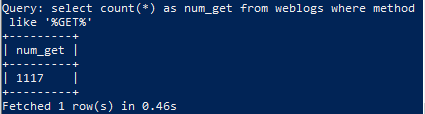
IST769 Impala, HBase and HCatalog

Your Name: Tajudeen Abdulazeez   
Your SUID: 69687-7373-0   
Your Email: toabdula@syr.edu   
Date Due:   
Homework #: 7

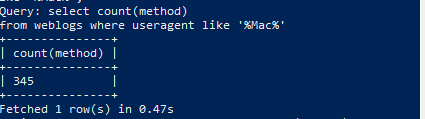
* From Impala, use the two external tables **weblogs** created from **clickstream/logs\_noheader** and **iplookup** created from **clickstream/iplookup\_noheader** you created in the previous assignment to complete this question. Use the impala shell to answer the following questions, making sure to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

select count(\*) as num\_get from weblogs where method like '%GET%';



* How many requests have Mac in the user agent?

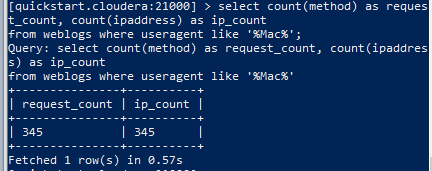
select count(method) from weblogs where useragent like '%Mac%';



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

select count(method) as request\_count, count(ipaddress) as ip\_count

from weblogs where useragent like '%Mac%';



* From the HBase shell, include the commands required to complete the following.
* Create a table named **computers** with column family **info**.

create "computers", "info" ;

* Issue HBase commands to write the following data to the table in the column family:

put "computers", 1,"info:Model", "Dell"

put "computers", 1,"info:GB\_Ram", 16

put "computers", 1,"info:TB\_Disk", 1

put "computers", 2,"info:Model", "IBM"

put "computers", 2,"info:GB\_Ram", 32

put "computers", 2,"info:TB\_Disk", 1.5

put "computers", 3,"info:Model", "HP"

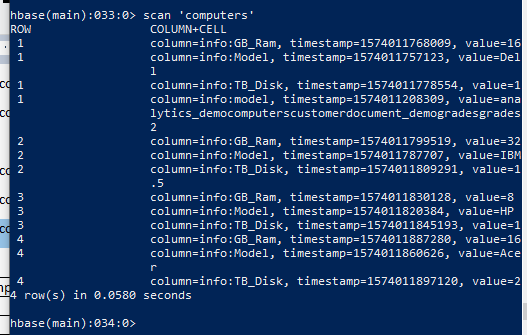
put "computers", 3,"info:GB\_Ram", 8

put "computers", 3,"info:TB\_Disk", 1

put "computers", 4,"info:Model", "Acer"

put "computers", 4,"info:GB\_Ram", 16

put "computers", 4,"info:TB\_Disk", 2



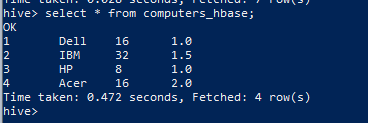
|  |  |  |  |
| --- | --- | --- | --- |
| Computer ID | Model | GB\_Ram | TB\_Disk |
| 1 | Dell | 16 | 1 |
| 2 | IBM | 32 | 1.5 |
| 3 | HP | 8 | 1 |
| 4 | Acer | 16 | 2 |

* 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.

create external table computers\_hbase (computer\_id int, model string, gb\_ram int,

tb\_disk double) stored by 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key,info:Model,info:GB\_Ram,info:TB\_Disk")

TBLPROPERTIES ("hbase.table.name"="computers");

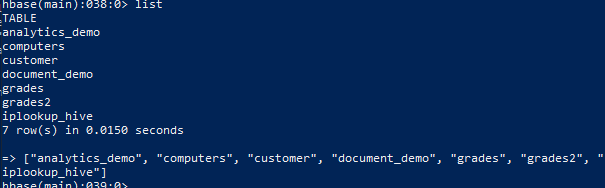


select sum(GB\_Ram) as total\_ram, sum(TB\_Disk) as total\_disk from computers\_hbase;

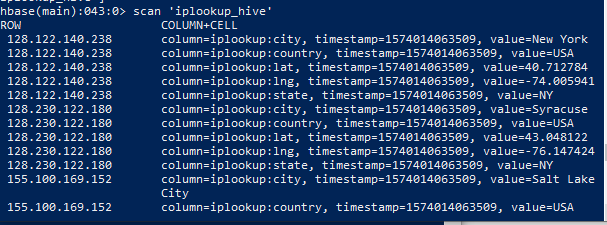


* Use Hive to load the **iplookup** table you created from **clickstream/iplookup\_noheader** into and HBase table, with IP address as key. Include the HQL Queries you wrote to make the table and load the data as the answer to your question.

create table iplookup\_hbase(ip string,country string,state string, city string, lat double, lng double) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH SERDEPROPERTIES ("hbase.columns.mapping"=":key,iplookup:country,iplookup:state,iplookup:city,iplooku p:lat, iplookup:lng") TBLPROPERTIES("hbase.table.name"="iplookup\_hive");



insert into iplookup\_hbase select \* from iplookup;



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

scan 'iplookup\_hive',{COLUMNS=>['iplookup:city','iplookup:state']}

