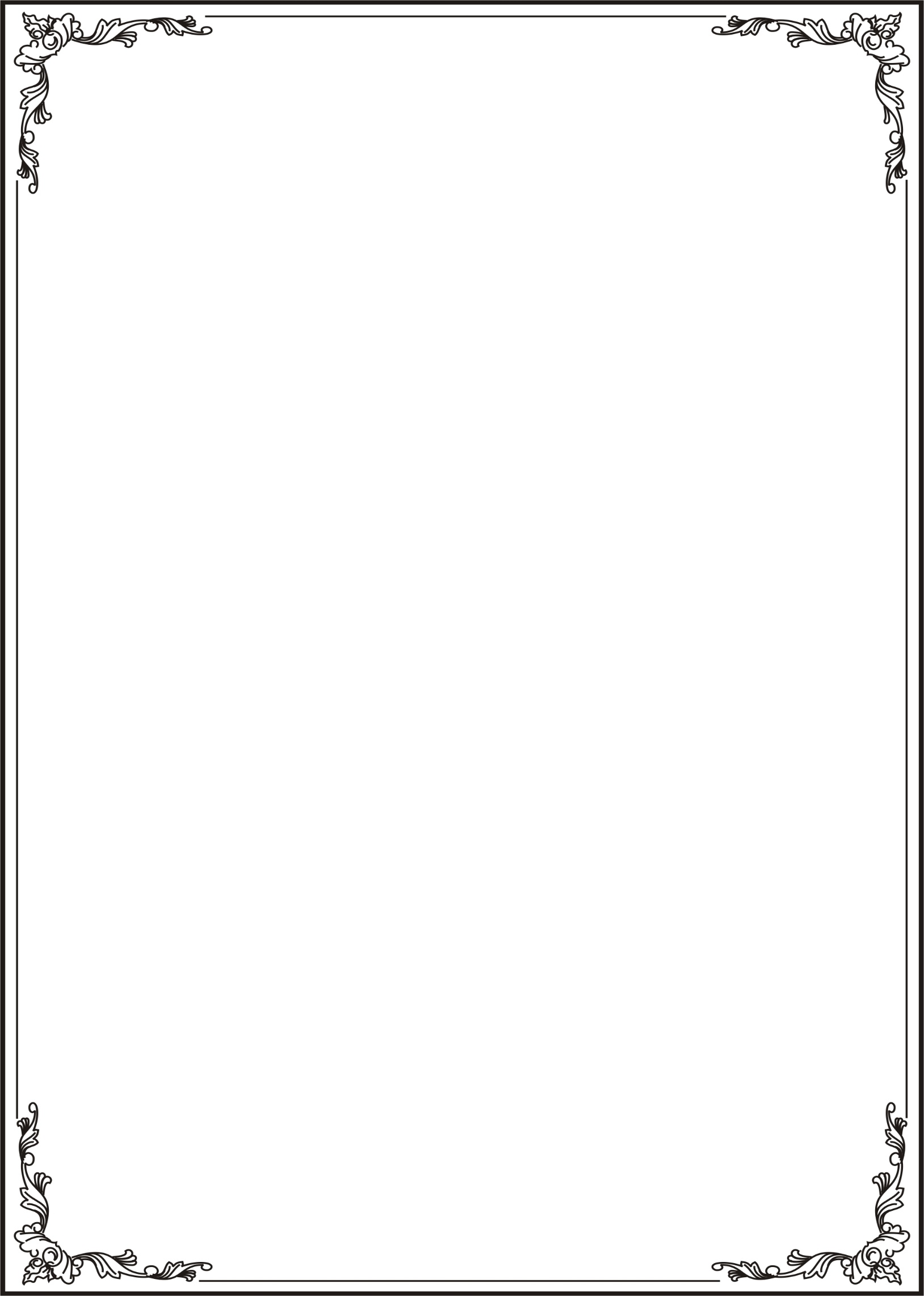
**ĐẠI HỌC QUỐC GIA TP HCM**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

**** □□



**ĐỒ ÁN**

Môn học: ĐIỆN TOÁN ĐÁM MÂY

Đề tài: Azure Synapse E2E Analytics

Giảng viên hướng dẫn: **Hà Lê Hoài Trung**

Lớp: Điện toán đám mây – IS402.N11.HTCL

Thành viên:

|  |  |  |
| --- | --- | --- |
| Lê Ngọc Tuấn | - | 19522466 |
| Lê Ngọc Minh Thư | - | 19522305 |
| Trần Nhật Khuê | - | 19521713 |
| Lê Văn Long | - | 19521783 |

**TP Hồ Chí Minh tháng 12 năm 2022**

**MỤC LỤC**

[**LỜI MỞ ĐẦU** 1](#_Toc122027980)

[**LỜI CẢM ƠN** 2](#_Toc122027981)

[**NHẬN XÉT CỦA GIÁO VIÊN** 3](#_Toc122027982)

[**CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ Azure Synapse Analytics** 4](#_Toc122027983)

[1. TỔNG QUAN 4](#_Toc122027984)

[1.1.Chức năng của Azure Synapse Analytics 4](#_Toc122027985)

[1.2. Hệ sinh thái của Azure Synapse Analytics 5](#_Toc122027986)

[1.2.Kiến trúc hệ thống của Azure Synapse Analytics 7](#_Toc122027987)

[1.4. Ưu và nhược điểm của Azure Synapse Analytics 9](#_Toc122027988)

[1.5. So sánh giữa Azure Synapse Analytics và Azure Data Factory 9](#_Toc122027989)

[**CHƯƠNG 2. CASE STUDY VÀ HƯỚNG GIẢI QUYẾT BÀI TOÁN** 11](#_Toc122027990)

[1. GIỚI THIỆU VỀ CASESTUDY 11](#_Toc122027991)

[2. HƯỚNG GIẢI QUYẾT BÀI TOÁN 12](#_Toc122027992)

[3. CÔNG NGHỆ SỬ DỤNG ĐỂ GIẢI QUYẾT BÀI TOÁN 12](#_Toc122027993)

[3.1 Thuận toán Random Forest 12](#_Toc122027994)

[3.2 Cloud Computing và End to End project 13](#_Toc122027995)

[3.3 Azure Synapse 13](#_Toc122027996)

[3.4 Power BI 13](#_Toc122027997)

[**CHƯƠNG 3. GIẢI QUYẾT BÀI TOÁN** 15](#_Toc122027998)

[1. MÔ TẢ VỀ DỮ LIỆU SỬ DỤNG 15](#_Toc122027999)

[2. TIỀN XỬ LÝ DỮ LIỆU 16](#_Toc122028000)

[3. THỰC HIỆN TRÊN AZURE SYNAPSE 18](#_Toc122028001)

[3. THỰC HIỆN TRÊN POWER BI 53](#_Toc122028002)

[**CHƯƠNG 4. TÀI LIỆU THAM KHẢO** 54](#_Toc122028003)

# **LỜI MỞ ĐẦU**

Cùng với xu thế toàn cầu hóa của việc sử dụng dữ liệu dựa trên đám mây, cung cấp nhiều giải pháp để đồng bộ hóa và xử lí dữ liệu lớn dựa trên các nền tảng khác nhau của cloud computing như Microsoft Azure, Amazon Web Services hay gọi tắc là AWS với trụ sở chính được đặt tại Seattle, Cung cấp hơn 90+ các loại sản phẩm và dịch vụ khác nhau trải dài các lĩnh vực như máy tính, cơ sở dữ liệu, phân tích, triển khai hoặc các công cụ cho nhà phát triển,… Từ đó ngày càng mở rộng khả năng tính toán, gián tiếp hỗ trợ các nghiên cứu đòi hỏi sức mạnh tính toán lớn.

Vậy nên, việc nghiên cứu và tìm hiểu những dịch của cloud computing là điều cần thiết để có thể đáp ứng các nhu cầu kịp thời của danh nghiệp, đưa ra các chiến lược phát triển đúng đắn dựa vào dữ liệu thu thập và phân tích được, và dựa vào những hiểu biết đó, nhóm chúng em lựa chọn đề tài Azure Synapse End to End Analytics để bước đầu tìm hiểu sâu về hệ thống điện toán đám mây.

# **LỜI CẢM ƠN**

Để hoàn thành bài báo cáo đồ án này, em xin tỏ lòng biết ơn sâu sắc đến giảng viên thầy Hà Lê Hoài Trung (Giảng viên môn Điện toán đám mây) đã tận tình giúp cho nhóm tích lũy, củng cố và có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này. Thầy đã trực tiếp hướng dẫn tận tình, chỉnh sửa và đóng góp nhiều ý kiến quý báu trong suốt quá trình của học của môn học này giúp nhóm hoàn thành tốt báo cáo môn học này của mình. Trong thời gian một học kỳ thực hiện đề tài, nhóm đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới, cũng như hành trang quý báu để nhóm em bước vào đời một cách vững trắc và tự tin. Từ đó, nhóm vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất.

Tuy nhiên, trong quá trình thực hiện, nhóm không tránh khỏi những thiếu sót. Chính vì vậy, nhóm rất mong nhận được những sự góp ý từ phía các thầy cô nhằm hoàn thiện những kiến thức mà nhóm đã học tập và là hành trang để nhóm thực hiện tiếp các đề tài khác trong tương lai. Xin chân thành cảm ơn Thầy!

# **NHẬN XÉT CỦA GIÁO VIÊN**

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

……………………………………………………………………………………………………………………………

# **CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ AZURE SYNAPSE ANALYTICS**

## TỔNG QUAN

* **Azure Synapse** là một Morden Data Warehouse chạy trực tuyến trên Microsoft Azure, nó cho phép các bạn lấy dữ liệu từ nhiều nguồn như các app mà các bạn làm, các Database như SQL Server, … đổ Data vào Azure Synapse, và đầu ra kết nối với các online service như Power BI Service để làm Phân tích dữ liệu, hoặc training data cho các model AI / Machine Learning.
* **Azure Synapse** còn thể gọi là một dịch vụ phân tích giúp tăng tốc thời gian để hiểu rõ hơn về các kho dữ liệu và hệ thống dữ liệu lớn. trước đây được gọi là Kho dữ liệu Azure, là một công cụ phân tích mới do Microsoft phát triển. Nhưng dịch vụ mới không chỉ thể hiện sự thay đổi tên mà còn là sự tiến hóa của cách thực hiện phân tích trong Azure.

Graphical user interface, application

Description automatically generated

### Chức năng của Azure Synapse Analytics

* Các chức năng mà nó cung cấp bao gồm:
* Kho dữ liệu: Công nghệ Kho dữ liệu Azure đã phổ biến để lưu trữ và quản lý dữ liệu để phân tích và ra quyết định, hiện thông qua các nhóm SQL.
* Công cụ Dữ liệu lớn: Với các nhóm Spark, các kỹ sư hiện có thể chạy các phân tích có thể mở rộng với ngôn ngữ Spark để xử lý Dữ liệu lớn với chúng.
* Luồng dữ liệu: Để phát triển luồng ETL sử dụng hoặc nhận dữ liệu trong Kho dữ liệu hoặc Hồ dữ liệu của bạn với cùng một công cụ được sử dụng với Nhà máy dữ liệu Azure.
* PowerBI: Kết nối dữ liệu của bạn trực tiếp với bộ dữ liệu PowerBI để hiển thị và phát triển các báo cáo và trang tổng quan của bạn một cách tối ưu và đơn giản.
* Nhà máy dữ liệu: Tận dụng các luồng Dữ liệu phân tích Synapse với các lợi thế của Nhà máy dữ liệu Azure cho các quy trình ETL mạnh mẽ nhất của bạn.
* Máy học: Tận dụng các quy trình Spark của bạn để phát triển các mô hình và hiển thị chúng thông qua Azure Machine Learning để các quy trình phân tích của bạn được kết nối với nhau, có thể mở rộng và khả dụng cao.

### 1.2. Hệ sinh thái của Azure Synapse Analytics

Hệ sinh thái của Azure Synapse Analytics có các thành phần như sau:

Diagram

Description automatically generated

* **Nguồn dữ liệu – Data Source**:

Đây là dạng dữ liệu trong kinh doanh như hệ thống ERP, hệ thống CRM…

Các nguồn dữ liệ thường có thể được định dạng với nhiều loại khác nhau chẳng hạn như: On - premises và Cloud, Cấu trúc, bán cấu trúc, Dữ liệu phi cấu trúc và Cơ sở dữ liệu, Tập tin và API

* **Nhập và điều phối dữ liệu - Ingestion**

Khi ta đã xác định được dạng dữ liệu kinh doanh mà ta muốn để dùng cho báo cáo và phân tích thì ta thường nhập dữ liệu đó vào một kho dữ liệu duy nhất và kết hợp dữ liệu đó từ nhiều nguồn khác nhau để có được dữ liệu hoàn chỉnh, cụ thể của khách hàng, sản phẩm cũng như quy trình kinh doanh của ta.

Các công cụ Azure ta thường dùng để thực hiện là:

* Azure Data Factory
* Azure Stream Analytics
* **Lưu trữ dữ liệu - Storage**

Kho dữ liệu trung tâm có thể được tạo ra và duy trì bằng nhiều phương pháp và định dạng khác nhau. Khi ta muốn truy cập nhanh vào dữ liệu theo real – time (thời gian thực) thì phương pháp Data Lake là phù hợp, trong khi kho dữ liệu được quản lý và mô hình hóa phù hợp với thông tin chi tiết hoàn chỉnh và nhiều trường hợp sử dụng như báo cáo, thông tin chi tiết. Các công cụ của Azure giúp ta thực hiện việc này:

* Azure Blob Storage
* Azure Data Lake
* Azure SQL Database
* **Phân tích dữ liệu - Analytics**

Khi dữ liệu của ta được nhập vào kho lưu trữ trung tâm thì được chuẩn bị cho báo cáo và phân tích. Ta có thể tận dụng các công cụ mạnh mẽ hiện giờ để có thể có được thông tin chi tiết, cụ thể về khách hàng, sản phẩm hoặc là quy trình kinh doanh của ta. Ngoài ra, ta cũng có thể dự đoán kết quả để doanh nghiệp của ta có thể phản ứng nhanh với những khó khăn, trục trặc xảy ra.

Các công cụ Azure giúp ta thực hiện:

* Azure Synapse
* Azure Databricks
* Power BI
* Azure Machine Learning

Một trong những công cụ này có vai trò quan trọng trọng trong kiến trúc dữ liệu tổng thể của ta. Khi tổ chức của ta phát triển và lớn mạnh trong Data và Analytics.

### Kiến trúc hệ thống của Azure Synapse Analytics

Kiến trúc Azure Synapse bao gồm 4 thành phần:

**Synapse SQL**: Hoàn thành phân tích dựa trên T – SQL.

* Nhóm SQL chuyên dụng (trả cho mỗi DWU được cung cấp).
* Nhóm SQL không có máy chủ (trả cho mỗi TB được xử lý).

**Spark**: Tích hợp sâu Apache Spark.

**Synapse Pipelines**: Tích hợp dữ liệu kết hợp.

**Studio**: Trải nghiệm người dùng hợp nhất

**- Synapse SQL**: Đây là khả năng thực hiện phân tích dựa tren T – SQL trong không gian làm việc của Synapse. Nó bao gồm hai mô hình tiêu thụ: chuyên dụng và phi máy chủ.

* Các nhóm SQL chuyên dụng được sử dụng cho các mô hình chuyên dụng và không gian làm việc có thể có bất kỳ số lượng pool nào trong số này.
* Nhóm SQL phi máy chủ được sử dụng cho các mô hình phi máy chủ và mọi không gian làm việc đều có một những pool này.

- **Spark**: Các nhóm Apache Spark phi máy chủ được tạo và sử dụng trong không gian làm việc Synapse để sử dụng phân tích Spark. Nó bao gồm các thành phần:

* Apache Spark cho Synapse.
* Apache Spark pool
* Ứng dụng Spark
* Spark session
* Định nghĩa Spark job
* **Synapse Pipelines**: Có các chức năng sau:
* Data Integration
* Data Flow
* Pipeline
* Activity
* Trigger
* Integration dataset

- **Synapse Studio/Workspace**: Đây là ranh giới cộng tác có thể bảo mật để thực hiện phân tích doanh nghiệp dựa trên Cloud trong Azure, được triển khai ở một khu vực cụ thể và cũng có hệ thống tệp cũng như tài khoản ADLS Gen2 liên quan đến lưu trữ dữ liệu tạm thời.

### 1.4. Ưu và nhược điểm của Azure Synapse Analytics

* **Azure Synapse Analytics** là một công cụ mạnh mẽ, đa chức năng trong bất kỳ môi trường quản lý dữ liệu hiện đại nào. Ưu điểm của công cụ này là tất cả trong một, nghĩa là, nó cung cấp một số cách làm việc và công nghệ trong một dịch vụ duy nhất, hợp lý hóa và thống nhất các quy trình phát triển và quản lý dữ liệu theo cách sáng tạo và vừa phải.
* **SQL Pool**
* Trong Synapse Analytics, Kho dữ liệu có thể được sử dụng thông qua các nhóm SQL, cho phép bạn truy vấn cơ sở dữ liệu thông qua các cụm có thể mở rộng cả về số lượng máy và kích thước của chúng.
* Ngoài những ưu điểm của việc mở rộng quy mô nhóm, bạn có thể chỉ định một mức xử lý và tài nguyên nhóm theo vai trò và loại truy vấn. Bạn có thể thêm bảo mật ở cấp hàng và cột để chỉ những người dùng được quy định mới có quyền truy cập nhất định.

## SO SÁNH GIỮA AZURE SYNAPSE ANALYTICS VÀ AZURE DATA FACTORY

|  |  |  |  |
| --- | --- | --- | --- |
| **Loại** | **Tính năng** | **Azure Data Factory** | **Azure Synapse Analytics** |
| **Thời gian chạy tích hợp** | Sử dụng thời gian chạy tích hợp SSIS và SSIS | Có hỗ trợ | Chỉ mới có bản xem trước |
| Hỗ trợ thời gian chạy tích hợp liên khu vực (Luồng dữ liệu) | Có hỗ trợ | Không có hỗ trợ |
| Chia sẻ thời gian chạy tích hợp | Có thể được chia sẻ trên các nhà máy dữ liệu khác nhau | Không hỗ trợ |
| **Hoạt động Pipelines** | Hoạt động đóng gói SSIS | Có hỗ trợ | Chỉ mới có bản xem trước |
| Hỗ trợ hoạt động Power Query | Có hỗ trợ | Không hỗ trợ |
| Hỗ trợ các thông số toàn cầu | Có hỗ trợ | Không hỗ trợ |
| **Thư viện mẫu và trung tâm thông tin** | Mẫu giải pháp | Thư viện mẫu Azure Data Factory | Trung tâm thông tin Synpase Workspace |
| **Tích hợp kho GIT** | Tích hợp GIT | Có hỗ trợ | Không có hỗ trợ |
| **Giám sát** | Giám sát các công việc Spark cho việc phân luồng dữ liệu | Không hỗ trợ | Hỗ trợ tận dụng các nhóm Synapse Spark |

# **CHƯƠNG 2. CASE STUDY VÀ HƯỚNG GIẢI QUYẾT BÀI TOÁN**

## 1. GIỚI THIỆU VỀ CASESTUDY

Hiện nay, cùng với sự phát triển của công nghệ thông tin, mọi thứ xung quanh cuộc sống của chúng ta đều được số hóa một cách triệt để và thành toán điện tử cũng không phải là ngoại lệ. Bên cạnh những lợi ích to lớn mà thanh toán điện tử mang lại, phương thức thanh toán mới này cũng tiểm ẩn nhiều rủi ro, trong đó rủi ro lớn nhất là việc lộ thông tin cá nhân cũng như thông tin thẻ,… Các đối tượng xấu có thể lợi dụng sơ hở của chúng ta để dùng những thông tin thanh toán này để thực hiện các giao dịch nặc danh, gây tổn thất nặng nề cho mỗi chúng ta.



