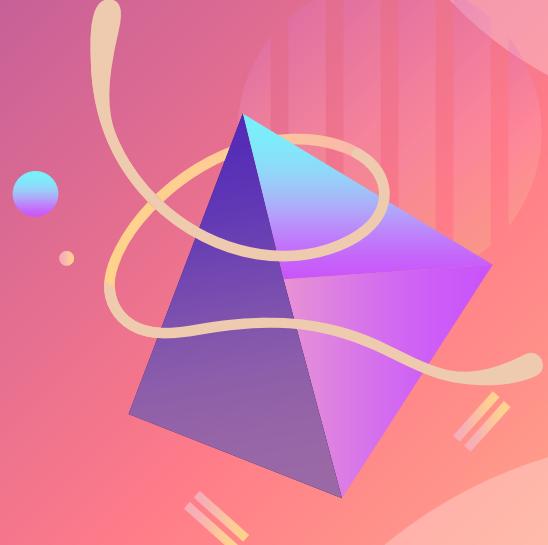


Azure Synapse Analysis Studio Overview

TS. Nguyễn Chí Kiên

Nhóm 1





Thành viên:

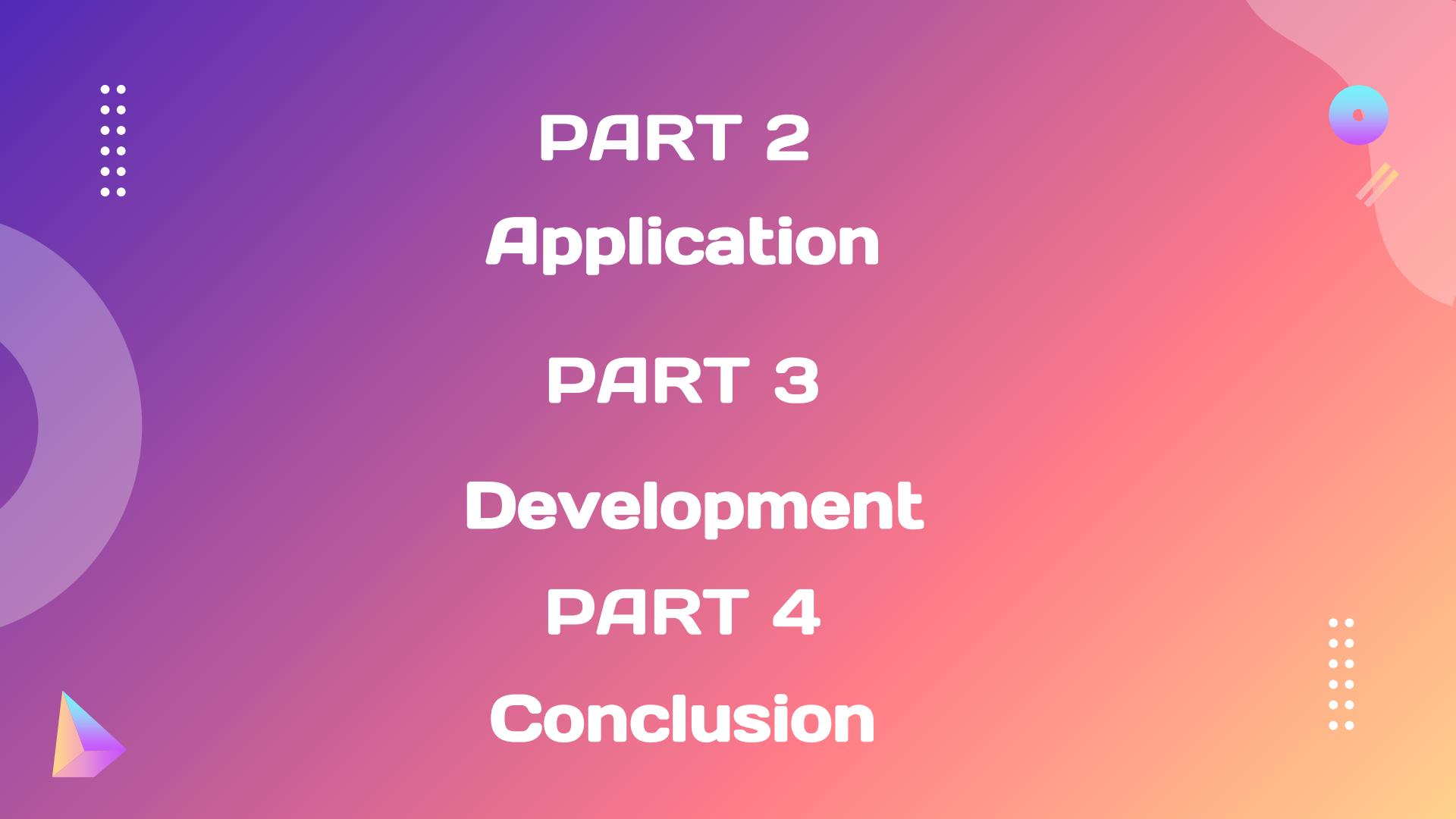
**Trần Thu Huyền – 21106211
Nguyễn Thanh Thùy Trang - 21084151**



TABLE OF CONTENTS

PART 1

I	INTRODUCTION	V	AZURE SYNAPSE ANALYTICS SQL ANALYTICS
II	AZURE SYNAPSE ANALYTICS MPP INTRO	VI	SQL ON DEMAND
III	AZURE SYNAPSE ANALYTICS STUDIO	VII	AZURE SYNAPSE ANALYTICS SPARK
IV	AZURE SYNAPSE ANALYTICS DATA INTEGRATION	VIII	INDUSTRY-LEADING SECURITY AND COMPLIANCE



PART 2

Application

PART 3

Development

PART 4

Conclusion





PART 1

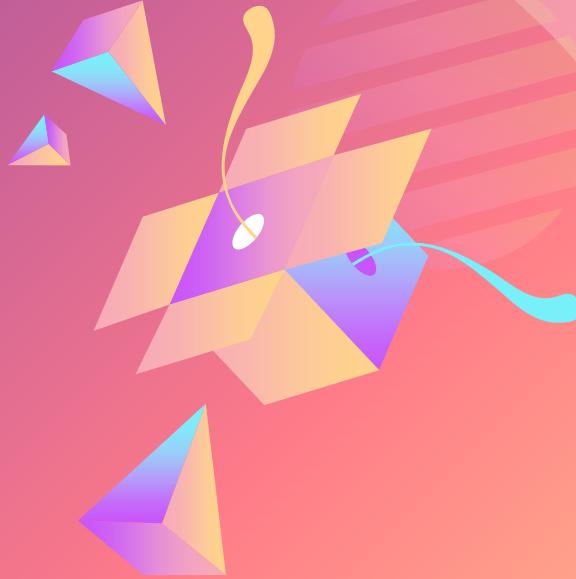
Tổng quan về Azure Synapse Analysis Studio



O1

INTRODUCTION

Giới thiệu về azure synapse analytics



I. Introduction

01

Azure Synapse Analytics là gì?

02

Mục đích chính

03

Điểm nổi bật

04

Kiến trúc Modern Data Warehouse

05

Kiến trúc Azure Synapse Analytics - Data Lakehouse

06

Các giai đoạn phát triển của Azure Synapse Analytics

07

Thành phần của Azure Synapse Analytics



O1 | Azure Synapse Analytics là gì?

Azure Synapse Analytics là dịch vụ phân tích tích hợp, kết hợp kho dữ liệu doanh nghiệp và phân tích Big Data. Nó cho phép truy vấn dữ liệu linh hoạt với tài nguyên serverless hoặc dedicated, hỗ trợ tiếp nhận, xử lý, quản lý và cung cấp dữ liệu trên một nền tảng thống nhất, phục vụ thông tin doanh nghiệp và học máy.

O2 | 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





03

Điểm nổi bật



Giá tốt nhất theo
hiệu suất tốt nhất
**Best in class price
per performance**

Bảo mật hàng
đầu ngành
**Industry-leading
security**

Thực thi truy vấn nhân
thức theo khối lượng
**Workload aware
query execution**

Tính linh hoạt của
dữ liệu
Data flexibility

Năng suất phát
triển
**Developer
productivity**



- Tối ưu hóa tài nguyên
- Tính toán linh hoạt



- Mã hóa dữ liệu
- Quản lý khóa
- Xác thực và quyền truy cập



- Tối ưu hóa truy vấn
- Phân bổ tài nguyên thông minh



- Hỗ trợ nhiều định dạng dữ liệu
- Tích hợp dễ dàng



- Công cụ phát triển mạnh mẽ
- Tích hợp CI/CD
- Hỗ trợ nhiều ngôn ngữ lập trình

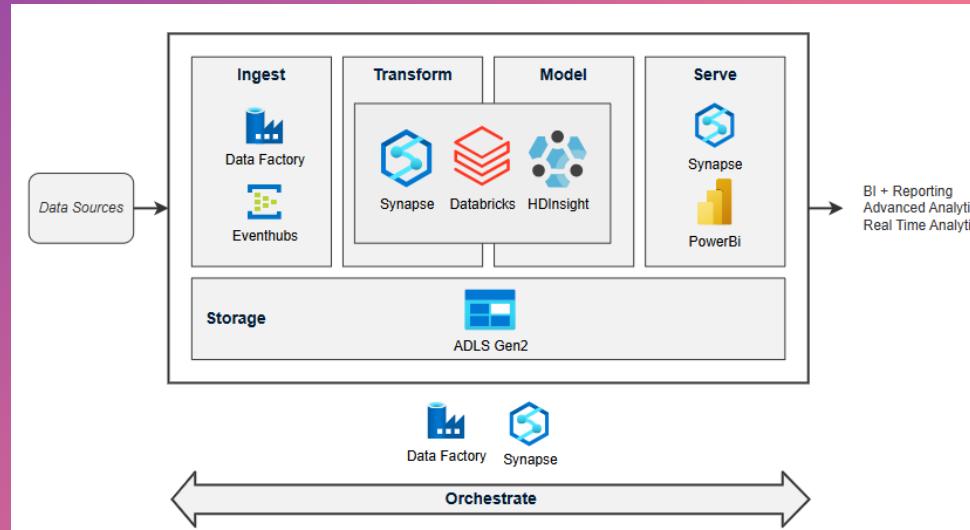


O4

Kiến trúc Modern Data Warehouse



- Thu thập dữ liệu từ nhiều nguồn (on-premises, cloud, SaaS)
- Xử lý qua nhiều bước riêng biệt và lưu trữ dữ liệu trong Azure Data Lake Storage





05 | Kiến trúc Azure Synapse Analytics - Data Lakehouse

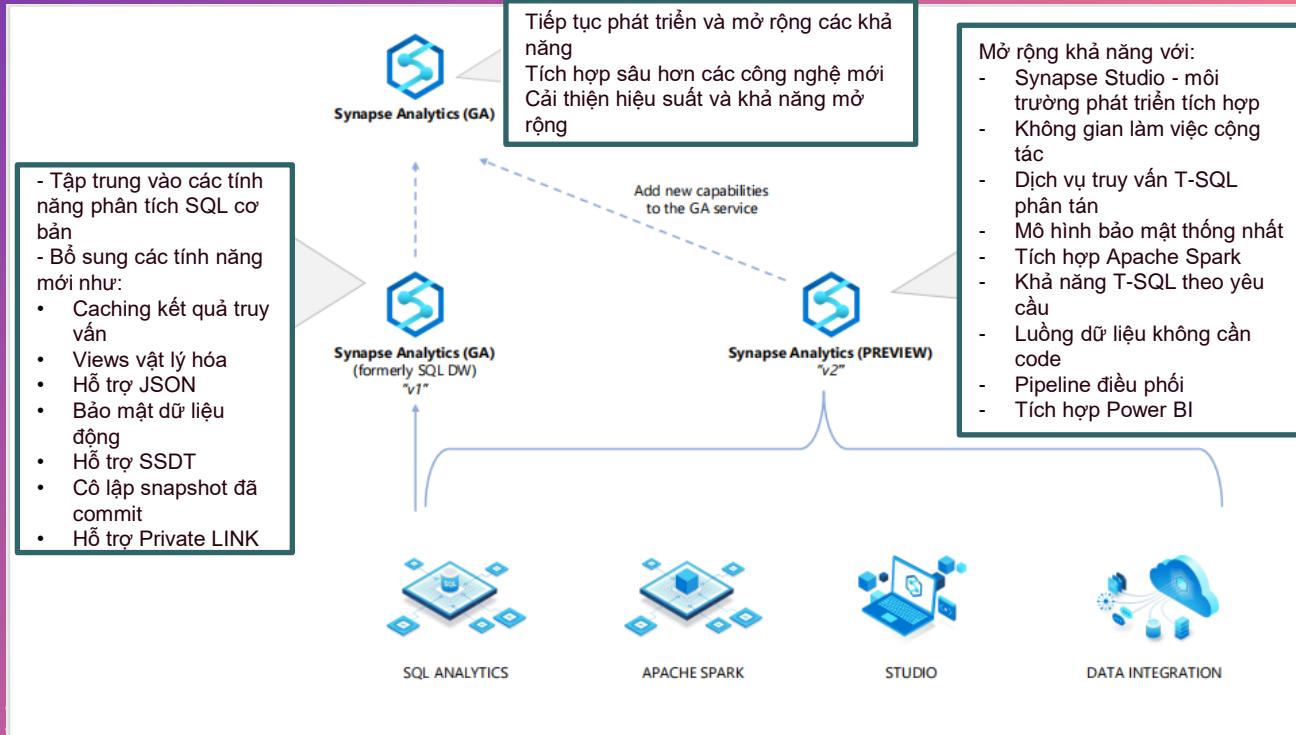


- Azure Synapse Analytics là một nền tảng tích hợp mạnh mẽ, kết hợp giữa kho dữ liệu truyền thống (Data Warehouse) và hồ dữ liệu (Data Lake), tạo thành kiến trúc Data Lakehouse
- Các Thành Phần Chính trong Kiến Trúc Data Lakehouse
 - Azure Data Lake
 - Azure Synapse Analytics
 - Data Integration
 - Data Warehouse
 - Analytics & Machine Learning.



