

Lab 2 Hadoop - AWS EMR

In this lab, you will create a simple Amazon EMR cluster in the AWS Management Console. After you create the cluster, you will submit a Hive script to process sample data stored in Amazon Simple Storage Service (Amazon S3). Please note that the cluster should be terminated after the lab!

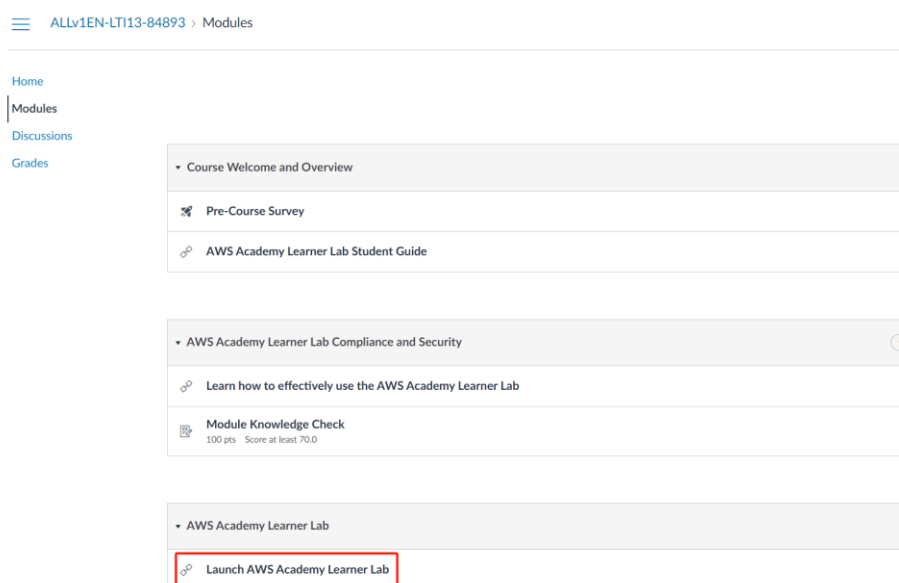
Task 1 Launch an Amazon EMR cluster

Task 1.1 Login AWS Academy account

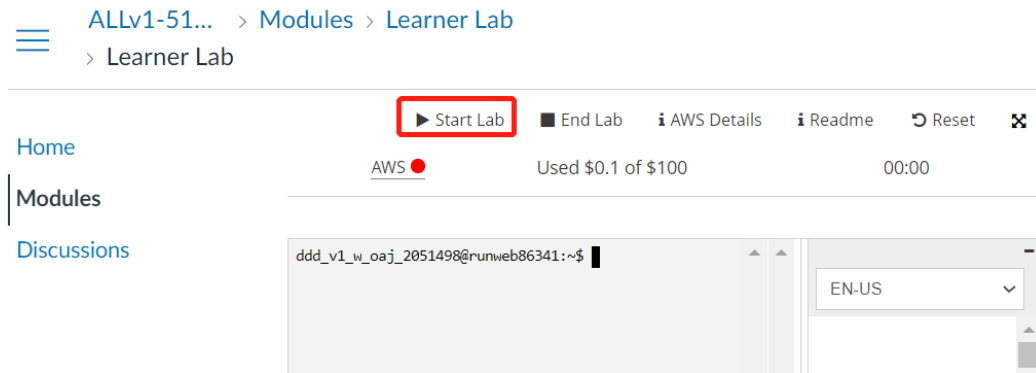
- If you haven't completed AWS Academy account registration.
- You should receive an email from AWS, and please follow the instructions. You may also refer to the Lab 1.
- Open Amazon Academy Canvas: <https://awsacademy.instructure.com/>
- Please Login as a Student.



