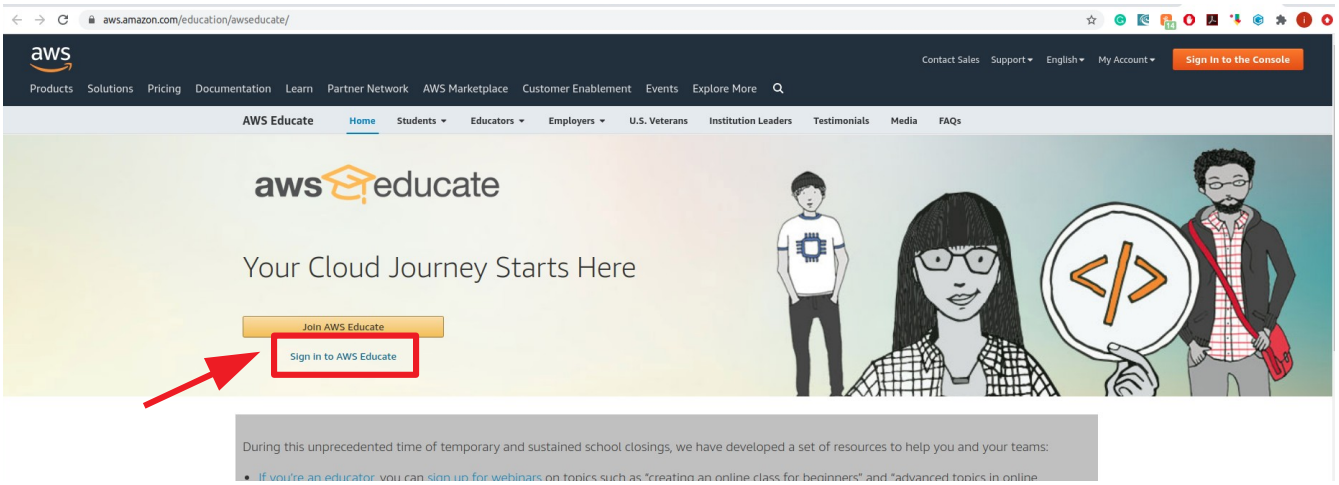
Instructions for using the Amazon Web Services(AWS) Platform for TP2

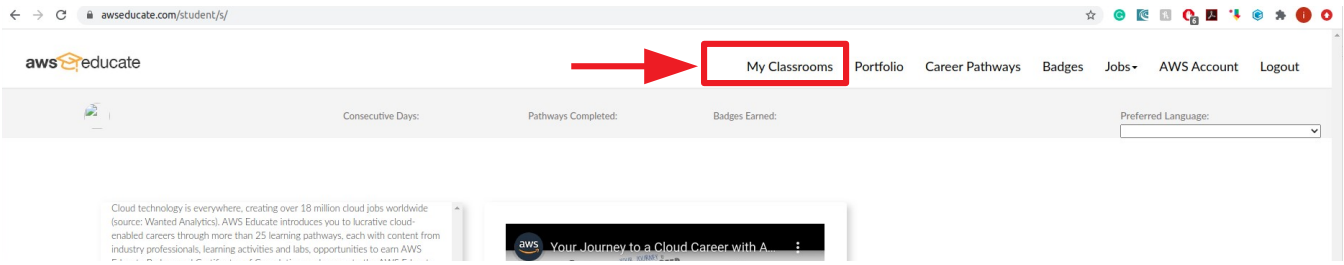
Dear students, you will find below the instructions on how to use the Amazon Web Services (AWS) required for the last part of TP2.

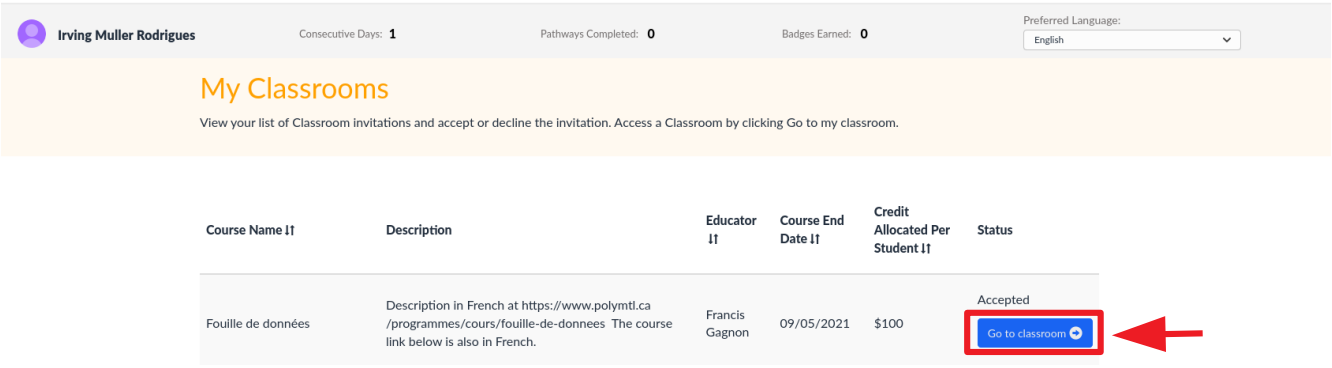
1. Activating your AWS education account

An email was sent to you by AWS Educate support. Use the provided link to set your password and open your educate account. Once you have completed this step, you can log in to the AWS Educate Student Portal using the following link:

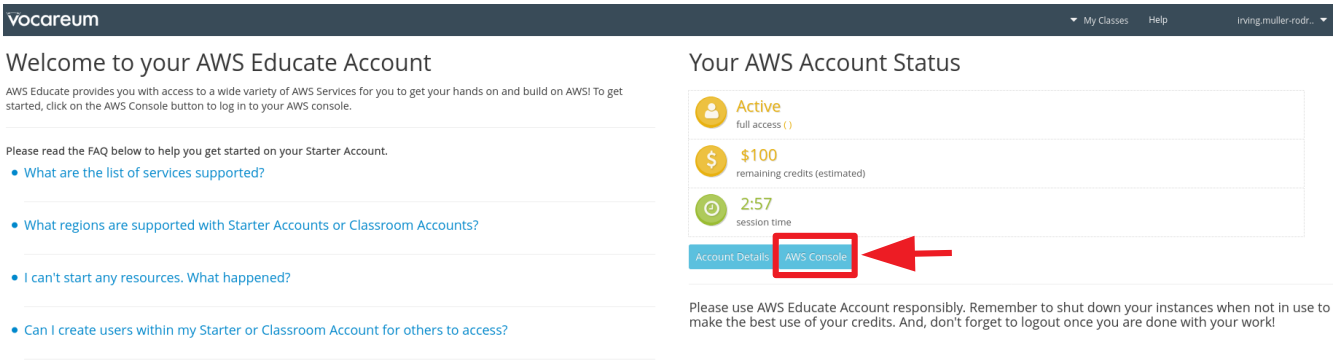
https://aws.amazon.com/education/awseducate/







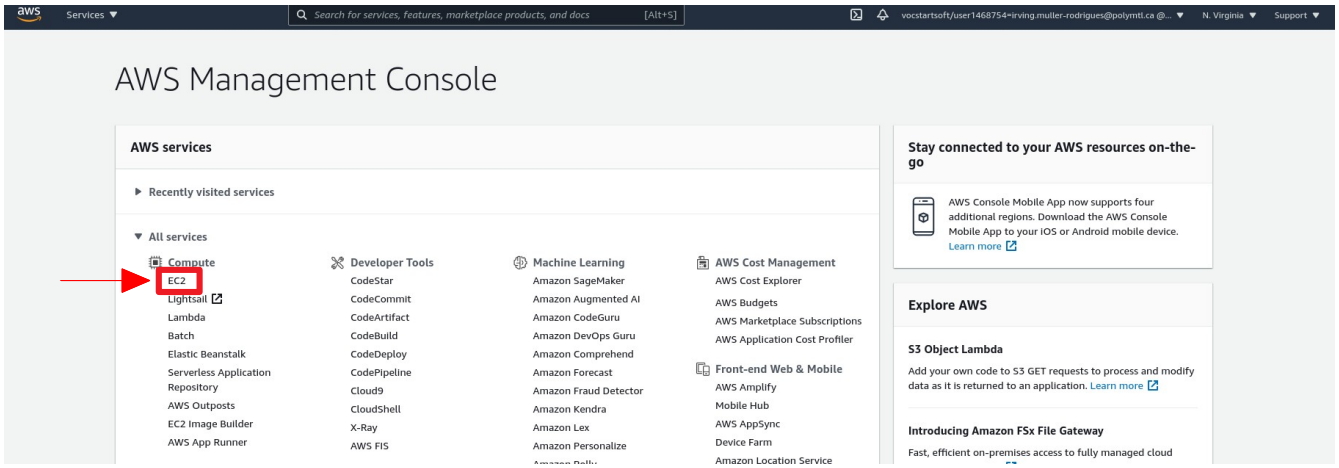
Now you can see your AWS account status. Select “AWS console” to access AWS Management Console:



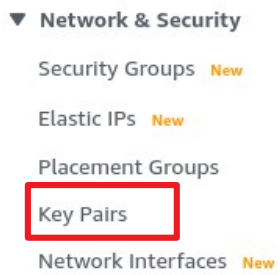
2. Create Key pair

Amazon AWS uses keys to encrypt and decrypt login information. At the basic level, a sender uses a public key to encrypt data, which its receiver then decrypts using another private key. These two keys, public and private, are known as a key pair. You need a key pair to be able to connect to your instances.

