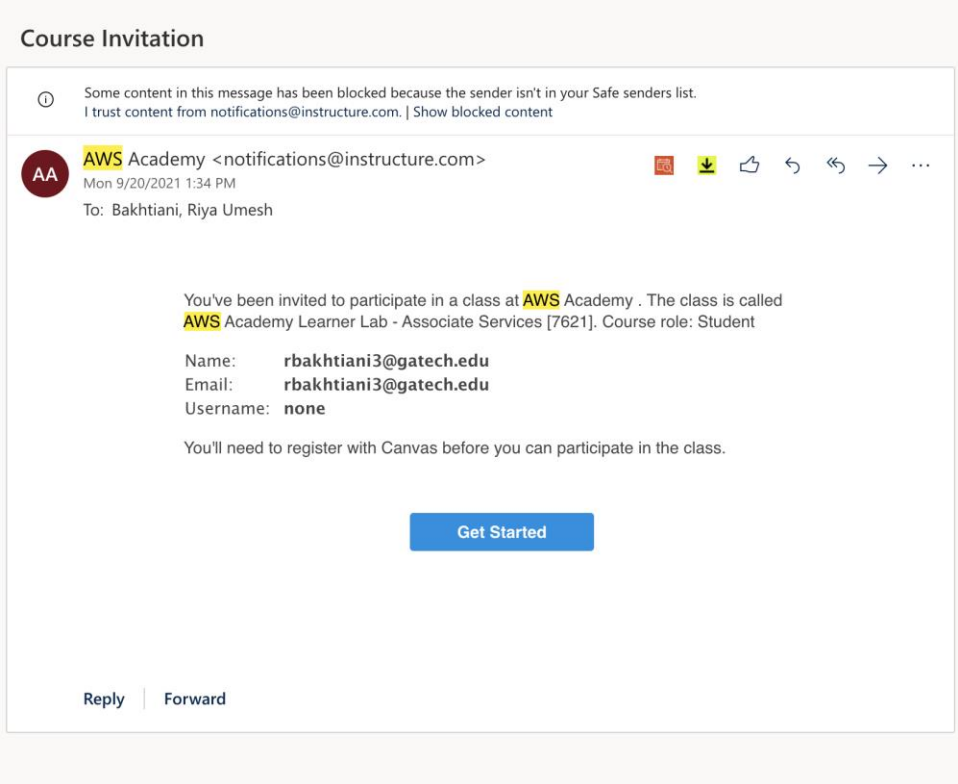
Fall 2022 Setup Guide [For Q3]

Getting Started

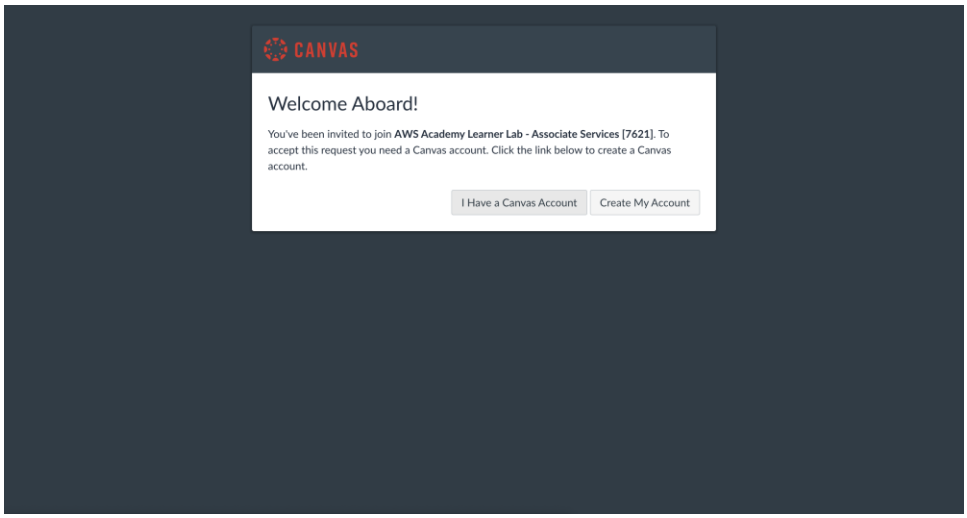
A video tutorial has been created to walk you through the steps 2-7 in this document. View it here: https://youtu.be/QyE2d-9lARc

1. Create an AWS Academy account

You will receive an email from notifications@instructure.com inviting you to participate in the **AWS Academy Learner Lab – Associate Services** course. Your AWS Academy allows you to access EC2, Elastic MapReduce and S3 storage. Click on the button to **Join AWS Academy** in the email to proceed.

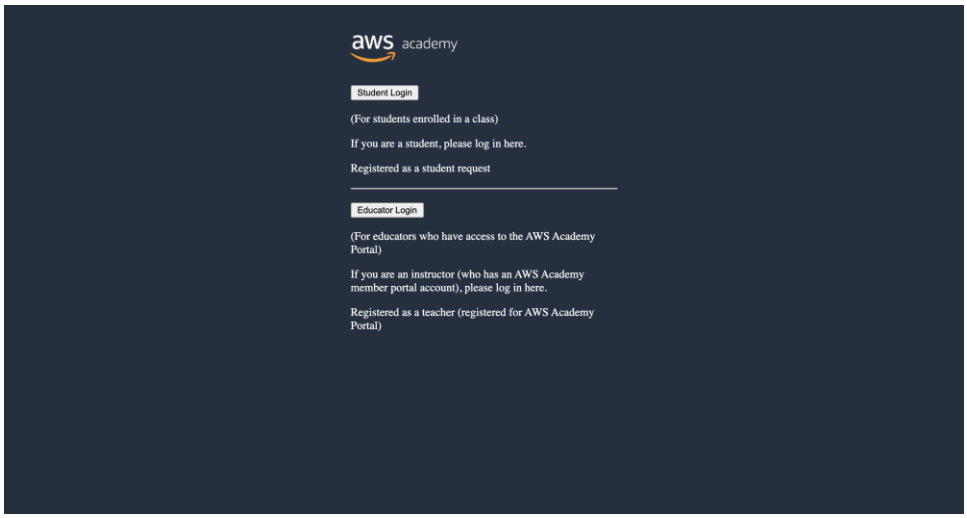


You will be taken to the Canvas page. When it appears, click on the “Create My Account” button.

