CSP 554—Big Data Technologies

Assignment #1 (Modules 01a & 01b, 12 points)

1. **Obtain our texts**

   The texts were downloaded from the Internet

2. **Read from (TW)**

   Read in brief about the Chapter 1

3. **(2 points) Submit very brief answers (or bullet points) to the following questions:**
   - **(Remote/Online Students Only) What location or time zone are you in when you attend the course?**
     I am a Main Campus Student. I am in Chicago and I follow Central Time Zone
   - **Describe any prior experience you might have with use of public cloud, data mining, machine learning, statistics, data science and big data.**
     During Fall 2018, I had taken CS422. In my Undergraduate I had completed a course certification on Hadoop. I am familiar with certain concepts of Big Data
   - **Share any big data interests and personal learning goals for the course.**
     When browsing through the syllabus, I came across concepts of Pig, Hive and Spark which is currently trending. I have a peak interest in learning them and applying them in the course
   - **Indicate if there are additional topics in the scope of the course of special interest to you.**
     The course is on AWS. It would be helpful if have a brief introduction on Microsoft Azure
   - **Do you have any anticipated personal issues such as expected absences or other necessary accommodations with course impact? (Of course, these will be held in strictest confidence.)**
     No

4. **Read article on "Blackboard" in Articles section**
   Article reading completed

5. **(5 points) Summarize the main points of the above article and your thoughts (questions you might want to ask the authors, areas where you disagree, other comments)**
   - No more than about ½ page single spaced

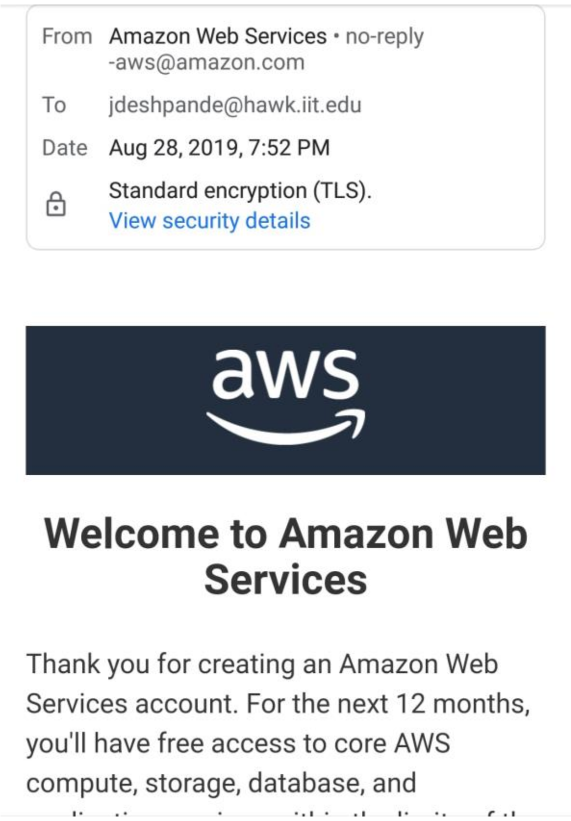
   From the article:

The article mainly talks about Google Flu Trends (GFT) started by Google and how it failed to predict influenza-like illness (ILI). The author says that the failure was due to big data hubris and algorithm dynamics. According to author, for big data hubris an example has been cited which states that if there is big data and small number of test cases there is a problem of overfitting. The author talks about temporal correlation stating that the mistakes repeated in previous week are repeated. To overcome this, the author states that use of traditional statistical methods needs to be done. The question I need to ask the author is that how reliable and accurate would the statistical methods be in predicting data? And if they were would it be able to work on big data. The other point highlighted by the author is that Google used GFT to correlate with CDC and there was an error where it considered the treatment for cold and flu same. The author talks about blue team where the algorithm is modified by the service provider according to the model and red team where it manipulates the data generating process for their own gains. The author states that companies cannot access data from Google due to Google's private policies but also says that the data should be transparent. So, to gain maximum throughput, big data methodologies and traditional methods should work together. They should not be allowed to access our private data and predict wrong results.

