ANALYTICS SYSTEMS ENGINEERING (MSDS 436)

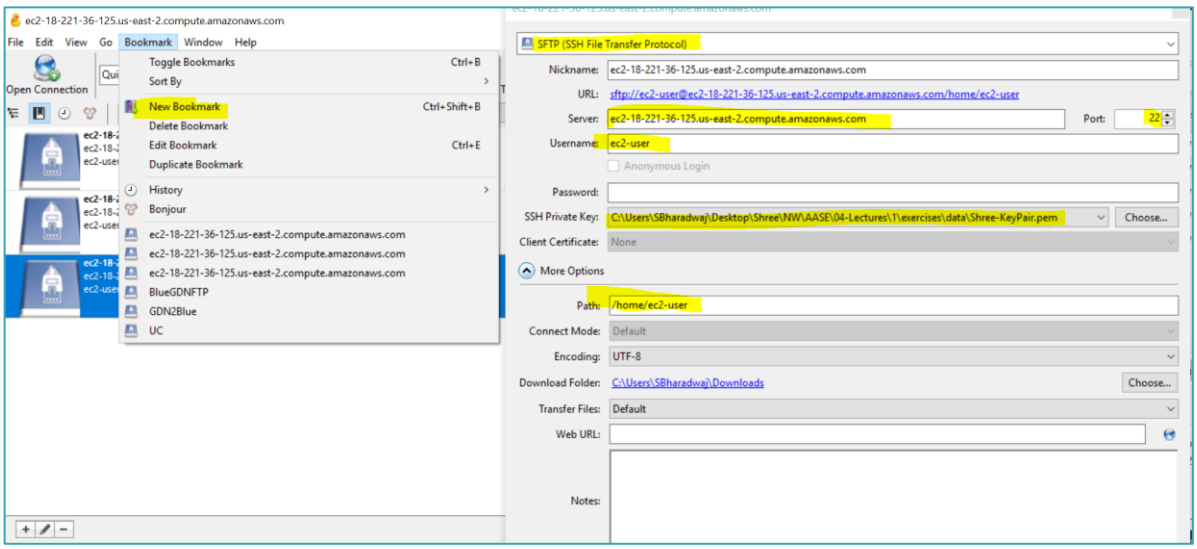
EXERCISE 4

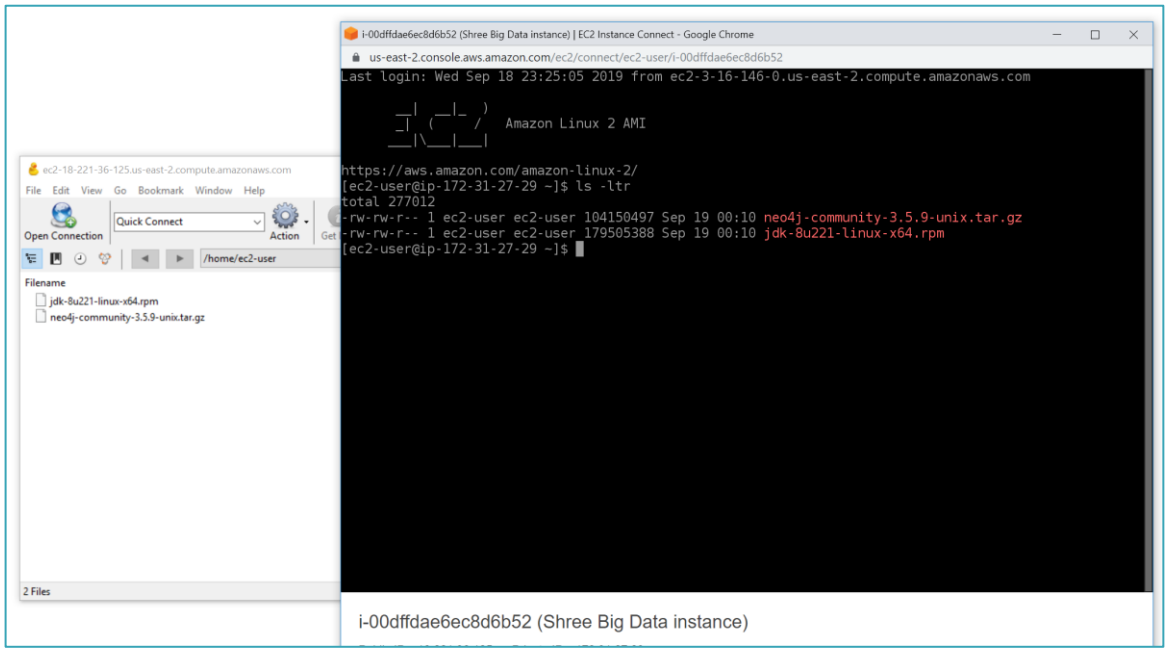
shreenidhi.bharadwaj@northwestern.edu | ChristopherFiore2015@u.northwestern.edu

CONTENTS

INSTALL NEO4J ON AWS EC2

1. Download the latest stable version of neo4j. The "Community Edition" fits well for development purposes. Do not forget to select the Linux version of the server. This will download a tar.gz file. Alternatively, a version of neo4J (**neo4j-community-3.5.9-unix.tar.gz)** is already added to the exercise folder within canvas which you can use readily.
2. Download the JDK file "**jdk-8u221-linux-x64.rpm**" from canvas and store it locally.
3. Copy both the **neo4j-community-3.5.9-unix.tar.gz & jdk-8u221-linux-x64.rpm** files to EC2 using Cyberduck.





4. Install java on the EC2 Instance and confirm it is installed successfully
    a. **sudo yum install -y jdk-8u221-linux-x64.rpm**
    b. **java -version**