Bên cạnh việc khuyến cáo người sử dụng thẻ phải cẩn thận với việc thanh toán, ngân hàng cung cấp thẻ cũng phải chịu trách nhiệm giải bài toán giao dịch mạo danh này, vậy họ có thê làm gì để giảm thiểu vấn nạn đang nhức nhối hiện nay?

## 2. HƯỚNG GIẢI QUYẾT BÀI TOÁN

Hầu hết người sử dụng các dịch vụ thanh toán điện tử đều có một thói quen ổn định, họ thường chi trả cho 1 số nhu cầu nhất định. Nhu cầu này thay dổi tùy theo độ tuổi, giới tính,… Thời gian thực hiện giao dịch cũng là một thói quen đáng để lưu tâm vì đa số chúng ta chỉ thực hiện các giao dịch vào các khoảng thời gian nhất định trong ngày. Sự bất thường trong thay đổi thói quen giao dịch là yếu tố quan trọng nhất để xác định xem giao dịch đó có phải là mạo danh hay không.

Nắm bắt được các yếu tố trên, bộ phận Data analyst và Data engineer của ngân hàng có thể dễ dàng xây dựng các mô hình Machine Learning để dự đoán xem giao dịch đang được thực hiện có phải là mạo danh hay không, đồng thời đưa ra giải pháp để hạn chế, ngăn chặn những giao dịch đó.

Một ngân hàng thường có rất nhiều Data analyst và Data engineer, họ không thể gặp mặt 24/24 để giải quyết các vấn đề như trên, việc gửi file qua lại trong suốt quá trình làm việc cũng gây ra bất tiện vô cùng lớn. Chính vì vậy, ngân hàng cần sử dụng các giải pháp Cloud Computing- Điện toán đám mây để giải quyết bài toán với hiệu suất tốt nhất.

## 3. CÔNG NGHỆ SỬ DỤNG ĐỂ GIẢI QUYẾT BÀI TOÁN

### 3.1 Thuận toán Random Forest

Random Forest là một thuật toán học có giám sát. Như tên gọi của nó, Rừng ngẫu nhiên sử dụng các cây (tree) để làm nền tảng.

Random Forest là một tập hợp của các Decision Tree, mà mỗi cây được chọn theo một thuật toán dựa vào ngẫu nhiên. Các điểm mạnh của Random Forest:

* Random Forest algorithm có thể sử dụng cho cả bài toán Classification và Regression
* Random Forest làm việc được với dữ liệu thiếu giá trị
* Khi Forest có nhiều cây hơn, chúng ta có thể tránh được việc Overfitting với tập dữ liệu
* Có thể tạo mô hình cho các giá trị phân loại

Dựa vào các đặc điểm trên, có thể kết luận Random Forest là một thuật toán tuyệt vời để dự đoán xem giao dịch đang thực hiện có phải là giao dịch mạo danh hay không. Bằng cách chọn các thành phần predictor là các cột giá trị nói lên thói quen giao dịch của người dùng, ta có thể đưa ra khẳng định xem giao dịch đó có phải do chính chủ thực hiện hay không với độ chính xác rất cao.

### 3.2 Cloud Computing và End to End project

Như đã nói ở phần trên, việc sử dụng Cloud Computing là vô cùng tiện lợi cho quá trình giải quyết bài toán này. Một trong những cách tối ưu nhất mà ngân hàng có thể lựa chọn, đó là dùng Cloud Computing để xây dựng một dự án End to End với phần đầu được triển khai bởi Data Engineer và phần cuối được đảm nhiệm bởi Data Analyst.

### 3.3 Azure Synapse

Azure Synapse, với những tính năng vượt trội đã được đề cập ở phần đầu của đồ án, là lựa chọn hàng đầu trong việc triển khai một dự án phân tích dữ liệu end to end.

Các Data engineer có thể thực hiện việc xử lí phân tích dữ liệu có trong data lake (dữ liệu raw được lưu trữ dưới nhiều định dạng khách nhau), qua đó chọn ra dữ liệu có ý nghĩa để tiến hành áp dụng các thuật toán Machine Learning

Một điểm đáng chú ý là Azure Synapse chỉ hỗ trợ mô hình Machine Learning đã được train sẵn và đã được chuyển đổi về ONNX (Open Neural Network Exchange).

ONNX một công cụ đóng vai trò như một trung gian hỗ trợ chuyển đổi mô hình học máy từ các framework khác nhau về dạng ONNX cung cấp nhờ đó giúp chúng ta chuyển đổi dễ dàng giữa các framework khác nhau. ONNX hỗ trợ chuyển đổi giữa nhiều framework phổ biến hiện nay như Keras, Tensorfow, Scikit-learn, Pytorch và XGBoost.

Mục tiêu của việc sử dụng Azure Synapse là để đưa ra các dự đoán bằng mô hình Machine Learning (Random Forest) và đưa ra tập dữ liệu để visualize bằng PowerBI.

### 3.4 Power BI

Power BI là một thành phần trong Power Platform của Microsoft. Ngoài Power BI, Microsoft Power Platform còn bao gồm Power Apps, Power Pages, Power Automate và Power Virtual Agents, là bộ các công cụ hỗ trợ cho phân tích dữ liệu, xây dựng giải pháp và tự động hóa quy trình, giúp các doanh nghiệp thúc đẩy kinh doanh.

Power BI đóng vai trò là một công cụ Kinh doanh thông minh (Business Intelligence) đem đến khả năng xử lý, tổ chức, phân tích và trực quan hóa dữ liệu, mang dữ liệu từ dạng thô trở thành các thông tin hữu ích hỗ trợ cho việc ra quyết định của các công ty, tổ chức, doanh nghiệp. Thể hiện của Power BI là các biểu đồ, bảng điều khiển (dashboard) không chỉ đẹp đẽ mà còn chứa đựng nhiều thông tin có giá trị giúp diễn giải, kiểm soát tình hình kinh doanh và dự báo cho tương lai.

# **CHƯƠNG 3. GIẢI QUYẾT BÀI TOÁN**

## MÔ TẢ VỀ DỮ LIỆU SỬ DỤNG

Bộ dữ liệu được thu thập trên Kaggle và mô tả về các mô phỏng giao dịch tín dụng hợp pháp và gian lận từ 01/1/2019 đến 31/12/2020.

Nó bao gồm thông tin thẻ tín dụng của 1000 khách hàng thực hiện giao dịch với nhóm 800 người.

**Kho dữ liệu**: Credit card transaction Fraud Detection Dataset.

**Link dataset**: https://www.kaggle.com/datasets/kartik2112/fraud-detection?fbclid=IwAR1hE6FXD2rkhdoO0dOnIn7P\_RjtONXMCytHD1pQClZwDp3PTSmJ93zkDsQ

**Các thuộc tính có trong kho dữ liệu:**

1. ID: mã key cho từng dòng (vd: 1,2,3,4,…)
2. trans\_date\_trans\_time: thời gian thực hiện giao dịch (vd: 2020-06-21 12:14:25)
3. cc\_num: mã thẻ của khách hàng - credit card number (vd: 2291163933867244)
4. merchant: tên thương gia (vd: fraud\_Kirlin and Sons)
5. category: loại thương gia (vd: personal\_care)
6. amt: amount – số tiền giao dịch (vd: 2.86)
7. first: tên chủ thẻ(vd: Jeff)
8. last: họ chủ thẻ (vd: Elliott)
9. gender: giới tính của chủ thẻ (vd: F-female or M-male)
10. street: địa chỉ của chủ thẻ (vd: 351 Darlene Green)
11. city: thành phố của chủ thẻ (vd: Columbia)
12. State: bang của chủ thẻ (vd: NC)
13. Zip: mã bưu điện của thẻ (vd: 29209)
14. Lat: vĩ độ của chủ thẻ (vd: 33.9659)
15. Long: kinh độ của chủ thẻ (vd: -80.9355)
16. city\_pop: dân số thành phố của chủ thẻ (vd: 333497)
17. job: nghề nghiệp của chủ thẻ (vd: Mechanical engineer)
18. Dob: ngày tháng năm sinh của chủ thẻ (vd: 1968-03-19)
19. trans\_num: mã giao dịch (vd: 2da90c7d74bd46a0caf3777415b3ebd3)
20. unix\_time : thời gian giao dịch (dạng unix) (vd: 1371816865)
21. merch\_lat: vĩ độ của thương gia (vd: 33.986391)
22. merch\_long: kinh độ của thương gia (vd: -81.200714)
23. is\_fraud: có bị lừa đảo hay không (vd: 0-no or 1: yes)

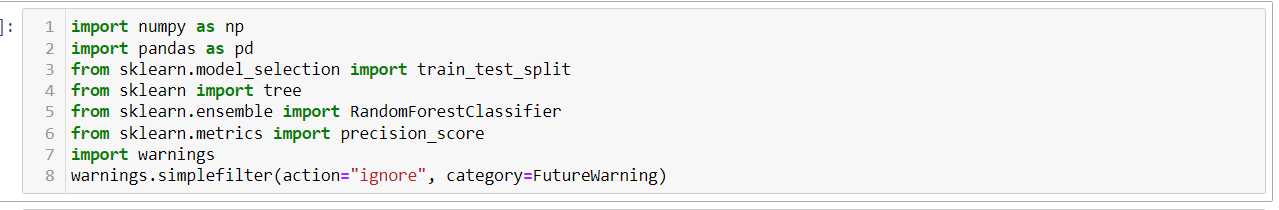
Trong 23 thuộc tính thì có 5 thuộc tính sẽ được sử dụng:

- Predict: amt, category, hour, dob.

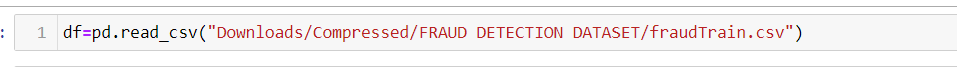
- Target: is\_fraud

## TIỀN XỬ LÝ DỮ LIỆU

Sử dụng Jupyter Notebook để loại bỏ các cột giá trị không sử dụng, chuyển các predictor về dạng nummeric như sau:

**

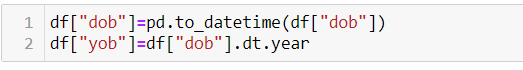
Import thư viện

**

Đọc file csv

**

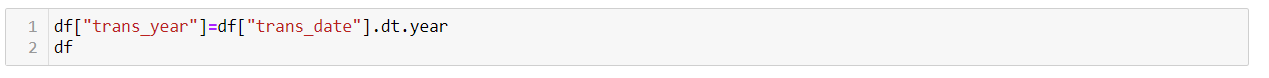
Xóa cột Unamed

**

Lấy năm sinh từ ngày sinh của người giao dịch

**

Lấy giờ giao dich trong cột ngày giờ giao dịch

**

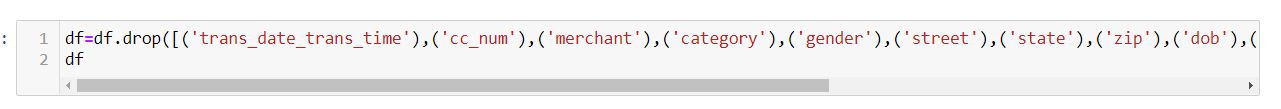
Lấy năm giao dịch để chia train set, test set

**

Chuyển cột city về dạng Nummeric

**

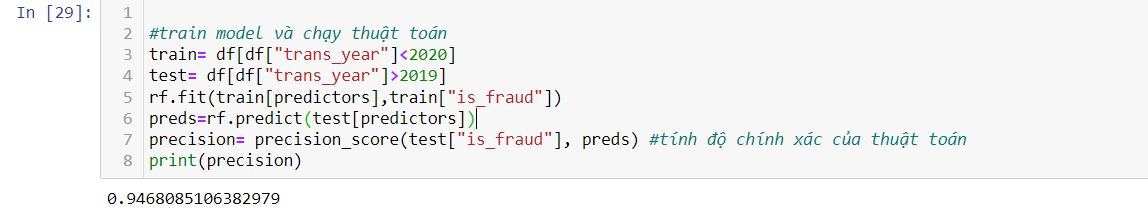
Chuyển cột category về dạng nummeric

**

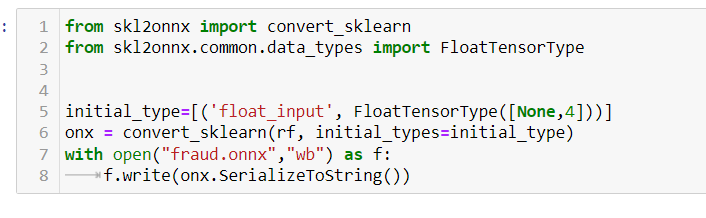
Xóa hết các cột không cần thiết

**

Chọn các thuộc tính predictor

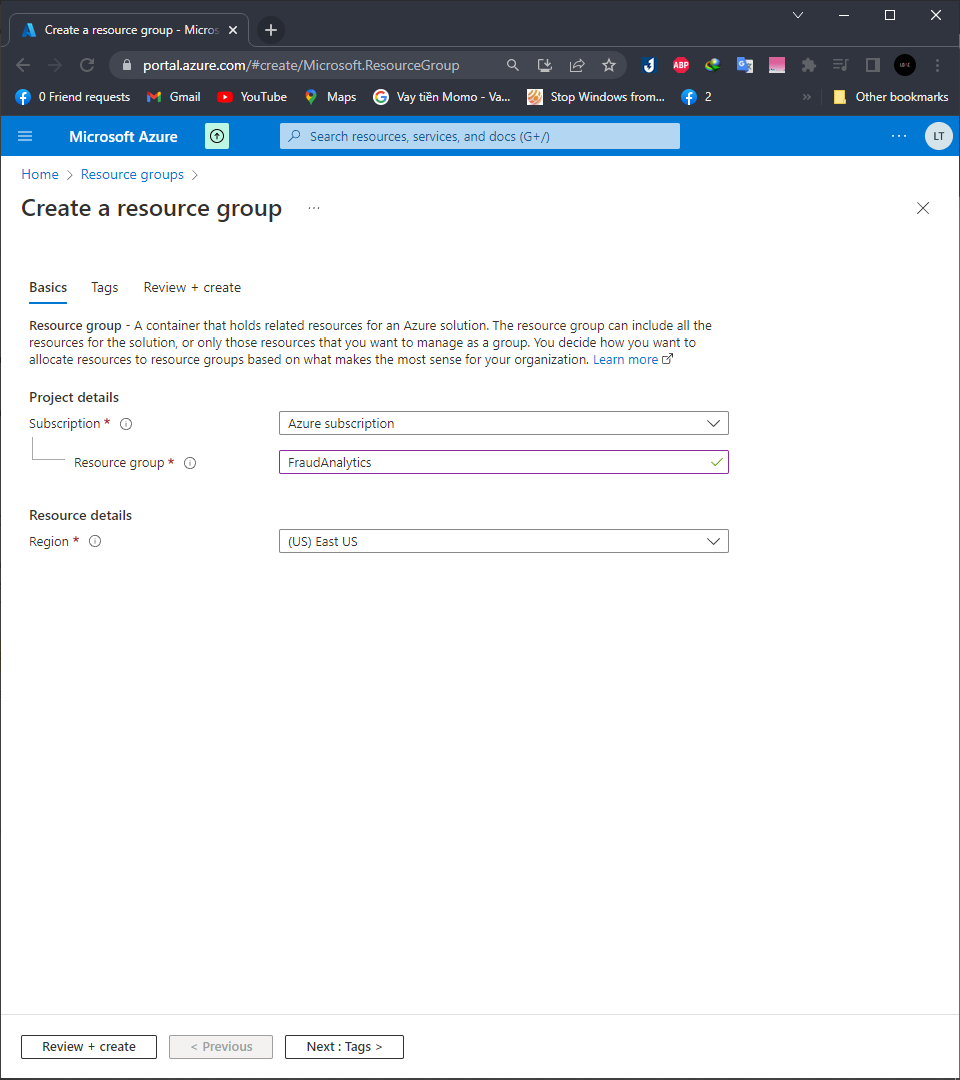
**

Train model và tính độ chính xác

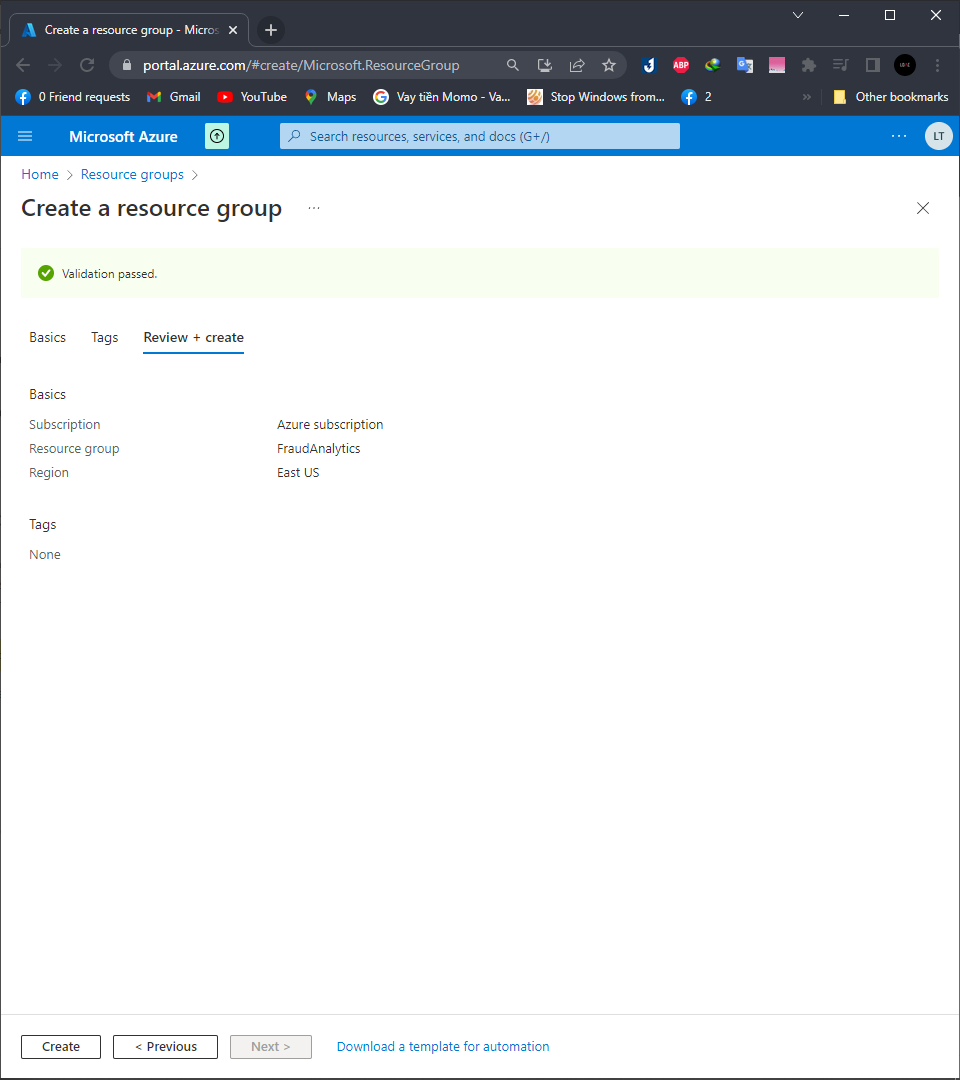
**

Chuyển model Random Forest về ONNX.

