

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH



BÁO CÁO CUỐI KỲ

Đề tài: Azure Synapse Analysis Studio

Lớp học phần: DHKHDL17A

Mã lớp học phần: 420300232902

Nhóm thực hiện: 01

GVHD: TS.Nguyễn Chí Kiên

Thành phố Hồ Chí Minh, 10 tháng 12 năm 2024

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH



BÁO CÁO CUỐI KỲ

Đề tài: Azure Synapse Analysis Studio

Thành viên

STT	MSSV	Họ và tên
1	21084151	Nguyễn Thanh Thùy Trang
2	21106211	Trần Thu Huyền

TP. Hồ Chí Minh, tháng 12 năm 2023

MỤC LỤC

PHẦN 1: GIỚI THIỆU VỀ AZURE SYNAPSE ANALYTICS.....	6
I. Azure Synapse Analytics là gì?	6
II. Mục đích chính	6
III. Các khái niệm cơ bản và thuật ngữ liên quan	6
IV. Điểm nổi bật	6
V. Lợi ích của Azure Synapse Analytics.	7
VI. Kiến trúc Modern Data Warehouse.....	8
VII. Kiến trúc Azure Synapse Analytics - Data Lakehouse.....	8
VIII. Các giai đoạn phát triển của Azure Synapse Analytics	9
IX. Thành phần của Azure Synapse Analytics.....	10
PHẦN 2: AZURE SYNAPSE ANALYTICS MPP INTRO.....	11
PHẦN 3: AZURE SYNAPSE ANALYTICS STUDIO	14
I. Create workspace	14
II. Synapse Studio Overview hub	18
III. Synapse Studio Data hub	19
IV. Synapse Studio Develop hub.....	24
V. Synapse Studio Integrate hub	32
VI. Synapse Studio Monitor hub	34
VII. Synapse Studio Manage hub	35
PHẦN 4: AZURE SYNAPSE ANALYTICS DATA INTEGRATION	40
I. Data integration	40
II. Data Movement	40
III. Pipelines.....	41
IV. Prep & Transform Data (Data flow).....	43
V. Triggers	43
VI. Manage – Linked Services	43
VII. Manage – Integration runtimes	44
PHẦN 5: AZURE SYNAPSE ANALYTICS SQL ANALYTICS	45
I. Analytics	45
II. Data Storage and Performance Optimizations	51
III. Performance Optimizations Workload Management.....	55
IV. Developer productivity	57

V. Maintenance.....	58
VI. Snapshots and restores	60
PHẦN 6: SQL ON DEMAND	60
I. Overview.....	60
II. Querying on storage.....	61
III. Querying CSV File	62
IV. Querying folders	63
V. Querying specific files	64
VI. Querying Parquet files	64
VII. Creating views	66
VIII. SQL On Demand – Querying JSON files	67
IX. Create External Table As Select.....	67
PHẦN 7: AZURE SYNAPSE ANALYTICS SPARK	68
I. Azure Synapse Apache Spark – Summary	68
II. Apache Spark	68
III. Motivation for Apache Spark.....	69
IV. What makes Spark fast	69
V. General Spark Cluster Architecture	70
VI. Spark Component Features	70
VII. Architecture Overview.....	71
VIII. Spark pool	72
PHẦN 8: INDUSTRY-LEADING SECURITY AND COMPLIANCE	78
I. Threat Protection (Bảo vệ môi đe dọa)	78
II. Network Security - Business requirements.....	83
III. Authentication - Business requirements.....	87
IV. Access Control - Business requirements.....	89
V. Data Protection - Business requirements.....	92
VI. Single Sign-On.....	95
PHẦN 9: APPLICATION	95
I.Setup workspace	95
II. Áp dụng SQL serverless	97
III. Áp dụng Notebook.....	99
PHẦN 10: DEVELOPMENT PROBLEM	101

PHẦN 11: CONCLUSION 109

PHẦN 1: GIỚI THIỆU VỀ AZURE SYNAPSE ANALYTICS

I. Azure Synapse Analytics là gì?

Azure Synapse Analytics là một dịch vụ phân tích không giới hạn, kết hợp giữa kho dữ liệu doanh nghiệp và phân tích Big Data. Nó cho phép bạn tự do truy vấn dữ liệu theo cách của bạn, sử dụng tài nguyên theo yêu cầu không máy chủ (serverless) hoặc tài nguyên đã được cung cấp (dedicated), với quy mô lớn. Azure Synapse kết hợp hai thế giới này với một trải nghiệm thống nhất để tiếp nhận, chuẩn bị, quản lý và phục vụ dữ liệu cho nhu cầu thông tin doanh nghiệp và học máy ngay lập tức.

II. Mục đích chính

- Giúp doanh nghiệp xử lý và phân tích khối lượng dữ liệu lớn nhanh chóng.
- Tạo ra các báo cáo chi tiết và phân tích dữ liệu theo thời gian thực, hỗ trợ đưa ra quyết định dựa trên dữ liệu. Các tính năng nổi bật của Synapse SQL

III. Các khái niệm cơ bản và thuật ngữ liên quan

- Workspace: Là môi trường làm việc trong Azure Synapse, nơi bạn có thể quản lý và tổ chức các tài nguyên phân tích của mình.
- SQL Pool: Bao gồm dedicated SQL pool và serverless SQL pool. Dedicated SQL pool cung cấp các tài nguyên tính toán cố định, trong khi serverless SQL pool cho phép bạn truy vấn dữ liệu mà không cần quản lý cơ sở hạ tầng.
- Apache Spark Pool: Cung cấp khả năng tính toán phân tán cho các tác vụ xử lý dữ liệu lớn, sử dụng Apache Spark.
- Data Explorer Pool: Được sử dụng để phân tích dữ liệu log và dữ liệu chuỗi thời gian.
- Pipeline: Là một chuỗi các hoạt động dữ liệu được tổ chức và thực thi theo thứ tự để chuyển đổi và tải dữ liệu.
- Linked Service: Là các kết nối đến các nguồn dữ liệu bên ngoài mà bạn muốn tích hợp vào Azure Synapse.
- Managed Private Endpoint: Là các điểm cuối riêng tư được quản lý, cho phép kết nối an toàn giữa Azure Synapse và các dịch vụ Azure khác.
- Integration Runtime: Là cơ sở hạ tầng tính toán được sử dụng để thực thi các hoạt động dữ liệu trong Azure Synapse.

IV. Điểm nổi bật

1. Best in class price performance (giá tốt nhất theo hiệu suất tốt nhất).

- Tối ưu hóa tài nguyên: Azure Synapse Analytics sử dụng các thuật toán tối ưu hóa để phân bổ tài nguyên một cách hiệu quả, giúp giảm chi phí mà vẫn đảm bảo hiệu suất cao.
- Tính toán linh hoạt: Dịch vụ cung cấp các tùy chọn tính toán linh hoạt, cho phép bạn chỉ trả tiền cho những gì bạn sử dụng, từ đó tối ưu hóa chi phí.

2. Industry-leading security (bảo mật hàng đầu ngành).

- Mã hóa dữ liệu: Azure Synapse sử dụng mã hóa dữ liệu ở cả trạng thái nghỉ và trong quá trình truyền tải, đảm bảo rằng dữ liệu của bạn luôn được bảo vệ.
- Quản lý khóa: Tích hợp với Azure Key Vault để quản lý và bảo vệ các khóa mã hóa, giúp tăng cường bảo mật.
- Xác thực và quyền truy cập: Sử dụng Azure Active Directory để quản lý quyền truy cập và xác thực người dùng, đảm bảo chỉ những người được ủy quyền mới có thể truy cập vào dữ liệu.

3. Workload aware query execution (thực thi truy vấn nhân thức theo khối lượng).

- Tối ưu hóa truy vấn: Azure Synapse có khả năng tự động tối ưu hóa các truy vấn dựa trên khối lượng công việc hiện tại, giúp cải thiện hiệu suất và giảm thời gian xử lý.
- Phân bổ tài nguyên thông minh: Dịch vụ có thể phân bổ tài nguyên tính toán một cách thông minh dựa trên nhu cầu của từng truy vấn, đảm bảo hiệu suất tối ưu.

4. Data flexibility (tính linh hoạt của dữ liệu).

- Hỗ trợ nhiều định dạng dữ liệu: Azure Synapse hỗ trợ nhiều định dạng dữ liệu khác nhau, từ dữ liệu có cấu trúc như SQL đến dữ liệu phi cấu trúc như JSON và Parquet.
- Tích hợp dễ dàng: Dịch vụ cho phép tích hợp dữ liệu từ nhiều nguồn khác nhau, bao gồm Azure Data Lake Storage, Azure Blob Storage, và các cơ sở dữ liệu bên ngoài.

5. Developer productivity (năng suất phát triển).

- Công cụ phát triển mạnh mẽ: Azure Synapse cung cấp các công cụ phát triển như Synapse Studio, giúp các nhà phát triển dễ dàng xây dựng, triển khai và quản lý các giải pháp phân tích dữ liệu.
- Tích hợp CI/CD: Hỗ trợ tích hợp với các công cụ CI/CD như Azure DevOps và GitHub, giúp tự động hóa quy trình phát triển và triển khai.
- Hỗ trợ nhiều ngôn ngữ lập trình: Dịch vụ hỗ trợ nhiều ngôn ngữ lập trình như SQL, Python, Scala, và .NET, giúp các nhà phát triển có thể làm việc với ngôn ngữ mà họ quen thuộc.

V. Lợi ích của Azure Synapse Analytics.

Bằng cách tận dụng các đối tác ISV, doanh nghiệp có thể mở rộng khả năng phân tích dữ liệu của mình trong Azure Synapse Analytics thông qua các công cụ chuyên biệt, tích hợp và giải pháp ngành, cùng với sự hỗ trợ từ các chuyên gia. Điều này giúp các tổ chức tối đa hóa khoản đầu tư vào Azure Synapse, đồng thời nâng cao khả năng phân tích dữ liệu, giúp đưa ra những quyết định kinh doanh tốt hơn.

Dưới đây là những lợi ích chính mà đối tác ISV có thể mang lại khi tích hợp với Azure Synapse Analytics:

- Azure Synapse Analytics là trung tâm phân tích dữ liệu: Nó kết nối với các công cụ khác trong hệ sinh thái Azure, cho phép thực hiện phân tích mạnh mẽ và linh hoạt.
- Azure Data Share: Đây là công cụ giúp chia sẻ dữ liệu giữa các hệ thống và đối tác khác nhau, hỗ trợ việc tích hợp dữ liệu từ nhiều nguồn khác nhau.
- Power BI: Được kết nối với Azure Synapse để tạo các bảng điều khiển và trực quan hóa dữ liệu, giúp doanh nghiệp dễ dàng nắm bắt và phân tích dữ liệu.
- Azure Machine Learning: Tích hợp công cụ học máy trong Azure Synapse để thực hiện các phân tích sâu và dự đoán từ dữ liệu, hỗ trợ việc ra quyết định dựa trên dữ liệu.
- Hệ sinh thái các công cụ và dịch vụ: Azure Synapse mở rộng khả năng phân tích với các đối tác như Informatica, Talend, Panoply, Azure Databricks, và Attunity, giúp tích hợp và tối ưu hóa quy trình phân tích.

Ngoài ra, còn có những lợi ích bổ sung sau:

- Khả năng tương thích ngược với Azure SQL Data Warehouse: Giúp tích hợp và điều phối dữ liệu dễ dàng, bảo đảm sự liên mạch giữa các hệ thống.
- Khả năng phân tích mở rộng trong Azure Synapse: Tạo ra các kịch bản mới cho các đối tác ISV, mở rộng các khả năng phân tích dữ liệu.
- Đảm bảo tính liên tục của dữ liệu: Khi tích hợp Azure Synapse với Azure Machine Learning và Power BI, giúp duy trì tính liên tục trong việc xử lý và phân tích dữ liệu.
- Giảm thiểu nỗ lực di chuyển: Các doanh nghiệp có thể tái sử dụng các nền tảng đối tác hiện có, giảm thiểu chi phí và thời gian chuyển đổi.

VII. Kiến trúc Modern Data Warehouse

- Thu thập dữ liệu từ nhiều nguồn (on-premises, cloud, SaaS)
- Xử lý qua nhiều bước riêng biệt:
 - INGEST (Thu thập) bằng Azure Data Factory
 - PREPARE (Chuẩn bị) bằng Data Factory/Databricks
 - TRANSFORM & ENRICH (Chuyển đổi & Làm giàu) bằng Data Factory/Databricks
 - SERVE (Phục vụ) bằng Azure SQL Data Warehouse
 - VISUALIZE (Trực quan hóa) bằng Power BI
- Lưu trữ dữ liệu trong Azure Data Lake Storage

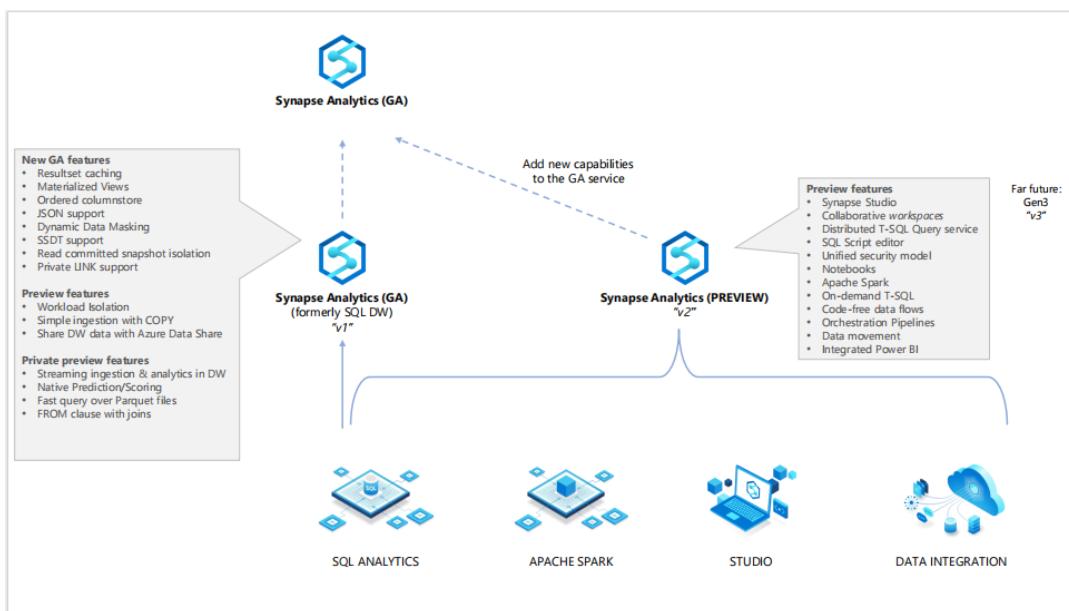
VIII. Kiến trúc Azure Synapse Analytics - Data Lakehouse

Azure Synapse Analytics là một nền tảng tích hợp mạnh mẽ, kết hợp giữa kho dữ liệu truyền thống (Data Warehouse) và hồ dữ liệu (Data Lake), tạo thành kiến trúc **Data Lakehouse**. Kiến trúc này cho phép quản lý và xử lý dữ liệu một cách linh hoạt, phục vụ cả nhu cầu phân tích tập trung lẫn phân tích phi tập trung.

Các Thành Phần Chính trong Kiến Trúc Data Lakehouse

- Azure Data Lake
- Azure Synapse Analytics
- Data Integration
- Data Warehouse
- Analytics & Machine Learning.

VIII. Các giai đoạn phát triển của Azure Synapse Analytics



1. Phiên bản v1 (formerly SQL DW)

- Tập trung vào các tính năng phân tích SQL cơ bản
- Bổ sung các tính năng mới như:
 - Caching kết quả truy vấn
 - Views vật lý hóa
 - Hỗ trợ JSON
 - Bảo mật dữ liệu động
 - Hỗ trợ SSDT
 - Cố lập snapshot đã commit
 - Hỗ trợ Private LINK

2. Phiên bản v2 (Preview)

Mở rộng khả năng với:

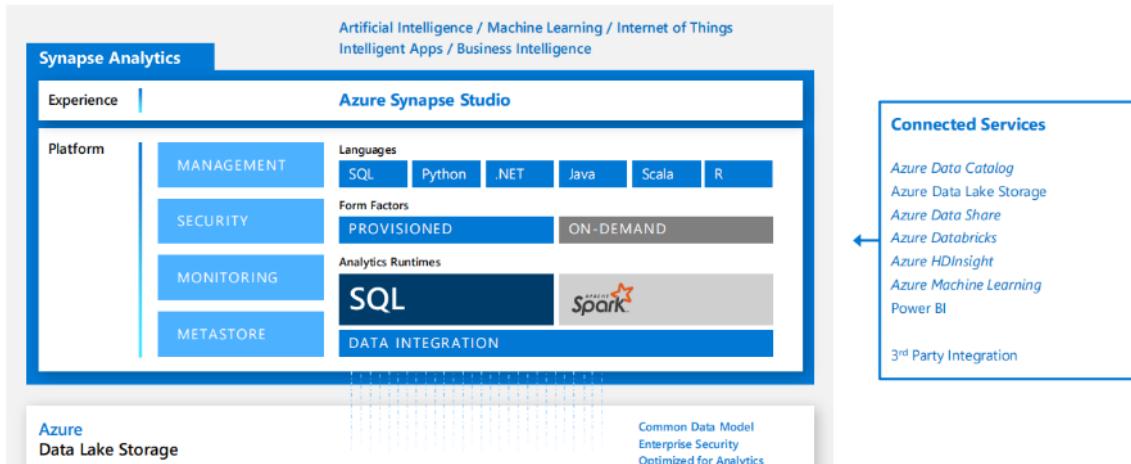
- Synapse Studio - môi trường phát triển tích hợp
- Không gian làm việc cộng tác
- Dịch vụ truy vấn T-SQL phân tán
- Mô hình bảo mật thống nhất
- Tích hợp Apache Spark
- Khả năng T-SQL theo yêu cầu
- Luồng dữ liệu không cần code

- Pipeline điều phối
- Tích hợp Power BI

3. Hướng tới tương lai (Gen3 "v3")

- Tiếp tục phát triển và mở rộng các khả năng
- Tích hợp sâu hơn các công nghệ mới
- Cải thiện hiệu suất và khả năng mở rộng

IX. Thành phần của Azure Synapse Analytics



1. Azure Synapse Analytics

Đây là trung tâm của dịch vụ, cung cấp một môi trường thống nhất để bạn quản lý, phân tích và trực quan hóa dữ liệu. Azure Synapse Studio

2. Azure Synapse Studio

- **Giao diện người dùng:** Đây là một giao diện web, giúp quản lý và phân tích dữ liệu một cách trực quan.
- **Không gian làm việc:** Cho phép nhiều người làm việc cùng nhau trên một dự án dữ liệu.
- **Hỗ trợ đa dạng:** có thể truy cập các cơ sở dữ liệu SQL, bảng Spark, chạy các script SQL, sử dụng notebook (hỗ trợ nhiều ngôn ngữ), tạo các luồng dữ liệu (Data Flows), pipeline (Data Integration), và thực hiện các tác vụ giám sát, bảo mật.
- **Kết nối:** Có thể kết nối với ADLS Gen2 và Power BI workspace.

3. Platform

- Các thành phần quản lý, bảo mật, giám sát và metastore để đảm bảo dữ liệu của bạn được bảo vệ và hoạt động hiệu quả.

4. Languages

- Hỗ trợ nhiều ngôn ngữ lập trình phổ biến như SQL, Python, .NET, Java, Scala và R, cho phép bạn lựa chọn công cụ phù hợp nhất với nhu cầu của mình.

5. Form Factors

- Cung cấp hai tùy chọn triển khai:
 - Provisioned: Môi trường được cung cấp sẵn với các tài nguyên được cấu hình trước.
 - On-Demand: Môi trường được mở rộng và thu nhỏ linh hoạt theo nhu cầu sử dụng.

6. Analytics Runtimes

- SQL: Dùng để xử lý các truy vấn SQL truyền thống, bao gồm cả xử lý batch và streaming.
- Spark: Dùng để xử lý dữ liệu lớn với các ngôn ngữ như Python, Scala và R.

7. Data Integration

- Cho phép tích hợp dữ liệu từ nhiều nguồn khác nhau, bao gồm cả dữ liệu trong Azure Data Lake.

8. Azure Data Lake Storage

- Là nơi lưu trữ dữ liệu lớn, được tối ưu hóa cho các hoạt động phân tích.

PHẦN 2: AZURE SYNAPSE ANALYTICS MPP INTRO

I. Kiến trúc của Azure Synapse Analytics

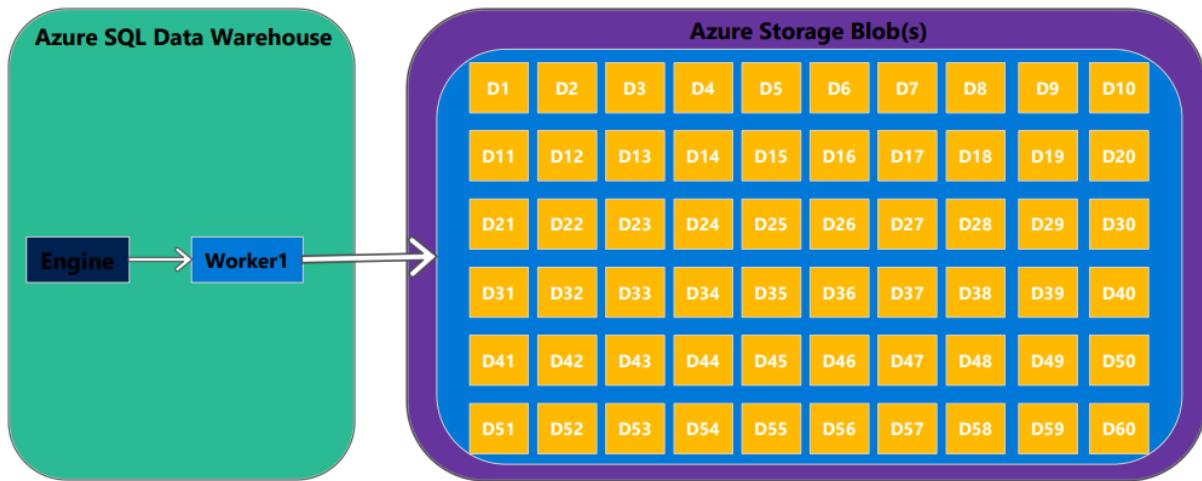
Azure Synapse Analytics (trước đây là SQL Data warehouse) sử dụng kiến trúc **Massively Parallel Processing (MPP)** để xử lý và phân tích dữ liệu quy mô lớn. MPP là một phương pháp trong đó nhiều nút (nodes) làm việc song song, chia nhỏ và phân phối các tác vụ tính toán, nhằm tăng tốc độ xử lý dữ liệu và cải thiện hiệu suất.

1. Giới thiệu về kiến trúc của MPP

- **Control Node:** Là trung tâm điều phối, nhận truy vấn từ người dùng và tối ưu hóa truy vấn, phân chia các tác vụ thành các đoạn nhỏ hơn và phân phối chúng cho các Compute Nodes.
- **Compute Nodes:** thực hiện các tác vụ tính toán. Mỗi Compute Node xử lý một phần dữ liệu được lưu trữ trên nó, sử dụng cơ chế phân vùng dữ liệu (**Data Distribution**) để lưu trữ dữ liệu một cách hiệu quả.
- **Storage Layer:** sử dụng Azure Data Lake Storage hoặc Blob Storage để lưu trữ dữ liệu một cách linh hoạt và có thể mở rộng, tách biệt giữa **Compute** và **Storage**, cho phép mở rộng độc lập khi cần thiết.

2. Architecture for DW100

Architecture for DW100



Kiến trúc DW100 trong Azure SQL Data Warehouse gồm 3 thành phần chính:

- **Azure SQL Data Warehouse (Engine):** Trung tâm điều khiển, quản lý metadata, lập kế hoạch truy vấn, và điều phối hoạt động giữa các thành phần.
- **Worker (Nút tính toán):** Xử lý dữ liệu từ các phân mảnh (slices) trong Azure Storage Blob. Mỗi nút tính toán đảm nhiệm một phần dữ liệu, hỗ trợ xử lý song song.
- **Azure Storage Blob(s):** Nơi lưu trữ dữ liệu, được chia nhỏ thành các phân mảnh, giúp tối ưu hóa phân phối và truy xuất.

Hoạt động:

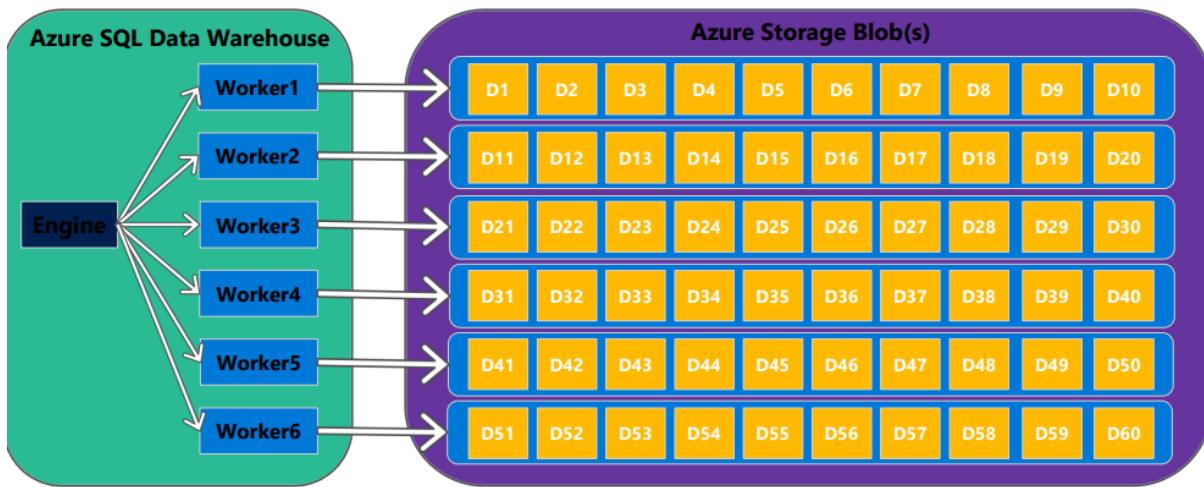
- Engine nhận truy vấn, lập kế hoạch, và phân công công việc đến các Worker.
- Các Worker xử lý dữ liệu và trả kết quả về cho Engine.
- Engine tổng hợp và trả kết quả cho người dùng.

Đặc điểm:

- Tăng tốc độ xử lý nhờ tính song song.
- Tiết kiệm chi phí với lưu trữ Azure Blob.
- Phù hợp cho phân tích dữ liệu lớn và truy vấn phức tạp.

3. Architecture for DW600

Architecture for DW600



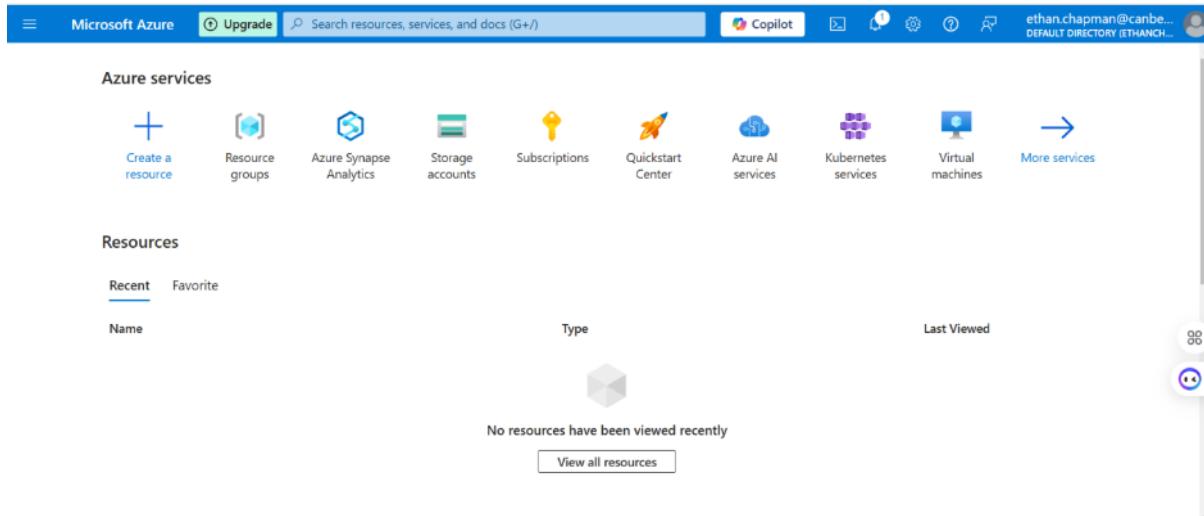
Về cơ bản Kiến Trúc của DW600 cũng giống DW100 tuy nhiên thì bên DW100 với 1 Worker còn DW600 là 6 Worker nên dẫn đến sự khác biệt sau :

Đặc điểm	DW100	DW600
Engine	1 Engine quản lý metadata và phân phối công việc.	Tương tự DW100, vẫn có 1 Engine làm nhiệm vụ điều phối.
Worker (Nút tính toán)	1 Worker duy nhất để xử lý dữ liệu.	6 Workers, mỗi Worker xử lý một nhóm phân mảnh dữ liệu.
Azure Storage Blob(s)	Dữ liệu được chia thành 60 phân mảnh .	Dữ liệu vẫn chia thành 60 phân mảnh , nhưng được phân phối đều cho 6 Workers.
Tính song song	Hạn chế vì chỉ có 1 Worker xử lý dữ liệu.	Mạnh mẽ hơn với 6 Workers xử lý song song.
Hiệu suất	Phù hợp cho khối lượng dữ liệu nhỏ.	Hiệu suất cao hơn, phù hợp với dữ liệu lớn và truy vấn phức tạp.
Ứng dụng	Dành cho các hệ thống nhỏ hoặc vừa.	Phù hợp với hệ thống lớn, yêu cầu xử lý nhanh và phân tích phức tạp.
Thời gian xử lý truy vấn	Chậm hơn do giới hạn ở 1 Worker.	Nhanh hơn nhờ chia công việc cho nhiều Workers.

PHẦN 3: AZURE SYNAPSE ANALYTICS STUDIO

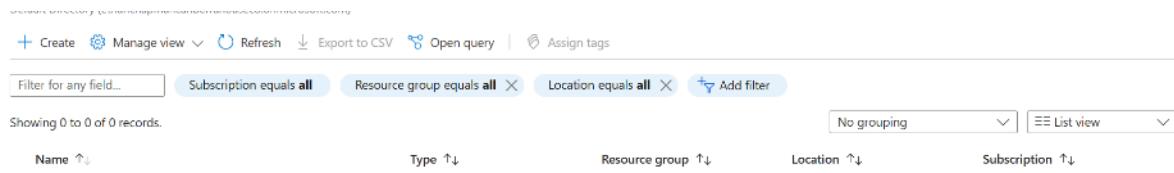
I. Create workspace

B1: Chọn Azure synapse Analytics



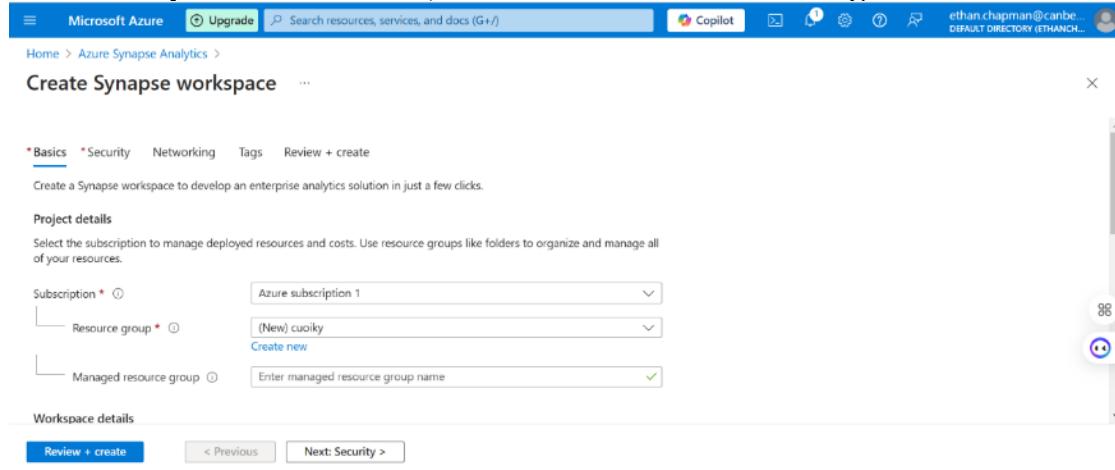
The screenshot shows the Microsoft Azure portal's main dashboard. At the top, there are links for 'Upgrade', 'Search resources, services, and docs (G+)', 'Copilot', and user information ('ethan.chapman@canbe...'). Below the search bar is a section titled 'Azure services' with various icons and links: 'Create a resource', 'Resource groups', 'Azure Synapse Analytics', 'Storage accounts', 'Subscriptions', 'Quickstart Center', 'Azure AI services', 'Kubernetes services', 'Virtual machines', and 'More services'. Under the 'Resources' section, it says 'Recent' and 'Favorite'. It lists 'Name', 'Type', 'Last Viewed', and a 'View all resources' button. A message states 'No resources have been viewed recently'.

B2: Chọn create



The screenshot shows a search results page for 'Create' in the Azure portal. The top navigation includes 'Create', 'Manage view', 'Refresh', 'Export to CSV', 'Open query', and 'Assign tags'. Below the search bar, there are filters: 'Filter for any field...', 'Subscription equals all', 'Resource group equals all', 'Location equals all', and 'Add filter'. The results table has columns for 'Name', 'Type', 'Resource group', 'Location', and 'Subscription'. A message at the top says 'Showing 0 to 0 of 0 records.'

B3: Chọn Review+create, khi đã hoàn thành các thông tin sau:



The screenshot shows the 'Create Synapse workspace' wizard. The first step, 'Basics', is selected. It asks to 'Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.' Under 'Project details', it says to 'Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.' The 'Subscription' dropdown is set to 'Azure subscription 1'. The 'Resource group' dropdown is set to '(New) cuoiky' with a 'Create new' link. The 'Managed resource group' input field is empty and labeled 'Enter managed resource group name'. At the bottom, there are buttons for 'Review + create' (highlighted in blue), '< Previous', and 'Next: Security >'.

Microsoft Azure Upgrade Search resources, services, and docs (G+)

ethan.chapman@canbe... DEFAULT DIRECTORY (ETHANCH...)

Home > Azure Synapse Analytics >

Create Synapse workspace

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name *	cuoiky-synapse
Region *	Southeast Asia
Select Data Lake Storage Gen2 *	<input checked="" type="radio"/> From subscription <input type="radio"/> Manually via URL
Account name *	(New) tranghuyen Create new
File system name *	(New) data Create new
<input checked="" type="checkbox"/> Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.	
Review + create < Previous Next: Security >	

- **Workspace name:** create new tranghuyen-synapse
- **Account name:** create new tranghuyenadls2
- **File System Name:** demo

Microsoft Azure Upgrade Search resources, services, and docs (G+)

ethan.chapman@canbe... DEFAULT DIRECTORY (ETHANCH...)

Home > Azure Synapse Analytics >

Create Synapse workspace

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name *	tranghuyen-synapse
Region *	Southeast Asia
Select Data Lake Storage Gen2 *	<input checked="" type="radio"/> From subscription <input type="radio"/> Manually via URL
Account name *	(New) tranghuyenadls2 Create new
File system name *	(New) demo Create new
<input checked="" type="checkbox"/> Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.	
Review + create < Previous Next: Security >	

B5: chọn create

Microsoft Azure Upgrade Search resources, services, and docs (G+)

ethan.chapman@canbe... DEFAULT DIRECTORY (ETHANCH...)

Home > Azure Synapse Analytics >

Create Synapse workspace

Validation succeeded

[Basics](#) [*Security](#) [Networking](#) [Tags](#) [Review + create](#)

Product Details

Azure Synapse Analytics workspace by Microsoft [Serverless SQL est. cost/TB](#) **5.00 USD**

[Terms of use](#) | [Privacy policy](#)

Terms

By clicking Create, I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional

[Create](#) [< Previous](#) [Next >](#) [Download a template for automation](#)

Kết quả

The screenshot shows the Microsoft Azure Synapse Analytics Overview page. At the top, it displays the deployment name: Microsoft.Azure.SynapseAnalytics-20241205214847. Below this, a message states "Deployment is in progress". It provides details about the deployment: Deployment name: Microsoft.Azure.SynapseAnalytics-20241205214847, Start time: 12/5/2024, 10:10:12 PM, Subscription: Azure subscription 1, Correlation ID: 23f15d41-d98a-47b8-b0eb-1..., and Resource group: tranghuyenRG. A table lists the resources being deployed:

Resource	Type
tranghuyen synapse	Synapse workspace
tranghuyenadls2/default/demo	Microsoft.Storage/storageAccounts/blobServices/c
tranghuyenadls2	Storage account

On the right side of the page, there are promotional banners for Microsoft Defender for Cloud, Free Microsoft tutorials, and Work with an expert.

Synapse Workspace

The screenshot shows the Microsoft Azure Synapse Workspace Overview page for the workspace "tranghuyen-synapse". The left sidebar includes links for Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, Analytics pools, Security, Monitoring, Automation, and Help. The main area displays the "Analytics pools" section, which includes a search bar and a table:

Name	Type	Size
Built-in	Serverless	Auto
Apache Spark pools	No pools provisioned	
Data Explorer pools	No pools provisioned	

Chọn + new dedicated SQL Pool để tạo SQL pool và chọn + New Apache Spark pool để tạo spark pool

SQL pool and spark pool

B6: Vào workspace tranhuyen-synapse -> chọn open synapse studio

II. Synapse Studio Overview hub

Synapse Analytics workspace
tranhuyen-synapse

New

Ingest
Perform a one-time or scheduled data load.

