INFSCI 2750 Miniproject 3

By Jing Pang (jip45@pitt.edu), Tian Xue (tix20@pitt.edu), Haoyang Qian (haq13@pitt.edu)

Part 1: Setting Up Cassandra

Setting up Cassandra on a single node on ubuntu linux with following instructions.

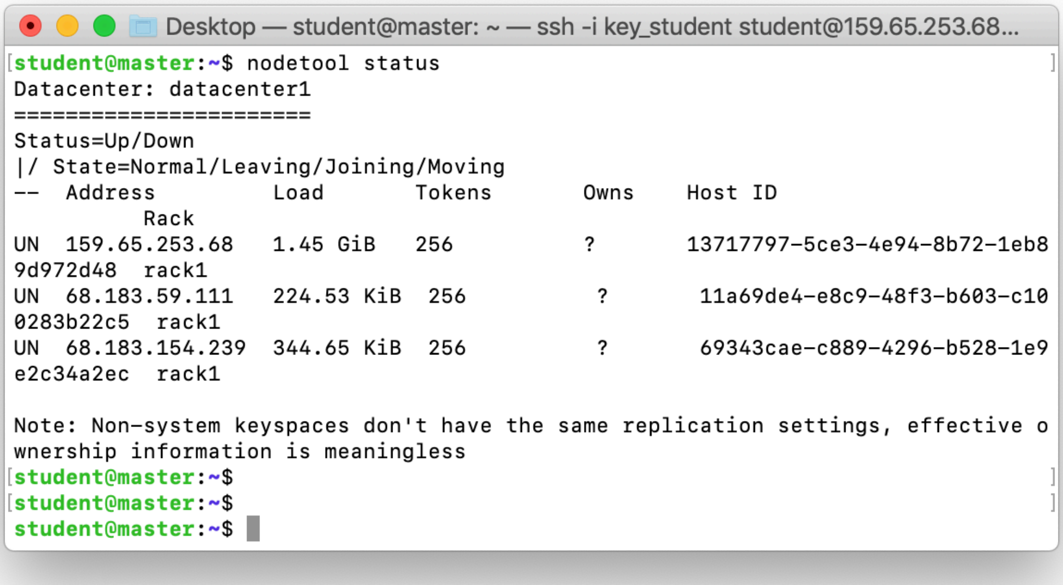
```
# install Cassandra
echo "deb http://www.apache.org/dist/cassandra/debian 311x main" | sudo tee -a
/etc/apt/sources.list.d/cassandra.sources.list
curl https://www.apache.org/dist/cassandra/KEYS | sudo apt-key add-
sudo apt-get update
sudo apt-get install cassandra
```

Change configure on all nodes

```
# read file
sudo nano /etc/cassandra/cassandra.yaml
# change file setting
- seeds: "master, slave-1, slave-2" (on all nodes)
listen_address: master (on master node)
listen_address: slave-1 (on slave-1 node)
listen_address: slave-2 (on slave-2 node)

rpc_address: master (on master node)
rpc_address: slave-1 (on slave-1 node)
rpc_address: slave-2 (on slave-2 node)
```

Then start the services on all nodes

```
# stop cassandra on all nodes
sudo service cassandra stop
# run cassandra
sudo cassandra -Rf
# check status
```



Desktop — student@master: ~ — ssh -i key_student student@159.65.253.68...

```
student@master:~$ nodetool status
Datacenter: datacenter1
=======================
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address         Load       Tokens       Owns    Host ID
          Rack
UN  159.65.253.68   1.45 GiB   256          ?       13717797-5ce3-4e94-8b72-1eb8
9d972d48  rack1
UN  68.183.59.111   224.53 KiB  256          ?       11a69de4-e8c9-48f3-b603-c10
0283b22c5  rack1
UN  68.183.154.239  344.65 KiB  256          ?       69343cae-c889-4296-b528-1e9
e2c34a2ec  rack1

Note: Non-system keyspaces don't have the same replication settings, effective o
wnership information is meaningless
student@master:~$
student@master:~$
student@master:~$
```

Part 2: Import Data into Cassandra

Test: Start CQL shell to see the correct setup

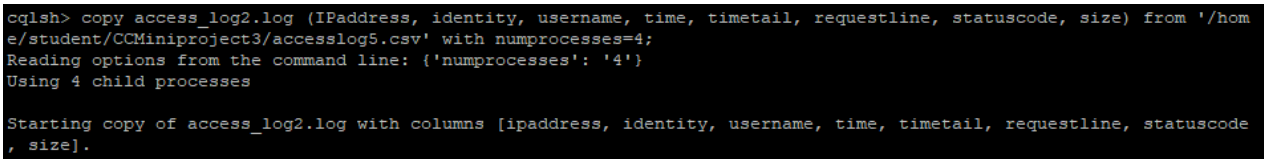
```
cqlsh master --request-timeout=600000
```

Before we upload the access_log file, we preprocess the file with several steps. We splited file into 5 pieces, and transferred each piece into a csv format. Then, we combined this set of files into a single file. At last, we upload this file to master node.