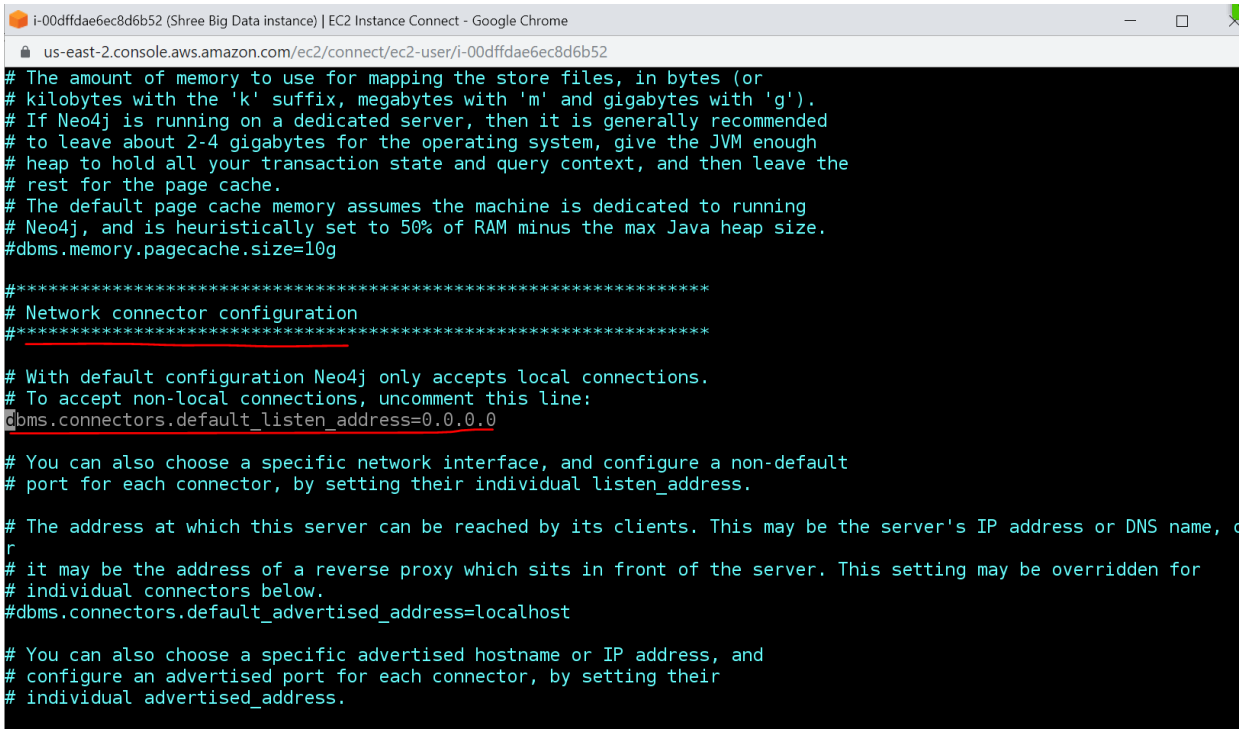
```
i-00dffdae6ec8d6b52 (Shree Big Data instance) | EC2 Instance Connect - Google Chrome
us-east-2.console.aws.amazon.com/ec2/connect/ec2-user/i-00dffdae6ec8d6b52
Installing:
 jdk1.8          x86_64          2000:1.8.0_221-fcs          /jdk-8u221-linux-x64          298 M

Transaction Summary
================================================================================================
Install  1 Package

Total size: 298 M
Installed size: 298 M
Downloading packages:
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : 2000:jdk1.8-1.8.0_221-fcs.x86_64                                                1/1
Unpacking JAR files...
        tools.jar...
        plugin.jar...
        javaws.jar...
        deploy.jar...
        rt.jar...
        jsse.jar...
        charsets.jar...
        localedata.jar...
  Verifying  : 2000:jdk1.8-1.8.0_221-fcs.x86_64                                                1/1

Installed:
  jdk1.8.x86_64 2000:1.8.0_221-fcs

Complete!
[ec2-user@ip-172-31-27-29 ~]$ java -version
java version "1.8.0_221"
Java(TM) SE Runtime Environment (build 1.8.0_221-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.221-b11, mixed mode)
[ec2-user@ip-172-31-27-29 ~]$
i-00dffdae6ec8d6b52 (Shree Big Data instance)
```

5. Install Neo4J on the EC2 Instance and move it to /usr/local and rename the folder to neo4j.
    a. **tar xvfz neo4j-community-3.5.9-unix.tar.gz**
    b. **sudo mv neo4j-community-3.5.9 /usr/local/neo4j**

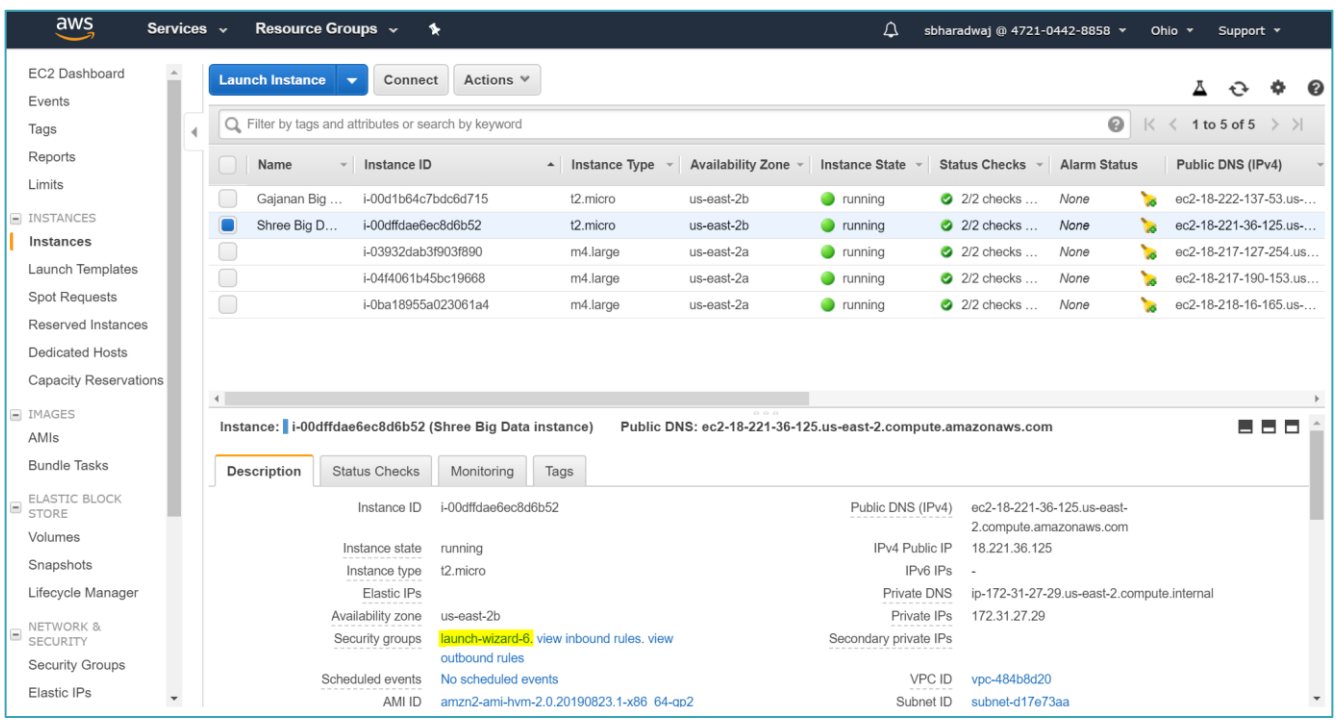
```
total 277012
drwxr-xr-x 10 ec2-user ec2-user       198 Aug 22 11:27 neo4j-community-3.5.9
-rw-rw-r--  1 ec2-user ec2-user 104150497 Sep 19 00:10 neo4j-community-3.5.9-unix.tar.gz
-rw-rw-r--  1 ec2-user ec2-user 179505388 Sep 19 00:10 jdk-8u221-linux-x64.rpm
[ec2-user@ip-172-31-27-29 ~]$ sudo mv neo4j-community-3.5.9 /usr/local/neo4j
[ec2-user@ip-172-31-27-29 ~]$
```

