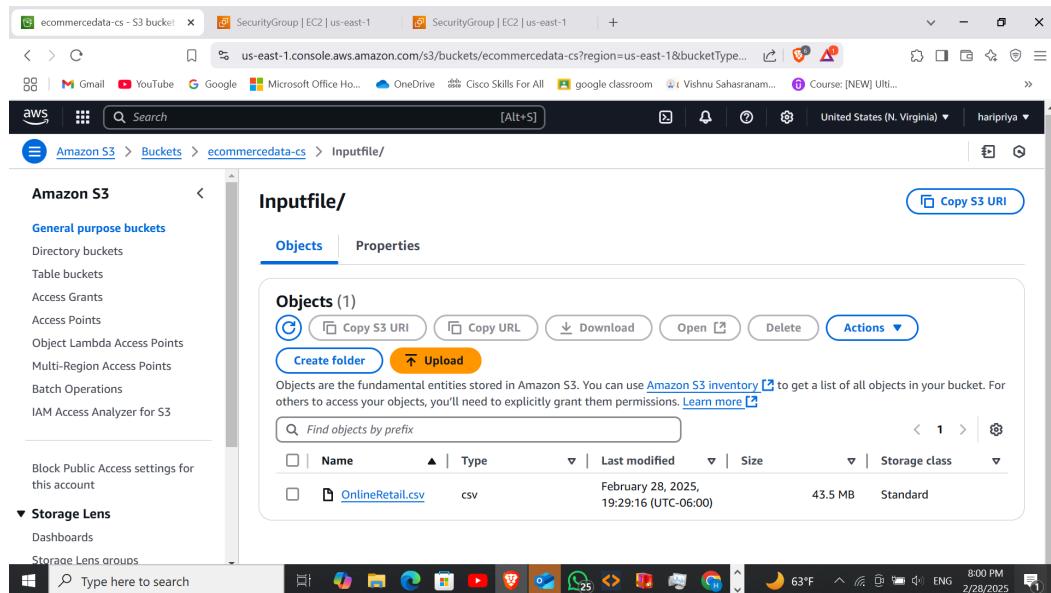


Big Data Driven Customer Segmentation In E-commerce

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 “ecommerce data - 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.



The screenshot shows the AWS S3 console interface. The left sidebar shows navigation options like 'Amazon S3', 'General purpose buckets', 'Storage Lens', and a search bar. The main area displays the 'Inputfile/' folder under the 'ecommerce data - cs' bucket. It shows one object, 'OnlineRetail.csv', which is a CSV file uploaded on February 28, 2025, at 19:29:16 (UTC-06:00), with a size of 43.5 MB and a standard storage class. There are buttons for 'Copy S3 URI', 'Actions', 'Create folder', and 'Upload'. A status bar at the bottom shows system icons and the date/time.

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.

The screenshot shows the AWS EMR console with the following details:

- Properties > Ecommerce > EMR**
- SecurityGroup | EC2 | us-east-1**
- SecurityGroup | EC2 | us-east-1**
- us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusterDetails/j-1C3UZQES5Y9VB**
- aws Search [Alt+S]**
- Gmail YouTube Google Microsoft Office Home OneDrive Cisco Skills For All google classroom Vishnu Sahasranam... Course: [NEW] Ulti...**
- United States (N. Virginia) haripriya**
- Amazon EMR > EMR on EC2: Clusters > Ecommerce**
- Your cluster "Ecommerce" has been successfully created.**
- Ecommerce** Updated less than a minute ago
- Summary**
- Cluster info**
 - Cluster ID: j-1C3UZQES5Y9VB
 - Cluster ARN: arn:aws:elasticmapreduce:us-east-1:127214202707:cluster/j-1C3UZQES5Y9VB
 - Cluster configuration: Instance groups
 - Capacity: 1 Primary | 1 Core | 1 Task
- Applications**
 - Amazon EMR version: emr-7.7.0
 - Installed applications: Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.5
- Cluster management**
 - Log destination in Amazon S3: aws-logs-127214202707-us-east-1/elasticmapreduce
 - Persistent application UIs: Spark History Server, YARN timeline server, Tez UI
- Status and time**
 - Status: Waiting
 - Creation time: February 28, 2025, 19:33 (UTC-06:00)
 - Elapsed time: 24 minutes, 19 seconds

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.

The screenshot shows the AWS EMR console with the following details:

- Properties > Ecommerce > EMR**
- SecurityGroup | EC2 | us-east-1**
- SecurityGroup | EC2 | us-east-1**
- us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusterDetails/j-1C3UZQES5Y9VB**
- aws Search [Alt+S]**
- Gmail YouTube Google Microsoft Office Home OneDrive Cisco Skills For All google classroom Vishnu Sahasranam... Course: [NEW] Ulti...**
- United States (N. Virginia) haripriya**
- Amazon EMR > EMR on EC2: Clusters > Ecommerce**
- On**
- Network and security**
- Network**
 - Virtual Private Cloud (VPC): vpc-0e714512323c2bb85
 - Subnet(s) and Availability Zone(s) (AZ): subnet-0c190b3fb68af44ec | us-east-1f
- EC2 security groups (firewall)**
 - Primary node: EMR managed security group sg-08dfc2919eb3a0bfb
 - Additional security groups: sg-04edf273fb120ab73
- Core and task nodes**
 - EMR managed security group sg-0d871a5b1265b38e9
- Security configuration**
 - Security configuration: None
 - EC2 key pair: Ecommerce key
- Permissions**
 - Service role for Amazon EMR: AmazonEMR-ServiceRole-20250216T182427
 - EC2 instance profile: AmazonEMR-InstanceProfile-20250216T182410
 - Custom automatic scaling role: Not configured

When clicked on 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 screenshot shows the AWS EC2 Security Groups Inbound rules page. The URL is <https://us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#SecurityGroup:group-id=sg-0d871a5b1265b38e9>. The page displays 7 inbound rules:

Name	Security group rule ID	IP version	Type	Protocol
-	sgr-0edfb02c2b14e1024	-	All UDP	UDP
-	sgr-0c1f6037c13b6d6a	-	All TCP	TCP
-	sgr-0ce31c27e26c02aef	-	All ICMP - IPv4	ICMP
-	sgr-0bca5f95f80b4c0f5	-	All ICMP - IPv4	ICMP
-	sgr-022371f7aaac4a8a	-	All UDP	UDP
-	sgr-09e30c57450782ff4	IPv4	SSH	TCP
-	sgr-06640612192c8862c	-	All TCP	TCP