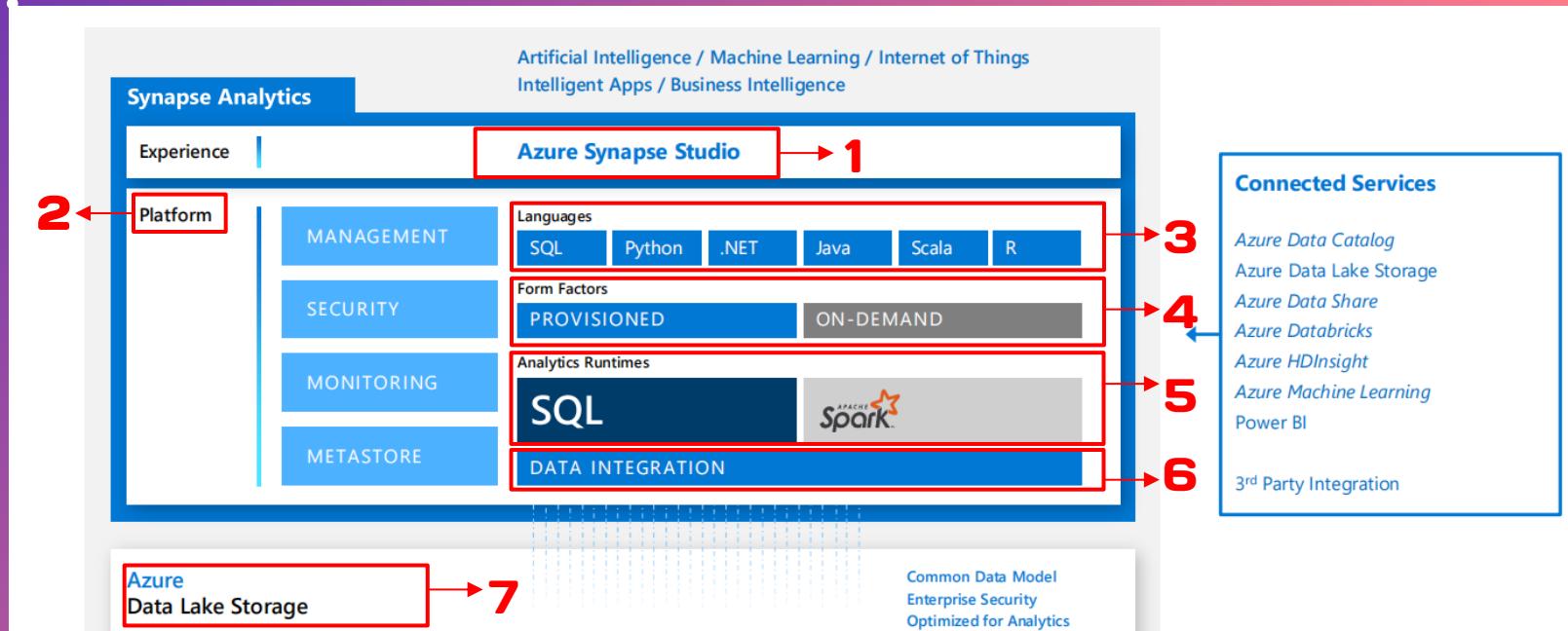
06

Các giai đoạn phát triển của Azure Synapse Analytics



O7

Thành phần của Azure Synapse Analytics



02

AZURE SYNAPSE ANALYTICS MPP INTRO





II. AZURE SYNAPSE ANALYTICS MPP INTRO

01

Kiến trúc của Azure Synapse Analytics

02

Kiến trúc DW100 và DW600

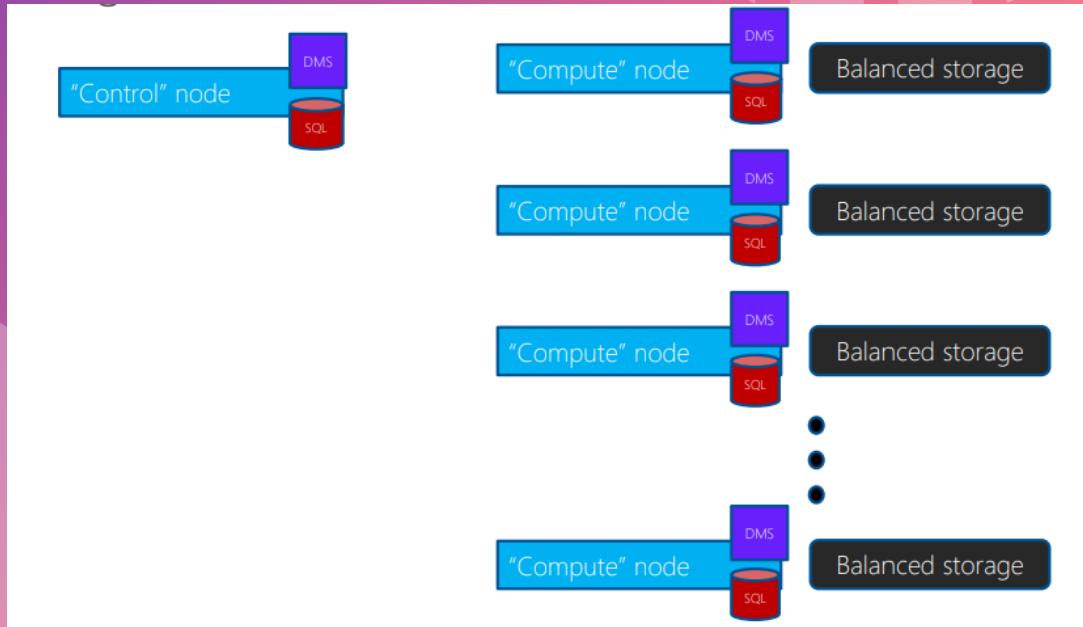




O1

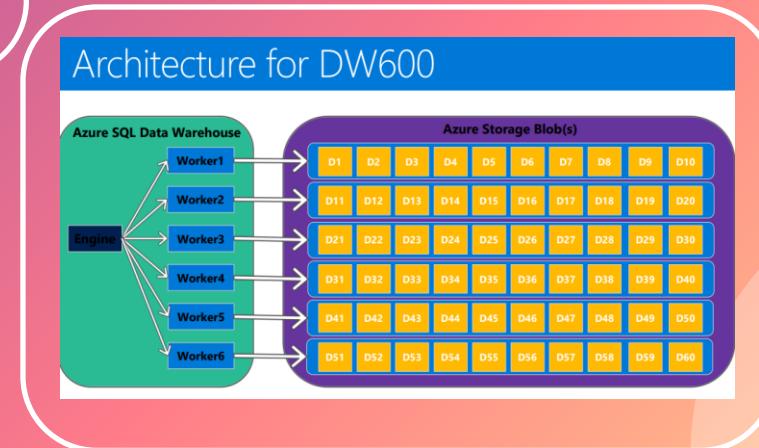
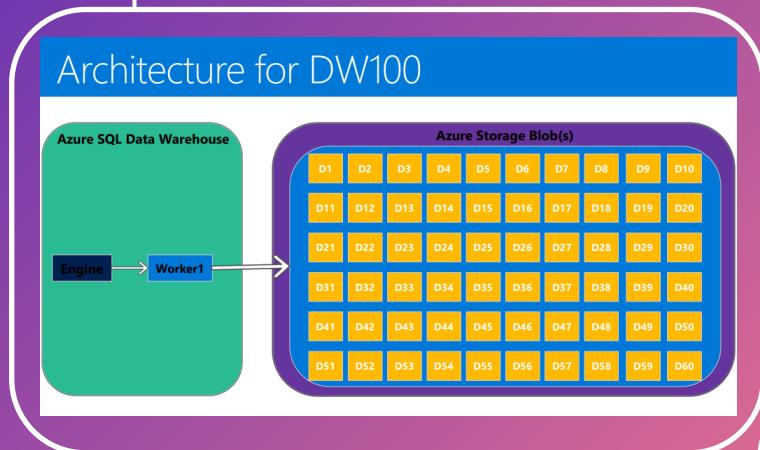
Kiến trúc của Azure Synapse Analytics

- Azure Synapse Analytics (trước đây là SQL Data warehouse) sử dụng kiến trúc Massively Parallel Processing (MPP), cho phép nhiều nút xử lý song song để tăng tốc độ và hiệu suất phân tích dữ liệu lớn.



• 02

Kiến trúc DW100 và DW600



03

AZURE SYNAPSE ANALYTICS STUDIO





III. AZURE SYNAPSE ANALYTICS STUDIO

01

Create workspace

02

Synapse Studio Overview hub

03

Synapse Studio Data hub

04

Synapse Studio Develop hub

05

Synapse Studio Integrate hub

06

Synapse Studio Monitor hub

07

Synapse Studio Manage hub

O1 | Create workspace

The screenshot shows the Microsoft Azure Synapse Analytics workspace creation interface. At the top left, there's a smaller window titled "tranhuyen-synapse" showing the "Overview" tab with a "Getting started" section containing a button to "Open Synapse Studio". A red diagonal line with an arrow points from this button down to the main workspace creation screen below.

The main screen has a blue header bar with the text "Microsoft Azure | Synapse Analytics > tranhuyen-synapse" and the user email "ethan.chapman@canberrahouse.co.uk" and "DEFAULT DIRECTORY".

The workspace name "tranhuyen-synapse" is displayed prominently in the center. Below it is a "New" button with a dropdown arrow.

On the left side, there's a vertical sidebar with icons for Home, Databases, Tables, Functions, Pipelines, and Storage.

The central area contains three main cards:

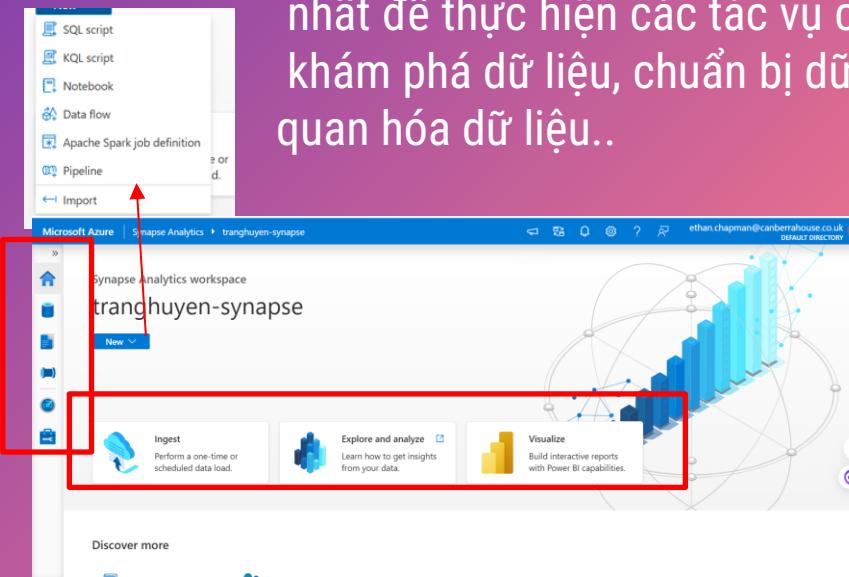
- Ingest**: Perform a one-time or scheduled data load.
- Explore and analyze**: Learn how to get insights from your data.
- Visualize**: Build interactive reports with Power BI capabilities.

At the bottom, there's a "Discover more" section with a "Data Lake" button.

O2

Synapse Studio Overview hub

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).

O3 | 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.