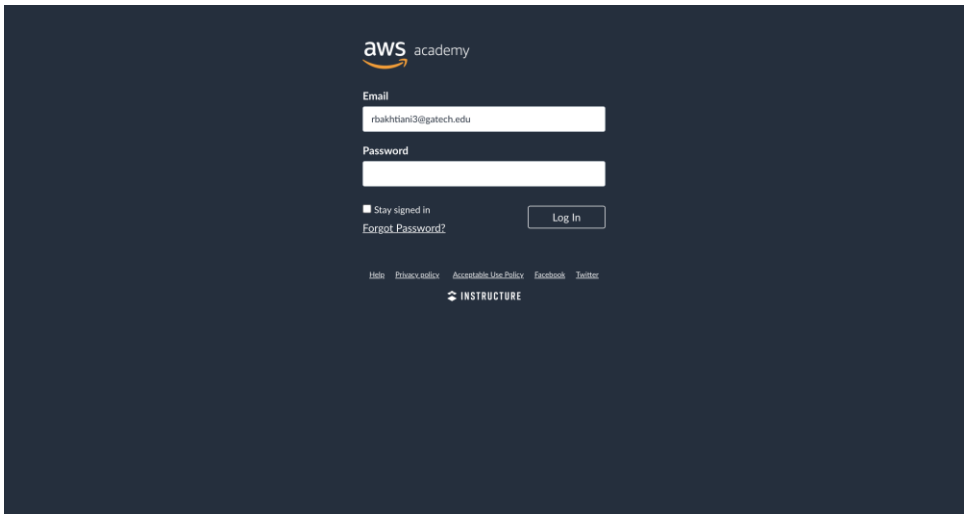


Now, fill in the requested information (e.g. e-mail address, password, etc.). Once submitting, you'll be able to log in to your account at the following URL: https://www.awsacademy.com/LMS_Login.

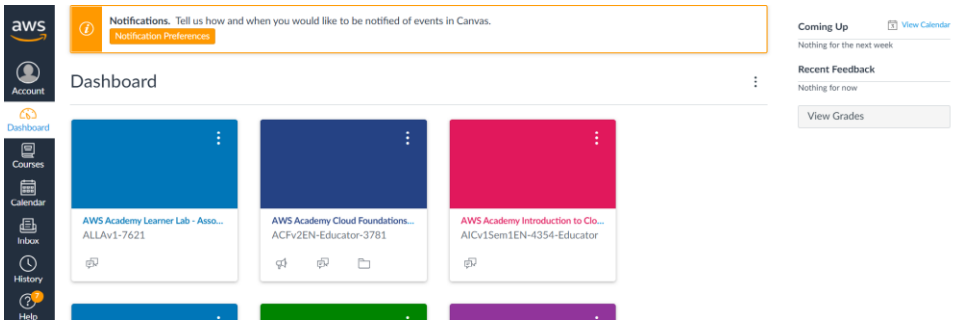
When you arrive at the login URL you will see a screen like this:



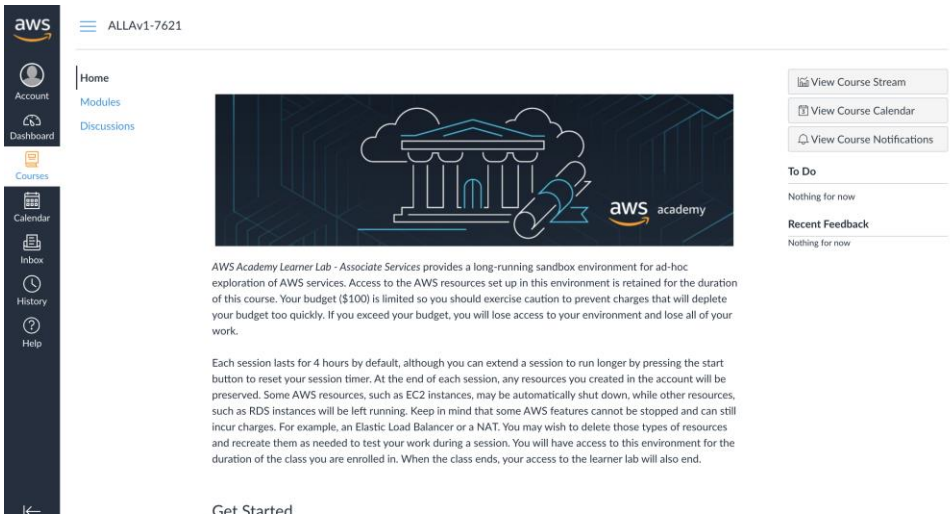
Click the "Student Login" button. Then, you will see the following screen:



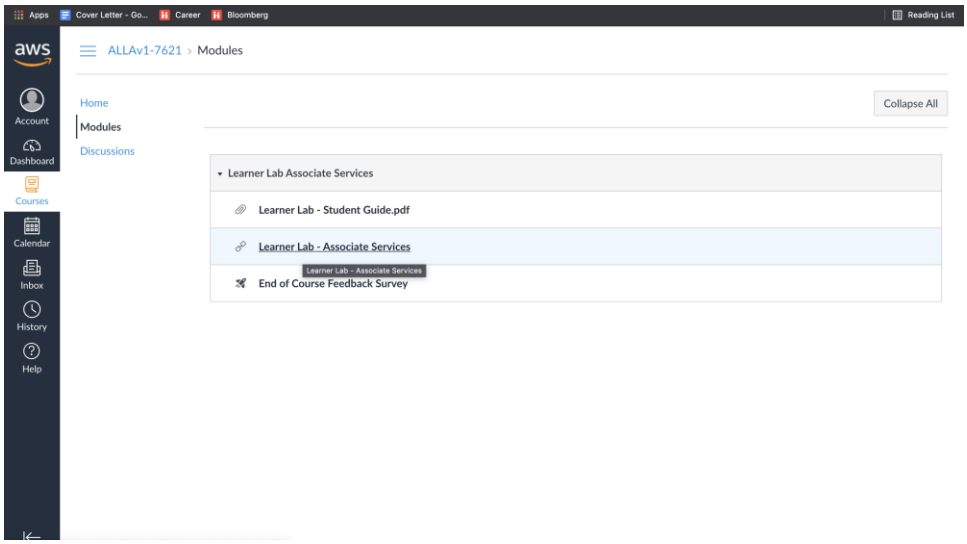
Please enter your e-mail address and password used when setting up the AWS Canvas account (not the Canvas account for the CSE 6242 course). Click the “Log In” button. You will then see the AWS Canvas home screen. Click on the “Dashboard” tab on the far left. You should see a screen like this one (although the number of courses on your screen will differ):



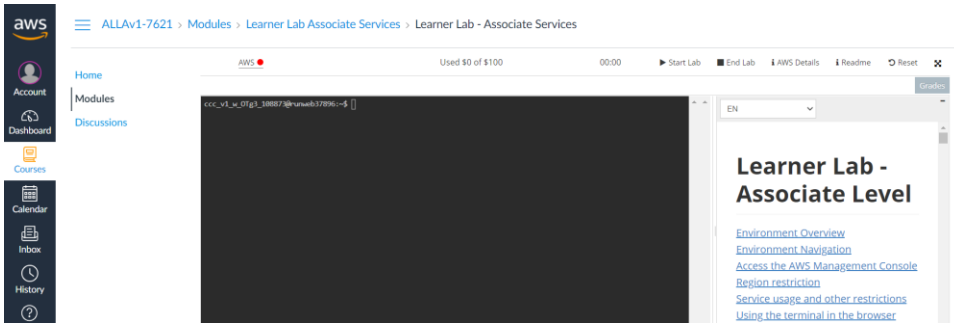
Click on the “AWS Academy Learner Lab” course button. You should see the following screen (the course homepage).



