Microsoft

GPS Data/AI Strategy FY23

Delivered by CSA Team

Franck Gaillard
Cloud Solution Architect
Data AI
frgail@microsoft.com

Narjes Majdoub
Cloud Solution Architect
Data AI
nmajdoub@microsoft.com

Ali Bouhaddou
Cloud Solution Architect
Data Analytics
albouhad@microsoft.com

Frederic Gisbert
Cloud Solution Architect
Data Analytics
frgisber@microsoft.com

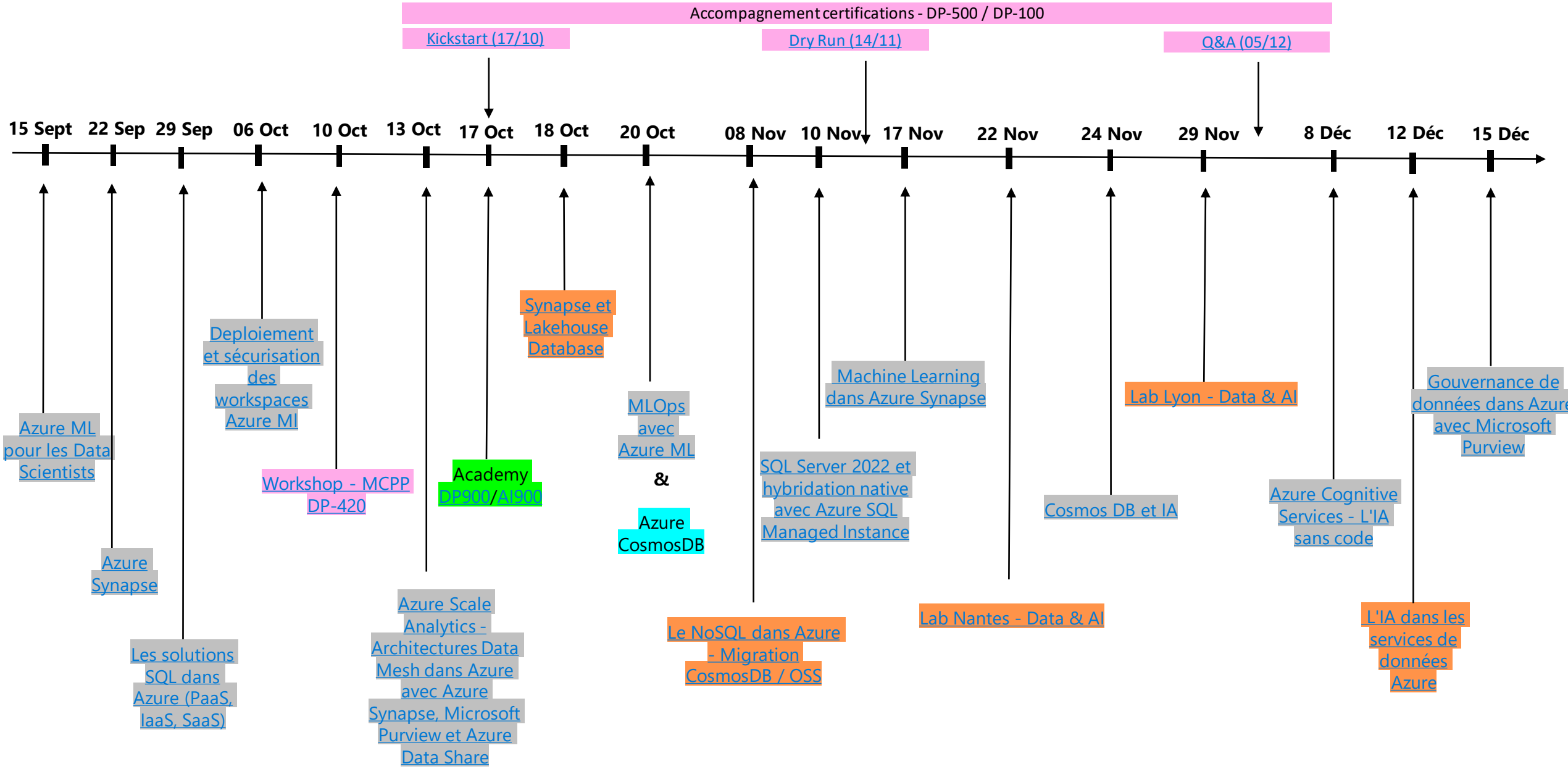
Azure Data & AI technical intensity plan