Explore and analyze
Learn how to get insights from your data.

Visualize
Build interactive reports with Power BI capabilities.

Discover more

Synapse Studio là một giao diện web tích hợp cho phép bạn phát triển, quản lý và giám sát các giải pháp phân tích dữ liệu trong Azure Synapse Analytics. Nó cung cấp một môi trường làm việc hợp nhất để thực hiện các tác vụ chính như nhập dữ liệu, khám phá dữ liệu, chuẩn bị dữ liệu, điều phối, và trực quan hóa dữ liệu.

Các tab chính trong Synapse Studio: Home (Trang chủ), Data (Dữ liệu), Develop (Phát triển), Integrate (Tích hợp), Monitor (Giám sát), Manage (Quản lý).

Synapse Analytics workspace
tranhuyen-synapse

New

SQL script

KQL script

Notebook

Data flow

Apache Spark job definition

Pipeline

Import

Discover more

Click New (Thả xuống)

- **SQL script:** Tạo tập lệnh SQL để truy vấn dữ liệu.
- **KQL script:** Tạo tập lệnh KQL (Kusto Query Language) để phân tích dữ liệu.
- **Notebook:** Tạo notebook (thường dùng Python hoặc Spark) để xử lý dữ liệu và phân tích.
- **Data flow:** Tạo luồng dữ liệu, giúp ETL (Extract, Transform, Load) và xử lý dữ liệu.

- Apache Spark job definition: Định nghĩa và quản lý các công việc Spark.
- Pipeline: Tạo quy trình ETL hoặc tích hợp dữ liệu.
- Import: Nhập pipeline hoặc tài nguyên từ các tệp bên ngoài.

III. Synapse Studio Data hub

1. Overview

Data hub trong Synapse Studio là nơi tập trung để quản lý và khám phá các tài sản dữ liệu của bạn. Nó cung cấp một giao diện thống nhất để làm việc với các loại dữ liệu khác nhau, bao gồm dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc.

2. Linked

Kết nối và tải dữ liệu từ nhiều nguồn khác nhau: Azure Blob Storage, Azure Data Lake Storage Gen2, Integration datasets.

B1: trong synapse studio đến data hub -> chọn linked

B2: Trong danh mục **Azure Data Lake Storage Gen2**, sẽ thấy một mục có tên **tranhuyen-synapse (Primary - tranhuyenadlsg2)**. Chọn vùng chứa có tên là **demo (Primary)**.

B3: Chọn có thể chọn upload or link từ integration từ table của workspace hoặc Blob storage account để đẩy dữ liệu vào.

chọn upload

B4: Chọn browse và chọn file NYCTaxiGreen.parquet -> chọn upload

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. On the left, there's a navigation pane with 'Data' selected, showing 'Workspace' and 'Linked' sections. A search bar at the top right has 'Search' entered. In the center, a 'demo' folder is selected under 'New SQL script'. To the right, the 'Upload Files' section is open, showing a file named 'NYCTaxiGreen.parquet' from the 'synapse' destination folder. A red box highlights the 'File Upload' area. Another red box highlights the 'Upload' button at the bottom of the upload dialog.

Kết quả

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface after the upload. The 'demo' folder now contains the 'NYCTaxiGreen.parquet' file, which is listed in the 'Content Type' column as '1.2 MB'. A red box highlights the file entry in the list.

Kết quả của dùng 3 cách

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The left sidebar is highlighted with a red box, showing various storage and dataset options like 'Azure Blob Storage', 'Sample Datasets', and 'Integration datasets'. The main workspace area shows a table named 'SqlPoolTable1' under an 'Azure Synapse dedicated SQL pool'. The 'Connection' tab is selected, showing 'SQL pool' set to 'SQLPool' and 'Schema' set to 'dbo.NYCTaxiGreen'. A red box highlights the connection settings.

a) Preview data (xem trước dữ liệu)

Click chuột phải vào file apriori.csv và chọn previews

The screenshot shows the Microsoft Azure Synapse Analytics interface. On the left, there's a sidebar with 'Data' selected, showing a workspace named 'tranhuyen-synapse'. Inside the workspace, there's a folder 'demo' containing a file named 'apriori.csv'. A red arrow points from the 'Preview' option in the context menu of 'apriori.csv' to the preview pane on the right. The preview pane displays the first few rows of the CSV file:

TRANSACTION	MILK	BREAD	BUTTER
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1

b) See basic file properties (xem properties của file)

Click chuột phải vào file NYCTaxiGreen.parquet và chọn properties

The screenshot shows the Microsoft Azure Synapse Analytics interface. On the left, there's a sidebar with 'Data' selected, showing a workspace named 'tranhuyen-synapse'. Inside the workspace, there's a folder 'synapse' containing a file named 'NYCTaxiGreen.parquet'. A red arrow points from the 'Properties' option in the context menu of 'NYCTaxiGreen.parquet' to the properties pane on the right. The properties pane displays the following information:

Name	NYCTaxiGreen.parquet
ABFS Path	abfss://tranhuyenadls.dfs.core.windows.net/synapse/nyctaxigreen.parquet
Last Modified	12/7/2024, 7:25:29 AM
Cache Control	max-age=0
Content Type	application/octet-stream
Content Hash (base64)	[REDACTED]

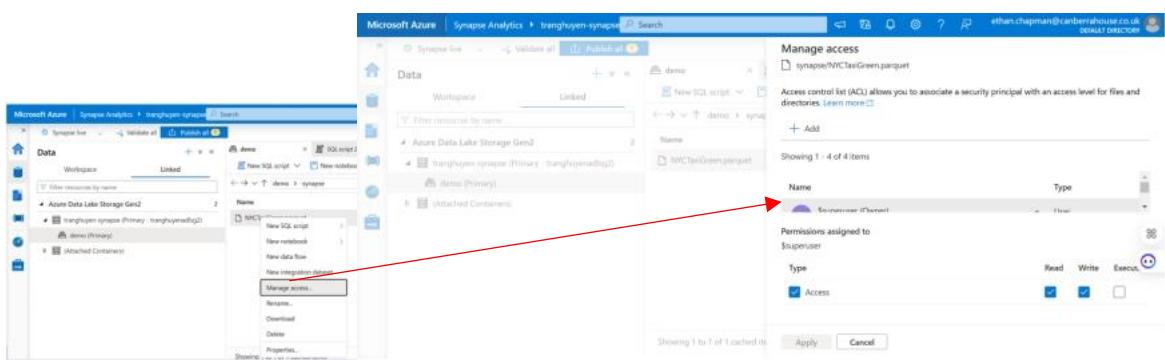
- Name: Tên của tệp dữ liệu, ở đây là synapse/NYCTaxiGreen.parquet. Điều này chỉ ra rằng tệp nằm trong thư mục synapse và tên tệp là NYCTaxiGreen.parquet.
- URL: Đường dẫn đầy đủ để truy cập tệp qua giao thức HTTP hoặc HTTPS. Ở đây, URL có dạng: <https://tranhuyenadls....>. Đây là đường dẫn công khai hoặc nội bộ trong Azure Blob Storage.
- ABFSS Path: Đường dẫn sử dụng giao thức Azure Blob File System (abfss://), dành cho việc truy cập dữ liệu trong Azure Data Lake từ các công cụ hoặc dịch vụ như Azure Synapse hoặc Spark. Thường dùng trong các script hoặc notebook để kết nối tới dữ liệu một cách an toàn.
- Last Modified: Thời gian chỉnh sửa cuối cùng của tệp. Ví dụ: 7:25:29 AM, 12/7/2024. Điều này giúp theo dõi phiên bản hoặc thời điểm tệp được cập nhật lần cuối.
- Cache Control: Chỉ định cách dữ liệu sẽ được lưu vào bộ nhớ đệm. Giá trị max-age=0 nghĩa là không cho phép lưu bộ nhớ đệm, tệp sẽ luôn được tải lại trực tiếp từ server.
- Content Type: Loại nội dung của tệp, được biểu thị dưới dạng MIME type. application/octet-stream là định dạng mặc định, chỉ ra đây là một luồng dữ liệu

nhi phân (binary stream), thường áp dụng cho các tệp không có định dạng cụ thể hoặc không được xác định.

- Content Disposition: Thuộc tính này xác định cách trình duyệt xử lý tệp khi tải về (ví dụ: mở trực tiếp hoặc lưu). Ở đây không có giá trị, nên trình duyệt sẽ sử dụng mặc định (thường là tải xuống).
- Content Encoding: Đây là cách dữ liệu được mã hóa để lưu trữ hoặc truyền tải (gzip, deflate, hoặc identity (không nén)). Giá trị trống cho thấy tệp không sử dụng bất kỳ mã hóa đặc biệt nào.
- Content language: Giá trị trống ở đây có nghĩa là ngôn ngữ không được chỉ định, thường vì tệp không liên quan đến ngôn ngữ (như tệp dữ liệu hoặc nhị phân).

c) Manage Access (quản lý truy cập)

Click chuột phải vào file **NYCTaxiGreen.parquet** và chọn manage access



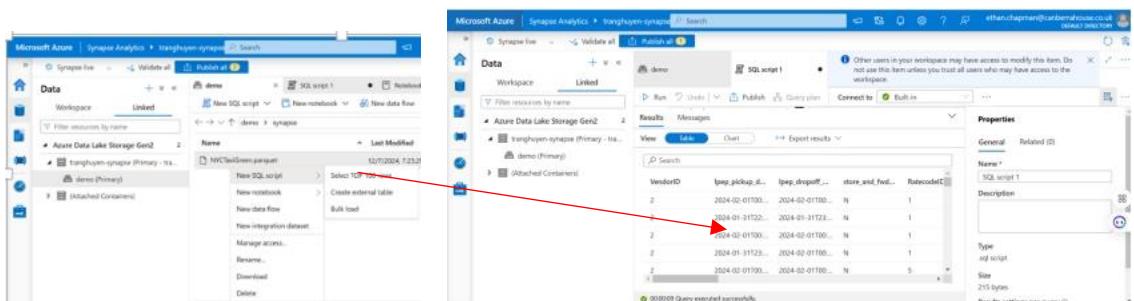
+ Add (Thêm người dùng hoặc nhóm): Nút này cho phép thêm một security principal (người dùng, nhóm, hoặc ứng dụng) và gán quyền cụ thể cho họ.

Danh sách quyền đã được gán: Name là Tên người dùng hoặc nhóm đã được gán quyền. Ví dụ: \$superuser là một vai trò hoặc người dùng được gán quyền. Type là Loại thực thể. Trong trường hợp này là User hoặc nhóm Role.

Các quyền được hiển thị gồm: Read cho phép đọc nội dung của tệp, Write cho phép ghi (chỉnh sửa hoặc thêm nội dung), Execute cho phép thực thi (truy cập thư mục hoặc chạy script nếu liên quan).

d) SQL script

B1: Click chuột phải chọn SQL script, chọn select 100 row-> run



B2: Nhập tên SQLpooltxgdedicated vào Properties và chọn Publish

The screenshot shows the Microsoft Azure Synapse Analytics Data studio interface. On the left, there's a sidebar with 'Data' and 'Linked' tabs, and a search bar. The main area has a 'Run' button, an 'Undo' button, a 'Publish' button (highlighted in yellow), and a 'Query plan' button. The 'Publish' button has a tooltip 'Publish [Ctrl+S]'. Below these are tabs for 'Results' and 'Messages'. The 'Results' tab shows a green checkmark and the message '00:00:09 Query executed successfully.' To the right, there's a 'General' tab for properties, where the 'Name' field is set to 'SQLpooltxgdedicated'. Other properties shown include 'Type: .sql script', 'Size: 215 bytes', and 'Results settings per query' with options 'First 5000 rows (default)' and 'All rows'.

e) Sử dụng notebook

B1: Chọn New Notebook -> chọn Load to DataFrame-> Trong bảng Notebook 2 mở ra, trong danh sách Attach to, chọn Sparkpool và đảm bảo rằng ngôn ngữ được thiết lập là PySpark (Python)-> run

This screenshot shows the Microsoft Azure Synapse Analytics Data studio interface. On the left, there's a sidebar with 'Data' and 'Linked' tabs, and a search bar. The main area has a 'Run' button, an 'Undo' button, a 'Publish' button (highlighted in yellow), and a 'Query plan' button. The 'Publish' button has a tooltip 'Publish [Ctrl+S]'. Below these are tabs for 'Results' and 'Messages'. The 'Results' tab shows a green checkmark and the message '4 min 23 sec The job completed successfully. Job execution succeeded. Spark 2 executors 8 cores'. To the right, there's a 'General' tab for properties, where the 'Name' field is set to 'SQLpooltxgdedicated'. Other properties shown include 'Type: .sql script', 'Size: 215 bytes', and 'Results settings per query' with options 'First 5000 rows (default)' and 'All rows'.

3. Workspace

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and the workspace name 'tranghuyen-synapse'. Below the navigation bar, there's a toolbar with 'Synapse live' dropdown, 'Validate all' button, and a 'Publish all' button with a yellow notification badge. The main area is divided into two main sections: 'Data' and 'Linked'. The 'Data' section is further divided into 'Workspace' and 'Linked'. Under 'Workspace', there's a 'Filter resources by name' search bar and a tree view of databases. The 'Lake database' section contains a 'default' database with a 'nyctaxi' schema containing 'passengercountstats' and 'trip' tables, and a 'Views' folder. The 'SQL database' section shows one entry. On the right side, there's a preview pane for a 'demo' dataset, showing a single row with a value of 1. Below the preview pane, there's a status indicator 'Not started'.

Có 3 loại cơ sở dữ liệu:

- **SQL Database:** Cơ sở dữ liệu quan hệ, dùng cho dữ liệu có cấu trúc, hỗ trợ cả **dedicated** và **serverless SQL pool**.
- **Lake Database:** Cơ sở dữ liệu trên **Azure Data Lake Storage**, lưu trữ và phân tích dữ liệu phi cấu trúc hoặc bán cấu trúc.
- **Data Explorer Database:** Cơ sở dữ liệu phân tích dữ liệu phi cấu trúc với **Kusto Query Language (KQL)**, tối ưu cho việc phân tích log và sự kiện thời gian thực.

IV. Synapse Studio Develop hub

1. Overview

Develop Hub là một nền tảng cung cấp trải nghiệm phát triển toàn diện cho việc truy vấn, phân tích và mô hình hóa dữ liệu. Dưới đây là một số điểm nổi bật:

- Cho phép viết và thực thi các truy vấn để lấy dữ liệu từ các nguồn khác nhau. Điều này giúp dễ dàng truy xuất thông tin cần thiết cho các phân tích và báo cáo.
- Cung cấp các công cụ và tính năng để phân tích dữ liệu, bao gồm các biểu đồ, đồ thị và các công cụ thống kê. Điều này giúp hiểu rõ hơn về dữ liệu và tìm ra các xu hướng hoặc mẫu.
- Hỗ trợ trong việc xây dựng các mô hình dữ liệu phức tạp, từ đó có thể dự đoán và đưa ra các quyết định dựa trên dữ liệu. Các mô hình này có thể bao gồm từ các mô hình thống kê đơn giản đến các mô hình học máy phức tạp.

```

1  SELECT TOP (100) [ORDERNUMBER]
2  ,[QUANTITYORDERED]
3  ,[PRICEEACH]
4  ,[ORDERLINENumber]
5  ,[SALES]
6  ,[ORDERDATE]
7  ,[STATUS]
8  ,[QTR_TD]
9  ,[MONTH_ID]
10  ,[YEAR_TD]
11  ,[PRODUCTLINE]
12  ,[MSRP]
13  ,[PRODUCTCODE]
14  ,[CUSTOMERNAME]
15  ,[PHONE]
16  ,[ADDRESSLINE1]
17  ,[ADDRESSLINE2]
18  ,[CITY]
19  ,[STATE]
20  ,[POSTALCODE]
21  ,[COUNTRY]

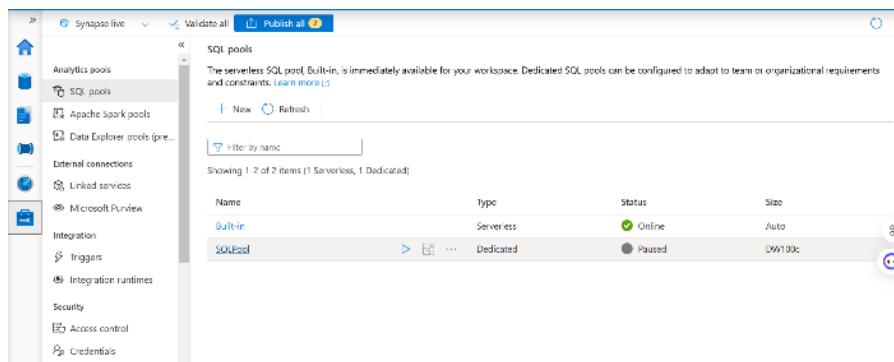
```

Results Messages
000000/ Query executed successfully.

2. SQL scripts

- Authoring SQL Scripts: Bạn có thể tạo và chỉnh sửa các script SQL một cách dễ dàng. Môi trường phát triển hỗ trợ viết mã với các tính năng như gợi ý cú pháp (IntelliSense) và kiểm tra lỗi cú pháp.
- Execute SQL Script: Bạn có thể thực thi các script SQL trên các SQL Pool đã được cung cấp hoặc SQL On-demand. Điều này cho phép bạn kiểm tra và chạy các truy vấn trực tiếp trong môi trường phát triển.
- Publish SQL Scripts: Tính năng này cho phép bạn xuất bản các script SQL riêng lẻ hoặc nhiều script cùng lúc thông qua tính năng "Publish all". Điều này rất hữu ích khi bạn cần triển khai các thay đổi lên môi trường sản xuất hoặc chia sẻ với nhóm của mình.
- Language Support and IntelliSense: Hỗ trợ nhiều ngôn ngữ và cung cấp tính năng IntelliSense để giúp bạn viết mã nhanh hơn và chính xác hơn. IntelliSense cung cấp gợi ý về cú pháp, tên bảng, cột và các hàm SQL, giúp giảm thiểu lỗi và tăng hiệu suất làm việc.

B1: Trong **Synapse Studio**, tại trang **Manage** (Quản lý), trong phần **SQL pools**, chọn hàng tương ứng với SQL pool chuyên dụng **SQLPool** và sử dụng biểu tượng để khởi động lại nó.



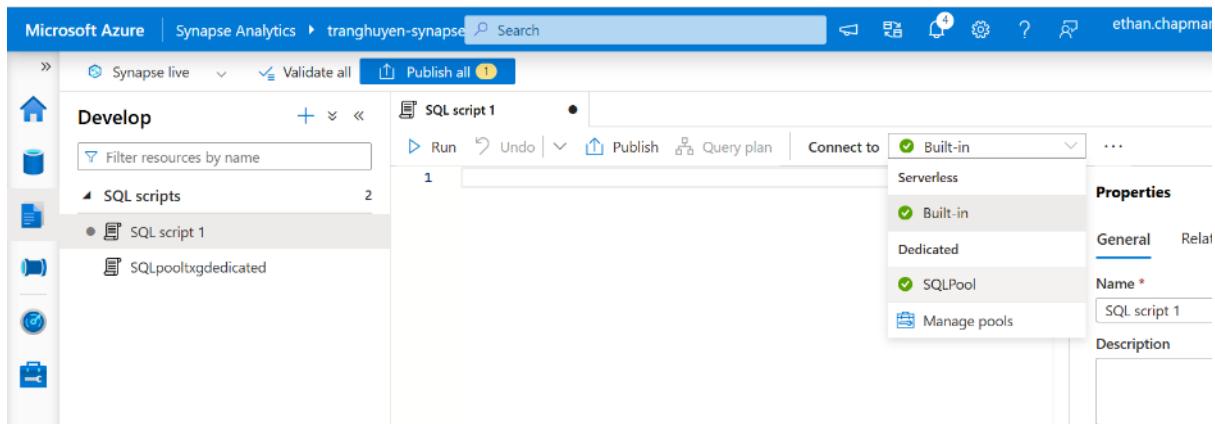
B2. Chờ cho SQL pool khởi động. Điều này có thể mất vài phút. Sử dụng nút **↻ Refresh** (Làm mới) để kiểm tra trạng thái định kỳ. Trạng thái sẽ hiển thị là **Online** khi sẵn sàng.

Name	Type	Status	Size
Built_in	Serverless	Online	Auto
SQLPool	Dedicated	Online	DW100c

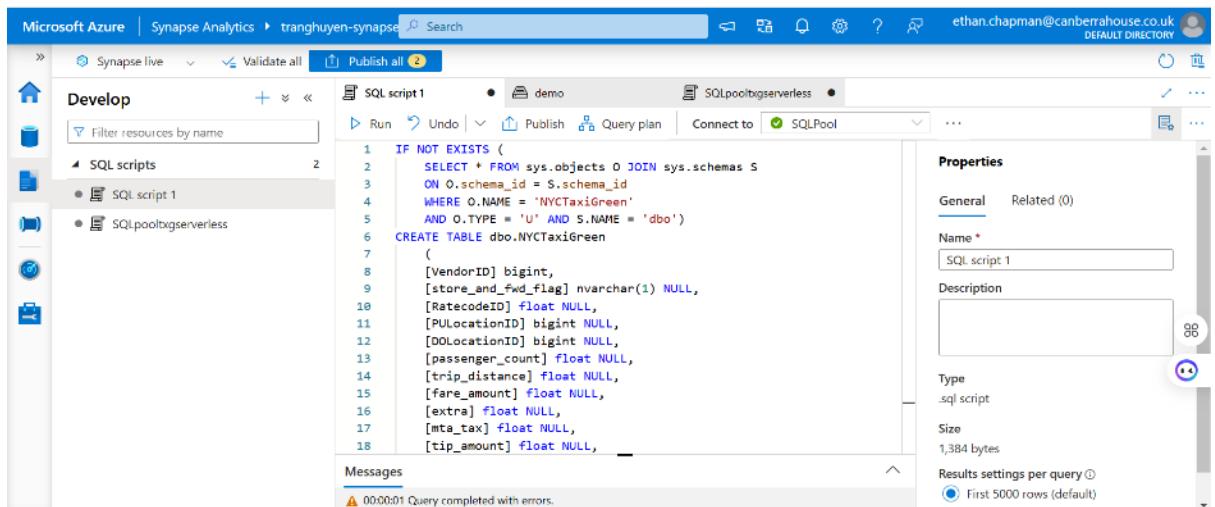
B3. Khi SQL pool đã khởi động, chọn trang **Data** (Dữ liệu); và trong tab **Workspace** (Không gian làm việc), mở rộng **SQL databases** (Cơ sở dữ liệu SQL) và kiểm tra xem **SQLPool** có được liệt kê hay không (nếu cần, sử dụng biểu tượng **↻** ở góc trên bên trái của trang để làm mới hiển thị).

B4: Develop hub, chọn nút + sau đó chọn create new SQL script.

B5: ở danh sách Connect to chọn SQLPool



B6: Nhập code tạo bảng NYCTaxiGreen và chọn run

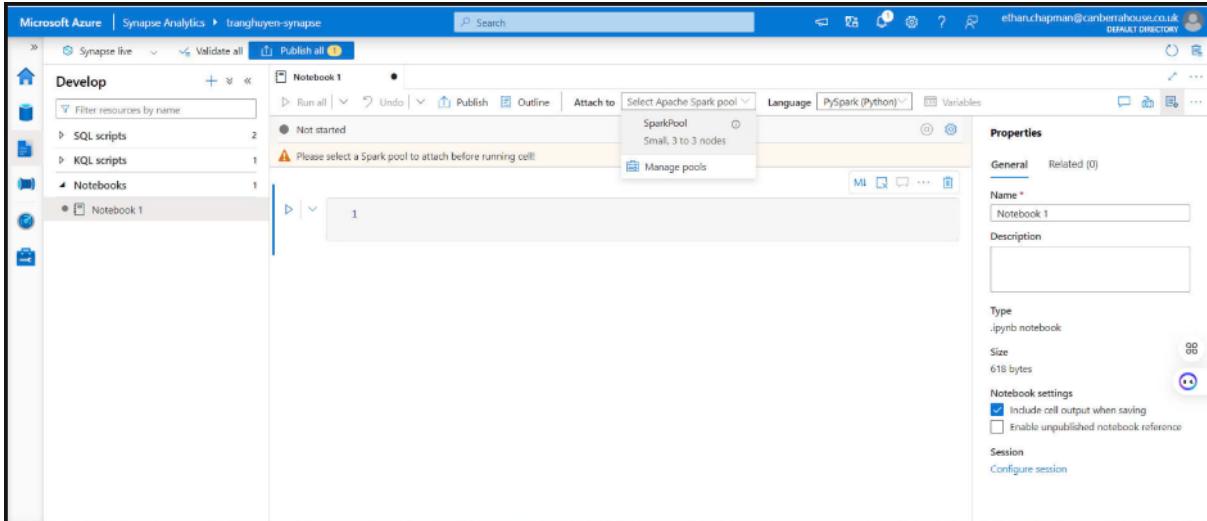


B7: Nhập tên lưu file script và chọn publish để lưu vào develop

3. KQL script

KQL (Kusto Query Language) là ngôn ngữ truy vấn dùng trong Azure Data Explorer và Azure Monitor để phân tích dữ liệu lớn nhanh chóng.

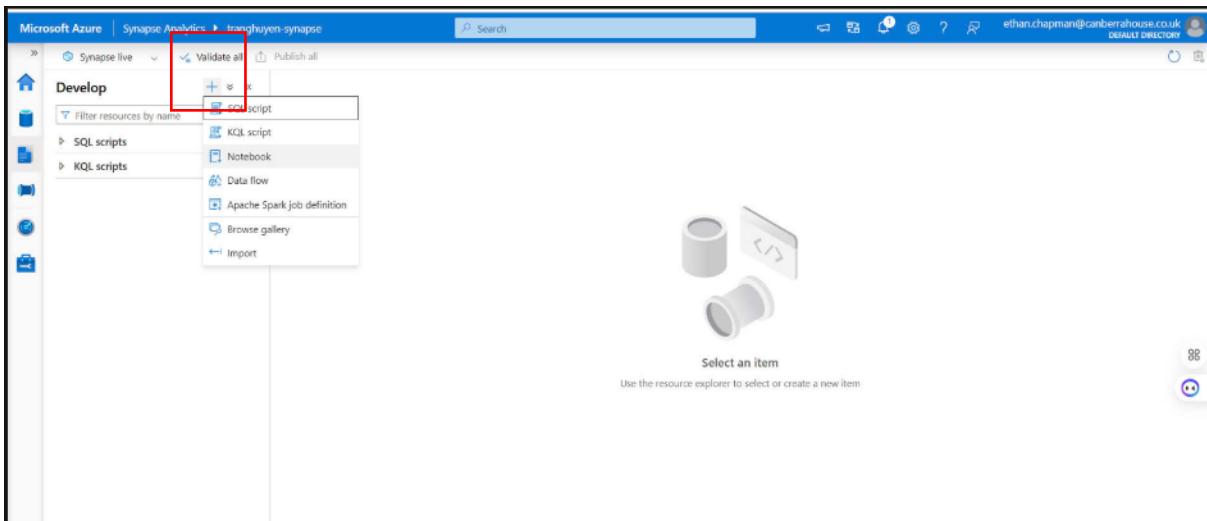
4. Notebooks



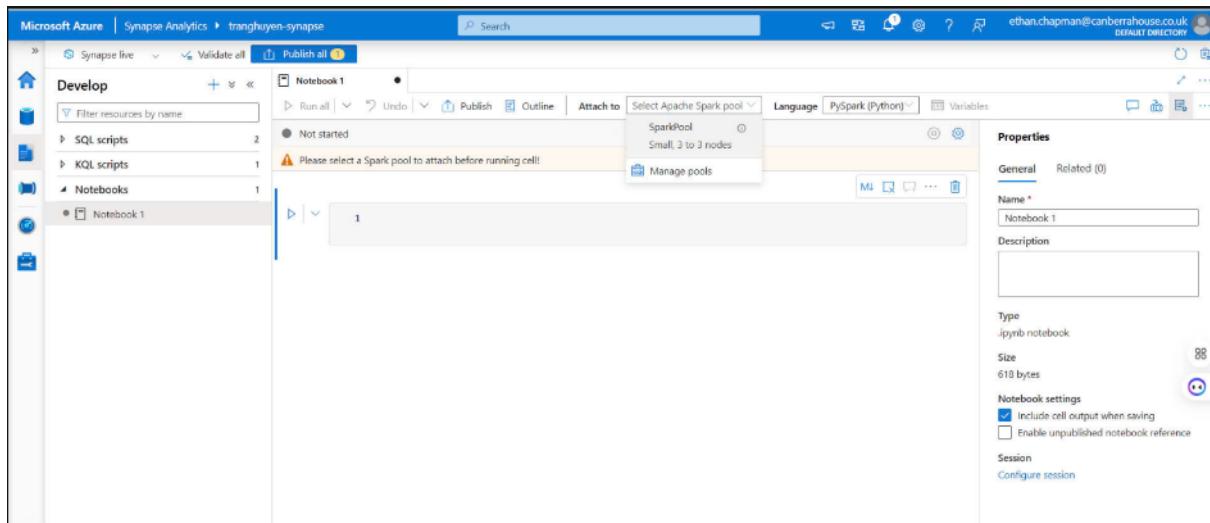
Tính năng của Develop Hub trong Azure Synapse Analytics. Đây là một công cụ mạnh mẽ cho phép bạn viết và thực thi mã trong nhiều ngôn ngữ khác nhau trong cùng một notebook. Dưới đây là một số tính năng chính:

- Hỗ trợ đa ngôn ngữ: Bạn có thể viết mã trong các ngôn ngữ khác nhau trong cùng một notebook bằng cách sử dụng lệnh % %<tên ngôn ngữ>.
- Bảng tạm thời: Bạn có thể tạo và sử dụng các bảng tạm thời giữa các ngôn ngữ khác nhau.
- Tính năng mã nâng cao: Develop Hub cung cấp các tính năng như tô sáng cú pháp, phát hiện lỗi cú pháp, hoàn thành mã, thuật toán thông minh và gấp mã để cải thiện trải nghiệm viết mã của bạn.
- Xuất kết quả: Bạn có thể xuất kết quả của các tính toán và phân tích của mình.

Tạo một note book bằng cách vào develop click vào dấu + và chọn Note book



Tại Attach to chọn SparkPool và chọn language là python



Tạo mới một cell code và Nhập code và chọn run

VendorID	lpep_pickup_datetime	lpep_dropoff_datetime	store_and_fwd_flag
2	2024-02-01 00:15:48	2024-01-01 00:24:00	N
2	2024-01-31 22:59:22	2024-01-31 23:27:14	N
2	2024-02-01 00:30:29	2024-02-01 00:35:32	N
2	2024-01-31 23:56:42	2024-02-01 00:06:53	N
2	2024-02-01 00:21:14	2024-02-01 00:31:16	N
2	2024-02-01 00:06:23	2024-02-01 00:10:10	N
2	2024-02-01 00:30:22	2024-02-01 00:33:43	N

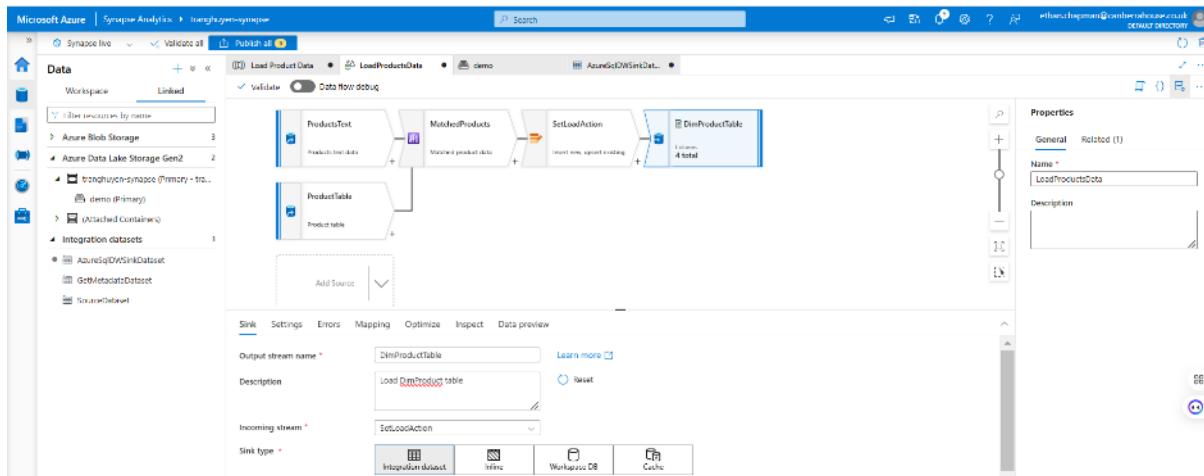
5. Data Flows

Data flow là một cách trực quan để chỉ định cách chuyển đổi dữ liệu. Cung cấp trải nghiệm không cần mã.

Khả năng của Data flow trong hệ thống xử lý dữ liệu:

- Handle upserts, updates, deletes on SQL sinks: khả năng xử lý việc thêm mới, cập nhật, hoặc xóa, dữ liệu trực tiếp trên bảng SQL đích.
- Add new partition methods: Hỗ trợ các phương pháp phân vùng mới nhằm tối ưu hóa hiệu suất khi xử lý lượng lớn dữ liệu.
- Add schema drift support: Cho phép xử lý schema drift – sự thay đổi cấu trúc dữ liệu (chẳng hạn như thêm/xóa cột) mà không cần sửa đổi thủ công cấu hình dữ liệu.
- Add file handling: Cung cấp khả năng quản lý file

- New inventory of functions: Thêm các hàm mới (ví dụ: hàm băm/hash để so sánh các hàng dữ liệu). Điều này hữu ích cho các hoạt động xác thực hoặc đồng bộ dữ liệu.
- Commonly used ETL patterns: Tích hợp các mẫu ETL thông dụng
- Data lineage: Hỗ trợ theo dõi nguồn gốc dữ liệu (data lineage)
- Implement commonly used ETL patterns as templates: Các mẫu ETL như SCD Type 1, Type 2 hoặc Data Vault có thể được cấu hình dưới dạng các template, giúp giảm thiểu thời gian thiết lập.

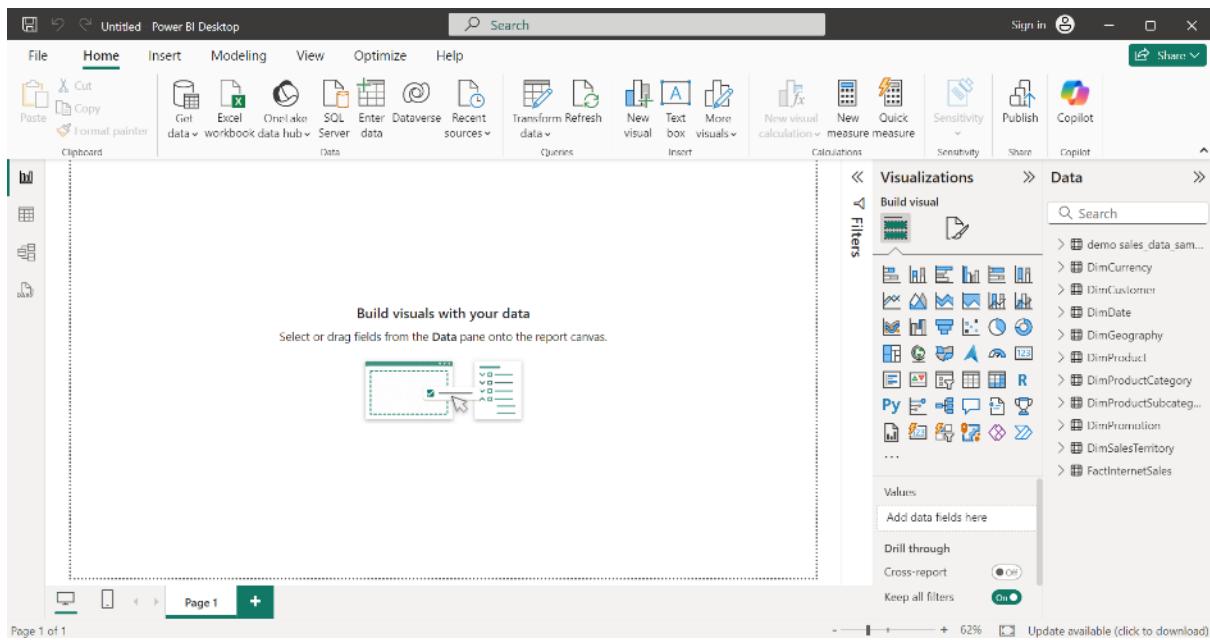


Ví dụ:

- ProductsText: Lấy dữ liệu thô về sản phẩm từ nguồn lưu trữ.
- ProductTable: Lấy dữ liệu sản phẩm hiện tại từ bảng SQL.
- MatchedProducts: So sánh dữ liệu mới với dữ liệu cũ, xác định các bản ghi mới hoặc đã thay đổi.
- SetLoadAction: Gắn nhãn cho các bản ghi (Insert hoặc Update).
- DimProductTable: Ghi dữ liệu đã xử lý vào bảng đích để sử dụng.

