CSE 4510/5310
Big Data

Instructor: Fitzroy Nembhard, Ph.D.

Setting up AWS S3, Loading Data into a Bucket and Performing Operations on the Data

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY®

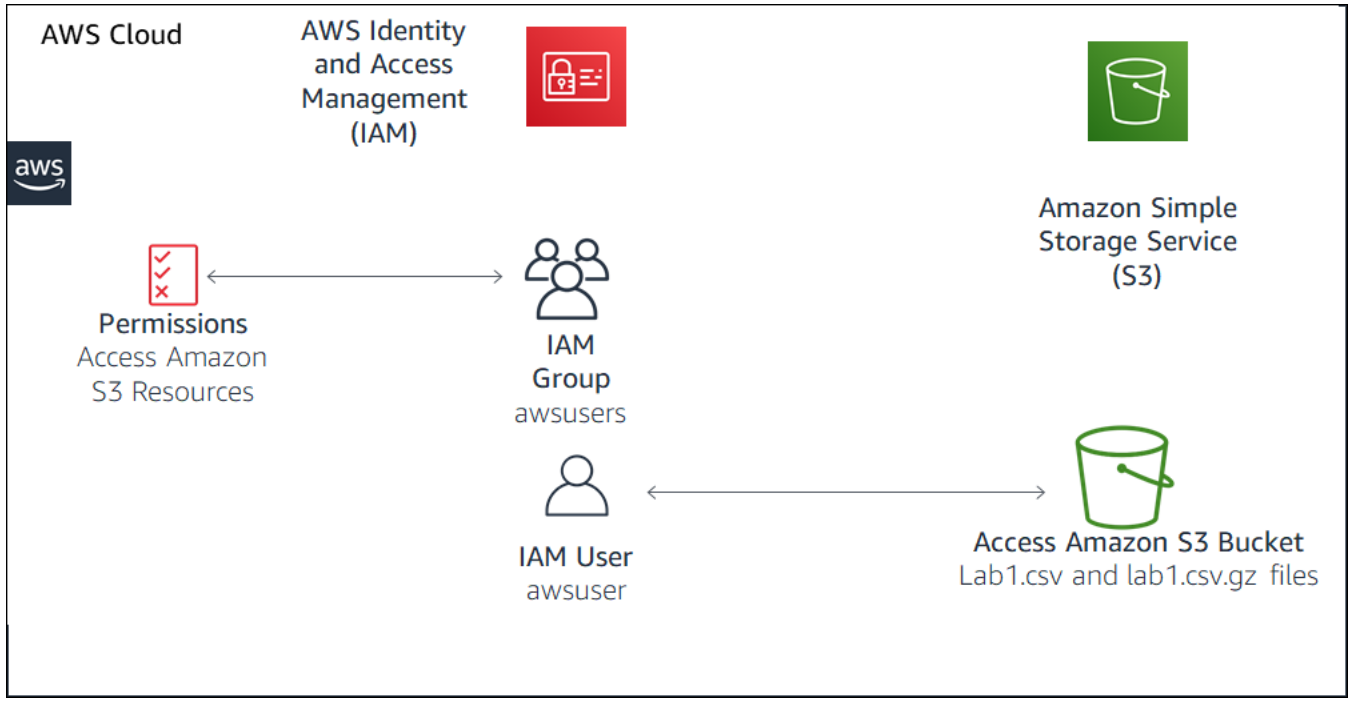
Goals

- To briefly explore AWS S3
- To upload data to S3 and perform basic SQL on the data

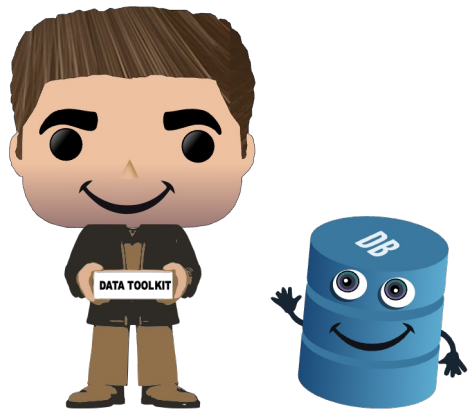
AWS S3

- Objectives
- Access Amazon S3 in the AWS Management Console
  - Secure an S3 bucket with IAM
  - Create a bucket with Amazon S3
  - Load data into an S3 bucket
  - Query an S3 bucket

Amazon Simple Storage Service (S3)

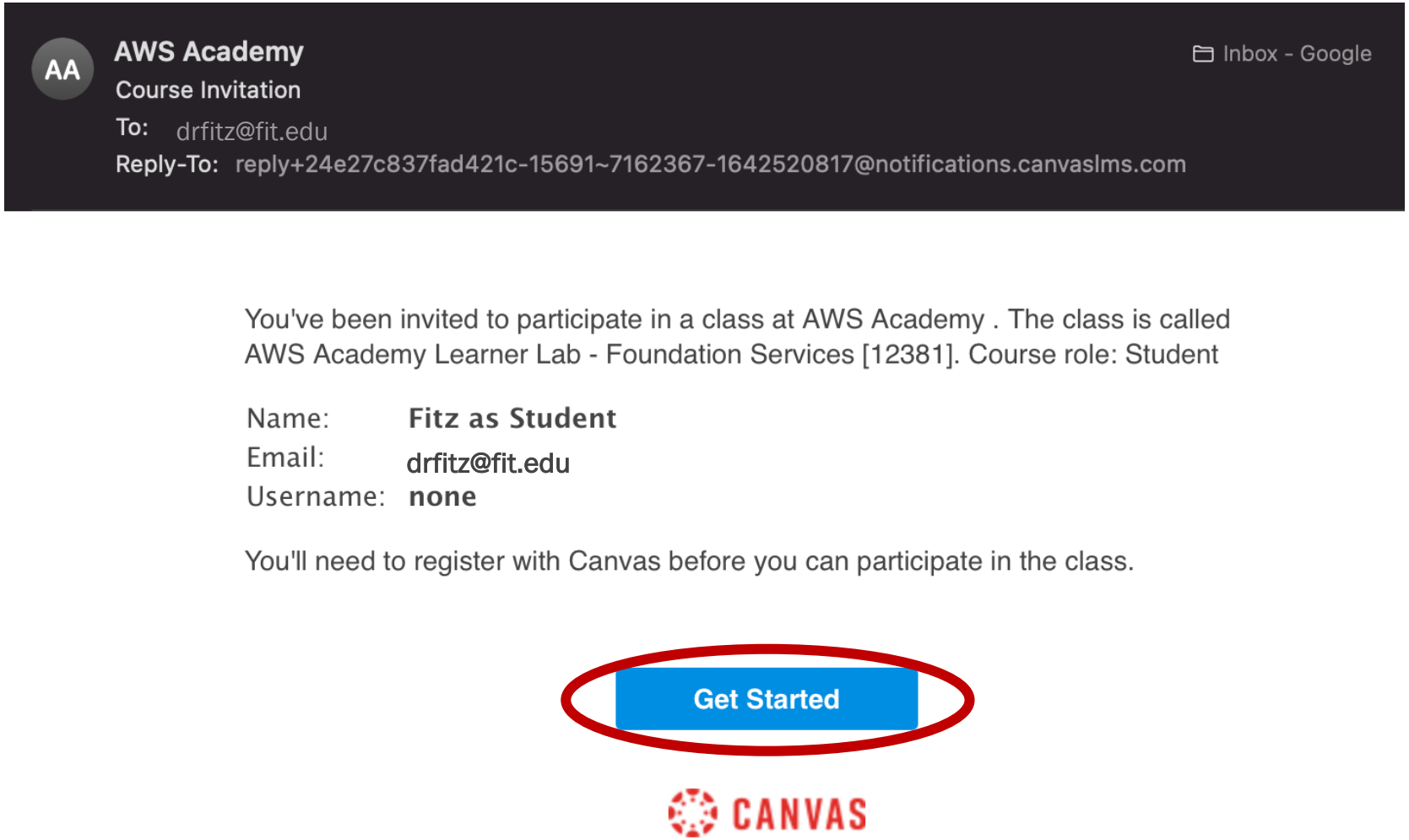


Hello,
I am Mr. Biggs! I specialize in Big Data. As you can see, I have my data toolkit, and I will be helping your set up and test AWS S3. Since Mr. DB specializes in databases, he may join us for DB-related topics.



Accept the invitation to join the AWS classroom

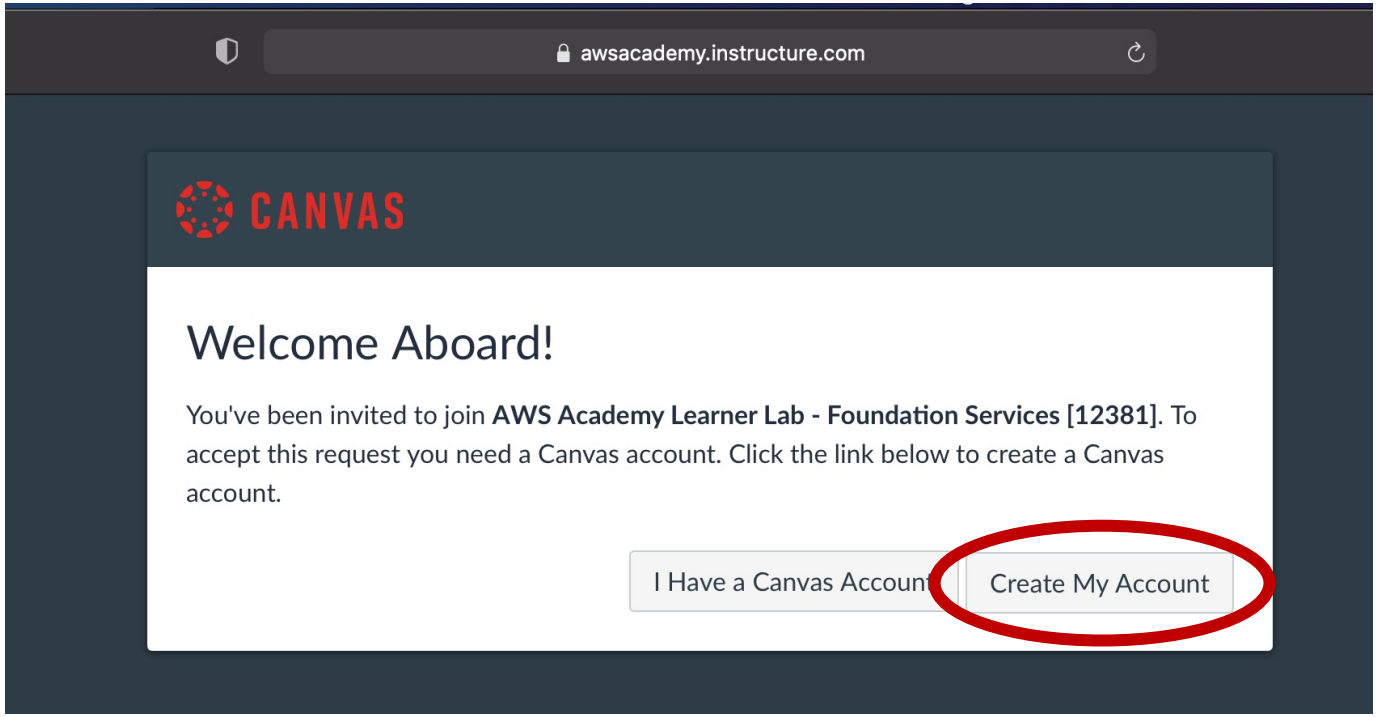
Join the class on AWS using the invitation email that should have come to your Florida Tech inbox



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Log in Using Canvas

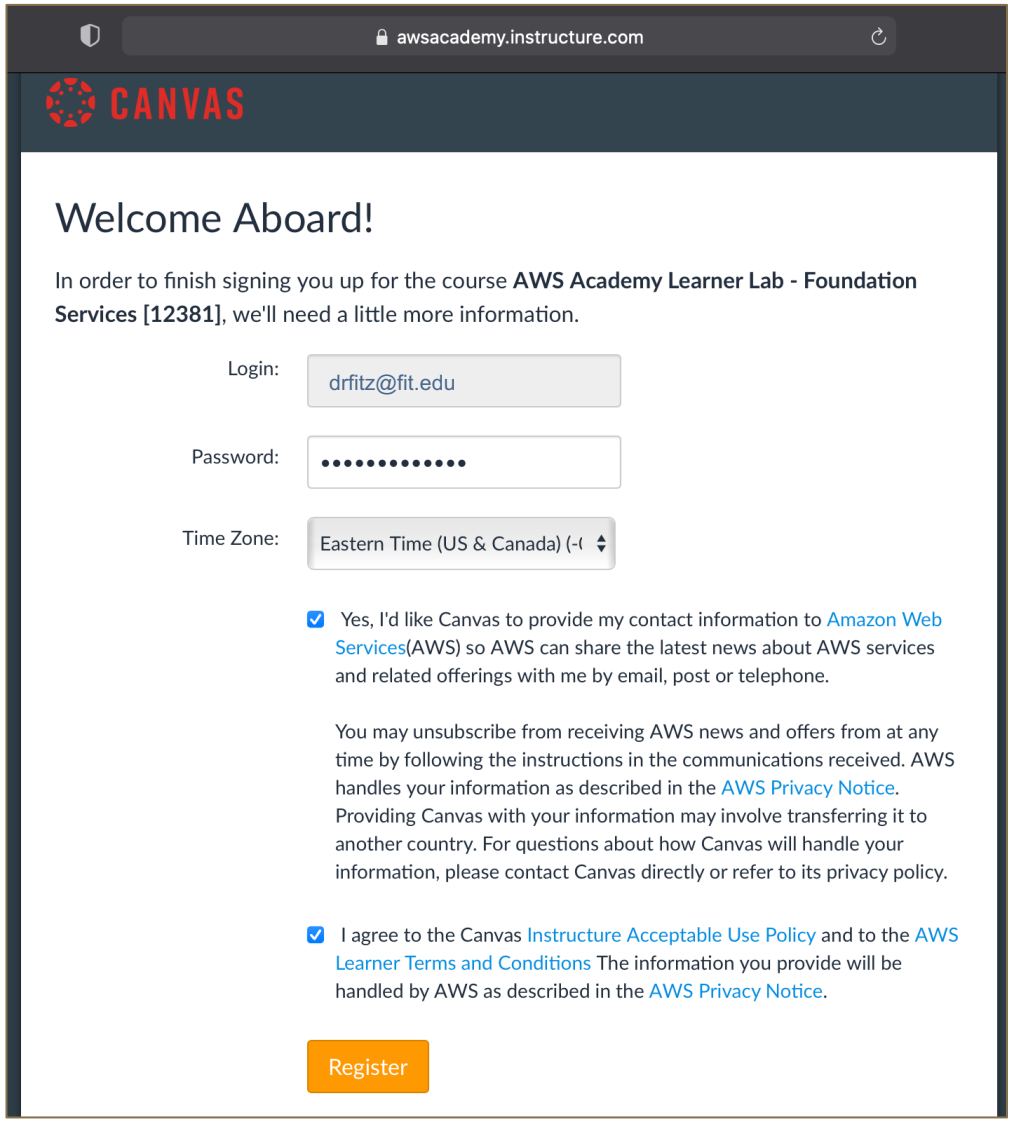
Your AWS Instructure account is different from your Florida Tech Canvas Account. If you have never used AWS Academy before, you will need to create an account by clicking the "Create My Account" button.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Creating a Canvas Account

Enter your credentials to register.

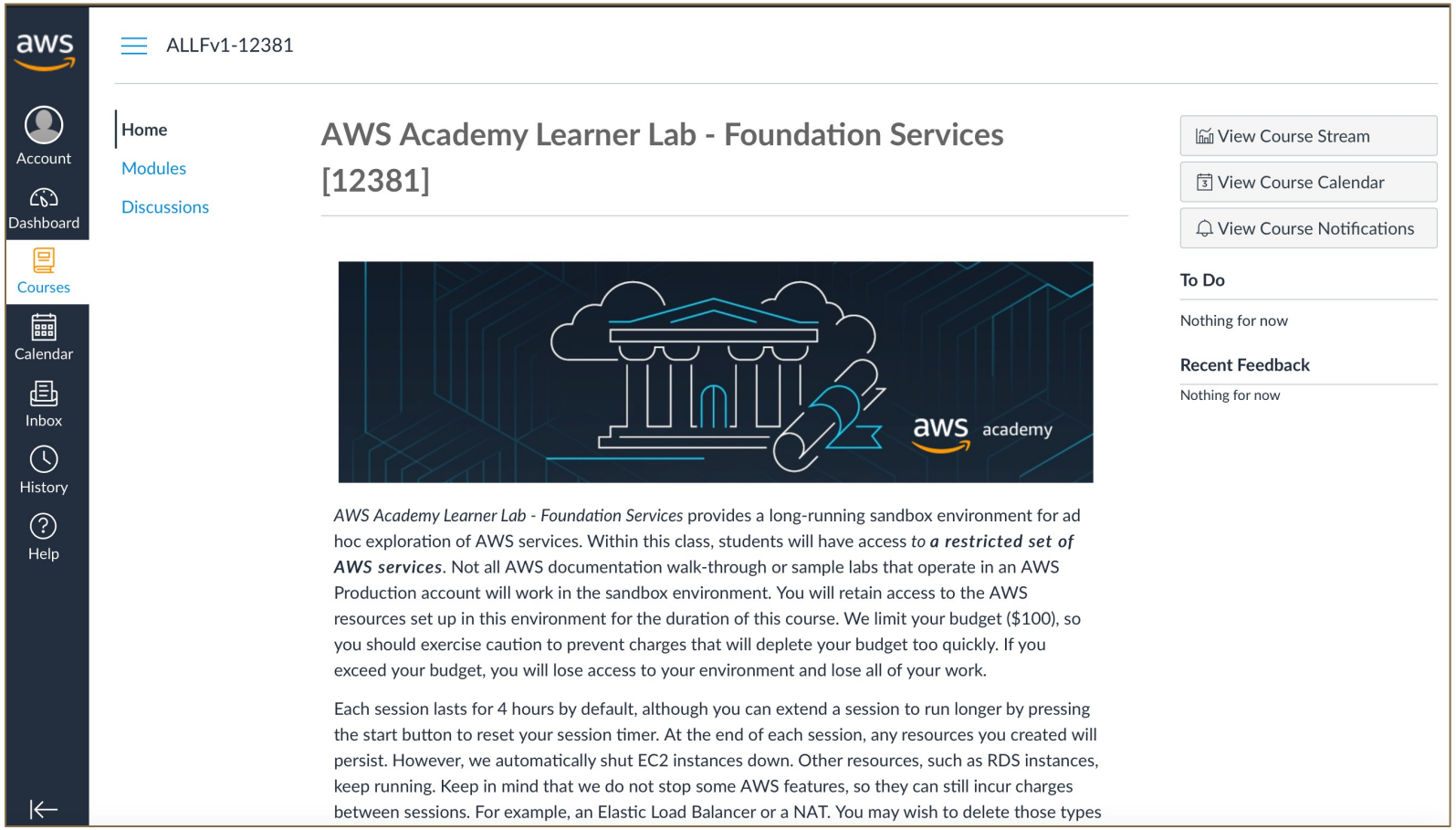


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Welcome Screen

After logging in to Canvas, you should see a screen that resembles the following.

This text displayed is dependent on the course selected. The Big Data Containers course will be a leraner lab whereas the general Big Data course will be named Data Analytics.