- **Submit via blackboard**
  Submitted through this document

6. **(5 points) Set up an Amazon Web Services (AWS) cloud account, if you don't already have one, and follow the tutorial about how to work with a storage service called S3. Since we will do most of our assignments using AWS, this will get you started. In a while we will come to understand S3 as one critical element of a big data processing architecture know as the "data lake."**

   Account setup done. Attached screenshot of confirmation email from AWS for reference.



From Amazon Web Services • no-reply -aws@amazon.com

To jdeshpande@hawk.iit.edu

Date Aug 28, 2019, 7:52 PM

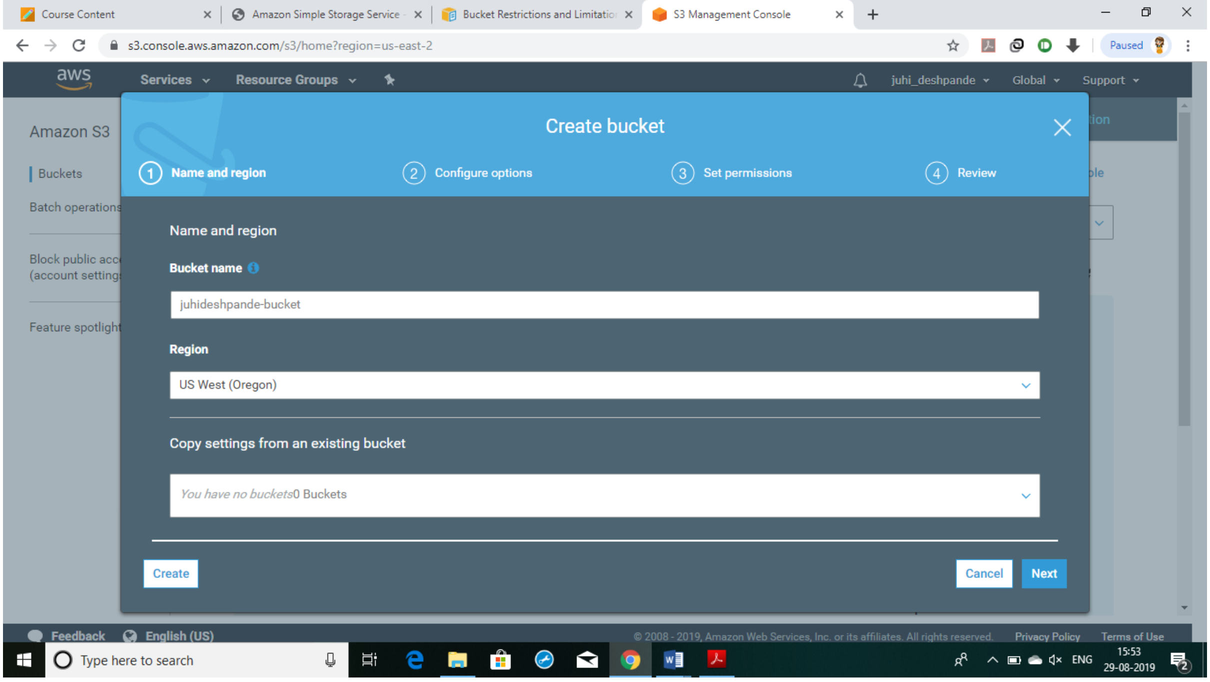
Standard encryption (TLS).
View security details

Welcome to Amazon Web Services

Thank you for creating an Amazon Web Services account. For the next 12 months, you'll have free access to core AWS compute, storage, database, and

