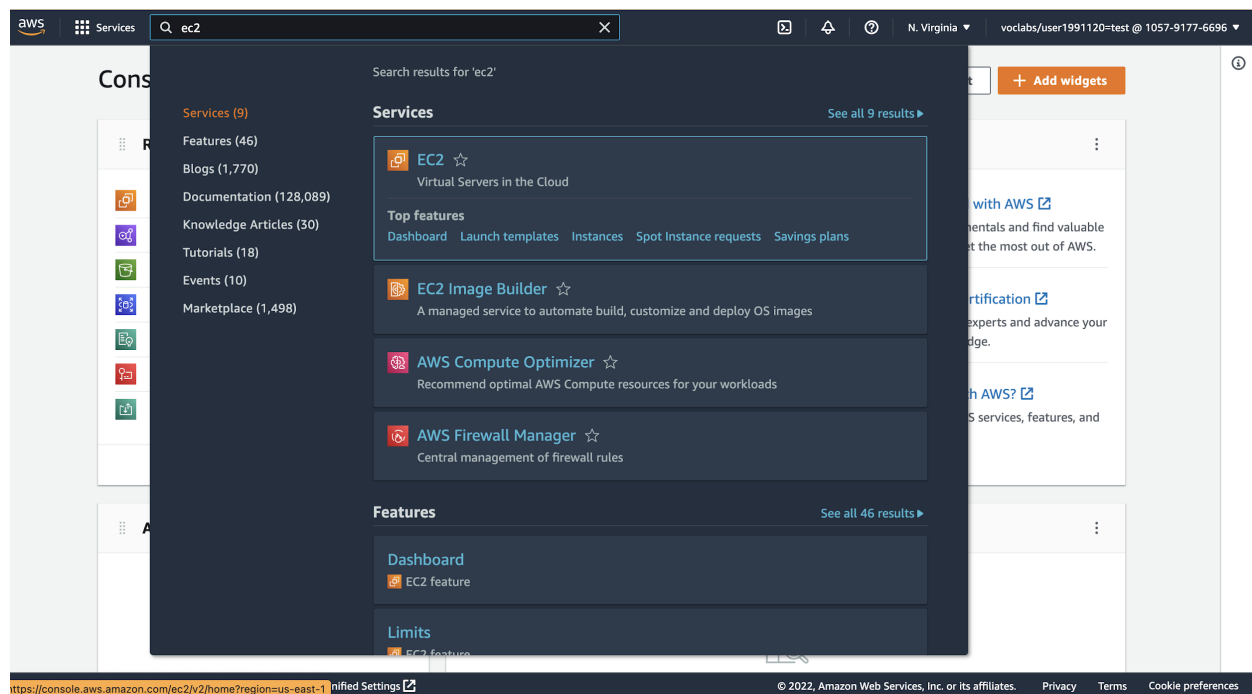


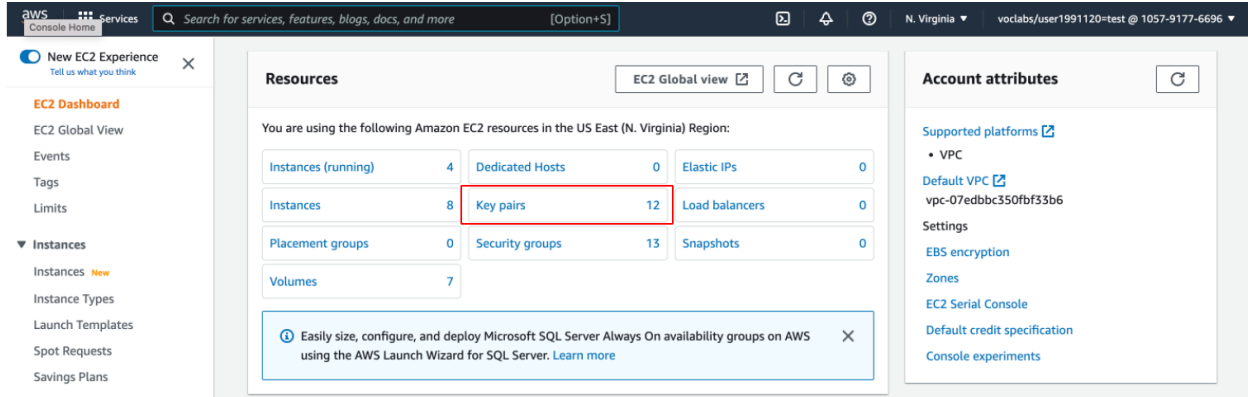
# Setting up Apache Spark instance and Jupyter Notebook

This document contains the steps to create an EMR cluster for using Apache Spark.

