**Stock Data Pipeline with Apache Airflow and PostgreSQL**

This project implements a data pipeline to fetch stock market data from an external API and store it in a PostgreSQL database using Apache Airflow for orchestration. It is fully containerized using Docker Compose, making it easy to set up and run.

# Features

Fetches stock data from a configurable API endpoint

Validates and filters incomplete or malformed records

Efficient batch insertion with UPSERT to avoid duplicates

Error handling with retry logic and logging

Scalable via connection pooling and batch inserts

Easily deployable using Docker Compose

Airflow DAG for scheduled and repeatable tasks

**Project Structure**

├── dags/

│ └── stock\_dag.py # Airflow DAG definition

├── scripts/

│ └── fetch\_stock\_data.py # ETL logic for fetching and storing data ├── .env # API URL, DB credentials, batch size etc.

├── docker-compose.yml # Defines services: Airflow, Postgres

└── README.md

# Setup Instructions

1. **Prerequisites**

Docker & Docker Compose installed

Port 5432 and 8080 available

API that returns stock data in the expected format (symbol, price, date)

1. **Configure Environment**

Create a .env file in the root directory:

API\_URL=<https://example.com/stock-api>DB\_HOST=postgres DB\_NAME=stocks DB\_USER=airflow

DB\_PASSWORD=airflow DB\_PORT=5432

MAX\_RETRIES=3

RETRY\_DELAY=5

BATCH\_SIZE=500

1. **Build and Run with Docker Compose**

docker-compose up --build

This will:

Start PostgreSQL and Airflow

Initialize the Airflow database

Create the Airflow admin user

Launch the webserver on [http://localhost:8080](http://localhost:8080/)

# Scheduling with Airflow

Once running, open your browser:

[http://localhost:8080](http://localhost:8080/) Username: airflow

Password: airflow

Enable the DAG named: stock\_data\_pipeline

This DAG is set to run hourly and will:

1. Fetch data from the API
2. Validate and filter it
3. Insert it into the stocktable with conflict resolution

# Testing the Pipeline

To manually trigger the pipeline:

1. Go to the Airflow UI
2. Turn the DAG on
3. Click the Trigger DAG button

Logs for each task can be viewed in the UI under each task instance.

# Database Schema

The data is stored in the stock\_data table:

CREATE TABLE stock\_data ( symbol TEXT, price NUMERIC, date DATE,

PRIMARY KEY (symbol, date)

);