Click on the “Modules” link on the near-left menu. A list of course modules will appear:

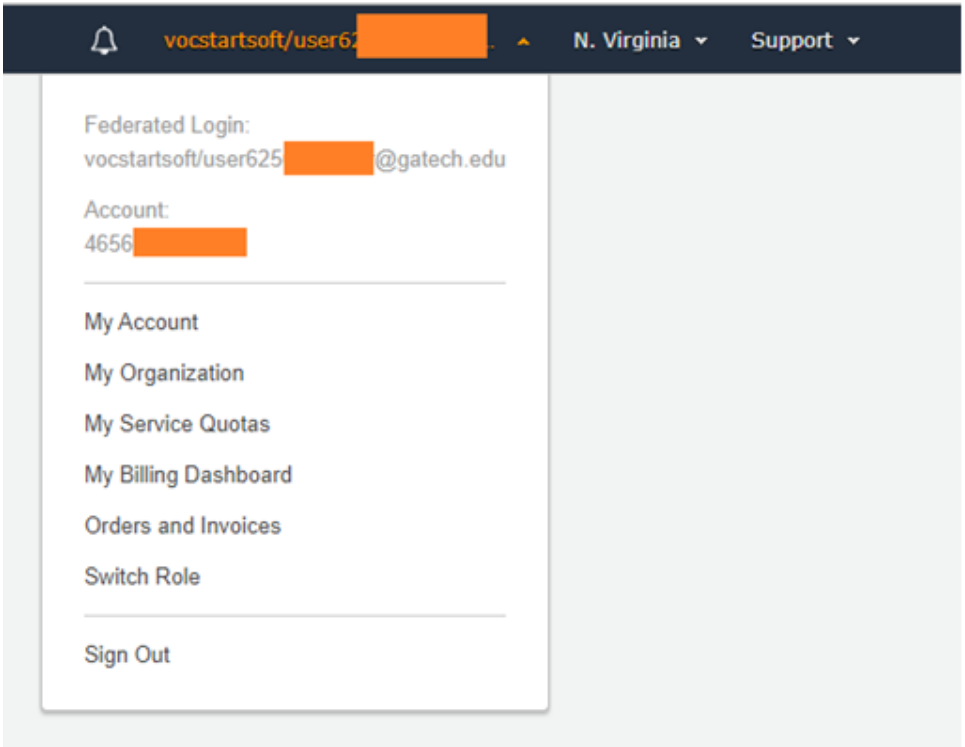


Click on the “Learner Lab” module. You’ll see a screen like the following screen:



Next, click on the “Start Lab” button on the top right. You’ll see an agreement from Vocareum. Scroll to the bottom of the agreement and click “I agree.” It will take a few minutes for the account to be set up. Once the red circle next to AWS turns green in top left, click the button and a new window will open.

You are now in the AWS management console! It will look something like this [right top corner].



If you have any problems with this process, please let the course staff know on EdStem via the dedicated AWS Setup thread.

2. Set up a CloudWatch Usage Alert

NOTE: There are known to be issues with setting up billing alerts 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 extra careful to make sure to close all clusters when not in use.

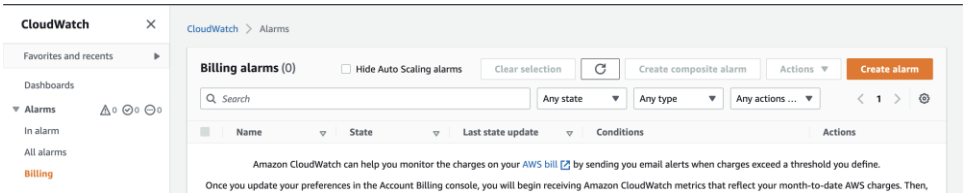
Make sure your region (in the upper right corner of the screen) is set to: **US East (N. Virginia)**. Test whether this email alert is working before scheduling in practice. That is, out of $100, when your credit balance goes below, say, $95, schedule a test alert and make sure it works. Remember this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

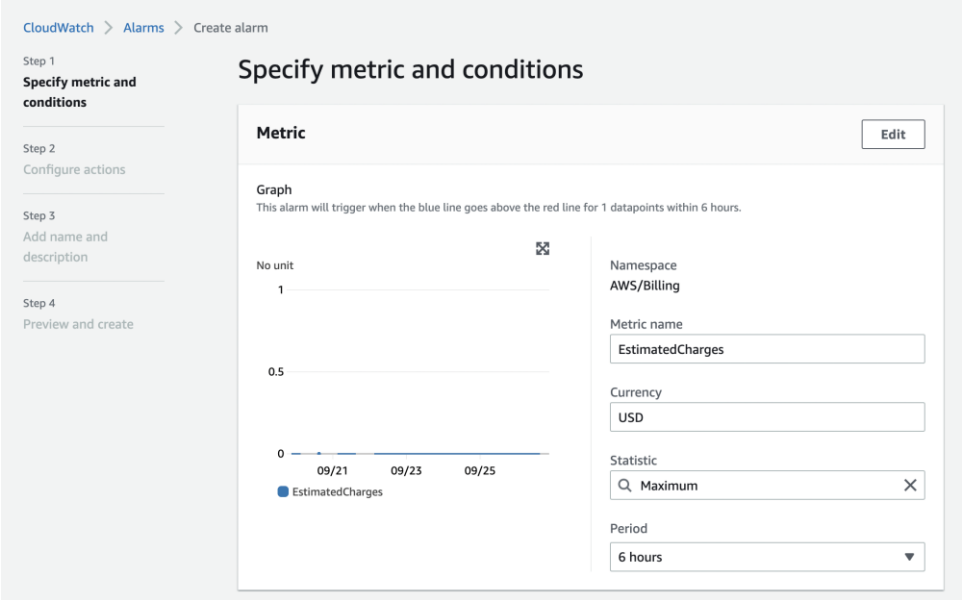
First, we need to create a custom alarm so that it tells you when you have spent money.

1. Open **CloudWatch** in the AWS Management Console. You can access CloudWatch by clicking on the “Services” drop down menu, or alternatively, type **CloudWatch** into the search bar at the top of the AWS console page.

