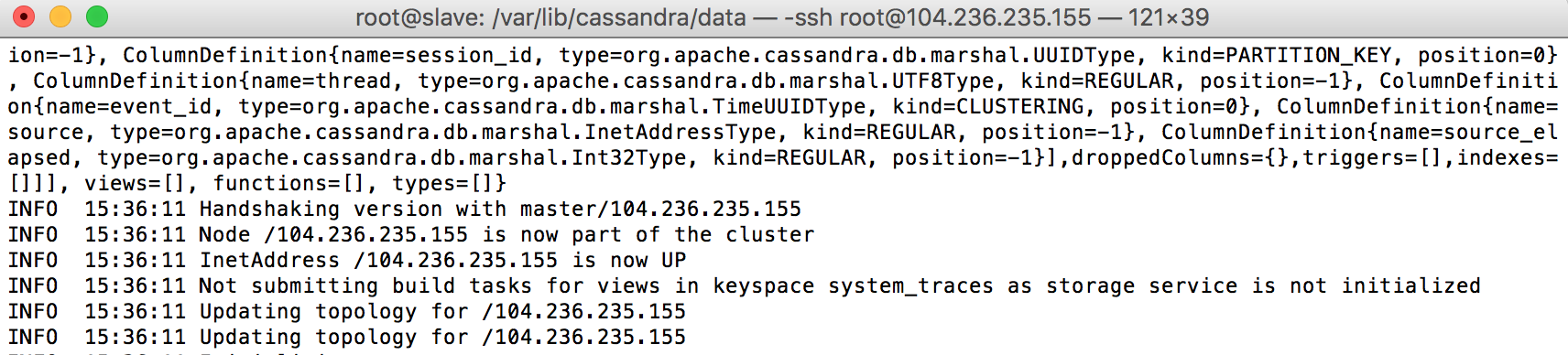