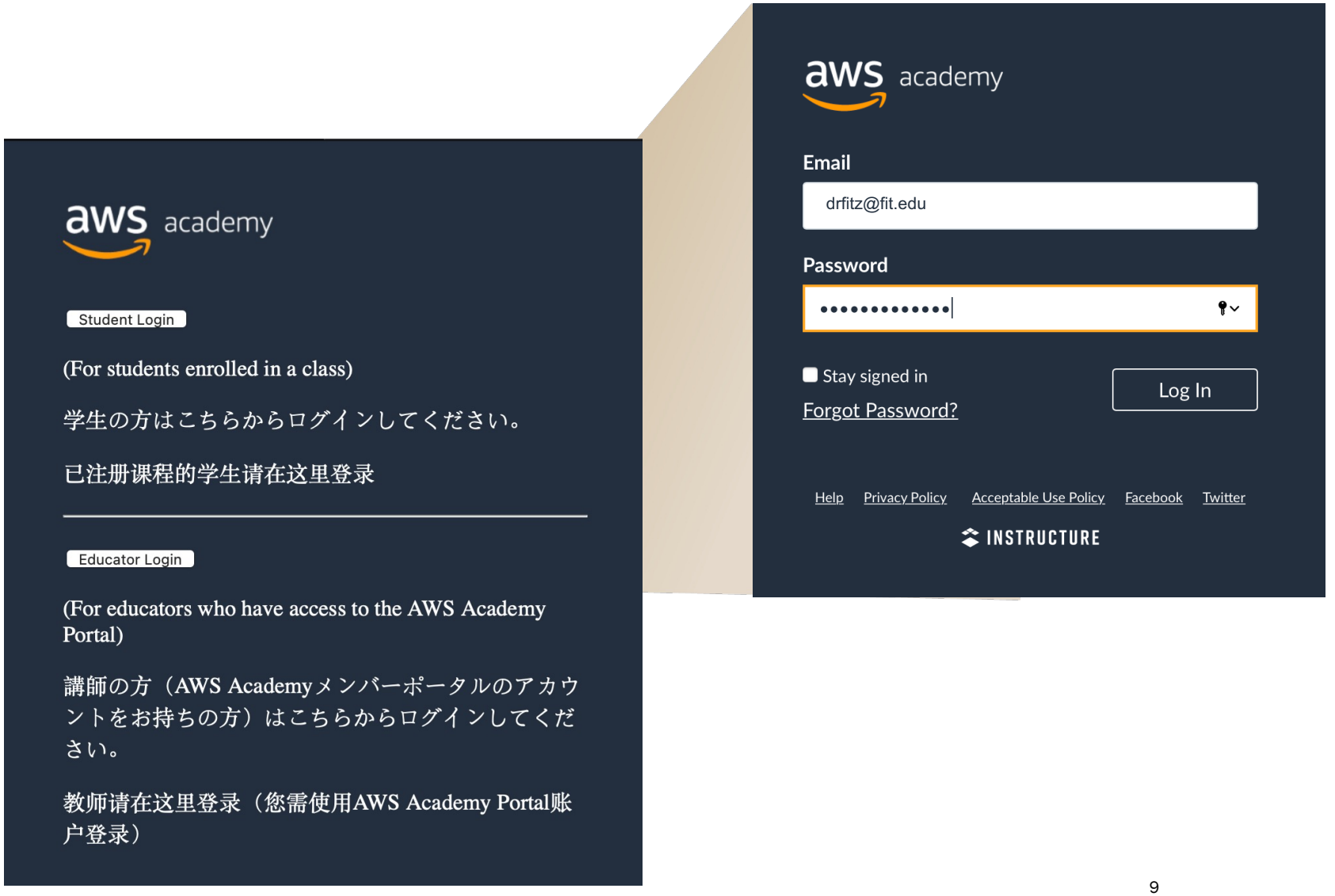


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Returning to AWS

To return to AWS, go to

https://www.awsacademy.com/LMS_Login



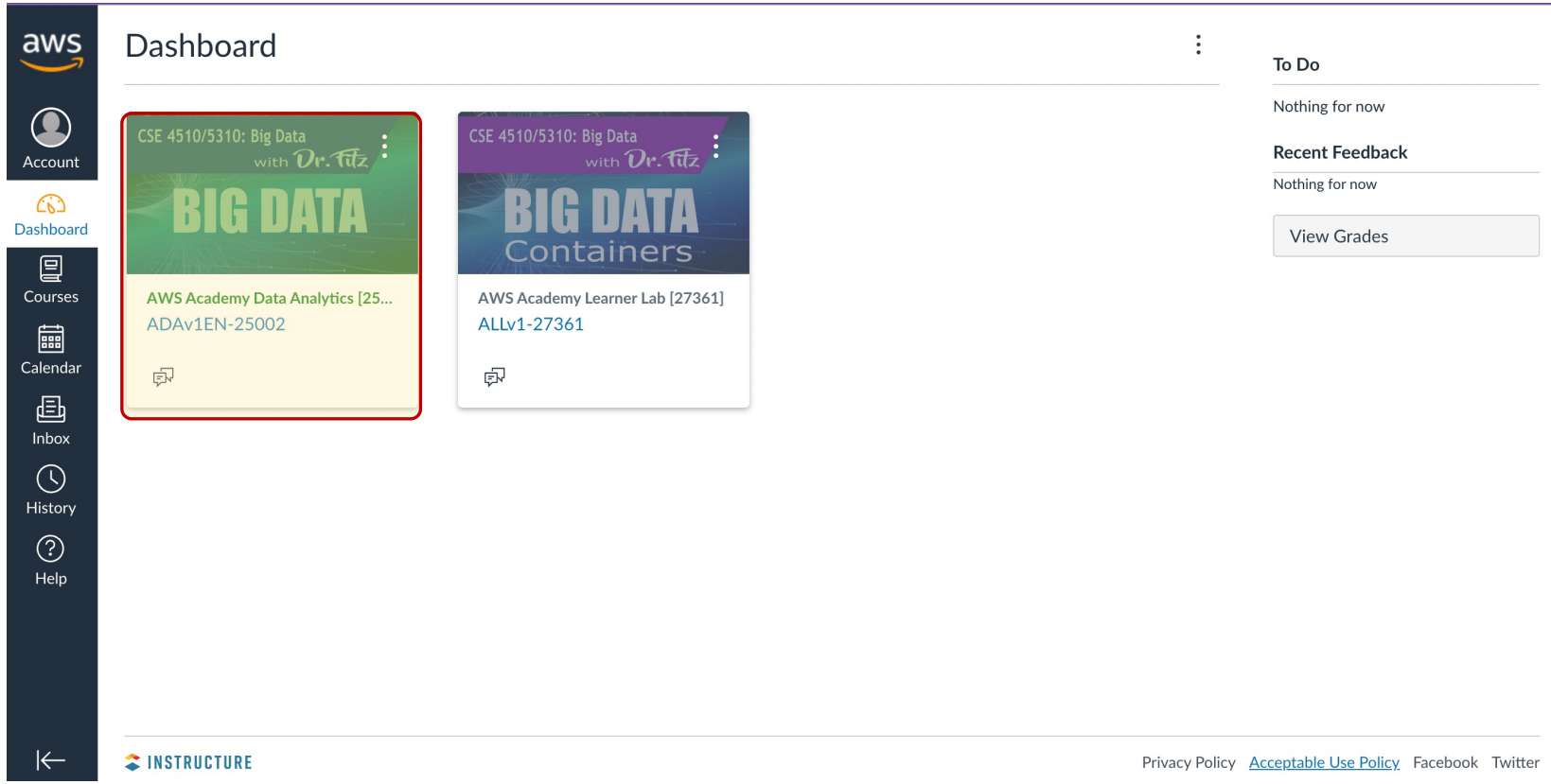
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Returning to AWS

The general Big Data Course has guided labs, which were created by AWS. Use this to explore various Big Data Platforms in a guided manner.

The Big Data Containers is for creating clusters and exploring various AWS Big Data platforms without guidance.

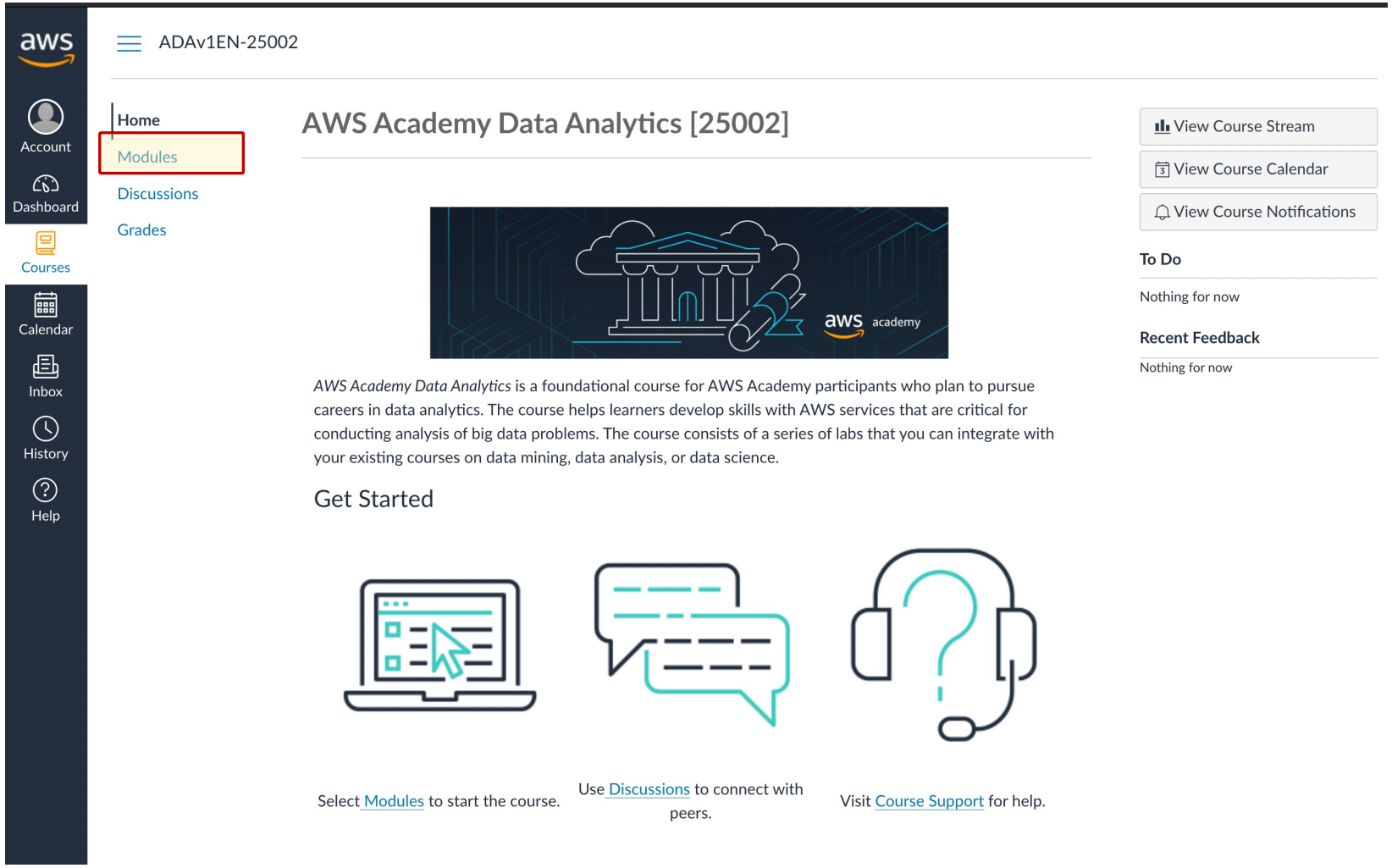
If you would like to work on your term project, I recommend you use the Big Data Containers.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Returning to AWS

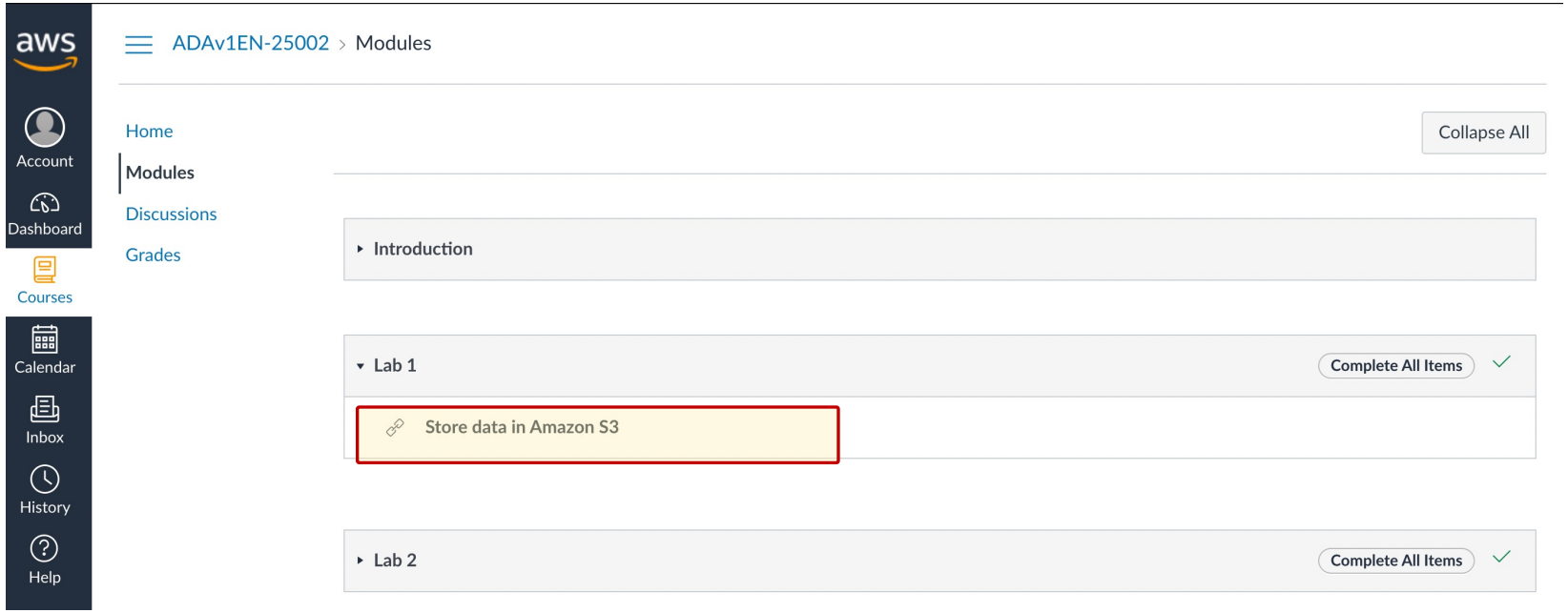
Click Modules to access course modules.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Returning to AWS

This screen shows the example of a guided lab that introduces students to Amazon S3.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

IAM Users

On the AWS Management Console, on the **Services** menu, choose **Services**.

From the list of services, choose **IAM**.

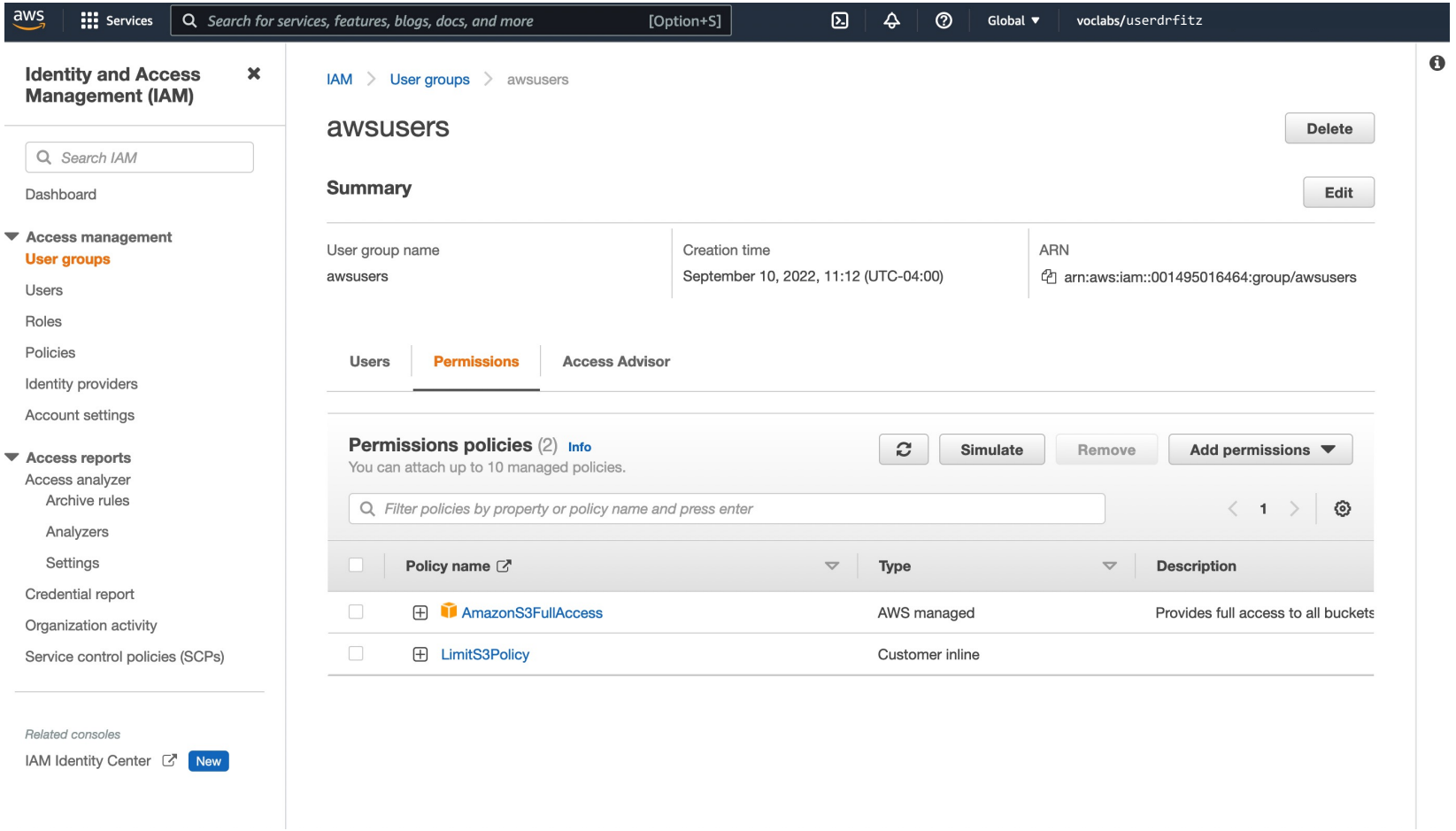
In the navigation pane, choose **Groups**.

Choose the awsusers group.

Choose the Permissions tab.

Notice that the *AmazonS3FullAccess* policy is attached to the group.

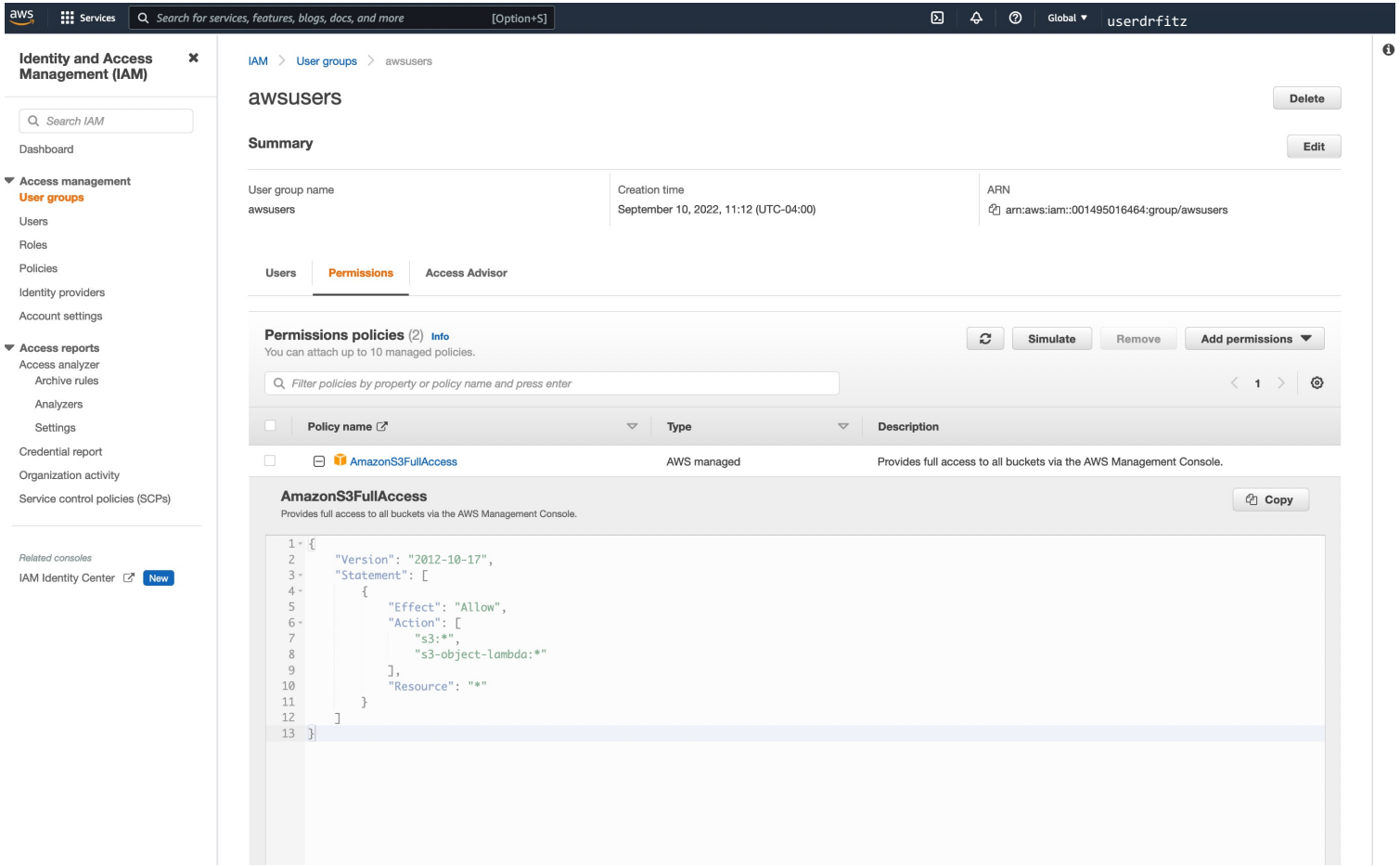
Choose Show Policy



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

IAM Users

The policy document is in JavaScript Object Notation (JSON) format. This policy states that users in that group are allowed to take all actions for Amazon S3 on all resources.



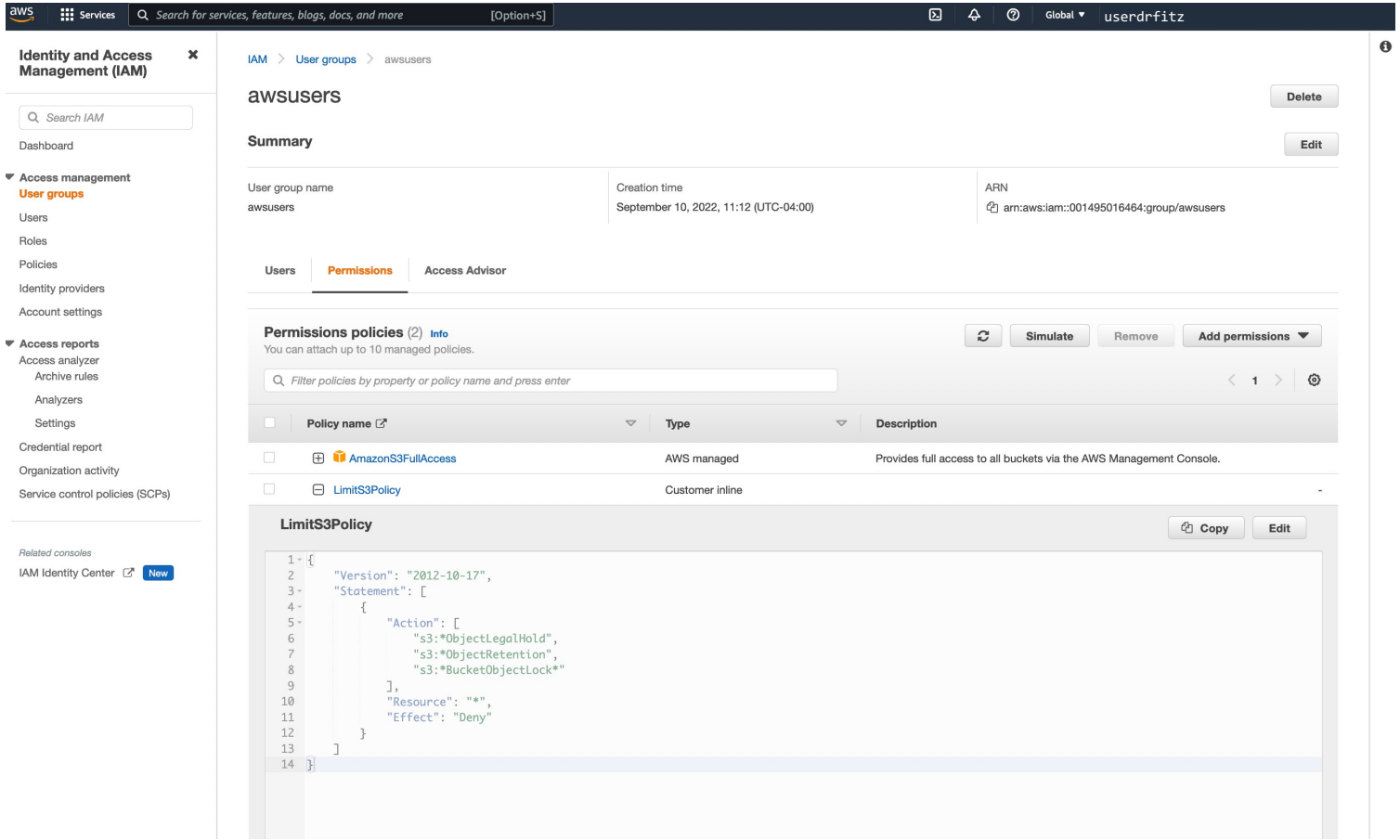
S3 Limit Policy

The policy document is in JSON format. This policy states that users in the group are **not** allowed to perform the following specified actions on S3 objects:

ObjectLegalHold – A legal hold prevents an object version from being overwritten or deleted.

ObjectRetention – A retention period determines how long an object is retained.

BucketObjectLock – When an object is locked, it cannot be deleted or overwritten.



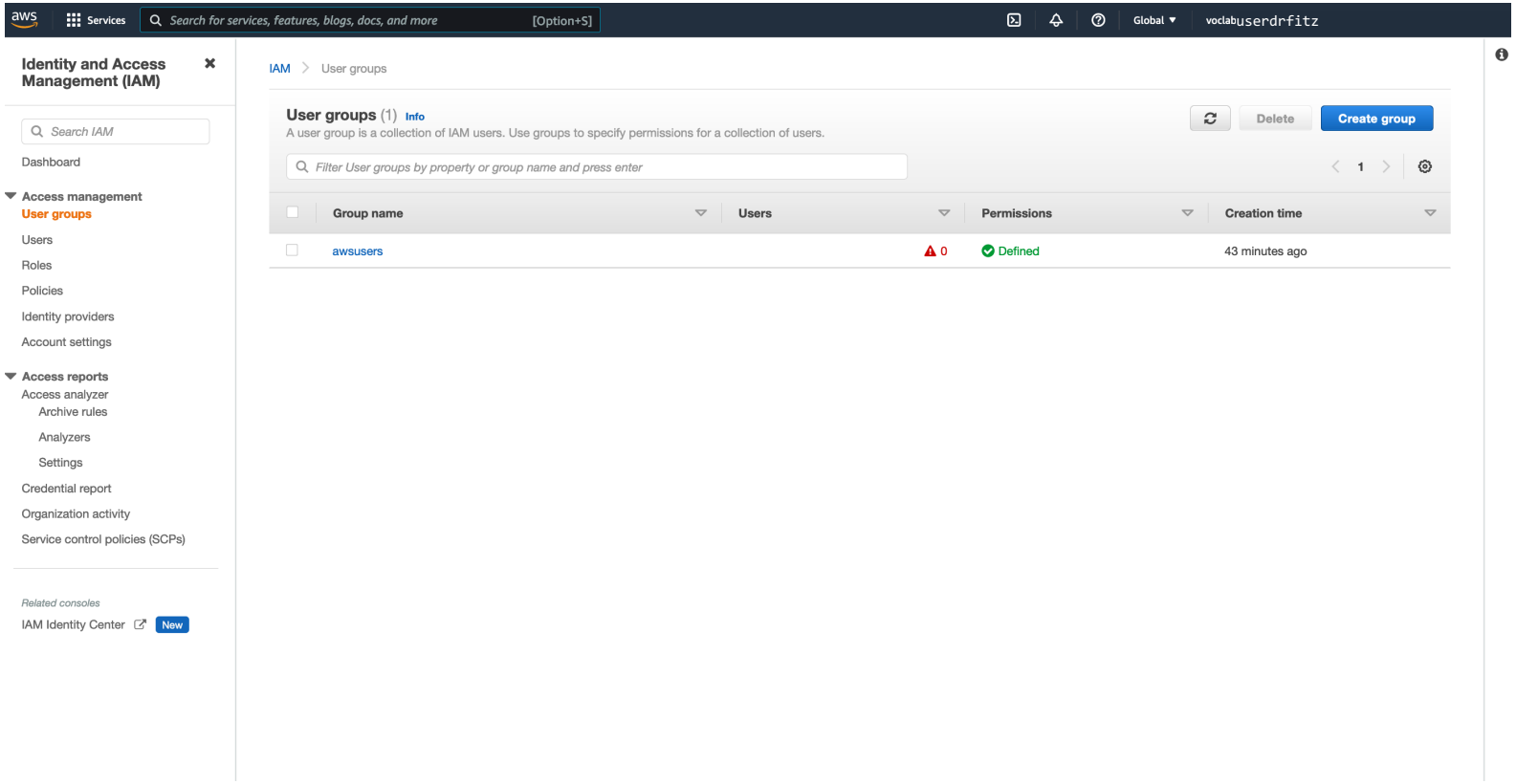
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add a user

In this task, you will add the *awsuser* to the *awsusers* group. You will also log out of the console and log back in to the console with the *awsuser* account and password.

In the navigation pane, choose **Groups**.

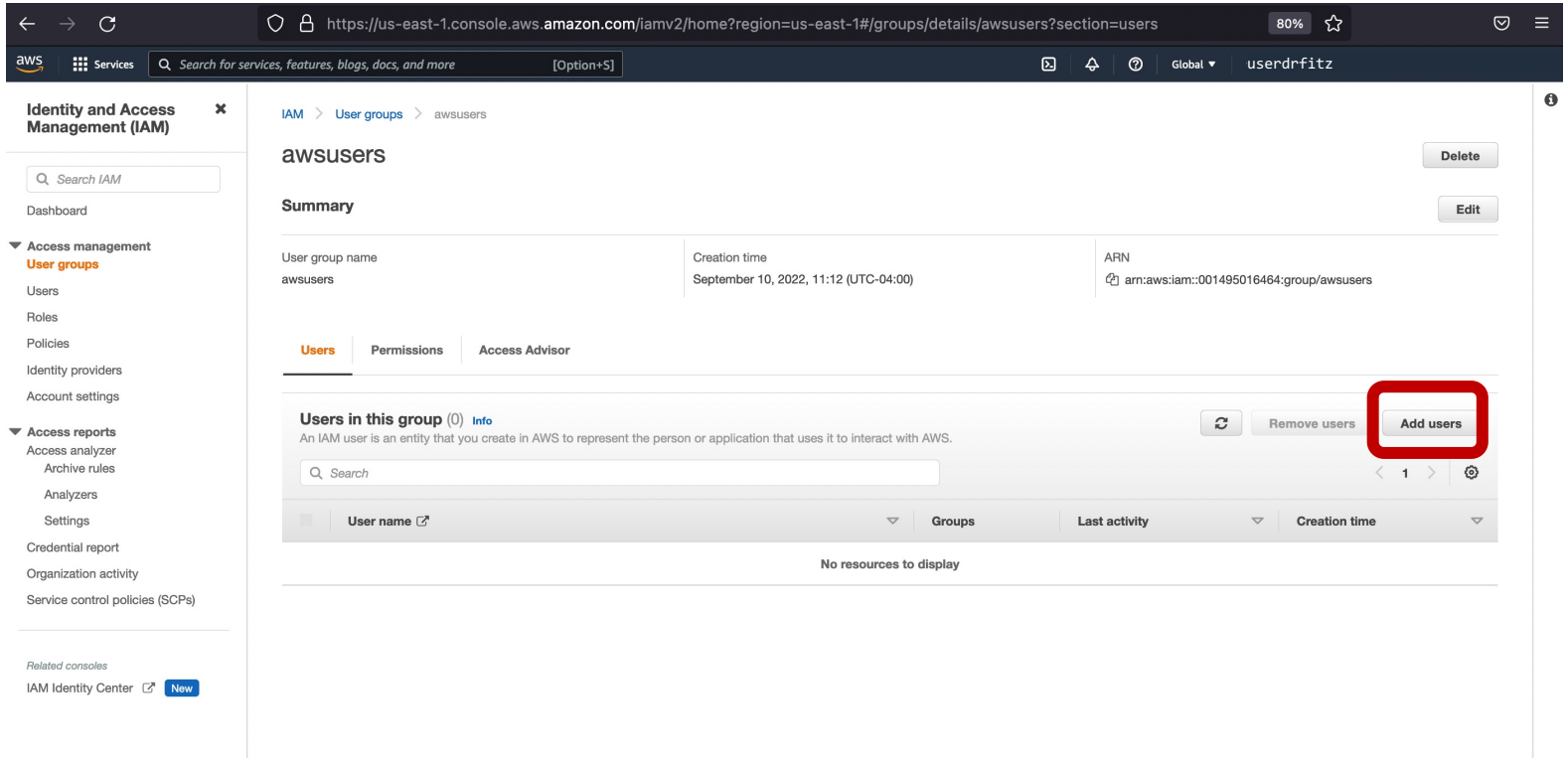
Select the **awsusers** group.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add a user

From the **Group Actions** menu, choose **Add Users** to **Group**.

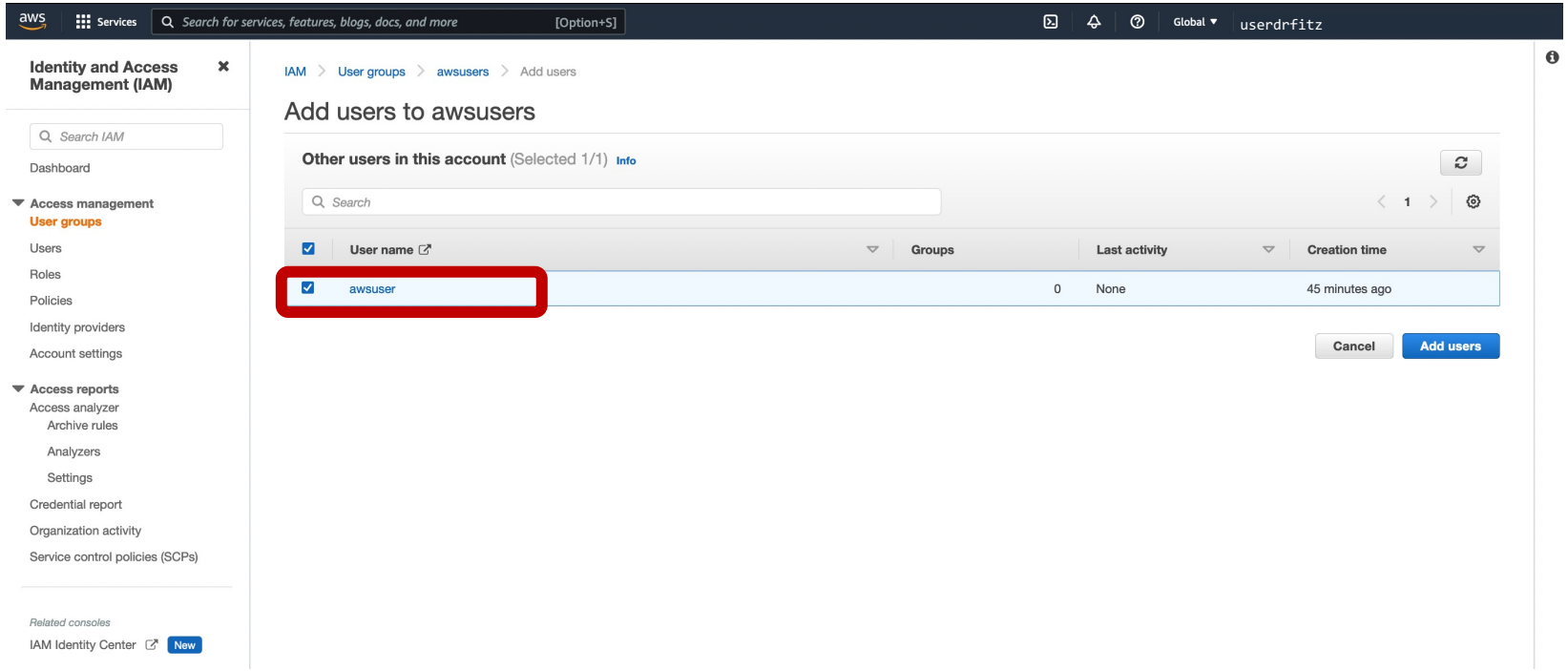


Add a user

Select the awsuser user.

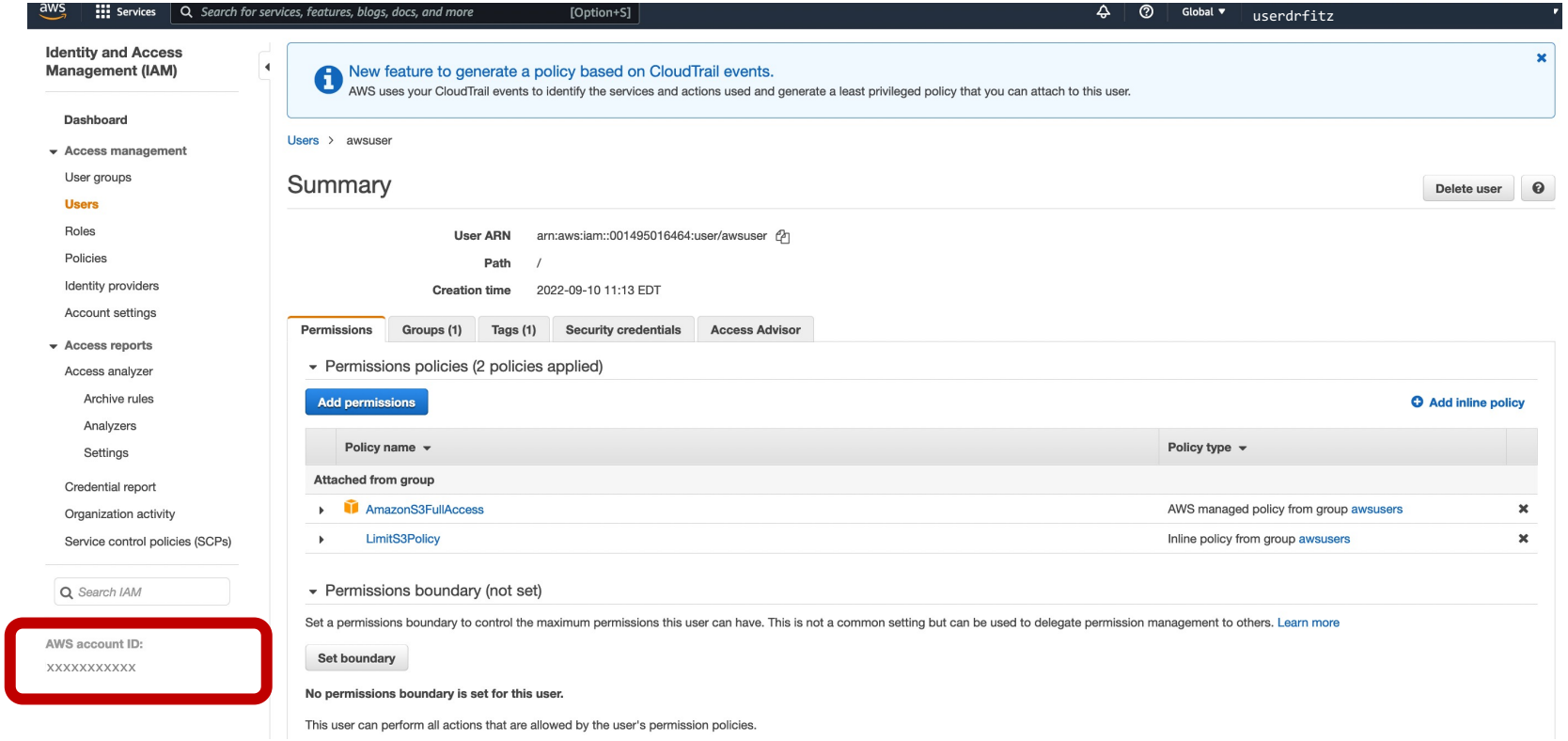
Choose Add Users.

From the navigation header, open the list of account actions and copy the account ID



Locate User ID

From the navigation header, open the list of account actions and copy the account ID.

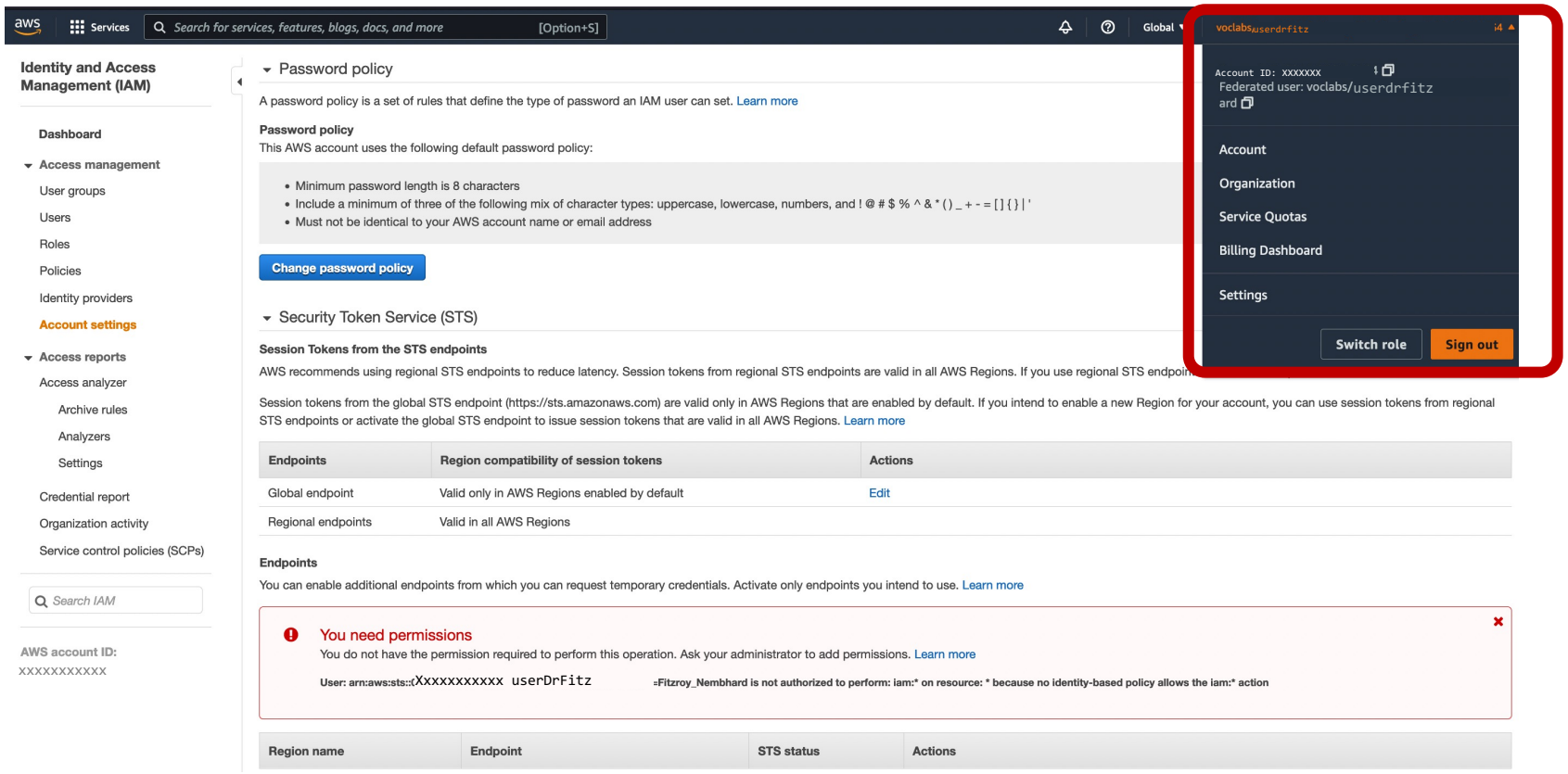


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Sign Out

In the list of account actions, choose **Sign Out**.

To sign back in with the *awsuser* credentials, choose **Sign In to the Console**.

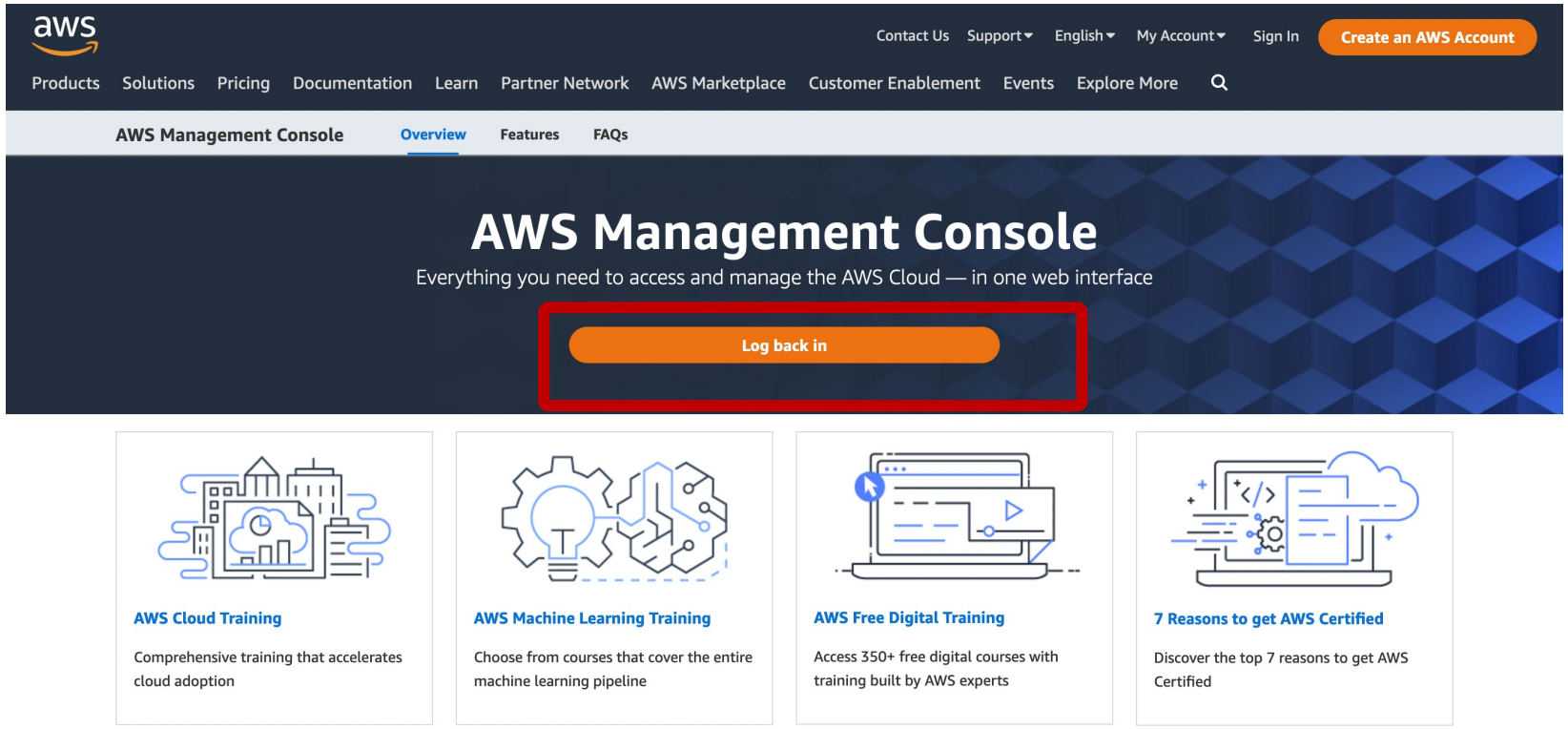


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Log in as IAM User

In the list of account actions, choose Sign Out.

To sign back in with the *awsuser* credentials, choose **Sign In to the Console**.



Log in as IAM User

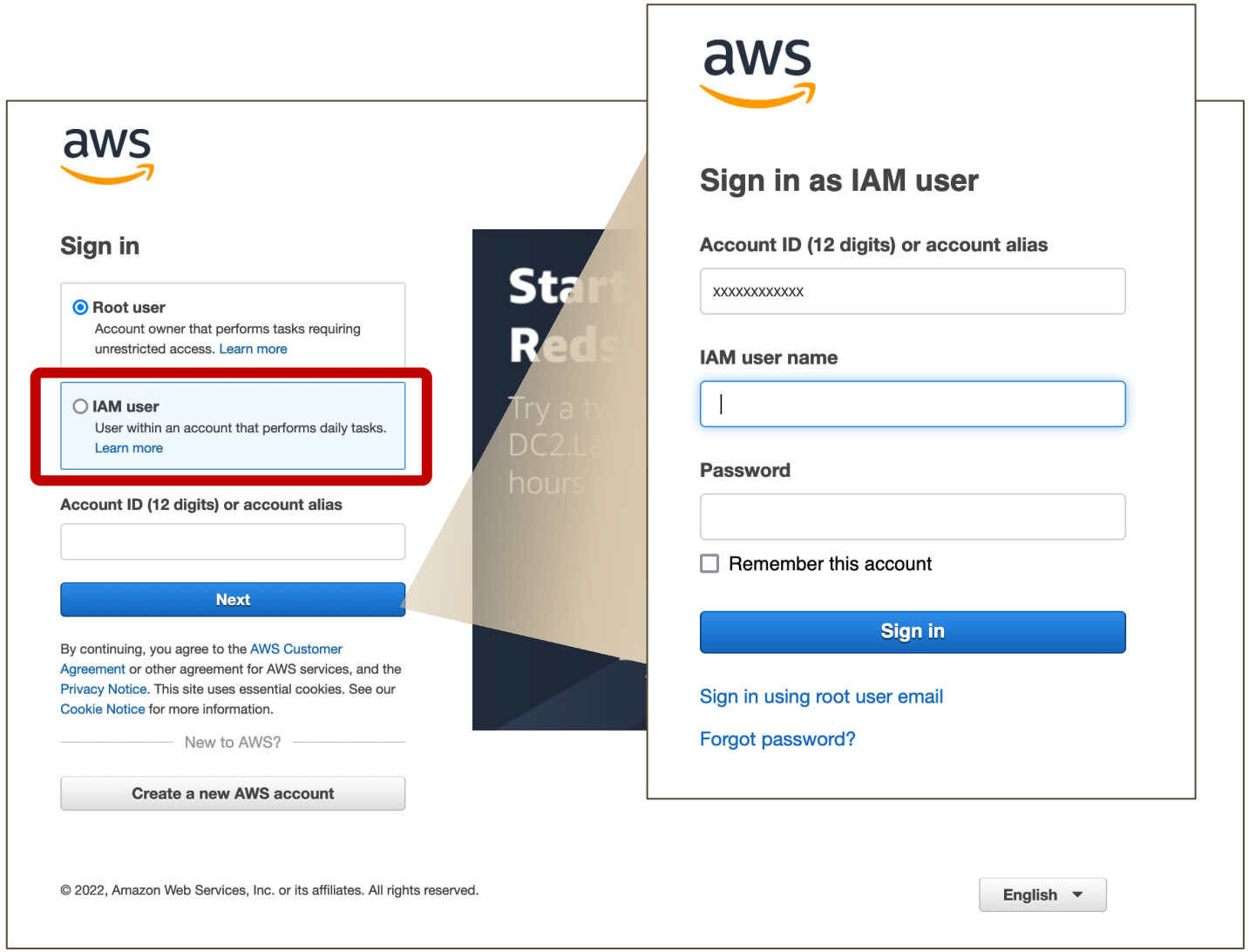
Select **IAM user** and then use the following information to sign in:

Note: Remove the dashes from the account number before you enter it.

Account: The account ID that you previously copied

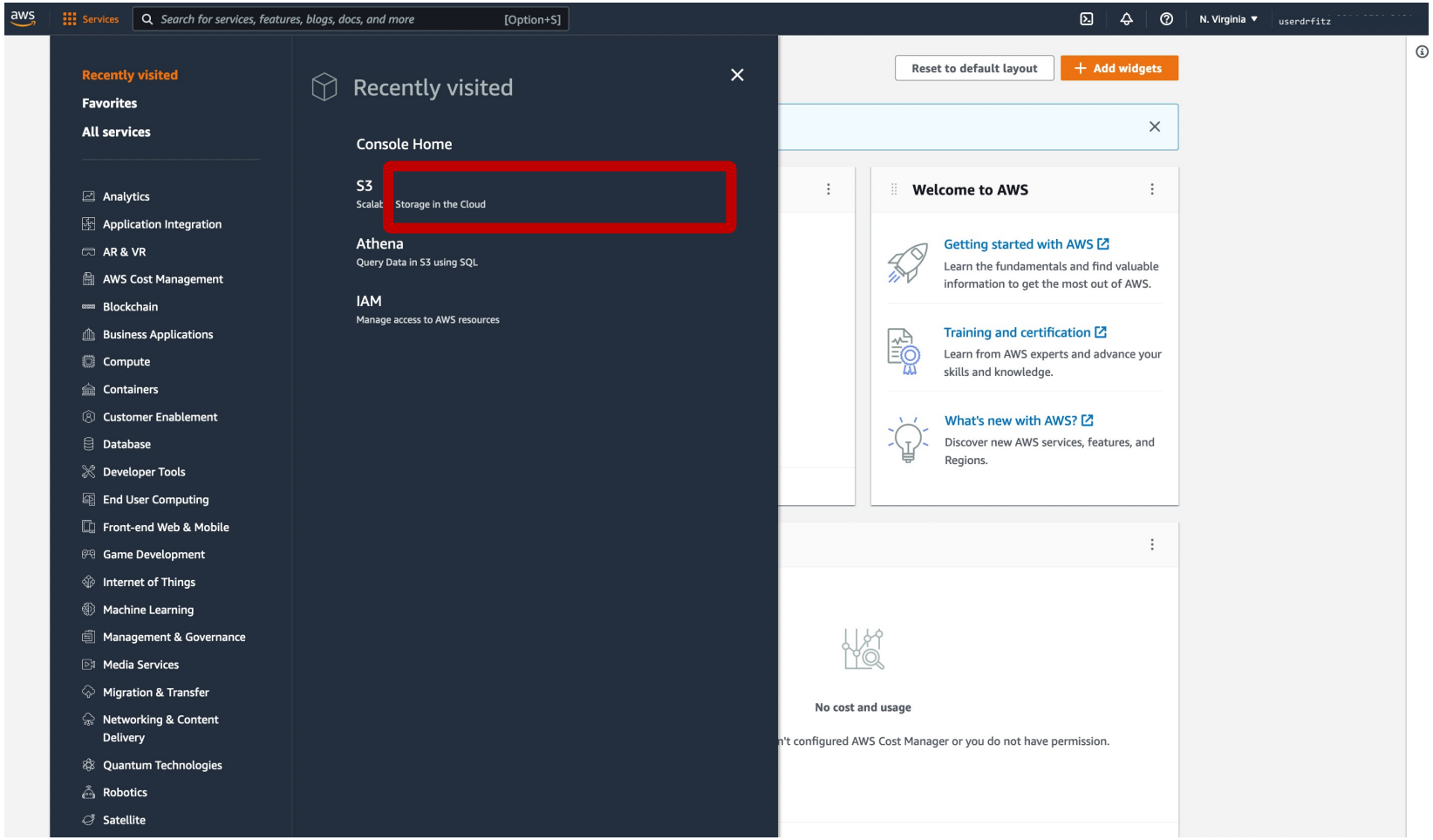
IAM user name: awsuser

Password: myP@ssW0rd



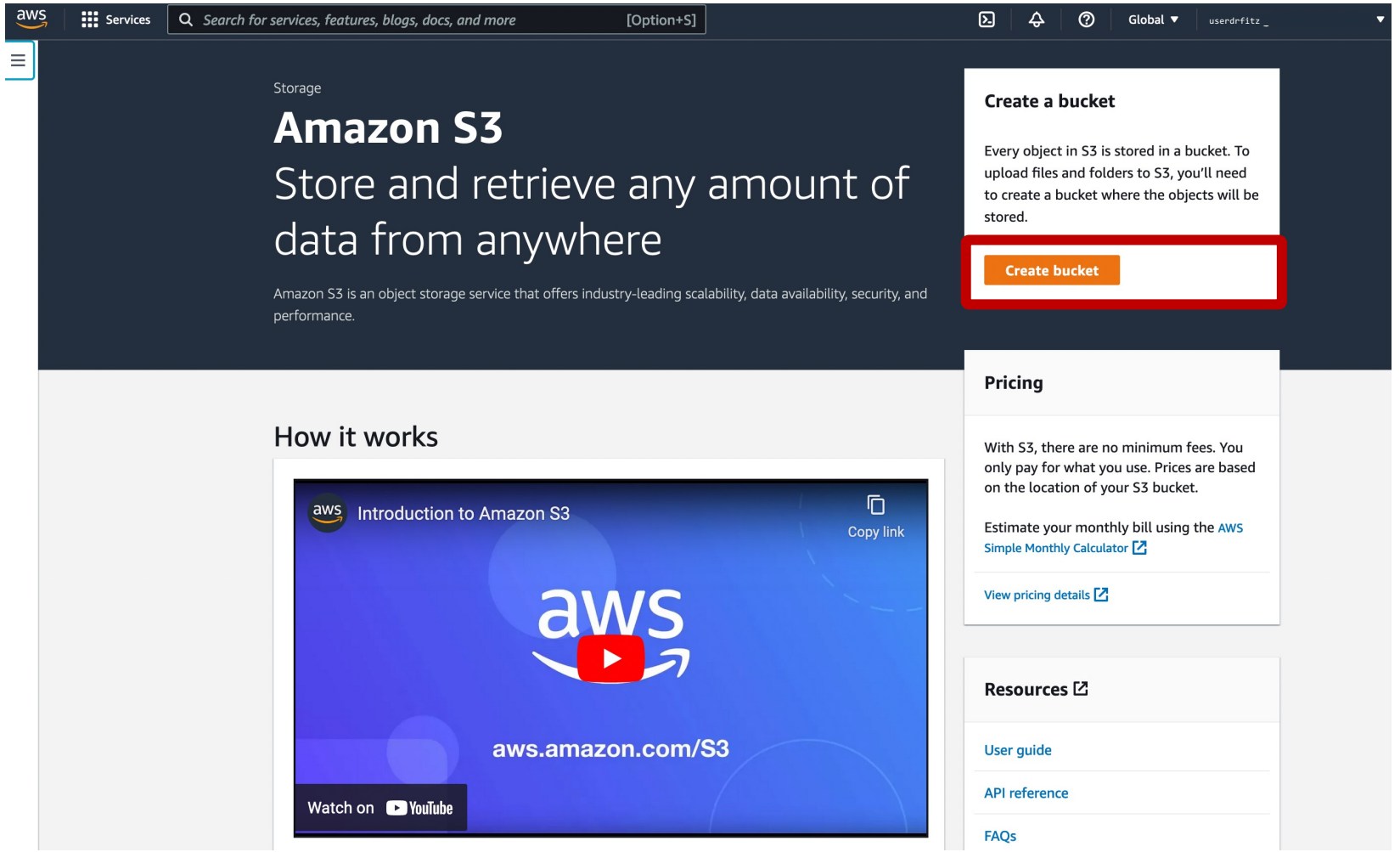
Select S3

Buckets and objects are the basic building blocks for Amazon S3. You create buckets and add objects to the buckets. Objects in Amazon S3 can be up to 5 TB.



Create a Bucket

Choose Create bucket.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

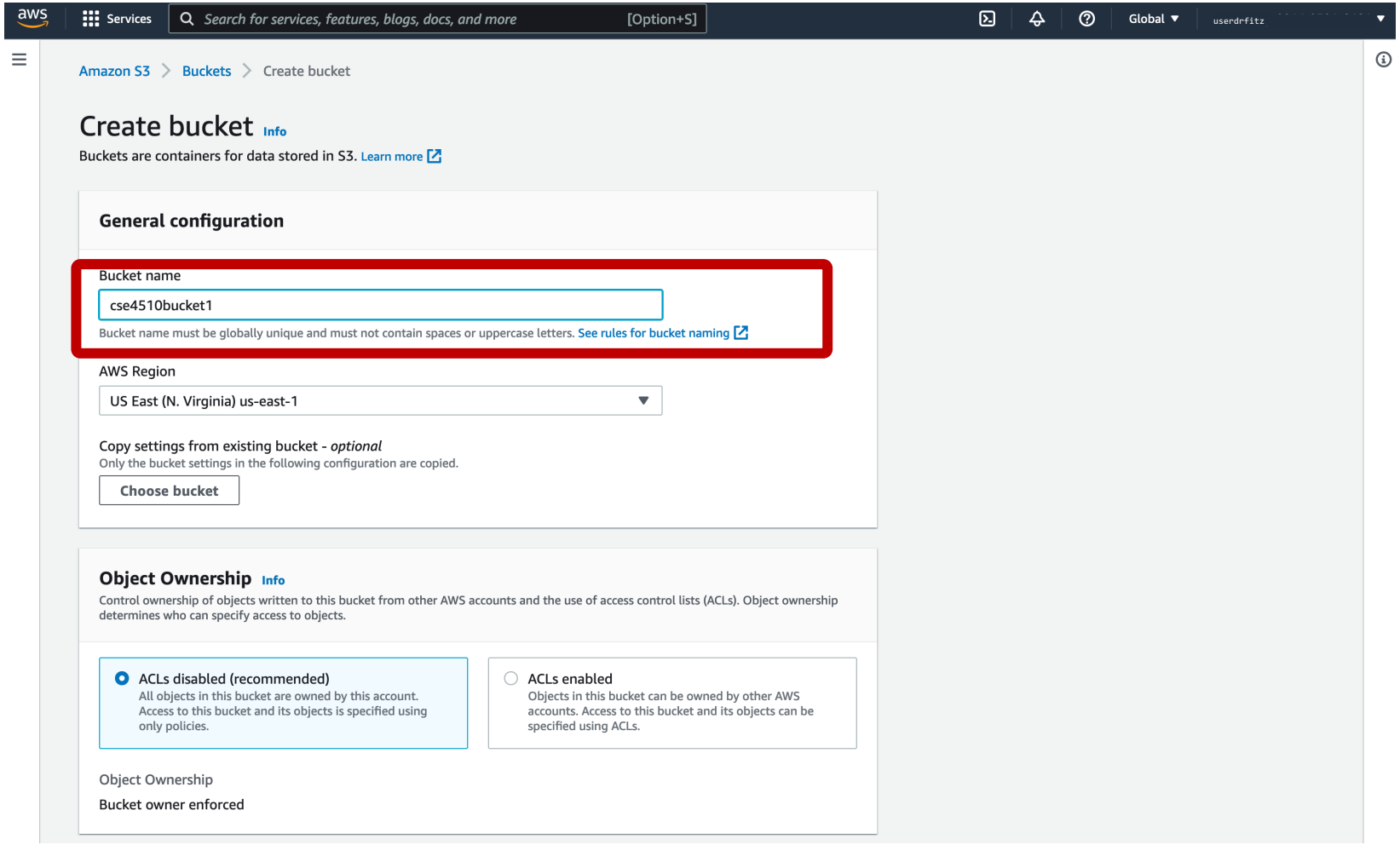
Create a Bucket

Enter a bucket name with three or more characters. Uppercase characters are not allowed.

Choose Create bucket.

Note: S3 bucket names must be unique across all buckets in Amazon S3. If you get a conflict with another bucket, add a digit and try again.

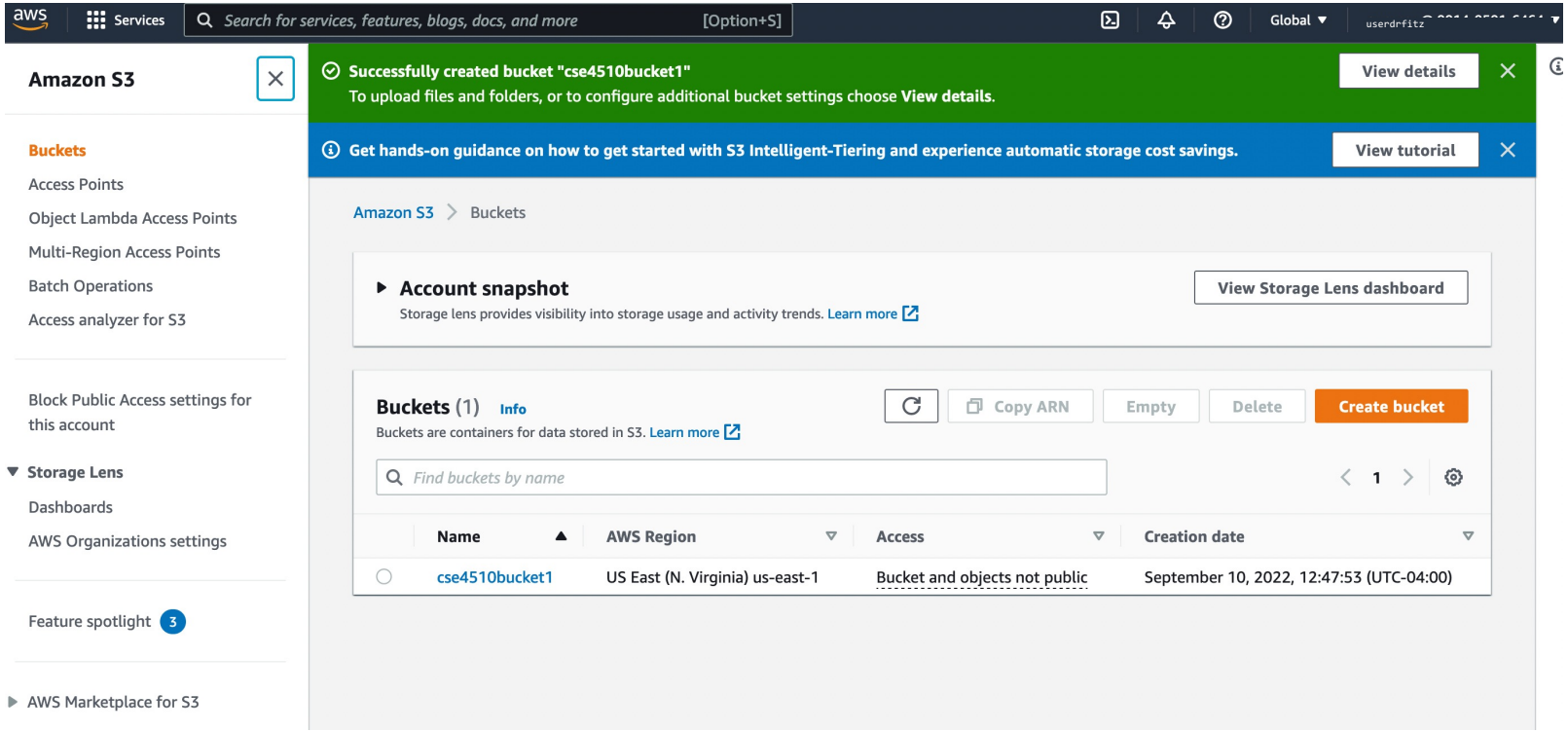
Note: Write down the bucket name because it will be used in future steps



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Create a Bucket

You should see a message that the Bucket was successfully created.

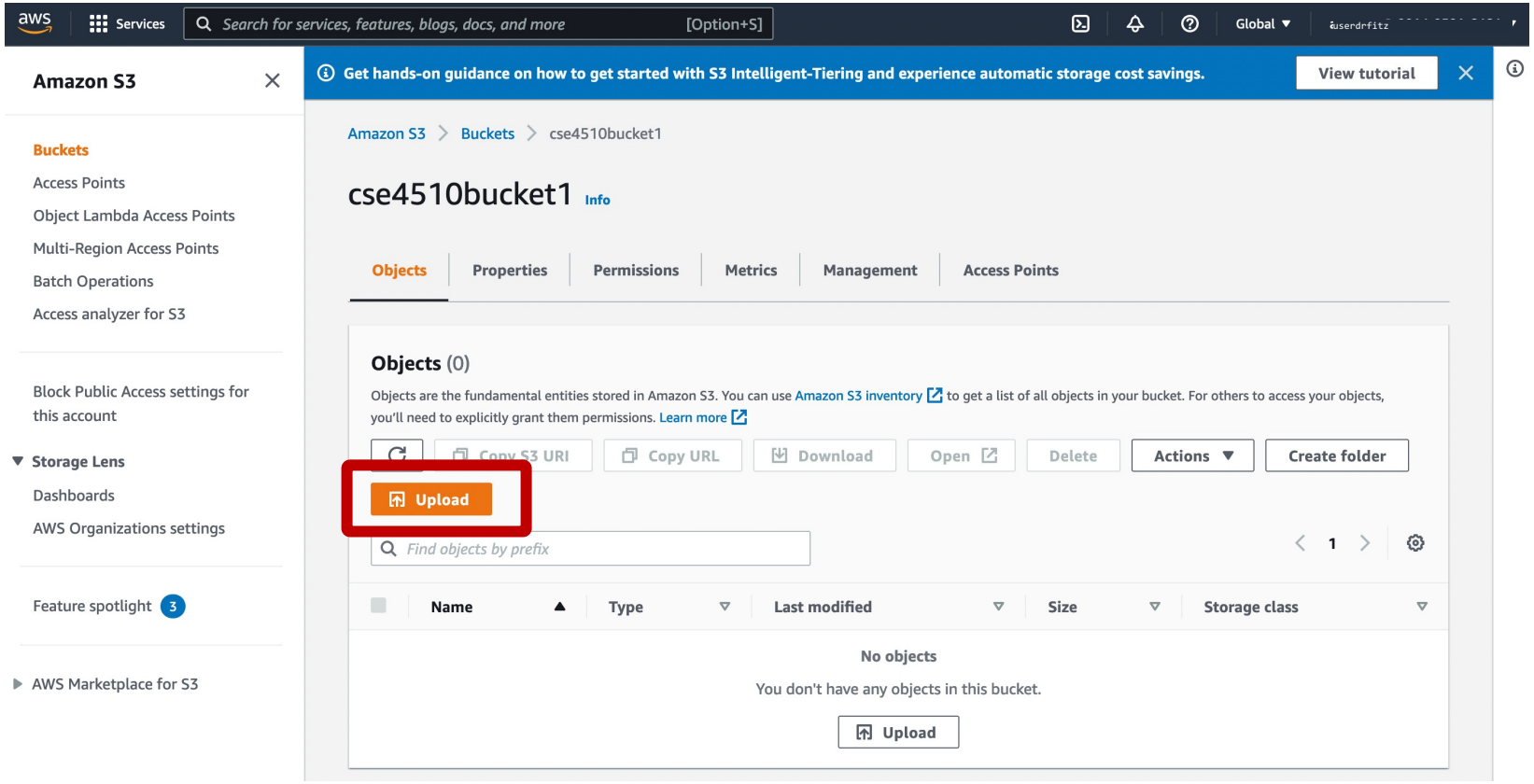


Upload Data

In this task, you will upload an object to the S3 bucket that you created. First, download the lab1.csv file from Canvas to a local directory.

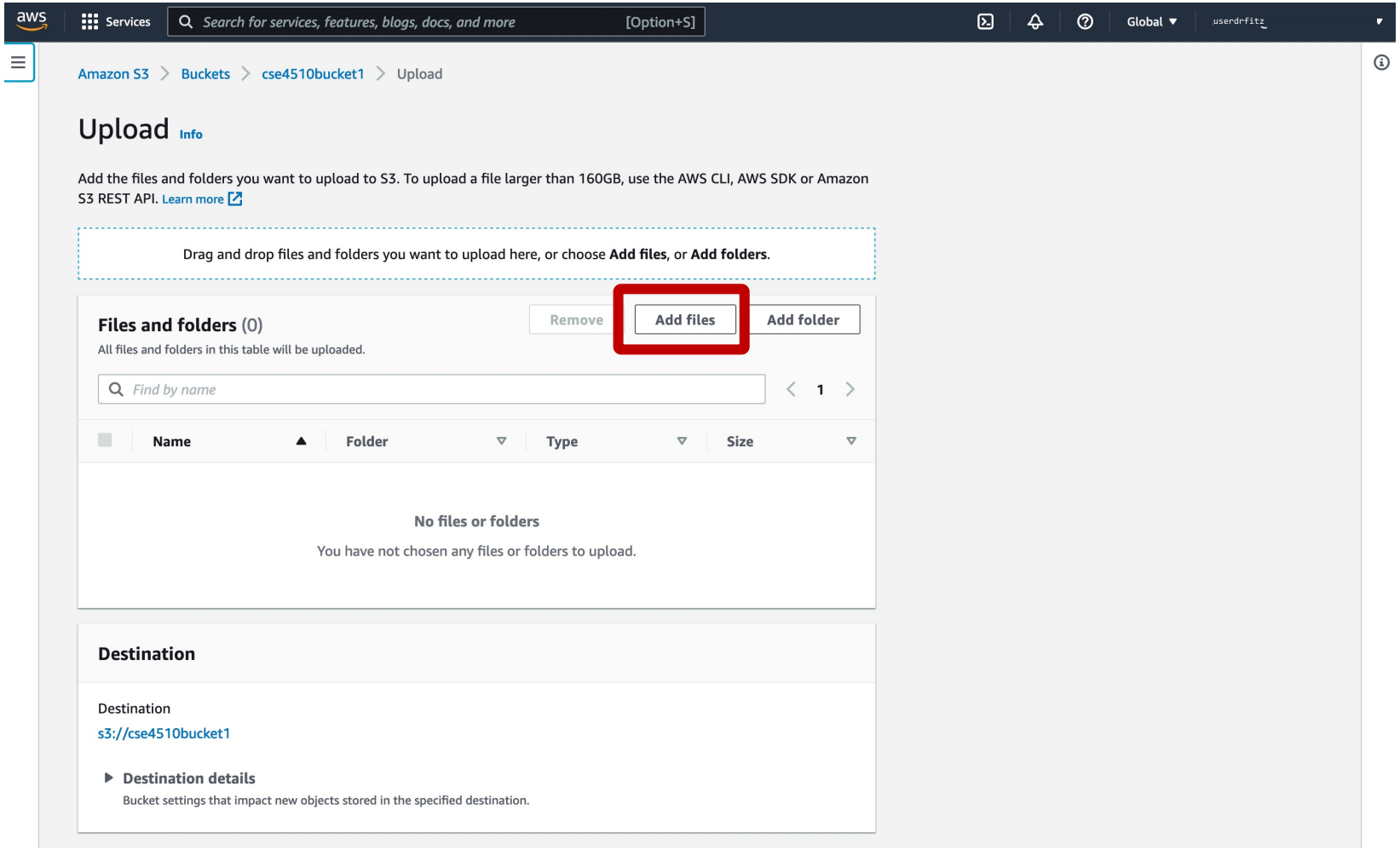
Choose the bucket that you created in the previous task.

In the Amazon S3 console, choose **Upload**.



Upload Data

Choose **Add files**.



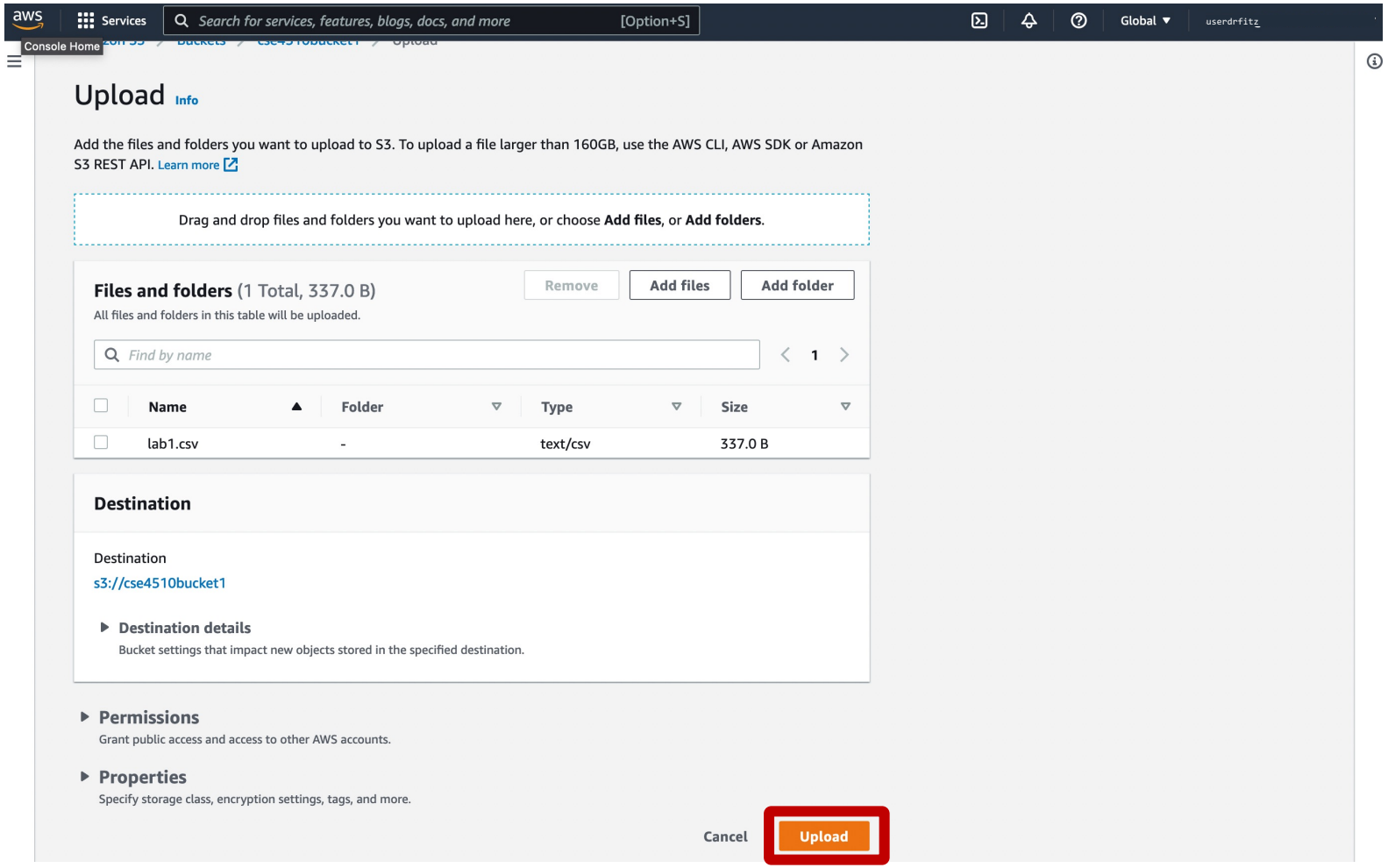
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Upload Data

Browse to the directory where you stored the lab1.csv file.

Choose the **lab1.csv** file.

Choose **Upload**.

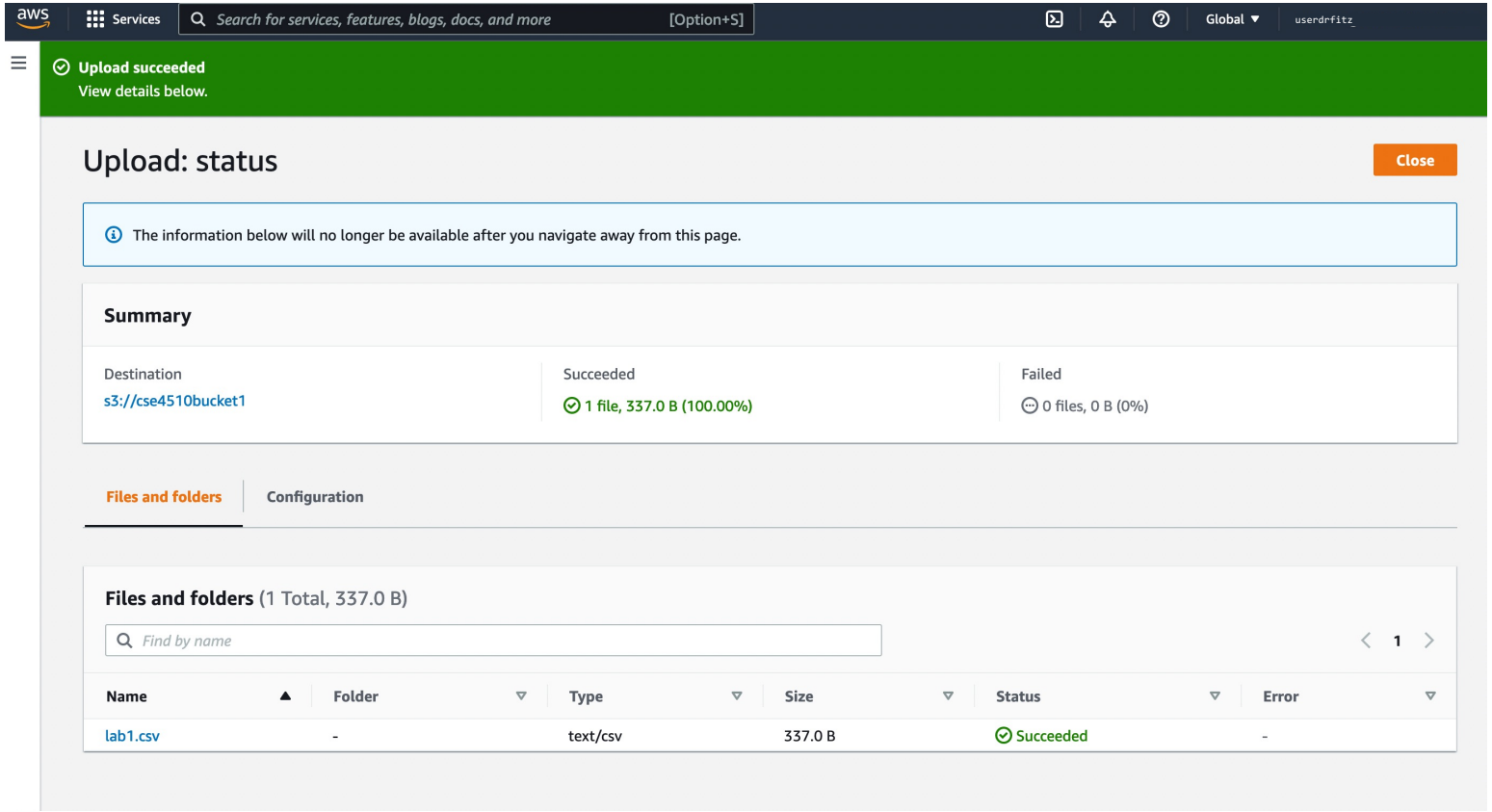


Upload Data

Browse to the directory where you stored the lab1.csv file.

Choose the **lab1.csv** file.

Choose **Upload**.



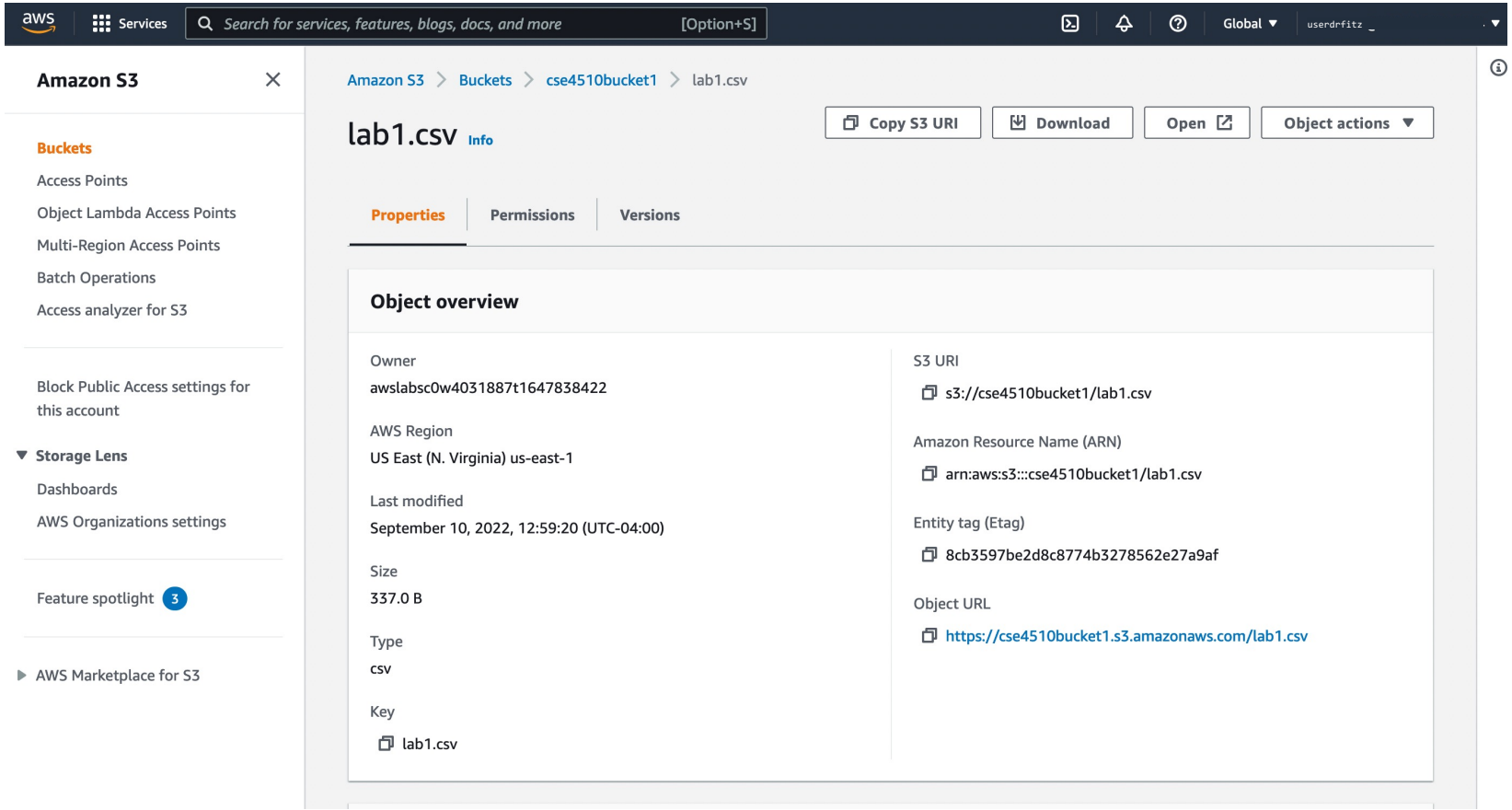
Review Uploaded File

In the Amazon S3 console, choose the **lab1.csv** file.

