



# GPS Data/AI Strategy FY23

Delivered by CSA Team  
22/09/2022

 Franck Gaillard  
Cloud Solution Architect  
Data AI  
[frgail@microsoft.com](mailto:frgail@microsoft.com)

 Narjes Majdoub  
Cloud Solution Architect  
Data AI  
[nmajdoub@microsoft.com](mailto:nmajdoub@microsoft.com)

 Ali Bouhaddou  
Cloud Solution Architect  
Data Analytics  
[albouhad@microsoft.com](mailto:albouhad@microsoft.com)

 Frederic Gisbert  
Cloud Solution Architect  
Data Analytics  
[frgisber@microsoft.com](mailto:frgisber@microsoft.com)



# Azure Data & AI technical intensity plan

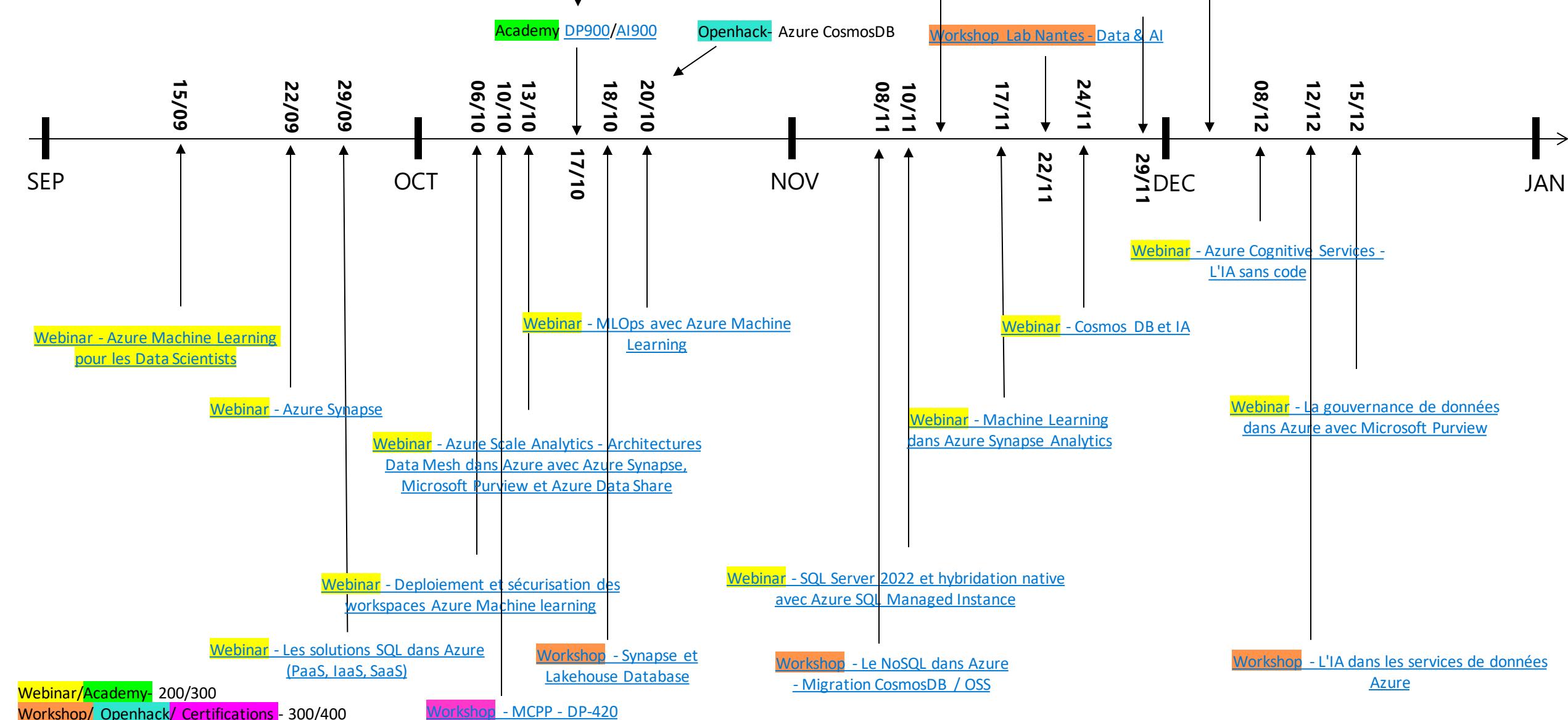
- From June 2022 to June 2023
- Focus on "Azure Data & AI" tech intensity
- Many content, from L100 Beginner to L400 Expert level
  - Academy L100
  - Webinar L200/L300
  - Workshop L300/L400
  - Certification kickstart L300/L400
  - Openhack / Microhack L400

Kickstart (17/10)

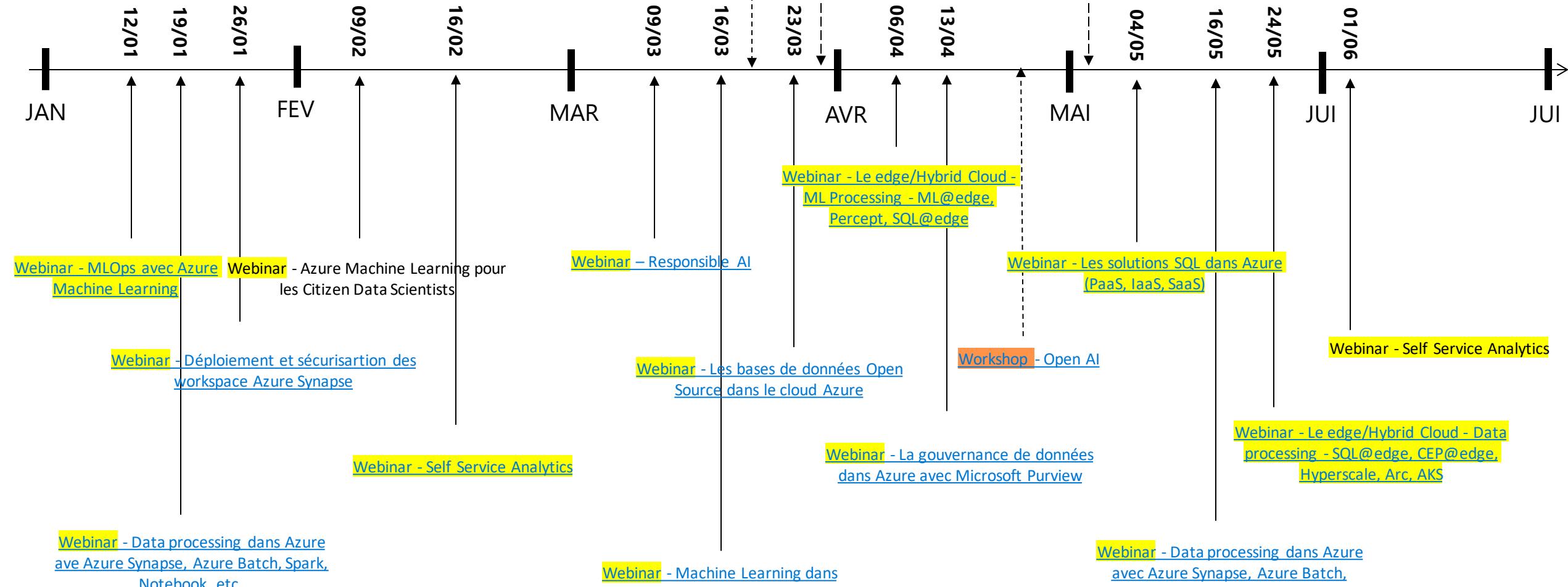
Dry Run (14/11)

Q&amp;A (05/12)

# Plan GPS Data/AI global FY23 (H1)



# Plan GPS Data/AI global FY23 (H2)



# Liste des évènements de type Webinar 2H

Event Webinar (Les jeudis de la Data & AI) - L200/300	Date	Duration (min)	Link
Azure Machine Learning pour les Data Scientists	15/09/2022	120	<a href="https://msevents.microsoft.com/event?id=2454281594">https://msevents.microsoft.com/event?id=2454281594</a>
Azure Synapse	22/09/2022	120	<a href="https://msevents.microsoft.com/event?id=857781749">https://msevents.microsoft.com/event?id=857781749</a>
Les solutions SQL dans Azure (PaaS, IaaS, SaaS)	29/09/2022	120	<a href="https://msevents.microsoft.com/event?id=502366997">https://msevents.microsoft.com/event?id=502366997</a>
Déploiement et sécurisation des workspaces Azure Machine learning	06/10/2022	120	<a href="https://msevents.microsoft.com/event?id=1505714138">https://msevents.microsoft.com/event?id=1505714138</a>
Azure Scale Analytics - Architectures Data Mesh dans Azure avec Azure Synapse, Microsoft Purview et Azure Data Share	13/10/2022	120	<a href="https://msevents.microsoft.com/event?id=139685175">https://msevents.microsoft.com/event?id=139685175</a>
MLOps avec Azure Machine Learning	20/10/2022	120	<a href="https://msevents.microsoft.com/event?id=1245885767">https://msevents.microsoft.com/event?id=1245885767</a>
SQL Server 2022 et hybridation native avec Azure SQL Managed Instance	10/11/2022	120	<a href="https://msevents.microsoft.com/event?id=145826476">https://msevents.microsoft.com/event?id=145826476</a>
Machine Learning dans Azure Synapse Analytics	17/11/2022	120	<a href="https://msevents.microsoft.com/event?id=3637723312">https://msevents.microsoft.com/event?id=3637723312</a>
Azure Cosmos DB et IA	24/11/2022	120	<a href="https://msevents.microsoft.com/event?id=2646013445">https://msevents.microsoft.com/event?id=2646013445</a>
Azure et les Services Cognitifs	08/12/2022	120	<a href="https://msevents.microsoft.com/event?id=3772037220">https://msevents.microsoft.com/event?id=3772037220</a>
La gouvernance de données dans Azure avec Microsoft Purview	15/12/2022	120	<a href="https://msevents.microsoft.com/event?id=1499560981">https://msevents.microsoft.com/event?id=1499560981</a>
MLOps avec Azure Machine Learning	12/01/2023	120	<a href="https://msevents.microsoft.com/event?id=4115194515">https://msevents.microsoft.com/event?id=4115194515</a>
	19/01/2023	120	<a href="https://msevents.microsoft.com/event?id=1537241181">https://msevents.microsoft.com/event?id=1537241181</a>
Data processing dans Azure ave Azure Synapse, Azure Batch, Spark, Notebook, etc.	26/01/2023	120	<a href="https://msevents.microsoft.com/event?id=1806467748">https://msevents.microsoft.com/event?id=1806467748</a>
Déploiement et sécurisation des workspace Azure Synapse	09/02/2023	120	En cours
Azure Machine Learning pour les Citizen Data Scientists	16/02/2023	120	<a href="https://msevents.microsoft.com/event?id=1401519679">https://msevents.microsoft.com/event?id=1401519679</a>
L'IA responsable avec Azure machine learning	09/03/2023	120	<a href="https://msevents.microsoft.com/event?id=2072953112">https://msevents.microsoft.com/event?id=2072953112</a>
Machine Learning dans Azure Synapse Analytics	16/03/2023	120	<a href="https://msevents.microsoft.com/event?id=3413014857">https://msevents.microsoft.com/event?id=3413014857</a>
Les bases de données Open Source dans le cloud Azure	23/03/2023	120	<a href="https://msevents.microsoft.com/event?id=2727487131">https://msevents.microsoft.com/event?id=2727487131</a>
Hybridation des services de Machine Learning Azure	06/04/2023	120	<a href="https://msevents.microsoft.com/event?id=1624914222">https://msevents.microsoft.com/event?id=1624914222</a>
La gouvernance de données dans Azure avec Microsoft Purview	13/04/2023	120	<a href="https://msevents.microsoft.com/event?id=3909342839">https://msevents.microsoft.com/event?id=3909342839</a>
Les solutions SQL dans Azure (PaaS, IaaS, SaaS)	04/05/2023	120	<a href="https://msevents.microsoft.com/event?id=1162207895">https://msevents.microsoft.com/event?id=1162207895</a>
	16/05/2023	120	<a href="https://msevents.microsoft.com/event?id=3517068442">https://msevents.microsoft.com/event?id=3517068442</a>
Data processing dans Azure ave Azure Synapse, Azure Batch, Spark, Notebook, etc.	24/05/2023	120	<a href="https://msevents.microsoft.com/event?id=2996507398">https://msevents.microsoft.com/event?id=2996507398</a>
Self Service Analytics	01/06/2023	120	En cours

Total

25

# Liste des évènements de type Workshop/Prepa Cert/Academy

Event Workshop L300/400	Date	Duration (min)	Link
Synapse et Lakehouse Database	18/10/2022	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u</a>
Le NoSQL dans Azure - Migration CosmosDB / OSS	08/11/2022	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u</a>
Lab Lyon - Data & AI	22/11/2022	240	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUMIZZORETORSWjcyTERYRkJGTIFFUjaUi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUMIZZORETORSWjcyTERYRkJGTIFFUjaUi4u</a>
Lab Nantes - Data & AI	29/11/2022	240	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUMIZZORETORSWjcyTERYRkJGTIFFUjaUi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUMIZZORETORSWjcyTERYRkJGTIFFUjaUi4u</a>
L'IA dans les services de données Azure	12/12/2022	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u</a>
Open AI	H2	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFSDVLR0cyS1kwWS4u</a>

Event Academy, kickstart certifications, workshop certifications	Date	Duration (min)	Link
MCPP - DP-420	10/10/2022	420	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUMkJSIRKSU1RRFA0OVgzSFdtSTY0RE9WQ4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUMkJSIRKSU1RRFA0OVgzSFdtSTY0RE9WQ4u</a>
Micro Hack CosmosDB	20/10/2022	420	<a href="#">H1 - Inscriptions PTA</a>
Academy DP900	17-21/10/2022	300	<a href="https://msevents.microsoft.com/event?id=3250818161">https://msevents.microsoft.com/event?id=3250818161</a>
Academy AI900	17-21/10/2022	300	<a href="https://msevents.microsoft.com/event?id=2717528090">https://msevents.microsoft.com/event?id=2717528090</a>
Kickstart DP-500	17/10/2022	60	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTO2Uk85V0cxQzM3TE9ZRS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTO2Uk85V0cxQzM3TE9ZRS4u</a>
Dry Run DP-500	14/11/2022	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTO2Uk85V0cxQzM3TE9ZRS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTO2Uk85V0cxQzM3TE9ZRS4u</a>
Q&A DP-500	05/12/2022	90	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTO2Uk85V0cxQzM3TE9ZRS4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTO2Uk85V0cxQzM3TE9ZRS4u</a>
Kickstart DP-100	17/10/2022	60	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI30UNMFYxSkRIMi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI30UNMFYxSkRIMi4u</a>
Dry Run DP-100	14/11/2022	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI30UNMFYxSkRIMi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI30UNMFYxSkRIMi4u</a>
Q&A DP-100	05/12/2022	90	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI30UNMFYxSkRIMi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI30UNMFYxSkRIMi4u</a>
Kickstart DP-203	17/10/2022	60	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyOk5SVjFBUFczNktCUFpLMi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyOk5SVjFBUFczNktCUFpLMi4u</a>
Dry Run DP-203	14/11/2022	120	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyOk5SVjFBUFczNktCUFpLMi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyOk5SVjFBUFczNktCUFpLMi4u</a>
Q&A DP-203	05/12/2022	90	<a href="https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyOk5SVjFBUFczNktCUFpLMi4u">https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHB3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyOk5SVjFBUFczNktCUFpLMi4u</a>

Total

16

# Azure Synapse Analytics

La modélisation de Datawarehouse n'a plus de secret pour vous ? Et si on parlait de la nouvelle approche de modélisation d'objets dans Azure Synapse. Nous parcourrons les notions de virtualisation, de lake database et de l'orientation Spark de la plateforme.

## Agenda (2h)

### Service overview / updates

- Data Intégration
- Data Engineering
- Data Warehousing
- Data Science
- Observational Analytics
- Business Intelligence
- Gouvernance



Ali Bouhaddou  
Cloud Solution Architect  
Data Analytics



Frederic Gisbert  
Cloud Solution Architect  
Data Analytics



Franck Gaillard  
Cloud Solution Architect  
Data AI



Narjes Majdoub  
Cloud Solution Architect  
Data AI

# Azure Synapse Analytics

Service overview / updates March 2022



From Data to Intelligence

Drive a data culture and power a new  
class of data first applications

everyone | every decision | at any scale

# Azure Synapse and Power BI



Data  
Integration



Analytics



Business  
Intelligence



# Azure Synapse Analytics

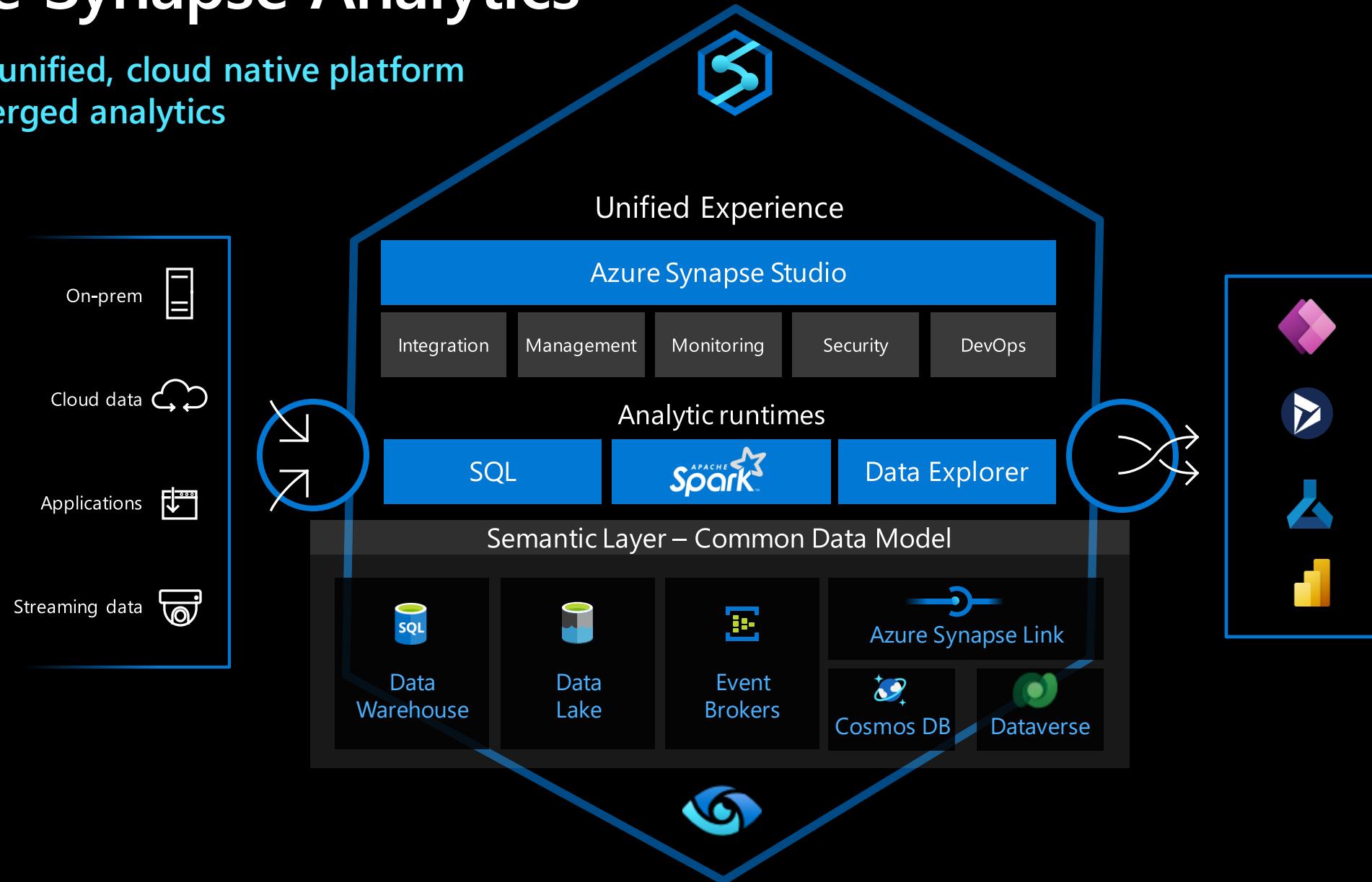
The first unified, cloud native platform for converged analytics



Azure Synapse is the only unified platform for analytics, blending big data, data warehousing, and data integration into a single cloud native service for end-to-end analytics at cloud scale.

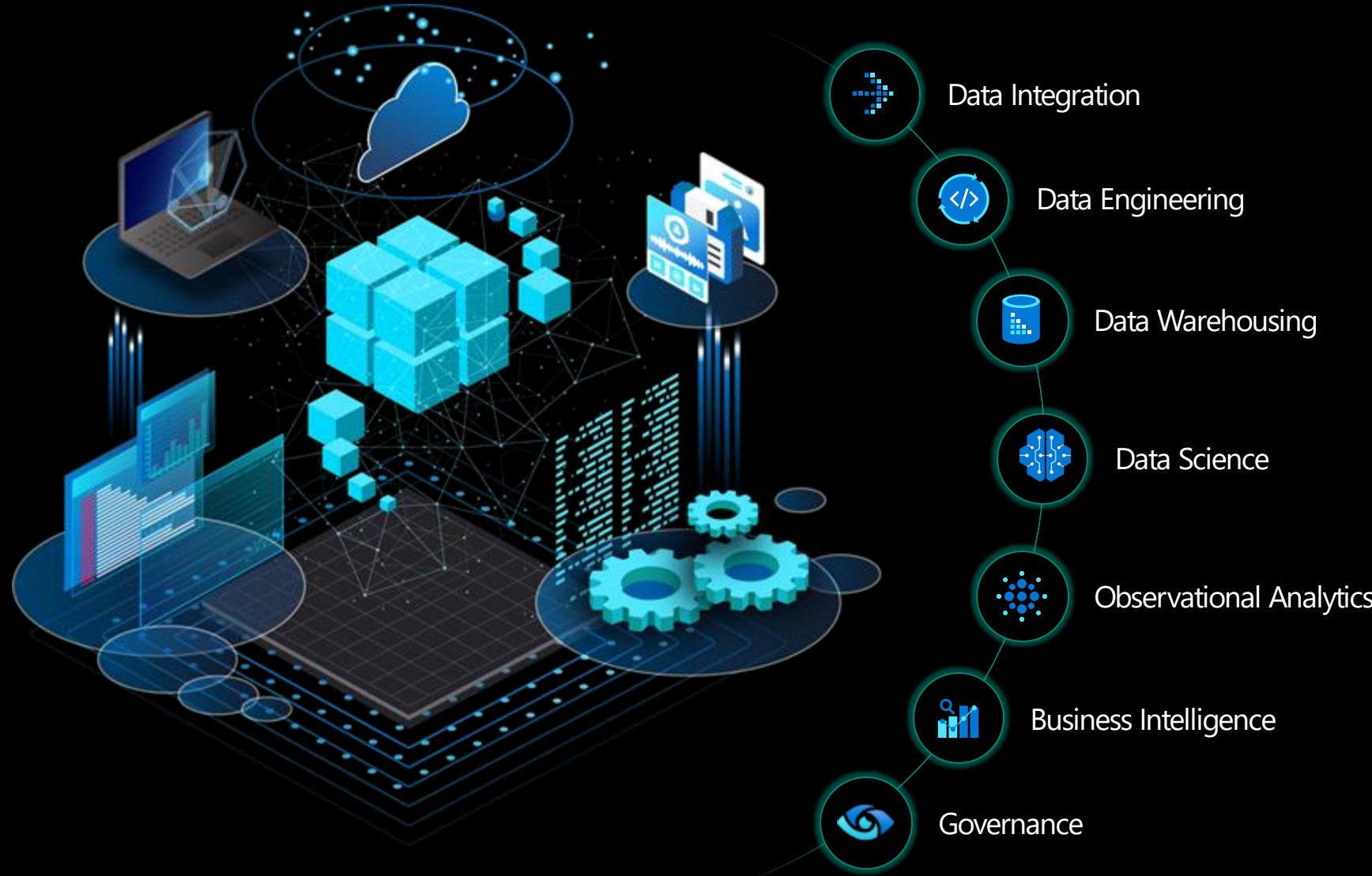
# Azure Synapse Analytics

The first **unified, cloud native platform**  
for converged analytics





# Synapse + Power BI





Data Integration

## Data Integration



Over 100 connectors to ingest  
data from a variety of platforms

---

Integrate from On-Premise, PaaS, and SaaS

---

Batch and Real-time data integration

---

Secure hybrid connectivity

---

Code-free development environment

## Generally Available

# 100+ Connectors

Connect on data sources in Azure, on-premise, other clouds, and SaaS applications

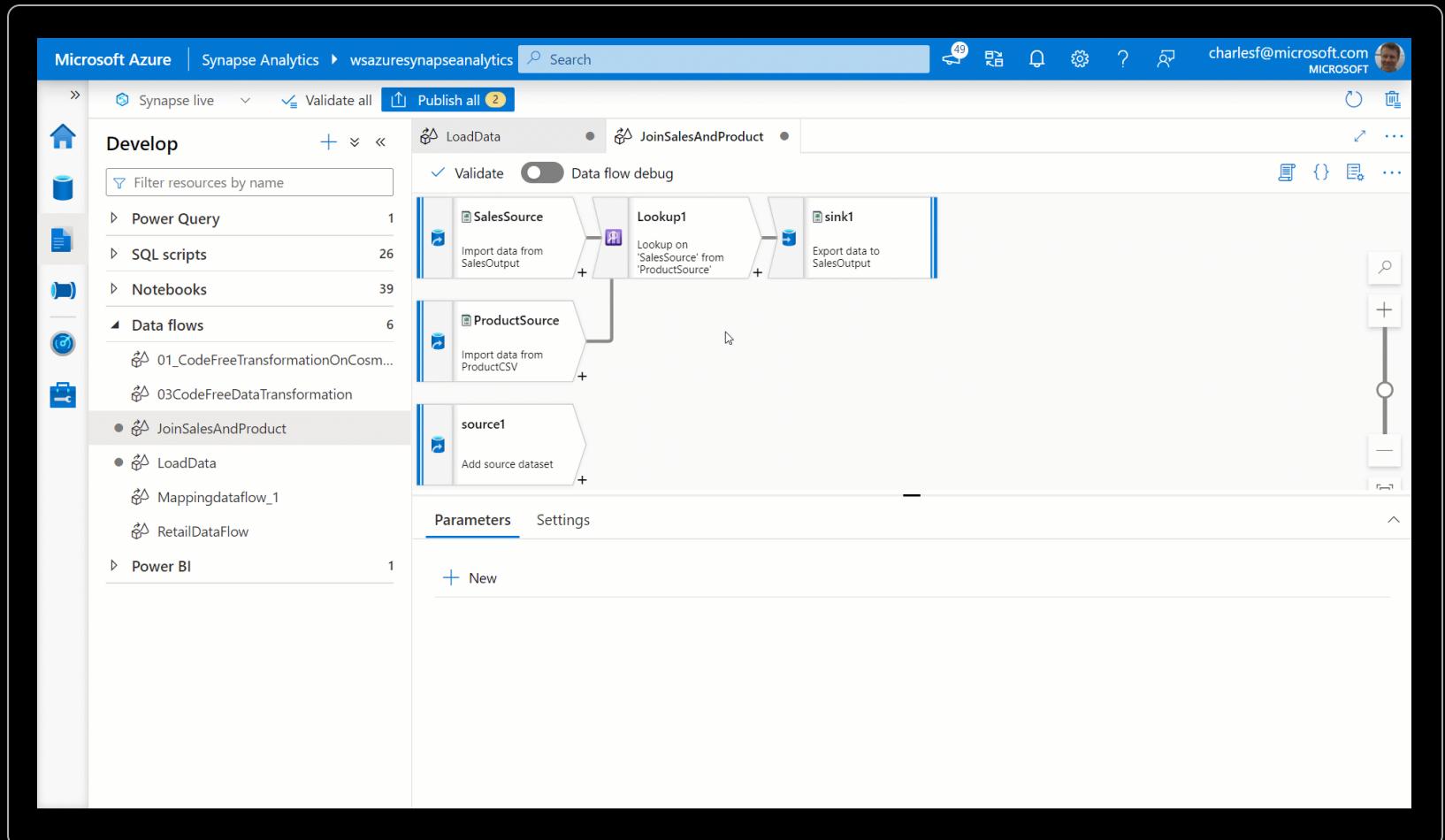
The screenshot shows the Microsoft Azure Synapse Analytics Data Flow blade. On the left, the 'Develop' sidebar lists various resources: Power Query (1), SQL scripts (26), Notebooks (39), Data flows (6), Power BI (1). The 'Data flows' section is expanded, showing a list of data flows: 01\_CodeFreeTransformationOnCosm..., 03CodeFreeDataTransformation, JoinSalesAndProduct, LoadData, Mappingdataflow\_1, RetailDataFlow. The 'LoadData' item is selected and highlighted in grey. The main workspace displays a single data flow step named 'LoadData' with a single input node labeled 'SourceData' which has '0 total' columns. Below the workspace, the 'Source settings' tab is active, showing the 'Output stream name' as 'SourceData', 'Source type' set to 'Integration dataset', and a dropdown for 'Dataset' with 'Select...' highlighted. Other tabs include 'Source options', 'Projection', 'Optimize', 'Inspect', and 'Data preview'. A 'Description' button is also present.

## Generally Available

# Code-free Data Flows

Enables developers to rapidly integrate data from a variety of sources

Execute on Spark for large scale processing



# Code-free Data Flows



Handle upserts, updates,  
deletes on sql sinks



Add new partition methods



Add schema drift support



Add file handling (move files  
after read, write files to file  
names described in rows etc)



New inventory of functions  
(for e.g. Hash functions for  
row comparison)



Commonly used ETL  
patterns(Sequence  
generator/Lookup  
transformation/SCD...)



Data lineage – Capturing sink  
column lineage & impact  
analysis(invaluable if this is  
for enterprise deployment)

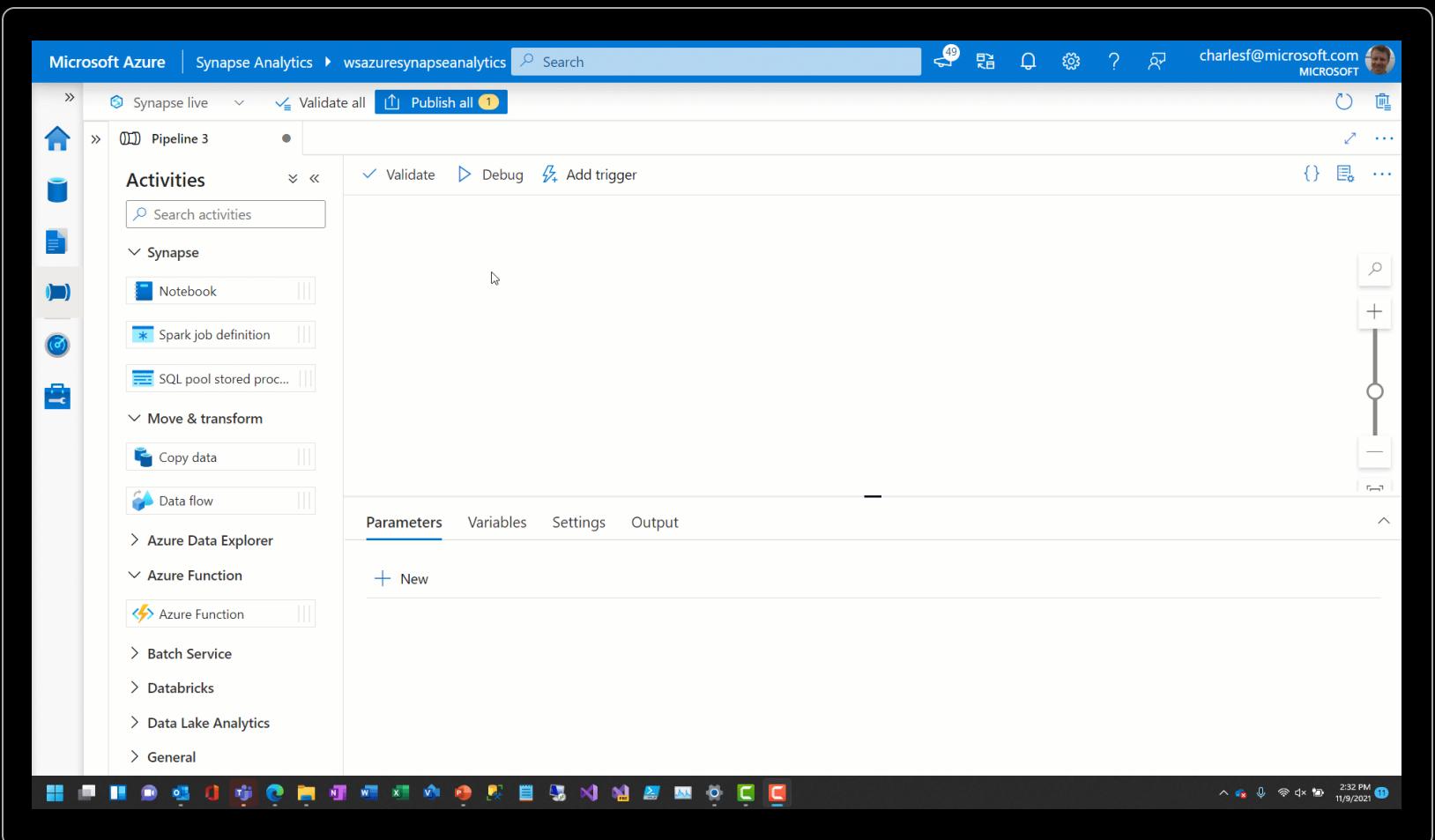


Implement commonly used  
ETL patterns as  
templates(SCD Type1, Type2,  
Data Vault)

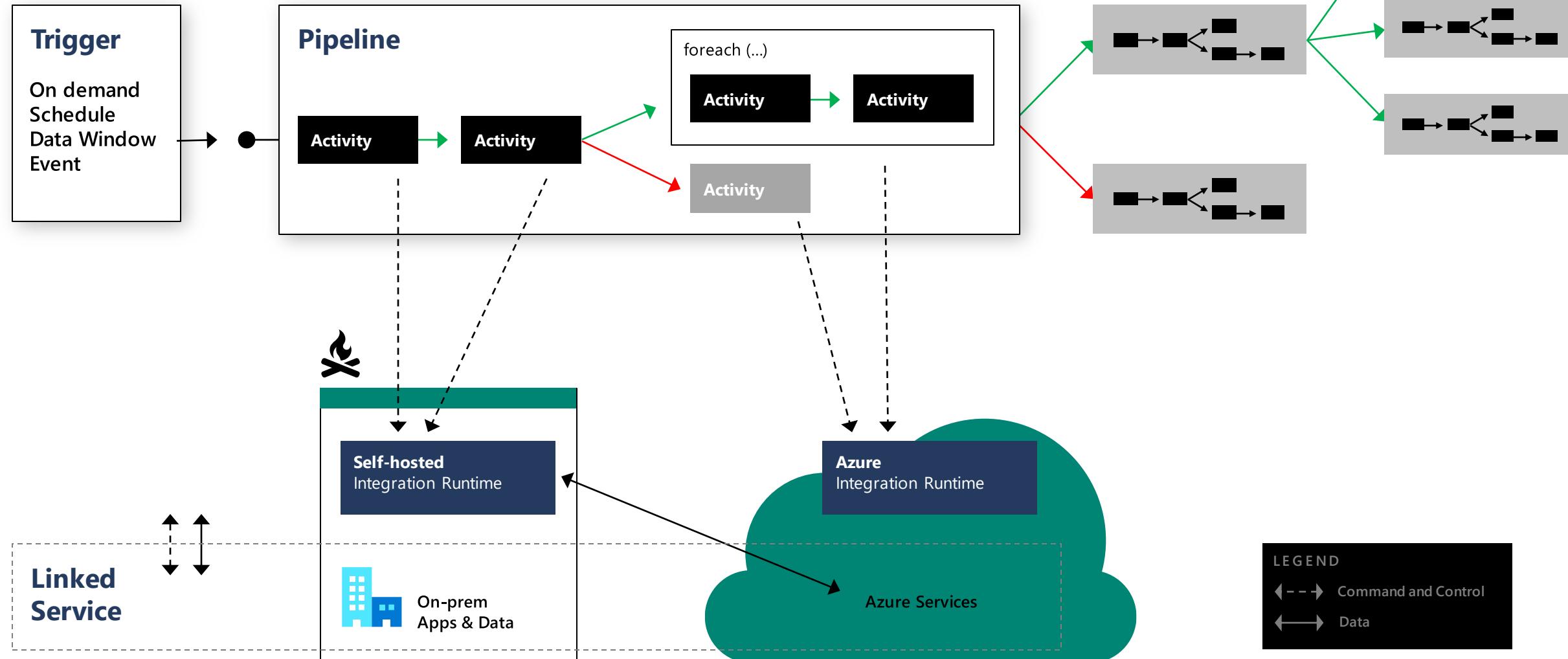
## Generally Available

# Pipeline Orchestration

Code-free experience for  
orchestrating a sequence of  
data integration tasks



# Pipeline Orchestration

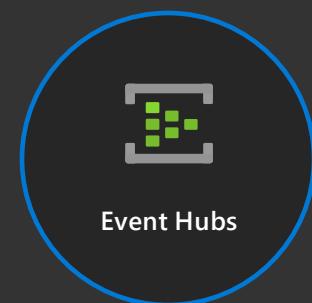


## Generally Available

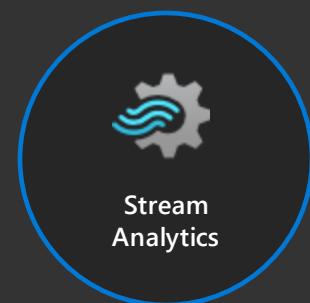
# Real-time Streaming Data Integration

Enables IoT data streams from event brokers to load directly into the data warehouse or data lake

Analyze data in-flight with temporal T-SQL queries in Stream Analytics



Event Hubs

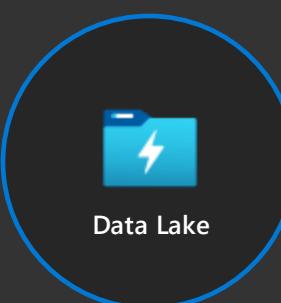


Stream  
Analytics

SQL Query  
Language



Data  
Warehouse



Data  
Lake

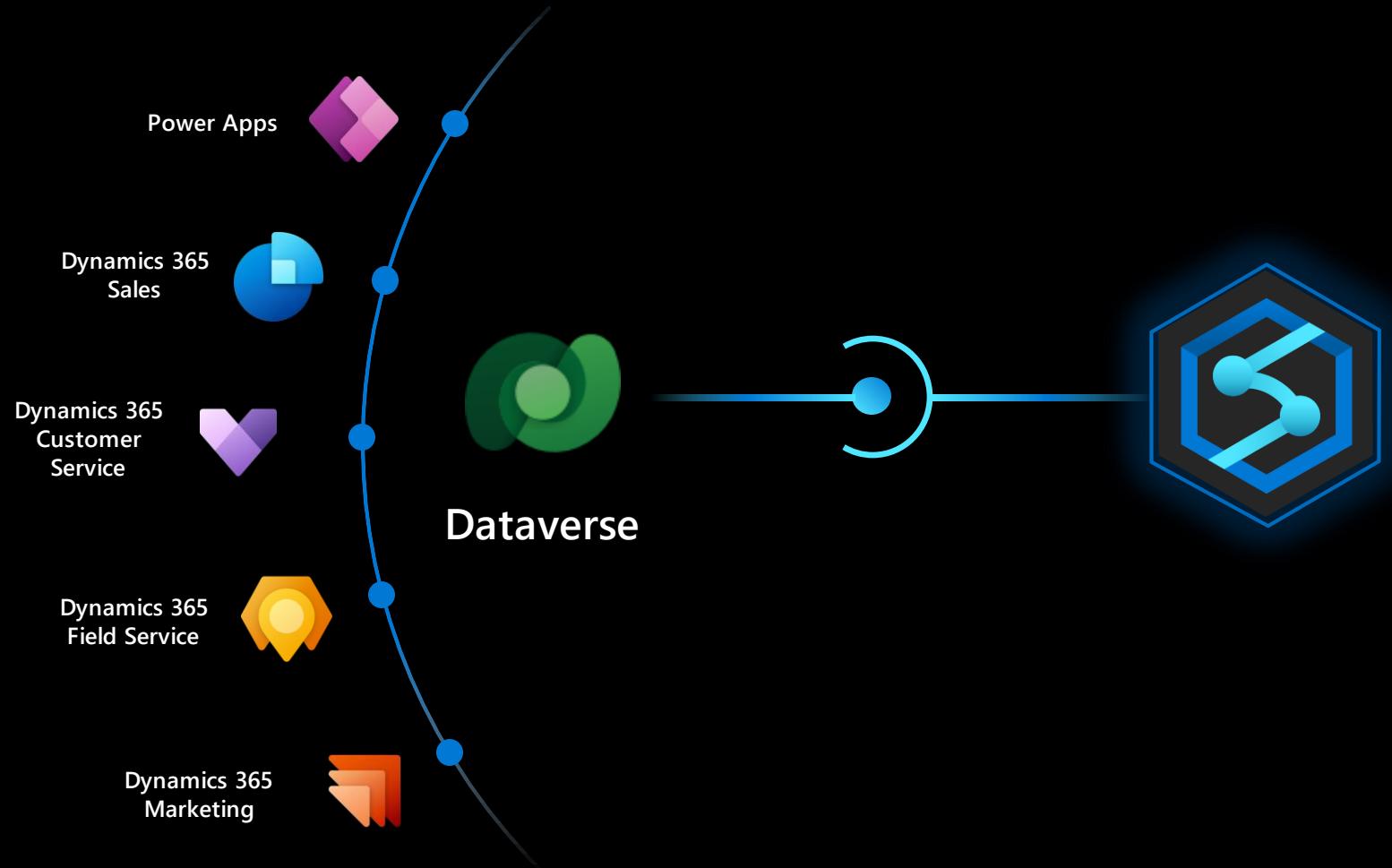
## Generally Available

November 2021

# Synapse Link for Dataverse

One-click integration of D365  
data into Synapse  
for analytics

No data pipelines required



## Public Preview

Q2 2022

# Parquet & Virtual Network Support for Dataverse

Parquet columnar file format optimizes query performance for user queries

Enables customers to apply Virtual Network security to Dataverse connection



## Public Preview

Q2 2022  
(SQL Server 2022 and  
Azure SQL Database)

# Synapse Link for Microsoft SQL

---

Near real-time operational  
analytics in Synapse

No data pipelines required

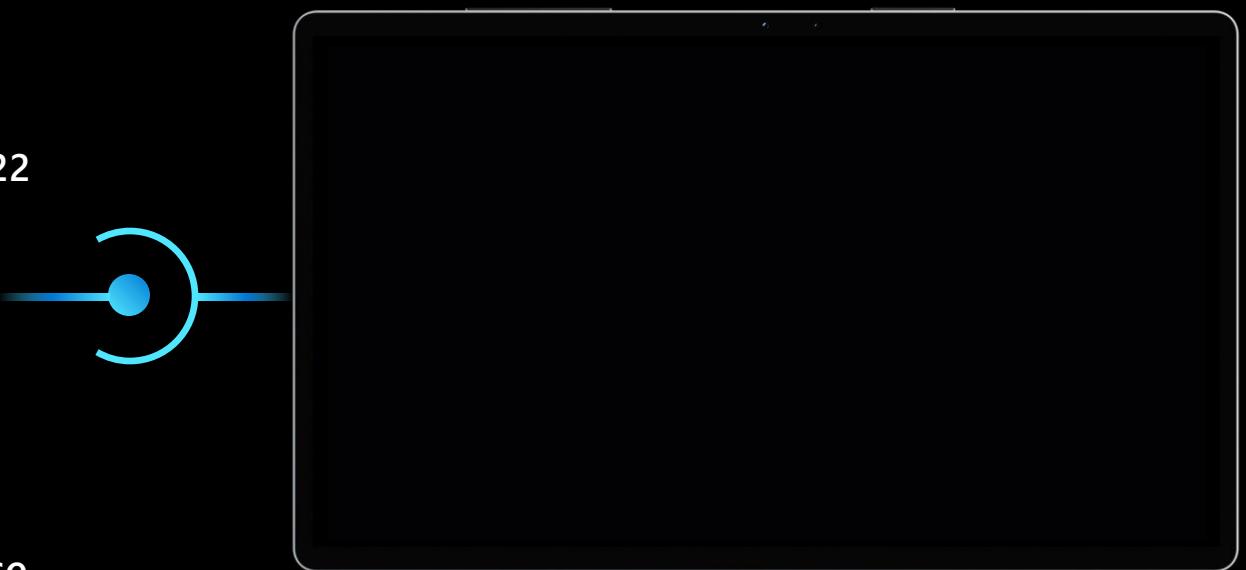
Hybrid integration for SQL  
Server running on-premise or  
in other clouds



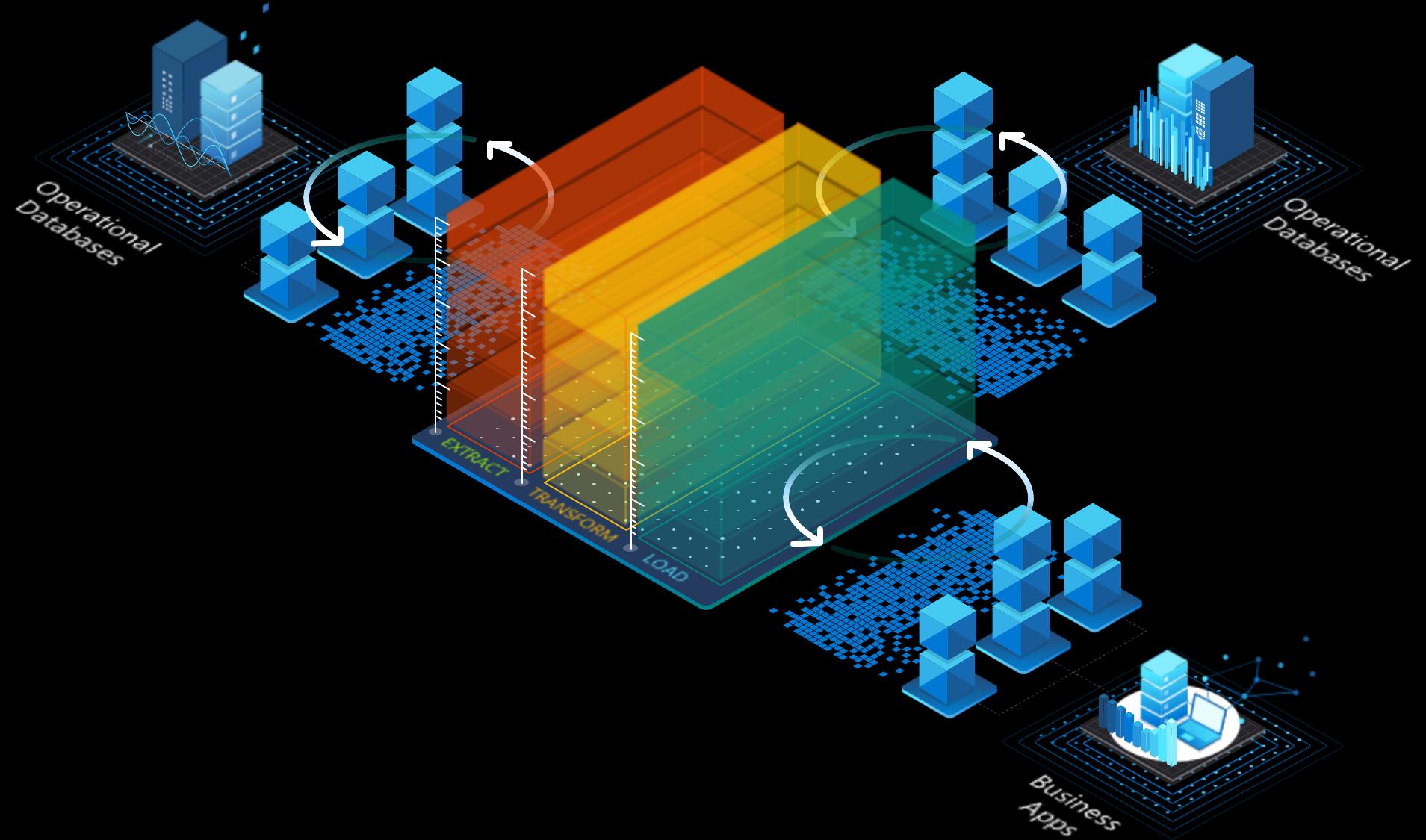
SQL Server 2022  
Public Preview Q2 2022



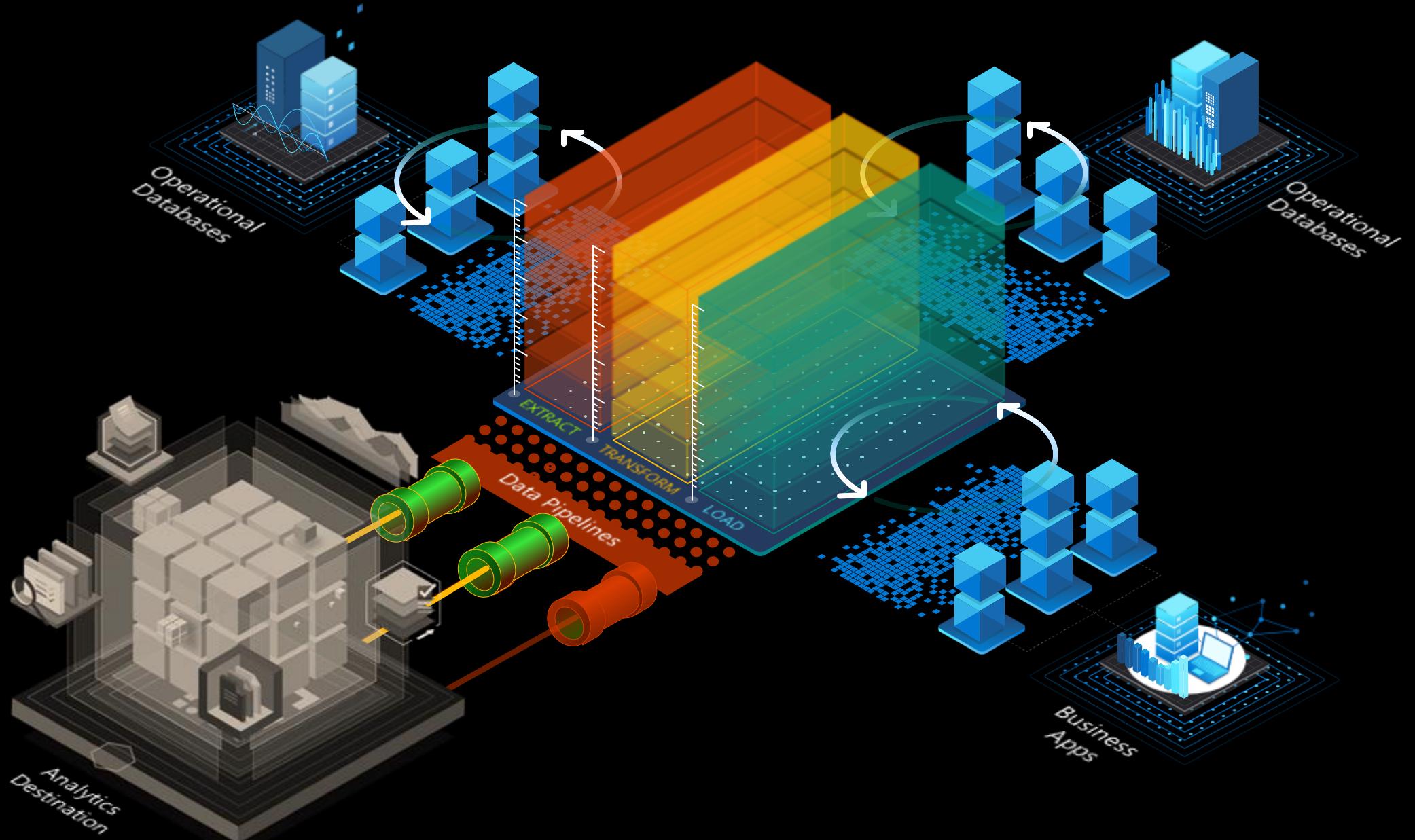
Azure SQL Database  
Public Preview Q2 2022



# Preparing data to be analytics-ready requires various processes



These processes result in additional issues that stall data insights



# Azure Synapse Link

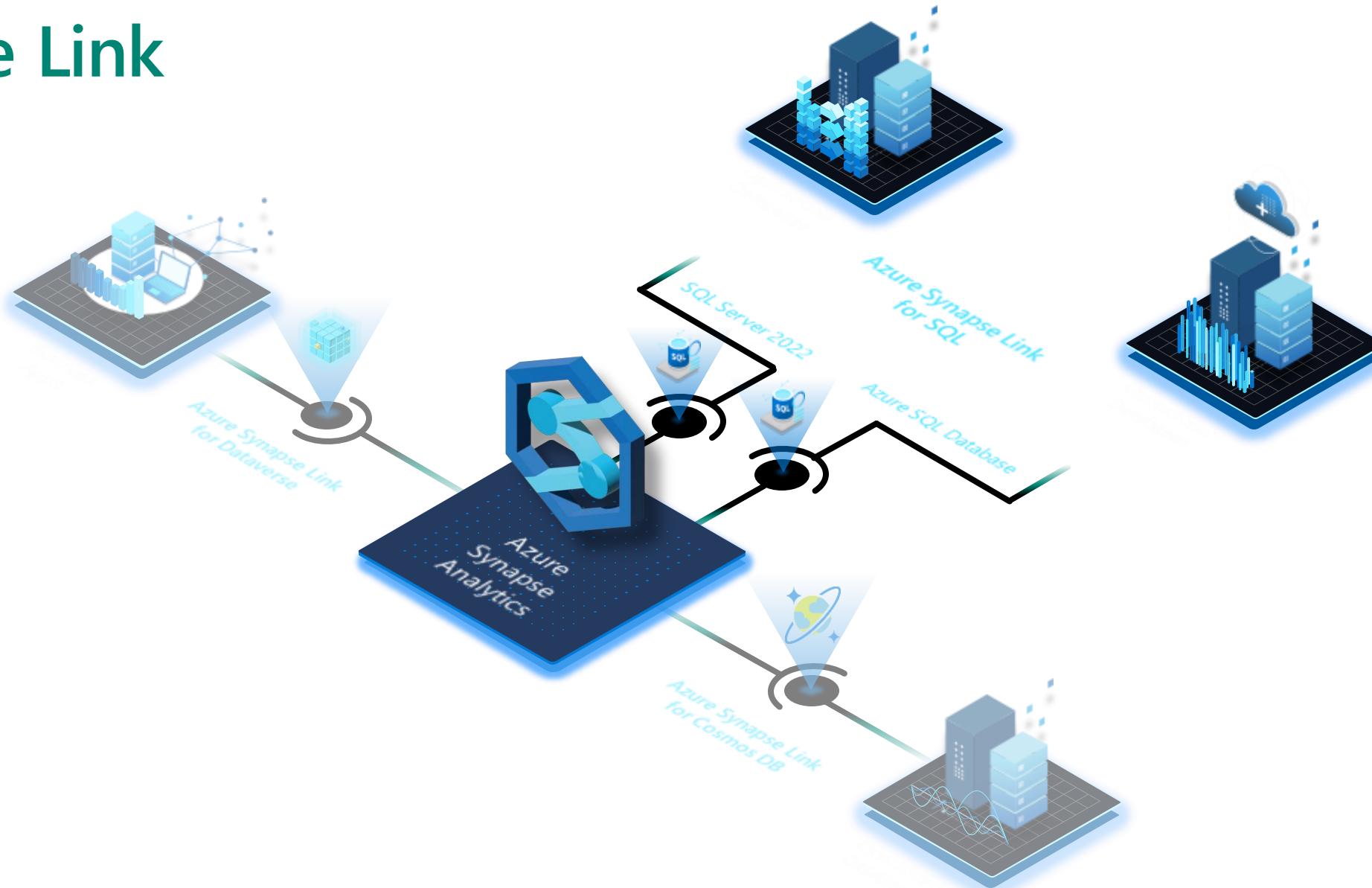




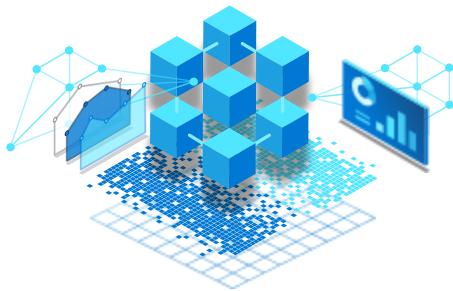
## Azure Synapse Link

Go from after-the-fact analysis to near real-time insights by eliminating barriers between data and analytics. Automatically move data from both operational databases and business applications without time-consuming ETL. Deliver critical insights by easily connecting separate systems, bringing the power of analytics to every data-connected team.

# Azure Synapse Link



# Azure Synapse Link for SQL



Simplified  
experience

With a streamlined setup, transfer data from operational systems seamlessly



Centralized  
data

Consolidate data from multiple sources into a single analytics solution



Near real-time  
insights

Maximize impact with automated systems, capturing and updating data in near real-time

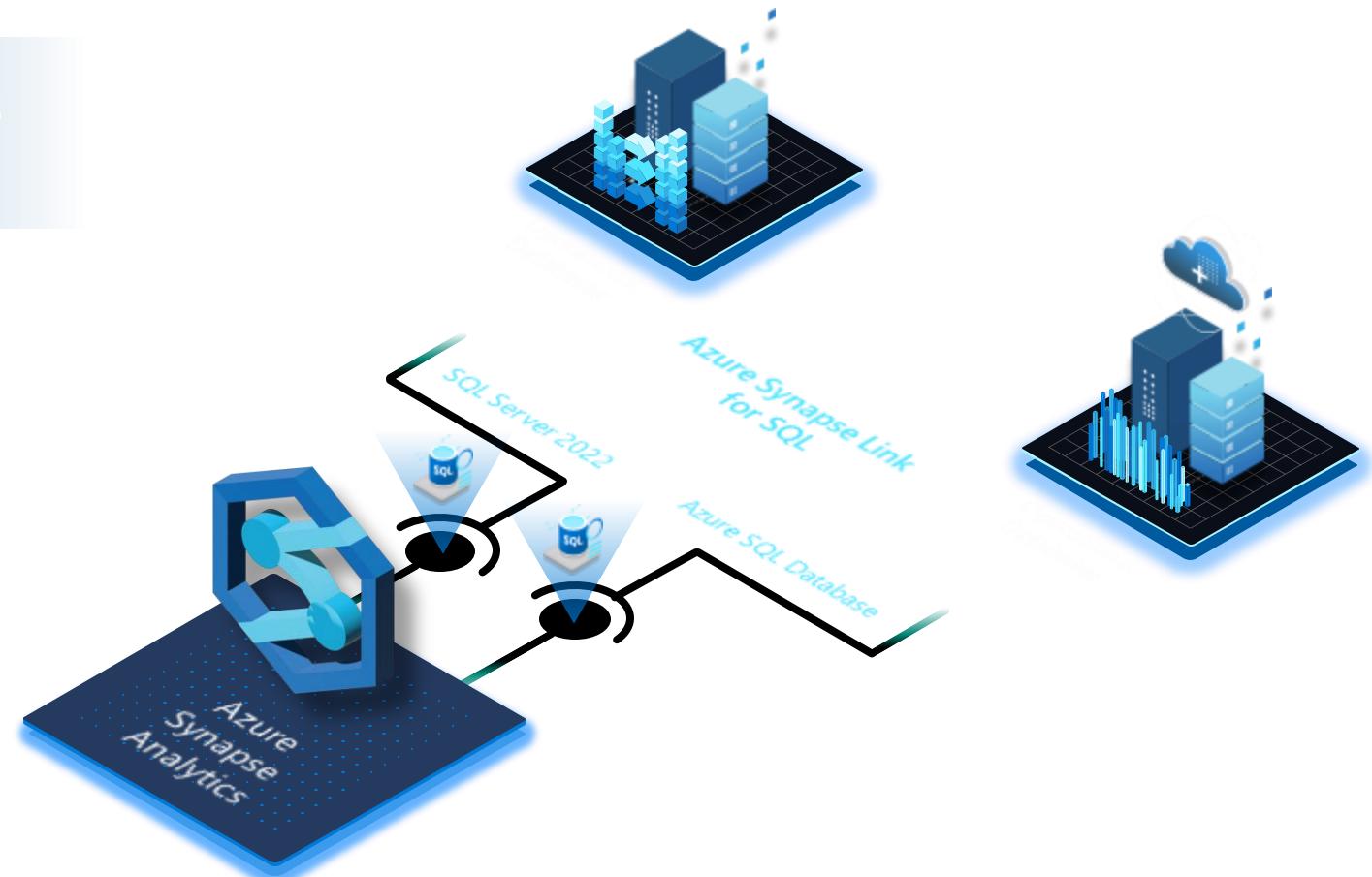


Minimized  
overhead

Record data changes without stretching source systems

# Azure Synapse Link for SQL

Automate data transfer and eliminate barriers  
with Azure Synapse Link for SQL



# Preparing data to be near real-time insights

Synapse live Validate all Publish all

## Integrate

- +
- Pipeline
- Link connection (Preview)**
- Copy Data tool
- Browse gallery
- Import from pipeline template

Filter resources by name

Link connection

New link connection

Synapse Link connection allows you to perform analytical processing on your operational data. [Learn more](#)

Source database settings  
Select a source and the set of tables you would like to include in the link connection.

Source type \* SQL Server

Source linked service \* SqlServer22

Source tables \*  
Filter by source table name

Name	Preview
dbo.table1	
dbo.Table2	
dbo.Table3	
dbo.Table4	

Show 1-11 of 11 items (1 selected)

Continue Cancel

New link connection

Connection settings  
Provide a name and select compute settings for the link connection.

Link connection name \* sql2gettingstart

Core count \* 4 (+ 4 Driver cores)

Landing zone linked service \* GettingStartedLZ

Landing zone folder path \* landingzone

Landing zone SAS token \*   
[Generate token](#)

This PREVIEW feature is licensed to you as part of your Azure subscription. By clicking "OK" you agree to the [Preview Terms](#) and [Privacy Statement](#).

OK Back Cancel

Synapse live Validate all Publish all

## Integrate

- +
- Link connection**
- Linkconnection1
- Linkconnection-azuresq0
- Linkconnection-demo

Stop New table Refresh

Filter by source table name

Source table	Target table	Distribution Type	Distribution Count
dbo.Table_1	dbo.Table_1_db2	Round robin	-
dbo.Table_2	dbo.Table_2_db2	Round robin	-



## Public Preview

Q2 2022

# SSIS Integration Runtime for Azure Synapse

Running packages deployed into SSIS catalog (SSISDB) hosted by Azure SQL Database server/Managed Instance (Project Deployment Model)

The screenshot shows the Azure Synapse Analytics studio interface. On the left, there is a navigation sidebar with the following items:

- Analytics pools
- SQL pools
- Apache Spark pools
- Data Explorer pools (preview)
- External connections
- Linked services
- Azure Purview
- Integration** (highlighted with a red box labeled 1)
- Triggers
- Integration runtimes** (highlighted with a red box labeled 2)
- Security
- Access control
- Credentials
- Managed private endpoints
- Code libraries
- Workspace packages
- Source control
- Git configuration

The main content area is titled "Integration runtimes". It contains a brief description: "The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment." Below this is a table listing two existing runtimes:

Name	Type	Sub-type	Status	Related	Region
AutoResolveIntegrationRuntime	Azure	Public	Running	2	Auto Resolve
FilesystemSSISR	Azure-SSIS	---	Running	1	Southeast Asia

At the top right of the main content area, there are three buttons: "+ New" (highlighted with a red box labeled 3), "Refresh", and a "Filter by name" search bar.

Running packages deployed into Azure Files (Package Deployment Model)

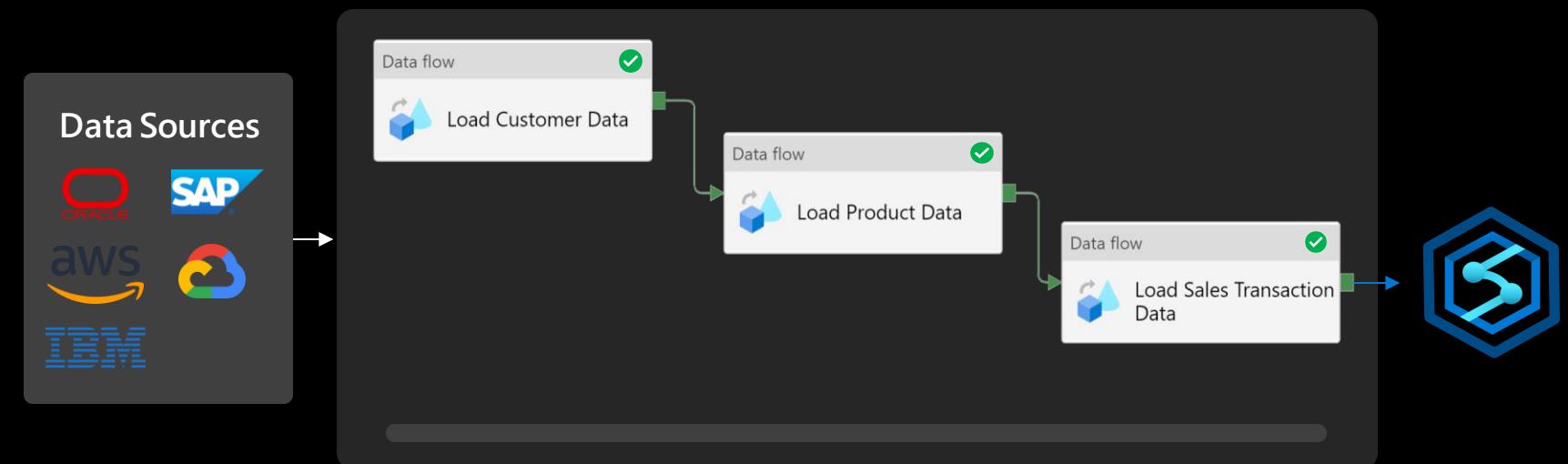
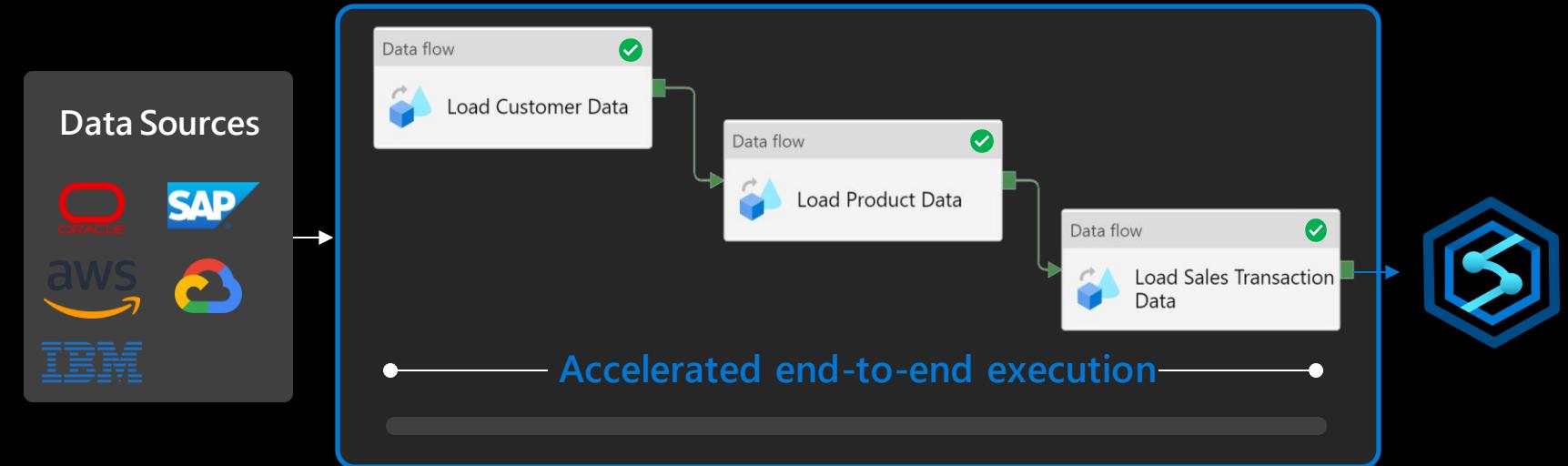
## Generally Available

October 2021

# Accelerated Data Pipelines

Cluster time-to-live enables  
near instant start of data flow  
pipelines for faster data  
integration

Data is available to the  
business faster to enable  
more timely decision making







Data Engineering

## Data Engineering



**Scalable Spark engine**

**Industry standard languages**

**Delta Lake Enabled**

**Azure DevOps integration**

## Public Preview

Q2 2022

# Spark 3.2

---

Enables developers can leverage the latest innovations in the Spark ecosystem

### Pandas (Koalas) integration

A highly popular and flexible library with broad industry adoption

### Adaptive Query Execution (AQE) enabled by default

Significant improvements in query performance out-of-the-box

### Small Query execution improvements

Small queries run faster due to reduced initialization overhead

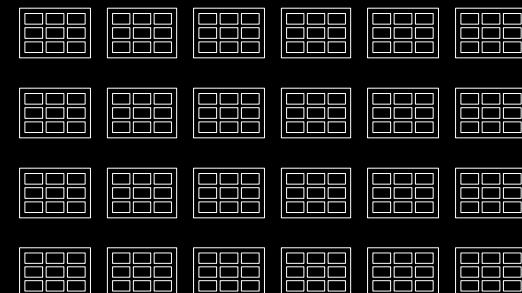
## Public Preview

Q2 2022

# Delta Lake Performance Enhancements

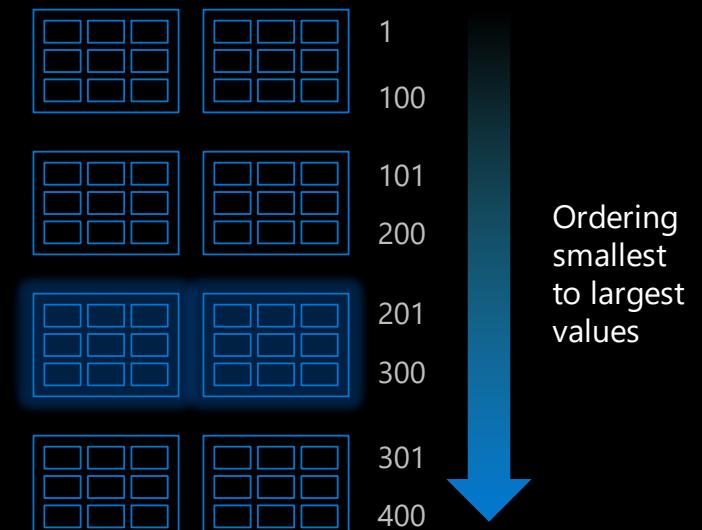
Improved performance and  
reduced cost with support for  
OPTIMIZE and Z-ORDER

## OPTIMIZE



Improve query performance  
by coalescing small files into  
larger ones

## Z-ORDER



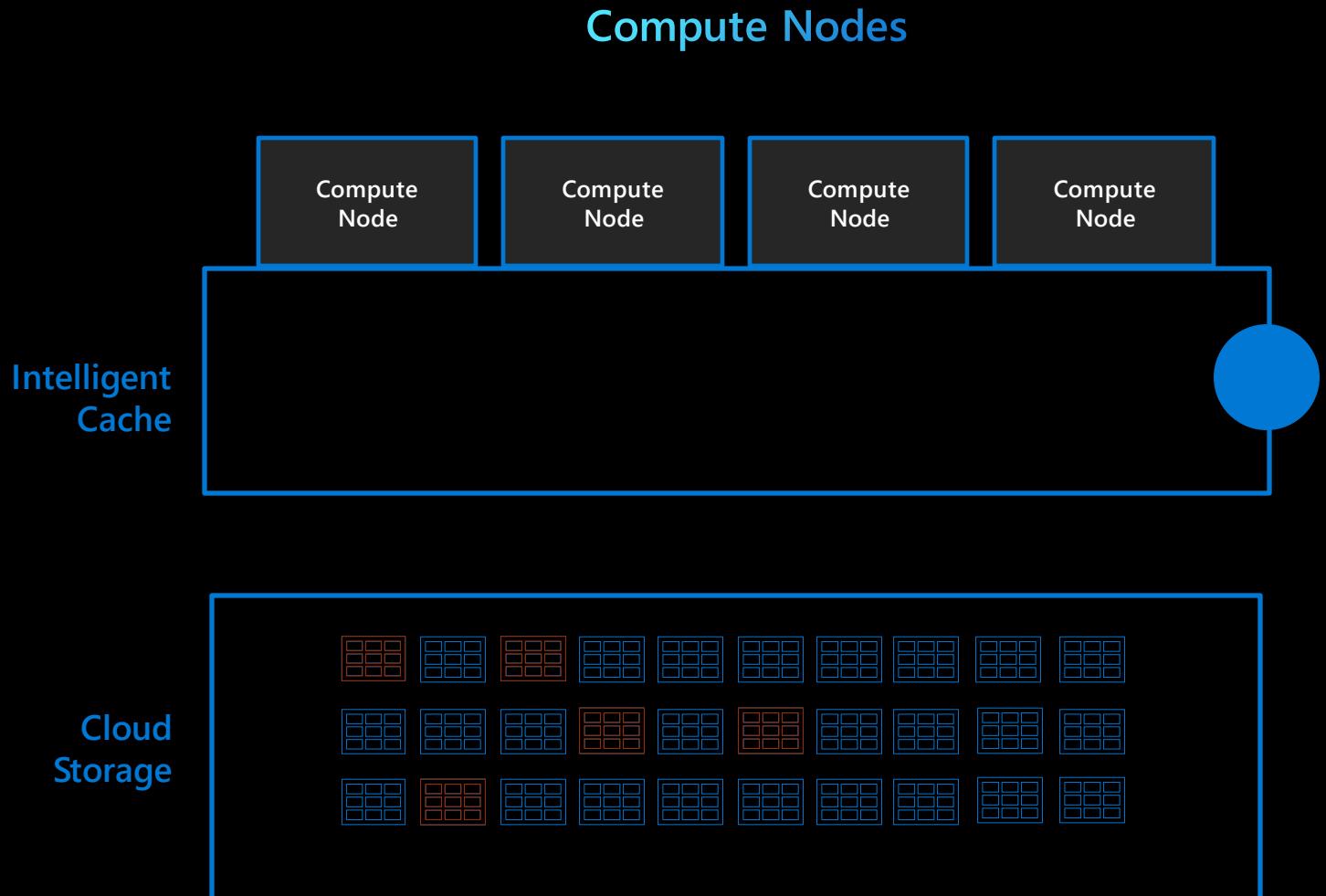
Improves filter query performance  
by ordering data for fast lookups on  
large datasets

## Public Preview

Q1 2022

# Spark Cashing Enhancements

Intelligent cache automatically  
detects changes in data to  
ensure data is fresh and  
results are accurate







Data Warehousing

## Data Warehousing



**Cornerstone of enterprise analytics for decades**

**Industry standard SQL language**

**Structured and semi-structured data**

**Broad ecosystem of applications**

**Fine-grained data security**

**Data models tailored to business consumption**

## Generally Available

### Dedicated & Serverless SQL

Elastic clusters with in-memory caching provide enterprise class performance combined with cloud economics



Serverless



Dedicated

## Generally Available

# Most Complete Workload Management

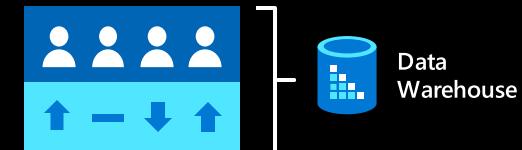
Scale-in to maximize output with the predictable cost

Scale-out to leverage cloud scale resources for spikes in demand

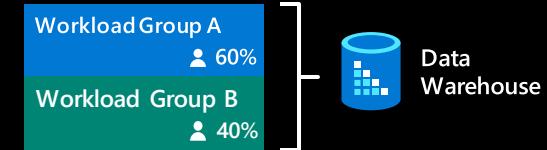
### Scale-In

- Predictable cost
- Prioritize higher value work
- Prevents global contention

### Workload Importance



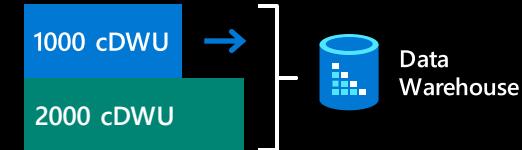
### Workload Isolation



### Scale-Out

- Add compute for variable workloads
- Pause compute when idle

### Elastic Cluster (Scale Up)



## Generally Available

# Complete Data Protection

Democratize data compliantly with fine-grained access controls and multi-level encryption

## Category

Category	Feature
Data Protection	Data in transit
	Data encryption at rest
	Data discovery and classification
Access Control	Object level security (tables/views)
	Row level security
	Column level security
	Dynamic data masking
	Column level encryption
Authentication	SQL login
	Azure active directory
	Multi-factor authentication
Network Security	Managed virtual network
	Custom virtual network
	Firewall
	Azure ExpressRoute
	Azure Private Link
Threat protection	Threat detection
	Auditing
	Vulnerability assessment
Isolation	Dedicated metadata store
	Hosted in customer tenant

## Generally Available

# Democratize ML predictions with SQL

In-engine ML scoring provides interactive query response times without any data leaving the system and no additional scoring cost



```
SELECT d.*, p.Score FROM PREDICT(MODEL = @onnx_model, ...)
```

### Synapse SQL



Model



Data



Predictions



T-SQL Language



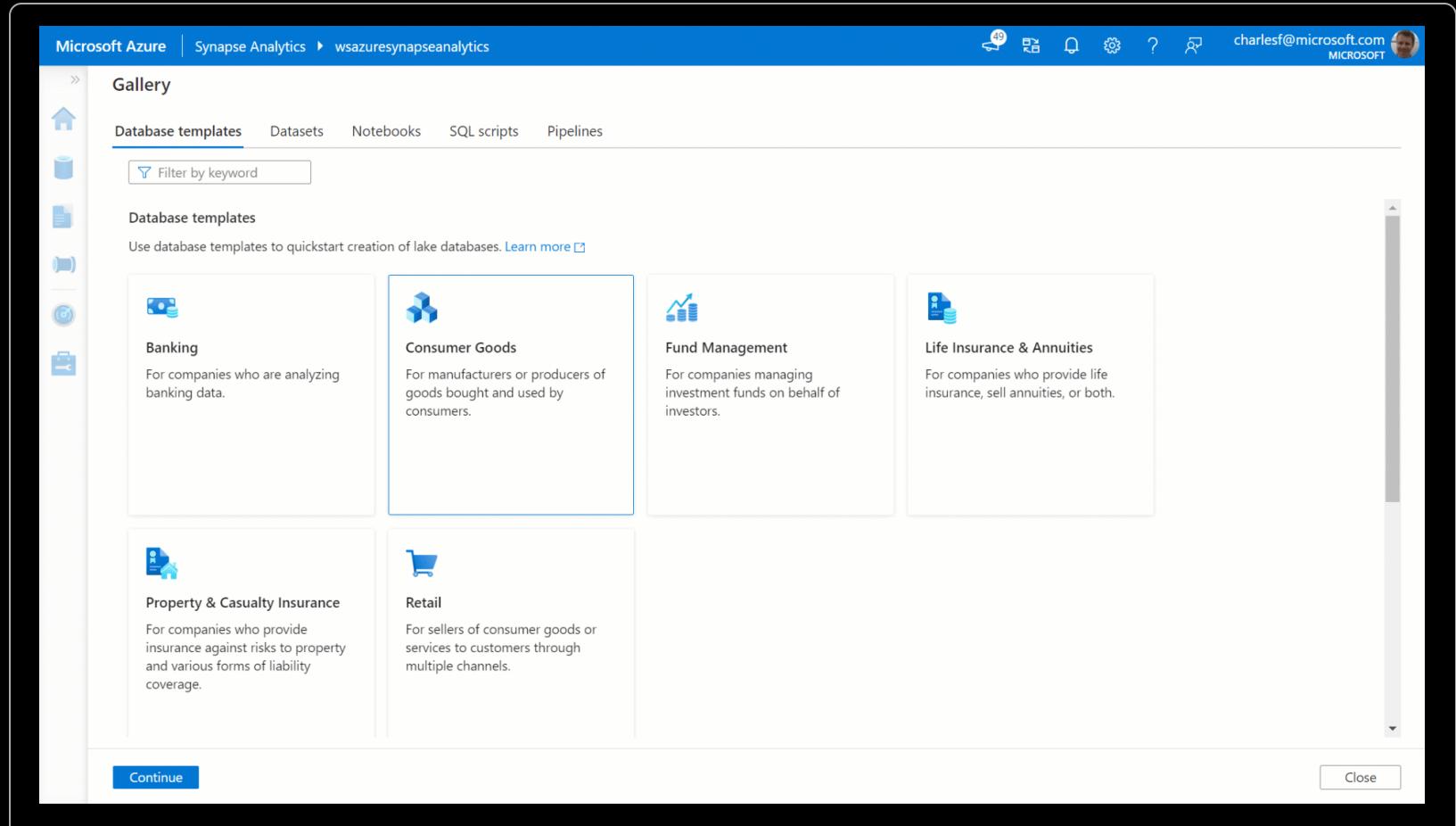
## Generally Available

Q2 2022

# Database template

Built-in database templates

Low code database designer



The screenshot shows the Microsoft Azure Synapse Analytics Gallery interface. The top navigation bar includes 'Microsoft Azure' and 'Synapse Analytics' with a workspace name 'wsazuresynapseanalytics'. The left sidebar has a 'Gallery' section with icons for 'Database templates', 'Datasets', 'Notebooks', 'SQL scripts', and 'Pipelines'. Below the sidebar, the 'Database templates' tab is selected. A 'Filter by keyword' search bar is present. The main area displays six database template cards:

- Banking**: For companies who are analyzing banking data.
- Consumer Goods**: For manufacturers or producers of goods bought and used by consumers.
- Fund Management**: For companies managing investment funds on behalf of investors.
- Life Insurance & Annuities**: For companies who provide life insurance, sell annuities, or both.
- Property & Casualty Insurance**: For companies who provide insurance against risks to property and various forms of liability coverage.
- Retail**: For sellers of consumer goods or services to customers through multiple channels.

At the bottom of the screen are 'Continue' and 'Close' buttons.

## Public Preview

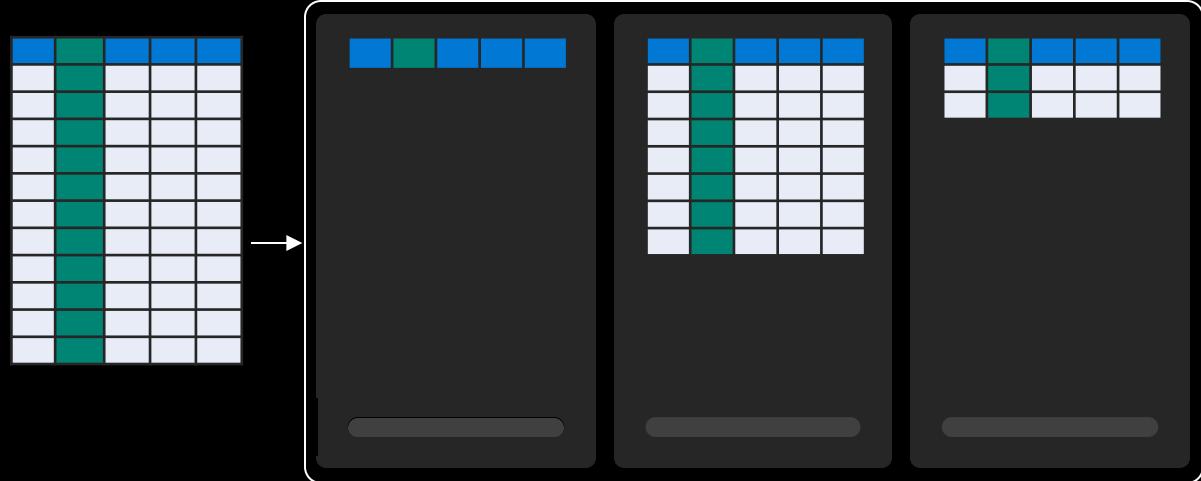
Q2 2022

# Multi-column distribution table sharing

Improved query performance and easier migrations

## Uneven Distribution (Skew)

```
CREATE TABLE SalesTransactions (
    WITH DISTRIBUTION =
        (HASH(ProductKey))
```



## Balanced Distribution

```
CREATE TABLE SalesTransactions (
    WITH DISTRIBUTION =
        (HASH(ProductKey, RegionKey))
```



Balanced Distribution: Queries execute faster

## Public Preview

Q2 2022

# MERGE SQL Statement

Improved performance and easier migration by executing INSERT, UPDATE, and DELETE functionality in a single statement

INSERT ...

UPDATE ...

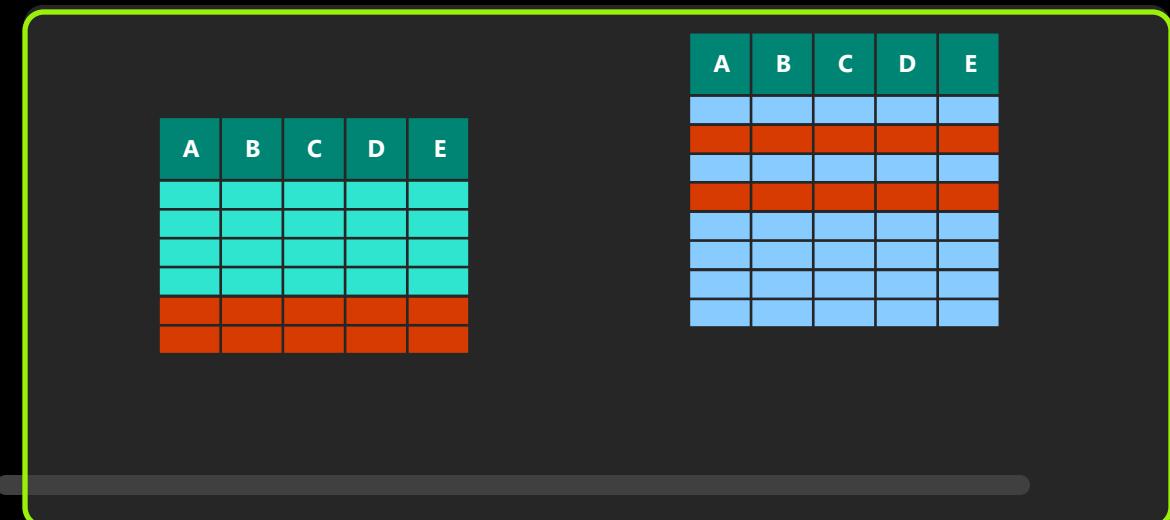
DELETE ...



MERGE ...

A	B	C	D	E
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Red	Red	Red	Red	Red
Red	Red	Red	Red	Red
Red	Red	Red	Red	Red
Red	Red	Red	Red	Red

A	B	C	D	E
Blue	Blue	Blue	Blue	Blue
Red	Red	Red	Red	Red
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue
Blue	Blue	Blue	Blue	Blue



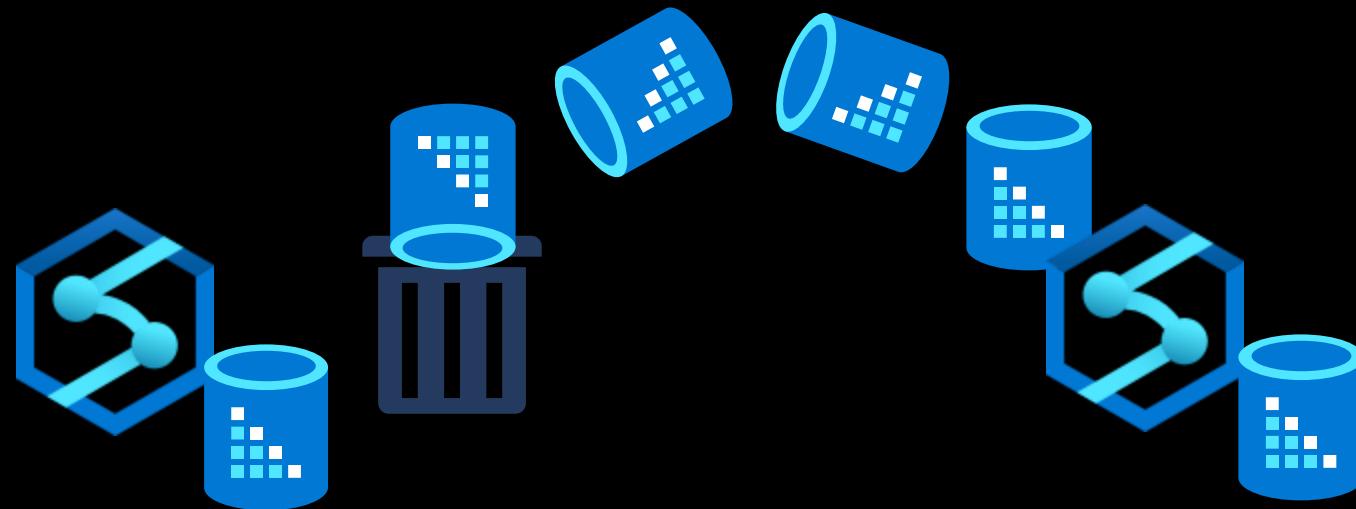
## Public Preview

Q2 2022

# Recover data warehouse from dropped server

---

No longer need support ticket to recover after an accidental server or workspace drop



`Restore-AzSynapseSqlPool -FromDroppedSqlPool`

## Public Preview

Q2 2022

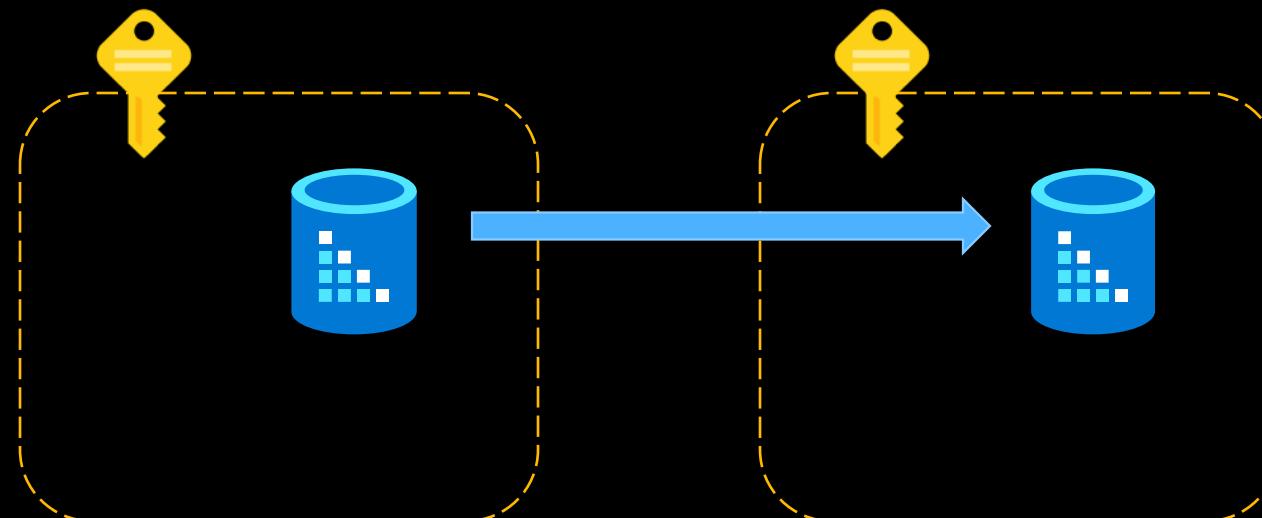
# Cross subscription restore

---

Restore data warehouse  
across subscription boundary  
in one command without  
workarounds

Unlock Dev/Test scenarios and  
simplify billing at the  
subscription level

PowerShell and API support  
available



## Public Preview

Q2 2022

# Distribution Advisor

Recommendation system that analyses the chosen queries or past run query data to provide suggestions on data distribution which can improve performance

ProductSales

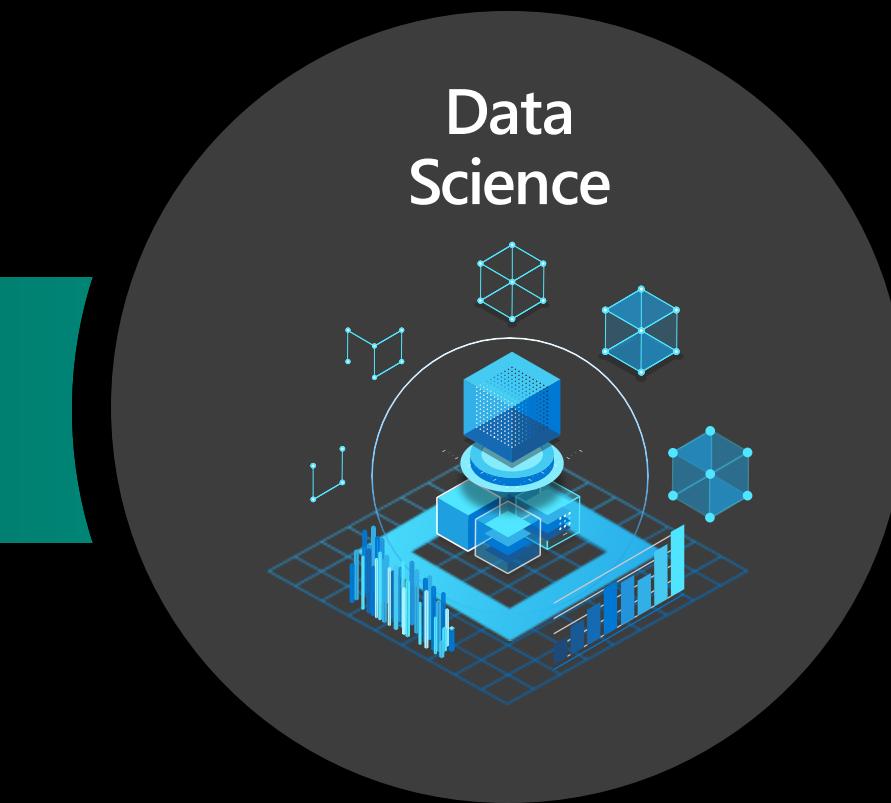
AccountID	SalesAmt	...
47	\$1,234.36	...
36	\$2,345.47	...
14	\$3,456.58	...
25	\$4,567.69	...
48	\$5,678.70	...
37	\$6,789.81	...
...	...	...

```
CREATE TABLE ProductSales
WITH (DISTRIBUTION=HASH(AccountID))
AS ...
```





Data Science



**Industry standard languages such as PySpark**

---

**Rich ecosystem of ML tools on Spark**

---

**Code-first and Code-free Auto ML**

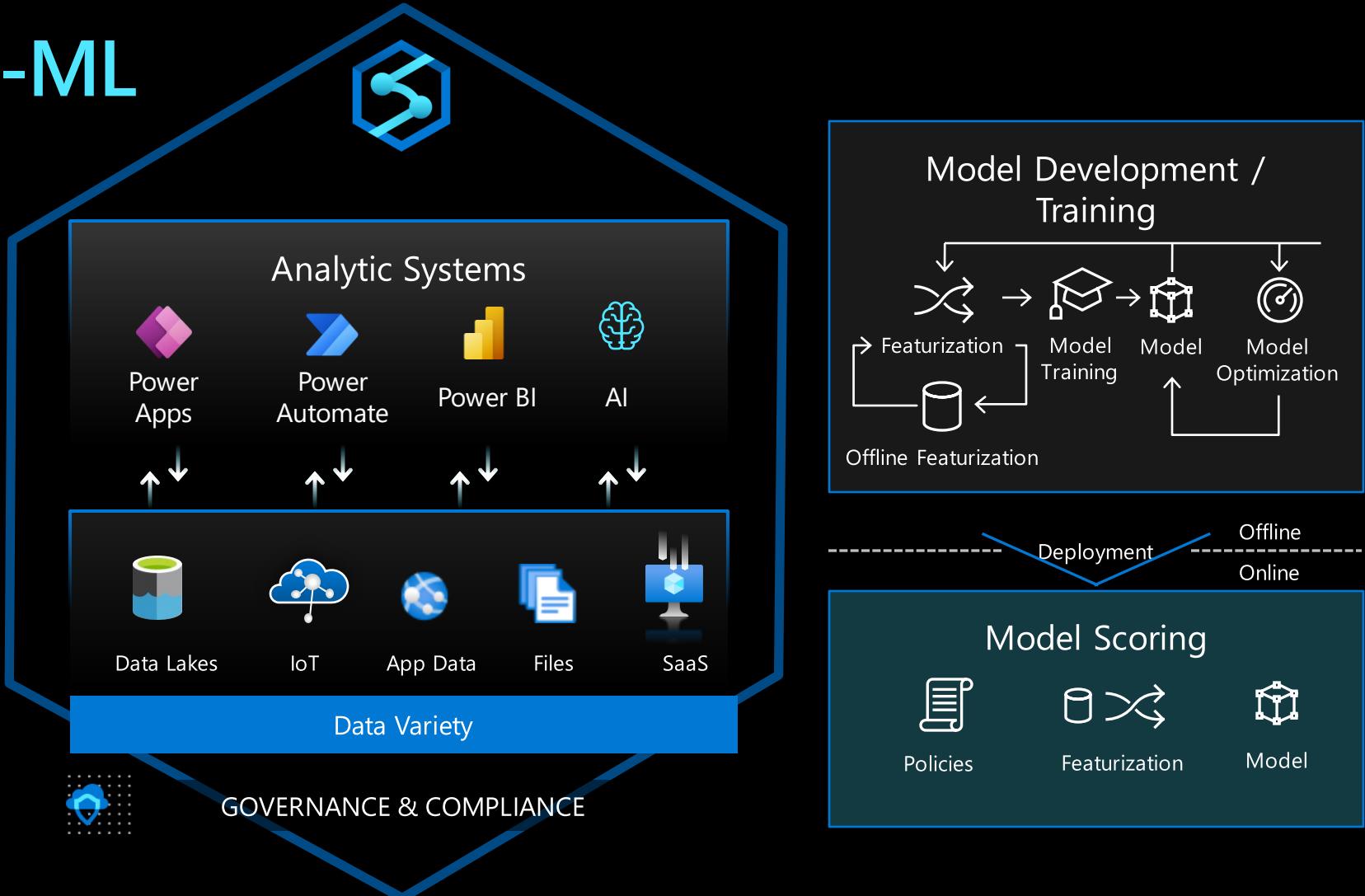
---

**Hardware accelerated execution with GPUs**

# Why Synapse Enterprise Grade-ML

## Productizing AI

- Train in the cloud within the Hub
- Scoring with operational systems
- Governance everywhere (models, lineage)
- Ethical AI
- Control over deployment
- Deployment across Apps, BI, Processes
- Exchange of models (ONNX)
- Enabling Reinforcement learning



## Generally Available

# Notebook Development Experience

Empower data scientists will a familiar Notebook based development interface

The screenshot shows the Microsoft Azure Synapse Analytics interface. The left sidebar displays a tree view of notebooks, categorized into 'Develop', 'SQL scripts', 'Notebooks', 'MS Conf notebooks', and 'Test notebooks'. The 'Notebooks' section contains several demo notebooks, with '020 Surface Sales Forecasting with Synapse' currently selected. The main workspace shows a Python script titled '020 Surface Sales Fo...'. The code is as follows:

```
1 from pandas.tseries.frequencies import to_offset
2 from azureml.core._vendor.azureml.client.core.common import metrics
3 from matplotlib import pyplot as plt
4 from automl.client.core.common import constants
5
6 def align_outputs(y_predicted, X_trans, X_test, y_test, target_column_name,
7     predicted_column_name='predicted',
8     horizon_colname='horizon_origin'):
9
10    if (horizon_colname in X_trans):
11        df_fcst = pd.DataFrame({predicted_column_name: y_predicted,
12                               horizon_colname: X_trans[horizon_colname]})  

13    else:
14        df_fcst = pd.DataFrame({predicted_column_name: y_predicted})
15
16    # y and X outputs are aligned by forecast() function contract
17    df_fcst.index = X_trans.index
18
19    # align original X_test to y_test
20    X_test_full = X_test.copy()
21    X_test_full[target_column_name] = y_test
22
23    # X_test_full's index does not include origin, so reset for merge
24    df_fcst.reset_index(inplace=True)
25    X_test_full = X_test_full.reset_index().drop(columns='index')
26    together = df_fcst.merge(X_test_full, how='right')
27
28    # drop rows where prediction or actuals are nan
29    clean = together[[target_column_name,
30                      predicted_column_name]].notnull().all(axis=1)
31
32    return(clean)
33
34 X_test[time_column_name] = pd.to_datetime(X_test[time_column_name])
35 df_all = align_outputs(y_predictions, X_trans, X_test, y_test, target_column_name)
36
# use automl metrics module
```

## Generally Available

# Built-in Cognitive Services

Enables simple integration of pre-built machine learning models.

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, the 'Data' section lists 'Lake database', 'retaildata' (selected), 'surfacesalesdb', and 'SQL database'. Inside 'retaildata', there are several tables: 'myparquettable', 'myparquettable2', 'myparquettable3', 'myparquettable5', 'myparquettable6', and 'retailsales' (selected). In the center, a code editor displays Python code for interacting with an Azure ML workspace:

```
1 import azureml.core
2 import pandas as pd
3 import numpy as np
4 import logging
5 from azureml.core.workspace import Workspace
6 from azureml.core import Experiment
7 from azureml.core.experiment import Experiment
8 from azureml.train.autoencoder import AutoEncoder
9 import os
10 subscription_id = os.getenv('AZUREML_SUBSCRIPTION_ID')
11 resource_group = os.getenv('AZUREML_RESOURCE_GROUP')
12 workspace_name = os.getenv('AZUREML_WORKSPACE_NAME')
13 workspace_region = os.getenv('AZUREML_WORKSPACE_REGION')
14
15 ws = Workspace(subscription_id, resource_group, workspace_name, workspace_region)
16 ws.write_config()
17
18 experiment_name = 'autoencoder'
19 experiment = Experiment(ws, experiment_name)
20 output = {}
21 output['Subscription ID'] = subscription_id
22 output['Workspace'] = ws
23 output['SKU'] = ws.sku
24 output['Resource Group'] = resource_group
25 output['Location'] = workspace_region
26 output['Run History Name'] = experiment_name
27 pd.set_option('display.max_rows', 10)
28 outputDf = pd.DataFrame([output])
```

To the right, there are sections for 'Predict with a model' (selected) and 'Choose a pre-trained model'. Under 'Predict with a model', it says 'retailsales'. Below that, under 'Choose a pre-trained model', is a section for 'Azure Cognitive Services' with two options: 'Anomaly Detector' and 'Sentiment Analysis'. Both options have descriptions and 'Learn more' links. At the bottom right are 'Continue' and 'Cancel' buttons.

## Public Preview

# Automated Machine Learning

No-code training for ML models empowers everyone with data science

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a sidebar with icons for Home, Databases, Tables, Views, and SQL databases. The main area shows a 'Data' workspace with a 'rawdata' folder containing several parquet files (myparquettable1 through myparquettable6) and a 'retailsales' folder. A code editor window displays Python code for initializing an AzureML workspace and experiment, and outputting results. To the right, the 'Train a new model' wizard is open, showing the 'retailsales' dataset selected. It asks to choose a model type, listing 'Classification', 'Regression', and 'Time series forecasting'. Each option has a brief description and an example. At the bottom right of the wizard are 'Continue' and 'Cancel' buttons.

```
1 import azureml.core
2 import pandas as pd
3 import numpy as np
4 import logging
5 from azureml.core.workspace import Workspace
6 from azureml.core import Experiment
7 from azureml.core.experiment import Experiment
8 from azureml.train.automl import AutoMLConfig
9 import os
10 subscription_id = os.getenv('AZUREML_SUBSCRIPTION_ID')
11 resource_group = os.getenv('AZUREML_RESOURCE_GROUP')
12 workspace_name = os.getenv('AZUREML_WORKSPACE_NAME')
13 workspace_region = os.getenv('AZUREML_WORKSPACE_REGION')
14
15 ws = Workspace(subscription_id=subscription_id,
16                 resource_group=resource_group,
17                 workspace_name=workspace_name,
18                 workspace_region=workspace_region)
19
20 experiment_name = 'auto'
21 experiment = Experiment(ws, experiment_name)
22 output = {}
23 output['Subscription ID'] = ws.subscription_id
24 output['Workspace'] = ws.workspace_name
25 output['Resource Group'] = ws.resource_group
26 output['Location'] = ws.location
27 pd.set_option('display.max_rows', 10)
28 outputDf = pd.DataFrame([output])
```

## Generally Available

# Industry Standard Open Ecosystem

Open file formats enable  
easy integration with other  
data services

Industry standard  
languages make it easy for  
developers to get started

```
1 import azureml.core
2
3 from azureml.core import Experiment, Workspace, Dataset, Datastore
4 from azureml.train.automl import AutoMLConfig
5 from azureml.data.dataset_factory import TabularDatasetFactory
6
7 subscription_id = "58f8824d-32b0-4825-9825-02fa6a801546"
8 resource_group = "plangadr"
9 workspace_name = "anlwsdemos"
10 experiment_name = "wsazuresynapseanalytics-retailsales-20210216065932"
11
12 ws = Workspace(subscription_id = subscription_id, resource_group = resource_group, workspace_name = workspace_name)
13 experiment = Experiment(ws, experiment_name)
14
15 df = spark.sql("SELECT * FROM retaildata.retailsales")
16
17 datastore = Datastore.get_default(ws)
18 dataset = TabularDatasetFactory.register_spark_dataframe(df, datastore, name = experiment_name + "-dataset")
19
20 automl_config = AutoMLConfig(spark_context = sc,
21                             task = "regression")
```

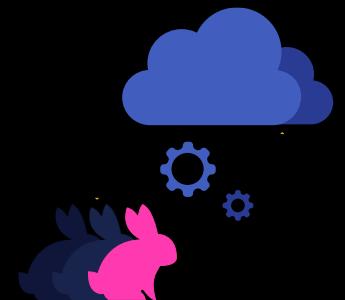
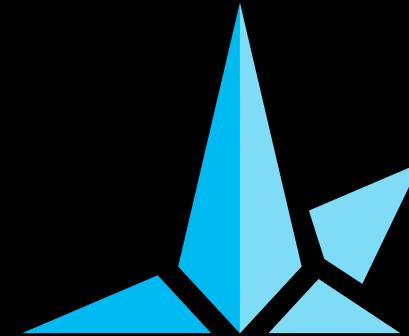
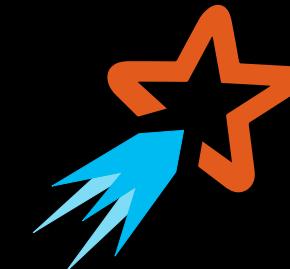
## Generally Available

### Synapse ML Library GA

Open-Source Spark Library

Distributed ML algorithms for Apache Spark, e.g.:

- Linear models & reinforcement learning: Vowpal Wabbit
- Gradient boosted trees: LightGBM
- Novelty detection: Isolation Forests



<https://aka.ms/spark>

Cognitive Services Integration for Spark

Language Binding Generators: Python, R, Java (, .NET)

## Public Preview

# Simplify model scoring at scale on Spark

Drastically simplifies handover of models from producer to consumer for “in-engine” scoring. Use machine Learning models from Azure ML in Synapse without moving any data

```
Prediction = spark.sql("""  
    SELECT  
        PREDICT('{MODEL_ID}', *) AS preds  
    FROM data  
""")
```

Synapse Spark

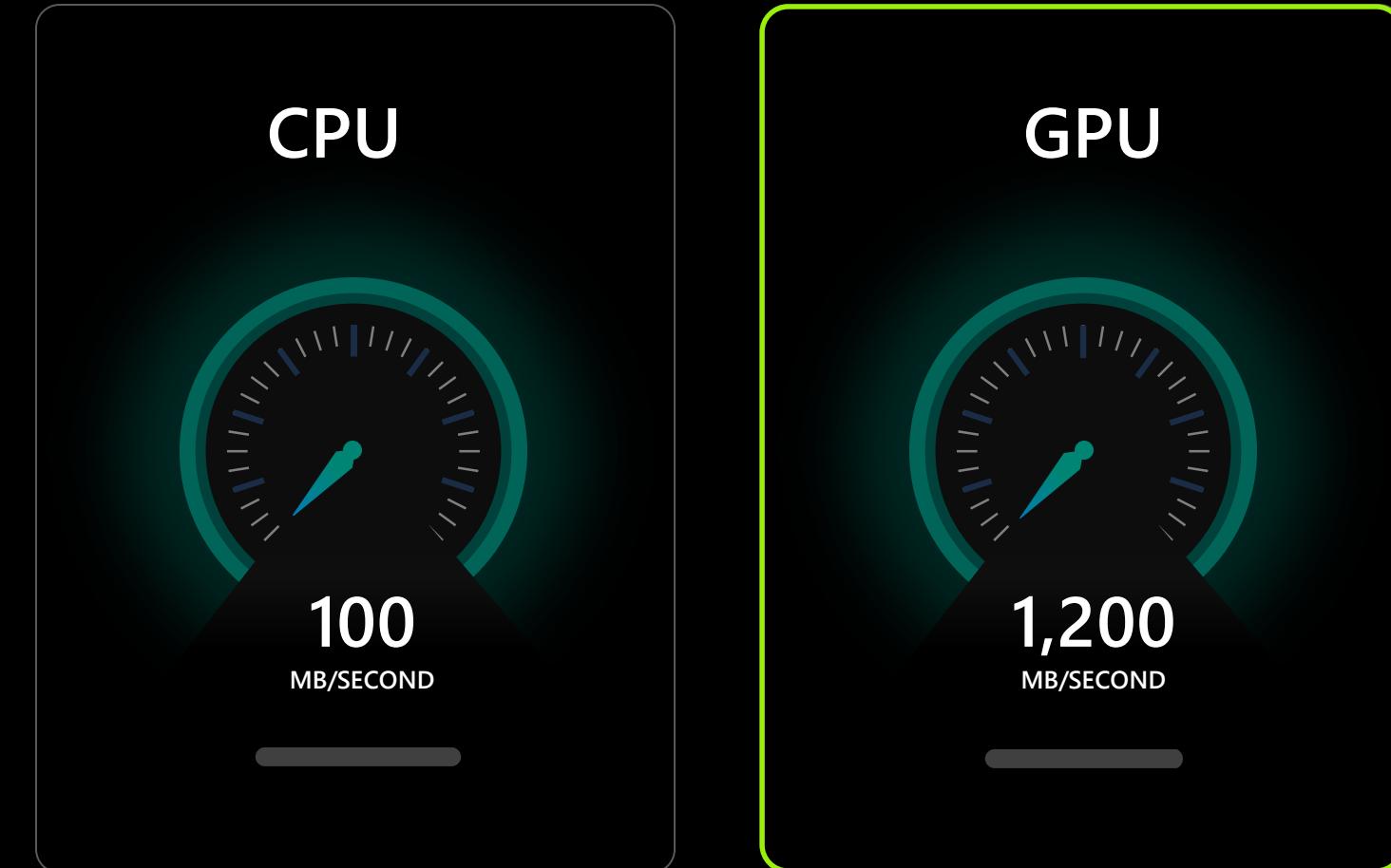


mlflow

## Public Preview

# GPU Accelerated Workloads

Accelerates data transformation and reduces ML model training time by dramatically increasing throughput vs. traditional CPU



Private Preview

Q2 2022



# R Language Support

Enables data scientists to apply the industry standard R language to developing ML models

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a sidebar with various notebooks and scripts listed under categories like 'Develop', 'SQL scripts', and 'Notebooks'. The main area displays two R code snippets in a notebook:

```
1 printSchema(df)
2
3 ✓ - Command executed in 160 ms on 4/03/18 PM, 11/03/21
```

```
1 root
2   |-- _id: string (nullable = true)
3   |-- Date: string (nullable = true)
4   |-- AveragePrice: string (nullable = true)
5   |-- TotalVolume: string (nullable = true)
6   |-- ABSD: string (nullable = true)
7   |-- 4229: string (nullable = true)
8   |-- 4230: string (nullable = true)
9   |-- Total_Bags: string (nullable = true)
10  |-- Small_Bags: string (nullable = true)
11  |-- Large_Bags: string (nullable = true)
12  |-- Merge_Bags: string (nullable = true)
13  |-- type: string (nullable = true)
14  |-- year: string (nullable = true)
15  |-- region: string (nullable = true)
```

```
1 head(select(df, #\$year))
2
3 ✓ - Command executed in 1 sec 40 ms on 4/03/24 PM, 11/03/21
```

Below the code, the output shows the 'year' column values: 1. 2015, 2. 2015, 3. 2015.





Observational Analytics

## Observational Analytics



What is it?

Semi-structured: text, json, time series

Machine generated or machine recorded human interactions

Mass volume

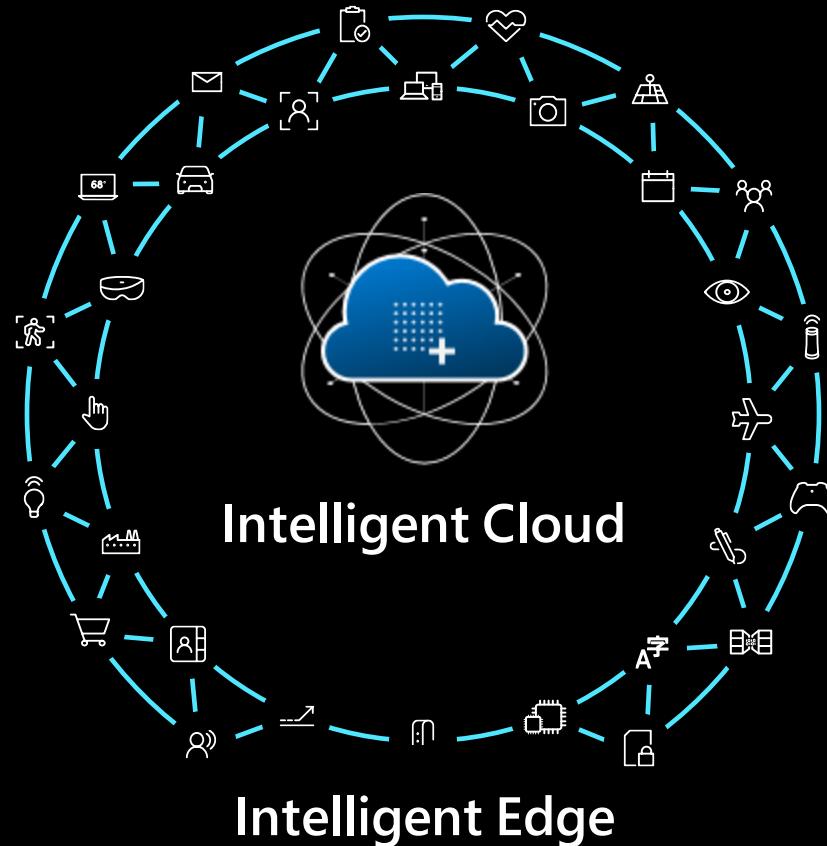
High velocity

Few large fact tablesstreams

# Observational Data

**50 BN**  
connected devices  
by 2030

**175 ZB**  
total amount of  
data by 2025



## Observational Data



Why is it challenging to analyze?

Looking for unpredictable phenomena

Constantly changing schema

Near real time visibility required

Analytics systems costs are often prohibitive

Frequently changing business questions

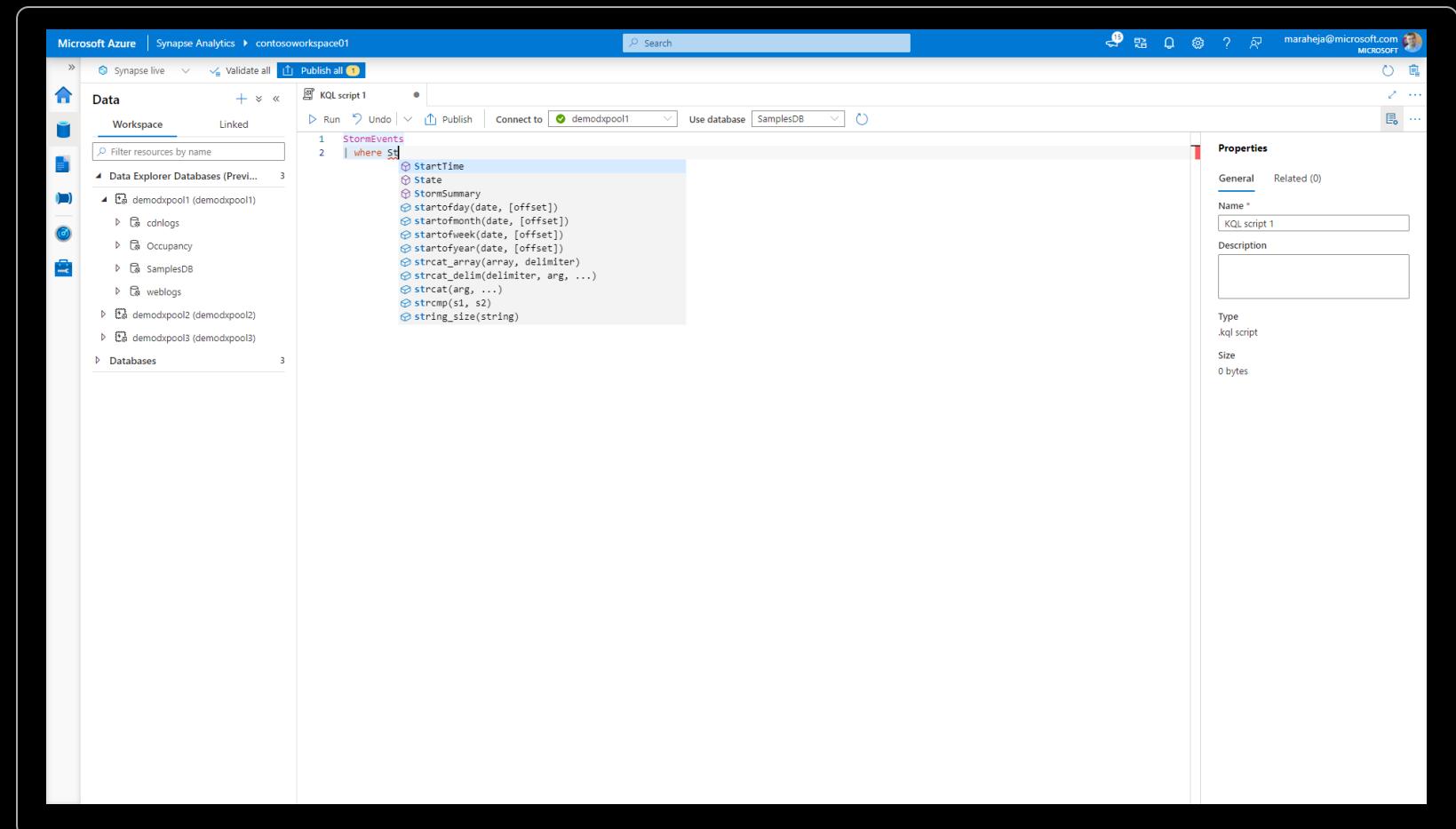
Near real-time analytics on  
Observational data at petabyte scale

## Public Preview

November 2021

# Synapse Data Explorer Engine

Industry leading free-text and semi-structured data indexing for sub second observational analytics

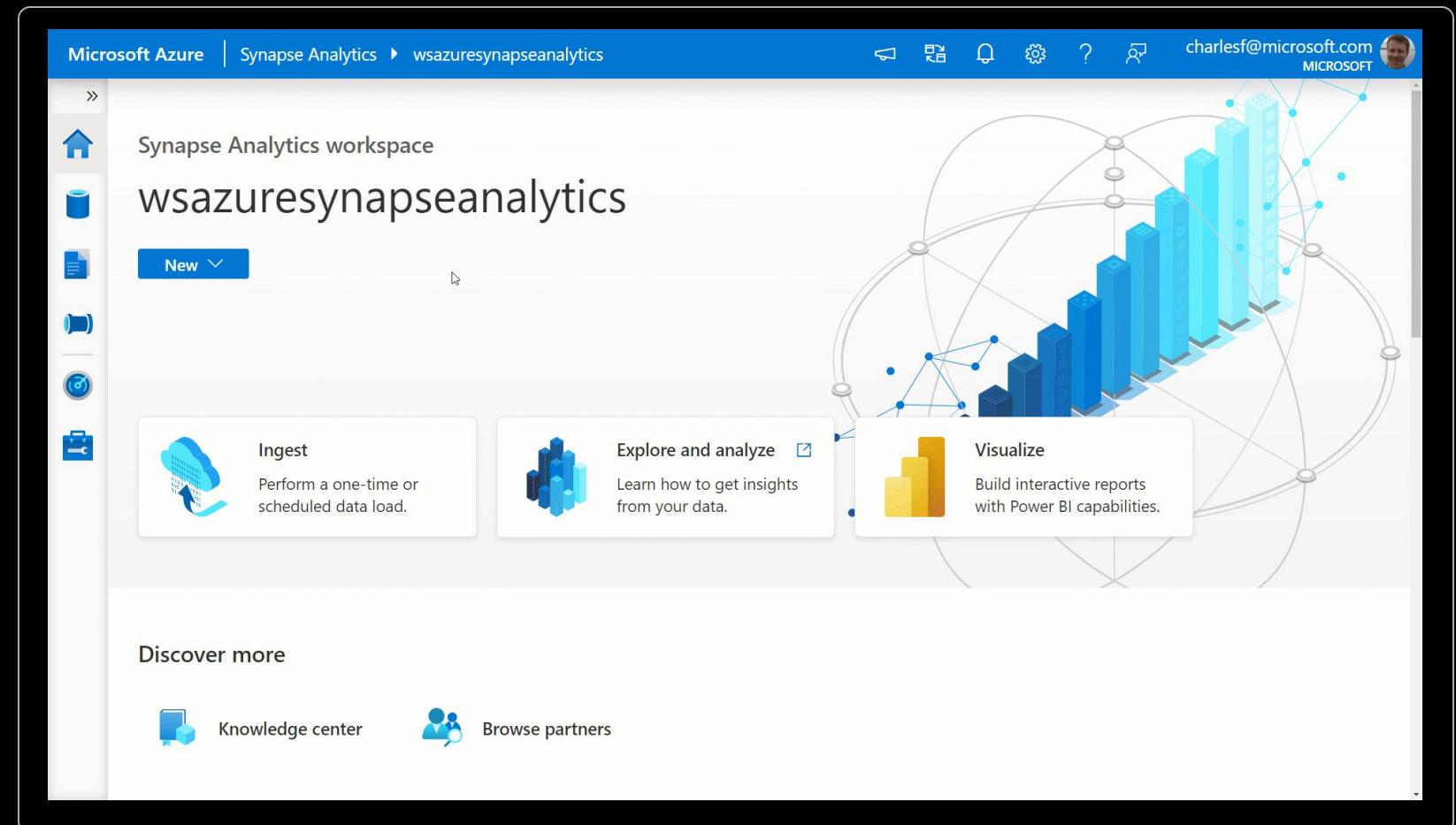


## Public Preview

Q4 2021

# Auto-Pause and Private Link for Data Explorer Clusters

Reduce costs by enabling clusters to automatically pause based on pre-defined timeout



Microsoft Azure | Synapse Analytics > wsazuresynapseanalytics

Synapse Analytics workspace  
wsazuresynapseanalytics

New

Ingest

Explore and analyze

Visualize

Discover more

Knowledge center

Browse partners

## Public Preview

Q4 2021

# 100,000 databases in a cluster

---

Enable developers to build large scale multi-tenant solutions with cluster compute reuse across workloads

Azure Data Explorer Cluster





Modélisation des objets logiques et "Lake Database"

# Approche moderne de données - Lakehouse

- Nouvelle approche d'objets métiers
- Exploiter les bénéfices des solutions « Cloud Native » comme serverless, autoscale, on-demand....
- Notion forte de croisement entre le Datawarehouse, le Datalake et les modèles d'analyse
- **La Datalake est vu comme une variable d'entrée et non une source de donnée**

## La finalité de l'approche

Exposer des objets métiers dynamiques

Exposer des objets métiers logiques

La matérialisation des objets n'est plus nécessaire

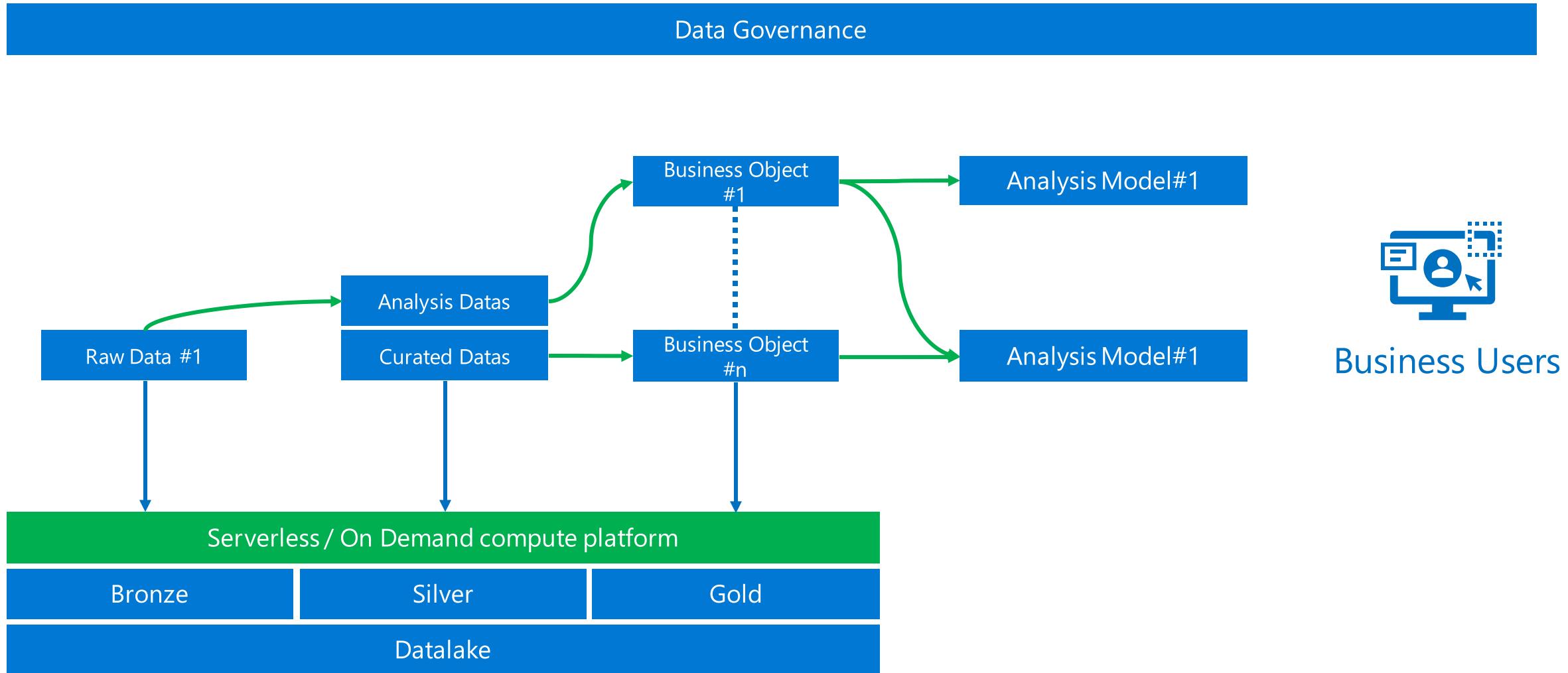
S'affranchir des notions de scalabilité et performance

S'affranchir de la consommation ou non de l'objet exposé

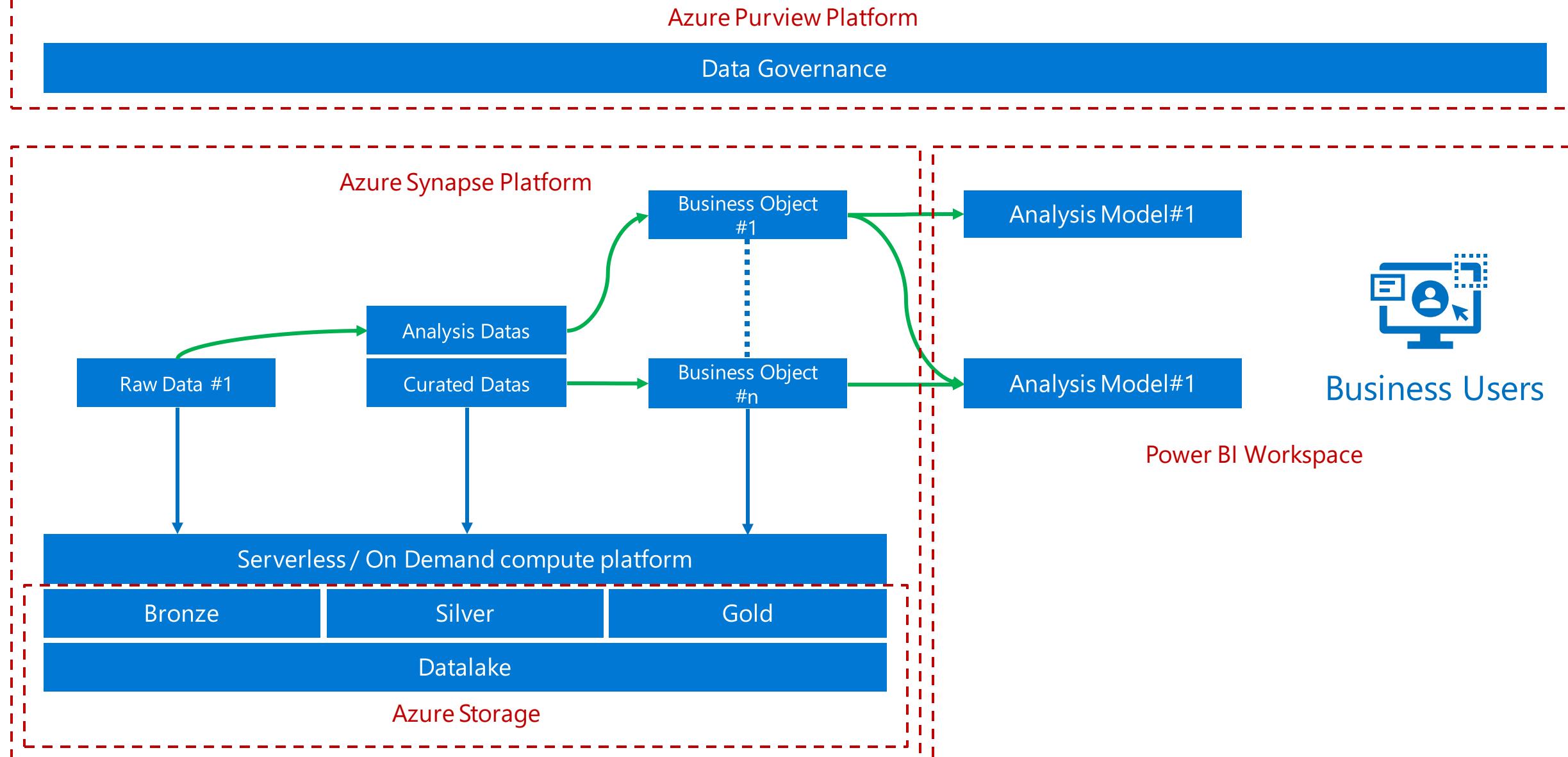
# Explications et retour d'expérience

- Besoin d'expliquer finement l'approche analytique moderne
- Besoin d'expliquer les nouvelles avancées du cloud
  - Détachement de la scalabilité des services, plus de "capacity planning"
  - Détachement de la notion de performance
  - Consommation "on demand" de la BI moderne
  - Les limites d'une telle approche
- **Réflexion sur "où" est calculé l'indicateur, modèles hybrides/composite**
- Couts variables pour les métiers (dépendant de l'utilisation), besoin de changer l'approche budgétaire.

# Objets logiques



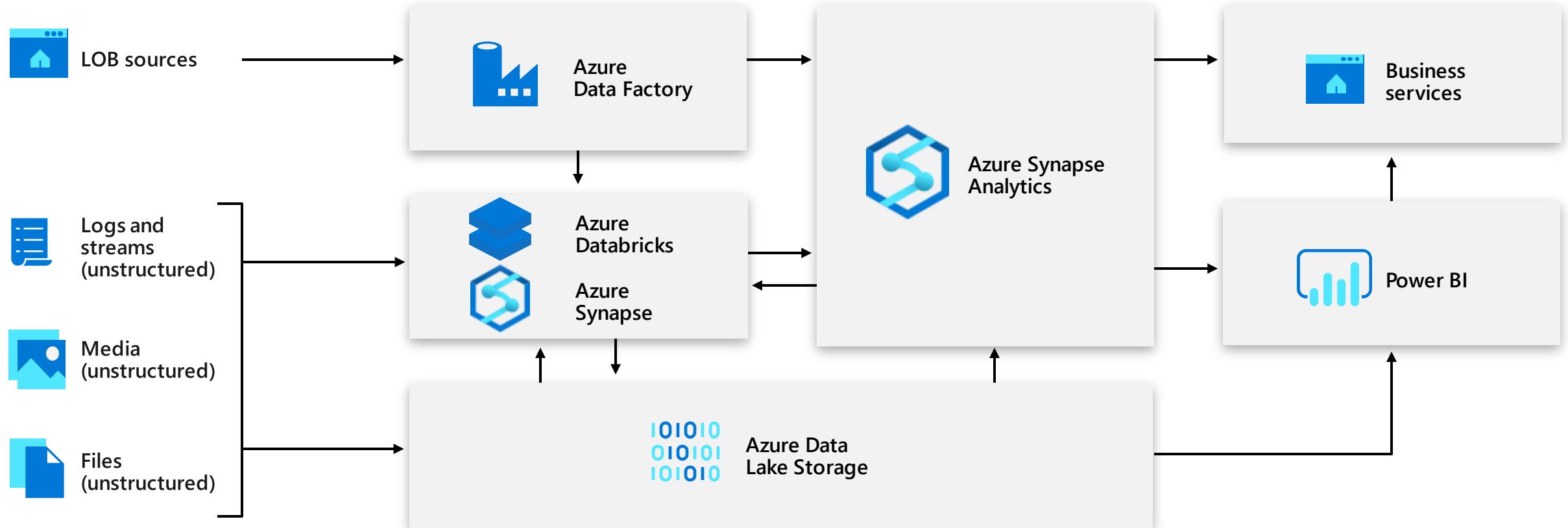
# Objets logiques



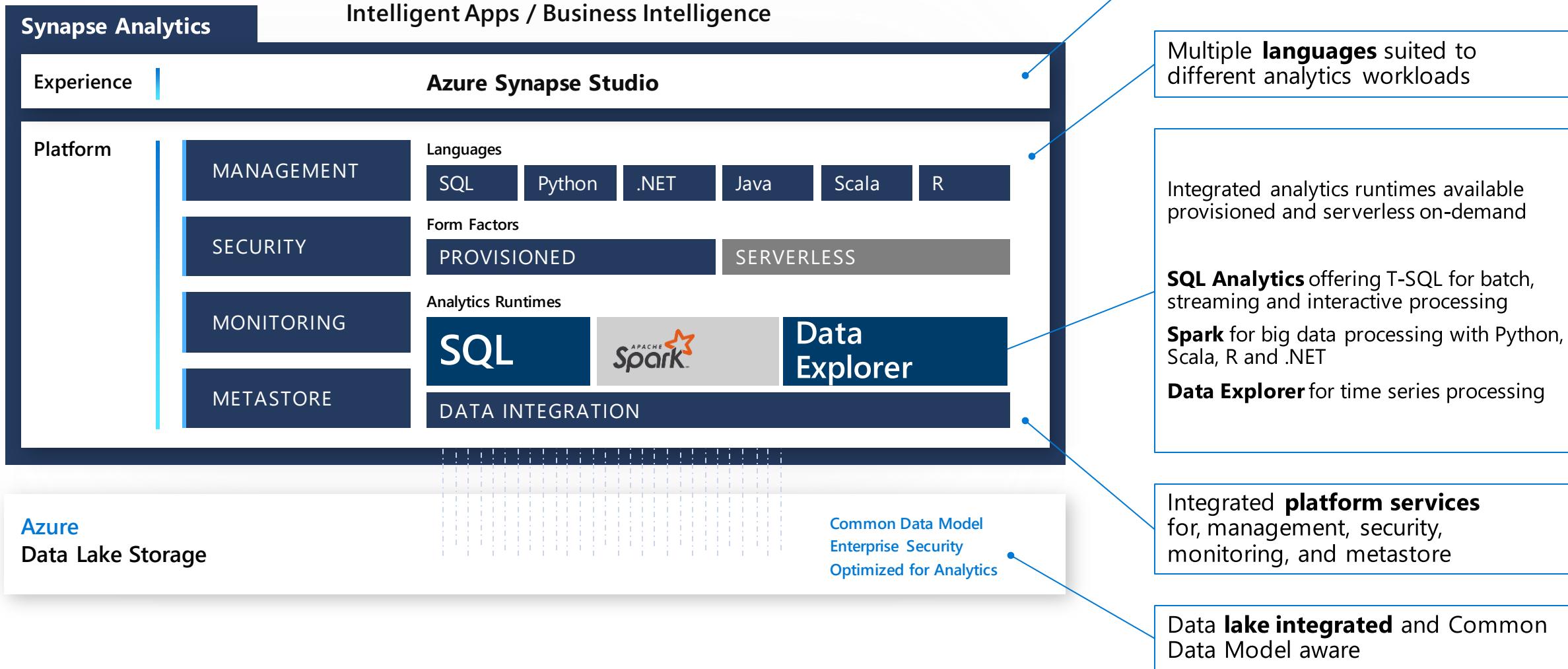
# Les zones de l'approche "Datalake(house)"

- Le "Datalake" se décompose en trois zones distinctes
- Ces zones sont réparties sur trois comptes de stockage Azure
- Zone **Bronze**
  - Cette zone correspond au dépôt des données brutes issues des systèmes amonts
  - Cette zone n'a pas de gouvernance spécifique
- Zone **Silver**
  - Cette zone est une zone de données partiellement nettoyées et analysables
  - L'accès est réservé aux "Data Engineer et Data Scientist" dans des approches de recherche et analyse
  - La gouvernance est minimaliste mais existe
- Zone **Gold**
  - Cette zone est la zone d'exposition des informations
  - Toutes données présentent dans celle-ci est une donnée vérifiée et exploitable dans l'entreprise (notion de trust)
  - Ces données sont exposées dans un outil de gouvernance d'accès dans l'entreprise
  - Gouvernance forte de la structure du datalake
  - Tout objet consommé dans l'entreprise provient de cette zone

# L'approche moderne d'exposition de données



# Moteurs de la plateforme Azure Synapse



# Performance



Elastic Architecture



Columnar Storage



Columnar Ordering



Table Partitioning



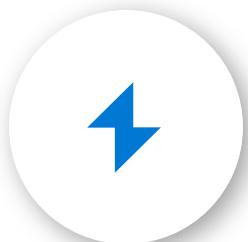
Secondary Indexes



Hash Distribution



Materialized Views



Resultset Cache

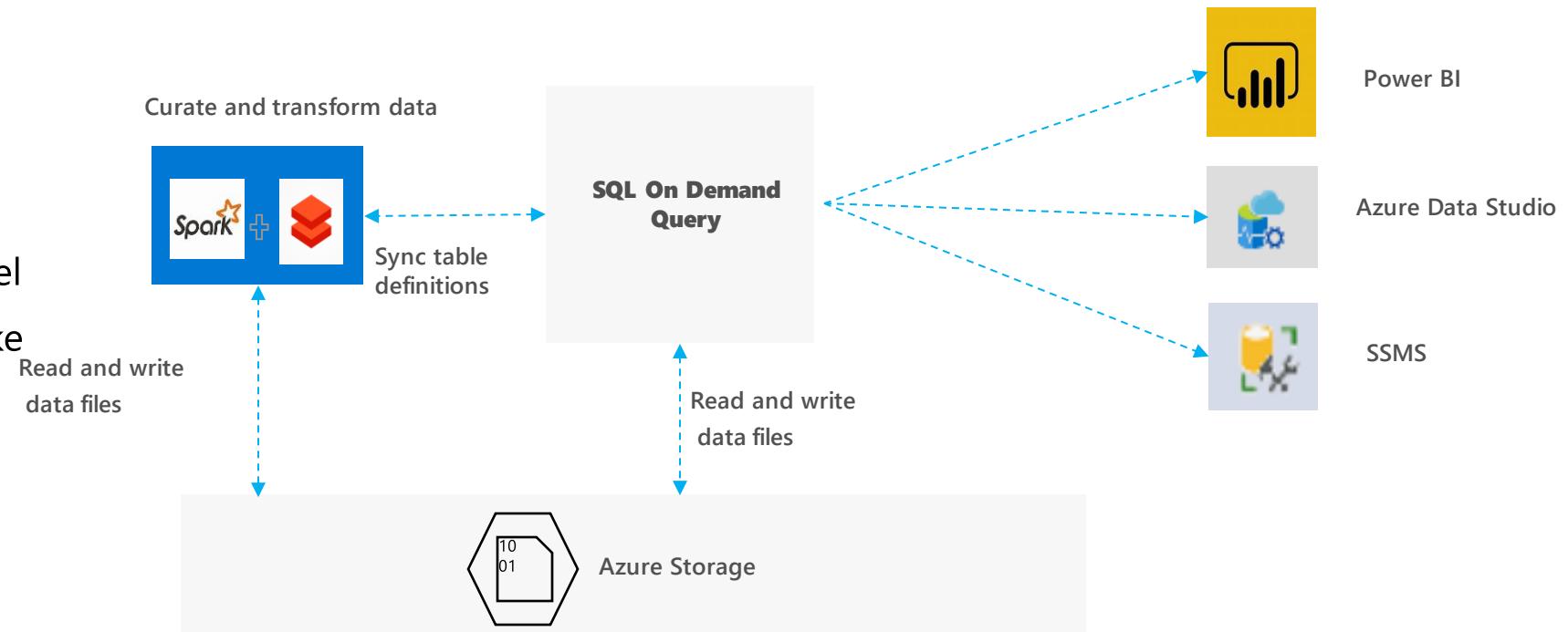
# Transform with Synapse SQL Serverless / Pricing

## Overview

An interactive query service that provides T-SQL queries over high scale data in Azure Storage.

## Benefits

- Pay-per-query with serverless model
- Query data in-place on the data lake with T-SQL (no ETL)
- Supports various file formats (Parquet, CSV, JSON)
- Integrates with Databricks, HDInsight, PowerBI, and the shared Synapse metastore



# Synapse SQL Serverless – Querying on storage

The screenshot illustrates the Microsoft Azure Synapse Analytics interface, specifically focusing on querying data stored in Azure Storage.

**Left Panel (Storage Explorer):**

- Shows a hierarchical view of storage accounts and containers:
  - Storage accounts:** prlangaddemo (Primary) containing filesystem, holidaydatacontainer, isdweatherdatacontainer, nyctlc, prlangaddemo, tmpcontainer, and wwidporters.
  - Databases:** prlangadSQLDW (SQL pool), default (SQL on-demand), default (Spark).
  - Datasets:** None.

**Middle Panel (File Explorer):**

- Shows a file tree under the nyctlc container:
  - puYear=2015/puMonth=3/part-00133-1e210938564719836543-aea5b543-5e83-4a7d-8d31-69f72c50b05d-15253-1.c000.snappy.parquet
- A context menu is open over the parquet file, with the "New SQL script" option highlighted.

**Right Panel (Query Editor):**

- The "SQL Analytics on-demand" dropdown is selected (highlighted with a red box).
- The query window contains the following T-SQL code:

```
1 SELECT
2     TOP 100 *
3 FROM
4     OPENROWSET(
5         BULK 'https://prlangaddemo.dfs.core.windows.net/nyctlc/yellow/puYear=2015/puMonth=3/part-00133-tid-210938564719836543-aea5b543-5e83-4a7d-8d31-69f72c50b05d-15253-1.c000.snappy.parquet'
6         FORMAT='PARQUET'
7     ) AS nyc;
```
- The results pane shows a table with several columns: VENDORID, TPEPICKUPDATETIME, TPEPDROPOFFDATETIME, PASSENGERCOUNT, TRIPDISTANCE, PULOCATIONID, DLOCATIONID, STARTLON, STARTLAT, and ENDLON.
- The status bar at the bottom indicates: "00:01:00 Query executed successfully."

# Synapse SQL Serverless – Querying CSV File

## Overview

Uses OPENROWSET function to access data

## Benefits

Ability to read CSV File with

- no header row, Windows style new line
- no header row, Unix-style new line
- header row, Unix-style new line
- header row, Unix-style new line, quoted
- header row, Unix-style new line, escape
- header row, Unix-style new line, tab-delimited
- without specifying all columns

```
SELECT *
FROM OPENROWSET(
    BULK 'https://XXX.blob.core.windows.net/csv/population/population.csv',
    FORMAT = 'CSV',
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n'
)
WITH (
    [country_code] VARCHAR (5) COLLATE Latin1_General_BIN2,
    [country_name] VARCHAR (100) COLLATE Latin1_General_BIN2,
    [year] smallint,
    [population] bigint
) AS [r]
WHERE
    country_name = 'Luxembourg'
    AND year = 2017
```

	country_code	country_name	year	population
1	LU	Luxembourg	2017	594130

# Synapse SQL Serverless – Querying specific files

## Overview

**filename** – Provides file name that originates row result

**filepath** – Provides full path when no parameter is passed or part of path when parameter is passed that originates result

## Benefits

Provides source name/path of file/folder for row result set

```
SELECT
    r.filepath() AS filepath,
    ,r.filepath(1) AS [year],
    ,r.filepath(2) AS [month],
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
    BULK 'https://XXX.blob.core.windows.net/csv/taxi/yellow_tripdata_2017-1*.csv',
    FORMAT = 'CSV',
    FIRSTROW = 2
)
WITH (
    vendor_id INT,
    pickup_datetime DATETIME2,
    dropoff_datetime DATETIME2,
    passenger_count SMALLINT,
    trip_distance FLOAT,
    <...columns>
) AS [r]
WHERE r.filepath(1) IN ('2017')
    AND r.filepath(2) IN ('10', '11', '12')
GROUP BY r.filepath(), r.filepath(1), r.filepath(2)
ORDER BY filepath
```

## Example of filename function

```
SELECT
    r.filename() AS [filename],
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
    BULK 'https://XXX.blob.core.windows.net/csv/taxi/yellow_tripdata_2017-1*.csv',
    FORMAT = 'CSV',
    FIRSTROW = 2
)
WITH (
    vendor_id INT,
    pickup_datetime DATETIME2,
    dropoff_datetime DATETIME2,
    passenger_count SMALLINT,
    trip_distance FLOAT,
    <...columns>
) AS [r]
```

GROUP BY r.filename()

ORDER BY [filename]

	filename	rows
1	yellow_tripdata_2017-10.csv	9768815
2	yellow_tripdata_2017-11.csv	9284803
3	yellow_tripdata_2017-12.csv	9508276

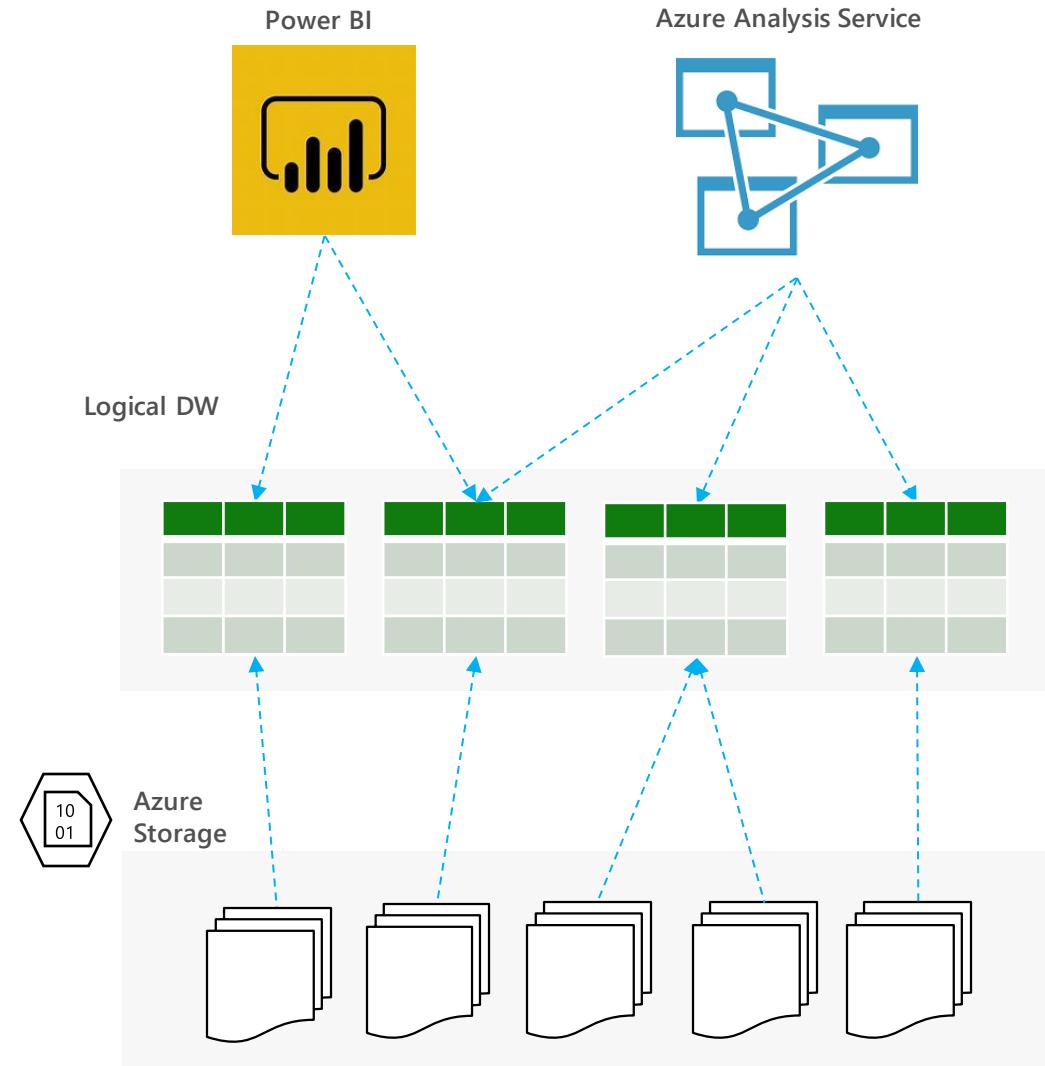
# Synapse serverless SQL pool as a logical data warehouse

## Overview

Logical relational layer on top of physical files in Azure Storage.

## Benefits

- Abstract physical storage and file formats using well understandable relational concepts such as tables and views.
- Direct connector to Azure storage for large ecosystem of BI tools
- BI tools that use SQL can work with files on storage
  - Analytic tools use external tables that represent proxy to actual files.
  - No need for custom connectors in BI tools.
- Provides complex data processing (joining and aggregation) on top of raw files.
- Apply enterprise-ready security model and access control using battle-tested SQL Server permission model on top of Azure storage files



# Logical Data Warehouse views

## Overview

serverless SQL pool logical data warehouse views are created on external files placed in customer Azure storage

## Benefits

Create SQL views on externally stored data

Access files using the view from various tools and language

Leverage rich T-SQL language to process and analyze data in external files exposed via views

Create PowerBI reports on the views created on external data

```
USE [mydbname]
GO

DROP VIEW IF EXISTS populationView
GO

CREATE VIEW populationView AS
SELECT *
FROM OPENROWSET(
    BULK 'https://XYZ.blob.core.windows.net/csv/population/\*.csv',
    FORMAT = 'CSV',
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n'
)
WITH (
    [country_code] VARCHAR (5),
    [country_name] VARCHAR (100),
    [year] smallint,
    [population] bigint
) AS [r]
```

```
SELECT
    country_name, population
FROM populationView
WHERE
    [year] = 2019
ORDER BY
    [population] DESC
```

	country_name	population
1	China	1389618778
2	India	1311559204
3	United States	331883986
4	Indonesia	264935824
5	Pakistan	210797836
6	Brazil	210301591
7	Nigeria	208679114
8	Bangladesh	161062905
9	Russia	141944641
10	Mexico	127318112

# Creating views

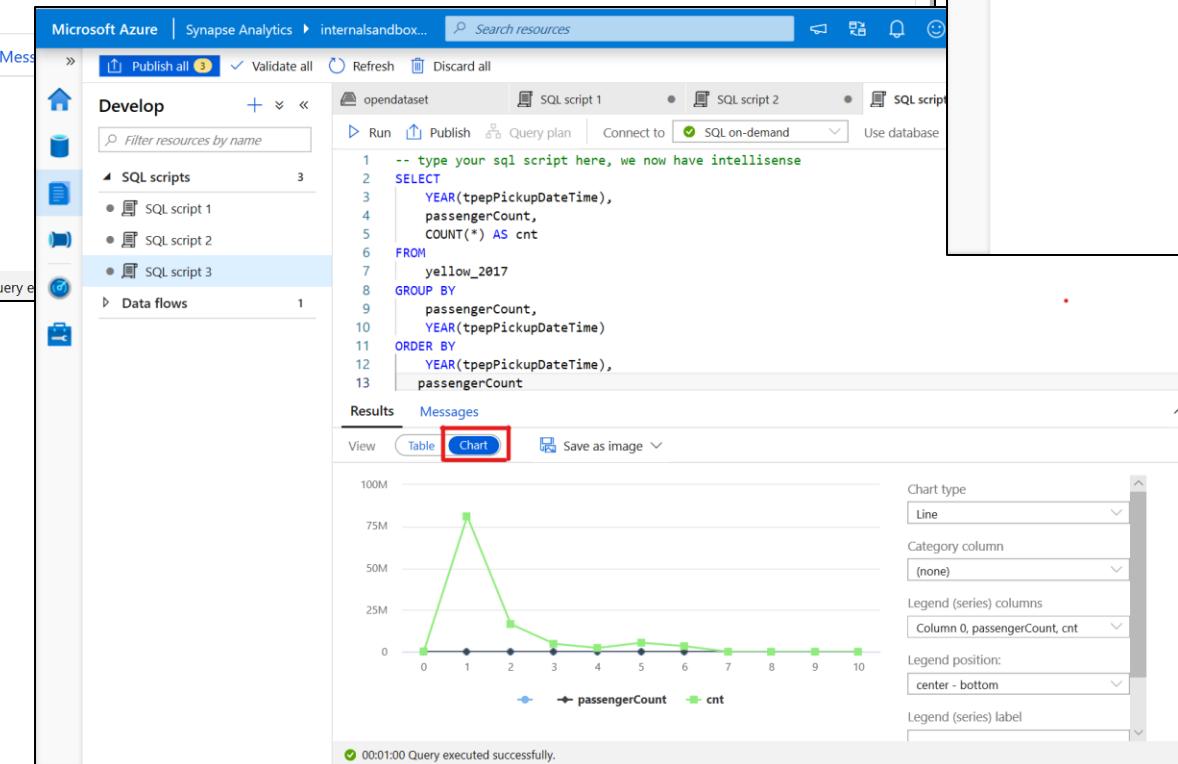
Microsoft Azure | Synapse Analytics > internalsandbox...

Data

opendataset

Connect to  Use database DefSQLOnDemand

```
1 CREATE VIEW yellow_2017 AS
2 Select *
3 FROM
4 OPENROWSET(
5      BULK 'https://internalsandboxwe.dfs.core.windows.net/opendataset/nyctlc/yellow/puYear=2017/\*/\*',
6      FORMAT='PARQUET'
7 ) AS [r];
```



Microsoft Azure | Synapse Analytics > internalsandbox...

Develop

opendataset

Connect to  Use database DefSQLOnDemand

```
1 -- type your sql script here, we now have intellisense
2 SELECT
3     YEAR(tpepPickupDateTime),
4     passengerCount,
5     COUNT(*) AS cnt
6 FROM
7     yellow_2017
8 GROUP BY
9     passengerCount,
10    YEAR(tpepPickupDateTime)
11 ORDER BY
12    YEAR(tpepPickupDateTime),
13    passengerCount
```

Results

View

00:01:00 Query executed successfully.

(NO COLUMN NAME)	PASSENGERCOUNT	CNT
2017	0	166086
2017	1	81034075
2017	2	16545571
2017	3	4748869
2017	4	2257813
2017	5	5407319

# Logical Data Warehouse - tables

## Overview

Create external tables that reference external files in your serverless SQL pool logical data warehouse

## Benefits

Create external tables that reference set of files on Azure storage.

Join and transform multiple tables in the same query.

Enables you to analyze external files with the same experience that you have in classic databases.

Manage column statistics in external tables.

Manage access rights per table.

Create PowerBI reports on the views created on external data

```
USE [mydbname]
GO

DROP TABLE IF EXISTS dbo.Population
GO

CREATE EXTERNAL TABLE dbo.Population (
    country_code VARCHAR(5) COLLATE Latin1_General_BIN2,
    country_name VARCHAR(100) COLLATE Latin1_General_BIN2,
    year smallint,
    population bigint
)
WITH(
    LOCATION = '/csv/population/population-*/*.csv',
    DATA_SOURCE = MyAzureStorage,
    FILE_FORMAT = MyAzureCSVFormat
)
```

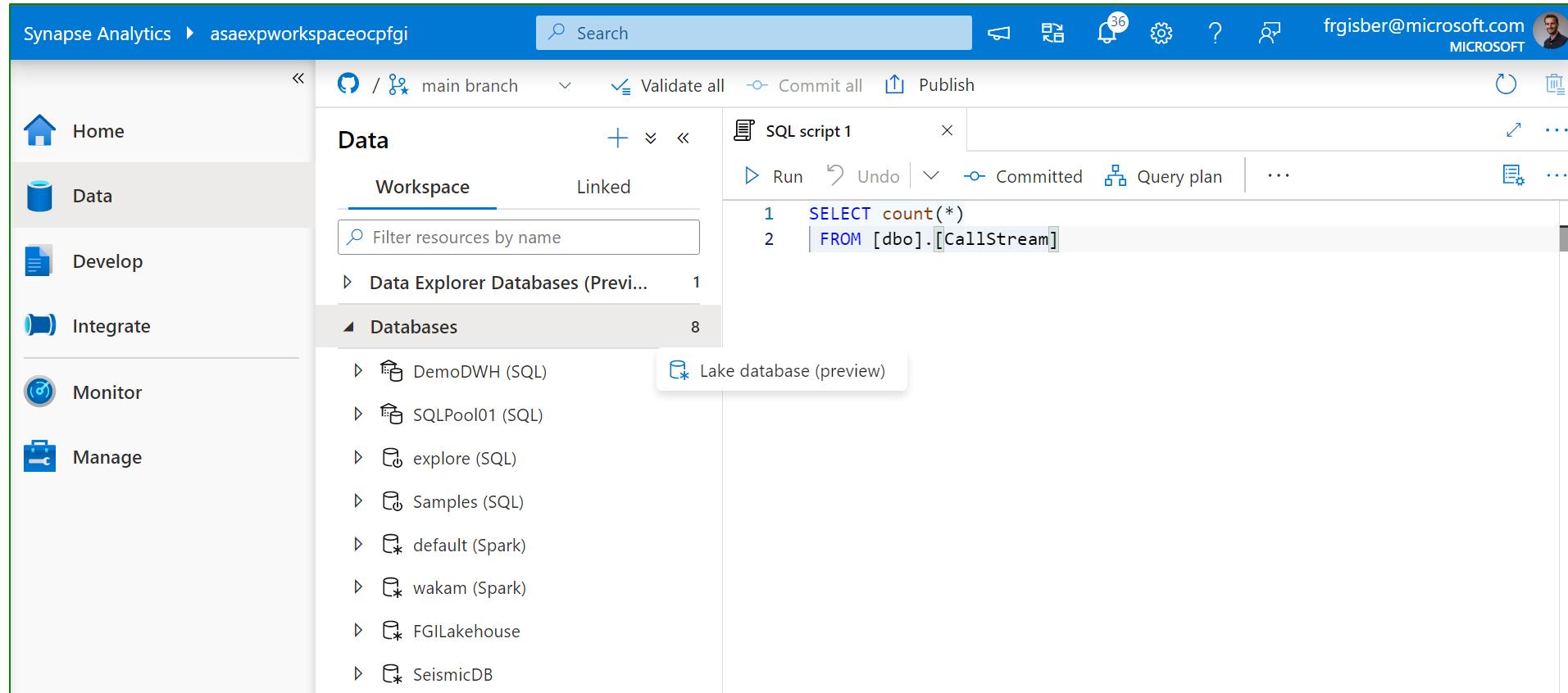
```
CREATE STATISTICS stat_country_name
ON dbo.Population(country_name);
```

```
SELECT
    country_name, population
FROM population
WHERE year = 2019
ORDER BY population DESC
```

	country_name	population
1	China	1389618778
2	India	1311559204
3	United States	331883986
4	Indonesia	264935824
5	Pakistan	210797836
6	Brazil	210301591
7	Nigeria	208679114
8	Bangladesh	161062905
9	Russia	141944641
10	Mexico	127318112

# Lakehouse Database Engine

- New Lakehouse Database capabilities based on Synapse Spark



# Data Lakehouse database

- Based on entities templates or custom design

The screenshot shows the 'Add from template' interface in Microsoft Synapse Analytics. The top navigation bar includes 'Synapse Analytics', a search bar, and user information 'frgisber@microsoft.com MICROSOFT'. The main area is titled 'Add from template' with a dropdown menu showing 'Database 1'. A 'Filter by keyword' input field is present. The interface displays a grid of 16 templates, each with an icon and a brief description:

Icon	Name	Description
Agriculture	Agriculture	For companies engaged in growing crops, raising livestock and dairy production.
Automotive	Automotive	For companies manufacturing automobiles, heavy vehicles, tires, and other automotive components.
Banking	Banking	For companies who are analyzing banking data.
Consumer Goods	Consumer Goods	For manufacturers or producers of goods bought and used by consumers.
Energy & Commodity Trading	Energy & Commodity Trading	For traders of energy, commodities, or carbon credits.
Freight & Logistics	Freight & Logistics	For companies providing freight and logistics services.
Fund Management	Fund Management	For companies managing investment funds on behalf of investors.
Genomics	Genomics	For companies acquiring and analyzing genomic data about human beings or other species.
Life Insurance & Annuities	Life Insurance & Annuities	For companies who provide life insurance, sell annuities, or both.
Manufacturing	Manufacturing	For companies engaged in discrete manufacturing of a wide range of products.
Oil & Gas	Oil & Gas	For companies involved in various phases of the Oil & Gas value chain.
Pharmaceuticals	Pharmaceuticals	For companies engaged in creating, manufacturing, and marketing pharmaceutical and bio-pharmaceutical products and medical devices.
Property & Casualty Insurance	Property & Casualty Insurance	For companies who provide insurance against risks to property and various forms of liability coverage.
R&D and Clinical Trials	R&D and Clinical Trials	For companies involved in research and development and clinical trials of pharmaceutical products or devices.
Retail	Retail	For sellers of consumer goods or services to customers through multiple channels.
Utilities	Utilities	For gas, electric and water utilities and power generators and water desalination.

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

Microsoft Azure | Synapse Analytics > asaexpworkspaceocfgi

main branch | Validate all | Commit all | Publish | Search

**Data** + << << / >> main branch

**Workspace** Linked

**Tables** << Filter resources by name

**SQL script 1** Database 1 SeismicDB

**Committed**

**Tables**

**SeismicShot**

- 123 SeismicShotId PK
- SeismicShotTimestamp
- SeismicShotEnergySourceA...
- SeismicShotLocationId
- GeographicAreaId
- GeographicAreaPolygonVer...
- SeismicShotSourceArrayTy...
- SeismicShotRecordingPolar...
- SeismicShotSeismicEnergyT...
- SeismicShotSeismicSensorT...

**SeismicData AcquisitionEvent**

- 123 SeismicDataAcquisitionEve... PK
- SeismicDataAcquisitionEve...
- SeismicDataAcquisitionEve...
- SeismicDataAcquisitionEve...
- SeismicDataAcquisitionEve...
- SurveyMethodTypeid
- NumberOfChannels

**ShotFile**

- 123 SeismicDataFileId PK
- 123 SeismicShotId PK,FK
- ShotFileNote

**ShotChannel**

- 123 SeismicShotId PK,FK
- ShotChannelId PK
- SeismicDataFileId PK,FK
- PeriodStartTimestamp PK
- PeriodEndTimestamp
- SeismicChannelTypeid FK
- ShotChannelNote

**General** **Columns** **Relationships**

**Columns**

Name	Keys	Description	Nullability	Data type	Format / Length
SeismicDataFileId	PK	The unique identifier of a seismic data file.	Null	integer	
SeismicShotId	PK, FK	The unique identifier of a seismic shot.	Null	integer	
ShotFileNote	PK	A note, comment or additional information regarding the shot file.	Null	string	1024

```

    graph TD
        SS[SeismicShot] --> SDAE[SeismicData AcquisitionEvent]
        SDAE --> SF[ShotFile]
        SF --> SC[ShotChannel]
    
```

# Spécificités des moteurs SQL

- Moteur SQL dédié
  - Puissance dédiée scalable horizontalement
  - Moteur MPP (Massivement parallèle)
  - 60 nœuds de stockages et n nœuds de calcul
  - Réponse linéaire des requêtes
  - Exposition d'objets
    - Logiques sous la forme de tables externes
    - Logiques sous la forme de vues (pouvant être matérialisées, donc objets physiques)
    - Physiques sous la forme de tables ou vues avec données embarquées
- Moteur SQL Serverless
  - Puissance "on demand"
  - Moteur non MPP
  - Scalabilité automatique et non linéaire (mais prédictible)
  - Exposition d'objets **logiques**

# Spécificités du moteurs Spark

- Le moteur Spark est disponible dans le service Azure Synapse
- Ce moteur possède deux "form factor"
  - Cluster Spark à disposition dans un format managé, "auto scalable" et activé à la demande
  - Objets Spark disponibles en mode "serverless"
- Possibilité de matérialiser les résultats des travaux de recherches, analyses, etc. dans des tables Spark
- Les tables Spark sont exposées en mode "Serverless"
- L'accès est réalisé via le point d'accès SQL Serverless **sans que le moteur Spark soit démarré**

# Power BI Embedded pour l'exposition et REST API

PowerBI REST APIs <https://docs.microsoft.com/en-us/rest/api/power-bi>

PowerBI Playground <https://microsoft.github.io/PowerBI-JavaScript/demo/v2-demo/index.html#>

## Getting started

[Set up your Power BI embedding environment](#)

[Power BI JavaScript API wiki](#)

[Power BI embedding documentation](#)

## Useful links

[Power BI Embedded on Azure](#)

[Power BI community](#)

[Power BI Ideas - APIs and embedding](#)

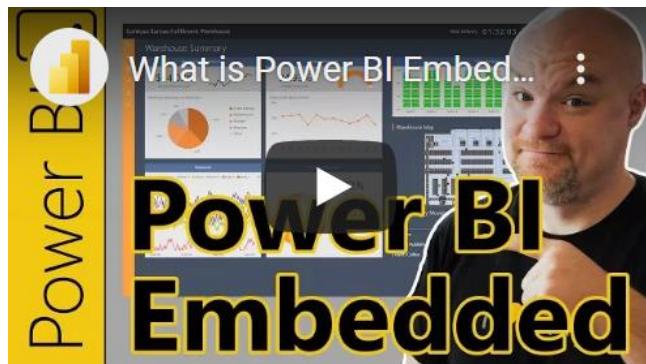
## Support

[Power BI Embedded FAQ](#)

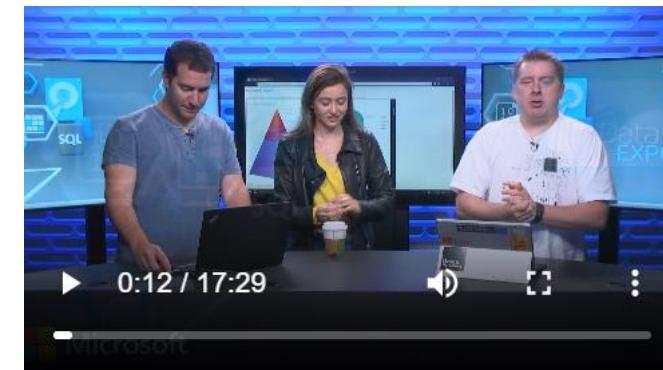
[Power BI Embedded troubleshooting](#)

[Power BI Support](#)

## What is Power BI Embedded



## Microsoft Power BI Embedded update



## Microsoft Power BI Embedded update

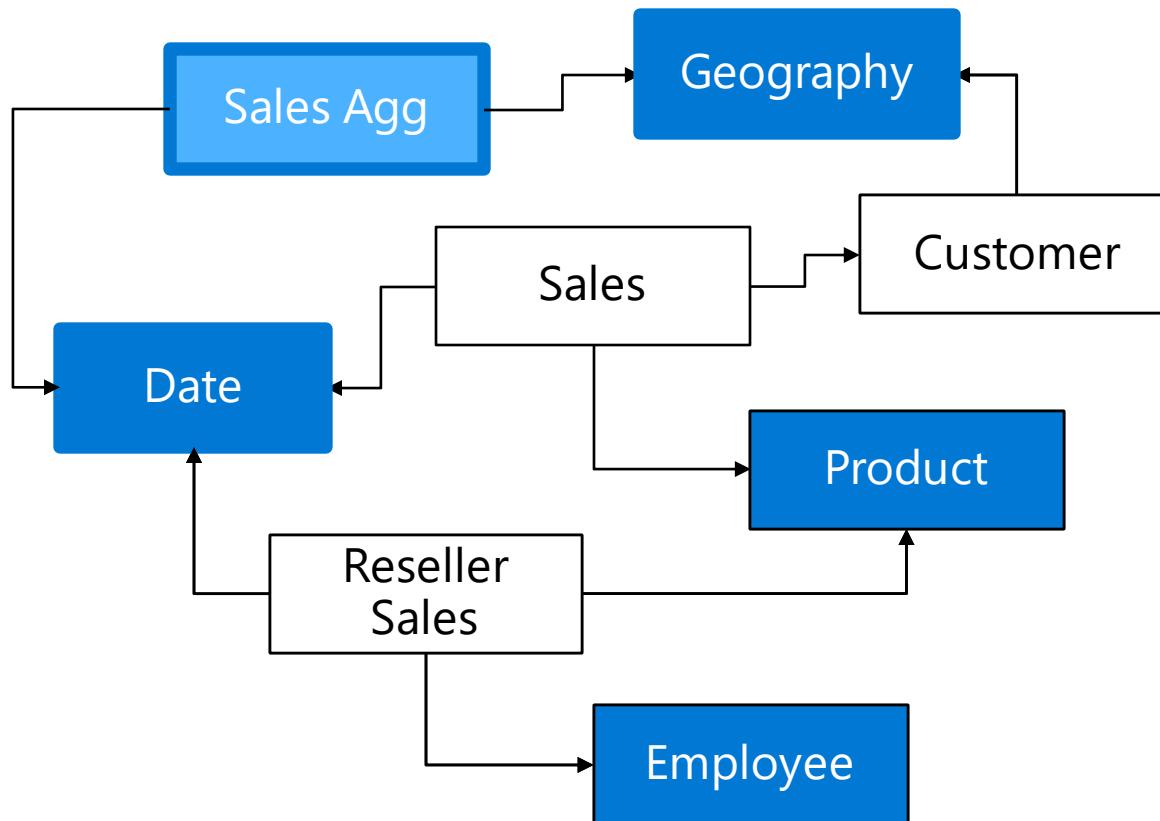


<https://playground.powerbi.com/>

# Rapports Dynamiques et Paginés

- Deux approches de rapports consommant les objets
- Rapport / Dashboard dynamiques PowerBI
  - Rapports très graphiques
  - Modèle de donnée embarqué ou délégué
  - Accès aux objets logiques dans une approche hybride (modèle composite)
  - Très adapté aux exploitations de résultats calculés, agrégés, etc.
  - Requêtes générées par l'outil
  - Utilisation de Power BI Desktop
- Rapport Paginés
  - Rapports graphiques, paramétrables
  - Adaptés aux approches de rapports sur les données de détails (listing, recherche d'information très fines)
  - Contrôle totale de la requête
  - Utilisation de Power BI Report Designer

# Import, Direct Query et Agrégats



```
SummarizeColumns(  
    Date[Year],  
    Geography[City],  
    "Sales", Sum(Sales[Amount])  
)
```

## DirectQuery

```
SELECT [Year],  
       [Name],  
       SUM([Amount]) AS [Amount]  
FROM   [Sales]  
INNER JOIN [Date] ON ...  
INNER JOIN [Customer] ON ...  
GROUP BY [Year],  
        [Name]
```





Business Intelligence

# Business Intelligence

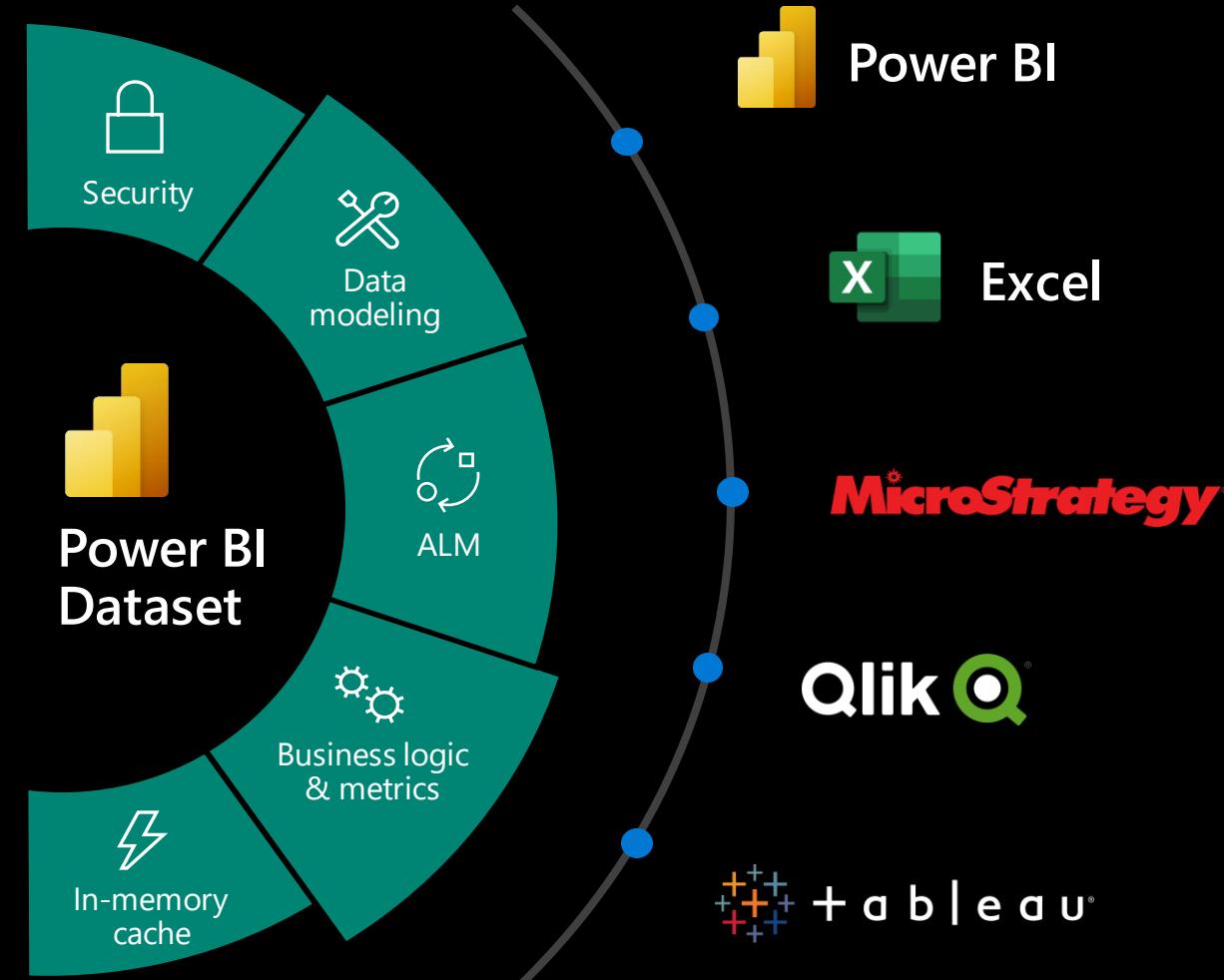


Generally Available

## World's leading OLAP engine

---

Blazing fast performance with connectivity for a variety of data visualization applications

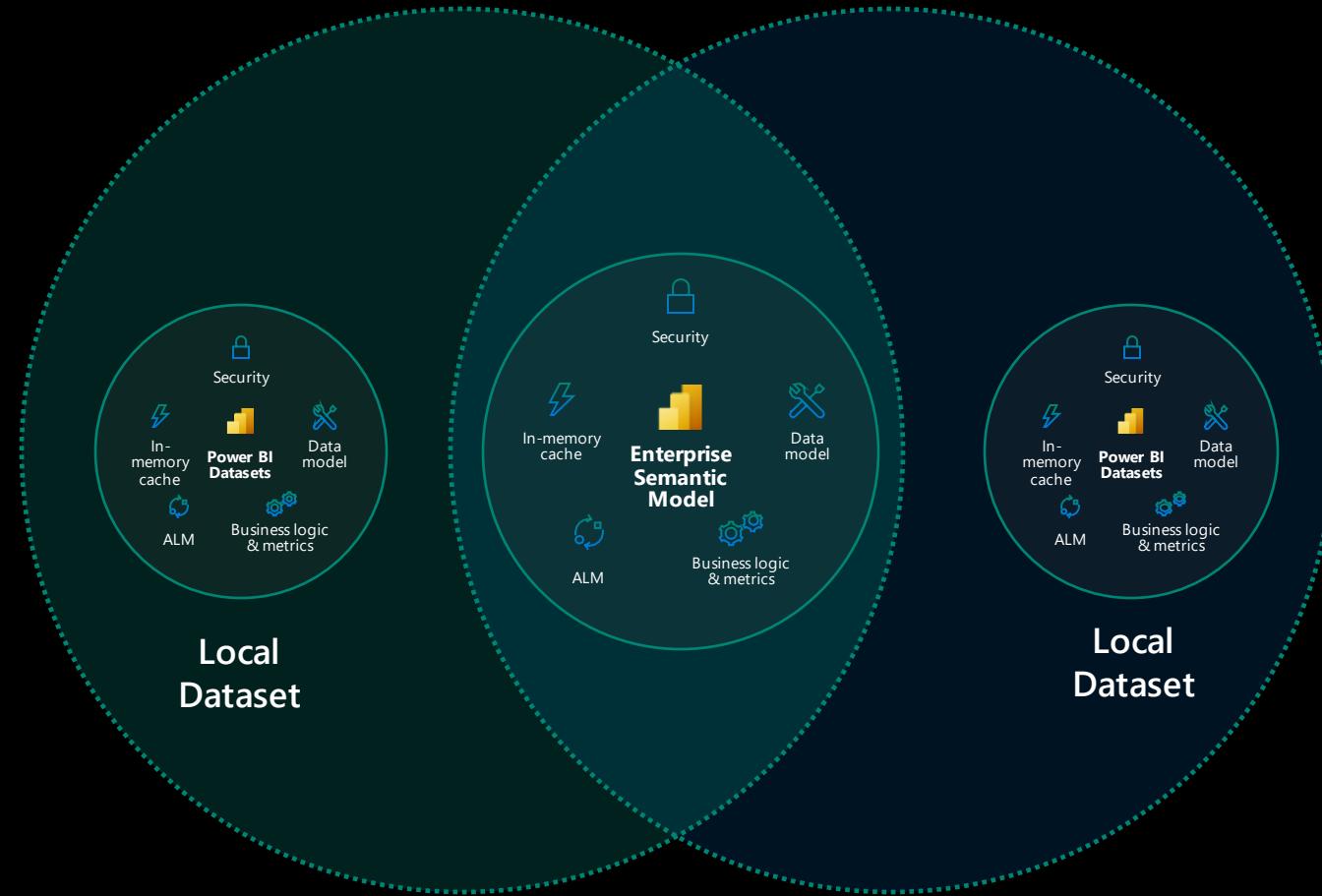


## Generally Available

# Combine enterprise and local datasets

---

Seamless evolution from self-service BI to an enterprise semantic model for company wide adoption



## Public Preview

# Composite models

Seamlessly combine and extend self-service BI with corporate BI models

The screenshot shows the Power BI Desktop interface with a report titled "Sales Report". The report features three main KPIs at the top: "Recharge Amount" (\$8.6M), "Sales" (\$94.12M), and "Pipeline" (\$78.71M). Below these are two visualizations: a bar chart showing the distribution of sales by hour of the day (0 to 23) and a card visualization for "Dynamics 365" showing a sequence of numbers from 0 to 6. A data table is also present, listing daily sales figures from April 24 to May 5, 2020.

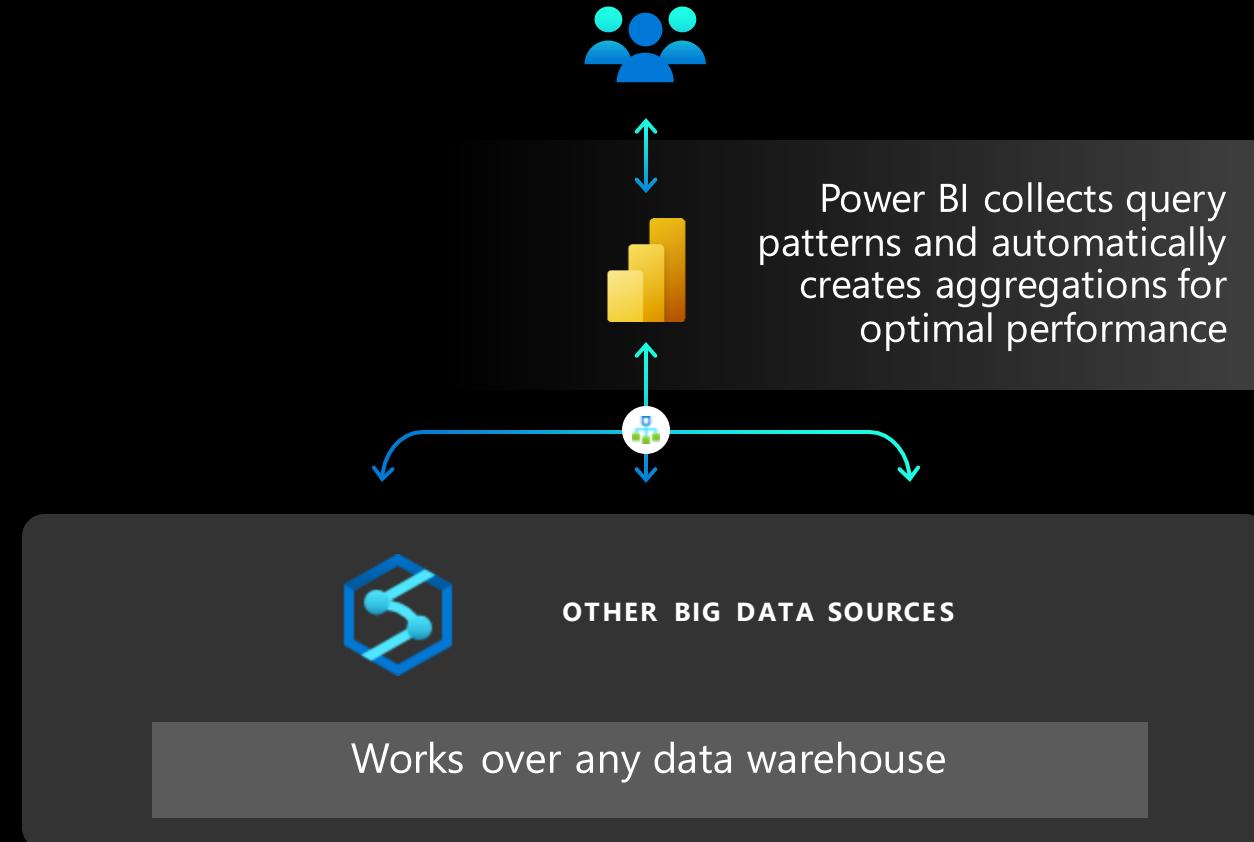
Date	Sales
4/24/2020	\$7,092,263.04
4/25/2020	\$7,627,960.2
4/26/2020	\$7,832,140.77
4/27/2020	\$7,417,329.26
4/28/2020	\$6,768,925.96
4/29/2020	\$6,478,403.22
4/30/2020	\$6,762,418.12
5/1/2020	\$7,211,322.3
5/2/2020	\$7,593,083.56
5/3/2020	\$7,883,966.76
5/4/2020	\$7,480,390.01
5/5/2020	\$6,983,100.07

The Power BI Desktop ribbon is visible at the top, showing tabs like File, Home, Insert, Modeling, View, Help, and External Tools. The right side of the screen displays the "Visualizations" and "Fields" panes, which are used for managing the composite model components.

## Public Preview

# Automatic Aggregations

Automatically learns about customer usage patterns and create aggregates to optimize performance and reduce cost



## Public Preview

# Automatic Aggregations

AI driven self-optimizing performance improvement

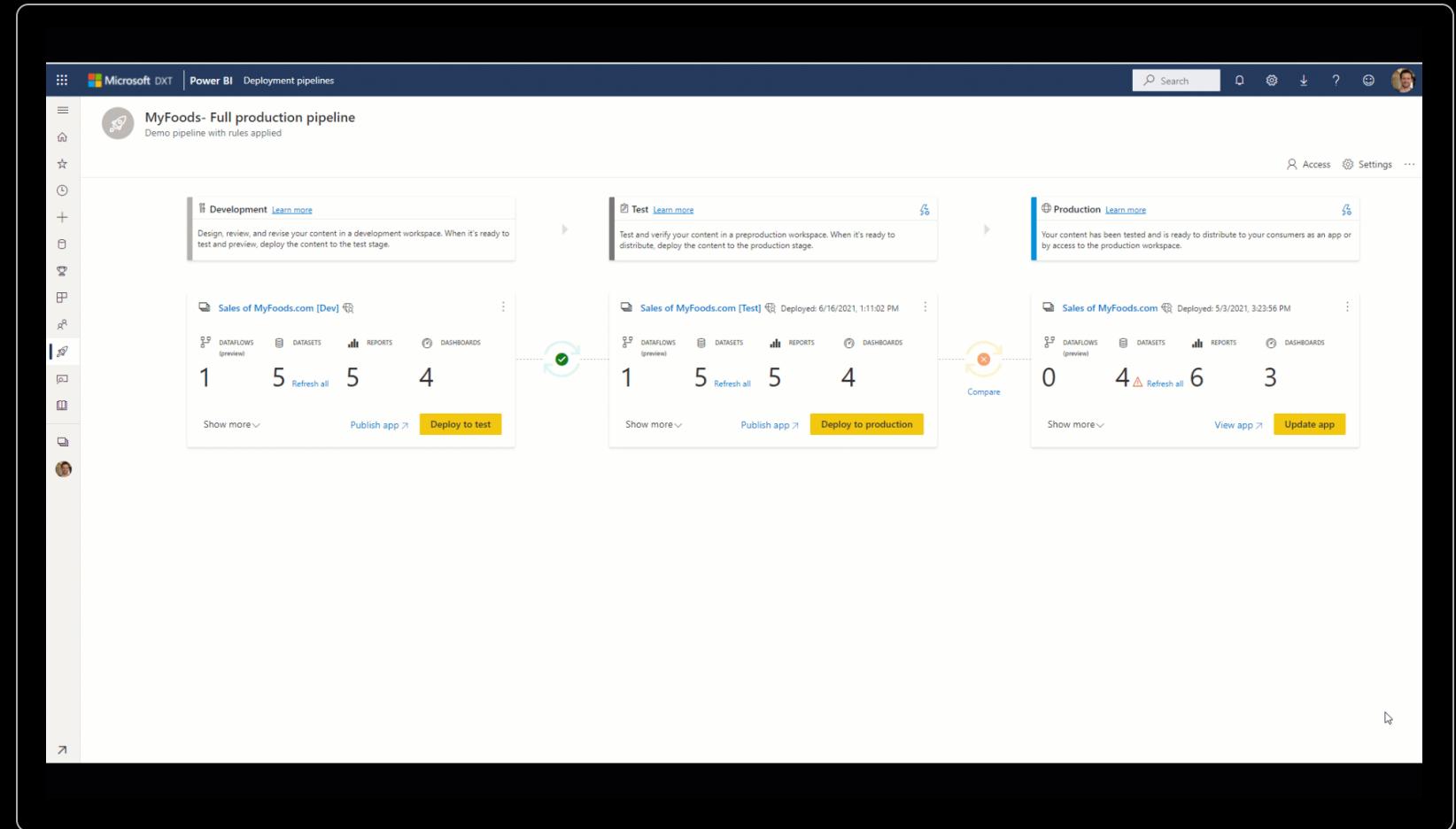
The screenshot shows the Power BI interface for the 'NY Taxi Analysis' dataset. The left sidebar contains navigation icons for Home, Favorites, Recent, New, Create a pipeline, Datasets + dataflows, and Reports. The main content area displays the dataset list with the following details:

Name	Type	Owner	Refreshed	Next refresh	Endorsement
MS Finance1	Dataset	NY Taxi Analysis	4/11/21, 8:52:51 PM	N/A	—
MS Finance2	Dataset	NY Taxi Analysis	4/27/21, 5:21:16 AM	N/A	—
Taxi1	Dataset	NY Taxi Analysis	4/12/21, 8:01:06 AM	6/30/21, 8:00:00 AM	—
Taxi2	Dataset	NY Taxi Analysis	4/27/21, 5:21:16 AM	N/A	—

## Generally Available

# Controlled change management

Power BI deployment pipelines enable efficient and reusable release processes

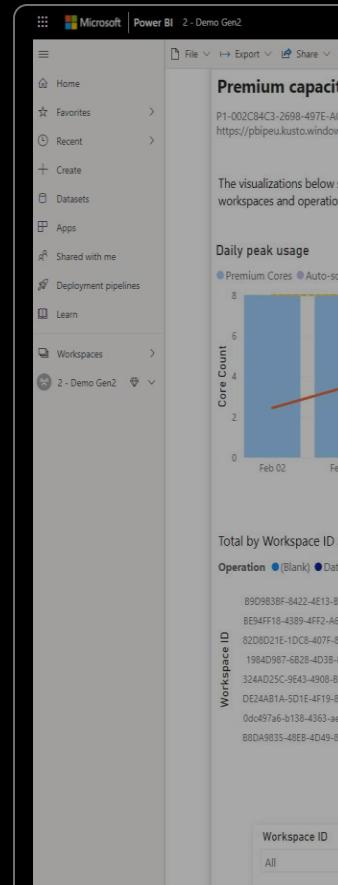


## Generally Available

# Premium Gen2

Unmatched large-scale analytics with simple low-overhead administration

Over 50%  
Premium Customer Nodes  
running on Gen 2 two weeks  
after release



André Kamman @AndreKamman · Oct 14

Went from carefully scheduling refreshes and still getting memory errors regularly to REFRESH ALL THE THINGS! Power BI Premium users, move to Gen2, do it now! My biggest model went from a 70 minute processing time to 12!

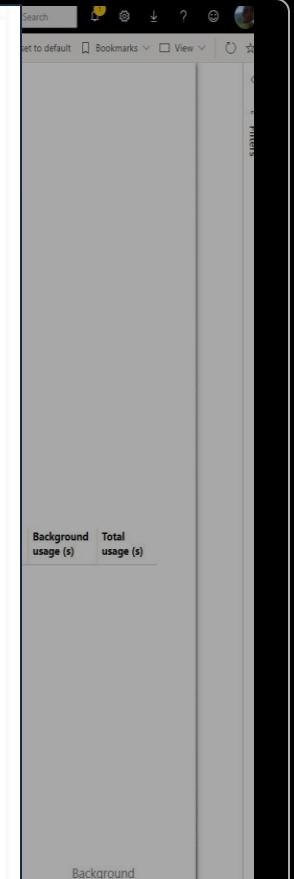


2

2

27

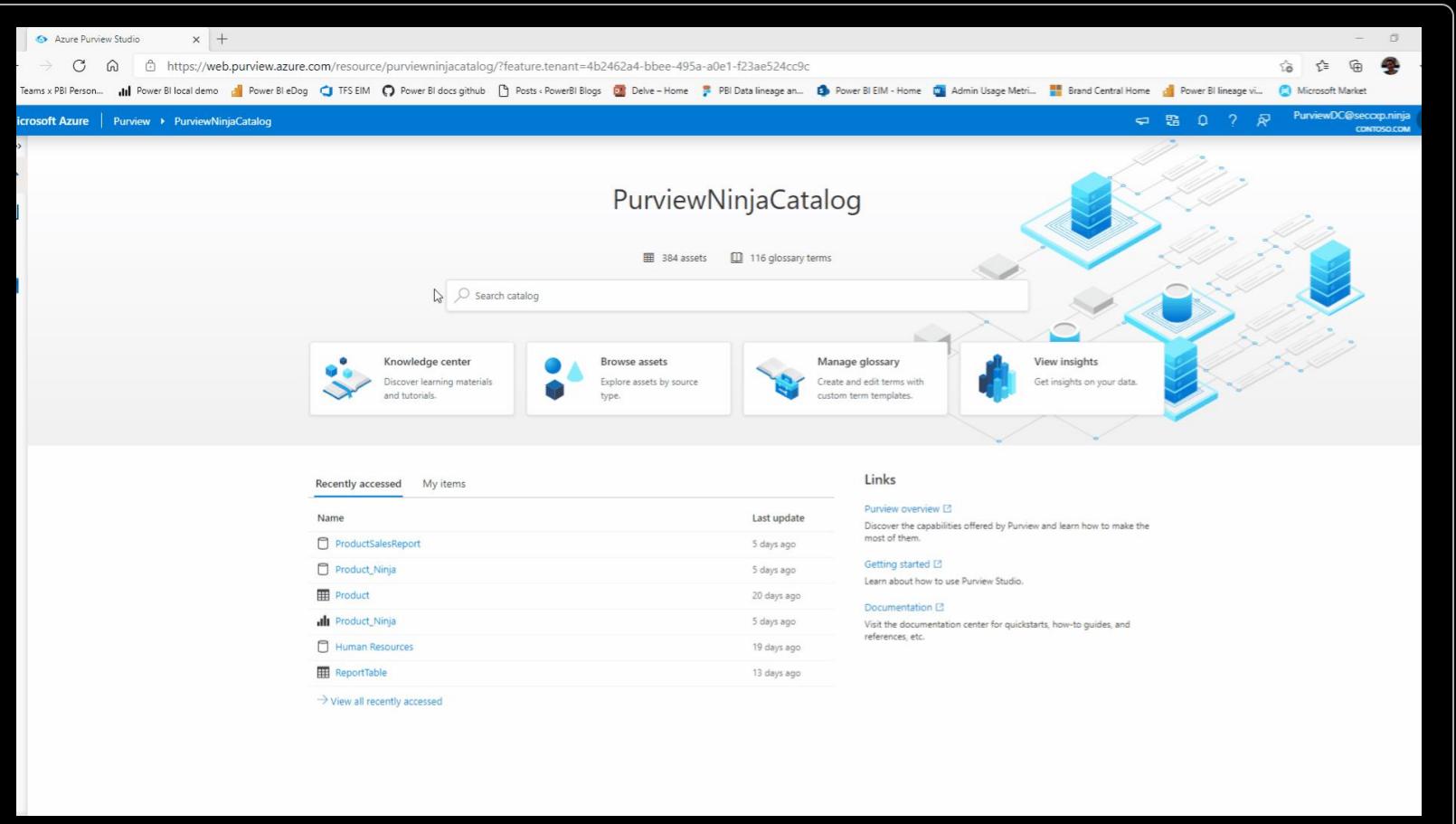
↑

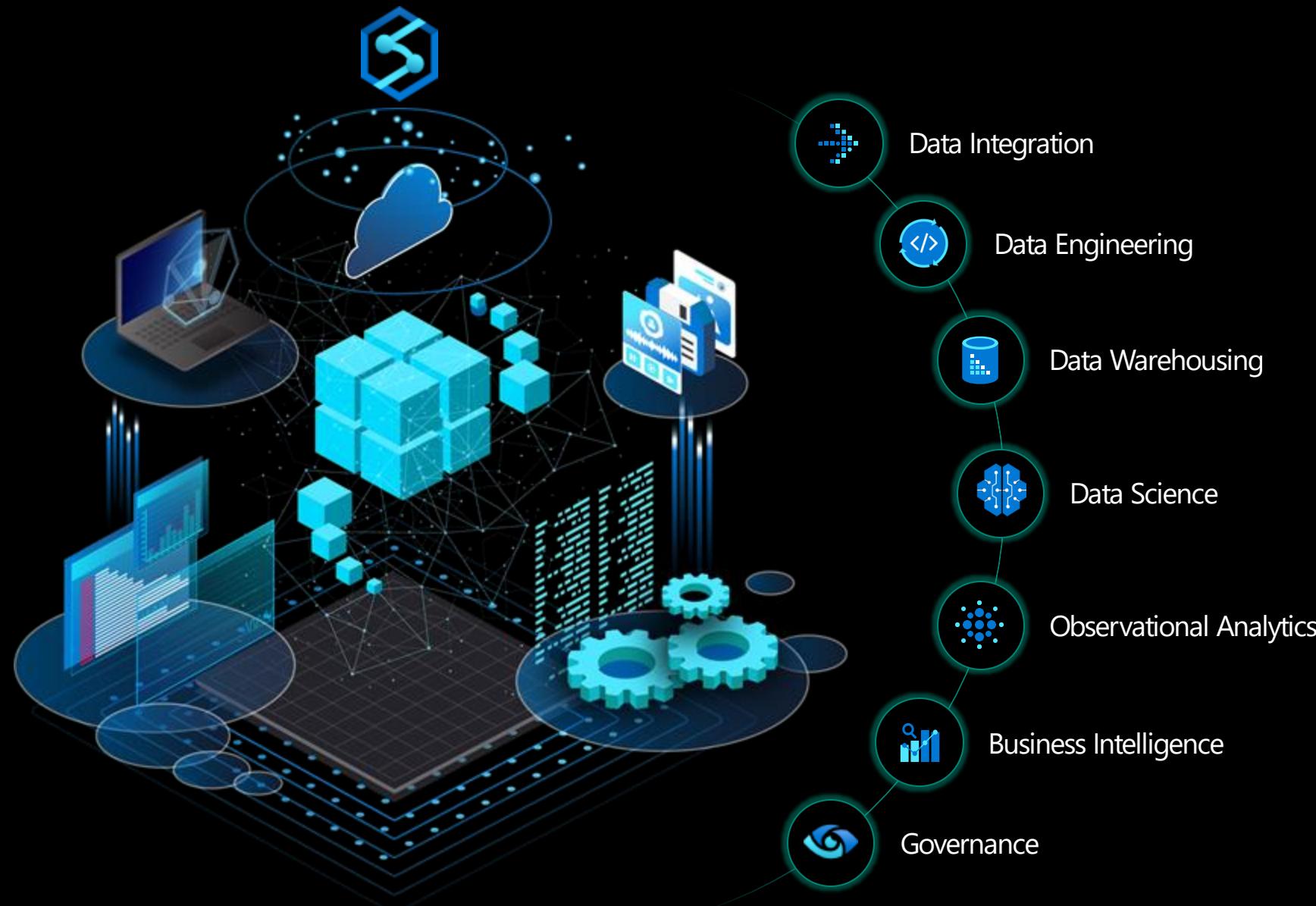


## Generally Available

# Power BI + Azure Purview

Enhanced governance and cataloging capabilities integrated with Power BI







Governance

# Governance



## Generally Available

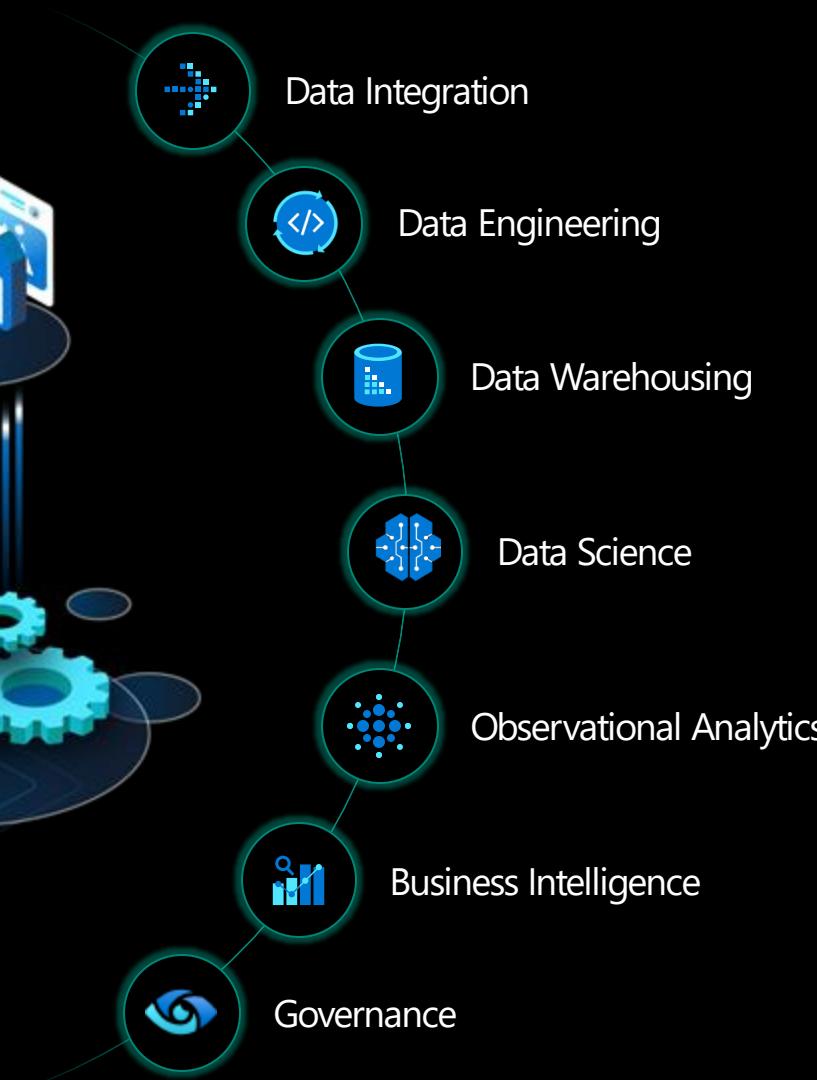
# Integrated Catalog search in Synapse

Enables developers to search for data assets across the entire data estate to analyze with Synapse

The screenshot shows the Microsoft Azure Synapse Analytics Data Explorer interface. The left sidebar has a 'Data' section with tabs for 'Workspace' (selected) and 'Linked'. A search bar at the top right says 'Search'. Below the tabs is a 'Filter resources by name' input field. Under 'Lake database', there are three items: 'default', 'retaildata', and 'surfacesalesdb'. Under 'SQL database', there are six items: 'Contoso\_Health (SQL)', 'newpoll (SQL)', 'NYCTaxi\_Pool (SQL)', 'Predict\_Pool (SQL)', 'Streaming\_Pool (SQL)', and 'WWI\_Pool (SQL)'. To the right of the list is a 'Select an item' section with two cylinders and a monitor icon, and the text 'Use the resource explorer to select or create a new item'.



# Synapse + Power BI





Partners make more possible



# Enregistrez vous dès maintenant au prochain Webinars Data AI

Event Webinar (Les jeudis de la Data & AI) - L200/300	Date	Duration (min)	Link
Azure Machine Learning pour les Data Scientists	15/09/2022	120	<a href="https://msevents.microsoft.com/event?id=2454281594">https://msevents.microsoft.com/event?id=2454281594</a>
Azure Synapse	22/09/2022	120	<a href="https://msevents.microsoft.com/event?id=857781749">https://msevents.microsoft.com/event?id=857781749</a>
Les solutions SQL dans Azure (PaaS, IaaS, SaaS)	29/09/2022	120	<a href="https://msevents.microsoft.com/event?id=502366997">https://msevents.microsoft.com/event?id=502366997</a>
Déploiement et sécurisation des workspaces Azure Machine learning	06/10/2022	120	<a href="https://msevents.microsoft.com/event?id=1505714138">https://msevents.microsoft.com/event?id=1505714138</a>
Azure Scale Analytics - Architectures Data Mesh dans Azure avec Azure Synapse, Microsoft Purview et Azure Data Share	13/10/2022	120	<a href="https://msevents.microsoft.com/event?id=139685175">https://msevents.microsoft.com/event?id=139685175</a>
MLOps avec Azure Machine Learning	20/10/2022	120	<a href="https://msevents.microsoft.com/event?id=1245885767">https://msevents.microsoft.com/event?id=1245885767</a>
SQL Server 2022 et hybridation native avec Azure SQL Managed Instance	10/11/2022	120	<a href="https://msevents.microsoft.com/event?id=145826476">https://msevents.microsoft.com/event?id=145826476</a>
Machine Learning dans Azure Synapse Analytics	17/11/2022	120	<a href="https://msevents.microsoft.com/event?id=3637723312">https://msevents.microsoft.com/event?id=3637723312</a>
Azure Cosmos DB et IA	24/11/2022	120	<a href="https://msevents.microsoft.com/event?id=2646013445">https://msevents.microsoft.com/event?id=2646013445</a>
Azure et les Services Cognitifs	08/12/2022	120	<a href="https://msevents.microsoft.com/event?id=3772037220">https://msevents.microsoft.com/event?id=3772037220</a>
La gouvernance de données dans Azure avec Microsoft Purview	15/12/2022	120	<a href="https://msevents.microsoft.com/event?id=1499560981">https://msevents.microsoft.com/event?id=1499560981</a>
MLOps avec Azure Machine Learning	12/01/2023	120	<a href="https://msevents.microsoft.com/event?id=4115194515">https://msevents.microsoft.com/event?id=4115194515</a>
	19/01/2023	120	<a href="https://msevents.microsoft.com/event?id=1537241181">https://msevents.microsoft.com/event?id=1537241181</a>
Data processing dans Azure ave Azure Synapse, Azure Batch, Spark, Notebook, etc.	26/01/2023	120	<a href="https://msevents.microsoft.com/event?id=1806467748">https://msevents.microsoft.com/event?id=1806467748</a>
Déploiement et sécurisation des workspace Azure Synapse	09/02/2023	120	<a href="#">En cours</a>
Azure Machine Learning pour les Citizen Data Scientists	16/02/2023	120	<a href="https://msevents.microsoft.com/event?id=1401519679">https://msevents.microsoft.com/event?id=1401519679</a>
L'IA responsable avec Azure machine learning	09/03/2023	120	<a href="https://msevents.microsoft.com/event?id=2072953112">https://msevents.microsoft.com/event?id=2072953112</a>
Machine Learning dans Azure Synapse Analytics	16/03/2023	120	<a href="https://msevents.microsoft.com/event?id=3413014857">https://msevents.microsoft.com/event?id=3413014857</a>
Les bases de données Open Source dans le cloud Azure	23/03/2023	120	<a href="https://msevents.microsoft.com/event?id=2727487131">https://msevents.microsoft.com/event?id=2727487131</a>
Hybridation des services de Machine Learning Azure	06/04/2023	120	<a href="https://msevents.microsoft.com/event?id=1624914222">https://msevents.microsoft.com/event?id=1624914222</a>
La gouvernance de données dans Azure avec Microsoft Purview	13/04/2023	120	<a href="https://msevents.microsoft.com/event?id=3909342839">https://msevents.microsoft.com/event?id=3909342839</a>
Les solutions SQL dans Azure (PaaS, IaaS, SaaS)	04/05/2023	120	<a href="https://msevents.microsoft.com/event?id=1162207895">https://msevents.microsoft.com/event?id=1162207895</a>
	16/05/2023	120	<a href="https://msevents.microsoft.com/event?id=3517068442">https://msevents.microsoft.com/event?id=3517068442</a>
Data processing dans Azure ave Azure Synapse, Azure Batch, Spark, Notebook, etc.	24/05/2023	120	<a href="https://msevents.microsoft.com/event?id=2996507398">https://msevents.microsoft.com/event?id=2996507398</a>
Hybridation des services de données Azur	01/06/2023	120	<a href="#">En cours</a>