Steps to Set Up an AWS EMR Cluster

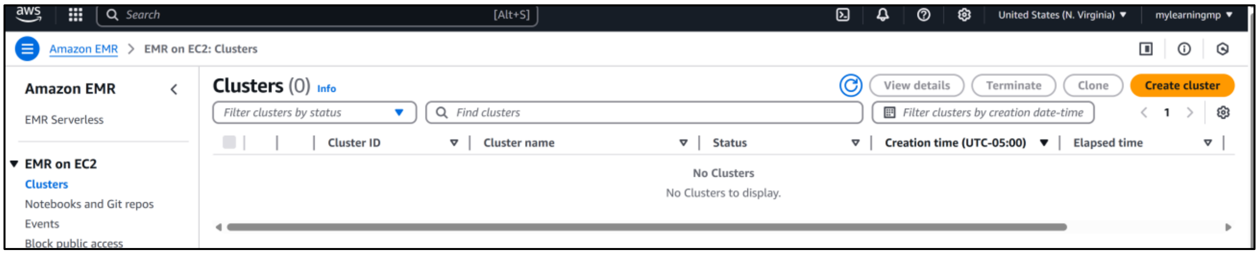
EMR Cluster can be setup from AWS Console, SDK or AWS CLI. For the blog we will cover setup via **AWS Console**.

Step 1: Log in to AWS Console

1. Login to AWS Console.
2. Once logged in, search for "EMR" in the search bar and navigate to **Amazon EMR.**

Step 2: Create EMR Cluster.

1. In the EMR Cluster, click on "Create Cluster"

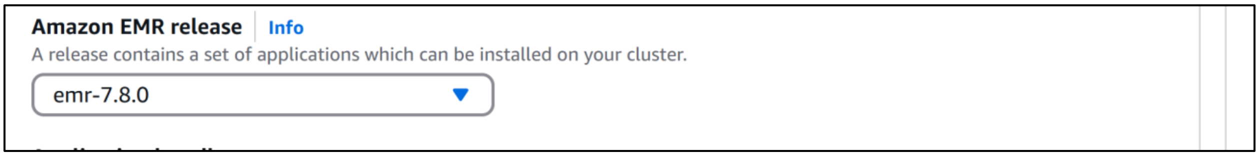


No Clusters

No Clusters to display.

Step 3: Configure the Software and Applications:

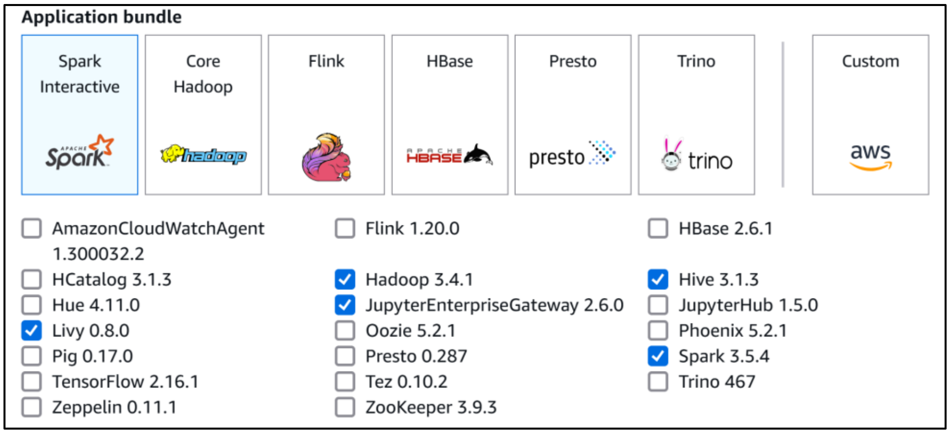
1. Choose the EMR Release version: We are going to choose **emr-7.8.0**



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

emr-7.8.0 ▾

2. Choose the Application bundle: From the available choices we choose "Spark". Checkmarks are chosen automatically based on the application choice you make.



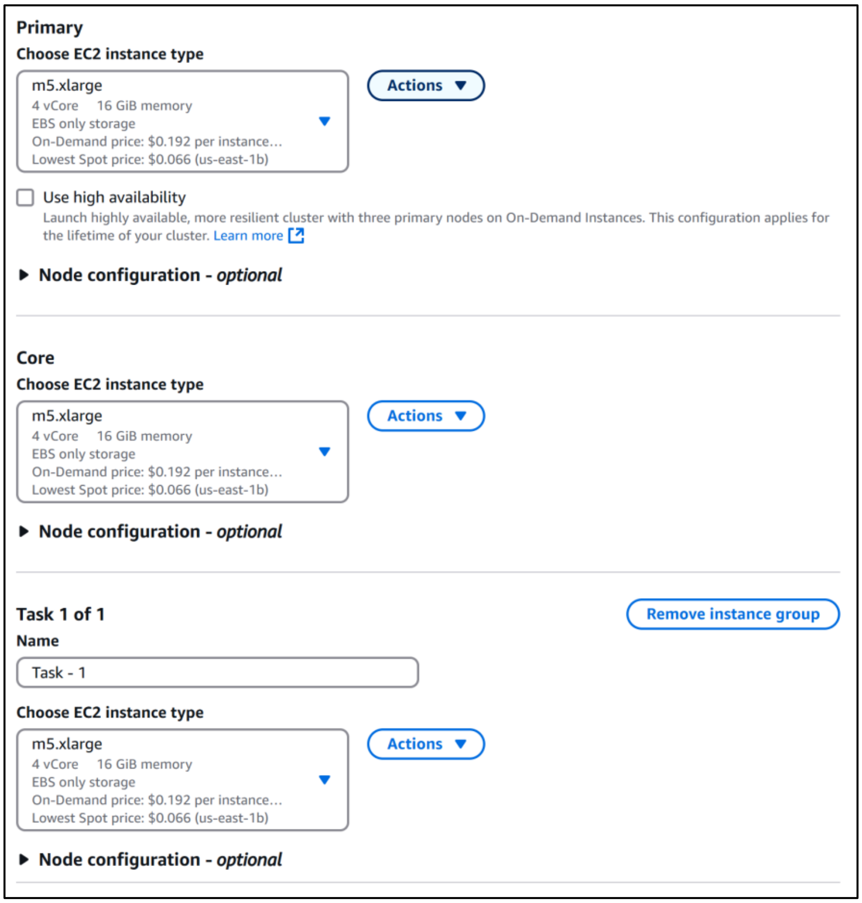
☐ AmazonCloudWatchAgent 1.300032.2
☐ HCatalog 3.1.3
☐ Hue 4.11.0
☑ Livy 0.8.0
☐ Pig 0.17.0
☐ TensorFlow 2.16.1
☐ Zeppelin 0.11.1
☐ Flink 1.20.0
☑ Hadoop 3.4.1
☑ JupyterEnterpriseGateway 2.6.0
☐ Oozie 5.2.1
☐ Presto 0.287
☐ Tez 0.10.2
☐ ZooKeeper 3.9.3
☐ HBase 2.6.1
☑ Hive 3.1.3
☐ JupyterHub 1.5.0
☐ Phoenix 5.2.1
☑ Spark 3.5.4
☐ Trino 467

Step 4: Configure Instance Types and Cluster Size :

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

Chose the Primary or Master Node:

1. **Choose EC2 Instance Types**:
    1. Master Node: Choose an instance type like m5-xlarge
    2. Core Nodes : Something like r5.xlarge for memory intensive applications like Spark.
    3. Task Nodes : If your workload needs extra processing power without additional storage , make use of Task node for optimizing Cluster performance.
2. Cluster Sizing: For optimal performance and cost-effectiveness, tailor your cluster size to your specific needs. Define the initial number of instances for each node type, considering both your data volume and processing intensity. EMR's built-in auto-scaling functionality provides the flexibility to automatically scale the cluster up or down based on real-time workload fluctuations, preventing unnecessary resource overhead.


Actions ▼
Actions ▼
Remove instance group
Actions ▼
▶ Node configuration - *optional*
▶ Node configuration - *optional*
▶ Node configuration - *optional*

Primary

Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage
On-Demand price: $0.192 per instance...
Lowest Spot price: $0.066 (us-east-1b)

□ Use high availability
Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. Learn more

Core

Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage
On-Demand price: $0.192 per instance...
Lowest Spot price: $0.066 (us-east-1b)

Task 1 of 1

Name

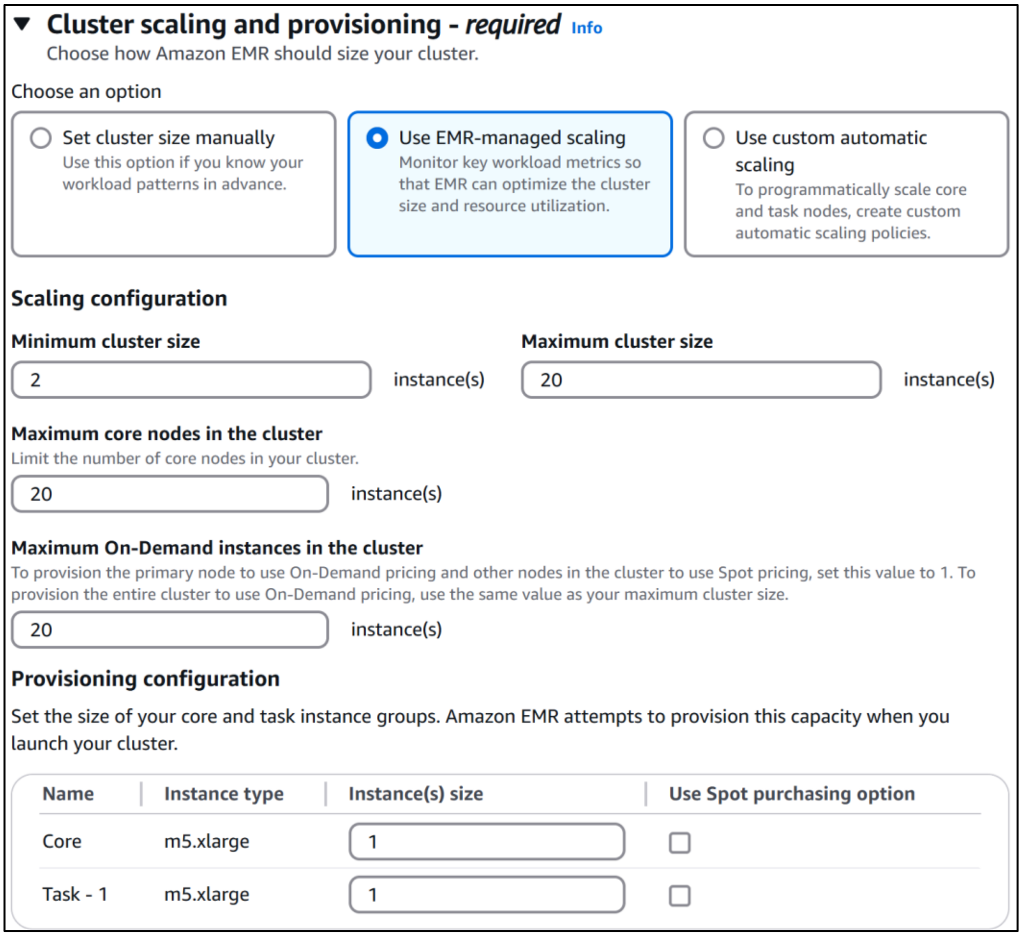
Task - 1

Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage
On-Demand price: $0.192 per instance...
Lowest Spot price: $0.066 (us-east-1b)

Step 5: Enabling EMR-Managed Scaling for Auto-Scaling

Chosen below is the “Use EMR-managed scaling” , since we want EMR to dynamically auto scale based on workload. Here you can also specify the Min and Max Cluster Size and also choose Max Core Nodes and On-Demand instances in cluster.
You can also specify number of instance that the cluster starts with under the “Provisioning Configuration” setting.



▼ **Cluster scaling and provisioning - *required*** Info

Choose how Amazon EMR should size your cluster.

Choose an option

○ Set cluster size manually
Use this option if you know your workload patterns in advance.

● Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.

○ Use custom automatic scaling
To programmatically scale core and task nodes, create custom automatic scaling policies.

Scaling configuration

Minimum cluster size
2 instance(s)

Maximum cluster size
20 instance(s)

Maximum core nodes in the cluster
Limit the number of core nodes in your cluster.
20 instance(s)

Maximum On-Demand instances in the cluster
To provision the primary node to use On-Demand pricing and other nodes in the cluster to use Spot pricing, set this value to 1. To provision the entire cluster to use On-Demand pricing, use the same value as your maximum cluster size.
20 instance(s)

Provisioning configuration

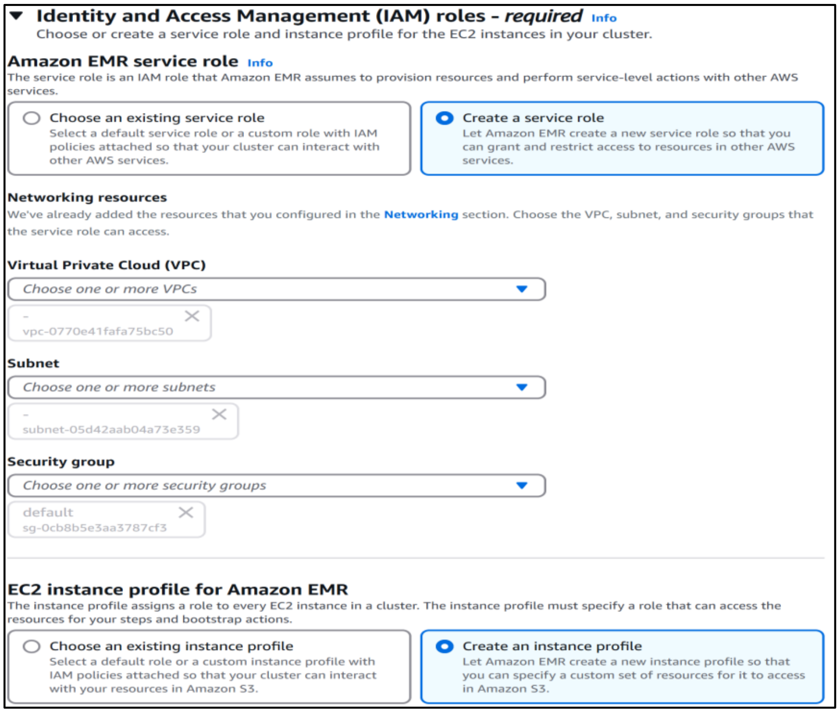
Set the size of your core and task instance groups. Amazon EMR attempts to provision this capacity when you launch your cluster.

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m5.xlarge	1	□
Task - 1	m5.xlarge	1	□

Step 6: Networking and Security Configuration

1. **VPN and Network Configuration:** Choose the appropriate VPC and Firewall for your EMR Setup.
2. **IAM Role configuration :**
   Choose or create new IAM Role granting necessary permissions to the clusters to interact with other AWS services like S3, Dynamo DB etc.
   3. EMR Default Role: Allows cluster to interact with S3 and Dynamobo DB and other services.
   4. EMR Default EC2 Role : Grants EC2 instances in the cluster necessary permission

In our case we choose Amazon to create the Service Roles and Atoscaling Role for EC2. Please see our preferences below.



Identity and Access Management (IAM) roles - *required* 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.

Networking resources

We've already added the resources that you configured in the **Networking** section. Choose the VPC, subnet, and security groups that the service role can access.

Virtual Private Cloud (VPC)

Choose one or more VPCs

-
vpc-0770e41fafa75bc50

Subnet

Choose one or more subnets

-
subnet-05d42aab04a73e359

Security group

Choose one or more security groups

default
sg-0cb8b5e3aa3787cf3

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.

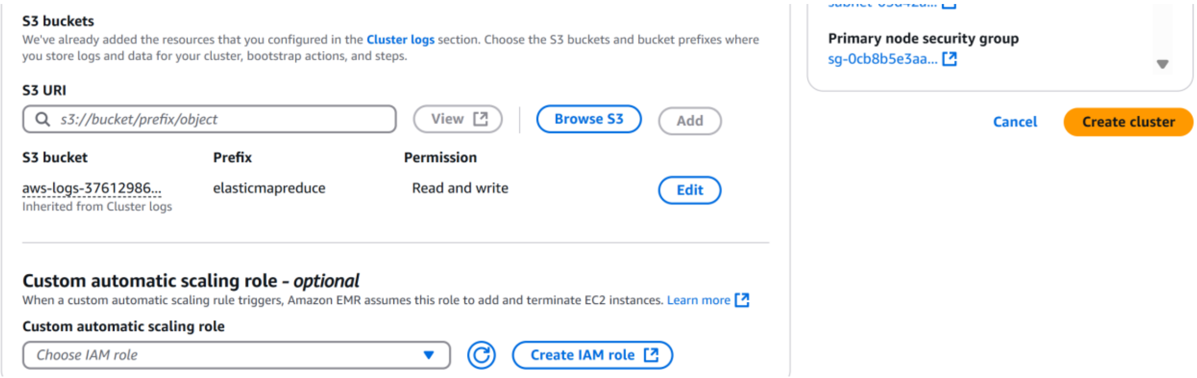
● Create an instance profile
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Step 7 : Logs:

Configure AWS S3 location for logging.

Step 8: Review and Create Cluster:

1. Review all configuration.
2. After a through review hit the “Create Cluster”



S3 buckets

We've already added the resources that you configured in the Cluster logs section. Choose the S3 buckets and bucket prefixes where you store logs and data for your cluster, bootstrap actions, and steps.

S3 URI: s3://bucket/prefix/object

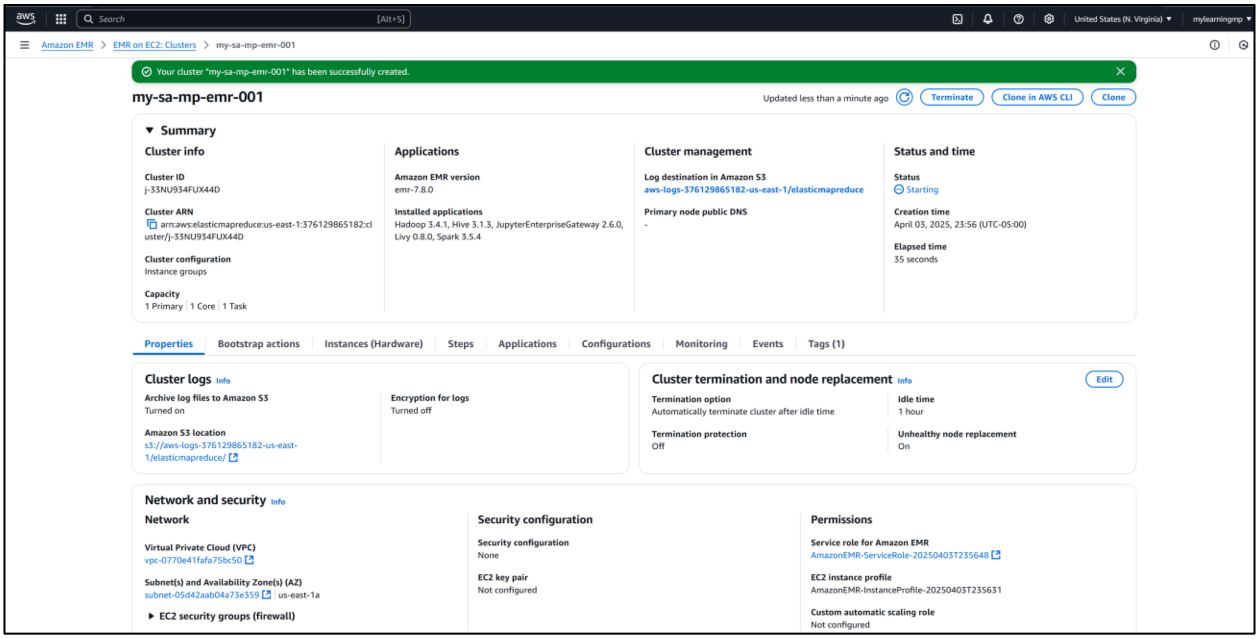
S3 bucket	Prefix	Permission
aws-logs-37612986... Inherited from Cluster logs	elasticmapreduce	Read and write

Custom automatic scaling role - *optional*

When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. Learn more

Custom automatic scaling role

Primary node security group: sg-0cb8b5e3aa...



Amazon EMR > EMR on EC2: Clusters > my-sa-mp-emr-001

Your cluster "my-sa-mp-emr-001" has been successfully created.

my-sa-mp-emr-001

Updated less than a minute ago

Summary

Cluster info

Cluster ID: j-33NU934FUX44D

Cluster ARN: arn:aws:elasticmapreduce:us-east-1:376129865182:cluster/j-33NU934FUX44D

Cluster configuration: Instance groups

Capacity: 1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version: emr-7.8.0

Installed applications: Hadoop 3.4.1, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.4

Cluster management

Log destination in Amazon S3: aws-logs-376129865182-us-east-1/elasticmapreduce

Primary node public DNS: -

Status and time

Status: Starting

Creation time: April 03, 2025, 23:56 (UTC-05:00)

Elapsed time: 35 seconds

Cluster logs

Archive log files to Amazon S3: Turned on

Encryption for logs: Turned off

Amazon S3 location: s3://aws-logs-376129865182-us-east-1/elasticmapreduce/

Cluster termination and node replacement

Termination option: Automatically terminate cluster after idle time

Idle time: 1 hour

Termination protection: Off

Unhealthy node replacement: On

Network and security

Network

Virtual Private Cloud (VPC): vpc-0770e41fafa75bc50

Subnet(s) and Availability Zone(s) (AZ): subnet-05d42aab04a73e359 | us-east-1a

EC2 security groups (firewall)

Security configuration

Security configuration: None

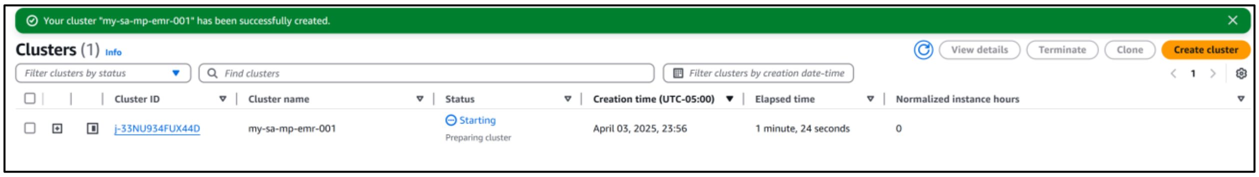
EC2 key pair: Not configured

Permissions

Service role for Amazon EMR: AmazonEMR-ServiceRole-20250403T235648

EC2 instance profile: AmazonEMR-InstanceProfile-20250403T235631

Custom automatic scaling role: Not configured


Info
View details
Terminate
Clone
Create cluster
Filter clusters by status
Find clusters
Filter clusters by creation date-time
1

Your cluster "my-sa-mp-emr-001" has been successfully created.

Clusters (1)

Cluster ID	Cluster name	Status	Creation time (UTC-05:00)	Elapsed time	Normalized instance hours
j-33NU934FUX44D	my-sa-mp-emr-001	Starting Preparing cluster	April 03, 2025, 23:56	1 minute, 24 seconds	0

Step 9: Manage and Monitor the Cluster.

1. Once your cluster is up and running, you can monitor its status via the **EMR Dashboard**.
2. Use **AWS CloudWatch** for real-time monitoring of cluster metrics, resource utilization, and job progress.

3. You can also SSH into the master node if you need to directly manage or troubleshoot the cluster.