TRƯỜNG ĐẠI HỌC KINH TẾ

**KHOA THỐNG KÊ – TIN HỌC**

–––––––––––––––––––––––––––––––



**BÁO CÁO THỰC TẬP NGHỀ NGHIỆP**

**NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

**CHUYÊN NGÀNH QUẢN TRỊ HỆ THỐNG THÔNG TIN**

**TRỰC QUAN HÓA DỮ LIỆU ĐƯA RA XU HƯỚNG PHÁT TRIỂN CỦA NGÀNH CÔNG NGHIỆP ĐIỆN ẢNH HOLLYWOOD**

Sinh viên thực hiện : Hồ Thị Hồng Thắm

Nguyễn Lê Hoài Long

Phạm Tấn Thiên

Lớp : 45K21.1

Đơn vị thực tập : TMA Solutions Bình Định

Cán bộ hướng dẫn : Nguyễn Bá Văn

Nguyễn Khương Duy

Nguyễn Tấn Vỹ

Giảng viên hướng dẫn : ThS. Cao Thị Nhâm

**Đà Nẵng, 8/2022**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên: Hồ Thị Hồng Thắm

Lớp: 45K21.1

Khoa: Thống kê – Tin học

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thực tập từ ngày: 06/06/2022 đến ngày: 12/08/2022

Tại: TMA Solutions Bình Định

Địa chỉ: 12 Đại lộ Khoa học, Thung lũng Sáng tạo, Ghềnh Ráng, Tp. Qui Nhơn, Bình Định.

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

**1. Về thái độ, ý thức, đạo đức, kỷ luật**

**2. Kiến thức chuyên môn**

**3. Khả năng hòa nhập và thích nghi với công việc**

**4. Trách nhiệm, sáng tạo trong công việc**

**5. Các nhận xét khác**

**Đánh giá chung:**

**Điểm:**

Bình Định, ngày .......tháng ......năm 2022

**Xác nhận của đơn vị thực tập**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên: Nguyễn Lê Hoài Long

Lớp: 45K21.1

Khoa: Thống kê – Tin học

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thực tập từ ngày: 06/06/2022 đến ngày: 12/08/2022

Tại: TMA Solutions Bình Định

Địa chỉ: 12 Đại lộ Khoa học, Thung lũng Sáng tạo, Ghềnh Ráng, Tp. Qui Nhơn, Bình Định.

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

**1. Về thái độ, ý thức, đạo đức, kỷ luật**

**2. Kiến thức chuyên môn**

**3. Khả năng hòa nhập và thích nghi với công việc**

**4. Trách nhiệm, sáng tạo trong công việc**

**5. Các nhận xét khác**

**Đánh giá chung:**

**Điểm:**

Bình Định, ngày .......tháng ......năm 2022

**Xác nhận của đơn vị thực tập**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên: Phạm Tấn Thiên

Lớp: 45K21.1

Khoa: Thống kê – Tin học

Trường: Đại học Kinh tế - Đại học Đà Nẵng

Thực tập từ ngày: 06/06/2022 đến ngày: 12/08/2022

Tại: TMA Solutions Bình Định

Địa chỉ: 12 Đại lộ Khoa học, Thung lũng Sáng tạo, Ghềnh Ráng, Tp. Qui Nhơn, Bình Định.

Sau quá trình thực tập tại đơn vị của sinh viên, chúng tôi có một số nhận xét, đánh giá như sau:

**1. Về thái độ, ý thức, đạo đức, kỷ luật**

**2. Kiến thức chuyên môn**

**3. Khả năng hòa nhập và thích nghi với công việc**

**4. Trách nhiệm, sáng tạo trong công việc**

**5. Các nhận xét khác**

**Đánh giá chung:**

**Điểm:**

Bình Định, ngày .......tháng ......năm 2022

**Xác nhận của đơn vị thực tập**

# LỜI CẢM ƠN

Lời cảm ơn đầu tiên nhóm em xin gửi đến quý thầy, cô giáo trong khoa Thống kê – Tin học trường Đại học Kinh tế - Đại học Đà Nẵng lời cảm ơn chân thành vì đã tận tâm giảng dạy và truyền đạt những kiến thức, kinh nghiệm quý báu cho nhóm. Đặc biệt nhóm xin gửi lời cảm ơn sâu sắc nhất đến cô Cao Thị Nhâm – người đã tận tình hướng dẫn, giúp đỡ nhóm hoàn thành chuyên đề báo cáo thực tập này.

Nhóm xin chân thành cảm ơn ban lãnh đạo cùng với các phòng ban, các cô chú, anh chị trong Công ty TMA Solutions Bình Định – đơn vị đã tiếp nhận và đã tạo điều kiện thuận lợi cho nhóm tiếp cận thực tế, nắm bắt được các quy trình làm việc.

Cuối cùng nhóm xin cảm ơn các anh Nguyễn Khương Duy, Nguyễn Bá Văn, Nguyễn Tấn Vỹ - là những Mentors hướng dẫn trực tiếp các thành viên trong nhóm tại công ty TMA. Các anh đã tận tình giúp đỡ, cung cấp những tài liệu cho chúng em để hoàn thành bài báo cáo trong quá trình thực tập mảng Data Engineer tại công ty.

Vì thời gian và kiến thức còn hạn hẹp nên bài báo cáo không thể tránh khỏi những thiếu sót, rất mong sự góp ý của công ty, quý thầy cô để các thành viên trong nhóm rút kinh nghiệm và hoàn thành tốt hơn.

Nhóm chúng em xin chân thành cảm ơn!

# LỜI CAM ĐOAN

Nhóm chúng em xin cam đoan dự án “Thu thập dữ liệu, xử lý và đánh giá ngành công nghiệp điện ảnh” là kết quả nghiên cứu của nhóm dưới sự hướng dẫn của giảng viên Cao Thị Nhâm và các anh mentors Nguyễn Khương Duy, Nguyễn Bá Văn, Nguyễn Tấn Vỹ. Ngoài ra không có bất cứ sự sao chép nào của người khác.

Đề tài, nội dung báo cáo thực tập là sản phẩm mà nhóm em đã nỗ lực nghiên cứu trong suốt quá trình thực tập tại công ty TMA Solutions Bình Định. Các dữ liệu trong bài báo cáo là hoàn toàn trung thực. Chúng em xin chịu hoàn toàn trách nhiệm, kỷ luật của nhà trường đề ra nếu có vấn đề nào xảy ra.

Quy định:

Soạn thảo trên trang A4 (trang dọc, lề trái: 3.5cm; trên, phải, dưới: 2.5 cm), Font Times New Roman, canh đều 2 bên, size 13, cách dòng 1.5, cách đoạn trên 6pt, cách đoạn dưới 3pt, hàng đâu tiên lùi vào 1.27 cm; hình và bảng soạn theo caption, chèn trích dẫn chéo (Cross-reference) cho bảng và hình; các danh mục hình, bảng, mục lục làm tự động; đánh số trang như file mẫu (bìa không có số trang, danh mục + mục lục số trang theo i, ii, iii…, nội dung chính theo 1,2,3…)

# MỤC LỤC

[NHẬN XÉT CỦA DOANH NGHIỆP THỰC TẬP ii](#_Toc98339732)

[LỜI CẢM ƠN iii](#_Toc98339733)

[LỜI CAM ĐOAN iv](#_Toc98339734)

[MỤC LỤC vi](#_Toc98339735)

[DANH MỤC HÌNH ẢNH viii](#_Toc98339736)

[DANH MỤC BẢNG BIỂU ix](#_Toc98339737)

[DANH MỤC CÁC TỪ VIẾT TẮT x](#_Toc98339738)

[LỜI MỞ ĐẦU 1](#_Toc98339739)

[CHƯƠNG 1. TỔNG QUAN … 2](#_Toc98339740)

[1.1. Mục 1.1 2](#_Toc98339741)

[1.1.1. Mục 1.1.1 2](#_Toc98339742)

[1.1.2. Mục 1.1.2 2](#_Toc98339743)

[1.2. Mục 1.2 2](#_Toc98339744)

[CHƯƠNG 2. LÝ THUYẾT 3](#_Toc98339745)

[2.1. Mục 2.1 3](#_Toc98339746)

[2.1.1. Mục 2.1.1 3](#_Toc98339747)

[2.1.2. Mục 2.1.2 3](#_Toc98339748)

[2.2. Mục 2.2 3](#_Toc98339749)

[CHƯƠNG 3. TRIỂN KHAI … 4](#_Toc98339750)

[3.1. Mục 3.1 4](#_Toc98339751)

[3.1.1. Mục 3.1.1 4](#_Toc98339752)

[3.1.2. Mục 3.1.2 4](#_Toc98339753)

[3.2. Mục 3.2 4](#_Toc98339754)

[CHƯƠNG 4. KẾT QUẢ 5](#_Toc98339755)

[4.1. Mục 4.1… 5](#_Toc98339756)

[4.2. Mục 4.2… 5](#_Toc98339757)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 6](#_Toc98339758)

[TÀI LIỆU THAM KHẢO 7](#_Toc98339759)

[PHỤ LỤC 8](#_Toc98339760)

*(Mục lục này chỉ là ví dụ)*

# DANH MỤC HÌNH ẢNH

[Hình 1.1 Ngôn ngữ lập trình Python 2](#_Toc98336120)

[Hình 4.1 Kiến trúc của mô hình 5](#_Toc98336121)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1 Kiến trúc 5](#_Toc74235471)

# DANH MỤC CÁC TỪ VIẾT TẮT

**AI** : Artificial Intelligence

**…**

# LỜI MỞ ĐẦU

1. **Mục tiêu nghiên cứu của đề tài**

Trong thời đại chuyển đổi số, dữ liệu được xem là tài sản quan trọng nhất. Ngày nay, người sở hữu dữ liệu được xem là người giàu có. Cuộc đua giành dữ liệu hiện đang diễn ra tại mọi nơi, đứng đầu là những tập đoàn dữ liệu khổng lồ như Google, Facebook... Thông qua những dịch vụ mà những tập đoàn này cung cấp, họ tiến hành thu thập thông tin, dữ liệu của chúng ta nhằm phục vụ cho những mục đích kinh doanh. Và hòa vào dòng chảy phát triển mạnh mẽ đó là những công việc liên quan tới dữ liệu ra đời. Trong đó không thể không kể đến Data Engineer - kỹ sư dữ liệu.

Data Engineer sẽ là người tiến hành thu thập những dữ liệu có ý nghĩa sau đó xử lý và hoàn thiện những dữ liệu thô thành những dữ liệu có ý nghĩa sẵn sàng phục vụ cho những công việc liên quan cần sử dụng dữ liệu để phân tích. Vậy Data Engineer làm những gì? Data Engineer là ai? Để làm rõ điều này, nhóm sẽ thực hiện công việc của một DE thông qua một báo cáo mang tên “TRỰC QUAN HÓA DỮ LIỆU ĐƯA RA XU HƯỚNG PHÁT TRIỂN CỦA NGÀNH CÔNG NGHIỆP ĐIỆN ẢNH HOLLYWOOD”. Bài báo cáo là tổng hợp sơ lược về công việc của một DE, đưa ra những hình dung cơ bản về nghề DE. Đồng thời, bài báo cáo còn sở hữu giá trị phân tích, đánh giá thông qua những dữ liệu đã được thu thập trong quá trình hoàn thành bài báo cáo.

1. **Nhiệm vụ của đề tài**

Thông qua quá trình thực tập tại công ty để thực hiện bài báo cáo, các thành viên trong nhóm nâng cao hiểu biết về ngành Data Engineer và có cho mình những hình dung trực quan về nghề nghiệp của “tương lai” này. Đồng thời, đây là khoảng thời gian để mỗi cá nhân nâng cao khả năng tự học, tự tìm tòi, khả năng làm việc nhóm và làm việc cá nhân. Trong quá trình thực tập tại doanh nghiệp, các thành viên trong nhóm được trải nghiệm môi trường làm việc tại công ty công nghệ lớn, được va chạm với công việc thực tế, được học tập những đồng nghiệp đi trước và nhận được sự chia sẻ nhiệt tình từ những anh chị tiền bối.

Bài báo cáo sẽ giúp các thành viên nhóm vận dụng những kiến thức đã được học vào thực tế, từ có có nhận thức sâu sắc hơn về nghề nghiệp. Quá trình thực hiện bài báo cáo sẽ giúp mỗi thành viên xây dựng tư duy làm việc, các nhận diện và giải quyết yêu cầu công việc, nên lấy dữ liệu ra sao, lấy bằng cách nào, xử lý dữ liệu như thế nào cho đúng, dữ liệu có insight gì, cần tìm insight dữ liệu ra sao, xử lý dữ liệu làm sao để đảm bảo toàn vẹn dữ liệu, phân tích và đánh giá dữ liệu dựa trên những yếu tố nào và câu chuyện ẩn chứa phía sau dữ liệu là gì…

1. **Kết cấu của đề tài**

Đề tài được tổ chức gồm phần mở đầu, IV chương nội dung và phần kết luận.