6. Open "neo4j.conf" using vi and locate "Network connector configuration" section to Uncomment (delete "#", use x by placing cursor on #) in front of "dbms.connectors.default_listen_address=0.0.0.0" on line 54 and save the file by typing (ESC, :wq! + Enter)
    a. **vi /usr/local/neo4j/conf/neo4j.conf**

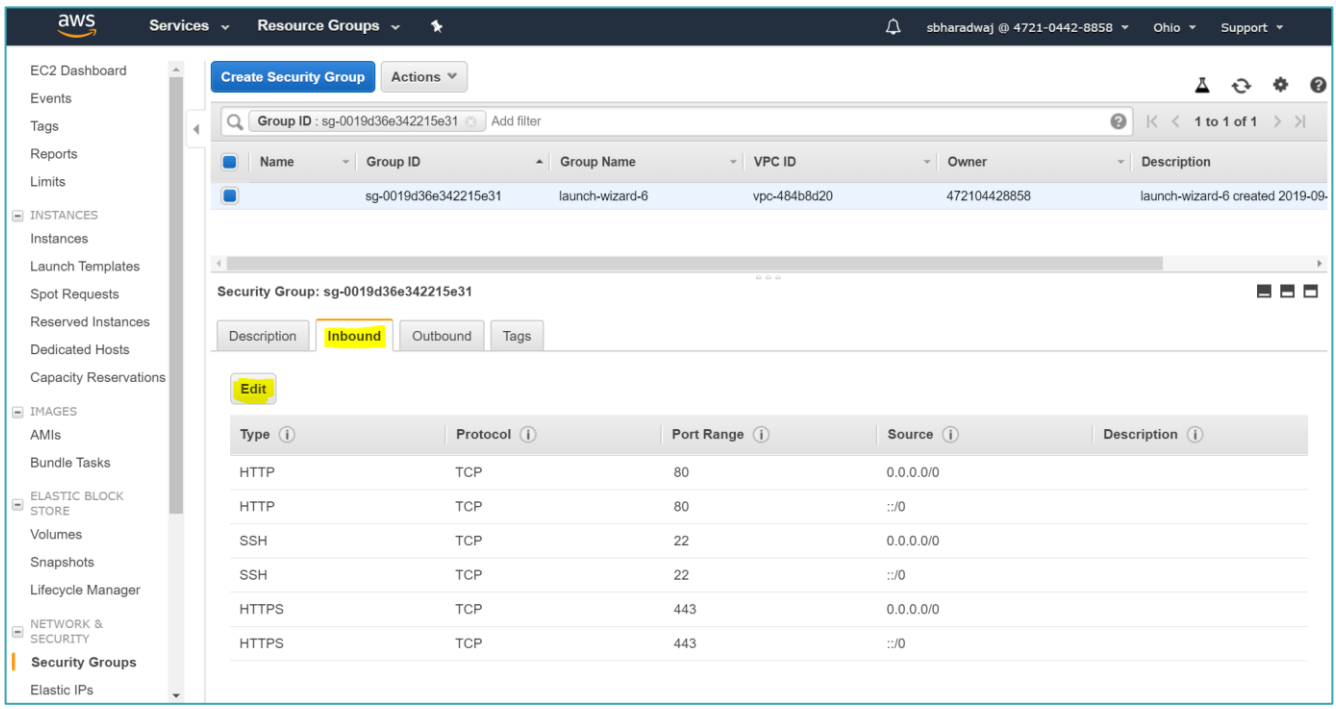
    Note: This will let you connect to Neo4j from your local machine



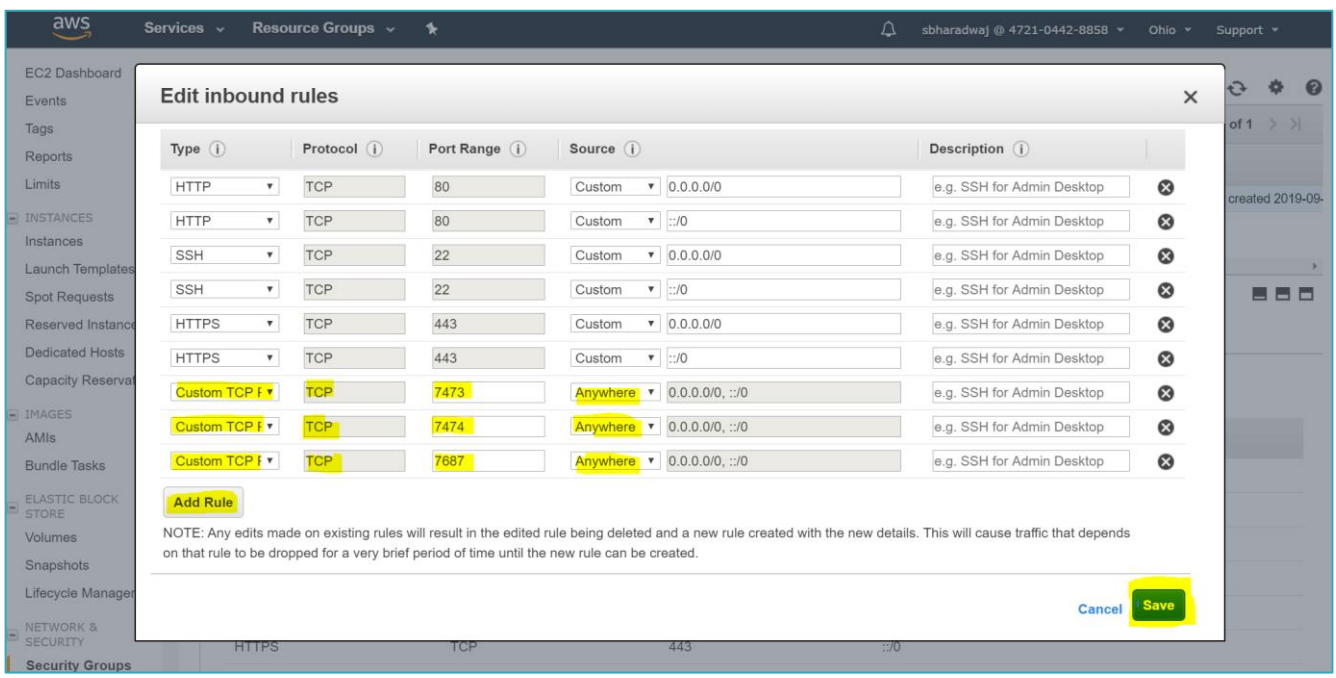
7. To Access to Neo4j from your laptop we will need to open specific ports in AWS EC2 instance. Locate the security group of your EC2 instance & click on it



