Project Statement

End-to-End E-Commerce Analytics, Data Engineering & AI Platform using Databricks

---

Problem Statement

Modern e-commerce organizations generate massive volumes of data across customer interactions, transactions, inventory, and marketing channels. While data availability is high, many organizations struggle to convert raw, fragmented data into timely, decision-ready insights due to:

- Siloed data systems
- Inconsistent data quality
- Limited scalability of analytics pipelines

As a result, business teams lack a single, trusted source of truth for analytics, while data teams face challenges in supporting advanced analytics, machine learning, and AI use cases at scale.

To address these challenges, the organization requires a **unified, production-ready data platform** that can:

- Ingest large-scale raw data from multiple sources
- Clean, standardize, and model data reliably
- Enable SQL-based analytics for business stakeholders
- Support advanced analytics, machine learning, and AI workflows
- Scale efficiently with growing data volume and user demand

In addition to descriptive analytics, the business aims to leverage historical customer behavior data to **predict user intent**, improve recommendations, optimize marketing spend, and enhance customer engagement.

Project Description

This project demonstrates how an e-commerce organization can leverage the **Databricks Lakehouse architecture** to build a scalable, end-to-end analytics and AI platform.

Starting from raw e-commerce interaction data, the solution implements a **multi-layered Medallion architecture (Bronze → Silver → Gold)** to ensure:

- High data quality
- Strong governance
- Reusable, analytics-ready datasets

Business-ready datasets are exposed for SQL analytics and reporting, while curated **Gold-layer feature tables** are prepared to support downstream machine learning workflows.

The platform integrates data engineering, analytics, and AI workloads using:

- Databricks
- Delta Lake
- Spark SQL
- Databricks Workflows
- MLflow

This architecture closely mirrors **modern, production-grade enterprise data platforms** used in real-world organizations.

Primary AI Use Case: User–Item Conversion Prediction

As a representative AI use case, this project focuses on predicting the likelihood that a user will convert (make a purchase) for a given product based on historical behavioral signals, including:

- Page views
- Add-to-cart actions
- Interaction intensity
- Engagement depth

AI Framing

- **ML Task Type:** Binary Classification
- **Input Features:**

    - Interaction score (weighted user actions)
    - Funnel depth (behavioral progression)
    - Engagement score (business-level abstraction)
- **Target Variable:** converted
    - 0 = No purchase
    - 1 = Purchase
- **Model Output:** Probability of conversion per user–item pair

Business Impact

- Prioritize high-intent users for targeted promotions
- Improve product recommendation strategies
- Optimize marketing spend by focusing on users with higher conversion likelihood

---

Solution Architecture Overview

1. Data Ingestion (Bronze Layer)

- Ingest raw e-commerce events and product data
- Store immutable raw data in Delta Lake Bronze tables
- Preserve data for traceability, auditing, and reprocessing

2. Data Cleaning & Standardization (Silver Layer)

- Remove duplicates and handle missing values
- Apply business rules and validation logic
- Enrich datasets with behavioral and categorical features

3. Exploratory Data Analysis & Validation

- Analyze customer behavior and interaction patterns
- Validate assumptions before downstream analytics and machine learning

4. Data Modeling & Analytics (Gold Layer)

- Create analytics-ready, business-friendly datasets

- Enable scalable SQL queries, KPI reporting, and dashboards
- Prepare curated feature tables for machine learning

5. Machine Learning & AI

- Perform feature engineering on Gold-layer data
- Train and evaluate machine learning models
- Track experiments and manage models using MLflow
- Persist predictions in Delta tables for downstream analytics and reporting

---

Business Impact

- Enables data-driven decision-making through a single, trusted data platform
- Reduces data silos and manual reporting efforts
- Supports scalable analytics and AI workflows
- Translates customer behavior data into actionable business insights

---

Key Learning Outcomes

- Designed and implemented an end-to-end Lakehouse architecture on Databricks
- Applied real-world data engineering best practices
- Built ML-ready feature pipelines from large-scale datasets
- Integrated analytics and machine learning workflows using MLflow

---