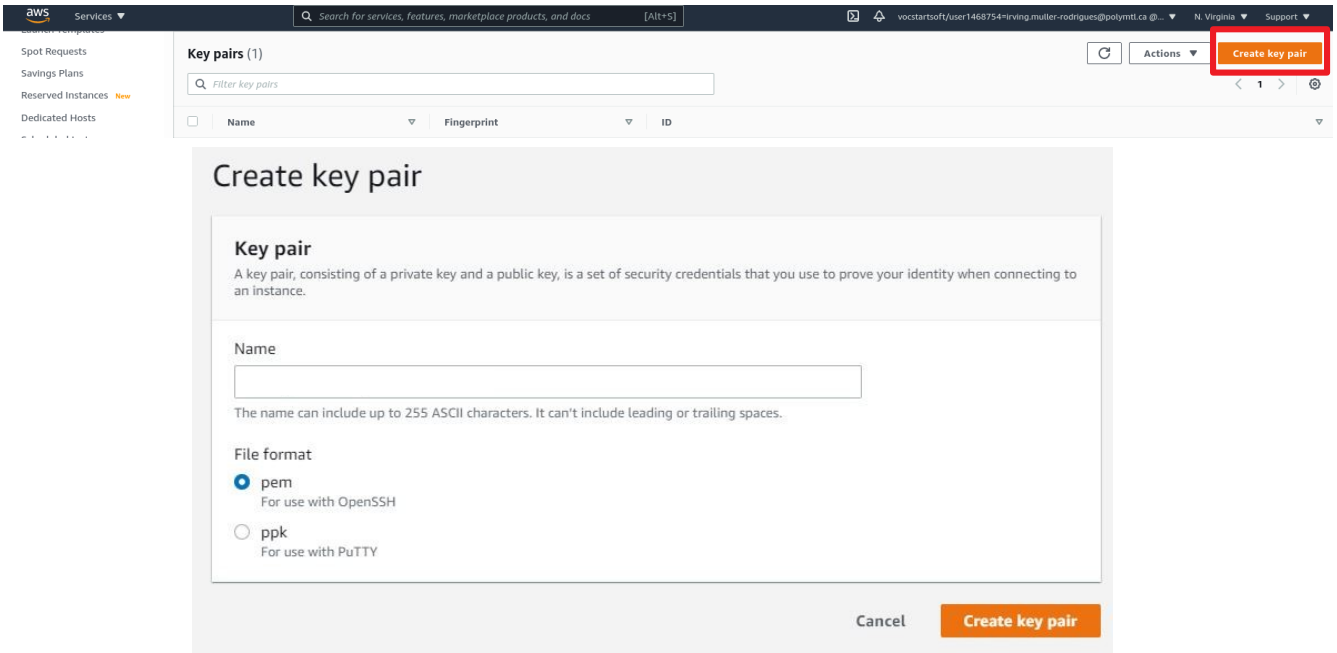
Go to the services section and select EC2.



Click on the “key pairs” button in the left bar and then choose Create Key Pair:



Click on Create key pair button, enter a name like my_key and select “pem” file format for use with OpenSSH.



Note: Because Amazon EC2 doesn't keep a copy of your private key, there is no way to recover a private key if you lose it. However, there can still be a way to connect to instances that use a lost key pair.

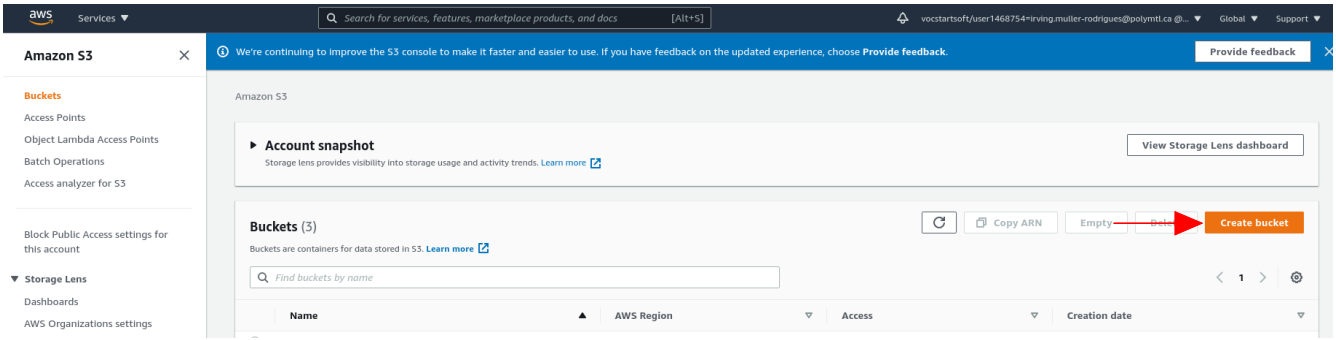
Create a new key pair and download it to your system. This key is required to connect to the EC2 instance. Then change the file permission of the downloaded security key in your pc and move it to a secure location (In Unix based OS: “Users/username/.ssh/”).

```
mv ~/Downloads/aws_ec2_security.pem /home/username/.ssh/
chmod 400 aws_ec2_security.pem
```

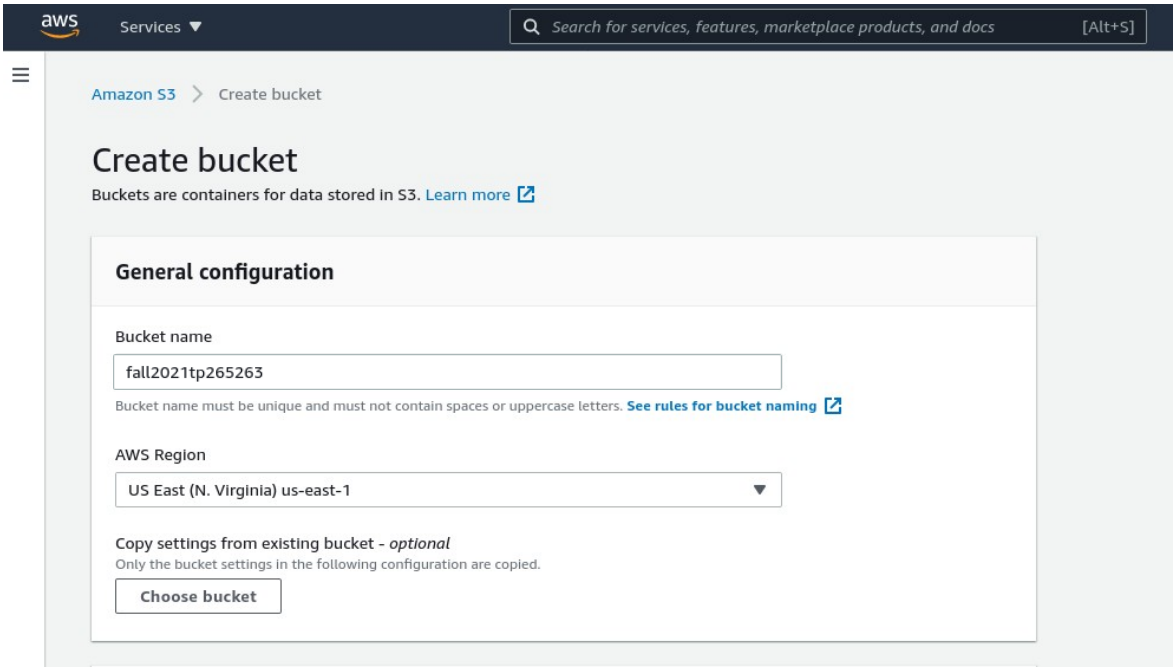
Note: Don’t forget to replace username by your actual username.

