

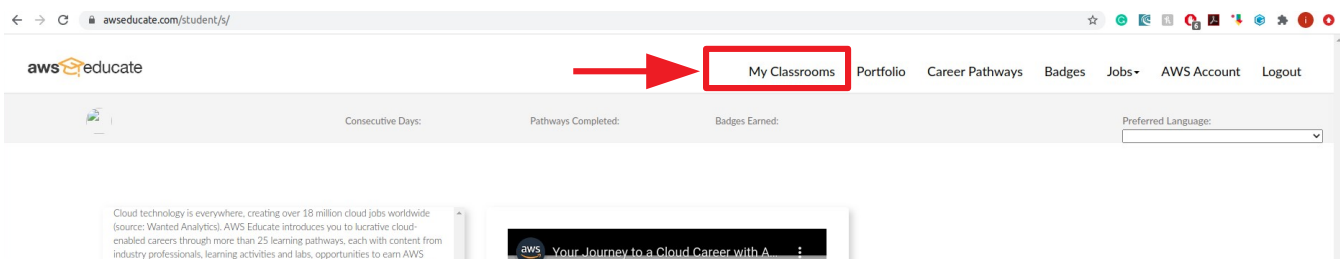
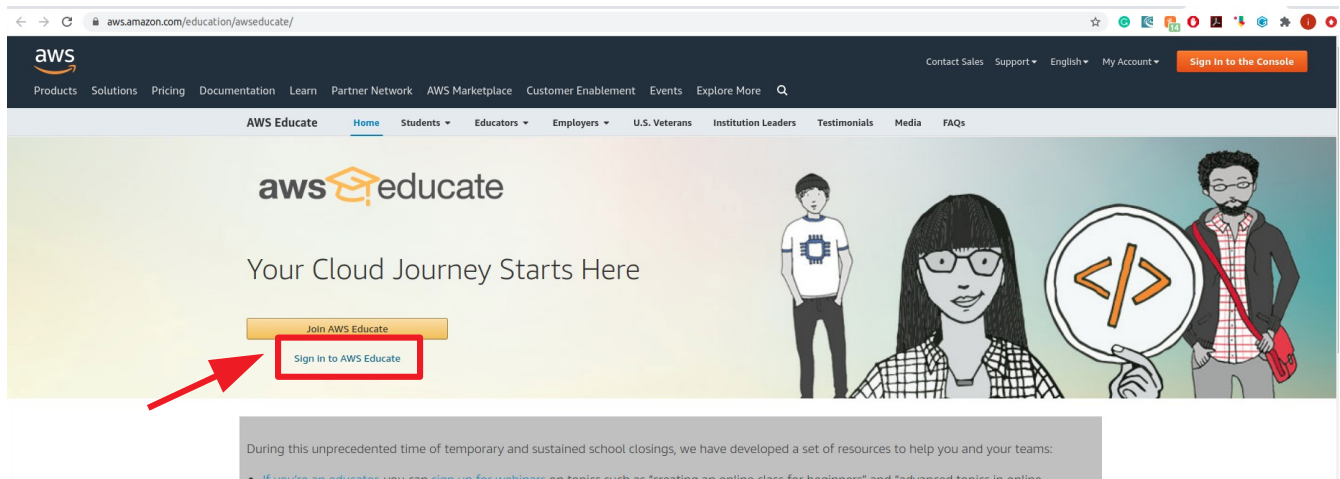
# Instructions for using the Amazon Web Services(AWS) Platform for TP2

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/>



**Irving Muller Rodrigues**

Consecutive Days: **1**

Pathways Completed: **0**

Badges Earned: **0**

Preferred Language: English

## My Classrooms

View your list of Classroom invitations and accept or decline the invitation. Access a Classroom by clicking Go to my classroom.

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

**vocareum**
My Classes
Help
irving.muller-rodr...

### Welcome to your AWS Educate Account

AWS Educate provides you with access to a wide variety of AWS Services for you to get your hands on and build on AWS! To get started, click on the AWS Console button to log in to your AWS console.

Please read the FAQ below to help you get started on your Starter Account.

- What are the list of services supported?
- What regions are supported with Starter Accounts or Classroom Accounts?
- I can't start any resources. What happened?
- Can I create users within my Starter or Classroom Account for others to access?

