Mini Project- BDA

Anoushka Shinde Fall sem'24

Abstract

This paper presents the design and implementation of a comprehensive big data pipeline leveraging Amazon Web Services (AWS) for data storage and visualization, complemented by a Linux-based PySpark environment for distributed data processing. The implementation utilizes a substantial Flipkart e-commerce dataset containing over 5.7 million records, demonstrating the pipeline's capability to handle large-scale data processing requirements in a real-world context.

Introduction

E-commerce platforms generate massive amounts of data that require sophisticated processing and analysis capabilities. This project addresses the challenges of handling such large-scale data by implementing a robust big data pipeline. The solution combines the scalability of AWS services with the distributed processing power of PySpark, creating a comprehensive system for e-commerce data analysis.

The project utilizes a pre-crawled Flipkart dataset found on Kaggle, representing one of India's largest e-commerce platforms. The dataset encompasses:

- Size: 8.75 MB
- Records: 5.7+ million entries
- Format: Structured data with clear categorical and numerical attributes

The dataset contains crucial e-commerce metrics and attributes:

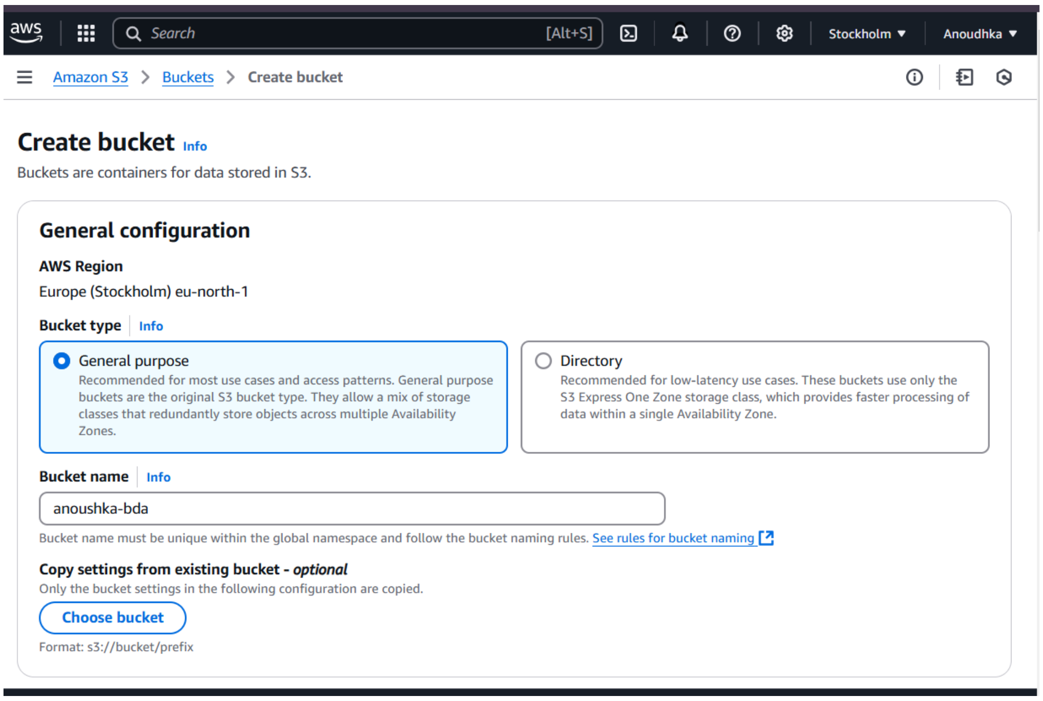
- **Product Identification**
  - _id: Unique identifier assigned to each product
  - brand: Product manufacturer or brand name
  - category/sub_category: Hierarchical product classification
- **Pricing Information**
  - actual_price: Original product price
  - selling_price: Discounted retail price
  - discount: Percentage reduction from original price
- **Performance Metrics**
  - average_rating: Aggregate customer satisfaction score
  - out_of_stock: Inventory availability indicator
  - seller: Vendor identification

This paper details the technical implementation, challenges encountered, and solutions developed throughout the project lifecycle, providing insights into building robust big data solutions for e-commerce applications.

Methodology

The initial phase involved setting up the AWS infrastructure for data storage:

- Created an S3 bucket for storing both raw and processed data
- Manually uploaded the raw dataset to the designated S3 bucket
- Established an Amazon Linux EC2 instance for processing
- Configured secure access using .pem to ppk conversion via PuttyGen
- Connected to EC2 instance using public IPv4 address through Putty



Create bucket Info

Buckets are containers for data stored in S3.

General configuration

AWS Region

Europe (Stockholm) eu-north-1

Bucket type Info

(●) General purpose
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

( ) Directory
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name Info

anoushka-bda

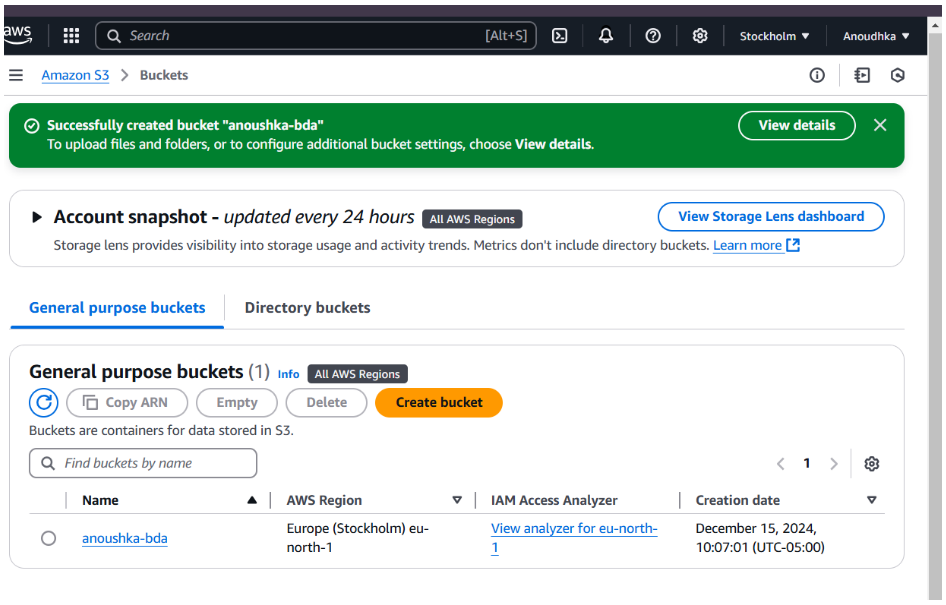
Bucket name must be unique within the global namespace and follow the bucket naming rules. See rules for bucket naming

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Choose bucket

Format: s3://bucket/prefix



Successfully created bucket "anoushka-bda"

To upload files and folders, or to configure additional bucket settings, choose **View details**.

View details

Account snapshot - *updated every 24 hours* All AWS Regions

View Storage Lens dashboard

Storage lens provides visibility into storage usage and activity trends. Metrics don't include directory buckets. Learn more

General purpose buckets | Directory buckets

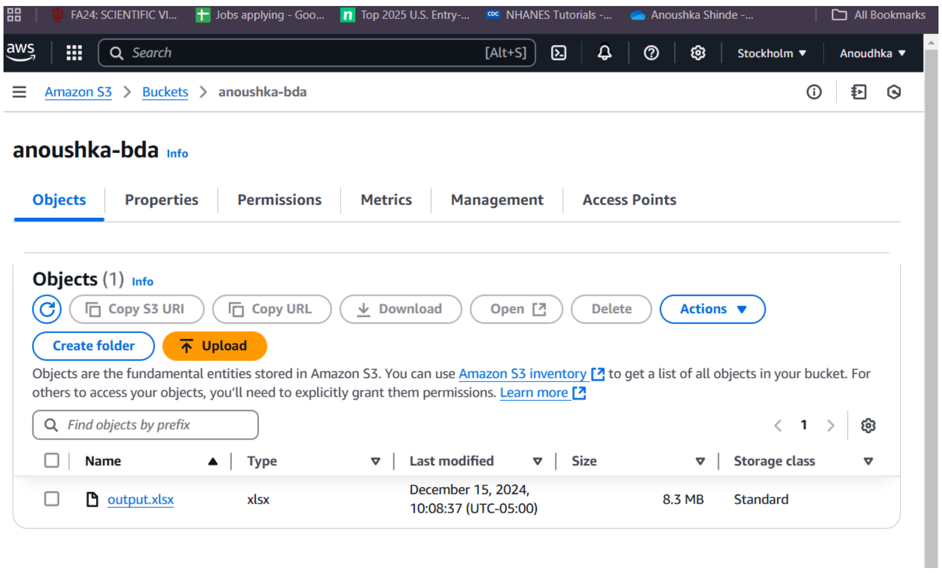
General purpose buckets (1) Info All AWS Regions

Copy ARN | Empty | Delete | Create bucket

Buckets are containers for data stored in S3.

Find buckets by name

Name	AWS Region	IAM Access Analyzer	Creation date
anoushka-bda	Europe (Stockholm) eu-north-1	View analyzer for eu-north-1	December 15, 2024, 10:07:01 (UTC-05:00)



Amazon S3 > Buckets > anoushka-bda

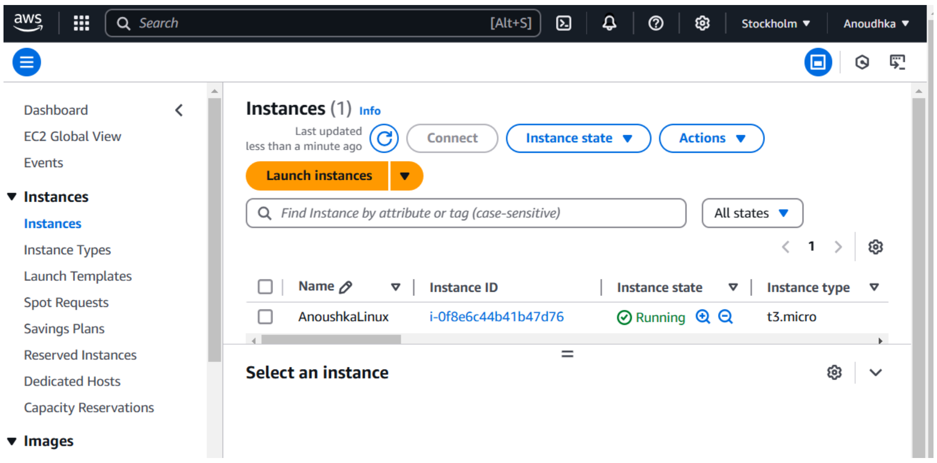
anoushka-bda

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (1)

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

Name	Type	Last modified	Size	Storage class
output.xlsx	xlsx	December 15, 2024, 10:08:37 (UTC-05:00)	8.3 MB	Standard

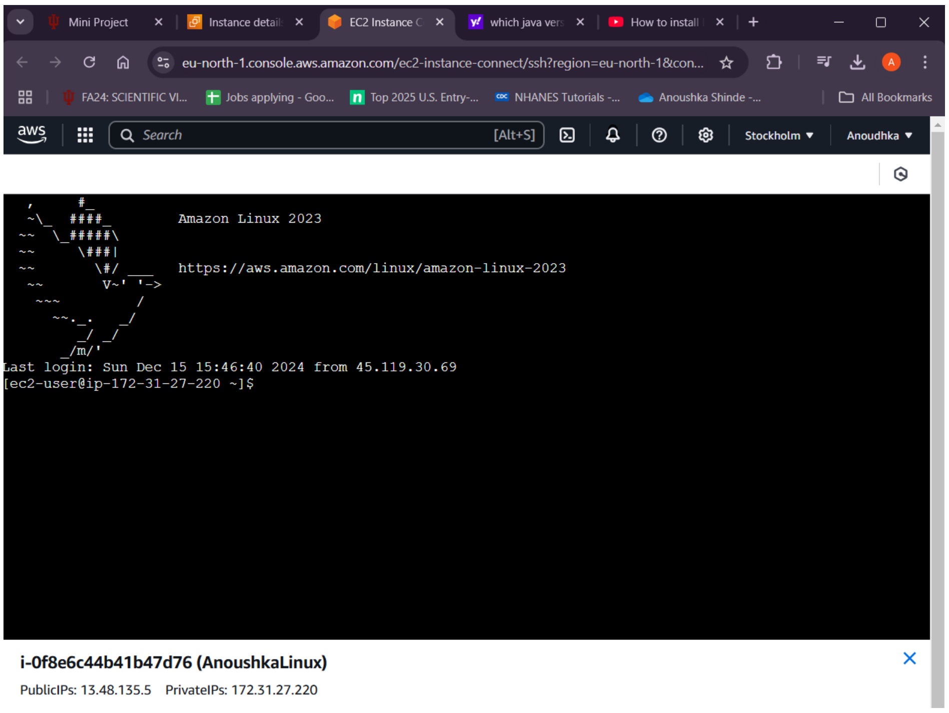


Instances (1)

Last updated less than a minute ago

Name	Instance ID	Instance state	Instance type
AnoushkaLinux	i-0f8e6c44b41b47d76	Running	t3.micro

Select an instance

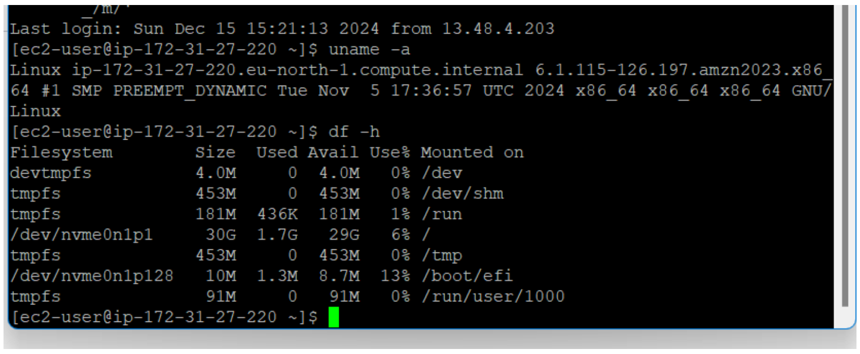


eu-north-1.console.aws.amazon.com/ec2-instance-connect/ssh?region=eu-north-1&con...

```
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Sun Dec 15 15:46:40 2024 from 45.119.30.69
[ec2-user@ip-172-31-27-220 ~]$
```

i-0f8e6c44b41b47d76 (AnoushkaLinux)

PublicIPs: 13.48.135.5 PrivateIPs: 172.31.27.220



```
        _/m/'
Last login: Sun Dec 15 15:21:13 2024 from 13.48.4.203
[ec2-user@ip-172-31-27-220 ~]$ uname -a
Linux ip-172-31-27-220.eu-north-1.compute.internal 6.1.115-126.197.amzn2023.x86_
64 #1 SMP PREEMPT_DYNAMIC Tue Nov  5 17:36:57 UTC 2024 x86_64 x86_64 x86_64 GNU/
Linux
[ec2-user@ip-172-31-27-220 ~]$ df -h
Filesystem        Size  Used Avail Use% Mounted on
devtmpfs          4.0M     0  4.0M   0% /dev
tmpfs             453M     0  453M   0% /dev/shm
tmpfs             181M  436K  181M   1% /run
/dev/nvme0n1p1     30G  1.7G   29G   6% /
tmpfs             453M     0  453M   0% /tmp
/dev/nvme0n1p128   10M  1.3M  8.7M  13% /boot/efi
tmpfs              91M     0   91M   0% /run/user/1000
[ec2-user@ip-172-31-27-220 ~]$ 
```

The PySpark environment setup encountered and resolved several technical challenges:

rm -rf ~/temp_install

mkdir -p ~/temp_install

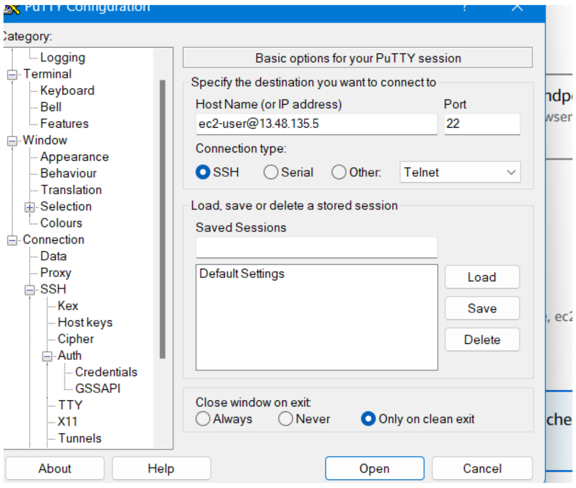
chmod 777 ~/temp_install

TMPDIR=~/temp_install pip3 install --no-cache-dir 'pyspark==3.2.0'

This configuration addressed initial space constraints by:

- Cleaning up the temporary directory
- Creating a new directory with explicit permissions
- Installing a specific PySpark version with temporary directory specification

I also set up Putty for accessing it through my desktop. I did this by converting the .pem file to a .ppk file using PuttyGen and then accessing it through Putty.



Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): ec2-user@13.48.135.5

Port: 22

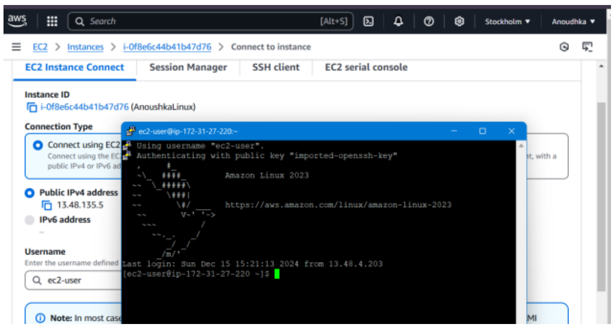
Connection type: [✓ SSH] Serial Other: Telnet

Load, save or delete a stored session

Saved Sessions

Default Settings

Close window on exit: Always Never [✓ Only on clean exit]



Instance ID: i-0f8e6c44b41b47d76 (AnoushkaLinux)

Connection Type: [✓ Connect using EC2]

[✓ Public IPv4 address] 13.48.135.5

IPv6 address

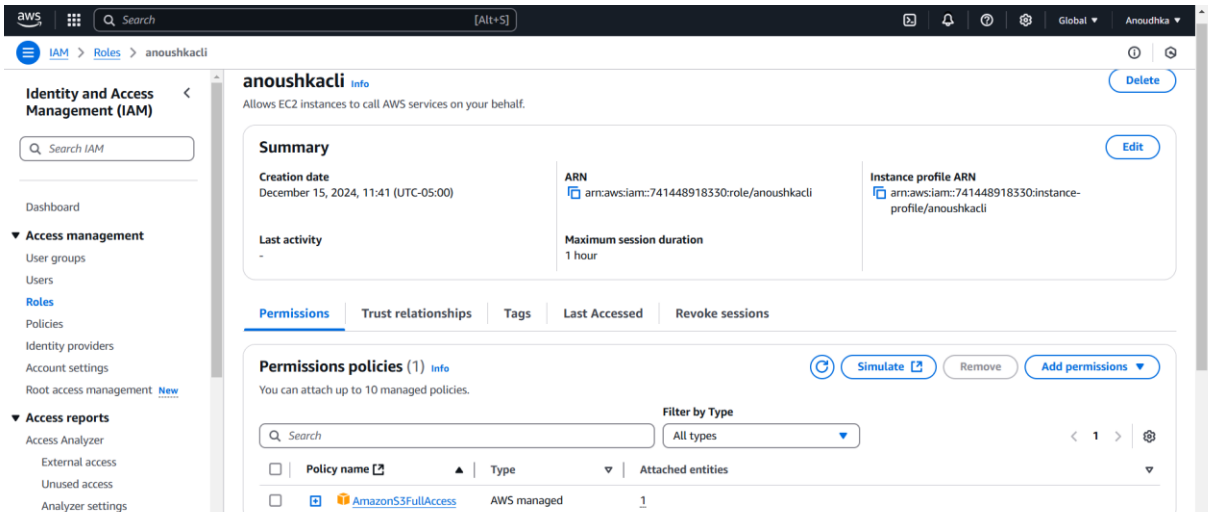
Username: ec2-user

ec2-user@ip-172-31-27-220:~

```
Using username "ec2-user".
Authenticating with public key "imported-openssh-key"
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Sun Dec 15 15:21:13 2024 from 13.48.4.203
[ec2-user@ip-172-31-27-220 ~]$ 
```

Another issue I faced is that I tried to configure aws using aws configure command. And put in random values. Then I forgot about it and created an IAM role. I momentarily forgot that I had listed values and so to fix this I put –

```
aws configure list
rm -rf ~/.aws/credentials ~/.aws/config
```



anoushkacli Info

Allows EC2 instances to call AWS services on your behalf.

Summary

Creation date: December 15, 2024, 11:41 (UTC-05:00)

ARN: arn:aws:iam::741448918330:role/anoushkacli

Instance profile ARN: arn:aws:iam::741448918330:instance-profile/anoushkacli

Last activity: -

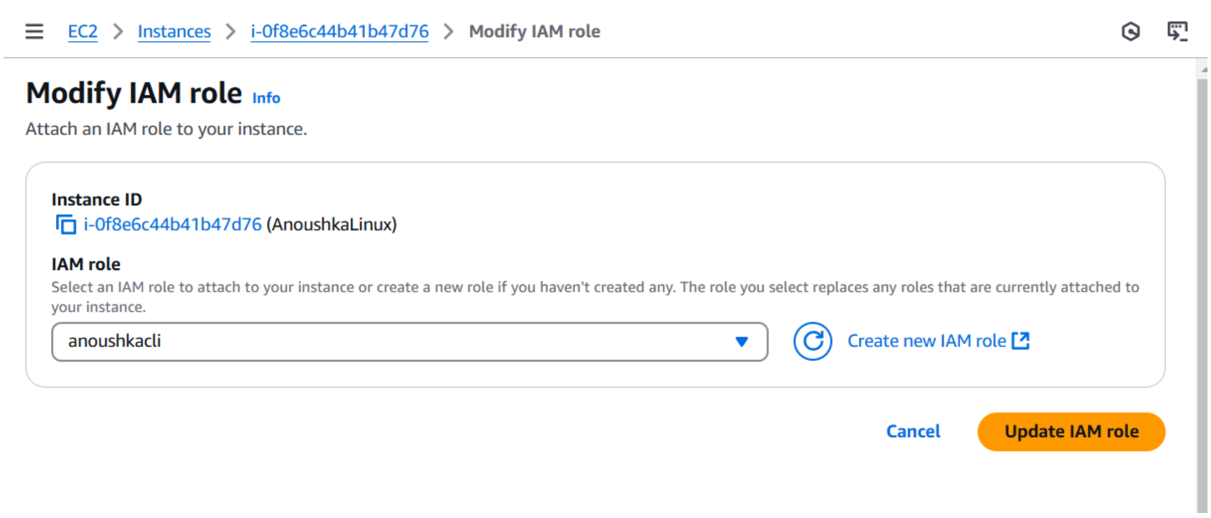
Maximum session duration: 1 hour

Permissions policies (1) Info

You can attach up to 10 managed policies.

Filter by Type: All types

Policy name	Type	Attached entities
AmazonS3FullAccess	AWS managed	1



Modify IAM role Info

Attach an IAM role to your instance.

Instance ID

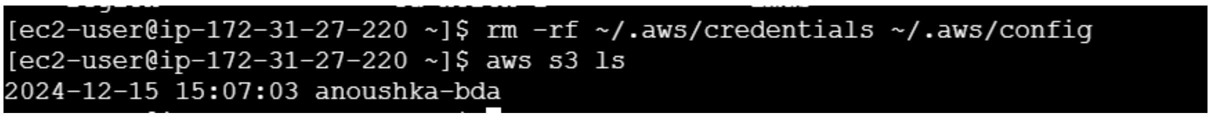
i-0f8e6c44b41b47d76 (AnoushkaLinux)

IAM role

Select an IAM role to attach to your instance or create a new role if you haven't created any. The role you select replaces any roles that are currently attached to your instance.

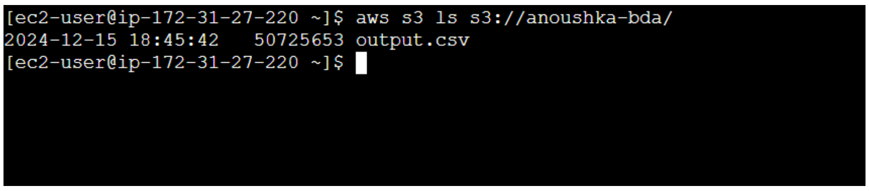
anoushkacli

I then verified the s3 access from ec2.



```
[ec2-user@ip-172-31-27-220 ~]$ rm -rf ~/.aws/credentials ~/.aws/config
[ec2-user@ip-172-31-27-220 ~]$ aws s3 ls
2024-12-15 15:07:03 anoushka-bda
```

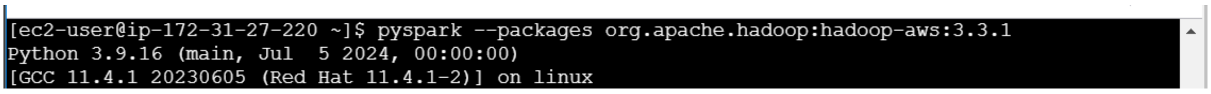
Run the following AWS CLI command to list the bucket:



```
[ec2-user@ip-172-31-27-220 ~]$ aws s3 ls s3://anoushka-bda/
2024-12-15 18:45:42   50725653 output.csv
[ec2-user@ip-172-31-27-220 ~]$ 
```

