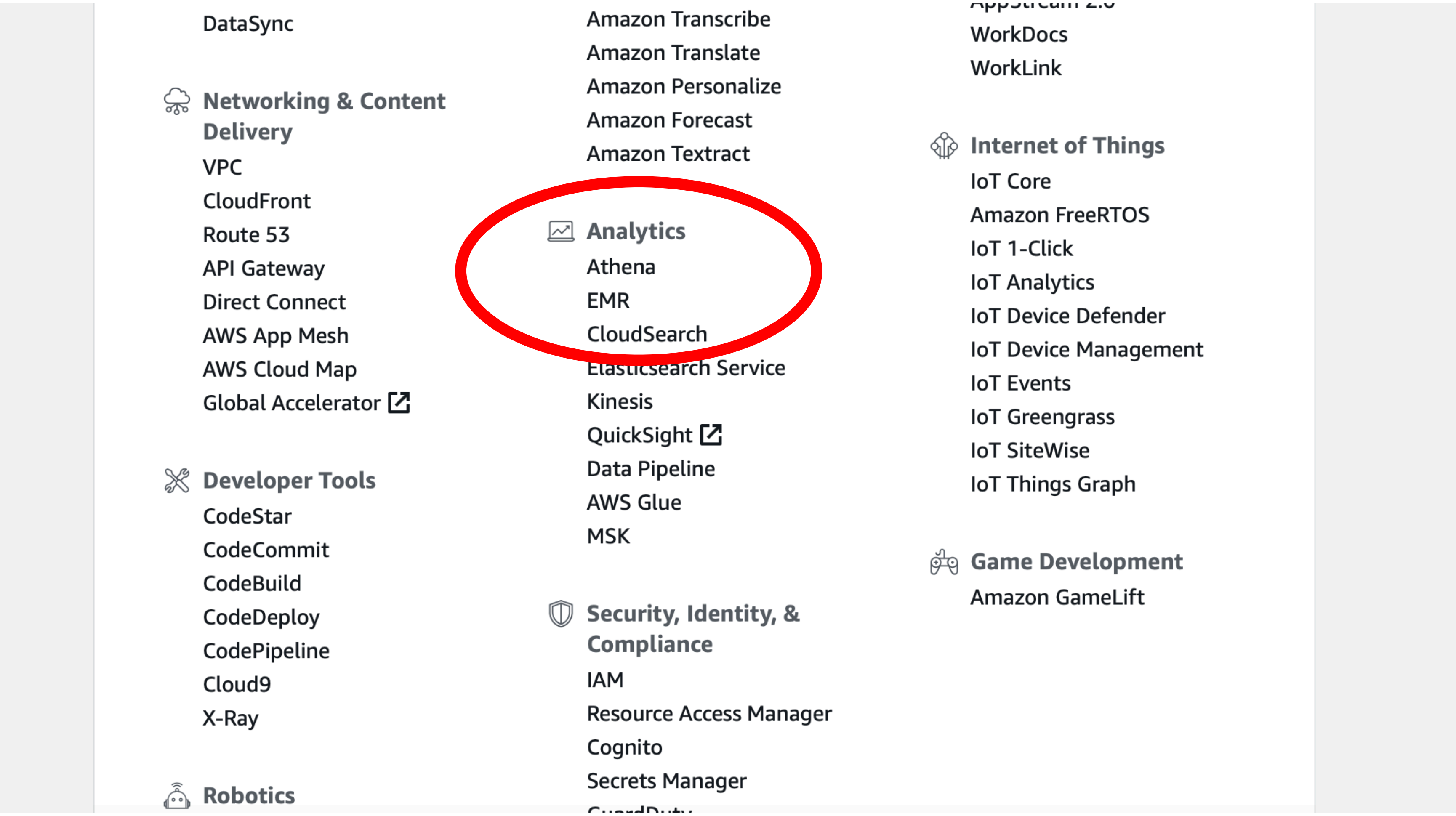
Project

Data at Scale

Create a Cluster

- In AWS Management Console:
  - Select EMR under Analytics



DataSync

Networking & Content Delivery
VPC
CloudFront
Route 53
API Gateway
Direct Connect
AWS App Mesh
AWS Cloud Map
Global Accelerator

Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
Cloud9
X-Ray

Robotics

Amazon Transcribe
Amazon Translate
Amazon Personalize
Amazon Forecast
Amazon Textract

Analytics
Athena
EMR
CloudSearch
Elasticsearch Service
Kinesis
QuickSight
Data Pipeline
AWS Glue
MSK

Security, Identity, & Compliance
IAM
Resource Access Manager
Cognito
Secrets Manager
GuardDuty

AppStream 2.0
WorkDocs
WorkLink

Internet of Things
IoT Core
Amazon FreeRTOS
IoT 1-Click
IoT Analytics
IoT Device Defender
IoT Device Management
IoT Events
IoT Greengrass
IoT SiteWise
IoT Things Graph

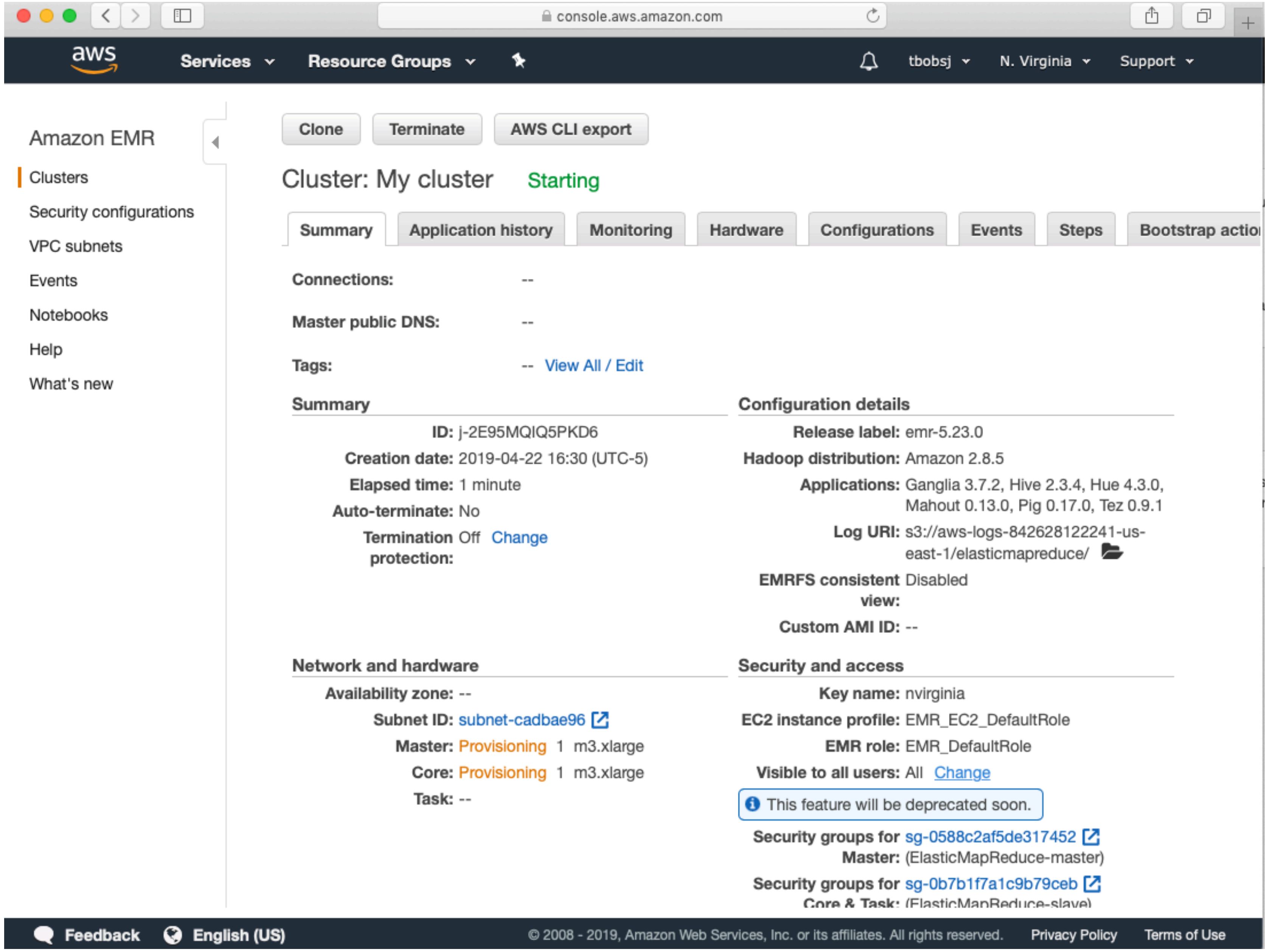
Game Development
Amazon GameLift

Creating a Cluster