### Your AWS Account Status

Active
full access ( )

\$100
remaining credits (estimated)

2:57
session time

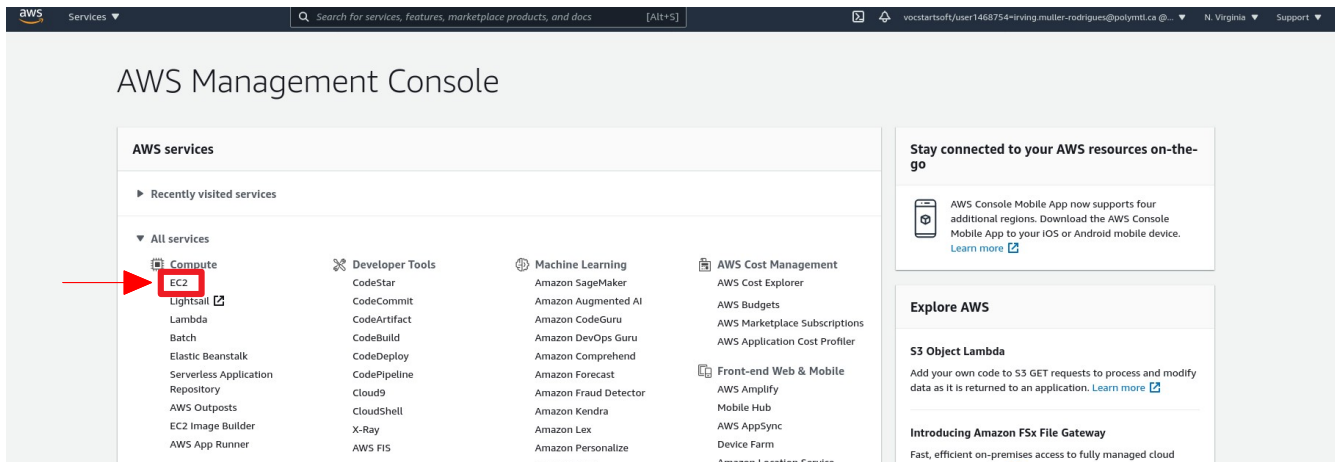
Account Details
[AWS Console](#)

Please use AWS Educate Account responsibly. Remember to shut down your instances when not in use to make the best use of your credits. And, don't forget to logout once you are done with your work!

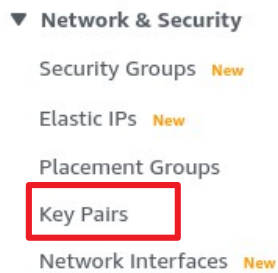
## 2. Create Key pair

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.

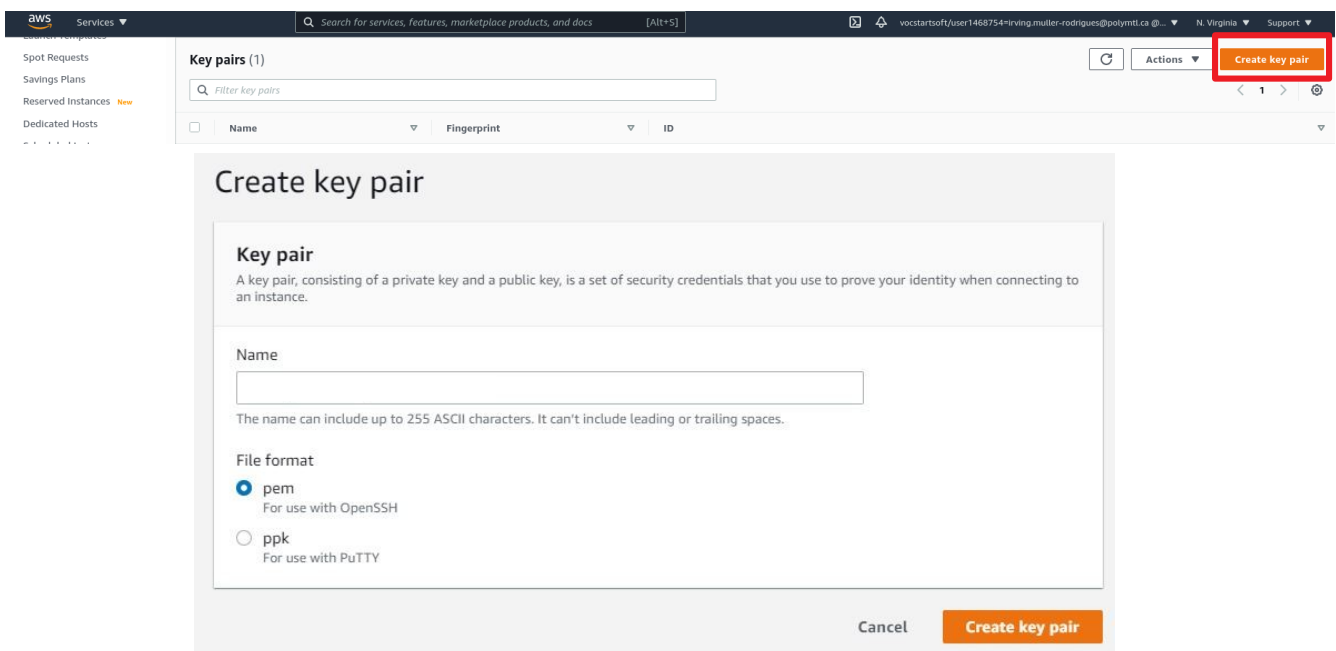
Go to the services section and select EC2.



