

#### Spring 2023 Setup Guide [For Q3]

#### **Getting Started**

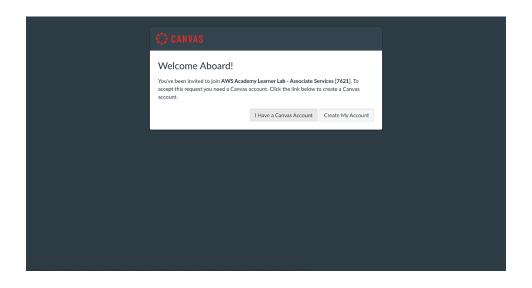
A video tutorial has been created to walk you through the steps 2-7 in this document. View it here: <u>https://youtu.be/QyE2d-9IARc</u>

#### 1. Create an AWS Academy account

You will receive an email from <u>notifications@instructure.com</u> 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.

Cour	se Invitation
Ū	Some content in this message has been blocked because the sender isn't in your Safe senders list. I trust content from notifications@instructure.com.   Show blocked content
AA	AWS Academy < notifications@instructure.com>       Image: I
	You've been invited to participate in a class at <mark>AWS</mark> Academy . The class is called AWS Academy Learner Lab - Associate Services [7621]. Course role: Student
	Name: <b>rbakhtiani3@gatech.edu</b> Email: <b>rbakhtiani3@gatech.edu</b> Username: <b>none</b>
	You'll need to register with Canvas before you can participate in the class.
	Get Started
	Reply Forward

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: <u>https://awsacademy.instructure.com/login/canvas</u>

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

į	academy
	Email Assimitsvi3@getech.edu
P	Password
	Stay signed in Log In
	iicite Prinavyaalay Acceptate Uos Foldy Bacebook Totiter

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):

aws	Notifications. Tell us how and w     Notification Preferences	when you would like to be notified of even	ts in Canvas.		Coming Up 📳 View Calendar Nothing for the next week
Account	Dashboard			:	Recent Feedback Nothing for now
Dashboard					View Grades
Courses	:	:	:		
Calendar					
E Inbox	AWS Academy Learner Lab - Asso ALLAv1-7621	AWS Academy Cloud Foundations ACFv2EN-Educator-3781	AWS Academy Introduction to Clo AICv1Sem1EN-4354-Educator		
History	ą	4 Ø D	-tip		
? Help					

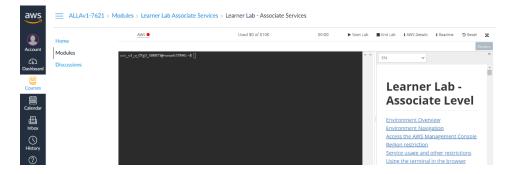
Click on the "AWS Academy Learner Lab" course button. You should see the following screen (the course homepage).

aws	ALLAv1-7621		
	Home		i View Course Stream
Account	Modules		I View Course Calendar
Dashboard	Discussions		Q View Course Notifications
Courses			To Do
			Nothing for now
Calendar			Recent Feedback
		<b>Y</b>	Nothing for now
		AWS Academy Learner Lab - Associate Services provides a long-running sandbox environment for ad-hoc exploration of AWS services. Access to the AWS resources set up in this environment is retained for the duration	
History		of this course. Your budget (\$100) is limited so you should exercise caution to prevent charges that will deplete	
⑦ Help		your budget too quickly. If you exceed your budget, you will lose access to your environment and lose all of your work.	
		Each session lasts for 4 hours by default, although you can extend a session to run longer by pressing the start	
		button to reset your session timer. At the end of each session, any resources you created in the account will be preserved. Some AWS resources, such as EC2 instances, may be automatically shut down, while other resources,	
		such as RDS instances will be left running. Keep in mind that some AWS features cannot be stopped and can still	
		incur charges. For example, an Elastic Load Balancer or a NAT. You may wish to delete those types of resources and recreate them as needed to test your work during a session. You will have access to this environment for the	
		duration of the class you are enrolled in. When the class ends, your access to the learner lab will also end.	
_I←		Get Started	

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

	Cover Letter - Go H Caree		🗄 Reading List
Account CO Dashboard	Home Modules Discussions		Collapse All
Courses			
Inbox U History (?) Help		ぷ End of Course Feedback Survey	
14-			

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.

AWS ● Used \$0 of \$100 03:15 ► Start Lab ■ End Lab i AWS Details i Readme C Reset 🗙

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

	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 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)**. <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.

- 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**.

Billing alarms (0) Hide Auto Scaling alarms Clear selection C: Create composite alarm Act	tions V Create alarm
Q. Search Any state V Any type V Any actions	
Name v State v Last state update v Conditions	Actions
Amazon ClaudWatch can help you monitor the charges on your AMS hill 28 by sending you small alerts when charges avoid a th	breshold you define
	Q. Search Any state V Any type V Any actions

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

Specify metric and conditions	Specify metric and condition	ns
itep 2	Metric	Edit
Configure actions	<b>Graph</b> This alarm will trigger when the blue line goes above the red	line for 1 datapoints within 6 hours.
dd name and lescription	No unit	Namespace
tep 4 review and create	1	AWS/Billing Metric name EstimatedCharges
	0.5	Currency USD
	0	Statistic Q. Maximum
		Period
		6 hours

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**.

Threshold type			
• Static Use a value as a thresh	old	<ul> <li>Anomaly detection</li> <li>Use a band as a thresh</li> </ul>	
Whenever EstimatedChar Define the alarm condition.	ges is		
• Greater > threshold	Greater/Equal >= threshold	Lower/Equal <= threshold	C Lower < threshold
<b>than</b> Define the threshold value.	_		
50 Must be a number	USD		

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 **Next**.

Alarm state trigger Define the alarm state that will trigger thi	s action.	Remove
• In alarm The metric or expression is outside of the defined threshold.	OK The metric or expression is within the defined threshold.	<ul> <li>Insufficient data</li> <li>The alarm has just started or not enough data is available.</li> </ul>
Send a notification to the following	•	
Define the SNS (Simple Notification Servi	ce) topic that will receive the notification.	
<ul> <li>Select an existing SNS topic</li> </ul>		
Create new topic		
	counts	
<ul> <li>Create new topic</li> <li>Use topic ARN to notify other ac</li> </ul>	counts	
<ul> <li>Create new topic</li> <li>Use topic ARN to notify other ac</li> <li>Create a new topic</li> </ul>	counts	
<ul> <li>Create new topic</li> <li>Use topic ARN to notify other ac</li> <li>Create a new topic</li> </ul>	counts	
Create new topic Use topic ARN to notify other ac Create a new topic The topic name must be unique. Notify-Me		s()
Create new topic Use topic ARN to notify other ac Create a new topic The topic name must be unique. Notify-Me	counts	s (_).
Create new topic Use topic ARN to notify other ac Create a new topic The topic name must be unique. Notify-Me SNS topic names can contain only alphane Email endpoints that will receive the	umeric characters, hyphens (-) and underscore:	_
Create new topic Use topic ARN to notify other ac Create a new topic The topic name must be unique. Notify-Me SNS topic names can contain only alphane Email endpoints that will receive the	umeric characters, hyphens (-) and underscore:	_
Create new topic Use topic ARN to notify other ac Create a new topic The topic name must be unique. Notify-Me SNS topic names can contain only alphane Email endpoints that will receive the	umeric characters, hyphens (-) and underscore:	_

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

dd name and description			
Name and description			
Alarm name			
Cost Exceeded \$50			
Alarm description - optional			
Alarm description			
Up to 1024 characters (0/1024)			
	Cancel	Previous	Next

7. On the **Preview and create** screen, scroll to the bottom click Create Alarm

	Cance	ncel	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  $\rightarrow$  Storage.

In the S3 console, click on Create Bucket.

Buckets (3)		C 🗇 Copy ARN	Empty	Delete	Create bucke
Buckets are containers for d	ata stared in CZ Learn mar				31

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.

nazon S3	Create bucket
	e bucket
<b>Gener</b> Bucket r	al configuration
cse62	42-gburdell3-logging
Bucket na	ame must be unique and must not contain spaces or uppercase letters. See rules for bucket naming 🔀
AWS Re	gion
US Eas	st (N. Virginia) us-east-1
Only the	ttings from existing bucket - optional bucket settings in the following configuration are copied. ose 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)

Amazon S3 > Create bucket	
reate bucket	
ckets are containers for data stored in S3. Learn more 🔀	
General configuration	
Bucket name	
cse6242-gburdell3	
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	

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-gburdell3	
cse6242-gburdell3	
Objects Properties Permissions Metrics Management Access Points	
6 Secold down to <b>Server Assess Legging</b> and disk Edit	
6. Scroll down to <b>Server Access Logging</b> and click <b>Edit.</b>	
Server access logging Log requests for access to your bucket. Learn more 🔀	Ed

Server access logging			
Disabled			

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

Click Save Changes

Edit server access logging			
Server access logging Log requests for access to your bucket. Learn more			
Server access logging			
O Disable			
O Enable			
▲ By enabling server access logging, S3 console will automatically update ye to include access to the S3 log delivery group.	our bucket	access con	trol list (ACL)
Target bucket			
s3://cse6242-gburdell3-logging		Brows	e <b>S</b> 3
Format: s3://bucket/prefix			
	-	ancel	Save changes
	C	ancel	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 left menu. Click "Create Notebook".

Amazon EMR	EMR Serverless is now GA With EMR Serverless, get th provisioning, quick job start	e benefits of Amazon El							e for popular frameworks along with easy	×
EMR Studio										
EMR Serverless 🗹 New	Notebooks									
EMR on EC2	Use EMR notebooks based on J	upyter to analyze data ir	teractively with live	code, narrative tr	ext, visualiza	ations, and	more. Create	and attach note	books to Amazon EMR clusters running	
Clusters	Hadoop, Spark, and Livy. Noteb	ooks run free of charge a	and are saved in Ama	zon S3 indepen	dently of clu	isters. Star	idard billing fo	or clusters and A	mazon S3 apply. Learn more	
Notebooks	Create notebook View	details Open in Ju	upyterLab Ope	en in Jupyter	Start	Stop	Delete			
Git repositories	Filter: All notebooks v Filter	r notebooks	0 notebooks (al	loaded) C						
Security configurations	Name		Status	Cluster		c	reation time	UTC-4)	Last modified A	
Block public access										
VPC subnets										

- 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 <u>name</u>. It can be anything you want.
  - b. Select the checkbox to "Create a cluster."
  - c. For Instance, choose m5.xlarge and change 1 to 4.

#### Name and configure your notebook

N-1-1	
Notebook name*	hw3-notebook
	Names may only contain alphanumeric characters, hyphens (-), or underscores ().
Description	
	256 characters max.
Cluster*	Choose an existing cluster
	Create a cluster ()
	Cluster name: NotebookCluster
	Release: emr-5.36.0
	Applications: Hadoop, Spark, Livy, Hive, JupyterEnterpriseGateway
	Instance: 4 m5.xlarge ~
	EMR role EMR_DefaultRole [2] Use EMR_DefaultRole_V2
	EC2 instance profile EMR_EC2_DefaultRole [7]
	EC2 key pair: Proceed without an EC2 key   >
	Auto-termination
	Terminate cluster when it is idle after $\fbox{1}$ $\checkmark$ hours $\fbox{0}$ $\checkmark$ minutes
Security groups	Use default security groups 1
	Choose security groups (vpc-085b13e55afd1ff54)

Name your notebook, choose a cluster or create one, and customize configuration options if desired. Learn more

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

AWS service role*	LabRole V 0	
	● Make sure this role has the required permissions. Learn more Z	
Notebook location*	Choose an S3 location where files for this notebook are saved.	
	Use a location that EMR creates	
	Choose an existing S3 location in us-east-1	
	s3://cse6242-gburdell3/	
<ul> <li>Git repository</li> </ul>	Link to a Git repository	
► Tags 🚯		
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**.

# Notebook: hw3-notebook Ready Work

**Open in JupyterLab** 

**Open in Jupyter** 

Stop

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.

	+ 12	<b>±</b>	G	\$ <sup>*</sup>	🖸 Launcher	
0	Name			Last Modified		
	📕 pyspark-notebook.i	oynb		2 minutes ago	No	t

- 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.

e cu		1
	Select Kernel	
•	Start Preferred Kernel	}
	✓ PySpark	
	Python 3	
	Use No Kernel	
	No Kernel	
	Start Other Kernel	
;()	Spark	
her ⊁⊡⊡	× ■ q3_pyspark.ipynb ×	

#### HW3 - Q3 [35 pts]

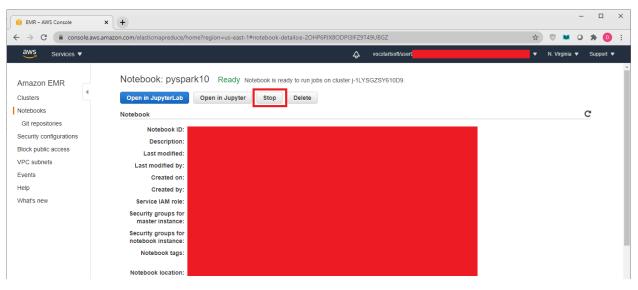
- 5. Run the first code cell which will import PySpark, followed by a cell which should contain **sc** to start the Spark Application so you can start programming the assignment.
- 6. Once you have finished coding, right click on the file name in the directory on the left and select download to download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

PySpark ()

# 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.

EMR - AWS Console × + C  C  C  C  C  C  C  C  C  C  C  C  C						
→ C 🔒 console.aw	A'S Console x + + + + + + + + + + + + + + + + + +					
Services <b>v</b>	x       +         amazon.com/elasticmapreduce/home?region=us-east-1#cluster/list:       x					
	Write and debug your Spark applications in real tim	e using EMR Notebooks.	Learn more			
mazon EMR	Create cluster View details Clone	Terminate				
otebooks	Filter: All clusters	11 cluster	s (all loaded)			
it repositories curity configurations	Name	ID	Status	Creation time (UTC-4) 👻	Elapsed time	
ck public access	NotebookCluster	j-1LYSGZSY610D9		2020-09-27 20:40 (UTC-4)	1 hour	0
C subnets	NotebookCluster	j-30L43FLF03GR5		2020-09-24 16:53 (UTC-4)	12 minutes	32
lp	NotebookCluster	j-1N3BWYW0YYBFX		2020-09-24 16:44 (UTC-4)	3 minutes	0
at's new	NotebookCluster	j-27LR3KDFZQHES		2020-09-24 16:28 (UTC-4)	35 minutes	8
	NotebookCluster	j-3FEJHELA97WM0		2020-09-24 16:24 (UTC-4)	41 seconds	0
	NotebookCluster	j-24PN9NHH6YZ36		2020-09-24 16:21 (UTC-4)	1 minute	0
	NotebookCluster	j-3LJ3RC1EAJ02Q		2020-09-24 16:20 (UTC-4)	1 minute	0
	NotebookCluster	j-1K6C9BA2K7RHC		2020-09-23 21:12 (UTC-4)	14 minutes	8
	NotebookCluster	j-26YZIDKBW3Y76		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

3. In the popup, select 'Terminate'.

ust	Terminate clusters X	TC-4
<u>ust</u>	Are you sure you want to terminate this cluster?	TC-4
lust	j-1LYSGZSY610D9 (NotebookCluster)	TC-4
<u>ust</u>	Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.	TC-4
ust ust	Cancel Terminate	тс-4) тс-4)
	Terminated with errors	

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.

Create	cluster	View details	Clone	Terminate		
Filter:	All clusters				1 cluster (all loaded)	
	Nan	1e				
•	Note	bookCluster				

- 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.

### Troubleshooting

1. If you get an error saying your credentials are invalid, when logging in



Click the logout link, and then try again. If that still does not work, try clearing your cache and cookies in your browser. You can also use Incognito mode in your browser or a completely different browser if none of those options work.

- Cluster status: Terminated with errors. The requested instance type m5.xlarge is not supported in the requested availability zone. Learn more at <u>https://docs.aws.amazon.com/console/elasticmapreduce/ERROR\_noinstancetype</u>
- To fix: choose a different instance type with similar capabilities. You do not have to use the suggested instance type. It is just a suggestion.