Bucket creation

Step 1: Creating a bucket



Create bucket

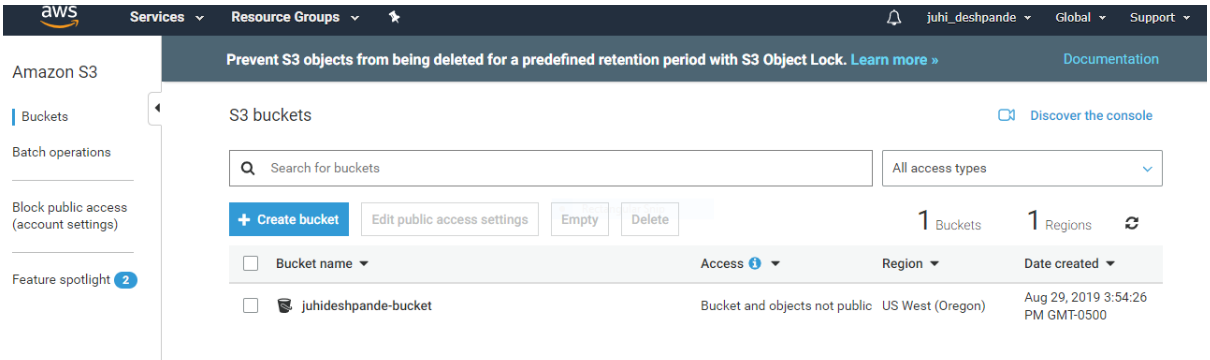
Name and region

Bucket name: juhideshpande-bucket

Region: US West (Oregon)

Copy settings from an existing bucket: *You have no buckets*0 Buckets

Step 2: Bucket created



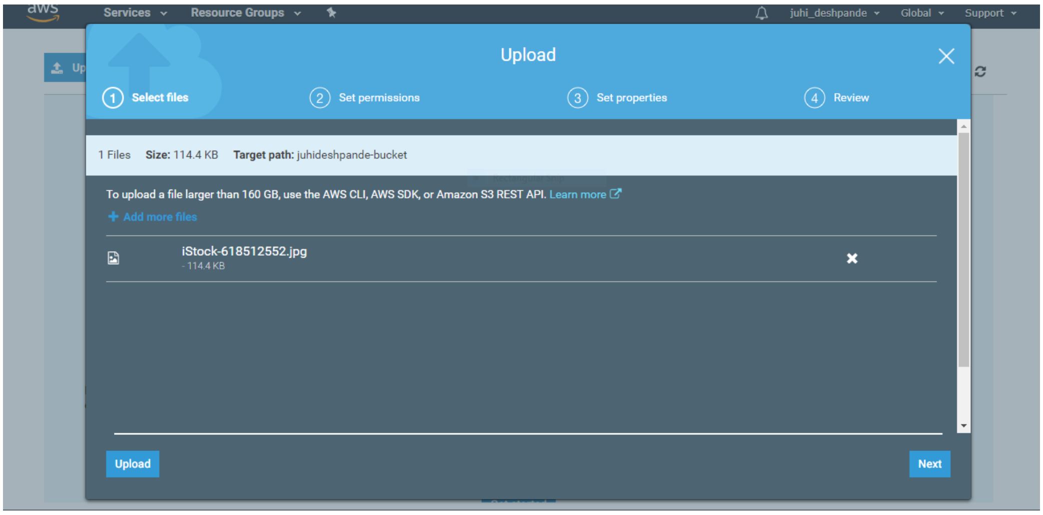
Prevent S3 objects from being deleted for a predefined retention period with S3 Object Lock. Learn more »

S3 buckets

1 Buckets 1 Regions

☐	Bucket name	Access	Region	Date created
☐	juhideshpande-bucket	Bucket and objects not public	US West (Oregon)	Aug 29, 2019 3:54:26 PM GMT-0500

Step 3: Adding object (image file) to bucket



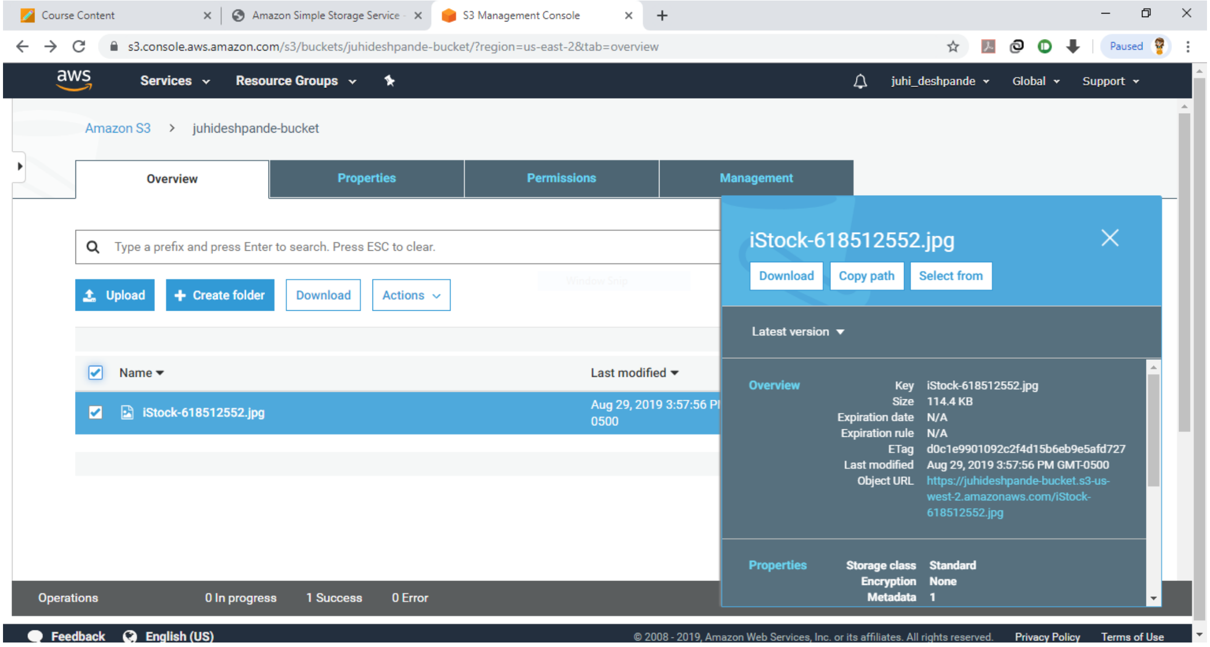
Upload

1 Files **Size:** 114.4 KB **Target path:** juhideshpande-bucket

To upload a file larger than 160 GB, use the AWS CLI, AWS SDK, or Amazon S3 REST API. Learn more

iStock-618512552.jpg
- 114.4 KB

Step 4: Downloading and viewing the object



☑ iStock-618512552.jpg — Aug 29, 2019 3:57:56 PM -0500

iStock-618512552.jpg

Download | Copy path | Select from

Latest version

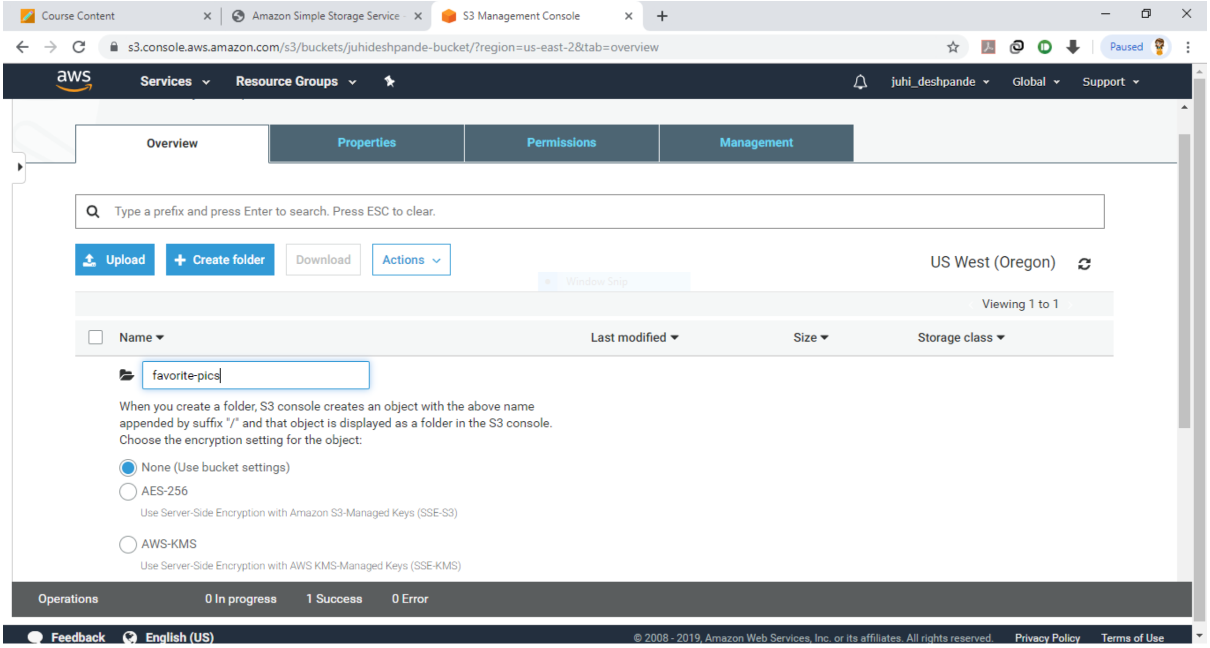
Overview

Key	iStock-618512552.jpg
Size	114.4 KB
Expiration date	N/A
Expiration rule	N/A
ETag	d0c1e9901092c2f4d15b6eb9e5afd727
Last modified	Aug 29, 2019 3:57:56 PM GMT-0500
Object URL	https://juhideshpande-bucket.s3-us-west-2.amazonaws.com/iStock-618512552.jpg

Properties

Storage class	Standard
Encryption	None
Metadata	1

Step 5: Creating a folder to move the object



favorite-pics

When you create a folder, S3 console creates an object with the above name appended by suffix "/" and that object is displayed as a folder in the S3 console. Choose the encryption setting for the object:

● None (Use bucket settings)

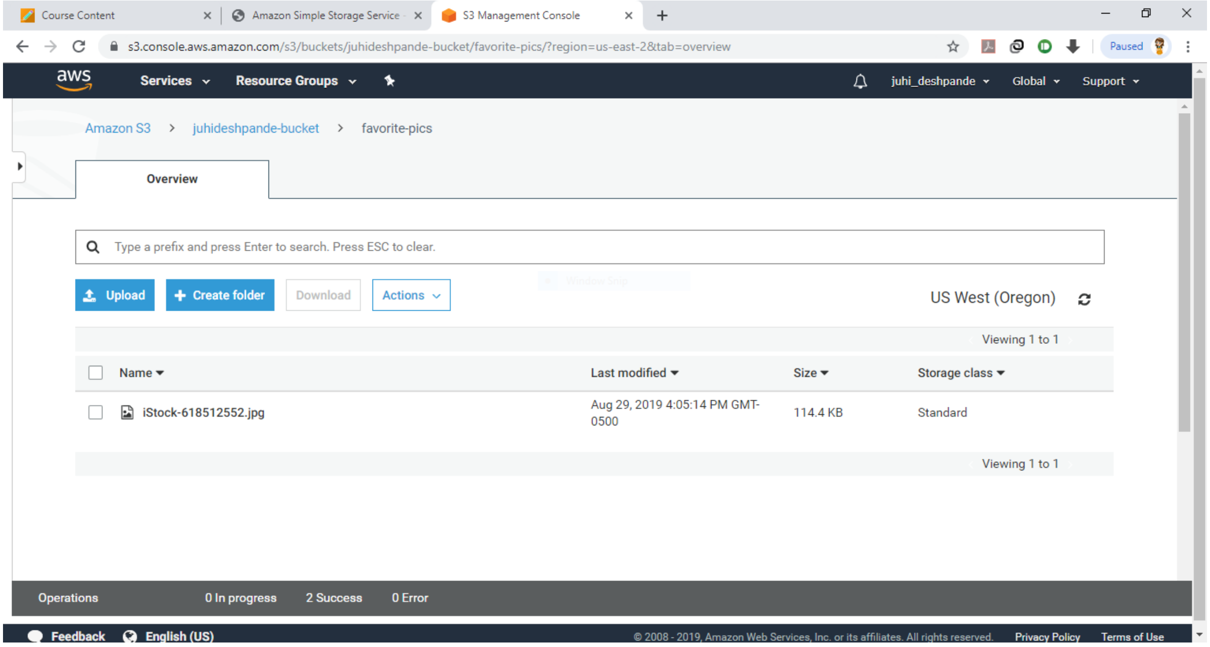
○ AES-256

Use Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)

○ AWS-KMS

Use Server-Side Encryption with AWS KMS-Managed Keys (SSE-KMS)

