Fall 2020 Setup Guide [For Q3]

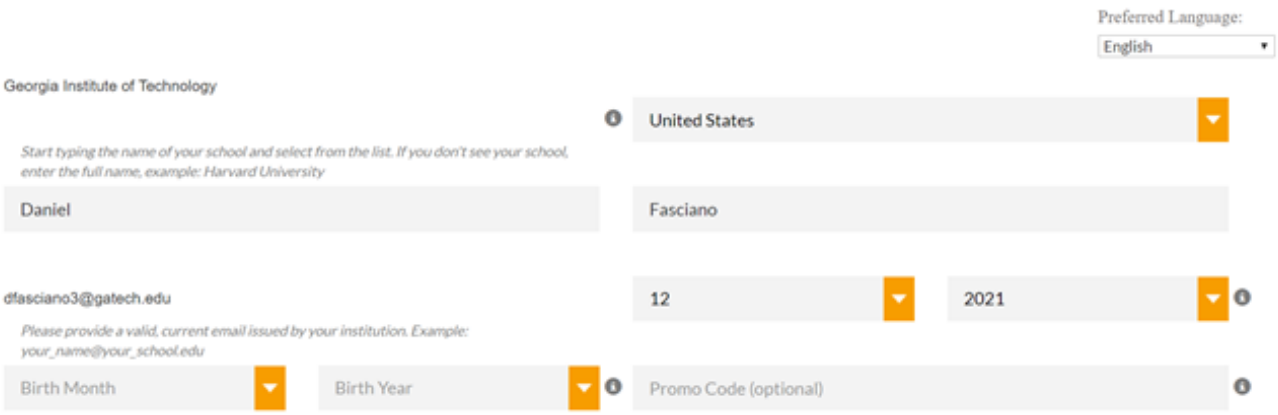
Getting Started

A video tutorial has been created to walk you through the steps in this document. You can view it here.

1. Create an AWS Educate account

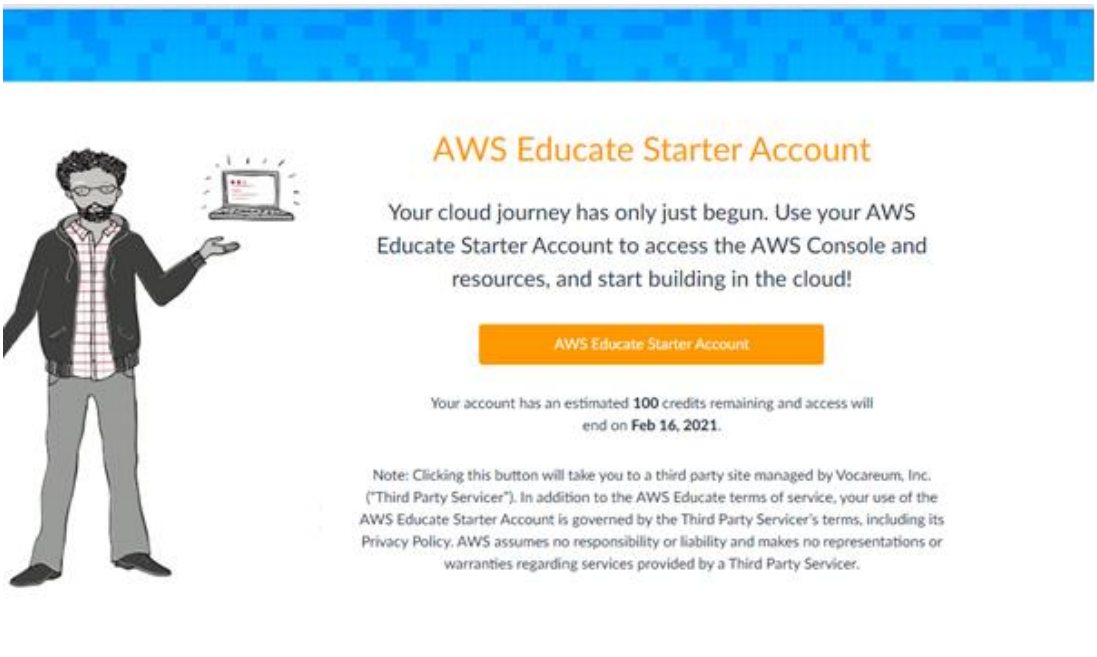
You will receive an email from support@awseducate.com inviting you to complete the **AWS Educate** application process. Your AWS Educate allows you to access EC2, Elastic MapReduce and S3 storage. Click on the link to **Join AWS Educate** in the email to proceed.

Now, fill in the requested information. Then click next.