- 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

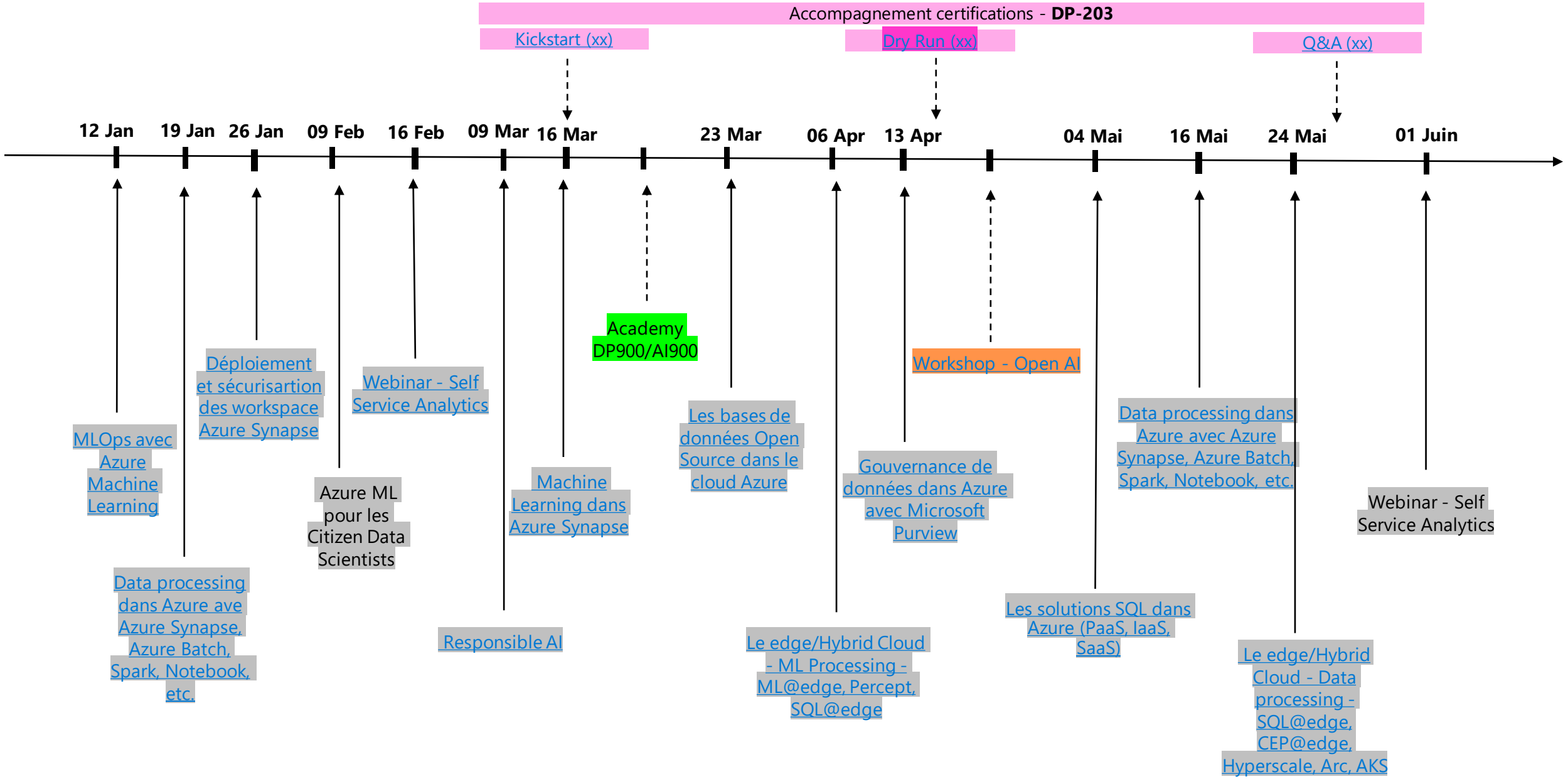
Data & AI events timeline – H1

Webinar/Academy- L 200/300
Workshop/ Openhack/ Certifications – L 300/400



Data & AI events timeline – H2

Webinar/Academy- L 200/300
Workshop/ Openhack/ Certifications – L 300/400



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

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	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFYSDVLR0cyS1kwWS4u
Le NoSQL dans Azure - Migration CosmosDB / OSS	08/11/2022	120	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFYSDVLR0cyS1kwWS4u
Lab Lyon - Data & AI	22/11/2022	240	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUMlZZOURET0RSWjcyTERYRkJGTlFFUjJaUi4u
Lab Nantes - Data & AI	29/11/2022	240	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUMlZZOURET0RSWjcyTERYRkJGTlFFUjJaUi4u
L'IA dans les services de données Azure	12/12/2022	120	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFYSDVLR0cyS1kwWS4u
Open AI	H2	120	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdURE1RMVgwTDNISTE1TDFYSDVLR0cyS1kwWS4u