* Mở đầu
* **Chương 1**: Cơ sở lý thuyết
* **Chương 2**: Tổng quan
* **Chương 3**: Triển khai
* **Chương 4**: Kết quả đạt được
* Kết luận và hướng phát triển

# CƠ SỞ LÝ THUYẾT

## Giới thiệu về Data Engineer

Data Engineer hay còn gọi là Kỹ sư dữ liệu, người này được xem là người xây dựng, kiểm tra và duy trì kiến trúc tổng hợp, lưu trữ và xuất dữ liệu từ những app hay hệ thống được tạo bởi Software Engineer. Đồng thời, DE là người thực hiện cải biến các quy trình thiết lập dữ liệu để mô hình hóa, khai thác và sản xuất dữ liệu. Nói một cách ngắn gọn, DE là người thu thập, tổng hợp dữ liệu từ nhiều nguồn thành một nguồn duy nhất. Hệ thống được xây dựng từ DE dùng phục vụ cho các phòng ban nghiệp vụ, Data Analyst, Data Scientist, Business Intelligence phân tích dữ liệu để hỗ trợ cho việc ra quyết định.

Công việc chủ yếu của một DE bao gồm: Cấu trúc dữ liệu nâng cao (distributed computing), lập trình đồng thời (concurrent programming), kiến thức sâu về một số công cụ mới như: Hadoop, Spark, Kafka, Hive…, tạo ETL/ data pipelines.

R và Python chính là 2 ngôn ngữ phổ biến nhất đối với nhân viên DE. Làm việc với 2 ngôn ngữ này, họ sẽ thường xuyên sử dụng các packages như ggplot2 để thực hiện trực quan hóa dữ liệu trong R hoặc thao tác dữ liệu thư viện Pandas Python. Tại một số công ty, ngôn ngữ Scala cũng được chọn làm ngôn ngữ chính dùng để xử lý dữ liệu. Đồng thời, SAS hay SPSS giúp xử lý tốt công việc. Ngoài ra, Power Bi, Tableau, Rapidminer, Matlab, Excel, Gephi cũng là những công cụ đắc lực của một DE.

## Tổng quan về Microsoft Azure

### Microsoft Azure là gì?

Azure Microsoft là một nền tảng điện toán đám mây (cloud computing platform) và một cổng trực tuyến (online portal) cho phép truy cập và quản lý các dịch vụ (service) và tài nguyên (resource) đám mây do Microsoft cung cấp.

### Cloud Computing (Điện toán đám mây)

1. Cloud Computing là gì?

Cloud Computing là việc cung cấp các dịch vụ điện toán hoàn toàn qua Internet. Hay nói đúng hơn là việc cung cấp tài nguyên phù hợp với nhu cầu người dùng hoàn toàn thông qua Internet.

1. Ưu điểm của Cloud Computing

* Khả năng mở rộng linh hoạt về quy mô.
* Khả năng phân bổ nguồn lực một cách nhanh chóng.
* Khả năng duy trì hoạt động trong các lỗi của bộ phận và dịch vụ.
* Khả năng phục hồi sau thảm họa. Chúng sao chép các tài nguyên tại các địa điểm khác nhau để có thể hoạt động ngay cả trong thời gian thảm họa.
* Khả năng giữ cho các tài nguyên có thể truy cập được trong một thời gian dài với thời gian ngừng hoạt động tối thiểu.

1. Các dạng Cloud Computing

* Public Cloud: Là loại hình được sở hữu và vận hành bởi nhà cung cấp dịch vụ trung gian. Với tất cả phần cứng, phần mềm và cơ sở hạ tầng hỗ trợ khác đều do nhà cung cấp sở hữu và quản lý.
* Private Cloud: Là một dạng tài nguyên được sử dụng riêng cho một tổ chức hoặc doanh nghiệp nào đó. Lúc này các tài nguyên của hệ thống công nghệ thông tin trong doanh nghiệp đều được chia sẻ về hệ thống máy chủ tính toán, giúp doanh nghiệp tận dụng được các tài nguyên này một cách hiệu quả và bảo mật nhất, cũng như tối ưu được nguồn tài chính, giảm thiểu rủi ro đầu tư cho chi phí cố định.
* Hybrid Cloud: Là sự kết hợp giữa cả hai hình thức trên. Chúng được ràng buộc với nhau bằng công nghệ để đảm bảo sự riêng tư. Lựa chọn này sẽ giúp cho doanh nghiệp có nhiều lựa chọn hơn, giúp tối ưu hóa cơ sở hạ tầng hiện tại.

### Azure Blod Storage

#### Blod Storage là gì?

Lưu trữ Azure Blob là giải pháp lưu trữ đối tượng của Microsoft cho đám mây.

Bộ lưu trữ Blob được tối ưu hóa để lưu trữ một lượng lớn dữ liệu phi cấu trúc. Dữ liệu phi cấu trúc là dữ liệu không tuân theo mô hình hoặc định nghĩa dữ liệu cụ thể, chẳng hạn như dữ liệu văn bản hoặc dữ liệu nhị phân

#### Blod Storage Resources

Blod Storage Resources gồm:

* Storage account.
* Container trong Storage account.
* Blod trong container.

#### Storage account: là tài khoản Storage Azure của bạn, nơi đây chứa tất cả các đối tượng dữ liệu như: blobs, files, queues, tables, và disks. Dữ liệu trong này có thể truy cập được từ mọi nơi trên thế giới qua HTTP hoặc HTTPS. Dữ liệu trong storage account của bạn có tính bền, khả dụng cao, an toàn và có khả năng mở rộng lớn.

Các loại Storage account:

* General-purpose v2 accounts: Một loại tài khoản lưu trữ cơ bản cho: blobs, files, queues, và tables. Và được đề xuất cho hầu hết các trường hợp sử dụng Azure Storage
* General-purpose v1 accounts: dùng để sử dụng mở rộng cho *purpose v2 accounts* khi cần
* BlockBlobStorage accounts: Tài khoản lưu trữ cao cấp có hiệu suất cao cho: block blobs và append blobs.
* FileStorage accounts: tài khoản lưu trữ chỉ dành cho file có hiệu suất cao
* BlobStorage accounts

#### Container trong Storage account: Là một vùng chứa để tập hợp các blob, tương tự như một thư mục trong hệ thống tệp. Một tài khoản lưu trữ có thể không giới hạn số lượng container và một container có thể lưu trữ không giới hạn số lượng các blobs.

#### Blod trong container: Azure storage hỗ trợ 3 loại blob:

* Block blobs: lưu trữ dữ liệu dạng text và dữ liệu nhị phân.
* Append blobs: lý tưởng cho việc ghi dữ liệu từ máy ảo
* Page blobs: lưu trữ các tệp truy cập ngẫu nhiên có kích thước lên đến 8 TB

### Azure Key Vault

Azure Keyvault là một dịch vụ lưu trữ các khóa, config cho website. Để tăng độ bảo mật, thường thì chỉ những khóa nào có mã khóa hay mật khẩu mới được sử dụng, và chi phí của nó thì không hề rẻ. Để sử dụng thì cũng rất đơn giản nếu bạn hiểu cách nó hoạt động.

Azure Keyvault không có config để bạn tích hợp trực tiếp vào project, muốn sử dụng bạn phải tạo một App Registration, và trong Azure Keyvault bạn tiến hành cấp quyền cho App Registration sử dụng. Bạn dùng config của App Registration để cấu hình vào project.

### Azure SQL Database

1.2.5.1. Azure SQL Database là gì?

Azure SQL Database là dịch vụ cơ sở dữ liệu thông minh, có thể mở rộng trên cloud. Cung cấp SQL Server Engine có tính tương thích rộng nhất và đem lại 212% ROI. Di chuyển các ứng dụng hiện tại và các ứng dụng mới xây dựng lên Azure cho các workloads SQL Server quan trọng của bạn.

Azure SQL Database hiện đang kích hoạt được Azure Arc. Bạn có thể chạy Azure SQL Database trên hạ tầng tại chỗ hoặc trên bất kỳ dịch vụ đám mây nào. Bạn vẫn có thể nhận được các lợi ích của Azure như tự động hóa, hỗ trợ không giới hạn, quản trị hợp nhất và tiết kiệm chi phí.

1.2.5.2. Các tính năng chính của Azure SQL Database

* Chuyển đổi dễ dàng

Tăng tốc quá trình di chuyển SQL Server trên on-premise mà không cần thay đổi mã với Managed Instance. Luôn sử dụng các tính năng mới nhất của SQL Server trên cloud mà không cần vá lỗi hay nâng cấp. Sử dụng chính sách Azure Hybrid Benifit để giảm chi phí.

Tận hưởng khả năng triển khai bổ sung linh hoạt với Single Database cho các ứng dụng SQL được tạo trên cloud hoặc Eslatic Pool cho các ứng dụng nhiều tenant.

* Tích hợp Machine learning

Mang lại hiệu suất và độ ổn định cao nhất với công nghệ AI an toàn , đáng tin cậy và đã được kiểm chứng. Tối đa hóa hiệu suất ứng dụng của bạn với các đề xuất tự động điều chỉnh tùy chỉnh. Sử dụng Intelligent Insights để giám sát và phát hiện các sự cố làm giảm hiệu suất.

* Quy mô lớn, tính khả dụng cao

Cân chỉnh ứng dụng của bạn theo nhu cầu với tính khả dụng lên tới 99.95%. Tận dụng Hyperscale với hỗ trợ lên tới 100TB và khôi phục nhiều cơ sở dữ liệu lớn trong vài phút. Cân chỉnh tính toán và lưu trữ tài nguyên độc lập mang tới tính linh hoạt cao và giảm thiểu chi phí. Tích hợp độ khả dụng cao trong Azure SQL Database đảm bảo rằng cơ sở dữ liệu sẽ không bao giờ bị lỗi trong kiến trúc phần mềm của bạn. Tối ưu hóa chi phí cho database với mô hình sử dụng gián đoạn, không tính trước với tầng tính toán không cần máy chủ.

* Bảo vệ dữ liệu nâng cao
* Bảo vệ database của bạn khỏi các mối nguy hại bằng kiểm soát truy cập chi tiết, công nghệ Always Encryted và khả năng chống các mối đe dọa cao cấp.
* Phát hiện, theo dõi và khắc phục các lỗ hổng tiềm ẩn với một bảng điều khiển duy nhất.
* Azure SQL Database đáp ứng các tiêu chuẩn tuân thủ nghiêm ngặt với công nghệ kiểm tra quản lý và bảo vệ thông tin tích hợp sẵn.
* Mang lại bảo vệ liên tục với các thông tin chuyên sâu từ Azure Security Center.
* Bảo mật nhiều lớp được cung cấp bởi Microsoft cho các Datacenter, hạ tầng và hệ thống.

### Azure Databricks

Đây là nền tảng phân tích dựa trên Apache Spark được tối ưu hóa cho đám mây Microsoft Azure. Với Azure Databricks, bạn có thể tạo các cụm Apache Spark được quản lý hoàn toàn. Bạn có thể động các cụm tỷ lệ tự động. Bạn cũng có thể tạo các cụm máy chủ và cũng có thể chia sẻ chúng giữa các nhóm và dễ dàng khám phá dữ liệu với sự trợ giúp của sổ ghi chép bằng R, Python, Scala và SQL. Bạn có thể sử dụng trang tổng quan tương tác để tạo báo cáo động. Nó có các khía cạnh bảo mật khi nói đến kiểm soát truy cập dựa trên vai trò và tích hợp với Azure Active Directory.

Apache Spark là một công cụ đa ngôn ngữ để thực thi kỹ thuật dữ liệu, khoa học dữ liệu và học máy trên các máy hoặc nút đơn.

Các tính năng chính:

-       Dữ liệu hàng loạt/ Truyền trực tuyến: Hợp nhất việc xử lý dữ liệu của bạn theo lô và phát trực tuyến theo thời gian thực, sử dụng ngôn ngữ ưa thích của bạn: Python, SQL, Scala, Java hoặc R.

-       Phân tích SQL: Thực thi các truy vấn ANSI SQL phân tán, nhanh chóng để lập bảng điều khiển và báo cáo đặc biệt. Chạy nhanh hơn hầu hết các kho dữ liệu.

-       Khoa học dữ liệu ở quy mô: Thực thi phân tích dữ liệu khám phá (EDA) trên dữ liệu quy mô petabyte mà không cần phải lấy mẫu xuống.

## Giới thiệu về ngôn ngữ lập trình Python và Scala

### Ngôn ngữ lập trình Python

1.3.1.1. Python là gì?

Là một ngôn ngữ lập trình bậc cao, do Guido van Rossum sang lập năm 1991. Có hình thức sáng sủa, cấu trúc rõ ràng, dễ đọc, dễ học và dễ nhớ, thuận tiện cho người mới học lập trình. Cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Thư viện phong phú, có cộng đồng lớn.

1.3.1.2. Các tính năng của Python

* Python là một ngôn ngữ lập trình dễ học, dễ đọc. Python tăng cường sử dụng từ khóa tiếng Anh, hạn chế các ký hiệu và cấu trúc cú pháp so với các ngôn ngữ khác.
* Mã nguồn của Python tương đối dễ để bảo trì và duy trì và có khả năng mở rộng.
* Python có một tiêu chuẩn thư viện rộng, nền tảng tương thích trên nhiều hệ điều hành hiện nay như: UNIX, Windows, và Macintosh. Đây cũng là một trong những điểm mạnh đáng kể của Python.
* Python có thể được sử dụng như ngôn ngữ script, hoặc ngôn ngữ biên dịch, nhờ đó có thể build các chương trình lớn trên nó.

### Ngôn ngữ lập trình Scala

1.3.2.1. Scala là gì?

Scala là ngôn ngữ lập trình đa mẫu hình, được thiết kế tích hợp các tính năng của lập trình hướng đối tượng và lập trình hàm. Scala chạy trên máy ảo Java nên nó có thể đọc được các thư viện của Java.

1.3.2.2. Những ưu điểm của Scala

* Scala ngắn gọn, súc tích, dễ đọc hiểu và dễ học. Các cấu trúc của Scala rất nhẹ. Cho phép lập trình viên có thể giảm kích thước của code ít nhất hai hoặc ba lần so với Java. Vì thế nó giúp cho việc viết code nhanh hơn và việc maintain cũng dễ hơn.
* Nó rất chính xác vì nó được trang bị hệ thống phát hiện và tránh nhiều lỗi ứng dụng tại thời điểm biên dịch.
* Ngôn ngữ này có tính mở rộng. Nó cung cấp một ngôn ngữ máy độc đáo, dễ dàng thêm thư viện, hỗ trợ việc lập trình dựa trên cá thư viện hàm có sẵn.

## Giới thiệu về PowerBI Desktop

Power BI Desktop là một ứng dụng miễn phí, cài đặt trên máy tính cục bộ cho phép kết nối, chuyển đổi và trực quan hóa dữ liệu. Với Power BI Desktop, có thể kết nối với nhiều nguồn dữ liệu khác nhau và kết hợp chúng (thường được gọi là mô hình hóa) thành một mô hình dữ liệu. Mô hình dữ liệu này cho phép xây dựng các bảng dữ liệu kết nổi và biểu đồ từ đó có thể chia sẻ dưới dạng báo cáo. Hầu hết người dùng làm việc trong các dự án kinh doanh thông minh sử dụng Power BI Desktop để tạo báo cáo, sau đó sử dụng Power BI Service để chia sẻ báo cáo cho những người khác.

Các ứng dụng và tính năng phổ biến nhất của Power BI Desktop bao gồm:

* Kết nối với dữ liệu (Connect to data)
* Chuyển đổi và làm sạch dữ liệu đó, để tạo mô hình dữ liệu (Transform & Clean data to create data model)
* Tạo hình ảnh, chẳng hạn như biểu đồ hoặc đồ thị, cung cấp các trình bày trực quan về dữ liệu (Create visuals)
* Tạo báo cáo là tập hợp các hình ảnh, trên một hoặc nhiều trang báo cáo (Create Report)
* Tải báo cáo lên Power BI Service để chia sẻ cho những người khác (Publish to Power BI Service to share report)

Có ba chế độ xem có sẵn trong Power BI Desktop. Các chế độ xem, được hiển thị theo thứ tự xuất hiện, như sau:

1. Báo cáo: Trong chế độ xem này, bạn tạo báo cáo và biểu đồ, nơi phần lớn thời gian thiết kế của bạn được sử dụng.
2. Dữ liệu: Trong chế độ xem này, bạn thấy các bảng, các measure hay calculated columns và dữ liệu khác được sử dụng trong mô hình dữ liệu liên kết với báo cáo của bạn. Từ đó, bạn có thể chuyển đổi dữ liệu để sử dụng tốt nhất trong mô hình của báo cáo.
3. Mô hình: Trong dạng xem này, bạn thấy và quản lý mối quan hệ giữa các bảng trong mô hình dữ liệu của mình.

## Các bước và công cụ

### Cào dữ liệu - Crawl

Mô tả công việc cào dữ liệu:

* Crawl là cào dữ liệu (Crawl Data) là một thuật ngữ không còn mới trong Marketing, và SEO. Vì Crawl là kỹ thuật mà các con robots của các công cụ tìm kiếm sử dụng như: Google, Bing Yahoo…
* Công việc chính của crawl là thu thập dữ liệu từ một trang bất kỳ. Rồi tiến hành phân tích mã nguồn HTML để đọc dữ liệu. Và lọc ra theo yêu cầu người dùng hoặc dữ liệu mà Search Engine yêu cầu.
* Trước tiên, thu thập dữ liệu webpage tại các URL. Từ các page này, tìm thấy các siêu liên kết đến nhiều URL khác và thêm các liên kết mới tìm được vào danh sách các trang cần thu thập thông tin tiếp theo. Sau đó sẽ truy cập vào URL các trang có thông tin quan trọng và bắt đầu cào những dữ liệu quan trọng phù hợp với yêu cầu.

Phần mềm sử dụng:

* Phần mềm: Visual Code
* Ngôn ngữ lập trình: Python
* Thư viện hỗ trợ: Beautifulsoup, Pandas

### ETL - Lưu chuyển, làm sạch và chuẩn hóa dữ liệu

ETL là viết tắt của Extract – Transform – Load, tạm được dịch là Trích xuất – biến đổi – tải. Đây là quy trình chung dùng để sao chép dữ liệu từ một hay nhiều nguồn vào hệ thống đích đại diện cho dữ liệu khác với nguồn dữ liệu. Trong quy trình ETL, dữ liệu sẽ được trải qua các bước:

* Extract/ Trích xuất: Đây là quá trình đọc dữ liệu từ cơ sở dữ liệu. Trong giai đoạn này, nó chọn dữ liệu nguồn hoặc siêu dữ liệu và dữ liệu kinh doanh đến từ các nguồn khác nhau. Dữ liệu có thể ở bất kỳ định dạng nào như tệp xml, tệp phẳng hoặc bất kỳ tệp cơ sở dữ liệu nào. Tại giai đoạn này, dữ liệu gốc được kiểm tra. Dữ liệu sẽ được kiểm tra để xem có bất kỳ lỗi nào hay không. Giai đoạn Extarct sẽ kiểm tra các lỗi hoặc sự không nhất quán của dữ liệu bằng cách sử dụng một số kỹ thuật trí tuệ nhân tạo. Trong ngắn hạn, nó xác minh xem chất lượng của sản phẩm có đạt hay không. Sau đó dữ liệu sẽ được chuyển tiếp tới giai đoạn tiếp theo để tiếp tục được xử lý.
* Transform/ Biến đổi: Đây là giai đoạn thứ hai trong ETL. Chuyển đổi là quá trình trong đó định dạng ban đầu được chuyển đổi thành định dạng bắt buộc mà người dùng muốn. Chuyển đổi là mô hình hóa hoặc thay đổi dữ liệu theo yêu cầu của người dùng. Ví dụ những thay đổi có thể là thay đổi về số cột hoặc số hàng.
* Load/ Tải: Tải dữ liệu và xác nhận số lượng hàng đã được xử lý. Quá trình này còn được biết tới như quá trình ghi chép dữ liệu vào cơ sở dữ liệu đích.

Như vậy, với những DE chuyên thực hiện công việc ETL, bạn sẽ tiến hành lưu chuyển, trao đổi dữ liệu dữ các nơi lưu trữ khác nhau nhằm mục đích đối chiều, bổ sung, thậm chí là backup dữ liệu. Đồng thời, các DE thực hiện làm sạch và chuẩn hóa dữ liệu bằng cách loại bỏ dữ liệu dư thừa, dữ liệu rác và chuyển các dữ liệu về các định dạng nhất định. Sau khi làm sạch, dữ liệu được đưa về tập trung, load lên một kho chứa duy nhất là Data Warehouse. Khi dữ liệu đã được lưu trữ tại đây có nghĩa là dữ liệu đã được thiết kế theo mô hình chuyên biệt sẵn sàng cho việc phân tích hoặc khôi phục dữ liệu trong trường hợp cần thiết.

Các công cụ giúp các DE có thể thực hiện quá trình ETL có thể là: SQL, SSIS, Data Stage, Bash jobs, Python, Scala, DataBricks…

### Trực quan hóa dữ liệu – Dashboard

Một công việc không thể thiếu của DE đó là phân tích và trực quan hóa dữ liệu. Đối với các công việc về phân tích dữ liệu nhiều người thường nghĩa ngay tới nghề DA (Data Analyst), tuy nhiên đối với DE việc biết về phân tích và trực quan hóa dữ liệu là điều hết sức cần thiết. Từ những dữ liệu được lưu trữ tại các Data Warehouse hay các Database, các DE sẽ sử dụng các công cụ như BI Tools để lấy dữ liệu có ý nghĩa và tiến hành trực quan hóa thông qua các dashboard và phân tích. Đối với những yêu cầu truy vấn đưa ra phân tích phức tạp hơn, hay những dữ liệu khó trực quan hóa hơn, DE cần áp dụng các mô hình thuật toán, sử dụng các code độc lập hoặc nhúng dữ liệu vào các tool để thực hiện trực quan hóa dữ liệu từ đó đưa ra phân tích và đánh giá dữ liệu.

Các công cụ phổ biến được các DE sử dụng để thực hiện Visualization bao gồm: R, Python, Oracle BI, Power BI, Tableau…

# TRIỂN KHAI

## Crawl

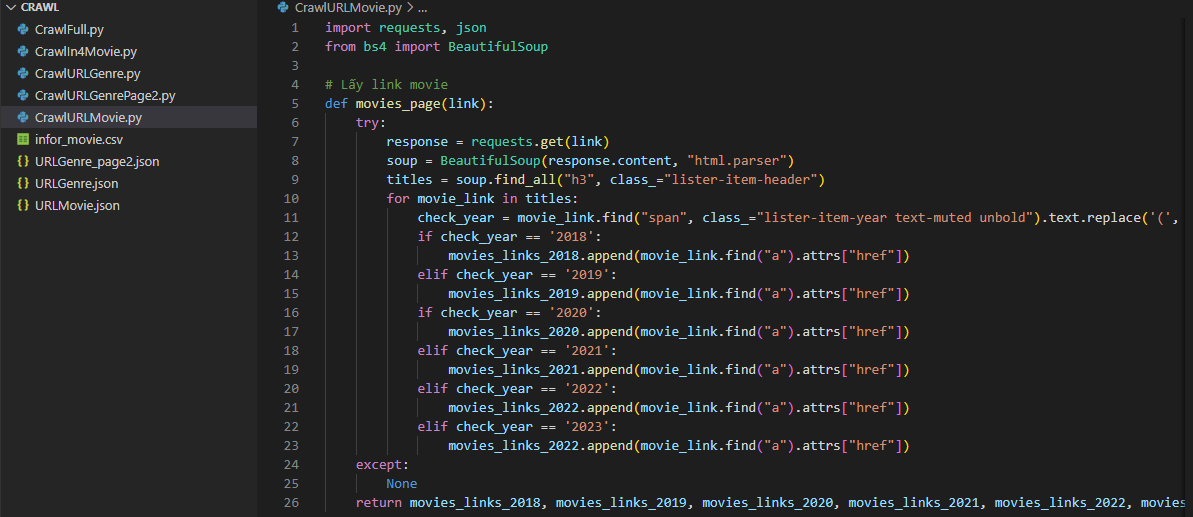
* Bước 1: Crawling URL từng thể loại thể loại phim (trang 1) và đưa vào file json (URLGenre.json)

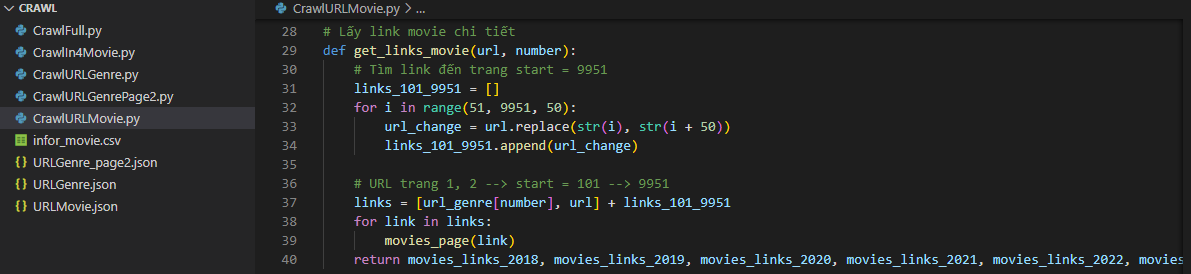


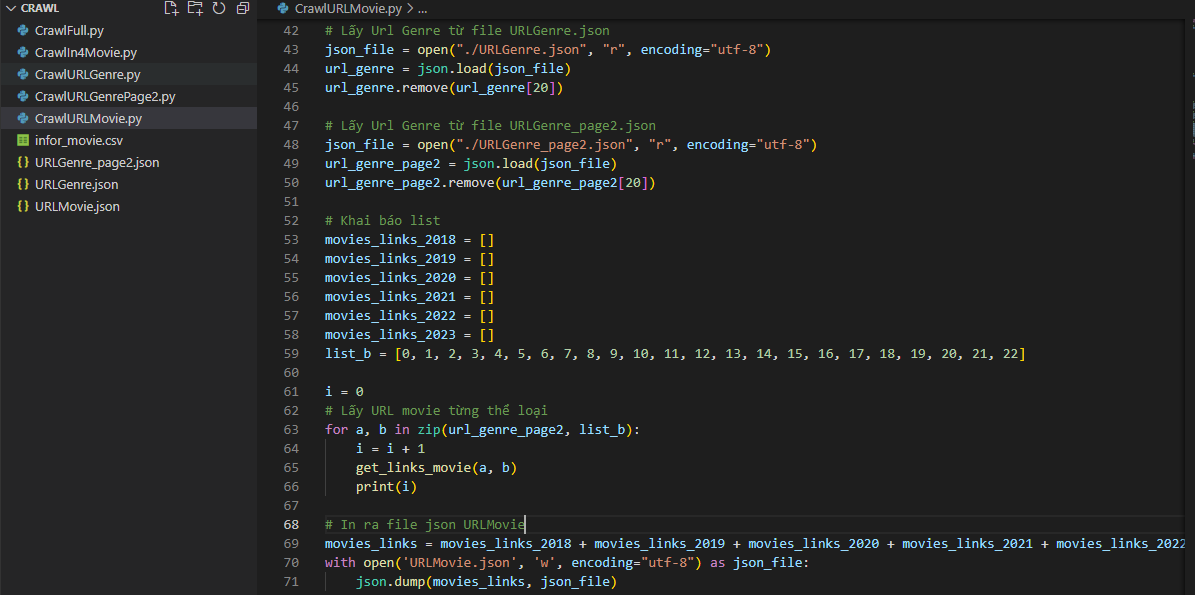
* Bước 2: Từ file json (URLGenre.json) lấy URL đưa vào và tiếp tục lấy URL thể loại phim (trang 2) và đưa vào file json URLGenre\_page2.json



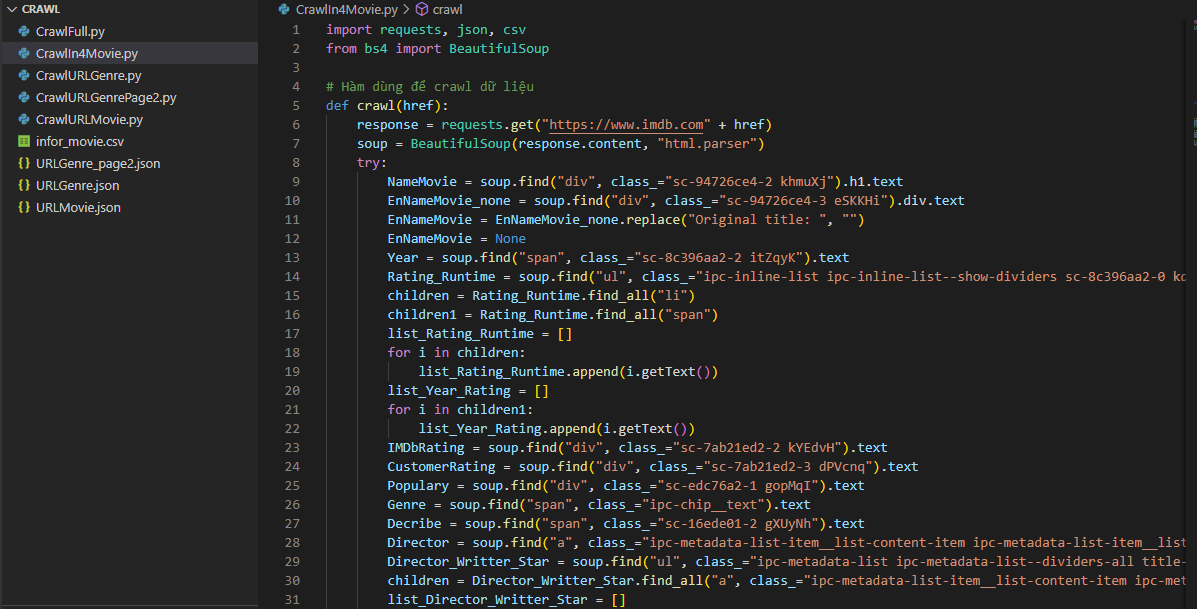
* Bước 3: Từ file json (URLGenre.json) lấy URL từng bộ phim dẫn đến trang phim và đưa vào URLMovie.json

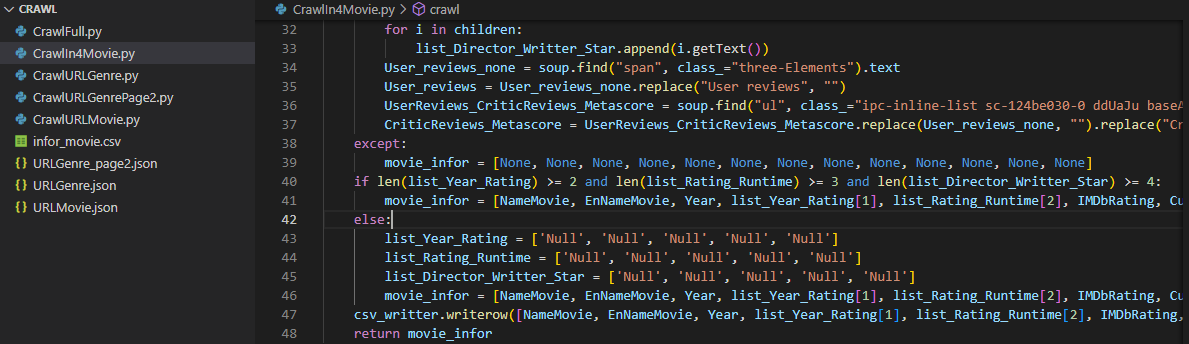


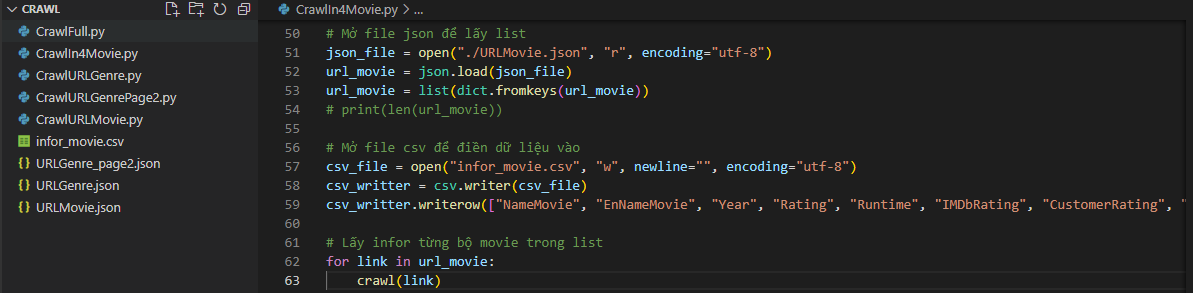




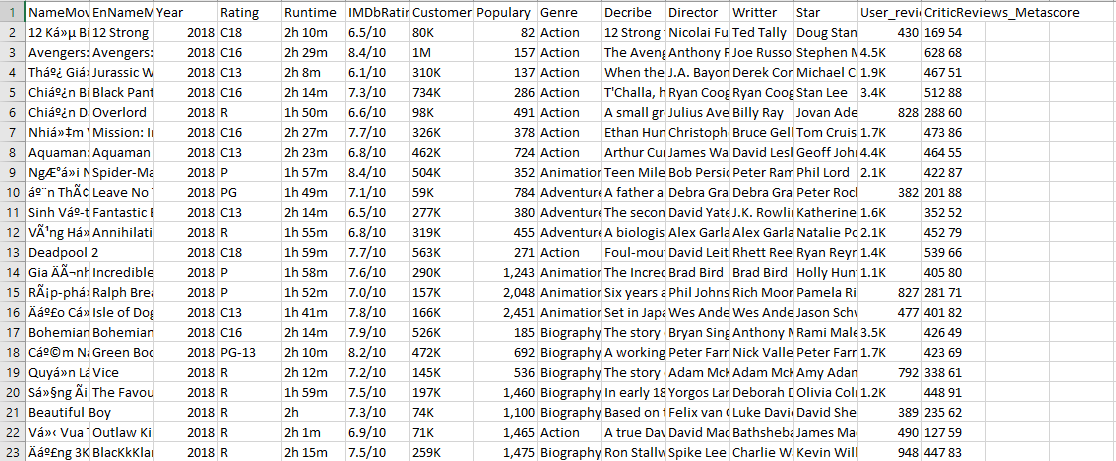
* Bước 4: Từ file json (URLMovie.json) lấy ra URL và truy cập vào trang phim để lấy thông tin sau đó xuất ra file CSV (infor\_movie.csv)



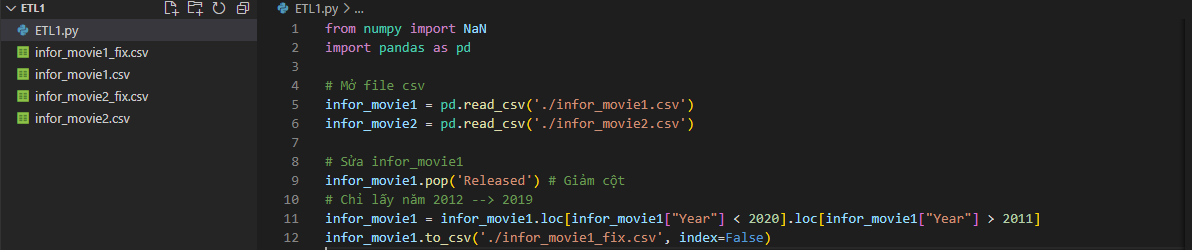


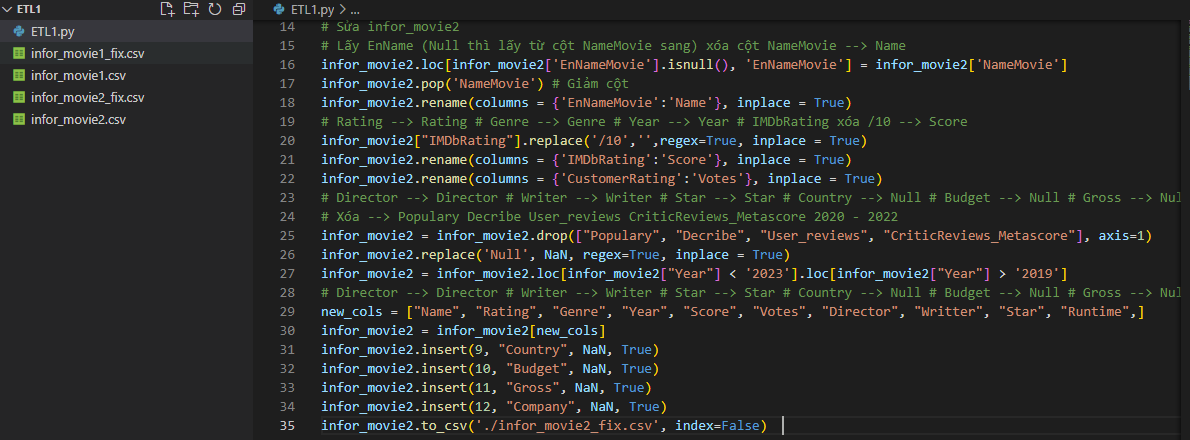


* File CSV sau khi crawling

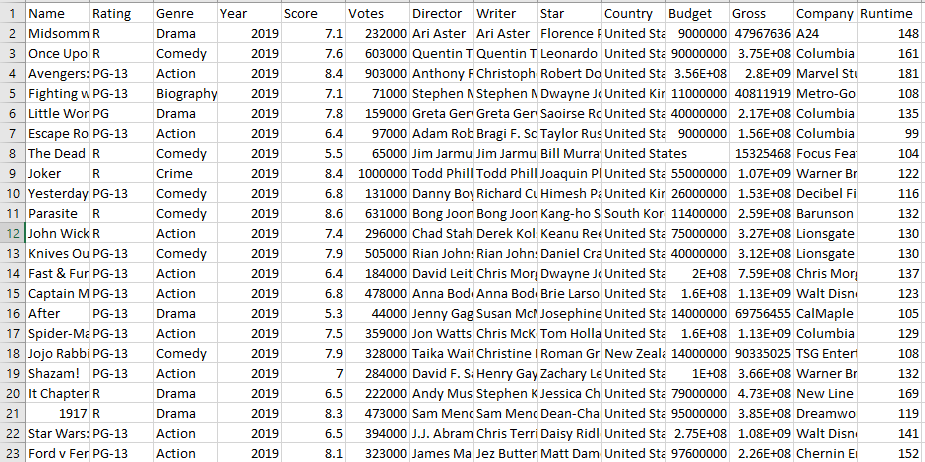


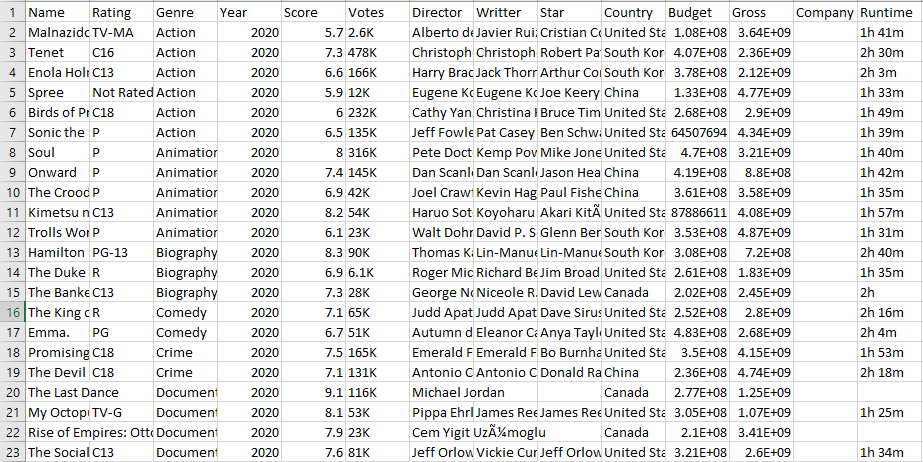
* Bước 5: Phân tách các cột và trường để thuận tiện cho việc xử lí dữ liệu





* Bảng sau khi phân tách hoàn chỉnh, hoàn thành công việc crawl dữ liệu và thu được 2 bảng dữ liệu (infor\_movie1\_fix.csv, infor\_movie2\_fix.csv) sau đó đưa đến cho ETL bắt đầu công việc xử lý

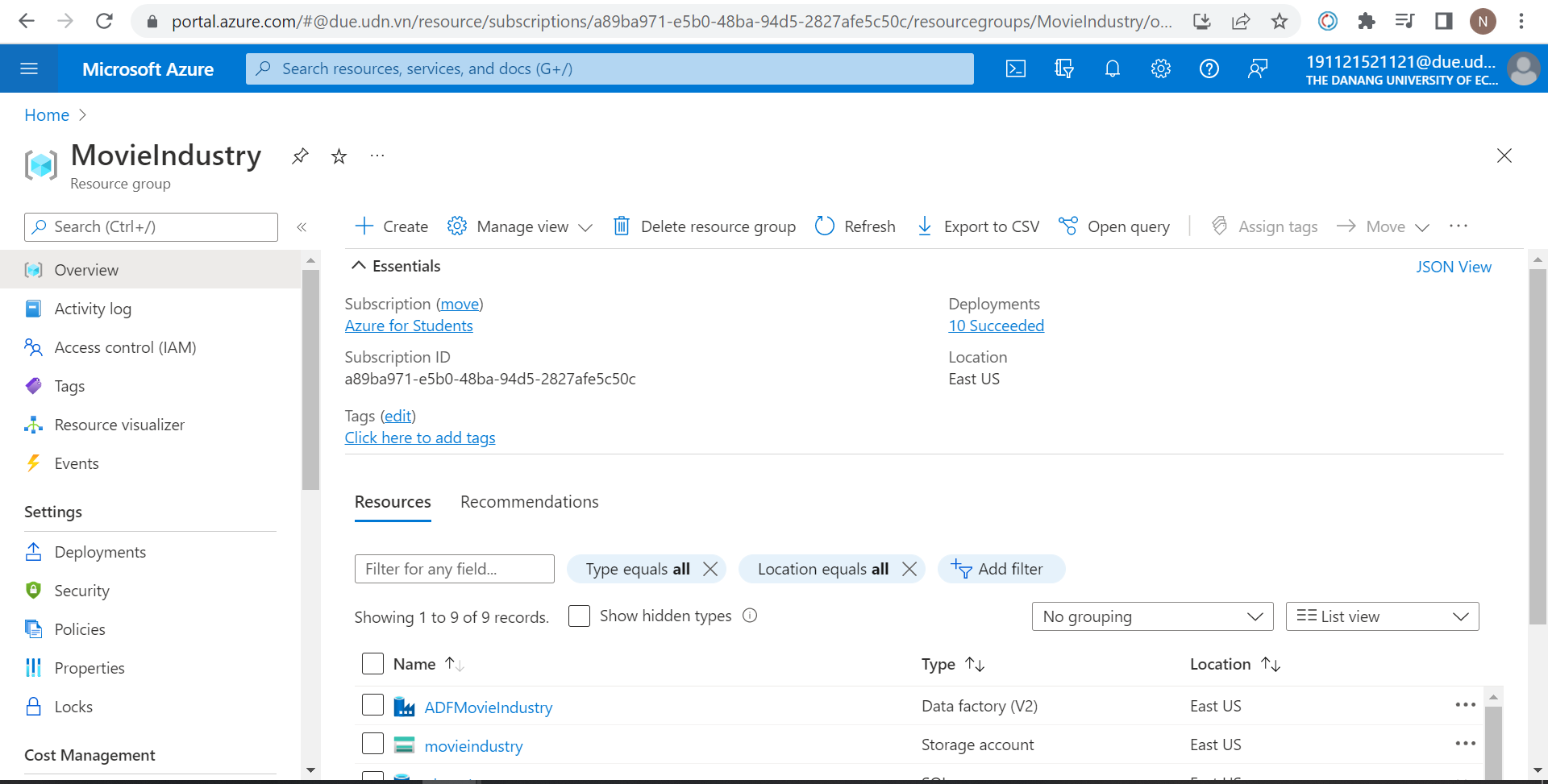




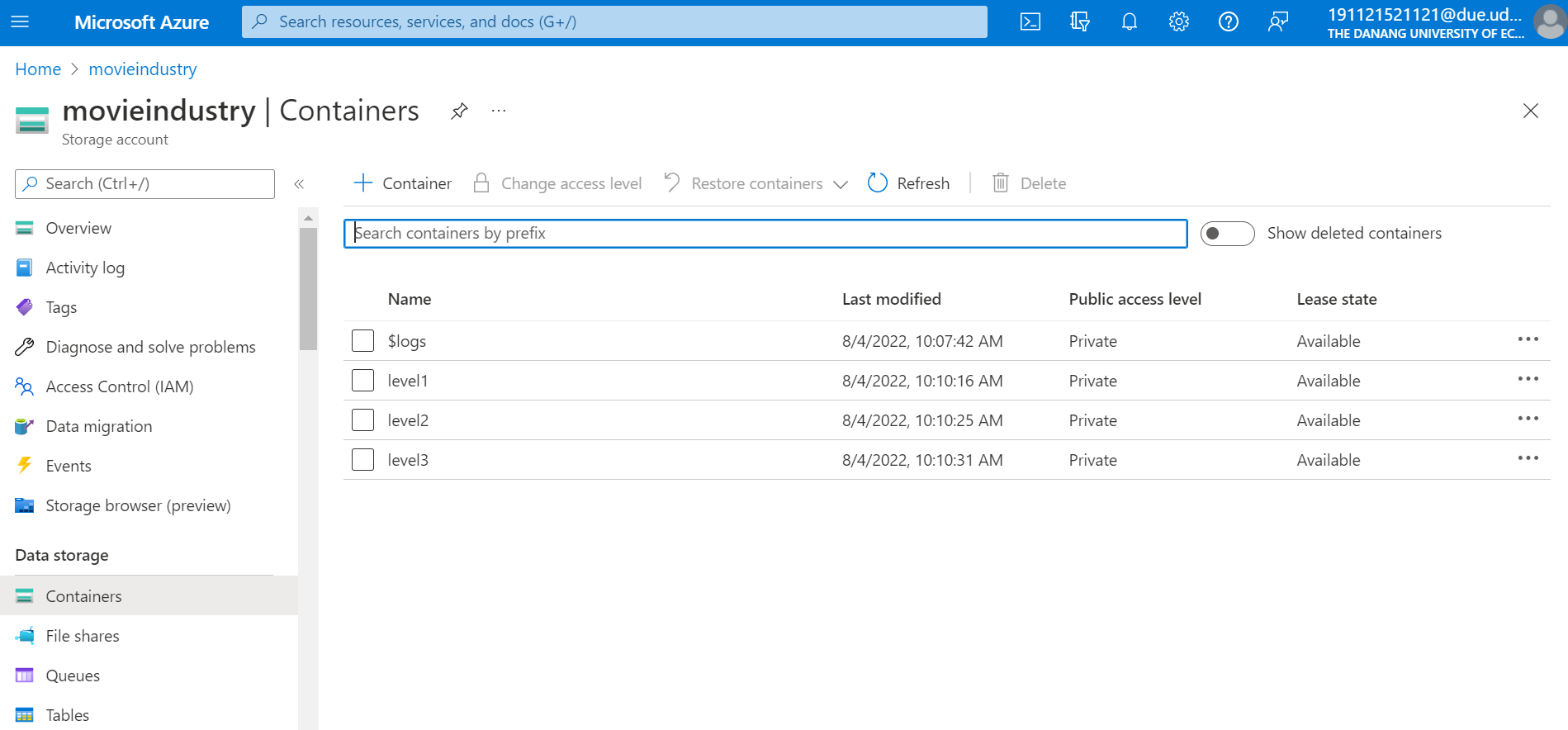
## ETL

Sau khi thực hiện cào dữ liệu, nhóm có dữ liệu thô từ năm 2020 đến năm 2022 và một bộ dữ liệu chưa được xử lý từ năm 2012 đến năm 2020. Để xử lý dữ liệu, nhóm thực hiện xử lý trên nền tảng cloud – Azure của Microsoft.

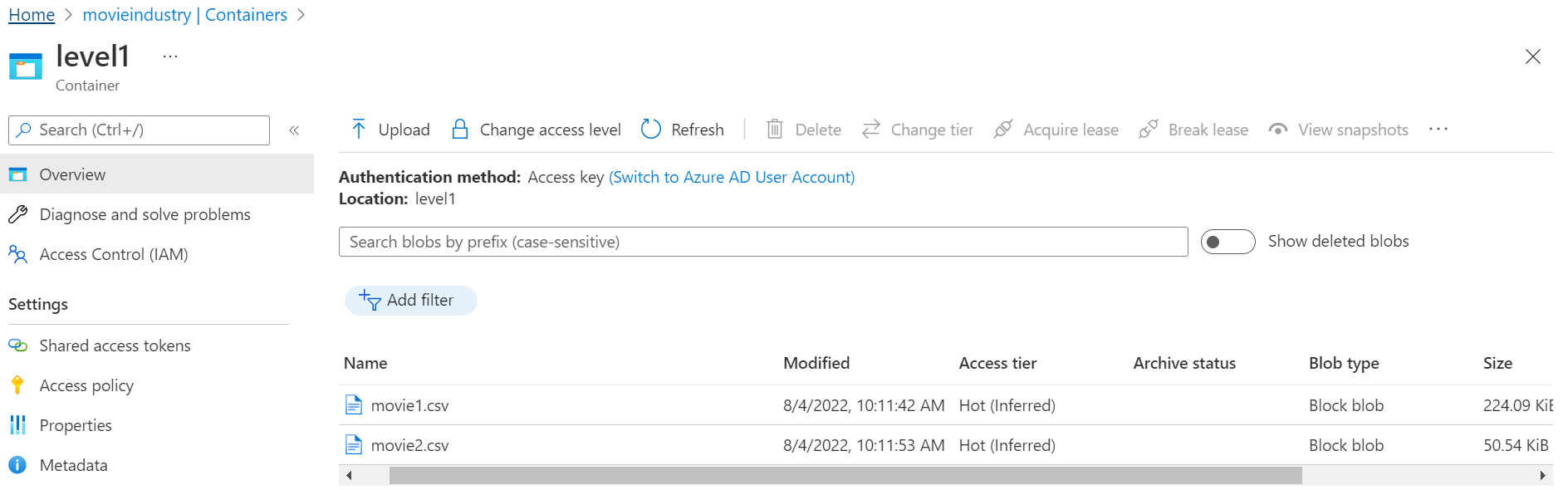
Đầu tiên, thực hiện trên tài khoản Microsoft Azure Portal, tiến hành xây dựng một Resourse Group tên là “MovieIndustry” để làm nền tảng và giúp quản lý những dịch vụ trong quá trình làm.



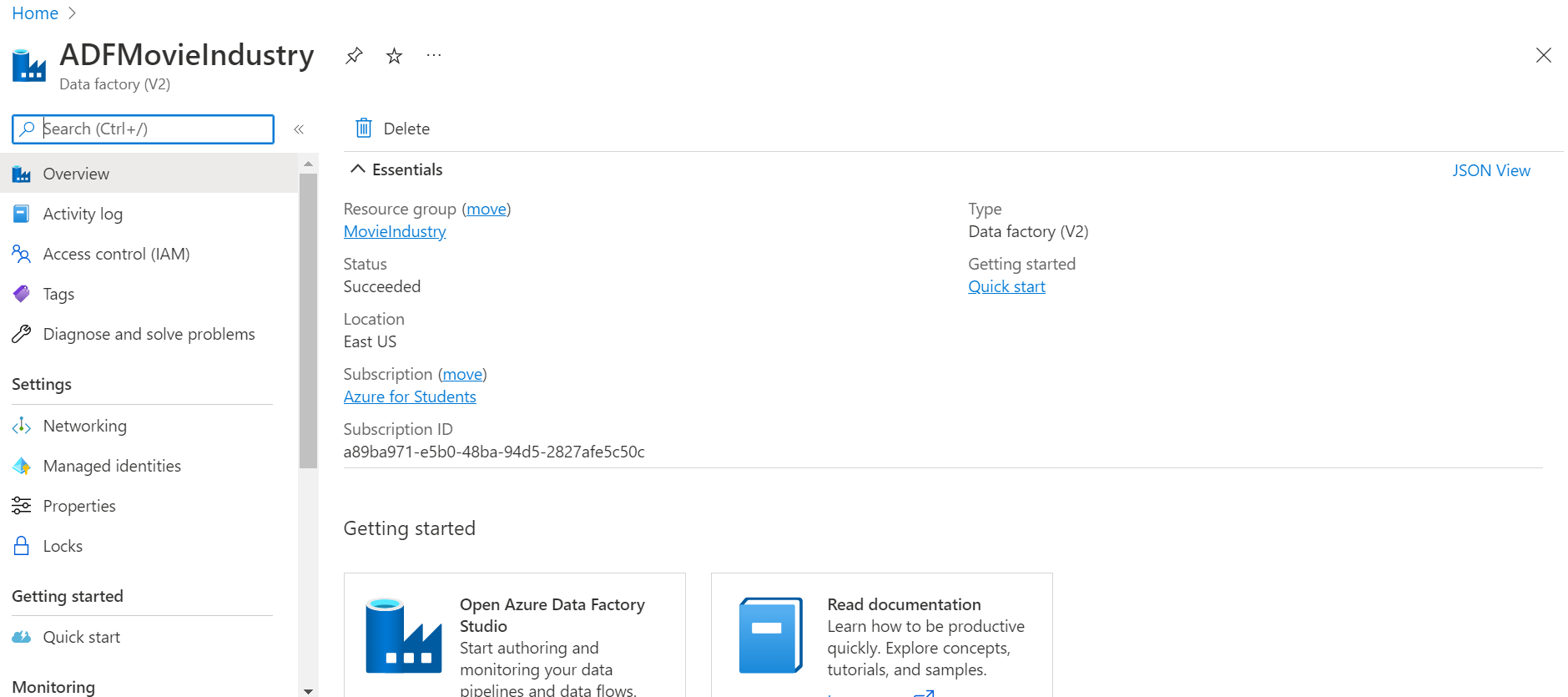
Nhóm tiến hành lưu trữ các file dữ liệu trên Azure Data Lake Storage v2 (ADLS). Tại đây, thực hiện tạo 3 containers (file system): level1, level2, level3 trên storage account để lưu trữ dữ liệu thành các vùng riêng biệt. Tại container level1, sẽ là nơi lưu trữ dữ liệu thô. Container level 2 sẽ là nơi tạo nhóm copy dữ liệu thô từ level 1 và thực hiện xử lý dữ liệu (ETL) tại đây thông qua DataBricks. Level 3 là nơi lưu trữ dữ liệu đã được làm sạch và sẵn sàng để phục vụ cho quá trình trực quan hóa dữ liệu.

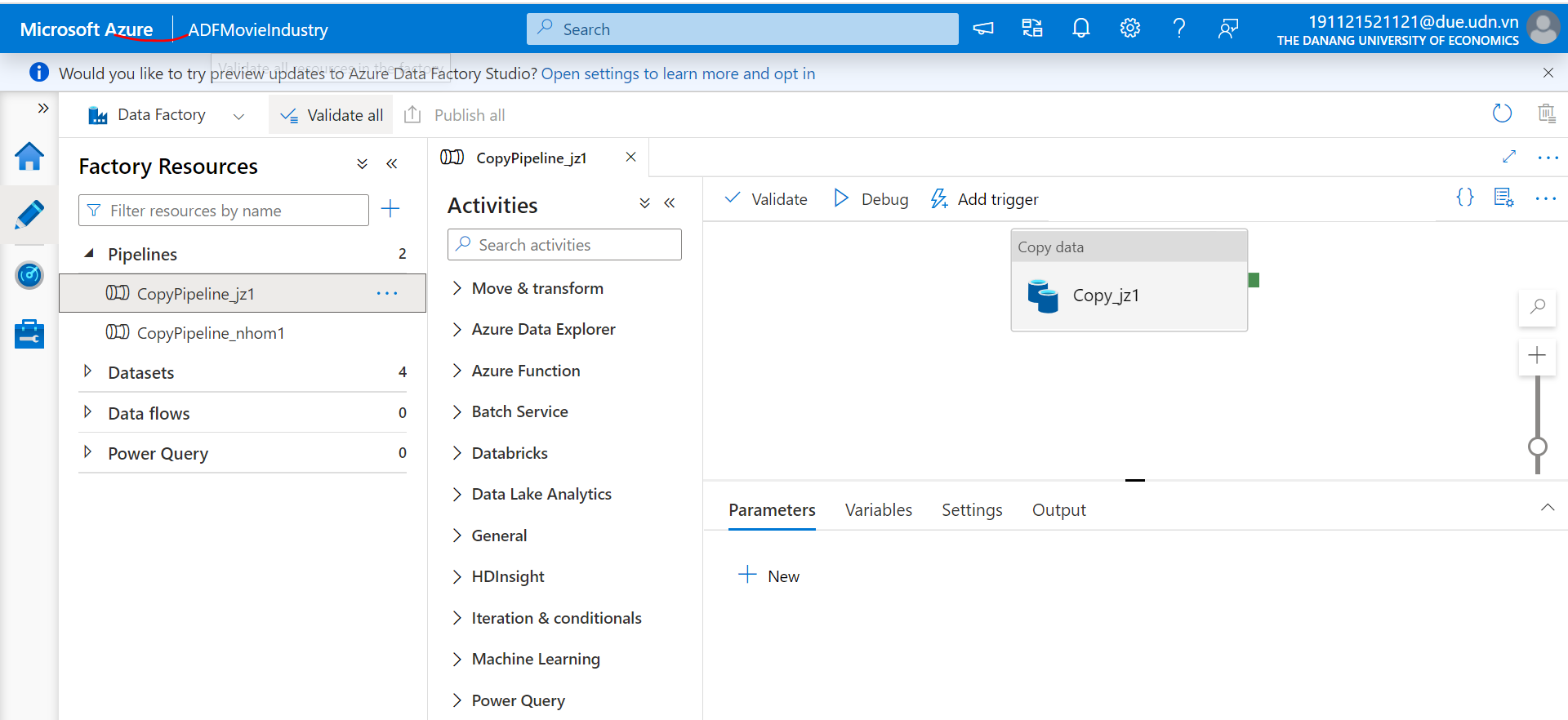


Dữ liệu thô được đưa lên lưu trữ tại Level 1

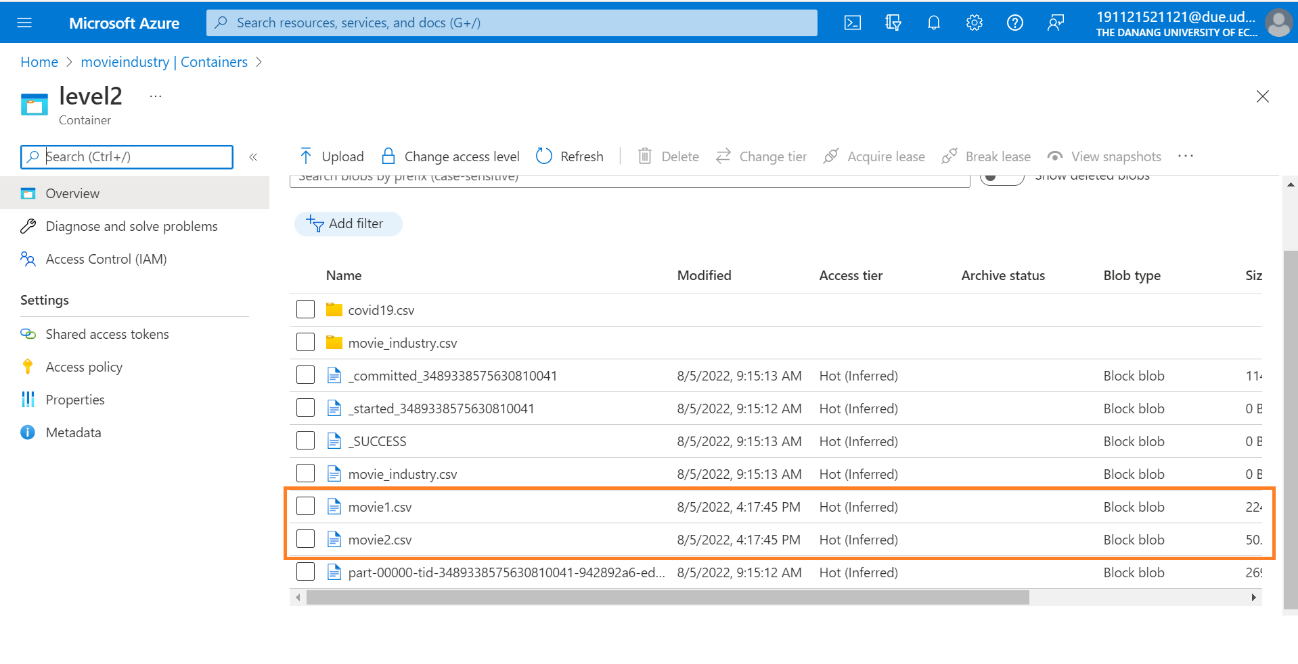


Để bước qua xử lý dữ liệu tại level 2, nhóm thực hiện tạo một Data Factory để di chuyển dữ liệu từ level 1 qua level 2.

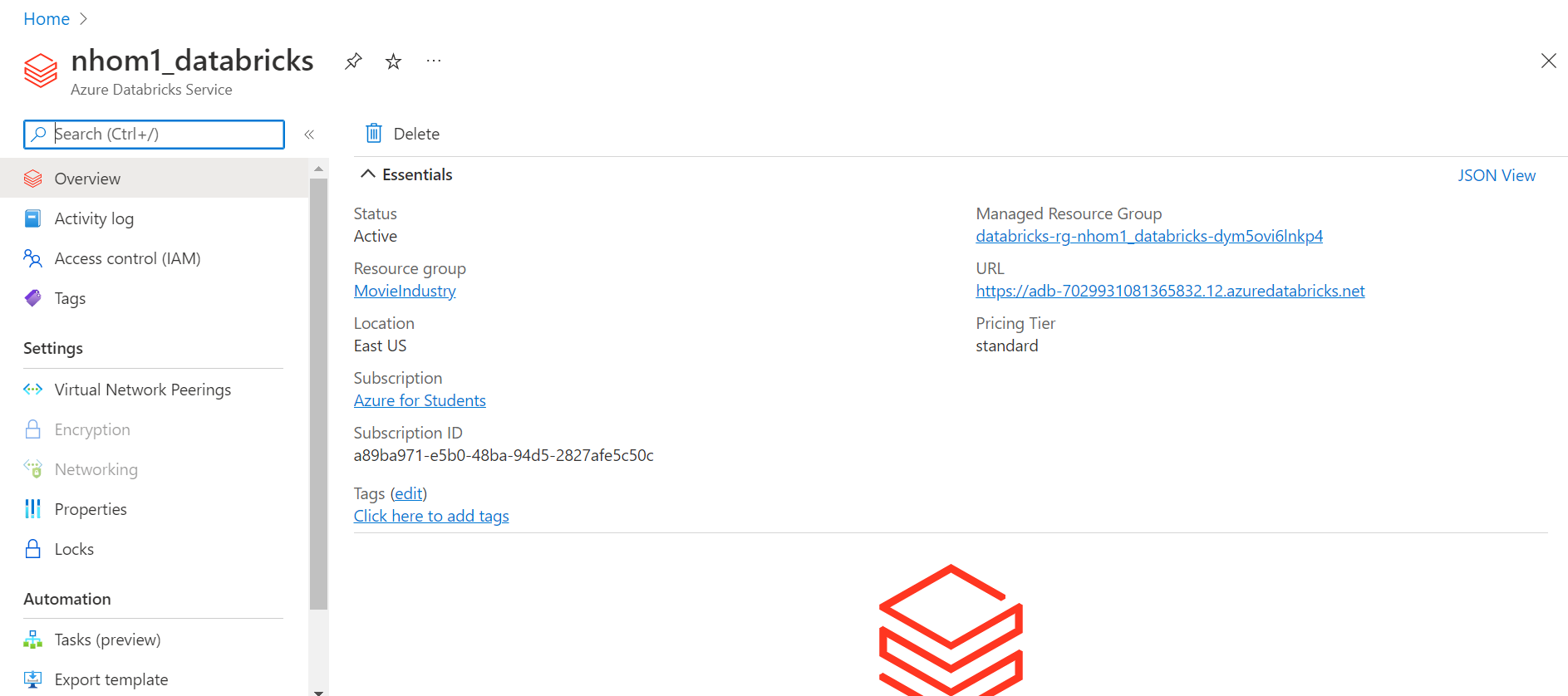


Để chuyển dữ liệu, sử dụng công cụ Copy Data để tạo một đường dẫn sao chép dữ liệu từ một thư mục trong bộ lưu trữ Azure Blob sang một thư mục khác.

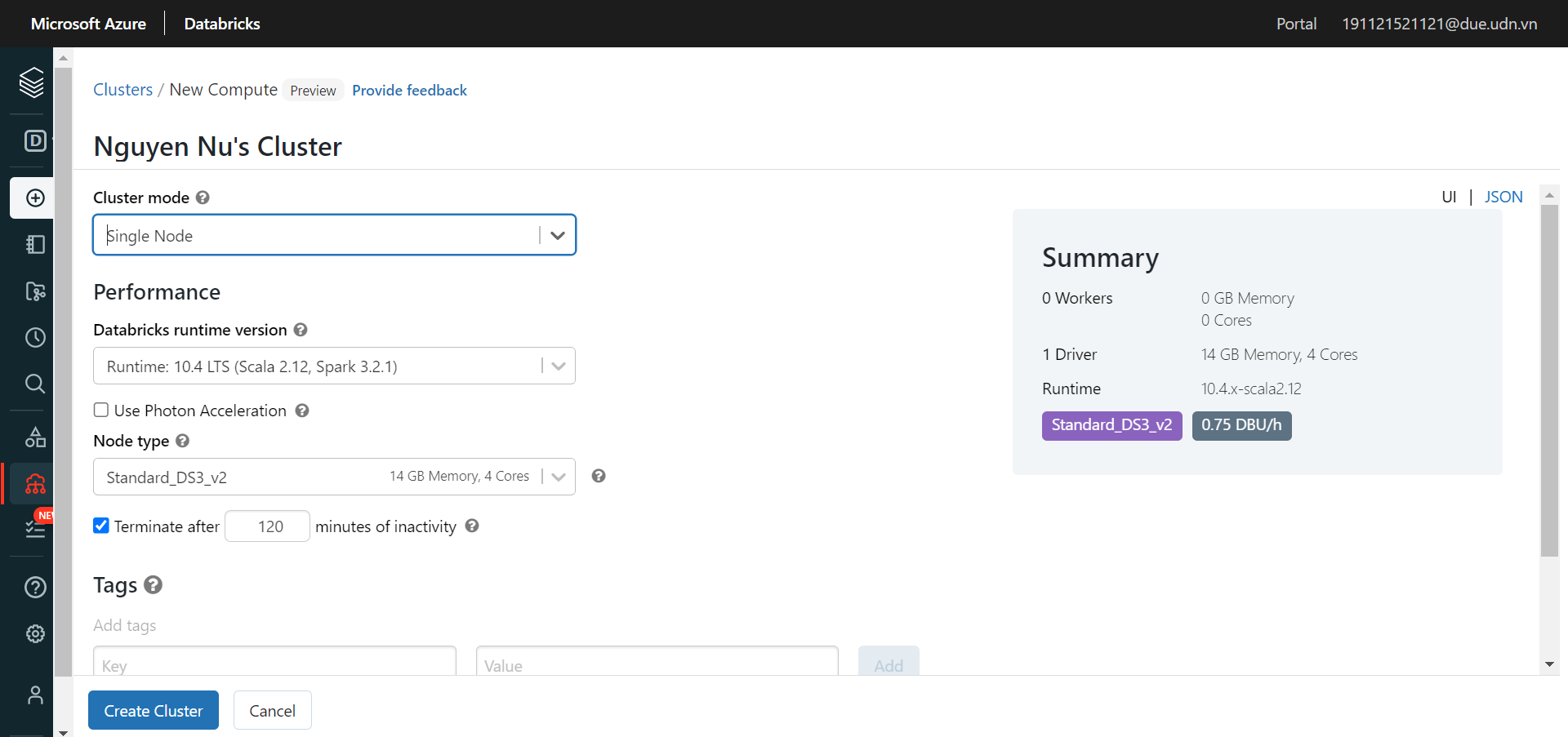
Sau khi copy, tại level 2 đã có 2 file dữ liệu



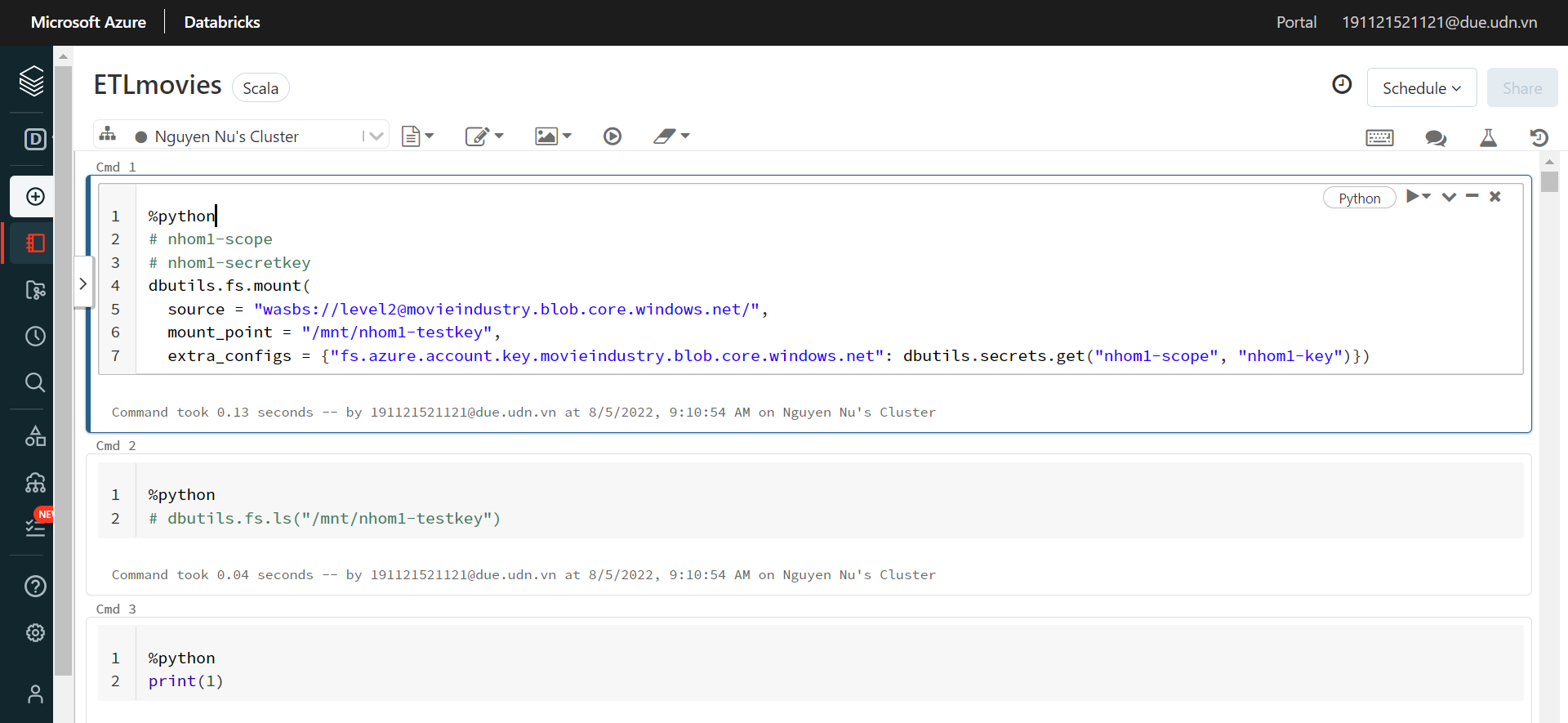
Để xử lý 2 file dữ liệu này, nhóm thực hiện tạo DataBricks có tên “” để thực hiện ETL



Sau đó, tiến hành tạo Lanch WorkSpace để tạo Cluster và sử dụng.

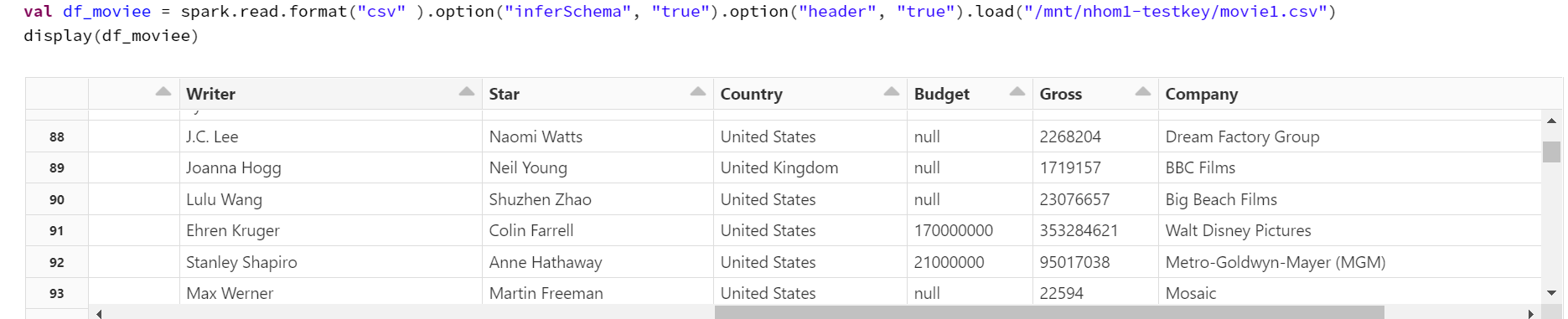


Tiếp theo, tiến hành tạo Notebook để thực hiện xử lý đọc dữ liệu và ETL dữ liệu trên Notebook



Tại đây, nhóm thực hiện xử lý dữ liệu bằng ngôn ngữ Scala và Python. Hiện tại, dữ liệu đang bị null tại các trường dữ liệu với số lượng như sau:

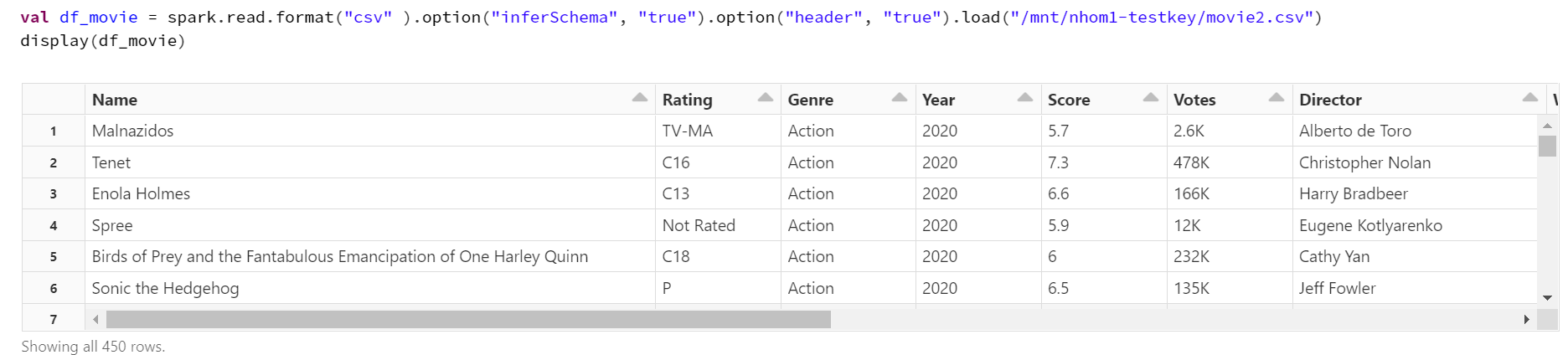
Dữ liệu 2012 – 2019



Dữ liệu có số lượng null như sau:



Dữ liệu 2019 – 2022

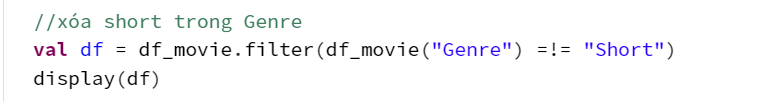


Dữ liệu có số lượng null như sau:

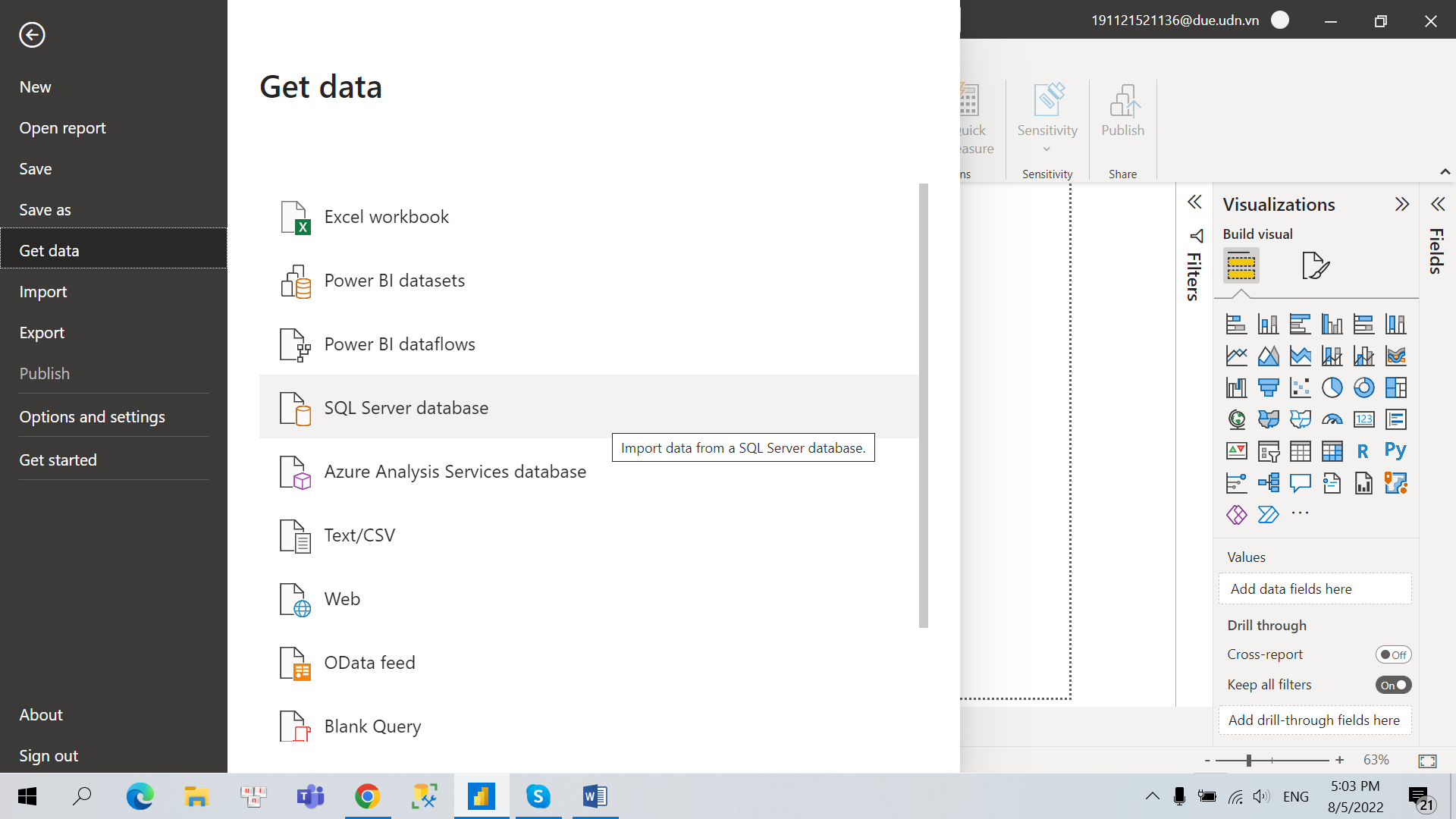


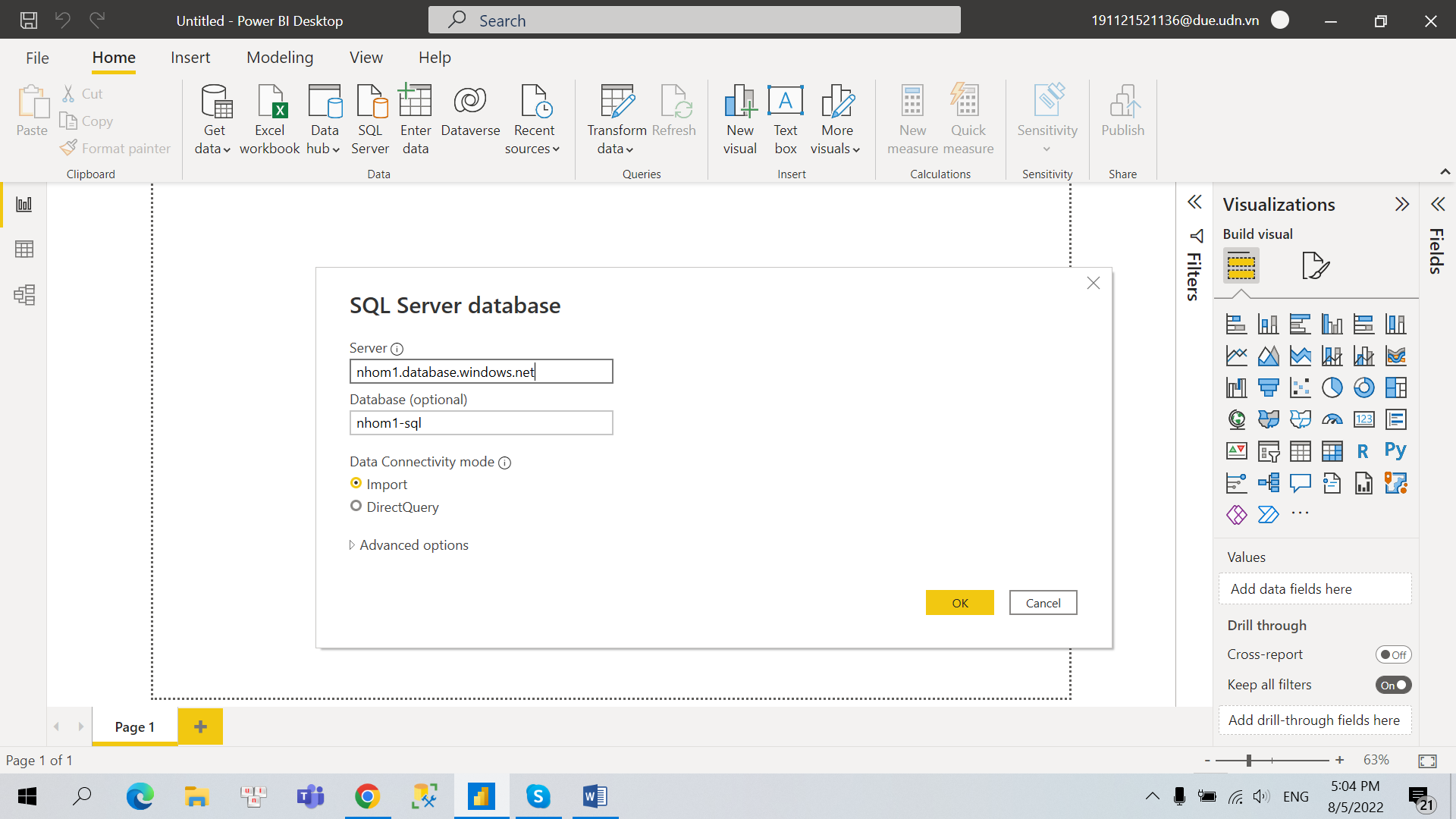


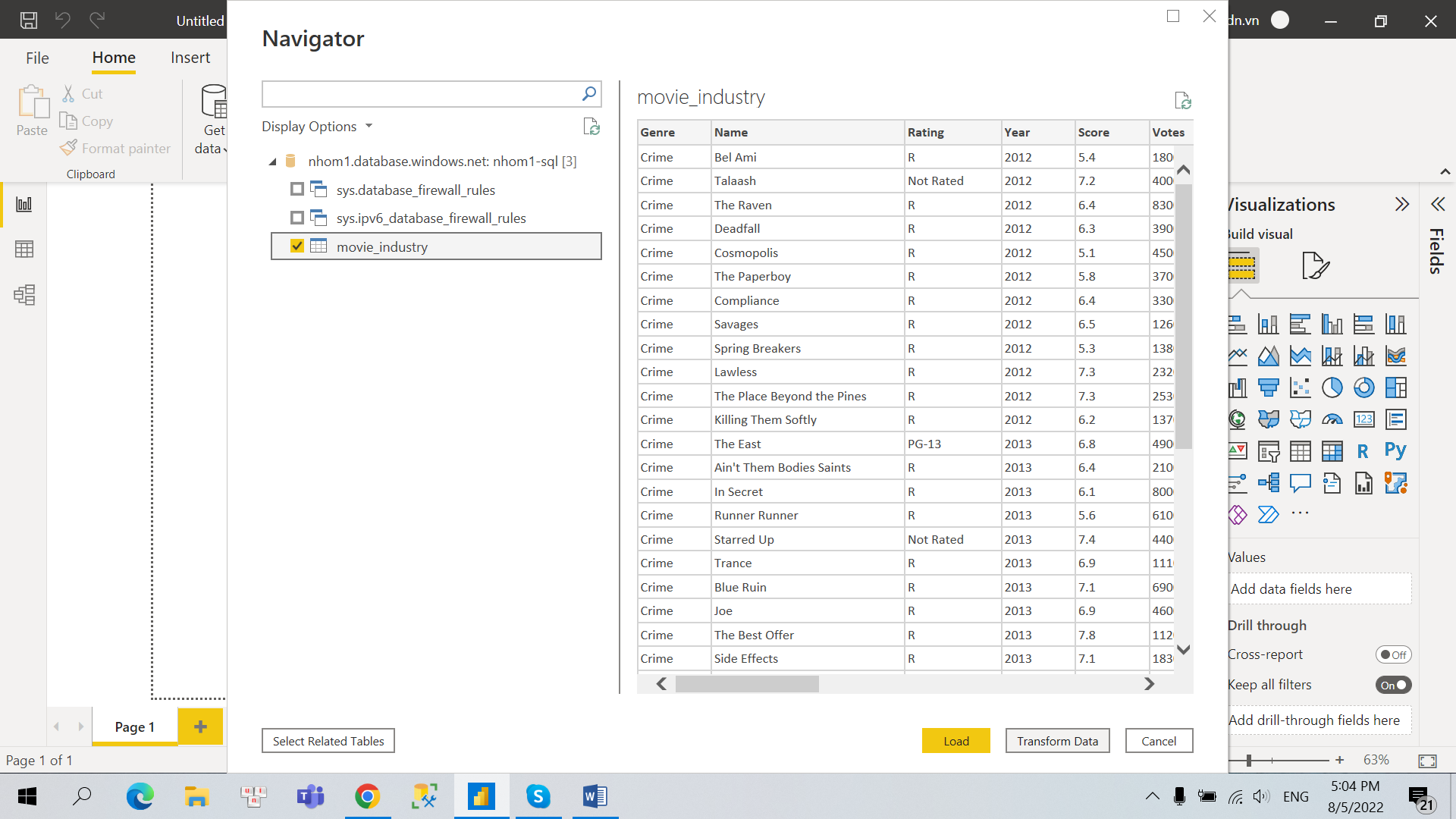
Đồng thời, giữa hai dataset có các kiểu dữ liệu không cùng định dạng ở các trường: Votes, Runtime. Ngoài ra, trong dataset có thể loại phim là Short không cùng giá trị phân tích nên sẽ bị loại bỏ.

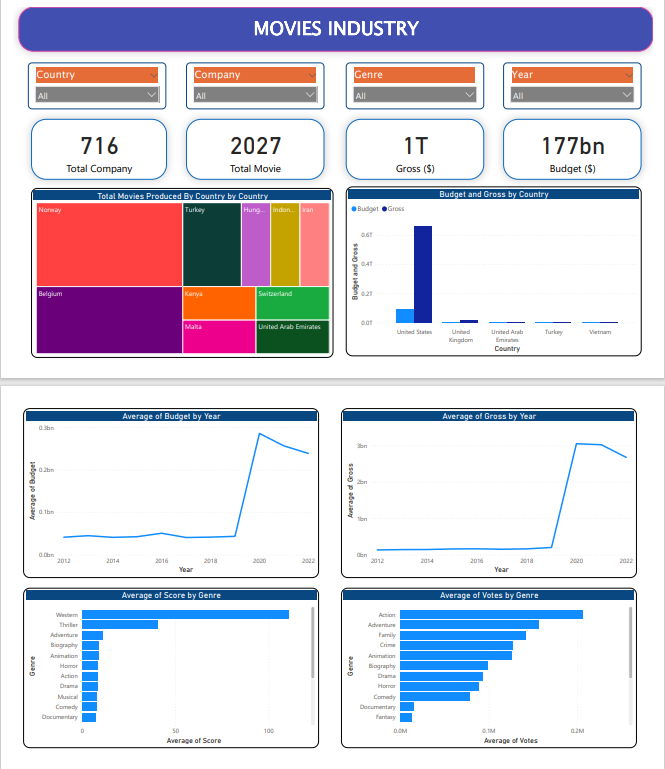


## Dashboard









## PowerBI

# KẾT QUẢ

## Mục 4.1…

## Mục 4.2…

Kết quả được xây dựng dựa trên ngôn ngữ lập trình thể hiện như (Bảng 4.1).

Bảng 4.1 Kiến trúc

|  |  |  |
| --- | --- | --- |
| Lớp | Ý nghĩa | Tham số |
|  |  |  |
|  |  |  |
|  | | |

Hình 4.1 Kiến trúc của mô hình

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Đề tài đã thực hiện được …

# TÀI LIỆU THAM KHẢO

* + - 1. Tên tác giả, Tên tác giả… *Tên công trình*. Nơi xuất bản, năm, trang
      2. Tên tác giả, Tên tác giả… *Tên công trình*. Nơi xuất bản, năm, trang

# PHỤ LỤC