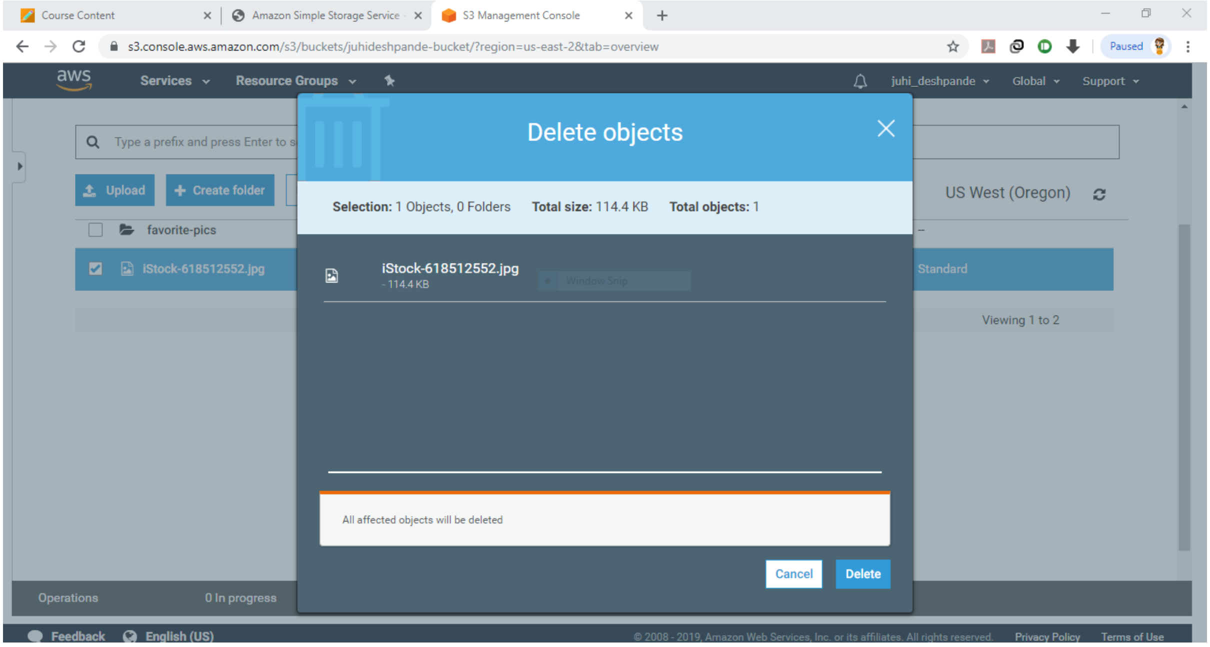
Step 6: Created the folder. Copied the image from juhideshpande-bucket and pasted to favorite-pics folder



Course Content
Amazon Simple Storage Service
S3 Management Console
s3.console.aws.amazon.com/s3/buckets/juhideshpande-bucket/favorite-pics/?region=us-east-2&tab=overview
Services
Resource Groups
juhi_deshpande
Global
Support
Amazon S3 > juhideshpande-bucket > favorite-pics
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
Download
Actions
US West (Oregon)
Viewing 1 to 1
Operations
0 In progress
2 Success
0 Error
Feedback
English (US)
© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Name	Last modified	Size	Storage class
iStock-618512552.jpg	Aug 29, 2019 4:05:14 PM GMT-0500	114.4 KB	Standard

