

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA HỆ THÔNG THÔNG TIN



BÁO CÁO BÀI TẬP QUÁ TRÌNH  
MÔN HỌC: ĐIỆN TOÁN ĐÁM MÂY

ĐỀ TÀI

Sử dụng Azure Synapse Analytics để truy vấn Data Lake  
(Using Azure Synapse Analytics to Query Data Lake)

Lớp: IS402.O21.HTCL

GVHD: ThS. Hà Lê Hoài Trung

Nhóm sinh viên thực hiện:

Phan Chí Cường 21520673

Nguyễn Minh Duy 21522005

TP. HỒ CHÍ MINH, 2024

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA HỆ THÔNG THÔNG TIN



BÁO CÁO BÀI TẬP QUÁ TRÌNH  
MÔN HỌC: ĐIỆN TOÁN ĐÁM MÂY

ĐỀ TÀI

Sử dụng Azure Synapse Analytics để truy vấn Data Lake  
(Using Azure Synapse Analytics to Query Data Lake)

Lớp: IS402.O21.HTCL

GVHD: ThS. Hà Lê Hoài Trung

Nhóm sinh viên thực hiện:

Phan Chí Cường 21520673

Nguyễn Minh Duy 21522005

TP. HỒ CHÍ MINH, 2024

## MỤC LỤC

CHƯƠNG 1. HỒ DỮ LIỆU .....	6
1.1. Tổng quan.....	6
1.2. Ứng dụng.....	10
1.3. Ưu điểm .....	10
1.4. Dịch vụ lưu trữ Data Lake Trên Azure – Azure Data Lake StoraGE .....	12
1.5. So sánh Data Lake và Data WareHouse.....	14
1.6. Ứng dụng.....	16
CHƯƠNG 2. AZURE SYNAPSE ANALYTICS.....	17
2.1. Tổng quan.....	17
2.2. Kiến trúc của Azure Synapse .....	18
2.3. Xây dựng Azure Synapse Analytics .....	33
2.4. So sánh Data Lake và Data WareHouse.....	44
CHƯƠNG 3. XÂY DỰNG ETL PIPELINE TRÊN AZURE DATA FACTORY .....	46
3.1. Trường hợp 1 .....	46
3.2. Trường hợp 2 .....	50
Tài liệu tham khảo .....	59

## Phụ lục hình ảnh

Hình 1 Data lake .....	7
Hình 2 Luồng dữ liệu của data lake .....	9
Hình 3. Ứng dụng của data lake.....	10
Hình 4. Azure Data Lake Storage .....	14
Hình 5 So sánh data lake và data warehouse.....	15
Hình 6. Dịch vụ Azure Synapse Analytics.....	17
Hình 7 Các ứng dụng sử dụng dịch vụ Azure Synapse Analytics.....	18
Hình 8. Kiến trúc của Azure Synapse Analytics.....	19
Hình 9. Dedicated SQL pool và Serverless SQL pool .....	21
Hình 10. Apache spark.....	24
Hình 11. Sự khác biệt của Apache spark so với mô hình MapReduce truyền thống.....	25
Hình 12. Synapse Pipeline .....	27
Hình 13. Các thành phần của Synapse pipeline.....	29
Hình 14. Giao diện của Synapse Studio .....	30
Hình 15. Azure Synapse Link.....	32
Hình 16. Hệ thống lưu trữ dữ liệu truyền thống.....	33
Hình 17. Kiến trúc của hồ dữ liệu .....	34
Hình 18. Các dịch vụ để giả lập máy chủ data lake .....	35
Hình 19. Cấu hình máy chủ.....	35
Hình 20. Kịch bản chạy local data lake trên máy chủ .....	36
Hình 21. Upload tệp làm việc lên trình lưu trữ đối tượng MinIO.....	37
Hình 22. Thực hiện truy vấn câu lệnh đơn giản với đối tượng .....	38
Hình 23. Resource groups - các tài nguyên trên Microsoft Azure .....	40
Hình 24. Tải tệp làm việc lên container của Synapse workspace .....	40
Hình 25. Tạo SAS Token và URL cho container chứa tệp làm việc .....	41
Hình 26. Tạo Linked services đến container nơi chứa tệp làm việc .....	41
Hình 27. Thủ truy vấn 100 giá trị đầu tiên của tệp làm việc .....	42

Hình 28. Thực hiện truy vấn câu lệnh tương đương trên máy chủ truyền thống để so sánh.....	42
Hình 29. Câu lệnh tạo view trên Azure Synapse Studio .....	44
Hình 30. Tạo view cho database db .....	44
Hình 31. Cấu hình linked service với máy chủ Windows.....	46
Hình 32. Các tệp trên máy chủ Windows .....	47
Hình 33. Tạo copy data activity với source là folder "branch" nơi chứa các tệp cần merge .....	47
Hình 34. Tệp tin đích sẽ được lưu trên container .....	48
Hình 35. Chạy pipeline để merge các tệp lại.....	48
Hình 36. Chi tiết của lệnh chạy pipeline .....	49
Hình 37. Tệp tin đích sau khi gộp .....	49
Hình 38. Chi tiết của tệp đích sau khi gộp dữ liệu .....	50
Hình 39. Pipeline để tách dữ liệu dựa trên thuộc tính cột.....	51
Hình 40. Thực thi luồng xử lý .....	51
Hình 41. Kết quả chi tiết của luồng xử lý.....	52
Hình 42. Luồng xử lý để tách và chuyển đổi tệp tin thành định dạng khác.....	53
Hình 43. Để tạo 1 mảng Male trong Json, chúng ta sẽ tạo 1 cột mới chứa object cũ.....	54
Hình 44. Lọc những đối tượng unique trong dataset .....	55
Hình 45 Kết quả thực thi luồng xử lý .....	56
Hình 46. Kết quả chi tiết của luồng xử lý.....	56
Hình 47. Chiều dài của mảng sau khi chỉ chọn những đối tượng unique .....	57
Hình 48. Tệp tin sau khi được chuyển đổi định dạng.....	58

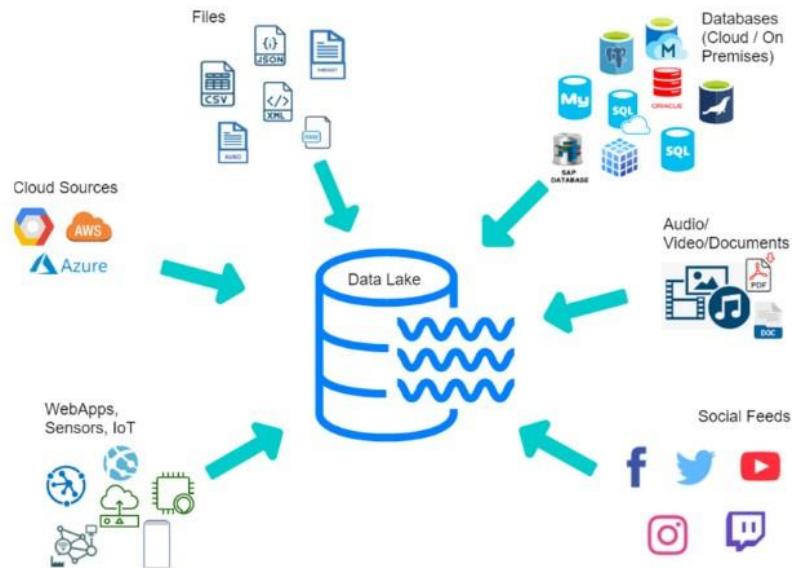
# CHƯƠNG 1. HỒ DỮ LIỆU

## 1.1. Tổng quan

**Khái niệm:**

- Hồ dữ liệu là một hệ thống hoặc kho lưu trữ dữ liệu được lưu trữ trong định dạng tự nhiên/nguyên thủy, thường là các đối tượng blob hoặc tệp.
- Hồ dữ liệu thường là một cửa hàng dữ liệu duy nhất bao gồm bản sao thô của dữ liệu hệ thống nguồn, dữ liệu cảm biến, dữ liệu xã hội, v.v., và dữ liệu đã được biến đổi sử dụng cho các nhiệm vụ như báo cáo, trực quan hóa, phân tích nâng cao và học máy.
- Một hồ dữ liệu có thể bao gồm:
  - Dữ liệu có cấu trúc từ các cơ sở dữ liệu quan hệ (hàng và cột),
  - Dữ liệu bán cấu trúc (CSV, log, XML, JSON),
  - Dữ liệu không cấu trúc (email, tài liệu, PDF),
  - Dữ liệu nhị phân (hình ảnh, âm thanh, video).
- Một hồ dữ liệu có thể được thiết lập "tại chỗ" (trong các trung tâm dữ liệu của một tổ chức) hoặc "trong đám mây" (sử dụng các dịch vụ đám mây từ các nhà cung cấp như Amazon, Microsoft, Oracle Cloud, hoặc Google).

Vậy, **Azure Data Lake** là một dịch vụ lưu trữ và phân tích dữ liệu có khả năng mở rộng. Dịch vụ này được lưu trữ trên Azure, đám mây công cộng của Microsoft.



Hình 1 Data lake

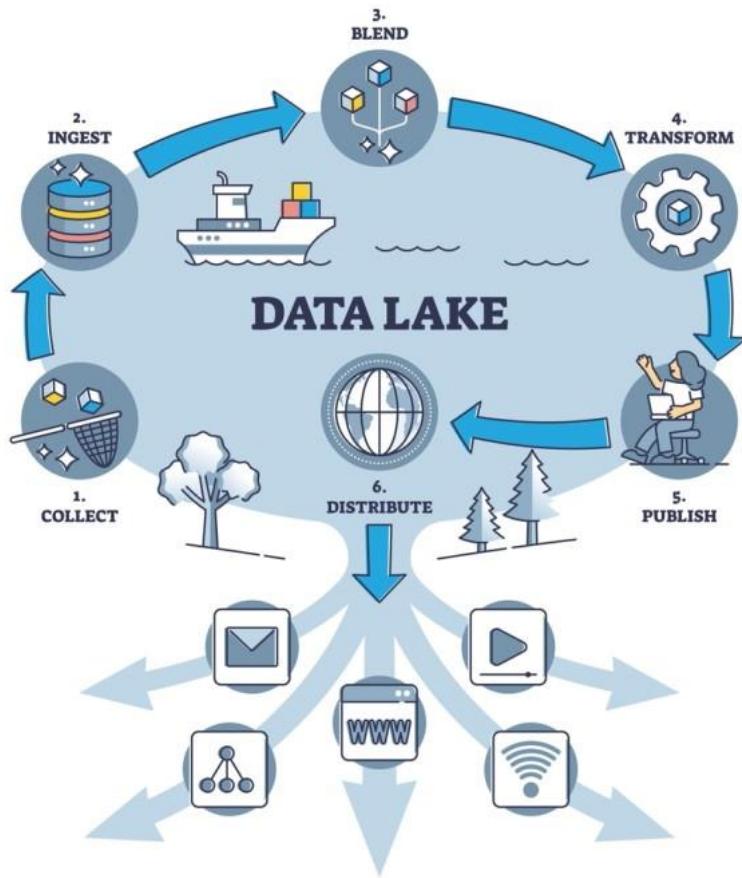
## Tại sao hồ dữ liệu (data lake) quan trọng đối với doanh nghiệp?

- Trong thế giới hiện đại, kết nối cao và lấy phân tích làm định hướng, các tổ chức không thể tồn tại mà thiếu các giải pháp hồ dữ liệu. Điều này là do các tổ chức phụ thuộc vào nền tảng hồ dữ liệu toàn diện, chẳng hạn như Azure Data Lake, để duy trì dữ liệu thô được hợp nhất, tích hợp, bảo mật và dễ dàng truy cập.
- Các công cụ lưu trữ có thể mở rộng, như Azure Data Lake Storage, cho phép lưu trữ và bảo vệ dữ liệu tại một vị trí trung tâm, giúp loại bỏ các silo dữ liệu riêng biệt và tối ưu hóa chi phí.
- Hồ dữ liệu đặt nền tảng cho nhiều loại khối lượng công việc, bao gồm:
  - Xử lý dữ liệu lớn (big data processing)
  - Truy vấn SQL
  - Khai thác văn bản (text mining)
  - Phân tích luồng dữ liệu (stream analytics)
  - Học máy (machine learning)
- Dữ liệu sau khi được xử lý có thể được sử dụng cho nhu cầu trực quan hóa dữ liệu và báo cáo cụ thể.

- Một nền tảng dữ liệu hiện đại, đầu-cuối như Azure Synapse Analytics có khả năng đáp ứng đầy đủ các nhu cầu của kiến trúc dữ liệu lớn, trung tâm là hồ dữ liệu.

### **Lịch sử ra đời Azure Data Lake:**

Dịch vụ Azure Data Lake đã được phát hành vào ngày 16 tháng 11 năm 2016. Nó dựa trên COSMOS, được sử dụng để lưu trữ và xử lý dữ liệu cho các ứng dụng như Azure, AdCenter, Bing, MSN, Skype và Windows Live. COSMOS có một công cụ truy vấn giống SQL gọi là SCOPE, trên đó U-SQL đã được xây dựng.



Hình 2 Luồng dữ liệu của data lake

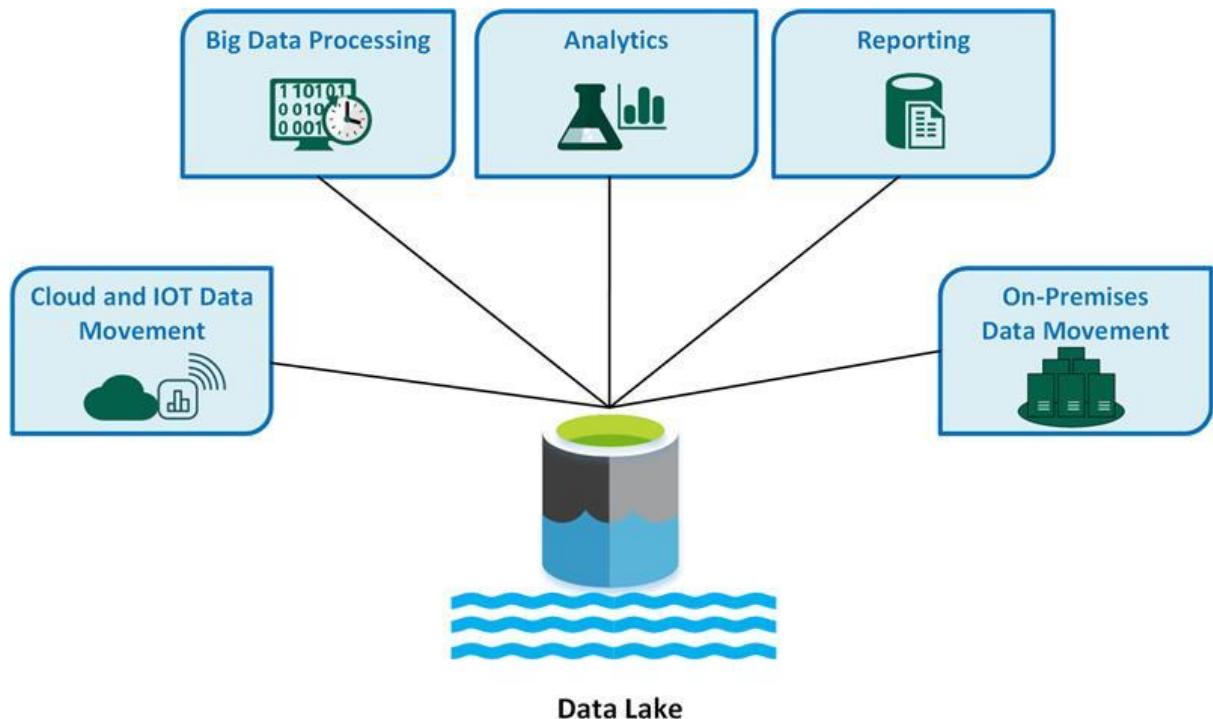
Hồ dữ liệu cung cấp nền tảng có khả năng mở rộng và bảo mật, cho phép các doanh nghiệp:

- Thu thập bất kỳ dữ liệu nào từ bất kỳ hệ thống nào với bất kỳ tốc độ nào – ngay cả khi dữ liệu đến từ hệ thống tại chỗ, đám mây, hoặc hệ thống tính toán tại biên.
- Lưu trữ bất kỳ loại hoặc khối lượng dữ liệu nào với độ trung thực cao.
- Xử lý dữ liệu theo thời gian thực hoặc theo lô.
- Phân tích dữ liệu sử dụng SQL, Python, R, hoặc bất kỳ ngôn ngữ nào khác, dữ liệu của bên thứ ba, hoặc ứng dụng phân tích dữ liệu.

## 1.2. Ứng dụng

Dưới đây là các trường hợp ứng dụng chính của hồ dữ liệu:

- Di chuyển dữ liệu đám mây và IoT
- Xử lý dữ liệu lớn
- Phân tích
- Báo cáo
- Di chuyển dữ liệu tại chỗ



Hình 3. Ứng dụng của data lake

## 1.3. Ưu điểm

**Ưu điểm của hồ dữ liệu:**

- Dữ liệu không bao giờ bị loại bỏ, bởi vì dữ liệu được lưu trữ trong định dạng thô của nó. Điều này đặc biệt hữu ích trong môi trường dữ liệu lớn, khi bạn có thể không biết trước những hiểu biết nào có thể có từ dữ liệu.
- Người dùng có thể khám phá dữ liệu và tạo các truy vấn của riêng mình.
- Có thể nhanh hơn so với các công cụ ETL truyền thống.
- Linh hoạt hơn so với kho dữ liệu, bởi vì nó có thể lưu trữ dữ liệu không có cấu trúc và bán cấu trúc.

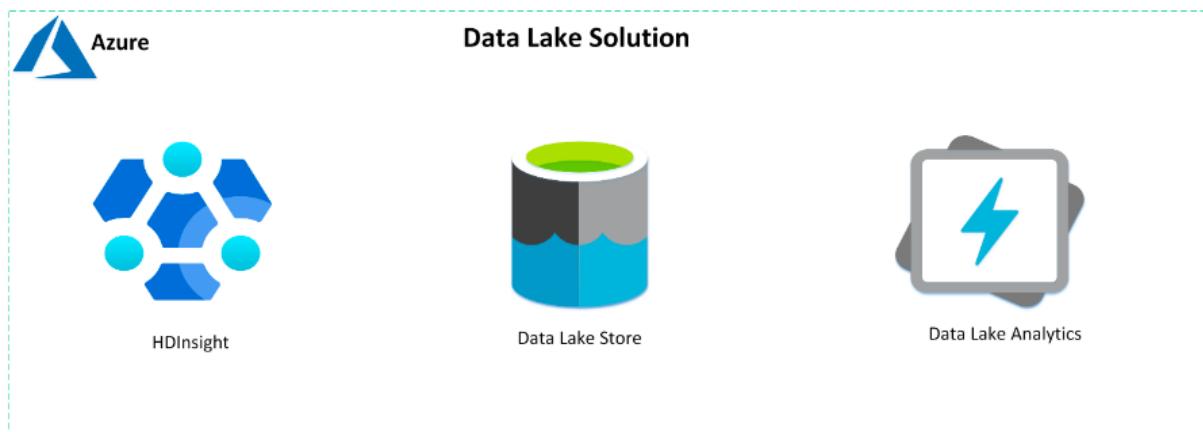
Một giải pháp hồ dữ liệu hoàn chỉnh bao gồm cả lưu trữ và xử lý. Lưu trữ hồ dữ liệu được thiết kế để chịu lỗi, khả năng mở rộng vô hạn, và nhập dữ liệu với tốc độ cao cho dữ liệu có hình dạng và kích thước khác nhau. Xử lý hồ dữ liệu bao gồm một hoặc nhiều công cụ xử lý được xây dựng với những mục tiêu này trong tâm trí, và có thể hoạt động trên dữ liệu được lưu trữ trong hồ dữ liệu ở quy mô lớn.

### **Xây dựng giải pháp hồ dữ liệu sử dụng các dịch vụ do Azure cung cấp:**

- *Azure HDInsight:* Sử dụng Azure HDInsight để thiết lập một dịch vụ phân tích mã nguồn mở, được quản lý đầy đủ, trong môi trường đám mây dành cho doanh nghiệp. HDInsight hỗ trợ nhiều công nghệ trong hệ sinh thái Hadoop như Apache Spark, HBase, Storm, Kafka và nhiều công nghệ khác, cho phép bạn xử lý và phân tích dữ liệu lớn một cách linh hoạt.
- *Azure Data Lake Store:* Lựa chọn Azure Data Lake Store như một kho lưu trữ tương thích với Hadoop, có khả năng mở rộng cao. ADLS được thiết kế để lưu trữ lượng dữ liệu lớn và hỗ trợ lưu trữ dữ liệu ở định dạng thô, cho phép bạn tối ưu hóa việc lưu trữ và truy cập dữ liệu mà không cần lo lắng về kích thước hay định dạng dữ liệu.

- *Azure Data Lake Analytics*: Tận dụng Azure Data Lake Analytics như một dịch vụ công việc phân tích theo yêu cầu để đơn giản hóa việc phân tích dữ liệu lớn. ADLA cho phép bạn viết truy vấn sử dụng U-SQL, một ngôn ngữ truy vấn mở rộng từ SQL với tích hợp mạnh mẽ từ C#, giúp xử lý dữ liệu phức tạp một cách hiệu quả và linh hoạt.

Kết hợp ba dịch vụ này giúp xây dựng một giải pháp hồ dữ liệu mạnh mẽ trên Azure, với khả năng lưu trữ và xử lý dữ liệu lớn, cung cấp công cụ phân tích linh hoạt và mở rộng, đáp ứng nhu cầu của doanh nghiệp về việc thu thập, lưu trữ, quản lý và phân tích dữ liệu ở quy mô lớn.



#### 1.4. Dịch vụ lưu trữ Data Lake Trên Azure – Azure Data Lake Storage

**Azure Storage** là dịch vụ được sử dụng để lưu trữ dữ liệu trên Microsoft Azure. Trong ngữ cảnh của Data Lake, bạn có thể lưu trữ dữ liệu trong Azure Data Lake Storage thông qua các dịch vụ sau:

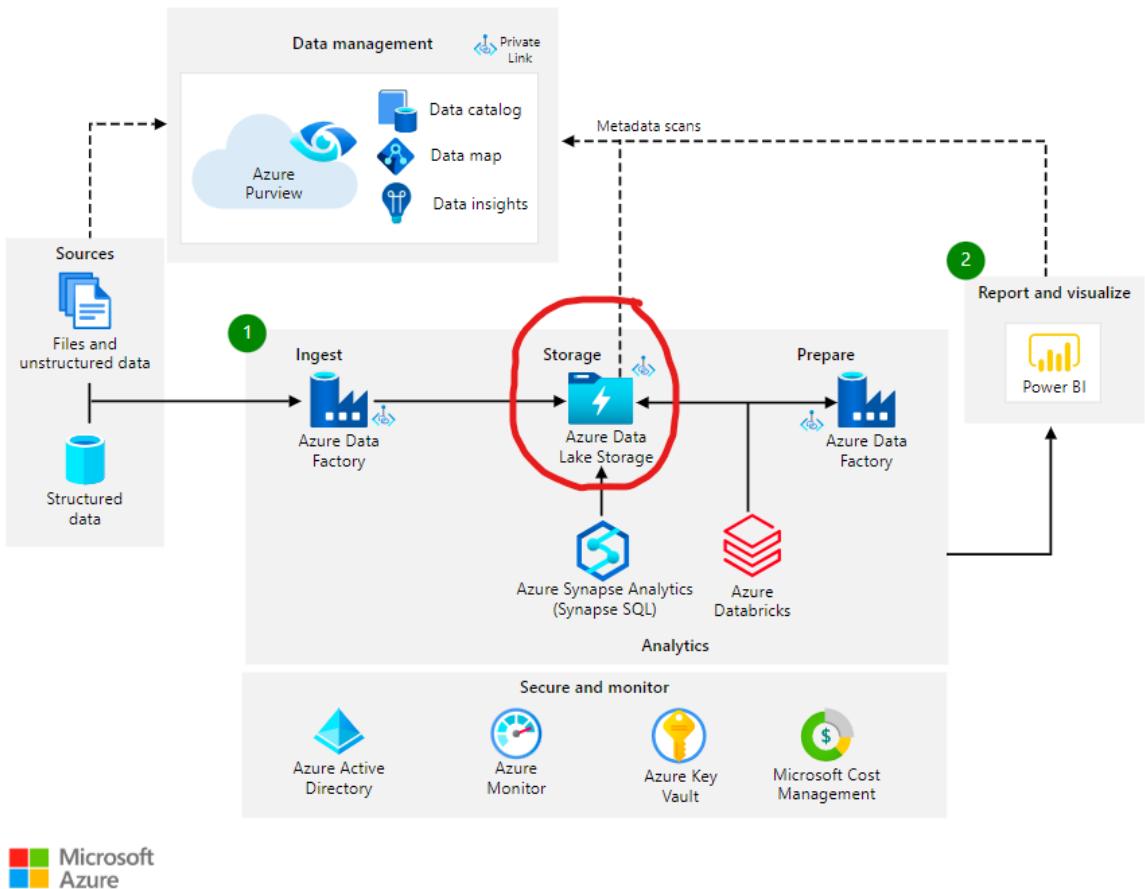
- *Azure Blobs*: Dịch vụ này giúp lưu trữ nhiều loại dữ liệu khác nhau trên đám mây. Blob được tối ưu để lưu trữ một lượng lớn các dữ liệu không có cấu trúc.

trúc như văn bản, dữ liệu nhị phân, hình ảnh, video, và tài liệu. Bạn có thể truy cập các đối tượng lưu trữ trong Blob thông qua HTTP hoặc HTTPS.

Blob thường được sử dụng để:

- Lưu ảnh, tài liệu.
- Lưu các file cần cho truy cập phân tán.
- Streaming video và audio.
- Lưu trữ dữ liệu dùng để sao lưu và phục hồi.
- Lưu trữ dữ liệu dùng để phân tích.

- *Azure Files*: Dịch vụ này cho phép bạn thiết lập một mạng lưới để chia sẻ file và có thể truy cập sử dụng giao thức Server Message Block (SMB). Các máy ảo (VMs) có thể chia sẻ cùng các file với quyền đọc và ghi. Các file cũng có thể truy cập thông qua các REST API hoặc các thư viện khác nhau.
- *Azure Queue*: Dịch vụ này được sử dụng để lưu trữ và lấy lại các tin nhắn. Queue có thể lên đến kích thước 64KB và một queue có thể chứa hàng triệu tin nhắn. Các queue thường được sử dụng để lưu trữ danh sách các tin nhắn cần được xử lý bất đồng bộ.
- *Azure Tables*: Dịch vụ này dùng để lưu trữ các dữ liệu NoSQL.



Hình 4. Azure Data Lake Storage

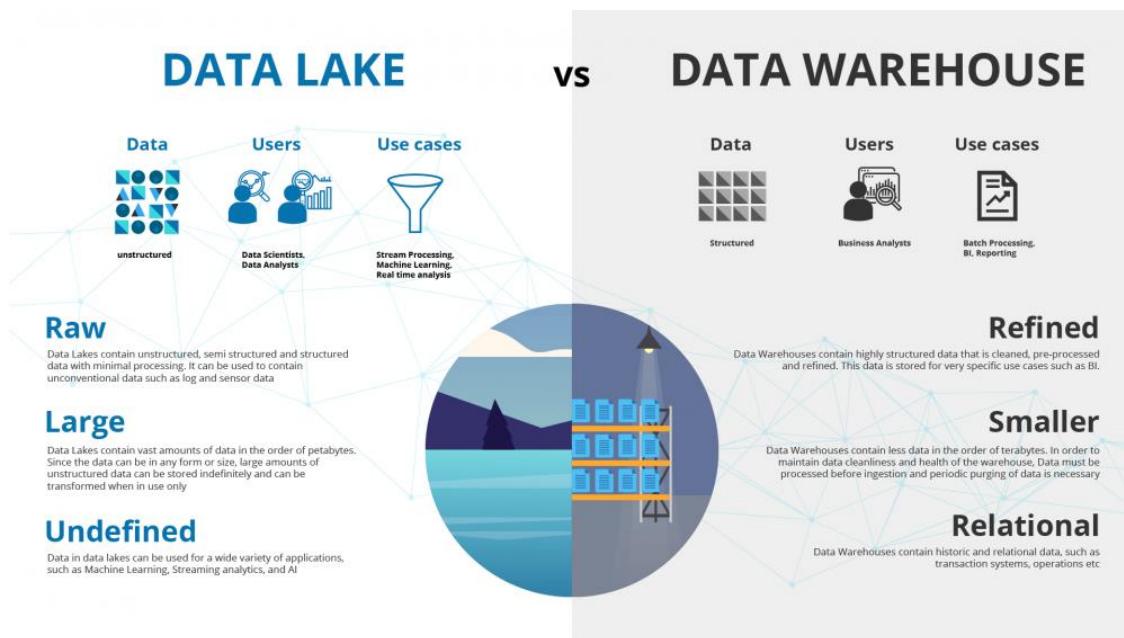
## 1.5. So sánh Data Lake và Data Warehouse

Mặc dù hồ dữ liệu và kho dữ liệu giống nhau ở chỗ đều lưu trữ và xử lý dữ liệu, nhưng mỗi hồ đều có những đặc điểm riêng và do đó có trường hợp sử dụng riêng. Đó là lý do tại sao một tổ chức cáp doanh nghiệp thường đưa hồ dữ liệu và kho dữ liệu vào hệ sinh thái phân tích của họ. Cả hai kho lưu trữ đều hoạt động cùng nhau để tạo thành một hệ thống an toàn, toàn diện để lưu trữ, xử lý và có thời gian tìm hiểu sâu hơn.

- Hồ dữ liệu thu thập cả dữ liệu quan hệ và phi quan hệ từ nhiều nguồn khác nhau—ứng dụng kinh doanh, ứng dụng di động, thiết bị IoT, mạng xã hội hoặc phát trực tuyến—mà không cần phải xác định cấu trúc hoặc lược đồ của

dữ liệu cho đến khi dữ liệu được đọc. Lược đồ khi đọc đảm bảo rằng mọi loại dữ liệu đều có thể được lưu trữ ở dạng thô. Do đó, các hồ dữ liệu có thể chứa nhiều loại dữ liệu khác nhau, từ có cấu trúc, bán cấu trúc đến không cấu trúc, ở mọi quy mô. Bản chất linh hoạt và có thể mở rộng của chúng khiến chúng trở nên cần thiết để thực hiện các dạng phân tích dữ liệu phức tạp bằng cách sử dụng các loại công cụ xử lý điện toán khác nhau như Apache Spark hoặc Azure Machine Learning.

- Ngược lại, kho dữ liệu có bản chất là quan hệ. Cấu trúc hoặc lược đồ được mô hình hóa hoặc xác định trước theo yêu cầu kinh doanh và sản phẩm được quản lý, tuân thủ và tối ưu hóa cho các hoạt động truy vấn SQL. Trong khi hồ dữ liệu chứa dữ liệu của tất cả các loại cấu trúc, bao gồm dữ liệu thô và chưa được xử lý, kho dữ liệu lưu trữ dữ liệu đã được xử lý và chuyển đổi với mục đích cụ thể, sau đó có thể được sử dụng để tạo nguồn báo cáo phân tích hoặc hoạt động. Điều này làm cho kho dữ liệu trở nên lý tưởng để tạo ra các dạng phân tích BI được tiêu chuẩn hóa hơn hoặc để phục vụ trường hợp sử dụng kinh doanh đã được xác định.



Hình 5 So sánh data lake và data warehouse

## 1.6. Ứng dụng

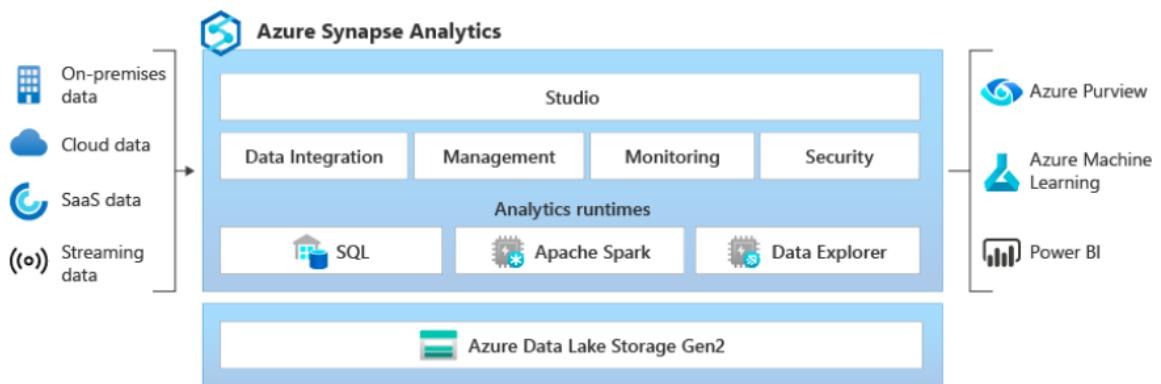
### Một vài ứng dụng của hồ dữ liệu:

- *Truyền thông phát trực tuyến.* Các công ty phát trực tuyến dựa trên đăng ký thu thập và xử lý thông tin chi tiết về hành vi khách hàng, mà họ có thể sử dụng để cải thiện thuật toán đề xuất của mình.
- *Tài chính:* Các công ty đầu tư sử dụng dữ liệu thị trường mới nhất, được thu thập và lưu trữ theo thời gian thực, để quản lý rủi ro danh mục đầu tư một cách hiệu quả.
- *Y tế:* Các tổ chức y tế dựa vào dữ liệu lớn để cải thiện chất lượng chăm sóc cho bệnh nhân. Các bệnh viện sử dụng lượng lớn dữ liệu lịch sử để tinh gọn lộ trình bệnh nhân, dẫn đến kết quả tốt hơn và giảm chi phí chăm sóc.
- *Bán lẻ đa kênh:* Các nhà bán lẻ sử dụng hồ dữ liệu để thu thập và tổng hợp dữ liệu đến từ nhiều điểm chạm, bao gồm di động, xã hội, trò chuyện, truyền miệng và trực tiếp.
- *IoT:* Các cảm biến phần cứng tạo ra lượng lớn dữ liệu bán cấu trúc đến không cấu trúc về thế giới vật lý xung quanh. Hồ dữ liệu cung cấp một kho lưu trữ trung tâm cho thông tin này để lưu trữ cho phân tích tương lai.
- *Chuỗi cung ứng kỹ thuật số:* Hồ dữ liệu giúp các nhà sản xuất tổng hợp dữ liệu kho hàng phân tán, bao gồm hệ thống EDI, XML và JSON.
- *Bán hàng:* Các nhà khoa học dữ liệu và kỹ sư bán hàng thường xây dựng các mô hình dự đoán để giúp xác định hành vi của khách hàng và giảm thiểu tỷ lệ rời bỏ tổng thể

## CHƯƠNG 2. AZURE SYNAPSE ANALYTICS

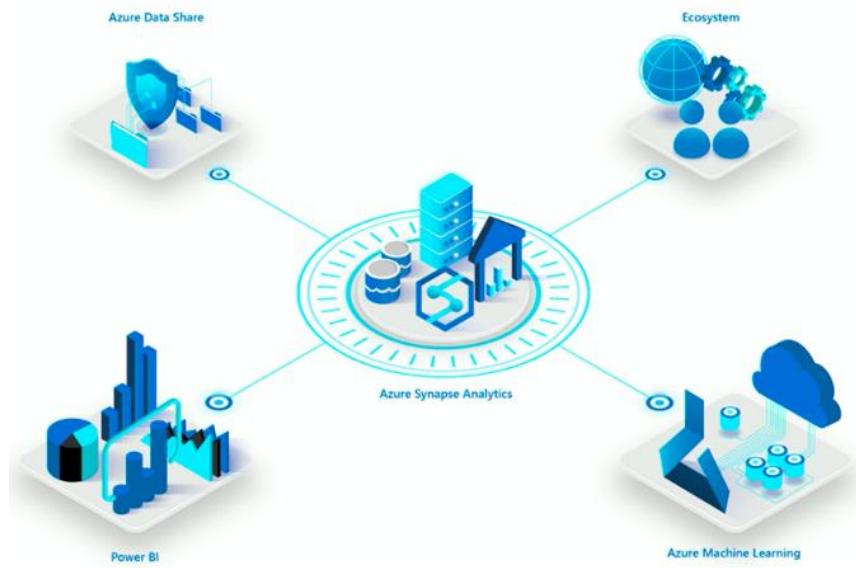
### 2.1. Tổng quan

Azure Synapse là một dịch vụ phân tích doanh nghiệp giúp tăng tốc thời gian để nhận được thông tin từ các kho dữ liệu và hệ thống dữ liệu lớn. Azure Synapse kết hợp những công nghệ SQL tốt nhất được sử dụng trong kho dữ liệu doanh nghiệp, công nghệ Spark được sử dụng cho dữ liệu lớn, Data Explorer cho phân tích nhát ký và chuỗi thời gian, Pipelines cho tích hợp dữ liệu và ETL/ELT, và tích hợp sâu với các dịch vụ Azure khác như Power BI, CosmosDB, và AzureML.



Hình 6. Dịch vụ Azure Synapse Analytics

Nó cung cấp một môi trường thống nhất bằng cách kết hợp kho dữ liệu của SQL, khả năng phân tích dữ liệu lớn của Spark và các công nghệ tích hợp dữ liệu để dễ dàng di chuyển dữ liệu giữa cả hai và từ các nguồn dữ liệu bên ngoài. Chúng tôi có thể nhập, chuẩn bị, quản lý và cung cấp dữ liệu cho nhu cầu BI và máy học trước mắt một cách dễ dàng với Azure Synapse Analytics.



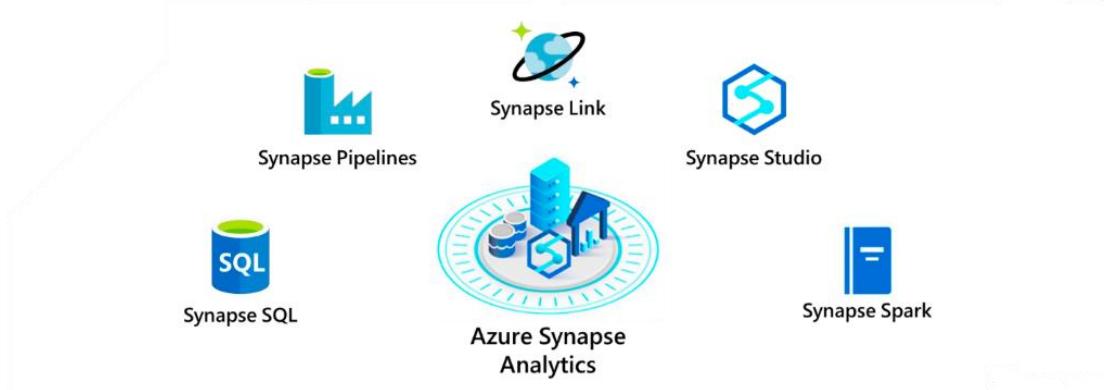
Hình 7 Các ứng dụng sử dụng dịch vụ Azure Synapse Analytics

Như vậy, sau khi lưu trữ các dữ liệu lên hồ sơ dữ liệu Azure Data Lake Storage Gen2, Azure Synapse sẽ là dịch vụ hỗ trợ người dùng và doanh nghiệp tiếp tục quá trình phân tích và khai thác những dữ liệu đó phục vụ kinh doanh

## 2.2. Kiến trúc của Azure Synapse

Kiến trúc Azure Synapse Analytics bao gồm các thành phần:

- *Synapse SQL*
- *Synapse Spark*
- *Synapse Pipelines*
- *Synapse Link*
- *Synapse Studio*



Hình 8. Kiến trúc của Azure Synapse Analytics

**Synapse SQL:** là một dịch vụ phân tích dữ liệu lớn để truy vấn và phân tích dữ liệu. Nó là hệ thống truy vấn phân tán cho phép lưu trữ dữ liệu và ảo hóa dữ liệu. Synapse SQL dựa trên T-SQL (Transact SQL) để truyền dữ liệu. Nó giúp phân tích dữ liệu lớn và sử dụng các giải pháp học máy.

- Có hai thành phần quan trọng trong Synapse SQL: Lớp tính toán(Compute layer) được tách biệt với Lớp lưu trữ(Storage layer) cho phép mở rộng quy mô tính toán độc lập với dữ liệu trong hệ thống
- Lớp tính toán(Compute layer), bao gồm 2 mô hình:
  - Dedicated Synapse SQL Pool
    - Đơn vị quy mô là một trùu tượng của sức mạnh tính toán được biết đến như là một đơn vị kho dữ liệu.
    - Nhận được sức mạnh tính toán thông qua các nguồn tài nguyên tính toán riêng biệt. Nó bao gồm một động cơ xử lý song song khối lượng lớn (MPP) nổi tiếng. Đối với Dedicated Synapse SQL Pools, đơn vị kho dữ liệu (DWU) là đơn vị của quy mô, được kế thừa từ Azure SQL Data Warehouse. Đây là một đơn

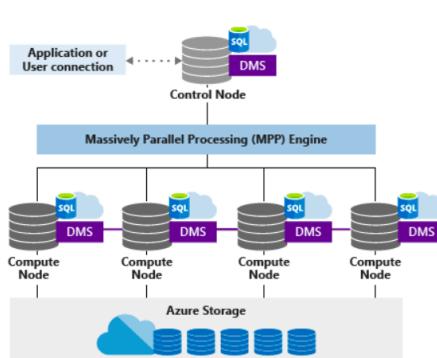
vị trùu tượng để đo lường sức mạnh tính toán cho Dedicated Synapse SQL Pools. Dedicated Synapse SQL sử dụng kiến trúc dựa trên nút. Có hai loại nút khác nhau trong kiến trúc của nó. Loại nút đầu tiên là nút điều khiển, đây là điểm nhập duy nhất cho bất kỳ ứng dụng hoặc người dùng nào. Một khi một ứng dụng hoặc người dùng được kết nối với một nút điều khiển trong Dedicated Synapse SQL Pool, họ có thể phát ra các lệnh T-SQL, sau đó sẽ được gửi đến động cơ MPP để phân phối. Động cơ sẽ phân phối các lệnh hoặc truy vấn đến nhiều nút tính toán, là loại nút thứ hai trong kiến trúc Azure Synapse SQL.

- Serverless Synapse SQL Pool

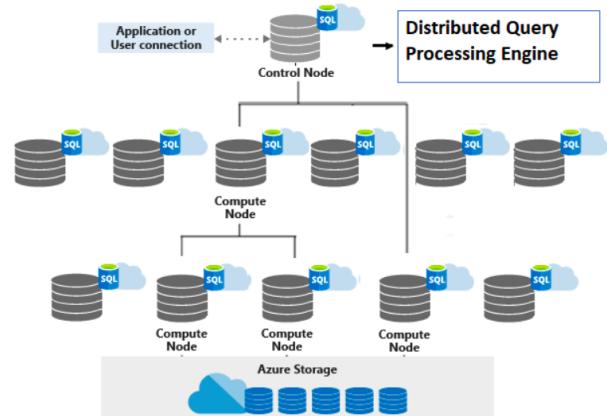
- Việc mở rộng quy mô được thực hiện tự động để đáp ứng yêu cầu tài nguyên của truy vấn. Khi cấu trúc thay đổi theo thời gian bằng cách thêm, loại bỏ các nút hoặc sự cố chuyển đổi, nó thích ứng với những thay đổi và đảm bảo rằng truy vấn có đủ tài nguyên và hoàn thành thành công. Ví dụ hình bên dưới, cho thấy SQL pool không dùng máy chủ sử dụng bốn nút tính toán để thực hiện một truy vấn.
- Không có sức mạnh tính toán riêng biệt, không giống như Dedicated Synapse SQL Pools. Dựa vào các truy vấn hoặc lệnh của bạn, nó sẽ cung cấp các tài nguyên tính toán cho bạn một cách ẩn danh. Vì vậy, không cần phải lo lắng về việc cung cấp tài nguyên tính toán trước, điều này là cần thiết cho Dedicated Synapse SQL Pools. Serverless Synapse SQL Pools sử dụng động cơ xử lý truy vấn phân tán (DQP), khác một chút so với động cơ MPP được sử dụng bởi Dedicated Synapse SQL Pools.

Nó cũng không sử dụng khái niệm DWU, mà dựa vào kích thước của dữ liệu được xử lý bởi truy vấn của bạn để tính phí. Serverless Synapse SQL Pools cũng sử dụng kiến trúc dựa trên nút. Ở đây cũng có hai loại nút khác nhau. Loại nút đầu tiên là nút điều khiển, là điểm nhập duy nhất cho bất kỳ ứng dụng hoặc người dùng nào. Một khi một ứng dụng hoặc người dùng được kết nối với nút điều khiển trong một Serverless Synapse SQL Pool, họ có thể phát ra các lệnh T-SQL, sau đó sẽ được gửi đến động cơ DQP để phân phối.

### Dedicated SQL pool



### Serverless SQL pool



Hình 9. Dedicated SQL pool và Serverless SQL pool

- Synapse SQL sử dụng kiến trúc dựa trên nút. Các ứng dụng kết nối và phát ra các lệnh T-SQL tới một Nút Điều Khiển, đây là điểm nhập duy nhất cho Synapse SQL.
  
  
  
- Nút Điều Khiển Azure Synapse SQL sử dụng một động cơ truy vấn phân tán để tối ưu hóa các truy vấn cho việc xử lý song song, sau đó chuyển các hoạt động sang các Nút Tính Toán để thực hiện công việc một cách song song.

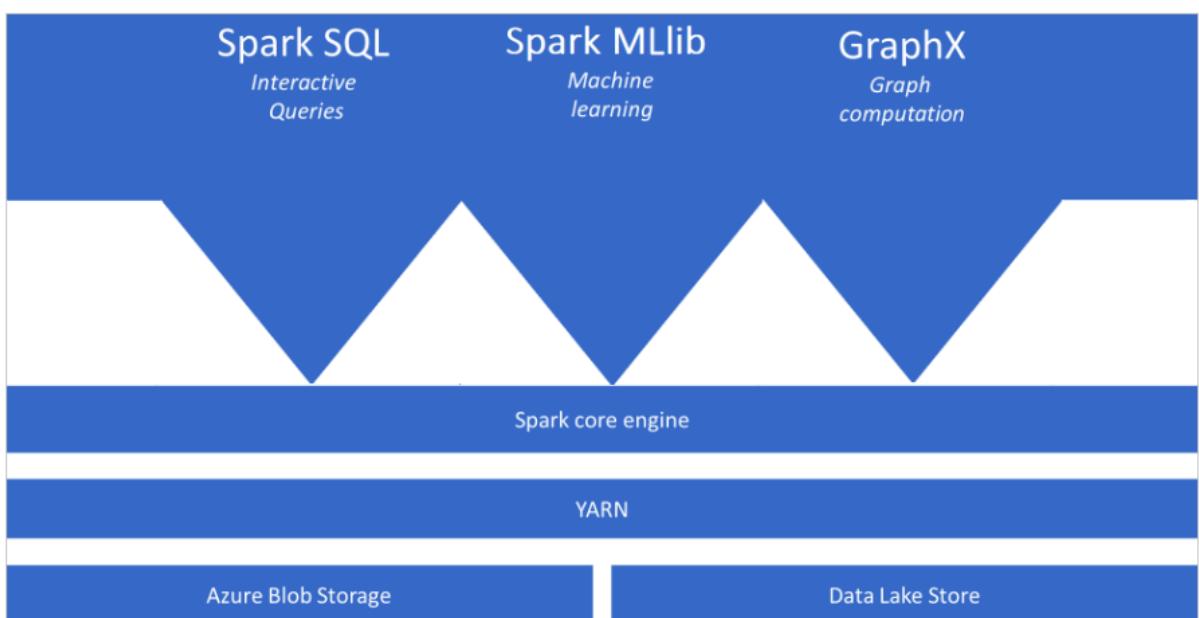
- Nút Điều Khiển của SQL pool không dùng máy chủ sử dụng Động cơ Xử lý Truy vấn Phân tán (DQP) để tối ưu hóa và điều phối việc thực thi phân tán của truy vấn người dùng bằng cách chia nó thành các truy vấn nhỏ hơn sẽ được thực thi trên các Nút Tính Toán. Mỗi truy vấn nhỏ được gọi là nhiệm vụ và đại diện cho đơn vị thực thi phân tán. Nó đọc tệp từ lưu trữ, kết hợp kết quả từ các nhiệm vụ khác, nhóm hoặc sắp xếp dữ liệu lấy từ các nhiệm vụ khác.
- Các Nút Tính Toán lưu trữ tất cả dữ liệu người dùng trong Azure Storage và thực thi các truy vấn song song. Dịch vụ Di Chuyển Dữ liệu (DMS) là một dịch vụ nội bộ cấp hệ thống di chuyển dữ liệu giữa các nút khi cần thiết để thực hiện các truy vấn song song và trả về kết quả chính xác.
- Với việc tách biệt lưu trữ và tính toán, khi sử dụng Synapse SQL, có thể hưởng lợi từ việc kích thước sức mạnh tính toán độc lập bất kể nhu cầu lưu trữ. Đối với SQL pool không dùng máy chủ, việc mở rộng quy mô được thực hiện tự động, trong khi đối với SQL pool chuyên dụng, có thể:
  - Tăng hoặc giảm sức mạnh tính toán, trong một SQL pool chuyên dụng, mà không cần di chuyển dữ liệu.
  - Tạm dừng khả năng tính toán trong khi để dữ liệu nguyên vẹn, vì vậy bạn chỉ phải trả tiền cho lưu trữ.
  - Tiếp tục khả năng tính toán trong giờ hoạt động.
- Lớp lưu trữ(Storage layer):
  - Bất kể quyết định sử dụng lựa chọn lớp Tính toán nào, cần phải có một lớp Lưu trữ để lưu trữ dữ liệu đã được xử lý.
  - Có hai lựa chọn chính cho lớp Lưu trữ: Azure Blob Storage hoặc Azure Data Lake Storage Gen2; Azure Data Lake Storage Gen1

không được hỗ trợ như một lớp Lưu trữ trong Azure Synapse Analytics.

- Lớp Lưu trữ độc lập với lớp Tính toán, vì vậy cả khi lớp Tính toán không hoạt động, dữ liệu vẫn sẽ có sẵn, vì nó được lưu trữ an toàn trong lớp Lưu trữ. Sẽ không mất bất kỳ dữ liệu nào khi bạn tắt lớp Tính toán sau khi lưu trữ dữ liệu trong lớp Lưu trữ.
- Serverless Synapse SQL Pool chỉ cho phép truy vấn dữ liệu được lưu trữ trong lớp Lưu trữ của bạn; không thể nhập dữ liệu vào lớp Lưu trữ sử dụng Serverless Synapse SQL Pool. Tuy nhiên, khi sử dụng Dedicated Synapse SQL Pool, hoàn toàn có thể truy vấn dữ liệu cũng như nhập dữ liệu vào lưu trữ. Khi nhập dữ liệu vào lớp Lưu trữ, dữ liệu sẽ được phân chia thành các phân phối. Điều này giúp đạt được hiệu suất tối ưu từ lớp Lưu trữ.
- Có ba mẫu phân chia khác nhau có sẵn cho bạn, bao gồm Hash, Round Robin và Replicate.
- Đối với Serverless Synapse SQL Pools, chúng ta có một lựa chọn bổ sung cho lớp Lưu trữ - Azure Cosmos DB Analytical Store. Tuy nhiên, điều này chỉ có thể khi sử dụng Azure Synapse Link (đang trong giai đoạn xem trước).

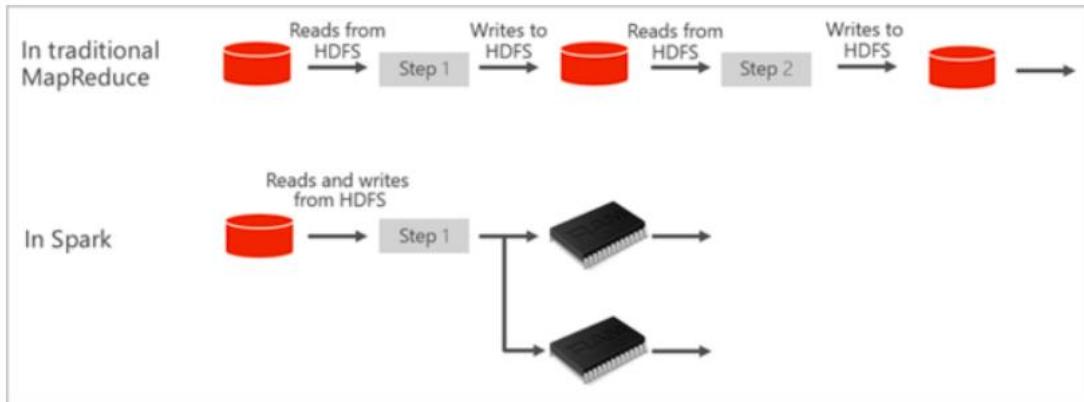
Azure Synapse sử dụng Azure Data Lake Storage Gen2 (ADLS Gen2) làm giải pháp lưu trữ dữ liệu cấp độ tiếp theo để hỗ trợ phân tích dữ liệu khối lượng lớn. ADLS Gen2 kết hợp các tính năng của ADLS Gen1 (như bảo mật cấp độ tệp, mở rộng quy mô và ngữ nghĩa hệ thống tệp) với các tính năng Lưu trữ Azure Blob như lưu trữ theo cấp độ, khắc phục thảm họa và tính sẵn sàng cao.

**Synapse Spark or Apache Spark:** Apache Spark là một khung xử lý song song hỗ trợ xử lý trong bộ nhớ để tăng cường hiệu suất của các ứng dụng phân tích dữ liệu lớn. Apache Spark trong Azure Synapse Analytics là một trong những triển khai của Microsoft về Apache Spark trên đám mây. Azure Synapse giúp việc tạo và cấu hình một nhóm Apache Spark không máy chủ trong Azure trở nên dễ dàng. Các nhóm Spark trong Azure Synapse tương thích với Azure Storage và Azure Data Lake Generation 2 Storage. Vì vậy có thể sử dụng các nhóm Spark để xử lý dữ liệu được lưu trữ trong Azure.



Hình 10. Apache spark

- Apache Spark cung cấp các nguyên tắc cơ bản cho việc tính toán cụm trong bộ nhớ. Công việc Spark có thể tải và lưu trữ dữ liệu vào bộ nhớ và truy vấn nó nhiều lần. Tính toán trong bộ nhớ nhanh hơn nhiều so với ứng dụng dựa trên đĩa. Spark cũng tích hợp với nhiều ngôn ngữ lập trình để bạn có thể thao tác với các tập dữ liệu phân tán giống như các bộ sưu tập cục bộ. Không cần phải cấu trúc mọi thứ dưới dạng các hoạt động banded và giảm.



Hình 11. Sự khác biệt của Apache spark so với mô hình MapReduce truyền thống

- Các Spark pool trong Azure Synapse cung cấp một dịch vụ Spark được quản lý hoàn toàn. Các lợi ích của việc tạo một Spark pool trong Azure Synapse Analytics bao gồm:
  - Tốc độ và hiệu quả
  - Dễ dàng sáng tạo
  - Dễ sử dụng
  - API REST
  - Hỗ trợ cho Azure Data Lake Storage thế hệ 2
  - Tích hợp với IDE của bên thứ ba
  - Thư viện Anaconda được tải sẵn
  - Khả năng mở rộng
- Nhóm Spark trong Azure Synapse bao gồm các thành phần sau:
  - Lõi Spark. Bao gồm Spark Core, Spark SQL, GraphX và MLlib.
  - Anaconda
  - Apache Livy
  - sô ghi chép tương tác

Các trường hợp sử dụng Apache Spark trong Azure Synapse Analytics

- *Data Engineering/Data Preparation*

- Apache Spark bao gồm nhiều tính năng ngôn ngữ để hỗ trợ chuẩn bị và xử lý khôi lượng lớn dữ liệu để dữ liệu có thể trở nên có giá trị hơn và sau đó được sử dụng bởi các dịch vụ khác trong Azure Synapse Analytics. Tính năng này được kích hoạt thông qua nhiều ngôn ngữ (C#, Scala, PySpark, Spark SQL) và cung cấp các thư viện để xử lý và kết nối.

- *Machine Learning*

- Apache Spark đi kèm với MLlib, một thư viện máy học được xây dựng dựa trên Spark mà bạn có thể sử dụng từ nhóm Spark trong Azure Synapse Analytics. Nhóm Spark trong Azure Synapse Analytics cũng bao gồm Anaconda, một bản phân phối Python với nhiều gói khác nhau dành cho khoa học dữ liệu, bao gồm cả học máy. Khi kết hợp với tính năng hỗ trợ tích hợp sẵn cho máy tính xách tay, bạn sẽ có một môi trường để tạo các ứng dụng học máy.

- *Streaming Data*

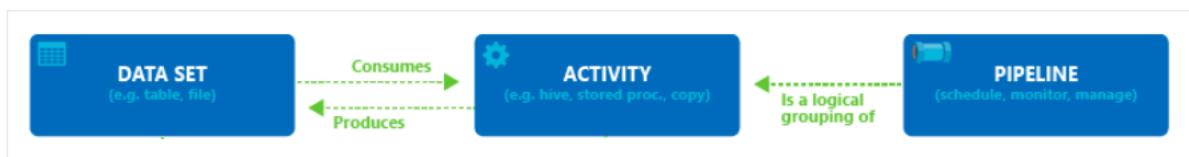
- Synapse Spark hỗ trợ phát trực tuyến có cấu trúc Spark miễn là bạn đang chạy phiên bản được hỗ trợ của bản phát hành thời gian chạy Azure Synapse Spark. Tất cả các công việc đều được hỗ trợ để sống trong bảy ngày. Điều này áp dụng cho cả công việc hàng loạt và công việc phát trực tuyến, đồng thời nhìn chung, khách hàng tự động hóa quá trình khởi động lại bằng cách sử dụng Chức năng Azure.

## Synapse Pipelines

- Một Data Factory hoặc Synapse Workspace có thể có một hoặc nhiều pipeline. Một pipeline là một nhóm logic của các hoạt động cùng thực hiện

một nhiệm vụ. Ví dụ, một pipeline có thể chứa một tập hợp các hoạt động để nhập và làm sạch dữ liệu log, sau đó kích hoạt một luồng dữ liệu ánh xạ để phân tích dữ liệu log. Pipeline cho phép bạn quản lý các hoạt động như một tập hợp thay vì từng cái một. Bạn triển khai và lên lịch cho pipeline thay vì các hoạt động một cách độc lập.

- Azure Data Factory và Azure Synapse Analytics có ba nhóm hoạt động: hoạt động di chuyển dữ liệu, hoạt động biến đổi dữ liệu và hoạt động kiểm soát. Một hoạt động có thể lấy không hoặc nhiều tập dữ liệu đầu vào và tạo ra một hoặc nhiều tập dữ liệu đầu ra. Biểu đồ sau cho thấy mối quan hệ giữa pipeline, hoạt động và tập dữ liệu:

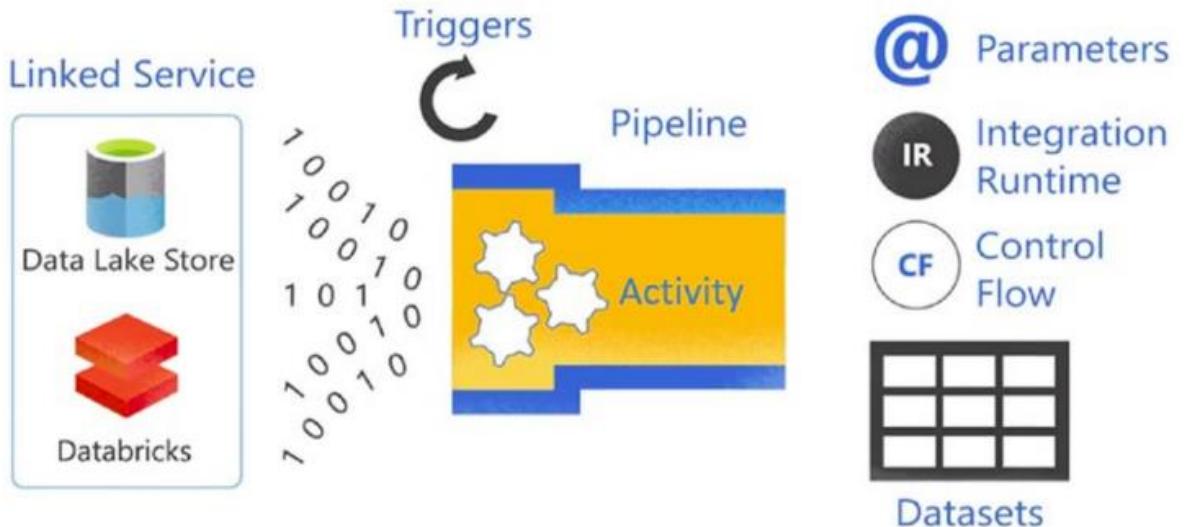


Hình 12. Synapse Pipeline

- Tập dữ liệu đầu vào đại diện cho dữ liệu đầu vào cho một hoạt động trong pipeline, và tập dữ liệu đầu ra đại diện cho dữ liệu đầu ra cho hoạt động đó. Các tập dữ liệu xác định dữ liệu trong các kho lưu trữ dữ liệu khác nhau, như bảng, tệp, thư mục và tài liệu. Sau khi bạn tạo một tập dữ liệu, bạn có thể sử dụng nó với các hoạt động trong một pipeline. Ví dụ, một tập dữ liệu có thể là tập dữ liệu đầu vào/đầu ra của một hoạt động Sao chép hoặc một hoạt động HDInsight Hive.

Gồm các thành phần:

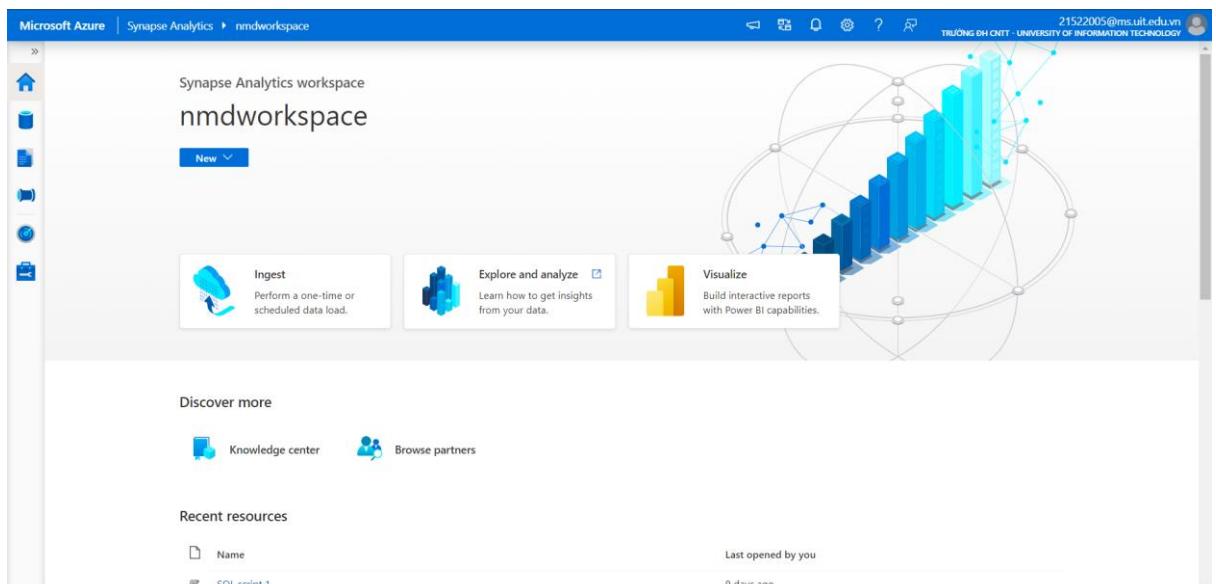
- Data movement activities(Hoạt động di chuyển dữ liệu): Hoạt động Sao chép trong Data Factory sao chép dữ liệu từ một kho lưu trữ nguồn đến một kho lưu trữ đích. Dữ liệu từ bất kỳ nguồn nào cũng có thể được ghi vào bất kỳ nguồn đích nào.
- Data transformation activities(Hoạt động chuyển đổi dữ liệu): Azure Data Factory và Azure Synapse Analytics hỗ trợ các hoạt động biến đổi dữ liệu sau đây có thể được thêm vào một cách cá nhân hoặc được kết nối với một hoạt động khác.
- Control flow activities(Các hoạt động kiểm soát luồng)
  - Ngoài ra, các Pipeline của Synapse là một dịch vụ tích hợp dữ liệu và ETL/ELT dựa trên đám mây là một phần cốt lõi của kiến trúc Azure Synapse Analytics. Bạn có thể tạo các luồng công việc dựa trên dữ liệu để di chuyển và biến đổi dữ liệu theo tỷ lệ lớn bằng cách sử dụng các Pipeline của Synapse. Bạn có thể xây dựng các quy trình ETL hoặc ELT phức tạp để di chuyển và biến đổi dữ liệu bằng cách sử dụng giao diện đồ họa người dùng (GUI). Đó là một dịch vụ dựa trên GUI hoàn toàn không mã; do đó, nó cho phép bạn phát triển các quy trình một cách trực quan.



Hình 13. Các thành phần của Synapse pipeline

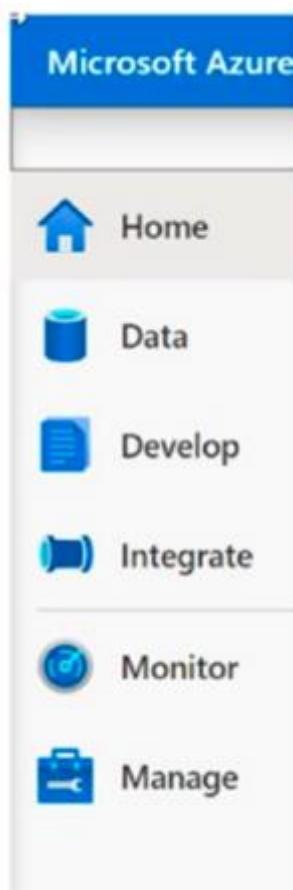
**Synapse Studio:** cung cấp một không gian làm việc cho việc chuẩn bị dữ liệu, quản lý dữ liệu, khám phá dữ liệu, data warehousing doanh nghiệp, big data và các nhiệm vụ AI. Các kỹ sư dữ liệu có thể sử dụng một môi trường trực quan không mã để quản lý các đường ống dữ liệu. Các quản trị cơ sở dữ liệu có thể tự động hóa tối ưu hóa truy vấn. Các nhà khoa học dữ liệu có thể xây dựng các bản chứng minh trong vài phút. Các nhà phân tích kinh doanh có thể truy cập các bộ dữ liệu một cách an toàn và sử dụng Power BI để xây dựng các bảng điều khiển trong vài phút - tất cả đều trong khi sử dụng cùng một dịch vụ phân tích.

Đây là một ranh giới hợp tác có thể bảo mật để thực hiện phân tích doanh nghiệp dựa trên đám mây trong Azure và được triển khai trong một khu vực cụ thể và cũng có một tài khoản và hệ thống tệp ADLS Gen2 liên kết để lưu trữ dữ liệu tạm thời.



Hình 14. Giao diện của Synapse Studio

### Các thành phần trong Azure Synapse Workspace and Studio



- Hub đầu tiên là Home hub, được hiển thị ngay khi bạn khởi chạy Synapse Studio. Ở trung tâm của Home hub là các liên kết đến việc nhập, khám phá, phân tích và trực quan hóa dữ liệu của bạn. Đây đều là các phím tắt đến các công cụ khác nhau có sẵn trong Azure Synapse Analytics.
- Tiếp theo là Data hub, cho phép truy cập vào cơ sở dữ liệu Synapse SQL không máy chủ cũng như cơ sở dữ liệu Synapse SQL Dành riêng. Nó cũng cung cấp quyền truy cập vào nguồn dữ liệu bên ngoài và các dịch vụ liên kết khác bạn đã tạo.
- Hub thứ ba là Develop hub, mà bạn sẽ có thể viết các tập lệnh SQL, ghi chú Synapse, báo cáo, và các công việc khác. Đây là nơi công việc phát triển thực sự sẽ diễn ra, và tất cả các vai trò dữ liệu bao gồm kỹ sư dữ liệu, nhà phân tích dữ liệu, nhà khoa học dữ liệu, và những người khác sẽ truy cập vào hub này để viết mã của họ.
- Hub Integrate chủ yếu là để đưa bạn đến giao diện Synapse Pipelines. Cảm nhận và giao diện giống với Azure Data Factory, vì Synapse Pipelines có kiến trúc tương tự Azure Data Factory. Đây là một hub quan trọng từ góc độ tích hợp vì bạn không cần phải đến Azure Data Factory cho bất kỳ nhu cầu tích hợp dữ liệu và điều phối nào. Hub Integrate cung cấp tất cả các tùy chọn đó trong chính Synapse Studio.
- Hub Monitor dành để cung cấp các tùy chọn theo dõi các pipeline, xem trạng thái của IRs, xem các công việc Synapse Spark, và những điều tương tự, cùng với chi tiết lịch sử về các hoạt động đã diễn ra trong không gian làm việc của bạn.

- Hub Manage cung cấp các tùy chọn để quản lý các SQL pool của Synapse, các pool Spark của Synapse, các dịch vụ liên kết, runtime tích hợp...

**Synapse Link:** với SQL cho phép phân tích gần thời gian thực trên dữ liệu vận hành trong Azure SQL Database hoặc SQL Server 2022. Với tích hợp mượt mà giữa các cửa hàng vận hành bao gồm Azure SQL Database và SQL Server 2022 và Azure Synapse Analytics, Azure Synapse Link for SQL cho phép bạn chạy các kịch bản phân tích, thông tin kinh doanh và học máy trên dữ liệu vận hành của bạn với tối thiểu ảnh hưởng đến cơ sở dữ liệu nguồn với công nghệ change feed mới.

Hình ảnh sau cho thấy tích hợp Azure Synapse Link với Azure SQL DB, SQL Server 2022 và Azure Synapse Analytics:



Hình 15. Azure Synapse Link

Azure Synapse Link for SQL cung cấp trải nghiệm quản lý hoàn toàn và dễ dàng để đưa dữ liệu vận hành vào Azure Synapse Analytics dedicated SQL pools. Điều này được thực hiện bằng cách sao chép liên tục dữ liệu từ Azure SQL Database hoặc SQL Server 2022 với độ nhạy cảm cao. Bằng cách sử dụng Azure Synapse Link for SQL, có thể nhận được các lợi ích sau:

- Ảnh hưởng tối thiểu đến công việc vận hành
- Giảm phức tạp với không cần quản lý công việc ETL
- Thông tin gần thời gian thực vào dữ liệu vận hành

### 2.3. Xây dựng Azure Synapse Analytics

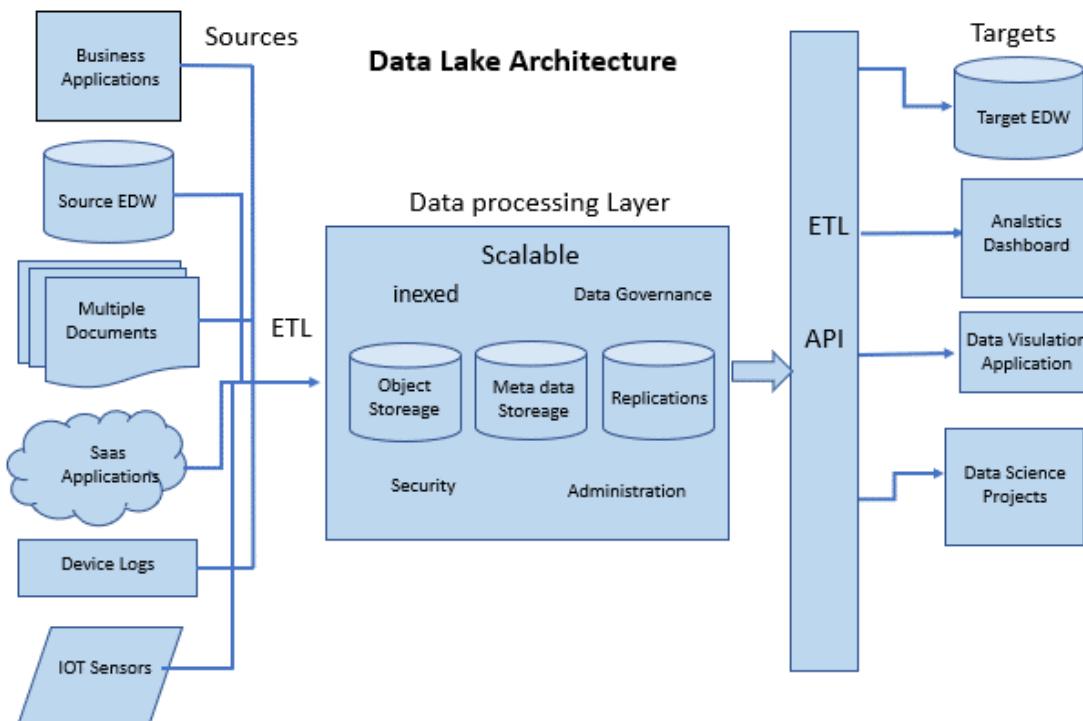
Trong thực tế, ta thấy vẫn có các doanh nghiệp đầu tư để xây dựng các hệ thống để lưu trữ dữ liệu (được gọi là data center).



Hình 16. Hệ thống lưu trữ dữ liệu truyền thống

Như vậy, chúng ta đã biết được cấu trúc của data processing layer trong data lake ở chương 1 cơ bản gồm:

- *Object storage.*
- *Meta data storage*
- *Replications*

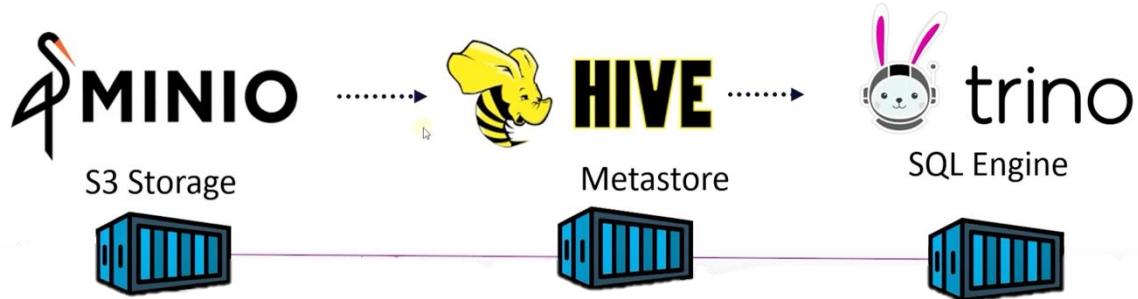


Hình 17. Kiến trúc của hồ dữ liệu

Để thử nghiệm hiệu suất khi sử dụng 1 Local Data lake, mình đã cài đặt 1 demo data lake trên máy cá nhân với những công cụ sau đây:

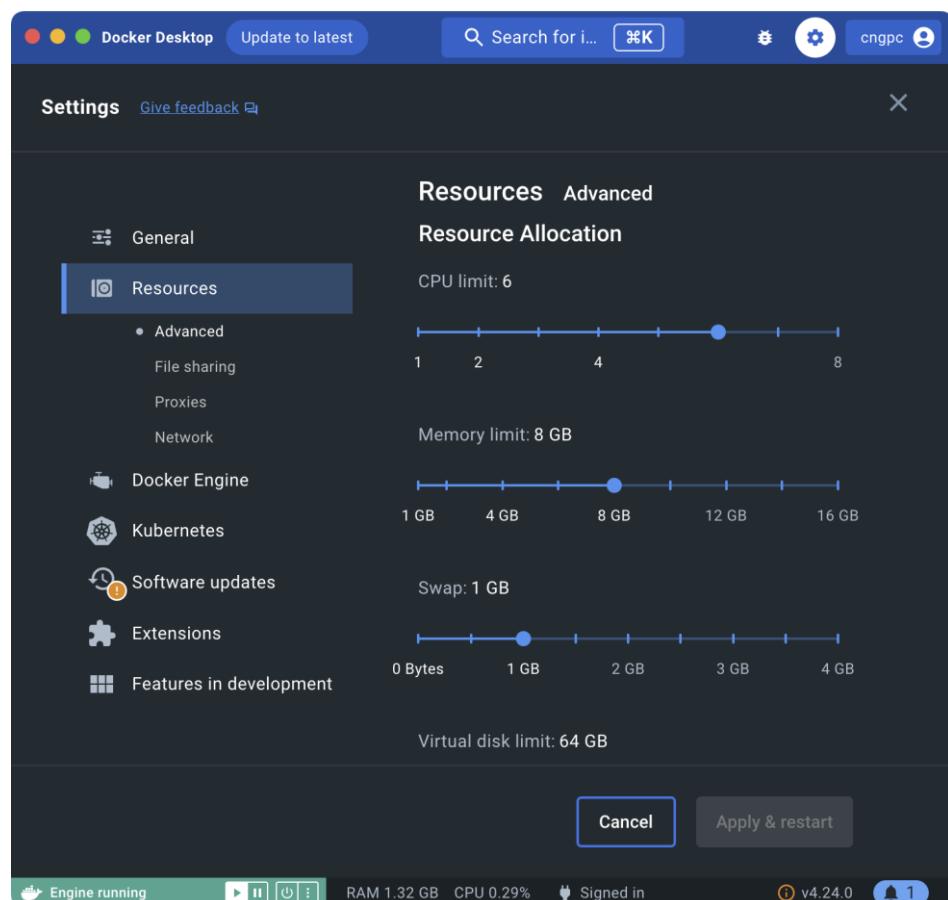
- *MinIO*: hệ thống lưu trữ đối tượng tương thích S3, được sử dụng để lưu trữ dữ liệu trên Data lake, các đơn vị lưu trữ đối tượng tương tự AWS
- *Hive*: đóng vai trò như 1 metastore, cho phép Trino truy cập và thao tác dữ liệu lên các tệp lưu trữ trên MinIO
- *Trino*: công cụ truy vấn SQL phân tán, có khả năng truy cập và thao tác dữ liệu từ nhiều hệ thống lưu trữ. Trong hệ thống này, trino sử dụng Hive metastore để hỗ trợ truy cập dữ liệu SQL đến các tệp được lưu trữ trong MinIO

- *MariaDB*: hệ thống quản lý CSDL quan hệ (RDBMS) để lưu trữ dữ liệu có cấu trúc, cụ thể là Hive metastore.



Hình 18. Các dịch vụ để già lập máy chủ data lake

Để cài đặt các dịch vụ này, nhóm đã sử dụng Docker, bên cạnh đó cấu hình 1 máy chủ với 6 nhân CPU, 8GB Ram.



Hình 19. Cấu hình máy chủ

File cấu hình cho các container hoạt động trong local datalake:

```

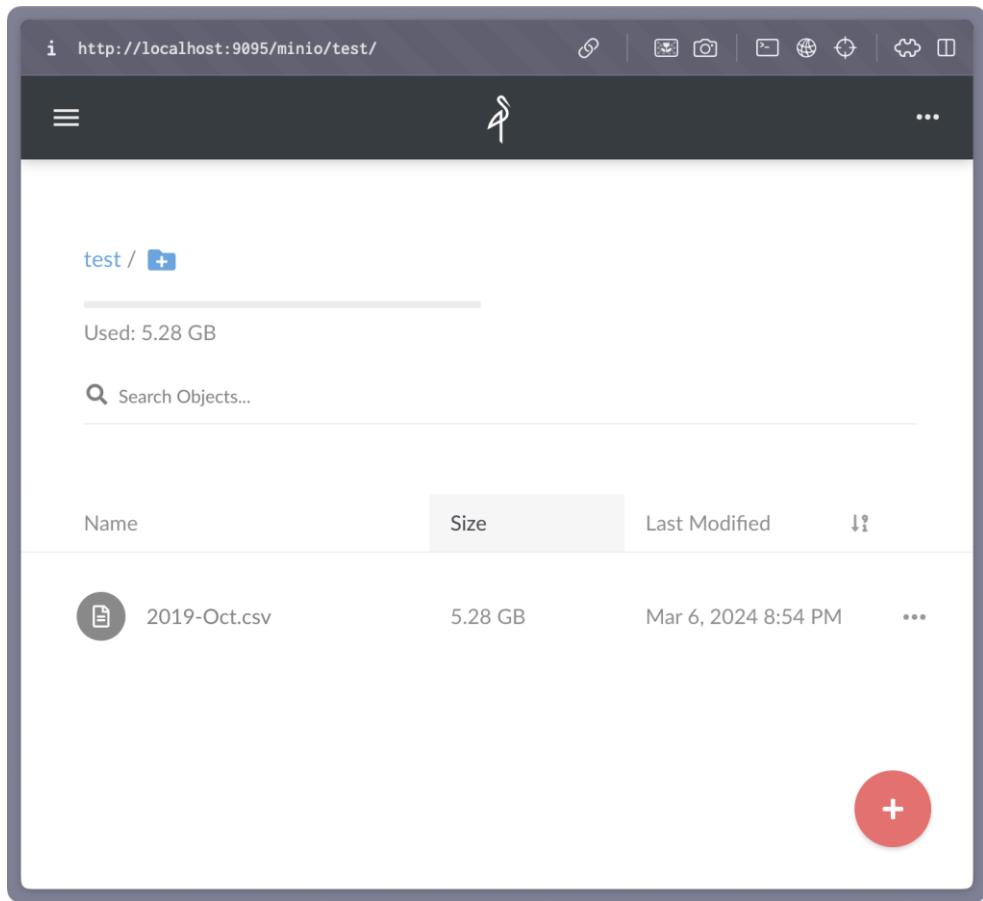
1  version: '3.7'
2  services:
3    trino:
4      hostname: trino
5      image: 'trinodb/trino:351'
6      ports:
7        - '8085:8080'
8      volumes:
9        - ./etc:/usr/lib/trino/etc:ro
10     networks:
11       - trino-network
12
13   mariadb:
14     hostname: mariadb
15     image: mariadb:10.5.8
16     ports:
17       - 3306:3306
18     environment:
19       MYSQL_ROOT_PASSWORD: admin
20       MYSQL_USER: admin
21       MYSQL_PASSWORD: admin
22       MYSQL_DATABASE: metastore_db
23     networks:
24       - trino-network
25
26   hive-metastore:
27     hostname: hive-metastore
28     image: 'bitsondata/hive-metastore:latest'
29     ports:
30       - '9083:9083'
31     volumes:
32       - ./conf/metastore-site.xml:/opt/apache-hive-metastore-3.0.0-bin/conf/metastore-site.xml:ro
33     environment:
34       METASTORE_DB_HOSTNAME: mariadb
35     depends_on:
36       - mariadb
37     networks:
38       - trino-network
39
40   minio:
41     hostname: minio
42     image: 'minio/minio:RELEASE.2021-01-08T21-18-21Z'
43     container_name: minio
44     ports:
45       - '9095:9000'
46     volumes:
47       - minio-data:/data
48     environment:
49       MINIO_ACCESS_KEY: minio_access_key
50       MINIO_SECRET_KEY: minio_secret_key
51     command: server /data
52     networks:
53       - trino-network
54
55   volumes:
56     minio-data:
57       driver: local
58
59   networks:
60     trino-network:
61       driver: bridge
62

```

Hình 20. Kịch bản chạy local data lake trên máy chủ

Local data lake này được sử dụng để lưu trữ 1 file CSV 5GB mô tả hành vi của người tiêu dùng trên 1 website bán sản phẩm công nghệ.

## Tiến hành upload file dữ liệu cần truy vấn lên MinIO



Hình 21. Upload tệp làm việc lên trình lưu trữ đối tượng MinIO

Nhóm đã thực hiện 1 câu truy vấn đơn giản trên dataset và đây là kết quả:

```

-- In ra các mã sản phẩm của apple và lượt mua của người dùng cho sp đó
SELECT product_id, count(user_id)
FROM minio.test.cus_behavior
WHERE event_type = 'purchase' and brand = 'apple'
GROUP BY product_id
LIMIT 100;

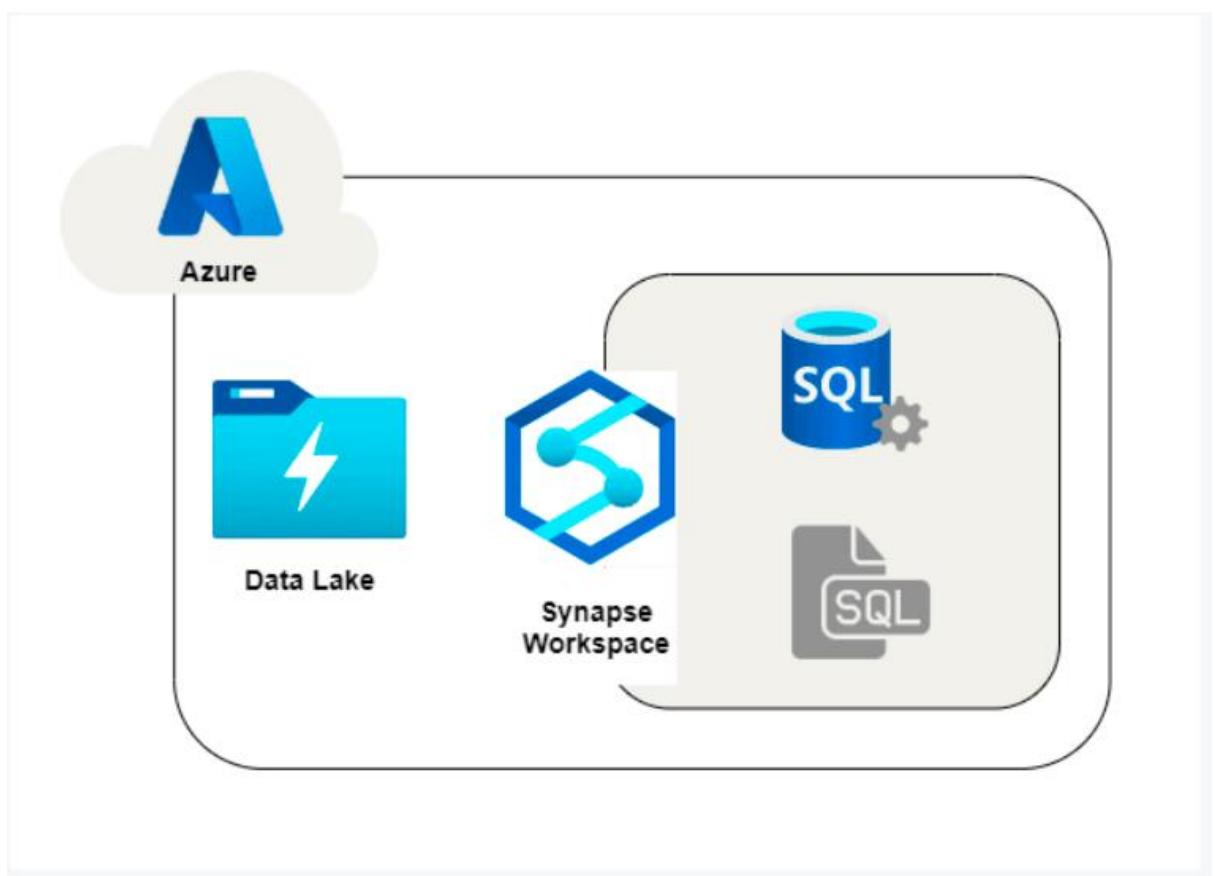
```

product_id	_col1
86	18901489
87	1595976
88	18900778
89	18180793
90	1093928
91	4894056
92	4894055
93	1093312
94	1095139
95	1082548
96	1084356
97	1291388
98	1394297
99	1093318
100	1397055

Hình 22. Thực hiện truy vấn câu lệnh đơn giản với đối tượng

Như vậy, kết quả truy vấn cho 1 câu lệnh đơn giản được trả về sau **6 phút**. Đây được coi là 1 khoảng thời gian khá lâu vì có thể gây ảnh hưởng đến hiệu suất phân tích dữ liệu, tăng chi phí đầu tư ban đầu cho 1 hệ thống mạnh hơn,...

Tương tự, nhóm sẽ thực hiện truy vấn file CSV này trên Azure Synapse Analytics:



Tiến hành tạo Azure Synapse Analytics Workspace để thực hiện truy vấn dữ liệu và để đẩy dữ liệu từ phía máy chủ lên cũng như thực hiện quy trình ETL.

Resource groups

demoresource

Subscription (move) Azure for Students

Subscription ID a25b1cac-fe61-4bcc-a9a1-9db3e27e0dbf

Location Southeast Asia

Tags (edit) Add tags

Resources Recommendations

Name	Type	Location
nmd	Storage account	Southeast Asia
nmdfactory1	Data factory (V2)	East US
<b>nmdworkspace</b>	Synapse workspace	Southeast Asia

Hình 23. Resource groups - các tài nguyên trên Microsoft Azure

Tiến hành upload file CSV 5GB mô tả hành vi của người tiêu dùng trên 1 website bán sản phẩm công nghệ trong Container ở phần Storage Accounts. Ở đây container chứa dữ liệu tên là "synapsesdemo"

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
2019-Oct.csv	3/8/2024, 6:54:05 PM	Hot (Inferred)		Block blob	5.28 GiB	Available
Pivot_table_Data.xls	3/8/2024, 6:08:30 PM	Hot (Inferred)		Block blob	63 Kib	Available
RoadTrafficAccidents_2014.csv	3/9/2024, 9:55:42 AM	Hot (Inferred)		Block blob	371.09 Kib	Available
sales.csv	3/9/2024, 11:30:48 AM	Hot (Inferred)		Block blob	69.94 Kib	Available

Hình 24. Tải tệp làm việc lên container của Synapse workspace

Để truy cập dữ liệu trên Container, ta cần tạo SAS Token và đường dẫn URL đến Container. Người dùng trên Azure sẽ được chọn permissions theo nhu cầu cá nhân,

ở đây, nhóm sẽ chọn tạo token được nhận tất cả quyền để thuận tiện cho demo truy vấn cũng như phân tích dữ liệu.

The screenshot shows the Microsoft Azure Storage accounts interface. A specific container named 'synapsedemo' is selected. On the left, there's a sidebar with options like Overview, Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens (which is currently selected and highlighted in blue), Manage ACL, Access policy, Properties, and Metadata. In the main content area, there's a search bar at the top followed by a detailed configuration form for generating a SAS token. The form includes fields for Signing method (Account key selected), Signing key (Key 1 selected), Stored access policy (None), Permissions (10 selected), Start and expiry date/time, Allowed IP addresses, Allowed protocols (HTTPS only selected), and a large 'Generate SAS token and URL' button at the bottom which is highlighted with a red box.

Hình 25. Tạo SAS Token và URL cho container chứa tệp làm việc

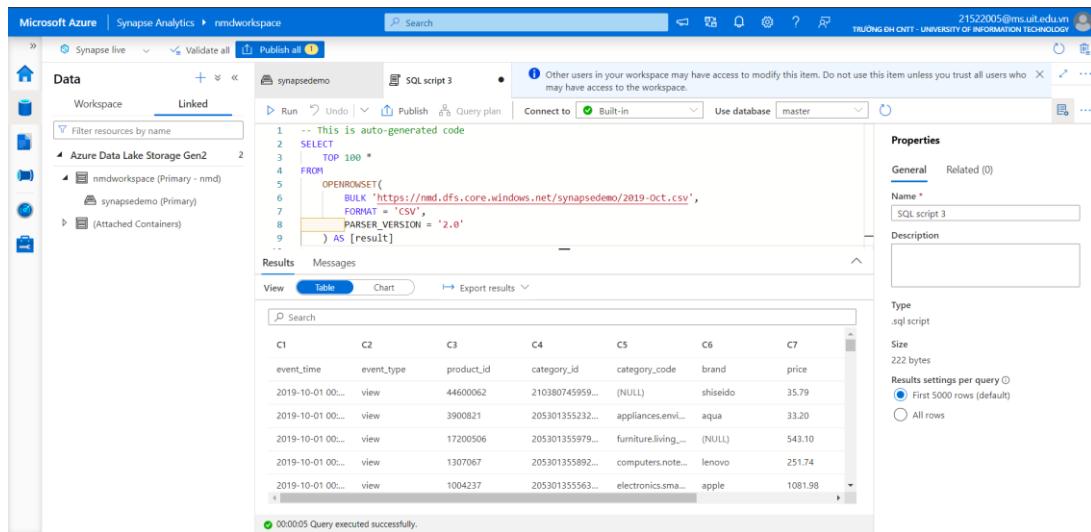
Tại Synapse Studio, trong phần Manage, ta tạo linked service đến container bằng SAS Token và URL.

The screenshot shows the Microsoft Synapse Studio interface under the 'Manage' section. On the left, there's a sidebar with options like Analytics pools, External connections, Linked services (which is currently selected and highlighted in blue), Microsoft Purview, Integration, Security, Access control, Managed private endpoints, Configurations + libraries, Workspace packages, Data flow libraries, and Apache Spark configurations. The main content area shows a list of existing linked services. On the right, a modal dialog box is open for creating a new linked service to 'Azure Data Lake Storage Gen2'. The dialog box has fields for Name (set to 'nmdworkspace-WorkspaceDefaultStorage'), Description, Connect via integration runtime (AutoResolveIntegrationRuntime selected), Authentication type (System Assigned Managed Identity), Account selection method (Enter manually selected), and URL (set to 'https://nmd.dfs.core.windows.net'). At the bottom of the dialog box, there are 'Close' and 'Test connection' buttons, with the 'Test connection' button highlighted with a red box.

Hình 26. Tạo Linked services đến container nơi chứa tệp làm việc

Lúc này đã có file CSV đã được đưa vào Azure Data Storage Gen2 vì đã có Linked Services để kết nối

Tiến hành truy vấn đơn giản để thử nghiệm



```

-- This is auto-generated code
SELECT
TOP 100 *
FROM
OPENROWSET(
    BULK 'https://nmd.dfs.core.windows.net/synapsedemo/2019-Oct.csv',
    FORMAT = 'CSV',
    PARSE_VERSION = '2.0'
) AS [result]

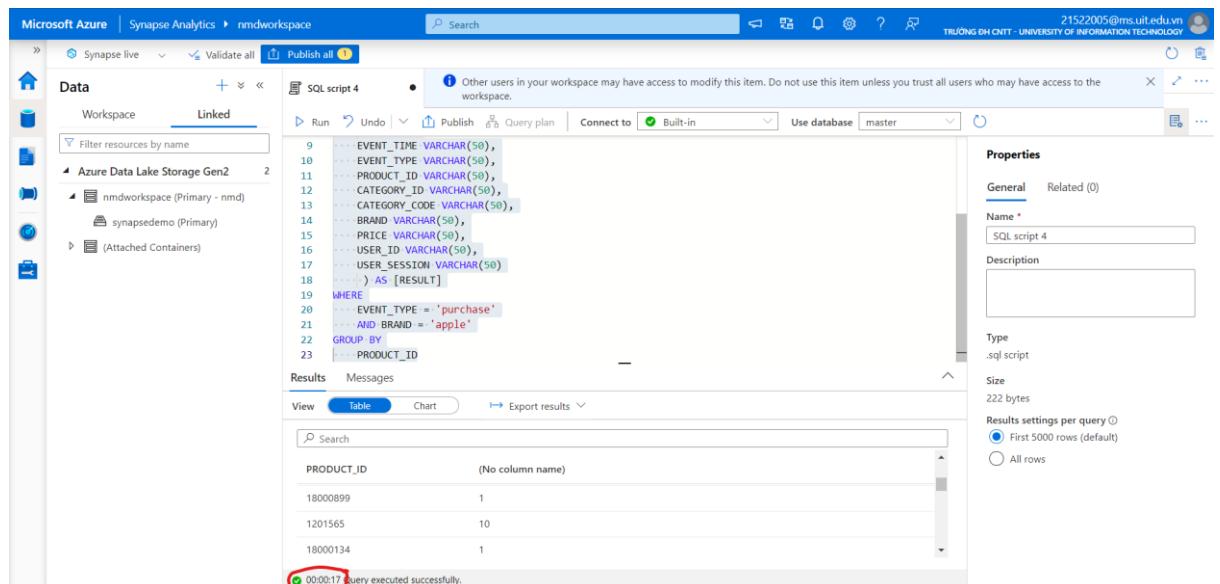
```

C1	C2	C3	C4	C5	C6	C7
event_time	event_type	product_id	category_id	category_code	brand	price
2019-10-01 00:...	view	4460062	210390745959...	(NULL)	shiseido	35.79
2019-10-01 00:...	view	3900821	205301355232...	appliances.env...	aqua	33.20
2019-10-01 00:...	view	17200506	205301355979...	furniture.living...	(NULL)	543.10
2019-10-01 00:...	view	1307067	205301355892...	computers.note...	lenovo	251.74
2019-10-01 00:...	view	1004237	205301355563...	electronics.sma...	apple	1081.98

00:00:05 Query executed successfully.

Hình 27. Thử truy vấn 100 giá trị đầu tiên của tệp làm việc

Và câu truy vấn tương tự trên môi trường local, để so sánh với môi trường trên local



```

EVENT_TIME VARCHAR(50),
EVENT_TYPE VARCHAR(50),
PRODUCT_ID VARCHAR(50),
CATEGORY_ID VARCHAR(50),
CATEGORY_CODE VARCHAR(50),
BRAND VARCHAR(50),
PRICE VARCHAR(50),
USER_ID VARCHAR(50),
USER_SESSION VARCHAR(50)
) AS [RESULT]
WHERE
EVENT_TYPE = 'purchase'
AND BRAND = 'apple'
GROUP BY
PRODUCT_ID

```

PRODUCT_ID	(No column name)
18000899	1
1201565	10
18000134	1

00:00:17 query executed successfully.

Hình 28. Thực hiện truy vấn câu lệnh tương đương trên máy chủ truyền thống để so sánh

Nhận xét: câu truy vấn với thời gian khá nhanh trên môi trường cloud (17s), bé hơn rất nhiều so với môi trường local (6 phút), như vậy cho thấy được hiệu suất phân tích dữ liệu rất nhanh trên môi trường cloud. Khi truy vấn trên môi trường Synapse Analytics, chúng ta không cần cài đặt cấu hình cho máy chủ vì lựa chọn này sẽ được tối ưu dựa trên tập tin làm việc. Hơn nữa, môi trường Synapse Analytics cũng sử dụng các công nghệ Spark để tổng hợp trước các phân tích phổ biến cũng như caching, lập lịch giúp tối ưu thực hiện truy vấn song song

### **Tạo view dựa trên data có sẵn**

Trong SQL Server, View là đoạn lệnh truy vấn đã được viết sẵn và lưu bên trong cơ sở dữ liệu. Một View thì bao gồm 1 câu lệnh SELECT và khi chạy View thì sẽ có kết quả giống như khi mở 1 Table. Tóm lại, View giống như một Table ảo. Bởi vì nó có thể tổng hợp dữ liệu từ nhiều Table để tạo thành 1 Table ảo.

View rất hữu dụng khi bạn muốn cho nhiều người truy cập ở các permission khác nhau. Cụ thể là:

- Hạn chế truy cập tới các Table cụ thể. Chỉ cho phép được xem qua View.
- Hạn chế truy cập vào vào Column của Table. Khi truy cập thông qua View bạn không thể biết được tên Column mà View đó truy cập vào.
- Liên kết các Column từ rất nhiều Table vào thành Table mới được thể hiện qua View.
- Trình bày các thông tin tổng hợp(VD: sử dụng function như COUNT, SUM, ...)

The screenshot shows the Azure Synapse Studio interface. In the left sidebar under 'Data', 'Workspace' is selected. The 'Views' section contains a node for 'dbo.cus\_behavior\_view'. A SQL script window titled 'synapsedemo' shows the following code:

```

1 CREATE VIEW cus_behavior_view
2 AS
3 SELECT *
4 FROM CUS_BEHAVIOR;

```

The 'Properties' pane on the right shows the following details:

- Name:** SQL script 1
- Type:** sql script
- Size:** 986 bytes
- Results settings per query:**
  - First 5000 rows (default)
  - All rows

The 'Results' tab indicates 'No results to show' and 'Your query yielded no displayable results'. A message at the bottom says '00:00:00 Query executed successfully.'

Hình 29. Câu lệnh tạo view trên Azure Synapse Studio

The screenshot shows the Azure Synapse Studio interface. In the left sidebar under 'Data', 'Workspace' is selected. The 'Views' section contains a node for 'dbo.cus\_behavior\_view'. A SQL script window titled 'SQL script 3' shows the following code:

```

1 SELECT TOP (100) [C1]
2 , [C2]
3 , [C3]
4 , [C4]
5 , [C5]
6 , [C6]
7 , [C7]
8 , [C8]
9 , [C9]
10 | FROM [dbo].[cus_behavior_view]

```

The 'Results' tab displays a table with 10 columns (C1-C9 and a final column) and 10 rows of data. The columns are labeled C1 through C7, C9, and C10. The data includes various product names and their prices.

C1	C2	C3	C4	C5	C6	C7	C9	C10
2019-10-01 00...	view	44600062	210380745959...	(NULL)	shiseido	35.79		
2019-10-01 00...	view	3900821	205301355232...	appliances.envi...	aqua	33.20		
2019-10-01 00...	view	17200506	205301355979...	furniture.living...	(NULL)	543.10		
2019-10-01 00...	view	1307067	205301355892...	computers.note...	lenovo	251.74		
2019-10-01 00...	view	1004237	205301355563...	electronics.sma...	apple	1081.98		
2019-10-01 00...	view	1004237	205301355563...	electronics.sma...	apple	1081.98		
2019-10-01 00...	view	1004237	205301355563...	electronics.sma...	apple	1081.98		
2019-10-01 00...	view	1004237	205301355563...	electronics.sma...	apple	1081.98		
2019-10-01 00...	view	1004237	205301355563...	electronics.sma...	apple	1081.98		
2019-10-01 00...	view	1004237	205301355563...	electronics.sma...	apple	1081.98		

A message at the bottom says '00:00:14 Query executed successfully.'

Hình 30. Tạo view cho database db

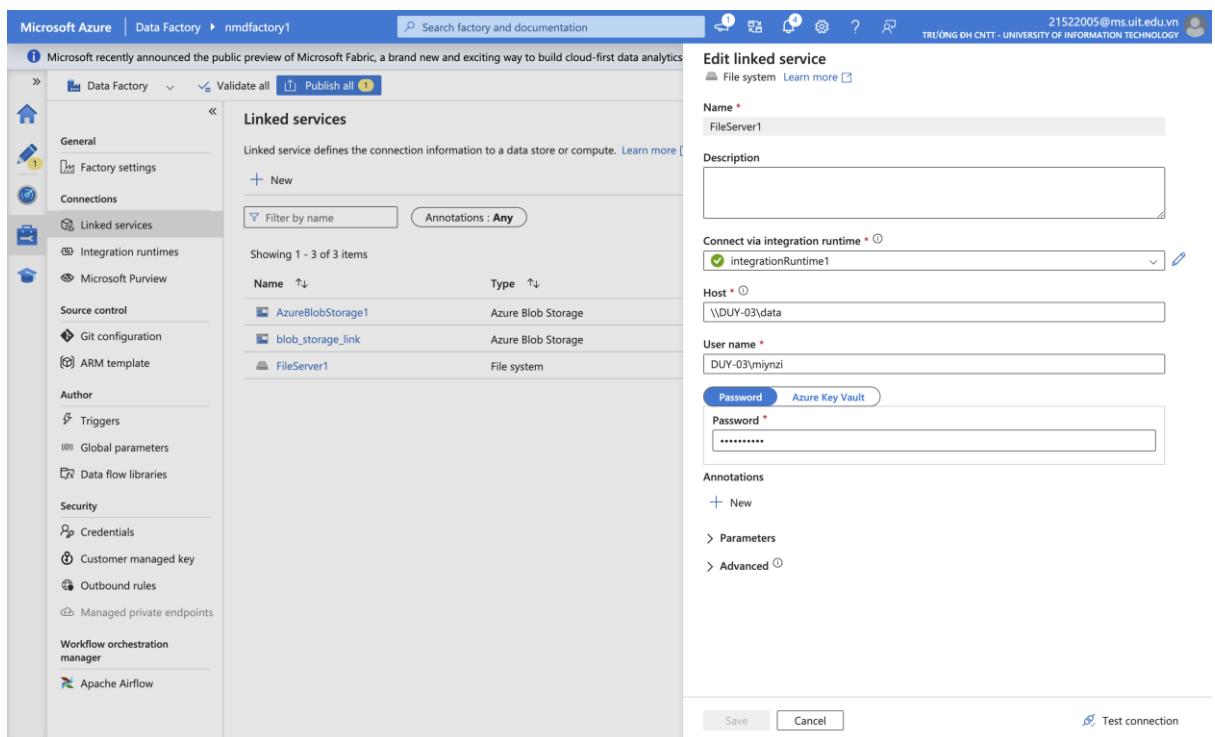
## 2.4. So sánh Data Lake và Data Warehouse



# CHƯƠNG 3. XÂY DỰNG ETL PIPELINE TRÊN AZURE DATA FACTORY

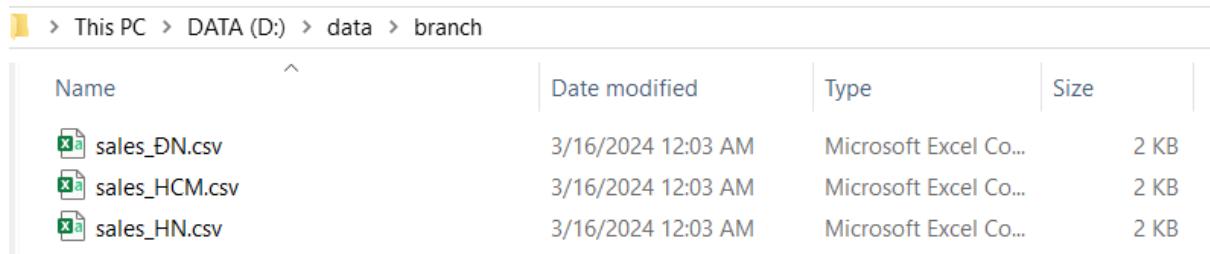
## 3.1. Trường hợp 1

Ví dụ 1 doanh nghiệp có 3 chi nhánh tại Hà Nội, HCM, Đà Nẵng. Dữ liệu bán hàng được lưu dưới định dạng CSV và doanh nghiệp cần phân tích dữ liệu trên cả 3 chi nhánh. Ở phần này, nhóm sẽ thực hiện sử dụng linked service là file system của máy chủ Windows, như vậy, dịch vụ Azure sẽ kết nối với thư mục làm việc đã được cấu hình trên máy Windows để lấy các dữ liệu từ thư mục đó



Hình 31. Cấu hình linked service với máy chủ Windows

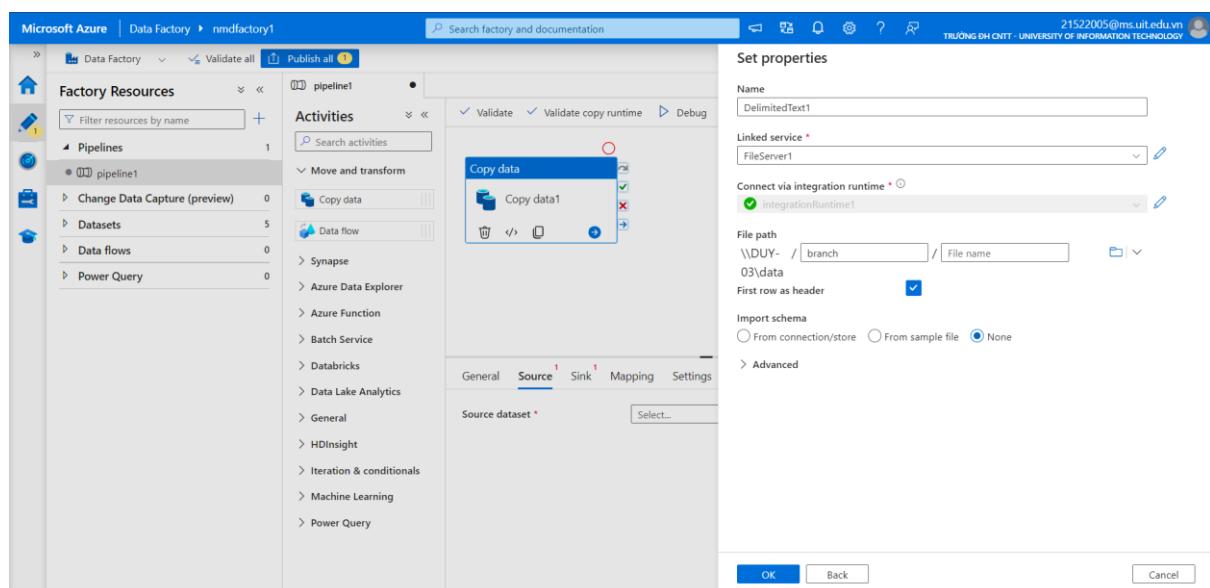
Ở đây, ta cần gộp dữ liệu ở 3 chi nhánh (trên local) để tổng hợp



Name	Date modified	Type	Size
sales_DN.csv	3/16/2024 12:03 AM	Microsoft Excel Co...	2 KB
sales_HCM.csv	3/16/2024 12:03 AM	Microsoft Excel Co...	2 KB
sales_HN.csv	3/16/2024 12:03 AM	Microsoft Excel Co...	2 KB

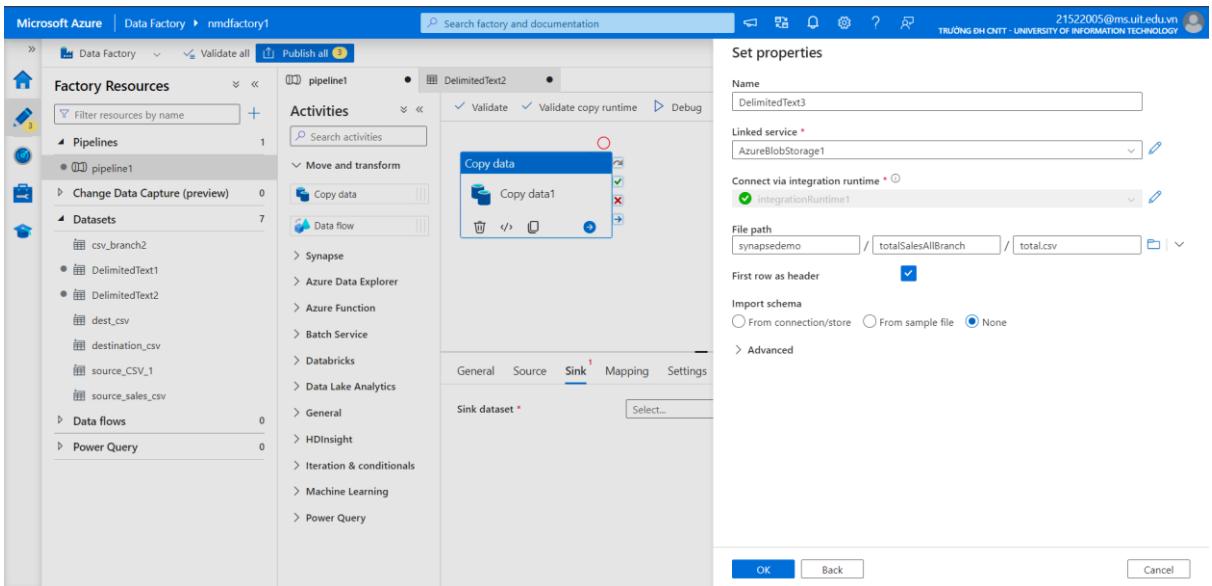
Hình 32. Các tệp trên máy chủ Windows

Tiến hành tạo Pipeline với Copy Data, với nguồn dữ liệu là cả 3 chi nhánh trên



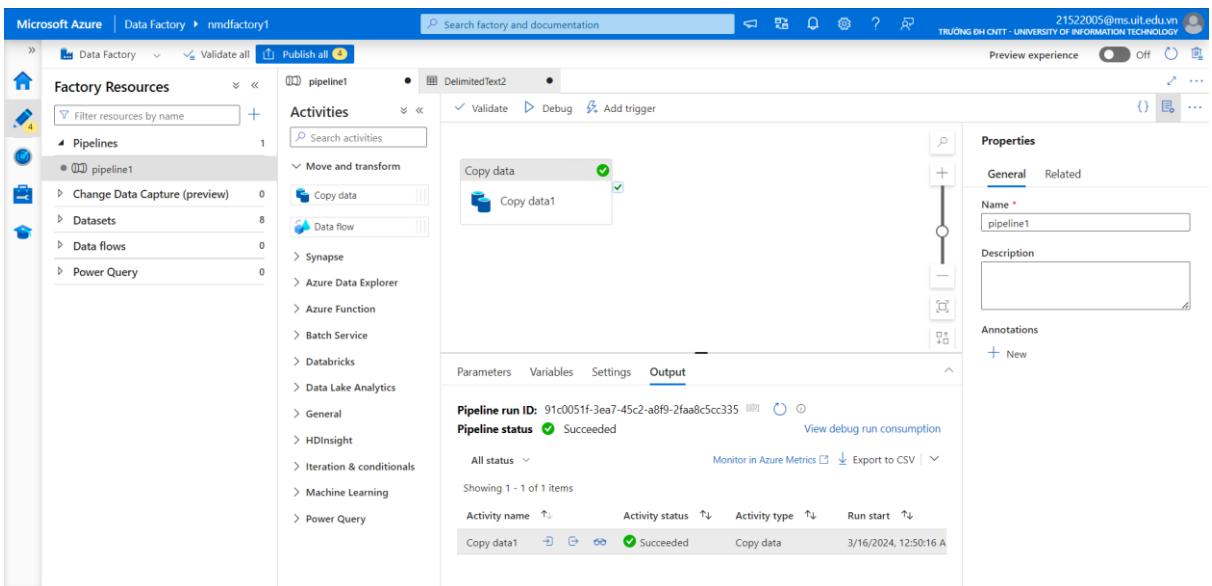
Hình 33. Tạo copy data activity với source là folder "branch" nơi chứa các tệp cần merge

Tiếp đến trỏ thư mục đến Container (Azure Blob Storage) để chứa file tổng hợp



Hình 34. Tệp tin đích sẽ được lưu trên container

Tiến hành Merges File lại với nhau, và kết quả là gộp thành công



Hình 35. Chạy pipeline để merge các tệp lại

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation pane includes 'Data Factory', 'Factory Resources', 'Pipelines' (selected), 'Datasets', 'Data flows', and 'Power Query'. The main area displays a 'Details' card for a pipeline run. The pipeline flow is shown as 'File system' → 'Succeeded' → 'Azure Blob Storage Region: Southeast Asia'. Below this, detailed metrics are provided:

File system	Azure Blob Storage
Data read: 3.538 KB	Data written: 3.692 KB
Files read: 3	Files written: 1
Rows read: 14	Rows written: 14
Peak connections: 1	

Copy duration: 00:00:10  
Throughput: 3.538 KB/s

File system → Azure Blob Storage details:

Start time	3/16/2024, 12:50:17 AM
Used parallel copies	3
Duration	00:00:10
Details	Working duration Total duration
Queue	00:00:05
Transfer	[ Listing source 00:00:00, Reading from source 00:00:00, Writing to sink 00:00:01 ]

Feedback: How satisfied or dissatisfied are you with the performance of this copy activity?

Hình 36. Chi tiết của lệnh chạy pipeline

Ta thu được file tổng hợp dữ liệu ở 3 chi nhánh

The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar lists 'Home', 'synapsesdemo' (Container), 'Overview', 'Diagnose and solve problems', and 'Access Control (IAM)'. The main area shows a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
total.csv	3/16/2024, 12:50:29 ...	Hot (Inferred)		Block blob	3.61 KiB	Available

Hình 37. Tệp tin đích sau khi gộp

The screenshot shows the Microsoft Azure Storage account interface for the 'synapsedemo' container. On the left, there's a sidebar with options like Overview, Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens, Manage ACL, Access policy, Properties, and Metadata. The main area displays a table titled 'totalSalesAllBranch/total.csv'. The table has columns: Name, File Type, Location, Size, Last Modified, Content Type, and ETag. There are four rows of data, each with a circled 'HCM' label in the last column.

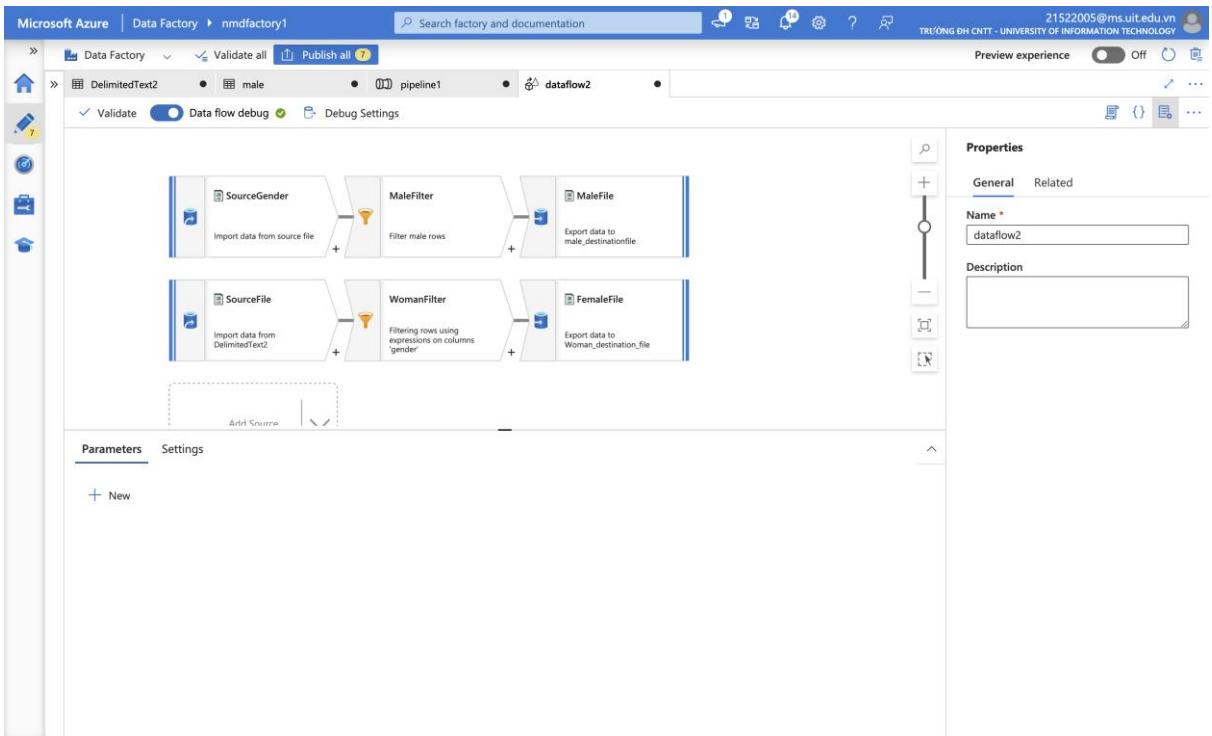
Name	File Type	Location	Size	Last Modified	Content Type	ETag	
file	Chicago	Illinois	60610	Central	OFF-BI-10002609	Office Supplies	Binders Avery Hidden Tab Dividers for Binding Systems 1.788 HCM
nsumer	Rochester	Minnesota	55901	Central	OFF-PA-10004040	Office Supplies	Paper Universal Premium White Copier/Laser Paper (20lb. and 87 Bright) 23.92 HCM
nsumer	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases Bush Somerset Collection Bookcase 261.96 HN
nsumer	Akron	Ohio	44312	East	OFF-PA-10002666	Office Supplies	Paper Southworth 25% Cotton Linen-Finish Paper & Envelopes 21.744 DN
nsumer	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.94 HN

Hình 38. Chi tiết của tệp đích sau khi gộp dữ liệu

### 3.2. Trường hợp 2

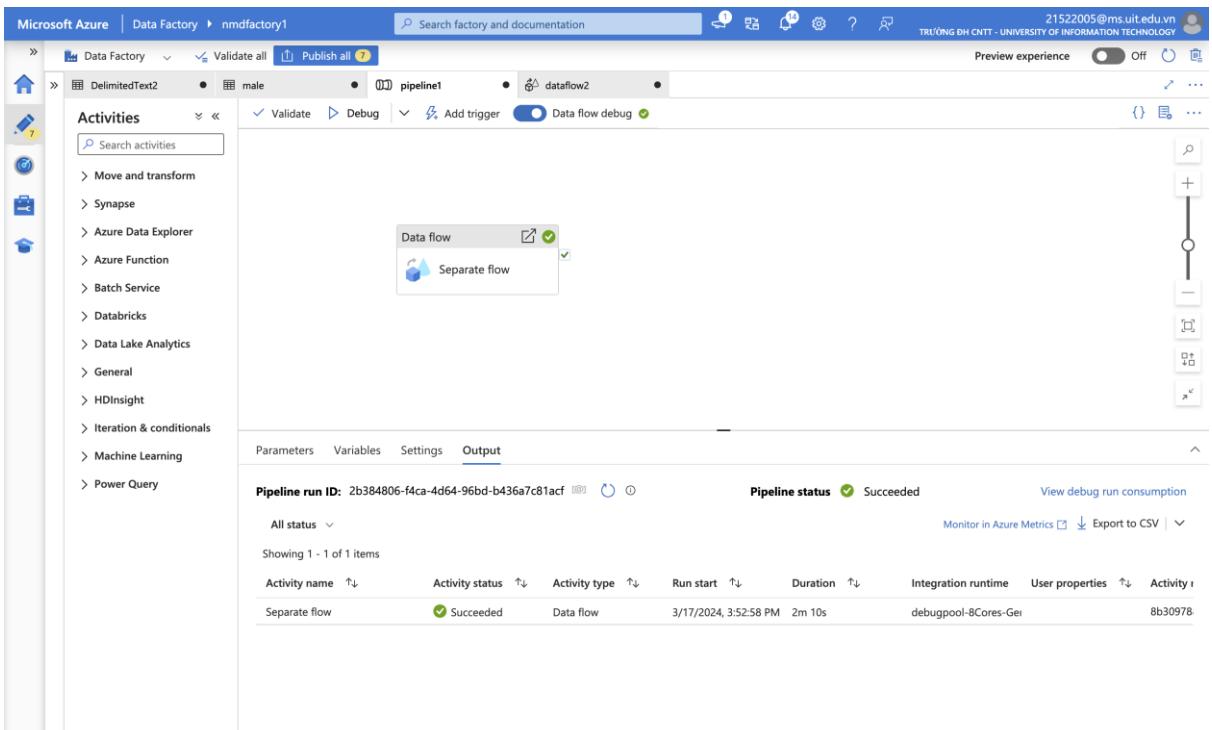
Giả sử tại 1 chi nhánh nào đó, chúng ta có thông tin của các cá nhân với giới tính khác nhau và bây giờ chúng ta cần tách riêng thông tin của nam và nữ để thực hiện phân tích dữ liệu.

Ở đây, nhóm thực hiện tạo 1 data flow để filter các bộ dữ liệu với điều kiện là nam và nữ để tách ra 2 file csv mới.



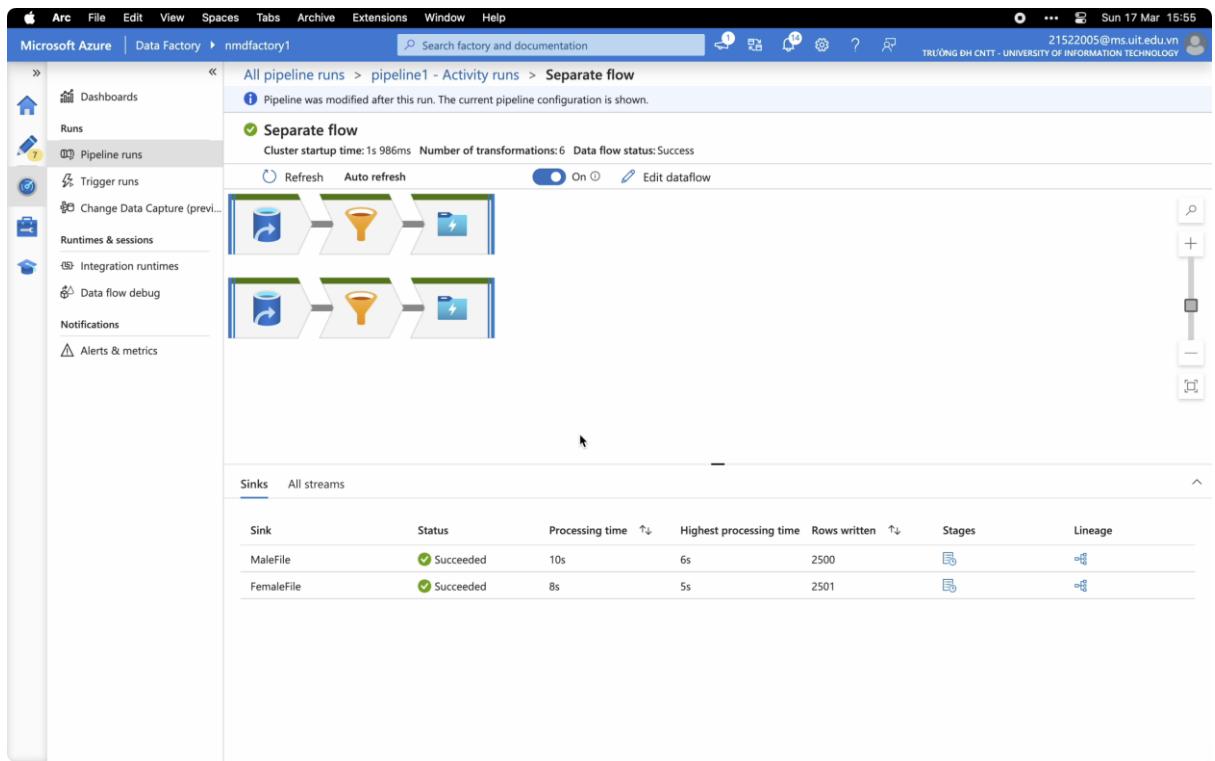
Hình 39. Pipeline để tách dữ liệu dựa trên thuộc tính cột

Thời gian thực thi cho data flow là 2 phút 10s



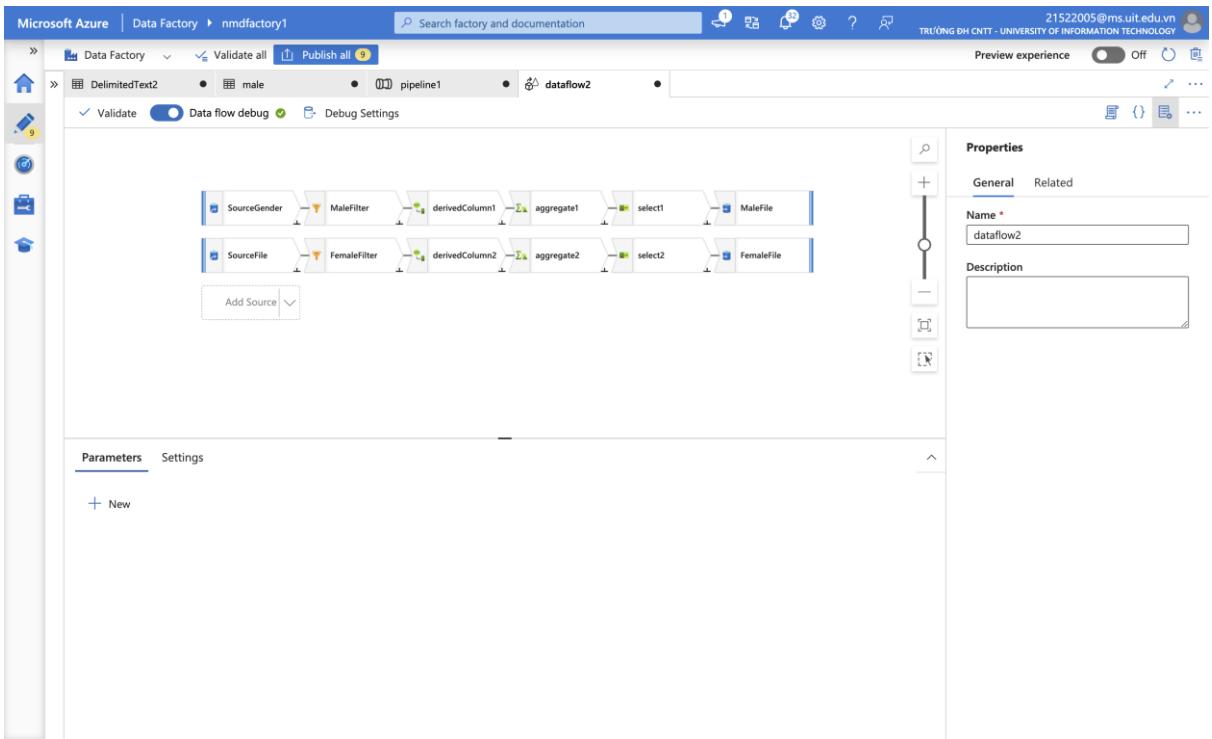
Hình 40. Thực thi luồng xử lý

Sau đây là kết quả thực thi pipeline



Hình 41. Kết quả chi tiết của luồng xử lý

Tiếp theo, nhóm sẽ thử phân tích các đặc điểm unique của mỗi giới tính và thử chuyển đổi bằng file CSV.



Hình 42. Luồng xử lý để tách và chuyển đổi tệp tin thành định dạng khác

Sau khi filter data source thì lúc này chúng ta có 2 đối tượng CSV nam và nữ. Vì vậy khi map sang file json, nhóm sẽ tạo 1 column mới tên là male và female để chứa các thuộc tính còn lại. Khi cột này chuyển sang định dạng JSON sẽ trở thành mảng Male và Female chứa các object tương ứng

The screenshot shows the Microsoft Azure Data Factory Dataflow expression builder interface. On the left, there's a sidebar with icons for Home, Edit, and Refresh, followed by a tree view under 'Derived Columns'. A '+' button for 'Create new' is visible. The main area has a 'Column name \*' input field containing 'Male'. Below it is an 'Expression' editor with a code snippet:

```
@{long_hair=long_hair,
 forehead_height_cm=forehead_height_cm,
 forehead_width_cm=forehead_width_cm,
 nose_wide=nose_wide,
 nose_long=nose_long,
 lips_thin=lips_thin,
 distance_nose_to_lip_long=distance_nose_to_lip_long,
```

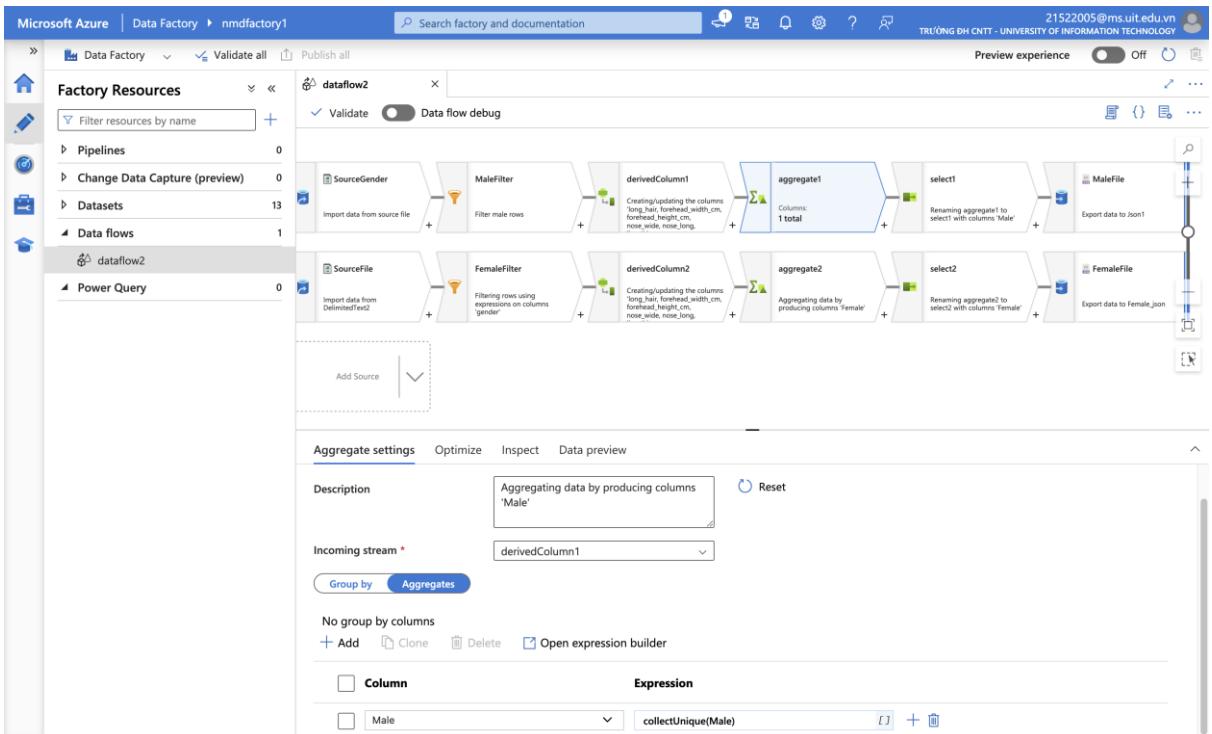
Below the expression editor is a toolbar with various operators (+, -, \*, /, ||, &&, !, ^, ==, !=, >, <, >=, <=, []). To the right of the expression editor is a 'Save' button.

On the left side of the main panel, there are two columns: 'Expression elements' (listing 'All', 'Functions', 'Input schema', 'Parameters', 'Cached lookup', 'Data flow library functions', and 'Locals') and 'Expression values' (listing 'long\_hair', 'forehead\_width\_cm', 'forehead\_height\_cm', 'nose\_wide', 'nose\_long', and 'lips\_thin').

At the bottom of the screen, there are buttons for 'Save and finish', 'Cancel', and 'Clear contents'.

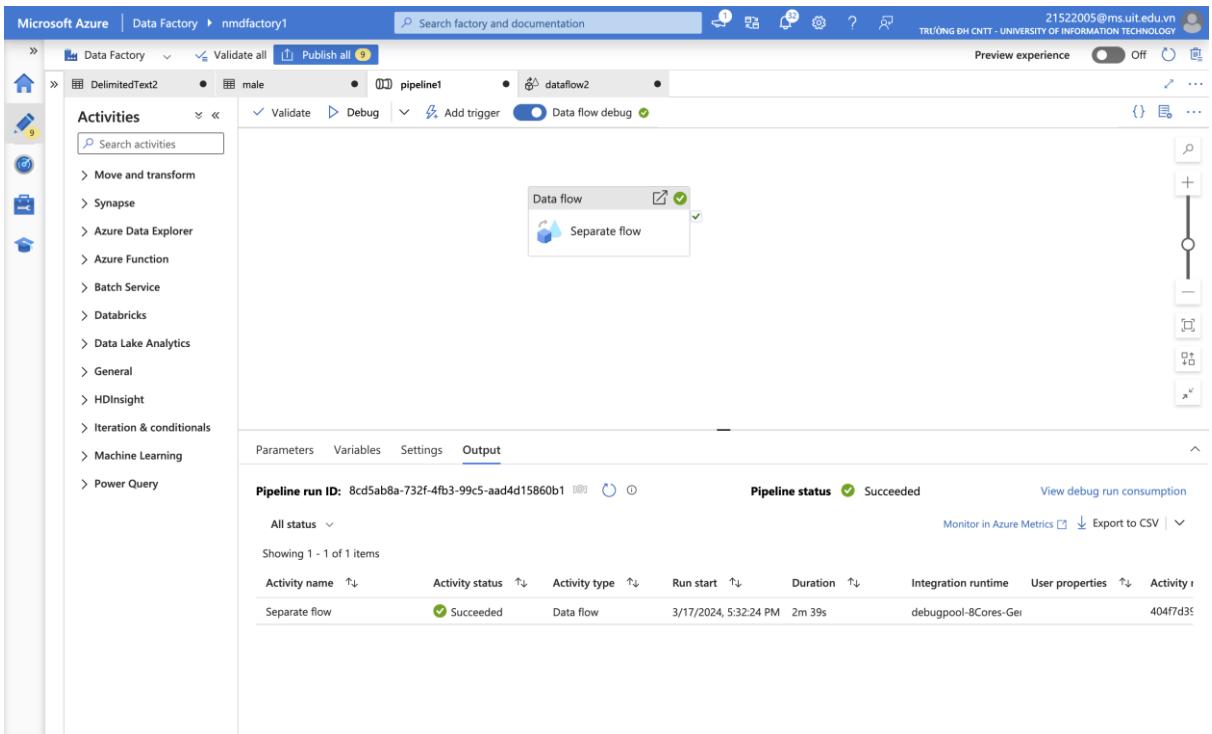
Hình 43. Để tạo 1 mảng Male trong Json, chúng ta sẽ tạo 1 cột mới chứa object cũ

Khi tổng hợp dữ liệu (aggregate data), nhóm đã kết hợp lấy những object unique để lọc những nhóm ngoại hình trùng lặp



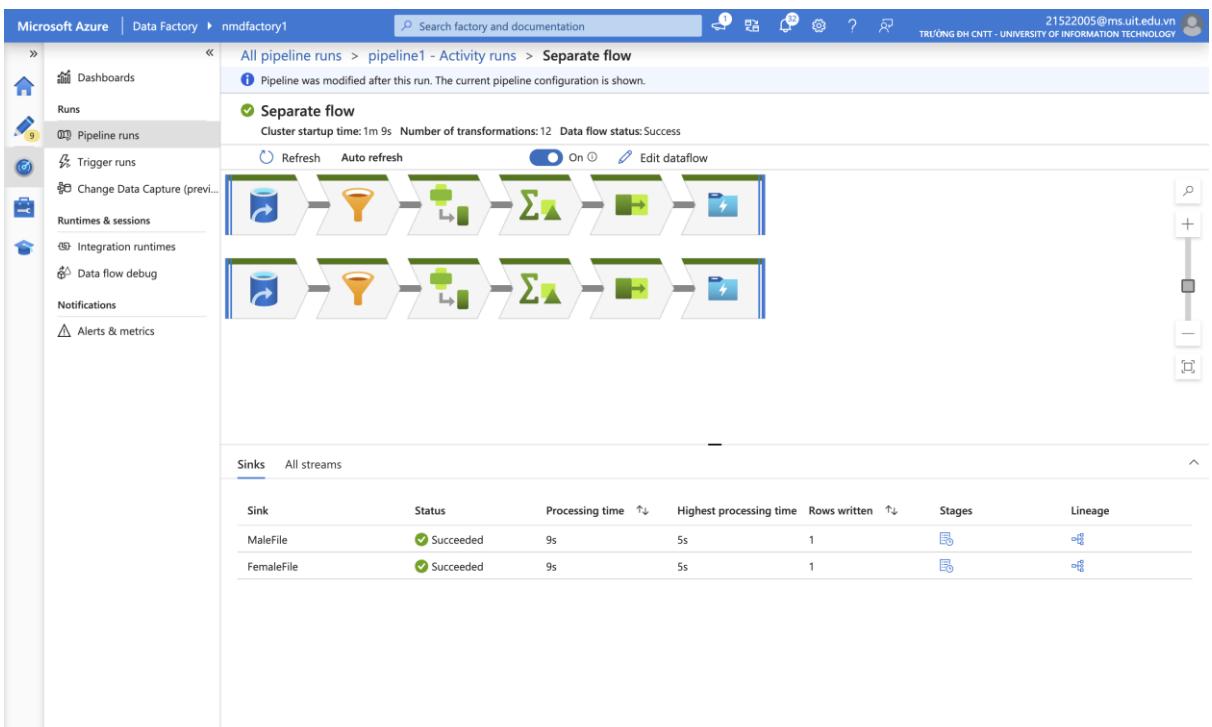
Hình 44. Lọc những đối tượng unique trong dataset

Kết quả chạy pipeline mất 2 phút 39 giây



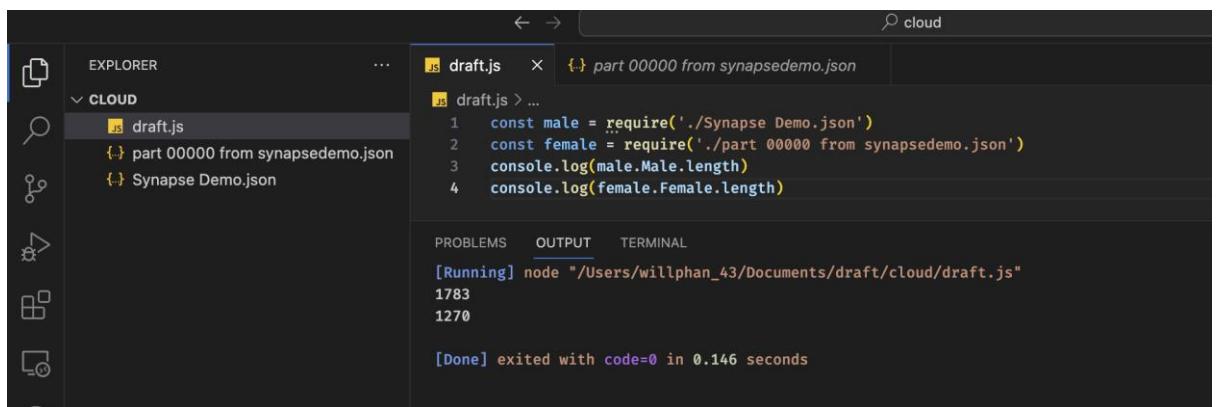
Hình 45 Kết quả thực thi luồng xử lý

Pipeline đã tạo ra 2 tệp đích Json cho male và female



Hình 46. Kết quả chi tiết của luồng xử lý

Bởi vì, pipeline chỉ collect những unique object nên so kết quả bên trên, ta thấy chiều dài của mảng đã giảm.



```
const male = require('./Synapse Demo.json')
const female = require('./part 00000 from synapsedemo.json')
console.log(male.Male.length)
console.log(female.Female.length)
```

[Running] node "/Users/willphan\_43/Documents/draft/cloud/draft.js"  
1783  
1270  
[Done] exited with code=0 in 0.146 seconds

Hình 47. Chiều dài của mảng sau khi chỉ chọn những đối tượng unique

Bây giờ, ta đã có 1 tệp tin dữ liệu mảng Male như sau:

```
[root@kali ~]# node ./script/writeJSON.js > ./data.json
```

```
{  
    Male: [  
        {  
            long_hair: '0',  
            forehead_height_cm: '5.2',  
            forehead_width_cm: '12.8',  
            nose_wide: '1',  
            nose_long: '0',  
            lips_thin: '1',  
            distance_nose_to_lip_long: '1',  
            gender: 1  
        },  
        {  
            long_hair: '0',  
            forehead_height_cm: '6.9',  
            forehead_width_cm: '12.7',  
            nose_wide: '0',  
            nose_long: '1',  
            lips_thin: '1',  
            distance_nose_to_lip_long: '1',  
            gender: 1  
        },  
        {  
            long_hair: '0',  
            forehead_height_cm: '5.5',  
            forehead_width_cm: '12.5',  
            nose_wide: '1',  
            nose_long: '0',  
            lips_thin: '1',  
            distance_nose_to_lip_long: '1',  
            gender: 1  
        },  
        {  
            long_hair: '0',  
            forehead_height_cm: '5.1'
```

Hình 48. Tệp tin sau khi được chuyển đổi định dạng

## Tài liệu tham khảo

### Sách tham khảo chính:

[Shiyal, B. \(2021\). Beginning Azure Synapse Analytics: Transition from Data Warehouse to Data Lakehouse. Apress.](#)

### Tài liệu về Data Lake:

- Azure. (n.d.). *What is a data lake?*. Retrieved from  
<https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake>
- AWS. (n.d.). *What is a data lake?*. Retrieved from  
<https://aws.amazon.com/vi/what-is/data-lake/>
- Google Cloud. (n.d.). *What is a data lake?*. Retrieved from  
<https://cloud.google.com/learn/what-is-a-data-lake>
- Databricks. (n.d.). *Discover data lakes*. Retrieved from  
<https://www.databricks.com/discover/data-lakes>
- Oracle. (n.d.). *What is a data lake?*. Retrieved from  
<https://www.oracle.com/th/big-data/data-lake/what-is-data-lake/>
- Microsoft Learn. (n.d.). *Data lake scenarios*. Retrieved from  
<https://learn.microsoft.com/en-us/azure/architecture/data-guide/scenarios/data-lake>
- Azure. (n.d.). *Solutions – Data lake*. Retrieved from  
<https://azure.microsoft.com/en-us/solutions/data-lake>

### Tài liệu về Azure Synapse Analytics và các thành phần:

- Microsoft Azure. (n.d.). *Azure Synapse Analytics*. Retrieved from  
<https://azure.microsoft.com/en-us/products/synapse-analytics>
- Microsoft Learn. (n.d.). *Azure Synapse Analytics documentation*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/>

- K21Academy. (n.d.). *Azure Synapse Analytics Overview*. Retrieved from <https://k21academy.com/microsoft-azure/data-engineer/azure-synapse-analytics/>
- YouTube – Azure Synapse. (n.d.). Retrieved from <https://youtube.com/@AzureSynapse>
- Integrate.io. (n.d.). *What is Azure Synapse Analytics?*. Retrieved from <https://www.integrate.io/blog/what-is-azure-synapse-analytics/>
- Microsoft Learn. (n.d.). *Introduction to Azure Synapse Analytics*. Retrieved from <https://learn.microsoft.com/en-us/training/modules/introduction-azure-synapse-analytics/>
- Microsoft Learn. (n.d.). *Data platform end-to-end scenario*. Retrieved from <https://learn.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end?tabs=portal>
- Microsoft Learn. (n.d.). *SQL architecture overview*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/overview-architecture>
- Microsoft Learn. (n.d.). *MPP architecture in Synapse SQL*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/massively-parallel-processing-mpp-architecture>
- Microsoft Learn. (n.d.). *Synapse SQL features overview*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/overview-features>
- Microsoft Learn. (n.d.). *Apache Spark in Synapse*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-overview>

- Microsoft Learn. (n.d.). *Data Factory concepts – Pipelines & Activities*. Retrieved from <https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities?tabs=data-factory>
- Microsoft Learn. (n.d.). *Get started with Synapse Pipelines*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/get-started-pipelines>
- Microsoft Learn. (n.d.). *SQL Synapse Link overview*. Retrieved from <https://learn.microsoft.com/en-us/azure/synapse-analytics/synapse-link/sql-synapse-link-overview>
- Microsoft Learn. (n.d.). *SQL Server views*. Retrieved from <https://learn.microsoft.com/en-us/sql/relational-databases/views/views?view=sql-server-ver16>