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.

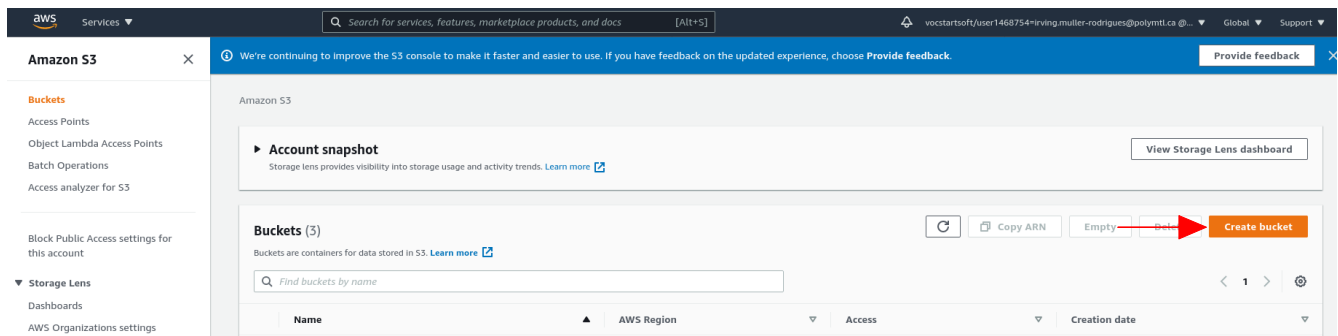
Create a new key pair and download it to your system. This key is required to connect to the EC2 instance. Then change the file permission of the downloaded security key in your pc and move it to a secure location (In Unix based OS: “Users/username/.ssh/”).

```
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 *Create Bucket*.

aws Services ▾ Search for services, features, marketplace products, and docs [Alt+S]

Amazon S3 > Create bucket

## Create bucket

Buckets are containers for data stored in S3. [Learn more](#)

### General configuration

Bucket name

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

### Tags (0) - optional

Track storage cost or other criteria by tagging your bucket. [Learn more](#)

No tags associated with this bucket.

Add tag

### Default encryption

Automatically encrypt new objects stored in this bucket. [Learn more](#)

Server-side encryption

☒ Disable  
☐ Enable

► Advanced settings

After creating the bucket you can upload files and folders to the bucket, and configure additional bucket settings.

Cancel Create 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]

vocstartsoft/user1468734-irving.muller-rodriguez@polymtl.ca @... 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

Amazon S3

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](#)

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

Create

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

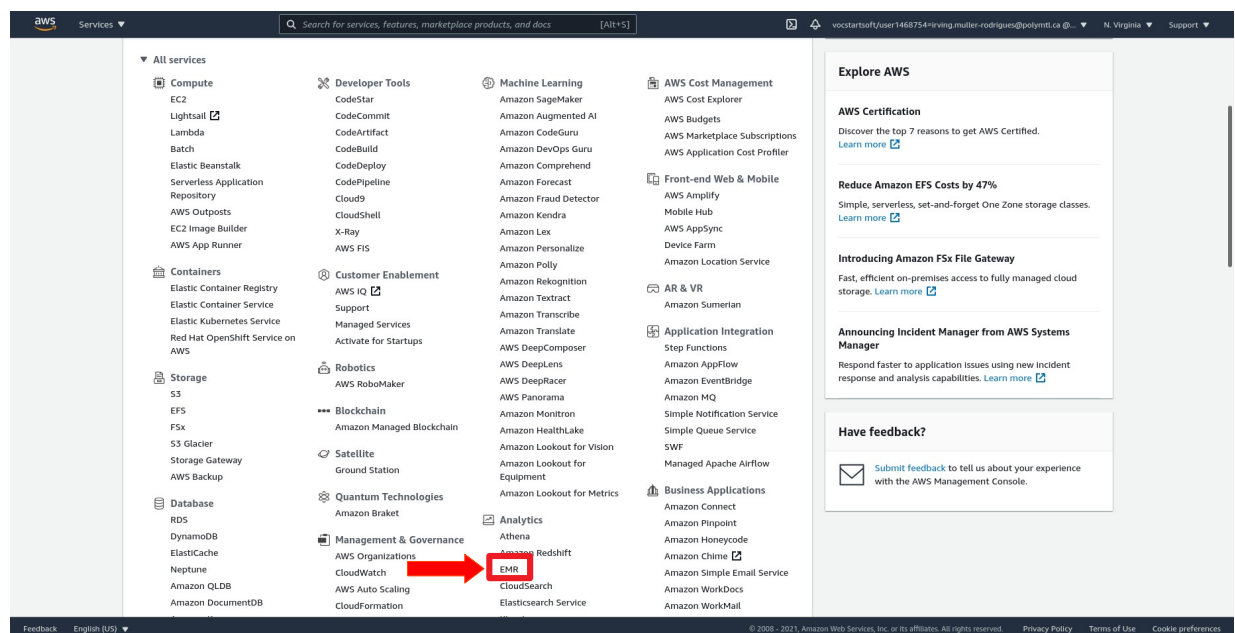
-

6

## 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*

The screenshot shows the AWS Management Console interface for creating an Amazon EMR cluster. The top navigation bar includes the AWS logo, a search bar, and user information. The left sidebar shows the 'Amazon EMR' service selected. The main content area displays the 'Create cluster' page with a table of existing clusters. A red arrow points to the 'Create cluster' button. Below the table, a red arrow points to the 'Go to advanced options' link.

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
My cluster	j-2TK8IKD5ON5B1	Terminated User request	2021-05-26 10:48 (UTC-4)	3 hours, 48 minutes	96
My cluster	j-1WQICN2KPM51	Terminated User request	2021-05-26 10:32 (UTC-4)	11 minutes	0
My cluster	j-ACEPX86TMYUQ	Terminated with errors Instance failure	2021-05-26 10:21 (UTC-4)	10 minutes	0
NotebookCluster	j-3FGD2ZOTKXYN0	Terminated with errors Instance failure	2021-05-25 22:22 (UTC-4)	4 minutes	0
My cluster	j-14V4ACLQAG34R	Terminated User request	2021-05-25 22:08 (UTC-4)	12 minutes	0
My cluster	j-1HCDTL0SSRR2H	Terminated with errors Validation error	2021-05-25 22:06 (UTC-4)	39 seconds	0
My cluster	j-102343MGY2EKW	Terminated User request	2021-05-25 21:32 (UTC-4)	33 minutes	24
My cluster	j-AS113UJH54BT	Terminated with errors Validation error	2021-05-25 21:21 (UTC-4)	36 seconds	0
NotebookCluster	j-1ITY9GHJU3YW1	Terminated User request	2021-05-25 21:03 (UTC-4)	6 minutes	0
My cluster	j-17OPERRC5GWYS	Terminated User request	2021-05-25 20:47 (UTC-4)	22 minutes	24
My cluster	j-21ZTQ1K090ZV9	Terminated with errors Validation error	2021-05-25 20:44 (UTC-4)	1 minute	0

The 'Create Cluster - Quickstart' page is shown below. It includes sections for General Configuration, Software configuration, Hardware configuration, and Security and access. A red arrow points to the 'Go to advanced options' link.

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.

### Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

### Software Configuration

Release **emr-5.33.0**

- |  |   |  |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 2.10.1                  | <input type="checkbox"/> Zeppelin 0.9.0         | <input type="checkbox"/> Livy 0.7.0            |
| <input checked="" type="checkbox"/> JupyterHub 1.1.0               | <input type="checkbox"/> Tez 0.9.2              | <input type="checkbox"/> Flink 1.12.1          |
| <input type="checkbox"/> Ganglia 3.7.2                             | <input type="checkbox"/> HBase 1.4.13           | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.7                     | <input type="checkbox"/> Presto 0.245.1         | <input type="checkbox"/> ZooKeeper 3.4.14      |
| <input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.1.0 | <input type="checkbox"/> MXNet 1.7.0            | <input type="checkbox"/> Sqoop 1.4.7           |
| <input type="checkbox"/> Mahout 0.13.0                             | <input checked="" type="checkbox"/> Hue 4.9.0   | <input type="checkbox"/> Phoenix 4.14.3        |
| <input type="checkbox"/> Oozie 5.2.0                               | <input checked="" type="checkbox"/> Spark 2.4.7 | <input type="checkbox"/> HCatalog 2.3.7        |
| <input type="checkbox"/> TensorFlow 2.4.1                          |   |  |

#### Multiple master nodes (optional)

- ☐ Use multiple master nodes to improve cluster availability. [Learn more](#)

#### 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": "jupyter-s3-conf",
    "Properties": {
```