Review the file properties for the file that you uploaded.

Note: You should get a message stating that versioning is not enabled for the bucket. This behavior is expected.

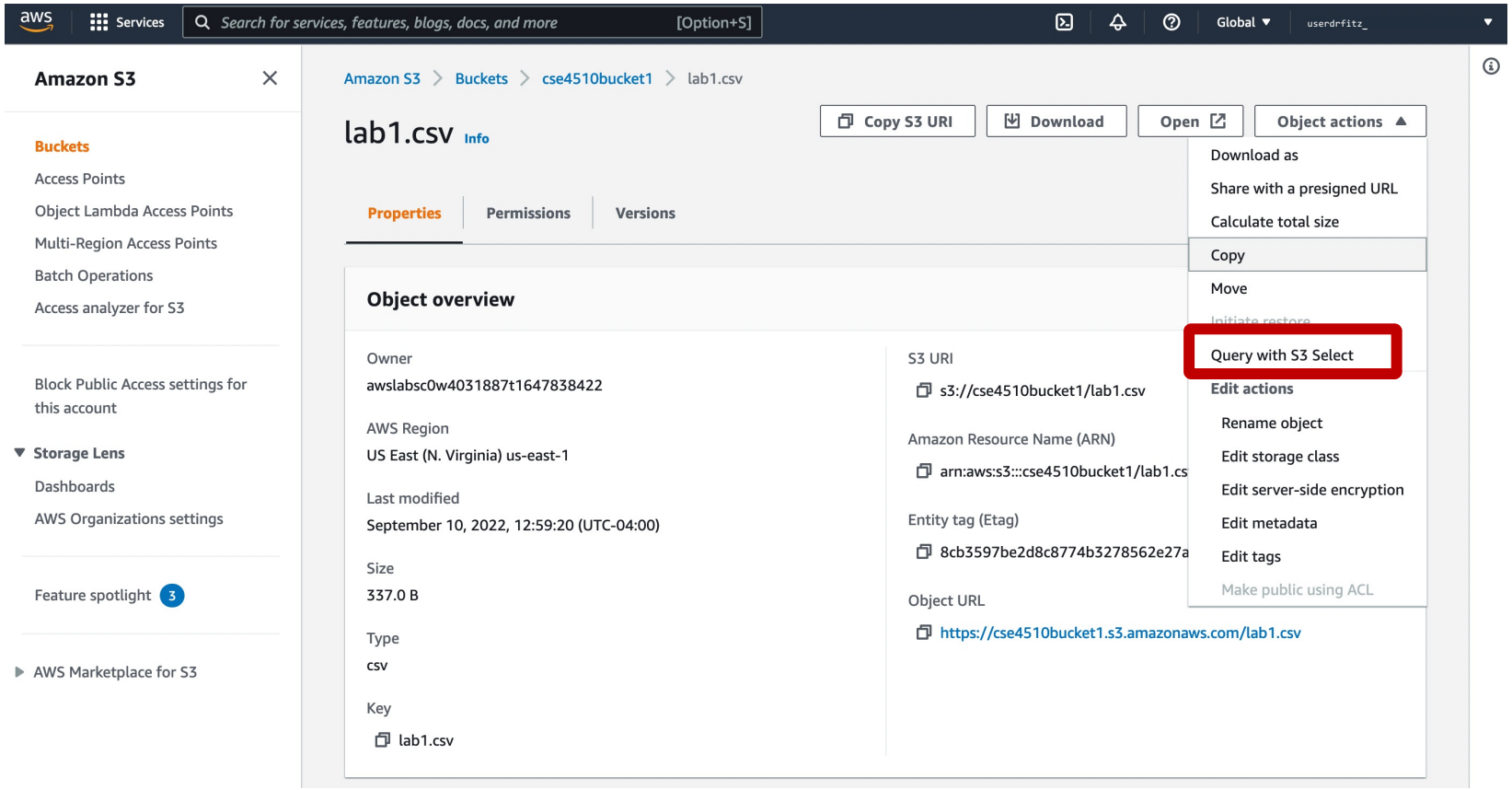
.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Query the object you uploaded

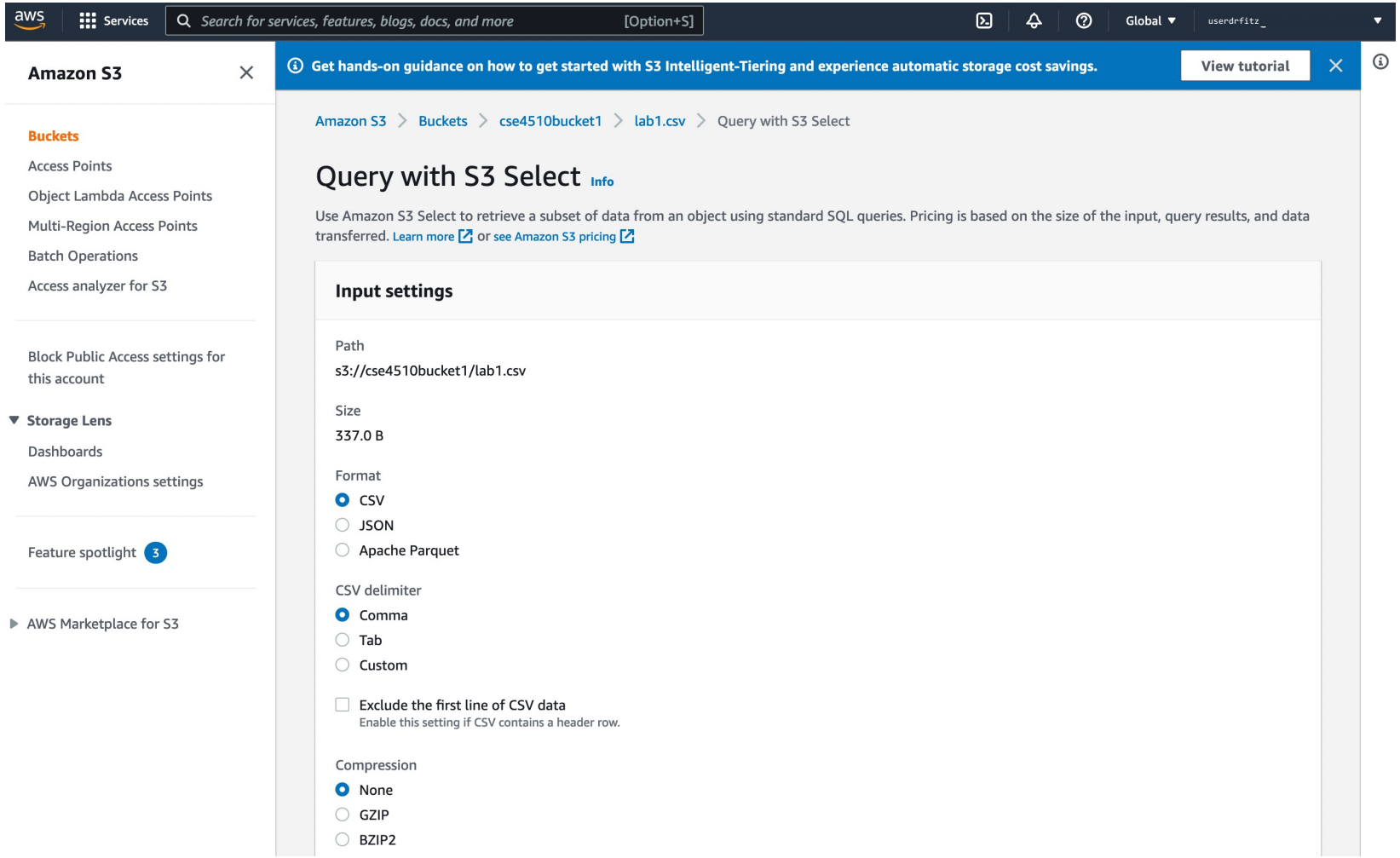
From the **Object actions** menu, choose **Query with S3 Select.**



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Query the object you uploaded

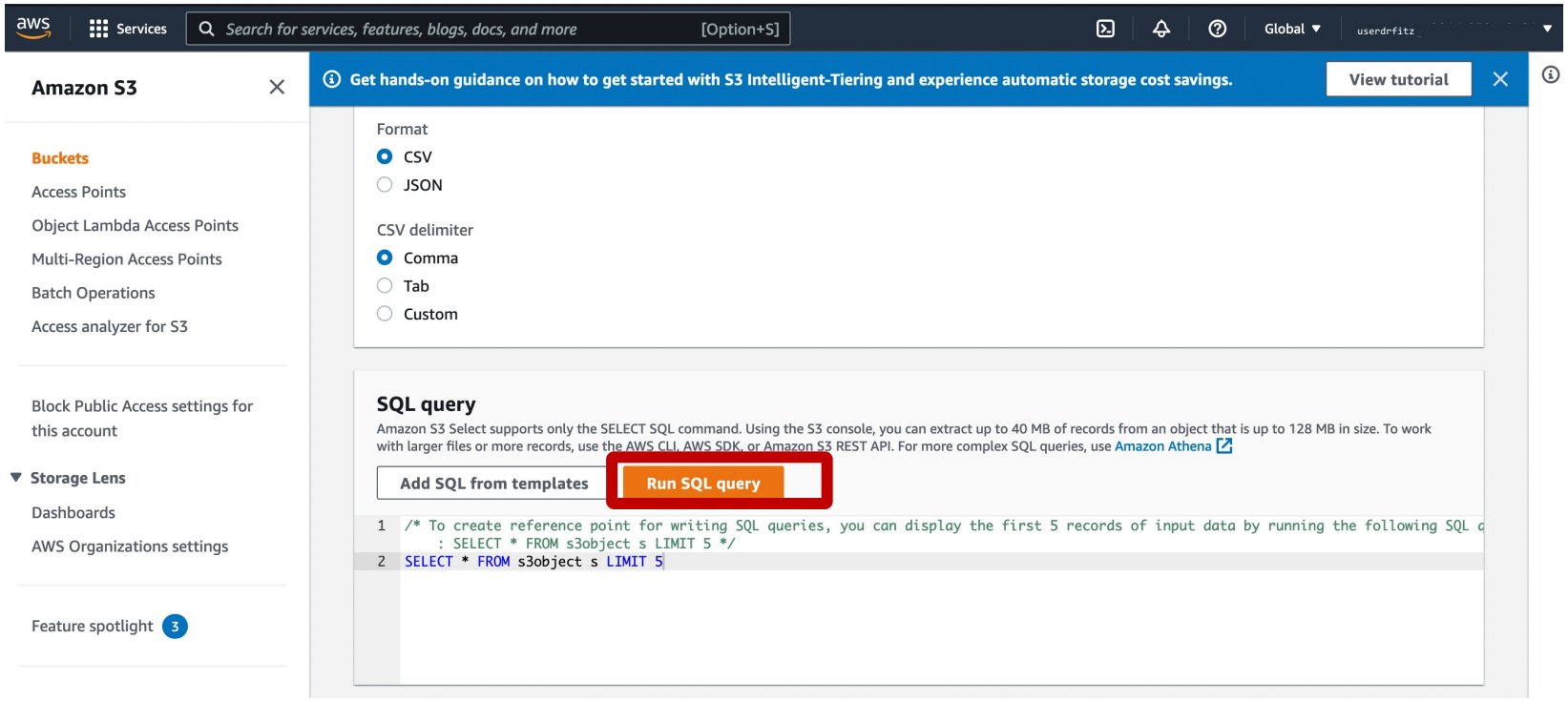
From the **Object actions** menu, choose **Query with S3 Select.**



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Query the object you uploaded

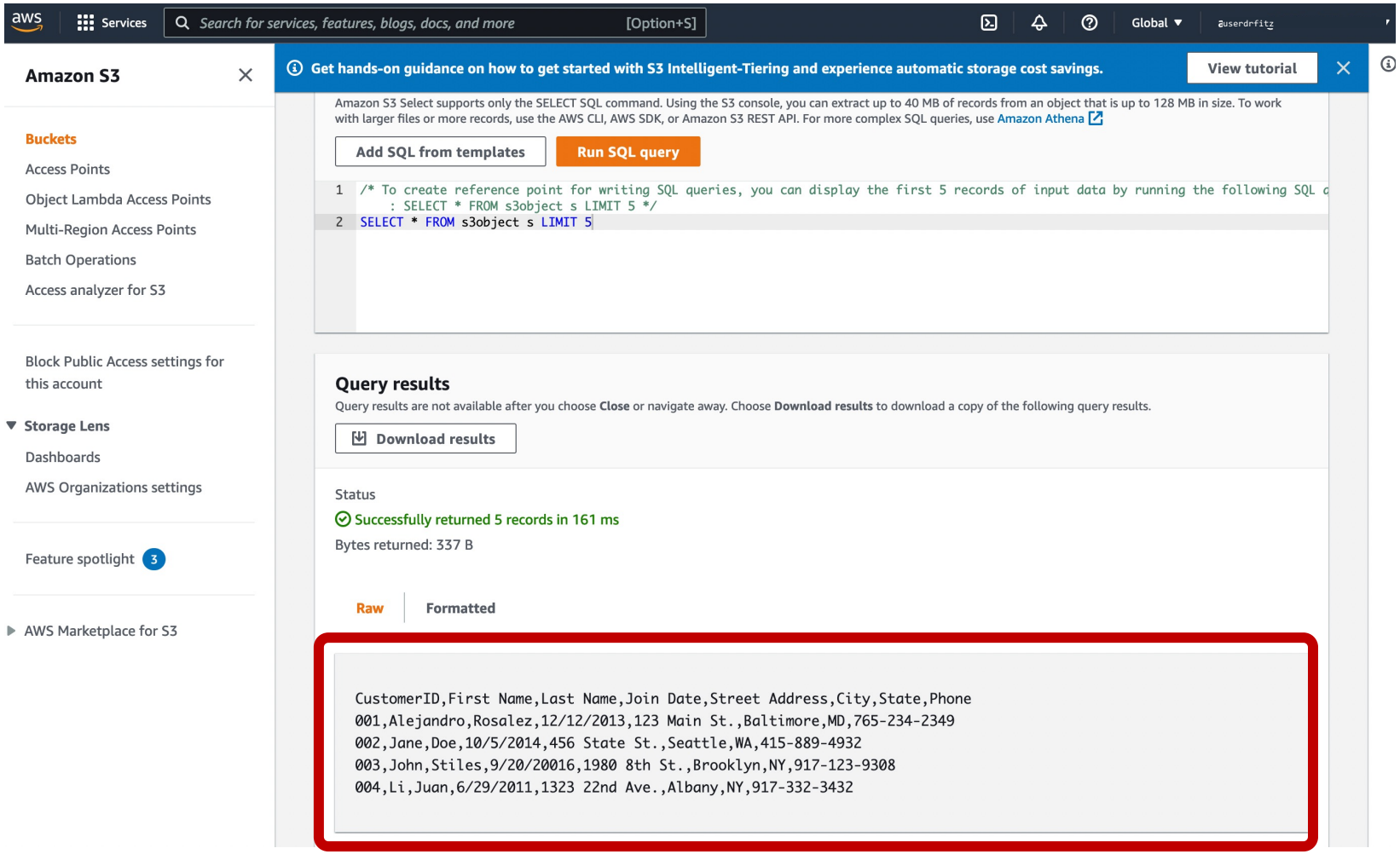
Scroll down the page and choose **Run SQL query**.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Query the object you uploaded

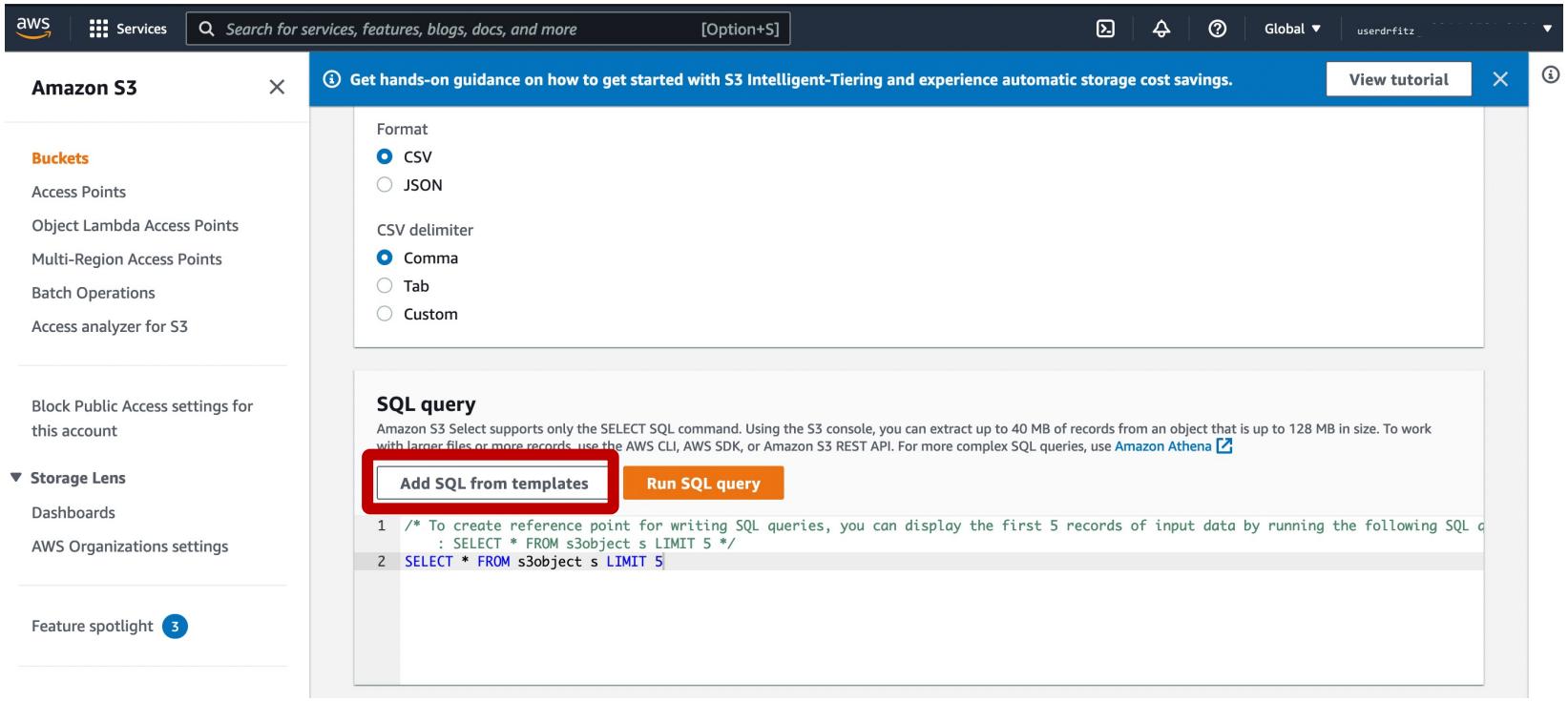
You should see the first few records from the file.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add SQL Template

Choose **Add SQL from templates**.

