BÁO CÁO BÀI TEST

1. Yêu cầu bài toán

- Đầu vào: 1 file Excel có 4 sheet dữ liệu (Q1, Q2, Q3, Q4).
- Thiết kế: Lưu toàn bộ dữ liệu gộp vào **1 bảng duy nhất trong MySQL**.
- ETL: Dùng **Python + Airflow** để đọc, chuẩn hoá và ghi dữ liệu.
- Tự động hoá: Lên lịch Airflow chạy job hàng ngày lúc **07:00 sáng**.
- Kết quả:
    - Source code (Python DAG + cấu hình).
    - Mô tả flow đang chạy.
    - Cách kiểm tra job thành công (Airflow UI, log, MySQL).

2. Thiết kế CSDL

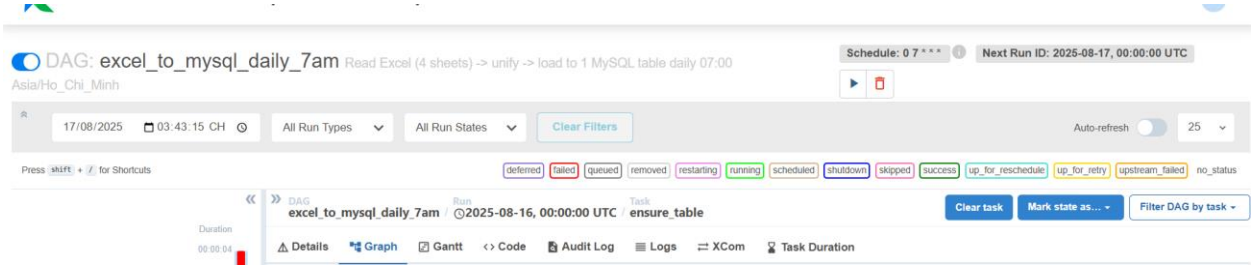
- Sử dụng MySQL, database de_test.
- Tạo bảng excel_combined có các trường chính:

```
CREATE TABLE excel_combined (
   id INT AUTO_INCREMENT PRIMARY KEY,
   source_sheet VARCHAR(50),
   load_date DATE,
   customer_id INT,
   order_date DATETIME,
   order_id VARCHAR(50),
   product_id INT,
   quantity INT,
   region VARCHAR(50),
   status VARCHAR(50),
   unit_price DECIMAL(10,2)
);
```

3. Flow ETL bằng Airflow

3.1 DAGs trong Airflow

DAG excel_to_mysql_daily_7am được bật và có lịch chạy mỗi ngày 07:00.

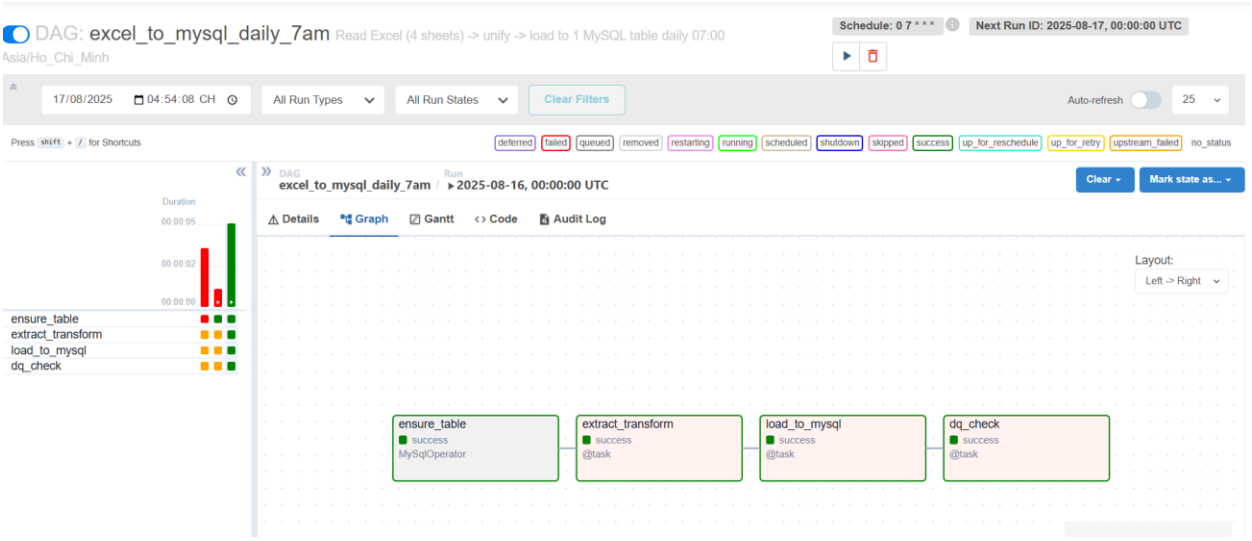


Hình 1. DAG Airflow bật ON

3.2 Cấu trúc DAG

Các task chính trong DAG:

- **extract_transform** → Đọc Excel, chuẩn hoá dữ liệu.
- **create_table** → Tạo bảng MySQL nếu chưa tồn tại.
- **load_to_mysql** → Insert dữ liệu vào bảng.
- **verify** → Kiểm tra số dòng sau khi load.