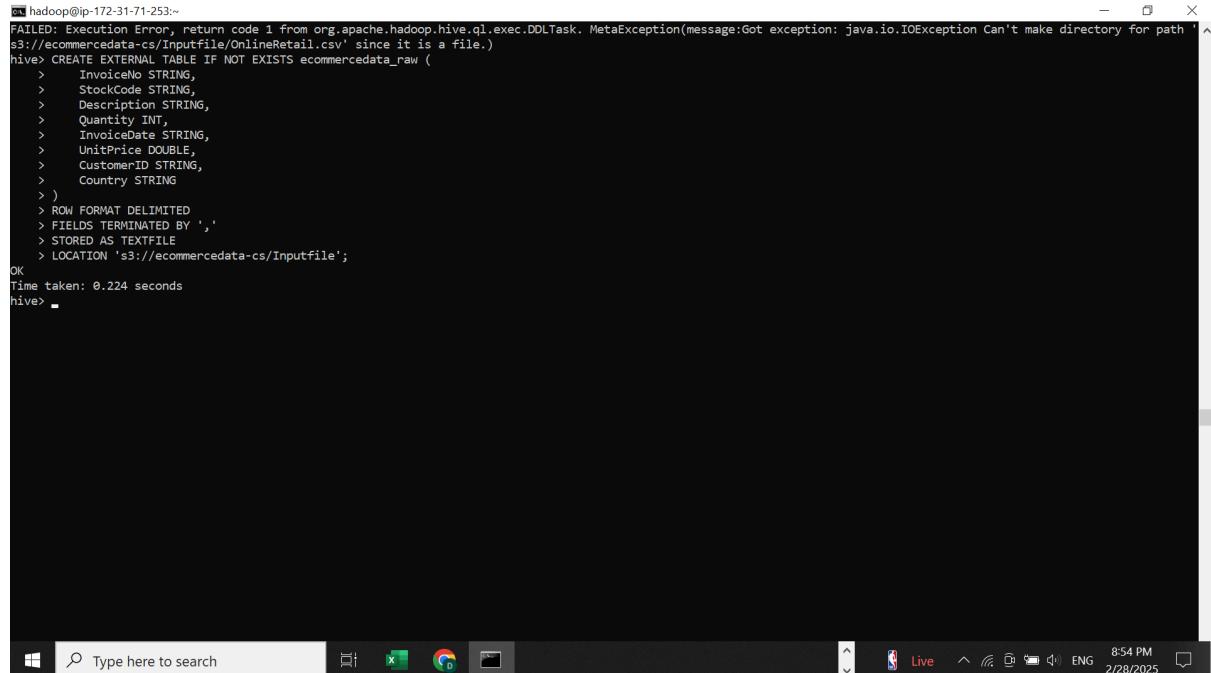
The same above process is repeated for core and task nodes EMR managed security groups.

The screenshot shows the AWS EC2 Security Groups Inbound rules page. The URL is <https://us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#SecurityGroup:group-id=sg-08dfc2919eb3a0fb>. The page displays 8 inbound rules:

Name	Security group rule ID	IP version	Type	Protocol
-	sgr-01bc9193813e13793	-	All TCP	TCP
-	sgr-0e648717414347dc0	-	All ICMP - IPv4	ICMP
-	sgr-004883c45df40b9bd	-	All UDP	UDP
-	sgr-0d096cbe6ecc7f5a	-	Custom TCP	TCP
-	sgr-0edccdf3c3723e392a	-	All UDP	UDP
-	sgr-0d209e2971b525eb1	-	All TCP	TCP
-	sgr-0bc5ad43559b6f7f9	-	All ICMP - IPv4	ICMP
-	sgr-09c275d1c8f5224b6	IPv4	SSH	TCP

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 “ecommerceadata_raw” in which i stored the data from online retail dataset by using the location of the S3 bucket as shown below “s3://ecommerceadata-cs/Inputfile”.

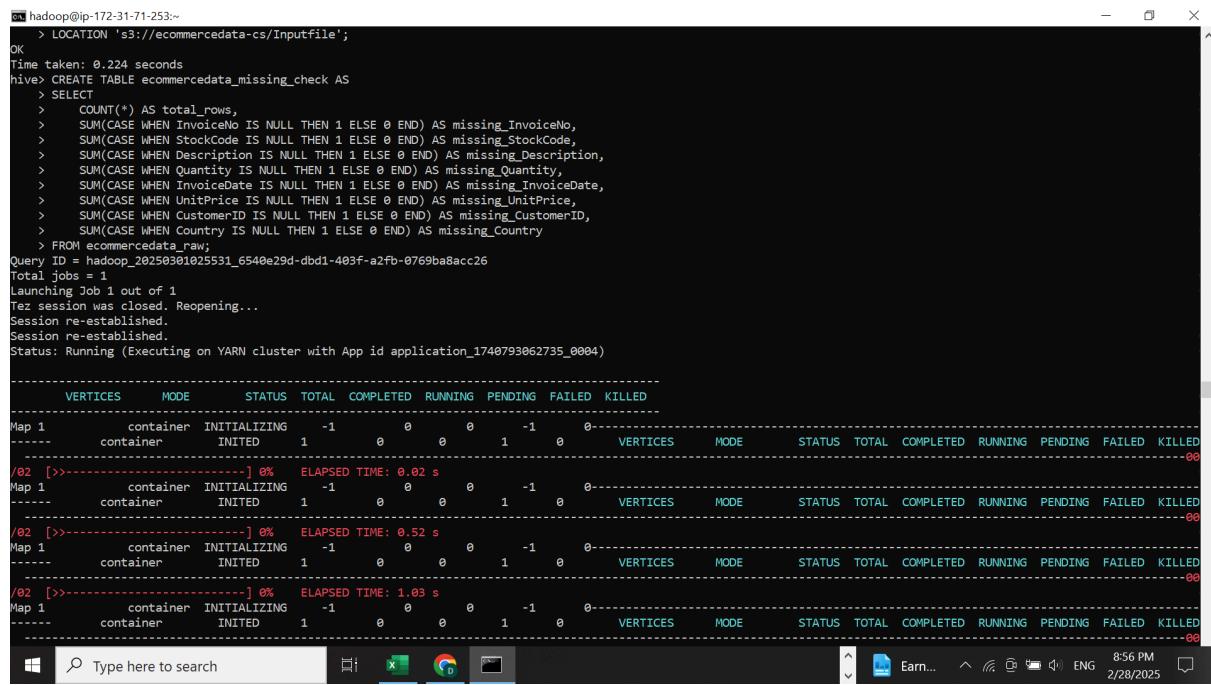


```

[hadoop@ip-172-31-71-253:~]
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. MetaException(message:Got exception: java.io.IOException Can't make directory for path 's3://ecommerceadata-cs/Inputfile/OnlineRetail.csv' since it is a file.)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS ecommerceadata_raw (
    >   InvoiceNo STRING,
    >   StockCode STRING,
    >   Description STRING,
    >   Quantity INT,
    >   InvoiceDate STRING,
    >   UnitPrice DOUBLE,
    >   CustomerID STRING,
    >   Country STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION 's3://ecommerceadata-cs/Inputfile';
OK
Time taken: 0.224 seconds
hive>

