

AWS EMR Instructions

These instructions walk you through the process of creating an initial Amazon EMR (Elastic Map Reduce) cluster using **Quick Create** options in the AWS Management Console.

Note, the EMR cluster you set up using these instructions is not meant for a production (secure) environment, and do not cover configuration options in depth. It is meant to help you set up a cluster for class purposes as quickly as possible.

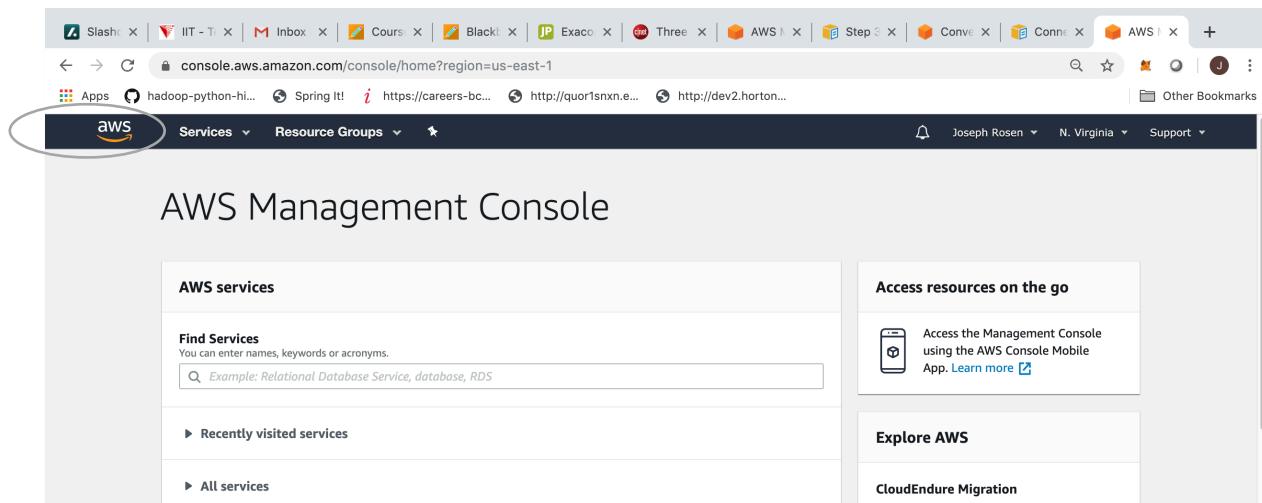
Charges accumulate for cluster you create at the per-second rate for Amazon EMR pricing. The cost will be minimal because the cluster should run for less than a couple of hours after the cluster is provisioned. So it is important that you decommission the cluster as instructed below after you are done with an assignment.

Step 1: Prerequisites

Before you begin setting up your Amazon EMR cluster, make sure that you have completed assignment #1, have an AWS account and understand the basics of working with S3 buckets and associated data objects.

Step 2: AWS Management Console

When you log in to AWS you are presented with the AWS Management Console page. Wherever you are on the site, you can always return to the management console page by clicking on the AWS logo at the top left.



Step 3: Finding Services

We will be making use of several AWS services including

- EC2 – provides computing capability in the form of virtual machines (servers)
- S3 – for object storage
- EMR – Elastic Map Reduce, the Hadoop cluster as a service

When you are on the AWS Management Console page (which we can always get to by clicking the AWS logo), you can find the main page for a service by doing one of the following

1. Type the name of the service whose web page you want to reach into the “Find Services” text box and press Enter/Return
2. If you typed in the name of or used a service recently you might be able to find its name by clicking on “Recently visited services” and then clicking on the name of the desired service
3. If you don’t recall the name of the service, then click on “All Services” to get a list and click on the service of interest.
4. Or you can always click on the word “Services” in the upper left of the management console to get a list of services and also type in the one you are looking for.

So in the following steps when you are requested to find some service, you can do the above.

Step 4: Create an Amazon EC2 Key Pair

You must have an Amazon Elastic Compute Cloud (Amazon EC2) key pair to connect to the nodes in your EMR cluster over a secure channel using the Secure Shell (SSH) protocol. We will understand more about SSH below.

1. Find the EC2 service page
2. In the navigation pane, under **NETWORK & SECURITY**, choose **Key Pairs**.

Note

The navigation pane is on the left side of the Amazon EC2 console. If you do not see the pane, it might be minimized; choose the arrow to expand the pane.

The screenshot shows the AWS EC2 Management Console. The left sidebar has a tree view with 'INSTANCES' expanded, showing 'Instances', 'Launch Templates', 'Spot Requests', 'Reserved Instances', 'Dedicated Hosts', 'Scheduled Instances', and 'Capacity Reservations'. Below this is another tree view with 'KEY PAIRS' expanded, showing 'Key Pairs'. The main content area is titled 'Resources' and displays the following statistics for the US East (N. Virginia) region:

	Value
0 Running Instances	0 Elastic IPs
0 Dedicated Hosts	0 Snapshots
0 Volumes	0 Load Balancers
2 Key Pairs	1 Security Groups
0 Placement Groups	

Below the statistics, there's a link to 'Learn more about the latest in AWS Compute from AWS re:Invent by viewing the EC2 Videos.'

The 'Create Instance' section is titled 'To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.' It includes a 'Launch Instance' button and a note: 'Note: Your instances will launch in the US East (N. Virginia) region.'

The 'Service Health' section shows 'Service Status' for 'US East (N. Virginia)' with a green checkmark and 'Availability Zone Status' for 'us-east-1a:' and 'us-east-1b:' both with green checkmarks.

The bottom of the page includes a 'Feedback' link, language selection ('English (US)'), and a copyright notice ('© 2008 - 2019, Amazon')).

3. Choose Create Key Pair.

The screenshot shows the AWS Lambda console with the search bar set to 'Search for services, features, marketplace products, and docs'. The top right shows user information ('jrosen') and region ('N. Virginia'). The main area is titled 'Key pairs' and contains a table with columns: Name, Fingerprint, and ID. A message at the bottom says 'No key pairs to display'. On the right side, there are 'Actions' and a 'Create key pair' button, which is circled in blue.

Then you should see the following form:

The screenshot shows the 'Create key pair' dialog box. At the top, it says 'Create key pair'. Below that is a 'Key pair' section with a detailed description: '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.' There are three tabs: 'Name', 'File format', and 'Tags (Optional)'. The 'Name' tab is active, showing a text input field with placeholder 'Enter key pair name' and a note: 'The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.' The 'File format' tab has two options: 'pem' (selected) and 'ppk'. The 'Tags (Optional)' tab shows a note: 'No tags associated with the resource.' At the bottom are 'Cancel' and 'Create key pair' buttons.

- For the key pair name, enter a name for the new key pair (something like emr-key-pair), and then choose **Create key pair**. Leave other options as they are, unless you are using Putty, then check ‘ppk.’

Create key pair

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.

File format
 pem
For use with OpenSSH
 ppk
For use with PuTTY

Tags (Optional)
No tags associated with the resource.

You can add 50 more tags.

Create key pair

- The private key file is automatically downloaded by your browser. The base file name is the name you specified as the name of your key pair, and the file name extension is .pem or “.cer” (or .ppk). Save the private key file in a safe place.

NOTE: Sometime AWS downloads the key file with a file suffix of “.cer” instead of “.pem” file, even if you request a “.pem” file. This is perfectly ok; the content of the file is the same. In this case, everywhere in the course, where you see instructions or examples refer to files with the “.pem” suffix, just substitute the “.cer” suffix and all will work as intended. Or, if you like you can just rename the file to use the “.pem” suffix, but this is not necessary. Remember, if you can’t find a “.pem” file on your computer check for a “.cer” file.

In most cases on the MAC the file will download to the directory
 /Users/<username>/Downloads

And on the PC the file will most likely download to

/c/Users/<username>/Downloads.

Note, the way I have written the path to the file is formatted for when using the git bash utility.

Important

This is the only chance for you to save the private key file. You'll need to provide the name of your key pair when you launch an instance and the corresponding private key each time you connect to the instance. But if you can create another by repeating the above steps.

6. So find the directory into which your .pem file has been downloaded and either keep it there or move it to another directory of your choice. You will need to know the path to this file.
7. Using the “terminal” program on the MAC or the “bash” utility on the PC execute the following command to set the permissions of your private key file so that only you can read it. Note, use the appropriate path and file name for your situation.

```
chmod 400 <path-to-file>/emr-key-pair.pem
```

Note, depending on the operating system used for your personal computer, the above may not work. Things might still be ok, but if not reach out to me.

Step 5: Launch Your Initial Amazon EMR Cluster

In this step, you launch your initial cluster by using **Quick Options** in the Amazon EMR console and leaving most options to their default values.

To launch the sample Amazon EMR cluster

1. Find the EMR console page
2. Choose **Create cluster**.

The screenshot shows the Amazon EMR console interface. On the left, a sidebar lists navigation options: Amazon EMR, EMR on EC2 (Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, Events), EMR on EKS (Virtual clusters), Help, and What's new. The main content area is titled "Welcome to Amazon Elastic MapReduce". It includes a brief description of the service, a message stating "You do not appear to have any clusters. Create one now:", and a prominent blue "Create cluster" button, which is circled in blue in the screenshot. Below this, there's a section titled "How Elastic MapReduce Works" with three icons: "Upload" (cloud with an orange arrow pointing up), "Create" (a network-like structure with a gear and an orange arrow), and "Monitor" (a monitor displaying a line graph with an orange arrow pointing down). Each icon has a corresponding text description: "Upload your data and processing application to S3.", "Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.", and "Monitor the health and progress of your cluster. Retrieve the output in S3.". At the bottom of each description is a "Learn more" link.

3. On the **Create Cluster - Quick Options** page, accept the default values except for the following fields (see figure on next page):

- Enter a **Cluster name** that helps you identify the cluster, for example, *My First EMR Cluster*.
- Under **Hardware configuration** choose:
 - The Instance type as: m4.large
 - The Number of instances as: 2
- Under **Security and access**, choose the **EC2 key pair** that you created in Create an Amazon EC2 Key Pair
-

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

General Configuration

Cluster name [Edit](#)

S3 folder [Edit](#)

Logging [Edit](#)

Launch mode Cluster Step execution [Edit](#)

Software configuration

Release [Edit](#)

Applications Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.5, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2 [Edit](#)

HBase: HBase 1.4.9 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.5, Hue 4.4.0, Phoenix 4.14.1, and ZooKeeper 3.4.14 [Edit](#)

Presto: Presto 0.220 with Hadoop 2.8.5 HDFS and Hive 2.3.5 Metastore [Edit](#)

Spark: Spark 2.4.3 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.1 [Edit](#)

Use AWS Glue Data Catalog for table metadata [Edit](#)

Hardware configuration

Instance type [Edit](#) The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. [Learn more](#)

Number of instances (1 master and 1 core nodes) [Edit](#)

Security and access

EC2 key pair [Edit](#) Learn how to create an EC2 key pair.

Permissions Default Custom [Edit](#)

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 [Edit](#)

EC2 instance profile [Edit](#)

[Cancel](#) [Create cluster](#)

Feedback English (US) © 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use Show All

4. Choose **Create cluster**.

Note your cluster is ready for use when, instead of “Starting” it says “Waiting Cluster ready after last step completed.” This could sometimes take 10+ minutes, so don’t worry.

The cluster status page with the cluster **Summary** appears (see below). You can use this page to monitor the progress of cluster creation and view details about cluster status. As cluster creation tasks finish, items on the status page update. You may need to choose the refresh icon (circular arrow) on the right or refresh your browser to receive updates.

The screenshot shows the AWS EMR Cluster Details page for a cluster named "my-first-emr-cluster". The cluster is currently in the "Starting" state. The page includes tabs for Summary, Application history, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. The Summary tab is active, displaying details such as ID, Creation date, Elapsed time, Auto-terminate setting, and termination protection. The Hardware tab is also highlighted with a blue oval. The Configurations tab shows configuration details like Release label, Hadoop distribution, Applications, Log URI, and EMRFS consistent view. The Security and access section shows Key name, EC2 instance profile, EMR role, and security groups. At the bottom, there are links for Feedback, English (US), Privacy Policy, Terms of Use, and a download link for emr-key-pair.pem.

Under **Network and hardware**, find the **Master** and **Core** instance status. The status goes from **Provisioning** to **Bootstrapping** to **Waiting** during the cluster creation process. For more information, see [Understanding the Cluster Lifecycle](#).

As soon as you see the links for **Security groups for Master** and **Security Groups for Core & Task (see below)**, you can move on to the next task, but you may want to wait until the cluster starts successfully and is in the **Waiting** state. The links are blue colored identifiers starting with “sg-” in the Security and Access Area of the page.

Under **Security and access** choose the **Security groups for Master** link

The screenshot shows the AWS EMR Cluster Summary page for a cluster named 'j-2MU1O79R5H57Q'. The 'Network and hardware' section is highlighted with a blue oval, showing details like Availability zone: us-east-1e, Subnet ID: subnet-9605f9aa, Master: Bootstrapping, Core: Provisioning, and Task: --. The 'Security and access' section is also highlighted with a blue oval, showing Key name: emr-key-pair, EC2 instance profile: EMR_EC2_DefaultRole, EMR role: EMR_DefaultRole, and security groups: sg-058def5512661472 (Master: (ElasticMapReduce-master)) and sg-004edcfa8cde8bedd (Core & Task: (ElasticMapReduce-slave)).

For more information about reading the cluster summary, see [View Cluster Status and Details](#).

Allow SSH Connections to the Cluster from Your Client

Security groups act as virtual firewalls to control inbound and outbound traffic to your cluster. When you create your first cluster, Amazon EMR creates the default Amazon EMR-managed security group associated with the master instance, **ElasticMapReduce-master**, and the security group associated with core and task nodes, **ElasticMapReduce-slave**. To reach the security groups of interest just click on the blue link associated with the Security group for Master entry and you should then see something like the following.

The screenshot shows the AWS Security Groups page with two entries listed:

Name	Security group ID	Security group name	VPC ID	Description
-	sg-01d0f8713a8ea2af5	ElasticMapReduce-slave	vpc-9f4ceae2	Slave group for Elastic ...
-	sg-0d8c6986dae379286	ElasticMapReduce-mas...	vpc-9f4ceae2	Master group for Elasti...

For more information about security groups, see [Control Network Traffic with Security Groups](#) and [Security Groups for Your VPC](#) in the *Amazon VPC User Guide*.

1. Choose **ElasticMapReduce-master** from the list. Select the ElasticMapReduce-master by clicking on its row.
2. On the bottom of the screen will appear tabs for this security group. Select the “Inbound rules” tab.

The screenshot shows the AWS Security Groups list and a detailed view for the security group `sg-0d8c6986dae379286 - ElasticMapReduce-master`.

Security Groups (1/2) Info

Search bar: `search: sg-0d8c6986dae379286`

Table columns: Name, Security group ID, Security group name, VPC ID, Description

Details for `sg-0d8c6986dae379286 - ElasticMapReduce-master`:

- Security group name: `ElasticMapReduce-master`
- Security group ID: `sg-0d8c6986dae379286`
- Description: `Master group for Elastic MapReduce created on 2021-01-27T17:54:53.770Z`
- VPC ID: `vpc-9f4ceae2`

The "Inbound rules" tab is selected.

When you see the “Edit inbound rules” button. Click on it.

The screenshot shows the Inbound rules configuration page for the security group `sg-0d8c6986dae379286 - ElasticMapReduce-master`.

Table columns: Type, Protocol, Port range, Source, Description - optional

Rules listed:

- Type: All TCP, Protocol: TCP, Port range: 0 - 65535, Source: `sg-01d0f8713a8ea2af5 (ElasticMapReduce-slave)`, Description: -
- Type: All TCP, Protocol: TCP, Port range: 0 - 65535, Source: `sg-0d8c6986dae379286 (ElasticMapReduce-master)`, Description: -

An "Edit inbound rules" button is circled in blue.

A new pane will appear allowing you to modify access rules. Scroll down to the bottom of the list where you will see the “Add rule” button. Select it.

The screenshot shows the AWS CloudFormation Access Rules interface. It displays three existing security group rules:

- Rule 1:** All UDP (Port range: 0 - 65535) to sg-0d8c6986dae37928 (Version 6).
- Rule 2:** All ICMP - IPv4 (Port range: All) to sg-01d0f8713a8ea2af (Version 5).
- Rule 3:** All ICMP - IPv4 (Port range: All) to sg-0d8c6986dae37928 (Version 6).

An "Add rule" button is located at the bottom left, circled in blue.

A line for you to enter a new access rule will appear:

The screenshot shows the AWS CloudFormation Access Rules interface with a new rule being added. The "Custom TCP" dropdown and the "Custom" dropdown next to it are circled in blue.

1. Select the field with label “Custom TCP” which pops up a list of options, select “SSH”. When you do the next field to its left will display the value “TCP” and the next field to the left of that will show “22”.
2. Now select the next field showing the value “Custom” which pops up a list from which you should select “My IP” which causes your IP to be the only one allowed to access your EMR cluster via SSH (or SCP). Scroll down a bit more, if needed, and click on the “Save rules” button.

The screenshot shows the AWS CloudFormation Access Rules interface with a new rule being added. The "SSH" dropdown is selected. A note at the bottom states: “⚠️ NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.”

The "Save rules" button is circled in blue at the bottom right.

Note, once you have set up this rule, in most cases when you create a new cluster, it will use the same security group, so you likely will not need to set up this rule again. But it always is good to check.

Step 6: Connect to the Master Node Using SSH

Secure Shell (SSH) is a network protocol you can use to create a secure connection to a remote computer. After you make a connection, the terminal on your local computer behaves as if it is running on the remote computer. Commands you issue locally run on the remote computer, and the command output from the remote computer appears in your terminal window.

When you use SSH with AWS, you are connecting to an EC2 instance, which is a virtual server running in the cloud. When working with Amazon EMR, the most common use of SSH is to connect to the EC2 instance that is acting as the master node of the cluster.

Using SSH to connect to the master node gives you the ability to monitor and interact with the cluster. You can issue Linux commands on the master node, run applications such as Hive and Pig interactively, browse directories, read log files, and so on.

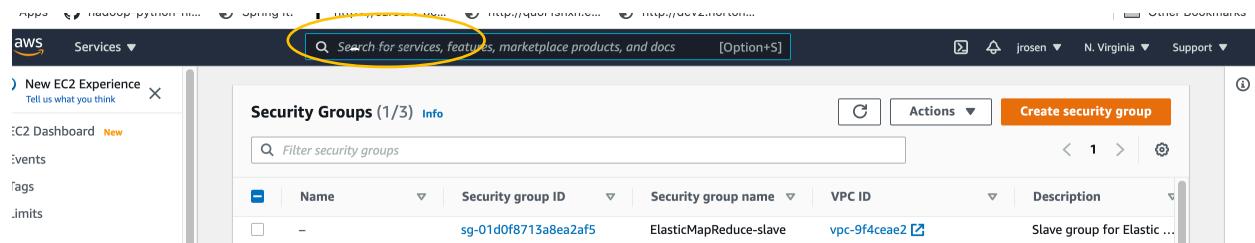
To connect to the master node using SSH, you need the public DNS name of the master node. In addition, the security group associated with the master node must have an inbound rule that allows SSH (TCP port 22) traffic from a source that includes the client where the SSH connection originates (something you did above).

Retrieve the Public DNS Name of the Master Node

You can retrieve the master public DNS name using the Amazon EMR console and the AWS CLI.

To retrieve the public DNS name of the master node using the Amazon EMR console

1. Find the EMR service page by typing EMR into the “Search for services.” Box and selecting EMR



2. On the Cluster List page, select the link for your cluster.

Create cluster	View details	Clone	Terminate		
Filter: All clusters		Filter clusters ...		1 cluster (all loaded) C	
Name	ID	Status	Creation time (UTC-6)	Elapsed time	Normalized instance hours
My cluster	j-1323ICIGXD599	Waiting Cluster ready	2021-01-27 14:47 (UTC-6)	39 minutes	0

3. Note the **Master public DNS** value that appears at the top of the **Cluster Details** page.

Clone **Terminate** **AWS CLI export**

Cluster: My cluster Waiting Cluster ready after last step completed.

Summary **Application user interfaces** **Monitoring** **Hardware** **Configurations** **Events**

Summary

ID: j-1323ICIGXD599
Creation date: 2021-01-27 14:47 (UTC-6)
Elapsed time: 40 minutes
After last step completes: Cluster waits
Termination protection: Off [Change](#)
Tags: -- [View All / Edit](#)
Master public DNS: [ec2-54-159-52-97.compute-1.amazonaws.com](#) 
[Connect to the Master Node Using SSH](#)

Configuration details **Application usage**
Release label: emr-5.32.0 **Persistent use**

To connect to the Master Node Using SSH and an Amazon EC2 Private Key

Open a terminal window the MAC or use the bash utility on the PC.

1. To establish a connection to the master node, type the following command.
 - a. Replace `ec2-###-##-##-##.compute-1.amazonaws.com` with the master public DNS name of your cluster
 - b. Replace `/<path-to-file>/mykeypair.pem` with the path (on your PC/Mac) and file name of your .pem file.

