MỤC LỤC

[**TỔNG QUAN DỰ ÁN** 2](#_Toc199177414)

[1. Mục tiêu dự án 2](#_Toc199177415)

[2. Mô hình dữ liệu (Star Schema) 2](#_Toc199177416)

[3. Xử lý thay đổi dữ liệu 2](#_Toc199177417)

[4. Kiến trúc Medallion 2](#_Toc199177418)

[5. Công nghệ & dịch vụ Azure 3](#_Toc199177419)

[6. Nguồn dữ liệu 3](#_Toc199177420)

[**CHUẨN BỊ HẠ TẦNG DEVELOPMENT** 4](#_Toc199177421)

[1. Resource Group 4](#_Toc199177422)

[2. Storage Account 4](#_Toc199177423)

[3. Azure Databricks 5](#_Toc199177424)

[4. Service Principal 6](#_Toc199177425)

[**PIPELINE XỬ LÝ DỮ LIỆU** 7](#_Toc199177426)

[1. Ingest (Bronze Layer) 7](#_Toc199177427)

[2. Transform (Silver Layer) 11](#_Toc199177428)

[3. Load (Gold Layer) 17](#_Toc199177429)

[**CI/CD VÀ TỰ ĐỘNG HÓA** 25](#_Toc199177430)

[1. Chuẩn bị hạ tầng Production 25](#_Toc199177431)

[2. Kết nối Git với Workspace Dev 27](#_Toc199177432)

[3. Thiết lập CI/CD Production 28](#_Toc199177433)

[4. Tự động hóa Lịch Chạy Job 28](#_Toc199177434)

[5. Kết nối Power BI với Azure Databricks 29](#_Toc199177435)

## **TỔNG QUAN DỰ ÁN**

### Mục tiêu dự án

* Tự động thu thập thông tin phim mới phát hành và phim đang thịnh hành mỗi tuần.
* Hỗ trợ người xem đưa ra gợi ý nhanh chóng, giúp lựa chọn phim phù hợp với thể loại yêu thích.

### Mô hình dữ liệu (Star Schema)

* Bảng Dimension
* dim\_movie – Thông tin phim (ID, tiêu đề, thời lượng…)
* dim\_genre – Thể loại phim
* dim\_company – Công ty sản xuất
* dim\_director – Đạo diễn
* dim\_date – Tuần phân tích
* Bảng Fact
* fact\_movies – Ghi nhận lượt xem, đánh giá, tương tác, liên kết các bảng dimension

### Xử lý thay đổi dữ liệu

* CDC (Change Data Capture): Chỉ ingest dữ liệu mới hoặc đã thay đổi dựa trên timestamp.
* SCD (Slowly Changing Dimensions): Sử dụng Type 1 (ghi đè) để quản lý lịch sử dữ liệu dimension.

### Kiến trúc Medallion

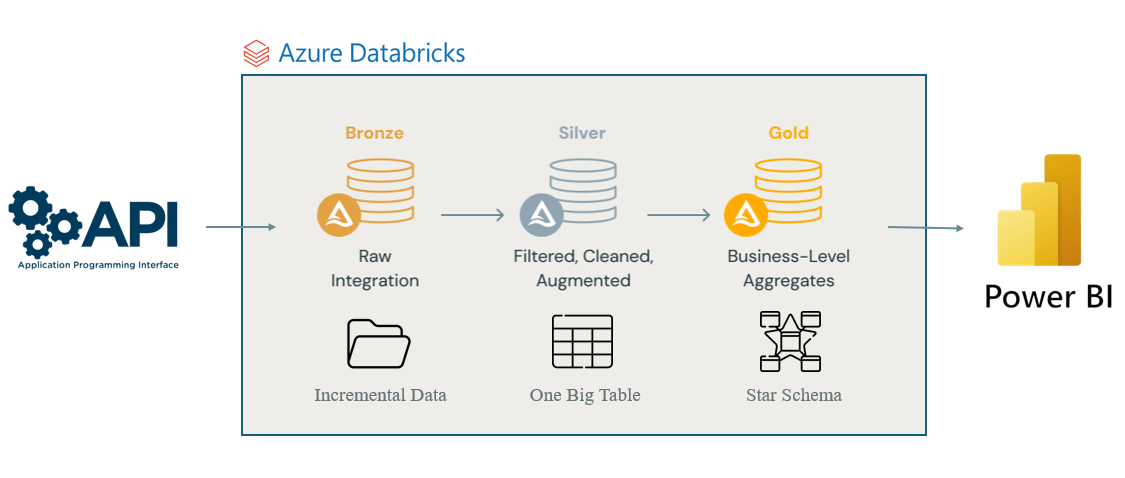
* Bronze – Dữ liệu thô ingest trực tiếp từ API (phim mới, phim xu hướng, thể loại)
* Silver – Dữ liệu đã làm sạch, chuẩn hóa schema
* Gold: Dữ liệu tổng hợp, sẵn sàng cho báo cáo và trực quan hóa.

### Công nghệ & dịch vụ Azure

* Azure Data Lake Storage Gen2 – Tạo 3 container (bronze, silver, gold).
* Azure Databricks – Sử dụng Delta Lake để lưu trữ, Databricks SQL để biến đổi.
* Quản lý bảo mật – Service Principal để cấp quyền truy cập an toàn cho Databricks.
* CI/CD – Sử dụng GitHub Actions để tự động deploy notebooks giữa môi trường Dev/Prod.

### Nguồn dữ liệu

* The Movie Database (TMDB) API: <https://www.themoviedb.org/>
* Cung cấp metadata phim: ngày phát hành, poster, lượt bình luận, điểm đánh giá…
* Hỗ trợ truy vấn incremental theo ngày/tuần để tối ưu băng thông và thời gian xử lý.

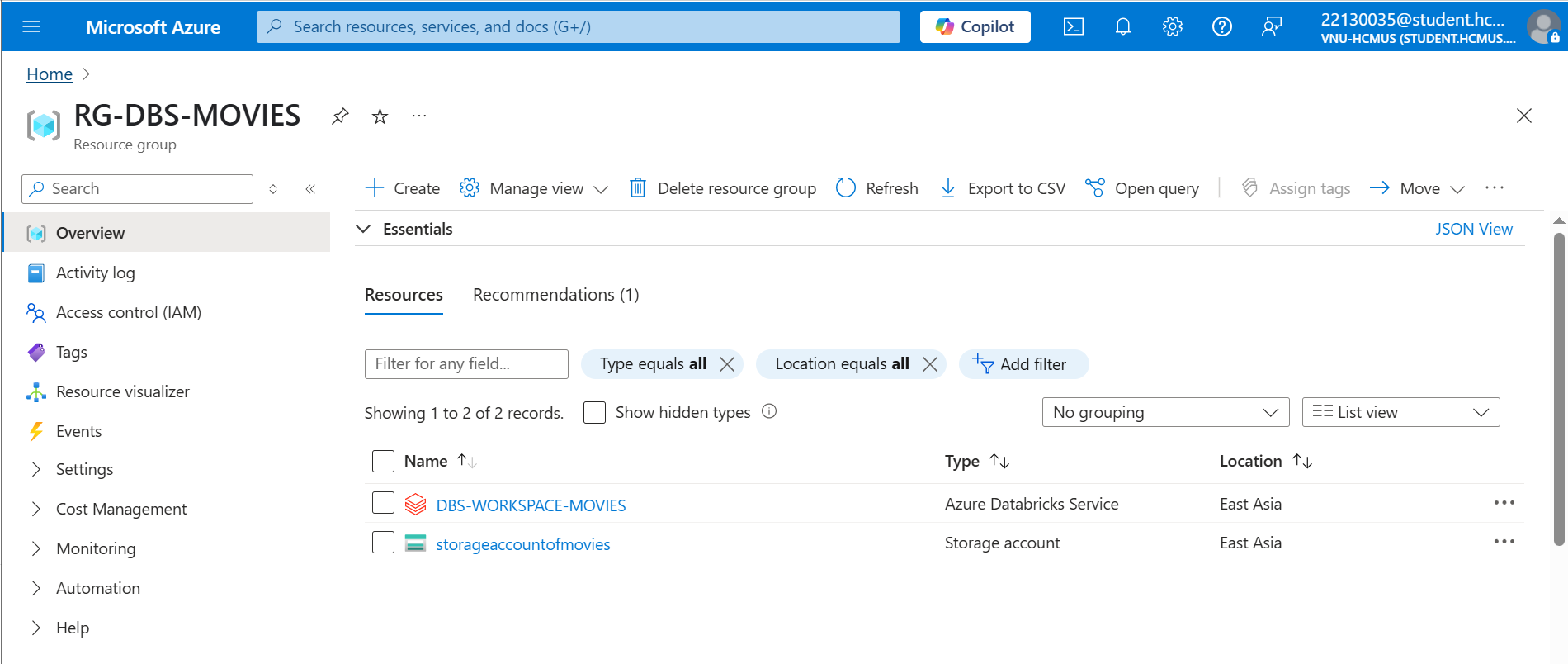


## **CHUẨN BỊ HẠ TẦNG DEVELOPMENT**

Trước khi triển khai pipeline, chúng ta cần thiết lập đầy đủ các tài nguyên Azure để đảm bảo môi trường sẵn sàng cho việc phát triển và vận hành.

### Resource Group

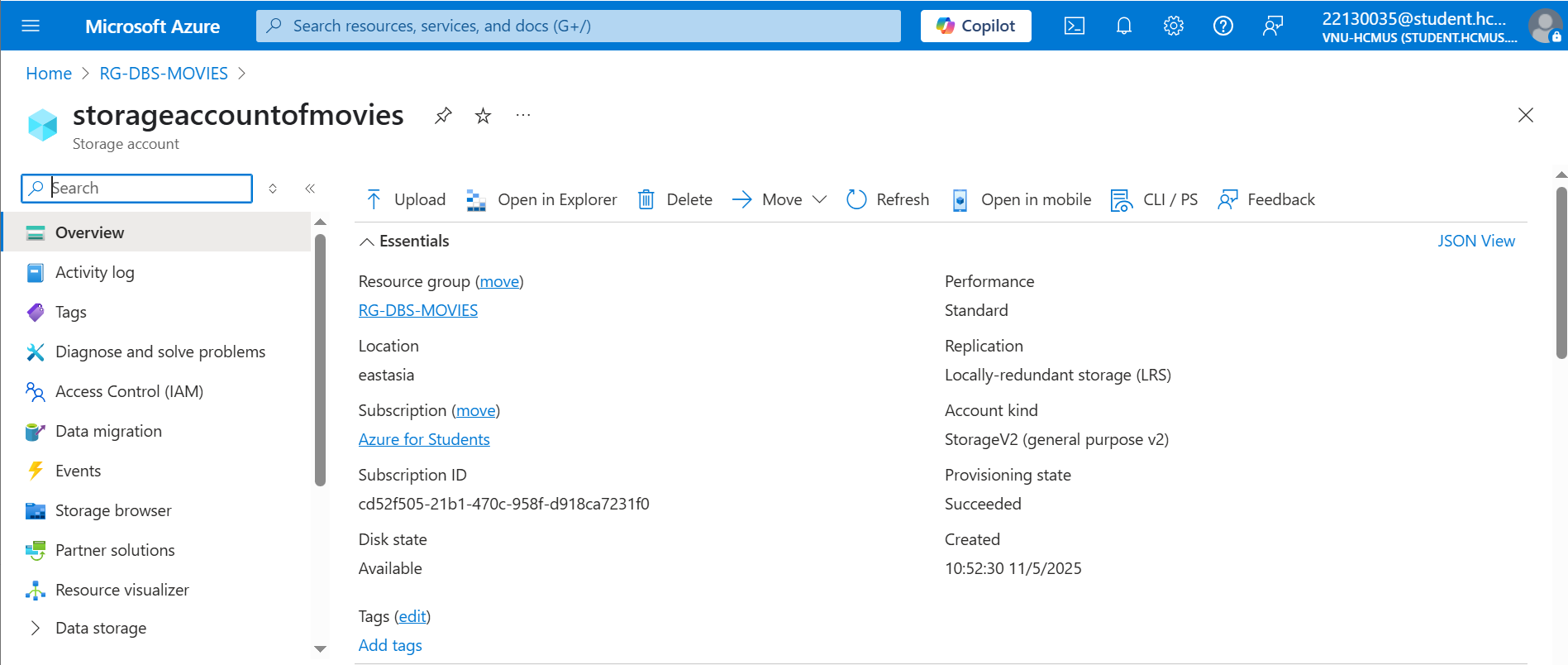
Tạo một Resource Group



Resource Group “RG-DBS-MOVIES” là vùng chứa tập trung các tài nguyên liên quan đến dự án, giúp chúng ta dễ dàng quản lý, theo dõi chi phí và quyền truy cập.