The screenshot shows the Microsoft Azure Synapse Studio interface. On the left, there's a navigation pane with 'Data' selected, showing 'Workspace' and 'Linked' tabs. Under 'Linked', several data sources are listed: 'Azure Blob Storage' (1 item), 'Sample Datasets' (1 item, containing 'nyc_tlc_green'), 'Azure Data Lake Storage Gen2' (2 items, containing 'tranhuyen-synapse (Primary - tra...' and 'demo (Primary)'), and 'Integration datasets' (1 item, containing 'SqlPoolTable1'). A red box highlights this list of linked datasets. To the right, a specific dataset named 'SqlPoolTable1' is selected from an 'Azure Synapse dedicated SQL pool'. The 'Connection' tab is active, showing 'SQL pool' dropdown set to 'SQLPool' (with a green checkmark) and 'Table' dropdown set to 'dbo.NYCTaxiGreen'. There are also 'Schema' and 'Parameters' tabs, along with 'Refresh' and 'Preview data' buttons.

03 | Synapse Studio Data hub

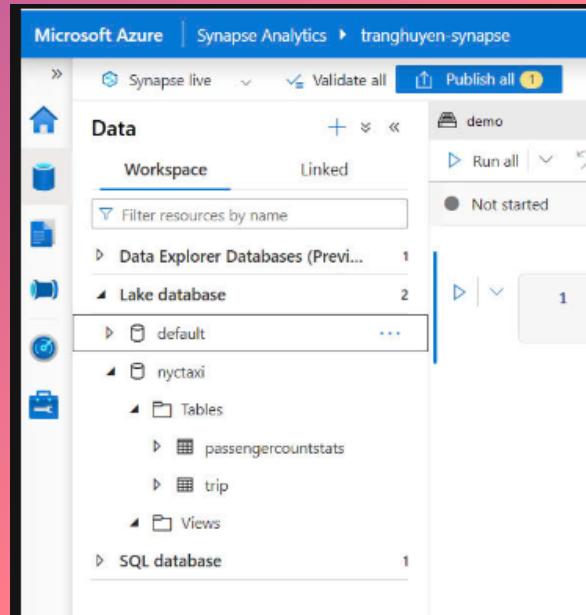
3. Workspace

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.



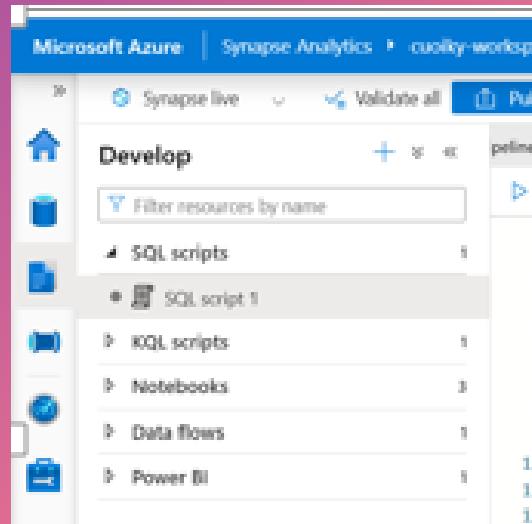


O4 | Synapse Studio Develop hub



1. Overview: Develop Hub là nền tảng hỗ trợ: Viết và thực thi truy vấn để lấy dữ liệu từ các nguồn khác nhau, cung cấp công cụ phân tích dữ liệu như biểu đồ, đồ thị và thống kê, hỗ trợ xây dựng mô hình dữ liệu và học máy để dự đoán và ra quyết định.

- Các công cụ và tài nguyên hỗ trợ quá trình phát triển và xử lý dữ liệu:





O4 | Synapse Studio Develop hub



2. SQL scripts

- Authoring SQL Scripts: Tạo và chỉnh sửa script SQL với hỗ trợ IntelliSense và kiểm tra lỗi cú pháp.
- Execute SQL Script: Thực thi script trên SQL Pool hoặc SQL On-demand.
- Publish SQL Scripts: Xuất bản các script lên môi trường sản xuất hoặc chia sẻ với nhóm.
- Language Support & IntelliSense: Hỗ trợ nhiều ngôn ngữ và gợi ý cú pháp giúp viết mã nhanh và chính xác.



O4 | Synapse Studio Develop hub

2. SQL scripts

B1: khởi động lại SQL pool

B2: Trong develop hub, chọn nút + sau đó chọn create new SQL script. -> connect với SQL pool

B3: ví dụ nhập code tạo bảng NYCTaxiGreen và chọn run

The screenshot shows the Microsoft Azure Synapse Analytics interface. In the top navigation bar, it says "Microsoft Azure | Synapse Analytics > tranghuyen-synapse". Below the navigation bar, there's a "Develop" section with a "SQL scripts" button highlighted. A dropdown menu is open under "SQL scripts", showing options like "SQL script", "KQL script", "Notebook", "Data flow", "Apache Spark job definition", "Browse gallery", and "Import".

This screenshot shows the "Develop" blade in Synapse Studio. On the left, there's a "Properties" panel with tabs for "General" and "Relational". Under "General", the "Name" field is set to "SQL script 1". The "Connect to" dropdown is set to "Built-in". The main area shows a list of resources: "SQL scripts" (2), "SQL script 1", and "SQLpoolbgserverless".

This screenshot shows the Microsoft Azure Synapse Analytics studio. The code editor contains the following SQL script:

```
1 IF NOT EXISTS (
2     SELECT * FROM sys.objects O JOIN sys.schemas S
3     WHERE O.name = 'NYCTaxiGreen'
4     AND O.type = 'U' AND S.name = 'dbo')
5     CREATE TABLE dbo.NYCTaxiGreen
6     (
7         [VendorID] bigint,
8         [store_and_fwd_flag] nvarchar(1) NULL,
9         [RatecodeID] float NULL,
10        [PULocationID] bigint NULL,
11        [DOLocationID] bigint NULL,
12        [trip_distance] float NULL,
13        [Fare_amount] float NULL,
14        [extra] float NULL,
15        [mta_tax] float NULL,
16        [tip_amount] float NULL,
```

The "Properties" panel on the right shows the script is named "SQL script 1" and has a type of "sql script". The status bar at the bottom indicates "00:00:01 Query completed with errors."



O4 | Synapse Studio Develop hub



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

Một số tính năng chính:

- Hỗ trợ đa ngôn ngữ: Viết mã trong nhiều ngôn ngữ khác nhau trong cùng một notebook.
- Bảng tạm thời: Tạo và sử dụng bảng tạm thời giữa các ngôn ngữ.
- Tính năng mã nâng cao: Tô sáng cú pháp, phát hiện lỗi, hoàn thành mã, thuần hóa và gấp mã.
- Xuất kết quả: Xuất kết quả phân tích và tính toán.

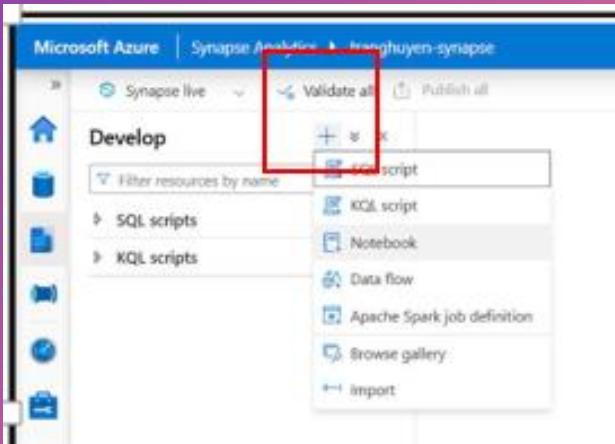


O4 | Synapse Studio Develop hub



4. Notebooks

B1: Tạo một note book bằng cách vào develop click vào dấu + và chọn Notebook



B2: Tại Attach to chọn SparkPool và tại language là python

B3: Tạo mới một cell code và Nhập code và chọn run

The screenshot shows the Microsoft Azure Synapse Studio Notebook interface. The 'Attach to' dropdown is set to 'SparkPool'. The 'Language' dropdown is set to 'PySpark (Python)'. The notebook content displays the following Python code:

```
spark.read.load('abfss://dem@tranhuyenadl2.dfs.core.windows.net/synapse/NYCtaxiGreen.parquet')
    .cache()
    .limit(100)
```

The notebook has run successfully, indicated by the green checkmark icon.

VendorID	lpep_pickup_datetime	lpep_dropoff_datetime	store_and_fwd_flag
2	2024-01-01 001155	2024-01-01 002420	N
2	2024-01-31 235922	2024-01-31 232714	N
2	2024-01-01 003029	2024-01-01 003532	N
2	2024-01-31 235642	2024-02-01 000653	N
2	2024-02-01 003114	2024-02-01 003116	N
2	2024-02-01 000923	2024-02-01 001010	N
2	2024-02-01 003022	2024-02-01 003343	N



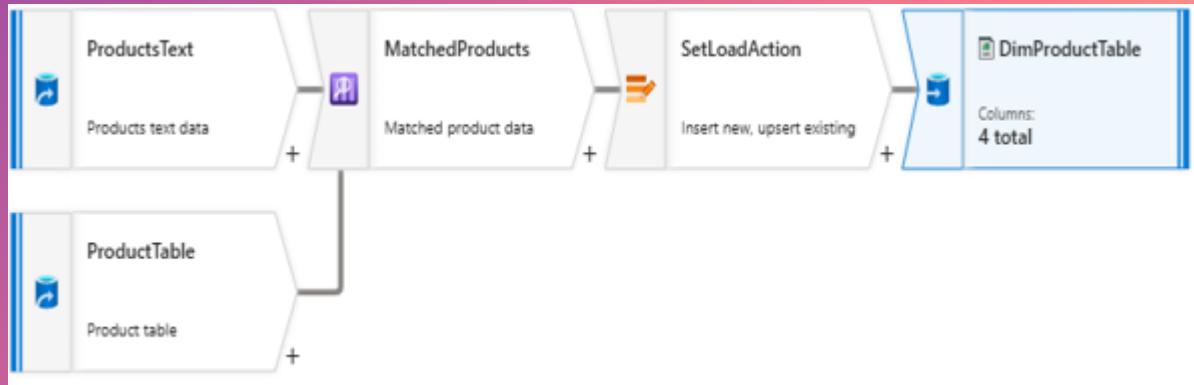
O4 | Synapse Studio Develop hub



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.

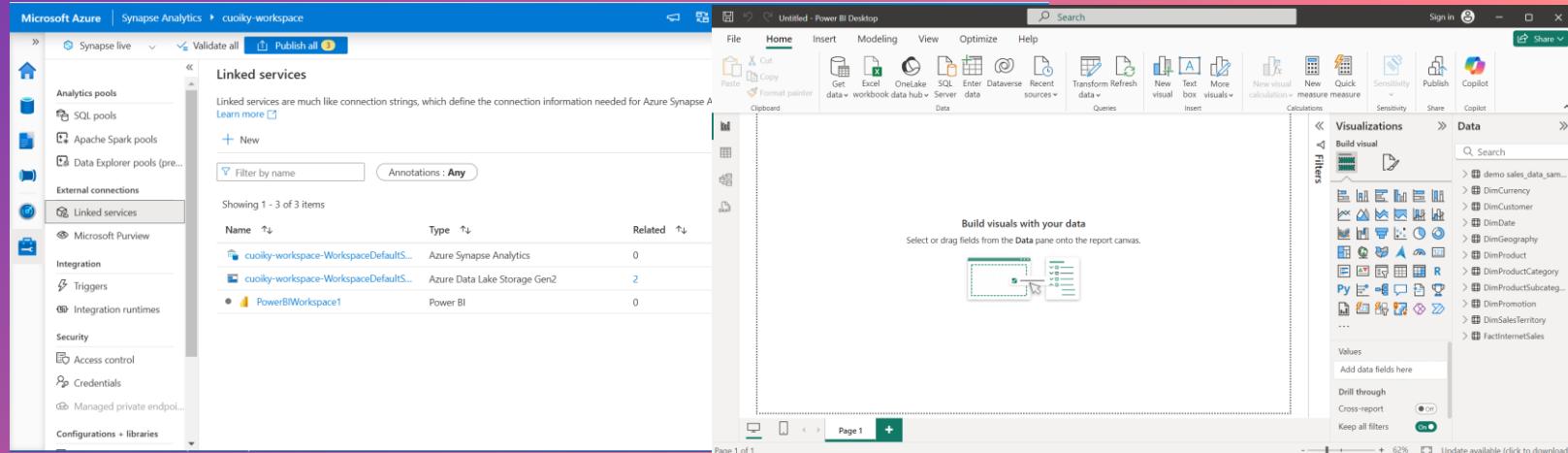
Data flow trong hệ thống xử lý dữ liệu hỗ trợ các tính năng như xử lý upserts, updates, deletes trên SQL sinks, phân vùng dữ liệu, xử lý schema drift, quản lý file, thêm hàm mới, tích hợp các mẫu ETL phổ biến, theo dõi nguồn gốc dữ liệu (data lineage) và sử dụng các template ETL như SCD Type 1, Type 2.