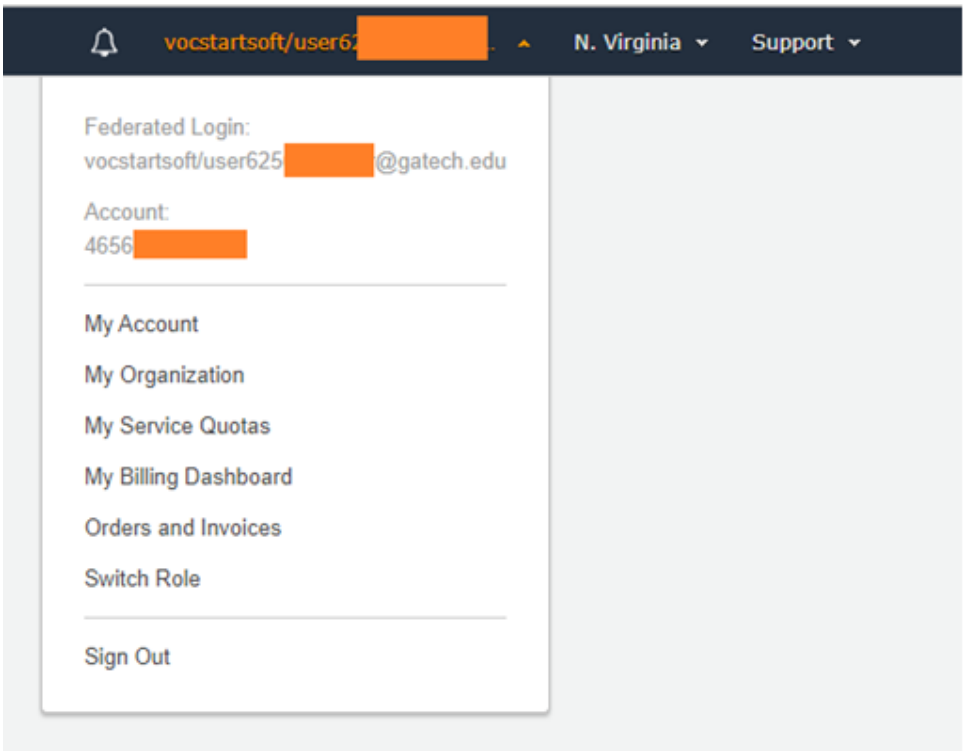


On the following screen check the box which says 'I Agree' then click Submit. You will receive an email which asks you to confirm your email, then you'll be able to log in to your account.

When you log in you will see a screen like this, so click the button to setup an AWS Educate Starter account to get your $100 of credits. This will take you to the Vocareum workbench where you can log into your account.



Once you log in, your dashboard [click AWS console] will look something like this [right top corner].



If you have any problems with this, or you receive an email from AWS saying that your application has been rejected, please contact the CSE6242 staff immediately and fill out this Google Form with your issue.

2. Set up a CloudWatch Usage Alert

NOTE: There known to be issues with setting up billing alerts in via CloudWatch in starter accounts. If you are not able to follow these steps, it is okay and you will still be able to complete the rest of the assignment, however you must be double careful to make sure to close all clusters when not in use.

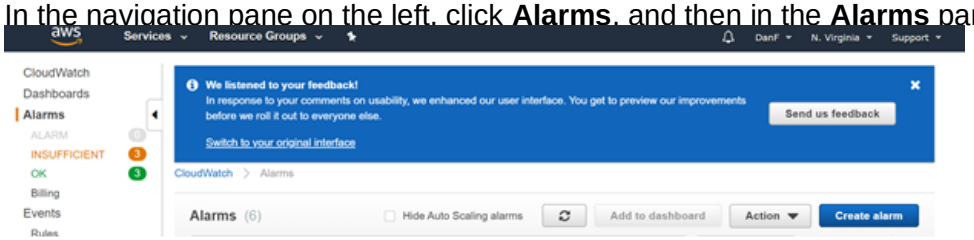
Make sure your region (in the upper right corner of the screen) is set to: **US East (N. Virginia)**. Test whether this email alert is working before scheduling in practice. That is, out of $100, when your credit balance goes below say $95, schedule a test alert and make sure it works. Remember this alert works only once. So, once you get an alert for $95, you schedule the next alert for $70 and the next one for $60 and so on.

Turn on Custom Alerts

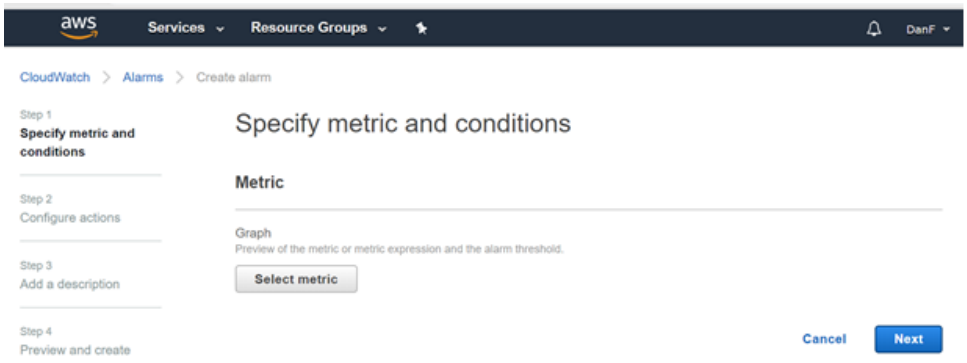
First, we need to create a custom alarm so that it tells you when you have spent money.

