The Brief: Building a Dockerized PySpark ETL Pipeline with Delta Tables

Objective

Your task is to create a simple ETL (Extract, Transform, Load) pipeline using PySpark, Docker, and Delta Tables. The goal is to read data from a CSV file, apply a basic transformation, and store the results in a Delta Table.
Requirements

Data Source:

- o Use the provided CSV file (data.csv) containing sample data (e.g., customer orders).

ETL Process:

- ▪ Read the data from the CSV file.
- ▪ Apply a transformation (e.g., calculate total order amount).
- ▪ Store the transformed data in a Delta Table.

Dockerization:

- • Create a Dockerfile to package your PySpark script and dependencies.
- • Build a Docker image.
- • Run the ETL process inside a Docker container.

Delta Table:

- o Initialize a Delta Table (you can use a local directory for simplicity).
- o Write the transformed data to the Delta Table.

Instructions

ii. Set up your development environment with Docker and PySpark.
iii. Write your PySpark script (**etl.py**) to perform the ETL process.
iv. Create a Dockerfile to package your script.
v. Build the Docker image.
vi. Run the ETL process inside a Docker container.
vii. Verify that the data is correctly stored in the Delta Table.