04 | Synapse Studio Develop hub

6. Power BI

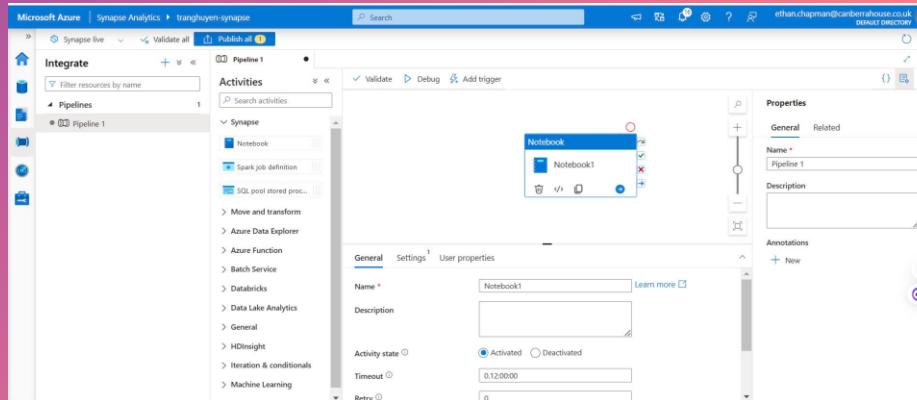
Azure Synapse Analytics cung cấp tích hợp mạnh mẽ với Power BI, cho phép tạo và quản lý báo cáo trực tiếp từ workspace. Tính năng này hỗ trợ phân tích dữ liệu trực quan và kết nối liền mạch giữa hai nền tảng.



05 | Synapse Studio Integrate hub

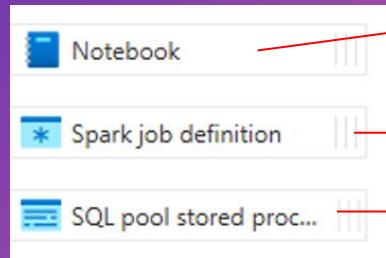
Azure Data Factory cung cấp khả năng tạo pipeline để thu thập, chuyển đổi và tải dữ liệu. Nó hỗ trợ hơn 90 trình kết nối tích hợp sẵn và các hoạt động như:

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





05 | Synapse Studio Integrate hub



Thực thi mã PySpark, Scala, SQL để phân tích hoặc xử lý dữ liệu.

Kích hoạt công việc Spark để xử lý dữ liệu lớn.

Thực thi stored procedure để xử lý dữ liệu trong SQL pool.

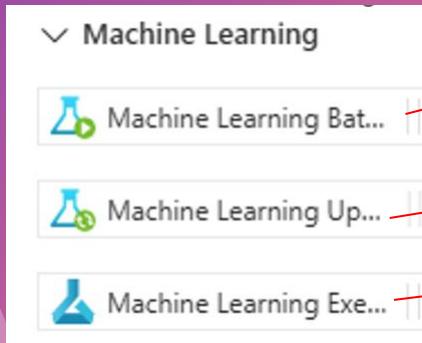
Tạo bản sao dữ liệu mà không thay đổi nguồn gốc

Move and transform

Copy data

Data flow

Di chuyển dữ liệu liên tục theo thời gian thực hoặc gần thời gian thực



Xử lý dữ liệu và huấn luyện mô hình trên lượng lớn dữ liệu theo lô.

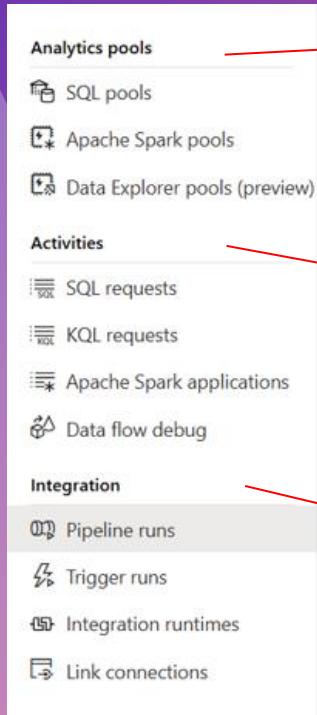
Cập nhật và cải tiến các mô hình học máy hiện tại.

Thực thi các mô hình học máy, như huấn luyện và dự đoán.

06

Synapse Studio Monitor hub

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



- 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.

- 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.

- 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.

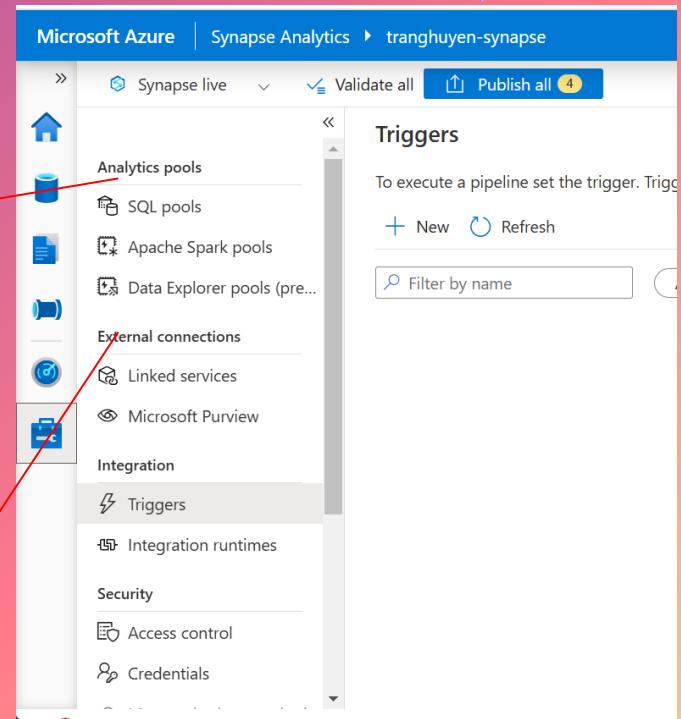
O7

Synapse Studio Manage hub

1. Overview

- SQL Pools: Quản lý cơ sở dữ liệu phân tích SQL.
- Apache Spark Pools: Quản lý cụm Spark để xử lý dữ liệu lớn.
- Data Explorer Pools: Quản lý pool dùng Data Explorer cho truy vấn thời gian thực.

- Linked Services: Azure Blob Storage, SQL Database, PowerBI,...
- Microsoft Purview: Giám sát và quản lý dữ liệu toàn bộ hệ sinh thái.



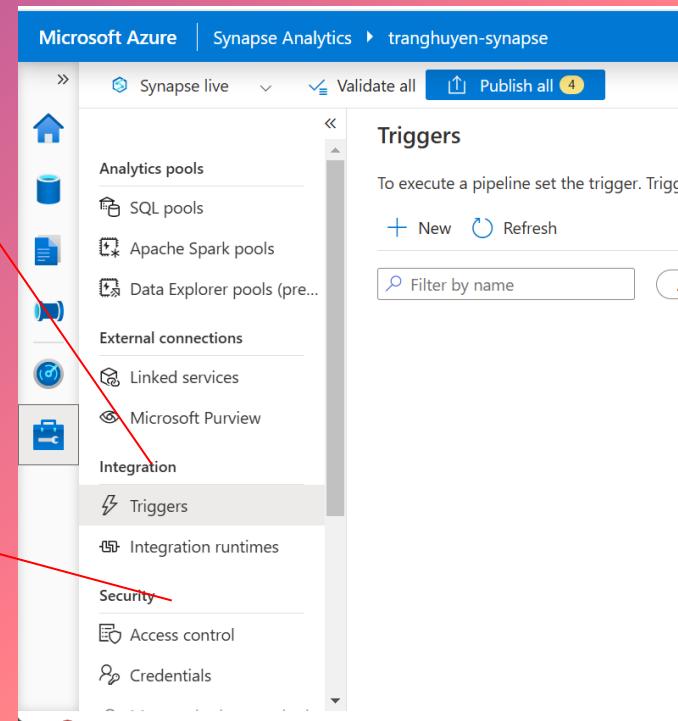
O7

Synapse Studio Manage hub

1. Overview

- Triggers: Giám sát, tạm dừng, khởi động lại, hoặc hủy bỏ quá trình thực thi pipeline.
- Integration Runtimes: Quản lý môi trường chạy dữ liệu, hỗ trợ tích hợp dữ liệu giữa các môi trường mạng khác nhau.

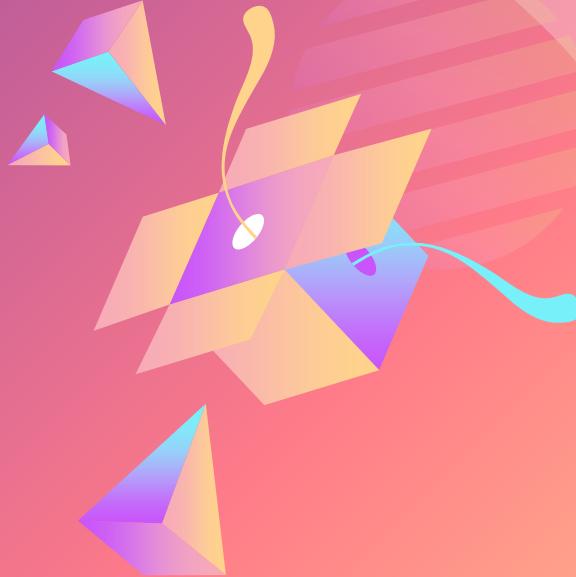
- Access Control: Quản lý quyền truy cập người dùng.
- Credentials: Quản lý thông tin xác thực.





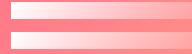
04

AZURE SYNAPSE ANALYTICS DATA INTEGRATION





IV. AZURE SYNAPSE ANALYTICS DATA INTEGRATION



01

Data integration Overview

02

Data Movement

03

Pipelines

04

Move & Transform Data

05

Triggers

06

Manage – Linked Services

07

Manage – Integration runtime.

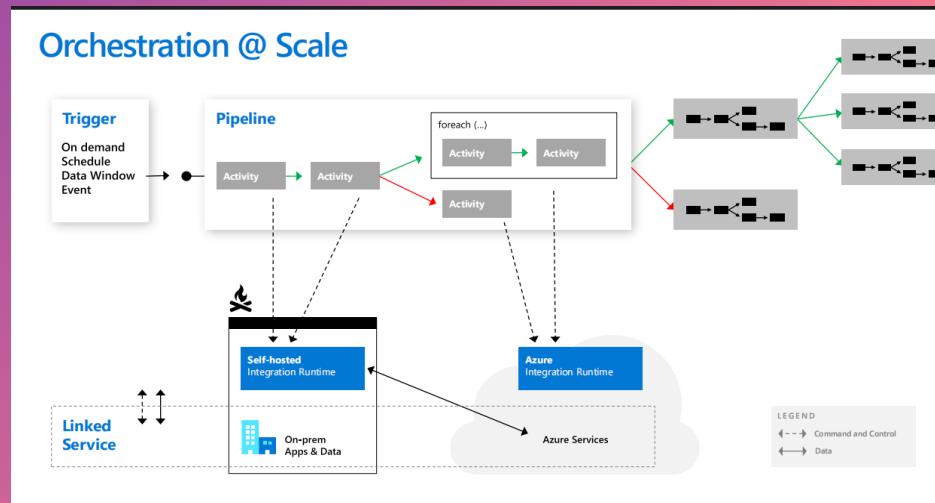




O1 | Synapse Studio Develop hub



Azure Data Factory (ADF) là dịch vụ đám mây của Microsoft Azure giúp tích hợp, chuyển đổi và di chuyển dữ liệu từ nhiều nguồn khác nhau (tại chỗ và đám mây). ADF hoạt động như một trung tâm điều phối, tự động hóa quy trình thu thập, chuẩn bị và xử lý dữ liệu, phục vụ mục tiêu phân tích và báo cáo của tổ chức.

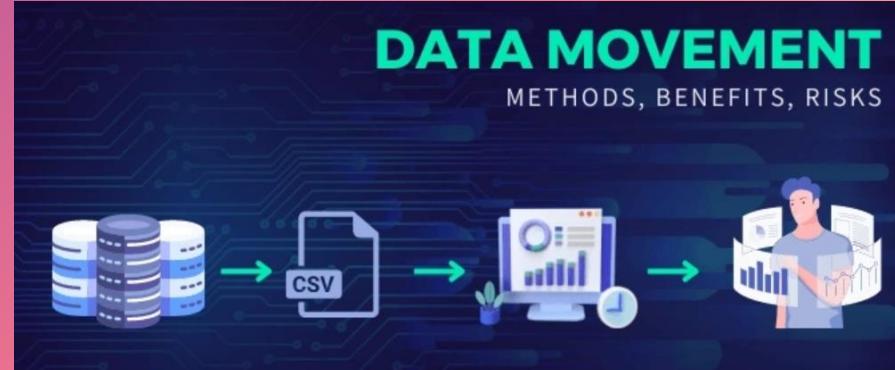




02 | Data Movement

Data Movement trong Azure Synapse Analytics cung cấp khả năng di chuyển dữ liệu giữa các hệ thống với tính năng mạnh mẽ và linh hoạt:

- Scalable: Tăng giảm tài nguyên xử lý theo khối lượng công việc, đạt tốc độ di chuyển lên tới 4 GB/s.
- Simple: Dễ dàng tạo và quản lý qua giao diện đồ họa hoặc mã (Python, .Net).
- Serverless: Azure tự động mở rộng và quản lý tài nguyên.
- Truy cập dữ liệu toàn diện: Hỗ trợ hơn 90 kết nối từ các nguồn đám mây
- Global Reach: hỗ trợ di chuyển dữ liệu giữa đám mây và hệ thống nội bộ (hybrid).





03

Pipelines



Copy data từ Azure data lake storage gen 2 vào database SQL để thực hiện transform chuyển đổi, xử lý dữ liệu

The screenshot shows the Microsoft Azure Synapse Analytics Pipelines interface. On the left, the navigation pane is open under the 'Integrate' section, specifically in the 'Pipelines' category. A pipeline named 'Load Product Data' is selected. The main workspace displays a 'Data flow' component with a single activity named 'LoadProducts'. The properties pane on the right is set to the 'General' tab, showing the pipeline's name as 'Load Product Data' and its description as empty. The 'Settings' tab is also visible, containing options like 'Logging level' (set to 'Verbose'), 'Sink properties', and 'Staging' settings.

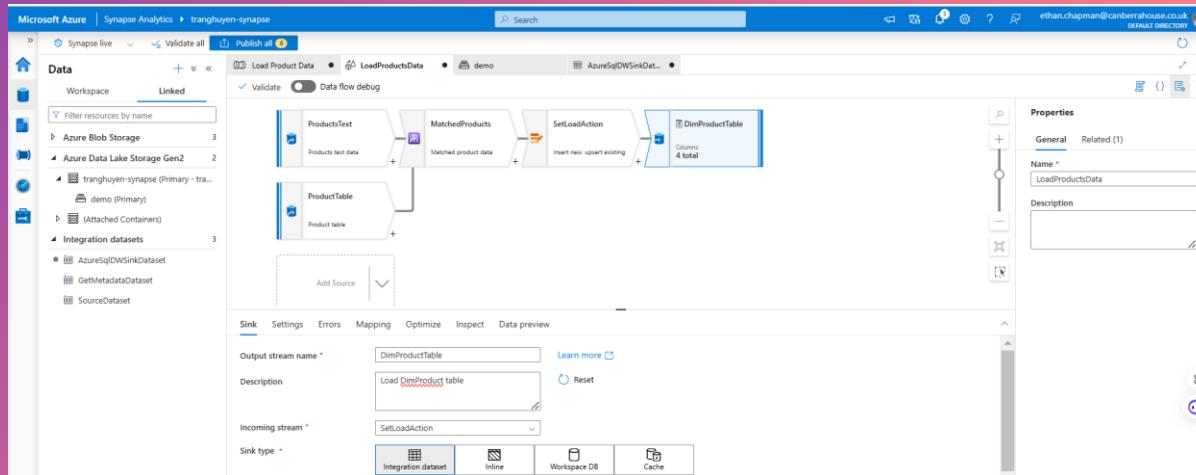


O4

Move & Transform Data



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).





05 | Triggers



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, có 4 loại trigger.

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	Schedule	Started	1	
* Trigger 1	Schedule	Stopped	0	



06

Manage – 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. Liên kết với powerBI để trực quan hóa dữ liệu.

The screenshot shows the Microsoft Azure Synapse Analytics interface. On the left, a sidebar lists various service categories: Analytics pools, External connections, Microsoft Purview, Integration, Triggers, Integration runtimes, Security, Access control, and Credentials. The 'Linked services' option is selected. The main area displays a table of existing linked services, with a red arrow pointing from the 'New' button at the top left of the table to a 'New linked service' dialog box on the right. The dialog box contains a grid of icons for different service providers, with the 'Power BI' icon highlighted by a red arrow. The table below shows five items:

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

The 'New linked service' dialog box contains a grid of service icons:

Icon	Service Name
PayPal (Preview)	PayPal
Phoenix	Phoenix
PostgreSQL	PostgreSQL
qb	QuickBooks (Preview)
Presta (Preview)	Presta
SAP BW	SAP BW
SAP BW via MDR	SAP BW via MDR
C4C	SAP Cloud for Customer
SAP ECC	SAP ECC
SAP HANA	SAP HANA

At the bottom of the dialog box are 'Continue' and 'Cancel' buttons.



07

Manage – Integration runtime



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



05

AZURE SYNAPSE ANALYTICS SQL ANALYTICS





V. AZURE SYNAPSE ANALYTICS SQL ANALYTICS

O1 Analytics

O2 Data Storage and Performance Optimizations

O3 Performance Optimizations Workload Management

O4 SQL Monitor with DMVs

O5 Developer productivity

O6 Maintenance

O7 Snapshots and restores





O1

Analytics

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

- Hệ thống lưu trữ tiên tiến bao gồm:
 - Columnstore Indexes
 - Table Partitions
 - Distributed Tables
 - Isolation Modes
 - Materialized Views
 - Nonclustered Indexes
 - Result-set Caching

- Truy vấn T-SQL cung cấp các tính năng mạnh mẽ như:

- Windowing Aggregates
- Approximate Execution (HyperLogLog)
- JSON Data Support

- Mô hình đối tượng SQL hoàn chỉnh, bao gồm:

- Tables
- Views
- Stored Procedures
- Functions





O1

Analytics



- 2. Windowing functions
- Aggregate functions
- Mệnh đề OVER

```
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 Chudukaktil	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	Tsolfias	1421810.9242	98055

- Ranking functions
- Analytical functions: PERCENTILE_DISC
- ROWS | RANGE: UNBOUNDED PRECEDING

```
-- 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



O1

Analytics

3. Approximate execution

- HyperLogLog accuracy
- APPROX_COUNT_DISTINCT

APPROX_COUNT_DISTINCT

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

100 %

Approx_Distinct_OrderKey
540

COUNT DISTINCT

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

100 %

Distinct_OrderKey
542

4. Group by options

- Group by with rollup
- Grouping sets

-- 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



O1

Analytics

5. Snapshot isolation

- Không thể đọc dữ liệu đã được sửa đổi nhưng chưa được cam kết

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

- Isolation levels (Cấp độ cô lập)
 - READ COMMITTED
 - REPEATABLE READ
 - SNAPSHOT
 - READ UNCOMMITTED
 - SERIALIZABLE
- Read_committed_snapshot
 - OFF (Mặc định): shared locks (khóa chia sẻ)
 - ON: row versioning



O1

Analytics

6. JSON data support

- Insert JSON data

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

- Read JSON data

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

- Modify and operate on JSON data

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

- Stored Procedures

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



02

Data Storage and Performance Optimizations



1. Database Tables

- Tables – Indexes

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

- Clustered Columnstore Index

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

- Tables – Distributions

```
CLUSTERED COLUMNSTORE INDEX,  
DISTRIBUTION = HASH([OrderId]) |  
ROUND ROBIN |  
REPLICATED
```

- Stored Procedures

```
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'
```

O2

Data Storage and Performance Optimizations



2. Database Views

- Materialized views

```
-- 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;
```

- COPY

```
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>')  
)
```

- Result-set caching

```
-- 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}
```

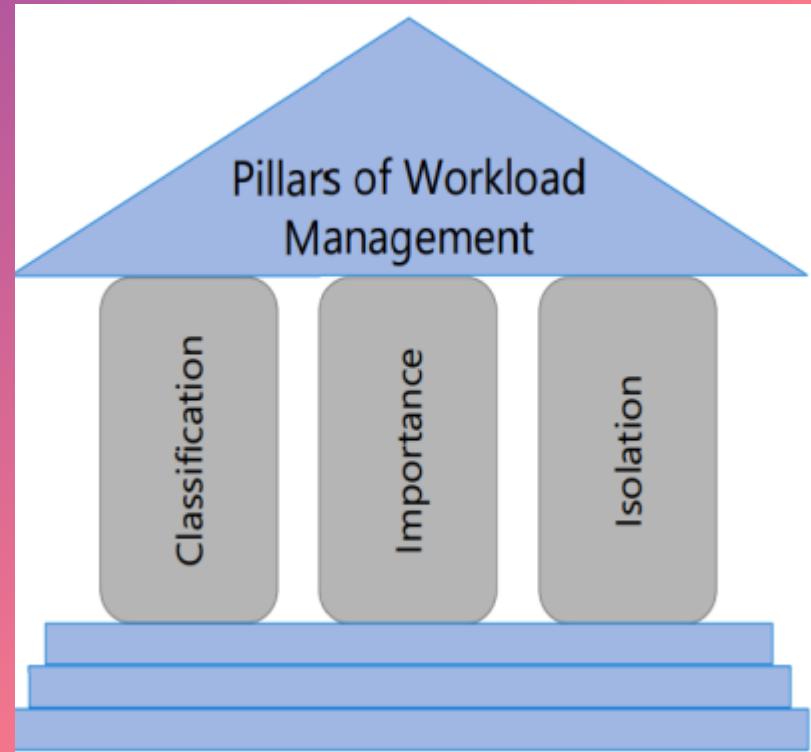
03

Performance Optimizations Workload Management



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.





03

Performance Optimizations Workload Management



1. Resource classes

```
/* 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';
```

2. Workload classification

```
CREATE WORKLOAD CLASSIFIER classifier_name
WITH
(
    [WORKLOAD_GROUP = '<Resource Class>']
    [IMPORTANCE = { LOW
                    |
                    BELOW_NORMAL
                    |
                    NORMAL
                    |
                    ABOVE_NORMAL
                    |
                    HIGH
                }
    ]
    [MEMBERNAME = 'security_account']
)
```

3. Workload importance

```
CREATE WORKLOAD CLASSIFIER National_Analyst
WITH
(
    [WORKLOAD_GROUP = 'smallrc']
    [IMPORTANCE = HIGH]
    [MEMBERNAME = 'National_Analyst_Login']
```

4. Workload Isolation

```
CREATE WORKLOAD GROUP group_name
WITH
(
    MIN_PERCENTAGE_RESOURCE = value
    , CAP_PERCENTAGE_RESOURCE = value
    , REQUEST_MIN_RESOURCE_GRANT_PERCENT = value
    [ [ , ] REQUEST_MAX_RESOURCE_GRANT_PERCENT = value ]
    [ [ , ] IMPORTANCE = {LOW | BELOW_NORMAL | NORMAL | ABOVE_NORMAL | HIGH} ]
    [ [ , ] QUERY_EXECUTION_TIMEOUT_SEC = value ]
)[ ; ]
```

O4

SQL Monitor with DMVs

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();
```

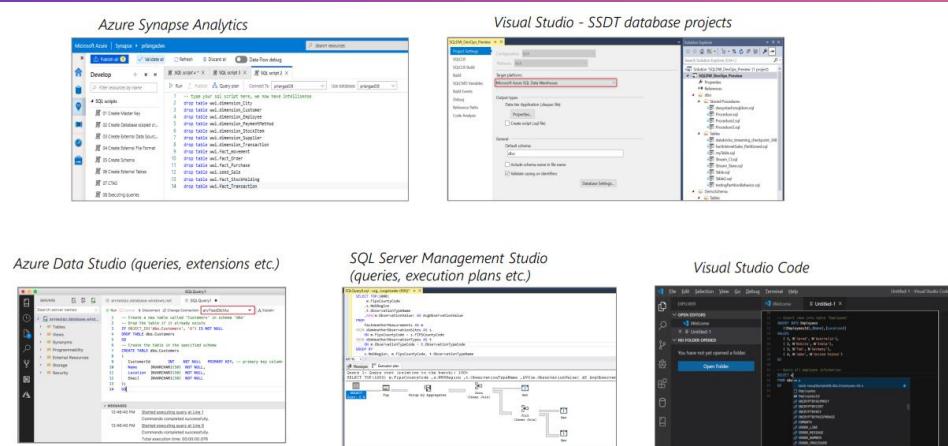
List all active queries

```
-- 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;
```

05

Developer productivity

1. Developer Tools



2. CI/CD Support

- Azure DevOps, cung cấp các tính năng sau:
 - Azure Pipelines
 - Azure Repos
 - Azure Test Plans
- Hỗ trợ tích hợp với các công cụ khác như Timetracker, Microsoft Teams, Slack, Jenkins, v.v.,

06

Maintenance

1. Maintenance windows

The screenshot shows the 'Maintenance Schedule (preview)' page in the Azure portal. It includes a note about maintenance windows, a section to choose a primary window (set to Saturday - Sunday), and two tables for primary and secondary maintenance windows. Both windows are set for Saturday at 03:00 UTC for 8 hours.

Primary maintenance window	Secondary maintenance window
Day: Saturday Start time: 03:00 UTC Time window: 8 hours	Day: Tuesday Start time: 13:00 UTC Time window: 8 hours

Schedule summary

Primary maintenance window: Saturday 03:00 UTC (8 hours)

Secondary maintenance window: Tuesday 13:00 UTC (8 hours)

2. Automatic statistics management

-- 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 }
```

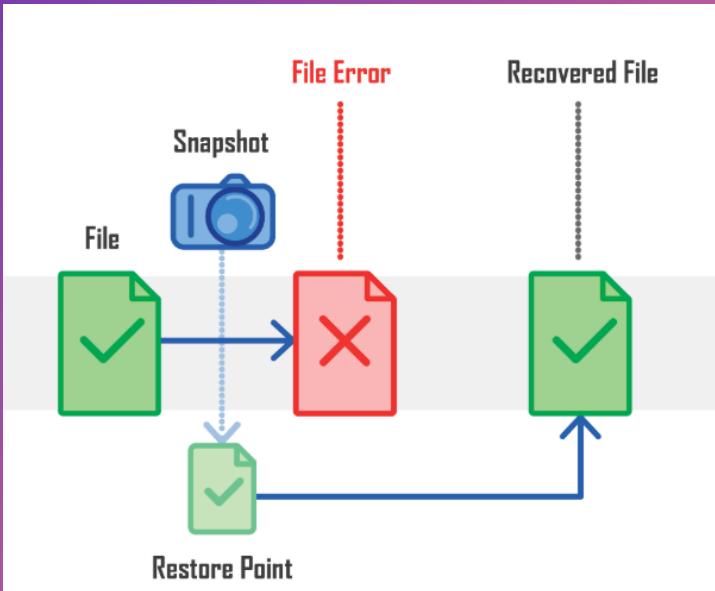
-- Check statistics settings for a database

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



07

Snapshots and restores



```
--View most recent snapshot time  
SELECT top 1 *  
FROM sys.pdw_loader_backup_runs  
ORDER BY run_id DESC;
```

06

SQL ON DEMAND





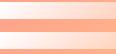
O1 | Overview

SQL On-Demand (Azure Synapse Serverless SQL Pools) là dịch vụ truy vấn tương tác, cho phép thực thi T-SQL trực tiếp trên dữ liệu lớn trong Azure Storage. Nó hỗ trợ khai thác dữ liệu hiệu quả mà không cần quản lý hạ tầng phức tạp.



O2 | Query

- Query on storage
- Query csv file
- Query json file
- Query parquet file
- Query folder





03

Create

Create views

```
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];
```

Create external table

```
-- Table to contain data
CREATE EXTERNAL TABLE dbo.ProcessedAccount (
    Id      varchar(100),
    SinkCreatedOn  DATETIME,
    SinkModifiedOn DATETIME,
    territorycode  varchar(100),
    parentaccountid varchar(100),
    parentaccountid_entitytype  varchar(100),
    msdyn_serviceterritory  varchar(100),
    msdyn_serviceterritory_entitytype  varchar(100),
    msdyn_billingaccount  varchar(100),
    msdyn_billingaccount_entitytype varchar(100),
    masterid  varchar(100),
    modifiedon DATETIME,
    masteraccountidname varchar(100),
    name      varchar(100),
    parentaccountidname varchar(100),
    createdon  DATETIME,
    description varchar(1000),
    accountid  varchar(100)
)
WITH (
    LOCATION='/ProcessedAccount.csv',
    DATA_SOURCE=AzureStorage,
    FILE_FORMAT=TextFile
);
```

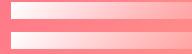
07

AZURE SYNAPSE ANALYTICS SPARK





V. AZURE SYNAPSE ANALYTICS SQL ANALYTICS



01

Azure Synapse Apache Spark –
Summary

02

Apache Spark

03

Motivation for Apache Spark

04

What makes Spark fast

05

General Spark Cluster
Architecture

06

Spark Component Features

07

Architecture Overview

08

Spark pool



O1

Azure Synapse Apache Spark – Summary



Azure Synapse Apache Spark là dịch vụ mạnh mẽ hỗ trợ xử lý và phân tích dữ liệu lớn trong Azure Synapse. Tính năng chính:

- Hỗ trợ ngôn ngữ: .NET Core 3.0, Python (PySpark), Scala, Java, Spark SQL, và R.
- Tích hợp Delta Lake: Hỗ trợ ACID và lưu trữ thời gian (time travel).
- Kết nối Azure Synapse: Azure Data Lake.
- Bảo mật: Đảm bảo bảo mật và xác thực người dùng.
- Tự động mở rộng/tạm dừng: Tối ưu hóa tài nguyên khi cần thiết.
- Notebook tích hợp: Hỗ trợ phát triển và phân tích dữ liệu.
- Tối ưu hóa tài nguyên: Xử lý nhanh, tiết kiệm chi phí.

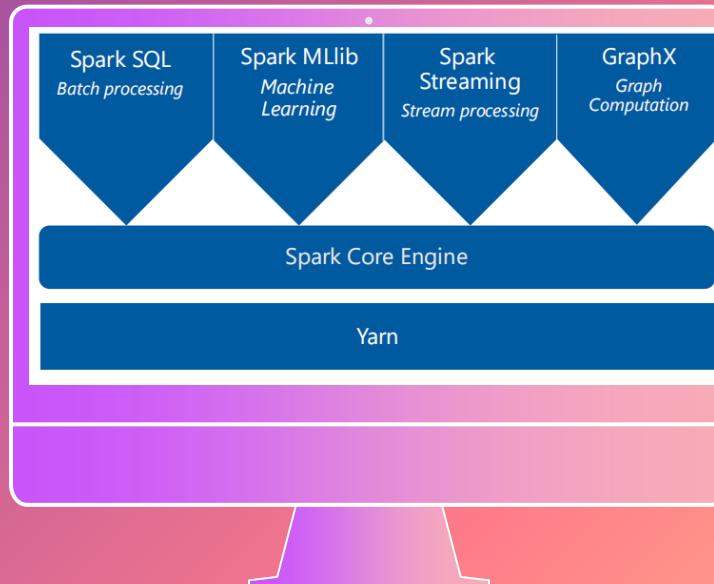


02

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:





03

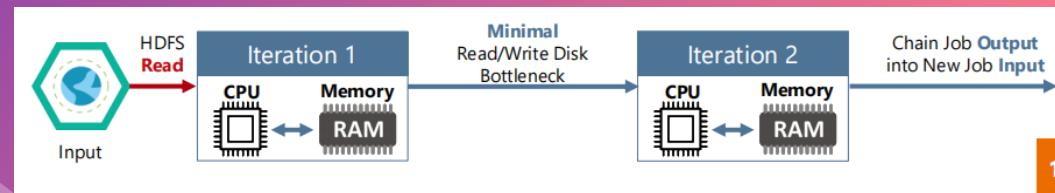
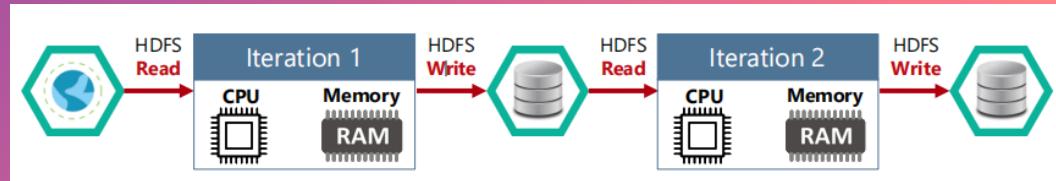
Motivation for Apache Spark



MapReduce: Đây là mô hình phổ biến trong Hadoop để xử lý dữ liệu lớn.

Vấn đề gặp phải:

- Tác vụ phức tạp: Xử lý nhiều jobs, truy vấn tương tác, xử lý sự kiện thời gian thực.
- MapReduce: Nhiều thao tác đọc/ghi đĩa (disk I/O) → tốc độ chậm, hiệu suất thấp.



Giải pháp: Apache Spark

- Lưu trữ in-memory: Giảm thao tác đĩa bằng cách giữ dữ liệu trong RAM.
- Engine phân tán: Tăng tốc xử lý dữ liệu lớn, nhanh hơn MapReduce nhiều lần.

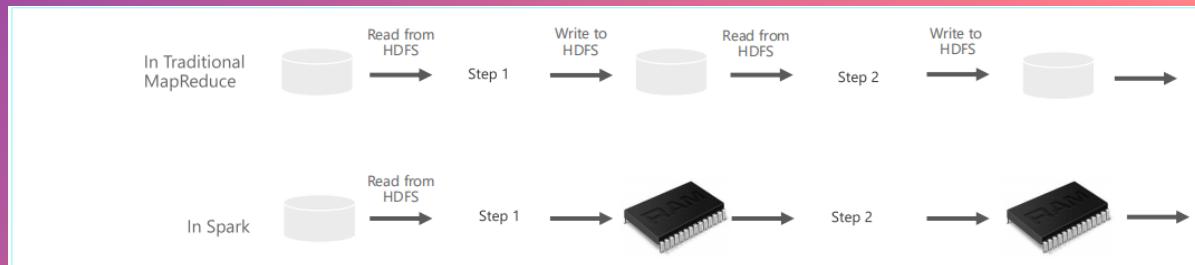


O4

What makes Spark fast



- Tính toán trong bộ nhớ: Spark lưu trữ tạm dữ liệu trong bộ nhớ, giúp truy vấn lặp nhanh hơn so với các hệ thống dựa trên ổ đĩa.
- Tích hợp Scala: Spark tích hợp với Scala, cho phép thao tác dữ liệu phân tán như dữ liệu cục bộ mà không cần map và reduce phức tạp.
- Chia sẻ dữ liệu nhanh: Dữ liệu được lưu trong bộ nhớ thay vì qua HDFS, giảm chi phí và không cần sao chép ba lần như Hadoop.



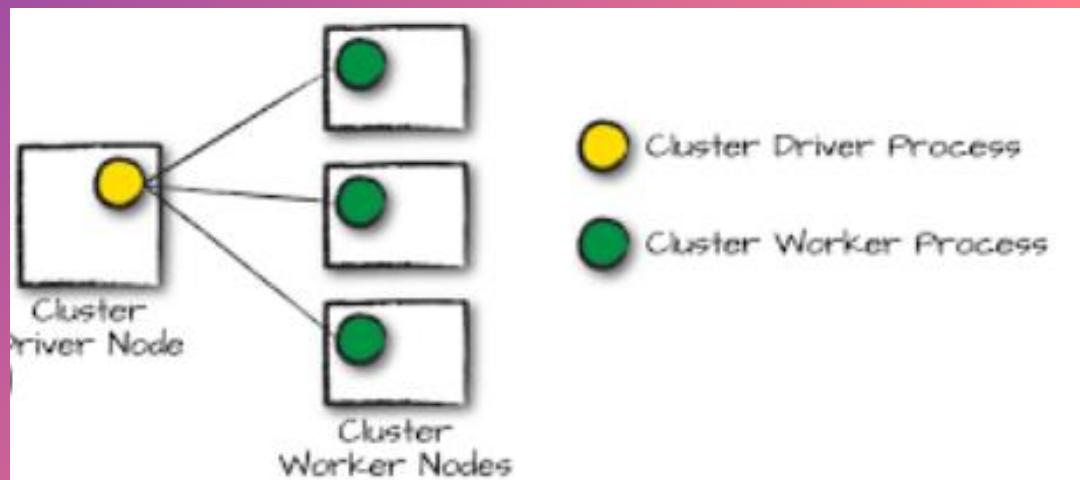


05

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.



06

Spark Component Features

APACHE SPARK ECOSYSTEM

Spark SQL

- Truy cập dữ liệu thống nhất qua SQL hoặc DataFrame APIs.
- Kết nối công cụ BI qua JDBC/ODBC.

Spark Streaming (Streaming)

- Xử lý micro-batch cho phân tích gần thời gian thực (IoT, Twitter, Kafka).
- Xuất dữ liệu theo lô đến các kho khác nhau.

MLlib (Machine learning)

- Thuật toán học máy: phân cụm, phân loại, hồi quy.
- Phân tích dự đoán và xây dựng ứng dụng thông minh.

GraphX (Graph Computation)

- Biểu diễn và phân tích đồ thị.
- Theo dõi kết nối giữa các nút, áp dụng cho giao thông, viễn thông, mạng xã hội, v.v.

SparkR (R on spark)

- Hỗ trợ R để phân tích dữ liệu lớn.
- Tích hợp DataFrames và các thư viện thống kê trong R.



07

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



O7

Architecture Overview

3. Cấu trúc

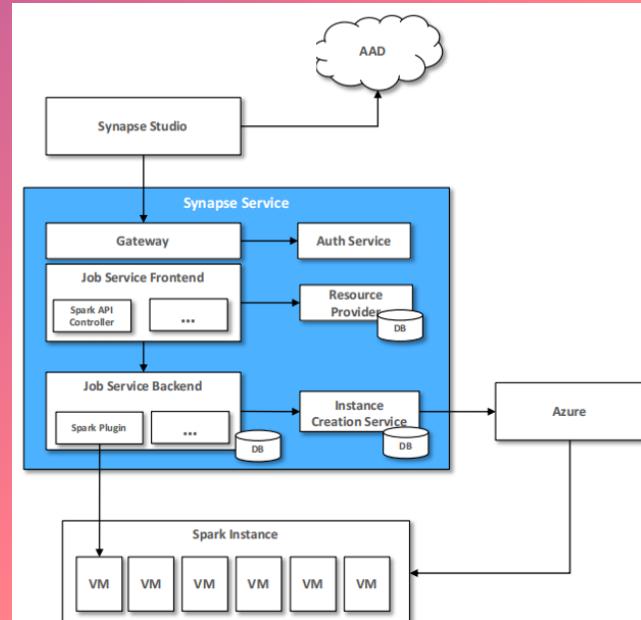
Synapse Job Service bao gồm hai phần chính:

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.





07

Architecture Overview



4. Quy trình tạo Spark cluster trong Azure Synapse

Người dùng tạo Synapse Workspace và Spark pool rồi mở Synapse Studio.

- Gắn Notebook vào Spark pool và nhập câu lệnh Spark.

Xác thực và chuyển tiếp yêu cầu:

- Notebook lấy token AAD và gửi yêu cầu tạo phiên Spark đến Synapse Gateway.
- Gateway xác thực và chuyển yêu cầu đến Spark Controller (Livy) trong Job Service.

Tạo phiên Spark:

- Job Service backend tạo 2 công việc: tạo cụm và tạo phiên Spark.
- Backend liên hệ với Synapse Resource Provider và Synapse Instance Service để tạo cụm.

Thực thi lệnh Spark:

- Khi cụm và phiên Spark được tạo, Notebook gửi các câu lệnh Spark đến Job Service.
- Job Service chuyển câu lệnh tới Livy endpoint trên cụm để thực thi.



08 | Spark pool



1. Tạo spark pool

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

New Apache Spark pool

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *

 ✓

Node size family

MemoryOptimized

Node size *

Small (4 vCores / 32 GB)

Autoscale *

Enabled Disabled

Number of nodes *

3 3

Dynamically allocate executors

Enabled Disabled

Estimated price

Est. cost per hour
1.74 to 1.74 USD
[View pricing details](#)

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

Apache Spark pools
sparkpool Apache Spark pool Small

08

Spark pool

2. User experience and languages

Tạo view và select view

The image displays two side-by-side screenshots of the Microsoft Azure Synapse Analytics workspace interface.

Notebook 1 (Left):

- Shows PySpark code:

```
1 df.createOrReplaceTempView('salesView')
```

 and

```
1 display(spark.sql('select* from salesView'))
```

.
- Output: "Command executed in 2 sec 867 ms by ethan.chapman on 1:35:05 PM, 12/15/24".
- Job execution status: "Job execution Succeeded Spark 2 executors 8 cores".
- Table output:

ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER
10107	30	95.7	2
10121	34	81.35	5
10134	41	94.74	2
10145	45	83.26	6

Notebook 2 (Right):

- Shows PySpark code:

```
1 display(spark.sql('select* from salesView'))
```

.
- Output: "Command executed in 2 sec 867 ms by ethan.chapman on 1:35:05 PM, 12/15/24".
- Job execution status: "Job execution Succeeded Spark 2 executors 8 cores".
- Chart output: A pie chart showing the distribution of sales by OrderNumber. The chart has several slices, with the largest slice labeled 10322.
- Chart configuration panel on the right:

 - Chart type: Pie chart
 - Key: ORDERNUMBER
 - Values: ORDERNUMBER
 - Series Group: ORDERNUMBER
 - Aggregation: Count

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



08 | Spark pool



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.

4. Spark ML Algorithms

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

- Phân loại và Hồi quy (Classification and Regression)
- Phân cụm (Clustering): k-means và streaming k-means
- Lọc cộng tác (Collaborative Filtering)

- Giảm chiều dữ liệu (Dimensionality Reduction)
- Khai thác mẫu thường xuyên (Frequent Pattern Mining): FP-growth
- Thống kê cơ bản (Basic Statistics): Tương quan

08 | Spark pool

5. Meachine learning

- Synapse Notebook:
 - Connect to AML workspace
 - Configure AML job to run on Synapse
 - Run AML job



```
files AML
Run all Undo Publish Outline Attach to sparkpool Language PySpark (Python) Variables
Ready
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)
[1] ✓ 1 min 21 sec - Apache Spark session started in 1 min 14 sec 828 ms. Command executed in 6 sec 926 ms by ethan.chapman on 2:38:40 PM, 12/15/24
*** SDK Version: 1.55.0

D | v
1 from azureml.core import Workspace
2 from azureml.core.workspace import WorkspaceException
3
4 try:
5     # Kết nối tới workspace
6     ws = Workspace(subscription_id="0e3ab6f7-120f-4b12-8d36-8ea6f30a4b25",
7                     resource_group="cuoiky",
8                     location="North Europe",
9                     workspace_name="cuoiky-workspace")
10
11     # In thông tin workspace nếu kết nối thành công
12     print("Workspace kết nối thành công!")
13     print("Thông tin workspace:")
14     print(f"Name: {ws.name}")
15     print(f"Location: {ws.location}")
16     print(f"Resource Group: {ws.resource_group}")
17
18 except WorkspaceException as e:
19     print("Lỗi: không thể tìm thấy workspace. Vui lòng kiểm tra các thông tin sau:")
20     print(f"1. Subscription ID có chính xác không?")
21     print(f"2. Resource Group có tồn tại không?")
22     print(f"3. Workspace Name có chính xác không?")
23     print(f"4. Bạn có đủ quyền truy cập vào workspace không?")
24     print(f"Chi tiết lỗi: {e}")
25 except Exception as ex:
26     print("Một lỗi không xác định đã xảy ra:")
27     print(f"{ex}")

1 # Import modules
2 import azureml.core
3 import pandas as pd
4 from azureml.core.authentication import ServicePrincipalAuthentication
5 from azureml.core.workspace import Workspace
6 from azureml.core.experiment import Experiment
[1] ✓ <1 sec - Command executed in 145 ms by ethan.chapman on 3:05:29 PM, 12/15/24
```

08

INDUSTRY-LEADING SECURITY AND COMPLIANCE





08

Industry-leading security and compliance



Data protection

- Truyền tải dữ liệu
- Phân loại dữ liệu
- Mã hóa dữ liệu

- Bảo mật cấp đối tượng (Bảng/View)
- Bảo mật cấp dòng
- Bảo mật cấp cột.
- Mặt nạ dữ liệu động

Access control

Authenticaiton

- Đăng nhập SQL
- Xác thực Azure Active Directory (AAD)
- Xác thực đa yếu tố



08

Industry-leading security and compliance



Network security

- Mạng ảo
- Tường lửa bảo mật
- Kết nối riêng Azure

- Phát hiện mối đe dọa
- Kiểm tra và bảo vệ mối đe dọa
- Đánh giá lỗ hổng

Threat protection



PART 2.

Application



Part 2. Application

01

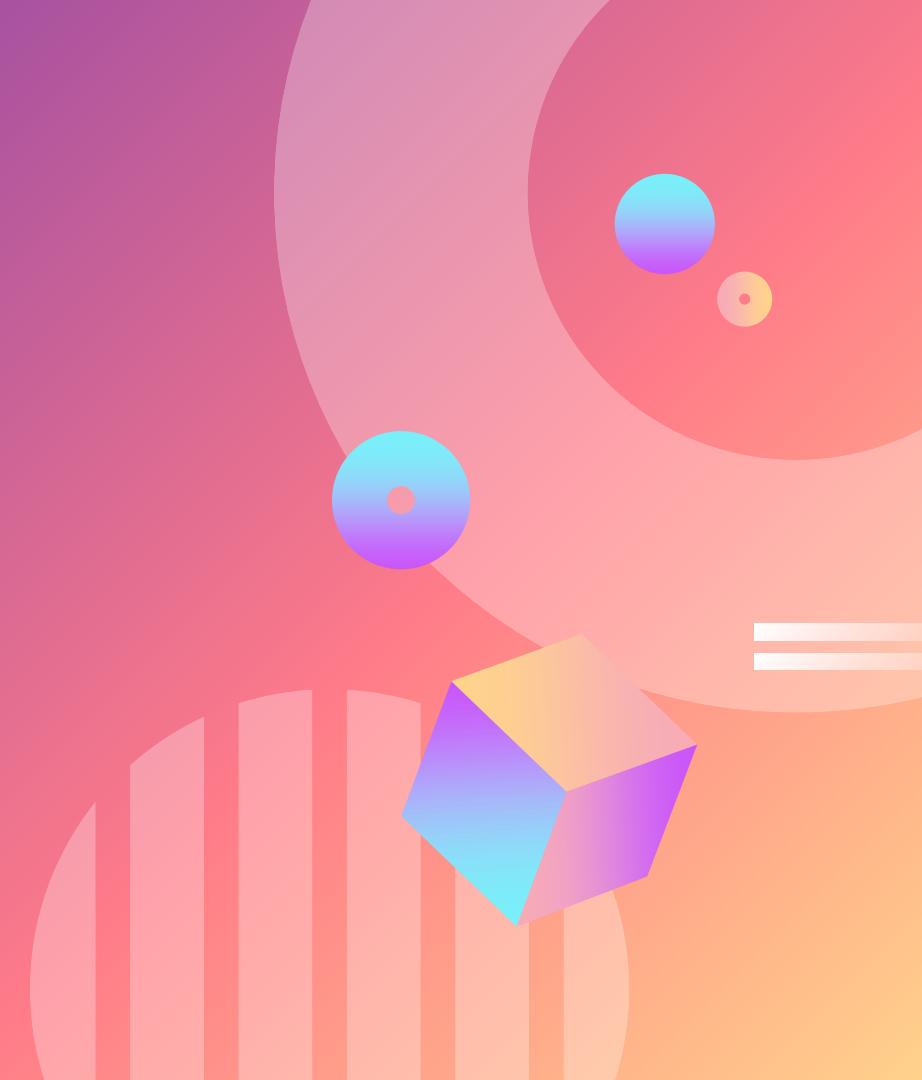
Setup workspace

02

SQL Serverless

03

Spark pool



O1

Setup workspace

The screenshot shows the Microsoft Azure Resource Groups page for the 'cuoiky' resource group. The top navigation bar includes 'Microsoft Azure', 'Upgrade', 'Search resources, services, and docs (G+)', 'Copilot', and user information 'ethan.chapman@canbe... DEFAULT DIRECTORY (ETHANCH...)'. The left sidebar lists 'Resource groups' with 'cuoiky' selected, and other options like 'Create', 'Manage view', and 'Essentials'. The main content area displays the 'Overview' tab for the 'cuoiky' group, featuring a search bar, a 'Create' button, and a 'Delete resource group' button. The 'Essentials' section shows a table of resources:

Name	Type	Location	Actions
cuoiky-workspace	Synapse workspace	Southeast Asia	...
sparkpool (cuoiky-workspace/sparkpool)	Apache Spark pool	Southeast Asia	...
sqlpool (cuoiky-workspace/sqlpool)	Dedicated SQL pool	Southeast Asia	...
tranghuyen	Storage account	Southeast Asia	...

At the bottom right of the table, there is a blue circular icon with a white question mark inside.

O2 | 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]
```

Results Messages

View Table Chart Export results ▾

Search

SalesOrderNu...	SalesOrderLine...	OrderDate	CustomerName	EmailAddress	Item
SO43701	1	2019-07-01	Christy Zhu	christy12@adv...	Mountain-100
SO43704	1	2019-07-01	Julio Ruiz	julio1@adventu...	Mountain-100
SO43705	1	2019-07-01	Curtis Lu	curtis9@advent...	Mountain-100
SO43700	1	2019-07-01	Ruben Prasad	ruben10@adve...	Road-650 Blac...
SO43703	1	2019-07-01	Albert Alvarez	albert7@adven...	Road-150 Red...
SO43697	1	2019-07-01	Cole Watson	cole1@adventu...	Road-150 Red...
SO43699	1	2019-07-01	Sydney Wright	sydney61@adv...	Mountain-100



03 | Spark pool

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))
```

B4: 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

03 | Spark pool

B5: Chia 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
```

B6: Xây dựng mô hình huấn luyện

```
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
View in monitoring | Open Spark UI
... Accuracy: 0.336742482063807
```

PART 3.



DEVELOPMENT PROBLEM

01

Mô tả bài toán

02

Phân tích dữ liệu tổng quan

03

Tiền xử lý dữ liệu

04

Chuẩn hóa dữ liệu

05

Triển khai học máy

06

Kết luận bài toán



O1

Mô tả 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. 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.

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.

O2 | Phân tích dữ liệu tổng quan

Bước 1: Đọc dữ liệu

```
1 %%pyspark
2 df = spark.read.load('abfss://files@tranghuyen.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 chanman 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

O2 | Phân tích dữ liệu tổng quan

B2: Thực hiện một số truy vấn cơ bản

```
▶ | ▾
  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|
+-----+
```

O3 | Tiền xử lý dữ liệu

a) Recency-đo khoảng thời gian giữa thời điểm khách hàng mua hàng lần cuối và ngày đầu tiên có đơn hàng 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_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
```

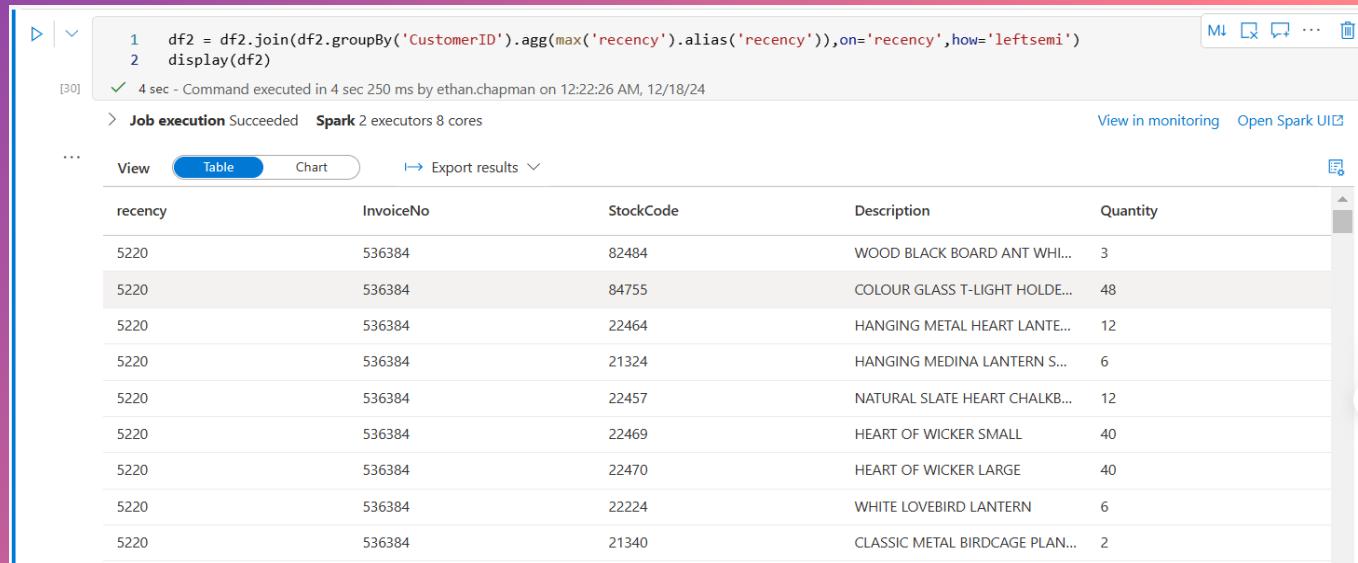
B2: Lấy giá trị thời gian mua của từng đơn hàng trừ đi from_date, giá trị sẽ được lưu tại cột 'recency'

```
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
```

O3 | Tiền xử lý dữ liệu

a) Recency-đo khoảng thời gian giữa thời điểm khách hàng mua hàng lần cuối và ngày đầu tiên có đơn hàng giá trị này càng lớn chứng tỏ khách hàng càng mua gần đây.

B3: Lấy lần cuối mà khách hàng mua hàng



The screenshot shows a Jupyter Notebook cell with the following content:

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

[30] ✓ 4 sec - Command executed in 4 sec 250 ms by ethan.chapman on 12:22:26 AM, 12/18/24

> Job execution Succeeded Spark 2 executors 8 cores

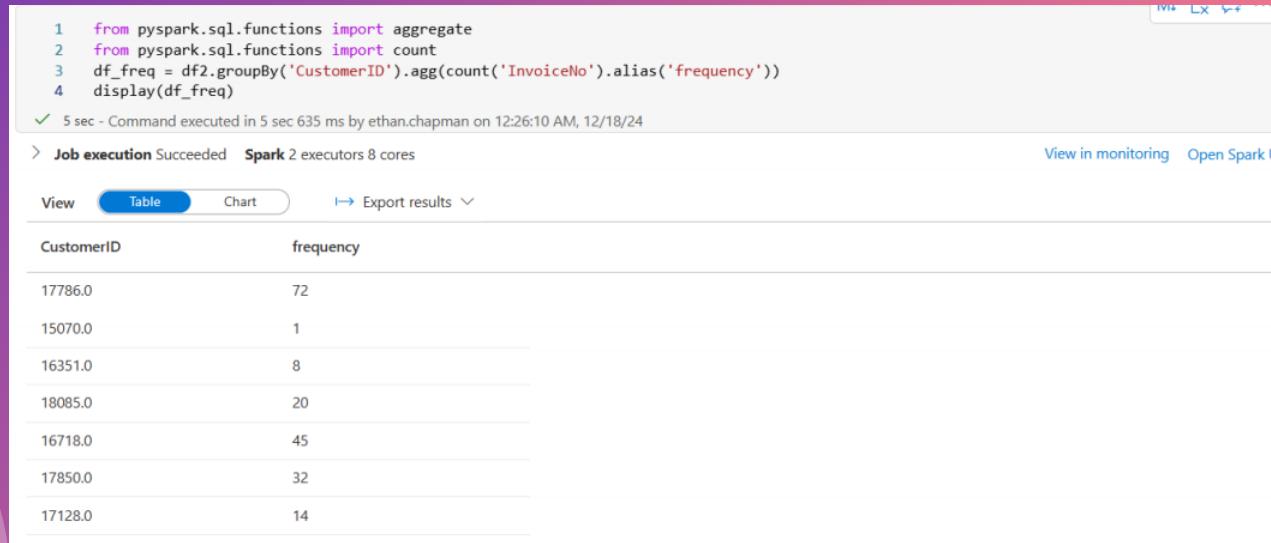
View in monitoring Open Spark UI

The cell displays a table of data with the following columns: recency, InvoiceNo, StockCode, Description, and Quantity. The data consists of 10 rows, each representing a purchase record with its corresponding details.

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

03 | Tiền xử lý dữ liệu

b) Frequency -Đo tần suất mà khách hàng mua hàng trong một khoảng thời gian nhất định



The screenshot shows a Jupyter Notebook cell with the following content:

```
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)
```

Execution results:

✓ 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 in monitoring Open Spark UI

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

03 | Tiết xử lý dữ liệu

c) Monetary-Đo giá trị đặt hàng của khách hàng.

```
1 # Tính số lượng và đơn giá của một lần mua hàng
2 m_val = df2.withColumn('TotalAmount', col("Quantity") * col("UnitPrice"))
3 display(m_val)
```

✓ 6 sec - Command executed in 5 sec 658 ms by ethan.chapman on 12:33:20 AM, 12/18/24

```
1 final_df = m_val.join(df3, on='CustomerID', how='inner')
```

```
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'))
```

d) lấy 4 trường dữ liệu sau để xây dựng mô hình 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
```

▶ | ▽
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	623.750000000002	16718.0
91080	32	126.38	17850.0

04

Chuẩn hóa dữ liệu

```
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)

✓ 9 sec - Command executed in 8 sec 922 ms by ethan.chapman on 12:54:54 AM, 12/18/24

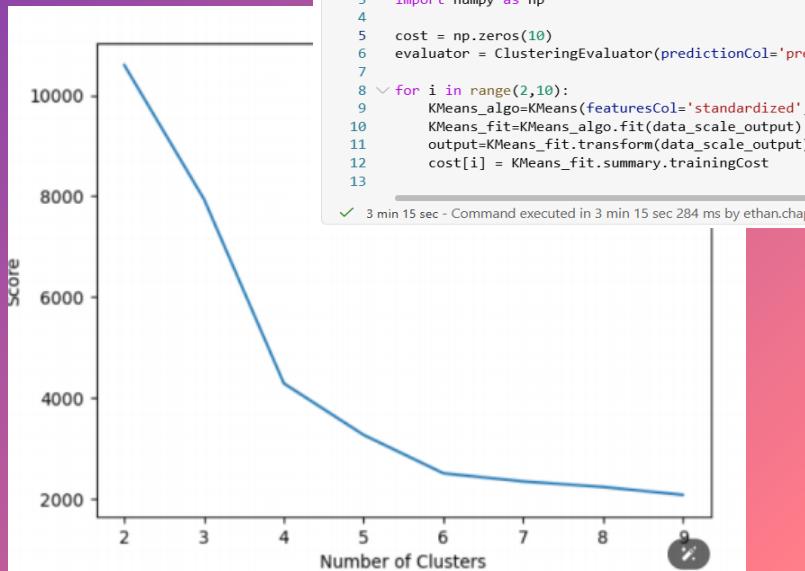
> **Job execution** Succeeded **Spark** 2 executors 8 cores

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

05 | Triển khai học máy_sử dụng thuật toán kmeans

a) Áp dụng quy tắc khuỷu tay để xác định số cụm

Ở đây thấy được số tâm cụm cần xác định là 3



```
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
```

✓ 3 min 15 sec - Command executed in 3 min 15 sec 284 ms by ethan.chapman on 1:03:47

```
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()
```

✓ <1 sec - Command executed in 660 ms by ethan.chapman on 1:03:47

05 | Triển khai học máy_sử dụng thuật toán kmeans

b) Train kmean

```
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
```

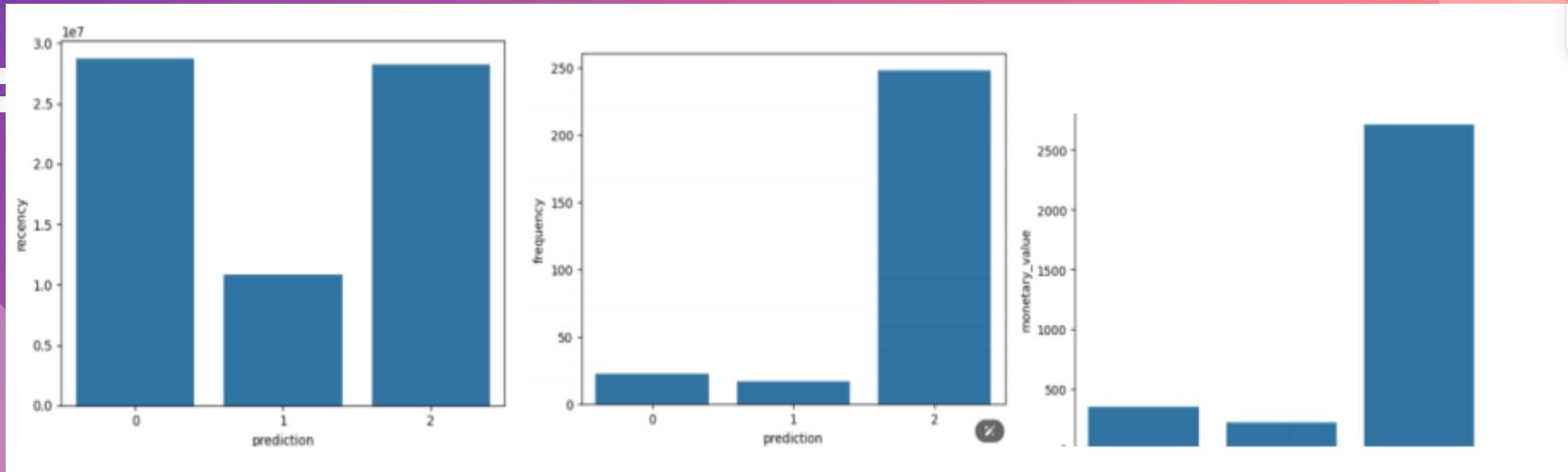
c) Dự đoán

```
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
```

CustomerID	recency	frequency	monetary_value	features	standardized	prediction
18170160	1	204.0	16553.0	[1.817016E7, 1.0, 2.0, ...]	[2.08327580864095...]	1
16952220	3	76.5	17536.0	[1.695222E7, 3.0, 7.0, ...]	[1.94363449902253...]	1
19549920	29	187.9199999999996	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, 8.0, ...]	[2.42074975937500...]	0

05 | Triển khai học máy_sử dụng thuật toán kmeans

b) Sử dụng matplotlib trực quan hóa phân khúc khách hàng



PART 4.

Conclusion

KẾT LUẬN

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.



Thank you

Cảm ơn thầy đã lắng nghe