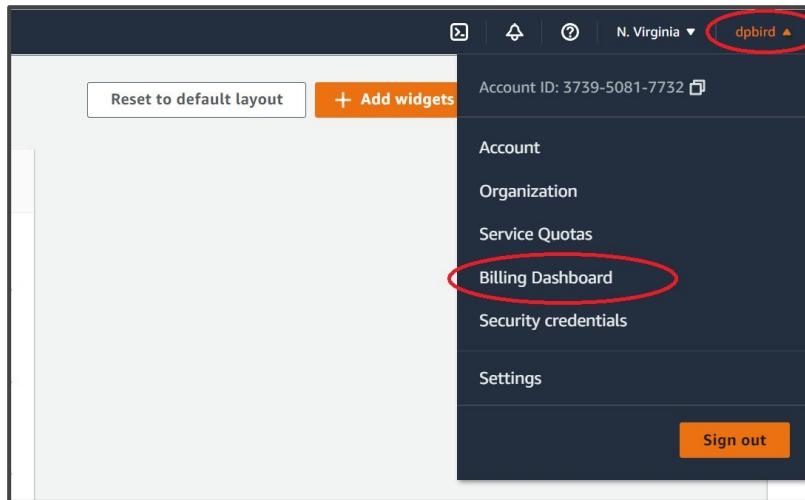


10-605 Homework 4 AWS Setup

Checking Credits

- All students should have \$100 put into their account.
- You should only need ~\$30 to complete Homework 4.

To check your credits, from your AWS Main Page:



Checking Credits

The screenshot shows the AWS Billing Credits page. The left sidebar has a 'Credits' link highlighted with a red circle. The main content area shows a summary table and a detailed list of credits.

Summary

Total amount remaining	Total amount used	Active credits
\$128.69	\$21.31	2

Credits

Expiration date	Credit name	Amount used	Amount remaining	Applicable products
07/31/2023	EDU_ENG_FY2021_CC_Q3_08_Carnegie_Mellon_University_50USD	\$21.31	\$28.69	See complete list of services
01/31/2023	EDU_E2W_FY2022_CC_Q3_09_CMU_100USD	\$0.00	\$100.00	See complete list of services

Checking Amount Spent

- Click Bills
 - Then you must click “Expand All” as it hides charges below:

The screenshot shows the AWS Billing interface. On the left, a sidebar lists various billing-related options: Home, Billing (with Bills selected and circled in red), Credits, Purchase orders, Cost & Usage Reports, Cost Categories, Cost allocation tags, Free Tier, Billing Conductor (with a dropdown arrow), Cost Management, Cost Explorer, Budgets, Budgets Reports, Savings Plans (with a dropdown arrow), Preferences, Billing preferences, Payment methods, Consolidated billing (with a dropdown arrow), and Tax settings. The main content area is titled 'Bills' and features a message box: 'Introducing the new AWS Bills page experience. We've redesigned the AWS Bills page so your AWS charges are clearer, and for you to find the right information you need. Try the new Bills page experience.' Below this is a date selector set to 'September 2022' and buttons for 'Download CSV' and 'Print'. A section titled 'Estimated Total' shows '\$0.00'. Under 'Details', there's a section for 'AWS Service Charges' with three items: 'Data Transfer' (\$0.00), 'Simple Storage Service' (\$0.00), and a button labeled '+ Expand All' (which is also circled in red). At the bottom, a small note explains that usage and recurring charges will be charged on the next billing date.

aws Services Search for services, features, blogs, docs, and more [Alt+S] Global ▾ dpbird ▾

Home

Billing **Bills**

Credits

Purchase orders

Cost & Usage Reports

Cost Categories

Cost allocation tags

Free Tier

Billing Conductor

Cost Management

Cost Explorer

Budgets

Budgets Reports

Savings Plans

Preferences

Billing preferences

Payment methods

Consolidated billing

Tax settings

Introducing the new AWS Bills page experience
We've redesigned the AWS Bills page so your AWS charges are clearer, and for you to find the right information you need. Try the new Bills page experience.

Date: September 2022

Download CSV Print

Estimated Total \$0.00

Credits

▶ Credits

Your invoiced total will be displayed once an invoice is issued.

+ Expand All

AWS Service Charges \$0.00

▶ Data Transfer \$0.00

▶ Simple Storage Service \$0.00

Usage and recurring charges for this statement period will be charged on your next billing date. Estimated charges shown on this page, or shown on any notifications that we send to you, may differ from your actual charges for this statement period. This is because estimated charges presented on this page do not include usage charges accrued during this statement period after the date you view this page. Similarly, information about estimated charges sent to you in a notification do not include usage charges accrued during this statement period after the date we send you the notification. One-time fees and subscription charges are assessed separately from usage and recurring charges, on the date that they occur. The charges on this page exclude taxes, unless it is listed as a separate line item. To access your tax information, contact your AWS Organization's management owner.

Checking Amount Spent

- Click Bills
 - Then you must click “Expand All” as it hides charges below:

The screenshot shows the AWS Bills page for September 2022. The sidebar on the left lists various billing options. The main area displays the 'Estimated Total' as \$0.00. Below this, under 'Details', the 'AWS Service Charges' section is expanded. It shows charges for Data Transfer (US East (N. Virginia)) and Simple Storage Service (No Region). A red box highlights the 'Collapse All' button at the top right of the details table.

Service	Description	Amount
Data Transfer	Bandwidth	\$0.00
	\$0.00 per GB - data transfer in per month \$0.00 per GB - data transfer out under the monthly global free tier	0.000000270 GB 0.000001 GB
Simple Storage Service	No Region	-\$0.02
	No Instance Type EDU_E2W_FY2022_CC_Q3_09_CMU_100USD	Credit -\$0.02
US East (N. Virginia)	Amazon Simple Storage Service Requests-Tier1	\$0.00
	\$0.005 per 1,000 PUT, COPY, POST, or LIST requests	5 000 Requests
	\$0.004 per 10,000 GET and all other requests	32 000 Requests
Amazon Simple Storage Service TimedStorage-ByteHrs	\$0.02	
\$0.023 per GB - first 50 TB / month of storage used	0.857 GB-Mo	

Usage and recurring charges for this statement period will be charged on your next billing date. Estimated charges shown on this page, or shown in any notifications that we send to you, may differ from your actual charges for this statement period. This is because estimated charges presented on this page do not include usage charges across all services, and they do not include recurring charges for services that are not yet active. The charges on this page are estimates only. The actual charges you incur will depend on the date you use the service, the date we send you the notification, and the specific usage and recurring charges on the date that they occur. The charges on this page exclude taxes, unless it is listed as a separate item. To access your tax information, contact your AWS Organization's management owner.

IMPORTANT: Creating a Budget Warning

- AWS is linked to your credit card you used to create the account. If you run over your \$100 allowance then it will automatically start charging your card.
- Every semester there is one student who leaves an instance or volume open for weeks because they neglected to set up warnings and they are left with a bill of literally **thousands** of dollars. The University is not responsible for this.
- PLEASE: Set up budget warnings and triple check that you have terminated all instances and volumes!

Creating a Budget Warning

The screenshot shows the AWS Billing Console Overview page. On the left, a sidebar menu lists various services: Home, Billing, Bills, Payments, Credits, Purchase orders, Cost & Usage Reports, Cost Categories, Cost allocation tags, Free Tier, Billing Conductor, Cost Management, Cost Explorer (with 'Budgets' highlighted), Budgets Reports, Savings Plans, Preferences, Billing preferences, Payment methods, Consolidated billing, and Tax settings. The main content area has a blue header bar with a search bar and a 'Submit feedback' button. Below the header, a message says 'You can now create a budget from a template, or explore hands-on tutorials to configure simple or advanced budgets. Learn more' with a link icon. The breadcrumb navigation shows 'Billing Console > Budgets > Overview'. The main section is titled 'Overview' with an 'Info' link. It contains a table titled 'Budgets (0) Info' with one row: 'Find a budget' and 'Show all budgets'. The table has columns: Name, Thresholds, Budget, Amount used, Forecasted amount, Current vs. budgeted, and Forecasted vs. budgeted. A note below the table says 'No budgets' and 'No budgets to display.' At the top right of the main content area are 'Download CSV' and 'Actions' buttons, and a prominent orange 'Create budget' button which is circled in red.

Creating a Budget Warning

Billing Console > Budgets > Create budget

Step 1
Choose budget type

Step 2
Set your budget

Step 3
Configure alerts

Step 4 - Optional
Attach actions

Step 5
Review

Choose budget type Info

Budget setup

Use a template (simplified)
Use the recommended configurations. You can change some configuration options after the budget is created.

Customize (advanced)
Customize a budget to set parameters specific to your use case. You can customize the time period, the start month, and specific accounts.

Budget types

Cost budget - Recommended
Monitor your costs against a specified dollar amount and receive alerts when your user-defined thresholds are met. Using cost budgets, the budgeted amount you set represents your expected cloud spend. For example, you can set a cost budget for a business unit and then add additional parameters such as the associated member accounts.

Usage budget
Monitor your usage of one or more specified usage types or usage type groups and receive alerts when your user-defined thresholds are met. Using usage budgets, the budgeted amount represents your expected usage. For example, you can use a usage budget to monitor the usage of certain services such as Amazon EC2 and Amazon S3.

Savings Plans budget
Track the utilization or coverage associated with your Savings Plans and receive alerts when your percentage drops below a threshold you define. Setting a coverage target lets you see how much of your instance usage is covered by Savings Plans, while setting a utilization target lets you see if your Savings Plans are unused or underutilized.

Reservation budget
Track the utilization or coverage associated with your reservations and receive alerts when your percentage drops below a threshold you define. Setting a coverage target lets you see how much of your instance usage is covered by reservations, while setting a utilization target lets you see if your reservations are unused or underutilized. Reservation alerts are supported for Amazon EC2, Amazon RDS, Amazon Redshift, Amazon ElastiCache, and Amazon Elasticsearch reservations.

Cancel

Next

Give your budget a name

Details

Budget name
Provide a descriptive name for this budget.

Names must be between 1-100 characters.

Set budget amount

Period
Daily budgets do not support enabling forecasted alerts, or daily budget planning.
 Monthly

Budget renewal type
 Recurring budget
Recurring budgets renew on the first day of every monthly billing period.
 Expiring budget
Expiring monthly budgets stop renewing at the end of the selected expiration month.

Start month End month

Budgeting method Info
 Fixed
Create a budget that tracks against a single monthly budgeted amount.
Enter your budgeted amount (\$)
Last month's cost: \$0.02

Budget scope Info
Add filtering and use advanced options to narrow the set of cost information tracked as part of this budget

Scope options

All AWS services (Recommended)
Track any cost incurred from any service for this account as part of the budget scope
 Filter specific AWS cost dimensions
Select specific dimensions to budget against. For example, you can select the specific service "EC2" to budget against.

Advanced options

Aggregate costs by
 Unblended costs

Creating a Budget Warning

Billing Console > Budgets > Create budget

Step 1

Choose budget type

Step 2

Set your budget

Step 3

Configure alerts

Step 4 - Optional

Attach actions

Step 5

Review

Configure alerts Info

▼ How budget alerts work



Why create budget alerts?

In order to be notified on the state of your budget, you can create up to 5 different alerts based on your budgeted amount. For example, create an alert to notify you when you have reached 75% of your budgeted amount.



How to get started?

Start by defining alert thresholds, then specify alert recipients and how you would like them to be notified. Alerts can be sent via email, AWS SNS, and AWS Chatbot.

Budget amount

Your budgeted amount: \$30.00

To change your budgeted amount, go back to step 2.

No alert thresholds created.

Add an alert threshold

Cancel

Previous

Next

Creating a Budget Warning

Step 4 - Optional
Attach actions

Step 5
Review

Why create budget alerts?
In order to be notified on the state of your budget, you can create up to 5 different alerts based on your budgeted amount. For example, create an alert to notify you when you have reached 75% of your budgeted amount.

How to get started?
Start by defining alert thresholds, then specify alert recipients and how you would like them to be notified. Alerts can be sent via email, AWS SNS, and AWS Chatbot.

Budget amount
Your budgeted amount: \$30.00
To change your budgeted amount, go back to step 2.

▼ Alert #1 Remove

Set alert threshold

Threshold When should this alert be triggered? Trigger How should this alert be triggered?

80 % of budgeted amount Actual

Summary: When your actual cost is greater than 80.00% (\$24.00) of your budgeted amount (\$30.00), the alert threshold will be exceeded.

Notification preferences
Select one or more notification preferences to receive alerts.

Email recipients
Specify the email recipients you want to notify when the threshold has exceeded.
dpbird@andrew.cmu.edu

Maximum number of email recipients is 10.

► Amazon SNS Alerts - Optional Info
► Amazon Chatbot Alerts - Optional

+ Add alert threshold

Cancel Previous Next

Set up an appropriate threshold of your budgeted amount. I used 80% →

Add your email address →

You can change this to "Projected" costs but this might cause AWS to warn you earlier than necessary, I'll keep it on "Actual"

Creating a Budget Warning

Billing Console > Budgets > Create budget

Step 1
Choose budget type

Step 2
Set your budget

Step 3
Configure alerts

Step 4 - Optional
Attach actions

Step 5
Review

Attach actions - *Optional* Info

▼ Using budgets actions



What is a budget action?
A budget action allows you to define and trigger cost saving responses to reinforce a cost-conscious culture. You have the option to attach actions that run whenever your alert threshold has been exceeded, such as stopping an EC2 instance from incurring any further costs. You can select the alerts to which you would like to attach actions, then define these actions.



How to get started?
To create a budget action, you will first need an alert threshold created from step 2. If you have already created an alert threshold select the type of action you want.

▼ Alert #1 (0 actions attached)

Threshold 80%	Email recipients dpbird@andrew.cmu.edu
Threshold measured against Actual Costs	Amazon SNS Not configured

Add action

Cancel Previous Next

Creating a Budget Warning

Step 1
Choose budget type

Step 2
Set your budget

Step 3
Configure alerts

Step 4 - Optional
Attach actions

Step 5
Review

Review Info

Step 1: Choose budget type

Budget type

Cost budget
Monitor your costs against a specified dollar amount and receive alerts when your user-defined thresholds are met.

Step 2: Set up your budget

Budget details

Name My AWS Budget	Start date Sep 2022	Budget amount \$30.00
Period Monthly	End date Oct 2022	

► Additional budget parameters

Step 3: Configure alerts

Alerts

Alert #1

Threshold
80% of budgeted amount

Threshold measured against
Actual costs

Step 4: Attach actions - optional

Actions

You have no budgets actions

Cancel Previous Create budget

Creating a Budget Warning

You can now create a budget from a [template](#), or explore hands-on [tutorials](#) to configure simple or advanced budgets. Learn more 

Your budget **My AWS Budget** has been created successfully.

Billing Console > Budgets > Overview

Overview 

Budgets (1) 

<input type="checkbox"/>	Name	Thresholds	Budget	Amount used	Forecasted amount	Current vs. budgeted	Forecasted vs. budgeted
<input type="checkbox"/>	My AWS Budget	 OK	\$30.00			0.00%	

 [Download CSV](#)  [Actions](#)  [Create budget](#)

< 1 > 

Budgets

Budgets Reports

Savings Plans 

Preferences

Billing preferences

Payment methods

Consolidated billing 

Tax settings

10-605 Setup for Homework Part A

- Setup S3 Bucket
- Setup IAM Roles
- Make a keypair
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Setup S3 Bucket

The screenshot shows the AWS search interface with the query 'S3' entered in the search bar. The results are categorized into 'Services' and 'Features'.

