

# Using AWS Educate For CSCE 676

This document is a quick guide on how to apply for an AWS Educate account, create a Hadoop/Spark cluster, create an S3 storage bucket and how to create a Jupyter notebook to run code against the created cluster.

Thanks to Amazon for providing free credits for this course!

## Table of Contents

Table of Contents	1
Applying to AWS Educate	2
Creating A Hadoop/Spark Cluster	5
Accessing the Cluster with SSH	13
Accessing Hadoop/Spark Web Interfaces	14
Accessing the cluster with a Jupyter Notebook	15
Creating An S3 Bucket	18
Warning! Be a Terminator!	19
Running Homework Jupyter Notebook	20
Warning: Save homework2.ipynb before termination!	20

# Applying to AWS Educate

To apply to join AWS Educate, follow the URL given in class and choose "Texas A&M University" and "Data Mining and Analysis - CSCE 676" course from the corresponding dropdown menus as show below. Click on an email confirmation link that will be sent to you. Your application should be approved within an hour.

**aws educate**  
Apply to join AWS Educate

**Step 2/3: Tell us about yourself**

Preferred Language: English

Texas A&M University  
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

Country

First Name

Last Name

Email  
Please provide a valid, current email issued by your institution. Example: your\_name@your\_school.edu

Graduation Month Graduation Year

Birth Month Birth Year

Promo Code (optional)

[Frequently Asked Questions](#)

**PLEASE MAKE SURE YOUR INSTITUTION IS CORRECT.**

Please click the box below to help assure that a person and not an automated program is submitting this application. If a set of letters is displayed enter them on the line. If you have any difficulty with the letters, you can click the reload icon to get a new set of letters, or click the headphones to hear audio of what to enter.

☐ I'm not a robot

RECAPTCHA  
Privacy - Terms


Please note that any personal information you provide will be treated in accordance with the [AWS Educate Terms and Conditions](#) and [AWS Privacy Notice](#)

**NEXT**

After your application gets approved, follow the link sent to you to set a password and to access your AWS Educate account. Once you have created your account and are able to access the starting page as shown below click on "Use an AWS Educate Starter Account" and then click "Create Starter Account"

awseducate

PortfolioCareer PathwaysBadgesJobs+AWS AccountLogout

Majid Alfifi

Consecutive Days: 1

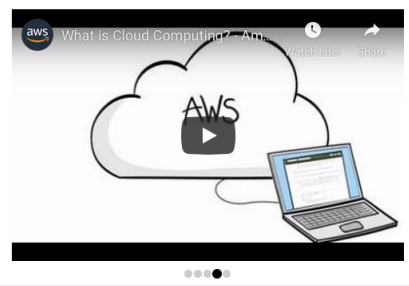
Pathways Completed: 0

Badges Earned: 0

Preferred Language: English

Cloud technology is everywhere, creating over 18 million cloud jobs worldwide (source: Wanted Analytics). AWS Educate introduces you to lucrative cloud-enabled careers through more than 25 learning pathways, each with content from industry professionals, learning activities and labs, opportunities to earn AWS Educate Badges and Certificates of Completion, and access to the AWS Educate Job Board. Coupled with courses at your school or through online providers, AWS Educate puts you on the pathway to your dream job in the clouds.

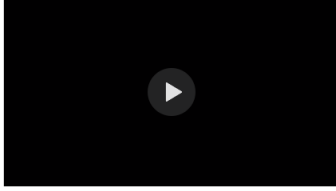
Begin your journey today!



Access AWS Services


[Use a personal AWS Account →](#)


[Use an AWS Educate Starter Account →](#)





Which option is right for me?

[Learn More →](#)

Portfolio

Career Pathways

Learn

Badges

Suggested Jobs

awseducate

PortfolioCareer PathwaysBadgesJobs+AWS AccountLogout



## I'd like to use an AWS Educate Starter Account


Choose an AWS Educate Starter Account to receive access to an AWS account with a preset limit on your spend on AWS services. An AWS Educate Starter Account is run and managed by a third party (Vocareum, Inc.) and the Starter Account runs in the Vocareum's environment on AWS. Starter Accounts are subject to a separate agreement between you and Vocareum under separate terms and conditions.

The AWS Educate Starter Account provides access to most but not all AWS services. Students at AWS Educate member institution will receive up to \$75 (US) of AWS credit per year in their AWS Educate Starter Account, and students at non-member institution will receive up to \$30 (US) of AWS credit per year.


You don't need a credit card to use a Starter Account because AWS promotional credits are already available in the account. When your usage of AWS services exceed the balance on the account, the account is closed and any running services or other resources on the account are lost.

Create Starter Account

or choose another option



PortfolioCareer PathwaysBadgesJobsAWS AccountLogout



## AWS Educate Starter Account

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!


[AWS Educate Starter Account](#)

Your account has an estimated **75** credits remaining and access will end on **Sep 19, 2020**.

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.

Clicking on "AWS Educate Starter Account" takes you to the following page which has the access to AWS Console! You can always access your account your email and the password you created above by visiting the following link:

<https://www.awseducate.com/signin>







My ClassesHelpalfifi@tamu.edu

## Welcome to AWS Educate Starter Account

Use your Starter Account to access to a wide variety of AWS Services and start building! Click on the AWS Console button to sign in and get started.

- What AWS services can I use in my Starter Account?
- What regions can I use with a Starter Account?
- Are Service Linked Roles supported?
- I can't start any resources. What happened?

## Your Starter Account Status

	<b>Active</b> full access (alfifi@tamu.edu)
	<b>\$75</b> credits (estimated)
	<b>364d 23:09:45</b> remaining term
	<b>2:56</b> session time

[Account Details](#) [AWS Console](#)

# Creating A Hadoop/Spark Cluster

Click on "AWS Console" on the home page shown above. You may need to allow pop-ups for the AWS Console to open. On the AWS Management Console click on All services under "Find Services" and select EMR (Elastic MapReduce). You can also search for it in the search box.

## AWS Management Console

### AWS services


#### Find Services

You can enter names, keywords or acronyms.

Q Example: Relational Database Service, database, RDS

#### ▼ All services

##### Compute

EC2  
Lightsail   
ECR  
ECS  
EKS  
Lambda  
Batch  
Elastic Beanstalk  
Serverless Application Repository

##### Storage

S3  
EFS  
FSx  
S3 Glacier  
Storage Gateway  
AWS Backup

##### Database

RDS  
DynamoDB  
ElastiCache  
Neptune  
Amazon Redshift  
Amazon QLDB  
Amazon DocumentDB

##### Developer Tools

CodeStar  
CodeCommit  
CodeBuild  
CodeDeploy  
CodePipeline  
Cloud9  
X-Ray

##### Robotics

AWS RoboMaker

##### Blockchain

Amazon Managed Blockchain

##### Satellite

Ground Station



##### Management & Governance

AWS Organizations  
CloudWatch  
AWS Auto Scaling  
CloudFormation  
CloudTrail  
Config  
OpsWorks  
Service Catalog  
Systems Manager

##### Machine Learning

Amazon SageMaker  
Amazon Comprehend  
AWS DeepLens  
Amazon Lex  
Machine Learning  
Amazon Polly  
Rekognition  
Amazon Transcribe  
Amazon Translate  
Amazon Personalize  
Amazon Forecast  
Amazon Textract  
AWS DeepRacer

##### Analytics

Athena  
**EMR**   
CloudSearch  
Elasticsearch Service  
Kinesis  
QuickSight   
Data Pipeline  
AWS Glue  
AWS Lake Formation  
MSK

##### Security, Identity, & Compliance

##### Mobile

AWS Amplify  
Mobile Hub  
AWS AppSync  
Device Farm

##### AR & VR

Amazon Sumerian


##### Application Integration

Step Functions  
Amazon EventBridge  
Amazon MQ  
Simple Notification Service  
Simple Queue Service  
SWF

##### Customer Engagement

Amazon Connect  
Pinpoint  
Simple Email Service

##### Business Applications

Alexa for Business  
Amazon Chime   
WorkMail

This will open the page below. Under EMR, we will be mainly working on two tabs:

- \* Clusters where we will build the Hadoop/Spark cluster, and
- \* Notebooks where we will create a Jupyter notebook to access the cluster and operate on it.

The screenshot shows the Amazon EMR console interface. On the left, the navigation menu has 'Clusters' and 'Notebooks' highlighted. The main area displays a welcome message and a 'Create cluster' button. Below this, a section titled 'How Elastic MapReduce Works' illustrates the workflow: Upload (uploading data to S3), Create (configuring the cluster), and Monitor (checking cluster health). The right sidebar provides additional information and links to various resources.

The steps to create a Hadoop/Spark cluster are as follows:

1. Click on the Create cluster button above.
2. Click on "Go to advanced options"

Create Cluster - Quick Options [Go to advanced options](#)

### General Configuration

Cluster name

☒ Logging

S3 folder

Launch mode ☒ Cluster ☐ Step execution

### Software configuration

Release

Applications ☒ Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.5, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

3. Chose the Hadoop, Hive, and Spark and also copy the following configuration line and paste it under "Edit software settings" a

```
classification=spark-defaults,properties=[spark.jars.packages=graphframes:graphframes:0.7.0-spark2.4-s_2.11]
```

## Create Cluster - Advanced Options [Go to quick options](#)

**Step 1: Software and Steps**

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

### Software Configuration

Release

<input checked="" type="checkbox"/> Hadoop 2.8.5	<input type="checkbox"/> Zeppelin 0.8.1	<input type="checkbox"/> Livy 0.6.0
<input type="checkbox"/> JupyterHub 0.9.6	<input type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.8.0
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 1.4.10	<input type="checkbox"/> Pig 0.17.0
<input checked="" type="checkbox"/> Hive 2.3.5	<input type="checkbox"/> Presto 0.220	<input type="checkbox"/> ZooKeeper 3.4.14
<input type="checkbox"/> MXNet 1.4.0	<input type="checkbox"/> Sqoop 1.4.7	<input type="checkbox"/> Mahout 0.13.0
<input type="checkbox"/> Hue 4.4.0	<input type="checkbox"/> Phoenix 4.14.2	<input type="checkbox"/> Oozie 5.1.0
<input checked="" type="checkbox"/> Spark 2.4.3	<input type="checkbox"/> HCatalog 2.3.5	<input type="checkbox"/> TensorFlow 1.13.1

Multi-master support

☐ Enable multi-master support

AWS Glue Data Catalog settings (optional)

☐ Use for Hive table metadata

☐ Use for Spark table metadata

Edit software settings

☒ Enter configuration ☐ Load JSON from S3

`classification=spark-defaults,properties=[spark.jars.packages=graphframes:graphframes:0.7.0-spark2.4-s_2.11]`

Add steps (optional)

Step type

☐ Auto-terminate cluster after the last step is completed

4. On "Step 2: Hardware" you can use the defaults (1 master node m5.xlarge, and 2 data nodes m5.xlarge). Use these for experimentation and learning but when you are ready to run your code on a bigger dataset, then you can create more core nodes (e.g. 10 nodes) which will cost more of your credits.