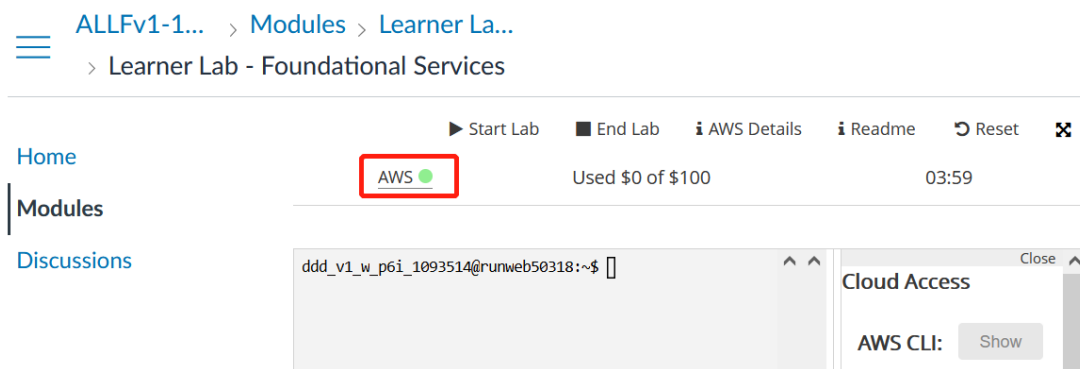
- Please go to the course: AWS Academy Learner Lab
- Under the Modules, please select Learner Lab



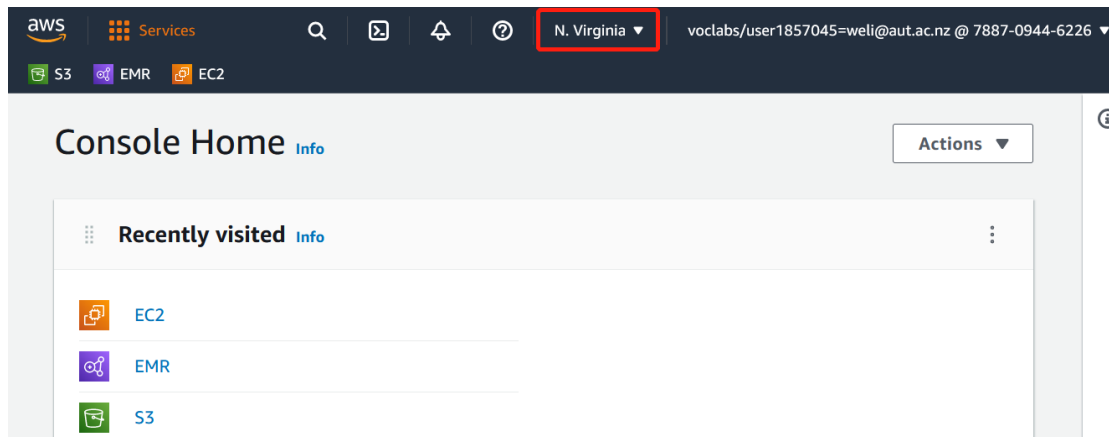
- Click Start Lab



- When you find the red dot becomes green, it means the AWS Services are ready to use.



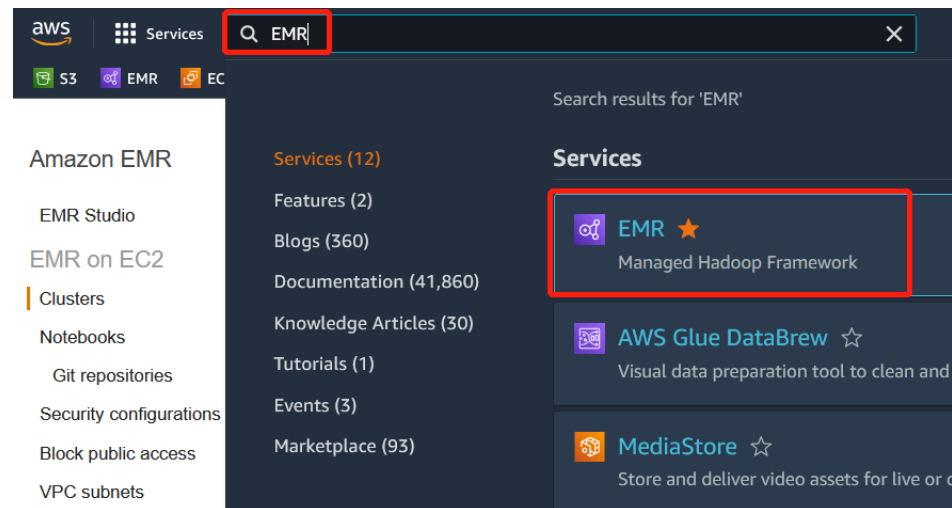
- Click AWS next to the green dot. You will see the screen below. You are not supposed to change the region, because the services for educate account are only supported in the *N. Virginia region*. You may find the connection a bit slow but no choice.



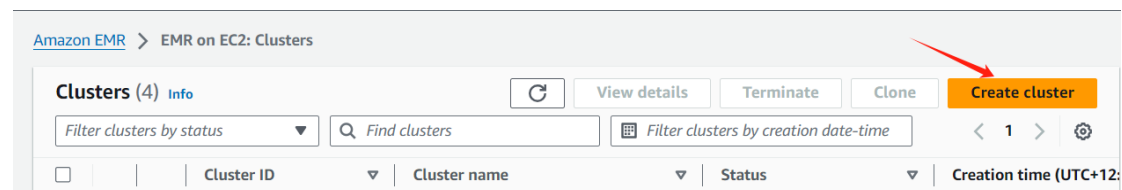
Task 1.2 Launch Your First Amazon EMR Cluster

In this task, you launch your first Amazon EMR cluster by using Quick Options in the Amazon EMR console and leaving most options to their default values. To learn more about these options, see [Summary of Quick Options](#) after the procedure. You can also select Go to advanced options to explore the additional configuration options available for a cluster.

Find EMR from the service list and click it.



Choose Create cluster.



Follow the below steps to give the options and create a cluster.

▼ **Name and applications - required** [Info](#)

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

Name

My cluster


Amazon EMR release [Info](#)

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


emr-7.1.0 ▼

Application bundle


Spark
Interactive




Core
Hadoop




Flink




HBase




Presto



Trino



Custom



☐ AmazonCloudWatchAgent
1.300032.2

☒ HCatalog 3.1.3

☒ Hue 4.11.0

☐ Livy 0.8.0

☐ Phoenix 5.1.3

☐ Spark 3.5.0

☒ Tez 0.10.2

☐ ZooKeeper 3.9.1

☐ Flink 1.18.1

☒ Hadoop 3.3.6

☐ JupyterEnterpriseGateway 2.6.0

☐ MXNet 1.9.1

☒ Pig 0.17.0

☐ Sqoop 1.4.7

☐ Trino 435

☐ HBase 2.4.17

☒ Hive 3.1.3

☐ JupyterHub 1.5.0

☐ Oozie 5.2.1

☐ Presto 0.284

☐ TensorFlow 2.11.0

☐ Zeppelin 0.10.1

AWS Glue Data Catalog settings

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

☐ Use for Hive table metadata

Operating system options [Info](#)

☒ Amazon Linux release

☐ Custom Amazon Machine Image (AMI)

☒ Automatically apply latest Amazon Linux updates

▼ Cluster configuration - *required* [Info](#)

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

☒ 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](#) [↗](#)

☐ Flexible instance fleets

Choose from the widest variety of provisioning options for the EC2 instances in your cluster. Diversify instance types and purchasing options, and use an allocation strategy. [Learn more](#) [↗](#)

Uniform instance groups

Primary

Choose EC2 instance type

m4.large

2 vCore 8 GiB memory EBS only storage

On-Demand price: - Lowest Spot price: -

Actions ▼

☐ 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: -

Actions ▼

► Node configuration - *optional*

Task 1 of 1

Remove instance group

Name

Task - 1

Choose EC2 instance type

m4.large

2 vCore 8 GiB memory EBS only storage

On-Demand price: - Lowest Spot price: -

Actions ▼

► Node configuration - *optional*

Cluster scaling and provisioning option [Info](#)

Amazon EMR console only supports EMR-managed scaling. To create a cluster with auto-scaling, use CLI or SDK.

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.

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	Size		Use Spot purchasing option
Core	m5.xlarge	<input type="text" value="1"/>	instance(s)	<input type="checkbox"/>
Task - 1	m5.xlarge	<input type="text" value="1"/>	instance(s)	<input type="checkbox"/>

Let's change the cluster termination time from 1 hour to 3 hours.

▼ Cluster termination and node replacement [Info](#)

Choose termination settings and protect your cluster from accidental shutdown.

Termination option

- ☐ Manually terminate cluster
- ☐ Automatically terminate cluster after last step ends
- ☒ Automatically terminate cluster after idle time (Recommended)

Idle time

Enter the time until your cluster terminates.

0 days ▼

Choose a time that is greater than 1 minute (00:01:00) and less than 7 days. The time is in hh:mm:ss (24-hour) format.

☐ Use termination protection

Protects your cluster from accidental termination. If on, you must first turn off protection to terminate the cluster. We recommend turning on termination protection for your long running clusters.

Unhealthy node replacement - *new* [Info](#)

☒ Turn on

Amazon EMR gracefully stops processes on unhealthy nodes to minimize data loss and job interruptions. It quickly replaces unhealthy nodes with new EC2 instances to keep your jobs running smoothly.

☐ Turn off

Amazon EMR adds unhealthy nodes to a denylist while keeping them in the cluster, allowing you continued access for troubleshooting.

▼ **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
 ▼ ↻

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
 ▼ ↻

After that, choose “Create cluster”, and you will find your cluster will be running after a few minutes. The cluster status page with the cluster Summary appears. You can use this page to monitor the progress of cluster creation and view details about cluster status. As cluster creation tasks finish, items on the status page update. You may need to choose the refresh icon on the right or refresh your browser to receive updates.

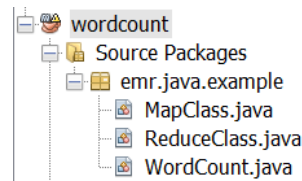
Click the cluster ID of your new cluster, and you will see its summary and settings.

Task 2 Elastic MapReduce Hadoop Job Using Custom Jar

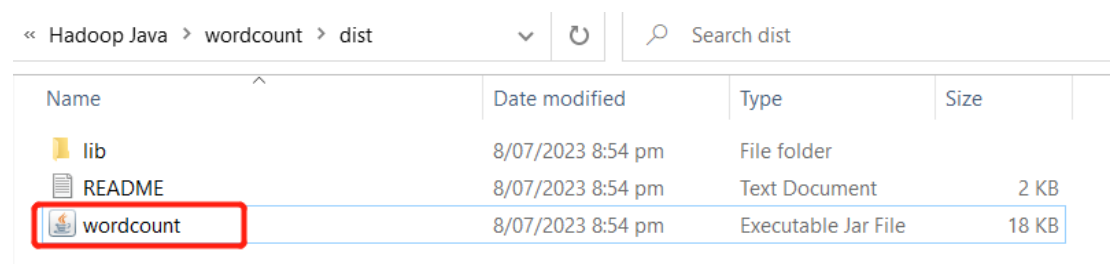
Download *input.txt* file, and *wordcount.jar* from Canvas, or prepare your own text file to be processed.

Task 2.1 Prepare Java map-reduce program (Optional)

Download *wordcount* project from Canvas, unzip it and then open this project using NetBeans IDE.



Walkthrough all the three classes and compile the project. *wordcount.jar* is a compiled version of this project.



« Hadoop Java > wordcount > dist

Name	Date modified	Type	Size
lib	8/07/2023 8:54 pm	File folder	
README	8/07/2023 8:54 pm	Text Document	2 KB
wordcount	8/07/2023 8:54 pm	Executable Jar File	18 KB

If you want to simulate MapReduce without the assistance of AWS EMR, please try *WordCount.java* under the *emr.simulation.wordcount* package.

Task 2.2 Create Amazon S3 Bucket

In this task, you need to specify an Amazon S3 bucket and create two folders to store the input data (files to be processed by your Java program) and your Java program – the jar file.

To access Amazon S3, please search S3 and click S3 from the result list.



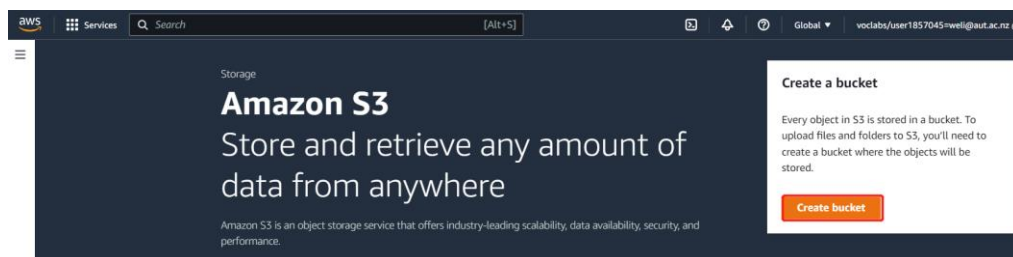
Because of Hadoop requirements, bucket and folder names that you use with Amazon EMR have the following limitations:

- They must contain only letters, numbers, periods (.), and hyphens (-).
- They cannot end in numbers.
- Bucket names must be unique *across all AWS accounts*.

For more information about creating a bucket, see [Create a Bucket](#) in the *Amazon Simple Storage Service Getting Started Guide*. After you create the bucket, choose it from the list and then choose Create folder, replace the New folder with a name that meets the requirements, and then choose Save.

The bucket and folder name used later in the tutorial is `s3://bigdatabucket2024/input` and `s3://bigdatabucket2024/jar`

Yours will be different, which means you cannot use bigdatabucket2024 if it has already been used by others.



Amazon S3 > Buckets > Create bucket

Create bucket [Info](#)

Buckets are containers for data stored in S3.

General configuration

AWS Region
US East (N. Virginia) us-east-1

Bucket type [Info](#)

☒ **General purpose**
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

☐ **Directory - New**
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name [Info](#)

Bucket name must be unique within the global namespace and follow the bucket naming rules. [See rules for bucket naming](#)

Copy settings from existing bucket - *optional*
Only the bucket settings in the following configuration are copied.

Format: s3://bucket/prefix

Use the default settings for the rest of the options.

General purpose buckets (2) Info All AWS Regions

Buckets are containers for data stored in S3.

Find buckets by name

	Name	AWS Region	IAM Access Analyzer	Creation date
<input type="radio"/>	aws-logs-731824951781-us-east-1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	June 11, 2024, 21:15:26 (UTC+12:00)
<input type="radio"/>	bigdatabucket2024	US East (N. Virginia) us-east-1	View analyzer for us-east-1	June 11, 2024, 21:34:39 (UTC+12:00)

Then, create two folders under the bucket: input and jar. Next, upload the *input.txt* to the input folder and *wordcount.jar* to the jar folder.

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	input/	Folder	-	-	-
<input type="checkbox"/>	jar/	Folder	-	-	-

Task 2.3 Execute the Java program using Map-reduce

Go back to the EMR service, select your cluster, and click “add step” under the Steps tab.

Amazon EMR > EMR on EC2: Clusters > My cluster

Updated less than a minute ago

My cluster

Summary

Cluster info	Applications	Cluster management	Status and time
<p>Cluster ID</p> <p>j-201JHIST308PX</p> <p>Cluster configuration</p> <p>Instance groups</p> <p>Capacity</p> <p>1 Primary 1 Core 1 Task</p>	<p>Amazon EMR version</p> <p>emr-7.1.0</p> <p>Installed applications</p> <p>Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, Pig 0.17.0, Tez 0.10.2</p>	<p>Log destination in Amazon S3</p> <p>aws-logs-731824951781-us-east-1/elasticmapreduce</p> <p>Persistent application UIs</p> <p>YARN timeline server</p> <p>Tez UI</p> <p>Primary node public DNS</p> <p>ec2-3-237-34-51.compute-1.amazonaws.com</p> <p>Connect to the Primary node using SSH</p> <p>Connect to the Primary node using SSM</p>	<p>Status</p> <p>Waiting</p> <p>Creation time</p> <p>June 11, 2024, 21:15 (UTC+12:00)</p> <p>Elapsed time</p> <p>26 minutes, 38 seconds</p>

Properties | Bootstrap actions | Instances (Hardware) | **Steps** | Applications | Configurations | Monitoring | Events | Tags (0)

Steps (0) Info

Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

Concurrent steps: 1

Filter steps by status

Find steps

Step ID	Status	Name	Log files	Creation time (UTC+12:00)
No matches				
We can't find a match				

Choose Custom JAR as the step type, and give a name for this step, e.g., wordcount. Specify the jar location by selecting it from your S3 directory.

Add step Info

Step settings

Type

☒ Custom JAR

Adds a step that enables you to write a custom script to process your data using the Java programming language.

☐ Streaming program

Adds a step that uses standard input to run mapper/reducer scripts and send results to standard output.

☐ Hive program

Adds a step that submits a Hive script for data warehouse interactions.

☐ Pig program

Adds a step that submits a Pig script for analyzing very large data sets.

☐ Shell script

Troubleshoot your cluster.

Name

Word count

JAR location

The JAR location may be a path into S3 or a fully qualified java class in the classpath.

s3://bigdatabucket2024/jar/wordcount.jar

View

Browse S3

Choose Amazon S3 location

S3 buckets > bigdatabucket2024 > jar

Objects (1/1)

Find objects

Key

wordcount.jar

Cancel

Choose

The arguments are very important. In the first line, place the main function of the program, which is `emr.java.example.WordCount`. Place the input and output folder to the second line and third line, respectively. *Please make sure that the "output" folder does NOT exist!*

In this example, we use the following argument:

```
emr.java.example.WordCount
s3://bigdatabucket2024/input/
s3://bigdatabucket2024/output
```

** replace the input and output augment with the folders that you have created.*

JAR location

The JAR location may be a path into S3 or a fully qualified java class in the classpath.

[View](#)

[Browse S3](#)

Arguments - optional [Info](#)

These are passed to the main function in the JAR. If the JAR does not specify a main class in its manifest file, you can specify another class name as the first argument.

```
emr.java.example.WordCount
s3://bigdatabucket2024/input/
s3://bigdatabucket2024/output
```

Click Add step, and you will see the status of the step *pending*. Give it around 30 seconds and refresh it. The status will be changed to *completed*.

Steps (1) [Info](#) [Refresh table](#) [Cancel steps](#) [Clone step](#) [Add step](#)

Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

Concurrent steps: 1 [↗](#)

[Filter steps by status](#) [< 1 >](#) [⚙](#)

<input type="checkbox"/>	Step ID	Status	Name	Log files ↗
<input type="checkbox"/>	s-1026155286WTL7F3UM6A	Pending	Word count	No logs created yet ↻

<input type="checkbox"/>	Step ID	Status	Name	Log files ↗
<input type="checkbox"/>	s-1026155286WTL7F3UM6A	Completed	Word count	No logs created yet ↻

Jar location
s3://bigdatabucket2024/jar/wordcount.jar

Action on failure
Continue

Permissions
-

Argument
[↗](#) emr.java.example.WordCount s3://bigdatabucket2024/input/ s3://bigdatabucket2024/output

Main class
-

When you return to your S3, you should be able to find the output folder. Open the output folder and explore the file's content (result).

Objects (3) [Info](#) [↻](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#)

[Create folder](#) [Upload](#)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[< 1 >](#) [⚙](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	input/	Folder	-	-	-
<input type="checkbox"/>	jar/	Folder	-	-	-
<input type="checkbox"/>	output/	Folder	-	-	-

Objects (4) [Info](#)

Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	_SUCCESS	-	June 11, 2024, 21:59:54 (UTC+12:00)	0 B	Standard
<input type="checkbox"/>	part-r-00000	-	June 11, 2024, 21:59:48 (UTC+12:00)	838.0 B	Standard
<input type="checkbox"/>	part-r-00001	-	June 11, 2024, 21:59:52 (UTC+12:00)	866.0 B	Standard
<input type="checkbox"/>	part-r-00002	-	June 11, 2024, 21:59:53 (UTC+12:00)	835.0 B	Standard

Check out the text of the file – part-r00000. If you create multiple nodes, you will have multiple outputs from different nodes. Here is the output from the text file:

```

part-r-00000
File Edit View
(HDFS) 1
(introduced 2
All 1
HDFS 1
Hive, 1
Java2
Oozie, 1
Pig, 1
Sqoop, 1
YARN 1
ZooKeeper, 1
allows 1
also 2
architecture 1
as 3
automatically 1
bandwidth 1
Ln 1, Col 1 836 characters 100% Unix (LF) UTF-8

```

Task 2.4 Terminate your cluster

If you don't want any extra costs incurred, never forget to terminate your cluster after using it!

Amazon EMR > EMR on EC2: Clusters

Clusters (1/1) [Info](#)

Filter clusters by status Find clusters Filter clusters by creation date-time

<input checked="" type="checkbox"/>	<input type="checkbox"/>	Cluster ID	Cluster name	Status	Creation time (UTC+12:00)
<input checked="" type="checkbox"/>	<input type="checkbox"/>	j-201JHIST308PX	My cluster	Waiting	June 11, 2024, 21:15

In the end, please go back to the Modules and click End Lab.

ALLv1-51345 > Modules > Learner Lab > Learner Lab

Home Modules Discussions

AWS Used \$0.1 of \$100 02:58 Start Lab End Lab

EN-US