Services

- S3** ☆ Scalable Storage in the Cloud
- S3 Glacier** ☆ Archive Storage in the Cloud
- Athena** ☆ Query Data in S3 using SQL
- AWS Snow Family** ☆ Large Scale Data Transport

Features

- Amazon S3 File Gateway** Storage Gateway feature
- Batch Operations**
- Buckets**

[See all 7 results ▶](#)

[See all 12 results ▶](#)

Services (7)

- Features (12)
- Blogs (1,132)
- Documentation (112,728)
- Knowledge Articles (30)
- Tutorials (7)
- Events (15)
- Marketplace (868)

Features (12)

- Amazon S3 File Gateway
- Batch Operations
- Buckets

Documentation (112,728)

Knowledge Articles (30)

Tutorials (7)

Events (15)

Marketplace (868)

See all 7 results ▶

See all 12 results ▶

Amou

Home

Billing

Bills

Payments

Credits

Purchase orders

Cost & Usage Reports

Cost Categories

Cost allocation tags

Free Tier

Billing Conductor

Cost Management

Cost Explorer

Budgets

Budgets Reports

Savings Plans

Preferences

Billing preferences

Payment methods

Consolidated billing

Tax settings

Setup S3 Bucket

Amazon S3 X We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose [Provide feedback](#). Provide feedback

Buckets Access Points Object Lambda Access Points Multi-Region Access Points Batch Operations Access analyzer for S3 Block Public Access settings for this account Storage Lens Dashboards AWS Organizations settings Feature spotlight: 3

Amazon S3 > Buckets

▶ Account snapshot Storage lens provides visibility into storage usage and activity trends. Learn more  View Storage Lens dashboard

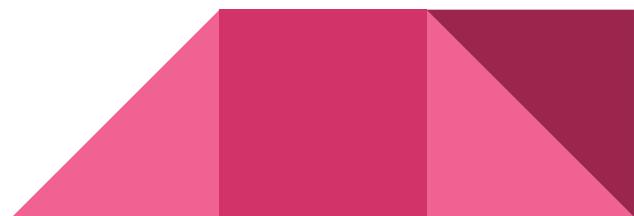
Buckets (1) Info Buckets are containers for data stored in S3. Learn more 

Find buckets by name < 1 >

Name	AWS Region	Access	Creation date
hw4605	US East (N. Virginia) us-east-1	Bucket and objects not public	October 21, 2021, 09:26:58 (UTC-04:00)

C Copy ARN Empty Delete Create bucket

▶ AWS Marketplace for S3



Setup S3 Bucket

Give your bucket a name. This name must be globally unique and not contain spaces or uppercase letters

The screenshot shows the 'Create bucket' wizard on the Amazon S3 service. The first step, 'General configuration', is active. It includes fields for 'Bucket name' (containing 'danielbirdbucket') and 'AWS Region' (set to 'US East (N. Virginia) us-east-1'). A note indicates that the bucket name must be globally unique and cannot contain spaces or uppercase letters. Below these fields is a section for 'Copy settings from existing bucket - optional', which includes a 'Choose bucket' button. The second step, 'Object Ownership', is shown below, with the 'ACLs disabled (recommended)' option selected. The third step, 'Block Public Access settings for this bucket', is at the bottom, with the 'Block all public access' checkbox checked.

10-605/805 MUST SET
AWS REGION TO "US
East (N. Virginia)
us-east-1" We need this
to get the Million Song
Dataset

Then click: "Create
Bucket"

Setup S3 Bucket

Amazon S3

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose [Provide feedback](#).

Successfully created bucket "danielbirdbucket". To upload files and folders, or to configure additional bucket settings choose [View details](#).

Buckets

Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards
AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets

▶ Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

[View Storage Lens dashboard](#)

Buckets (2) Buckets are containers for data stored in S3. Learn more

Q Find buckets by name

< 1 >

Name	AWS Region	Access	Creation date
danielbirdbucket	US East (N. Virginia) us-east-1	Bucket and objects not public	September 28, 2022, 13:03:01 (UTC-04:00)
hw4605	US East (N. Virginia) us-east-1	Bucket and objects not public	October 21, 2021, 09:26:58 (UTC-04:00)

[Create bucket](#)

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles
- Make a keypair
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Setup IAM Roles

- Identity and Access Management

The screenshot shows the AWS search interface with the query 'iam' entered in the search bar. The results are categorized into Services and Features.

Services (6)

- IAM: Manage access to AWS resources. This is the top result and is highlighted with a blue border. It includes a 'Top features' section with links to Groups, Users, Roles, Policies, and Access Analyzer.
- IAM Identity Center (successor to AWS Single Sign-On): Manage workforce user access to multiple AWS accounts and cloud applications.
- Resource Access Manager: Share AWS resources with other accounts or AWS Organizations.
- Amazon VPC IP Address Manager: Managed IP address management service.

Features (17)

- Groups: IAM feature.
- Roles: IAM feature.
- Access Analyzer

See all 6 results ▶

See all 17 results ▶

Amazon S3

- Buckets
 - Access Points
 - Object Lambda Access
 - Multi-Region Access
 - Batch Operations
 - Access analyzer for S3
- Marketplace (361)

Storage Lens

- Dashboards
- AWS Organizations

Feature spotlight (3)

AWS Marketplace

Setup IAM Roles

Screenshot of the AWS IAM Roles page.

The left sidebar shows the navigation menu:

- Identity and Access Management (IAM)
- Dashboard
- Access management
 - User groups
 - Roles** (highlighted with a red circle)
 - Policies
 - Identity providers
 - Account settings
- Access reports
 - Access analyzer
 - Archive rules
 - Analyzers
 - Settings
 - Credential report
 - Organization activity
- Service control policies (SCPs)

The main content area displays the Roles list:

Role name	Trusted entities	Last activity
AWSServiceRoleForEMRCleanup	AWS Service: elasticmapreduce (Service-Linked Role)	47 minutes ago
AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	
AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Role)	
EMR_AutoScaling_DefaultRole	AWS Service: application-autoscaling, and 1 more.	
EMR_DefaultRole	AWS Service: elasticmapreduce	2 hours ago
EMR_EC2_DefaultRole	AWS Service: ec2	2 hours ago
hwrole	AWS Service: ec2	335 days ago

A message on the right says: "You might not have any current roles".

At the top right of the table, there are "Delete" and "Create role" buttons, with "Create role" being highlighted with a red circle.

Below the table, there are three sections:

- Roles Anywhere** [Info](#)
Authenticate your non AWS workloads and securely provide access to AWS services.

Access AWS from your non AWS workloads
Operate your non AWS workloads using the same authentication and authorization strategy that you use within AWS.
- 
X.509 Standard
Use your own existing PKI infrastructure or use AWS Certificate Manager Private Certificate Authority to authenticate identities.
- 
Temporary credentials
Use temporary credentials with ease and benefit from the enhanced security they provide.

Setup IAM Roles

IAM > Roles > Create role

Step 1

Select trusted entity

Step 2

Add permissions

Step 3

Name, review, and create

Select trusted entity

Trusted entity type

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.

Web identity

Allows users federated by the specified external web identity provider to assume this role 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.

Common use cases:

EC2

Allows EC2 instances to call AWS services on your behalf.

Lambda

Allows Lambda functions to call AWS services on your behalf.

Use cases for other AWS services:

Choose a service to view use case ▾

Cancel

Next

Setup IAM Roles

Introducing the new IAM roles experience
We've redesigned the IAM roles experience to make it easier to use. [Let us know what you think.](#)

IAM > Roles > Create role

Step 1
Select trusted entity

Step 2
Add permissions

Step 3
Name, review, and create

Add permissions

Search for "S3FullAccess"

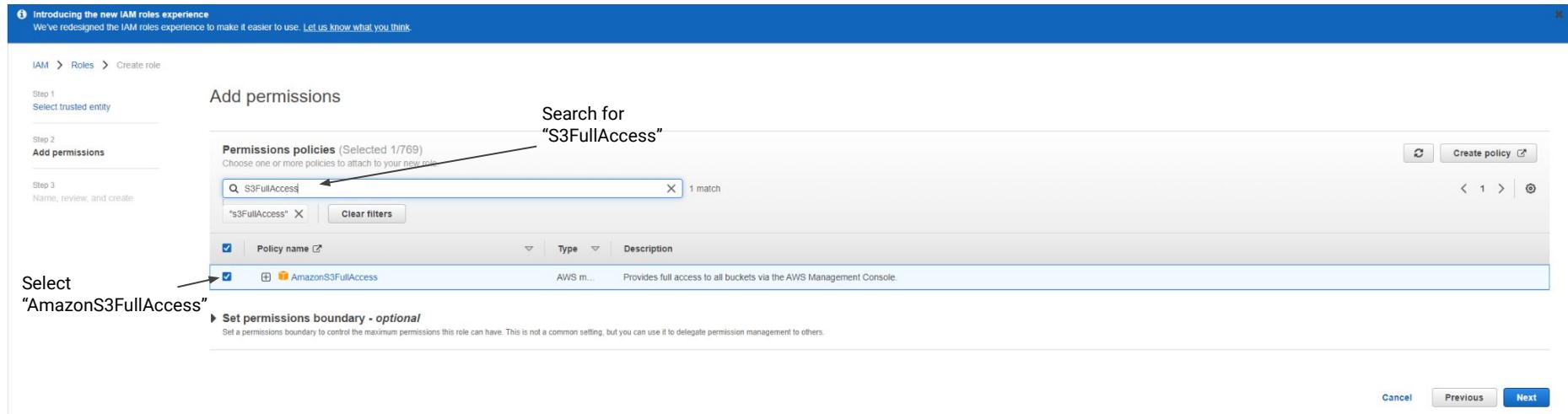
Permissions policies (Selected 1/769)
Choose one or more policies to attach to your new role

S3FullAccess X 1 match

AmazonS3FullAccess AWS m... Provides full access to all buckets via the AWS Management Console.

Set permissions boundary - optional
Set a permissions boundary to control the maximum permissions this role can have. This is not a common setting, but you can use it to delegate permission management to others.

Cancel Previous Next



This is allowing EC2 instances to have full access to all S3 Buckets!

Setup IAM Roles

Introducing the new IAM roles experience
We've redesigned the IAM roles experience to make it easier to use. Let us know what you think.

IAM > Roles > Create role

Step 1
Select trusted entity

Step 2
Add permissions

Step 3
Name, review, and create

Name, review, and create

Role details

Give your role a name

Role name
Enter a meaningful name to identify this role.
AVSMainRole

Description
Add a short explanation for this role.
Allows EC2 instances to call AWS services on your behalf.

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

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

Step 1: Select trusted entities

```
1+ [{}  
2+     "Version": "2012-10-17",  
3+     "Statement": [  
4+         {  
5+             "Effect": "Allow",  
6+             "Action": [  
7+                 "sts:AssumeRole"  
8+             ],  
9+             "Principal": [  
10+                {  
11+                    "Service": [  
12+                        "ec2.amazonaws.com"  
13+                    ]  
14+                }  
15+            ]  
16+        ]]
```

Step 2: Add permissions

Permissions policy summary

Policy name	Type	Attached as
AmazonS3FullAccess	AWS managed	Permissions policy

Setup IAM Roles

New! Securely access AWS services from your data center with IAM Roles Anywhere. [Learn more](#)

Role AWSMainRole created.

IAM > Roles

Roles (8) Info

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Role name	Trusted entities	Last activity
AWSMainRole	AWS Service: ec2	-
AWSServiceRoleForEMRCleanup	AWS Service: elasticmapreduce (Service-Linked Role)	58 minutes ago
AWSServiceRoleForSupport	AWS Service: support (Service-Linked Role)	-
AWSServiceRoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Role)	-
EMR_AutoScaling_DefaultRole	AWS Service: elasticmapreduce, and 1 more.	-
EMR_DefaultRole	AWS Service: elasticmapreduce	2 hours ago
EMR_EC2_DefaultRole	AWS Service: ec2	2 hours ago
hw4role	AWS Service: ec2	335 days ago

Roles Anywhere [Info](#)

Authenticate your non AWS workloads and securely provide access to AWS services.

Access AWS from your non AWS workloads
Operate your non AWS workloads using the same authentication and authorization strategy that you use within AWS.

X.509 Standard
Use your own existing PKI infrastructure or use AWS Certificate Manager Private Certificate Authority to authenticate identities.

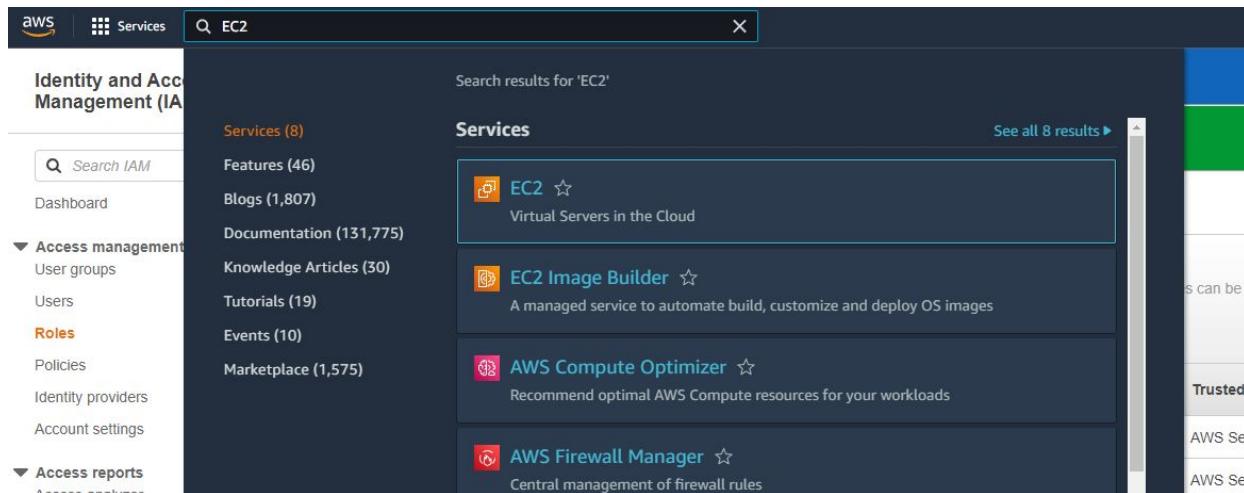
Temporary credentials
Use temporary credentials with ease and benefit from the enhanced security they provide.

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Make a keypair

- This is needed to access your EC2 instance and your EMR Cluster.
- It is important that it is private and you know where it is saved.
- First head to EC2



Make a keypair

The screenshot shows the AWS EC2 Key Pairs page. On the left, there's a sidebar with various navigation options like EC2 Dashboard, Instances, AMIs, Elastic Block Store, Network & Security, and Load Balancing. Under Network & Security, the 'Key Pairs' link is highlighted with a red circle. At the top right, there's a search bar and a 'Create key pair' button, which is also circled in red.

Name	Type	Created	Fingerprint	ID
hw4key	rsa	2021/10/21 09:33 GMT-4	70:d1:c7:97:55:5f:41:9c:97:6f:b7:bf:d0:a...	key-011a3f9b942e9f987
keyhw4	rsa	2021/10/27 10:05 GMT-4	3e:53:99:16:ce:4b:0a:6e:67:3e:e6:41:14:...	key-07af2e66699f563e8
testkeyexampledelete	rsa	2022/09/28 09:13 GMT-4	be:02:c1:2f:c2:04:e9:8a:6c:51:4ff:fc:4e:5...	key-066e4d9acf3af9af0

Make a keypair

The screenshot shows the 'Create key pair' wizard in the AWS Management Console. The 'Name' field contains 'myawskey'. The 'Key pair type' section has 'RSA' selected. The 'Private key file format' section has '.pem' selected. There are tabs for 'Info' and 'Advanced options' at the bottom right of the form.

Give your key a name,
I recommend not
using spaces

Select RSA

Select .pem

- Once you click “Create Key Pair”, this will download a .pem. Make sure you save this somewhere safe where you can locate it.

Make a keypair

New EC2 Experience Tell us what you think

Successfully created key pair

Key pairs (4) [Info](#)

Name	Type	Created	Fingerprint	ID
hw4key	rsa	2021/10/21 09:33 GMT-4	70:d1:c7:97:55:5f:41:9c:97:6f:b7:bf:d0:a...	key-011a3f9b942e9f987
keyhw4	rsa	2021/10/27 10:05 GMT-4	3e:53:99:16:ce:4b:0x:6e:67:3e:c6:41:14:...	key-07af2e66699f563e8
myawskey	rsa	2022/09/28 13:46 GMT-4	e7:71:3d:b6:8a:8e:17:9d:30:1c:1f:5e:d6:...	key-0f3329c9cc5558b66
testkeyexampledelete	rsa	2022/09/28 09:13 GMT-4	be:02:c1:2f:c2:04:e9:8a:6c:51:4f:fc:4e:5...	key-066e4d9a:f3af9af0

Actions [Create key pair](#)

EC2 Dashboard [EC2 Global View](#) [Events](#) [Tags](#) [Limits](#)

Instances [New](#) [Instance Types](#) [Launch Templates](#) [Spot Requests](#) [Savings Plans](#) [Reserved Instances New](#) [Dedicated Hosts](#) [Scheduled Instances](#) [Capacity Reservations](#)



10-605 Setup for Homework Part A

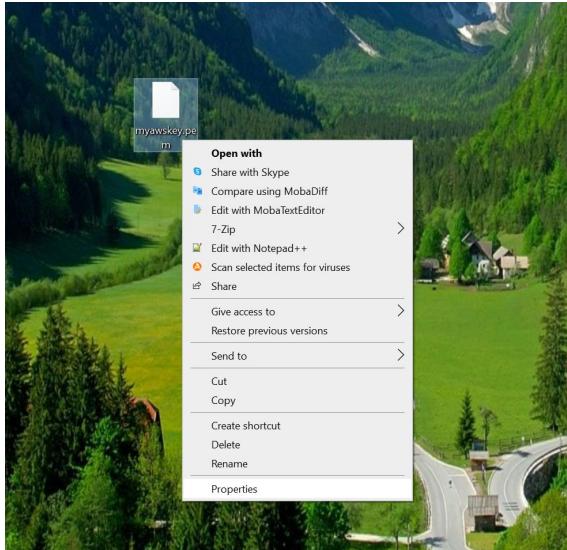
- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE Kind of...
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Make a keypair - DONE Kind of...

- Your .pem file is currently “Public” on your computer. When you try to access an instance with your .pem file it won’t allow you until you change the permissions.
- How you do this depends on whether you are using Windows or Mac

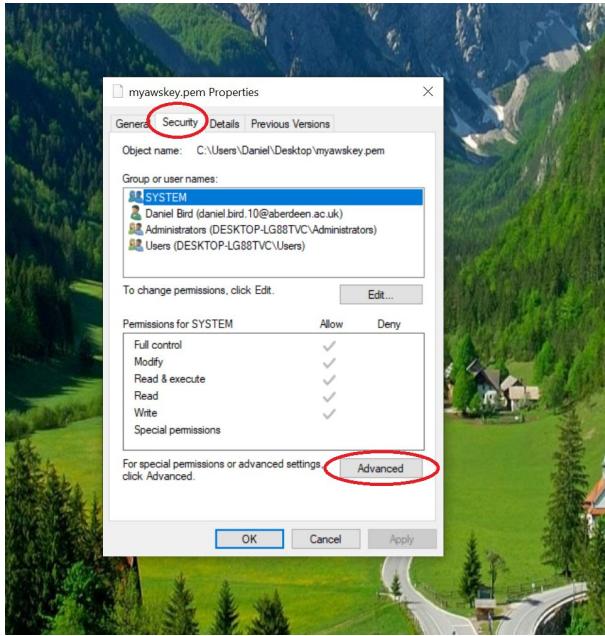
Make a keypair - Windows Instructions

- Locate the file, right click and select “Properties”



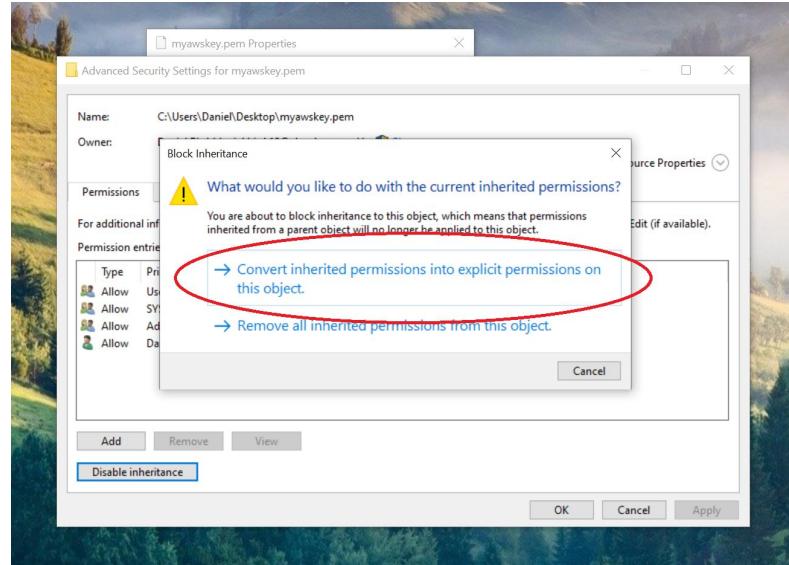
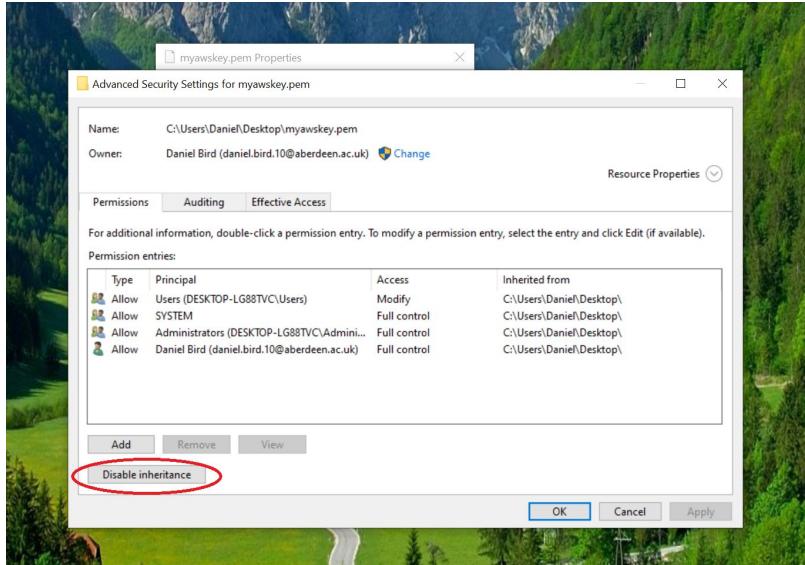
Make a keypair - Windows Instructions

- Click the “Security” tab, then “Advanced”.



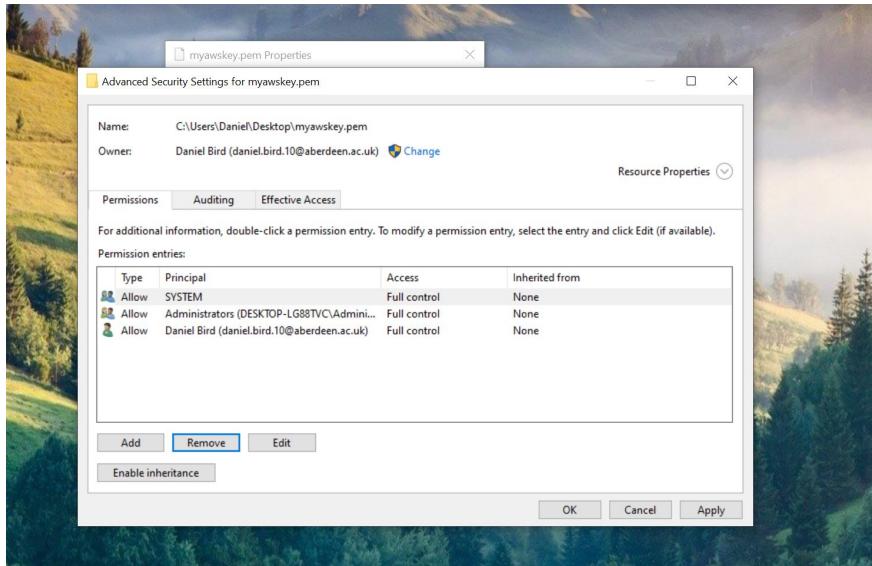
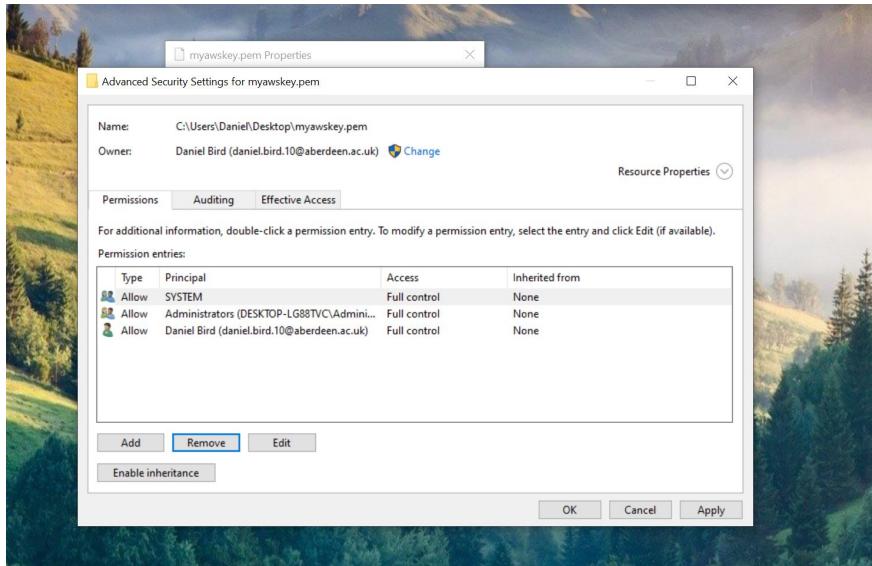
Make a keypair - Windows Instructions

- Click “Disable inheritance”, then “Convert inherited permissions into...”



Make a keypair - Windows Instructions

- Remove “Users” by clicking on it and then clicking the Remove button.



Make a keypair - Mac Instructions

- In the terminal run
 - chmod 400 #KEYPAIRNAME
 - So I would run:
 - chmod 400 myawskey.pem

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Setup EC2 Instance

- What is EC2? (Elastic Compute)
- It consists of the ability to:
 - Rent Virtual Machines (EC2)
 - Store data on virtual drives (EBS)
 - Other useful services which have been adopted by many companies who require cloud computing
- You can think of this as the main AWS offering.

Setup EC2 Instance

The screenshot shows the AWS EC2 Instances page. The left sidebar has a 'New EC2 Experience' header with a 'Tell us what you think' link. It includes sections for EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances (with 'Instances New' highlighted), Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances (with 'New' link), Dedicated Hosts, Scheduled Instances, Capacity Reservations, Images (with 'AMIs New' link), and AMI Catalog. The main content area has a 'Instances Info' header with a search bar, 'Connect' button, and dropdowns for Instance state, Actions, and Launch Instances. A red box highlights the 'Launch Instances' button. Below it, a message says 'No instances' and 'You do not have any instances in this region'. A large 'Launch instances' button is centered. At the bottom, a modal window titled 'Select an instance' is open, also with a red border.

New EC2 Experience X

EC2 Dashboard

EC2 Global View

Events

Tags

Limits

Instances

Instances New

Instance Types

Launch Templates

Spot Requests

Savings Plans

Reserved Instances New

Dedicated Hosts

Scheduled Instances

Capacity Reservations

Images

AMIs New

AMI Catalog

Elastic Block Store

Volumes New

Snapshots New

Lifecycle Manager New

Network & Security

Security Groups

Instances Info

Find instance by attribute or tag (case-sensitive)

Connect

Instance state

Actions

Launch Instances

No instances

You do not have any instances in this region

Launch instances

Select an instance

Setup EC2 Instance

EC2 > Instances > Launch an instance

Launch an instance Info

Amazon EC2 allows you to create virtual machines, or instances, that run on the AWS Cloud. Quickly get started by following the simple steps below.

Name and tags Info

Give your EC2 instance a name

Name Add additional tags

Application and OS Images (Amazon Machine Image) Info

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. Search or Browse for AMIs if you don't see what you are looking for below.

Search our full catalog including 1000s of application and OS images

Quick Start

Amazon Linux  macOS  Ubuntu  Windows  Red Hat  ... 

 Browse more AMIs
Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

Amazon Linux 2 AMI (HVM) - Kernel 5.10, SSD Volume Type
Free tier eligible

ami-026b57f3c383e2ec (64-bit (x86)) / ami-0636eac5d73e0e5d7 (64-bit (Arm))
Virtualization: hvm ENA enabled: true Root device type: ebs

Description

Summary

Number of instances Info
1

Software Image (AMI)
Amazon Linux 2 Kernel 5.10 AMI... [read more](#)
ami-026b57f3c383e2ec

Virtual server type (instance type)
t2.micro

Firewall (security group)
New security group

Storage (volumes)
1 volume(s) - 8 GiB

 **Free tier:** In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GiB of EBS storage, 2 million I/Os, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

Cancel **Launch instance**

Setup EC2 Instance

Search our full catalog including 1000s of application and OS images

Quick Start

Leave your quick start option as is

Amazon Linux macOS Ubuntu Windows Red Hat S Browse more AMIs Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

Amazon Linux 2 AMI (HVM) - Kernel 5.10, SSD Volume Type
ami-026b57f3c383c2eec (64-bit (x86)) / ami-0636ea5c5d73e0e5d7 (64-bit (Arm))
Virtualization: hvm ENA enabled: true Root device type: ebs

Free tier eligible

Description

Amazon Linux 2 Kernel 5.10 AMI 2.0.20220912.1 x86_64 HVM gp2

Architecture AMI ID

64-bit (x86) ami-026b57f3c383c2eec Verified provider

Make sure your architecture is 64-bit (x86)

▼ Instance type Info

Instance type

t2.micro Free tier eligible

Family: t2 1 vCPU 1 GiB Memory
On-Demand Linux pricing: 0.0116 USD per Hour
On-Demand Windows pricing: 0.0162 USD per Hour

Compare instance types

▼ Key pair (login) Info

You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch

▼ Summary

Number of instances [Info](#)
1

Software Image (AMI)
Amazon Linux 2 Kernel 5.10 AMI...[read more](#)
ami-026b57f3c383c2eec

Virtual server type (instance type)
t2.micro

Firewall (security group)
New security group

Storage (volumes)
1 volume(s) - 8 GiB

Free tier: In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GiB of EBS storage, 2 million I/Os, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

Cancel Launch instance

For development purposes you can use t2.micro. When actually running the million_song_prep we recommend using t2.medium or t2.large

Setup EC2 Instance

The screenshot shows the AWS EC2 instance creation wizard. It consists of two main panels.