1. Before you create your EMR cluster, you will need to create a key pair. This is because your EMR cluster will be running on EC2 instances and you will require a key pair to connect with your instance. Let us quickly revise how you can create your key pair using the console.
  - a. Once you have logged in to your AWS account, search for 'EC2' under 'Find Services' and click on it.

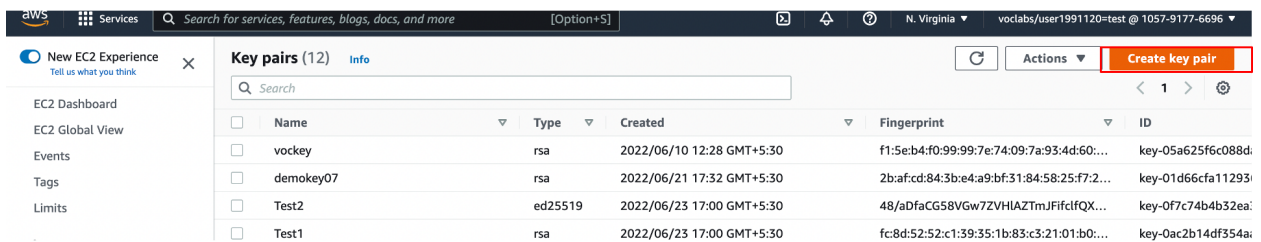


- b. Under 'Resources', click on 'Key Pairs'.



The screenshot shows the AWS Management Console interface. On the left, there's a sidebar with navigation options like 'EC2 Dashboard', 'Events', 'Tags', 'Limits', and 'Instances'. The main area is titled 'Resources' and shows a summary of EC2 resources in the 'US East (N. Virginia) Region'. A table lists various resources: Instances (running: 4, total: 8), Dedicated Hosts (0), Elastic IPs (0), Key pairs (12, highlighted with a red box), Load balancers (0), Placement groups (0), Security groups (13), Snapshots (0), and Volumes (7). On the right, there's a section for 'Account attributes' including 'Supported platforms', 'Default VPC', and 'Settings'.

c. Now click on 'Create key pair'



The screenshot shows the 'Key pairs (12)' page in the AWS Management Console. It features a search bar and a table of existing key pairs. The 'Create key pair' button in the top right corner is highlighted with a red box.

Name	Type	Created	Fingerprint	ID
vockey	rsa	2022/06/10 12:28 GMT+5:30	f1:5e:b4:f0:99:99:7e:74:09:7a:93:4d:60:...	key-05a625f6c088d...
demokey07	rsa	2022/06/21 17:32 GMT+5:30	2b:af:cd:84:3b:e4:a9:bf:31:84:58:25:f7:2:...	key-01d66cfa11293...
Test2	ed25519	2022/06/23 17:00 GMT+5:30	48/aDfaCG58VGw7ZVHIAZTmJFifclFQX...	key-0f7c74b4b32ea...
Test1	rsa	2022/06/23 17:00 GMT+5:30	fc:8d:52:52:c1:39:35:1b:83:c3:21:01:b0:...	key-0ac2b14df354a...

d. Give a name to your key pair. In our case, we have named it as RHEL and used the pem File format. Now click on 'Create key pair'.

**Create key pair** [Info](#)

**Key pair**  
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

**Name**  
  
The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

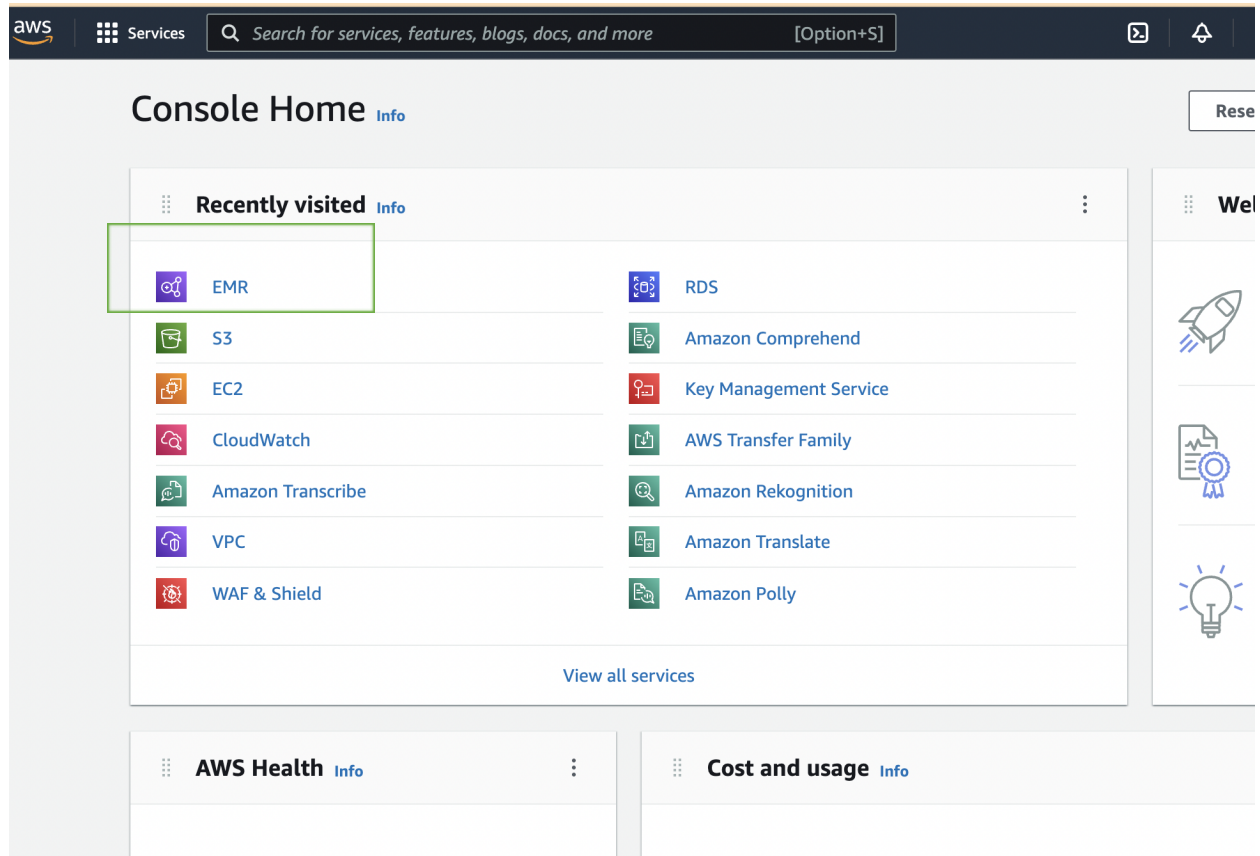
**Key pair type** [Info](#)  
☒ RSA  
☐ ED25519

**Private key file format**  
☒ .pem  
For use with OpenSSH  
☐ .ppk  
For use with PuTTY

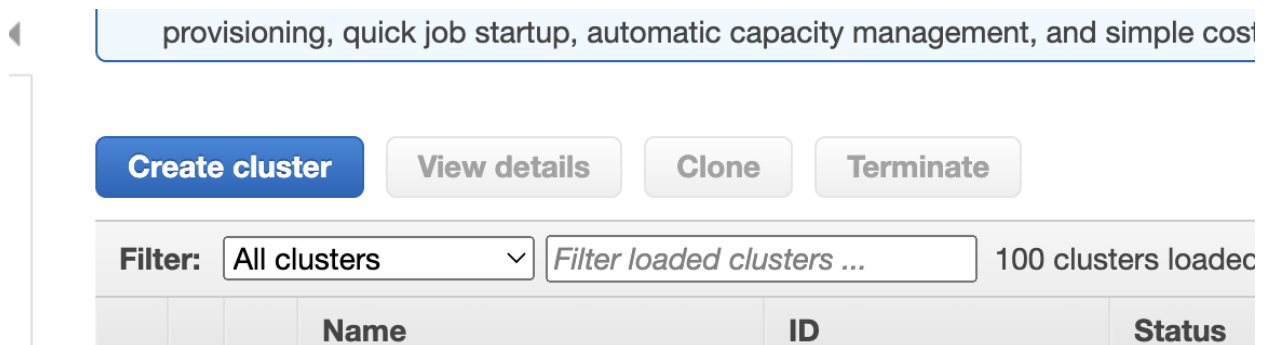
**Tags - optional**  
No tags associated with the resource.  
[Add new tag](#)  
You can add up to 50 more tags.

[Cancel](#) [Create key pair](#)

- e. Great! You now have your key pair and you can proceed with launching your EMR cluster
2. Click on the **Services** at the top of the AWS console and then click on the **EMR** service.



- Click on the **Create cluster** button and you will go to the cluster creation page.