```

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



```

[hadoop@ip-172-31-71-253:~]
> LOCATION 's3://ecommerceadata-cs/Inputfile';
OK
Time taken: 0.224 seconds
hive> CREATE TABLE ecommerceadata_missing_check AS
> SELECT
    >   COUNT(*) AS total_rows,
    >   SUM(CASE WHEN InvoiceNo IS NULL THEN 1 ELSE 0 END) AS missing_InvoiceNo,
    >   SUM(CASE WHEN StockCode IS NULL THEN 1 ELSE 0 END) AS missing_StockCode,
    >   SUM(CASE WHEN Description IS NULL THEN 1 ELSE 0 END) AS missing_Description,
    >   SUM(CASE WHEN Quantity IS NULL THEN 1 ELSE 0 END) AS missing_Quantity,
    >   SUM(CASE WHEN InvoiceDate IS NULL THEN 1 ELSE 0 END) AS missing_InvoiceDate,
    >   SUM(CASE WHEN UnitPrice IS NULL THEN 1 ELSE 0 END) AS missing_UnitPrice,
    >   SUM(CASE WHEN CustomerID IS NULL THEN 1 ELSE 0 END) AS missing_CustomerID,
    >   SUM(CASE WHEN Country IS NULL THEN 1 ELSE 0 END) AS missing_Country
    > FROM ecommerceadata_raw;
Query ID = hadoop_20250301025531_6540e29d-db11-403f-a2fb-9769ba8acc26
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITIALIZING -1 0 0 -1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
--- container INITED 1 0 0 1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
/02 [>>>] 0% ELAPSED TIME: 0.02 s
Map 1 container INITIALIZING -1 0 0 -1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
--- container INITED 1 0 0 1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
/02 [>>>] 0% ELAPSED TIME: 0.52 s
Map 1 container INITIALIZING -1 0 0 -1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
--- container INITED 1 0 0 1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
/02 [>>>] 0% ELAPSED TIME: 1.03 s
Map 1 container INITIALIZING -1 0 0 -1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
--- container INITED 1 0 0 1 0 ----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED

```

For cleaned data a new table has been created with the name ecommerceadata_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.

```
hive> CREATE TABLE ecommerceadata_cleaned AS
> SELECT * FROM ecommerceadata_raw
> WHERE InvoiceNo IS NOT NULL
> AND StockCode IS NOT NULL
> AND InvoiceDate IS NOT NULL
> AND CustomerID IS NOT NULL;
Query ID = hadoop_20250301025704_dcac2411-4060-459a-b998-f6a3c0e1d75d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)

-----  
 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 ..... container SUCCEEDED 3      3      0      0      0      0  
-----  
VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 10.52 s
```



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

```

hadoop@ip-172-31-71-253:~>
> ;
hive> CREATE TABLE ecommerce_data_filled_description AS
> SELECT InvoiceNo, StockCode,
> COALESCE>Description, 'Unknown') AS Description,
> Quantity, InvoiceDate, UnitPrice, CustomerID, Country
> FROM ecommerce_data_cleaned;
Query ID = hadoop_20250301025905_b6733d25-4448-4b12-a7fb-7d9c23d3ba71
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

RTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container SUCCEEDED 3 3 0 0 0 0  

-----  

VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 7.20 s  

-----  

Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommerce_data_filled_description
OK  

Time taken: 7.509 seconds
hive>

```

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

```

hadoop@ip-172-31-71-253:~>
Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommerce_data_filled_description
OK  

Time taken: 7.509 seconds
hive> CREATE TABLE ecommerce_data_description_check AS
> SELECT
> COUNT(*) AS missing_description_count,
> COUNT(CASE WHEN LOWER(Description) = 'unknown' THEN 1 ELSE NULL END) AS unknown_description_count,
> COUNT(CASE WHEN TRIM(Description) = '' THEN 1 ELSE NULL END) AS empty_description_count
> FROM ecommerce_data_filled_description;
Query ID = hadoop_20250301025953_58ec485e-b49c-4913-a917-9dfdeb3a5b4e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container INITIALIZING -1 0 0 0 -1 0 0  

-----  

Map 1 ..... container INITED 1 0 0 0 1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

/02 [>>>-----] 0% ELAPSED TIME: 0.00 s  

Map 1 ..... container INITED 3 0 0 0 3 0  

-----  

Map 1 ..... container INITED 1 0 0 0 1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

/02 [>>>-----] 0% ELAPSED TIME: 0.50 s  

Map 1 ..... container INITED 3 0 0 0 3 0  

-----  

Map 1 ..... container INITED 1 0 0 0 1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

/02 [>>>-----] 0% ELAPSED TIME: 1.01 s  

Map 1 ..... container INITED 3 0 0 0 3 0  

-----  

Map 1 ..... container INITED 1 0 0 0 1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

/02 [>>>-----] 0% ELAPSED TIME: 1.51 s  

Map 1 ..... container INITED 3 0 0 0 3 0  

-----  

Map 1 ..... container INITED 1 0 0 0 1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

/02 [>>>-----] 0% ELAPSED TIME: 2.02 s  

Map 1 ..... container INITED 3 0 0 0 3 0  

-----  

Map 1 ..... container INITED 1 0 0 0 1 0 VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

/02 [>>>-----] 0% ELAPSED TIME: 2.52 s  


```

```

cd hadoop@ip-172-31-71-253:~
Map 1      container    INITED     3      0      0      0      1      3      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-->-----] 0% ELAPSED TIME: 0.50 s
Map 1      container    INITED     3      0      0      0      1      3      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-->-----] 0% ELAPSED TIME: 1.01 s
Map 1      container    INITED     3      0      0      0      1      3      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-->-----] 0% ELAPSED TIME: 1.51 s
Map 1      container    INITED     3      0      0      0      1      3      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-->-----] 0% ELAPSED TIME: 2.02 s
Map 1      container    INITED     3      0      0      0      1      3      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-->-----] 0% ELAPSED TIME: 2.52 s
Map 1      container    RUNNING    3      0      1      2      0      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-->-----] 0% ELAPSED TIME: 3.03 s
Map 1      container    RUNNING    3      0      2      1      0      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED  3      3      0      0      0      0      0
Reducer 2 .. container SUCCEEDED  1      1      0      0      0      0      0
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 6.55 s
Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommerceadata_description_check
OK
Time taken: 6.907 seconds
hive>

```

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

```

cd hadoop@ip-172-31-71-253:~
hive> CREATE TABLE ecommerceadata_final_description AS
> SELECT InvoiceNo, StockCode,
>        CASE
>          WHEN Description IS NULL OR TRIM(Description) = '' OR LOWER(Description) = 'unknown'
>          THEN 'Not Available'
>          ELSE Description
>        END AS Description,
>        Quantity, InvoiceDate, UnitPrice, CustomerID, Country
> FROM ecommerceadata_filled_description;
Query ID = hadoop_20250301030332_12d8e5e9-3a17-49e3-a996-cdeca61226f5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)

----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED ----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
ICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED ----- container   RUNNING    3      1      2      0      0- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
DE     STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED 00/01 [=====>>>] 33% ELAPSED TIME: 6.53 s
----- VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED ----- Map 1 ...
..... container SUCCEEDED  3      3      0      0      0      0
ICES: 01/01 [=====>>>] 100% ELAPSED TIME: 7.95 s ----- Move
ng data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommerceadata_final_description----- ] 0% ELAPSED TIME: 6.03 s

```

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

```
c:\ hadoop@ip-172-31-71-253:~  
> _____  
> ;  
hive> CREATE TABLE ecommerce_data.date_check AS  
> SELECT DISTINCT InvoiceDate FROM ecommerce_data_final_description LIMIT 10;  
Query ID = hadoop_20250301030438_7e502501-82b6-428d-ac0-4740d5108426  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)  
  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITIALIZING -1 0 0 -1 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITED 3 0 0 3 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITED 3 0 0 3 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITED 3 0 0 3 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITED 3 0 0 3 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITED 3 0 0 3 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container INITED 3 0 0 3 0 0  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
Map 1 container RUNNING 3 0 1 2 0 0  
-----  
Windows Type here to search ↑ Earn... ⌂ ENG 9:05 PM 2/28/2025
```