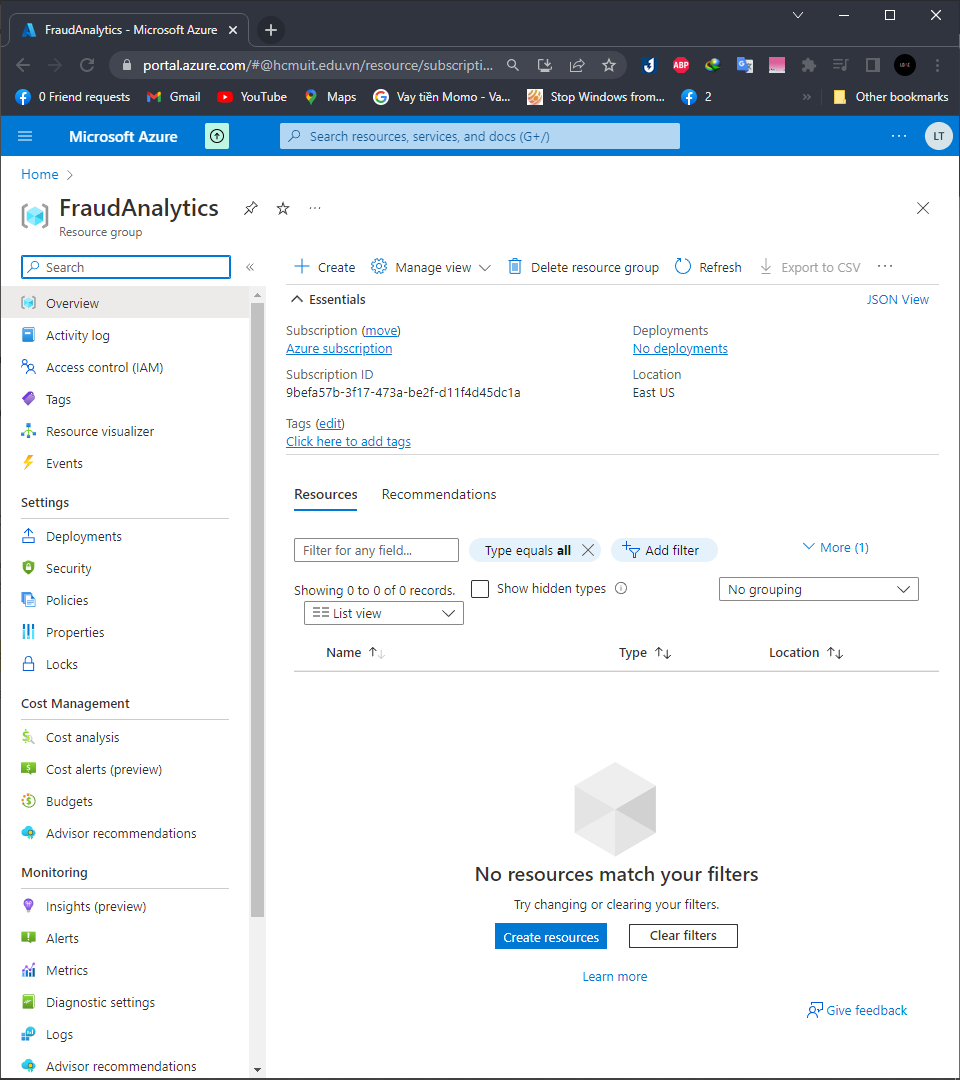
## 3. THỰC HIỆN TRÊN AZURE SYNAPSE

**

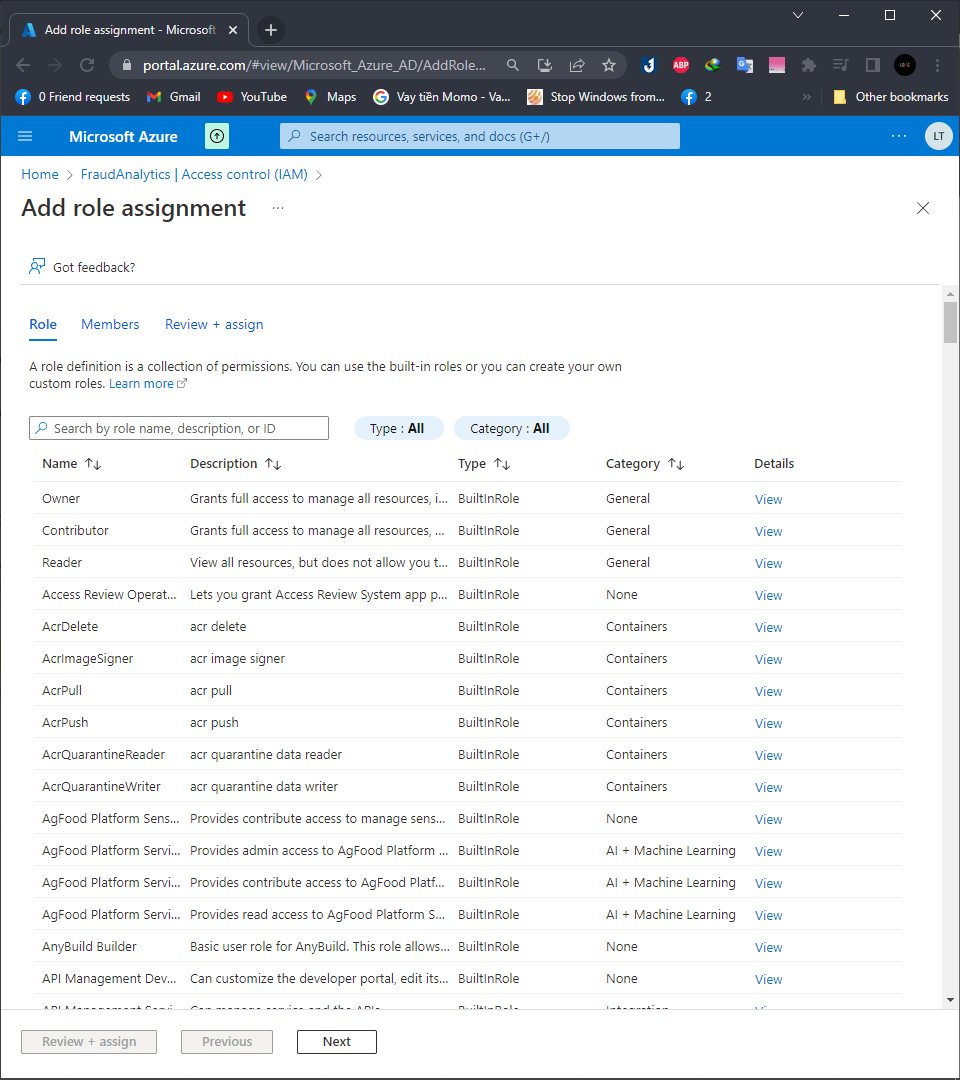
*Tạo resource group*

*\*

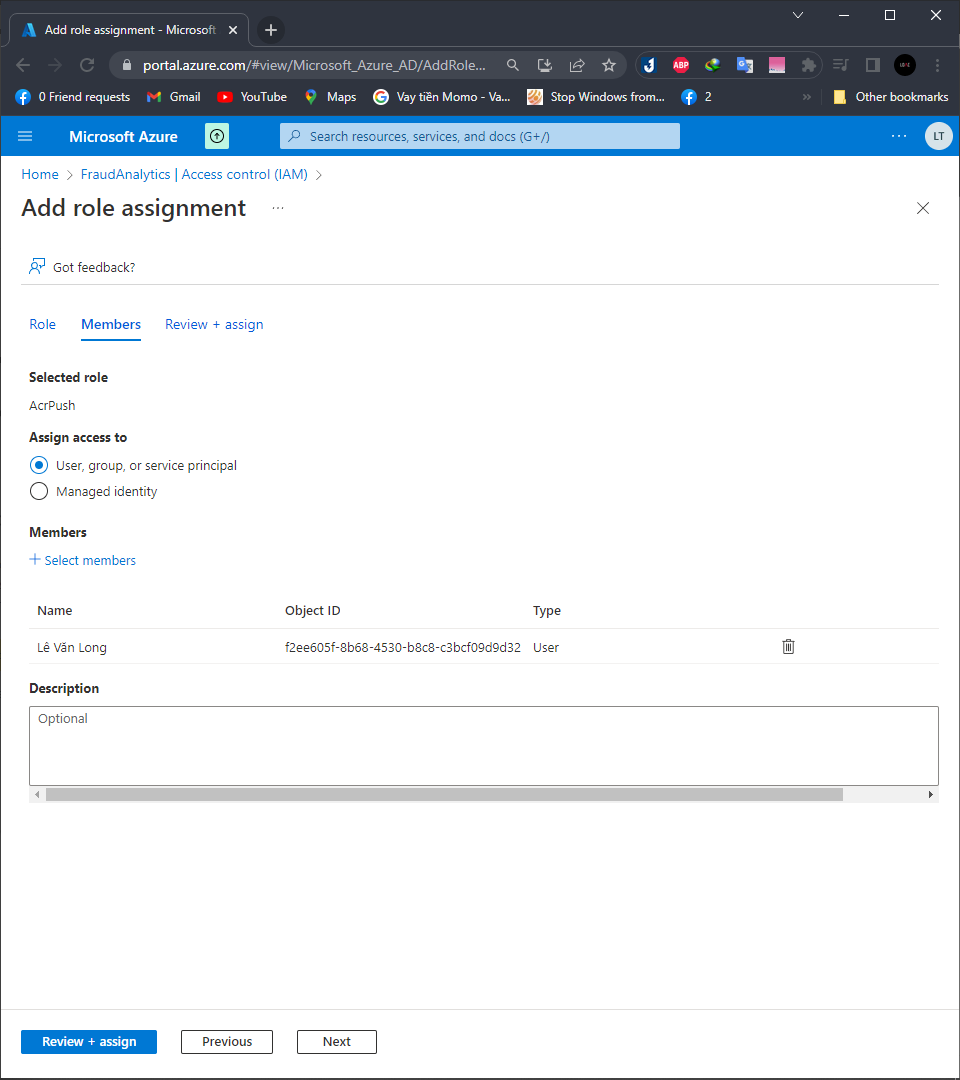
*Tạo Rerource group*

**

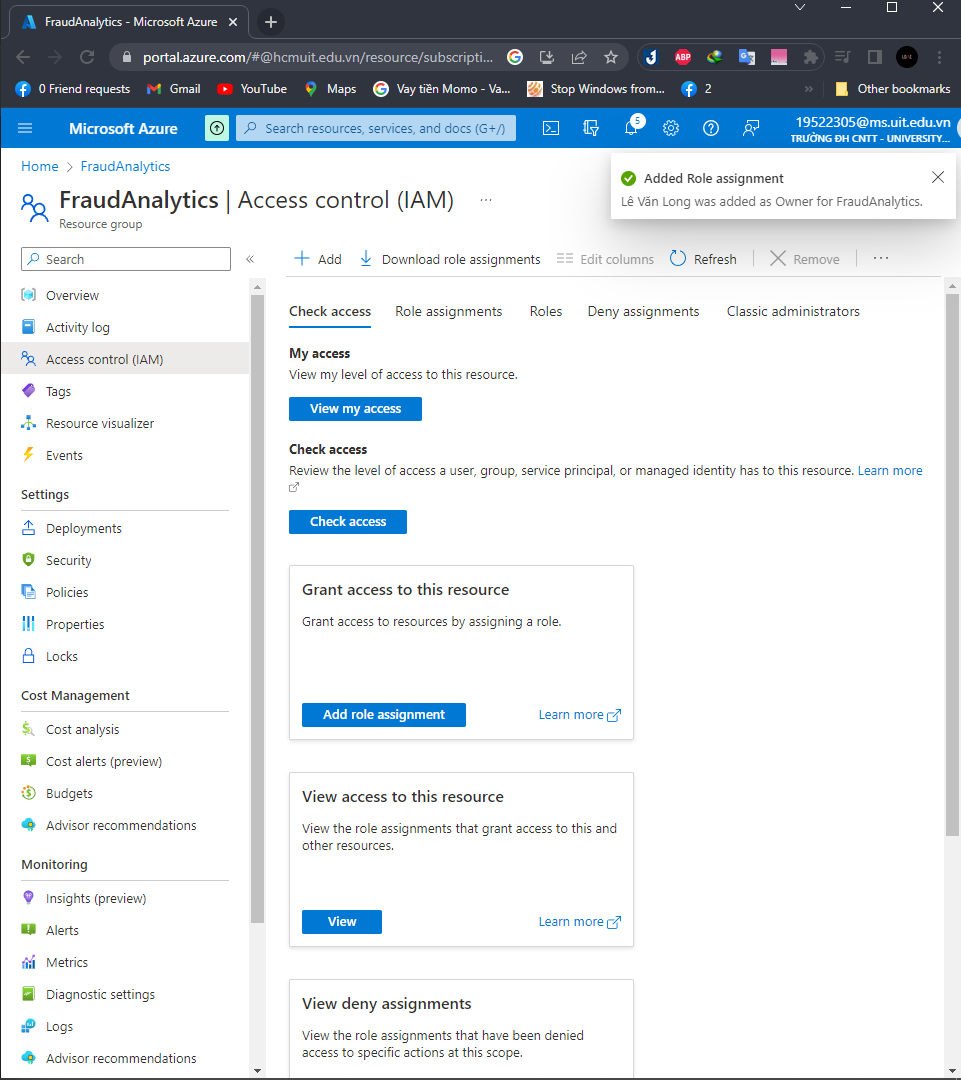
*Giao diện sau khi tạo xong resource group*

**

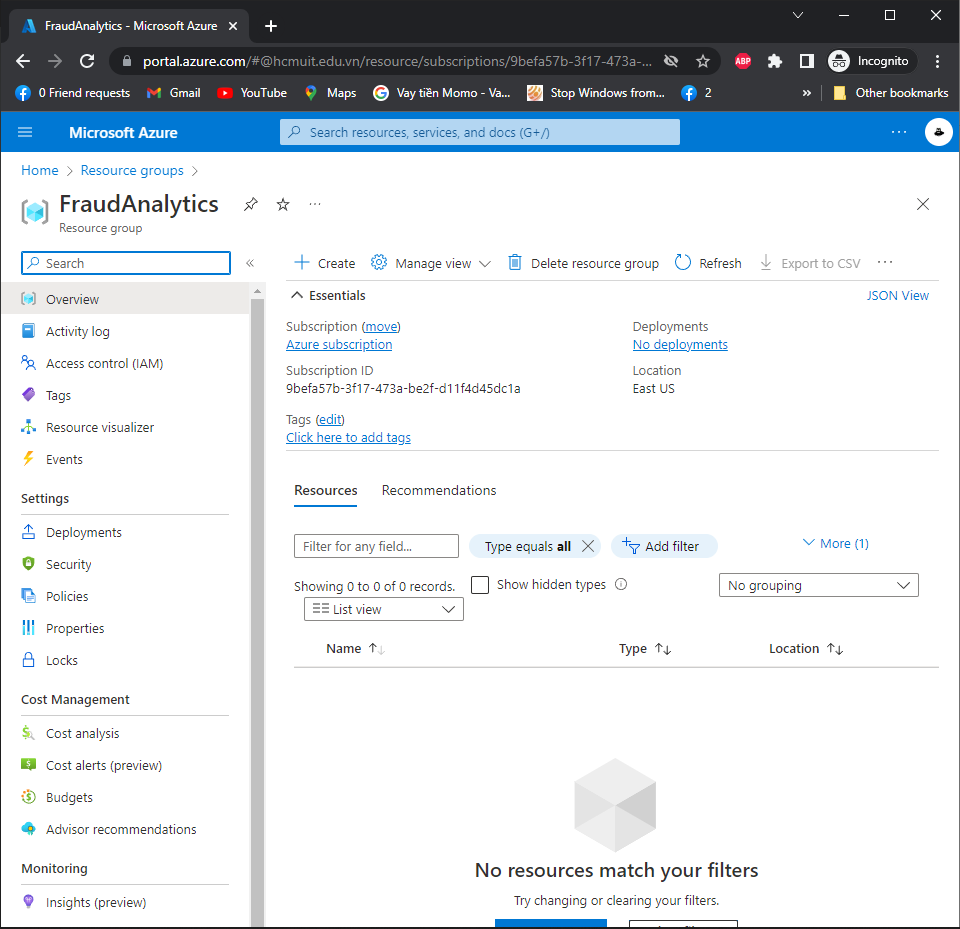
*Tạo role cho Resource group*

**

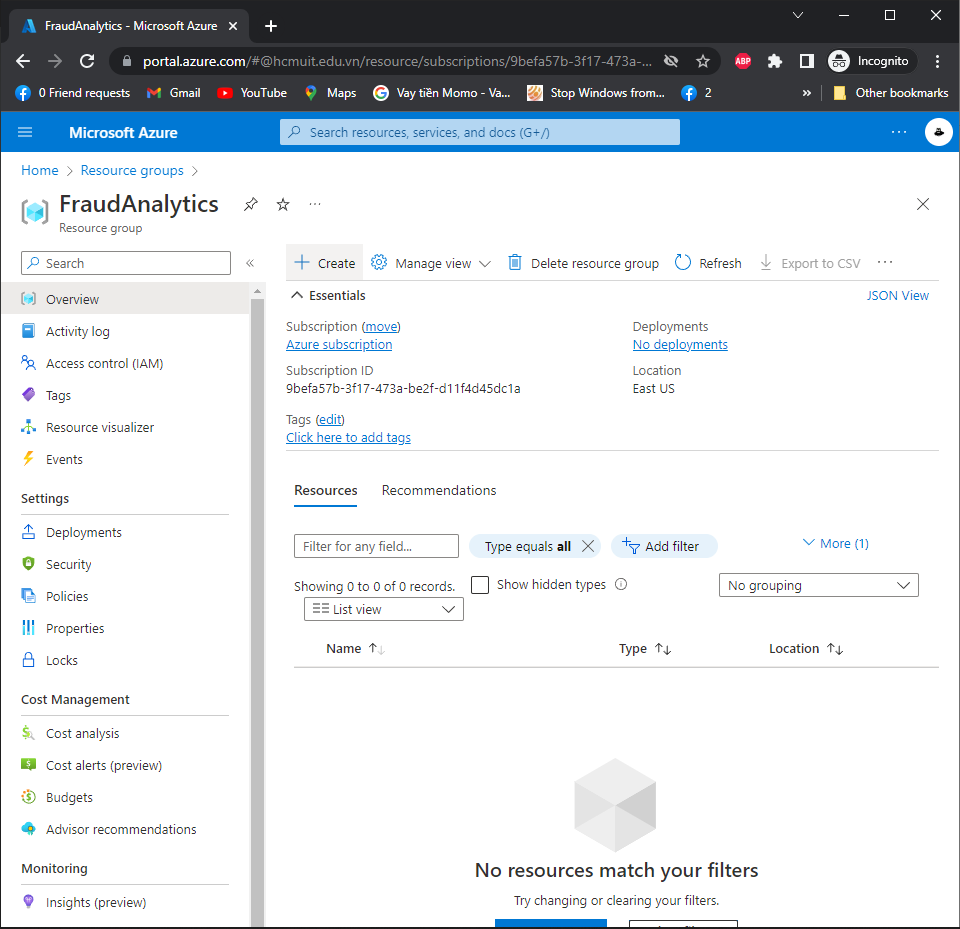
*Chọn tab member, tạo user với quyền owner*

