filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

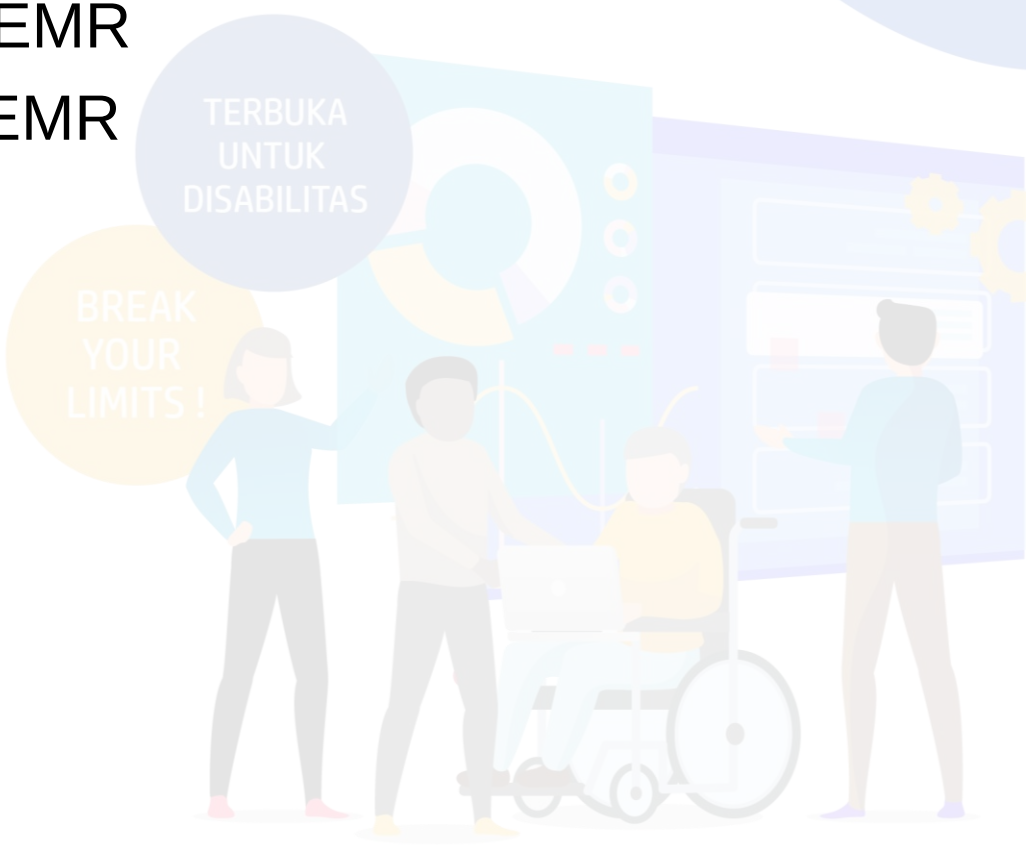
Big Data Analytics

EMR Lab: Hive, Hue, & Others

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)

Pokok Pembahasan

- Menggunakan Hive pada EMR
- Menggunakan Hue pada EMR
- Tugas



Lab Hive

Big Data Query

```
data.head()
```

	created_at	in_reply_to_status_id	id_str	retweeted	in_reply_to_user_id_str	coordinates	retweet_count	contributors	favorite_count
0	2019-07-22 05:41:45	NaN	1153178298081374208	False	NaN	NaN	0	NaN	0
1	2019-07-22 05:41:47	1.153177e+18	1153178304603545600	False	731167584.0	NaN	0	NaN	0
2	2019-07-22 05:41:48	NaN	1153178308716695552	False	NaN	NaN	0	NaN	0
3	2019-07-22 05:41:49	NaN	1153178311515766785						
4	2019-07-22 05:41:49	NaN	1153178313942282247						

favorited	in_reply_to_status_id_str	source	in_reply_to_user_id	user_id_str	user_screen_name	place	geo	text
False	NaN	Twitter for Android	NaN	513783027	fiercealien	NaN	NaN	Aaaakkk YukSay balik [illegible]
False	1.153177e+18	Twitter for iPhone	731167584.0	1003356794683416576	CatsRule98	NaN	NaN	@RonBrownstein That's why #IStandWithIlhan
False	NaN	Twitter for Android	NaN	1097334060	JejakaShahid	NaN	NaN	RT @mathsanova: mana pakai faceapp lagi htt...
False	NaN	Twitter for Android	NaN	1014117841	nurlailaeksa	NaN	NaN	RT @BadmintonTalk: #BtalkBWFRankPrediction MS ...
False	NaN	Twitter for Android	NaN	701611356496113664	sticvh	NaN	NaN	RT @Husen_Jafar: FaceApp haram? #RecehanDakwah...

Siapa user yang paling sering nge-tweet?

Tweet mana yang paling sering difavoritkan?

Aplikasi apa saja yang paling populer untuk nge-twit?

Total data : 5.186.814 baris (row)

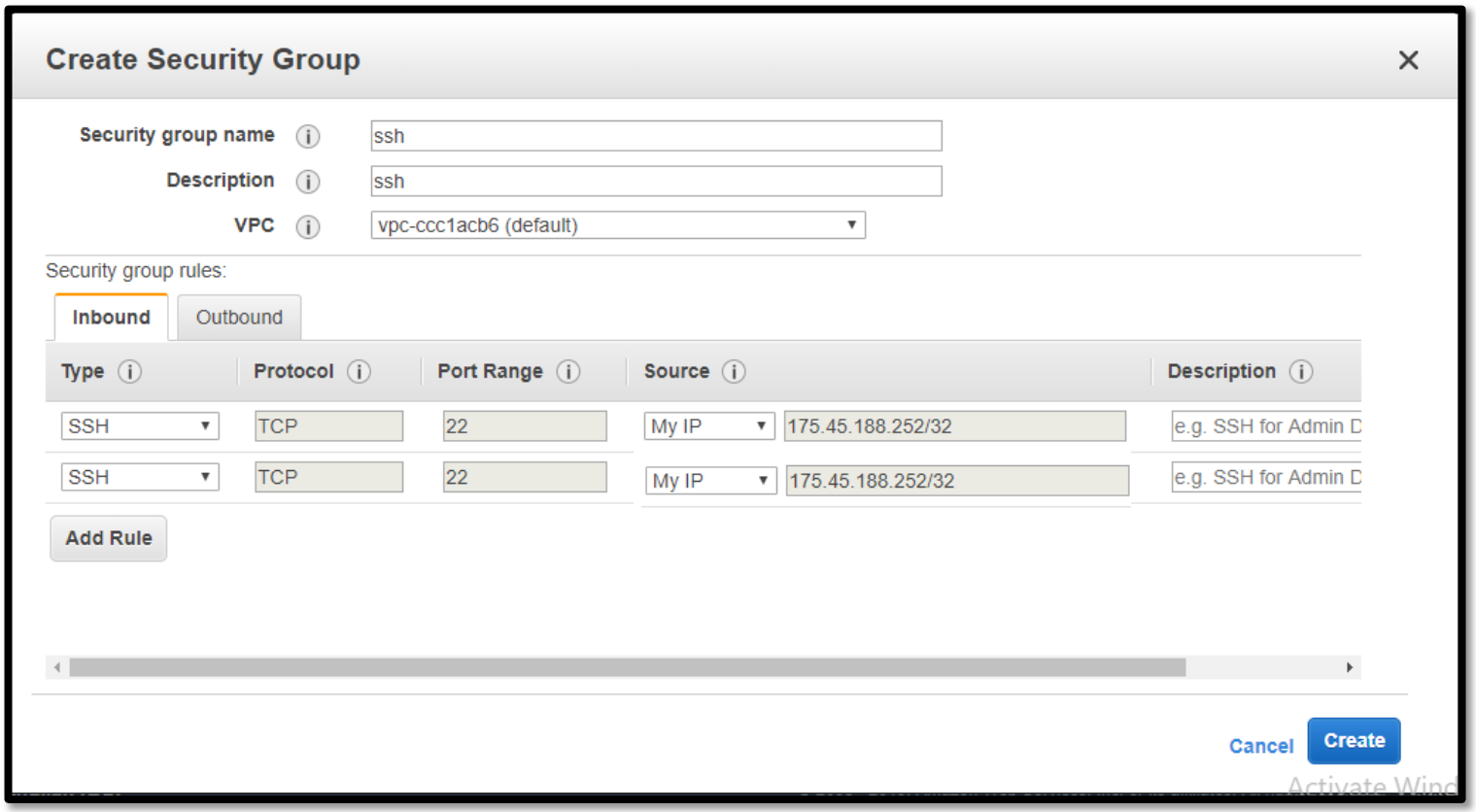
Dataset

Object URL :