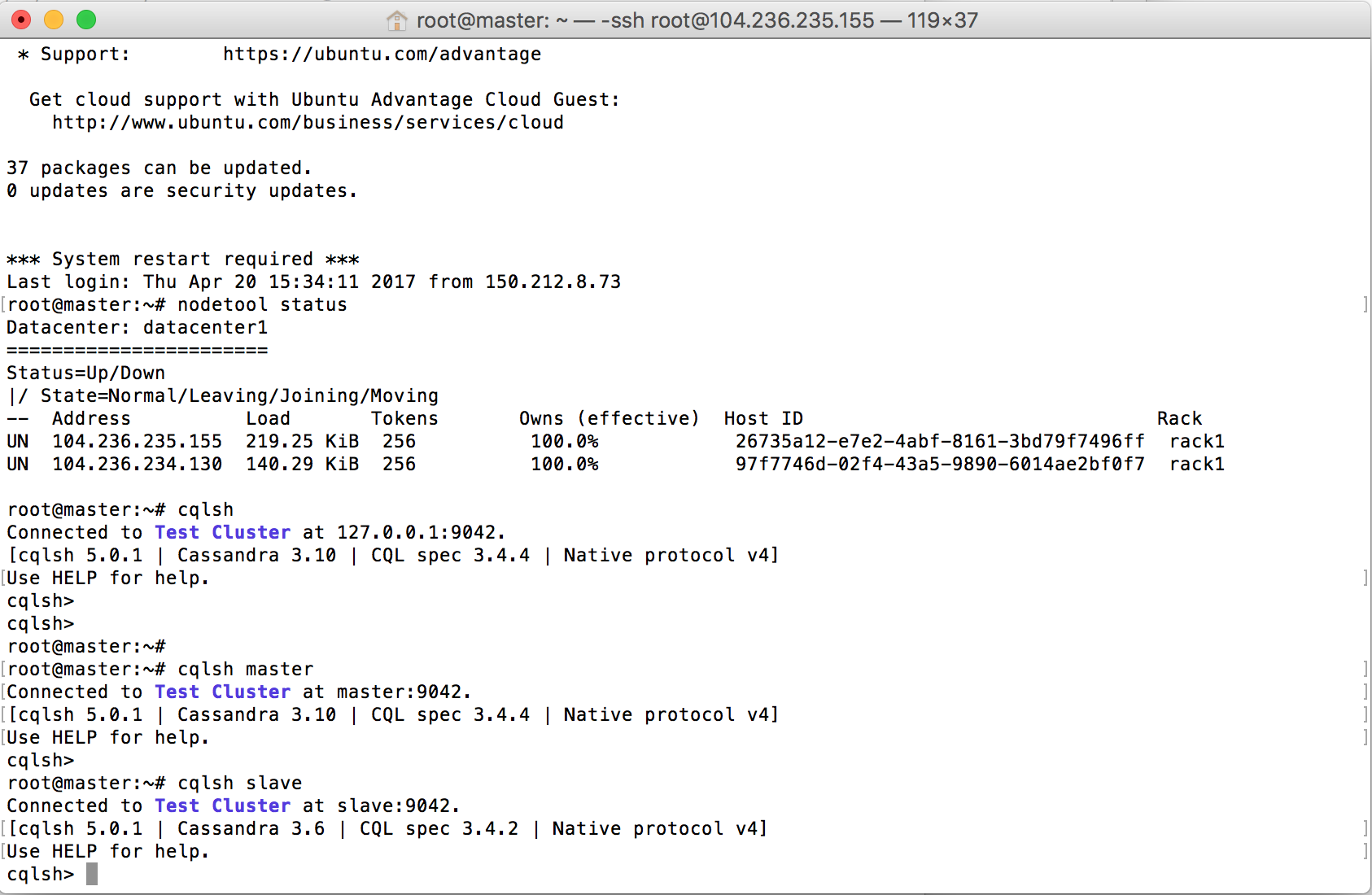
**INFSCI 2750: Cloud Computing**

**Mini Project 3**

**Deadline: April 27, 2017, 11:59 pm**

Part 1: Setting up Cassandra





for now the Cassandra is in the cluster mode.

Part 2: Import data into Cassandra

#first we should convert the access log file into modified csv file



Before we start to import data, we should open the Cassandra service.

1. Open a terminal and type ‘cassandra -Rf’
2. Then open another new terminal and type ‘cqlsh –request-time 600 localhost’

#then to import the data in the Cassandra, we should create at least one keyspace and one table

CREATE KEYSPACE test WITH REPLICATION = {

‘class’ : ‘SImpleStrategy’,

‘replication’ = 3 };

use test;

#then create a table called accesslog in keyspace test.

CREATE TABLE accesslog(

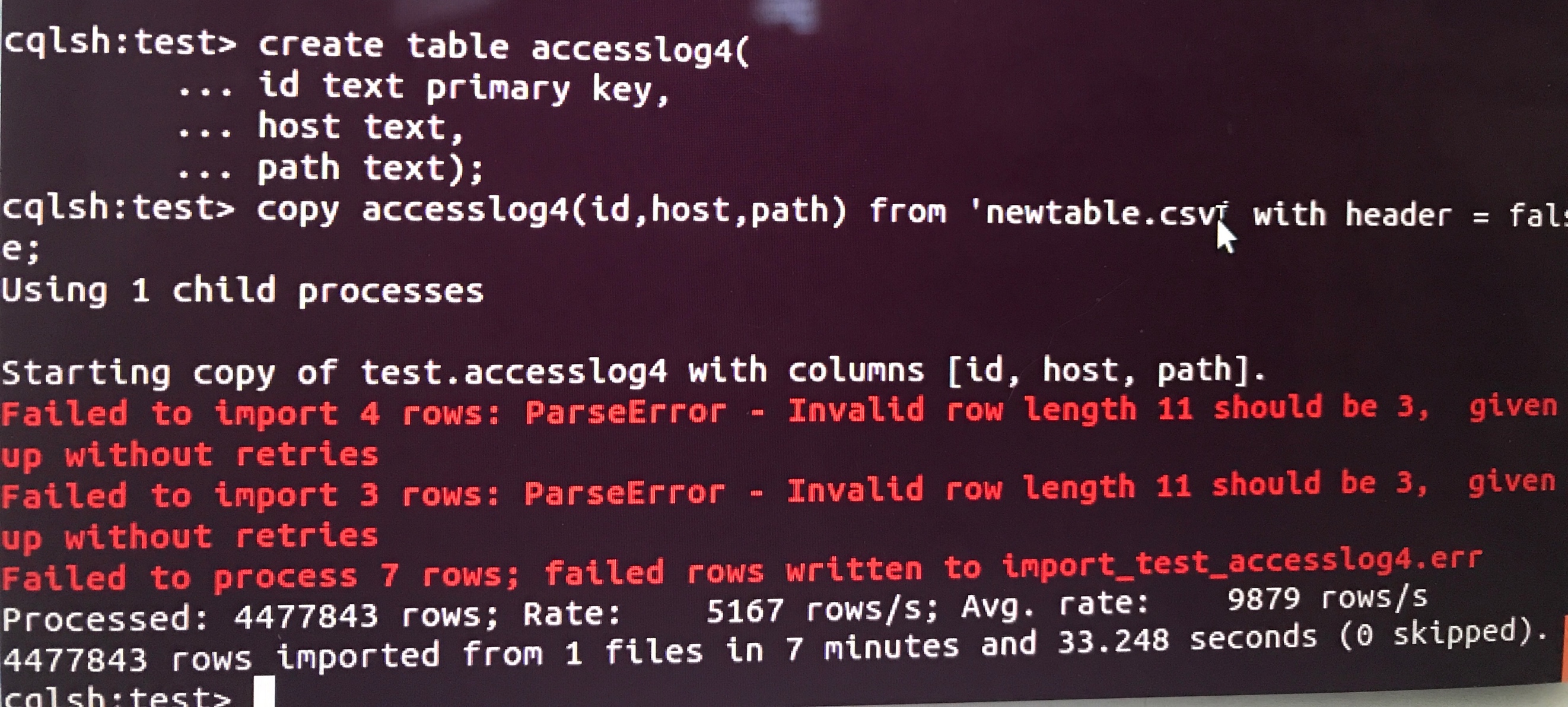
Id text PRIMARY KEY,

host text,

path text );