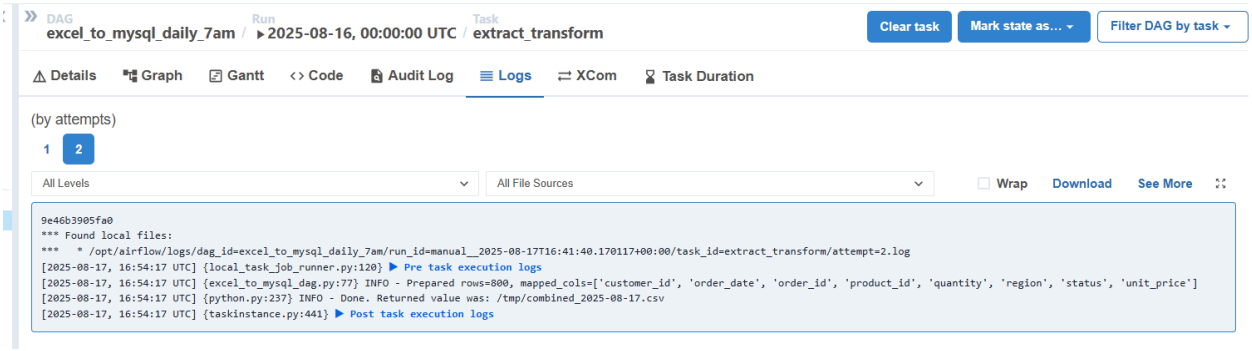


Hình 2. Graph view DAG

4. Kết quả chạy thực tế

4.1 Log Airflow

Trong log có thể thấy dữ liệu được đọc thành công từ 4 sheet, tổng cộng 800 dòng.



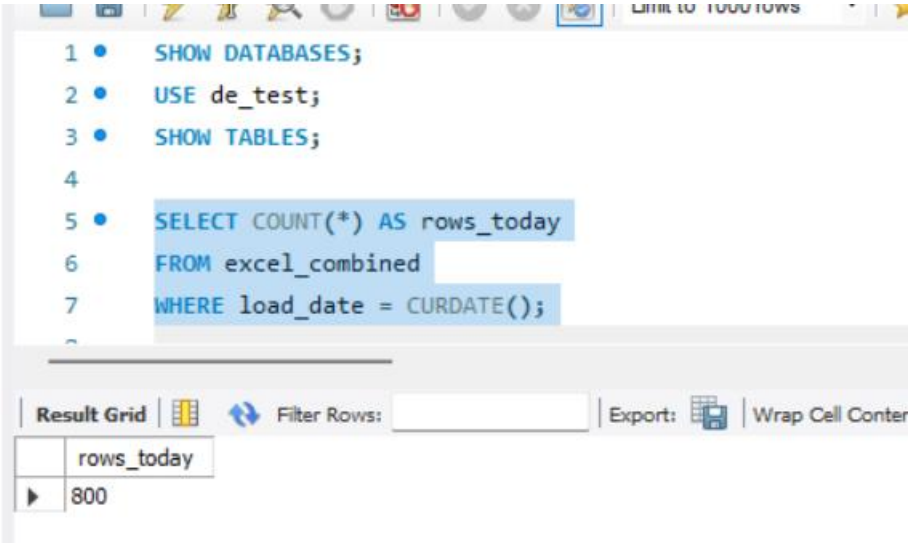
Hình 3. Log extract_transform

4.2 Kiểm tra MySQL

Sau khi load xong, kiểm tra trong MySQL:

Số bản ghi hôm nay:

SELECT COUNT(*) FROM excel_combined WHERE load_date = CURDATE();

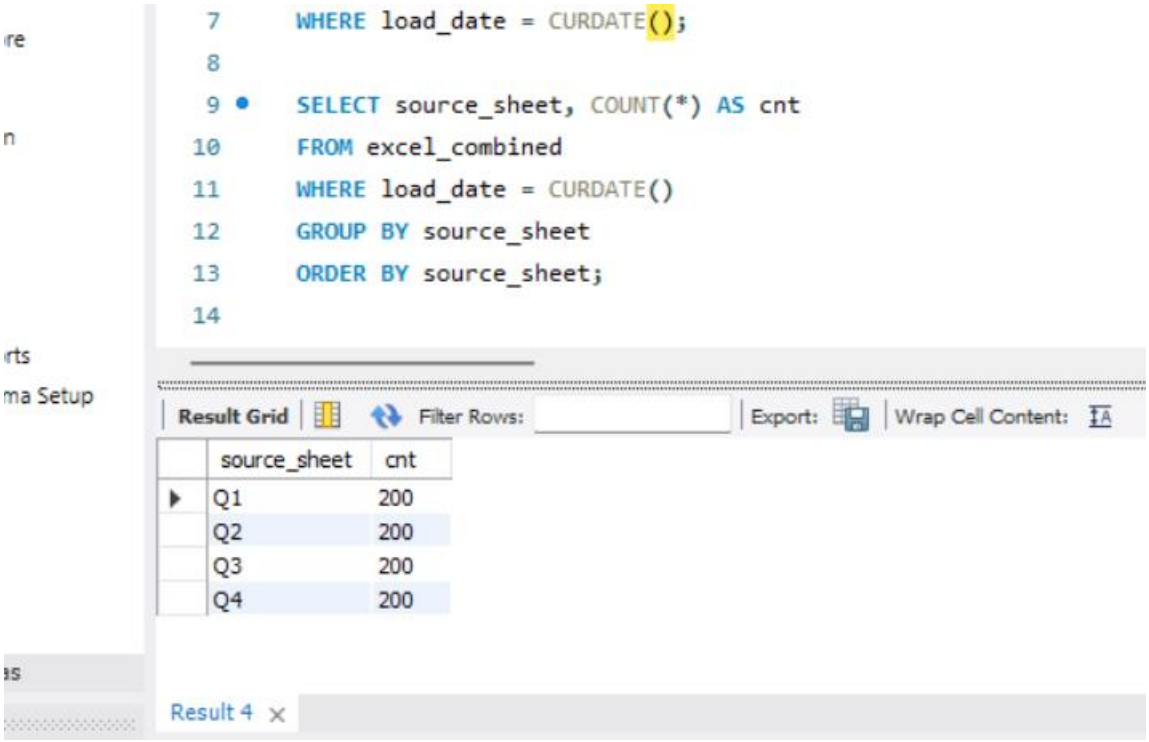


Hình 4. Kết quả query MySQL 1

Kết quả: **800 rows**.

Theo từng sheet:

```
SELECT source_sheet, COUNT(*)
FROM excel_combined
WHERE load_date = CURDATE()
GROUP BY source_sheet;
```

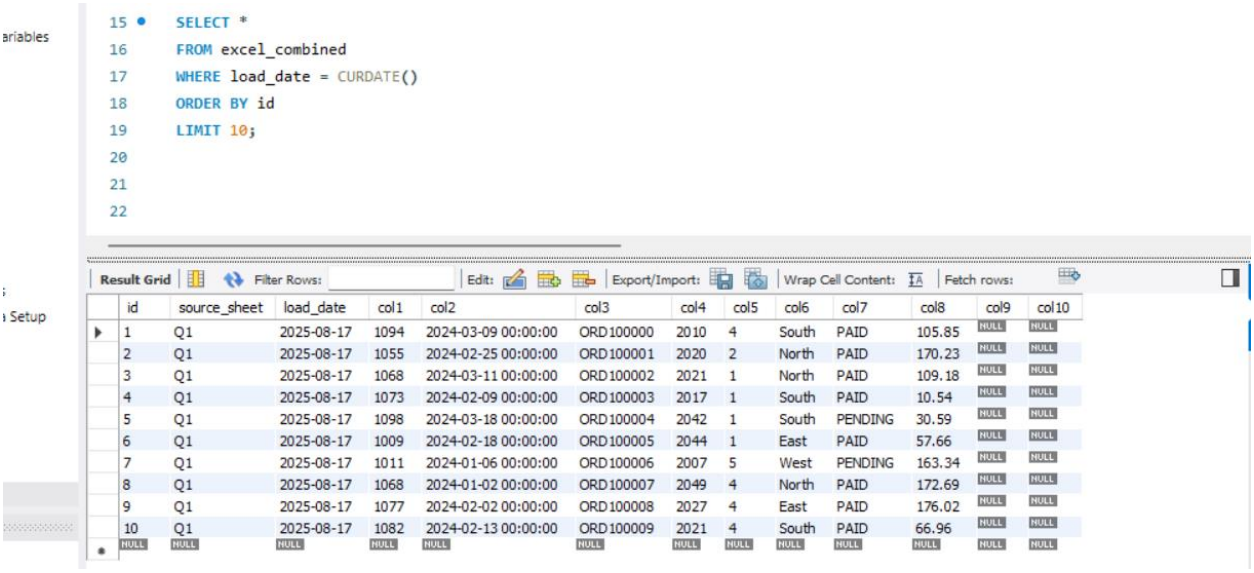


Hình 5. Kết quả query MySQL 2

Kết quả: Q1=200, Q2=200, Q3=200, Q4=200.

Xem 10 dòng đầu:

SELECT * FROM excel_combined LIMIT 10;



Hình 6. Kết quả query MySQL 3

5. Cách kiểm tra job

- Vào Airflow UI (localhost:8081).
- Check DAG excel_to_mysql_daily_7am → Tab **Graph** → tất cả task màu xanh = chạy thành công.
- Vào **Log** từng task để xem chi tiết.
- Mở MySQL Workbench hoặc dùng terminal query kiểm tra dữ liệu.

6. Kết luận

- ETL pipeline đã được thiết kế và chạy thành công: đọc file Excel nhiều sheet → load MySQL → xác thực dữ liệu.
- DAG đã được scheduling tự động chạy mỗi ngày 07:00 sáng.