1. Click **CloudWatch** in the AWS Management Console.

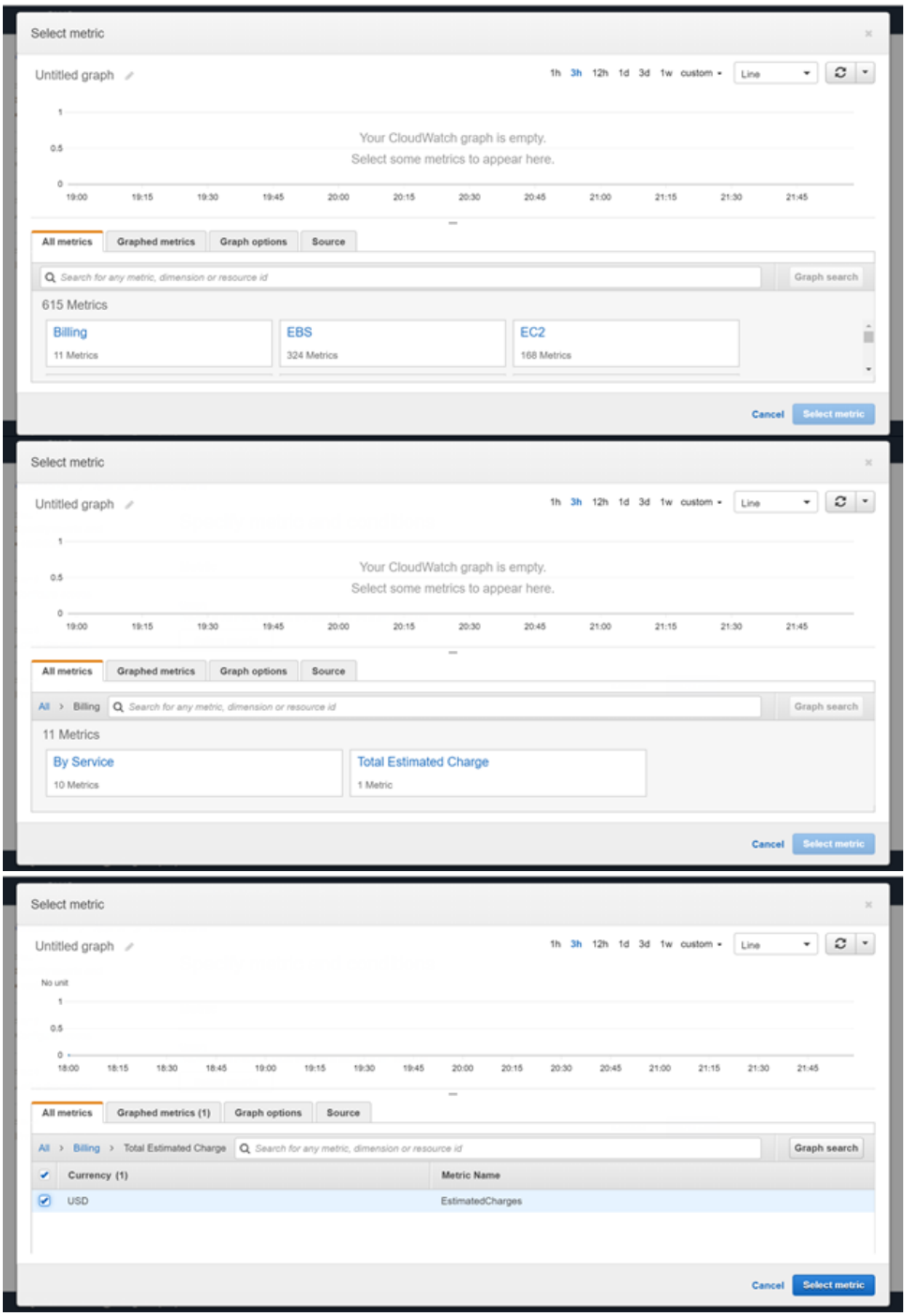
2. In the navigation pane on the left, click **Alarms**, and then in the **Alarms** pane, click **Create Alarm**.]



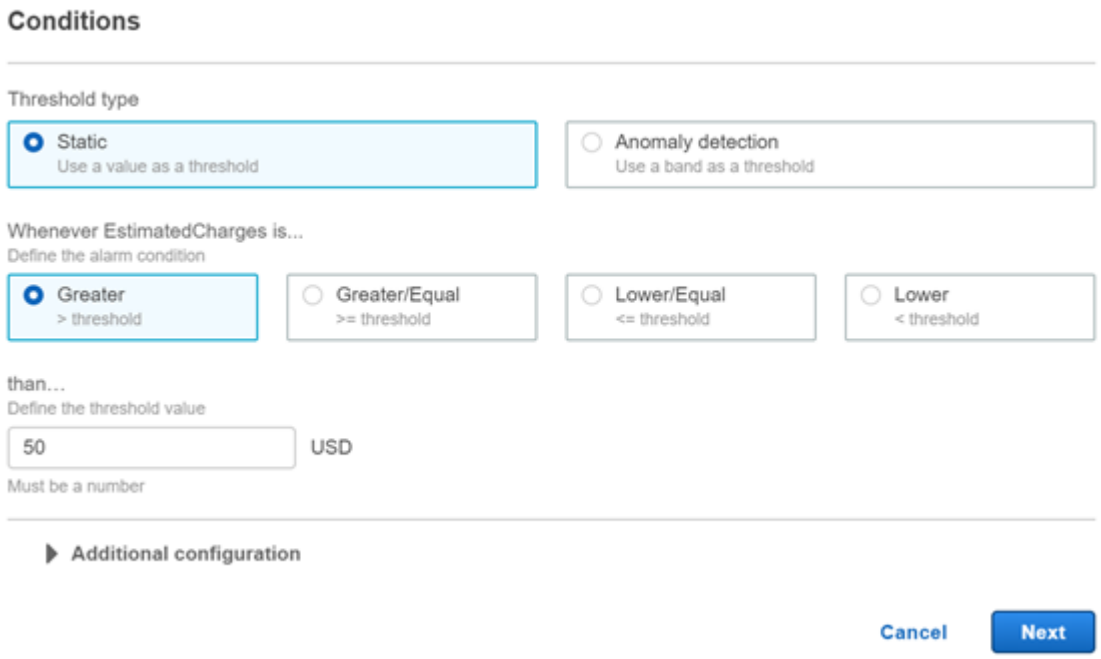
3. Click on **Select Metric**.



4. Under **All metrics**, select **Billing**, then total **Estimated Charge**. Select the checkbox, then click on **Select Metric**.

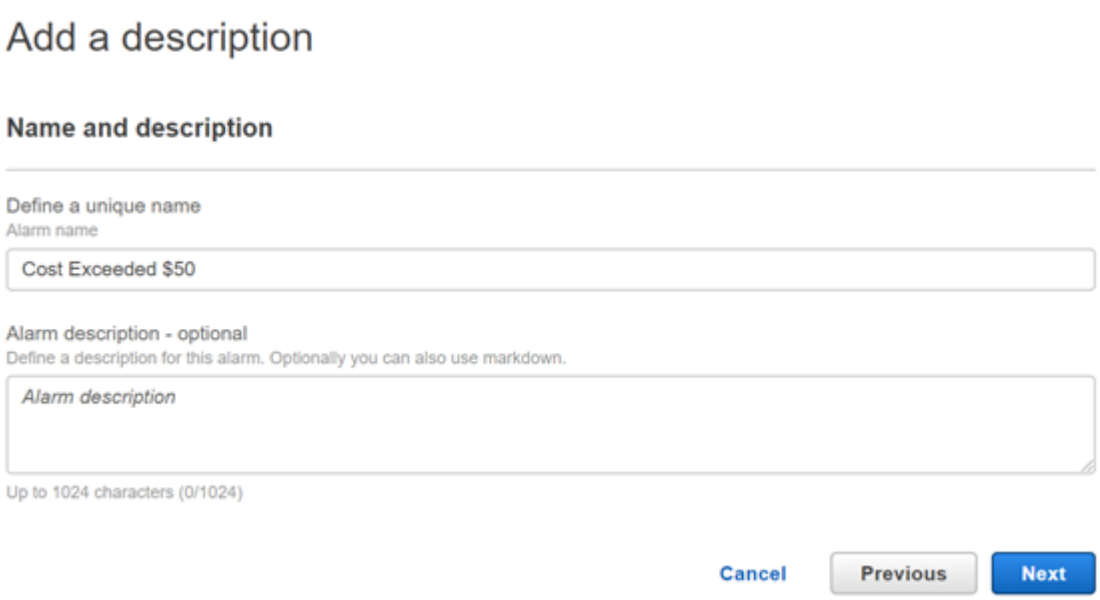


5. Set up your conditions as below, using default values, and typing 50 for the threshold value. Click next.



6. Make sure the alarm state is set to ‘in Alarm.’ Then, select Create a new topic, and enter a name and your email address, then click ’Create topic’. Scroll to the bottom of the screen and click next.

7. Enter a name for the alert and click next.



8. On this preview screen, scroll to the bottom click Create Alarm

You have now created an alert that will notify you when you have used $50. Consider creating a few additional alerts (e.g., $60, $70) so you will be well informed of your usage!

3. Create storage buckets on S3

We need S3 for two reasons:

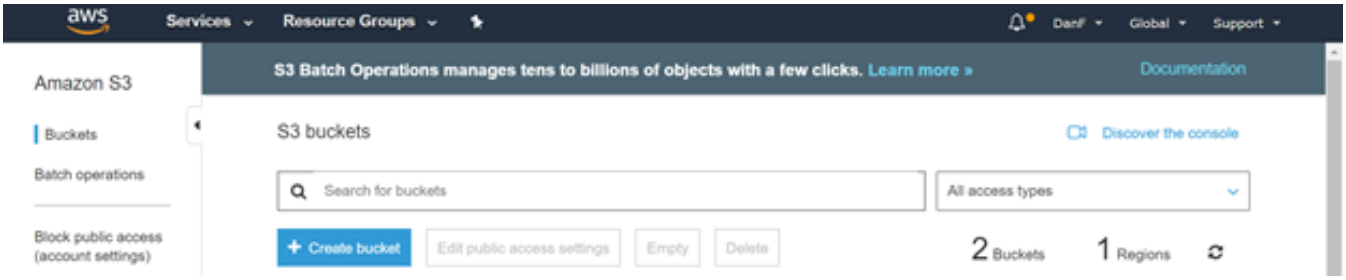
(1) An EMR (Elastic MapReduce) workflow requires the input data to be on S3.

(2) An EMR workflow output is always saved to S3.

Data (or objects) in S3 are stored in what we call "**buckets**". You can think of buckets as folders. All S3 buckets need to have unique names. You will need to create some buckets of your own to (1) store your EMR output; and (2) store your log files if you wish to debug your EMR runs. Once you have signed up, we will begin by creating the log bucket first.

1. In the AWS Management Console click on **S3** under **All services → Storage**.

   In the S3 console, click on **Create Bucket**.



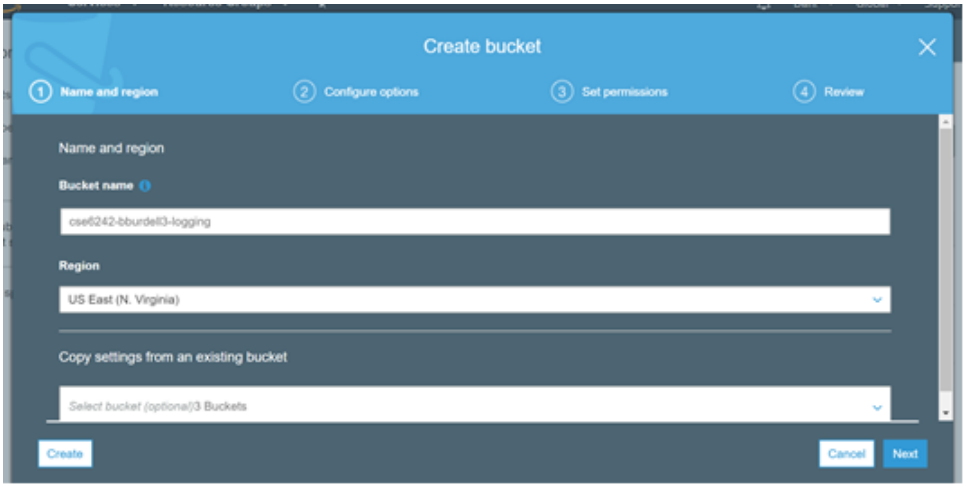
2. Create a logging bucket: Click on **Create Bucket** and enter the following details (bucket name and region) then click **Create** (not Next). Keep all other settings as the same.

   Bucket Name Format: cse6242-<GT username>-logging

   Example: cse6242-gburdell3-logging

   Region: US East (N. Virginia)

   VERY IMPORTANT: Please select **"US East (N. Virginia)**" only. If you have buckets in other regions, data transfer charges would apply.



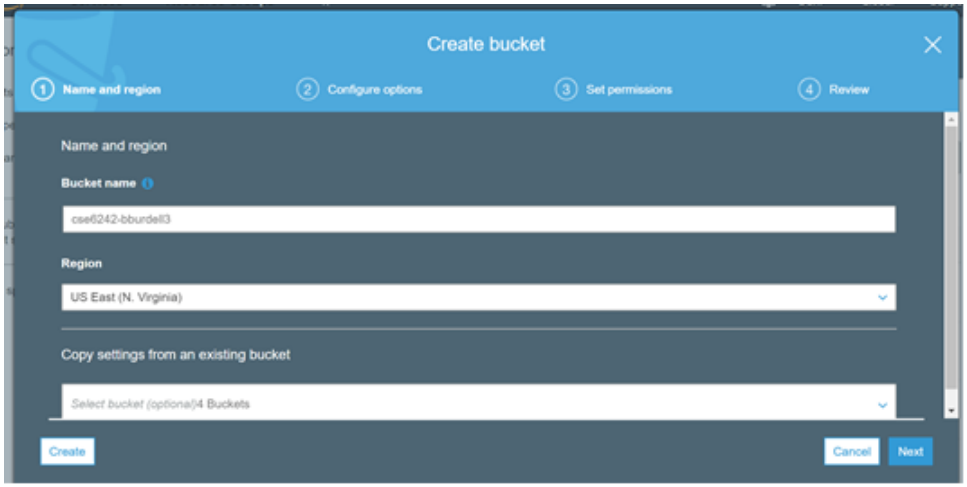
3. A new bucket will appear in the S3 console. Clicking on it will show you that it is empty.

4. Create the main bucket: Go back to the main screen (clicking on **Amazon S3**). Again, click on **Create Bucket** and enter the following details.

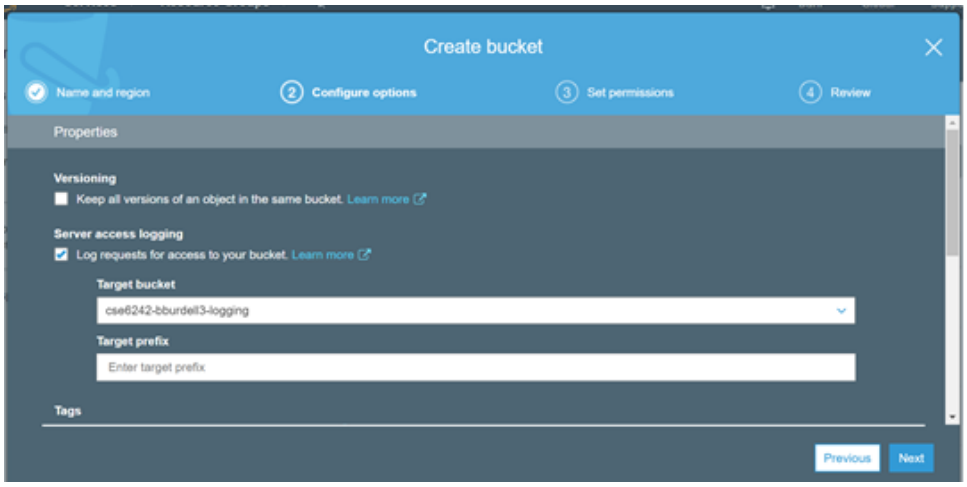
   Bucket Name Format: cse6242-<GT username>

   Example: cse6242-gburdell3

   Region: US East (N. Virginia)



5. Since we will link this bucket to our logging bucket, the regions for the two buckets should be the same. We will link our logging bucket to the one we are creating now, so click on **Next** and then select the check box for **Server access logging**.



6. Click on **Target Bucket and** select the 'logging bucket' you created. (e.g.: cse6242-gburdell3-logging) in the drop-down menu.

