Browsersync: connected

Caloudi 教育訓練課程

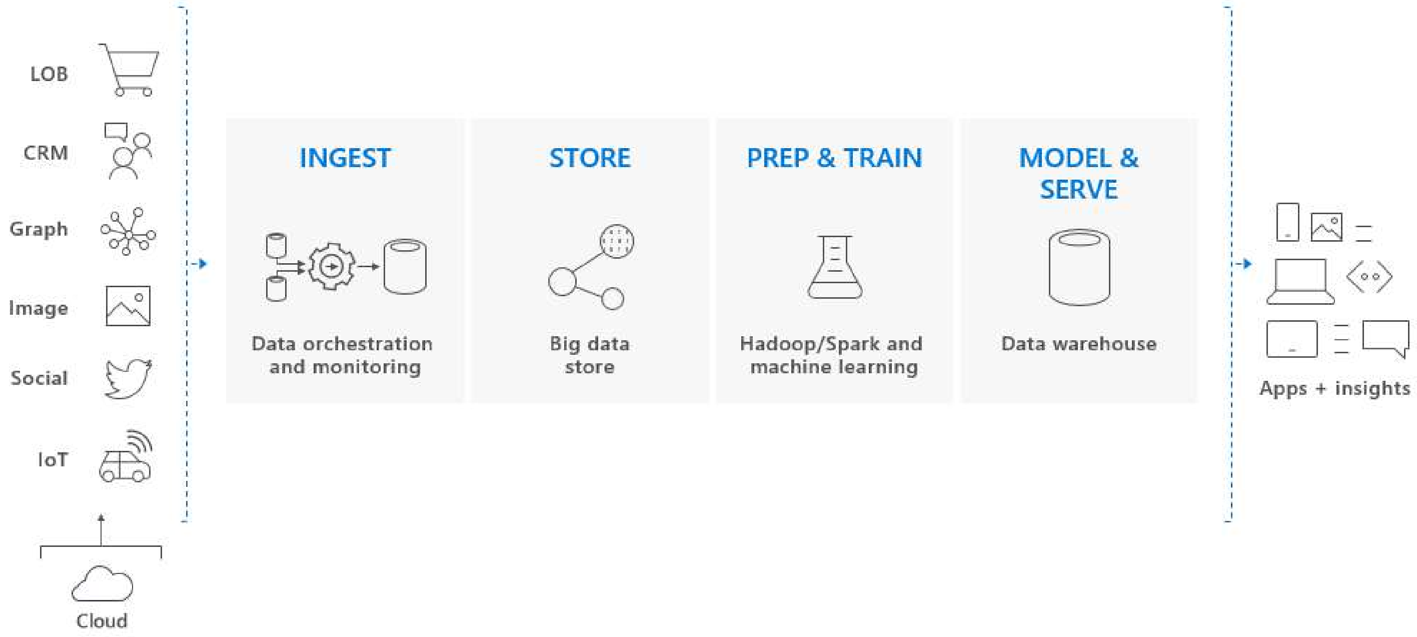
Azure Synapse Analytics Workshop

Agenda

- Data Analytics Evolution
- Data Lake Evolution
- Azure Synapse Analytics
- SQL DW / Dedicated SQL Pool
- Implementation Issues
- References

Data Analytics Evolution

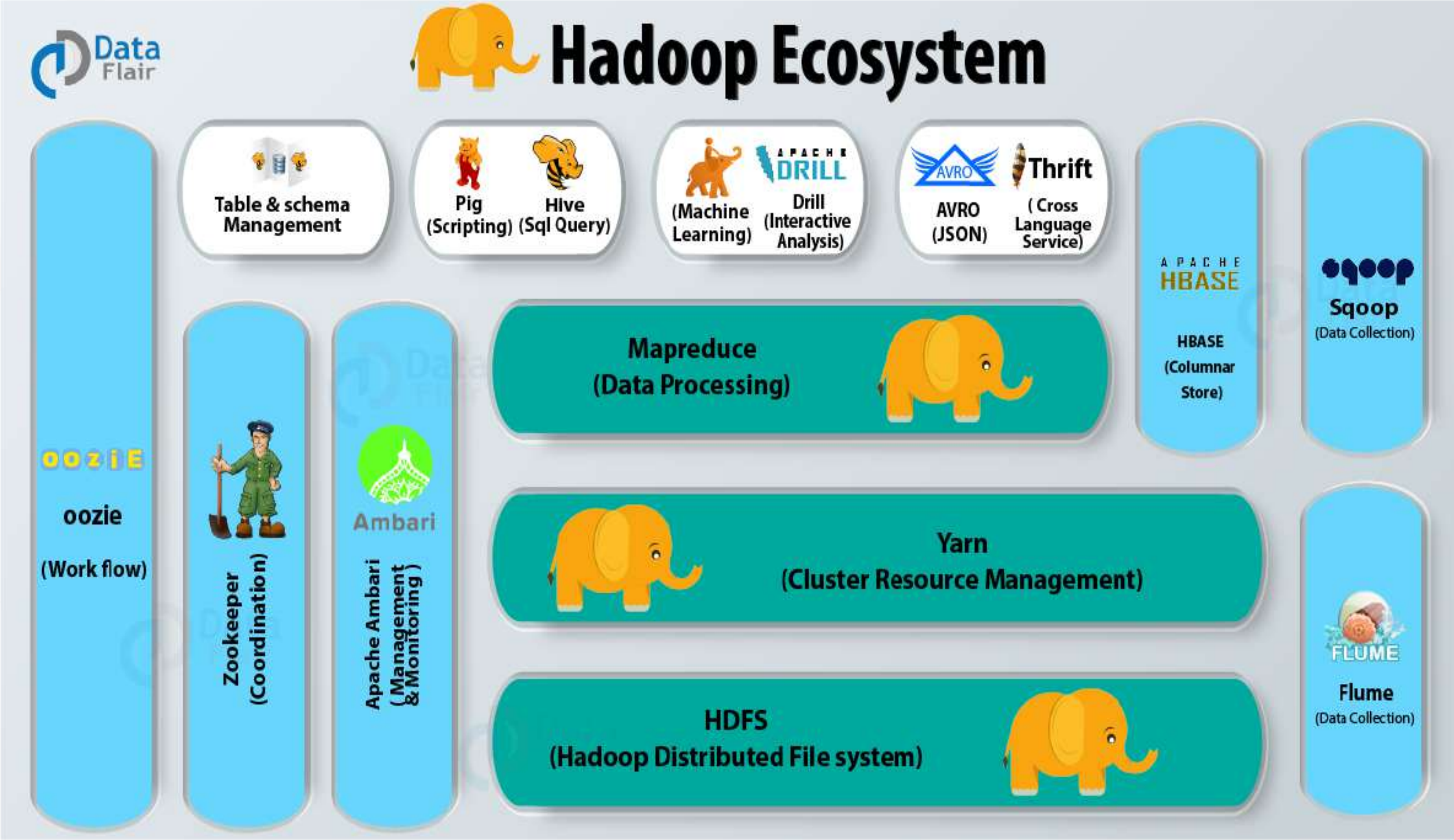
Data Warehouse Solution



參考資料：What is dedicated SQL pool (formerly SQL DW) in Azure Synapse Analytics?

Apache Hadoop Ecosystem

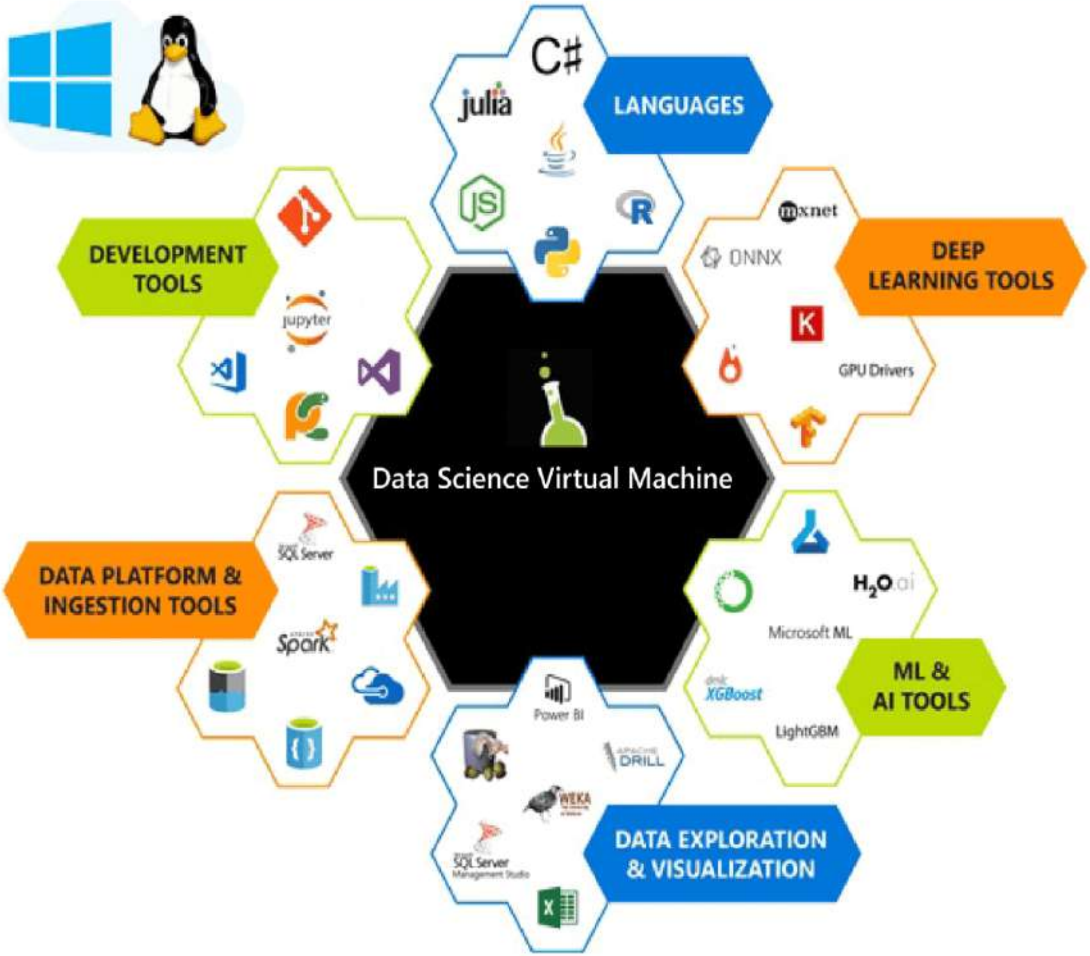
- 根據 Google 在 2003 年發表的 Google File System 論文而開發
- 主要包含 Hadoop Distributed File System (HDFS) 與 MapReduce Engine



參考資料：Hadoop Ecosystem – 15 Must Know Hadoop Components

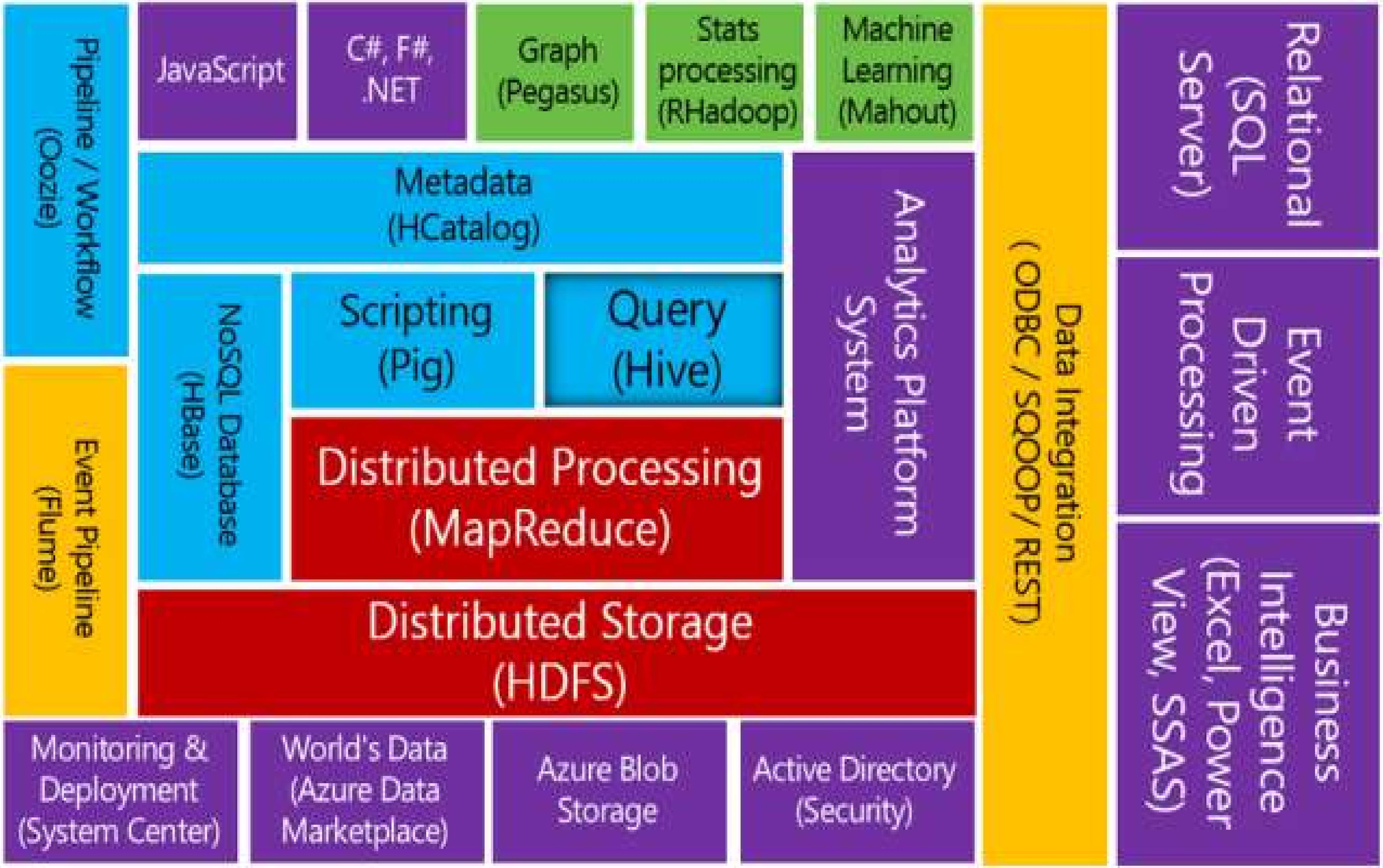
Browsersync: connected

Azure Data Science Virtual Machine



參考資料：Data Science Virtual Machines

Azure HDInsight Ecosystem



Pipeline / Workflow (Oozie)

JavaScript

C#, F#, .NET

Graph (Pegasus)

Stats processing (RHadoop)

Machine Learning (Mahout)

Data Integration (ODBC / SQOOP/ REST)

Relational (SQL Server)

Metadata (HCatalog)

Analytics Platform System

Event Driven Processing

NoSQL Database (HBase)

Scripting (Pig)

Query (Hive)

Event Pipeline (Flume)

Distributed Processing (MapReduce)

Business Intelligence (Excel, Power View, SSAS)

Distributed Storage (HDFS)

Monitoring & Deployment (System Center)

World's Data (Azure Data Marketplace)

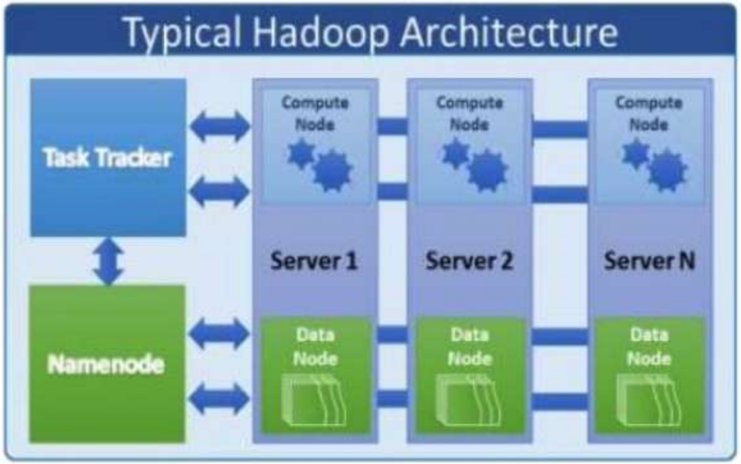
Azure Blob Storage

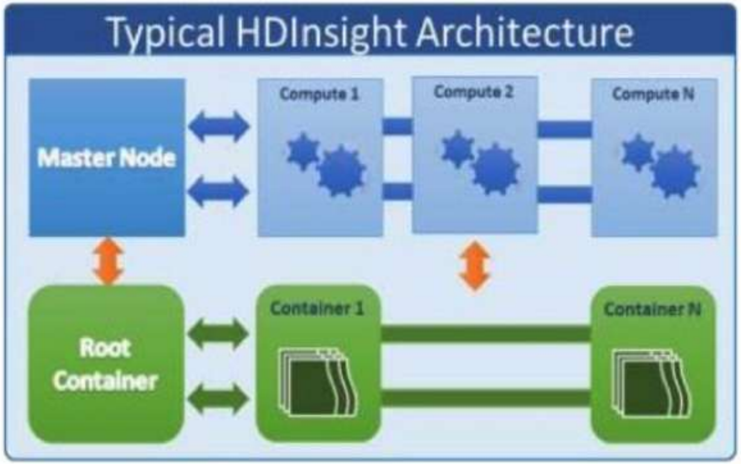
Active Directory (Security)

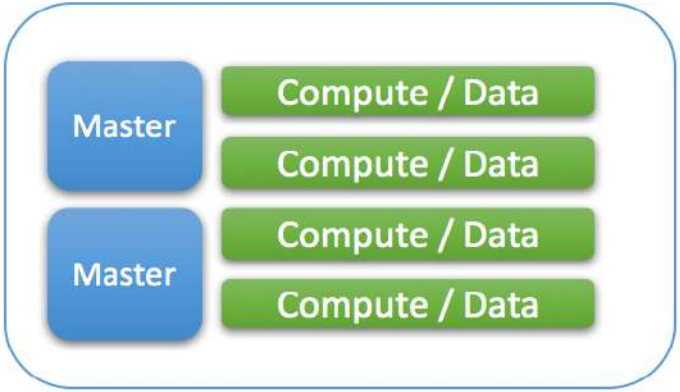
參考資料：[Big Data] Hadoop/HDInsight : concepts généraux

Browsersync: connected

Apache Hadoop vs. Azure HDInsight

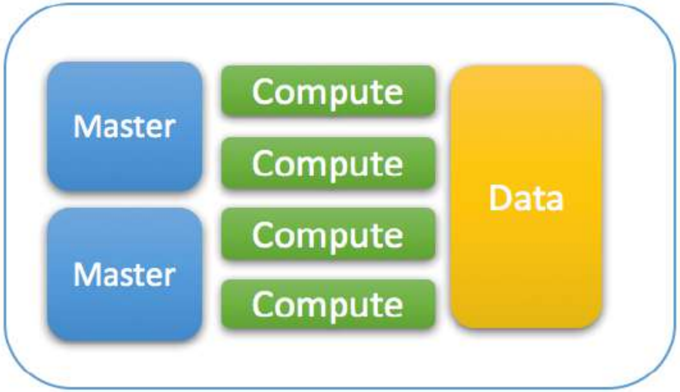






On-Premises Hadoop Architecture

vs



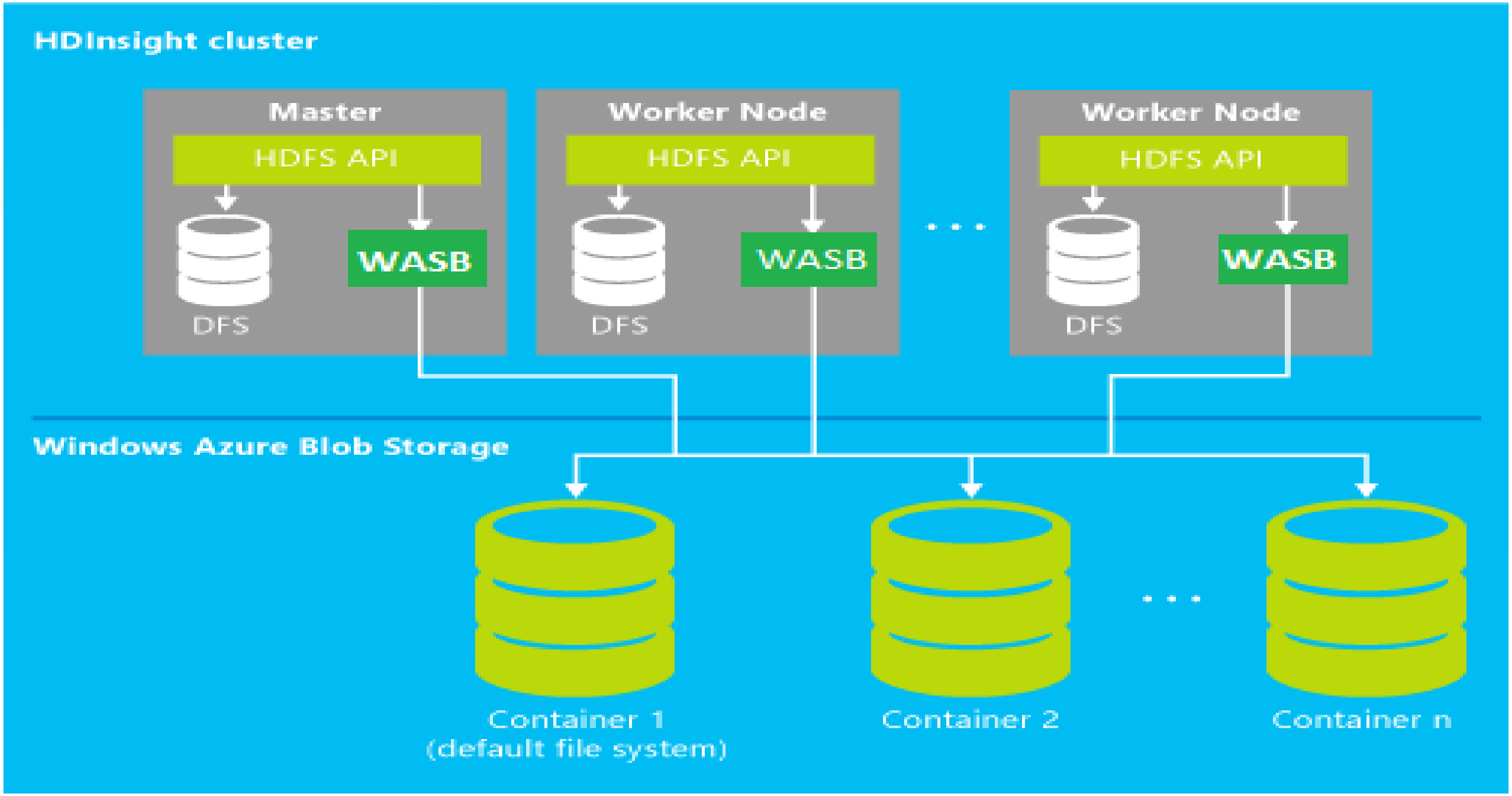
HDInsight Architecture

參考資料：
Introduction to Microsoft Azure HD Insight by Dattatrey Sindhol
HOW TO CHOOSE THE RIGHT AZURE HDINSIGHT CLUSTER

Azure Storage Type for Apache Hadoop

`wasb`：

- Windows Azure Storage Blob
- HDFS File System Driver over Azure Blob Storage

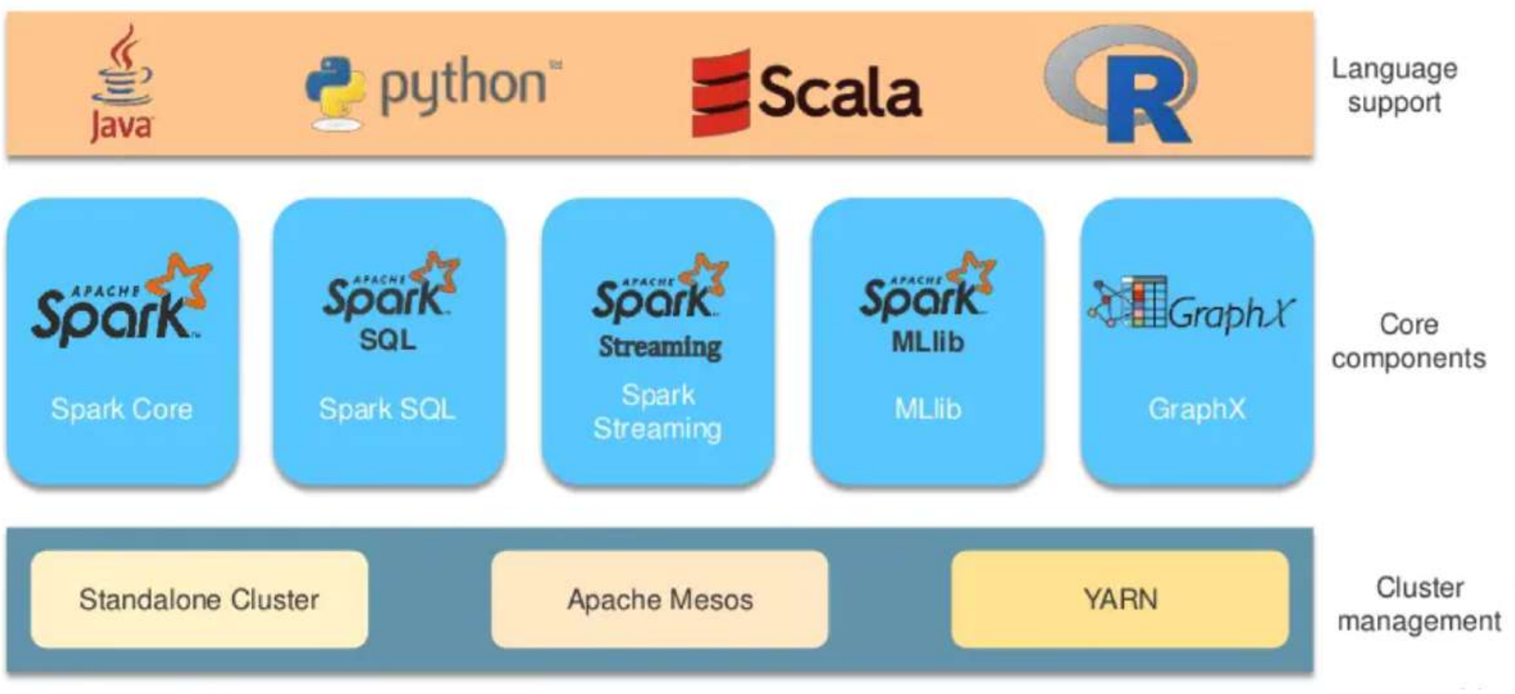


參考資料：What is HDInsight?

Browsersync: connected

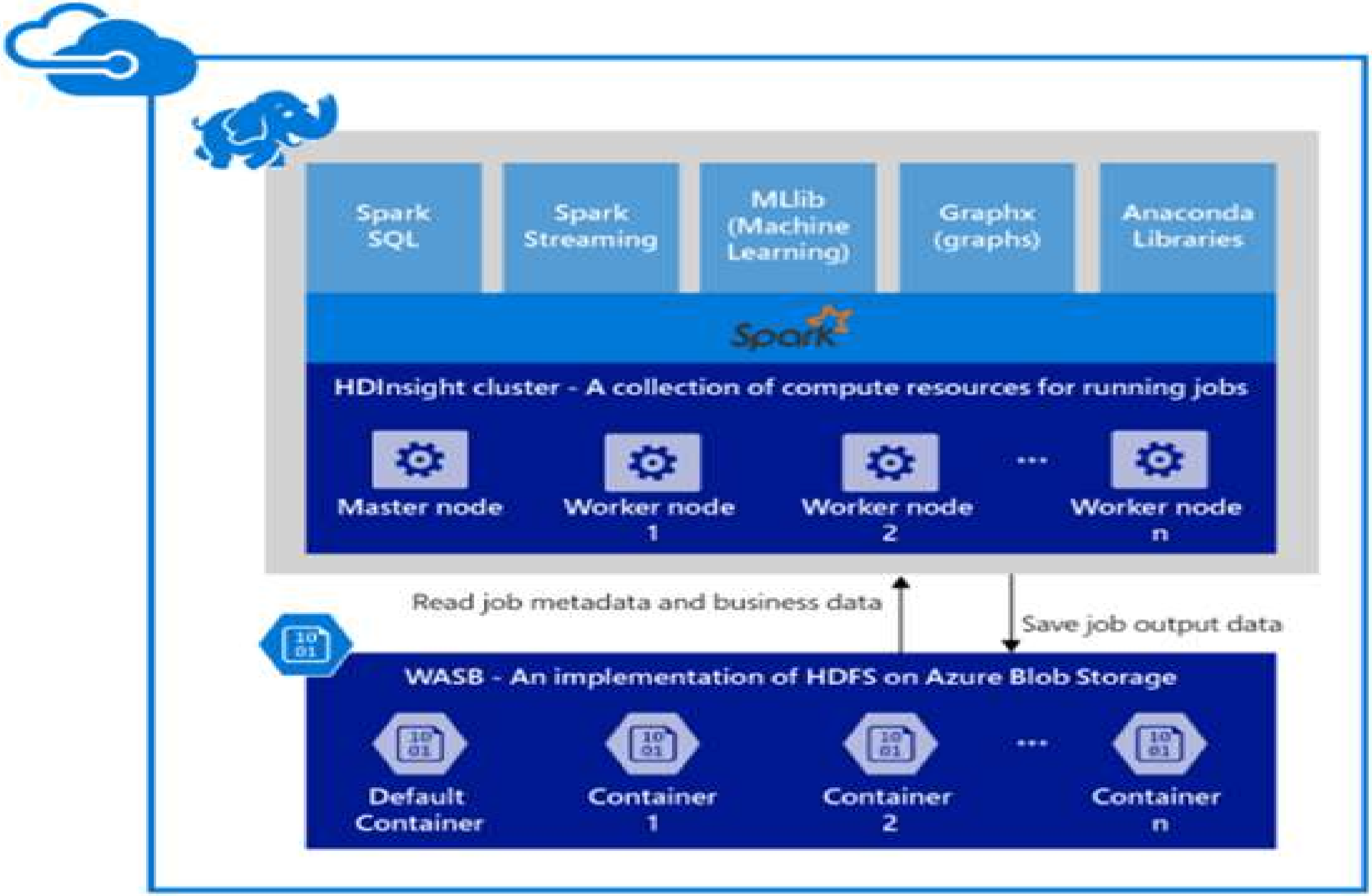
Apache Spark

- 2009 年開發
- 2010 年 Open Source
- 2013 年捐給 Apache Software Foundation



Apache Spark Tutorial | Spark Tutorial For Beginners | Apache Spark Architecture | Simplilearn

Apache Spark on Azure HDInsight



參考資料：大象跳舞系列之Spark on HDInsight (1)

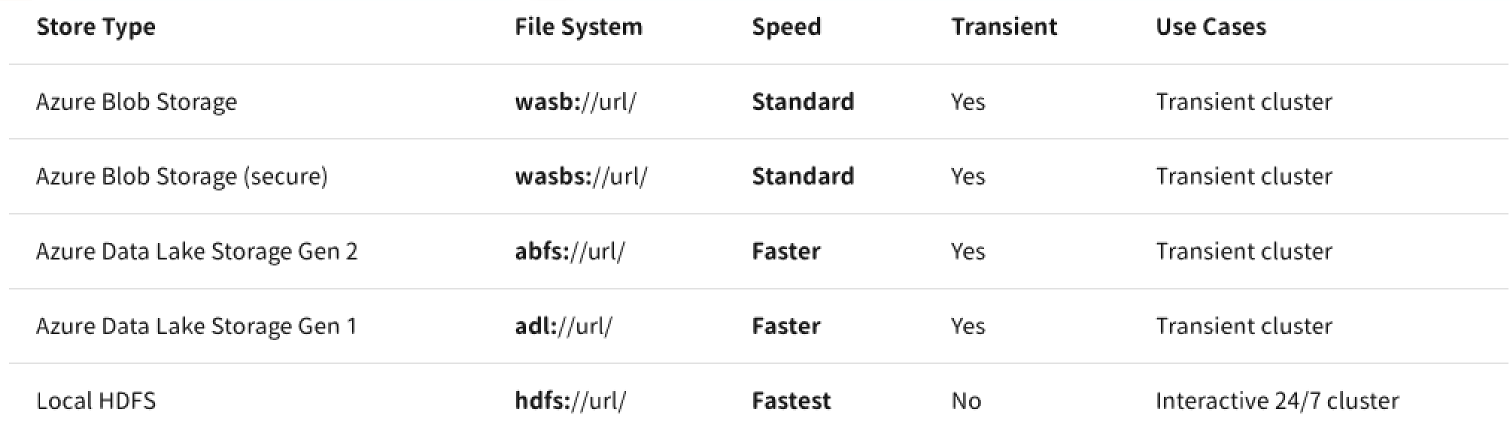
Azure Storage Type for Apache Spark

`wasb`：

- Windows Azure Storage Blob
- HDFS File System Driver over Azure Blob Storage

`abfs`：

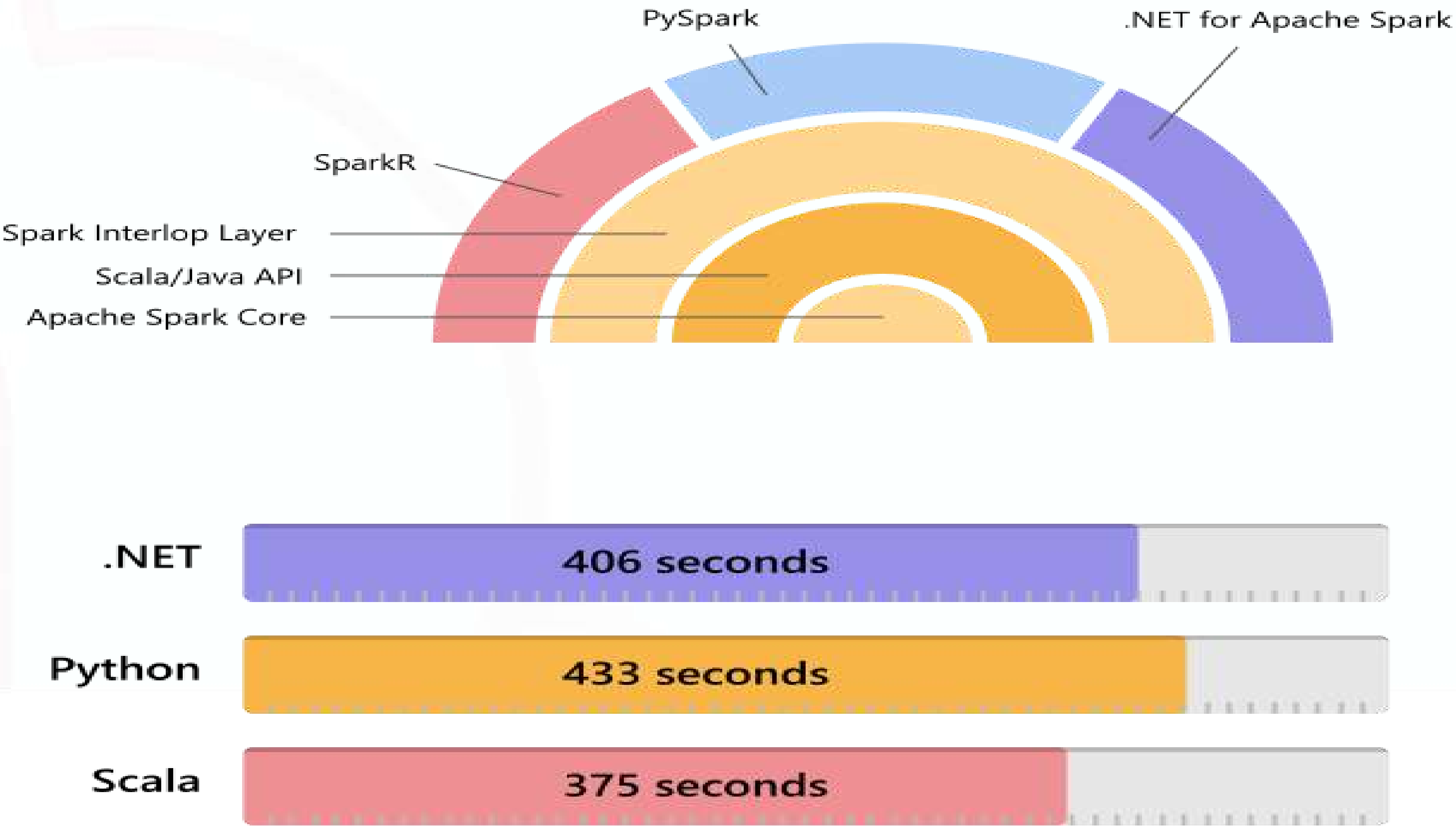
- Azure Blob File System
- HDFS File System Driver Compatible with Azure Data Lake Storage Gen2



Store Type	File System	Speed	Transient	Use Cases
Azure Blob Storage	**wasb:**//url/	**Standard**	Yes	Transient cluster
Azure Blob Storage (secure)	**wasbs:**//url/	**Standard**	Yes	Transient cluster
Azure Data Lake Storage Gen 2	**abfs:**//url/	**Faster**	Yes	Transient cluster
Azure Data Lake Storage Gen 1	**adl:**//url/	**Faster**	Yes	Transient cluster
Local HDFS	**hdfs:**//url/	**Fastest**	No	Interactive 24/7 cluster

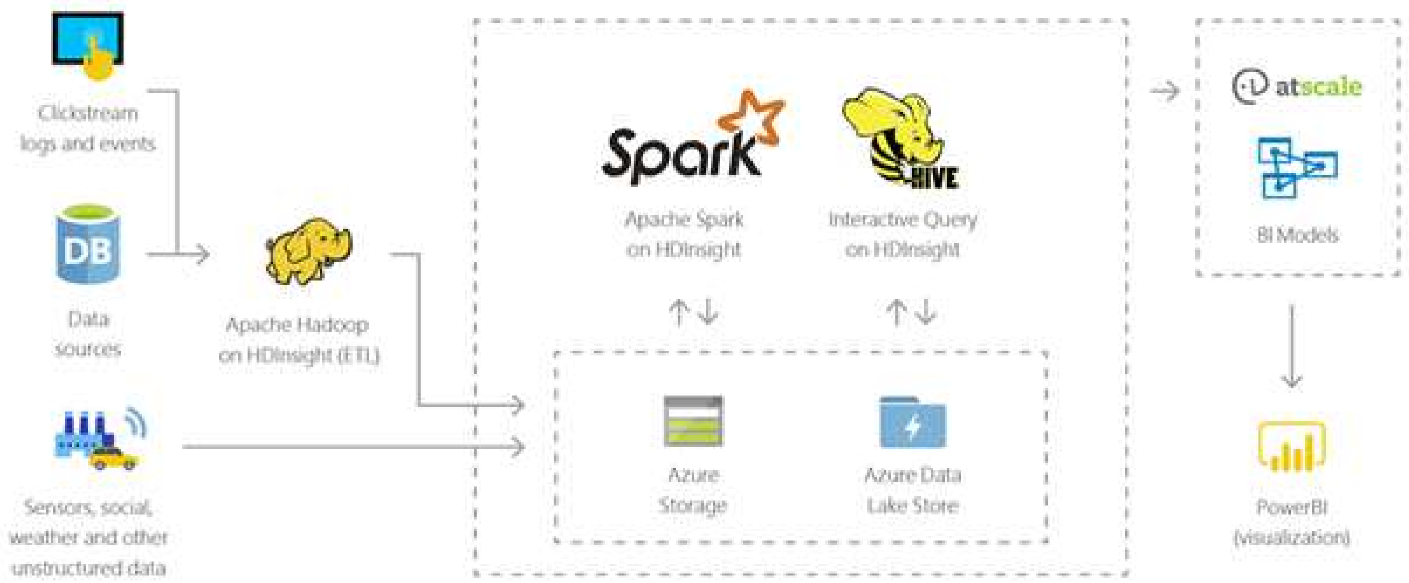
參考資料：Data storage optimization for Apache Spark

.NET for Apache Spark



參考資料：.NET for Apache® Spark™

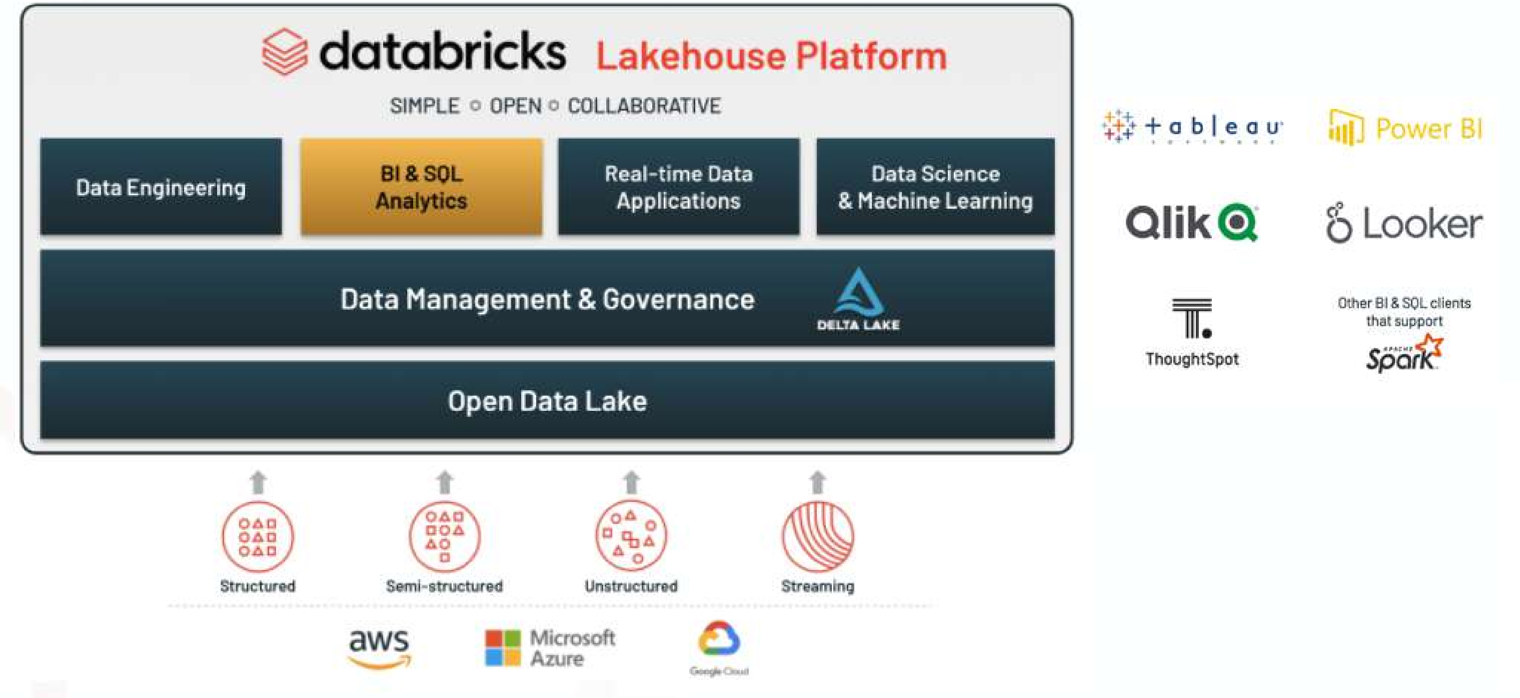
Azure HDInsight and Datawarehouse


Clickstream logs and events
DB
Data sources
Sensors, social, weather and other unstructured data
Apache Hadoop on HDInsight (ETL)
Spark
Apache Spark on HDInsight
HIVE
Interactive Query on HDInsight
Azure Storage
Azure Data Lake Store
atscale
BI Models
PowerBI (visualization)

參考資料：What is Azure HDInsight?

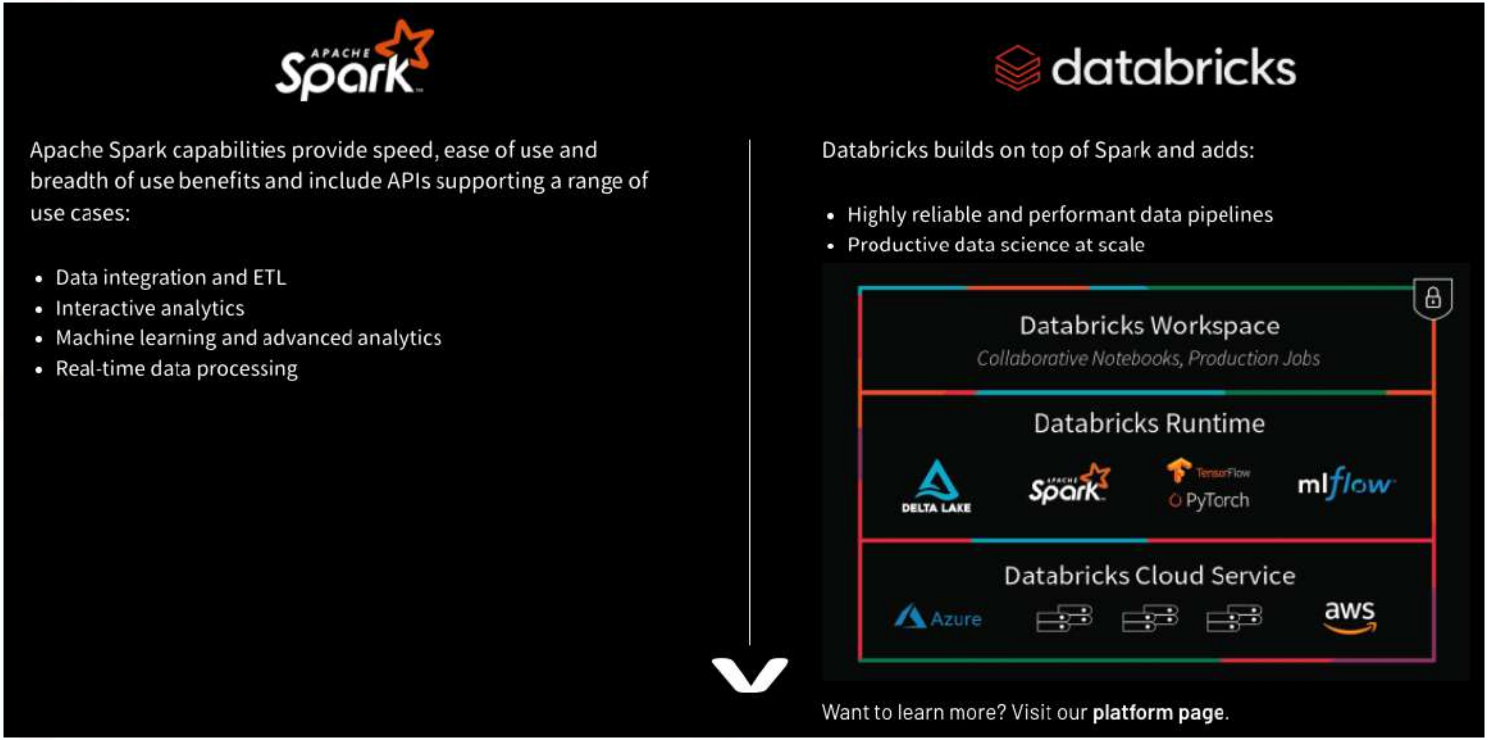
Databricks

- 由 Apache Spark 原始開發者建立
- 同時也開發 Delta Lake、MLflow、與 Koalas 等知名 Open Source 專案
- 2017 年 11 月推出與 Microsoft Azure 整合的 Azure Databricks 平台


databricks Lakehouse Platform
SIMPLE ○ OPEN ○ COLLABORATIVE
Data Engineering
BI & SQL Analytics
Real-time Data Applications
Data Science & Machine Learning
Data Management & Governance
DELTA LAKE
Open Data Lake
Structured
Semi-structured
Unstructured
Streaming
aws
Microsoft Azure
Google Cloud
tableau
Power BI
Qlik
Looker
ThoughtSpot
Other BI & SQL clients that support Spark

參考資料：Announcing the Launch of Databricks SQL

Apache Spark vs. Databricks



Apache Spark capabilities provide speed, ease of use and breadth of use benefits and include APIs supporting a range of use cases:

- Data integration and ETL
- Interactive analytics
- Machine learning and advanced analytics
- Real-time data processing

Databricks builds on top of Spark and adds:

- Highly reliable and performant data pipelines
- Productive data science at scale

Databricks Workspace

Collaborative Notebooks, Production Jobs

Databricks Runtime

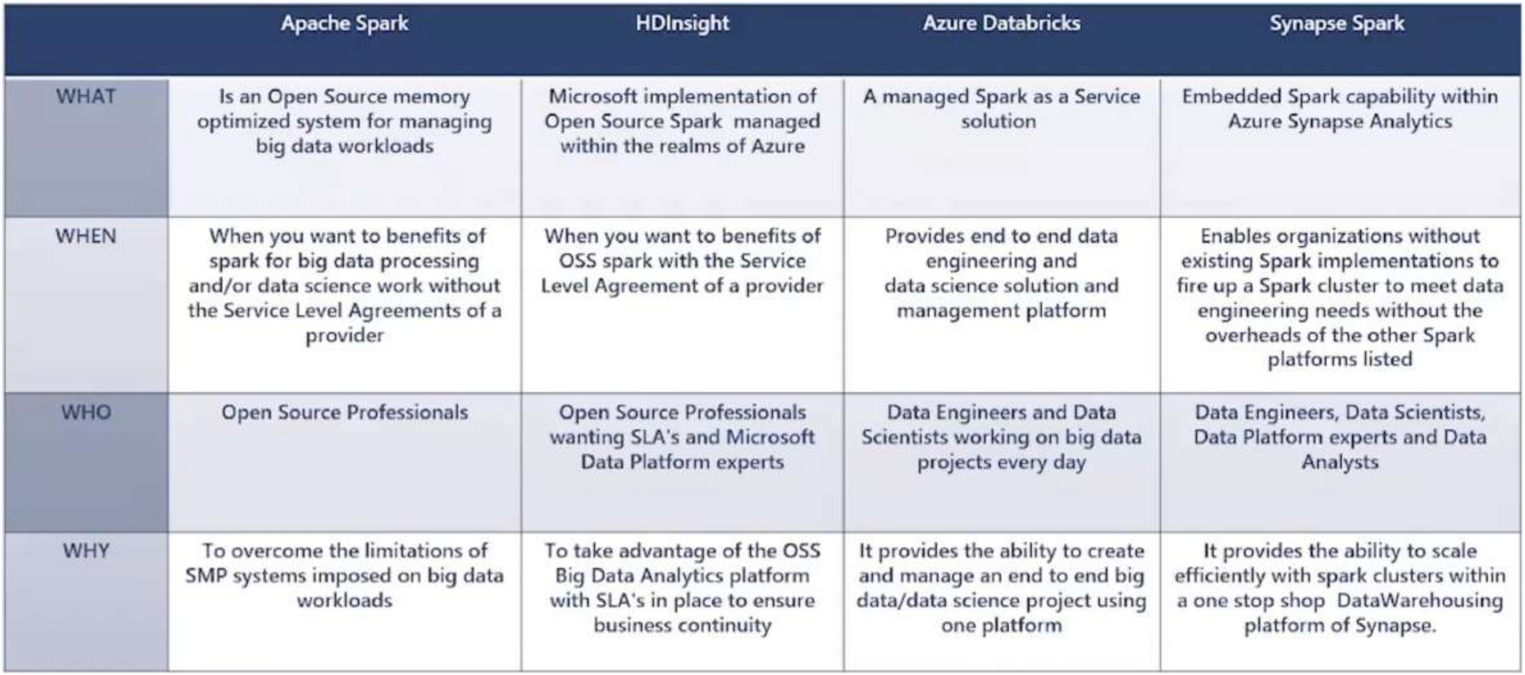
Databricks Cloud Service

Want to learn more? Visit our **platform page**.

參考資料：Comparing Apache Spark™ and Databricks

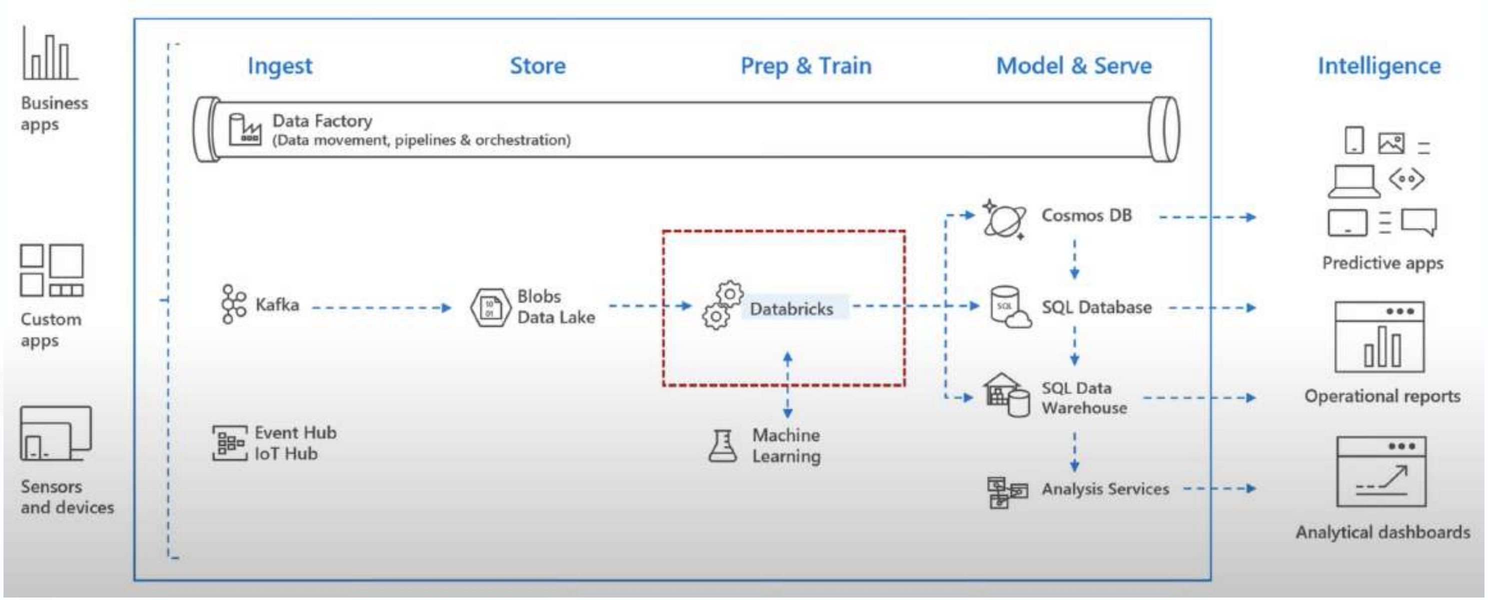
Spark Comparisons

Apache Spark vs. Azure HDInsight vs. Azure Databricks vs. Synapse Spark：



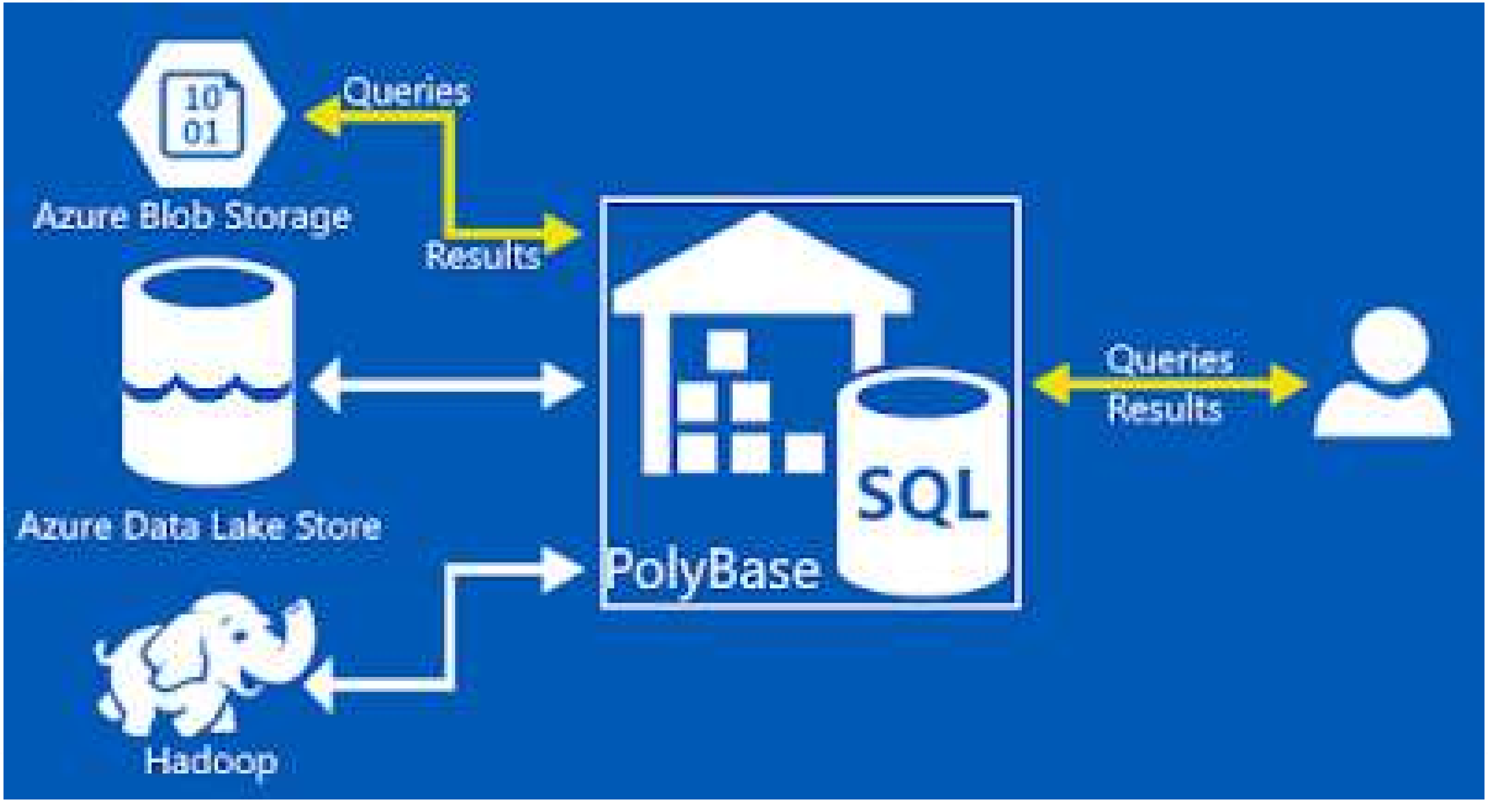
	Apache Spark	HDInsight	Azure Databricks	Synapse Spark
WHAT	Is an Open Source memory optimized system for managing big data workloads	Microsoft implementation of Open Source Spark managed within the realms of Azure	A managed Spark as a Service solution	Embedded Spark capability within Azure Synapse Analytics
WHEN	When you want to benefits of spark for big data processing and/or data science work without the Service Level Agreements of a provider	When you want to benefits of OSS spark with the Service Level Agreement of a provider	Provides end to end data engineering and data science solution and management platform	Enables organizations without existing Spark implementations to fire up a Spark cluster to meet data engineering needs without the overheads of the other Spark platforms listed
WHO	Open Source Professionals	Open Source Professionals wanting SLA's and Microsoft Data Platform experts	Data Engineers and Data Scientists working on big data projects every day	Data Engineers, Data Scientists, Data Platform experts and Data Analysts
WHY	To overcome the limitations of SMP systems imposed on big data workloads	To take advantage of the OSS Big Data Analytics platform with SLA's in place to ensure business continuity	It provides the ability to create and manage an end to end big data/data science project using one platform	It provides the ability to scale efficiently with spark clusters within a one stop shop DataWarehousing platform of Synapse.

