

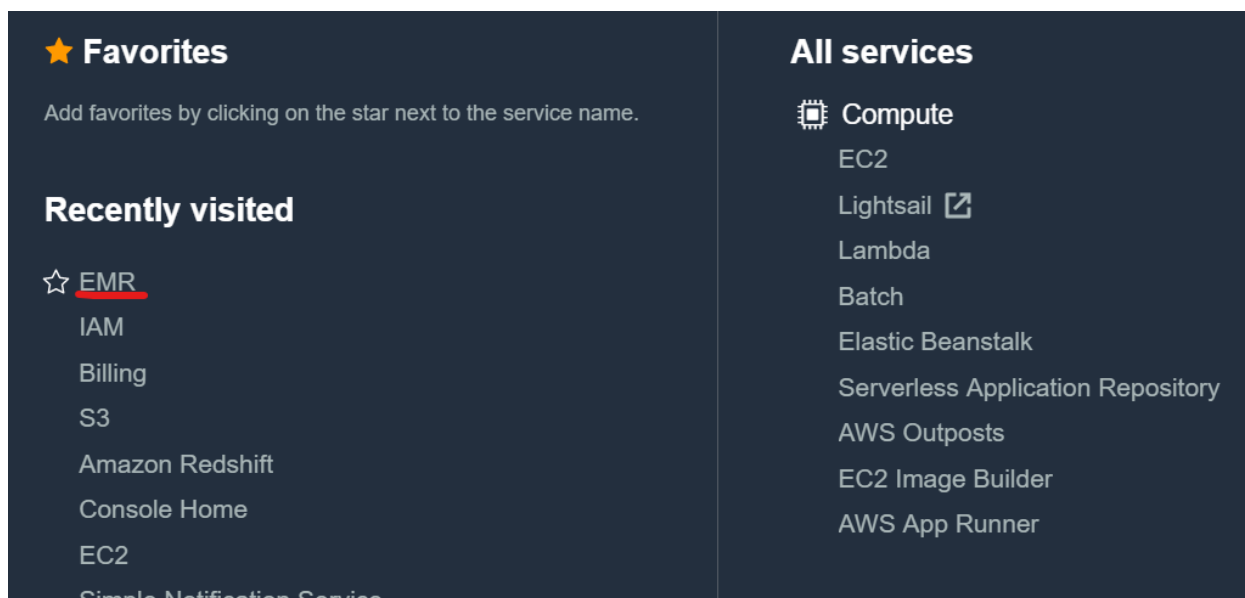
# Amazon EMR for Airflow

This document contains the steps to create an EMR cluster.

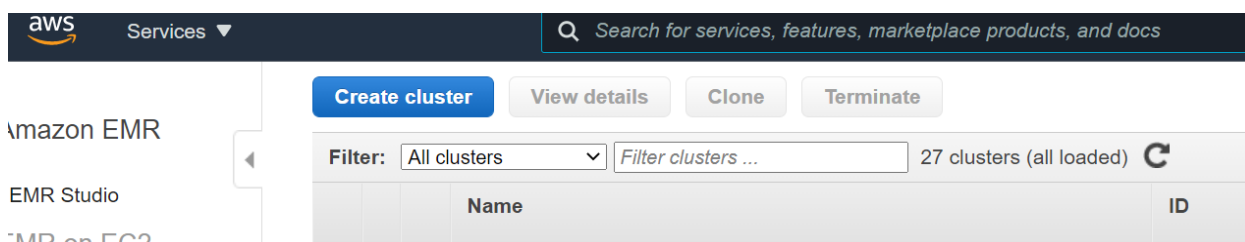
**Note:** Make sure that the location selected for running your account is **US East (N.Virginia)** us-east-1, else the EMR cluster will not launch.