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.

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. [Learn more about instance purchasing options](#)

Console options for automatic scaling have changed. [Learn more](#)

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 128 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 128 GiB Add configuration settings	7 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price

+ Add task instance group

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

General Options

Cluster name

☒ Logging  
S3 folder

☐ Log encryption  
☒ Debugging  
☒ Termination protection

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

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

Cancel

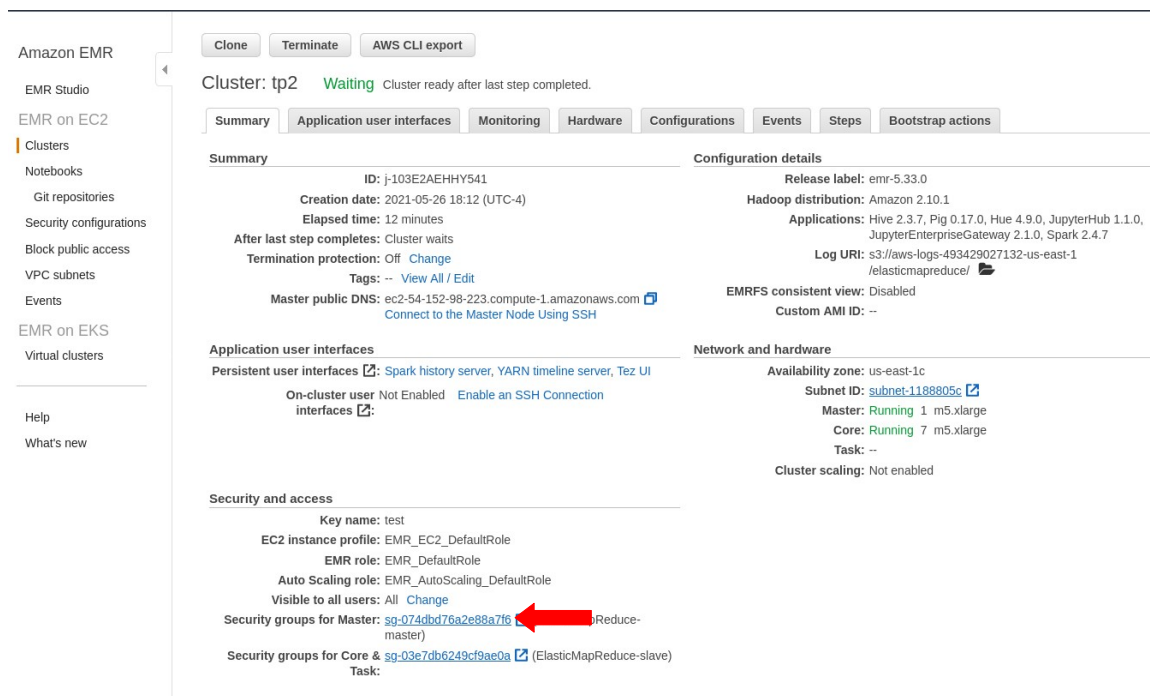
Note: It takes around 5 minutes to completely configure a cluster.

## 5. Run your notebook on AWS

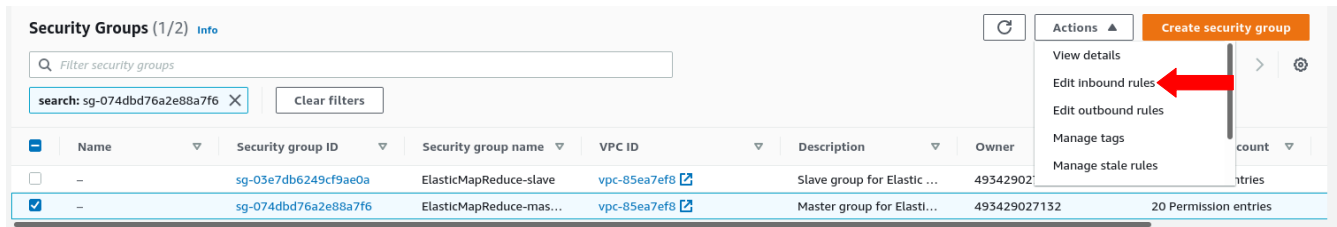
### 5.1. Setup SSH tunnel

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

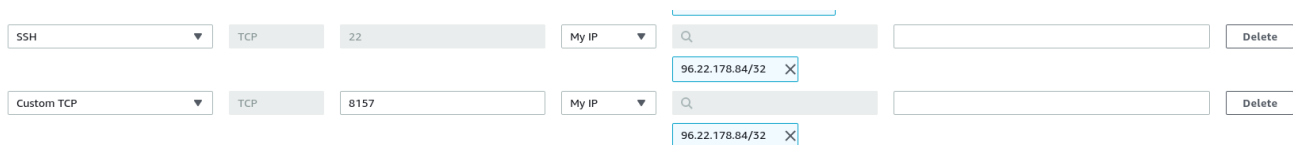
Click on your cluster in Amazon EMR page. Then, access the security group page by clicking on security groups of the master.



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*.

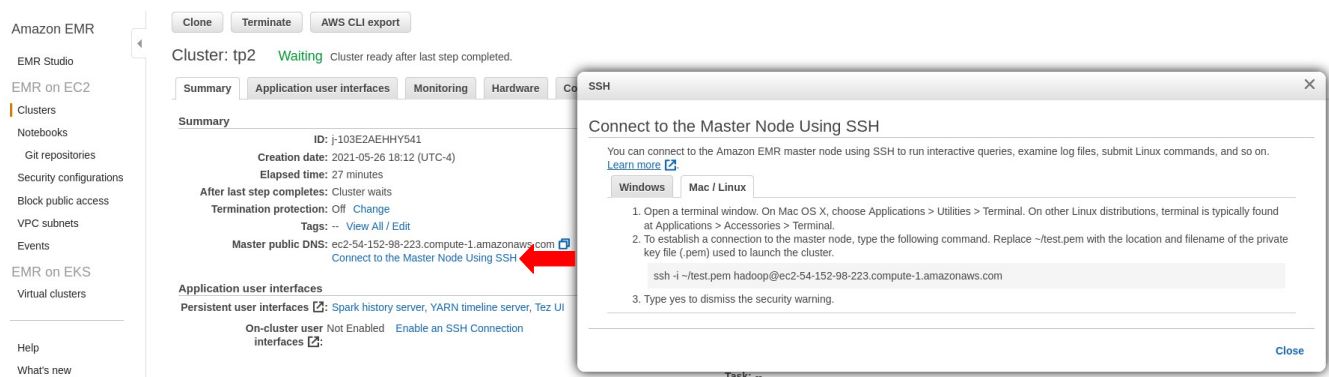


**Note:** Since your IP is probably dynamic, you might need update your IP during the TP.

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

```
ssh -i your_key.pem hadoop@ec2-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*

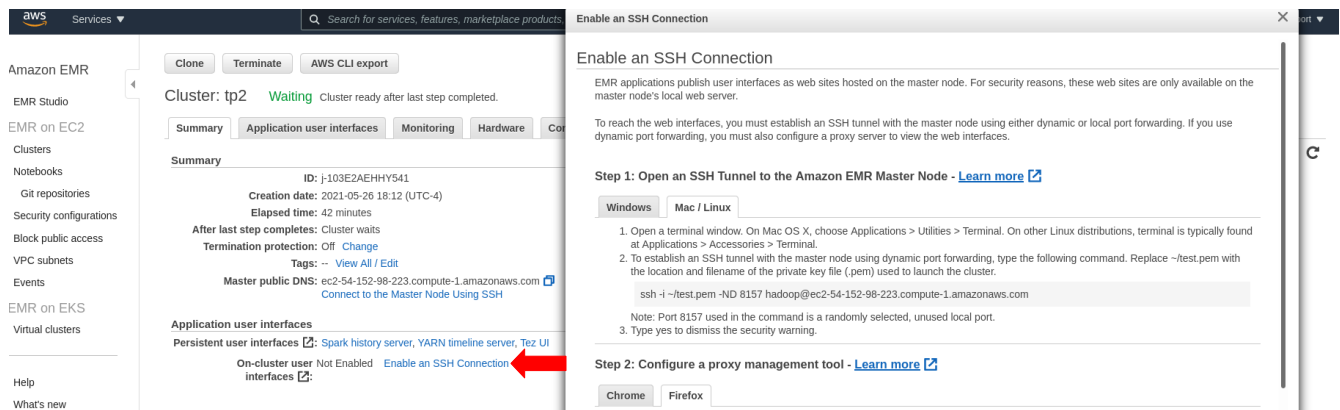


**Note:** “your\_key.pem” is the key created in the second section of this document.

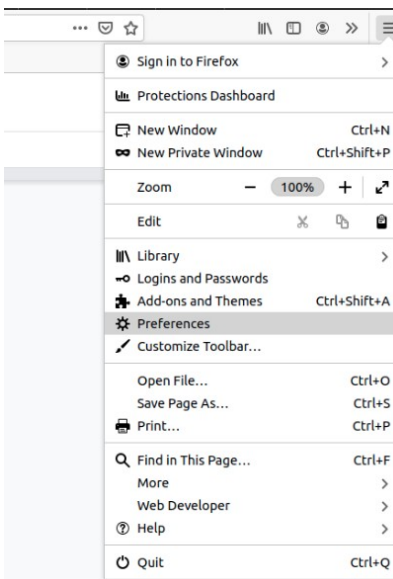