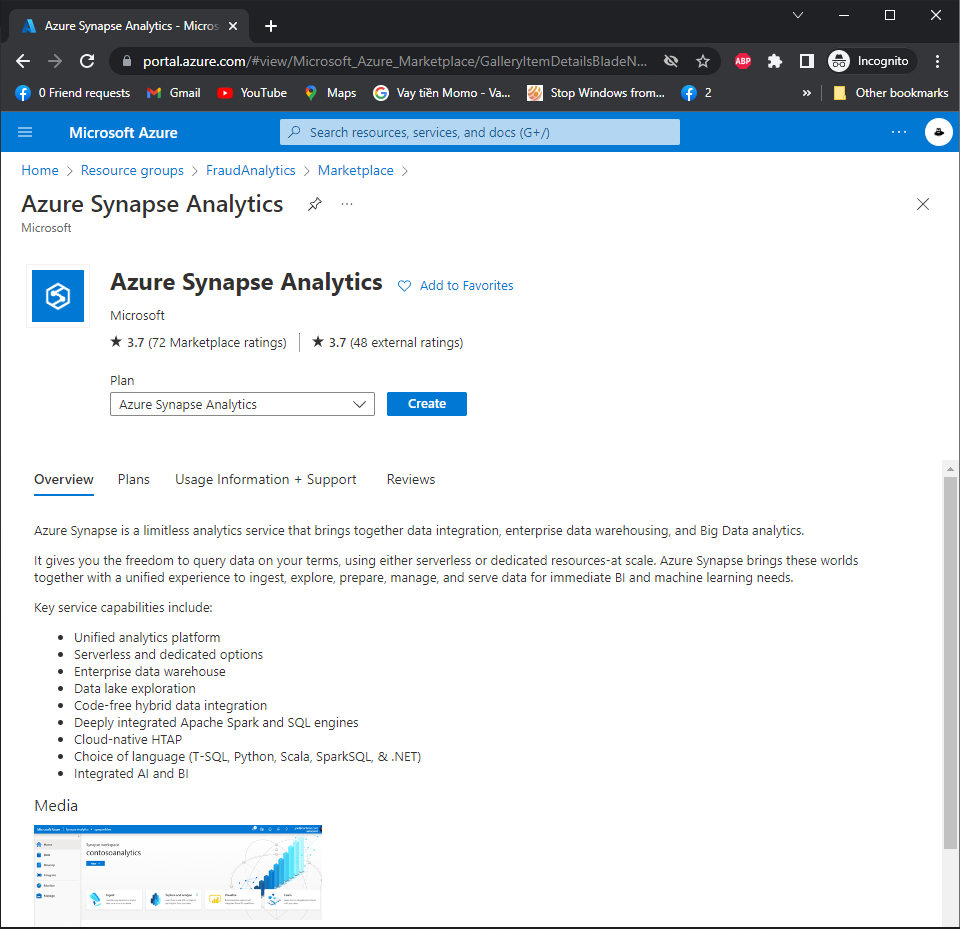
**

*Tạo user thành công*

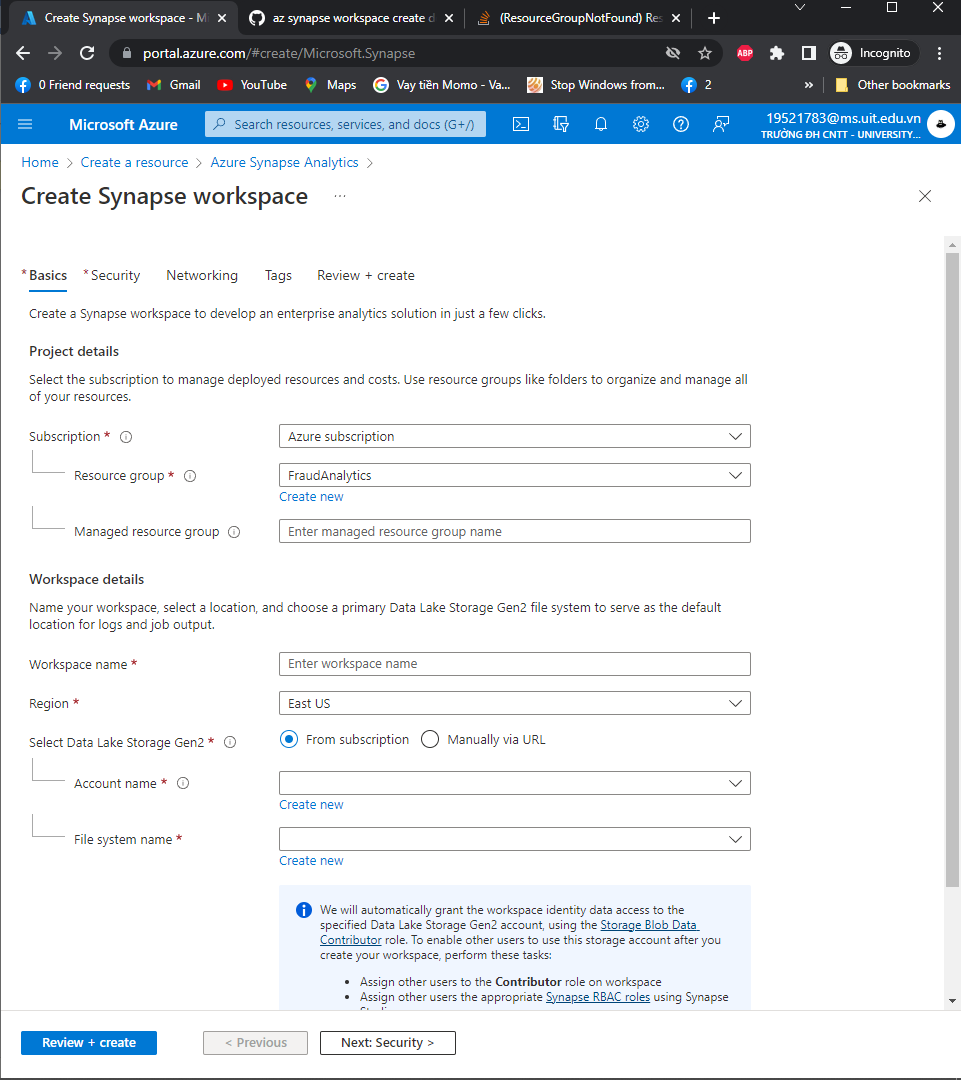
**

*Vào resource group*

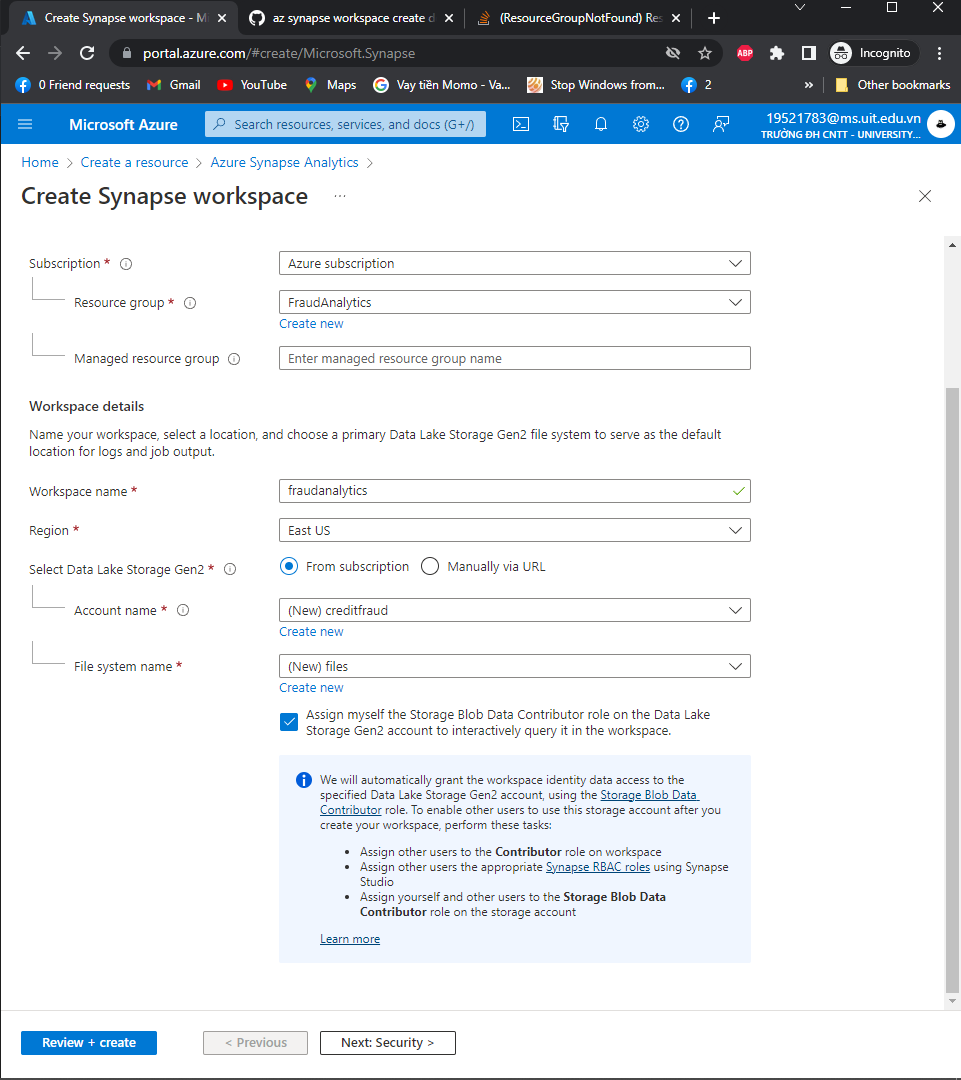
**

**

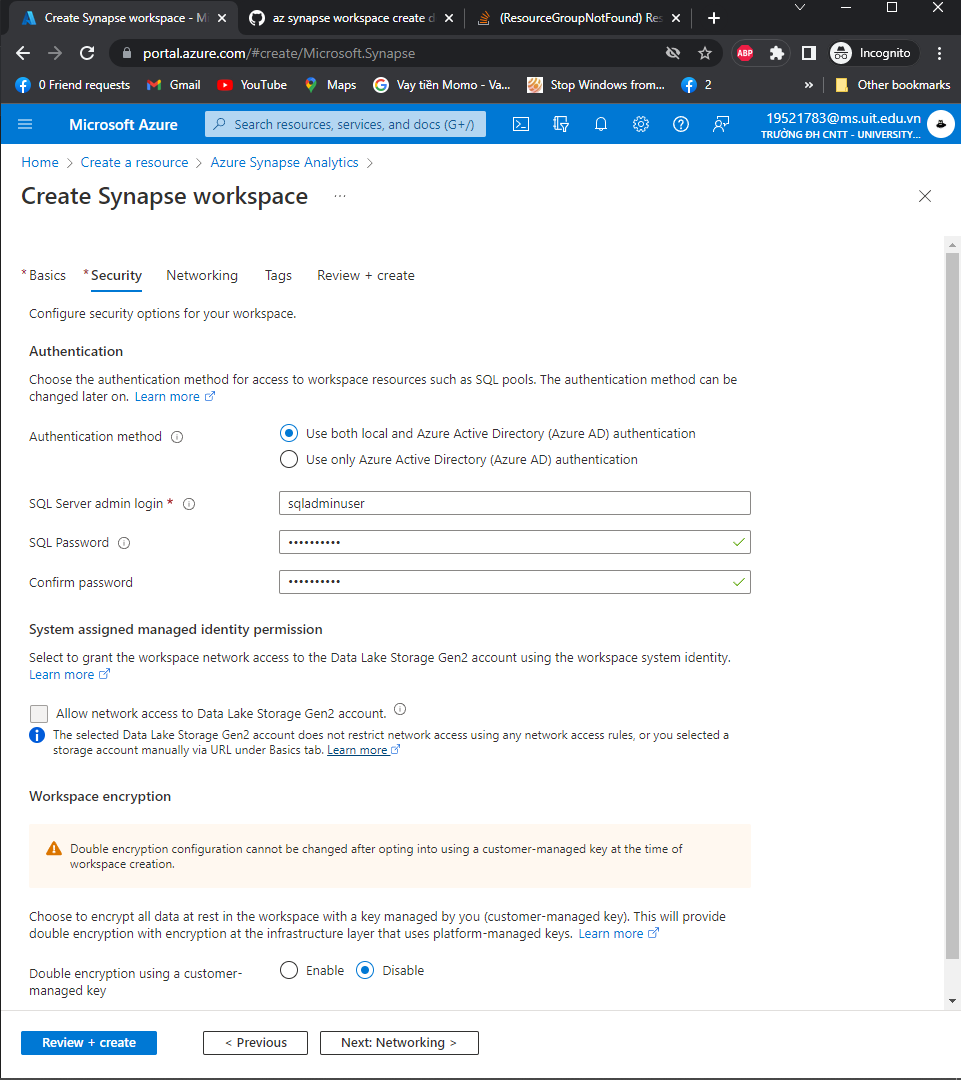
*Tạo Azure Synapse Analytics Workspace*

**

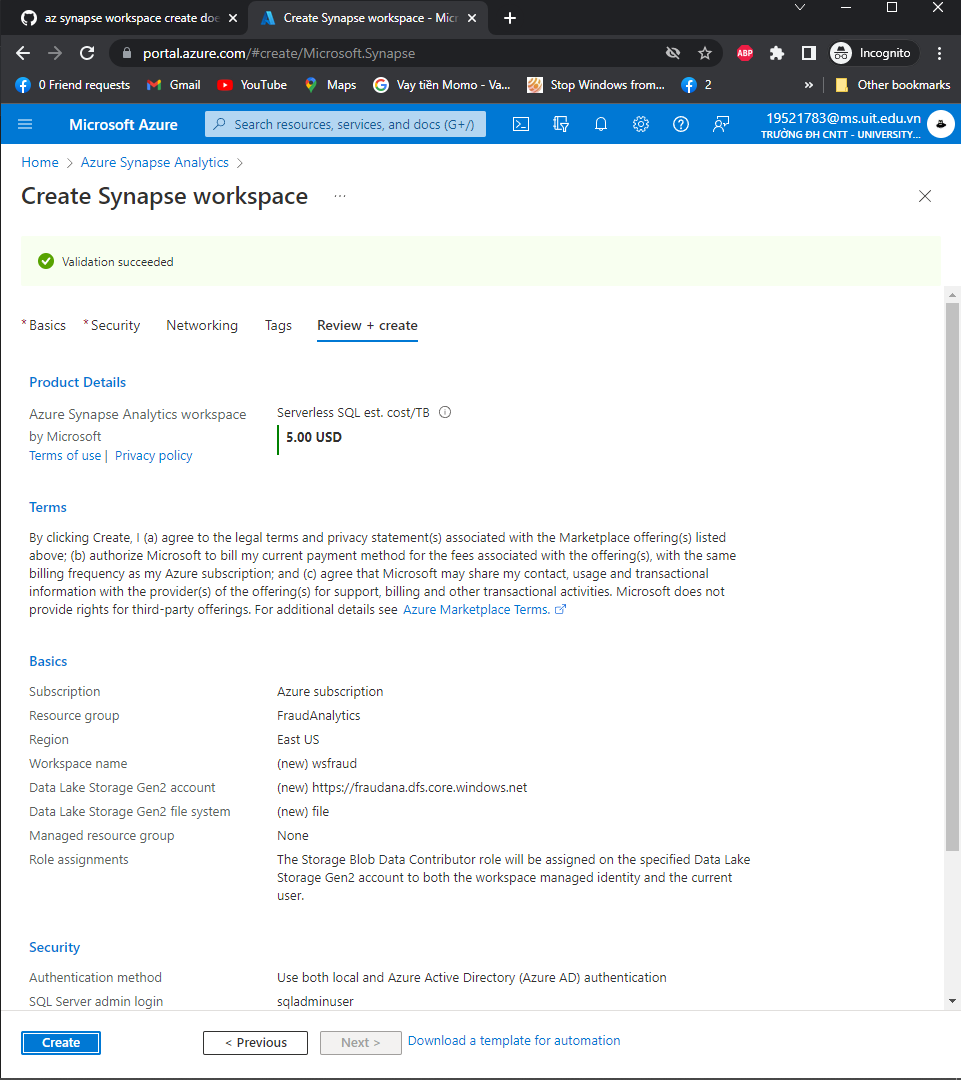
*Điền thông tin vào form*

**

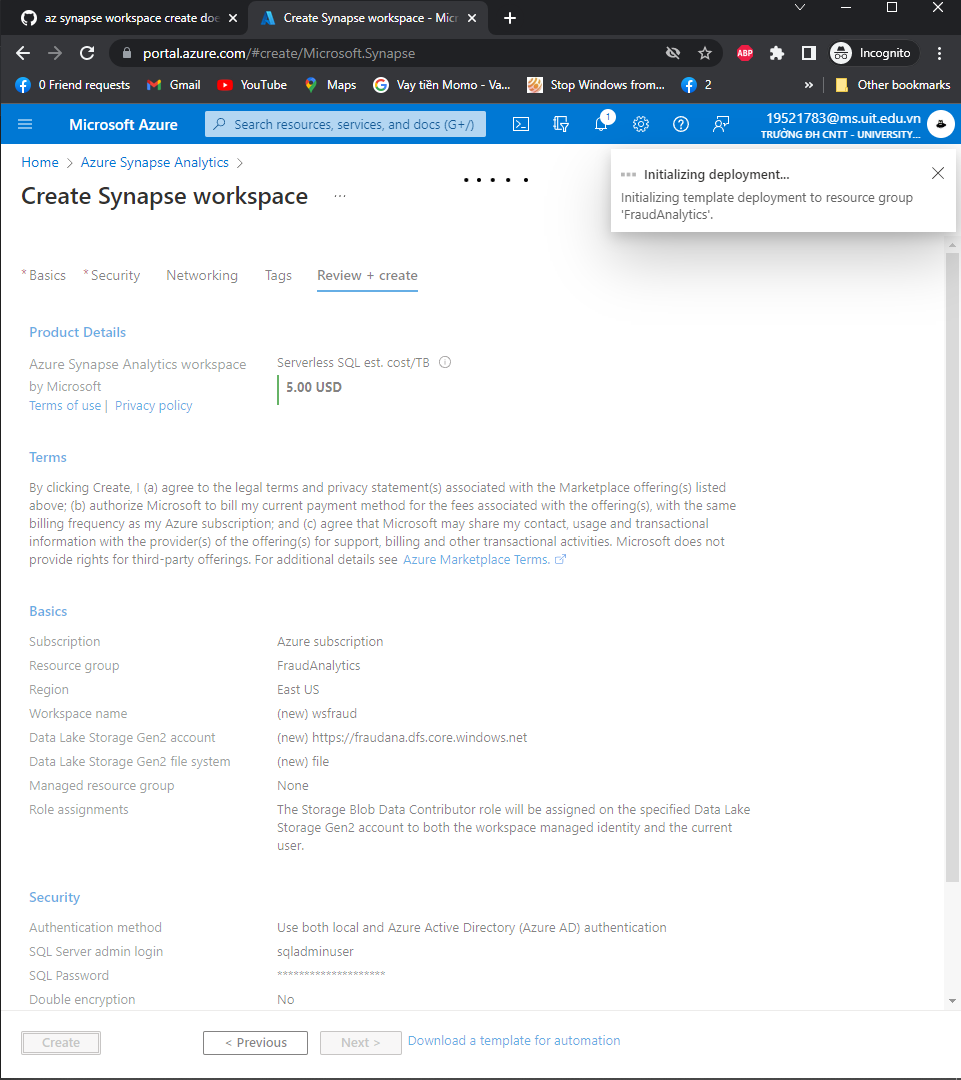
*Điền thông tin vào form*

**

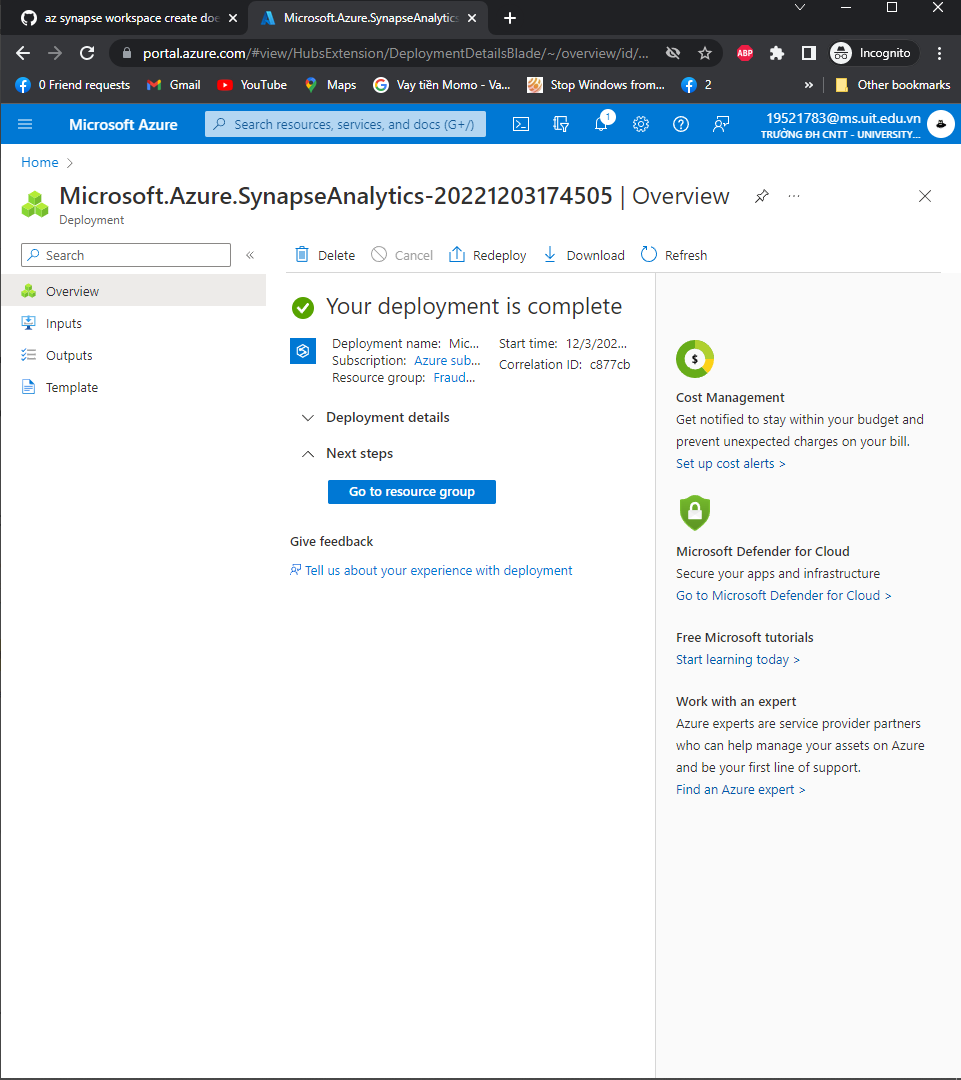
*Tạo sqladmin user*

**

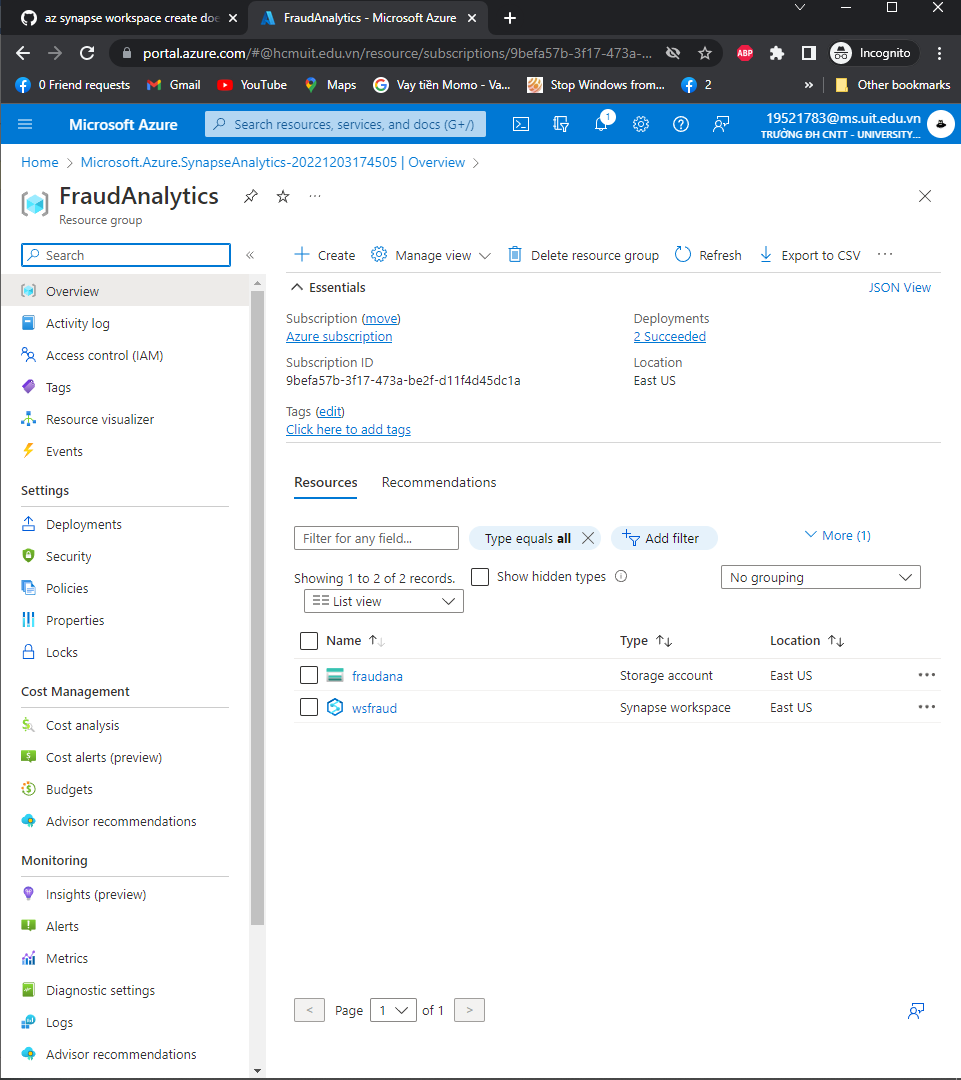
*Chọn create*

**

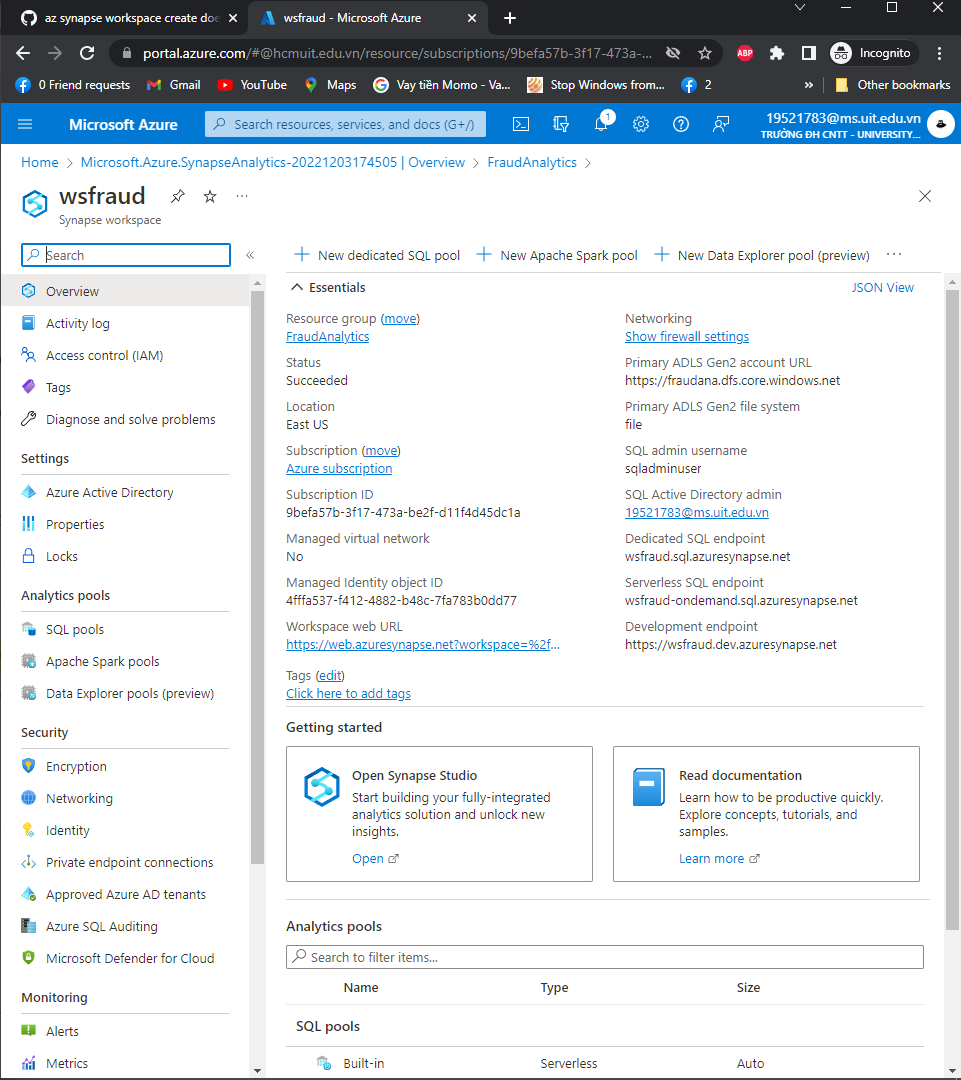
*Đợi Synapse workspace khởi tạo*

**

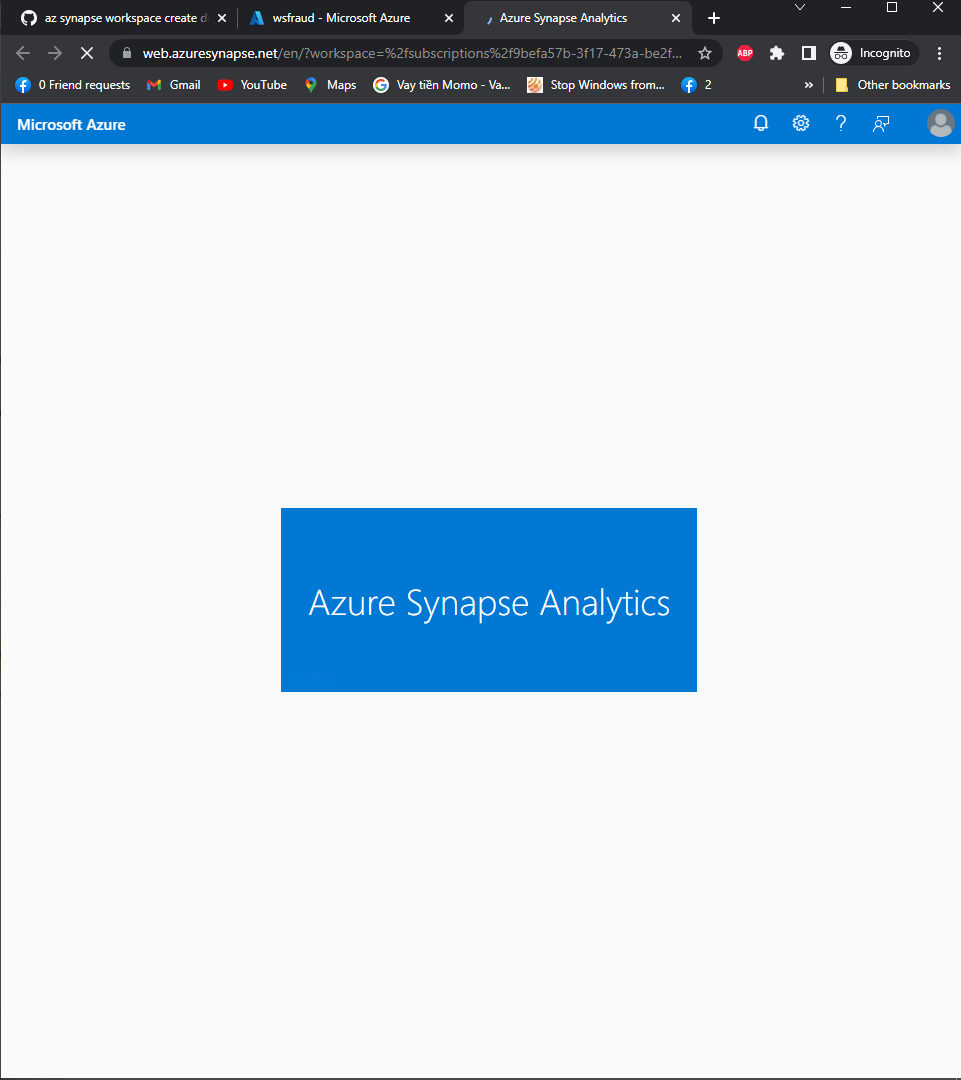
*Synapse workspace được tạo thành công*

**

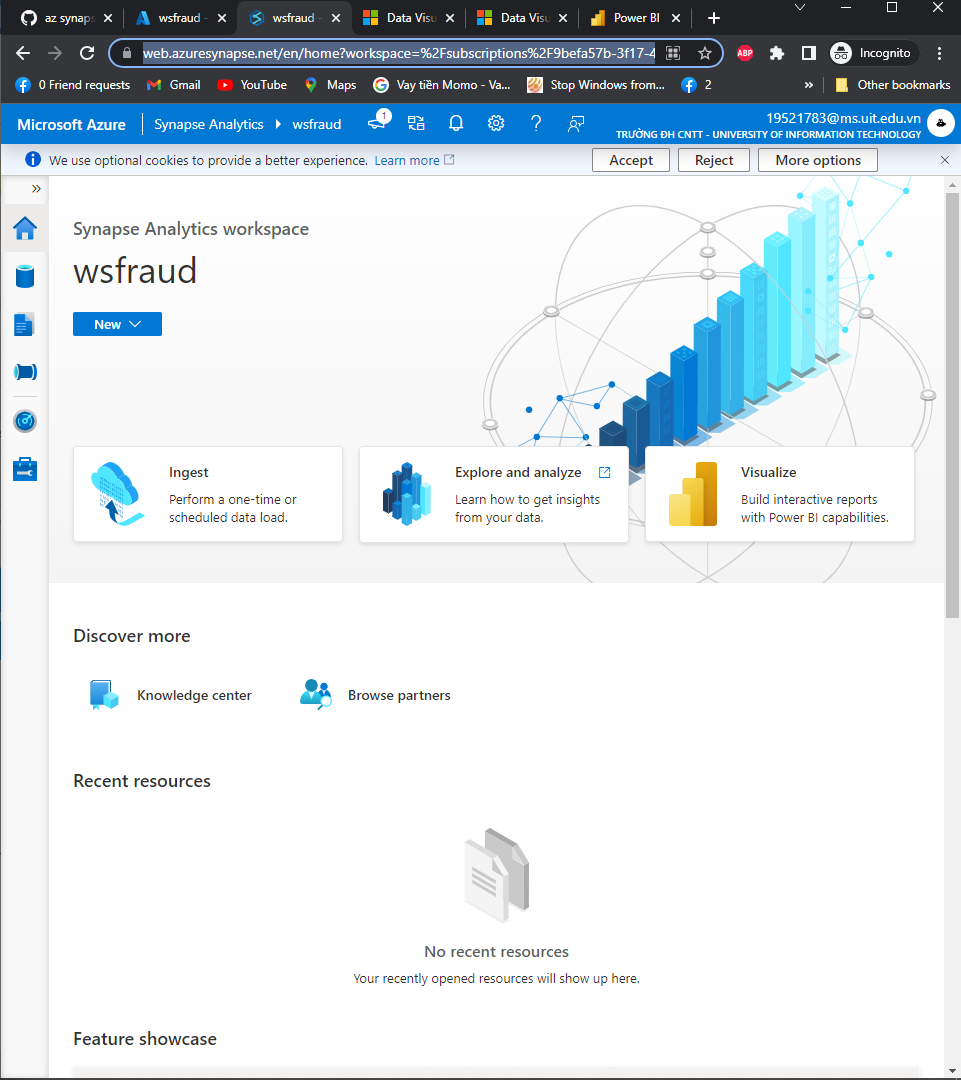
*Tại resource group sẽ có Storage account và synapse workspace*