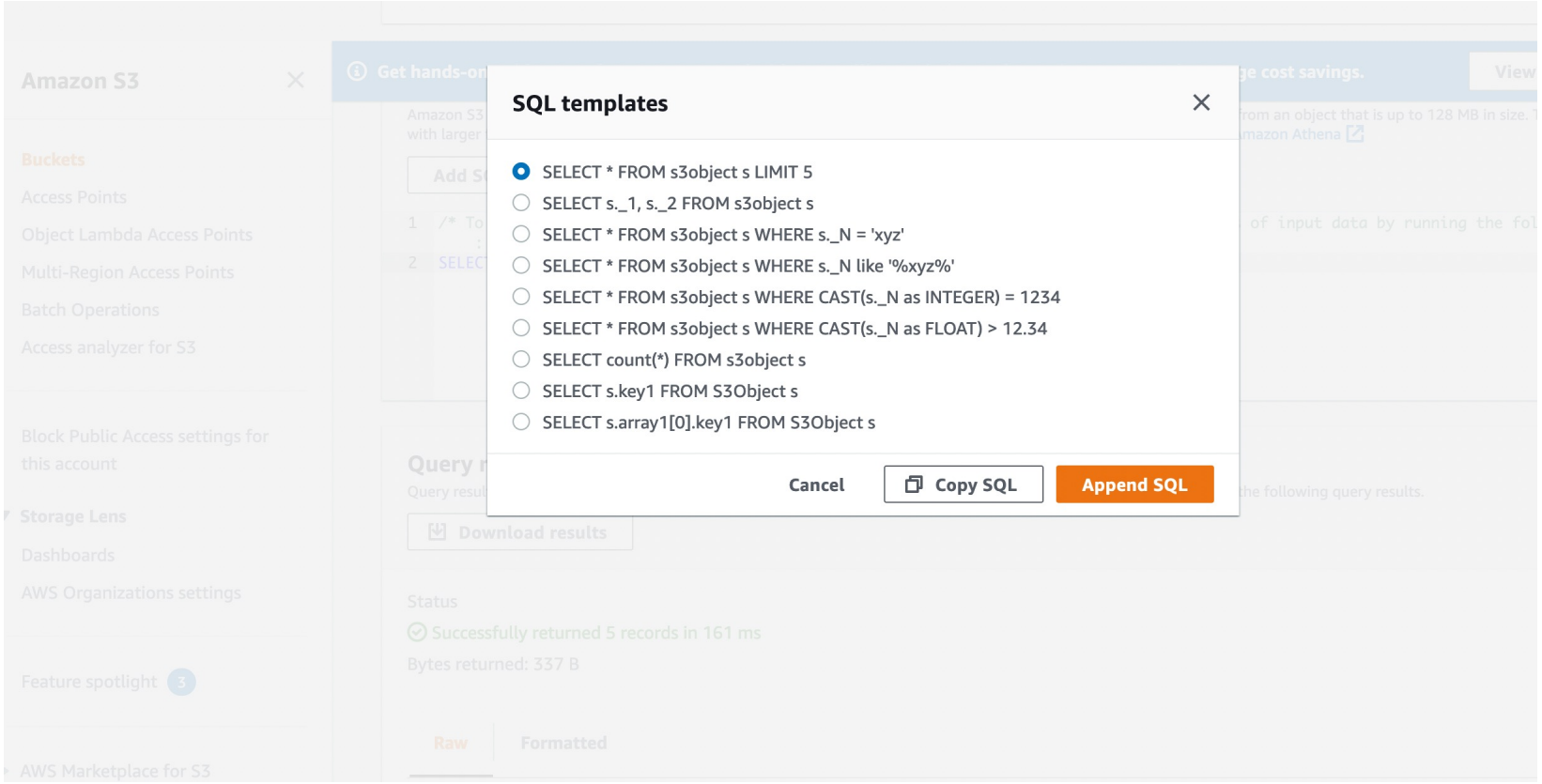


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add SQL Template

Choose SELECT COUNT * FROM s3object s.

Choose Copy SQL.



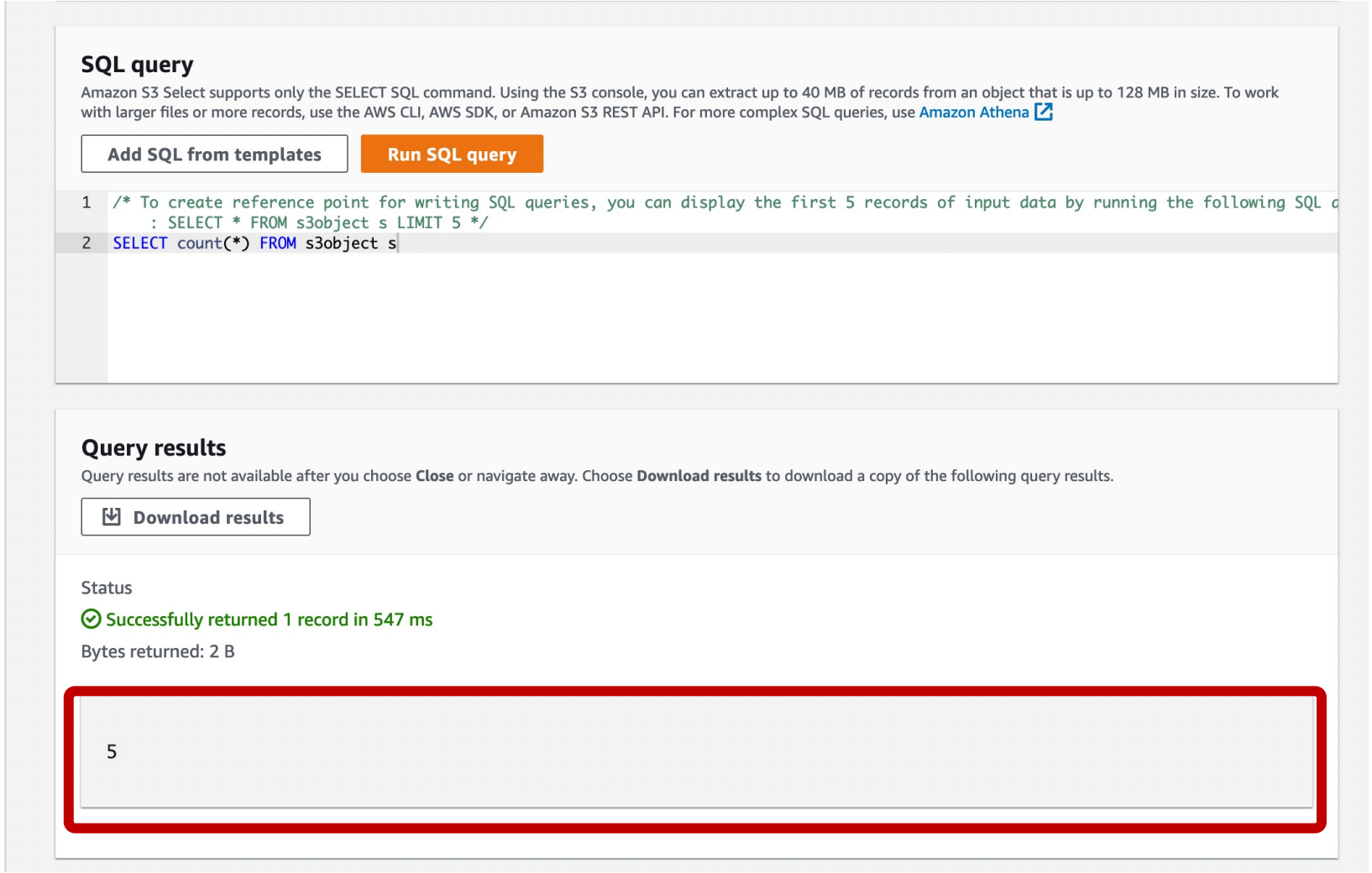
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Add SQL Template

Replace the previous query by deleting it and then paste the query you copied.

Choose Run SQL query.

In the **Result** pane, you should get the total number of records, which is 5.

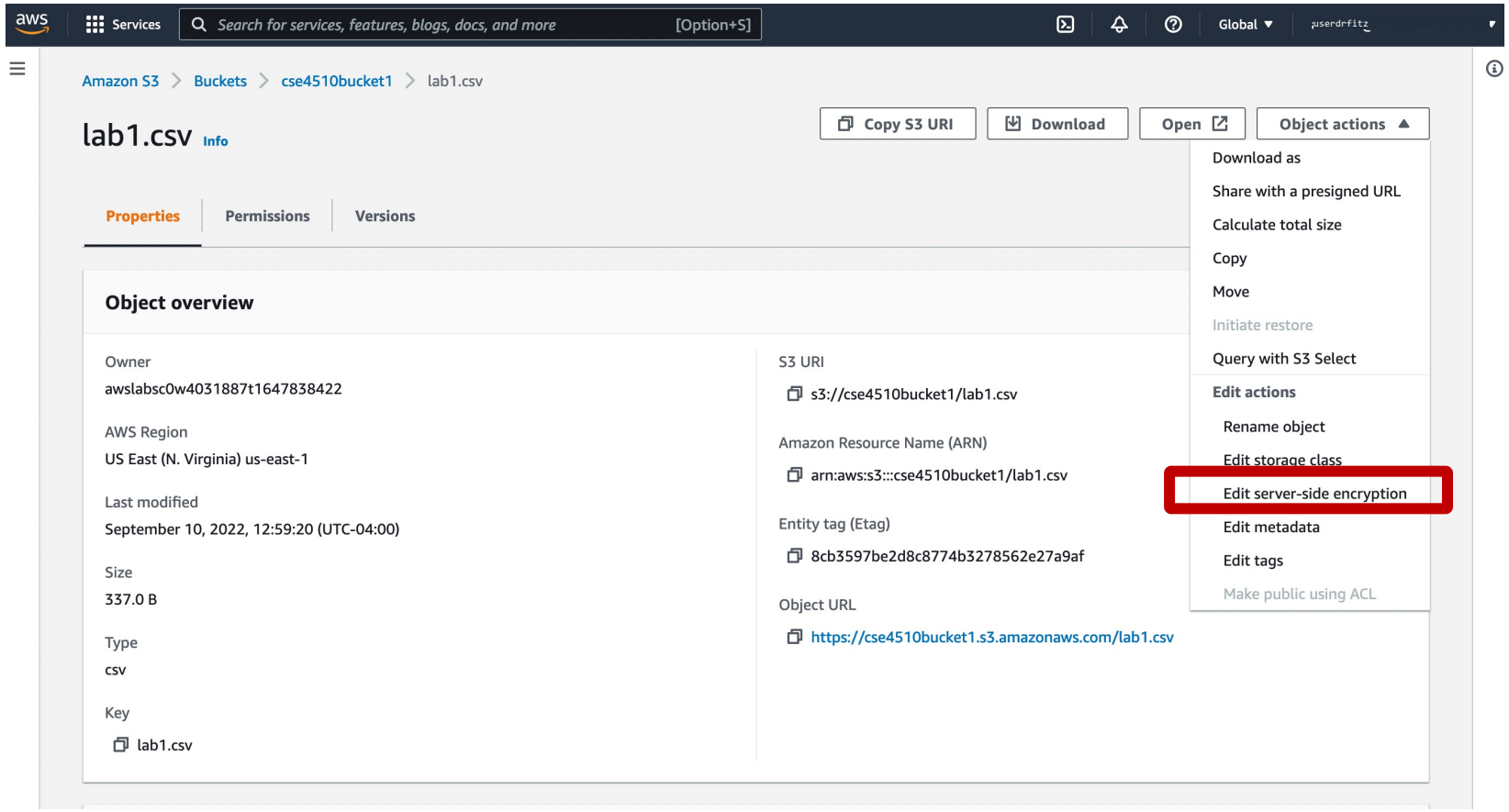


Change the encryption properties

You can set individual object properties—such as encryption at rest and storage class type—in the Amazon S3 console. Amazon S3 supports two kinds of encryption: Advanced Encryption Standard (AES)-256, and AWS Key Management Service (AWS KMS).

In the Amazon S3 console, choose the **lab1.csv** file.From the **Object actions** menu, choose **Edit server-side encryption**. Choose **Enable** and **Save changes**.To return to the object overview page, choose **Exit**.

From the **Object actions** menu, choose **Edit storage class**.Select **Intelligent-Tiering** and **Save changes**.

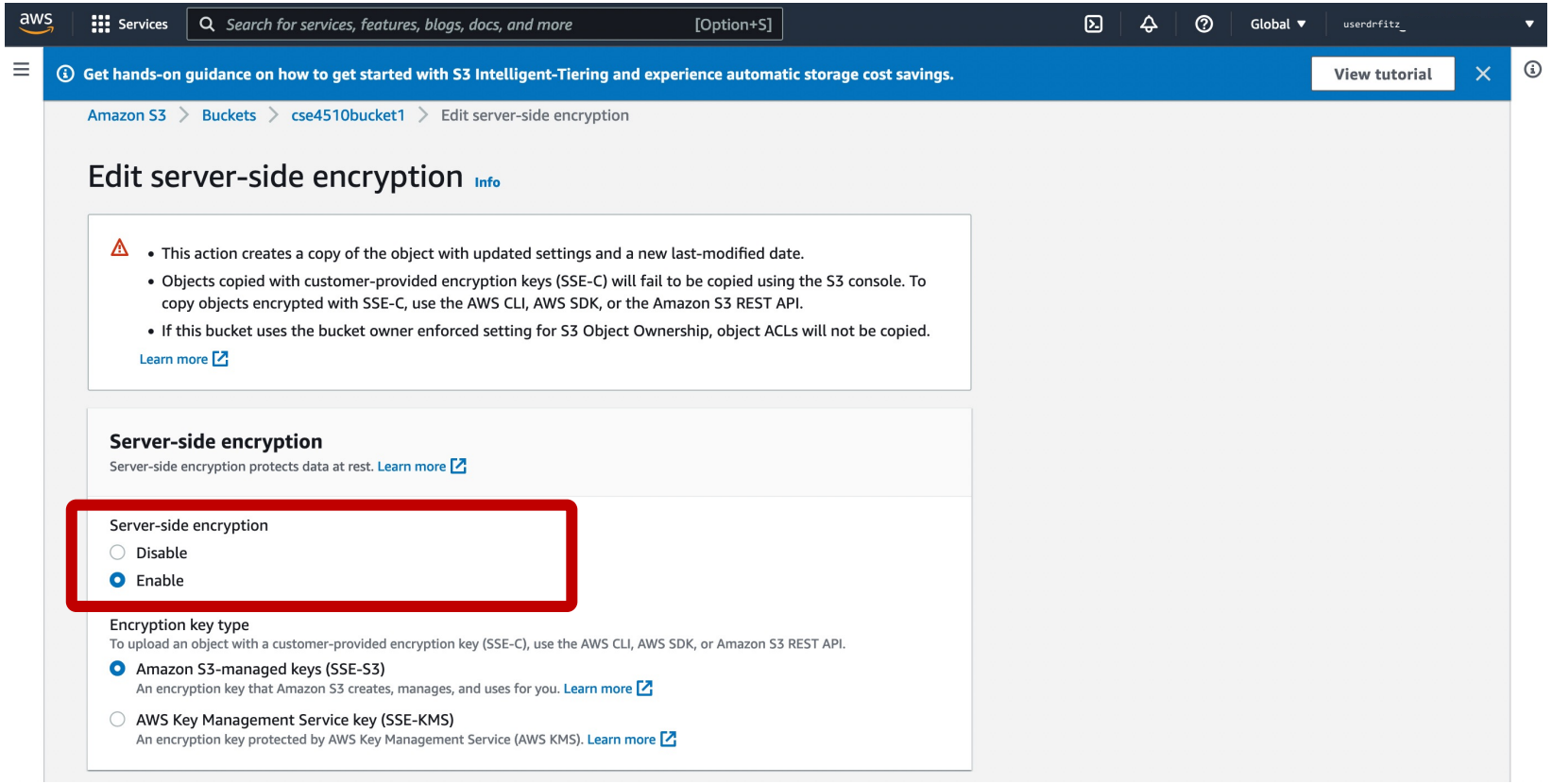


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Change the encryption properties

You can set individual object properties—such as encryption at rest and storage class type—in the Amazon S3 console. Amazon S3 supports two kinds of encryption: Advanced Encryption Standard (AES)-256, and AWS Key Management Service (AWS KMS).

In the Amazon S3 console, choose the **lab1.csv** file.From the **Object actions** menu, choose **Edit server-side encryption**. Choose **Enable** and **Save changes**.To return to the object overview page, choose **Exit**.

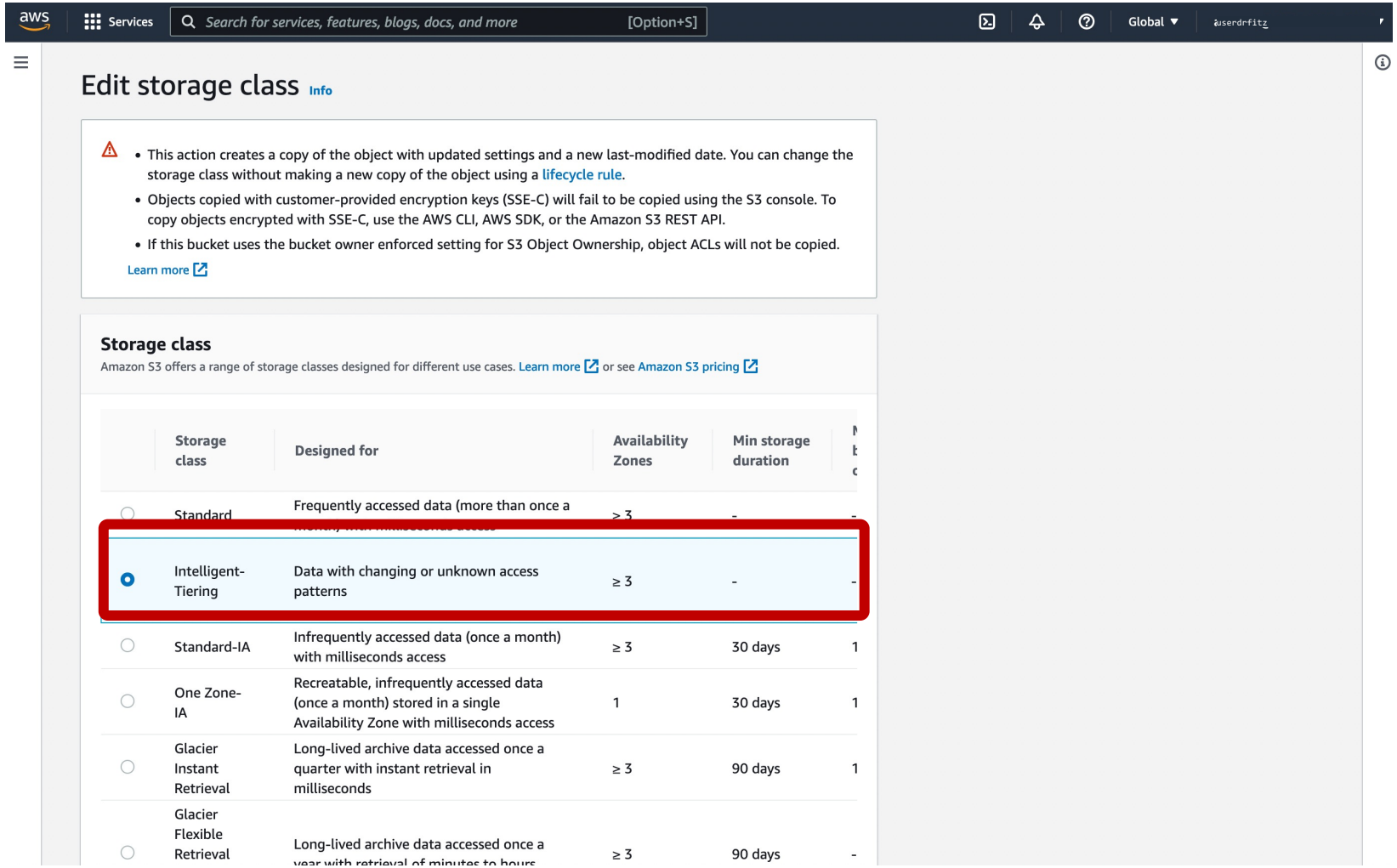


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Change the Storage Type properties

From the **Object actions** menu, choose **Edit storage class**.

Select **Intelligent-Tiering** and **Save changes**.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Storage Pricing

From the **Object actions** menu, choose **Edit storage class**.

Select **Intelligent-Tiering** and **Save changes**.

Region: US East (Ohio)

	Storage pricing
S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data	
First 50 TB / Month	$0.023 per GB
Next 450 TB / Month	$0.022 per GB
Over 500 TB / Month	$0.021 per GB
S3 Intelligent - Tiering* - Automatic cost savings for data with unknown or changing access patterns	
Monitoring and Automation, All Storage / Month (Objects > 128 KB)	$0.0025 per 1,000 objects
Frequent Access Tier, First 50 TB / Month	$0.023 per GB
Frequent Access Tier, Next 450 TB / Month	$0.022 per GB
Frequent Access Tier, Over 500 TB / Month	$0.021 per GB
Infrequent Access Tier, All Storage / Month	$0.0125 per GB
Archive Instant Access Tier, All Storage / Month	$0.004 per GB
S3 Intelligent - Tiering* - Optional asynchronous Archive Access tiers	
Archive Access Tier, All Storage / Month	$0.0036 per GB
Deep Archive Access Tier, All Storage / Month	$0.00099 per GB
S3 Standard - Infrequent Access** - For long lived but infrequently accessed data that needs millisecond access	
All Storage / Month	$0.0125 per GB
S3 One Zone - Infrequent Access** - For re-createable infrequently accessed data that needs millisecond access	
All Storage / Month	$0.01 per GB
S3 Glacier Instant Retrieval*** - For long-lived archive data accessed once a quarter with instant retrieval in milliseconds	
All Storage / Month	$0.004 per GB
S3 Glacier Flexible Retrieval (Formerly S3 Glacier)*** - For long-term backups and archives with retrieval option from 1 minute to 12 hours	
All Storage / Month	$0.0036 per GB
S3 Glacier Deep Archive*** - For long-term data archiving that is accessed once or twice in a year and can be restored within 12 hours	
All Storage / Month	$0.00099 per GB

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Upload Compressed Files

For big data scenarios, you should generally store compressed files in Amazon S3. Amazon S3 supports the .gzip and .bzip2 compression formats. Uploading a compressed file is essentially the same as uploading a file that is notebook compressed.

In this task, you will upload a file that is compressed as a .gzip file. First, download the compressed file from Canvas and save it to a local directory.

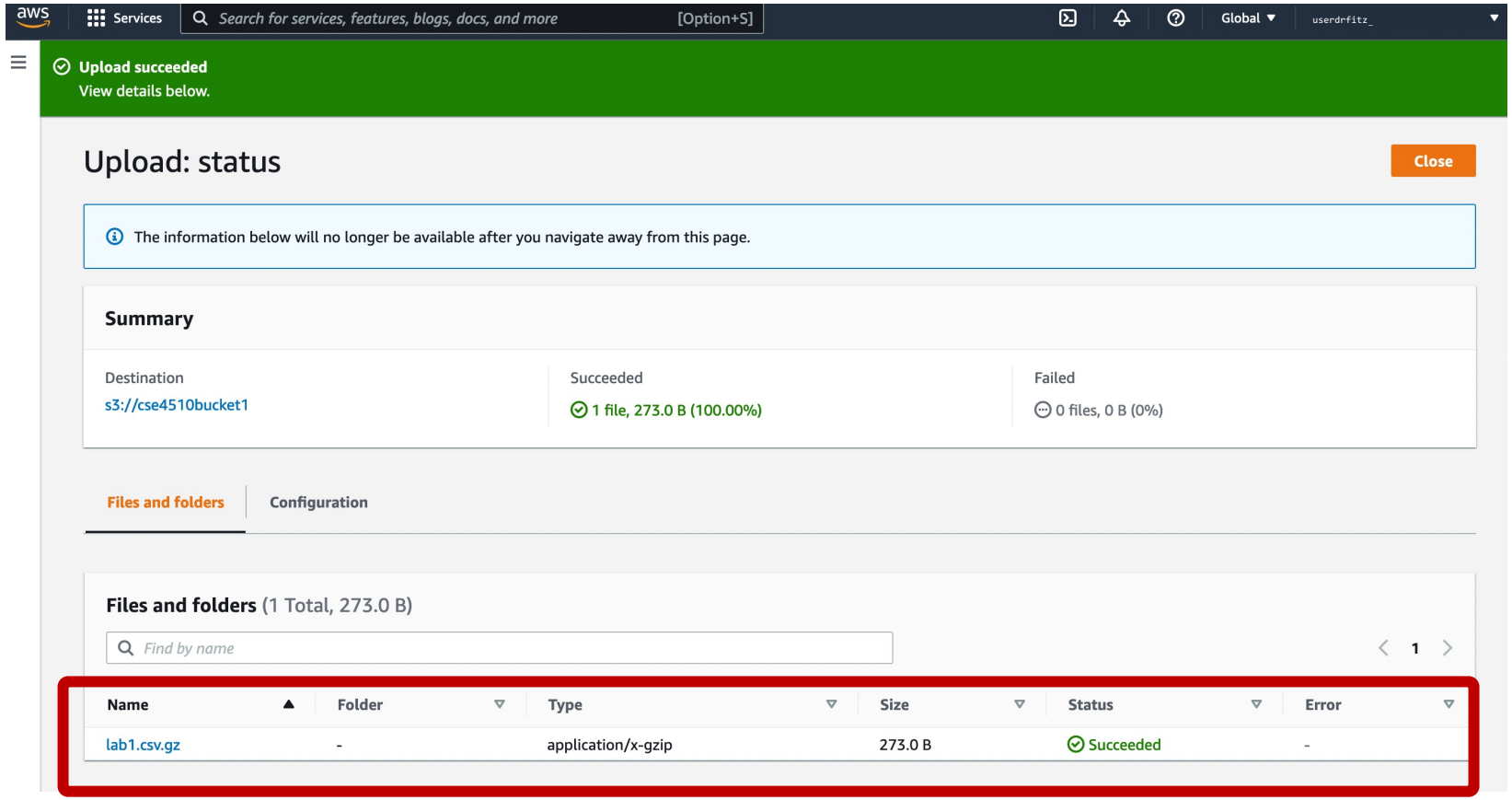
In the Amazon S3 console, choose your bucket from the breadcrumbs again.

Choose **Upload**.

Choose **Add files**, and choose the **lab1.csv.gz** file that you downloaded previously.

Choose **Upload**.

Select the **lab1.csv.gz** file.

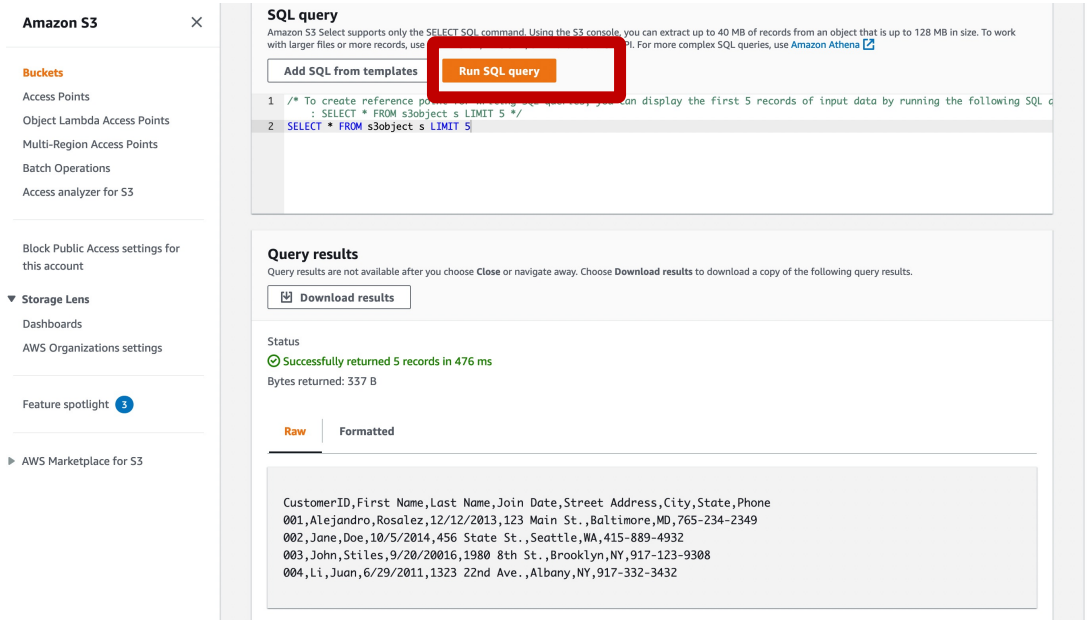


Querying Compressed Data in S3

From the **Object actions** menu, choose **Query with S3 Select.**

Scroll down the page and choose **Run SQL query.**

You should get results that demonstrate that you can query the compressed file in the same way as a non-compressed file



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Working with S3 via Code

THIS SECTION IS TARGETED TO THOSE USING THE BIG DATA CONTAINERS COURSE WITHOUT FOLLOWING AWS ACADEMY'S GUIDED LABS.

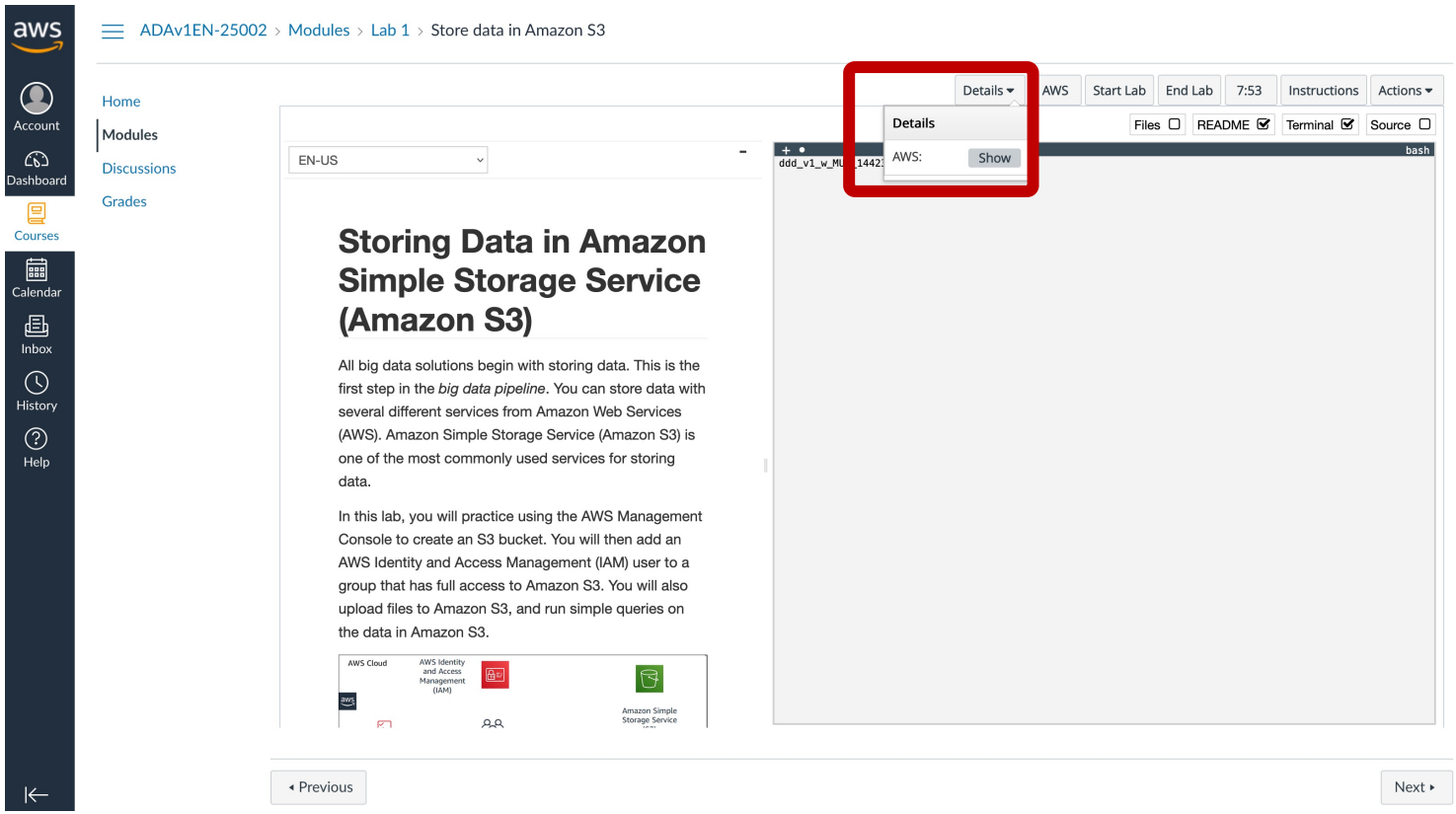
IF YOU HAVE TIME REMAINING ON YOUR GUIDED LAB, YOU MAY ALSO ACCESS THE PLATFORM VIA CODE.

Accessing Credentials & Starting a Session

First click start lab to start a lab.

Then click Details→Show to access your credentials.

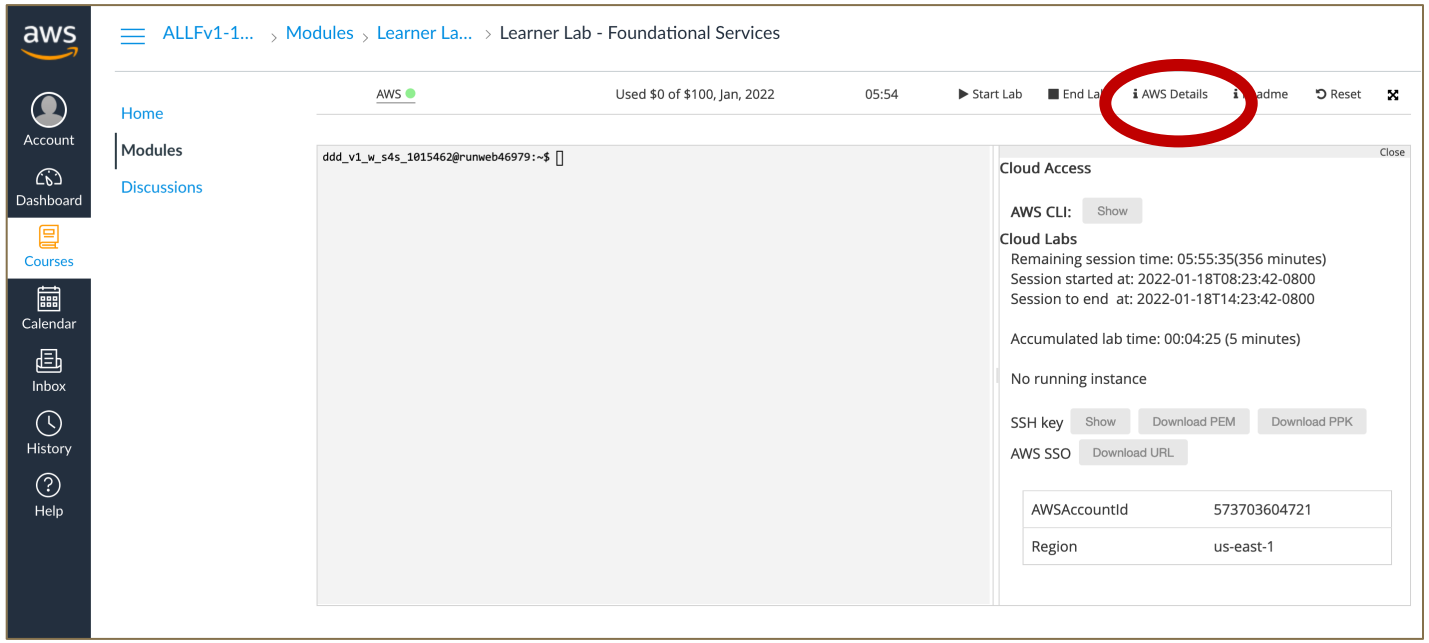
If you are working with the Big Data Containers course, your screen will rexemble the one on the next slide.



Access AWS Details

Click AWS details to access AWS information such as keys for accessing resources using SSH, etc.

Click Show to see your credentials for using the AWS CLI.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Set up Credentials

Create a file named credentials inside the folder ~/.aws

Run the following from your Terminal/cmd to create the file

cd ~ or cd %USERPROFILE% for Windows

mkdir .aws

cd .aws

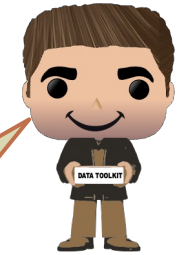
sudo nano credentials

Windows users:

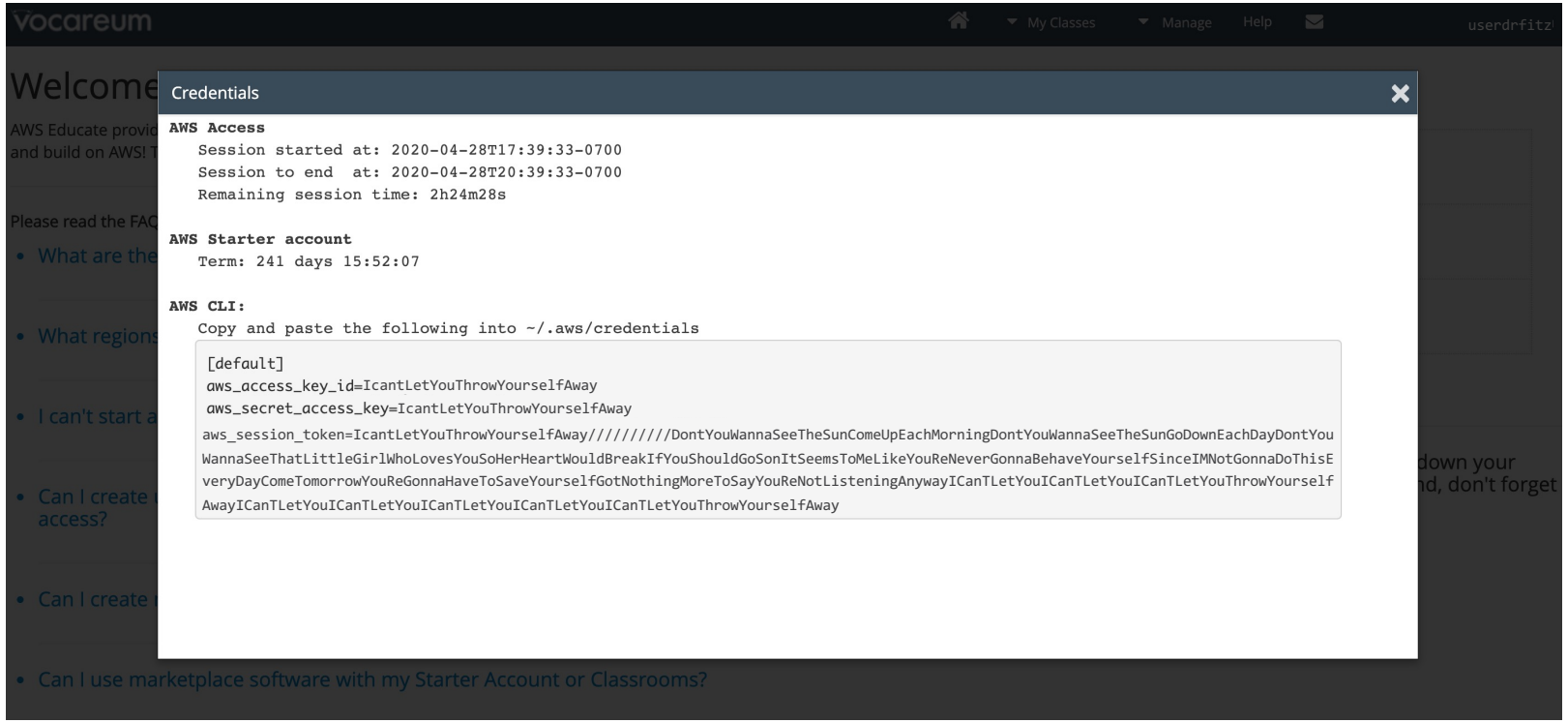
Do not use sudo

To get nano, Install the Chocolatey (https://chocolatey.org/) package manager and enter

choco install nano



Keep track of the session time and
Refresh temporary credentials five minutes before their expiration.



Set up Credentials & Starting a Session

You may also store your credentials in a file or hard-code them (not recommended).

You may also set up multiple sets of credentials (profiles) in the same file and specify the profile name when connecting to s3.

```
Approach 2 – Hard-code or read from a text file
import boto3 session = boto3.Session( 's3',
          aws_access_key_id=ACCESS_KEY,
          aws_secret_access_key=SECRET_KEY,
          aws_session_token=SESSION_TOKEN )
```

```
Approach 3 – Specify multiple accounts in a Shared Credentials File
[default]
aws_access_key_id=foo
aws_secret_access_key=bar

[dev]
aws_access_key_id=foo2
aws_secret_access_key=bar2

[prod]
aws_access_key_id=foo3
aws_secret_access_key=bar3
```

```
import boto3
session = boto3.Session(profile_name='dev')
dev_s3_client = session.client('s3')

dynamodb = session.resource('s3', region_name='us-east-1' )
```

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Working with S3 Buckets via Code

Use the following examples to access your S3 Bucket via Python code.

```
import boto3
# Creating the low level functional client
client = boto3.client(
    's3',
    aws_access_key_id='AKIA46SFIWN5AMWMDQVB',
    aws_secret_access_key='yuHNxlcbEx7b9Vs6QEo2KWiaAPxj/k6RdEY4DfeS',
    region_name='ap-south-1'
)

# Creating the high level object oriented interface
resource = boto3.resource(
    's3',
    aws_access_key_id='AKIA46SFIWN5AMWMDQVB',
    aws_secret_access_key='yuHNxlcbEx7b9Vs6QEo2KWiaAPxj/k6RdEY4DfeS',
    region_name='ap-south-1'
)
```

https://boto3.amazonaws.com/v1/documentation/api/latest/guide/s3-examples.html

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Working with S3 Buckets via Code

Use the following examples to access your S3 Bucket via Python code.

```
# Create the S3 object
obj = client.get_object(
    Bucket='sql-server-shack-demo-1',
    Key='sql-shack-demo.csv'
)

# Read data from the S3 object
data = pandas.read_csv(obj['Body'])

# Print the data frame
print('Printing the data frame...')
print(data)
```

https://boto3.amazonaws.com/v1/documentation/api/latest/guide/s3-examples.html

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

References

The following references were used to create this tutorial.

- AWS Academy
- https://boto3.amazonaws.com/v1/documentation/api/latest/guide/s3-examples.html

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY