INFSCI 2750: Cloud Computing

Mini Project 3

Jiamin Cheng (jic122@pitt.edu)

Zherui Cao ([zhc61@pitt.edu)](mailto:zhc61@pitt.edu))

Part 1: Setting up Cassandra: (50 points)

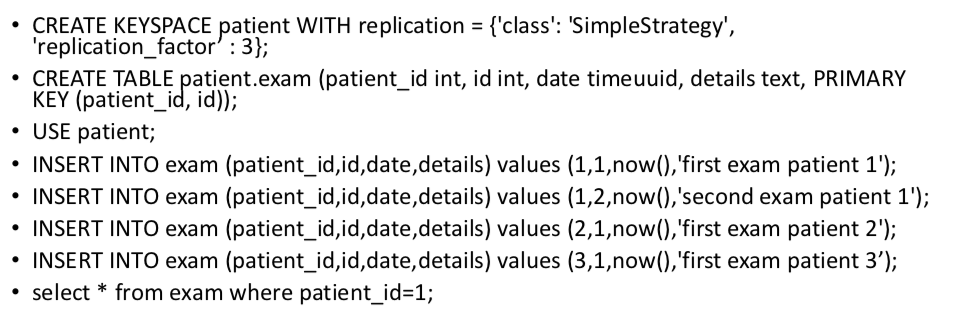
Firstly, we started cassandra on both machines.(listening 7199)



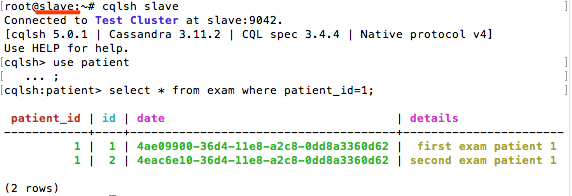


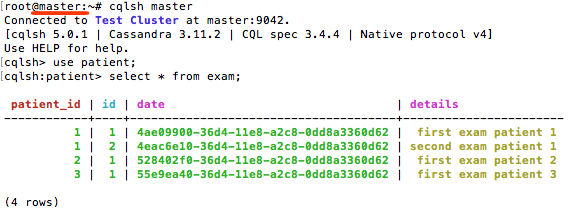
After excuting the

CQL test



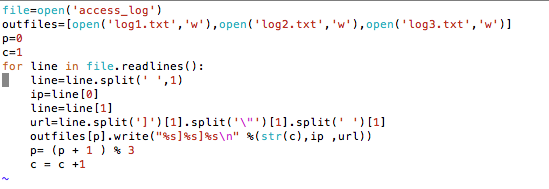
we can manipulate data from master or slave.





Part 2: Import Data into Cassandra (25 points)

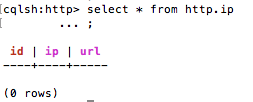
we divide the access\_log into three files so that we can use copy to import data. ( using  as delimiter because no  will be in the log files we produce)



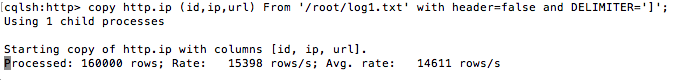
then we create keyspace and table

|  |
| --- |
| cqlsh> CREATE KEYSPACE http WITH replication = {'class': 'SimpleStrategy', 'replication\_factor' : 3};  cqlsh> use http;  cqlsh:http> create table http.ip (id text Primary key,ip text,url text);  cqlsh:http> copy http.ip (id,ip,url) From '/root/log1.txt' with header=false and DELIMITER=']';  cqlsh:http> copy http.ip (id,ip,url) From '/root/log2.txt' with header=false and DELIMITER=']';  cqlsh:http> copy http.ip (id,ip,url) From '/root/log3.txt' with header=false and DELIMITER=']'; |

select from the empty table



Importing data



Then we select again

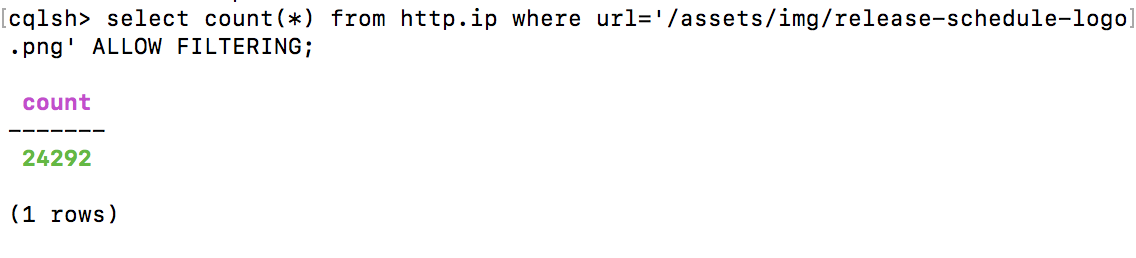


Part 3: Operate Data in Cassandra

The operation count is quite costly so we need to modify the read\_request\_timeout\_in\_ms to 30000 and request\_timeout\_in\_ms to 200000 and reboot the server the activate the configuration.Or it will give an readtimeout error.

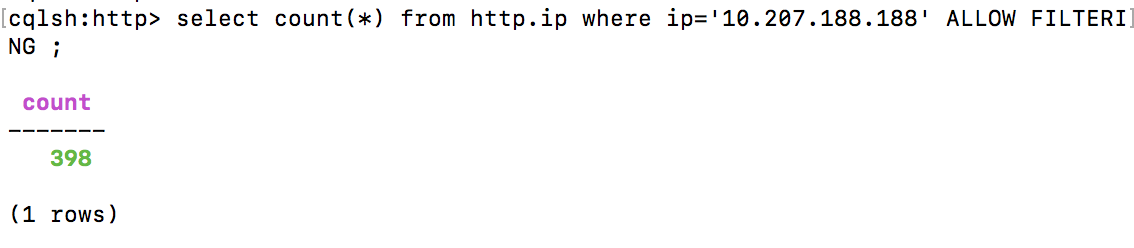
1.How many hits were made to the website item “/assets/img/release-schedule- logo.png”?

select count(\*) from http.ip where url='/assets/img/release-schedule-logo.png' ALLOW FILTERING;



2. How many hits were made from the IP: 10.207.188.188

select count(\*) from http.ip where ip='10.207.188.188' ALLOW FILTERING ;



3. Which path in the website has been hit most? How many hits were made to the path?

Since Cassandra do not have groupby function, so we first define functions:

CREATE OR REPLACE FUNCTION state\_group\_and\_count( state map<text, int>, type text )

CALLED ON NULL INPUT

RETURNS map<text, int>

LANGUAGE java AS '

Integer count = (Integer) state.get(type);

if (count == null)

count = 1;

else count++;

state.put(type, count);

return state; ' ;

CREATE OR REPLACE AGGREGATE group\_and\_count(text)

SFUNC state\_group\_and\_count

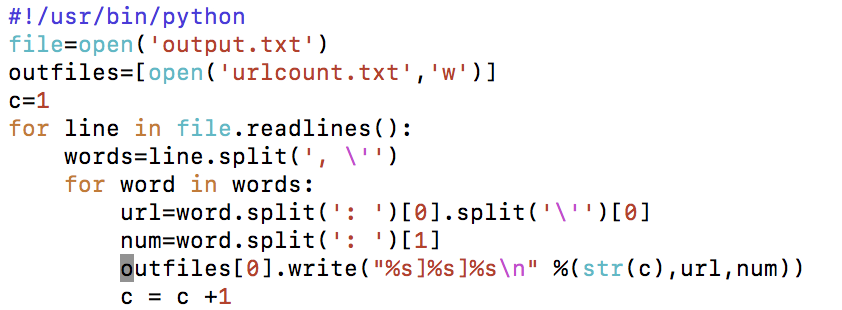
STYPE map<text, int>

INITCOND {};

Then we use the user-define functions to get the url-count pairs and output it in a external file:

cqlsh -e'SELECT group\_and\_count(url) from http.ip' --request-time 600 master> /root/output.txt

Then we we python to modify the output to import the data into Cassandra:



To find the url with most hits:

We define another function:

CREATE FUNCTION maxI(current int, candidate int)

CALLED ON NULL INPUT

RETURNS int LANGUAGE java AS

'if (current == null) return candidate; else return Math.max(current, candidate);' ;

CREATE AGGREGATE maxAgg(int)

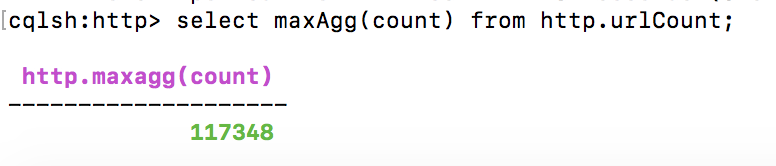
SFUNC maxI

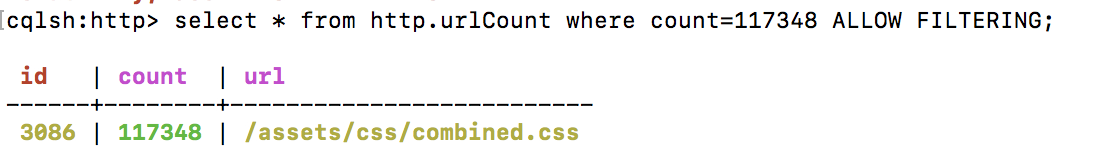
STYPE int

INITCOND null;

cqlsh:http> create table http.urlCount (id text Primary key,url text,count id);

copy http.urlCount(id,url,count) from '/root/urlcount.txt' with header=false and delimiter=']';





So /assets/css/combined.css has the most hits and the number is 117348.

4. Which IP accesses the website most? How many accesses were made by it?

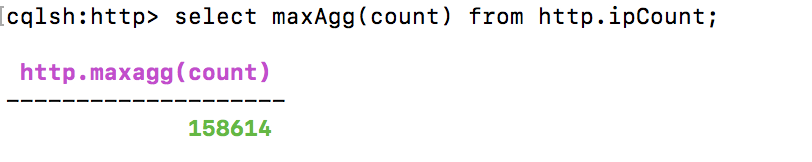
We can use the function defined in $3:

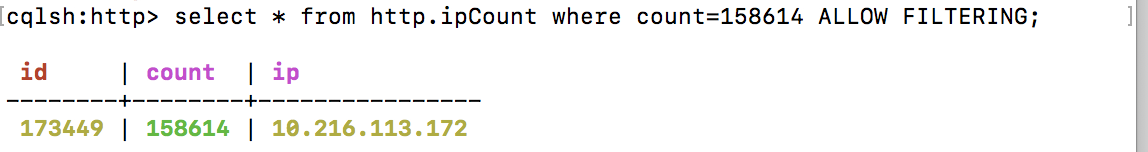
cqlsh -e'SELECT group\_and\_count(ip) from http.ip' --request-time 600 master> /root/output1.txt

After modified the data,we import the data:

cqlsh:http> create table http.ipCount(id text Primary key,ip text,count int);

copy http.ipCount(id,ip,count) from '/root/urlcount1.txt' with header=false and delimiter=']';





So the ip 10.216.113.172 give the most hits and the number is 158614.