Azure Databricks and Data Warehouse



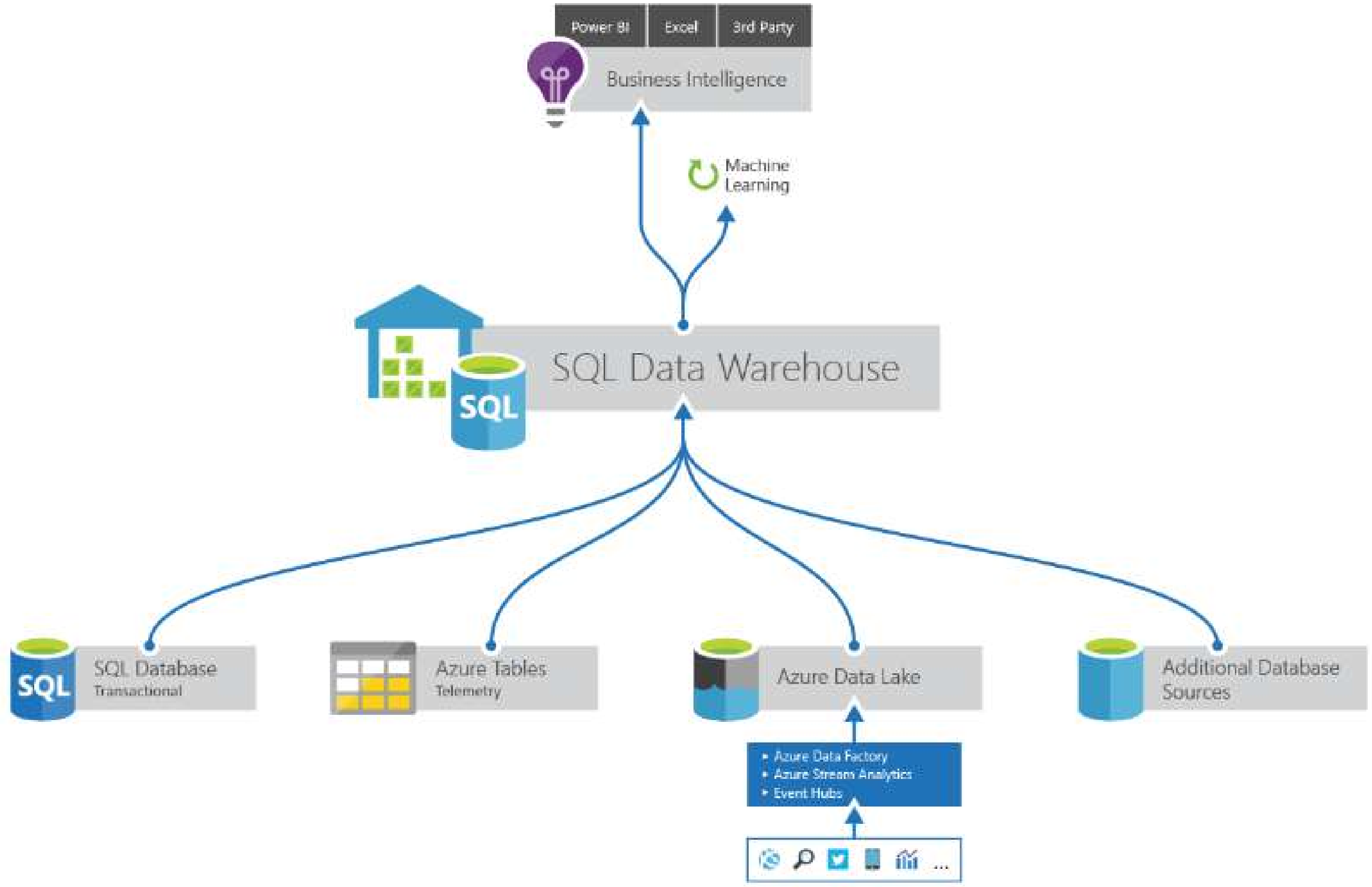
參考資料：How to connect Databricks to Azure Data Lake?

Microsoft SQL Server and PolyBase



參考資料：Use PolyBase to read Blob Storage in Azure SQL DW

Azure SQL Data Warehouse (SQL DW)



參考資料：Azure SQL Data Warehouse

Data Lake Evolution

Database / Data Warehouse / Data Lake

Database：

- 比較 Operational 的 Data
- Data 可以新增刪除修改

Data Warehouse：

- 比較 Analytical 的 Data，Quality 比較好的 Data
- Data 原則上是處理過的，不會刪除，不會修改，只會增加
- 通常是 Database 的 Data 累積一段時間之後，再轉到 Data Warehouse
- Subject-Oriented、Integrated、Time-Variant、與 Non-Volatile
- 是一種結構化的、可以分析查詢的 Transactional Data Copy
- 經常用來產生 Report 或 Dashboard

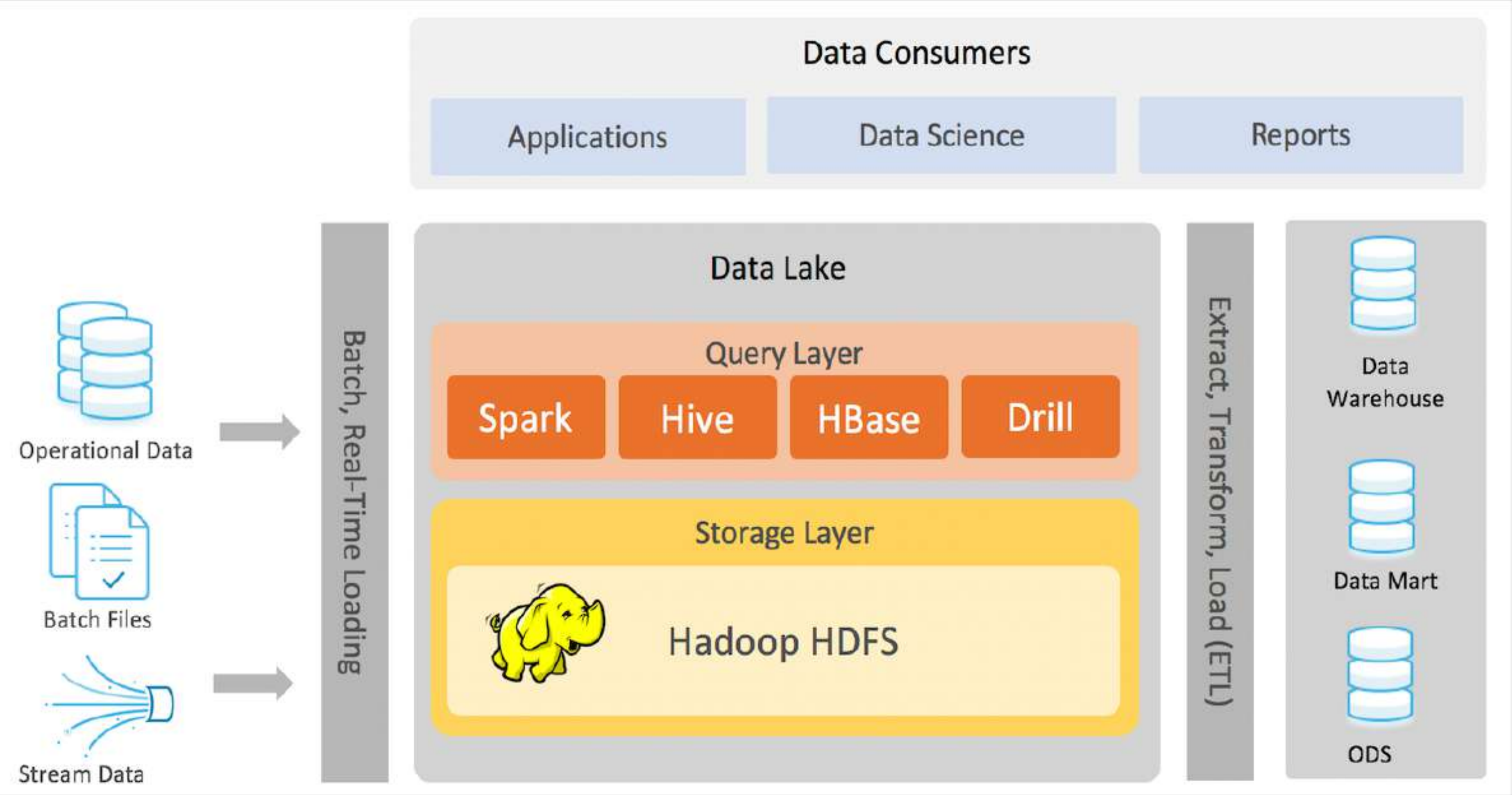
Data Lake：

- Repositories for Big Data in Its Native Form
- 儲存各種 Un-Structured Data / Semi-Structured / Structured Data 的大型 Data Storage
- 不預先定義 Data 儲存的目的，或者說儲存就是它的目的
- 會比較在乎成本效益

參考資料：資料倉儲的類型

Hadoop Data Lake Architecture

- Hadoop 是常見的一種 Data Lake 實現方式



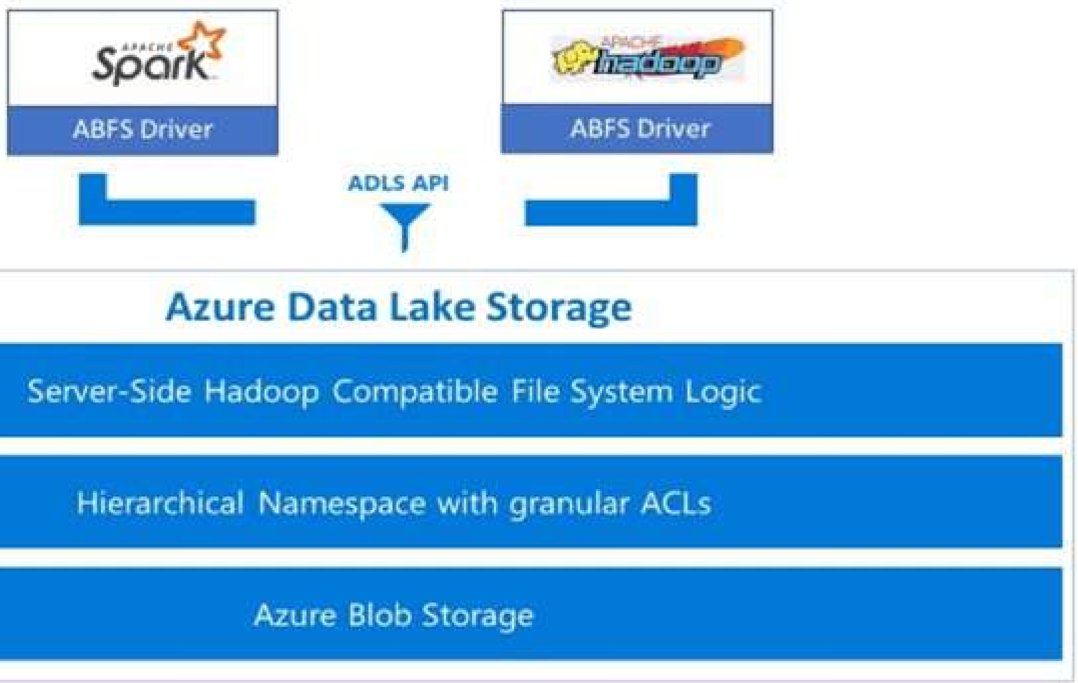
參考資料：Webinar: Data Lake Advances in the Age of Operational ML/AI

Browsersync: connected

Azure Data Lake Storage Architecture

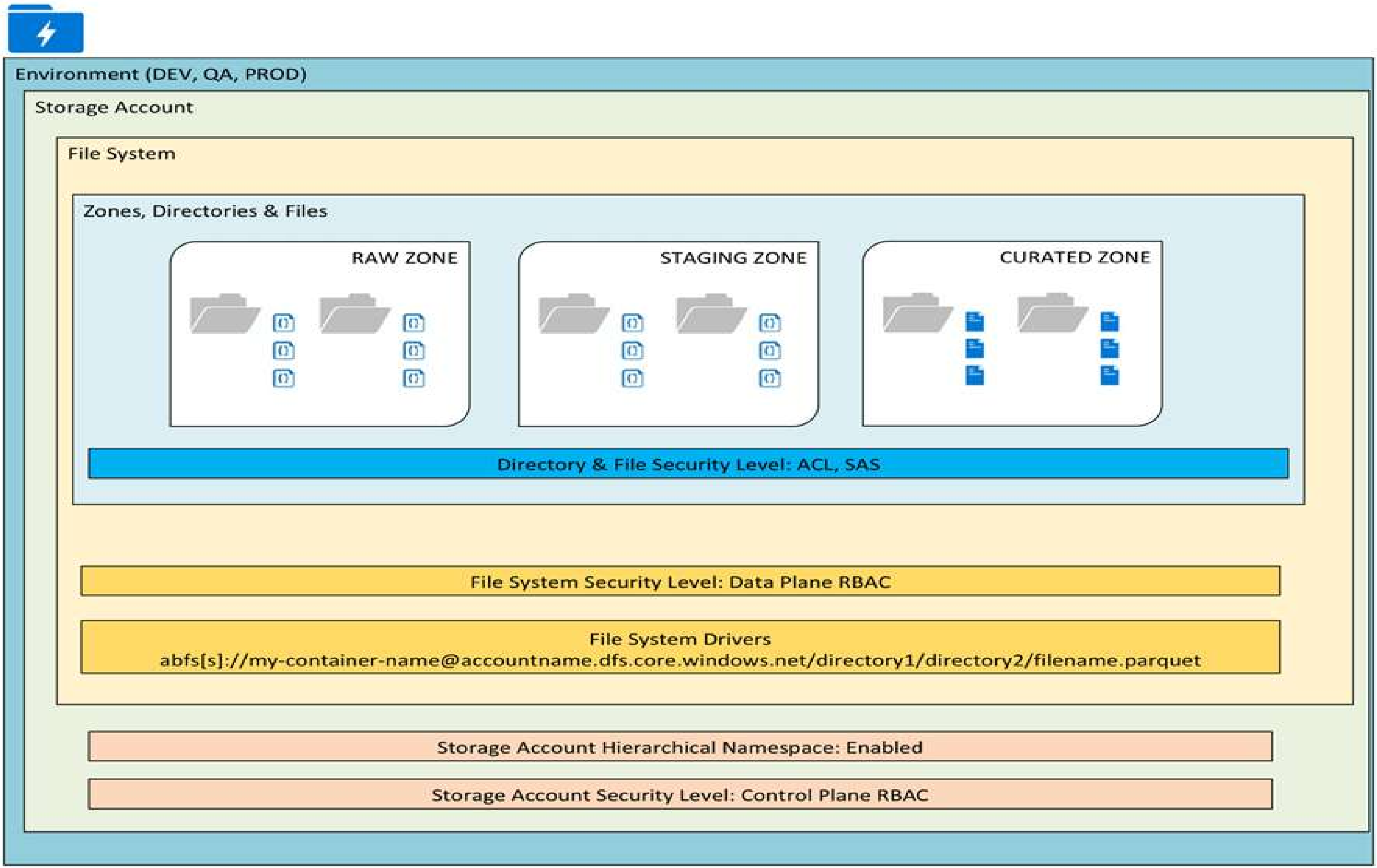
- File System Support
- Encryption of Data in Transit
- Encryption of Data at Rest
- Storage Account Firewall

- Compatible with HDFS
- Virtual Network Integration
- Role-Based Access Security
- Hierarchical Namespace



參考資料：What's new in Azure Data Lake Storage Gen2

Azure Data Lake Storage Layered Design

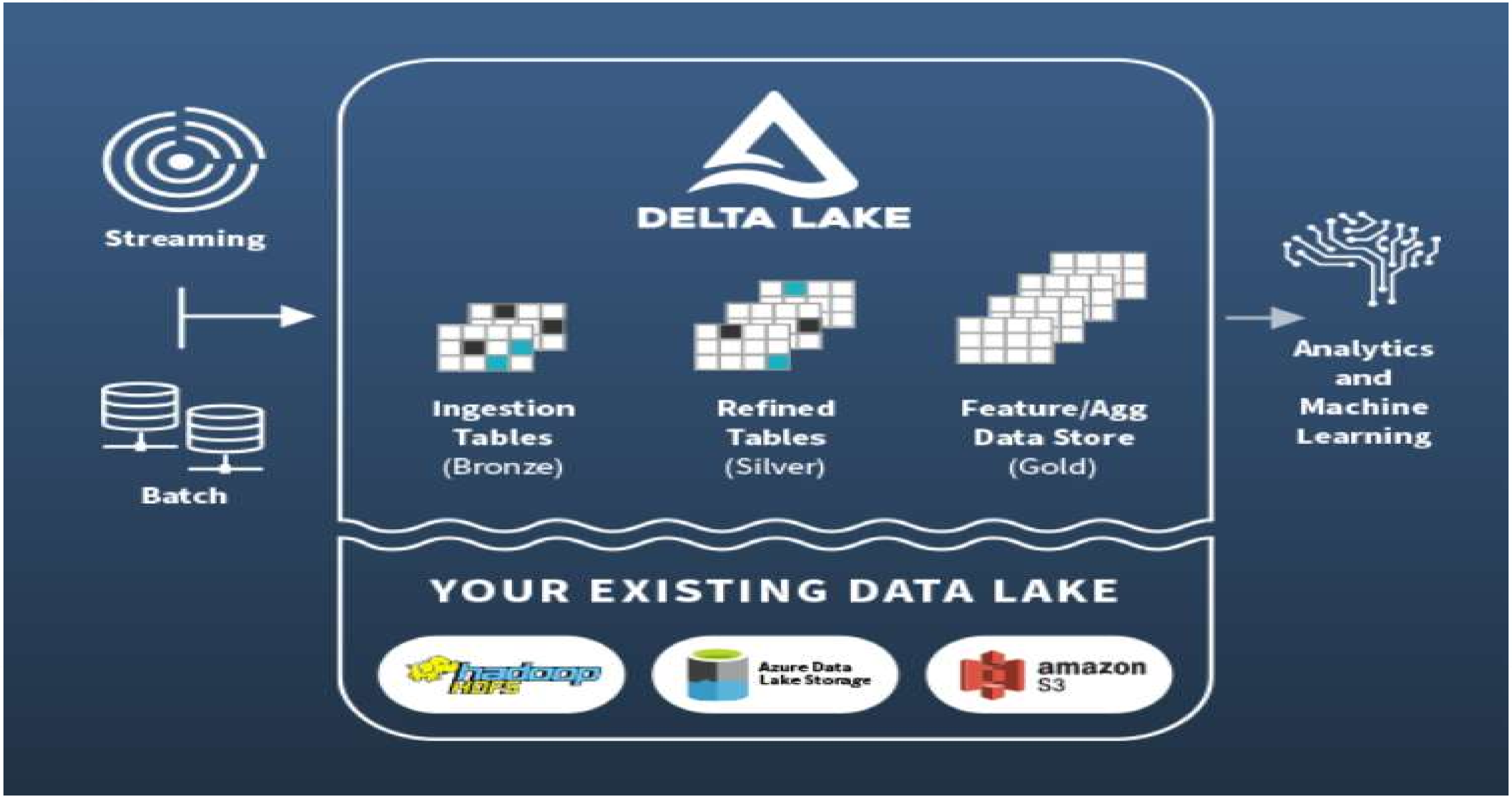

Environment (DEV, QA, PROD)
Storage Account
File System
Zones, Directories & Files
RAW ZONE
STAGING ZONE
CURATED ZONE
Directory & File Security Level: ACL, SAS
File System Security Level: Data Plane RBAC
File System Drivers
abfs[s]://my-container-name@accountname.dfs.core.windows.net/directory1/directory2/filename.parquet
Storage Account Hierarchical Namespace: Enabled
Storage Account Security Level: Control Plane RBAC

參考資料：Designing an Azure Data Lake Store Gen2

Browsersync: connected

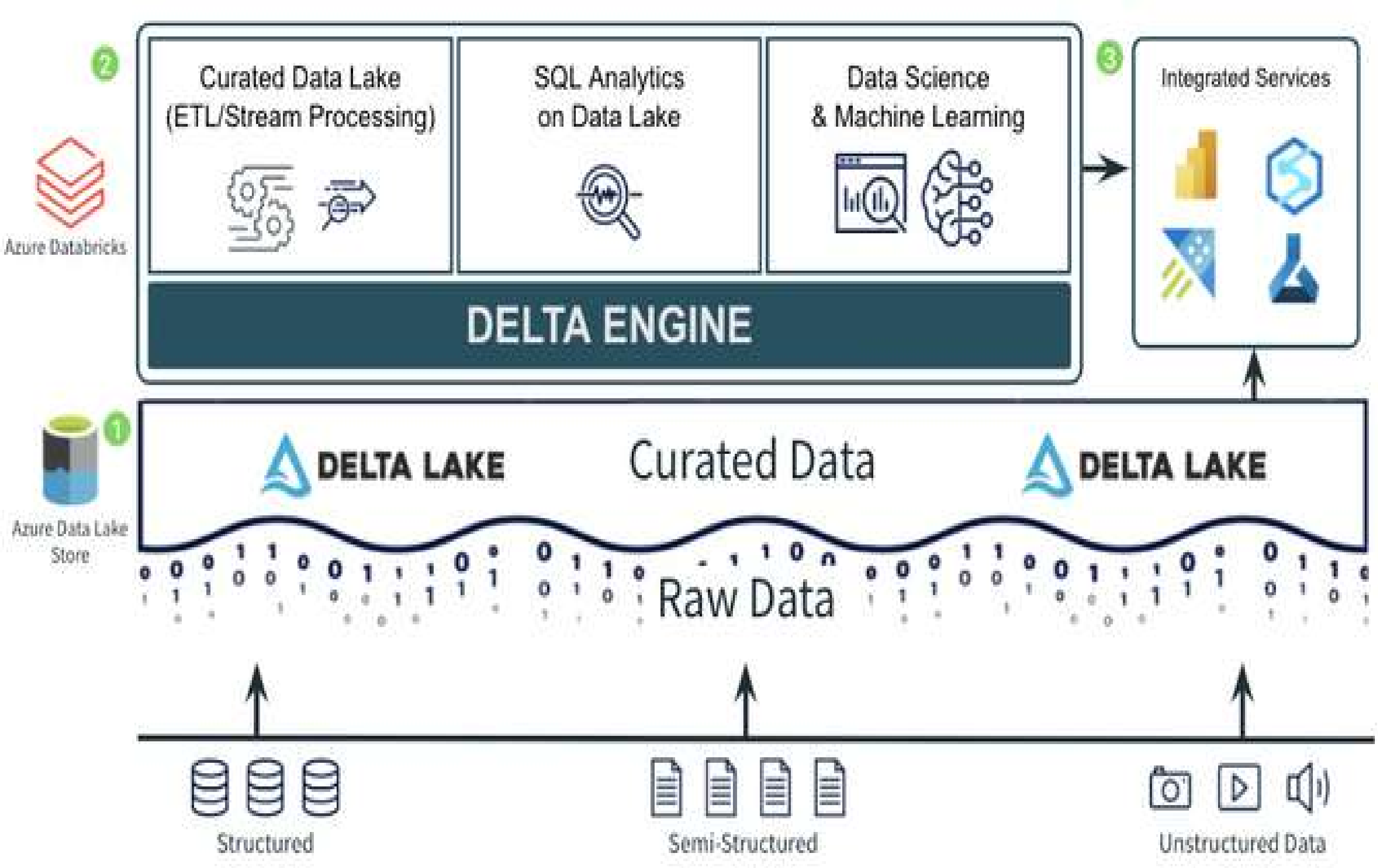
Databricks Delta Data Lake Architecture

- Databricks 的 Open Source Data Lake 實現方式



參考資料：Data Lake or Warehouse? Databricks Offers a Third Way

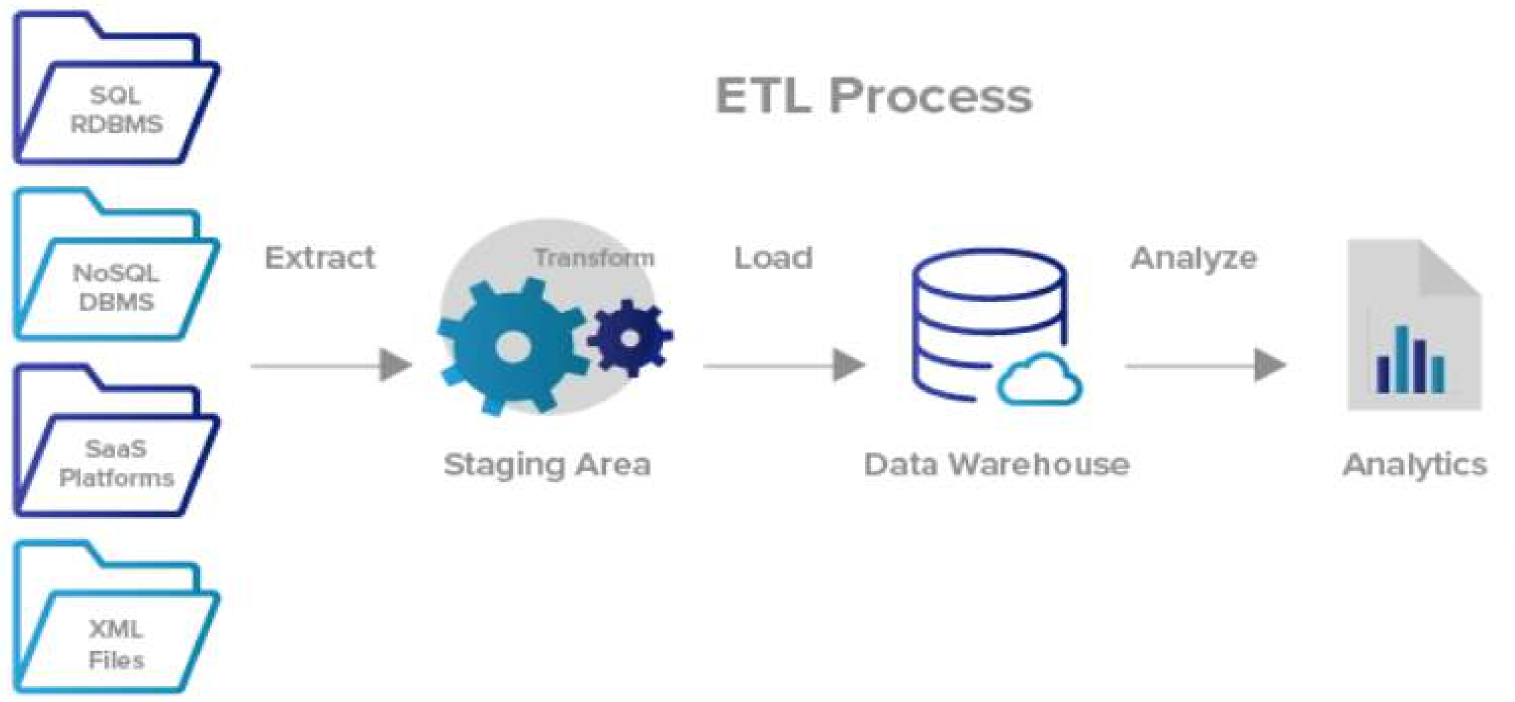
Data Lake + Warehouse = Lakehouse



參考資料：Simplify Your Lakehouse Architecture with Azure Databricks, Delta Lake, and Azure Data Lake Storage