Step 7: Deleting object from bucket



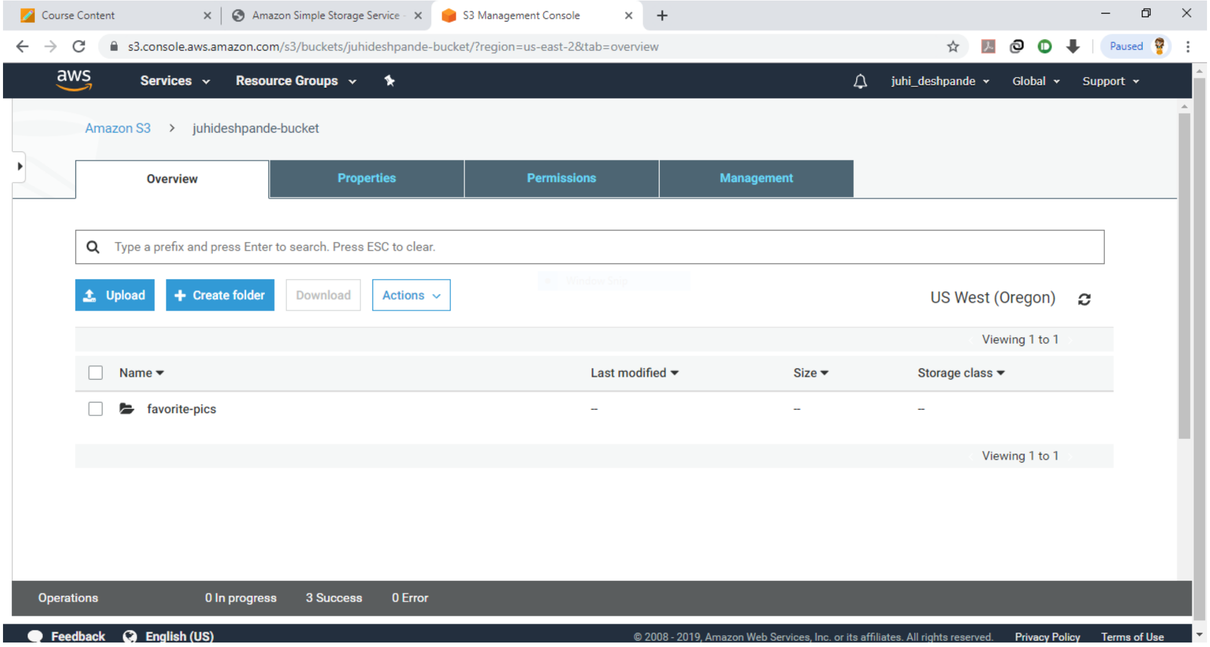
Delete objects

Selection: 1 Objects, 0 Folders Total size: 114.4 KB Total objects: 1

iStock-618512552.jpg
- 114.4 KB

All affected objects will be deleted

Step 8: Bucket after deletion of object



s3.console.aws.amazon.com/s3/buckets/juhideshpande-bucket/?region=us-east-2&tab=overview

Amazon S3 > juhideshpande-bucket

Overview | Properties | Permissions | Management

Type a prefix and press Enter to search. Press ESC to clear.

Upload | Create folder | Download | Actions

US West (Oregon)