6. Power BI

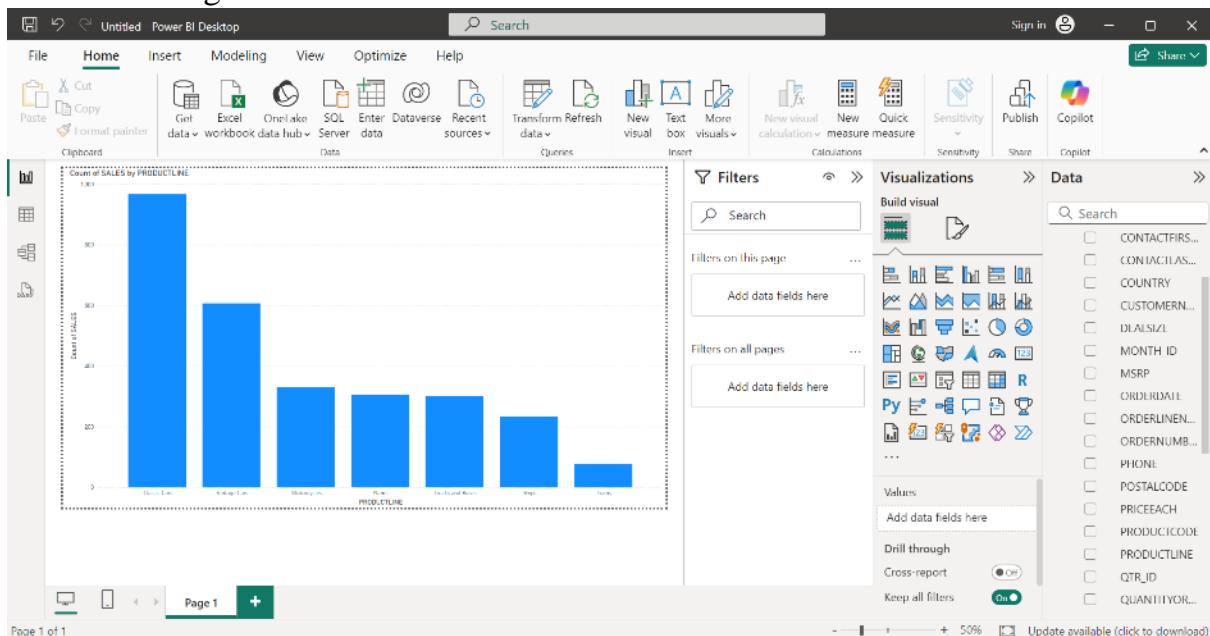
The screenshot shows the Microsoft Azure Synapse Analytics workspace under the 'cuoily workspace' tab. On the left, the 'Linked services' section is selected, displaying a list of three services: 'cuoily-workspace-WorkspaceDefault...', 'cuoily-workspace-WorkspaceDefault...', and 'PowerBIWorkspace1'. The 'PowerBIWorkspace1' service is highlighted with a yellow icon. The main pane shows a summary of the workspace, including the number of datasets (2), pipelines (1), and notebooks (0).



Azure Synapse Analytics cung cấp tích hợp mạnh mẽ với **Power BI**, cho phép tạo và quản lý báo cáo trực tiếp từ workspace. Tính năng này hỗ trợ phân tích dữ liệu trực quan và kết nối liền mạch giữa hai nền tảng.

Tính năng chính:

- Tạo báo cáo Power BI trong Synapse Workspace: người dùng có thể tạo báo cáo Power BI trực tiếp từ Synapse Analytics workspace mà không cần chuyển đổi nền tảng.



- Truy cập các báo cáo đã xuất bản: hỗ trợ truy cập và quản lý các báo cáo đã được xuất bản trong Power BI workspace.

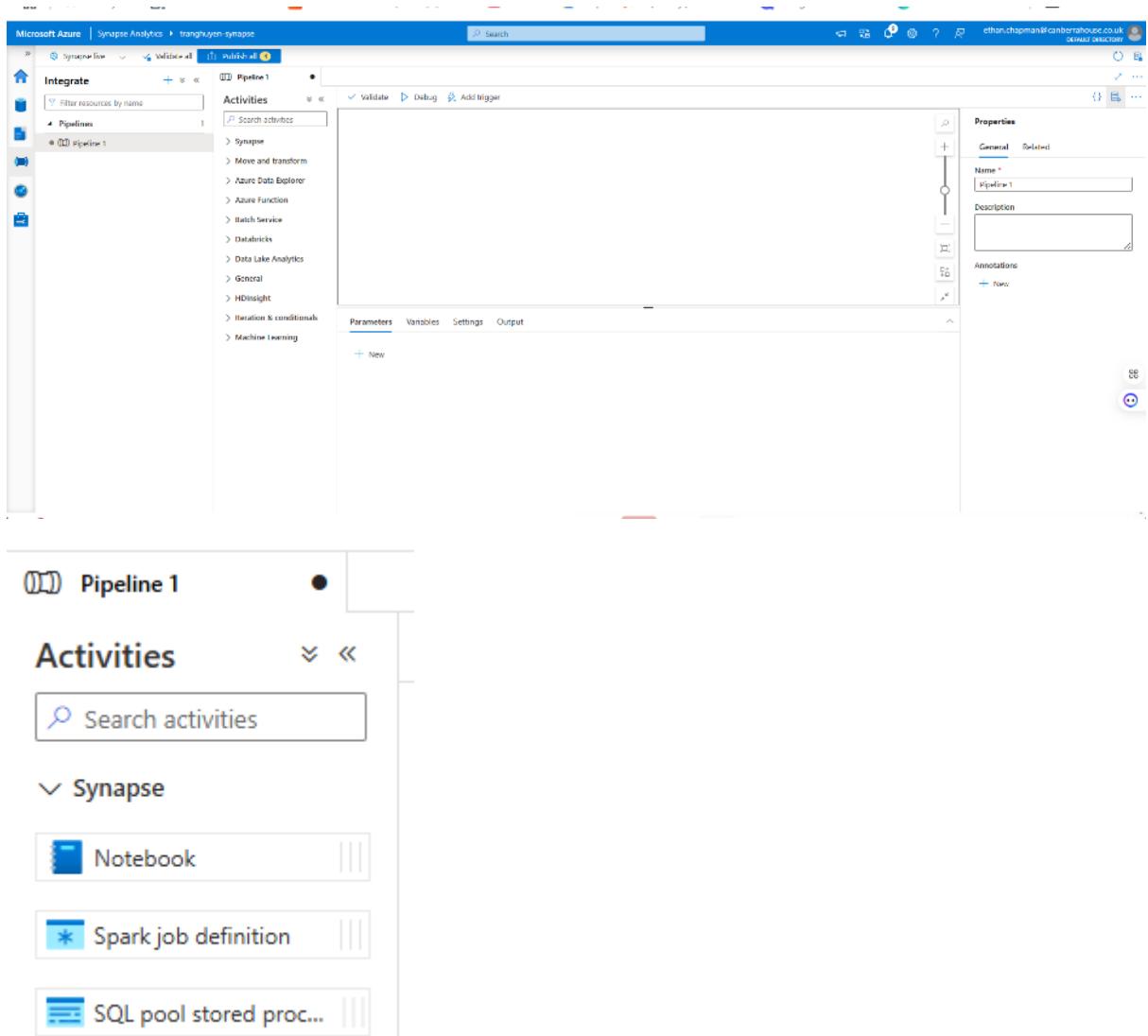
- Cập nhật báo cáo theo thời gian thực: khi dữ liệu trong Synapse Analytics được cập nhật, các thay đổi này sẽ được phản ánh ngay trong Power BI service.
- Khám phá và phân tích dữ liệu trực quan: Người dùng có thể thực hiện phân tích dữ liệu sâu bằng các công cụ trực quan mạnh mẽ của Power BI.

V. Synapse Studio Integrate hub

Cung cấp khả năng tạo các pipeline để thu thập, chuyển đổi và tải dữ liệu với hơn 90+ trình kết nối tích hợp sẵn.

Nó cung cấp một loạt các hoạt động mà pipeline có thể thực hiện, chẳng hạn như:

- Thu thập dữ liệu từ nhiều nguồn khác nhau (on-premises, cloud).
- Chuyển đổi dữ liệu với các thao tác phức tạp sử dụng Data Flow hoặc các dịch vụ xử lý dữ liệu bên ngoài (như Azure Databricks, Azure HDInsight).
- Tải dữ liệu vào kho lưu trữ đích như Azure SQL Database, Azure Data Lake Storage, hoặc Synapse Analytics.



Notebook:

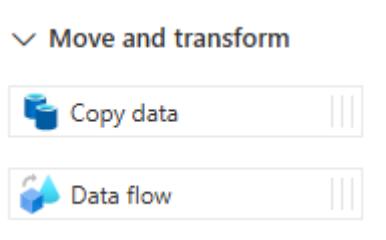
- Kết nối và thực thi các Notebook trong Azure Synapse Analytics.
- Thường dùng để chạy các bước phân tích dữ liệu hoặc xử lý dữ liệu với mã viết bằng PySpark, Scala hoặc SQL.

Spark job definition:

- Kích hoạt các công việc (job) Spark đã được định nghĩa trong Azure Synapse.
- Thích hợp để xử lý dữ liệu lớn, phân tán trên Spark Cluster.

SQL pool stored procedure:

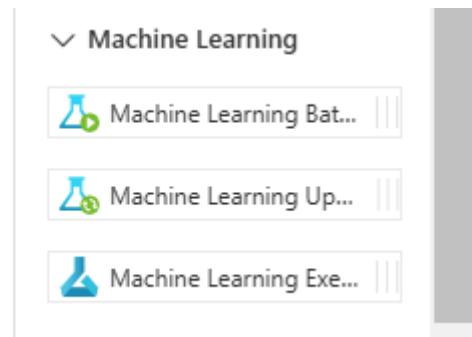
- Chạy các stored procedure (thủ tục lưu trữ) trong SQL pool của Azure Synapse.
- Được sử dụng để thực hiện các thao tác dữ liệu như truy vấn, cập nhật, hoặc xử lý dữ liệu trong cơ sở dữ liệu SQL.



Move and Transform: Di chuyển dữ liệu từ vị trí này sang vị trí khác, thay đổi định dạng hoặc cấu trúc dữ liệu trong quá trình xử lý, áp dụng các biến đổi như lọc, tổng hợp, hoặc làm sạch dữ liệu

Copy data: Tạo ra một bản sao chính xác của dữ liệu của bạn, Giữ nguyên nguồn dữ liệu gốc, hữu ích cho mục đích sao lưu hoặc tạo môi trường thử nghiệm, thường là một thao tác một lần hoặc theo lịch trình

Data flow: Sự di chuyển liên tục của dữ liệu giữa các hệ thống, truyền dữ liệu theo thời gian thực hoặc gần thời gian thực, thường được sử dụng cho dữ liệu trực tuyến hoặc đồng bộ hóa liên tục, có thể bao gồm các biến đổi tự động khi dữ liệu di chuyển



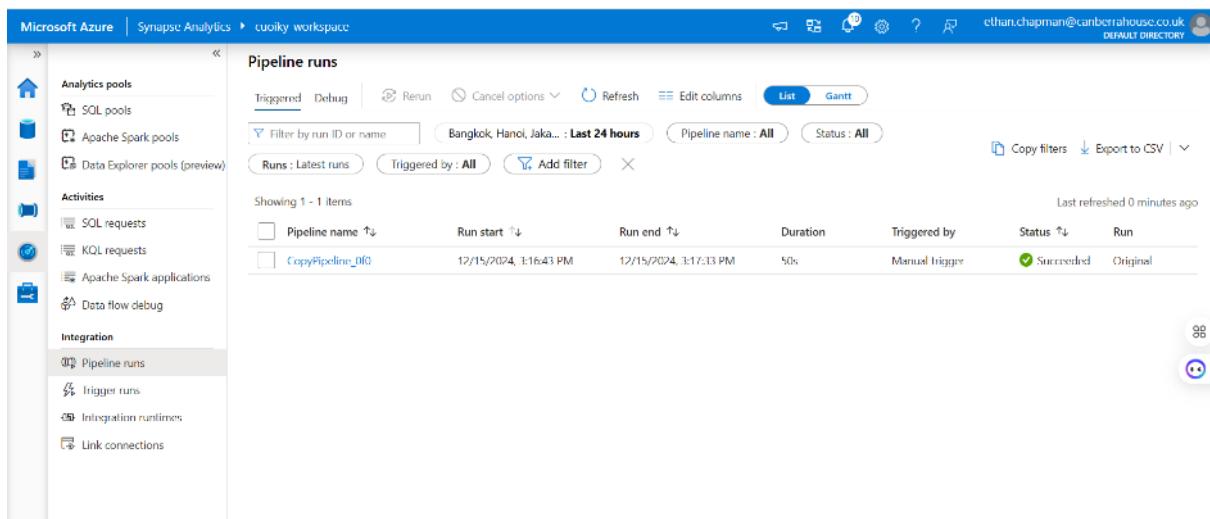
Machine Learning Batch (Học Máy theo Lô): Xử lý dữ liệu theo nhóm lớn, thường được sử dụng cho việc huấn luyện mô hình với dữ liệu lịch sử, phù hợp cho các tập dữ liệu lớn, thực hiện theo định kỳ hoặc theo lịch trình

Machine Learning Up (Cập nhật Học Máy): Cập nhật và tinh chỉnh mô hình học máy, Cho phép mô hình học từ dữ liệu mới, cải thiện độ chính xác của mô hình theo thời gian, thường được sử dụng để duy trì và nâng cao hiệu suất mô hình

Machine Learning Exe (Thực thi Học Máy): Thực thi và triển khai mô hình học máy, sử dụng mô hình đã được huấn luyện để dự đoán hoặc phân tích, áp dụng mô hình vào dữ liệu mới, tạo ra kết quả và dự đoán trong thời gian thực.

VI. Synapse Studio Monitor hub

Tính năng này cung cấp khả năng giám sát việc điều phối, hoạt động và tính toán tài nguyên.



The screenshot shows the Microsoft Azure Synapse Analytics Monitor hub interface. The left sidebar lists categories: Analytics pools (SQL pools, Apache Spark pools, Data Explorer pools (preview)), Activities (SQL requests, KQL requests, Apache Spark applications, Data flow debug), and Integration (Pipeline runs, Trigger runs, Integration runtimes, Link connections). The main area is titled 'Pipeline runs' with tabs for 'Triggered' and 'Debug'. It includes filters for 'Run ID or name' (Bangkok, Hanoi, Jakarta... : Last 24 hours), 'Pipeline name' (All), and 'Status' (All). Below the filters are buttons for 'Rerun', 'Cancel options', 'Refresh', 'Edit columns', and 'List' (selected) or 'Gantt'. A search bar and a refresh button are also present. The table below shows one item: 'CopyPipeline_0f0' with a run start of 12/15/2024, 1:16:43 PM, a run end of 12/15/2024, 1:17:13 PM, a duration of 50s, triggered by 'Manual trigger', status 'Succeeded', and run 'Original'. The table has columns: Pipeline name, Run start, Run end, Duration, Triggered by, Status, and Run. The status column for the row shows a green checkmark. The bottom right corner of the interface shows a small circular icon with a question mark and a refresh symbol.

Analytics Pools

- SQL Pools: Quản lý và theo dõi các truy vấn SQL trong Synapse.
- Apache Spark Pools: Theo dõi các ứng dụng Spark chạy trên nền tảng Synapse.
- Data Explorer Pools (Preview): Quản lý và theo dõi các hoạt động trong Data Explorer.

Activities

- SQL Requests: Hiển thị lịch sử và trạng thái các truy vấn SQL.
- KQL Requests: Theo dõi các truy vấn KQL (Kusto Query Language).
- Apache Spark Applications: Xem thông tin chi tiết về các ứng dụng Spark.
- Data Flow Debug: Gỡ lỗi và theo dõi các luồng dữ liệu.

Integration

- Pipeline Runs: Theo dõi lịch sử chạy của các pipeline, bao gồm trạng thái, thời gian bắt đầu và kết thúc, cũng như thời gian thực thi.
- Trigger Runs: Quản lý và theo dõi các trigger (tác vụ tự động) liên kết với pipeline.
- Integration Runtimes: Kiểm tra trạng thái và hiệu suất của các runtime tích hợp.

- Link Connections: Theo dõi và quản lý các kết nối dữ liệu.

VII. Synapse Studio Manage hub

1. Overview

The screenshot shows the Microsoft Azure Synapse Analytics Manage hub interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and a specific workspace name 'tranhuyen-synapse'. Below the navigation is a toolbar with icons for 'Synapse live' (dropdown), 'Validate all', and 'Publish all' (button with a count of 4). The main content area has a sidebar on the left with categories: 'Analytics pools' (SQL pools, Apache Spark pools, Data Explorer pools (preview)), 'External connections' (Linked services, Microsoft Purview), 'Integration' (Triggers, Integration runtimes), 'Security' (Access control, Credentials), and a bottom section for 'Data Lake Analytics' (Jobs, Pipelines, Datasets). The right panel is titled 'Triggers' with the sub-instruction 'To execute a pipeline set the trigger. Trig...'. It features a '+ New' button, a 'Refresh' button, and a 'Filter by name' search bar.

Analytics Pools

- SQL Pools: Quản lý tài nguyên và các cơ sở dữ liệu phân tích sử dụng SQL.
- Apache Spark Pools: Quản lý cụm Spark để xử lý dữ liệu lớn và phân tích dữ liệu phi cấu trúc.
- Data Explorer Pools (Preview): Quản lý các pool sử dụng Data Explorer cho các truy vấn thời gian thực.

External Connections

- Linked Services: Quản lý các kết nối đến các nguồn dữ liệu bên ngoài như Azure Data Lake, Azure SQL Database, hoặc REST API.
- Microsoft Purview: Tích hợp với Microsoft Purview để giám sát và quản lý dữ liệu trên toàn bộ hệ sinh thái.

Integration:

- Triggers: Tạo, quản lý, và theo dõi các trigger (bộ kích hoạt) để tự động hóa các pipeline. Ví dụ: Kích hoạt pipeline theo thời gian hoặc sự kiện.
- Integration Runtimes: Quản lý các môi trường chạy (runtimes) để di chuyển và chuyển đổi dữ liệu.

Security:

- Access Control: Quản lý quyền truy cập của người dùng và nhóm trong Azure Synapse Analytics.
- Credentials: Lưu trữ và quản lý thông tin xác thực để truy cập các dịch vụ và tài nguyên.

2. Linked services

Linked Services là một thành phần định nghĩa thông tin kết nối cần thiết để Azure Synapse Analytics có thể kết nối với các **nguồn dữ liệu bên ngoài** hoặc **tài nguyên tính toán**. Nó đóng vai trò là cầu nối giữa Azure Synapse và các hệ thống khác.

- Hỗ trợ hơn 90+ trình kết nối sẵn có (Pre-built Connectors): Azure Blob Storage, azure data Lake, SQL Database, HTTP, REST APIs, và nhiều dịch vụ của bên thứ ba (SAP, Oracle, Google BigQuery).
- Di chuyển dữ liệu dễ dàng trên các nền tảng (Cross-platform Data Migration): hỗ trợ ETL (Extract, Transform, Load) để tối ưu hóa quá trình xử lý và di chuyển dữ liệu.
- Biểu diễn tài nguyên dữ liệu hoặc tính toán (Data Store or Compute Resources)

Giới thiệu giao diện

The screenshot shows the Microsoft Azure Synapse Analytics interface. The left sidebar includes options like 'Analytics pools', 'External connections', 'Linked services' (which is highlighted), and 'Integration'. The main area displays a table of linked services:

Name	Type	Related	Annotations
Data_Warehouse	Azure Synapse Analytics	0	
nyc_tlc_green	Azure Blob Storage	0	
Products	HTTP	0	
tranhuyen-synapse-Workspace...	Azure Synapse Analytics	1	
tranhuyen-synapse-Workspace...	Azure Data Lake Storage Gen2	2	

Below the table is a grid of icons representing various pre-built connectors, including SAP BW, SAP HANA, and others.

- Menu bên trái bao gồm các mục chính như:

- o **Analytics pools:** SQL pools, Apache Spark pools, data explorer pool

- **External connections:** Bao gồm **Linked services**, nơi bạn có thể thiết lập kết nối với các nguồn dữ liệu bên ngoài.
- **Integration:** Các mục liên quan đến **Triggers** và **Pipelines** để tích hợp và tự động hóa dữ liệu.
- **Security:** Tùy chọn bảo mật như **Access control** và **Credentials**.

- Danh sách các Linked Services:

- Đây là danh sách các dịch vụ liên kết mà hệ thống đã cấu hình.
- Cột chính bao gồm:
 - **Name:** Tên của dịch vụ liên kết.
 - **Type:** Loại kết nối, ví dụ: Azure Blob Storage, HTTP, Azure Data Lake Storage Gen2, v.v.
 - **Related:** Hiển thị số liên kết hiện tại của dịch vụ đến các pipeline hoặc entity khác.
 - **Annotations:** Các ghi chú hoặc nhãn thêm (nếu có).

- **Tùy chọn thêm mới (New):** Nút **New** cho phép thêm một dịch vụ liên kết mới. Khi nhấn, bạn có thể chọn loại kết nối, cấu hình các tham số như URL, tài khoản, chứng thực, v.v.

- **Filter by name:** Thanh lọc (filter) để tìm kiếm một dịch vụ liên kết cụ thể theo tên.

- **Thanh công cụ phía trên (Toolbar):** Bao gồm các tùy chọn Validate all (Kiểm tra tính hợp lệ của tất cả các dịch vụ liên kết) và publish all (xuất bản hoặc áp dụng các thay đổi cấu hình dịch vụ).

- **Cột trạng thái (Annotations/Related):** Related (Số lượng liên kết mà dịch vụ liên kết đang được sử dụng), Annotations (Các nhãn hoặc chú thích để mô tả thêm).

3. Access Control

Access Control Management trong Azure Synapse giúp quản lý quyền truy cập vào các tài nguyên và thành phần trong workspace, áp dụng cho cả quản trị viên và người dùng thông thường. Đảm bảo rằng **chỉ những người được phép** mới có thể truy cập hoặc thực hiện các thao tác trên các tài nguyên trong workspace. Cho phép quản lý quyền trên các đối tượng như **pipeline**, **datasets**, **Spark pools**, và **code artifacts**.

Lợi ích (Benefits):

- Chia sẻ workspace với nhóm (Share workspace with the team):
- Tăng năng suất (Increases productivity):
- Quản lý quyền trên code artifacts và Spark pools (Manage permissions on code artifacts and Spark pools):

Name	Type	Role	Scope
Jeannette Chavis	User	Synapse Administrator	Workspace
tranhuyen-synapse	Service principal	Synapse Administrator	Workspace

4. Triggers

Triggers trong Azure Synapse Analytics được sử dụng để định nghĩa **thời điểm** một pipeline cần được thực thi. Chúng đóng vai trò quan trọng trong việc tự động hóa và lập lịch xử lý dữ liệu.

Lợi ích (Benefits):

- Tạo và quản lý (Create and Manage): Cho phép dễ dàng tạo và quản lý ba loại trigger chính:
 - Schedule Trigger: Kích hoạt pipeline theo lịch trình định kỳ.
 - Tumbling Window Trigger: Kích hoạt theo khoảng thời gian cố định và có khả năng lưu trữ trạng thái.
 - Event Trigger: Kích hoạt khi có sự kiện xảy ra (như dữ liệu mới được tải lên).
- Kiểm soát thực thi pipeline (Control Pipeline Execution): Cung cấp khả năng giám sát (monitor) và quản lý quá trình thực thi pipeline, có thể tạm dừng, khởi động lại hoặc hủy bỏ trigger nếu cần.

Triggers					
To execute a pipeline set the trigger. Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off.					
NAME ↑↓		TYPE ↑↓	STATUS ↑↓	NUMBER OF PIPELINES ↑↓	ANNOTATIONS ↑↓
* CopyParquetDataTrigger	II				Started 1
* Trigger 1					Stopped 0

5. Integration runtimes

Integration runtimes					
The Integration Runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environments:					
NAME ↑↓		TYPE ↑↓	SUB-TYPE ↑↓	STATUS ↑↓	REGION ↑↓
AutoResolveIntegrationRuntime			Azure	Public 	Running Auto Resolve

Integration Runtimes (IR) là hạ tầng tính toán trong Azure Synapse Analytics được sử dụng bởi **Pipelines** để hỗ trợ tích hợp dữ liệu giữa các môi trường mạng khác nhau. Nó đóng vai trò là cầu nối giữa **activities** (các hoạt động) và **linked services** (dịch vụ liên kết) trong quá trình xử lý dữ liệu.

Lợi ích (Benefits):

Hỗ trợ hai loại Integration Runtime chính:

- Azure Integration Runtime: Hạ tầng tính toán được quản lý hoàn toàn bởi Azure, Serverless (không cần cài đặt, quản lý hoặc bảo trì máy chủ). Sử dụng để xử lý dữ liệu trong đám mây Azure hoặc giữa các dịch vụ đám mây.
- Self-Hosted Integration Runtime: Sử dụng tài nguyên tính toán trên máy chủ nội bộ (on-premises) hoặc máy ảo (VM) trong mạng riêng, phù hợp với các tác vụ cần kết nối tới hệ thống nội bộ hoặc trong môi trường mạng giới hạn (VPN, ExpressRoute).

Cung cấp tính linh hoạt trong tích hợp dữ liệu: Hỗ trợ kịch bản tích hợp dữ liệu đám mây đến đám mây (cloud-to-cloud), đám mây đến nội bộ (hybrid), và hoàn toàn nội bộ (on-premises).

Dễ dàng cấu hình và quản lý: Azure cung cấp giao diện quản lý đơn giản để thiết lập và giám sát hiệu suất.

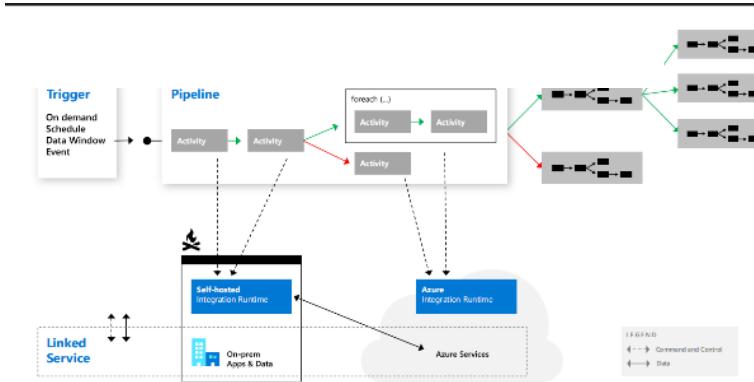
PHẦN 4: AZURE SYNAPSE ANALYTICS DATA INTEGRATION

I. Data integration

Azure Data Factory (ADF) là dịch vụ tích hợp và sắp xếp dữ liệu dựa trên đám mây do Microsoft Azure cung cấp. Dịch vụ này cung cấp nền tảng để thu thập, chuẩn bị, chuyển đổi và phân phối dữ liệu từ nhiều nguồn khác nhau đến các đích khác nhau. ADF cho phép các tổ chức tạo quy trình làm việc dựa trên dữ liệu và tự động hóa các tác vụ tích hợp và xử lý dữ liệu một cách hiệu quả.

Tổng quan về vai trò của nó trong tích hợp và xử lý dữ liệu

Azure Data Factory đóng vai trò quan trọng trong việc cho phép tích hợp và xử lý dữ liệu liền mạch trên nhiều nguồn khác nhau, cả tại chỗ và trên đám mây. Nó hoạt động như một trung tâm để quản lý và điều phối các hoạt động di chuyển và chuyển đổi dữ liệu, cho phép các tổ chức hợp nhất và xử lý dữ liệu để phân tích, báo cáo và các sáng kiến khác dựa trên dữ liệu.



II. Data Movement

Data Movement trong Azure Synapse Analytics cung cấp các khả năng mạnh mẽ để xử lý và di chuyển dữ liệu giữa các hệ thống khác nhau. Dưới đây là mô tả chi tiết các tính năng và lợi ích của **Data Movement**:

Lợi ích và tính năng của Data Movement:

- **Scalable (Có thể mở rộng):** Elasticity per job có thể tăng hoặc giảm tài nguyên xử lý dựa trên khối lượng công việc, giúp tối ưu hóa hiệu suất và chi phí. **Up to**

4 GB/s tốc độ di chuyển dữ liệu có thể đạt đến 4GB mỗi giây, cho phép xử lý dữ liệu với khối lượng lớn một cách nhanh chóng.

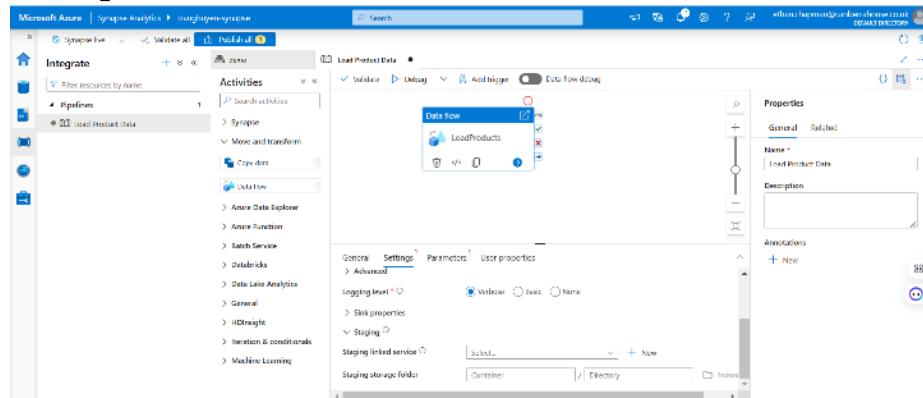
- **Simple (Đơn giản):** Visually author or via code (Python, .Net, etc)
- **Serverless (Không cần quản lý hạ tầng):** Không cần phải quản lý hoặc duy trì bất kỳ hạ tầng máy chủ nào. Azure cung cấp môi trường không máy chủ, tự động quản lý tài nguyên và mở rộng quy mô khi cần thiết, giúp giảm bớt gánh nặng quản lý hạ tầng cho người dùng.

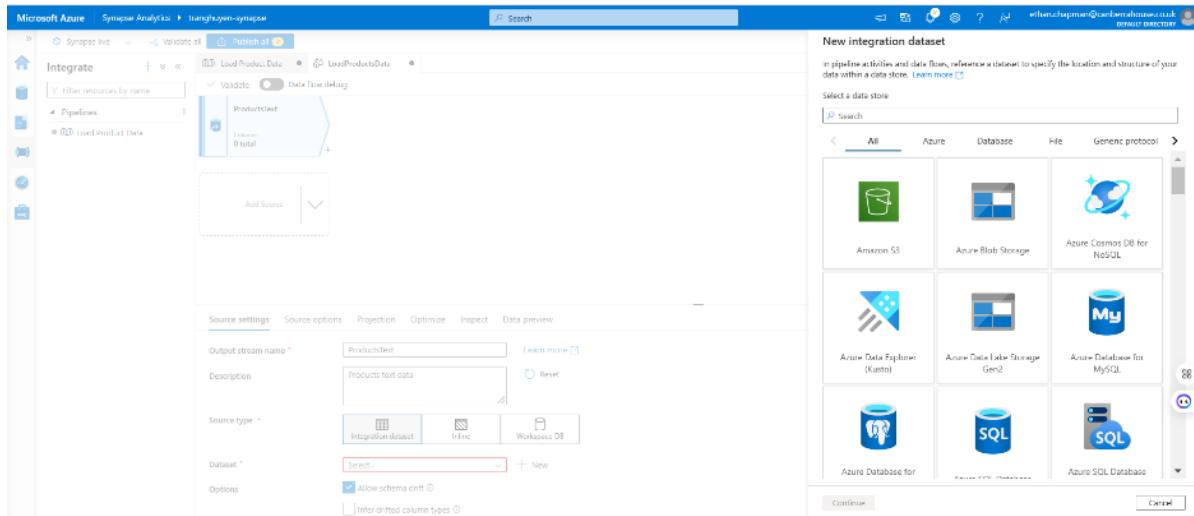
Access all your data (Truy cập toàn bộ dữ liệu): Azure Synapse cung cấp hơn 90 kết nối sẵn có, bao gồm cả các nguồn dữ liệu đám mây, on-premises và SaaS, cho phép bạn di chuyển dữ liệu từ và đến nhiều hệ thống khác nhau.

- **Cloud:** Dữ liệu từ các dịch vụ đám mây như Azure Blob Storage, AWS S3, Google Cloud Storage.
- **On-premises:** Dữ liệu từ các hệ thống nội bộ như cơ sở dữ liệu SQL Server, Oracle, SAP.
- **SaaS:** Dữ liệu từ các ứng dụng SaaS như Salesforce, Dynamics 365, Google Analytics.

Data Movement as a Service: Azure Synapse có 25 điểm hiện diện trên toàn cầu, giúp tăng cường khả năng truy cập và di chuyển dữ liệu ở phạm vi rộng lớn, tối ưu hóa hiệu suất di chuyển dữ liệu toàn cầu. Hỗ trợ sử dụng **Self-hosted Integration Runtime**, cho phép di chuyển dữ liệu giữa các hệ thống đám mây và hệ thống nội bộ (hybrid data movement), cung cấp tính linh hoạt khi xử lý dữ liệu từ các môi trường khác nhau.

III. Pipelines



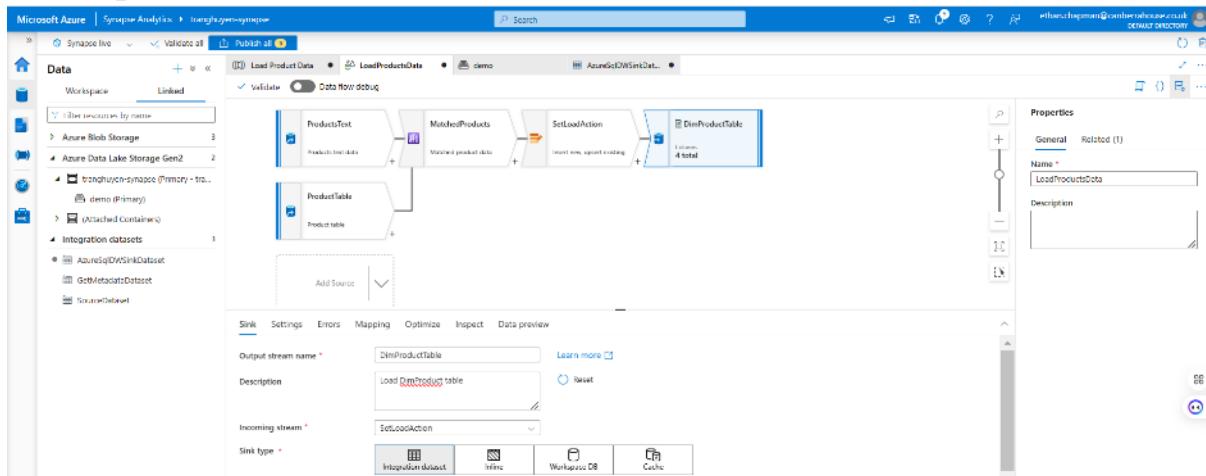


Pipelines trong Azure Synapse Analytics là một thành phần quan trọng trong việc xử lý và di chuyển dữ liệu giữa các nguồn và đích, cho phép người dùng tự động hóa quy trình tải dữ liệu. Pipelines cung cấp khả năng tải dữ liệu từ các tài khoản lưu trữ (storage accounts) đến các dịch vụ liên kết (linked services) mong muốn, giúp dễ dàng di chuyển dữ liệu giữa các hệ thống khác nhau. Dữ liệu có thể được tải bằng cách thực thi thủ công pipeline hoặc thông qua việc sử dụng cơ chế điều phối (orchestration) để tự động hóa quy trình.

Lợi ích của Pipelines:

- Hỗ trợ các mẫu tải dữ liệu phổ biến
- Tải dữ liệu song song (Parallel loading) : Pipelines hỗ trợ việc tải dữ liệu song song vào các kho dữ liệu như **data lake** hoặc các bảng **SQL**, giúp tối ưu hóa tốc độ và hiệu quả của quá trình tải dữ liệu, đặc biệt khi xử lý khối lượng lớn dữ liệu.
- Trải nghiệm phát triển đồ họa (Graphical development experience): Pipelines cung cấp một **trải nghiệm phát triển đồ họa** trực quan, giúp người dùng dễ dàng thiết kế và xây dựng các quy trình tải dữ liệu mà không cần phải viết mã. Giao diện đồ họa giúp giảm bớt sự phức tạp khi tạo ra các quy trình phức tạp và dễ dàng theo dõi trạng thái thực thi của pipeline..

IV. Prep & Transform Data (Data flow)



Data Flows (luồng dữ liệu): cho phép các tổ chức thiết kế trực quan và thực hiện các tác vụ chuyển đổi và xử lý dữ liệu bằng cách sử dụng phương pháp không cần mã (a code-free approach). Chúng cung cấp giao diện đồ họa để xác định transformations (chuyển đổi) dữ liệu, aggregations (thu thập), filtering (lọc), and schema mappings (ánh xạ lược đồ).

V. Triggers

NAME	TYPE	STATUS	NUMBER OF PIPELINES	ANNOTATIONS
* CopyParquetDataTrigger	Schedule	Started	1	
* Trigger 1	Schedule	Stopped	0	

Triggers: xác định lịch trình thực hiện hoặc sự kiện khởi tạo việc thực hiện một pipeline. Cho phép các tổ chức tự động hóa và lên lịch các quy trình tích hợp dữ liệu dựa trên các yêu cầu cụ thể của họ, có 4 loại trigger.

- **On-demand:** Kích hoạt thủ công khi người dùng yêu cầu.
- **Schedule:** Kích hoạt tự động theo lịch trình đã định.
- **Data Window:** Kích hoạt dựa trên phạm vi thời gian của dữ liệu.
- **Event:** Kích hoạt khi một sự kiện nhất định xảy ra.

VI. Manage – Linked Services

Dịch vụ liên kết (Linked Services): thiết lập kết nối đến nhiều nguồn dữ liệu và đích đến khác nhau, bao gồm cơ sở dữ liệu, hệ thống tệp, lưu trữ đám mây và ứng dụng SaaS. Chúng cung cấp thông tin chi tiết về cấu hình và xác thực cần thiết để kết nối và tương tác với các nguồn dữ liệu này.

Linked services

Linked services are much like connection strings, which define the connection information needed for Azure Synapse Analytics to connect to external resources.