8. Click on "inbound" tab & then click on "Edit" to add new rules



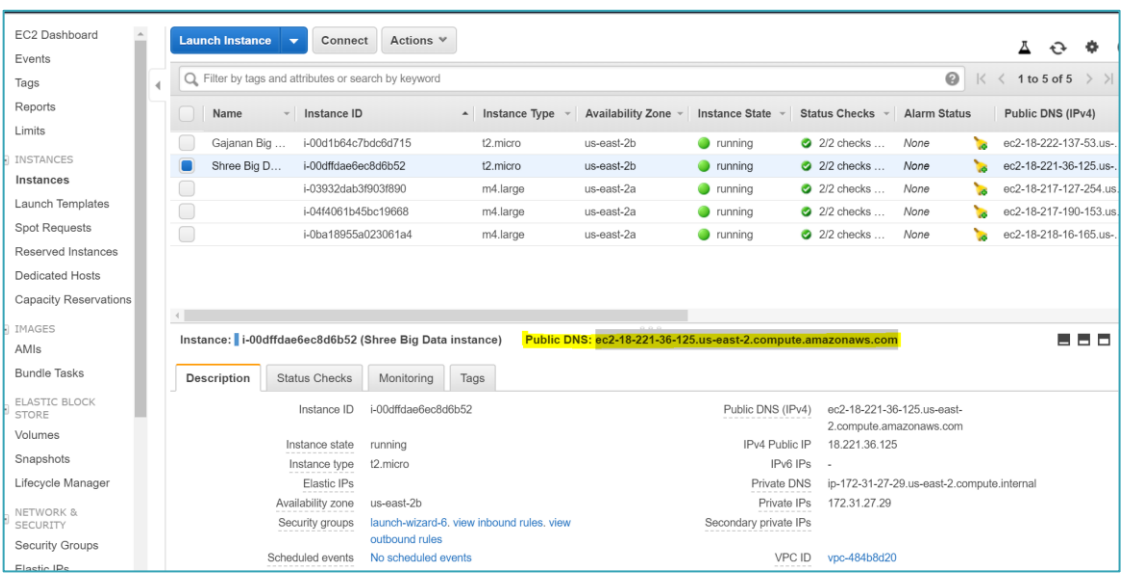
9. Add "Custom TCP rule" for ports 7473, 7474, 7687 and click "Save"



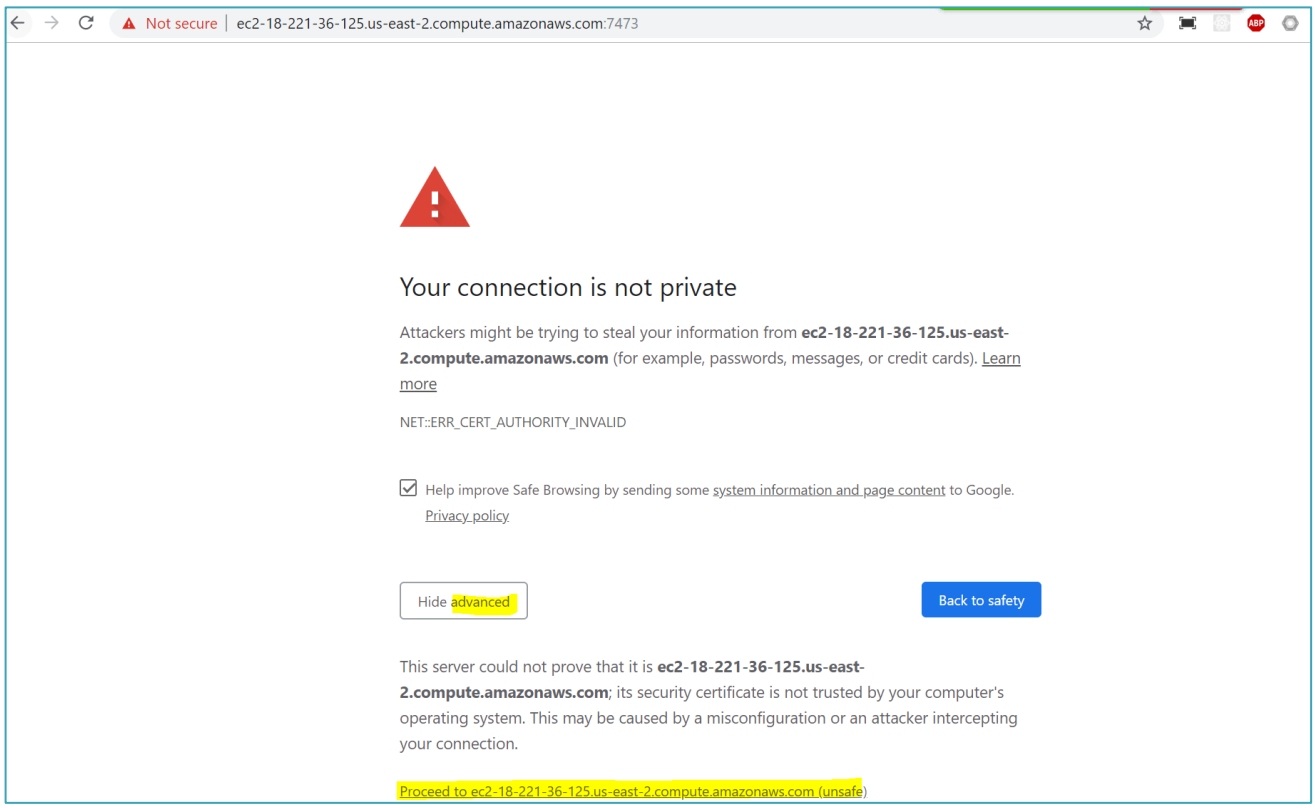
10. Start the Neo4J server
    a. sudo /usr/local/neo4j/bin/neo4j start
    b. tail -f /usr/local/neo4j/logs/neo4j.log (if you want to check the logs)

```
[ec2-user@ip-172-31-27-29 ~]$ sudo /usr/local/neo4j/bin/neo4j start
Active database: graph.db
Directories in use:
  home:         /usr/local/neo4j
  config:       /usr/local/neo4j/conf
  logs:         /usr/local/neo4j/logs
  plugins:      /usr/local/neo4j/plugins
  import:       /usr/local/neo4j/import
  data:         /usr/local/neo4j/data
  certificates: /usr/local/neo4j/certificates
  run:          /usr/local/neo4j/run
Starting Neo4j.
WARNING: Max 1024 open files allowed, minimum of 40000 recommended. See the Neo4j manual.
Started neo4j (pid 1113). It is available at http://0.0.0.0:7474/
There may be a short delay until the server is ready.
See /usr/local/neo4j/logs/neo4j.log for current status.
[ec2-user@ip-172-31-27-29 ~]$ tail -f /usr/local/neo4j/logs/neo4j.log
nohup: ignoring input
2019-09-19 00:58:35.198+0000 INFO  ======== Neo4j 3.5.9 ========
2019-09-19 00:58:35.227+0000 INFO  Starting...
2019-09-19 00:58:40.073+0000 INFO  Bolt enabled on 0.0.0.0:7687.
2019-09-19 00:58:43.245+0000 INFO  Started.
2019-09-19 00:58:45.465+0000 INFO  Remote interface available at http://localhost:7474/
```

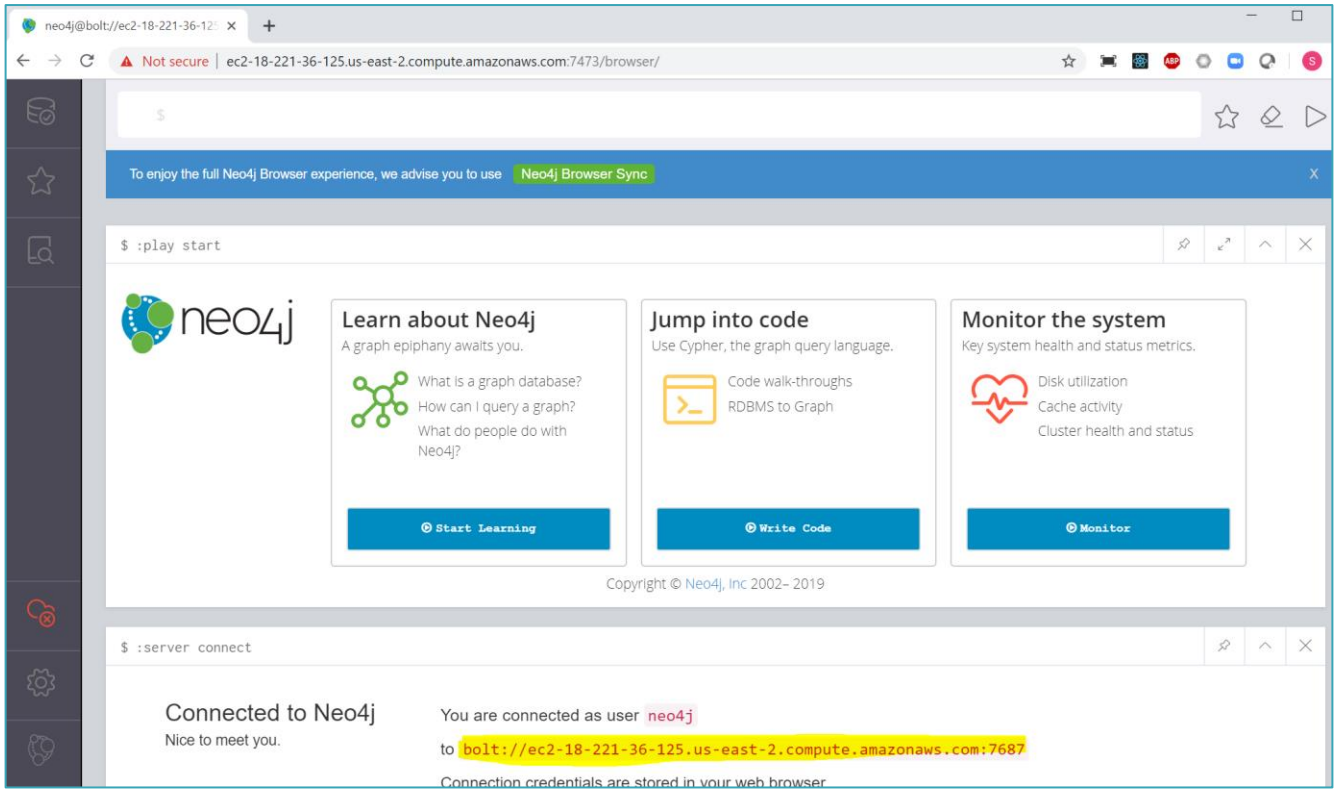
11. Neo4J Server setup is now complete. Connect to Neo4j server running in AWS EC2 instance by typing below URL on your laptop chrome browser
    - https://<YOUR_PUBLIC_DNS>:7473/
      E.g. https://ec2-18-221-36-125.us-east-2.compute.amazonaws.com:7473/



12. Once loaded connect to Neo4j using 'neo4j' as userid & password. On your first login, the neo4j app asks you to update the password, update the password to "root" & keep it handy for later use.



