

Fall 2021 Setup Guide [For Q3]

Getting Started

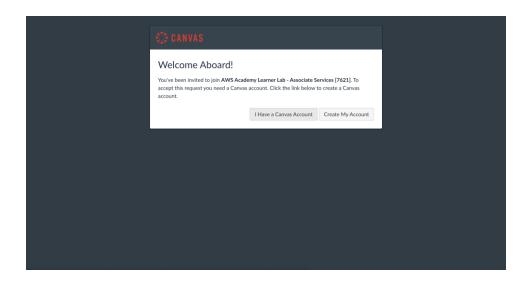
A video tutorial has been created to walk you through the steps 2-7 in this document. View it here: https://youtu.be/QyE2d-9IARc

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.

Coui	rse Invitation
()	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: Image: Im
	You've been invited to participate in a class at AWS Academy . The class is called AWS Academy Learner Lab - Associate Services [7621]. Course role: Student Name: rbakhtiani3@gatech.edu Email: rbakhtiani3@gatech.edu Username: none 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://www.awsacademy.com/LMS_Login</u>.

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



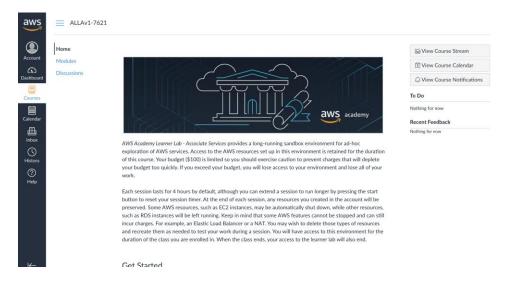
Click the "Student Login" button. Then, you will see the following screen:

aws academy
Email rbakhtiani3@gatech.edu
Password
Stay signed in Log In Forgot Password2
Hide Ethnoroutino decentation.BoocEulor Energiese Tantano SINSTRUCTURE

Please enter in 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 a Notification Preferences	when you would like to be notified of event	ts in Canvas.		Coming Up lothing for the next	Xiew Calendar
Account	Dashboard			:	ecent Feedback	τ.
Dashboard					View Grades	
Courses	:	:	:			
Calendar						
Enbox	AWS Academy Learner Lab - Asso ALLAv1-7621	AWS Academy Cloud Foundations ACFv2EN-Educator-3781	AWS Academy Introduction to Clo AICv1Sem1EN-4354-Educator			
Listory	φ.	4 p	ڻي ا			
?						

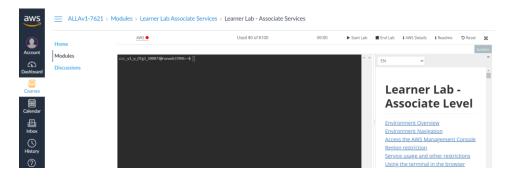
Click on the "AWS Academy Learner Lab – Associate Services" course button. You should see the following screen (the course homepage).



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

aws	Cover Letter - Go 🔢 Care		🗄 Reading List
Account	Home Modules		Collapse All
CS Dashboard	Discussions	Learner Lab Associate Services	
Courses		Learner Lab - Student Guide.pdf	
Inbox		Course Feedback Survey	
History ② Help			

Click on the "Learner Lab – Associate Services" 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." Then, click on "AWS" tab in the top left. A new window will open. You are now in the AWS management console! It will look something like this [right top corner].

vocstartsoft/use	er67 . 🔺	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 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. 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 "**In Alarms**"; then, in the **Alarms** pane, click **Create Alarm**.

aws	Services	*	Resource Groups ~ 🖌	7	DanF 👻	N. Virginia 👻	Support
CloudWatch Dashboards Alarms ALARM INSUFFICIENT	•		We listened to your feedback! In response to your comments on usability, we enhanced our user interface. You get to preview our improvem before we roll it out to everyone else. <u>Switch to your original interface</u>	ents	Se	nd us feedback	×
OK Billing	3	Cloud	Watch > Alarms				
Events		AI	arms (6)	A	ction 🔻	Create al	larm

3. Click the "Select Metric" button.

aws s	Services 🗸 Resource Groups 🤟 🛧	众 DanF ◄
CloudWatch > Alarm	ns > Create alarm	
Step 1 Specify metric and conditions	Specify metric and conditions	
Step 2	Metric	
Configure actions	Graph	
Step 3 Add a description	Preview of the metric or metric expression and the alarm threshold. Select metric	
Step 4 Preview and create	Cancel	Next

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

Select metric		×
Untitled graph 🥜	th 3h 12h 1d 3d 1w custom - Line	• 3 •
1		
1	Your CloudWatch graph is empty.	
0.5	Select some metrics to appear here.	
0		
19:00 19:15 19:30	19:45 20:00 20:15 20:30 20:45 21:00 21:15 21:30 21:	:45
All metrics Graphed metrics Graph	options Source	
Q. Search for any metric, dimension or resource	ce id	Graph search
615 Metrics		
Billing	EBS EC2	÷.
11 Metrics	324 Metrics 168 Metrics	
		*
	Cancel	Select metric
Select metric		×
Untitled graph 🥒	1h 3h 12h 1d 3d 1w custom - Line	• 0 •
Seedly		
1	Your CloudWatch graph is empty.	
0.5	Select some metrics to appear here.	
0		
19:00 19:15 19:30	19-45 20:00 20:15 20:30 20:45 21:00 21:15 21:30 21:	:45
All metrics Graphed metrics Graph	options Source	
All > Billing Q. Search for any metric, dime	insion or resource id	Graph search
11 Metrics		
By Service	Total Estimated Charge	
10 Metrics	1 Metric	
	Cancel	Select metric
Select metric		×
Untilled graph	1h 3h 12h 1d 3d 1w custom - Line	• 2 •
Untitled graph 🥒	In Sin 121 10 50 1W COSON - Line	
No unit		
0.5		
0		
18:00 18:15 18:30 18:45	18:00 19:15 18:30 19:45 20:00 20:15 20:30 20:45 21:00 21:15 21:30	21:45
All metrics Graphed metrics (1) Gra	aph options Source	
		Graph search
Currency (1)	Coercin for any metric, cirmension or resource id	- opri sesseri
USD 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 	d	Anomaly detection Use a band as a thresho	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		
/lust 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 a	iction		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	e) topic that will receive the notification	1	
 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 alphanu	meric characters, hyphens (-) and unc	derscores (_).	
Email endpoints that will receive the Add a comma-separated list of email addre		a subscription to the topic above.	
user@example.com			
user1@example.com, user2@example.com	n		
Create topic			

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

Add a description

Name and description

Cancel	Previous	Next
	Cancel	Cancel Previous

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 D Copy ARN Empty Delete	Create bucket
Buckets are containers for data stored in S3. Learn more	[2]	

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	
Lreate bucket uckets 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	
reate bucket	
ckets are containers for data stored in S	3. Learn more 🛃
General configuration	
Bucket name	
cse6242-gburdell3	
Bucket name must be unique and must not co	ontain spaces or uppercase letters. See rules for bucket naming 🔀
AWS Region	
US East (N. Virginia) us-east-1	▼
Copy settings from existing bucket - op	tional
Only the bucket settings in the following cont	
Choose bucket	
SHOUSE BUCKEL	

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	lell3			
cse624	2-gburde	ell3			
Objects	Properties	Permissions	Metrics	Management	Access Points

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

dit 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 updat to include access to the S3 log delivery group. Target bucket	te your bucket access control list (ACL)
s3://cse6242-gburdell3-logging	Browse S3
Format: s3://bucket/prefix	
	Cancel Save chang

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		Notebooks
Clusters	•	Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations
Notebooks		independently of clusters. Standard billing for clusters and Amazon S3 apply. Learn more 🗹
Git repositories		Create notebook View details Open in JupyterLab Open in Jupyter Start
Security configurations		Filter: All notebooks V Filter notebooks 7 notebooks (all loaded)
Block public access		Name
VPC subnets		etest5
Events		test4
Help		
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 <u>name</u>. It can be anything you want.
 - b. Select the checkbox to "Create a cluster."

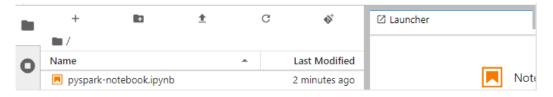
Note: It's okay if the release version in screenshot doesn't match.

- 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 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".

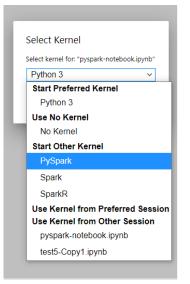
5. Get started with the skeleton

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.



ner ×	🖪 q3_pyspark.ipynb	×	
% □ □ ▶ ■	⊖ Markdown ∨	PySp	park 🔿

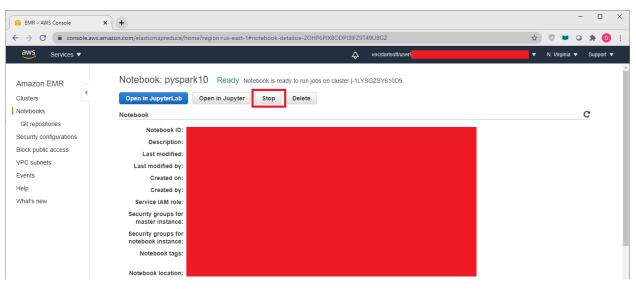
HW3 - Q3 [35 pts]

- 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.

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	× +					- 0	×
\leftrightarrow \rightarrow C \cong console.aw	s.amazon.com/elasticmapreduce/home?region=us	-east-1#cluster-list:			\$	V 🖬 O 🛸	•
aws Services ▼			¢ voc	startsoft/user	. 🔻	N. Virginia 🔻 Suppor	rt 🔻
	Write and debug your Spark applications in real	time using EMR Notebooks.	Learn more				
Amazon EMR	Create cluster View details Clor	Terminate					
Clusters	Filter: All clusters	11 cluster	s (all loaded)				
Git repositories Security configurations	Name	ID	Status	Creation time (UTC-4) 👻	Elapsed time	Normalized instance hours	s
Block public access	NotebookCluster	j-1LYSGZSY610D9	Waiting Cluster ready	2020-09-27 20:40 (UTC-4)	1 hour	0	
VPC subnets	NotebookCluster	j-30L43FLF03GR5	Terminated User request	2020-09-24 16:53 (UTC-4)	12 minutes	32	
Events 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 i	ite affiliates. All rights research	Privacy Policy Terms	s of Use

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 cl	usters V Filter clusters 1 cluster (all loaded)			
	Name			
	NotebookCluster			

- 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.