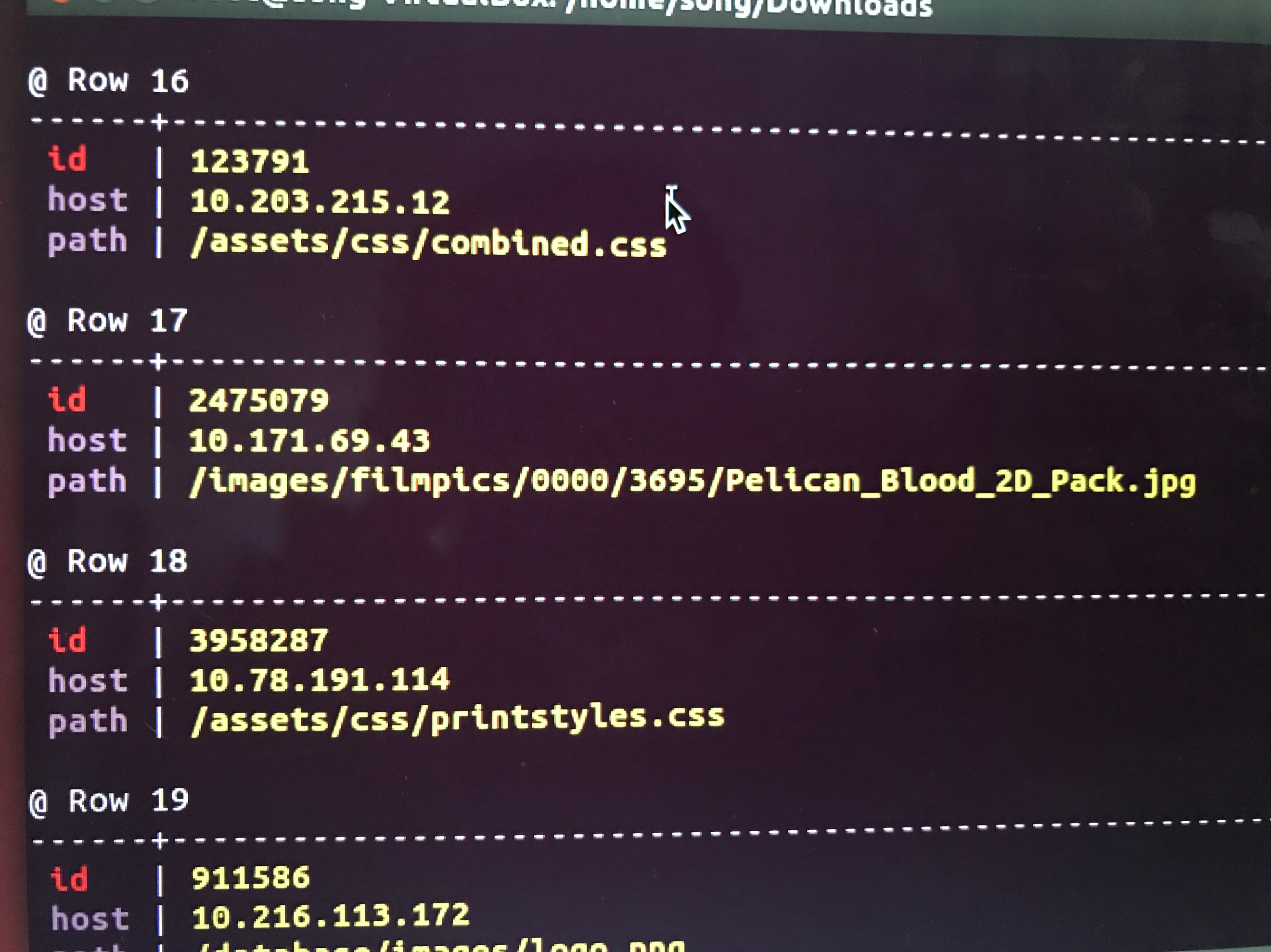
#then we copy the data from access1.csv to the table accesslog

COPY accesslog (id,host,request,path) FROM ‘access1.csv’ WITH header = FALSE;



the figure above is the screenshot of the sample imported data in accesslog table.