**

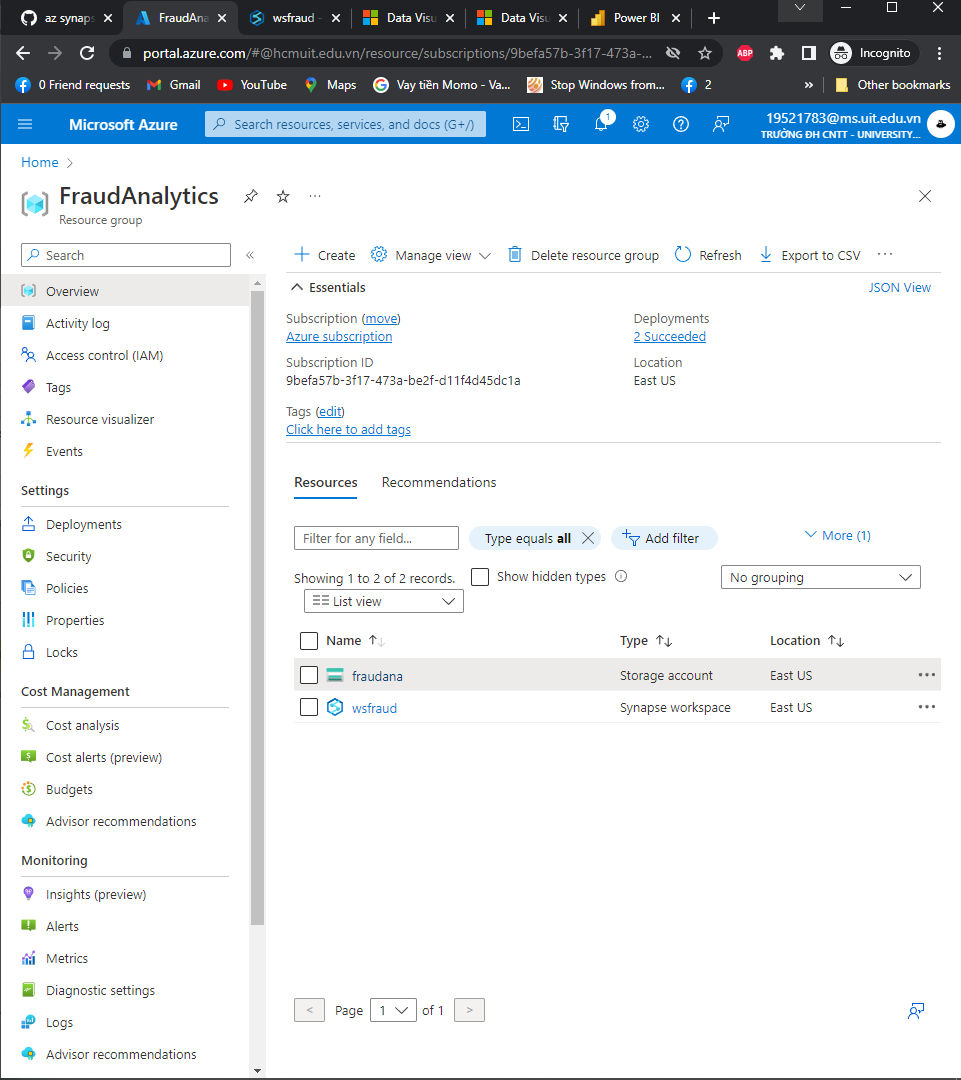
*Màn hình giao diện chính của Synapse workspace*

**

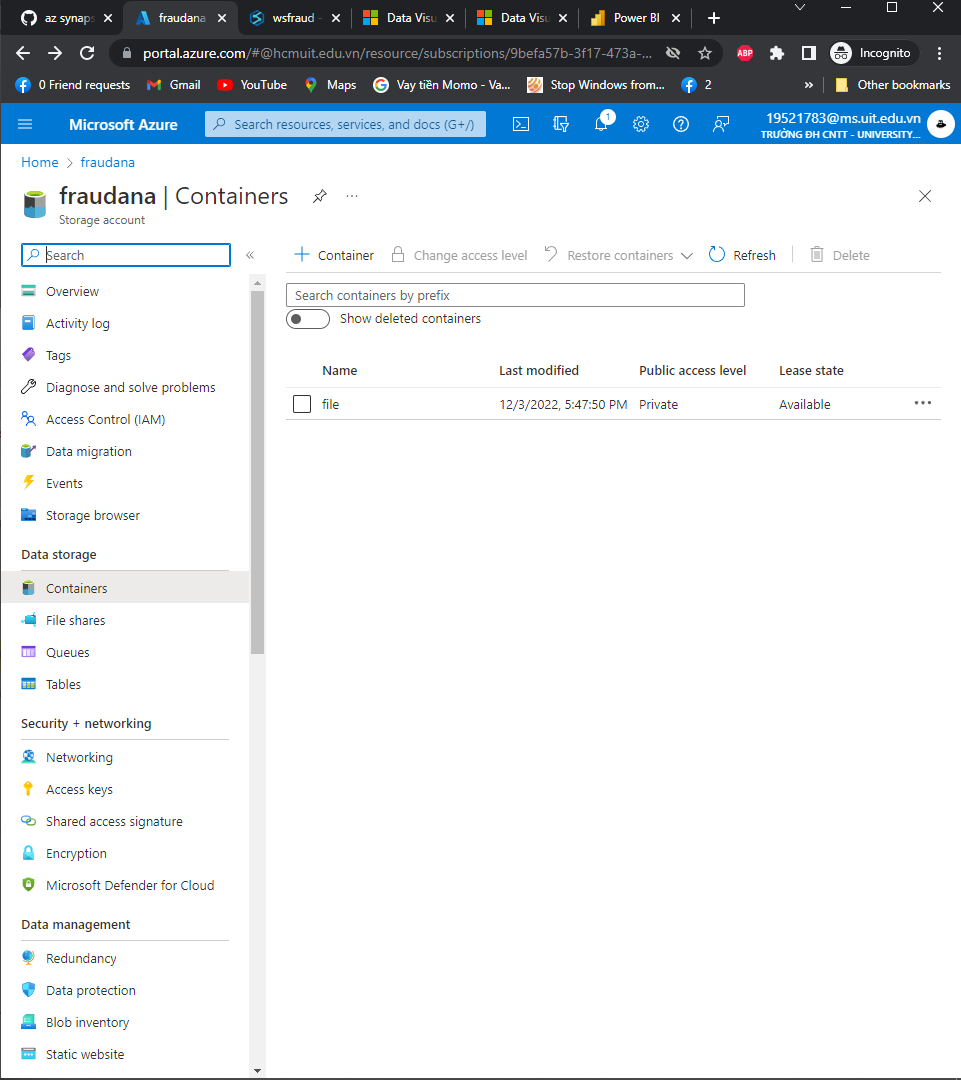
*Mở Synapse studio từ trang Synapse workspace*

**

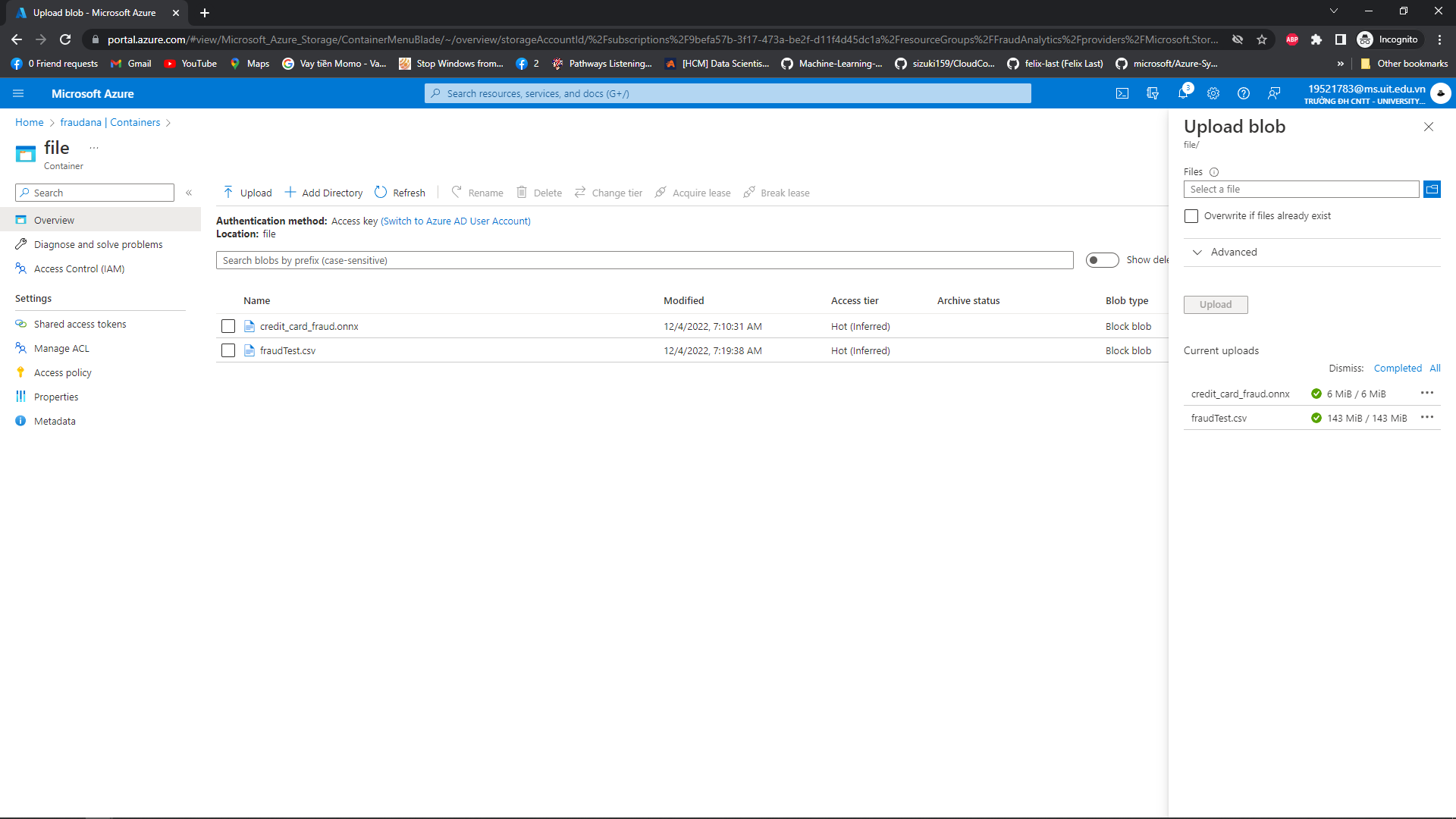
*Trang home của Synapse studio*

**

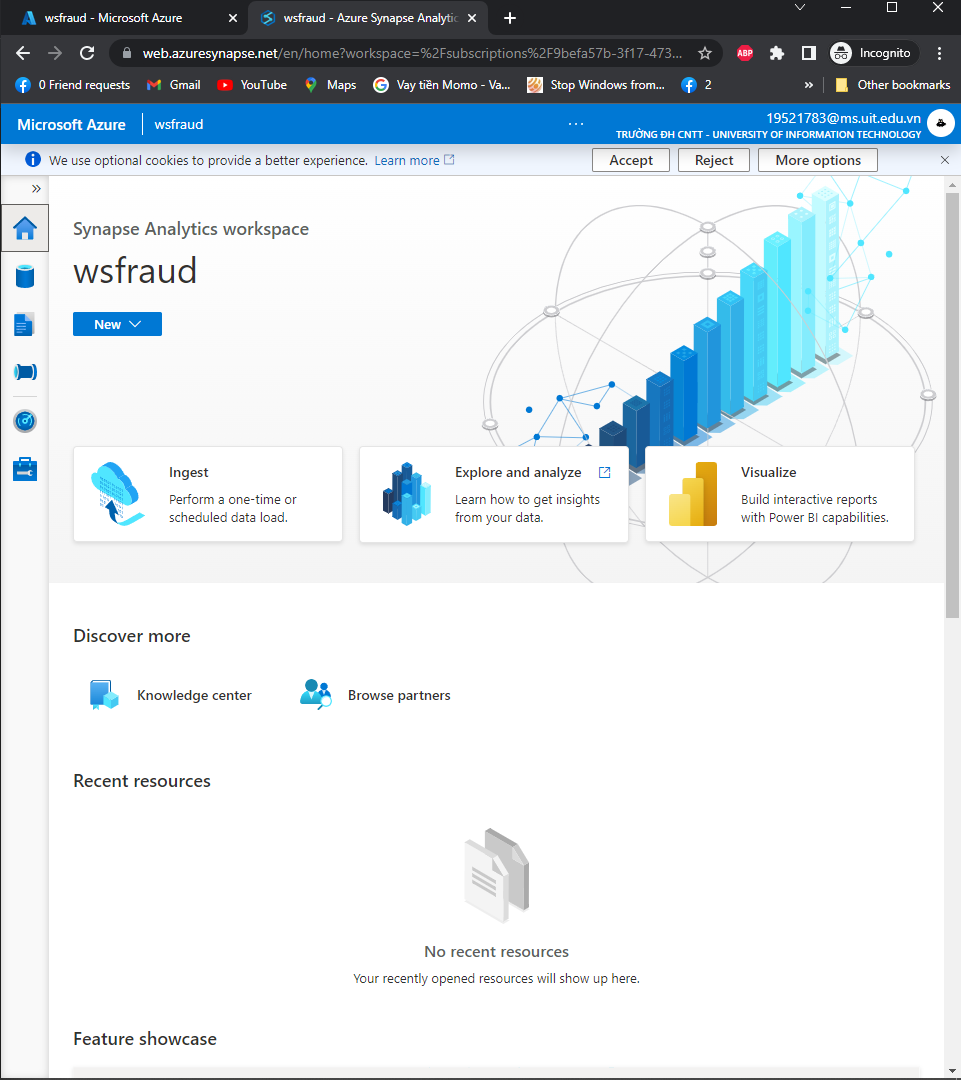
*Quay trở lại Storage account để tạo file trong container*

**

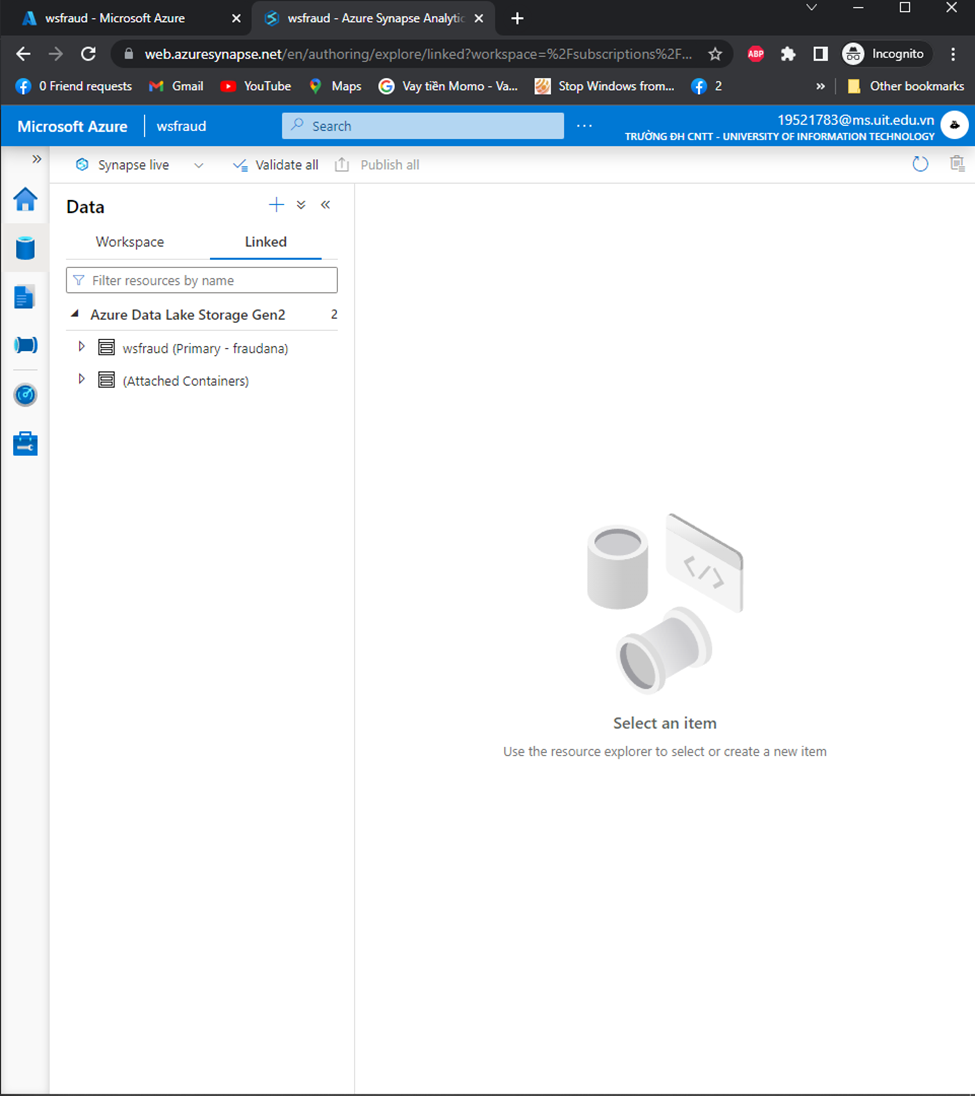
*Vào thư mục file đã tạo ở phần đầu*

**

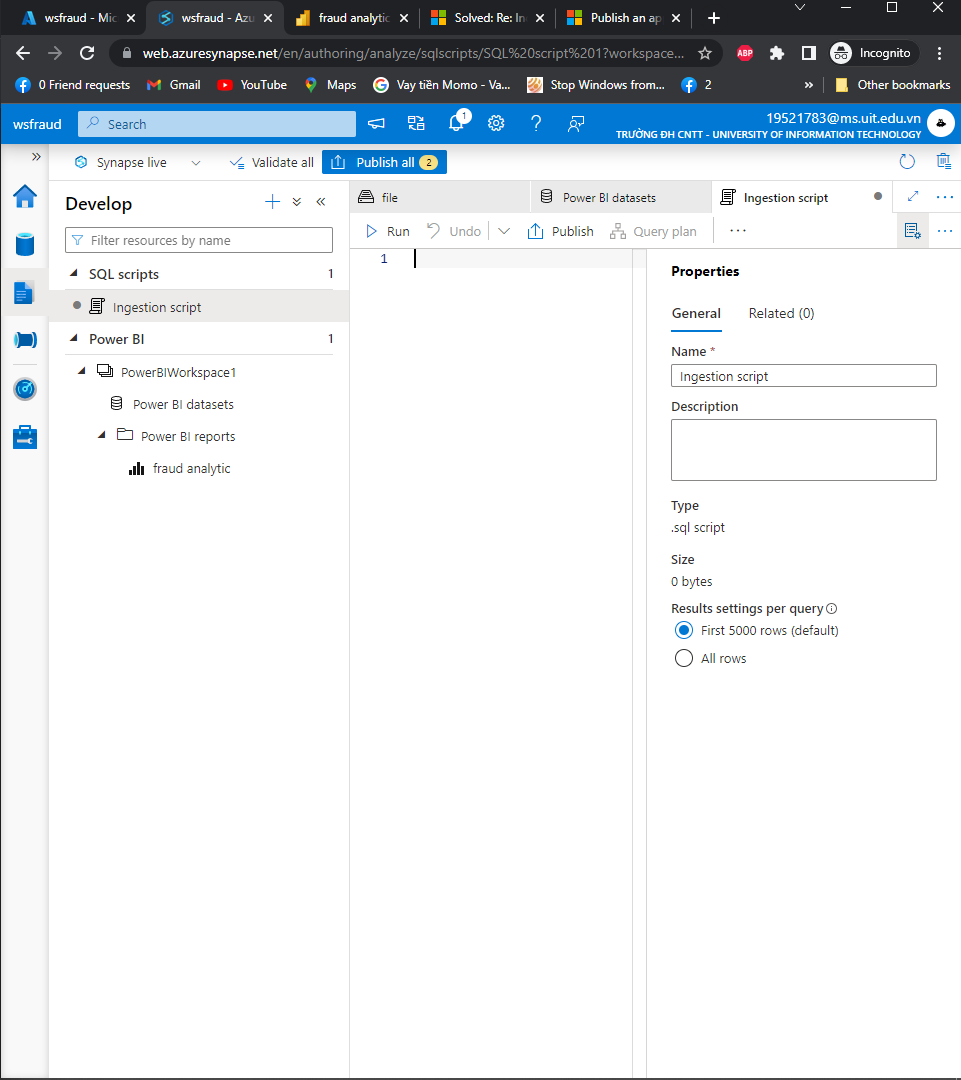
*Upload các file đã tiền xử lý dữ liệu lên Azure Synapse*

**

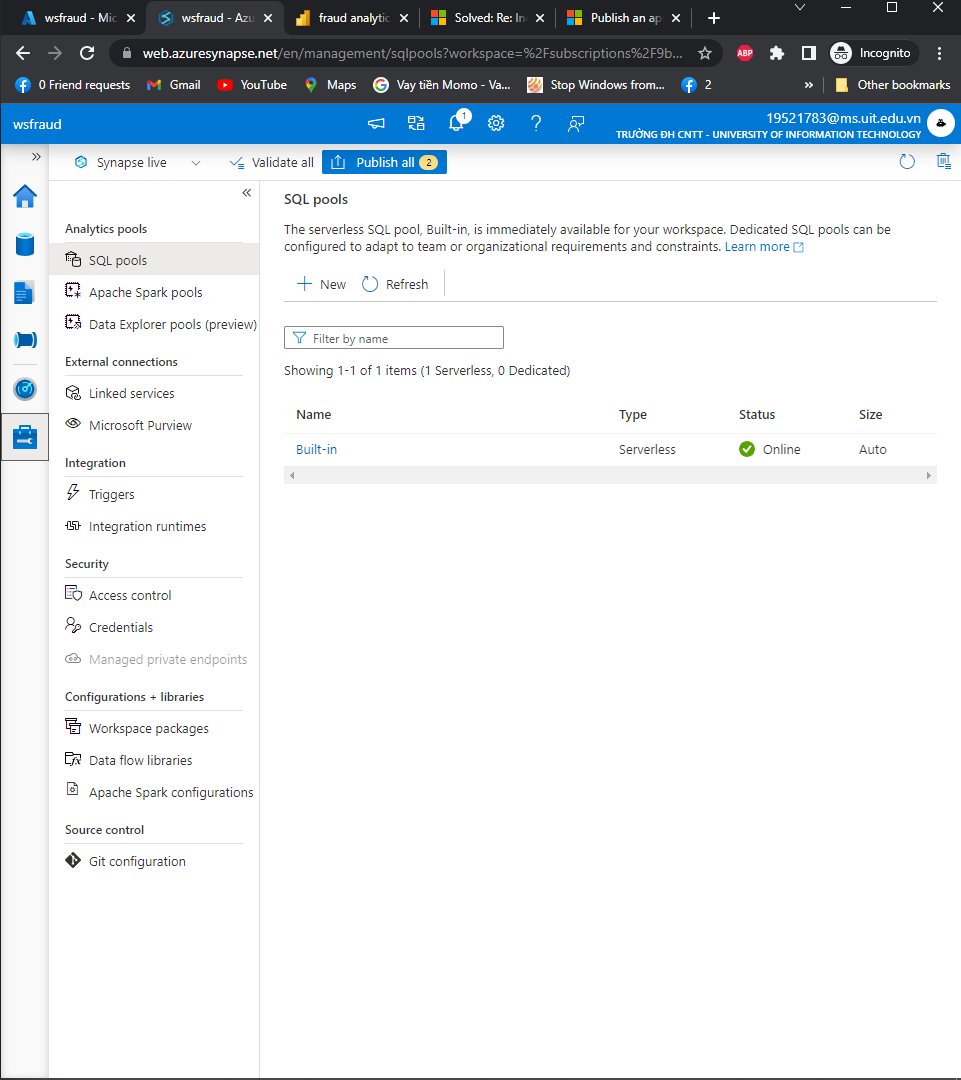
*Quay lại Synapse Studo*

**

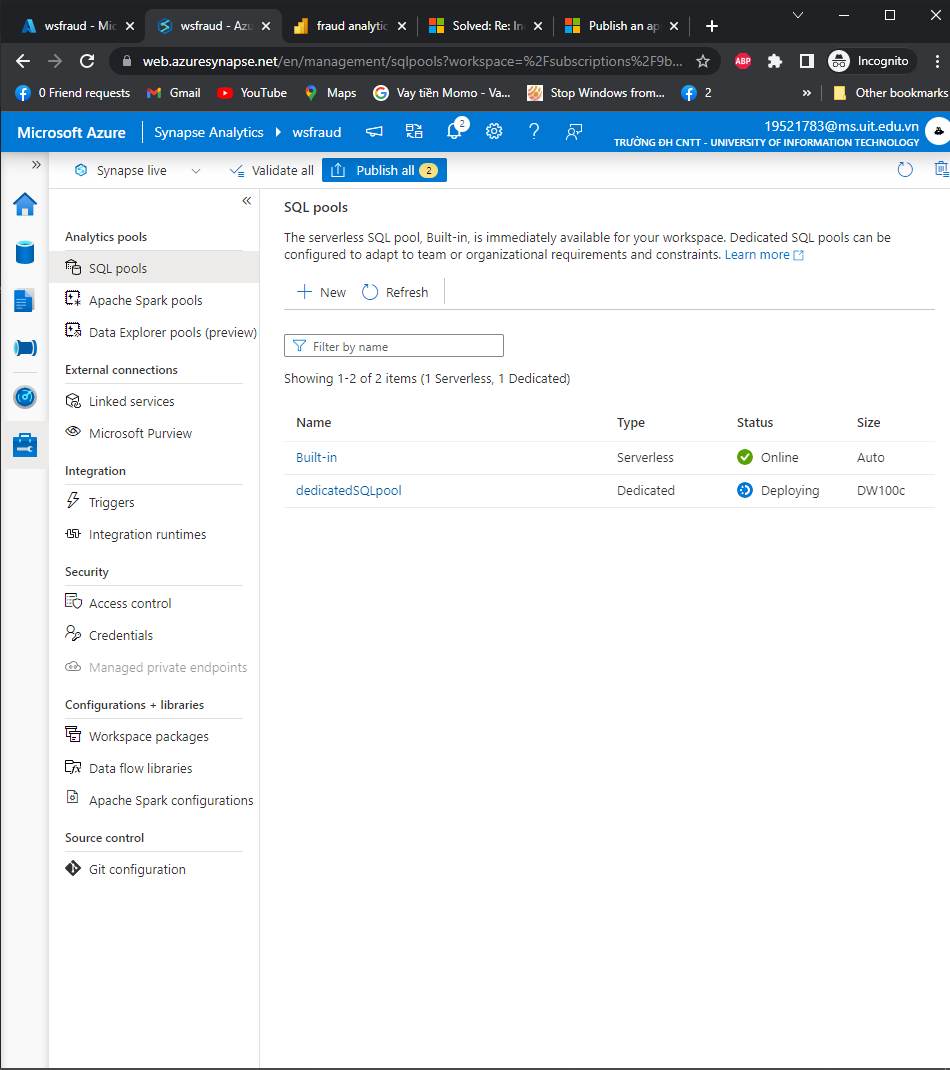
*Vào tab Data, chọn Linked*

**

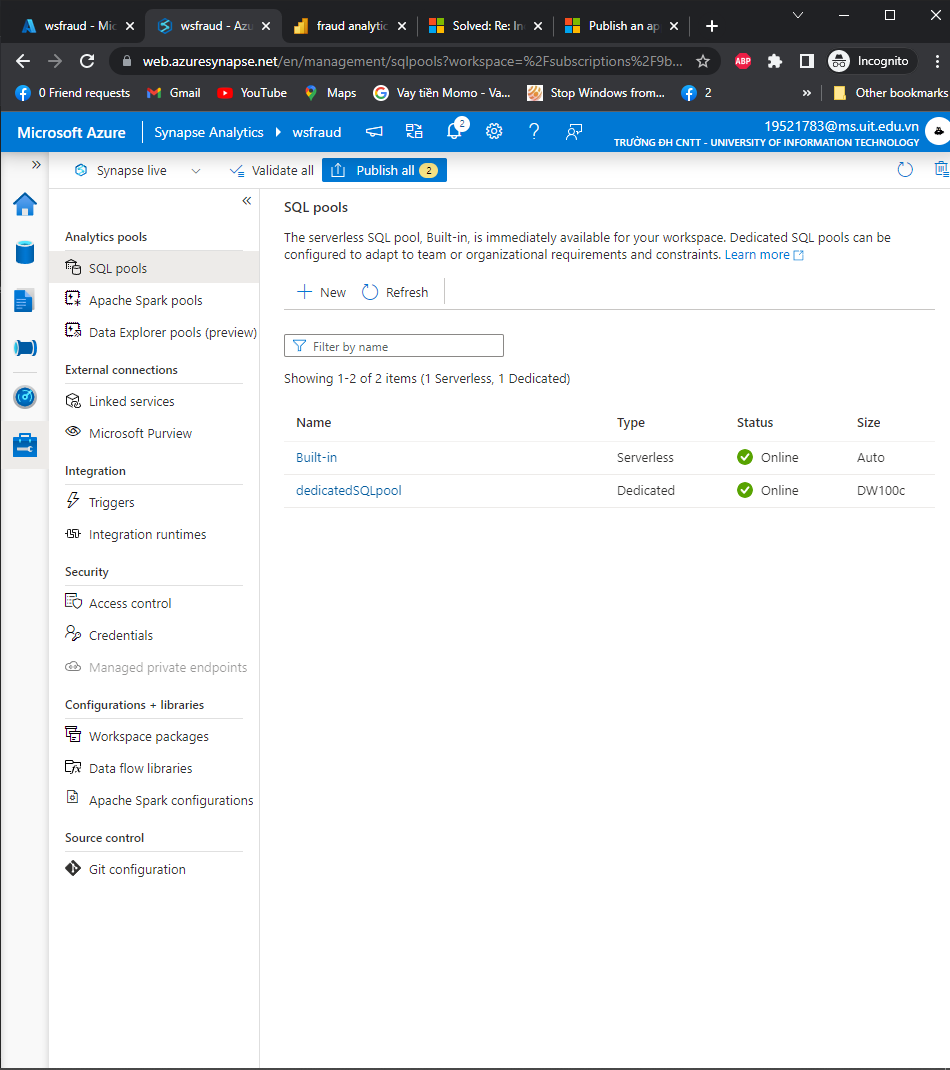
*Tạo TSQL Queries script*

**

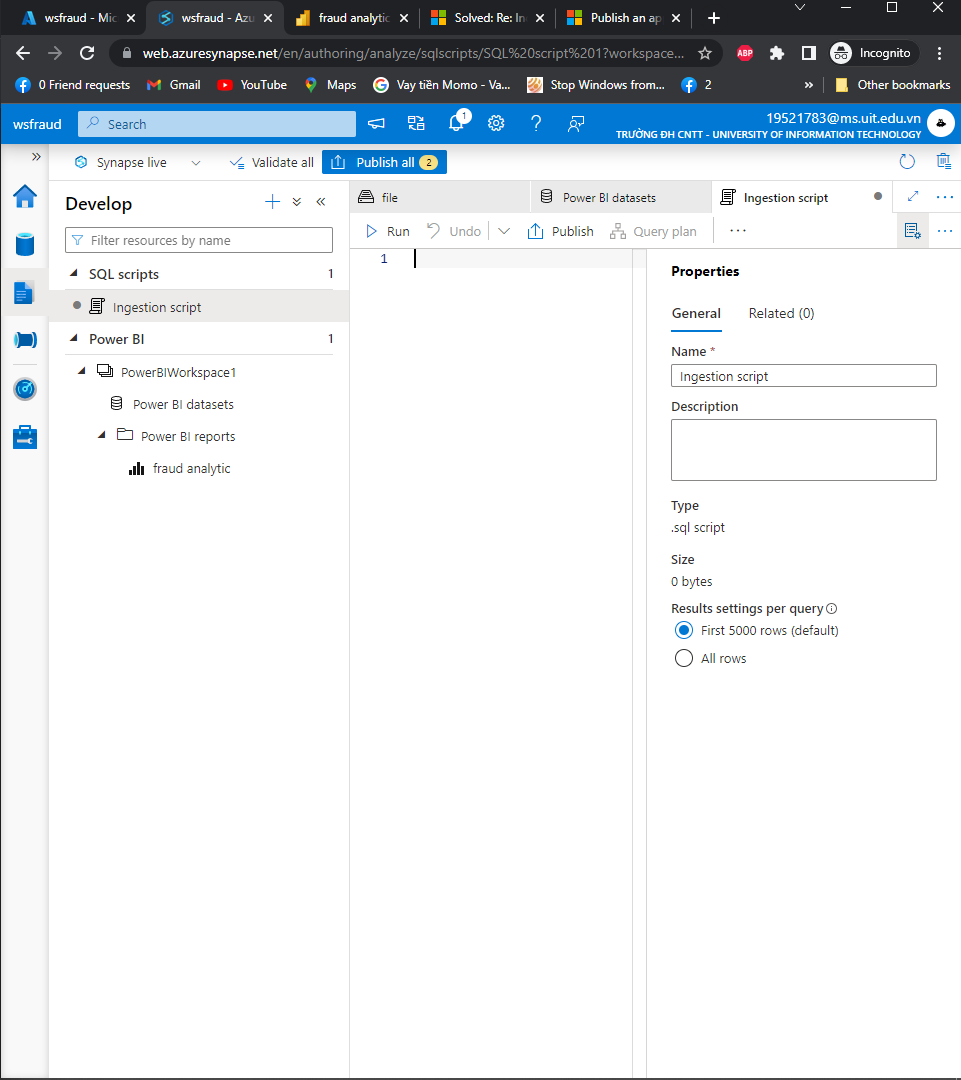
*Vào phần manage để tạo SQL Pool*

**

*Deploy Dedicated SQL Pool*

**

*Deploy Dedicated SQL Pool thành công*

**

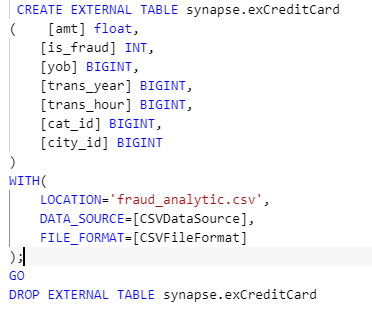
*Quay trở lại SQL Script và chạy các query sau*



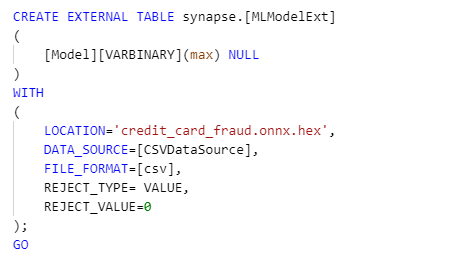
*Tạo schema và các thông tin credential*



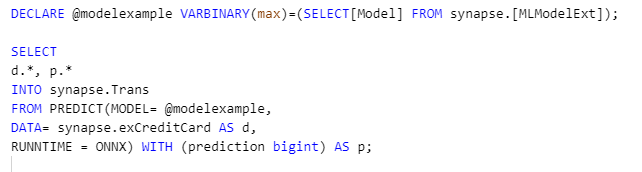
*Tạo datasource và file format*



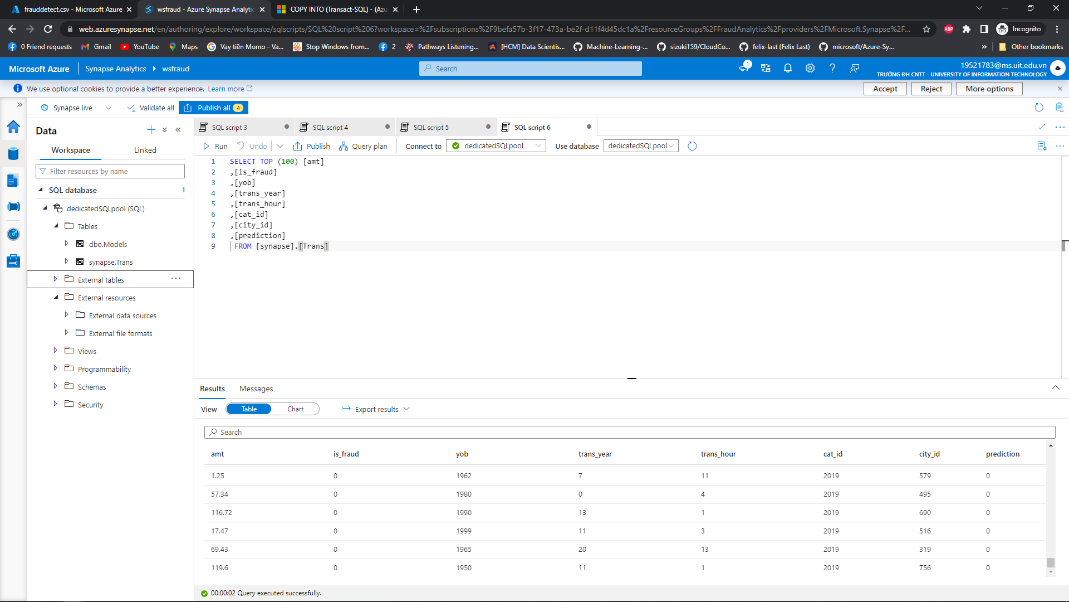
*Tạo external table CreditCard để chứa các thông tin trong dataset*



