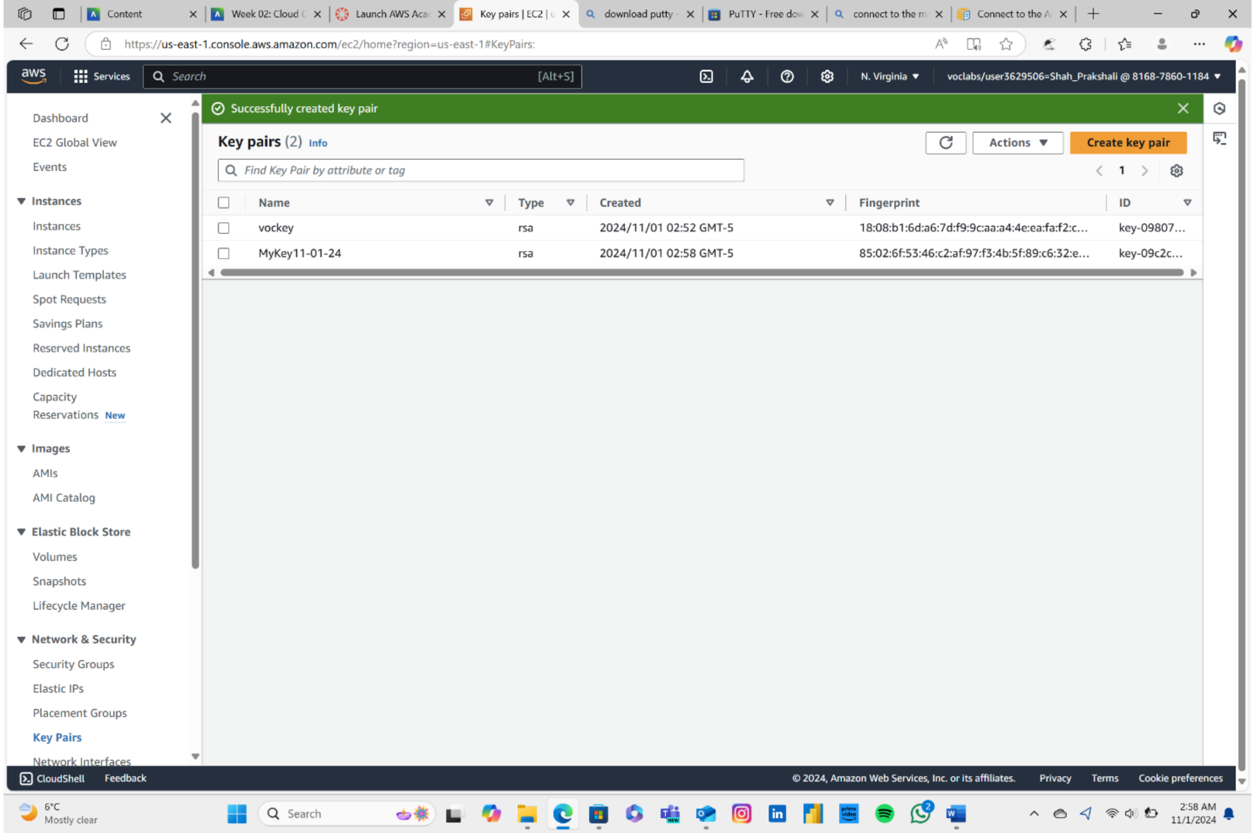
AWS EMR CLUSTER SETUP

Hnas-on Lab

- Task 1: Create a Security Key Pair



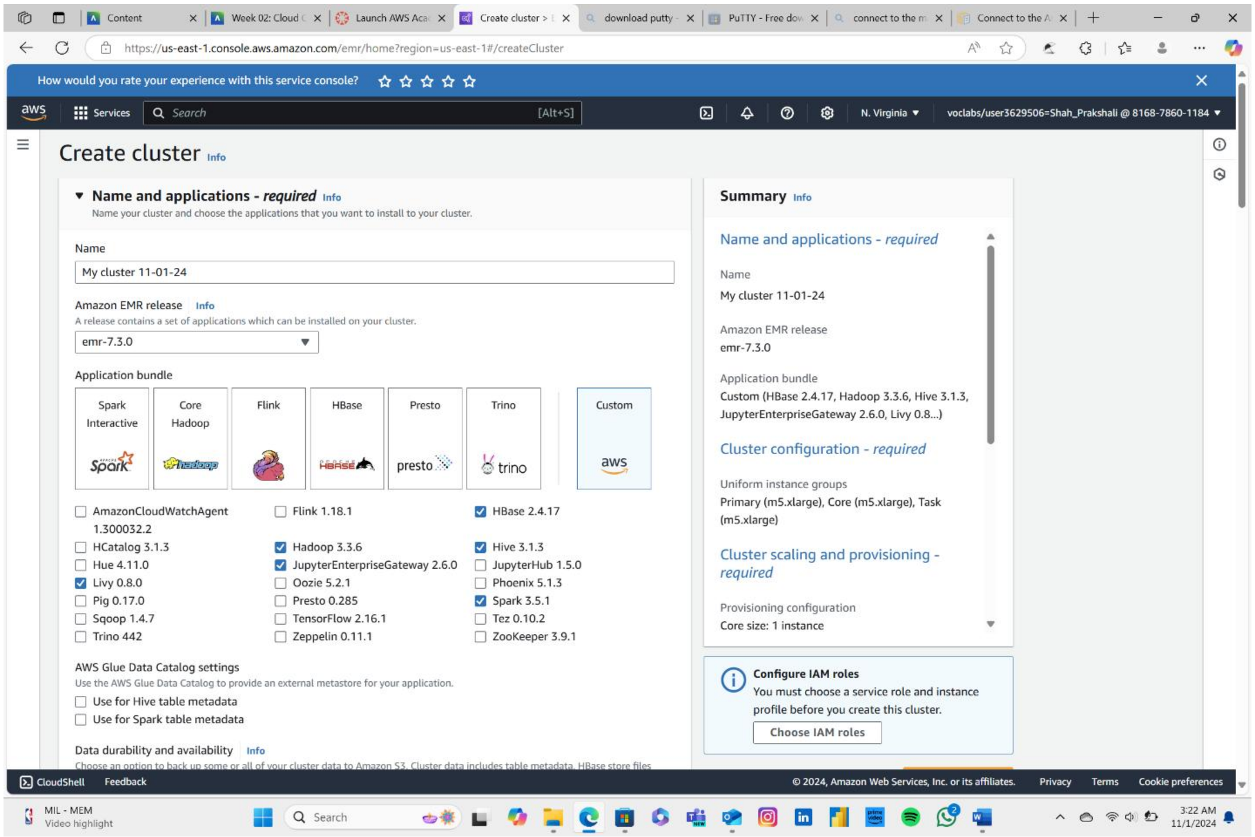
Successfully created key pair

Key pairs (2) Info

	Name	Type	Created	Fingerprint	ID
□	vockey	rsa	2024/11/01 02:52 GMT-5	18:08:b1:6d:a6:7d:f9:9c:aa:a4:4e:ea:fa:f2:c...	key-09807...
□	MyKey11-01-24	rsa	2024/11/01 02:58 GMT-5	85:02:6f:53:46:c2:af:97:f3:4b:5f:89:c6:32:e...	key-09c2c...

Task 2

- Task 2.1 Choose your software



Content
Week 02: Cloud
Launch AWS Aca
Create cluster
download putty
PuTTY - Free dow
connect to the m
Connect to the A
https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/createCluster
How would you rate your experience with this service console?
Services
Search
[Alt+S]
N. Virginia
voclabs/user3629506=Shah_Prakshali @ 8168-7860-1184
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
MIL - MEM
Video highlight
Search
3:22 AM
11/1/2024

Create cluster Info

Name and applications - *required* Info

Name your cluster and choose the applications that you want to install to your cluster.

Name: My cluster 11-01-24

Amazon EMR release Info

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

emr-7.3.0

Application bundle

□ AmazonCloudWatchAgent 1.300032.2
□ HCatalog 3.1.3
□ Hue 4.11.0
☑ Livy 0.8.0
□ Pig 0.17.0
□ Sqoop 1.4.7
□ Trino 442
□ Flink 1.18.1
☑ Hadoop 3.3.6
☑ JupyterEnterpriseGateway 2.6.0
□ Oozie 5.2.1
□ Presto 0.285
□ TensorFlow 2.16.1
□ Zeppelin 0.11.1
☑ HBase 2.4.17
☑ Hive 3.1.3
□ JupyterHub 1.5.0
□ Phoenix 5.1.3
☑ Spark 3.5.1
□ Tez 0.10.2
□ ZooKeeper 3.9.1

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

□ Use for Hive table metadata
□ Use for Spark table metadata

Data durability and availability Info

Choose an option to back up some or all of your cluster data to Amazon S3. Cluster data includes table metadata, HBase store files

Summary Info

Name and applications - *required*

Name
My cluster 11-01-24

Amazon EMR release
emr-7.3.0

Application bundle
Custom (HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8...)