New

Showing 1 - 5 of 5 items

Name	Type	Related	Annotations
Data_Warehouse	Azure Synapse Analytics	0	
nyc_llc_green	Azure Blob Storage	0	
Products	HTTP	0	
tranhuyen synapse Workspace	Azure Synapse Analytics	1	
tranhuyen-synapse-Workspace	Azure Data Lake Storage Gen2	2	

New linked service

PayPal [Preview]	Phoenix	PostgreSQL
Power BI	Presto [Preview]	QuickBooks [Preview]
REST	SAP BW	SAP BW
C4C	SAP ECC	SAP HANA
SAP Cloud for Commerce	SAP ECC	SAP HANA
SAP		

VII. Manage – Integration runtimes

Integration Runtime (IR) là phần tính toán giúp kết nối giữa các hoạt động trong pipeline và các linked services (dịch vụ liên kết), đảm bảo rằng dữ liệu được xử lý và di chuyển một cách hiệu quả giữa các nguồn dữ liệu khác nhau.

Lợi ích của Integration Runtime

- Azure Integration Runtime (AIR): Cung cấp môi trường tính toán serverless (không cần quản lý hạ tầng) được quản lý hoàn toàn trong Azure. Đây là một giải pháp dễ dàng và tự động hóa cho các tác vụ xử lý và di chuyển dữ liệu trong môi trường đám mây.
- Self-Hosted Integration Runtime (SHIR): Sử dụng tài nguyên tính toán từ máy chủ on-premises (tại chỗ) hoặc máy ảo (VM) trong mạng riêng. Cho phép kết nối dữ liệu từ môi trường on-premises hoặc mạng riêng.



(private network) với các dịch vụ đám mây mà không cần phải di chuyển tất cả dữ liệu lên đám mây.

PHẦN 5: AZURE SYNAPSE ANALYTICS SQL ANALYTICS

I. Analytics

Azure SQL Data Warehouse cung cấp hiệu suất hàng đầu với giá thành tối ưu, nhờ tích hợp với hệ sinh thái Azure và cải tiến từ SQL Server, mang lại:

- **Bộ nhớ đệm Gen2:** Dùng ổ NVMe tăng băng thông I/O cho truy vấn.
- **Mạng tăng tốc bằng FPGA:** Truyền dữ liệu lên đến 1GB/giây mỗi node.
- **Di chuyển dữ liệu tức thì:** Sử dụng đa lõi để di chuyển dữ liệu giữa các node hiệu quả.
- **Tối ưu hóa truy vấn:** Cải tiến liên tục trong tối ưu hóa truy vấn phân tán.

Azure Synapse là hệ thống phân tích đầu tiên và duy nhất thực hiện tất cả các truy vấn TPC-H ở quy mô 1 petabyte.

1. Comprehensive SQL functionality (Chức năng SQL toàn diện).

- Hệ thống lưu trữ tiên tiến bao gồm:
 - **Columnstore Indexes:** Tăng hiệu suất truy vấn cho dữ liệu lớn.
 - **Table Partitions:** Chia bảng thành các phần nhỏ để quản lý và truy vấn hiệu quả hơn.
 - **Distributed Tables:** Phân phối dữ liệu trên nhiều node để tối ưu hóa xử lý.
 - **Isolation Modes:** Hỗ trợ các mức độ lập để đảm bảo tính toàn vẹn dữ liệu.
 - **Materialized Views:** Lưu trữ kết quả truy vấn để cải thiện hiệu suất.
 - **Nonclustered Indexes:** Cải thiện tốc độ truy vấn không thay đổi cấu trúc bảng.
 - **Result-set Caching:** Lưu trữ kết quả truy vấn để truy cập nhanh hơn.
- **Truy vấn T-SQL** cung cấp các tính năng mạnh mẽ như:
 - **Windowing Aggregates:** Thực hiện tính toán tổng hợp (SUM, AVG, COUNT, v.v.) trên các cửa sổ dữ liệu cụ thể.
 - **Approximate Execution (HyperLogLog):** Hỗ trợ tính toán gần đúng, đặc biệt hữu ích với tập dữ liệu lớn.
 - **JSON Data Support:** Tích hợp và xử lý dữ liệu JSON trực tiếp trong truy vấn SQL.
- **Mô hình đối tượng SQL hoàn chỉnh**, bao gồm:
 - **Tables:** Lưu trữ dữ liệu có cấu trúc.
 - **Views:** Tạo bảng ảo dựa trên truy vấn để dễ dàng truy cập dữ liệu.
 - **Stored Procedures:** Tập hợp các câu lệnh SQL được lưu trữ và tái sử dụng.

- **Functions:** Các hàm định nghĩa để thực hiện các phép tính hoặc thao tác trên dữ liệu.

2. Windowing functions

- **Mệnh đề OVER:**

- Xác định một window hoặc tập hợp các hàng cụ thể trong kết quả truy vấn.
- Tính toán một giá trị cho mỗi hàng trong cửa sổ được chỉ định.

- **Aggregate functions:** COUNT, MAX, AVG, SUM, APPROX_COUNT_DISTINCT, MIN, STDEV, STDEVP, STRING_AGG, VAR, VARP, GROUPING, GROUPING_ID, COUNT_BIG, CHECKSUM_AGG

- **Ranking functions:** RANK, NTILE, DENSE_RANK, ROW_NUMBER

- **Analytical functions:** LAG, LEAD, FIRST_VALUE, LAST_VALUE, CUME_DIST, PERCENTILE_CONT, PERCENTILE_DISC, PERCENT_RANK

```
SELECT
    ROW_NUMBER() OVER(PARTITION BY PostalCode ORDER BY SalesYTD DESC
) AS "Row Number",
    LastName,
    SalesYTD,
    PostalCode
FROM Sales
WHERE SalesYTD <> 0
ORDER BY PostalCode;
```

Row Number	LastName	SalesYTD	PostalCode
1	Mitchell	4251368.5497	98027
2	Blythe	3763178.1787	98027
3	Carson	3189418.3662	98027
4	Reiter	2315185.611	98027
5	Vargas	1453719.4653	98027
6	Anzman-Wolfe	1352577.1325	98027
1	Pak	4116870.2277	98055
2	Varkey Chudukakti	3121616.3202	98055
3	Saraiva	2604540.7172	98055
4	Ito	2458535.6169	98055
5	Valdez	1827066.7118	98055
6	Mensa-Annan	1576562.1966	98055
7	Campbell	1573012.9383	98055
8	Tsoflias	1421810.9242	98055

```
-- PERCENTILE_CONT, PERCENTILE_DISC
SELECT DISTINCT Name AS DepartmentName
,PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY ph.Rate)
    OVER (PARTITION BY Name) AS MedianCont
,PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY ph.Rate)
    OVER (PARTITION BY Name) AS MedianDisc
FROM HumanResources.Department AS d
INNER JOIN HumanResources.EmployeeDepartmentHistory AS dh
    ON dh.DepartmentID = d.DepartmentID
INNER JOIN HumanResources.EmployeePayHistory AS ph
    ON ph.BusinessEntityID = dh.BusinessEntityID
WHERE dh.EndDate IS NULL;
```

DepartmentName	MedianCont	MedianDisc
Document Control	16.8269	16.8269
Engineering	34.375	32.6923
Executive	54.32695	48.5577
Human Resources	17.427850	16.5865

```
-- LAG Function
SELECT BusinessEntityID,
YEAR(QuotaDate) AS SalesYear,
SalesQuota AS CurrentQuota,
LAG(SalesQuota, 1,0) OVER (ORDER BY YEAR(QuotaDate)) AS PreviousQuota
FROM Sales.SalesPersonQuotaHistory
WHERE BusinessEntityID = 275 and YEAR(QuotaDate) IN ('2005','2006');
```

BusinessEntityID	SalesYear	CurrentQuota	PreviousQuota
275	2005	367000.00	0.00
275	2005	556000.00	367000.00
275	2006	502000.00	556000.00
275	2006	550000.00	502000.00
275	2006	1429000.00	550000.00
275	2006	1324000.00	1429000.00

- **ROWS | RANGE: PRECEDING, UNBOUNDING PRECEDING, CURRENT ROW, BETWEEN, FOLLOWING, UNBOUNDED FOLLOWING**

```
-- First_Value
SELECT JobTitle, LastName, VacationHours AS VacHours,
FIRST_VALUE(LastName) OVER (PARTITION BY JobTitle
ORDER BY VacationHours ASC ROWS UNBOUNDED PRECEDING ) AS
FewestVacHours
FROM HumanResources.Employee AS e
INNER JOIN Person.Person AS p
ON e.BusinessEntityID = p.BusinessEntityID
ORDER BY JobTitle;
```

JobTitle	LastName	VacHours	FewestVacHours
Accountant	Moreland	58	Moreland
Accountant	Seamans	59	Moreland
Accounts Manager	Liu	57	Liu
Accounts Payable Specialist	Tomic	63	Tomic
Accounts Payable Specialist	Sheperdigian	64	Tomic
Accounts Receivable Specialist	Poe	60	Poe
Accounts Receivable Specialist	Spoon	61	Poe
Accounts Receivable Specialist	Walton	62	Poe

3. Approximate execution

- **HyperLogLog accuracy**

Sẽ trả về kết quả với độ chính xác trung bình là 2% so với số lượng thực tế.

Ví dụ: Nếu COUNT(DISTINCT) trả về 1.000.000, thì HyperLogLog sẽ trả về một giá trị trong khoảng từ **999.736 đến 1.016.234**.

- **APPROX_COUNT_DISTINCT**

Trả về số lượng xấp xỉ các giá trị không null duy nhất trong một nhóm.

Trường hợp sử dụng: Xấp xỉ hành vi xu hướng sử dụng web.

-- Syntax

APPROX_COUNT_DISTINCT (expression)

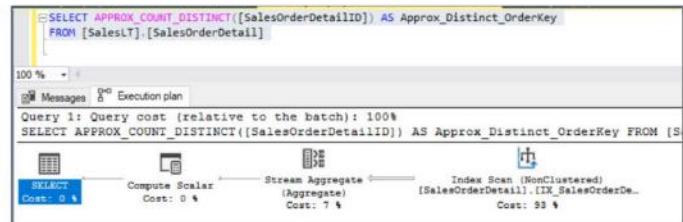
```
-- The approximate number of different order keys by order status from the orders table.
SELECT O_OrderStatus, APPROX_COUNT_DISTINCT(O_OrderKey) AS Approx_Distinct_OrderKey
FROM dbo.Orders
GROUP BY O_OrderStatus
ORDER BY O_OrderStatus;
```

APPROX_COUNT_DISTINCT

```
SELECT APPROX_COUNT_DISTINCT([SalesOrderDetailID]) AS Approx_Distinct_OrderKey
FROM [SalesLT].[SalesOrderDetail]
```

Results

Approx_Distinct_OrderKey
540

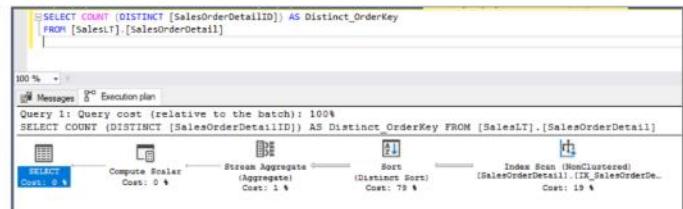


COUNT DISTINCT

```
SELECT COUNT(DISTINCT [SalesOrderDetailID]) AS Distinct_OrderKey
FROM [SalesLT].[SalesOrderDetail]
```

Results

Distinct_OrderKey
542



4. Group by options

Group by with rollup

- Tạo một nhóm cho mỗi tổ hợp của các biểu thức cột.
- Tổng hợp các kết quả thành subtotals (tổng phụ) và grand totals (tổng toàn bộ).
- Tính toán các phép tổng hợp dữ liệu phân cấp.

Grouping sets

- Kết hợp nhiều mệnh đề GROUP BY thành một mệnh đề GROUP BY duy nhất.
- Tương đương với UNION ALL của các nhóm được chỉ định.

-- GROUP BY ROLLUP Example --

```
SELECT Country,
Region,
SUM(Sales) AS TotalSales
FROM Sales
GROUP BY ROLLUP (Country, Region);
-- Results --
```

Country	Region	TotalSales
Canada	Alberta	100
Canada	British Columbia	500
Canada	NULL	600
United States	Montana	100
United States	NULL	100
NULL	NULL	700

5. Snapshot isolation

Tổng quan

- Quy định rằng các câu lệnh **không thể đọc dữ liệu đã được sửa đổi nhưng chưa được cam kết (commit)** bởi các giao dịch khác.
- Điều này ngăn chặn **dirty reads (đọc bẩn)**.

Isolation levels (Cấp độ cô lập)

```
ALTER DATABASE MyDatabase
SET ALLOW_SNAPSHOT_ISOLATION ON
```

```
ALTER DATABASE MyDatabase SET
READ_COMMITTED_SNAPSHOT ON
```

- **READ COMMITTED:** Đảm bảo chỉ đọc dữ liệu đã được cam kết.
- **REPEATABLE READ:** Đảm bảo dữ liệu được đọc sẽ không bị thay đổi bởi các giao dịch khác trong cùng một phiên đọc.
- **SNAPSHOT:** Sử dụng phiên bản dữ liệu tại thời điểm bắt đầu giao dịch, tránh xung đột dữ liệu.
- **READ UNCOMMITTED:** Cho phép **dirty reads** (đọc bẩn), nghĩa là dữ liệu chưa cam kết vẫn có thể bị đọc.
- **SERIALIZABLE:** Cấp độ cô lập cao nhất, đảm bảo các giao dịch diễn ra tuân tự mà không có xung đột.

Read _committed_snapshot

- **OFF (Mặc định):** Sử dụng **shared locks** (khóa chia sẻ) để ngăn giao dịch khác sửa đổi các hàng trong khi thực hiện đọc.
- **ON:** Sử dụng **row versioning** (phiên bản hàng) để cung cấp cho mỗi câu lệnh một snapshot (ảnh chụp nhanh) nhất quán của dữ liệu tại thời điểm bắt đầu câu lệnh. **Không sử dụng khóa** để bảo vệ dữ liệu khỏi việc bị cập nhật.

6. JSON data support

a) Insert JSON data

- Tổng quan

- Định dạng JSON cho phép biểu diễn cấu trúc dữ liệu phức tạp hoặc phân cấp trong bảng.
- Dữ liệu JSON được lưu trữ dưới dạng các cột tiêu chuẩn kiểu NVARCHAR trong bảng.

- Lợi ích

- Chuyển đổi các mảng đối tượng JSON thành định dạng bảng.
- Tối ưu hóa hiệu năng bằng cách sử dụng:
 - Clustered columnstore indexes (Chỉ mục lưu trữ dạng cột cụm)
 - Memory-optimized tables (Bảng tối ưu hóa bộ nhớ)

```
-- Create Table with column for JSON string
CREATE TABLE CustomerOrders
(
    CustomerId BIGINT NOT NULL,
    Country NVARCHAR(150) NOT NULL,
    OrderDetails NVARCHAR(3000) NOT NULL -- NVARCHAR column for JSON
) WITH (DISTRIBUTION = ROUND_ROBIN)

-- Populate table with semi-structured data
INSERT INTO CustomerOrders
VALUES
( 101, -- CustomerId
  'Bahrain', -- Country
  N'[{ "StoreId": "AW73565",
        "Order": { "Number": "SO43659",
                   "Date": "2011-05-31T00:00:00"
                 },
        "Item": { "Price": 2024.40, "Quantity": 1 }
      }]' -- OrderDetails
)
```

b) Read JSON data

- **Tổng quan:** Đọc dữ liệu JSON được lưu trong một cột dạng chuỗi bằng các phương pháp sau

- **ISJSON:** Kiểm tra xem văn bản có phải là JSON hợp lệ hay không.
- **JSON_VALUE:** Trích xuất một giá trị đơn lẻ (scalar value) từ chuỗi JSON.
- **JSON_QUERY:** Trích xuất một đối tượng JSON (JSON object) hoặc một mảng (array) từ chuỗi JSON.

```
-- Return all rows with valid JSON data
SELECT CustomerId, OrderDetails
FROM CustomerOrders
WHERE ISJSON(OrderDetails) > 0;
```

CustomerId	OrderDetails
101	N'[{ StoreId": "AW73565", "Order": { "Number": "SO43659", "Date": "2011-05-31T00:00:00" }, "Item": { "Price": 2024.40, "Quantity": 1 }}]'

```
-- Extract values from JSON string
SELECT CustomerId,
Country,
JSON_VALUE(OrderDetails,'$.StoreId') AS StoreId,
JSON_QUERY(OrderDetails,'$.Item') AS ItemDetails
FROM CustomerOrders;
```

CustomerId	Country	StoreId	ItemDetails
101	Bahrain	AW73565	{ "Price": 2024.40, "Quantity": 1 }

- **Lợi ích**

- Có khả năng lấy cả **các cột tiêu chuẩn** và cột dạng JSON.
- Thực hiện **phép tổng hợp (aggregation)** và **lọc (filter)** trên các giá trị JSON.

c) Modify and operate on JSON data

- **Tổng quan**

- Sử dụng **các cột tiêu chuẩn trong bảng** và các giá trị từ văn bản JSON trong cùng một truy vấn phân tích.
- Sửa đổi dữ liệu JSON bằng các phương pháp sau:
 - **JSON_MODIFY:** Thay đổi một giá trị trong chuỗi JSON.
 - **OPENJSON:** Chuyển đổi một tập hợp JSON thành các hàng và cột.

```
-- Modify Item Quantity value
UPDATE CustomerOrders SET OrderDetails =
JSON_MODIFY(OrderDetails,'$.OrderDetails.item.Quantity',2)
```

OrderDetails
N'[{ StoreId": "AW73565", "Order": { "Number": "SO43659", "Date": "2011-05-31T00:00:00" }, "Item": { "Price": 2024.40, "Quantity": 2 }}]'

```
-- Convert JSON collection to rows and columns
SELECT CustomerId,
StoreId,
OrderDetails.OrderDate,
OrderDetails.OrderPrice
FROM CustomerOrders
CROSS APPLY OPENJSON (CustomerOrders.OrderDetails)
WITH ( StoreId  VARCHAR(50) '$.StoreId',
OrderNumber  VARCHAR(100) '$.Order.Date',
OrderDate   DATETIME '$.Order.Date',
OrderPrice  DECIMAL '$.Item.Price',
OrderQuantity INT '$.Item.Quantity'
) AS OrderDetails
```

CustomerId	StoreId	OrderDate	OrderPrice
101	AW73565	2011-05-31T00:00:00	2024.40

- **Lợi ích**

- **Linh hoạt** trong việc cập nhật chuỗi JSON bằng T-SQL.
- Chuyển đổi **dữ liệu phân cấp** thành cấu trúc bảng **phẳng**.

d) Stored Procedures

- Tổng quan

- Là một nhóm bao gồm một hoặc nhiều câu lệnh SQL, hoặc tham chiếu đến **phương thức của Microsoft .NET Framework common runtime language (CLR)**.
- Thúc đẩy **tính linh hoạt** và **tính module hóa**.
- Hỗ trợ **tham số** và **lồng nhau**.

- Lợi ích

- Giảm lưu lượng mạng giữa máy chủ và máy khách, cải thiện hiệu năng.
- Tăng cường bảo mật.
- Dễ bảo trì.

II. Data Storage and Performance Optimizations

1. Database Tables

a. Tables – Indexes

Clustered Columnstore Index (Chỉ mục dạng cột cụm - Mặc định chính):

- Mức độ nén dữ liệu cao nhất.
- Hiệu năng truy vấn tổng thể tốt nhất.

Clustered Index (Chỉ mục cụm - Chính):

- Hiệu quả khi tra cứu một hoặc một vài hàng.

Heap (Không chỉ mục - Chính):

- Tải dữ liệu nhanh hơn và phù hợp để lưu trữ dữ liệu tạm thời.
- Tốt nhất cho các bảng tra cứu nhỏ.

```
CREATE PROCEDURE HumanResources.uspGetAllEmployees
AS
    SET NOCOUNT ON;
    SELECT LastName, FirstName, JobTitle, Department
    FROM HumanResources.vEmployeeDepartment;
GO

-- Execute a stored procedures
EXECUTE HumanResources.uspGetAllEmployees;
GO
-- Or
EXEC HumanResources.uspGetAllEmployees;
GO
-- Or, if this procedure is the first statement
-- within a batch:
HumanResources.uspGetAllEmployees;
```

```
-- Create table with index
CREATE TABLE orderTable
(
    OrderId INT NOT NULL,
    Date DATE NOT NULL,
    Name VARCHAR(2),
    Country VARCHAR(2)
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX |
    HEAP |
    CLUSTERED INDEX (OrderId)
);

-- Add non-clustered index to table
CREATE INDEX NameIndex ON orderTable (Name);
```

Nonclustered Indexes (Chỉ mục không cụm - Phụ):

- Hỗ trợ sắp xếp nhiều cột trong một bảng.
- Cho phép tạo nhiều chỉ mục không cụm trên một bảng duy nhất.

- Có thể được tạo trên bất kỳ loại chỉ mục chính nào (Clustered, Heap, hoặc Columnstore).
- Cải thiện hiệu suất cho các truy vấn tra cứu.

b. Clustered Columnstore Index

Ordered Columnstore Segments

- Tổng quan

- Các truy vấn đối với các bảng có các **phân đoạn dạng cột được sắp xếp theo thứ tự** có thể tận dụng khả năng loại bỏ **phân đoạn (segment elimination)** được cải thiện.
- Điều này giúp **giảm đáng kể thời gian** cần thiết để xử lý một truy vấn.

-- Create Table with Ordered Columnstore Index

```
CREATE TABLE sortedOrderTable
(
    OrderId INT NOT NULL,
    Date DATE NOT NULL,
    Name VARCHAR(2),
    Country VARCHAR(2)
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX ORDER (OrderId)
)

-- Create Clustered Columnstore Index on existing table
CREATE CLUSTERED COLUMNSTORE INDEX cciOrderId
ON dbo.OrderTable ORDER (OrderId)
```

c. Tables – Distributions

- **Round-robin distributed (Phân phối vòng tròn):** Phân phối các hàng của bảng đều nhau trên tất cả các phân phối một cách ngẫu nhiên.
- **Hash distributed (Phân phối băm):** Phân phối các hàng của bảng trên các Compute nodes bằng cách sử dụng một **hàm băm xác định** để gán mỗi hàng vào một phân phối cụ thể.
- **Replicated (Sao chép):** Một **bản sao đầy đủ của bảng** được truy cập trên mỗi Compute node.

d. Tables – Partitions

Tổng quan

- **Phân vùng bảng** chia dữ liệu thành các nhóm nhỏ hơn.
- Trong hầu hết các trường hợp, phân vùng được tạo trên **cột ngày tháng**.
- **Hỗ trợ trên tất cả các loại bảng**.

CREATE TABLE dbo.OrderTable

```
( 
    OrderId INT NOT NULL,
    Date DATE NOT NULL,
    Name VARCHAR(2),
    Country VARCHAR(2)
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX,
    DISTRIBUTION = HASH([OrderId]) | 
                    ROUND ROBIN | 
                    REPLICATED
);
```

- **RANGE RIGHT:** Sử dụng cho phân vùng theo **thời gian**.
- **RANGE LEFT:** Sử dụng cho phân vùng theo **số học**.

Lợi ích

- **Cải thiện hiệu quả và hiệu năng** khi tải và truy vấn dữ liệu bằng cách giới hạn phạm vi chỉ trong tập con dữ liệu.
- Mang lại **tăng cường hiệu năng truy vấn** đáng kể khi lọc trên khóa phân vùng, giúp loại bỏ các quét không cần thiết và giảm thao tác I/O.

```

CREATE TABLE partitionedOrderTable
(
    OrderId INT NOT NULL,
    Date DATE NOT NULL,
    Name VARCHAR(2),
    Country VARCHAR(2)
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX,
    DISTRIBUTION = HASH([OrderId]),
    PARTITION (
        [Date] RANGE RIGHT FOR VALUES (
            '2000-01-01', '2001-01-01', '2002-01-01',
            '2003-01-01', '2004-01-01', '2005-01-01'
        )
    )
);

```

2. Database Views

a. Materialized views

Tổng quan

- **Materialized view (Khung nhìn vật chất hóa)** thực hiện tính toán trước, lưu trữ và duy trì dữ liệu của nó như một bảng.
- Materialized view sẽ **tự động cập nhật** khi dữ liệu trong các bảng nguồn thay đổi. Đây là một **hoạt động đồng bộ**, xảy ra ngay khi dữ liệu thay đổi.
- Chức năng **bộ đệm tự động (auto caching)** cho phép Azure Synapse Analytics Query Optimizer sử dụng view có chỉ mục ngay cả khi view không được tham chiếu trong truy vấn.
- Các phép tính tổng hợp được hỗ trợ: **MAX, MIN, AVG, COUNT, COUNT_BIG, SUM, VAR, STDEV**.

```

-- Create indexed view
CREATE MATERIALIZED VIEW Sales.vw_Orders
WITH
(
    DISTRIBUTION = ROUND_ROBIN | HASH(ProductID)
)
AS
SELECT SUM(UnitPrice*OrderQty) AS Revenue,
       OrderDate,
       ProductID,
       COUNT_BIG(*) AS OrderCount
FROM Sales.SalesOrderDetail
GROUP BY OrderDate, ProductID;
GO

-- Disable index view and put it in suspended mode
ALTER INDEX ALL ON Sales.vw_Orders DISABLE;

-- Re-enable index view by rebuilding it
ALTER INDEX ALL ON Sales.vw_Orders REBUILD;

```

Lợi ích

- **Tự động và đồng bộ hóa việc làm mới dữ liệu** khi có thay đổi ở các bảng cơ sở. Không cần người dùng thực hiện thêm hành động.
- **Độ sẵn sàng cao và khả năng phục hồi tốt** giống như các bảng thông thường.

Materialized views-Recommendations

- EXPLAIN

Cung cấp kế hoạch truy vấn (query plan) cho SQL Data Warehouse mà **không cần chạy câu lệnh SQL**.

Cho phép xem chi phí ước tính của các thao tác trong truy vấn.

```
EXPLAIN WITH_RECOMMENDATIONS
select count(*)
from ((select distinct c_last_name, c_first_name, d_date
       from store_sales, date_dim, customer
      where store_sales.ss_sold_date_sk =
            date_dim.d_date_sk
        and store_sales.ss_customer_sk =
            customer.c_customer_sk
        and d_month_seq between 1194 and 1194+11)
     except
     (select distinct c_last_name, c_first_name, d_date
       from catalog_sales, date_dim, customer
      where catalog_sales.cs_sold_date_sk =
            date_dim.d_date_sk
        and catalog_sales.cs_bill_customer_sk =
            customer.c_customer_sk
        and d_month_seq between 1194 and 1194+11)
) top_customers
```

- EXPLAIN WITH_RECOMMENDATIONS

Cung cấp kế hoạch truy vấn kèm theo **đề xuất tối ưu hóa** để cải thiện hiệu suất của câu lệnh SQL.

b. COPY

- Tổng quan:

Sao chép dữ liệu từ nguồn đến đích.

- Lợi ích

- Truy xuất dữ liệu từ tất cả các tệp trong thư mục và tất cả các thư mục con của nó.
- Hỗ trợ nhiều vị trí từ cùng một tài khoản lưu trữ, được phân tách bằng dấu phẩy.
- Hỗ trợ Azure Data Lake Storage (ADLS) Gen 2 và Azure Blob Storage.
- Hỗ trợ các định dạng tệp CSV, PARQUET, ORC.

```
COPY INTO test_1
FROM
'https://XXX.blob.core.windows.net/customerdatasets/test_1.txt'
WITH (
    FILE_TYPE = 'CSV',
    CREDENTIAL=(IDENTITY= 'Shared Access Signature',
SECRET='<Your_SAS_Token>'),
    FIELDQUOTE = """",
    FIELDTERMINATOR=';',
    ROWTERMINATOR='0X0A',
    ENCODING = 'UTF8',
    DATEFORMAT = 'ymd',
    MAXERRORS = 10,
    ERRORFILE = '/errorsfolder/' --path starting from
the storage container,
    IDENTITY_INSERT
)
```

```
COPY INTO test_parquet
FROM
'https://XXX.blob.core.windows.net/customerdatasets/test.parquet'
WITH (
    FILE_FORMAT = myFileFormat
    CREDENTIAL=(IDENTITY= 'Shared Access Signature',
SECRET='<Your_SAS_Token>')
)
```

c. Result-set caching

- Tổng quan

- Lưu vào bộ nhớ đệm kết quả của truy vấn trong bộ lưu trữ của **Data Warehouse (DW)**. Điều này cho phép thời gian phản hồi nhanh khi thực hiện lại các truy vấn lặp đi lặp lại trên các bảng có dữ liệu thay đổi không thường xuyên.
- Bộ đếm kết quả vẫn được giữ nguyên ngay cả khi Data Warehouse tạm dừng và được khôi phục sau đó.

```
-- Turn on/off result-set caching for a database
-- Must be run on the MASTER database
ALTER DATABASE {database_name}
SET RESULT_SET_CACHING { ON | OFF }

-- Turn on/off result-set caching for a client session
-- Run on target data warehouse
SET RESULT_SET_CACHING {ON | OFF}

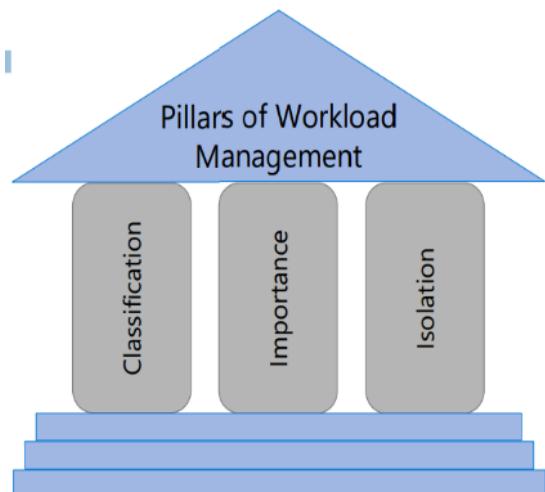
-- Check result-set caching setting for a database
-- Run on target data warehouse
SELECT is_result_set_caching_on
FROM sys.databases
WHERE name = {database_name}

-- Return all query requests with cache hits
-- Run on target data warehouse
SELECT *
FROM sys.dm_pdw_request_steps
WHERE command like '%DWResultCacheDb%'
AND step_index = 0
```

- Bộ nhớ đệm truy vấn sẽ bị **vô hiệu hóa và làm mới** khi dữ liệu của bảng hoặc mã truy vấn thay đổi.
- Bộ nhớ đệm kết quả sẽ bị loại bỏ định kỳ dựa trên thuật toán **TLRU** (**Time-aware Least Recently Used**).
- **Lợi ích**
 - **Cải thiện hiệu suất** khi cùng một kết quả được yêu cầu lặp lại nhiều lần.
 - **Giảm tải cho máy chủ** đối với các truy vấn được thực thi lặp lại.
 - **Cung cấp khả năng giám sát** việc thực thi truy vấn với trạng thái **cache hit** (truy vấn trùng bộ nhớ đệm) hoặc **cache miss** (không trùng bộ nhớ đệm).

III. Performance Optimizations Workload Management

- **Tổng quan:** Cơ chế này quản lý tài nguyên, đảm bảo việc sử dụng tài nguyên hiệu quả cao và tối đa hóa **lợi tức đầu tư (ROI)**.
- Ba trụ cột chính của **workload management**:
 - **Workload Classification:** Gán một yêu cầu vào một nhóm khối lượng công việc và thiết lập mức độ ưu tiên.
 - **Workload Importance:** Ảnh hưởng đến thứ tự mà một yêu cầu được cấp quyền truy cập vào tài nguyên.
 - **Workload Isolation:** Dành riêng tài nguyên cho một nhóm khối lượng công việc cụ thể.



1. Resource classes

- **Tổng quan:** Giới hạn tài nguyên **được xác định trước** cho một người dùng hoặc vai trò cụ thể.
 - **Lợi ích**
 - **Quản lý bộ nhớ hệ thống** được gán cho mỗi truy vấn.
 - Được sử dụng hiệu quả để **kiểm soát số lượng truy vấn đồng thời** có thể chạy trên kho dữ liệu.
-
- ```

/* View resource classes in the data warehouse */
SELECT name
FROM sys.database_principals
WHERE name LIKE '%rc%' AND type_desc = 'DATABASE_ROLE';

/* Change user's resource class to 'largerc' */
EXEC sp_addrolemember 'largerc', 'loaduser';

/* Decrease the loading user's resource class */
EXEC sp_droprolemember 'largerc', 'loaduser';

```

#### 2. Workload classification

- **Tổng quan**

- Ánh xạ các truy vấn với phân bô tài nguyên thông qua các **quy tắc được xác định trước**.
- Sử dụng cùng với **tầm quan trọng của khối lượng công việc (workload importance)** để chia sẻ tài nguyên hiệu quả giữa các loại khối lượng công việc khác nhau.
- Nếu một yêu cầu truy vấn không khớp với bộ phân loại (classifier), nó sẽ được gán vào nhóm khối lượng công việc mặc định (**smallrc resource class**).

#### - Lợi ích

- Ánh xạ các truy vấn với cả **Quản lý Tài nguyên (Resource Management)** và **Khái niệm Cô lập Khối lượng Công việc (Workload Isolation)**.
- Quản lý **nhóm người dùng** chỉ với một số bộ phân loại (classifiers) đơn giản.

### 3. Workload importance

- Các truy vấn vượt quá giới hạn đồng thời sẽ được đưa vào hàng đợi FiFo (First-In, First-Out).
- Theo mặc định, các truy vấn sẽ được giải phóng khỏi hàng đợi theo thứ tự đến trước, xử lý trước, khi tài nguyên khả dụng.
- Tầm quan trọng của khối lượng công việc (Workload Importance) cho phép các truy vấn có mức ưu tiên cao hơn được cấp tài nguyên ngay lập tức, bất kể vị trí trong hàng đợi.

### 4. Workload Isolation

- Phân bô tài nguyên cố định cho một nhóm khối lượng công việc (**workload group**).
- Gán mức sử dụng tối đa và tối thiểu cho các tài nguyên khác nhau dưới tải.
- Các điều chỉnh này có thể được thực hiện trực tiếp mà không cần đưa **SQL Analytics** về chế độ ngoại tuyến.

#### Lợi ích

- Dành riêng tài nguyên** cho một nhóm yêu cầu cụ thể.
- Giới hạn lượng tài nguyên** mà một nhóm yêu cầu có thể tiêu thụ.
- Tài nguyên được chia sẻ** sẽ được truy cập dựa trên mức độ quan trọng.
- Đặt giá trị timeout cho truy vấn:** Loại bỏ nhu cầu quản trị viên cơ sở dữ liệu (DBA) phải can thiệp để dừng các truy vấn chạy không kiểm soát.

```

CREATE WORKLOAD CLASSIFIER classifier_name
WITH
(
 [WORKLOAD_GROUP = '<Resource Class>'],
 [IMPORTANCE = { LOW
 |
 BELOW_NORMAL
 NORMAL
 ABOVE_NORMAL
 HIGH
 }
],
 [MEMBERNAME = 'security_account']
)
WORKLOAD_GROUP: maps to an existing resource class
IMPORTANCE: specifies relative importance of request
MEMBERNAME: database user, role, AAD Login or AAD group

```

## 5. Dynamic Management Views (DMVs)

- **Tổng quan:** **Dynamic Management Views (DMV)** là các truy vấn trả về thông tin về các đối tượng mô hình, hoạt động của máy chủ, và tình trạng sức khỏe của máy chủ.
- **Lợi ích**
  - **Cú pháp SQL đơn giản** – Dễ dàng viết và thực thi.
  - **Kết quả trả về dưới dạng bảng** – Giúp hiển thị trực quan và dễ hiểu.
  - **Dễ đọc và sao chép kết quả** – Thuận tiện trong việc phân tích và chia sẻ dữ liệu.

## 6. SQL Monitor with DMVs

**Tổng quan:** Cung cấp khả năng giám sát:

- Tất cả các phiên (sessions) đang mở và đã đóng.
- Số lượng phiên theo từng người dùng.
- Số lượng truy vấn đã hoàn thành theo từng người dùng.
- Tất cả các truy vấn đang hoạt động và đã hoàn thành.
- Các truy vấn chạy lâu nhất.
- Mức tiêu thụ bộ nhớ.

