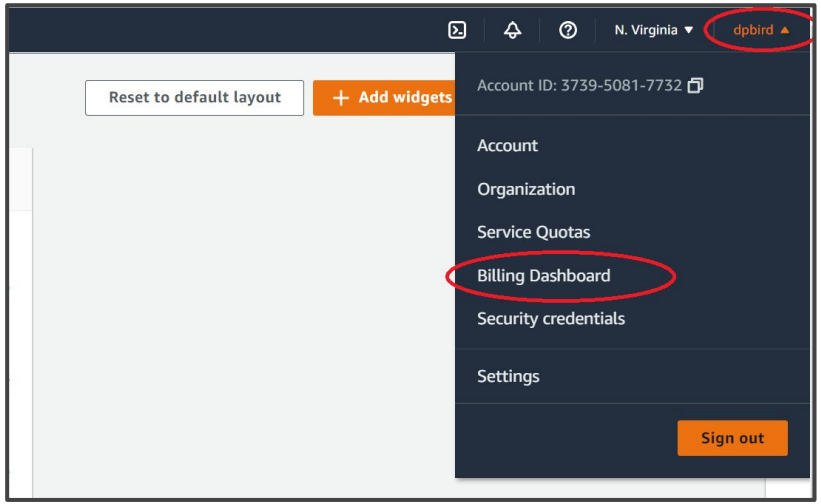
10-605 Homework 4 AWS Setup

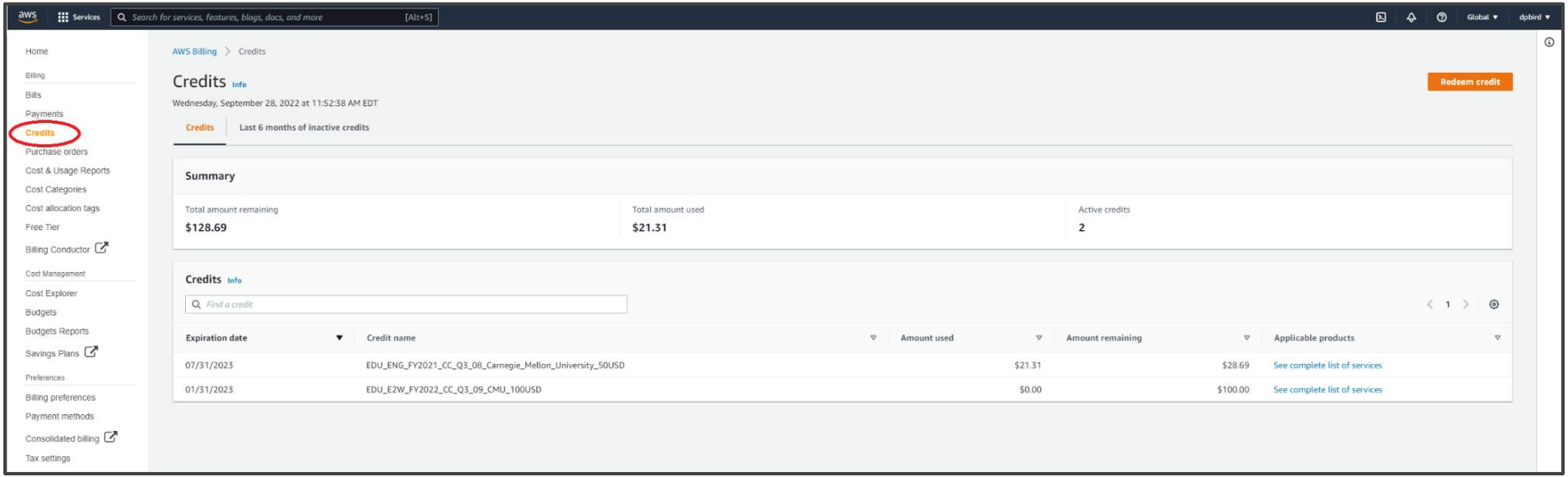
Checking Credits

- All students should have $100 put into their account.
- You should only need ~$30 to complete Homework 4.

To check your credits, from your AWS Main Page:

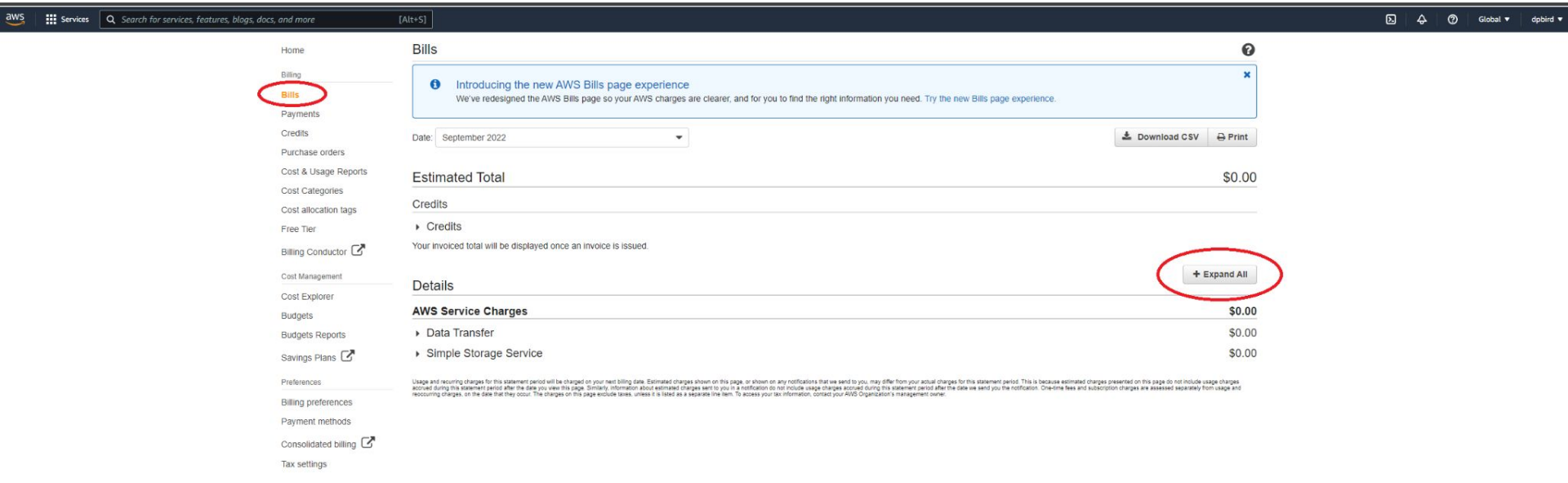


Checking Credits



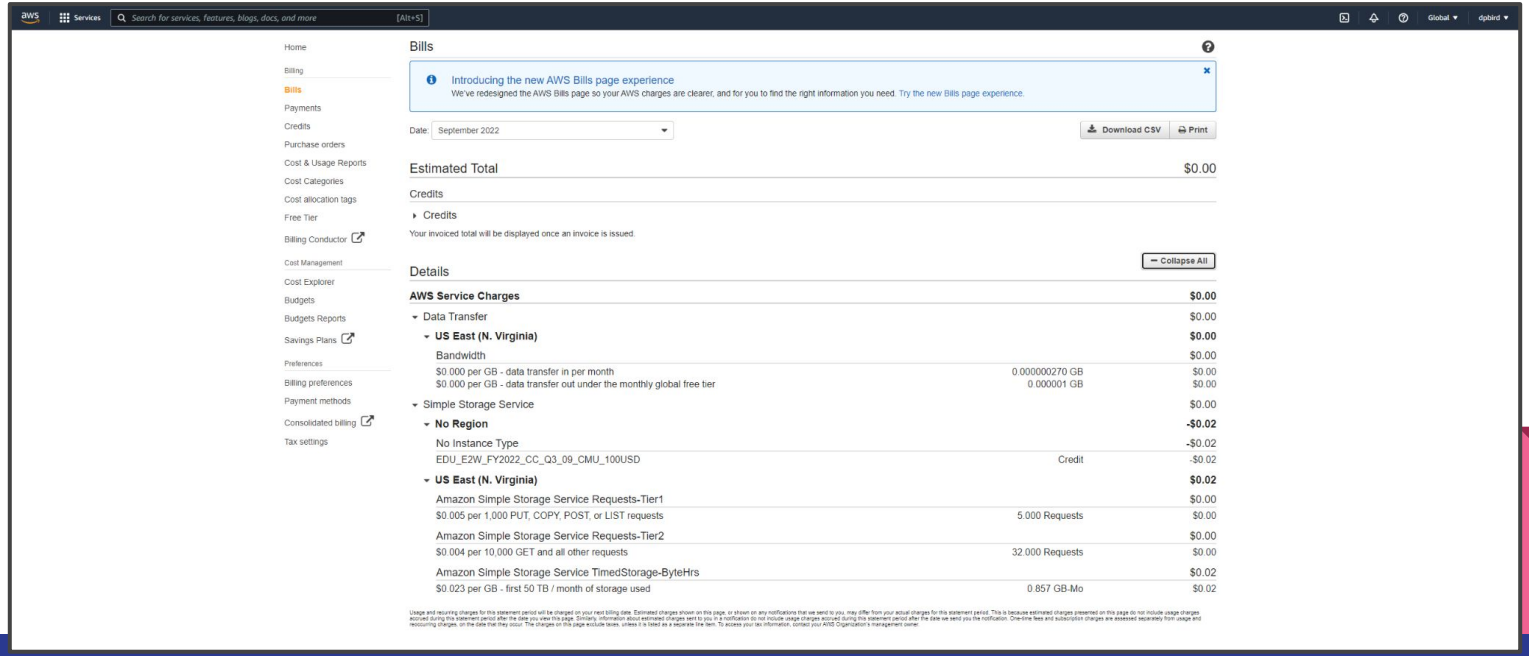
Checking Amount Spent

- Click Bills
  - Then you must click “Expand All” as it hides charges below:



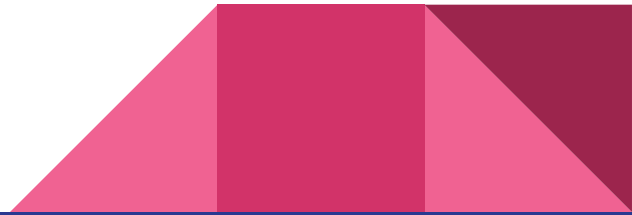
Checking Amount Spent

- Click Bills
  - Then you must click “Expand All” as it hides charges below:

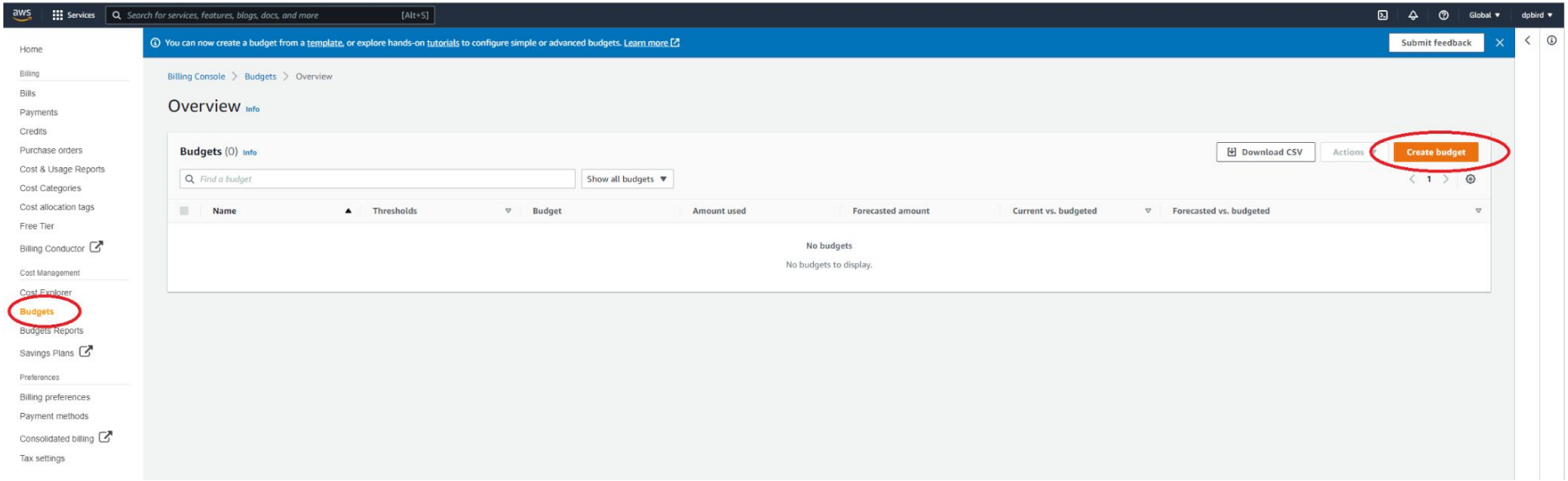


IMPORTANT: Creating a Budget Warning

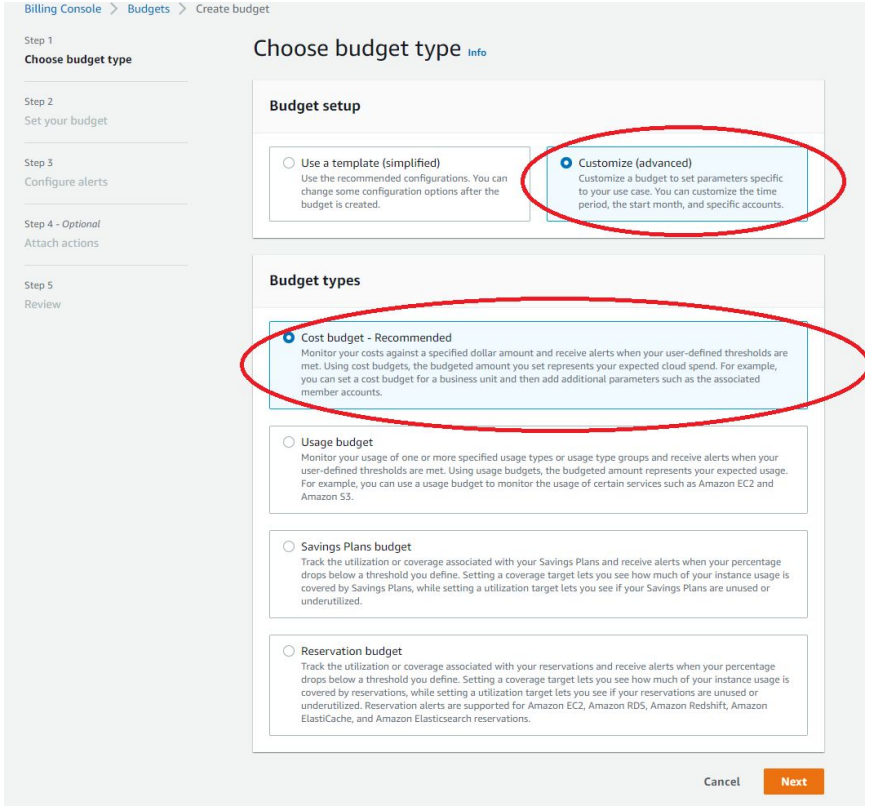
- AWS is linked to your credit card you used to create the account. If you run over your $100 allowance then it will automatically start charging your card.
- Every semester there is one student who leaves an instance or volume open for weeks because they neglected to set up warnings and they are left with a bill of literally **thousands** of dollars. The University is not responsible for this.
- PLEASE: Set up budget warnings and triple check that you have terminated all instances and volumes!



Creating a Budget Warning



Creating a Budget Warning

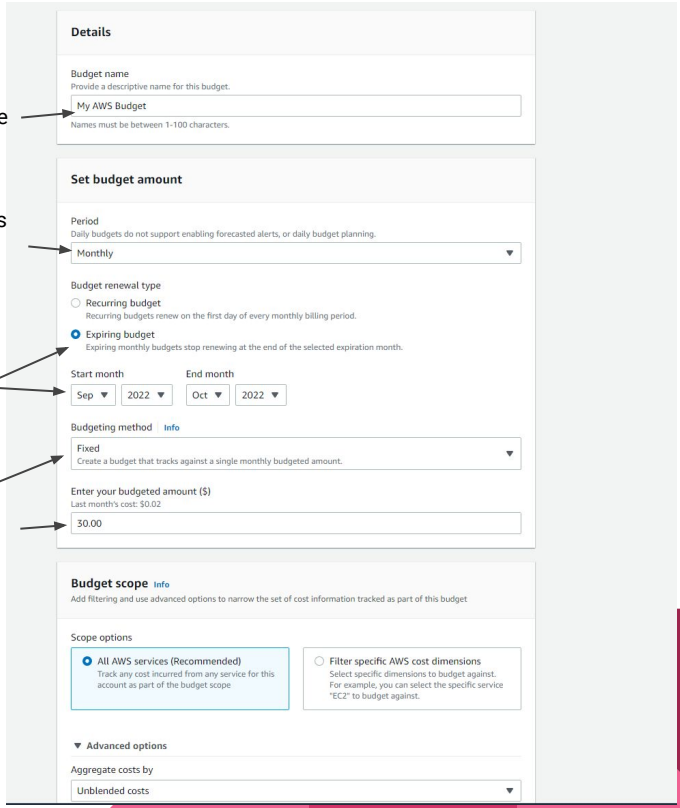


Give your budget a name

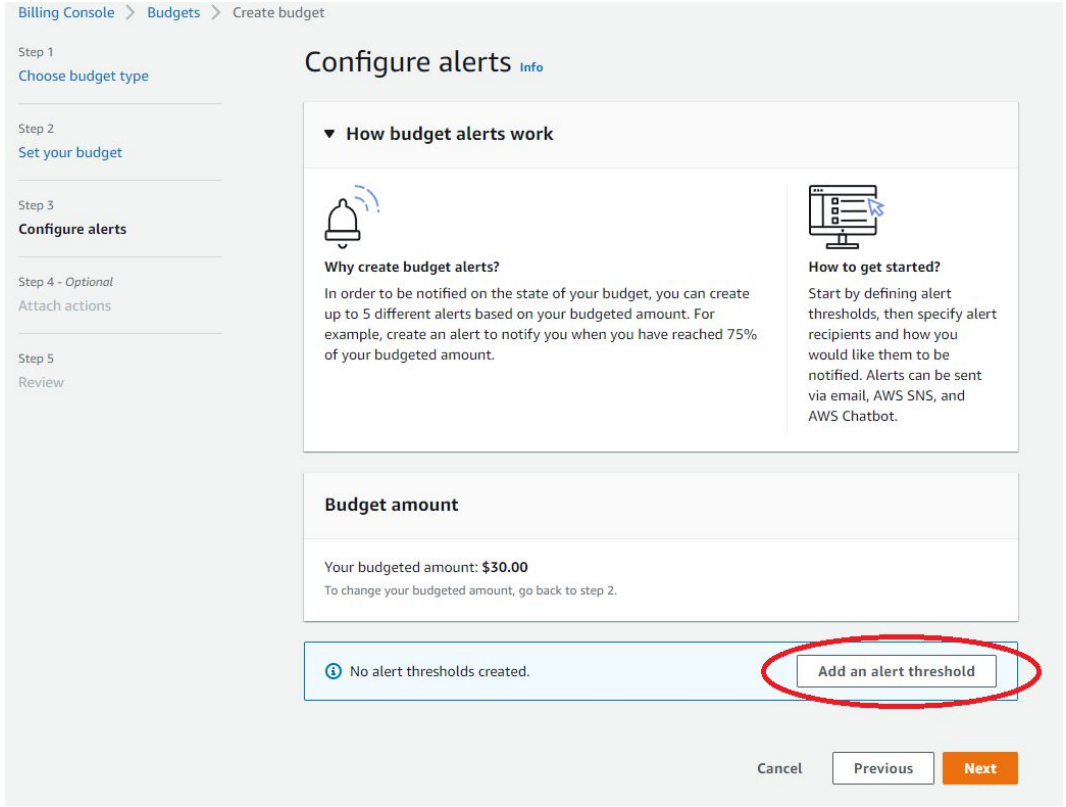
Since the assignment is only over 1 month select “Monthly”

You can choose “Expiring” and select the appropriate start and end month

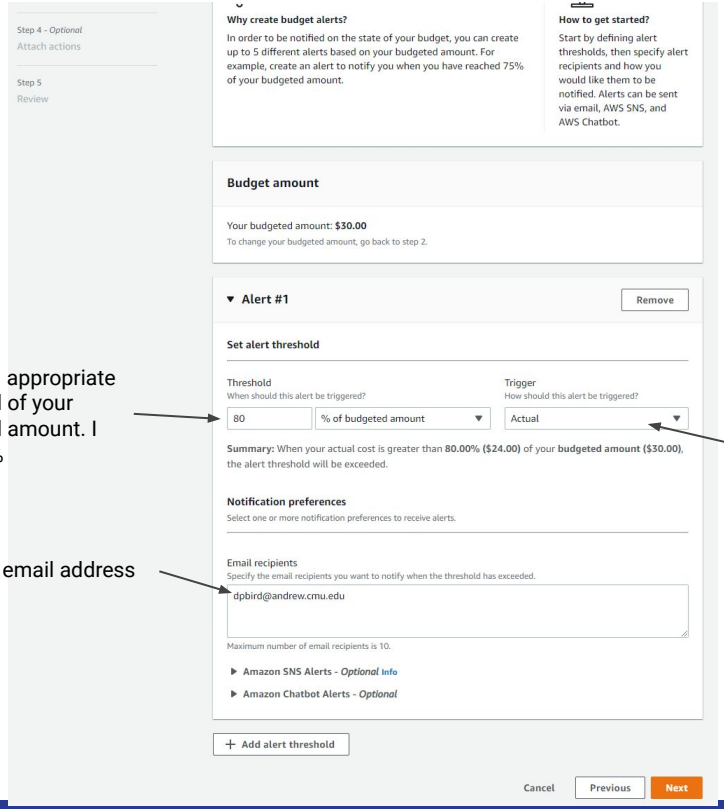
Choose “Fixed” and since the assignment can be completed with $30 I recommend using this as your Budgeted amount.



Creating a Budget Warning



Creating a Budget Warning

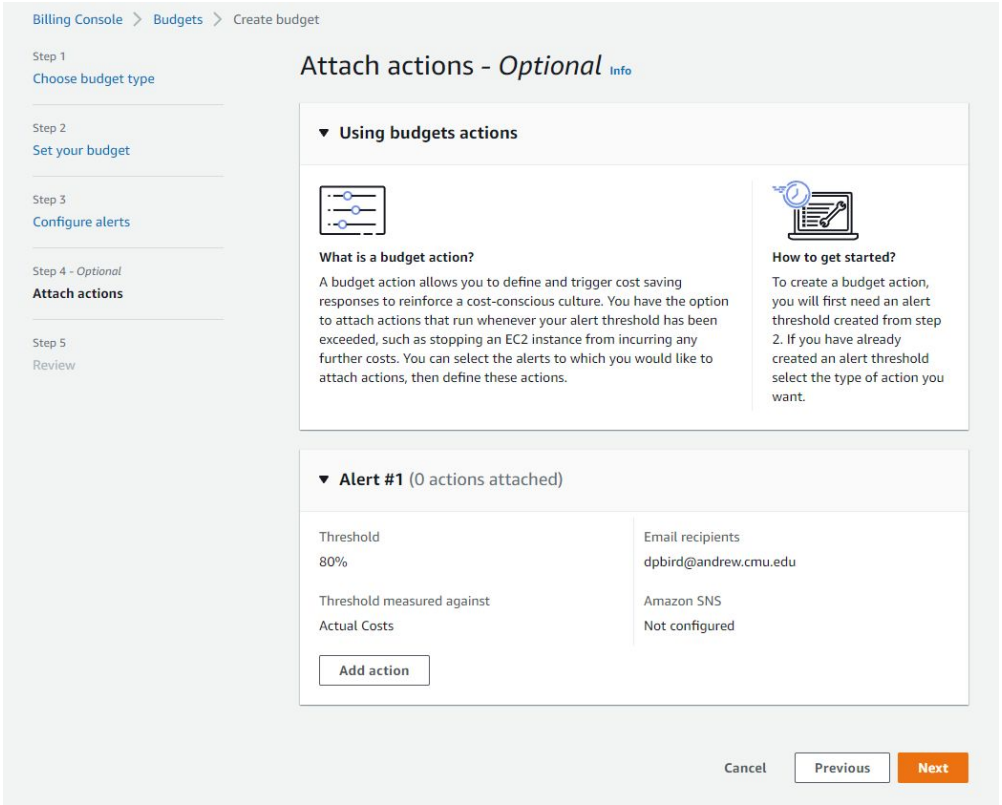


Set up an appropriate threshold of your budgeted amount. I used 80%

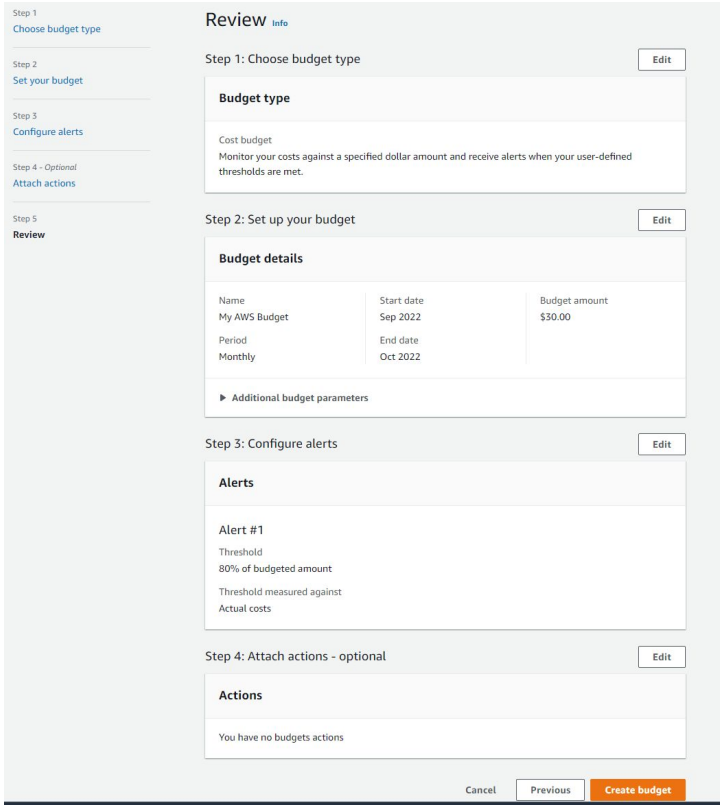
Add your email address

You can change this to “Projected” costs but this might cause AWS to warn you earlier than necessary, I’ll keep it on “Actual”

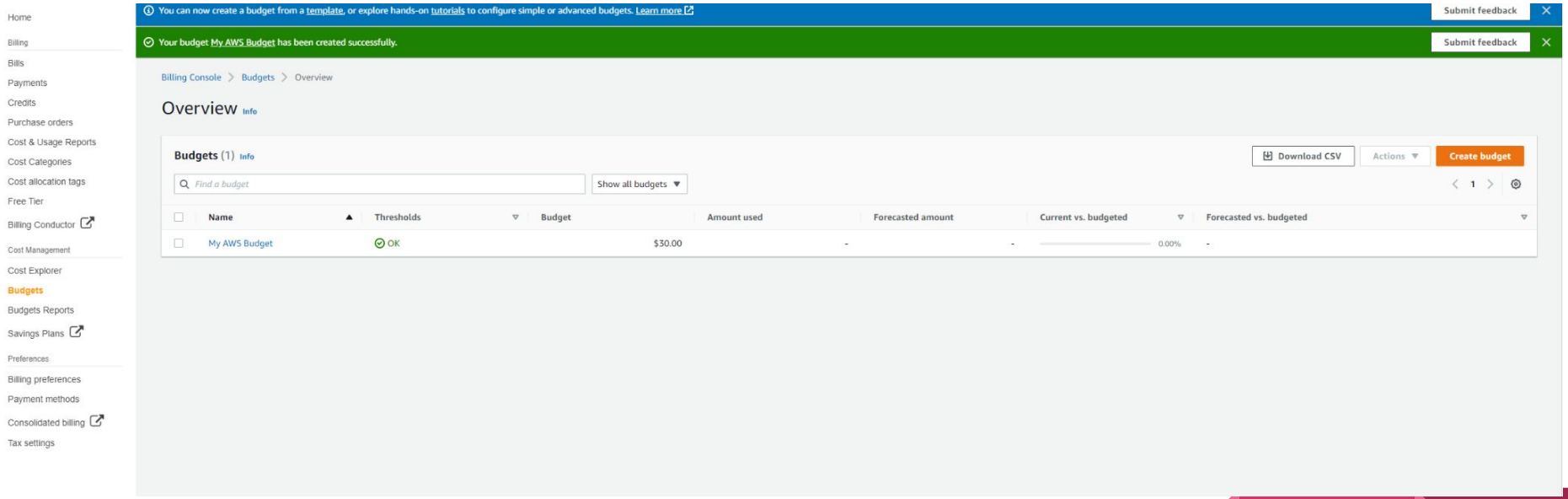
Creating a Budget Warning



Creating a Budget Warning

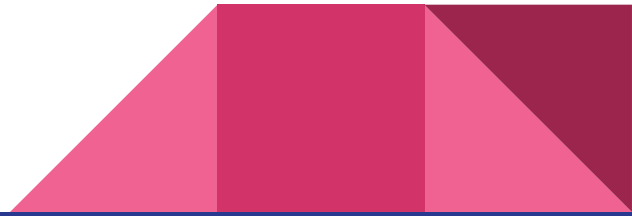


Creating a Budget Warning

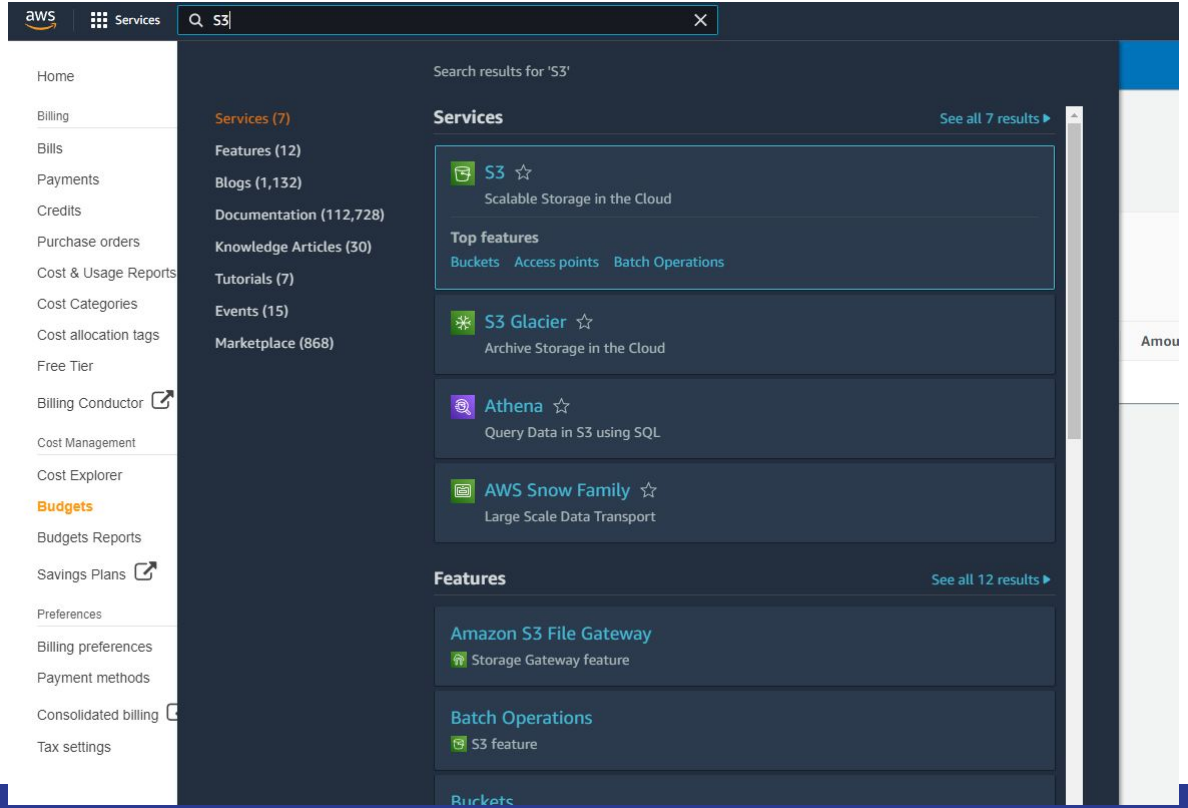


10-605 Setup for Homework Part A

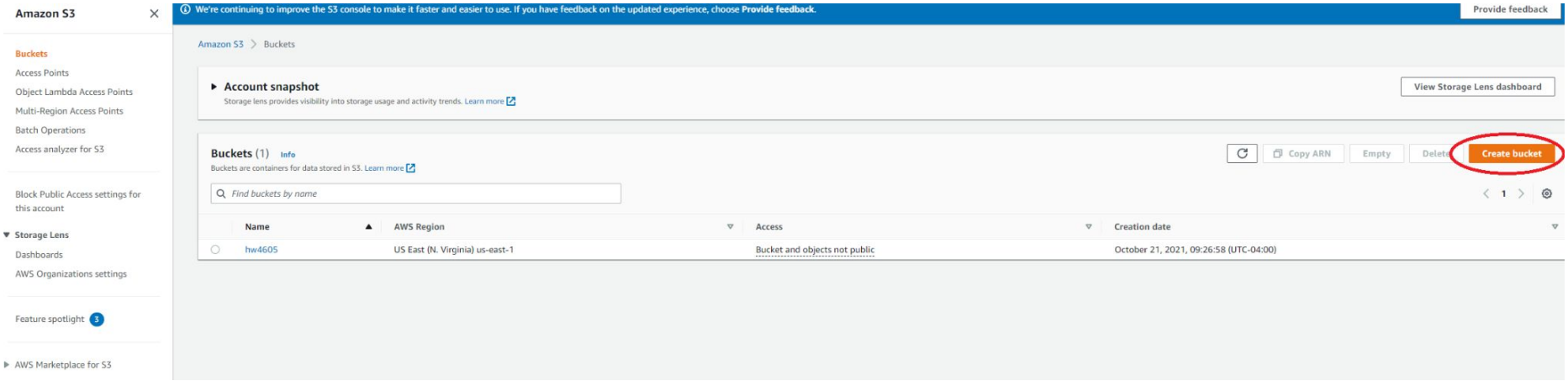
- Setup S3 Bucket
- Setup IAM Roles
- Make a keypair
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*



