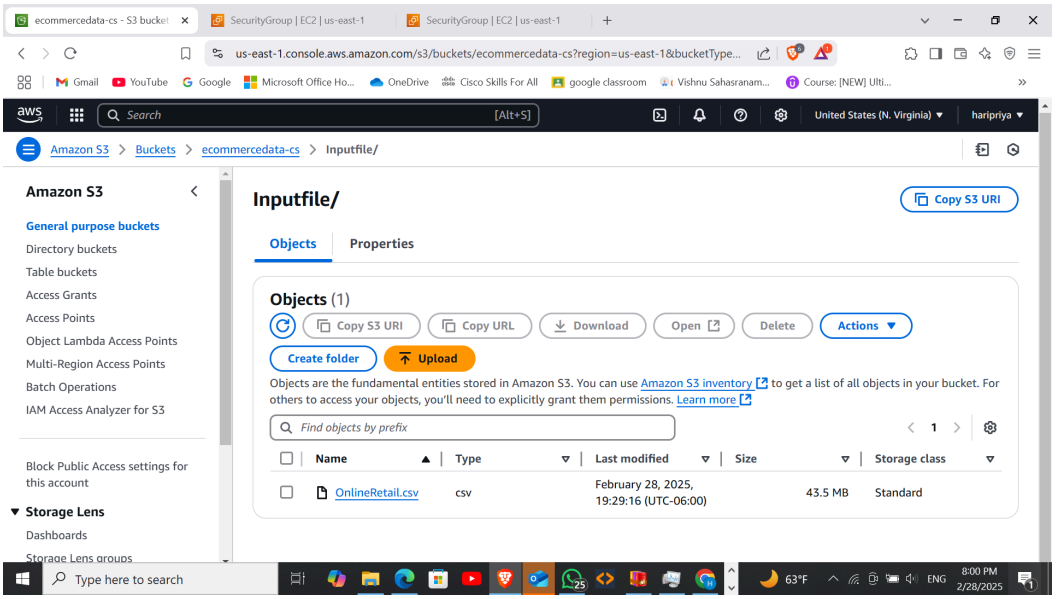
Big Data Driven Customer Segmentation In E-commerce

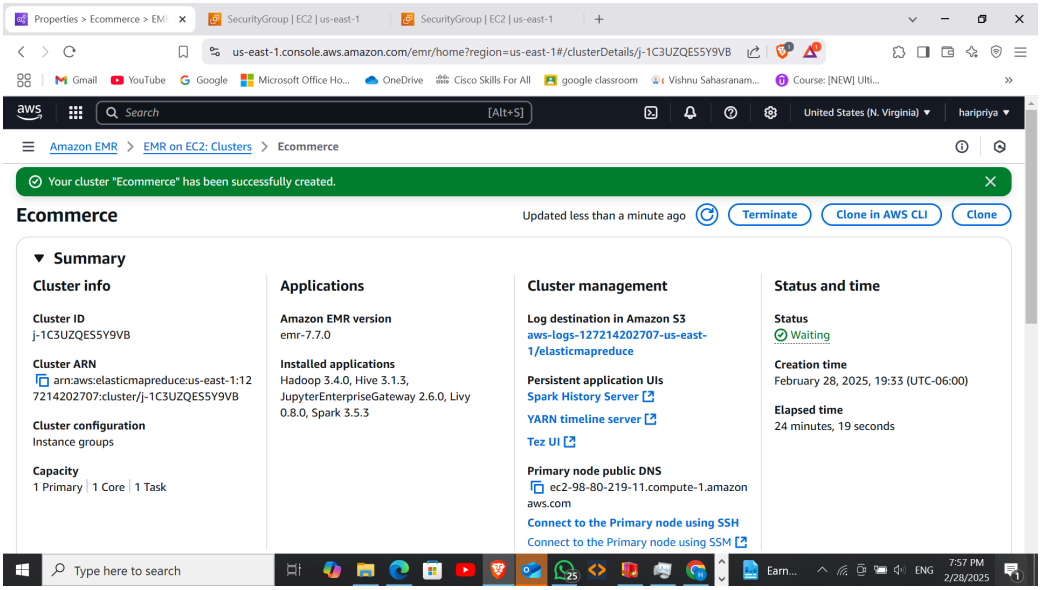
Prepared by - HariPriya Dasari Govindappa

DATA PREPROCESSING

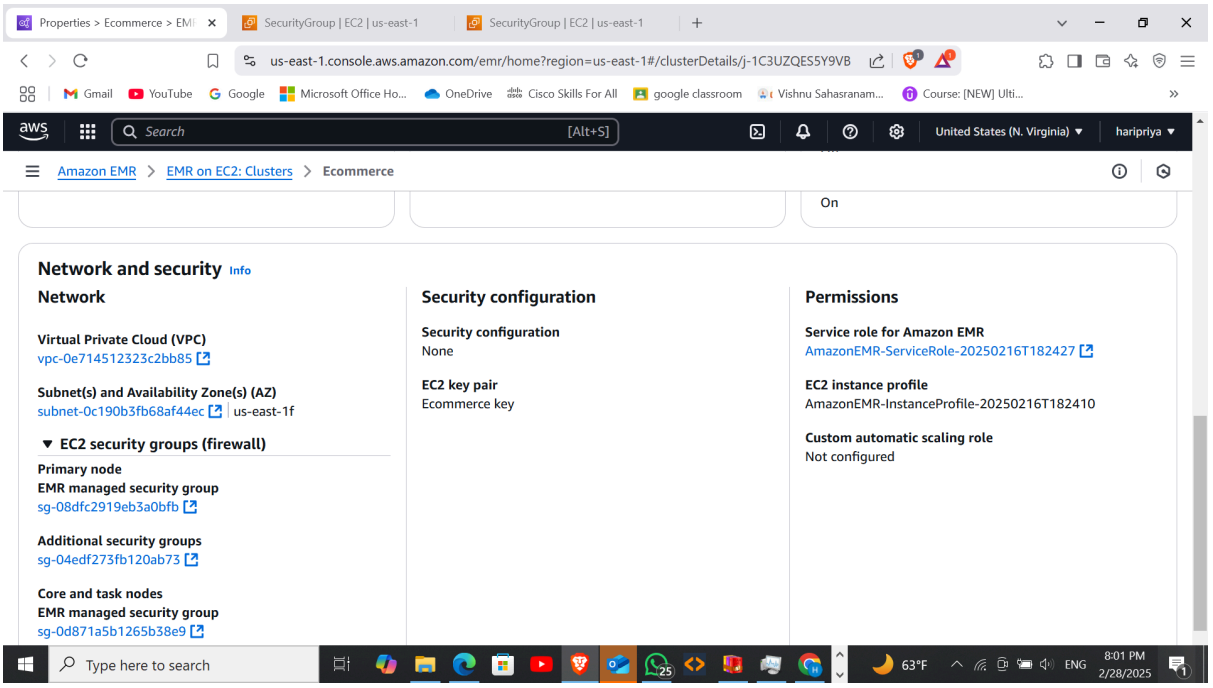
Open the Amazon S3 and then click on create a bucket then create a bucket in this i have created a bucket with name “ecommercedata - cs” once it is created i have created a folder inside this bucket with name “Inputfile” in that folder i have uploaded the the dataset OnlineRetail.csv into this input folder.



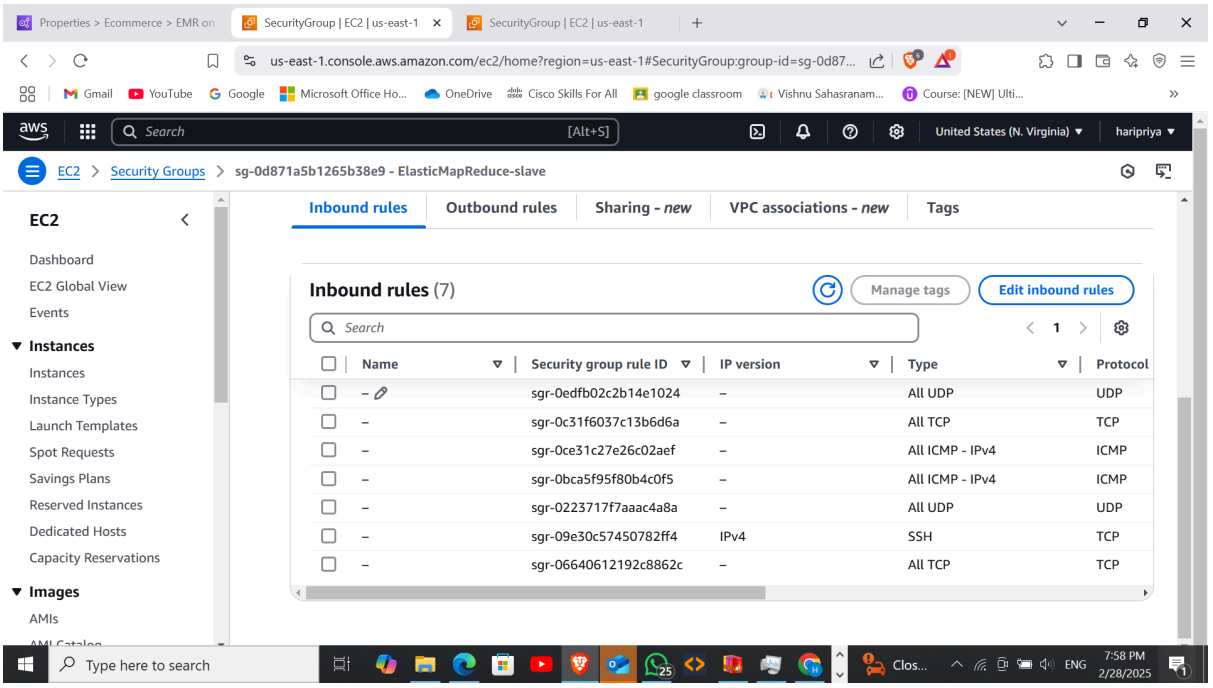
Once the Amazon S3 bucket has been created then now we will be creating the Amazon EMR cluster here. I have named my cluster as “Ecommerce” . With this cluster we will also create an EC2 key pair with which we will be able to download the “pem” file here, I have named my key pair as “Ecommerce key” . Once we create the cluster we have to wait until the status will go from “starting” to “waiting” as shown in the below image.



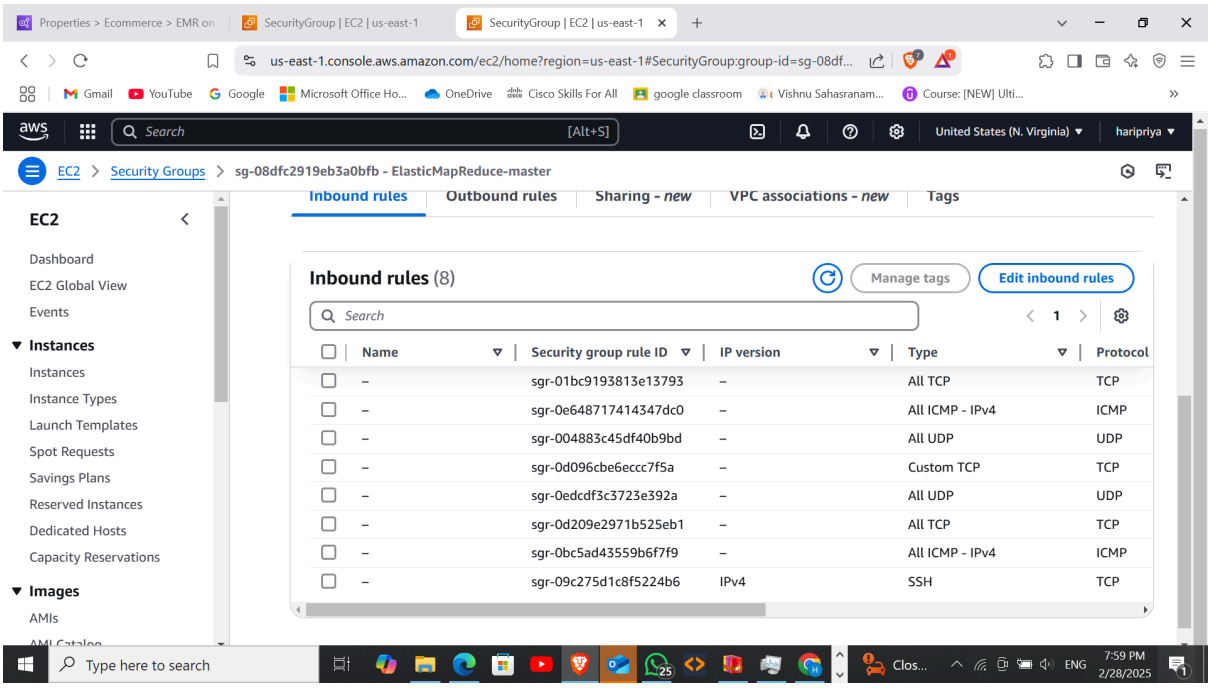
Once status turns to waiting then scroll down here we can see our selected EC2 key pair and also we can see the firewall links of primary and core and task nodes.



