Project Domain: FinTech – Financial Transaction & Fraud Analysis

Task: Build a Dockerized Airflow environment, process local transaction data, and upload it to Google Cloud Storage (GCS).

Goal: Get a basic Airflow setup running in Docker, process a local CSV file using Python/Pandas, and upload the processed data to a GCS bucket.

Prohibited (Automatic Failure):

No Docker Desktop: Use Terminal/CLI only (Colima for Mac).

No Local Execution: Airflow and Python scripts must run inside Docker containers managed by Airflow.

No Managed Cloud: No Cloud Composer or Dataproc.

No Manual Uploads: All data movement/processing must be automated via code.

Required Stack:

Containers: Docker CLI, Docker Compose.

Orchestration: Apache Airflow (Dockerized).

Processing: Python / Pandas (Dockerized, run by Airflow).

Cloud: Google Cloud Storage (GCS).

Required Flow:

1. Utilize a FinTech – Financial Transaction Dataset, downloaded from Kaggle to your local machine.

2. Design and implement an automated pipeline using the required stack to achieve the stated goal.

3. Prove the successful execution and outcome of your pipeline.

Deliverables:

Confluence Page: Architecture diagram, setup steps, and an Error/Resolution log.

AI Disclosure: List all prompts used and how you validated the code.

Live Demo: Trigger the DAG and show data reflecting in BigQuery.