https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv

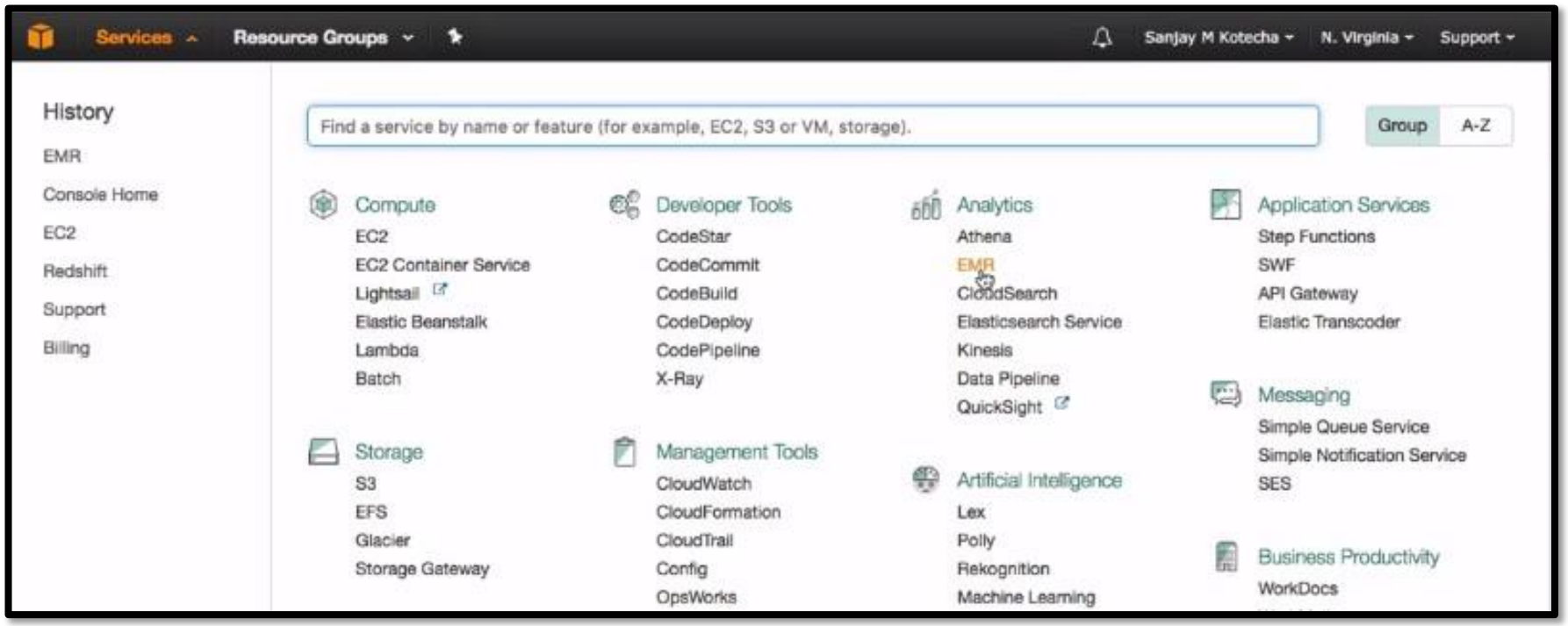
EMR Lab

1. Buatlah security group pada EC2 anda dengan nama SSH, dan juga pastikan inbound dan outbound rule seperti berikut :



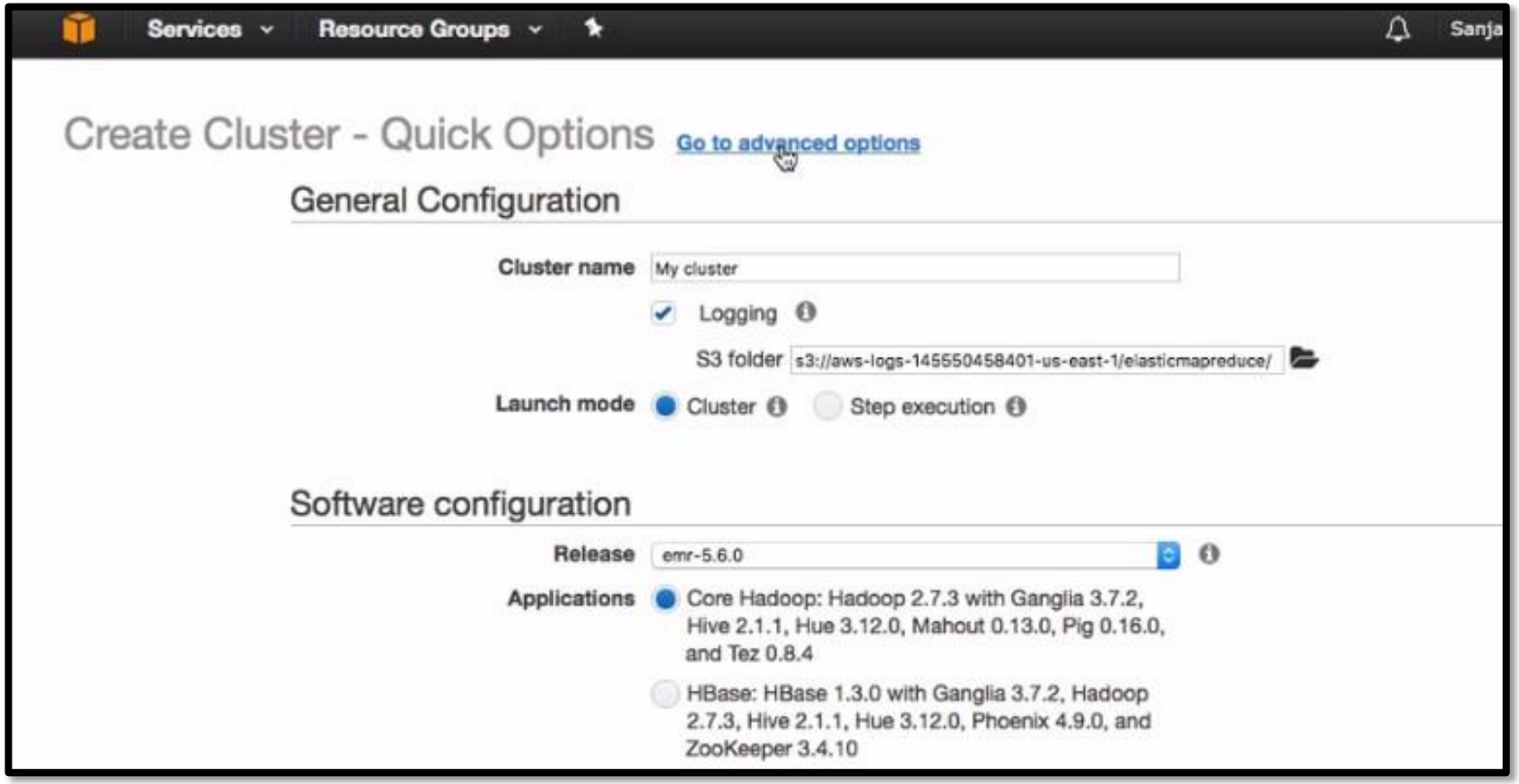
EMR Lab

2. Buka Service EMR pada Analytics



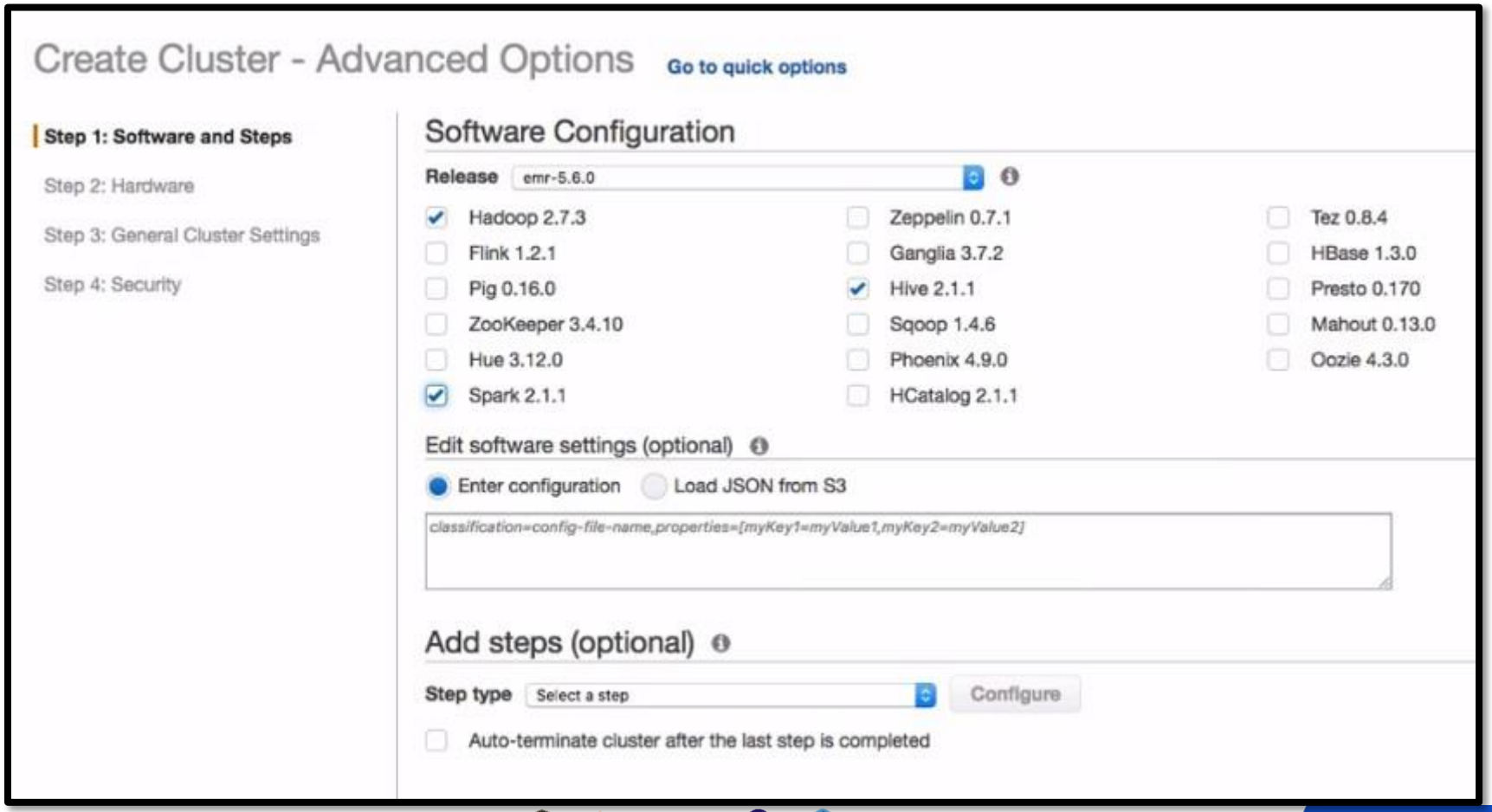
EMR Lab

3. Pada EMR Klik Create Cluster à go to advanced options



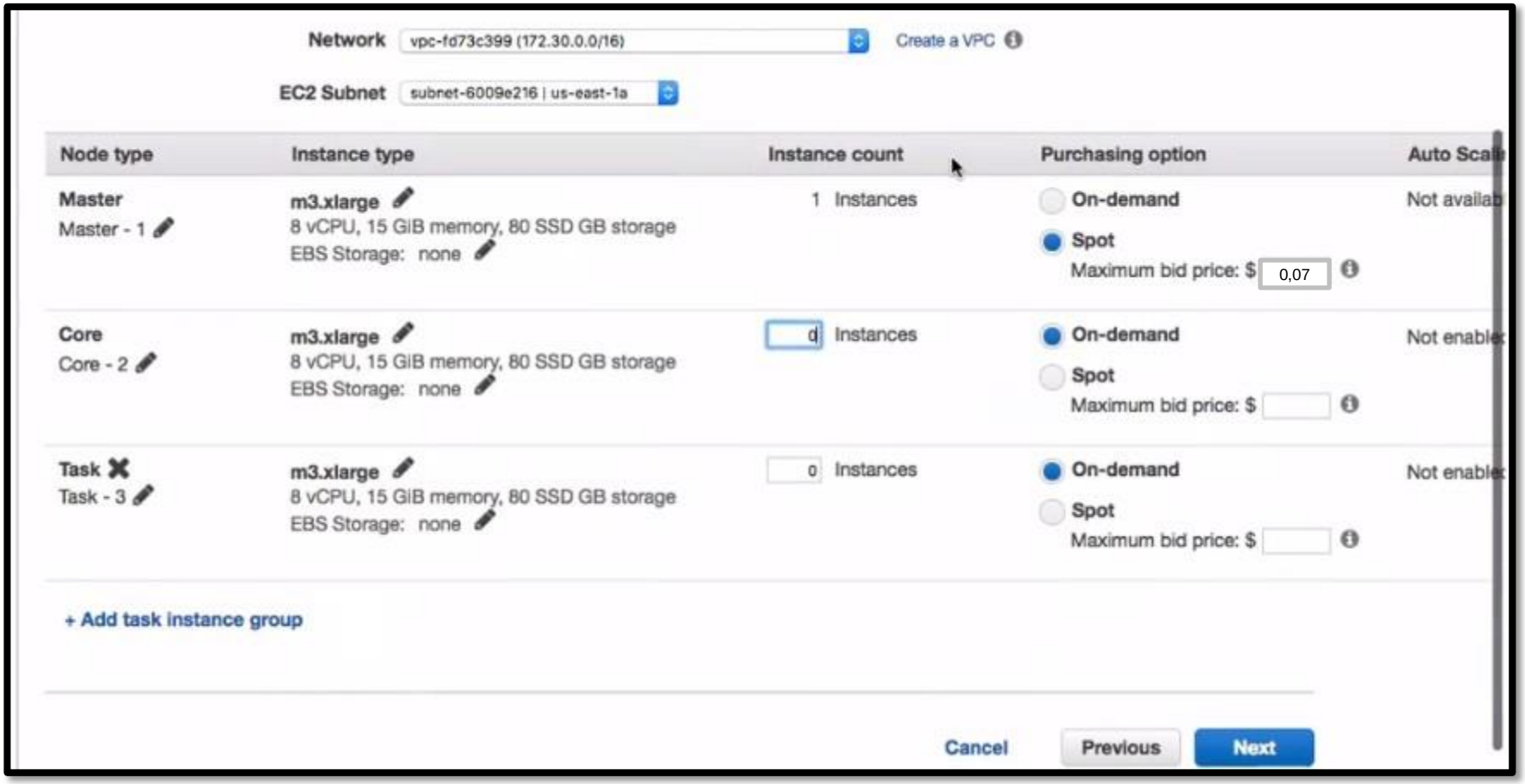
EMR Lab

4. Sesuaikan software configuration seperti ini, lalu klik Next à



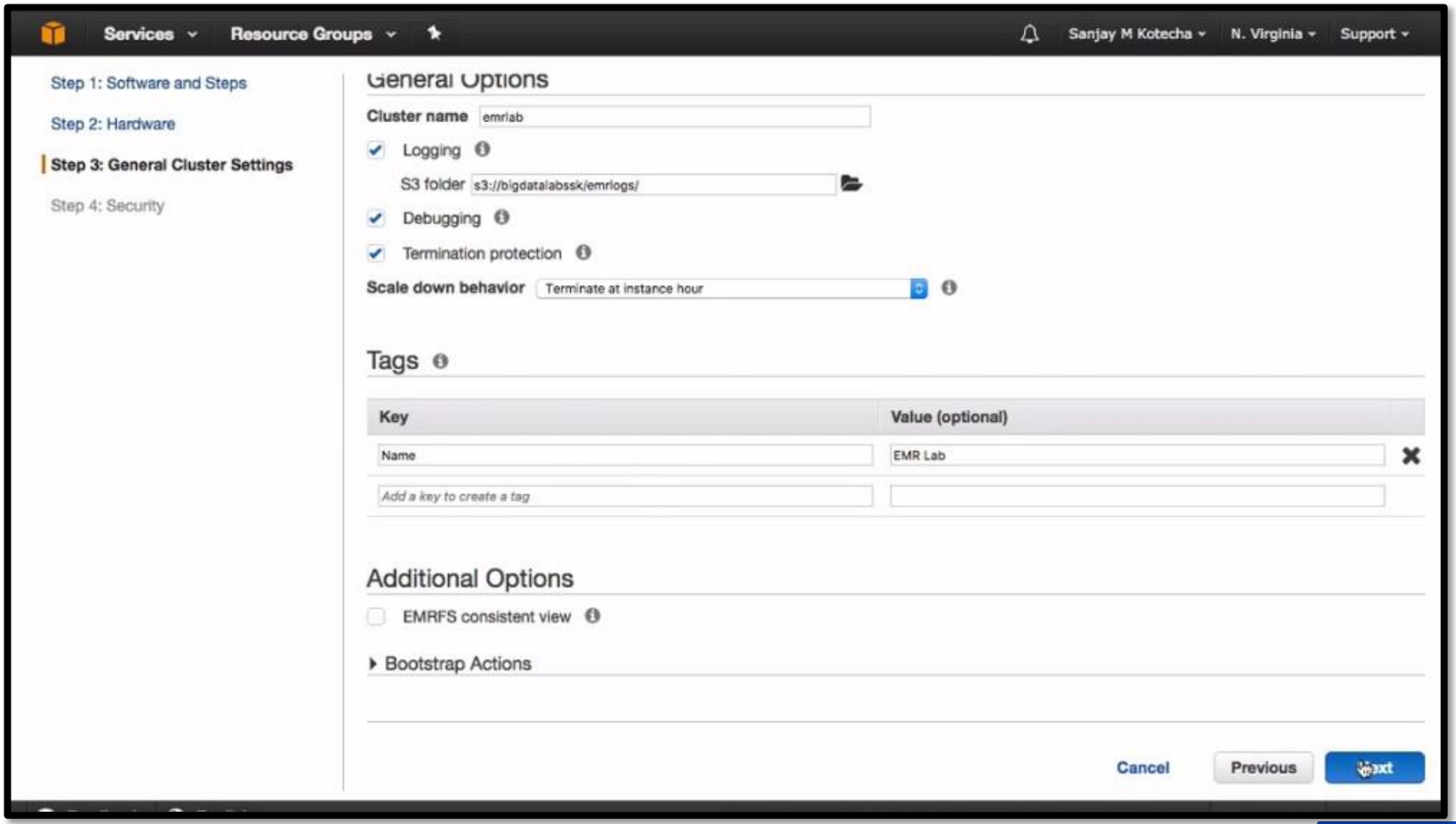
EMR Lab

5. Sesuaikan hardware configuration seperti ini, lalu klik Next à



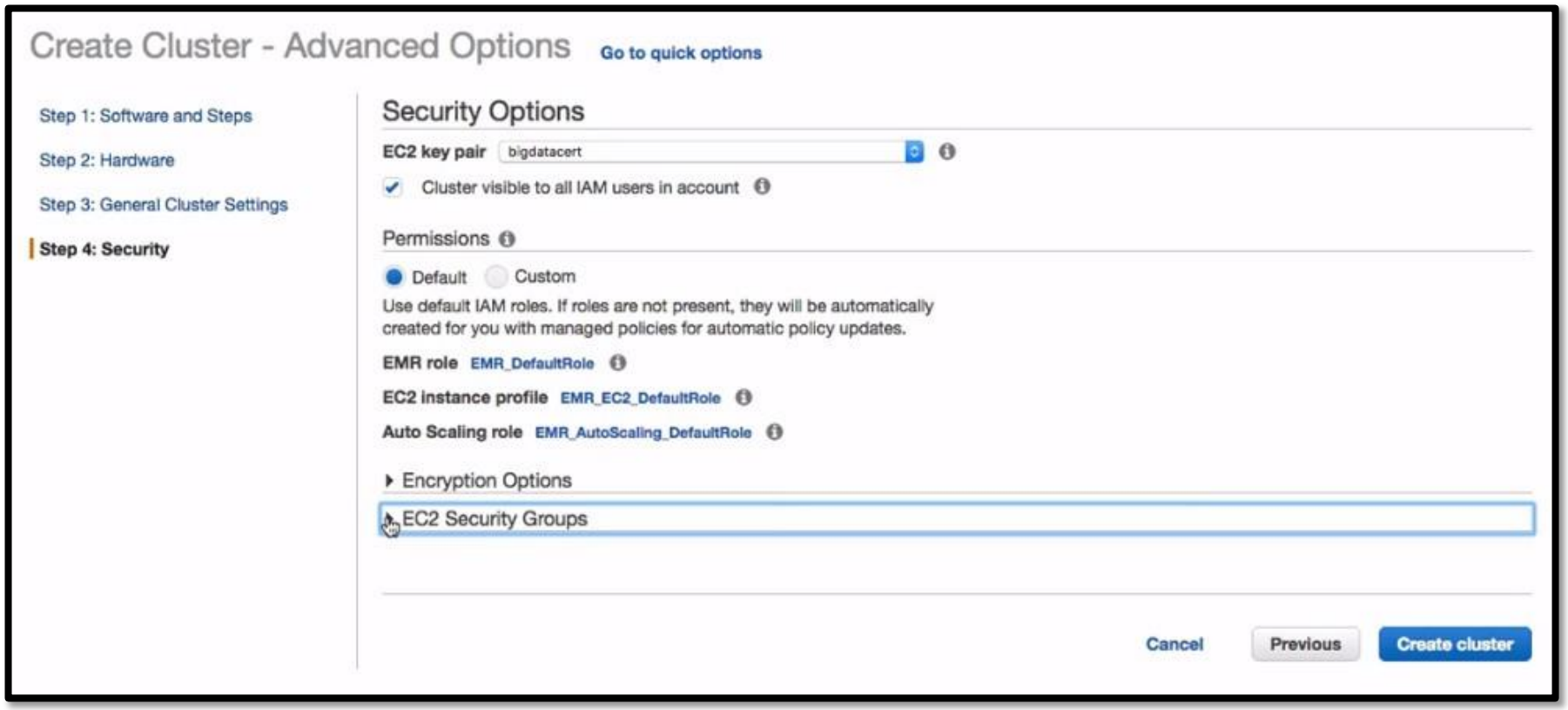
EMR Lab

6. Sesuaikan General Cluster setting seperti berikut, lalu klik Next à



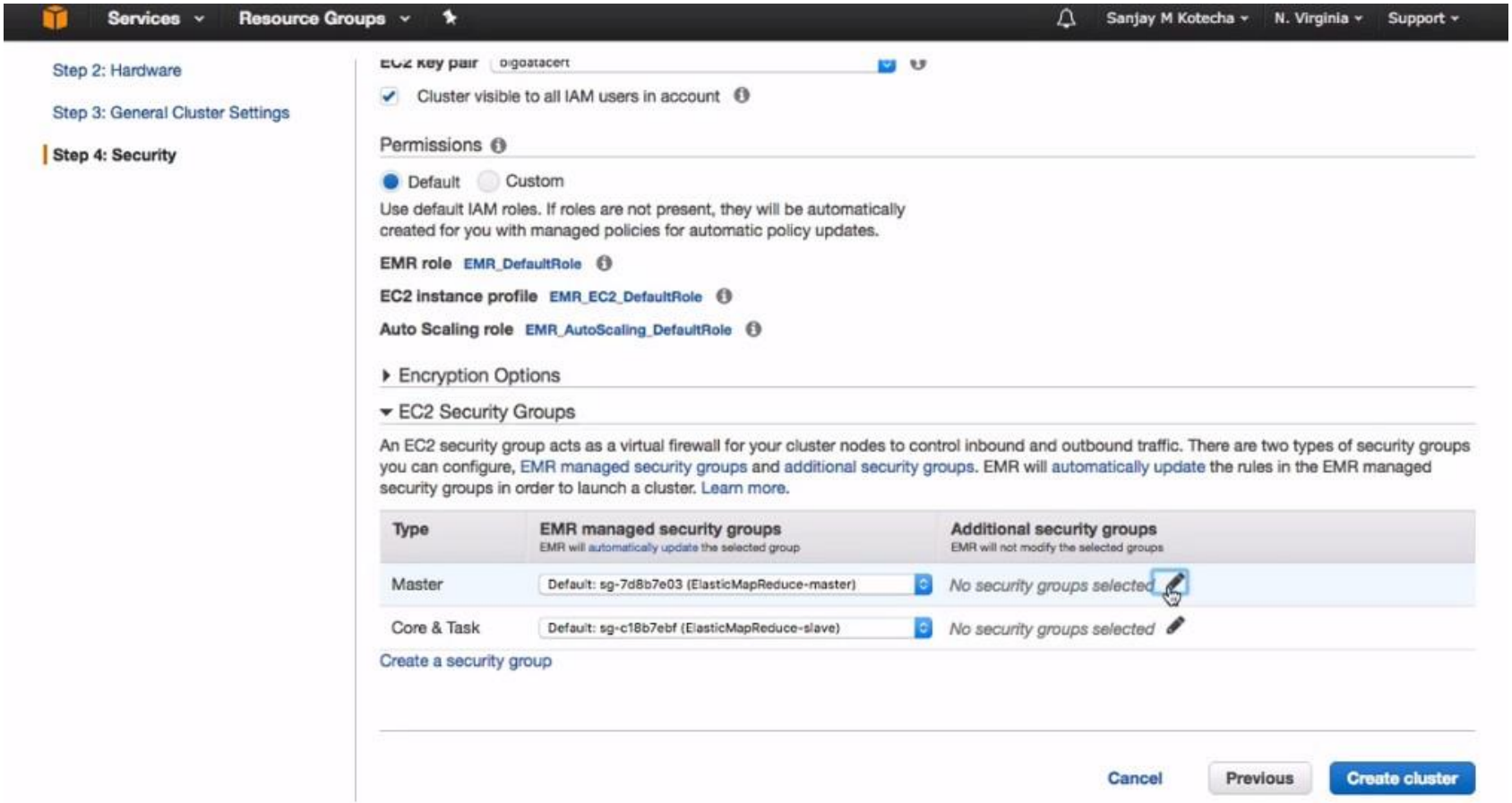
EMR Lab

7. Pada Security pilih EC2 Key pair yang sudah dibuat sebelumnya, lalu klik EC2 Security Options



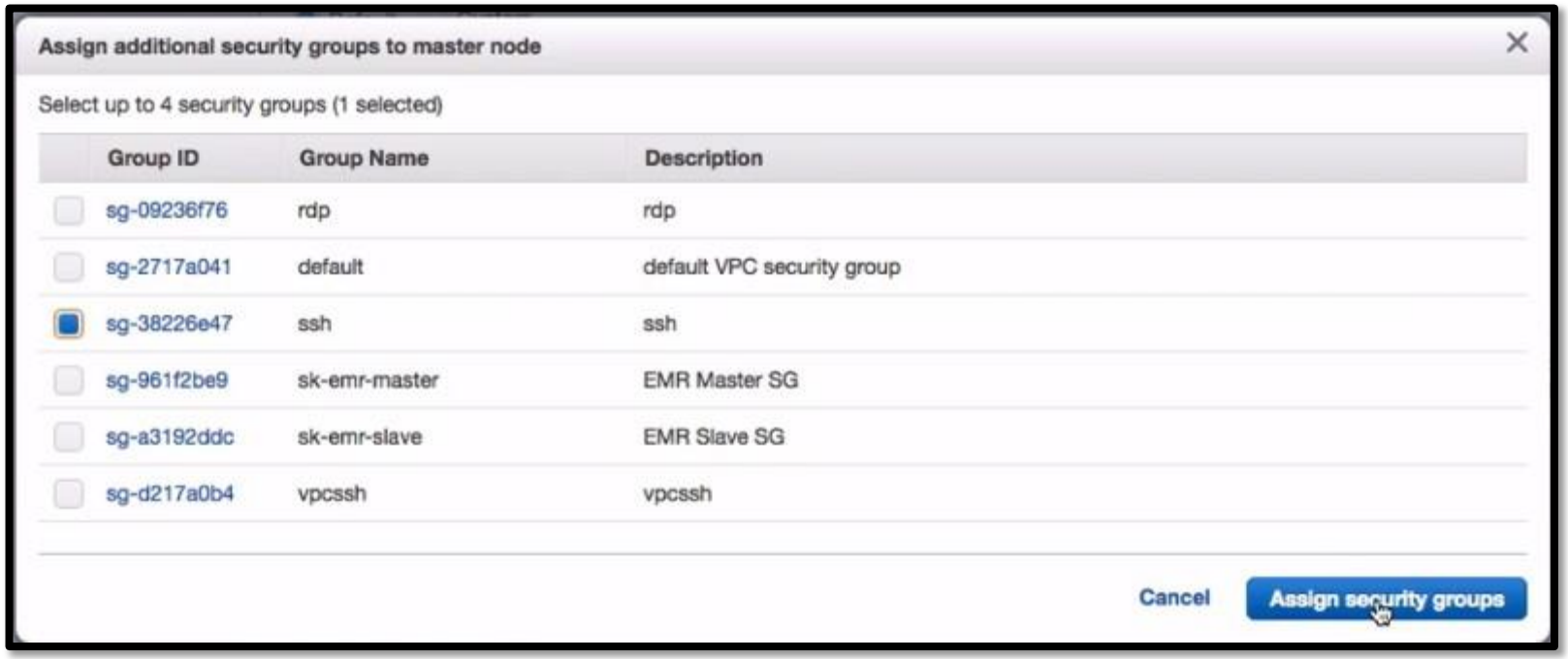
EMR Lab

8. Pada Master klik pilihan edit pada Additional security group



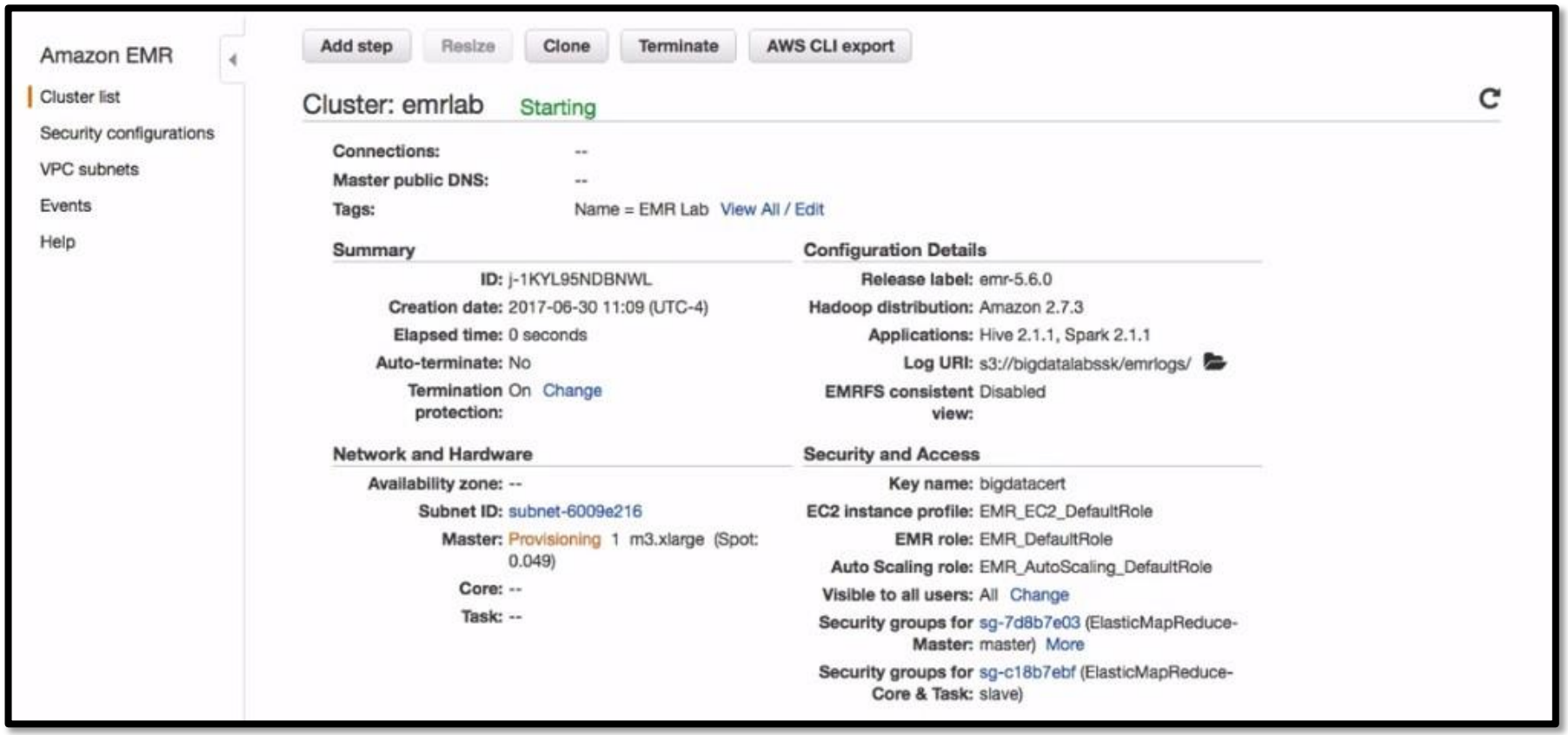
EMR Lab

9. Lalu, pilih ssh security group yang telah dibuat sebelumnya, dan klik Assign security group, Lalu klik Create Cluster



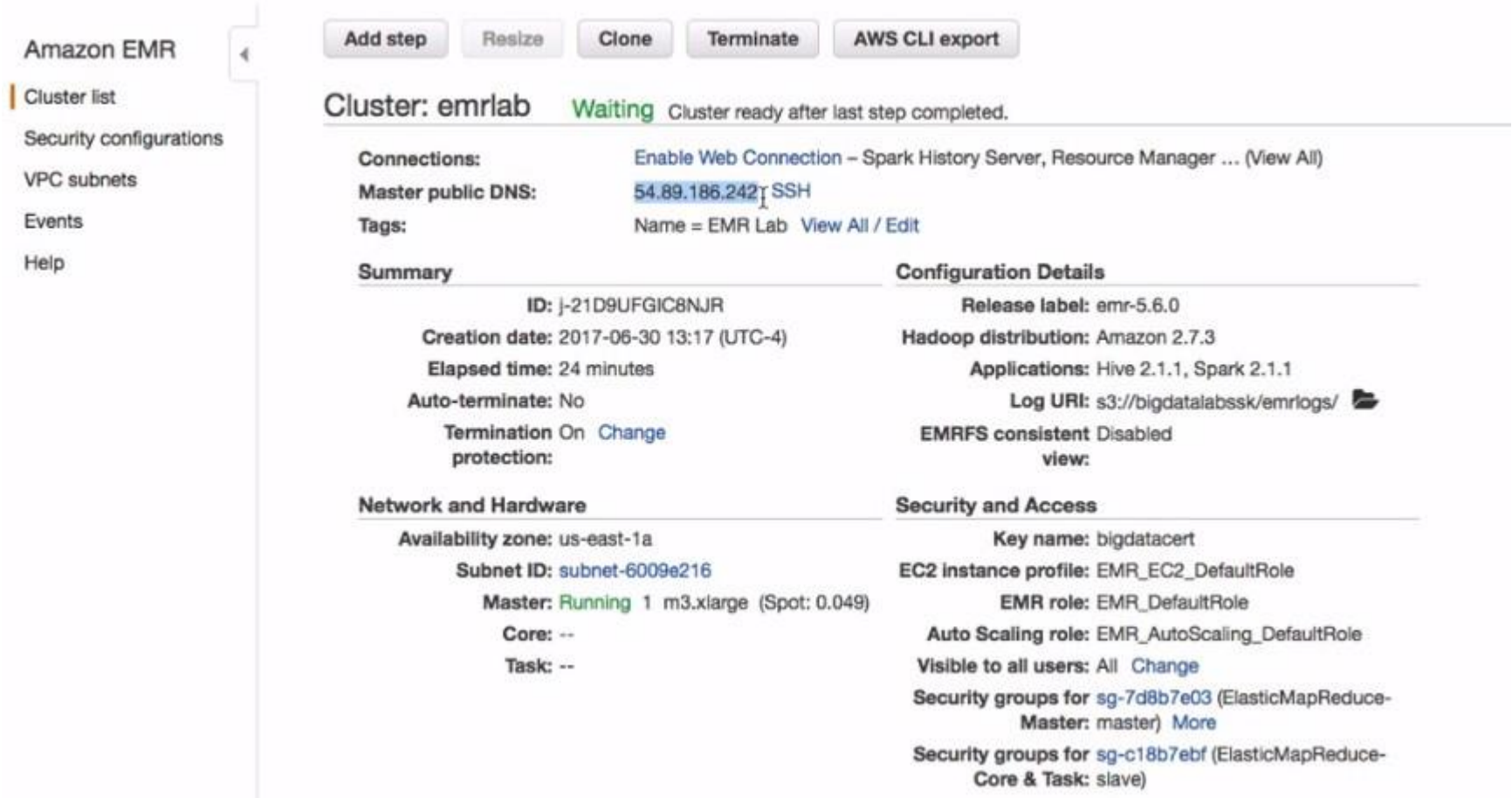
EMR Lab