SELECT \* FROM accesslog4

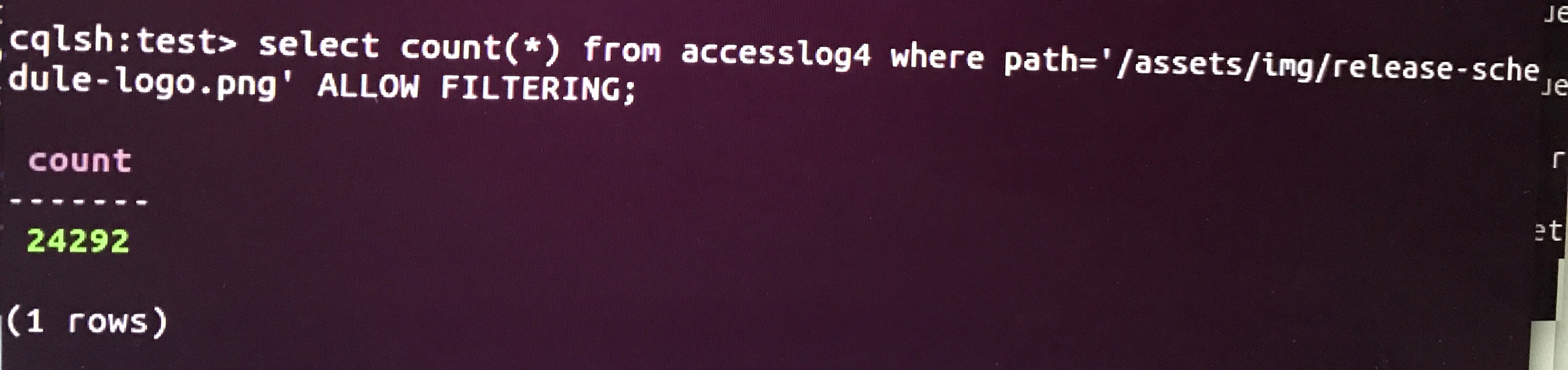


Part3: Operate data in Cassandra

1. how many hit were made to the website item ‘/assests/img/release-schedulelogo.png’

SELECT COUNT(\*) from accesslog4 WHERE path=’ ‘/assests/img/release-schedulelogo.png’

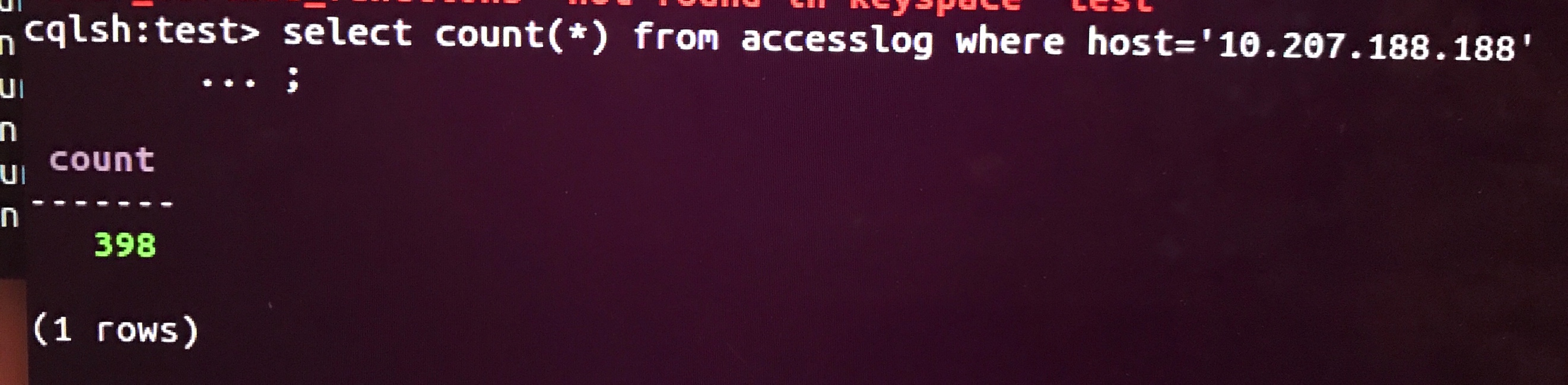
The result: there are 24292 hits



2.how many hits were made from the host ’10.207.188.188’

SELECT COUNT(\*) FROM accesslog WHERE host=’10.207.188.188’

The result: there are 398 hits



1. which path in the website has been hit the most and how many hits were made to the path?

Fot the groupby function, there are no groupby in the Cassandra, so in this part, we should create User Defined Function to implement the groupby and return the max count.