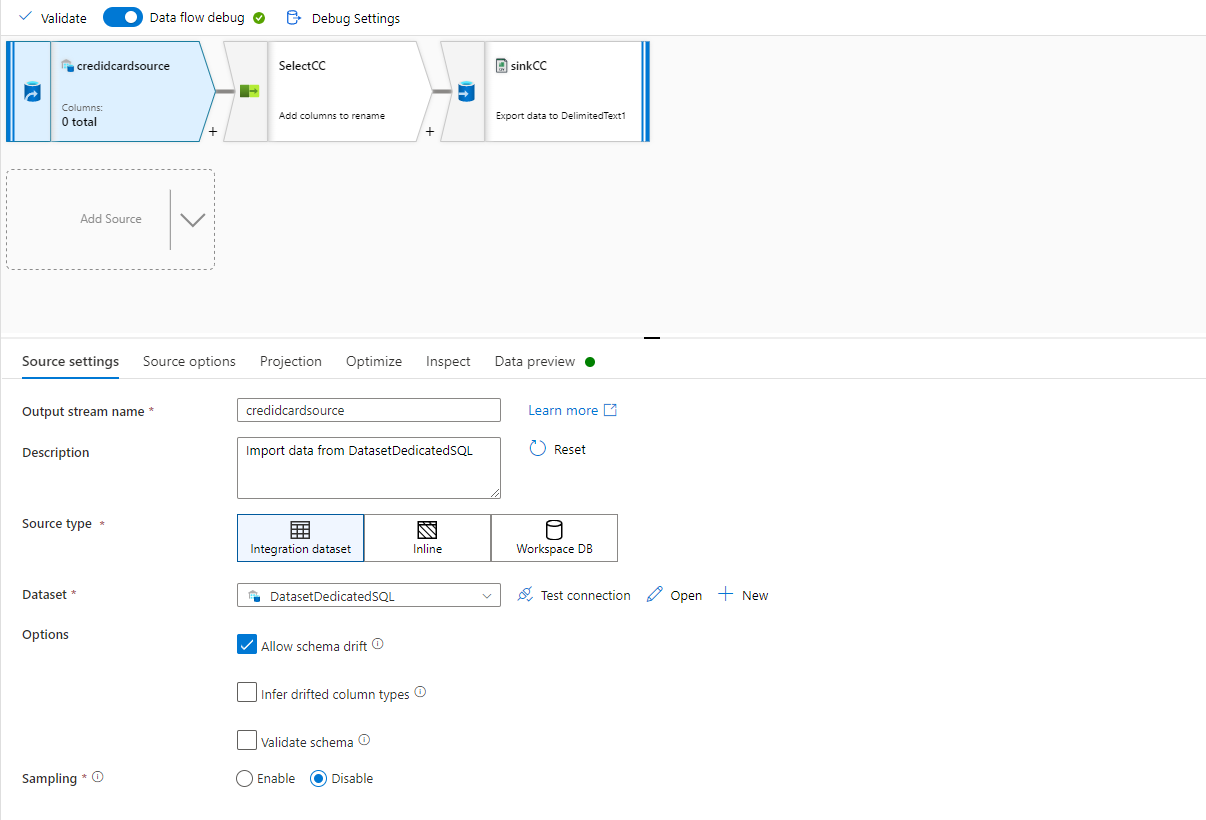
*Tạo external table MLModel để chứa model Random Forest đã build ở phần trước.*



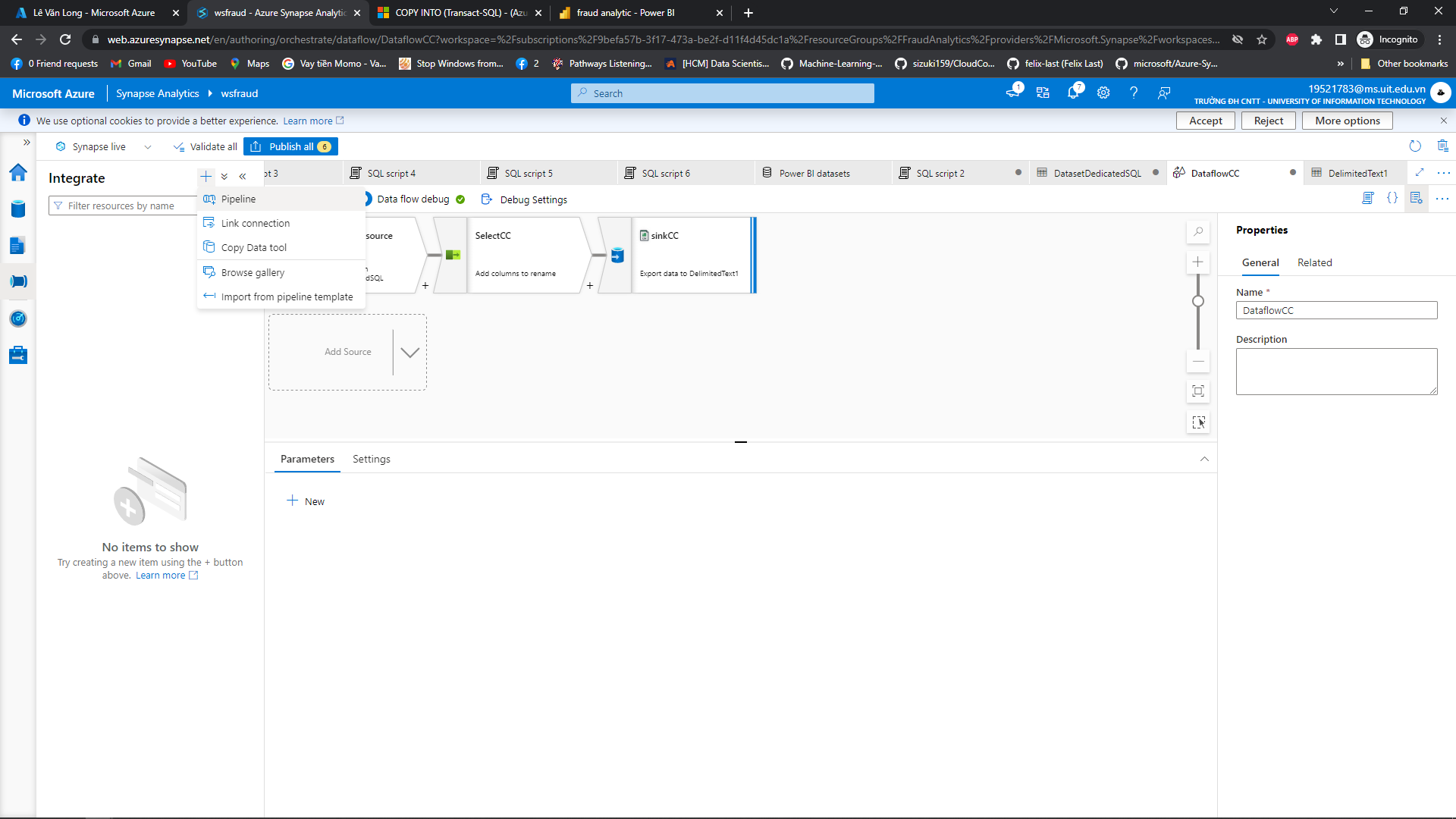
*Hàm để dự đoán bằng model Machine Learning.*



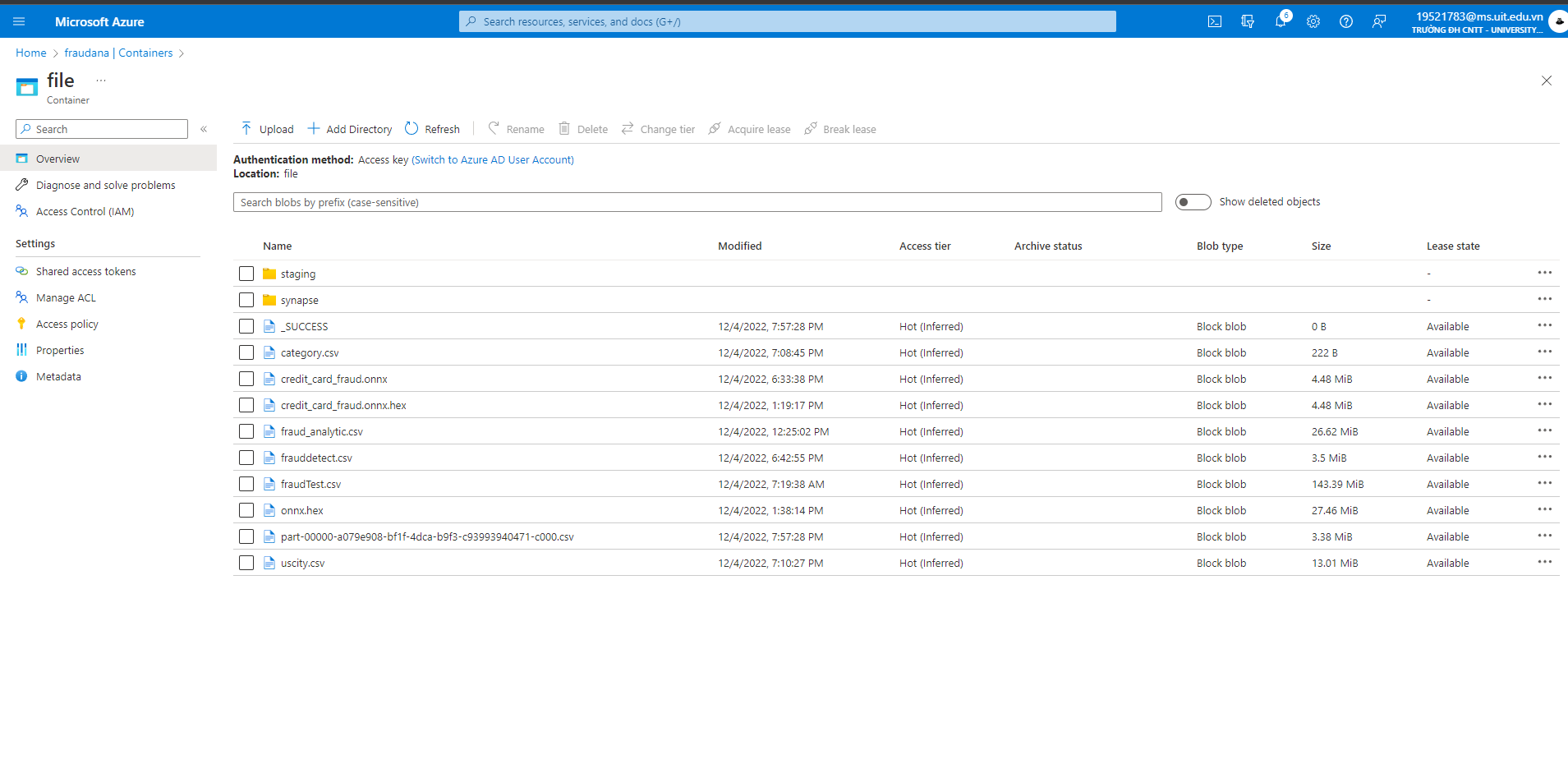
*Kết quả prediction*

**

*Tạo Data flow để tiến hành xuất kết quả dự đoán ra csv*

**

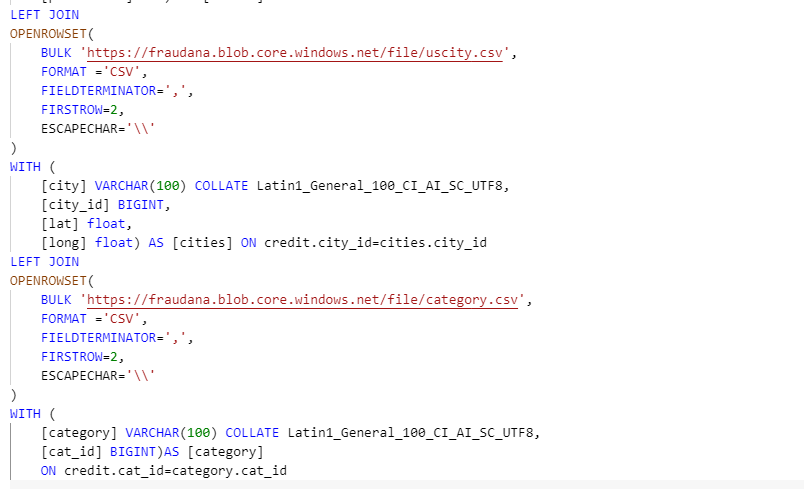
*Tạo Data Pipeline để xuất kết quả dự đoán ra file csv*

**

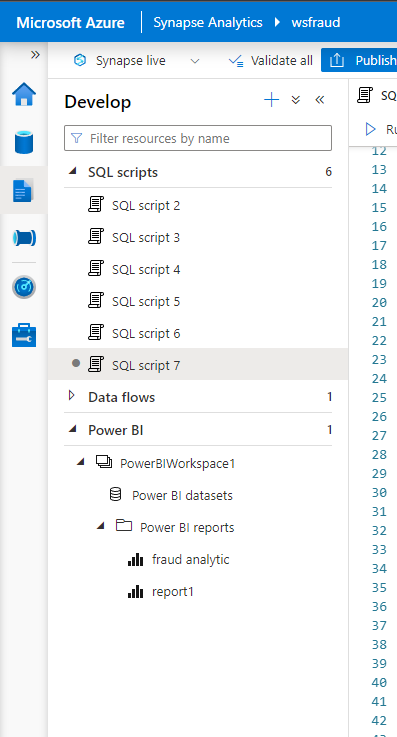
*File part0000… là file kết quả dự đoán được xuất ra*



*Tạo view để join các cột với nhau*

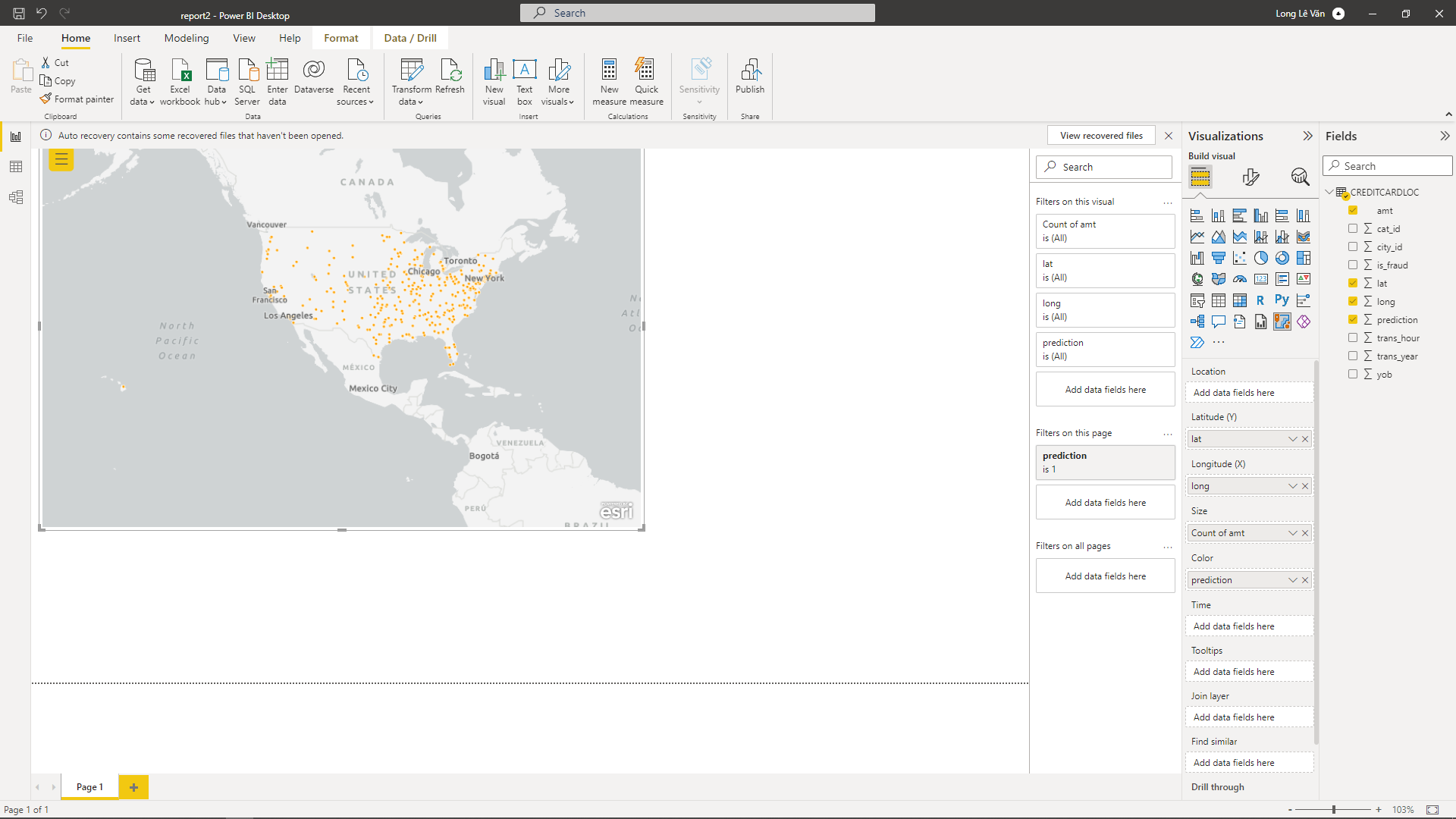


*Tạo view để join các cột với nhau*



*Chọn Power BI Dataset và tải template về.*

## THỰC HIỆN TRÊN POWER BI



*Tiến hành vẽ các biểu đồ và đưa ra kết quả phân tích*

# **CHƯƠNG 4. TÀI LIỆU THAM KHẢO**

1. What is Azure Synapse Analytics?: <https://learn.microsoft.com/vi-vn/azure/synapse-analytics/overview-what-is>
2. Tìm hiểu phân tích Azure Synapse: <https://morioh.com/p/386b2793d117>
3. Analytics end-to-end with Azure Synapse: <https://learn.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end?tabs=portal>