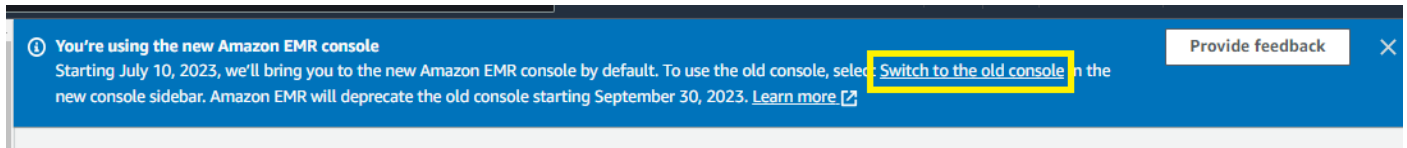


Important Note on Interface of EMR New VS Old Console

The following instructions detail the steps needed to work with EMR clusters using the new console.



The Amazon EMR console defaults to the new Amazon EMR console experience. If you want to go back to the old console, select '**Switch to the old console**' from the banner at the top of the console or from the side navigation and, you'll be redirected to the old EMR console.

To access the console, you may use the following links:

Old Console: <https://us-east-1.console.aws.amazon.com/elasticmapreduce/home?region=us-east-1>

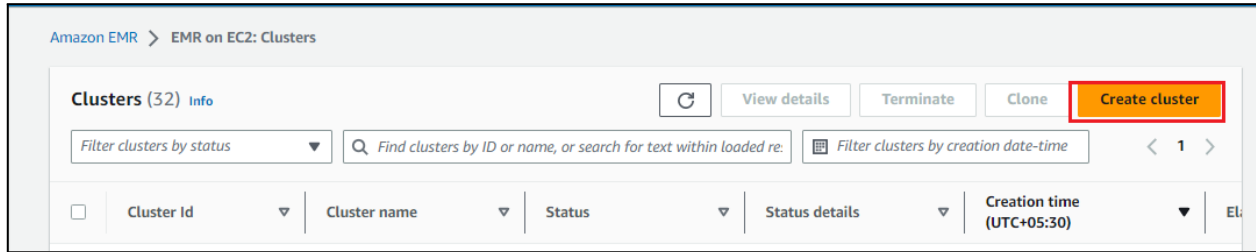
New Console: <https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1>

As you might note, the old console needs to have "**elasticmapreduce**" before "**/home?region=us-east-1**" in the link while the new console uses "**emr**".

The steps to create an EMR cluster in the new console is detailed below.

EMR CLUSTER CREATION

- You can access more information about the new console from this [link](#). Click on the 'Create Cluster' button.



- Now enter the details for the Cluster name, the EMR release/version, and the applications to be installed in the Cluster.

Create cluster Info

Name and applications Info

Name

Amazon EMR release Info

A release contains a set of applications which can be installed on your cluster.

Application bundle

Spark

Core Hadoop

HBase

Presto

Trino

Custom

Applications included in bundle

Spark 3.3.1 on Hadoop 3.3.3 YARN with and Zeppelin 0.10.1

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

☐ Use for Spark table metadata

Operating system options Info

☒ Amazon Linux release

☐ Custom Amazon Machine Image (AMI)

☒ Automatically apply latest Amazon Linux updates

- Click on the Custom option to configure the tools needed for your cluster.

Application bundle

Spark

Core Hadoop

HBase

Presto

Trino

Custom

▼ Customize your application bundle

Applications included in bundle

<input type="checkbox"/> Flink 1.16.0	<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 2.4.15
<input type="checkbox"/> HCatalog 3.1.3	<input checked="" type="checkbox"/> Hadoop 3.3.3	<input checked="" type="checkbox"/> Hive 3.1.3
<input checked="" type="checkbox"/> Hue 4.10.0	<input type="checkbox"/> JupyterEnterpriseGateway 2.6.0	<input type="checkbox"/> JupyterHub 1.5.0
<input type="checkbox"/> Livy 0.7.1	<input type="checkbox"/> MXNet 1.9.1	<input type="checkbox"/> Oozie 5.2.1
<input type="checkbox"/> Phoenix 5.1.2	<input checked="" type="checkbox"/> Pig 0.17.0	<input type="checkbox"/> Presto 0.278
<input type="checkbox"/> Spark 3.3.1	<input type="checkbox"/> Sqoop 1.4.7	<input type="checkbox"/> TensorFlow 2.11.0
<input type="checkbox"/> Tez 0.10.2	<input type="checkbox"/> Trino 403	<input type="checkbox"/> Zeppelin 0.10.1
<input type="checkbox"/> ZooKeeper 3.5.10		

- Choose the configuration of the Cluster.

Cluster configuration [Info](#)

Choose a configuration method for the primary, core, and task node groups for your cluster.

☒ Instance groups
Choose one instance type per node group

☐ Instance fleets
Choose any combination of instance types within each node group

Instance groups

Primary

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: - Lowest Spot price: -

Actions ▼

☐ Use multiple primary nodes
To improve cluster availability, use 3 primary nodes with the same configuration and bootstrap actions. You can not use multiple primary nodes with instance fleets.

► Node configuration - optional

Add core instance group

You must add a core instance group before you can add a task instance group.

► EBS root volume - optional

- In the next steps, you're required to configure the Networking settings for your EMR cluster. Select the pre-configured VPC, if any, or go with the defaults.

The screenshot shows the 'Networking' tab in the AWS EMR console. It includes sections for 'Virtual private cloud (VPC)', 'Subnet', and 'EC2 security groups (firewall)'. The VPC is set to 'vpc-08d90d01fd5e7e3d9' and the Subnet to 'subnet-01e9d68680dc030dd'. Under 'EC2 security groups', the 'Primary node' is set to 'ElasticMapReduce-Primary' and the 'Core and task nodes' are set to 'ElasticMapReduce-Core'. Both have dropdowns for 'Additional security groups - optional'.

- In the security configuration tab, select the key pair values. If you don't have any key pairs available, click on the "Create key pair" button.

The screenshot shows the 'Security configuration and EC2 key pair - optional' tab. It includes a section for 'Security configuration - optional' with a dropdown to 'Choose a security configuration' and buttons for 'Browse' and 'Create security configuration'. Below this, the 'Amazon EC2 key pair for SSH to the cluster - optional' section is highlighted with a red box. It contains a search bar with 'new' entered, a 'Browse' button, and a 'Create key pair' button. A dropdown menu is open showing the 'new' key pair with ID 'key-031c2c8e18a37a7e2' and the 'vockey' key pair with ID 'key-03e5c75c4141c229b'.

- In the IAM roles settings, use the following IAM roles

Service role: EMR_DefaultRole

EC2 Instance profile for Amazon EMR: EMR_EC2_DefaultRole

Identity and Access Management (IAM) roles [Info](#)

Choose or create a service role and instance profile for the EC2 instances in your cluster.

Amazon EMR service role [Info](#)

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

☒ Choose an existing service role
Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

☐ Create a service role
Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Service role

EMR_DefaultRole

▼

↻

EC2 instance profile for Amazon EMR

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

☒ Choose an existing instance profile
Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

☐ Create an instance profile
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile

EMR_EC2_DefaultRole

▼

↻

- Click on the “Create Cluster” button once the selected configuration is complete.

Amazon EC2 key pair for SSH to the cluster - optional [Info](#)

Identity and Access Management (IAM) roles [Info](#)

Choose or create a service role and instance profile for the EC2 instances in your cluster.

Amazon EMR service role [Info](#)

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

☒ Choose an existing service role

Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

☐ Create a service role

Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Service role

Summary [Info](#)

emr-6.10.0

Application bundle

Custom (Hadoop 3.3.3, Hive 3.1.3, Hue 4.10.0, JupyterEnterpriseGateway 2.6.0, JupyterHub...)

Cluster configuration

Instance groups

Primary (m4.xlarge)

Networking

Your cluster "EMR Cluster" has been successfully created.

Amazon EMR > EMR on EC2: Clusters > EMR Cluster

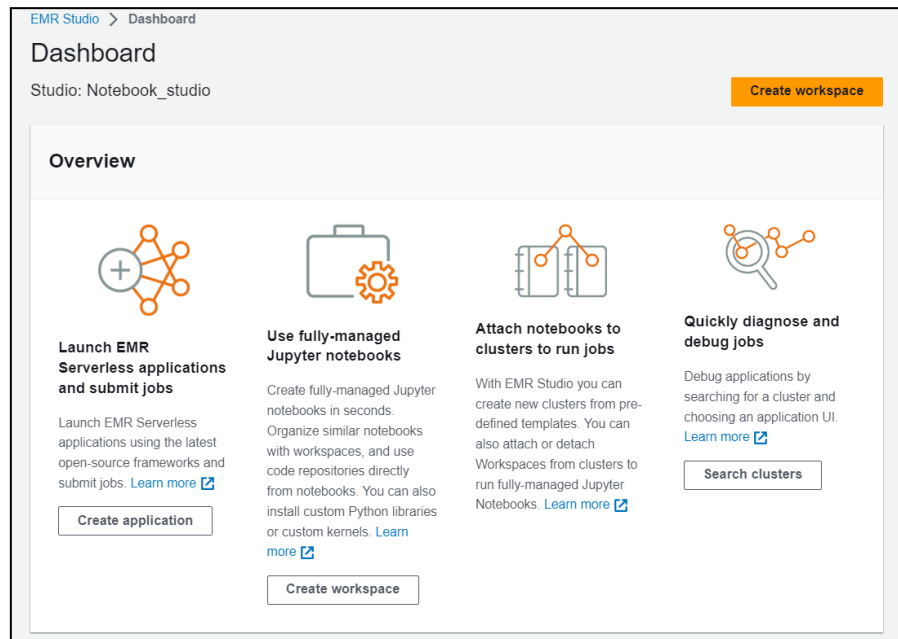
EMR Cluster Updated less than a minute ago

Summary

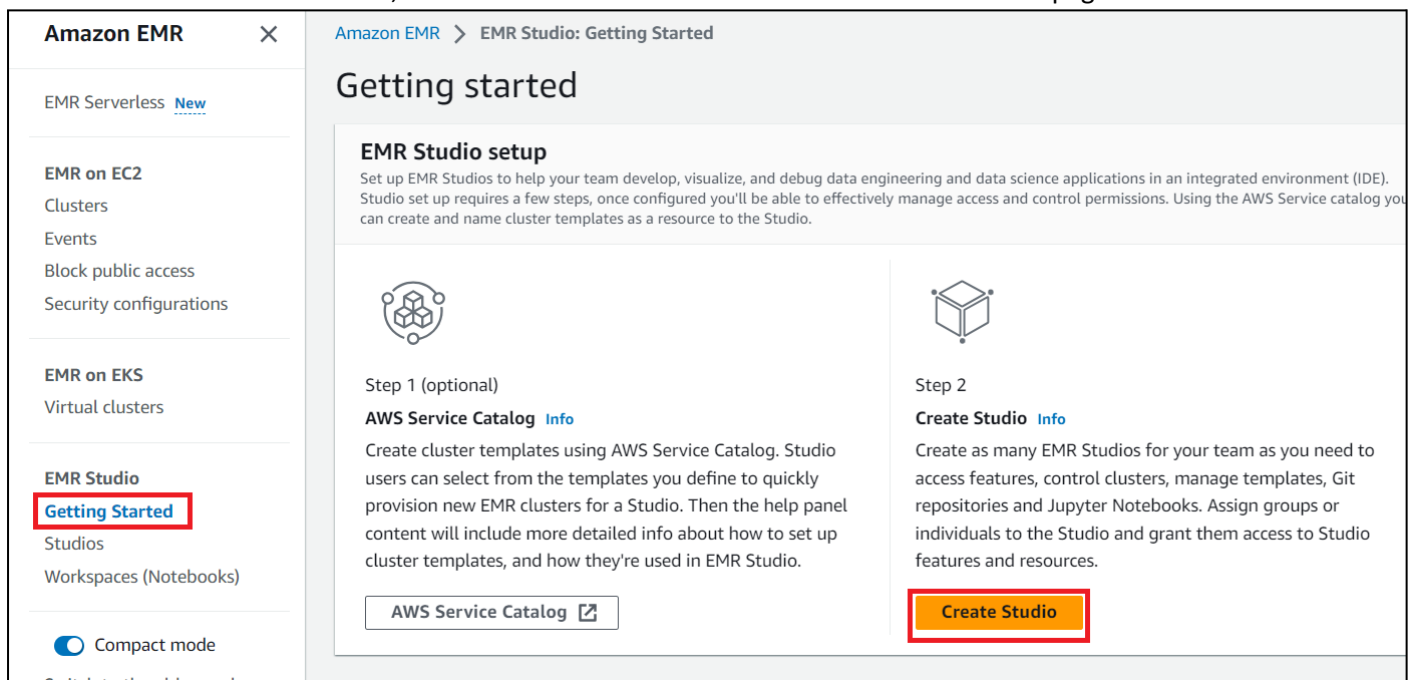
Cluster info	Applications	Cluster management	Status and time
Cluster ID j-HS06ENMV4U1Z	Amazon EMR version emr-6.10.0	Log destination in Amazon S3 aws-logs-811393130449-us-east-1/elasticmapreduce	Status Starting
Cluster configuration Instance groups	Installed applications Hadoop 3.3.3, Hive 3.1.3, Hue 4.10.0, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Pig 0.17.0, Spark 3.3.1	Primary node public DNS ec2-44-210-101-48.compute-1.amazonaws.com	Creation time May 24, 2023, 17:46 (UTC+05:30)
Capacity 1 Primary 0 Core 0 Task			Elapsed time 56 seconds

EMR Notebooks

In the new console, EMR Notebooks functionality has been replaced with EMR Studio which allows kernels and applications to run on EMR clusters.



- To start with EMR Studio, click on the 'Create Studio' button in the 'EMR Studio' page.



- Enter the name for the EMR Studio.

Amazon EMR > EMR Studio: Studios > Create a Studio

Create a Studio [Info](#)

Provide a unique name for your EMR Studio. This will be the name your Studio team sees when they log in. Set the networking and security configuration for the Studio and then assign roles and permissions. After you have created a Studio you will then be able to add users to it and refine permissions.

Studio name and description

Studio name

Notebook_studio

Use up to 256 characters (alphanumeric, hyphens, or underscores).

Description

Describe the Studio

256 characters maximum

Tags

No tags associated with the resource.

[Add new tag](#)

You can add 50 more tags.

- In the **Networking and Security** tab, select the VPC and subnet configurations for the EMR studio. If it's not showing up, create VPC and subnets by following the steps in the documentations available in the 'Introduction to Cloud' module.

Networking and security

VPC [Info](#)

Choose the VPC that EMR Studio can use to communicate with EMR clusters. Make sure the VPC is tagged with key `for-use-with-amazon-emr-managed-policies` and value `true`. To manage tags, use [VPC Dashboard](#).

vpc-08d90d01fd5e7e3d9

Subnets [Info](#)

Choose subnets that EMR Studio can use to communicate with EMR Clusters. Make sure each subnet is tagged with key `for-use-with-amazon-emr-managed-policies` and value `true`. To manage tags, use [VPC Dashboard](#).

Select subnets

subnet-07cc83a5f57b566a5 ✕

Security and access [Info](#)

Security groups act as firewalls with rules that allow network traffic between the EMR cluster and your workspace. You can use default security groups with the minimum required rules or specify your own security groups.

☒ **Default security group**

Select between enabling the user to only use the default EMR cluster or endpoint, or opt to include the ability to edit GIT repositories.

☐ **Custom security group**

Select from security groups for cluster or endpoints, or select a security group for the studio workspaces.

Default security group

☐ Enable clusters/endpoints and Git repository

☒ **Enable clusters/endpoints**

- In the Authentication setting, select 'AWS Identity and Access Management (IAM)' option and Service role as 'EMR_Notebooks_DefaultRole'

Authentication
Info

Choose an authentication method for your Studio.

☒ **AWS Identity and Access Management (IAM)**
Authenticate with single sign-on using IAM identity federation or IAM credentials.

☐ **AWS IAM Identity Center (successor to AWS Single Sign-On)**
Authenticate with single sign-on using IAM Identity Center (recommended to centrally manage access permissions for multiple AWS accounts).

Select identity provider - optional

Select your identity provider, which auto-populates your identity provider RelayState parameter name below. Provide only if you have enabled IAM federation and you want users to login via EMR Studio-generated URL.

Select your identity provider

Identity provider login URL - optional

Enter your identity provider login URL. Provide only if you have enabled IAM federation and you want users to login via EMR Studio-generated URL.

Enter the identity provider login URL

512 characters maximum

Identity provider RelayState parameter name - optional

This is the name of the RelayState parameter used by identity provider and differs based on the identity provider which you use. Provide only if you have enabled IAM federation and you want users to login via EMR Studio-generated URL.

Enter the RelayState parameter name

256 characters maximum

Service role

The service role defines the allowable actions for EMR Studio when provisioning resources. Examples of such actions include attaching a workspace to a cluster or accessing the S3 backup location for workspaces. [AWS IAM](#)

EMR_Notebooks_DefaultRole

- In the Workspace storage setting, choose the S3 location and click on the 'Create Studio' button.

Workspace storage

S3 Bucket

The S3 backup location for workspaces.

s3://bucket/prefix/object

View

Browse S3

Cancel

Create Studio

- The EMR Studio should shortly be created.

EMR on EKS

Virtual clusters

EMR Studio

Getting Started

Studios

Workspaces (Notebooks)

Compact mode

Amazon EMR > EMR Studio: Studios

Studios (2) Info

Find Studios by name or Studio access URL

View details

Edit

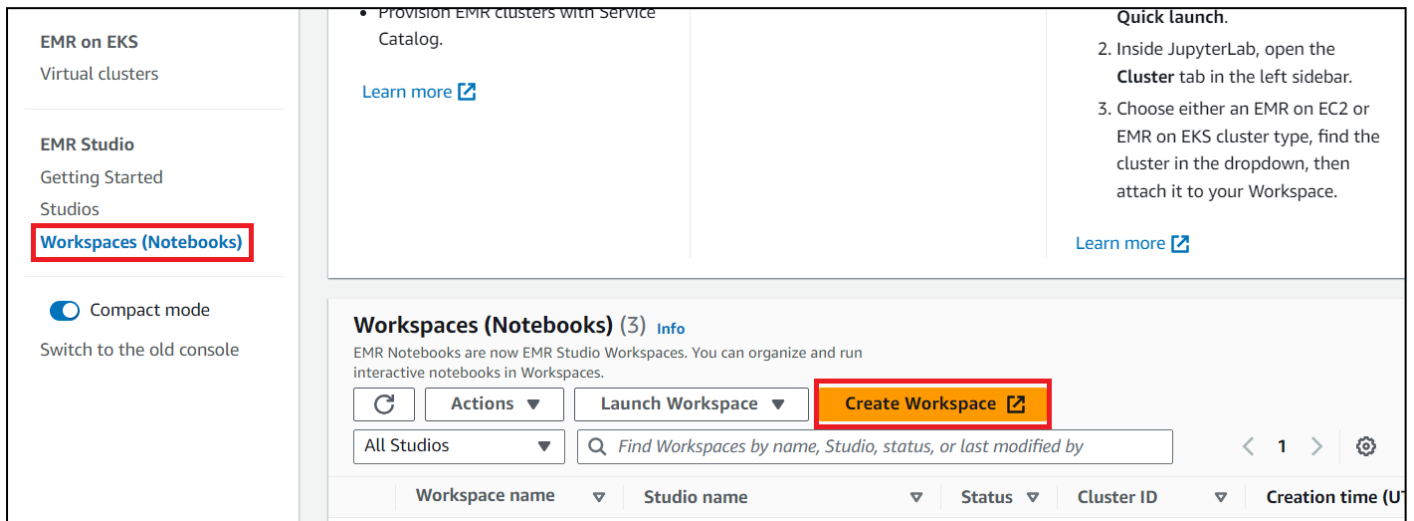
Delete

Create Studio

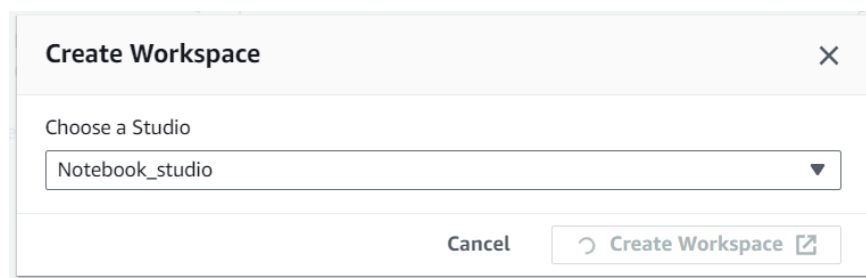
< 1 >

	Studio name	Creation time (UTC+05:30)	Authenticated by	Studio Access URL
<input type="radio"/>	Notebook_studio	July 04, 2023, 13:42	IAM	https://es-RXV30QXHHE5Z7559N1B...
<input type="radio"/>	Studio-vpc-08d90d01fd5e7e3d9-1	May 08, 2023, 17:50	IAM	https://es-4V9C3CIW35SCTVNO3XW...

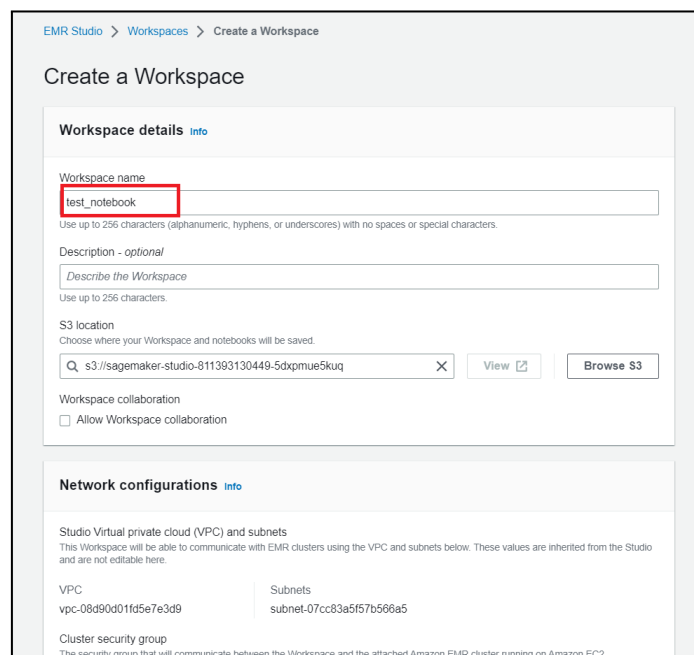
- Once the studio has been successfully created, click on the Workspaces(Notebook) option and click on the 'Create Workspace' button in the page.



- Select the Workspace Studio you've created.



- Enter the name of the Workspace in the space provided. The rest of the settings will be automatically populated from the settings provided in the Studio.



- Click on the 'Create Workspace' button to provision the workspace.

Network configurations [Info](#)

Studio Virtual private cloud (VPC) and subnets
This Workspace will be able to communicate with EMR clusters using the VPC and subnets below. These values are inherited from the Studio and are not editable here.

VPC
vpc-08d90d01fd5e7e3d9

Subnets
subnet-07cc83a5f57b566a5

Cluster security group
The security group that will communicate between the Workspace and the attached Amazon EMR cluster running on Amazon EC2.
sg-00ab0f816804da748 (DefaultEngineSecurityGroup)

Workspace security group
The security group that will allow the Workspace to route traffic to the internet and enable linking Git repositories to the Workspace.
sg-0bc515b490ab4bd22 (DefaultWorkspaceSecurityGroupWithoutGit)

After creating a Workspace, you can launch it to open up Jupyterlab. You can then attach it to an EMR compute cluster from the Cluster tab in the left sidebar within Jupyterlab.

Cancel **Create Workspace**

- Once the workspace is created, stop the workspace by clicking on the 'Actions' button and clicking on 'Stop'. Click on 'Launch Workspace' and then 'Launch with options'.

Workspaces (Notebooks) (1/4) [Info](#)

EMR Notebooks are now EMR Studio Workspaces. You can organize and run interactive notebooks in Workspaces.

All Studios ▼

Launch Workspace ▲

Quick launch [↗](#)

Launch with options [↗](#)

	Workspace name ▼	Studio name ▼	Status ▼	Cluster ID ▼	Creation time (UTC+05:30) ▼	Last modified by
●	test_notebook	Notebook_studio	Idle	j-1IVZ70GSLSA88	July 04, 2023, 13:44	use

- Select 'Launch in Jupyter' and the cluster id of the EMR cluster in the setting and click on 'Launch Workspace' button.

Launch Workspace with options ✕

☐ Launch in JupyterLab
Use the latest web-based interactive development environment for notebooks.

☒ **Launch in Jupyter**
Use the classic development environment for notebooks.

EMR cluster
Choose an EMR cluster to attach to your Workspace.
j-1IVZ70GSLSA88 (EMR Cluster) ▼

Cluster security group
Choose the security group that will communicate between the Workspace and the attached Amazon EMR cluster running on Amazon EC2.
sg-00ab0f816804da748 (DefaultEngineSecurityGroup) ▼

Workspace security group
Choose the security group that will allow the Workspace to route traffic to the internet and enable linking Git repositories to the Workspace.
sg-0bc515b490ab4bd22 (DefaultWorkspaceSecurityGroupWithoutGit) ▼

Cancel **Launch Workspace** [↗](#)

- The workspace will start shortly and a new jupyter page will appear.

Workspaces (Notebooks) (1/4) [Info](#)

EMR Notebooks are now EMR Studio Workspaces. You can organize and run interactive notebooks in Workspaces.

All Studios ▼

Find Workspaces by name, Studio, status, or last modified by

Workspace name ▼	Studio name ▼	Status ▼	Cluster ID ▼	Creation time (UTC+05:30) ▼
test_notebook	Notebook_studio	Starting	j-1IVZ70GSLSA88	July 04, 2023, 13:44

e-7xj1f8b6yarta5nhrfw903qrx.emrnotebooks-prod.us-east-1.amazonaws.com/e-7XJ1F8B6YARTA5NHRFW903QRX/tree?username=user2518953=henryisaac617+awsac22@gmail.com

jupyter

Files Running Clusters

Select items to perform actions on them.

0 /

test_notebook.ipynb

Upload

New

Refresh

Name

Last Modified

File size

2 hours ago

72 B

Disclaimer: All content and material on the upGrad website is copyrighted material, either belonging to upGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access print and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copies of this document, in part or full, saved to disc or to any other storage medium may only be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of the content for any other commercial/unauthorized purposes in any way which could infringe the intellectual property rights of upGrad or its contributors, is strictly prohibited.
- No graphics, images or photographs from any accompanying text in this document will be used separately for unauthorised purposes.
- No material in this document will be modified, adapted or altered in any way.
- No part of this document or upGrad content may be reproduced or stored in any other web site or included in any public or private electronic retrieval system or service without upGrad's prior written permission.
- Any rights not expressly granted in these terms are reserved.