Event Academy, kickstart certifications, workshop certifications	Date	Duration (min)	Link
MCPP - DP-420	10/10/2022	420	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUMkJISlRKSU1RRFA0OVgzSFdTSTY0RE9WQy4u
Micro Hack CosmosDB	20/10/2022	420	H1 - Inscriptions PTA
Academy DP900	17-21/10/2022	300	https://msevents.microsoft.com/event?id=3250818161
Academy AI900	17-21/10/2022	300	https://msevents.microsoft.com/event?id=2717528090
Kickstart DP-500	17/10/2022	60	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTQ2Uk85V0cxQzM3TE9ZRS4u
Dry Run DP-500	14/11/2022	120	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTQ2Uk85V0cxQzM3TE9ZRS4u
Q&A DP-500	05/12/2022	90	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUNEk3WFQ1TEdNNTQ2Uk85V0cxQzM3TE9ZRS4u
Kickstart DP-100	17/10/2022	60	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI3OUNMUFYxS1kRIMi4u
Dry Run DP-100	14/11/2022	120	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI3OUNMUFYxS1kRIMi4u
Q&A DP-100	05/12/2022	90	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUNDAxV0hSN0FHM1YzUzI3OUNMUFYxS1kRIMi4u
Kickstart DP-203	17/10/2022	60	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyQk5SVjFBUFczNktCUFpLMi4u
Dry Run DP-203	14/11/2022	120	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyQk5SVjFBUFczNktCUFpLMi4u
Q&A DP-203	05/12/2022	90	https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR3zwJTO3s11AuaqpNnBbrwdUOVFWOUVCNFcyQk5SVjFBUFczNktCUFpLMi4u

Microsoft

Machine learning in Azure Synapse

17/11/2022

Frederic Gisbert
Cloud Solution Architect
Data Analytics
frgisber@microsoft.com

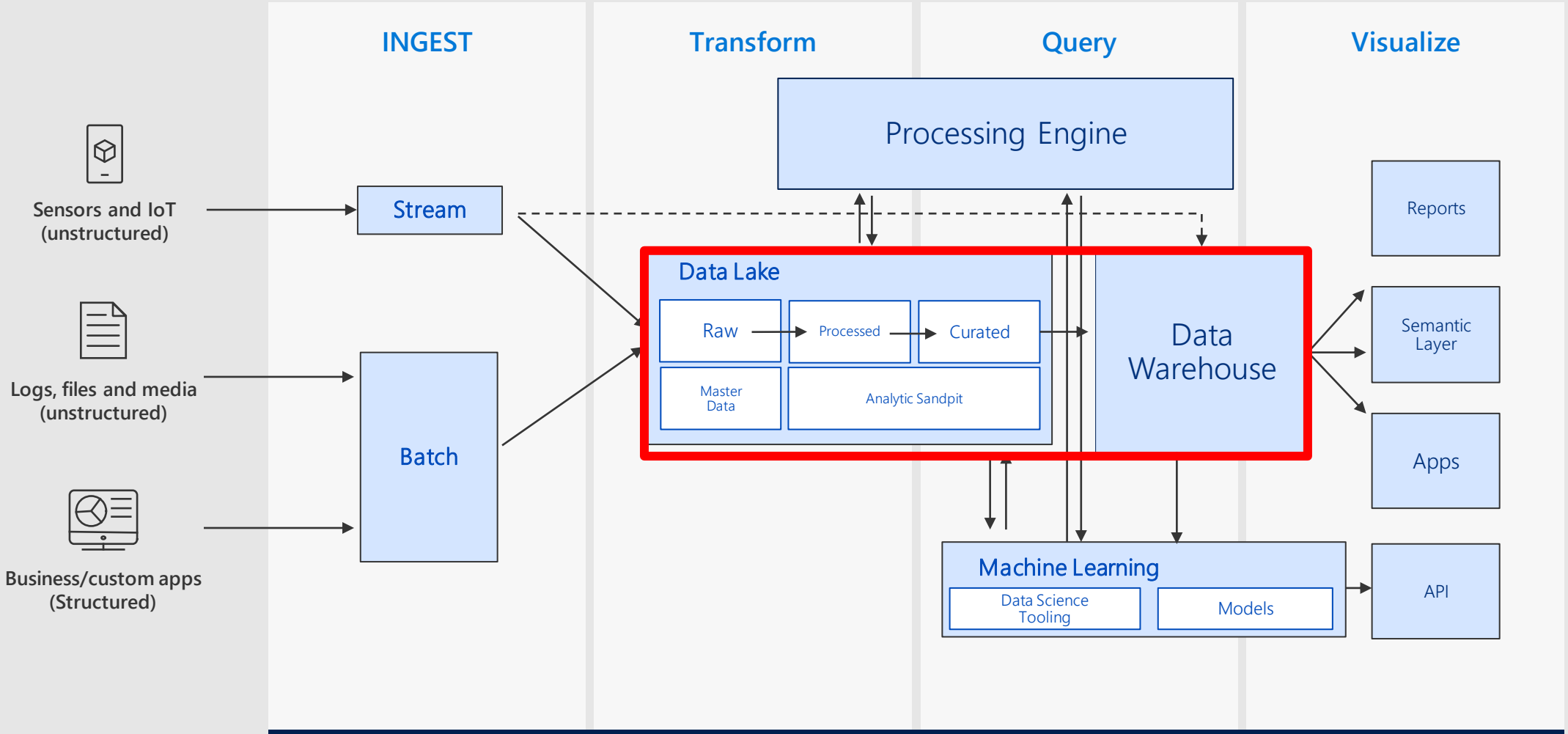
Narjes Majdoub
Cloud Solution Architect
Data AI
nmajdoub@microsoft.com