When clicked on the the primary node EMR managed security group the below page opens in this i have clicked on Inbound rules and new rule with SSH with IPv4 version

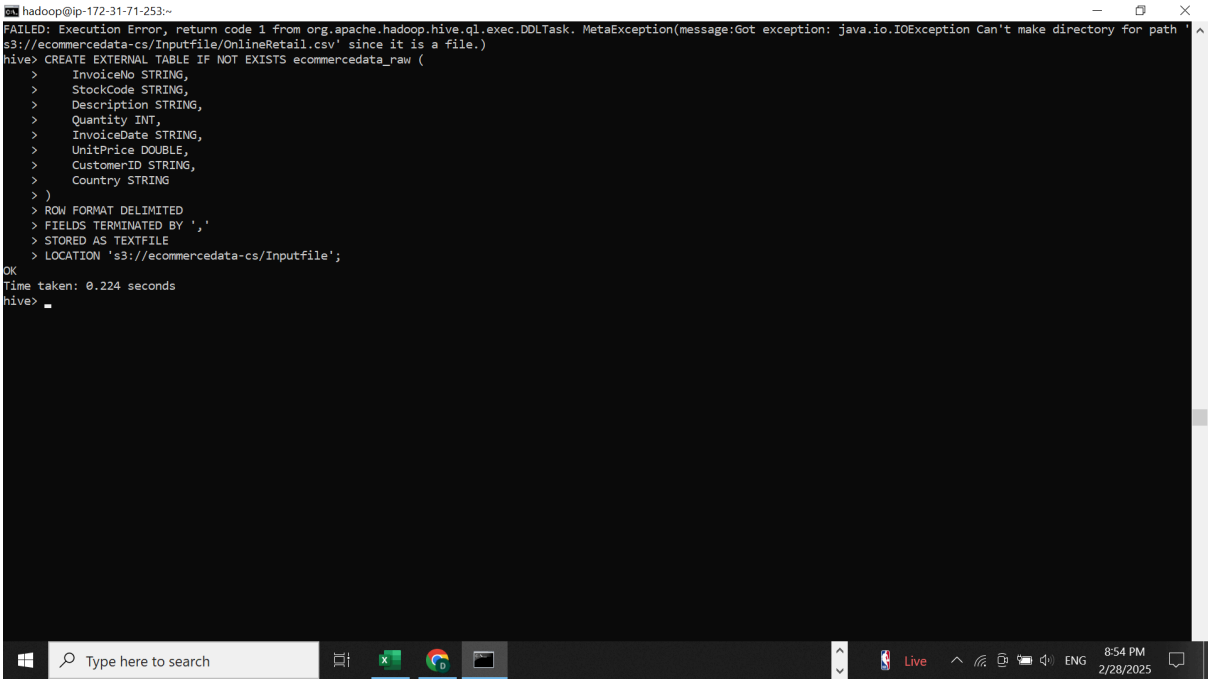


The same above process is repeated for core and task nodes EMR managed security groups.

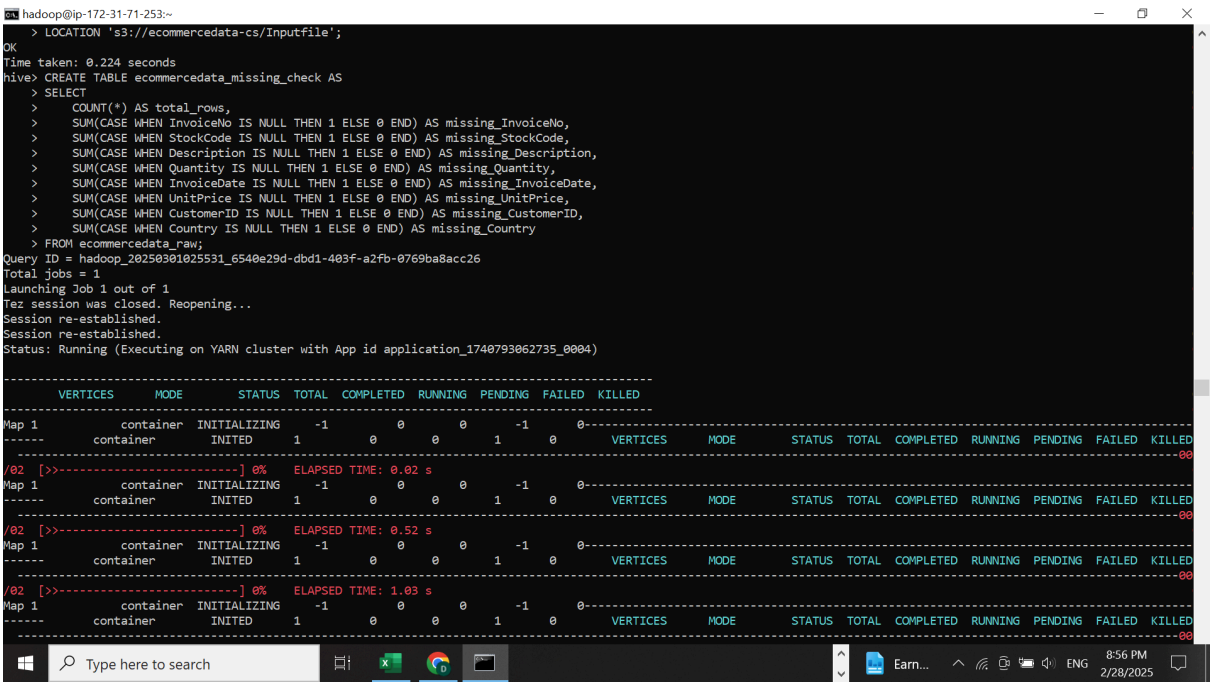


Then the for the next step i have worked on command prompt in that i used “ssh -i Ecommerce key.pem hadoop@Ec2-98-80-219-11.compute-1.amazonaws.com” and then type yes once i did this i got into EMR cluster then when i typed “hive” i was able to get into hive.

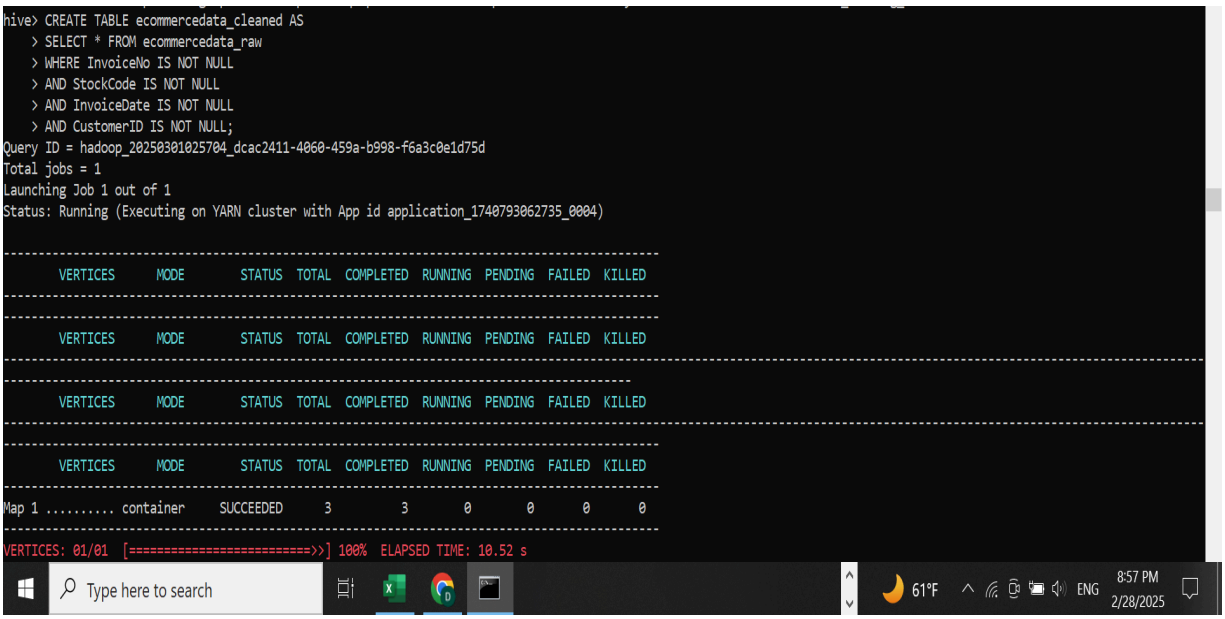
The below shown image shows the table i created with name “ecommercedata_raw” in which i stored the data from online retail dataset by using the location of the S3 bucket as shown below “s3://ecommercedata-cs/Inputfile”.



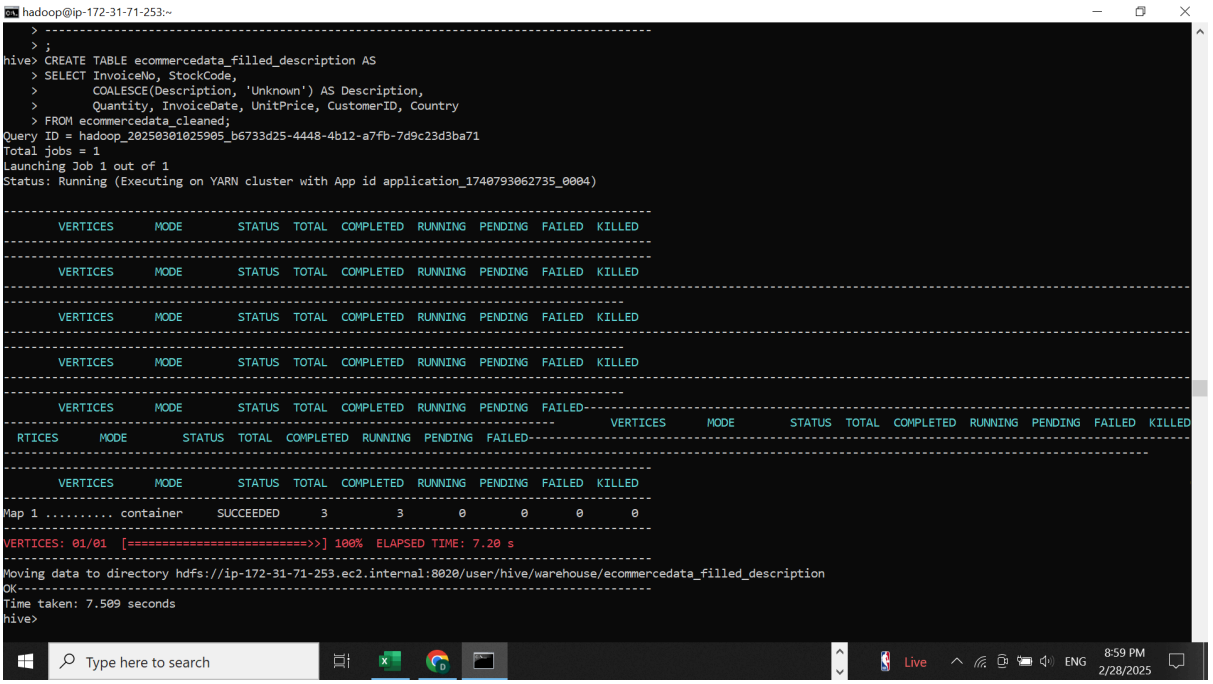
As shown in the below screenshot SUM is used to find the sum of total missing values in each column.



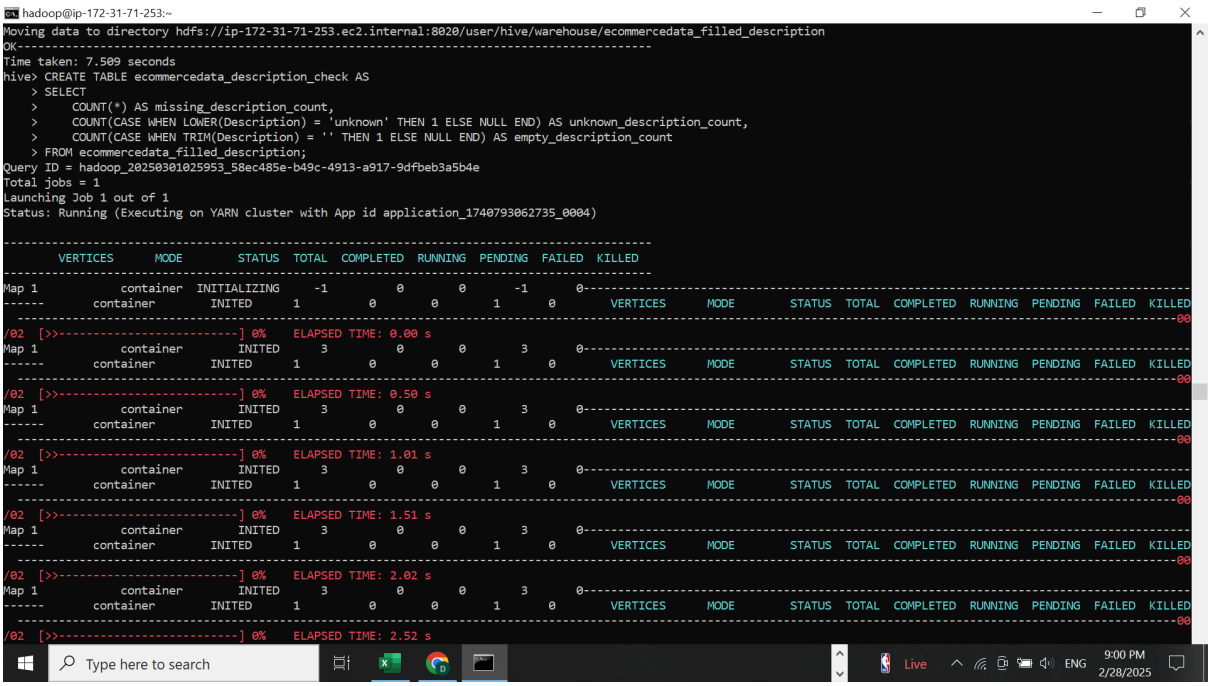
For cleaned data a new table has been created with the name ecommercedata_cleaned. From the previous step result it was found that InvoiceNo, StockCode, InvooiceDate, CustomerId and Description were the columns with the missing values. By giving column names with NOT NULL we will be able to retrieve the rows that are not null.