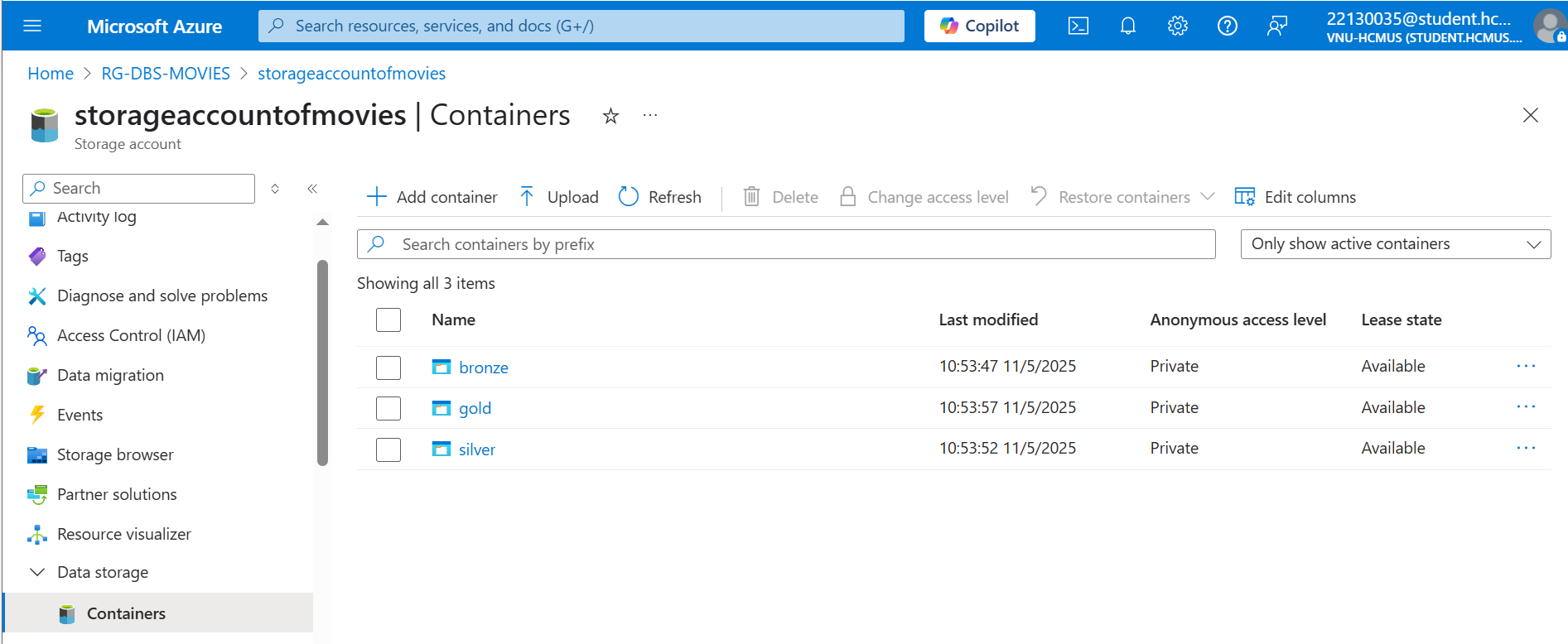
### Storage Account

Tạo một Storage Account



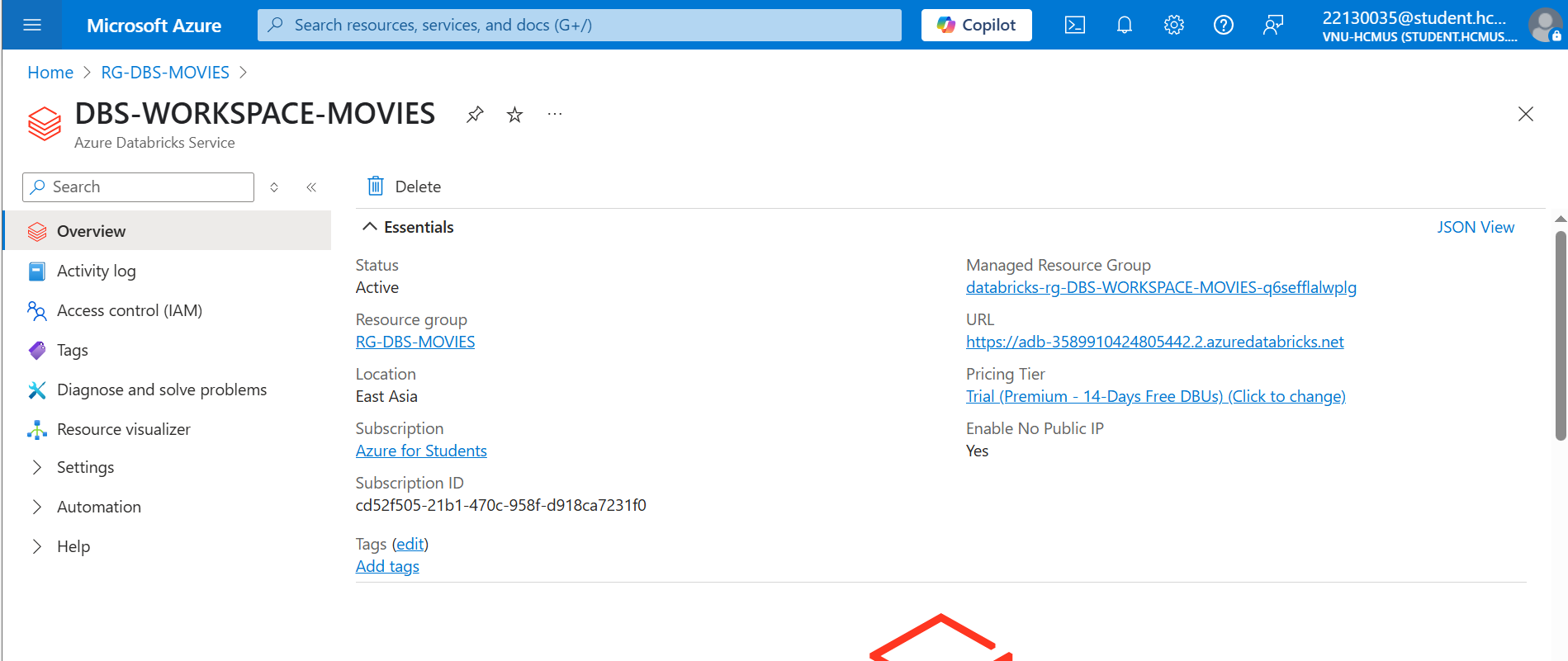
Storage Account “storageaccountofmovies” chịu trách nhiệm lưu trữ dữ liệu thô, trung gian và kết quả cuối cùng.

Tạo các Containers trong Data Lake



### Azure Databricks

Tạo một Azure Databricks Workspace

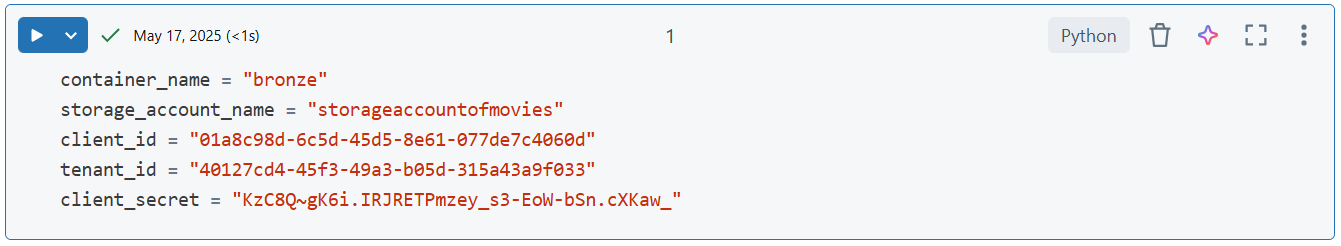


Workspace “DBS-WORKSPACE-MOVIES” là nơi chạy các notebook và job xử lý dữ liệu. Chúng ta sẽ cấu hình cluster và kết nối tới Data Lake Storage Gen 2.

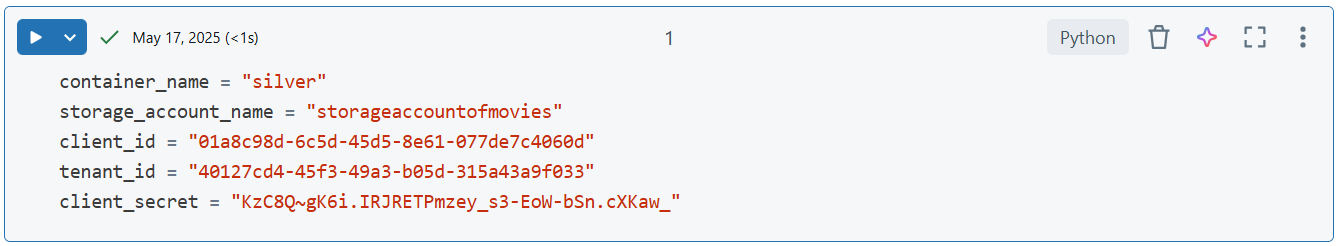
### Service Principal

Tạo Service Principal và cấp quyền Reader/Contributor cho Databricks tới Storage Account, đảm bảo authentication an toàn khi Spark job truy xuất dữ liệu.

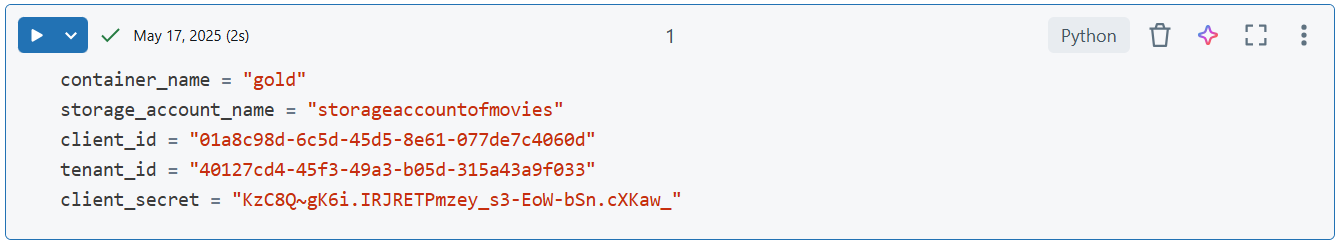
Bronze Layer



Silver Layer



Gold Layer

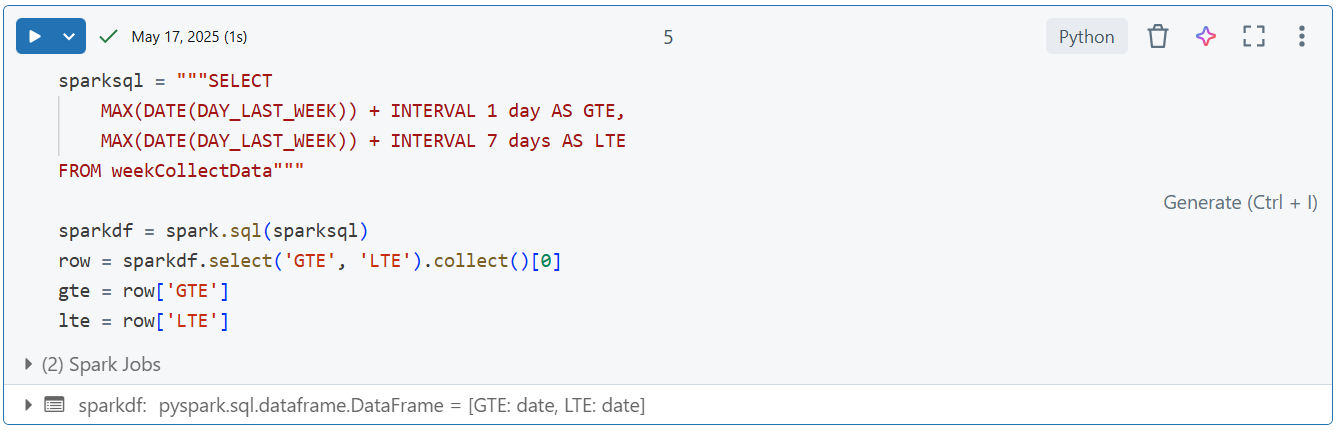


## **PIPELINE XỬ LÝ DỮ LIỆU**

### Ingest (Bronze Layer)

* Xác định khoảng thời gian tuần

Chạy Spark SQL trên bảng weekCollectData để tính ngày đầu tuần (GTE) và ngày cuối tuần (LTE).



* Thu thập dữ liệu
* Lấy danh sách thể loại (genres)

Gọi endpoint /genre/movie/list để tải về bảng genres và lưu vào /bronze/genres (mode overwrite).



* Lấy phim xu hướng & phim mới

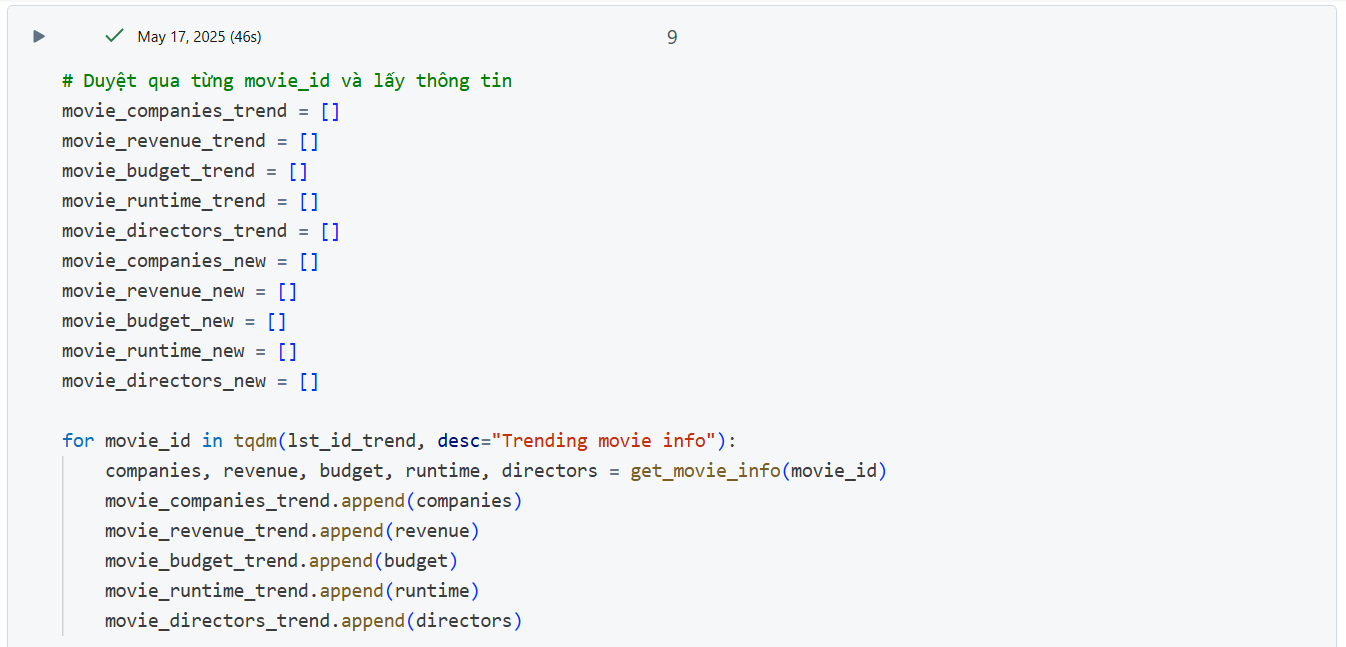
Gọi /trending/movie/week/

discover/movie?primary\_release\_date.gte=GTE&primary\_release\_date.lte=LTE



* Bổ sung thông tin chi tiết

Với mỗi movie\_id, dùng hai endpoint /movie/{id} và /movie/{id}/credits để gắn thêm companies, revenue, budget, runtime, directors.

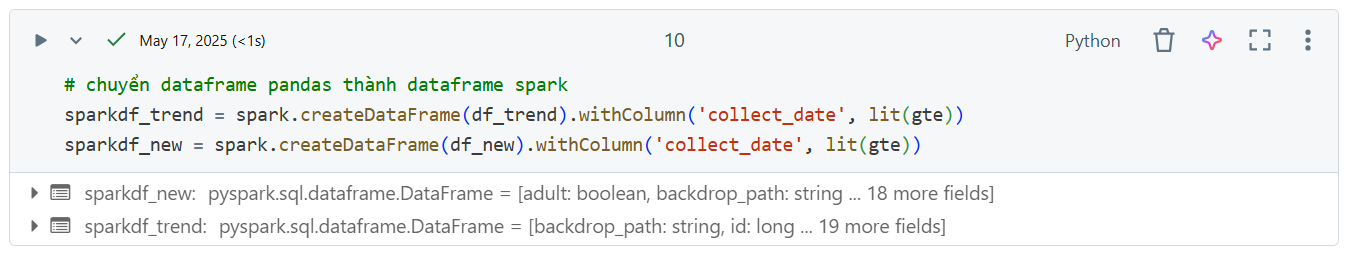




* Chuyển sang Spark DataFrame

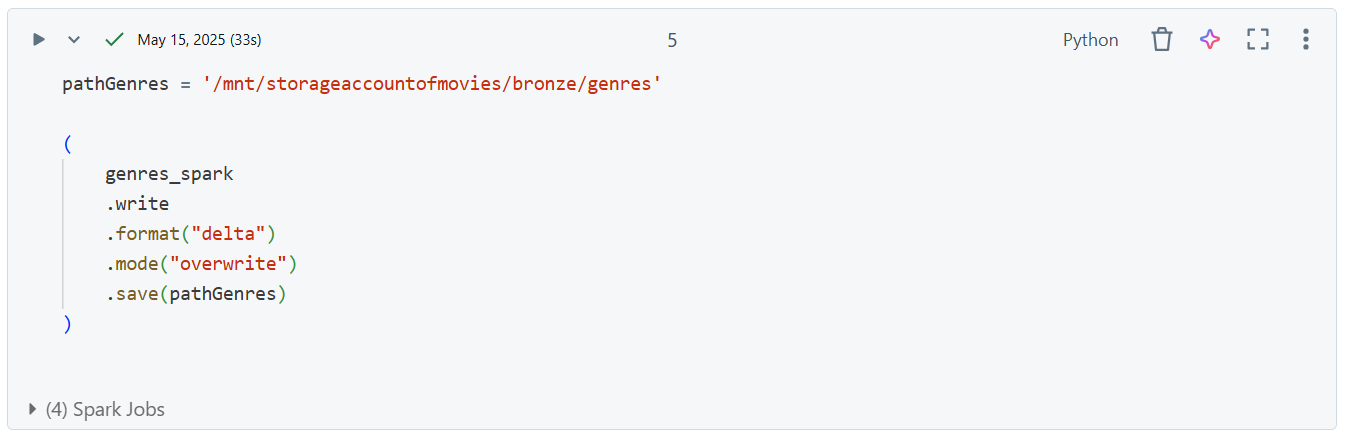
Biến đổi Pandas DataFrame → Spark DataFrame, thêm cột collect\_date = GTE

(áp dụng kỹ thuật CDC Timestamp)



* Ghi vào Bronze Layer & cập nhật lịch

Genres: lưu Delta table tại /bronze/genres (mode overwrite).



Movies: ghi hai Delta paths /bronze/DataTrend và /bronze/DataNew (mode append, mergeSchema=true).



Cập nhật tuần: chèn DAY\_FIRST\_WEEK=GTE, DAY\_LAST\_WEEK=LTE vào bảng weekCollectData.



### Transform (Silver Layer)

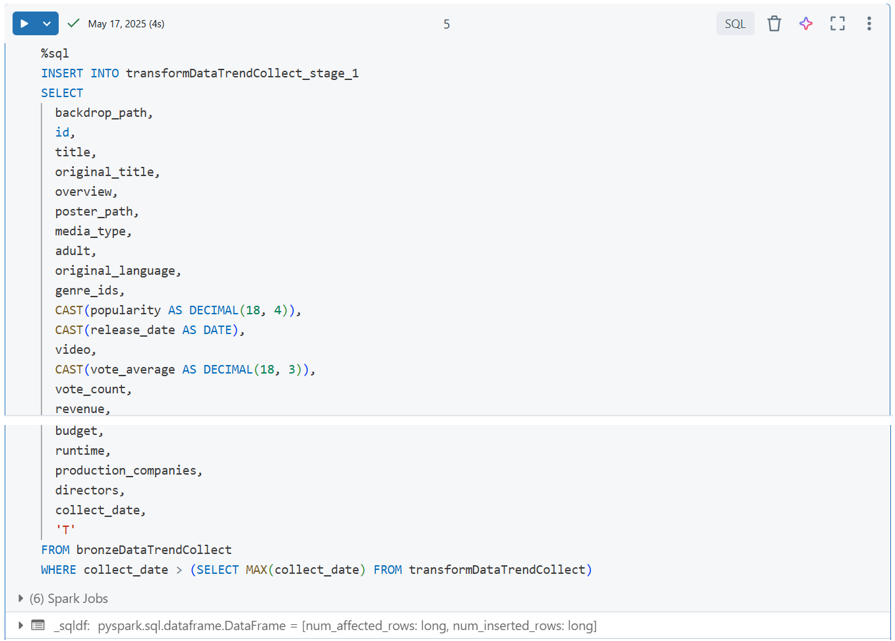
* Stage 1 – Chuẩn hóa & đánh dấu loại dữ liệu

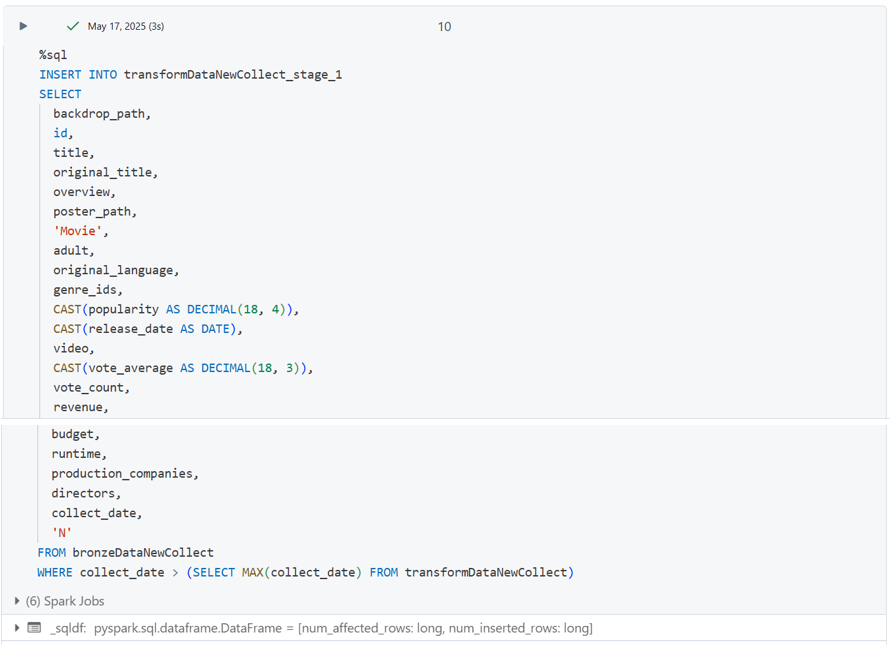
Lấy dữ liệu thô từ bronzeDataTrendCollect và bronzeDataNewCollect.

CAST lại kiểu cột:

* popularity → DECIMAL(18,4)
* vote\_average → DECIMAL(18,3)
* release\_date → DATE

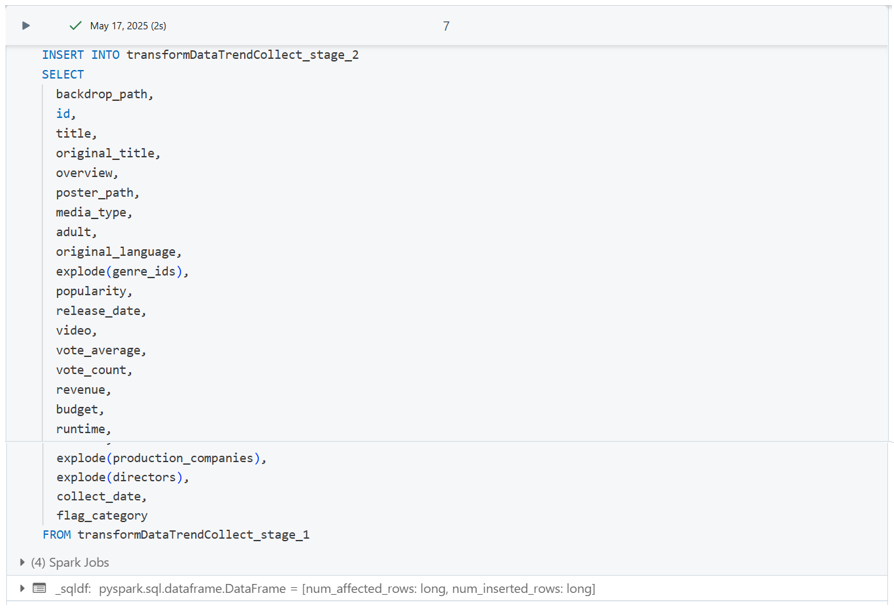
Thêm cột flag\_category bằng literal ('T' cho phim xu hướng, 'N' cho phim mới).





* Stage 2 – Phân rã mảng (Explode) thành dòng

Dùng explode(genre\_ids) tách mỗi phần tử trong mảng genre\_ids thành một dòng riêng.

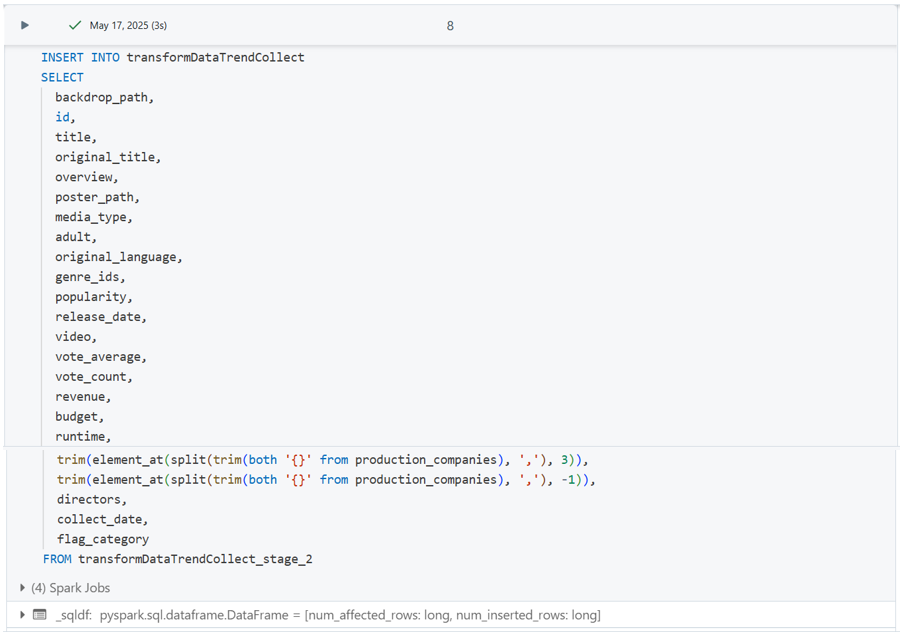
Tương tự với production\_companies và directors. 



* Stage 3 – Thêm dữ liệu vào bảng chính

Biến đổi cột production\_companies thành hai phần company và country.

Chèn dữ liệu vào transformDataTrendCollect và transformDataNewCollect.



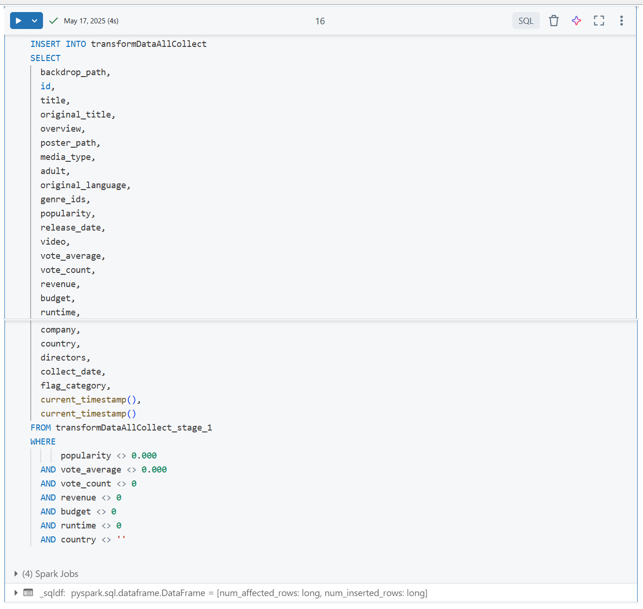


* Hợp nhất, lọc & tạo Silver Table chính

UNION hai bảng transformDataTrendCollect và transformDataNewCollect vào bảng tạm transformDataAllCollect\_stage\_1

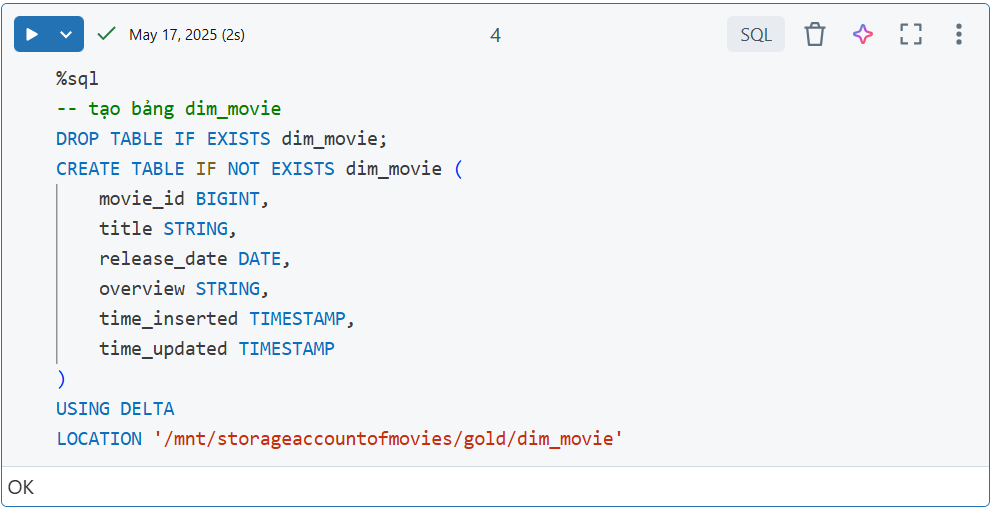
INSERT INTO transformDataAllCollect

* Lọc các bản ghi không hợp lệ (popularity, vote\_count, revenue,… ≠ 0; country ≠'')
* Thêm time\_inserted và time\_updated = current\_timestamp()



### Load (Gold Layer)

* Cấu hình các bảng dimensions và bảng fact:
* Dim\_movie



* Dim\_genre



* Dim\_company



* Dim\_director



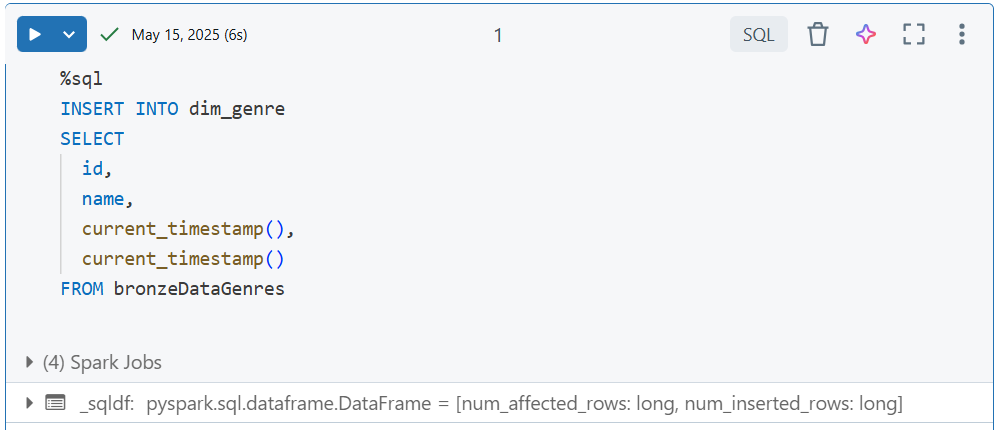
* Dim\_date



* Fact\_movies

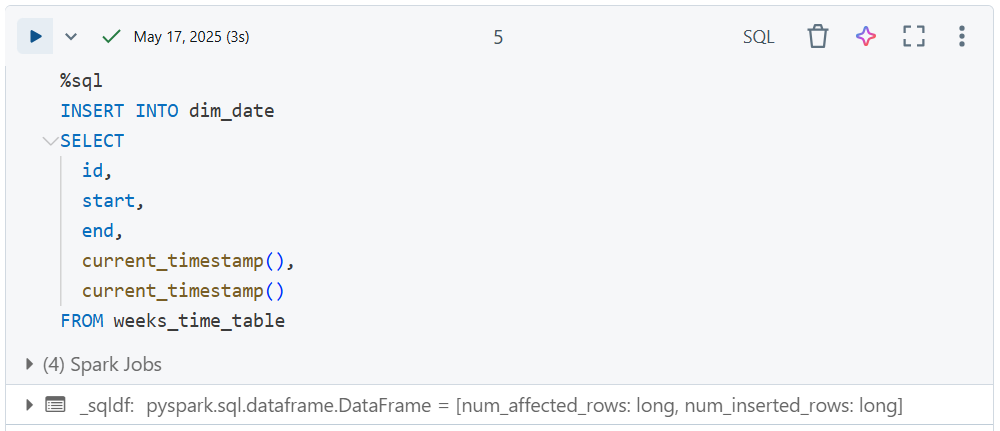


* Load dữ liệu vào các bảng Dimensions cố định (chỉ chạy một lần)
* Dim\_genre



* Dim\_date (2024/01/01 – 2024/12/31)

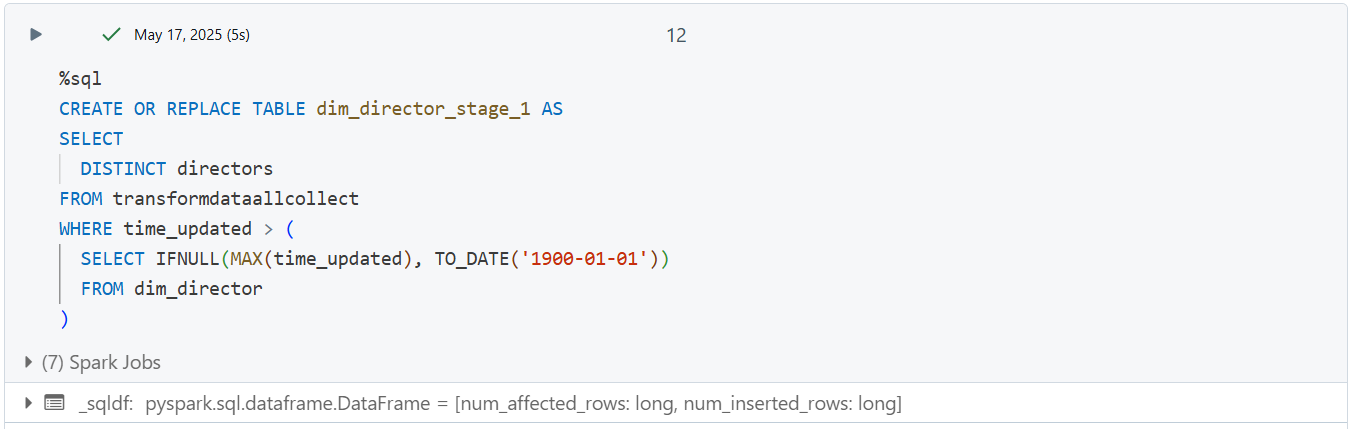
Start: ngày bắt đầu tuần, End: ngày kết thúc tuần



* Load dữ liệu vào các bảng Dimensions còn lại (lập lịch chạy cùng bảng Fact)
* Stage 1 – Chọn bản ghi mới / đã cập nhật
* Lấy các giá trị DISTINCT từ bảng transformDataAllCollect (title/company/directors).
* Chỉ chọn những bản ghi có time\_updated lớn hơn MAX(time\_updated) hiện tại trong bảng dimension (hoặc ngày mặc định 1900-01-01).
* Kết quả lưu vào bảng stage\_1.







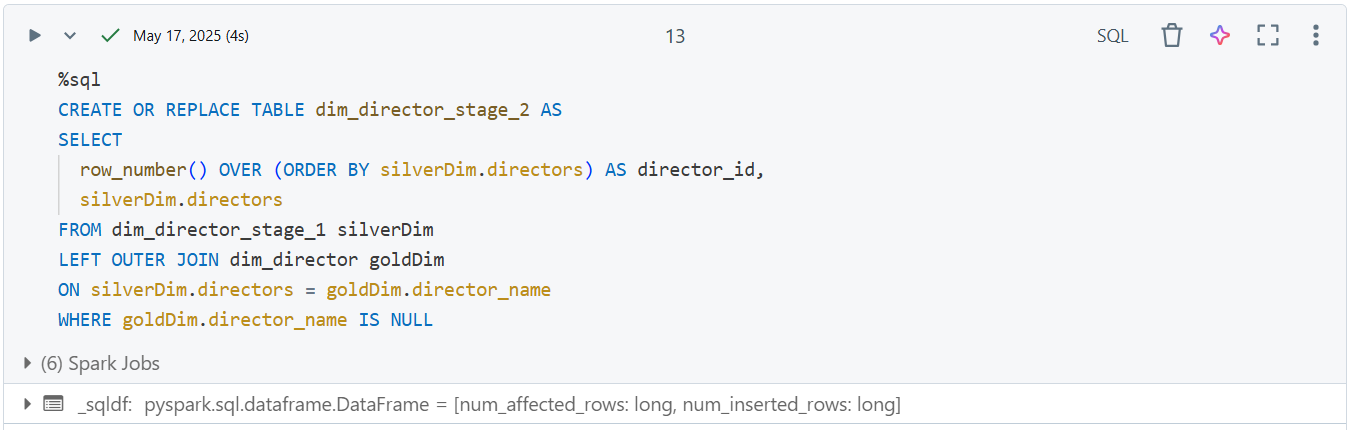
* Stage 2 – So sánh & gán ID tạm
* LEFT JOIN bảng stage\_1 với bảng dimension chính theo key

(title = title, company = company\_name, directors = director\_name).

* Với mỗi record:
* Nếu chưa có trong bảng chính → gán ID\_tạm = row\_number() OVER (…)
* Nếu đã có → giữ lại ID\_cũ từ bảng chính.
* Lưu kết quả vào stage\_2.



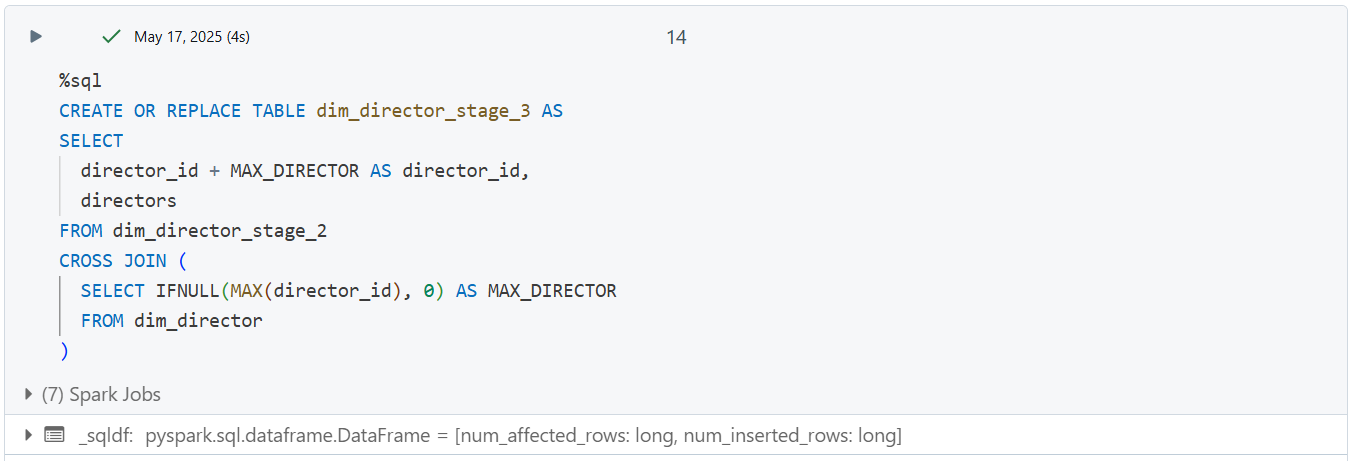




* Stage 3 – Tính ID cuối (surrogate key)
* CROSS JOIN với truy vấn lấy MAX(id) hiện tại từ bảng dimension chính.
* Với record mới (không khớp), tăng ID\_tạm + MAX(id) để tránh trùng.
* Với record cũ, giữ nguyên ID\_cũ.
* Kết quả lưu vào stage\_3.







* Stage 4 – MERGE (update/insert) vào bảng dimension
* Cú pháp MERGE INTO dim\_… USING stage\_3 ON key:
* WHEN MATCHED → UPDATE các cột

(country, release\_date, overview, time\_updated = current\_timestamp())

* WHEN NOT MATCHED → INSERT toàn bộ cột

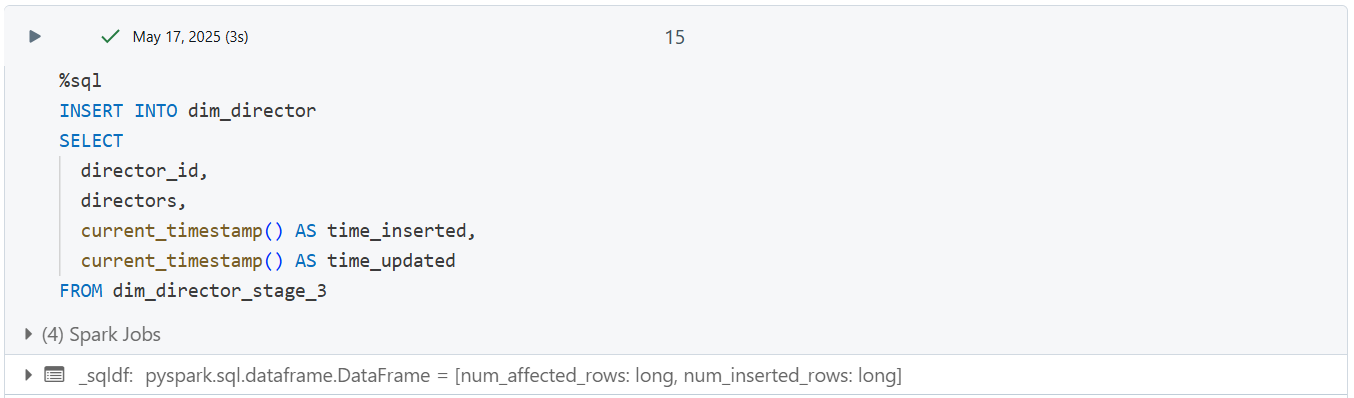
bao gồm surrogate key, fields, time\_inserted = current\_timestamp(), time\_updated = current\_timestamp().





Bảng Dim\_director không có cột bổ sung thông tin cho cột directors

nên không cần dùng MERGE INTO



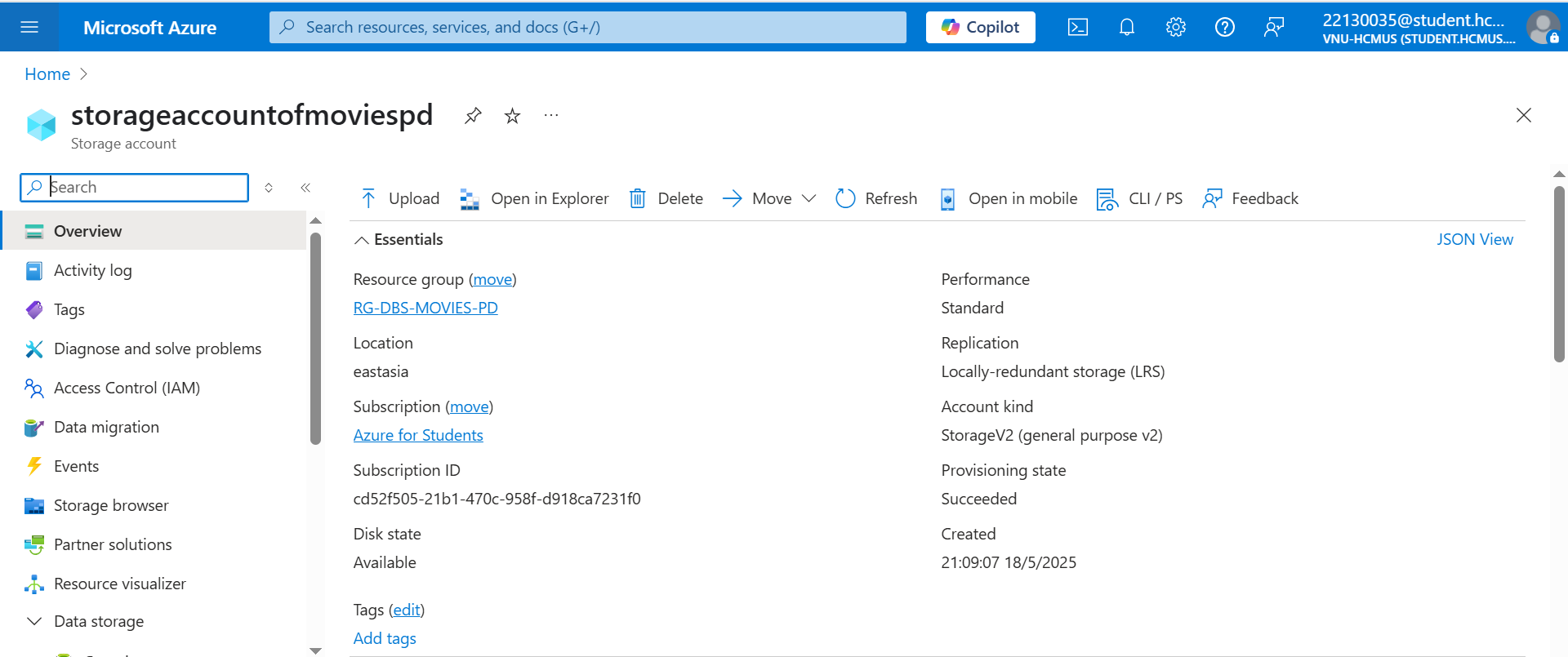
## **CI/CD VÀ TỰ ĐỘNG HÓA**

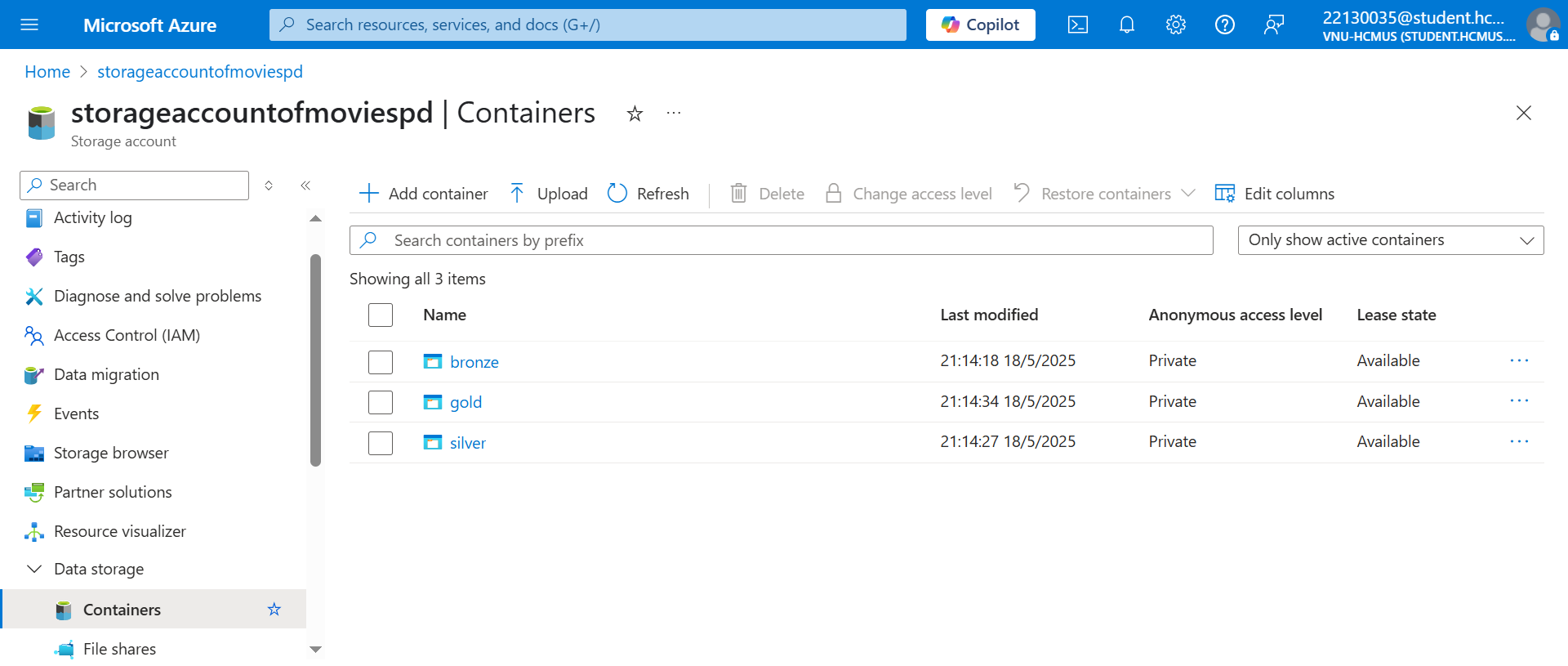
### Chuẩn bị hạ tầng Production

* Resource Group

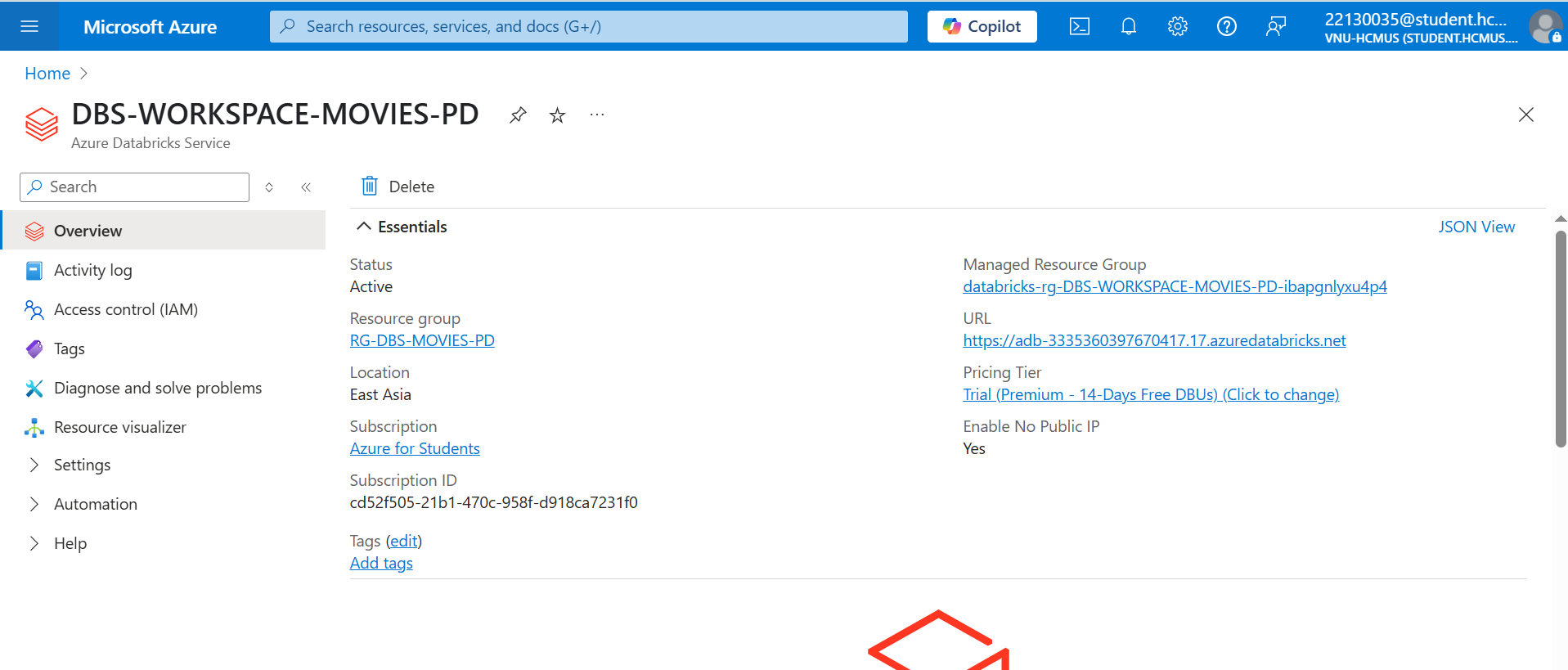


* Storage Account

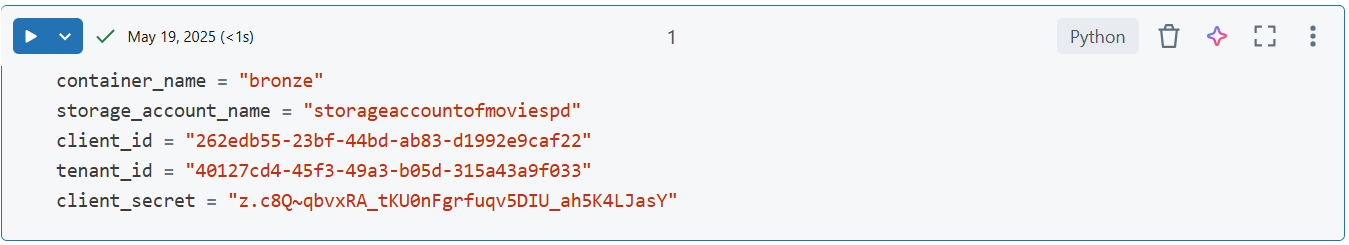


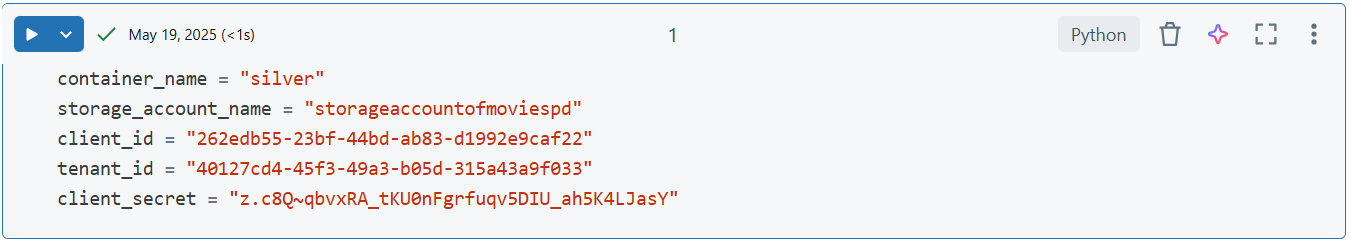


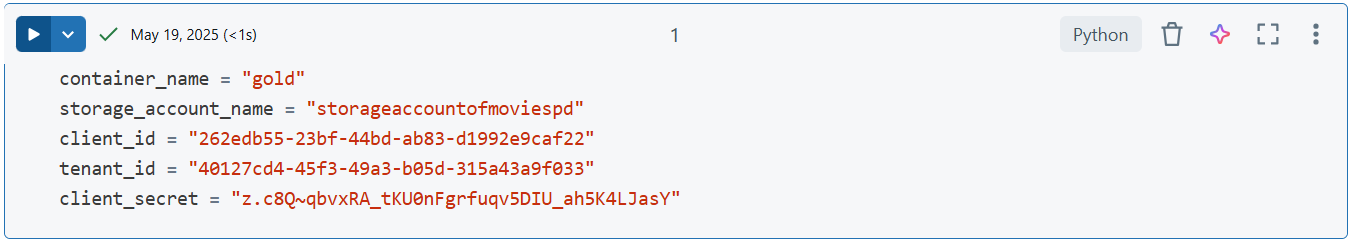
* Azure Databricks



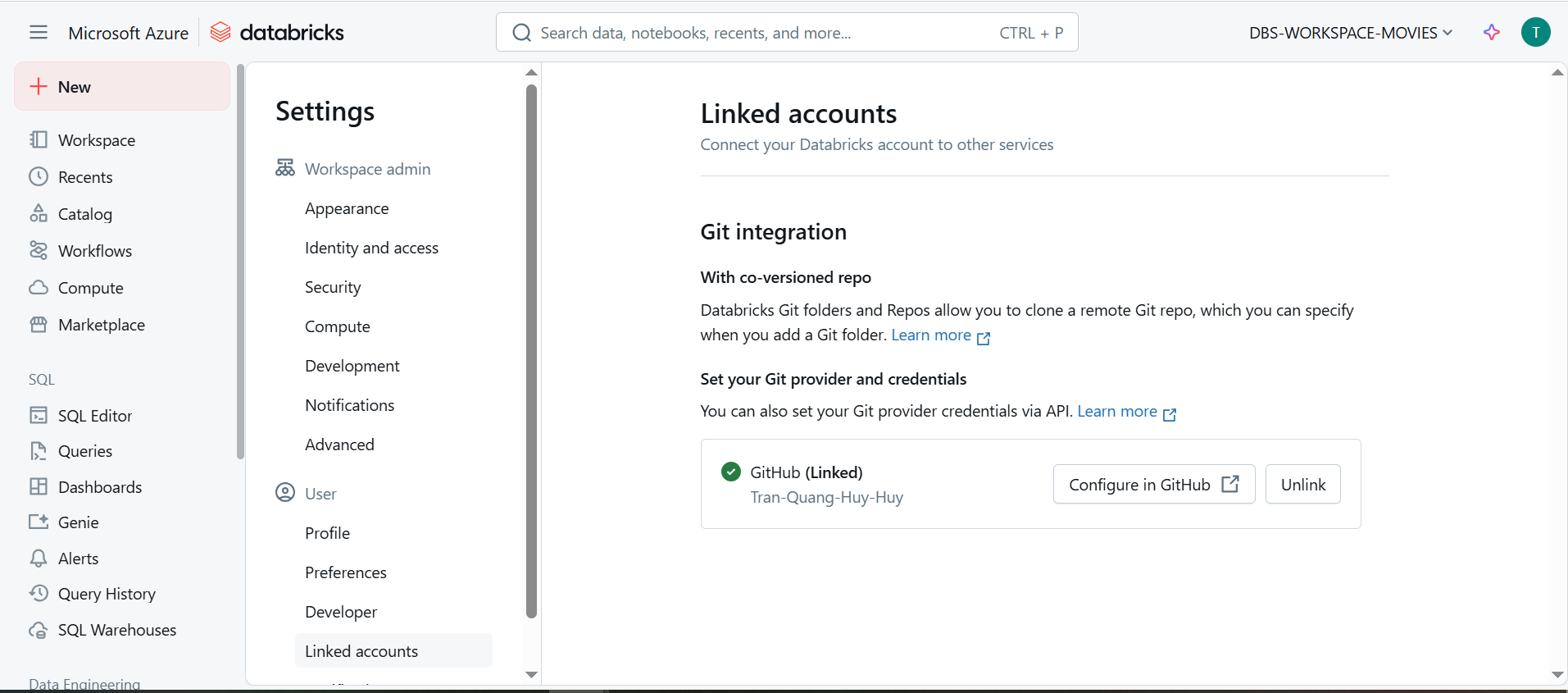
* Service Principle

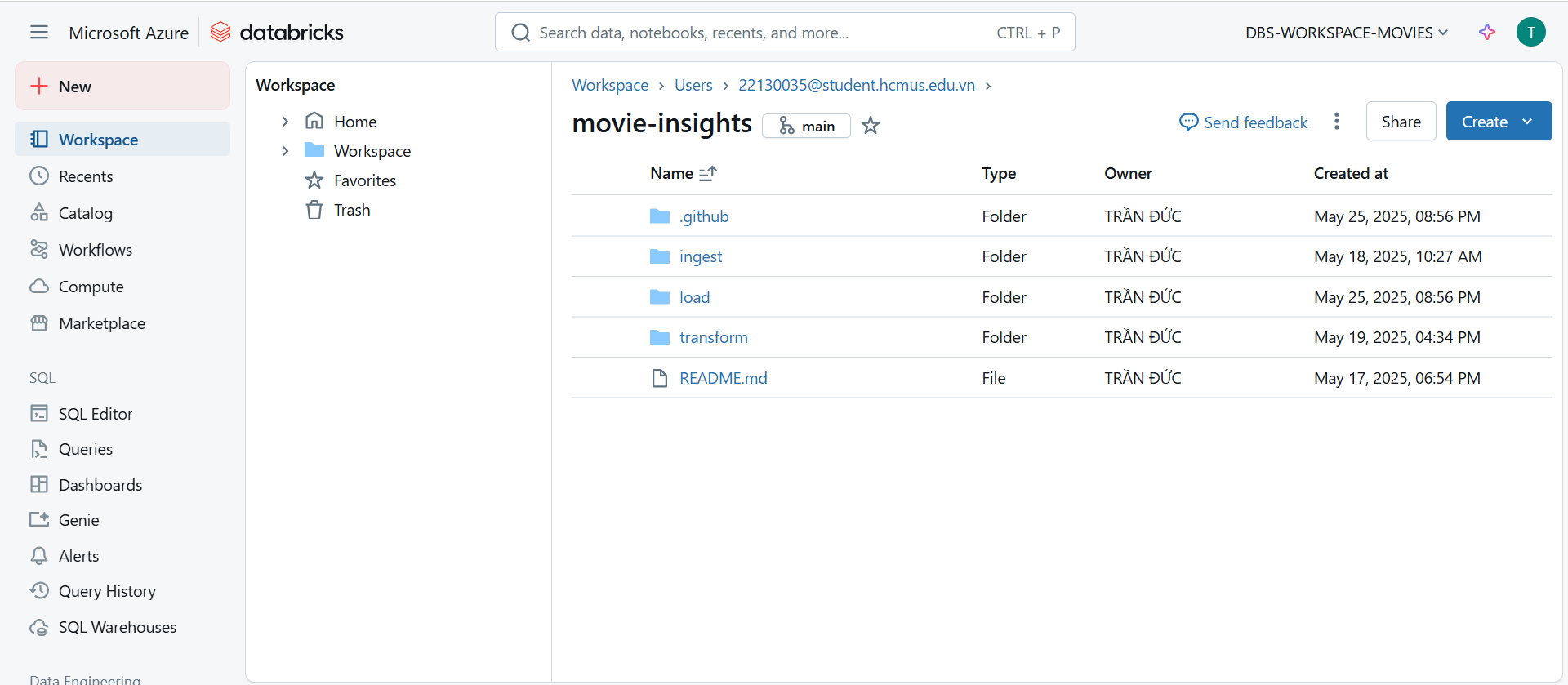




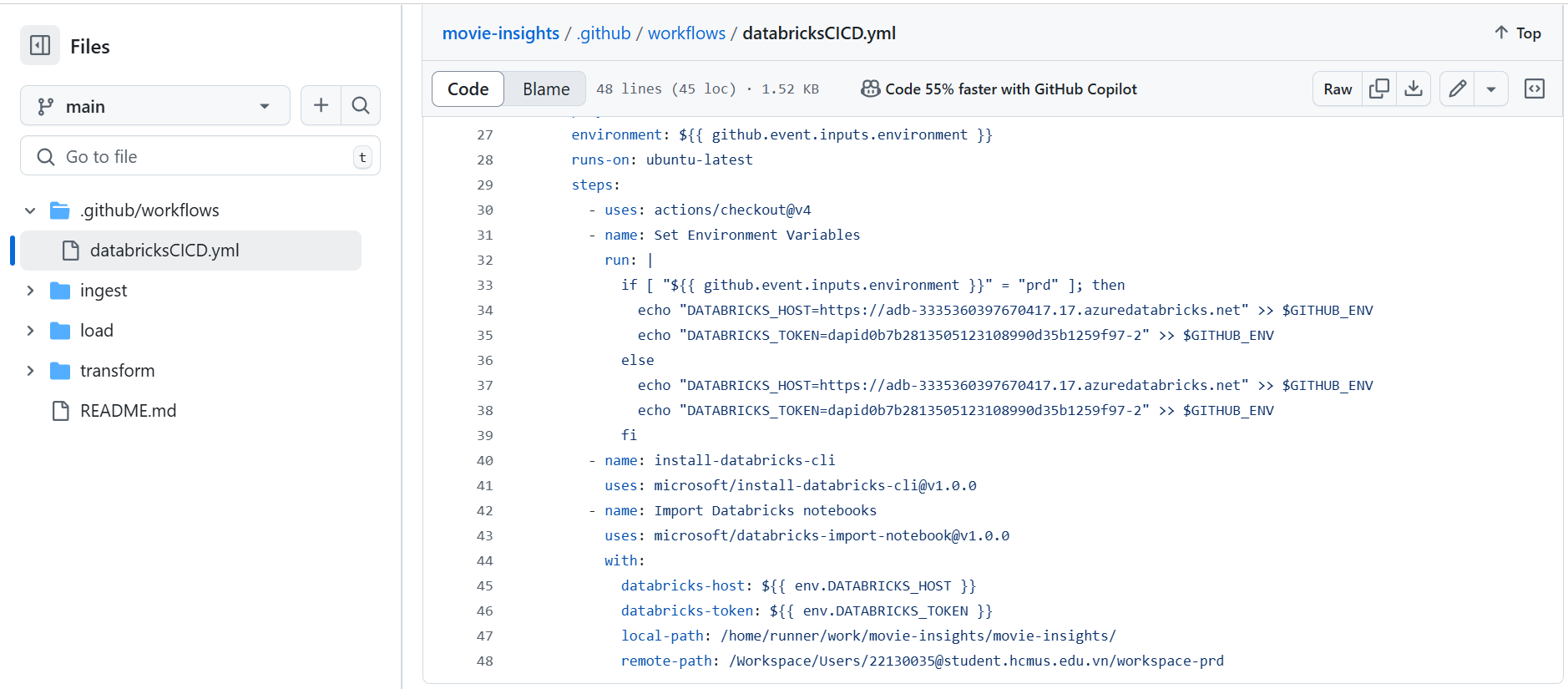


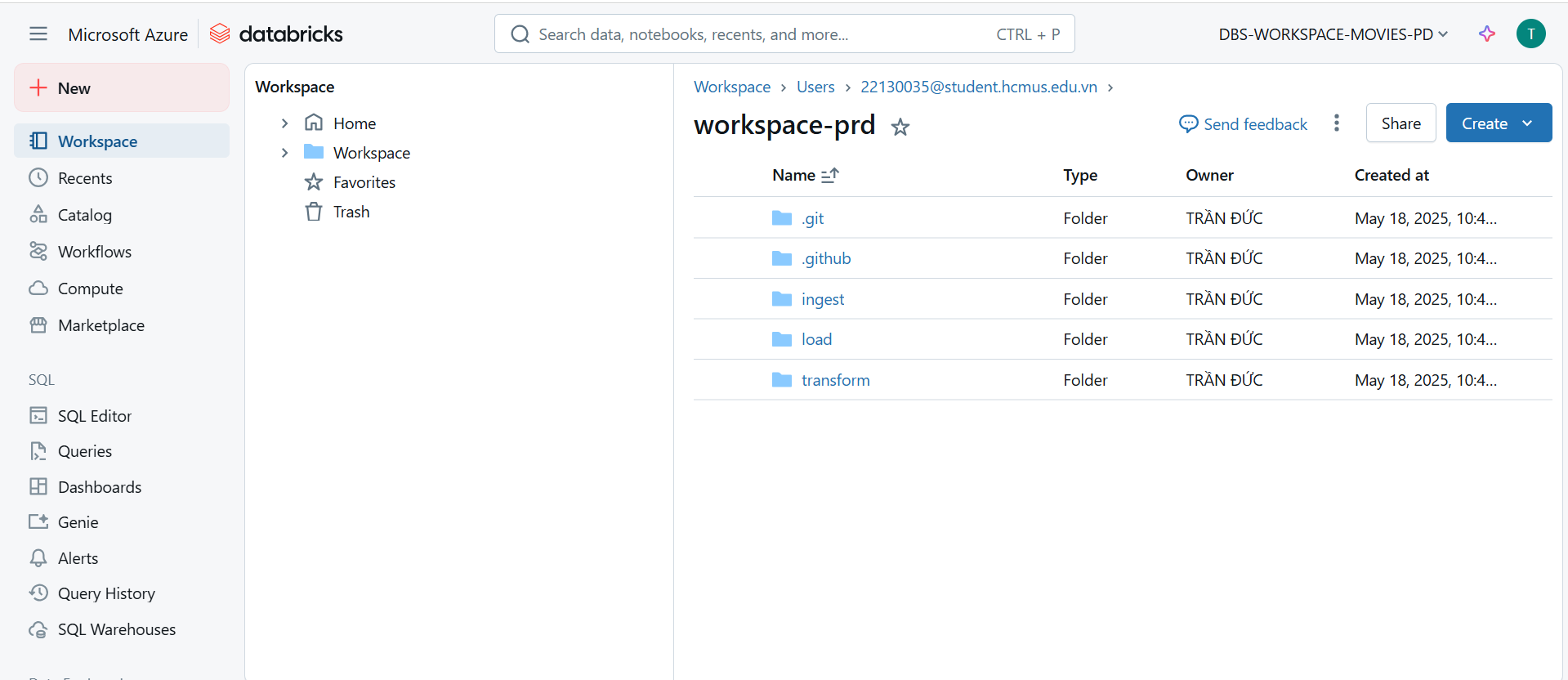
### Kết nối Git với Workspace Dev



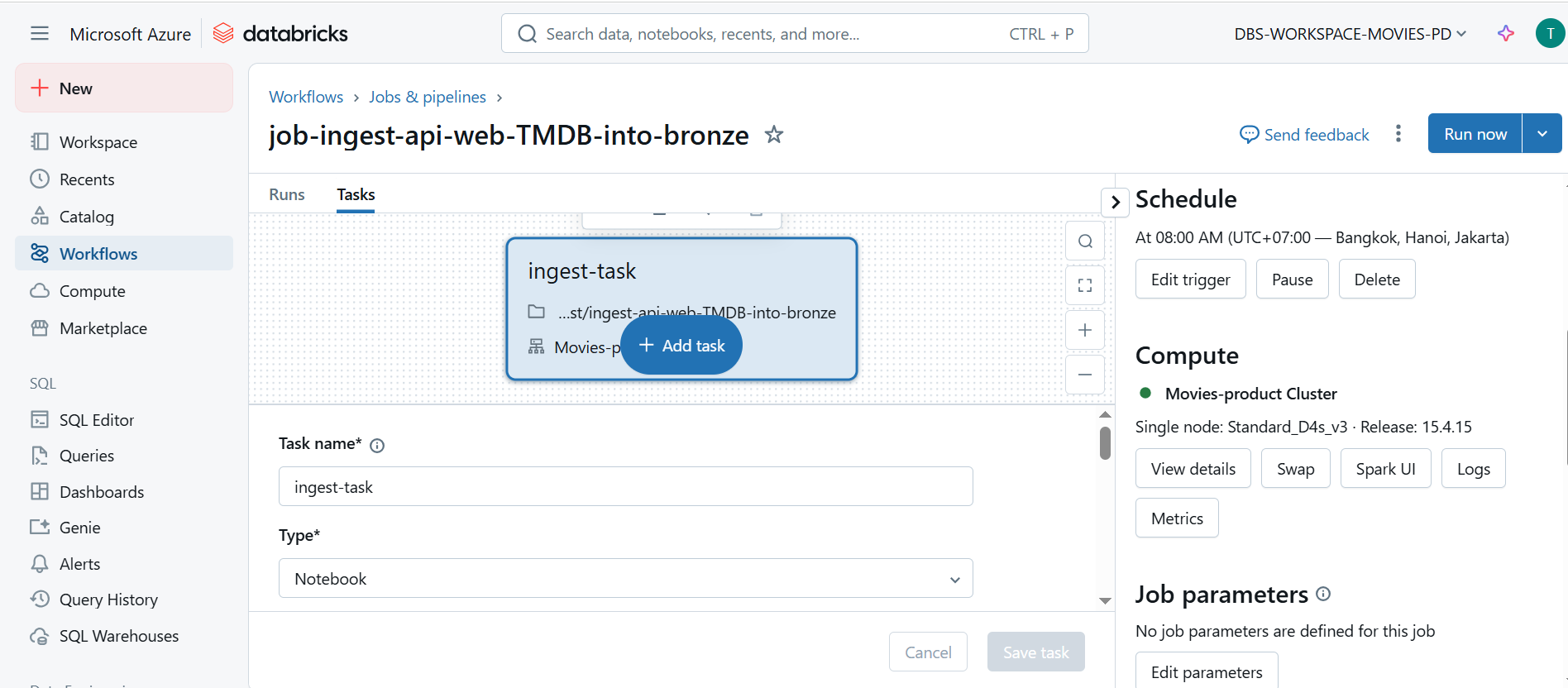


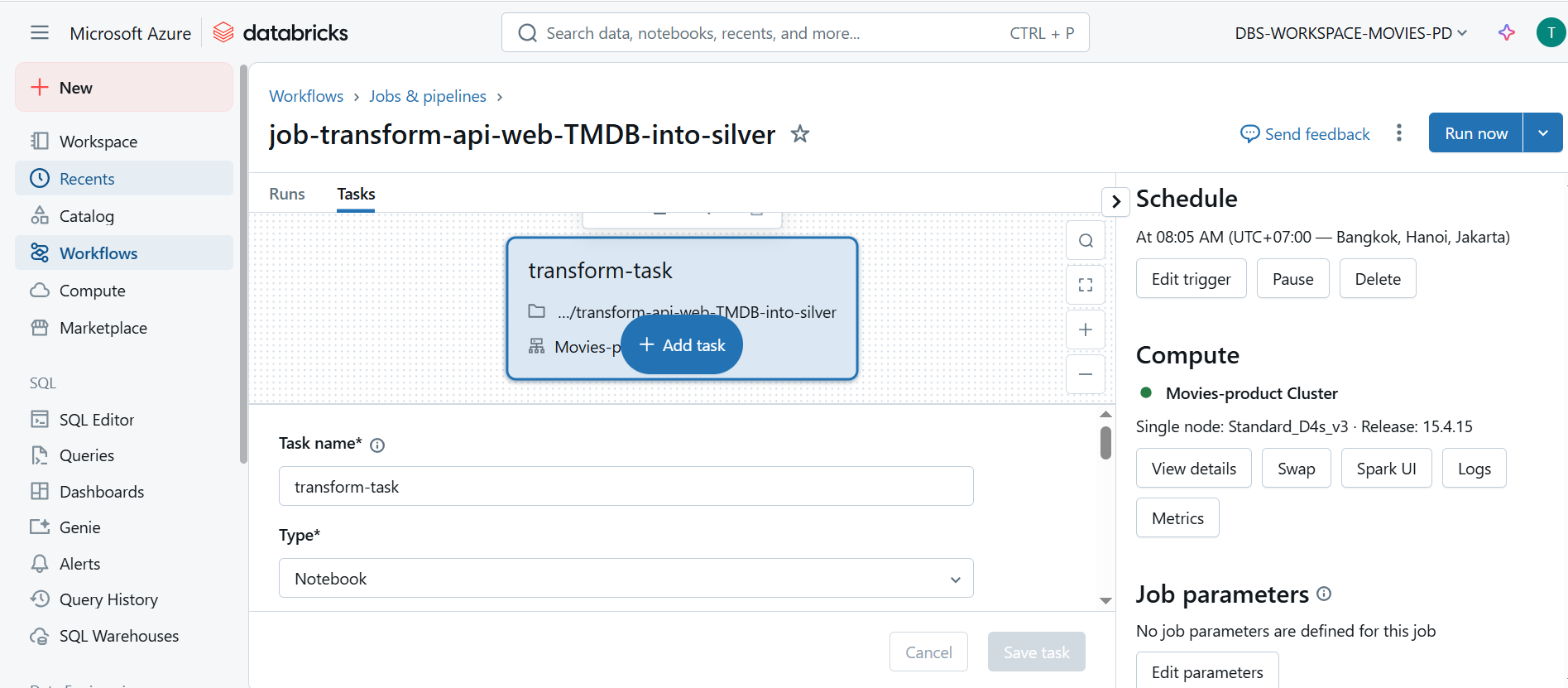
### Thiết lập CI/CD Production





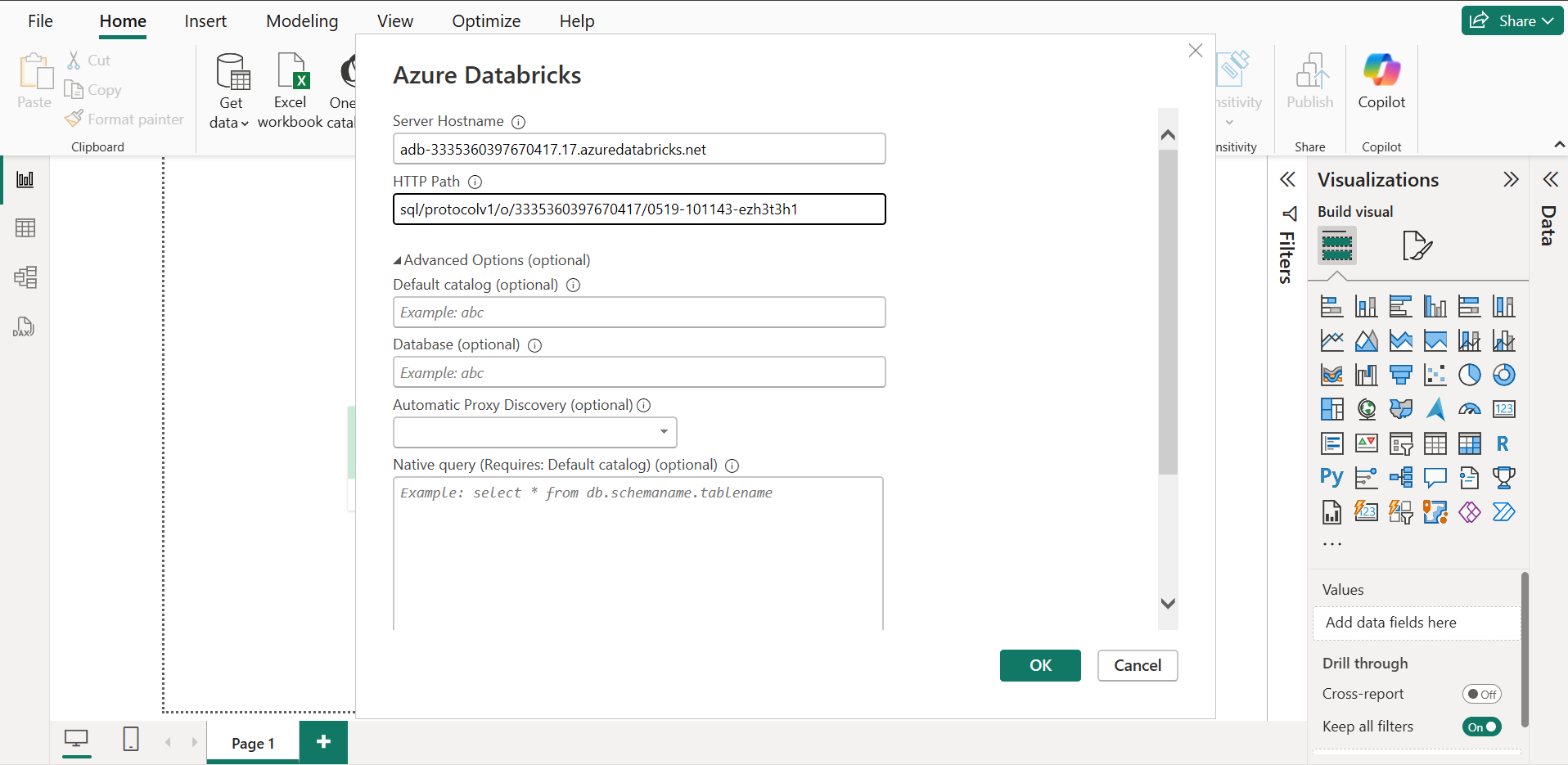