- To avoid costs for moving data, use the Northern Virginia data center

Creating a Cluster

- Go to AWS



Cluster: My cluster Starting

Connections: --

Master public DNS: --

Tags: -- View All / Edit

Summary

ID: j-2E95MQIQ5PKD6
Creation date: 2019-04-22 16:30 (UTC-5)
Elapsed time: 1 minute
Auto-terminate: No
Termination protection: Off Change

Configuration details

Release label: emr-5.23.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.1
Log URI: s3://aws-logs-842628122241-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

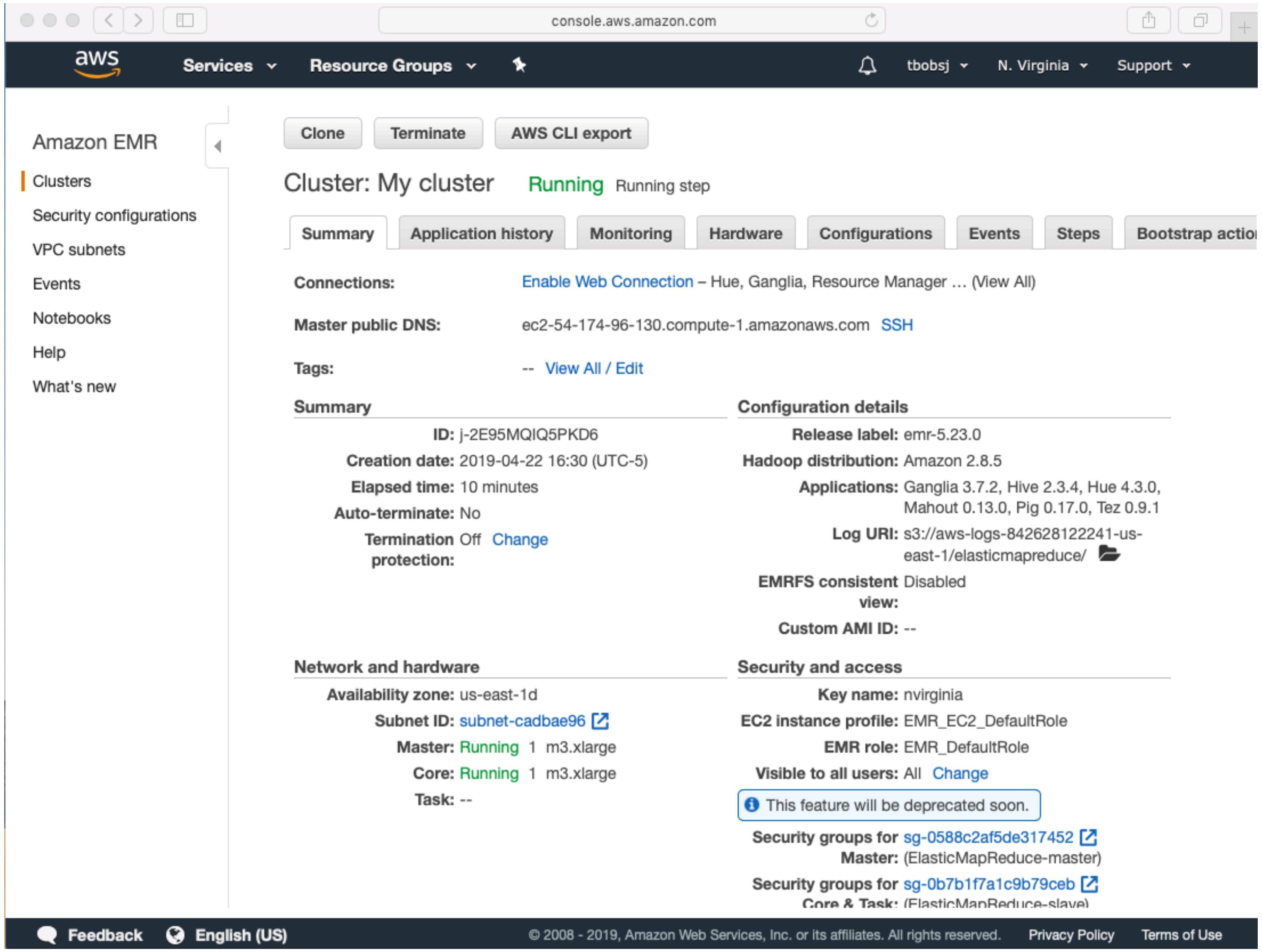
Availability zone: --
Subnet ID: subnet-cadbae96
Master: Provisioning 1 m3.xlarge
Core: Provisioning 1 m3.xlarge
Task: --

Security and access

Key name: nvirginia
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
This feature will be deprecated soon.
Security groups for Master: sg-0588c2af5de317452 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0b7b1f7a1c9b79ceb (ElasticMapReduce-slave)

Creating a Cluster

- After five (!) minutes



Cluster: My cluster Running Running step

Connections: Enable Web Connection – Hue, Ganglia, Resource Manager ... (View All)

Master public DNS: ec2-54-174-96-130.compute-1.amazonaws.com SSH

Tags: -- View All / Edit

Summary

ID: j-2E95MQIQ5PKD6
Creation date: 2019-04-22 16:30 (UTC-5)
Elapsed time: 10 minutes
Auto-terminate: No
Termination protection: Off Change

Configuration details

Release label: emr-5.23.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.1
Log URI: s3://aws-logs-842628122241-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

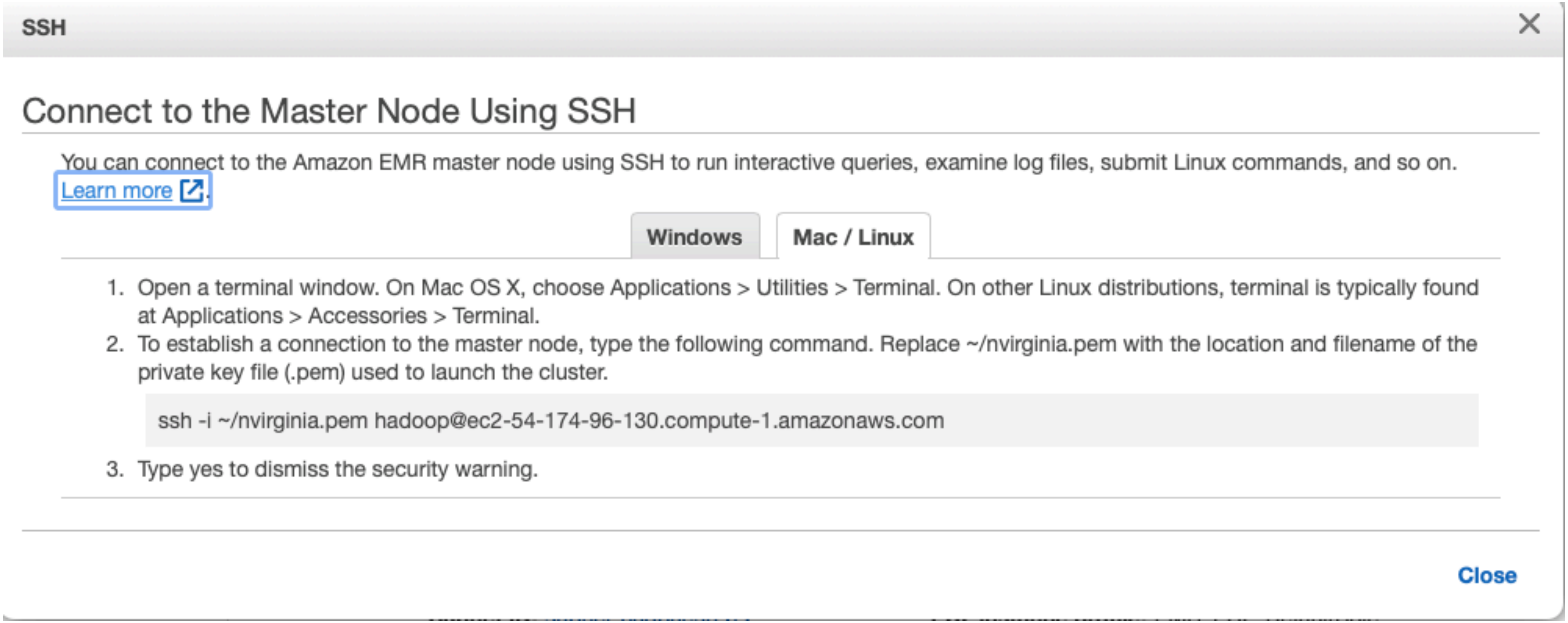
Availability zone: us-east-1d
Subnet ID: subnet-cadbae96
Master: Running 1 m3.xlarge
Core: Running 1 m3.xlarge
Task: --

Security and access

Key name: nvirginia
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
This feature will be deprecated soon.
Security groups for Master: sg-0588c2af5de317452 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0b7b1f7a1c9b79ceb (ElasticMapReduce-slave)

Creating a Cluster

- Click on the SSH link



Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more.

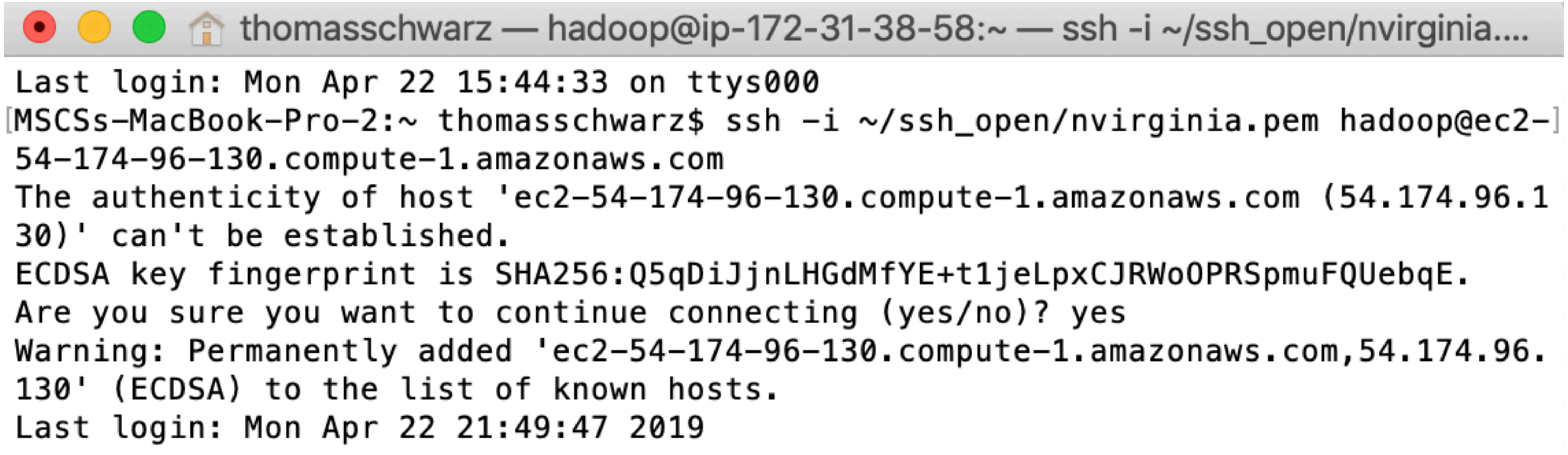
1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/nvirginia.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/nvirginia.pem hadoop@ec2-54-174-96-130.compute-1.amazonaws.com
```

3. Type yes to dismiss the security warning.

Connecting to the Cluster

- Enter the credential in your Command Window / Terminal
  - Be sure to specify the path to the private key that you used to obtain the cluster



thomasschwarz — hadoop@ip-172-31-38-58:~ — ssh -i ~/ssh_open/nvirginia....

```
Last login: Mon Apr 22 15:44:33 on ttys000
[MSCSs-MacBook-Pro-2:~ thomasschwarz$ ssh -i ~/ssh_open/nvirginia.pem hadoop@ec2-]
54-174-96-130.compute-1.amazonaws.com
The authenticity of host 'ec2-54-174-96-130.compute-1.amazonaws.com (54.174.96.1
30)' can't be established.
ECDSA key fingerprint is SHA256:Q5qDiJjnLHGdMfYE+t1jeLpxCJRWoOPRSpmuFQUebqE.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-54-174-96-130.compute-1.amazonaws.com,54.174.96.
130' (ECDSA) to the list of known hosts.
Last login: Mon Apr 22 21:49:47 2019
```

Getting the data

- Go to Kaggle
  - Get the data sets Black Friday and Gun-violence
  - download them to a known location
  - sftp into your cluster
  - upload the files to a directory of your chosing

```
MSCSs-MacBook-Pro-2:BlackFriday thomasschwarz$ sftp -i ../../ssh_open/nvirginia.pem hadoop@ec2-54-162-49-218.compute-1.amazonaws.com
Connected to hadoop@ec2-54-162-49-218.compute-1.amazonaws.com.
sftp> ls
data
sftp> lls
BlackFriday.csv
sftp> put BlackFriday.csv
Uploading BlackFriday.csv to /home/hadoop/BlackFriday.csv
BlackFriday.csv                                                                                    100%   24MB  23.6MB/s   00:01
sftp> lls
BlackFriday.csv               gun-violence-data_01-2013_03-2018.csv
sftp> ls
BlackFriday.csv    data
```

SFTP and FTP commands

Getting the data

- In the master, organize your data
  - Manually delete the first row with the description of the fields

Starting Pig

- Start pig

```
pig –x mapreduce
```

- Register piggybank

```
register file:/usr/lib/pig/lib/piggybank.jar
```

- Move data into Hadoop

```
grunt> sh ls data
BlackFriday.csv
gun-violence-data_01-2013_03-2018.csv
grunt> copyFromLocal ./data/BlackFriday.csv BlackFriday.csv
grunt> ls
hdfs://ip-172-31-40-153.ec2.internal:8020/user/hadoop/
BlackFriday.csv<r 1>  24956107
```

Load data into Pig

```
blackfriday = LOAD 'blackfriday' USING PigStorage(',')
AS (UserID: int, ProductID: chararray, Gender:
chararray, Age: chararray, Occupation: int,
CityCategory: chararray, Stay_In_Current_City: int,
Marital_Status: int, Product_Category_1: chararray,
Product_Category_2: chararray, Product_Category_3:
chararray, Purchase: int);

illustrate blackfriday;

simplified = FOREACH blackfriday GENERATE UserID,
Gender, Age, Purchase;
```

Getting the data

- Group by age and gender

```
gender_age = GROUP simplified by (Gender, Age);
```

- Calculate maximum and average purchase amounts per age-gender group

```
my_sum = FOREACH gender_age GENERATE group as
age_count, COUNT(simplified),
MAX(simplified.Purchase),
SUM(simplified.Purchase)/COUNT(simplified);
```

Getting the data

- Display the results:

```
dump my_sum;
```

Tasks

- Get the statistics from Black Friday for age, gender, age-gender, age-gender-CityCategory

- Use gun_violence_data to find
  - the city or county in Wisconsin with the highest number of gun incidents
  - the city in the country with the highest number of gun incidents