Left Panel: Key Pair Selection

- Key pair (login) Info:** A note says "You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch the instance."
- Key pair name - required:** A dropdown menu is open, showing "myawskey".
- Create new key pair:** A button to generate a new key pair.
- Network settings Info:** A note says "We need to choose our Subnet, so click the Edit button under Network Settings". An arrow points to the "Edit" button, which is circled in red.
- Subnet Info:** Shows "vpc-09b3d5ce9ddca5b53".
- No preference (Default subnet in any availability zone)**
- Auto-assign public IP Info:**
- Firewall (security groups) Info:** A note says "A security group is a set of firewall rules that control the traffic for your instance. Add rules to allow specific traffic to reach your instance." Two radio buttons are shown: "Create security group" (selected) and "Select existing security group".
- Rules:**
 - Allow SSH traffic from**: Helps you connect to your instance. Destination: Anywhere (0.0.0.0/0).
 - Allow HTTPS traffic from the internet**: To set up an endpoint, for example when creating a web server.
 - Allow HTTP traffic from the internet**: To set up an endpoint, for example when creating a web server.
- Warning:** "⚠ Rules with source of 0.0.0.0/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only."

Right Panel: Summary and Launch Step

- Summary:** Number of instances: 1
- Software Image (AMI):** Amazon Linux 2 Kernel 5.10 AMI... (with a "read more" link)
- Virtual server type (instance type):** t2.micro
- Firewall (security group):** New security group
- Storage (volumes):** 1 volume(s) - 8 GB
- Free tier information:** In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GB of EBS storage, 2 million I/Os, 1 GB of snapshots, and 100 GB of bandwidth to the internet.
- Buttons:** Cancel, Launch Instance (highlighted in orange).

Setup EC2 Instance

10-605 Must choose a subnet which is in us-east-1 otherwise you will not be able to get the data

Choose "Select existing security group"

Choose "default" from the "Select security groups" dropdown

Configure storage: 1x 8 GiB gp2 Root volume

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage

Summary: Number of instances 1 Software Image (AMI) Amazon Linux 2 Kernel 5.10 AMI...read more ami-026b57f3c383c2ec Virtual server type (instance type) Firewall (security group) default Storage (volumes) 1 volume(s) - 8 GiB

Launch Instance

The image shows the AWS EC2 instance setup process. On the left, under 'Network settings', a red box highlights the 'Subnet Info' section where a specific subnet is selected. Another red box highlights the 'Select existing security group' button in the 'Firewall (security group)' section. A third red box highlights the 'Select security groups' dropdown, with an annotation pointing to the 'default' option. On the right, the 'Summary' step shows the instance configuration: one instance, Amazon Linux 2 AMI, t2.micro instance type, and 8 GiB of storage. A tooltip for the 'Free tier' is visible. At the bottom right is a large orange 'Launch Instance' button.

Setup EC2 Instance

The screenshot shows the 'Advanced details' section of the EC2 instance setup. A red oval highlights the 'IAM instance profile' dropdown, which contains the value 'AWSMainRole'. An annotation above the dropdown reads: 'Under "Advanced details" set your IAM Instance Profile to be the one you created'.

Below the dropdown, there is a 'Create new IAM profile' button. The 'Hostname type' field is also circled in red. The 'Launch instance' button is highlighted with a red arrow and the text 'Then click Launch Instance!'. A callout box provides information about the Free tier:

Free tier: In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GB of EBS storage, 2 million I/Os, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

Other visible fields include:

- Purchasing option: Request Spot Instances (unchecked)
- Domain join directory: Select (dropdown), Create new directory (button)
- IAM instance profile: AWSMainRole (selected), Create new IAM profile (button)
- Hostname type: IP name (dropdown)
- DNS Hostname:
 - Enable IP name IPv4 (A record) DNS requests (checked)
 - Enable resource-based IPv4 (A record) DNS requests (checked)
 - Enable resource-based IPv6 (AAAA record) DNS requests (unchecked)
- Instance auto-recovery: Select (dropdown)
- Shutdown behavior: Select (dropdown)
- Stop - Hibernate behavior: Select (dropdown)

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - Launching...
 - Add an additional Volume
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Add an additional Volume to your EC2 Instance

The screenshot shows the AWS EBS Volumes page. On the left, there's a navigation sidebar with various services like Launch Templates, Spot Requests, and Auto Scaling. The 'Elastic Block Store' section is expanded, and 'Volumes' is selected, which is highlighted with a red circle. Below the sidebar, a table lists the volumes. The first row shows a volume with the following details:

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created	Availability Zone	Volume state	Alarm status
-	vol-0813c18439814128f	gp2	8 GiB	100	-	snap-07bb851...	2022/09/29 10:24 GMT-4	us-east-1a	In-use	No alarms

An arrow points from the text below to the 'Name' column of the first row. A red box highlights the 'Create volume' button in the top right corner of the table header.

You will notice the volume that our EC2 instance just created is here!

Add an additional Volume to your EC2 Instance

The screenshot shows the 'Create volume' wizard in the AWS Management Console. The current step is 'Volume settings'. Key configuration options include:

- Volume type:** General Purpose SSD (gp2)
- Size (GiB):** 500 (with a note: "The Million Song Dataset recommends 500GB of data")
- IOPS:** 1500 / 3000
- Throughput (MiB/s):** Not applicable
- Availability Zone:** us-east-1a (with a note: "The Availability Zone must match the subnet we chose when making the EC2 instance")
- Snapshot ID - optional:** Don't create volume from a snapshot
- Encryption:** Use Amazon EBS encryption as an encryption solution for your EBS resources associated with your EC2 instances. (checkbox: Encrypt this volume)
- Tags - optional:** A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter.

- We want this volume to come with our dataset on it. To do this we can use a SnapshotID
- Head to <http://millionsongdataset.com/>

Add an additional Volume to your EC2 Instance

The screenshot shows the homepage of the Million Song Dataset. At the top, there's a navigation bar with links: Home, Getting the dataset (which is highlighted with a red circle), Code, Tutorial, Tasks / Demos, More data, Forum, Contact / Cite, and Blog. Below the navigation bar, there's a sub-navigation menu with Home, Getting the dataset (again), Code, Tutorial, and News. The main content area has a heading "Getting the dataset". It includes an *Important note* about audio inclusion, a note about distribution logistics, and instructions for reviewing the dataset contents. It also lists universities that have a copy of the dataset. A section for AWS users provides instructions on how to attach an EBS volume to an EC2 instance. At the bottom, there's a terminal window showing the command-line steps to create a snapshot and mount it. A note at the very bottom says "The 493G partition at the end (of which only 272G used) is the MSD data."

- Click “Getting the Dataset” button on the top tab.
- Then copy the snap code: snap-5178cf30

Add an additional Volume to your EC2 Instance

EC2 > Volumes > Create volume

Create volume Info

Create an Amazon EBS volume to attach to any EC2 instance in the same Availability Zone.

Volume settings

Volume type Info
General Purpose SSD (gp2)

Size (GiB) Info
500
Min: 1 GiB, Max: 16384 GiB. The value must be an integer.

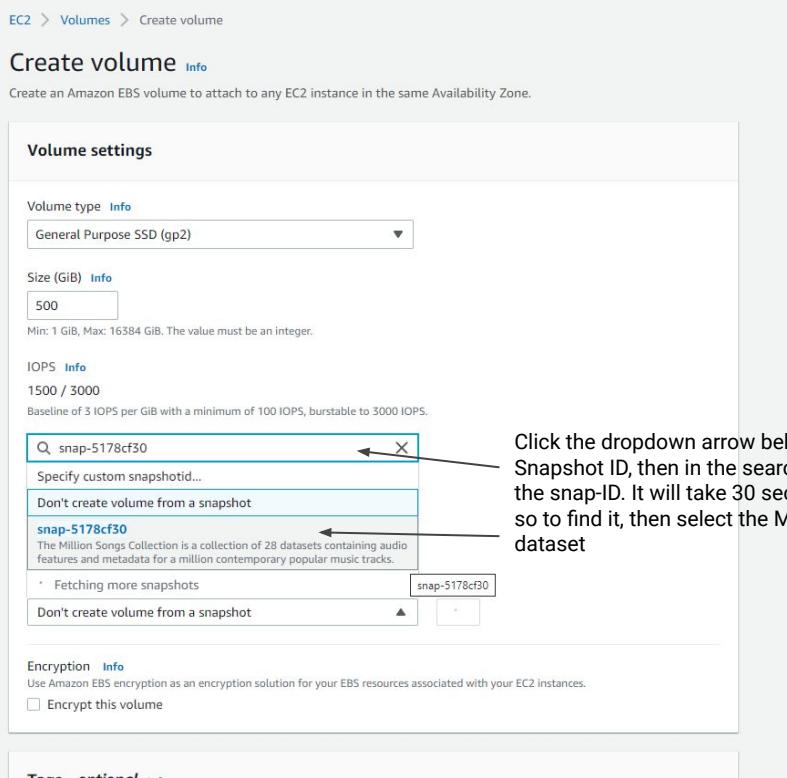
IOPS Info
1500 / 3000
Baseline of 3 IOPS per GiB with a minimum of 100 IOPS, burstable to 3000 IOPS.

Snapshot ID: snap-5178cf30

Click the dropdown arrow below Snapshot ID, then in the search bar add the snap-ID. It will take 30 seconds or so to find it, then select the Million song dataset

Encryption Info
Use Amazon EBS encryption as an encryption solution for your EBS resources associated with your EC2 instances.
 Encrypt this volume

Tags - optional Info
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter resources.



Then click create volume!

Add an additional Volume to your EC2 Instance

New EC2 Experience X

Tell us what you think

EC2 Dashboard

EC2 Global View

Events

Tags

Limits

Instances

- Instances New
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances New
- Dedicated Hosts
- Scheduled Instances
- Capacity Reservations

Images

- AMIs New
- AMI Catalog

Elastic Block Store

- Volumes New
- Snapshots New
- Lifecycle Manager New

Network & Security

- Security Groups
- Elastic IPs
- Placement Groups

Successfully created volume vol-07f5a9004b99188da.

Volumes (2)

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created	Availability Zone	Volume state	Alarm status
-	vol-0813c18439814128f	gp2	8 GiB	100	-	snap-07bb851...	2022/09/29 10:24 GMT-4	us-east-1a	In-use	No alarms
-	vol-07f5a9004b99188da	gp2	500 GiB	1500	-	snap-5178cf30	2022/09/29 10:43 GMT-4	us-east-1a	Creating	No alarms

Our new volume. It will take a while to start up

Select a volume above

The screenshot shows the AWS EC2 Volumes page. On the left, there's a navigation sidebar with links for EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances (with sub-links for Instances, Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances, Dedicated Hosts, Scheduled Instances, Capacity Reservations), Images (with sub-links for AMIs, AMI Catalog), and Elastic Block Store (with sub-links for Volumes, Snapshots, and Lifecycle Manager). The main content area has a green header bar stating "Successfully created volume vol-07f5a9004b99188da.". Below this is a table titled "Volumes (2)" showing two entries. The first entry is a 8 GiB gp2 volume with Name "-", Volume ID "vol-0813c18439814128f", Type "gp2", Size "8 GiB", IOPS "100", Throughput "-", Snapshot "snap-07bb851...", Created "2022/09/29 10:24 GMT-4", Availability Zone "us-east-1a", Volume state "In-use", and Alarm status "No alarms". The second entry is a 500 GiB gp2 volume with Name "-", Volume ID "vol-07f5a9004b99188da", Type "gp2", Size "500 GiB", IOPS "1500", Throughput "-", Snapshot "snap-5178cf30", Created "2022/09/29 10:43 GMT-4", Availability Zone "us-east-1a", Volume state "Creating", and Alarm status "No alarms". A callout arrow points from the text "Our new volume. It will take a while to start up" to the second volume row. At the bottom of the table, it says "Select a volume above".

Add an additional Volume to your EC2 Instance

A note on Volumes: This will be easily the most expensive part of your HW4, it is also the easiest to forget to terminate!

I recommend only setting this up when you plan to run `million_song_prep.py` and terminating the volume after it has finished running and you can see your csv files in your S3 bucket.

Add an additional Volume to your EC2 Instance

The screenshot shows the AWS EC2 Volumes page with two volumes listed:

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created	Availability Zone	Volume state	Alarm status
vol-0813c18439814128f	gp2	8 GiB	100	-	snap-07bb851...	2022/09/29 10:24 GMT-4	us-east-1a	In-use	No alarms	
vol-07f5a9004b99188da	gp2	500 GiB	1500	-	snap-5178cf30	2022/09/29 10:43 GMT-4	us-east-1a	Available	No alarms	

A callout arrow points to the "Available" status of the second volume. The text in the callout reads:

Once the volume is ready to be attached you will see the "Volume State" change to Available

New EC2 Experience Tell us what you think X

EC2 Dashboard
EC2 Global View
Events
Tags
Limits

Instances

- Instances New
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances New
- Dedicated Hosts
- Scheduled Instances
- Capacity Reservations

Images

- AMIs New
- AMI Catalog

Elastic Block Store

- Volumes New
- Snapshots New
- Lifecycle Manager New

Network & Security

- Security Groups
- Elastic IPs
- Placement Groups

Feedback Looking for language selection? Find it in the new Unified Settings ?

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences 10:49 AM

Add an additional Volume to your EC2 Instance

The screenshot shows the AWS EC2 Volumes page. On the left, there's a navigation sidebar with links like New EC2 Experience, EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances (with sub-links for Instances, Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances, Dedicated Hosts, Scheduled Instances, Capacity Reservations), Images (AMIs, AMI Catalog), and Elastic Block Store (Volumes, Snapshots, Lifecycle Manager). The main area displays a table of volumes with columns: Name, Volume ID, Type, Size, IOPS, Throughput, Snapshot, and Created. Two volumes are listed: one with 8 GiB and another with 500 GiB. The 500 GiB volume has a checkmark next to it. To the right of the table is an 'Actions' dropdown menu with options: Modify volume, Create snapshot, Create snapshot lifecycle policy, Delete volume, Attach volume, Detach volume, Force detach volume, Manage auto-enabled I/O, and Manage tags. The 'Actions' button and the 'Modify volume' option are highlighted with red circles.

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created
-	vol-0813c18439814128f	gp2	8 GiB	100	-	snap-07bb851...	2022/09/29 10:24 GMT-4
<input checked="" type="checkbox"/>	vol-07f5a9004b99188da	gp2	500 GiB	1500	-	snap-5178cf30	2022/09/29 10:43 GMT-4

Actions ▾

- Modify volume
- Create snapshot
- Create snapshot lifecycle policy
- Delete volume
- Attach volume
- Detach volume
- Force detach volume
- Manage auto-enabled I/O
- Manage tags

Volume ID: vol-07f5a9004b99188da

Details | Status checks | Monitoring | Tags

Details

Add an additional Volume to your EC2 Instance

EC2 > Volumes > vol-07f5a9004b99188da > Attach volume

Attach volume [Info](#)

Attach a volume to an instance to use it as you would a regular physical hard disk drive.