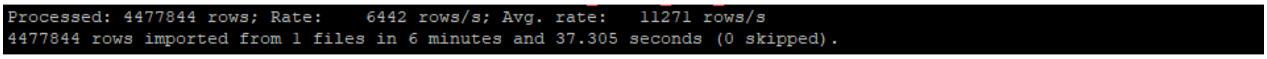
```
# upload file from local
scp -i key_student /Users/pangjing/Desktop/ccmini3/accesslog5.csv
student@159.65.253.68:~/CCMiniproject3/
```

Then, we login to the CQL shell, and use the COPY command to upload data into a table

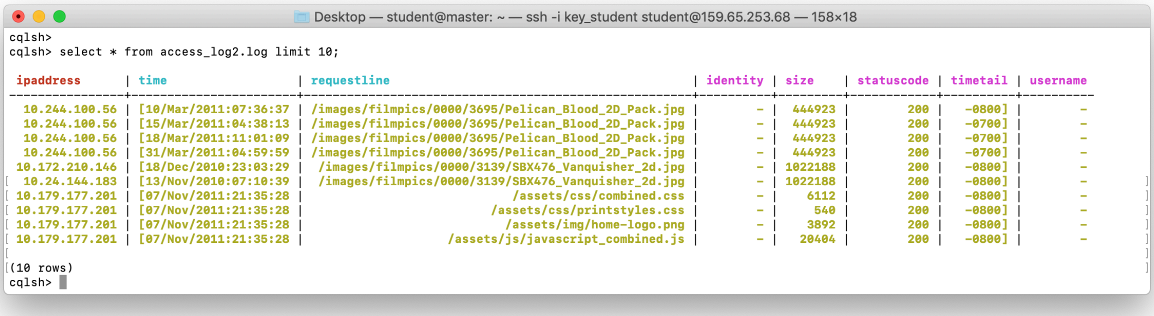
```
# login
cqlsh master --request-timeout=600000
# create keyspace
create keyspace access_log2
with replication = {
'class' : 'NetworkTopologyStrategy',
'datacenter1' : 1
};

# create table
create table access_log2.log (
IPaddress text,
identity text,
username text,
time text,
timetail text,
requestline text,
statuscode text,
size text,
primary key (IPaddress, time, requestline)
);

# copy data to table
COPY access_log2.log (IPaddress, identity, username, time, timetail,
requestline, statuscode, size)
FROM '/home/student/CCMiniproject3/accesslog5.csv' WITH numprocesses=4;
```



```
cqlsh> copy access_log2.log (IPaddress, identity, username, time, timetail, requestline, statuscode, size) from '/hom
e/student/CCMiniproject3/accesslog5.csv' with numprocesses=4;
Reading options from the command line: {'numprocesses': '4'}
Using 4 child processes

Starting copy of access_log2.log with columns [ipaddress, identity, username, time, timetail, requestline, statuscode
, size].
```



```
Processed: 4477844 rows; Rate:    6442 rows/s; Avg. rate:   11271 rows/s
4477844 rows imported from 1 files in 6 minutes and 37.305 seconds (0 skipped).
```



