**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**

**PHÂN TÍCH DỮ LIỆU VỀ DOANH SỐ GAME QUA CÁC NĂM**

Sinh viên thực hiện : Trương Thị Ngọc Viên

Nguyễn Bá Khánh

Lớp : 45K21.1

Đơn vị thực tập : TMA Solution

Cán bộ hướng dẫn : Nguyễn Thanh Bình

Lê Bá Thiện

Giảng viên hướng dẫn : Ths.Cao Thị Nhâm

**Đà Nẵng, 8/2022**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

Họ và tên sinh viên:

Lớp: Khoa: Trường:

Thực tập từ ngày: …./……/ 2022 đến ngày: ........./ .……./ 2022

Tại:

Địa chỉ:

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:

……….., ngày .......tháng ......năm 2022

Xác nhận của đơn vị thực tập

# LỜI CẢM ƠN

Mười tuần thực tập ngắn ngủi là cơ hội cho em tổng hợp và hệ thống hóa lại những kiến thức đã học về Data, đồng thời kết hợp với thực tế để nâng cao kiến thức chuyên môn. Tuy chỉ có mười tuần thực tập, nhưng qua quá trình thực tập, em đã được mở rộng tầm nhìn và tiếp thu rất nhiều kiến thức thực tế. Từ đó em nhận thấy, việc cọ sát thực tế là vô cùng quan trọng – nó giúp sinh viên xây dựng nền tảng lý thuyết được học ở trường vững chắc hơn. Trong quá trình thực tập, từ chỗ còn bở ngỡ cho đến thiếu kinh nghiệm, em đã gặp phải rất nhiều khó khăn nhưng với sự giúp đỡ tận tình của quý thầy cô khoa và sự nhiệt tình của các anh chị đến từ TMA Solutions đã giúp em có được những kinh nghiệm quý báu để hoàn thành tốt kì thực tập này cũng như viết lên bài báo cáo cuối kỳ. Em xin chân thành cám ơn.

Lời cám ơn đầu tiên em xin gửi đến ban lãnh đạo cùng các phòng ban, các anh chị trong TMA Solutions – đơn vị đã tiếp nhận và nhiệt tình tạo mọi điều kiện thuận lợi cho em tiếp cận, hiểu rõ về cách xử lý dữ liệu.

Em cũng xin gửi lời cám ơn chân thành đến, quý thầy cô khoa Thống Kê – Tin Học đã tận tâm giảng dạy và truyền đạt những kiến thức, kinh nghiệm quý báu cho em. Đặc biệt, em xin cán ơn cô Cao Thị Nhâm , người đã tận tình hướng dẫn em hoàn thành bài báo cáo này.

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 thầy cô và các bạn, để em rút kinh nghiệm và hoàn thành tốt hơn. Em xin chân thành cám ơn!

# LỜI CAM ĐOAN

Em xin cam đoan đề tài: “Báo cáo thực tập xử lý và phân tích xu hướng quảng cáo Game qua Video” là một công trình nghiên cứu độc lập dưới sự hướng dẫn của giáo viên hướng dẫn: Ths Cao Thị Nhâm. Ngoài ra không có bất cứ sự sao chép của người khác. Đề tài, nội dung báo cáo thực tập là sản phẩm mà em đã nỗ lực nghiên cứu trong quá trình học tập tại trường cũng như tham gia thực tập tại công ty TMA Solutions Bình Định. Các số liệu, kết quả trình bày trong báo cáo là hoàn toàn trung thực, em xin chịu hoàn toàn trách nhiệm, kỷ luật của bộ môn và nhà trường đề ra nếu như có vấn đề 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**

[**LỜI CẢM ƠN i**](#_heading=h.2et92p0)

[**LỜI CAM ĐOAN ii**](#_heading=h.tyjcwt)

[**DANH MỤC HÌNH ẢNH vi**](#_heading=h.1t3h5sf)

[**DANH MỤC BẢNG BIỂU vii**](#_heading=h.4d34og8)

[**DANH MỤC CÁC TỪ VIẾT TẮT viii**](#_heading=h.2s8eyo1)

[**LỜI MỞ ĐẦU 1**](#_heading=h.1ksv4uv)

[**CHƯƠNG 1. TỔNG QUAN LÝ THUYẾT 2**](#_heading=h.2jxsxqh)

[**1.1**](#_heading=h.3o7alnk) **Azure và Azure DataBricks** 2

[**1.1.1**](#_heading=h.23ckvvd)**Microsoft Azure *2***

[***1.1.2.***](#_heading=h.ihv636) ***Databrick sử dụng nền tảng Azure 7***

[***1.2.***](#_heading=h.32hioqz) ***Data Egineer*** 7

[***1.2.1.***](#_heading=h.1hmsyys) ***Giới thiệu tổng quan về data egineer 7***

[***1.2.2.***](#_heading=h.z337ya) ***Thu thập dữ liệu - Crawling 7***

[***1.2.3.***](#_heading=h.3j2qqm3) ***Xử lý dữ liệu – ETL (Extract - Transform – Load) 7***

[***1.2.4.***](#_heading=h.1y810tw) ***Trực quan hoá và phân tích dữ liệu bằng Power BI 9***

[**CHƯƠNG 2. TRIỂN KHAI … 11**](#_heading=h.41mghml)

[2.1.](#_heading=h.2grqrue) Mục 3.1 11

[***2.1.1.***](#_heading=h.vx1227) ***Mục 3.1.1 11***

[***2.1.2.***](#_heading=h.3fwokq0) ***Mục 3.1.2 11***

[2.2.](#_heading=h.1v1yuxt) Mục 3.2 11

[**CHƯƠNG 3. KẾT QUẢ 12**](#_heading=h.4i7ojhp)

[3.1.](#_heading=h.2xcytpi) Mục 4.1… 12

[3.2.](#_heading=h.1ci93xb) Mục 4.2… 12

[**KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 13**](#_heading=h.3as4poj)

[**TÀI LIỆU THAM KHẢO 14**](#_heading=h.2p2csry)

[**PHỤ LỤC 15**](#_heading=h.147n2zr)

# DANH MỤC HÌNH ẢNH

[**Hình 1.1 Ngôn ngữ lập trình Python 2**](#_heading=h.4f1mdlm)

[**Hình 4.1 Kiến trúc của mô hình 5**](#_heading=h.2bn6wsx)

# DANH MỤC BẢNG BIỂU

[**Bảng 4.1 Kiến trúc 5**](#_heading=h.3whwml4)

# DANH MỤC CÁC TỪ VIẾT TẮT

**AI : Artificial Intelligence**

**ADF : Azure Data Factory**

**DBFS : Databricks File System**

**ETL : Extrac Transform Loading**

# LỜI MỞ ĐẦU

Ngày nay với sự phát triển không ngừng của lĩnh vực Công nghệ thông tin đã thay đổi cuộc sống của con người. Từ các hoạt động kinh doanh, buôn bán, xây dựng hay đầu tư phát triển cũng đều ứng dụng Công nghệ thông tin vào trong quá trình vận hành. Những ứng dụng được xây dựng nhằm phục vụ cho quá trình này rất đa dạng các nền tảng như: Website, Mobi,… Các ứng dụng trong quá trình sử dụng sẽ sinh ra rất nhiều dữ liệu, từ những dữ liệu cố định được xậy dựng từ đầu, các dữ liệu sẽ được cập nhật xuyên suốt trong cả quá trình vận hành của ứng dụng.

Dữ liệu là một nguồn tai nguyên quý giá trong việc phân tích và xây dựng các hệ thống. Đặc biệt phải kể đến hoạt động kinh doanh, dữ liệu nắm giữ vai trò quan trọng trong mọi hoạt động, phát triển của doanh nghiệp từ việc phân tích, thấu hiểu, nghiên cứ thị trường, đưa ra các báo cáo, hỗ trợ xây dựng các hoạt động quản trị, … Nhưng để có dữ liệu thực hiện những điều này không phải đơn giản. Khi các nguồn để tổng hợp đến từ nhiều nơi khiến dữ liệu không đồng bộ về mặt cấu trúc, lỗi dữ liệu, … Giải quyết điều đó cần phải có hoạt động xử lý và làm sạch dữ liệu trước khi tiến hành phân tích, xây dựng hệ thống.

Thông qua những tìm hiểu lý thuyết cơ sở, nhóm thực hiện bài báo cáo “Phân tích xu hướng phát triển Game” nhằm làm rõ hoạt động xử lý dữ liệu, rút ra các kết luận liên quan đến dữ liệu thu được nhằm cải thiện hoạt động kinh doanh trong lĩnh vực sản xuất và phát triển Game. Báo cáo bao gồm một số nội dung chính:

Chương 1: Tổng quan về lý thuyết

Chương 2: Triển khai

Chương 3: Kết quả

# TỔNG QUAN LÝ THUYẾT

* 1. **Tổng quan về Data Engineer**

Ngày nay với sự phát triển không ngừng của khoa học, công nghê. Khi xã hội ngày càng phát triển thì càng có những công việc xuất hiện nhiều hơn. Đặc biệt là công việc liên quan đến công nghệ thông tin. Data engineer – kỹ sư dữ liệu là một thuật ngữ để chỉ những người làm việc trong lĩnh vực dữ liệu. Kỹ sư dữ liệu là người xây dựng một systems tổng hợp lưu trữ và xuất dữ liệu từ một ứng dụng, hệ thống đã được tạo ra trước đó. Các kỹ sư dữ liệu thực hiện các công việc chính:

* Làm việc liên quan đến hệ thống điện toán đám mây
* Lập trình
* Trao dồi kiến thức về các công cụ mới
* Thực hiện các công việc tạo ETL/data piplines
* Cấu trúc dữ liệu nâng cao

Những công cụ mà kỹ sư dữ liệu thường sử dụng là Databricks và Python. Họ sẽ phải thường xuyên sử dụng các packages như ggplot2 khi làm việc với các công cụ này, trực quan hóa dự liệu với Power BI hoặc thao tác với thư viện Pandas Python. Một phần khác các công cụ như SAS hay SPSS sẽ xử lý tốt các công việc liên quan đến dữ liệu

Data engineer thực hiện thu thập và lưu trữ dữ liệu được thiết kế và thực hiện, xử lý trước dữ liệu và chuyển đổi nó thành một định dạng có thể sử dụng được. Họ xây dựng dữ liệu và đảm bảo rằng dữ liệu trôi chảy. Để thực hiện được những điều này Data engineer cần có những kỹ năng cơ bản sau:

* Kỹ năng lập trình cơ bản
* Khả năng tập trung cao độ
* Kỹ năng phân tích logic
* Kỹ nẵng thiết kế báo cáo
* Kỹ năng giao tiếp, trình bày
* Biết cách quản lý thời gian, sắp xếp công việc
* Có thể kể câu chuyện từ số liệu

## Các công cụ thực hiện

* + 1. **Microsoft Azure**

1. **Tổng quan về Azure và nền tảng điện toán đám mây**

Hiện nay các nền tảng điện toán đám mây đang phát triển nhanh chóng và đóng vài trò quan trọng trong cơ sở hạ tầng công nghệ thông tin. Hai cái tên luôn được nhắc đến đầu tiên là Amazon Web Services và Microsoft Azure, tuy không được phủ sóng rộng như Amazon Web Services nhưng Azure vẫn được đánh giá cao với tốc độ phát triển vượt bật

Microsoft Azure là một nền tảng điện toán đám mây công cộng, cung cấp giải pháp gồm Cơ sở hạ tầng dưới dạng dịch vụ - (IaaS), nền tảng dưới dạng dịch vụ (PaaS) và phần mềm dưới dạng dịch vụ (SaaS). Azure có thể được sử dụng cho các dịch vụ như phân tích, ‘ảo hóa’ máy tính, lưu trữ mạng, …

* **Cơ sở hạ tầng dưới dạng dịch vụ - Infrastructure as a Service (IaaS):** Azure cho người sử dụng xây dựng và tự quản lý các máy ảo, không cần quan tâm đến viêc quản lý server. Giảm bớt chi phí vốn để tập trung hơn vào vấn đề kinh doanh.
* **Nền tảng dưới dạng dịch vụ - Platform as a Service (PaaS):** Azure quản lý hết từ các máy ảo được người dùng xây dựng đến Server, tự động thiết đặt thêm các dịch vụ cho người dùng. Chủ yếu dùng để lưu trữ database với tính năng backup cùng tính hữu dụng cao.
* **Phần mềm dưới dạng dịch vụ - Software as a Service (SaaS):** Cung cấp dịch vụ thông qua các phần mềm, người dùng tải phần mềm để sử dụng các dịch vụ. Các hệ thống đi kèm được lưu trữ trực tuyến và được quản lý toàn bộ bởi Azure.

Microsoft Azure có tốc độ hoạt đông rất nhanh, độ linh hoạt cao và giá cả phải chăng. Điều đó giúp cho Azure trở thành một dịch vụ đám mây công cộng tốt nhất trên thi trường hiện nay với một số dặc điểm chính:

* **Linh hoạt:** Di chuyển (tải về hoặc tải lên) tài nguyên máy tính qua lại máy chủ khi hoạt độn
* **Mở:** Hỗ trợ hầu hết mọi hệ điều hành, ngôn ngữ, công cụ
* **Đáng tin cậy:** Thỏa thuận mức dịch vụ (SLA) là 99,95% với độ tin cây cao cùng khả năng quản lý dịch vụ và bảo mật thông tin tốt.
* **Toàn cầu:** Dữ liệu được đặt trong các trung tâm dữ liệu mà người dùng có thể truy cập từ bất cứ nơi nào chỉ bằng Internet.
* **Tiết kiệm:** Chỉ trả tiền cho những gì bản sử dụng, giảm đi hoặc không còn các chi phí triển khai ban đầu.

Với nền tảng điện toán đám mây việc xây dựng các tài nguyên trên đó trở nên dễ dàng hơn:

* Ít chi phí hoặc không có chi phí ban đầu
* Không cẩn phải quản lý các khía cạnh như: Máy chủ vật lý, mạng, hệ thống làm lạnh, hệ thống phòng vệ, … vì đã có Azure hỗ trợ
* Chi trả khi sử dụng, hoặc tính phí khi chuyển sang các dịch vụ khác**.**

Những nền tảng điện toán đám mây mà Azure cung cấp:

* **Public Cloud:** Cung cấp các dịch vụ sử dụng thông qua web browser, chi phí thấp hơn, ít bảo dưỡng, các dịch vụ được cung cấp và quản lý bởi Microsoft Azure
* **Private Cloud:** Các cơ sở hạ tầng thuộc quyền quản lý của công ty, chính phủ hoặc các tổ chức tài chính, khi họ muốn có tính sở hữu cao, và cá nhân đối với các tài nguyên mà họ xây dựng trên đó, vì vậy tốn nhiều chi phí hơn.
* **Hybrid Cloud:** Kết nối Azure đến cơ sở hạ tầng trên máy của người dùng. Thừa hưởng cả hai điểm lợi từ việc xây dựng on premise và cloud, có tính linh hoạt cao. Nhưng cần thêm chi phí và bảo trì cho các tài nguyên được xây dựng

1. **Xây dựng tài nguyên trên Microsoft Azure**
2. **Resources và Resource group**

Với sự hổ trợ của Azure việc xây dựng các tài nguyên thông qua xây dựng máy ảo và quản lý nó trở nên đơn giản về dễ dàng hơn nhiều với người dùng hiện nay. Việc tạo ra các tài nguyên trong Azure tức ra tạo ra các Data center (Container) để lưu trữ các cơ sở hạ tầng được người dùng thiết đặt về sau. Các Data center liên kết với nhau với độ trễ thấp và tạo thành một vùng (Region)

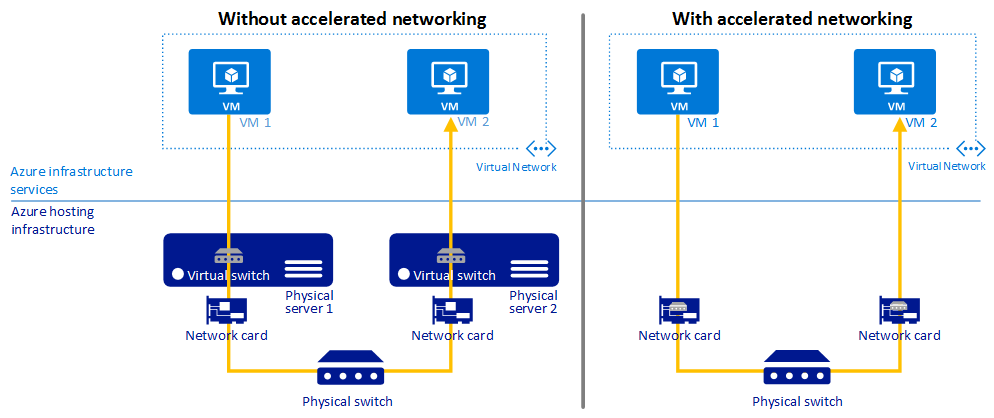
Các tài nguyên được xây dựng có thể là Website, máy ảo, cơ sở dữ liệu. Các tài nguyên được xây dựng có thể liên kết đến các tài nguyên khác tạo nên mạng lưới các tài nguyên. Để dễ dàng trong việc quản lý nhiều tài nguyên người dùng có thể gom các tài nguyên thành nhóm (Resource group).

Các máy ảo khi xây dựng được lưu trữ trong mạng ảo (Virtual Network). Mạng ảo bao gồm nhiều mạng con (Subnet). Máy ảo khi tạo nên sẽ được liên kết đến một giao diện mạng (Interface Network) có nhiệm vụ chính là cung cấp các địa chỉ Ip để các máy ảo có thể liên lạc được với nhau. Mỗi mạng ảo được chỉ định một địa chỉ IP cụ thể, các mạng con được cung cấp Ip thông qua Ip của mạng ảo.

Có 2 loại địa chỉ Ip được phân bổ cho các máy ảo:

* **Private Ip:** Ip được cung cấp thông qua Ip của các mạng con, thông qua Private Ip các máy ảo trong cùng 1 mạng có thể giao tiếp được với nhau
* **Public Ip:** phân bổ cho các máy ảo để giao tiếp với các máy ảo khác qua Internet

Trên thực tế việc giao tiếp giữa các mạng là không thể, phải tạo ra các kết nối mạng có thể nhìn thấy được (Virtual network Peering connection) từ đó các máy ảo có thể liên kết được với nhau



1. **Quản lý máy ảo**

Mạng máy ảo khi vào hoạt động sẽ được giám sát và quản lý thông qua một dịch vụ của Microsoft Azure là Network Security Groups (NSG). Các chức năng chính của NSG:

* Giúp người dùng có thể giám sát lưu lượng vào các mạng con hoặc các máy ảo
* Thiết lập luật bên trong và bên ngoài lưu thông
* Cho phép những lưu thông nào rõ ràng (đến từ cổng nào, giao thức kết nối nào, thực hiện hành động nào?) mới có thể đi qua và kết nối với các máy ảo
* Những luật được xây dựng phải có nguồn, đích

Các ứng dụng (ví dụ: Website, app, …) có thể được xây dựng trên một máy ảo nào đó. Khi người dùng muốn truy cập vào Website cũng đồng nghĩa truy cập đến máy ảo. Đối với một cá nhân thì không ảnh hưởng, nhưng khi nhiều cá nhân cùng truy cập vào thì việc xử lý tải sẽ trở nên khó khăn. Bộ mở rộng quy mô (Scale Sets) là một giải quyết cho vấn đề này. Các máy ảo khi đó sẽ được quản lý theo quy mô của bộ dịch vụ, bộ quy mô sẽ tự động thêm máy ảo để mở rộng quy mô phục vụ cho nhu cầu truy cập vào của người dùng.

Bộ mở rộng (Scale Sets) thực hiện các chức năng như:

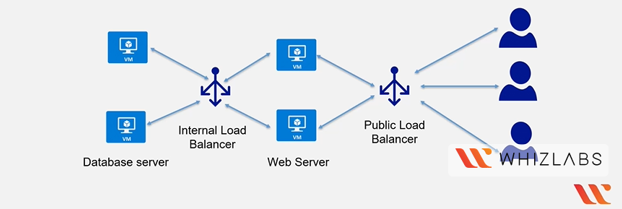
* Là dịch vụ cho phép tạo ra và thiết đặt những máy ảo giống hệt nhau
* Cho phép mở rộng quy mô theo yêu cầu
* Nâng cao tính sẵn có cho cấu trúc cơ sở hạ tầng được xây dựng
* Có thể đặt máy ảo sau bộ mở rộng hoặc bộ cân bằng tải

Một dịch vụ quản lý khác cũng được Azure cung cấp cho người dùng giúp quản lý lưu lượng, chuyển yêu cầu của người dùng phân bổ đều cho các máy ảo phụ trợ tránh trường hợp một máy ảo phụ trợ bị quá tải yêu cầu truy cập thông qua bộ cân bẳng tải (Azure Load Balancer).

Có 2 loại cân bằng tải người dùng có thể tạo ra để quản lý lưu lượng trên Azure bao gồm:

* **Internal Load Balancer:** cân bằng tải nội bộ điều phối lưu lượng giữa máy ảo với nhau
* **Public Load Balancer:** cân bằng tải giữa người dùng với máy ảo

Ngoài ra còn có những nhóm phụ trợ khác (đây cũng là những máy ảo với chức năng khác) giúp xác định máy ảo, thiết đặt bộ mở rộng, xác định những giao thức, số cổng và khoảng thời gian cho mỗi lần tải.



1. **Azure Database**

Azure SQL Database là nên tảng dịch vụ được quản lý bởi Azure được triển khai linh hoạt với nhiều mức giá, giúp truy cập đến các chức năng của Microsoft SQL Server. Khi triển khai SQL trên Virtual Machine sẽ có đầy đủ các chức năng, dễ dàng vận chuyển data từ máy của người dùng lên nền tảng đám mây. Sử dụng các công cụ truyền thống như lưu trên cơ sở dữ liệu dưới máy của người dừng và có thể kết nối bằng SQL Server Management

Azure SQL Data Warehouse: dùng để lưu trữ dữ liệu lớn. Dữ liệu trong data Warehouse được lưu trữ trong các bảng quan hệ và dùng cách lưu theo tên cột. Sử dụng dịch vụ lưu trữ ở Warehouse giúp giảm bớt các chi phí, cải thiện hiệu suất truy vấn. Tại đây lưu trữ các dữ liệu sạch sử dụng cho mục đích phân tích và báo cáo.

Một dịch vụ khác được Azure cung cấp là dịch vụ không máy chủ (Serverless). Đối với giải pháp lưu trữ này người dùng không cần quan tâm đến máy chủ, cũng như vấn đề bảo trì (một ví dụ điển hình là DataBricks), dữ liệu người dùng được đưa lên từ máy tính cá nhân nhưng thật sự không được lưu trữ trên server.

### Databrick sử dụng nền tảng Azure

Với sự phát triển xã hội hiện nay, đặc biệt là trong xu hướng chuyển đổi số của cuộc cách mạng công nghiệp 4.0, mọi hoạt động dần dần được được internet hóa. Lượng thông tin trao đổi trên Internet ngày càng đa dạng và không ngừng tăng. Chính vì vậy, dữ liệu ngày càng lớn và đa dạng. Big Data lớn về số lượng, dữ liệu phức tạp có thể có cấu trúc hoặc không có cấu trúc (noSQL). Những yếu tố này làm cho Big Data khó bắt giữ lại, khai phá và quản lý nếu dùng các phương thức truyền thống.

Dù mới phát triển gần đây nhưng Databricks là một nền tảng khoa học dữ liệu, kỹ thuật dữ liệu và phân tích dữ liệu cộng tác dựa trên đám mây, kết hợp tốt data warehouses,  data lakes và lakehouse với nhau.

Databrick là nơi hỗ trợ xử lý big data được xây dựng từ Apache Spark hỗ trợ xử lý dữ liệu với tốc độ nhanh và toàn vẹn. Databrick hỗ trợ trên ba nền tảng lớn hiện nay là Microsoft Azure, amazon web services và google cloud. Azure Databricks là dịch vụ triển khai Databricks trên nền tảng Azure, cung cấp khả năng autoscale, tương tác với các thành viên khác dễ dàng thông qua workspace. Azure Databricks hỗ trợ nhiều ngôn ngữ như Java, Python, Scala, R,... Tuy nhiên, vì được xây dựng dựa trên lõi là Apache Spark nên Databricks hỗ trợ mạnh mẽ cho ngôn ngữ Scala với tốc độ xử lý nhanh và kho thư viện lớn.

Azure Databricks hiện đã trở thành một nền tảng phân tích dữ liệu được tối ưu hóa cho nền tảng dịch vụ đám mây Microsoft Azure. Azure Databricks cung cấp ba môi trường:

* Databricks SQL
* Databricks data science and engineering
* Databricks machine learning

Để sử dụng databrick, trước tiên cần tạo một cluster, sau đó đẩy file data vào DBFS để tạo môi trường làm việc. Sau khi chuẩn bị xong tạo một notebook mới để tiến hành xử lý dữ liệu.

* + 1. **Visual Studio Code và ngôn ngữ Python**

- Python: ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng được tạo ra năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức sáng sủa, cấu trúc dễ rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc python cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Ban đầu Python được phát triển để chạy trên nên Unix. Dần theo thời gian mở rộng sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Linux và các hệ điều hành khác thuộc họ Unix. Đặc biệt được sử dụng rỗng rãi trong lĩnh vực dữ liệu, phát triển trí tuệ nhân tạo (AI), trích xuất thông tin từ dữ liệu và máy học dạy máy tính tự động học hỏi từ dữ liệu và đưa ra các dự đoán chính xác. Có thể sử dụng Python cho các tác vụ khoa học dữ liệu sau:

+ Sửa và loại bỏ dữ liệu không chính xác, hay còn được gọi là làm sach dữ liệu

+ Trích xuất và chọn lọc các đặc điểm đa dạng của dữ liệu

+ Gắn nhãn dữ liệu, gán tên có ý nghĩa cho dữ liệu

+ Tìm các số liệu thống kê khác nhau từ dữ liệu

+ Trực quan hóa dữ liệu bằng cách sử dụng các biểu đồ và đồ thị, chẵn hạn như biểu đồ đường, biểu đồ cột, biểu đồ tần suất và biểu đồ tròn.

Sử dụng các thư việnc ML (Machine Learning) của Python để đào tạo các mô hình ML và xây dựng các công cụ phân loại giúp phân loại dữ liệu một cách chính xác. Những chuyên gia từ nhiều lĩnh vực sử dụng những công cụ phân loại dựa trên Python để thực hiện các tác vụ phân loại, chẵng hạn như phân loại hình ảnh, văn bản cũng như lưu lượng truy cập mạng, nhan dạng giọng nói và nhận diện khuân mặt. Các nhà khoa học dữ liệu cũng sử dụng Python cho deep learning, một kỹ thuật ML nâng cao.

* **Visual Studio Code:** là ứng dụng cho phép biên tập, soạn thảo các đoạn code để hỗ trợ trong quá trình thực hiện xây dựng và thiết kế chương trình một cách hiệu quả và nhanh chóng. Trình soạn thảo này hỗ trơ trên hầu hết các nền tảng như MacOS, Windows, Linux. Hơn thế nữa VS Code (Visual Studio Code) còn có khả năng tương thích với những máy tính có cấu hình tầm trung vẫn có thể sử dụng được dễ dàng.

Hỗ trợ đa dạng các chức năng Debug, đi kèm với Git, có Syntax Highlighting. Đặc biệt là tự hoàn thành mã thông minh, Snippets, và khả năng cải tiến mã nguồn. Nhờ tính năng tùy chỉnh, Visual Studio Code cũng cho phép người dùng thay đổi Theme, phím tắt,…

Visual Studio Code được chia thành 2 phiên bản Visual Studio Enterprise và Visual Studio Professional, các phiên bản cao cấp có tính phí được sử dụng bởi các công ty chuyên về lập trình. Bên cạnh đó Microsoft cũng cho ra mắt phiên bản Community của phần mềm, cung cấp cho người dùng những tíng năng cơ bản nhất, phù hợp với đối tượng lập trình không chuyên, mới tiếp cận với công nghệ

Sở hữu nhiều tính năng:

+ Đa nền tảng

+ Đa ngôn ngữ lập trình

+ Hỗ trợ website

+ Kho tiện ích mở rộng phong phú

+ Lưu trữ phân cấp

+ Màn hình đa nhiệm

+ Hỗ trợ Git

* 1. **Quy trình thực hiện**

### Thu thập dữ liệu – Crawling/Scraping

Thu thập dữ liệu (Web Scraping/Web Scraping) là thuật ngữ chỉ việc tìm kiếm và thu thập các thông tin trên một website. Quét hay cào dữ liệu được sử dụng khá phổ biến trong hoạt động SEO web thuộc lĩnh vực marketing online hay trong lĩnh vực công nghệ thông tin.

Kỹ thuật Crawl/Scrap dữ liệu có ý nghĩa quan trọng đối với website. Nó thực hiện việc thu thập dữ liệu từ một website nào đó rồi tổng hợp, phân tích và bóc tách thông tin theo yêu cầu mà người dùng đặt ra hay theo các dữ liệu mà công cụ tìm kiếm yêu cầu. Sau khi thực hiện xong, dữ liệu sẽ được lưu vào database.

Web Crawling và Web Scraping đều là hai khái niệm liên quan đến việc thu thập dữ liệu từ các website nhưng cũng có vài điểm khác biệt.

Crawl lấy thông tin từ website qua đường link web được nhập vào, thông qua các lệnh được lập trình bởi một ngôn ngữ lập trình nào đó (Python, Java, C#) để lấy dữ liệu. Việc crawl dữ liệu cũng tương tự như hoạt động của robots của các công cụ tìm kiếm (Google, Bing, Baidu,…) thường làm để phục vụ cho việc tìm kiếm.

Scrap cũng thực hiện việc tìm kiếm và thu thập thông tin như vây nhưng khác với Crawl, Scrap không thu thập toàn bộ thông tin của một trang web mà chỉ thu thập những thông tin cần thiết, phù hợp với mục tiêu của người sử dụng. Trong Web Scraping cũng sử dụng WebCrawl (bot) để thu thập dữ liệu, kết hợp với Data Extraction (trích xuất dữ liệu) để tập trung vào những nội dung thật sự cần thiết.

Những ưu điểm của việc thực hiện crawl data:

* Giảm thiểu thời gian và công sức cho quá trình thu thập khối lượng lớn thông tin và dữu liệu
* Có thể dễ dàng tổng hợp thông tin từ nhiều trang web khác nhau. Sau đó thực hiện đối chiếu, so sánh và làm các thuật toán để phân tích thị trường, tiết kiệm thời gian, công sức và hiệu quả hơn.
* Giúp lấy thông tin nhanh chóng, chính xác và hữu hiệu

Dù khá hiệu quả trong việc lấy dữu liệu nhưng Crawl vẫn có một số các hạn chế:

* Khi website thay đổi cấu trúc HTML thì phải thực hiện update, cập nhật để chương trình crawl phù hơp với những thay đổi đó.
* Trong thực tế việc crawl có thể coi là vi phạm các quyền về bảo mật thông tin, nên các trang web uy tín và có chất lượng thường sẽ không thực hiện crawl được

### Xử lý dữ liệu – ETL (Extract - Transform – Load)

ETL là viết tắt của trích xuất, biến đổi và tải, là một quá trình tích hợp dữ liệu kết hợp dữ liệu từ nhiều nguồn dữ liệu vào một kho dữ liệu nhất quán, duy nhất được tải vào  kho dữ liệu  hoặc hệ thống đích khác. ETL làm sạch và sắp xếp dữ liệu theo cách giải quyết các nhu cầu thông minh kinh doanh cụ thể, như báo cáo hàng tháng, nhưng nó cũng có thể giải quyết các phân tích nâng cao hơn, có thể cải thiện quy trình back-end hoặc trải nghiệm người dùng cuối. ETL thường được một tổ chức sử dụng để:

* Trích xuất dữ liệu từ các hệ thống kế thừa
* Xóa dữ liệu để cải thiện chất lượng dữ liệu và thiết lập tính nhất quán
* Tải dữ liệu vào cơ sở dữ liệu đích

Dữ liệu thô sau khi crawl cần được ETL để trở thành một tập dữ liệu sạch, toàn vẹn và có ý nghĩa. Dữ liệu thô thường sẽ chứa các giá trị Null, duplicate, negative,... Tuỳ vào mỗi dataset mà cần có cách tiến hành xử lý dữ liệu phù hợp. Công cụ để tiến hành ETL cũng khá đa dạng và phụ thuộc vào yêu cầu cũng như mục đích sử dụng của người dùng cuối. Tuy nhiên, những công cụ phổ biến được dùng để ETL hiện nay là ngôn ngữ: R, Python, Scala, Java, SQL,...

Trong thời đại 4.0 lượng dữ liệu đồ sộ khiến cho việc xử lý làm sạch phải diễn ra trên môi trường cloud tích hợp đa ngôn ngữ với bộ nhớ lớn đảm bảo quá trình xử lý dữ liệu diễn ra nhanh, tiện, chính xác. Hiện nay, có rất nhiều nền tảng cloud phục vụ cho việc ETL big data, nhưng lớn nhất phải kể đến AWS, Azure và Google Cloud.

Dữ liệu thô sau khi được trích xuất sẽ trải qua quá trình xử lý dữ liệu. Tại đây, dữ liệu được chuyển đổi và hợp nhất cho trường hợp sử dụng phân tích dự kiến ​​của nó. Giai đoạn này có thể bao gồm các nhiệm vụ sau:

* Lọc, làm sạch, bỏ trùng lặp, xác thực dữ liệu.
* Thực hiện các phép tính, dịch hoặc tóm tắt dữ liệu dựa trên dữ liệu thô như thay đổi tiêu đề hàng và cột để có tính nhất quán, chuyển đổi kiểu dữ liệu, chỉnh sửa chuỗi văn bản, v.v.
* Thực hiện kiểm tra để đảm bảo chất lượng và toàn vẹn của dữ liệu.
* Xóa, mã hóa hoặc bảo vệ dữ liệu do các cơ quan quản lý ngành hoặc chính phủ quản lý.
* Định dạng dữ liệu thành các bảng hoặc bảng nối để phù hợp với lược đồ của data warehouse.

Dữ liệu tại data warehouse thuộc lớp gold đã có ý nghĩa và có thể phục vụ cho mục đích cao hơn của data science hoặc data analyst. Do đó, ETL vô cùng quan trọng, nếu quá trình ETL không tốt,sẽ ảnh hưởng lớn đến việc nghiên cứu, phân tích và đánh giá dữ liệu.

### Phân tích và trực quan hoá dữ liệu bằng công cụ Power BI

Trong thời đại nhiều dữ liệu của chúng ta, việc hiểu cách phân tích và rút ra ý nghĩa thực sự từ những hiểu biết kỹ thuật số của doanh nghiệp chúng ta là một trong những động lực chính dẫn đến thành công.

Phân tích dữ liệu là quá trình thu thập, mô hình hóa và phân tích dữ liệu để rút ra những hiểu biết sâu sắc hỗ trợ việc ra quyết định. Có một số phương pháp và kỹ thuật để thực hiện phân tích tùy thuộc vào ngành và mục đích của cuộc điều tra. Tất cả các phương pháp khác nhau này chủ yếu dựa trên hai lĩnh vực cốt lõi: nghiên cứu định lượng và định tính.

Với sự trợ giúp của các kỹ thuật khác nhau như phân tích thống kê, hồi quy, mạng nơ-ron, phân tích văn bản, v.v., bạn có thể bắt đầu phân tích và thao tác dữ liệu của mình để rút ra các kết luận có liên quan. Ở giai đoạn này, bạn tìm thấy các xu hướng, mối tương quan, các biến thể và các mẫu có thể giúp bạn trả lời các câu hỏi bạn nghĩ đến đầu tiên trong giai đoạn xác định. Các công nghệ khác nhau trên thị trường hỗ trợ các nhà nghiên cứu và người dùng doanh nghiệp trung bình quản lý dữ liệu của họ. Một số trong số đó bao gồm phần mềm trực quan và trí tuệ kinh doanh, phân tích dự đoán, khai thác dữ liệu, trong số những phần mềm khác. Phân tích dữ liệu có 4 mức độ chính từ thấp đến cao gồm:

* Phân tích mô tả: cho chúng ta biết điều gì đã xảy ra. Loại phân tích này giúp mô tả hoặc tóm tắt dữ liệu định lượng bằng cách trình bày các số liệu thống kê.
* Phân tích chuẩn đoán: cho biết vì sao những điều này lại xảy ra, loại phân tích này giúp hiểu sâu về nguyên nhân của những gì đã xảy ra trong quá khứ.
* Phân tích dự đoán: sử dụng dữ liệu để hình thành các dự báo về tương lai. Bằng những nguyên nhân đã được tìm ra ở phân tích chuẩn đoán có thể dự đoán điều gì sẽ xảy ra trong tương lai.
* Phân tích đánh giá: là bước phân tích khó nhất đòi hỏi kinh nghiệm và tầm nhìn khách quan, bởi vì sau 3 bước phân tích trên, người phân tích sẽ đưa ra những nhận xét và lời khuyên đề nghị cho doanh nghiệp,...

Ở lớp phân tích mô tả có thể thực hiện bằng Power BI. Tuy nhiên từ lới chuẩn đoán trở đi phải kết hợp hoặc sử dụng thuần thuật toán để đánh giá, nghiên cứu dữ liệu. Đối với lớp chuẩn đoán trở đi thì cần Machine Learning để phân tích.

# CHƯƠNG 2. TRIỂN KHAI

**2.1 Đặt vấn đề**

Những năm gần đây với sự phát triển không ngừng của trò chơi điện tử đã khiến cho lĩnh vực đầu tư này trở nên nóng bỏng hơn bao giờ hết. Trò chơi điện tử giờ đây không còn được nhắc đến với những lời lẻ tiêu cực, chỉ những thứ vô bổ, tốn thời gian. Trò chơi điện tử giờ đây đã trở thành một công cụ nơi mà nhà đầu tư dễ dàng thu lợi nhuận và người chơi có thể đem lại thu nhập cho mình từ việc chơi tựa game đó. Các nền tảng trực tuyến nổi tiếng hiện nay như Nimo TV, Facebook, Youtube cũng chú trọng nhiều đến mảng trò chơi, thông qua các nền tảng này người chơi có thể livestream để chia sẽ các khoảnh khắc của mình và kiếm tiền từ lược theo dõi và lược xem. Khi chỉ dừng lại ở đó trò chơi điện tử nói chung và thể thao điện tử nói riêng ngày càng được công nhận, điển hình như Seagame 31 lần đầu tiên trò chơi điện tử được đưa vào trở thành một bộ môn tranh tài như bao bộ môn khác. Góp mặt tại Seagame với 8 bộ môn và 10 nội dung thi đấu như: Liên minh huyển thoại tốc chiến, Liên minh huyền thoại, PUBG Mobile, Fifa Online 4, … đã cho thấy một bộ mặt khác trong lĩnh vực esport này.

Đầu tư trong lĩnh vữ trò chơi điện tử đã và đang cho thấy một bức tranh lợi nhuận to lớn, nhưng việc đầu tư và phát triển như thế nào và ở đâu thì lại là một bài toán nan giải. Để tạo ra một tựa game và phát triển nó để nhiều người biết đến và thu lại được lợi nhuận trên thực tế là điều không đơn giản. Có rất nhiều các dự án phát triển về game đã được triển khai những kết quả lại không được như mong đợi, khiến các nhà đầu tư thua lỗ lớn như: “Elemon”, ”Heroes Empires”, “Space Cryto”. Để có thể thành công trong việc xây dựng và phát hành một tựa game bên cạnh các công nghệ để xây dựng, lập trình nên nó cũng cần phải quan tâm đến việc nghiên cứu xu hướng phát triển của những tựa game qua các năm, tìm hiểu nền tảng hay thể loại nào đã và đang được yêu thích từ đó dễ dàng tiếp cận với người dùng và đem lại lợi nhuận cao.

**2.2 Giải quyết vấn đề:**

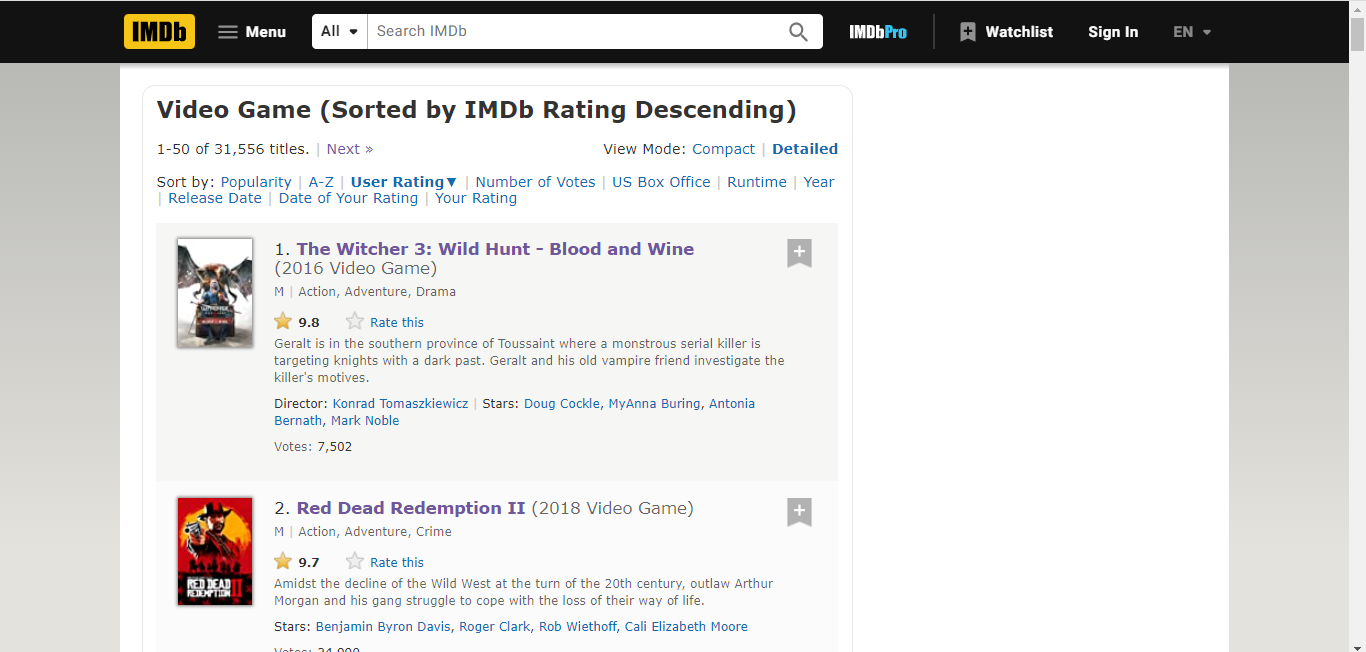
2.2.1 Scraping data

Để giải quyết vấn đề như trên, thực hiện scraping bằng ngôn ngữ Python trên trình soạn thảo Visual Studio Code nhằm thu thập những dữ liệu cần thiết cho quá trình phân tích xu hướng phát triển Game

Các bước thực hiện để scraping data:

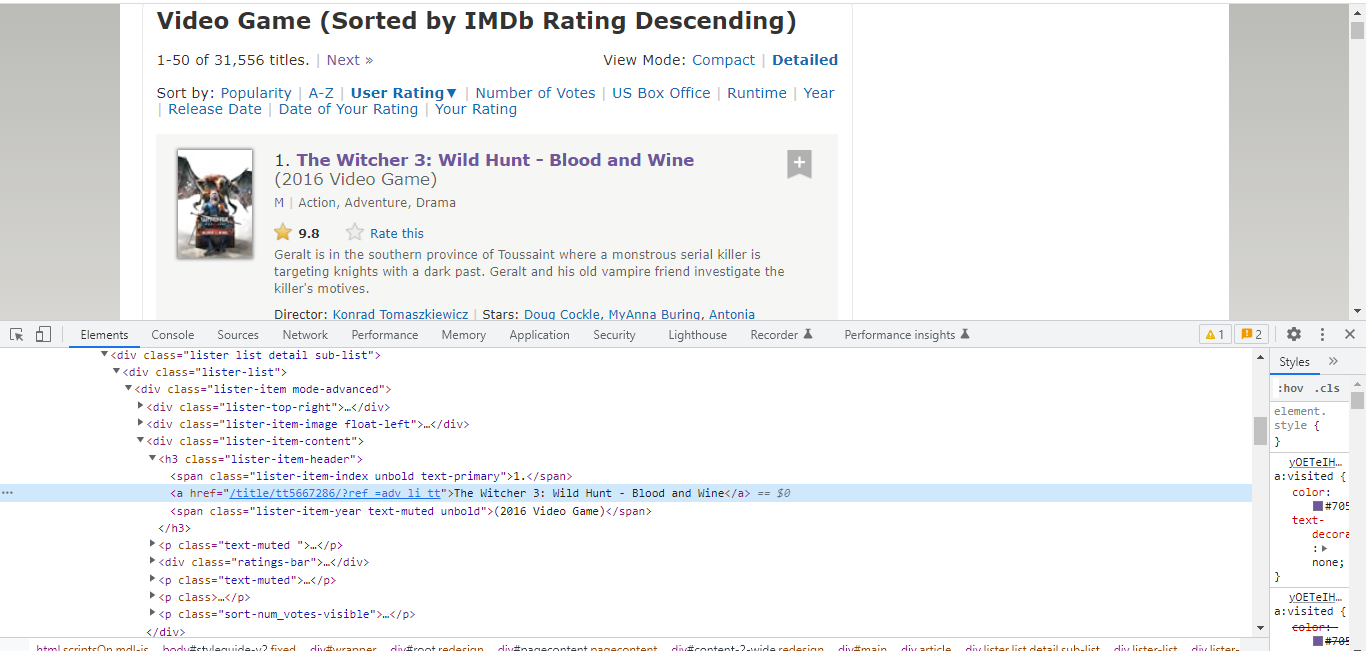
* + - 1. Phân tích
* Khảo sát website, lựa chọn các trang web về game có thể request để lấy thông tin thông tin, ưu tiên những website có cấu trúc đơn giản, thân thiện với người dùng
* Chọn website về Video game trên IMDB

https://www.imdb.com/search/title/?title\_type=video\_game&sort=user\_rating,desc&start=01&ref\_=adv\_nxt



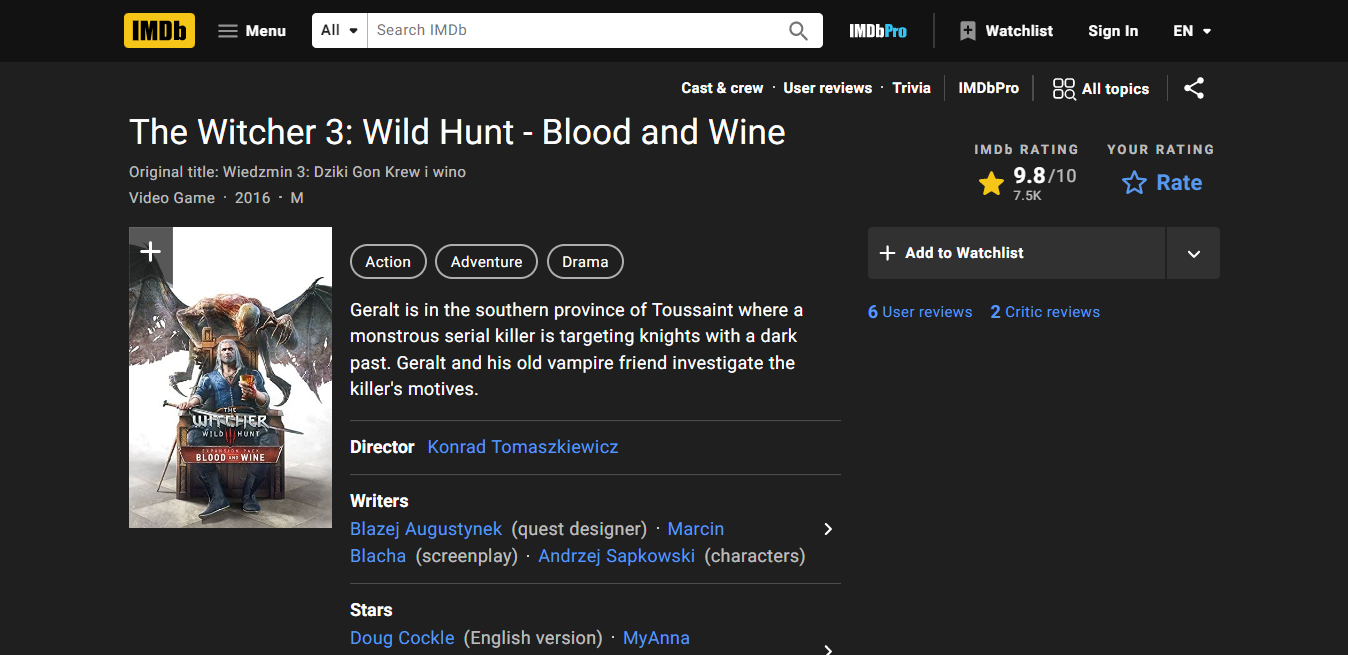
*Website để lấy dữ liệu*

Phân tích cấu trúc HTML của trang web bằng phím f12 hoặc chuột phải 🡪 Inspect



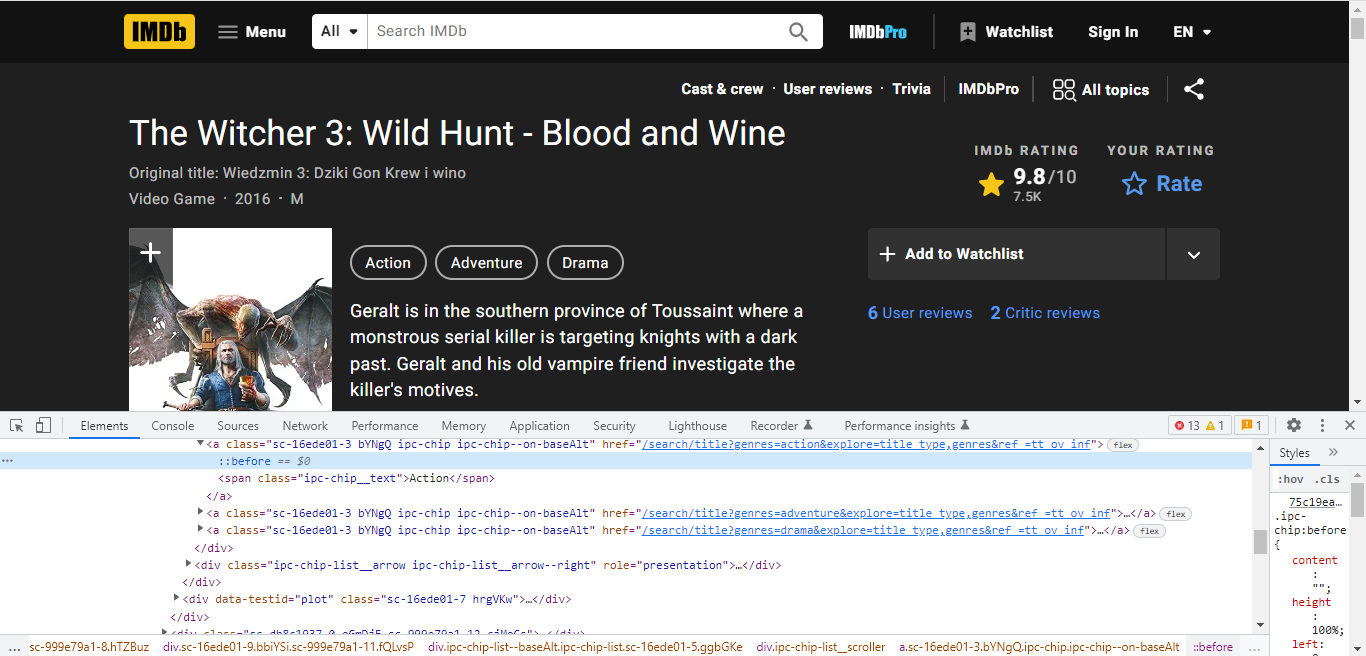
*Phân tích cấu trúc trang web*

* Chọn các thông tin cần để scraping (tên, thể loại, thời gian, quốc gia, rate, vote, …)



*Lựa chọn thông tin cần lấy trên website*

* Tìm những thẻ chứa các đoạn text của thông tin và xác định nó thông qua các thuộc tính như: class, data-testid, …



*Xác định các thẻ chứa nội dung cần lấy*

* + - 1. Viết code

- Import các thư viện hỗ trợ cho quá trình scraping

+ BeautifulSoup: phân tích nội dung của website

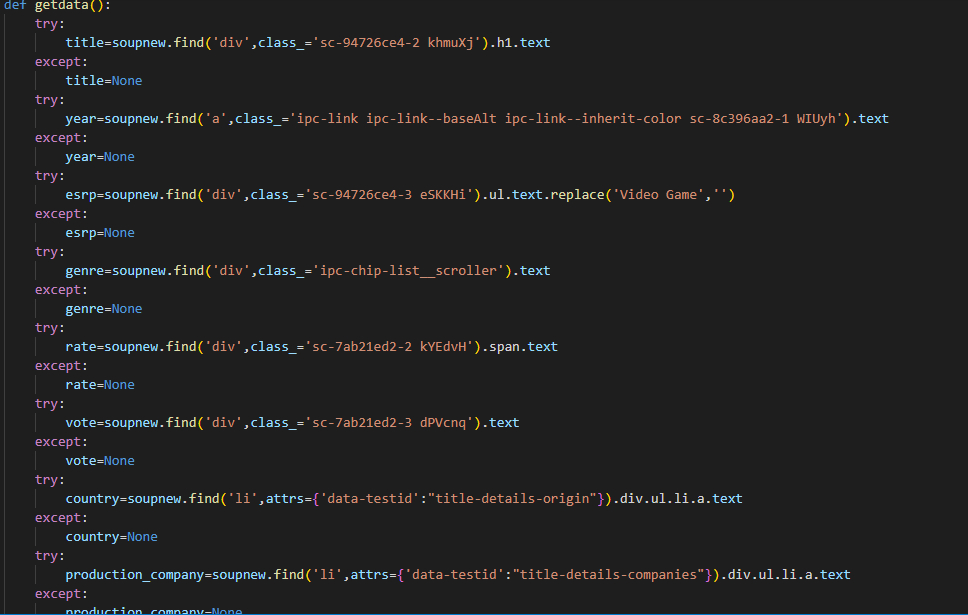
+ urllib.request: truy cập đến website

+ csv: tạo file lưu dữ liệu



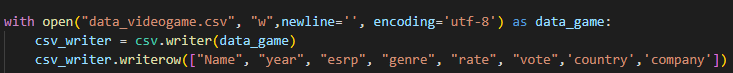
*Import thư viện*

- Tạo hàm để lấy dữ liệu thông qua các thẻ chứa nội dung và các thẻ được xác định thông qua các thuộc tính



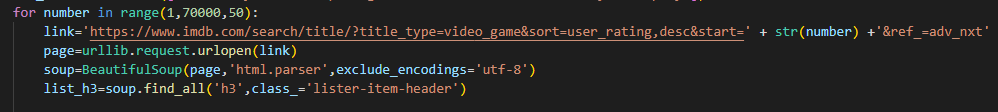
*Tạo hàm get data*

-Tạo file csv với tên “data\_videogame”, “w” mở file để ghi nội dung nhưng trước đó sẽ xóa hết nội dung hiện có của file nếu có, chuẩn encoding = “utf-8” chuyển đổi dữ liệu thành định dạng cần thiết cho nhu cầu xử lý thông tin. Viết từng dòng vào file csv với tên của các cột tương ứng: Tên, năm, chuẩn game, thể loại, đánh giá, lượt xem, quốc gia, công ty phát hành.



*Mở file csv*

-Qua vòng lặp for lấy số tương ứng với đường link trên trang web để request, dùng thư viện BeautifulSoup để phân tích cấu trúc html (html.parser) và tìm tất cả các thẻ h3 chứa link để trang cần lấy thông tin

**

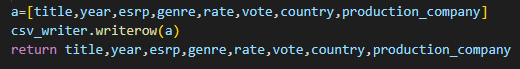
*Tìm thẻ h3 chứa link vào trang có thông tin cần lấy*

- Dùng vòng lặp for chạy qua lần lượt các thẻ h3 tìm được và lấy phần text trong phần href, gán đoạn text vào trong cấu trúc đường link và request đến đó. Gọi lại hàm đã tạo từ trước để lấy những dữ liệu như đã đề ra

****

*Tuy cập vào đường dẫn, lấy dữ liệu*

- Những dữ liệu lấy về sẽ được gán vào list và viết từng dòng tương ứng vào file csv



*Ghi dữ liệu vào file csv*

**2.2.Xử lý dữ liệu(ETL)**

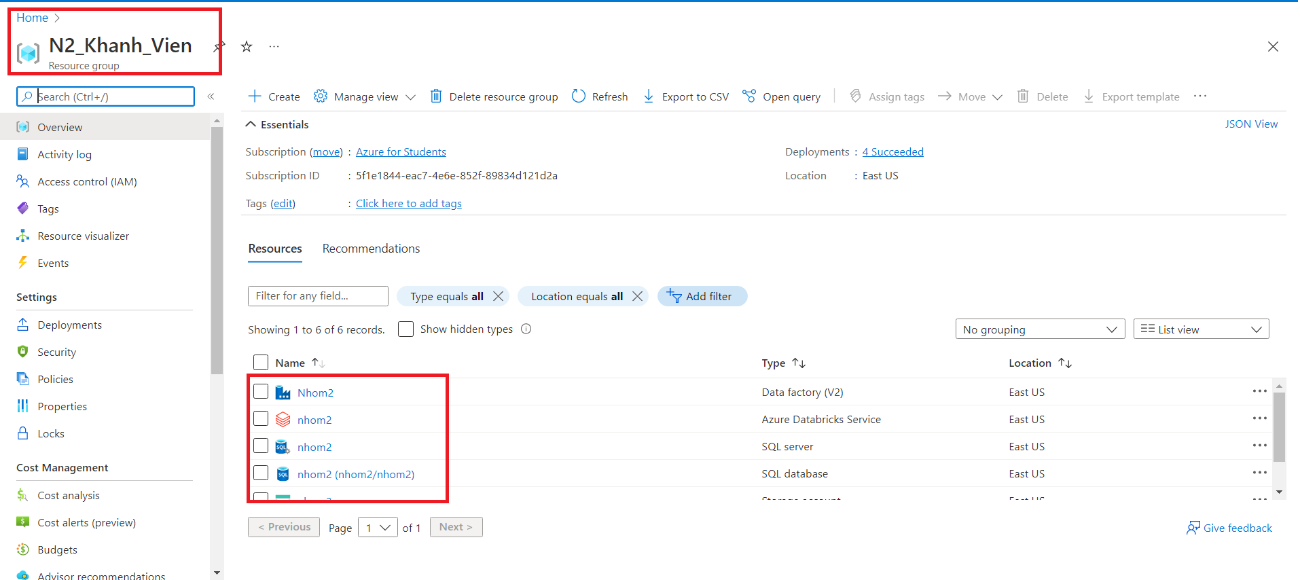
Dữ liệu thô sau khi cào về vẫn còn bẩn(Null, Duplicate, Negative Values,…) dưới định dạng file .csv sẽ được đưa lên Azure Microsoft để tiến hành xử lý. Quy trình xử lý dữ liệu trên Azure sẽ qua các bước:

* Bước 1: Tạo Resource Group để nhóm các tài nguyên, tools.
* Bước 2: Tạo Databricks, Data Factory(ADF), Database SQL, Storage Account trên nhóm Resource Group vừa tạo.
* Bước 4: Tạo Container chứa dữ liệu 3 mức ở Storage Account và upload dữ liệu vào Container level 1.
* Bước 5: Chuyển dữ liệu từ level 1 vào level 2 bằng ADF.
* Bước 6: Kết nối Container level2 với Databricks và xử lý dữ liệu.
* Bước 7: Đổ data sạch về Container level 3 thông qua ADF.
* Bước 8: Chuyển data ở Container level 3 vào SQL Database.

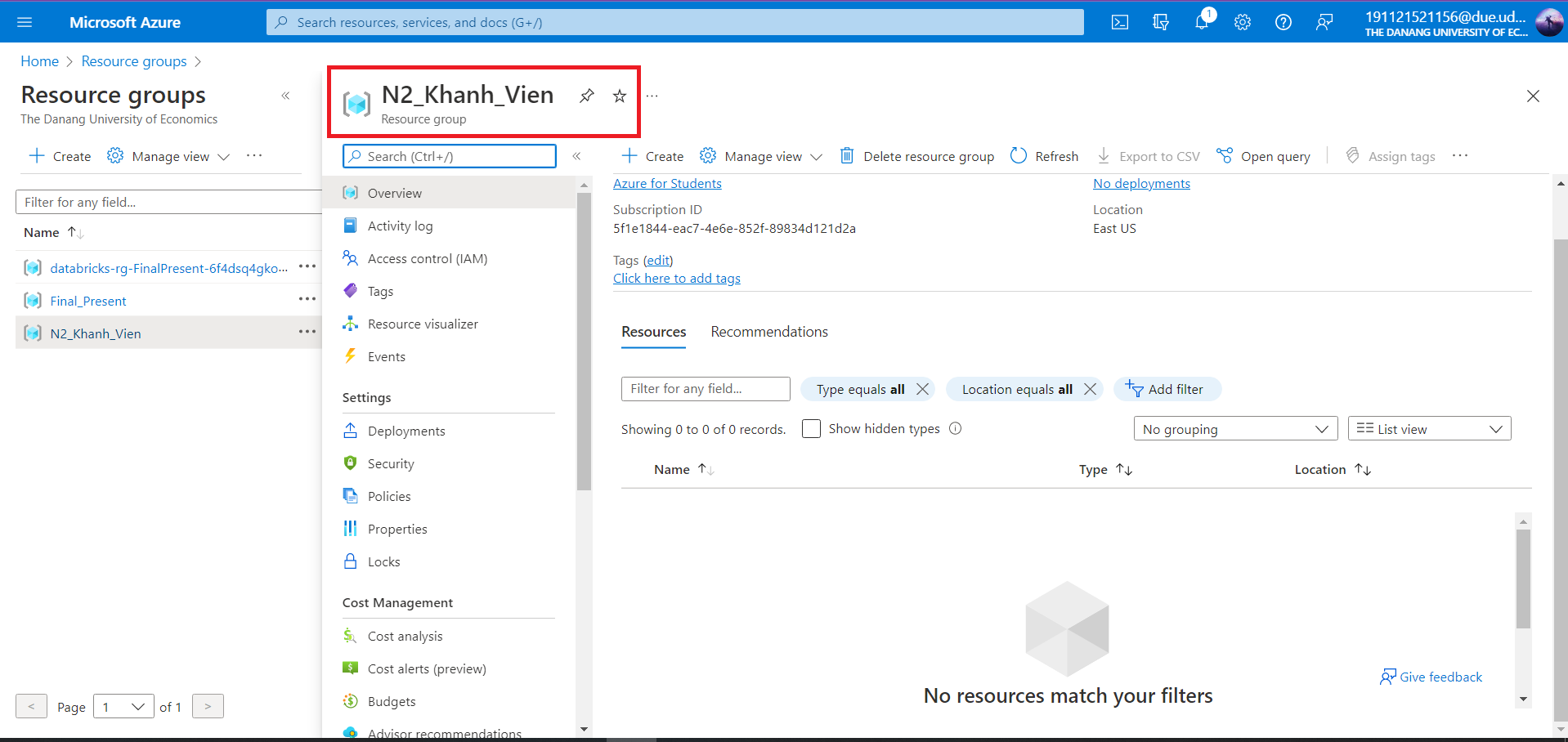
Dưới đây là từng bước cụ thể để tiến hành ETL dữ liệu:

1. Tạo Resource Group

Để có thể sử dụng đầy đủ các chức năng của Azure Microsoft, nhóm tiến hành đăng kí tài khoản dùng thử miễn phí 30 ngày cho sinh viên.



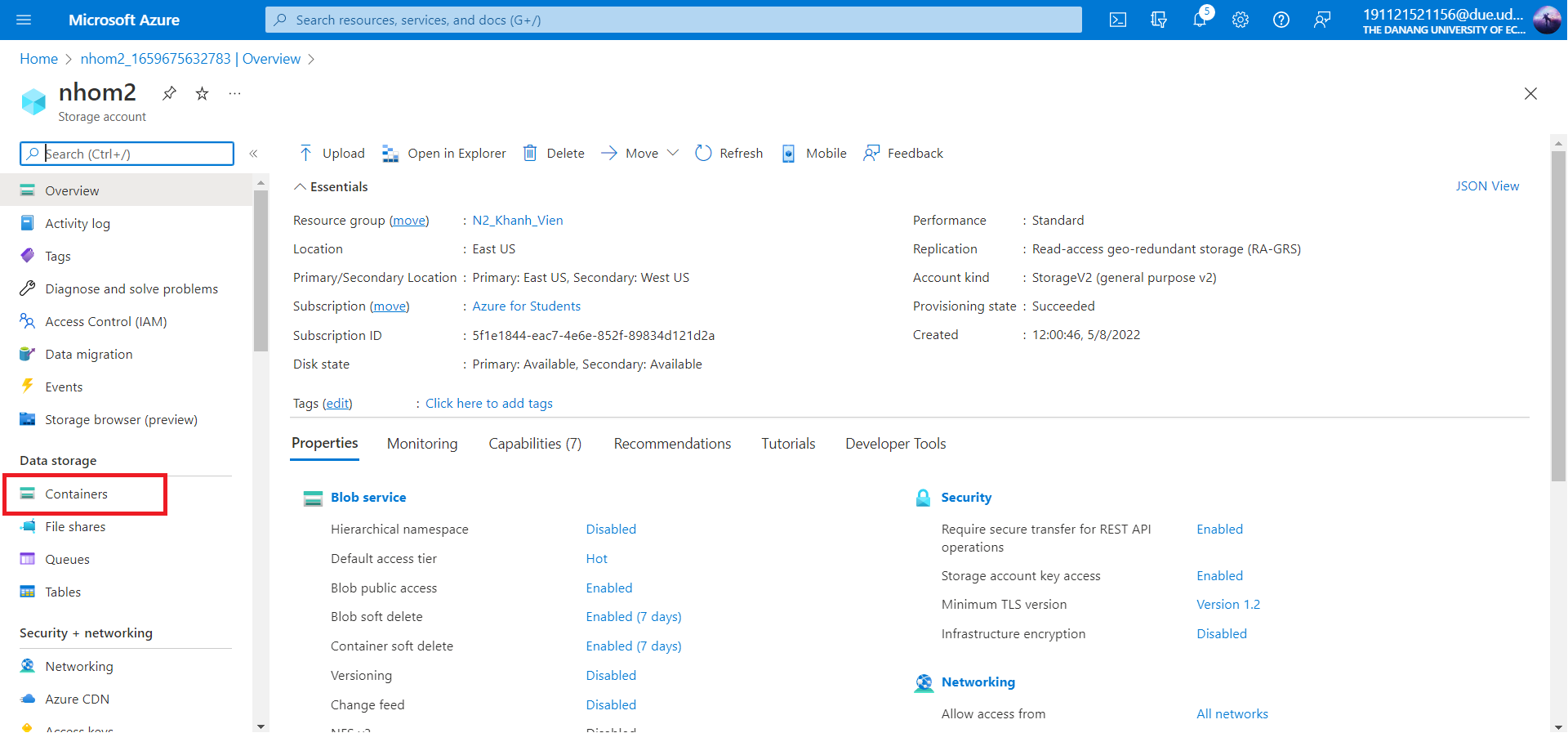
*Nhóm chức năng trong Group“N2\_Khanh\_Vien”*



*Tạo Resource Group đặt tên là “N2\_Khanh\_Vien”*

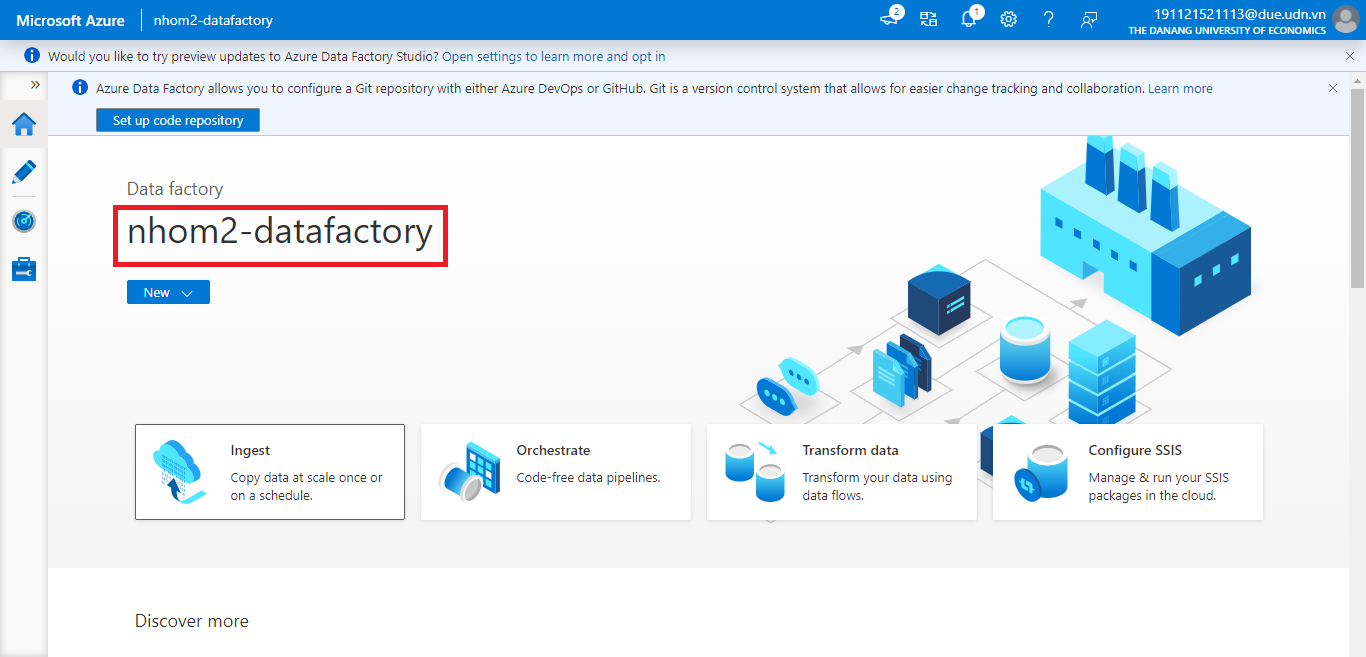
1. Tạo Databricks, Secret Scope, Key vault, Data Factory, SQL Database, Storage Account

Trước tiên, tạo một Storage Account để chứa dữ liệu. Tất cả các chức năng đều nhóm theo Resource Group “N2\_Khanh\_Vien”.

**

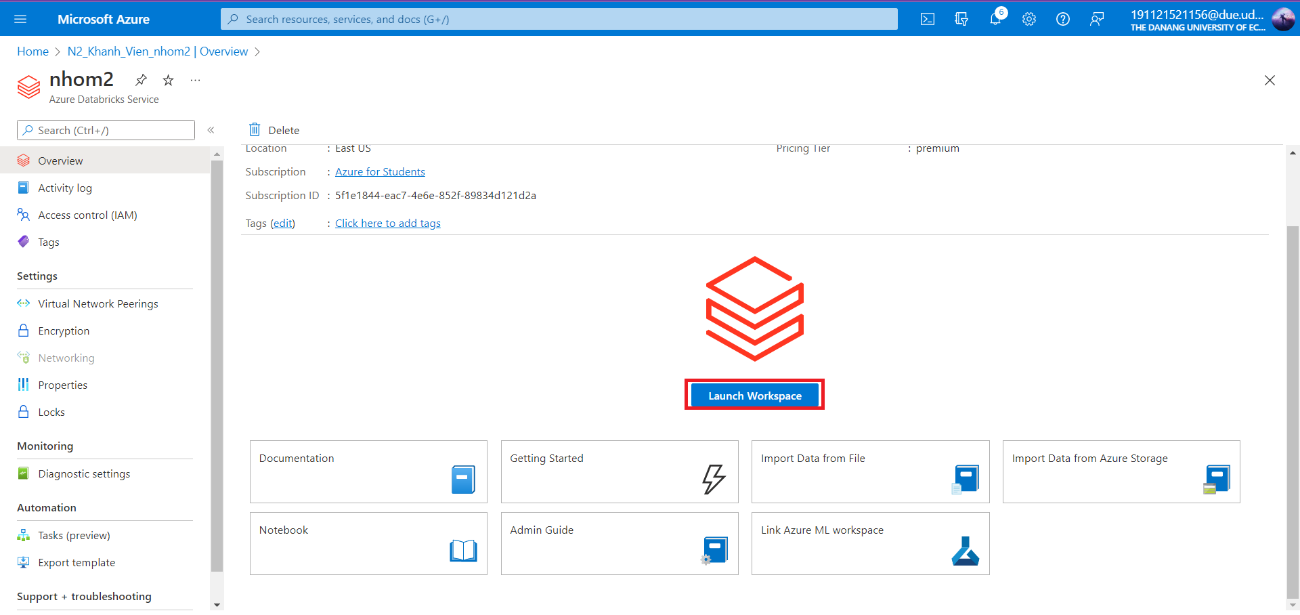
*Container* *chứa dữ liệu nằm trong Storage Account*

Sau đó tạo một Data Factory để tiến hành quản lý – liên kết – lấy dữ liệu theo dạng đường ống(Pieline).



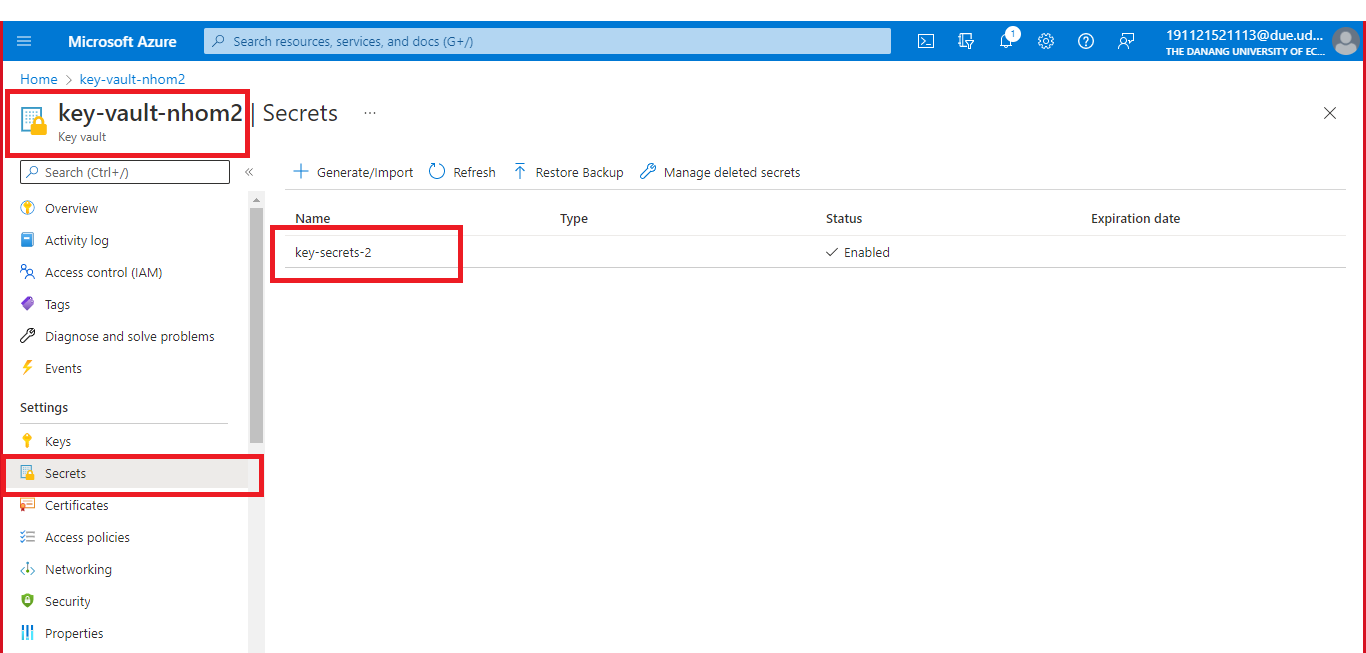
*Tạo Data Factory “nhom 2-datafactory”*

Tạo một Databricks để tiến hành ETL dữ liệu thô vừa Scrawling về.



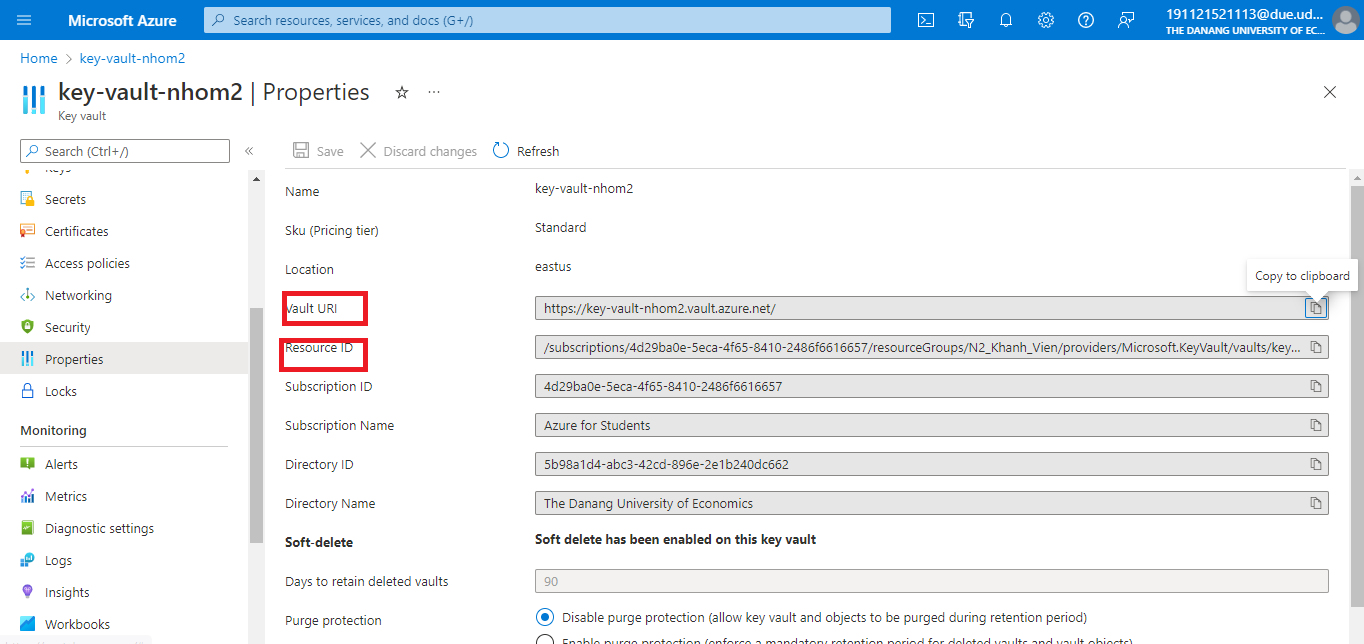
*Databricks*

Tạo Key Secret trong Key Vault nhằm tạo một lớp bảo mật, giúp che dấu đường dẫn đến Storage trong trường hợp kết nối từ databricks để đưa dữ liệu sau khi xử lý về lại storage

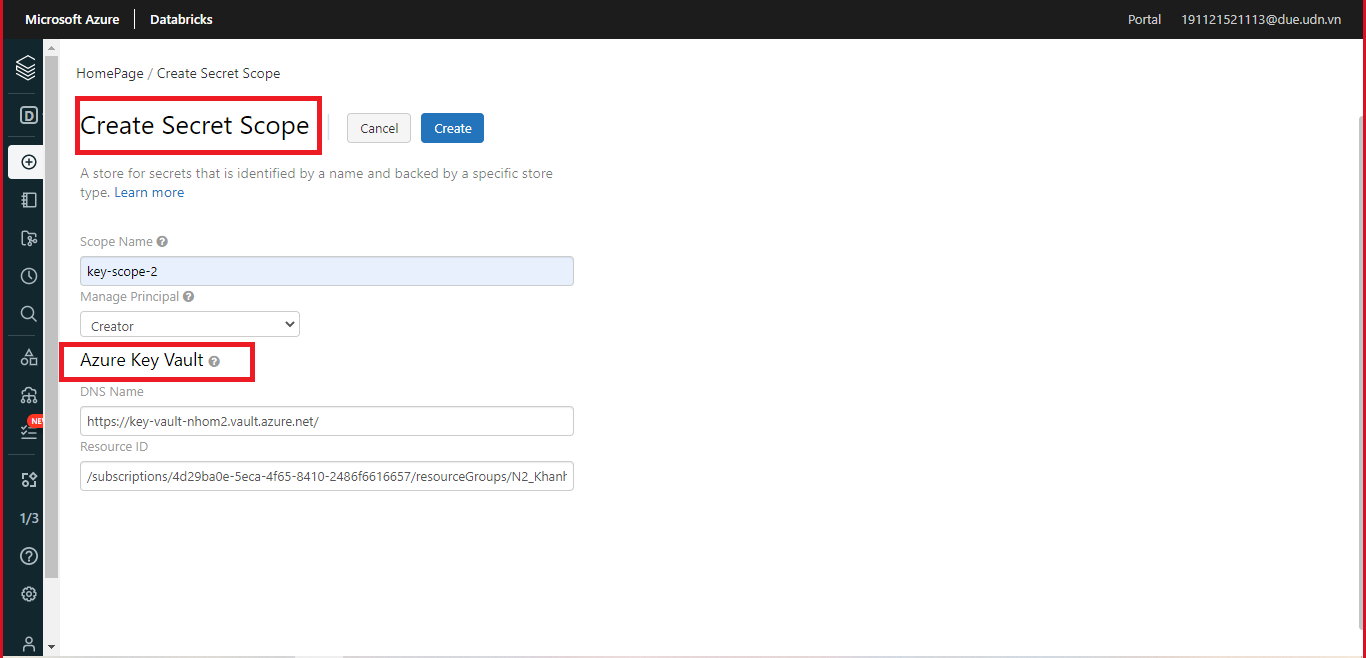


*Tạo Key Secret trong Key Vault*

Tạo môt vùng secret trong databrics thông qua các thuộc tính của Key Vault được tạo thành

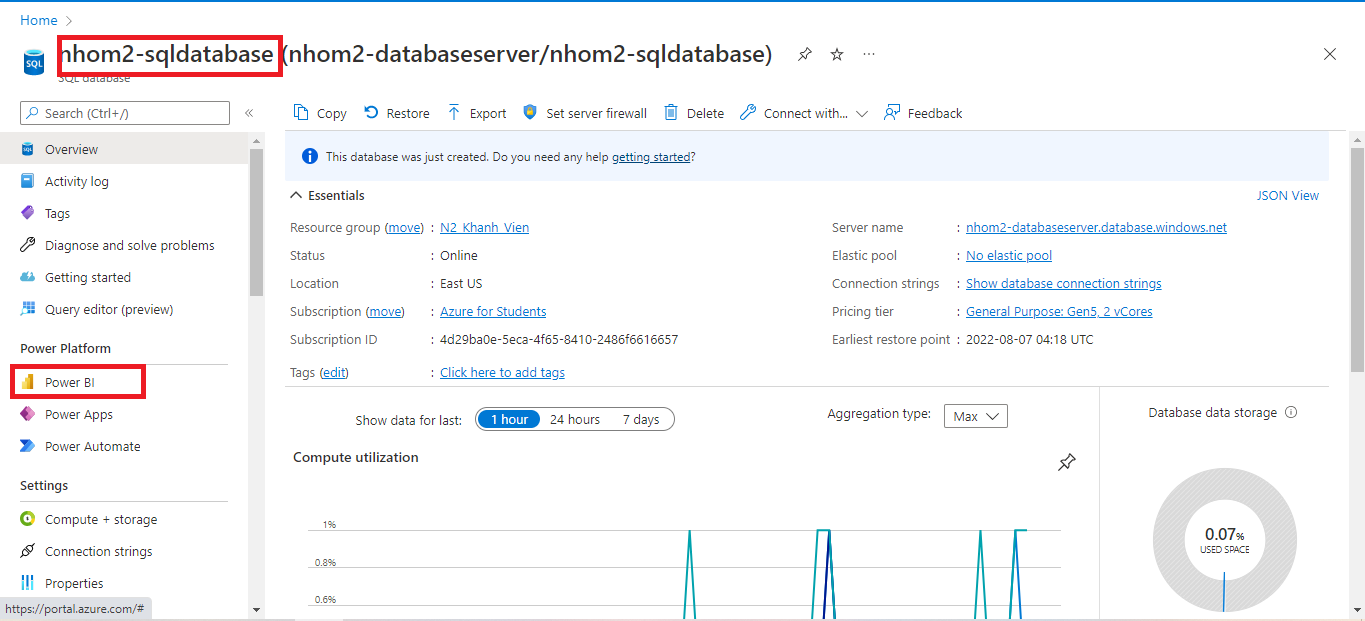


*Thông tin của Key Vault*



*Tạo Secret Scope*

Tạo một SQL Database để chứa dữ liệu sạch sau khi ETL tại Databricks và visualize trên Power BI.

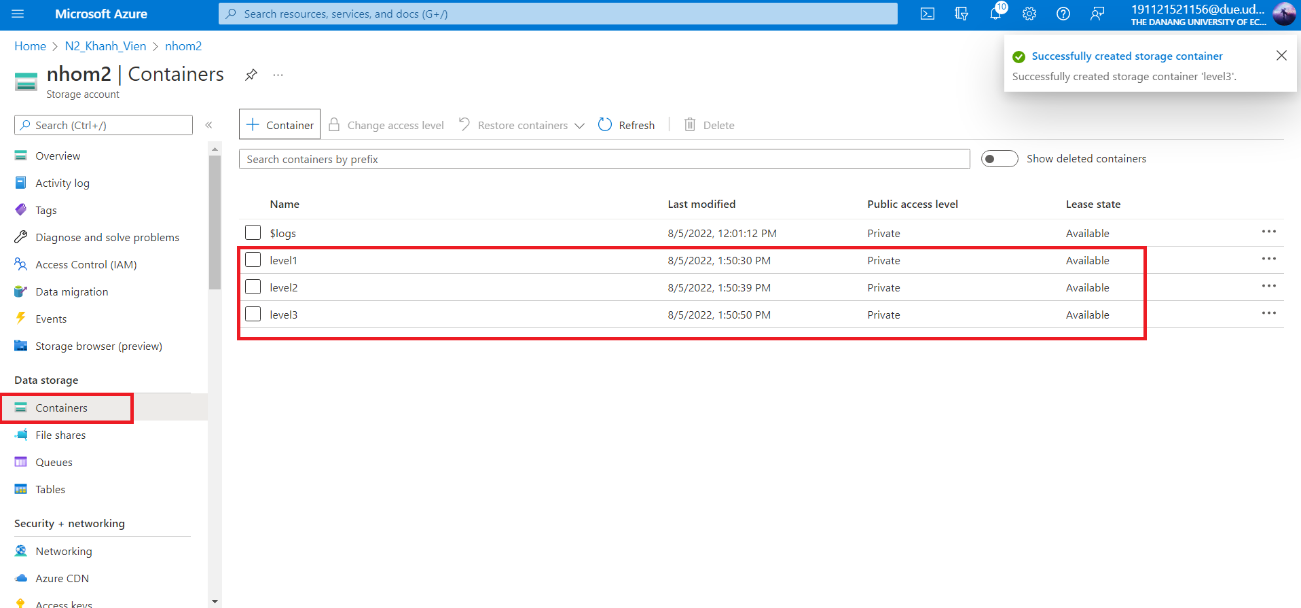


*SQL Database tích hợp Power BI*

1. Tạo Container chứa dữ liệu 3 mức

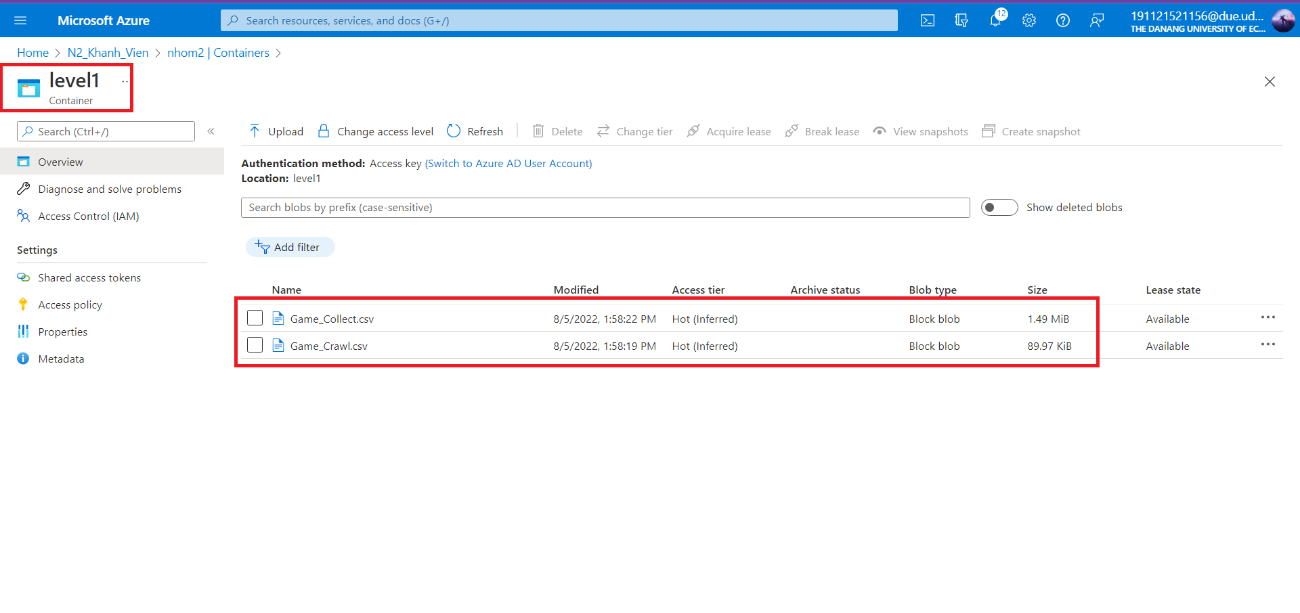
Trong mục Container Storage Account chứa dữ liệu 3 lớp gồm: level 1, level2 và level 3. Mỗi level chứa những dữ liệu như sau:

* Level 1: Chứa dữ liệu thô sau khi clean về và dữ liệu tìm được trên Kangel
* Level 2: Chứa dữ liệu copy từ level1 thông qua ADF
* Level 3: chứa dữ liệu sạch sau khi đã được ETL tại Databricks



*Tạo 3 Container chứa dữ liệu 3 mức*

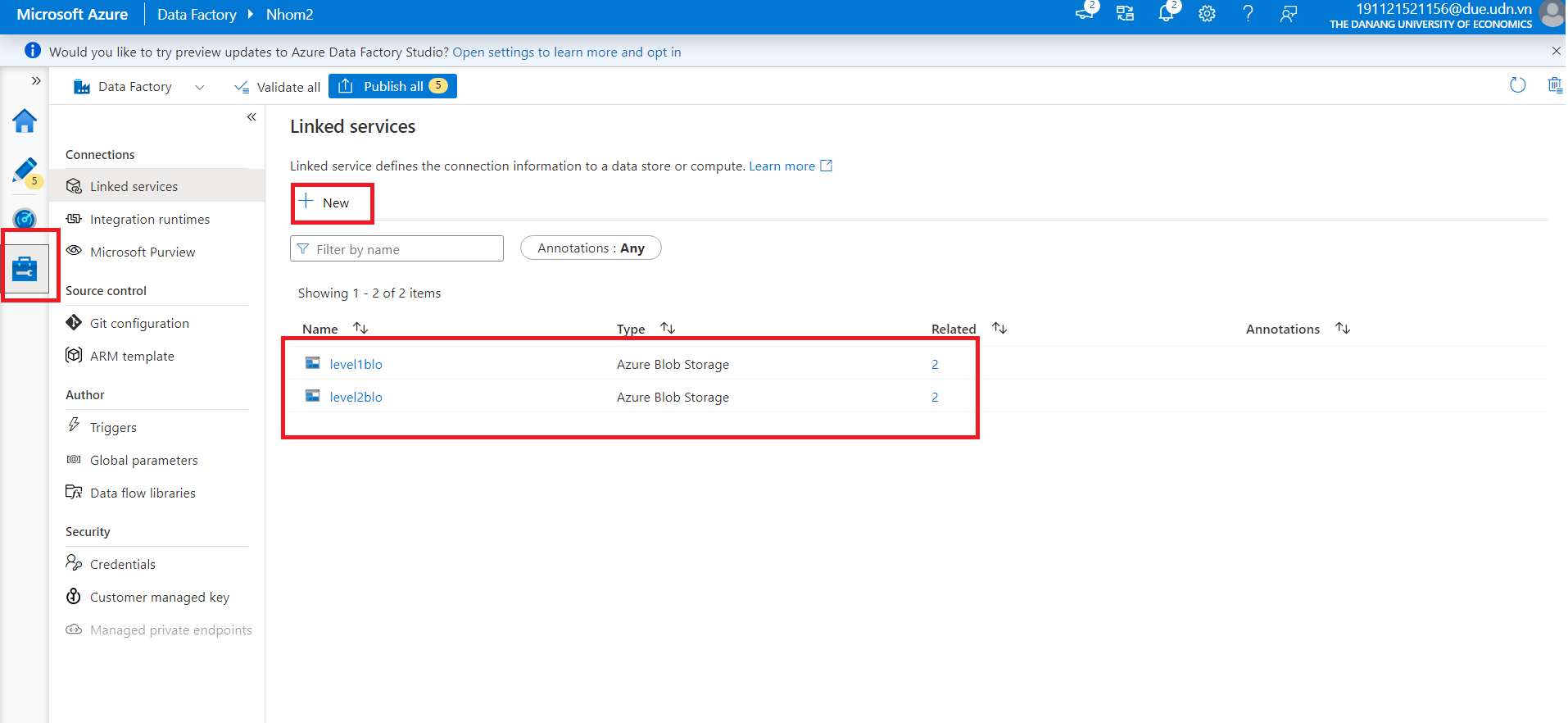
Trong Container level1 Upload 2 file dataset thô gồm dataset Crawling được và dataset tìm được trên Kangel.



*Upload dataset vào Container “level1”*

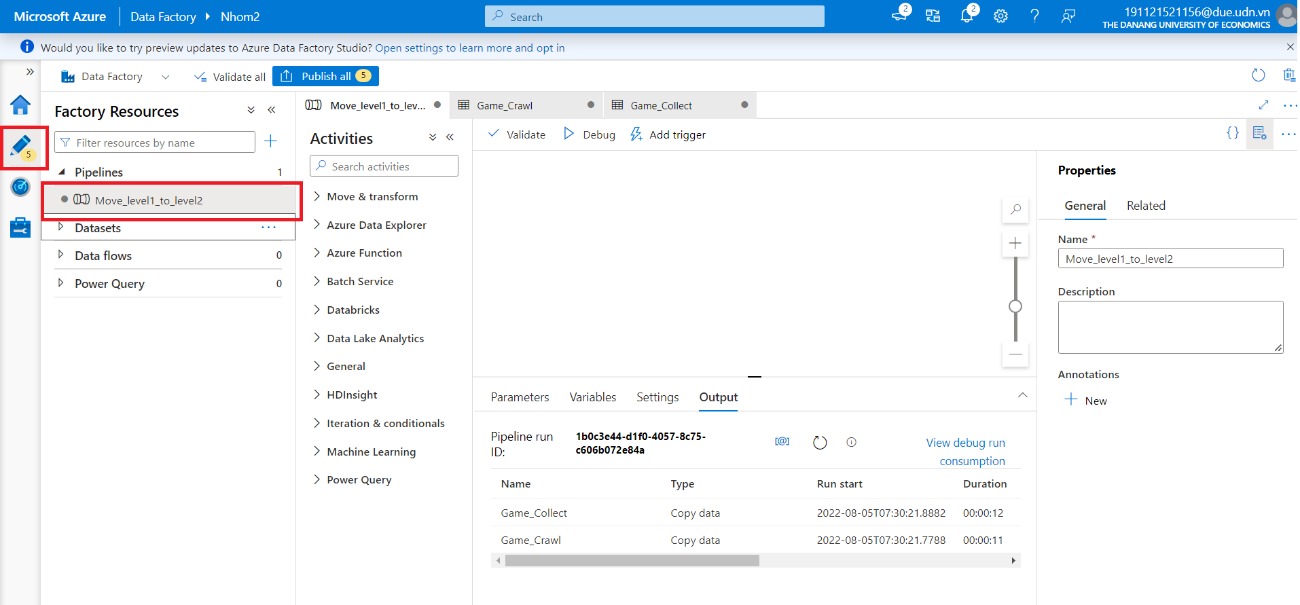
1. Copy dataset ở level1 đến level bằng ADF

Để tạo dataset copy trong Container level2 phải tạo Linked Services để kết nối dataset trong container level1 đến ADF. Sau đó tạo một Linkes Services để lấy dataset từ ADF về Container level2.



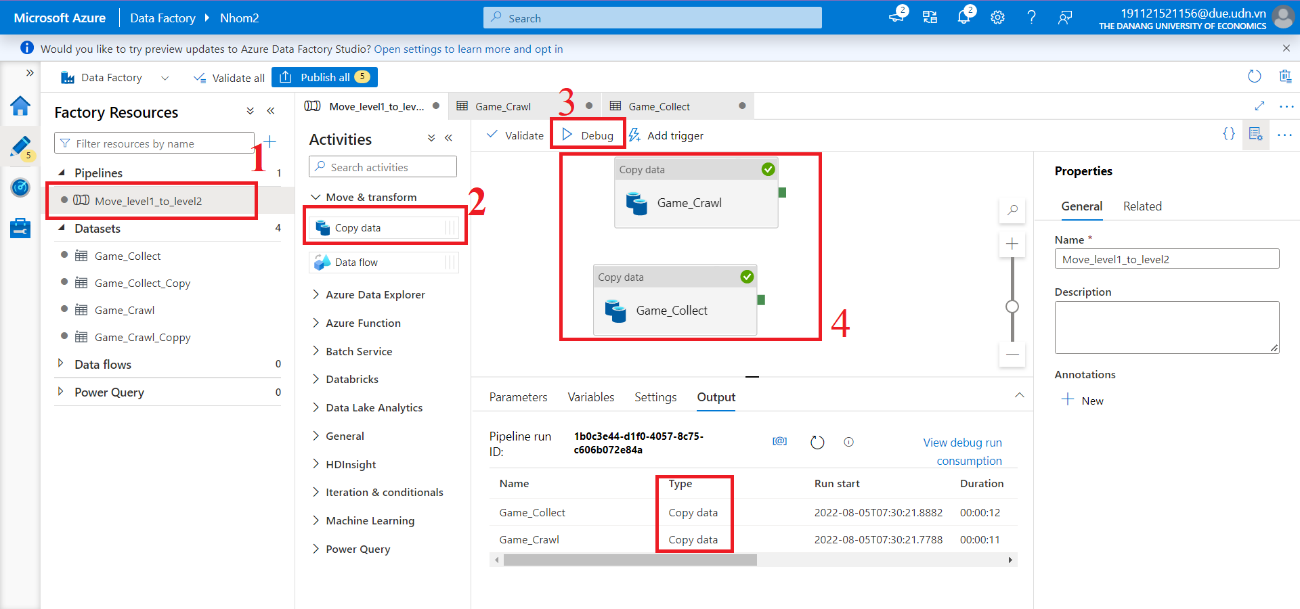
*Tạo Linked Services*

Sau đó tạo Pieline để tiến hành Copy data

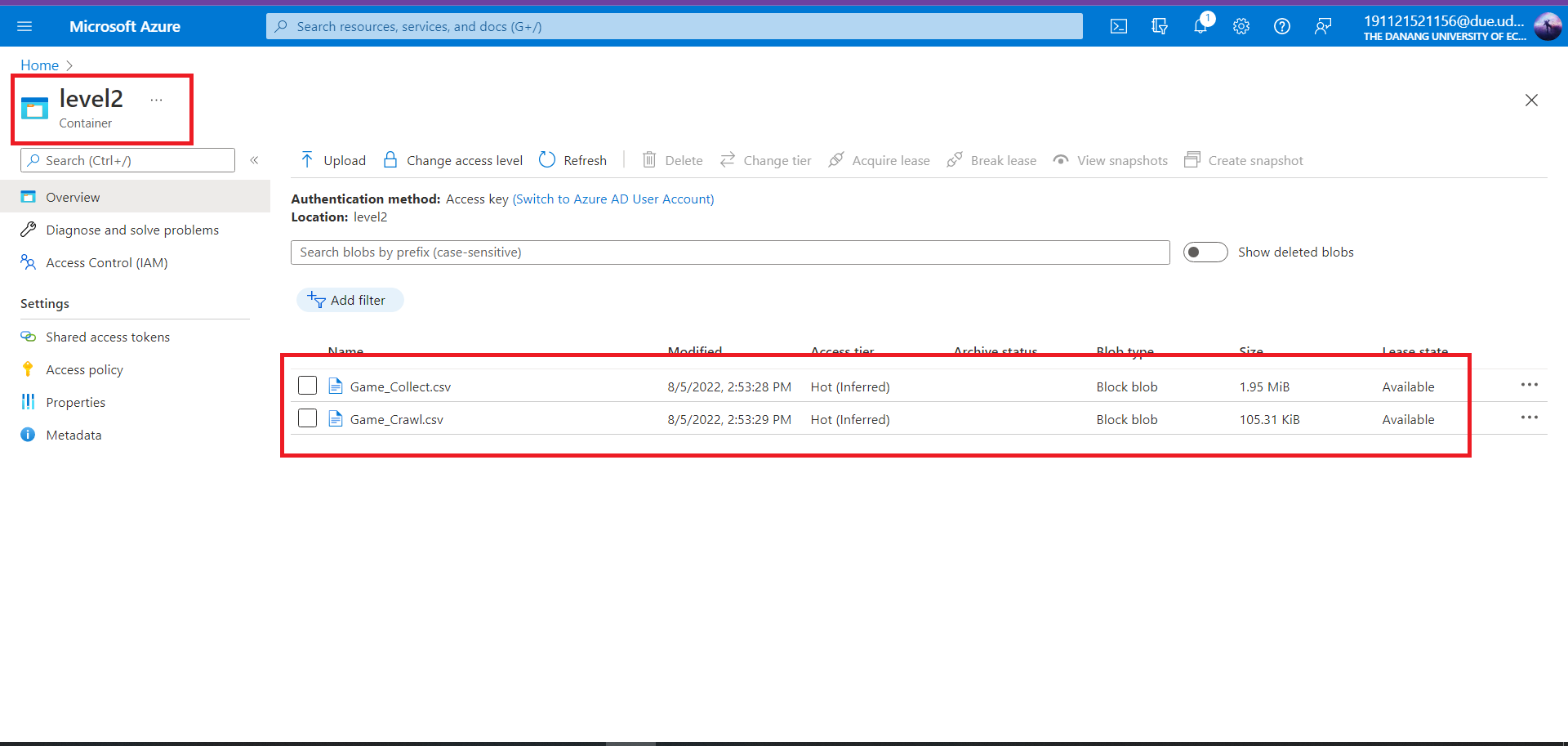
**

*Tạo Pieline*

Tiến hành thao tác copy data trên Pieline vừa tạo.

**

*Copy data đến Container level2 trên Pieline vừa tạo*

**

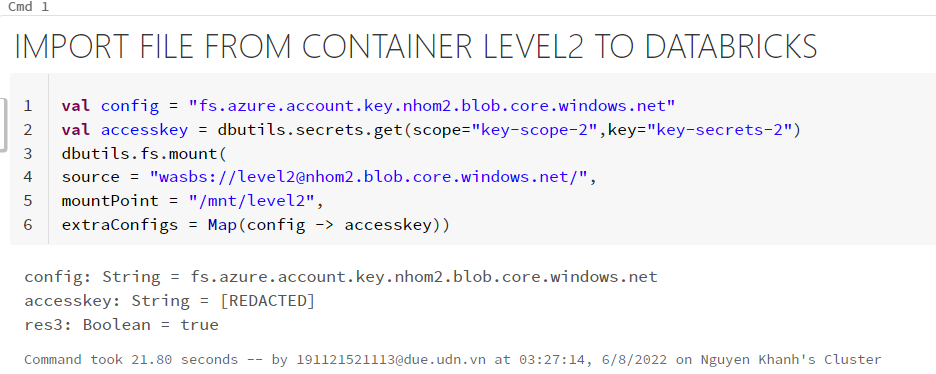
*Copy dataset thành công*

Dữ liệu sau khi copy trên ADF sẽ xuất hiện tại Container “level2”.

1. Kết nối Container với Databricks bằng ADF và xử lý dữ liệu

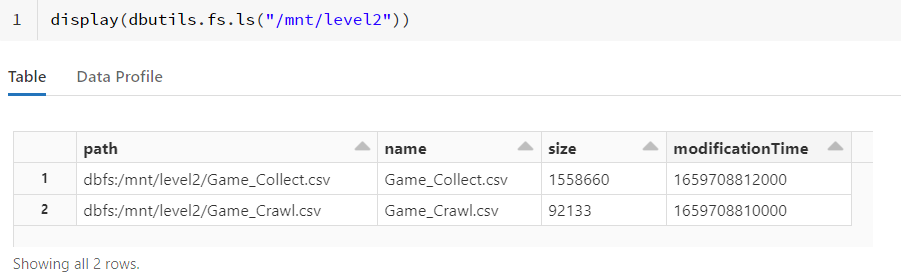
Để tạo ra một file dữ liệu sạch dưới định dạng .csv trên nền tảng Databricks đối với dataset thô mà nhóm Scrawl được thì cần phải trải qua nhiều bước tiến hành khác nhau.

Trước tiên, tiến hành lấy dataset từ Container “level2” về và đổ vào dataframe thông qua đường nối. Ở đây, để bảo mật dữ liệu, tránh bị đánh cắp data khi key của Storange Account bị đánh mất ra ngoài, nhóm thay thế bằng key secret được tạo trong key vault để bảo mật.



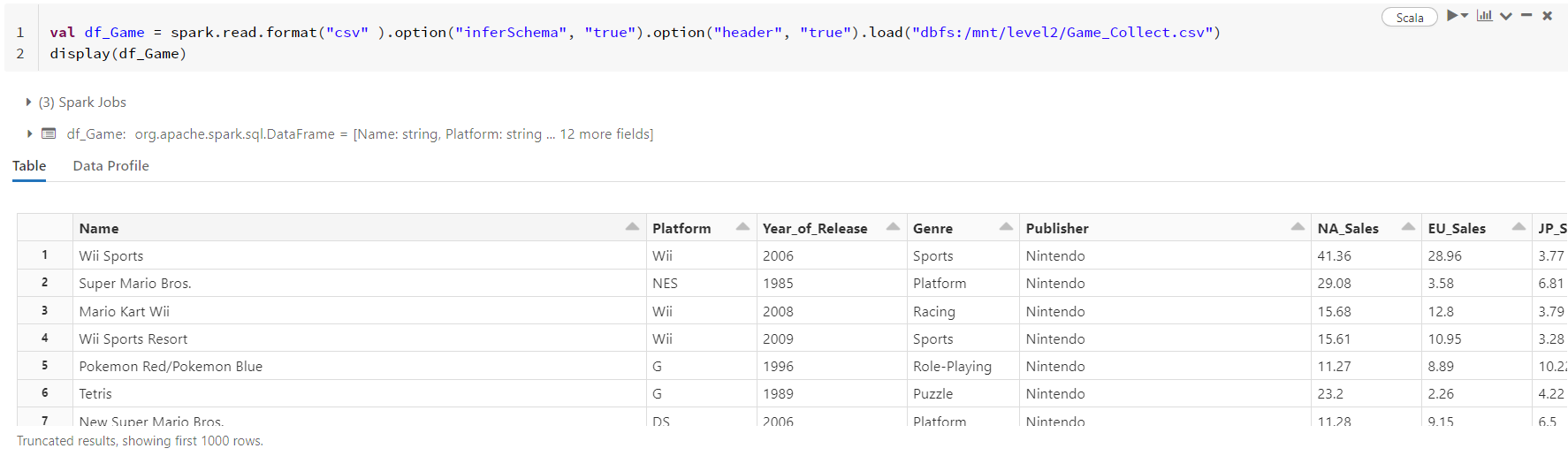
*Nối databricks với container level2*

Sau khi Databricks nối với container, tiến hành kiểm tra và đọc file.

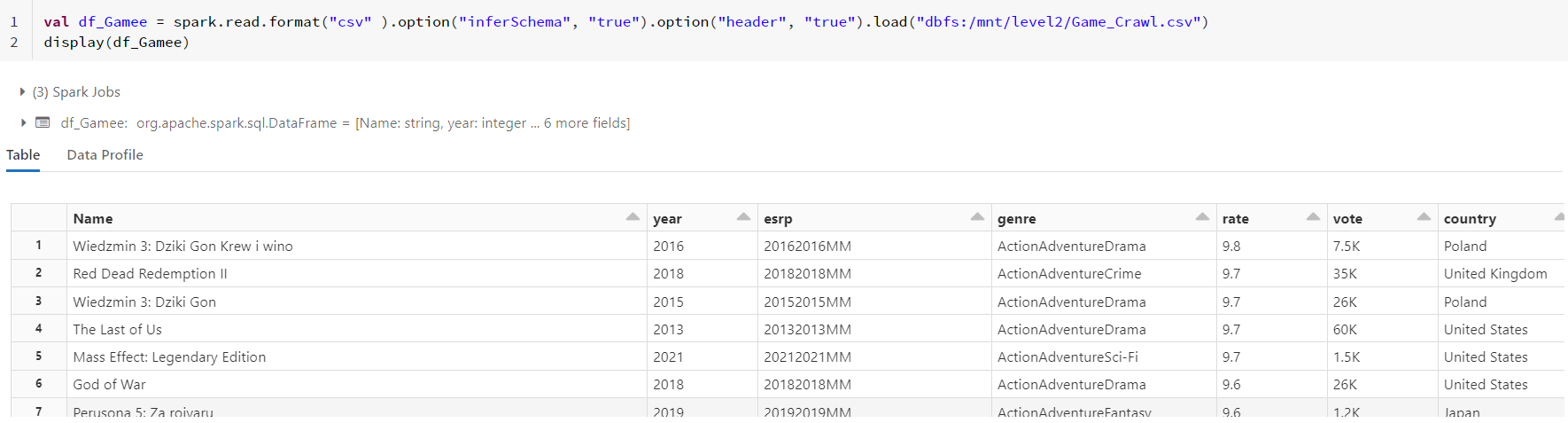


*Đọc File*

Tiếp đến là gán từng dataframe cho mỗi file dataset. Ở đây, nhóm có 2 dataset gồm: 1 dataset crawling được và 1 dataset cũng về Video Game tìm kiếm trên Kangel( vì dataset tìm được ít dữ liệu định lượng, khó visualize).



*Dataset tìm được trên Kangel*

**

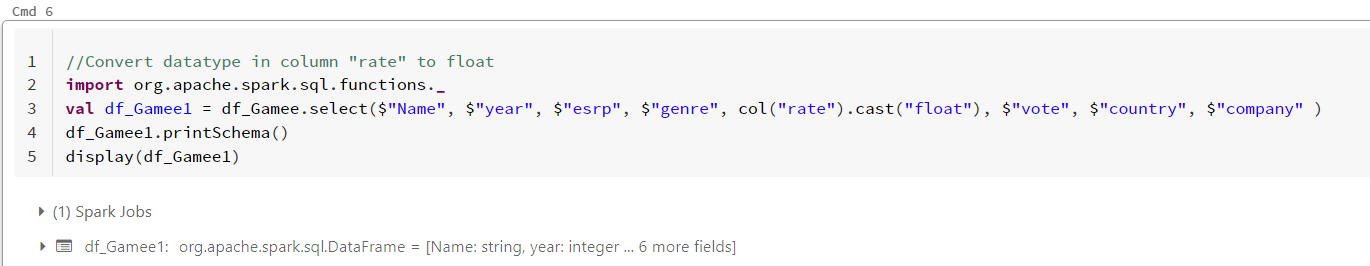
*Dataset sau khi Crawl*

Kế tiếp, sau khi đã có dataframe chứa dữ liệu của từng dataset, nhóm tiến hành ETL. Trước tiên sẽ xử lý dữ liệu ở dataframe chứa dataset crawl được. Đối với bước này nhóm tiến hành kiểm tra kiểu dữ liệu.



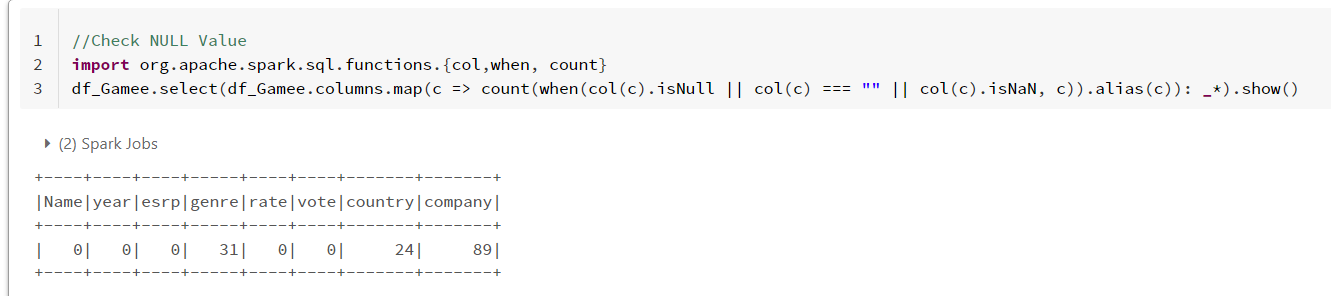
*Kiểm tra kiểu dữ liệu*

Sau khi kiểm tra, cột rate có kiểu dữ liệu Double, để phù hợp hơn khi đổ vào SQL Database nhóm chuyển thành kiểu Float.

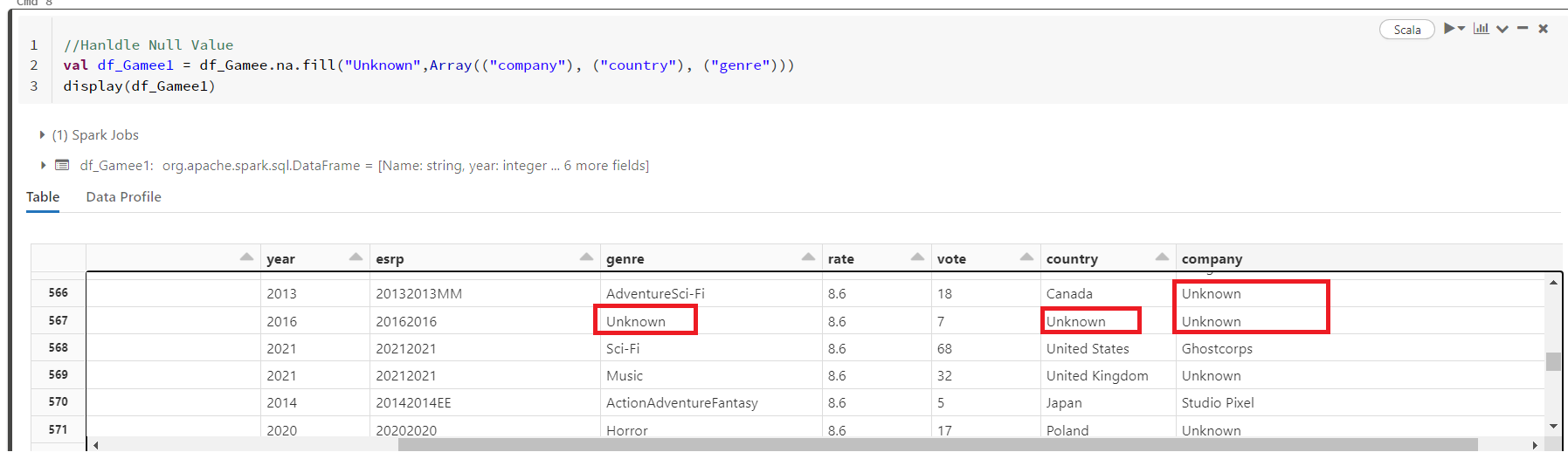


*Chuyển kiểu dữ liệu của cột rate*

Tiếp đến, kiểm tra xem dataset có Null, NaN, hoặc trống không.

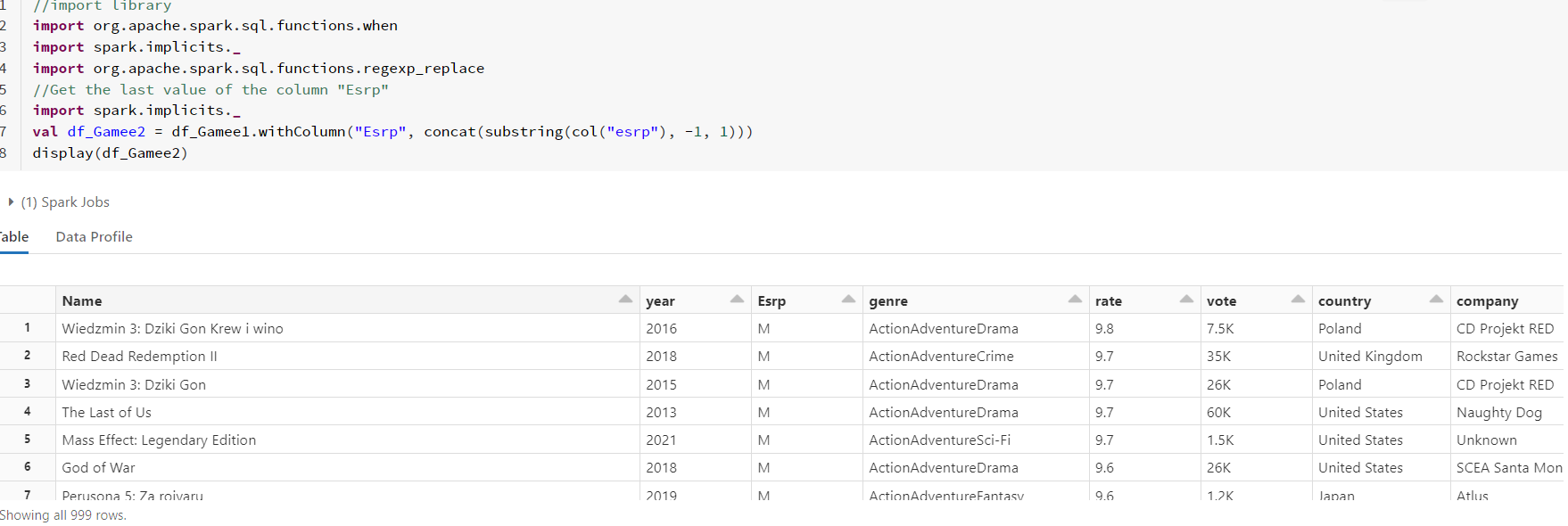


*Kiểm tra Null value*

Dựa vào kết quả nhận được dễ dàng nhận thấy, data có 3 column chứa dữ liệu Null. Đối với những cột này mang giá trị định tính, những giá trị Null sẽ được thay thế bằng “Unknown”.

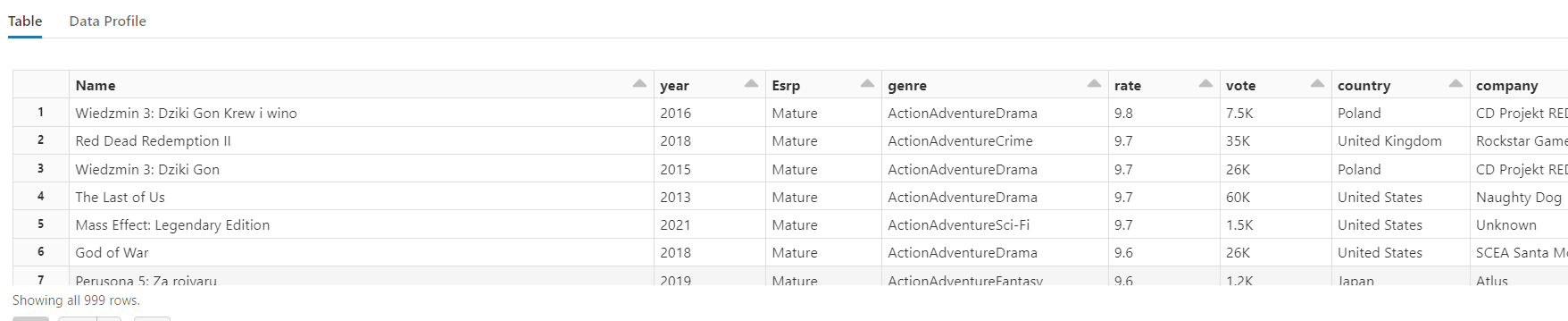
*Cột chứa dữ liệu Null*

Nhìn vào data mà nhóm Crawl được sẽ bị lỗi ở 2 cột là Esrp và Genre. Để xử lý lỗi này, nhóm tiến hành lấy giá trị cuối ở cột Esrp(mỗi chuẩn game sẽ có kí hiệu là: M, E10+, E,…) và cắt chuỗi ở cột Genre để phân loại game.

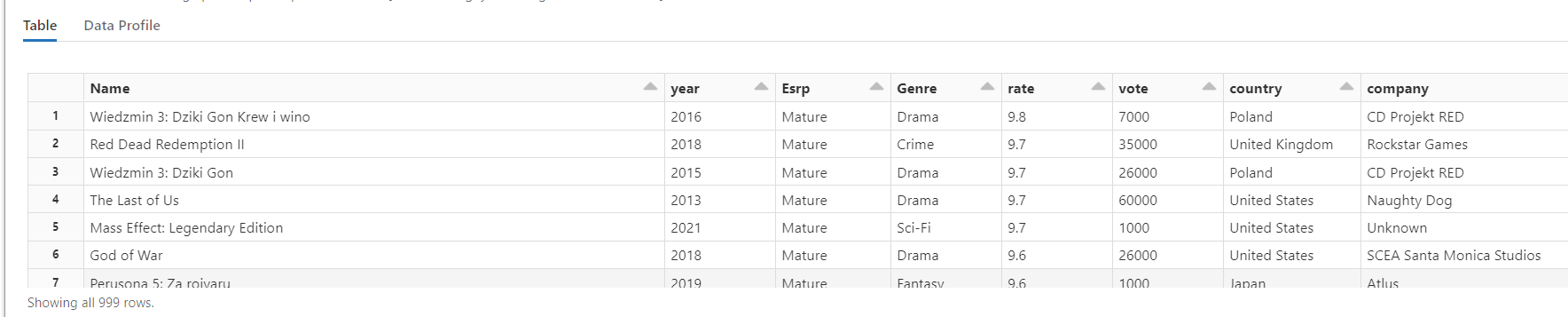


*Xử lý dữ liệu cột Esrp*

Sau khi cắt lấy giá trị cuối ở cột Esrp và Genre, để dễ dàng quan đọc hiểu data cho người dùng và visualize được trực quan, nhóm tiến hành thay thế kí tự.

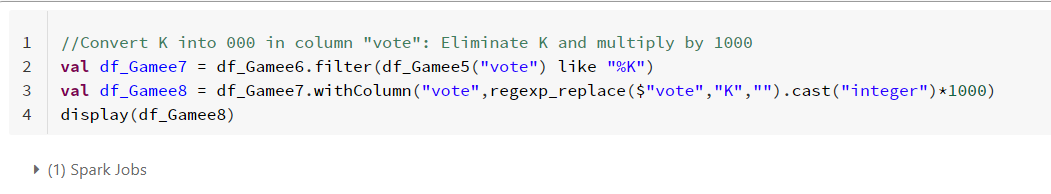


*Cột “Esrp” sau khi thay thế*



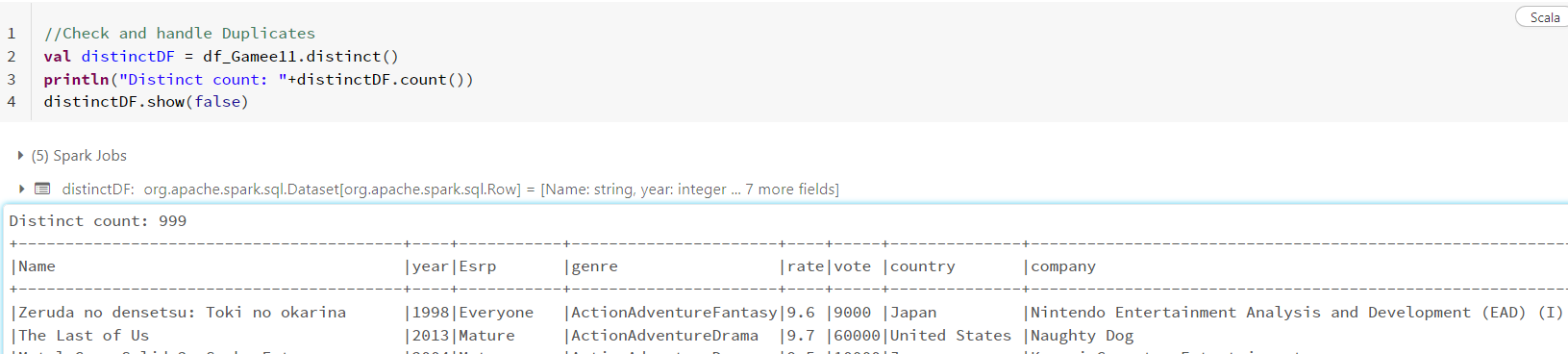
*Cột “Genre” sau khi thay thế*

Tiếp đến, ở cột Vote, để có thể so sánh lượt bình chọn, nhóm tiến hành đổi giá trị từ dạng định lượng sang định tính bằng cách bỏ K và nhân cho 1000 đối với những value có K.



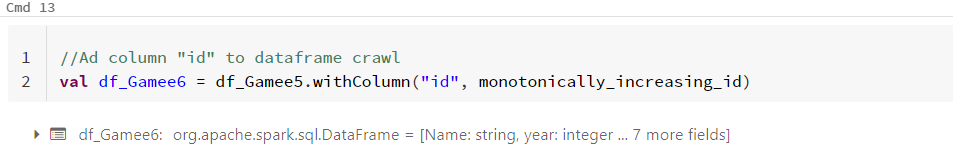
*Xử lý dữ liệu cột Vote*

Tiếp đến, nhóm tiến hành kiểm tra những dòng dữ liệu trùng lặp, tuy nhiên dataset không có dữ liệu bị Duplicates, nên sẽ không xử lý.



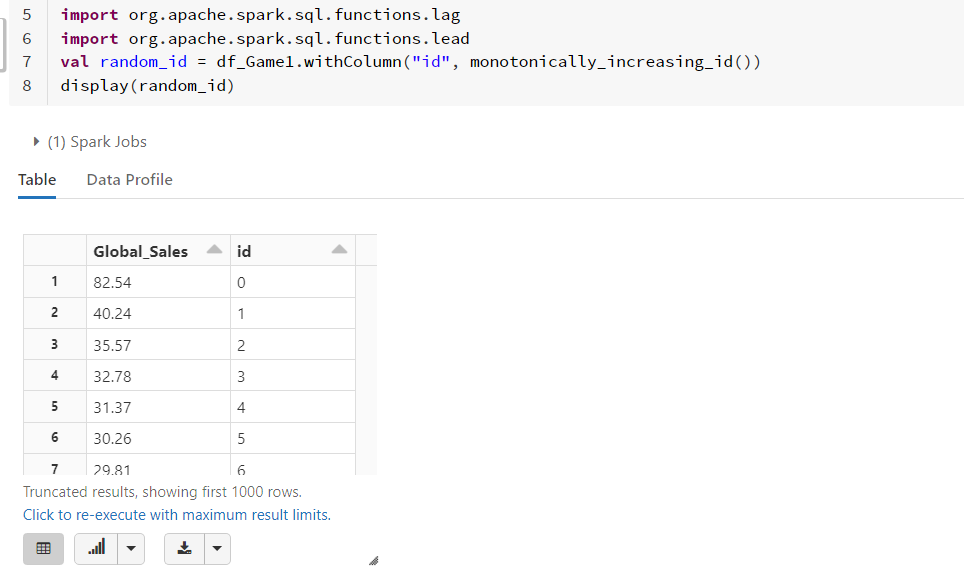
*Kiểm tra dữ liệu trùng lặp*

Cuối cùng, tạo một cột id tự động từ 1 đến 999 để làm khoá tham chiếu join với dataframe chứa dataset tìm được trên Kangel.



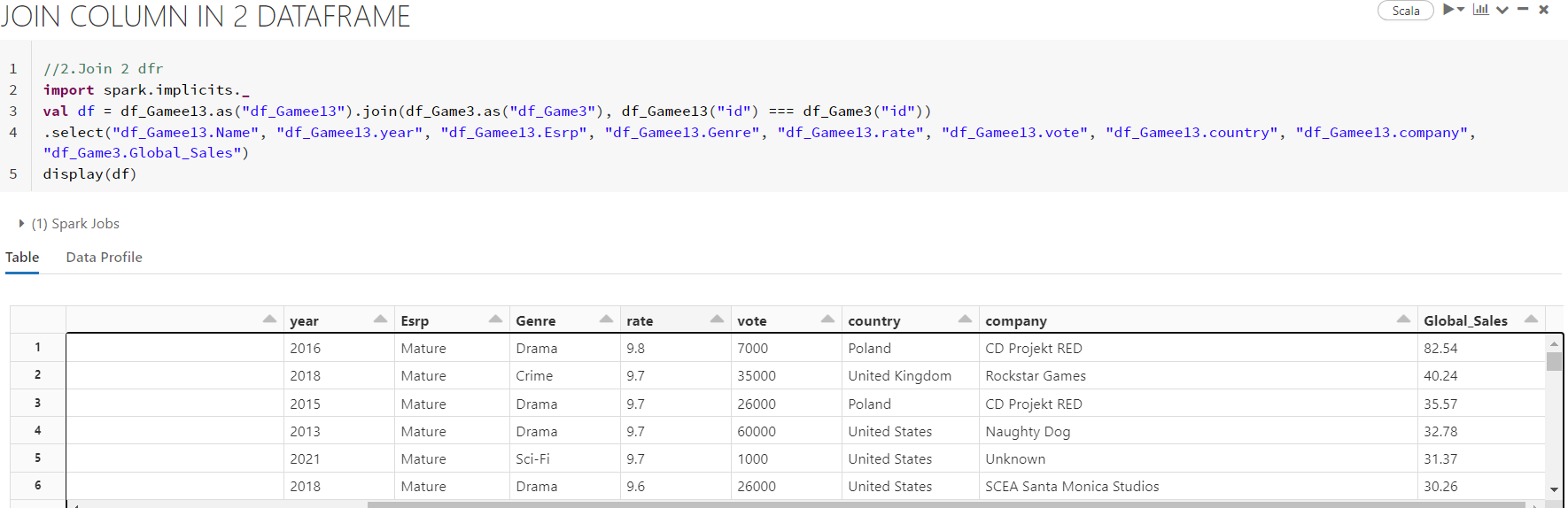
*Thêm cột id*

Dữ liệu tìm được trên Kangel đã sạch nên không cần xử lý, chỉ tạo thêm một column id tự động để join với dataset crawl đã xử lý sạch ở trên.



*Thêm cột id cho data trên Kaggle*

Tiếp đến là join 2 dataframe theo cột id đã tạo.



*Join 2 dataframe*

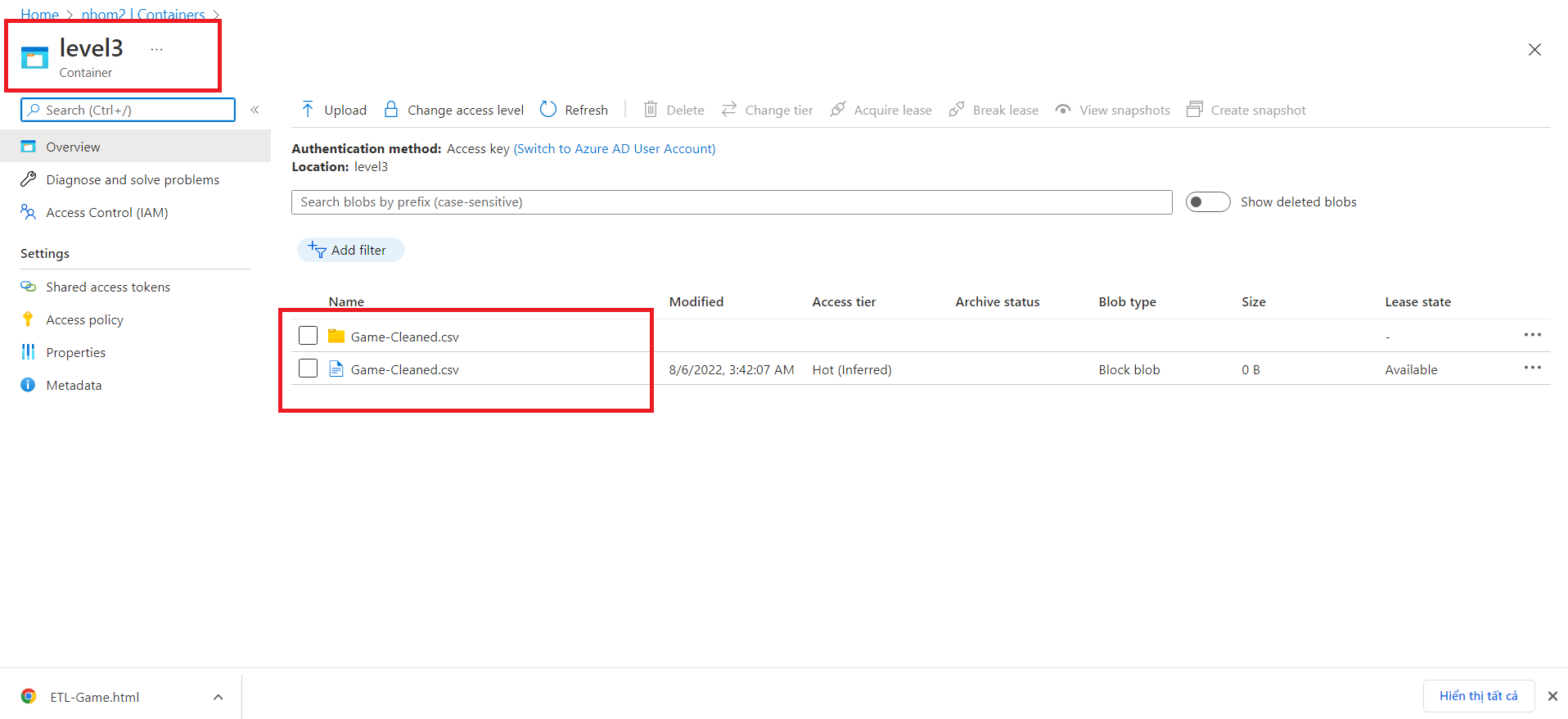
1. Chuyển dữ liệu sạch từ Databricks về Container qua ADF

Kiểm tra lại một lần nữa, dataframe cuối cùng đảm bảo đủ dòng và cột thì tiến hành xuất dataframe thành file có định dạng .csv đến Container chứa dữ liệu sạch là “level3”.



*Lệnh thực hiện chuyển dữ liệu về container- level3*

Kiểm tra lại trên Container “level3” đã có dataset nghĩa là hoàn thành.

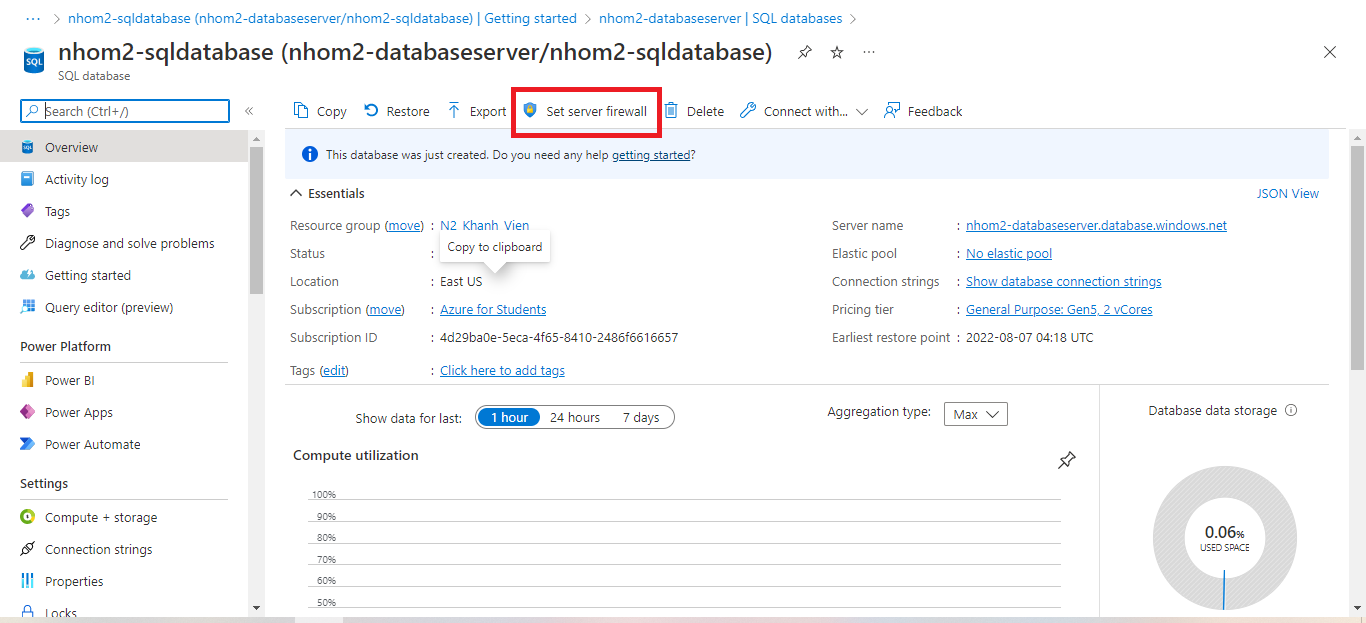


*Chuyển dữ liệu hoàn thành*

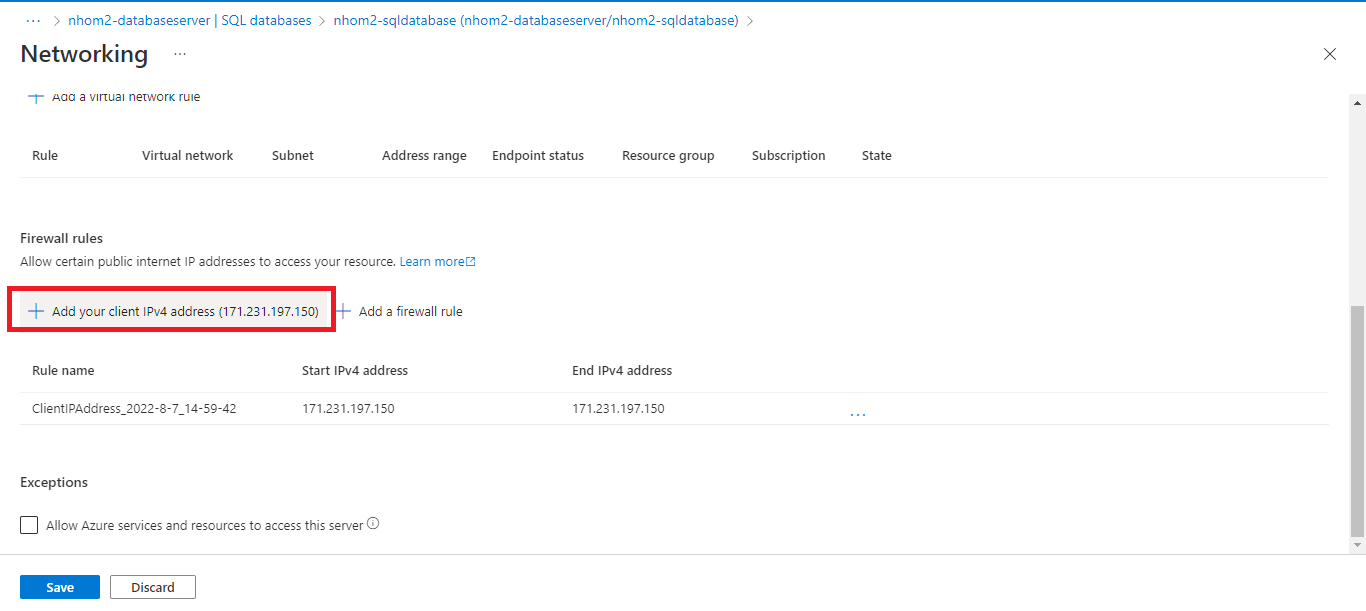
1. Chuyển dữ liệu sạch tại Container vào SQL Database

Để chuyển dữ liệu từ Container level 3 vào SQL Data Base, nhóm tiếp tục tạo một Pieline trong ADF. Sau đó tiến hành chọn data copy. Click vào Data Copy tiến hành chọn Source kết nối đến container level3 và Destination là Database đổ ra dữ liệu.

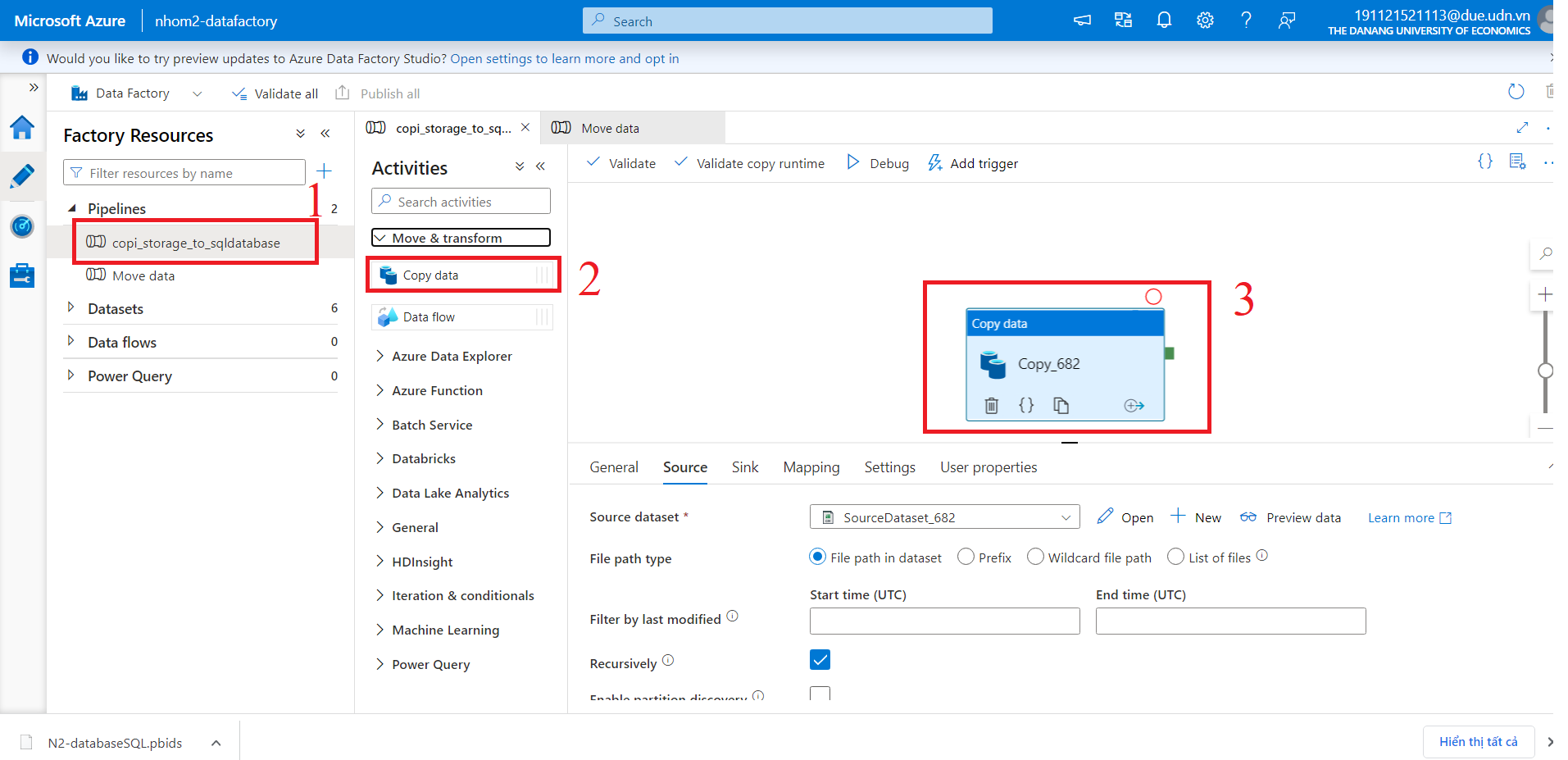
Phải thêm vào Ipv4 cho database để có thể connect từ container đến thông qua phần Set Server Firewall



*Set Serve Firewall*

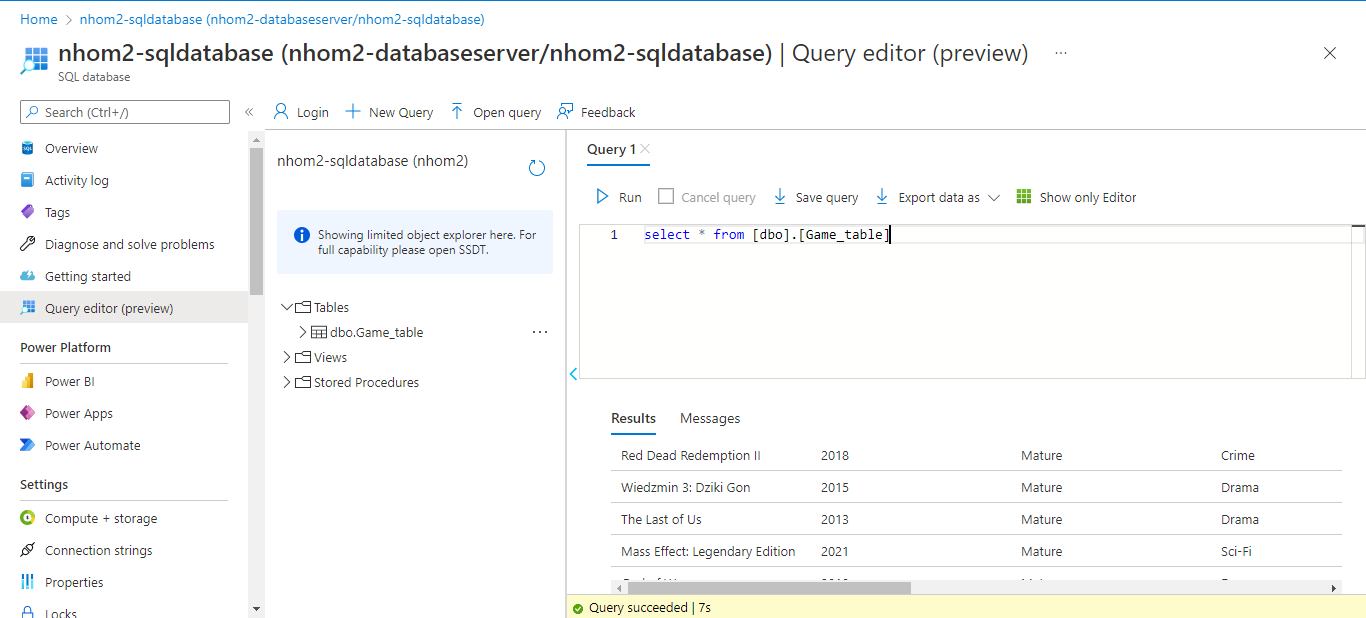


*Thêm vào Client ip*



*Tạo pipeline*

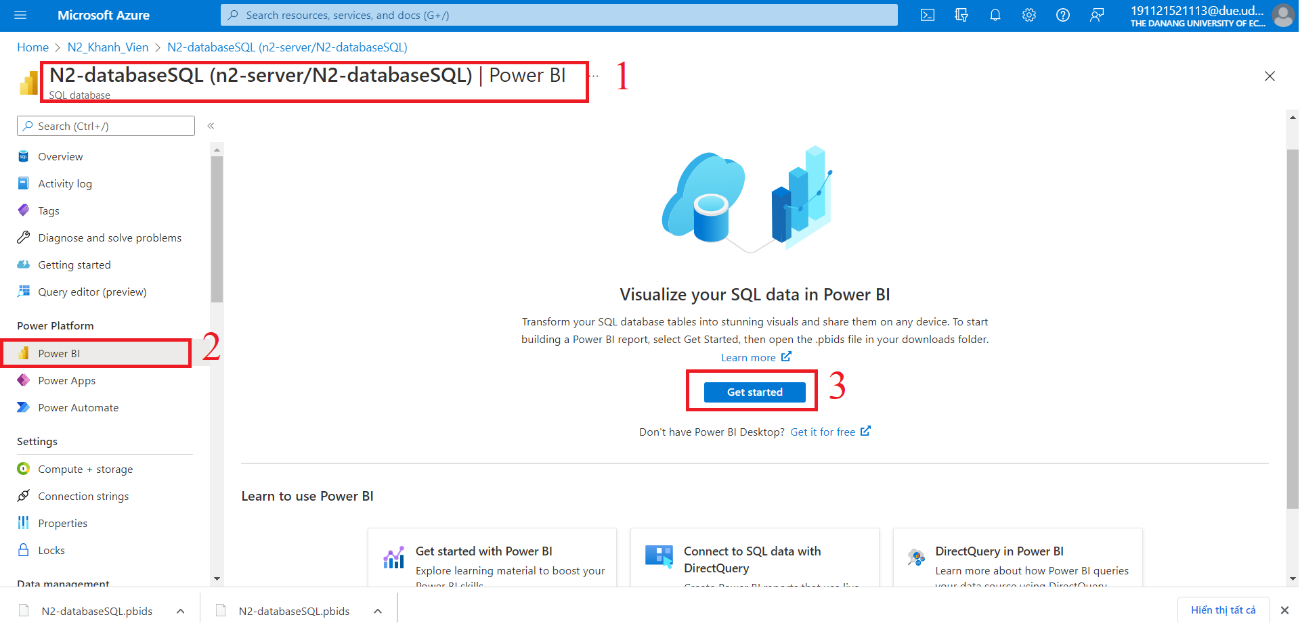
Sau khi set up kiểm tra lại trên SQL DataBase.



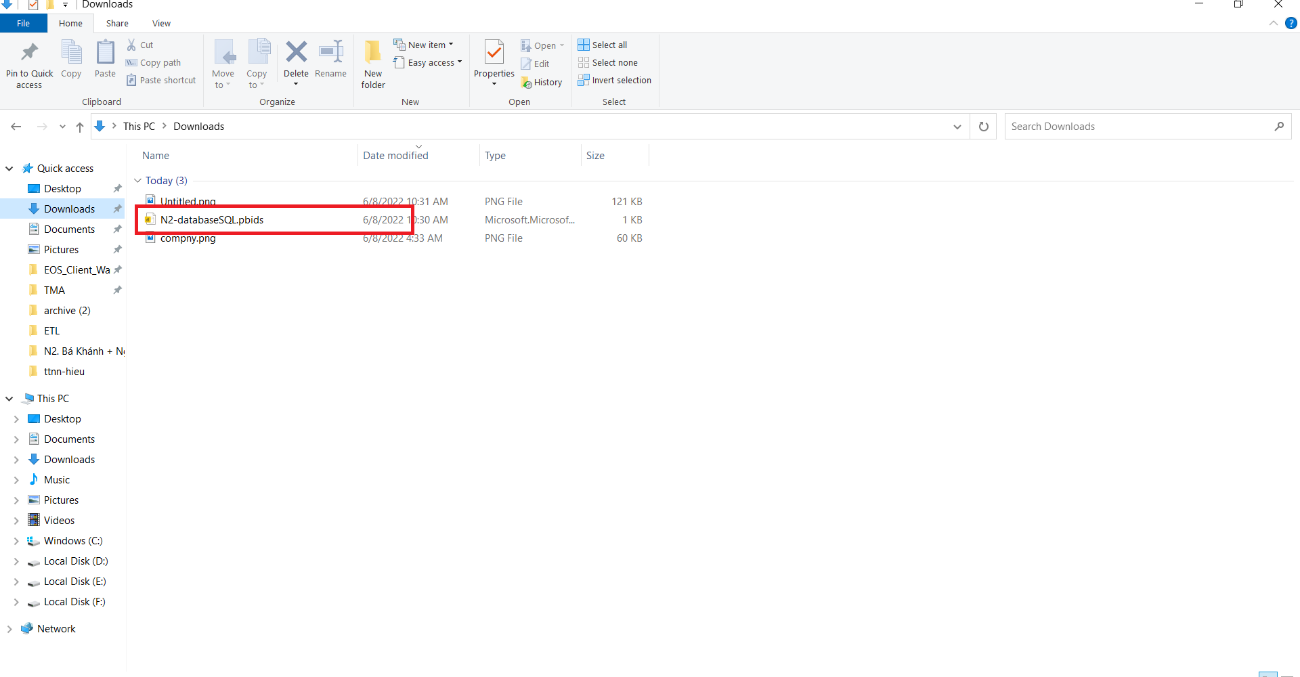
*Đã đưa dữ liệu thành công đến sql database*

**2.3. Trực quan hoá dữ liệu bằng Power BI**

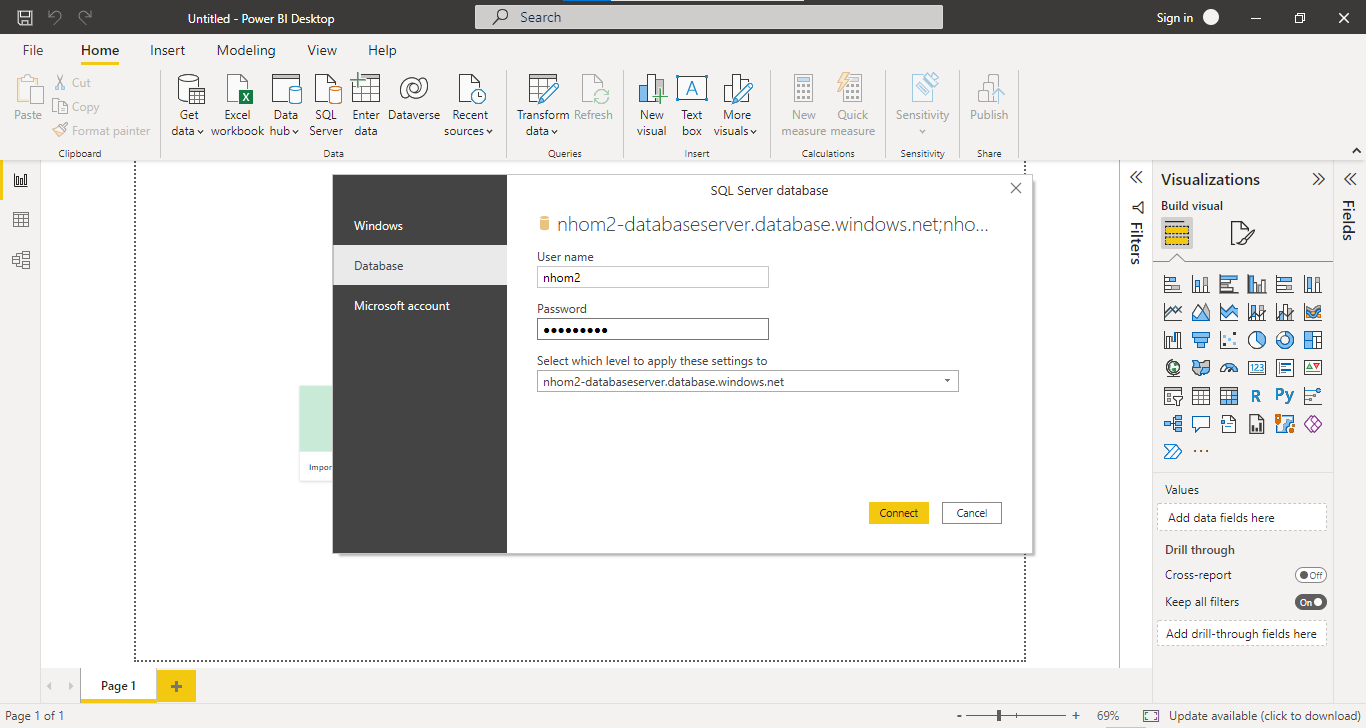
Để visualize dataset vừa tạo bằng Power BI, nhóm tiến hành dowload file Power BI có chứa data vừa chuyển tại Database SQL về và kết nối.

****

Sau đó sẽ có một file Power BI tải về, tại đây nhóm tiến hành kết nối và visualize.



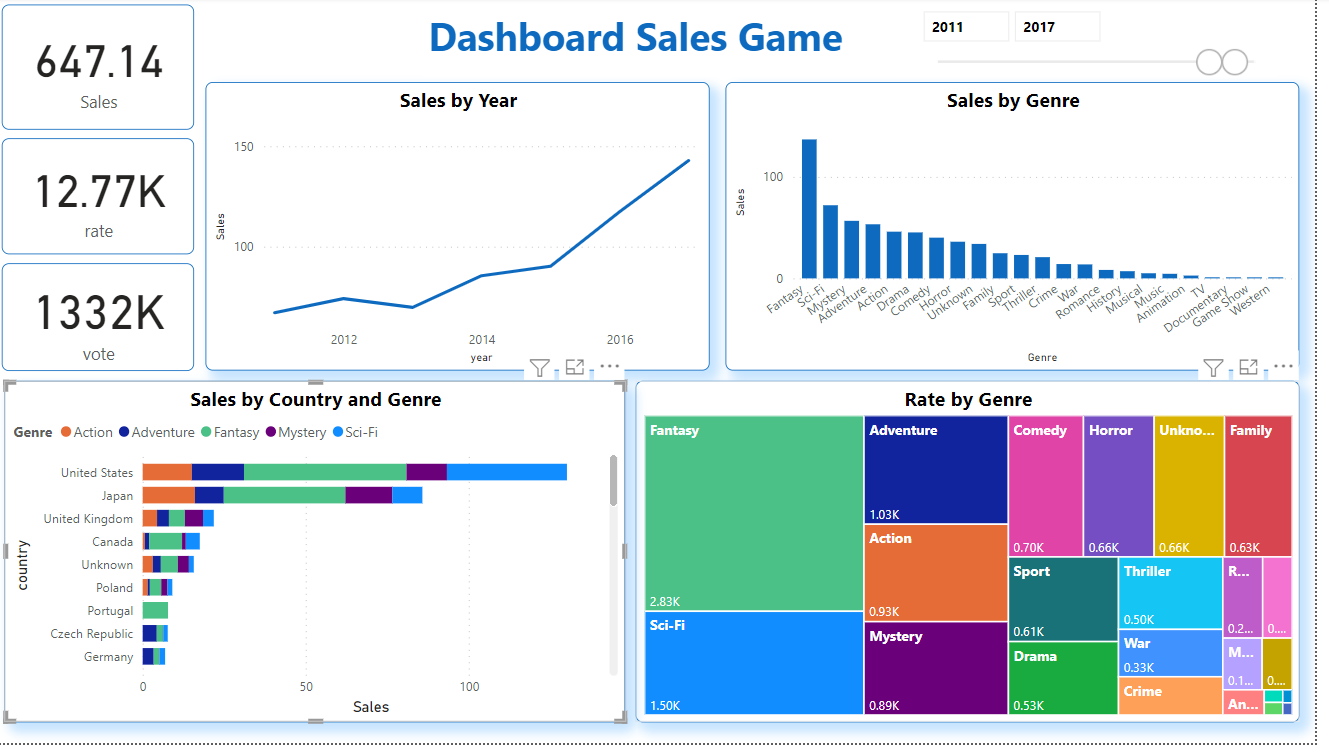
*File Power BI tải về*



*Kết nối đến SQL Database bằng tài khoản xác thực đã tạo trong SQL Servver*

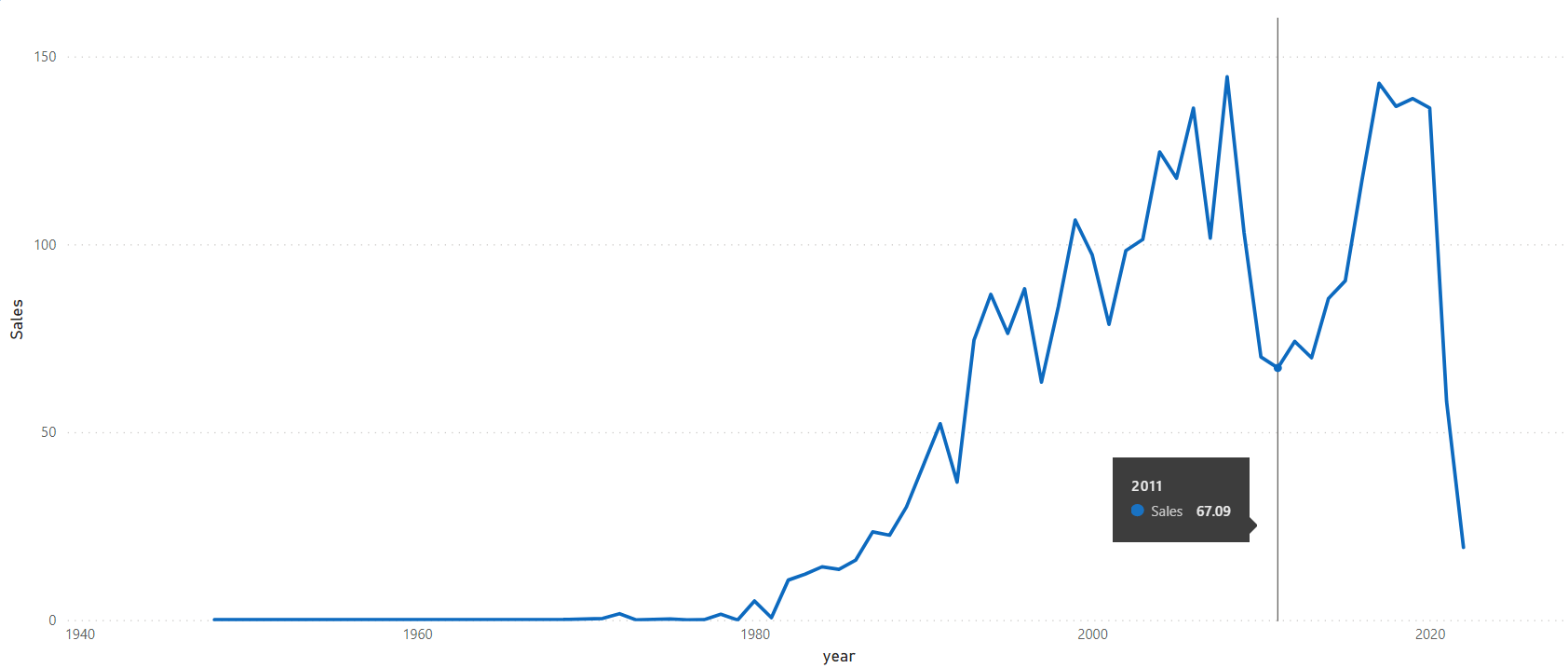
# Chương 3 KẾT QUẢ

Mục đích trực quan hoá dữ liệu thông qua dataset của nhóm nhằm tìm kiếm thể loại game đang mang lại doanh thu cao và sẽ là xu hướng trong những năm kế đến dựa trên kết quả của những năm qua. Dưới đây là bảng dashboard chung mà nhóm đã xây dựng được từ nguồn dataset đã ETL.



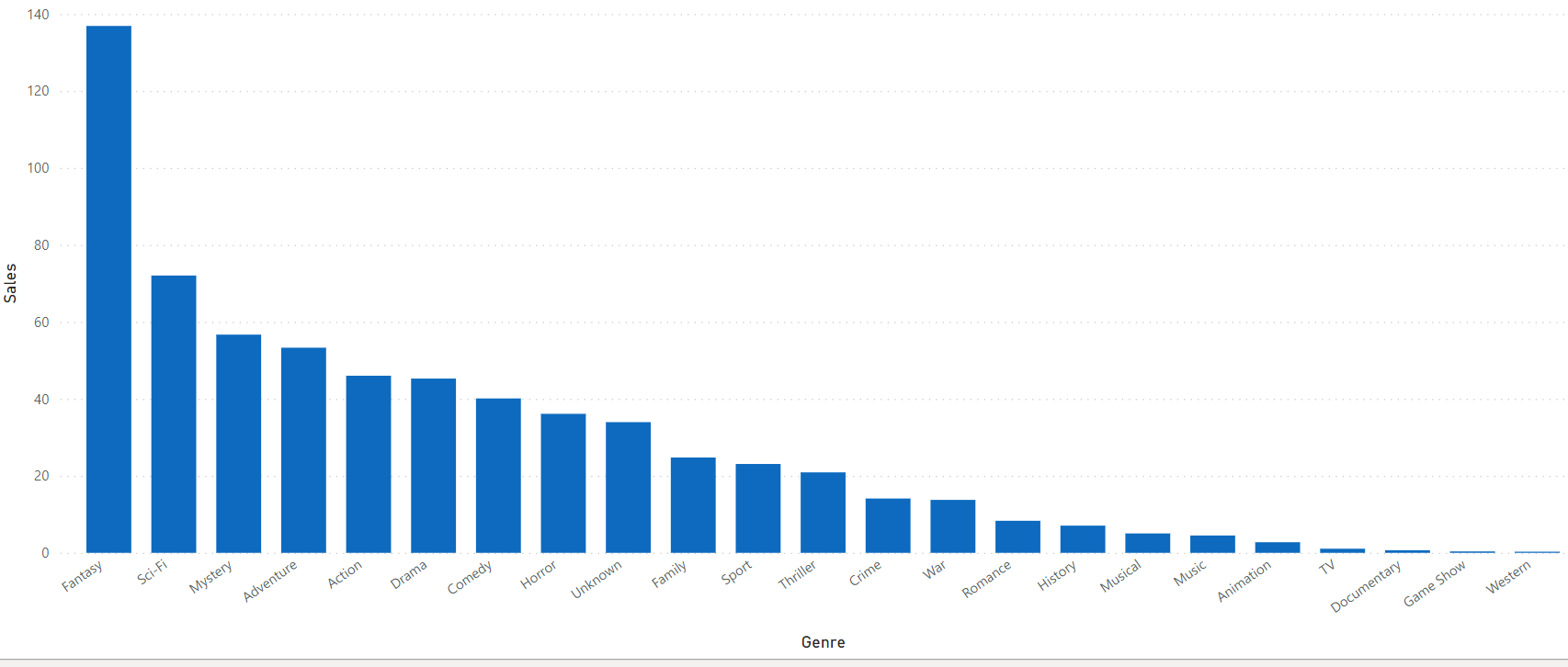
*Dashboard overview*

Trước tiên, dataset thu thập thông tin game qua các năm từ 1948 – 2022, tuy nhiên nhóm chỉ chọn khoảng từ 2011 – 2017 là thời gian mà game phát triển vượt bậc, đem đến doanh thu tổng cao nhất, đây là khoảng thời gian mà thị trường game sau khoảng thời gian giảm mạnh về doanh thu, từ mốc gần như thấp nhất (2011) đến phát triển đỉnh điểm (2017).



*Char: Sales by year*

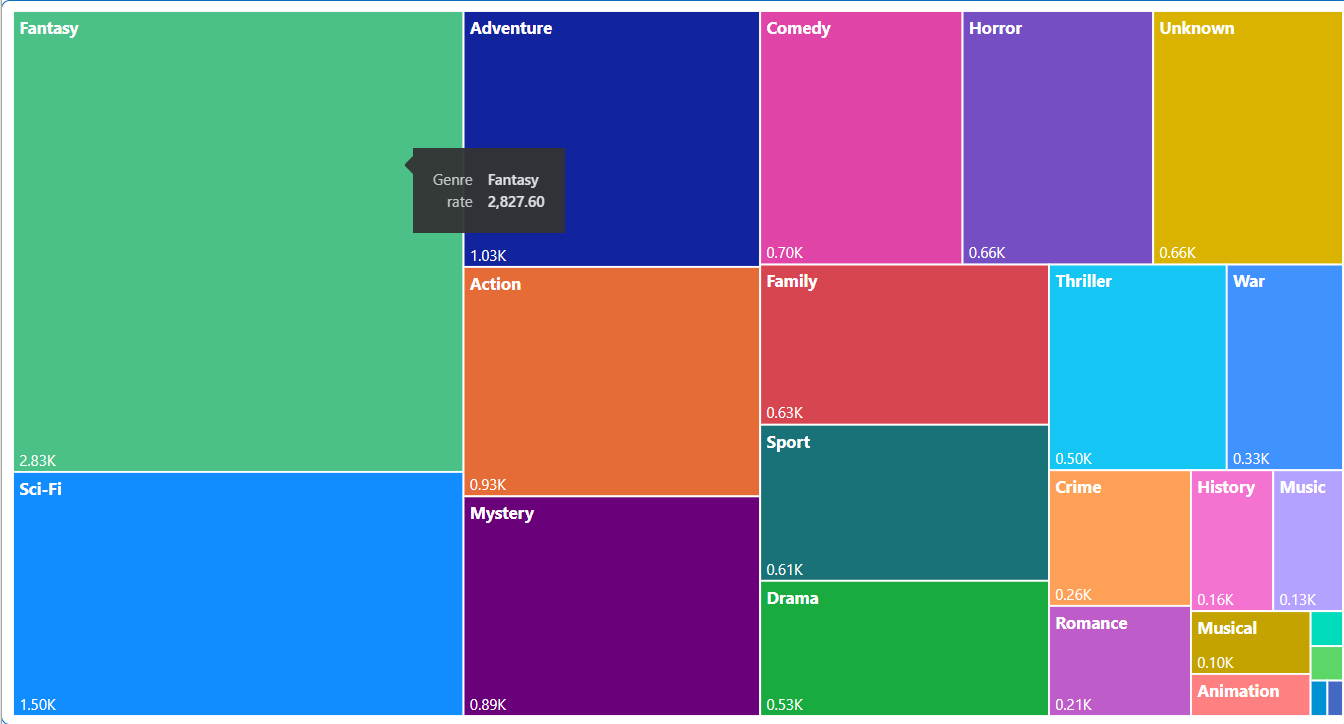
Sau khi chọn mốc thời gian cụ thể để phân tích, nhóm tiến hành phân tích thể loại game đã có doanh thu cao và lý do vì sao thể loại game này lại có mức doanh thu cao đến vậy.



*Char: Sales by Genre*

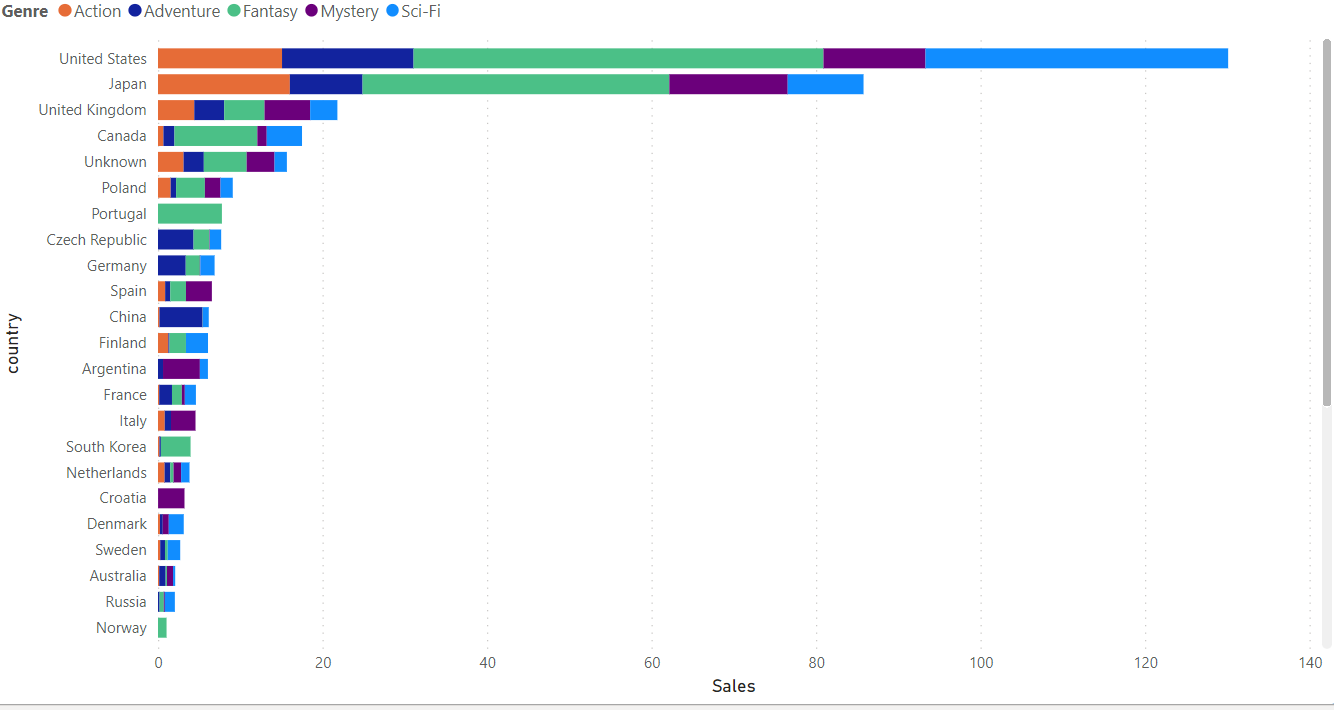
Trong giai đoạn 2011 – 2017, nhìn vào biểu đồ cột trên, có thể dễ dàng nhận thấy thể loại game “Fantasy” có doanh thu vô cùng cao, gần như gấp nhiều lần so với những loại game còn lại.

Để lý giải vì sao trong giai đoạn 2011-2017 loại game “Fantasy” có lượng doanh thu khủng đến vậy, nhóm đã tiến hành phân tích dự trên lượt đánh giá.



*Char: Rate by Genre*

Ngoài ra, xem xét trên khía cạnh quốc gia để nhận thấy thể loại game nào mang lại doanh thu cao thì nhóm tiến hành phân tích doanh thu theo quốc giá.



*Char: Sales by Contry and Genre*

Có thể thấy, đối với hầu hết các quốc gia, đặc biệt là những quốc gia lớn và đi đầu trong lĩnh game như “United States”, Japan,… thì “Fantasy” vẫn có mức doanh thu cao nhất”.

Từ đó, có thể dễ dàng nhận thấy loại game “Fantasy” có tổng lượt đánh giá cao gấp đôi, chính vì được yêu thích nhiều nên thể loại game này được ưa chuộng và mang lại doanh thu khủng. Có thể thấy đây là thể loại game có nhiều khả năng sẽ là xu hướng game của những năm sau.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Thông qua những kiến thức cơ bản về thu thập, xử lý và phân tích dữ liệu nhóm đã thực hiện bài báo cáo sơ bộ nhằm phân tích doanh thu của Game qua các năm nhằm tiềm ra đầu là loại game đáng được đầu tư nhất.

Thông qua quá trình báo cáo nhóm đã thực hiện được:

-Tìm hiểu thực tế về data engineer

-Biết cách thu thập dữ liệu từ website

-Sử dụng các công cụ điện toán đám mây nhằm lưu trữ xử lý và phân tích dữ liệu

-Ứng dụng story telling ở mức cơ bản để phân tích dữ liệu nhằm trả lời yêu cầu của bài toán đặt ra

Do thời gian thực tập còn ngắn và kiến thức có hạn, nếu có thời gian nghiên cứ và tìm hiểu thêm nhóm sẽ thực hiện cải thiện một số điểm còn hạn chế trong báo cáo:

-Tạo các pipeline tự động nhằm sao chép dữ liệu và đưa dữ liệu đến sql database cho các lần sau

-Ứng dụng machine learning nhằm thực hiện các phân tích nâng cao và dự đoán xu hướng với dữ liệu

-Sử dụng các kỹ thuật nâng cao trong thu thập và xử lý nhằm nâng cao hiệu suất.

# 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