Now, run the following command to create a tunnel:

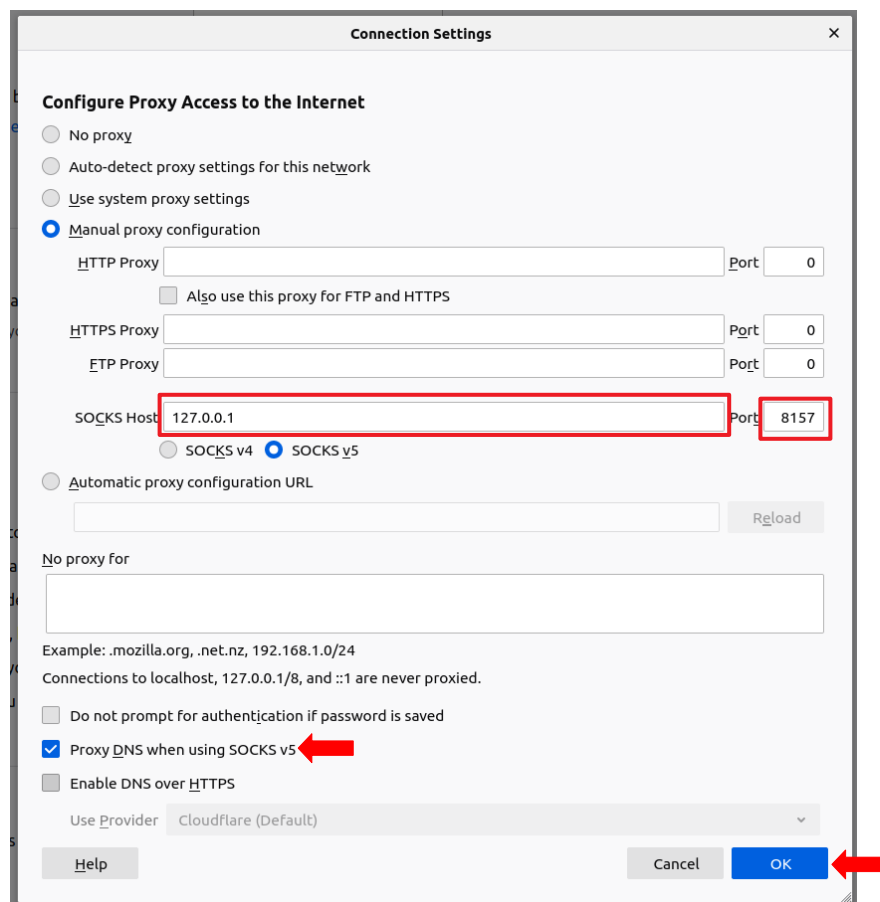
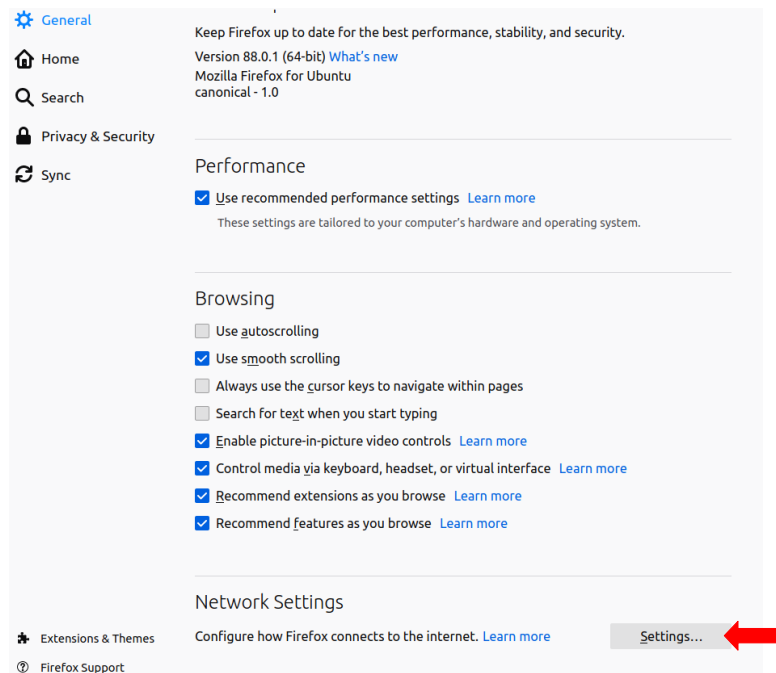
```
ssh -i .ssh/your_key.pem -ND 8157 hadoop@ecX-XX-XXX-XX-XXX.compute-1.amazonaws.com
```

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.





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
```

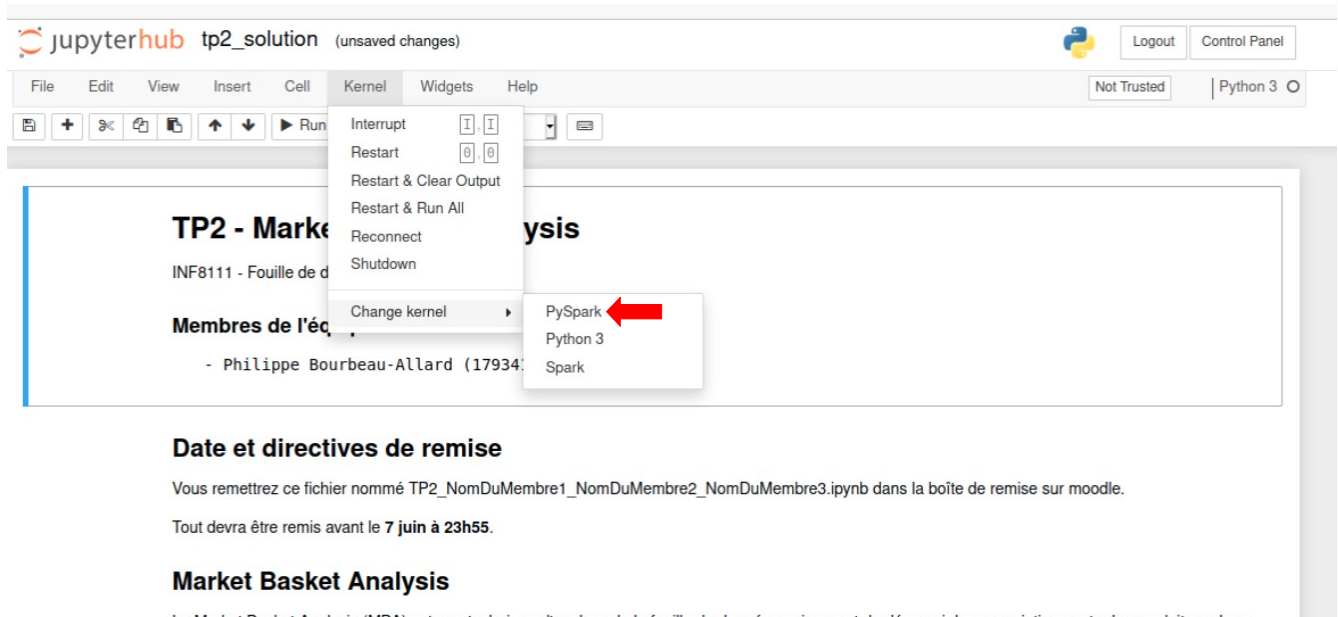
**Note:** replace `ecX-XX-XXX-XX-XXX.compute-1.amazonaws.com` by the correct master url

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()

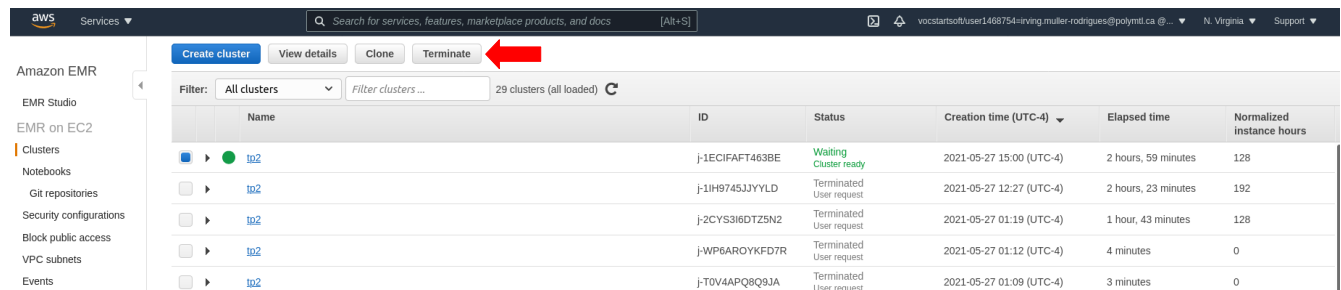
# Your code

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*.



The screenshot shows the AWS Management Console for Amazon EMR. The left sidebar contains navigation links for Amazon EMR, EMR Studio, EMR on EC2, Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, and Events. The main content area shows the 'Clusters' page with a filter set to 'All clusters' and 29 clusters loaded. A table lists the clusters with columns for Name, ID, Status, Creation time (UTC-4), Elapsed time, and Normalized instance hours. The first cluster, j-1ECIFAF463BE, is in 'Waiting Cluster ready' status. The other four clusters are in 'Terminated User request' status. The 'Terminate' button is highlighted with a red arrow.

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
td2	j-1ECIFAF463BE	Waiting Cluster ready	2021-05-27 15:00 (UTC-4)	2 hours, 59 minutes	128
td2	j-1IH9745JJYYLD	Terminated User request	2021-05-27 12:27 (UTC-4)	2 hours, 23 minutes	192
td2	j-2CYS3I6DTZ5N2	Terminated User request	2021-05-27 01:19 (UTC-4)	1 hour, 43 minutes	128
td2	j-WP6AR0YKFD7R	Terminated User request	2021-05-27 01:12 (UTC-4)	4 minutes	0
td2	j-T0V4APQ8Q9JA	Terminated User request	2021-05-27 01:09 (UTC-4)	3 minutes	0

**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*.