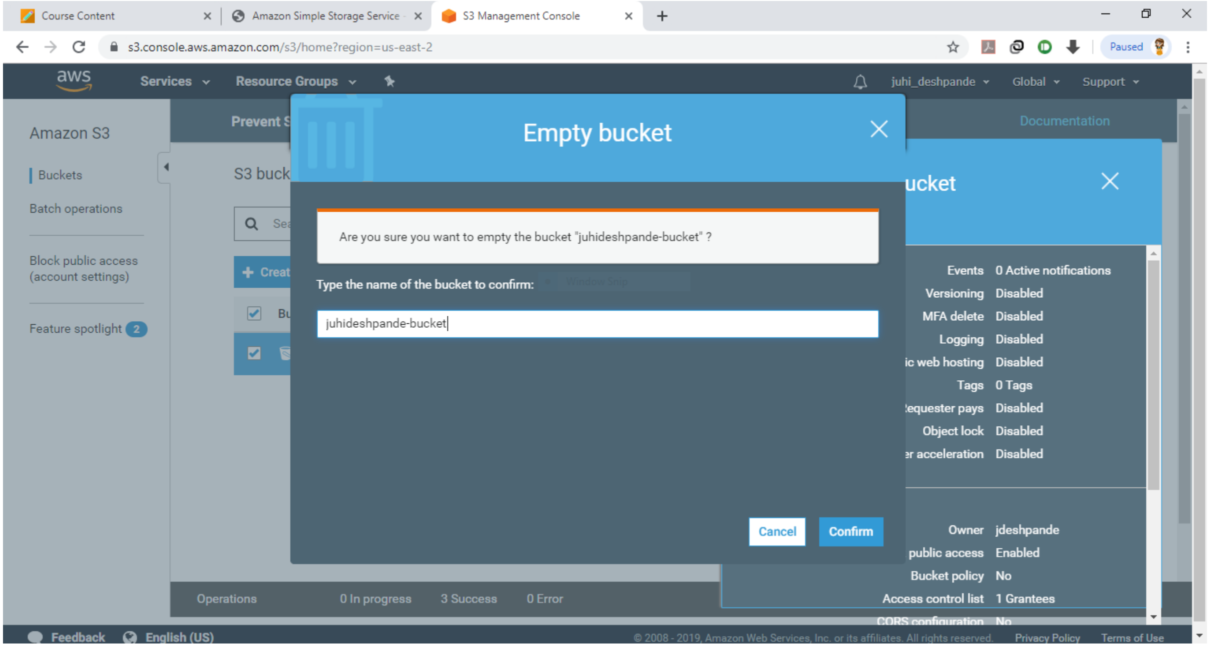
Viewing 1 to 1

Name	Last modified	Size	Storage class
favorite-pics	–	–	–

Viewing 1 to 1

Operations 0 In progress 3 Success 0 Error

Step 9: Screenshot only how to empty bucket



s3.console.aws.amazon.com/s3/home?region=us-east-2

Empty bucket

Are you sure you want to empty the bucket "juhideshpande-bucket" ?

Type the name of the bucket to confirm:

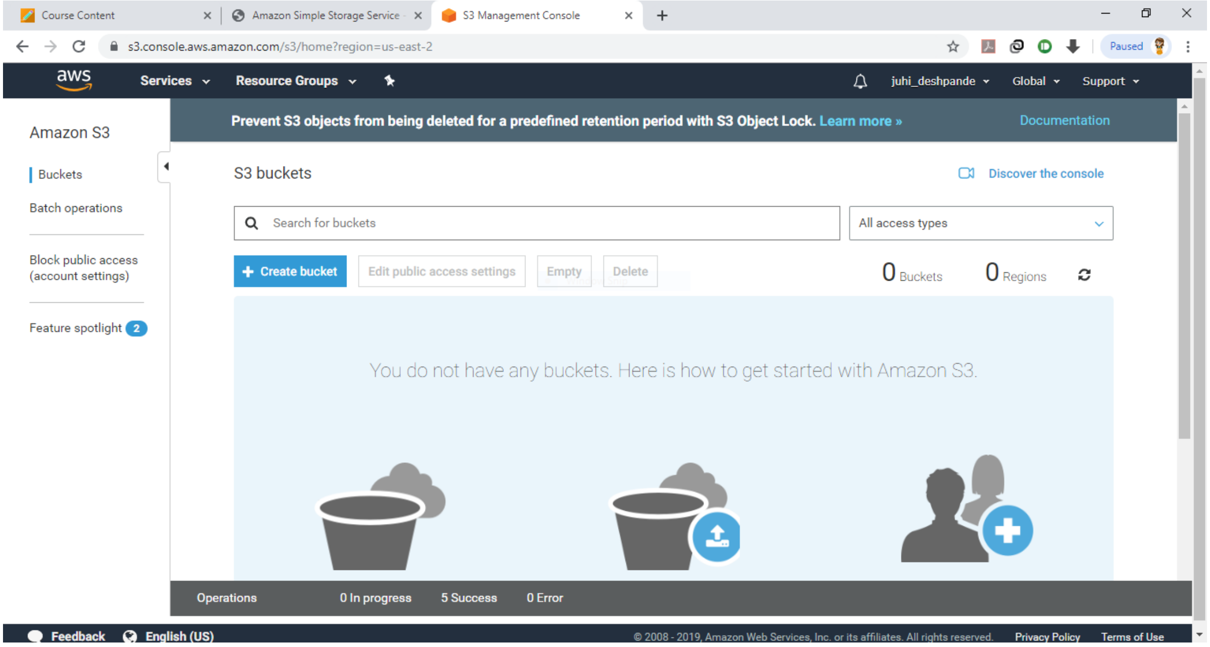
juhideshpande-bucket

Cancel | Confirm

Events	0 Active notifications
Versioning	Disabled
MFA delete	Disabled
Logging	Disabled
ic web hosting	Disabled
Tags	0 Tags
equester pays	Disabled
Object lock	Disabled
er acceleration	Disabled
Owner	jdeshpande
public access	Enabled
Bucket policy	No
Access control list	1 Grantees
CORS configuration	No

Operations 0 In progress 3 Success 0 Error

Step 10: Deleted the bucket instead of emptying. Screenshot displays the deleted bucket



s3.console.aws.amazon.com/s3/home?region=us-east-2

Prevent S3 objects from being deleted for a predefined retention period with S3 Object Lock. Learn more »

S3 buckets

0 Buckets 0 Regions

You do not have any buckets. Here is how to get started with Amazon S3.

Operations 0 In progress 5 Success 0 Error

© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.