For MACOS or Linux, something like:

```
ssh -i /path/to/emr-key-pair.pem hadoop@ ec2-###-##-##-##.compute-1.amazonaws.com
```

For Windows, something like;

```
ssh -i c:/path/to/emr-key-pair.pem hadoop@ ec2-###-##-##-##.compute-1.amazonaws.com
```

Important

You must use the login name hadoop when you connect to the Amazon EMR master node; otherwise, you may see an error similar to Server refused our key.

- When you enter this properly you should see

```
MacBook-Pro-3:~ nachdaph$ ssh -i /Users/nachdaph/csp55-spring-2021/keys/emr-key-pair.pem hadoop@ec2-54-159-52-97.compute-1.amazonaws.com
Warning: Identity file /Users/nachdaph/csp55-spring-2021/keys/emr-key-pair.pem not accessible: No such file or directory.
The authenticity of host 'ec2-54-159-52-97.compute-1.amazonaws.com (54.159.52.97)' can't be established.
ECDSA key fingerprint is SHA256:jmkTz2XSI/dwExwUy4M58vxbw4S0wfsxRWp+qyOGZEM.
Are you sure you want to continue connecting (yes/no/[fingerprint])? █
```

- You might see a warning. The warning states that the authenticity of the host you are connecting to cannot be verified. If needed, type yes to continue.
- When you are done working on the master node (as you might be at the end of an assignment), type the following command to close the SSH connection.

```
exit
```

Step 7: Terminate the Cluster and Delete the Bucket

After you complete your homework assignment or other project work, you may want to terminate your cluster and delete your Amazon S3 bucket to avoid additional charges.

Terminating your cluster terminates the associated Amazon EC2 instances and stops the accrual of Amazon EMR charges. Amazon EMR preserves metadata information about completed clusters for your reference, at no charge, for two months. The console does not provide a way to delete terminated clusters so that they aren't viewable in the console. Terminated clusters are removed from the cluster when the metadata is removed.

To terminate the cluster

1. Find the EMR service
2. Choose **Clusters**, then choose your cluster.

The screenshot shows the AWS EMR Clusters page. On the left, there's a sidebar with options like Clusters, Security configurations, Block public access, VPC subnets, Events, Notebooks, Help, and What's new. The 'Clusters' option is highlighted with a blue oval. The main area displays a table of clusters with the following columns: Name, ID, Status, Creation time (UTC-5), Elapsed time, and Normalized instance hours. A blue oval highlights the 'Status' column header and the first cluster row. The cluster table contains the following data:

Name	ID	Status	Creation time (UTC-5)	Elapsed time	Normalized instance hours
emrtest4	j-4N7YGRW1UIWB	Terminated User request	2019-08-25 17:59 (UTC-5)	1 hour, 25 minutes	32
my-first-emr-cluster	j-2MU1O79R5H57Q	Terminated User request	2019-07-14 14:11 (UTC-5)	1 hour, 11 minutes	16
emrtest3	j-214BQNUSH85FQ	Terminated with errors Instance failure	2019-07-09 20:48 (UTC-5)	1 day, 22 hours	376
emrtest2	j-1GCML6GBWNWU0	Terminated User request	2019-07-09 20:27 (UTC-5)	9 minutes	0
emrtest1	j-3ODXKFUN674MI	Terminated User request	2019-07-07 11:22 (UTC-5)	1 hour, 40 minutes	16

3. Choose Terminate:

The screenshot shows the AWS EMR Cluster details page for the cluster 'emrtest4'. The cluster is listed as 'Terminated' with the reason 'Terminated by user request'. There are tabs for Summary, Application history, Monitoring, Hardware, Configurations, Events, Steps, and Bootstrap actions. The Summary tab is selected. It shows the cluster ID (j-4N7YGRW1UIWB), Release label (emr-5.26.0), Hadoop distribution (Amazon 2.8.5), Applications (Ganglia 3.7.2, Hive 2.3.5, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2), Creation date (2019-08-25 17:59 UTC-5), End date (2019-08-25 19:25 UTC-5), and Elapsed time (1 hour, 25 minutes). A blue oval highlights the 'Terminate' button in the top navigation bar. The sidebar on the left is identical to the one in the previous screenshot.

To delete the cluster logging output bucket

1. Find the S3 service
2. Choose the EMR bucket from the list, so that the whole bucket row is selected.
3. Choose delete bucket, type the name of the bucket, and then click **Confirm**.

For more information about deleting folders and buckets, go to [How Do I Delete an S3 Bucket](#) in the *Amazon Simple Storage Service Getting Started Guide*.