

Spring 2021 Setup Guide [For Q3]

Getting Started

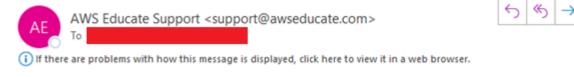
A video tutorial has been created to walk you through the steps in this document. View it here.

Note: This video is from Fall 2020, so some UI elements may have changed. The instructions below should be up to date.

1. Create an AWS Educate account

You will receive an email from <u>support@awseducate.com</u> 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





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 <u>FAQs</u>. We're excited to have you on board!

Join AWS Educate

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

2:20 PM

				Preferred Lang	guage:
				English	,
Georgia Institute of Technology					_
	0	United States			•
Start typing the name of your school and select from the list. If you don't see your sch enter the full name, example: Harvard University	hool,				
Daniel		Fasciano			
dfasciano3@gatech.edu		12	2021		- O
Please provide a valid, current email issued by your institution. Example: your_name@your_school.edu					
Birth Month Birth Year	- 0	Promo Code (optional)			0

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].

A vocstartsoft/us	ser62 . 🔺	N. Virginia 👻	Support 👻
Federated Login:			
vocstartsoft/user625	@gatech.edu		
Account:			
4656			
My Account			
My Organization			
My Service Quotas			
My Billing Dashboard			
Orders and Invoices			
Switch Role			
Sign Out			

If you have any problems with this, or you receive an email from AWS saying that your application has been rejected, please let the course staff know on Piazza via the dedicated AWS Setup thread.

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)**. <u>Test</u> <u>whether this email alert is working before scheduling in practice</u>. 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.]

aws	Services	- Resource Groups - 🐆), DanF = N. Virginia = Support =
CloudWatch Dashboards Alarms ALARM INSUFFICIENT OK	(() () () () () () () () () () () () ()	We listened to your feedback! In response to your comments on usability, we enhanced our user interface. You get to preview our improveme before we roll it out to everyone else. Switch to your original interface CoudWatch > Alarms	nts Send us feedback
Billing Events		Alarms (6) Hide Auto Scaling alarms 2 Add to dashboard	Action 👻 Create alarm
Rules Click on 1	ا مام	ct Metric.	
	Jeie		
2005	-		
aws	Sen	ices 🗸 Resource Groups 🤟 🛠	û DanF ◄
CloudWatch >	Alarms	> Create alarm	
Step 1 Specify metric a conditions	nd	Specify metric and conditions	
Step 2		Metric	
Configure actions		Graph Preview of the metric or metric expression and the alarm threshold.	
Step 3 Add a description		Select metric	
Step 4 Preview and creat	te		Cancel Next

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

3.

Select metric	×
Untitled graph	1h 3h 12h 1d 3d 1w custom - Line 👻 📿 👻
1	Veux Plaud Malah arank is amake
0.5	Your CloudWatch graph is empty. Select some metrics to appear here.
0	euros como momos so appear nero:
19:00 19:15 19:30 19	145 20:00 20:15 20:30 20:45 21:00 21:15 21:30 21:45
All metrics Graphed metrics Graph optic	ons Source
Q Search for any metric, dimension or resource id	Graph search
615 Metrics	[F00
Billing 11 Metrics	EBS EC2
	UNIV INVENTION 1.00 (INVENTION
	Cancel Select metric
Select metric	х
Untitled graph 🥒	1h 3h 12h 1d 3d 1w custom - Line 👻 📿 👻
1	
0.5	Your CloudWatch graph is empty.
	Select some metrics to appear here.
0 19:00 19:15 19:30 19	145 20:00 20:15 20:30 20:45 21:00 21:15 21:30 21:45
	-
All metrics Graphed metrics Graph optic	ons Source
All > Billing Q Search for any metric, dimension	or resource id Graph search
11 Metrics	
By Service	Total Estimated Charge
10 Metrics	1 Metric
	Cancel Select metric
Select metric	х
Untitled graph 🥒	1h 3h 12h 1d 3d 1w custom - Line 💌 🖸 💌
No unit	
1	
0.5	
0	0 19:15 19:30 19:45 20:00 20:15 20:30 20:45 21:00 21:15 21:30 21:45
IN. IV. IN. IV. 10.40 10.00	
All metrics Graphed metrics (1) Graph o	options Source
All > Billing > Total Estimated Charge Q, Sea	Inch for any metric, dimension or resource id
Currency (1)	Metric Name
🕑 USD	EstimatedCharges
	Cancel Select metric

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

Conditions

 Static Use a value as a threshold 	• Static Use a value as a threshold		old
Whenever EstimatedCharge Define the alarm condition)s is		
 Greater > threshold 	Greater/Equal	C Lower/Equal	C Lower < threshold
han Define the threshold value			
50	USD		
fust be a number			

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 Servi	ce) 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 alphan	umeric characters, hyphens (-) and und	ierscores ().	
Email endpoints that will receive the Add a comma-separated list of email addr		subscription to the topic above.	
user@example.com			
user1@example.com, user2@example.co	m		
Create topic			

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

Add a description

Name and description

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



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 \rightarrow Storage.