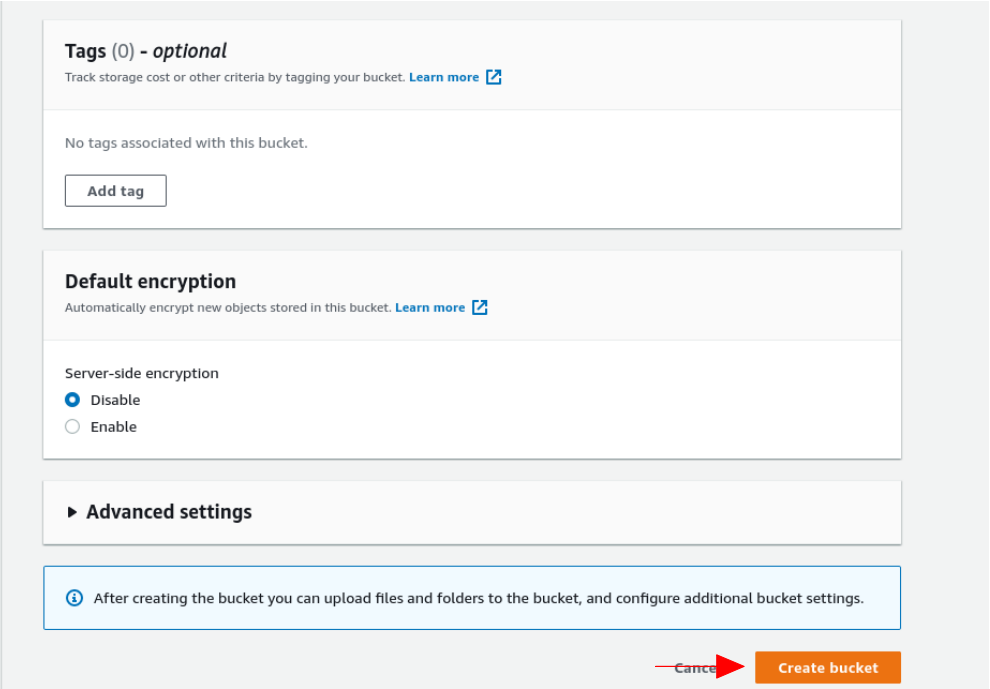
3. Create Bucket and Upload files

The notebook and dataset will be stored on Amazon S3. First, you will create a bucket on S3. For that, you have to access https://s3.console.aws.amazon.com/s3/ and, then, click on button Create Bucket.

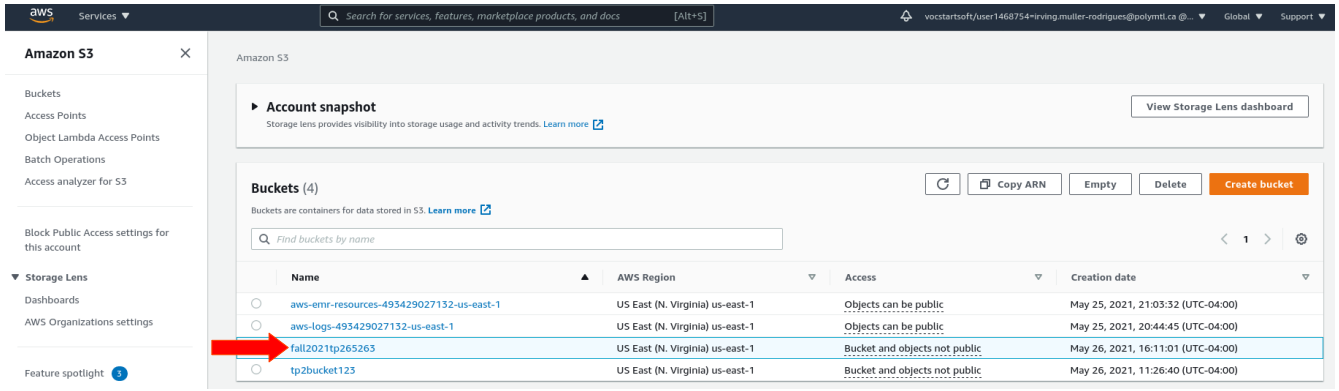


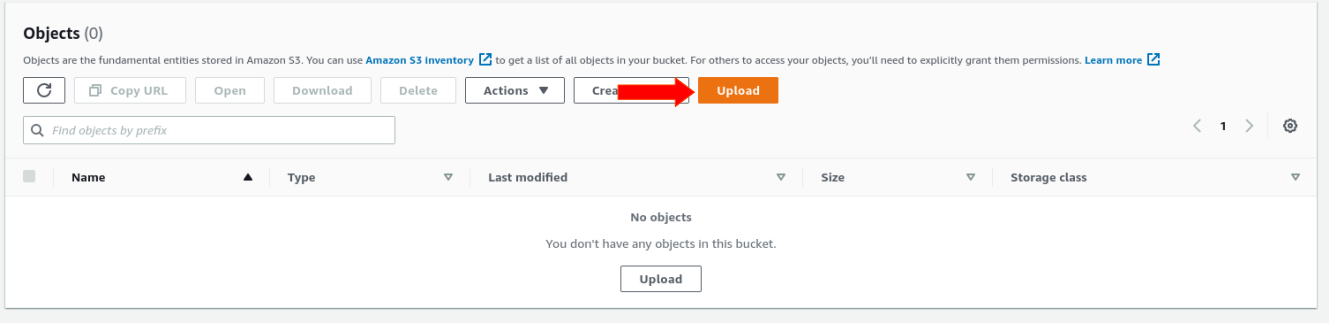
Choose a bucket name (e.g., fall2021tp265263) that is **unique** and select the region ***US East (N. Virginia) us-east-1***. Click on C*reate Bucket.*

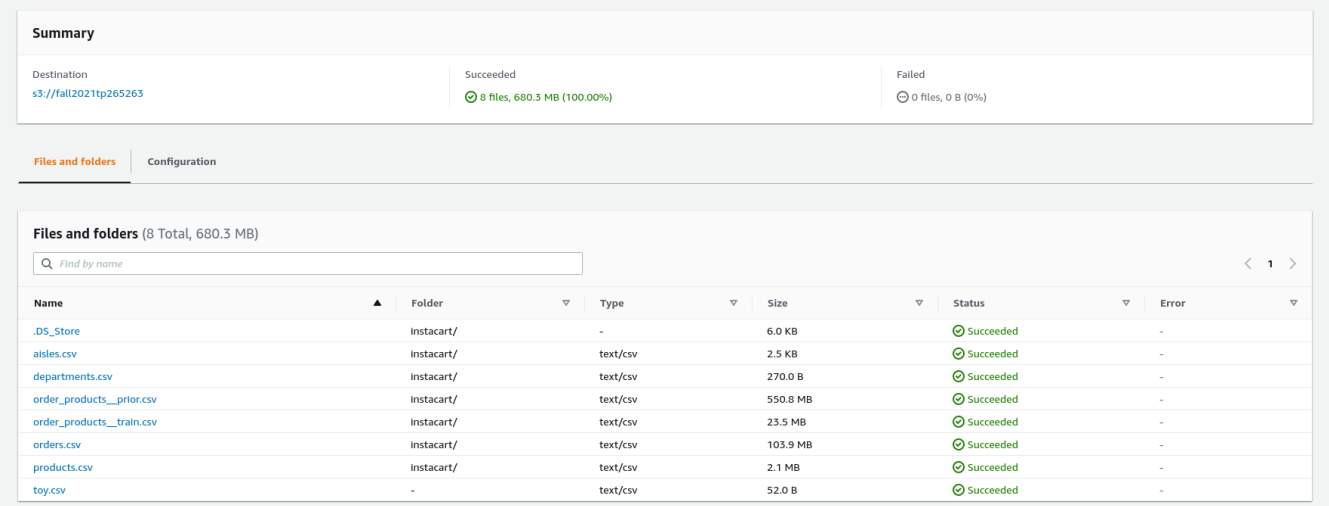




Click on the bucket name that you have created. Then, click on Upload to upload the **toy.csv** and **instacart dataset**.


