

PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

Problem Statement:

Develop a scalable, end-to-end data analytics platform that ingests, transforms, analyses, and visualizes Amazon product sales data using Databricks and AWS services. The goal is to empower stakeholders with real-time insights into product performance, customer behaviour, price trends, and to enable machine learning-driven sales forecasting and anomaly detection.

Dataset:

Source: Kaggle

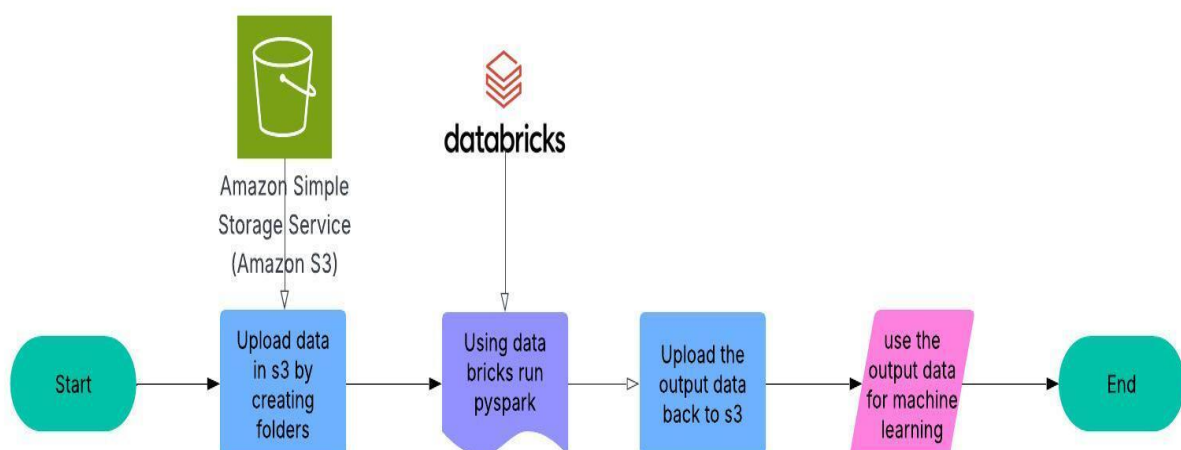
Title: Amazon Products Dataset

Description: This dataset contains Amazon product listings including product names, categories, prices, ratings, and reviews.

Business Goals:

1. Identify top-selling product categories and high-demand products.
2. Detect pricing anomalies that could impact sales or customer trust.
3. Predict customer ratings for new or updated products.
4. Analyse average rating trends over time across categories.
5. Enable real-time alerting and dashboard updates for business stakeholders.

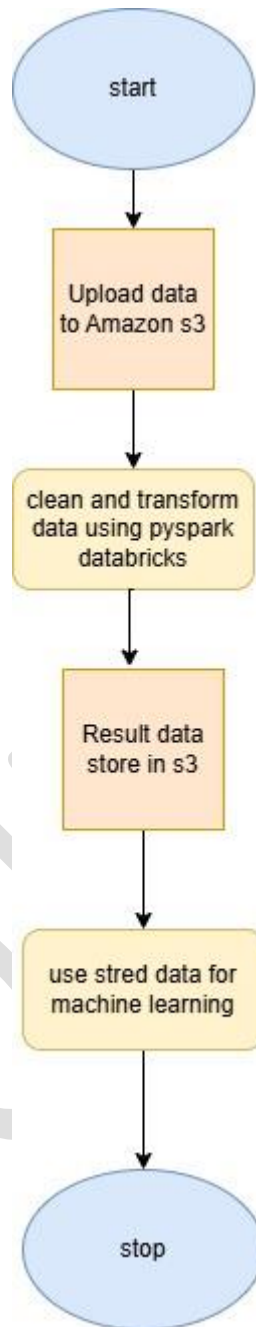
Data Architecture:



PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

Data flow:



Data Analysis Tasks:

1. BRONZE LAYER:

- Raw CSV uploaded to s3://myhemap2-b/bronze/
- Created sub folders – Bronze, Silver, Gold.

PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

aws

Search

[Alt+S]

United States (N. Virginia)

cloud_user @ 3814-9203-2511

Amazon S3

Buckets

Create bucket

US East (N. Virginia) us-east-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

myhemap2-b

Bucket names must be 3 to 63 characters and unique within the global namespace. Bucket names must also begin and end with a letter or number. Valid characters are a-z, 0-9, periods (.), and hyphens (-). [Learn More](#)

Copy settings from existing bucket - optional

Only the bucket settings in the following configuration are copied.