- On "Step 3: General Cluster Settings", click on "Bootstrap Actions". From the drop-down menu, choose "Custom action" and then click "Configure and add"

#### ▼ Bootstrap Actions

Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize the cluster.

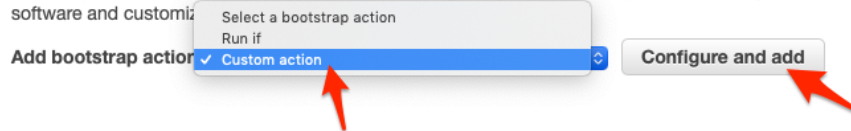
Add bootstrap action

Select a bootstrap action

Run if

✓ Custom action

Configure and add



In the script location type `s3://caverlee/init.sh`

Add Bootstrap Action

×

Bootstrap action type


Custom action

Name

Custom action

Script location

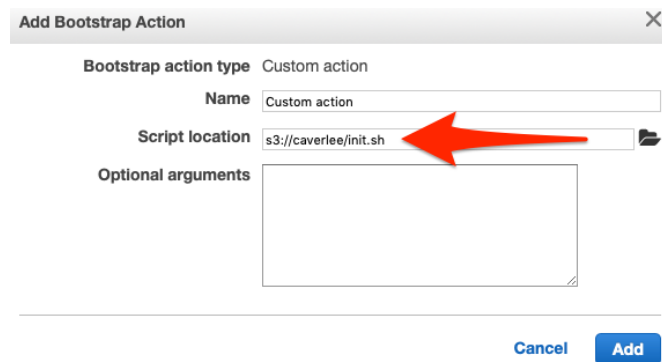
s3://caverlee/init.sh



Optional arguments

Cancel

Add



Click add. This script will install Jupyter on the master node and download "start-jupyter.sh" script to use later.



6. On "Step 4: Security". We need to do two things:

1. Create an EC2 Key Pair, and
2. Allow SSH access.

To create an EC2 Key Pair, click on the link shown and follow instructions.

## Create Cluster - Advanced Options [Go to quick options](#)

[Step 1: Software and Steps](#)

[Step 2: Hardware](#)

[Step 3: General Cluster Settings](#)

**Step 4: Security**

### Security Options

EC2 key pair No key pairs found

☒ Cluster visible to all IAM users in account

#### Permissions

☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#)

EC2 instance profile [EMR\\_EC2\\_DefaultRole](#)

Auto Scaling role [EMR\\_AutoScaling\\_DefaultRole](#)

#### Security Configuration

##### EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#).

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	<a href="#">Create ElasticMapReduce-master</a>	No security groups selected
Core & Task	<a href="#">Create ElasticMapReduce-slave</a>	No security groups selected

[Create a security group](#)

**No EC2 key pair has been selected. You will not be able to SSH to this cluster. [Learn how to create an EC2 Key Pair.](#)**

[Cancel](#)

[Previous](#)

[Create cluster](#)

After you have created the key select it in the EC2 key pair dropdown list:

## Create Cluster - Advanced Options [Go to quick options](#)

[Step 1: Software and Steps](#)

[Step 2: Hardware](#)

[Step 3: General Cluster Settings](#)

**Step 4: Security**

### Security Options

EC2 key pair ✓ Proceed without an EC2 key pair

☒ Cluster visible to all IAM users in account

#### Permissions

☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#)

EC2 instance profile [EMR\\_EC2\\_DefaultRole](#)

Auto Scaling role [EMR\\_AutoScaling\\_DefaultRole](#)

To allow SSH access to your cluster click on "Create a security group".

## Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security**

### Security Options

EC2 key pair

☒ Cluster visible to all IAM users in account

#### Permissions

☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#)

EC2 instance profile [EMR\\_EC2\\_DefaultRole](#)

Auto Scaling role [EMR\\_AutoScaling\\_DefaultRole](#)

#### Security Configuration

##### EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#).

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	<input type="text" value="Create ElasticMapReduce-master"/>	No security groups selected
Core & Task	<input type="text" value="Create ElasticMapReduce-slave"/>	No security groups selected

[Create a security group](#)

Cancel

Previous

Create cluster

Select the "default" security group and choose edit inbound rules as shown below

The screenshot shows the AWS VPC console. On the left, the 'VPC Dashboard' sidebar is visible. The main area shows a list of security groups. The 'default' security group (sg-40c7a011) is selected. A context menu is open over the selected group, with a red arrow pointing to the 'Edit inbound rules' option. Another red arrow points to the 'default' security group in the list.