Count sessions by user

```
--count sessions by user
SELECT login_name, COUNT(*) as session_count FROM
sys.dm_pdw_exec_sessions where status = 'Closed' and session_id <> session_id() GROUP BY login_name;
```

List all open sessions

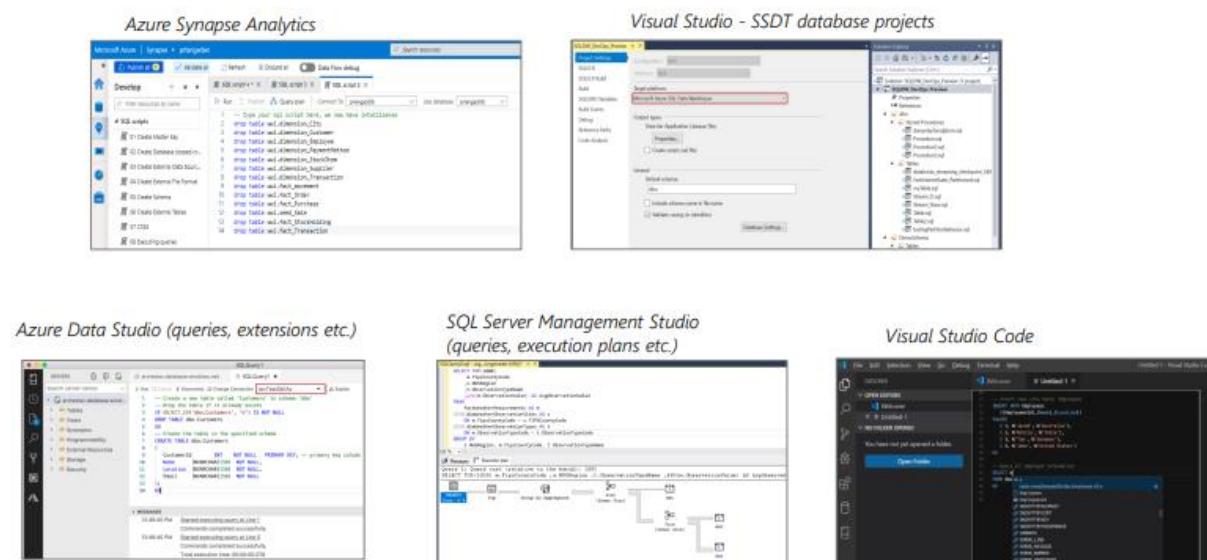
```
-- List all open sessions
SELECT * FROM sys.dm_pdw_exec_sessions where status <> 'Closed'
and session_id <> session_id();
```

List all active queries

```
-- List all active queries
SELECT * FROM sys.dm_pdw_exec_requests WHERE status not in
('Completed','Failed','Cancelled') AND session_id <> session_id()
ORDER BY submit_time DESC;
```

## IV. Developer productivity

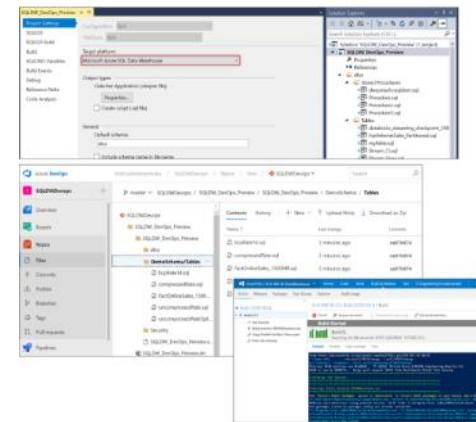
### 1. Developer Tools





## 2. Continuous integration and delivery (CI/CD)

- **Tổng quan:** Hỗ trợ dự án cơ sở dữ liệu trong **SQL Server Data Tools (SSDT)** cho phép các nhóm phát triển hợp tác trên một kho dữ liệu có kiểm soát phiên bản, đồng thời theo dõi, triển khai và kiểm tra các thay đổi về lược đồ (**schema changes**).
- **Lợi ích:** Hỗ trợ dự án cơ sở dữ liệu bao gồm tích hợp cấp cao với **Azure DevOps**, cung cấp các tính năng sau:



- **Azure Pipelines:**

- Chạy các quy trình CI/CD (Tích hợp liên tục và Triển khai liên tục) trên mọi nền tảng (Linux, macOS, và Windows).

- **Azure Repos:**

- Lưu trữ các tệp dự án trong **kiểm soát nguồn (source control)**.

- **Azure Test Plans:**

- Chạy các bài kiểm tra tự động để xác minh các bản cập nhật và sửa đổi lược đồ.

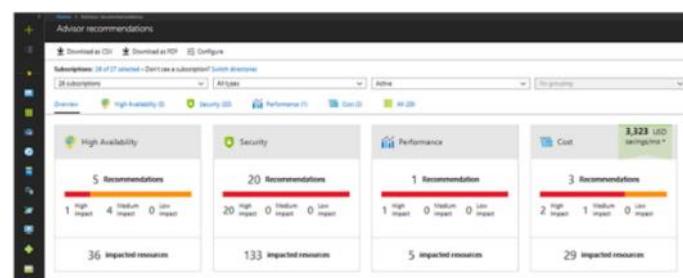
- **Hệ sinh thái bên thứ ba đang phát triển:**

- Hỗ trợ tích hợp với các công cụ khác như **Timetracker, Microsoft Teams, Slack, Jenkins**, v.v., để bổ sung cho các quy trình làm việc hiện tại.

## V. Maintenance

### 1. Azure Advisor recommendations

- **Suboptimal Table Distribution:** Giảm chuyển động dữ liệu, thực hiện sao chép bảng (replicate tables).
- **Data Skew:** Lựa chọn khóa phân phối băm (hash-distribution key) mới.

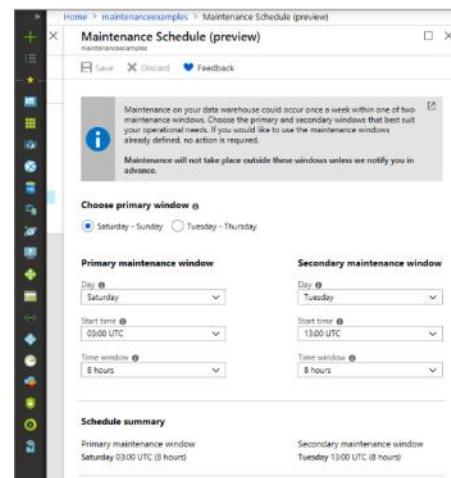


- **Slowest distribution limits performance:** Tìm cách tối ưu hóa các phân phối chậm để giảm ảnh hưởng tới toàn bộ hệ thống.
- **Cache Misses:** Cung cấp thêm tài nguyên, thêm dung lượng xử lý.
- **Tempdb Contention:** Tăng quy mô hoặc cập nhật llop tài nguyên người dùng (user resource class).
- **Suboptimal Plan Selection:** Tạo mới hoặc cập nhật thông kê bảng (table statistics).

## 2. Maintenance windows

### Tổng quan

- Chọn một **khoảng thời gian** để thực hiện nâng cấp.
- Chọn **cửa sổ thời gian chính và phụ** trong vòng bảy ngày.
- **Khoảng thời gian** có thể từ **3 đến 8 giờ**.
- Nhận **Thông báo trước 24 giờ** cho các sự kiện bảo trì.



### Lợi ích

- **Đảm bảo nâng cấp** được thực hiện theo lịch trình của bạn.
- **Lập kế hoạch dễ dàng** cho các tác vụ chạy dài hạn.
- **Luôn được thông báo** về thời điểm bắt đầu và kết thúc bảo trì.

## 3. Automatic statistics management

### Tổng quan

- **Thống kê (Statistics)** được tự động tạo và duy trì cho SQL pool.
- Các truy vấn đến được phân tích, và các thống kê cột riêng lẻ được tạo trên các cột nhằm cải thiện ước tính số lượng (cardinality estimates) để tăng hiệu suất truy vấn.
- Thống kê được tự động cập nhật khi dữ liệu trong các bảng cơ sở bị thay đổi. Mặc định, các cập nhật này là **đồng bộ (synchronous)** nhưng có thể cấu hình thành **không đồng bộ (asynchronous)**.
- Thống kê được coi là lỗi thời khi:
  - Có sự thay đổi dữ liệu trên bảng rõ ràng.

-- Turn on/off auto-create statistics settings

```
ALTER DATABASE {database_name}
SET AUTO_CREATE_STATISTICS { ON | OFF }
```

-- Turn on/off auto-update statistics settings

```
ALTER DATABASE {database_name}
SET AUTO_UPDATE_STATISTICS { ON | OFF }
```

-- Configure synchronous/asynchronous update

```
ALTER DATABASE {database_name}
SET AUTO_UPDATE_STATISTICS_ASYNC { ON | OFF }
```

-- Check statistics settings for a database

```
SELECT is_auto_create_stats_on,
 is_auto_update_stats_on,
 is_auto_update_stats_async_on
FROM sys.databases
```

- Số lượng hàng trong bảng tại thời điểm tạo thống kê là **500 hoặc ít hơn**, và **hơn 500 hàng** đã được cập nhật.
- Số lượng hàng trong bảng tại thời điểm tạo thống kê là **hơn 500**, và **hơn 500 + 20% số hàng** đã được cập nhật.

## VII. Snapshots and restores

### Tổng quan

- Bản sao tự động của trạng thái kho dữ liệu.
- Được tạo trong suốt ngày hoặc kích hoạt thủ công.
- Có sẵn trong tối đa 7 ngày, ngay cả sau khi kho dữ liệu bị xóa. Thời gian RPO là 8 giờ đối với việc khôi phục từ ảnh chụp nhanh.
- Khôi phục theo khu vực dưới 20 phút, bất kể kích thước dữ liệu.
- Ảnh chụp nhanh và sao lưu địa lý cho phép khôi phục giữa các khu vực.
- Ảnh chụp nhanh tự động và sao lưu địa lý được bật theo mặc định.

### Lợi ích

- Ảnh chụp nhanh bảo vệ khỏi việc hỏng hoặc xóa dữ liệu.
- Khôi phục nhanh chóng để tạo các bản sao phát triển/kiểm thử dữ liệu.
- Ảnh chụp nhanh thủ công bảo vệ các sửa đổi lớn.
- Sao lưu địa lý sao chép một trong các ảnh chụp nhanh tự động mỗi ngày vào bộ lưu trữ RA-GRS. Điều này có thể được sử dụng trong trường hợp xảy ra thảm họa để khôi phục kho dữ liệu SQL của bạn sang một khu vực mới. Thời gian RPO là 24 giờ cho việc khôi phục địa lý.

## PHẦN 6: SQL ON DEMAND

### I. Overview

SQL On-Demand (hay còn gọi là Azure Synapse Serverless SQL Pools) là một dịch vụ truy vấn tương tác, cho phép người dùng thực thi các truy vấn T-SQL trực tiếp trên dữ liệu lớn được lưu trữ trong Azure Storage. Dịch vụ này giúp khai thác dữ liệu hiệu quả mà không cần quản lý cơ sở hạ tầng phức tạp.

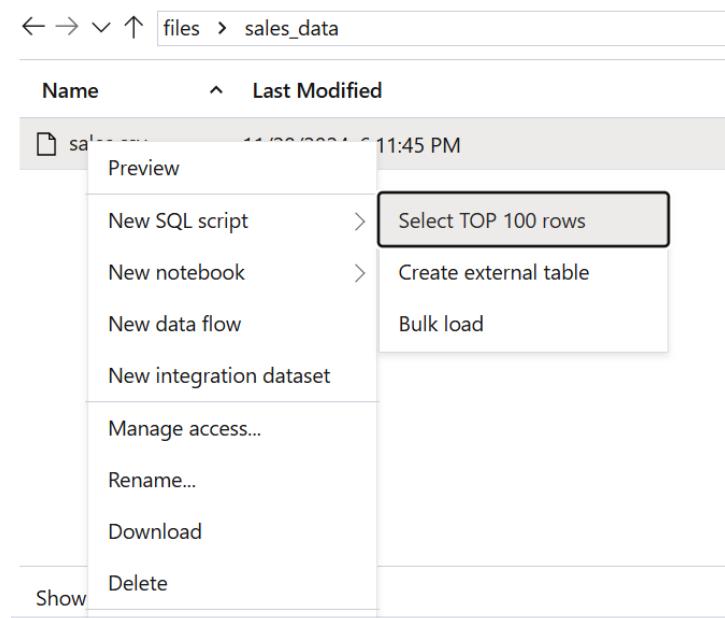
Lợi ích:

- Kiến trúc không máy chủ (Serverless: không yêu cầu thiết lập hoặc quản lý cơ sở hạ tầng, tự động mở rộng và hoạt động linh hoạt dựa trên khối lượng công việc).
- Thanh toán theo truy vấn :chỉ trả phí cho số lượng truy vấn thực thi, giúp giảm thiểu chi phí so với các hệ thống kho dữ liệu truyền thống.

- Không cần ETL : có thể truy vấn trực tiếp dữ liệu ở dạng thô (như Parquet, CSV, JSON) mà không cần phải thực hiện các bước Extract, Transform, Load.
- Bảo mật: cung cấp các tính năng bảo mật mạnh mẽ để đảm bảo dữ liệu an toàn, hỗ trợ kiểm soát truy cập chi tiết theo dòng, cột.
- Tích hợp dữ liệu : kết nối liền mạch với các công cụ phân tích dữ liệu lớn như Databricks và HDInsight.
- Cú pháp T-SQL quen thuộc: sử dụng các câu lệnh T-SQL phổ biến để truy vấn và phân tích dữ liệu, phù hợp với những người đã quen làm việc với SQL Server.
- Hỗ trợ nhiều định dạng dữ liệu: làm việc với dữ liệu có cấu trúc và bán cấu trúc trong các định dạng Parquet, CSV, và JSON.
- Hỗ trợ hệ sinh thái BI :tích hợp dễ dàng với các công cụ Business Intelligence như Power BI, Tableau, và Excel, hỗ trợ trực quan hóa dữ liệu và lập báo cáo.

## II. Querying on storage

SQL On-Demand (Serverless SQL Pools) hỗ trợ **truy vấn trực tiếp** trên các tệp lưu trữ trong Azure Storage hoặc Azure Data Lake Storage bằng cách sử dụng các câu lệnh T-SQL quen thuộc. Điều này giúp khai thác dữ liệu mà không cần tải dữ liệu vào kho hoặc thực hiện các bước ETL phức tạp.



```

1 -- This is auto-generated code
2 SELECT
3 TOP 100 *
4 FROM
5 OPENROWSET(
6 BULK 'https://datalake04seg9a.dfs.core.windows.net/files/sales_data/sales.csv',
7 FORMAT = 'CSV',
8 PARSER_VERSION = '2.0'
9) AS [result]
10

```

Và chọn run

| C1        | C2                      | C3             | C4        |
|-----------|-------------------------|----------------|-----------|
| ProductID | ProductName             | Category       | ListPrice |
| 771       | Mountain-100 Silver, 38 | Mountain Bikes | 3399.9900 |
| 772       | Mountain-100 Silver, 42 | Mountain Bikes | 3399.9900 |
| 773       | Mountain-100 Silver, 44 | Mountain Bikes | 3399.9900 |
| 774       | Mountain-100 Silver, 48 | Mountain Bikes | 3399.9900 |
| 775       | Mountain-100 Black, 38  | Mountain Bikes | 3374.9900 |

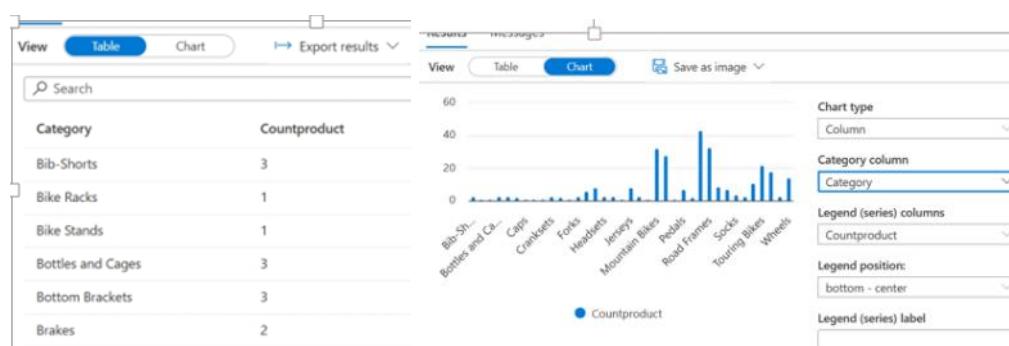
### III. Querying CSV File

SQL On-Demand (Serverless SQL Pools in Azure Synapse Analytics) sử dụng chức năng **OPENROWSET** để truy cập dữ liệu trực tiếp từ các tệp trong Azure Storage hoặc Azure Data Lake mà không cần tải dữ liệu vào trước.

The screenshot shows the Azure Synapse Analytics Data blade. On the left, there's a navigation menu with options like Home, Data, Develop, Integrate, Monitor, and Manage. The 'Data' option is selected. In the main area, there's a 'Synapse live' button, a 'Validate all' button, and a 'Publish all' button with a progress bar. Below these are sections for 'Data' (Workspace and Linked), 'Data湖' (files, New SQL script, New notebook, New data flow, More), and 'Integration datasets'. A search bar at the top right says 'Search'. The 'files' section shows a list of files under 'sales\_data': 'sales.csv' (Last Modified: 11/28/2024, 6:11:45 PM, Content Type: CSV, Size: 16.2 KB). At the bottom, it says 'Showing 1 to 1 of 1 cached items'.

Viết câu truy vấn đếm số lượng theo loại sản phẩm và run

```
-- This is auto-generated code
SELECT
 Category, COUNT(*) as Countproduct
FROM
 OPENROWSET(
 BULK 'https://datalake04seg9a.dfs.core.windows.net/files/sales',
 FORMAT = 'CSV',
 HEADER_ROW = TRUE,
 PARSER_VERSION = '2.0'
) AS [result]
GROUP BY Category
```



#### IV. Querying folders

**SQL On Demand** (thường dùng trong Azure Synapse) cho phép bạn truy vấn dữ liệu trực tiếp từ các tệp hoặc thư mục trong các hệ thống lưu trữ dữ liệu như Azure Data Lake hoặc Blob Storage mà không cần di chuyển dữ liệu vào cơ sở dữ liệu. Hàm **OPENROWSET** là một công cụ quan trọng được sử dụng để truy cập dữ liệu từ nhiều tệp hoặc thư mục.

-Truy vấn tất cả các tệp CSV trong một thư mục

```
SELECT *
FROM OPENROWSET(
 BULK 'https://tranghuyend1sg2.blob.core.windows.net/demo/synapse/*.csv',
 FORMAT = 'CSV'

)
```

-Truy vấn các tệp trong nhiều thư mục con

```
SELECT *
FROM OPENROWSET(
 BULK 'https://tranghuyend1sg2.blob.core.windows.net/demo/*/*Sales-*.csv',
 FORMAT = 'CSV'

)
```

Truy vấn một tệp cụ thể

```

SELECT *
FROM OPENROWSET(
 BULK 'https://tranhuyendlsg2.blob.core.windows.net/demo/Sales.csv',
 FORMAT = 'CSV')

```

## V. Querying specific files

Khi truy vấn dữ liệu trong SQL On Demand (Azure Synapse), có thể sử dụng các thuộc tính như **filename** và **filepath** để xác định nguồn gốc của từng hàng dữ liệu trong tập kết quả. Điều này đặc biệt hữu ích khi làm việc với nhiều tệp hoặc cấu trúc thư mục phức tạp.

**Filename:** Cung cấp tên tệp chứa dữ liệu mà mỗi hàng được trích xuất.

```

SELECT
 filename() AS FileName,
 COUNT(*) AS Rows
FROM OPENROWSET(
 BULK 'https://tranhuyendlsg2.blob.core.windows.net/demo/*/*.csv',
 FORMAT = 'CSV',
 PARSER_VERSION = '2.0'
)
GROUP BY filename();

```

|               |        |
|---------------|--------|
| Sales-Jan.csv | 123456 |
| Sales-Feb.csv | 110000 |
| Sales-Apr.csv | 98765  |
| Sales-May.csv | 104321 |

**Filepath:** Cung cấp đường dẫn đầy đủ hoặc một phần đường dẫn chứa dữ liệu.

```

SELECT
 filepath() ASFullPath,
 COUNT(*) AS RowCount
FROM OPENROWSET(
 BULK 'https://tranhuyendlsg2.blob.core.windows.net/demo/*/*.csv',
 FORMAT = 'CSV',
 PARSER_VERSION = '2.0'
)
AS [result]
GROUP BY filepath();

```

|                                                                    |        |
|--------------------------------------------------------------------|--------|
| https://tranhuyendlsg2.blob.core.windows.net/demo/Q1/Sales-Jan.csv | 100000 |
| https://tranhuyendlsg2.blob.core.windows.net/demo/Q1/Sales-Feb.csv | 95000  |
| https://tranhuyendlsg2.blob.core.windows.net/demo/Q2/Sales-Apr.csv | 123456 |
| https://tranhuyendlsg2.blob.core.windows.net/demo/Q2/Sales-May.csv | 98765  |

## VI. Querying Parquet files

SQL On-Demand (Serverless SQL Pools) sử dụng hàm OPENROWSET để truy cập dữ liệu trực tiếp từ Azure Storage hoặc Azure Data Lake. Điều này cho phép truy vấn dữ liệu dễ dàng mà không cần phải tải vào hệ thống quản lý cơ sở dữ liệu hoặc thực hiện các bước ETL phức tạp.

## Lợi ích:

- Bạn có thể chọn chỉ những cột cần thiết từ tệp dữ liệu, thay vì phải tải toàn bộ dữ liệu không cần thiết.
- OPENROWSET có khả năng tự động phát hiện và sử dụng tên cột cùng kiểu dữ liệu từ tệp nguồn, giúp đơn giản hóa cấu hình truy vấn.
- Khi làm việc với dữ liệu được lưu trữ theo cấu trúc phân vùng (partitioned data), bạn có thể lọc các phân vùng bằng cách sử dụng hàm filepath().

The image consists of two screenshots of the Microsoft Azure Synapse Analytics Data workspace interface.

The top screenshot shows the Data workspace navigation pane on the left with "Linked" selected. Under "Linked", there is a folder named "tranhuyen-synapse (Primary - tra...)" containing a subfolder "demo (Primary)". A context menu is open over a file named "NYCTaxiGreen.parquet" in the "demo" folder. The menu options include: New SQL script, New notebook, New data flow, New integration dataset, Select TOP 100 rows (which is highlighted with a red box), Create external table, Bulk load, Manage access..., Rename..., Download, and Delete. The file details show it was last modified on 12/7/2024, 7:25:29 AM and is 1.2 MB in size.

The bottom screenshot shows the SQL pool editor. The top bar includes "Synapse live", "Validate all", "Publish all", "Data", "Workspace", "Linked", "SQL script 1", "Notebook 1", and "Properties...". The main area displays the following T-SQL code:

```
1 e-- This is auto-generated code
2
3 SELECT
4 TOP 100
5 *
6 FROM
7 OPENROWSET(
8 BULK 'https://tranhuyenadls2.dfs.core.windows.net/demo/synapse'
9 FORMAT = 'PARQUET'
10) AS [result]
```

To the right of the code, the "General" properties panel is visible, showing:

- Name: SQLpooltxgdedicated
- Description: (empty)
- Type: sql script
- Size: 215 bytes
- Results settings per query:
  - First 5000 rows (default) (radio button selected)
  - All rows

At the bottom of the editor, the status bar says "00:00:09 Query executed successfully."

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a navigation pane with 'Data' selected, showing 'Workspace' and 'Linked' sections. Under 'Linked', it lists 'Azure Data Lake Storage Gen2' and 'tranhuyen-synapse (Primary - tra...)' which contains a 'demo' container. The main area displays a table titled 'SQL script 1' with the following data:

| VendorID | lpep_pickup_d... | lpep_dropoff...  | store_and_fwd... | RatecodeID |
|----------|------------------|------------------|------------------|------------|
| 2        | 2024-02-01T00... | 2024-02-01T00... | N                | 1          |
| 2        | 2024-01-31T22... | 2024-01-31T23... | N                | 1          |
| 2        | 2024-02-01T00... | 2024-02-01T00... | N                | 1          |
| 2        | 2024-01-31T23... | 2024-02-01T00... | N                | 1          |
| 2        | 2024-02-01T00... | 2024-02-01T00... | N                | 5          |

Below the table, a message says '00:00:09 Query executed successfully.' To the right, there's a 'Properties' panel with tabs for 'General' and 'Related (0)'. The 'General' tab shows the name is 'SQL script 1', type is 'sql script', and size is '215 bytes'.

## VII. Creating views

SQL On Demand cho phép tạo views sử dụng các truy vấn SQL trực tiếp. Views trong SQL On Demand hoạt động giống như các views trong SQL Server hoặc các hệ thống quản lý cơ sở dữ liệu khác. Views là các đối tượng cơ sở dữ liệu ảo, được sử dụng để lưu trữ một truy vấn SQL và có thể được truy vấn giống như một bảng thông thường.

Lợi ích:

- Tái sử dụng truy vấn: Views cho phép tái sử dụng các truy vấn SQL phức tạp mà không phải viết lại nhiều lần.

```

1 CREATE DATABASE cuoiky_db
2 Go
3 USE cuoiky_db
4 GO
5 CREATE VIEW SalesView AS
6 SELECT *
7 FROM OPENROWSET(
8 BULK 'https://tranhuyen.dfs.core.windows.net/files/sales_data/sales.csv',
9 FORMAT = 'CSV',
10 FIELDTERMINATOR = ',',
11 ROWTERMINATOR = '\n'
12)
13 WITH (
14 [SalesOrderNumber] VARCHAR(20),
15 [SalesOrderLineNumber] INT,
16 [OrderDate] DATETIME,
17 [CustomerName] VARCHAR(100),
18 [EmailAddress] VARCHAR(100),
19 [Item] VARCHAR(100),
20 [Quantity] INT,
21 [UnitPrice] DECIMAL(18,2),
22 [TaxAmount] DECIMAL(18,2)
23) AS [Sales];

```

- Truy vấn dễ dàng: Views cung cấp cách thức đơn giản để thực hiện truy vấn dữ liệu từ nhiều bảng hoặc tệp.

```

Use cuoiky_db
GO
SELECT * FROM SalesView

```

- Quản lý dễ dàng: quản lý các truy vấn phức tạp một cách dễ dàng hơn khi lưu trữ chúng dưới dạng views.

## VIII. SQL On Demand – Querying JSON files

- Sử dụng **OPENROWSET** để đọc file JSON từ Azure Blob Storage.
- Lợi ích
  - Đọc JSON trực tiếp: Không cần ETL hoặc tải dữ liệu trước.
  - OPENJSON: Dễ dàng trích xuất dữ liệu từ mảng hoặc đối tượng JSON.
  - JSON\_VALUE và JSON\_QUERY: Phân tích và xử lý dữ liệu JSON linh hoạt.

## IX. Create External Table As Select

Tạo một bảng ngoài (external table) trong cơ sở dữ liệu và sau đó xuất kết quả của câu lệnh `SELECT` vào bảng ngoài đó. Bảng ngoài chỉ tồn tại trong suốt thời gian thực thi câu lệnh truy vấn, giúp xuất dữ liệu một cách linh hoạt.

### Các Bước Tạo Bảng Ngoài và Xuất Dữ Liệu

B1: Tạo Master Key: Đầu tiên, bạn cần tạo master key để đảm bảo việc mã hóa dữ liệu, đặc biệt nếu bạn sử dụng các thông tin xác thực hay kết nối được mã hóa.

```
CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'HuyenTrang123@';
```

B2: ạo Thông Tin Xác Thực (Credentials): Tiếp theo, bạn cần định nghĩa thông tin xác thực để truy cập vào nguồn dữ liệu ngoài, ví dụ như tài khoản lưu trữ Azure hoặc các hệ thống dữ liệu ngoài khác.

```
CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential
WITH IDENTITY = 'huyentrang', SECRET = '1234';
```

B3: Tạo Nguồn Dữ Liệu Ngoài (External Data Source): Định nghĩa nguồn dữ liệu ngoài nơi dữ liệu sẽ được xuất. Đây có thể là Azure Blob Storage hoặc một cụm Hadoop. Bước này giúp liên kết cơ sở dữ liệu của bạn với hệ thống lưu trữ ngoài.

```
CREATE EXTERNAL DATA SOURCE MyAzureStorage
WITH (TYPE = BLOB_STORAGE,
 LOCATION = ' https://tranghuyen.dfs.core.windows.net ',
 CREDENTIAL = AzureStorageCredential);
```

B4: Tạo Định Dạng Dữ Liệu Ngoài (External Data Format): Xác định định dạng dữ liệu sẽ được xuất, ví dụ như nếu bạn xuất dữ liệu dưới dạng file CSV, bạn cần chỉ định định dạng này.

```
CREATE EXTERNAL FILE FORMAT yAzureCSVFormat
WITH (FORMAT_TYPE = DELIMITEDTEXT,
 FORMAT_OPTIONS (FIELD_TERMINATOR = ',', STRING_DELIMITER = ''));
```

B5: Tạo Bảng Ngoài: Cuối cùng, bạn tạo bảng ngoài để lưu trữ kết quả của câu lệnh `SELECT`. Cần chỉ định nguồn dữ liệu ngoài và định dạng file.

```
CREATE EXTERNAL TABLE dbo.FactInternetSalesNew
WITH(
 LOCATION = '/files/sales_data/sales',
 DATA_SOURCE = MyAzureStorage,
 FILE_FORMAT = yAzureCSVFormat
)
```

B6: Sử Dụng SELECT Để Xuất Dữ Liệu: sau khi tạo bảng ngoài, bạn có thể sử dụng câu lệnh SELECT để xuất dữ liệu từ cơ sở dữ liệu vào

```
SELECT *
FROM sales
INTO EXTERNAL TABLE dbo.FactInternetSalesNew;
```

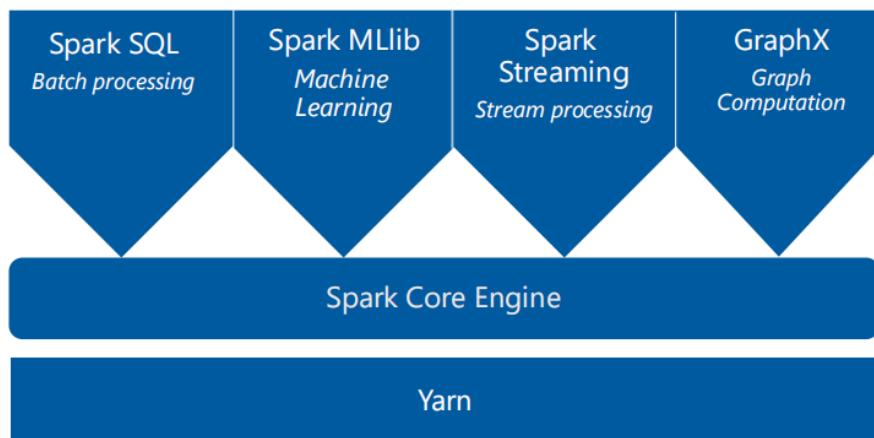
## PHẦN 7: AZURE SYNAPSE ANALYTICS SPARK

### I. Azure Synapse Apache Spark – Summary

Azure Synapse Apache Spark là một dịch vụ mạnh mẽ trong hệ sinh thái Azure Synapse, hỗ trợ xử lý dữ liệu lớn và phân tích dữ liệu hiệu quả. Dưới đây là các tính năng chính:

- Hỗ trợ nhiều ngôn ngữ: .NET Core 3.0, Python (PySpark), Scala, Java, Spark SQL, và R (dự kiến đầu năm 2020).
- Tích hợp với Delta Lake 0.4: Quản lý dữ liệu với tính năng ACID và lưu trữ thời gian (time travel).
- Tích hợp với các dịch vụ Azure Synapse: Liên kết chặt chẽ với các dịch vụ khác như SQL Pools và Azure Data Lake.
- Bảo mật tích hợp: Đảm bảo bảo mật và xác thực người dùng.
- Tự động mở rộng và tạm dừng: Tối ưu hóa tài nguyên với tính năng tự động mở rộng và tạm dừng cluster khi không sử dụng.
- Giao diện người dùng tích hợp: Hỗ trợ notebook dựa trên Jupyter cho việc phát triển và phân tích dữ liệu.
- Tối ưu hóa tài nguyên: Sử dụng tài nguyên hiệu quả, giúp xử lý nhanh và tiết kiệm chi phí.

### II. Apache Spark



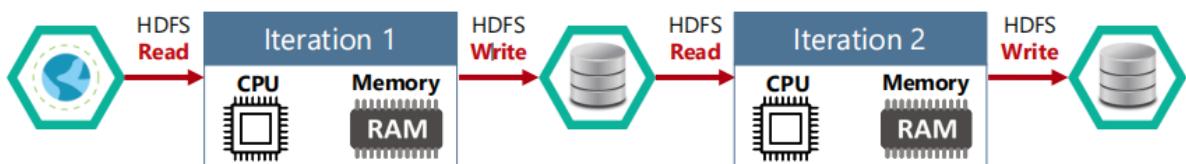
Một framework mã nguồn mở, thống nhất, xử lý dữ liệu song song dành cho phân tích Dữ liệu Lớn chính là Apache Spark.

Đặc điểm chính của Apache Spark:

- Batch Processing: Xử lý dữ liệu theo lô.
- Interactive SQL: Hỗ trợ truy vấn SQL tương tác qua Spark SQL.
- Real-time Processing: Xử lý dữ liệu thời gian thực với Spark Streaming.
- Machine Learning: Thực hiện học máy thông qua thư viện MLlib.
- Deep Learning: Hỗ trợ các tác vụ học sâu.
- Graph Processing: Phân tích dữ liệu dạng đồ thị qua GraphX.

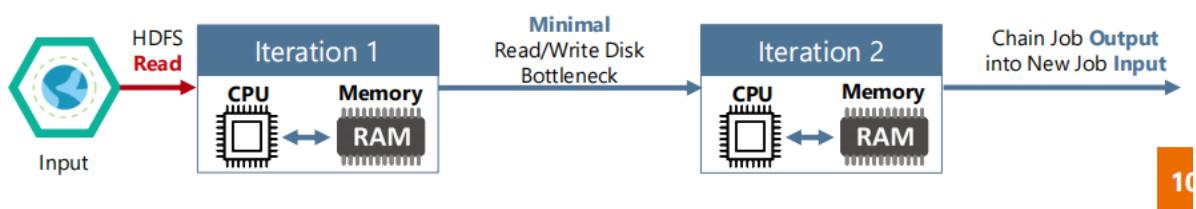
### III. Motivation for Apache Spark

- Cách tiếp cận truyền thống



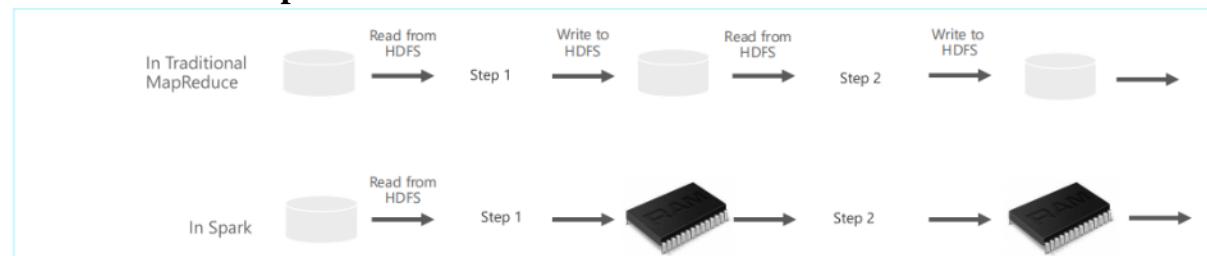
- MapReduce: Đây là mô hình phổ biến trong Hadoop để xử lý dữ liệu lớn.
- Vấn đề gặp phải:

- Với các tác vụ phức tạp như: Xử lý nhiều công việc (jobs), truy vấn tương tác (interactive query), xử lý sự kiện thời gian thực (online event-hub processing).
- MapReduce phải thực hiện rất nhiều thao tác đọc/ghi I/O trên đĩa (disk I/O), dẫn đến tốc độ chậm và hiệu suất thấp.
- Giải pháp: Apache Spark



- Lưu dữ liệu trong bộ nhớ (in-memory): Spark hạn chế thao tác với ổ đĩa bằng cách giữ dữ liệu trong bộ nhớ RAM trong suốt quá trình xử lý.
- Công cụ thực thi phân tán mới: Spark sử dụng engine phân tán giúp tăng tốc độ xử lý dữ liệu lớn gấp nhiều lần so với MapReduce.

### IV. What makes Spark fast



- Tính toán cụm trong bộ nhớ: Spark cung cấp các tính năng tính toán trong cụm **bộ nhớ**. Một công việc Spark có thể **nạp và lưu trữ tạm** dữ liệu trong bộ nhớ và truy vấn lặp lại nhanh hơn rất nhiều so với các hệ thống dựa trên ổ đĩa.

- Tích hợp với Scala: Spark tích hợp với ngôn ngữ lập trình Scala, cho phép bạn thao tác các tập dữ liệu phân tán giống như các tập hợp cục bộ. Không cần phải cấu trúc mọi thứ thành các thao tác map và reduce.
- Chia sẻ dữ liệu nhanh hơn: Việc chia sẻ dữ liệu giữa các thao tác nhanh hơn vì dữ liệu được lưu trữ trong bộ nhớ: Trong Hadoop truyền thống, dữ liệu được chia sẻ thông qua HDFS, điều này rất tốn kém. HDFS duy trì ba bản sao dữ liệu, spark lưu trữ dữ liệu trong bộ nhớ **mà không cần sao chép**.

## V. General Spark Cluster Architecture

- Driver điều phối công việc và thu thập kết quả từ các worker nodes.
- Worker nodes đọc/ghi dữ liệu từ HDFS và lưu trữ dữ liệu chuyển đổi trong bộ nhớ dưới dạng RDDs.
- Spark có thể chạy trên các máy ảo (VMs) trong đám mây như AWS, Google Cloud, và Azure.

## VI. Spark Component Features

### Spark SQL

- Truy cập dữ liệu thông nhất: Truy vấn các tập dữ liệu có cấu trúc bằng SQL hoặc DataFrame APIs
- Ngôn ngữ truy vấn nhanh và quen thuộc cho tất cả dữ liệu doanh nghiệp của bạn
- Sử dụng các công cụ BI để kết nối và truy vấn thông qua trình điều khiển JDBC hoặc ODBC

### Spark Streaming

- Xử lý sự kiện micro-batch cho phân tích thời gian gần thực
- Ví dụ: Các thiết bị Internet of Things (IoT), nguồn cấp Twitter, Kafka (event hub), v.v.
- Công cụ Spark thực hiện một số hành động hoặc xuất dữ liệu theo lô đến các kho dữ liệu khác nhau

### MLlib/SparkML

- Phân tích dự đoán và quy định
- Các thuật toán học máy cho: phân cụm, phân loại, hồi quy, v.v.
- Thiết kế ứng dụng thông minh từ các mô hình thống kê và thuật toán có sẵn

### GraphX

- Biểu diễn và phân tích các hệ thống được thể hiện bằng các nút đồ thị
- Theo dõi các kết nối giữa các nút đồ thị
- Áp dụng cho các trường hợp sử dụng trong vận tải, viễn thông, mạng lưới đường bộ, mô hình hóa các mối quan hệ cá nhân, mạng xã hội, v.v.

## VII. Architecture Overview

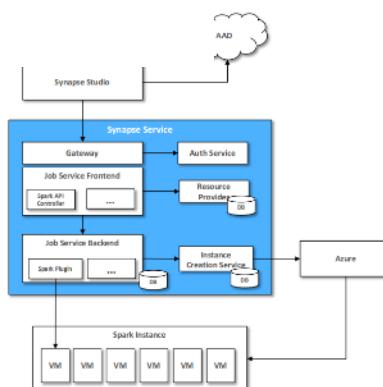
### 1. Định nghĩa

Synapse Job Service là dịch vụ quản lý và điều phối các công việc (jobs) trong môi trường Azure Synapse Analytics.

### 2. Chức năng chính

- Quản lý việc tạo và thực thi các công việc Spark
- Điều phối việc tạo cụm (cluster) và phiên làm việc Spark
- Xử lý yêu cầu từ người dùng và chuyển tiếp đến các thành phần liên quan

### 3. Cấu trúc



Synapse Job Service bao gồm hai phần chính:

#### a) Frontend:

- Tiếp nhận yêu cầu từ Synapse Gateway
- Chuyển tiếp yêu cầu đến backend
- Giao tiếp trực tiếp với Livy để thực thi câu lệnh Spark

#### b) Backend:

- Tạo các công việc cụ thể (ví dụ: tạo cụm, tạo phiên Spark)
- Tương tác với Synapse Resource Provider để lấy thông tin chi tiết về Workspace và Spark pool.
- Ủy quyền yêu cầu tạo cụm cho Synapse Instance Service.
- Lưu trữ thông tin về điểm cuối Livy cho mỗi cụm được tạo.

### 4. Quy trình làm việc:

- Người dùng tạo Synapse Workspace và Spark pool và khởi chạy Synapse Studio.
- Người dùng gắn Notebook vào Spark pool và nhập một hoặc nhiều câu lệnh Spark (các khối mã).
- Client Notebook lấy token người dùng từ AAD và gửi yêu cầu tạo phiên Spark đến Synapse Gateway.

- Synapse Gateway xác thực yêu cầu và xác nhận quyền truy cập vào Workspace và Spark pool, sau đó chuyển tiếp nó đến bộ điều khiển Spark (Livy) được lưu trữ trong phần giao diện của Synapse Job Service.
- Giao diện Job Service chuyển tiếp yêu cầu đến phần backend của Job Service, tạo ra hai công việc - một để tạo cụm và một để tạo phiên Spark.
- Backend của Job Service liên hệ với Synapse Resource Provider để lấy thông tin chi tiết về Workspace và Spark pool, và ủy quyền yêu cầu tạo cụm cho Synapse Instance Service.
- Khi instance được tạo, backend của Job Service chuyển tiếp yêu cầu tạo phiên Spark đến điểm cuối Livy trong cụm.
- Khi phiên Spark được tạo, client Notebook gửi các câu lệnh Spark đến giao diện của Job Service.
- Giao diện Job Service lấy điểm cuối Livy thực tế cho cụm được tạo cho người dùng cụ thể từ backend và gửi câu lệnh trực tiếp đến Livy để thực thi.

## 5.Quy trình tạo Spark cluster trong Azure Synapse

- **Synapse Job Service** gửi yêu cầu đến **Cluster Service** để tạo BBC cluster dựa trên mô tả trong Spark pool.
- **Cluster Service** gửi yêu cầu đến Azure bằng **Azure SDK** để tạo các **VMs** (bao gồm VM chính và VM bổ sung) với **VHD chuyên biệt**.
- **VHD chuyên biệt** chứa các dịch vụ cần thiết cho loại cluster (ví dụ: Spark) cùng với cơ chế **prefetch** để tối ưu hóa.
- Khi **VM khởi động**, **Node Agent** gửi **heartbeat** đến **Cluster Service** để lấy cấu hình node.
- Các node được khởi tạo và gán vai trò dựa trên **heartbeat đầu tiên**.
- **Các node thừa** sẽ bị xóa ngay sau **heartbeat đầu tiên**.
- Khi **Cluster Service** xác định cluster đã sẵn sàng, nó trả về **Livy endpoint** cho **Job Service**.

## VIII. Spark pool

### 1. Tạo spark pool

Vào synapse azure analytics-> chọn cuoiky-workspace -> creating a Spark pool.

Microsoft Azure Upgrade Search resources, services, and docs (G+)

Home > Resource groups > cuoiky > cuoiky-workspace >

## New Apache Spark pool

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name \*  ✓

Node size family

Node size \*

Autoscale \*  Enabled  Disabled

Number of nodes \*

Dynamically allocate executors  Enabled  Disabled

Estimated price  [View pricing details](#)

[Review + create](#) [Next: Additional settings >](#)

### Apache Spark pools

 sparkpool

Apache Spark pool

Small

B2: Create Notebook on files in storage: click chuột phải sales\_data\_sampate.csv chọn newnotebook-> load DataFrame-> run

Microsoft Azure Synapse Analytics > cuoiky-workspace

Data Workspace Linked

SQL database 

files      More

← → ↑ ↓ files > sales\_data

| Name           | Last Modified     | Content type | Size   |
|----------------|-------------------|--------------|--------|
| sales_data.csv | 2024, 11:43:28 AM |              | 3.1 MB |

New SQL script  New notebook New data flow New integration dataset Manage access... Rename... Download Delete

<https://web.azure.com/resourceId=...>

Microsoft Azure Synapse Analytics > cuoiky-workspace

Data Workspace Linked

SQL database 

Notebook 2     Attach to sparkpool Language PySpark (Python) Variables

Please wait a few minutes while your session starts.

```

1 %pyspark
2 df = spark.read.load('abfss://tranghuyen.dfs.core.windows.net/sales_data_sample.csv', format='csv')
3 # If header exists uncomment line below
4 # header=True
5
6 display(df.limit(10))

```

- Canceled by ethan.chapman on 12:51:20 PM, 12/15/24

```

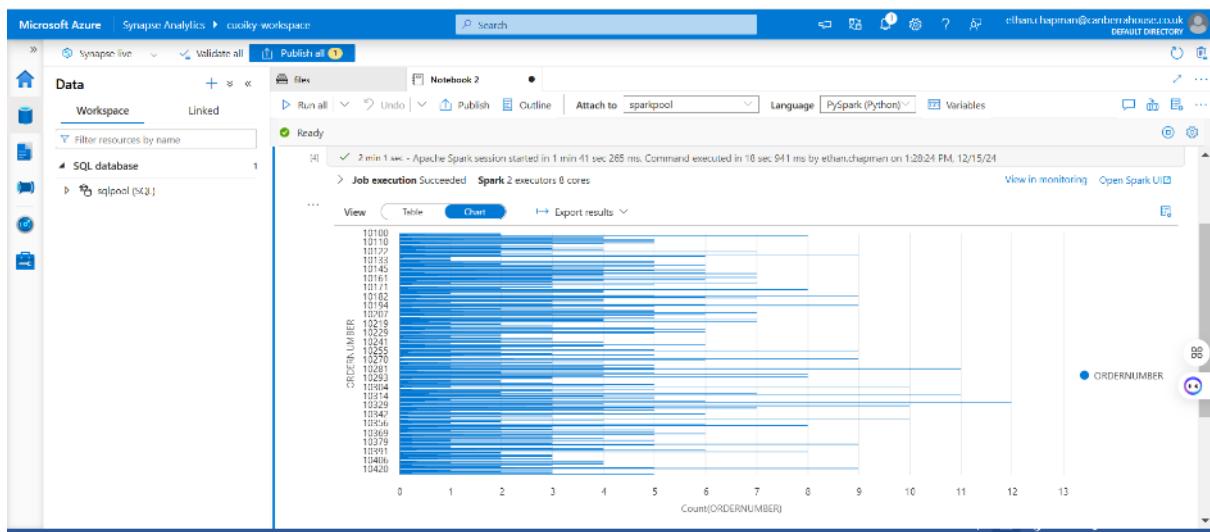
%%pyspark
df = spark.read.load('abfss://files@tranhuyen.dfs.core.windows.net/sales_data_sample.csv', format='csv')
If header exists uncomment line below
header=True
display(df.limit(10))

```

Job execution succeeded. Spark 2 executors 8 cores.

| ORDERNUMBER | QUANTITYORDERED | PRICEEACH | ORDERLINENUMBER | SALES   |
|-------------|-----------------|-----------|-----------------|---------|
| 10107       | 30              | 95.7      | 2               | 2071    |
| 10121       | 34              | 81.35     | 5               | 2765.9  |
| 10134       | 41              | 94.74     | 2               | 3884.34 |
| 10145       | 45              | 83.26     | 6               | 3746.7  |
| 10159       | 49              | 100       | 14              | 5205.27 |
| 10168       | 36              | 96.66     | 1               | 3479.76 |

Chọn format Chart:



## 2. User experience and languages

Tạo view và select view

```

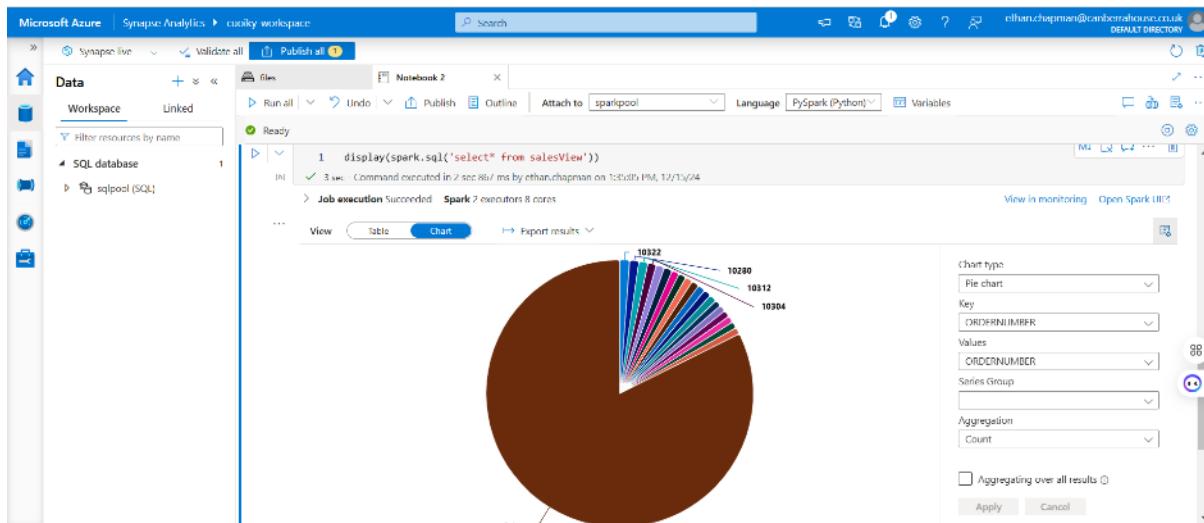
df.createOrReplaceTempView('salesView')

```

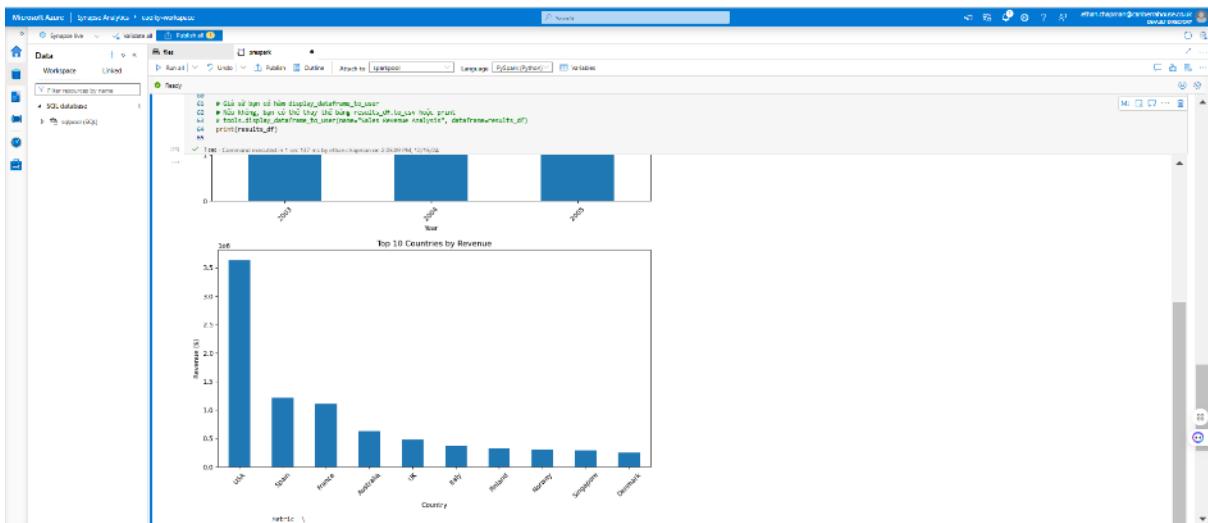
Command executed in 208 ms by ethan.chapman on 1:30:19 PM, 12/15/24

Job execution succeeded. Spark 2 executors 8 cores.

| ORDERNUMBER | QUANTITYORDERED | PRICEEACH | ORDERLINENUMBER | SALES   | ORD    |
|-------------|-----------------|-----------|-----------------|---------|--------|
| 10107       | 30              | 95.7      | 2               | 2071    | 2/2/24 |
| 10121       | 34              | 81.35     | 5               | 2765.9  | 5/7/24 |
| 10134       | 41              | 94.74     | 2               | 3884.34 | 7/1/24 |
| 10145       | 45              | 83.26     | 6               | 3746.7  | 8/2/24 |



## Phân tích doanh thu theo năm và doanh thu theo khu vực



### 3. Library Management – Python

Tính năng này cho phép khách hàng thêm thư viện Python mới ở cấp độ Spark pool. Điều này giúp quản lý các thư viện và phụ thuộc cụ thể cho các cụm Spark một cách dễ dàng hơn.

#### Lợi Ích

- Yêu cầu đầu vào: Quá trình thêm thư viện được đơn giản hóa bằng cách sử dụng tệp requirements.txt theo định dạng pip freeze.
- Thêm thư viện mới: Người dùng có thể dễ dàng thêm thư viện mới vào Spark pool.
- Cập nhật thư viện hiện có: Người dùng có thể cập nhật phiên bản của các thư viện đã cài đặt trong cụm của họ.
- Tạo cụm: Thư viện được chỉ định trong tệp yêu cầu sẽ tự động được cài đặt khi tạo cụm.
- Nhiều pool: Người dùng có thể chỉ định tệp yêu cầu khác nhau cho từng pool trong cùng một workspace, cho phép quản lý thư viện tùy chỉnh cho mỗi pool.

### 4. Spark ML Algorithms

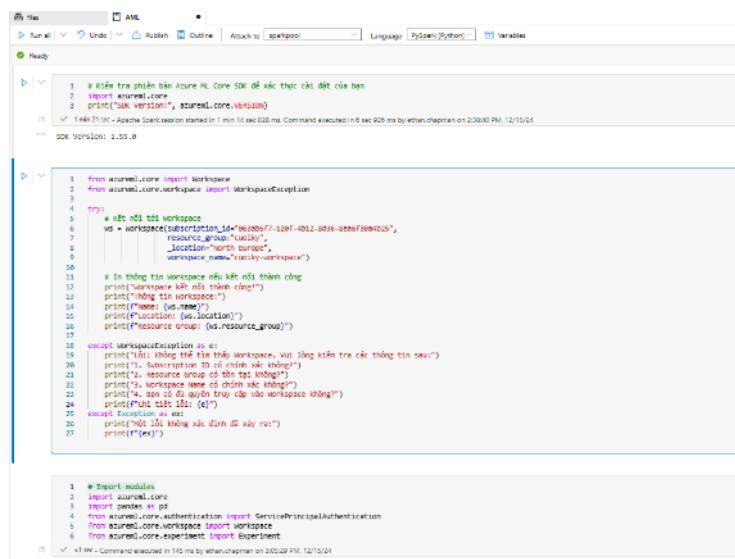
Các thuật toán học máy được tích hợp sẵn trong Spark ML

- ❖ Phân loại và Hồi quy (Classification and Regression):
  - Các mô hình tuyến tính (SVMs, hồi quy logistic, hồi quy tuyến tính)
  - Naive Bayes
  - Cây quyết định
  - Tập hợp các cây (Random Forest, Gradient-Boosted Trees)
  - Hồi quy đanding nhiệt
- ❖ Phân cụm (Clustering).
  - k-means và streaming k-means
  - Hỗn hợp Gaussian

- Phân cụm lắp lại Power (PIC)
- Phân bổ Dirichlet ẩn (LDA)
- ❖ Lọc cộng tác (Collaborative Filtering): bình phương tối thiểu xen kẽ (ALS)
- ❖ Giảm chiều dữ liệu (Dimensionality Reduction).
  - SVD (Phân rã giá trị riêng)
  - PCA (Phân tích thành phần chính)
- ❖ Khai thác mẫu thường xuyên (Frequent Pattern Mining):
  - FP-growth (Tăng trưởng mẫu thường xuyên)
  - Luật kết hợp.
- ❖ Thống kê cơ bản (Basic Statistics):
  - Thống kê tóm tắt
  - Tương quan
  - Lấy mẫu phân tầng
  - Kiểm định giả thuyết
  - Tạo dữ liệu ngẫu nhiên

## 5. Meachine learning

Synapse Notebook: Connect to AML workspace



```

1 # Kiểm tra phiên bản Azure ML Core SDK để xác thực cài đặt của bạn
2 import azureml.core
3 print("SDK version:", azureml.core.VERSION)
4
5 # !ml login - Apache Spark session started in 1 min 50 sec 620 ms. Command executed in 6 sec 420 ms by ethan.chapman on 2:08:00 PM, 12/15/24
6
7 SDK VERSION: 1.55.0
8
9
10 from azureml.core import Workspace
11 from azureml.core.workspace import WorkspaceException
12
13 try:
14 # Kết nối tới workspace
15 ws = Workspace(subscription_id="0e3bb0f7-32ef-4012-8d36-8e66f3894025",
16 resource_group="cuocvay",
17 location="north europe",
18 workspace_name="traininy-aml-synapse")
19
20 # In thông tin về workspace
21 print("Tên workspace: ", ws.name)
22 print("Subscription ID: ", ws.subscription_id)
23 print("Resource group: ", ws.resource_group)
24 print("Location: ", ws.location)
25 print("Resource group: ", ws.resource_group)
26
27 except WorkspaceException as e:
28 print("Không thể kết nối đến workspace, vui lòng kiểm tra các thông tin sau:")
29 print("1. Subscription ID có chính xác không?")
30 print("2. Resource group có tồn tại không?")
31 print("3. Địa chỉ IP của bạn có truy cập vào workspace không?")
32 print("4. Bạn có đủ quyền truy cập vào workspace không?")
33 print("5. Phiên bản SDK có lỗi không? [E]")
34
35 except Exception as e:
36 print("Lỗi không xác định đã xảy ra!")
37
38 print("OK")

```

Synapse Notebook: Configure AML job to run on Synapse

Synapse Notebook: Run AML job

## PHẦN 8: INDUSTRY-LEADING SECURITY AND COMPLIANCE

### I. Threat Protection (Bảo vệ môi đe dọa)

Làm thế nào để liệt kê và theo dõi các lỗ hổng SQL tiềm ẩn?

Giảm thiểu mọi cấu hình bảo mật sai trước khi chúng trở thành vấn đề nghiêm trọng.

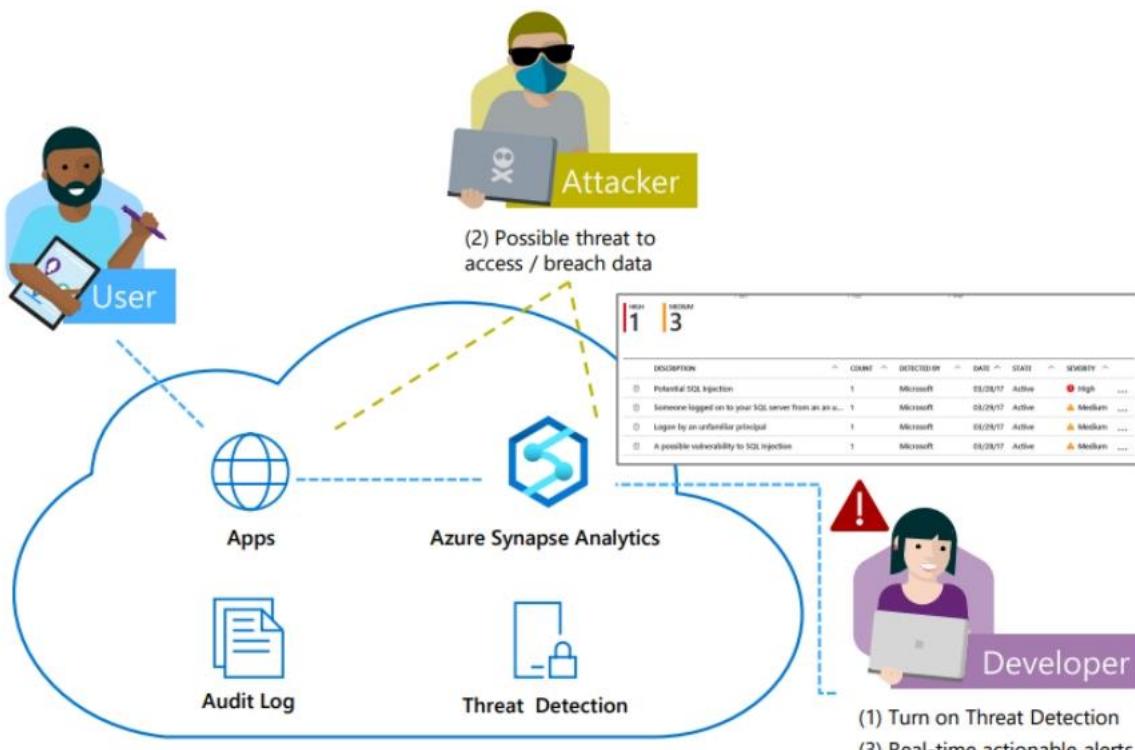
Làm thế nào để phát hiện và cảnh báo về hoạt động đáng ngờ trong cơ sở dữ liệu?

Phát hiện và xử lý mọi tấn công rò rỉ dữ liệu hoặc SQL Injection.



#### 1. Threat Detection

**SQL threat detection:** Phát hiện và điều tra hoạt động bất thường trong cơ sở dữ liệu.



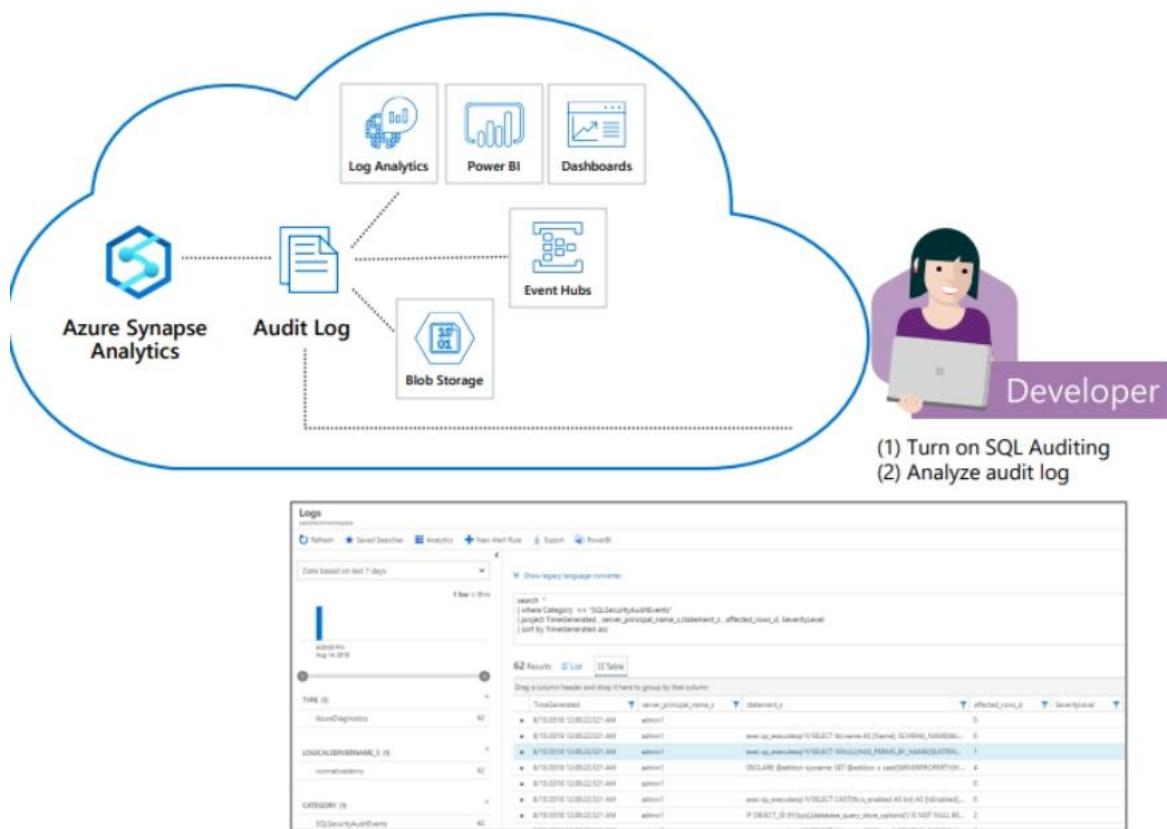
- (1) Bật phát hiện môi đe dọa.
  - (2) Nguy cơ truy cập hoặc xâm phạm dữ liệu.
  - (3) Cảnh báo theo thời gian thực có thể hành động.
- Phát hiện các cuộc tấn công SQL Injection tiềm ẩn.

- Phát hiện truy cập bất thường và hoạt động rò rỉ dữ liệu.
- Cung cấp cảnh báo có thể hành động để điều tra và khắc phục.
- Xem tất cả cảnh báo cho Azure Tenant thông qua Azure Security Center.

## 2. Auditing (Kiểm tra)

**SQL auditing in Azure Log Analytics and Event Hubs:** Hiểu sâu về nhật ký kiểm tra cơ sở dữ liệu.

- **Azure Log Analytics:** Lưu trữ và phân tích nhật ký kiểm tra SQL, giúp phát hiện các hoạt động bất thường và bảo mật.
- **Azure Event Hubs:** Chuyển tiếp nhật ký kiểm tra đến các hệ thống phân tích bên ngoài, tích hợp với công cụ giám sát thời gian thực.



- (1) Bật tính năng kiểm tra SQL (SQL Auditing).
- (2) Phân tích nhật ký kiểm tra.

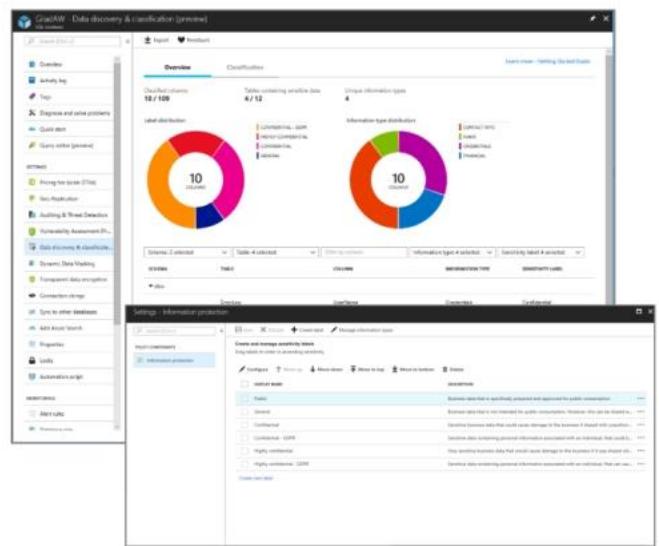
Có thể cấu hình thông qua chính sách kiểm tra (audit policy).

- Nhật ký kiểm tra SQL có thể lưu trữ tại: Tài khoản Azure Storage, Azure Log Analytics, Azure Event Hubs.
- Cung cấp bộ công cụ phong phú để: Điều tra cảnh báo bảo mật, theo dõi truy cập vào dữ liệu nhạy cảm.

### 3. Data Discovery & Classification

**SQL Data Discovery & Classification:** Phát hiện, phân loại, bảo vệ và theo dõi truy cập vào dữ liệu

- Tự động phát hiện các cột chứa dữ liệu nhạy cảm.
- Thêm nhãn dữ liệu nhạy cảm lâu dài.
- Kiểm tra và phát hiện truy cập vào dữ liệu nhạy cảm.
- Quản lý nhãn cho toàn bộ Azure Tenant thông qua Azure Security Center. nhạy cảm.



### SQL Data Discovery & Classification -

setup Bước 1: Bật Advanced Data Security trên máy chủ SQL logic.

Bước 2: Sử dụng các gợi ý hoặc phân loại thủ công để xác định tất cả các cột nhạy cảm trong bảng của bạn.

| SCHEMA          | TABLE                      | COLUMN          | INFORMATION TYPE | SENSITIVITY LABEL |
|-----------------|----------------------------|-----------------|------------------|-------------------|
| externalstaging | dimUSFIPSCode              | StatePostalCode | Contact Info     | Confidential      |
| externalstaging | dimWeatherObservationSites | StatePostalCode | Contact Info     | Confidential      |
| externalstaging | factDroughtMeasurements    | StatePostalCode | Contact Info     | Confidential      |
| externalstaging | factWaterUsageMeasurements | StatePostalCode | Contact Info     | Confidential      |

## SQL Data Discovery & Classification – audit sensitive data access

Bước 1: Cấu hình tính năng kiểm tra (auditing) cho kho dữ liệu (Data Warehouse) mục tiêu. Điều này có thể được áp dụng cho một kho dữ liệu riêng lẻ hoặc tất cả các cơ sở dữ liệu trên một máy chủ.

Bước 2: Điều hướng đến nhật ký kiểm tra (audit logs) trong tài khoản lưu trữ và tải xuống các tệp nhật ký .xel về máy cục bộ.

DOWNLOADED XE LOG FILES TO LOCAL MACHINE.

**sqldbauditlogs**  
Container

Search (Ctrl+ /)

Upload Refresh Change access level Delete Acquire lease Break lease

Authentication method: Access key (Switch to Azure AD User Account)  
Location: sqldbauditlogs / ayotestserver / ayotestdw / SqlDbAuditing\_Audit / 2019-04-02

Search blobs by prefix (case-sensitive)

| NAME               | MODIFIED             | ACCESS TIER    | BLOB TYPE   | SIZE    | LEASE STATE |
|--------------------|----------------------|----------------|-------------|---------|-------------|
| 01_34_30_090_0.xls | 4/1/2019, 6:34:31 PM | Hot (Inferred) | Append blob | 7.5 KiB | Available   |

Bước 3: Mở nhật ký bằng Extended Events Viewer trong SQL Server Management Studio (SSMS). Cấu hình trình xem để bao gồm cột data\_sensitivity\_information.

02\_26\_37\_736\_0.xls

Displaying 24785 Events

| name        | timestamp                   | affected_rows | application_name             | client_ip | data_sensitivity_information | database_name |
|-------------|-----------------------------|---------------|------------------------------|-----------|------------------------------|---------------|
| audit_event | 2019-02-26 18:38:35.7892923 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.7661039 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.7052286 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.6873633 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.6680990 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.6490621 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.6292824 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.6110493 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.5911164 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.5739871 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.5557121 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.5393015 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.5213010 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.5032121 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.4856126 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.4675595 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.4487751 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |
| audit_event | 2019-02-26 18:38:35.4290439 | 0             | .Net SqlClient Data Provider | 10.0.0.4  |                              | master        |

Event: audit\_event (2019-02-26 18:38:35.6680990)

Details

|                              |                                                                                   |
|------------------------------|-----------------------------------------------------------------------------------|
| Field                        | Value                                                                             |
| action_id                    | 1178681924                                                                        |
| additional_information       | <login_information><error_code>18456</error_code><error_st...</login_information> |
| affected_rows                | 0                                                                                 |
| application_name             | .Net SqlClient Data Provider                                                      |
| audit_schema_version         | 1                                                                                 |
| class_type                   | 16964                                                                             |
| client_ip                    | 10.0.0.4                                                                          |
| connection_id                | F1AD6457-9F40-409B-B43C-E638AAF47902                                              |
| data_sensitivity_information |                                                                                   |
| database_name                | master                                                                            |
| database_principal_id        | -1                                                                                |
| database_principal_name      |                                                                                   |
| duration_milliseconds        | 0                                                                                 |
| event_time                   | 2019-02-26 18:38:35.6743004                                                       |
| host_name                    | usgsvm089                                                                         |
| is_column_permission         | False                                                                             |
| object_id                    | 5                                                                                 |
| object_name                  | master                                                                            |

## II. Network Security - Business requirements

Làm thế nào để triển khai công lập mạng?

Dữ liệu ở các mức độ bảo mật khác nhau cần được truy cập từ các địa điểm khác nhau.

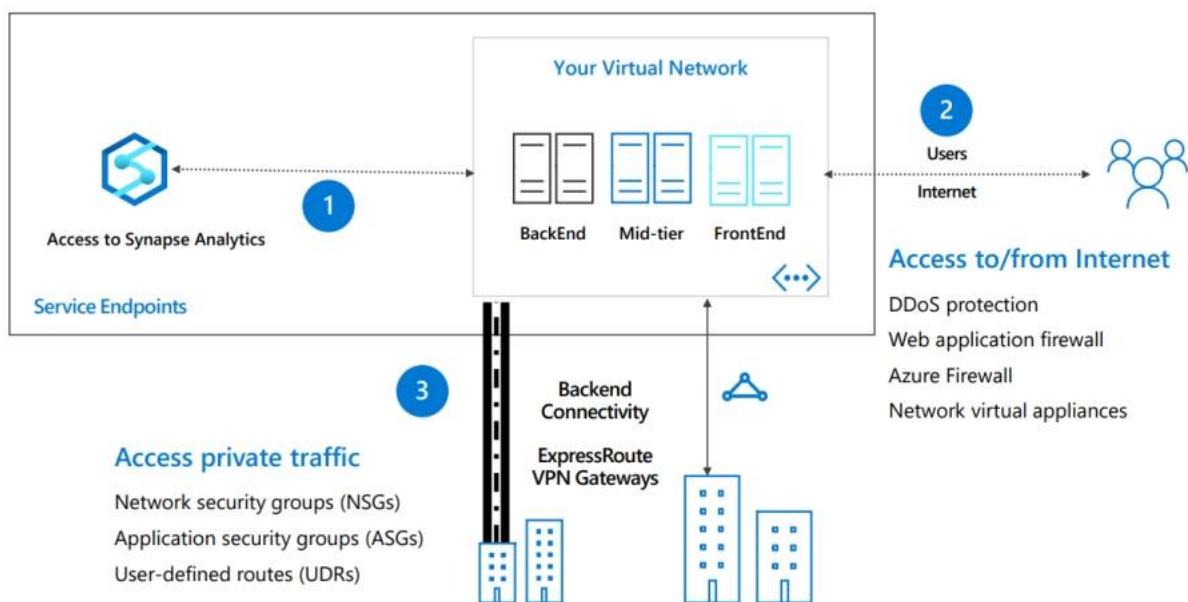
Làm thế nào để đạt được sự phân tách?

Không cho phép truy cập từ các thực thể bên ngoài ranh giới bảo mật mạng của công ty.



### 1. Application-access patterns

#### Azure networking: application-access patterns



- **Truy cập Synapse Analytics (Service Endpoints):** Người dùng hoặc dịch vụ trong mạng ảo (Virtual Network - VNet) kết nối tới Synapse Analytics qua Service Endpoints.
- **Truy cập đến/từ Internet:** Người dùng bên ngoài hoặc dịch vụ từ internet kết nối vào hệ thống thông qua các lớp bảo vệ.
- **Công cụ bảo vệ:**
  - **DDoS Protection:** Ngăn chặn các cuộc tấn công từ chối dịch vụ.

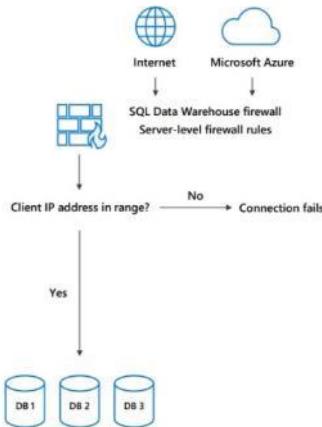
- **Web Application Firewall (WAF)**: Bảo vệ ứng dụng khỏi các tấn công phổ biến như SQL Injection hoặc XSS.
  - **Azure Firewall**: Giám sát và kiểm soát toàn bộ lưu lượng đến/đi qua mạng.
  - **Network Virtual Appliances**: Thiết bị ảo hóa giúp tăng cường bảo mật mạng.
- **Truy cập lưu lượng riêng tư (Access Private Traffic)**: Lưu lượng nội bộ giữa các thành phần trong mạng ảo (BackEnd, Mid-Tier, FrontEnd) được quản lý và bảo vệ.
- **Công cụ sử dụng:**
  - **Network Security Groups (NSGs)**: Quy định các quy tắc cho phép/chặn lưu lượng dựa trên IP, cổng và giao thức.
  - **Application Security Groups (ASGs)**: Quản lý bảo mật theo nhóm ứng dụng.
  - **User-defined Routes (UDRs)**: Định nghĩa đường đi tùy chỉnh cho lưu lượng.
- **Kết nối Backend (ExpressRoute và VPN Gateways)**: Dữ liệu giữa Azure và mạng nội bộ của tổ chức được truyền qua kết nối riêng.
- **Công cụ sử dụng:**
  - **ExpressRoute**: Kết nối trực tiếp giữa mạng nội bộ và Azure mà không qua internet công cộng, đảm bảo bảo mật và hiệu suất cao.
  - **VPN Gateways**: Kết nối an toàn qua VPN nếu không có ExpressRoute.