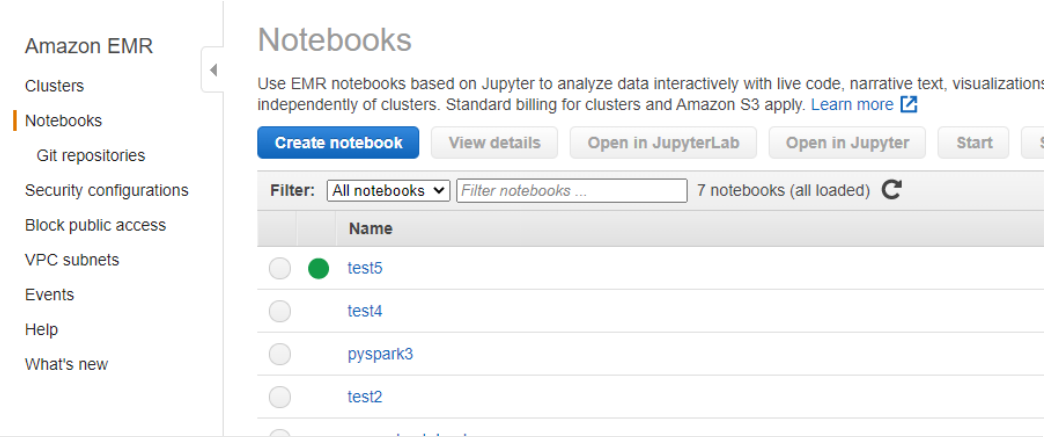
7. Keeping the settings as they are in the next 2 tabs, Hit **Next** twice and press **Create Bucket**.

We are done with creating S3 buckets at this point.

4. Launch a Notebook

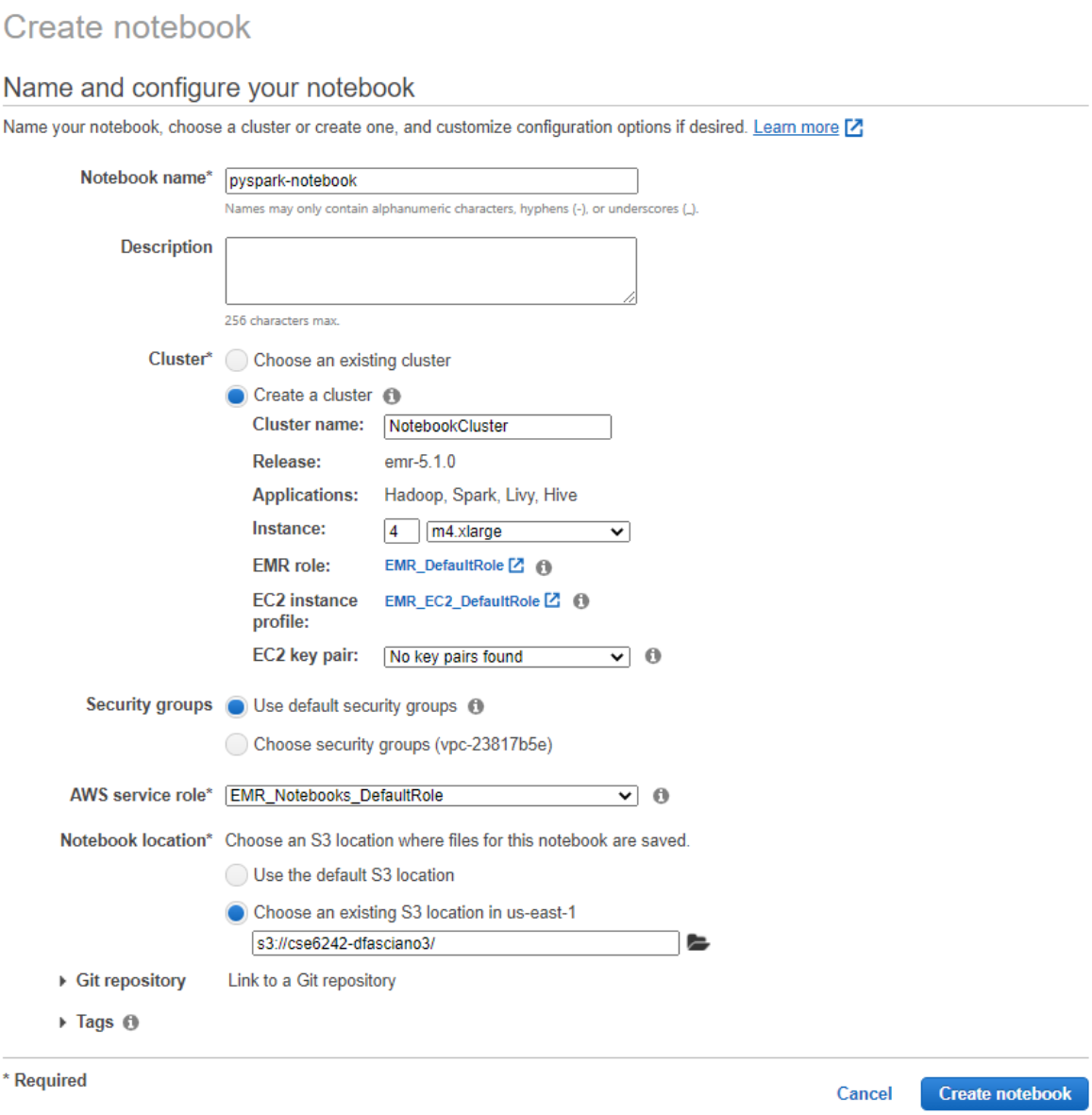
This section will cover launching a Notebook in Amazon EMR. For further information about notebooks in EMR, click here.