10. Akhirnya Cluster telah dibuat



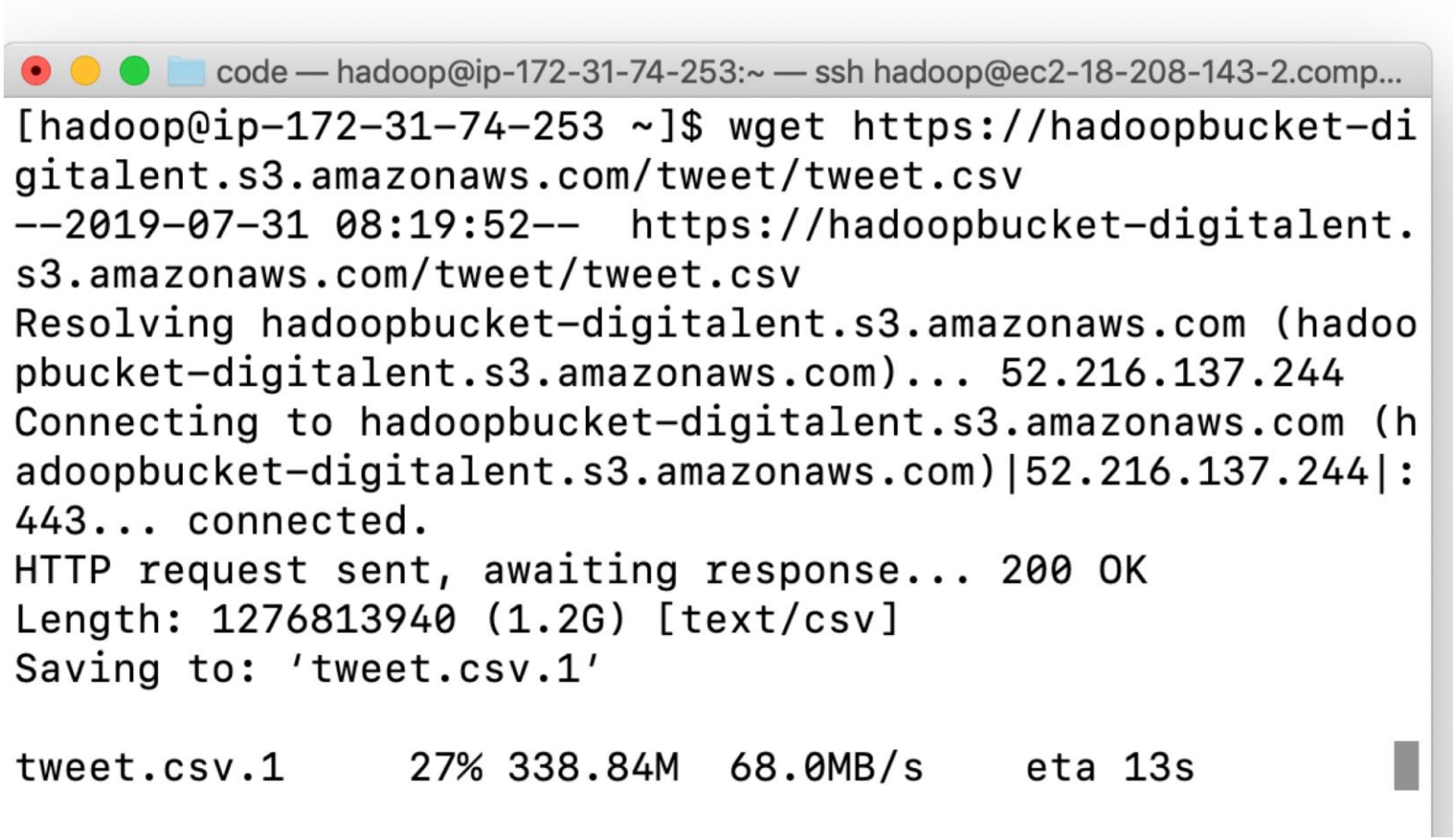
EMR Lab

11. Copy Master public DNS, dan buka sesion terminal SSH



Download dataset

11. wget https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv



Copy data ke HDFS

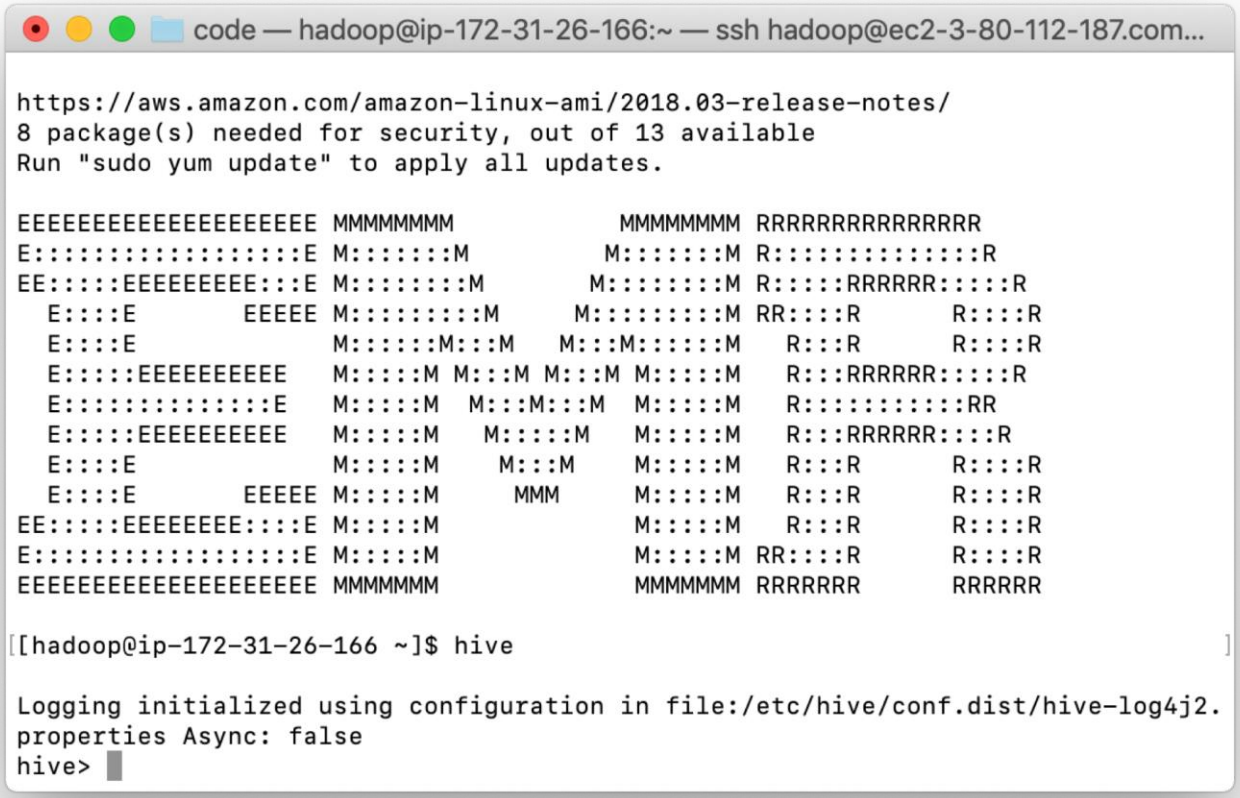
11. hdfs dfs -put tweet.csv /user/hadoop/

code — hadoop@ip-172-31-74-253:~ — ssh hadoop@ec2-18-208-143-2.comp...

```
pbucket-digitalent.s3.amazonaws.com)... 52.217.38.204
Connecting to hadoopbucket-digitalent.s3.amazonaws.com (h
adoopbucket-digitalent.s3.amazonaws.com)|52.217.38.204|:4
43... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1276813940 (1.2G) [text/csv]
Saving to: 'tweet.csv'

tweet.csv           100%[==================>]   1.19G  51.4
MB/s    in 21s

2019-07-31 07:20:20 (58.0 MB/s) - 'tweet.csv' saved [1276
813940/1276813940]

[hadoop@ip-172-31-74-253 ~]$ hdfs dfs -put tweet.csv /use
r/hadoop/tweet.csv
```

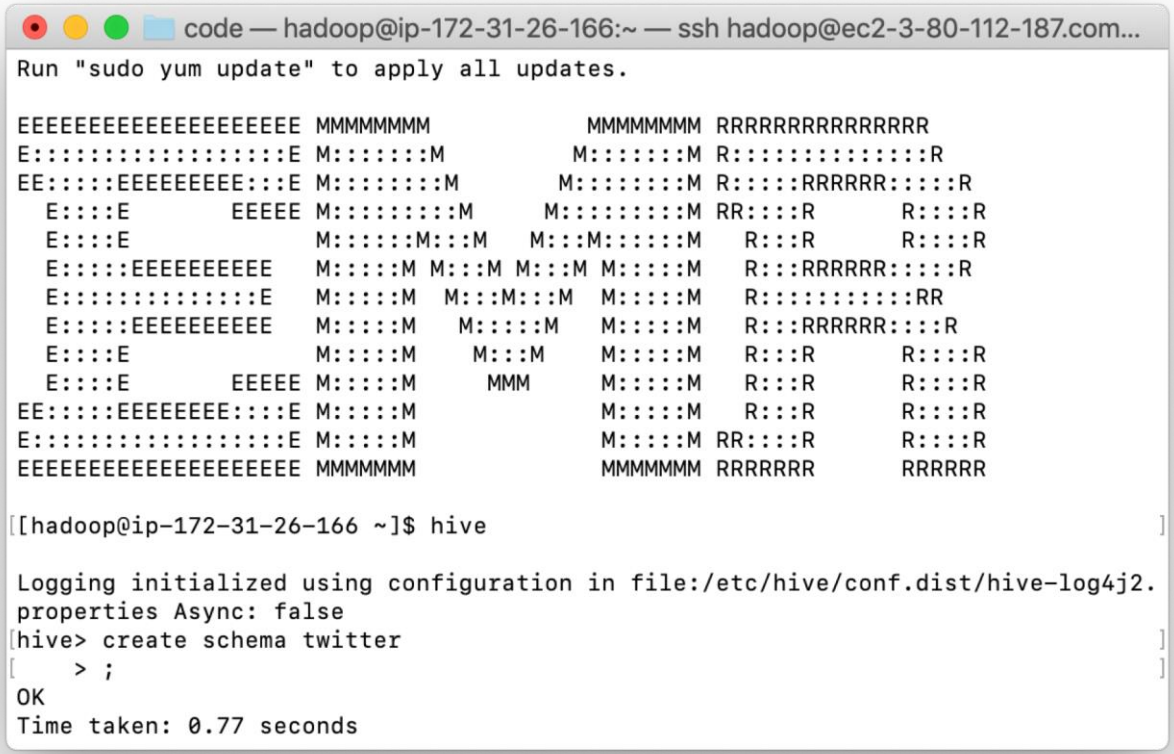
Menggunakan Hive

12. Ketik ‘hive’ di terminal



Membuat schema di EMR

13. Ketik ‘create schema twitter;’



Import data dari S3

14. Perintah : https://pastebin.com/6P1D7Pym

```
properties Async: false
hive> CREATE EXTERNAL TABLE twitter.tweet (created_at STRING, in_reply_to_status
_id STRING, id_str STRING, retweeted BOOLEAN, in_reply_to_user_id_str STRING, co
ordinates STRING, retweet_count INT, contributors STRING, favorite_count INT,fav
orited BOOLEAN, in_reply_to_status_id_str STRING, source STRING, in_reply_to_use
r_id STRING, user_id_str STRING, user_screen_name STRING, place STRING, geo STRI
NG, text STRING) LOCATION 's3://hadoopbucket-digitalent/tweet/' TBLPROPERTIES ("
skip.header.line.count"="1");
OK
Time taken: 4.782 seconds
hive>
```

Cek banyak baris pada tabel tweet

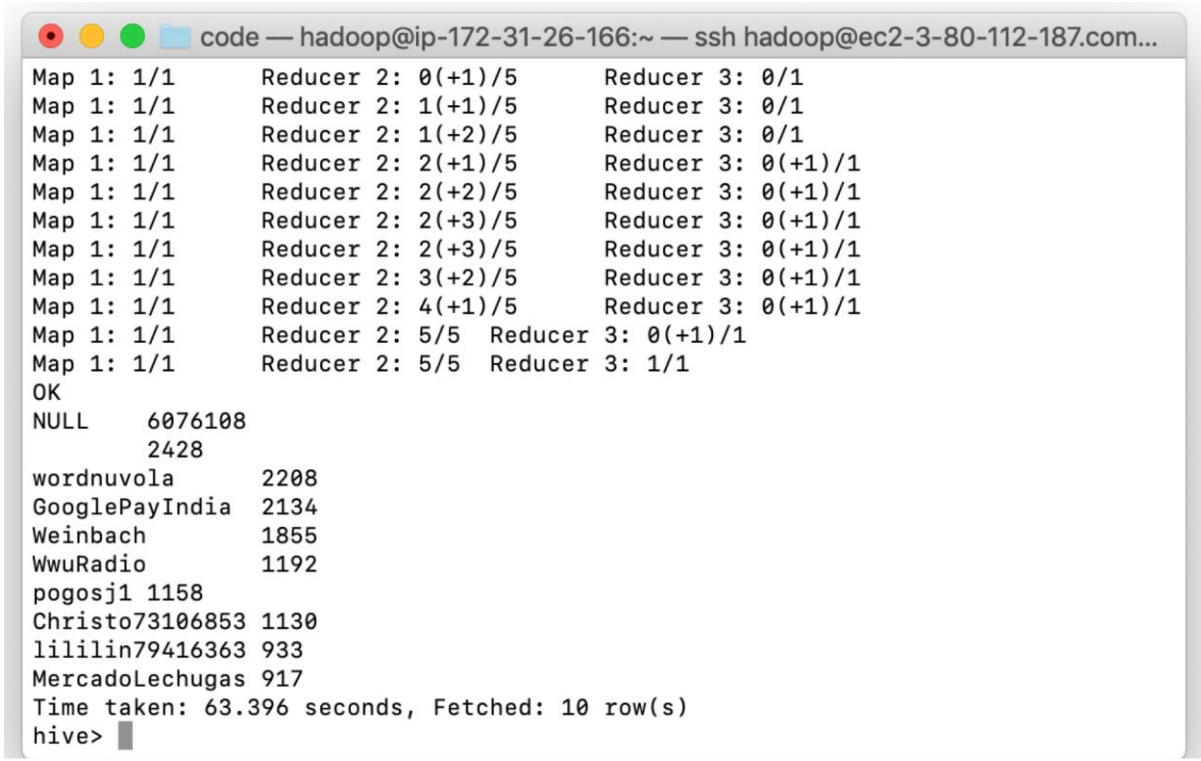
14. Perintah : select count(*) from tweet;

code — hadoop@ip-172-31-26-166:~ — ssh hadoop@ec2-3-80-112-187.com...

```
hive> select count(*) from tweet;
Query ID = hadoop_20190731043722_455b0eff-8a29-46d6-bcd7-36d5dc38f7b1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1564545374633
_0003)

Map 1: -/-      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0/1      Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 0(+1)/1  Reducer 2: 0/1
Map 1: 1/1      Reducer 2: 0(+1)/1
Map 1: 1/1      Reducer 2: 1/1
OK
11262928
Time taken: 31.412 seconds, Fetched: 1 row(s)
```

User paling sering nge-tweet

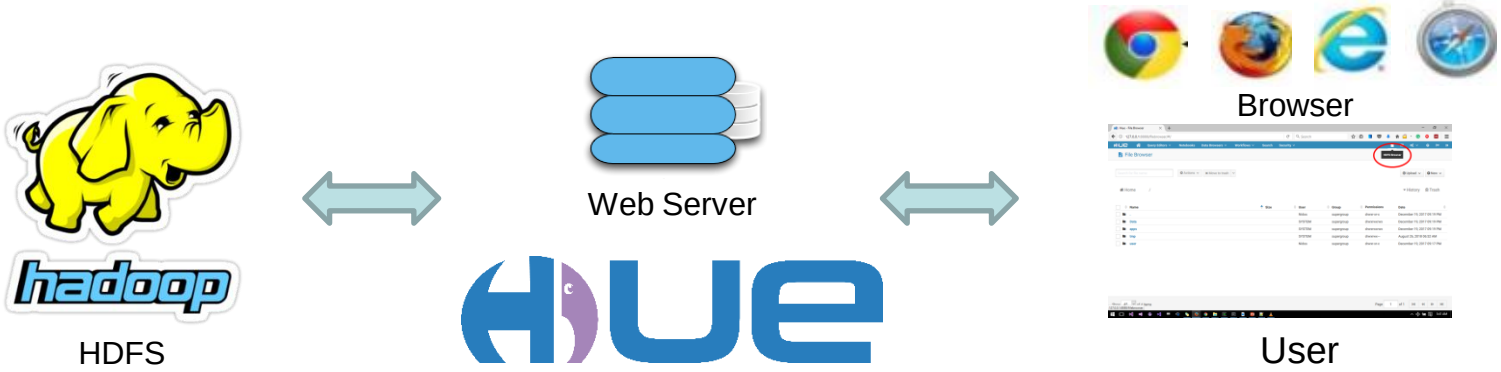
15. Perintah : select user_screen_name,count(*) as c from tweet group by user_screen_name order by c desc limit 10;



Lab Hue

Tentang Hue

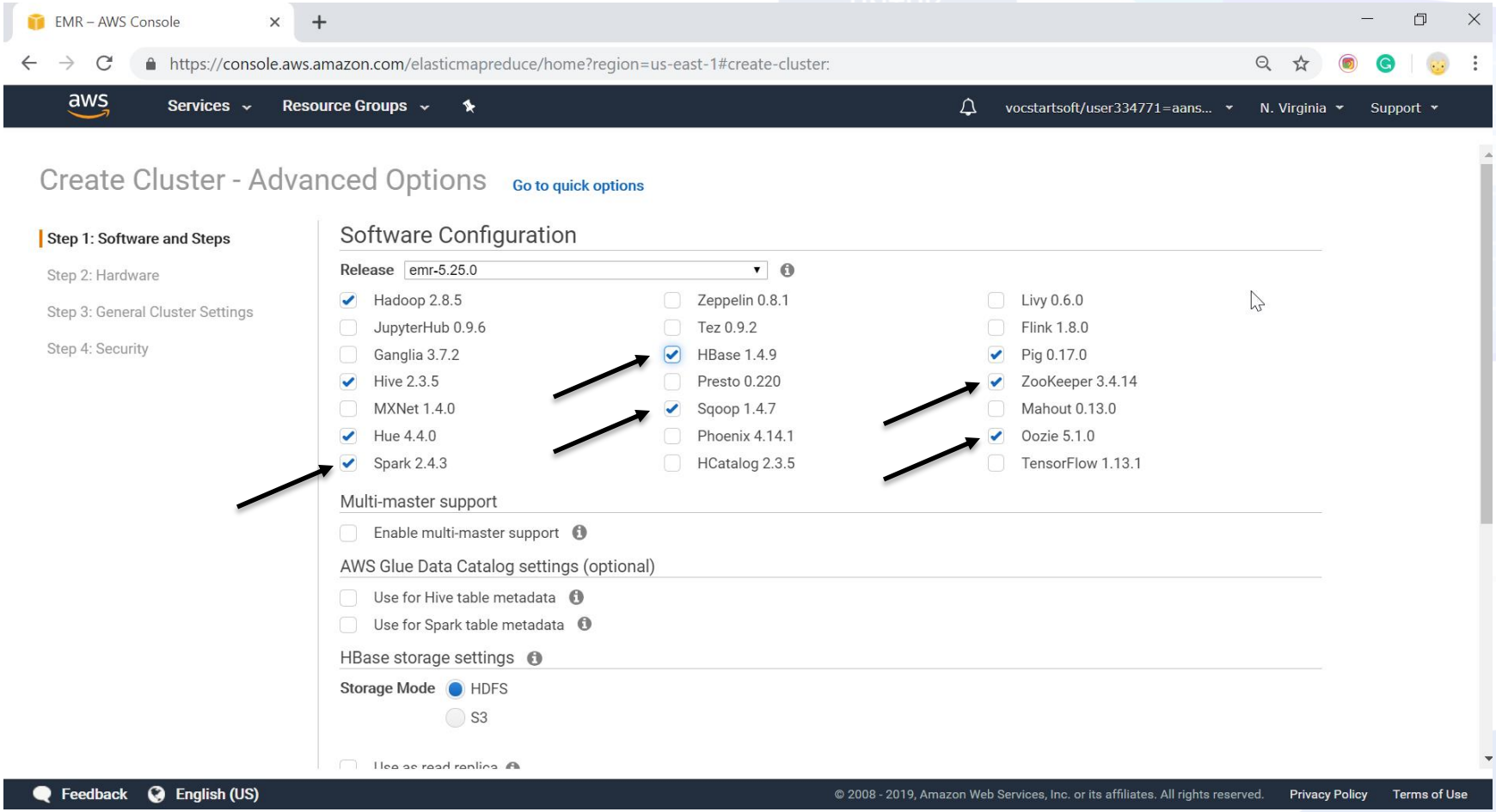
- Hue (*Hadoop User Experience*) adalah antarmuka berbasis Web untuk memudahkan dalam memonitoring atau memanage (create, delete, edit, etc) data HDFS pada Apache Hadoop dan beberapa fungsi lainnya, serta bisa diinstal di pc/notebook dengan versi hadoop manapun.



- Membuat “any users” untuk lebih fokus pada big data processing.

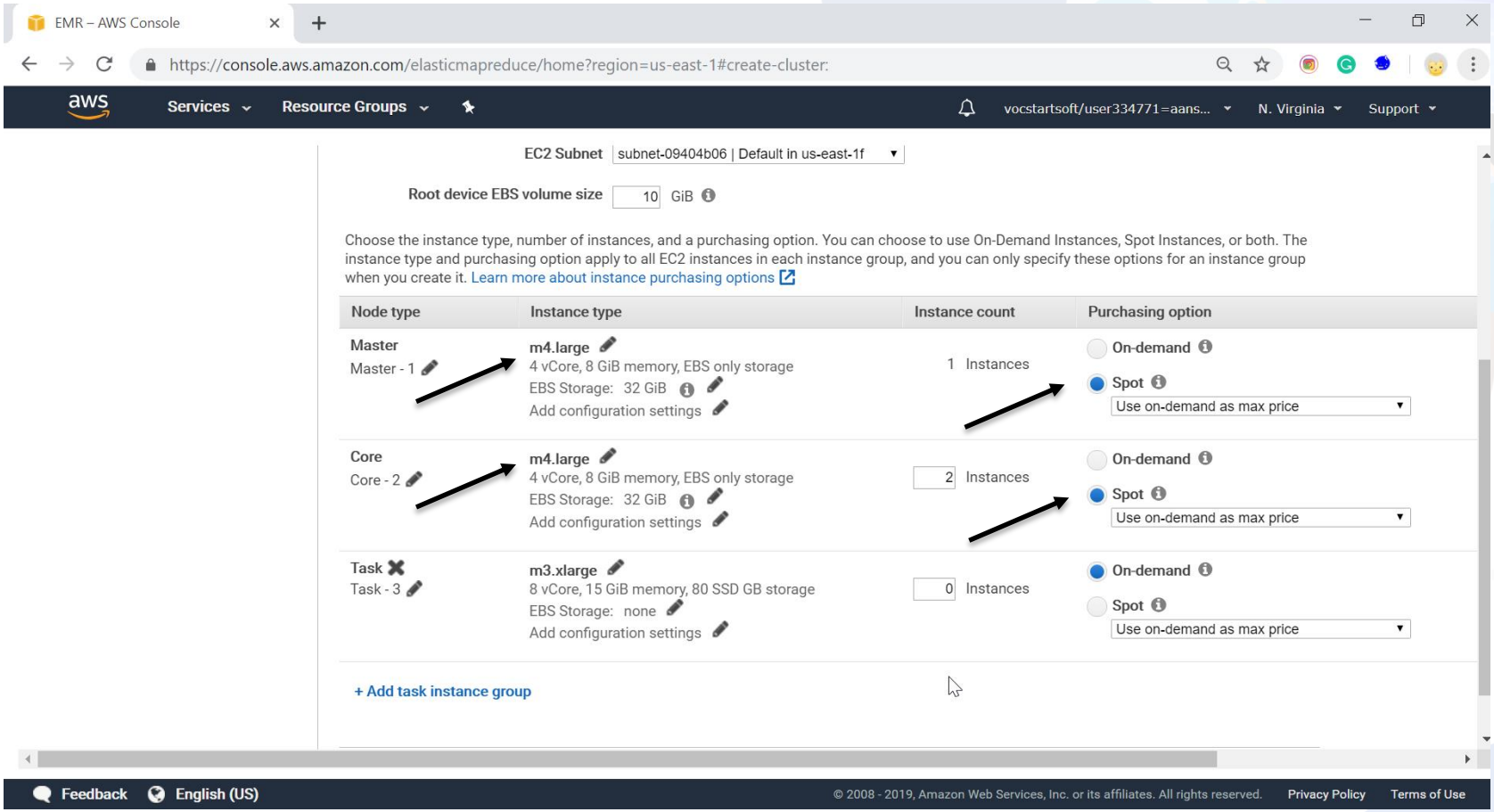
Tutorial Hue

Buat cluster pada layanan EMR, beri centang seperti pada gambar dibawah, kemudian next



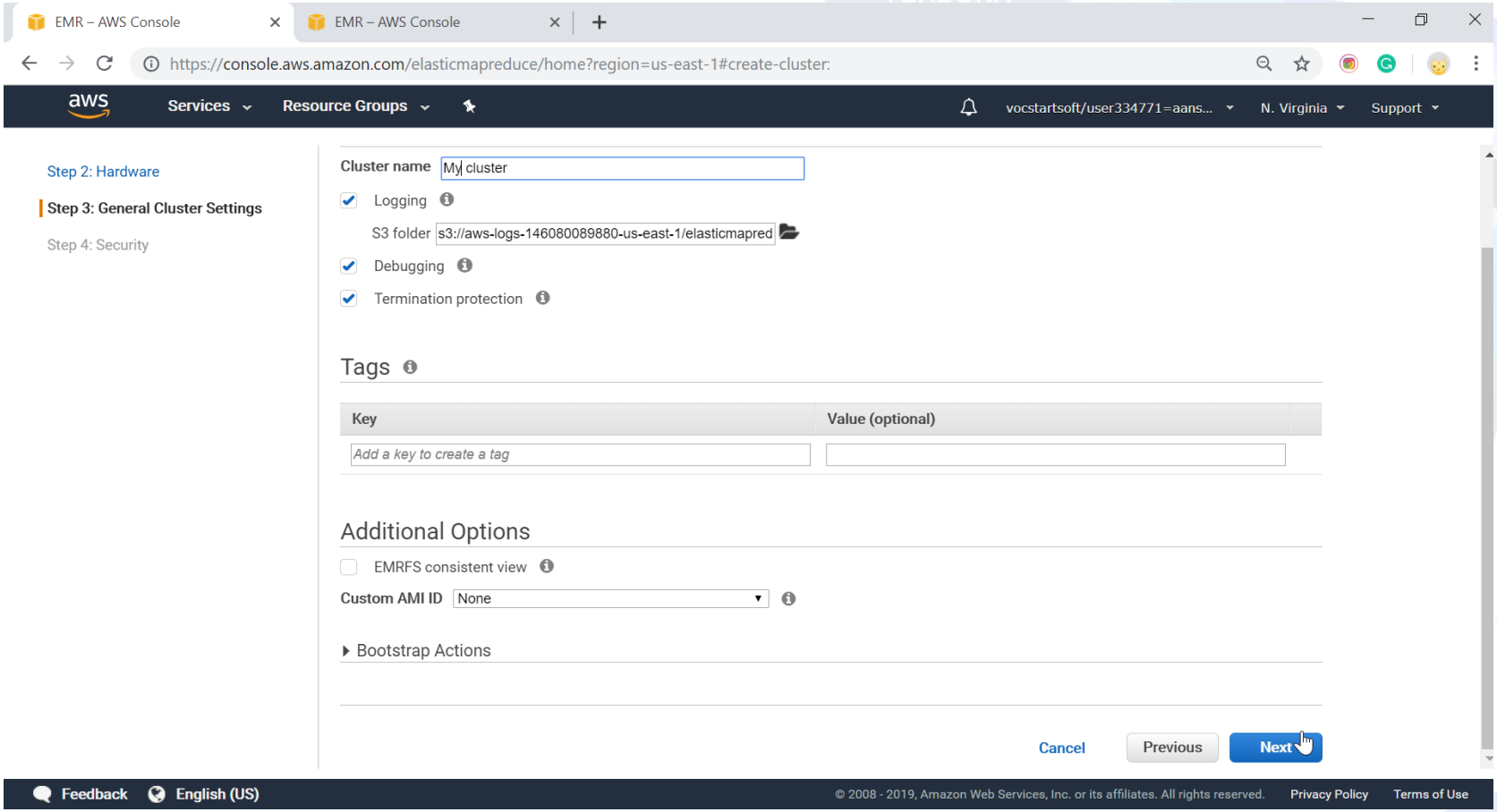
Tutorial Hue

Berikan mark Spot untuk master dan core node dan ubah instance type



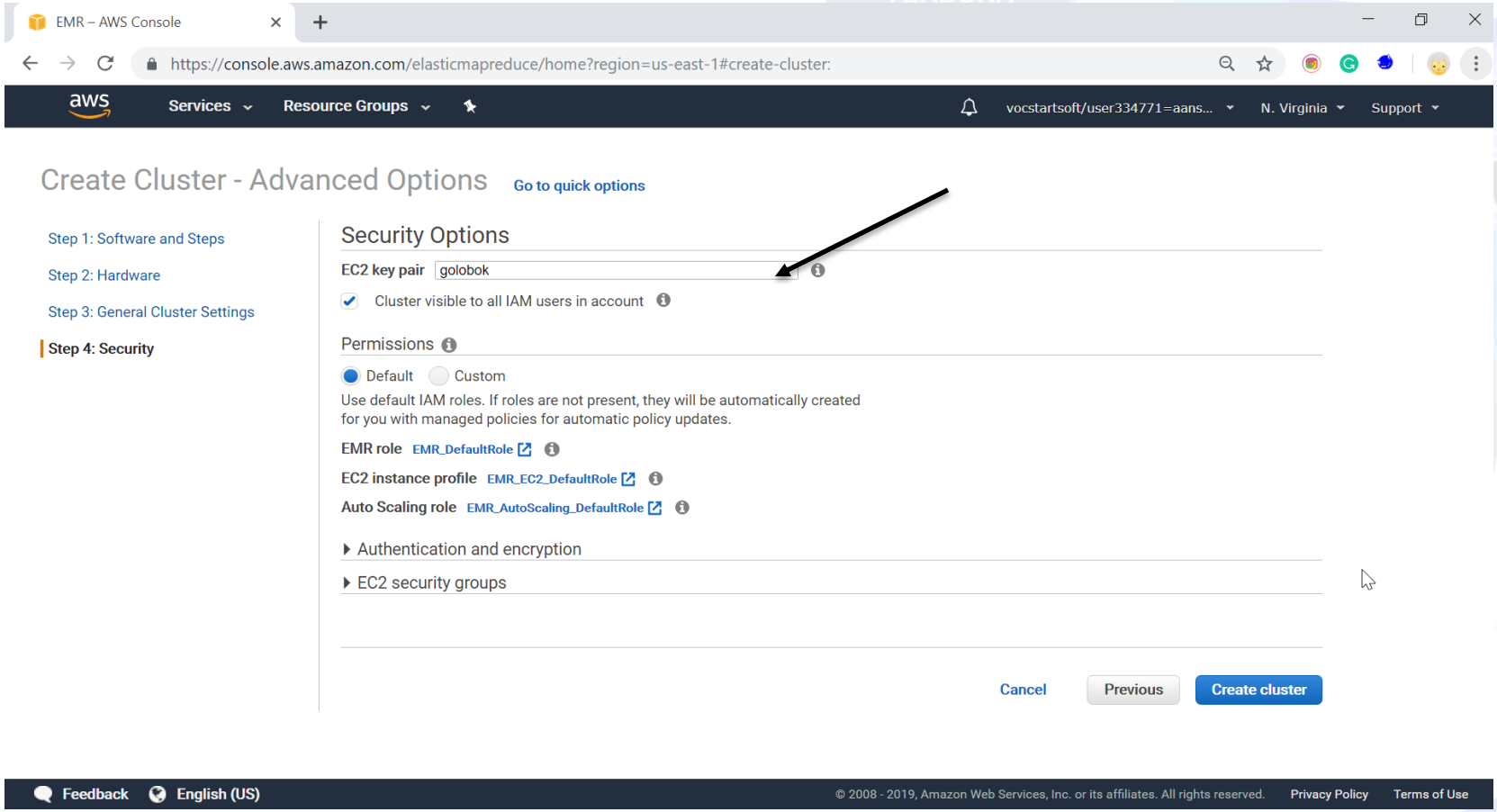
Tutorial Hue

Hingga muncul tampilan berikut, lanjutkan dengan klik Next



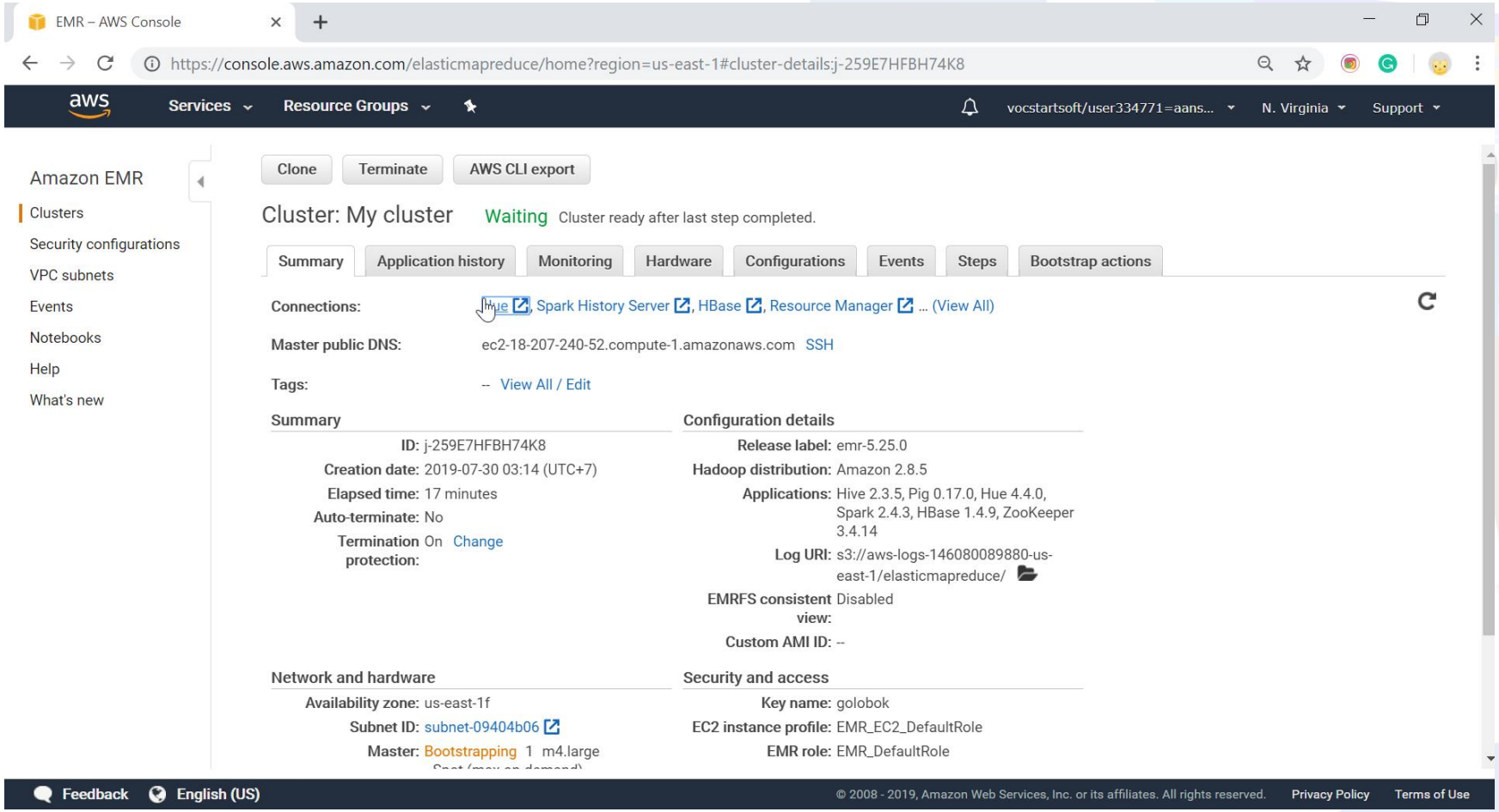
Tutorial Hue

Pilih EC2 key-pair yang pernah dibua, kemudian Create Cluster



Tutorial Hue

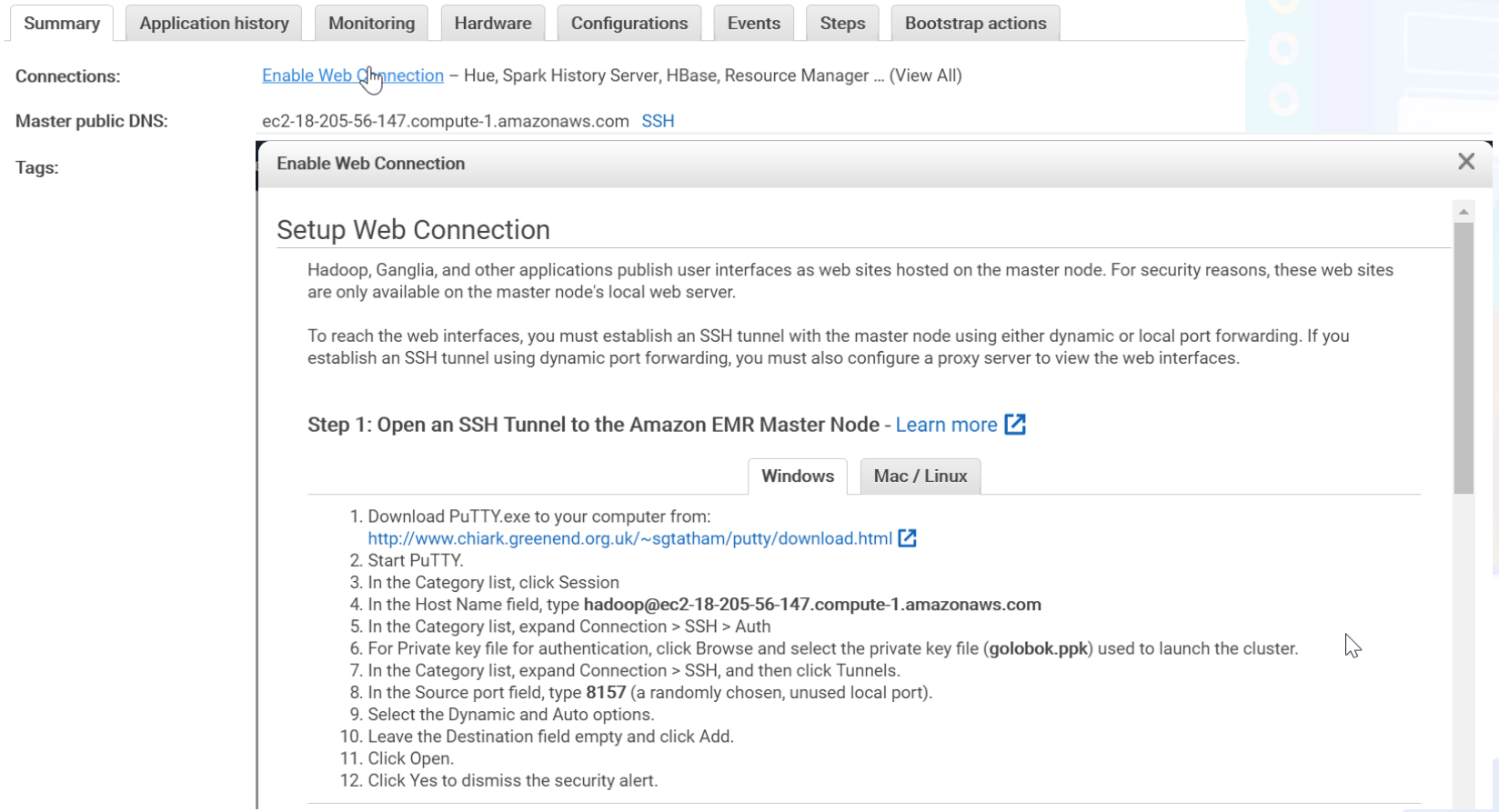
Tunggu status Starting menjadi Waiting. Kemudian pilih Hue



Tutorial Hue

Apabila pilihan Hue tidak tersedia, ikuti tutorial berikut:

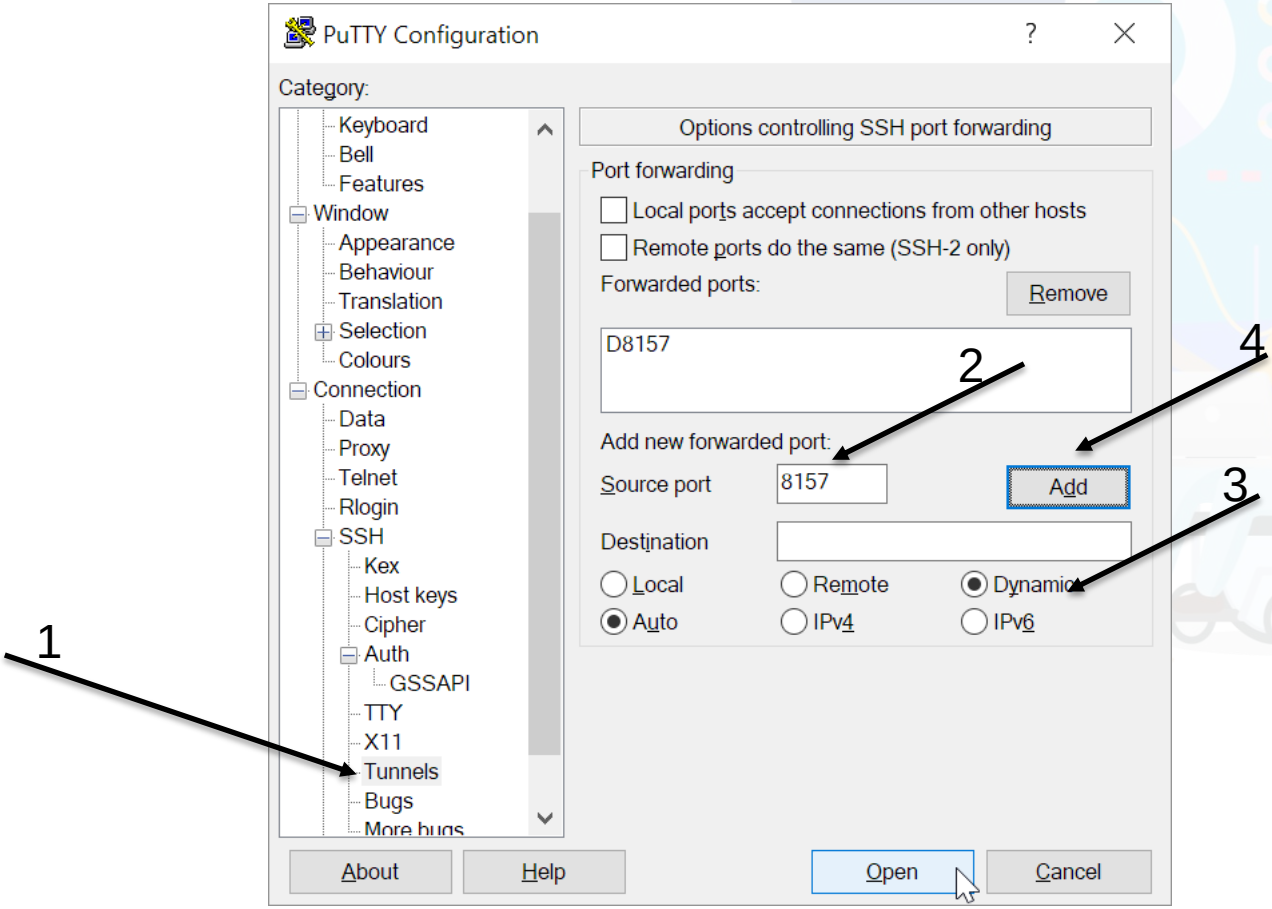
- **Klik Enable Web Connection dan ikuti langkah-langkahnya**



Tutorial Hue

Buka putty dan masukkan konfigurasi seperti biasa,

Tambahkan pengaturan dibawah ini :



Tutorial Hue

Buat file baru dengan nama “foxyproxy-settings.xml”,

Buka dan paste script dibawah kemudian simpan.

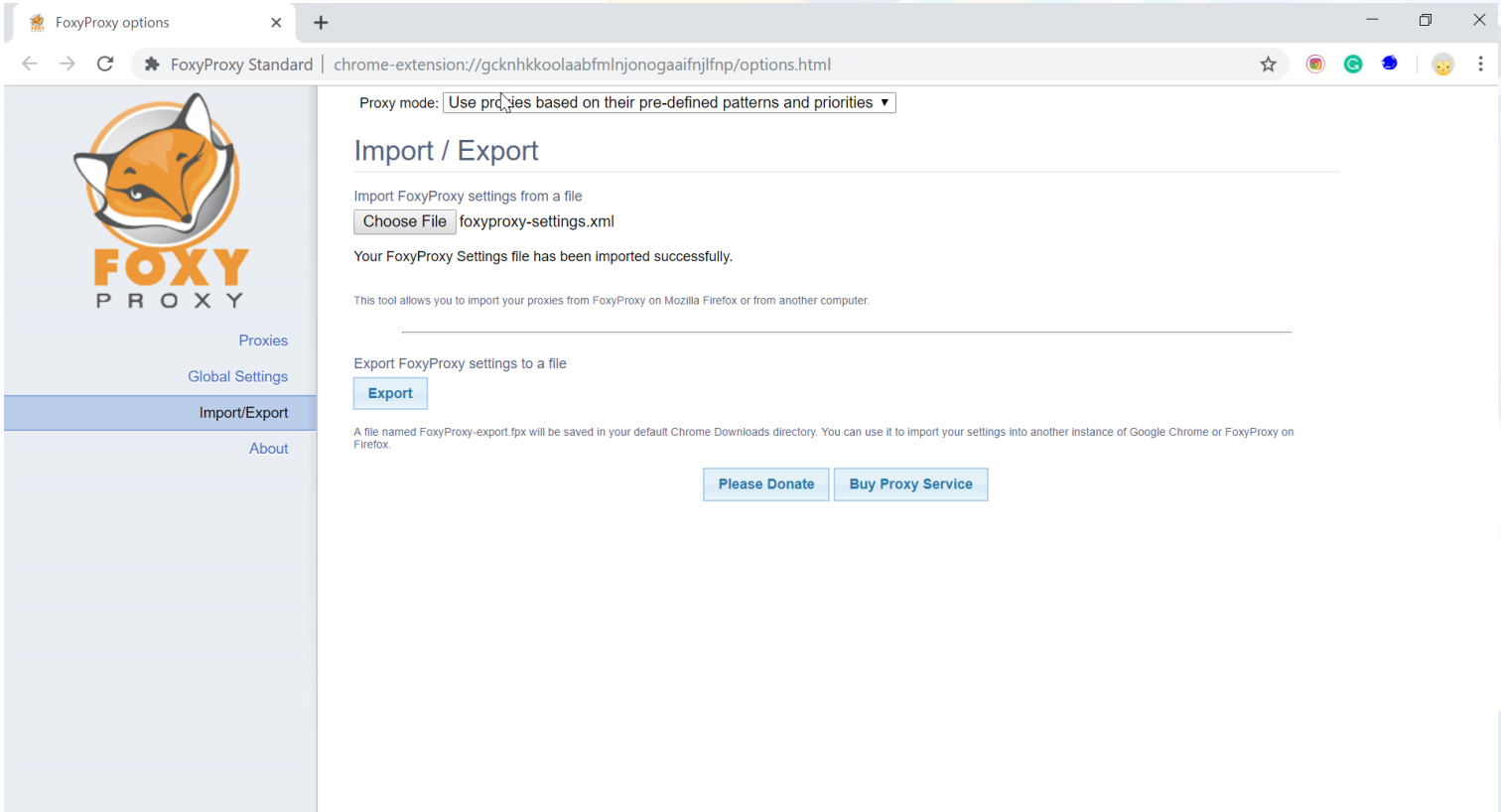
```
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
    <proxies>
        <proxy name="emr-socks-proxy" id="2322596116" notes="" fromSubscription="false" enabled="true"
mode="manual" selectedTabIndex="2" lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5"
proxyDNS="true" noInternalIPs="false" autoconfMode="pac" clearCacheBeforeUse="false" disableCache="false"
clearCookiesBeforeUse="false" rejectCookies="false">
            <matches>
                <match enabled="true" name="*ec2*.amazonaws.com*" pattern="*ec2*.amazonaws.com*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*ec2*.compute*" pattern="*ec2*.compute*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="10.*" pattern="http://10.*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*10*.amazonaws.com*" pattern="*10*.amazonaws.com*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*10*.compute*" pattern="*10*.compute*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*.compute.internal*" pattern="*.compute.internal*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*.ec2.internal*" pattern="*.ec2.internal*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
            </matches>
            <manualconf host="localhost" port="8157" socksversion="5" isSocks="true" username="" password=""
domain="" />
        </proxy>
    </proxies>
</foxyproxy>
```

Tutorial Hue

- **Buka chrome, tambahkan ekstensi berikut :**

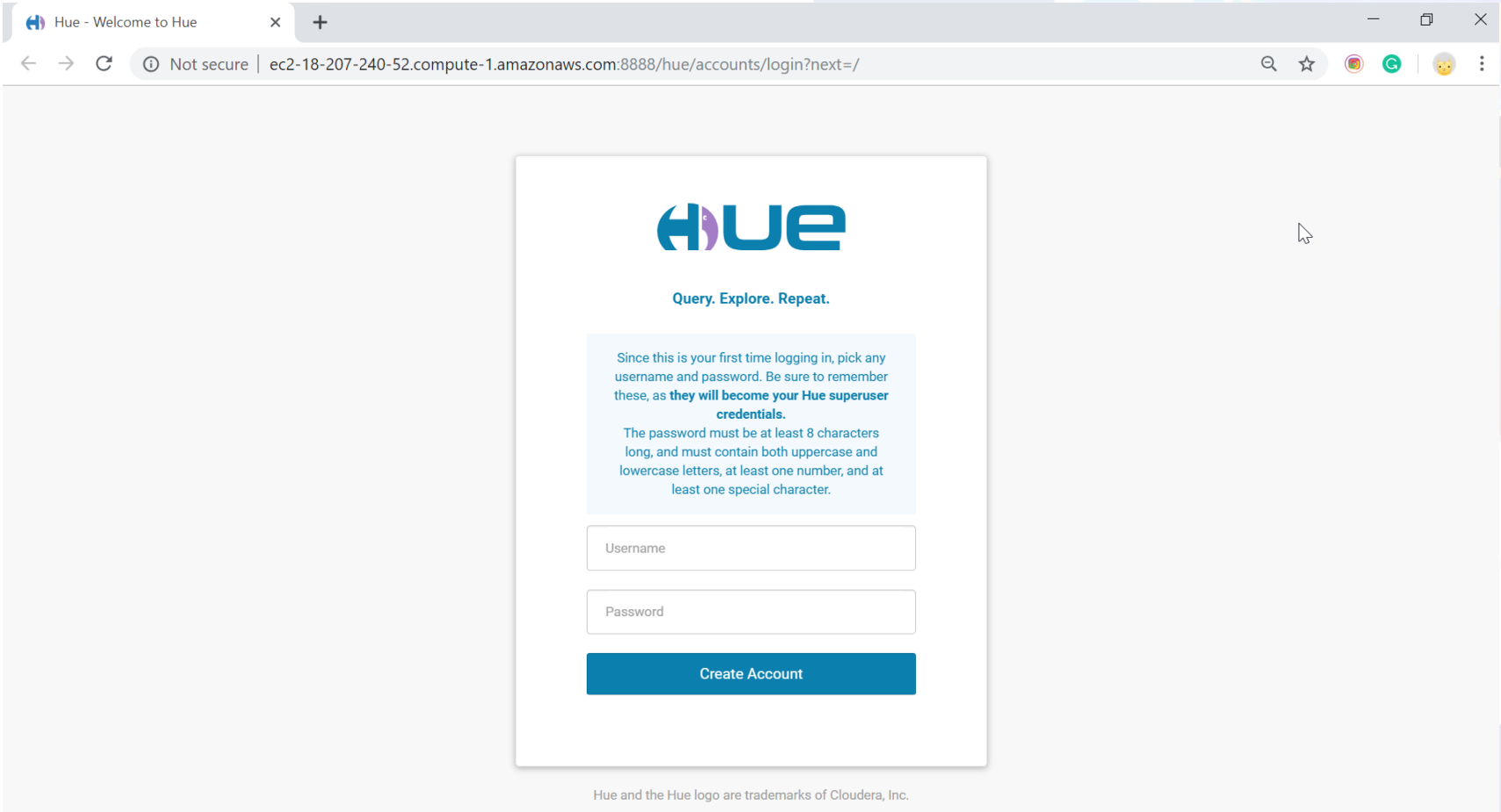
https://chrome.google.com/webstore/search/foxy%20proxy

- **Klik kanan ikon foxyproxy – options – Import/Export - Choose File - foxyproxy-settings.xml yang telah dibuat , kemudian pilih Add.**



Tutorial Hue

Untuk pertama kali, buat username dan password yang akan digunakan untuk login kedalam Hue

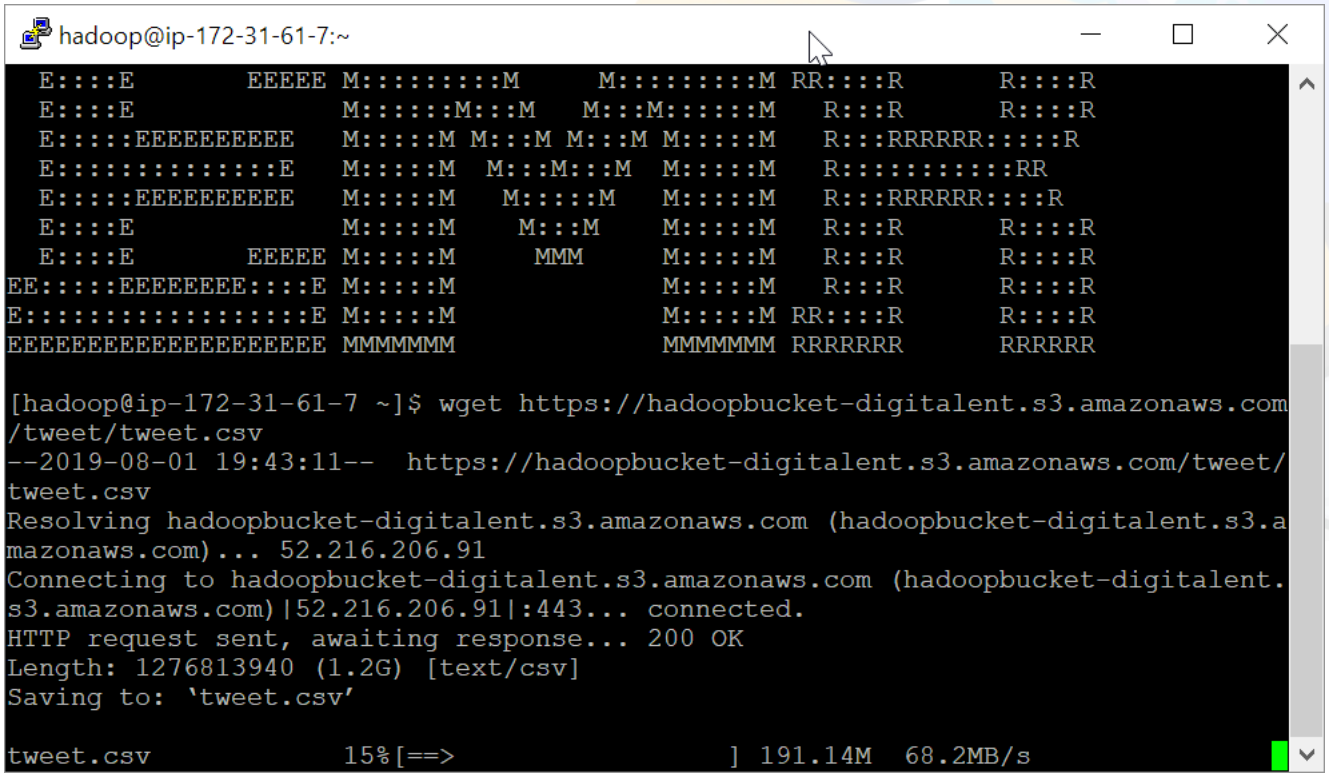


Tutorial Hue

Unduh data yang akan digunakan, disini menggunakan data twitter.csv

Ketikkan perintah berikut pada putty:

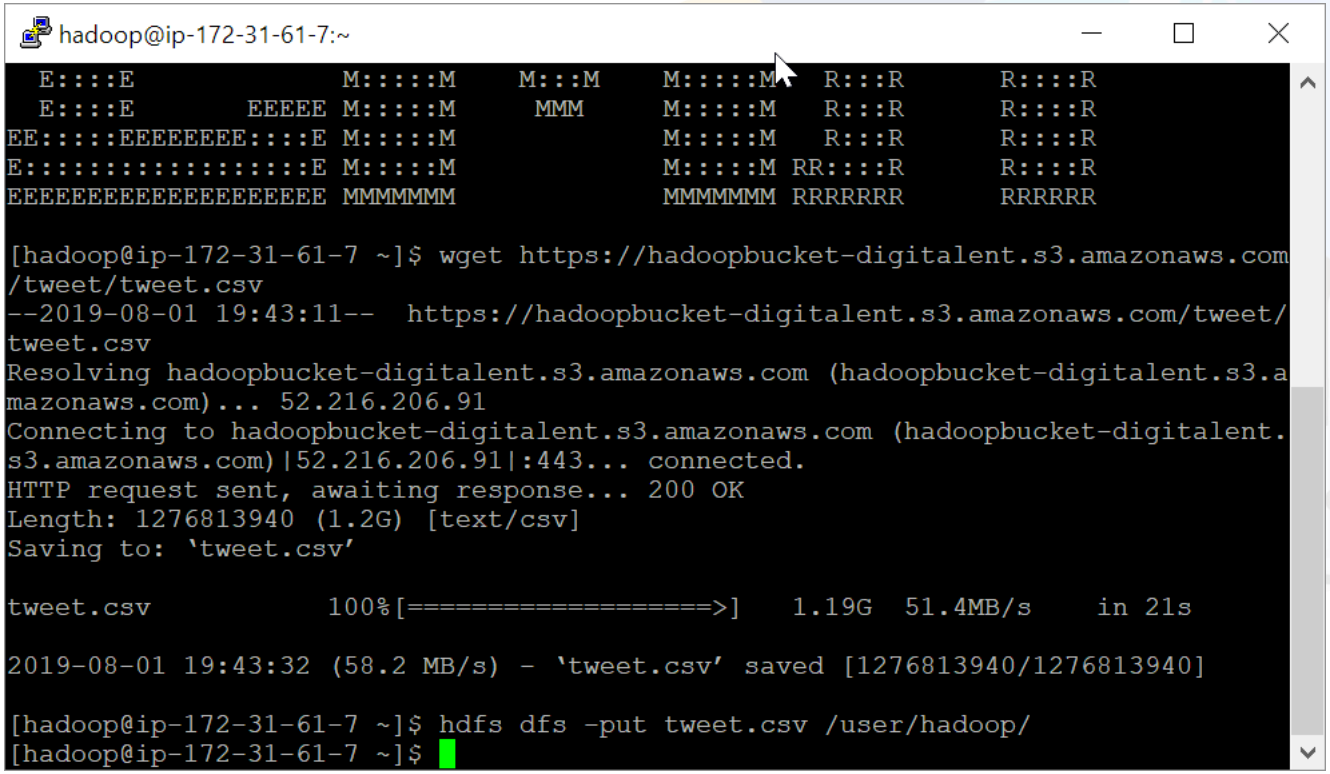
```
wget https://hadoopbucket-digitalent.s3.amazonaws.com/tweet/tweet.csv
```



Tutorial Hue

Salin file twitter.csv kedalam /user/hadoop/:

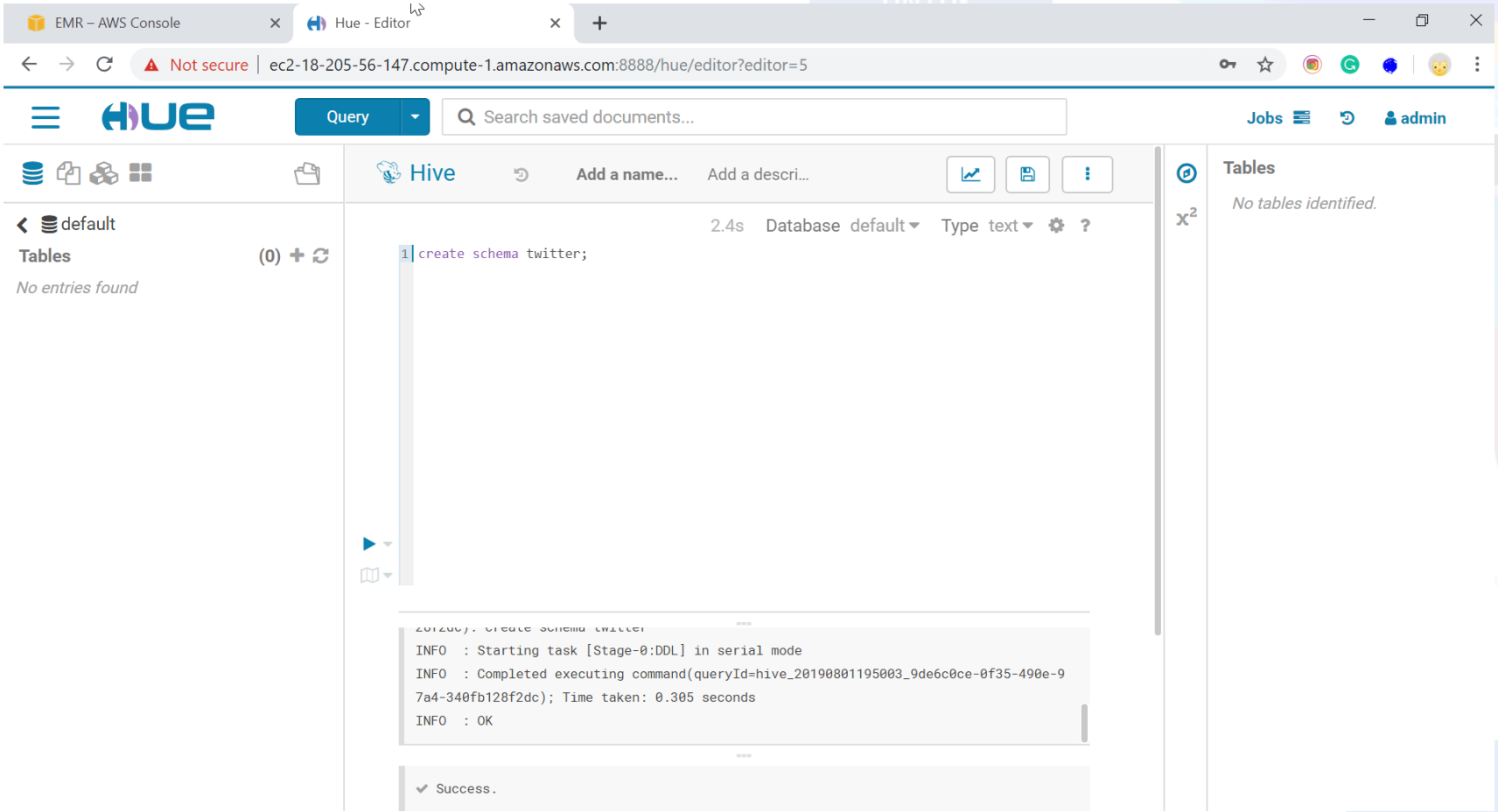
```
Hdfs dfs -put tweet.csv /user/Hadoop/
```



Tutorial Hue

Buka Hue kembali dan ketikkan perintah berikut:

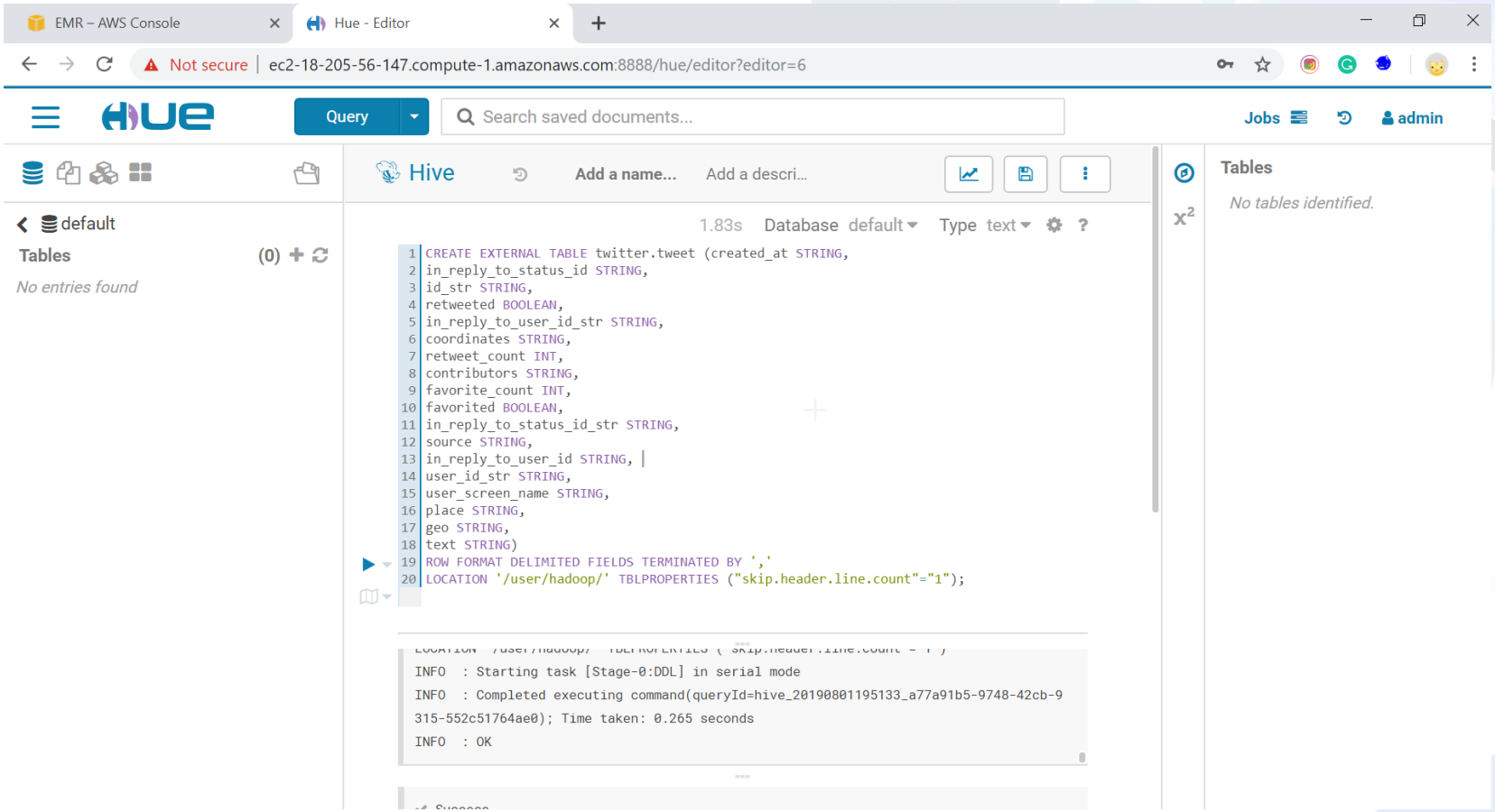
```
Create schema twitter;
```



Tutorial Hue

Create external table dengan perintah kueri dibawah:

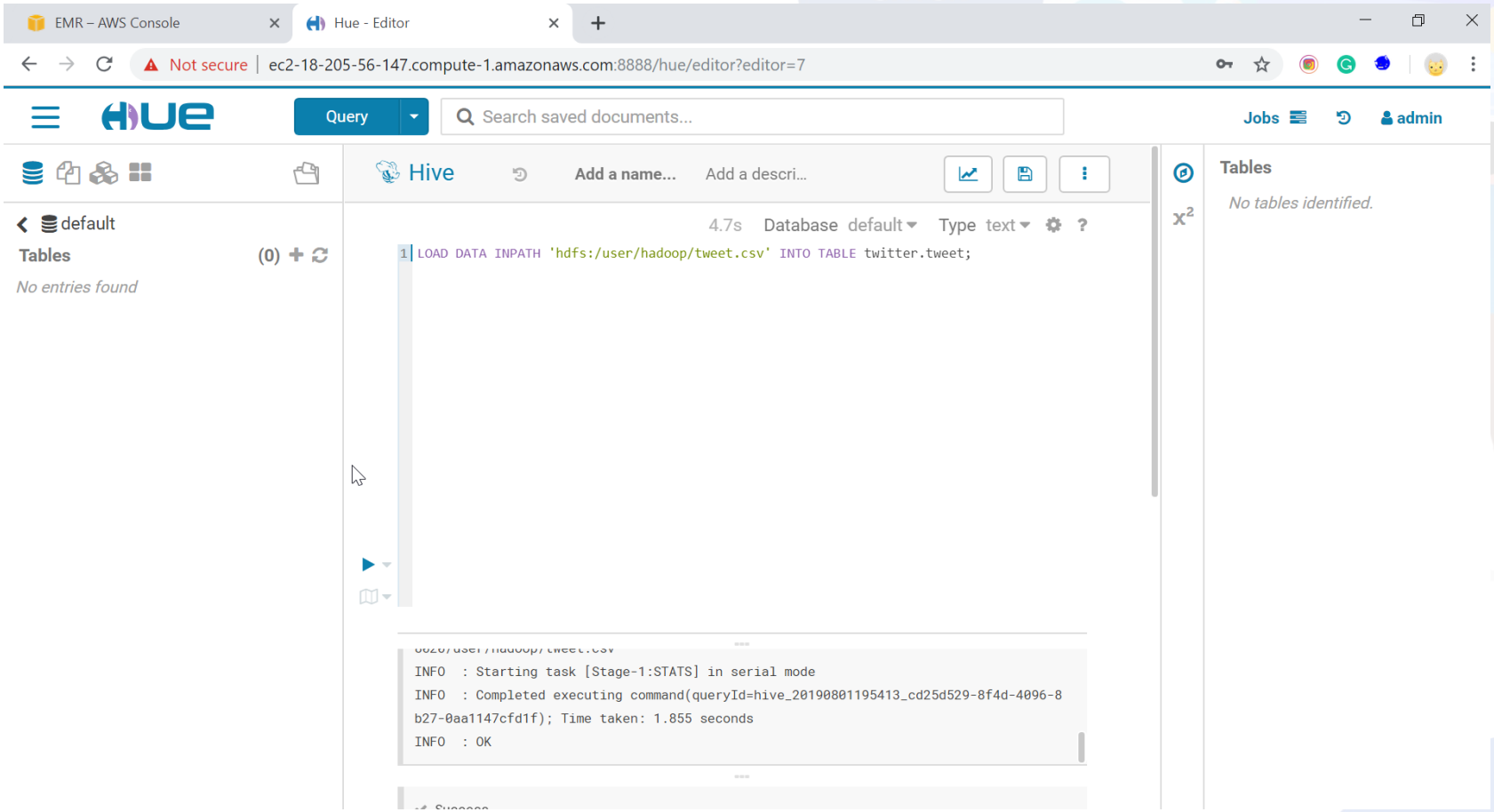
https://pastebin.com/6P1D7Pym



Tutorial Hue

Kemudian load data kedalam tabel:

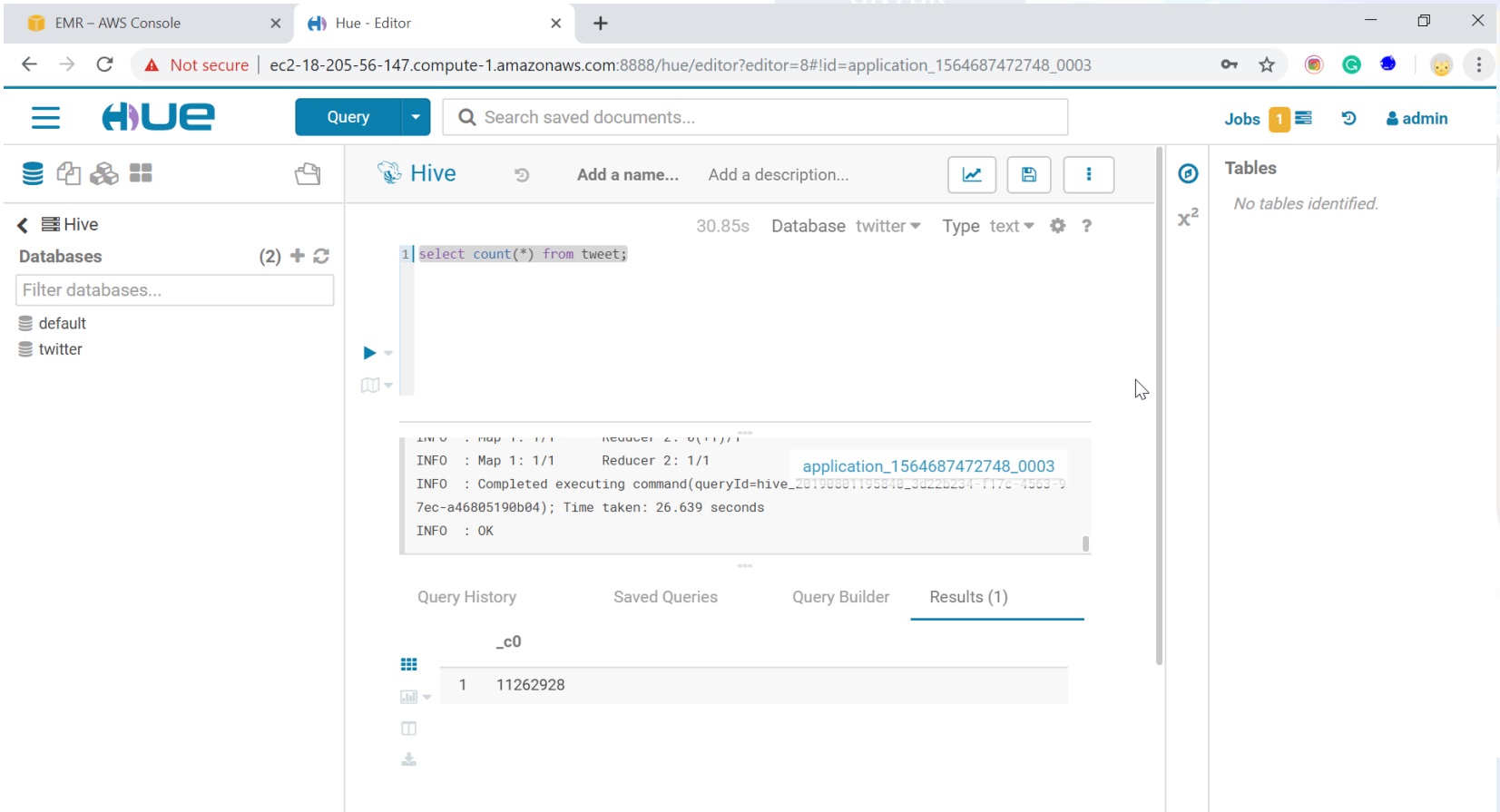
LOAD DATA INPATH 'hdfs:/user/hadoop/tweet.csv' INTO TABLE twitter.tweet;



Tutorial Hue

Coba kueri:

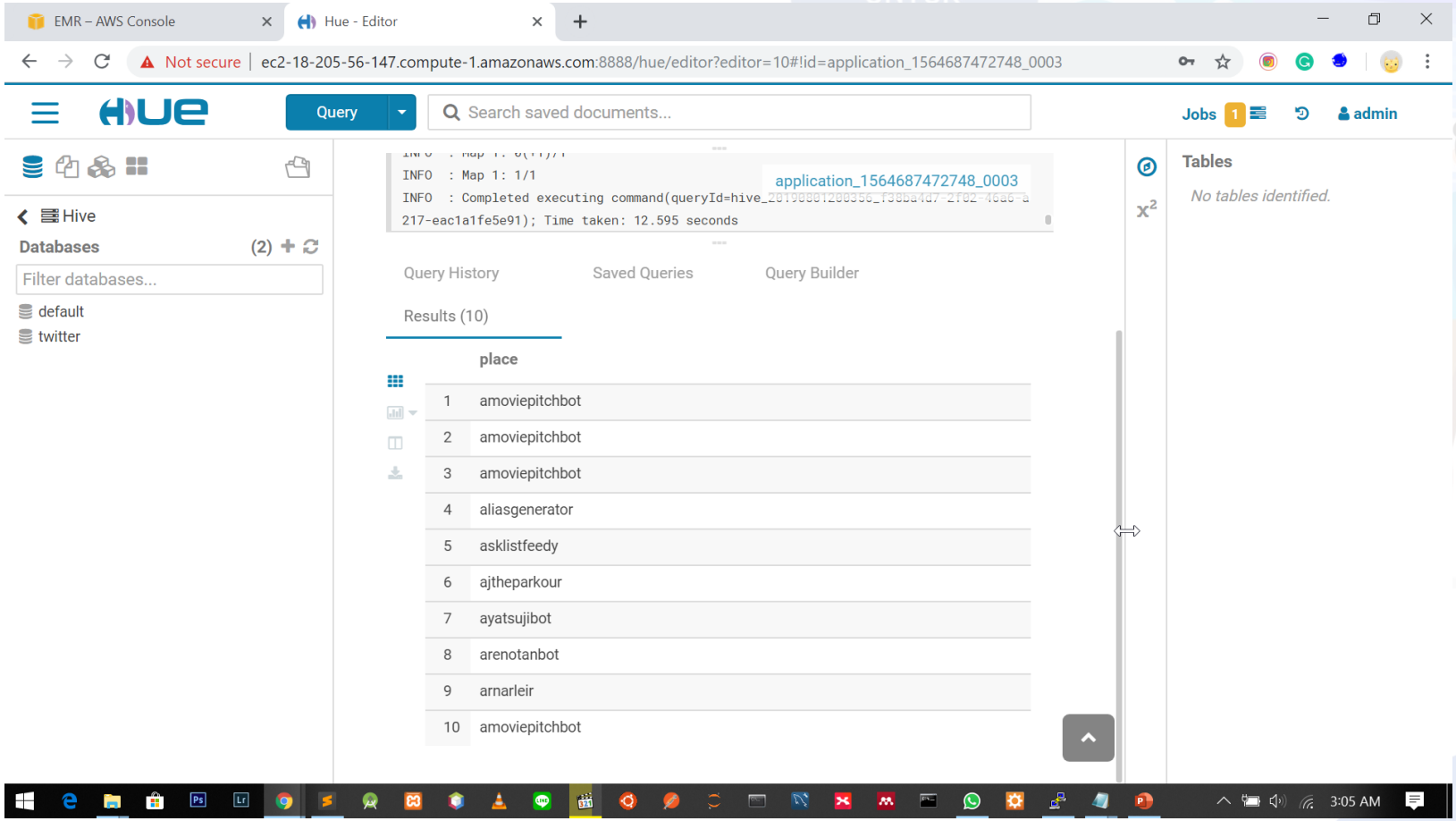
select count(*) from tweet;



Tutorial Hue

Coba kueri:

select place from tweet where place like 'a%' limit 10;



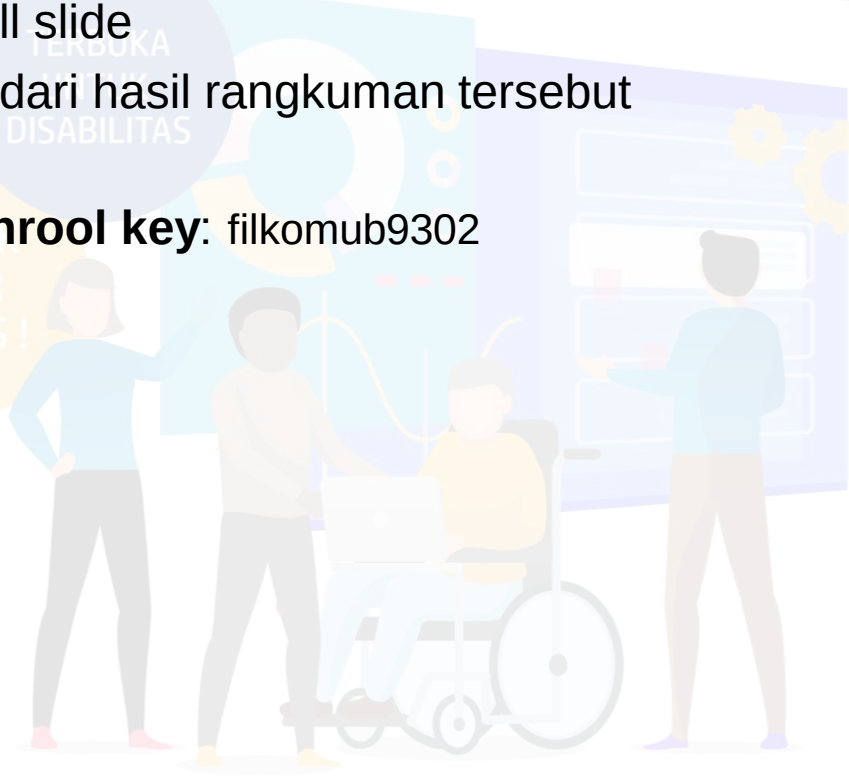
Penyedia Otentikasi

1. LDAP
2. PAM
3. SPNEGO
4. OpenID
5. Oauth
6. SAML2

Tugas Individu

1. Buatlah rangkuman materi dengan cara berikut:
   - Lalukan ulang seperti yang ada di All slide
   - Cek plagiasi diturnitin (tiap minggu) dari hasil rangkuman tersebut

     > Register ke turnitin
     > Masukkan **id class**: 21563495 & **enrool key**: filkomub9302



filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics

Terimakasih

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)