Stock Data Pipeline with Airflow, PostgreSQL, and Visualization

Introduction

In this project, I built a data pipeline to automatically fetch and store daily stock prices for Apple Inc. (AAPL). The pipeline uses Apache Airflow to schedule and execute data fetching tasks, PostgreSQL as the storage backend, and Python with matplotlib for data visualization. This system enables continuous updating of stock data and easy visual analysis. This was a great learning opportunity to understand how to use Airflow to collect stock data automatically, save it in a PostgreSQL database, and then visualize the data using Python (as we learned in the Data Engineering and Wrangling module). I also gained experience using pgAdmin to manage the database easily and learned how to troubleshoot common connection and password issues. Overall, this project helped me understand how different tools work together to build a simple data pipeline from start to finish. It gave me hands-on experience with real-world tools, which enhanced my learning during the semester.

Project Workflow and Implementation

1. Data Pipeline with Airflow

- I created an Airflow DAG (stock_dag.py) that runs daily.

- The DAG downloads stock data using the yfinance library.

- Data includes daily open, high, low, close, and volume.

- The data is then inserted into a PostgreSQL database in a table named stock_prices.

- The DAG avoids inserting duplicate records by using PostgreSQL's ON CONFLICT DO NOTHING feature.

2. Database Setup and Management

- PostgreSQL serves as the relational database to store stock data.

- The database user airflow with password airflow connects to a database named stockdb.

- I used pgAdmin, a graphical interface, to manage and inspect the database.

- In pgAdmin, I encountered password authentication issues, which I resolved by resetting the user password using SQL commands executed as a superuser.

3. Data Visualization

- To visualize the stored stock data, I wrote a Python script (plot_stock.py) that:

    o Connects to the PostgreSQL database.

- o Queries historical closing prices ordered by date.
  - o Plots the closing prices over time using matplotlib.
- This visualization confirms the data is correctly fetched and stored.

Reflection

Working through this project helped me understand the full lifecycle of a data pipeline from data ingestion, storage, to consumption and visualization. Here are some key takeaways:

- **Airflow's power** lies in automating workflows, but configuration (like database connectivity) requires attention to detail.

- **Database authentication and access** can be tricky, especially managing credentials and access permissions. Using tools like pgAdmin simplifies exploration but may require troubleshooting connection issues.

- **Visualization scripts** are essential for quick validation and insights from stored data.

- Handling environment-specific issues (like hostname differences between Docker containers and local machines) is a crucial practical skill.

- Learning to install and troubleshoot Python packages (psycopg2, matplotlib, pandas) was important to move from backend data handling to frontend visualization.

Figure 1 Docker Containers

Figure 2 docker exec -it stock-pipeline-postgres-1 psql -U airflow -d stockdb

```
stockdb=# CREATE TABLE IF NOT EXISTS stock_prices (
    date DATE PRIMARY KEY,
    open FLOAT,
    high FLOAT,
    low FLOAT,
    close FLOAT,
    volume BIGINT
);
CREATE TABLE
stockdb=# \q
PS C:\Users\vivie> |
```
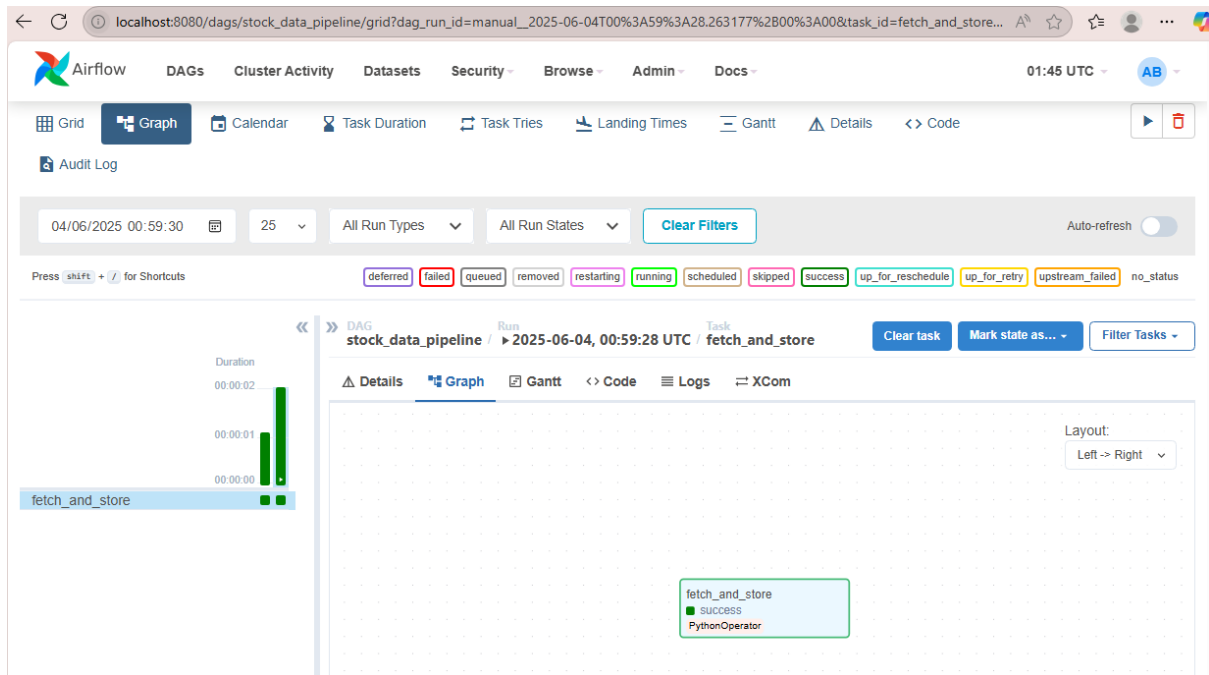
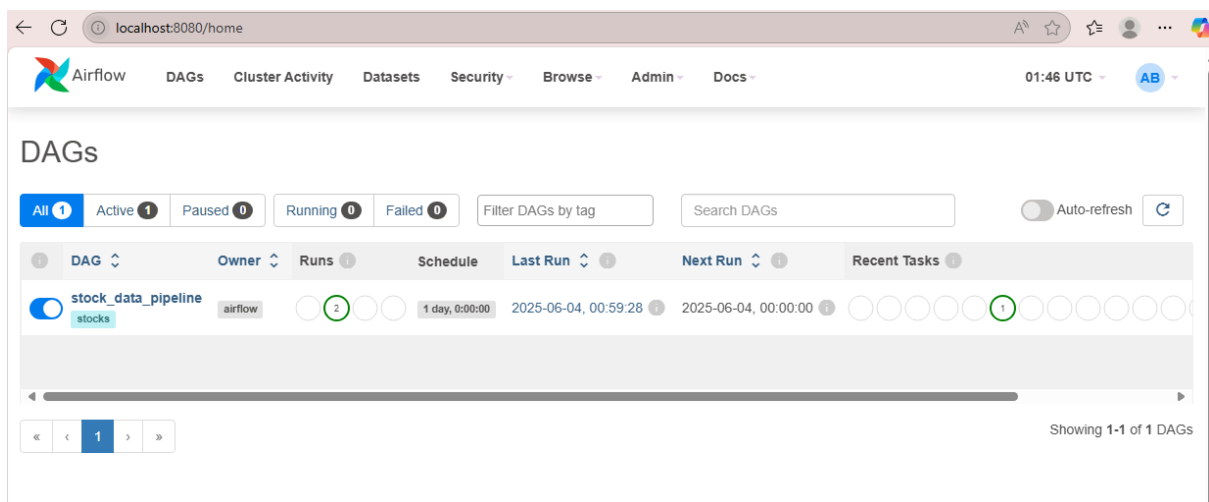*Figure 3 stock_prices table*

Figure 4 Airflow DAG

Figure 5 Airflow DAGs overview

First Run airflow, db

```
PS C:\Users\vivie> cd C:\Users\vivie\stock_data_pipeline\stock-pipeline
```

```
PS C:\Users\vivie\stock_data_pipeline\stock-pipeline> docker-compose run airflow
airflow db init

time="2025-06-04T01:58:06+02:00" level=warning
msg="C:\\Users\\vivie\\stock_data_pipeline\\stock-pipeline\\docker-compose.yml: the
attribute `version` is obsolete, it will be ignored, please remove it to avoid potential
confusion"

[+] Creating 1/1

 ✔ Container stock-pipeline-postgres-1  Running                                        0.0s


/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812
DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the
sql_alchemy_conn option in [database] - the old setting has been used, but please
update your config.

/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:738
DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the
sql_alchemy_conn option in [database] - the old setting has been used, but please
update your config.

/home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:194
DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the
sql_alchemy_conn option in [database] - the old setting has been used, but please
update your config.

/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:71
DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the
sql_alchemy_conn option in [database] - the old setting has been used, but please
update your config.

/home/airflow/.local/lib/python3.9/site-
packages/airflow/cli/commands/db_command.py:47 DeprecationWarning: `db init` is
deprecated.  Use `db migrate` instead to migrate the db and/or airflow connections
create-default-connections to create the default connections

DB: postgresql+psycopg2://airflow:***@postgres/stockdb

[2025-06-03T23:58:14.163+0000] {migration.py:216} INFO - Context impl
PostgresqlImpl.

[2025-06-03T23:58:14.164+0000] {migration.py:219} INFO - Will assume transactional
DDL.

INFO  [alembic.runtime.migration] Context impl PostgresqlImpl.
```

```
INFO  [alembic.runtime.migration] Will assume transactional DDL.

INFO  [alembic.runtime.migration] Running stamp_revision  -> 88344c1d9134

WARNI [airflow.models.crypto] empty cryptography key - values will not be stored
encrypted.

Initialization done

PS C:\Users\vivie\stock_data_pipeline\stock-pipeline> docker-compose down

time="2025-06-04T01:58:23+02:00" level=warning
msg="C:\\Users\\vivie\\stock_data_pipeline\\stock-pipeline\\docker-compose.yml: the
attribute `version` is obsolete, it will be ignored, please remove it to avoid potential
confusion"

[+] Running 3/3

 ✓ Container stock-pipeline-airflow-1   Removed                              0.1s

 ✓ Container stock-pipeline-postgres-1  Removed                              0.7s

 ✓ Network stock-pipeline_default     Removed                              0.8s

PS C:\Users\vivie\stock_data_pipeline\stock-pipeline> docker-compose up --build

time="2025-06-04T01:58:29+02:00" level=warning
msg="C:\\Users\\vivie\\stock_data_pipeline\\stock-pipeline\\docker-compose.yml: the
attribute `version` is obsolete, it will be ignored, please remove it to avoid potential
confusion"

time="2025-06-04T01:58:29+02:00" level=warning msg="Found orphan containers
([stock-pipeline-airflow-run-876bea568647]) for this project. If you removed or renamed
this service in your compose file, you can run this command with the --remove-orphans
flag to clean it up."

[+] Running 3/3

 ✓ Network stock-pipeline_default     Created                              0.0s

 ✓ Container stock-pipeline-postgres-1  Created                              0.2s

 ✓ Container stock-pipeline-airflow-1   Created                              0.1s

Attaching to airflow-1, postgres-1

postgres-1  |

postgres-1  | PostgreSQL Database directory appears to contain a database; Skipping
initialization
```

```