In the S3 console, click on Create Bucket.

Buckets (3)	C Opy ARN Empty Delete	Create bucket
Buckets are containers for data stored in S3. Learn n		

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 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.

mazon S3 > Create bucket
Create bucket suckets are containers for data stored in S3. Learn more 🔀
General configuration
Bucket name
cse6242-gburdell3-logging
Bucket name must be unique and must not contain spaces or uppercase letters. See rules for bucket naming 🔀
AWS Region
US East (N. Virginia) us-east-1
Copy settings from existing bucket - optional Only the bucket settings in the following configuration are copied. Choose bucket

- 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)

nazon S3	> Create bucket	
	bucket ontainers for data stored in S3. Learn more	
Genera Bucket na	l configuration	
cse6242	-gburdell3	
Bucket nam	e must be unique and must not contain spaces or uppercase letters. See rules for bucket naming 🔀	
AWS Regi	n	
US East	(N. Virginia) us-east-1	
Only the bu	ngs from existing bucket - <i>optional</i> cket settings in the following configuration are copied. e bucket	

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.

Amazon S3	cse6242-gburd	ell3			
cse624	2-gburde	ell3			
Objects	Properties	Permissions	Metrics	Management	Access Points

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

Server access logging Log requests for access to your bucket. Learn more	Edit
Server access logging Disabled	

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

Click Save Changes

dit server access logging	
Server access logging Log requests for access to your bucket. Learn more	
Server access logging Disable Enable	
A By enabling server access logging, S3 console will automatically update you to include access to the S3 log delivery group.	ur bucket access control list (ACL)
Target bucket	
s3://cse6242-gburdell3-logging	Browse S3
Format: s3://bucket/prefix	
	Cancel 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 <u>here</u>.

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

Amazon EMR		Notebooks	
Clusters Notebooks Git repositories	•		alyze data interactively with live code, narrative text, visualization: clusters and Amazon S3 apply. Learn more 🔀 Open in JupyterLab Open in Jupyter Start S
Security configurations		Filter: All notebooks	
Block public access		Name	
VPC subnets		🔵 🌒 test5	
Events		test4	
Help		muchark?	
What's new		pyspark3	
		test2	

- 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."

Note: It's okay if the release version in screenshot doesn't match.

- c. For <u>instance type</u>, select **m5.xlarge** (This will likely be the default). You can also change the <u>number of instances</u> 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 <u>AWS service role</u>, 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 <u>Notebook location</u>, 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, choos	e a cluster or create o	ne, and customize configuration options	if desired. Learn more 🗹	
Notebook name*	pyspark-notebook			
		in alphanumeric characters, hyphens (-), or u	underscores (_).	
Description]		
	256 characters max.			
Cluster*	Choose an existi	ng cluster		
	Create a cluster	0		
	Cluster name:	NotebookCluster		
	Release:	emr-5.32.0		
	Applications:	Hadoop, Spark, Livy, Hive, JupyterEnterpriseGateway		
	Instance:	4 m5.xlarge		
	EMR role:	EMR_DefaultRole 🔀 🚯		
	EC2 instance profile:	EMR_EC2_DefaultRole 🖸 🚯		
	EC2 key pair:	No key pairs found 💙 🕚		
Security groups	OUse default secu	rity groups 🚯		
	Choose security	groups (vpc-c8fe2fb5)		
AWS service role*	EMR_Notebooks_De	efaultRole V		
Notebook location*	Choose an S3 locati	on where files for this notebook are save	d.	
	Use a location th	at EMR creates 0		
	Choose an existi	ng S3 location in us-east-1		
	s3://cse6242-gbu	rdell3/	5	
▶ Git repository	Link to a Git reposite	bry		
Fags				
* 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.

	+		±	C	\$ [÷]	🛙 Launcher	
	m /						
0	Name			*	Last Modified		
	🖪 pyspark-notebo	ook.ipynb			2 minutes ago		Note

- 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.