Desktop — student@master: ~ — ssh -i key_student student@159.65.253.68 — 158×18

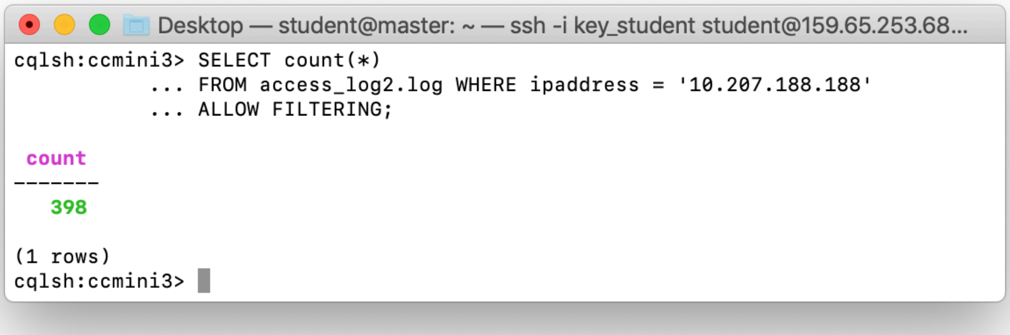
```
cqlsh>
cqlsh> select * from access_log2.log limit 10;

 ipaddress      | time                   | requestline                                            | identity | size    | statuscode | timetail | username
----------------+------------------------+--------------------------------------------------------+----------+---------+------------+----------+----------
  10.244.100.56 | [10/Mar/2011:07:36:37 | /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg |        - |  444923 |        200 |   -0800] |        -
  10.244.100.56 | [15/Mar/2011:04:38:13 | /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg |        - |  444923 |        200 |   -0700] |        -
  10.244.100.56 | [18/Mar/2011:11:01:09 | /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg |        - |  444923 |        200 |   -0700] |        -
  10.244.100.56 | [31/Mar/2011:04:59:59 | /images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg |        - |  444923 |        200 |   -0700] |        -
 10.172.210.146 | [18/Dec/2010:23:03:29 |  /images/filmpics/0000/3139/SBX476_Vanquisher_2d.jpg |        - | 1022188 |        200 |   -0800] |        -
  10.24.144.183 | [13/Nov/2010:07:10:39 |  /images/filmpics/0000/3139/SBX476_Vanquisher_2d.jpg |        - | 1022188 |        200 |   -0800] |        -
 10.179.177.201 | [07/Nov/2011:21:35:28 |                              /assets/css/combined.css |        - |    6112 |        200 |   -0800] |        -
 10.179.177.201 | [07/Nov/2011:21:35:28 |                           /assets/css/printstyles.css |        - |     540 |        200 |   -0800] |        -
 10.179.177.201 | [07/Nov/2011:21:35:28 |                             /assets/img/home-logo.png |        - |    3892 |        200 |   -0800] |        -
 10.179.177.201 | [07/Nov/2011:21:35:28 |                     /assets/js/javascript_combined.js |        - |   20404 |        200 |   -0800] |        -

(10 rows)
cqlsh>
```

Part 3: Operate Data in Cassandra

Solve problems with Cassandra

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

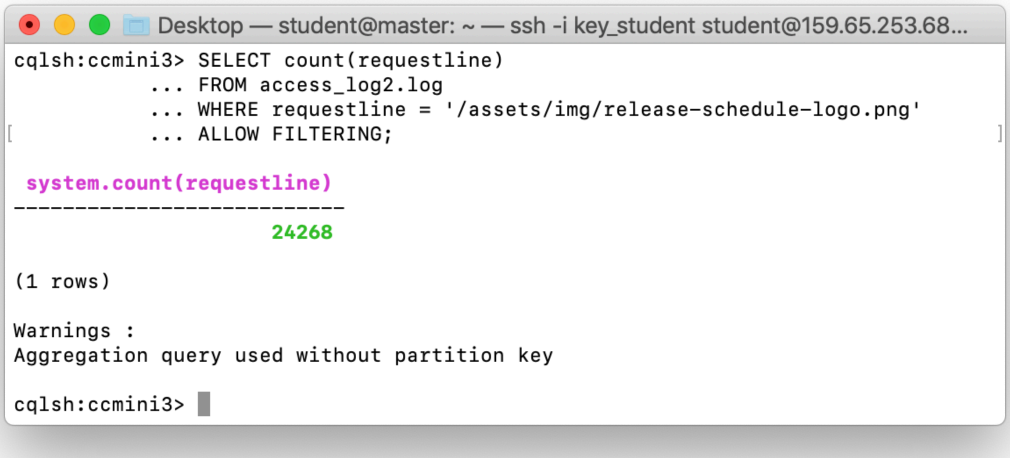
```
SELECT count(*)
FROM access_log2.log
WHERE requestline = '/assets/img/release-schedule-logo.png'
ALLOW FILTERING;
```