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

```
hive> CREATE TABLE ecommerce_data_fixed_date AS  
> SELECT InvoiceNo, StockCode, Description, Quantity,  
>       FROM_UNIXTIME(UNIX_TIMESTAMP(InvoiceDate, 'MM/dd/yyyy HH:mm')) AS InvoiceDate,  
>       UnitPrice, CustomerID, Country  
> FROM ecommerce_data_final_description;  
Query ID = hadoop_20250301030608_60f8c1ad-2142-47cb-a388-c60d8fb7852f  
Total jobs = 1  
Launching Job 1 out of 1  
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)  
  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  
-----  
Windows Type here to search ↑ Earn... ⌂ ENG 9:06 PM 2/28/2025
```

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

```
hadoop@ip-172-31-71-253:~
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0

VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 8.46 s

Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommercefixed_date

OK

Time taken: 8.711 seconds

```
hive> CREATE TABLE ecommercefixed_data_invalid_values AS
> SELECT * FROM ecommercefixed_data
> WHERE Quantity < 0 OR UnitPrice < 0;
Query ID = hadoop_20250301030707_2af4506b-0520-4680-99ff-15b19e87a2e0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0

VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 7.43 s

Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommercefixed_data_invalid_values

OK



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.

```
hadoop@ip-172-31-71-253:~
```

Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0

VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 7.43 s

Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommercefixed_data_invalid_values

OK

Time taken: 7.715 seconds

```
hive> CREATE TABLE ecommercefixed_data_final_cleaned AS
> SELECT * FROM ecommercefixed_data
> WHERE Quantity >= 0 AND UnitPrice >= 0;
Query ID = hadoop_20250301030743_4e914aca-6db4-4082-ac07-ebe718181b33
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0004)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	3	3	0	0	0	0

VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 7.79 s

Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommercefixed_data_final_cleaned

OK



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

```

hadoop@ip-172-31-71-253:~ 
581237 23509 MINI PLAYING CARDS FUN FAIR 2 2011-12-08 10:40:00 0.42
5443 United Kingdom 2
581310 23109 PACK OF SIX LED TEA LIGHTS 1 2011-12-08 11:43:00 2.89
6442 United Kingdom 2
581405 20975 12 PENCILS SMALL TUBE RED RETROSPOT 1 2011-12-08 13:50:00
65 13521 United Kingdom 2
581412 22199 FRYING PAN RED RETROSPOT 1 2011-12-08 14:38:00 1.25
4415 United Kingdom 2
581414 22326 ROUND SNACK BOXES SET OF4 WOODLAND 1 2011-12-08 14:39:00
95 14730 United Kingdom 2
581456 35964 FOLKART CLIP ON STARS 2 2011-12-08 18:42:00 0.39 17530
United Kingdom 2
581514 22075 6 RIBBONS ELEGANT CHRISTMAS 24 2011-12-09 11:20:00 0.39
7754 United Kingdom 2
Time taken: 15.988 seconds, Fetched: 4805 row(s)
hive> CREATE TABLE ecommerceedata_no_duplicates AS
> SELECT DISTINCT * FROM ecommerceedata_final_cleaned;
Query ID = hadoop_20250301034151_4c161ae4-13c4-4f77-976f-7865ded92089
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0005)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITIALIZING -1 0 0 0 -1 0
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITED 6 0 0 0 6 0
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITED 3 0 0 0 3 0
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container RUNNING 3 2 1 0 0 0
Reducer 2 ..... container RUNNING 6 0 2 4 0 0
VERTICES: 00/02 [=====>-----] 22% ELAPSED TIME: 7.57 s
Windows Type here to search 9:41 PM 2/28/2025

```

Performing normalization on Numerical columns - Quantity, UnitPrice.

```

hadoop@ip-172-31-71-253:~ 
Time taken: 16.955 seconds
hive> SELECT
> MIN(Quantity) AS min_quantity, MAX(Quantity) AS max_quantity,
> MIN(UnitPrice) AS min_price, MAX(UnitPrice) AS max_price
> FROM ecommerceedata_no_duplicates;
Query ID = hadoop_20250301034221_8544abb5-ba3d-47f8-96e6-9a779f5ab7d4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0005)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITIALIZING -1 0 0 0 -1 0
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITED 1 0 0 0 1 0
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 container INITED 4 0 0 0 4 0
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 4 4 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 02/02 [=====>-----] 100% ELAPSED TIME: 6.81 s
OK
1 74215 0.0 13541.33
Time taken: 7.018 seconds, Fetched: 1 row(s)
hive> 
Windows Type here to search 9:42 PM 2/28/2025

```

Apply Scaling with The formula : $X' = (X - X_{min}) / (X_{max} - X_{min})$

```
hadoop@ip-172-31-71-253:~$ 
OK
1    74215  0.0    13541.33
Time taken: 7.018 seconds, Fetched: 1 row(s)
hive> CREATE TABLE ecommercedata_scaled AS
> SELECT InvoiceNo, StockCode, Description,
>        (Quantity - min_q) / (max_q - min_q) AS Quantity_Scaled,
>        InvoiceDate,
>        (UnitPrice - min_p) / (max_p - min_p) AS UnitPrice_Scaled,
>        CustomerID, Country
> FROM ecommercedata_no_duplicates
> CROSS JOIN (SELECT
>             MIN(Quantity) AS min_q, MAX(Quantity) AS max_q,
>             MIN(UnitPrice) AS min_p, MAX(UnitPrice) AS max_p
>           FROM ecommercedata_no_duplicates) stats;
Warning: Map Join MAPJOIN[13][bigTable[?]] in task 'Map 1' is a cross product
Query ID = hadoop_20250301034436_d3647ca2-ef00-401f-b457-0d2bfff1c468
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1740793062735_0005)

-----  

VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 2      container INITIALIZING   -1      0      0      -1      0      0
-----  

VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 2      container INITED       4      0      0      4      0      0
-----  

VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 2 ..... container SUCCEEDED     4      4      0      0      0      0
Reducer 3 ..... container SUCCEEDED     1      1      0      0      0      0
Map 1 ..... container SUCCEEDED     4      4      0      0      0      0
-----  

VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 13.42 s
-----  

Moving data to directory hdfs://ip-172-31-71-253.ec2.internal:8020/user/hive/warehouse/ecommercedata_scaled
OK
Time taken: 13.906 seconds      INITED      4      0      0      0      1      4      0      0      0
-----  

hive>      container INITED       1      0      0      0      0      0      0      0      0
-----  

Windows Taskbar: Type here to search, File Explorer, Google Chrome, File Manager, Task View, Taskbar icons, System tray showing 60°F, ENG, 945 PM, 2/28/2025
```

