Module 2 Project Assignment: 8-Aug-25.

User Story

As a business intelligence analyst at an e-commerce company,
I want to access clean, well-structured sales, customer, and product data in a single warehouse,
so that I can quickly generate reports and uncover trends without manually cleaning or joining multiple raw data files.

Problem Statement

The company's sales, customer, and product data is scattered across multiple raw CSV files, each with inconsistent formats, missing values, and duplicate entries. This fragmentation makes it time-consuming for analysts to create accurate reports, delays decision-making, and increases the risk of errors in business insights. There is no centralized, analytics-ready dataset to support performance tracking, customer segmentation, or product sales analysis.

Use Case

Title: Building an analytics-ready e-commerce data warehouse

Actors:

- Data Engineering Team (responsible for ingestion, transformation, and loading)
- Business Intelligence Analysts (consumers of cleaned data)
- Management Team (decision-makers using insights)

Scenario:

1. Data engineers ingest raw sales, customer, product, seller, and payment data from multiple CSV sources.
2. They design and implement a **star schema** in a central data warehouse.
3. ETL/ELT pipelines populate **dimension tables** (customers, products, sellers, dates) and a **fact table** (sales).
4. Data quality checks ensure referential integrity, remove duplicates, and validate numerical ranges.
5. Analysts query the warehouse to:
   - Track monthly sales trends.
   - Identify top-performing products and sellers.
   - Segment customers based on purchase behavior.
6. Management uses these insights to guide marketing spend, inventory planning, and seller partnerships.