Setup S3 Bucket

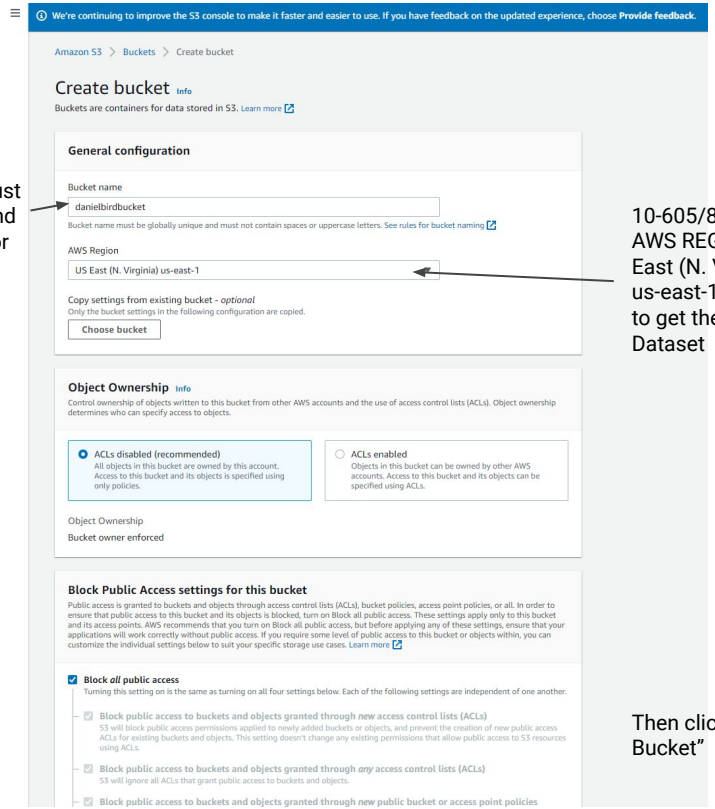


Setup S3 Bucket



Setup S3 Bucket

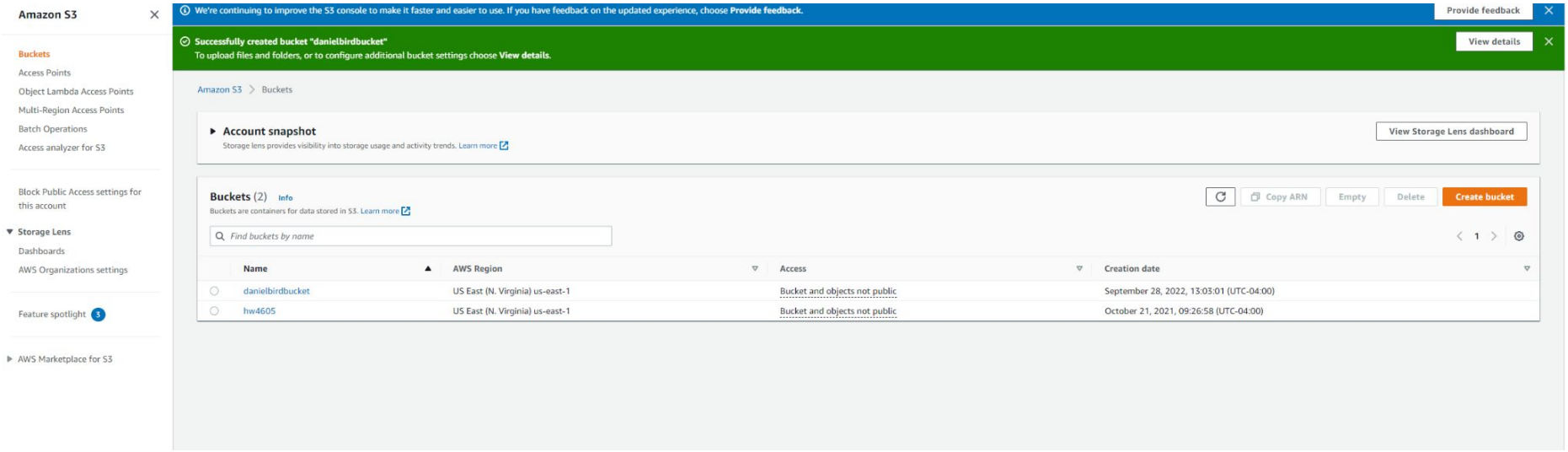
Give your bucket a name. This name must be globally unique and not contain spaces or uppercase letters



10-605/805 **MUST** SET AWS REGION TO “US East (N. Virginia) us-east-1” We need this to get the Million Song Dataset

Then click: “Create Bucket”

Setup S3 Bucket

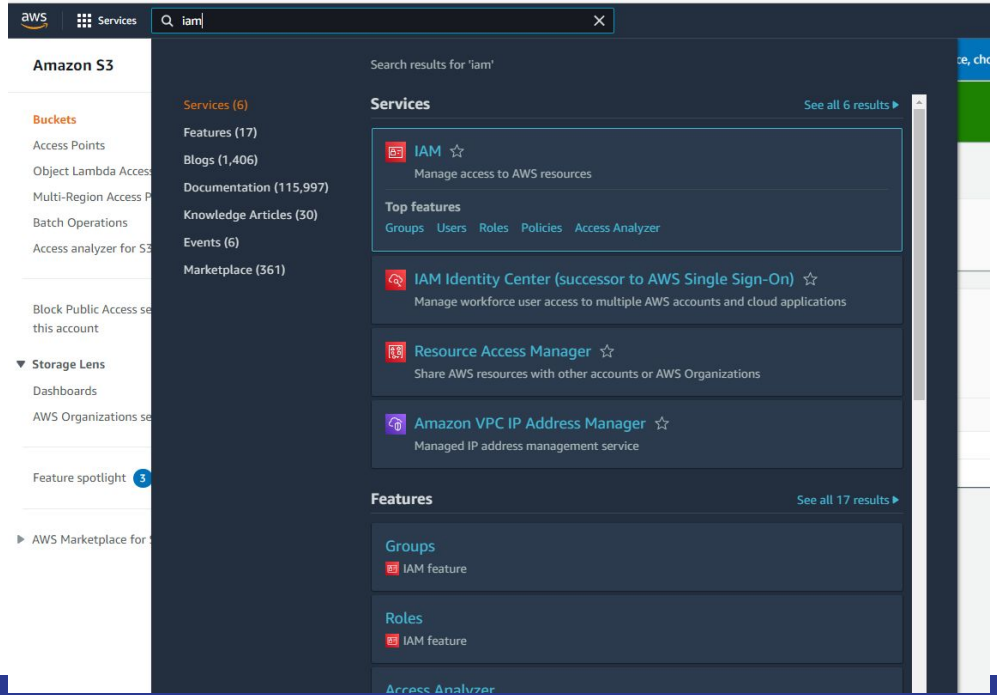


10-605 Setup for Homework Part A

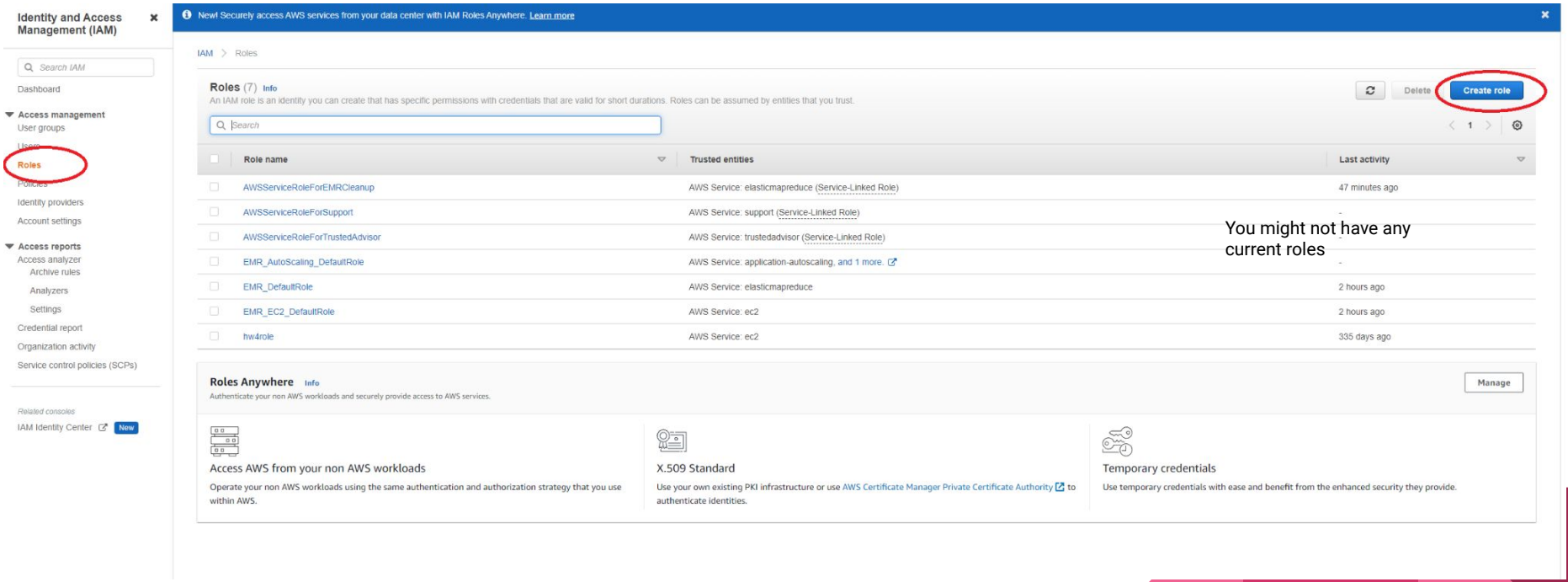
- Setup S3 Bucket - DONE
- Setup IAM Roles
- Make a keypair
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*

Setup IAM Roles

- Identity and Access Management

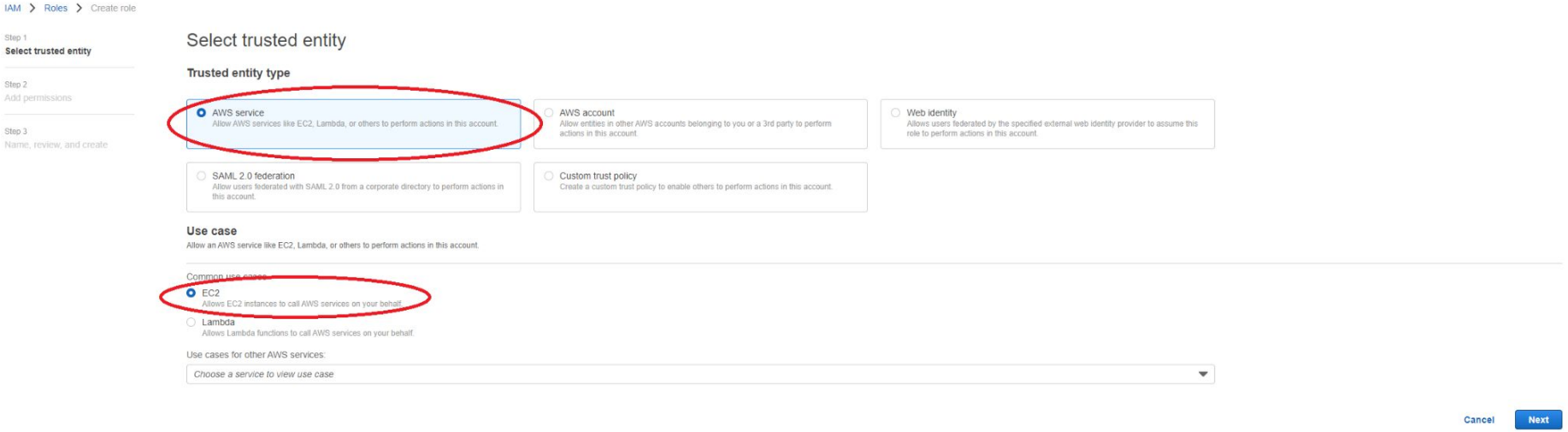


Setup IAM Roles

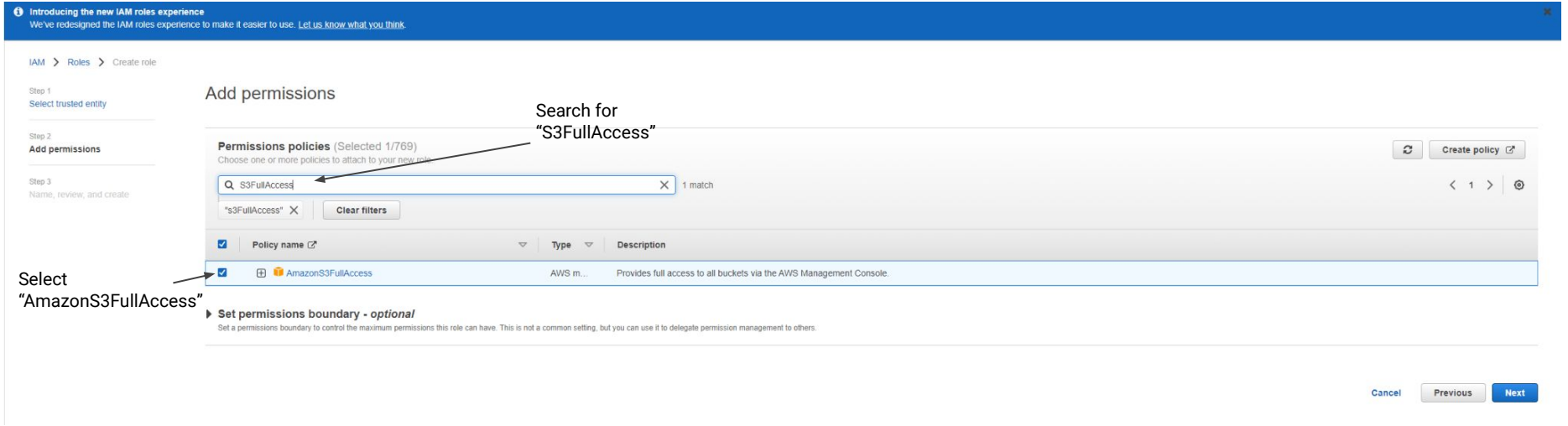


You might not have any current roles

Setup IAM Roles

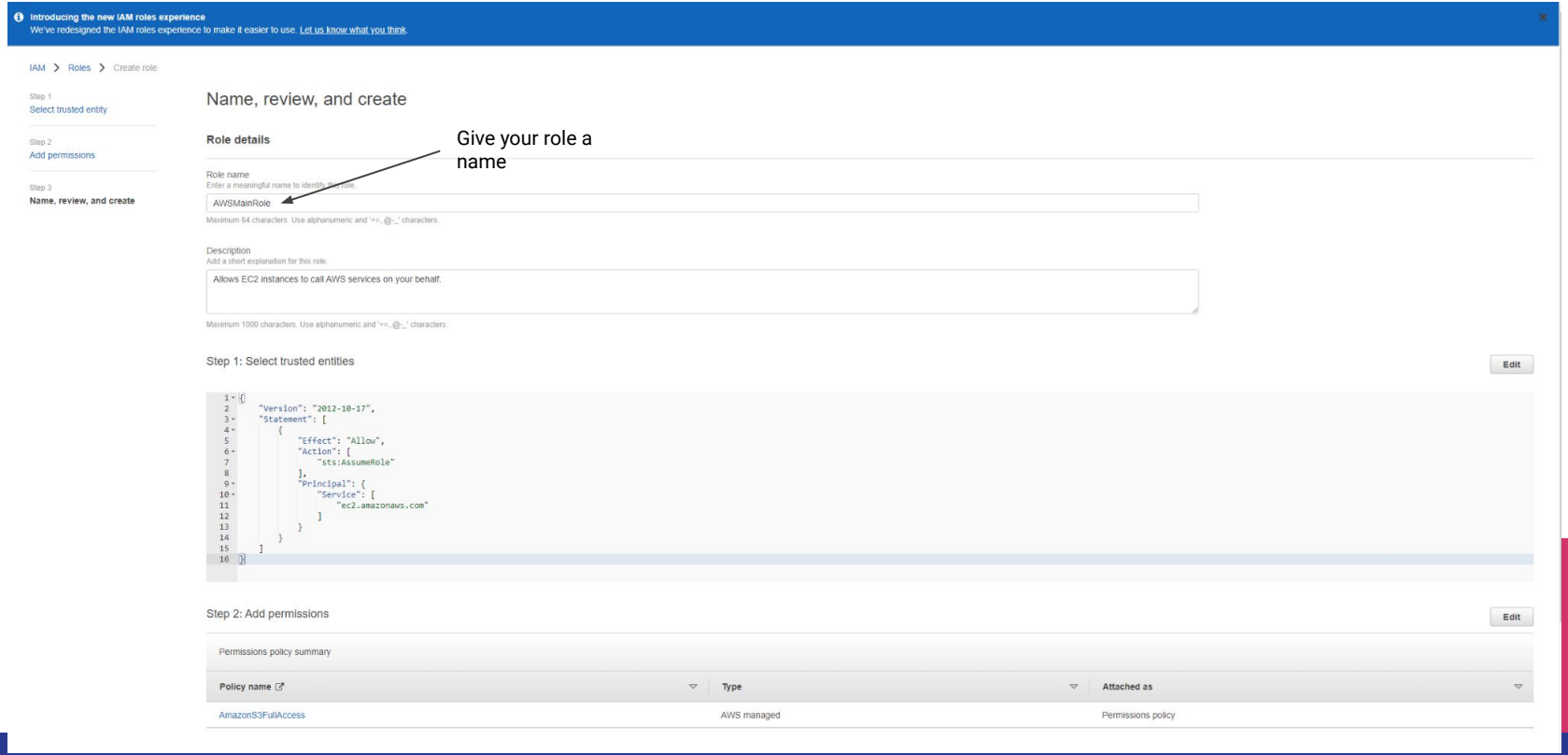


Setup IAM Roles

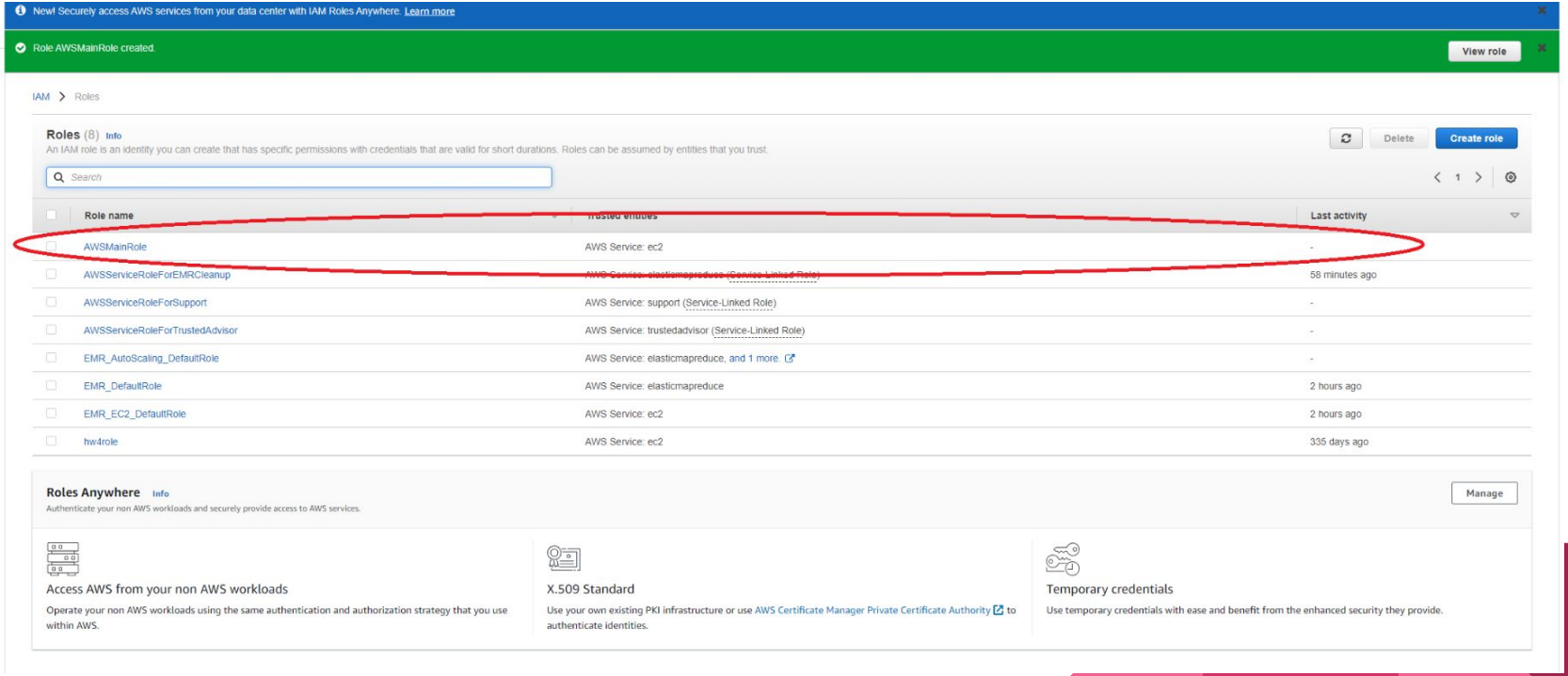


This is allowing EC2 instances to have full access to all S3 Buckets!

Setup IAM Roles

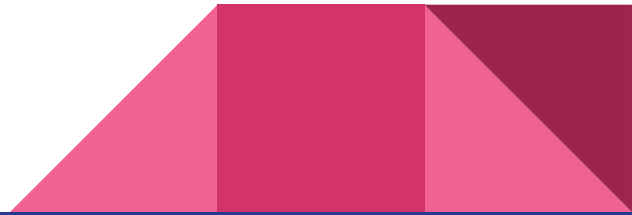


Setup IAM Roles



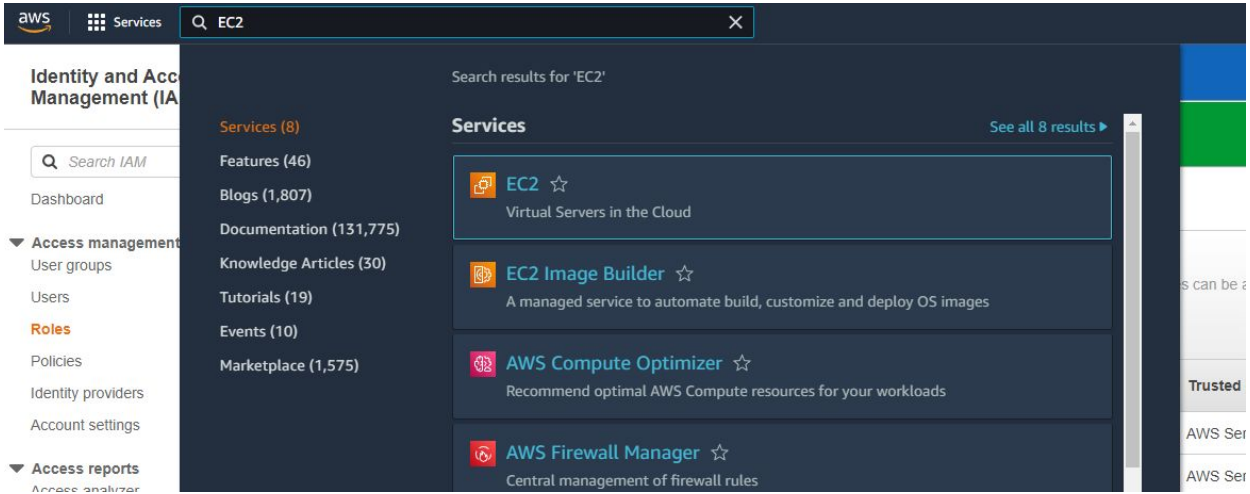
10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*

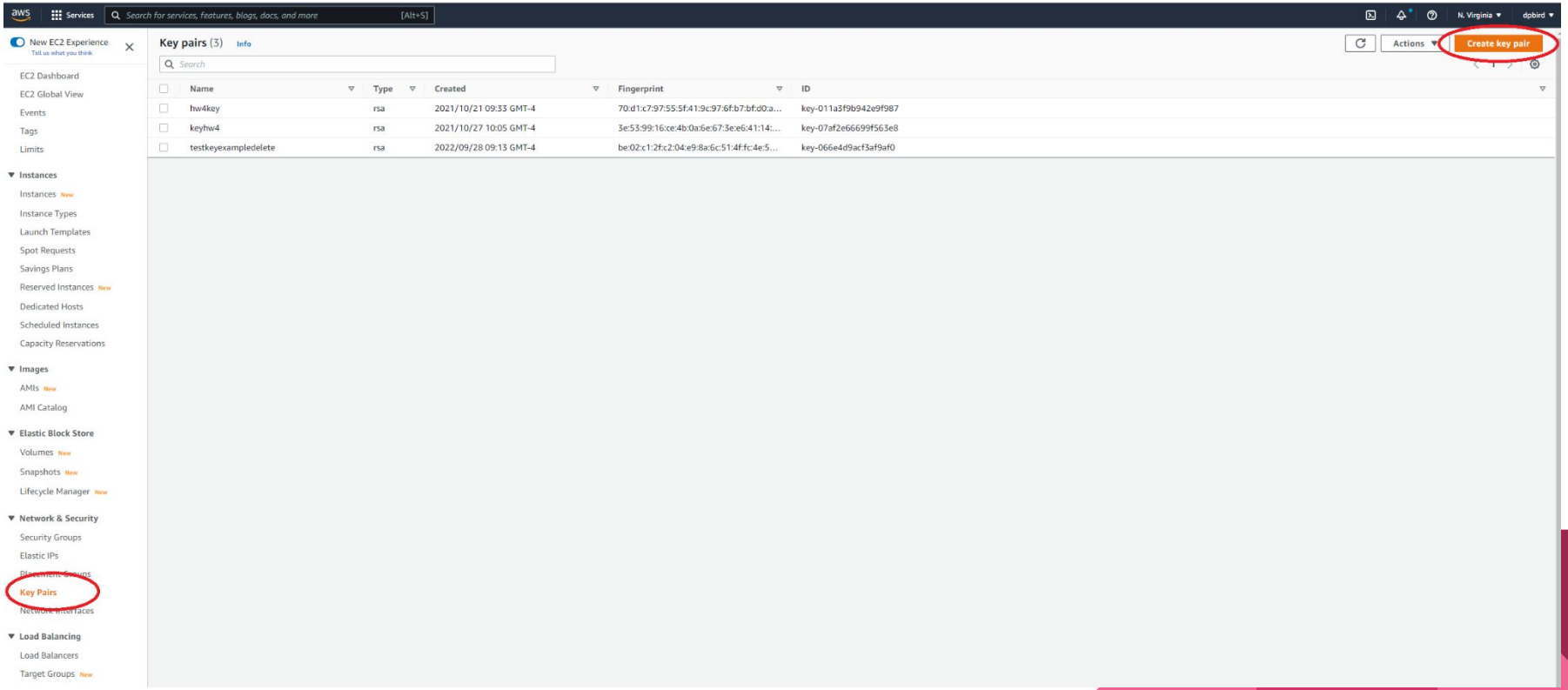


Make a keypair

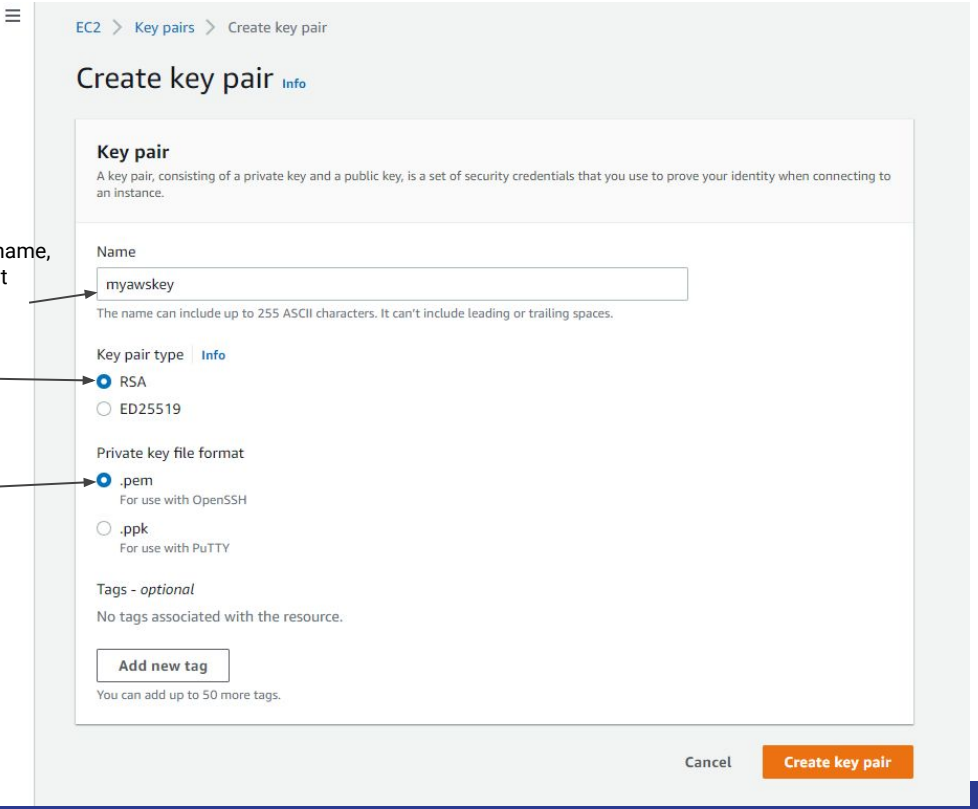
- This is needed to access your EC2 instance and your EMR Cluster.
- It is important that it is private and you know where it is saved.
- First head to EC2



Make a keypair



Make a keypair



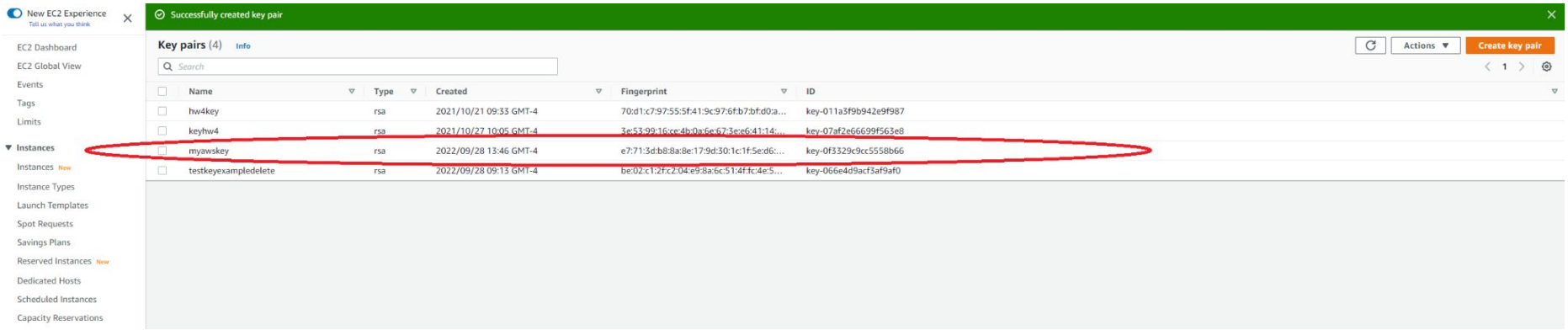
Give your key a name, I recommend not using spaces

Select RSA

Select .pem

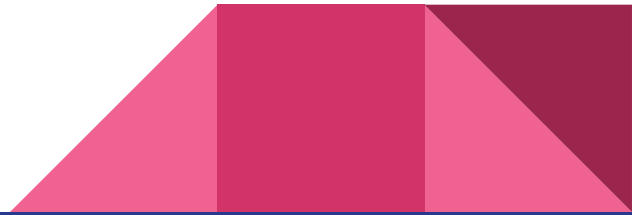
- Once you click “Create Key Pair”, this will download a .pem. Make sure you save this somewhere safe where you can locate it.