- Once you click on the 'Create cluster' button, you need to click on the **Go to advanced options** link.

automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

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

### General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

5. It will take you to the following page. In the 'Release' column, you will be choosing the **emr-6.7.0** version for your EMR cluster. You will now configure the software applications that you need for your EMR cluster. For HBase, you need to install the following services -

- **Hadoop**
- **Spark**
- **Zeppelin**
- **JupyterHub**
- **JupyterEnterpriseGateway**
- **Livy**

### Software Configuration

Release  ⓘ

<input checked="" type="checkbox"/> Hadoop 3.2.1	<input checked="" type="checkbox"/> Zeppelin 0.10.0	<input checked="" type="checkbox"/> Livy 0.7.1
<input checked="" type="checkbox"/> JupyterHub 1.4.1	<input type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.14.2
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 2.4.4	<input type="checkbox"/> Pig 0.17.0
<input type="checkbox"/> Hive 3.1.3	<input type="checkbox"/> Presto 0.272	<input type="checkbox"/> ZooKeeper 3.5.7
<input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.1.0	<input type="checkbox"/> MXNet 1.8.0	<input type="checkbox"/> Sqoop 1.4.7
<input type="checkbox"/> Hue 4.10.0	<input type="checkbox"/> Phoenix 5.1.2	<input type="checkbox"/> Trino 378
<input type="checkbox"/> Oozie 5.2.1	<input checked="" type="checkbox"/> Spark 3.2.1	<input type="checkbox"/> HCatalog 3.1.3
<input type="checkbox"/> TensorFlow 2.4.1		

6. Click on the **Next** button at the end of the page.
7. In this page, scroll down to the 'Cluster Nodes' and 'Instances' section. You will now have to click on the **Cross** button to the right of the **Task Node**, and then under **Core Node**, you will need to type **0** under Instances. Do not create a multiple node cluster, otherwise, you might consume the entire budget in a single day.

## 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
<b>Master</b> Master - 1	<b>m5.xlarge</b> 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
<b>Core</b> Core - 2	<b>m5.xlarge</b> 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
<b>Task</b> Task - 3	<b>m5.xlarge</b> 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price

+ Add task instance group

8. Now, you need to click on the pencil button to the right of **m5.xlarge**.

## 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
<b>Master</b> Master - 1	<b>m4.xlarge</b> 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 40 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
<b>Core</b> Core - 2	<b>m4.large</b> 2 vCore, 8 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price

+ Add task instance group

9. Now, you need to select the **m4.xlarge** instance type under the window that appears, and then click on the **Save** button. You can use 'Control + F' keys and search for m4.xlarge.

Instance types				
<input type="radio"/>	m2.4xlarge	8	68.4	1690 SSD
<input type="radio"/>	m3.xlarge	4	15	80 SSD
<input type="radio"/>	m3.2xlarge	8	30	160 SSD
<input type="radio"/>	m4.large	2	8	EBS only
<input checked="" type="radio"/>	m4.xlarge	4	16	EBS only
<input type="radio"/>	m4.2xlarge	8	32	EBS only
<input type="radio"/>	m4.4xlarge	16	64	EBS only
<input type="radio"/>	m4.10xlarge	40	160	EBS only
<input type="radio"/>	m4.16xlarge	64	256	EBS only
<input type="radio"/>	m5.xlarge	4	16	EBS only
<input type="radio"/>	m5.2xlarge	8	32	EBS only
<input type="radio"/>	m5.4xlarge	16	64	EBS only

Cancel
Save

10. Now, you need to click on the pencil button to the right of the EBS storage.



11. Next, you need to configure the EBS volume for your EMR cluster. Click on the 'Volume type' and select the **General Purpose SSD (GP2)** option, and then under the 'Size' column, type **40**. Remove any other EBS volumes if present. After this, you can now click on the **Done** button. Thereafter, you can click on the **Next button** for this step of the advanced options as well.

Add EBS volumes

☒ EBS-Optimized instance

Volume type	Size (GiB)	IOPS	Throughput (MB/sec)	Volumes per instance
General Purpose SSD (GP2)	40	120/3000	Not Applicable	1

Min: 1 GiB, Max: 16384 GiB

Add EBS volumes

Cancel

Done

12. Now, in this step like the previous method, type the Cluster name that you want for your EMR cluster. Also, uncheck the 'Termination protection' option as this is not needed for this EMR instance. You can now click on the **Next** button.

General Options

Cluster name

PySparkCluster

☒ Logging

S3 folder

s3://aws-logs-105791776696-us-east-1/elasticma

☐ Log encryption

☒ Debugging

☐ Termination protection

13. Now in this step, you just need to select the EC2 key pair that you had created previously, and after that, you can click on the **Create cluster** button.



## Security Options

**EC2 key pair** RHEL ⓘ

☒ Cluster visible to all IAM users in account ⓘ

**Permissions** ⓘ

☐ Default ☒ Custom

Select custom roles to tailor permissions for your cluster.

**EMR role** EMR\_DefaultRole ⓘ

**EC2 instance profile** EMR\_EC2\_DefaultRole ⓘ

**Auto Scaling role** Proceed without role ⓘ

▸ 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 <a href="#">automatically update</a> the selected group	Additional security groups EMR will not modify the selected groups
Master	Default: sg-00fcc219431b5c0a6 (ElasticMapReduce~)	No security groups selected ✎
Core & Task	Default: sg-0a724c5cbac439160 (ElasticMapReduce~)	No security groups selected ✎

[Create a security group](#)

Cancel Previous Create cluster

14. Thereafter, the cluster will start setting up.

Amazon EMR
EMR Studio
EMR Serverless New
EMR on EC2
Clusters

EMR Serverless is now GA.  
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

Clone
Terminate
AWS CLI export

Cluster: PySparkCluster Starting

Once you have the EMR running, follow the steps to launch jupyter notebook:

## Creating a Jupyter Notebook on an EMR Cluster

Once you have created and started your EMR cluster, you can then create a new Jupyter Notebook wherein you can write your Spark jobs. Note that these notebooks are persisted on S3 even after you terminate the EMR cluster, so you don't have to worry about creating a Jupyter Notebook again from scratch. You can follow the steps below to create a Jupyter Notebook:

**Step 1:** First, you need to go to the 'Notebooks' link in the left navigation pane under Amazon EMR.




**Step 2:** On clicking the link, your screen will appear as shown below. Now, to create a new Jupyter Notebook, simply click on the 'Create notebook' button.

### Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create independently of clusters. Standard billing for clusters and Amazon S3 apply. [Learn more](#)

Create notebookView detailsOpen in JupyterLabOpen in JupyterStartStopDelete

**Filter:** All notebooks ▾  2 notebooks (all loaded) 

	Name
<input type="radio"/>	SparkNotebook
<input type="radio"/>	MyNewNoteBook

Once you click on the button, the following page will open on your screen.

## Create notebook

### Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. [Learn more](#)

**Notebook name\***

Names may only contain alphanumeric characters, hyphens (-), or underscores (\_).

**Description**

256 characters max.

**Cluster\***

☒ Choose an existing cluster

☐ Create a cluster
 

?

**Security groups**

☒ Use default security groups
 

?

☐ Choose security groups
 

?

**AWS service role\***

?

**Notebook location\***

Choose an S3 location where files for this notebook are saved.

☒ Use the default S3 location
 

s3://aws-emr-resources-864328032829-us-east-1/notebooks/

☐ Choose an existing S3 location in us-east-1
 

?

**Git repository**

Link to a Git repository

**Tags**

?

\* Required

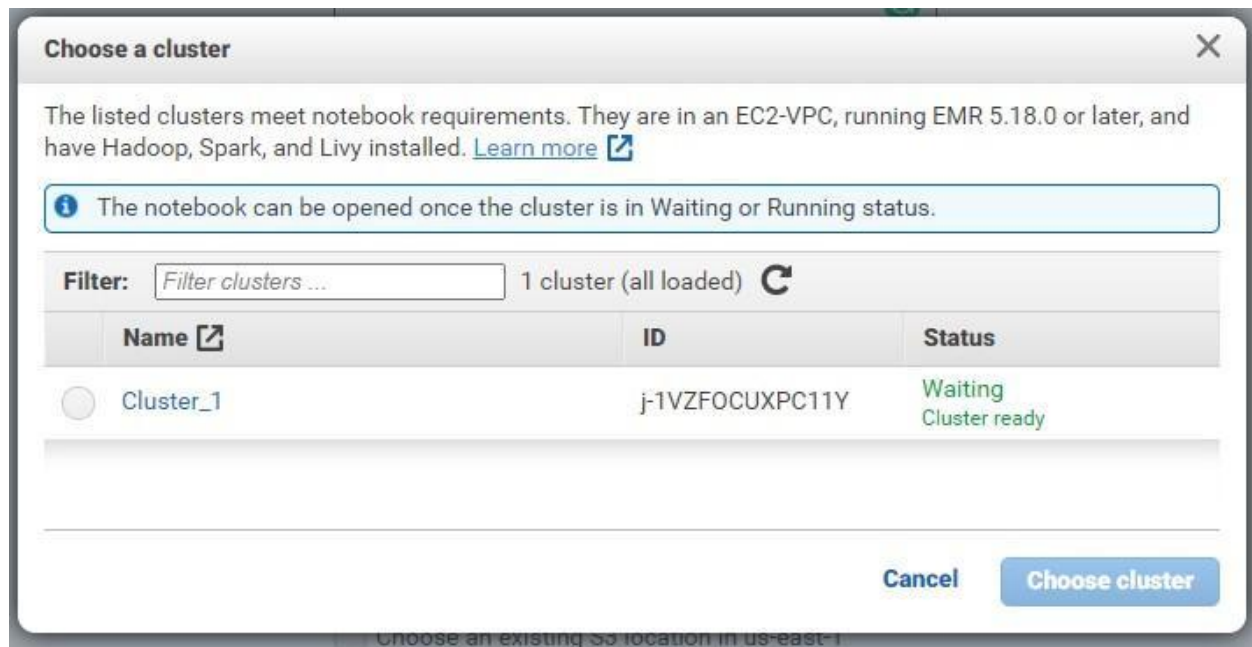
**Step 3:** Here, you can write a 'Notebook name' for your Jupyter Notebook. Under 'Description', you can write a few lines to describe the notebook that you are creating.

**Notebook name\***

Names may only contain alphanumeric characters, hyphens (-), or underscores (\_).

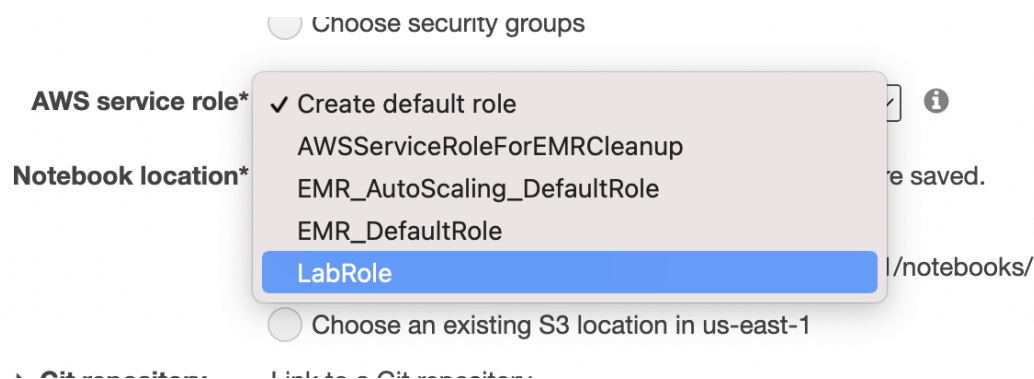
**Description**

**Step 4:** Now, under 'Cluster', click the radio button for 'Choose an existing cluster' and then click on the 'Choose' button. As soon as you do that, a pop-up will appear, showing you the list of all currently running EMR clusters.



**Step 5:** In this step, just select the EMR cluster that you have created and then click on the 'Choose cluster' button at the bottom right corner of the page.

**Step 6:** Next, in AWS service role select '**LabRole**'



**Step 6:** After this step, you can keep the other settings as default and click on the 'Create notebook' button at the bottom right of the page. As soon as you do this, the following page will open on your screen.

Notebook: Notebook1\_test **Starting** Starting notebook for cluster j-1VZF0CUXPC11Y.

[Open in JupyterLab](#) [Open in Jupyter](#) [Stop](#) [Delete](#)

## Notebook

**Notebook ID:** e-6CPAK9W6ZUVCOVW0DQ41N6Z4P  
**Description:** This is a test notebook.  
**Last modified by:** ...federated-user/rishav.talwar@upgrad.com  
**Created on:** 2020-09-04 00:01 (UTC+5:30)  
**Created by:** ...federated-user/rishav.talwar@upgrad.com  
**Service IAM role:** EMR\_Notebooks\_DefaultRole  
**Notebook tags:** creatorUserId = 864328032829-rishav.talwar@upgrad.com [View All / Edit](#)  
**Notebook location:** s3://aws-emr-resource-864328032829-us-east-1/notebooks/

## Cluster

**Cluster:** Cluster\_1  
**Cluster ID:** j-1VZF0CUXPC11Y  
**Cluster status:** **Waiting** Cluster ready after last step completed.  
**Cluster tags:** --  
**Step logs:** --

## Git repositories

The repository can be linked to a notebook once the notebook is ready. Make sure your cluster, service role and security groups have the required settings. [Learn more](#)

[Link new repository](#) [Unlink repository](#)

Repository name	URL	Branch	Link status	Failure reason
-----------------	-----	--------	-------------	----------------

This means the Notebook has been created and is now starting. Shortly after, the notebook will show at the top with status 'Ready'.

**Step 7:** Finally, after the status of the Jupyter Notebook shows 'Ready', you can launch the notebook. To do this, simply click on the '**Open in Jupyter**' button.

Notebook: Notebook1\_test **Ready** Notebook is ready to run jobs on cluster j-1VZF0CUXPC11Y.

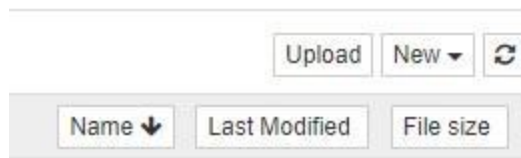
[Open in JupyterLab](#) [Open in Jupyter](#) [Stop](#) [Delete](#)

This will open the familiar Jupyter UI. Here, you can start creating your own Jupyter Notebooks as you have done in the previous module on Spark.

**Note:** Please note that if you want to work with Apache Spark, you will need to set the kernel of your notebook to ""

You can also upload any Jupyter Notebook that you want easily. For this, simply click on the 'Upload' button to the top right on the Jupyter UI.

	Name	Last Modified	File size
<input type="checkbox"/>	Notebook1_test.ipynb	10 minutes ago	72 B
<input type="checkbox"/>	Project1.ipynb	in a few seconds	14.1 kB



A Windows 'Open' dialogue box will appear. From here, you can simply find the location of your Jupyter Notebook and then click on the 'Open' button. After this, you will see that the name has been appended to the list of notebooks. You now need to click on the 'Upload' button next to your notebook file.

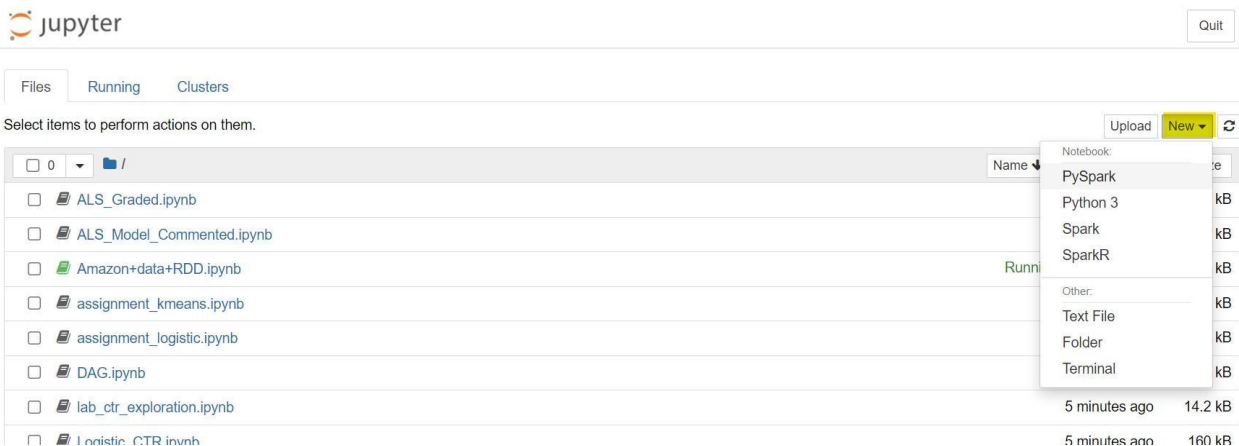


This will upload the Jupyter Notebook file to your EMR notebooks folder.

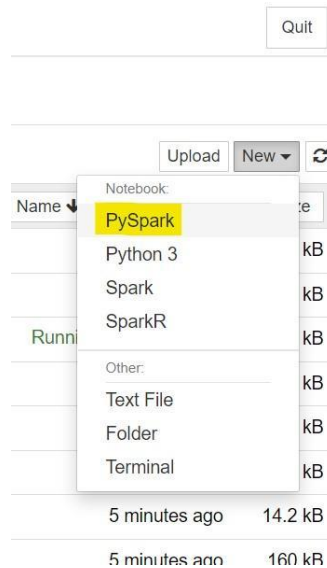
## Using Jupyter Notebook with Apache Spark

If you want to create a new Jupyter notebook to be used with Apache Spark then you need to follow these steps:

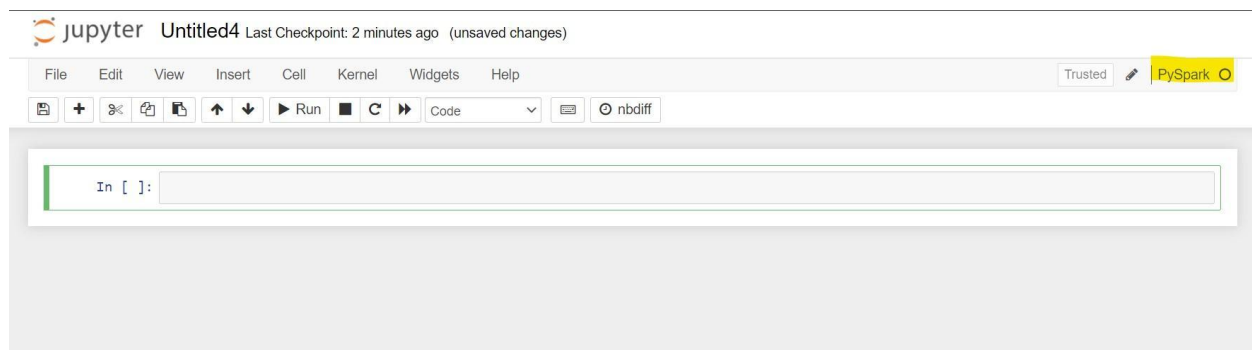
- Click on the **new button** to the top right of the Jupyter homepage.



- You will then have to select **PySpark** as your kernel in the drop-down menu. Click on PySpark

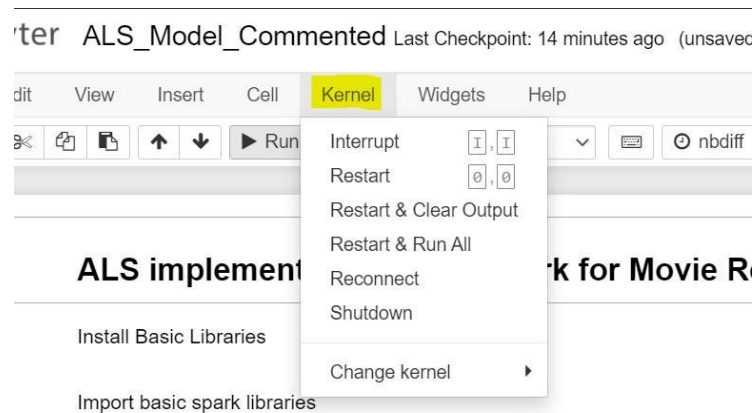


- This will then open a new window where your PySpark notebook will open up.

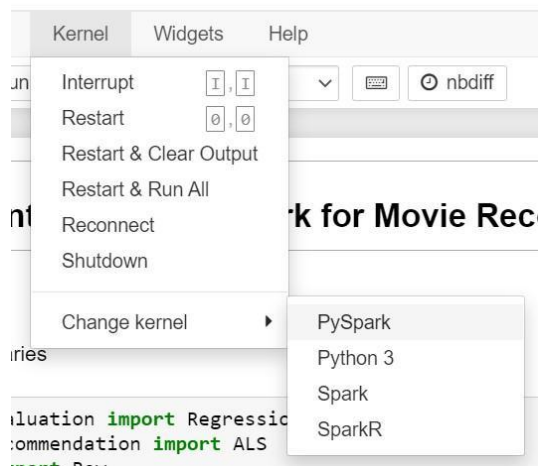


**Note:** Please note that if you want to work with Apache Spark on an older notebook or a notebook that you just uploaded to Jupyter, you might need to set the kernel of your notebook to “PySpark”. You can do this by following these steps:

- Click on **Kernel** on the top menu of your notebook as shown in the image below.



- Go to Change kernel at the end of this menu and then click on PySpark. This will change your kernel to PySpark in a few seconds. You can then start working on your PySpark Jupyter Notebook



## Stopping a Jupyter Notebook

You can also stop the Notebook whenever you want by simply clicking on the **Stop** button in the Notebook UI, as shown below:

Notebook: Notebook1\_test **Ready** Notebook is ready to run jobs on cluster j-1VZFOCUXPC11Y.

[Open in JupyterLab](#)

[Open in Jupyter](#)

[Stop](#)

[Delete](#)

Notebook



And, if you need to resume your Jupyter Notebook, then you can do so by going to the notebooks list, selecting your notebook, and clicking on the **Start** button.

## Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create and independently of clusters. Standard billing for clusters and Amazon S3 apply. [Learn more](#)

Create notebook
View details
Open in JupyterLab
Open in Jupyter
Start
Stop
Delete

Filter: All notebooks  3 notebooks (all loaded)

	Name
<input checked="" type="radio"/>	Notebook1_test
<input type="radio"/>	SparkNotebook
<input type="radio"/>	MyNewNoteBook