Agenda

1. Introduction
2. Data acquisition & preparation
3. Modeling capabilities
4. Scoring & deployment

Introduction

Cloud Data Warehouse – Conceptual



sqlbits

Lakehouse Is Born

- Databricks white paper 2020
- Arose after identifying the DW + DL might not be the silver bullet
- Lack of transaction support
- Hard to enforce data quality
- Complicated to mix appends, updates and dele
- Data Swamps not Data Lakes
- Multiple storage layers

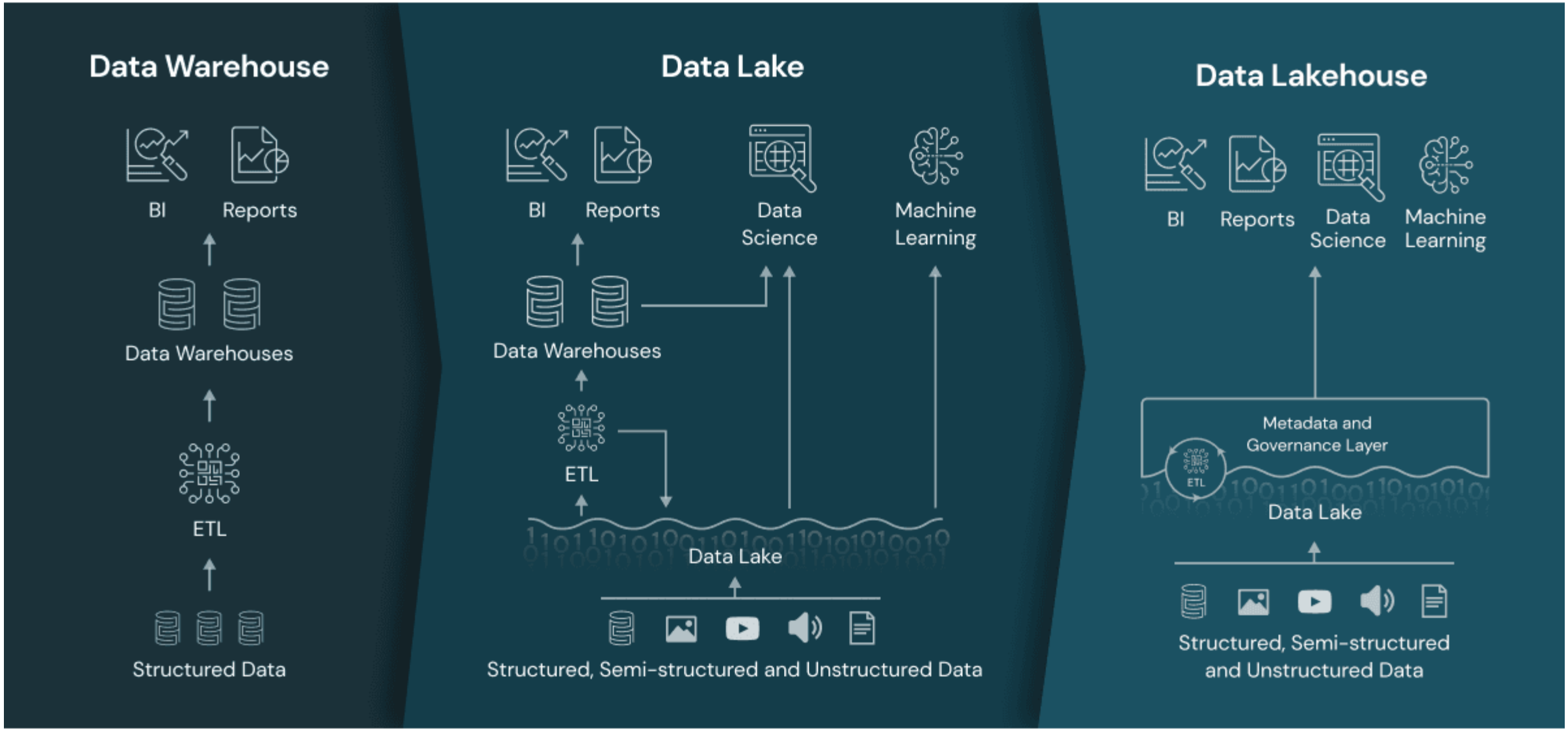
Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

Michael Armbrust[1], Ali Ghodsi[1,2], Reynold Xin[1], Matei Zaharia[1,3]
[1]Databricks, [2]UC Berkeley, [3]Stanford University

quality and governance downstream. In this architecture, a small

Source: https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

From BI to AI

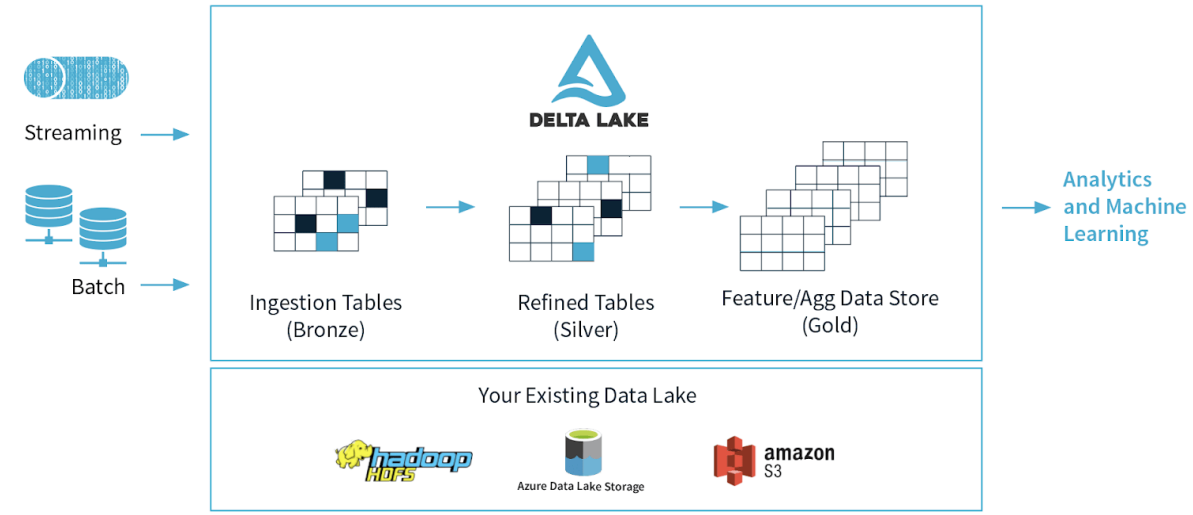


Source: https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

Delta Lake – the foundations of the Lakehouse

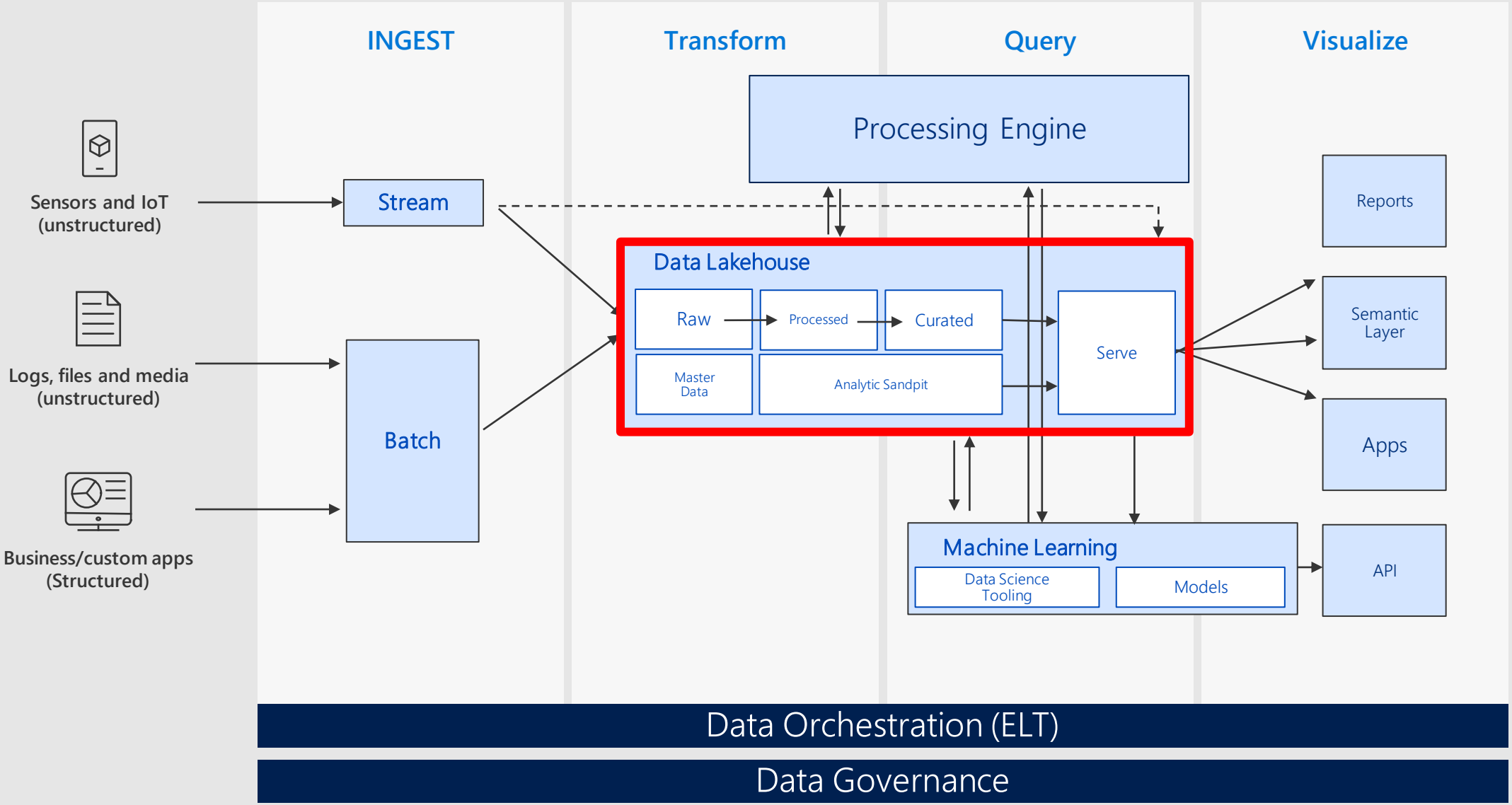
- ACID transactions
- Time travel
- Streaming and batch unification
- Schema enforcement
- Upserts and deletes (MERGE)
- Performance improvement

Delta Lake is an open-source project that enables building a Lakehouse Architecture with compute engines including **Spark, Synapse Serverless**, PrestoDB, Flink, Trino, and Hive and APIs for Scala, Java, Rust, Ruby, and Python.



Delta Lake - Reliable Data Lakes at Scale

Data Lakehouse - Conceptual



Introducing Azure Synapse Analytics

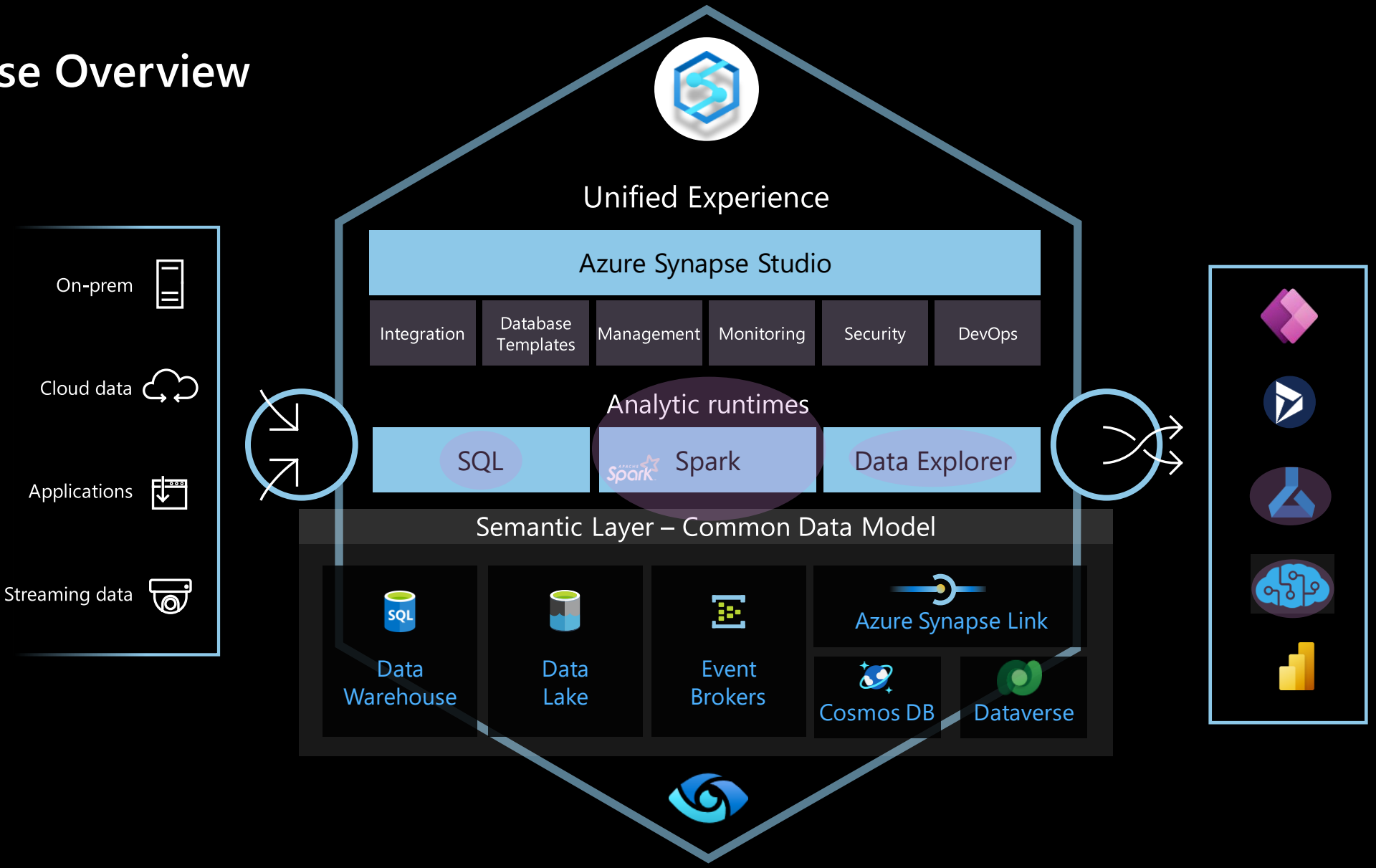
Azure Synapse Analytics

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.