1. Go to Amazon EMR. Select Notebooks on the right menu. Click “Create Notebook”.



2. Make sure the region specified in the top-right corner of the page is **N. Virginia.** Otherwise click on it and from the drop-down choose N. Virginia.

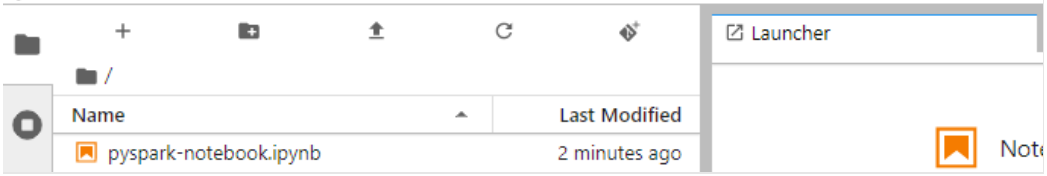
3. We will now fill out the various configuration fields to create a new Notebook:
    a. Give your notebook a name. It can be anything you want.
    b. Select the checkbox to “Create a cluster.”
    c. For instance type, select **m5.xlarge** (This will likely be the default)**.** You can also change the number of instances used, so select **4**. You can experiment with other instance types and numbers of clusters to see the impact on performance, but there are many which are not eligible to be used on a starter account, so they may result in errors when attempting to create a notebook.
    d. For AWS service role, select **EMR_Notebooks_DefaultRole**. If this is your first time running EMR, it may also give you the option to “Create Default Role”, which you should do in this case.
    e. For Notebook location, select the s3 bucket (eg: s3://cse6242-gburdell3) you created earlier.
    f. Your settings should look something like this. Once you have confirmed this, select “Create Notebook”.



5. Get started with the skeleton

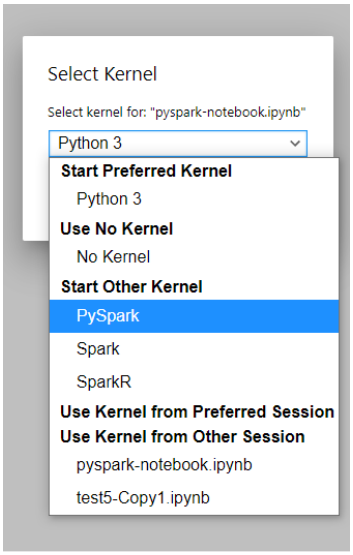
In this section we will upload the skeleton file to the notebook and run our first cell.

1. Once your notebook has finished instantiating and has the status of ‘Ready’, (this will take several minutes), click “Open in JupyterLab”.
2. In the left bar, click the arrow with a line under it to upload a file and upload the pyspark.ipynb file provided in the skeleton.



3. Double click on the newly added file to open it.

4. In the screen that gives you the option to Select a kernel, choose PySpark. If this pop up does not appear, select the Kernel in the top right of the screen to cause this pop up to appear.

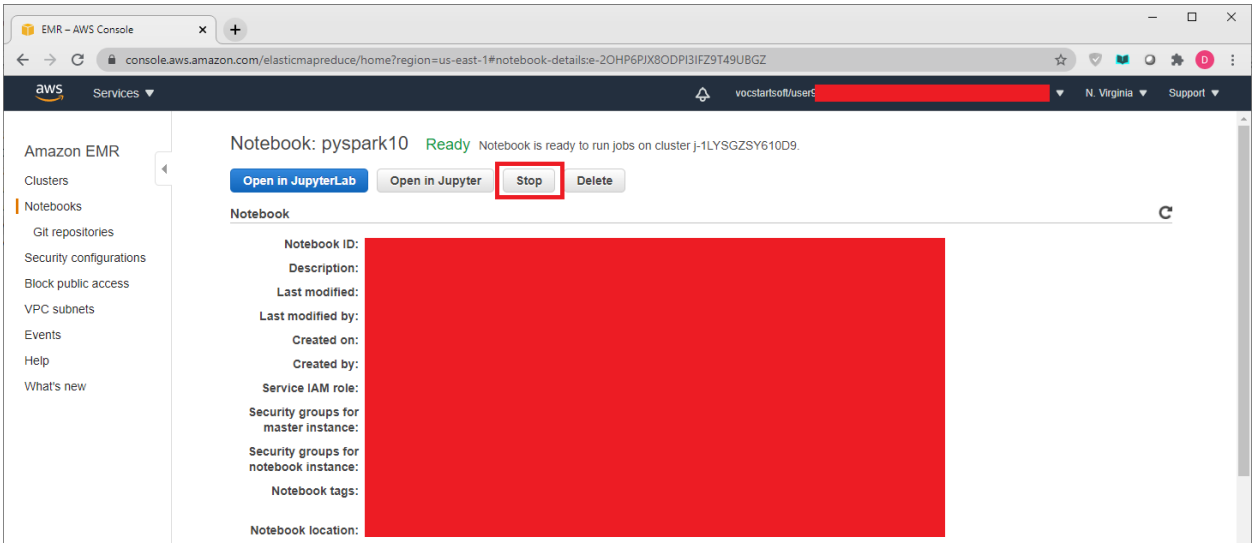


5. Run the first code cell, which should contain sc to start the Spark Application so you can start programming the assignment.
6. Once you have finished coding, right click on the file name in the directory on the left and select download to download it. It will also be saved in your S3 bucket,

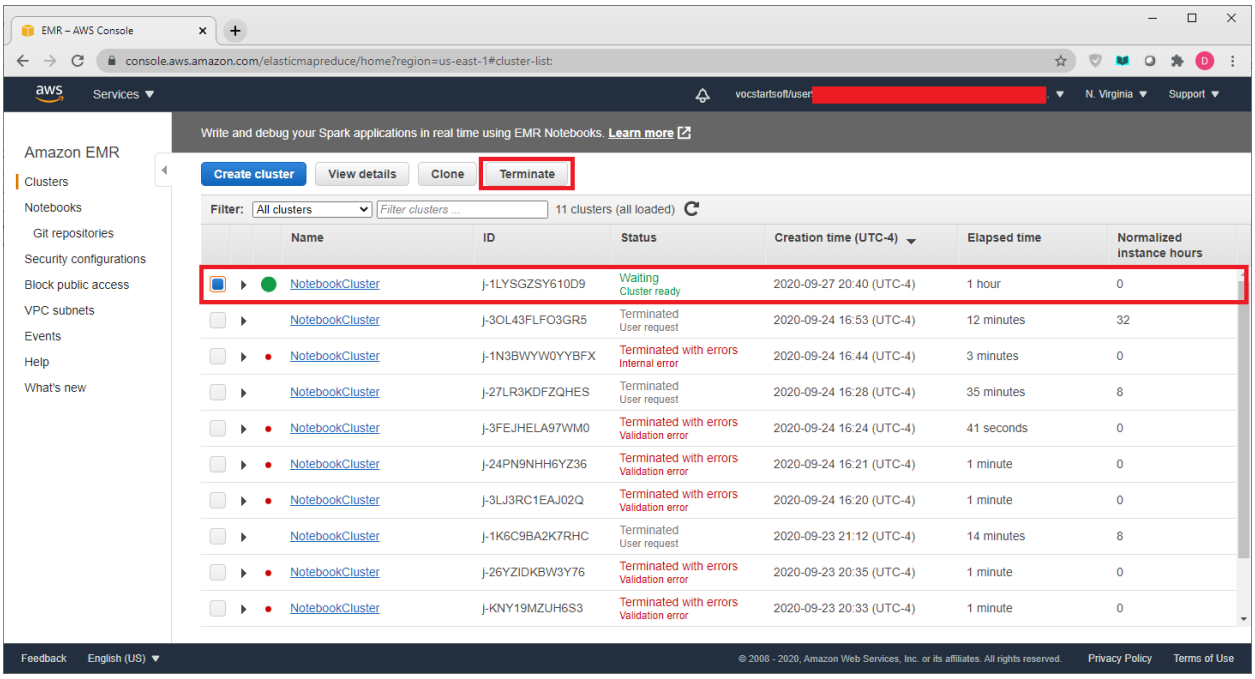
6. Terminating All Clusters

WARNING: It is very important that you do not leave clusters running when not working on your workbook. Costs can go up quickly and use up your credits.

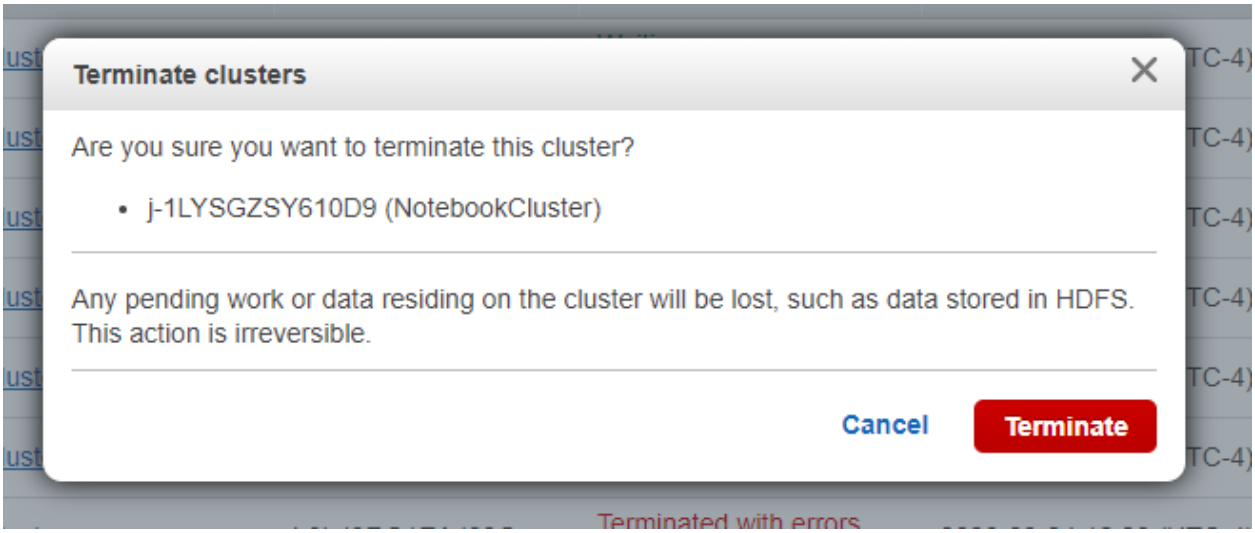
1. Back on your Notebook's page in EMR, click 'Stop'.



2. Now click on “Clusters” in the side bar on the left. Click the check box next to your running cluster (the one with the green circle) and click “Terminate”.



3. In the popup, select ‘Terminate’.



You have now closed all your clusters and will no longer be accruing charges!