Choose bucket

Format: s3://bucket/prefix

Object Ownership

Amazon S3

Buckets

Create bucket

Format: s3://bucket/prefix

Object Ownership

Info

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

☐ ACLs disabled (recommended)

All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

☒ ACLs enabled

Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

⚠ We recommend disabling ACLs, unless you need to control access for each object individually or to have the object writer own the data they upload. Using a bucket policy instead of ACLs to share data with users outside of your account simplifies permissions management and auditing.

Object Ownership

☐ Bucket owner preferred

If new objects written to this bucket specify the bucket-owner-full-control canned ACL, they are owned by the bucket owner. Otherwise, they are owned by the object writer.

☒ Object writer

The object writer remains the object owner.

CloudShell

Feedback

© 2025 Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie preferences

aws

Search

[Alt+S]

United States (N. Virginia)

cloud_user @ 3814-9203-2511

Amazon S3

Buckets

Create bucket

☐ Block all public access

Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.

☐ Block public access to buckets and objects granted through new access control lists (ACLs)

S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.

☐ Block public access to buckets and objects granted through any access control lists (ACLs)

S3 will ignore all ACLs that grant public access to buckets and objects.

☐ Block public access to buckets and objects granted through new public bucket or access point policies

S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.

☐ Block public and cross-account access to buckets and objects through any public bucket or access point policies

S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.

⚠ Turning off block all public access might result in this bucket and the objects within becoming public

AWS recommends that you turn on block all public access, unless public access is required for specific and verified use cases such as static website hosting.

☒ I acknowledge that the current settings might result in this bucket and the objects within becoming public.

PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

The first screenshot shows the 'myhemap2-b' bucket policy being edited. A green notification bar at the top states 'Successfully edited bucket policy.' Below it, the JSON policy is displayed in a code editor. A 'Copy' button is visible on the right.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "PublicReadForAllObjects",
      "Effect": "Allow",
      "Principal": "*",
      "Action": "s3:GetObject",
      "Resource": "arn:aws:s3:::myhemap2-b/*"
    }
  ]
}
```

The second screenshot shows an 'Upload succeeded' notification for the file 'Amazon-Products.csv' (179.9 MB) in the 'bronze/' folder. Below the notification, the 'Files and folders' tab is active, showing a table with one entry: 'Amazon-Products.csv' (text/csv, 179.9 MB, Succeeded).

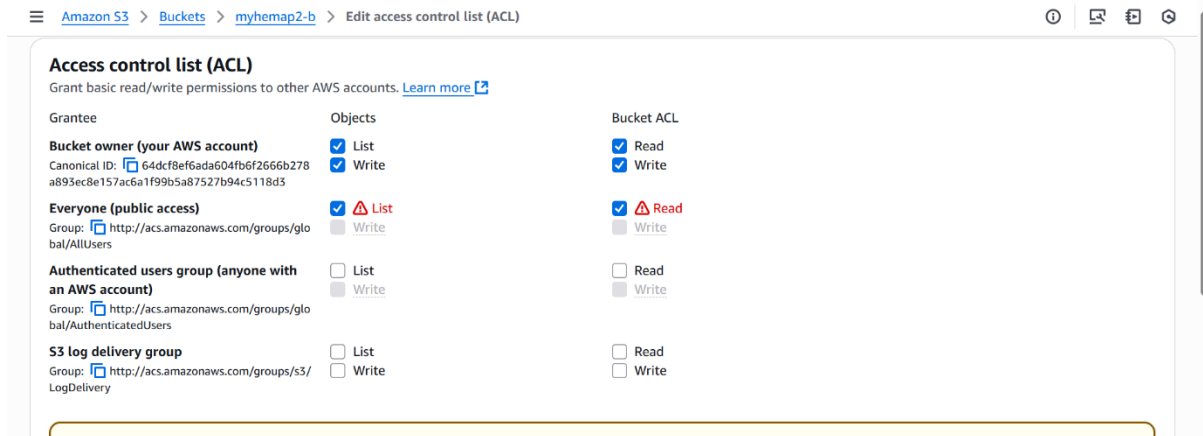
Name	Folder	Type	Size	Status	Error
Amazon-Products.csv	-	text/csv	179.9 MB	Succeeded	-

The third screenshot shows the 'myhemap2-b' bucket overview. The 'Objects' tab is selected, displaying a list of three folders: 'bronze/', 'gold/', and 'silver/'. Above the list are various action buttons like 'Copy S3 URI', 'Download', 'Delete', 'Create folder', and 'Upload'.

Name	Type	Last modified	Size	Storage class
bronze/	Folder	-	-	-
gold/	Folder	-	-	-
silver/	Folder	-	-	-

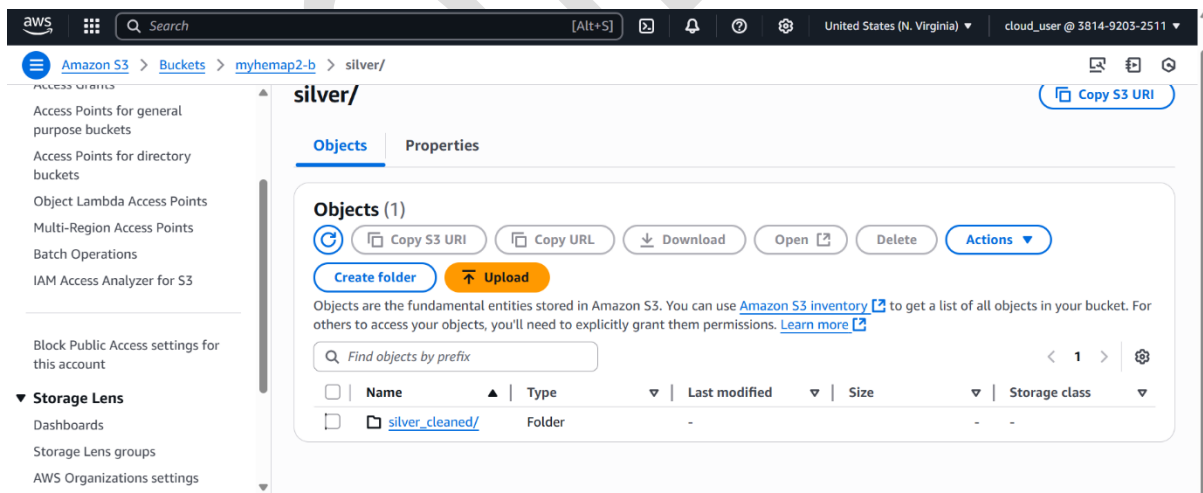
PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS



2. SILVER LAYER:

- Clean price (remove symbols, convert to float)
- Convert ratings to numeric
- Remove duplicates & fix nulls
- Standardize `main_category` / `sub_category`



PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

Amazon S3 > Buckets > myhemap2-b > silver/ > silver_cleaned/

Objects (11)

Copy S3 URI Copy URL Download Open Delete Actions

Create folder Upload

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

Find objects by prefix

Name	Type	Last modified	Size	Storage class
_committed_5200607549135708385	-	10:24:16 (UTC-04:00)	728.0 B	Standard
_started_5200607549135708385	-	July 9, 2025, 10:24:12 (UTC-04:00)	0 B	Standard
_SUCCESS	-	July 9, 2025, 10:24:17 (UTC-04:00)	0 B	Standard
part-00000-tid-5200607549135708385	-			

3. GOLD LAYER:

- Aggregate: Average rating, total reviews per category
- Exported as CSV to `s3://myhemap2-b/gold/`

Amazon S3 > Buckets > myhemap2-b > gold/

gold/

Copy S3 URI

Objects Properties

Objects (1)

Copy S3 URI Copy URL Download Open Delete Actions

Create folder Upload

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

Find objects by prefix

Name	Type	Last modified	Size	Storage class
_analysis-report/	Folder	-	-	-

Amazon S3 > Buckets > myhemap2-b > gold/ > analysis-report/

analysis-report/

Copy S3 URI

Objects (11)

Copy S3 URI Copy URL Download Open Delete Actions

Create folder Upload

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

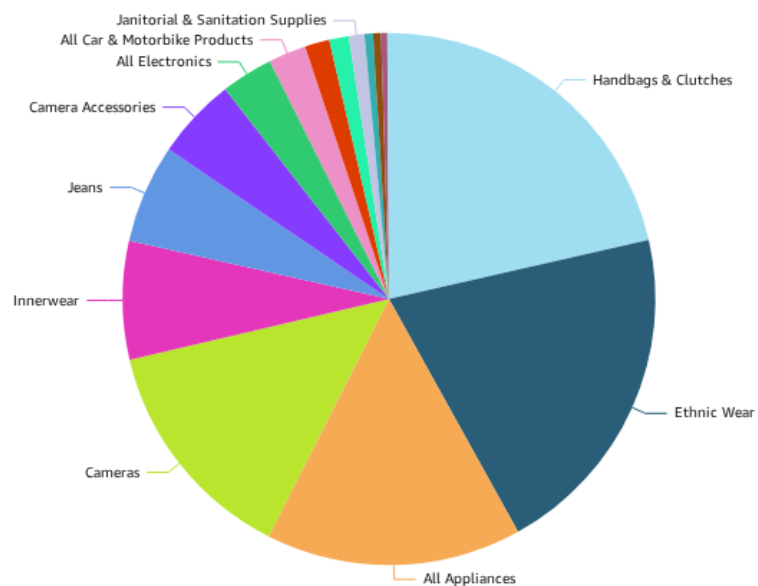
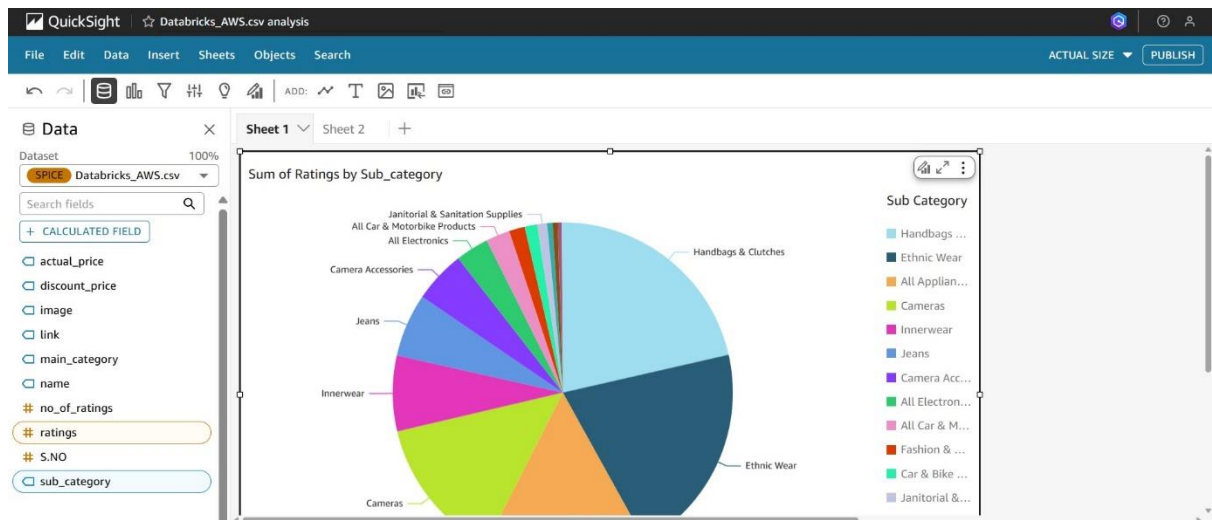
Find objects by prefix

Name	Type	Last modified	Size	Storage class
_started_8521954063017852233	-	July 9, 2025, 10:30:02 (UTC-04:00)	0 B	Standard
_SUCCESS	-	July 9, 2025, 10:30:09 (UTC-04:00)	0 B	Standard
part-00000-tid-8521954063017852233-e9ffaaf2-3225-4808-bc2c-	csv	July 9, 2025, 10:30:05 (UTC-04:00)	14.2 MB	Standard

PROJECT 2 DOCUMENTATION

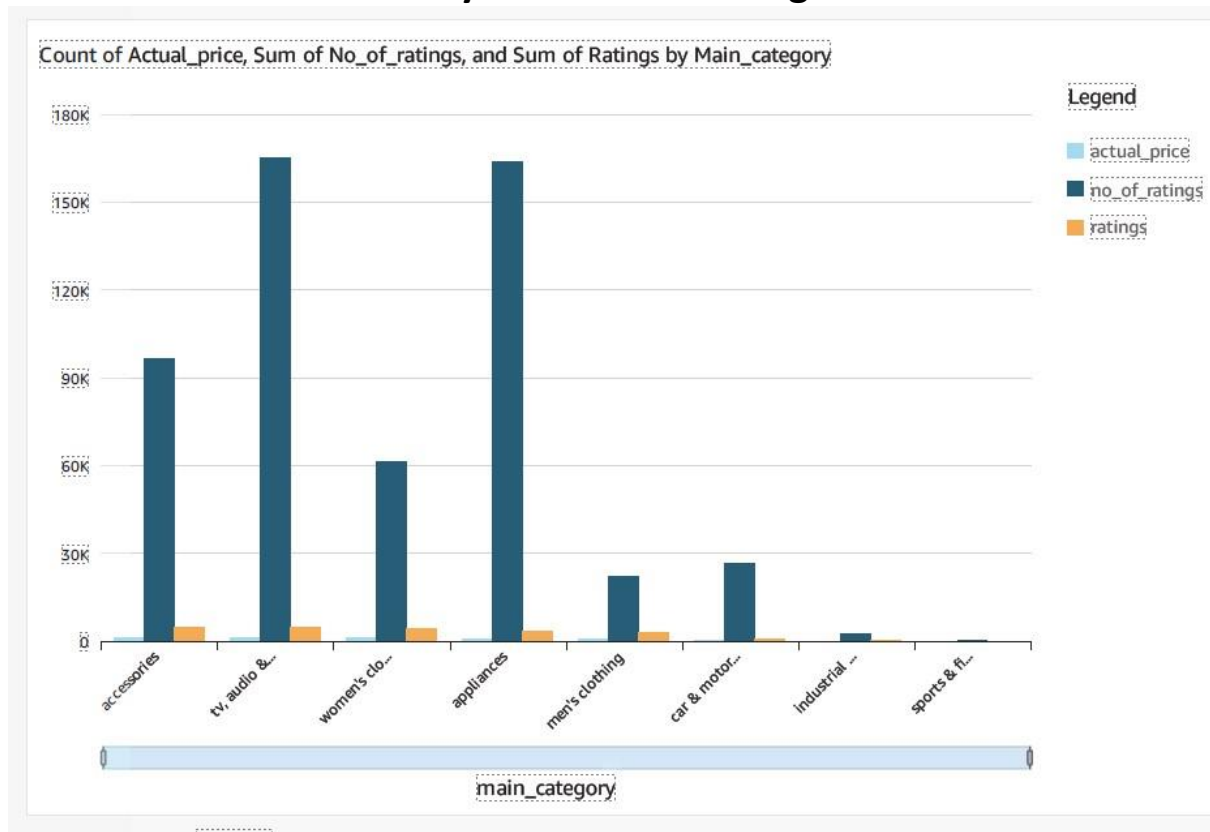
Advanced Product Analytics Platform Using Databricks & AWS

DASHBOARD ANALYSIS:



PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS



MACHINE LEARNING TASKS:

Predict Product Ratings (Regression Model)

Goal: Estimate the customer rating for new or modified products before they go live.

Approach:

- Model Type: Linear Regression / XGBoost Regressor
- Features Used:
 - price
 - discount_price
 - sub_category (One-hot encoded)

PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

```
# Evaluate
predictions = lr_model.transform(test)
predictions.select("ratings", "prediction").show(10)

from pyspark.ml.evaluation import RegressionEvaluator

evaluator = RegressionEvaluator(labelCol="ratings", predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)
print(f"RMSE: {rmse}")
```

```
+-----+-----+
|ratings| prediction|
+-----+-----+
| 5.0|3.7387652322825478|
| 4.5|3.7418476840189707|
| 4.2|3.7732326898620507|
| 5.0| 3.738484996082761|
| 4.3|3.7418476752802334|
| 3.4|3.7482928104087234|
| 4.2| 3.807980365989092|
| 3.4| 3.775754681782681|
| 3.5|3.7387652018717423|
| 4.5|3.7387652018717423|
+-----+-----+
only showing top 10 rows

RMSE: 0.758829945631222
```

CONCLUSION:

The Advanced Product Analytics Platform successfully delivers a robust, scalable solution for processing and analysing Amazon product data in real-time. By leveraging the power of AWS and Databricks, the project achieved the following:

Key Outcomes:

- **Efficient Data Ingestion**
Ingested raw CSV files into Amazon S3 (Bronze Layer) and catalogued them with AWS Glue for seamless access.
- **Cleaned & Standardized Data**
Transformed raw product listings into clean, structured datasets (Silver Layer) using PySpark and Databricks.
- **Aggregated Insights**
Derived valuable business metrics in the Gold Layer, such as average ratings, top-selling categories, and total reviews.
- **Machine Learning Integration**
Built and tested ML models for:
 - Rating Prediction
- **Visualization:**
Delivered dynamic, real-time dashboards via Amazon Quick Sight, enabling stakeholders to make data-driven decisions.