**Prerequisites for EMR setup** – EC2 key pair already set up from the previous EC2 instance setup.

1. Click on the **Services** at the top of the AWS console and then click on the **EMR** service.



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



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

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

### General Configuration

Cluster name

☒ Logging ⓘ

S3 folder

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

4. It will take you to the following page. In the 'Release' column, you will be choosing the **emr-5.30.1** version for your EMR cluster. You will now configure the software applications that you need for your EMR cluster. For Airflow, you need to install the following services -

- ☐ **Hadoop**
- ☐ **Hive**
- ☐ **Spark**
- ☐ **Sqoop**

### Software Configuration

Release  ⓘ

<input checked="" type="checkbox"/> Hadoop 2.8.5	<input type="checkbox"/> Zeppelin 0.8.2	<input type="checkbox"/> Livy 0.7.0
<input type="checkbox"/> JupyterHub 1.1.0	<input type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.10.0
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 1.4.13	<input type="checkbox"/> Pig 0.17.0
<input checked="" type="checkbox"/> Hive 2.3.6	<input type="checkbox"/> Presto 0.232	<input type="checkbox"/> ZooKeeper 3.4.14
<input type="checkbox"/> MXNet 1.5.1	<input checked="" type="checkbox"/> Sqoop 1.4.7	<input type="checkbox"/> Mahout 0.13.0
<input type="checkbox"/> Hue 4.6.0	<input type="checkbox"/> Phoenix 4.14.3	<input type="checkbox"/> Oozie 5.2.0
<input checked="" type="checkbox"/> Spark 2.4.5	<input type="checkbox"/> HCatalog 2.3.6	<input type="checkbox"/> TensorFlow 1.14.0

5. 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. The **Task** and **Core Node count should be 0**. 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](#)

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

Node type	Instance type	Instance count	Purchasing option
<b>Master</b> Master Instance Group	<b>m4.xlarge</b> 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Edit 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: 32 GiB Add configuration settings	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price

- 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

☐ m2.4xlarge
 8
 68.4
 1690 SSD

☐ m3.xlarge
 4
 15
 80 SSD

☐ m3.2xlarge
 8
 30
 160 SSD

☐ m4.large
 2
 8
 EBS only

☒ m4.xlarge
 4
 16
 EBS only

☐ m4.2xlarge
 8
 32
 EBS only

☐ m4.4xlarge
 16
 64
 EBS only

☐ m4.10xlarge
 40
 160
 EBS only

☐ m4.16xlarge
 64
 256
 EBS only

☐ m5.xlarge
 4
 16
 EBS only

☐ m5.2xlarge
 8
 32
 EBS only

☐ m5.4xlarge
 16
 64
 EBS only

Cancel

Save

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

m4.xlarge

4 vCore, 16 GiB memory, EBS only storage

EBS Storage: 64 GiB

Edit configuration settings

- 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

- 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
upgrad\_emr

☒ Logging ⓘ

S3 folder
s3://aws-logs-367134191692-us-east-1/elasticmaprec

☐ Log encryption ⓘ

☒ Debugging ⓘ

☐ Termination protection ⓘ

11. 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	<span>Default: sg-00fcc219431b5c0a6 (ElasticMapReduce-</span>	No security groups selected ✎
Core & Task	<span>Default: sg-0a724c5cb4e439160 (ElasticMapReduce-</span>	No security groups selected ✎

[Create a security group](#)

Cancel Previous Create cluster

12. Thereafter, the cluster will start setting up.

Cluster: upgrad\_emr Starting Configuring cluster software

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

### Summary

ID: j-24L8PEMY6A07  
 Creation date: 2021-07-17 20:05 (UTC+5:30)  
 Elapsed time: 7 minutes  
 After last step completes: Cluster waits  
 Termination protection: Off [Change](#)  
 Tags: -- [View All](#) / [Edit](#)  
 Master public DNS: ec2-34-207-142-34.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

### Configuration details

Release label: emr-5.30.1  
 Hadoop distribution: Amazon 2.8.5  
 Applications: Hive 2.3.6, Hue 4.6.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2  
 Log URI: s3://aws-logs-367134191692-us-east-1/elasticmapreduce/ [View](#)  
 EMRFS consistent view: Disabled  
 Custom AMI ID: --

### Application user interfaces

Persistent user interfaces [View](#): --

On-cluster user interfaces [View](#): Not Enabled [Enable an SSH Connection](#)

### Network and hardware

Availability zone: us-east-1a  
 Subnet ID: [subnet-0085edc222d4dad91](#)  
 Master: Bootstrapping 1 m4.xlarge  
 Core: --  
 Task: --  
 Cluster scaling: Not enabled

### Security and access

Key name: RHEL  
 EC2 instance profile: EMR\_EC2\_DefaultRole  
 EMR role: EMR\_DefaultRole  
 Visible to all users: All [Change](#)  
 Security groups for Master: [sg-00fcc219431b5c0a6](#) (ElasticMapReduce-master)  
 Security groups for Core & Task: [sg-0a724c5cb4e439160](#) (ElasticMapReduce-slave)

## Steps to follow before performing SSH to the master node:

1. Under the cluster information page, click on the **security groups of the master node**.

### Security and access

Key name: RHEL

EC2 instance profile: EMR\_EC2\_DefaultRole

EMR role: EMR\_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-00fcc219431b5c0a6](#) [🔗](#) (ElasticMapReduce-master)

Security groups for Core & Task: [sg-0a724c5cb4e439160](#) [🔗](#) (ElasticMapReduce-slave)

2. Click on the 'security group' and you will land on a similar page. Here, click on the security group of the **Elastic MapReduce-master node** as highlighted in the image below.

Security Groups (2) <a href="#">Info</a>				
<input type="text" value="Filter security groups"/>				
<div> <div>search: sg-00fcc219431b5c0a6 X</div> <div>Clear filters</div> </div>				
<input type="checkbox"/>	Name ▼	Security group ID ▼	Security group name ▼	VPC ID ▼
<input type="checkbox"/>	-	<a href="#">sg-00fcc219431b5c0a6</a>	ElasticMapReduce-master	<a href="#">vpc-069954fb4011801ca</a> <a href="#">🔗</a>
<input type="checkbox"/>	-	<a href="#">sg-0a724c5cb4e439160</a>	ElasticMapReduce-slave	<a href="#">vpc-069954fb4011801ca</a> <a href="#">🔗</a>

- Clicking on the security group will land you on the corresponding security information page. Click on 'Edit inbound rules' to add a new rule.

sg-00fcc219431b5c0a6 - ElasticMapReduce-master

Actions

Details

Security group name ElasticMapReduce-master	Security group ID sg-00fcc219431b5c0a6	Description Master group for Elastic MapReduce created on 2021-03-16T22:01:30.781Z	VPC ID vpc-069954fb4011801ca
Owner 367134191692	Inbound rules count 21 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules

Outbound rules

Tags

Inbound rules (21)

Filter security group rules

Manage tags

Edit inbound rules

<

1

>

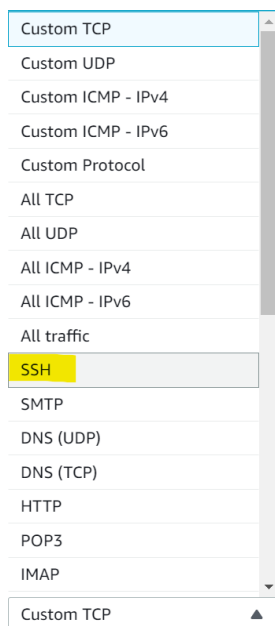
- This will take you to a list of existing rules, where you have the option to delete the existing rules [clicking on delete on the extreme right-hand side] or add a new rule by clicking on **Add rule** towards the bottom of all the rules. Clicking on the 'add rule' will add a new row as shown in the figure below.

sgr-049e72c7217cd27d1	All TCP	TCP	0 - 65535	Custom	72.21.198.64/29		Delete
sgr-0aa1d8567077864f9	All ICMP - IPv4	ICMP	All	Custom	sg-00fcc219431b5c0a6		Delete
sgr-0522ea2084a15dc38	Custom TCP	TCP	8443	Custom	sg-0a724c5cb4e439160		Delete
sgr-0e4552cd2b814d5be	Custom TCP	TCP	8443	Custom	207.171.167.26/32		Delete
-	Custom TCP	TCP	0	Custom	72.21.217.0/24		Delete

Add rule



Under the type field of the newly added row, select **SSH**.

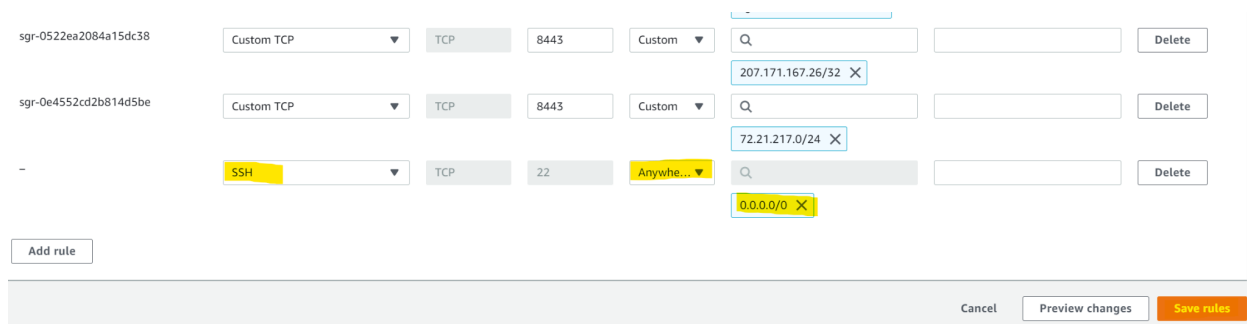


The **Source** will be '**Anywhere-IPv4**' for this rule. For frequent testing, you can avoid using My IP address and choose 'Anywhere-IPv4' while adding rules in the Security Group.

After adding this rule, you need to add another rule for port **8082**. Click on Add rule again and for this rule, select **Custom TCP** as the type, type **8082** under port, and then for Source select **My IP**. Your rule for port 8082 should look like this.



After you have set these rules, do not forget to click on **Save rules** at the bottom of the window.



You can now set up the **YARN parameter configurations** and then log in to your instance.

**Note: Avoid choosing My IP for the EMR cluster for SSH.**

### What happens if you choose My IP?

If you Choose **My IP** under the source field, this will automatically load your IP address in the adjacent blank column.

My IP ▼

117.216.241.70/32 ✕

On adding the rule and choosing the appropriate options as shown below, click on 'Save rule' [at the bottom of the screen] to successfully add the rule.

All UDP ▼	UDP	0 - 65535	Custom ▼	<input type="text"/> <div>sg-08b9414f92bc12611 ✕</div>	<input type="text"/>	Delete
All ICMP - IPv4 ▼	ICMP	All	Custom ▼	<input type="text"/> <div>sg-024a40edec2182c0d ✕</div>	<input type="text"/>	Delete
All ICMP - IPv4 ▼	ICMP	All	Custom ▼	<input type="text"/> <div>sg-08b9414f92bc12611 ✕</div>	<input type="text"/>	Delete
SSH ▼	TCP	22	My IP ▼	<input type="text"/> <div>117.216.241.70/32 ✕</div>	<input type="text"/>	Delete

Add rule

Now, under the list of inbound rules appearing under the master node security group, you can see the newly added rule.

Inbound rules					Edit inbound rules
Type	Protocol	Port range	Source	Description - optional	
All TCP	TCP	0 - 65535	sg-024a40edec2182c0d (ElasticMapReduce-master)	-	
All TCP	TCP	0 - 65535	sg-08b9414f92bc12611 (ElasticMapReduce-slave)	-	
SSH	TCP	22	117.216.241.70/32	-	
Custom TCP	TCP	8443	207.171.167.25/32	-	
Custom TCP	TCP	8443	54.240.217.8/29	-	
Custom TCP	TCP	8443	72.21.196.64/29	-	
Custom TCP	TCP	8443	72.21.198.64/29	-	
Custom TCP	TCP	8443	54.240.217.16/29	-	

On adding this rule, it enables you to perform an SSH to the master node of the cluster.

**But each time you are cloning** the cluster or connecting to the cluster after restarting the laptop, the first thing to do is to edit the security groups and update your current IP address.

All TCP

TCP

0 - 65535

Custom

Q

Delete

All TCP

TCP

0 - 65535

Custom

Q

Delete

SSH

TCP

22

My IP

Q

Delete

As shown in the figure, you need to edit the rule corresponding to the **SSH type** rule and change the option from custom to **My IP** and click on **Save rules**.

This will ensure that you are performing a successful SSH or successfully cloning to a new cluster.

### Important – General Practice:

To avoid having to clone a cluster or restart the laptop every time, a common practice followed while studying or internal testing is to select the option '**anywhere**' for **SSH** instead of custom or **My IP**. In the real-world development environment, this should be avoided because it makes the cluster vulnerable and any IP address can access the cluster.