Basic details

Volume ID
 vol-07f5a9004b99188da

Availability Zone
us-east-1a

Instance [Info](#)

Only instances in the same Availability Zone as the selected volume are displayed.

Device name [Info](#)

Recommended device names for Linux: /dev/sda1 for root volume. /dev/sd[f-p] for data volumes.

ⓘ Newer Linux kernels may rename your devices to /dev/xvdf through /dev/xvd

internally, even when the device name entered here (and shown in the details) is /dev/sdf through /dev/sdp.

Add an additional Volume to your EC2 Instance

New EC2 Experience X

Tell us what you think

EC2 Dashboard

EC2 Global View

Events

Tags

Limits

Instances

- Instances New
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances New
- Dedicated Hosts
- Scheduled Instances
- Capacity Reservations

Images

- AMIs New
- AMI Catalog

Elastic Block Store

- Volumes New
- Snapshots New
- Lifecycle Manager New

Network & Security

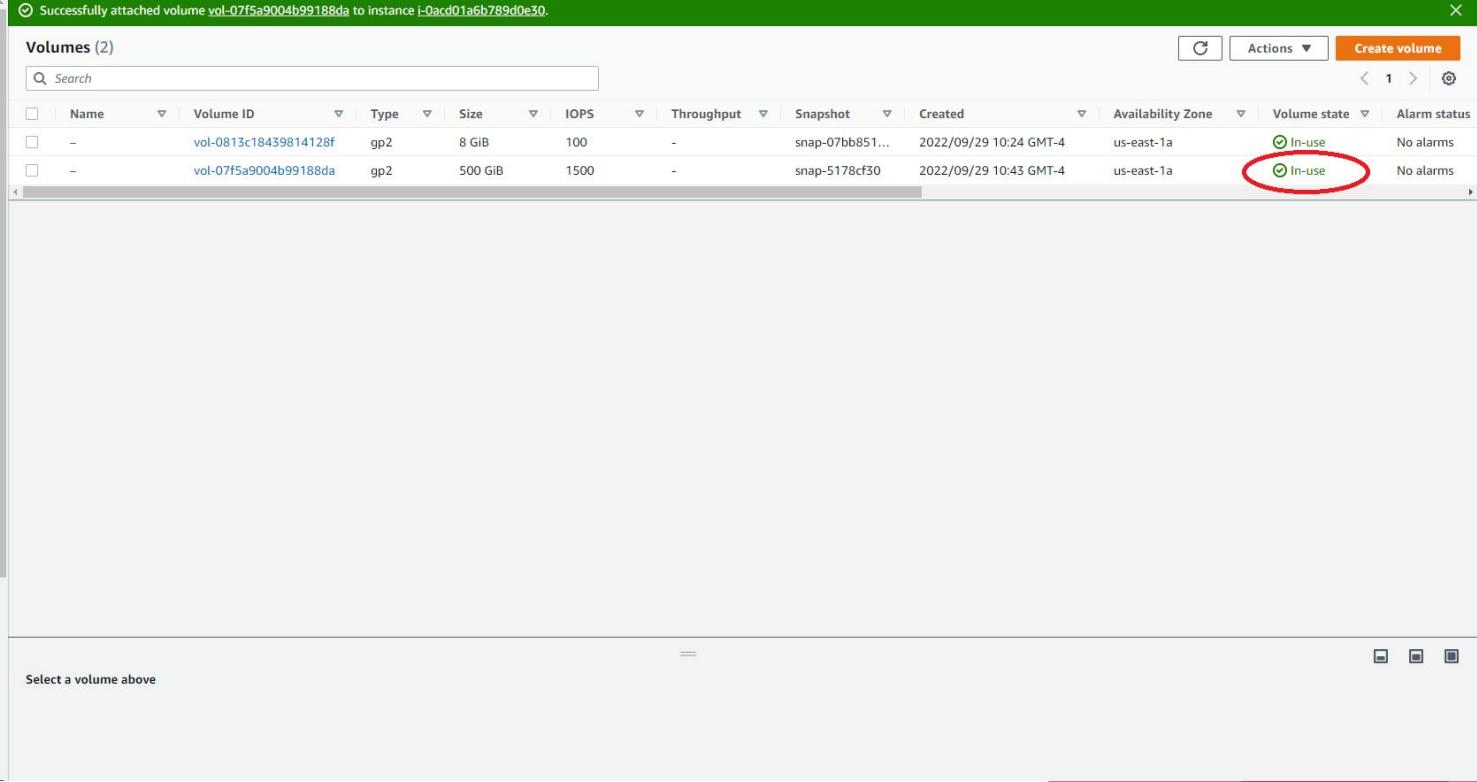
- Security Groups
- Elastic IPs

Volumes (2)

Successfully attached volume vol-07f5a9004b99188da to instance i-0acd01a6b789d0e30.

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created	Availability Zone	Volume state	Alarm status
-	vol-0813c18439814128f	gp2	8 GiB	100	-	snap-07bb851...	2022/09/29 10:24 GMT-4	us-east-1a	In-use	No alarms
-	vol-07f5a9004b99188da	gp2	500 GiB	1500	-	snap-5178cf30	2022/09/29 10:43 GMT-4	us-east-1a	In-use	No alarms

Select a volume above



10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
 - Add an additional Volume - DONE
- Setup Security Groups
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Setup Security Groups

The screenshot shows the AWS Management Console interface for setting up security groups.

Left Sidebar:

- Services: Instances, New, Launch Templates, Spot Requests, Savings Plans, Reserved Instances, Dedicated Hosts, Scheduled Instances, Capacity Reservations.
- Images: AMIs, New, AMI Catalog.
- Elastic Block Store: Volumes, New (circled), Snapshots, New.
- Network & Security: Security Groups, Elastic IPs, Placement Groups, Key Pairs, Network Interfaces.
- Load Balancing: Load Balancers, Target Groups, New.
- Auto Scaling: Launch Configurations, Auto Scaling Groups.

Main Content Area:

Security Groups (1/3) Info

Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules count
-	sg-03cb5fd362cbc1497	ElasticMapReduce-mas...	vpc-09b3d5ce9ddca5b53	Master group for Elasti...	373950817732	18 Permission entries	1 Permission entry
-	sg-07218735e8883cef8	ElasticMapReduce-slave	vpc-09b3d5ce9ddca5b53	Slave group for Elastic ...	373950817732	6 Permission entries	1 Permission entry
<input checked="" type="checkbox"/>	sg-009f243c921d70e1b	default	vpc-09b3d5ce9ddca5b53	default VPC security gr...	373950817732	3 Permission entries	1 Permission entry

Select your “default” security group

Click “Inbound rules” in the window here

sg-009f243c921d70e1b - default

Details Inbound rules Outbound rules Tags

You can now check network connectivity with Reachability Analyzer

Run Reachability Analyzer

Inbound rules (3)

Manage tags Edit inbound rules

Filter security group rules

Name	Security group rule...	IP version	Type	Protocol	Port range	Source	Description

Setup Security Groups

Post Note: If you are getting an issue when adding inbound rules to a default security group, try deleting all of the current rules (Even if they are blank) and then clicking "New rule", this should let you update the rules properly.

EC2 > Security Groups > sg-009f243c921d70e1b - default > Edit inbound rules

Edit inbound rules Info

Inbound rules control the incoming traffic that's allowed to reach the instance.

Add both of these rules.

Inbound rules <small>Info</small>					
Security group rule ID	Type <small>Info</small>	Protocol <small>Info</small>	Port range <small>Info</small>	Source <small>Info</small>	Description - optional <small>Info</small>
sgr-0fc604f2d6166035b	Custom TCP	TCP	9443	My IP	<input type="text"/> MY IP <button>Delete</button>
sgr-01821c8a25be3b992	SSH	TCP	22	My IP	<input type="text"/> MY IP <button>Delete</button>

Add rule Cancel Preview changes

- The first rule will be used later in part B and will let us access Jupyter Notebooks
- The second is needed so that we can access our EC2 or later our EMR Instances

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
 - Add an additional Volume - DONE
- Setup Security Groups - DONE
- Get access and run million_song_prep.py - Convert Data from .h5 to .csv
- *Important Note: Make sure your AWS is using US East (N. Virginia) us-east-1
We need this to get the Million Song Dataset*

Get access to our EC2 Instance

The screenshot shows the AWS EC2 Instances page. On the left, there's a navigation sidebar with sections like EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances (with 'Instances' highlighted), Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances (New), Dedicated Hosts, Scheduled Instances, Capacity Reservations, Images (AMIs New, AMI Catalog), Elastic Block Store (Volumes New, Snapshots New, Lifecycle Manager New), and Network & Security (Security Groups, Elastic IPs). The main content area has a title 'Instances (1) Info' and a search bar. A table lists one instance: 'MyEC2' (Instance ID: i-0acd01a6b789d0e30, Status: Running, Type: t2.micro, Checks: 2/2 passed, Alarms: No alarms, Zone: us-east-1a, Public DNS: ec2-34-229-76-76.com..., Public IP: 34.229.76.76). Below the table is a modal titled 'Select an instance'.

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 ...	Elastic IP
MyEC2	i-0acd01a6b789d0e30	Running	t2.micro	2/2 checks passed	No alarms	us-east-1a	ec2-34-229-76-76.com...	34.229.76.76	-

Get access to our EC2 Instance

New EC2 Experience X
Tell us what you think.

EC2 Dashboard
EC2 Global View
Events
Tags
Limits
Instances
Instances New
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances New
Dedicated Hosts
Scheduled Instances
Capacity Reservations

Images
AMIs New
AMI Catalog

Elastic Block Store
Volumes New
Snapshots New
Lifecycle Manager New

Network & Security
Security Groups
Elastic IPs
Placement Groups

EC2 > Instances > i-0acd01a6b789d0e30

Instance summary for i-0acd01a6b789d0e30 (MyEC2) Info

Updated less than a minute ago

Value	Description
Instance ID	i-0acd01a6b789d0e30 (MyEC2)
Public IPv4 address	34.229.76.76 open address
Private IP4 addresses	172.31.19.128
IPv6 address	-
Instance state	Running
Private IP DNS name (IPv4 only)	ip-172-31-19-128.ec2.internal
Public IPv4 DNS	ec2-34-229-76-76.compute-1.amazonaws.com open address
Hostname type	IP name: ip-172-31-19-128.ec2.internal
Answer private resource DNS name	IPv4 (A)
Instance type	t2.micro
Elastic IP addresses	-
Auto-assigned IP address	34.229.76.76 [Public IP]
VPC ID	vpc-09b3d5ce9ddca5b53
AWS Compute Optimizer finding	Opt-in to AWS Compute Optimizer for recommendations. Learn more
IAM Role	AWSMainRole
Subnet ID	subnet-0793d858c9dbe4441
Auto Scaling Group name	-
Platform	Amazon Linux (Inferred)
AMI ID	ami-026b57f3c383c2eec
Monitoring	disabled
Platform details	Linux/UNIX
AMI name	amzn2-ami-kernel-5.10-hvm-2.0.20220912.1-x86_64-gp2
Termination protection	Disabled
Stop protection	Disabled
Launch time	Thu Sep 29 2022 10:24:35 GMT-0400 (Eastern Daylight Time) (about 1 hour)
AMI location	amazon/amzn2-ami-kernel-5.10-hvm-2.0.20220912.1-x86_64-gp2
Instance auto-recovery	Default
Lifecycle	normal
Stop-hibernate behavior	disabled
AMI Launch index	0
Key pair name	myawskey
State transition reason	-

Connect (highlighted) Instance state ▾ Actions ▾

Details Security Networking Storage Status checks Monitoring Tags

Instance details Info

Platform	Amazon Linux (Inferred)	AMI ID	ami-026b57f3c383c2eec	Monitoring	disabled
Platform details	Linux/UNIX	AMI name	amzn2-ami-kernel-5.10-hvm-2.0.20220912.1-x86_64-gp2	Termination protection	Disabled
Stop protection	Disabled	Launch time	Thu Sep 29 2022 10:24:35 GMT-0400 (Eastern Daylight Time) (about 1 hour)	AMI location	amazon/amzn2-ami-kernel-5.10-hvm-2.0.20220912.1-x86_64-gp2
Instance auto-recovery	Default	Lifecycle	normal	Stop-hibernate behavior	disabled
AMI Launch index	0	Key pair name	myawskey	State transition reason	-

Get access to our EC2 Instance

The screenshot shows the 'Connect to instance' page for an EC2 instance. The instance ID is i-0acd01a6b789d0e30 (MyEC2). The 'SSH client' tab is highlighted with a red circle. Below it, there are four steps to connect:

1. Open an SSH client.
2. Locate your private key file. The key used to launch this instance is myawskey.pem.
3. Run this command, if necessary, to ensure your key is not publicly viewable. Notice the warning about making sure the key is not able to be accessed publicly! (See Slide 32)
chmod 400 myawskey.pem
4. Connect to your instance using its Public DNS:
ec2-34-229-76-76.compute-1.amazonaws.com

Example command:
ssh -i "myawskey.pem" ec2-user@ec2-34-229-76-76.compute-1.amazonaws.com

Note: In most cases, the guessed user name is correct. However, read your AMI usage instructions to check if the AMI owner has changed the default AMI user name.

You'll want to copy this and update the "myawskey.pem" to be the location of your key.

Get access to our EC2 Instance

- Launch either Windows Powershell or Terminal (Mac) and paste the ssh command:

```
[ec2-user@ip-172-31-19-128 ~]# Windows PowerShell  
Copyright (C) Microsoft Corporation. All rights reserved.  
  
Try the new cross-platform PowerShell https://aka.ms/pscore6  
  
PS C:\Users\dpbird> cd ..\Desktop\  
PS C:\Users\dpbird\Desktop> ssh -i "myawskey.pem" ec2-user@ec2-34-229-76-76.compute-1.amazonaws.com  
The authenticity of host 'ec2-34-229-76-76.compute-1.amazonaws.com (34.229.76.76)' can't be established.  
ECDSA key fingerprint is SHA256:g7pONSLltH7Ky4S4z2JQ0C4PkBeoHz904KkcngnBPgo.  
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes  
Warning: Permanently added 'ec2-34-229-76-76.compute-1.amazonaws.com,34.229.76.76' (ECDSA) to the list of known hosts.  
  
_ _ | _ _ / )  
| ( _ _ / Amazon Linux 2 AMI  
_ _ \_\_ | _ |  
  
https://aws.amazon.com/amazon-linux-2/  
[ec2-user@ip-172-31-19-128 ~]$
```

Get access to our EC2 Instance

- Once logged in you will be able to install all of the necessary packages and run the million_song_prep.py file. e.g:
 - pip3 install matplotlib
- Mount your 500gb Volume:
 - lsblk - *Lists the mountable drives so you can find the name of the 500gb drive*
 - mkdir songs - *creates a folder called songs which we mount the drive to*
 - sudo mount /dev/#volumename songs - *Mounts the drive to songs folder*
- A concrete guide to PART A this will be on Friday October 14th

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
 - Add an additional Volume - DONE
- Setup Security Groups - DONE
- Get access - DONE
- **Terminating your EC2 instance and Volume!**
- run million_song_prep.py - Convert Data from .h5 to .csv - Recitation

Terminating your EC2 instance and Volume!

The screenshot shows the AWS EC2 Instances page. On the left, there's a sidebar with various navigation links like EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances, Images, Elastic Block Store, Network & Security, and Discontinued Features. The main area displays a table of instances. A red circle highlights the checkbox next to the instance name 'MyEC2'. Another red circle highlights the 'Instance state' dropdown menu in the top right corner of the table header.

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 IP	Elastic IP
MyEC2	i-0acd01a6b789d0e30	Running	t2.micro	2/2 checks passed	No alarms	us-east-1a	ec2-34-229-76-76.com...	34.229.76.76	-

Instance: i-0acd01a6b789d0e30 (MyEC2)

Details | Security | Networking | Storage | Status checks | Monitoring | Tags

Instance summary

Instance ID i-0acd01a6b789d0e30 (MyEC2)	Public IPv4 address 34.229.76.76 open address	Private IPv4 addresses 172.31.19.128
IPv6 address -	Instance state Running	Public IPv4 DNS ec2-34-229-76-76.compute-1.amazonaws.com open address
Hostname type IP name: ip-172-31-19-128.ec2.internal	Private IP DNS name (IPv4 only) ip-172-31-19-128.ec2.internal	Elastic IP addresses -
Answer private resource DNS name IPv4 (A)	Instance type t2.micro	AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recommendations. Learn more
Auto-assigned IP address 34.229.76.76 [Public IP]	VPC ID vpc-09b3d5ce9ddca5b53	

Terminating your EC2 instance and Volume!

Stopping the Instance is
NOT the same as
Terminating it

Instances (1/1) Info

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
MyEC2	i-0acd01a6b789d0e30	Running	t2.micro	2/2 checks passed	No alarms	us-east-1a

Actions ▾

- Stop instance
- Start instance
- Reboot instance
- Terminate instance

Public IPv4 ... Elastic IP

34.229.76.76 -

Terminate instance

Instance: i-0acd01a6b789d0e30 (MyEC2)

Details Security Networking Storage Status checks Monitoring Tags

Instance summary Info

Instance ID i-0acd01a6b789d0e30 (MyEC2)	Public IPv4 address 34.229.76.76 open address	Private IPv4 addresses 172.31.19.128
IPv6 address -	Instance state Running	Public IPv4 DNS ec2-34-229-76-76.compute-1.amazonaws.com open address
Hostname type IP name: ip-172-31-19-128.ec2.internal	Private IP DNS name (IPv4 only) ip-172-31-19-128.ec2.internal	Elastic IP addresses -
Answer private resource DNS name IPv4 (A)	Instance type t2.micro	AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recommendations. Learn more
Auto-assigned IP address 34.229.76.76 [Public IP]	VPC ID vpc-09b3d5ce9ddca5b53	

Terminating your EC2 instance and Volume!

The screenshot shows the AWS EC2 Instances page. On the left, there's a navigation sidebar with links like EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances (selected), Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances (New), Dedicated Hosts, Scheduled Instances, Capacity Reservations, Images (AMIs New, AMI Catalog), Elastic Block Store (Volumes New, Snapshots New, Lifecycle Manager New), and Network & Security (Security Groups, Elastic IPs, Placement Groups).

The main area displays a table of instances. One instance, "MyEC2" (i-0acd01a6b789d0e30), is selected and shown in more detail below the table. The instance is running, t2.micro, and has 2/2 checks passed.

A modal window titled "Terminate instance?" is open over the instance details. It contains a warning message: "⚠️ On an EBS-backed instance, the default action is for the root EBS volume to be deleted when the instance is terminated. Storage on any local drives will be lost." Below the message, it asks, "Are you sure you want to terminate these instances?". It lists the instance ID "i-0acd01a6b789d0e30 (MyEC2)". There's a "Clean up associated resources" section with a "Delete EBS volumes" link. A confirmation message at the bottom says, "To confirm that you want to terminate the instances, choose the terminate button below. Terminating the instance cannot be undone." At the bottom of the modal are "Cancel" and "Terminate" buttons, with "Terminate" being highlighted with a red circle.

Below the modal, the instance details are shown again, including its state (Running), type (t2.micro), and network information (IPv4 DNS: ec2-34-229-76-76.compute-1.amazonaws.com, Public IP: 34.229.76.76). There are also sections for Hostname type, Answer private resource DNS name, VPC ID, and AWS Compute Optimizer finding.

Terminating your EC2 instance and Volume!

The EC2 instance has now been Terminated. Stopping it is not enough as you will still be charged. Once it is terminated everything on the instance will be lost. So make sure that your data, after processing, is all in the S3 bucket.

Once you delete the instance it will take a little, during that time you will not be able to terminate the 500gb volume we created, as it is still attached. Once the EC2 instance is terminated the option to Delete the volume will become available.

You need to delete your volume separately! It is vitally important that you do not forget this!!!

Terminating your EC2 instance and Volume!

The screenshot shows the AWS EC2 Volumes page. On the left, there's a navigation sidebar with links like EC2 Dashboard, EC2 Global View, Events, Tags, Limits, Instances, Images, Elastic Block Store, Network & Security, and Documentation. The main area displays a table titled "Volumes (1/1)" with one row. The row contains a checkbox (which is checked), a dash icon, the Volume ID "vol-07f5a9004b99188da", the Type "gp2", Size "500 GiB", IOPS "1500", Throughput "-", Snapshot "snap-5178cf30", and Created "2022/09/29 10:43 GMT-4". To the right of the table is a "Actions" menu with options: Modify volume, Create snapshot, Create snapshot lifecycle policy, Delete volume (which is highlighted with a red circle), Attach volume, Detach volume, Force detach volume, Manage auto-enabled I/O, and Manage tags.

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created
-	vol-07f5a9004b99188da	gp2	500 GiB	1500	-	snap-5178cf30	2022/09/29 10:43 GMT-4

Actions ▾

- Modify volume
- Create snapshot
- Create snapshot lifecycle policy
- Delete volume**
- Attach volume
- Detach volume
- Force detach volume
- Manage auto-enabled I/O
- Manage tags

Volume ID: vol-07f5a9004b99188da

Details | Status checks | Monitoring | Tags

Details

Terminating your EC2 instance and Volume!

The screenshot shows the AWS EC2 Volumes page. A modal dialog box titled "Delete vol-07f5a9004b99188da?" is displayed. The dialog contains a warning message: "After you delete a volume, its data is permanently deleted and the volume can no longer be attached to an instance." Below the message is a question: "Are you sure that you want to delete vol-07f5a9004b99188da?". At the bottom right of the dialog, there are two buttons: "Cancel" and a red-highlighted "Delete" button.

New EC2 Experience X

Tell us what you think

EC2 Dashboard

EC2 Global View

Events

Tags

Limits

Instances

- Instances New
- Instance Types
- Launch Templates
- Spot Requests
- Savings Plans
- Reserved Instances New
- Dedicated Hosts
- Scheduled Instances
- Capacity Reservations

Images

- AMIs New
- AMI Catalog

Elastic Block Store

- Volumes New
- Snapshots New
- Lifecycle Manager New

Network & Security

- Security Groups
- Elastic IPs
- Placement Groups

Volumes (1/1)

Name	Volume ID	Type	Size	IOPS	Throughput	Snapshot	Created	Availability Zone	Volume state	Alarm status
-	vol-07f5a9004b99188da	gp2	500 GiB	1500	-	snap-5178cf30	2022/09/29 10:43 GMT-4	us-east-1a	Available	No alarms

Actions ▼ Create volume

Volume ID: vol-07f5a9004b99188da

Details Status checks Monitoring Tags

Details

Delete vol-07f5a9004b99188da?

After you delete a volume, its data is permanently deleted and the volume can no longer be attached to an instance.

Are you sure that you want to delete vol-07f5a9004b99188da?

Cancel Delete

One more time: Make sure your converted data is all in the S3 bucket before you delete your volume and terminate your EC2 instance.

10-605 Setup for Homework Part A

- Setup S3 Bucket - DONE
- Setup IAM Roles - DONE
- Make a keypair - DONE
- Setup EC2 Instance - DONE
 - Add an additional Volume - DONE
- Setup Security Groups - DONE
- Get access - DONE
- **Terminating your EC2 instance and Volume! - DONE**
- run million_song_prep.py - Convert Data from .h5 to .csv - Recitation

10-605 Setup for Homework Part B

- Setup EMR Cluster
- Log into Master Node
- Get Access to Jupyter Notebooks Interface

Setup EMR Cluster

- What is EMR? (Elastic MapReduce)
- It is a managed cluster platform typically used for large data services
- It lets us set up a master node and worker nodes similar to what we have been discussing in the course, allowing us to run Hadoop and Apache Spark to work with large datasets!

Setup EMR Cluster

The screenshot shows the AWS Management Console search results for the term 'emr'. The search bar at the top contains 'emr'. The left sidebar shows navigation links for EC2 services like EC2 Dashboard, Instances, and Images. The main search results are categorized into 'Services' and 'Features'.

Services

- EMR** ☆ Managed Hadoop Framework
- AWS Glue DataBrew** ☆ Visual data preparation tool to clean and normalize data for analytics and machine learning
- MediaStore** ☆ Store and deliver video assets for live or on-demand media workflows
- MediaConvert** ☆ Convert file-based content for broadcast and multiscreen delivery

See all 9 results ▶

Features

- Private registry**
Elastic Container Registry feature
- Repositories**
Elastic Container Registry feature

Account attributes

- Supported platforms
- VPC
- Default VPC
- Regions
- EBS encryption
- Zones
- EC2 Serial Console
- Default credit specification
- Console experiments

Explore AWS

- Enable Best Price-Performance with AWS Graviton2 powered EC2 instances for up to 40% better price performance for a broad range of cloud workloads. [Learn more](#)
- Save Up to 45% on ML Inference with EC2 Inf1 instances provide high performance

Setup EMR Cluster

Amazon EMR

EMR Studio

EMR Serverless  New

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

 **EMR Serverless** is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless](#) 

Create cluster  View details  Clone  Terminate 

Filter: All clusters  Filter clusters ...  1 cluster (all loaded) 

	Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
<input type="checkbox"/>  Clusty		j-2BPSW6DGU1Y1K	Terminated User request	2022-09-28 10:43 (UTC-4)	17 minutes	16

Setup EMR Cluster

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless.](#)

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

General Configuration

Cluster name: My cluster

Logging

S3 folder: s3://aws-logs-373950817732-us-east-1/elasticmap

Launch mode: Cluster Step execution

Software configuration

Release: emr-5.36.0

Applications:

- Core Hadoop: Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- Presto: Presto 0.267 with Hadoop 2.10.1 HDFS and Hive 2.3.9 Metastore
- Spark: Spark 2.4.8 on Hadoop 2.10.1 YARN and Zeppelin 0.10.0

Use AWS Glue Data Catalog for table metadata

Hardware configuration

Setup EMR Cluster

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless](#).

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release emr-5.36.0

- Hadoop 2.10.1
- JupyterHub 1.4.1
- Ganglia 3.7.2
- Hive 2.3.9
- JupyterEnterpriseGateway 2.1.0
- Mahout 0.13.0
- Oozie 5.2.1
- TensorFlow 2.4.1

- Zeppelin 0.10.0
- Tez 0.9.2
- HBase 1.4.13
- Presto 0.267
- MXNet 1.8.0
- Hue 4.10.0
- Spark 2.4.8

We just want to select:

- Hadoop 2.10.1
- JupyterHub 1.4.1
- Spark 2.4.8

No other choices should be selected

Multiple master nodes (optional)

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

AWS Glue Data Catalog settings (optional)

- Use for Spark table metadata

Edit software settings

- Enter configuration
- Load JSON from S3

```
"Classification": "spark",
"Properties": {
  "maximizeResourceAllocation": "true"
}
```

In the edit software setting we must add the following:

Edit Software Setting

- Configuring spark:
<https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark-configure.html>
- Add:

```
[  
  {  
    "Classification" : "spark",  
    "Properties" : {  
      "maximizeResourceAllocation" : "true"  
    }  
  }  
]
```

Hit Next to get to step 2:

Setup EMR Cluster

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless](#).

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration [i](#)

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

Cluster Composition

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

Instance group configuration

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

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

Networking

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Launch the cluster into a VPC with a public, private or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Launch the cluster into a VPC with a public, private, or

Nothing to do here, just scroll down.

Setup EMR Cluster

Screenshot of the AWS EMR Cluster Nodes and Instances configuration page.

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. [Learn more about instance purchasing options](#)

Console options for automatic scaling have changed. [Learn more](#)

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	1 Instances	<input checked="" type="radio"/> On-demand Details <input type="radio"/> Spot Details Use on-demand as max price
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	1 Instances	<input checked="" type="radio"/> On-demand Details <input type="radio"/> Spot Details Use on-demand as max price
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB	0 Instances	<input checked="" type="radio"/> On-demand Details <input type="radio"/> Spot Details Use on-demand as max price

+ Add task instance group

Total core and task units: 1 Total units

While you are still developing your notebook I recommend using just 1 Core instance, you can increase this once you have worked it all through on the subsample of the data and you want to run it on the entire Million song dataset. Maybe 2-4, be sure to keep an eye on your spending

Setup EMR Cluster

Task - 3

4 vCore, 16 GiB memory, EBS only storage
EBS Storage: 64 GiB
[Add configuration settings](#)

Instances 0 Instances

Spot
[Use on-demand as max price](#)

[+ Add task instance group](#)

Total core and task units 1 Total units

Cluster scaling

Adjust the number of Amazon EC2 instances available to an EMR cluster via EMR-managed scaling or a custom automatic scaling policy. [Learn more](#)

Cluster scaling Enable Cluster Scaling

Auto-termination

Select a time to have the cluster terminate after the cluster becomes idle. Choose a minimum of 1 minute or a max of 24 hours. [Learn more](#)

Auto-termination Enable auto-termination

EBS Root Volume

Specify the root device volume size up to 100 GiB. This sizing applies to all instances in the cluster. [Learn more](#)

Root device EBS volume size GiB

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

I recommend turning Auto-termination off. If you leave your cluster idle for the specified amount of time it will automatically terminate the cluster, if it terminates before you download your ipynb then you will lose it.

Setup EMR Cluster

aws Services Search for services, features, blogs, docs, and more [Alt+S]

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

General Options

Cluster name Name your cluster

Logging Disable Logging

Termination protection Disable Termination Protection

Tags

Key	Value (optional)
Add a key to create a tag	

Additional Options

EMRFS consistent view

Operating System Options

Amazon Linux Release 2.0.20220426.0

Custom AMI ID Enter an AMI ID
 Update all installed packages on reboot (recommended)

▶ Bootstrap Actions

Hit Next to get to step 4:

Setup EMR Cluster

Screenshot of the AWS Create Cluster - Advanced Options page, Step 4: Security.

The page shows the following sections:

- Security Options:** Includes an EC2 key pair dropdown (set to "myawskey") and a checkbox for "Cluster visible to all IAM users in account". A red circle highlights the EC2 key pair dropdown.
- Permissions:** Includes "Default" and "Custom" radio buttons. A note says "Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates." Below are dropdowns for "EMR role" (set to "EMR_DefaultRole"), "EC2 instance profile" (set to "EMR_EC2_DefaultRole"), and "Auto Scaling role" (set to "EMR_AutoScaling_DefaultRole").
- EC2 security groups:** A section titled "Expand the EC2 Security Groups Menu" contains two dropdowns: "Master" (set to "Default: sg-03cb5fd362cbc1497 (ElasticMapReduce)") and "Core & Task" (set to "Default: sg-07218735e8883cef8 (ElasticMapReduce)"). A red circle highlights the "Master" dropdown. A note below says: "An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#)".
- Additional security groups:** A table showing "Additional security groups" with two rows: "Master" and "Core & Task", both with "No security groups selected". A red circle highlights the "Master" row. A note to the right says: "We need to add our "default" security group to be additional security groups on both the Master and the Core & Task".

Page navigation and footer:

- Top bar: AWS logo, Services, Search for services, features, blogs, docs, and more [Alt+S], Notifications, N. Virginia, dpbird.
- Bottom footer: Feedback, Looking for language selection? Find it in the new Unified Settings, © 2022, Amazon Web Services, Inc. or its affiliates., Privacy, Terms, Cookie preferences.

Setup EMR Cluster

Screenshot of the AWS Create Cluster - Advanced Options interface, showing the Security Options step.

The EC2 key pair dropdown is set to "myawskey". The checkbox "Cluster visible to all IAM users in account" is checked.

The "Assign additional security groups to master node" dialog is open, showing one selected security group:

Group ID	Group Name	Description
sg-009f243c921d70e1b	default	default VPC security group

The "Assign security groups" button is highlighted with a red circle and the number 8.

Below the dialog, the cluster configuration shows:

Type	EMR managed security groups	Additional security groups
Master	Default: sg-03cb5fd362cbc1497 (ElasticMapReduce)	No security groups selected
Core & Task	Default: sg-07218735e8883cef8 (ElasticMapReduce)	No security groups selected

A note at the bottom left says: "managed security groups in order to launch a cluster. [Learn more](#)".

Feedback: Looking for language selection? Find it in the new Unified Settings. © 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preference

Setup EMR Cluster

Screenshot of the AWS EMR Cluster Setup Step 4: Security page.

The page shows security options and permissions for the cluster. The "EC2 key pair" is set to "myawskey". The "Cluster visible to all IAM users in account" checkbox is checked. Under "Permissions", the "Default" radio button is selected. It notes that default IAM roles will be automatically created if not present. The "EMR role" is set to "EMR_DefaultRole". The "EC2 instance profile" and "Auto Scaling role" are also listed.

The "EC2 security groups" section is expanded, showing two entries:

Type	EMR managed security groups	Additional security groups
Master	Default: sg-03cb5fd362cbc1497 (ElasticMapRed...	sg-009f243c921d70e1b (default)
Core & Task	Default: sg-07218735e8883cef8 (ElasticMapRed...	sg-009f243c921d70e1b (default)

A red circle highlights the "Additional security groups" row for the Master group, with an arrow pointing to the text "DONE!".

At the bottom right, the "Create cluster" button is highlighted with a red circle.

Page footer: Feedback, Looking for language selection? Find it in the new Unified Settings, © 2022, Amazon Web Services, Inc. or its affiliates., Privacy, Terms, Cookie preferences.

Setup EMR Cluster

AWS Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia dpbird

Amazon EMR

EMR Studio

EMR Serverless New

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Clone Terminate AWS CLI export

Cluster: My EMR Cluster Starting It will take a few minutes to setup! Once it's ready this will stay "Waiting"

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-3MK1GJSC9OQS4
Creation date: 2022-09-29 14:25 (UTC-4)
Elapsed time: 0 seconds
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: --

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: JupyterHub 1.4.1, Spark 2.4.8
Log URI: --
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20220426.0 Learn more

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: --

Network and hardware

Availability zone: --
Subnet ID: subnet-032f2fdf83b62a1b9
Master: Provisioning 1 m5.xlarge
Core: Provisioning 1 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: myawskey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

Setup EMR Cluster

AWS Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia dpbird

Amazon EMR Cluster: My EMR Cluster Waiting Cluster ready to run steps.

Clone Terminate AWS CLI export

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-3MK1GJSC9OQS4
Creation date: 2022-09-29 14:25 (UTC-4)
Elapsed time: 7 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-54-172-237-153.compute-1.amazonaws.com Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: JupyterHub 1.4.1, Spark 2.4.8
Log URI: --
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20220426.0 Learn more

Application user interfaces

Persistent user interfaces: Spark history server, YARN timeline server
On-cluster user interfaces: Not Enabled Enable an SSH Connection

Network and hardware

Availability zone: us-east-1b
Subnet ID: subnet-032f2fdf83b62a1b9
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 1 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: myawskey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

10-605 Setup for Homework Part B

- Setup EMR Cluster - DONE
- Log into Master Node
- Get Access to Jupyter Notebooks Interface

Log into Master Node

- Now the cluster is up we will log into the Master node to install some packages which we will need in the notebook.
- Open up Windows Powershell or Terminal (Mac)

Log into Master Node

AWS Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia dpbird

Amazon EMR Clusters Notebooks Git repositories Security configurations Block public access VPC subnets Events EMR on EKS Virtual clusters Help What's new

Clone Terminate AWS CLI export

Cluster: My EMR Cluster Waiting Cluster ready to run steps.

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-3MK1GJSC9OQS4
Creation date: 2022-09-29 14:25 (UTC-4)
Elapsed time: 7 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-54-172-227-152.compute-1.amazonaws.com Click here to Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: JupyterHub 1.4.1, Spark 2.4.8
Log URI: --
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20220426.0 [Learn more](#)

Application user interfaces

Persistent user interfaces: Spark history server, YARN timeline server
On-cluster user Not Enabled Enable an SSH Connection interfaces:

Network and hardware

Availability zone: us-east-1b
Subnet ID: [subnet-032f2fdf83b62a1b9](#)
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 1 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: myawskey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole

Log into Master Node

The screenshot shows the AWS EMR console interface. On the left, there's a sidebar with navigation links like Amazon EMR, EMR Studio, EMR Serverless, EMR on EC2, Clusters, Notebooks, Git repositories, Security configurations, Block public access, VPC subnets, Events, EMR on EKS, Virtual clusters, Help, and What's new. The main area shows a cluster named "My EMR Cluster" in the "Waiting" state. A modal window titled "SSH" is open, providing instructions for connecting to the master node using SSH. It offers two options: "Windows" and "Mac / Linux". The "Mac / Linux" option is selected and highlighted with a red circle. Below it, a note says "Click Mac/Linux here (Even if you are using Windows)". The "Windows" tab is also circled in red. The modal contains three steps: 1. Open a terminal window. 2. Establish a connection using the command "ssh -i ~/myawskey.pem hadoop@ec2-54-172-237-153.compute-1.amazonaws.com". This step is circled in red, and a callout arrow points to a "Copy this to your terminal" button. 3. Type "yes" to dismiss the security warning. At the bottom of the modal, there's a "Close" button and status information: "Cluster scaling: Not enabled" and "Auto-termination: Not enabled". The footer of the page includes links for Feedback, Unified Settings, and various AWS services like S3, Lambda, and CloudWatch.

Amazon EMR

EMR Studio

EMR Serverless New

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Services

Search for services, blogs, docs, and more

[Alt+S]

Clone Terminate AWS CLI export

Cluster: My EMR Cluster Waiting Cluster ready to run steps.

SSH

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. [Learn more](#)

Windows **Mac / Linux**

Click Mac/Linux here (Even if you are using Windows)

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/myawskey.pem with the location and filename of the private key file (.pem) used to launch the cluster.

`ssh -i ~/myawskey.pem hadoop@ec2-54-172-237-153.compute-1.amazonaws.com`

3. Type yes to dismiss the security warning.

Cluster scaling: Not enabled

Auto-termination: Not enabled

Security and access

Key name: myawskey

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Feedback Looking for language selection? Find it in the new [Unified Settings](#)

© 2022, Amazon Web Services, Inc. or its affiliates.

Privacy Terms Cookie preferences

Log into Master Node

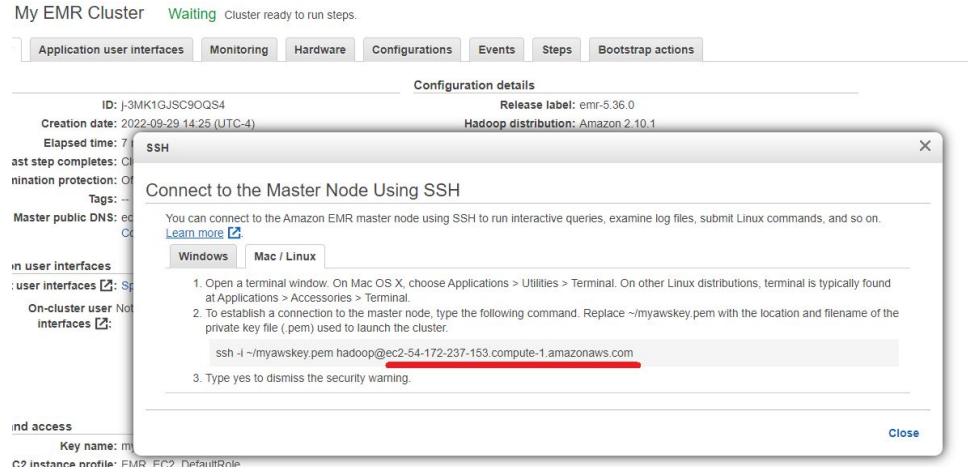
```
hadoop@ip-172-31-42-105:~  
Windows PowerShell  
Copyright (C) Microsoft Corporation. All rights reserved.  
Try the new cross-platform PowerShell https://aka.ms/pscore6  
  
S C:\Users\dpbird> cd ..\Desktop  
S C:\Users\dpbird\Desktop> ssh -i myawskey.pem hadoop@ec2-54-172-237-153.compute-1.amazonaws.com  
The authenticity of host 'ec2-54-172-237-153.compute-1.amazonaws.com (54.172.237.153)' can't be established.  
ECDSA key fingerprint is SHA256:Ndsxgnd1Dwv8Z3vITeFeuG/bgvWljlL7z3KhcC6J4.  
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes  
Warning: Permanently added 'ec2-54-172-237-153.compute-1.amazonaws.com,54.172.237.153' (ECDSA) to the list of known hosts.  
Last login: Thu Sep 29 18:48:19 2022  
_ _| _ _|_)  
_ | ( _ / Amazon Linux 2 AMI  
_ | \_\_|_ |  
  
https://aws.amazon.com/amazon-linux-2/  
0 package(s) needed for security, out of 42 available  
Run "sudo yum update" to apply all updates.  
  
EEEEEEEEEEEEEE MMMMMMM M:::::M R:::::RRRRRRRRRRRRR  
E:::::::E:::::E M:::::::M M:::::::M R:::::RRRRRRR:::::R  
E:::::E EEEEE M:::::::M M:::::::M RR:::::R R:::::R  
E:::::E:::::E M:::::::M M:::::::M M:::::::M R:::::RRRRRRR:::::R  
E:::::E:::::E M:::::::M M:::::::M M:::::::M R:::::RRRRRRR:::::R  
E:::::E:::::E M:::::::M M:::::::M M:::::::M R:::::RRRRRRR:::::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  
E:::::::E:::::E M:::::::M M:::::::M R:::::R R:::::R  
EEEEEEEEEEEEEE MMMMMMM RRRRRRR  
hadoop@ip-172-31-42-105 ~]$
```

Log into Master Node

- You will want to **sudo** install matplotlib
 - sudo pip3 install matplotlib - *you need to use sudo so that the workers have access to it*
- Then you can log out.

10-605 Setup for Homework Part B

- Setup EMR Cluster - DONE
- Log into Master Node - DONE
- Get Access to Jupyter Notebooks Interface
 - Before we start this we want to grab this from the cluster:



So mine is:
ec2-54-172-237-153.compute-1.amazonaws.com

Get Access to Jupyter Notebooks Interface

- The Jupyter Notebook Interface can only be accessed via SSH tunneling.

The screenshot shows the AWS EMR Cluster Overview page for a cluster named "My EMR Cluster" which is currently "Waiting". The "Application user interfaces" tab is selected, highlighted with a red circle. Below it, the "On-cluster application user interfaces" section lists several services with their corresponding User Interface URLs:

Application	User Interface URL	Status
HDFS Name Node	http://ec2-54-172-237-153.compute-1.amazonaws.com:50070/	SSH tunnel not enabled
JupyterHub	https://ec2-54-172-237-153.compute-1.amazonaws.com:9443/	SSH tunnel not enabled
Spark History Server	http://ec2-54-172-237-153.compute-1.amazonaws.com:18080/	SSH tunnel not enabled
Resource Manager	http://ec2-54-172-237-153.compute-1.amazonaws.com:8088/	SSH tunnel not enabled

A red arrow points from the text "We want to access this!" to the "JupyterHub" URL. The entire screenshot is set against a background gradient of blue and pink.

Get Access to Jupyter Notebooks Interface

EC2 > Security Groups > sg-009f243c921d70e1b - default > Edit inbound rules

Edit inbound rules Info

Inbound rules control the incoming traffic that's allowed to reach the instance.

Inbound rules <small>Info</small>	Type <small>Info</small>	Protocol <small>Info</small>	Port range <small>Info</small>	Source <small>Info</small>	Description - optional <small>Info</small>	
sgr-0fc604f2d6166035b	Custom TCP	TCP	9443	My IP	<input type="text"/> <small>Search</small>	<input type="button" value="Delete"/>
sgr-01821c8a25be3b992	SSH	TCP	22	My IP	<input type="text"/> <small>Search</small>	<input type="button" value="Delete"/>

Add rule

Recall back when we set up our Security Group's inbound rules we add this port. This is going to allow us to tunnel to the Jupyter Notebooks interface

Add rule Cancel Preview changes Save rules

Get Access to Jupyter Notebooks Interface

- Remember 3 slides ago when I told you to grab
 - `ec2-54-172-237-153.compute-1.amazonaws.com`
- From your terminal:
 - `ssh -i myawskey.pem -N -L 1232:ec2-54-172-237-153.compute-1.amazonaws.com:9443`
`hadoop@ec2-54-172-237-153.compute-1.amazonaws.com`
- The `1232` will be my localhost number for later, you can make this any 4 digit number.
- The `9443` is the port number of the JupyterHub

Get Access to Jupyter Notebooks Interface

```
Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\dpbird> cd .\Desktop\
PS C:\Users\dpbird\Desktop> ssh -i myawskey.pem -N -L 1232:ec2-54-172-237-153.compute-1.amazonaws.com:9443 hadoop@ec2-54-172-237-153.compute-1.amazonaws.com
```

- This will appear like it has frozen, but it hasn't it is holding the tunnel for you.
- Now you can head to
 - <https://localhost:1232/>
- It will warn you that your connection is not private. You want to continue to localhost.

Note: Some users' computers do not automatically add the https portion, make sure you include it!

Get Access to Jupyter Notebooks Interface

⚠ Not secure | <https://localhost:1232>



This is what I see in microsoft Edge, other browsers may vary. Some students with Chrome found success clicking on the page and typing "thisisunsafe" without the "" just into the page.



Your connection isn't private

Attackers might be trying to steal your information from **localhost** (for example, passwords, messages, or credit cards).

NET::ERR_CERT_AUTHORITY_INVALID

[Hide advanced](#)

[Go back](#)

This server couldn't prove that it's **localhost**; its security certificate is not trusted by your computer's operating system. This may be caused by a misconfiguration or an attacker intercepting your connection.