CREATE OR REPLACE FUNCTION group\_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\_max(text)

SFUNC group\_count,

STYPE map<text, int>

INITICOND {};

Then we export the query result to the csv file as result1.csv

And then we create a new table to import the result1

Then we use maxAgg(count) to find the maximum count

#find the maximum count

CREATE FUNCTION maxI(current int,candidate int)

CALLED  ON NULL INPUT

RETURN int LANGUAGE java AS

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

CREATE AGGREGATE maxAgg(int)

SFUNC maxI

STYPE int

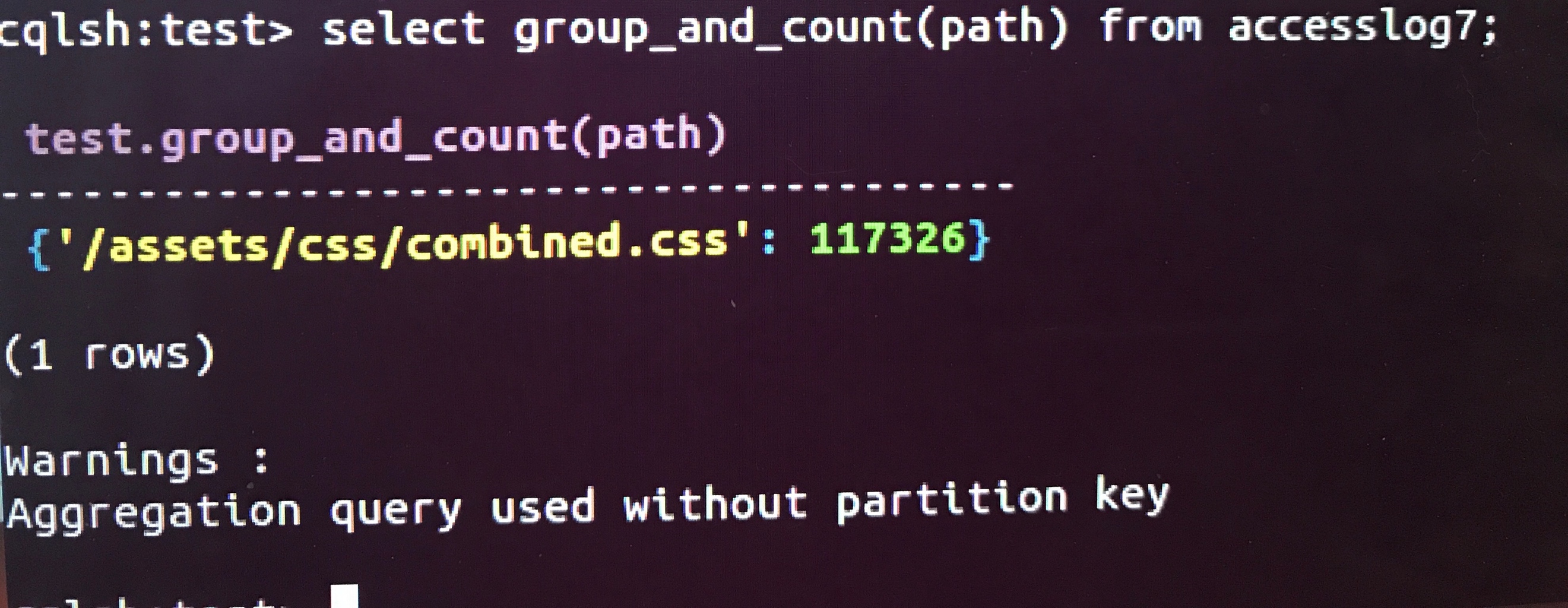
INITCOND null;

Then we execute this aggregate function in the cqlsh with

SELECT maxAgg(count) FROM result1;

The result is showed below:

The path ‘/assets/css/combind.css’ has been hit the most there are 117348 hits were made by this path.



1. Which IP access the website most and how many access were made by it?

Same with the question 3 use the same aggregate function to find the IP.

We still need to export a query result to the new csv file result2

SELECT maxAgg(count) FROM result2;

the result is showed below:

the ’10.216.113.172’ access the website most and there are 158614 were made by this IP address

