COMP 330/543 Lab 1: Using Amazon EMR to Run a Hadoop Job

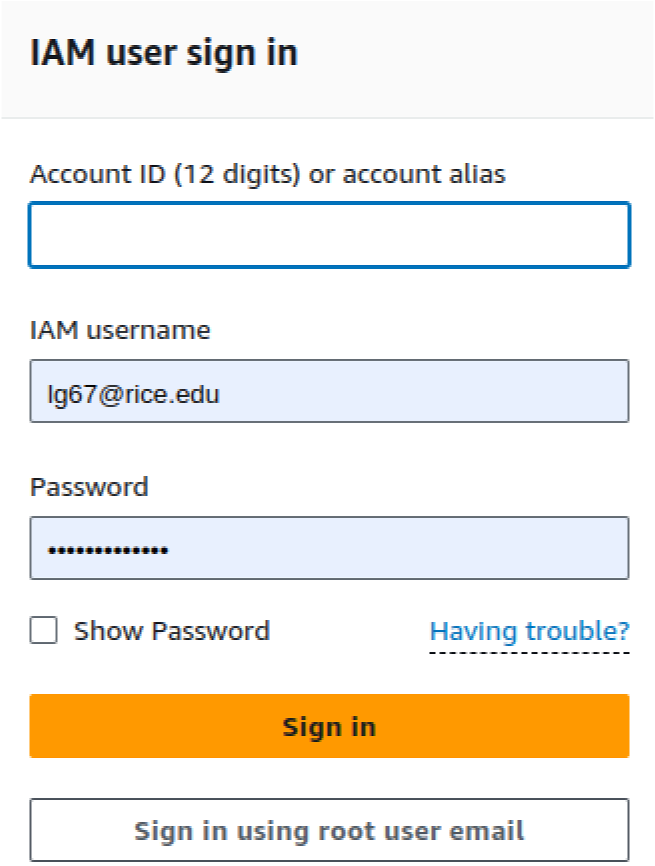
Note: this assumes you have previously signed up for an Amazon account.

In this lab, you will:

1. Create a Hadoop cluster using Amazon AWS.
2. Compile a Hadoop MapReduce program using the Java compiler (Hadoop is a popular open source MapReduce tool).
3. Connect to your AWS cluster using SSH (Secure Shell).
4. Load data into Hadoop's Distributed File System (HDFS).
5. Run your Hadoop program to process the data.

Task 1: Start Up a Hadoop Cluster

1. Go to Amazon's AWS website (aws.amazon.com) and click on **"Sign in to console"**. Sign in with your username and password.
    a. If you get a Login page asking for **"IAM user sign in",** click on the **"Sign in using root user mail"** at the bottom.



IAM user sign in

Account ID (12 digits) or account alias

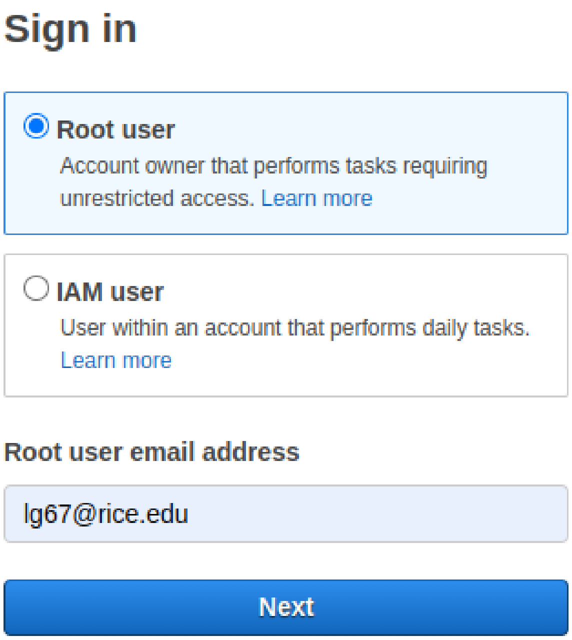
IAM username

lg67@rice.edu

Password

••••••••••••••

b. You should then get a different Login form that looks as follows



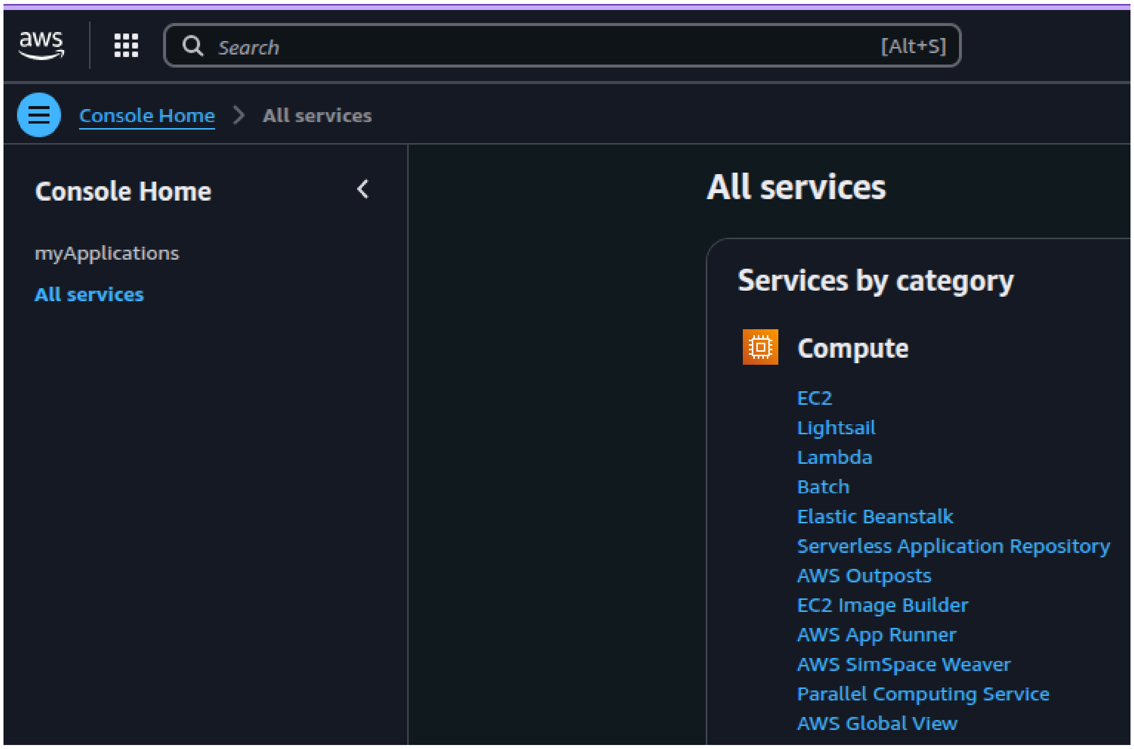
Sign in

● Root user
Account owner that performs tasks requiring unrestricted access. Learn more

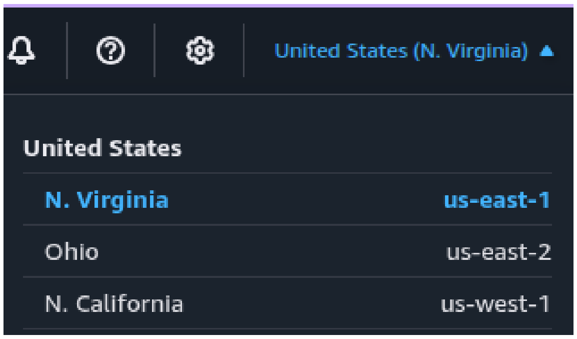
○ IAM user
User within an account that performs daily tasks. Learn more

Root user email address: lg67@rice.edu

2. Click on the hamburger menu icon in the top left corner, followed by **“All Services”**, and finally on **“EC2”** which can be found under the **“Compute”** section.

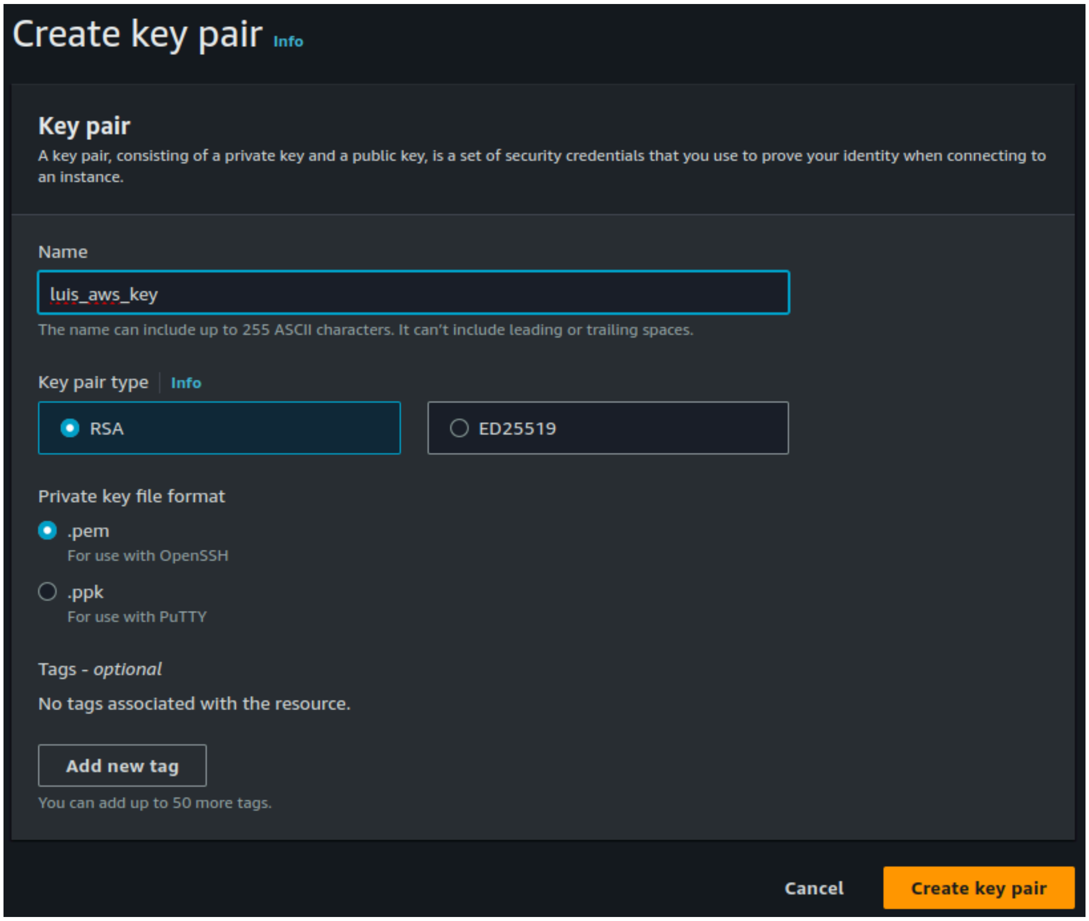


3. In the top right corner, you will find a dropdown menu to select a region. Make sure you select **“United States (N. Virginia) us-east-1”**.



4. Next, you will create a **“Key Pair”** that will allow you to connect securely to the cluster. In the left-hand side menu click on **“Key Pairs”** under **“Network & Security”**.

5. Click the orange “Create key pair” button in the top right corner.
   a. Pick a name that is likely unique to you.
   b. Select RSA as the Key pair type.
   c. Select **“.pem”** for the format if using Mac/Linux or **“.ppk”** if using Windows.



Create key pair Info

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

luis_aws_key

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type | Info

● RSA ○ ED25519

Private key file format

● .pem
For use with OpenSSH

○ .ppk
For use with PuTTY

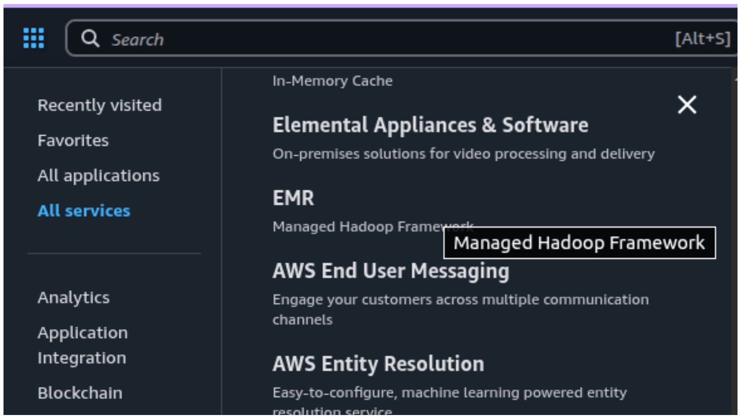
Tags - *optional*

No tags associated with the resource.

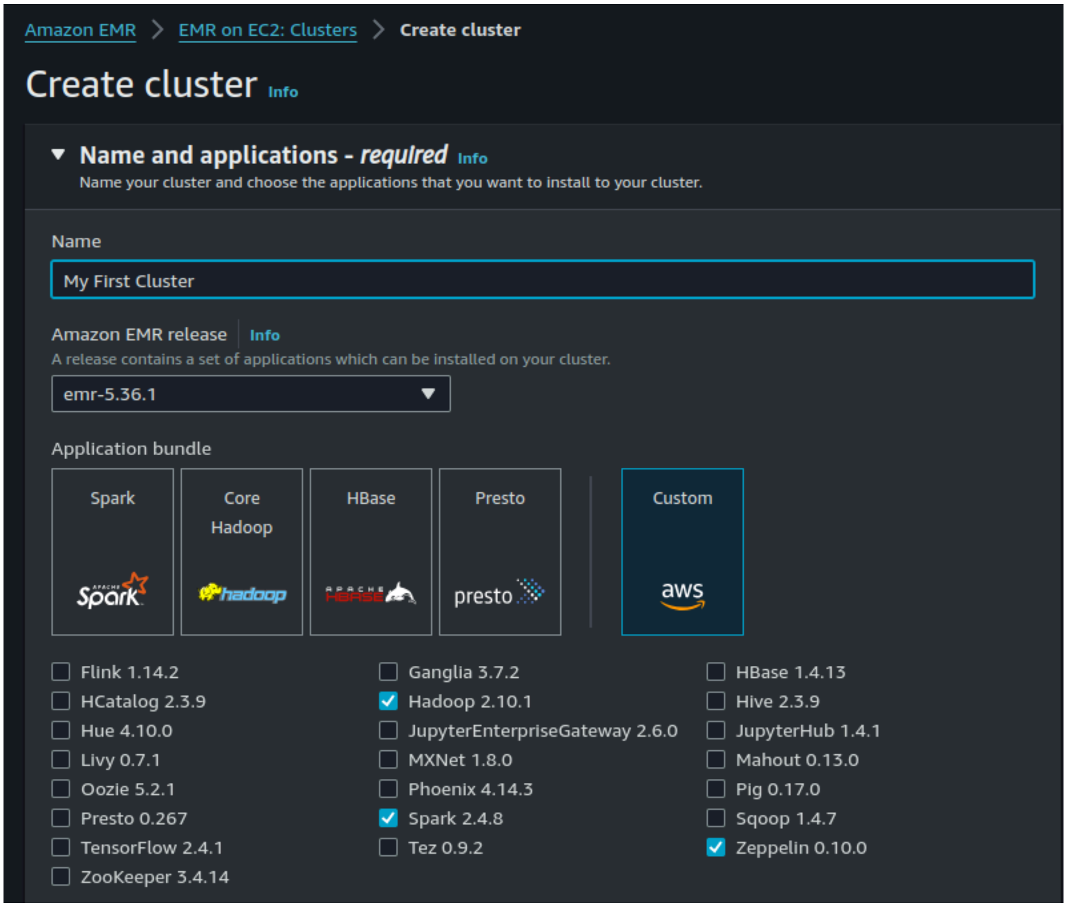
You can add up to 50 more tags.

6. When you click on “Create key pair”, a file named with your key will automatically be downloaded into your computer. The file name will be the same name you chose for your key and its extension will be either .pem or .ppk.
   a. **Make sure to save it somewhere you can find it!**

7. Now it’s time to create a cluster where you can run Hadoop programs. Again, go to “Services” and “All Services” but this time look for “EMR” (Elastic Map Reduce)



8. Click on the orange “Create cluster” button in the top right corner and select the following options:


Amazon EMR > EMR on EC2: Clusters > Create cluster

Create cluster Info

▼ Name and applications - *required* Info
Name your cluster and choose the applications that you want to install to your cluster.

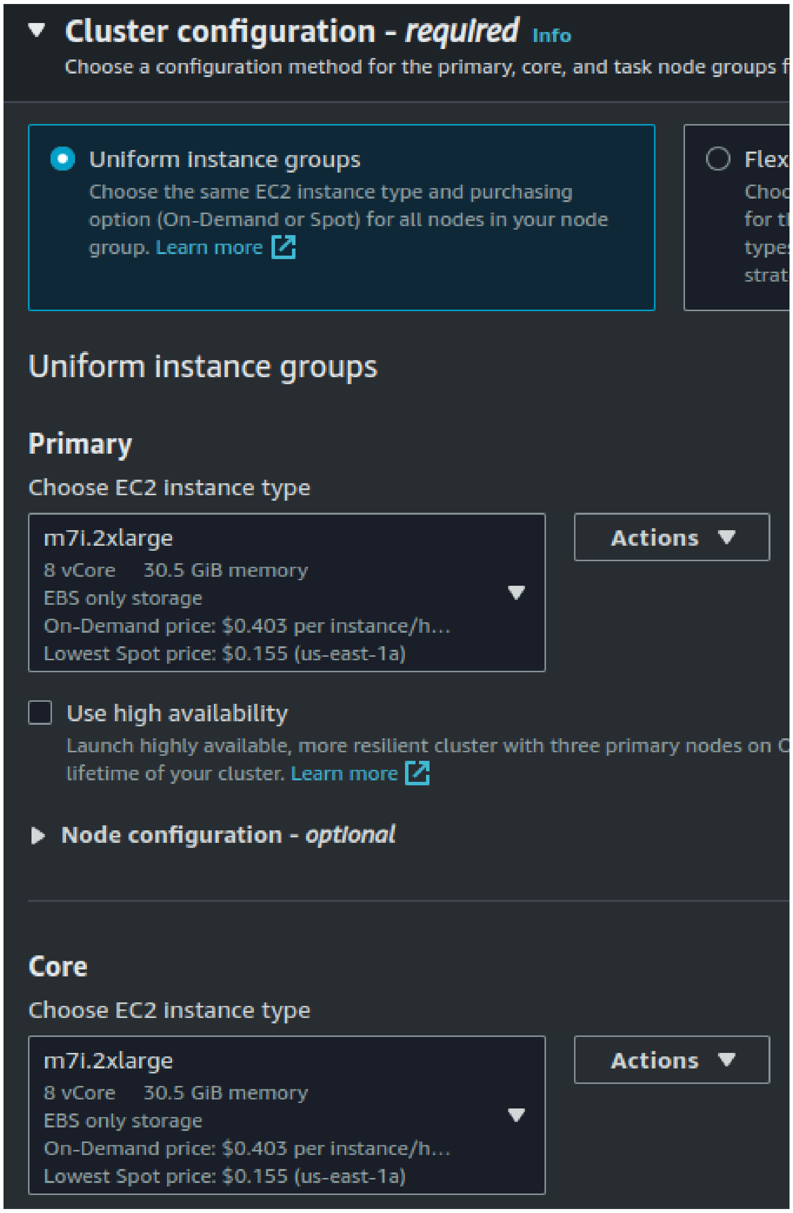
Name
My First Cluster

Amazon EMR release Info
A release contains a set of applications which can be installed on your cluster.
emr-5.36.1 ▼

Application bundle

☐ Flink 1.14.2
☐ HCatalog 2.3.9
☐ Hue 4.10.0
☐ Livy 0.7.1
☐ Oozie 5.2.1
☐ Presto 0.267
☐ TensorFlow 2.4.1
☐ ZooKeeper 3.4.14
☐ Ganglia 3.7.2
☑ Hadoop 2.10.1
☐ JupyterEnterpriseGateway 2.6.0
☐ MXNet 1.8.0
☐ Phoenix 4.14.3
☑ Spark 2.4.8
☐ Tez 0.9.2
☐ HBase 1.4.13
☐ Hive 2.3.9
☐ JupyterHub 1.4.1
☐ Mahout 0.13.0
☐ Pig 0.17.0
☐ Sqoop 1.4.7
☑ Zeppelin 0.10.0

9. For the **“Primary”** an **“Core”** instance groups you can choose ***m7i.2xlarge***. These machines cost ~40 cents per hour each.



▼ Cluster configuration - *required* Info

Choose a configuration method for the primary, core, and task node groups f

● Uniform instance groups
Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. Learn more

○ Flex
Choo
for th
types
strat

Uniform instance groups

Primary

Choose EC2 instance type

m7i.2xlarge
8 vCore 30.5 GiB memory
EBS only storage
On-Demand price: $0.403 per instance/h…
Lowest Spot price: $0.155 (us-east-1a)

□ Use high availability
Launch highly available, more resilient cluster with three primary nodes on C
lifetime of your cluster. Learn more

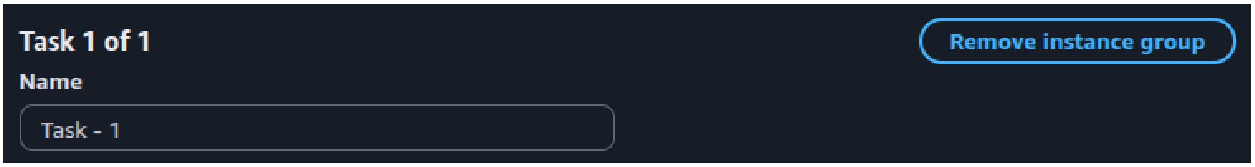
▶ Node configuration - *optional*

Core

Choose EC2 instance type

m7i.2xlarge
8 vCore 30.5 GiB memory
EBS only storage
On-Demand price: $0.403 per instance/h…
Lowest Spot price: $0.155 (us-east-1a)

10. If any additional tasks are shown (“Task - 1” in the picture below), you should remove them by clicking the **“Remove instance group button”**.

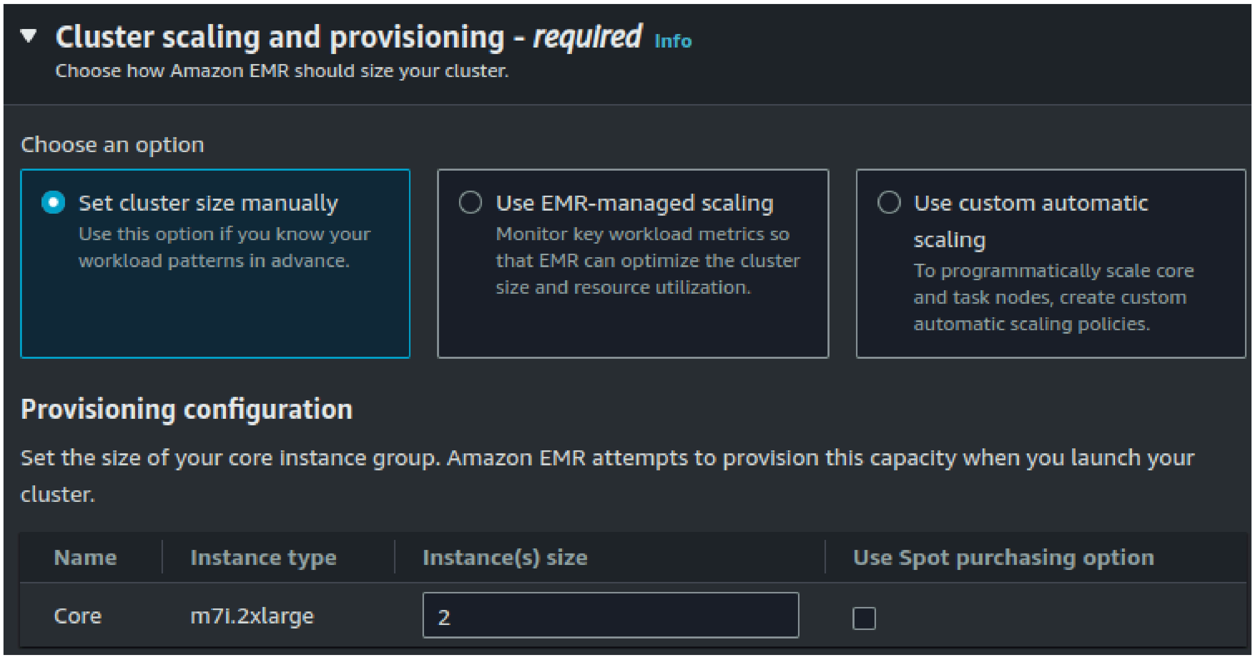


Task 1 of 1

Name

Task - 1

11. Set your cluster size to have two worker nodes



▼ **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.

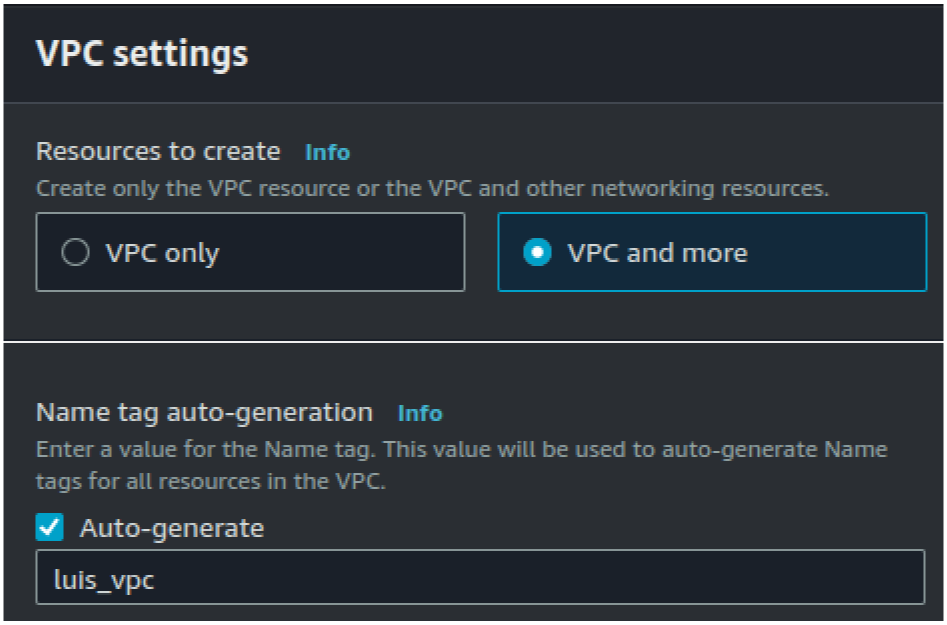
○ Use custom automatic scaling
To programmatically scale core and task nodes, create custom automatic scaling policies.

Provisioning configuration

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

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m7i.2xlarge	2	□

12. Under **“Networking”**, click on **“Create VPC”** which will open a new tab in your browser.
    a. Here, click on the orange **“Create VPC”** button at the top, select **“VPC and more”** and name your VPC.



VPC settings

Resources to create Info

Create only the VPC resource or the VPC and other networking resources.

○ VPC only

● VPC and more

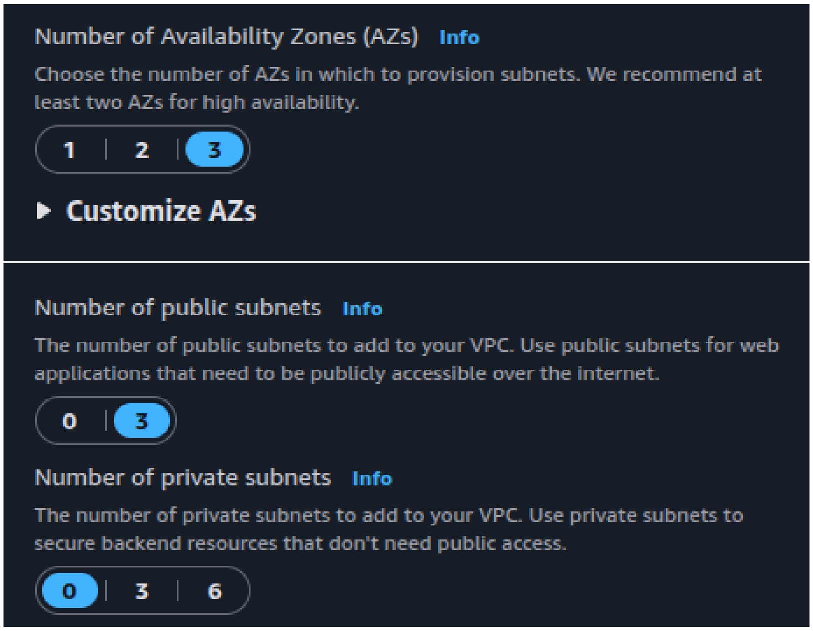
Name tag auto-generation Info

Enter a value for the Name tag. This value will be used to auto-generate Name tags for all resources in the VPC.

☑ Auto-generate

luis_vpc

b. Then choose **3** under **“Number of Availability Zones”** and **set the number of private subnets to 0.**



Number of Availability Zones (AZs) Info

Choose the number of AZs in which to provision subnets. We recommend at least two AZs for high availability.

1 | 2 | [✓ 3]

▶ Customize AZs

Number of public subnets Info

The number of public subnets to add to your VPC. Use public subnets for web applications that need to be publicly accessible over the internet.

0 | [✓ 3]

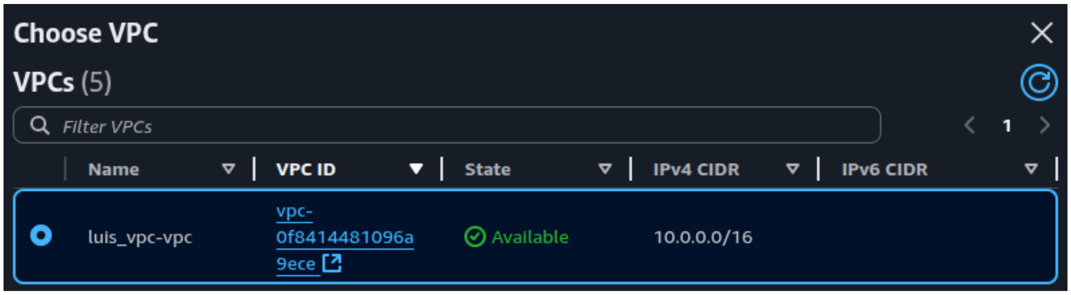
Number of private subnets Info

The number of private subnets to add to your VPC. Use private subnets to secure backend resources that don't need public access.

[✓ 0] | 3 | 6

c. Finally, click on **“Create VPC”** at the bottom.

13. Go back to the previous tab and select your newly created VPC for your cluster. If your new VPC does not show, simply click on the reload button in the top right corner of the pop-up window.

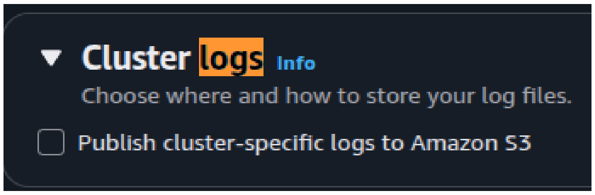


Choose VPC

VPCs (5)

	Name	VPC ID	State	IPv4 CIDR	IPv6 CIDR
●	luis_vpc-vpc	vpc-0f84144810969a9ece	✓ Available	10.0.0.0/16	

14. Under **“Cluster logs”**, disable the checkbox “P***ublish cluster-specific logs to Amazon S3”***

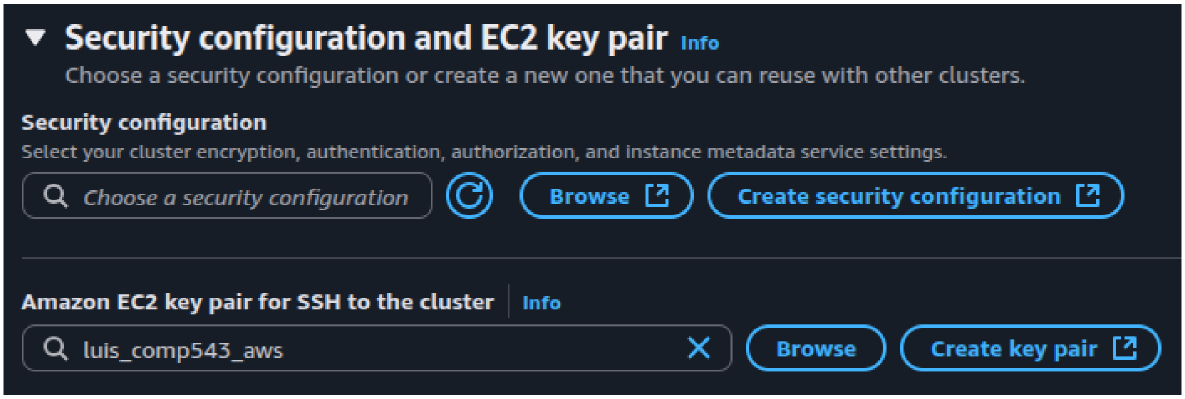


▼ Cluster logs Info

Choose where and how to store your log files.

□ Publish cluster-specific logs to Amazon S3

15. Under “Security configuration and EC2 key pair”, select the key pair created previously.



▼ Security configuration and EC2 key pair Info

Choose a security configuration or create a new one that you can reuse with other clusters.

Security configuration

Select your cluster encryption, authentication, authorization, and instance metadata service settings.

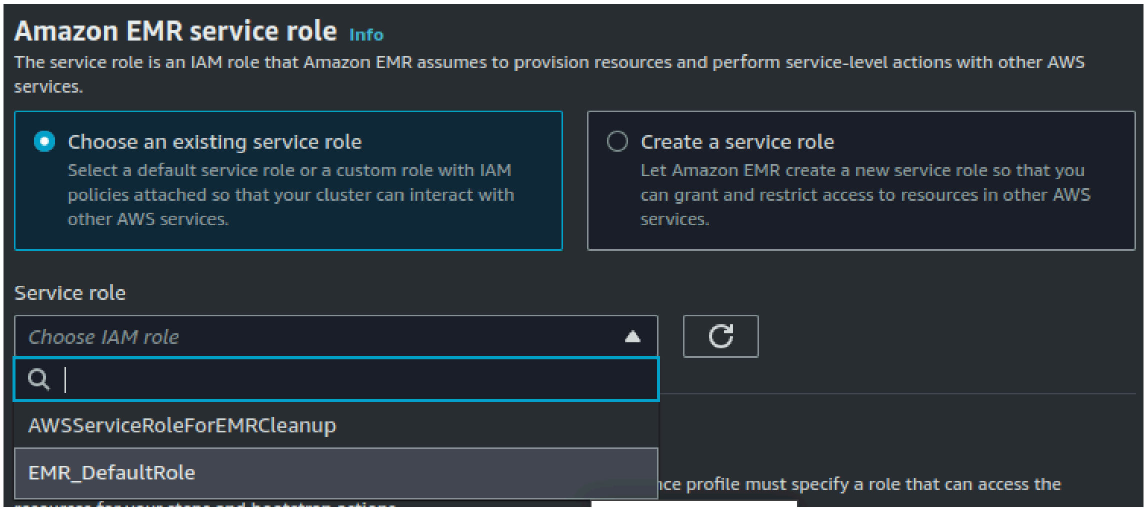
Choose a security configuration

Amazon EC2 key pair for SSH to the cluster Info

luis_comp543_aws

16. Under **“Identitiy and Access Management (IAM) roles”**

   a. Select **“EMR_DefaultRole”** for your **“Amazon EMR service role”**. (Note: if that option does not show see below!)



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

Choose IAM role

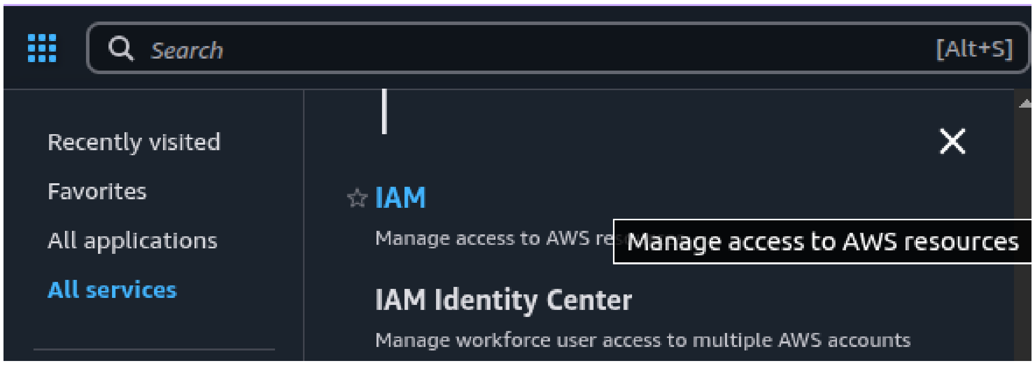
AWSServiceRoleForEMRCleanup

EMR_DefaultRole

ce profile must specify a role that can access the

   b. If “EMR_DefaultRole” is not there, you will need to create it.

      i. Open a new AWS tab/window and again go to **“All Services”** but this time look for **“IAM”**.

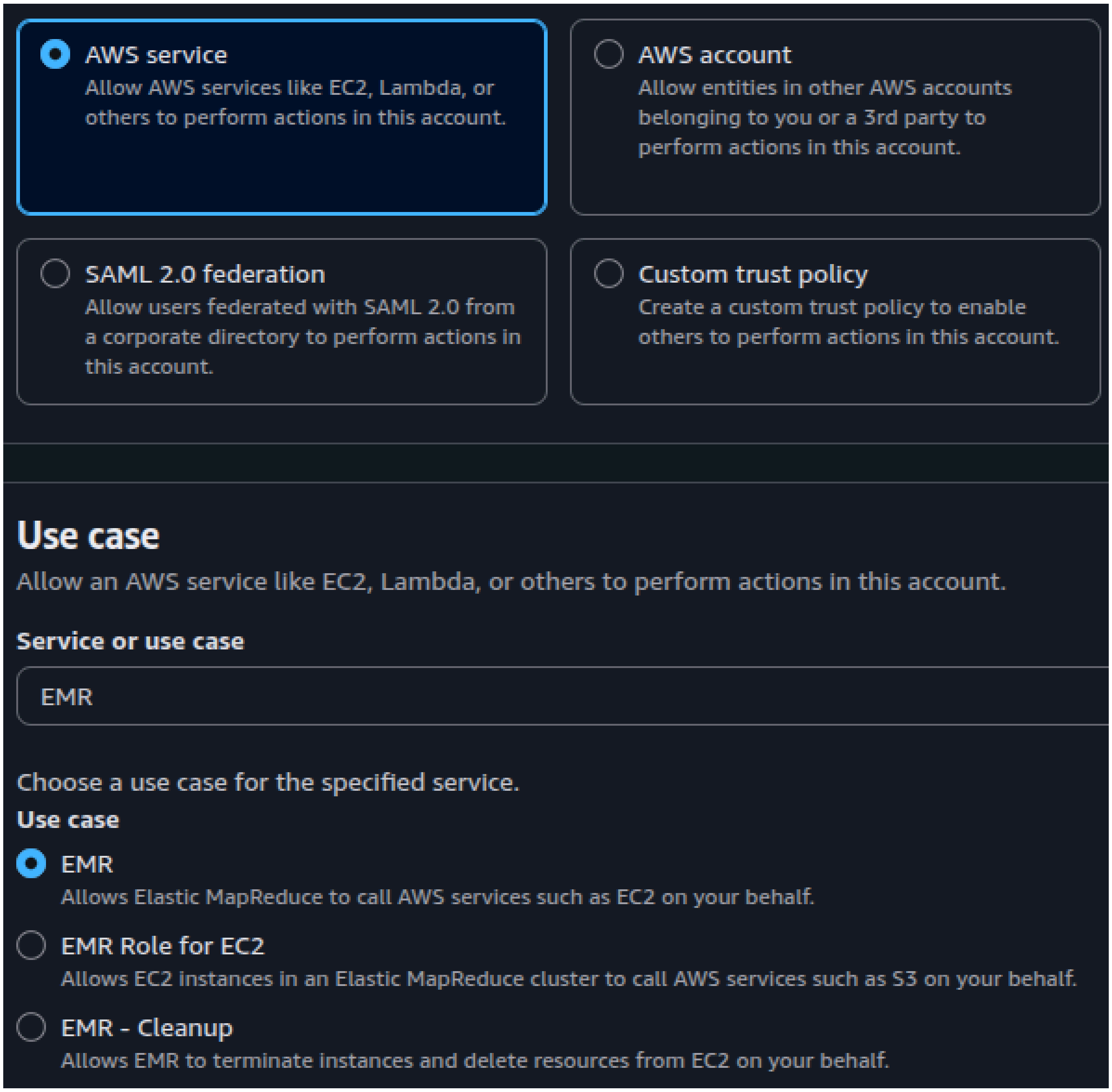


☆ IAM
Manage access to AWS re
Manage access to AWS resources

IAM Identity Center
Manage workforce user access to multiple AWS accounts

ii. In the left-hand side, under **“Access Management”**, click on **“Roles”** and then click the orange **“Create role”** button in the top right corner.

iii. In the next page, select **“AWS service”** as the entity type and **“EMR”** as the service or use case.



● AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.

○ AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.

○ SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.

○ Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.

Use case
Allow an AWS service like EC2, Lambda, or others to perform actions in this account.

Service or use case
EMR

Choose a use case for the specified service.
Use case

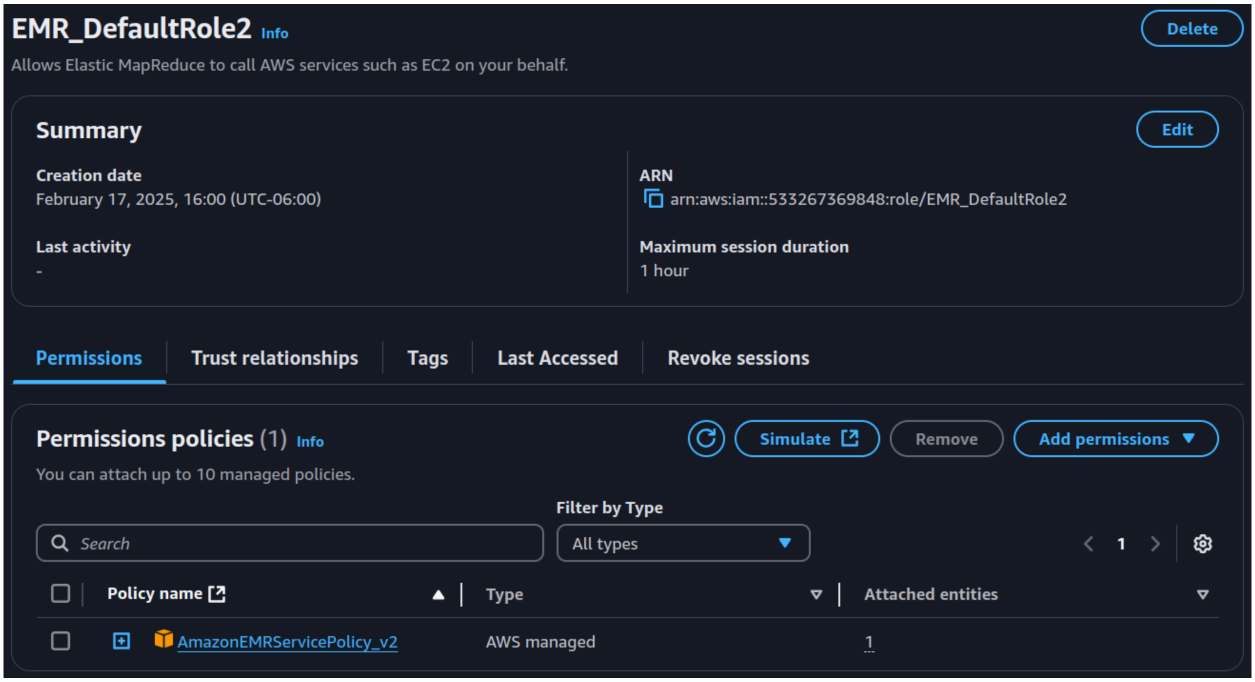
● EMR
Allows Elastic MapReduce to call AWS services such as EC2 on your behalf.

○ EMR Role for EC2
Allows EC2 instances in an Elastic MapReduce cluster to call AWS services such as S3 on your behalf.

○ EMR - Cleanup
Allows EMR to terminate instances and delete resources from EC2 on your behalf.

iv. Click “next” a couple times and then name your role as **“EMR_DefaultRole”**.

v. Finish by clicking **“Create role”**. The new **“EMR_DefaultRole”** should now appear in the Roles listing. **Click on it** and it should take you to the following page:



EMR_DefaultRole2

Allows Elastic MapReduce to call AWS services such as EC2 on your behalf.

Summary

Creation date
February 17, 2025, 16:00 (UTC-06:00)

Last activity
-

ARN
arn:aws:iam::533267369848:role/EMR_DefaultRole2

Maximum session duration
1 hour

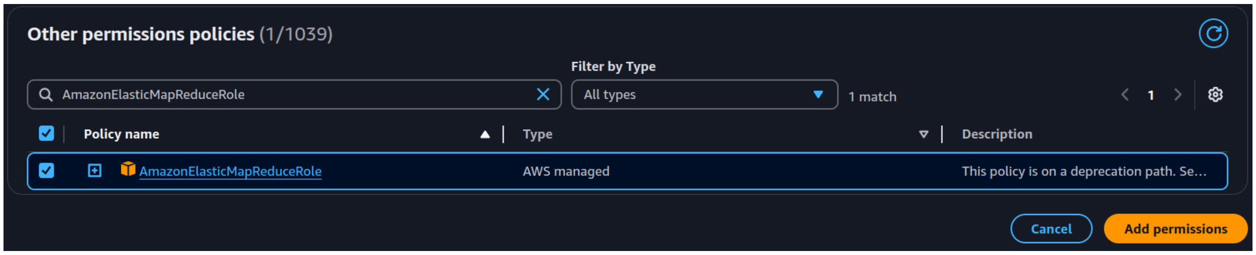
Permissions policies (1)

You can attach up to 10 managed policies.

Filter by Type

□	Policy name	Type	Attached entities
□	AmazonEMRServicePolicy_v2	AWS managed	1

vi. Click on **"Add permissions -> Attach policies"**. Then, in the next page's search bar, search for ***"AmazonElasticMapReduce Role"***, click the check box next to it and then click on **"Add permissions"**:



Other permissions policies (1/1039)

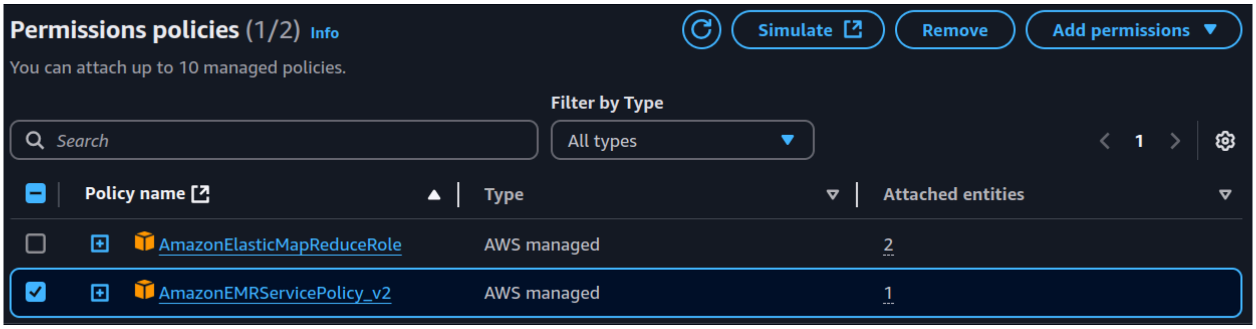
Filter by Type

AmazonElasticMapReduceRole

1 match

☑	Policy name	Type	Description
☑	AmazonElasticMapReduceRole	AWS managed	This policy is on a deprecation path. Se...

vii. Back in the EMR_DefaultRole page, remove the ***"AmazonEMRServicePolicy_v2"*** permission policy by clicking the check box next to it and then the **"Remove"** button.



Permissions policies (1/2)

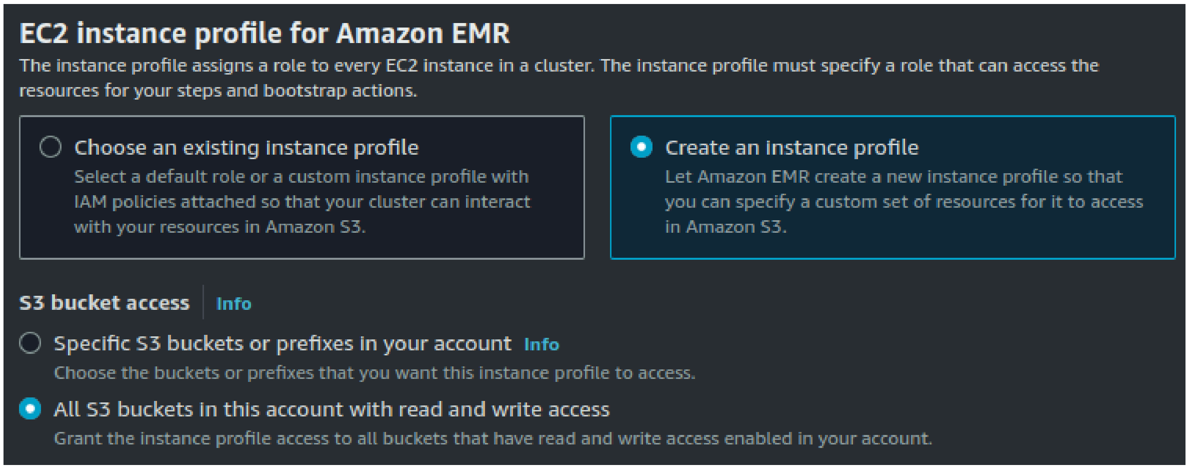
You can attach up to 10 managed policies.

Filter by Type

▬	Policy name	Type	Attached entities
□	AmazonElasticMapReduceRole	AWS managed	2
☑	AmazonEMRServicePolicy_v2	AWS managed	1

viii. You can now select the **“EMR_DefaultRole”** for your cluster. (You may need to hit the reload arrow next to the selector).

c. Under **“EC2 instance profile for Amazon EMR”**, choose **“Create an instance profile”** and give it access to all S3 buckets in your account.



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.

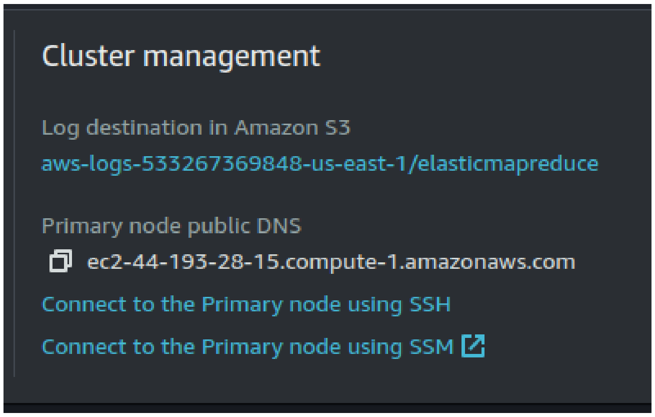
● Create an instance profile
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

S3 bucket access Info

○ Specific S3 buckets or prefixes in your account Info
Choose the buckets or prefixes that you want this instance profile to access.

● All S3 buckets in this account with read and write access
Grant the instance profile access to all buckets that have read and write access enabled in your account.

17. Finally, click on the **“Create cluster”** button on the right-hand side. It will take a few minutes for your cluster to be created and the instances to start.
    a. You will know that your cluster has finished booting up when you see the option to **“Connect to the Primary node using SSH”** under the Cluster management column.



Cluster management

Log destination in Amazon S3
aws-logs-533267369848-us-east-1/elasticmapreduce

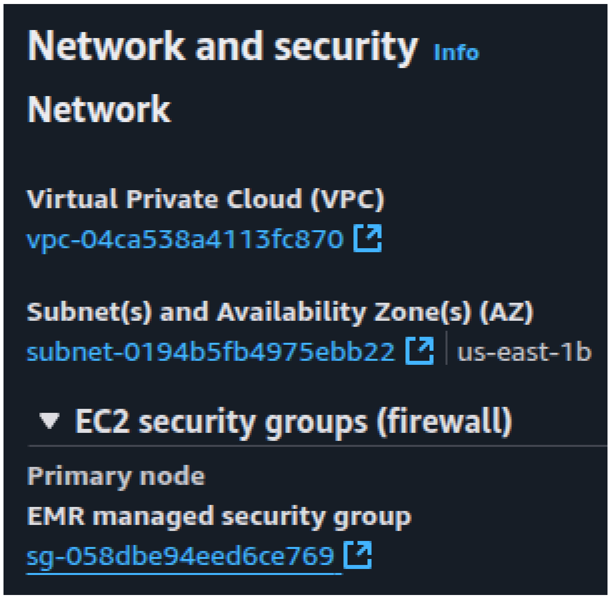
Primary node public DNS
ec2-44-193-28-15.compute-1.amazonaws.com

Connect to the Primary node using SSH

Connect to the Primary node using SSM

18. In order to connect to your cluster, you must make it so that you can connect via SSH. In your cluster’s main page, go to **“Network and Security”** and click on **“EC2 security groups (firewall)”**.

a. Next click on the link below **“Primary node, EMR managed security group”**



Network and security Info

Network

Virtual Private Cloud (VPC)
vpc-04ca538a4113fc870

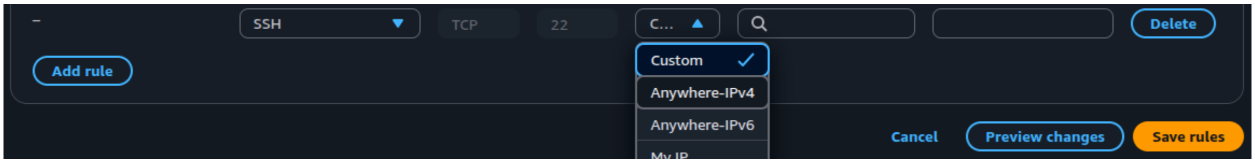
Subnet(s) and Availability Zone(s) (AZ)
subnet-0194b5fb4975ebb22 | us-east-1b

▼ EC2 security groups (firewall)

Primary node
EMR managed security group
sg-058dbe94eed6ce769

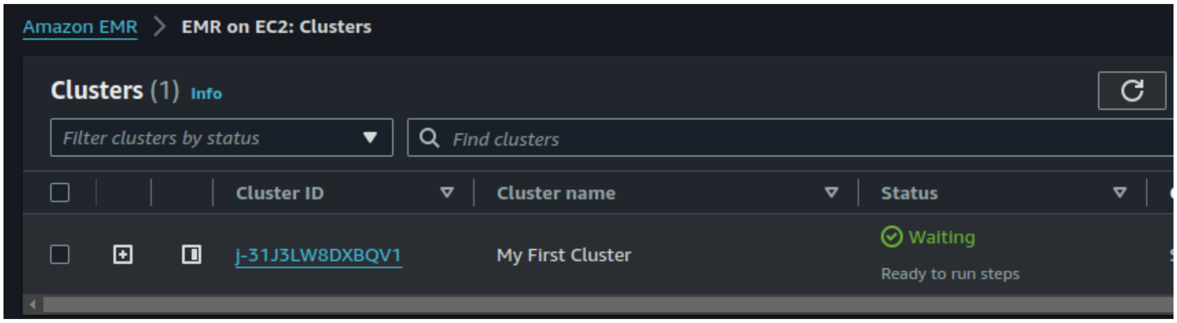
b. In the next page, click on the **“Edit inbound rules”** button located on the right-hand side of the screen.

c. Next, click on **“Add rule”**, and select **“SSH”** for the type and **“Anywhere-IPv4”** as the Source.



d. Click on **“Save Rules”** to apply the changes.

19. If you ever want to get to the page that lists all your EMR clusters, just click the **“AWS”** logo in the top left corner and then search for **“EMR”**.



Amazon EMR > EMR on EC2: Clusters

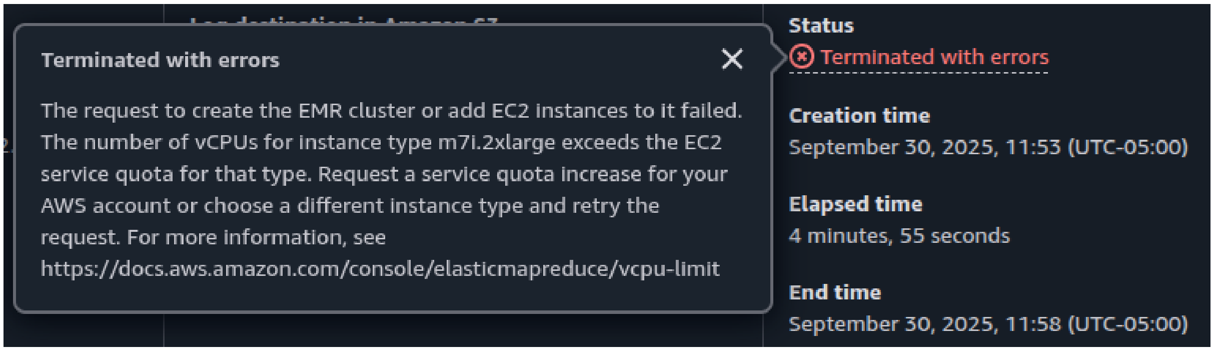
Clusters (1) Info

Filter clusters by status ▼ | Find clusters

□			Cluster ID	Cluster name	Status
□	⊞	◫	J-31J3LW8DXBQV1	My First Cluster	✓ Waiting Ready to run steps

20. **Shutdown your cluster** by selecting it and clicking on the “Terminate” button in the top right. You are ready to move onto Task 2.

21. Some of you **may encounter a not-so-uncommon issue** when attempting to create your clusters with the specifications detailed here. After a few minutes, your cluster will terminate and you will see a message like the following:



Terminated with errors

The request to create the EMR cluster or add EC2 instances to it failed. The number of vCPUs for instance type m7i.2xlarge exceeds the EC2 service quota for that type. Request a service quota increase for your AWS account or choose a different instance type and retry the request. For more information, see https://docs.aws.amazon.com/console/elasticmapreduce/vcpu-limit

Status
Terminated with errors

Creation time
September 30, 2025, 11:53 (UTC-05:00)

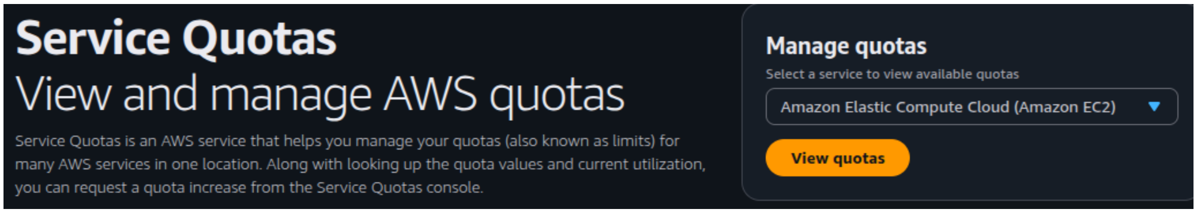
Elapsed time
4 minutes, 55 seconds

End time
September 30, 2025, 11:58 (UTC-05:00)

The issue stems from the fact that sometimes Amazon assigns new accounts with a smaller quota for standard type (A, C, D, H, I, M, R, T Z) instances.

You can check your assigned quota by using the search bar on top to search for **“Service Quotas”** which should take you to your service quotas dashboard.

Use the search box in the **“Manage quotas”** box to look for **“Amazon EC2”** and then click on **“View quotas”**.

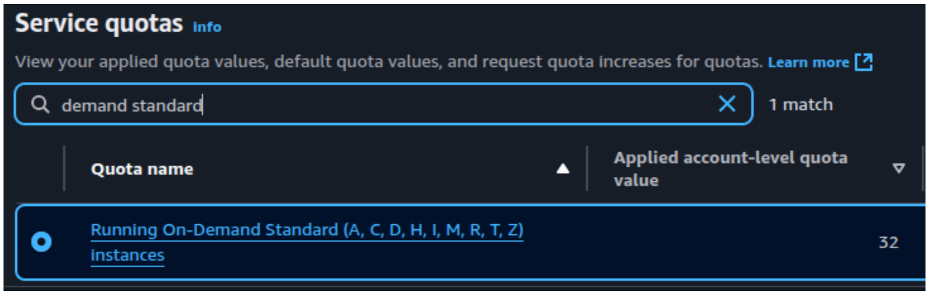


Service Quotas
View and manage AWS quotas

Service Quotas is an AWS service that helps you manage your quotas (also known as limits) for many AWS services in one location. Along with looking up the quota values and current utilization, you can request a quota increase from the Service Quotas console.

Manage quotas
Select a service to view available quotas
Amazon Elastic Compute Cloud (Amazon EC2)

In the next page, search for “demand” in the search bar and then click on **“Running On-Demand Standard (A, C, D, H, I, M, R, T, Z) instances:**



Service quotas Info
View your applied quota values, default quota values, and request quota increases for quotas. Learn more

demand standard — 1 match

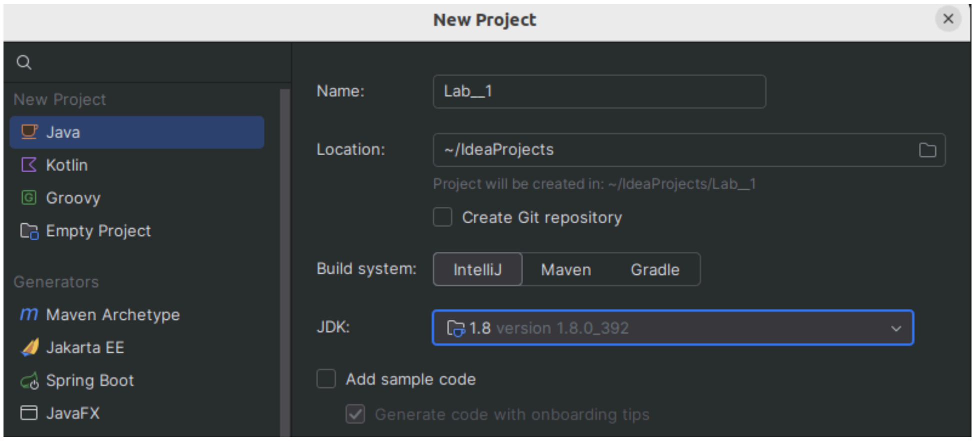
Quota name	Applied account-level quota value
● Running On-Demand Standard (A, C, D, H, I, M, R, T, Z) instances	32

If your assigned quota is less than 24, you will not be able to create the cluster with the specified requirements. You should use the **“Request increase at account level”** button in the top right to request a quota of at least 24 units.

Task 2: Compile a Hadoop Program

Note: This assumes you have installed the IntelliJ IDE in your computer. If you used DataGrip for Assignment 1, IntelliJ is also included with your student license.

1. Start by booting up IntelliJ and creating a new Java project.
    a. **Make sure you are using Java 8** (version 1.8) as other versions can have compatibility issues.



Name: Lab_1

Location: ~/IdeaProjects

Project will be created in: ~/IdeaProjects/Lab_1

□ Create Git repository

Build system: IntelliJ | Maven | Gradle

JDK: 1.8 version 1.8.0_392

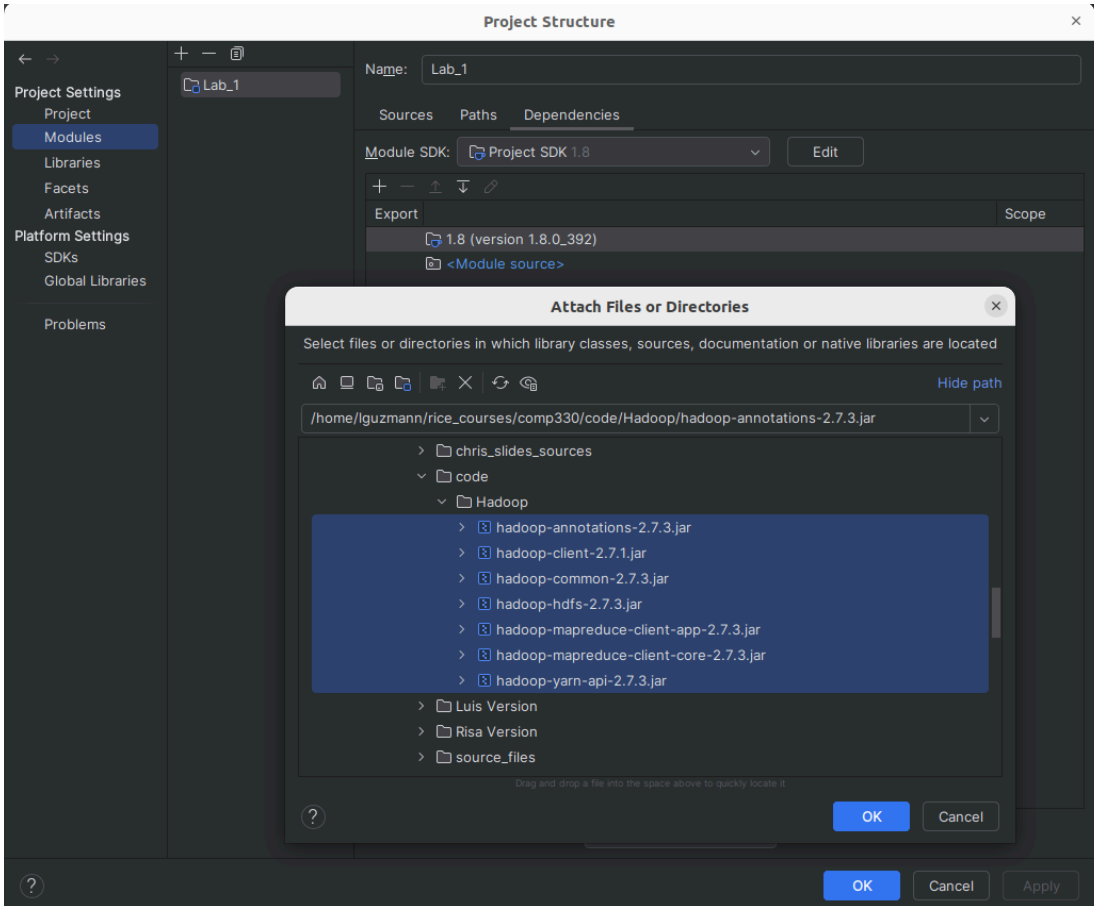
□ Add sample code

☑ Generate code with onboarding tips

    b. Note that **if you don't have Java version 1.8** you can download it directly from IntelliJ. Open a new Java project. Click on the JDK drop down menu. Then click on "Download JDK". A list of available versions will show (it may take a little bit of time to load), select version 1.8 (near the bottom). Any vendor should be fine (e.g., Amazon Coretto 1.8.0_422). Click download and finish creating the project.
    c. Here are another couple of options to download the Java 1.8 JDK: Option 1, Option 2.

2. Download and unzip all the JAR files included here (make sure there are 7 of them). Put them into a directory in your computer.

3. Add the JAR files to your project by going to **"Files" -> "Project Structure" -> "Modules" -> "Dependencies"** and clicking the **"+"** symbol
    a. Make sure that all 7 files are added, sometimes the ***"hadoop-client-2.7.1.jar"* does not get added** for some reason.
        i. If this happens then repeat the process but add the ***"hadoop-client-2.7.1.jar"*** file individually. Select **"Classes"** in the small window that pops up.
    b. Make sure you click on the **"Apply"** button.



Name: Lab_1

Module SDK: Project SDK 1.8

1.8 (version 1.8.0_392)

<Module source>

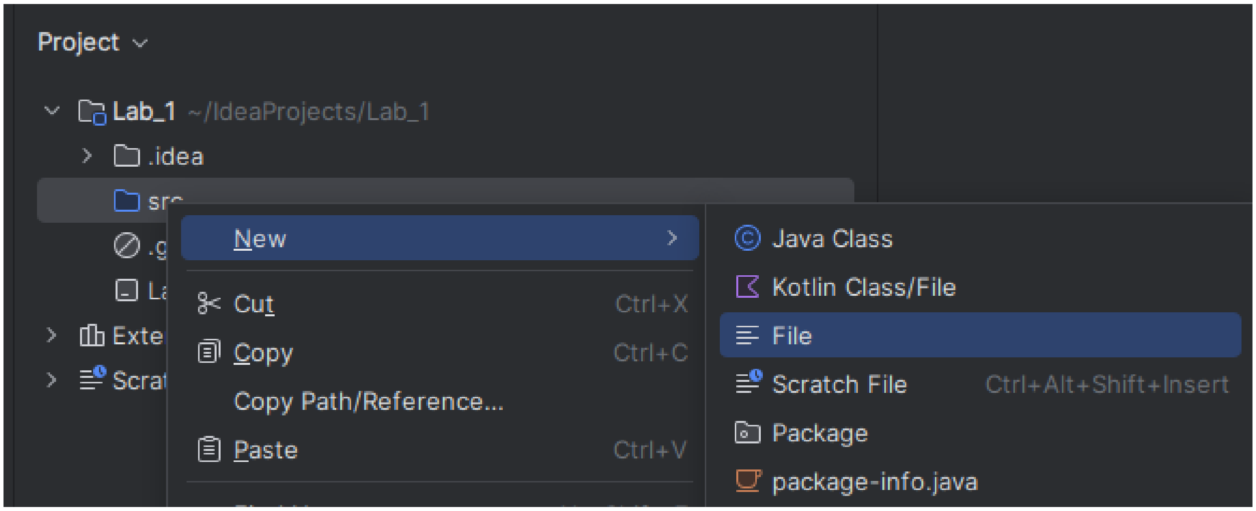
Select files or directories in which library classes, sources, documentation or native libraries are located

/home/lguzmann/rice_courses/comp330/code/Hadoop/hadoop-annotations-2.7.3.jar

```
chris_slides_sources
code
  Hadoop
    hadoop-annotations-2.7.3.jar
    hadoop-client-2.7.1.jar
    hadoop-common-2.7.3.jar
    hadoop-hdfs-2.7.3.jar
    hadoop-mapreduce-client-app-2.7.3.jar
    hadoop-mapreduce-client-core-2.7.3.jar
    hadoop-yarn-api-2.7.3.jar
Luis Version
Risa Version
source_files
```

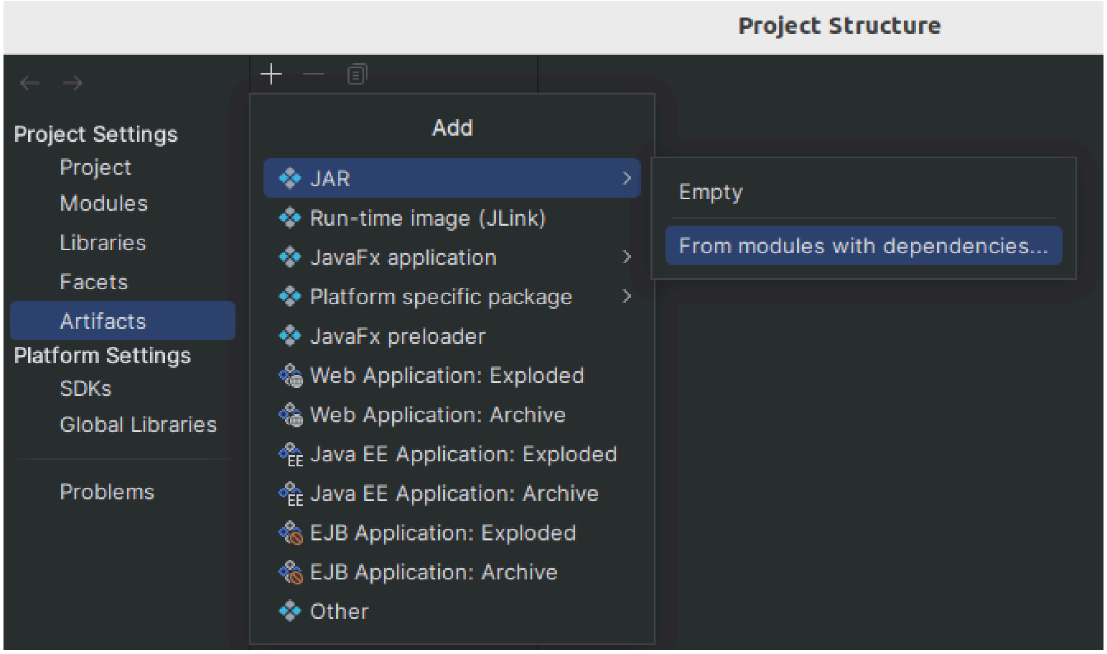
Drag and drop a file into the space above to quickly locate it

4. Next, create a new file called **WordCount.java** in the **src** folder of your project.
    a. **Capitalization is important** here so make sure that the name you use is an exact match.

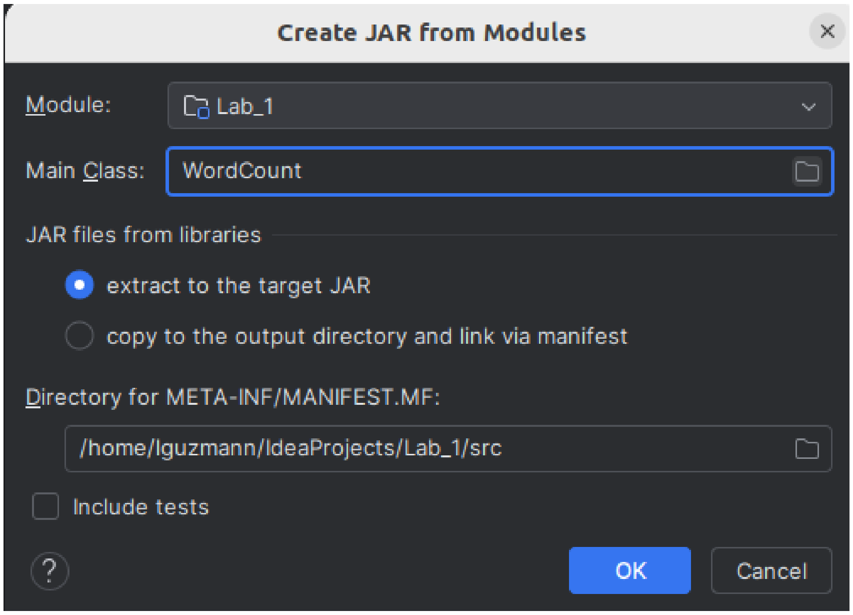


5. In the ZIP file that included all the JAR files, there was a ***WordCount.java***. Copy and paste the contents of this file into the WordCount.java file in your project.

6. Next, we will create a JAR artifact of our WordCount program that can be run in the EMR cluster.
   a. Go to **"File" -> "Project Structure" -> "Artifacts"**, click on the **"+"** and select **"JAR" -> "From modules with dependencies"**.



   b. In the next pop up, select **"WordCount"** as the main class and click **"OK"**. Click **"Apply"** and then **"OK"**.



Module: Lab_1

Main Class: WordCount

JAR files from libraries

● extract to the target JAR

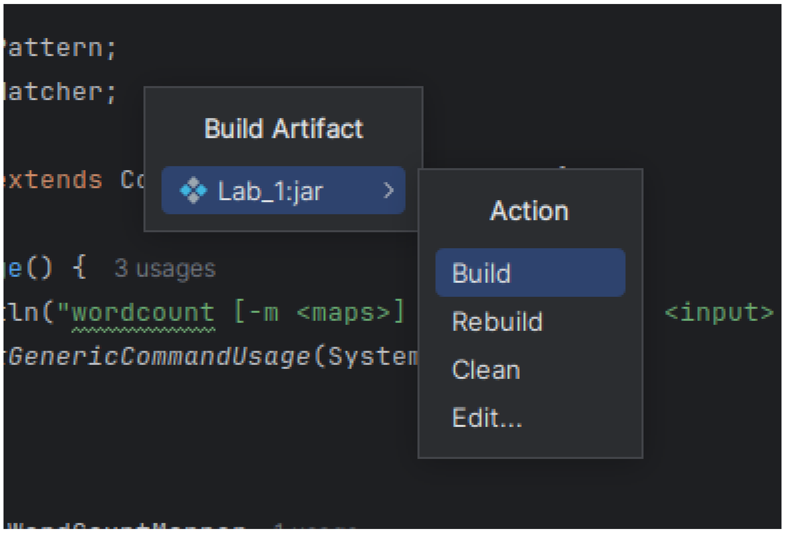
○ copy to the output directory and link via manifest

Directory for META-INF/MANIFEST.MF:

/home/lguzmann/IdeaProjects/Lab_1/src

☐ Include tests

c. Next, go to **"Build" -> "Build Artifacts"** and select **"Build"** in the pop up window.



```
attern;
atcher;

xtends Co

e() {  3 usages
ln("wordcount [-m <maps>]                              <input>
GenericCommandUsage(System
```

d. A new ***[Project Name].jar*** file will be created in the the following folder:

   i. ***[Project Name]/out/artifacts/[Project Name]***

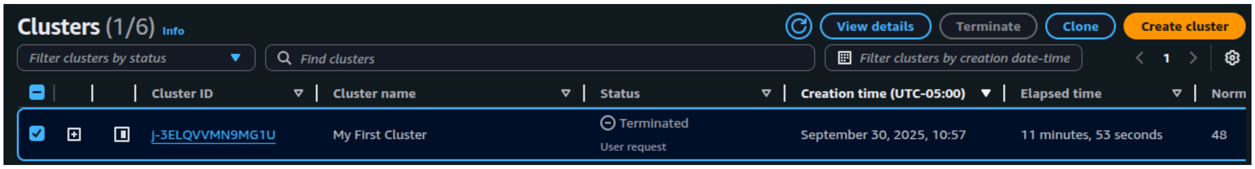
7. You are now ready to move on to Task 3.

Task 3: Connect to the EMR Cluster through SSH

Time to go back to AWS. Once your cluster is up and running, you will want to connect to the primary node so that you can run Hadoop jobs on it.

1. If you followed the instructions in Task 1, you will have terminated your cluster and won't be able to use it anymore. Fortunately, you do not need to repeat the whole process as you can easily create a new clone of your previous cluster.

   a. Go to your cluster page and clone it by selecting it and clicking on the **"Clone"** button located on the top right corner.



			Cluster ID	Cluster name	Status	Creation time (UTC-05:00)	Elapsed time	Norm
☑	⊞	◫	j-3ELQVVMN9MG1U	My First Cluster	⊖ Terminated User request	September 30, 2025, 10:57	11 minutes, 53 seconds	48

   b. Next, just click on the orange **"Clone cluster"** button on the right side of the screen and your clone will begin booting up.

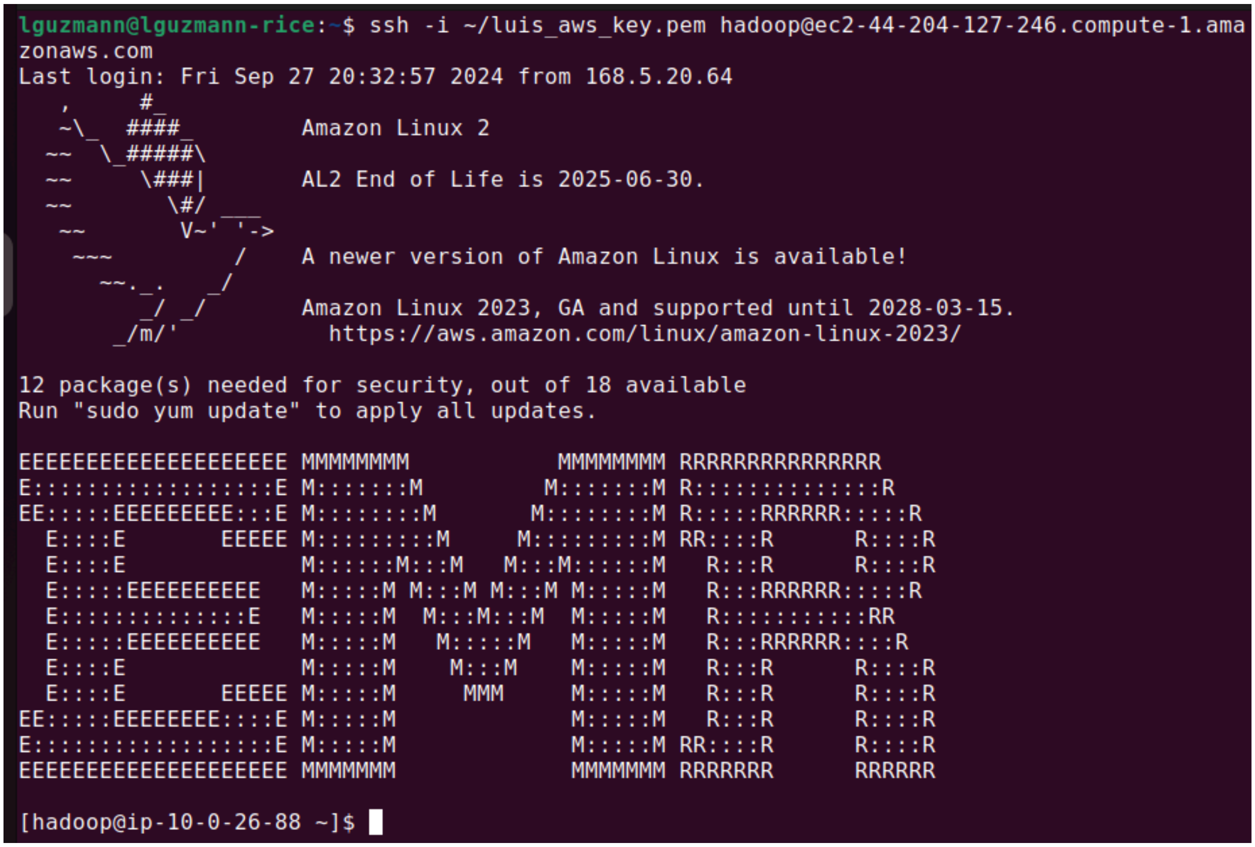
c. Remember that you won’t be able to connect into your cluster until you see the option to **“Connect to the Primary node using SSH”**.

d. Once the option appears, click on it to obtain the ssh command used to connect to your cluster.

2. The next step will be platform dependent,
    a. If you are using **MacOS/Linux**, simply open a terminal and paste the ssh command.
        i. Keep in mind that the way the command is structured by default is to assume that your key pair file (*.pem* file) is **located in your Home directory**. If your *.pem* file is not there, the command will not work. Simply, **change the relative path** used by the command to point to the location of your *.pem* file.
        ii. **IMPORTANT:** If you get an error message saying that you are using an **“Unprotected Private Key File”**, all you need to do is change the permissions on your .pem file by running the following command:

            chmod 600 [Your Key Name].pem

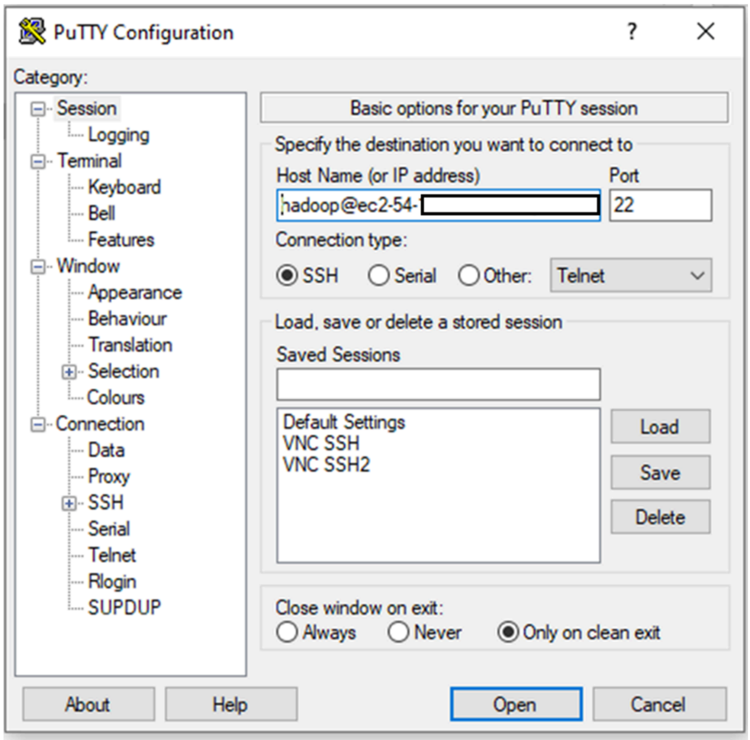
        iii. You should be able to successfully connect to your cluster



```
lguzmann@lguzmann-rice:~$ ssh -i ~/luis_aws_key.pem hadoop@ec2-44-204-127-246.compute-1.ama
zonaws.com
Last login: Fri Sep 27 20:32:57 2024 from 168.5.20.64
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

12 package(s) needed for security, out of 18 available
Run "sudo yum update" to apply all updates.

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-10-0-26-88 ~]$ █
```

b. If you are using Windows, we will assume you are using PuTTy.

   i. In the left-hand side menu, click on “Connection” -> “SSH” -> “Auth”, then click “Browse” to select the private key file (.ppk file).



Specify the destination you want to connect to

Host Name (or IP address): hadoop@ec2-54-[redacted]

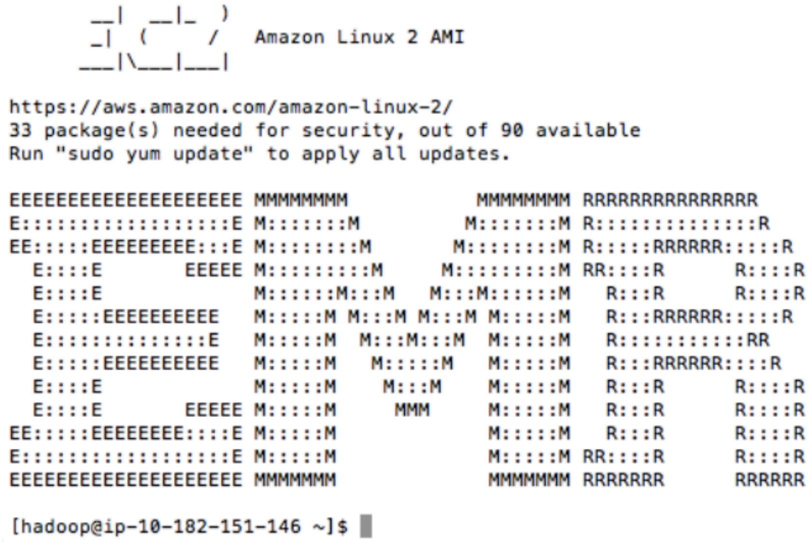
Port: 22

Connection type: ● SSH ○ Serial ○ Other: Telnet

Close window on exit: ○ Always ○ Never ● Only on clean exit

   ii. Select your key and click “Open”, then go back to the “Session” and enter your machine information (“hadoop@...”, you get this from your cluster page in AWS) and click on “Open”.

   iii. You should be able to successfully connect to your cluster.



```
       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
33 package(s) needed for security, out of 90 available
Run "sudo yum update" to apply all updates.

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-10-182-151-146 ~]$ █
```

Note on PuTTy time out:

If PuTTy is timing out. Do these steps:

1. Go to EC2 console
2. Click Instances on Left
3. Select your instance
4. In the Description tab, locate Security Groups and click the available group link for “Primary Node”
5. Click edit button on Inbound tab
6. Click Add Rule and select SSH for type, Port Range 22, and Source Anywhere

Task 4: Load Data into Hadoop’s Distributed File System (HDFS)

Next we need to transfer the JAR file we created in Task 2 to the EMR cluster.

1. Remember that your JAR file will be located in a folder

   [Project Name]/out/artifacts/[Project Name]_jar/[Project Name].jar

   So, for instance, if your project name is Lab1, the path will be

   Lab1/out/artifacts/Lab1_jar/Lab1.jar

   Note: These examples are, of course, relative paths that assume your current working directory is the folder where IntelliJ saves projects.

2. How you transfer the file **into** your cluster will be platform dependent.

   a. **If you are using MacOS/Linux** you can use either the ***scp*** or ***sftp*** commands.

      i. Using scp (Secure CoPy), it is a single-line command as follows:

scp -i [path to key] [path to jar] [cluster ip address]:

To me this looked like:

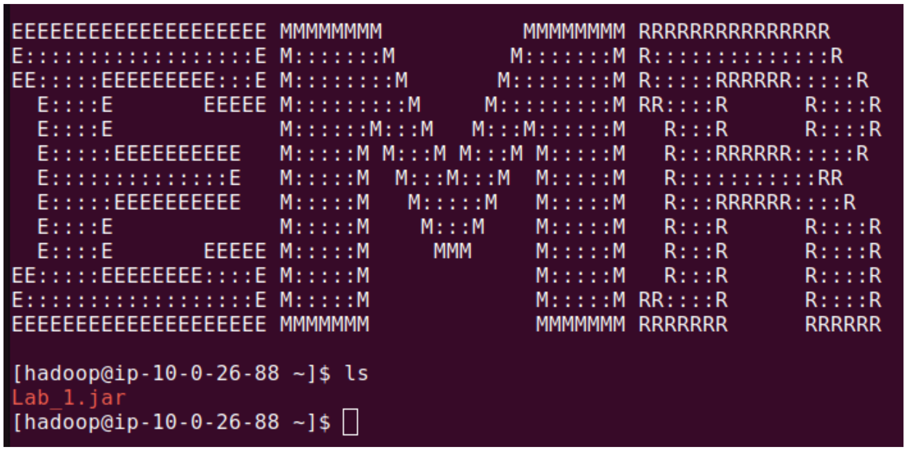
scp -i ~/luis_aws_key.pem ~/IdeaProjects/Lab_1/out/artifacts/Lab_1_jar/Lab_1.jar hadoop@ec2-44-204-127-246.compute-1.amazonaws.com:

IMPORTANT: You should execute the command above **in your local computer’s**

terminal, not in the cluster's terminal. Remember that you are trying to copy something **FROM your computer INTO your cluster**.

ii. Using sftp (Secure File-Transfer Protocol), it is a multi-step process:
   1. *sftp -i [path to key] [cluster ip address]*
   2. *put "[path to jar]"*
   3. *exit*

iii. Either way, the JAR file should be copied into your cluster's home directory. You can corroborate this with the ***ls*** command in your cluster's terminal.



```
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-10-0-26-88 ~]$ ls
Lab_1.jar
[hadoop@ip-10-0-26-88 ~]$ 
```

b. If you are using Windows, download WinSCP. Start it up and click on the "Advanced" button, then "Authentication" and browse to find your .ppk file.

   After installing WinSCP, it may import from PuTTY, click Ok in that case.

   1. In WinSCP, enter your EC2 instance public DNS name into ***Host*** *name* box.

   2. Enter ***"hadoop"*** as a user name

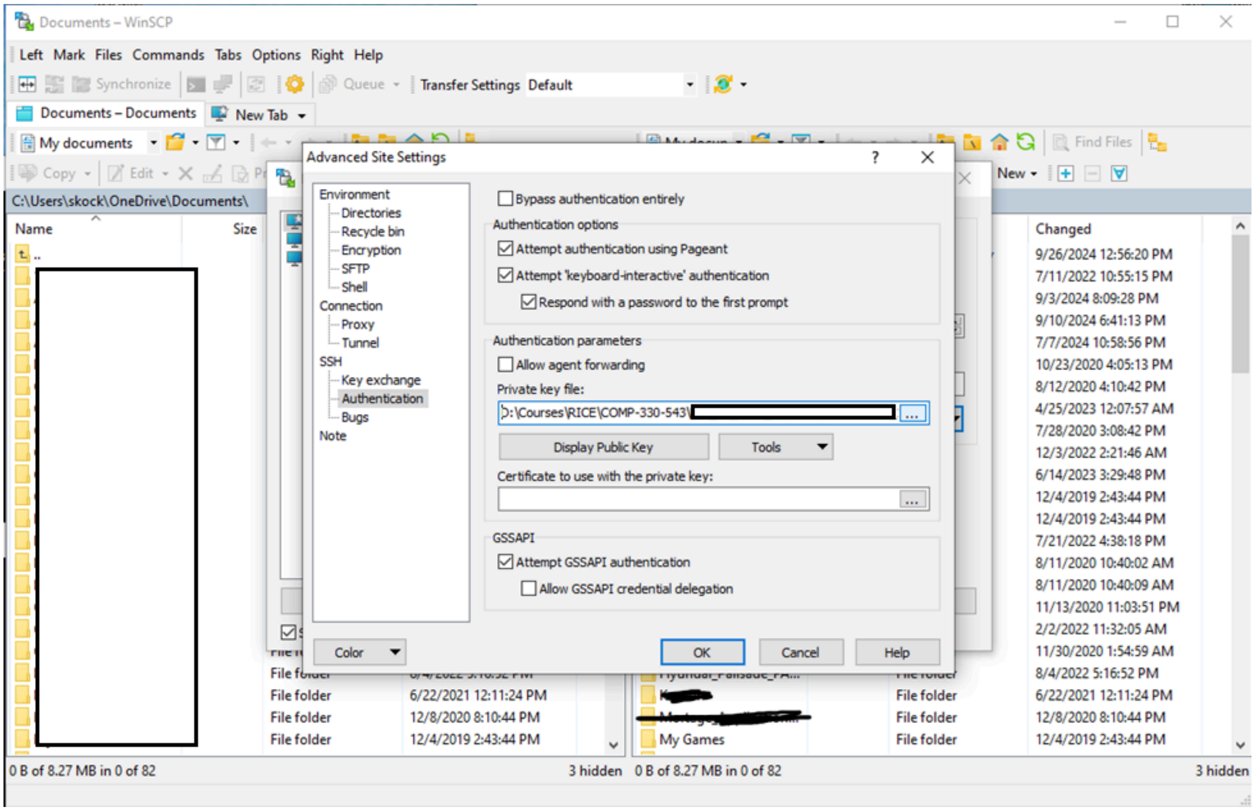
   3. Press the ***Advanced*** button to open advanced site settings dialog and go to ***SSH -> Authentication***.

   4. In the ***Private key*** file-box select your .ppk file

   5. Submit these settings by clicking the ***OK*** button

   6. Save your site settings using ***Save*** button.

/



C:\Users\skock\OneDrive\Documents\

[redacted]

☐ Bypass authentication entirely

Authentication options

☑ Attempt authentication using Pageant

☑ Attempt 'keyboard-interactive' authentication

☑ Respond with a password to the first prompt

Authentication parameters

☐ Allow agent forwarding

Private key file: D:\Courses\RICE\COMP-330-543\[redacted]

Certificate to use with the private key:

GSSAPI

☑ Attempt GSSAPI authentication

☐ Allow GSSAPI credential delegation

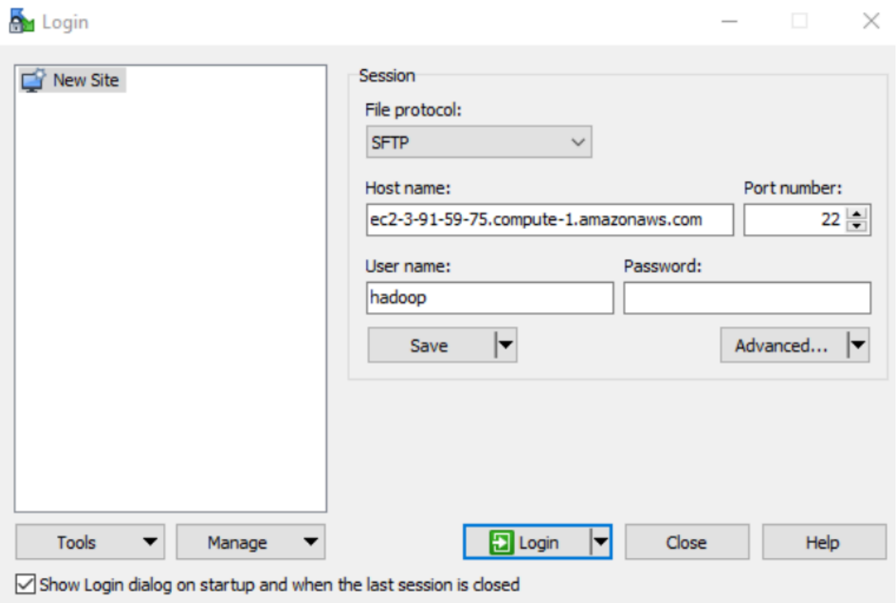
[redacted]

[redacted]

My Games

0 B of 8.27 MB in 0 of 82 3 hidden

c. Enter the primary node address, “hadoop” for the username, and press “Login”



File protocol: SFTP

Host name: ec2-3-91-59-75.compute-1.amazonaws.com

Port number: 22

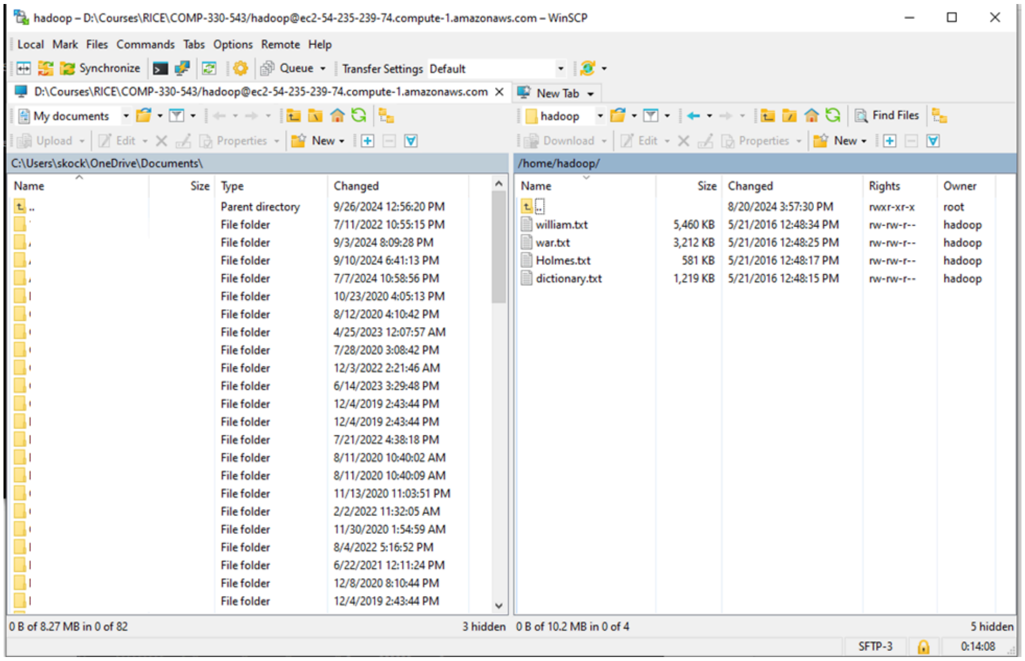
User name: hadoop

Password:

☑ Show Login dialog on startup and when the last session is closed

d. WinSCP will connect to the cluster and you can use its graphical interface to transfer files to it.

Notice that WinSCP shows the uploaded files(Task 4.e): under /home/hadoop


hadoop – D:\Courses\RICE\COMP-330-543/hadoop@ec2-54-235-239-74.compute-1.amazonaws.com – WinSCP
Local Mark Files Commands Tabs Options Remote Help
Synchronize
Queue
Transfer Settings Default
D:\Courses\RICE\COMP-330-543/hadoop@ec2-54-235-239-74.compute-1.amazonaws.com
New Tab
My documents
Upload
Edit
Properties
New
hadoop
Find Files
Download

C:\Users\skock\OneDrive\Documents\

Name	Size	Type	Changed
..		Parent directory	9/26/2024 12:56:20 PM
		File folder	7/11/2022 10:55:15 PM
		File folder	9/3/2024 8:09:28 PM
		File folder	9/10/2024 6:41:13 PM
		File folder	7/7/2024 10:58:56 PM
		File folder	10/23/2020 4:05:13 PM
		File folder	8/12/2020 4:10:42 PM
		File folder	4/25/2023 12:07:57 AM
		File folder	7/28/2020 3:08:42 PM
		File folder	12/3/2022 2:21:46 AM
		File folder	6/14/2023 3:29:48 PM
		File folder	12/4/2019 2:43:44 PM
		File folder	12/4/2019 2:43:44 PM
		File folder	7/21/2022 4:38:18 PM
		File folder	8/11/2020 10:40:02 AM
		File folder	8/11/2020 10:40:09 AM
		File folder	11/13/2020 11:03:51 PM
		File folder	2/2/2022 11:32:05 AM
		File folder	11/30/2020 1:54:59 AM
		File folder	8/4/2022 5:16:52 PM
		File folder	6/22/2021 12:11:24 PM
		File folder	12/8/2020 8:10:44 PM
		File folder	12/4/2019 2:43:44 PM

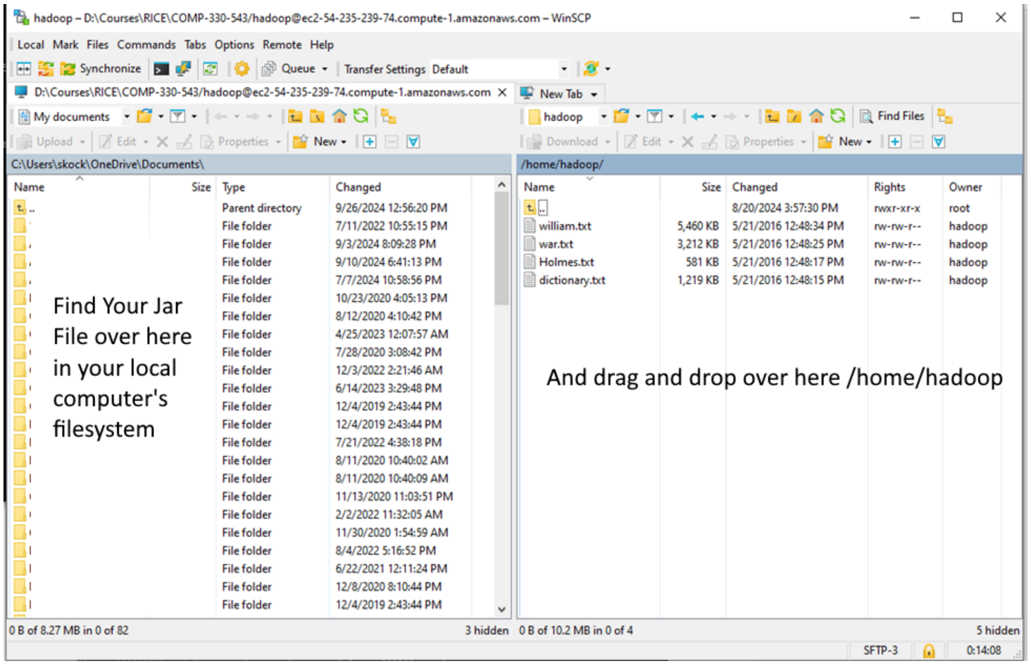
0 B of 8.27 MB in 0 of 82 3 hidden

/home/hadoop/

Name	Size	Changed	Rights	Owner
..		8/20/2024 3:57:30 PM	rwxr-xr-x	root
william.txt	5,460 KB	5/21/2016 12:48:34 PM	rw-rw-r--	hadoop
war.txt	3,212 KB	5/21/2016 12:48:25 PM	rw-rw-r--	hadoop
Holmes.txt	581 KB	5/21/2016 12:48:17 PM	rw-rw-r--	hadoop
dictionary.txt	1,219 KB	5/21/2016 12:48:15 PM	rw-rw-r--	hadoop

0 B of 10.2 MB in 0 of 4 5 hidden

SFTP-3 0:14:08


hadoop – D:\Courses\RICE\COMP-330-543/hadoop@ec2-54-235-239-74.compute-1.amazonaws.com – WinSCP
Local Mark Files Commands Tabs Options Remote Help
Synchronize
Queue
Transfer Settings Default
D:\Courses\RICE\COMP-330-543/hadoop@ec2-54-235-239-74.compute-1.amazonaws.com
New Tab
My documents
Upload
Edit
Properties
New
hadoop
Find Files
Download

Find Your Jar File over here in your local computer's filesystem

And drag and drop over here /home/hadoop

C:\Users\skock\OneDrive\Documents\

Name	Size	Type	Changed
..		Parent directory	9/26/2024 12:56:20 PM
		File folder	7/11/2022 10:55:15 PM
		File folder	9/3/2024 8:09:28 PM
		File folder	9/10/2024 6:41:13 PM
		File folder	7/7/2024 10:58:56 PM
		File folder	10/23/2020 4:05:13 PM
		File folder	8/12/2020 4:10:42 PM
		File folder	4/25/2023 12:07:57 AM
		File folder	7/28/2020 3:08:42 PM
		File folder	12/3/2022 2:21:46 AM
		File folder	6/14/2023 3:29:48 PM
		File folder	12/4/2019 2:43:44 PM
		File folder	12/4/2019 2:43:44 PM
		File folder	7/21/2022 4:38:18 PM
		File folder	8/11/2020 10:40:02 AM
		File folder	8/11/2020 10:40:09 AM
		File folder	11/13/2020 11:03:51 PM
		File folder	2/2/2022 11:32:05 AM
		File folder	11/30/2020 1:54:59 AM
		File folder	8/4/2022 5:16:52 PM
		File folder	6/22/2021 12:11:24 PM
		File folder	12/8/2020 8:10:44 PM
		File folder	12/4/2019 2:43:44 PM

0 B of 8.27 MB in 0 of 82 3 hidden

/home/hadoop/

Name	Size	Changed	Rights	Owner
..		8/20/2024 3:57:30 PM	rwxr-xr-x	root
william.txt	5,460 KB	5/21/2016 12:48:34 PM	rw-rw-r--	hadoop
war.txt	3,212 KB	5/21/2016 12:48:25 PM	rw-rw-r--	hadoop
Holmes.txt	581 KB	5/21/2016 12:48:17 PM	rw-rw-r--	hadoop
dictionary.txt	1,219 KB	5/21/2016 12:48:15 PM	rw-rw-r--	hadoop

0 B of 10.2 MB in 0 of 4 5 hidden

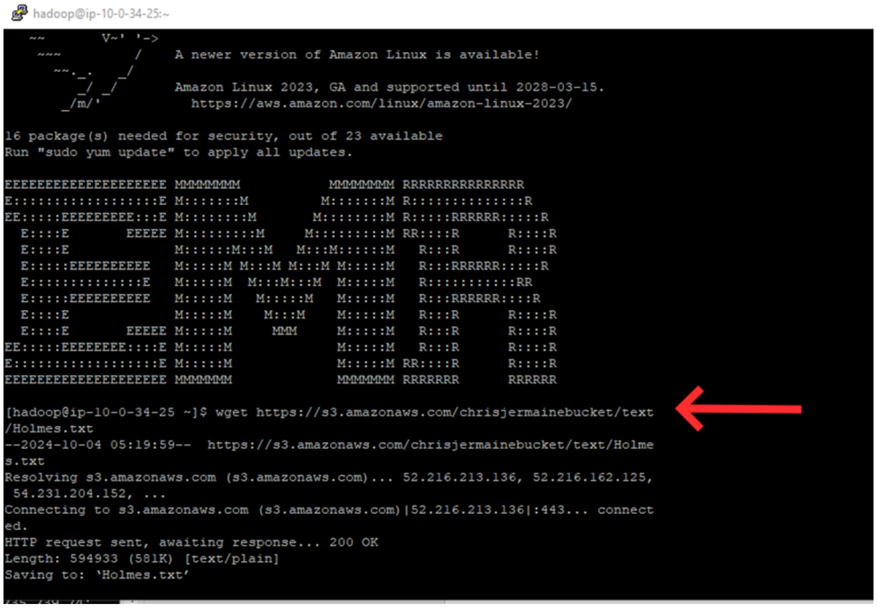
SFTP-3 0:14:08

e. Next we will load some text files into the cluster to use with our WordCount program.

   i. Simply copy and paste the following commands into your cluster terminal

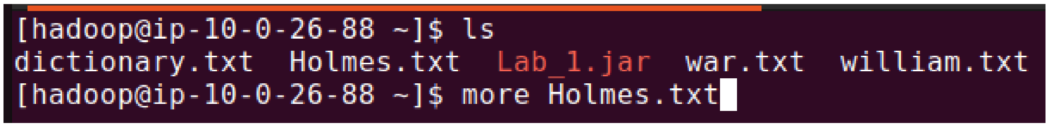
```
wget https://s3.amazonaws.com/luisguzmannateras/text/Holmes.txt
wget https://s3.amazonaws.com/luisguzmannateras/text/dictionary.txt
wget https://s3.amazonaws.com/luisguzmannateras/text/war.txt
wget https://s3.amazonaws.com/luisguzmannateras/text/william.txt
```

If you are working on Windows, right click on your PuTTY terminal to paste the commands.



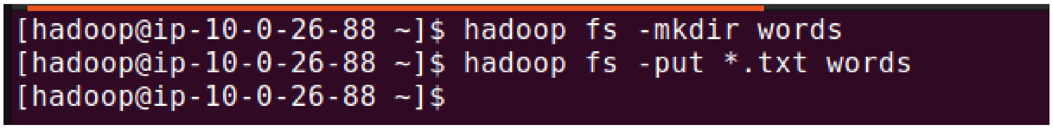
```
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

16 package(s) needed for security, out of 23 available
Run "sudo yum update" to apply all updates.

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-10-0-34-25 ~]$ wget https://s3.amazonaws.com/chrisjermainebucket/text
/Holmes.txt
--2024-10-04 05:19:59--  https://s3.amazonaws.com/chrisjermainebucket/text/Holme
s.txt
Resolving s3.amazonaws.com (s3.amazonaws.com)... 52.216.213.136, 52.216.162.125,
 54.231.204.152, ...
Connecting to s3.amazonaws.com (s3.amazonaws.com)|52.216.213.136|:443... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 594933 (581K) [text/plain]
Saving to: ‘Holmes.txt’
```

   ii. This will download four large text files into your cluster. You can look at the contents of these files by using the “more” or “less” commands (q to stop, space bar to see more).



```
[hadoop@ip-10-0-26-88 ~]$ ls
dictionary.txt  Holmes.txt  Lab_1.jar  war.txt  william.txt
[hadoop@ip-10-0-26-88 ~]$ more Holmes.txt
```

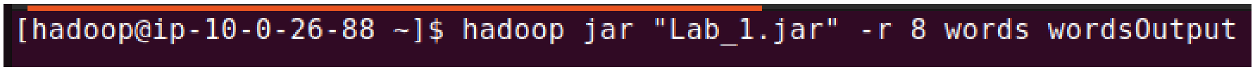
   iii. Then, we need to **load this files into the HDFS**, run the following:



```
[hadoop@ip-10-0-26-88 ~]$ hadoop fs -mkdir words
[hadoop@ip-10-0-26-88 ~]$ hadoop fs -put *.txt words
[hadoop@ip-10-0-26-88 ~]$
```

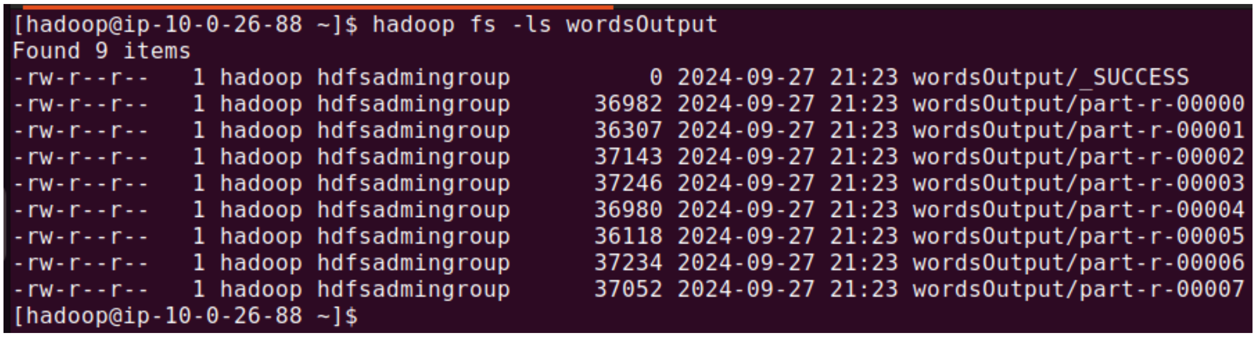
Task 5: Run a Hadoop Program in the EMR Cluster

1. At last, we can run our WordCount program on the files in the HDFS



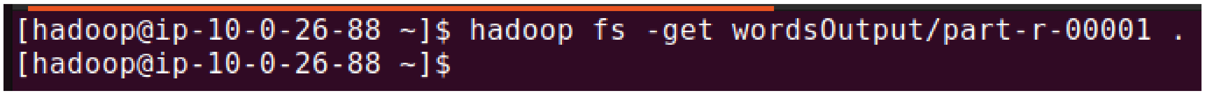
```
[hadoop@ip-10-0-26-88 ~]$ hadoop jar "Lab_1.jar" -r 8 words wordsOutput
```

2. To check the results, type:



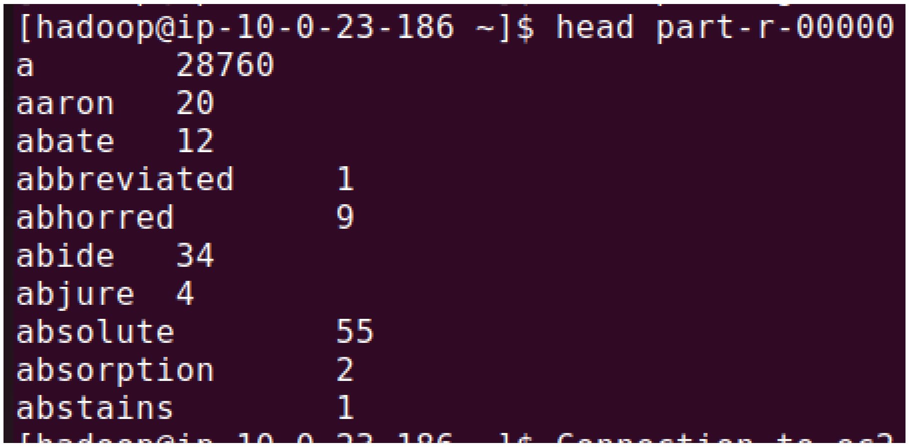
```
[hadoop@ip-10-0-26-88 ~]$ hadoop fs -ls wordsOutput
Found 9 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2024-09-27 21:23 wordsOutput/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup      36982 2024-09-27 21:23 wordsOutput/part-r-00000
-rw-r--r--   1 hadoop hdfsadmingroup      36307 2024-09-27 21:23 wordsOutput/part-r-00001
-rw-r--r--   1 hadoop hdfsadmingroup      37143 2024-09-27 21:23 wordsOutput/part-r-00002
-rw-r--r--   1 hadoop hdfsadmingroup      37246 2024-09-27 21:23 wordsOutput/part-r-00003
-rw-r--r--   1 hadoop hdfsadmingroup      36980 2024-09-27 21:23 wordsOutput/part-r-00004
-rw-r--r--   1 hadoop hdfsadmingroup      36118 2024-09-27 21:23 wordsOutput/part-r-00005
-rw-r--r--   1 hadoop hdfsadmingroup      37234 2024-09-27 21:23 wordsOutput/part-r-00006
-rw-r--r--   1 hadoop hdfsadmingroup      37052 2024-09-27 21:23 wordsOutput/part-r-00007
[hadoop@ip-10-0-26-88 ~]$
```

3. Note that it's OK if yours looks a bit different. To copy some of the results from the HDFS into the master node of your cluster, type:



```
[hadoop@ip-10-0-26-88 ~]$ hadoop fs -get wordsOutput/part-r-00001 .
[hadoop@ip-10-0-26-88 ~]$
```

4. Now, you can take a look at the counts using the *more* command again, simply type:



```
[hadoop@ip-10-0-23-186 ~]$ head part-r-00000
a       28760
aaron   20
abate   12
abbreviated     1
abhorred        9
abide   34
abjure  4
absolute        55
absorption      2
abstains        1
[hadoop@ip-10-0-23-186 ~]$ Connection to ec2
```

5. **Copy and paste some of the counts you got to Canvas** in order to get points for the Lab.

IMPORTANT TASK 6: SHUT DOWN YOUR CLUSTER

1. Never leave your cluster running when you are not using it, **YOU ARE BEING CHARGED!**

2. Remember to shut off your cluster by clicking the **"Terminate"** button in your EMR page.

3. After you kill them, **make sure** that they are dead. Go to **"All services" -> "EC2"** and look for **"Instances (running)"**. There should not be any. If they are still there, click on **"Running Instances"**. Then click the checkbox next to each of your machines, and under **"Actions" -> "Instance State"** choose **"Terminate"**. Only log out after you have verified from the EC2 page that you have no running instance.

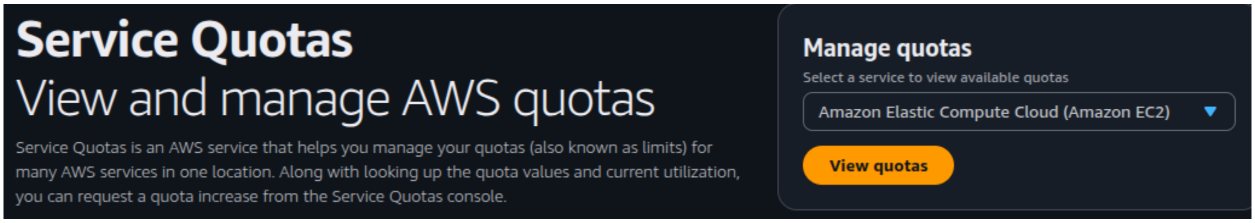
4. It is a good idea to set a termination option under **"Cluster termination and node replacement"**. You can set your cluster to terminate after a specific amount of idle time. However, if you are connected through SSH into your cluster, it won't be considered idle and will continue to run.

Extra Task: Request Quota Increase for G-type Instances

For Assignment 6, we will need to make use of G-type EC2 instances that have GPUs available. However, chances are you are going to have to request for a quota increase for this type of instance because new accounts usually are assigned a default quota of 0.

First, log in to your AWS console and use the search bar on top to search for **"Service Quotas"** which should take you to your service quotas dashboard.

Use the search box in the **"Manage quotas"** box to look for **"Amazon EC2"** and then click on "View quotas".



Service Quotas

View and manage AWS quotas

Service Quotas is an AWS service that helps you manage your quotas (also known as limits) for many AWS services in one location. Along with looking up the quota values and current utilization, you can request a quota increase from the Service Quotas console.

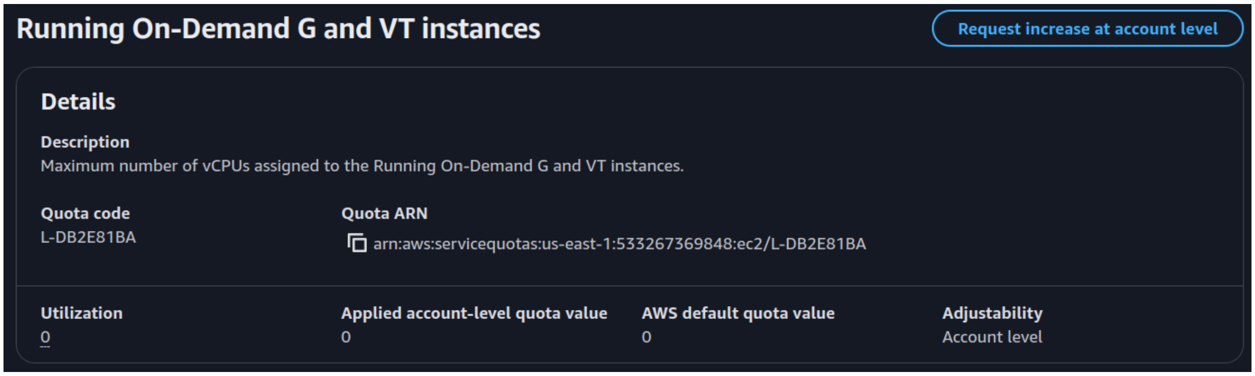
Manage quotas

Select a service to view available quotas

Amazon Elastic Compute Cloud (Amazon EC2)

In the next page, type "G and VT" in the search bar and then click on **"Running On-Demand G and VT instances"** which should take you to a separate page where you can see your current allotted quota for G-type instances.

Important Note: Double check that you do **NOT** ask for a quota increase for **SPOT** instances. It should be for **ON-DEMAND** instances.



Running On-Demand G and VT instances

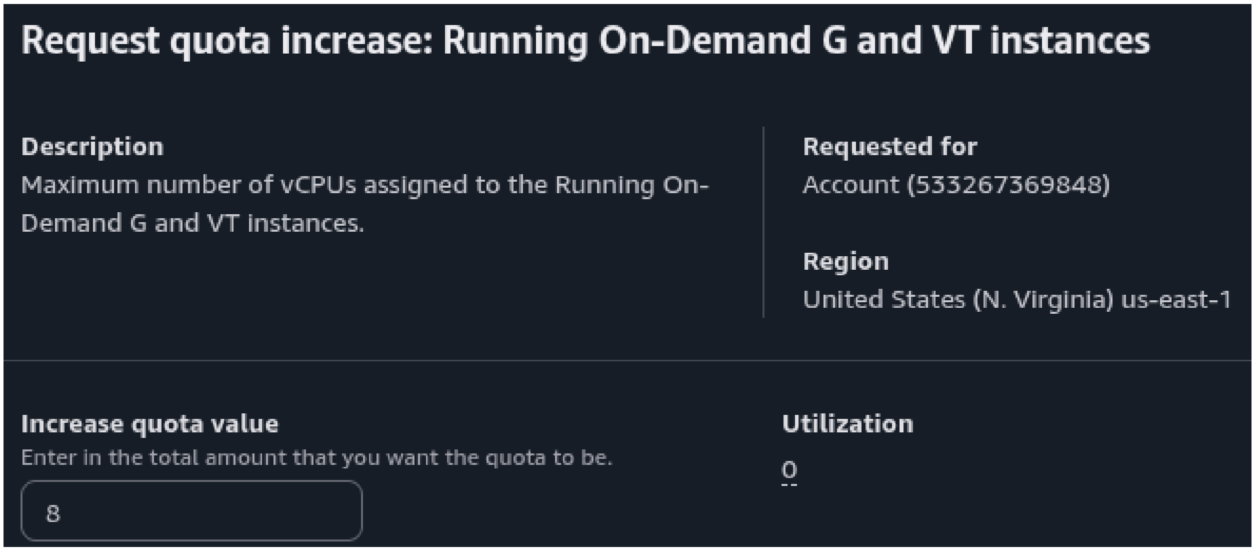
Details

Description
Maximum number of vCPUs assigned to the Running On-Demand G and VT instances.

Quota code	Quota ARN
L-DB2E81BA	arn:aws:servicequotas:us-east-1:533267369848:ec2/L-DB2E81BA

Utilization	Applied account-level quota value	AWS default quota value	Adjustability
0	0	0	Account level

Your current quota will likely be zero and you will have to use the **“Request increase at account level”** button to make a request. Ask for a quota of 8 and click on Request.



Request quota increase: Running On-Demand G and VT instances

Description
Maximum number of vCPUs assigned to the Running On-Demand G and VT instances.

Requested for
Account (533267369848)

Region
United States (N. Virginia) us-east-1

Increase quota value
Enter in the total amount that you want the quota to be.
8

Utilization
0

Most of the time, quota-increase requests are approved fairly quickly (2~3 hrs) and without incident.

However, every once in a while Amazon will deny your request and you will have to talk to their customer support in order to get it approved. This process can be slow and take a few interactions with them. As such, we recommend you submit this request early so that you may be already approved by the time Assignment 6 comes along.