[Continue to localhost \(unsafe\)](#)

Get Access to Jupyter Notebooks Interface

← → C | Not secure | <https://localhost:1232/hub/login?next=%2Fhub%2F>

jupyterhub



The screenshot shows a web browser window with the address bar indicating a non-secure connection (Not secure) and the URL <https://localhost:1232/hub/login?next=%2Fhub%2F>. The title bar says "jupyterhub". The main content is a "Sign in" form with an orange header. It has two input fields: "Username" and "Password", both with placeholder text. Below the fields is an orange "Sign in" button.

- Once through you will be in JupyterHub
- Username: jovyan
- Password: jupyter

Get Access to Jupyter Notebooks Interface



The screenshot shows the JupyterHub interface with a red circle highlighting the 'Upload' button in the top right corner of the file list area.

Files

Select items to perform actions on them.

	Name	Last Modified	File size
<input type="checkbox"/>	jupyterhub-proxy.pid	an hour ago	2 kB
<input type="checkbox"/>	jupyterhub.sqlite	seconds ago	102 kB
<input type="checkbox"/>	jupyterhub_cookie_secret	an hour ago	65 B

Logout Control Panel

Upload New

- Here you can upload your hw4.ipynb handout.
- **IMPORTANT:** Before you close your terminal or terminate your EMR Cluster make sure you download your hw4.ipynb otherwise you will lose all changes!!!

Get Access to Jupyter Notebooks Interface

⚠ Not secure | <https://localhost:1232/user/jovyan/notebooks/hw4.ipynb>

The screenshot shows a Jupyter Notebook interface with the following elements:

- Header:** "jupyterhub hw4 Last Checkpoint: a few seconds ago (unsaved changes)".
- Kernel Status:** "Python 3" (highlighted with a red circle).
- Kernel Menu:** A dropdown menu from the "Kernel" button in the top navigation bar. It includes options like "Interrupt", "Restart", "Restart & Clear Output", "Restart & Run All", "Reconnect", "Shutdown", and "Change kernel". The "PySpark" option is highlighted with a red arrow pointing to it.
- Notebook Content:** A section titled "CMU Machine Learning Homework 4 - Part A: Working with Large Datasets at Scale". It contains text about data loading and a note about grading.
- Code Cell:** An "In []:" cell containing Python code for starting a PySpark session and importing necessary libraries.
- Text Block:** A note at the bottom stating that Matplotlib and other useful Python libraries are not pre-installed on the cluster.

Get Access to Jupyter Notebooks Interface

The screenshot shows a Jupyter Notebook interface with the following details:

- Title Bar:** jupyterhub hw4old Last Checkpoint: 2 minutes ago (unsaved changes)
- Toolbar:** File, Edit, View, Insert, Cell, Kernel, Widgets, Help
- Cell Toolbar Buttons:** In, +, %, Run, Cell, Code
- Kernel Selection:** Trusted PySpark (highlighted with a red circle and arrow)
- Message:** All set!
- Section Headers:** CMU Machine Learning with Large Datasets, Homework 4 - Machine Learning at Scale
- Note:** Note that we've included code that does data loading and preparation for you. You could take a brief look to learn about how to specify a schema when loading data, or just run them all and start from "Part B Begins" after adding your S3 bucket name to cmd 8.
- Text:** Note that we will not be autograding this notebook because of the open-ended nature of it (although you will have to submit this notebook). To make grading easier and to learn about your thought process, throughout the notebook, we include questions you have to answer in your writeup. Whenever this happens, there is a ★ symbol.
- Section 0:** Start a Spark Session and Install Libraries
- In [1]:**

```
# You are highly recommended to select the "PySpark" kernel instead of python kernel,
# Otherwise you need to modify this cell to get pyspark working.

from pyspark.sql import *

sc = spark.sparkContext
print(f'num executors: {sc.getConf().get("spark.executor.instances")}')
```
- Output:** Starting Spark application
- Table:** YARN Application ID, Kind, State, Spark UI, Driver log, User, Current session?

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
1	application_1664896745988_0002	pyspark	idle	Link	Link	None	✓

- Text:** SparkSession available as 'spark'.

Get Access to Jupyter Notebooks Interface

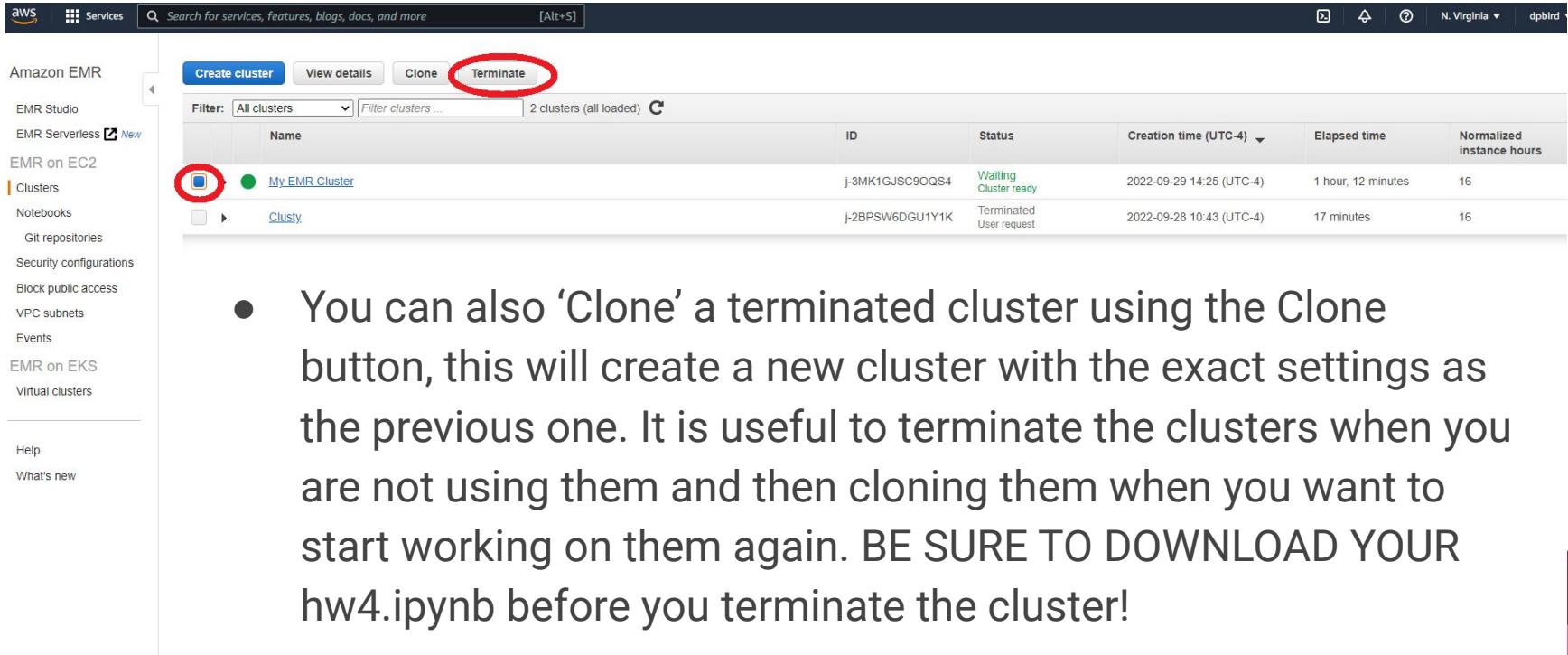
```
In [ ]: #-----  
#Set to False after writing your code and run end-to-end on full data  
#-----  
load_subset = True  
#-----  
#Add your s3 bucket name below:  
#-----  
s3_bucket_name = ''  
  
if load_subset:  
    df = spark.read.format("csv")\  
        .option("header", "false")\  
        .option("nanValue", "nan")\  
        .option("quote", "\")\  
        .option("escape", "\\")\  
        .schema(schema)\  
        .load(f"s3://{{s3_bucket_name}}/processed/A_1.csv")  
else:  
    df = spark.read.format("csv")\  
        .option("header", "false")\  
        .option("nanValue", "nan")\  
        .option("quote", "\")\  
        .option("escape", "\\")\  
        .schema(schema)\  
        .load(f"s3://{{s3_bucket_name}}/processed/*.csv")  
  
print('loaded {} records'.format(df.count()))
```

You are going to want to add your S3 bucket name here, so that it can get the data!

10-605 Setup for Homework Part B

- Setup EMR Cluster - DONE
- Log into Master Node - DONE
- Get Access to Jupyter Notebooks Interface - DONE
- Terminating and Cloning EMR Clusters

Terminating and Cloning EMR Clusters



The screenshot shows the AWS EMR console interface. At the top, there's a navigation bar with the AWS logo, a search bar, and a services menu. Below the navigation bar, the left sidebar lists various EMR-related services like Studio, Serverless, and EC2. The main content area displays a table of clusters. The table has columns for Name, ID, Status, Creation time (UTC-4), Elapsed time, and Normalized instance hours. Two clusters are listed: 'My EMR Cluster' (ID: J-3MK1GJSC9OQS4, Status: Waiting Cluster ready) and 'Clusty' (ID: J-2BPSW6DGU1Y1K, Status: Terminated User request). A red circle highlights the 'Terminate' button at the top of the table header. Another red circle highlights the cluster ID 'J-3MK1GJSC9OQS4' for 'My EMR Cluster'.

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
My EMR Cluster	J-3MK1GJSC9OQS4	Waiting Cluster ready	2022-09-29 14:25 (UTC-4)	1 hour, 12 minutes	16
Clusty	J-2BPSW6DGU1Y1K	Terminated User request	2022-09-28 10:43 (UTC-4)	17 minutes	16

- You can also ‘Clone’ a terminated cluster using the Clone button, this will create a new cluster with the exact settings as the previous one. It is useful to terminate the clusters when you are not using them and then cloning them when you want to start working on them again. BE SURE TO DOWNLOAD YOUR hw4.ipynb before you terminate the cluster!

Finally: Deleting your S3 Buckets

- S3 Buckets are much cheaper than the volumes we created for the EC2 instance so keeping these running the whole duration of the homework isn't costly. I recommend only deleting these once you have concluded the homework, this way you don't have to run Part A everytime you want to continue Part B.

Finally: Deleting your S3 Buckets

Amazon S3 X

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose [Provide feedback](#).

Earn an AWS Learning Badge to showcase your knowledge of S3.

Amazon S3 > Buckets

Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (2) [Info](#)

Buckets are containers for data stored in S3. Learn more

Find buckets by name

Name	AWS Region	Access	Creation date
<input checked="" type="radio"/> danielbirdbucket	US East (N. Virginia) us-east-1	Bucket and objects not public	September 28, 2022, 13:03:01 (UTC-04:00)
<input type="radio"/> hw-4605	US East (N. Virginia) us-east-1	Bucket and objects not public	October 21, 2021, 09:26:58 (UTC-04:00)

[View Storage Lens dashboard](#)

[Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight:

AWS Marketplace for S3

Name	AWS Region	Access	Creation date
<input checked="" type="radio"/> danielbirdbucket	US East (N. Virginia) us-east-1	Bucket and objects not public	September 28, 2022, 13:03:01 (UTC-04:00)
<input type="radio"/> hw-4605	US East (N. Virginia) us-east-1	Bucket and objects not public	October 21, 2021, 09:26:58 (UTC-04:00)

Finally: Deleting your S3 Buckets



ⓘ We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose [Provide feedback](#).

Amazon S3 > Buckets > danielbirdbucket > Delete bucket

Delete bucket Info



- Deleting a bucket cannot be undone.
- Bucket names are unique. If you delete a bucket, another AWS user can use the name.

[Learn more](#)

Delete bucket "danielbirdbucket"?

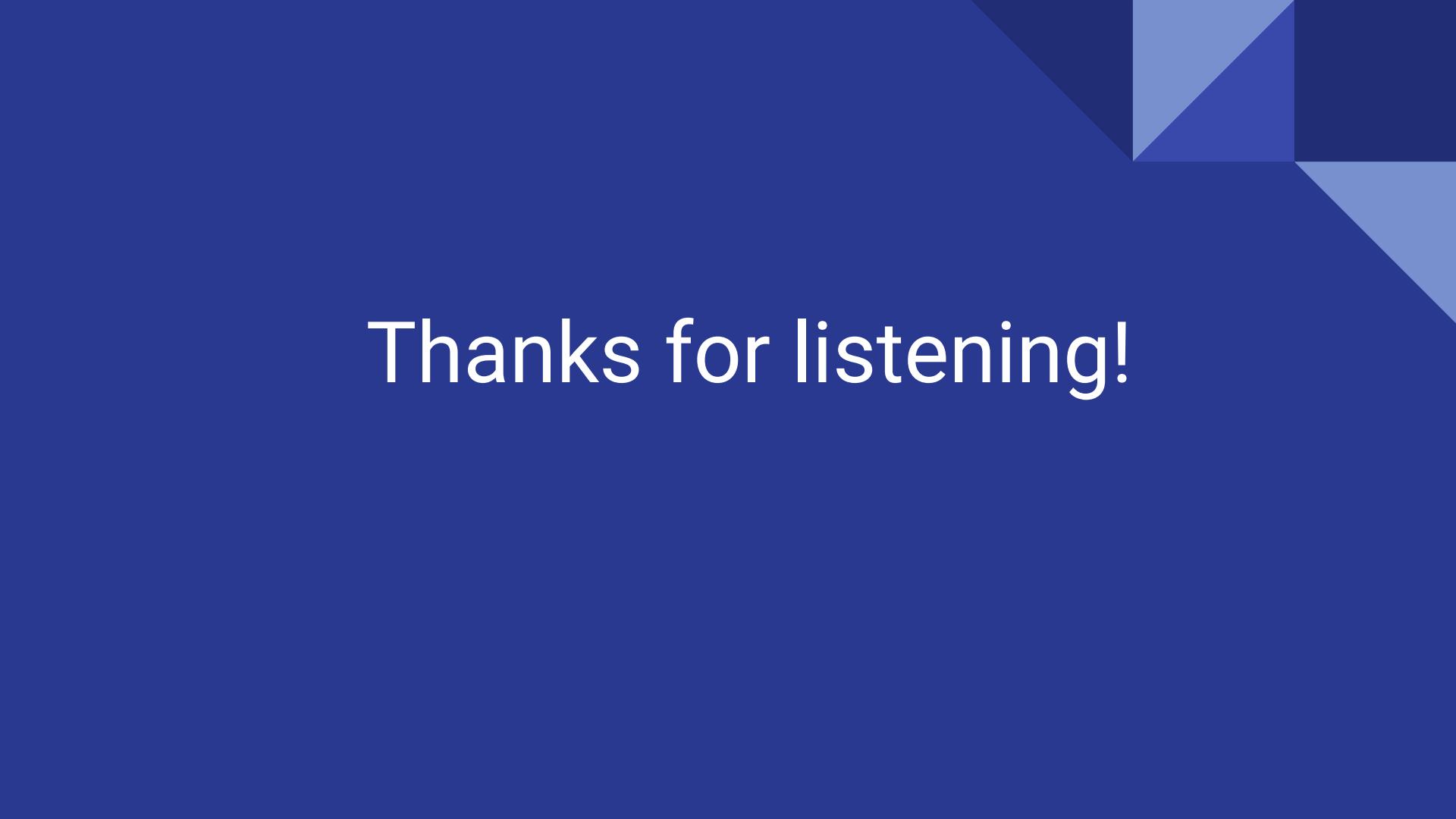
To confirm deletion, enter the name of the bucket in the text input field.

danielbirdbucket

You need to type the name of the bucket
in order to be able to delete it

Cancel

Delete bucket



Thanks for listening!