = FOREACH remove GENERATE reqdate, reqtime, method, uri, ipaddress, useragent;



STORE Result INTO 'clickstream/logs\_noheader' USING PigStorage('\t');



4.

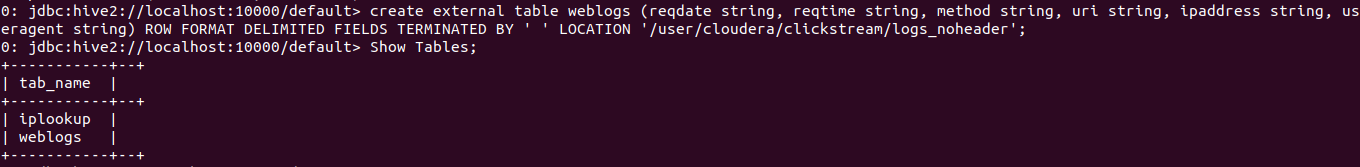
create database clickstream;

use clickstream;

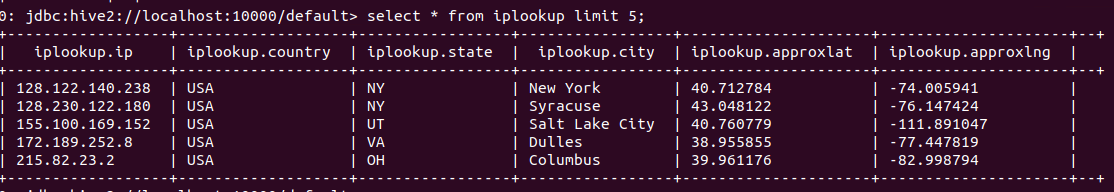
create external table iplookup (IP string, Country string, State string, City string, ApproxLat string, ApproxLng string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/cloudera/clickstream/iplookup\_noheader';

create external table weblogs (reqdate string, reqtime string, method string, uri string, ipaddress string, useragent string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LOCATION '/user/cloudera/clickstream/logs\_noheader';

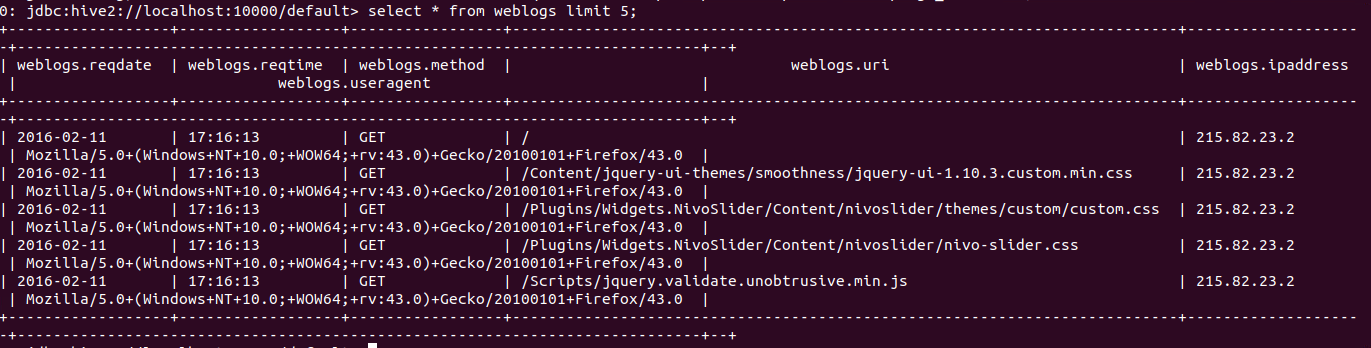
Show tables;



Select \* from iplookup limit 5;

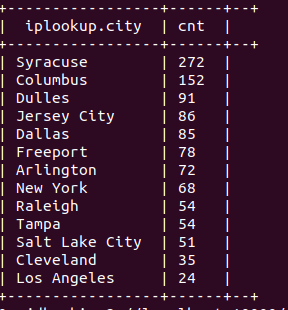


Select \* from weblogs limit 5;



5.

select iplookup.city, count(\*) as cnt from iplookup join weblogs on (iplookup.ip = weblogs.ipaddress) group by (iplookup.city) order by cnt DESC;



## Student Reflection:

1. What new information have I learned from completing this assignment?

Hadoop, Pig, Hive, HCatalog

1. What were my challenges / roadblocks I encountered on the way to completing it?

Operation of Linux System

1. To be better prepared to attempt this assignment I should \_\_\_\_\_\_\_ ?

Review

1. If I could make this assignment better, I would \_\_\_\_\_?

Google

1. Rate your comfort level with completing this assignment:

1.

1 ==> I can do this on my own and explain how to do it.

2 ==> I can do this on my own without any help.

3 ==> I can do this with help or guidance from others. If you choose this level, please list those who helped you.

4 ==> I don't understand this at all yet and need extra help. If you choose this, please try to articulate that which you do not understand.