### Tự động hóa Lịch Chạy Job

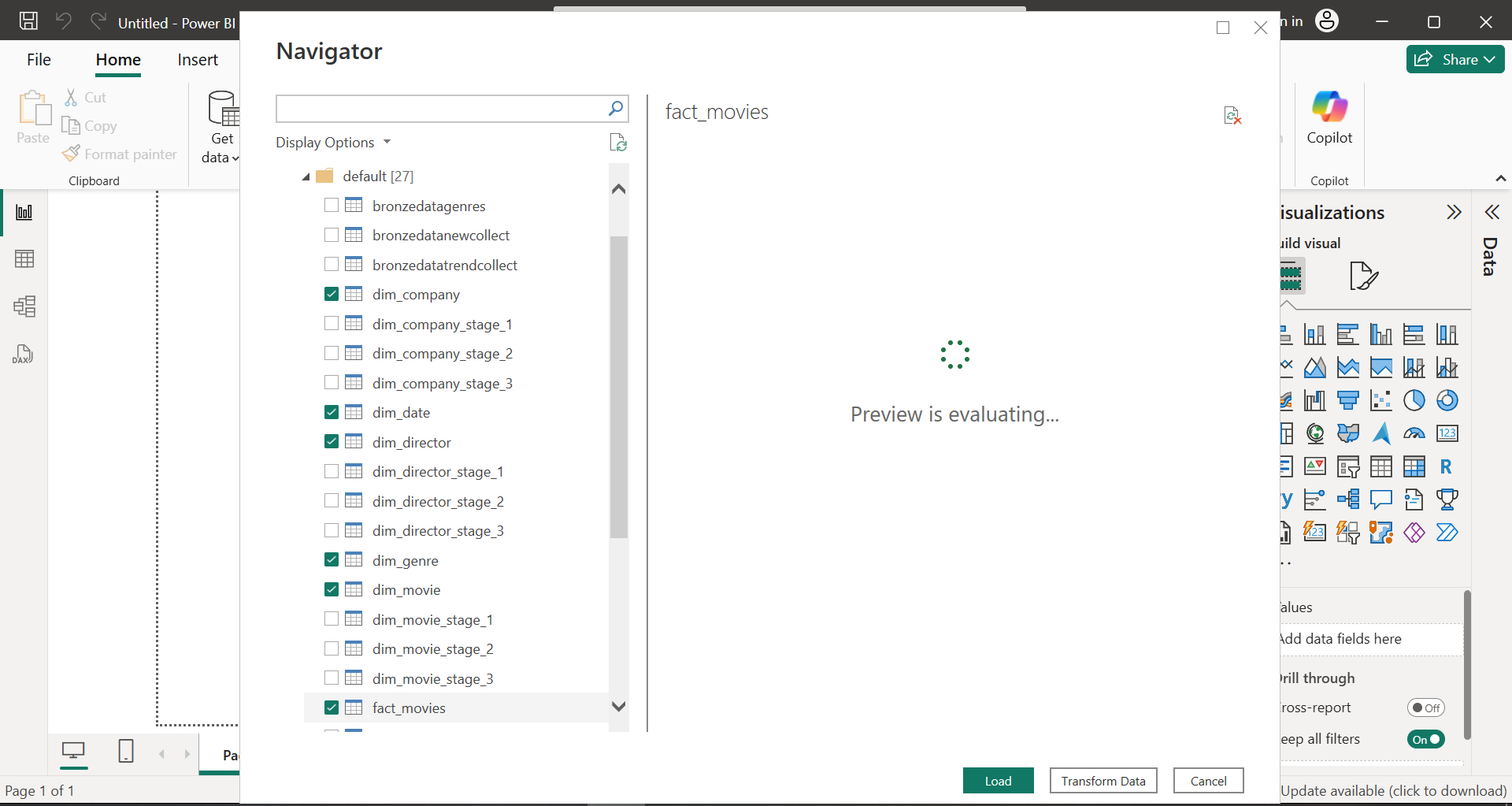




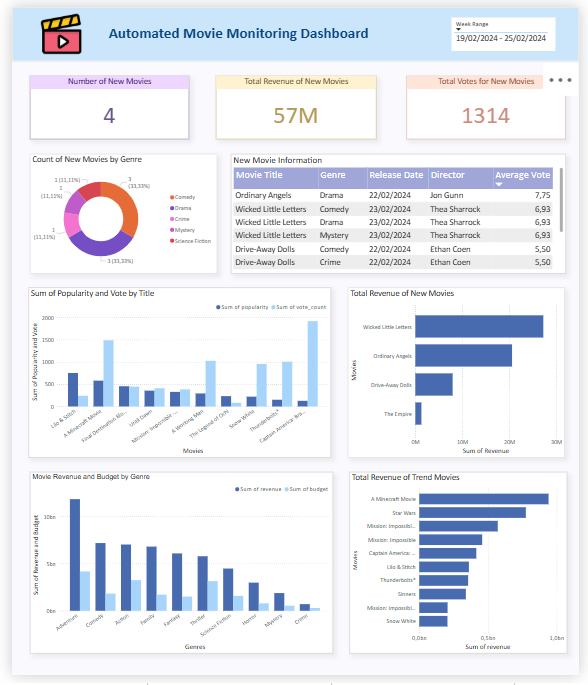


### Kết nối Power BI với Azure Databricks





***Automated Movie Monitoring Dashboard***



Cảm ơn các Anh/Chị, các bạn đã dành thời gian theo dõi project này. Rất mong nhận được những góp ý quý báu để hoàn thiện hơn trong tương lai. Kính chúc các Anh/Chị, các bạn có một ngày làm việc hiệu quả và nhiều niềm vui!