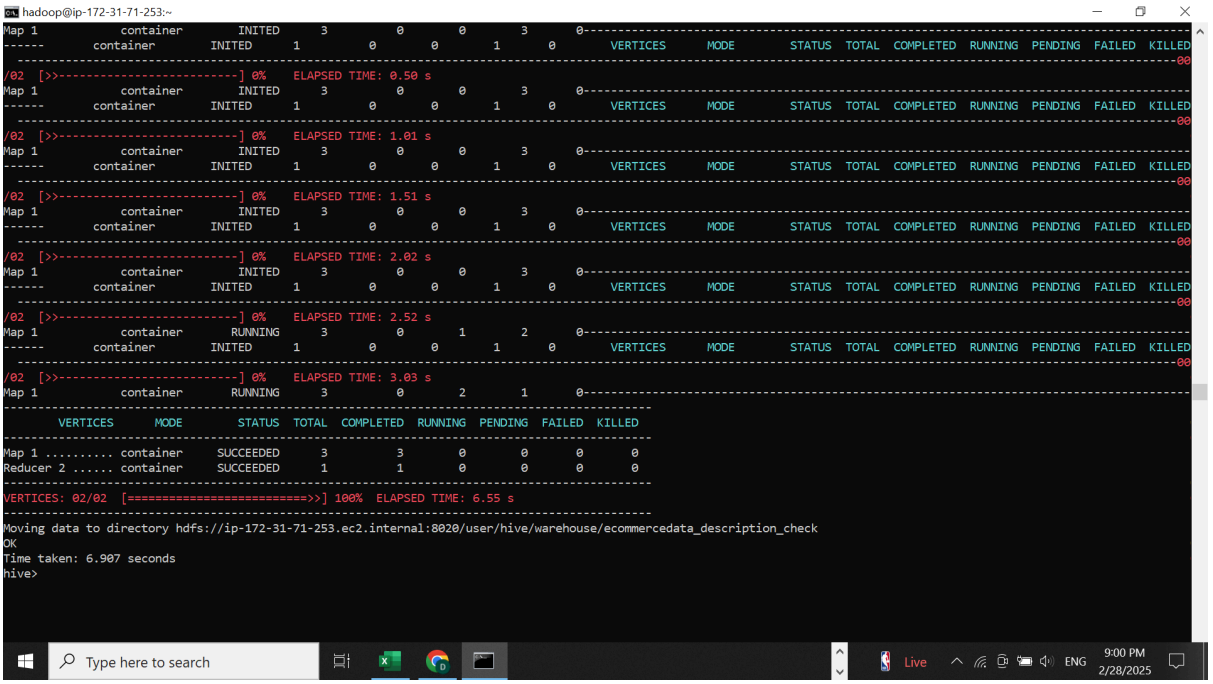


In this step COALESCE is being used to replace null values with “Unknown” in the description.

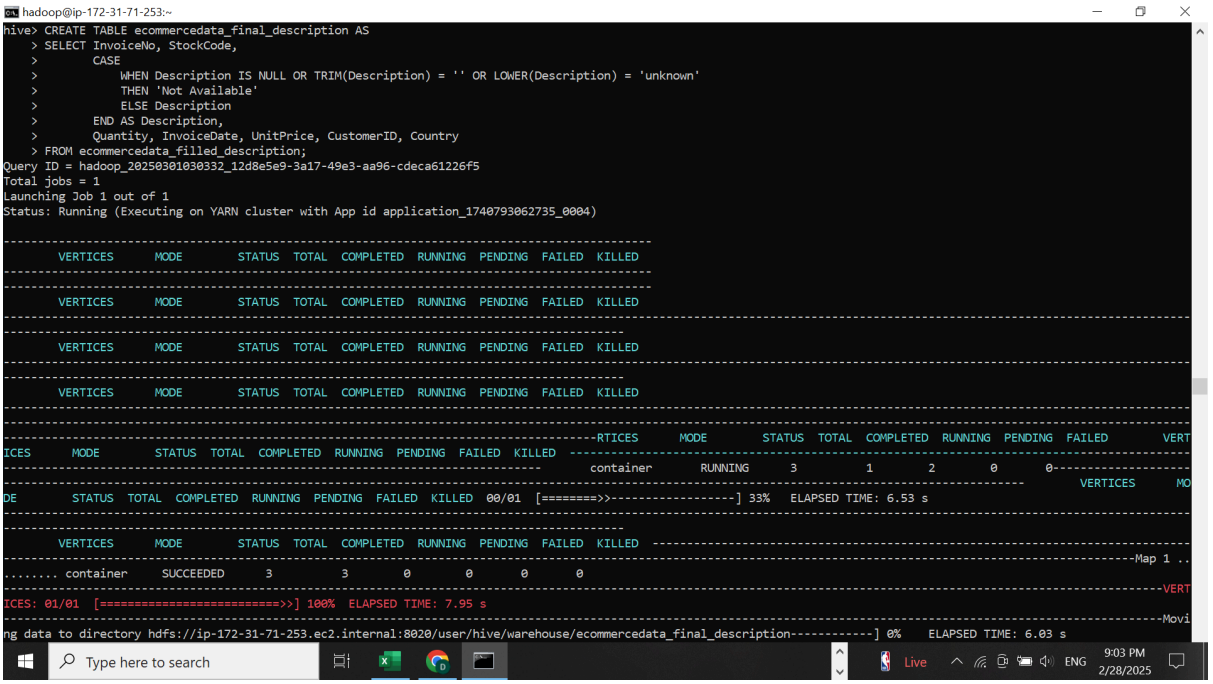


The below query is written to find the count of unknown values or null or missing values in the description column.

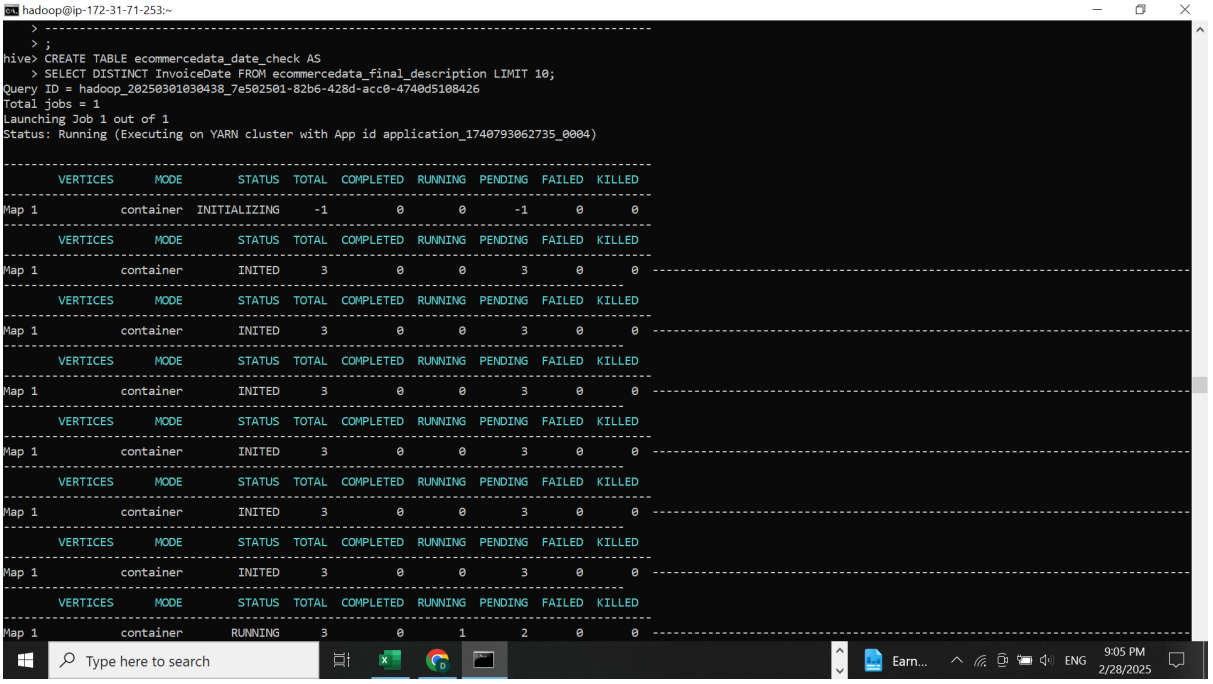




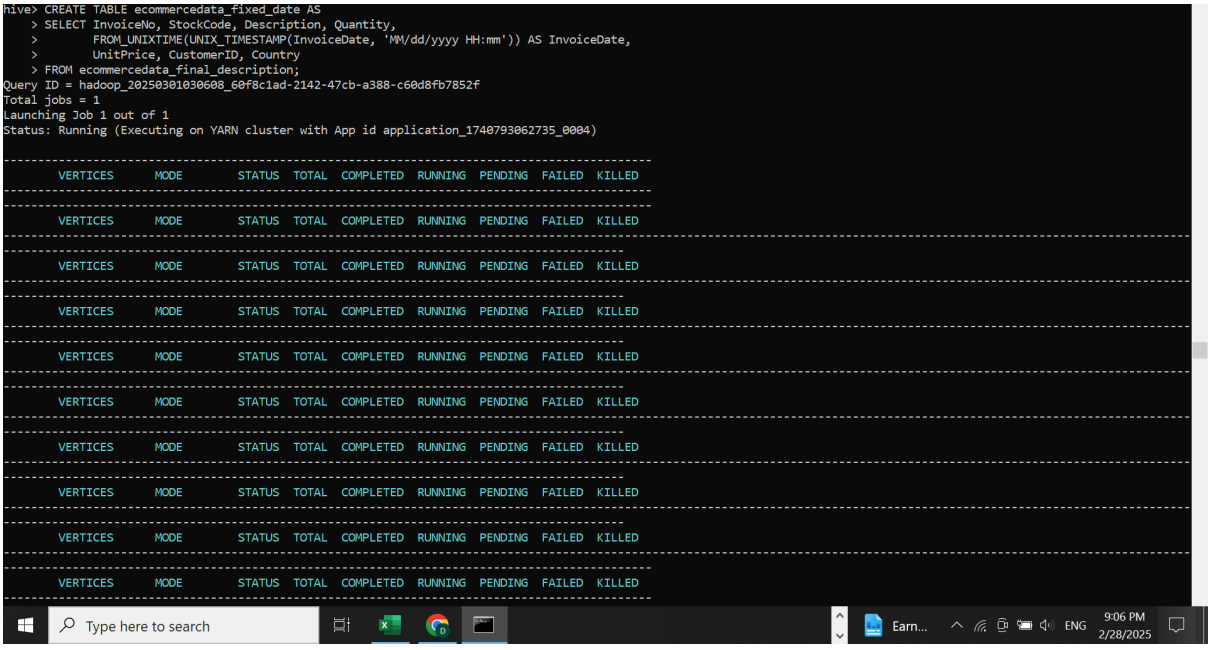
In this step the data cleaning process is applied for description column by replacing the null, missing and the unknown values with “not available”.



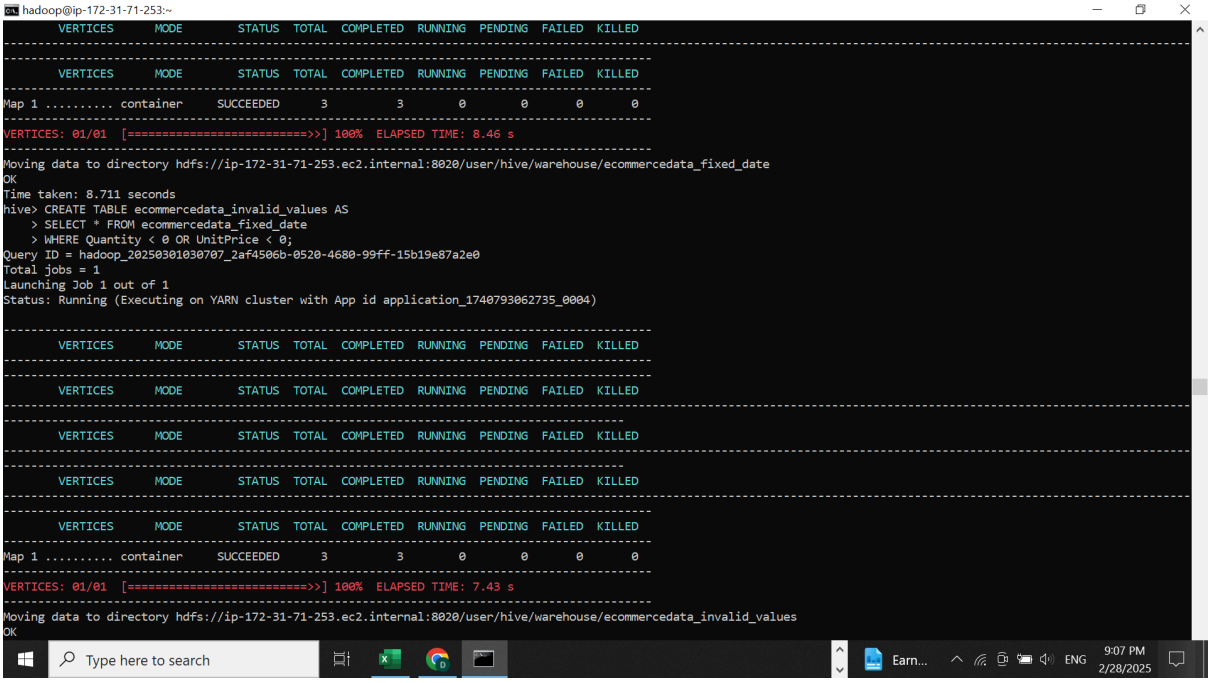
The below code is written to check the current date format.



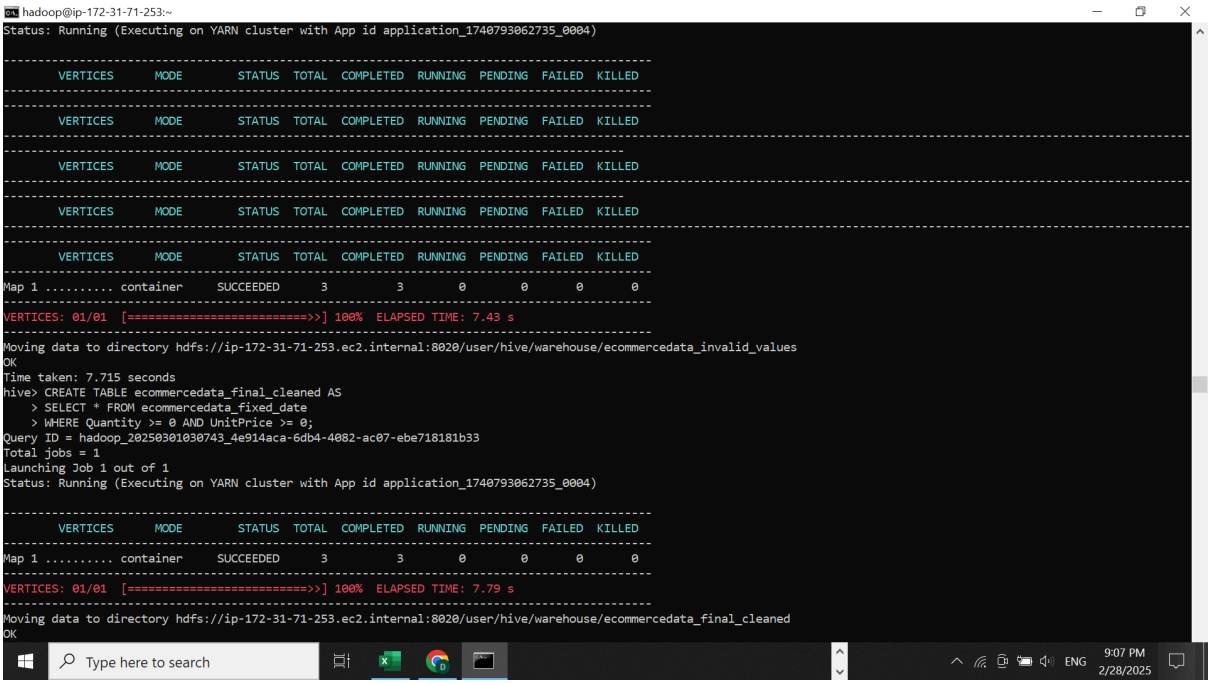
For reformatting the date to “YYYY/MM/DD HH:MM” the below code is written.



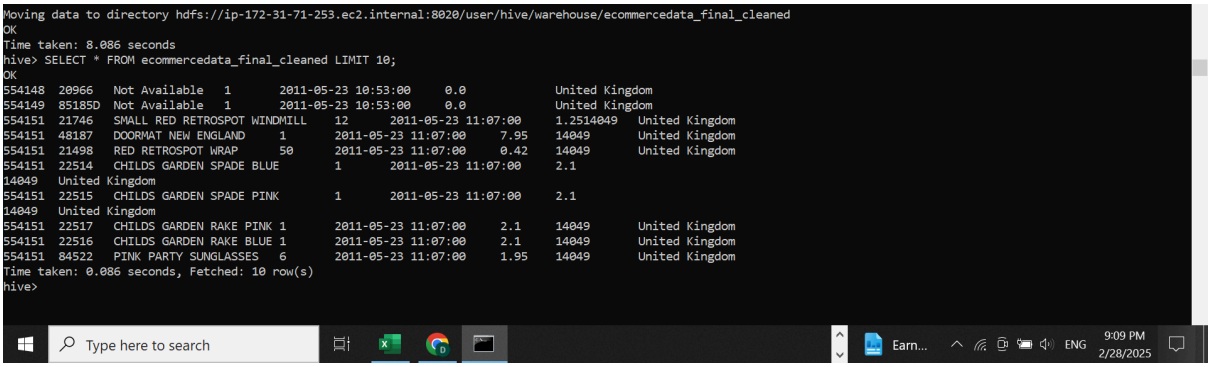
In this step Quantity <0 and UnitPrice <0 is used to check for the values available on the Quantity and UnitPrice columns.

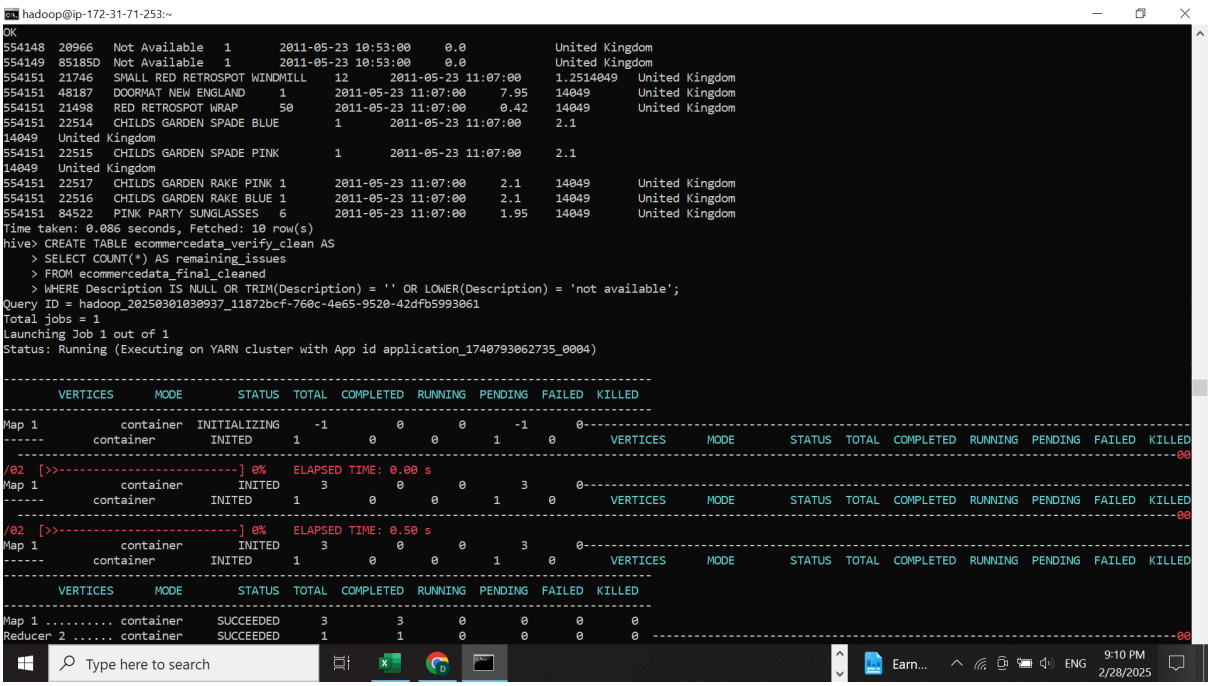


Now that from the above step as we found negative values, to remove those invalid rows with invalid numbers we are using the below query.

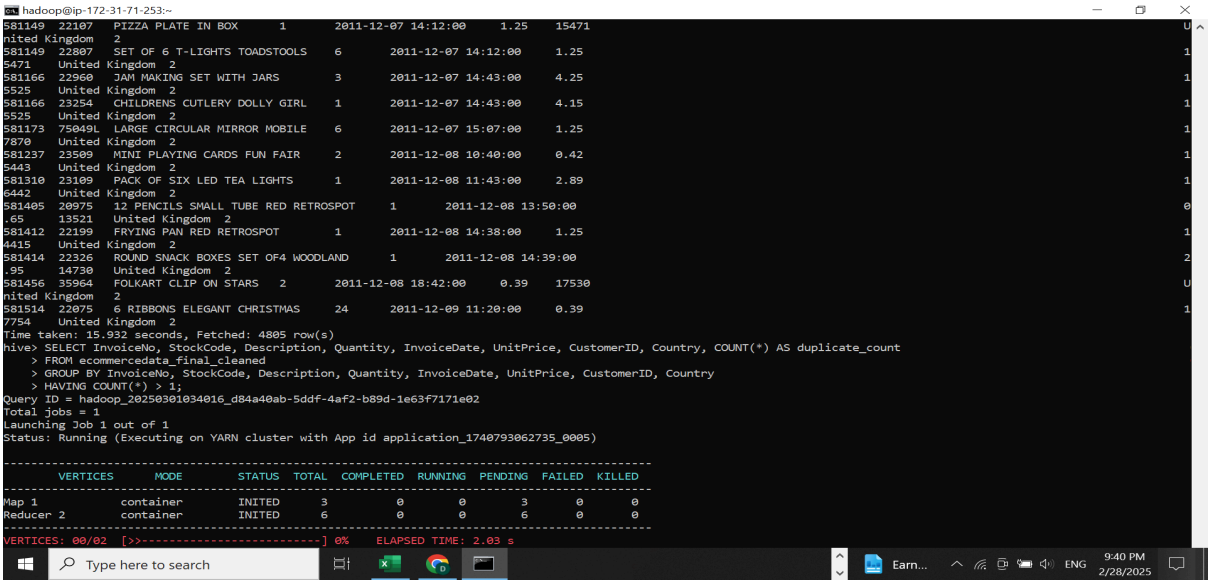


A query to display the cleaned data for first 10 rows

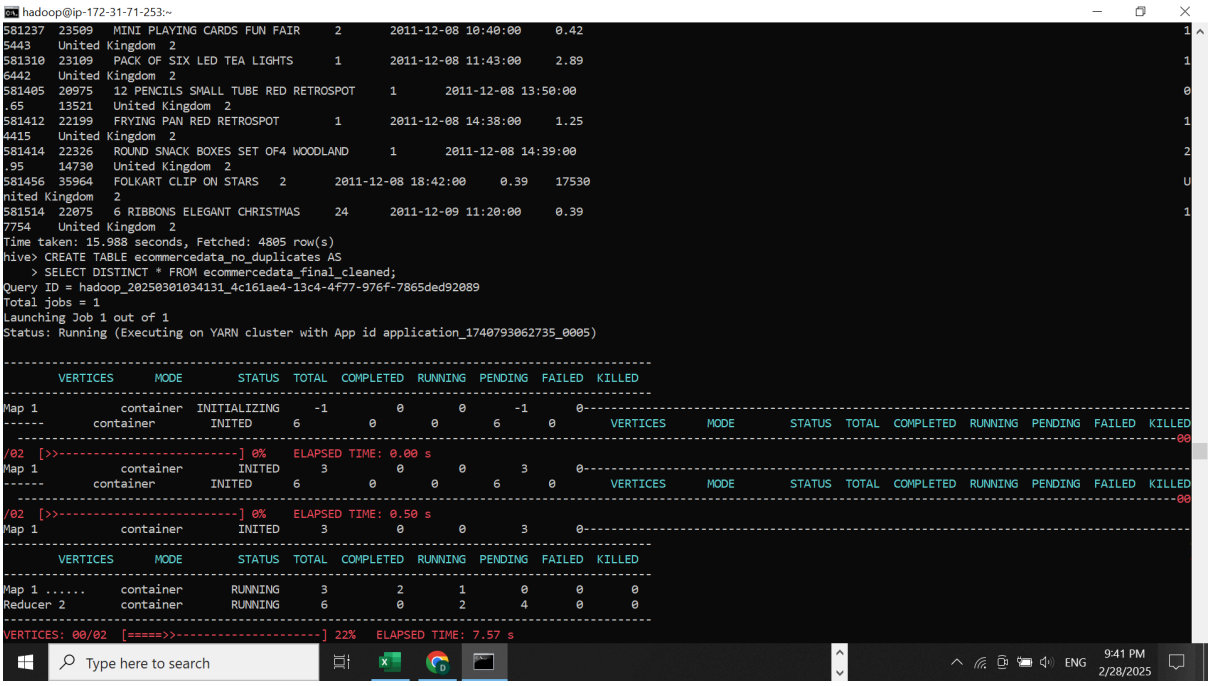




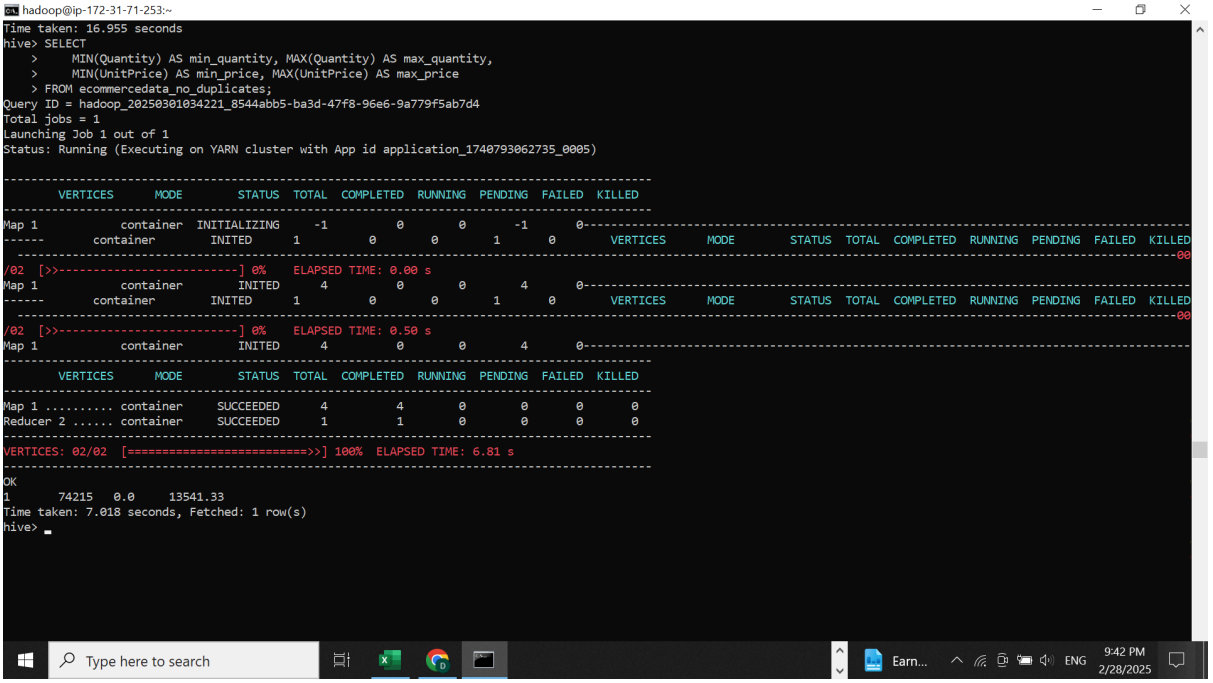
A query to check for duplicate Records.



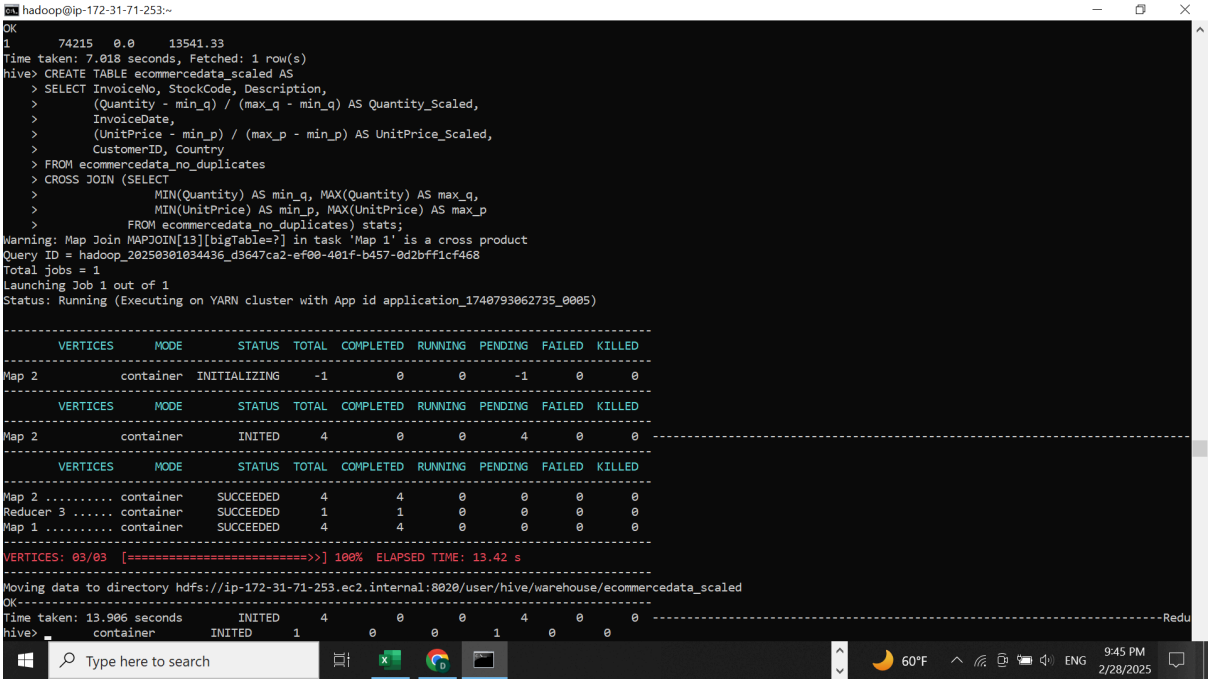
In the below query we will be able to remove all the duplicate records.

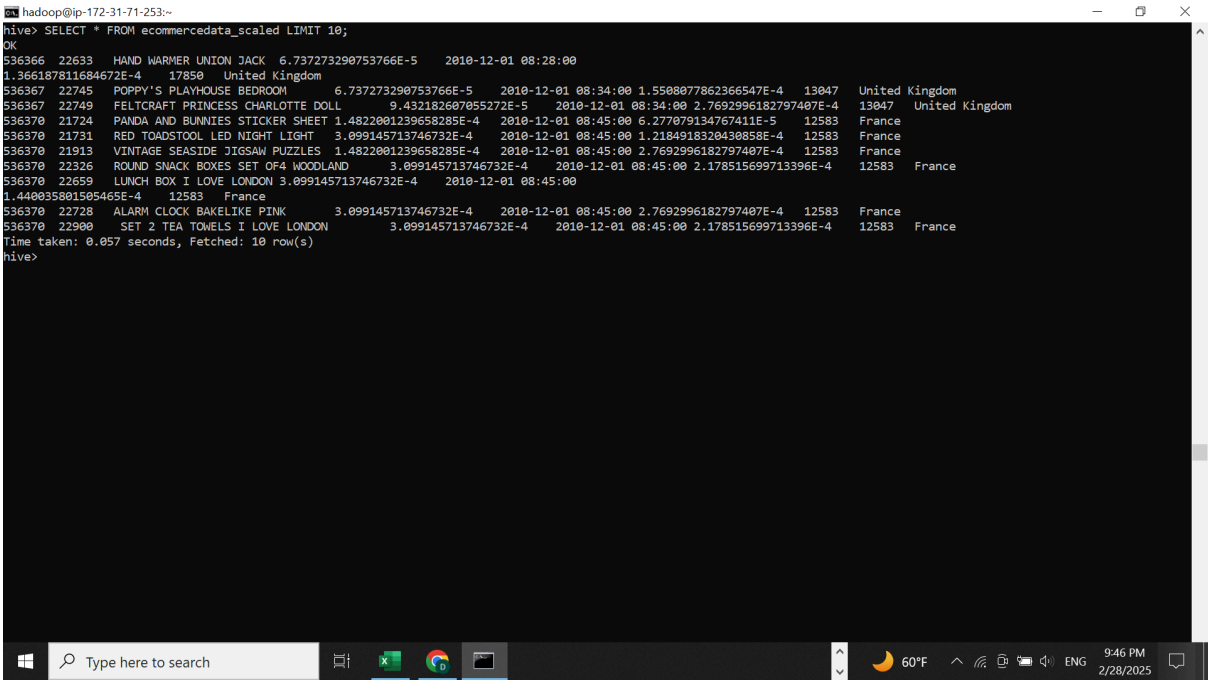


Performing normalization on Numerical columns - Quantity, UnitPrice.

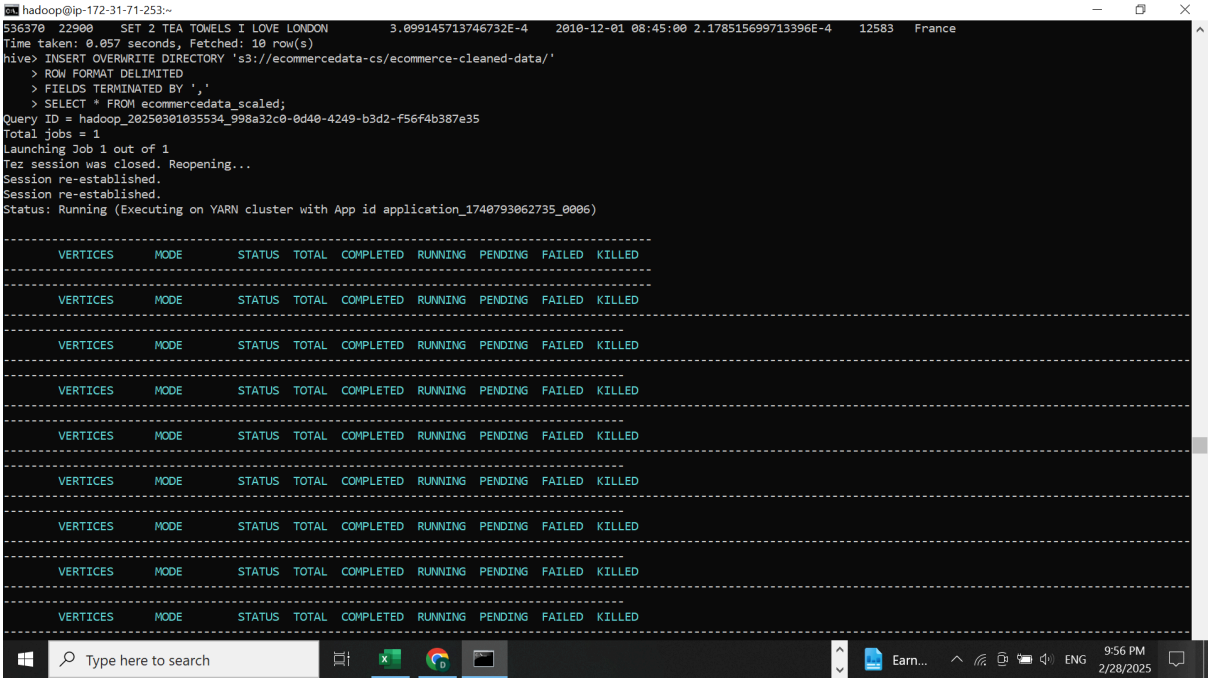


Apply Scaling with The formula : X′= (X - Xmin)/(Xmax - Xmin)

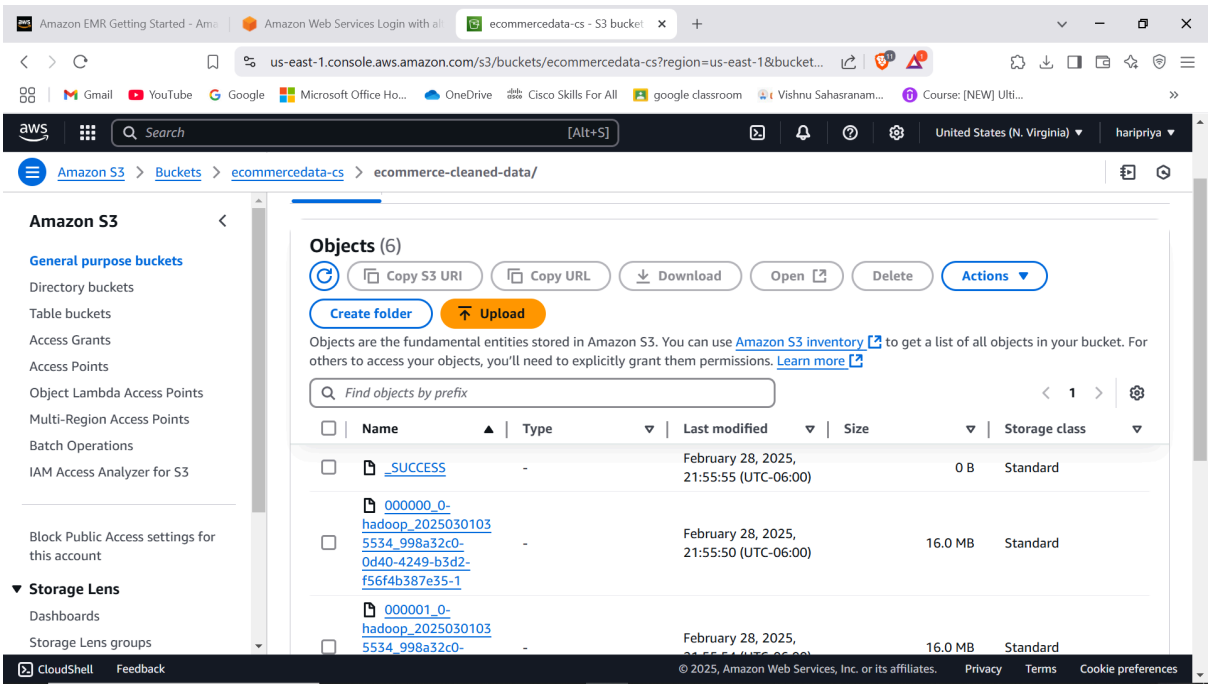


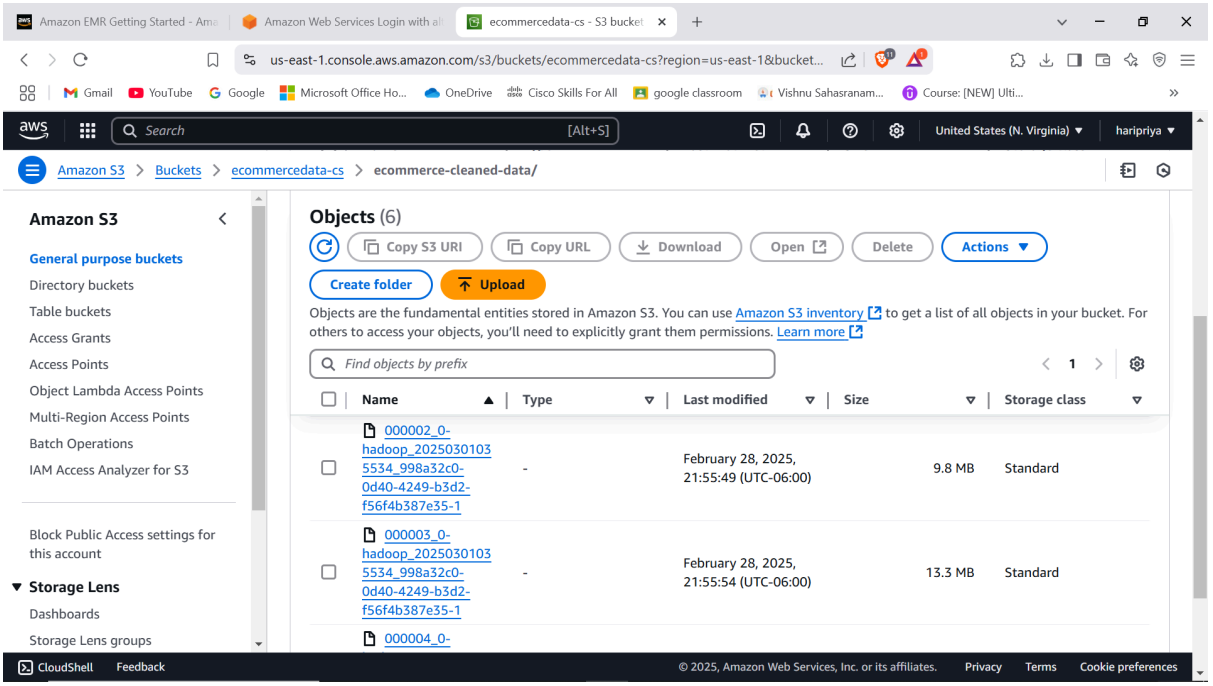


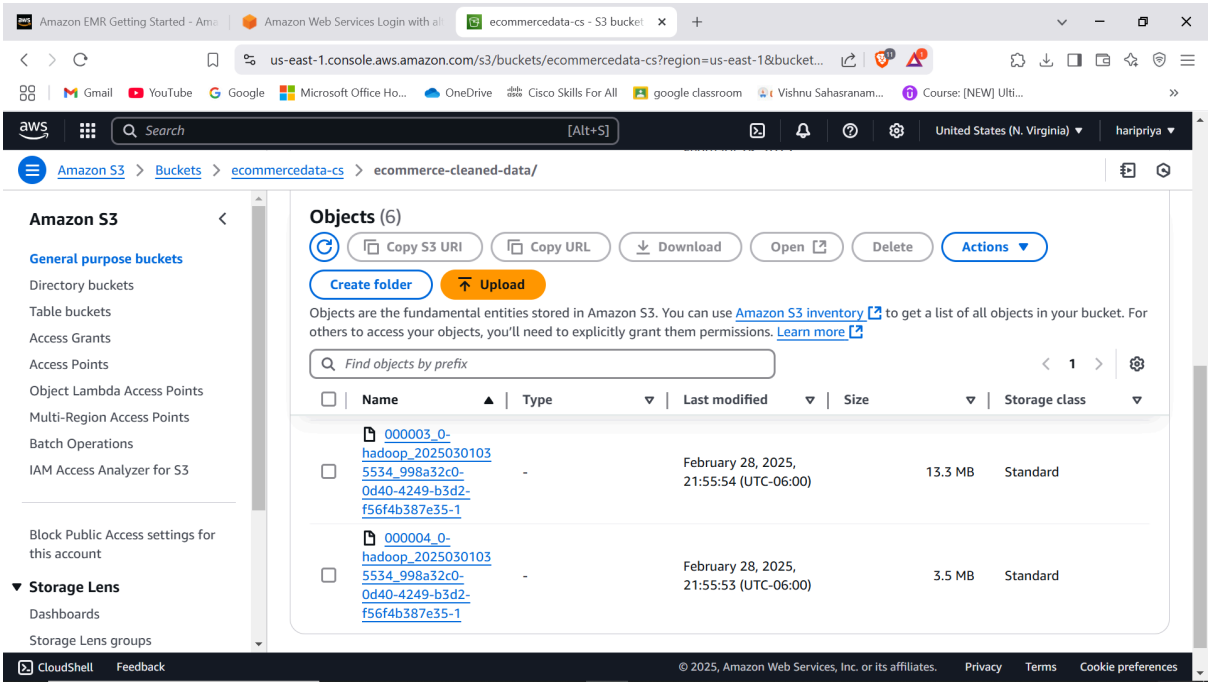
The below code of OVERWRITE DIRECTORY is used for storing the cleaned data back into the S3 bucket.



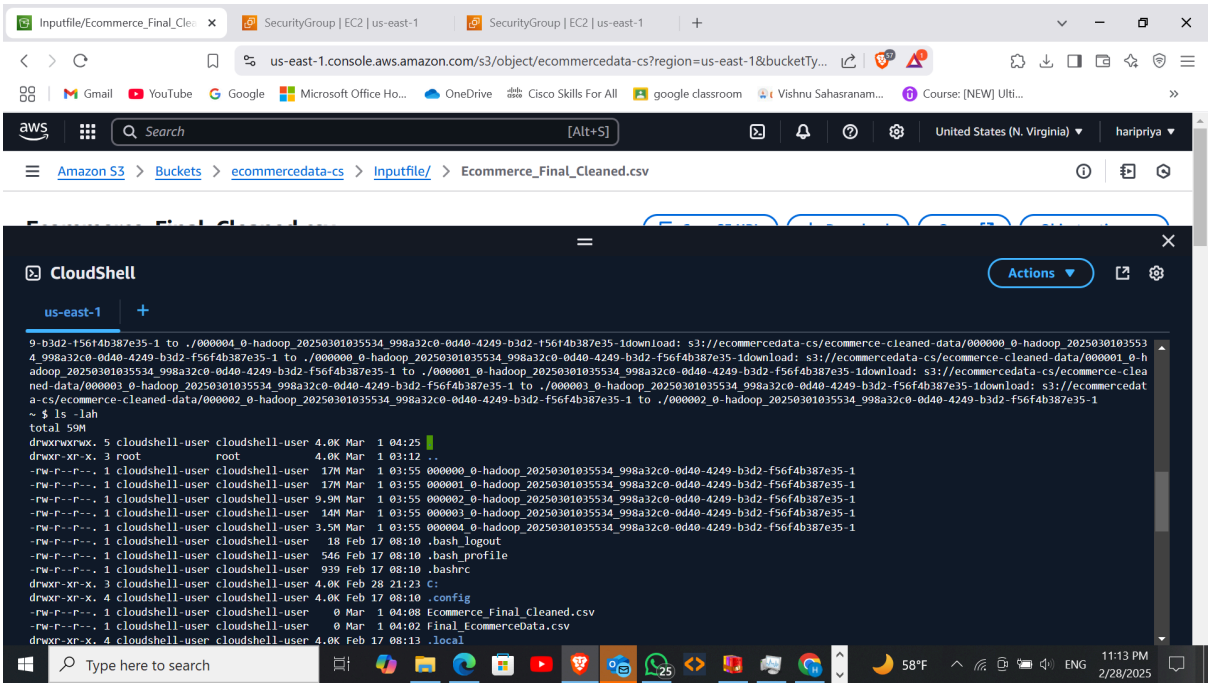
Below are the datafiles that were added back to S3 Bucket.

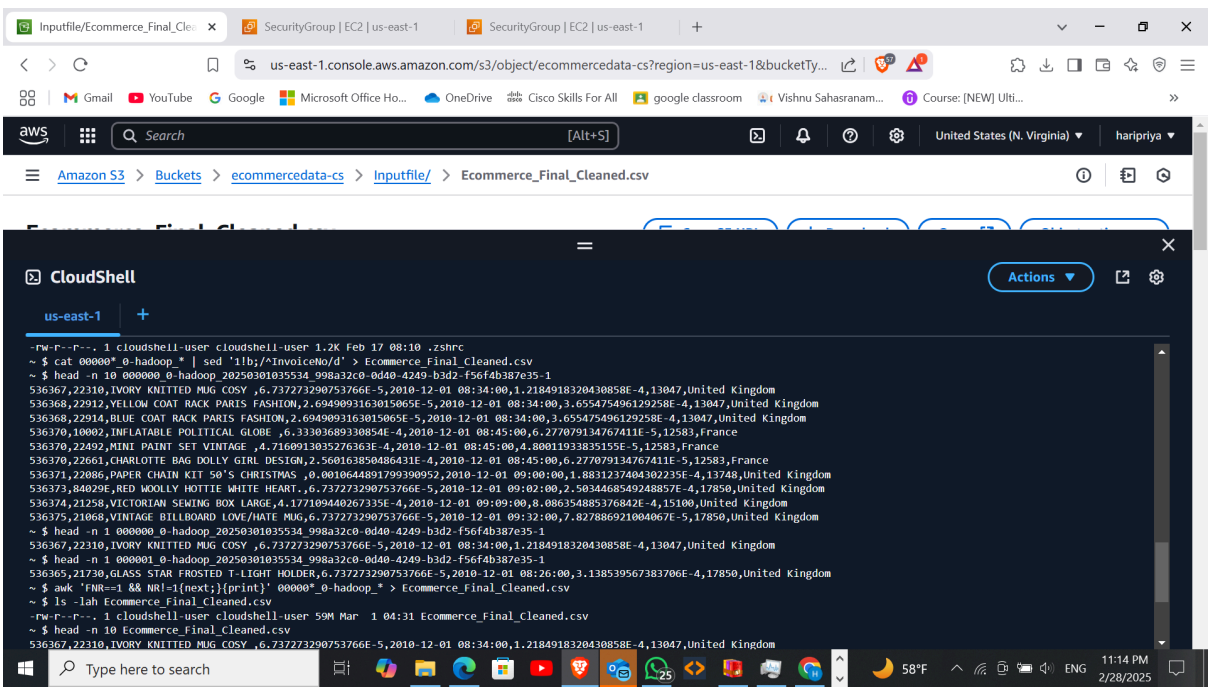


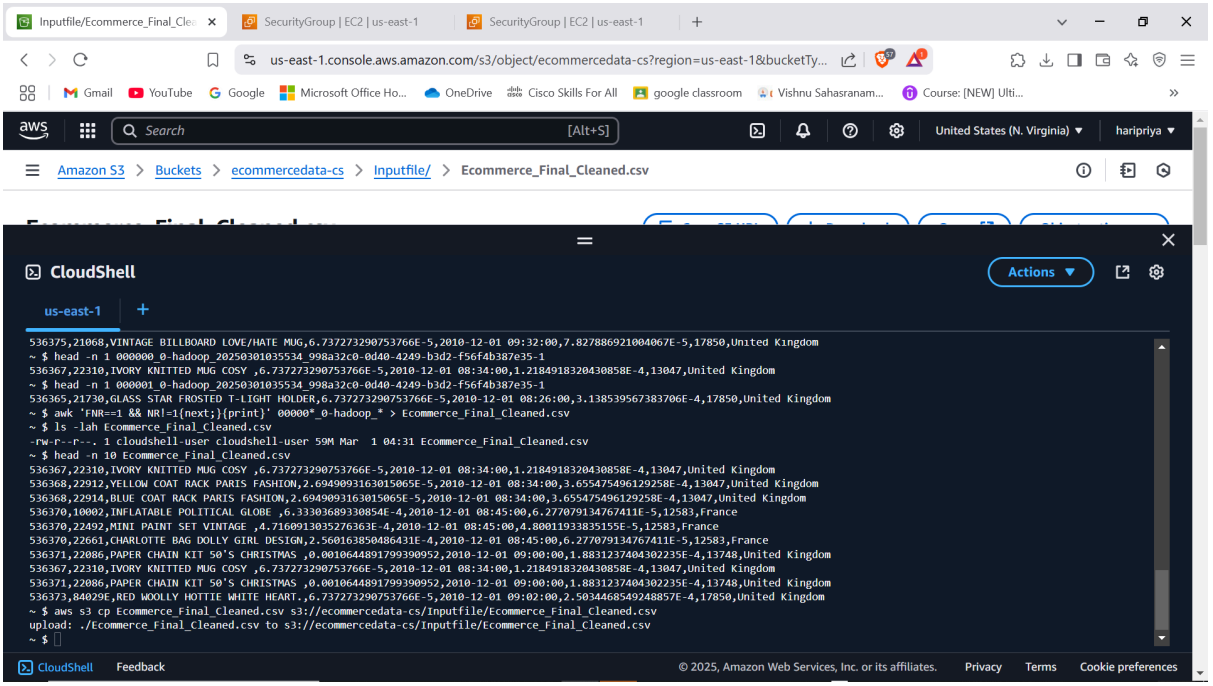

Amazon EMR Getting Started - Ama
Amazon Web Services Login with al
ecommercedata-cs - S3 bucket
us-east-1.console.aws.amazon.com/s3/buckets/ecommercedata-cs?region=us-east-1&bucket...
Gmail
YouTube
Google
Microsoft Office Ho...
OneDrive
Cisco Skills For All
google classroom
Vishnu Sahasranam...
Course: [NEW] Ulti...
Search
[Alt+S]
United States (N. Virginia)
haripriya
Amazon S3 > Buckets > ecommercedata-cs > ecommerce-cleaned-data/
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (6)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
Storage class
000002_0-hadoop_2025030103 5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
February 28, 2025, 21:55:49 (UTC-06:00)
9.8 MB
Standard
000003_0-hadoop_2025030103 5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
February 28, 2025, 21:55:54 (UTC-06:00)
13.3 MB
Standard
000004_0-
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon EMR Getting Started - Ama
Amazon Web Services Login with al
ecommercedata-cs - S3 bucket
us-east-1.console.aws.amazon.com/s3/buckets/ecommercedata-cs?region=us-east-1&bucket...
Gmail
YouTube
Google
Microsoft Office Ho...
OneDrive
Cisco Skills For All
google classroom
Vishnu Sahasranam...
Course: [NEW] Ulti...
Search
[Alt+S]
United States (N. Virginia)
haripriya
Amazon S3 > Buckets > ecommercedata-cs > ecommerce-cleaned-data/
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (6)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
Storage class
000003_0-hadoop_2025030103 5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
February 28, 2025, 21:55:54 (UTC-06:00)
13.3 MB
Standard
000004_0-hadoop_2025030103 5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
February 28, 2025, 21:55:53 (UTC-06:00)
3.5 MB
Standard
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

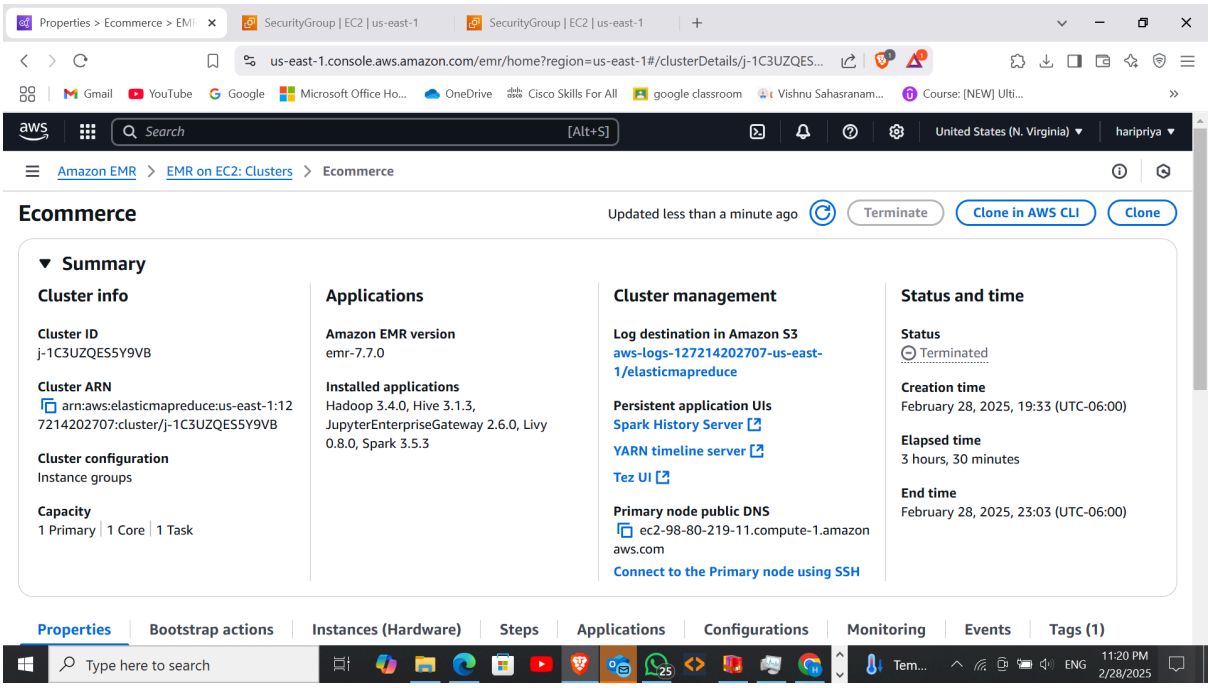
As the data set is huge the output data was stored in multiple files as shown in the above screenshots, to merge them the below process in done in the AWS CloudShell





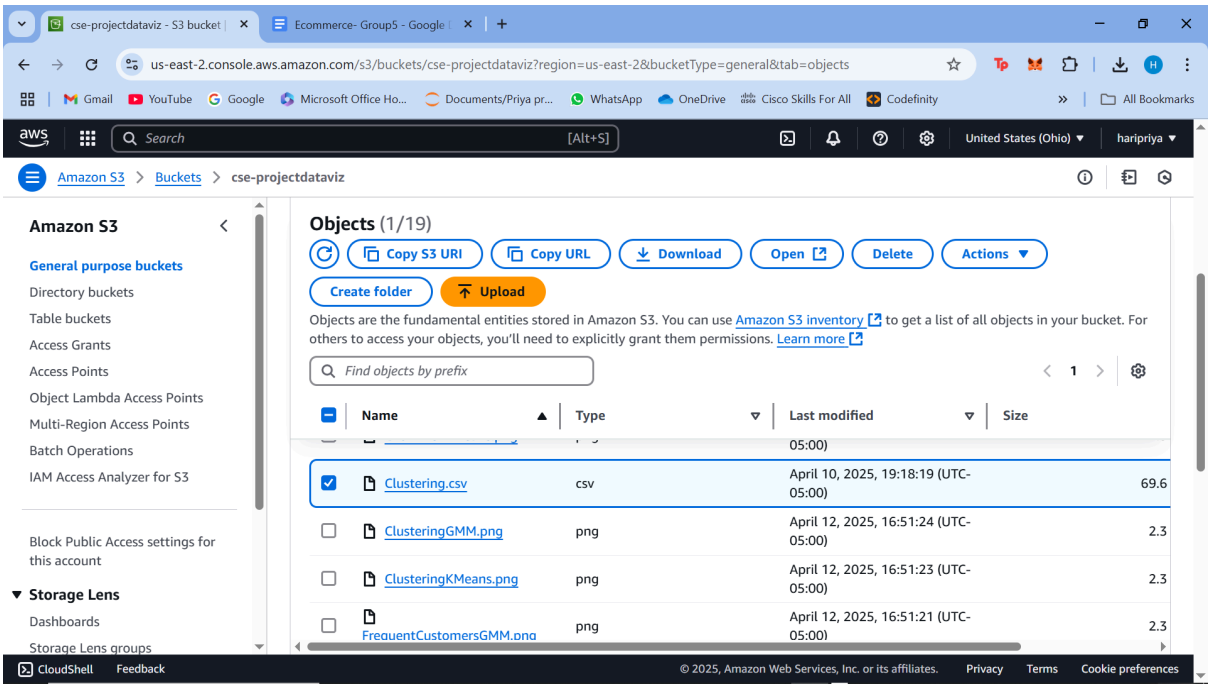


Once the whole process is done terminate the cluster.



DATA VISUALIZATION

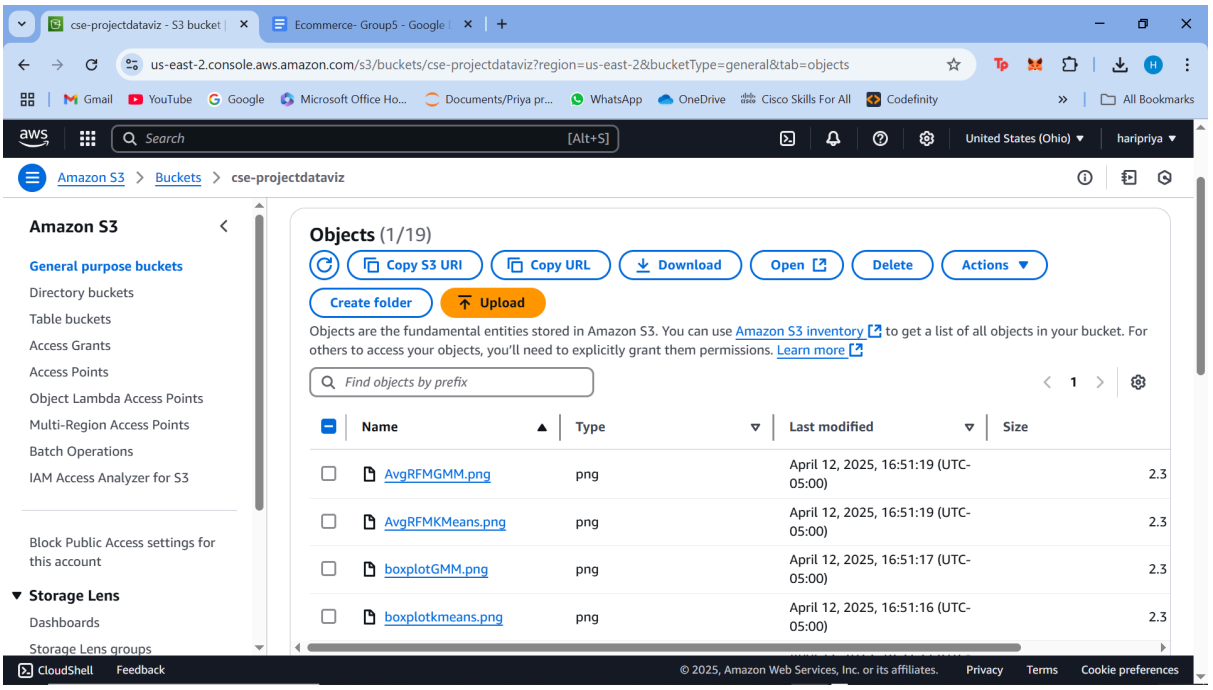
S3 bucket was created and i added the csv file which was given by my teammate after application of clustering Algorithms K-means and GMM

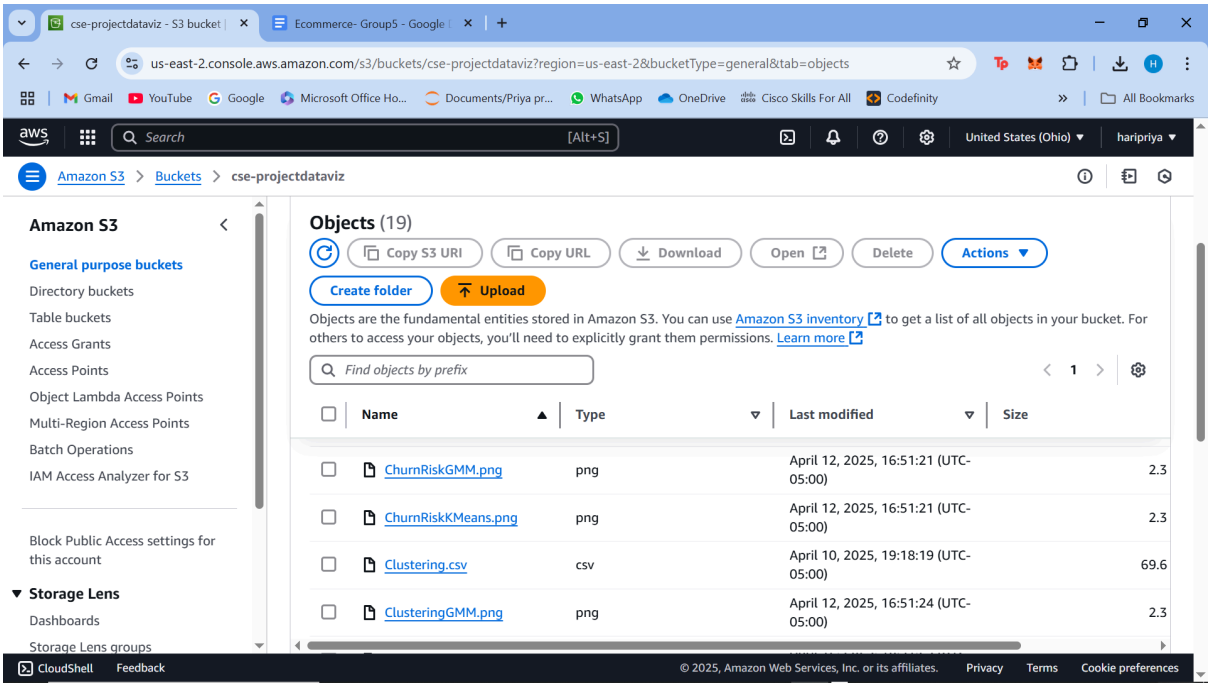


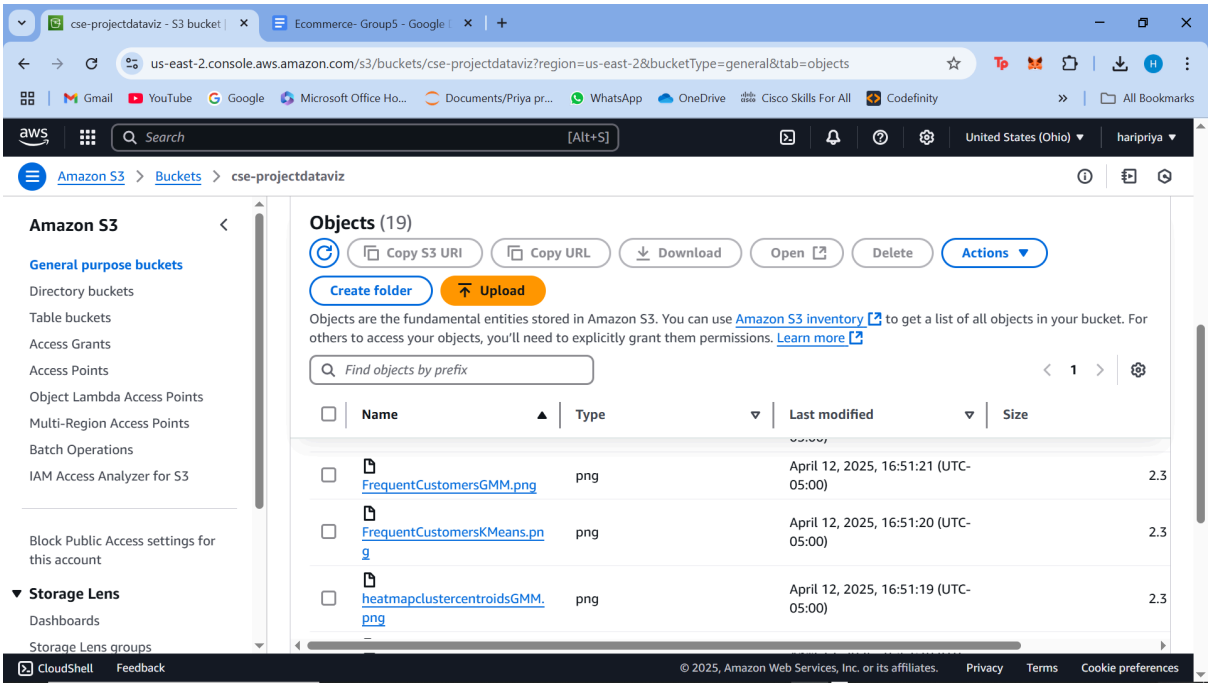
USED

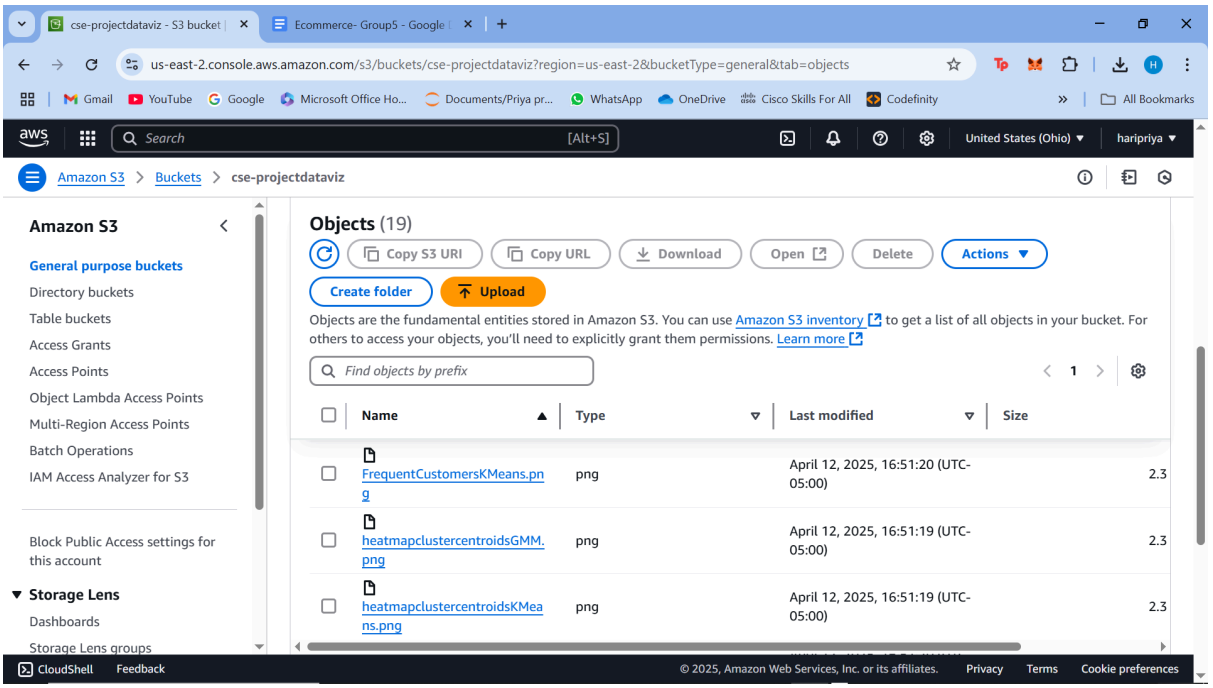
```
# Save and upload buffer = BytesIO()
plt.savefig(buffer, format='png')
buffer.seek(0)
key = 'ClusteringKMeans.png'
s3.upload_fileobj(buffer, bucket_name, key)
plt.close()
print(f"Saved Clustering of K-Means to S3: s3://{bucket_name}/{key}")
```