Desktop — student@master: ~ — ssh -i key_student student@159.65.253.68...

```
cqlsh:ccmini3> SELECT count(*)
           ... FROM access_log2.log WHERE ipaddress = '10.207.188.188'
           ... ALLOW FILTERING;

 count
-------
   398

(1 rows)
cqlsh:ccmini3>
```

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

```
SELECT count(*)
FROM access_log2.log WHERE ipaddress = '10.207.188.188'
ALLOW FILTERING;
```


Desktop — student@master: ~ — ssh -i key_student student@159.65.253.68...

```
cqlsh:ccmini3> SELECT count(requestline)
           ... FROM access_log2.log
           ... WHERE requestline = '/assets/img/release-schedule-logo.png'
           ... ALLOW FILTERING;

 system.count(requestline)
---------------------------
                     24268

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:ccmini3>
```

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

   Construct a new csv file containning the path and counts by using the Java program. Upload this new csv file to master node.

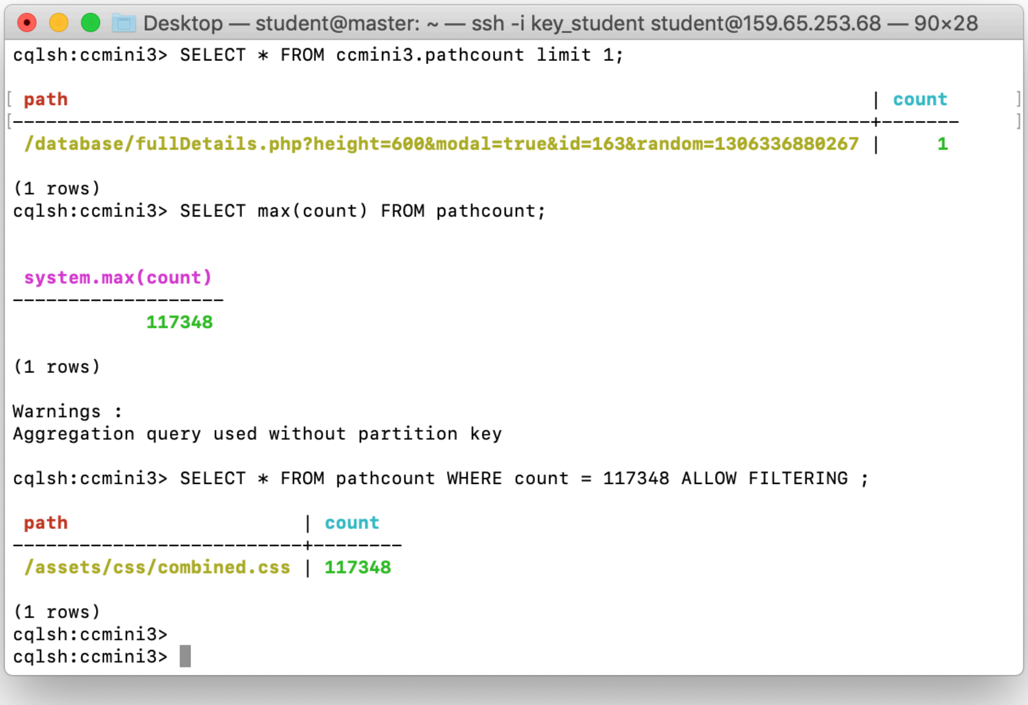
```
# upload from local computer to master node
scp -i key_student /Users/pangjing/IdeaProjects/CCMini3/out/pathcount.csv
student@159.65.253.68:~/ccmini3/
```

   Import csv file to cassandra

```
# create keyspace
CREATE KEYSPACE ccmini3 WITH replication = {'class': 'SimpleStrategy',
'replication_factor' : 3};
# create table
CREATE TABLE pathcount (path text, count int, PRIMARY KEY (path, count) )
;
# import data
COPY pathcount(path,count) FROM '~/ccmini3/pathcount.csv' WITH DELIMITER =
' ' AND HEADER = TRUE;
```

Run calculation to find match in CQL shell

```
SELECT * FROM ccmini3.pathcount limit 1;
SELECT max(count) FROM pathcount;
SELECT * FROM pathcount WHERE count = 117348 ALLOW FILTERING ;
```



Desktop — student@master: ~ — ssh -i key_student student@159.65.253.68 — 90×28

```
cqlsh:ccmini3> SELECT * FROM ccmini3.pathcount limit 1;

 path                                                                        | count
-----------------------------------------------------------------------------+-------
 /database/fullDetails.php?height=600&modal=true&id=163&random=1306336880267 |     1

(1 rows)
cqlsh:ccmini3> SELECT max(count) FROM pathcount;


 system.max(count)
-------------------
            117348

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:ccmini3> SELECT * FROM pathcount WHERE count = 117348 ALLOW FILTERING ;

 path                       | count
----------------------------+--------
 /assets/css/combined.css | 117348

(1 rows)
cqlsh:ccmini3>
cqlsh:ccmini3>
```

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

   Construct a new csv file containinng the path and counts by using the Java program. Upload this new csv file to master node.

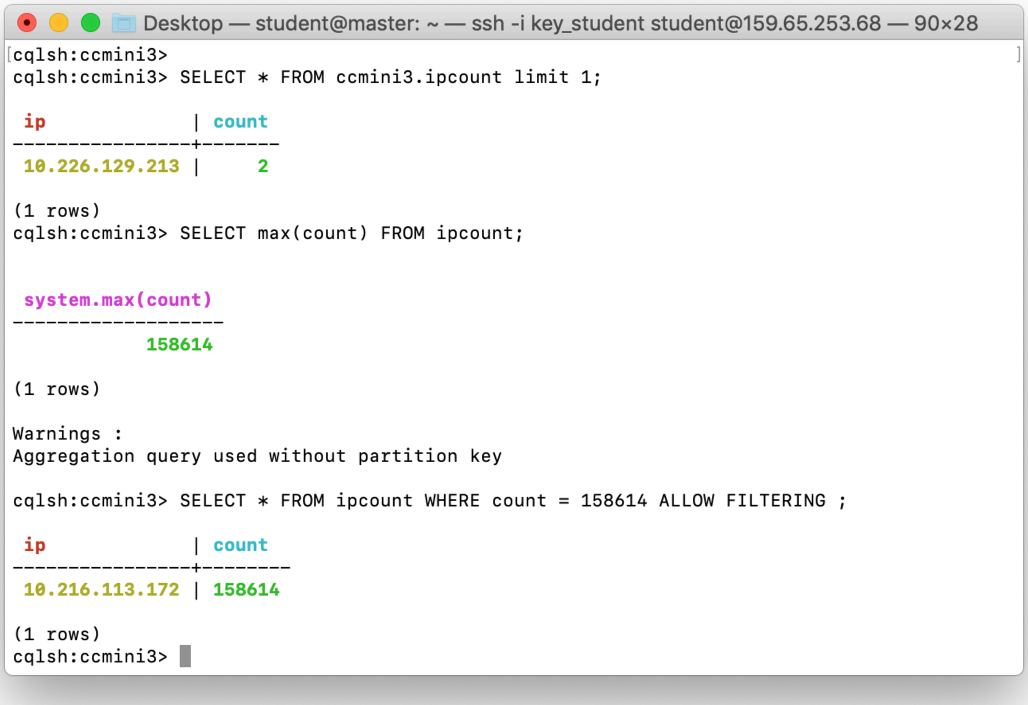
```
# upload from local computer to master node
scp -i key_student /Users/pangjing/IdeaProjects/CCMini3/out/ipcount.csv
student@159.65.253.68:~/ccmini3/
```

Import csv file to cassandra

```
# create table
CREATE TABLE ipcount (ip text, count int, PRIMARY KEY (ip, count) ) ;
# import data
COPY ipcount(ip,count) FROM '~/ccmini3/ipcount.csv' WITH DELIMITER = ' '
AND HEADER = TRUE;
```

Run calculation to find match in CQL shell

```
SELECT * FROM ccmini3.ipcount limit 1;
SELECT max(count) FROM ipcount;
SELECT * FROM ipcount WHERE count = 158614 ALLOW FILTERING ;
```



Desktop — student@master: ~ — ssh -i key_student student@159.65.253.68 — 90×28

```
cqlsh:ccmini3>
cqlsh:ccmini3> SELECT * FROM ccmini3.ipcount limit 1;

 ip             | count
----------------+-------
 10.226.129.213 |     2

(1 rows)
cqlsh:ccmini3> SELECT max(count) FROM ipcount;


 system.max(count)
-------------------
            158614

(1 rows)

Warnings :
Aggregation query used without partition key

cqlsh:ccmini3> SELECT * FROM ipcount WHERE count = 158614 ALLOW FILTERING ;

 ip             | count
----------------+--------
 10.216.113.172 | 158614

(1 rows)
cqlsh:ccmini3> 
```