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

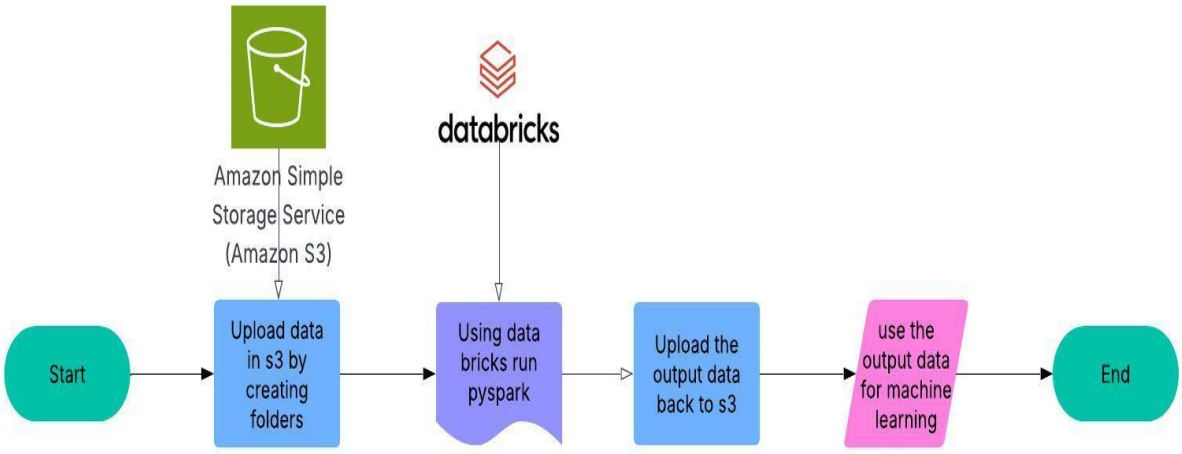
Title: Amazon Products Dataset

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. Analys e average rating trends over time across categories.
5. Enable real-time alerting and dashboard updates for business stakeholders.

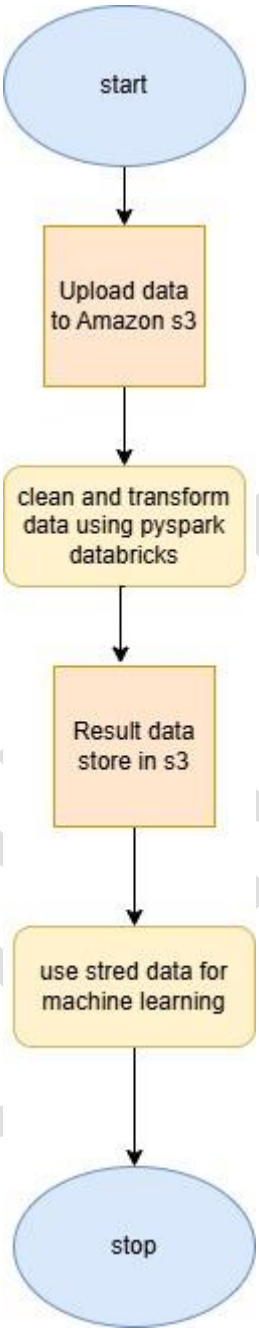
Data Architecture:



PROJECT 2 DOCUMENTATION

Advanced Product Analytics Platform Using Databricks & AWS

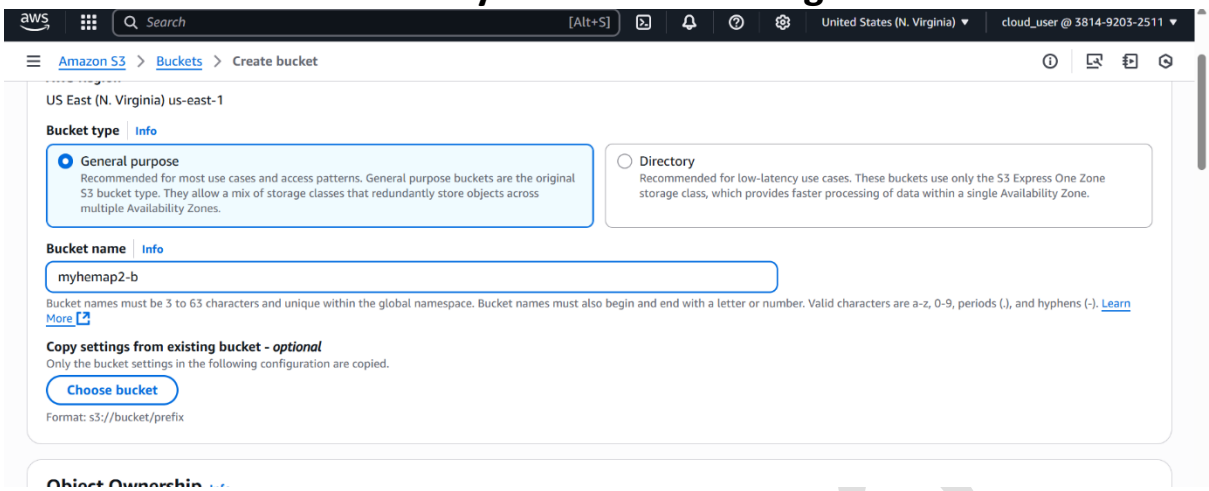
Data flow:

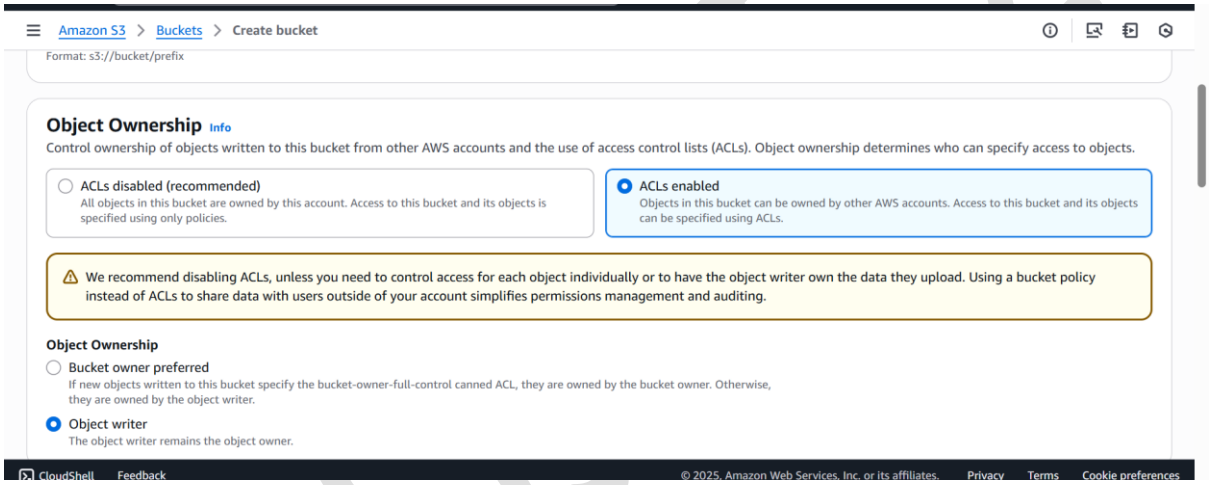


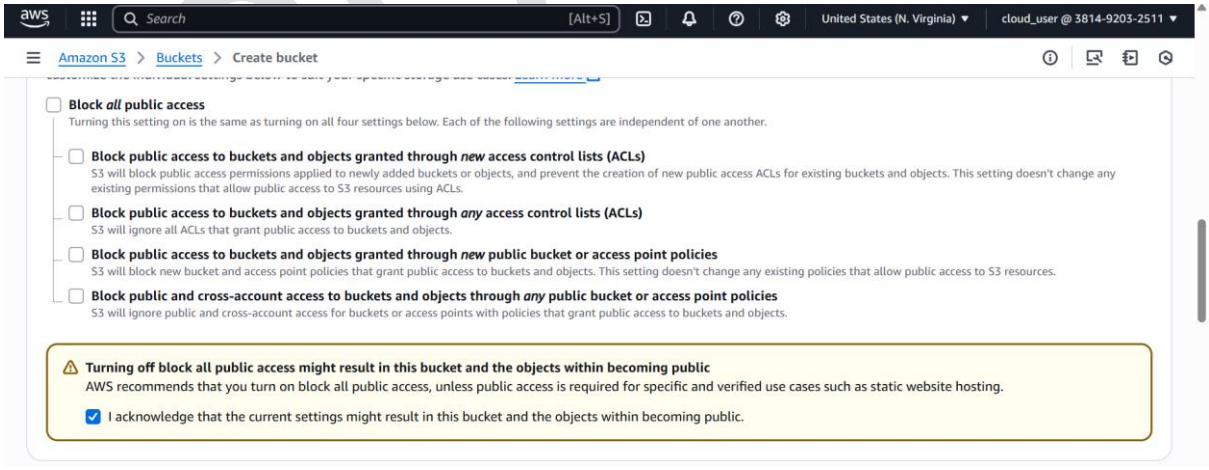
Data Analysis Tasks:

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

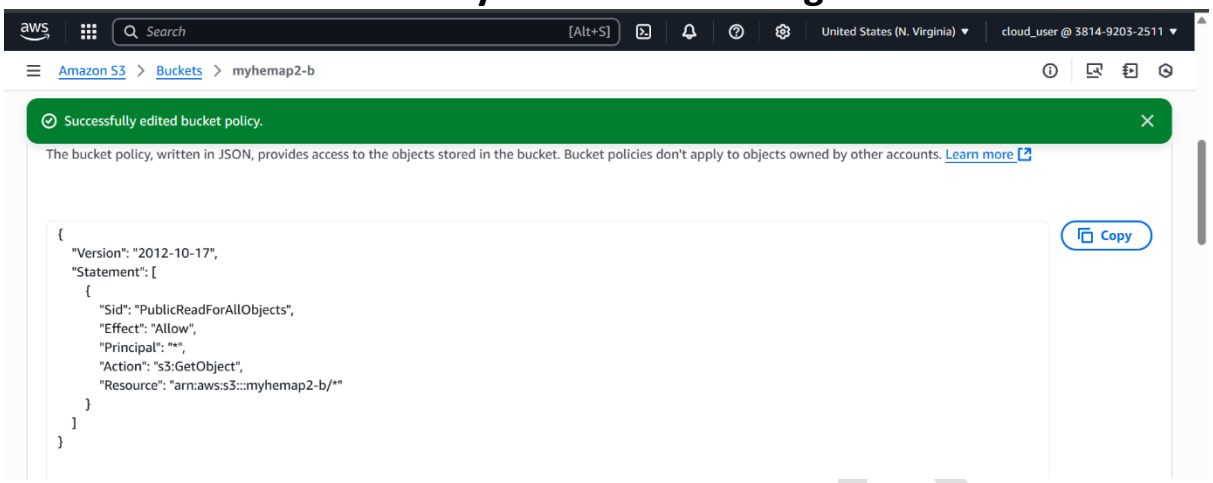
PROJECT 2 DOCUMENTATION
Advanced Product Analytics Platform Using Databricks & AWS

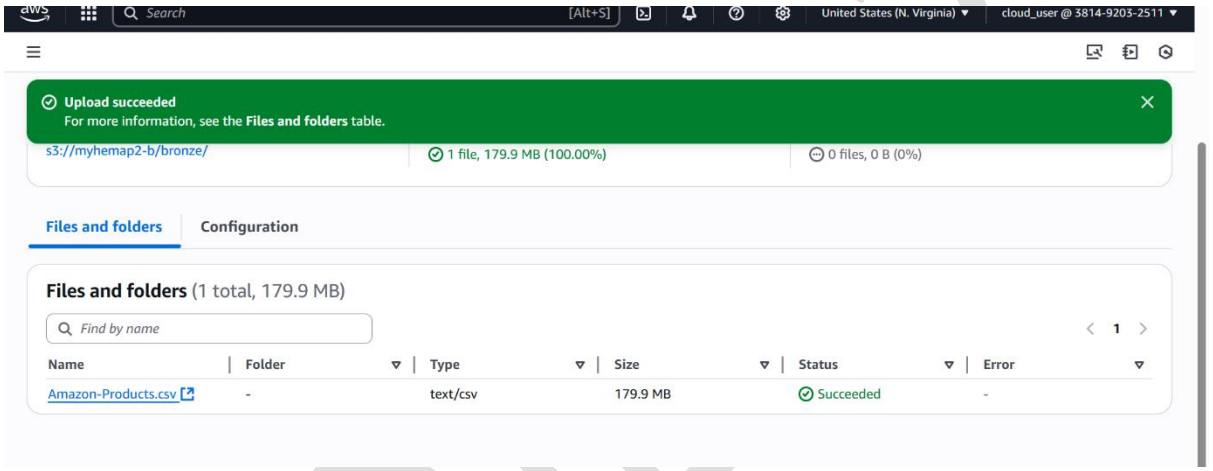


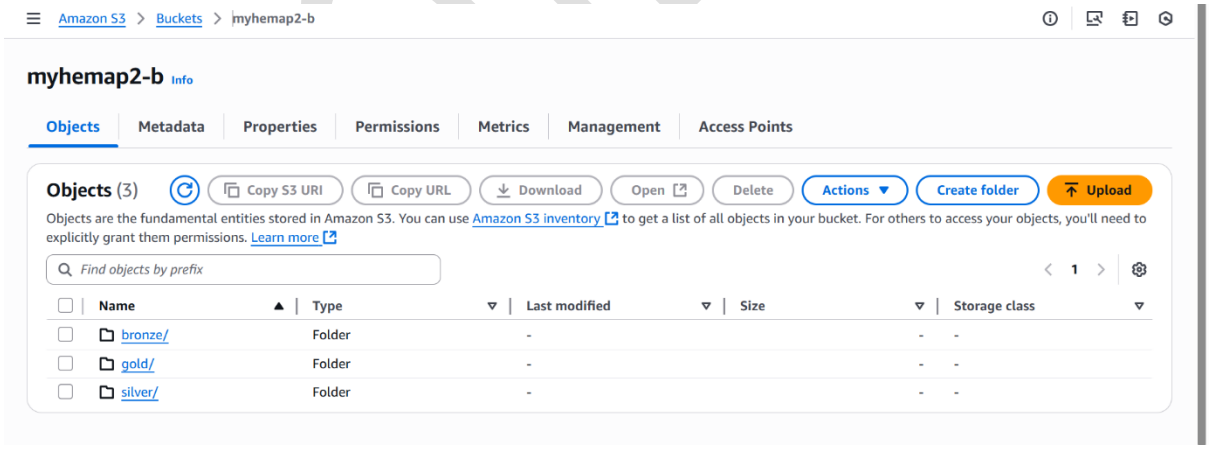


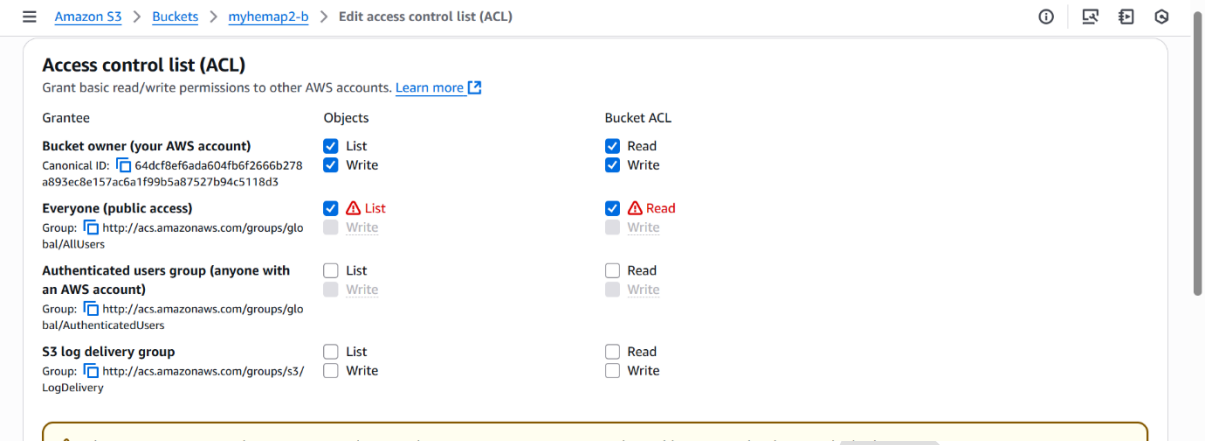


PROJECT 2 DOCUMENTATION
Advanced Product Analytics Platform Using Databricks & AWS



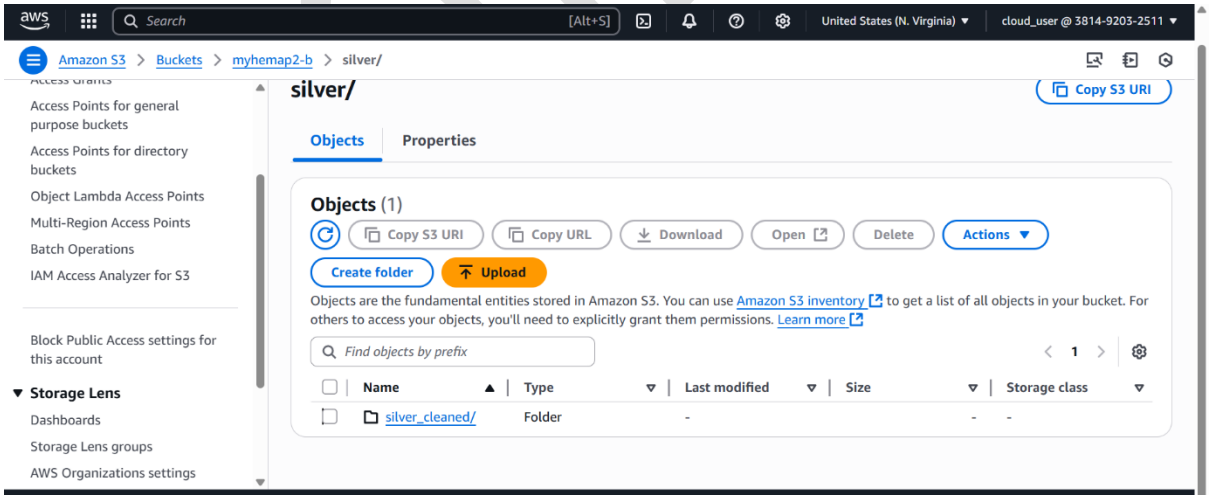




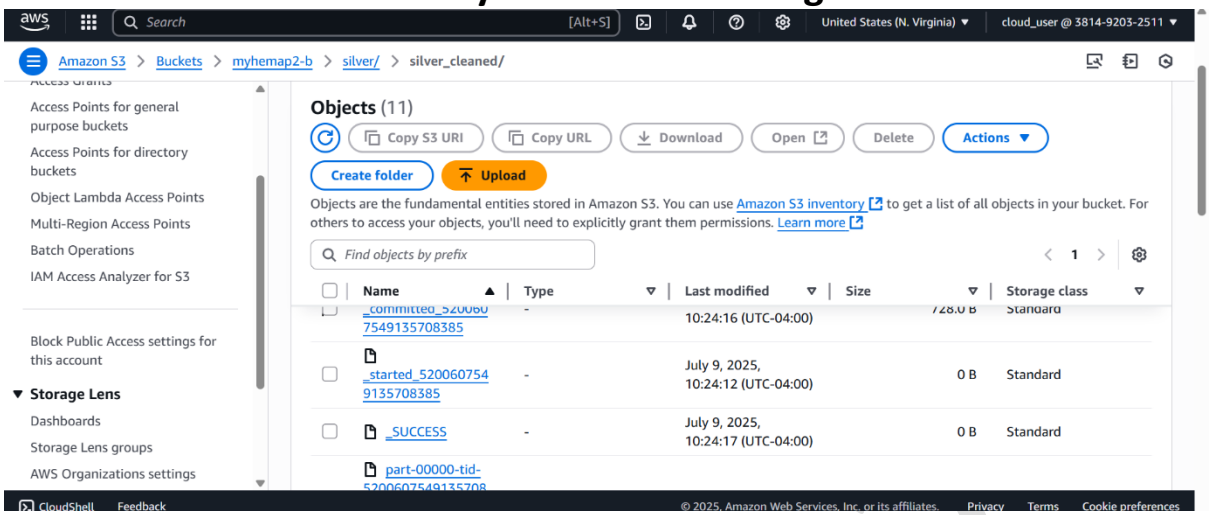


2. SILVER LAYER:

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

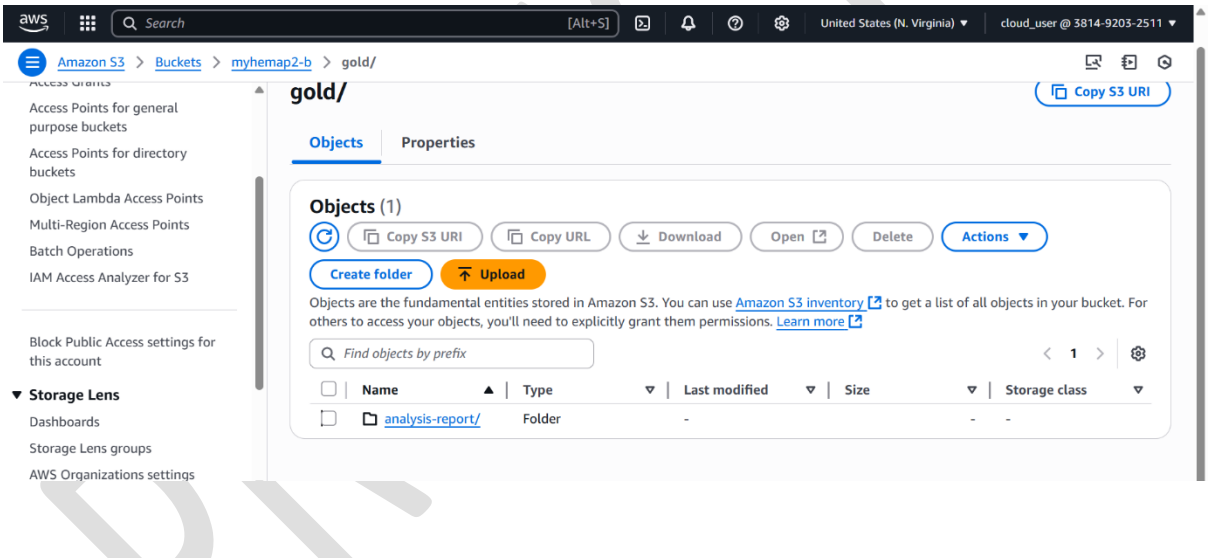


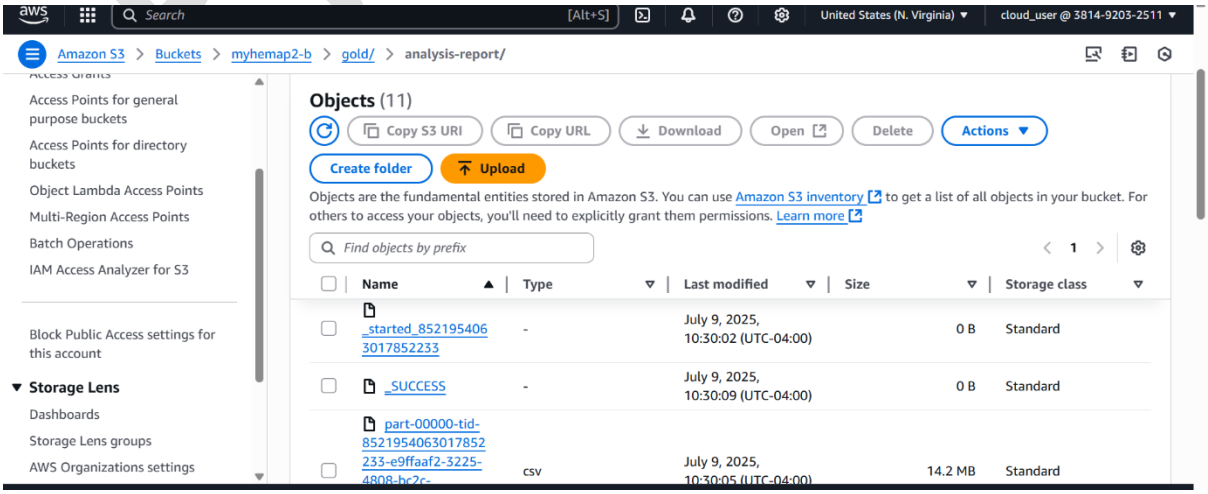
PROJECT 2 DOCUMENTATION
Advanced Product Analytics Platform Using Databricks & AWS



3. GOLD LAYER:

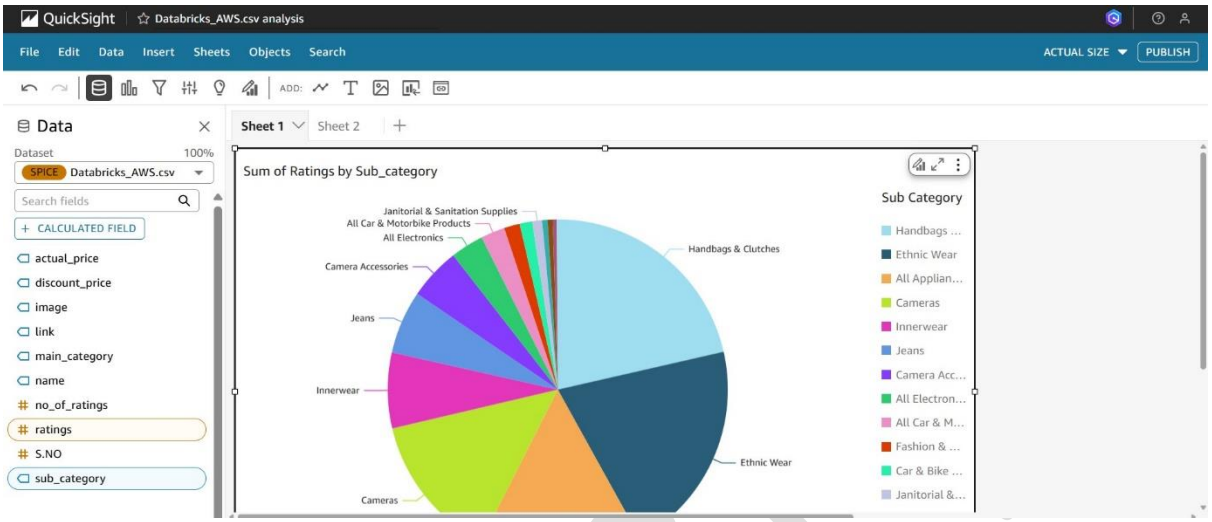
- Aggregate: Average rating, total reviews per category
- Exported as CSV to `s3://myhemap2-b/gold/`

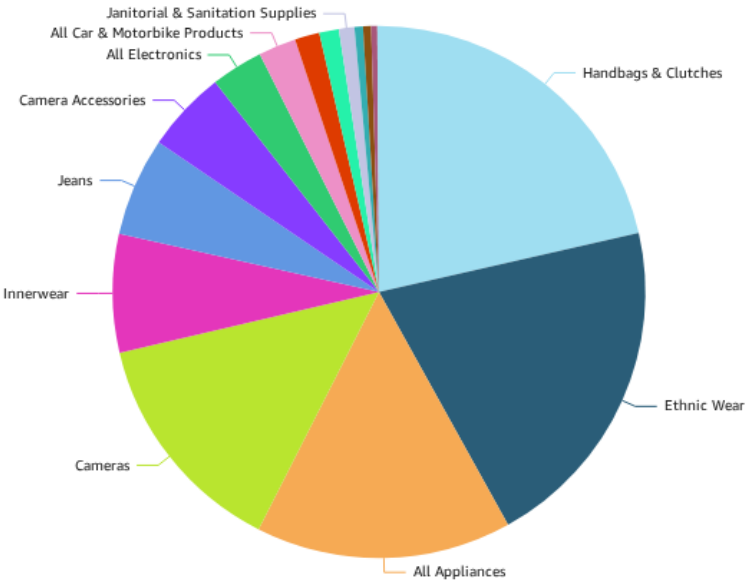




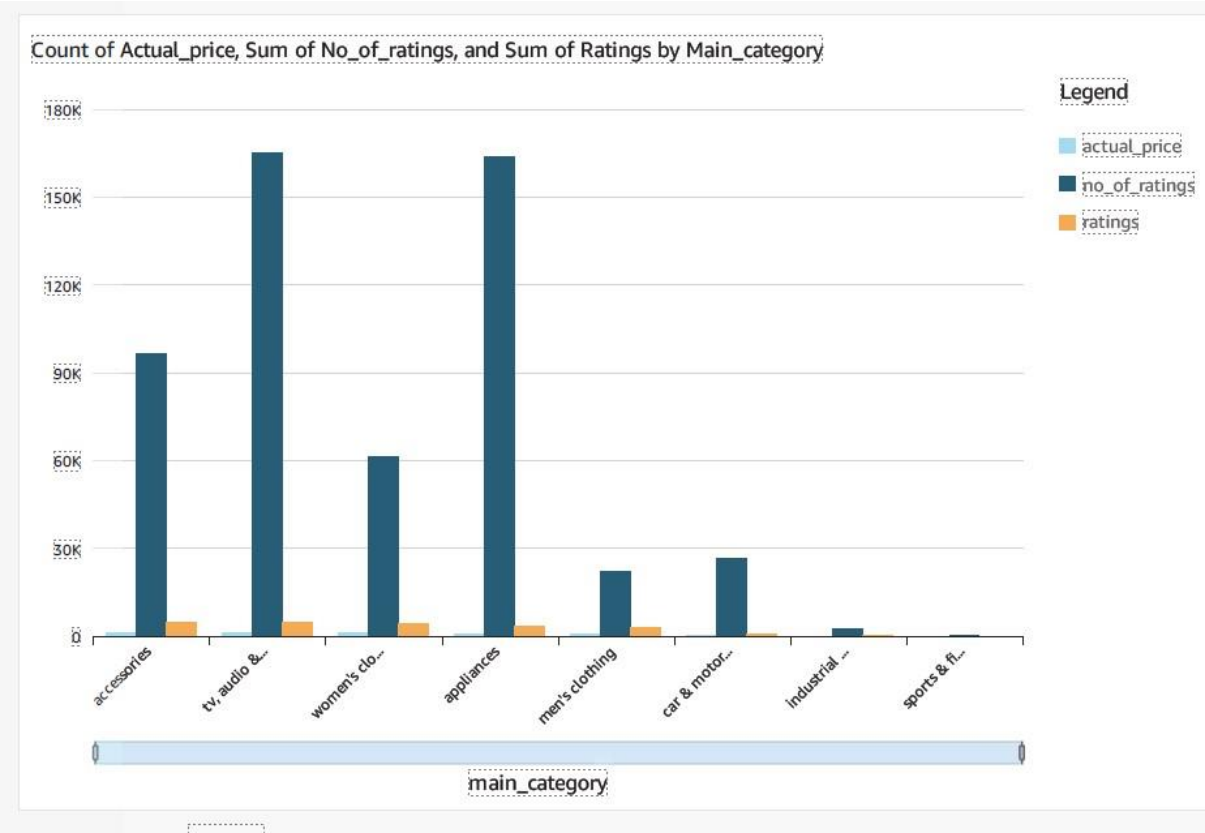
PROJECT 2 DOCUMENTATION
Advanced Product Analytics Platform Using Databricks & AWS

DASHBOARD ANALYSIS:





PROJECT 2 DOCUMENTATION
Advanced Product Analytics Platform Using Databricks & AWS



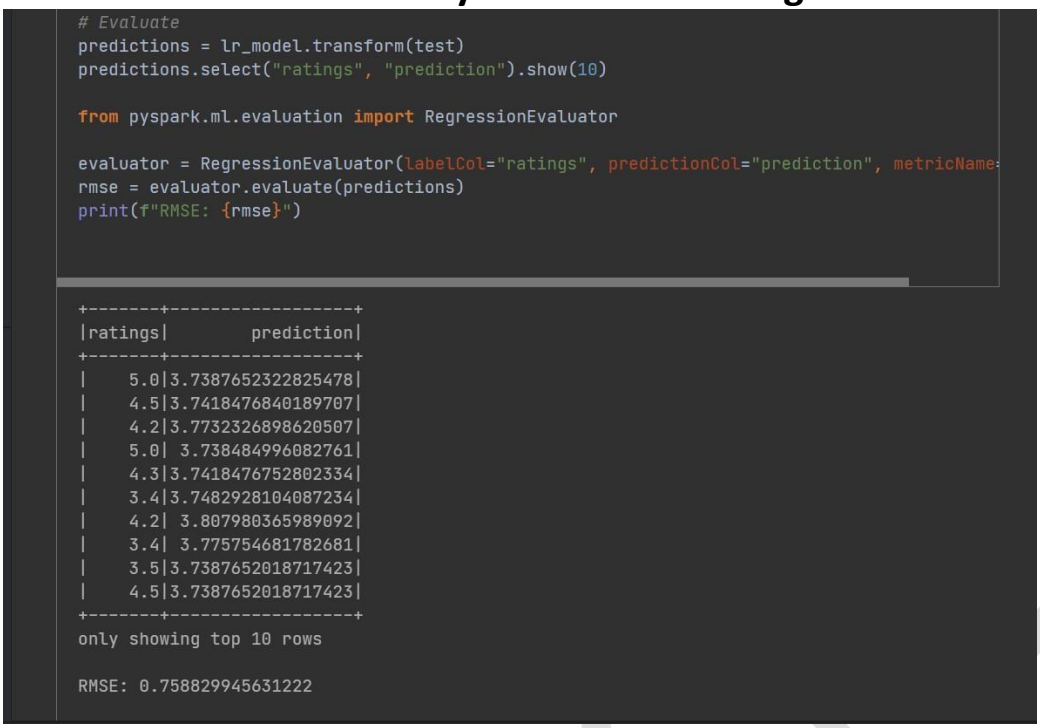
MACHINE LEARNING TASKS:

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)



```
# 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 = 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.