Then click "Edit Rules" -> "Add Rule" and type **22 0-65535** in the Port Range and choose "Anywhere" for the Source. Click "Save rules". Note there is security risk but since this will live for only few hours after which you will terminate the cluster, it should be fine for this homework.

[Security Groups](#) > Edit inbound rules

### Edit inbound rules

Inbound rules control the incoming traffic that's allowed to reach the instance.

Type	Protocol	Port Range	Source		Description	
Custom TCP...	TCP	0-65535	Anywhere	0.0.0.0/0, ::/0	e.g. SSH for Admin Desktop	<input type="button" value="X"/>

Add Rule

Finally in "Step 4: Security" click on "Additional security groups" links and add the "default" security group you just edited

## Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security**

### Security Options

EC2 key pair [Proceed without an EC2 key pair](#)

☒ Cluster visible to all IAM users in account

#### Permissions

☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#)

EC2 instance profile [EMR\\_EC2\\_DefaultRole](#)

Auto Scaling role [EMR\\_AutoScaling\\_DefaultRole](#)

#### Security Configuration

##### EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#).

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	Default: sg-00ef00b2c1f3eb5a1 (ElasticMapReduce-mast)	<a href="#">sg-4bc7a611</a> (default)
Core & Task	Default: sg-0078f752b600760b2 (ElasticMapReduce-sla)	<a href="#">sg-4bc7a611</a> (default)

[Create a security group](#)

No EC2 key pair has been selected, so you will not be able to SSH to this cluster or connect to HUE (unless you are using a VPN). [Learn how to create an EC2 Key Pair.](#)

Cancel

Previous

Create cluster

7. Now you are ready to click "Create cluster"

## Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security**

### Security Options

EC2 key pair [Proceed without an EC2 key pair](#)

☒ Cluster visible to all IAM users in account

#### Permissions

☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#)

EC2 instance profile [EMR\\_EC2\\_DefaultRole](#)

Auto Scaling role [EMR\\_AutoScaling\\_DefaultRole](#)

#### Security Configuration

##### EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#).

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	Default: sg-00ef00b2c1f3eb5a1 (ElasticMapReduce-mast)	<a href="#">sg-4bc7a611</a> (default)
Core & Task	Default: sg-0078f752b600760b2 (ElasticMapReduce-sla)	<a href="#">sg-4bc7a611</a> (default)

[Create a security group](#)

No EC2 key pair has been selected, so you will not be able to SSH to this cluster or connect to HUE (unless you are using a VPN). [Learn how to create an EC2 Key Pair.](#)

Cancel

Previous

Create cluster

The cluster will take few minutes to start.

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Help

What's new

Clone

Terminate

AWS CLI export

Cluster: My cluster Starting

Summary

Application history

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Connections:

[Enable Web Connection](#) – Spark History Server, Resource Manager ... (View All)

Master public DNS:

ec2-35-171-47-16.compute-1.amazonaws.com [SSH](#)

Tags:

-- [View All / Edit](#)

Summary

ID: j-C95AWF15W9YL

Creation date: 2019-09-20 14:31 (UTC-5)

Elapsed time: 1 minute

Auto-terminate: No

Termination protection: On [Change](#)

Network and hardware

Availability zone: us-east-1f

Subnet ID: [subnet-59504656](#)

Master: Provisioning 1 m5.xlarge

Core: Provisioning 2 m5.xlarge

Task: --

Configuration details

Release label: emr-5.26.0

Hadoop distribution: Amazon 2.8.5

Applications: Hive 2.3.5, Spark 2.4.3

Log URI: s3://aws-logs-718366129819-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled

Custom AMI ID: --

Security and access

Key name: csce676

EC2 instance profile: EMR\_EC2\_DefaultRole

EMR role: EMR\_DefaultRole

Auto Scaling role: EMR\_AutoScaling\_DefaultRole

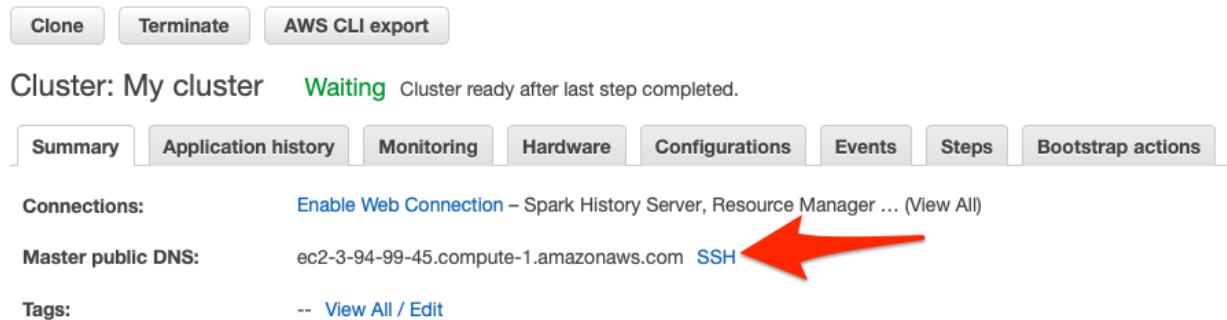
Visible to all users: All [Change](#)

Security groups for [sg-00ef00b2c1f3eb5a1](#)  
Master: (ElasticMapReduce-master)

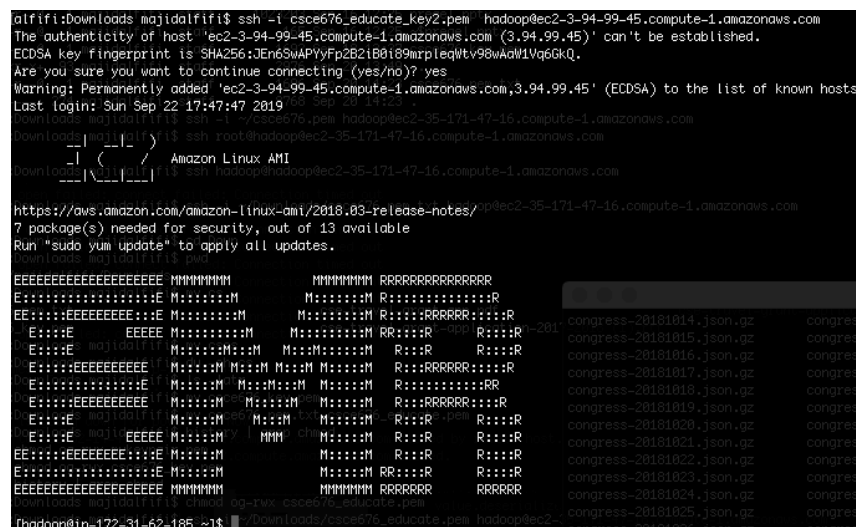
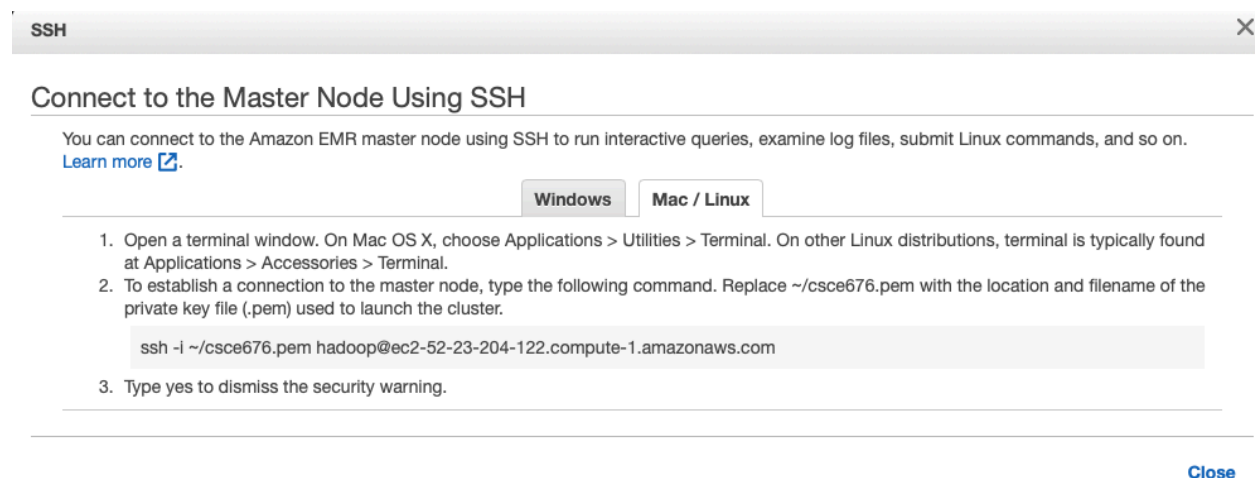
Security groups for [sg-0078f752b600760b2](#)  
Core & Task: (ElasticMapReduce-slave)

## Accessing the Cluster with SSH

You can SSH to the master node (or any of the worker nodes if you allow SSH access) and run any unix commands you like, or investigate logs, and so on. To do so, click on the "SSH" link as shown below



This will open a window like the following that will show you the domain name of the master node so you can ssh to it



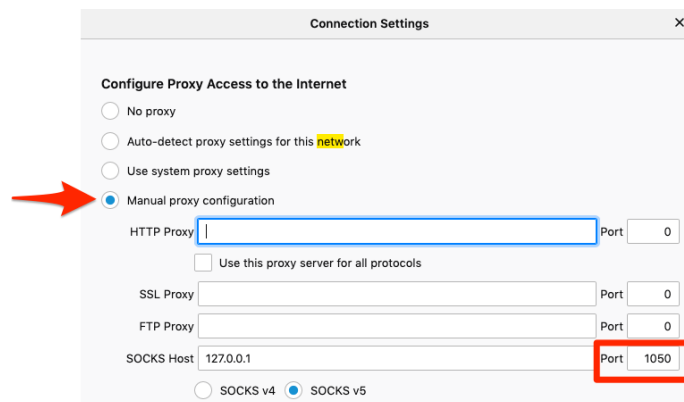
# Accessing Hadoop/Spark Web Interfaces

Note: If you have opened all ports on the servers you can access the web interfaces directly using the public domain name without the steps below.

Hadoop/Spark provide web interfaces to investigate their status and logs. There are couple of ways to access those interfaces but the easiest maybe to use SSH tunnel and SOCKS proxy.

The steps are as follows:

1. When connecting to the master node with SSH, us -D option as follows `ssh -D 1050`  
`ssh -D 1050 -i csce676.pem hadoop@ec2-34-239-180-133.compute-1.amazonaws.com`
2. In your browser, forward all traffic through this port (1050) which makes your browser work as though it was running on the remote Amazon server. I use Firefox for this purpose because it's easy to setup SOCKs and I have this browser dedicated for my cluster work. In Firefox, set the connection settings as follows:



3. Find out the IP of the master node and use it to access the web interfaces shown in the table below. (Hint: you can find the ip after you login in the prompt name. For example, `[hadoop@ip-172-31-62-185] $`

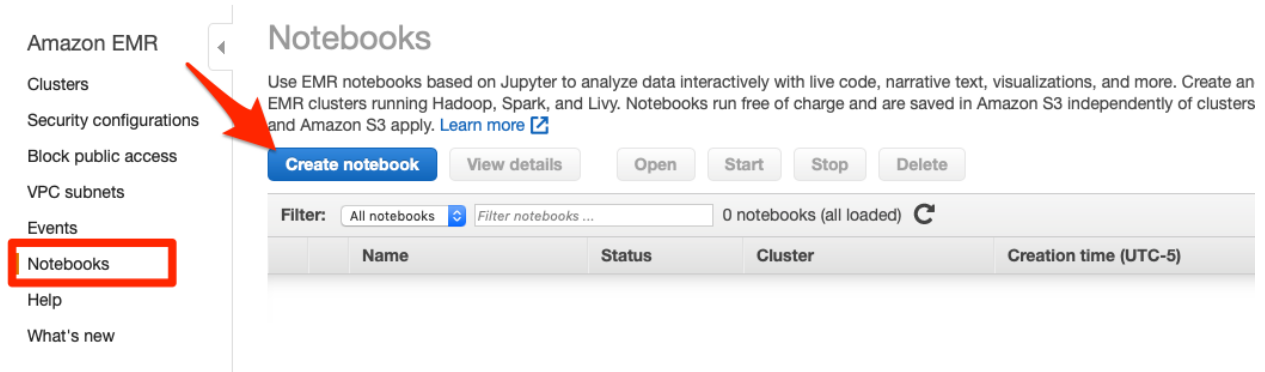
Name of interface	URI
YARN ResourceManager	<code>http://master-ip:8088/</code>
Hadoop HDFS NameNode	<code>http://master-ip:50070/</code>
Spark HistoryServer	<code>http://master-ip:18080/</code>

Note: if you want to consider other ways to access the web interfaces and for a longer list of interfaces, consult the [AWS documentation](#).

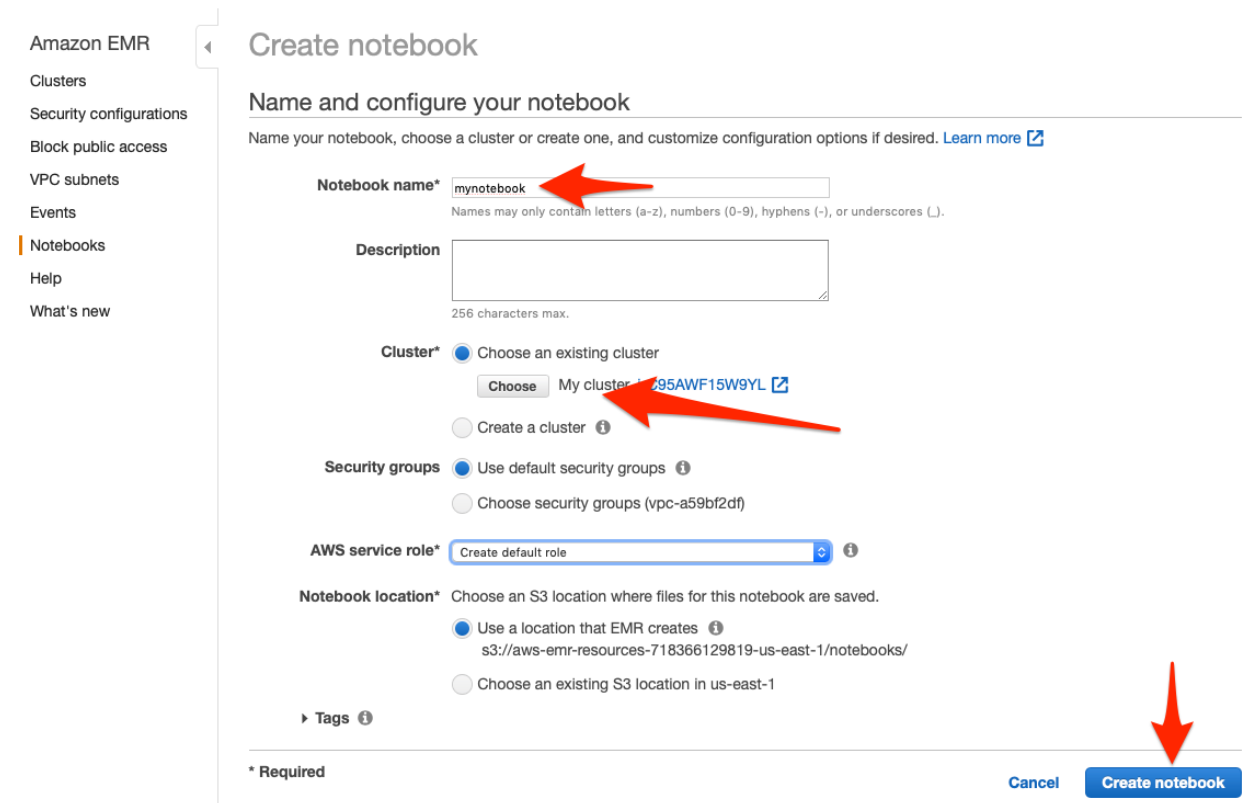
# Accessing the cluster with a Jupyter Notebook

Ignore this section (doesn't work any more). Check the last section in this guide for an alternative way.

Click on the Notebooks Tab and then click on the "Create a notebook" button:



Choose a name for your notebook and choose the cluster you created in the previous step as the cluster associated with this notebook. Finally click "create notebook".



Note: your notebook will be saved in S3 and you will always find it there. However, you should create a cluster only when you need to work on your notebook and then terminate the cluster

immediately after you are done to avoid wasting credits on an unused cluster. You can stop the notebook and leave it there for your next time to continue working where you left.

The screenshot shows the Amazon EMR console interface. On the left is a navigation menu with options: Amazon EMR, Clusters, Security configurations, Block public access, VPC subnets, Events, Notebooks (highlighted with an orange bar), Help, and What's new. The main content area displays details for a notebook named 'mynotebook', which is in a 'Ready' state. A red arrow points to the 'Open' button at the top of the notebook details. Below the 'Open' button are 'Stop' and 'Delete' buttons. The notebook details include: Notebook ID (e-DGQ3IGL3V35HQHV7S1YKW38C), Description (--), Last modified (3 seconds ago), Last modified by (...assumed-role/vocstartsoft/user443936=alfifi@tamu.edu), Created on (2019-09-20 14:50 (UTC-5)), Created by (...assumed-role/vocstartsoft/user443936=alfifi@tamu.edu), Service IAM role (EMR\_Notebooks\_DefaultRole), Notebook tags (creatorUserId = AROA2OQPVH2N6RGEBGCQI:user443936=alfifi@tamu.edu), and Notebook location (s3://aws-emr-resources-718366129819-us-east-1/notebooks/). Below the notebook details is a 'Cluster' section showing: Cluster (My cluster), Cluster Id (j-C95AWF15W9YL), Cluster status (Waiting), Cluster tags (--), and Step logs (s3://aws-logs-718366129819-us-east-1/elasticmapreduce/).

After opening the notebook, choose PySpark Kernel and you are now ready to run Spark code against your cluster!

The screenshot shows the Jupyter Notebook interface for 'mynotebook'. The top bar includes the Jupyter logo, the notebook name 'mynotebook', and '(autosaved)'. Below the top bar is a menu bar with File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. The 'Kernel' menu is open, showing options: Interrupt, Restart, Restart & Clear Output, Restart & Run All, Reconnect, Shutdown, and Change kernel. The 'Change kernel' option is selected, opening a sub-menu with options: PySpark, Python 3, Spark, and SparkR. The main area of the notebook shows a code cell with the text 'In [\*]: tweets = spark.read...' and a status bar at the bottom indicating 'Starting Spark app...'.

You will find a sample notebook in Piazza Resources to get you started.



```
In [1]: tweets = spark.read.json("s3://us-congress-tweets/congress-20181003.json.gz")
```

» Spark Job Progress

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1569008117578_0001	pyspark	idle	<a href="#">Link</a>	<a href="#">Link</a>	✓

SparkSession available as 'spark'.

```
In [3]: tweets.select("user.screen_name", "text").show()
```

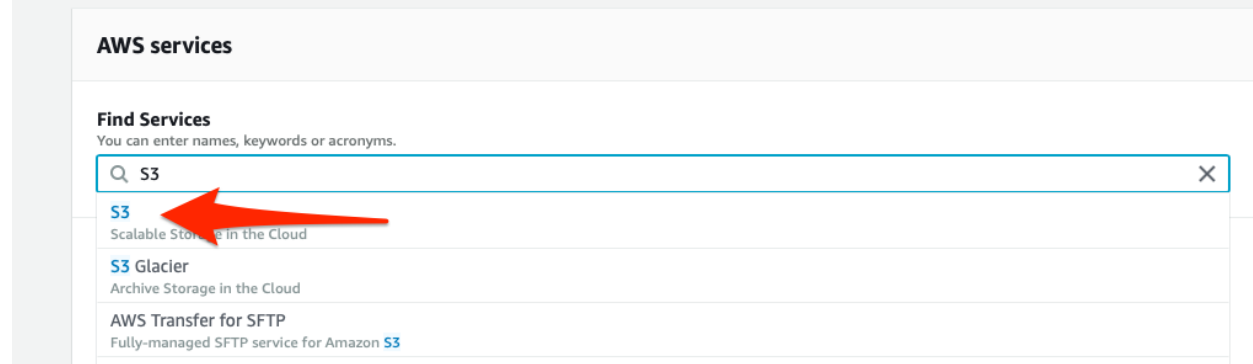
» Spark Job Progress

```
+-----+-----+
| screen_name | text |
+-----+-----+
| kikilezigoto | RT @namek237: - T... |
| chriswoillini | @charliekirk11 Pl... |
| michele5411 | ... she'll vote t... |
| aneesajv | RT @SenatorDurbin... |
| AcidRayneStorm | RT @Johnoco656060... |
| Jmooretrumpgirl | @lisamurkowski @S... |
| AzLakeHouse | RT @ChuckGrassley... |
| Srk195lmn | RT @RonWyden: Evi... |
| jannsloan | RT @CheriJacobus:... |
| burcham_don | RT @TODAYshow: "D... |
| EhHannah | @Keith1156 @canda... |
| JESUSFALFONS01 | RT @TrulyTrumpett... |
| Mo_An2016 | RT @DananaMama: @... |
| Stumpcuttr | #CoonsAndFlake Ne... |
| shoop_judy | RT @ChuckGrassley... |
| SheilaUtz1 | @RepAdamSchiff @H... |
| guernsey_robert | RT @RepAdamSchiff... |
| LymeLadytrump | RT @LawrenceBuck1... |
| trumpATeam | @realDonaldTrump ... |
| pamelasengle1 | RT @SenSchumer: A... |
+-----+-----+
only showing top 20 rows
```