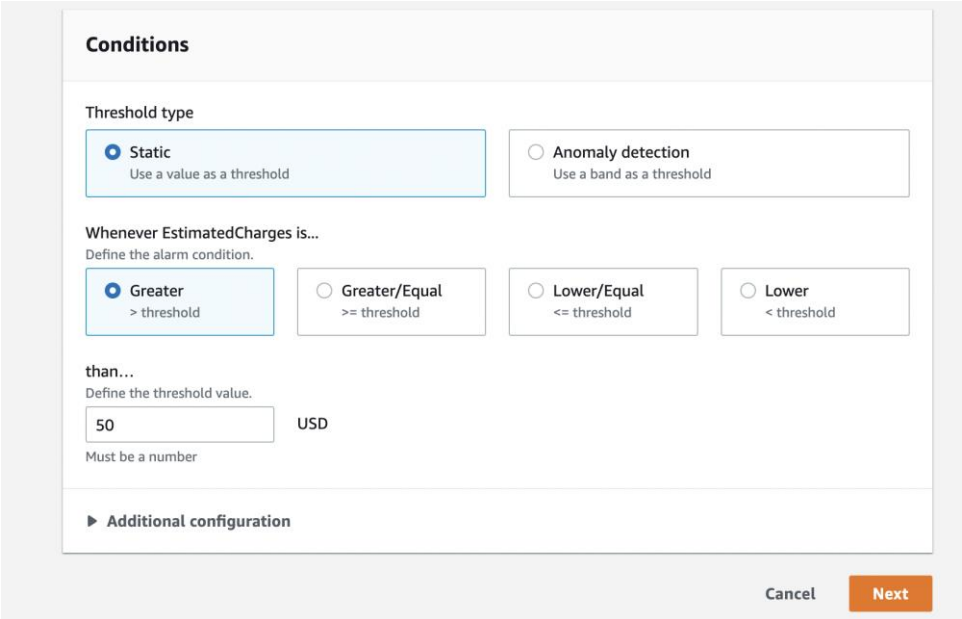
2. In the navigation pane on the left, click **Alarms**, and then **Billing**; then click **Create Alarm**.



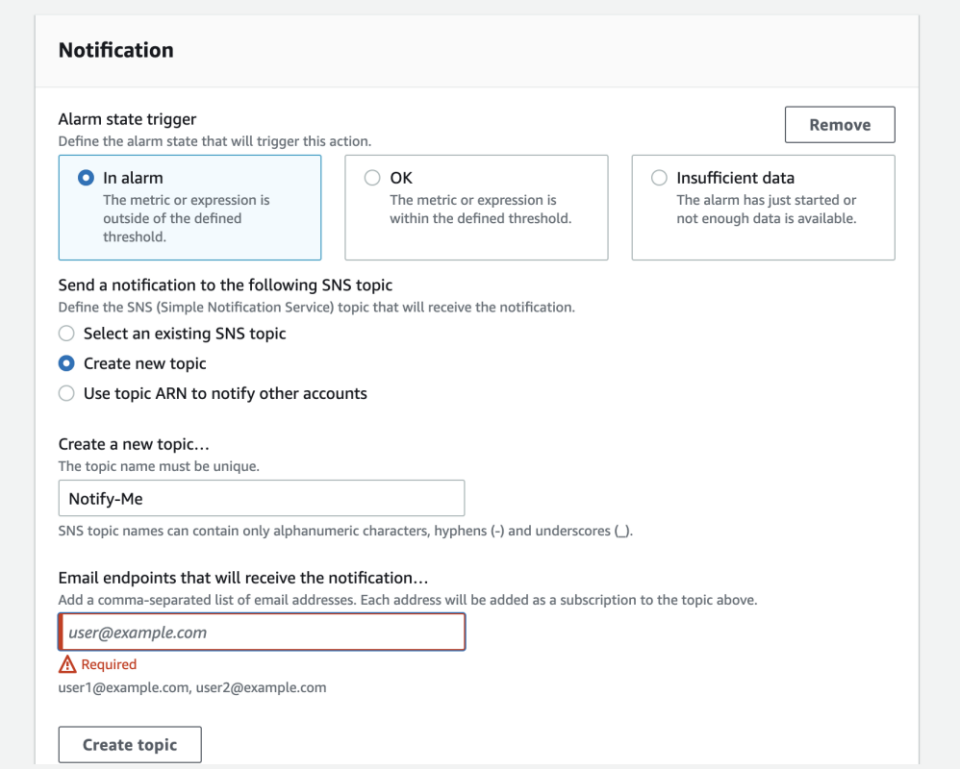
3. Keep the default metric **EstimatedCharges** and scroll down to **Conditions.**



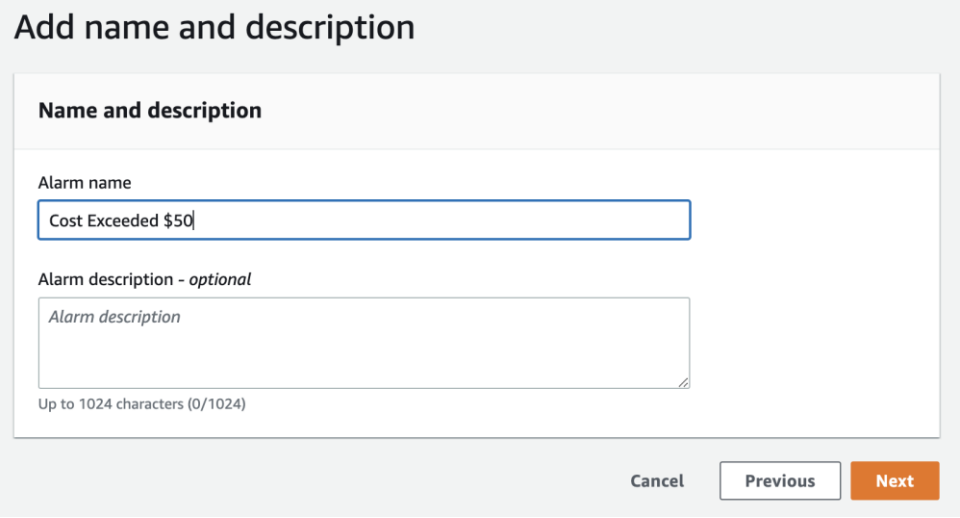
4. Use the default values, but use **50** as the threshold value. This means you will get an alarm once you have spent half of your credits. Click **Next**.



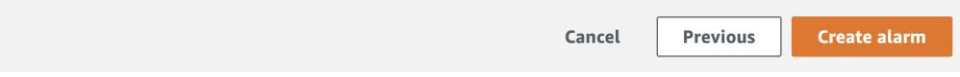
5. Make sure the alarm state is set to **In Alarm**. Then, select **Create a new topic**, and enter a topic name and your email address, then click **Create topic**. Scroll to the bottom of the screen and click N**ext**.



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



7. On the **Preview and create** 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:

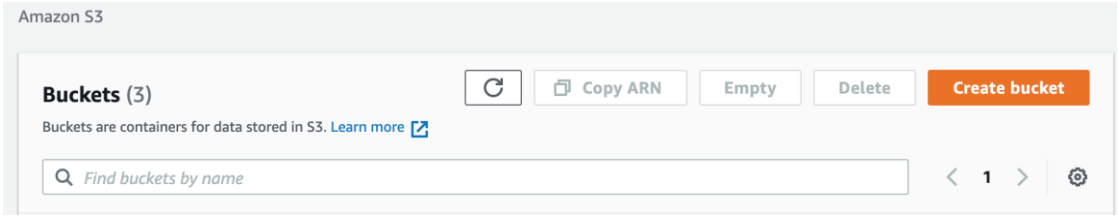
(1) An EMR (Elastic MapReduce) workflow requires the input data to be on S3.

(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**.



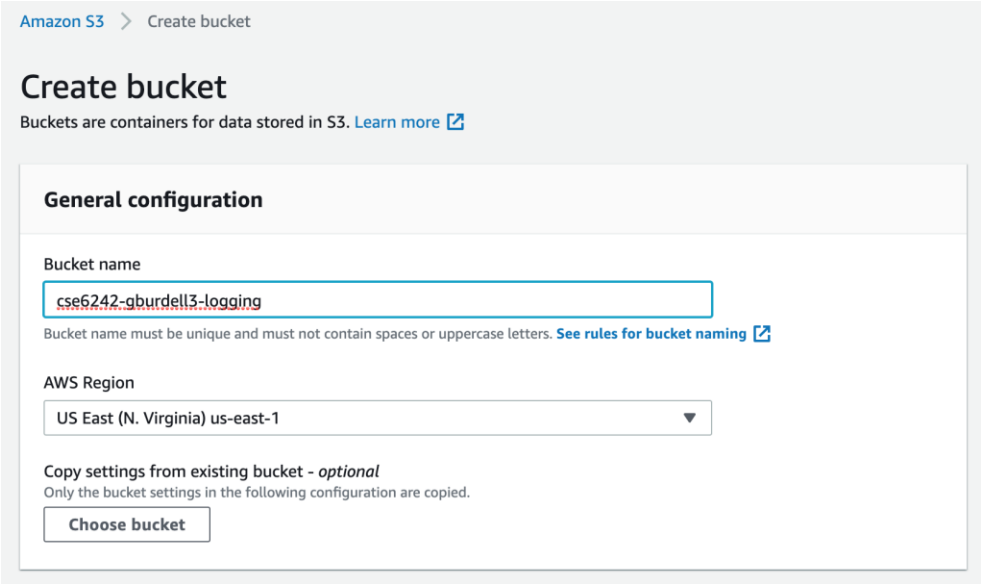
2. Create a logging bucket: Enter the following details (bucket name and region) then click **Create Bucket** at the bottom of the screen. Keep all other settings the same.

   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.



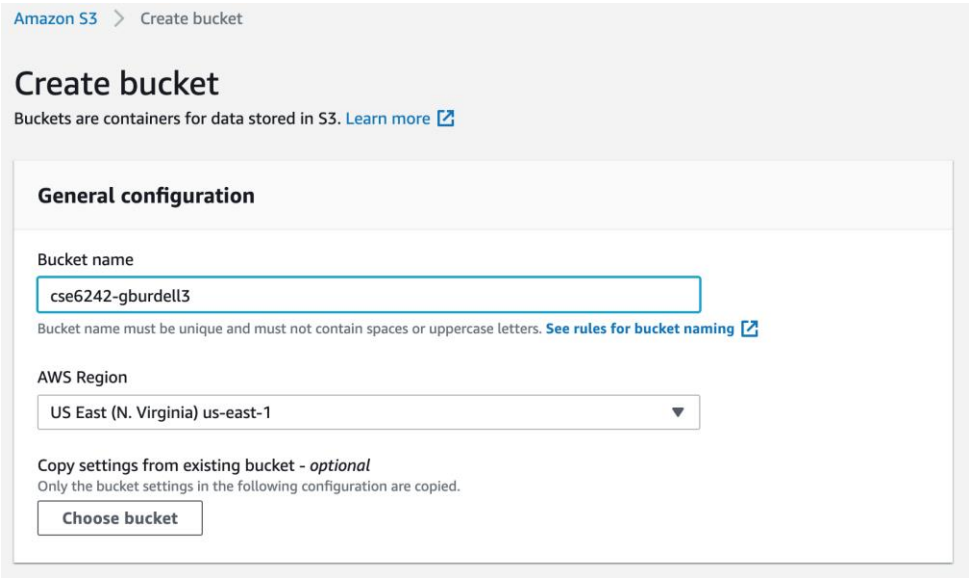
3. A new bucket will appear in the S3 console. Clicking on it will show you that it is empty.

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.

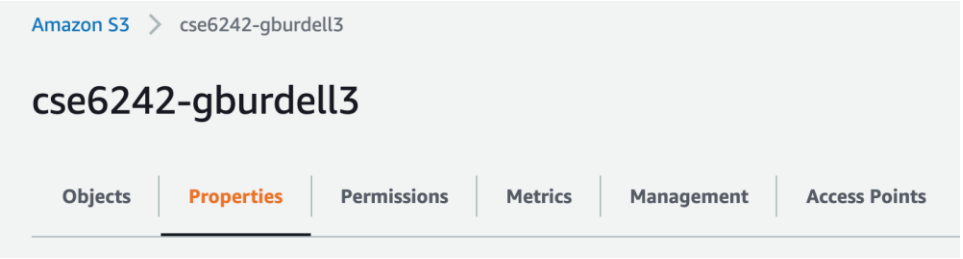
   Bucket Name Format: cse6242-<GT username>

   Example: cse6242-gburdell3

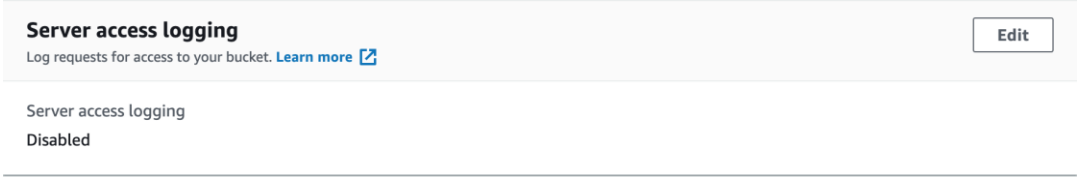
   Region: US East (N. Virginia)



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. Once the bucket is created, click on the bucket on the main screen and select the properties tab.

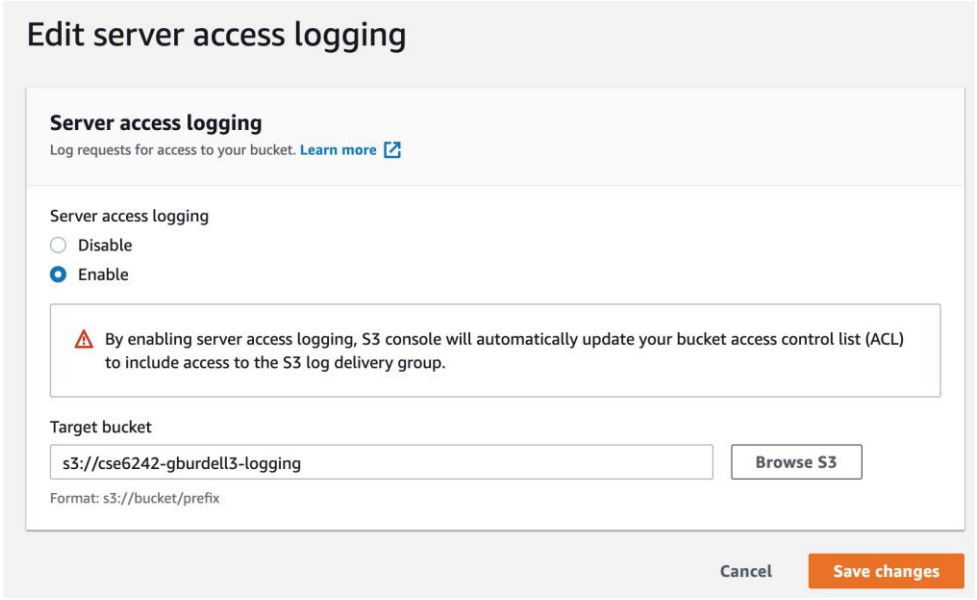


6. Scroll down to **Server Access Logging** and click **Edit.**



7. Select **Enable**, and then make the Target Bucket the logging bucket created in step 2.

   Click **Save Changes**

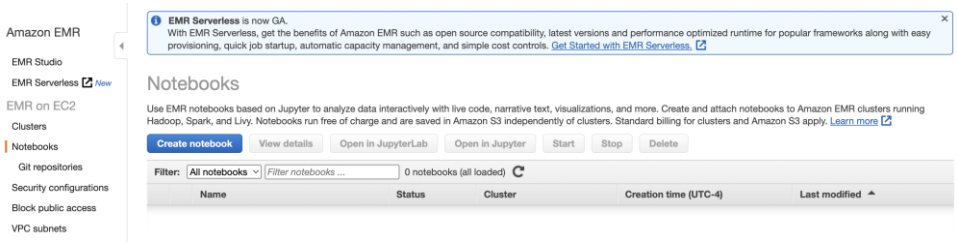


We are done with creating S3 buckets at this point.

4. Launch a Notebook

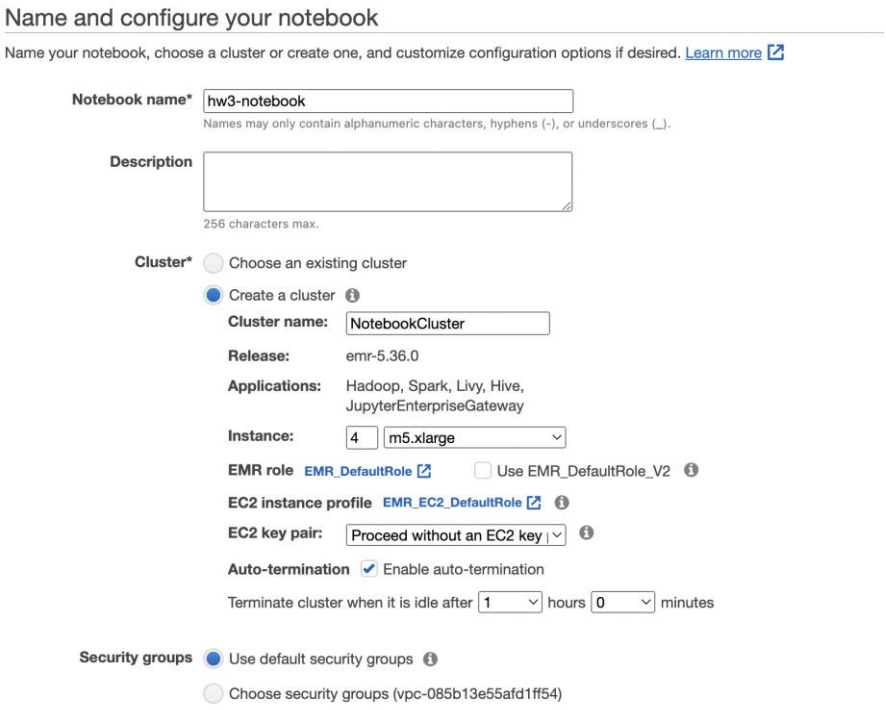
This section will cover launching a Notebook in Amazon EMR. For further information about notebooks in EMR, click here.

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

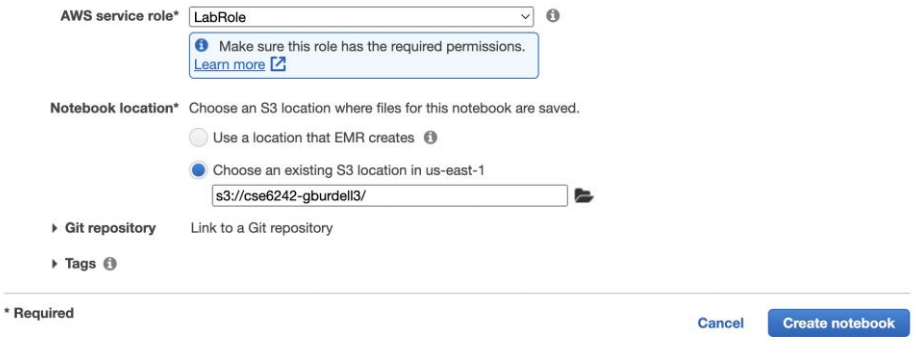


2. Make sure the region specified in the top-right corner of the page is **N. Virginia.** Otherwise click on it and from the drop-down choose N. Virginia.

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 I**nstance**, choose **m5.xlarge** and change 1 to 4.



    d. For AWS service role, select **LabRole**.
    e. For Notebook location, select the s3 bucket (eg: s3://cse6242-gburdell3) you created earlier.
    f. Once you have confirmed this, select “Create Notebook”.

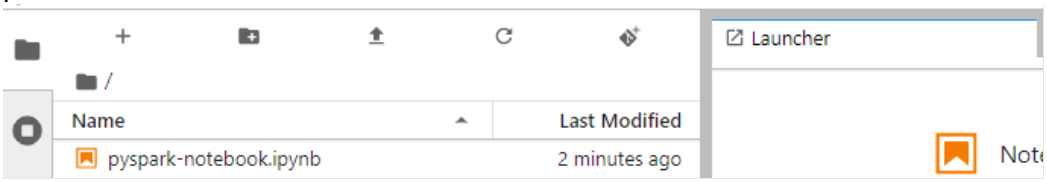


5. Get started with the skeleton

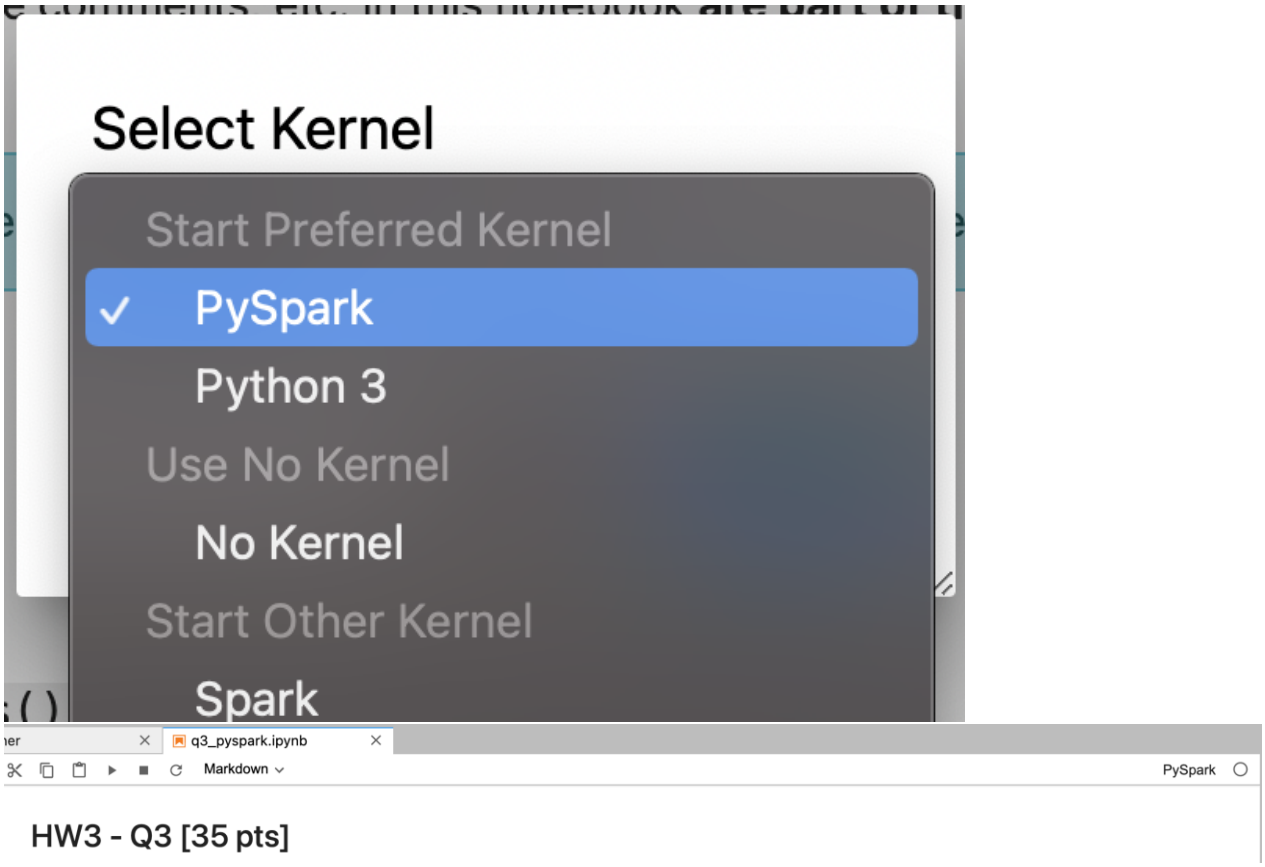
In this section we will upload the skeleton file to the notebook and run our first cell.

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 q3.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.



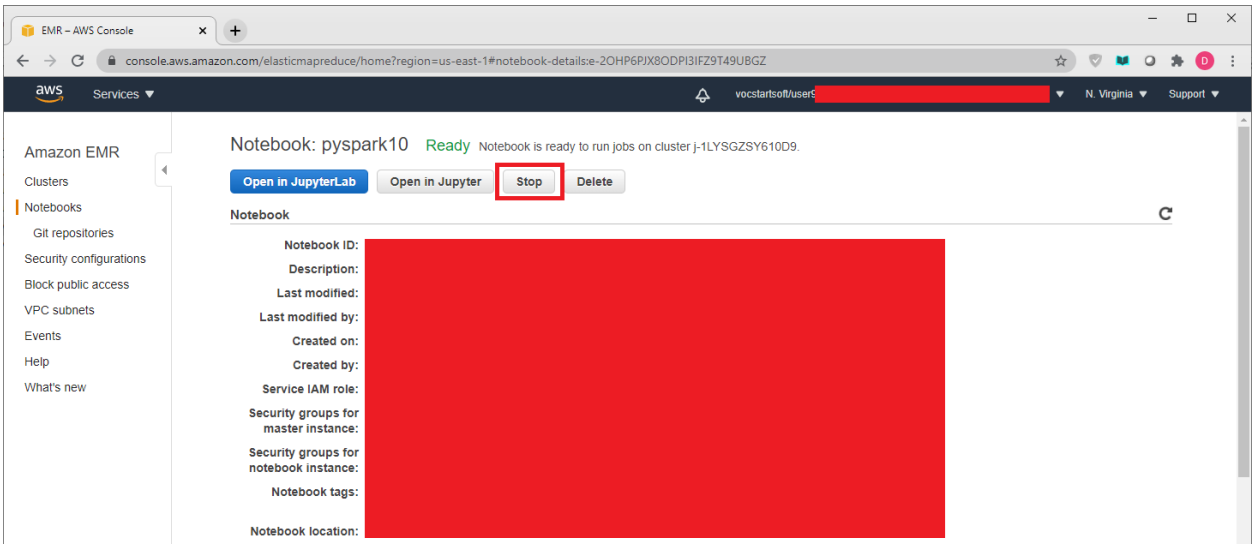
5. Run the first code cell which will import PySpark, followed by a cell which should contain **sc** to start the Spark Application so you can start programming the assignment.
6. Once you have finished coding, right click on the file name in the directory on the left and select download to download 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.

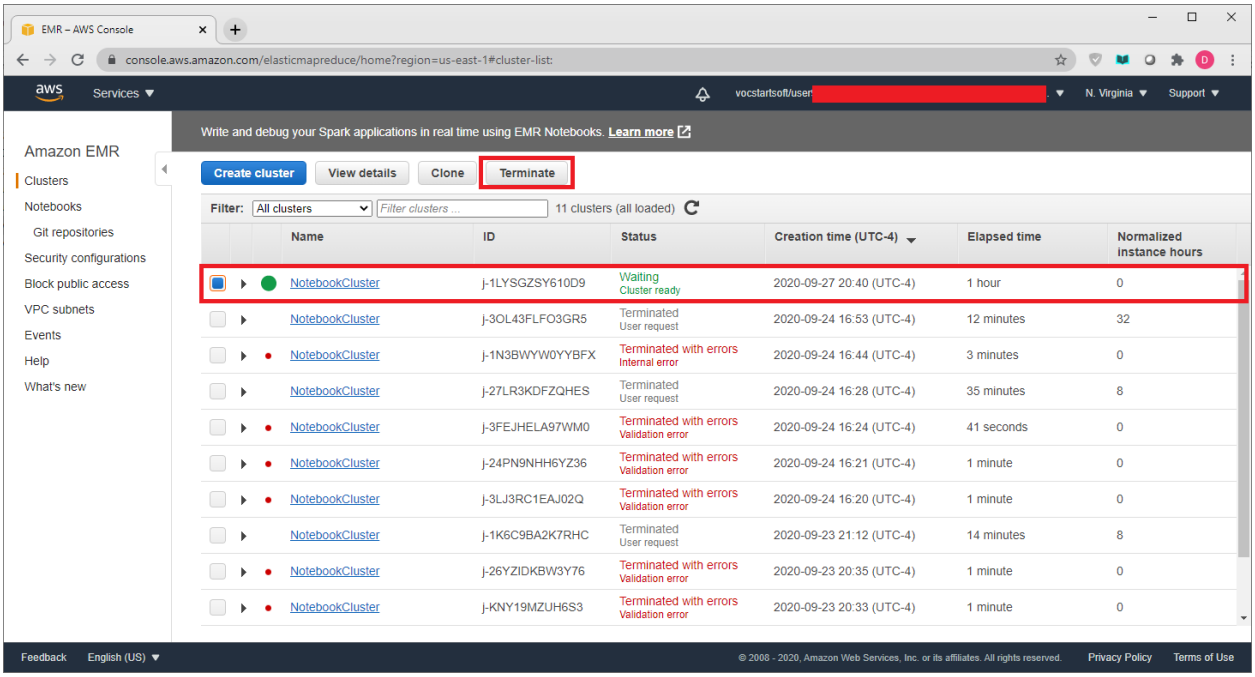
NOTE: The AWS billing report can be as much as six hours behind. It may take up to six hours after terminating all clusters before the billing report stops increasing.

1. After saving your Notebook, back on your Notebook's page in EMR, click 'Stop'.

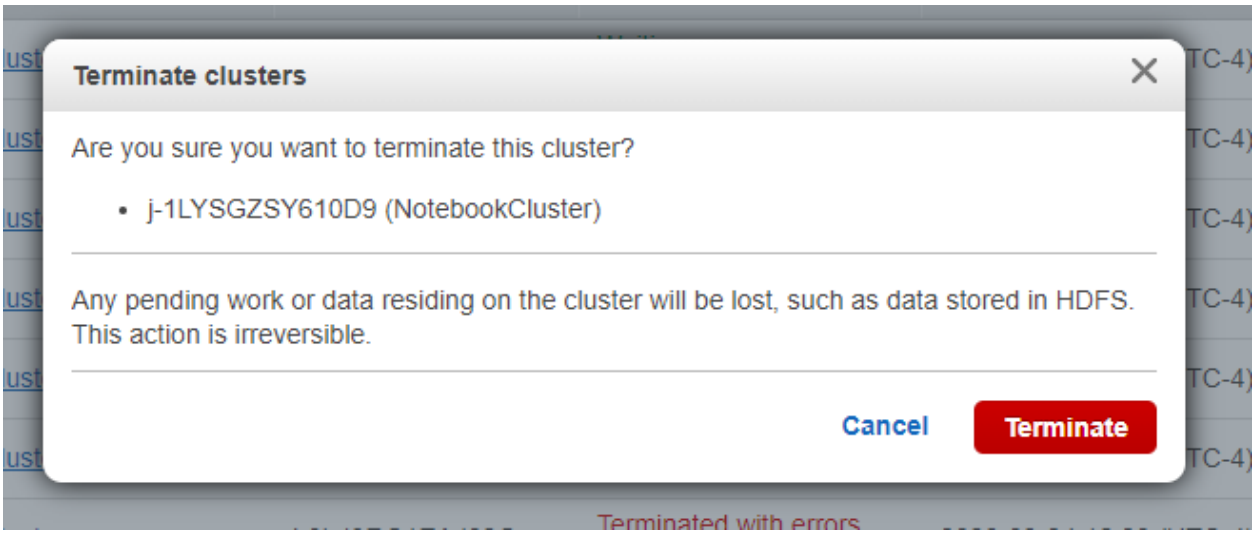


2. Now click on “Clusters” in the side bar on the left. Click on “Active Clusters.” Click the check box next to your running cluster (the one with the green circle) and click “Terminate”.

Note: You may have to refresh your screen for the cluster to show up.



3. In the popup, select ‘Terminate’.

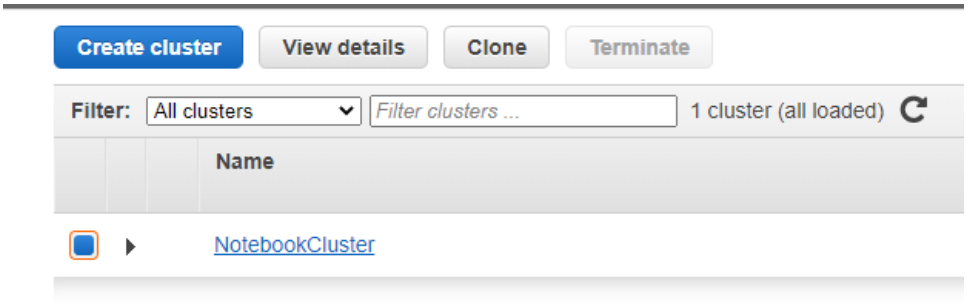


You have now closed all your clusters and will no longer be accruing charges!

7. Restarting an Old Cluster

If you stopped your cluster and took a break and want to start the assignment again, there is a quick and easy way to do so.

1. Clone the old terminated cluster. Click on “Clusters” and select “Terminated Clusters” from the drop down menu.



2. It will then ask if you would like to copy the setting from the old cluster. Click Yes.
3. Confirm the settings and Start the cloned cluster, waiting 5-10 minutes for it to spin up.
4. You will then have to start your old notebook and attach it to the running cluster.