## 2. Firewall (Tường lửa)

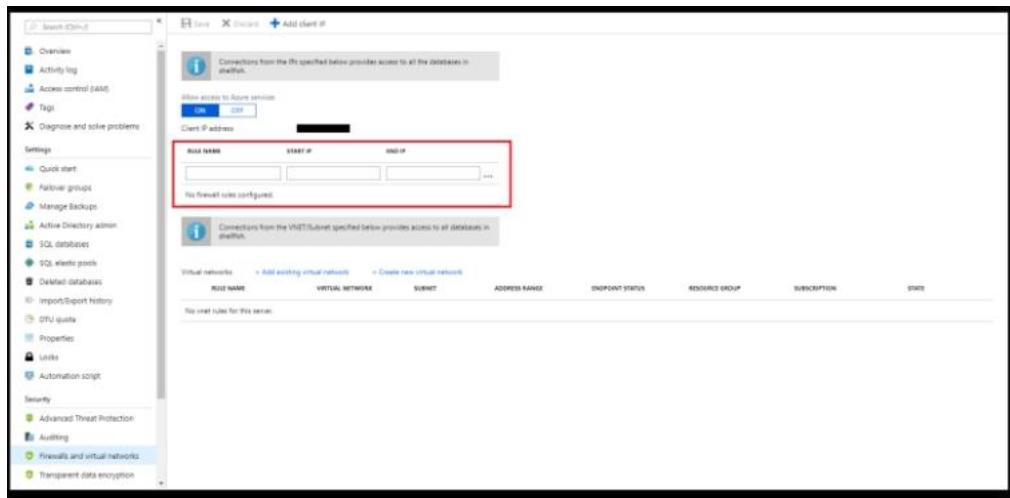
### Securing with firewalls

- Tổng quan

- **Mặc định:** Tất cả truy cập vào Azure Synapse Analytics đều bị chặn bởi tường lửa.
  - **Quản lý:** Tường lửa quản lý các quy tắc mạng ảo dựa trên các service endpoints của mạng ảo.
- **Quy tắc**
- Cho phép IP cụ thể hoặc dải IP được đưa vào danh sách trắng.
  - Cho phép các ứng dụng Azure kết nối.



### Firewall configuration on the portal:



- **Mặc định:** Azure chặn tất cả các kết nối bên ngoài đến **cổng 1433** để đảm bảo bảo mật.
- **Cấu hình kết nối:** Thực hiện các bước sau:

1. Truy cập vào **Azure Synapse Analytics Resource**.
2. Chọn **Server name**.
3. Vào mục **Firewalls and Virtual Networks** để cấu hình:
  - Thêm địa chỉ IP được phép truy cập.
  - Kích hoạt hoặc tùy chỉnh các quy tắc mạng ảo.

## Firewall configuration using REST API: Quản lý quy tắc tường lửa qua REST API

### 1. Yêu cầu xác thực:

- Các thao tác quản lý quy tắc tường lửa qua REST API cần được xác thực.
- Xem thêm tại: *Authenticating Service Management Requests*.

```
PUT
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/firewallRules/{firewallRuleName}?api-version=2014-04-01REQUEST BODY
{
 "properties": {
 "startIpAddress": "0.0.0.3",
 "endIpAddress": "0.0.0.3"
 }
}

DELETE
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/firewallRules/{firewallRuleName}?api-version=2014-04-01

GET
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/firewallRules/{firewallRuleName}?api-version=2014-04-01
```

### 2. Các thao tác REST API:

- Tạo hoặc cập nhật quy tắc cấp máy chủ: Sử dụng phương thức **PUT**.
- Xóa quy tắc cấp máy chủ: Sử dụng phương thức **DELETE**.
- Liệt kê các quy tắc tường lửa: Sử dụng phương thức **GET**.

## Firewall configuration using PowerShell/T-SQL

### Windows PowerShell Azure cmdlets

```
New-AzureRmSqlServerFirewallRule
Get-AzureRmSqlServerFirewallRule
Set-AzureRmSqlServerFirewallRule
```

### Transact SQL

```
sp_set_firewall_rule
sp_delete_firewall_rule
```

```
PS Allow external IP access to SQL DW
PS C:\> New-AzureRmSqlServerFirewallRule
 -ResourceGroupName "myResourceGroup" `
 -ServerName $servername `
 -FirewallRuleName "AllowSome" `
 -StartIpAddress "0.0.0.0" `
 -EndIpAddress "0.0.0.0"

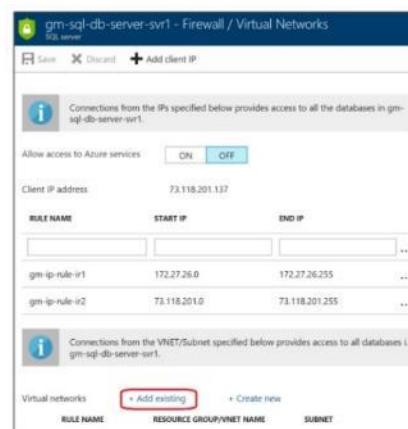
-- T-SQL Allow external IP access to SQL DW
EXECUTE sp_set_firewall_rule
 @name = N'ContosoFirewallRule',
 @start_ip_address = '192.168.1.1',
 @end_ip_address = '192.168.1.10'
```

## Virtual networks service endpoints

Cấu hình kết nối với Azure Synapse Analytics

### Các bước cấu hình:

- Truy cập **Azure Synapse Analytics Resource**.
- Chọn **Server name**.
- Đi tới mục **Firewalls and Virtual Networks**:
  - Thêm các quy tắc mạng để cho phép kết nối.
  - Cấu hình các địa chỉ IP hoặc quy tắc mạng ảo.



- **Tùy chọn thay thế:** Có thể sử dụng **REST API** hoặc **PowerShell** để tự động hóa quá trình này.

**Lưu ý:**

- Mặc định, **máy ảo (VMs)** trên subnet không thể giao tiếp với SQL Data Warehouse.
- Để cho phép, cần cấu hình **Virtual Network Service Endpoint**, sau đó tham chiếu endpoint này trong quy tắc mạng.

### III. Authentication - Business requirements

*Làm thế nào để cấu hình Azure Active Directory (AAD) với Azure Synapse Analytics?*

Tôi muốn kiểm soát bổ sung dưới dạng xác thực yếu tố (MFA).

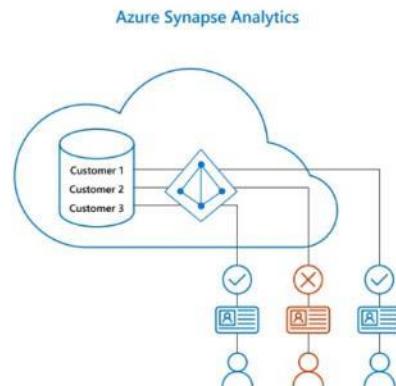
*Làm thế nào để cho phép các tài khoản không thuộc Microsoft xác thực? Azure Active Directory trust architecture*



#### 1. Azure Active Directory authentication Tổng quan

- **Quản lý danh tính người dùng** tại một nơi duy nhất.
- Kích hoạt truy cập vào **Azure Synapse Analytics** và các dịch vụ Microsoft khác thông qua **Azure Active Directory (AAD)**.

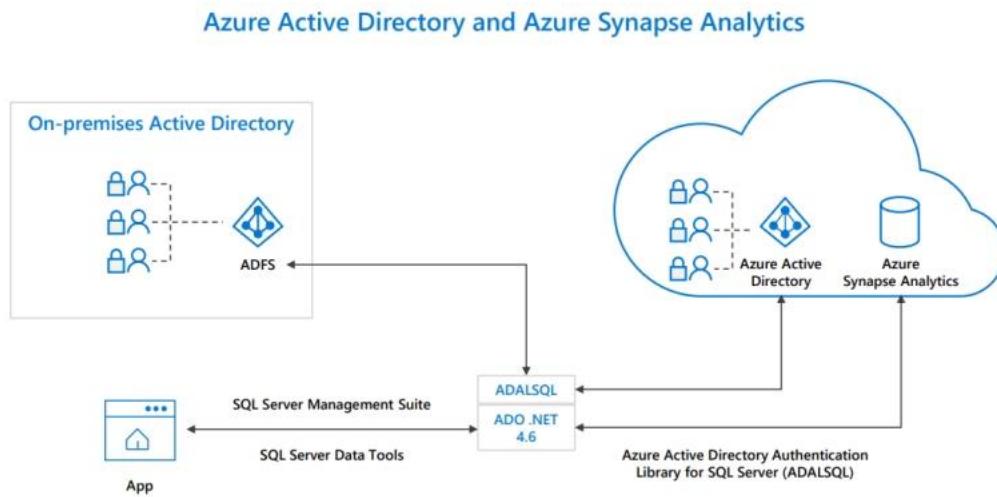
**Lợi ích:**



- **Thay thế xác thực SQL Server:** Sử dụng AAD thay vì xác thực truyền thống dựa trên tài khoản SQL Server.
- **Giảm sự phân tán danh tính:** Hạn chế việc tạo và quản lý nhiều danh tính trên các cơ sở dữ liệu khác nhau.
- **Dễ dàng xoay vòng mật khẩu:** Thực hiện xoay vòng mật khẩu tập trung tại AAD.
- **Quản lý quyền với nhóm AAD:** Quản lý quyền cơ sở dữ liệu bằng cách sử dụng **nhóm AAD** bên ngoài.

- **Loại bỏ lưu trữ mật khẩu:** Không cần lưu trữ mật khẩu trong hệ thống, tăng cường bảo mật.

## Azure Active Directory trust architecture



- **On-premises Active Directory:**
  - Người dùng và ứng dụng trong môi trường **Active Directory nội bộ** (On-premises) kết nối thông qua **ADFS** (Active Directory Federation Services) để liên kết với Azure Active Directory.
- **Azure Active Directory (AAD):**
  - Quản lý danh tính và quyền truy cập người dùng trên đám mây.
  - Kết nối và cung cấp quyền truy cập vào **Azure Synapse Analytics**.
- **ADO.NET và ADALSQL:**
  - Thư viện xác thực **Azure Active Directory Authentication Library for SQL Server (ADALSQL)**:
    - Hỗ trợ các ứng dụng (SQL Server Management Suite, SQL Server Data Tools) kết nối với Azure Synapse thông qua danh tính AAD.
- **Kết nối ứng dụng:**
  - Ứng dụng sử dụng **ADO.NET 4.6** hoặc công cụ SQL để kết nối với Synapse, đảm bảo xác thực an toàn thông qua AAD.

## 2. SQL authentication

### Tổng quan về phương thức xác thực

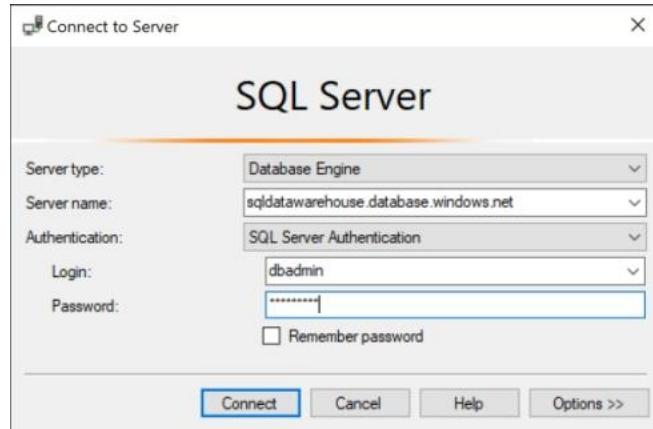
#### – Phương thức xác thực:

- Sử dụng **tên người dùng (username)** và **mật khẩu (password)**.

```
-- Connect to master database and create a login
CREATE LOGIN ApplicationLogin WITH PASSWORD = 'Strong_password'
CREATE USER ApplicationUser FOR LOGIN ApplicationLogin;

-- Connect to SQL DW database and create a database user
CREATE USER DatabaseUser FOR LOGIN ApplicationLogin;
```

- **Quản trị viên máy chủ (Server Admin):**
  - Khi tạo máy chủ logic cho kho dữ liệu, bạn đã chỉ định một **tài khoản quản trị viên máy chủ** với tên người dùng và mật khẩu.
  - Tài khoản này có quyền xác thực và truy cập bất kỳ cơ sở dữ liệu nào trên máy chủ với tư cách **chủ sở hữu cơ sở dữ liệu (database owner)**.
- **Tạo người dùng và vai trò:**
  - Bạn có thể tạo **đăng nhập người dùng (user logins)** và **vai trò (roles)** bằng cú pháp SQL quen thuộc.



#### IV. Access Control - Business requirements

*Làm thế nào để hạn chế quyền truy cập dữ liệu nhạy cảm cho người dùng cụ thể trong cơ sở dữ liệu?*

*Làm thế nào để đảm bảo người dùng chỉ có quyền truy cập dữ liệu liên quan?*

Ví dụ, trong một bệnh viện, chỉ nhân viên y tế mới được phép xem dữ liệu bệnh nhân liên quan đến họ—không phải dữ liệu của tất cả bệnh nhân.



##### 1. Object-level security (tables, views, and more)

##### Tổng quan về GRANT

- **Quản lý quyền truy cập: GRANT**: GRANT được sử dụng để cấp quyền cho các bảng, views, stored procedures và functions cụ thể.
- **Ngăn chặn truy vấn trái phép**: Bằng cách chỉ định quyền, ngăn chặn các truy vấn không được phép đối với một số bảng.
- **Đơn giản hóa bảo mật**: Triển khai bảo mật ở **cấp cơ sở dữ liệu**, thay vì cấp ứng dụng, giúp dễ dàng quản lý và thiết kế hệ thống hơn.

```
-- Grant SELECT permission to user RosaQdM on table Person.Address in the AdventureWorks2012 database
GRANT SELECT ON OBJECT::Person.Address TO RosaQdM;
GO

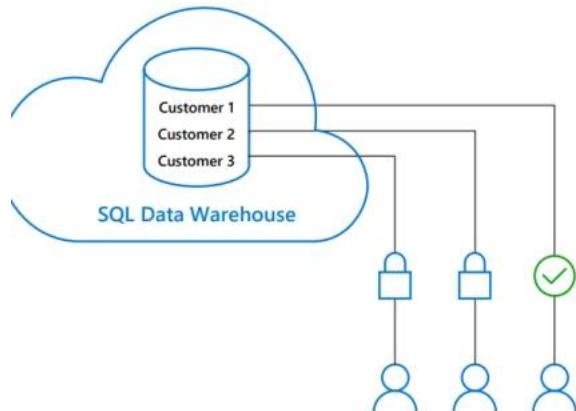
-- Grant REFERENCES permission on column BusinessEntityID in view HumanResources.vEmployee to user Wanida
GRANT REFERENCES(BusinessEntityID) ON OBJECT::HumanResources.vEmployee TO Wanida WITH GRANT OPTION;
GO

-- Grant EXECUTE permission on stored procedure HumanResources.uspUpdateEmployeeHireInfo to an application role called Recruiting11
USE AdventureWorks2012;
GRANT EXECUTE ON OBJECT::HumanResources.uspUpdateEmployeeHireInfo TO RECRUITING11;
GO
```

## 2. Row-level security (RLS)

### Tổng quan về kiểm soát truy cập chi tiết

- **Kiểm soát truy cập hàng dữ liệu:** Quản lý quyền truy cập ở **mức độ chi tiết (rows)** trong bảng cơ sở dữ liệu.
- **Ngăn truy cập trái phép:** Đảm bảo bảo mật khi nhiều người dùng chia sẻ cùng một bảng.
- **Không cần lọc kết nối:** Loại bỏ nhu cầu triển khai lọc kết nối trong các ứng dụng đa tenant.
- **Quản lý dễ dàng:** Thực hiện thông qua **SQL Server Management Studio (SSMS)** hoặc **SQL Server Data Tools (SSDT)**.
- **Logic bảo mật rõ ràng:** Dễ dàng xác định và thực thi logic bảo mật trong cơ sở dữ liệu và gắn liền với bảng.



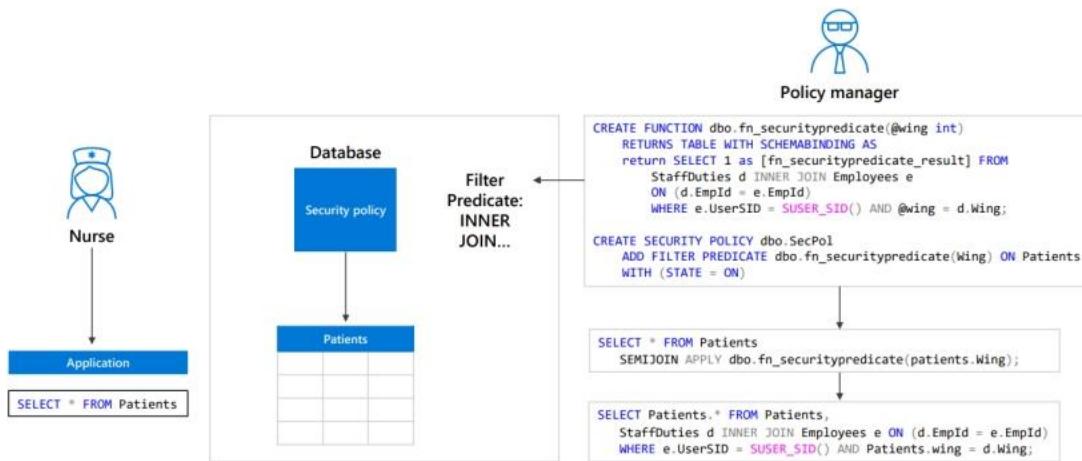
### Tạo chính sách

- **Filter predicates:** Lọc các hàng được phép truy cập trong các thao tác đọc (SELECT), cập nhật (UPDATE), và xóa (DELETE).
- **Ví dụ:** Sử dụng cú pháp **CREATE SECURITY POLICY** để triển khai chính sách bảo mật

```
-- The following syntax creates a security policy with a filter predicate for the
Customer table
CREATE SECURITY POLICY [FederatedSecurityPolicy]
ADD FILTER PREDICATE [rls].[fn_securitypredicate]([CustomerId])
ON [dbo].[Customer];

-- Create a new schema and predicate function, which will use the application user ID
-- stored in CONTEXT_INFO to filter rows.
CREATE FUNCTION rls.fn_securitypredicate (@AppUserId int)
RETURNS TABLE
WITH SCHEMABINDING
AS
RETURN (
 SELECT 1 AS fn_securitypredicate_result
 WHERE
 DATABASE_PRINCIPAL_ID() = DATABASE_PRINCIPAL_ID('dbo') -- application context
 AND CONTEXT_INFO() = CONVERT(VARBINARY(128), @AppUserId);
GO
```

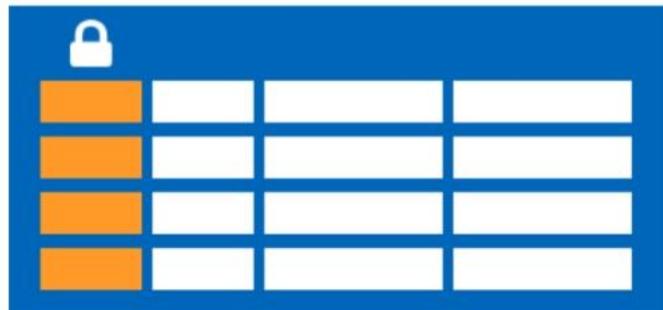
### Ba bước tạo chính sách bảo mật:



1. Tạo predicate và chính sách bảo mật: Quản trị viên chính sách sử dụng T-SQL để tạo filter predicate và chính sách bảo mật, gắn kết predicate với bảng Patients.
2. Người dùng ứng dụng truy vấn dữ liệu: Ví dụ: Y tá thực hiện truy vấn SELECT trên bảng Patients.
3. Chính sách bảo mật áp dụng tự động: Chính sách bảo mật tự động viết lại truy vấn, áp dụng filter predicate để đảm bảo chỉ dữ liệu được phép mới hiển thị.

### 3. Column-level security

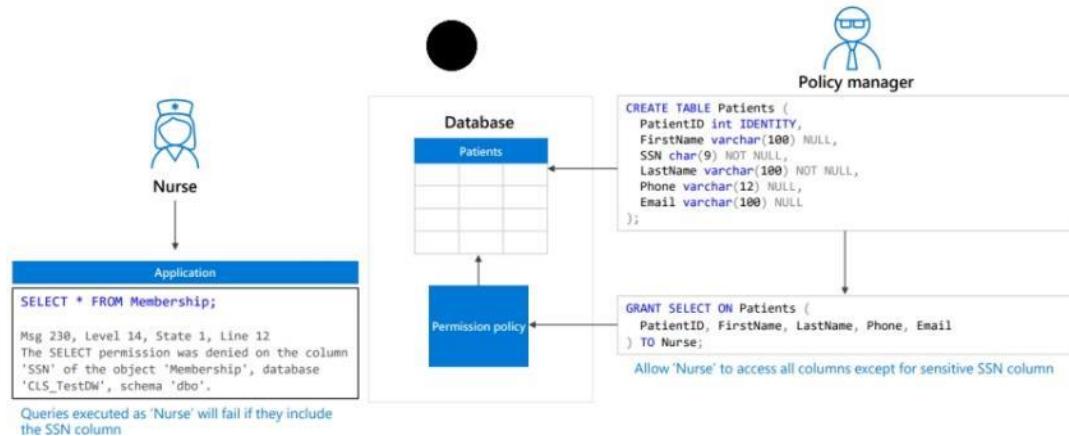
- Kiểm soát quyền truy cập **cột**: Hạn chế truy cập vào các cột trong bảng dựa trên nhóm khách hàng hoặc ngữ cảnh thực thi.
- Đơn giản hóa bảo mật: Áp dụng logic hạn chế ở tầng cơ sở dữ liệu thay vì tầng ứng dụng.
- Quản lý: Sử dụng lệnh GRANT trong T-SQL.
- Hỗ trợ xác thực: Tương thích với cả Azure Active Directory (AAD) và SQL authentication.



Ba bước tạo chính sách quyền truy cập:

1. Tạo chính sách quyền: Quản trị viên chính sách dùng T-SQL để tạo chính sách quyền, gắn chính sách với bảng Patients và nhóm cụ thể.

- Người dùng ứng dụng truy vấn dữ liệu: Ví dụ: Y tá thực hiện truy vấn SELECT trên bảng Patients.
- Chính sách quyền ngăn truy cập dữ liệu nhạy cảm: Chính sách tự động áp dụng, chặn quyền truy cập vào các dữ liệu nhạy cảm.

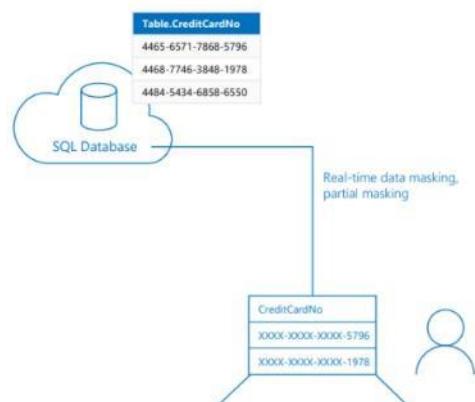


## V. Data Protection - Business requirements

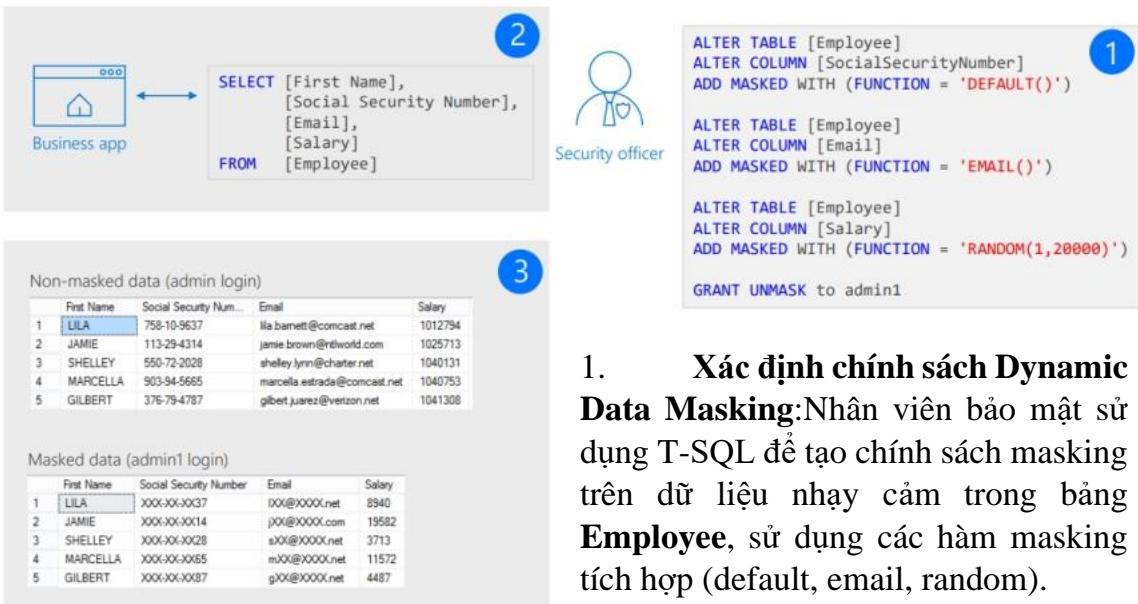
### 1. Dynamic Data Masking

#### Tổng quan

- Ngăn chặn lạm dụng dữ liệu nhạy cảm:** Ân dữ liệu khỏi người dùng không được phép.
- Cấu hình dễ dàng:** Thực hiện trên Azure Portal mới.
- Chính sách linh hoạt:** Áp dụng ở mức bảng và cột cho nhóm người dùng được xác định.
- Masking dữ liệu theo thời gian thực:** Áp dụng khi trả về kết quả truy vấn, dựa trên chính sách.
- Nhiều chức năng masking:** Hỗ trợ ẩn toàn phần hoặc một phần cho các loại dữ liệu nhạy cảm (số thẻ tín dụng, SSN, v.v.).



## Ba bước chính:



2. **Người dùng ứng dụng truy vấn:** Người dùng ứng dụng thực hiện truy vấn SELECT trên bảng Employee.
  3. **Ẩn dữ liệu nhạy cảm:** Chính sách Dynamic Data Masking làm mờ dữ liệu nhạy cảm trong kết quả truy vấn đối với người dùng không có quyền cao.

## 2. Types of data encryption

| Mã hóa Dữ liệu          | Công nghệ Mã hóa                                              | Giá trị Khách hàng                                                              |
|-------------------------|---------------------------------------------------------------|---------------------------------------------------------------------------------|
| Khi truyền (In transit) | Transport Layer Security (TLS) 1.2                            | Bảo vệ dữ liệu giữa máy khách và máy chủ, ngăn chặn tấn công man-in-the-middle. |
| Khi lưu trữ (At rest)   | Transparent Data Encryption (TDE) cho Azure Synapse Analytics | Bảo vệ dữ liệu trên đĩa. Azure quản lý khóa, giúp dễ dàng tuân thủ quy định.    |

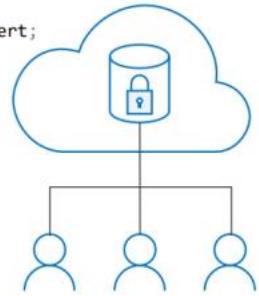


### 3. Transparent data encryption (TDE)

#### Tổng quan

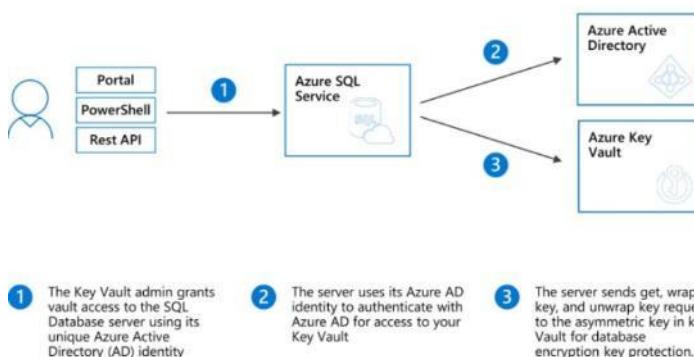
- Tất cả dữ liệu khách hàng được mã hóa khi lưu trữ.
- TDE (Transparent Data Encryption) thực hiện mã hóa và giải mã dữ liệu, tệp nhât ký theo thời gian thực trong quá trình I/O.
- Khóa mã hóa có thể được quản lý bởi dịch vụ (Service) hoặc người dùng (User).
- Thay đổi ứng dụng được giữ ở mức tối thiểu.
- Mã hóa/giải mã dữ liệu diễn ra một cách minh bạch trên trình điều khiển máy khách được kích hoạt TDE.
- Tuân thủ nhiều luật, quy định, và hướng dẫn trong các ngành công nghiệp khác nhau.

```
USE master;
GO
CREATE MASTER KEY ENCRYPTION BY PASSWORD = '<UseStrongPasswordHere>';
go
CREATE CERTIFICATE MyServerCert WITH SUBJECT = 'My DEK Certificate';
go
USE MyDatabase;
GO
CREATE DATABASE ENCRYPTION KEY
WITH ALGORITHM = AES_128
ENCRYPTION BY SERVER CERTIFICATE MyServerCert;
go
ALTER DATABASE MyDatabase
SET ENCRYPTION ON;
GO
```



#### Key Vault: Lợi ích của Khóa do Người dùng Quản lý (User Managed Keys)

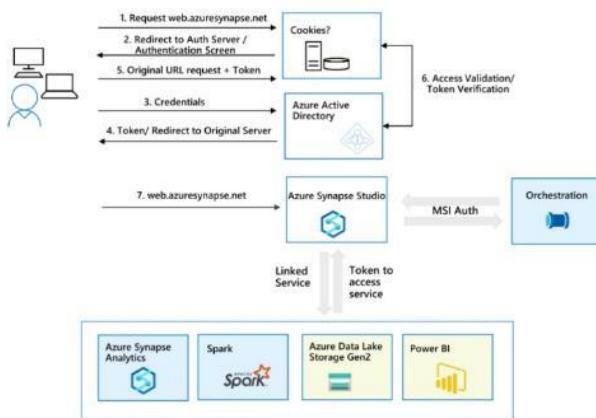
- Kiểm soát tốt hơn: Chủ động kiểm soát ai có quyền truy cập vào dữ liệu và khi nào.
- Khả dụng cao và mở rộng: Kho lưu trữ khóa trên đám mây với độ sẵn sàng và khả năng mở rộng cao.
- Quản lý khóa tập trung: Phân tách việc quản lý khóa và dữ liệu, giúp bảo mật và tổ chức hiệu quả hơn.
- Cấu hình linh hoạt: Thực hiện qua Azure Portal, PowerShell, và REST API.



- (1) Quản trị viên Key Vault cấp quyền truy cập kho lưu trữ (vault) cho máy chủ cơ sở dữ liệu SQL thông qua Azure Active Directory (AD) với danh tính duy nhất của nó.

- (2) Máy chủ sử dụng danh tính Azure AD của mình để xác thực với Azure AD và truy cập Key Vault của bạn.
- (3) Máy chủ gửi các yêu cầu get key (lấy khóa), wrap key (gói khóa), và unwrap key (mở gói khóa) đến Key Vault để bảo vệ khóa mã hóa cơ sở dữ liệu.

## VI. Single Sign-On



- **Implicit Authentication:**
  - Người dùng cung cấp thông tin đăng nhập một lần để truy cập Azure Synapse Workspace.
- **AAD Authentication (Azure Active Directory):**
  - Azure Synapse Studio yêu cầu token để truy cập từng dịch vụ liên kết dưới danh nghĩa người dùng.
  - Một token riêng được cấp cho các dịch vụ: ADLS Gen2 (Azure Data Lake Storage Gen2), azure Synapse Analytics, power BI, spark Livy API (cho Spark), management.azure.com (cấp phát tài nguyên), dev.workspace.net (phát triển artifacts), graph endpoints.
- **MSI Authentication (Managed Service Identity):** Được sử dụng cho các tác vụ tự động hóa trong Orchestration, giúp loại bỏ nhu cầu lưu trữ thông tin đăng nhập.

## PHẦN 9: APPLICATION

### I. Setup workspace

Chuẩn bị dữ liệu

```
Biến mặc định
$synapseWorkspace = "cuoiky-workspace" # Tên Synapse workspace
$dataLakeAccountName = "tranhuyen" # Tên Azure Data Lake Storage
$sparkPool = "sparkpool" # Tên Spark pool
$sqlDatabaseName = "sqlpool" # Tên SQL pool
$subscriptionName = "Azure subscription 1" # Tên Subscription
$subscriptionID = "063ab6f7-120f-4b12-8d36-8ea6f30a4b25" # Subscription ID
$region = "Southeast Asia" # Khu vực hoạt động
$sqlUser = "SQLUser"
$sqlPassword = "Chanchan123@"

Xác nhận subscription
Select-AzSubscription -SubscriptionId $subscriptionID
az account set --subscription $subscriptionID
- setup workspace với powershell
```

B1: Sử dụng nút [>\_] bên phải thanh tìm kiếm ở phía trên cùng của trang để tạo một Cloud Shell mới trong cổng thông tin Azure.

B2: Trong cửa sổ PowerShell, nhập các lệnh sau để sao chép (clone) repository Bigdata

```
git clone https://github.com/NguyenThanhThuyTrang/BigData cuoiky
```

B3: Sau khi repo đã được sao chép nhập lệnh sau để chạy script setup.ps1 có trong đó

```
cd cuoiky/cuoiky
./setup.ps1
```

Khi script thiết lập hoàn tất, trong cổng Azure, hãy truy cập vào group storage cuoiky đã tạo: trong đó chứa cuoiky-workspace một Apache Spark pool và một Dedicated SQL pool, storage account.

The screenshot shows the Azure Resource Groups interface. The 'cuoiky' resource group is selected. It lists four resources: 'cuoiky workspace' (Synapse workspace), 'sparkpool (cuoiky-workspace/sparkpool)' (Apache Spark pool), 'sqlpool (cuoiky-workspace/sqlpool)' (Dedicated SQL pool), and 'tranghuyen' (Storage account). All resources are located in Southeast Asia.

## II. Áp dụng SQL serverless

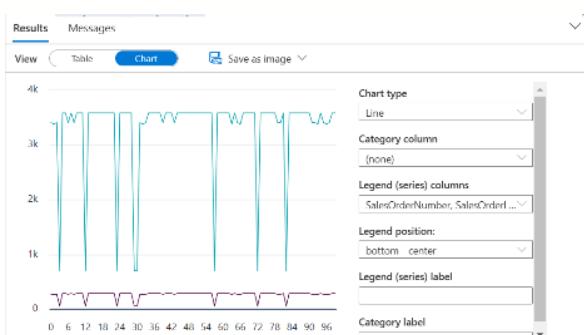
### --1. Lấy 100 dòng đầu tiên

**SELECT**

```
TOP 100 *
FROM
OPENROWSET(
 BULK 'https://tranghuyen.dfs.core.windows.net/files/sales_data/sales.csv',
 FORMAT = 'CSV',
 HEADER_ROW = TRUE,
 PARSER_VERSION = '2.0'
) AS [result]
```

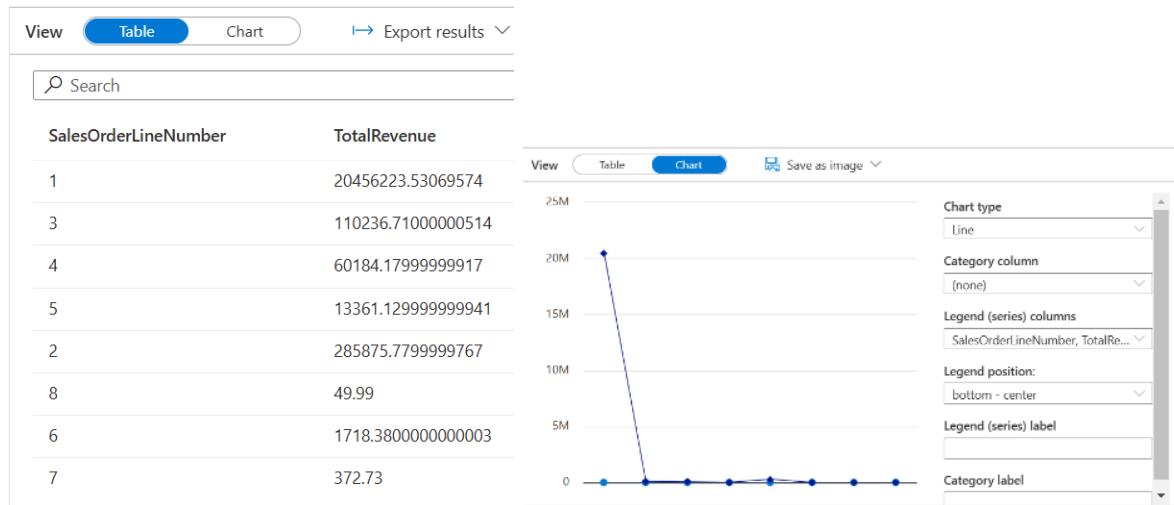
The screenshot shows the Azure Data Explorer results table. It displays the first 100 rows of sales data from a CSV file. The columns include SalesOrderNumber, SalesOrderLineNumber, OrderDate, CustomerName, EmailAddress, and Item.

| SalesOrderNumber | SalesOrderLineNumber | OrderDate  | CustomerName   | EmailAddress                  | Item          |
|------------------|----------------------|------------|----------------|-------------------------------|---------------|
| SO43701          | 1                    | 2019-07-01 | Christy Zhu    | christy12@adventure-works.com | Mountain-100  |
| SO43704          | 1                    | 2019-07-01 | Julio Ruiz     | julio1@adventure-works.com    | Mountain-100  |
| SO43705          | 1                    | 2019-07-01 | Curtis Lu      | curtis9@adventure-works.com   | Mountain-100  |
| SO43700          | 1                    | 2019-07-01 | Ruben Prasad   | ruben10@adventure-works.com   | Road-650 Blak |
| SO43703          | 1                    | 2019-07-01 | Albert Alvarez | albert7@adventure-works.com   | Road-150 Red  |
| SO43697          | 1                    | 2019-07-01 | Cale Walton    | cale1@adventure-works.com     | Road-150 Red  |
| SO43699          | 1                    | 2019-07-01 | Sydney Wright  | sydney61@adventure-works.com  | Mountain-100  |



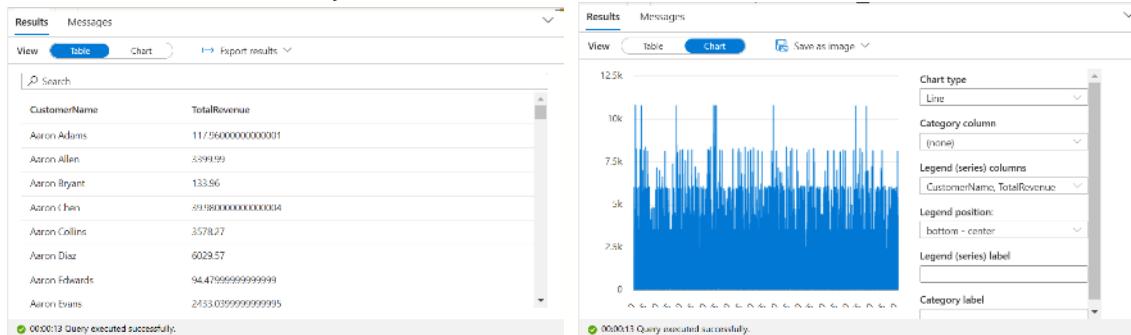
### -- 2. Tổng doanh thu theo mỗi đơn hàng

```
SELECT SalesOrderLineNumber,
 SUM(Quantity * UnitPrice) AS TotalRevenue
FROM OPENROWSET(
 BULK 'https://tranghuyen.dfs.core.windows.net/files/sales_data/sales.csv',
 FORMAT = 'CSV',
 HEADER_ROW = TRUE,
 PARSER_VERSION = '2.0'
) AS sales GROUP BY SalesOrderLineNumber;
```



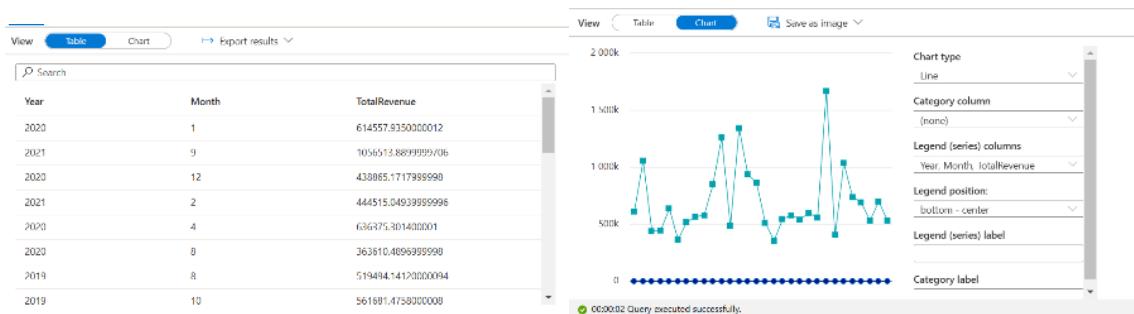
-- 3. Doanh thu theo khách hàng

```
SELECT CustomerName,
 SUM(Quantity * UnitPrice) AS TotalRevenue
 FROM OPENROWSET(
 BULK 'https://tranghuyen.dfs.core.windows.net/files/sales_data/sales.csv',
 FORMAT = 'CSV', HEADER_ROW = TRUE, PARSER_VERSION = '2.0'
) AS sales
 GROUP BY CustomerName;
```



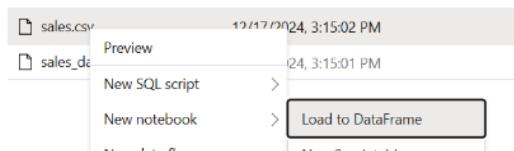
-- 4. Doanh thu theo tháng

```
SELECT YEAR(OrderDate) AS Year, MONTH(OrderDate) AS Month, SUM(Quantity * UnitPrice) AS TotalRevenue
 FROM OPENROWSET(
 BULK 'https://tranghuyen.dfs.core.windows.net/files/sales_data/sales.csv',
 FORMAT = 'CSV',
 HEADER_ROW = TRUE,
 PARSER_VERSION = '2.0'
) AS sales
 GROUP BY YEAR(OrderDate), MONTH(OrderDate);
```

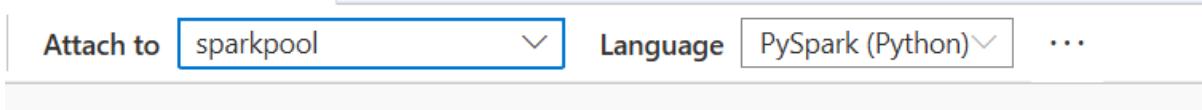


### III. Áp dụng Notebook

B1: click chuột phải chọn như sau và chọn + code để thêm cell code.

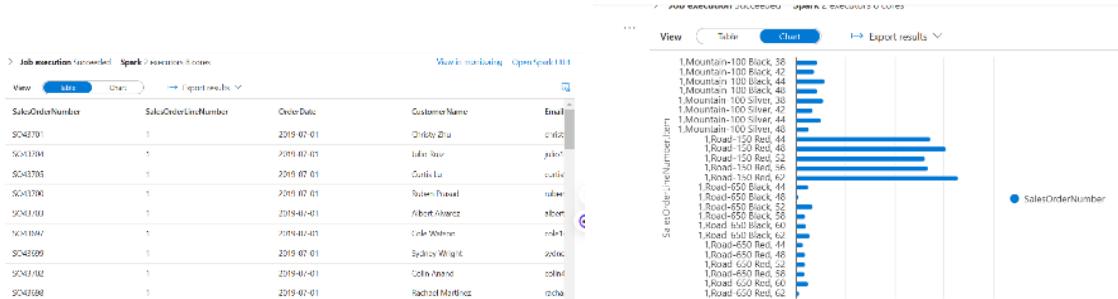


B2: chọn sparkpool và ngôn ngữ là python



B3: đọc dữ liệu

```
%pyspark
Xem 10 dòng đầu tiên
df = spark.read.load('abfss://files@tranghuyen.dfs.core.windows.net/sales_data/sales.csv', format='csv'
If header exists uncomment line below
##, header=True
)
display(df.limit(10))
```



B3: tiền xử lý dữ liệu

```

1 # 2. Tiền xử lý dữ liệu
2 from pyspark.ml.feature import StringIndexer
3
4 # Chuyển đổi các cột phân loại thành các chỉ số số
5 indexer = StringIndexer(inputCol="Item", outputCol="label")
6 df = indexer.fit(df).transform(df)

```

✓ 8 sec - Command executed in 7 sec 134 ms by ethan.chapman on 5:12:23 PM, 12/17/24

B4: Sử dụng VectorAssembler để kết hợp các cột đặc trưng thành một vector duy nhất để sử dụng trong mô hình.

```

1 from pyspark.ml.feature import VectorAssembler
2 from pyspark.sql.functions import col
3
4 # Chuyển các cột 'Quantity' và 'UnitPrice' sang kiểu số (float hoặc integer)
5 df = df.withColumn("Quantity", col("Quantity").cast("int"))
6 df = df.withColumn("UnitPrice", col("UnitPrice").cast("float"))
7
8 # Các cột đặc trưng bạn muốn sử dụng cho mô hình
9 feature_columns = ['Quantity', 'UnitPrice'] # Thêm các cột khác nếu cần
10
11 assembler = VectorAssembler(inputCols=feature_columns, outputCol="features")
12 df = assembler.transform(df)

```

B5: Chia dữ liệu thành tập train và test

```

1 # 3. Chia dữ liệu thành tập huấn luyện và kiểm tra
2 train_df, test_df = df.randomSplit([0.8, 0.2], seed=1234)
3
4 # Hiển thị 10 dòng đầu tiên của tập huấn luyện và tập kiểm tra
5 train_df.show(10)
6 test_df.show(10)

```

✓ 4 sec - Command executed in 4 sec 40 ms by ethan.chapman on 5:22:52 PM, 12/17/24

|         | SalesOrderNumber | SalesOrderLineNumber                 | OrderDate                  | CustomerName   | EmailAddress         |                         |
|---------|------------------|--------------------------------------|----------------------------|----------------|----------------------|-------------------------|
| Item    | Quantity         | UnitPrice                            | TaxAmount                  | label          | features             |                         |
| 3578.27 | SO436971         | 286.26161                            | 1 2019-07-01               | Cole Watson    | cole1@adventure-w... | Road-150 Red, 62  1     |
|         |                  | 28.0 [1.0,3578.2700195...]           |                            |                |                      |                         |
|         | SO436991         |                                      | 1 2019-07-01               | Sydney Wright  | sydney61@adventur... | Mountain-100 Silv...  1 |
|         | 3399.99          | 271.9992 104.0 [1.0,3399.9899902...] |                            |                |                      |                         |
|         | SO437001         |                                      | 1 2019-07-01               | Ruben Prasad   | ruben10@adventure... | Road-650 Black, 62  1   |
|         | 699.0982         | 55.9279                              | 90.0 [1.0,699.09820556...] |                |                      |                         |
|         | SO437021         |                                      | 1 2019-07-01               | Colin Anand    | colin45@adventure... | Road-150 Red, 44  1     |
|         | 3578.27          | 286.26161                            | 31.0 [1.0,3578.2700195...] |                |                      |                         |
|         | SO437031         |                                      | 1 2019-07-01               | Albert Alvarez | albert7@adventure... | Road-150 Red, 62  1     |
|         | 3578.27          | 286.26161                            | 28.0 [1.0,3578.2700195...] |                |                      |                         |
|         | SO437041         |                                      | 1 2019-07-01               | Julio Ruiz     | julio1@adventure-... | Mountain-100 Blac...  1 |
|         | 3374.99          | 269.9992                             | 99.0 [1.0,3374.9899902...] |                |                      |                         |
|         | SO437051         |                                      | 1 2019-07-01               | Curtis Lu      | curtis9@adventure... | Mountain-100 Silv...  1 |
|         | 3399.99          | 271.9992                             | 98.0 [1.0,3399.9899902...] |                |                      |                         |

B5: Tiến hành huấn luyện mô hình.

```

1 from pyspark.ml.classification import RandomForestClassifier
2 from pyspark.ml.evaluation import MulticlassClassificationEvaluator
3
4 # Khởi tạo mô hình phân loại
5 rf = RandomForestClassifier(labelCol="label", featuresCol="features")
6
7 # Huấn luyện mô hình
8 rf_model = rf.fit(train_df)
9
10 # Dự đoán trên tập kiểm tra
11 predictions = rf_model.transform(test_df)
12
13 # Đánh giá mô hình
14 evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="accuracy")
15 accuracy = evaluator.evaluate(predictions)
16 print(f"Accuracy: {accuracy}")

```

[18] ✓ 7 sec - Command executed in 7 sec 118 ms by ethan.chapman on 5:37:08 PM, 12/17/24

> Job execution Succeeded Spark 2 executors 8 cores

... Accuracy: 0.336742482063807

Độ chính xác của mô hình: 33.6%

## PHẦN 10: DEVELOPMENT PROBLEM

### I. Giới thiệu bài toán

Tập dữ liệu Oneline\_Retail.csv bao gồm các đơn đặt hàng được thực hiện ở các quốc gia khác nhau từ tháng 12 năm 2010 đến tháng 12 năm 2011.

Cùng xem xét ý nghĩa các trường dữ liệu trong file

- InvoiceNo: ID của đơn hàng, nếu ID bắt đầu bằng chữ "c" thể hiện đơn hàng đó bị hủy (Cancel)
- StockCode: Mã sản phẩm
- Description: Tên sản phẩm
- Quantity: Số lượng sản phẩm trên đơn đặt hàng
- InvoiceDate: Ngày và giờ khi đơn hàng được tạo
- UnitPrice: Giá sản phẩm trên mỗi đơn vị, tính bằng pound
- CustomerID: ID của khách hàng
- Country: Quốc gia nơi khách hàng cư trú

**Bài toán đặt ra:** Phân khúc khách hàng là một chiến lược trong kinh doanh và marketing, giúp doanh nghiệp chia khách hàng thành các nhóm dựa trên hành vi, thói quen hoặc đặc điểm chung. Điều này giúp xây dựng các chiến lược quảng cáo và chăm sóc khách hàng phù hợp.

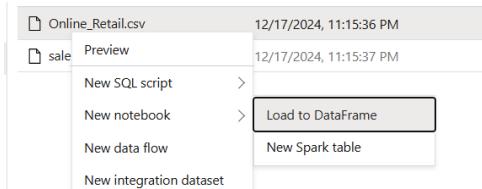
Nếu bạn ghé vào cửa hàng quần áo X vào mỗi dịp đầu tháng, có thể bạn được xếp vào nhóm những người tiêu tiền đầu tháng, nghèo cuối tháng, hay lương được trả vào cuối tháng, thích mua sắm hàng tháng, thích mua quần áo,... Và khi đó, công ty X sẽ có những quảng cáo với mác "Chỉ dành riêng cho bạn" hay "Duy nhất trong ngày hôm nay" và những quảng cáo này sẽ xuất hiện cho những người cùng nhóm với bạn, xuất hiện vào đầu tháng, các mặt hàng sale chủ yếu là quần áo, ...

Trong bài toán này, nhóm em sẽ thực hiện phân khúc khách hàng dựa trên K-means với tập dữ liệu trên.

## II. Phân tích dữ liệu tổng quan

B1: Đọc dữ liệu: click chuột phải vào Online\_Retail.csv-> new notebook->

Load to DataFrame -> run



```
1 %%pyspark
2 df = spark.read.load('abfss://filestranghuyen.dfs.core.windows.net/sales_data/OnelineRetail.csv', format='csv'
3 , header=True
4)
5 display(df.limit(10))
```

✓ 3 sec - Command executed in 2 sec 960 ms by ethan.chapman on 11:52:04 PM, 12/17/24

| InvoiceNo | StockCode | Description                     | Quantity | InvoiceDate         |
|-----------|-----------|---------------------------------|----------|---------------------|
| 536365    | 85123A    | WHITE HANGING HEART T-LIGH...   | 6        | 2010-12-01 08:26:00 |
| 536365    | 71053     | WHITE METAL LANTERN             | 6        | 2010-12-01 08:26:00 |
| 536365    | 84406B    | CREAM CUPID HEARTS COAT HA...   | 8        | 2010-12-01 08:26:00 |
| 536365    | 84029G    | KNITTED UNION FLAG HOT WAT...   | 6        | 2010-12-01 08:26:00 |
| 536365    | 84029E    | RED WOOLLY HOTTIE WHITE HE...   | 6        | 2010-12-01 08:26:00 |
| 536365    | 22752     | SET 7 BABUSHKA NESTING BOXES    | 2        | 2010-12-01 08:26:00 |
| 536365    | 21730     | GLASS STAR FROSTED T-LIGHT H... | 6        | 2010-12-01 08:26:00 |
| 536366    | 22633     | HAND WARMER UNION JACK          | 6        | 2010-12-01 08:28:00 |
| 536366    | 22632     | HAND WARMER RED POLKA DOT       | 6        | 2010-12-01 08:28:00 |
| 536367    | 84879     | ASSORTED COLOUR BIRD ORNA...    | 32       | 2010-12-01 08:34:00 |

B2: Thực hiện một số truy vấn cơ bản

```
1 # Đếm xem có bao nhiêu dòng dữ liệu
2 df.count()
```

✓ 4 sec - Command executed in 4 sec 527 ms by ethan.chapman on 11:55:21 PM, 12/17/24

> Job execution Succeeded Spark 2 executors 8 cores

[View in monitoring](#) [Open Spark UI](#)

541909

```
1 # Có bao nhiêu khách hàng
2 df.select('CustomerID').distinct().count()
```

✓ 5 sec - Command executed in 5 sec 537 ms by ethan.chapman on 11:55:27 PM, 12/17/24

> Job execution Succeeded Spark 2 executors 8 cores

[View in monitoring](#) [Open Spark UI](#)

4373

```

1 from pyspark.sql.functions import countDistinct, desc
2
3 # Quốc gia nào có số lượng khách hàng thế nào
4 df.groupBy('Country') \
5 .agg(countDistinct('CustomerID').alias('country_count')) \
6 .orderBy(desc('country_count')) \
7 .show()

```

✓ 4 sec - Command executed in 4 sec 138 ms by ethan.chapman on 11:57:48 PM, 12/17/24

> Job execution Succeeded Spark 2 executors 8 cores [View in monitoring](#) [Open Spark UI](#)

| Country        | country_count |
|----------------|---------------|
| United Kingdom | 3950          |
| Germany        | 95            |
| France         | 87            |
| Spain          | 31            |
| Belgium        | 25            |
| Switzerland    | 21            |
| Portugal       | 19            |
| Italy          | 15            |

```

1 from pyspark.sql.functions import max
2 # Ngày có đơn hàng gần đây nhất
3 df.select(max("date")).show()

```

[17] ✓ 3 sec - Command executed in 3 sec 23 ms by ethan.chapman on 12:01:50 AM, 12/18/24

> Job execution Succeeded Spark 2 executors 8 cores [View in monitoring](#) [Open Spark UI](#)

| max(date)           |
|---------------------|
| 2011-12-09 12:50:00 |

```

1 from pyspark.sql.functions import min
2 # Ngày đầu tiên có đơn hàng
3 df.select(min("date")).show()

```

[19] ✓ 2 sec - Command executed in 1 sec 981 ms by ethan.chapman on 12:02:08 AM, 12/18/24

> Job execution Succeeded Spark 2 executors 8 cores [View in monitoring](#) [Open Spark UI](#)

| min(date)           |
|---------------------|
| 2010-12-01 08:26:00 |

### III. Tiền xử lý dữ liệu

#### 1. Recency-đo thời điểm mà khách hàng đã mua hàng lần cuối

- Là khoảng thời gian giữa thời điểm khách hàng mua hàng lần cuối và ngày đầu tiên có đơn hàng.
- Rõ ràng giá trị này càng lớn chứng tỏ khách hàng càng mua gần đây.

B1: Tạo 1 cột mới, đặt giá trị của tất cả cột đó là ngày đầu tiên có đơn hàng. Cột này có tên "from\_date".

```

1 from pyspark.sql.functions import lit
2 df = df.withColumn("from_date", lit("2010-12-01 08:26:00"))

```

✓ <1 sec - Command executed in 190 ms by ethan.chapman on 12:14:49 AM, 12/18/24

> Job execution Succeeded Spark 2 executors 8 cores [View in monitoring](#) [Open Spark](#)

View [Table](#) [Chart](#) [Export results](#)

| UnitPrice | CustomerID | Country        | date                | from_date           |
|-----------|------------|----------------|---------------------|---------------------|
| 2.55      | 17850.0    | United Kingdom | 2010-12-01 08:26:00 | 2010-12-01 08:26:00 |
| 3.39      | 17850.0    | United Kingdom | 2010-12-01 08:26:00 | 2010-12-01 08:26:00 |
| 2.75      | 17850.0    | United Kingdom | 2010-12-01 08:26:00 | 2010-12-01 08:26:00 |
| 3.39      | 17850.0    | United Kingdom | 2010-12-01 08:26:00 | 2010-12-01 08:26:00 |
| 3.39      | 17850.0    | United Kingdom | 2010-12-01 08:26:00 | 2010-12-01 08:26:00 |

B2: Lấy giá trị thời gian mua của từng đơn hàng trừ đi from\_date, ta sẽ biết đơn hàng đó được đặt cách ngày đầu tiên có đơn hàng là bao nhiêu (theo đơn vị timestamp), giá trị sẽ được lưu tại cột 'recency'.

```
1 from pyspark.sql.functions import col
2 df = df.withColumn('from_date',to_timestamp("from_date", 'yy-MM-dd HH:mm'))
3 df2 = df.withColumn('from_date',to_timestamp(col("from_date"))).withColumn('recency',col("date").cast("long") - col('from_date').cast("long"))
✓ <1 sec - Command executed in 183 ms by ethan.chapman on 12:19:06 AM, 12/18/24
```

1 display(df2)

✓ 2 sec - Command executed in 2 sec 221 ms by ethan.chapman on 12:20:18 AM, 12/18/24

> Job execution Succeeded Spark 2 executors 8 cores

View Table Chart Export results

| CustomerID | Country        | date                | from_date           | recency |
|------------|----------------|---------------------|---------------------|---------|
| 17850.0    | United Kingdom | 2010-12-01 08:28:00 | 2010-12-01 08:26:00 | 120     |
| 17850.0    | United Kingdom | 2010-12-01 08:28:00 | 2010-12-01 08:26:00 | 120     |
| 13047.0    | United Kingdom | 2010-12-01 08:34:00 | 2010-12-01 08:26:00 | 480     |

B3: Mỗi khách hàng có thể mua nhiều lần vào nhiều mốc thời gian khác nhau, chúng ta chỉ quan tâm lần cuối cùng họ mua, vì vậy cần xử lý lại cột 'recency'.

1 df2 = df2.join(df2.groupBy('CustomerID').agg(max('recency').alias('recency')),on='recency',how='leftsemi')

✓ 4 sec - Command executed in 4 sec 250 ms by ethan.chapman on 12:22:26 AM, 12/18/24

[30] > Job execution Succeeded Spark 2 executors 8 cores

View Table Chart Export results

| recency | InvoiceNo | StockCode | Description                    | Quantity |
|---------|-----------|-----------|--------------------------------|----------|
| 5220    | 536384    | 82484     | WOOD BLACK BOARD ANT WHI...    | 3        |
| 5220    | 536384    | 84755     | COLOUR GLASS T-LIGHT HOLDE...  | 48       |
| 5220    | 536384    | 22464     | HANGING METAL HEART LANTE...   | 12       |
| 5220    | 536384    | 21324     | HANGING MEDINA LANTERN S...    | 6        |
| 5220    | 536384    | 22457     | NATURAL SLATE HEART CHALKB...  | 12       |
| 5220    | 536384    | 22469     | HEART OF WICKER SMALL          | 40       |
| 5220    | 536384    | 22470     | HEART OF WICKER LARGE          | 40       |
| 5220    | 536384    | 22224     | WHITE LOVEBIRD LANTERN         | 6        |
| 5220    | 536384    | 21340     | CLASSIC METAL BIRDCAGE PLAN... | 2        |

## 2. Frequency -Đo tần suất mà khách hàng mua hàng trong một khoảng thời gian nhất định.

Khách hàng mua hàng thường xuyên được xem là có giá trị hơn so với những khách hàng mua hàng ít lần.

Phần này thì chúng ta sẽ tính tần suất một khách hàng mua một đồ gì đó. Chúng ta chỉ cần nhóm theo từng ID khách hàng và đếm số mặt hàng họ đã mua.

```

1 from pyspark.sql.functions import aggregate
2 from pyspark.sql.functions import count
3 df_freq = df2.groupby("CustomerID").agg(count("InvoiceNo").alias('frequency'))
4 display(df_freq)

```

[33] ✓ 5 sec - Command executed in 5 sec 635 ms by ethan.chapman on 12:26:10 AM, 12/18/24  
 > Job execution Succeeded Spark 2 executors 8 cores

View Table Chart Export results

| CustomerID | frequency |
|------------|-----------|
| 17786.0    | 72        |
| 15070.0    | 1         |
| 16351.0    | 8         |
| 18085.0    | 20        |
| 16718.0    | 45        |
| 17850.0    | 32        |
| 17128.0    | 14        |
| 16499.0    | 1         |

Nối nó vào dataframe

```

1 df3 = df2.join(df_freq, on="CustomerID", how="inner")

```

[34] ✓ <1 sec - Command executed in 184 ms by ethan.chapman on 12:28:15 AM, 12/18/24

### 3. Monetary-Đo giá trị đặt hàng của khách hàng.

Khách hàng đặt hàng có giá trị cao hơn được xem là có giá trị cao hơn so với những khách hàng đặt hàng có giá trị thấp.

B1: Tính số tổng tiền của một lần mua hàng.

```

1 # Tính số lượng và đơn giá của một lần mua hàng
2 m_val = df3.withColumn("TotalAmount", col("Quantity") * col("UnitPrice"))
3 display(m_val)

```

[35] ✓ 6 sec - Command executed in 5 sec 658 ms by ethan.chapman on 12:33:20 AM, 12/18/24  
 > Job execution Succeeded Spark 2 executors 8 cores

View Table Chart Export results

| Country        | date                | from_date           | frequency | TotalAmount        |
|----------------|---------------------|---------------------|-----------|--------------------|
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 19.35              |
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 31.200000000000003 |
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 19.799999999999997 |
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 17.700000000000003 |
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 35.400000000000006 |
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 58.0               |
| United Kingdom | 2010-12-01 09:53:00 | 2010-12-01 08:26:00 | 13        | 102.0              |

B2: Tính tổng số tiền mà khách hàng đã chi

```

1 from pyspark.sql.functions import sum
2 # Tính tổng số tiền mà khách hàng đã chi
3 m_val = m_val.groupBy('CustomerID').agg(sum('TotalAmount').alias('monetary_value'))

```

```

1 display(m_val)
✓ 7 sec - Command executed in 7 sec 248 ms by ethan.chapman on 12:36:54 AM, 12/18/24
> Job execution Succeeded Spark 2 executors 8 cores

```

View    **Table**    Chart    Export results

| CustomerID | monetary_value     |
|------------|--------------------|
| 15396.0    | 288.17999999999995 |
| 12535.0    | 344.90000000000003 |
| 17786.0    | 278.74             |
| 13067.0    | 115.46000000000002 |
| 13514.0    | 152.20000000000002 |
| 16083.0    | 979.6400000000001  |

### B3: Gộp dữ liệu vào DataFrame

```

1 final_df = m_val.join(df3, on='CustomerID', how='inner')
✓ <1 sec - Command executed in 415 ms by ethan.chapman on 12:38:59 AM, 12/18/24

```

| CustomerID | monetary_value | recency | InvoiceNo | StockCode |
|------------|----------------|---------|-----------|-----------|
| 18074.0    | 489.6          | 5220    | 536384    | 82484     |
| 18074.0    | 489.6          | 5220    | 536384    | 84755     |
| 18074.0    | 489.6          | 5220    | 536384    | 22464     |
| 18074.0    | 489.6          | 5220    | 536384    | 21324     |
| 18074.0    | 489.6          | 5220    | 536384    | 22457     |
| 18074.0    | 489.6          | 5220    | 536384    | 22469     |
| 18074.0    | 489.6          | 5220    | 536384    | 22470     |
| 18074.0    | 489.6          | 5220    | 536384    | 22224     |
| 18074.0    | 489.6          | 5220    | 536384    | 21340     |
| 18074.0    | 489.6          | 5220    | 536384    | 22189     |
| 18074.0    | 489.6          | 5220    | 536384    | 22427     |
| 18074.0    | 489.6          | 5220    | 536384    | 22428     |

### B4: Lấy 4 trường trong dữ liệu đã tính toán trên để xây dựng model dự đoán

```

1 final_df = final_df.select(['recency', 'frequency', 'monetary_value', 'CustomerID']).distinct()
[46] ✓ <1 sec - Command executed in 179 ms by ethan.chapman on 12:43:35 AM, 12/18/24

```

▶ [47]

```

1 display(final_df)
[47] ✓ 11 sec - Command executed in 10 sec 991 ms by ethan.chapman on 12:44:04 AM, 12/18/24
 > Job execution Succeeded Spark 2 executors 8 cores

```

View Table Chart Export results ▾

| recency | frequency | monetary_value    | CustomerID |
|---------|-----------|-------------------|------------|
| 4860540 | 72        | 278.74            | 17786.0    |
| 97020   | 1         | 106.2             | 15070.0    |
| 3549600 | 8         | 153.9             | 16351.0    |
| 3739860 | 20        | 386.0499999999995 | 18085.0    |
| 362220  | 45        | 623.7500000000002 | 16718.0    |
| 91080   | 32        | 126.38            | 17850.0    |

#### IV. Chuẩn hóa dữ liệu

Mỗi trường dữ liệu có một đơn vị khác nhau, nếu không chuẩn hóa thì chắc chắn sẽ nảy sinh nhiều vấn đề về sau.

```

1 from pyspark.ml.feature import VectorAssembler
2 from pyspark.ml.feature import StandardScaler
3
4 assemble=VectorAssembler(inputCols=[
5 'recency','frequency','monetary_value'
6], outputCol='features')
7
8 assembled_data=assemble.transform(final_df)
9
10 scale=StandardScaler(inputCol='features',outputCol='standardized')
11 data_scale=scale.fit(assembled_data)
12 data_scale_output=data_scale.transform(assembled_data)
13

```

✓ 13 sec - Command executed in 13 sec 64 ms by ethan.chapman on 12:52:36 AM, 12/18/24

▶ [50]

```

1 data_scale_output.select('standardized').show(2, truncate=False)
[50] ✓ 9 sec - Command executed in 8 sec 922 ms by ethan.chapman on 12:54:54 AM, 12/18/24
 > Job execution Succeeded Spark 2 executors 8 cores

```

... +-----+  
| standardized |  
+-----+  
|[ 0.9623273639691915, 0.2323775559933276, 0.14005912310985658 ] |  
|[ 0.6272129860810431, 0.43893538354295214, 0.11962548724953012 ] |  
+-----+  
only showing top 2 rows

## V. Triển khai học máy

### 1. Lựa chọn số簇 cụm

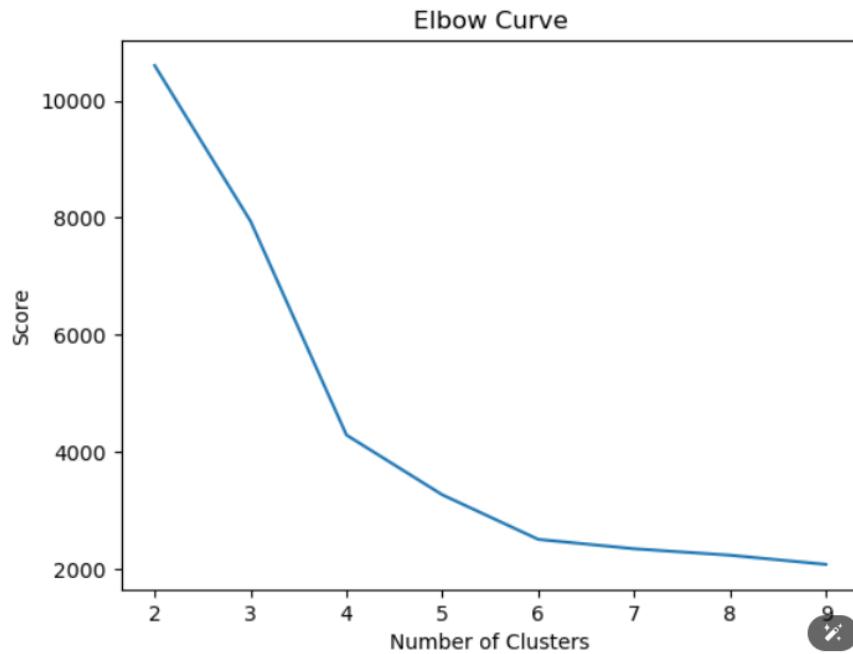
Chạy K-means với nhiều簇 và trực quan hóa kết quả, tìm ra điểm uốn giống như khủy tay và lựa chọn điểm này. ưu tiên k càng nhỏ càng tốt

```
1 from pyspark.ml.clustering import KMeans
2 from pyspark.ml.evaluation import ClusteringEvaluator
3 import numpy as np
4
5 cost = np.zeros(10)
6 evaluator = ClusteringEvaluator(predictionCol='prediction', featuresCol='standardized', metricName='silhouette', distanceMeasure='squaredEuclidean')
7
8 for i in range(2,10):
9 KMeans_algo=KMeans(featuresCol='standardized', k=i)
10 KMeans_fit=KMeans_algo.fit(data_scale_output)
11 output=KMeans_fit.transform(data_scale_output)
12 cost[i] = KMeans_fit.summary.trainingCost
13
```

✓ 3 min 15 sec - Command executed in 3 min 15 sec 284 ms by ethan.chapman on 1:03:47 AM, 12/18/24

```
1 import pandas as pd
2 import pylab as pl
3 df_cost = pd.DataFrame(cost[2:])
4 df_cost.columns = ["cost"]
5 new_col = range(2,10)
6 df_cost.insert(0, 'cluster', new_col)
7 pl.plot(df_cost.cluster, df_cost.cost)
8 pl.xlabel('Number of Clusters')
9 pl.ylabel('Score')
10 pl.title('Elbow Curve')
11 pl.show()
12
```

✓ <1 sec - Command executed in 660 ms by ethan.chapman on 1:04:12 AM, 12/18/24



sẽ thấy được điểm k cần tìm ở đây là 3

### 2. Triển khai số簇 cụm với k=3

B1: Training K-means

```

1 #train
2 kmeans_algo=KMeans(featuresCol='standardized', k=3)
3 kmeans_fit=kmeans_algo.fit(data_scale_output)
4

✓ 24 sec - Command executed in 24 sec 154 ms by ethan.chapman on 1:07:41 AM, 12/18/24

```

## B2: Dự đoán

```

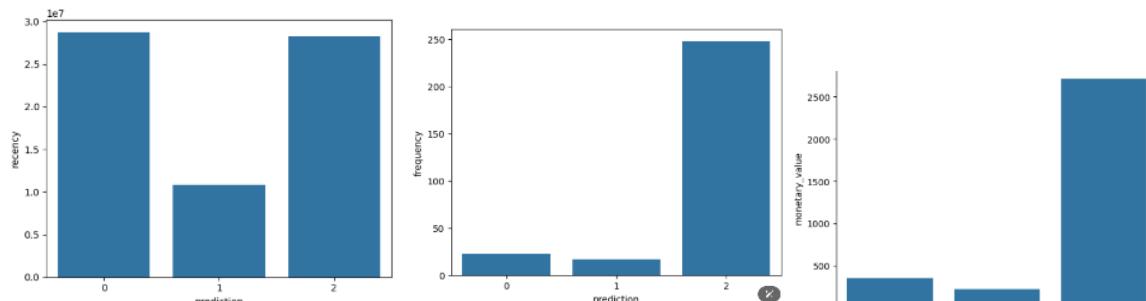
1 preds=kmeans_fit.transform(data_scale_output)
2 preds.show(5)

✓ 7 sec - Command executed in 7 sec 155 ms by ethan.chapman on 1:08:00 AM, 12/18/24

```

|                                                                                                  | recency frequency | monetary_value CustomerID | features | standardized prediction |
|--------------------------------------------------------------------------------------------------|-------------------|---------------------------|----------|-------------------------|
| 18170160   1   204.0   16553.0   [1.817016E7, 1.0, 2...   [2.08327580864095...   1               |                   |                           |          |                         |
| 16952220   3   76.5   17536.0   [1.695222E7, 3.0, 7...   [1.94363449902253...   1                |                   |                           |          |                         |
| 19549920   29   187.91999999999996   14722.0   [1.954992E7, 29.0, ...   [2.24147037763376...   1 |                   |                           |          |                         |
| 20747940   65   299.31   13827.0   [2.074794E7, 65.0, ...   [2.37882778583864...   0             |                   |                           |          |                         |
| 21113580   4   870.0   17353.0   [2.111358E7, 4.0, 8...   [2.42074975937500...   0               |                   |                           |          |                         |

## B3: sử dụng matplotlib để trực quan hóa phân khúc khách hàng



Nhóm 0: Nhóm này có tần suất mua hàng tương đối cao, cao trội hơn 3 nhóm còn lại, lần truy cập gần nhất cũng tương đối cao, giá trị tiền mua hàng tương đối nhỏ, cho thấy là một đối tượng đa số là cá nhân, hướng tới các sản phẩm giá rẻ

Nhóm 1: Nhóm này có cả 3 chỉ số lần truy cập gần nhất, tần suất mua hàng và tổng tiền mua hàng rất thấp, không có quá nhiều hi vọng là khách hàng tiềm năng, khả năng cao sẽ ngừng mua hàng trong thời gian tới.

Nhóm 2: Tần suất đặt hàng rất cao, gần đây lại đặt phổ biến, lượng tiền mua hàng cao vượt trội nhiều lần so với các nhóm khác. Nhóm này khả năng là các doanh nghiệp có xu hướng mua các loại hàng có giá trị cao hoặc mua có số lượng lớn.

## PHẦN 11: CONCLUSION

Azure Synapse Analytics là một nền tảng mạnh mẽ, kết hợp giữa khả năng xử lý dữ liệu lớn và phân tích dữ liệu trong một hệ sinh thái thống nhất. Với các tính năng nổi bật như

tích hợp đa dạng, hiệu suất tối ưu, bảo mật hàng đầu, và hỗ trợ nhiều công cụ phát triển, Synapse đã chứng minh giá trị to lớn trong việc đáp ứng các nhu cầu phân tích hiện đại.

Việc áp dụng Azure Synapse Analytics mang lại nhiều lợi ích quan trọng cho doanh nghiệp, từ khả năng phân tích thời gian thực, quản lý và trực quan hóa dữ liệu hiệu quả, đến khả năng tích hợp với các công cụ học máy và trực quan hóa như Power BI. Điều này không chỉ hỗ trợ đưa ra các quyết định dựa trên dữ liệu mà còn tối ưu hóa các quy trình kinh doanh.

Nhìn về tương lai, Azure Synapse Analytics sẽ tiếp tục phát triển với các công nghệ mới như Gen3, mở rộng hơn nữa khả năng phân tích và hiệu suất. Do đó, việc triển khai và ứng dụng Synapse trong các tổ chức không chỉ là một chiến lược ngắn hạn mà còn là một bước đi lâu dài hướng tới sự phát triển bền vững dựa trên dữ liệu.