postgres-1  |

postgres-1  | 2025-06-03 23:58:30.265 UTC [1] LOG:  starting PostgreSQL 13.21 (Debian
13.21-1.pgdg120+1) on x86_64-pc-linux-gnu, compiled by gcc (Debian 12.2.0-14)
12.2.0, 64-bit

postgres-1  | 2025-06-03 23:58:30.265 UTC [1] LOG:  listening on IPv4 address "0.0.0.0",
port 5432

postgres-1  | 2025-06-03 23:58:30.265 UTC [1] LOG:  listening on IPv6 address "::", port
5432

postgres-1  | 2025-06-03 23:58:30.345 UTC [1] LOG:  listening on Unix socket
"/var/run/postgresql/.s.PGSQL.5432"

postgres-1  | 2025-06-03 23:58:30.356 UTC [27] LOG:  database system was shut down
at 2025-06-03 23:58:23 UTC

postgres-1  | 2025-06-03 23:58:30.363 UTC [1] LOG:  database system is ready to accept
connections

airflow-1   |

airflow-1   | /home/airflow/.local/lib/python3.9/site-
packages/airflow/configuration.py:812 DeprecationWarning: The sql_alchemy_conn
option in [core] has been moved to the sql_alchemy_conn option in [database] - the old
setting has been used, but please update your config.

airflow-1   | /home/airflow/.local/lib/python3.9/site-
packages/airflow/configuration.py:738 DeprecationWarning: The sql_alchemy_conn
option in [core] has been moved to the sql_alchemy_conn option in [database] - the old
setting has been used, but please update your config.

airflow-1   | /home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:194
DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the
sql_alchemy_conn option in [database] - the old setting has been used, but please
update your config.

airflow-1   | Defaulting to user installation because normal site-packages is not
writeable

airflow-1   | /home/airflow/.local/lib/python3.9/site-
packages/airflow/models/base.py:71 DeprecationWarning: The sql_alchemy_conn
option in [core] has been moved to the sql_alchemy_conn option in [database] - the old
setting has been used, but please update your config.
```

```
airflow-1  | Requirement already satisfied: pandas in
/home/airflow/.local/lib/python3.9/site-packages (from -r /requirements.txt (line 1))
(2.1.4)

airflow-1  | Collecting yfinance (from -r /requirements.txt (line 2))

airflow-1  |   Downloading yfinance-0.2.61-py2.py3-none-any.whl.metadata (5.8 kB)

airflow-1  | Requirement already satisfied: psycopg2-binary in
/home/airflow/.local/lib/python3.9/site-packages (from -r /requirements.txt (line 3))
(2.9.9)

airflow-1  | Requirement already satisfied: numpy<2,>=1.22.4 in
/home/airflow/.local/lib/python3.9/site-packages (from pandas->-r /requirements.txt
(line 1)) (1.24.4)

airflow-1  | Requirement already satisfied: python-dateutil>=2.8.2 in
/home/airflow/.local/lib/python3.9/site-packages (from pandas->-r /requirements.txt
(line 1)) (2.8.2)

airflow-1  | Requirement already satisfied: pytz>=2020.1 in
/home/airflow/.local/lib/python3.9/site-packages (from pandas->-r /requirements.txt
(line 1)) (2023.3.post1)

airflow-1  | Requirement already satisfied: tzdata>=2022.1 in
/home/airflow/.local/lib/python3.9/site-packages (from pandas->-r /requirements.txt
(line 1)) (2023.4)

airflow-1  | Requirement already satisfied: requests>=2.31 in
/home/airflow/.local/lib/python3.9/site-packages (from yfinance->-r /requirements.txt
(line 2)) (2.31.0)

airflow-1  | Collecting multitasking>=0.0.7 (from yfinance->-r /requirements.txt (line 2))

airflow-1  |   Downloading multitasking-0.0.11-py3-none-any.whl.metadata (5.5 kB)

airflow-1  | Requirement already satisfied: platformdirs>=2.0.0 in
/home/airflow/.local/lib/python3.9/site-packages (from yfinance->-r /requirements.txt
(line 2)) (3.11.0)

airflow-1  | Collecting frozendict>=2.3.4 (from yfinance->-r /requirements.txt (line 2))

airflow-1  |   Downloading frozendict-2.4.6-cp39-cp39-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (23 kB)

airflow-1  | Collecting peewee>=3.16.2 (from yfinance->-r /requirements.txt (line 2))

airflow-1  |   Downloading peewee-3.18.1.tar.gz (3.0 MB)
```

airflow-1 | [2025-06-03T23:58:36.944+0000] {configuration.py:2065} INFO - Creating new FAB webserver config file in: /opt/airflow/webserver_config.py

airflow-1 | ━━ ━━━━━━━━━━━━━━━━━━ 3.0/3.0 MB 14.5 MB/s eta 0:00:00

airflow-1 | Installing build dependencies: started

airflow-1 | /home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.

airflow-1 | /home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:738 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.

airflow-1 | /home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:194 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.

airflow-1 | /home/airflow/.local/lib/python3.9/site-packages/airflow/www/app.py:86 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.

airflow-1 | /home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:71 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.

airflow-1 | /home/airflow/.local/lib/python3.9/site-packages/flask_limiter/extension.py:336 UserWarning: Using the in-memory storage for tracking rate limits as no storage was explicitly specified. This is not recommended for production use. See: https://flask-limiter.readthedocs.io#configuring-a-storage-backend for documentation about configuring the storage backend.

airflow-1 | [2025-06-03T23:58:39.105+0000] {override.py:1369} INFO - Inserted Role: Admin

airflow-1 | [2025-06-03T23:58:39.111+0000] {override.py:1369} INFO - Inserted Role: Public

airflow-1 | [2025-06-03T23:58:39.113+0000] {override.py:868} WARNING - No user yet created, use flask fab command to do it.

airflow-1 | [2025-06-03T23:58:39.144+0000] {override.py:1769} INFO - Created Permission View: can edit on Passwords

airflow-1 | [2025-06-03T23:58:39.150+0000] {override.py:1820} INFO - Added Permission can edit on Passwords to role Admin

airflow-1 | [2025-06-03T23:58:39.157+0000] {override.py:1769} INFO - Created Permission View: can read on Passwords

airflow-1 | [2025-06-03T23:58:39.162+0000] {override.py:1820} INFO - Added Permission can read on Passwords to role Admin

airflow-1 | [2025-06-03T23:58:39.174+0000] {override.py:1769} INFO - Created Permission View: can edit on My Password

airflow-1 | [2025-06-03T23:58:39.179+0000] {override.py:1820} INFO - Added Permission can edit on My Password to role Admin

airflow-1 | [2025-06-03T23:58:39.185+0000] {override.py:1769} INFO - Created Permission View: can read on My Password

airflow-1 | [2025-06-03T23:58:39.190+0000] {override.py:1820} INFO - Added Permission can read on My Password to role Admin

airflow-1 | [2025-06-03T23:58:39.202+0000] {override.py:1769} INFO - Created Permission View: can edit on My Profile

airflow-1 | [2025-06-03T23:58:39.207+0000] {override.py:1820} INFO - Added Permission can edit on My Profile to role Admin

airflow-1 | [2025-06-03T23:58:39.213+0000] {override.py:1769} INFO - Created Permission View: can read on My Profile

airflow-1 | [2025-06-03T23:58:39.218+0000] {override.py:1820} INFO - Added Permission can read on My Profile to role Admin

airflow-1 | [2025-06-03T23:58:39.246+0000] {override.py:1769} INFO - Created Permission View: can create on Users

airflow-1 | [2025-06-03T23:58:39.251+0000] {override.py:1820} INFO - Added Permission can create on Users to role Admin

airflow-1 | [2025-06-03T23:58:39.257+0000] {override.py:1769} INFO - Created Permission View: can read on Users

```
airflow-1   | [2025-06-03T23:58:39.263+0000] {override.py:1820} INFO - Added
Permission can read on Users to role Admin

airflow-1   | [2025-06-03T23:58:39.268+0000] {override.py:1769} INFO - Created
Permission View: can edit on Users

airflow-1   | [2025-06-03T23:58:39.273+0000] {override.py:1820} INFO - Added
Permission can edit on Users to role Admin

airflow-1   | [2025-06-03T23:58:39.281+0000] {override.py:1769} INFO - Created
Permission View: can delete on Users

airflow-1   | [2025-06-03T23:58:39.285+0000] {override.py:1820} INFO - Added
Permission can delete on Users to role Admin

airflow-1   | [2025-06-03T23:58:39.297+0000] {override.py:1769} INFO - Created
Permission View: menu access on List Users

airflow-1   | [2025-06-03T23:58:39.302+0000] {override.py:1820} INFO - Added
Permission menu access on List Users to role Admin

airflow-1   | [2025-06-03T23:58:39.313+0000] {override.py:1769} INFO - Created
Permission View: menu access on Security

airflow-1   | [2025-06-03T23:58:39.317+0000] {override.py:1820} INFO - Added
Permission menu access on Security to role Admin... (35 KB left)
```

Problems with storing the data

AttributeError: 'Series' object has no attribute 'date'

https://chatgpt.com/share/683f92f3-b0cc-800c-9ff6-c209f3067a51
This conversation with ChatGPT helped me further understand the Error and fix it in my stock_dag.py file.