EXL Round 2 – Mini Project Instructions

■ Objective

Build an end-to-end data pipeline using raw CSV files. This includes data cleaning, transformation, modeling, ETL, SQL analytics, and data quality checks.

■ Files Provided

- users.csv – Dirty user master data (nulls, duplicates, inconsistent casing)
- products.csv – Product data with messy categories and SKU formats
- orders.csv – 20k+ raw orders (mixed date formats, invalid numeric fields)
- order_items.csv – Item-level data (dirty SKUs, wrong quantities, duplicates)

■ Tasks Breakdown

- Data Cleaning & Transformation
- Build Star Schema (fact_orders, dim_users, dim_products, dim_date)
- Implement ETL Logic (PySpark)
- Write Analytical SQL Queries (6 mandatory queries)
- Create Data Quality Checks
- Prepare Final Report
- Record a 3-minute Explanation Video

■ Final Deliverables

- Cleaned datasets (CSV)
- PySpark Notebook
- Star Schema Diagram (ERD)
- SQL Query & Outputs
- Data Quality Summary
- 3-minute Video Link

■ Evaluation Rubric (100 Marks)

- 20 – Data Cleaning & Preparation
- 20 – Data Model Design
- 10 – ETL Pipeline Implementation
- 20 – SQL Analytics
- 10 – Data Quality Checks
- 20 – Video Explanation