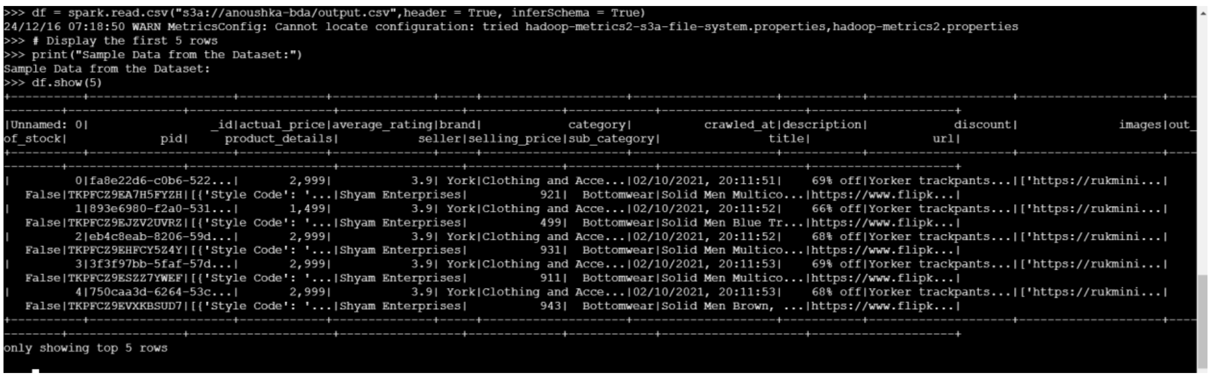
PySpark was initialized with necessary AWS support:

pyspark --packages org.apache.hadoop:hadoop-aws:3.3.1

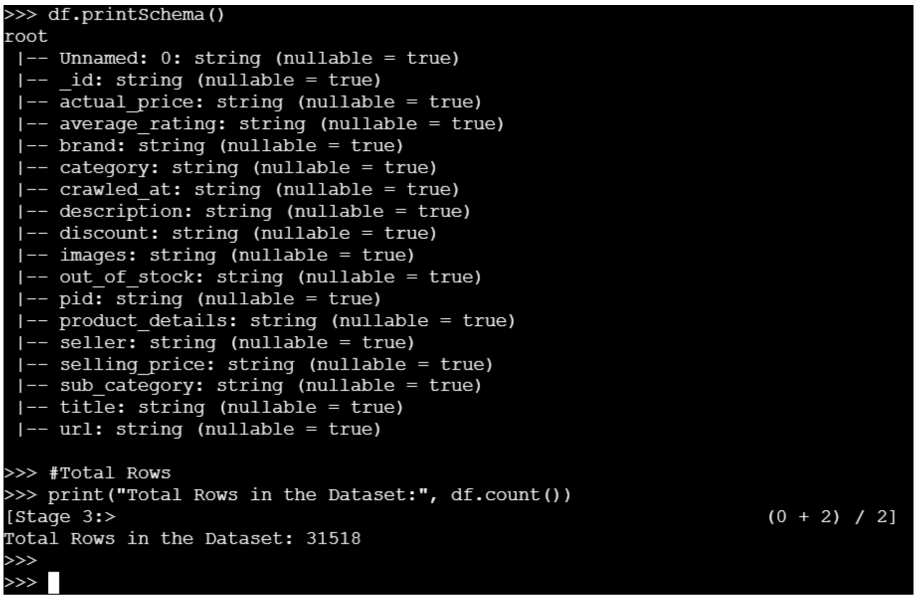


```
[ec2-user@ip-172-31-27-220 ~]$ pyspark --packages org.apache.hadoop:hadoop-aws:3.3.1
Python 3.9.16 (main, Jul  5 2024, 00:00:00)
[GCC 11.4.1 20230605 (Red Hat 11.4.1-2)] on linux
```

And then I inspected the dataset for common information.



```
>>> df = spark.read.csv("s3a://anoushka-bda/output.csv",header = True, inferSchema = True)
24/12/16 07:18:50 WARN MetricsConfig: Cannot locate configuration: tried hadoop-metrics2-s3a-file-system.properties,hadoop-metrics2.properties
>>> # Display the first 5 rows
>>> print("Sample Data from the Dataset:")
Sample Data from the Dataset:
>>> df.show(5)
+----------+--------------------+------------+--------------+-----+--------------------+--------------------+-----------+--------------------+--------------------+----
--------+--------------------+--------------------+------------------+------------+--------------------+--------------------+
|Unnamed: 0|                 _id|actual_price|average_rating|brand|            category|          crawled_at|description|            discount|              images|out_
of_stock|                 pid|     product_details|            seller|selling_price|sub_category|               title|                 url|
+----------+--------------------+------------+--------------+-----+--------------------+--------------------+-----------+--------------------+--------------------+----
--------+--------------------+--------------------+------------------+------------+--------------------+--------------------+
|         0|fa8e22d6-c0b6-522...|       2,999|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:51|    69% off|Yorker trackpants...|['https://rukmini...|
   False|TKPFCZ9EA7H5FYZH|[{'Style Code': '...|Shyam Enterprises|         921|  Bottomwear|Solid Men Multico...|https://www.flipk...|
|         1|893e6980-f2a0-531...|       1,499|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:52|    66% off|Yorker trackpants...|['https://rukmini...|
   False|TKPFCZ9EJZV2UVRZ|[{'Style Code': '...|Shyam Enterprises|         499|  Bottomwear|Solid Men Blue Tr...|https://www.flipk...|
|         2|eb4c8eab-8206-59d...|       2,999|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:52|    68% off|Yorker trackpants...|['https://rukmini...|
   False|TKPFCZ9EHFCY5Z4Y|[{'Style Code': '...|Shyam Enterprises|         931|  Bottomwear|Solid Men Multico...|https://www.flipk...|
|         3|3f3f97bb-5faf-57d...|       2,999|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:53|    69% off|Yorker trackpants...|['https://rukmini...|
   False|TKPFCZ9ESZZ7YWEF|[{'Style Code': '...|Shyam Enterprises|         911|  Bottomwear|Solid Men Multico...|https://www.flipk...|
|         4|750caa3d-6264-53c...|       2,999|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:53|    68% off|Yorker trackpants...|['https://rukmini...|
   False|TKPFCZ9EVXKBSUD7|[{'Style Code': '...|Shyam Enterprises|         943|  Bottomwear|Solid Men Brown, ...|https://www.flipk...|
+----------+--------------------+------------+--------------+-----+--------------------+--------------------+-----------+--------------------+--------------------+----
--------+--------------------+--------------------+------------------+------------+--------------------+--------------------+
only showing top 5 rows
```



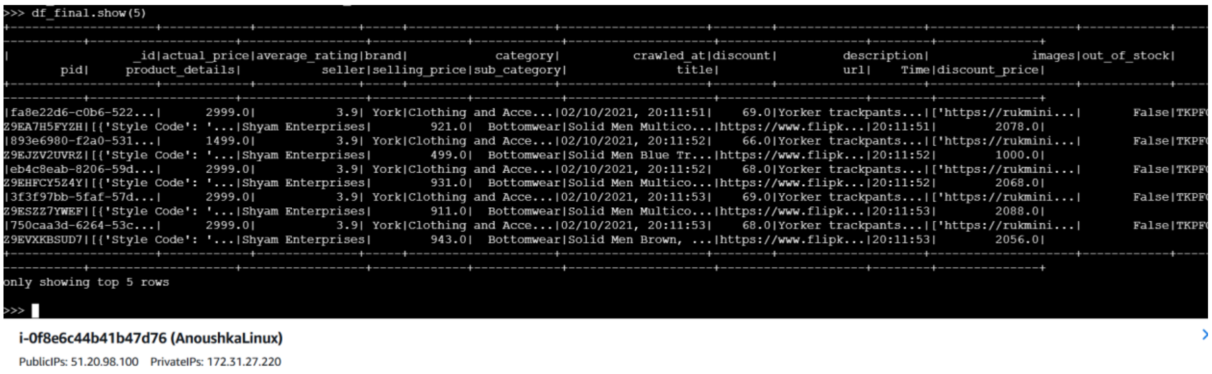
```
>>> df.printSchema()
root
 |-- Unnamed: 0: string (nullable = true)
 |-- _id: string (nullable = true)
 |-- actual_price: string (nullable = true)
 |-- average_rating: string (nullable = true)
 |-- brand: string (nullable = true)
 |-- category: string (nullable = true)
 |-- crawled_at: string (nullable = true)
 |-- description: string (nullable = true)
 |-- discount: string (nullable = true)
 |-- images: string (nullable = true)
 |-- out_of_stock: string (nullable = true)
 |-- pid: string (nullable = true)
 |-- product_details: string (nullable = true)
 |-- seller: string (nullable = true)
 |-- selling_price: string (nullable = true)
 |-- sub_category: string (nullable = true)
 |-- title: string (nullable = true)
 |-- url: string (nullable = true)

>>> #Total Rows
>>> print("Total Rows in the Dataset:", df.count())
[Stage 3:>                                                          (0 + 2) / 2]
Total Rows in the Dataset: 31518
>>>
>>>
```

i-0f8e6c44b41b47d76 (AnoushkaLinux)

The data cleaning process involved several key steps:

1. Null Value Management
    - o Implemented systematic null value removal
2. Data Type Conversion
    - o Converted actual_price and selling_price to numeric format
    - o Standardized discount column by removing 'off' suffix
3. Column Standardization
    - o Corrected column name mismatches between description and discount
    - o Implemented proper column renaming

4. Feature Engineering New columns created:
   - Crawled_time: Timestamp information
   - discounted_price: How much price was discounted



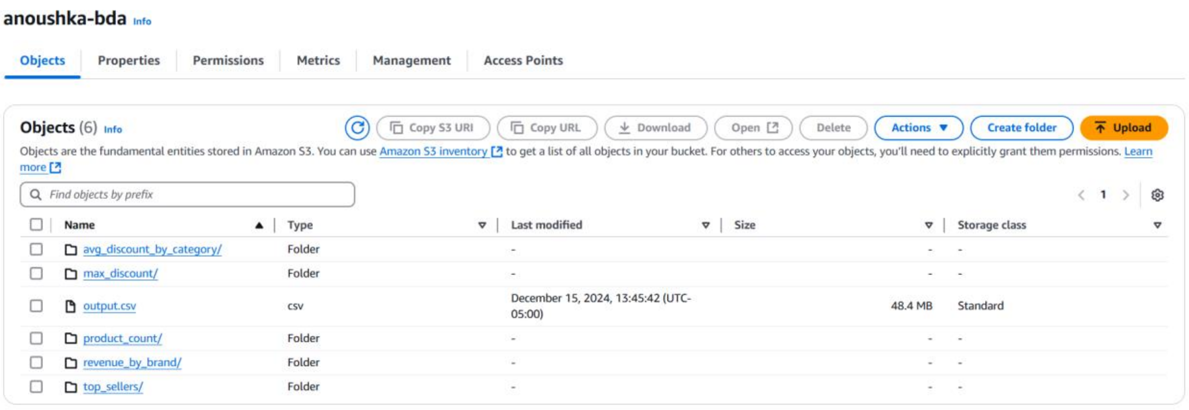
```
>>> df_final.show(5)
+--------------------+------------+--------------+-----+--------------------+--------------------+--------+--------------------+--------------------+------------+----
-----------+--------------------+-----------------+-------------+------------+--------------------+--------------------+--------+--------------+
|                 _id|actual_price|average_rating|brand|            category|          crawled_at|discount|         description|              images|out_of_stock|
        pid|     product_details|           seller|selling_price|sub_category|               title|                 url|    Time|discount_price|
+--------------------+------------+--------------+-----+--------------------+--------------------+--------+--------------------+--------------------+------------+----
-----------+--------------------+-----------------+-------------+------------+--------------------+--------------------+--------+--------------+
|fa8e22d6-c0b6-522...|      2999.0|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:51|    69.0|Yorker trackpants...|['https://rukmini...|       False|TKPF
Z9EA7H5FYZH|[{'Style Code': '...|Shyam Enterprises|        921.0|  Bottomwear|Solid Men Multico...|https://www.flipk...|20:11:51|        2078.0|
|893e6980-f2a0-531...|      1499.0|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:52|    66.0|Yorker trackpants...|['https://rukmini...|       False|TKPF
Z9EJZV2UVRZ|[{'Style Code': '...|Shyam Enterprises|        499.0|  Bottomwear|Solid Men Blue Tr...|https://www.flipk...|20:11:52|        1000.0|
|eb4c8eab-8206-59d...|      2999.0|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:52|    68.0|Yorker trackpants...|['https://rukmini...|       False|TKPF
Z9EHFCY5Z4Y|[{'Style Code': '...|Shyam Enterprises|        931.0|  Bottomwear|Solid Men Multico...|https://www.flipk...|20:11:52|        2068.0|
|3f3f97bb-5faf-57d...|      2999.0|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:53|    69.0|Yorker trackpants...|['https://rukmini...|       False|TKPF
Z9ESZZ7YWEF|[{'Style Code': '...|Shyam Enterprises|        911.0|  Bottomwear|Solid Men Multico...|https://www.flipk...|20:11:53|        2088.0|
|750caa3d-6264-53c...|      2999.0|           3.9| York|Clothing and Acce...|02/10/2021, 20:11:53|    68.0|Yorker trackpants...|['https://rukmini...|       False|TKPF
Z9EVXKBSUD7|[{'Style Code': '...|Shyam Enterprises|        943.0|  Bottomwear|Solid Men Brown, ...|https://www.flipk...|20:11:53|        2056.0|
+--------------------+------------+--------------+-----+--------------------+--------------------+--------+--------------------+--------------------+------------+----
-----------+--------------------+-----------------+-------------+------------+--------------------+--------------------+--------+--------------+

only showing top 5 rows

>>> 
```

i-0f8e6c44b41b47d76 (AnoushkaLinux)

PublicIPs: 51.20.98.100 PrivateIPs: 172.31.27.220

The project implemented two comprehensive aggregation approaches, with results stored directly in the S3 bucket:

1. Python-based Aggregations: The Python implementation focused on key business metrics and was executed through a dedicated .py script. The following aggregations were performed:
   - Revenue by brand: Calculated total revenue generation across different brands
   - Average discount by category: Analyzed discount patterns within product categories
   - Top sellers analysis: Identified and ranked the most successful sellers
   - Product count metrics: Quantified product distribution
   - Maximum discount price calculations: Determined highest discount offerings



anoushka-bda Info

Objects (6) Info

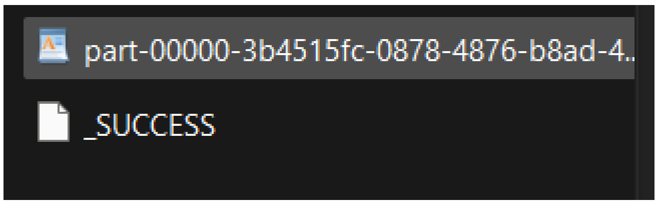
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Find objects by prefix

