Case Study: Scalable Retail Data Pipeline & ETL Design

Background

XYZ Retail operates hundreds of stores nationwide, generating vast volumes of sales, inventory, and product data daily. Currently, these datasets (stored in CSVs and local systems) are fragmented and inconsistent, making it difficult to answer critical business questions such as:

- Which products drive the highest revenue across regions?
- Which stores are underperforming?
- Which products are nearing stockouts and need replenishment?

The leadership team seeks a centralized analytics platform powered by scalable data pipelines that can clean, integrate, and transform raw data into actionable insights while supporting future automation and real-time ingestion.

Data Sources

- sales.csv: transaction_id, store_id, product_id, quantity_sold, sale_date, price_sold
- inventory.csv: store_id, product_id, stock_quantity, last_updated
- stores.csv: store_id, store_name, region, city
- products.csv: product_id, product_name, category, cost_price

Objectives

As a Data Engineer, you are expected to:

1. Design and implement a scalable ETL pipeline for data cleaning, integration, and transformation.
2. Generate analytical outputs that provide business insights.
3. Propose a future-ready architecture (scalable, automated, and cloud-compatible).

Tasks

Task 1 – Data Cleaning & Transformation

- Remove duplicates and handle missing values.
- Standardize datetime fields (sale_date, last_updated).
- Normalize text fields (e.g., product/store names).

Task 2 – Data Integration

- Merge all datasets into a unified schema:
  transaction_id, store_name, city, region, product_name, category, quantity_sold, price_sold, stock_quantity, sale_date

Task 3 – Analytics & Aggregation

- Compute total sales per store (last month).
- Identify top 5 selling products.
- Flag products with low stock (<10) and propose reorder needs.
- *(Optional)* Calculate gross profit per product.

Task 4 – SQL Queries

- Sales by region.
- Top 3 stores by revenue.
- Number of unique products sold per category.

Task 5 – Pipeline & Architecture

- Design an ETL workflow diagram covering ingestion, cleaning, transformation, and aggregation.
- *(Optional)* Suggest cloud-based architecture (AWS/Azure/GCP) and real-time extension (Kafka/Spark Streaming).

Deliverables

- Python/PySpark/SQL scripts for all tasks.
- Aggregated output tables (CSV or DB dump).
- ETL/pipeline diagram.
- Documentation: approach, cleaning strategy, integration logic, architecture decisions.

Scoring Rubric

Criteria	Weightage	Evaluation Focus
Data Cleaning & Quality	20%	Correct handling of duplicates, missing values, formatting & standardization
Data Integration Accuracy	20%	Consistency & correctness of merged dataset, schema design
Analytics & Insights	20%	Accuracy of KPIs (sales, top products, low stock, profit)
Code Quality & Efficiency	15%	Readability, modularity, scalability of Python/SQL scripts
Pipeline Design & Scalability	15%	ETL workflow clarity, extensibility, cloud/real-time readiness
Documentation & Communication	10%	Clear explanation of approach, architecture justification, business alignment