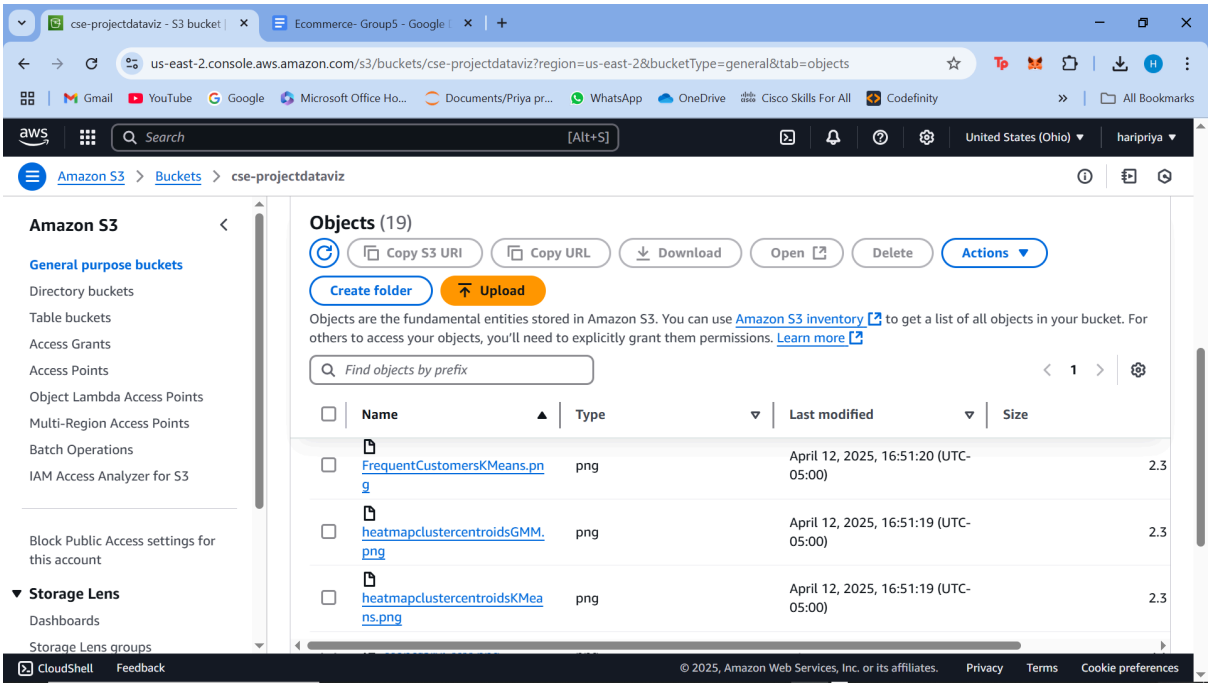
To upload the created visualizations back to S3 bucket

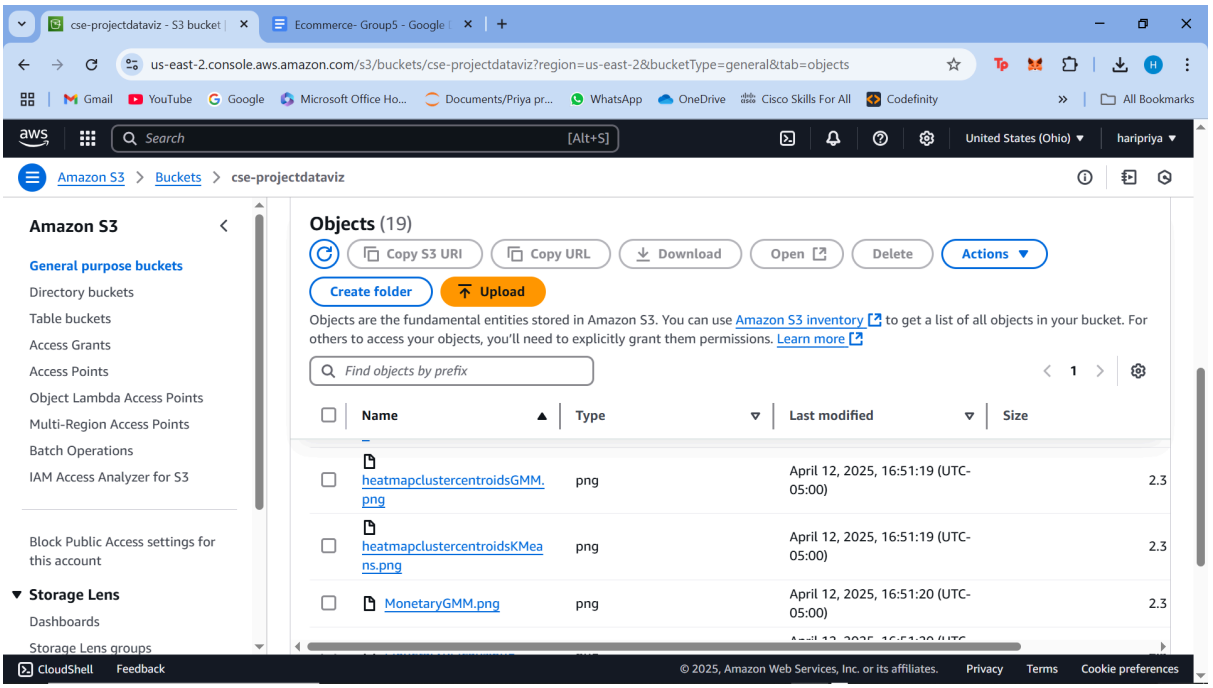


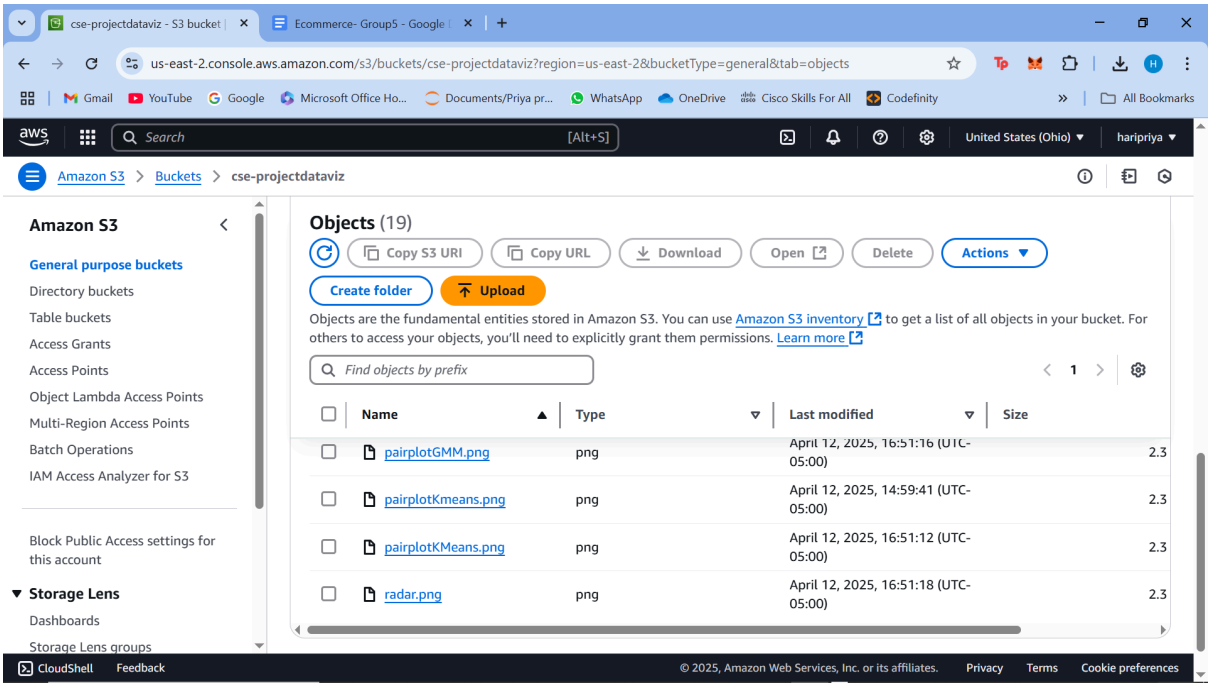



cse-projectdataviz - S3 bucket
Ecommerce- Group5 - Google
us-east-2.console.aws.amazon.com/s3/buckets/cse-projectdataviz?region=us-east-2&bucketType=general&tab=objects
Gmail
YouTube
Google
Microsoft Office Ho...
Documents/Priya pr...
WhatsApp
OneDrive
Cisco Skills For All
Codefinity
All Bookmarks
aws
Search
[Alt+S]
United States (Ohio)
haripriya
Amazon S3
Buckets
cse-projectdataviz
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (19)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
FrequentCustomersGMM.png
png
April 12, 2025, 16:51:21 (UTC-05:00)
2.3
FrequentCustomersKMeans.png
png
April 12, 2025, 16:51:20 (UTC-05:00)
2.3
heatmapclustercentroidsGMM.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


cse-projectdataviz - S3 bucket
Ecommerce- Group5 - Google
us-east-2.console.aws.amazon.com/s3/buckets/cse-projectdataviz?region=us-east-2&bucketType=general&tab=objects
Gmail
YouTube
Google
Microsoft Office Ho...
Documents/Priya pr...
WhatsApp
OneDrive
Cisco Skills For All
Codefinity
All Bookmarks
aws
Search
[Alt+S]
United States (Ohio)
haripriya
Amazon S3
Buckets
cse-projectdataviz
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (19)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
FrequentCustomersKMeans.png
png
April 12, 2025, 16:51:20 (UTC-05:00)
2.3
heatmapclustercentroidsGMM.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
heatmapclustercentroidsKMeans.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


cse-projectdataviz - S3 bucket
Ecommerce- Group5 - Google
us-east-2.console.aws.amazon.com/s3/buckets/cse-projectdataviz?region=us-east-2&bucketType=general&tab=objects
Search [Alt+S]
United States (Ohio)
haripriya
Amazon S3 > Buckets > cse-projectdataviz
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (19)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
FrequentCustomersKMeans.png
png
April 12, 2025, 16:51:20 (UTC-05:00)
2.3
heatmapclustercentroidsGMM.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
heatmapclustercentroidsKMeans.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


cse-projectdataviz - S3 bucket
Ecommerce- Group5 - Google
us-east-2.console.aws.amazon.com/s3/buckets/cse-projectdataviz?region=us-east-2&bucketType=general&tab=objects
Search [Alt+S]
United States (Ohio)
haripriya
Amazon S3 > Buckets > cse-projectdataviz
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (19)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
heatmapclustercentroidsGMM.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
heatmapclustercentroidsKMeans.png
png
April 12, 2025, 16:51:19 (UTC-05:00)
2.3
MonetaryGMM.png
png
April 12, 2025, 16:51:20 (UTC-05:00)
2.3
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


cse-projectdataviz - S3 bucket
Ecommerce- Group5 - Google
us-east-2.console.aws.amazon.com/s3/buckets/cse-projectdataviz?region=us-east-2&bucketType=general&tab=objects
Gmail
YouTube
Google
Microsoft Office Ho...
Documents/Priya pr...
WhatsApp
OneDrive
Cisco Skills For All
Codefinity
All Bookmarks
Search
[Alt+S]
United States (Ohio)
haripriya
Amazon S3
Buckets
cse-projectdataviz
Amazon S3
General purpose buckets
Directory buckets
Table buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Objects (19)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
pairplotGMM.png
png
April 12, 2025, 16:51:16 (UTC-05:00)
2.3
pairplotKmeans.png
png
April 12, 2025, 14:59:41 (UTC-05:00)
2.3
pairplotKMeans.png
png
April 12, 2025, 16:51:12 (UTC-05:00)
2.3
radar.png
png
April 12, 2025, 16:51:18 (UTC-05:00)
2.3
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences