IST769 Homework Submission Template

Basic Information

Your Name: Mark Roberts  
Your SUID: 598273961  
Your Email: mrober04@syr.edu  
Date Due: August 17, 2021   
Homework #: 6

Your Answers:

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.

|  |
| --- |
| CODE |
| ip\_lookup = LOAD 'clickstream/iplookup/\*' USING PigStorage(',') AS (IP:chararray, Country:chararray, State:chararray, City:chararray, ApproxLat:float, ApproxLng:float);  DUMP ip\_lookup;  Ip\_lookup\_rows = FILTER ip\_lookup by IP != 'IP';  DUMP ip\_lookup\_rows;  Ip\_lookup\_sorted = ORDER ip\_lookup\_rows BY IP;  DUMP ip\_lookup\_sorted;  STORE ip\_lookup\_sorted INTO 'clickstream/iplookup\_noheader' USING PigStorage(',');  fs -cat clickstream/iplookup\_noheader/\*; |
| SCREENSHOT/OUTPUT |
|  |

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.

|  |
| --- |
| CODE |
| ip\_lookup\_grouped = GROUP ip\_lookup\_sorted BY (State);  ip\_lookup\_count = FOREACH ip\_lookup\_grouped GENERATE group, COUNT(ip\_lookup\_sorted.IP) AS Count;  ip\_lookup\_count\_sorted = ORDER ip\_lookup\_count by Count DESC;  DUMP ip\_lookup\_count\_sorted; |
| SCREENSHOT/OUTPUT |
|  |

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.
2. 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.

|  |
| --- |
| CODE |
| 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);  DESCRIBE logs;  log\_rows = FILTER logs BY NOT STARTSWITH(reqdate, '#');  DUMP log\_rows;  log\_cols = FOREACH log\_rows GENERATE reqdate, reqtime, method, uri, ipaddress, useragent;  STORE log\_cols INTO 'clickstream/logs no header' USING PigStorage('\t');  fs -cat clickstream/logs\_noheader/\*; |
| SCREENSHOT/OUTPUT |
|  |

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.

|  |
| --- |
| CODE |
| show databases;  show tables;  describe web\_logs;  CREATE DATABASE IF NOT EXISTS clickstream;  USE clickstream;  show tables;  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/';  SELECT COUNT(\*) FROM weblogs;  CREATE EXTERNAL TABLE iplookup (IP string, Country string, State string, City string, ApproxLat float, ApproxLng float) ROW FORMATED DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/cloudera/clickstream/iplookup\_noheader/';  SELECT COUNT(\*) FROM iplookup;  SELECT City, COUNT(\*) AS NumberOfRequests FROM iplookup INNER JOIN weblogs on weblogs.ipaddress = iplookup.IP GROUP BY City ORDER BY NumberOfRequests DESC; |
| SCREENSHOT/OUTPUT |
|  |