# Creating An S3 Bucket

To avoid losing data when destroying a cluster, you can store your results in S3 but to do so you need to first create a bucket. To do that you go back to the AWS Management Console and this time choose S3 service rather than EMR. This is to store output of your operations. The Jupyter file itself will be stored by default in S3 without creating a specific bucket for it.

## AWS Management Console



Click "Create bucket" and follow the steps. You can use defaults for all steps. You should then have a bucket like the following:

S3 buckets Discover the console

Search for buckets All access types

[+ Create bucket](#) [Edit public access settings](#) [Empty](#) [Delete](#) 1 Buckets 1 Regions [Refresh](#)

<input type="checkbox"/> Bucket name	Access <span>i</span>	Region	Date created
<input type="checkbox"/> csce677	Bucket and objects not public	US East (N. Virginia)	Sep 20, 2019 3:52:59 PM GMT-0500

You can now store your data produced in the Jupyter notebook to this bucket which will survive cluster terminations. For example, in PySpark you could extract tweet ids and store them in S3 as follows:

```
tweets.select("id").write.csv("s3://mybucket/mytweets")
```

You could also write the output to the cluster HDFS as follows but keep in mind these will be deleted with you terminate the cluster:

```
tweets.select("id").write.csv("hdfs://mybucket/mytweets")
```

## Warning! Be a Terminator!

You should create a cluster before starting your work and terminate it immediately after you are done to avoid wasting credits. Your notebook will be save in S3 and next time you open the notebook there is an option to choose a new cluster to associate to the notebook.

You can use the AWS Glue Data Catalog as your external Hive metastore for [Apache Spark](#), [Apache Hive](#), and [Presto](#) workloads on Amazon EMR release 5.10.0 and later. To get started, simply select the AWS Glue Data Catalog for table metadata when creating your cluster.

Create clusterView detailsCloneTerminate

Filter: All clusters  1 cluster (all loaded)

	Name	ID	Status	Creation time (UTC-5) ▾	Elapsed time
▶  My cluster		j-C95AWF15W9YL	Waiting Cluster ready	2019-09-20 14:31 (UTC-5)	1 hour, 17 minutes

# Running Homework Jupyter Notebook

ssh to the master node of your cluster and run the following commands. If you haven't started working on the homework, then you can download the homework from S3 to the master node as follows:

```
$ aws s3 cp s3://caverlee/homework2.ipynb .
```

After that, you can start Jupyter in the terminal as follows:

```
$ ./start-jupyter.sh
```

You can then access it on the web your master node public domain and the port is 8880. The password is csce676.

```
start-jupyter.sh
[hadoop@ip-172-31-55-236 ~]$ aws s3 cp s3://caverlee/homework2.ipynb .
download: s3://caverlee/homework2.ipynb to ./homework2.ipynb
[hadoop@ip-172-31-55-236 ~]$ ./start-jupyter.sh
[I 23:19:16.880 NotebookApp] Writing notebook server cookie secret to /home/hadoop/.local/share/jupyter/runtime/notebook_cookie_secret
[I 23:19:17.034 NotebookApp] Serving notebooks from local directory: /home/hadoop
[I 23:19:17.034 NotebookApp] The Jupyter Notebook is running at:
[I 23:19:17.034 NotebookApp] http://(ip-172-31-55-236 or 127.0.0.1):8880/
[I 23:19:17.034 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
```

The screenshot shows the Jupyter Notebook web interface in a browser. The address bar indicates the URL is `http://ec2-18-206-15-103.compute-1.amazonaws.com`. The page displays the Jupyter logo and a login form with the password `csce676` and a "Log in" button. Below the login form, there is a "Files" tab selected, showing a list of files in the current directory. The files listed are `homework2.ipynb` (19.8 kB, 5 hours ago) and `start-jupyter.sh` (522 B, 27 minutes ago).

Name	Last Modified	File size
homework2.ipynb	5 hours ago	19.8 kB
start-jupyter.sh	27 minutes ago	522 B

## Warning: Save homework2.ipynb before termination!

Remember to download your edited homework2.ipynb to your machine or create a bucket on s3 and upload there because when you terminal the cluster it will be deleted.