Name	Type	Last modified	Size	Storage class
avg_discount_by_category/	Folder	-	-	-
max_discount/	Folder	-	-	-
output.csv	csv	December 15, 2024, 13:45:42 (UTC-05:00)	48.4 MB	Standard
product_count/	Folder	-	-	-
revenue_by_brand/	Folder	-	-	-
top_sellers/	Folder	-	-	-

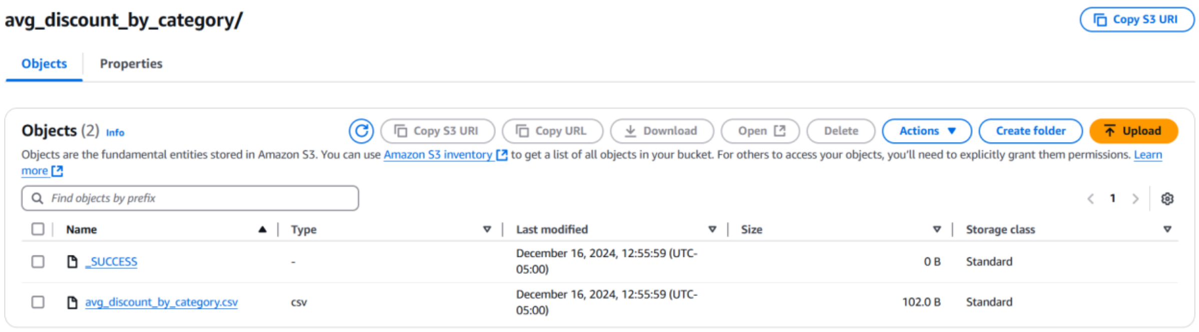
Implementation Note: The output generation process created folders containing two files:

- A success indicator file
- A CSV file containing the aggregation results


part-00000-3b4515fc-0878-4876-b8ad-4..
_SUCCESS

Post-processing steps included:

- Manual download of generated files from S3
- Systematic renaming of CSV files to reflect their specific tasks
- Organization into a dedicated 'python' folder within the output directory
- Re-upload of the organized files to the S3 bucket



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

Name	Type	Last modified	Size	Storage class
_SUCCESS	-	December 16, 2024, 12:55:59 (UTC-05:00)	0 B	Standard
avg_discount_by_category.csv	csv	December 16, 2024, 12:55:59 (UTC-05:00)	102.0 B	Standard

All the outputs can be found here.

Results –

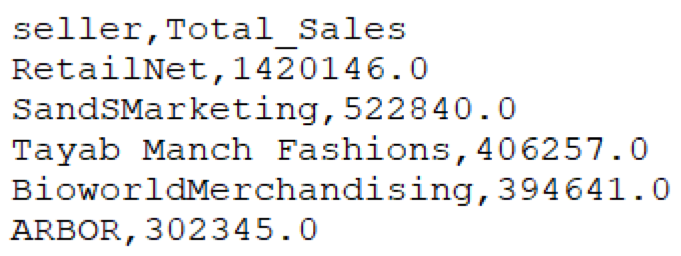
Revenue by brand

```
brand,Total_Revenue
K,14736.0
fas,1495.0
ECKO Unl,522840.0
ertugrul ghazi c,5519.0
SATDEVANGIKHADIBHAND,73026.0
Marca Disa,122597.0
Aeload,749.0
A J STYL,23512.0
Jagdish Garmen,10541.0
NEXXE,8328.0
Gracew,153270.0
Pramukhraj Enterpri,185.0
CupidSto,160296.0
Musk,276.0
ATTIITU,75466.0
The Dry Ca,14838.0
DISCOUNT OUTL,9538.0
Mylifestylebazz,798.0
BS S,769.0
Royalty Retail And Expo,6444.0
Man,2160.0
GLAUBEN PO,1895.0
EYETWIST,2457.0
style acco,3865.0
Vector,145229.0
LA SMI,234.0
U.S. POLO ASS,83391.0
She,349.0
Jack Roy,529.0
ANGI,280.0
Bone,47882.0
NEXT,1136.0
Shoef,218824.0
MARATH,1285.0
Fairdea,20707.0
Clo,13899.0
```

Avg discount by category

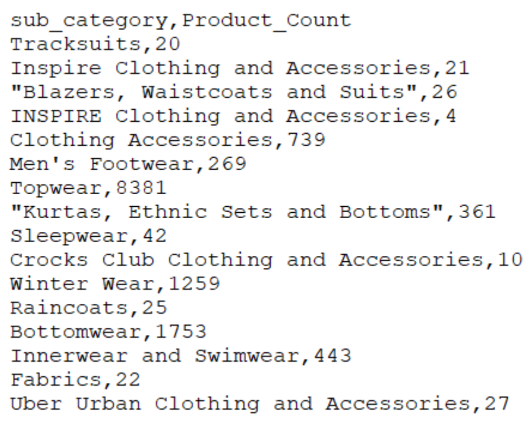
```
category,Avg_Discount_Percentage
Footwear,29.3271375464684
Clothing and Accessories,49.51085052920125
```

Top sellers



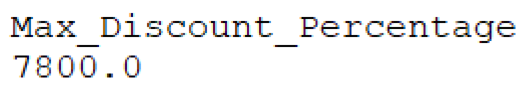
```
seller,Total_Sales
RetailNet,1420146.0
SandSMarketing,522840.0
Tayab Manch Fashions,406257.0
BioworldMerchandising,394641.0
ARBOR,302345.0
```

Product count



```
sub_category,Product_Count
Tracksuits,20
Inspire Clothing and Accessories,21
"Blazers, Waistcoats and Suits",26
INSPIRE Clothing and Accessories,4
Clothing Accessories,739
Men's Footwear,269
Topwear,8381
"Kurtas, Ethnic Sets and Bottoms",361
Sleepwear,42
Crocks Club Clothing and Accessories,10
Winter Wear,1259
Raincoats,25
Bottomwear,1753
Innerwear and Swimwear,443
Fabrics,22
Uber Urban Clothing and Accessories,27
```

Max discount price



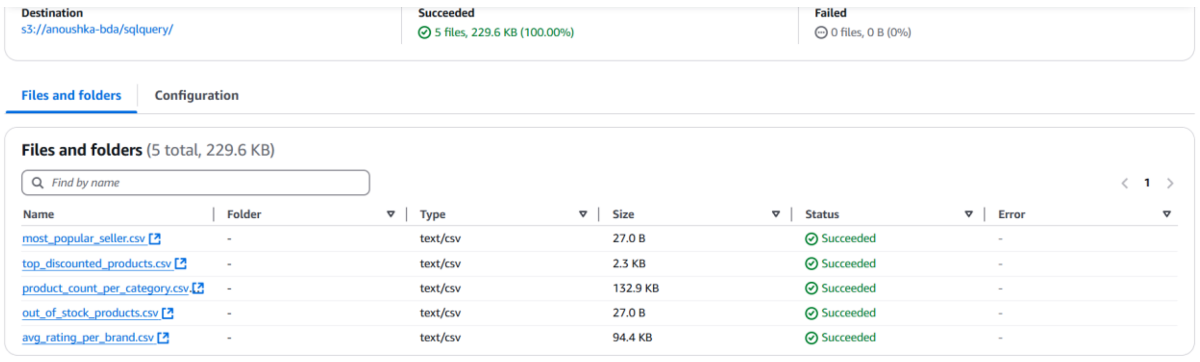
```
Max_Discount_Percentage
7800.0
```

2. SQL-based Aggregations: SQL aggregations were implemented through PySpark's SQL interface, providing different analytical perspectives:

   - Top discounted products: Identification of products with highest discount percentages
   - Average rating per brand: Brand performance analysis based on customer ratings
   - Product count per category: Detailed category-wise inventory analysis
   - Most popular seller analysis: Seller performance metrics and rankings
   - Out-of-stock products: Inventory status tracking and analysis

The SQL implementation was documented in a separate .py file, ensuring reproducibility and maintainability of the analysis pipeline. Each aggregation was designed to provide specific business insights while maintaining efficient processing of the large-scale dataset.

Similarly, I completed the code for sql in pyspark. The entire code has been pasted in a .py file that is present in the attachments.



Destination
s3://anoushka-bda/sqlquery/

Succeeded
5 files, 229.6 KB (100.00%)

Failed
0 files, 0 B (0%)

Files and folders | Configuration

Files and folders (5 total, 229.6 KB)

Find by name

Name	Folder	Type	Size	Status	Error
most_popular_seller.csv	-	text/csv	27.0 B	Succeeded	-
top_discounted_products.csv	-	text/csv	2.3 KB	Succeeded	-
product_count_per_category.csv	-	text/csv	132.9 KB	Succeeded	-
out_of_stock_products.csv	-	text/csv	27.0 B	Succeeded	-
avg_rating_per_brand.csv	-	text/csv	94.4 KB	Succeeded	-

Results –

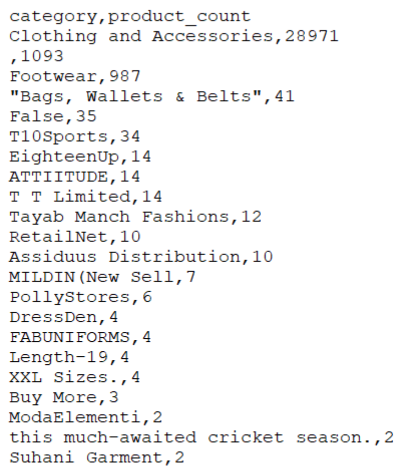
Top Discounted products

```
_id,title,actual_price,selling_price,discount
86d0e25f-ddbe-5807-ac46-070ecd14db48,Printed Men Polo Neck Blue
T-ShirtÂ Â (Pack of 3),5500.0,950.0,
57b8f120-68e2-57b4-a961-12cbef8974d0,"Printed, Striped Men Round
Neck Blue T-Shirt",4999.0,800.0,
8ec86a28-1200-56c6-9dba-4c9e0f6d1fd9,Full Sleeve Solid Men Padded
Jacket,4999.0,995.0,
619f5e66-c5be-50e0-8bf5-a78d44b9b394,Striped Men V Neck
Multicolor T-Shirt,4200.0,799.0,
e8932882-c6ce-5f0d-b0ef-f330c5186e27,Solid Men Round Neck Light
Blue T-ShirtÂ Â (Pack of 5),3995.0,799.0,"Refresh your wardrobe
with the Premium Collection of Cotton Round Neck T-shirts from
Scott International. These t-shirts are made of Pure Cotton and
provide comfort all day. Melange Fabric, Elegant Stitching and
finish, Soft Neck Tape, Comfort Fit makes this product an ideal
daily wear and casual wear. Combine with Trousers or Denims for a
cool Casual look. Combo pack is ideal value for money."
65b654a6-1a14-55c5-b497-56eff5d4066a,"Solid Men Round Neck
Maroon, Blue, Light Blue, Dark Blue, Black T-ShirtÂ Â (Pack of
5)",3995.0,799.0,
9e466d9e-f3ef-5516-8f2e-3000b83b90bd,"Solid Men Round Neck
Maroon, Blue, Light Blue, Dark Blue, Black T-ShirtÂ Â (Pack of
5)",3995.0,799.0,
e926483d-80d3-5c8d-86c5-76ae4b3781f6,Solid Men Round Neck
Multicolor T-ShirtÂ Â (Pack of 5),3995.0,799.0,"Refresh your
wardrobe with the Premium Collection of Cotton Round Neck T-
shirts from Scott International. These t-shirts are made of Pure
Cotton and provide comfort all day. Melange Fabric, Elegant
Stitching and finish, Soft Neck Tape, Comfort Fit makes this
product an ideal daily wear and casual wear. Combine with
Trousers or Denims for a cool Casual look. Combo pack is ideal
```

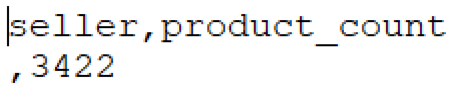
Avg rating per brand

```
brand,avg_rating
62,61.0
MILD,5.0
VARTe,5.0
VIKING.INE,4.640000000000001
SORA,4.609459459459455
COL,4.55
Modest Ci,4.5
PixF,4.468922305764411
Cots,4.456357388316171
GYMBROTHE,4.407142857142857
Royalta,4.4
Ziko,4.4
JUARI BE A GENTLEM,4.4
Neon Ro,4.4
COOL CRU,4.4
LOUIS MONAR,4.3999999999999995
Bofri,4.366666666666667
Mega Sty,4.357142857142857
Zack Fo,4.354651162790697
Onei,4.302531645569618
Bracev,4.3
U.S. Polo Ass,4.3
ud fabr,4.3
Fresh Fe,4.299999999999999
Kt Kun,4.266666666666667
U.S.Polo As,4.23548387096774
Gracew,4.23118811881188
Aeload,4.225
limited colou,4.210169491525429
Truemo,4.204166666666668
ALQI,4.200000000000001
ANGI,4.2
MARATH,4.2
BOYT,4.2
Mohan Reta,4.2
```

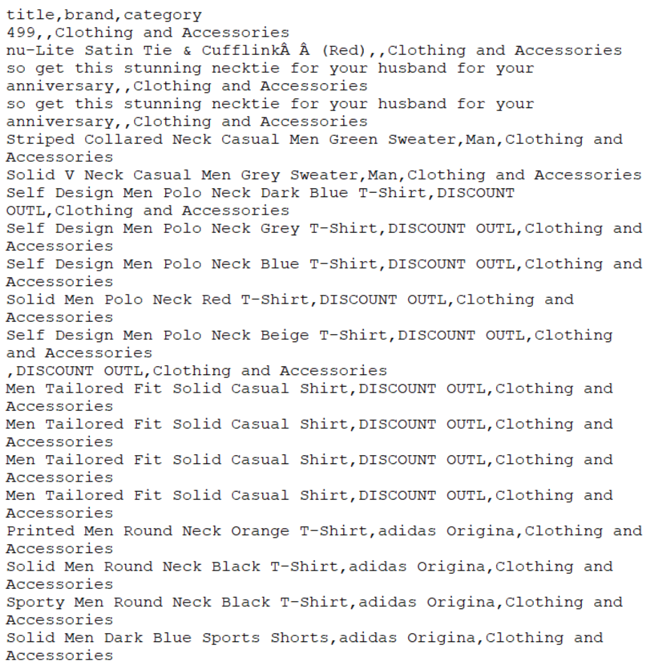
Product count per category


```
category,product_count
Clothing and Accessories,28971
,1093
Footwear,987
"Bags, Wallets & Belts",41
False,35
T10Sports,34
EighteenUp,14
ATTIITUDE,14
T T Limited,14
Tayab Manch Fashions,12
RetailNet,10
Assiduus Distribution,10
MILDIN(New Sell,7
PollyStores,6
DressDen,4
FABUNIFORMS,4
Length-19,4
XXL Sizes.,4
Buy More,3
ModaElementi,2
this much-awaited cricket season.,2
Suhani Garment,2
```

Most popular seller


```
seller,product_count
,3422
```

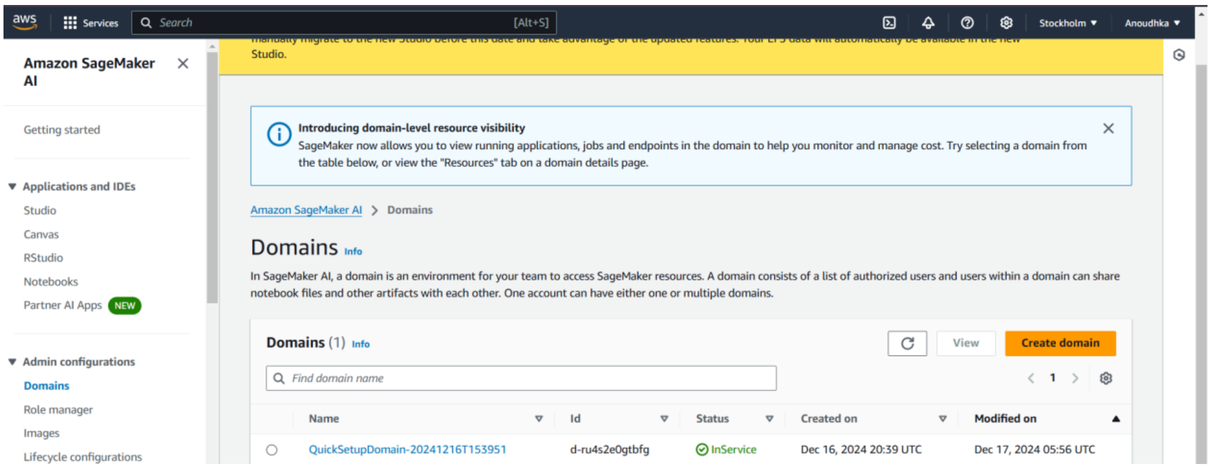
Out of stock products


```
title,brand,category
499,,Clothing and Accessories
nu-Lite Satin Tie & CufflinkÂ Â (Red),,Clothing and Accessories
so get this stunning necktie for your husband for your
anniversary,,Clothing and Accessories
so get this stunning necktie for your husband for your
anniversary,,Clothing and Accessories
Striped Collared Neck Casual Men Green Sweater,Man,Clothing and
Accessories
Solid V Neck Casual Men Grey Sweater,Man,Clothing and Accessories
Self Design Men Polo Neck Dark Blue T-Shirt,DISCOUNT
OUTL,Clothing and Accessories
Self Design Men Polo Neck Grey T-Shirt,DISCOUNT OUTL,Clothing and
Accessories
Self Design Men Polo Neck Blue T-Shirt,DISCOUNT OUTL,Clothing and
Accessories
Solid Men Polo Neck Red T-Shirt,DISCOUNT OUTL,Clothing and
Accessories
Self Design Men Polo Neck Beige T-Shirt,DISCOUNT OUTL,Clothing
and Accessories
,DISCOUNT OUTL,Clothing and Accessories
Men Tailored Fit Solid Casual Shirt,DISCOUNT OUTL,Clothing and
Accessories
Men Tailored Fit Solid Casual Shirt,DISCOUNT OUTL,Clothing and
Accessories
Men Tailored Fit Solid Casual Shirt,DISCOUNT OUTL,Clothing and
Accessories
Men Tailored Fit Solid Casual Shirt,DISCOUNT OUTL,Clothing and
Accessories
Printed Men Round Neck Orange T-Shirt,adidas Origina,Clothing and
Accessories
Solid Men Round Neck Black T-Shirt,adidas Origina,Clothing and
Accessories
Sporty Men Round Neck Black T-Shirt,adidas Origina,Clothing and
Accessories
Solid Men Dark Blue Sports Shorts,adidas Origina,Clothing and
Accessories
```

The project leveraged AWS SageMaker Autopilot for implementing machine learning capabilities, focusing on automated model training and evaluation without direct code intervention. This approach enabled efficient model development while maintaining high standards of accuracy and reliability.The initial setup phase involved several critical steps to ensure proper integration with AWS services:

1. Domain Configuration
    - A new SageMaker domain was established to provide a controlled environment for machine learning operations

- The environment was configured with a specialized IAM role to manage necessary permissions and access controls
- Security protocols were implemented to ensure safe data handling and model training


Studio.

Introducing domain-level resource visibility
SageMaker now allows you to view running applications, jobs and endpoints in the domain to help you monitor and manage cost. Try selecting a domain from the table below, or view the "Resources" tab on a domain details page.

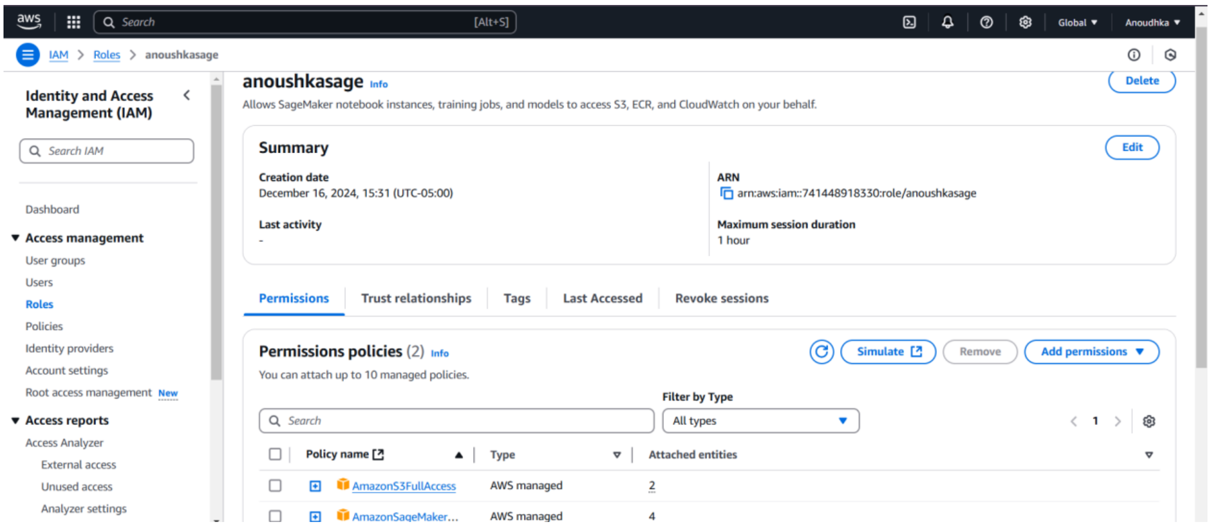
Amazon SageMaker AI > Domains

Domains Info

In SageMaker AI, a domain is an environment for your team to access SageMaker resources. A domain consists of a list of authorized users and users within a domain can share notebook files and other artifacts with each other. One account can have either one or multiple domains.

Domains (1) Info

Name	Id	Status	Created on	Modified on
QuickSetupDomain-20241216T153951	d-ru4s2e0gtbfg	InService	Dec 16, 2024 20:39 UTC	Dec 17, 2024 05:56 UTC


IAM > Roles > anoushkasage

anoushkasage Info

Allows SageMaker notebook instances, training jobs, and models to access S3, ECR, and CloudWatch on your behalf.

Summary

Creation date
December 16, 2024, 15:31 (UTC-05:00)

ARN
arn:aws:iam::741448918330:role/anoushkasage

Last activity
-

Maximum session duration
1 hour

Permissions policies (2) Info

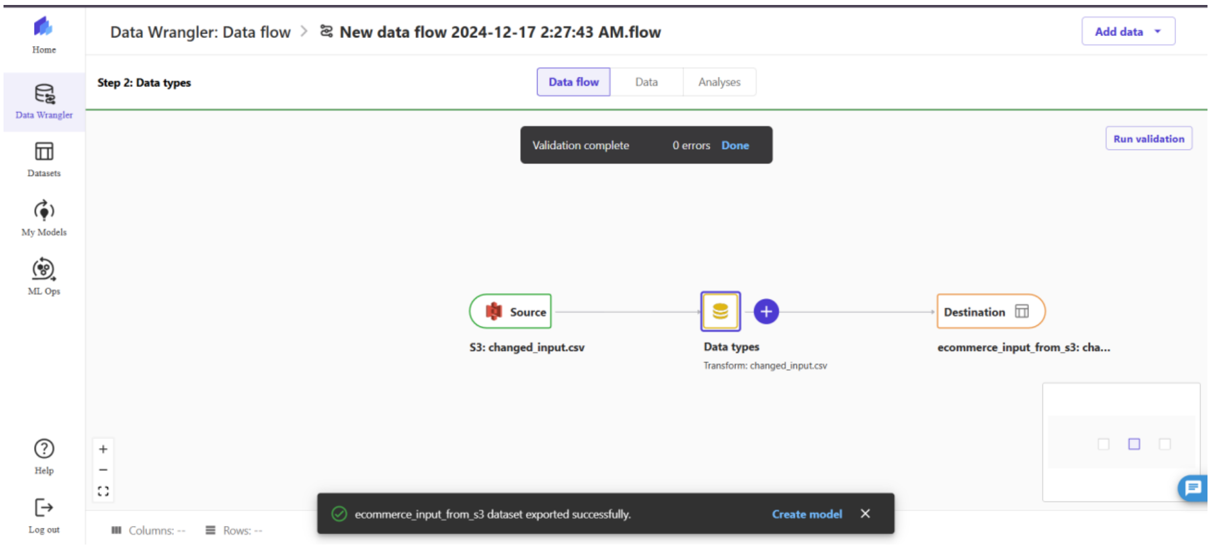
You can attach up to 10 managed policies.

Filter by Type: All types

Policy name	Type	Attached entities
AmazonS3FullAccess	AWS managed	2
AmazonSageMaker...	AWS managed	4

The data integration process focused on seamlessly connecting the processed dataset with SageMaker Autopilot:

1. Data Import Process
   - Established a direct connection to the processed dataset stored in the S3 bucket
   - Implemented automated data loading procedures to ensure efficient data transfer
   - Verified data integrity throughout the import process



Data Wrangler: Data flow > New data flow 2024-12-17 2:27:43 AM.flow

Step 2: Data types

Validation complete 0 errors Done

S3: changed_input.csv

Data types
Transform: changed_input.csv

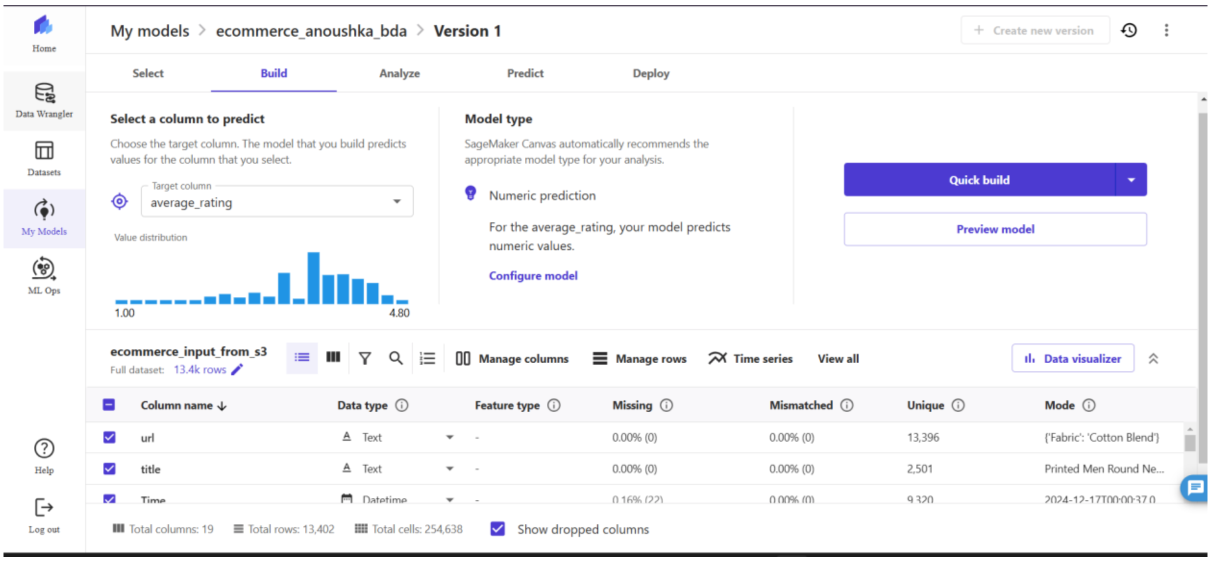
ecommerce_input_from_s3: cha...

ecommerce_input_from_s3 dataset exported successfully. Create model

Columns: -- Rows: --

The automated machine learning process was executed through a structured approach:

1. Model Training Configuration
    - Carefully selected appropriate target variables for prediction
    - Configured the AutoML process to explore multiple model architectures
    - Established evaluation criteria for model performance assessment



My models > ecommerce_anoushka_bda > Version 1

Select a column to predict

Choose the target column. The model that you build predicts values for the column that you select.

Target column: average_rating

Value distribution

Model type

SageMaker Canvas automatically recommends the appropriate model type for your analysis.

Numeric prediction

For the average_rating, your model predicts numeric values.

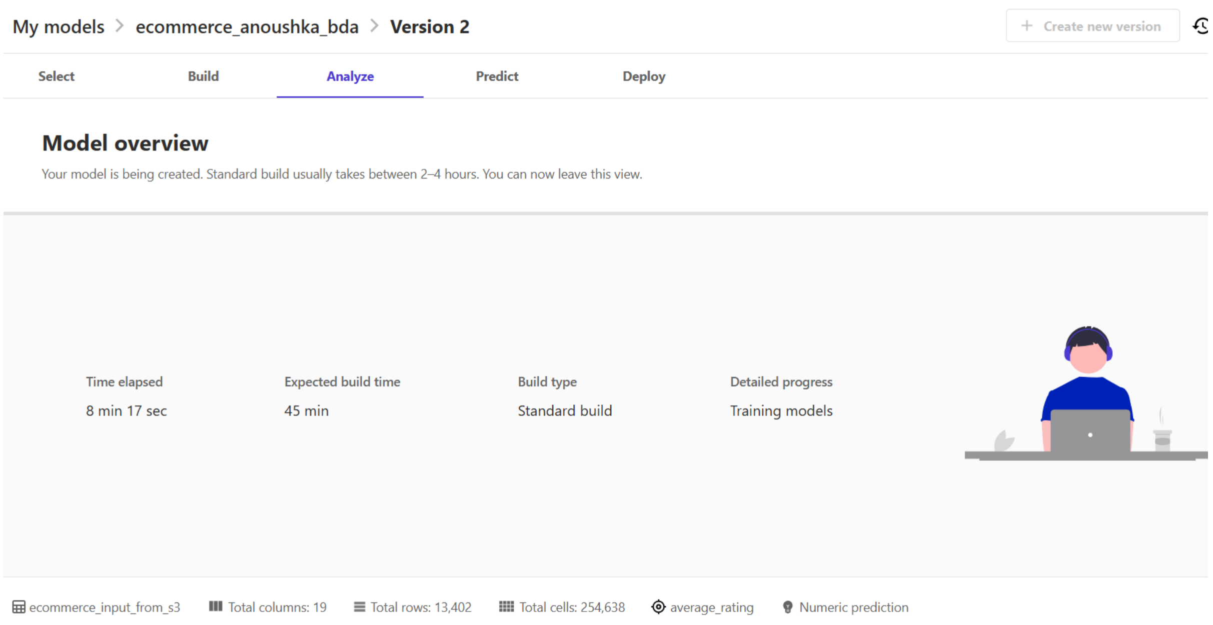
Configure model

ecommerce_input_from_s3
Full dataset: 13.4k rows

Column name	Data type	Feature type	Missing	Mismatched	Unique	Mode
url	Text	-	0.00% (0)	0.00% (0)	13,396	{'Fabric': 'Cotton Blend'}
title	Text	-	0.00% (0)	0.00% (0)	2,501	Printed Men Round Ne...
Time	Datetime	-	0.16% (22)	0.00% (0)	9,320	2024-12-17T00:00:37.0

Total columns: 19 Total rows: 13,402 Total cells: 254,638 Show dropped columns

2. Training Execution
    - Initiated the automated model training process
    - Monitored the training progress across multiple model iterations
    - Documented computational resource utilization and training duration



My models > ecommerce_anoushka_bda > Version 2

+ Create new version

Select | Build | Analyze | Predict | Deploy

Model overview

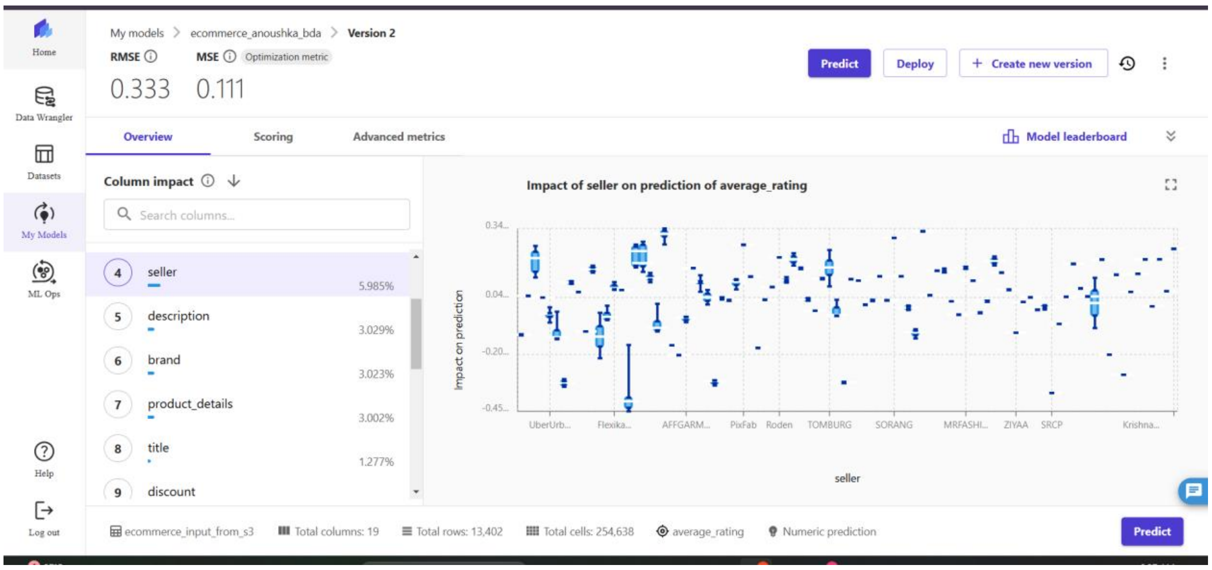
Your model is being created. Standard build usually takes between 2–4 hours. You can now leave this view.

Time elapsed	Expected build time	Build type	Detailed progress
8 min 17 sec	45 min	Standard build	Training models

ecommerce_input_from_s3 | Total columns: 19 | Total rows: 13,402 | Total cells: 254,638 | average_rating | Numeric prediction

3. Model Evaluation

   - Analyzed the model leaderboard to identify top-performing algorithms
   - Evaluated comprehensive performance metrics across all trained models
   - Documented comparative analysis of different model architectures



My models > ecommerce_anoushka_bda > Version 2

RMSE 0.333 | MSE (Optimization metric) 0.111

Predict | Deploy | + Create new version

Overview | Scoring | Advanced metrics | Model leaderboard

Column impact

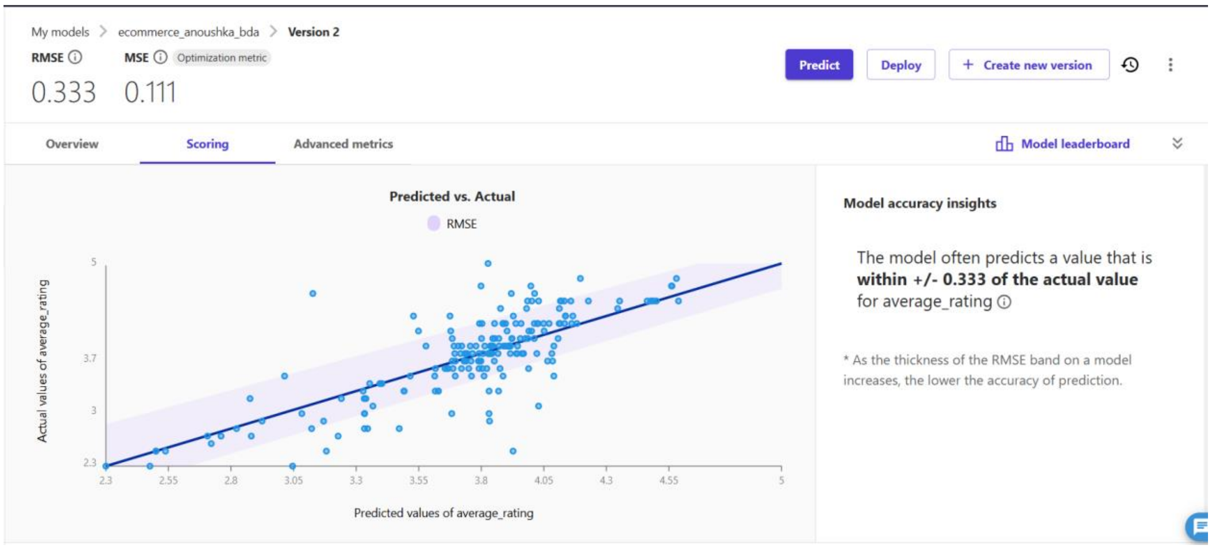
Search columns...

4	seller	5.985%
5	description	3.029%
6	brand	3.023%
7	product_details	3.002%
8	title	1.277%
9	discount	

Impact of seller on prediction of average_rating

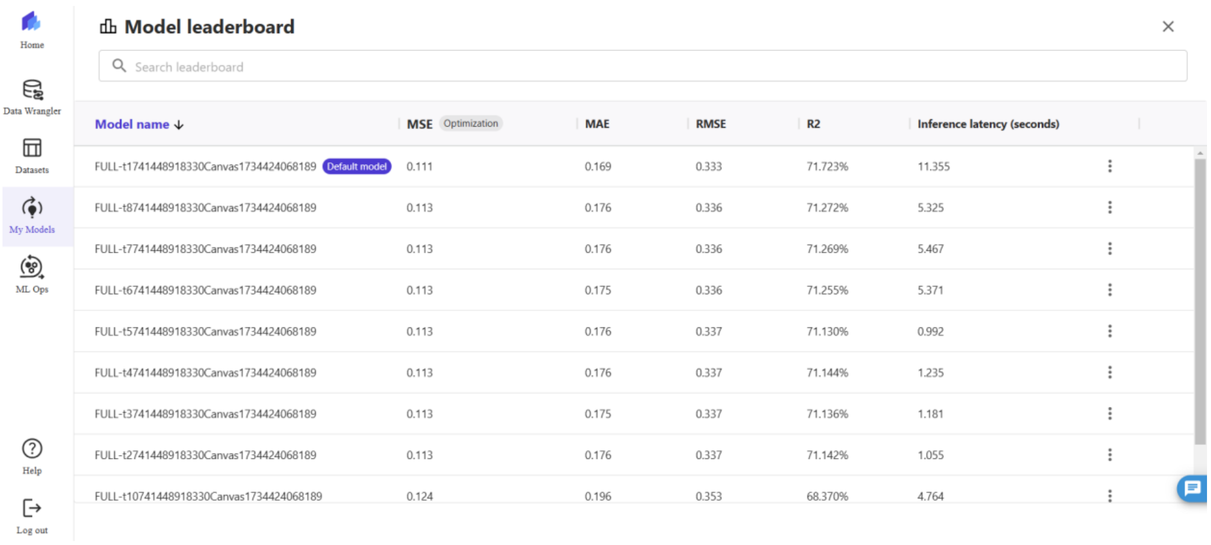
ecommerce_input_from_s3 | Total columns: 19 | Total rows: 13,402 | Total cells: 254,638 | average_rating | Numeric prediction

Predict



The model often predicts a value that is **within +/- 0.333 of the actual value** for average_rating

* As the thickness of the RMSE band on a model increases, the lower the accuracy of prediction.



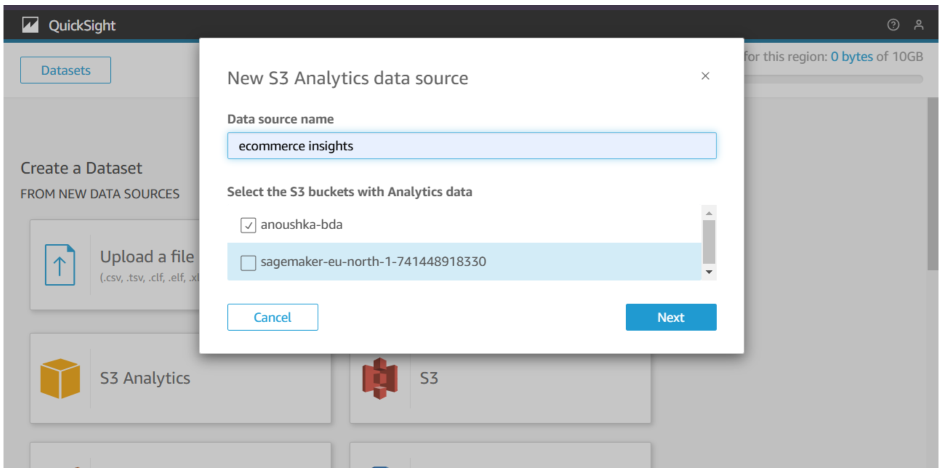
Model leaderboard

Model name ↓	MSE Optimization	MAE	RMSE	R2	Inference latency (seconds)
FULL-t1741448918330Canvas1734424068189 Default model	0.111	0.169	0.333	71.723%	11.355
FULL-t8741448918330Canvas1734424068189	0.113	0.176	0.336	71.272%	5.325
FULL-t7741448918330Canvas1734424068189	0.113	0.176	0.336	71.269%	5.467
FULL-t6741448918330Canvas1734424068189	0.113	0.175	0.336	71.255%	5.371
FULL-t5741448918330Canvas1734424068189	0.113	0.176	0.337	71.130%	0.992
FULL-t4741448918330Canvas1734424068189	0.113	0.176	0.337	71.144%	1.235
FULL-t3741448918330Canvas1734424068189	0.113	0.175	0.337	71.136%	1.181
FULL-t2741448918330Canvas1734424068189	0.113	0.176	0.337	71.142%	1.055
FULL-t10741448918330Canvas1734424068189	0.124	0.196	0.353	68.370%	4.764

The best model out of these was the default model because it has :

- **Lowest MSE and RMSE**: Indicates the model has the least error in predictions.
- **Highest R^2**: Shows the model explains the highest percentage of the variance in the data.

To visualise this data I created an account with Amazon Quicksight and gave access to my S3 bucket in the beginning of initialization itself. Then I created a new dataset where data was sourced from my S3 bucket.



New S3 Analytics data source

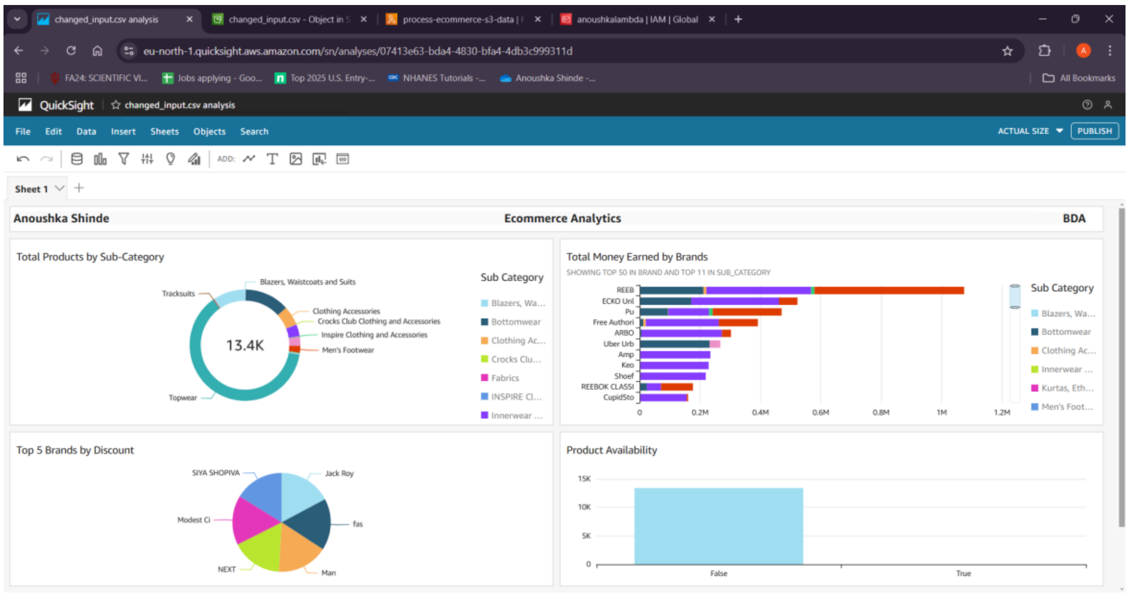
Data source name

ecommerce insights

Select the S3 buckets with Analytics data

☑ anoushka-bda

☐ sagemaker-eu-north-1-741448918330

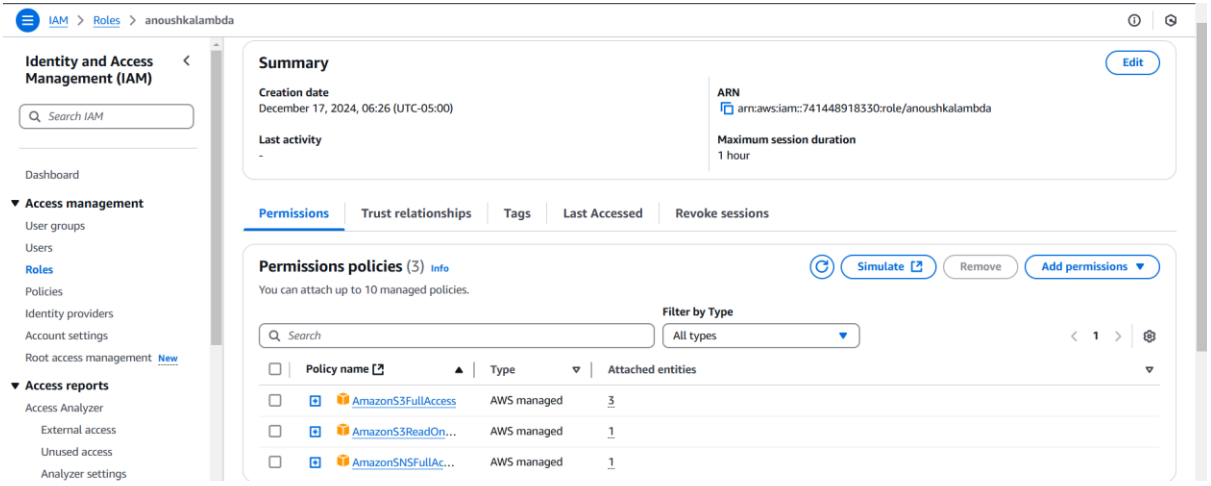


From the above visualizations we can summarise that:

- **Total Products by Sub-Category**:
  The **"Topwear"** and **"Blazers, Waistcoats and Suits"** categories dominate the total product count, accounting for a significant portion of the inventory. This indicates a higher focus on these sub-categories within the dataset.
- **Total Money Earned by Brands**:
  The **REEB** and **ECKO Unl** brands have the highest total earnings, significantly outpacing other brands. Brands such as **Pu** and **Free Authori** also contribute to overall sales, but their earnings are comparatively lower. This suggests that REEB and ECKO Unl are key revenue drivers.
- **Top 5 Brands by Discount**:
  The brands **SIYA SHOPIVA**, **Jack Roy**, and **fas** are offering the highest total discounts. This could indicate a strategy to attract more customers through higher discounts or possibly clear inventory.
- **Product Availability**:
  A significant number of products are **in stock** (True), while a smaller portion is marked as **out of stock** (False). This suggests that the inventory is well-maintained, with most products currently available to customers

BONUS

I also automated the entire pipeline. I started by creating a role that will be accessed by aws lambda to get data from s3 and send out push notifications using sns.



IAM > Roles > anoushkalambda

Identity and Access Management (IAM)

Summary

Creation date
December 17, 2024, 06:26 (UTC-05:00)

Last activity
-

ARN
arn:aws:iam::741448918330:role/anoushkalambda

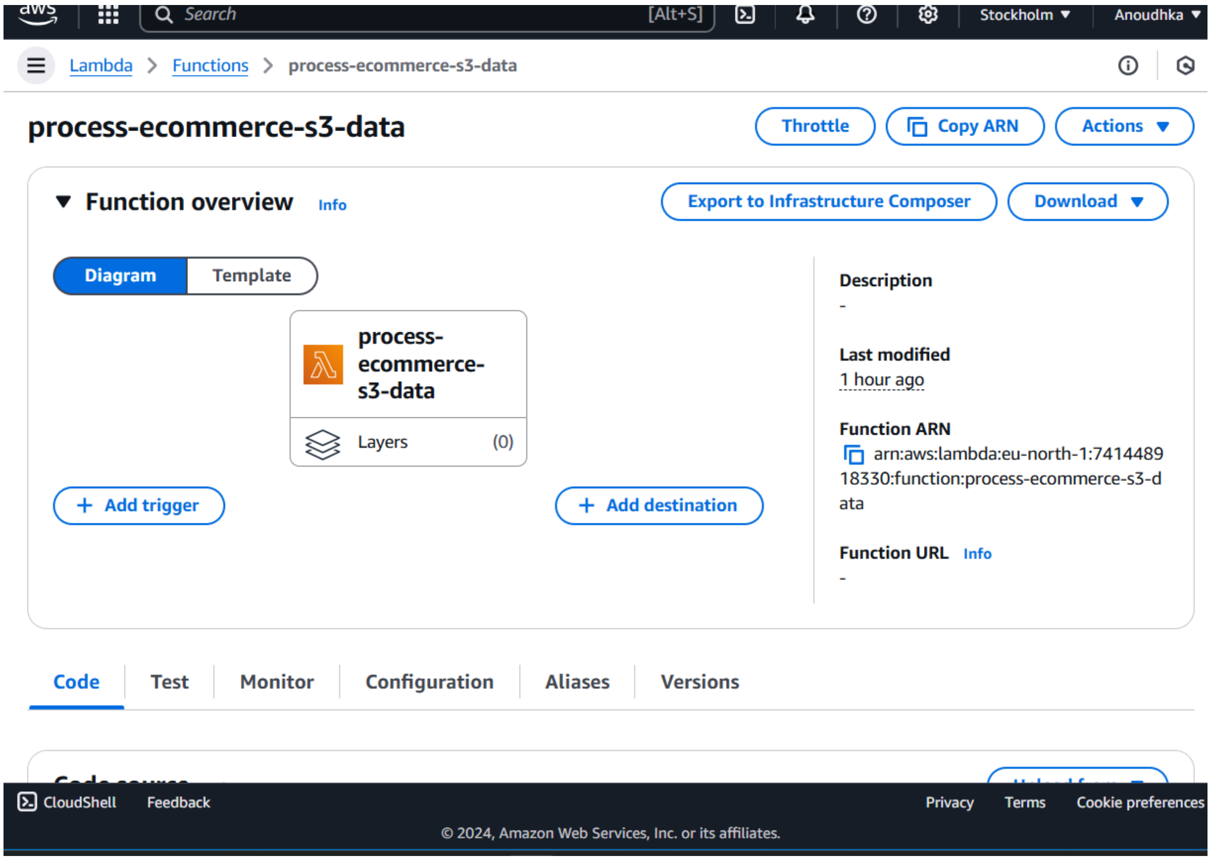
Maximum session duration
1 hour

Permissions policies (3)
You can attach up to 10 managed policies.

Filter by Type: All types

Policy name	Type	Attached entities
AmazonS3FullAccess	AWS managed	3
AmazonS3ReadOn...	AWS managed	1
AmazonSNSFullAc...	AWS managed	1

I also created a new function in Lambda. The script for processing can be found in the attachments. I attached the IAM I created to this function.



Lambda > Functions > process-ecommerce-s3-data

process-ecommerce-s3-data

Function overview

Description
-

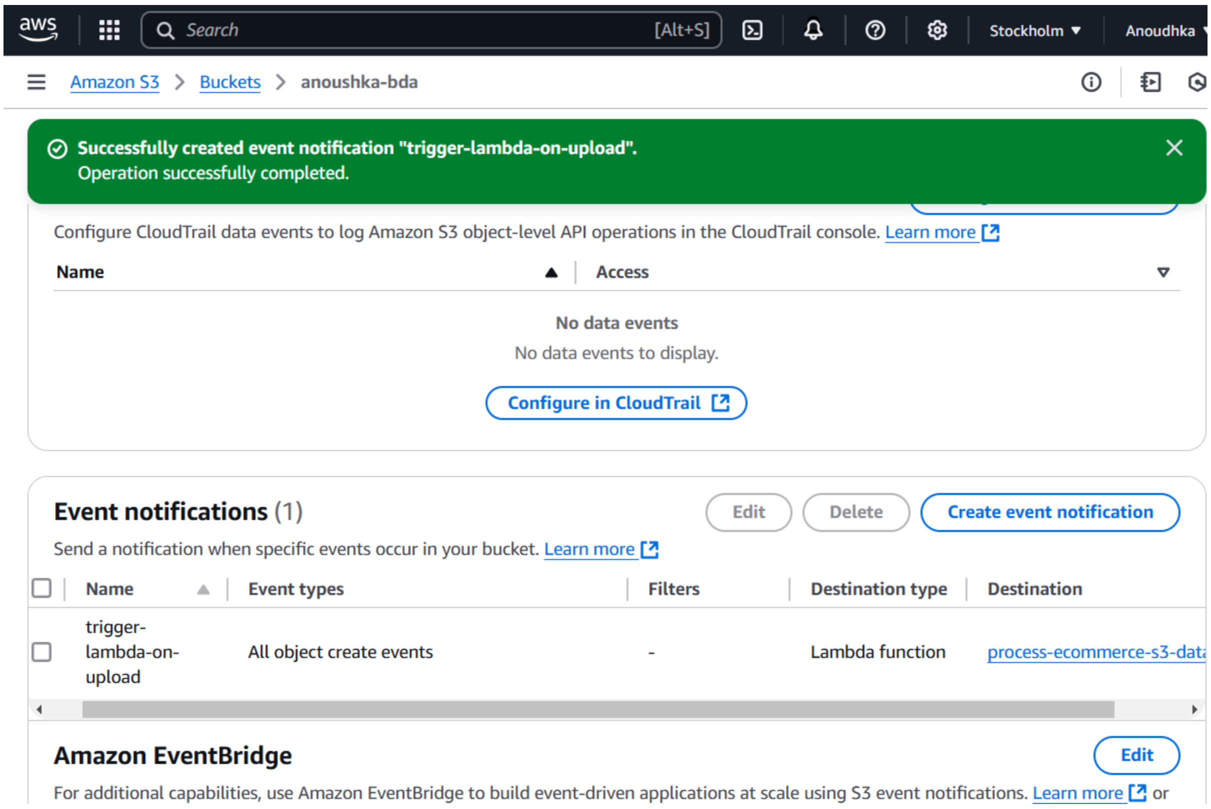
Last modified
1 hour ago

Function ARN
arn:aws:lambda:eu-north-1:741448918330:function:process-ecommerce-s3-data

Function URL
-

© 2024, Amazon Web Services, Inc. or its affiliates.

I then update the permissions of my s3 bucket to include s3 event notification with the trigger destination being the lambda function that I set up.



Successfully created event notification "trigger-lambda-on-upload".
Operation successfully completed.

Configure CloudTrail data events to log Amazon S3 object-level API operations in the CloudTrail console. Learn more

Name	Access

No data events
No data events to display.

Event notifications (1)

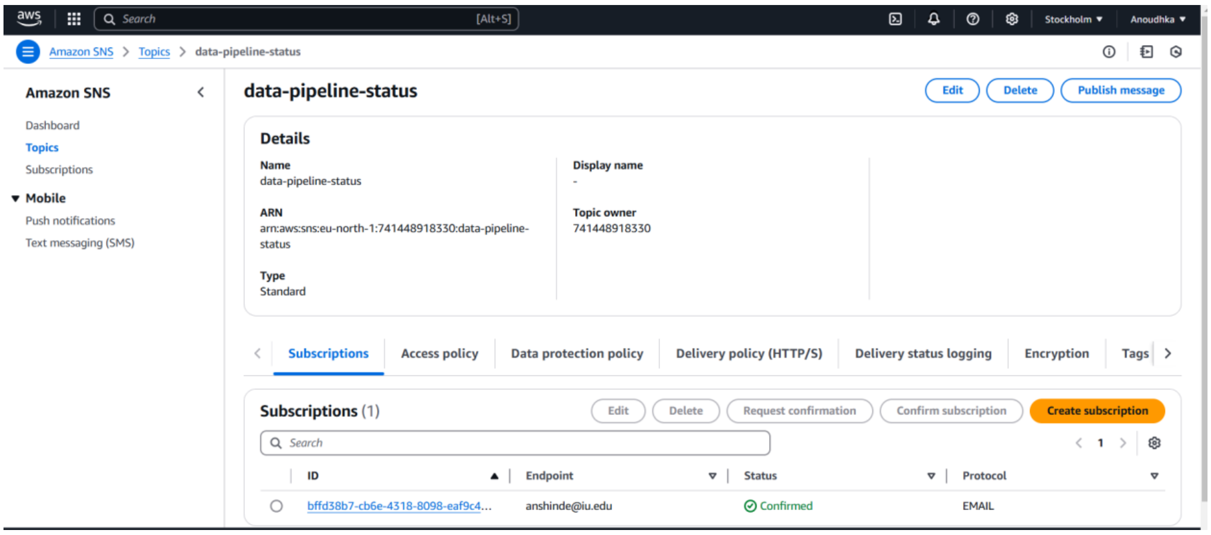
Send a notification when specific events occur in your bucket. Learn more

Name	Event types	Filters	Destination type	Destination
trigger-lambda-on-upload	All object create events	-	Lambda function	process-ecommerce-s3-data

Amazon EventBridge

For additional capabilities, use Amazon EventBridge to build event-driven applications at scale using S3 event notifications. Learn more or

Then, I wrote the lambda code that will read the file from S3, process or transform the data (simple operation like format validation) and save processed data back to a different S3 bucket. You can find the code attached in the zip file. After we deploy the lambda function, we will create an SNS topic.



data-pipeline-status

Details

Name
data-pipeline-status

Display name
-

ARN
arn:aws:sns:eu-north-1:741448918330:data-pipeline-status

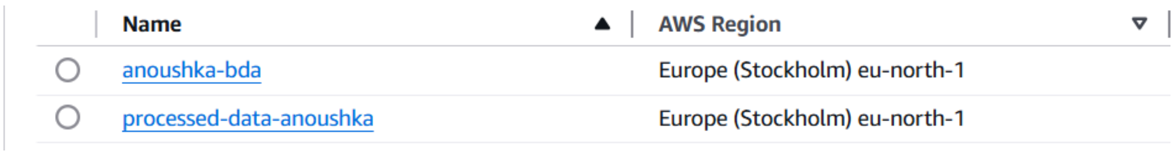
Topic owner
741448918330

Type
Standard

Subscriptions (1)

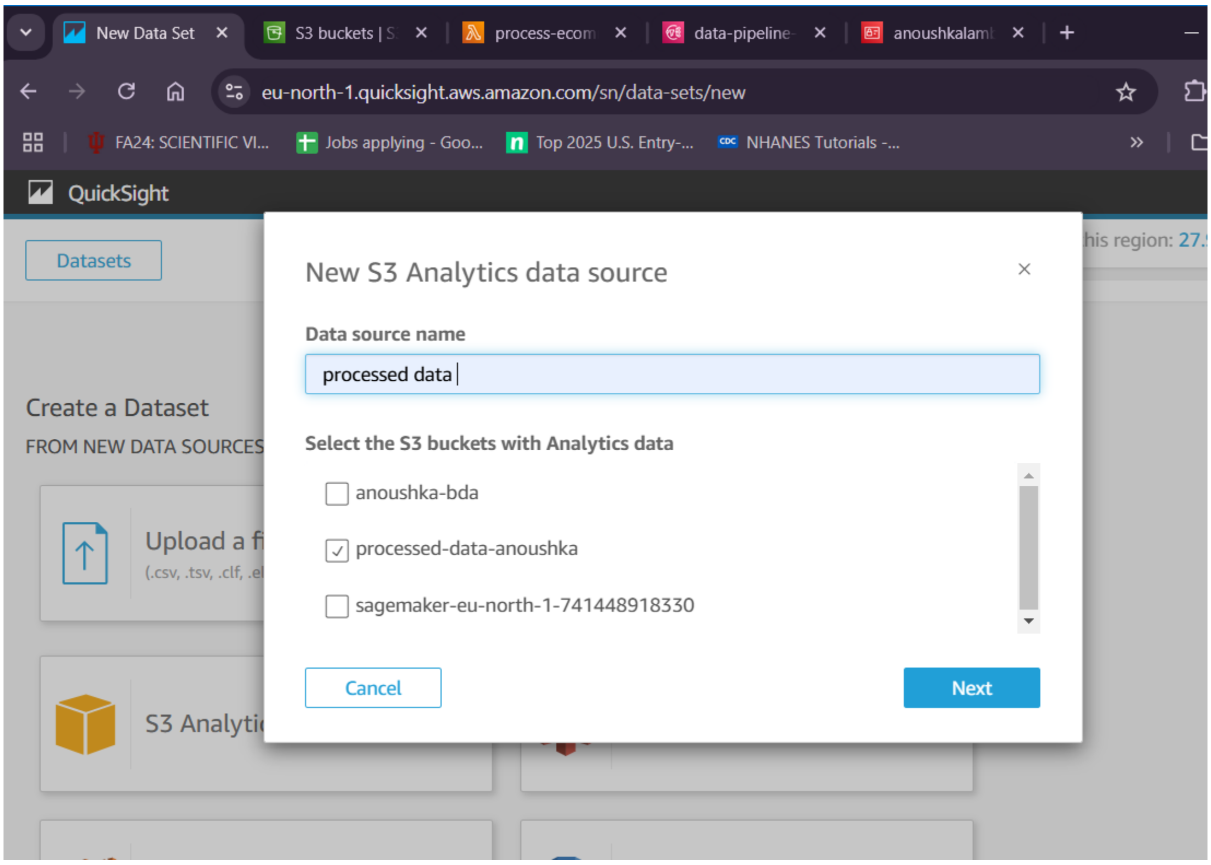
ID	Endpoint	Status	Protocol
bffd38b7-cb6e-4318-8098-eaf9c4...	anshinde@iu.edu	Confirmed	EMAIL

Additionally, I created a new bucket to store all the data coming from the lambda function and linked it to quick insight.



Name	AWS Region
anoushka-bda	Europe (Stockholm) eu-north-1
processed-data-anoushka	Europe (Stockholm) eu-north-1

Finally, I connected it to QuickSIght for Real Time Insights.


New Data Set | S3 buckets | S | process-ecom | data-pipeline | anoushkalam

eu-north-1.quicksight.aws.amazon.com/sn/data-sets/new

FA24: SCIENTIFIC VI... | Jobs applying - Goo... | Top 2025 U.S. Entry-... | NHANES Tutorials -...

QuickSight

Datasets

Create a Dataset

FROM NEW DATA SOURCES

Upload a fi (.csv, .tsv, .clf, .el

S3 Analyti

New S3 Analytics data source

Data source name

processed data

Select the S3 buckets with Analytics data

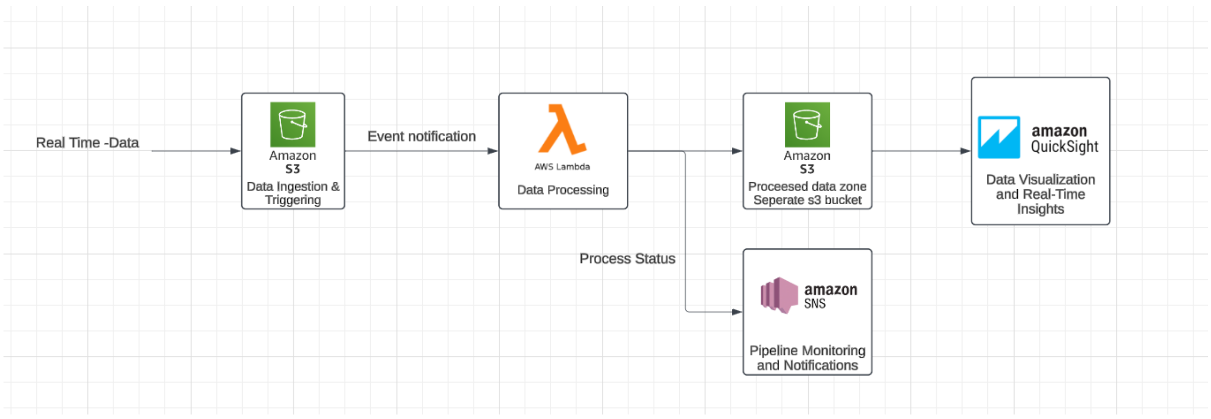
- [ ] anoushka-bda
- [x] processed-data-anoushka
- [ ] sagemaker-eu-north-1-741448918330

Cancel | Next

Architecture Summary

1. **Data Ingestion & Triggering: S3 Event Notifications** for triggering.
2. **Data Processing: AWS Lambda** for lightweight processing.
3. **Data Storage: Amazon S3** for raw and processed data.
4. **Notifications: AWS SNS** for pipeline status alerts.
5. **Data Visualization: Amazon QuickSight** (with SPICE) for dashboards.

Architecture diagram of the pipeline



Results

The implementation of the big data pipeline yielded significant insights into various aspects of the e-commerce platform's operations and performance metrics. The analysis revealed several key findings across multiple dimensions of the business.

The AWS SageMaker Autopilot implementation demonstrated robust predictive capabilities for average rating prediction. The model achieved notable performance metrics:

- Root Mean Square Error (RMSE) of 0.333
- Mean Square Error (MSE) of 0.111

The predicted vs. actual plot shows a strong positive correlation between predicted and actual values, with most predictions falling within the confidence band. The model consistently predicts values within ±0.333 of the actual average rating, indicating reliable performance for practical applications.

The analysis of product distribution revealed significant insights into inventory composition. A total of approximately 13.4K products were analyzed across various sub-categories. Topwear emerged as the dominant category, representing the largest segment. Tracksuits and Blazers/Waistcoats/Suits formed significant secondary categories. Clothing Accessories and specialized categories like Crocks Club showed moderate representation.

Revenue analysis across brands revealed clear market leaders. REEB emerged as the top-performing brand in terms of revenue generation. ECKO Unltd showed strong performance as the second-highest revenue generator. A clear tiering of brand performance was observed, with significant gaps between top performers and middle-tier brands. Brands like Amp and Keo maintained steady mid-level performance.

The product availability analysis presented crucial operational insights. A substantial portion of the inventory showed as 'False' in the availability metric, suggesting potential stock management challenges. The disparity between available and unavailable products indicates an opportunity for inventory optimization.

Finally, the model's feature importance analysis revealed valuable insights into rating predictions. Seller information emerged as a significant predictor, contributing 5.985% to the model's decisions. Product description and brand showed similar importance levels (approximately 3%). The impact of sellers on rating predictions varied significantly, with some sellers showing consistent positive impact while others demonstrated more variable effects

Conclusion

This comprehensive analysis of the Flipkart e-commerce dataset through our implemented big data pipeline has revealed several crucial insights for business operations and strategy. The machine learning model's performance demonstrates the viability of automated rating prediction systems, while the business metrics highlight clear opportunities for optimization.

The variation in brand performance and product category distribution suggests opportunities for targeted inventory management and marketing strategies. The significant impact of sellers on product ratings indicates the importance of seller quality management and potential for optimization of seller partnerships.

The implementation successfully demonstrated the scalability and effectiveness of combining AWS services with PySpark for large-scale e-commerce data analysis. The insights generated provide actionable intelligence for business decision-making, particularly in areas of inventory management, brand partnerships, and seller relations.

References:

Kaggle: https://www.kaggle.com/datasets/aaditshukla/flipkart-fasion-products-dataset/code

QuickSight Documentation: https://docs.aws.amazon.com/quicksight/latest/user/quickstart-createanalysis.html

SageMaker Auto Tutorial: https://aws.amazon.com/tutorials/machine-learning-tutorial-automatically-create-models/