Cluster configuration - *required*

Uniform instance groups
Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

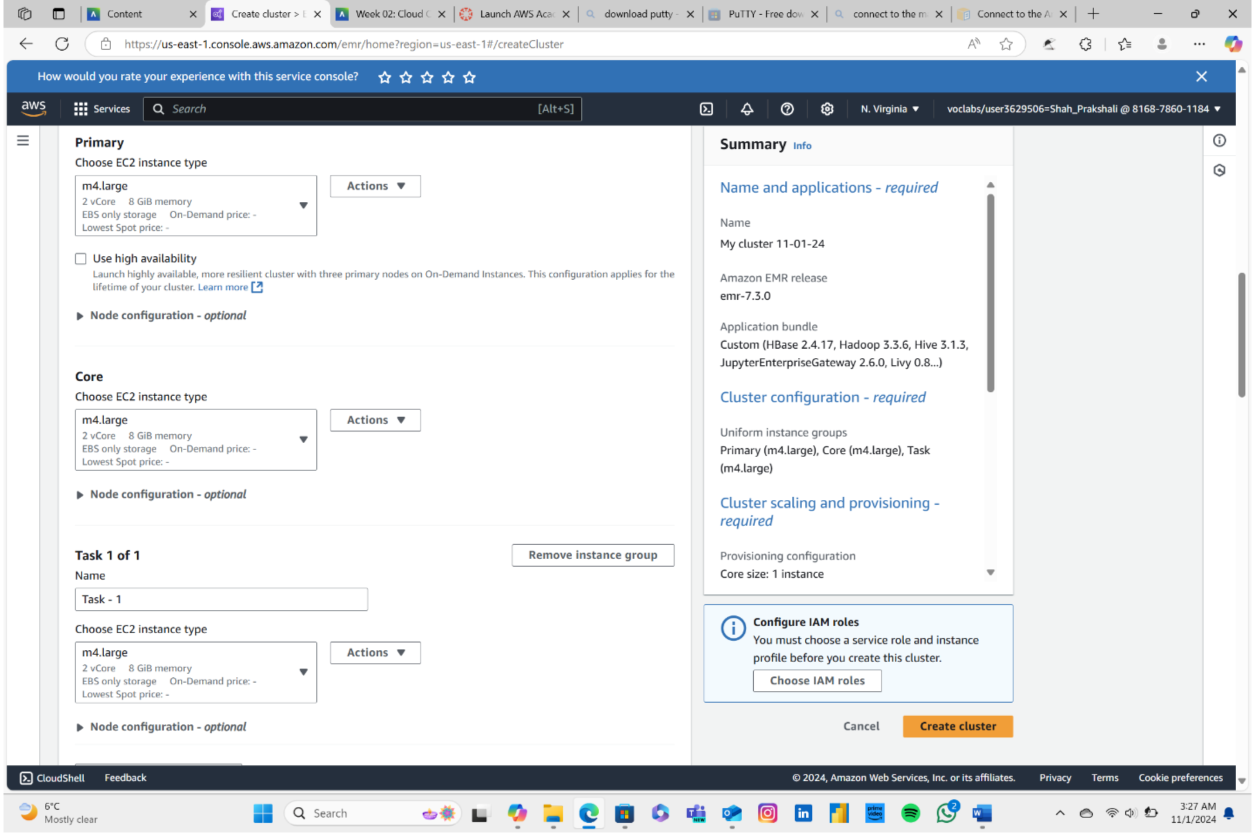
Cluster scaling and provisioning - *required*

Provisioning configuration
Core size: 1 instance

Configure IAM roles

You must choose a service role and instance profile before you create this cluster.

- Task 2.2 Configure Instance Group, instance type



https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/createCluster

How would you rate your experience with this service console? ☆ ☆ ☆ ☆ ☆

N. Virginia ▼ voclabs/user3629506=Shah_Prakshali @ 8168-7860-1184 ▼

Primary

Choose EC2 instance type

m4.large
2 vCore 8 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

□ 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

▶ Node configuration - *optional*

Core

Choose EC2 instance type

m4.large
2 vCore 8 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ Node configuration - *optional*

Task 1 of 1

Name

Task - 1

Choose EC2 instance type

m4.large
2 vCore 8 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ Node configuration - *optional*

Summary Info

Name and applications - *required*

Name

My cluster 11-01-24

Amazon EMR release

emr-7.3.0

Application bundle

Custom (HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8...)

Cluster configuration - *required*

Uniform instance groups

Primary (m4.large), Core (m4.large), Task (m4.large)

Cluster scaling and provisioning - *required*

Provisioning configuration

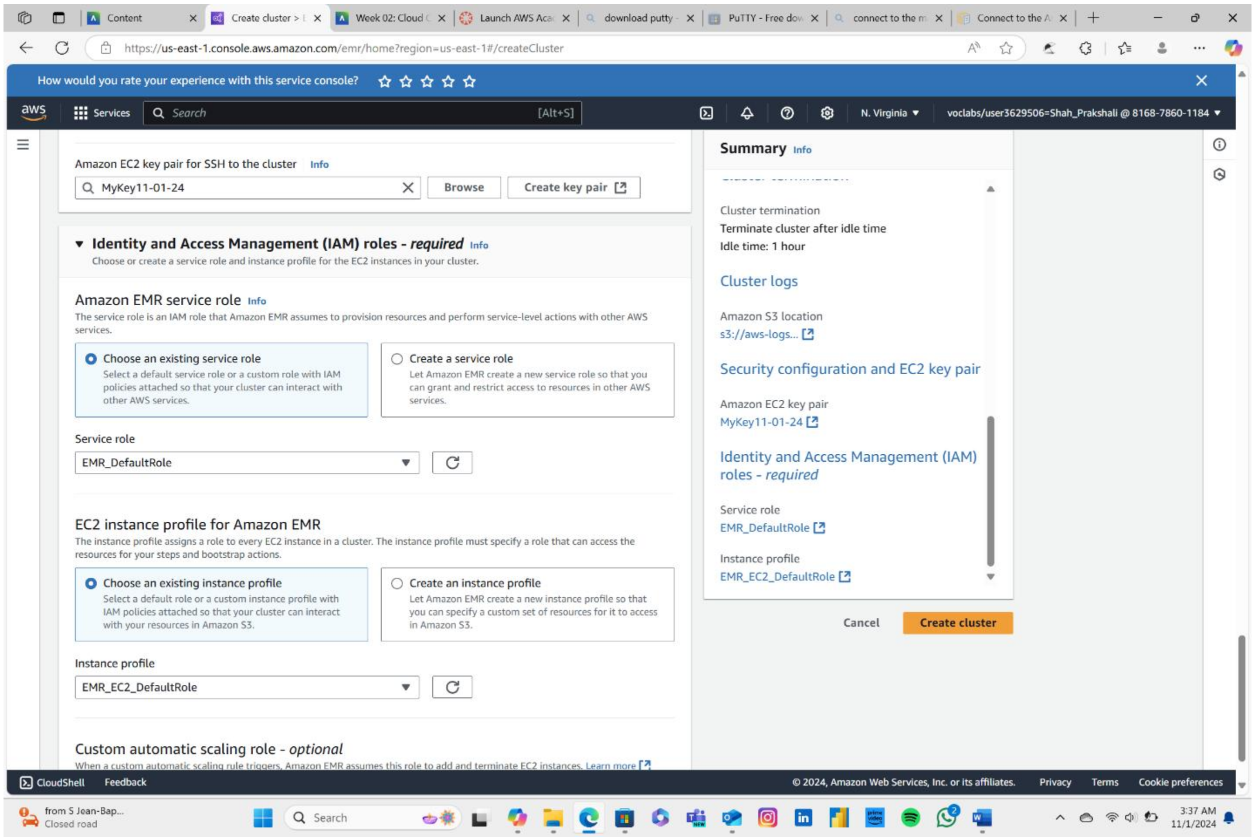
Core size: 1 instance

Configure IAM roles

You must choose a service role and instance profile before you create this cluster.

© 2024, Amazon Web Services, Inc. or its affiliates.

- Task 2.3 Configure Security



https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/createCluster

How would you rate your experience with this service console?

N. Virginia ▾ voclabs/user3629506=Shah_Prakshali @ 8168-7860-1184 ▾

Amazon EC2 key pair for SSH to the cluster Info

MyKey11-01-24 ✕ Browse Create key pair

▼ **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.

Service role

EMR_DefaultRole ▼

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.

Instance profile

EMR_EC2_DefaultRole ▼

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

Summary Info

Cluster termination
Terminate cluster after idle time
Idle time: 1 hour

Cluster logs

Amazon S3 location
s3://aws-logs...

Security configuration and EC2 key pair

Amazon EC2 key pair
MyKey11-01-24

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

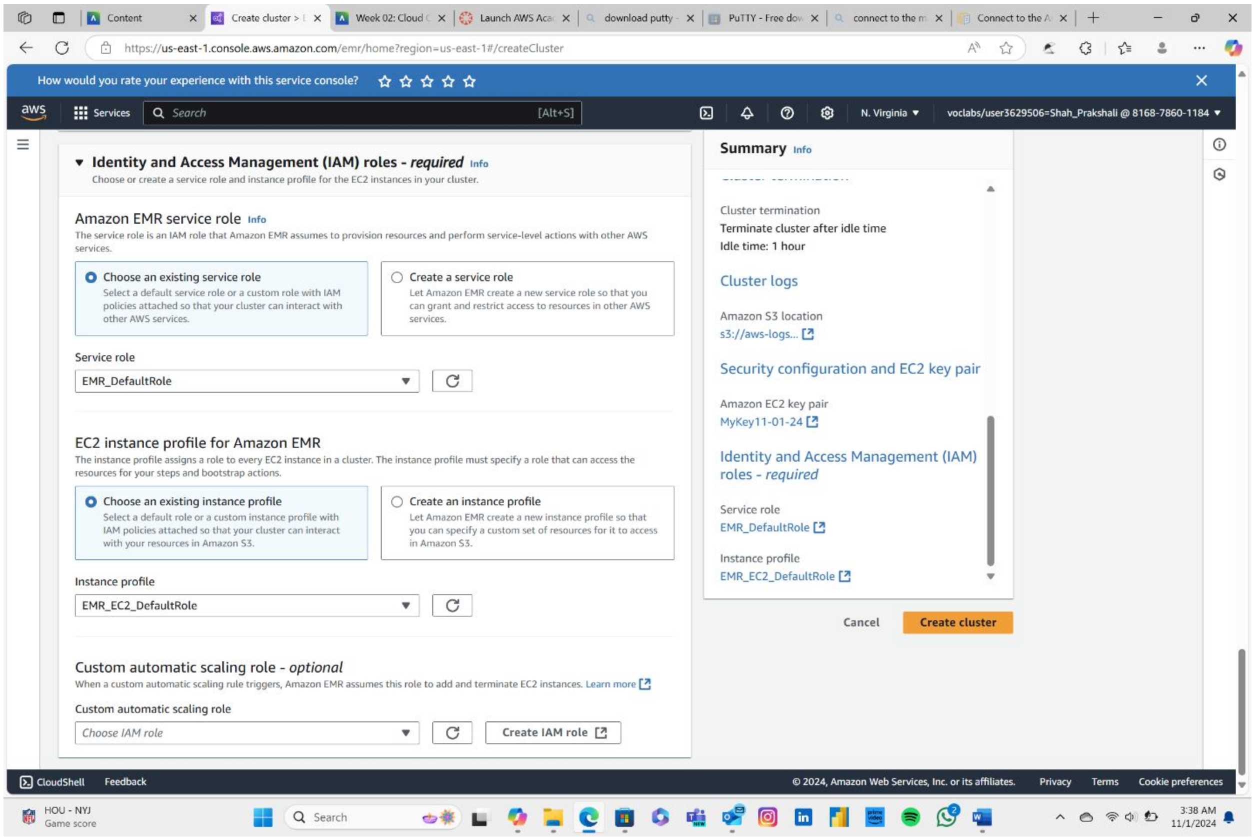
Service role
EMR_DefaultRole

Instance profile
EMR_EC2_DefaultRole

Cancel Create cluster

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

- Task 2.4 Configure IAM Role


Content
Create cluster > E
Week 02: Cloud C
Launch AWS Acad
download putty
PuTTY - Free dow
connect to the m
Connect to the A
https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/createCluster
How would you rate your experience with this service console? ☆ ☆ ☆ ☆ ☆
aws
Services
Search [Alt+S]
N. Virginia ▼
voclabs/user3629506=Shah_Prakshali @ 8168-7860-1184 ▼

▼ **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.

Service role: EMR_DefaultRole ▼

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.

Instance profile: EMR_EC2_DefaultRole ▼

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: *Choose IAM role* ▼

Summary Info

Cluster termination
Terminate cluster after idle time
Idle time: 1 hour

Cluster logs

Amazon S3 location
s3://aws-logs...

Security configuration and EC2 key pair

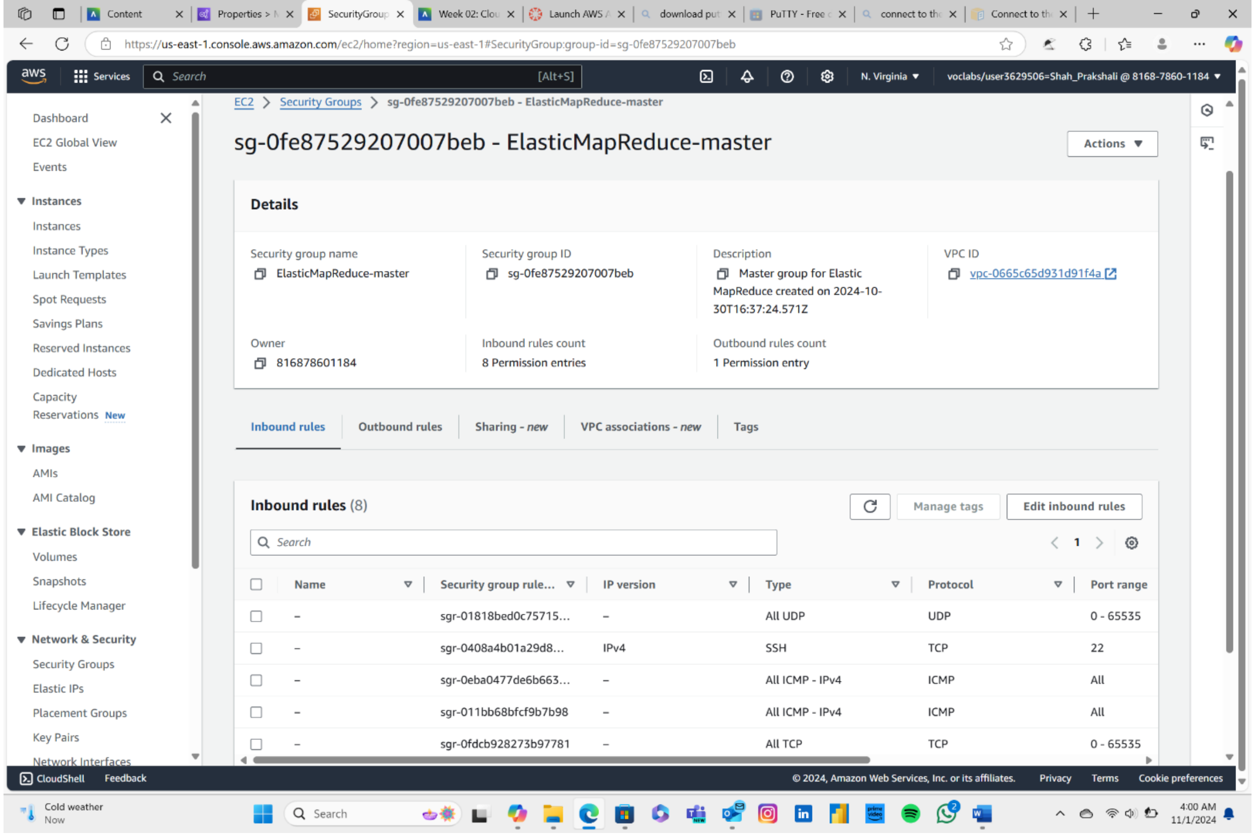
Amazon EC2 key pair
MyKey11-01-24

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

Service role
EMR_DefaultRole

Instance profile
EMR_EC2_DefaultRole

- Task 2.5 Configure Security Group Inbound Rules



sg-0fe87529207007beb - ElasticMapReduce-master

Details

Security group name: ElasticMapReduce-master

Security group ID: sg-0fe87529207007beb

Description: Master group for Elastic MapReduce created on 2024-10-30T16:37:24.571Z

VPC ID: vpc-0665c65d931d91f4a

Owner: 816878601184

Inbound rules count: 8 Permission entries

Outbound rules count: 1 Permission entry

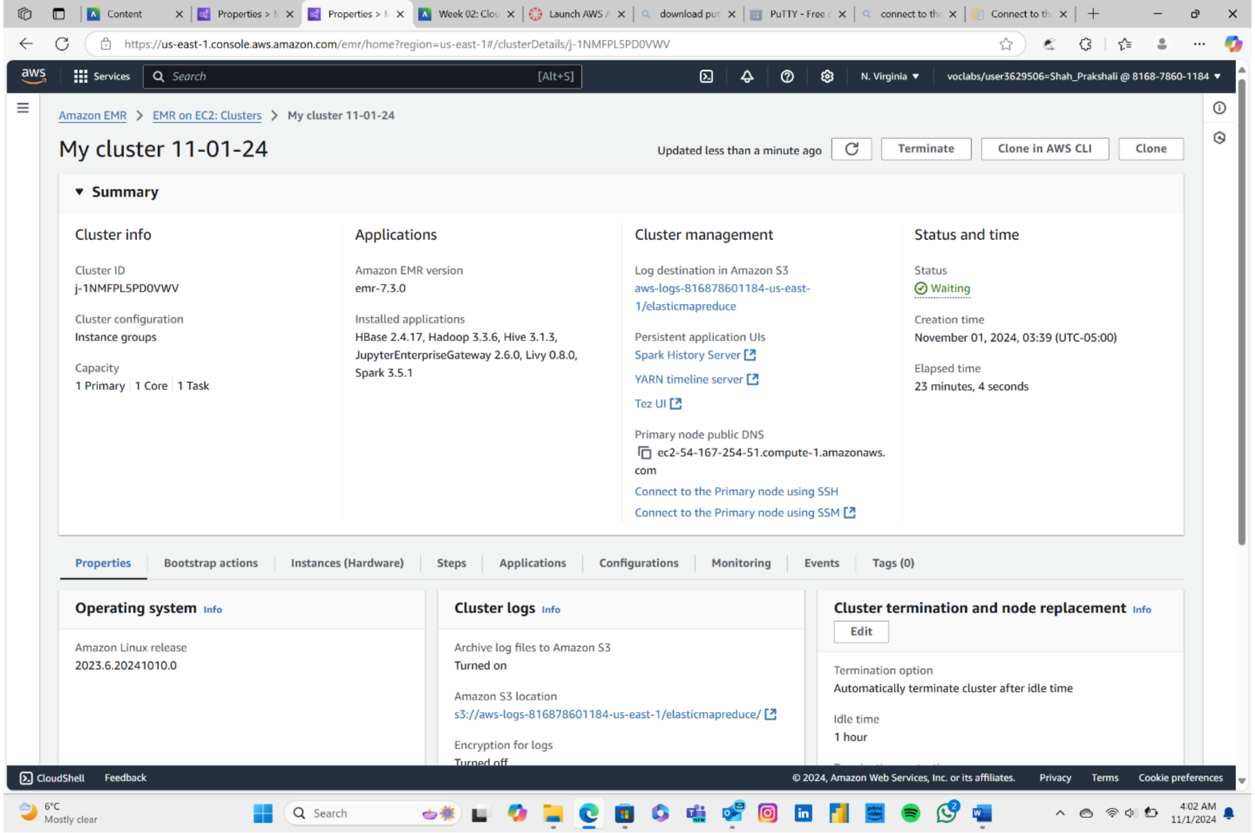
Inbound rules (8)

□	Name	Security group rule...	IP version	Type	Protocol	Port range
□	–	sgr-01818bed0c75715...	–	All UDP	UDP	0 - 65535
□	–	sgr-0408a4b01a29d8...	IPv4	SSH	TCP	22
□	–	sgr-0eba0477de6b663...	–	All ICMP - IPv4	ICMP	All
□	–	sgr-011bb68bfcf9b7b98	–	All ICMP - IPv4	ICMP	All
□	–	sgr-0fdcb928273b97781	–	All TCP	TCP	0 - 65535

Cold weather
Now

4:00 AM
11/1/2024

- Task 2.6 Waiting for EMR to be provisioned



https://us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusterDetails/j-1NMFPL5PD0VWV

Amazon EMR > EMR on EC2: Clusters > My cluster 11-01-24

My cluster 11-01-24

Updated less than a minute ago

▼ Summary

Cluster info

Cluster ID
j-1NMFPL5PD0VWV

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version
emr-7.3.0

Installed applications
HBase 2.4.17, Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.1

Cluster management

Log destination in Amazon S3
aws-logs-816878601184-us-east-1/elasticmapreduce

Persistent application UIs
Spark History Server
YARN timeline server
Tez UI

Primary node public DNS
ec2-54-167-254-51.compute-1.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
✓ Waiting

Creation time
November 01, 2024, 03:39 (UTC-05:00)

