Introduction

This project develops a cloud-hosted retail analytics and recommendation prototype using the UCI Online Retail dataset, addressing three main questions: who to sell to, what to sell, and when. The workflow runs fully in Apache Spark, covering ETL, analytics, and machine learning at scale.

Traditional single-machine notebooks cannot handle millions of rows or reproduce full pipelines reliably. They face limits in memory, fault tolerance, and scalability. The goal is to create a **reproducible, distributed system** that performs large-scale data processing efficiently in the cloud.

Objectives and Features

The project integrates data engineering and machine learning into one Spark-based workflow:

1. **ETL:** Clean invoices, remove cancellations, and store data as month-partitioned Parquet.
2. **Analytics:**
    - (1) Market-basket co-occurrence with support, confidence, and lift;
    - (2) Customer segmentation using RFM metrics and K-Means (k=4);
    - (3) Recommendation system with ALS (implicit feedback) evaluated by Precision, Recall, HitRate, NDCG, and MRR.
3. **Outputs:** Visual reports (PNG) and data tables (CSV) for reuse and presentation.

Technical Solution

The system runs on a GCP Compute Engine Spark cluster (1 master, 2 workers). Data and outputs are stored in HDFS/Object Storage, processed with PySpark SQL, MLlib (K-Means & ALS), and custom metrics. All workflows are containerized, ensuring repeatability and easy scaling.

Current results:

- ALS@10: Precision=0.031, Recall=0.047, HitRate=0.243, NDCG=0.170, MRR=0.090.
- K-Means (k=4):
    - Seg1 High-Value (R=6.6, F=82.5, M=127,338, n=13)
    - Seg0 Loyal (R=15.0, F=22.3, M=12,709, n=204)

- Seg3 Promising (R=43.4, F=3.7, M=1,358, n=3,060)
- Seg2 Dormant (R=248.2, F=1.6, M=478, n=1,061)
  ~20% of SKUs drive ~80% of revenue (Pareto pattern).

Cloud Benefits & Cost

Spark provides distributed computation, durability, and scalability. With preemptible nodes, costs stay low.
Estimated monthly cost (120 hours, 50 GiB total):

- Master = $18.16; Worker 1 = $10.78; Worker 2 = $10.78 → Total = $39.73/month.

Architecture Flow

Workflow

Raw CSV lands in storage; ETL writes Parquet. Jupyter submits jobs to the master; executors on workers run SQL/ML stages. Results (tables/plots) are saved to a shared volume and synchronised back to storage for hand-off to BI or apps.

This workflow demonstrates a scalable, cost-efficient, and reproducible big data pipeline ready for future retail intelligence applications.