Extract - Transform - Load (ETL)

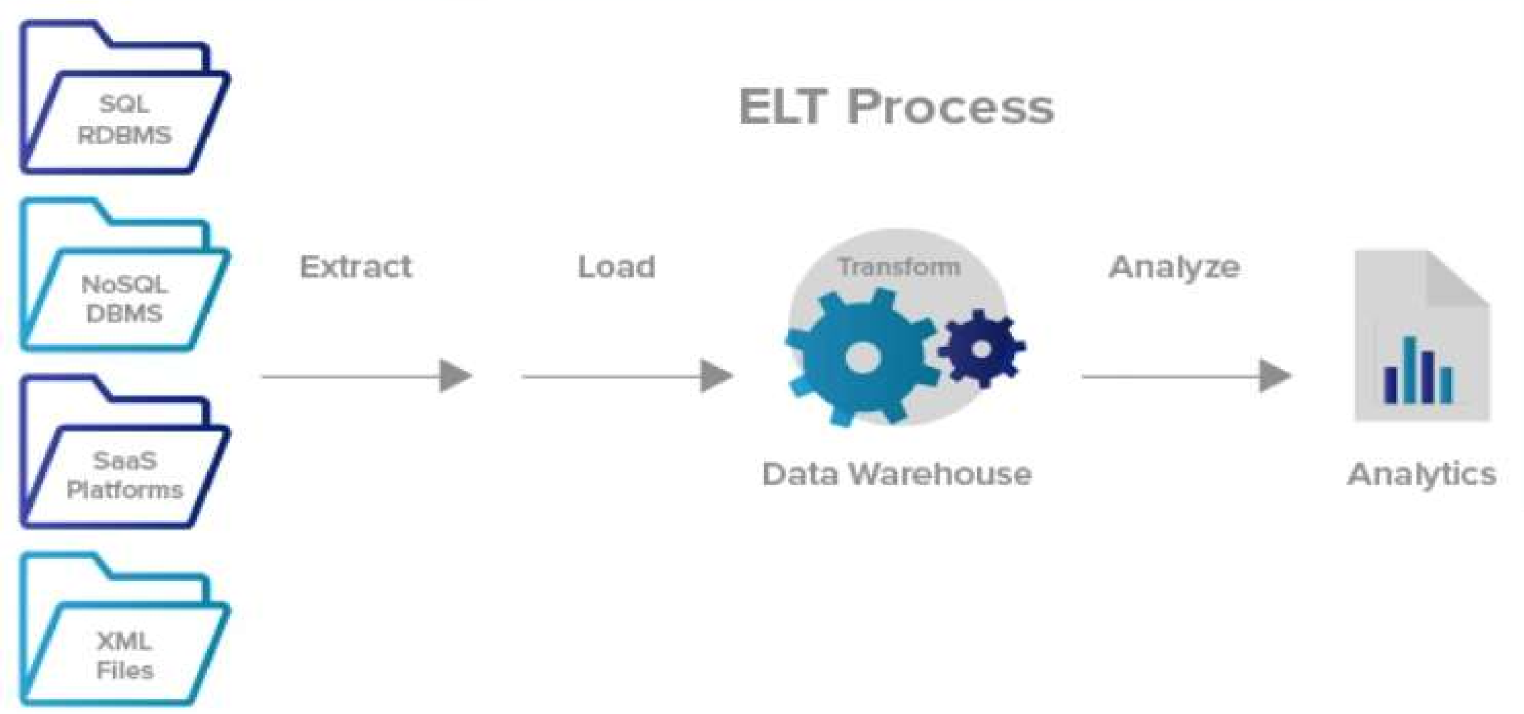
- 習慣上會先把 Data Extract 到所謂的 Staging Area，進行 Transform，再 Load 到 Data Warehouse 裡頭
- 通常處理 On-Premises 的 Relational Structured Data
- 比較不支援 Data Lake 的作法



參考資料：ETL vs ELT: 5 Critical Differences

Extract - Load - Transform (ELT)

- Data 直接 Extract 之後就 Load 到 Data Warehouse，然後在 Data Warehouse 裡面進行 Transform
- 比較貼近在 Cloud 同時處理 Structured 與 Unstructured Data 的想法
- 比較支援 Data Lake 的作法
- 也比較適合處理 Big Data 進行分析



參考資料：ETL vs ELT: 5 Critical Differences

Browsersync: connected

Azure Synapse Analytics

Azure Synapse Analytics

Limitless Analytics Service with Unmatched Time to Insight.

- SQL DW：Data Warehouse
- Apache Spark / Azure Databricks：Big Data Analytics
- Azure Synapse：Data Ware House + Big Data Analytics

Azure Synapse
Structured and unstructured data
Data warehousing and big data analytics
Cloud-native

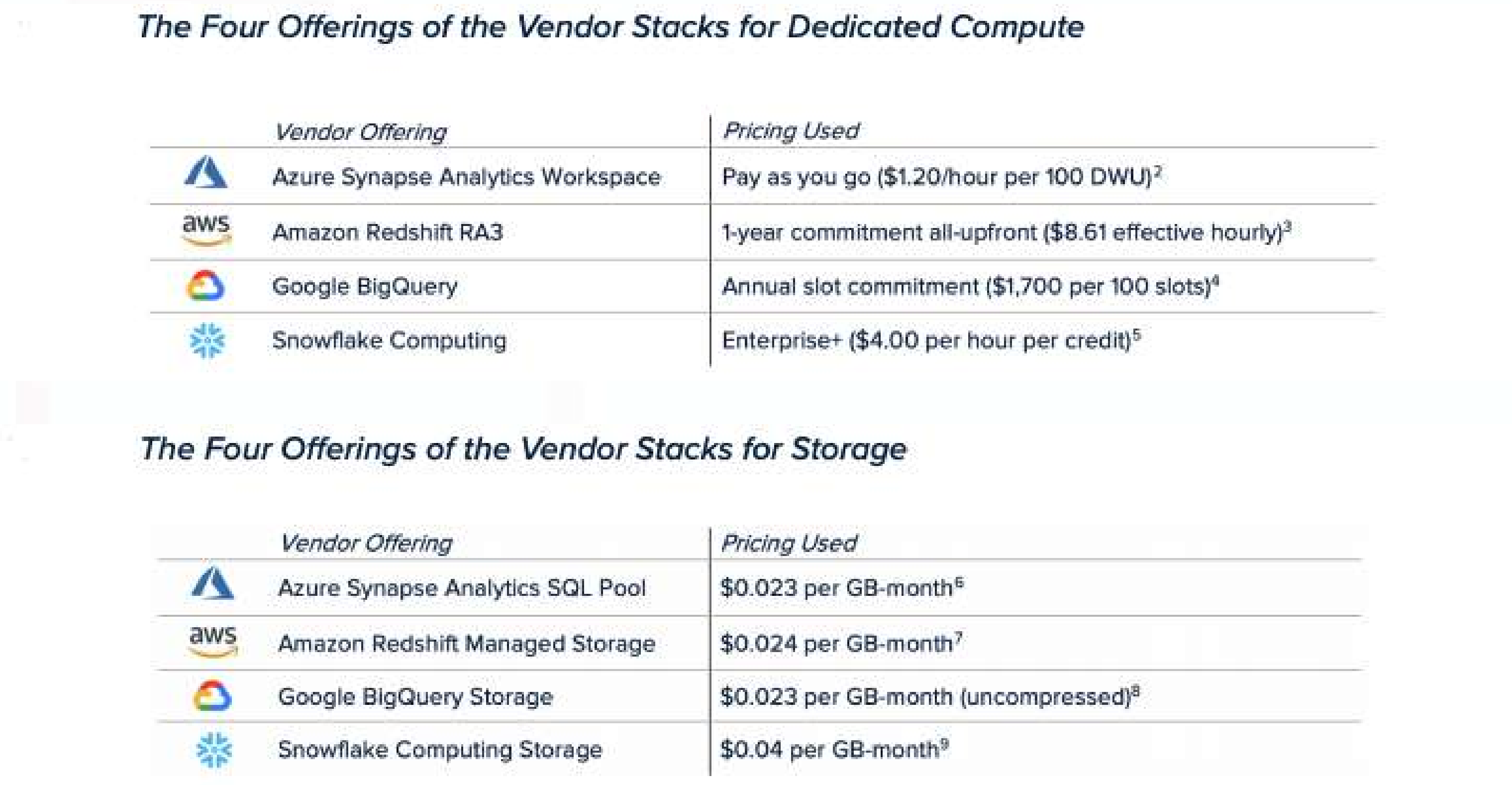
參考資料：微軟推出Azure Synapse Analytics以串連分析系統、資料倉儲

Azure Synapse Studio

- Web-Based Single Hub of Azure Synapse Analytics
- Ingest / Explore / Prepare / Train / Analyze / Visualize Data
- 不需要知道 Data Schema 就可以進行 Data Exploration
- 透過 Knowledge Center 提供無止盡的 Learning Material / Sample Code / Template / Azure Open Dataset / Pipeline
- 內建 Pipeline Template 可以整合各種 Data Source
- 可以建立 Data Pipeline 進行 ETL
- 內建 AI / ML 功能，不需要另外整合 Azure ML 或是 Cognitive Services
- 以前的 SQL DW 變成現在 Built-in 的 Serverless SQL Pool
- 可以根據專案需求自己建立一或多個 Dedicated SQL Pool
- SQL Pool 以 Columnar Format 儲存 Data，搭配 `CLUSTERED COLUMNSTORE INDEX` 會有更好的 Performance
- SQL Pool 直接可以使用 `PREDICT` Function 在 Query 中進行 Prediction
- Apache Spark Pool 可以使用 C# / Python / Scala / Spark SQL 語言，支援 Spark ML 與 Spart Streaming
- 整合 AutoML，而且可以直接使用 Serverless Apache Spark Pool
- 可以存取 Azure Machine Learning Model Registry 直接進行 Model Scoring，不需要將 Data 搬進搬出
- 透過 Cognitive Services 可以 Enrich Data
- 支援與 Jupyter 相同的 Notebook 格式
- 整合 Power BI，可以直接使用 Power BI 的 Dataset / Report / Dashboard 功能

Why Azure Synapse Analytics

Implementing an end-to-end analytics solution in Azure costs up to 13 percent less compared to AWS, up to 49 percent less compared to Google.



The Four Offerings of the Vendor Stacks for Dedicated Compute

	Vendor Offering	*Pricing Used*
	Azure Synapse Analytics Workspace	Pay as you go ($1.20/hour per 100 DWU)[2]
aws	Amazon Redshift RA3	1-year commitment all-upfront ($8.61 effective hourly)[3]
	Google BigQuery	Annual slot commitment ($1,700 per 100 slots)[4]
	Snowflake Computing	Enterprise+ ($4.00 per hour per credit)[5]

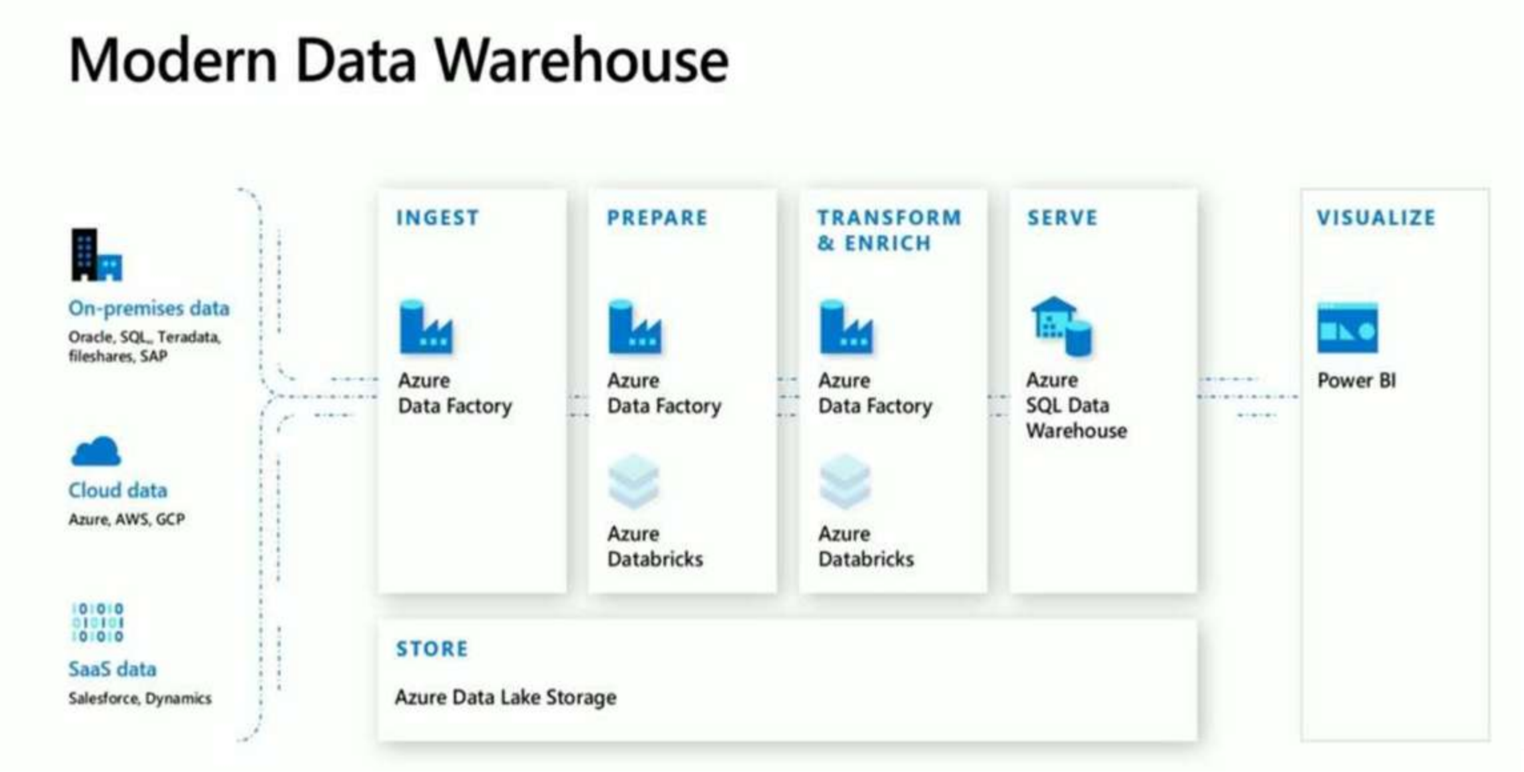
The Four Offerings of the Vendor Stacks for Storage

	Vendor Offering	*Pricing Used*
	Azure Synapse Analytics SQL Pool	$0.023 per GB-month[6]
aws	Amazon Redshift Managed Storage	$0.024 per GB-month[7]
	Google BigQuery Storage	$0.023 per GB-month (uncompressed)[8]
	Snowflake Computing Storage	$0.04 per GB-month[9]

參考資料：Cloud Analytics Platform Total Cost of Ownership
參考資料：Analytics in Azure is up to 14x faster and costs 94% less than other cloud providers. Why go anywhere else?

Browsersync: connected

Before vs. After Azure Synapse Analytics



Modern Data Warehouse

On-premises data
Oracle, SQL, Teradata, fileshares, SAP

Cloud data
Azure, AWS, GCP

SaaS data
Salesforce, Dynamics

INGEST
Azure Data Factory

PREPARE
Azure Data Factory
Azure Databricks

TRANSFORM & ENRICH
Azure Data Factory
Azure Databricks

SERVE
Azure SQL Data Warehouse

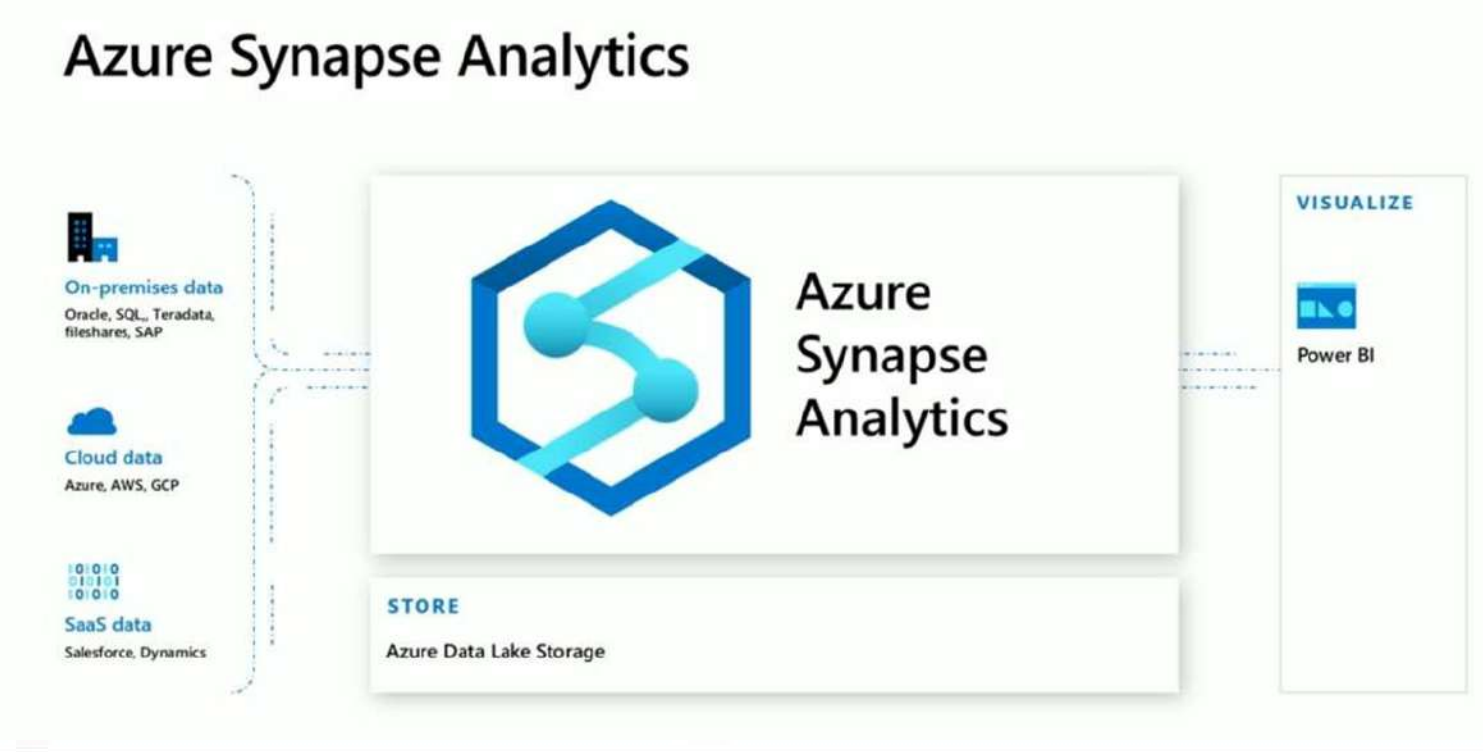
VISUALIZE
Power BI

STORE
Azure Data Lake Storage

參考資料：Microsoft Ignite postmortem: Cutting through the complexity

Browsersync: connected

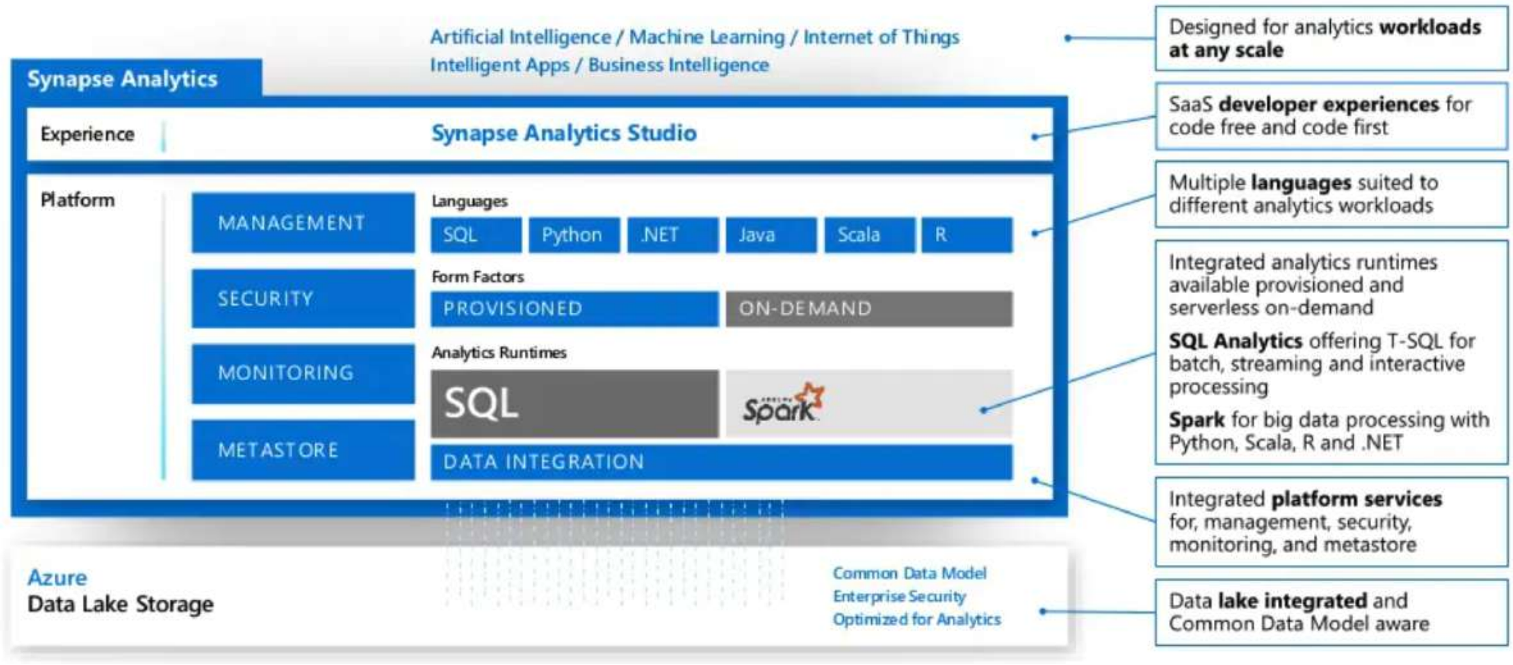
Before vs. After Azure Synapse Analytics



參考資料：Microsoft Ignite postmortem: Cutting through the complexity

Azure Synapse Analytics Architecture

- Data Storage 集中存放在 Azure Data Lake Storage Gen2 之上
- ETL/ELT 提供比 Azure Data Factory 更簡單好用的 Synapse Pipeline
- Data Analysis 提供 SQL-Based Runtime 與 Apache Spark Runtime
- Security 直接整合 Azure Active Directory
- 透過 Linked Service 架構整合各種 Data Source 與 Cloud Service
- 所有一切功能都整合在 Synapse Studio 之內



Designed for analytics **workloads at any scale**

SaaS **developer experiences** for code free and code first

Multiple **languages** suited to different analytics workloads

Integrated analytics runtimes available provisioned and serverless on-demand

SQL Analytics offering T-SQL for batch, streaming and interactive processing

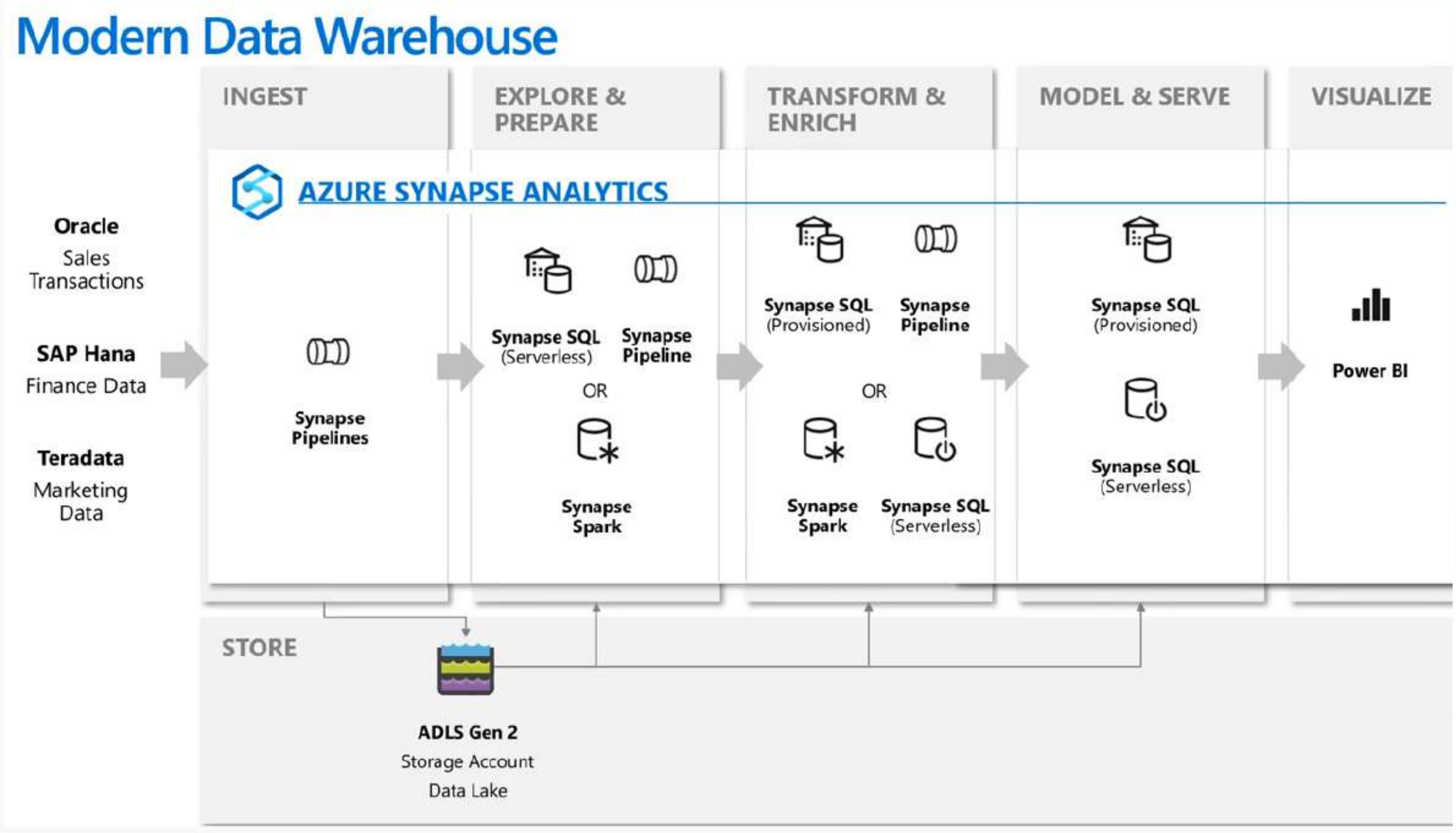
Spark for big data processing with Python, Scala, R and .NET

Integrated **platform services** for, management, security, monitoring, and metastore

Data **lake integrated** and Common Data Model aware

參考資料：Azure Synapse Analytics Overview (r2)

Before vs. After (Detailed Edition)

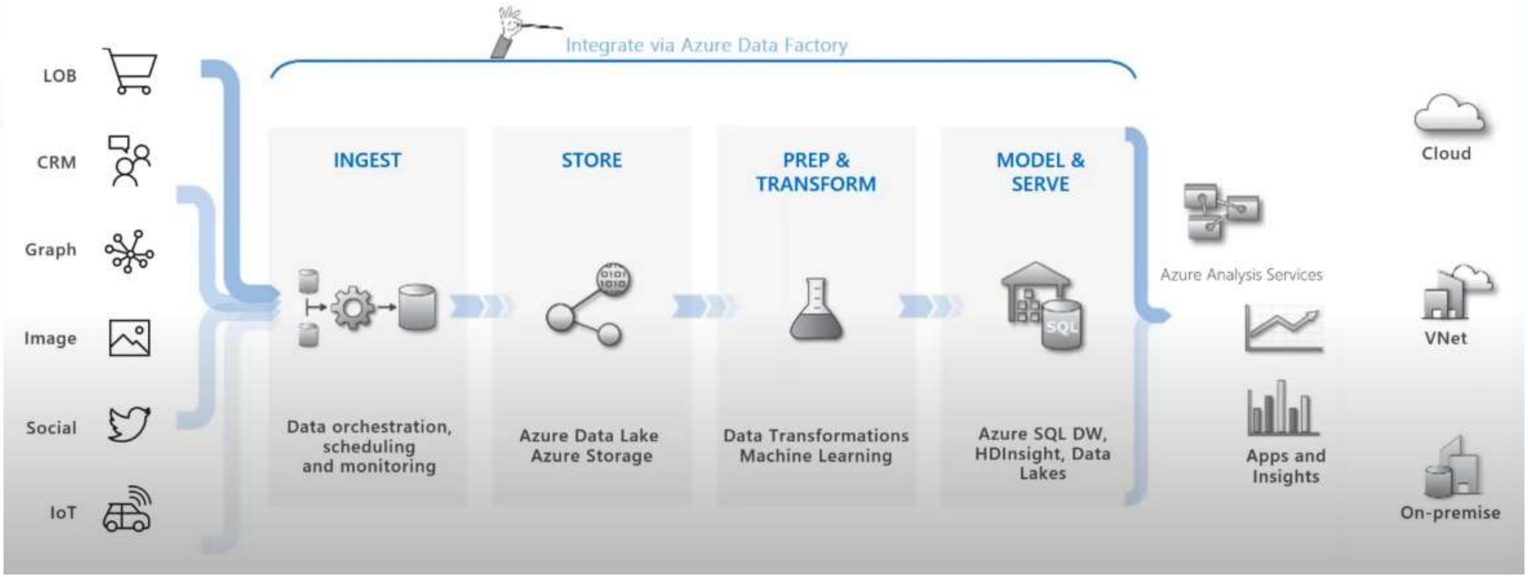


參考資料：Azure Synapse Analytics For Beginner

Browsersync: connected

Data Warehouse/Analytics Solution

- Data Source：File / File Server / Database Server / Web / IoT
- Ingest：Synapse Pipeline / SQL BULK COPY
- Store：Azure Data Lake Storage Gen2
- Prepare & Train：Apache Spark Pool
- Model & Serve：Serverless SQL Pool (Metadata) / Dedicated SQL Pool (Data)
- Apps + Insights：Power BI / Tableau / Azure Service / O365 Service / Others

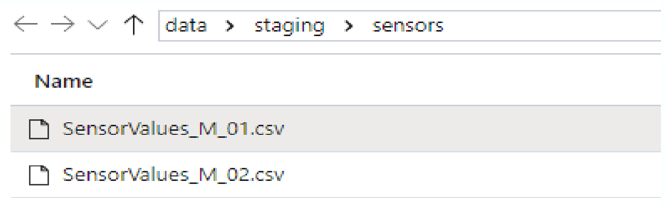


Integrate via Azure Data Factory
LOB
CRM
Graph
Image
Social
IoT
INGEST
Data orchestration, scheduling and monitoring
STORE
Azure Data Lake Azure Storage
PREP & TRANSFORM
Data Transformations Machine Learning
MODEL & SERVE
SQL
Azure SQL DW, HDInsight, Data Lakes
Azure Analysis Services
Apps and Insights
Cloud
VNet
On-premise

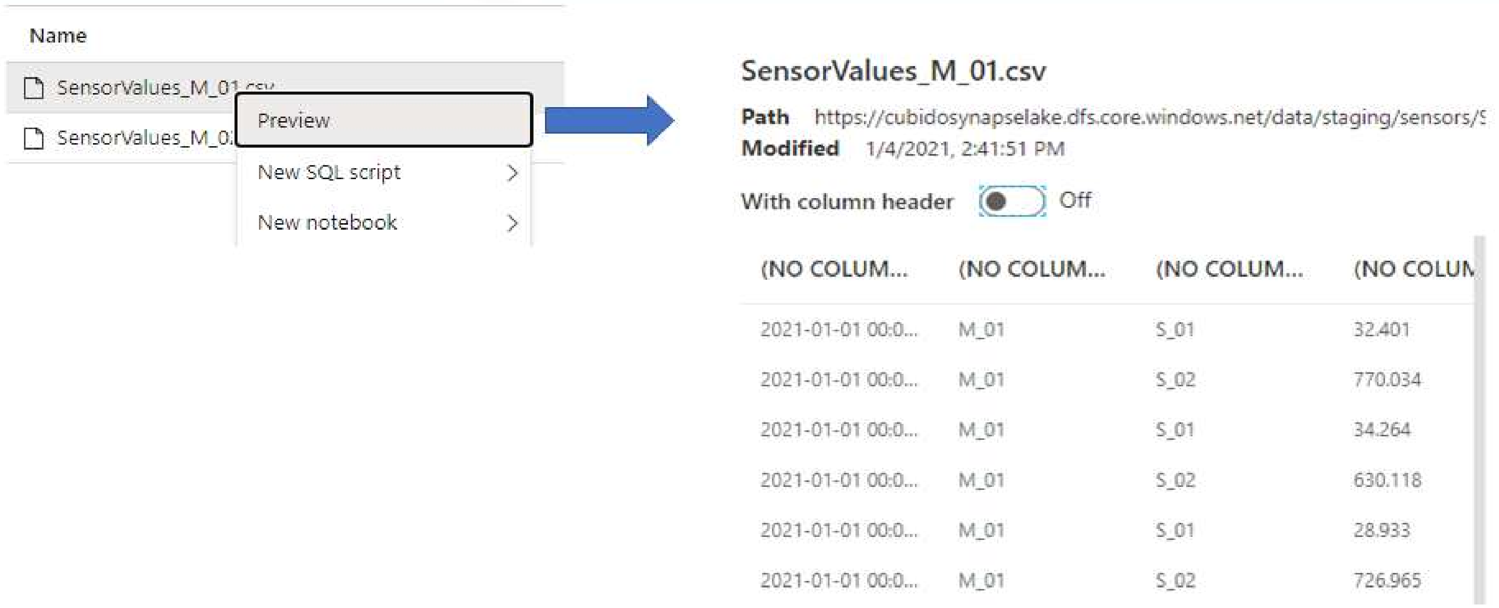
參考資料：What is dedicated SQL pool (formerly SQL DW) in Azure Synapse Analytics?

Explore Data

- 收集：Azure Stream Analytics
- 儲存：Azure Data Lake Storage
- 預覽：Synapse Studio



SensorValues_M_01.csv
SensorValues_M_02.csv



SensorValues_M_01.csv

Path https://cubidosynapselake.dfs.core.windows.net/data/staging/sensors/S

Modified 1/4/2021, 2:41:51 PM

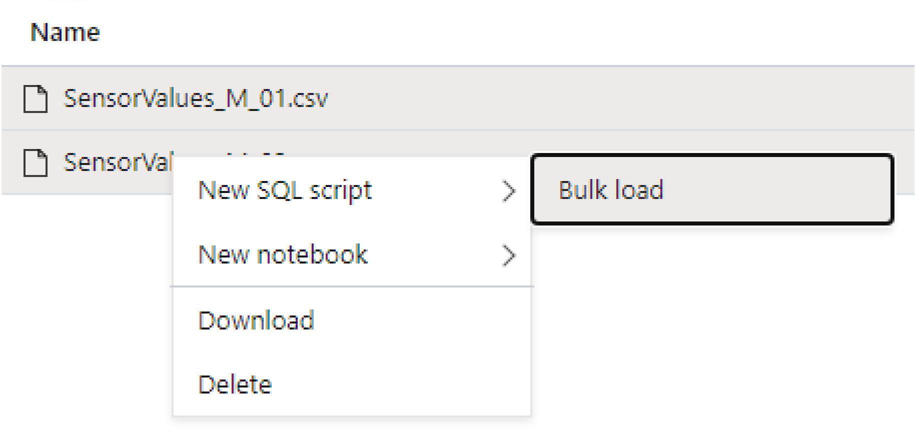
With column header Off

(NO COLUM...	(NO COLUM...	(NO COLUM...	(NO COLUM
2021-01-01 00:0...	M_01	S_01	32.401
2021-01-01 00:0...	M_01	S_02	770.034
2021-01-01 00:0...	M_01	S_01	34.264
2021-01-01 00:0...	M_01	S_02	630.118
2021-01-01 00:0...	M_01	S_01	28.933
2021-01-01 00:0...	M_01	S_02	726.965

參考資料：Azure Synapse Analytics end-to-end: From your database and sensors to a Power BI dashboard

Ingest Data

- SQL BULK COPY

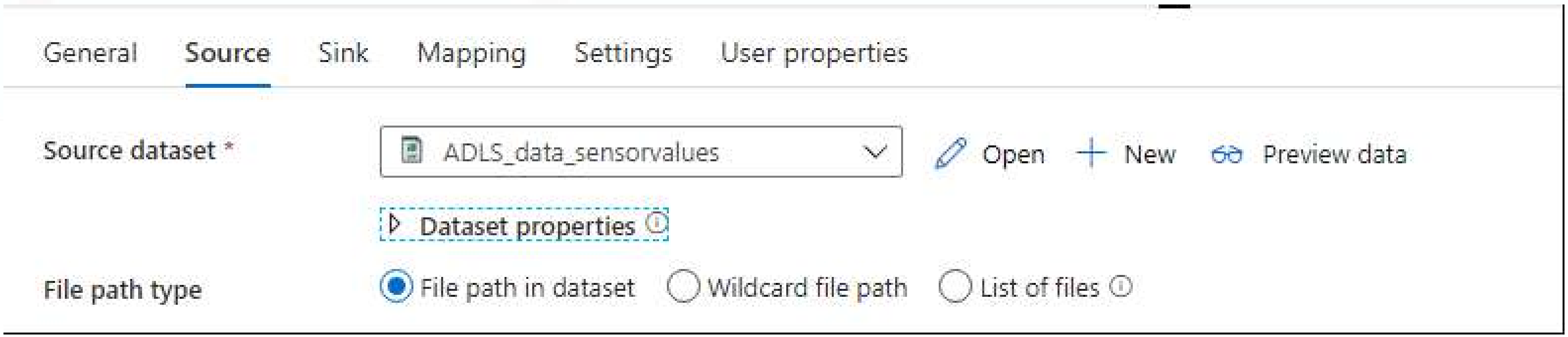


Name

SensorValues_M_01.csv

SensorVal

- Synapse Pipeline



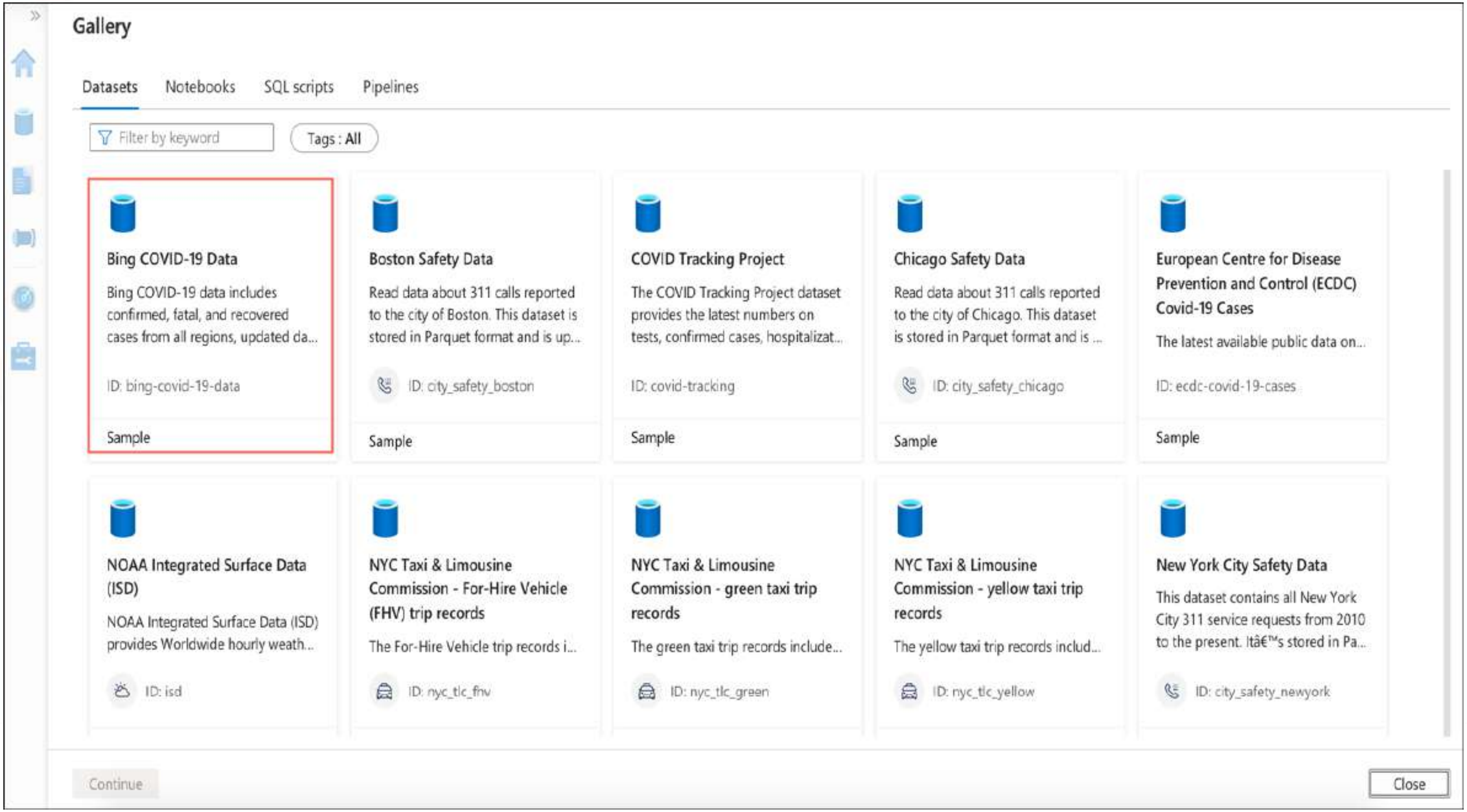
Source dataset * ADLS_data_sensorvalues

File path type ● File path in dataset ○ Wildcard file path ○ List of files

參考資料：Azure Synapse Analytics end-to-end: From your database and sensors to a Power BI dashboard

Browsersync: connected

Knowledge Center Gallery



Bing COVID-19 Data
Bing COVID-19 data includes confirmed, fatal, and recovered cases from all regions, updated da...
ID: bing-covid-19-data
Sample

Boston Safety Data
Read data about 311 calls reported to the city of Boston. This dataset is stored in Parquet format and is up...
ID: city_safety_boston
Sample

COVID Tracking Project
The COVID Tracking Project dataset provides the latest numbers on tests, confirmed cases, hospitalizat...
ID: covid-tracking
Sample

Chicago Safety Data
Read data about 311 calls reported to the city of Chicago. This dataset is stored in Parquet format and is ...
ID: city_safety_chicago
Sample

European Centre for Disease Prevention and Control (ECDC) Covid-19 Cases
The latest available public data on...
ID: ecdc-covid-19-cases
Sample

NOAA Integrated Surface Data (ISD)
NOAA Integrated Surface Data (ISD) provides Worldwide hourly weath...
ID: isd

NYC Taxi & Limousine Commission - For-Hire Vehicle (FHV) trip records
The For-Hire Vehicle trip records i...
ID: nyc_tlc_fhv

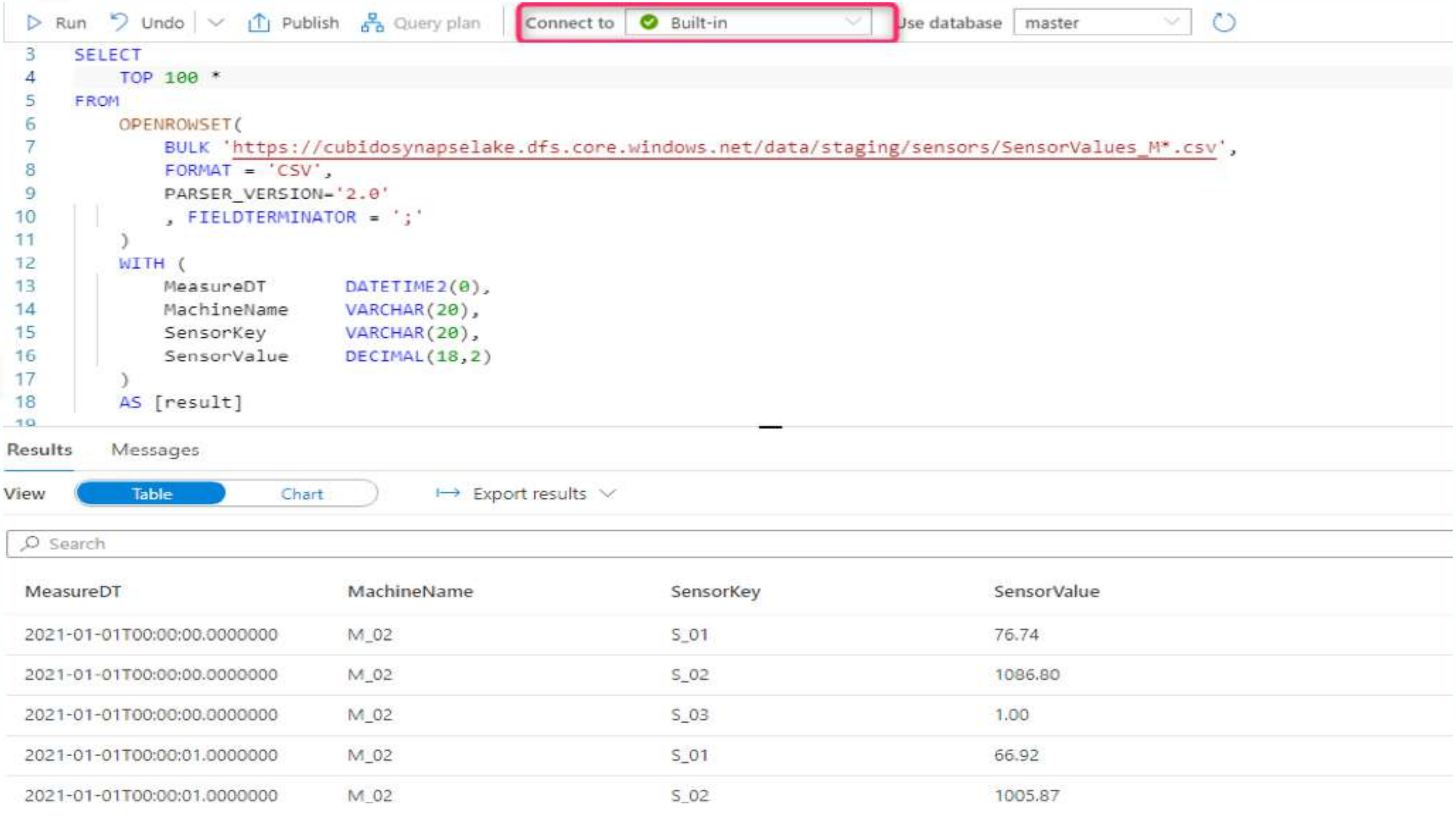
NYC Taxi & Limousine Commission - green taxi trip records
The green taxi trip records include...
ID: nyc_tlc_green

NYC Taxi & Limousine Commission - yellow taxi trip records
The yellow taxi trip records includ...
ID: nyc_tlc_yellow

New York City Safety Data
This dataset contains all New York City 311 service requests from 2010 to the present. Itâ€™s stored in Pa...
ID: city_safety_newyork

Browsersync: connected

Query Data



```
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://cubidosynapselake.dfs.core.windows.net/data/staging/sensors/SensorValues_M*.csv',
        FORMAT = 'CSV',
        PARSER_VERSION='2.0'
        , FIELDTERMINATOR = ';'
    )
    WITH (
        MeasureDT       DATETIME2(0),
        MachineName     VARCHAR(20),
        SensorKey       VARCHAR(20),
        SensorValue     DECIMAL(18,2)
    )
    AS [result]
```

MeasureDT	MachineName	SensorKey	SensorValue
2021-01-01T00:00:00.0000000	M_02	S_01	76.74
2021-01-01T00:00:00.0000000	M_02	S_02	1086.80
2021-01-01T00:00:00.0000000	M_02	S_03	1.00
2021-01-01T00:00:01.0000000	M_02	S_01	66.92
2021-01-01T00:00:01.0000000	M_02	S_02	1005.87

參考資料：Azure Synapse Analytics end-to-end: From your database and sensors to a Power BI dashboard

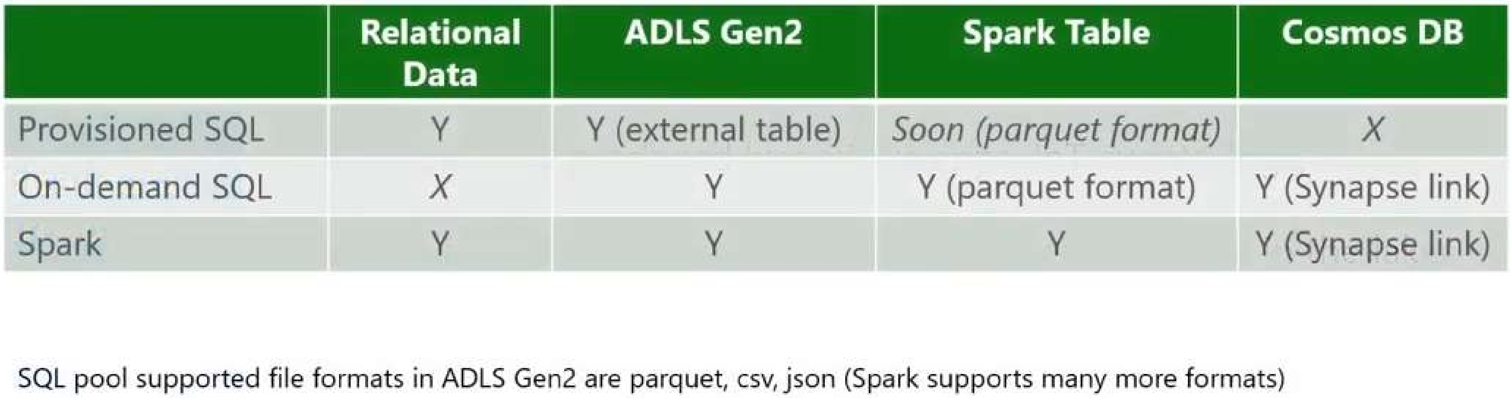
Different Options of Compute / Query

Azure Synapse Analytics 整合了：

- Data Lake
- Data Warehouse
- Big Data Analytics

Azure Synapse Analytics 提供三種 Query Service：

- Dedicated SQL Pool (原來的 SQL DW，底下的 Provisoned SQL)
- Serverless On-Demand SQL Pool (支援 Unstructured Data)
- Serverless Apache Spark Pool (Distributed In-Memory Computing)

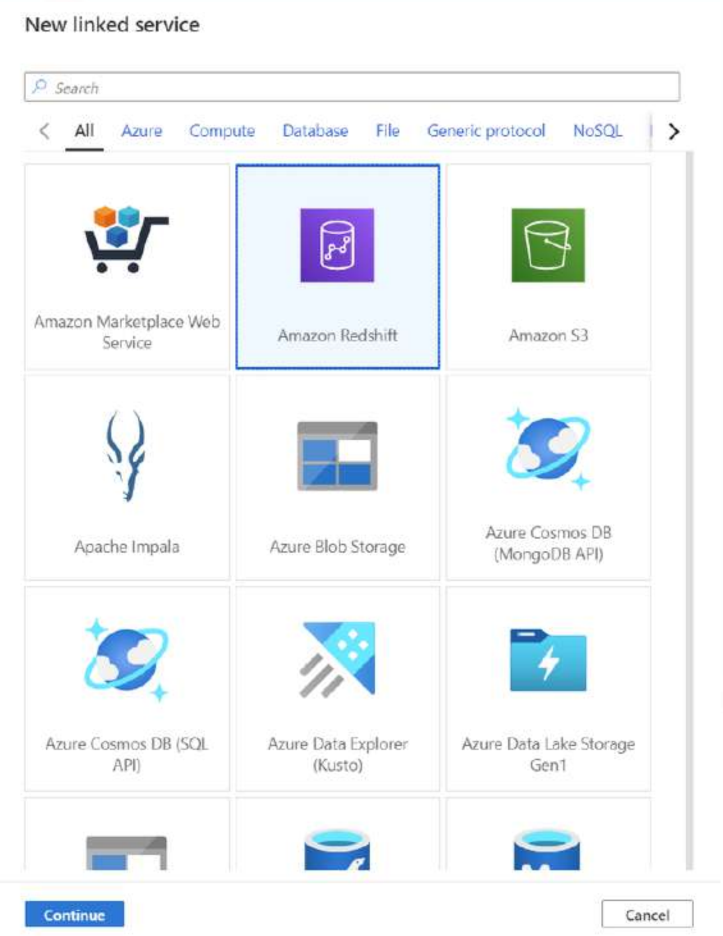


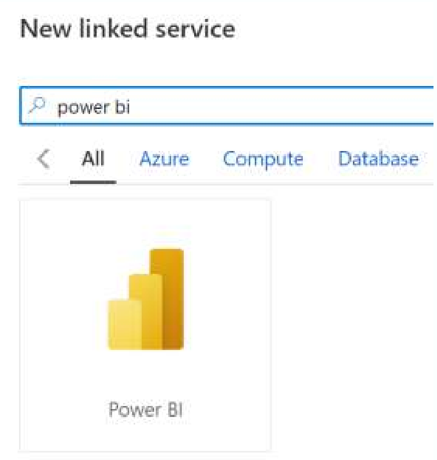
	Relational Data	ADLS Gen2	Spark Table	Cosmos DB
Provisioned SQL	Y	Y (external table)	*Soon (parquet format)*	*X*
On-demand SQL	*X*	Y	Y (parquet format)	Y (Synapse link)
Spark	Y	Y	Y	Y (Synapse link)

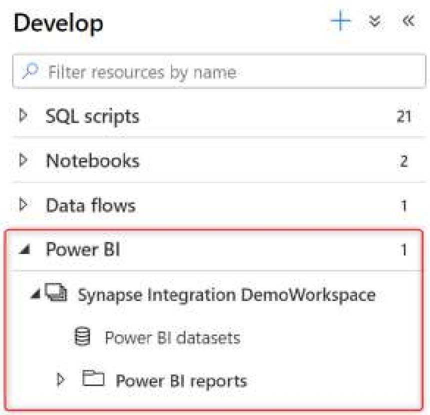
SQL pool supported file formats in ADLS Gen2 are parquet, csv, json (Spark supports many more formats)

參考資料：Azure Synapse Analytics For Beginner

Linked Service

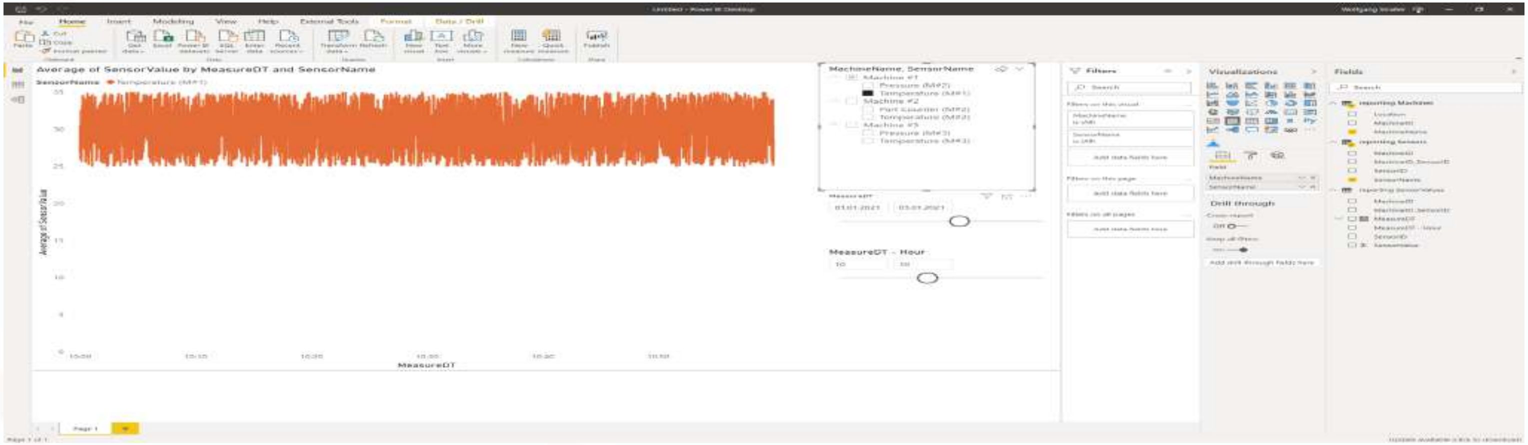


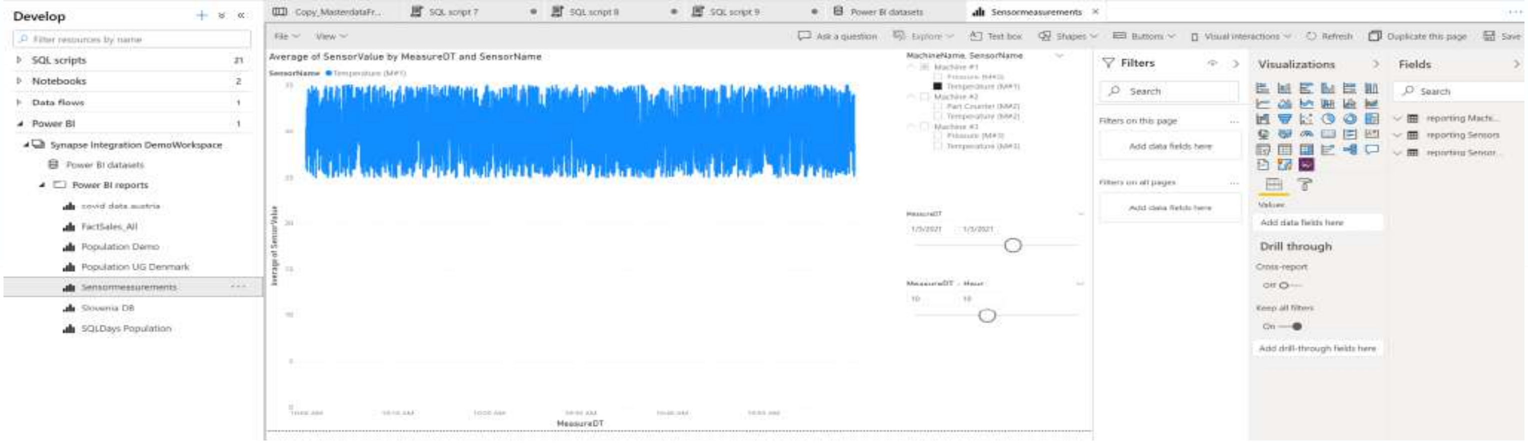




參考資料：Azure Synapse Analytics end-to-end: From your database and sensors to a Power BI dashboard

Analyze & Visualize Data



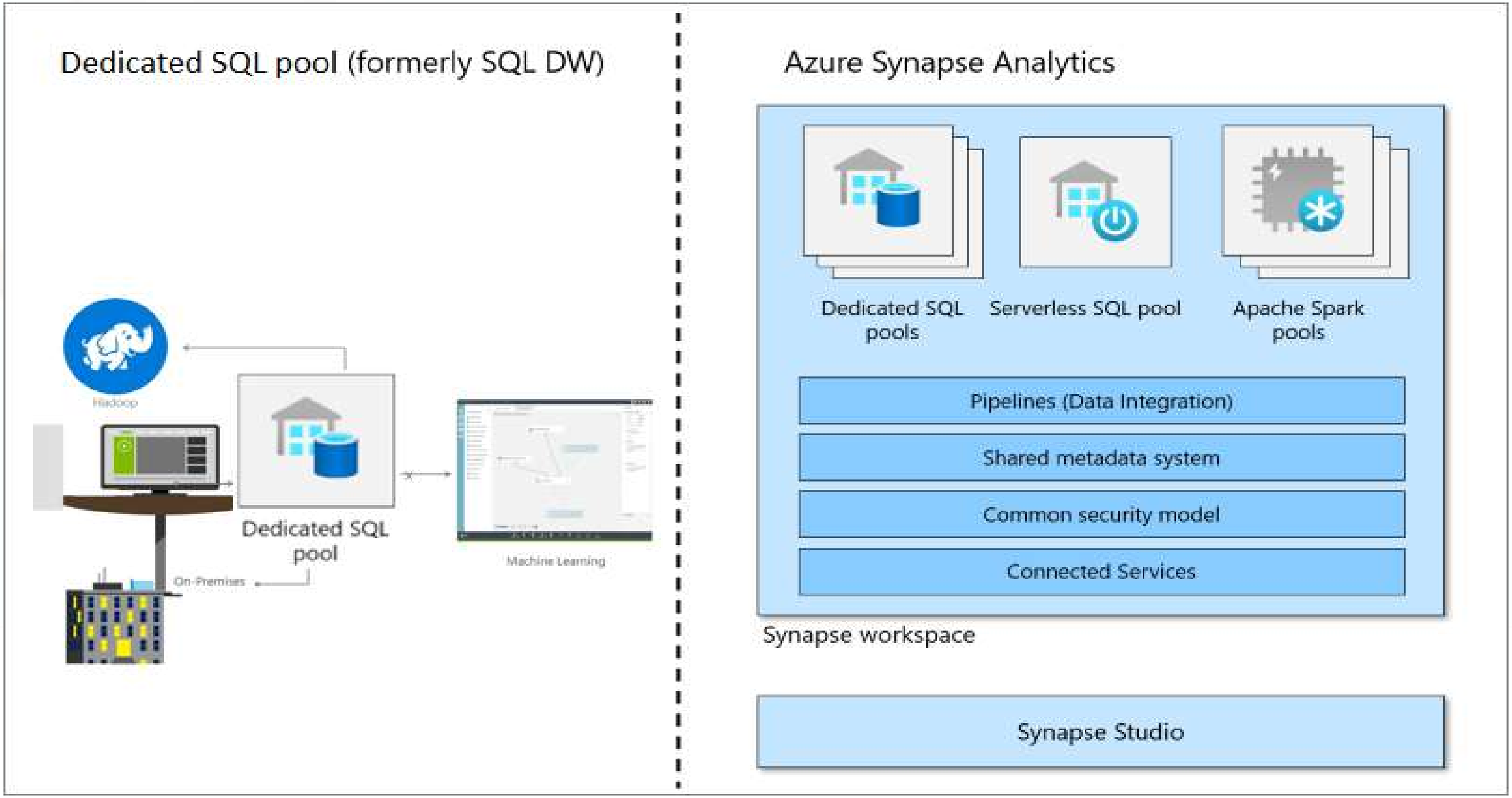


參考資料：Azure Synapse Analytics end-to-end: From your database and sensors to a Power BI dashboard

SQL DW / Dedicated SQL Pool

SQL DW vs. Azure Synapse Analytics

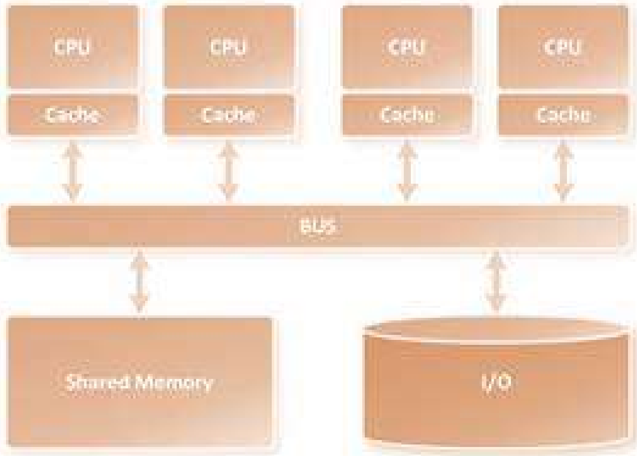
- 之前叫 SQL Data Warehouse
- 現在是 Azure Synapse Analytics 裡面的一個 Dedicated SQL Pool



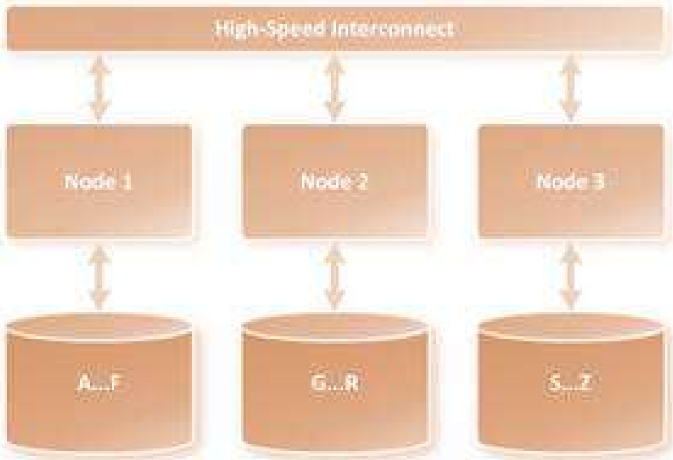
參考資料：
Dedicated SQL pool (formerly SQL DW) architecture in Azure Synapse Analytics
Azure SQL Data Warehouse

Shared-Memory vs. Shared-Nothing

- SQL DW 採用 Massively Parallel Processing 架構

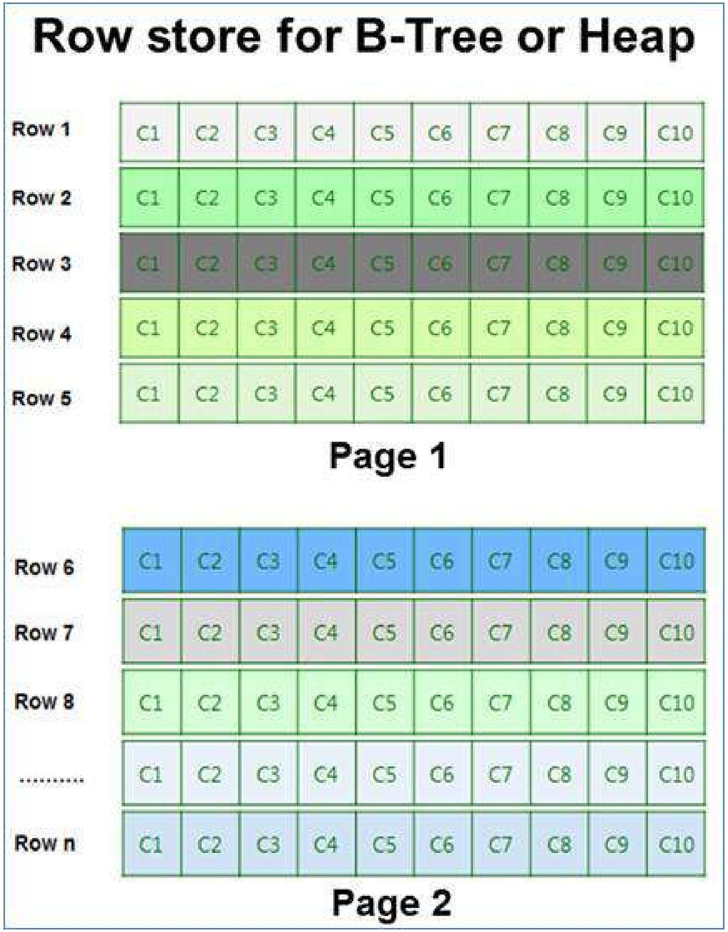


VS



參考資料：Getting Started with Azure SQL Data Warehouse - Part 1

Row Store vs. Column Store



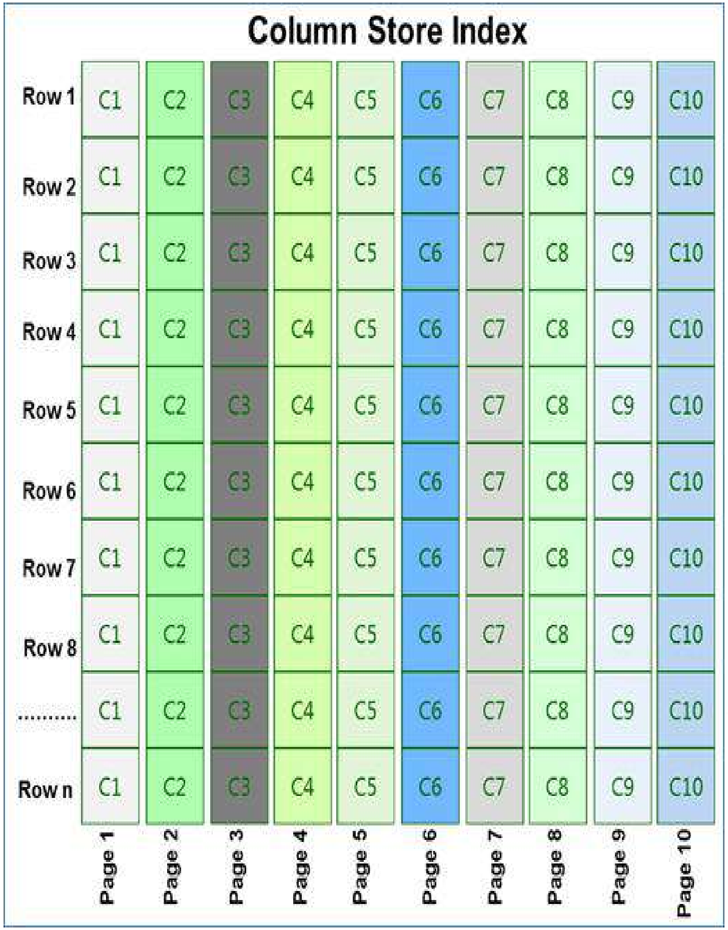
Row store for B-Tree or Heap

Row 1	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 2	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 3	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 4	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 5	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10

Page 1

Row 6	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 7	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 8	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
....	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row n	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10

Page 2



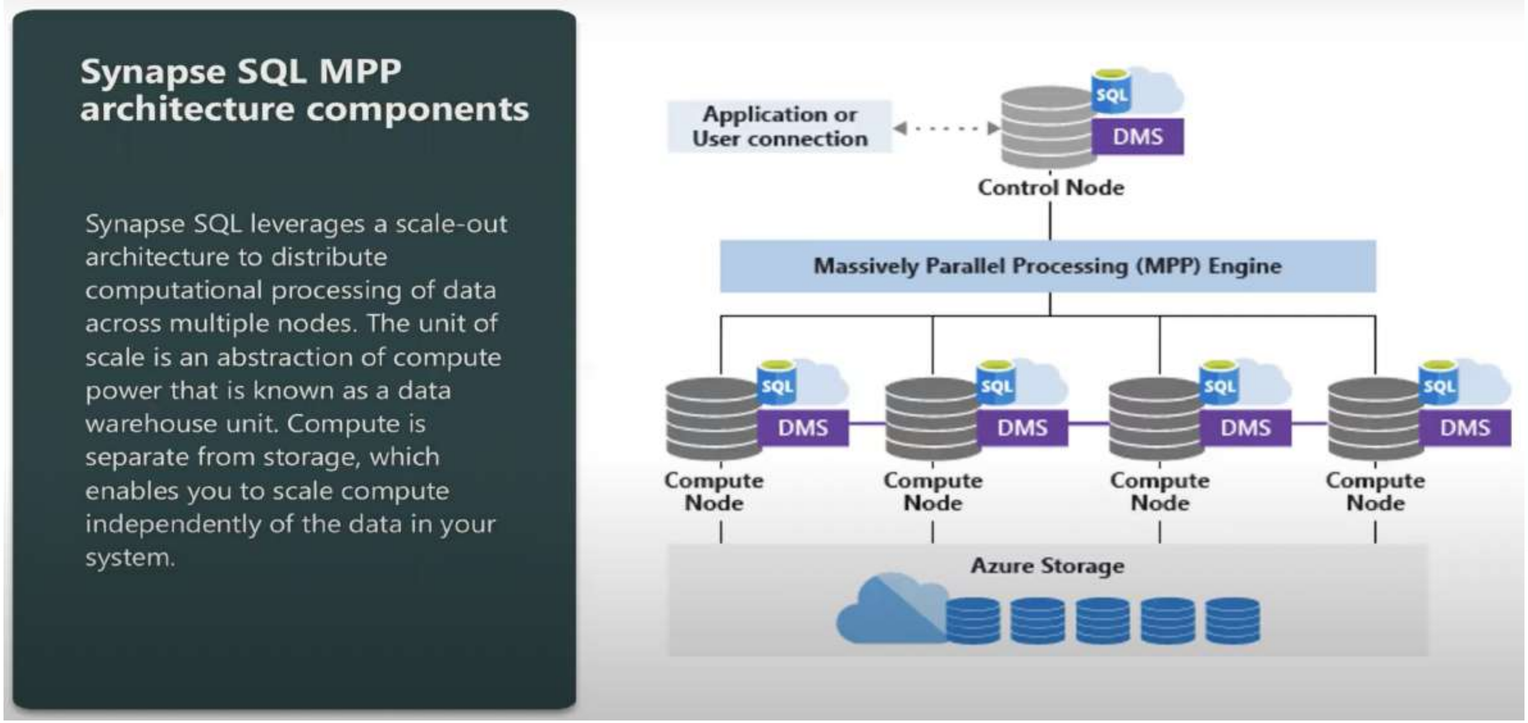
Column Store Index

Row 1	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 2	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 3	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 4	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 5	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 6	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 7	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row 8	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
....	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Row n	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	Page 1	Page 2	Page 3	Page 4	Page 5	Page 6	Page 7	Page 8	Page 9	Page 10

參考資料：Understanding New Column Store Index of SQL Server 2012

Separated Compute/Storage

- 由 1 個 Control Node (Master) + 1 或 多個 Compute Node (Slave) 組成
- Compute Node 數目由 Data Warehouse Unit (DWU) 決定
- 支援 PolyBase T-SQL Query Language
- 透過 Distributed Query Engine 進行 High-Performance Analytics



Synapse SQL MPP architecture components

Synapse SQL leverages a scale-out architecture to distribute computational processing of data across multiple nodes. The unit of scale is an abstraction of compute power that is known as a data warehouse unit. Compute is separate from storage, which enables you to scale compute independently of the data in your system.

參考資料：Dedicated SQL pool (formerly SQL DW) architecture in Azure Synapse Analytics

DWU vs. SLO vs. Performance Tier

Data Warehouse Unit (DWU)：

- 是 CPU、Memory、與 I/O 的組合
- Abstract Normalized Measure of Compute Resources and Performance
- 調整 Service Level Objective 就會改變 Data Warehouse Unit
- 改變 Data Warehouse Unit 不會影響 Storage Cost

Service Level Objective (SLO)：

- Scalability Setting of Cost and Performance Level
- Gen1 的 SLO 以 Data Warehouse Unit (DWU) 量測
- Gen2 的 SLO 以 Compute Data Warehouse Unit (cDWU) 量測

Performance Tier：

- 分爲 Gen1 (`DW100`-`DW6000`) 與 Gen2 (`DW100c`-`DW30000c`) 兩種
- Gen2 提供 Local Disk-Based Cache，所以 Performance 更好
- 之前使用 Gen1 部分原因是 Gen2 只支援比較高的 DWU (Gen2 現在支援 `DW100c`)
- SQL-DW DWU x 7.5 = SQL Server Required DTU
- SQL-DW cDWU x 9 = SQL Server Required DTU
- SQL Server DTU 上限預設 `54,000` (`DW6000c`)，更高就要開 Support Ticket
- SQL DB Contributor 與 SQL Server Contributor 這兩種 Role 才能調整

參考資料：
Capacity limits for dedicated SQL pool in Azure Synapse Analytics
Azure SQL Data Warehouse Gen2 now supports lower compute tiers

Service Level vs. Compute Nodes

Performance Level	Compute Nodes	Distributions / Compute Node	Memory / Data Warehouse (GB)
DW100c	1	60	60
DW200c	1	60	120
DW300c	1	60	180
DW400c	1	60	240
DW500c	1	60	300
DW1000c	2	30	600
DW1500c	3	20	900
DW2000c	4	15	1200
DW2500c	5	12	1500
DW3000c	6	10	1800
DW5000c	10	6	3000
DW6000c	12	5	3600
DW7500c	15	4	4500
DW10000c	20	3	6000
DW15000c	30	2	9000
DW30000c	60	1	18000

Browsersync: connected

Dedicated SQL Pool Pricing

Service Level	DWU	Pay As You Go (隨用隨付)	1 Year RC (節省約 37%)	3 Year RC (節省約 65%)
DW100c	100	NT$54.399/hour	NT$34.2709/hour	NT$19.0391/hour
DW200c	200	NT$108.797/hour	NT$68.5417/hour	NT$38.0781/hour
DW300c	300	NT$163.195/hour	NT$102.8125/hour	NT$57.1172/hour
DW400c	400	NT$217.593/hour	NT$137.0833/hour	NT$76.1562/hour
DW500c	500	NT$271.991/hour	NT$171.3541/hour	NT$95.1952/hour
DW1000c	1000	NT$543.982/hour	NT$342.7081/hour	NT$190.3904/hour
DW1500c	1500	NT$815.972/hour	NT$514.0621/hour	NT$285.5856/hour
DW2000c	2000	NT$1,087.963/hour	NT$685.4161/hour	NT$380.7808/hour
DW2500c	2500	NT$1,359.953/hour	NT$856.7702/hour	NT$475.9759/hour
DW3000c	3000	NT$1,631.944/hour	NT$1,028.1242/hour	NT$571.1711/hour
DW5000c	5000	NT$2,719.906/hour	NT$1,713.5403/hour	NT$951.9518/hour
DW6000c	6000	NT$3,263.887/hour	NT$2,056.2483/hour	NT$1,142.3422/hour
DW7500c	7500	NT$4,079.858/hour	NT$2,570.3104/hour	NT$1,427.9277/hour
DW10000c	10000	NT$5,439.811/hour	NT$3,427.0805/hour	NT$1,903.9036/hour
DW15000c	15000	NT$8,159.716/hour	NT$5,140.6207/hour	NT$2,855.8553/hour
DW30000c	30000	NT$16,319.431/hour	NT$10,281.2413/hour	NT$5,711.7106/hour

參考資料：Azure Synapse Analytics定價

Apache Spark Pool Pricing

以分計費：

Type	Price	Free Quantity
Memory Optimized	NT$4.646 / vCore-Hour	**2021-07-31 前** 120 Free vCore-Hours / Month

Data Exploration/Warehousing Pricing

Serverless / Dedicated 分別計費：

Type	Price	Free Quantity
Serverless	NT$202.866 / TB Data Processed	**2021-07-31 前** 10 TB of Free Queries / Month
Dedicated	與 Dedicated SQL Pool 相同計費	

參考資料：Azure Synapse Analytics定價

Browsersync: connected

SLO/DWU Verification (PowerShell)

在 PowerShell 或是 Azure Cloud Shell 手動安裝 `Az.Synapse` Module：

```
Install-Module -Name Az.Synapse

Untrusted repository
You are installing the modules from an untrusted repository... install the modules from 'PSGallery'?
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is "N"): A
```

建立後檢視：

```
PS /home/kc_su> Get-AzSynapseSqlPool -WorkspaceName synapseworkspacem00 -Name sqlpoolm00

ResourceGroupName        :
WorkspaceName            : synapseworkspacem00
SqlPoolName              : sqlpoolm00
Sku                      : DW100c
MaxSizeBytes             : 263882790666240
Collation                : SQL_Latin1_General_CP1_CI_AS
SourceDatabaseId         :
RecoverableDatabaseId    :
ProvisioningState        : Succeeded
Status                   : Online
RestorePointInTime       : 1/1/0001 12:00:00 AM
CreateMode               :
CreationDate             : 7/15/2021 3:37:08 PM
Tags                     : {}
TagsTable                :
Location                 : southeastasia
Id                       : /subscriptions/.../workspaces/synapseworkspacem00/sqlPools/sqlpoolm00
Type                     : Microsoft.Synapse/workspaces/sqlPools
```

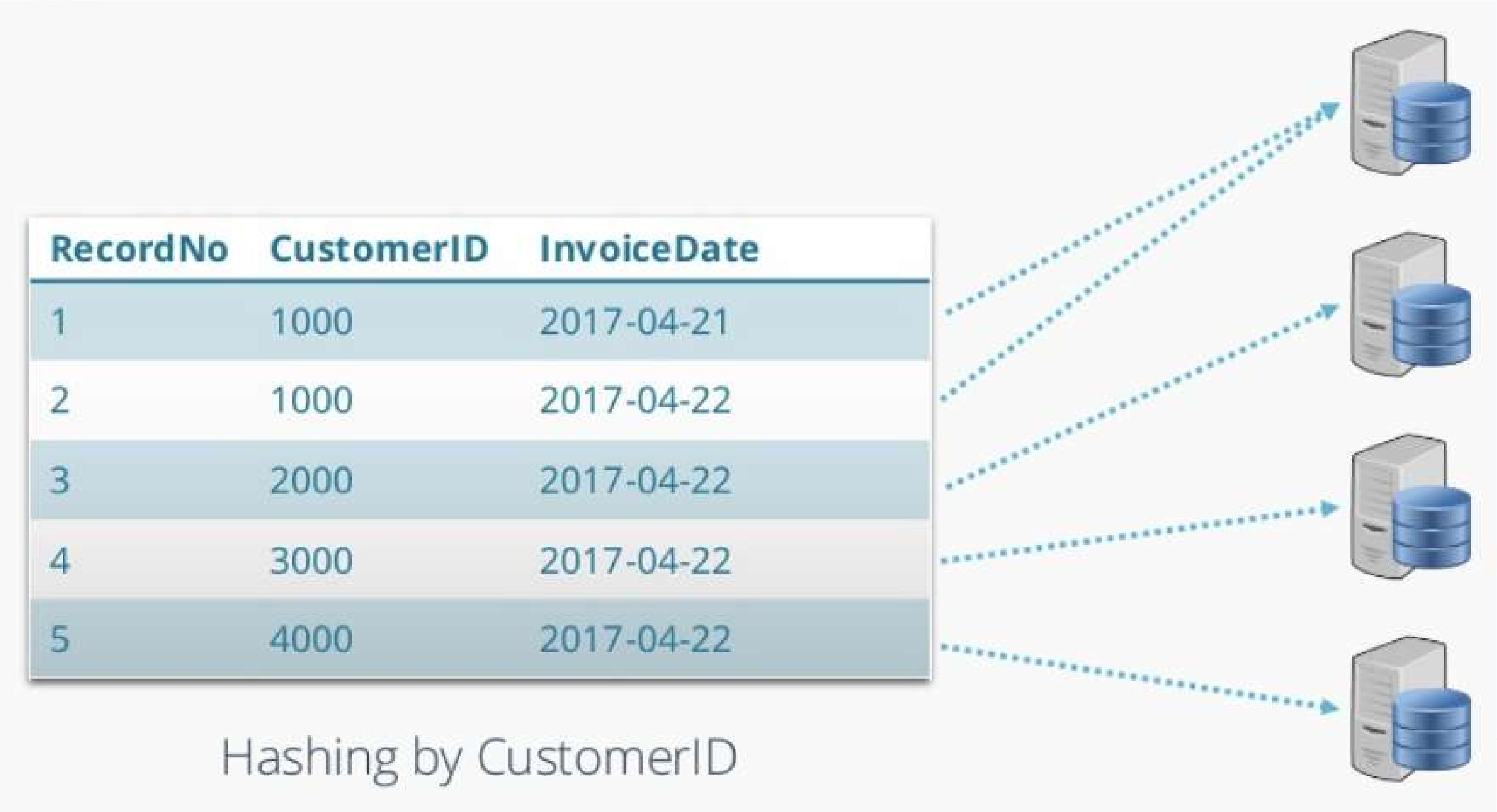
Data Distrubution (Sharding Pattern)

SQL-DW 載入資料的時候，最多會分成 60 個 Distribution：

- Hash
  - Distributed Using Hash Algorithm
  - Equal Values Hashed to the Same Distribution
  - Optimal for Joins and Aggregations on Large Fact Tables
- Round Robin
  - Distributed Evenly But Randomly
  - Not Requiring Knowledge about Data or Queries
  - Optimal for Large Tables without Good Hash Columns or Varied Queries
  - Fast Performance as a Staging Table for Loads
- Replicate
  - All Data Present on Every Node
  - Simplifying Query Plans and Reducing Data Movement
  - Best for Small Lookup Tables

參考資料：Design and Manage Azure SQL Data Warehouse

Hash

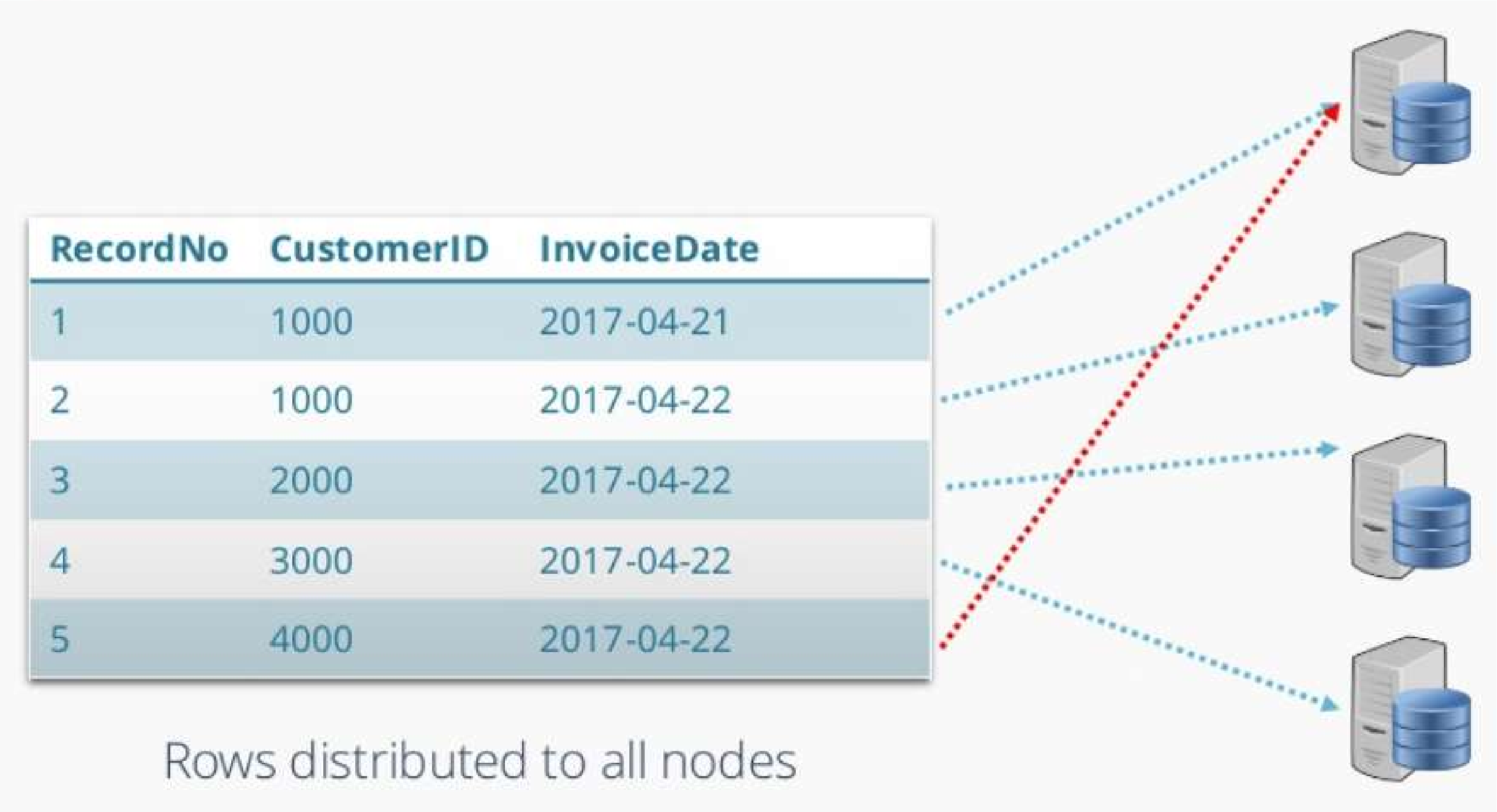


RecordNo	CustomerID	InvoiceDate
1	1000	2017-04-21
2	1000	2017-04-22
3	2000	2017-04-22
4	3000	2017-04-22
5	4000	2017-04-22

Hashing by CustomerID

参考資料：Azure SQL Data Warehouse

Round Robin



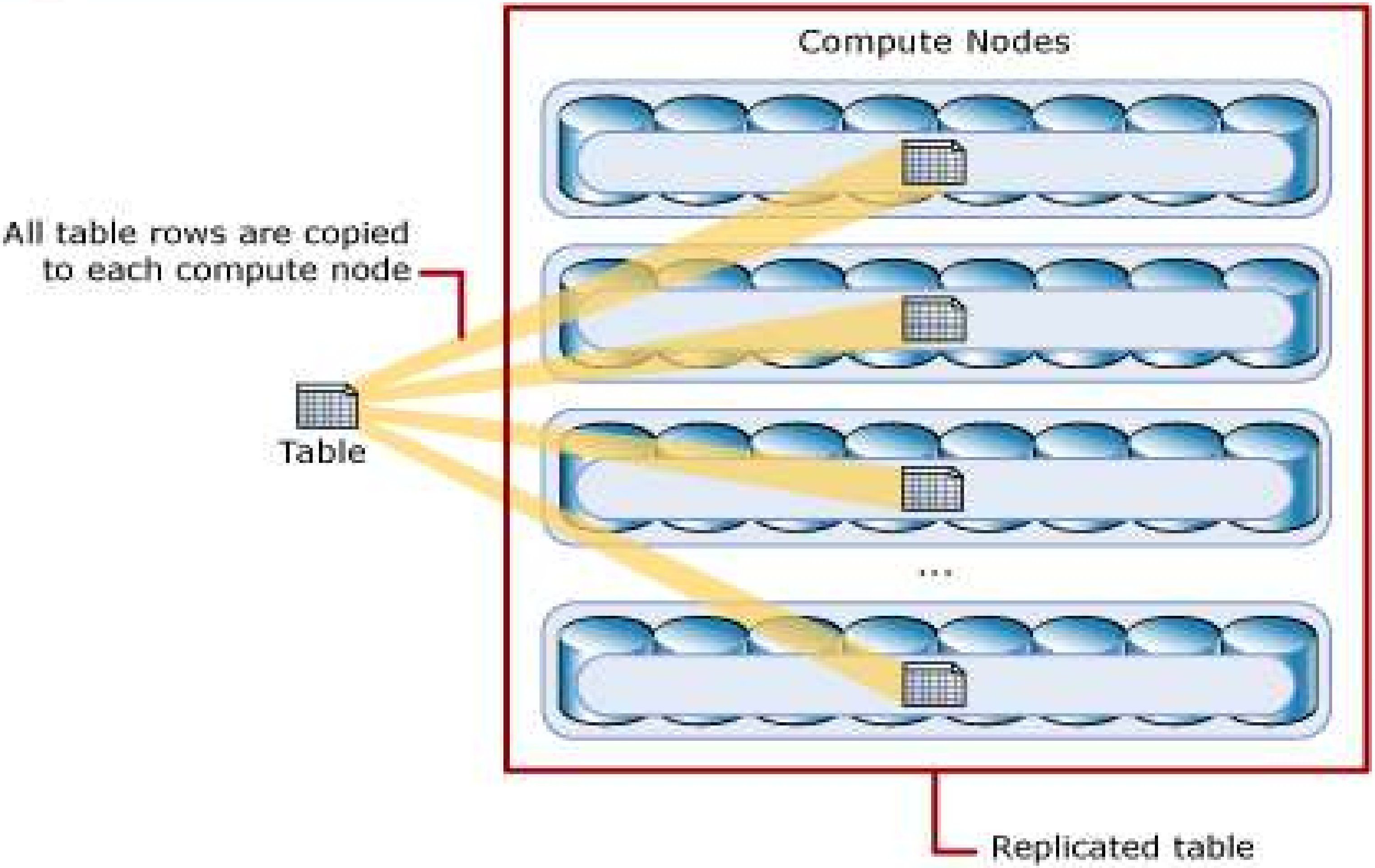
RecordNo	CustomerID	InvoiceDate
1	1000	2017-04-21
2	1000	2017-04-22
3	2000	2017-04-22
4	3000	2017-04-22
5	4000	2017-04-22

Rows distributed to all nodes

參考資料：Azure SQL Data Warehouse

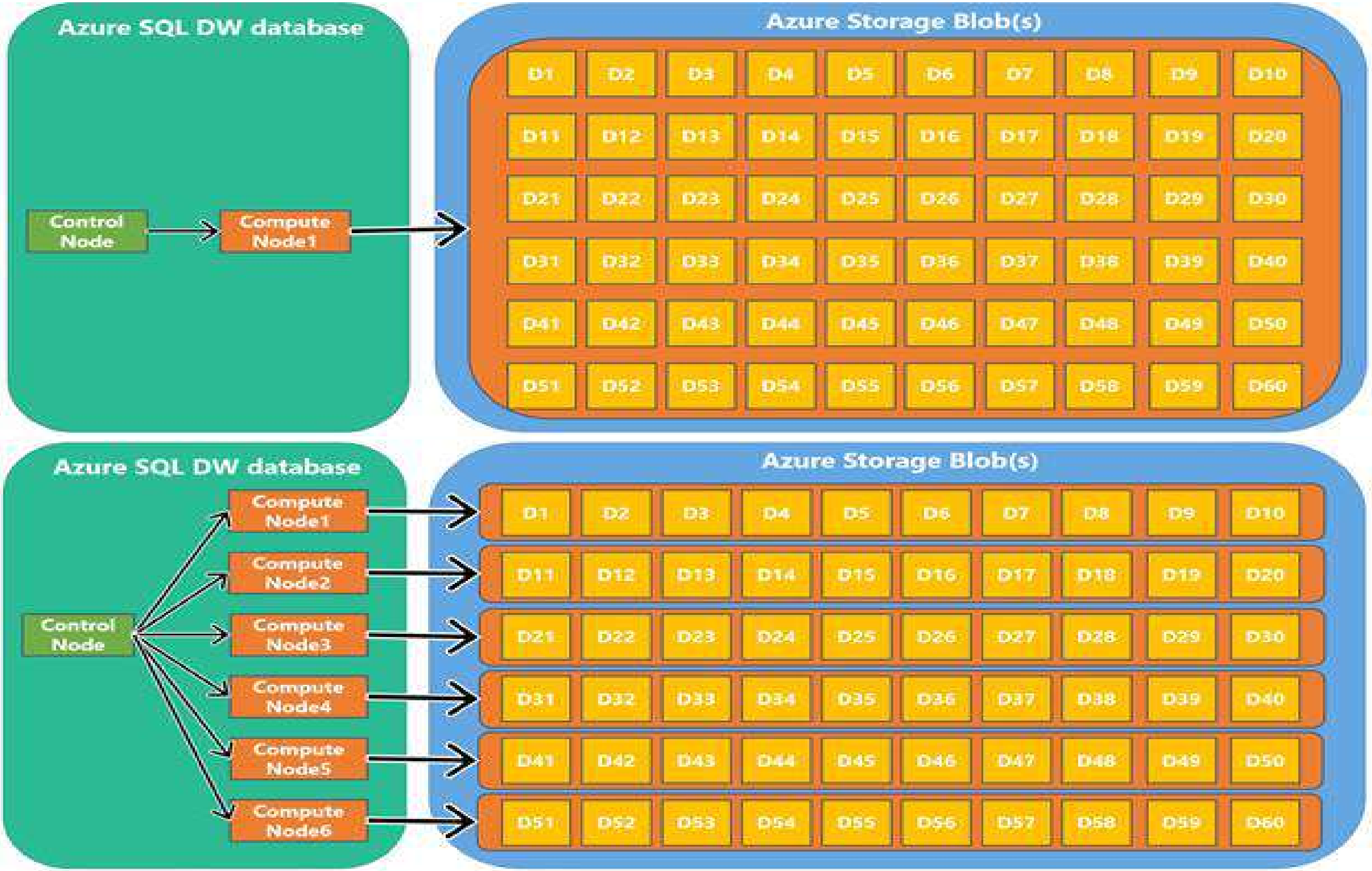
Browsersync: connected

Replicate



參考資料：Dedicated SQL pool (formerly SQL DW) architecture in Azure Synapse Analytics

Compute Node vs. Data Distribution



D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
D31	D32	D33	D34	D35	D36	D37	D38	D39	D40
D41	D42	D43	D44	D45	D46	D47	D48	D49	D50
D51	D52	D53	D54	D55	D56	D57	D58	D59	D60

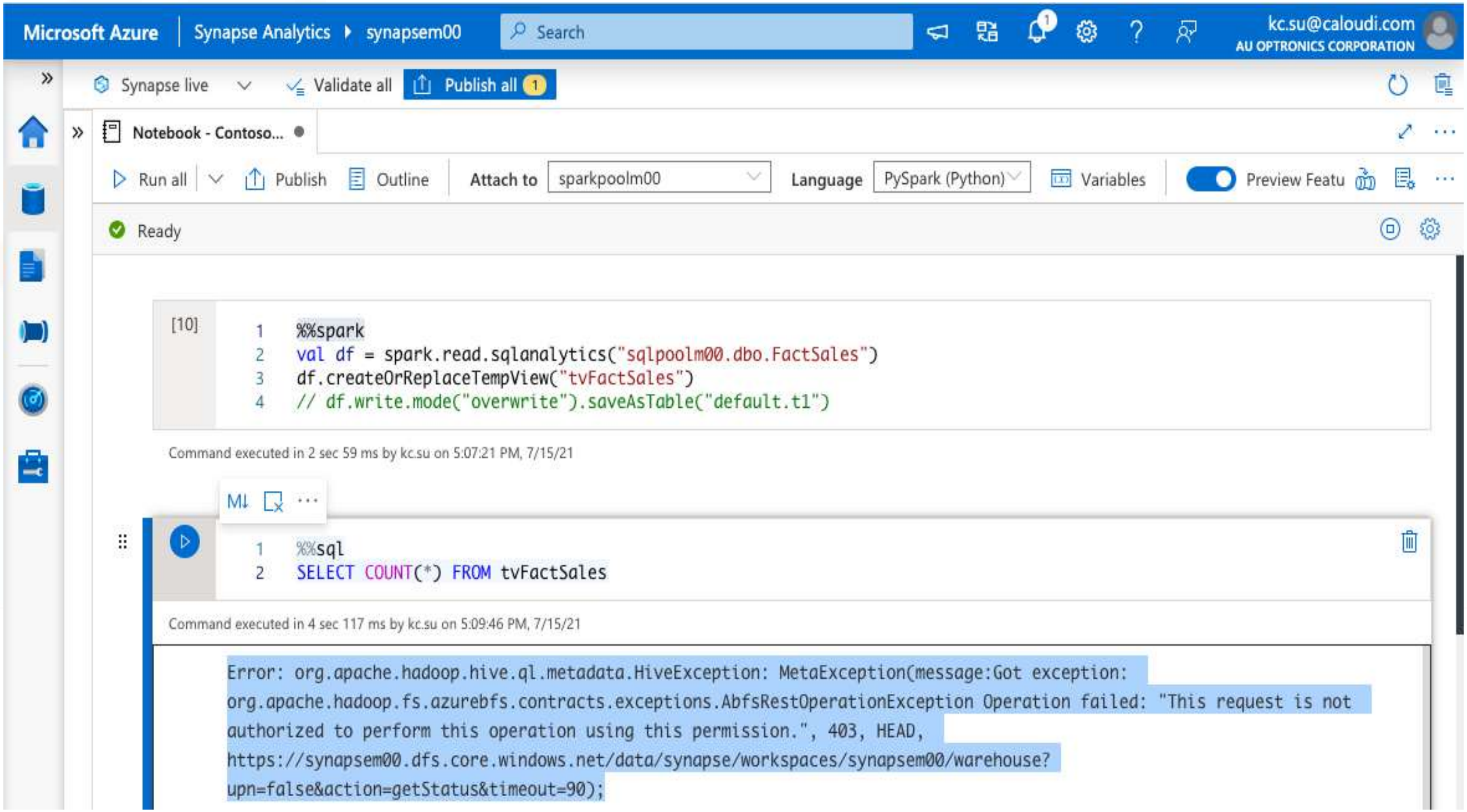
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
D31	D32	D33	D34	D35	D36	D37	D38	D39	D40
D41	D42	D43	D44	D45	D46	D47	D48	D49	D50
D51	D52	D53	D54	D55	D56	D57	D58	D59	D60

參考資料：Getting Started with Azure SQL Data Warehouse - Part 2

Implementation Issues

Permission Error

- 爲了存取 Azure Data Lake Storage Gen2 Storage Account，需要 Storage Blob Data Contributor 權限



```
%%spark
val df = spark.read.sqlanalytics("sqlpoolm00.dbo.FactSales")
df.createOrReplaceTempView("tvFactSales")
// df.write.mode("overwrite").saveAsTable("default.t1")
```

Command executed in 2 sec 59 ms by kc.su on 5:07:21 PM, 7/15/21

```
%%sql
SELECT COUNT(*) FROM tvFactSales
```

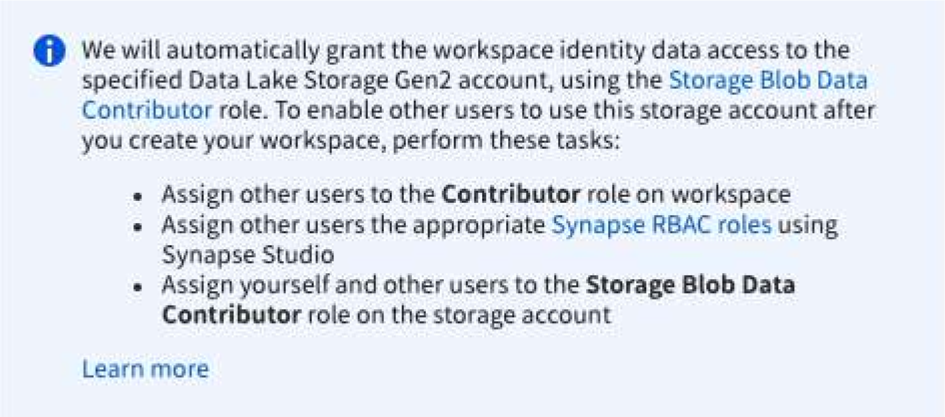
Command executed in 4 sec 117 ms by kc.su on 5:09:46 PM, 7/15/21

```
Error: org.apache.hadoop.hive.ql.metadata.HiveException: MetaException(message:Got exception:
org.apache.hadoop.fs.azurebfs.contracts.exceptions.AbfsRestOperationException Operation failed: "This request is not
authorized to perform this operation using this permission.", 403, HEAD,
https://synapsem00.dfs.core.windows.net/data/synapse/workspaces/synapsem00/warehouse?
upn=false&action=getStatus&timeout=90);
```

參考資料：Grant permissions to workspace managed identity

Problem and Answer

```
Error: org.apache.hadoop.hive.ql.metadata.HiveException: MetaException(
message:
  Got exception: org.apache.hadoop.fs.azurebfs.contracts.exceptions.AbfsRestOperationException
  Operation failed: "This request is not authorized to perform this operation using this permission.",
  403, HEAD,
  https://synapsem00.dfs.core.windows.net/data/synapse/workspaces/synapsem00/warehouse
      ?upn=false&action=getStatus&timeout=90);
```



We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the Storage Blob Data Contributor role. To enable other users to use this storage account after you create your workspace, perform these tasks:

- Assign other users to the **Contributor** role on workspace
- Assign other users the appropriate Synapse RBAC roles using Synapse Studio
- Assign yourself and other users to the **Storage Blob Data Contributor** role on the storage account

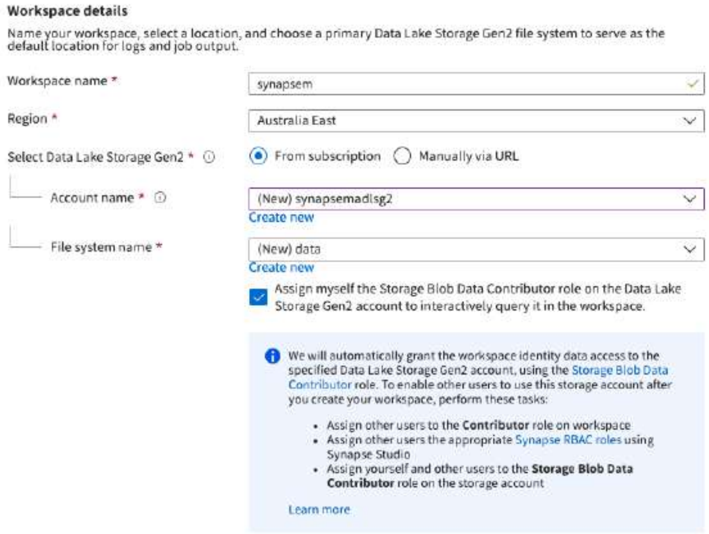
Learn more

Browsersync: connected

Subscription Owner vs. Non-Owner

底下這個選項只有在目前的 User 擁有 Subscription 的 Owner Role 時才會出現：

Assign myself the **Storage Blob Data Contributor** role on the Data Lake Storage Gen2 account to interactively query it in the 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 * synapsem

Region * Australia East

Select Data Lake Storage Gen2 * ● From subscription ○ Manually via URL

Account name * (New) synapsemadlsg2
Create new

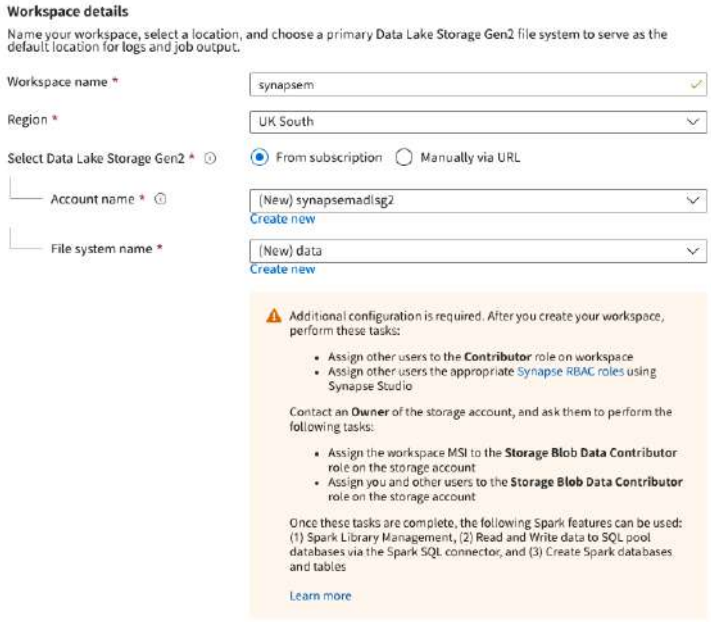
File system name * (New) data
Create new

☑ Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.

We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the Storage Blob Data Contributor role. To enable other users to use this storage account after you create your workspace, perform these tasks:

- Assign other users to the **Contributor** role on workspace
- Assign other users the appropriate Synapse RBAC roles using Synapse Studio
- Assign yourself and other users to the **Storage Blob Data Contributor** role on the storage account

Learn more



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 * synapsem

Region * UK South

Select Data Lake Storage Gen2 * ● From subscription ○ Manually via URL

Account name * (New) synapsemadlsg2
Create new

File system name * (New) data
Create new

Additional configuration is required. After you create your workspace, perform these tasks:

- Assign other users to the **Contributor** role on workspace
- Assign other users the appropriate Synapse RBAC roles using Synapse Studio

Contact an **Owner** of the storage account, and ask them to perform the following tasks:

- Assign the workspace MSI to the **Storage Blob Data Contributor** role on the storage account
- Assign you and other users to the **Storage Blob Data Contributor** role on the storage account

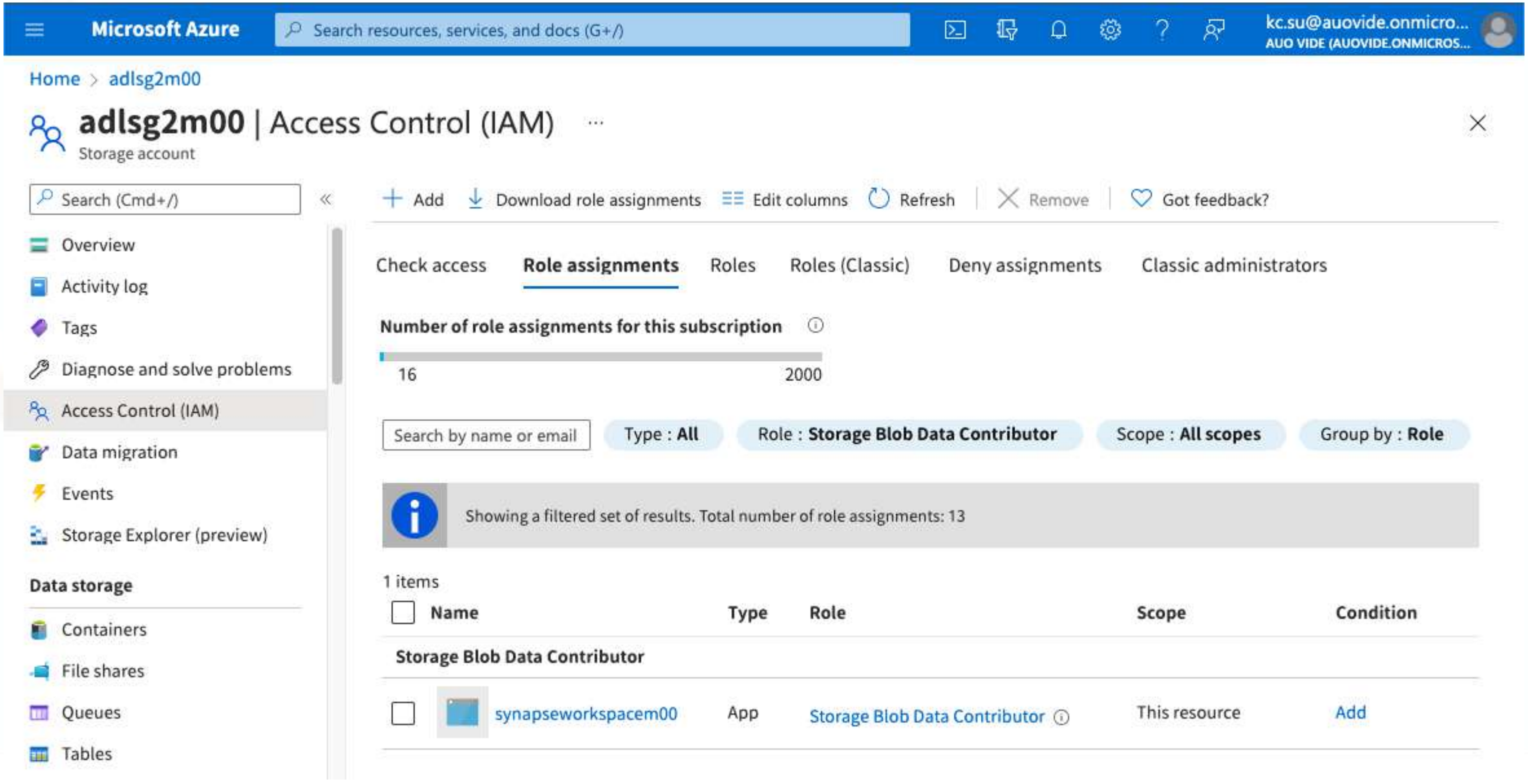
Once these tasks are complete, the following Spark features can be used: (1) Spark Library Management, (2) Read and Write data to SQL pool databases via the Spark SQL connector, and (3) Create Spark databases and tables

Learn more

參考資料：Quickstart: Create a Synapse workspace

Browsersync: connected

Storage Account -> Workspace



Microsoft Azure | Search resources, services, and docs (G+/) | kc.su@auovide.onmicro... AUO VIDE (AUOVIDE.ONMICROS...

Home > adlsg2m00

adlsg2m00 | Access Control (IAM) ...

Storage account

Search (Cmd+/) «

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage Explorer (preview)

Data storage

- Containers
- File shares
- Queues
- Tables

+ Add | Download role assignments | Edit columns | Refresh | Remove | Got feedback?

Check access | **Role assignments** | Roles | Roles (Classic) | Deny assignments | Classic administrators

Number of role assignments for this subscription

16 2000

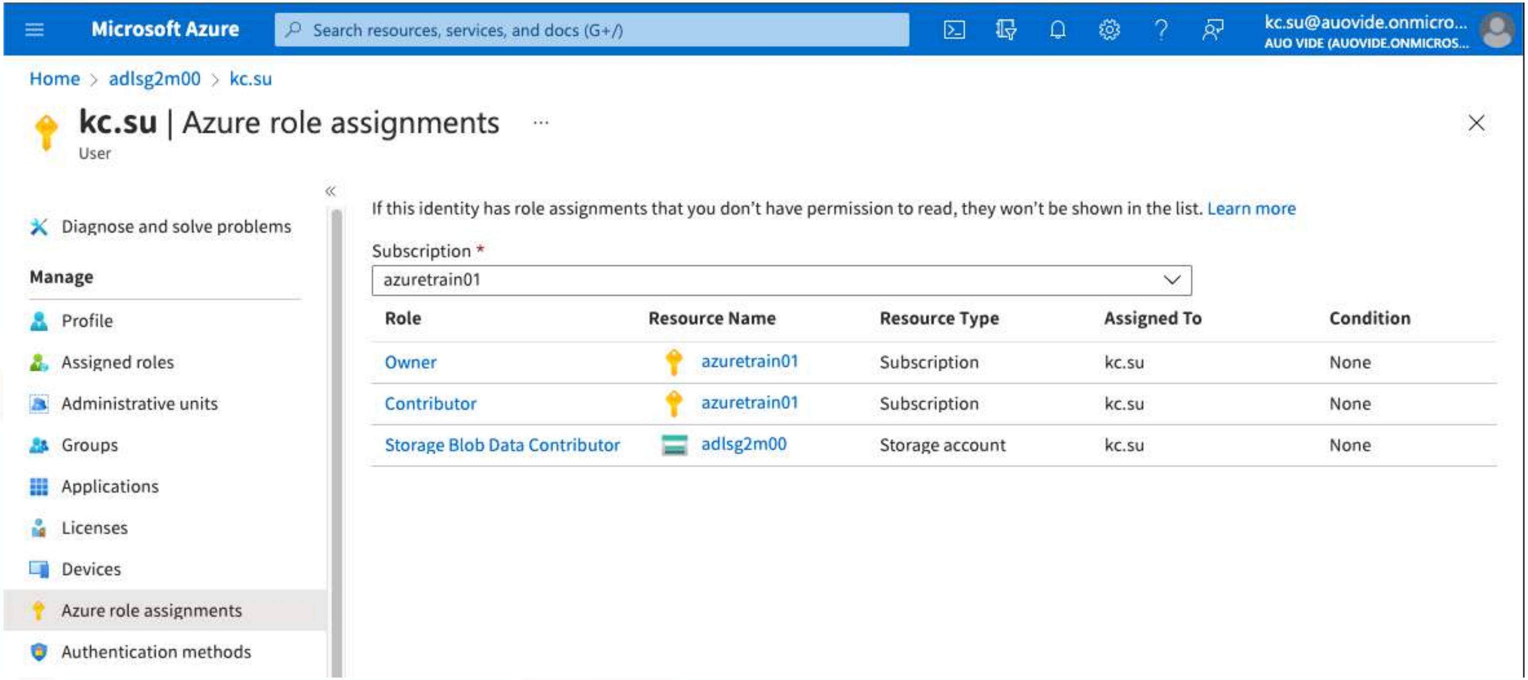
Search by name or email | Type : **All** | Role : **Storage Blob Data Contributor** | Scope : **All scopes** | Group by : **Role**

Showing a filtered set of results. Total number of role assignments: 13

1 items

☐ Name	Type	Role	Scope	Condition
Storage Blob Data Contributor				
☐ synapseworkspacem00	App	Storage Blob Data Contributor	This resource	Add

Storage Account -> User



If this identity has role assignments that you don't have permission to read, they won't be shown in the list. Learn more

Subscription *

azuretrain01

Role	Resource Name	Resource Type	Assigned To	Condition
Owner	azuretrain01	Subscription	kc.su	None
Contributor	azuretrain01	Subscription	kc.su	None
Storage Blob Data Contributor	adlsg2m00	Storage account	kc.su	None

Browsersync: connected

PolyBase OPENROWSET

- 可以存取 Azure Storage 裡面的 File Data
- 可以讀入 File Data，以 Relational Structure (A Set of Rows) 傳回

```sql
SELECT *
FROM OPENROWSET(
  BULK 'https://pandemicdatalake.blob.core.windows.net/.../covid-19/ecdc_cases/latest/ecdc_cases.csv',
  FORMAT = 'CSV',
  PARSER_VERSION = '2.0',
  HEADER_ROW = TRUE
) as [r]
```

```sql
SELECT TOP 1 *
FROM OPENROWSET(
  BULK 'https://azureopendatastorage.blob.core.windows.net/.../us_population/year=20*/*.parquet',
  FORMAT='PARQUET'
) WITH (
  [stateName] VARCHAR (50),
  [population] BIGINT
) AS [r]
```

```sql
SELECT TOP 1 *
FROM OPENROWSET(
  BULK 'https://azureopendatastorage.blob.core.windows.net/.../us_population/year=20*/*.parquet',
  FORMAT = 'DELTA'
) AS [r]
```

參考資料：How to use OPENROWSET using serverless SQL pool in Azure Synapse Analytics

Browsersync: connected

Serverless SQL Pool vs. Collation

- Serverless SQL Pool 可以將 UTF-8 Data 當成 `VARCHAR` Field 讀入，但是要注意 Collation 的影響
- 如果 `VARCHAR` Field 沒有指定 Collation，就會將 UTF-8 Character 強迫轉換爲 Plain `CARCHAR` Character，可能會造成轉換錯誤
- 這類錯誤會發生在 `OPENROWSET` 沒有 `WITH` 子句，或是 `OPENROWSET` 與 External Table 傳回 `VARCHAR` Field 時沒有指定 Collation
- 這個議題對 `NVARCHAR` Field 不適用，因爲 `NVARCHAR` Field 的轉換跟 Collation 無關，但是轉換成 `NVARCHAR` Field 會有一點 Performance 上的問題
- 改了 Database Colation 之後，既有的 External Table 必須重新建立才行

解決方式就是在 `CREATE DATABASE` 時指定 UTF-8 Collation：

```
CREATE DATABASE Database名稱 COLLATE Latin1_General_100_BIN2_UTF8
```

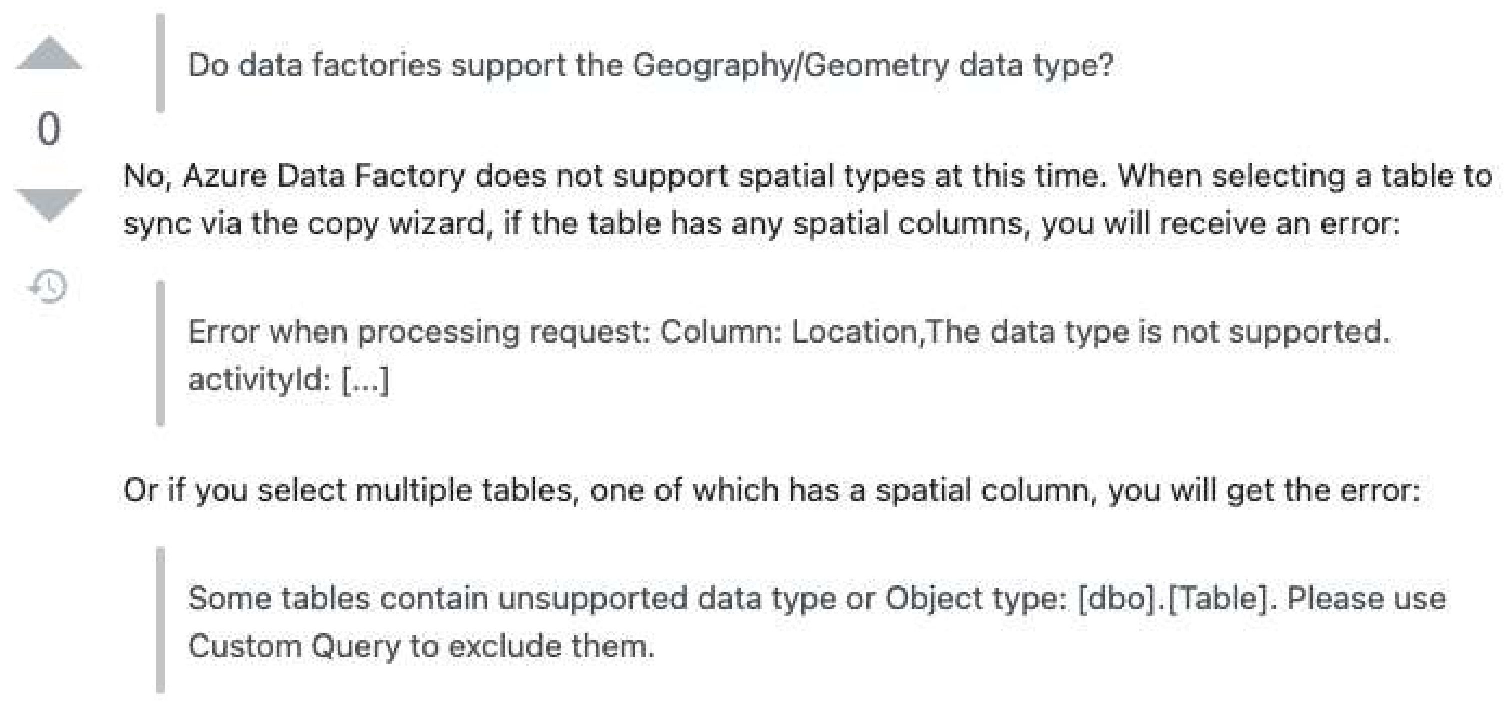
或是在 `OPENROWSET` 傳回的每個 `VARCHAR` Field 指定 UTF-8 Collation：

```
SELECT TOP 10 *
FROM OPENROWSET(
  BULK 'https://azureopendatastorage.blob.core.windows.net/.../us_population/year=20*/*.parquet',
  FORMAT='PARQUET'
) WITH (
  [stateName] VARCHAR (50) COLLATE Latin1_General_100_BIN2_UTF8 ,
  [population] BIGINT
) AS [r]
```

參考資料：
定序與 Unicode 支援
Always use UTF-8 collations to read UTF-8 text in serverless SQL pool

Azure Data Factory vs. Spatial Data Types

- Azure Data Factory 處理 SQL Server 的 Spatial Data Type 常常有問題
- Microsoft Contoso BI Demo Dataset 的 `DimStore` Table 的 `GeoLocation` 與 `Geometry` 兩個 Field 會匯入失敗
- 解決方式就是先刪掉這兩個 Field



> Do data factories support the Geography/Geometry data type?

No, Azure Data Factory does not support spatial types at this time. When selecting a table to sync via the copy wizard, if the table has any spatial columns, you will receive an error:

> Error when processing request: Column: Location,The data type is not supported. activityId: [...]

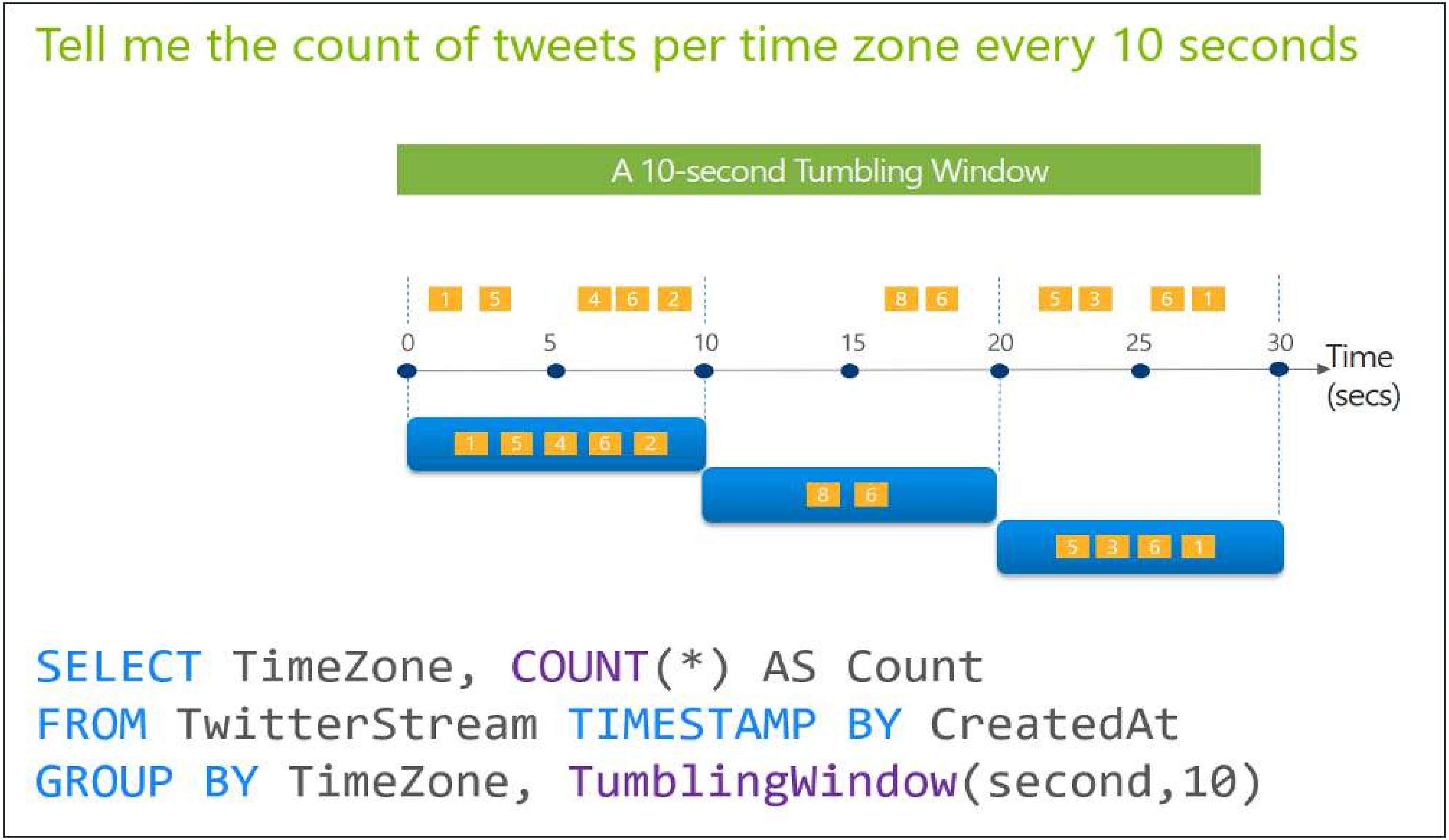
Or if you select multiple tables, one of which has a spatial column, you will get the error:

> Some tables contain unsupported data type or Object type: [dbo].[Table]. Please use Custom Query to exclude them.

參考資料：Azure Data Factory Geography or Geometry Data types

Tumbling Window

A series of fixed-sized, non-overlapping and contiguous time intervals.



Tell me the count of tweets per time zone every 10 seconds

```sql
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

參考資料：Tumbling Window (Azure Stream Analytics)

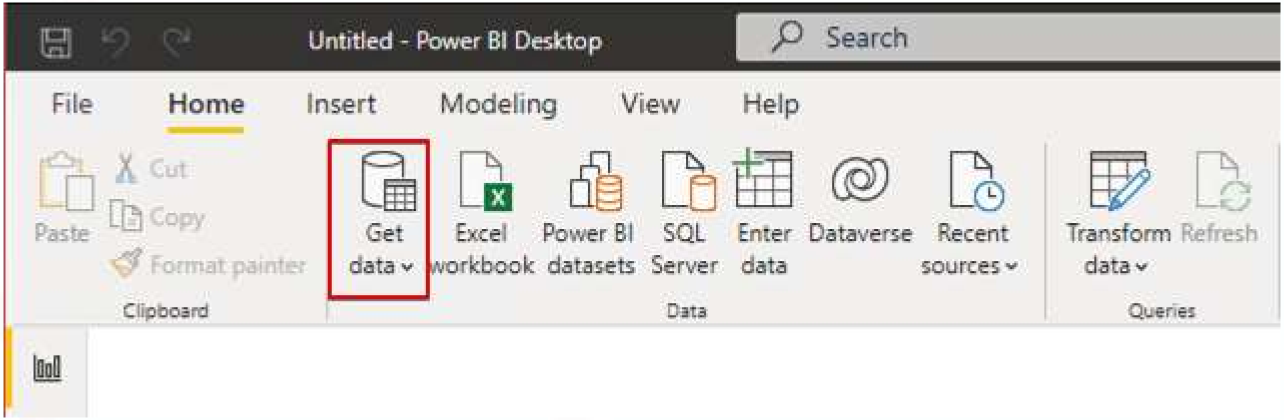
Power BI .pbids File

- `.pbids` 其實是 JSON 格式的檔案
- 所以也可以透過 Power BI 正常的 Get Data 程序連上 Data Source

```
{
  "version": "0.1",
  "connections": [
      {
          "details": {
              "protocol": "tds",
              "address": {
                  "server": "synapseworkspacem00.sql.azuresynapse.net",
                  "database": "sqlpoolm00"
              }
          },
          "storageMode": "DirectQuery"
      }
  ]
}
```

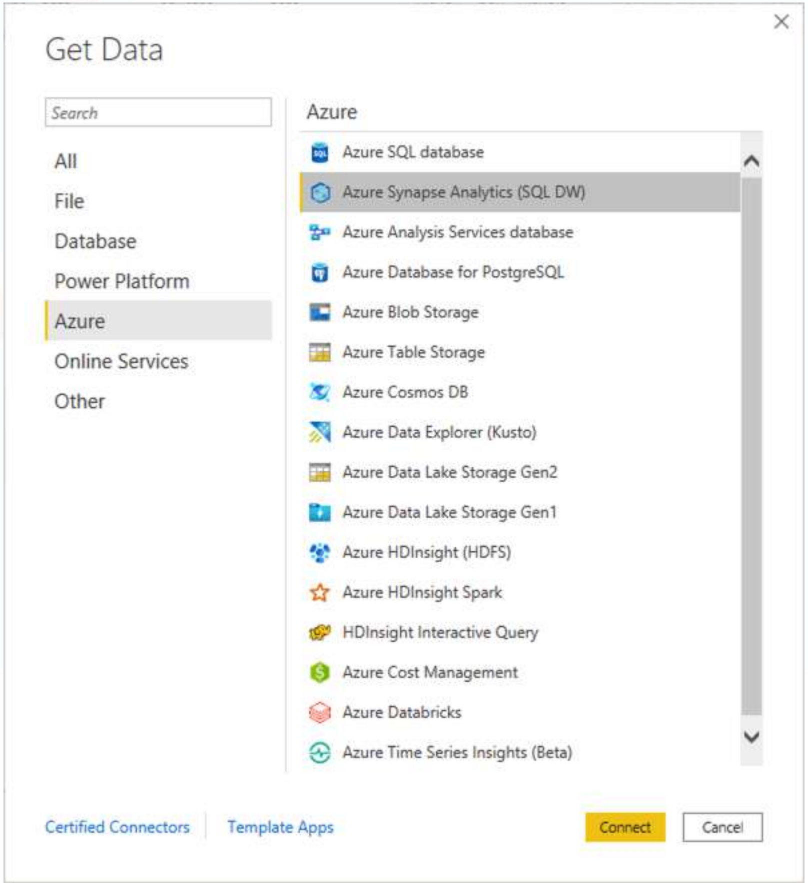
Power BI Get Data (01)

- 按下 Get data 圖示



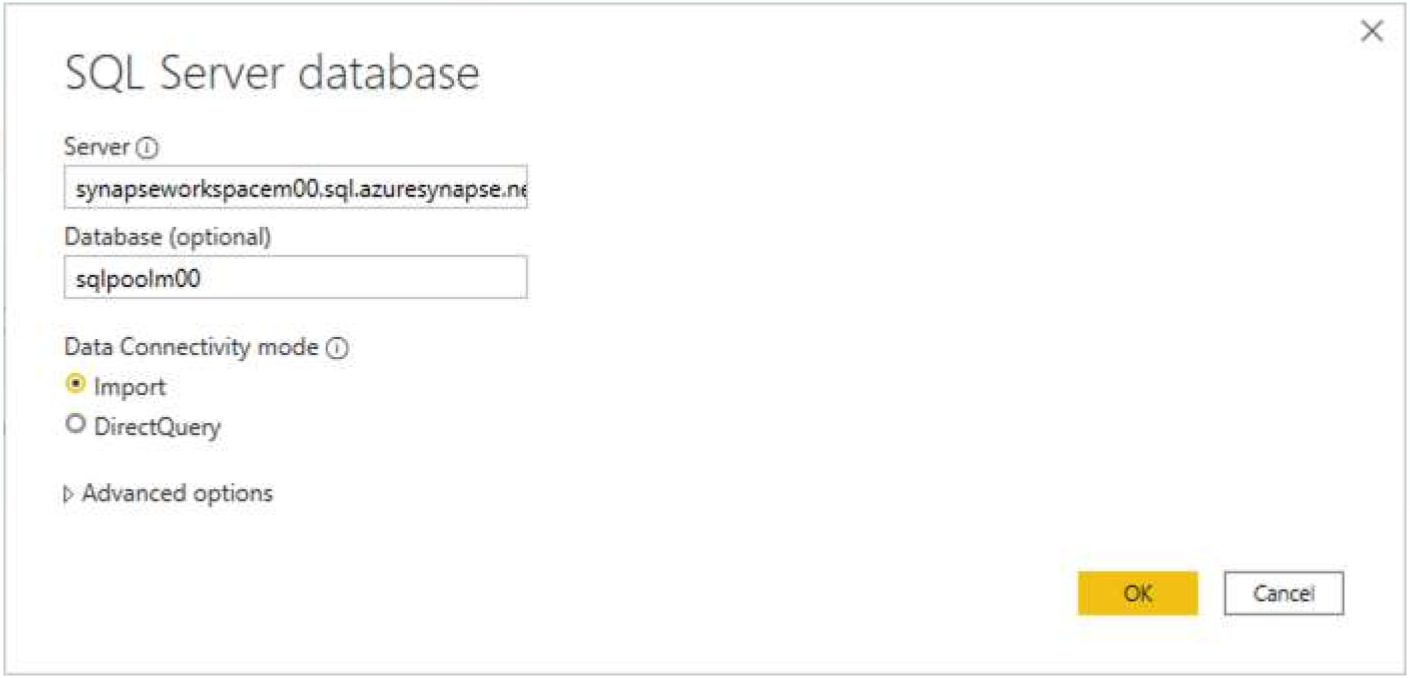
Power BI Get Data (02)

- 選取 Azure Synapse Analytics (SQL DW)，按下 Connect 按鈕



Power BI Get Data (03)

- Server 輸入 Dedicated SQL endpoint `synapseworkspace員工編號.sql.azuresynapse.net`
- Database 輸入 Dedicated SQL Pool `sqlpoolm00`
- Data Connectivity mode 選取 Import
- 然後按下 OK 按鈕



SQL Server database

Server ⓘ

synapseworkspacem00.sql.azuresynapse.ne

Database (optional)

sqlpoolm00

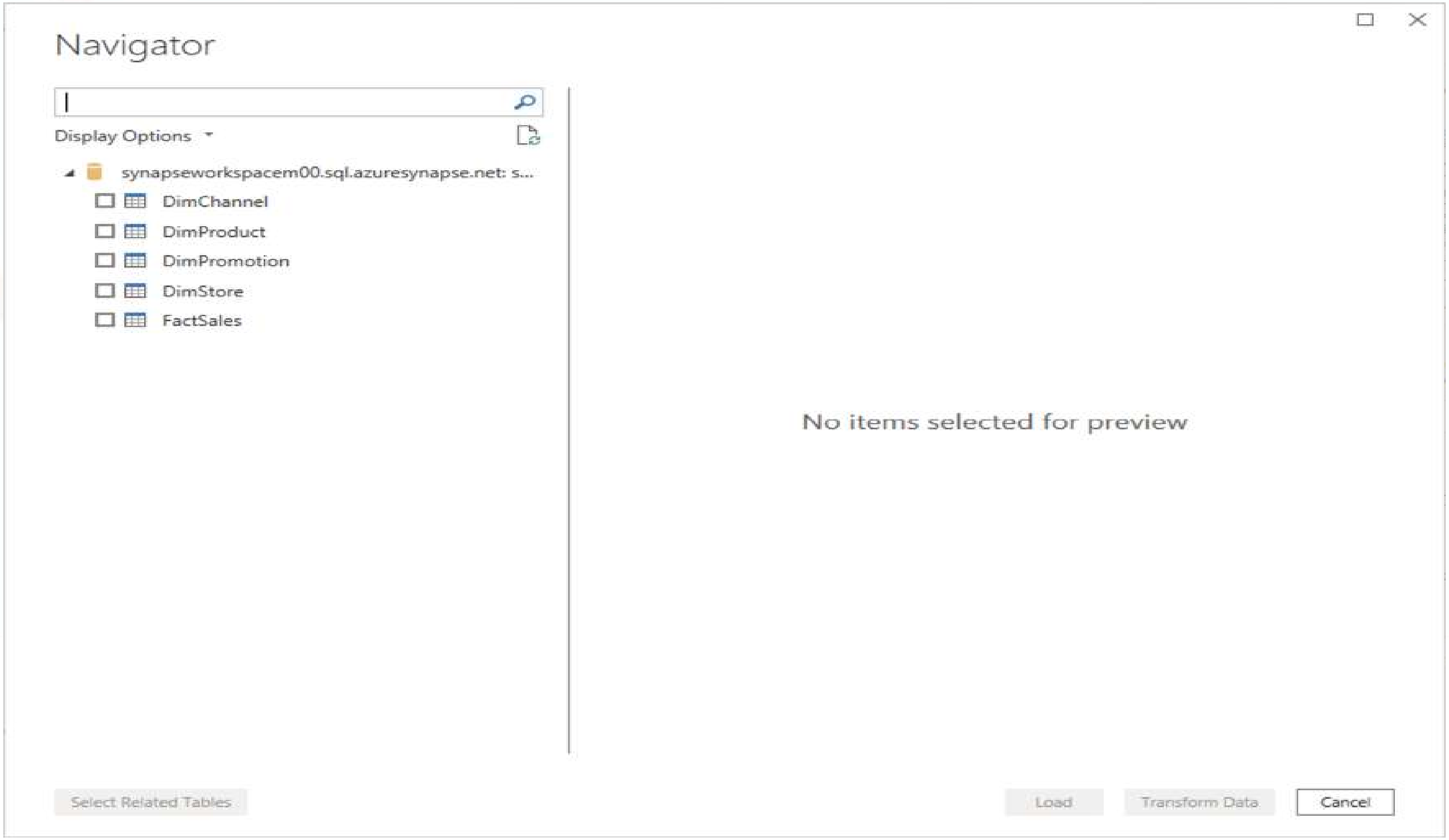
Data Connectivity mode ⓘ

● Import

○ DirectQuery

Power BI Get Data (04)

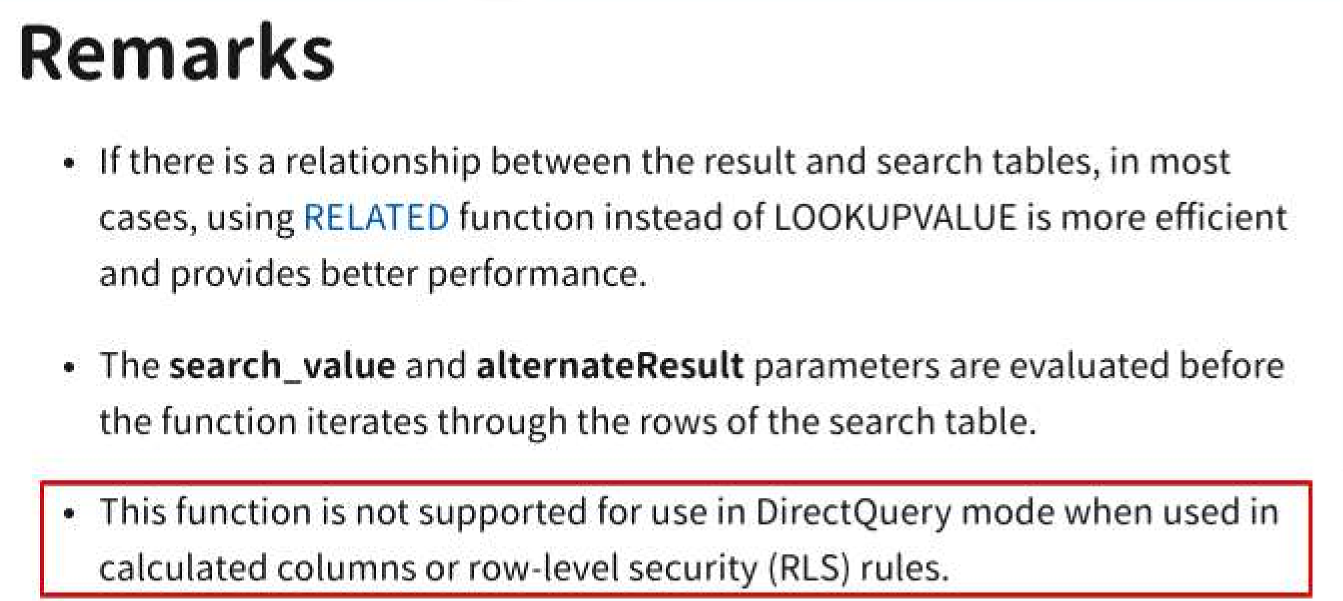
- 之後就跟按下 `.pbids` 檔案的畫面一樣



Power BI Import vs. DirectQuery

Import 與 DirectQuery：

- 所有 Data Source 都支援 Import，少部分 Data Source 才支援 DirectQuery
- Import 會真的把 Data 匯入，所以一開始 Load Data 比較慢
- DirectQuery 只是建立 Connection，所以一開始 Load Data 很快就結束
- 重點是：DirectQuery 不支援 `LOOKUPVALUE()` Function



Remarks

- If there is a relationship between the result and search tables, in most cases, using RELATED function instead of LOOKUPVALUE is more efficient and provides better performance.
- The **search_value** and **alternateResult** parameters are evaluated before the function iterates through the rows of the search table.
- This function is not supported for use in DirectQuery mode when used in calculated columns or row-level security (RLS) rules.

參考資料：
LOOKUPVALUE
Use DirectQuery in Power BI Desktop

References

Browsersync: connected

Reference Materials

- Azure Synapse Analytics
- What is dedicated SQL pool (formerly SQL DW) in Azure Synapse Analytics?
- Transact-SQL statements
- Az.Synapse
- Get Started with Azure Synapse Analytics
- Load data from Azure Data Lake Storage into dedicated SQL pools in Azure Synapse Analytics
- Load Contoso retail data into dedicated SQL pools in Azure Synapse Analytics