Services
Search for services, features, marketplace products, and docs
[Alt+S]
Global
Support
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more
View Storage Lens dashboard
Buckets (4)
Buckets are containers for data stored in S3. Learn more
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
Name
AWS Region
Access
Creation date
aws-emr-resources-493429027132-us-east-1
US East (N. Virginia) us-east-1
Objects can be public
May 25, 2021, 21:03:32 (UTC-04:00)
aws-logs-493429027132-us-east-1
US East (N. Virginia) us-east-1
Objects can be public
May 25, 2021, 20:44:45 (UTC-04:00)
fall2021tp265263
US East (N. Virginia) us-east-1
Bucket and objects not public
May 26, 2021, 16:11:01 (UTC-04:00)
tp2bucket123
US East (N. Virginia) us-east-1
Bucket and objects not public
May 26, 2021, 11:26:40 (UTC-04:00)

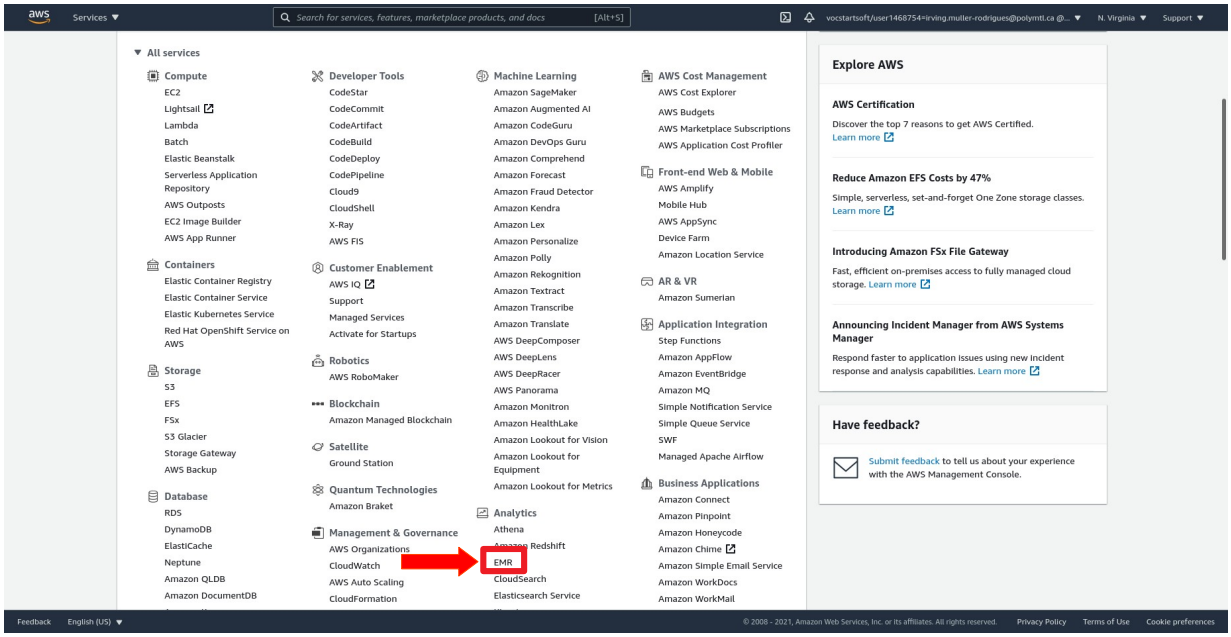

Objects (0)
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 Inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Copy URL
Open
Download
Delete
Actions
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
No objects
You don't have any objects in this bucket.
Upload


Summary
Destination
s3://fall2021tp265263
Succeeded
8 files, 680.3 MB (100.00%)
Failed
0 files, 0 B (0%)
Files and folders
Configuration
Files and folders (8 Total, 680.3 MB)
Find by name
Name
Folder
Type
Size
Status
Error
.DS_Store
instacart/
-
6.0 KB
Succeeded
aisles.csv
instacart/
text/csv
2.5 KB
Succeeded
departments.csv
instacart/
text/csv
270.0 B
Succeeded
order_products__prior.csv
instacart/
text/csv
550.8 MB
Succeeded
order_products__train.csv
instacart/
text/csv
23.5 MB
Succeeded
orders.csv
instacart/
text/csv
103.9 MB
Succeeded
products.csv
instacart/
text/csv
2.1 MB
Succeeded
toy.csv
-
text/csv
52.0 B
Succeeded

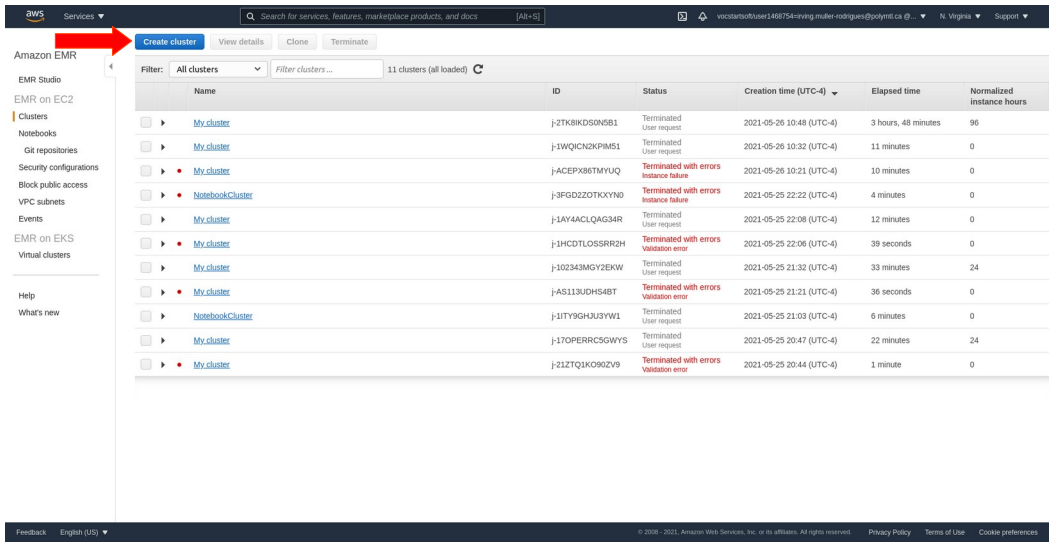
4. Create cluster

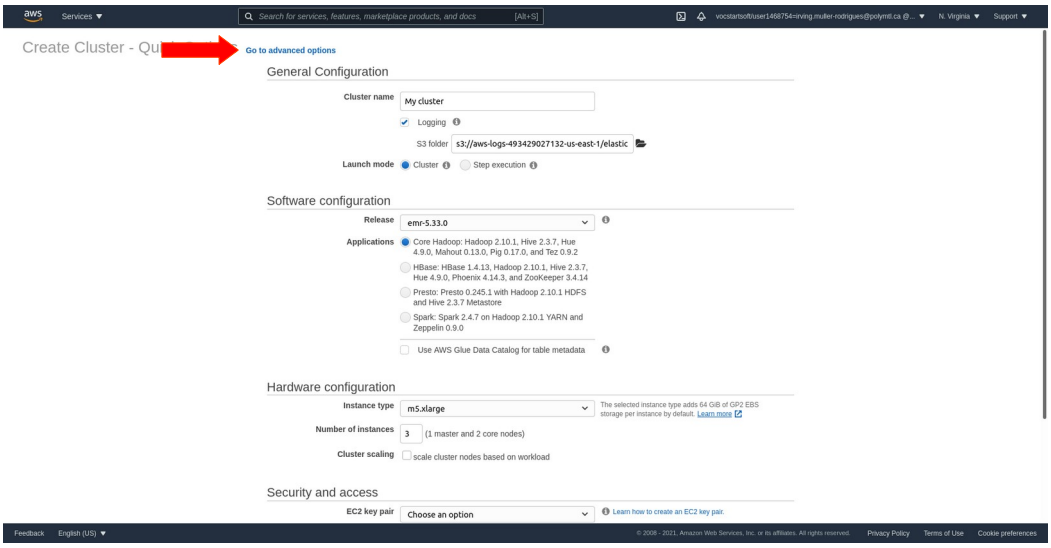
Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as `Apache Spark`, `Apache Hive`, `Apache HBase`, `Apache Flink`, `Apache Hudi`, and `Presto`. Amazon EMR makes it easy to set up, operate, and scale your big data environments by automating time-consuming tasks like provisioning capacity and tuning clusters.

Go to the services section and select **EMR**.