Make a keypair



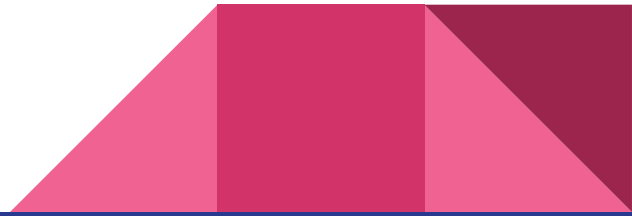
10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE Kind of…
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*



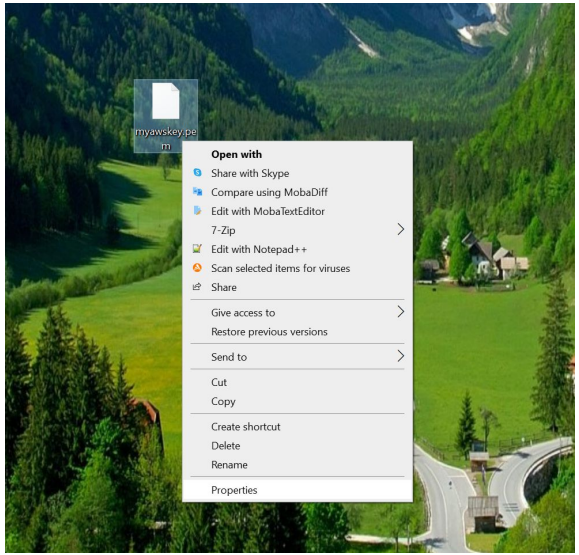
Make a keypair - DONE Kind of…

- Your .pem file is currently “Public” on your computer. When you try to access an instance with your .pem file it won’t allow you until you change the permissions.
- How you do this depends on whether you are using Windows or Mac



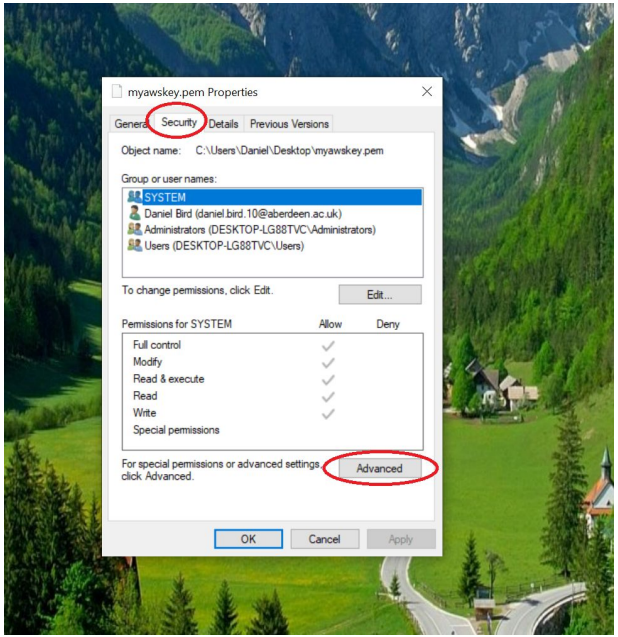
Make a keypair - Windows Instructions

- Locate the file, right click and select “Properties”



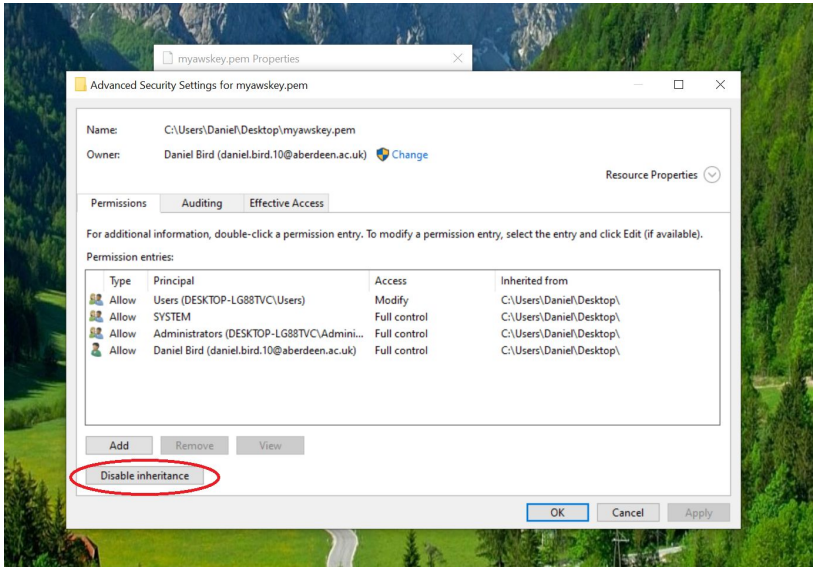
Make a keypair - Windows Instructions

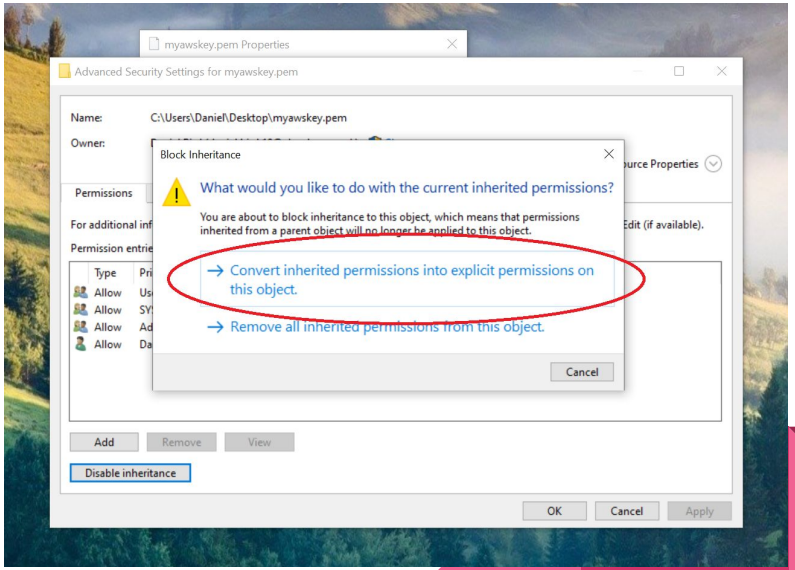
- Click the “Security” tab, then “Advanced”.



Make a keypair - Windows Instructions

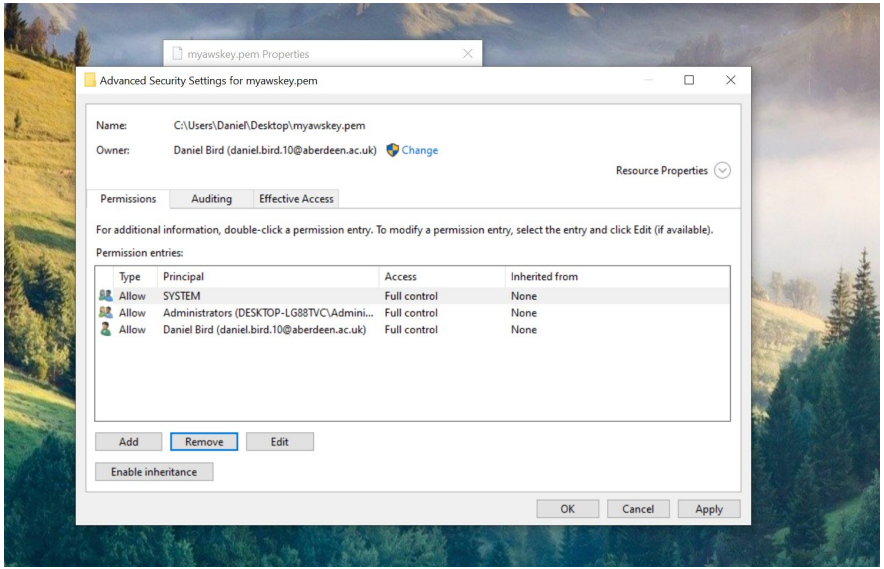
- Click “Disable inheritance”, then “Convert inherited permissions into…”





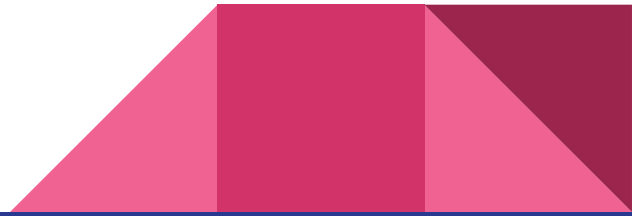
Make a keypair - Windows Instructions

- Remove “Users” by clicking on it and then clicking the Remove button.



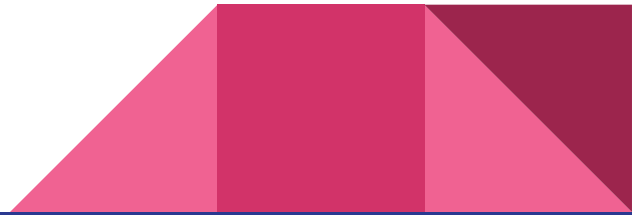
Make a keypair - Mac Instructions

- In the terminal run
  - chmod 400 #KEYPAIRNAME
  - So I would run:
  - chmod 400 myawskey.pem



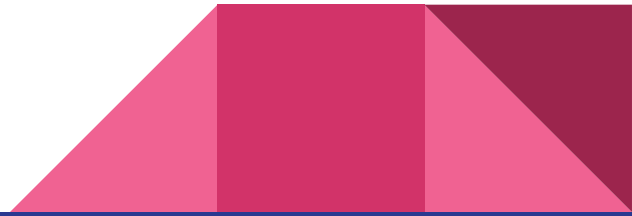
10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*

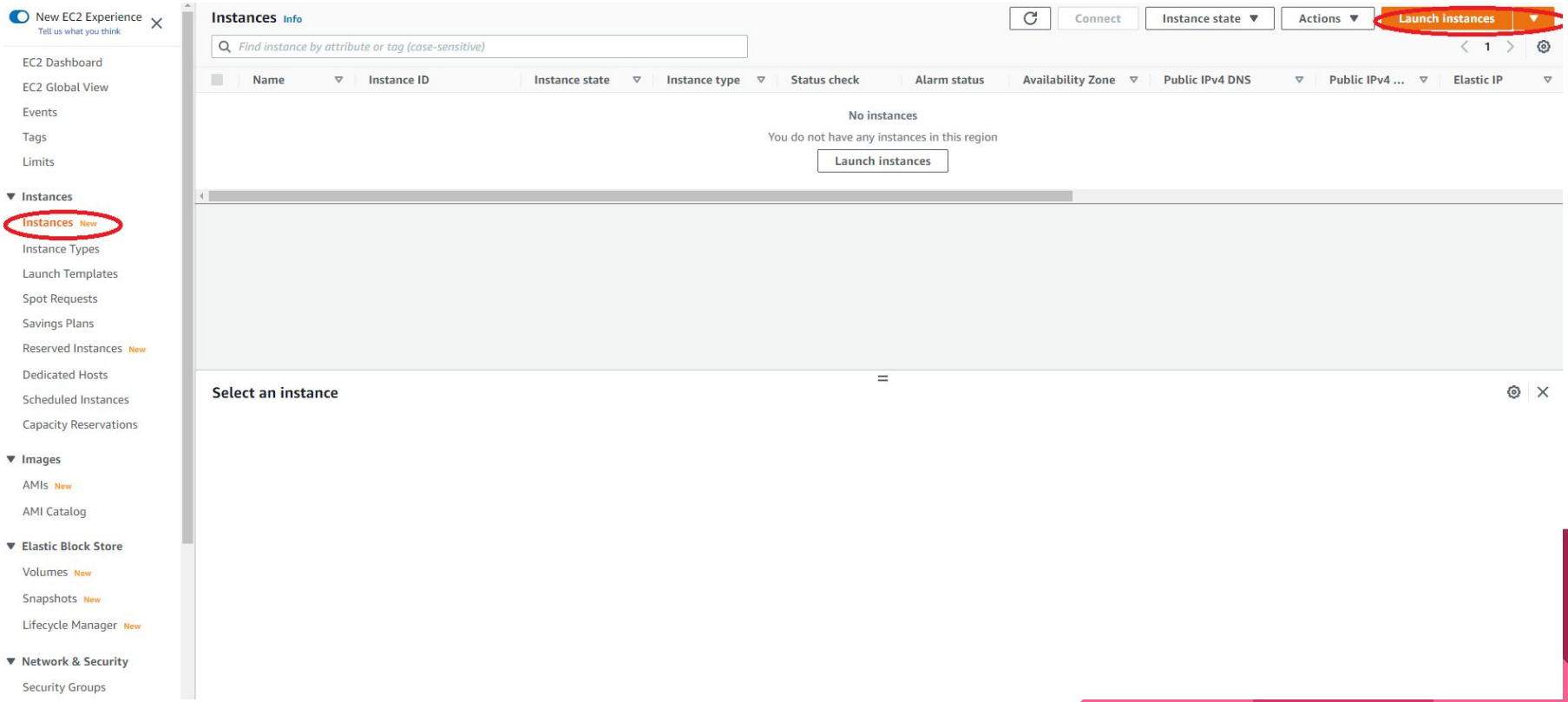


Setup EC2 Instance

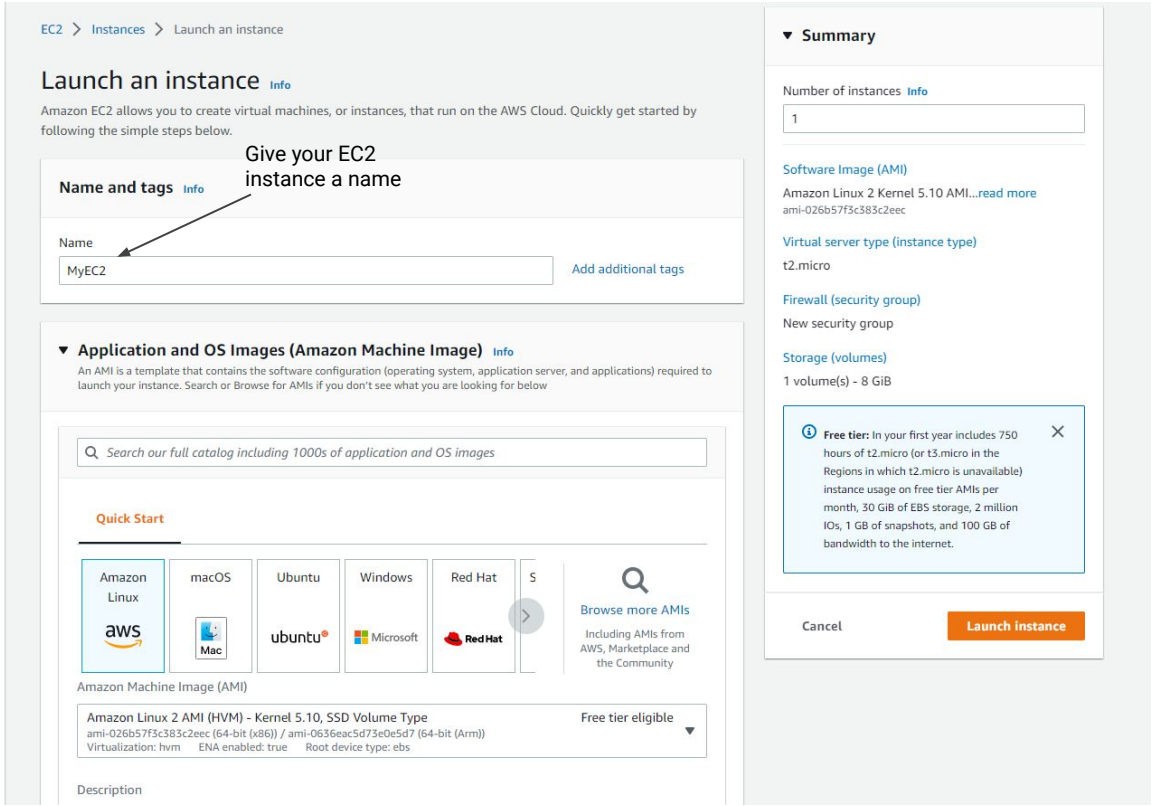
- What is EC2? (Elastic Compute)
- It consists of the ability to:
  - Rent Virtual Machines (EC2)
  - Store data on virtual drives (EBS)
  - Other useful services which have been adopted by many companies who require cloud computing
- You can think of this as the main AWS offering.



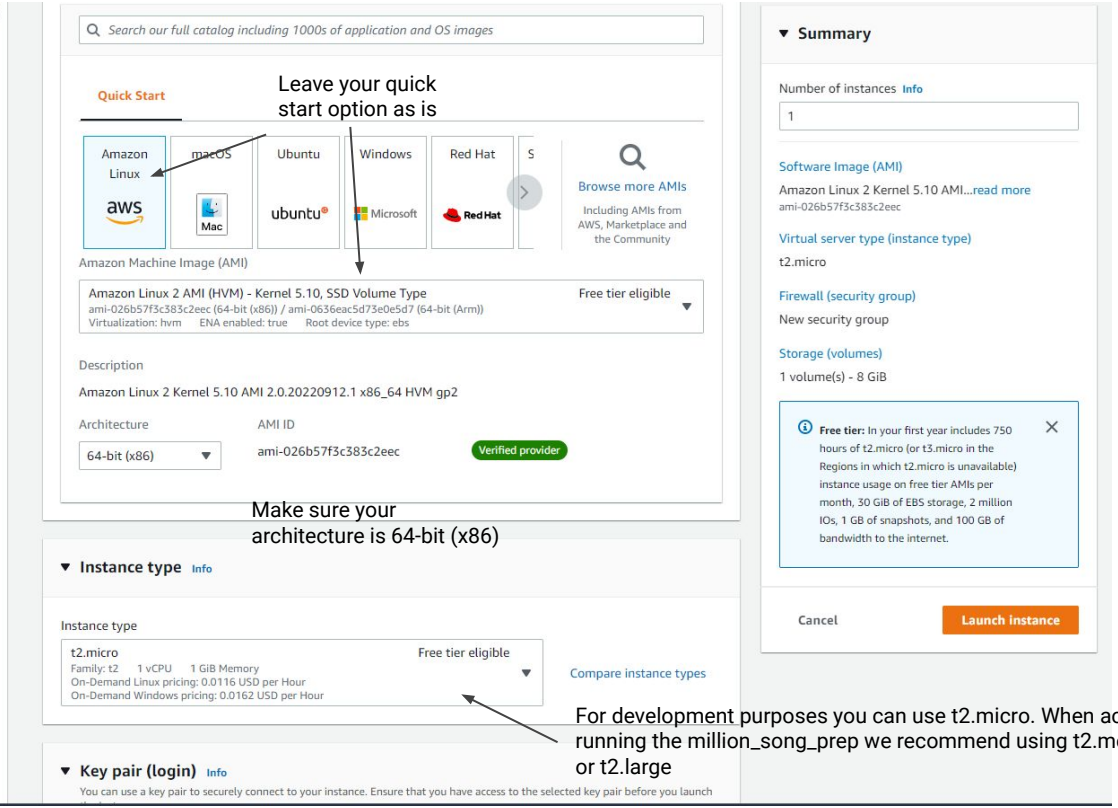
Setup EC2 Instance



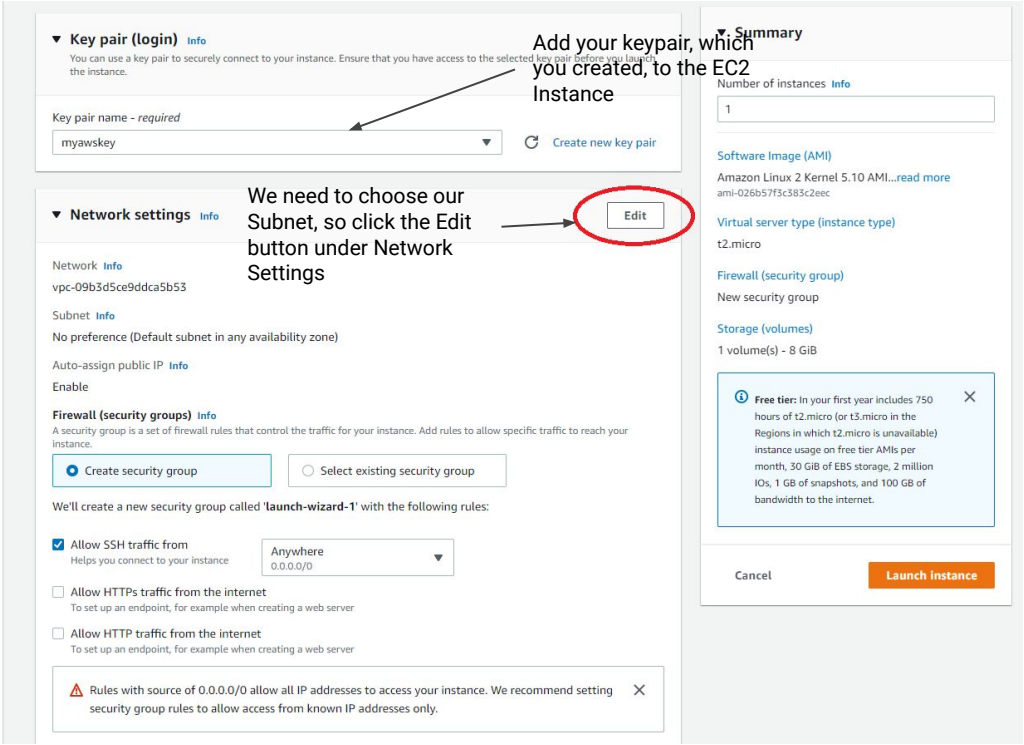
Setup EC2 Instance



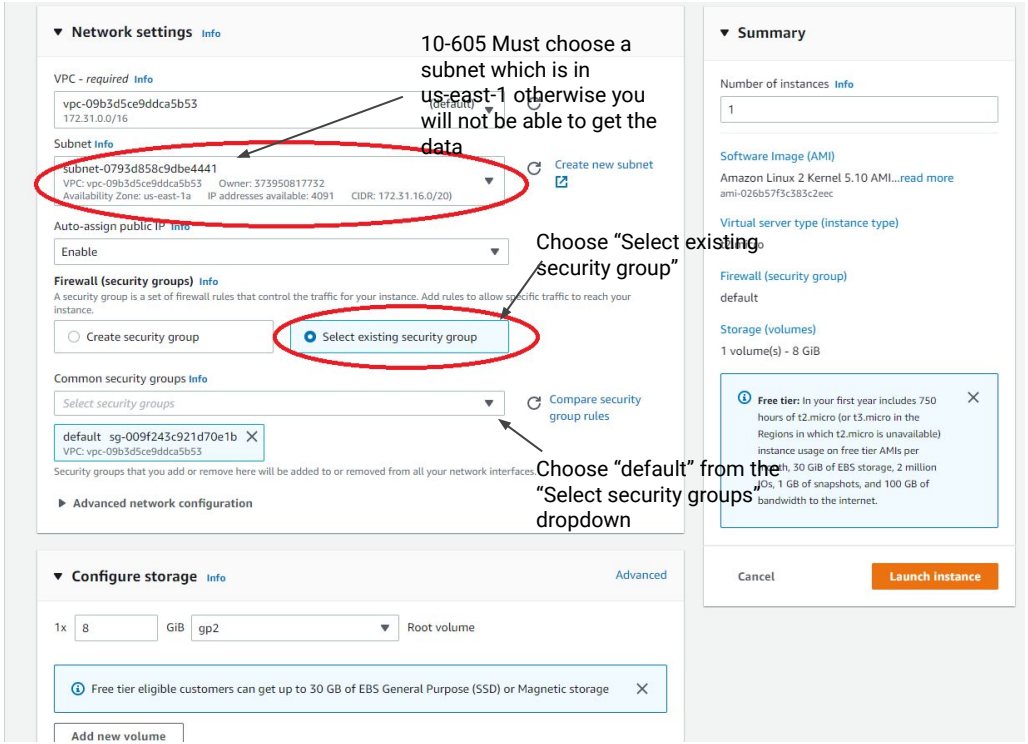
Setup EC2 Instance



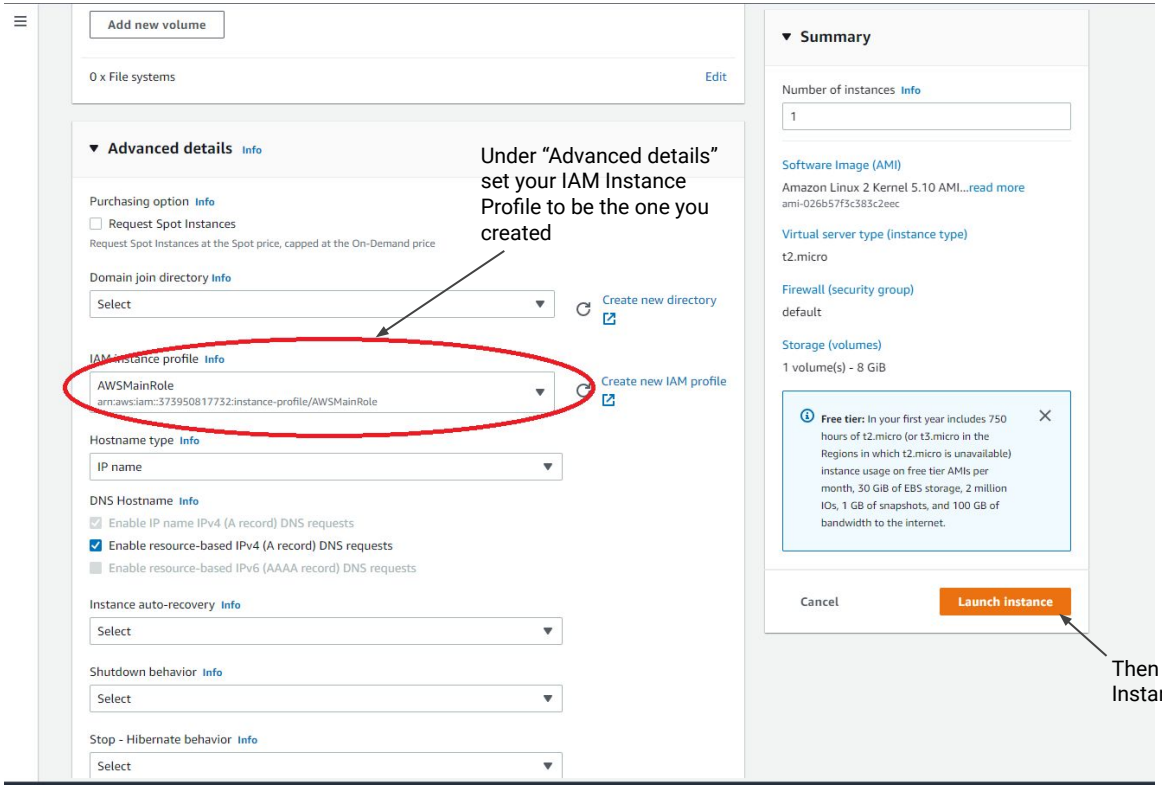
Setup EC2 Instance



Setup EC2 Instance



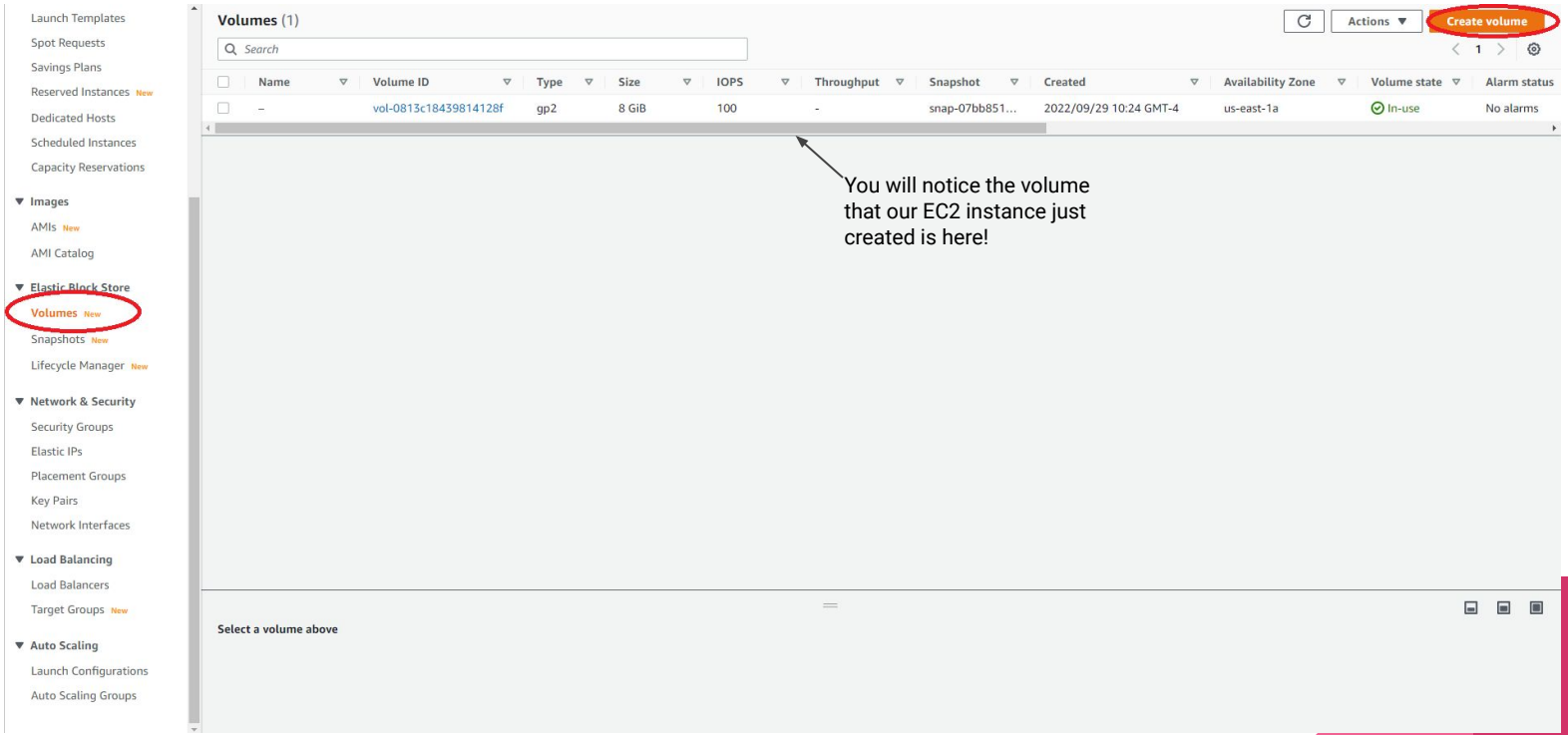
Setup EC2 Instance



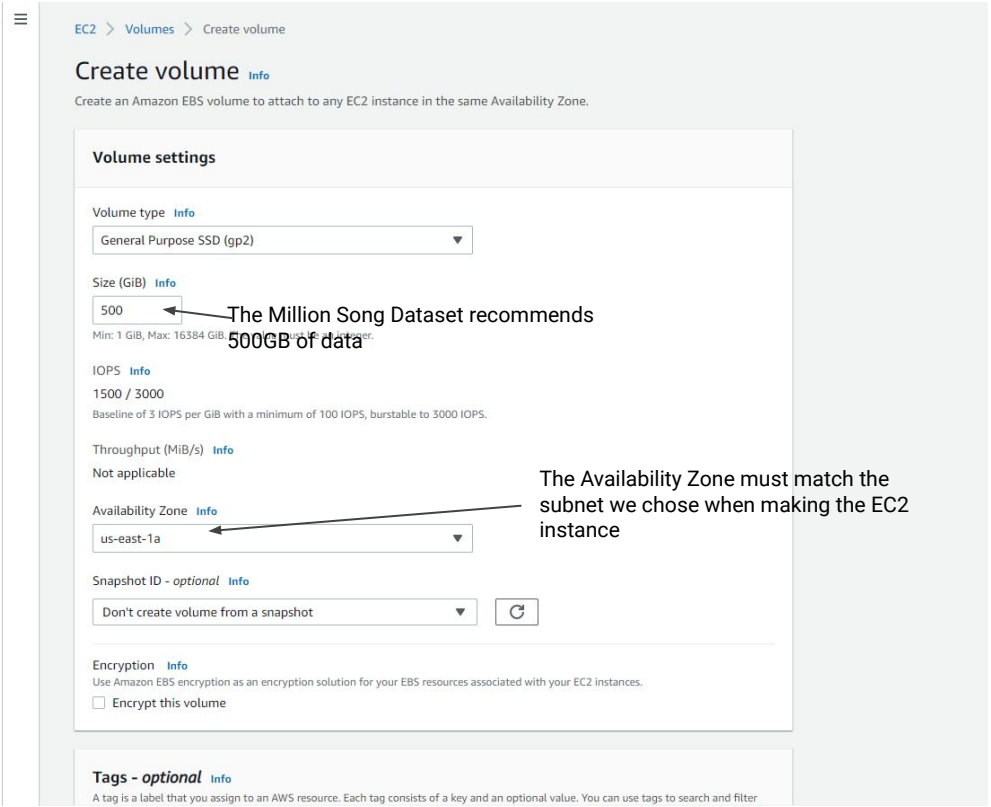
10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - Launching…
  - Add an additional Volume
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*

Add an additional Volume to your EC2 Instance

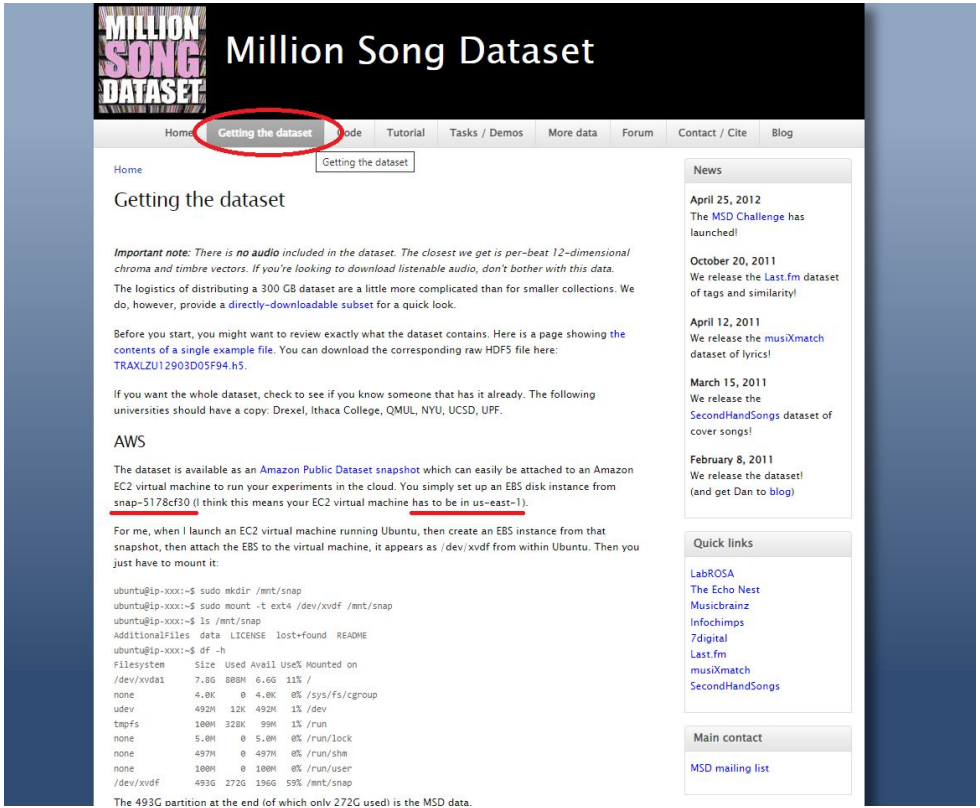


Add an additional Volume to your EC2 Instance



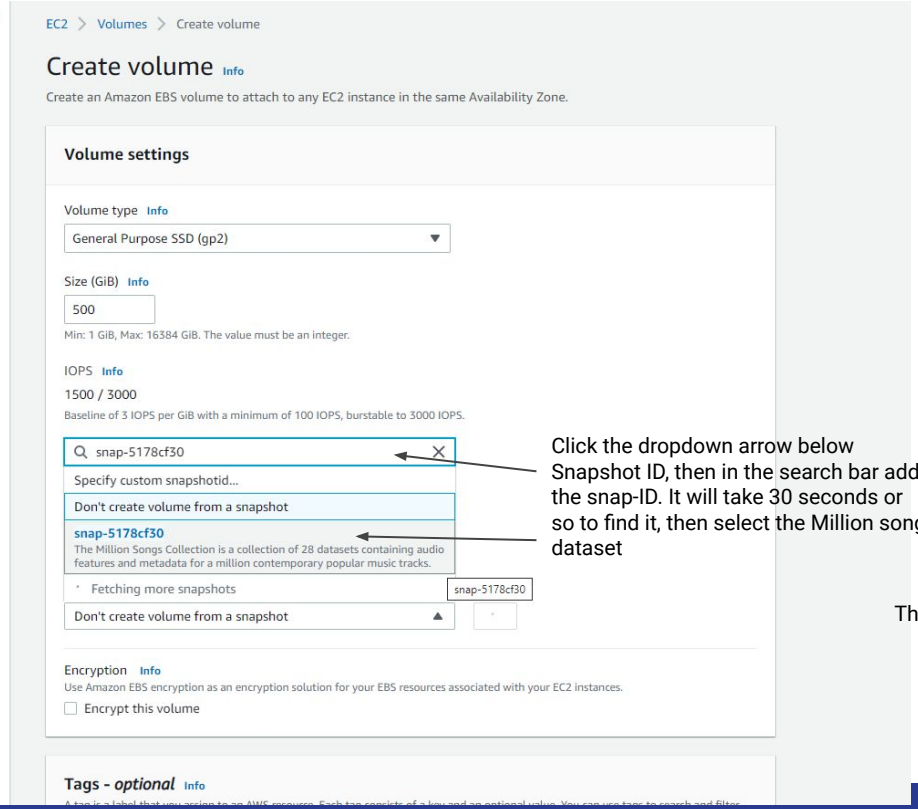
- We want this volume to come with our dataset on it. To do this we can use a SnapshotID
- Head to http://millionsongdataset.com/

Add an additional Volume to your EC2 Instance



- Click “Getting the Dataset” button on the top tab.
- Then copy the snap code: snap-5178cf30

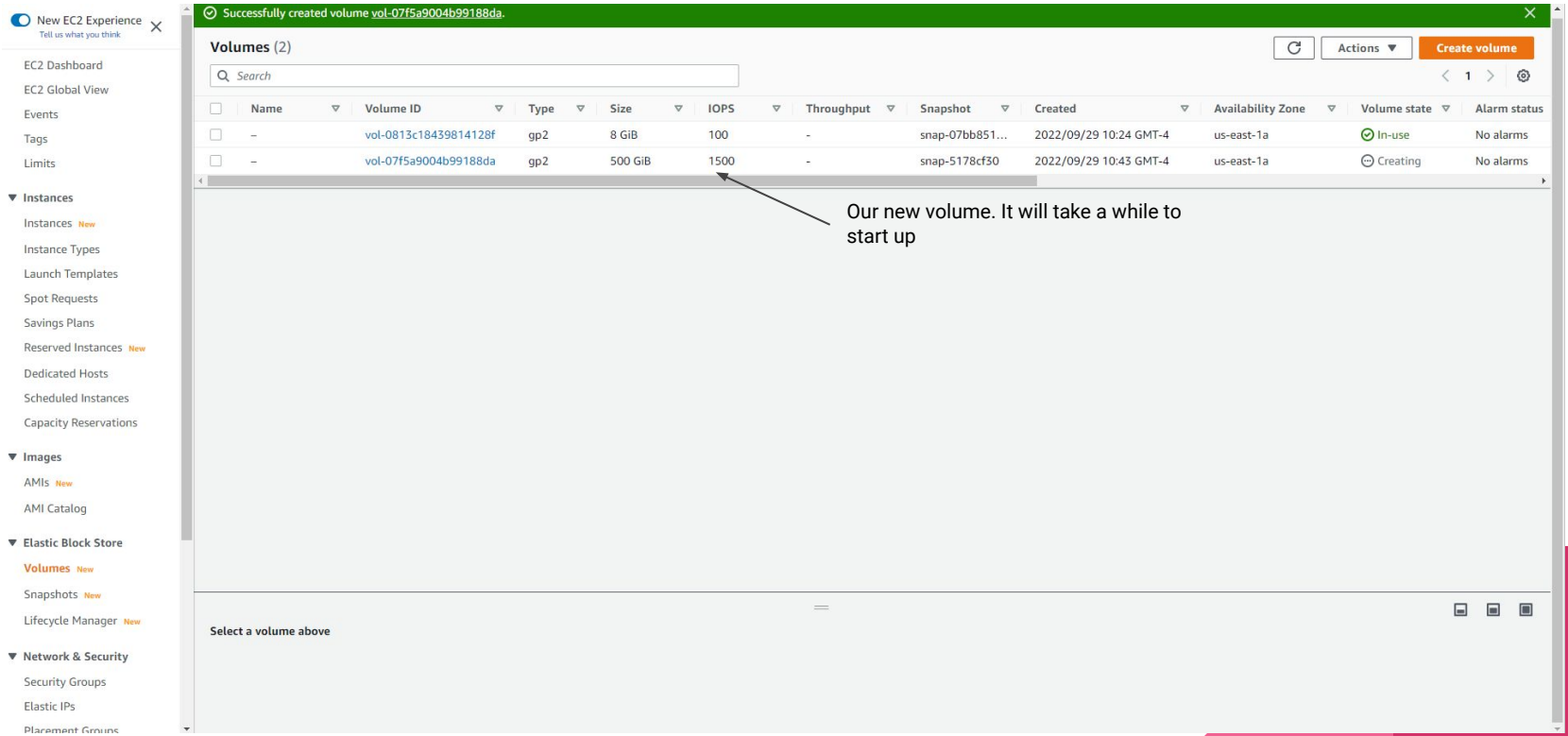
Add an additional Volume to your EC2 Instance



Click the dropdown arrow below Snapshot ID, then in the search bar add the snap-ID. It will take 30 seconds or so to find it, then select the Million song dataset

Then click create volume!

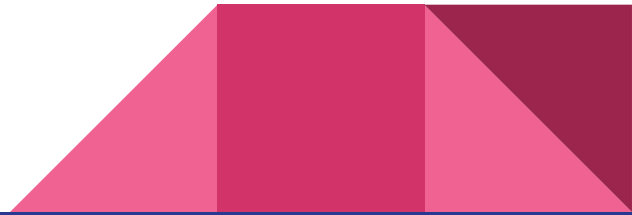
Add an additional Volume to your EC2 Instance



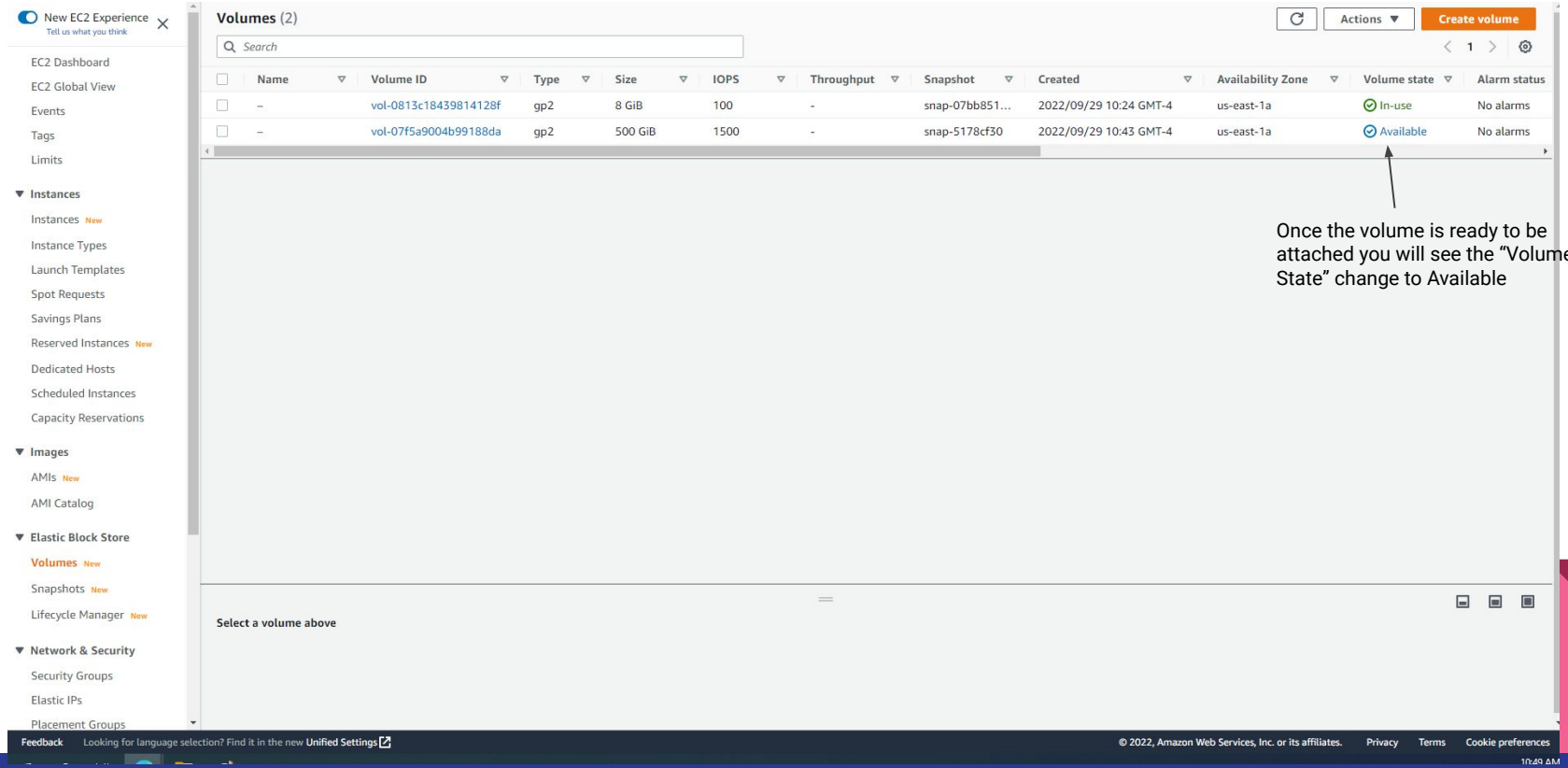
Add an additional Volume to your EC2 Instance

A note on Volumes: This will be easily the most expensive part of your HW4, it is also the easiest to forget to terminate!

I recommend only setting this up when you plan to run million_song_prep.py and terminating the volume after it has finished running and you can see your csv files in your S3 bucket.

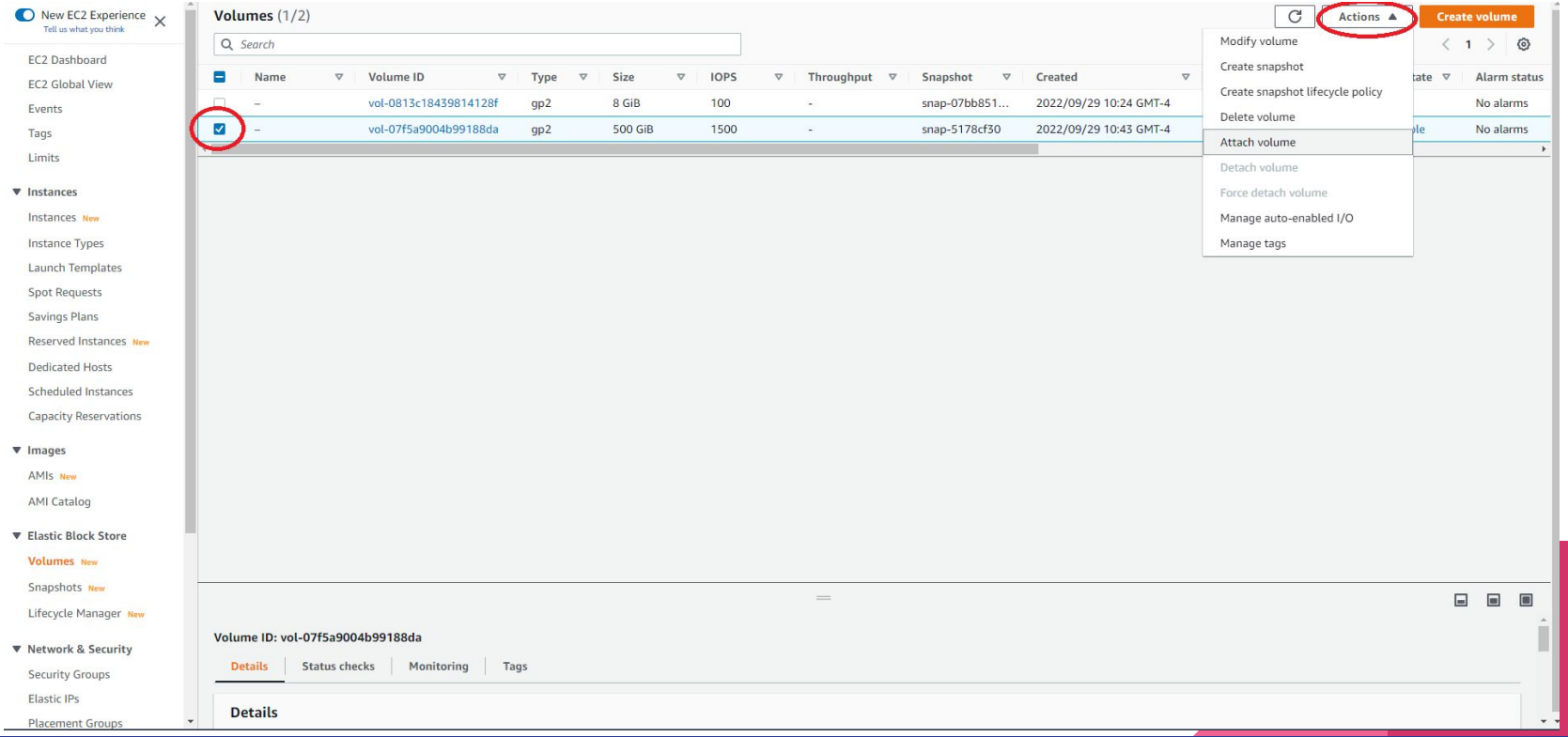


Add an additional Volume to your EC2 Instance

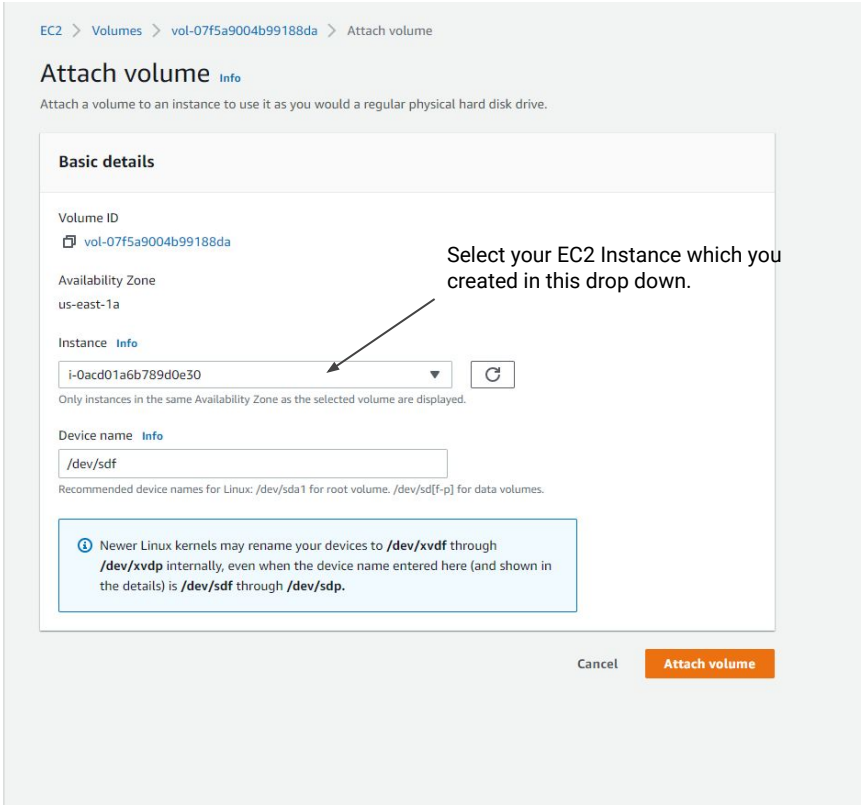


Once the volume is ready to be attached you will see the “Volume State” change to Available

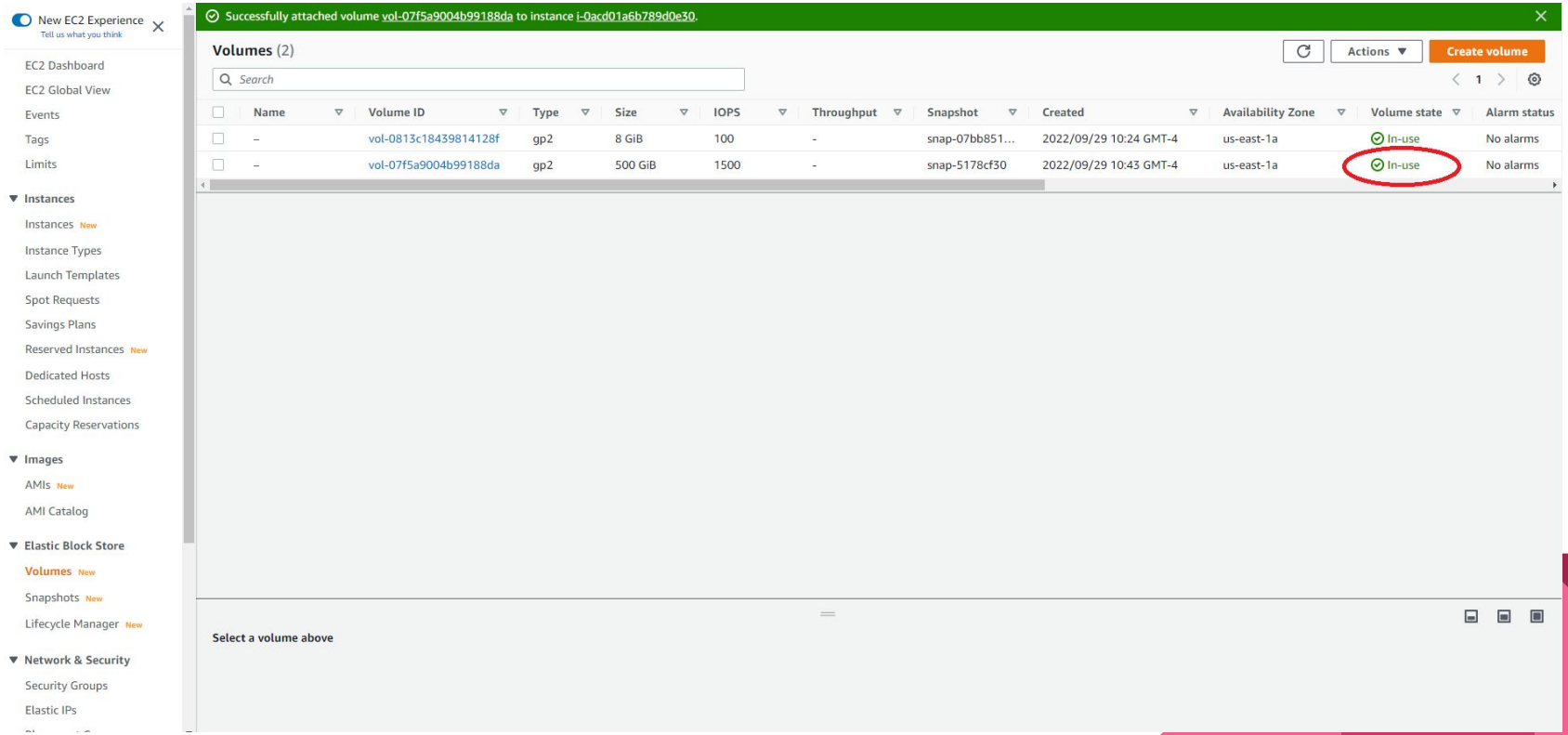
Add an additional Volume to your EC2 Instance



Add an additional Volume to your EC2 Instance



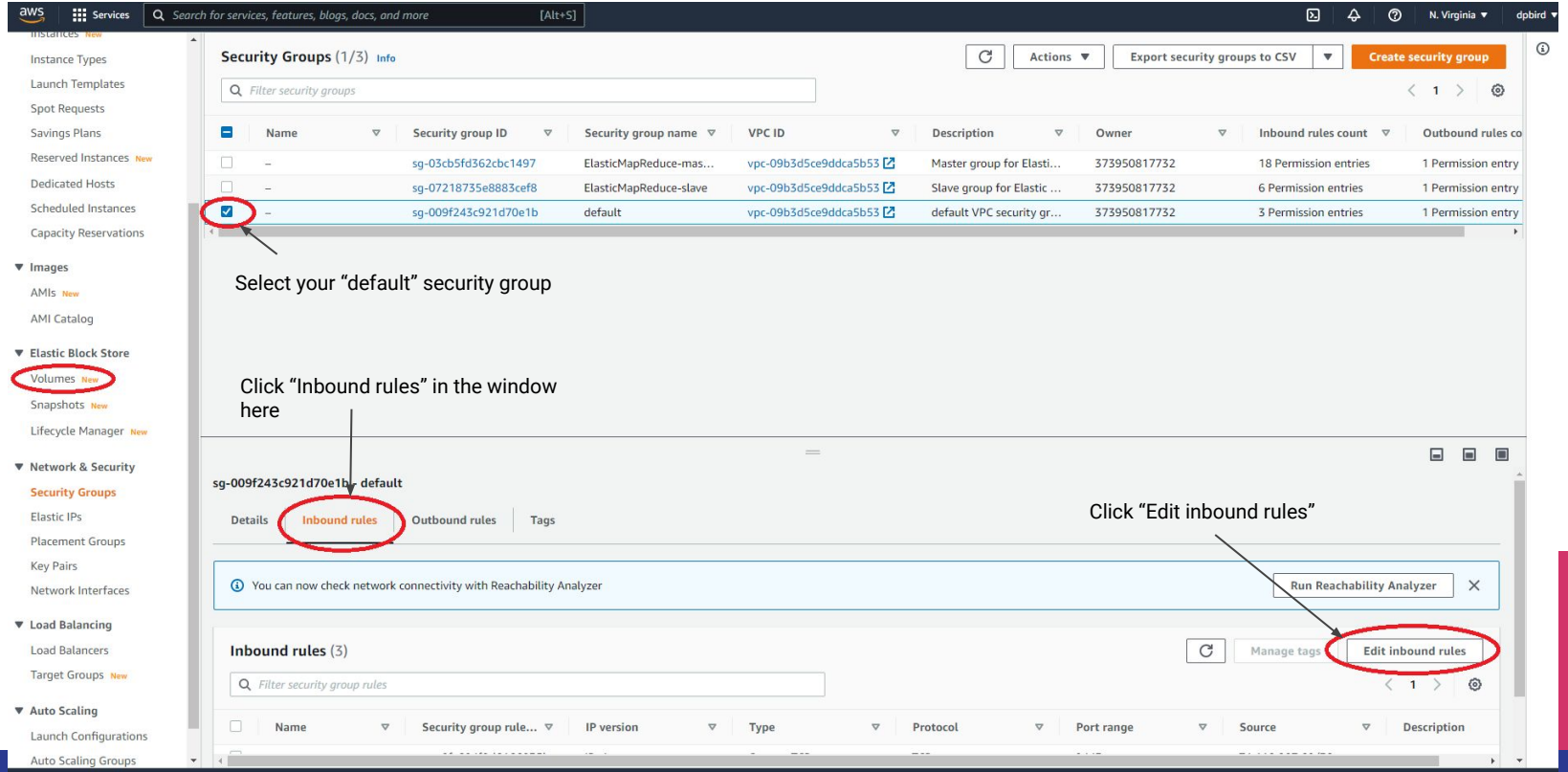
Add an additional Volume to your EC2 Instance



10-605 Setup for Homework Part A

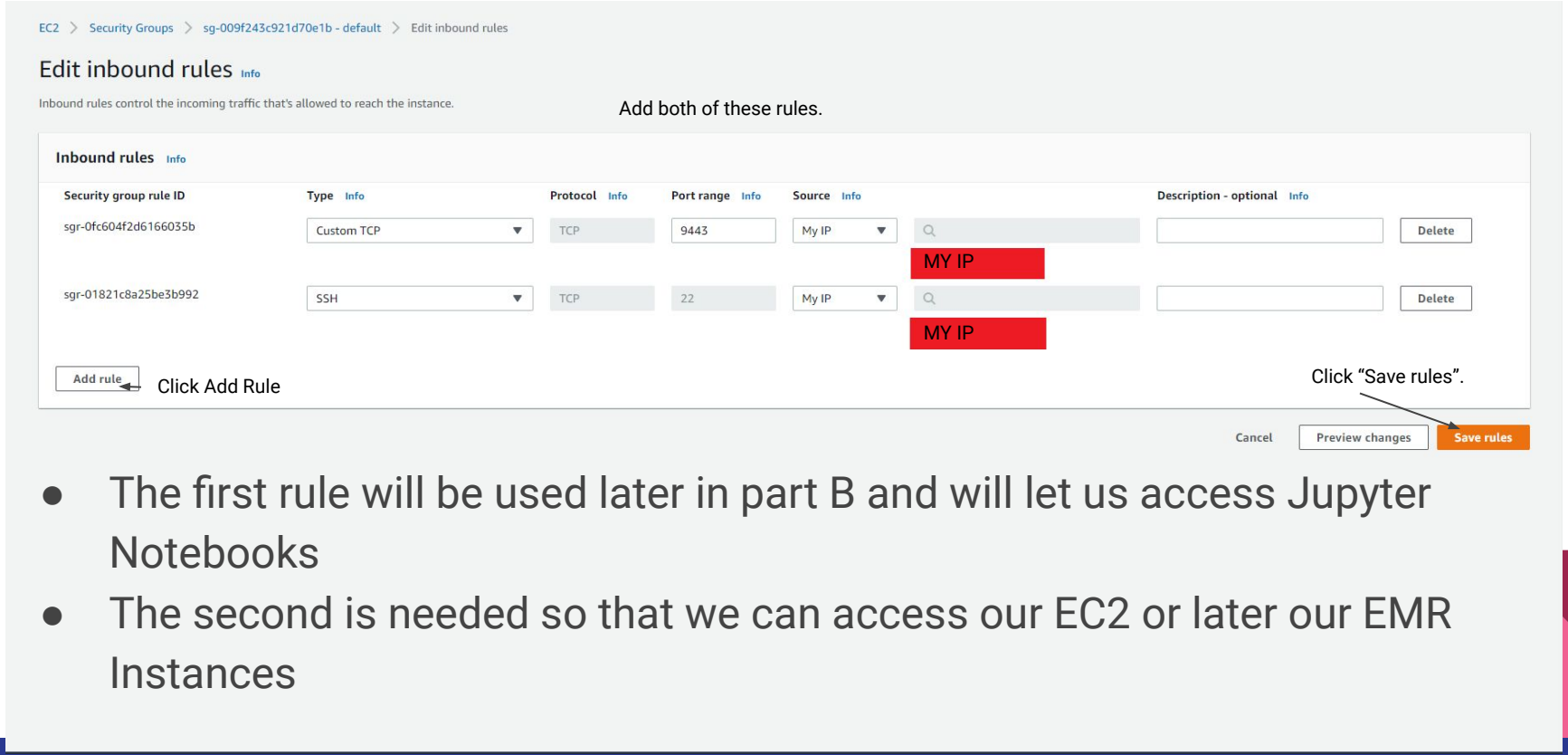
- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
  - Add an additional Volume - DONE
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*

Setup Security Groups



Setup Security Groups

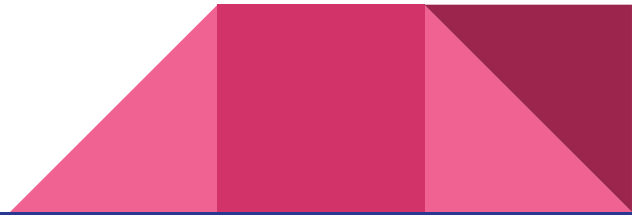
Post Note: If you are getting an issue when adding inbound rules to a default security group, try deleting all of the current rules (Even if they are blank) and then clicking “New rule”, this should let you update the rules properly.



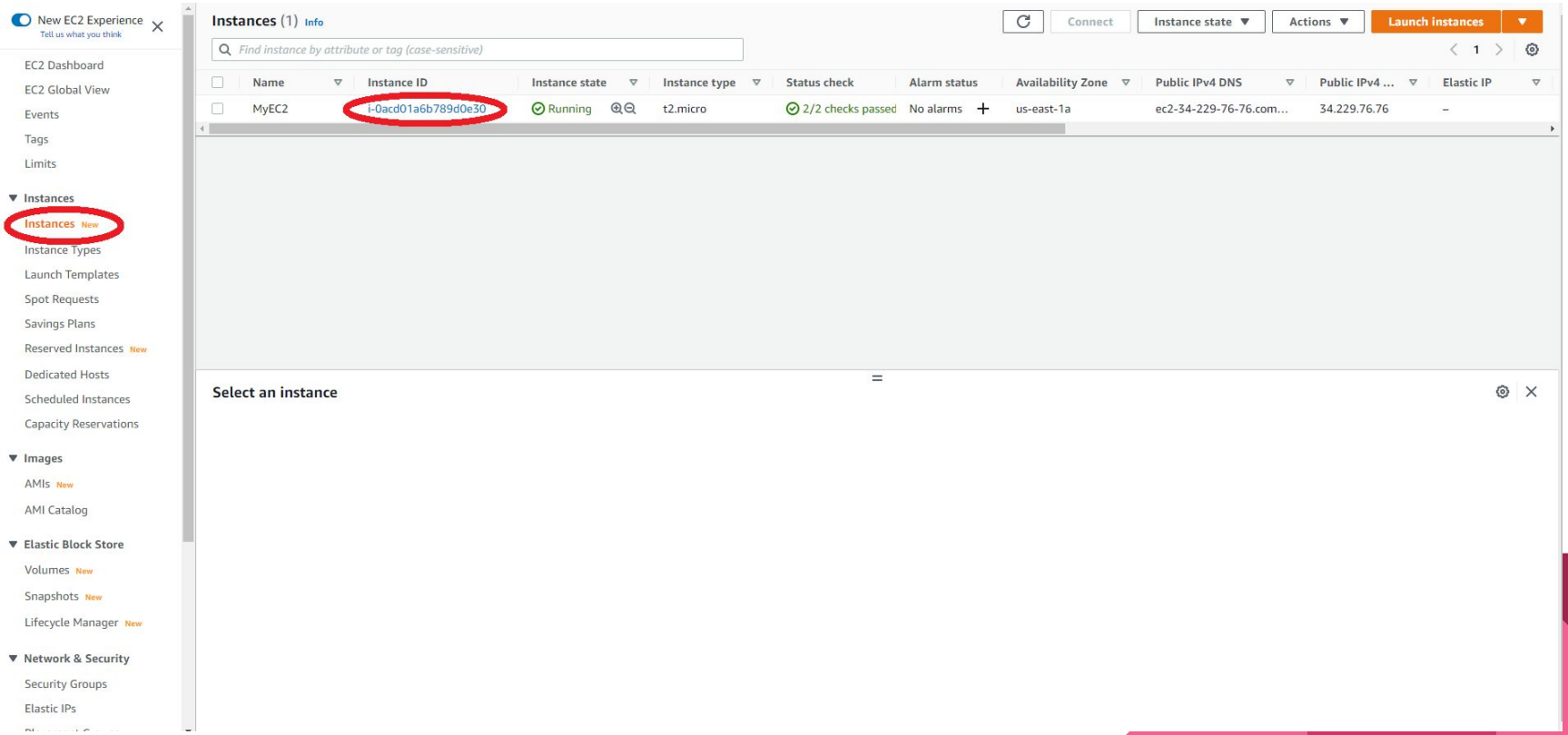
- The first rule will be used later in part B and will let us access Jupyter Notebooks
- The second is needed so that we can access our EC2 or later our EMR Instances

10-605 Setup for Homework Part A

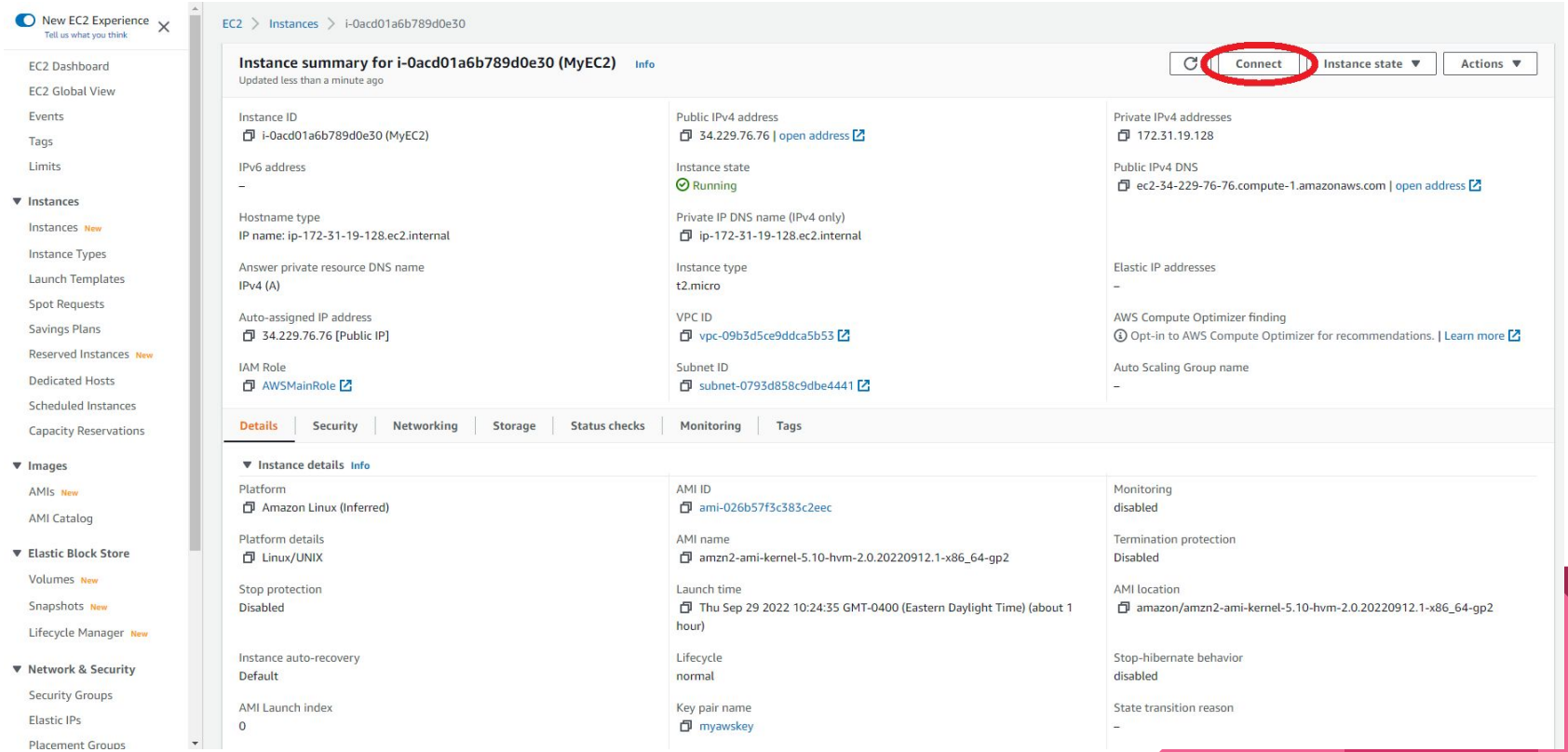
- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
  - Add an additional Volume - DONE
- Setup Security Groups - DONE
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1 We need this to get the Million Song Dataset*



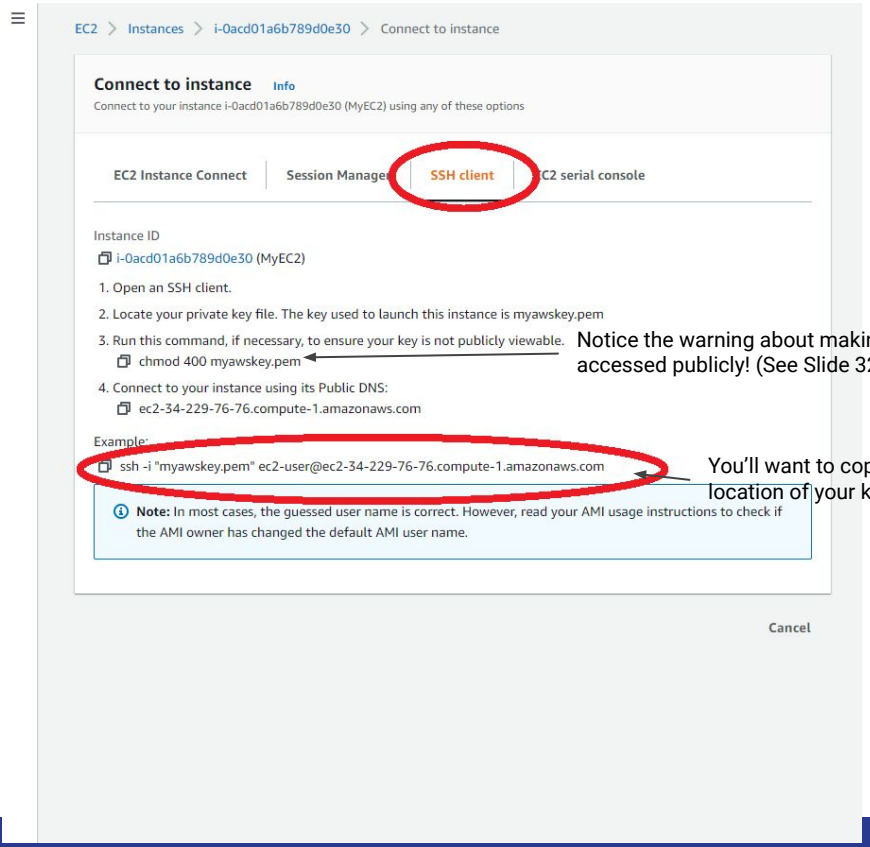
Get access to our EC2 Instance



Get access to our EC2 Instance



Get access to our EC2 Instance

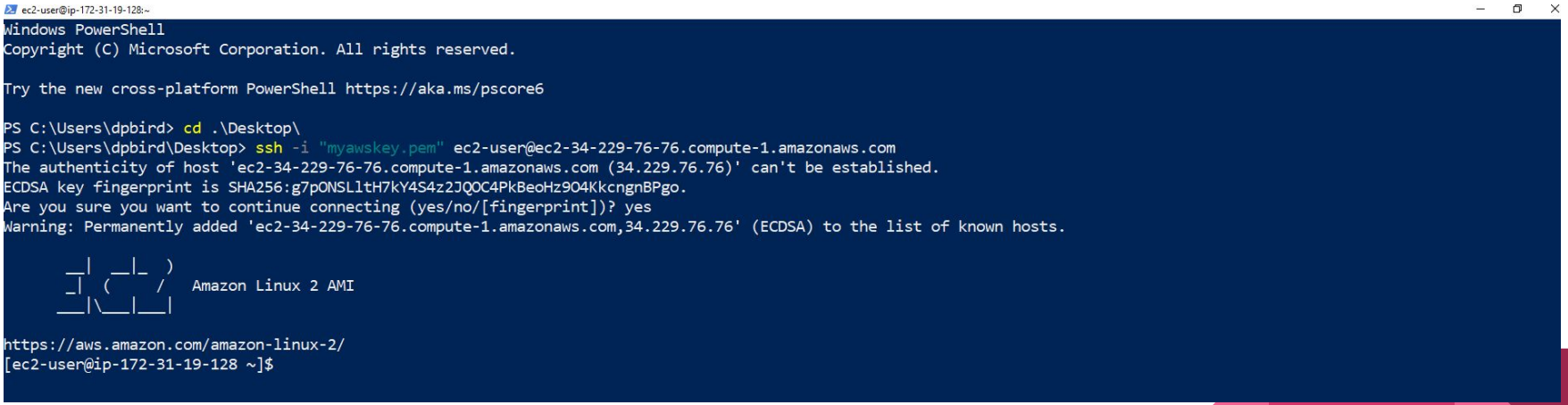


Notice the warning about making sure the key is not able to be accessed publicly! (See Slide 32)

You'll want to copy this and update the "myawskey.pem" to be the location of your key.

Get access to our EC2 Instance

- Launch either Windows Powershell or Terminal (Mac) and paste the ssh command:

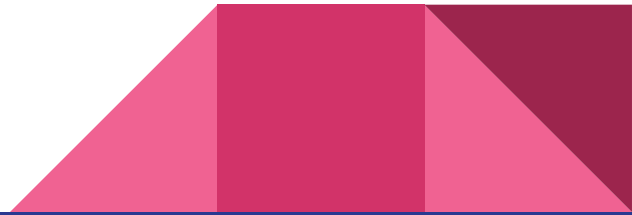


Get access to our EC2 Instance

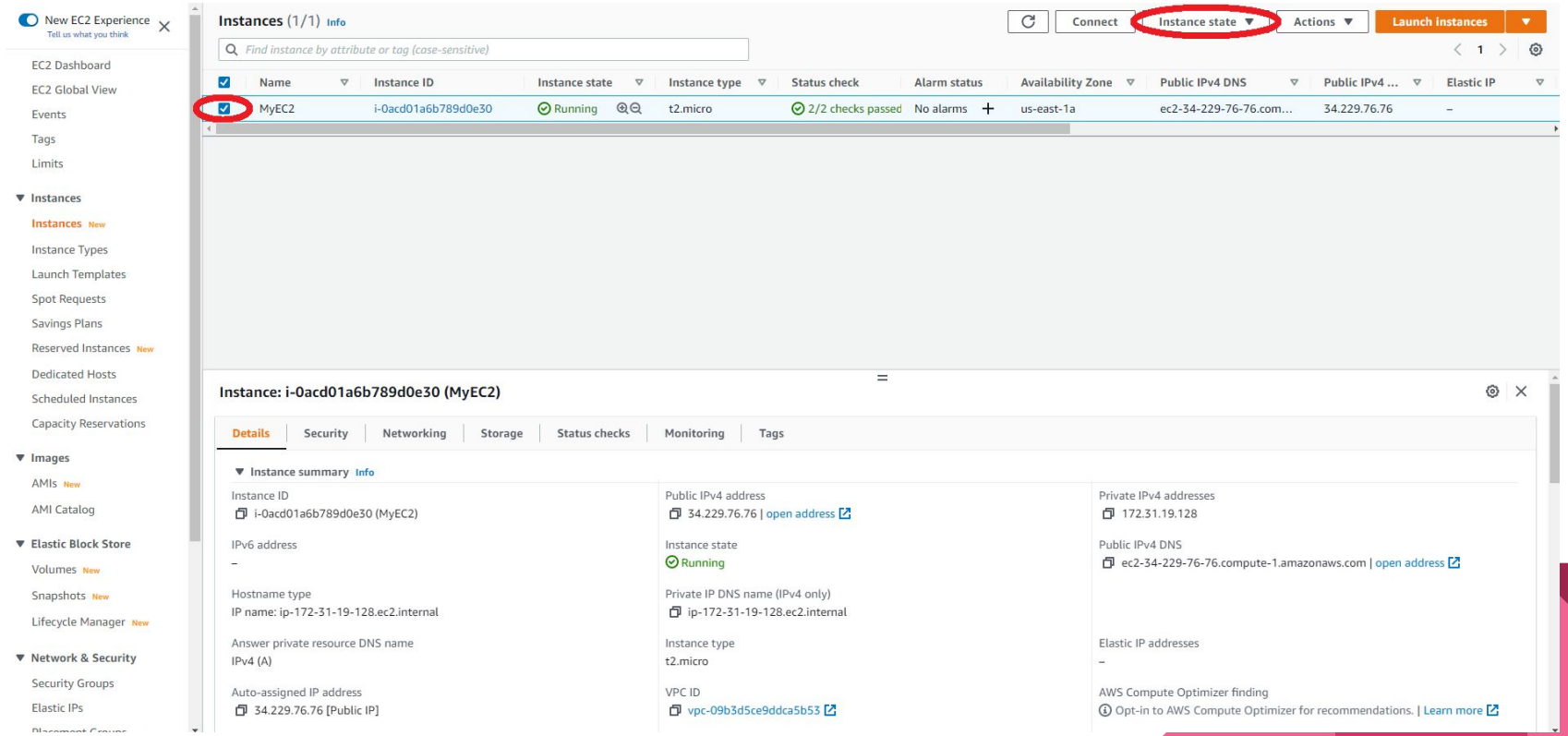
- Once logged in you will be able to install all of the necessary packages and run the million_song_prep.py file. e.g:
  - pip3 install matplotlib
- Mount your 500gb Volume:
  - lsblk - *Lists the mountable drives so you can find the name of the 500gb drive*
  - mkdir songs - *creates a folder called songs which we mount the drive to*
  - sudo mount /dev/#volumename songs - *Mounts the drive to songs folder*
- A concrete guide to PART A this will be on Friday October 14th

10-605 Setup for Homework Part A

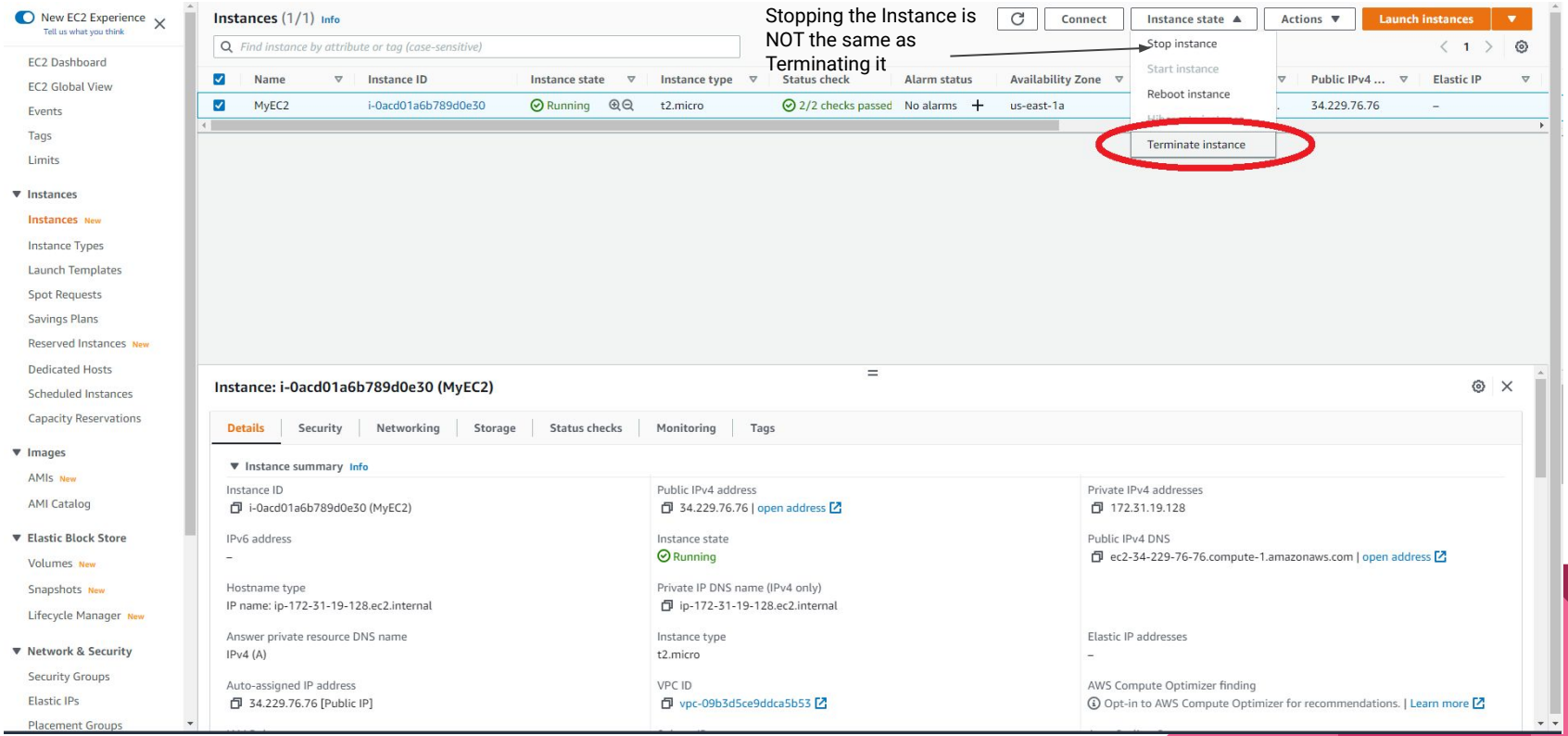
- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
  - Add an additional Volume - DONE
- Setup Security Groups - DONE
- Get access - DONE
- **Terminating your EC2 instance and Volume!**
- run million_song_prep.py - Convert Data from .h5 to .csv - Recitation



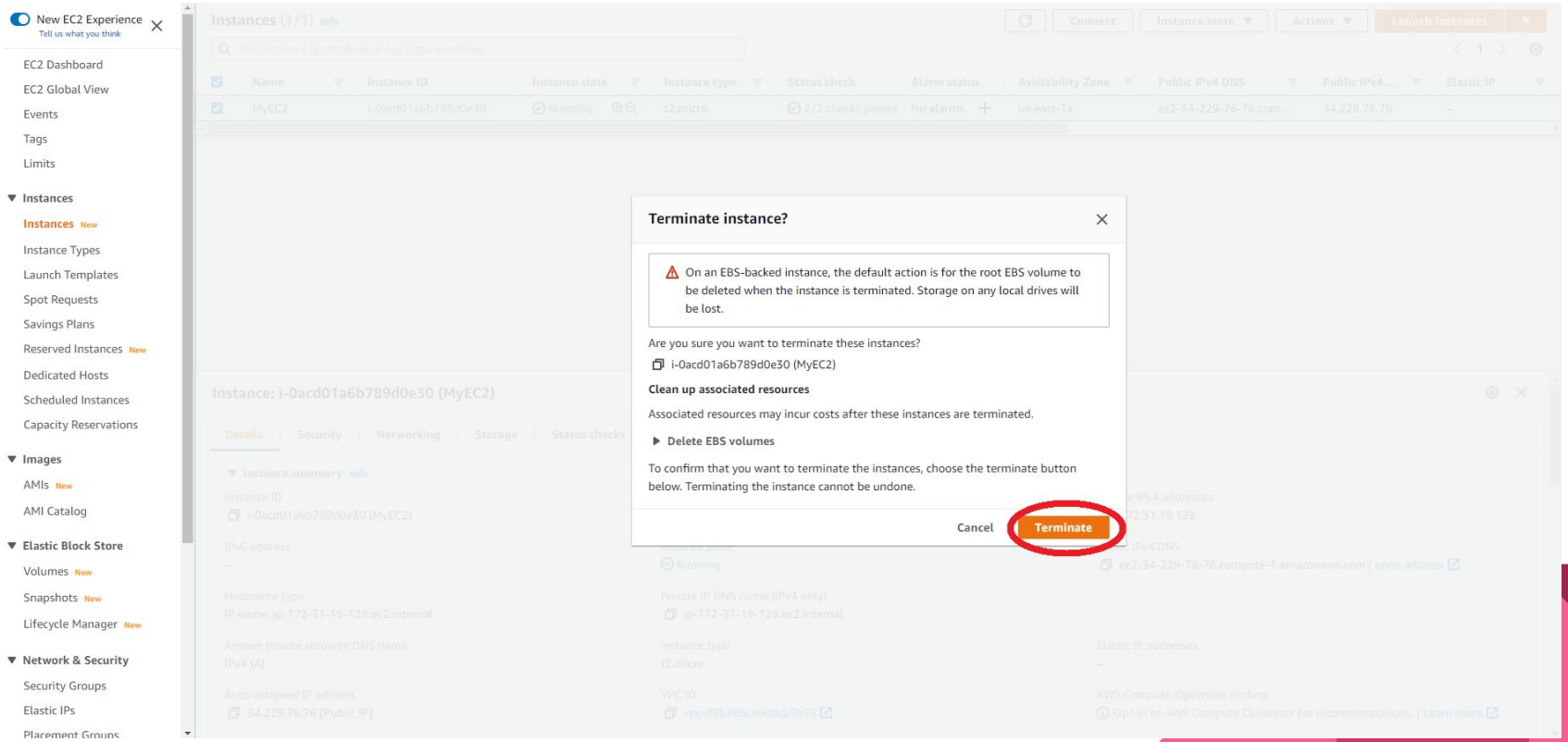
Terminating your EC2 instance and Volume!



Terminating your EC2 instance and Volume!



Terminating your EC2 instance and Volume!



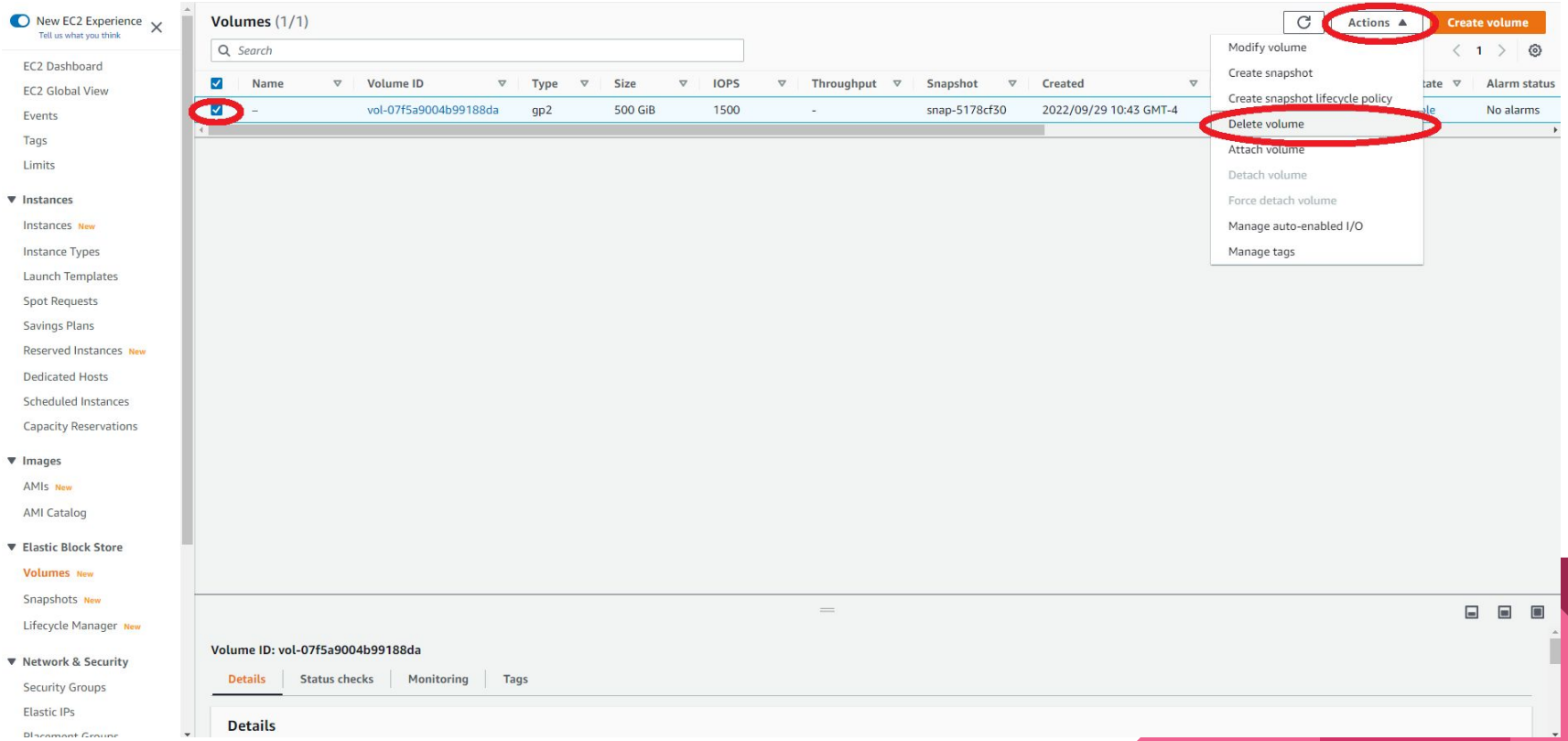
Terminating your EC2 instance and Volume!

The EC2 instance has now been Terminated. Stopping it is not enough as you will still be charged. Once it is terminated everything on the instance will be lost. So make sure that your data, after processing, is all in the S3 bucket.

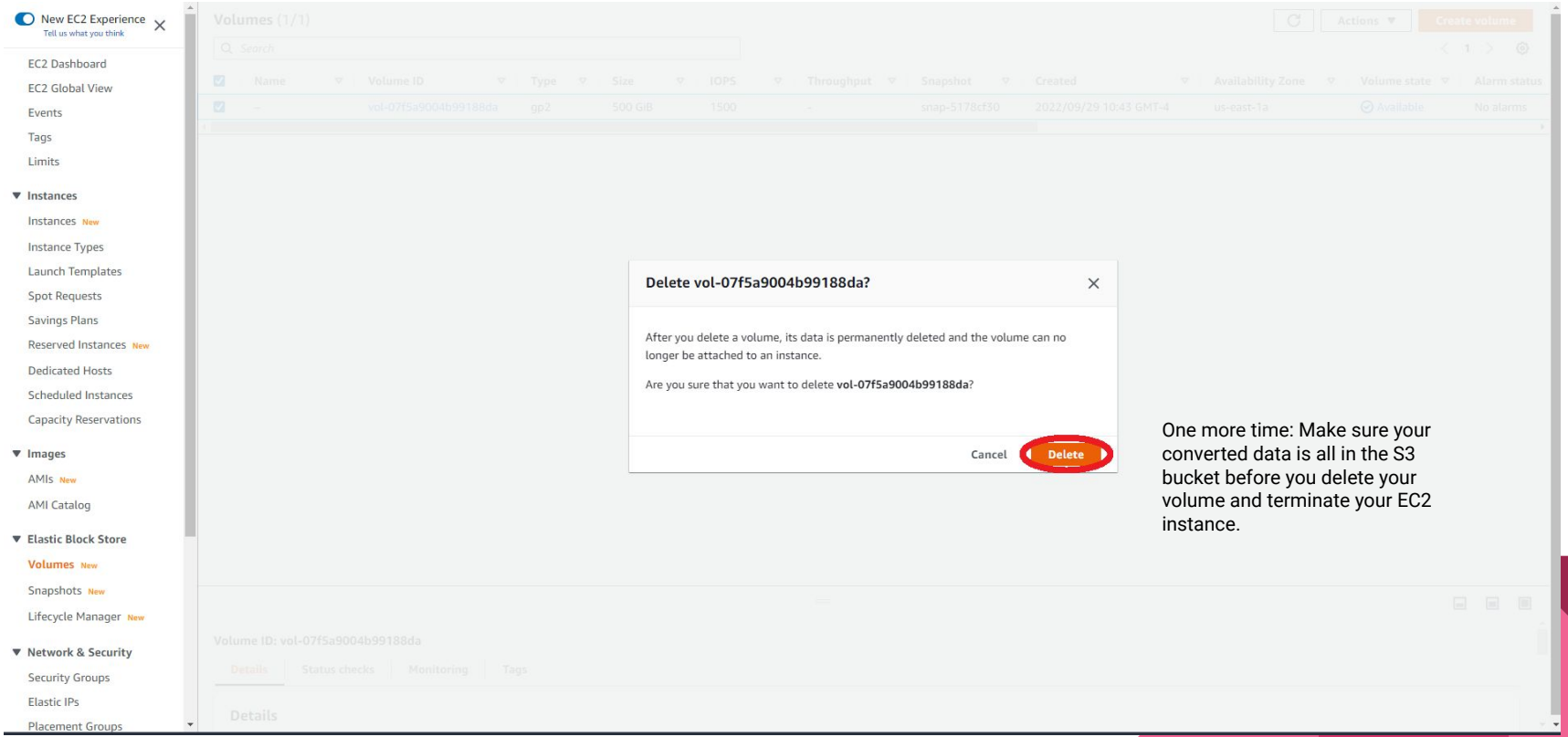
Once you delete the instance it will take a little, during that time you will not be able to terminate the 500gb volume we created, as it is still attached. Once the EC2 instance is terminated the option to Delete the volume will become available.

You need to delete your volume separately! It is vitally important that you do not forget this!!!

Terminating your EC2 instance and Volume!



Terminating your EC2 instance and Volume!



One more time: Make sure your converted data is all in the S3 bucket before you delete your volume and terminate your EC2 instance.

10-605 Setup for Homework Part A

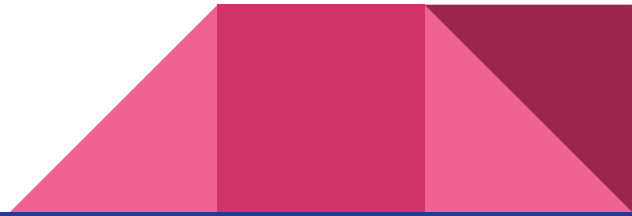
- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
  - Add an additional Volume - DONE
- Setup Security Groups - DONE
- Get access - DONE
- **Terminating your EC2 instance and Volume! - DONE**
- run million_song_prep.py - Convert Data from .h5 to .csv - Recitation

10-605 Setup for Homework Part B

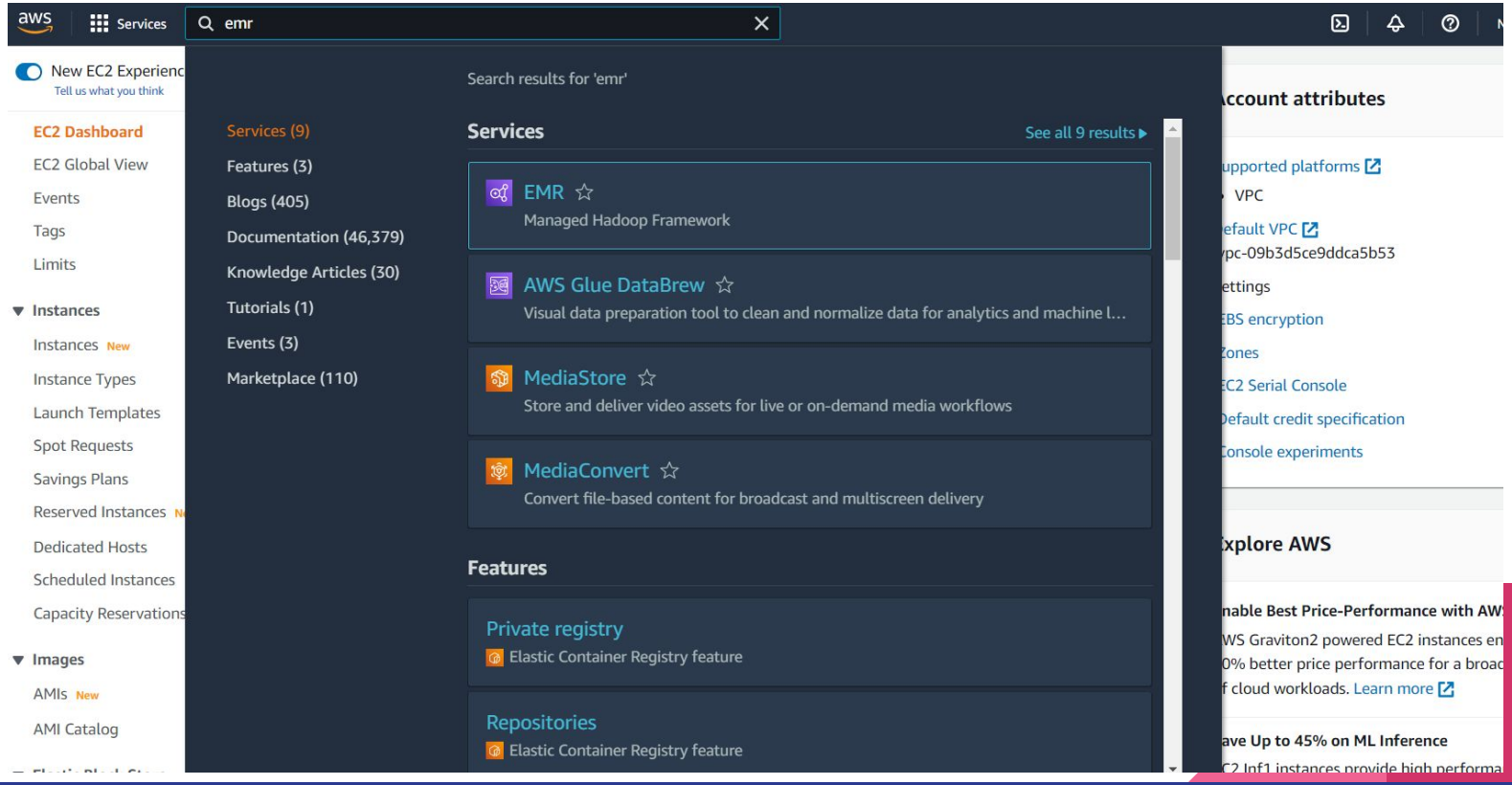
- Setup EMR Cluster
- Log into Master Node
- Get Access to Jupyter Notebooks Interface

Setup EMR Cluster

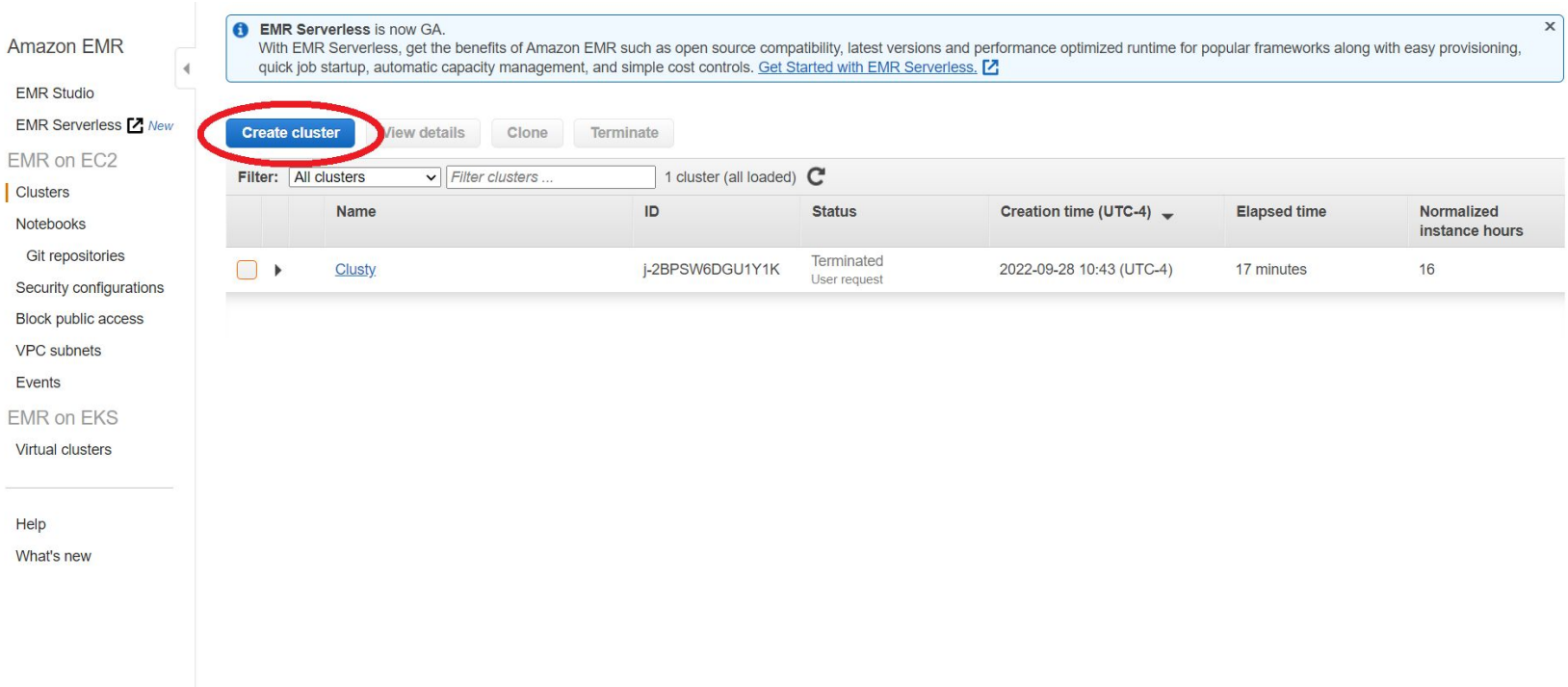
- What is EMR? (Elastic MapReduce)
- It is a managed cluster platform typically used for large data services
- It lets us set up a master node and worker nodes similar to what we have been discussing in the course, allowing us to run Hadoop and Apache Spark to work with large datasets!



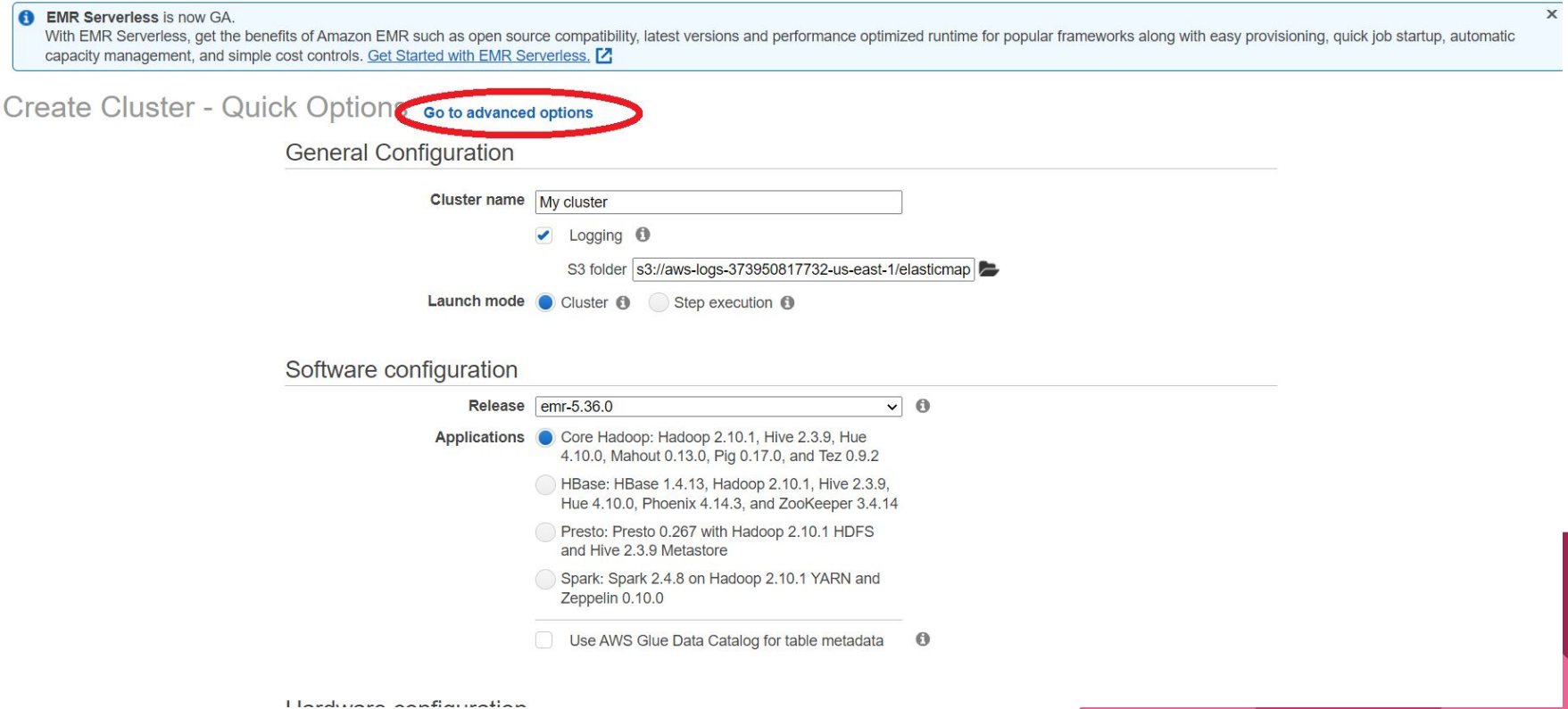
Setup EMR Cluster



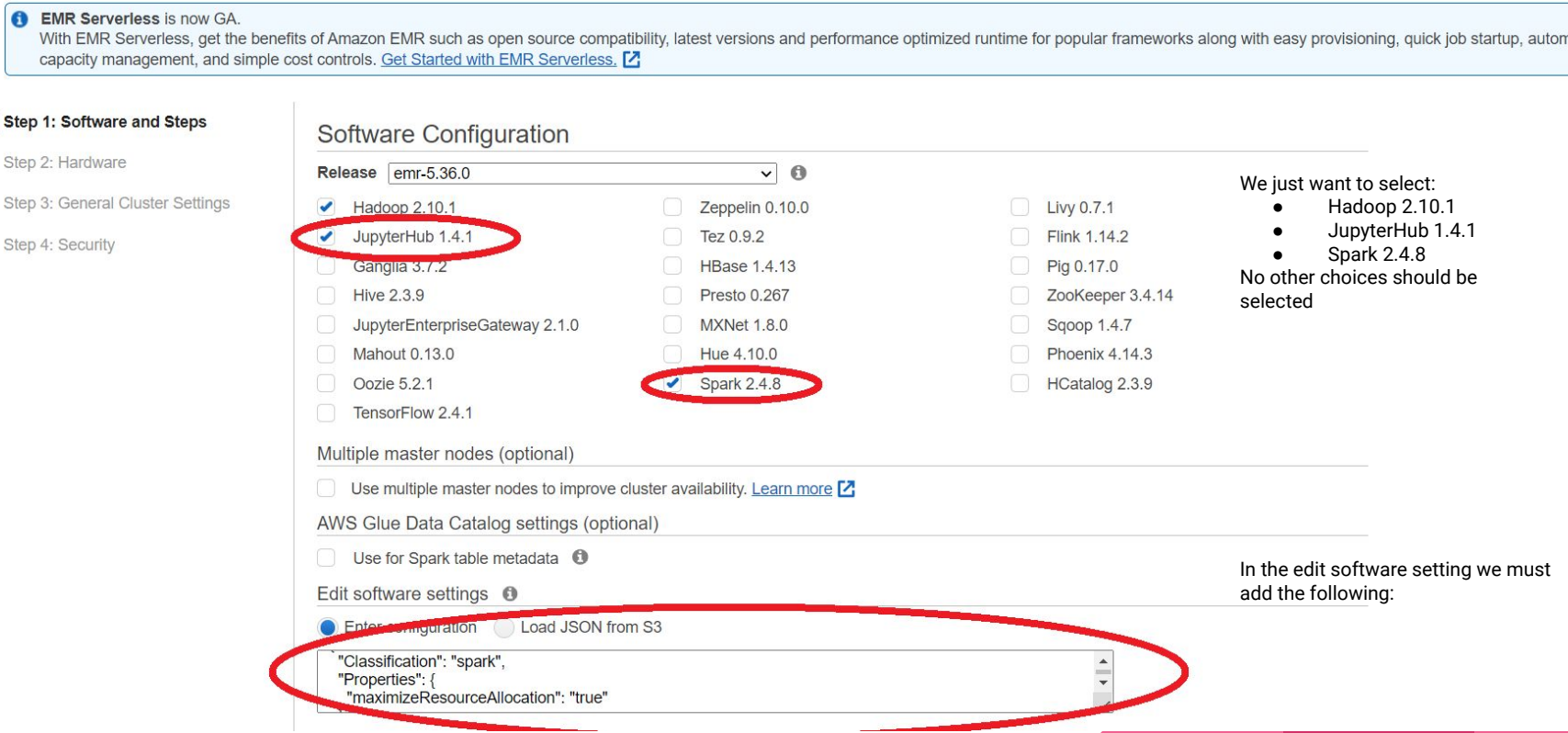
Setup EMR Cluster



Setup EMR Cluster



Setup EMR Cluster



We just want to select:

- Hadoop 2.10.1
- JupyterHub 1.4.1
- Spark 2.4.8

No other choices should be selected

In the edit software setting we must add the following:

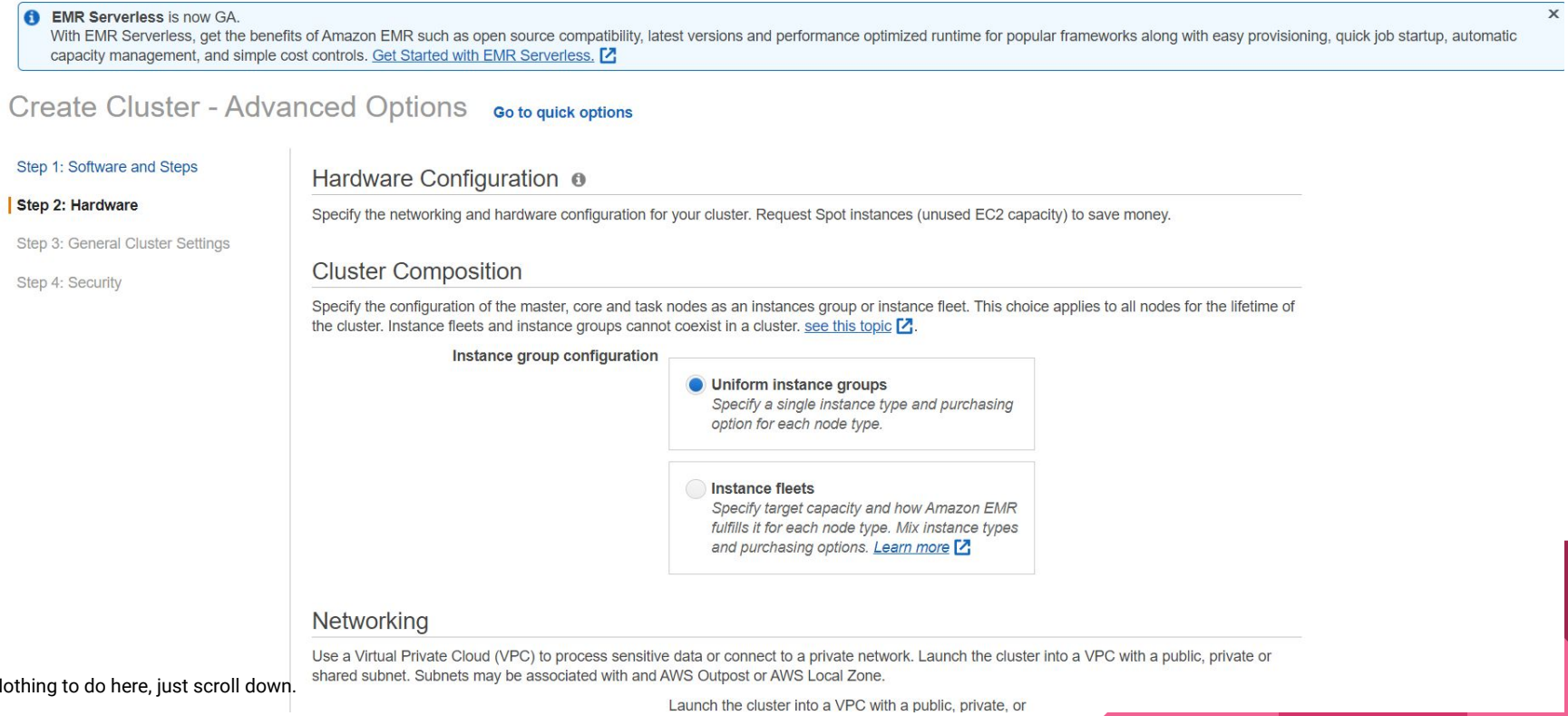
Edit Software Setting

- Configuring spark: https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark-configure.html
- Add:

```
[
  {
    "Classification" : "spark",
    "Properties" : {
      "maximizeResourceAllocation" : "true"
    }
  }
]
```

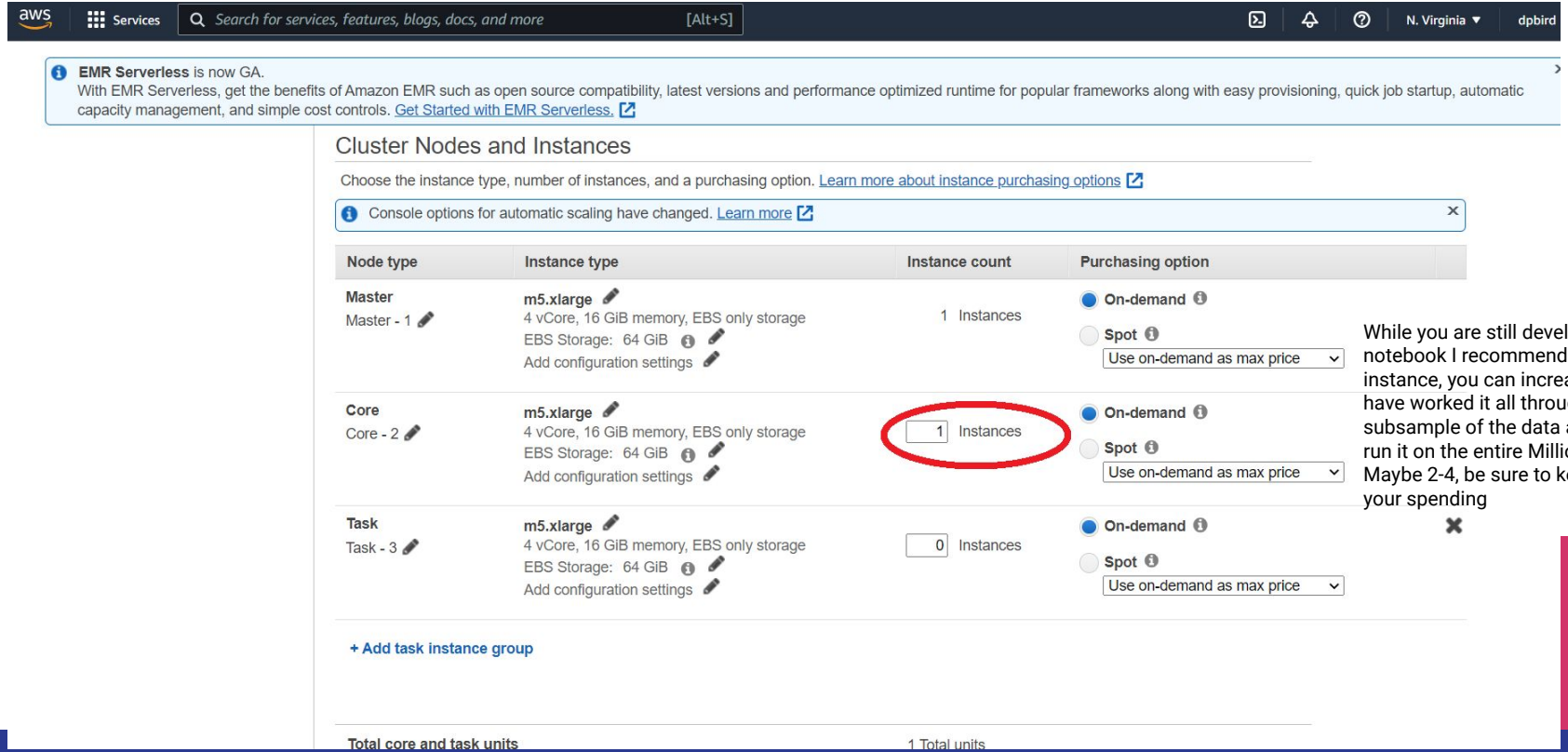
Hit Next to get to step 2:

Setup EMR Cluster



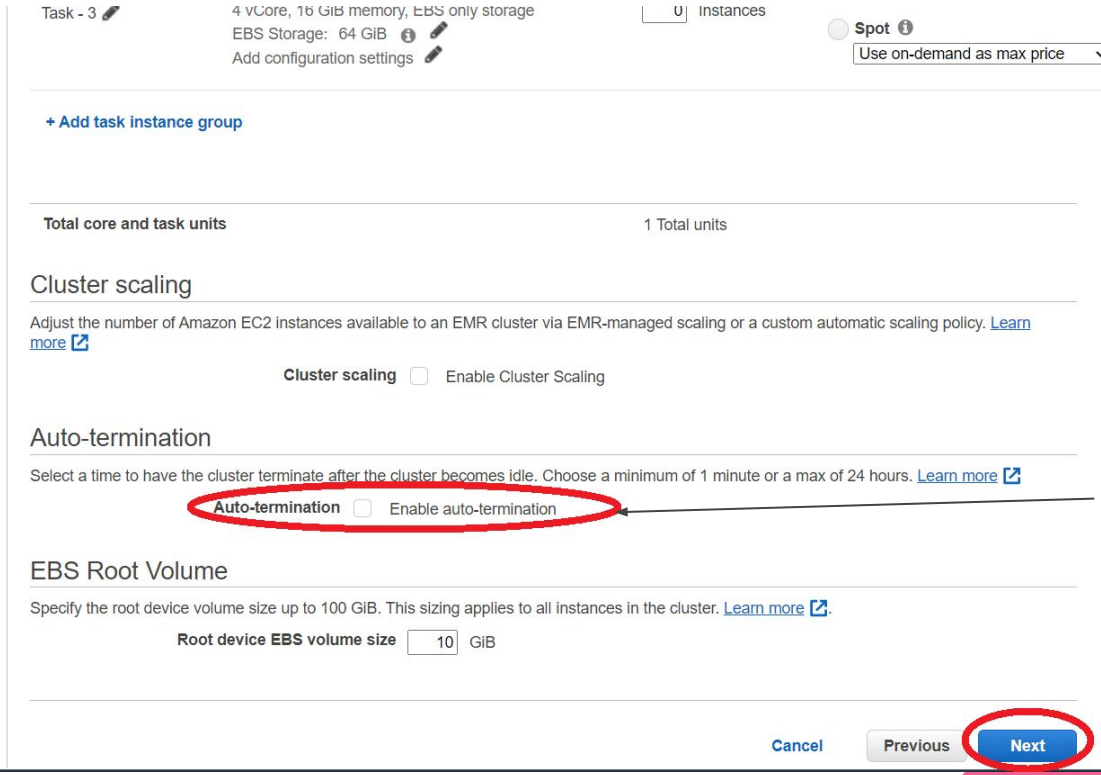
Nothing to do here, just scroll down.

Setup EMR Cluster



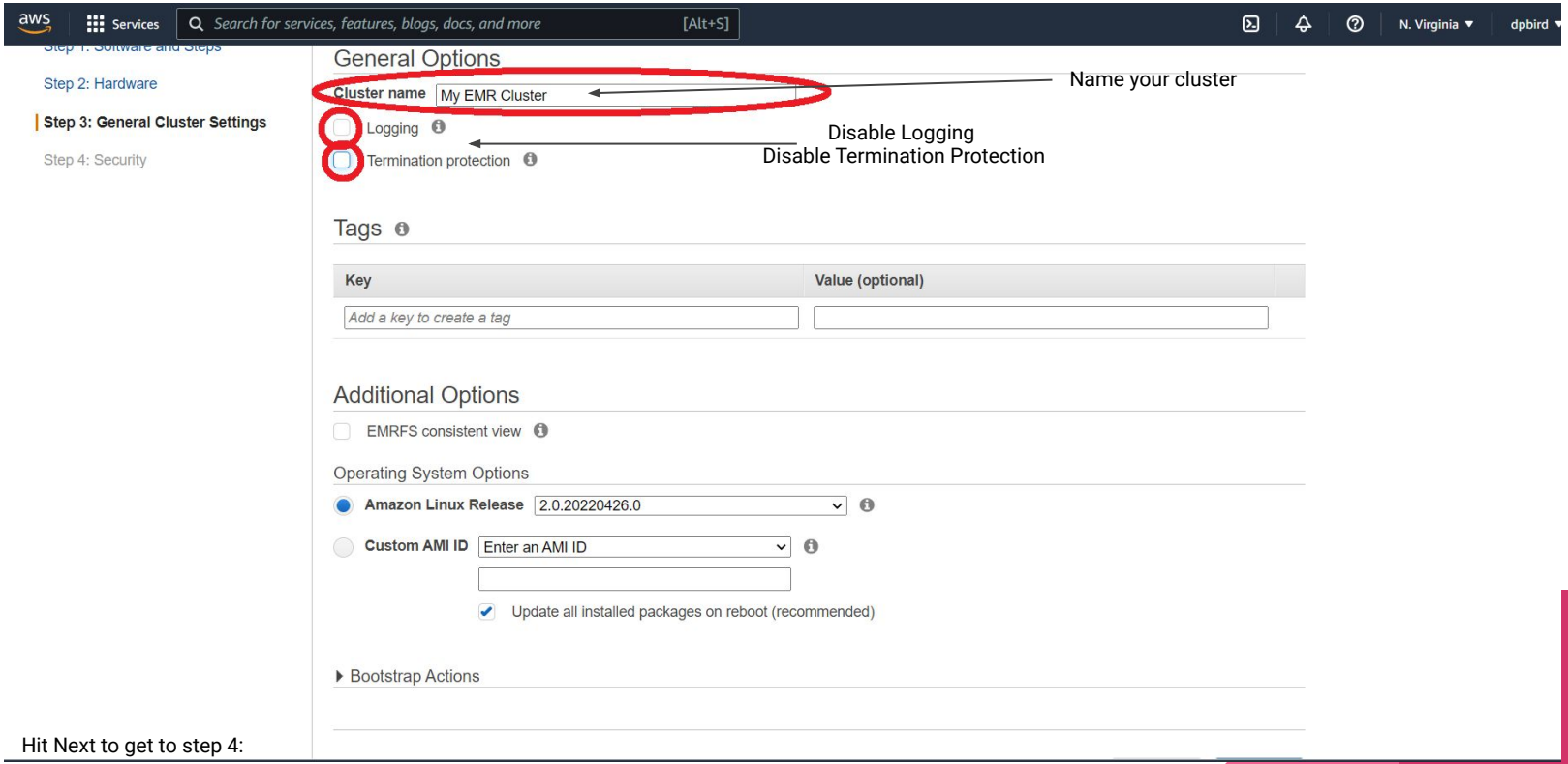
While you are still developing your notebook I recommend using just 1 Core instance, you can increase this once you have worked it all through on the subsample of the data and you want to run it on the entire Million song dataset. Maybe 2-4, be sure to keep an eye on your spending

Setup EMR Cluster



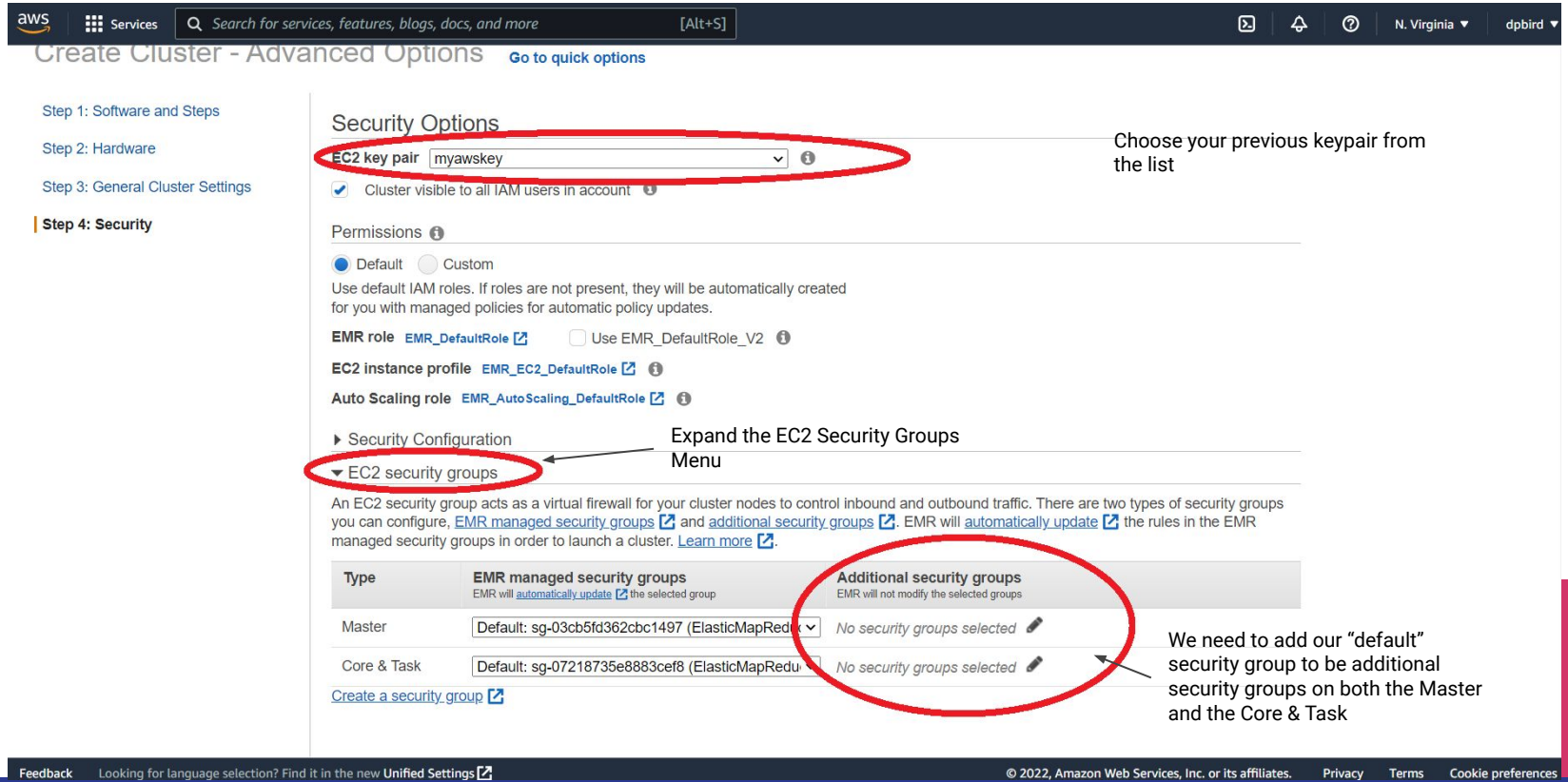
I recommend turning Auto-termination off. If you leave your cluster idle for the specified amount of time it will automatically terminate the cluster, if it terminates before you download your ipynb then you will lose it.

Setup EMR Cluster



Hit Next to get to step 4:

Setup EMR Cluster

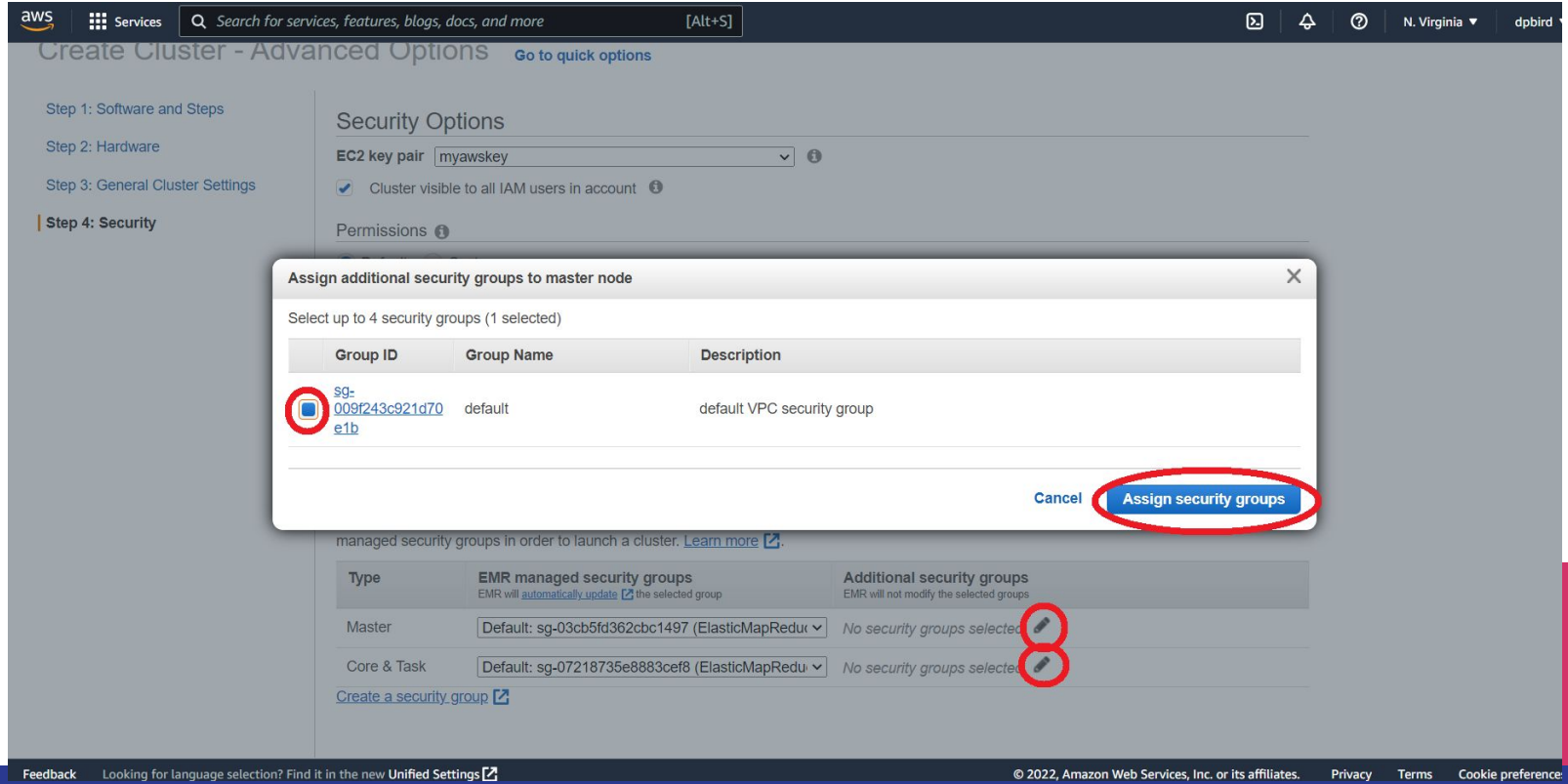


Choose your previous keypair from the list

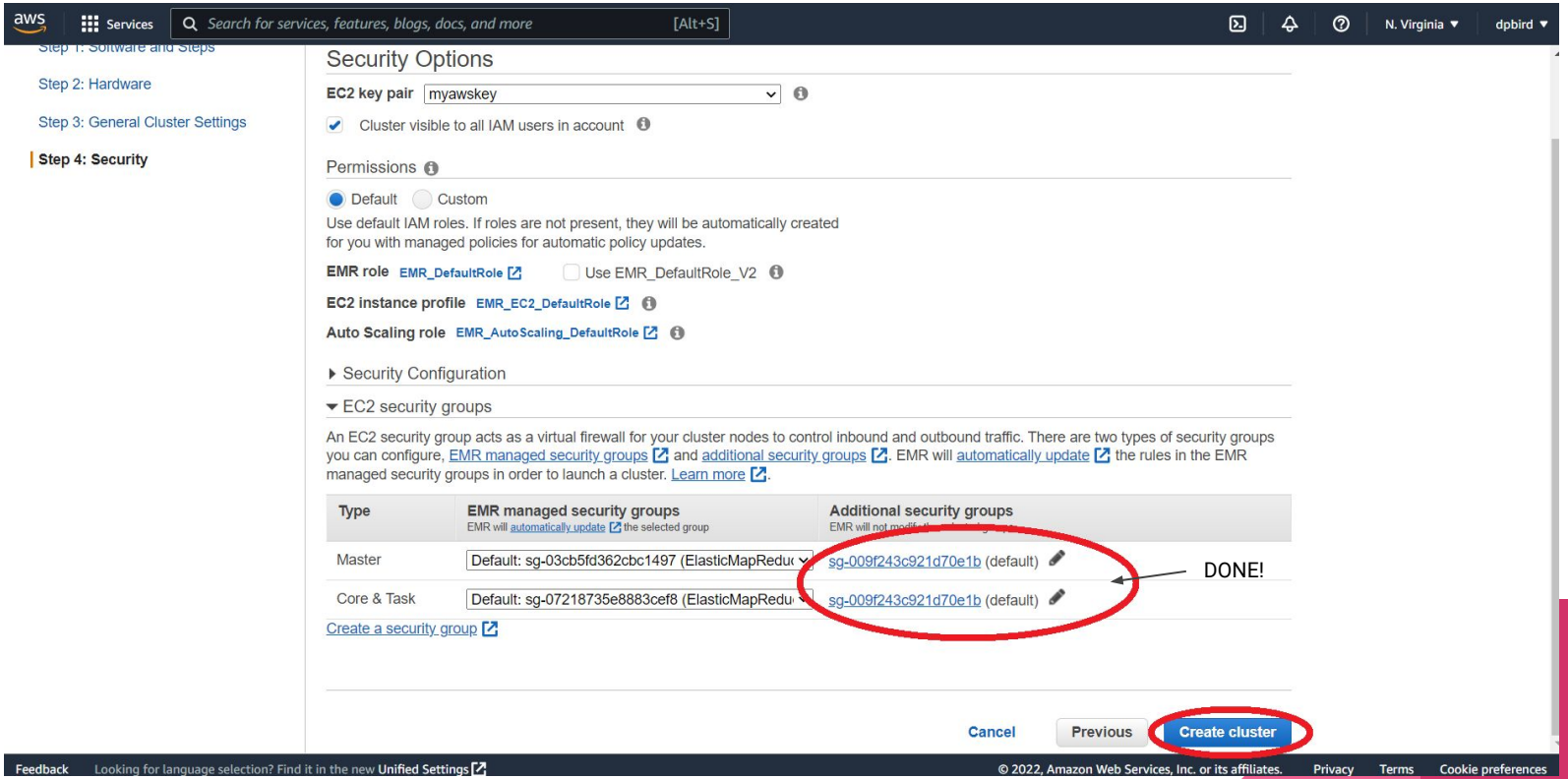
Expand the EC2 Security Groups Menu

We need to add our “default” security group to be additional security groups on both the Master and the Core & Task

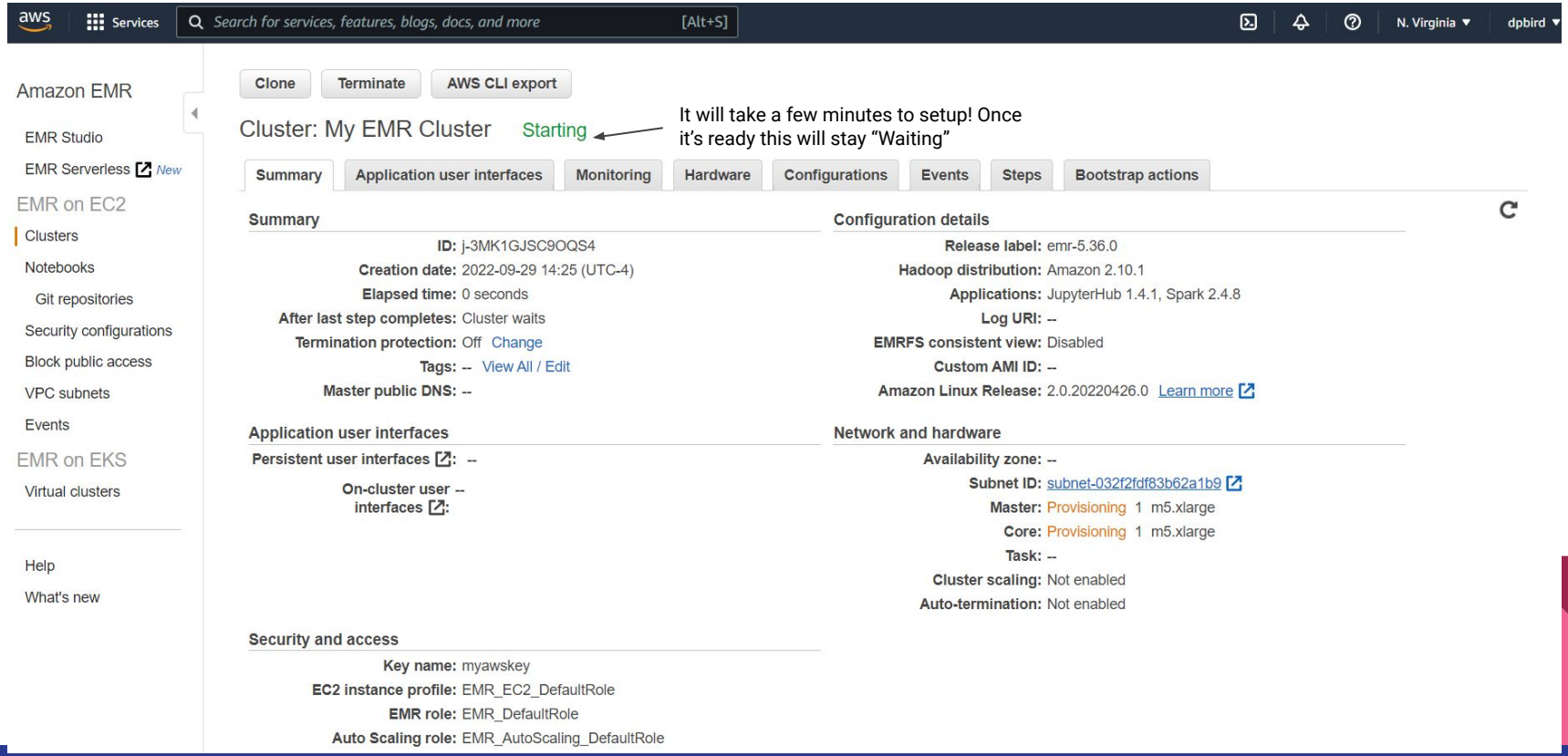
Setup EMR Cluster



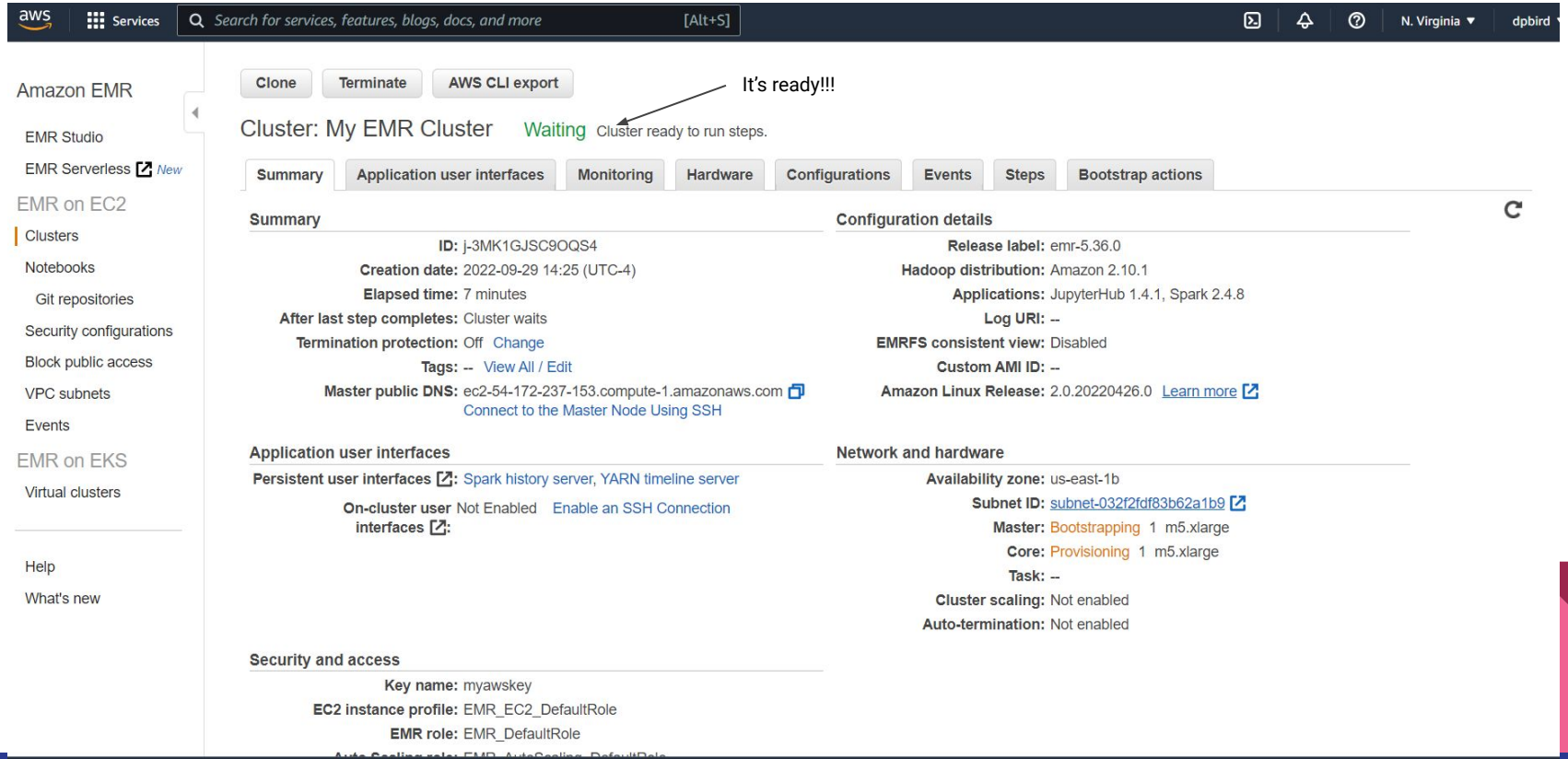
Setup EMR Cluster



Setup EMR Cluster



Setup EMR Cluster

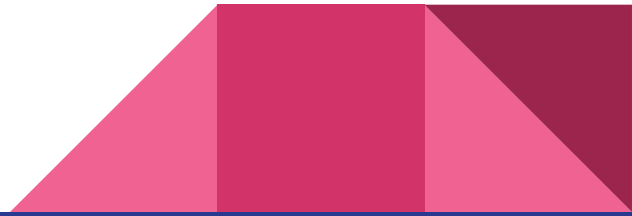


10-605 Setup for Homework Part B

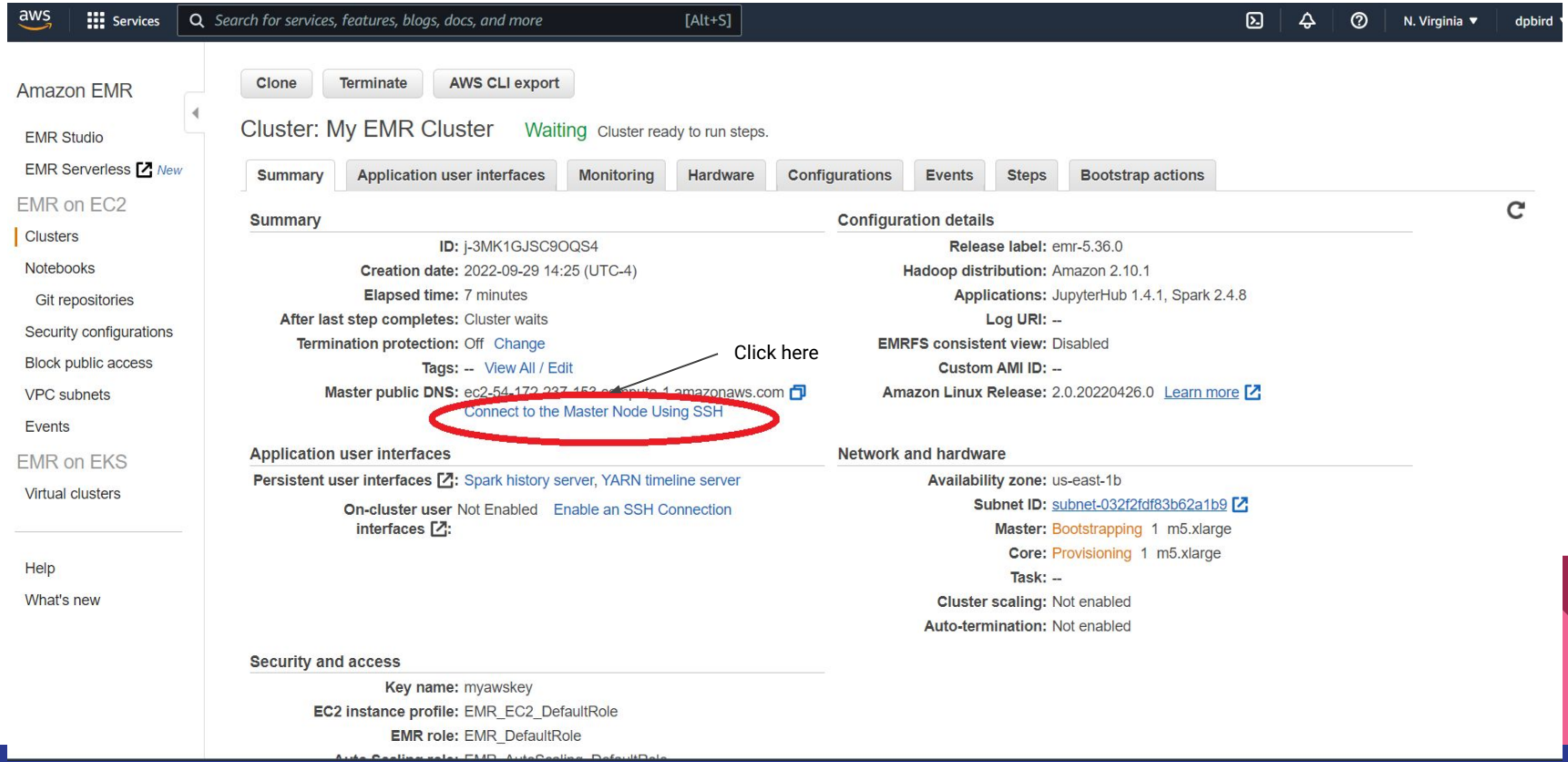
- Setup EMR Cluster - DONE
- Log into Master Node
- Get Access to Jupyter Notebooks Interface

Log into Master Node

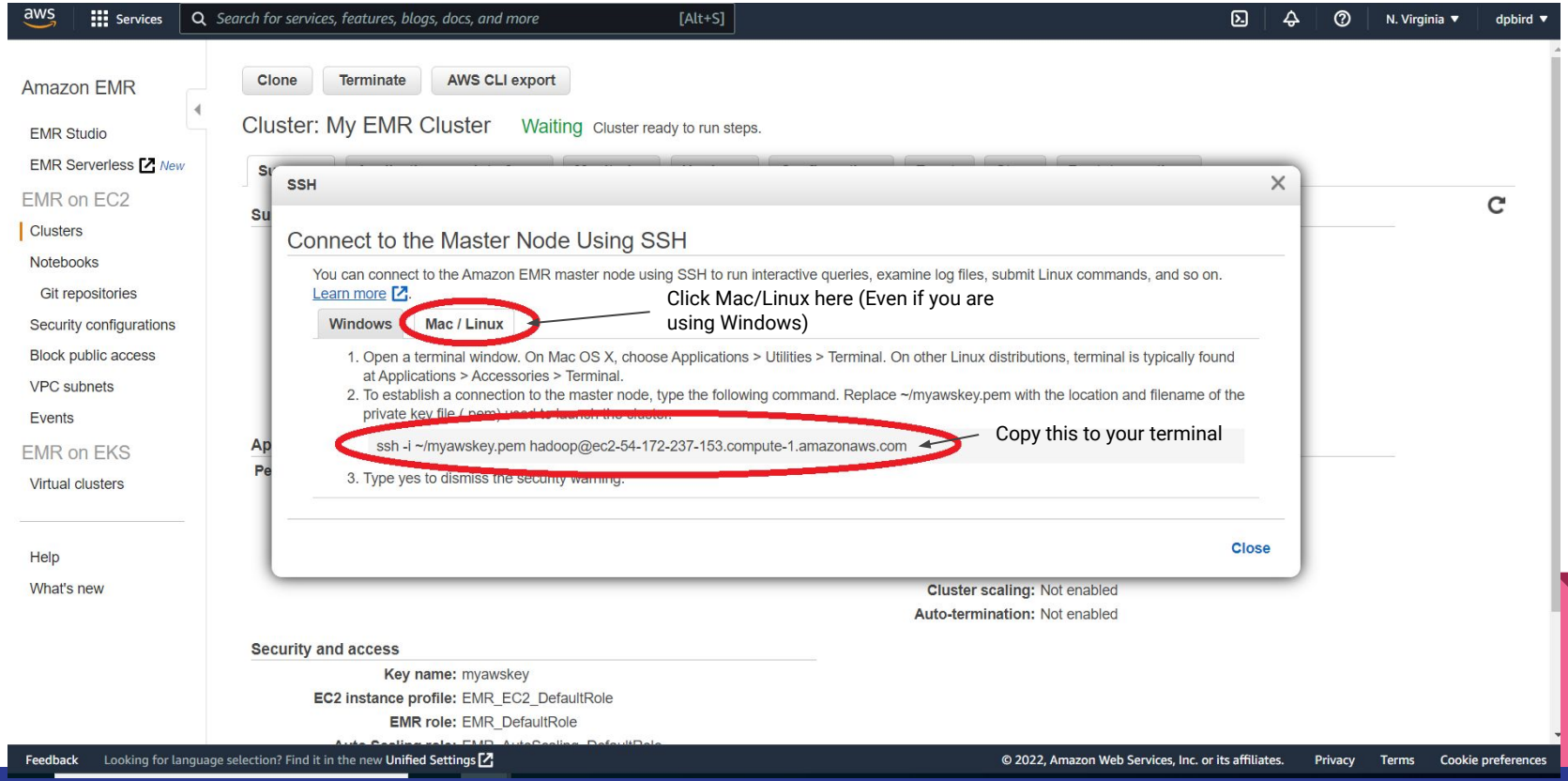
- Now the cluster is up we will log into the Master node to install some packages which we will need in the notebook.
- Open up Windows Powershell or Terminal (Mac)



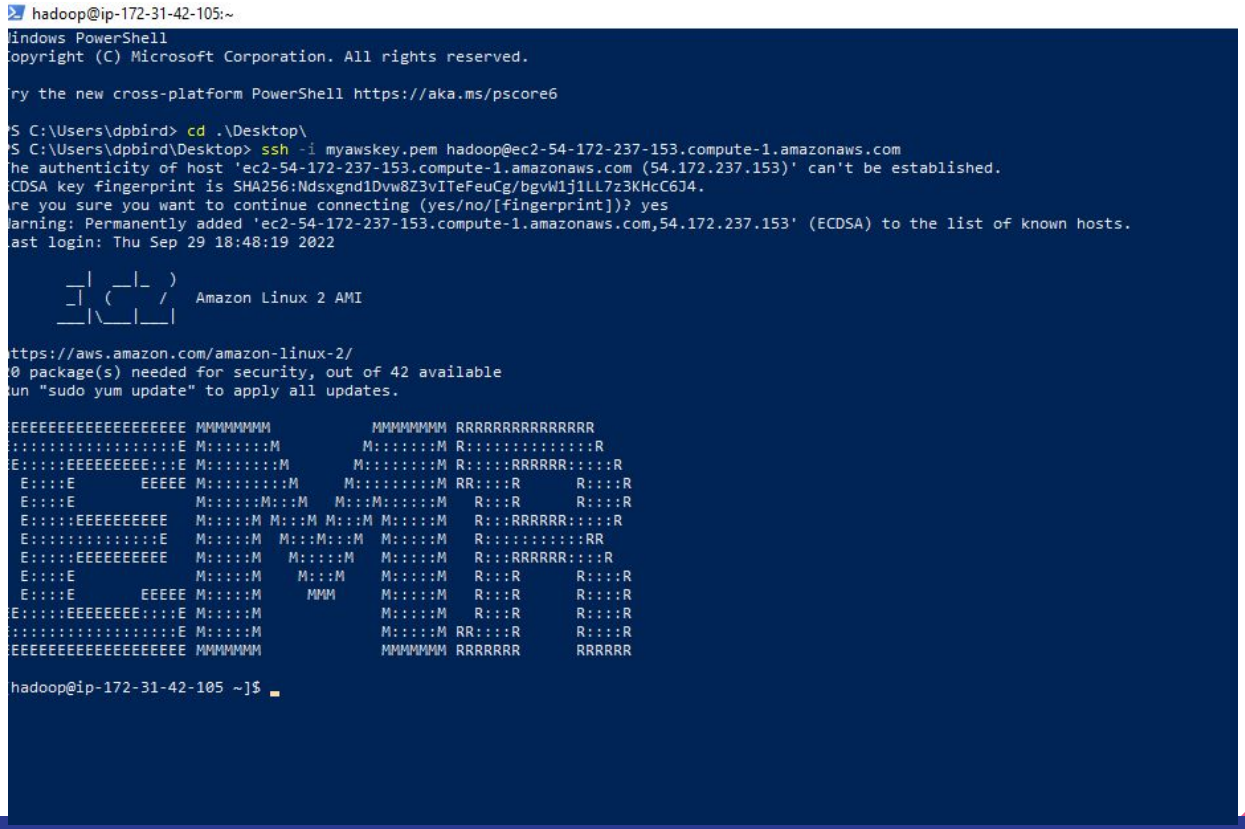
Log into Master Node



Log into Master Node

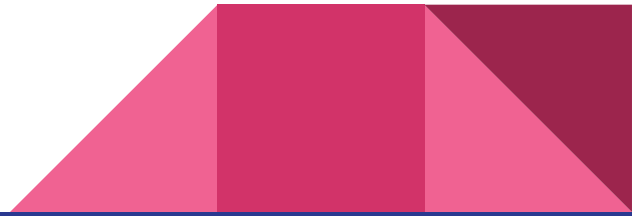


Log into Master Node



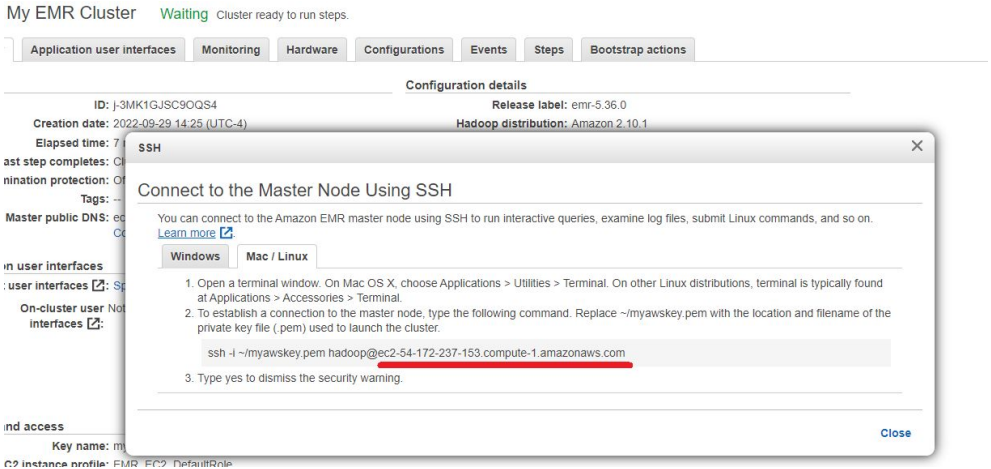
Log into Master Node

- You will want to **sudo** install matplotlib
  - sudo pip3 install matplotlib - *you need to use sudo so that the workers have access to it*
- Then you can log out.



10-605 Setup for Homework Part B

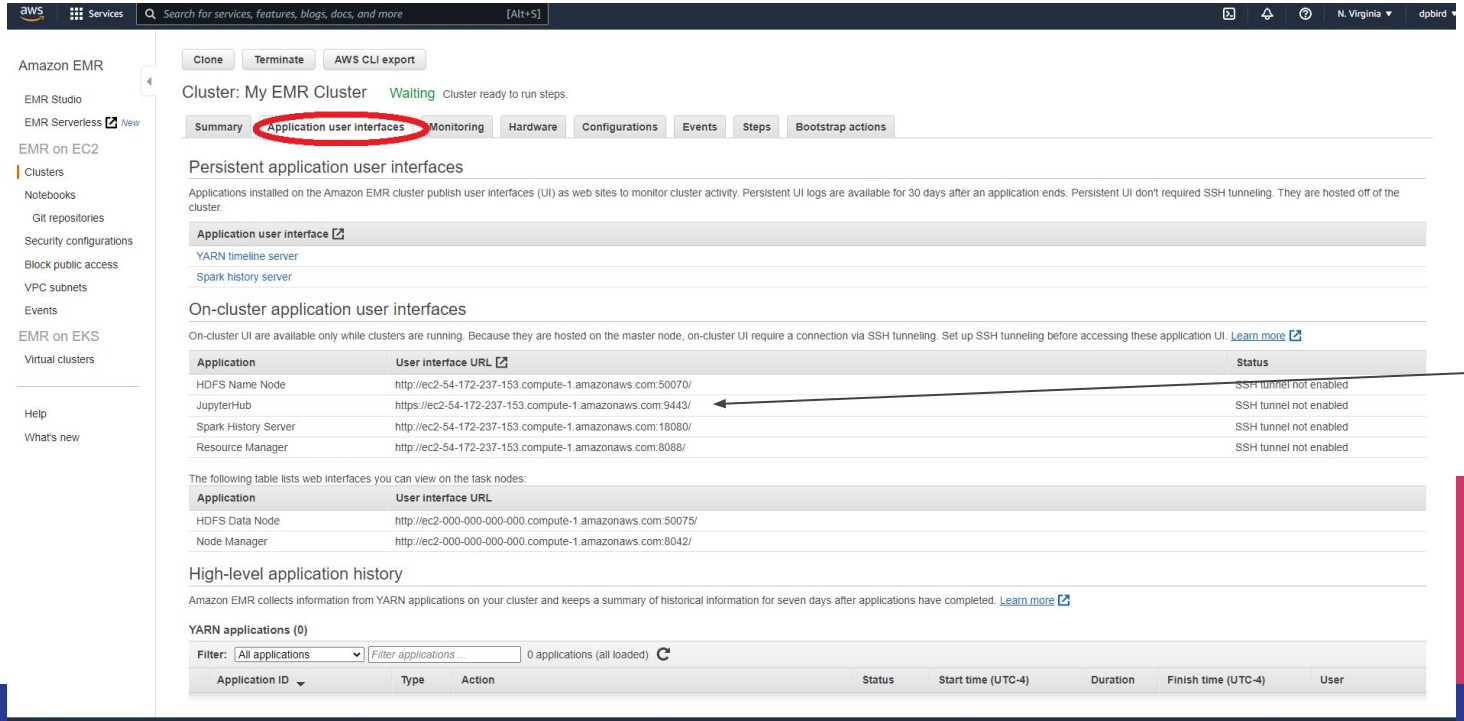
- Setup EMR Cluster - DONE
- Log into Master Node - DONE
- Get Access to Jupyter Notebooks Interface
  - Before we start this we want to grab this from the cluster:



So mine is:
ec2-54-172-237-153.compute-1.amazonaws.com

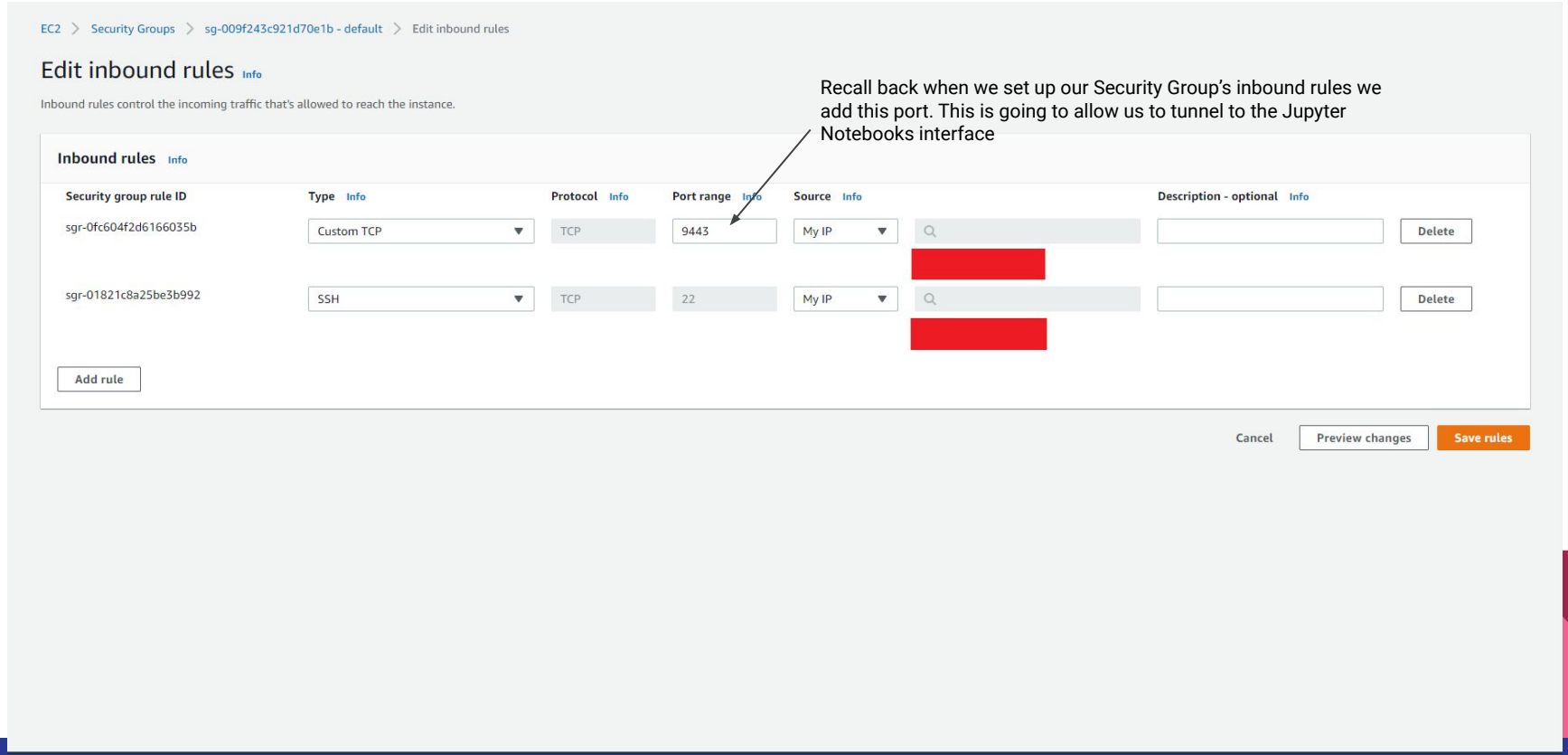
Get Access to Jupyter Notebooks Interface

- The Jupyter Notebook Interface can only be accessed via SSH tunneling.



We want to access this!

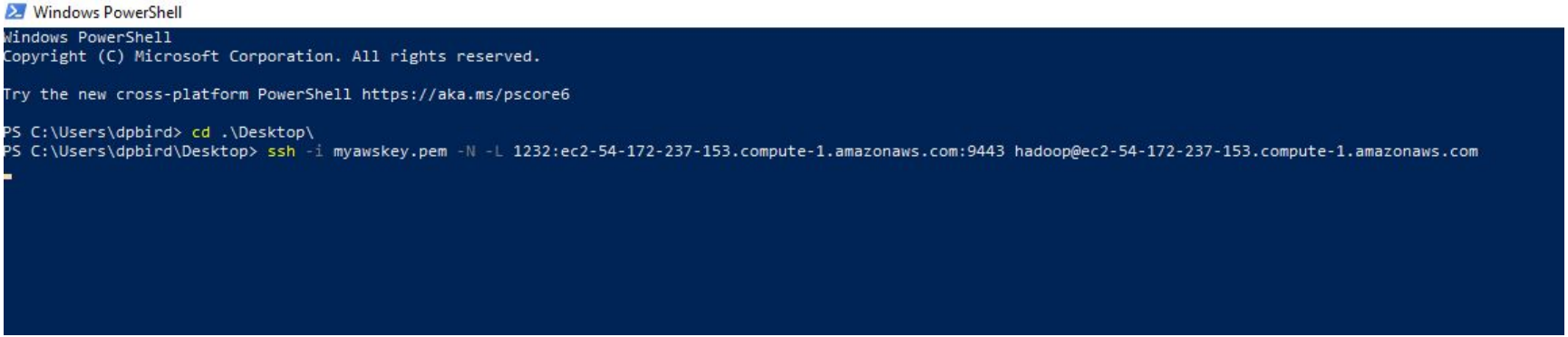
Get Access to Jupyter Notebooks Interface



Get Access to Jupyter Notebooks Interface

- Remember 3 slides ago when I told you to grab
  - ec2-54-172-237-153.compute-1.amazonaws.com
- From your terminal:
  - ssh -i myawskey.pem -N -L 1232:ec2-54-172-237-153.compute-1.amazonaws.com:9443 hadoop@ec2-54-172-237-153.compute-1.amazonaws.com
- The 1232 will be my localhost number for later, you can make this any 4 digit number.
- The 9443 is the port number of the JupyterHub

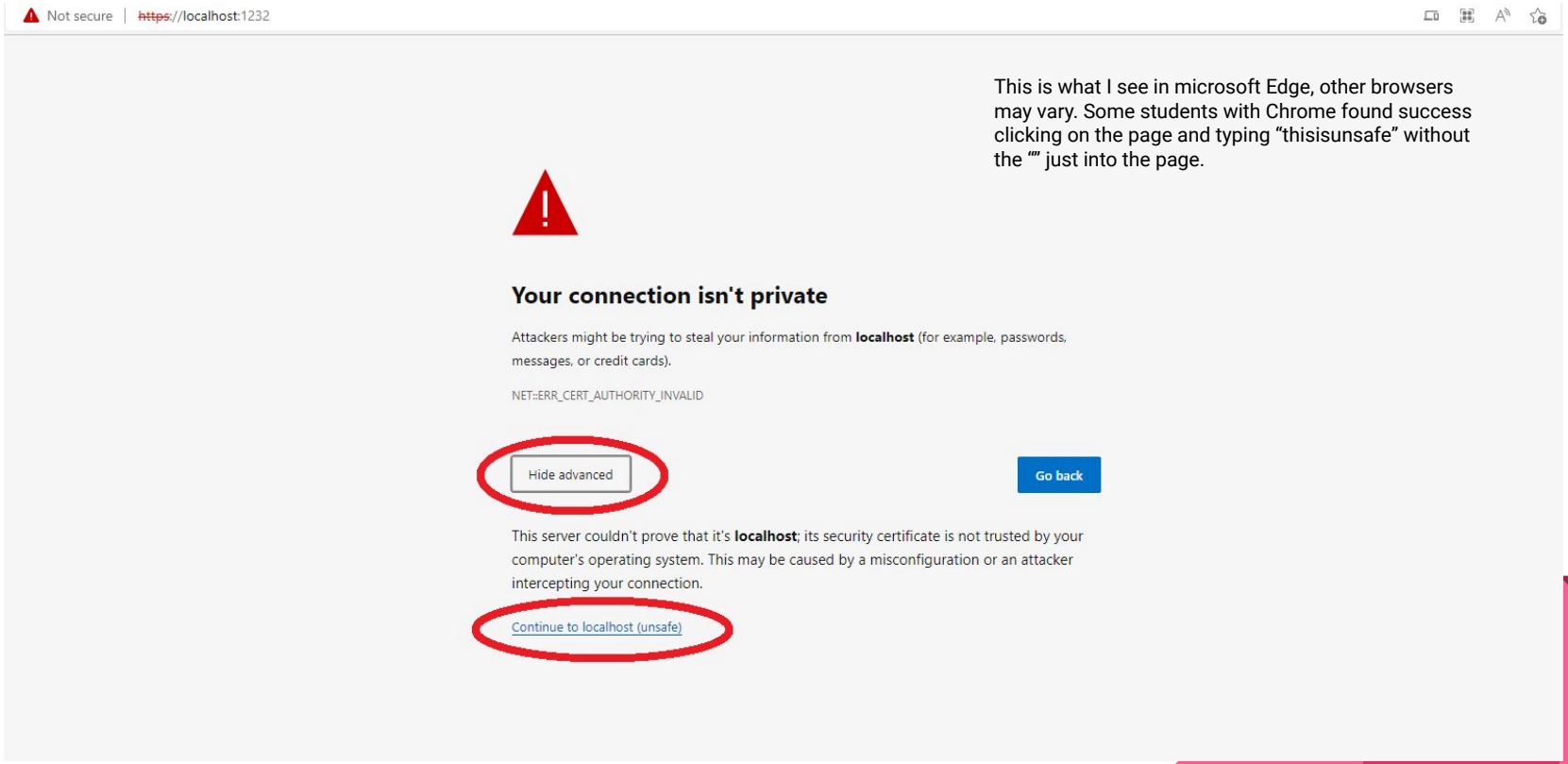
Get Access to Jupyter Notebooks Interface



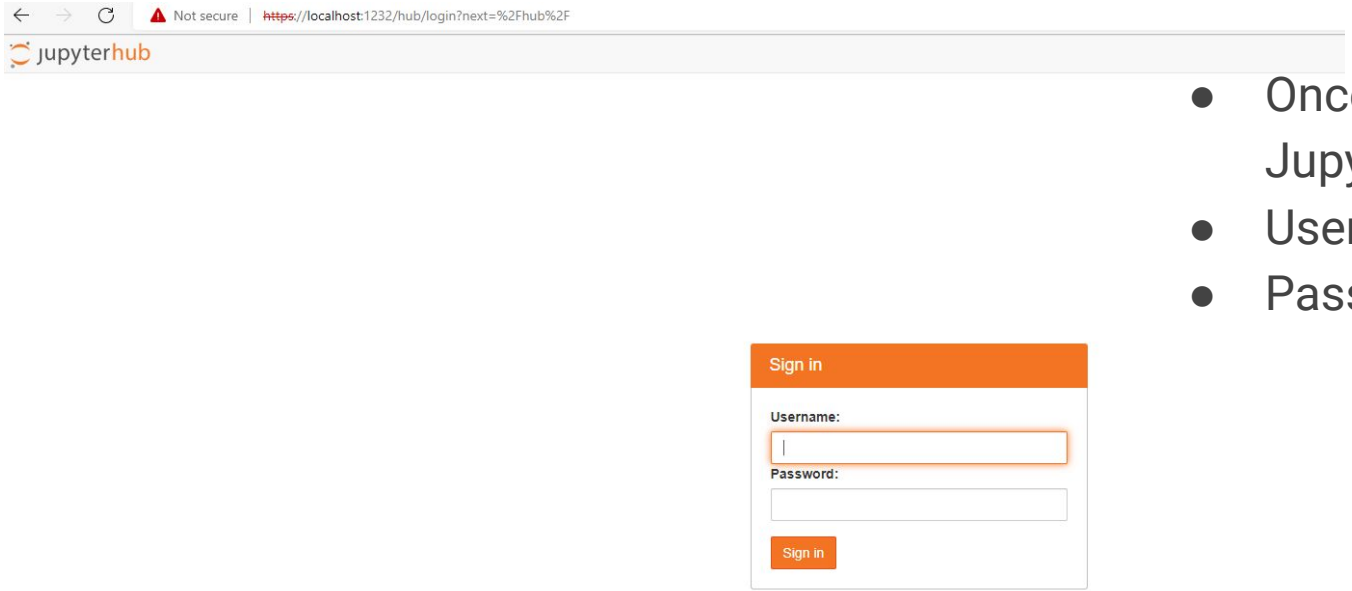
- This will appear like it has frozen, but it hasn't it is holding the tunnel for you.
- Now you can head to
  - https://localhost:1232/
- It will warn you that your connection is not private. You want to continue to localhost.

Note: Some users' computers do not automatically add the https portion, make sure you include it!

Get Access to Jupyter Notebooks Interface



Get Access to Jupyter Notebooks Interface

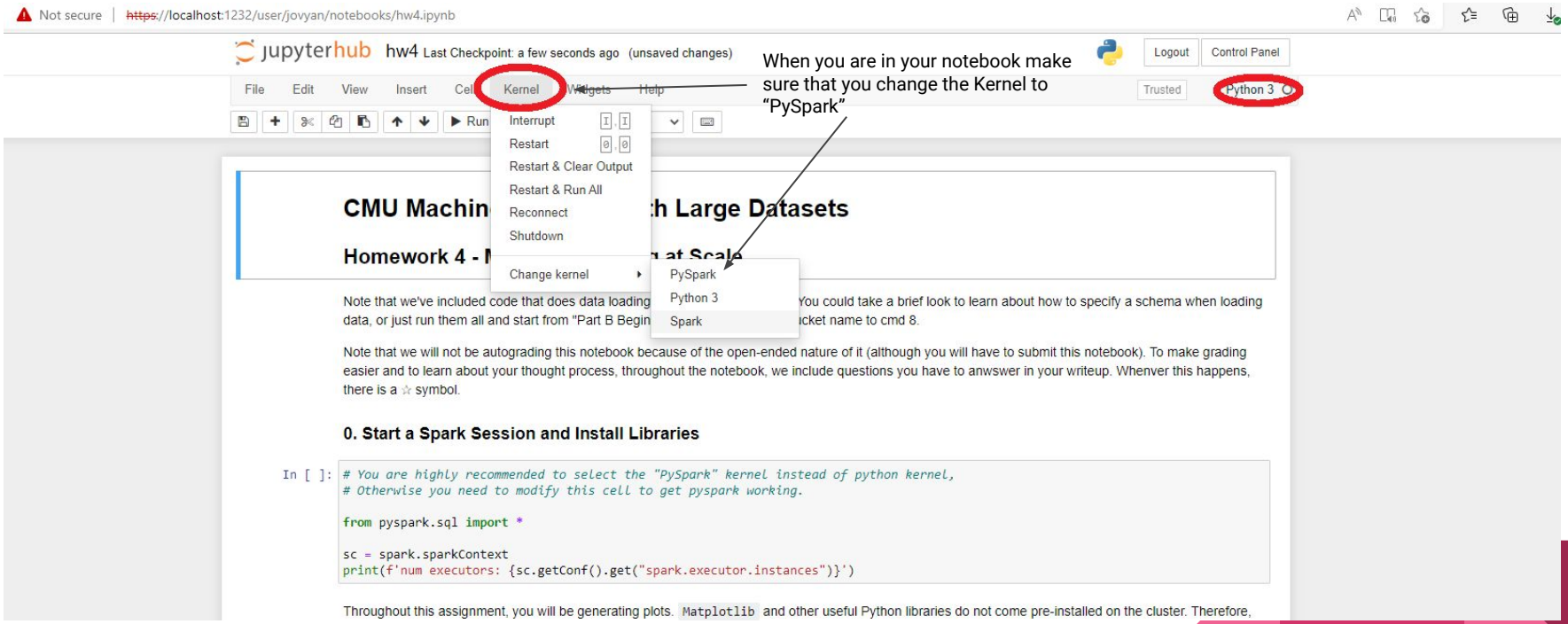


- Once through you will be in JupyterHub
- Username: jovyan
- Password: jupyter

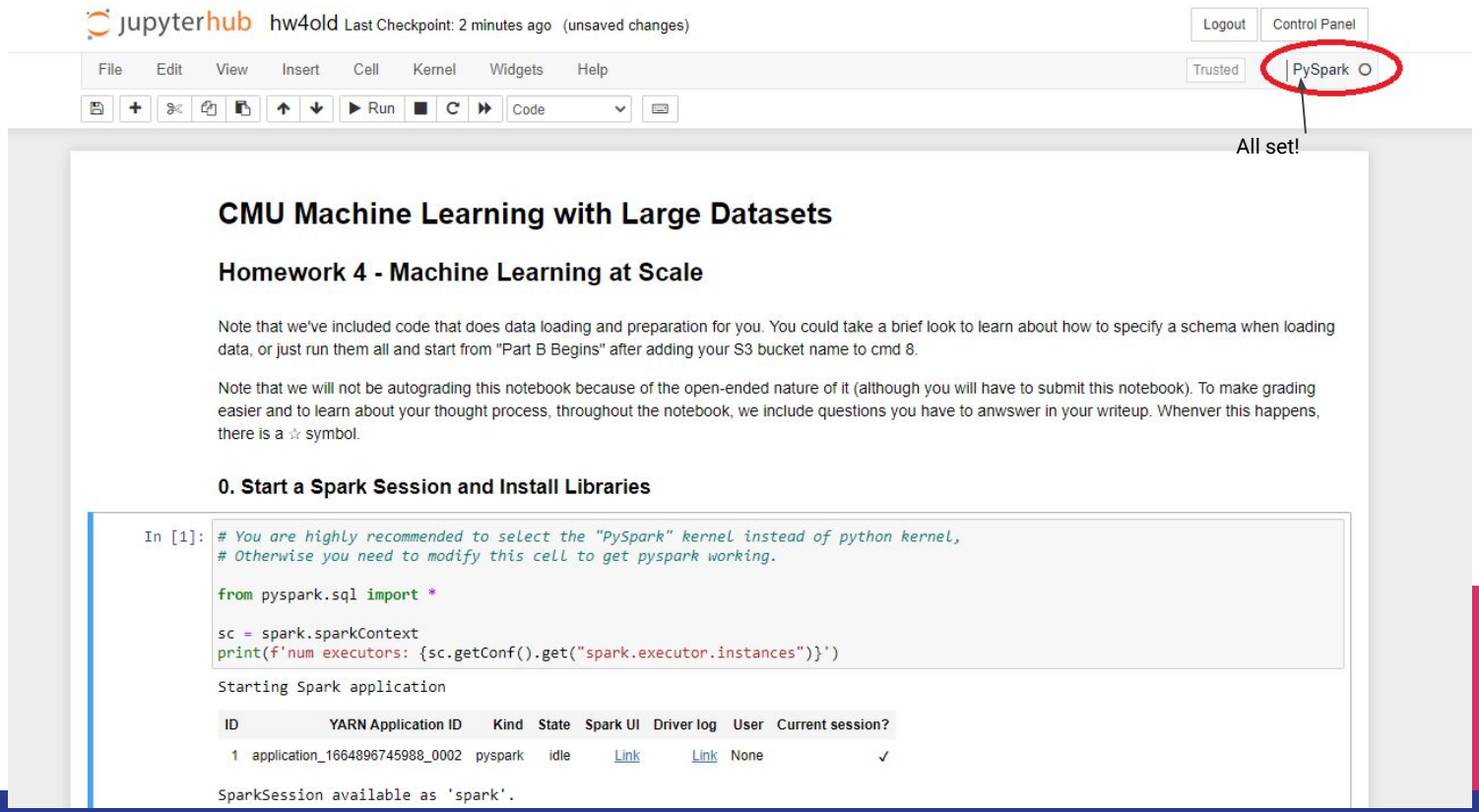
Get Access to Jupyter Notebooks Interface

- Here you can upload your hw4.ipynb handout.
- IMPORTANT: Before you close your terminal or terminate your EMR Cluster make sure you download your hw4.ipynb otherwise you will lose all changes!!!

Get Access to Jupyter Notebooks Interface



Get Access to Jupyter Notebooks Interface



Get Access to Jupyter Notebooks Interface

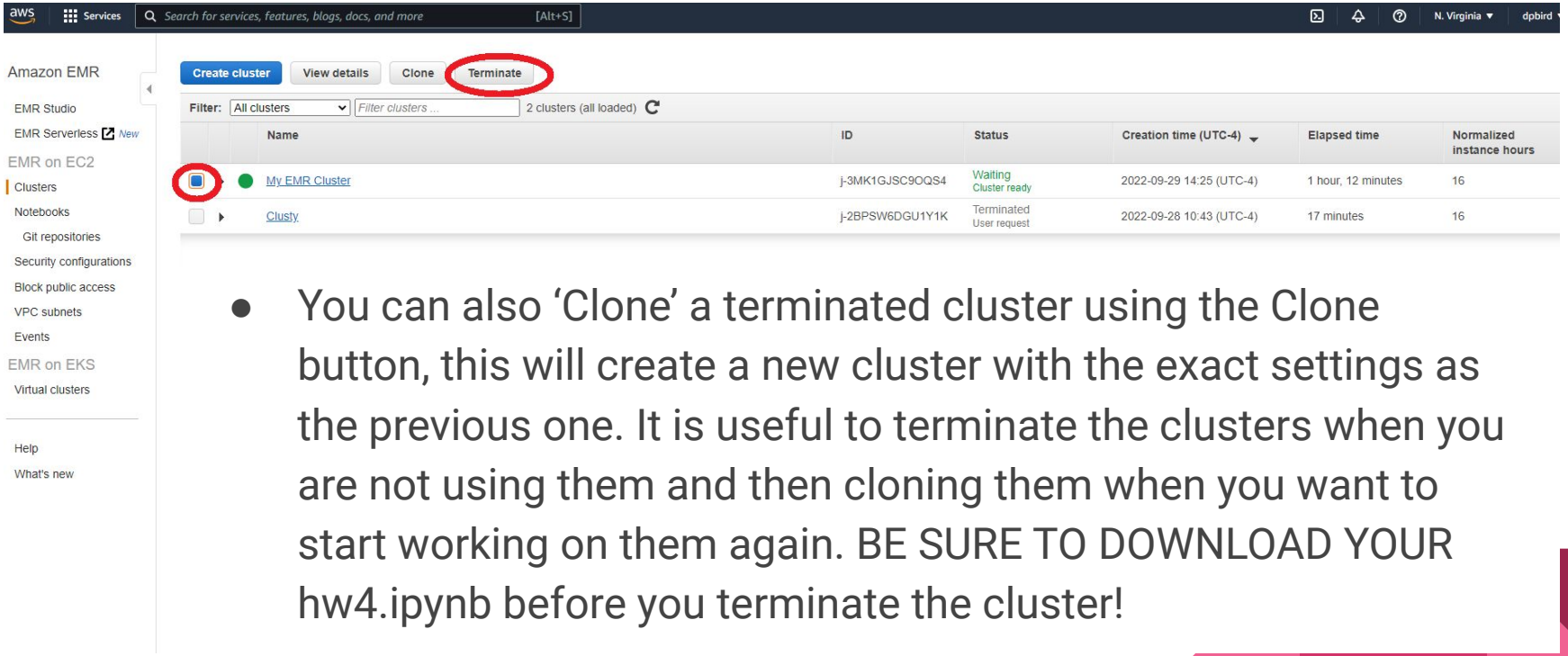
In [ ]:
```
#----------------------------------------------------------------------
#Set to False after writing your code and run end-to-end on full data
#----------------------------------------------------------------------
load_subset = True
#----------------------------
#Add your s3 bucket name below:
#----------------------------
s3_bucket_name = ''

if load_subset:
    df = spark.read.format("csv")\
        .option("header", "false")\
        .option("nanValue", "nan")\
        .option("quote", "\"")\
        .option("escape", "\"")\
        .schema(schema)\
        .load(f"s3://{s3_bucket_name}/processed/A_1.csv")
else:
    df = spark.read.format("csv")\
        .option("header", "false")\
        .option("nanValue", "nan")\
        .option("quote", "\"")\
        .option("escape", "\"")\
        .schema(schema)\
        .load(f"s3://{s3_bucket_name}/processed/*.csv")

print('loaded {} records'.format(df.count()))
```

You are going to want to add your S3 bucket name here, so that it can get the data!

10-605 Setup for Homework Part B

- Setup EMR Cluster - DONE
- Log into Master Node - DONE
- Get Access to Jupyter Notebooks Interface - DONE
- Terminating and Cloning EMR Clusters

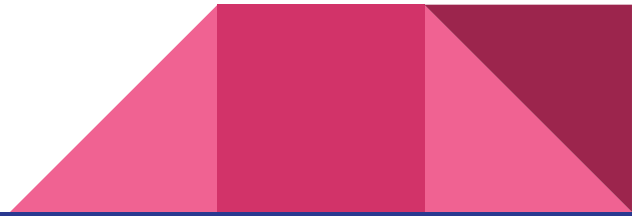
Terminating and Cloning EMR Clusters



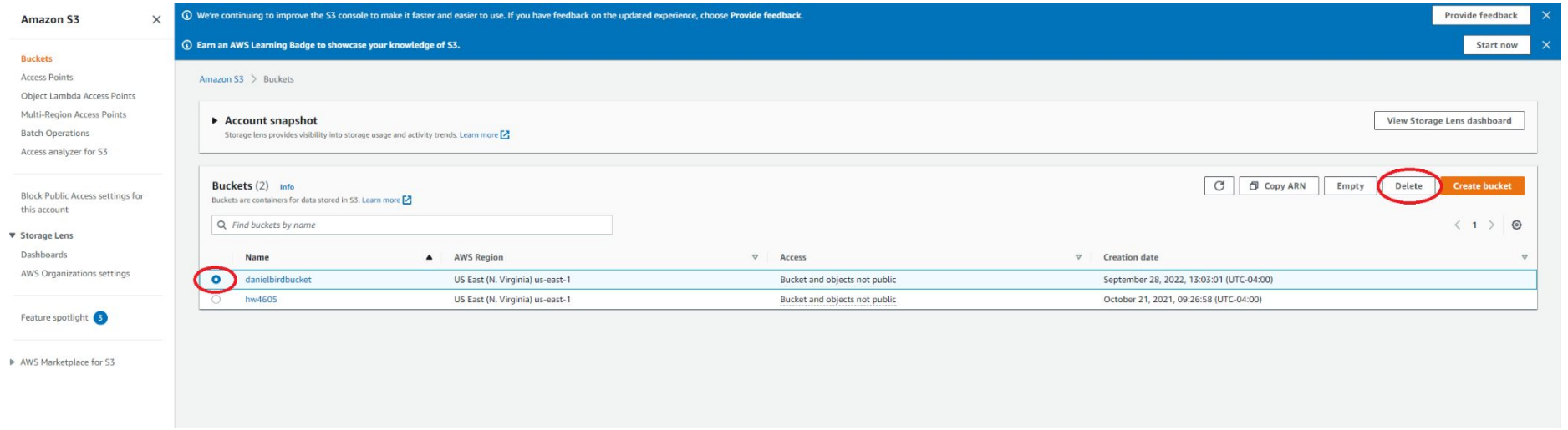
- You can also ‘Clone’ a terminated cluster using the Clone button, this will create a new cluster with the exact settings as the previous one. It is useful to terminate the clusters when you are not using them and then cloning them when you want to start working on them again. BE SURE TO DOWNLOAD YOUR hw4.ipynb before you terminate the cluster!

Finally: Deleting your S3 Buckets

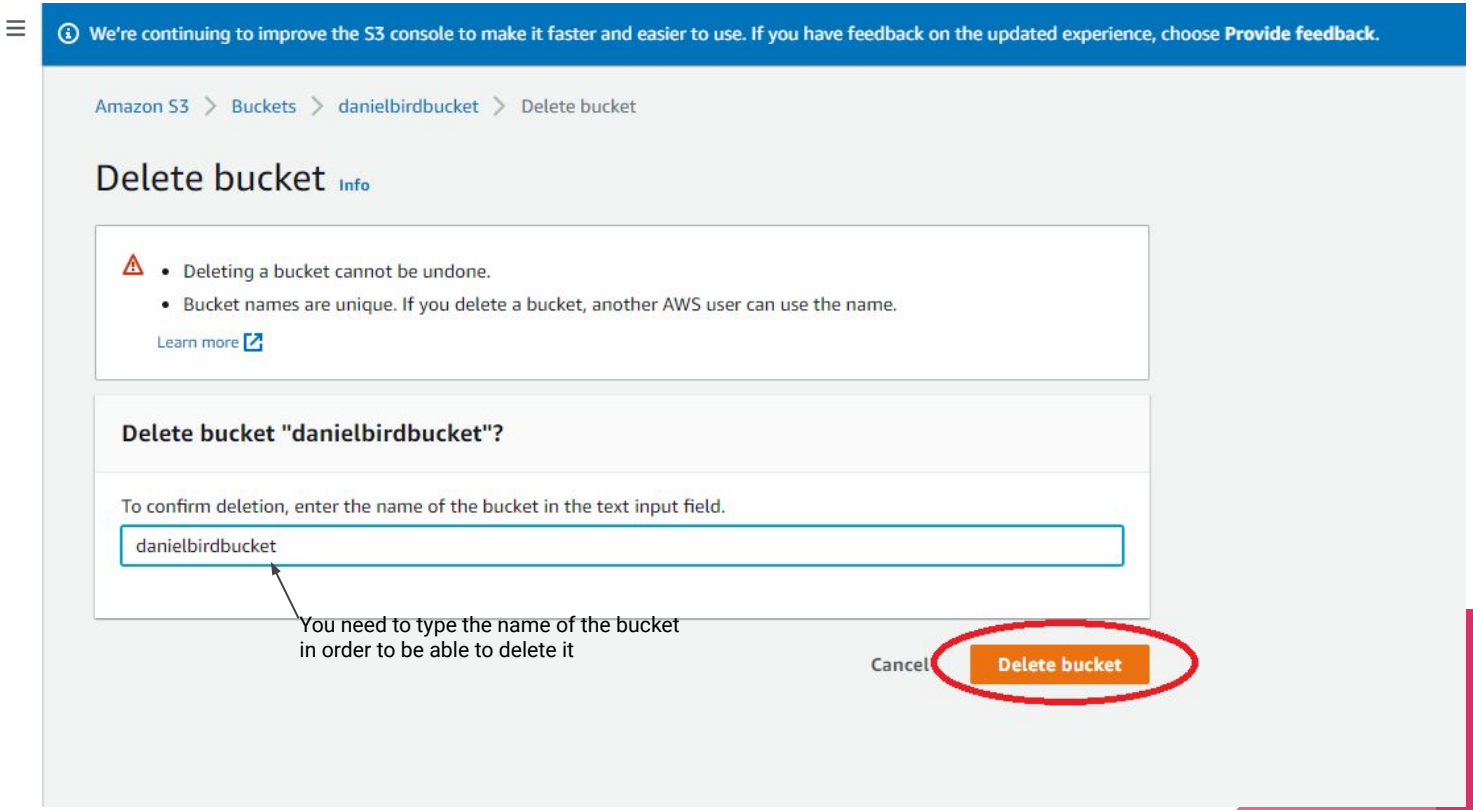
- S3 Buckets are much cheaper than the volumes we created for the EC2 instance so keeping these running the whole duration of the homework isn't costly. I recommend only deleting these once you have concluded the homework, this way you don't have to run Part A everytime you want to continue Part B.



Finally: Deleting your S3 Buckets



Finally: Deleting your S3 Buckets



Thanks for listening!