

EMR Serveless

I am Roles

1. Create role for EMR notebook. [emr_notebook_role.json]
 - Create AmazonElasticMapReducePolicy. [AmazonElasticMapReduceEditorsRolepolicy.json]
 - Create AmazonS3FullAccessPolicy. [s3_full_access_policy.json]

Attach the above 2 policies to the **EMR notebook Role**.

2. Create role for EMR serverless Execution. [emr_serverless_role.json]
 - Create emr serverless policy.[emr_serverless_policy.json]

Attach the above policy to **EMR serverless Execution Role**.

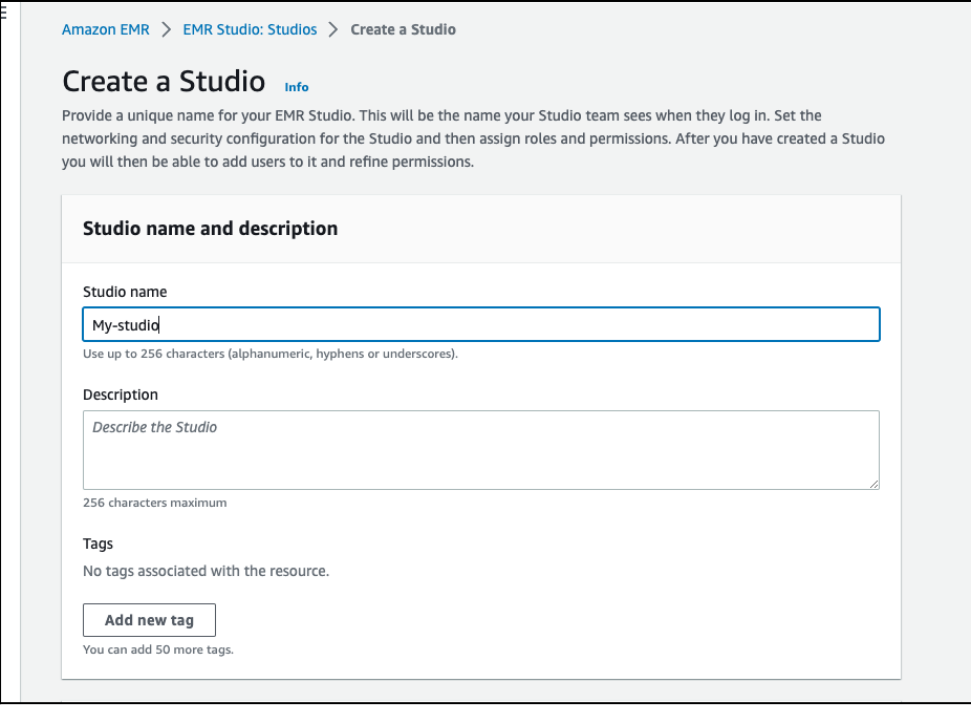
S3 Bucket Name : new-s3-bucket-cli-test

- Under s3 bucket create a folder called scripts to keep the Wordcount .py script or hive script [hive_statements.sql]
- Create another folder called query_results to store the outputs.
- Create another folder to upload a csv file required for running the script.

For submitting Spark job using UI Directly:

Go to the AWS console -> type EMR -> Click -> choose EMR Studio -> create studio to manage your applications.

Provide a name for the studio name.



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

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.

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](#).

Select a VPC ▼

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 ▼

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.

☐ **Customised 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

Under network and security. -> Choose the default VPC or your VPC. -> choose 2 or 3 subnets in the subnets field -> rest can be left with default options.

Authentication and IAM roles [Info](#)

Authentication [Info](#)
Choose an authentication method for your Studio.

☒ **AWS Identity and Access Management (IAM)**
Authenticate with single sign-on using AWS IAM identity Federation or AWS 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 log in 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 log in via an 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 the identity provider and it differs depending on which identity provider you use. Provide only if you have enabled IAM federation and you want users to log in 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](#)

Select a service role ▼

Under the authentication and i am roles -> in service field -> choose the created **EMR notebook Role**.

Workspace storage

S3 bucket
The S3 location where the Workspaces under this Studio will be backed up.

Q s3://bucket/prefix/object

View

Browse S3

Cancel Create Studio

Under workspace and storage -> In s3 Bucket field provide the bucket name directly or browse the bucket name-> click on create Studio.

Successful studio is created !

In the AWS EMR console -> click on EMR serveless.

Amazon EMR

- EMR Studio
- EMR Serverless New
- EMR on EC2

Amazon EMR has launched a new console experience. [Learn more](#) or [switch to the new console](#)

Create cluster View details Clone Terminate

Choose the Studio name you have created and click manage applications -> a tab will open take to the applications page.

Manage applications

To manage EMR Serverless applications, you need EMR Studio. Since you already have multiple studios in this region, choose one studio and then choose Manage applications to launch your studio and manage applications.

EMR Studio

Choose a studio ▼

Manage applications

Create Application to submit the Spark job or hive job.



Create an application.

Create application

You can submit multiple data processing jobs to an application. Each application corresponds to a specific big-data-processing framework and [EMR release version](#). You do not incur charges for creating an application.

Application settings

Name

May include up to 64 alphanumeric, underscore, hyphen, forward slash, hash, and period characters.

Type

Release version

Architecture

Choose an instruction set architecture (ISA) option for your application.

☒ **x86_64**
This architecture uses x86 processors and is compatible with most third-party tools and libraries.

☐ **arm64 - new**
This architecture uses AWS Graviton2 processors. You might have to recompile some third-party tools and libraries.

1. Under the application settings in the name provide a name for the application.
2. Under the type field choose spark or Hive Job.
3. Under the Release version choose the version from emr-6.6.0 to emr-6.9.0. Emr-6.9.0 is the recent version.
4. Leave the Architecture field with the default option.

Under the Application setup option leave with the default settings or choose custom settings to customise the application according to the job requirements. -> click on create application.

Application setup options [Info](#)

☒ **Choose default settings**
Get started quickly with pre-initialized capacity ready for running 1 job.

☐ **Choose custom settings**
Control all settings including pre-initialized capacity, network connection etc.

Default settings

Pre-Initialized capacity

Spark drivers

1

Size of driver

4 vCPUs, 16 GB memory, 20 GB disk

Spark executors

2

Size of executor

4 vCPUs, 16 GB memory, 20 GB disk

Application limits
400 vCPUs, 3000 GB memory, 20000 GB disk

Application behavior

Auto start application

Automatically starts on job submission

Auto stop application

Automatically stops after application is idle for 15 minutes

Network connections
No network connectivity to resources in your VPC

Tags
No tags

[Cancel](#)
[Create application](#)

Once the application created Successfully -> click on the app ->

my-new-app

[Refresh](#)
[Start application](#)
[Stop application](#)
[Action](#)

Application details

Application ID	ARN	Type
00f78hb964ic3l0p	arn:aws:emr-serverless:eu-west-1:880145880630/applications/00f78hb964ic3l0p	Spark
Status	Creation time	Release version
Stopped	Fri, 20 Jan 2023 17:41:18 GMT (UTC +0:00)	emr-6.9.0
	Last updated	Architecture
	Fri, 20 Jan 2023 18:12:56 GMT (UTC +0:00)	x86_64

[Properties](#)
[Job runs](#)
[Tags](#)

Job runs (4) [Info](#)

[Refresh](#)
[View application UIs](#)
[Clone job](#)
[Cancel job](#)
[Submit job](#)

Click on the submit job .

Job details
[Info](#)

Name

spark-job

Runtime role

The IAM role assumed by the job. This role must have permissions to access your data sources, targets, scripts, and any libraries used by the job. [Learn more](#)

serveless2ndclirole

Script location

The location of the main JAR or Python script in Amazon S3 that you want to run.

S3 URI

s3://emr-serveless/scripts/wordcount.py

View

Browse S3

Script arguments

An array of arguments passed to your main JAR or Python script. Your code should handle reading these parameters. Each argument in the array must be separated by a comma.

["argument_value_2", "argument_value_2" ...]

► Spark properties - optional

[Info](#)

Additional configuration properties that you can specify for each job. Amazon EMR uses default application properties to help you get started quickly.

► Job configuration - optional

[Info](#)

Job configurations allow you to override the default configurations for applications.

► Additional settings - optional

[Info](#)

Under job details application.

1. Provide a name for the spark job.
2. Choose the created **EMR serverless Execution Role**.
3. Under the script location filed provide the path for the wordcount .py.

► Tags - optional

[Info](#)

Labels that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Cancel

Submit job

Click on submit job.

Job runs (4) Info

Find job runs by name, ID, tags, or status

Any run status

Start time in last 30 days

< 1 >

Job run name

Run status

Job run ID

Start time (UTC +0:00)

End time (UTC +0:00)

Run time

new_spark_job

Success

00f78hk3lc3a8t0p

20 Jan 2023, 17:56

20 Jan 2023, 17:57

0 min, 60 secs

My_First-Spark-Job-clone

Failed

00f78hhid3vvbd0p

20 Jan 2023, 17:52

20 Jan 2023, 17:52

0 min, 11 secs

Under jobs will be able to see the job status like -> starting, Scheduling, running, failed or successful.

In the s3 bucket will be able to see the logs of the submitted spark job.

Submitting Hive Job

For submitting hive job create application -> choose -> Type ->hive and version -> create a hive application.

For submitting hive job -> job details -> give a name -> provide sql script uri.

▼ Job configuration [Info](#)

Job configurations allow you to override the default configurations for applications.

☒ Edit in JSON

☐ Load JSON from Amazon S3

1 ▼ {

2 ▼ "applicationConfiguration": [

3 ▼ {

4 "classification": "hive-site",

5 "configurations": {},

6 ▼ "properties": {

7 "hive.exec.scratchdir": "s3://DOC-EXAMPLE_BUCKET

8 "hive.metastore.warehouse.dir": "s3://DOC

9 }

10 }

11]

12 }

JSON Ln 1, Col 1

Errors: 0 Warnings: 0

[Copy](#)

Provide the s3 bucket name in the json and submit the job.

Submitting a spark job using AWS CLI.

1. To create s3 bucket

```
aws s3api create-bucket \
  --bucket kavitha-s3-test \
  --region eu-west-1 \
  --create-bucket-configuration LocationConstraint=eu-west-1
```

2. Create folder in s3 bucket and upload the word count.py file

```
aws s3api put-object --bucket kavitha-s3-test --key SparkScript/wordcount.py --body
/Users/kavitharajendran/Documents/wordcount.py
```

3. To create IAM policy.

```
aws iam create-policy --policy-name glue-policy --policy-document
file:///Users/kavitharajendran/Documents/emr-serverless/Glue.yml or json
```

Note: Provide the policy name and the path of the file yml or json file can be used.

4. To create an IAM Role.

```
aws iam create-role --role-name AwsEmrServerlessRole--assume-role-policy-document
file:///Users/kavitharajendran/Documents/emr-serverless/emr_notebook_role.yml or json
```

Note: provide the name and the path of the file.

5. To attach the policy with IAM role.

```
aws iam attach-role-policy --policy-arn
arn:aws:iam::880145880830:policy/S3fullaccesstocli-bucket --role-name
AwsEmrServerlessRoledemo
```

Note: Provide the arn of the policy and the role name to attach the policy to role.

6. To Create an application.

```
aws emr-serverless create-application \
  --release-label emr-6.6.0 \
  --type "SPARK" \
  --name was-cli-spark-application-test
```

Provide the name for the spark application, in type provide spark or hive job and provide the version.

7. For submitting the spark job.

```
aws emr-serverless start-job-run \
  --application-id 00f760th3gmvg0p --execution-role-arn
arn:aws:iam::880145880830:role/serveless2ndclirole --name my-application\
  --job-driver '{
    "sparkSubmit": {
```



```

    "entryPoint": "s3://kavitha-s3-test/scripts/wordcount.py",
    "entryPointArguments": ["s3://kavitha-s3-test/SparkScripts/output"],
    "sparkSubmitParameters": "--conf spark.executor.cores=1 --conf
spark.executor.memory=4g --conf spark.driver.cores=1 --conf spark.driver.memory=4g --conf
spark.executor.instances=1"
  }
}'

```

Note : Provide the name of the application, provide the arn of the **EMR serverless Execution Role** and application name.

8. To get the details of the created application

```
aws emr-serverless get-application \
```

```
--application-id 00f76naiqncul30p
```

Note: Provide application ID : "00f76naiqncul30p"

9. To get the status of the running job

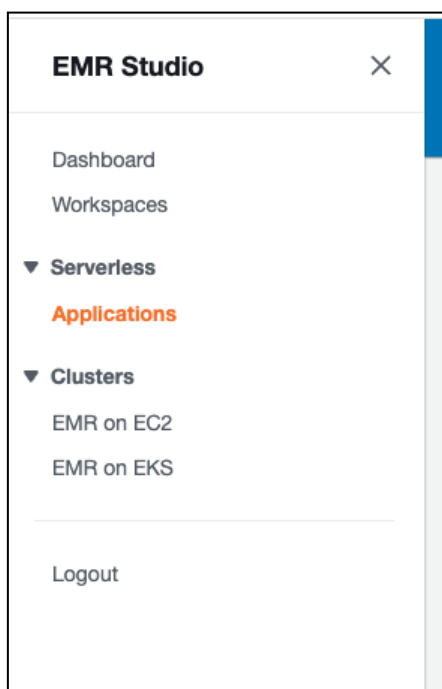
```
aws emr-serverless get-job-run \
```

```
--application-id 00f76oth3gmvog0p\
```

```
--job-run-id 00f76phhr1vv9h0p
```

Note: Provide application ID : "00f76oth3gmvog0p" and job run id: "00f76phhr1vv9h0p"

To create a Jupyter notebook :



Choose workspace - > create workspace.

EMR Studio > Workspaces > Create a Workspace

Create a Workspace [Info](#)

Name your Workspace, then choose a subnet and the S3 location to save your notebooks. See configuration options for additional ways to customize the Workspace.

Workspace definitions

Workspace name

Description - optional

S3 location
Choose where your Workspace and notebooks will be saved.

☒ Allow Workspace Collaboration

Advanced configuration
To run your fully-managed Jupyter Notebook, you need to attach the Workspace to an EMR cluster. You can create a new cluster or select a cluster template. Some of the advanced configuration options require access permissions from your administrator.

☐ Attach Workspace to an EMR cluster
Run your Workspace by choosing a cluster from a list of preset, running clusters.

☐ Create an EMR cluster
Provision a new EMR cluster for your Workspace.

☐ Use a cluster template
Provision a new EMR cluster from a pre-defined template.

☐ Attach Workspace to an EMR on EKS cluster
Attach an EKS cluster to your Workspace to submit and manage jobs.

Cancel **Create Workspace**

1. Provide a name for the workspace.
2. Create an EMR cluster or attach workspace to an EMR cluster which is created already.
3. Note: if you are attaching to the cluster which is created separately . make sure you choose (jupyterHub,hadoop, JupyterenterpriseGateway,spark,livy,hue,pig)
4. Click create workspace.

BWS Services Search [Option+S]

Amazon EMR has launched a new console experience. [Learn more](#) or [switch to the new console](#)

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 2.10.1	<input 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 1.4.13	<input checked="" type="checkbox"/> Pig 0.17.0
<input checked="" type="checkbox"/> Hive 2.3.9	<input type="checkbox"/> Presto 0.267	<input type="checkbox"/> ZooKeeper 3.4.14
<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"/> Mahout 0.13.0	<input checked="" type="checkbox"/> Hue 4.10.0	<input type="checkbox"/> Phoenix 4.14.3
<input type="checkbox"/> Oozie 5.2.1	<input checked="" type="checkbox"/> Spark 2.4.8	<input type="checkbox"/> HCatalog 2.3.9
<input type="checkbox"/> TensorFlow 2.4.1		

Click on the created workspace to launch the JupyterLab

EMR Studio

Workspace note successfully launched in a new browser tab.

EMR Studio > Workspaces

Studio: studio-2

Workspaces (1) [Info](#)

[Actions](#)

Workspace name	Status	Last modified by	Creation time (UTC+00:00)	Last modified (UTC+00:00)
<input type="radio"/> note	Ready	kavitha.rajendran@bigspark.dev	January 20, 2023, 13:54	January 23, 2023, 11:30