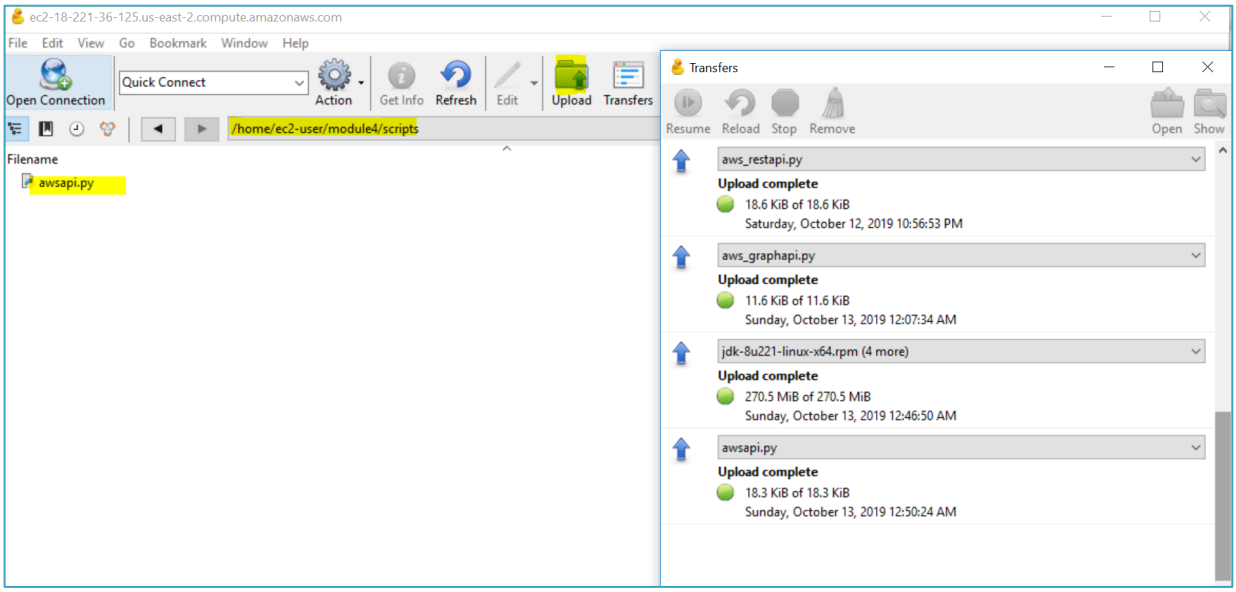
After successful login you should see a screen like below:



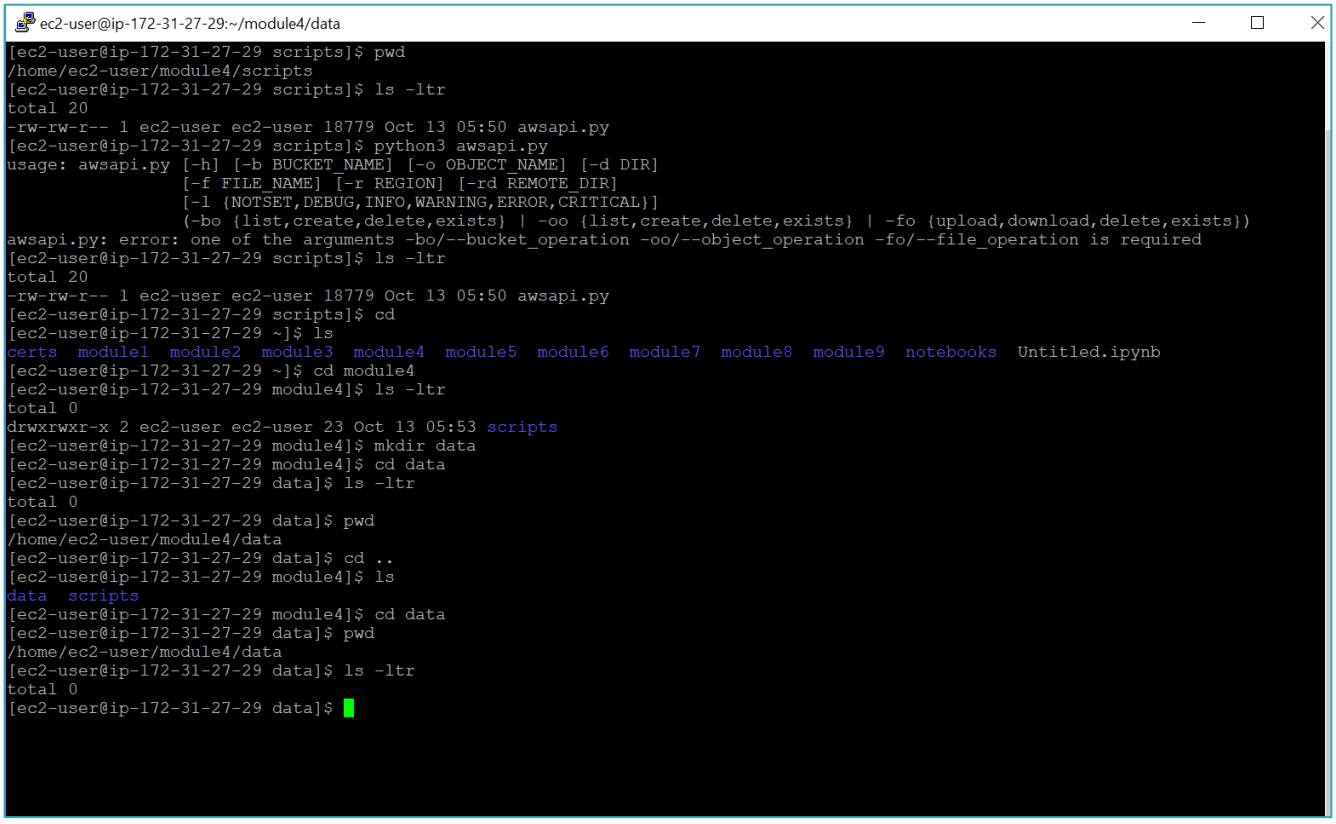
Note: Neo4J installation is now complete.

IMPORTING DATA INTO Neo4J

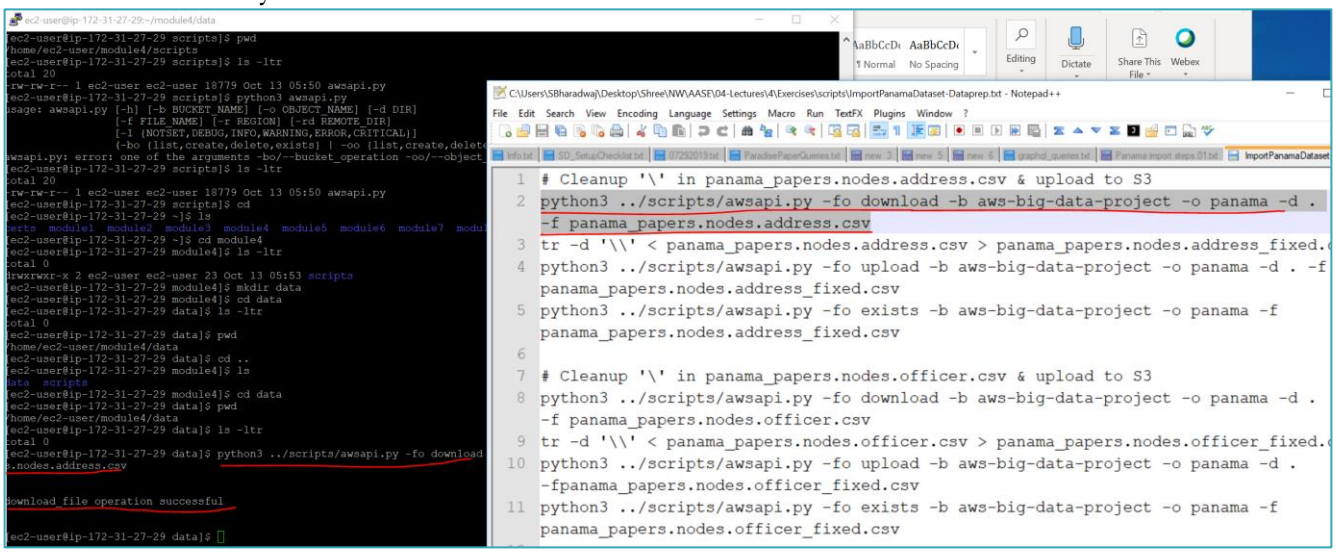
1. Using Cyberduck please upload the scripts into the module4/scripts folder within EC2. Note: If the scripts folder does not exist then create it.



2. Login to EC2 instance and create the data folder within the module4 folder.
   mkdir data
   cd data



3. Within the data folder, use the commands given in the ImportPanamaDataset-Dataprep.txt to prep the data for importing it into Neo4J.
   Note: Please verify the counts as indicated in the text file.

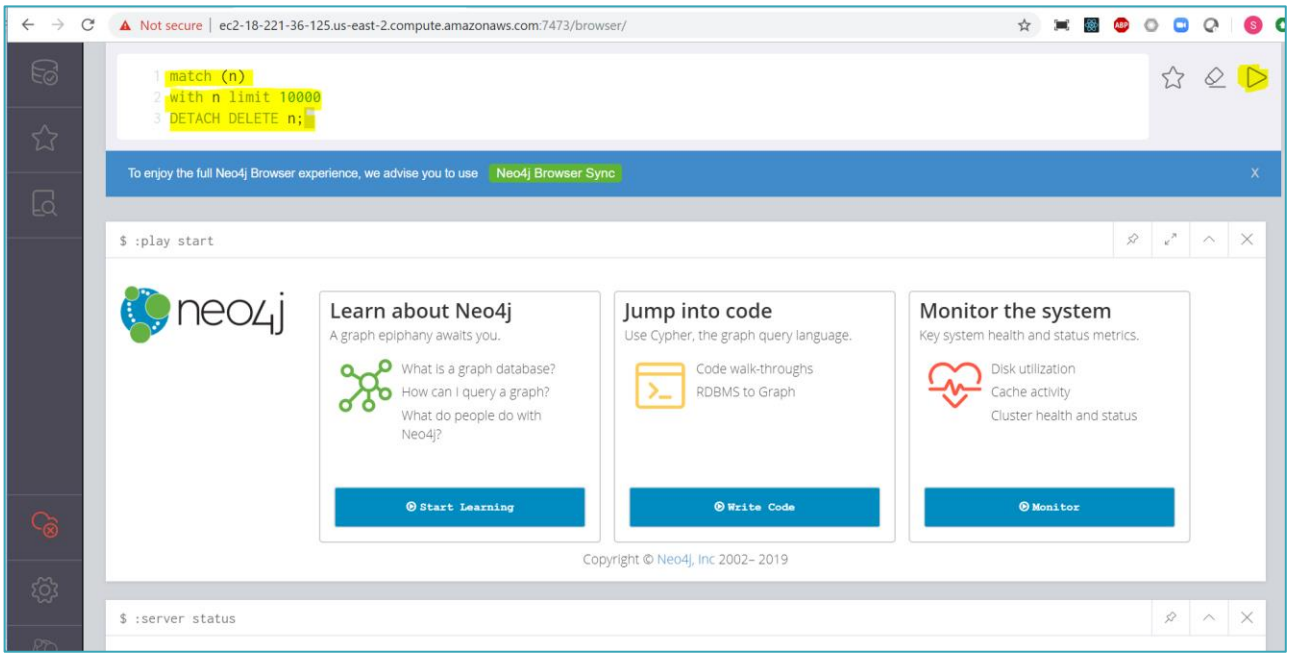


VISUALIZING CONNECTED DATA IN NEO4J

1. Login to Neo4J database
   https://<EC2_PUBLIC_DNS>:7473/browser/ with credentials (user : neo4j)
   Ex: https://ec2-18-221-36-125.us-east-2.compute.amazonaws.com:7473/

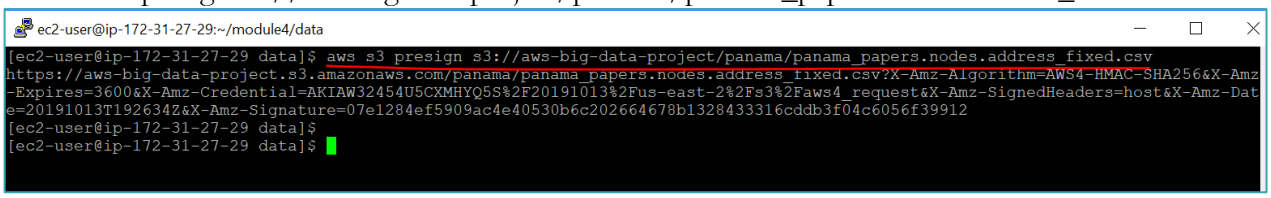
2. Delete all existing data using the cypher query below
   match (n)
   with n limit 10000

DETACH DELETE n;