```
hadoop@ip-172-31-71-253:~$ 
hive> SELECT * FROM ecommercedata_scaled LIMIT 10;
OK
536366  22633  HAND WARMER UNION JACK  6.737273290753766E-5  2010-12-01 08:28:00
1.366187811684672E-4  17850  United Kingdom
536367  22745  POPPY'S PLAYHOUSE BEDROOM  6.737273290753766E-5  2010-12-01 08:34:00 1.5508077862366547E-4  13047  United Kingdom
536367  22749  FELTCRAFT PRINCESS CHARLOTTE DOLL  9.432182607055272E-5  2010-12-01 08:34:00 2.7692996182797407E-4  13047  United Kingdom
536370  21724  PANDA AND BUNNIES STICKER SHEET 1.4822001239658285E-4  2010-12-01 08:45:00 6.277079134767411E-5  12583  France
536370  21731  RED TOADSTOOL LED NIGHT LIGHT  3.099145713746732E-4  2010-12-01 08:45:00 1.2184918320430858E-4  12583  France
536370  21913  VINTAGE SEASIDE JIGSAW PUZZLES 1.4822001239658285E-4  2010-12-01 08:45:00 2.7692996182797407E-4  12583  France
536370  22326  ROUND SNACK BOXES SET OF4 WOODLAND  3.099145713746732E-4  2010-12-01 08:45:00 2.178515699713396E-4  12583  France
536370  22659  LUNCH BOX I LOVE LONDON 3.099145713746732E-4  2010-12-01 08:45:00
1.440035801505465E-4  12583  France
536370  22728  ALARM CLOCK BAKELIKE PINK  3.099145713746732E-4  2010-12-01 08:45:00 2.7692996182797407E-4  12583  France
536370  22900  SET 2 TEA TOWELS I LOVE LONDON  3.099145713746732E-4  2010-12-01 08:45:00 2.178515699713396E-4  12583  France
Time taken: 0.057 seconds, Fetched: 10 row(s)
hive>
```

Windows Taskbar: Type here to search, File Explorer, Google Chrome, File Manager, Task View, Taskbar icons, System tray showing 60°F, ENG, 946 PM, 2/28/2025

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

```

hadoop@ip-172-31-71-23:~ 
536370 22900 SET 2 TEA TOWELS I LOVE LONDON 3.099145713746732E-4 2010-12-01 08:45:00 2.178515699713396E-4 12583 France
Time taken: 0.057 seconds. Fetched: 10 row(s)
hive> INSERT OVERWRITE DIRECTORY 's3://ecommercecleaneddata/ecommerce-cleaned-data/'
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ','
> SELECT * FROM ecommercecleaneddata_scaled;
Query ID = hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1740793062735_0006)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
```

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

Amazon S3 > Buckets > ecommercecleaneddata > ecommerce-cleaned-data/

Objects (6)

- [Copy S3 URI](#)
- [Copy URL](#)
- [Download](#)
- [Open](#)
- [Delete](#)
- [Actions](#)
- [Create folder](#)
- [Upload](#)

Objects:

Name	Type	Last modified	Size	Storage class
_SUCCESS	-	February 28, 2025, 21:55:55 (UTC-06:00)	0 B	Standard
00000_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:50 (UTC-06:00)	16.0 MB	Standard
00001_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:51 (UTC-06:00)	16.0 MB	Standard

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with navigation links like 'Amazon S3', 'General purpose buckets', 'Storage Lens', and 'CloudShell'. The main area is titled 'Objects (6)' and lists six files. The files are:

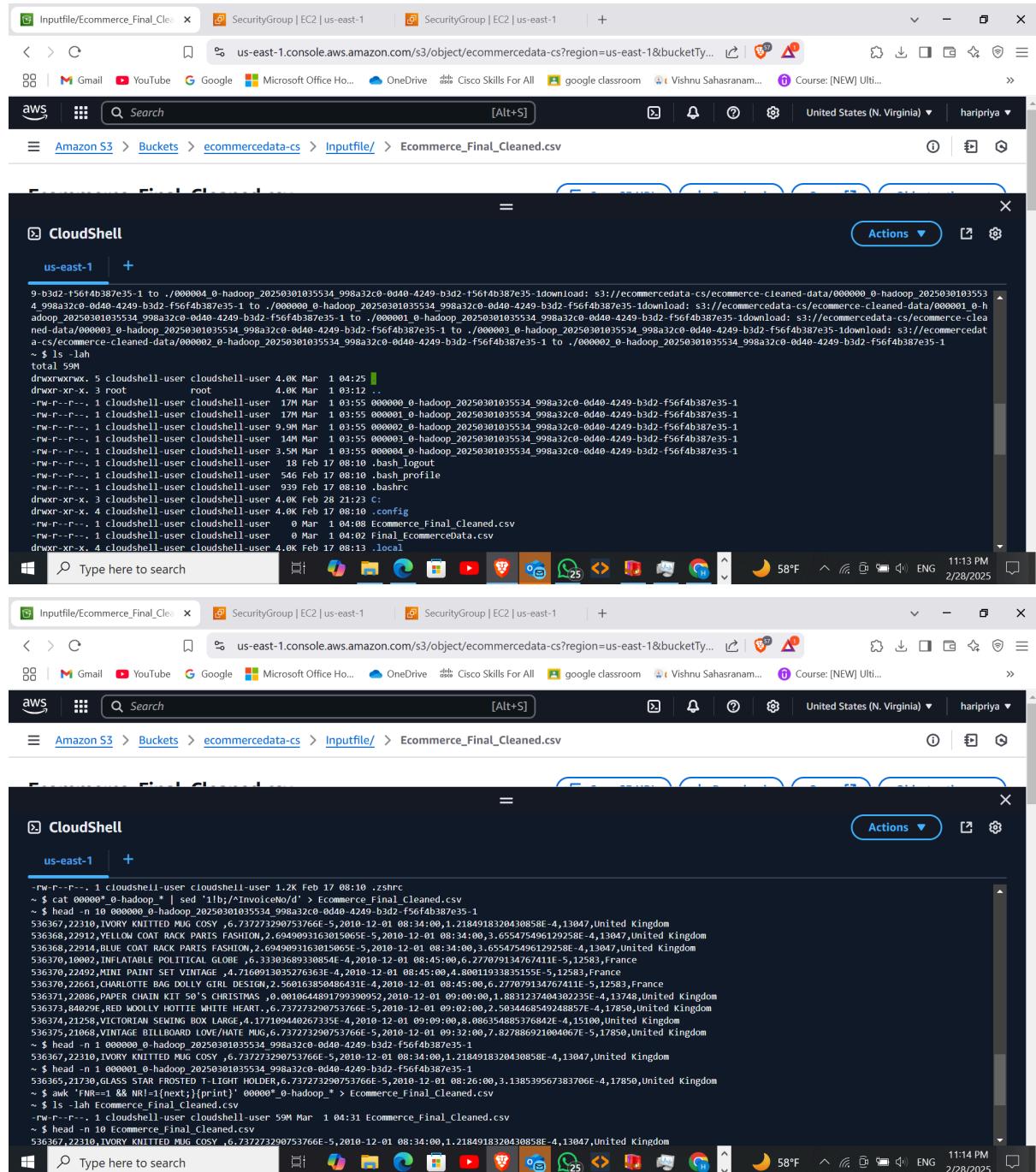
Name	Type	Last modified	Size	Storage class
000002_0-hadoop_2025030103	-	February 28, 2025, 21:55:49 (UTC-06:00)	9.8 MB	Standard
5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:54 (UTC-06:00)	13.3 MB	Standard
000003_0-hadoop_2025030103	-	February 28, 2025, 21:55:54 (UTC-06:00)	13.3 MB	Standard
5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:54 (UTC-06:00)	13.3 MB	Standard
000004_0-hadoop_2025030103	-	February 28, 2025, 21:55:53 (UTC-06:00)	3.5 MB	Standard
5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:53 (UTC-06:00)	3.5 MB	Standard

At the bottom right of the main area, there are links for 'Privacy', 'Terms', and 'Cookie preferences'.

This screenshot is nearly identical to the one above, showing the same list of six objects in the 'ecommerce-cleaned-data' bucket. The only difference is in the file names, where the second and third entries have been swapped. The first and fourth entries remain the same, and the fifth and sixth entries also remain the same.

Name	Type	Last modified	Size	Storage class
000002_0-hadoop_2025030103	-	February 28, 2025, 21:55:49 (UTC-06:00)	9.8 MB	Standard
5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:54 (UTC-06:00)	13.3 MB	Standard
000003_0-hadoop_2025030103	-	February 28, 2025, 21:55:54 (UTC-06:00)	13.3 MB	Standard
5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:54 (UTC-06:00)	13.3 MB	Standard
000004_0-hadoop_2025030103	-	February 28, 2025, 21:55:53 (UTC-06:00)	3.5 MB	Standard
5534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1	-	February 28, 2025, 21:55:53 (UTC-06:00)	3.5 MB	Standard

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 is done in the AWS CloudShell



```
ls -lah
total 59M
drwxrwxrwx. 5 cloudshell-user cloudshell-user 4.0K Mar  1 04:25 .
drwxr-xr-x. 3 root          root      4.0K Mar  1 03:12 ..
-rw-r--r--. 1 cloudshell-user cloudshell-user 17M Mar  1 03:55 000000_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
-rw-r--r--. 1 cloudshell-user cloudshell-user 17M Mar  1 03:55 000001_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
-rw-r--r--. 1 cloudshell-user cloudshell-user 9.9M Mar  1 03:55 000002_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
-rw-r--r--. 1 cloudshell-user cloudshell-user 14M Mar  1 03:55 000003_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
-rw-r--r--. 1 cloudshell-user cloudshell-user 3.5M Mar  1 03:55 000004_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
-rw-r--r--. 1 cloudshell-user cloudshell-user 18 Feb 17 08:10 .bash_logout
-rw-r--r--. 1 cloudshell-user cloudshell-user 546 Feb 17 08:10 .bash_profile
-rw-r--r--. 1 cloudshell-user cloudshell-user 939 Feb 17 08:10 .bashrc
drwxr-xr-x. 4 cloudshell-user cloudshell-user 4.0K Feb 17 08:10 .config
-rw-r--r--. 1 cloudshell-user cloudshell-user 0 Mar  1 04:08 Ecommerce_Final_Cleaned.csv
-rw-r--r--. 1 cloudshell-user cloudshell-user 0 Mar  1 04:02 Final_EcommerceData.csv
drwxr-xr-x. 4 cloudshell-user cloudshell-user 4.0K Feb 17 08:13 .local

ls -lah
total 59M
-rw-r--r--. 1 cloudshell-user cloudshell-user 1.2K Feb 17 08:10 .zshrc
~ $ cat 00000*_0-hadoop* | sed '1b;/^InvoiceNo/D' > Ecommerce_Final_Cleaned.csv
~ $ head -n 10 000000_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
536367,22310,IVORY KNITTED MUG COSY ,6,737273290753766E-5,2010-12-01 08:34:00,1,2184918320430858E-4,13047,United Kingdom
536368,22912,YELLOW COAT RACK PARTS FASHION,2,6940993163015065E-5,2010-12-01 08:34:00,3,655475496129258E-4,13047,United Kingdom
536369,10002,INFLATABLE POLITICAL GLOBE ,6,3303699330854E-4,2010-12-01 08:45:00,6,27707913476411E-5,17583,France
536370,22492,MINI PAINT SET VINTAGE ,4,7166991303527636E-4,2010-12-01 08:45:00,4,8001193835155E-5,12583,France
536371,22661,CHARLOTTE BAB DOLLY GIRL DESIGN,2,56016385048641E-4,2010-12-01 08:45:00,6,27707913476411E-5,12583,France
536372,22886,PAPER CHAIN KIT 50'S CHRISTMAS ,0,001064489179939095E,2010-12-01 09:00:00,1,881217404302235E-4,13748,United Kingdom
536373,84029E,RED WOOLLY HOTTIE WHITE HEART,6,737273290753766E-5,2010-12-01 09:02:00,2,503468549248857E-4,17850,United Kingdom
536374,21258,VICTORIAN SEWING BOX LARGE,4,17109440467335E-4,2010-12-01 09:09:00,8,086354885376842E-4,15160,United Kingdom
536375,21068,VINTAGE BILLBOARD LOVE/HATE MUG ,6,737273290753766E-5,2010-12-01 09:32:00,7,827886921004067E-5,17850,United Kingdom
~ $ head -n 1 000000_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
536367,22310,IVORY KNITTED MUG COSY ,6,737273290753766E-5,2010-12-01 08:34:00,1,2184918320430858E-4,13047,United Kingdom
~ $ head -n 1 000001_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56f4b387e35-1
536368,21730,GLASS STAR FROSTED T-LIGHT HOLDER,6,737273290753766E-5,2010-12-01 08:26:00,3,138539567383706E-4,17850,United Kingdom
~ $ awk '{FN=1 && NR!=1}{print}' 00000*_0-hadoop_> Ecommerce_Final_cleaned.csv
~ $ ls -lah Ecommerce_Final_cleaned.csv
-rw-r--r--. 1 cloudshell-user cloudshell-user 59M Mar  1 04:31 Ecommerce_Final_cleaned.csv
536367,22310,IVORY KNITTED MUG COSY ,6,737273290753766E-5,2010-12-01 08:34:00,1,2184918320430858E-4,13047,United Kingdom
```

```

536375,21068,VINTAGE BILLBOARD LOVE/HATE MUG,6,737223290753766E-5,2010-12-01 09:32:00,7,827886921004067E-5,17850,United Kingdom
~ $ head -n 1 000000_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56fd3b387e35-1
536367,22310,IVORY KNITTED MUG COSY ,6,737223290753766E-5,2010-12-01 08:34:00,1,2184918320430858E-4,13047,United Kingdom
~ $ head -n 1 000001_0-hadoop_20250301035534_998a32c0-0d40-4249-b3d2-f56fd3b387e35-1
536365,21730,GLASS STAR FROSTED T-LIGHT HOLDER,6,737223290753766E-5,2010-12-01 08:26:00,3,138539567383706E-4,17850,United Kingdom
~ $ awk '{NR==1 && NR!=1}{print}' 000000_0-hadoop_> Ecommerce_Final_Cleaned.csv
~ $ ls -lah Ecommerce_Final_Cleaned.csv
-rw-r--r--. 1 cloudshell-user cloudshell-user 59M Mar 1 04:31 Ecommerce_Final_Cleaned.csv
~ $ head -n 10 Ecommerce_Final_Cleaned.csv
536367,22310,IVORY KNITTED MUG COSY ,6,737223290753766E-5,2010-12-01 08:34:00,1,2184918320430858E-4,13047,United Kingdom
536368,22911,YELLOW COAT RACK PARIS FASHION,2,6940993163015065E-5,2010-12-01 08:34:00,3,655475496129258E-4,13047,United Kingdom
536368,22914,BLUE COAT RACK PARIS FASHION,2,6940993163015065E-5,2010-12-01 08:34:00,3,655475496129258E-4,13047,United Kingdom
536370,22492,MINI PAINT SET VINTAGE ,4,71689913035276363E-4,2010-12-01 08:45:00,4,80011933835155E-5,12583,France
536370,22492,CHARLOTTE BAG DOLLY GIRL DESIGN,2,560163850486431E-4,2010-12-01 08:45:00,6,277079134767411E-5,12583,France
536371,22086,PAPER CHAIN KIT 50'S CHRISTMAS ,0,0010644891799390952,2010-12-01 09:00:00,1,8831237404302235E-4,13748,United Kingdom
536367,22310,IVORY KNITTED MUG COSY ,6,737223290753766E-5,2010-12-01 08:45:00,6,277079134767411E-5,12583,France
536371,22086,PAPER CHAIN KIT 50'S CHRISTMAS ,0,0010644891799390952,2010-12-01 09:00:00,1,8831237404302235E-4,13748,United Kingdom
536373,84029E,RED WOOLLY HOTTIE WHITE HEART ,6,737223290753766E-5,2010-12-01 09:02:00,2,5034468549248857E-4,17850,United Kingdom
~ $ aws s3 cp Ecommerce_Final_Cleaned.csv s3://ecommerceadata-cs/Inputfile/Ecommerce_Final_Cleaned.csv
upload: ./Ecommerce_Final_Cleaned.csv to s3://ecommerceadata-cs/Inputfile/Ecommerce_Final_Cleaned.csv
~ $ 

```

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

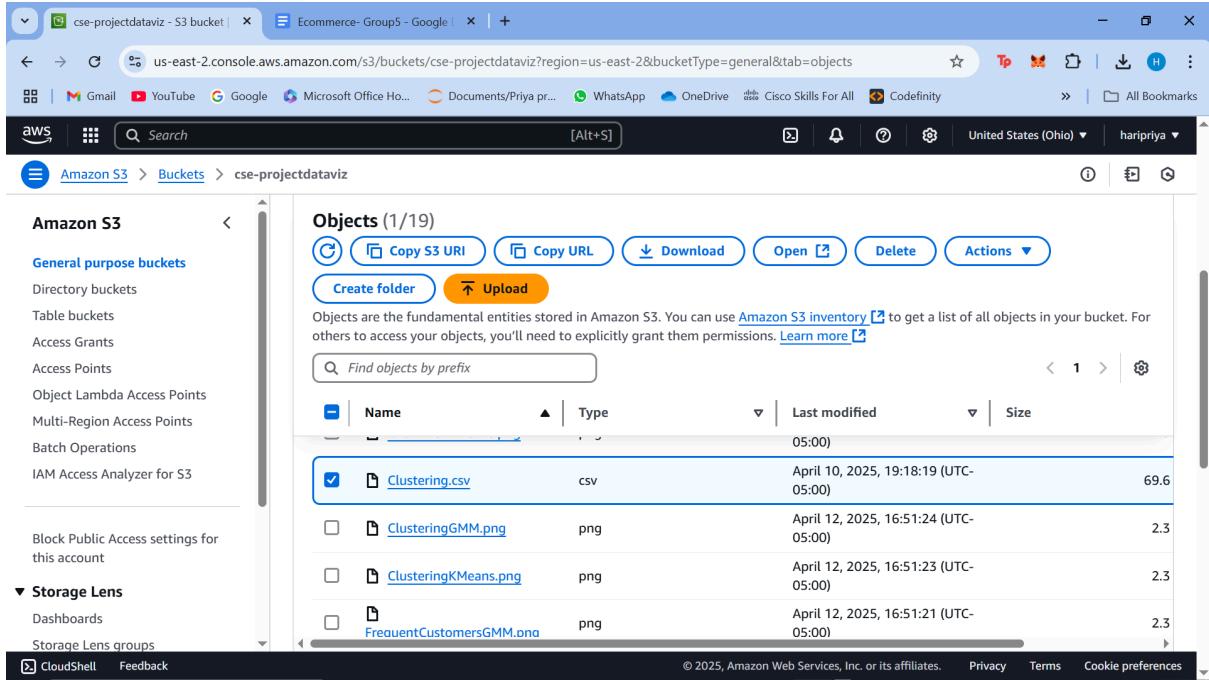
Once the whole process is done terminate the cluster.

Cluster info		Applications	Cluster management	Status and time
Cluster ID j-1C3UZQES5Y9VB		Amazon EMR version emr-7.0	Log destination in Amazon S3 aws-logs-127214202707-us-east-1/elasticmapreduce	Status Terminated
Cluster ARN arn:aws:elasticmapreduce:us-east-1:2714202707:cluster/j-1C3UZQES5Y9VB		Installed applications Hadoop 3.4.0, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.3	Persistent application UIs Spark History Server YARN timeline server Tez UI	Creation time February 28, 2025, 19:33 (UTC-06:00)
Cluster configuration Instance groups			Primary node public DNS ec2-98-80-219-11.compute-1.amazonaws.com	Elapsed time 3 hours, 30 minutes
Capacity 1 Primary 1 Core 1 Task			Connect to the Primary node using SSH	End time February 28, 2025, 23:03 (UTC-06:00)

Properties Bootstrap actions Instances (Hardware) Steps Applications Configurations Monitoring Events Tags (1)

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



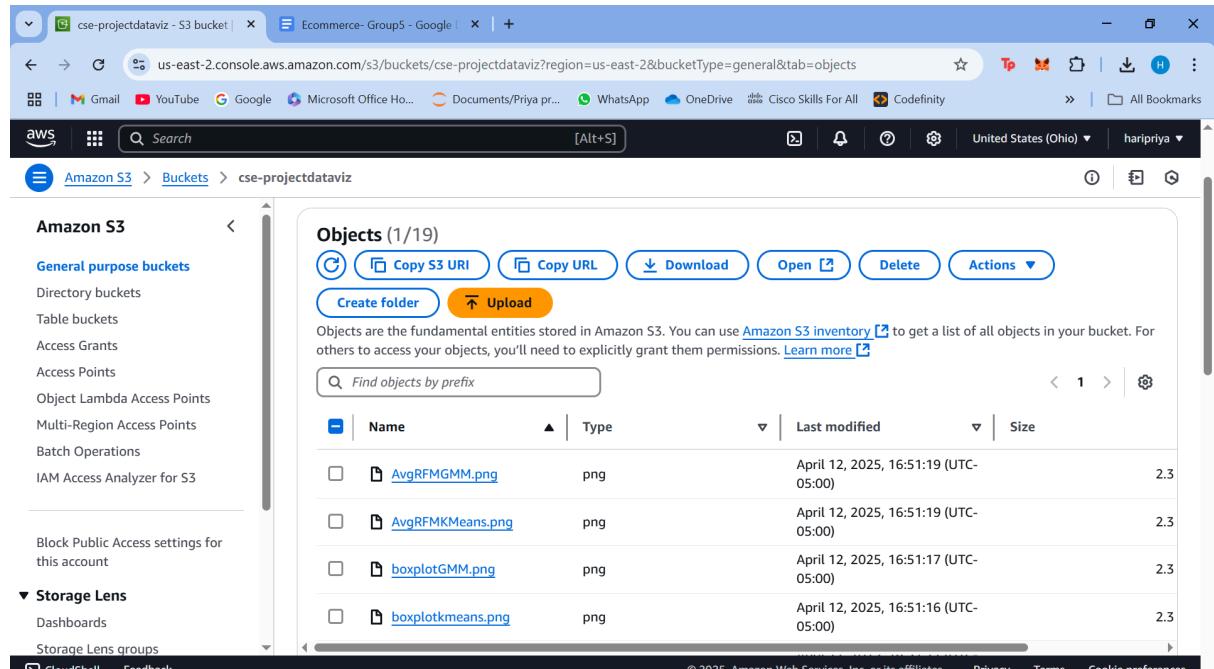
The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with 'Amazon S3' navigation and various service links like General purpose buckets, IAM Access Analyzer for S3, and Storage Lens. The main area is titled 'Objects (1/19)' and lists four items:

Name	Type	Last modified	Size
Clustering.csv	csv	April 10, 2025, 19:18:19 (UTC-05:00)	69.6
ClusteringGMM.png	png	April 12, 2025, 16:51:24 (UTC-05:00)	2.3
ClusteringKMeans.png	png	April 12, 2025, 16:51:23 (UTC-05:00)	2.3
FrequentCustomersGMM.png	png	April 12, 2025, 16:51:21 (UTC-05:00)	2.3

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



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

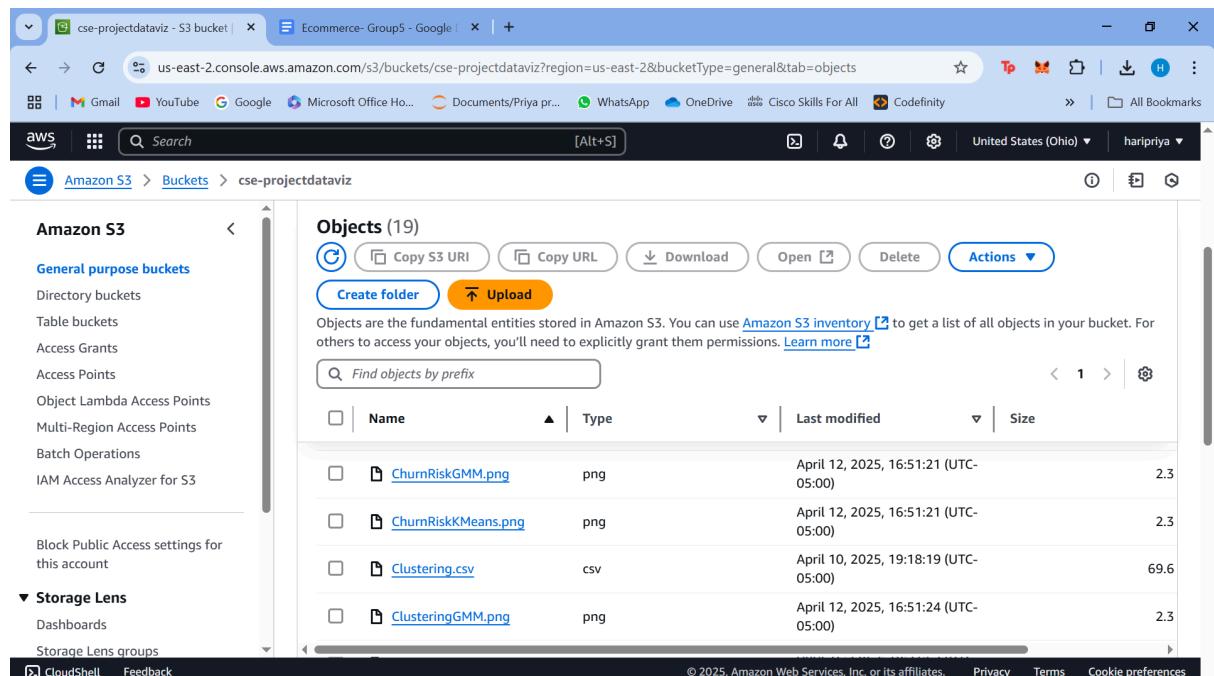
- Dashboards
- Storage Lens groups

CloudShell Feedback

Objects (1/19)

Name	Type	Last modified	Size
AvgRFMGMM.png	png	April 12, 2025, 16:51:19 (UTC-05:00)	2.3
AvgRFMKMeans.png	png	April 12, 2025, 16:51:19 (UTC-05:00)	2.3
boxplotGMM.png	png	April 12, 2025, 16:51:17 (UTC-05:00)	2.3
boxplotkmeans.png	png	April 12, 2025, 16:51:16 (UTC-05:00)	2.3

© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences



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

CloudShell Feedback

Objects (19)

Name	Type	Last modified	Size
ChurnRiskGMM.png	png	April 12, 2025, 16:51:21 (UTC-05:00)	2.3
ChurnRiskKMeans.png	png	April 12, 2025, 16:51:21 (UTC-05:00)	2.3
Clustering.csv	csv	April 10, 2025, 19:18:19 (UTC-05:00)	69.6
ClusteringGMM.png	png	April 12, 2025, 16:51:24 (UTC-05:00)	2.3

© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Screenshot of the AWS S3 console showing the contents of the 'cse-projectdataviz' bucket.

The left sidebar shows the navigation path: Amazon S3 > Buckets > cse-projectdataviz.

The main area displays 19 objects:

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

Actions buttons available for each object include Copy S3 URI, Copy URL, Download, Open, Delete, and Actions.

Screenshot of the AWS S3 console showing the contents of the 'cse-projectdataviz' bucket.

The left sidebar shows the navigation path: Amazon S3 > Buckets > cse-projectdataviz.

The main area displays 19 objects:

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

Actions buttons available for each object include Copy S3 URI, Copy URL, Download, Open, Delete, and Actions.

Screenshot of the AWS S3 console showing the contents of the 'cse-projectdataviz' bucket.

The left sidebar shows the navigation path: Amazon S3 > Buckets > cse-projectdataviz.

The main area displays 19 objects:

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

Actions buttons available for each object include Copy S3 URI, Copy URL, Download, Open, Delete, and Actions.

Screenshot of the AWS S3 console showing the contents of the 'cse-projectdataviz' bucket.

The left sidebar shows the navigation path: Amazon S3 > Buckets > cse-projectdataviz.

The main area displays 19 objects:

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

Actions buttons available for each object include Copy S3 URI, Copy URL, Download, Open, Delete, and Actions.

Screenshot of the AWS S3 console showing the contents of the 'cse-projectdataviz' bucket.

The left sidebar shows the 'Amazon S3' navigation path: Amazon S3 > Buckets > cse-projectdataviz. It also lists 'General purpose buckets' and 'Storage Lens' options.

The main content area displays the 'Objects (19)' section. A table lists four objects:

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

Actions buttons include: Copy S3 URI, Copy URL, Download, Open, Delete, Actions, Create folder, and Upload.

Footer: © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences