



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.

Action Required: You have been invited to join AWS Educate



AWS Educate Support <support@awseducate.com>

To

[Redacted email address]



2:20 PM

If there are problems with how this message is displayed, click here to view it in a web browser.



Mahdi Roozbahani from **Georgia Institute of Technology** has invited you to join AWS Educate, where you can find AWS content to help with classwork and connect to self-paced labs and training resources. Please sign up within **30 days**, or this invitation will expire.

If you have questions about AWS Educate, review our [FAQs](#). We're excited to have you on board!

[Join AWS Educate](#)

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

Preferred Language:
English

Georgia Institute of Technology

United States

Start typing the name of your school and select from the list. If you don't see your school, enter the full name, example: Harvard University

Daniel

Fasciano

dfasciano3@gatech.edu

12

2021

Please provide a valid, current email issued by your institution. Example: your_name@your_school.edu

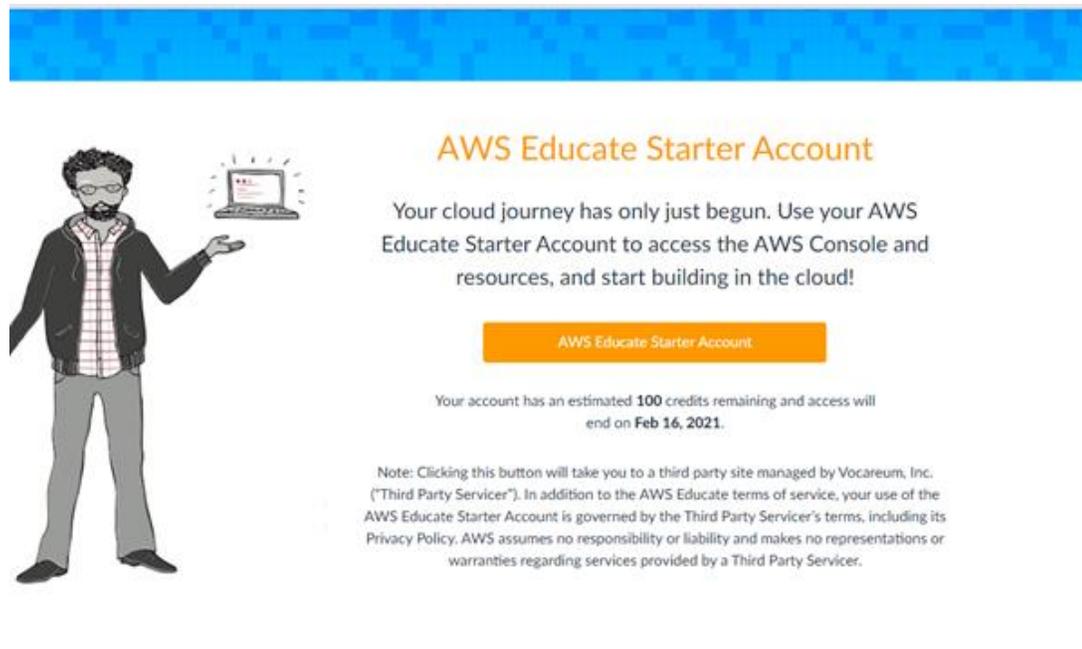
Birth Month

Birth Year

Promo Code (optional)

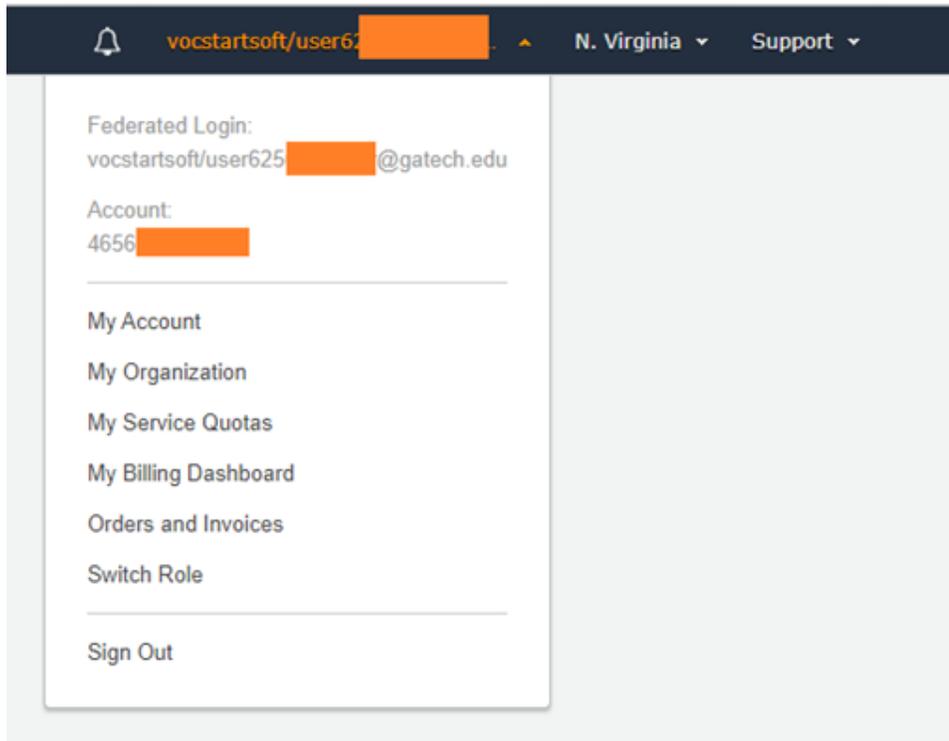
On the following screen check the box which says 'I Agree' then click Submit. You will receive an email which asks you to confirm your email, then you'll be able to log in to your account.

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.



The image shows a landing page for the AWS Educate Starter Account. On the left, there is a cartoon illustration of a man with a beard and glasses, wearing a dark jacket over a plaid shirt and grey pants, pointing towards a laptop. The laptop screen displays the AWS logo. The main heading is "AWS Educate Starter Account" in orange. Below the heading, the text reads: "Your cloud journey has only just begun. Use your AWS Educate Starter Account to access the AWS Console and resources, and start building in the cloud!". A prominent orange button labeled "AWS Educate Starter Account" is centered below the text. Underneath the button, it states: "Your account has an estimated 100 credits remaining and access will end on Feb 16, 2021." At the bottom, there is a small note: "Note: Clicking this button will take you to a third party site managed by Vocareum, Inc. ('Third Party Servicer'). In addition to the AWS Educate terms of service, your use of the AWS Educate Starter Account is governed by the Third Party Servicer's terms, including its Privacy Policy. AWS assumes no responsibility or liability and makes no representations or warranties regarding services provided by a Third Party Servicer."

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.

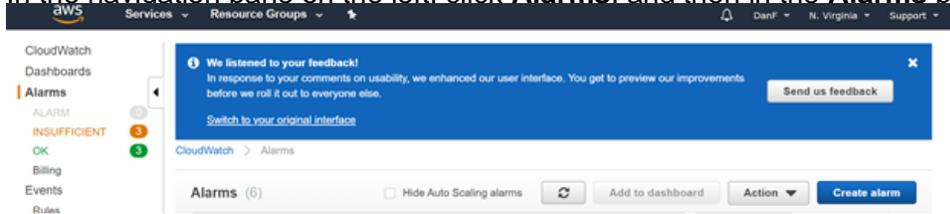
Make sure your region (in the upper right corner of the screen) is set to: **US East (N. Virginia)**. [Test whether this email alert is working before scheduling in practice](#). That is, out of \$100, when your credit balance goes below say \$95, schedule a test alert and make sure it works. Remember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

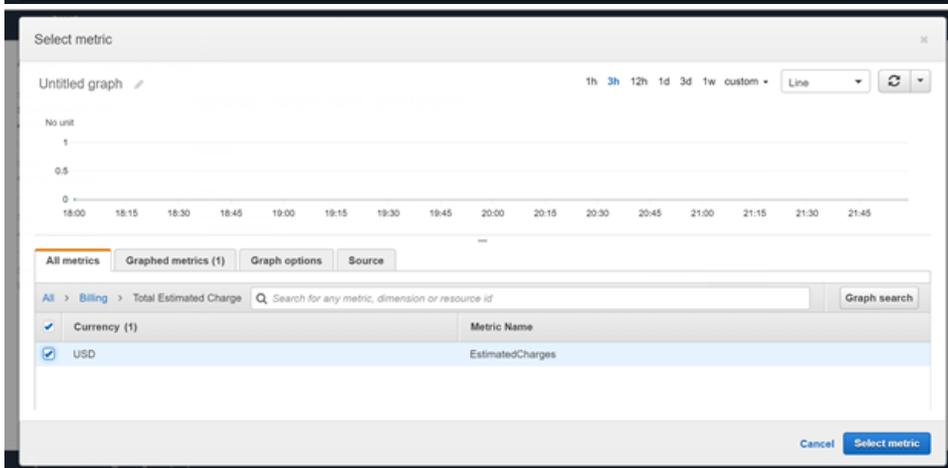
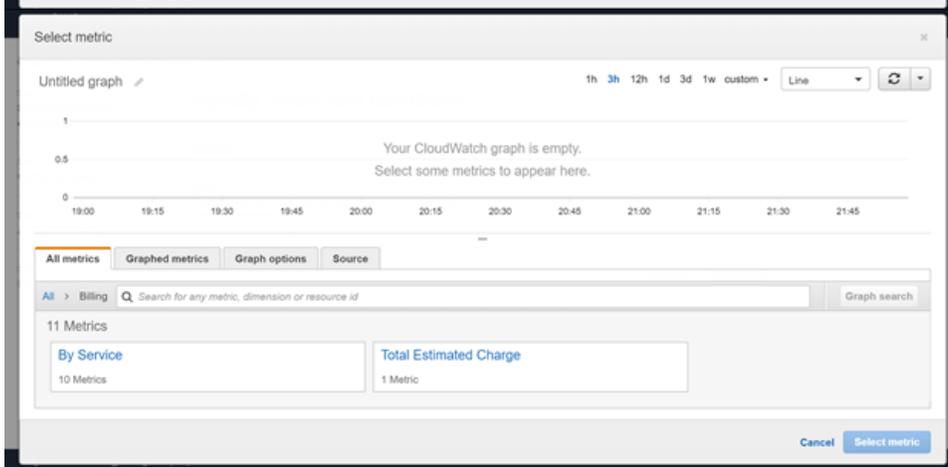
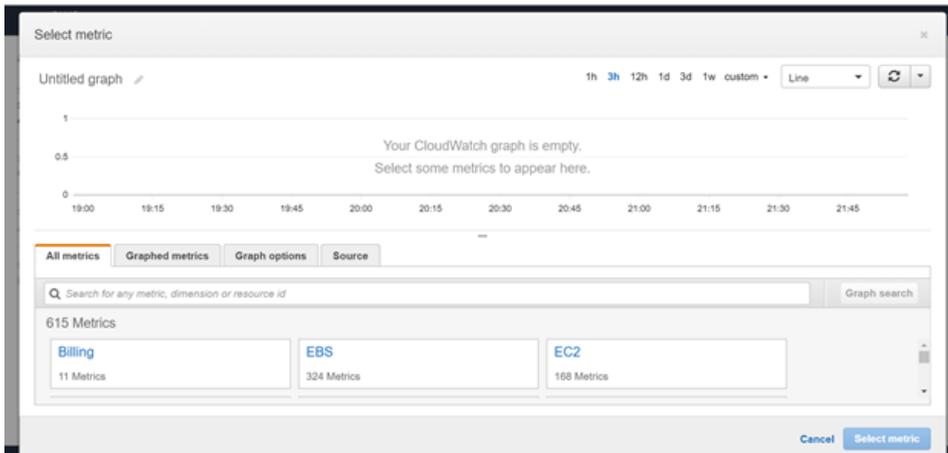
2. In the navigation pane on the left, click **Alarms**, and then in the **Alarms** pane, click **Create Alarm.**



3. Click on **Select Metric.**



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



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

Conditions

Threshold type

Static

Use a value as a threshold

Anomaly detection

Use a band as a threshold

Whenever EstimatedCharges is...

Define the alarm condition

Greater

> threshold

Greater/Equal

>= threshold

Lower/Equal

<= threshold

Lower

< threshold

than...

Define the threshold value

50

USD

Must be a number

► Additional configuration

Cancel

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.

Whenever this alarm state is...

Define the alarm state that will trigger this action

Remove

in Alarm

The metric or expression is outside of the defined threshold.

OK

The metric or expression is within the defined threshold.

INSUFFICIENT_DATA

The alarm has just started or not enough data is available.

Select an SNS topic

Define the SNS (Simple Notification Service) topic that will receive the notification

Select an existing SNS topic

Create new topic

Use topic ARN

Create a new topic...

The topic name must be unique.

Notify-Me

SNS topic names can contain only alphanumeric characters, hyphens (-) and underscores (_).

Email endpoints that will receive the notification...

Add a comma-separated list of email addresses. Each address will be added as a subscription to the topic above.

user@example.com

user1@example.com, user2@example.com

Create topic

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

Add a description

Name and description

Define a unique name

Alarm name

Cost Exceeded \$50

Alarm description - optional

Define a description for this alarm. Optionally you can also use markdown.

Alarm description

Up to 1024 characters (0/1024)

Cancel

Previous

Next

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

Cancel

Previous

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

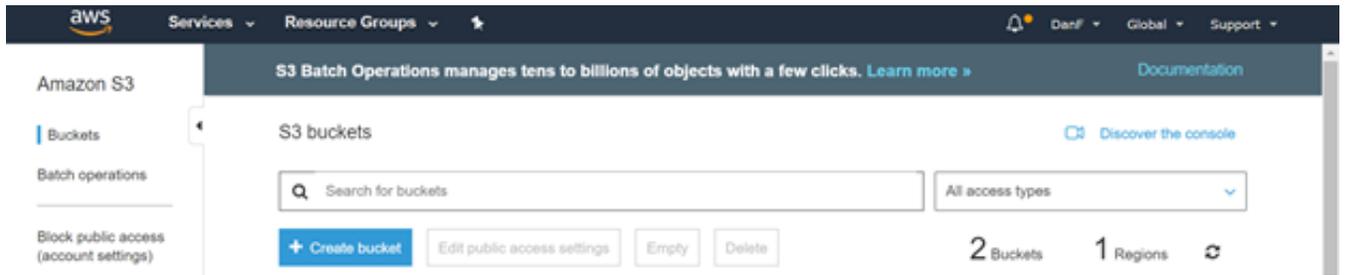
We need S3 for two reasons:

- (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**.



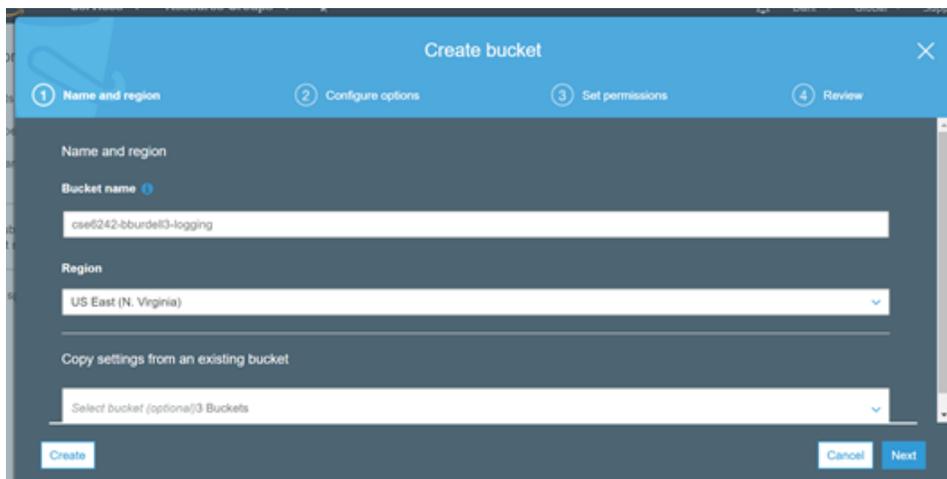
2. Create a logging bucket: Click on **Create Bucket** and enter the following details (bucket name and region) then click **Create** (not Next). Keep all other settings as the same.

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.

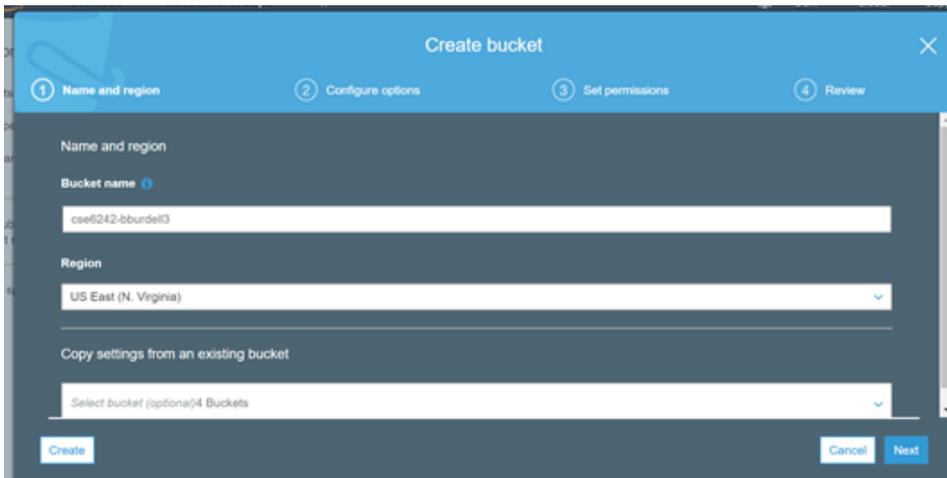


3. A new bucket will appear in the S3 console. Clicking on it will show you that it is empty.
4. Create the main bucket: Go back to the main screen (clicking on **Amazon S3**). Again, click on **Create Bucket** and enter the following details.

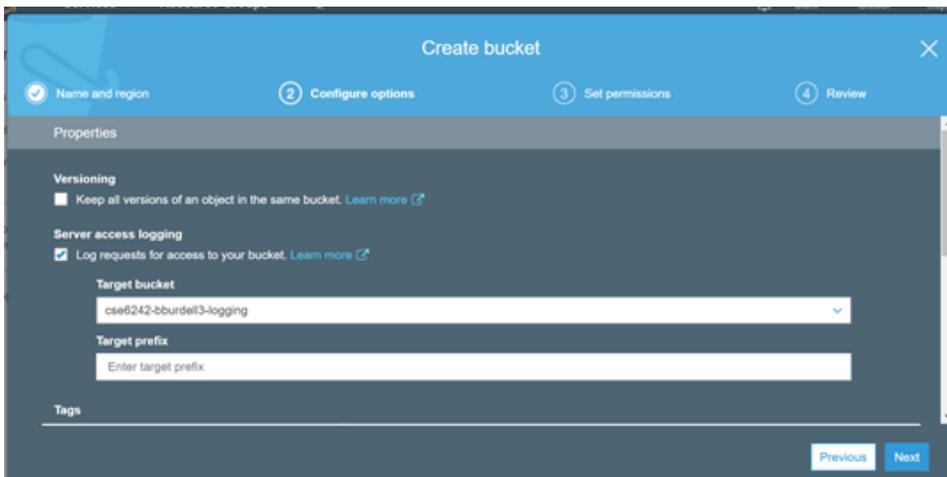
Bucket Name Format: cse6242-<GT username>

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**.



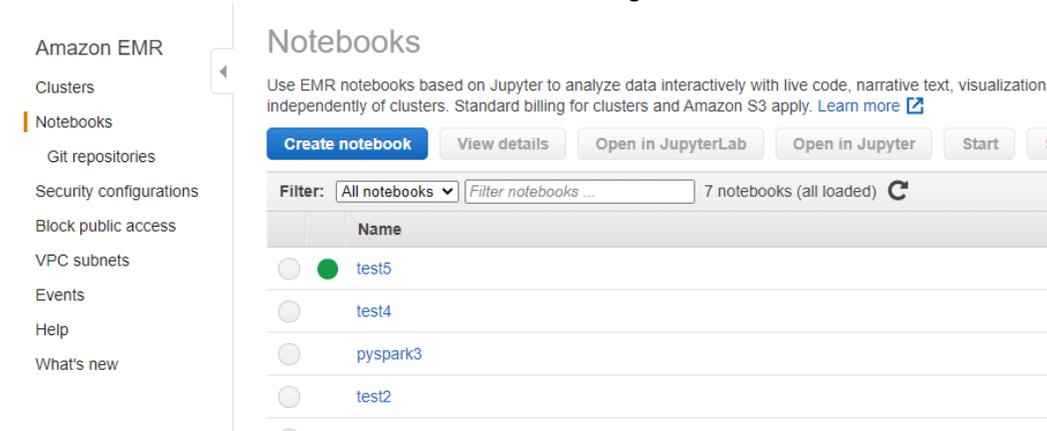
6. Click on **Target Bucket** and select the 'logging bucket' you created. (e.g.: cse6242-gburdell3-logging) in the drop-down menu.
7. Keeping the settings as they are in the next 2 tabs, Hit **Next** twice and press **Create Bucket**.

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”.

Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. [Learn more](#)

Notebook name*
Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

Description
256 characters max.

Cluster* Choose an existing cluster
 Create a cluster **i**

Cluster name:

Release: emr-5.1.0

Applications: Hadoop, Spark, Livy, Hive

Instance:

EMR role: [EMR_DefaultRole](#) **i**

EC2 instance profile: [EMR_EC2_DefaultRole](#) **i**

EC2 key pair: **i**

Security groups Use default security groups **i**
 Choose security groups (vpc-23817b5e)

AWS service role* **i**

Notebook location* Choose an S3 location where files for this notebook are saved.
 Use the default S3 location
 Choose an existing S3 location in us-east-1

▶ **Git repository** [Link to a Git repository](#)

▶ **Tags** **i**

* Required

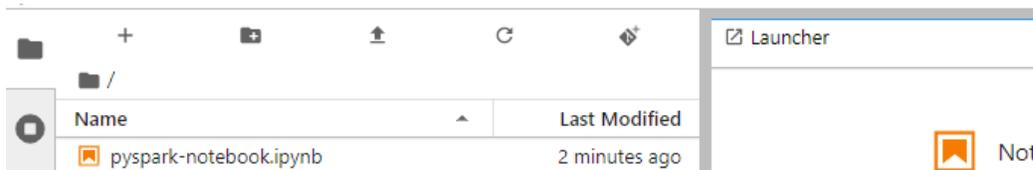
Cancel

Create notebook

5. Get started with the skeleton

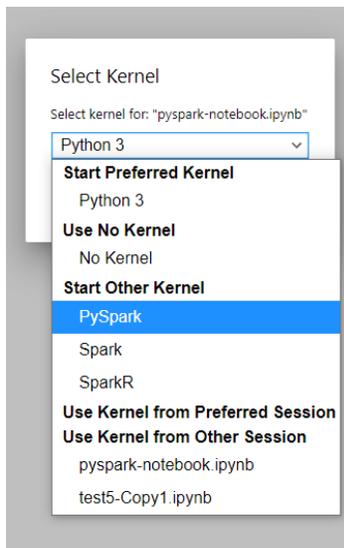
In this section we will upload the skeleton file to the notebook and run our first cell.

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.

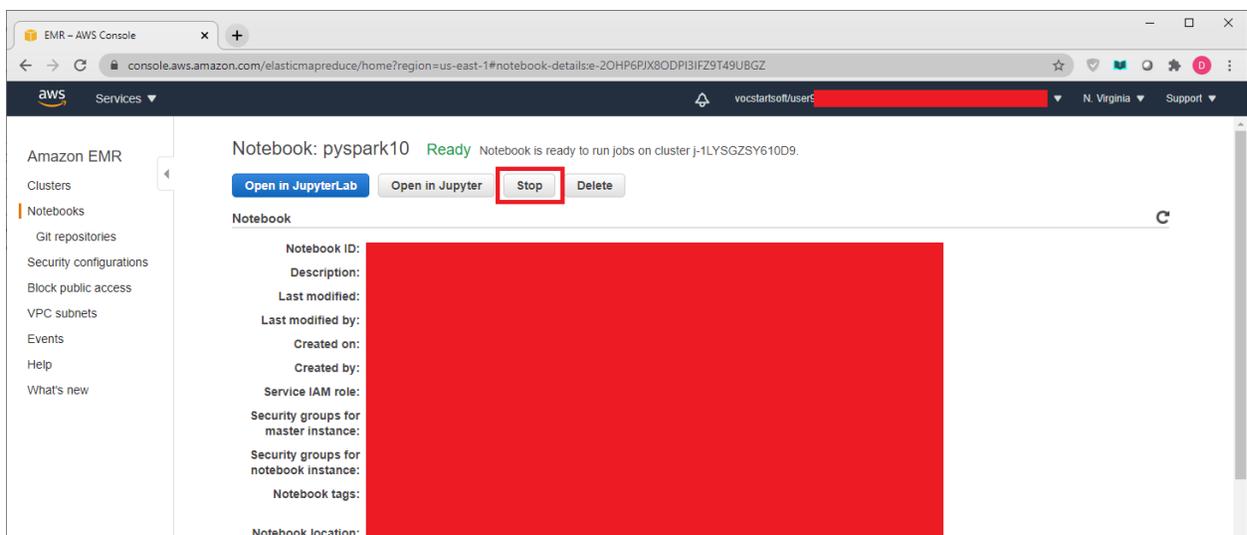


5. Run the first code cell, which should contain `sc` to start the Spark Application so you can start programming the assignment.
6. Once you have finished coding, right click on the file name in the directory on the left and select download to download it. It will also be saved in your S3 bucket,

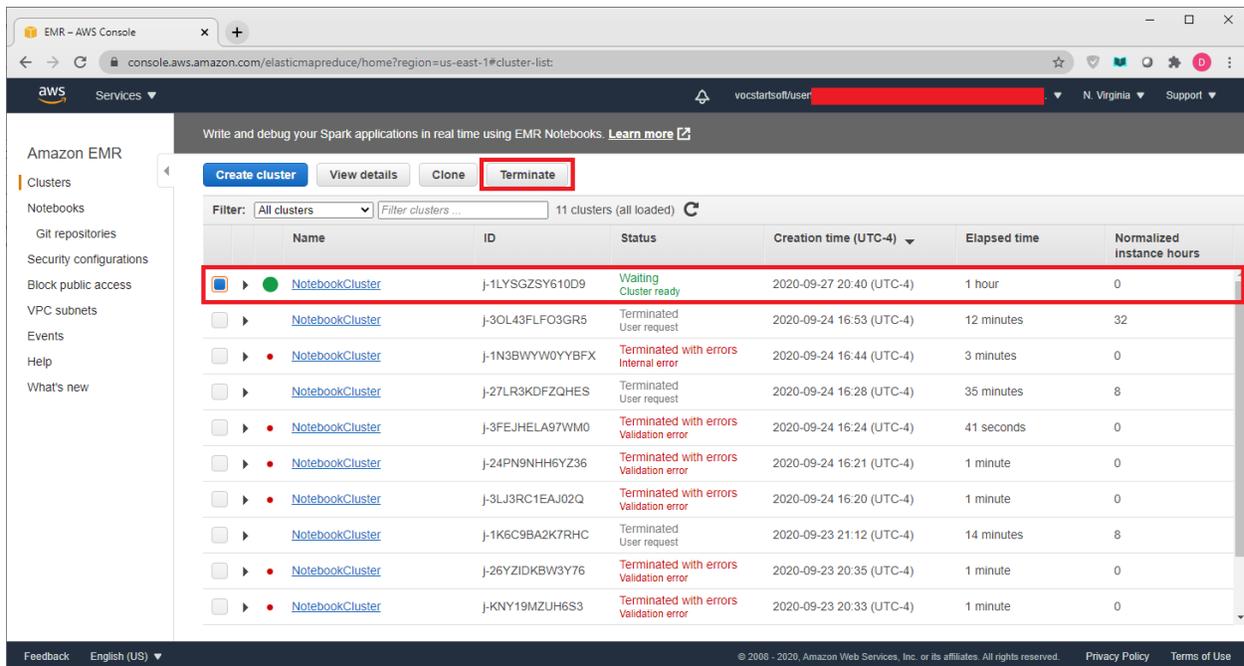
6. Terminating All Clusters

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.

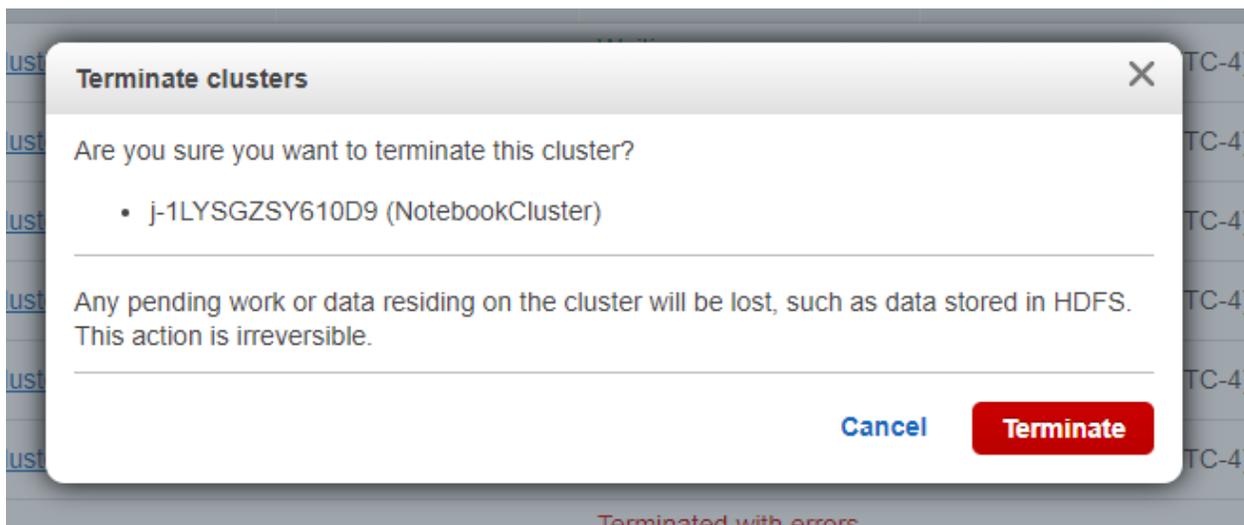
1. Back on your Notebook's page in EMR, click 'Stop'.



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!