Python 3	b" ✓
Start Preferred Kernel	
Python 3	
Use No Kernel	
No Kernel	
Start Other Kernel	
PySpark	
Spark	
SparkR	
Use Kernel from Preferred Set Use Kernel from Other Session pyspark-notebook.ipynb	
test5-Copy1.ipynb	

HW3 - Q3 [35 pts]

- 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.

PySpark O

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

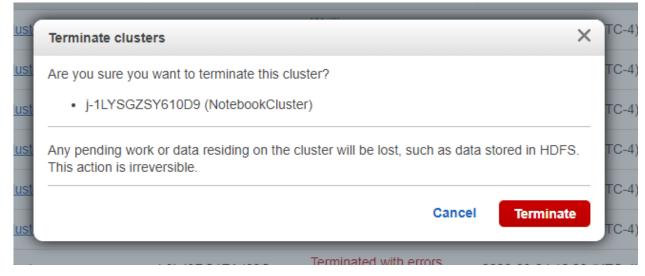
👔 EMR – AWS Console	x +	- 🗆 ×
\leftrightarrow \rightarrow C \oplus console.	aws.amazon.com/elasticmapreduce/home?region=us-east-1#notebook-details:e-2OHP6PJX8ODPI3IFZ9T49UBGZ	🖈 🛡 🖬 O 🌟 🚺 E
aws Services ▼		▼ N. Virginia ▼ Support ▼
Amazon EMR Clusters Notebooks Git repositories Security configurations Block public access	Notebook: pyspark10 Ready Notebook is ready to run jobs on cluster j-1LYSGZSY610D9. Open in JupyterLab Open in Jupyter Stop Delete Notebook Notebook ID: Description: Last modified: Description:	c
VPC subnets Events Help What's new	Last modified by: Created on: Created by: Service IAM role: Security groups for master instance: Security groups for notebook instance: Notebook tags: Notebook location:	

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".

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

育 EMR – AWS Console ← → C	x + s.amazon.com/elasticmapreduce/home?region=us-ea	ist-1#cluster-list:			\$	V 🖬 O 🛪 D
aws Services V			ې voc	startsoft/user	. •	N. Virginia 🔻 Support 🔻
Amazon EMR Clusters	Write and debug your Spark applications in real tim	Terminate				
Notebooks Git repositories Security configurations	Filter: All clusters V Filter clusters Name	ID 11 cluster	s (all loaded) C	Creation time (UTC-4) 👻	Elapsed time	Normalized instance hours
Block public access	NotebookCluster	J-1LYSGZSY610D9	Waiting Cluster ready	2020-09-27 20:40 (UTC-4)	1 hour	0
VPC subnets Events	NotebookCluster	j-30L43FLF03GR5	Terminated User request	2020-09-24 16:53 (UTC-4)	12 minutes	32
Help	NotebookCluster	j-1N3BWYW0YYBFX	Terminated with errors Internal error	2020-09-24 16:44 (UTC-4)	3 minutes	0
What's new	NotebookCluster	j-27LR3KDFZQHES	Terminated User request	2020-09-24 16:28 (UTC-4)	35 minutes	8
	NotebookCluster	j-3FEJHELA97WM0	Terminated with errors Validation error	2020-09-24 16:24 (UTC-4)	41 seconds	0
	NotebookCluster	j-24PN9NHH6YZ36	Terminated with errors Validation error	2020-09-24 16:21 (UTC-4)	1 minute	0
	NotebookCluster	j-3LJ3RC1EAJ02Q	Terminated with errors Validation error	2020-09-24 16:20 (UTC-4)	1 minute	0
	NotebookCluster	j-1K6C9BA2K7RHC	Terminated User request	2020-09-23 21:12 (UTC-4)	14 minutes	8
	NotebookCluster	j-26YZIDKBW3Y76	Terminated with errors Validation error	2020-09-23 20:35 (UTC-4)	1 minute	0
	NotebookCluster	j-KNY19MZUH6S3	Terminated with errors Validation error	2020-09-23 20:33 (UTC-4)	1 minute	0
Feedback English (US) 🔻				008 - 2020. Amazon Web Services. Inc. or it		Privacy Policy Terms of U

3. In the popup, select 'Terminate'.

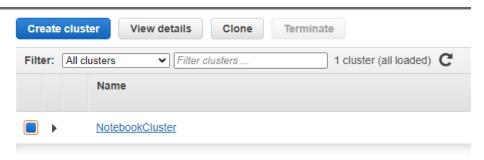


You have now closed all your clusters and will no longer be accruing charges!

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.



- 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.