Advanced Product Analytics Platform Using Databricks & AWS

Bootcamp - AWS Data Engineering Project - 2

1. Objective

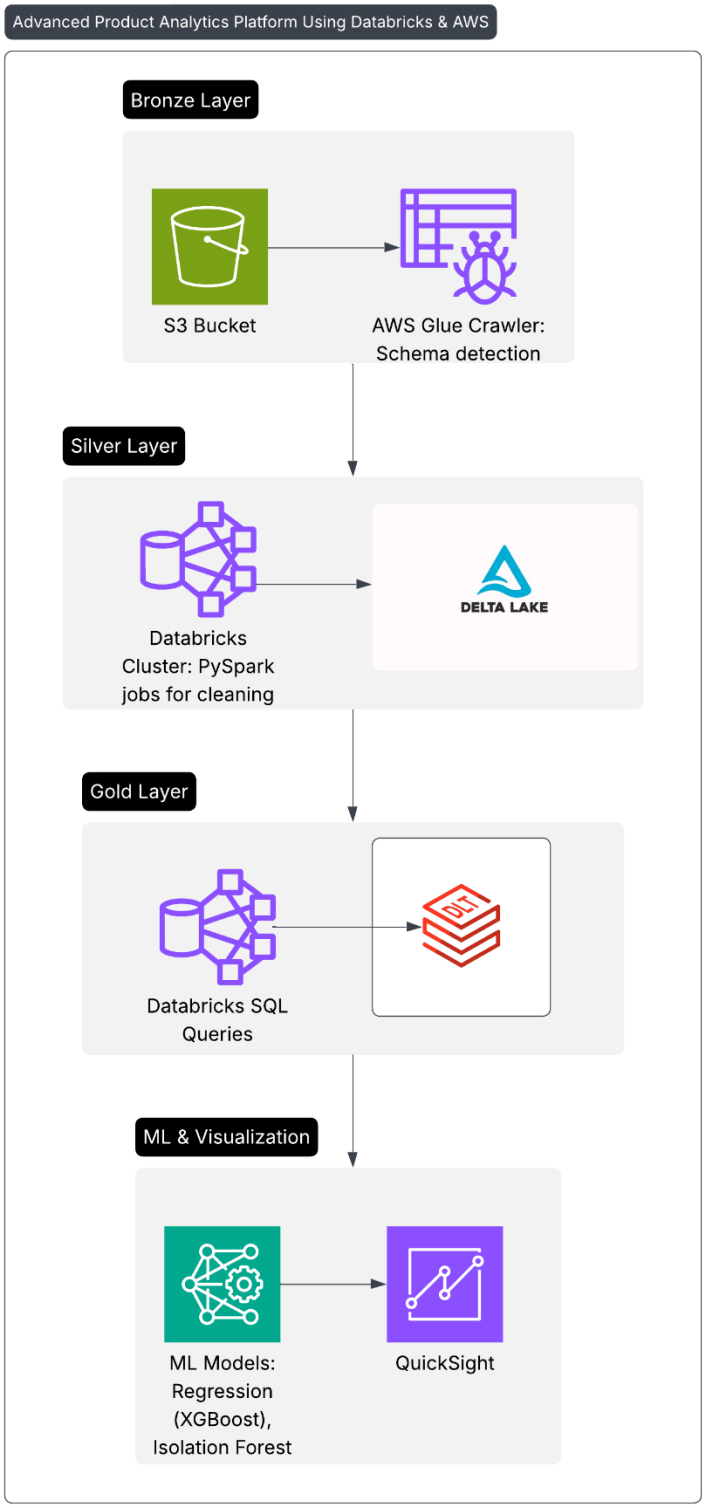
The platform aims to:

- Ingest, transform, and analyze Amazon product data to derive actionable insights.
- Deploy machine learning models for sales forecasting, rating prediction, and anomaly detection.
- Empower stakeholders with interactive dashboards (QuickSight).

Key Outcomes:

- Identified top-selling categories and high-demand products.
- Detected pricing outliers using Isolation Forest.
- Predicted product ratings via XGBoost regression.
- Aggregated metrics (avg. rating, reviews, price trends) by category.

2. System Architecture



3. Prerequisites

- AWS Account with permissions for S3, Glue, Lambda, QuickSight.
- Databricks Workspace (AWS integration).
- Dataset: Amazon Products Dataset (Kaggle).
- Libraries: PySpark, Pandas, Scikit-learn, XGBoost.

4. Component Breakdown

Component	Purpose
S3 Buckets	Storage for raw (Bronze), cleaned (Silver), and aggregated (Gold) data.
AWS Glue	Schema detection and cataloging for raw data.
Databricks	ETL (PySpark), ML model training, and Delta Lake management, Model deployment.
QuickSight	Visualization of aggregated metrics and predictions.

5. Design Decisions

- Delta Lake: Chosen for ACID compliance, time travel, and schema evolution.
- PySpark over Pandas: Scalability for large datasets.
- Isolation Forest: Unsupervised anomaly detection for pricing outliers.
- Modular Layers: Bronze-Silver-Gold architecture ensures data quality and reusability

6. Data Flow

1. Ingestion: CSV -> S3 Bronze (via manual upload, Glue Crawler).
2. Cleaning: Databricks processes raw data -> Silver (Delta).
3. Aggregation: Silver -> Gold (PySpark aggregations).
4. ML: Gold data -> Train models -> Predictions -> S3/QuickSight.
5. Visualization: QuickSight connects to Gold layer for dashboards.

7. Security & Compliance

- IAM Roles: Least-privilege access for Glue, Databricks.
- S3 Encryption: Server-side encryption (SSE-S3) for data at rest.

- Delta Lake: Audit logs for data changes.
- Compliance: Follows AWS Well-Architected Framework (Data Protection Pillar).

8. Monitoring & Quality

- AWS CloudWatch: Logs for Lambda/Glue jobs.
- Databricks Jobs: Scheduled runs with failure alerts.
- Data Quality Checks:
    - Null checks in Silver layer (na.drop()).
    - Validation of numeric fields (e.g., ratings between 1-5).