Click on *Create cluster*. Then, click on *Go to advanced options*



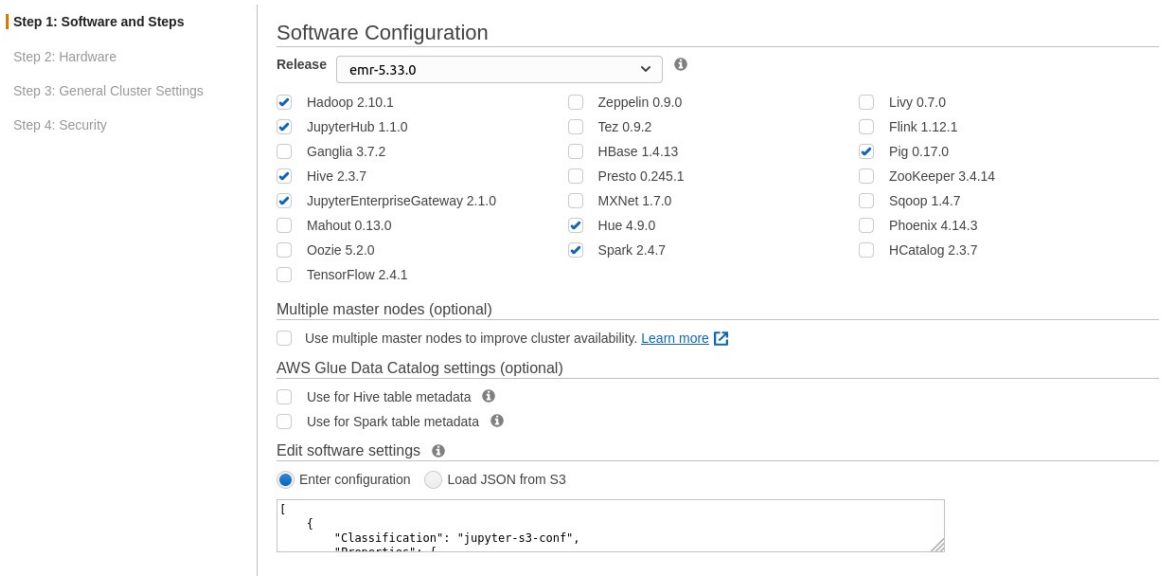


In section Software Configuration, select the release *emr-33.0* and check the checkbox *Hadoop 2.10.1*, *JupyterHub 1.1.0*, *Hive 2.3.7*, *JupyterEnterpriseGateway 2.1.0*, *Hue 4.9.0*, *Spark 2.4.7*, and *Pig 0.17.0*.

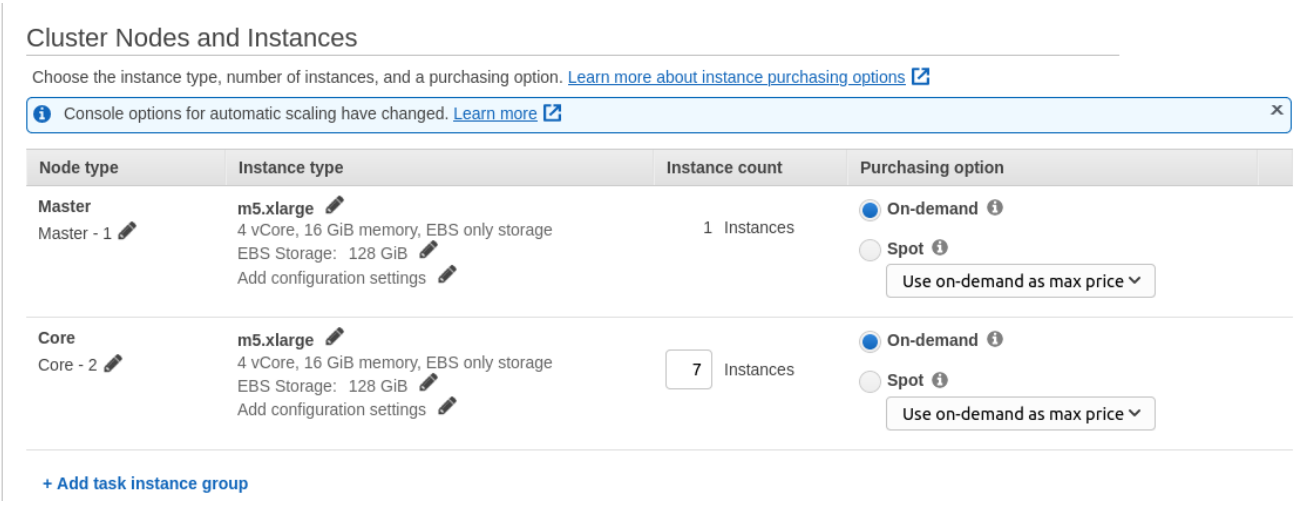
In section *Edit software settings*, add the following configuration to the TextArea:

```
[
    {
        "Classification": "jupyter-s3-conf",
        "Properties": {
            "s3.persistence.enabled": "true",
            "s3.persistence.bucket": "fall2021tp265263"
        }
    }
]
```

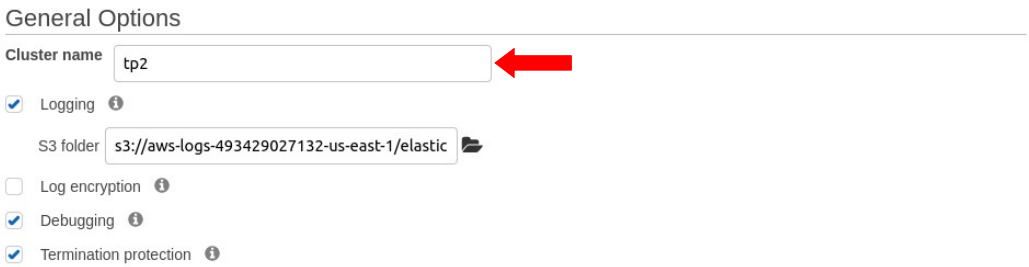
Note: Don't forget to replace **`s3.persistence.bucket`** with the bucket name created in the last section.



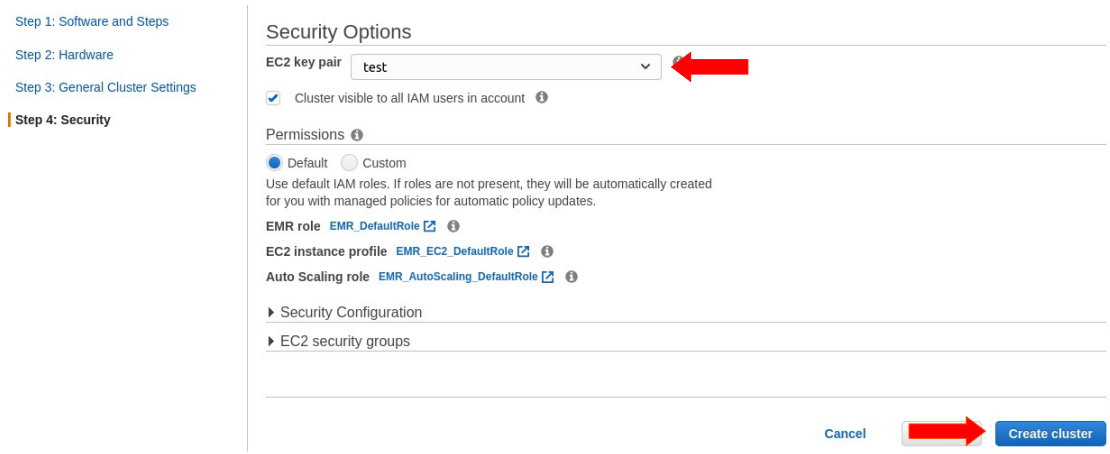
Click on next button to go to the **step 2 Hardware.** In step 2, use the setup bellow (1 master m5.xlarge and 7 master m5.xlarge) and click on next button.



In step 3, choose the cluster name (e.g., tp2) and click next to go the last step.



In step 4, choose your EC2 key pair and the create cluster.



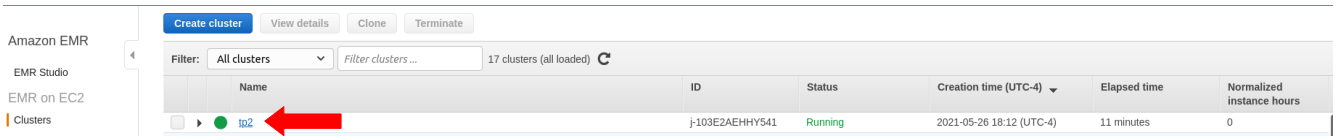
Note: It takes around 5 minutes to completely configure a cluster.

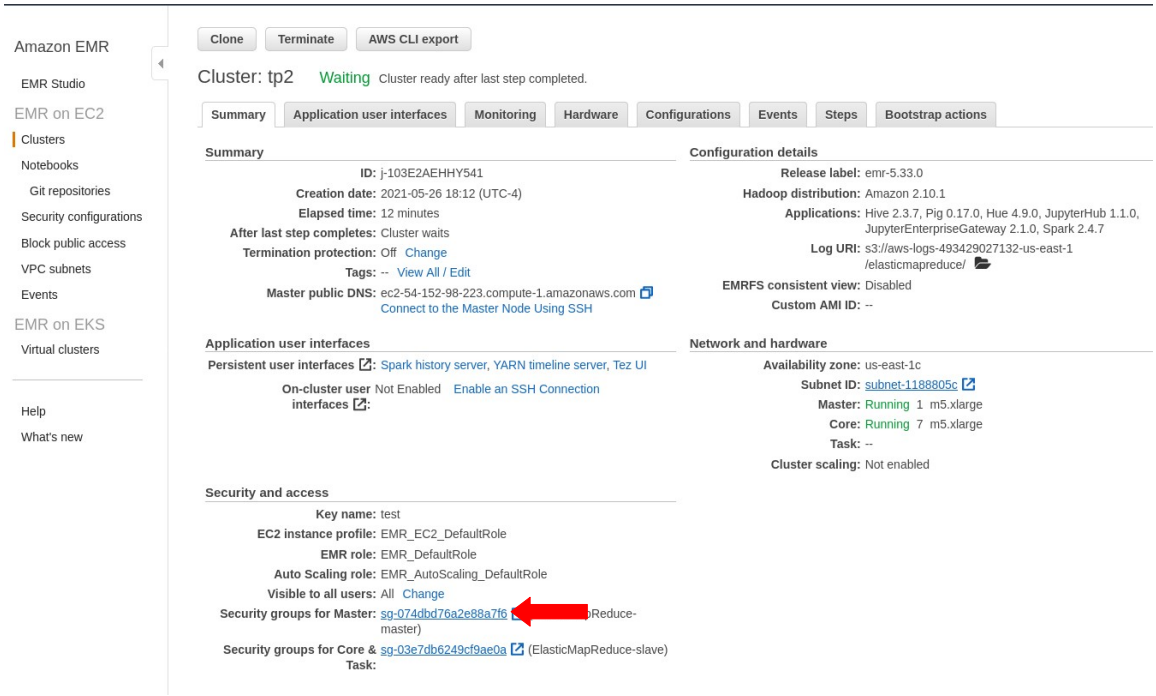
5. Run your notebook on AWS

5.1. Setup SSH tunnel

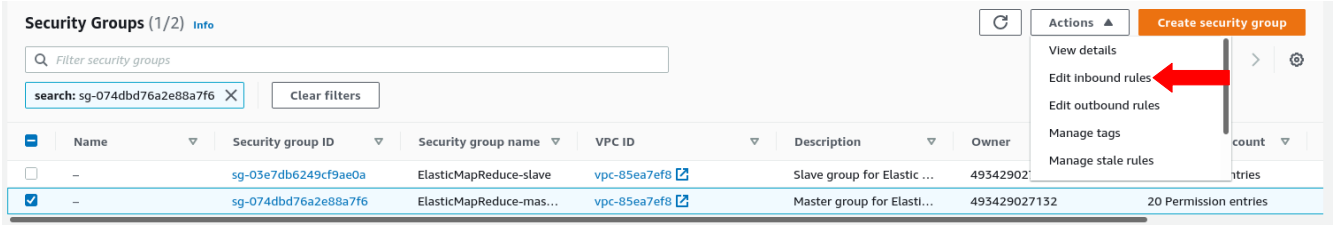
First, we need to add two inbound permissions that will allow us to connect to the master through SSH tunnel.

Click on your cluster in Amazon EMR page. Then, access the security group page by clicking on security groups of the master.

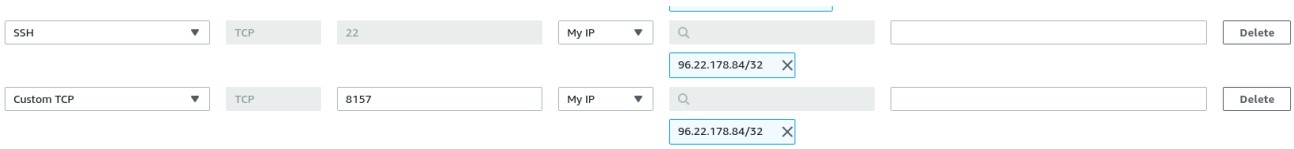




Select the security group id of the group name *ElasticMapReduce-master.* Then, click *Actions* and click on *Edit Inbound rules.*



Add two rules: 1) `SSH port_range=22 source=My Ip`; and 2) `Custom TP port_range=8157 source=My Ip`. Click on *Save rules.*

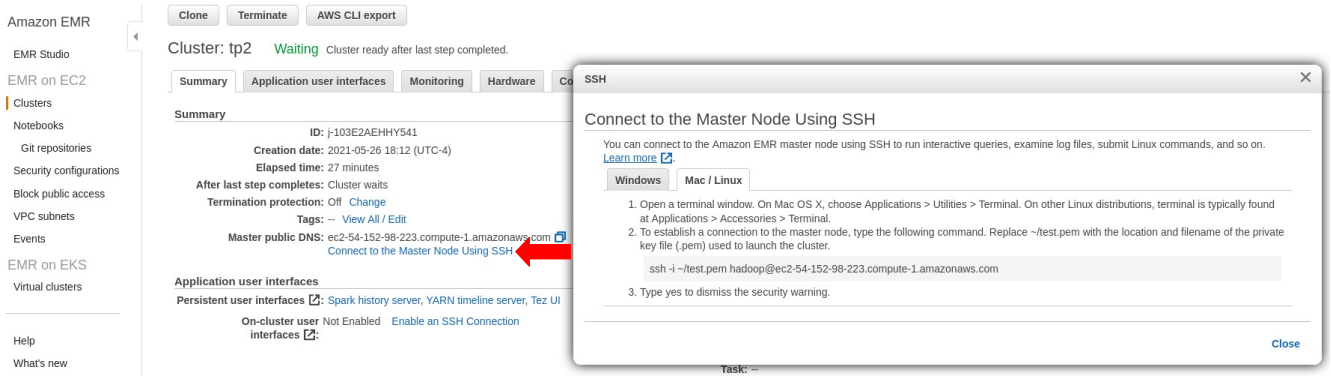


Note: Since your IP is probably dynamic, you might need update your IP during the TP.

Now, you can try to access the master by running the following command:

```
ssh -i your_key.pem  hadoop@ecX-XX-X-XX-XXX.compute-1.amazonaws.com
```

You can find the server url and more information about SSH by clicking on *Connect to the Master Node Using SSH*

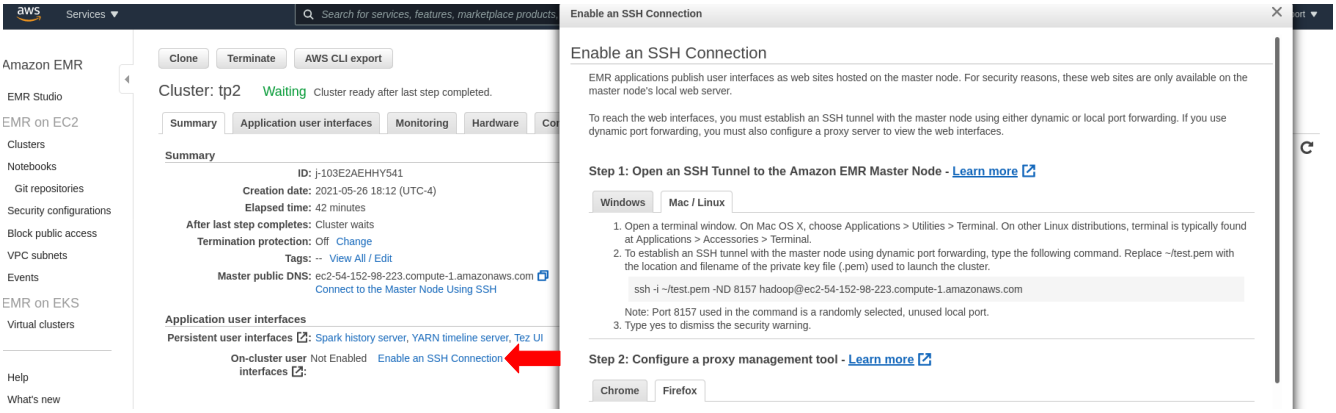


Note: “your_key.pem” is the key created in the second section of this document.

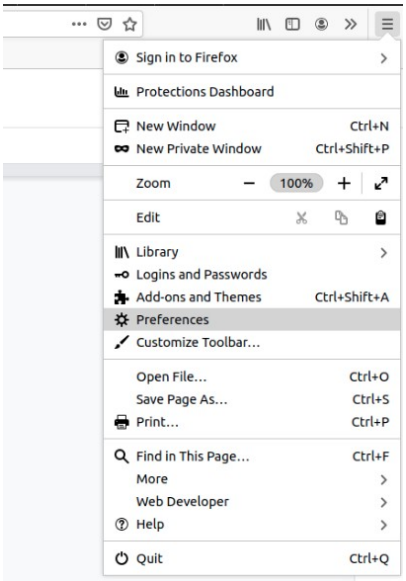
Now, run the following command to create a tunnel:

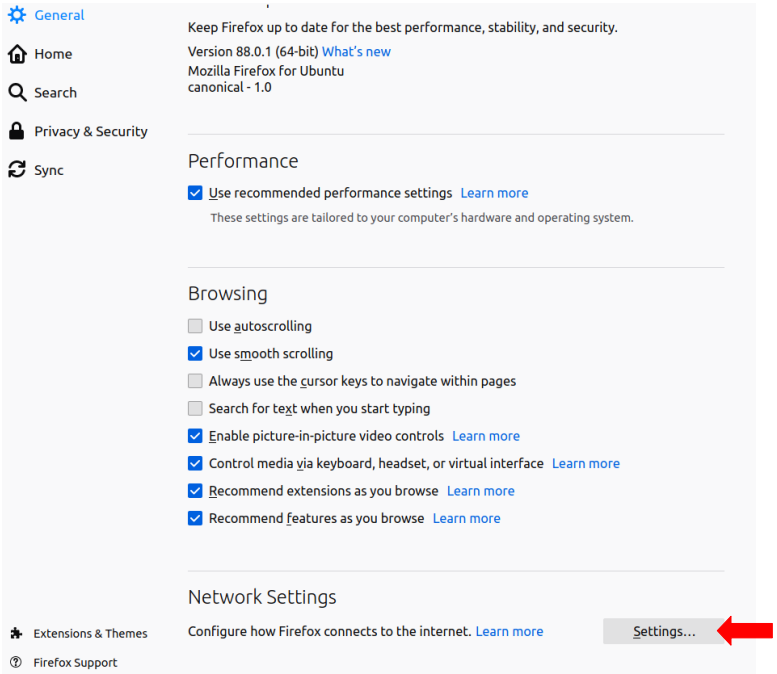
```
ssh -i .ssh/your_key.pem -ND 8157 hadoop@ecX-XX-XXX-XX-XXX.compute-1.amazonaws.com
```

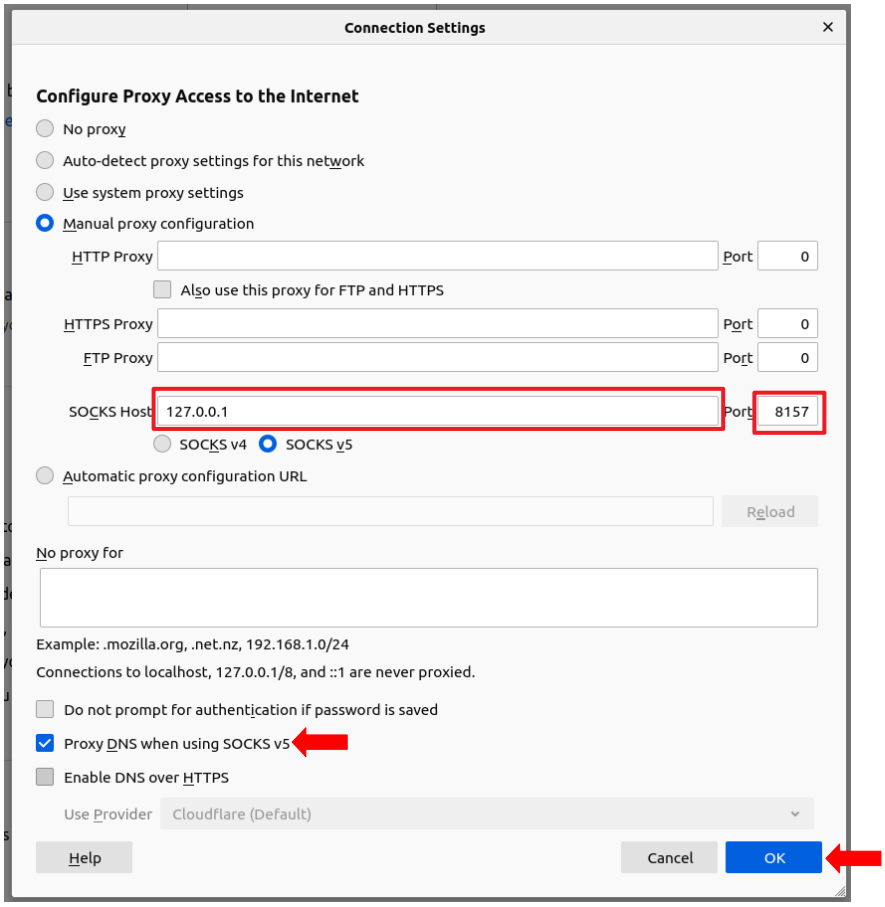
You can find the server url and more information about SSH Tunnel by clicking on *Connect to the Master Node Using SSH*



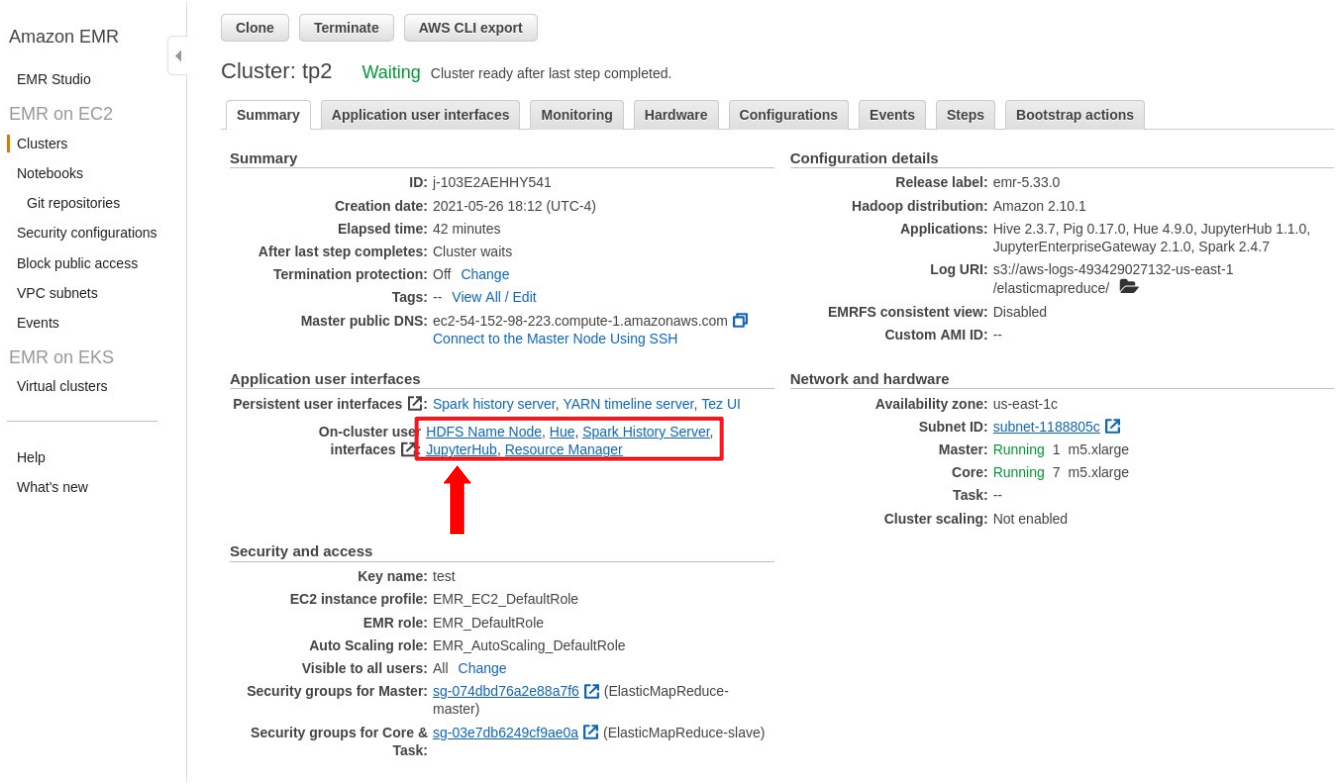
You can easily configure a proxy on Firefox. Click on *Preferences*, scroll down until Network settings, and click on *Settings*. Select *Manual proxy configuration* and enter 127.0.0.1 in *Socks Host* and 8157 in *Port*. Check the option *Proxy DNS when using Socks v5* and click on OK.




General
Home
Search
Privacy & Security
Sync
Extensions & Themes
Firefox Support
Keep Firefox up to date for the best performance, stability, and security.
Version 88.0.1 (64-bit) What's new
Mozilla Firefox for Ubuntu
canonical - 1.0
Performance
Use recommended performance settings Learn more
These settings are tailored to your computer's hardware and operating system.
Browsing
Use autoscrolling
Use smooth scrolling
Always use the cursor keys to navigate within pages
Search for text when you start typing
Enable picture-in-picture video controls Learn more
Control media via keyboard, headset, or virtual interface Learn more
Recommend extensions as you browse Learn more
Recommend features as you browse Learn more
Network Settings
Configure how Firefox connects to the internet. Learn more
Settings...


Connection Settings
Configure Proxy Access to the Internet
No proxy
Auto-detect proxy settings for this network
Use system proxy settings
Manual proxy configuration
HTTP Proxy
Port 0
Also use this proxy for FTP and HTTPS
HTTPS Proxy
Port 0
FTP Proxy
Port 0
SOCKS Host 127.0.0.1
Port 8157
SOCKS v4
SOCKS v5
Automatic proxy configuration URL
Reload
No proxy for
Example: .mozilla.org, .net.nz, 192.168.1.0/24
Connections to localhost, 127.0.0.1/8, and ::1 are never proxied.
Do not prompt for authentication if password is saved
Proxy DNS when using SOCKS v5
Enable DNS over HTTPS
Use Provider
Cloudflare (Default)
Help
Cancel
OK

After setting up the proxy, you should be able to see the *On-cluster user interface*. Click on JupyterHub to access jupyter notebook.



You can also access JupyterHub by creating a Local Port Forwarding. Run the following command to create a tunnel:

```
ssh -i test.pem -N -L 8157:ecX-XX-XXX-XX-XXX.compute-1.amazonaws.com:9443
hadoop@ecX-XX-XXX-XX-XXX.compute-1.amazonaws.com
```

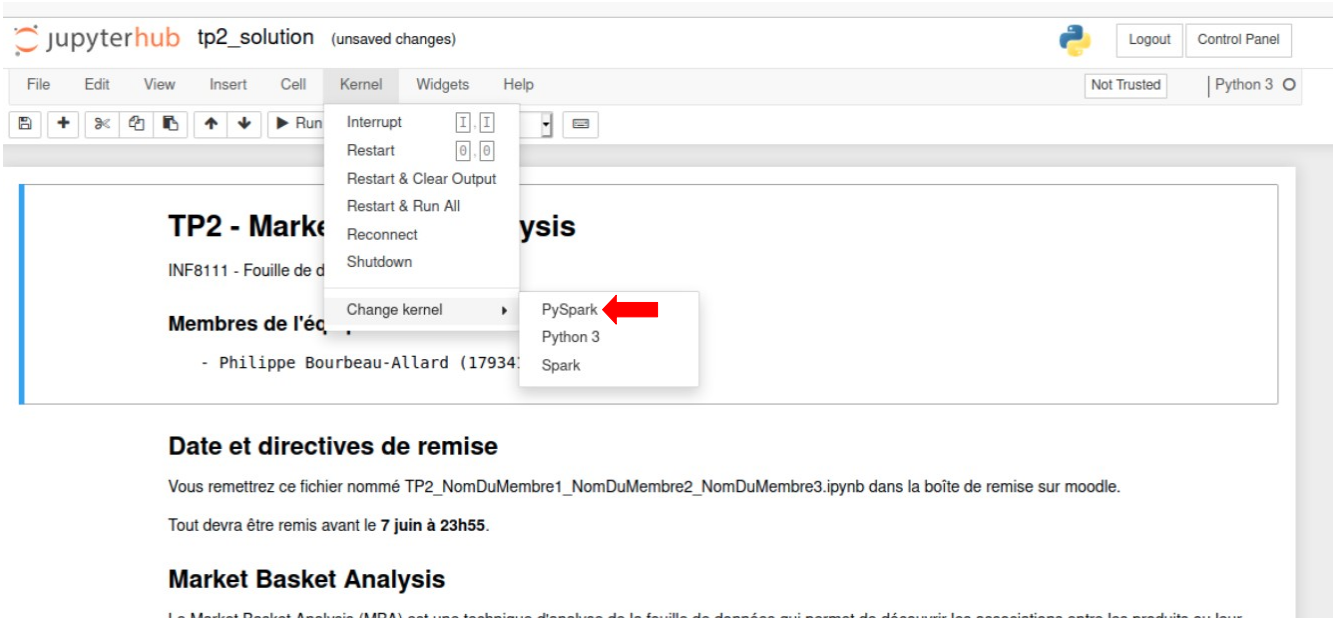
Note: replace ecX-XX-XXX-XX-XXX.compute-1.amazonaws by the correct master url

To open the JupyterHub in your browser, type **https://localhost:8157/** in the address bar.

5.2. JupyterHub

Access JupyterHub in your browser. The browser will display an alert. Ignore this alert and access the link. The **username** is **jovyan** and **password** is **jupyter**.

Upload your notebook to the JupyterHub. Access the notebook and change the kernel to *PySpark*.



To run your code on AWS, you have to change the path of toy.csv and instacart in the notebook.

```
toy = spark.read.csv('s3://fall2021tp265263/toy.csv', header=True)
```

```
df_order_prod =
spark.read.csv('s3://fall2021tp265263/instacart/order_products__train.csv
', header=True, sep=',', inferSchema=True)


df_orders = spark.read.csv('s3://fall2021tp265263/instacart/orders.csv',
header=True, sep=',', inferSchema=True)

df_products =
spark.read.csv('s3://fall2021tp265263/instacart/products.csv',
header=True, sep=',', inferSchema=True)

's3://fall2021tp265263/instacart/order_products__test.csv'
```

Note: Don't forget to replace `fall2021tp265263` with the correct bucket name.

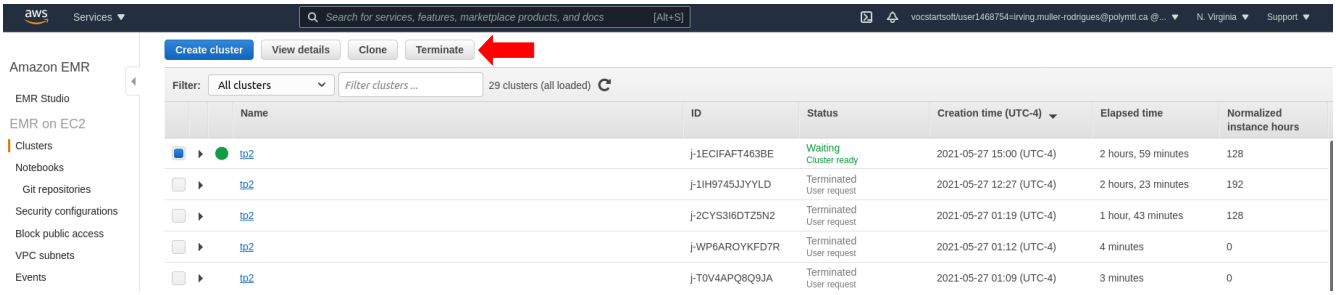
%%time **does not correctly work on JupyterHub**. You can compute the time to run a cell by using the package time. For instance:

```
from time import time

start = time()

# Your code

print(f"Total time: {time() - start} seconds")
```

6. Terminate a Cluster

Different of EC2 instance, you cannot stop or shutdown a cluster. To make the best use of your credits, you have to **terminate** the cluster after running a notebook.

For that, go to https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1. Then, select your cluster and click on *Terminate*.



Note: If you use all your credits, your data will be lost. Thus, be careful to always terminate a cluster after using it.

You can easily create the same cluster by selecting a cluster and clicking on *Clone*.