Elapsed time
23 minutes, 4 seconds

Operating system Info

Amazon Linux release
2023.6.20241010.0

Cluster logs Info

Archive log files to Amazon S3
Turned on

Amazon S3 location
s3://aws-logs-816878601184-us-east-1/elasticmapreduce/

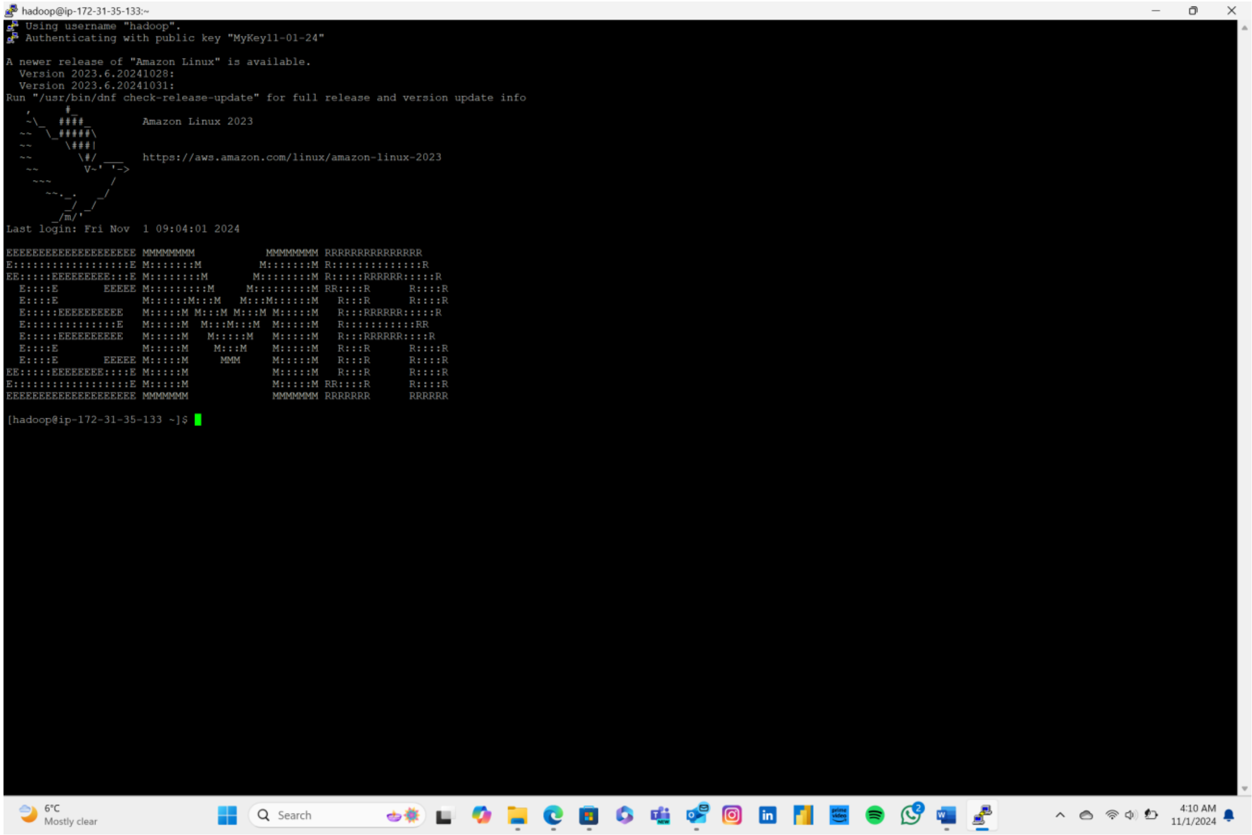
Encryption for logs
Turned off

Cluster termination and node replacement Info

Termination option
Automatically terminate cluster after idle time

Idle time
1 hour

- Task 3 Connect to the EMR



```
Using username "hadoop".
Authenticating with public key "MyKey11-01-24"

A newer release of "Amazon Linux" is available.
  Version 2023.6.20241028:
  Version 2023.6.20241031:
Run "/usr/bin/dnf check-release-update" for full release and version update info
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Fri Nov  1 09:04:01 2024

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::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   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:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-35-133 ~]$
```

A. Reflection

Diving into the exercise of connecting to and logging into the Hadoop cluster was a comprehensive learning experience.

What I Learned

Configuring Security Key Pairs: I understood the significance of creating and managing security key pairs, which are essential for secure access to the EC2 instances.

Setting Up EMR Clusters: I learned how to navigate the AWS Management Console to set up an EMR cluster with specific software configurations like Hadoop, Spark, Hive, and HBase. Choosing the appropriate instance types (m4.large) was crucial for optimal performance.

Inbound Rules for SSH Access: Configuring the security group's inbound rules to allow SSH access from any IP address was a pivotal step, ensuring seamless connectivity.

Challenges Faced

Inbound Rules Configuration: One major challenge was ensuring the correct configuration of inbound rules to allow SSH access. Initially, I encountered issues with connecting to the EMR cluster due to misconfigured rules.

Instance Type Selection: Choosing the correct instance types for the primary, core, and task nodes also posed a challenge as selecting the wrong instance could impact performance.

Connectivity Issues: There were moments when connecting via Putty required rechecking the host name and ensuring the key was correctly saved and selected.

Resolutions

Referring to Instructions: Whenever I faced an issue, I revisited the detailed instructions and video guides provided in the task list. This ensured that I adhered to the steps meticulously.

Trial and Error: For configuring inbound rules, I used a trial-and-error approach, modifying settings until the SSH connection was successful.

Patience and Precision: Ensuring precision in following each step, from creating the security key pair to configuring the EMR cluster, played a vital role in overcoming challenges.

This exercise not only enhanced my technical skills in cloud-based infrastructure management but also underscored the importance of meticulous attention to detail and problem-solving in dynamic environments.