IST769 Homework Submission Template

Basic Information

Your Name: Mark Roberts  
Your SUID: 598273961  
Your Email: mrober04@syr.edu  
Date Due: August 24, 2021   
Homework #: 7

Your Answers:

1. From Impala, use the two external tables **weblogs** created from **clickstream/logs\_noheader** and **iplookup** created from **clickstream/iplookup\_noheader** you created in the previous assignment to complete this question. Use the impala shell to answer the following questions, making sure to include the SELECT query you used to answer it.
   1. How many GET and POST requests are there in the weblogs?
   2. How many requests have Mac in the user agent?

How many hosts (ip addresses) have Mac in the user agent?

|  |
| --- |
| CODE |
| impala-shell  CONNECT localhost;  SHOW databases;  USE clickstream;  show tables;  a.  SELECT COUNT(\*) FROM weblogs WHERE method IN ('GET', 'POST');  SELECT COUNT(\*) FROM weblogs WHERE method NOT IN ('GET', 'POST');  SELECT method, COUNT(\*) FROM weblogs GROUP BY method;  b.  SELECT COUNT(\*) FROM weblogs WHERE useragent LIKE '%Mac%';  SELECT COUNT(\*) FROM weblogs WHERE useragent NOT LIKE '%Mac%';  SELECT COUNT(\*) FROM weblogs WHERE useragent LIKE '%mac%';  c.  SELECT COUNT(DISTINCT IP)  FROM weblogs AS WEB  INNER JOIN iplookup AS IP ON IP.IP = WEB.ipaddress  WHERE useragent LIKE '%Mac%'; |
| SCREENSHOT/OUTPUT |
|  |
|  |

1. From the HBase shell, include the commands required to complete the following.
   1. Create a table named **computers** with column family **info**.
   2. Issue HBase commands to write the following data to the table in the column family:

|  |  |  |  |
| --- | --- | --- | --- |
| Computer ID | Model | GB\_Ram | TB\_Disk |
| 1 | Dell | 16 | 1 |
| 2 | IBM | 32 | 1.5 |
| 3 | HP | 8 | 1 |
| 4 | Acer | 16 | 2 |

|  |
| --- |
| CODE |
|  |
| SCREENSHOT/OUTPUT |
|  |

1. From the Hive shell, write an HQL statement to create an external Hive table from the HBase **computers** table. Then write a hive query to add up the total ram and disk across all computers. Your answer should include all HQL statements.

|  |
| --- |
| CODE |
| beeline -u jdbc:hive2://localhost:10000/default -n cloudera -p cloudera --silent=true  USE clickstream;  CREATE EXTERNAL TABLE computers  (  id int,  model string,  ram int,  disk float  )  STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'  WITH  SERDEPROPERTIES ('hbase.columns.mapping'=':key,info:model,info:ram,info:disk')  TBLPROPERTIES ('hbase.table.name'='computers');  SELECT \* FROM computers; |
| SCREENSHOT/OUTPUT |
|  |

1. Use Hive to load the **iplookup** table you created from **clickstream/iplookup\_noheader** into and HBase table, with IP address as key. Include the HQL Queries you wrote to make the table and load the data as the answer to your question.

|  |
| --- |
| CODE |
| USE clickstream;  CREATE TABLE iplookup\_hbase  (  IP string,  Country string,  State string,  City string,  ApproxLat float,  ApproxLng float  )  STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'  WITH  SERDEPROPERTIES ('hbase.columns.mapping'=':key,ip:Country,ip:State,ip:City,ip:ApproxLat,ip:ApproxLng')  TBLPROPERTIES ('hbase.table.name'='iplookup\_hbase', 'hbase.mapred.output.outputtable'='iplookup\_hbase');  INSERT OVERWRITE TABLE iplookup\_hbase SELECT \* FROM iplookup;  scan 'iplookup\_hbase' |
| SCREENSHOT/OUTPUT |
|  |

1. From the HBase shell, write an HBase query to retrieve the city and state columns for all rows in the **iplookup** table.

|  |
| --- |
| CODE |
| hbase shell  scan 'iplookup\_hbase', {COLUMNS=>['ip:City','ip:State']} |
| SCREENSHOT/OUTPUT |
|  |