



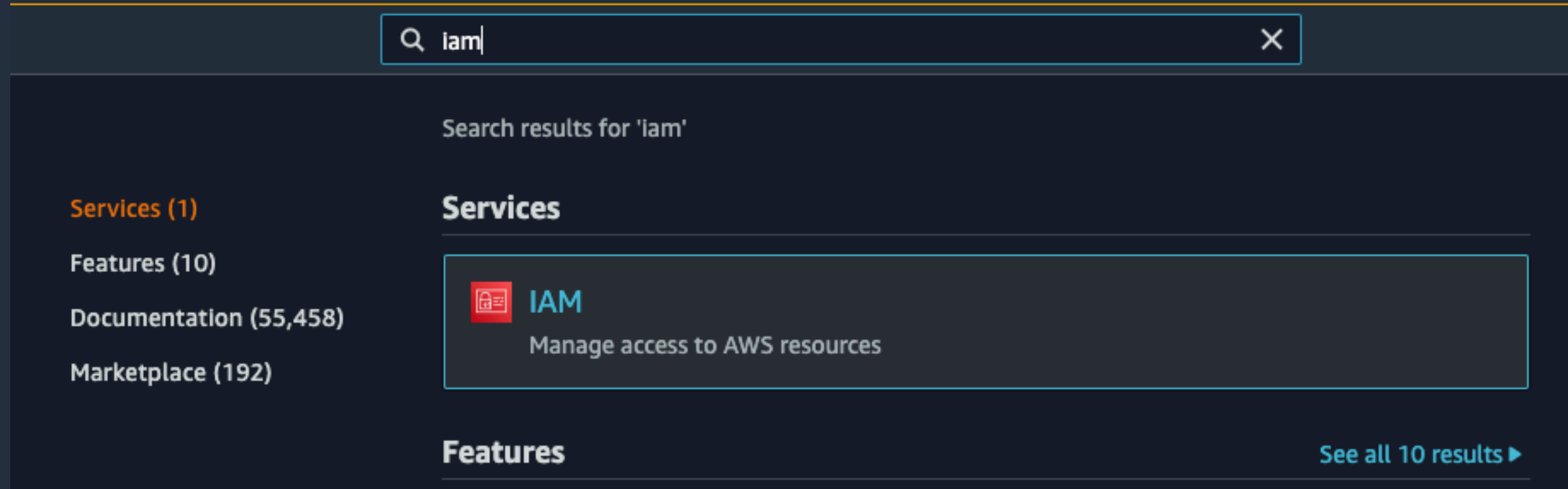
Taming EMR on AWS

Guidance Pack

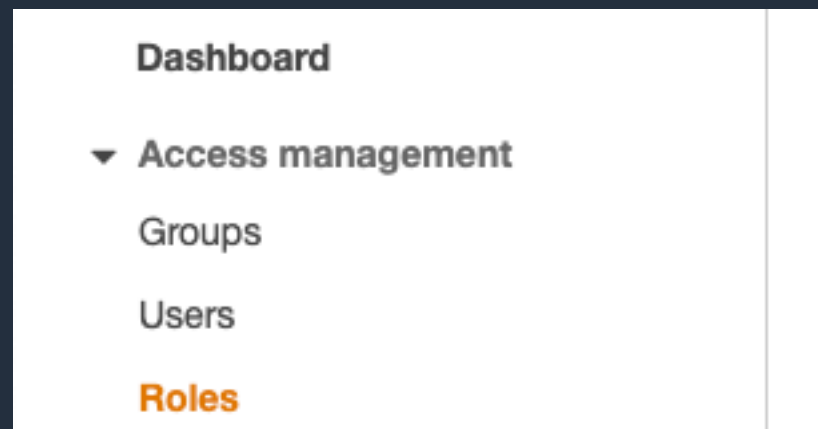


How to launch an EMR cluster

First we need to create a Role, open the IAM console

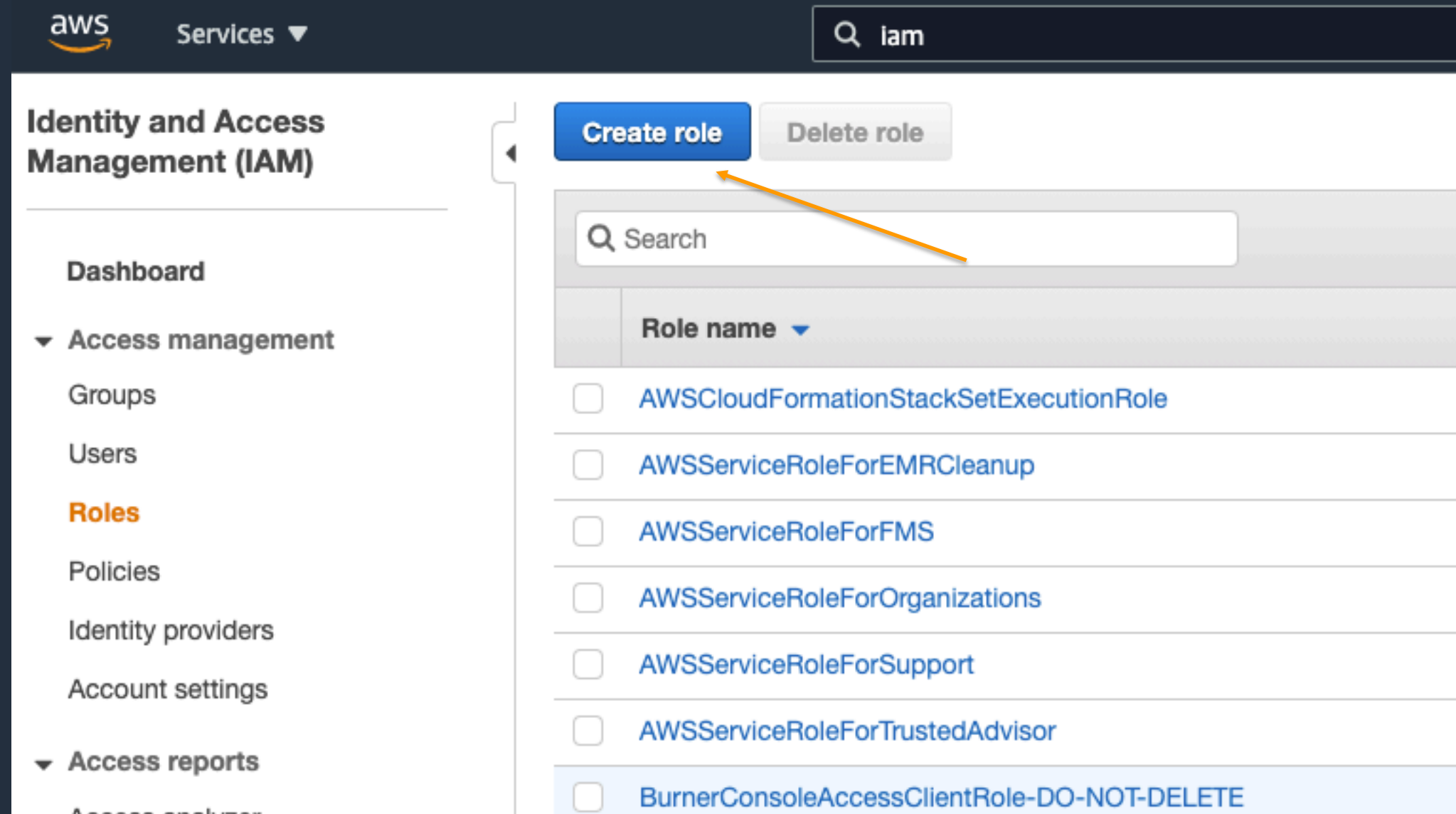


And select the Roles



How to launch an EMR cluster

Click in Create role



The screenshot shows the AWS IAM console interface. At the top, the AWS logo and 'Services' dropdown are visible. A search bar contains the text 'iam'. The left sidebar shows the 'Identity and Access Management (IAM)' section with a list of options: Dashboard, Access management (expanded), Groups, Users, Roles (highlighted in orange), Policies, Identity providers, Account settings, Access reports (expanded), and Access analyzer. The main content area has a 'Create role' button (blue) and a 'Delete role' button (grey). Below these buttons is a search bar labeled 'Search' and a table of roles. An orange arrow points from the 'Create role' button to the search bar. The table has a header 'Role name' and lists several roles with checkboxes:

Role name
<input type="checkbox"/> AWSCloudFormationStackSetExecutionRole
<input type="checkbox"/> AWSServiceRoleForEMRCleanup
<input type="checkbox"/> AWSServiceRoleForFMS
<input type="checkbox"/> AWSServiceRoleForOrganizations
<input type="checkbox"/> AWSServiceRoleForSupport
<input type="checkbox"/> AWSServiceRoleForTrustedAdvisor
<input type="checkbox"/> BurnerConsoleAccessClientRole-DO-NOT-DELETE


How to launch an EMR cluster


Select AW Service as trusted entity, EC2 as common use cases and click in Next


Create role


1234

Select type of trusted entity

**AWS service**
EC2, Lambda and others

**Another AWS account**
Belonging to you or 3rd party

**Web identity**
Cognito or any OpenID provider

**SAML 2.0 federation**
Your corporate directory

Allows AWS services to perform actions on your behalf. [Learn more](#)

Choose a use case

Common use cases

EC2

Allows EC2 instances to call AWS services on your behalf.

Lambda

Allows Lambda functions to call AWS services on your behalf.

* Required

Cancel

Next: Permissions

How to launch an EMR cluster

Search and select AmazonEC2RoleforSSM, after that search and select AmazonElasticMapReduceforEC2Role. Click in Next

Create role


1234

▼ Attach permissions policies


Choose one or more policies to attach to your new role.

Create policy

Filter policies ▼Showing 1 result

	Policy name ▼	Used as
<input checked="" type="checkbox"/>	 AmazonEC2RoleforSSM	Permissions policy (1)

Filter policies ▼

	Policy name ▼
<input checked="" type="checkbox"/>	 AmazonElasticMapReduceforEC2Role

How to launch an EMR cluster

In the Tags session click in Next, In the Review add EMR_EC2_ROLE as Role name.

Check if both policies(AmazonEC2RoleforSSM, AmazonElasticMapReduceforEC2Role) are selected and if Trusted entities is correct, after that click in Create Role

Create role

1 2 3 **4**

Review

Provide the required information below and review this role before you create it.

Role name*



Use alphanumeric and '+=,.-_' characters. Maximum 64 characters.

Role description

Maximum 1000 characters. Use alphanumeric and '+=,.-_' characters.

Trusted entities AWS service: ec2.amazonaws.com

Policies

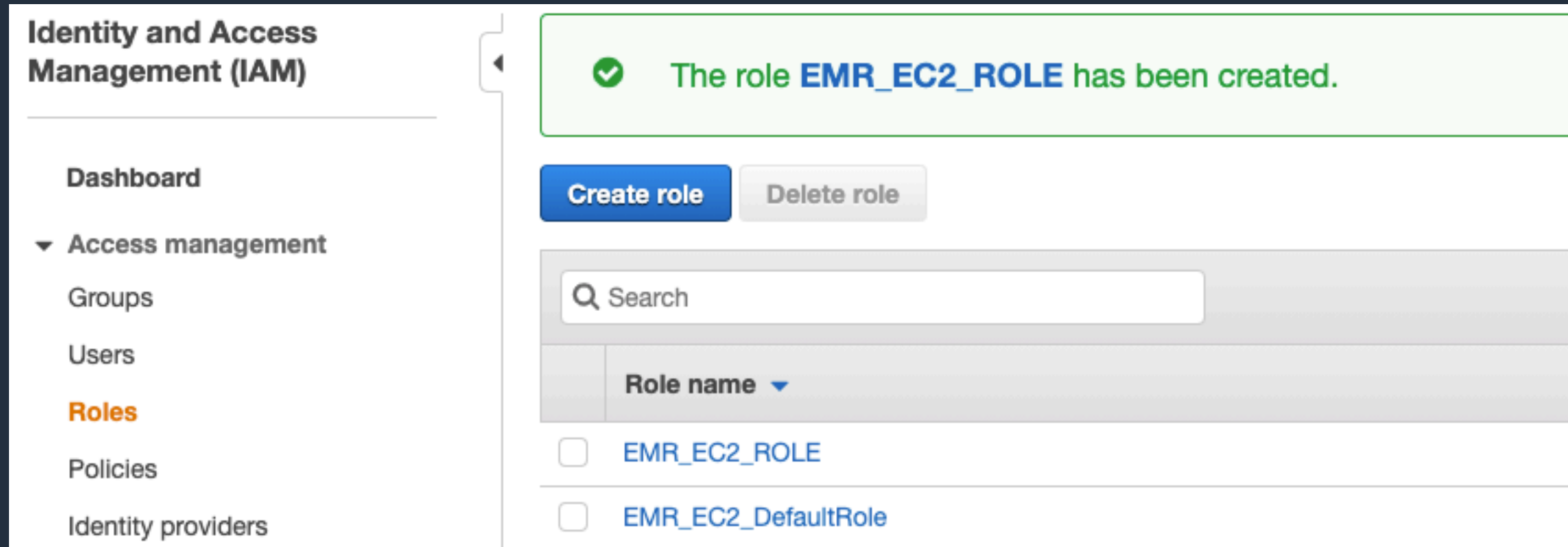
-  [AmazonEC2RoleforSSM](#)
-  [AmazonElasticMapReduceforEC2Role](#)

* Required

[Cancel](#) [Previous](#) [Create role](#)

How to launch an EMR cluster

Check if you can see the role EMR_EC2_ROLE in the Role name List

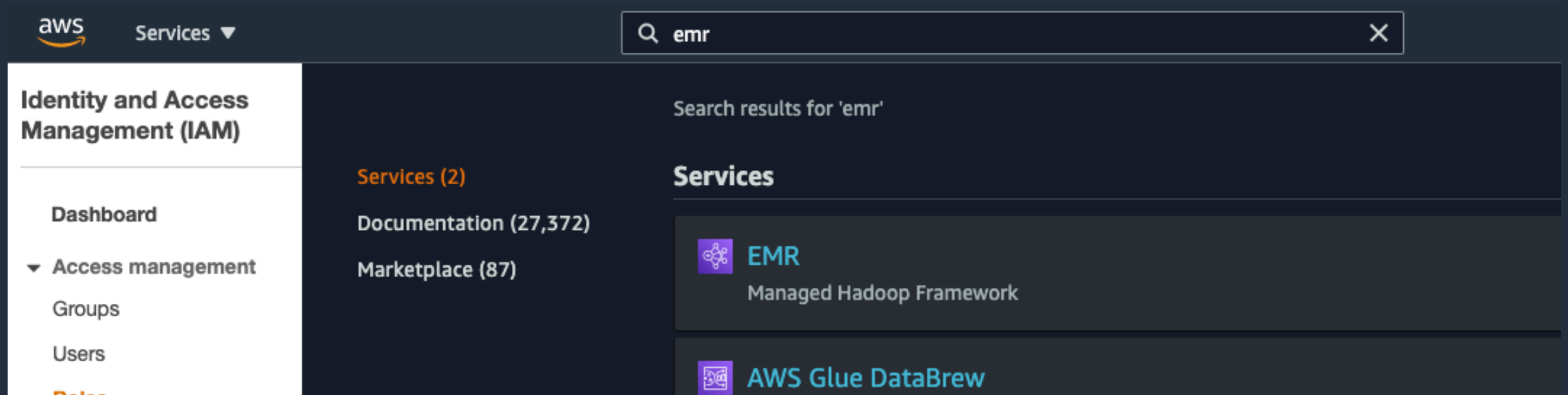


The screenshot displays the AWS Identity and Access Management (IAM) console. On the left, the navigation pane shows the 'Roles' section highlighted. The main content area features a green success message: 'The role EMR_EC2_ROLE has been created.' Below this, there are 'Create role' and 'Delete role' buttons. A search bar is present, and a table lists the roles. The table has a header 'Role name' and two entries: 'EMR_EC2_ROLE' and 'EMR_EC2_DefaultRole', each with an unchecked checkbox to its left.

	Role name
<input type="checkbox"/>	EMR_EC2_ROLE
<input type="checkbox"/>	EMR_EC2_DefaultRole

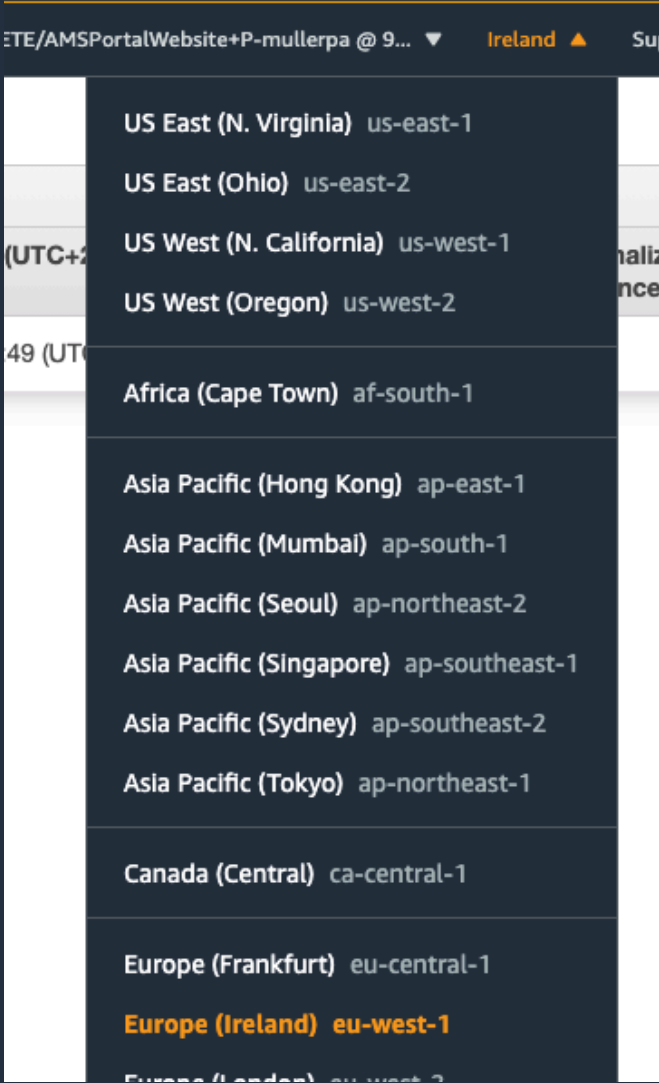
How to launch an EMR cluster

On the AWS console, select search and select EMR from the list of services



How to launch an EMR cluster

Select the Ireland Region



How to launch an EMR cluster

Click on
"Create cluster"

Amazon EMR

EMR on EC2

- Clusters
- Notebooks
 - Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events

EMR on EKS

- Virtual clusters

Help

What's new

Welcome to Amazon Elastic MapReduce


Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

[Create cluster](#)


How Elastic MapReduce Works

Upload




Upload your data and processing application to S3.

Create



Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

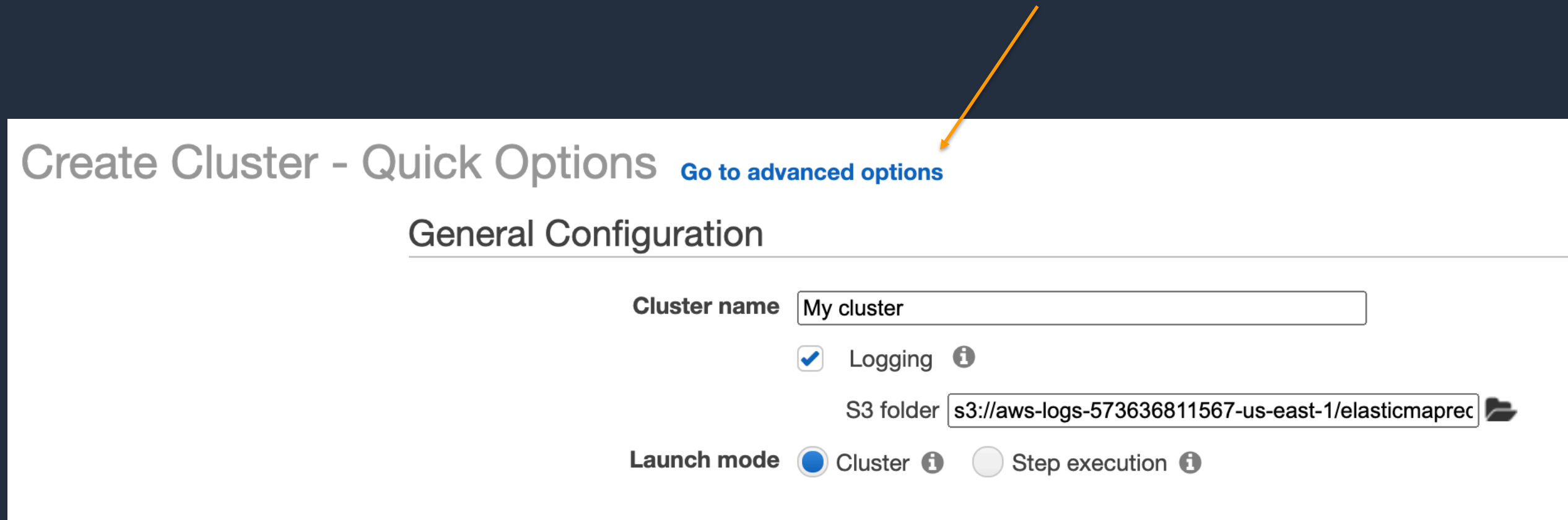
Monitor



Monitor the health and progress of your cluster. Retrieve the output in S3.

How to launch an EMR cluster

On the create cluster page, you'll see an option to "Go to advanced options". Click on that




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

General Configuration

Cluster name

☒ **Logging** ⓘ

S3 folder 

Launch mode ☒ **Cluster** ⓘ ☐ **Step execution** ⓘ

How to launch an EMR cluster

During the Advanced Options, you will be able to select the applications you will be using during this course. Please select EMR release 5.32.0 and applications:

- Hadoop
- Ganglia
- Hive
- Spark
- Tez

Then click “Next”

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 2.8.5	<input type="checkbox"/> Zeppelin 0.8.2	<input type="checkbox"/> Livy 0.7.0
<input type="checkbox"/> JupyterHub 1.1.0	<input checked="" type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.10.0
<input checked="" 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 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

Multiple master nodes (optional)
☐ Use multiple master nodes to improve cluster availability. [Learn more](#)

AWS Glue Data Catalog settings (optional)
☐ Use for Hive table metadata
☐ Use for Spark table metadata

Edit software settings
☒ Enter configuration ☐ Load JSON from S3

`classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]`

How to launch an EMR cluster

In the Hardware
Configuration just
click in Next

[Step 1: Software and Steps](#)

Step 2: Hardware


[Step 3: General Cluster Settings](#)

[Step 4: Security](#)

Hardware Configuration

Specify the networking and hardware configuration for your cluster. Request Spot instances (unused EC2 capacity) to save money.

Cluster Composition

Specify the configuration of the master, core and task nodes as an instances group or instance fleet. This choice applies to all nodes for the lifetime of the cluster. Instance fleets and instance groups cannot coexist in a cluster. [see this topic](#) .

Instance group configuration




Uniform instance groups

Specify a single instance type and purchasing option for each node type.



Instance fleets

Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#) .

How to launch an EMR cluster

On the “General Cluster Settings” page, you can name your cluster and leave the rest of The config as it is. Click “Next”.

Firefox File Edit View History Bookmarks Tools Window Help

EMR - AWS Console

https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster: 80%

aws Services Resource Groups

LakeFormationAdmin @ 8378-... N. Virginia Support

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

General Options

Cluster name

☒ Logging ⓘ
S3 folder

☐ Log encryption ⓘ

☒ Debugging ⓘ

☒ Termination protection ⓘ

Tags ⓘ

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

Additional Options

☐ EMRFS consistent view ⓘ

Custom AMI ID

▶ Bootstrap Actions

[Cancel](#) [Previous](#) [Next](#)

Feedback English (US)

© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

How to launch an EMR cluster

In the Security Tab, in the Permissions session, select Custom

And in the EC2 Instance profile select EMR_EC2_ROLE and click in Create Cluster

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

Security Options

EC2 key pair No key pairs found ⓘ

☒ Cluster visible to all IAM users in account ⓘ

Permissions ⓘ

☐ Default ☒ Custom

Select custom roles to tailor permissions for your cluster.

EMR role Create default role ⓘ

EC2 instance profile EMR_EC2_ROLE ⓘ

Auto Scaling role Proceed without role ⓘ

▶ Security Configuration

▶ EC2 security groups

ⓘ No EC2 key pair has been selected, so you will not be able to SSH to this cluster or connect to HUE (unless you are using a VPN). [Learn how to create an EC2 Key Pair.](#)

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

How to launch an EMR cluster

Your cluster will now be starting, and resources will be provisioning. Please allow 5-10min for your cluster to provision successfully. The status will move from starting, to waiting.

Clone

Terminate

AWS CLI export

Cluster: My cluster Starting Configuring cluster software

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

ID: j-Q2S1VU993QDY

Creation date: 2021-02-11 15:49 (UTC+2)

Elapsed time: 5 minutes

After last step completes: Cluster waits

Termination protection: On [Change](#)

Tags: -- [View All / Edit](#)

Master public DNS:
ec2-54-246-247-154.eu-west-1.compute.amazonaws.com [🔗](#)

[Connect to the Master Node Using SSH](#)

Application user interfaces

Persistent user interfaces [🔗](#): --

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

Configuration details

Release label: emr-5.32.0

Hadoop distribution: Amazon 2.10.1

Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.8.0

Log URI: s3://aws-logs-144364850537-eu-west-1/elasticmapreduce/ [🔗](#)

EMRFS consistent view: Disabled

Custom AMI ID: --

Network and hardware

Availability zone: eu-west-1a

Subnet ID: [subnet-16c78d4c](#) [🔗](#)

Master: Bootstrapping 1 m5.xlarge

Core: Provisioning 2 m5.xlarge

Task: --

Cluster scaling: Not enabled

⚠️ Core - 2: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-verification@amazon.com. We appreciate your patience..

⚠️ Master - 1: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-verification@amazon.com. We appreciate your patience..

How to launch an EMR cluster

To connect to your cluster select the Hardware Tab, and click in the ID for the Master Node

Cluster: My cluster **Running** Running step

Summary Application user interfaces Monitoring **Hardware** Configurations Events Steps Bootstrap actions

Add task instance group

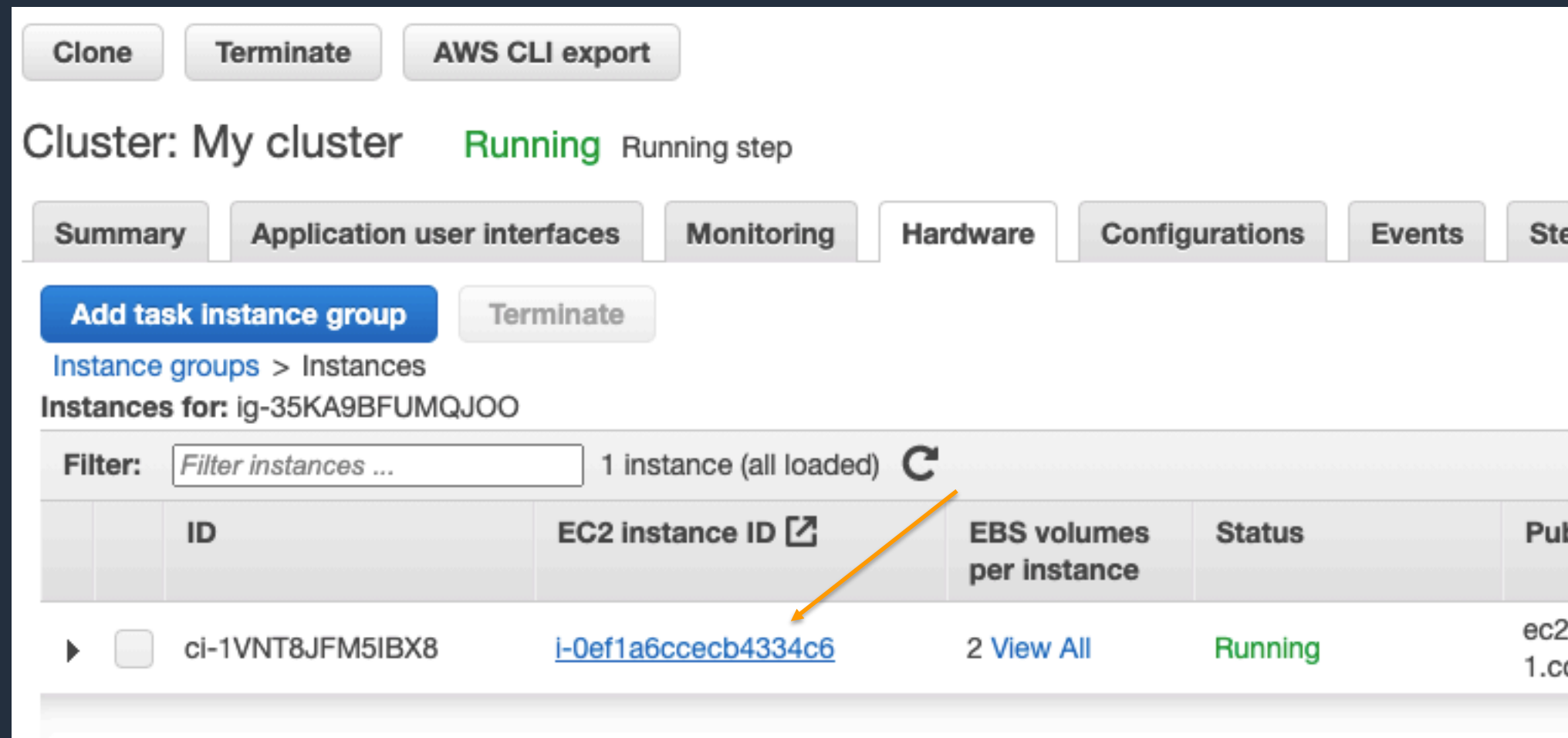
Instance groups

Filter: 2 instance groups (all loaded)

ID	Status	Node type & name	Instance type
▶ ig-KOD8UM2RQ9IE	Running	CORE Core - 2	m5.xlarge 4 vCore, 16 GiB memory EBS Storage: 64 GiB
▶ ig-35KA9BFUMQJOO	Running	MASTER Master - 1	m5.xlarge 4 vCore, 16 GiB memory EBS Storage: 64 GiB

How to launch an EMR cluster

Click in the EC2 Instance ID Link



Clone Terminate AWS CLI export


Cluster: My cluster **Running** Running step


Summary Application user interfaces Monitoring Hardware Configurations Events Ste

Add task instance group Terminate

Instance groups > Instances

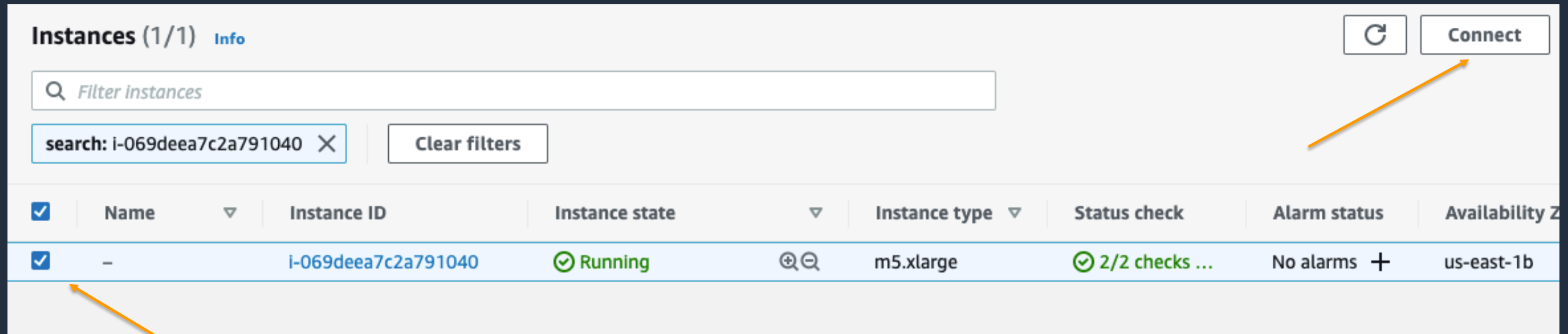
Instances for: ig-35KA9BFUMQJOO

Filter: 1 instance (all loaded) 

	ID	EC2 instance ID 	EBS volumes per instance	Status	Pub
▶ <input type="checkbox"/>	ci-1VNT8JFM5IBX8	i-0ef1a6ccebcb4334c6	2 View All	Running	ec2 1.co

How to launch an EMR cluster

In the EC2 Instance Page, select the instance ID and Click in Connect



The screenshot shows the AWS Management Console 'Instances' page. At the top, there's a header 'Instances (1/1)' with an 'Info' link. Below this is a search bar labeled 'Filter instances' and a search filter box containing 'search: i-069deea7c2a791040' with a 'Clear filters' button. On the right, there are 'Refresh' and 'Connect' buttons. An orange arrow points from the 'Connect' button to the 'Connect' button. Below the filters is a table of instances. The first instance is selected, indicated by a blue checkmark in the first column. An orange arrow points from the bottom left towards the selected instance. The table has columns: Name, Instance ID, Instance state, Instance type, Status check, Alarm status, and Availability Zone. The instance details are: Name (hyphen), Instance ID (i-069deea7c2a791040), Instance state (Running with a green checkmark), Instance type (m5.xlarge), Status check (2/2 checks ... with a green checkmark), Alarm status (No alarms +), and Availability Zone (us-east-1b).

<input checked="" type="checkbox"/>	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
<input checked="" type="checkbox"/>	-	i-069deea7c2a791040	Running	m5.xlarge	2/2 checks ...	No alarms +	us-east-1b

How to launch an EMR cluster

Select the Session Manager option and click in Connect

EC2 > Instances > i-069deea7c2a791040 > Connect to instance

Connect to instance [Info](#)
Connect to your instance i-069deea7c2a791040 using any of these options

EC2 Instance Connect

Session Manager

SSH client

Session Manager usage:

- Connect to your instance without SSH keys or a bastion host.
- Sessions are secured using an AWS Key Management Service key.
- You can log session commands and details in an Amazon S3 bucket or CloudWatch Logs log group.
- Configure sessions on the Session Manager [Preferences](#) page.

Cancel

Connect

How to launch an EMR cluster

In the terminal change run the below commands

`sudo su - hadoop`

You need to work with the hadoop user

```
sh-4.2$ whoami
ssm-user
sh-4.2$ sudo su - hadoop
Last login: Thu Feb 11 10:05:37 UTC 2021 on pts/0

EEEEEEEEEEEEEEEEEEEE MMMMMMM                      MMMMMMM RRRRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M                      M::::::::M R:::::::::R
EE:::::EEEEEEEEEE::E M::::::::M                      M::::::::M R::::RRRRRR:::R
  E::::E          EEEEE M::::::::M                      M::::::::M RR::::R      R::::R
  E::::E          M:::::M:::M  M:::M:::M  M:::M:::M  R:::R      R::::R
  E:::::EEEEEEEEEE  M:::::M M:::M M:::M M:::M  R:::RRRRRR:::R
  E:::::~::~:~::~:E  M:::::M  M:::M:::M  M:::~::~:~::~:R:::::~::~:RR
  E:::::EEEEEEEEEE  M:::::M  M:::::M  M:::~::~:~::~:R:::RRRRRR:::R
  E::::E          M:::::M  M:::M  M:::~::~:~::~:R:::R      R::::R
  E::::E          EEEEE M:::::M  MMM  M:::~::~:~::~:R:::R      R::::R
EE:::::EEEEEEEEEE::E M:::::M                      M:::::M  R:::R      R::::R
E:::::~::~:~::~:E  M:::::M                      M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM                      MMMMMMM RRRRRRRR      RRRRRR

[hadoop@ip-172-31-33-10 ~]$ whoami
hadoop
[hadoop@ip-172-31-33-10 ~]$
```