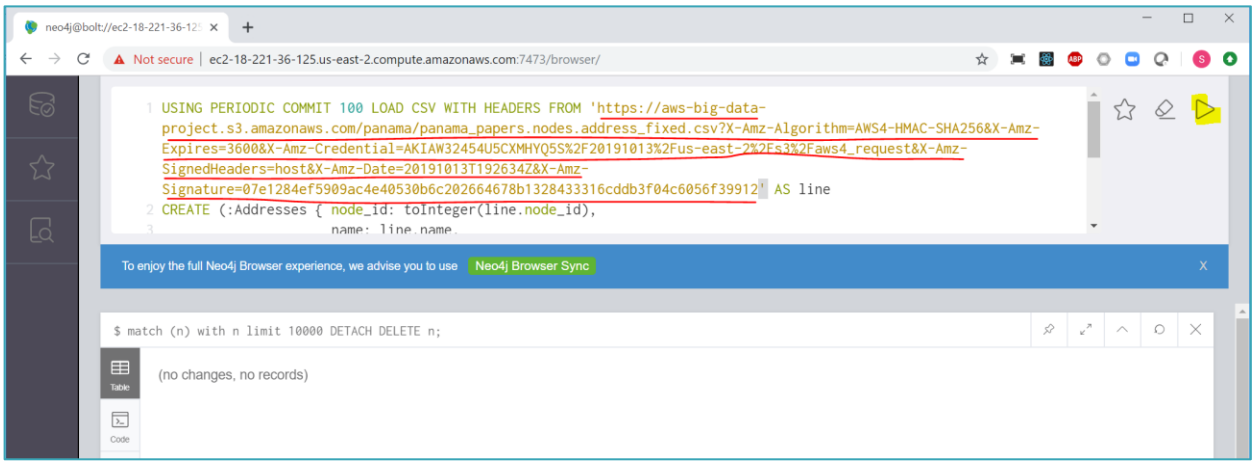


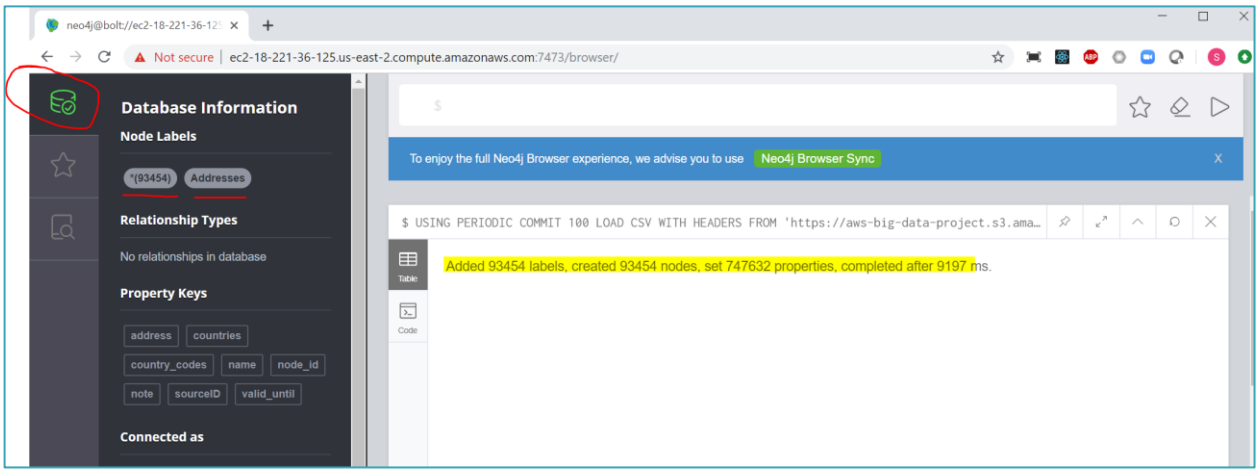
3. Load the data/files from S3 into Neo4J. To do so, we will need
    a) Create a pre authenticated URL
    b) Use the pre authenticated URL to import data into Neo4J

a) Create a pre authenticated URL for panama_papers.nodes.address_fixed.csv
aws s3 presign s3://<s3 bucket> /panama/panama_papers.nodes.address_fixed.csv
Ex: aws s3 presign s3://aws-big-data-project/panama/panama_papers.nodes.address_fixed.csv



b) Use the pre authenticated URL (output from the previous query) to import panama_papers.nodes.address_fixed.csv data into Neo4J.
Note: Refer to ImportPanamaDataset-UsingCypher.txt



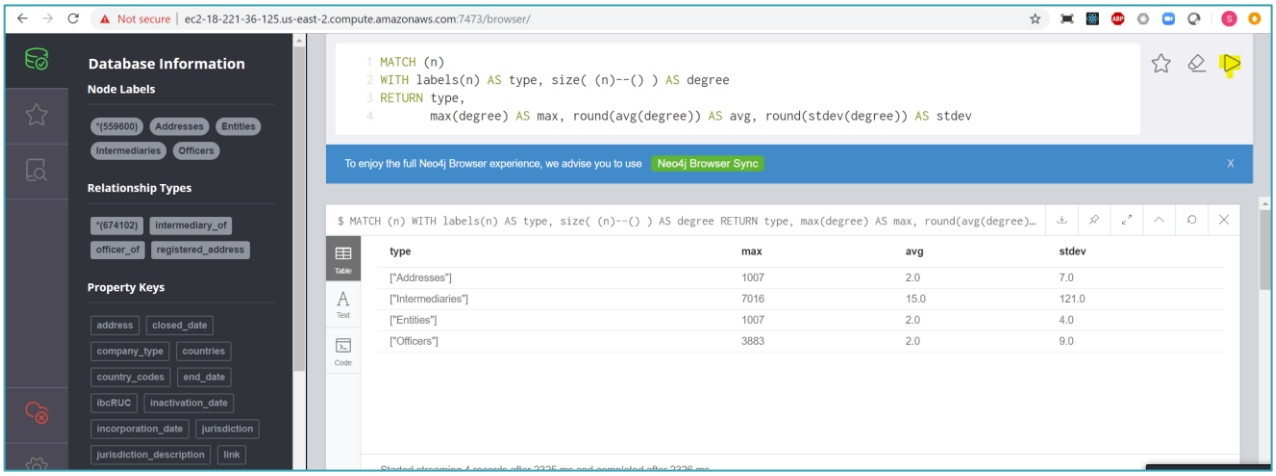


c) Repeat the same process to import rest of the data from S3 into Neo4J. Please refer to the ImportPanamaDataset-UsingCypher.txt script to do the same.

d) Run Queries to visualize connected dataset to derive insights such as basic statistics
MATCH (n)
RETURN labels(n), count(*)
ORDER BY count(*) DESC

Another insight : The distribution of degrees for our entities shows a typical power law: there are some addresses which have 127 companies registered and some people who have almost 90 shell companies registered to their name.

```
MATCH (n)
WITH labels(n) AS type, size((n)--() ) AS degree
RETURN type,
      max(degree) AS max, round(avg(degree)) AS avg, round(stdev(degree)) AS stdev
```



Exercise 4 is now complete