Microsoft

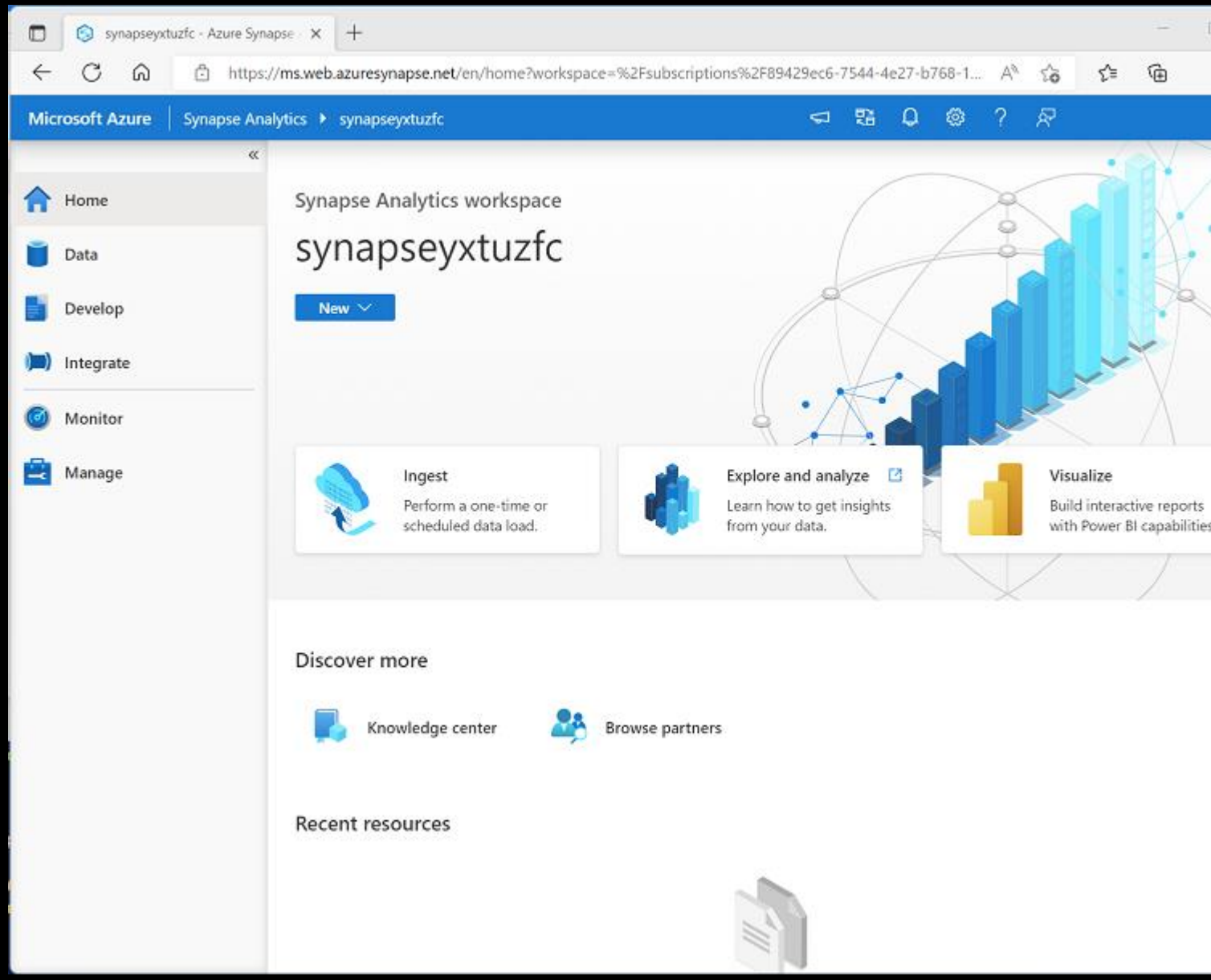
Synapse Overview



Azure Synapse Analytics workspace

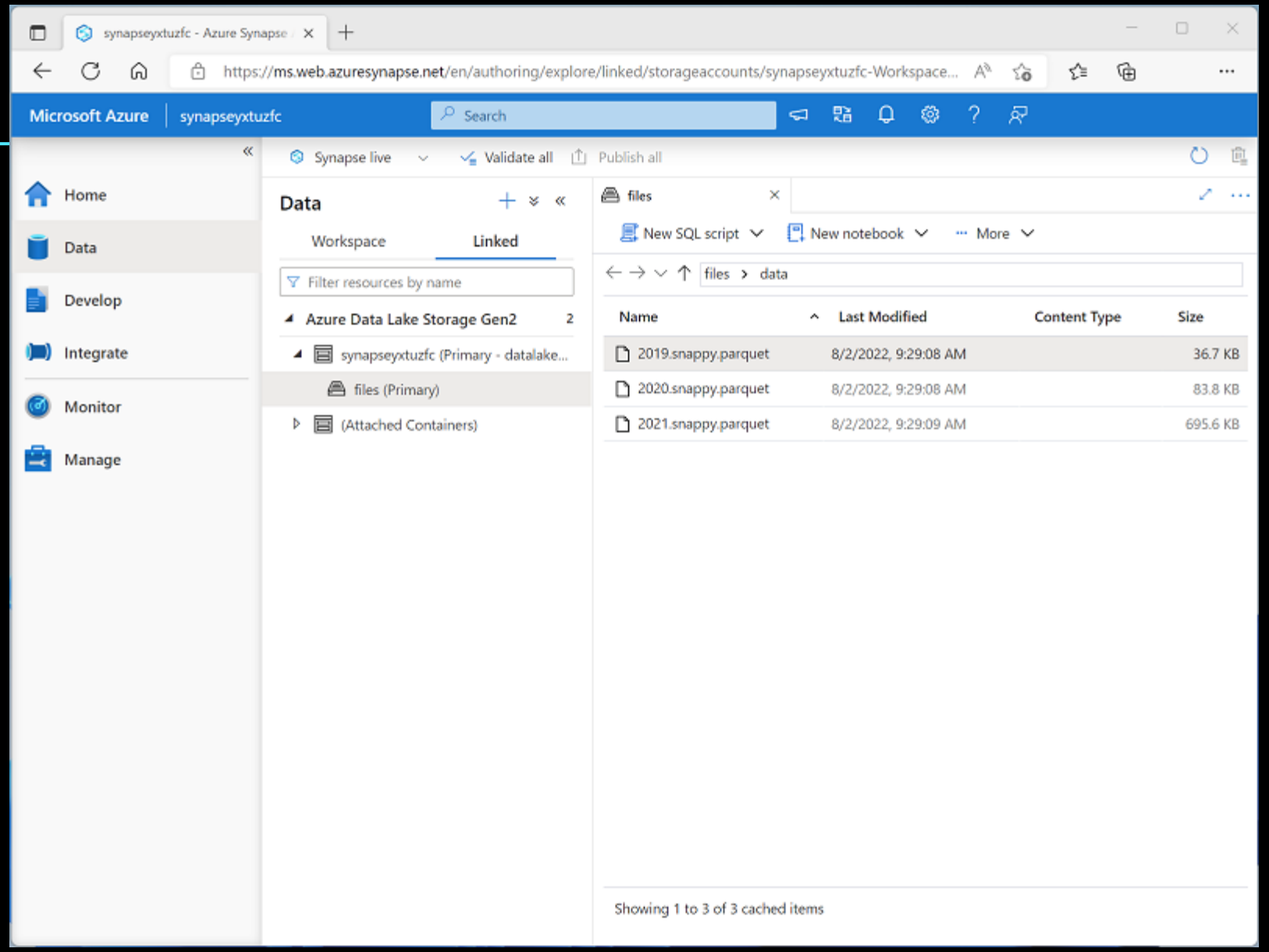
Integrated console to manage, monitor, and administer all components.

Graphical utility for managing the services and data resources



Working with files in a data lake

Storage repository that allows you to store your data in native format without having to first structure the data at any scale



Serverless SQL Pools

Flexible consumption models

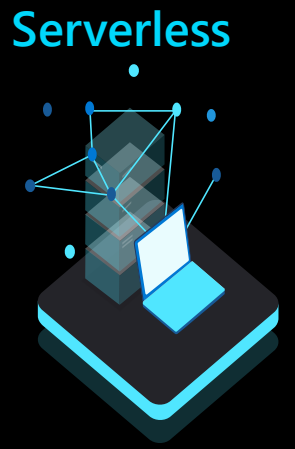
Serverless pay-per-query ideal for logical data warehouse (lakehouse) ,ad-hoc data lake exploration and transformation

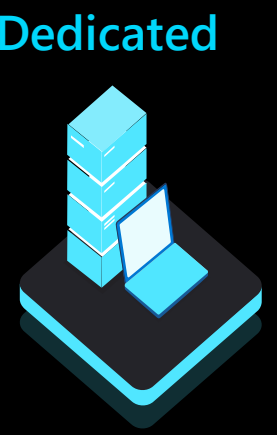
T-SQL code directly on the data lake

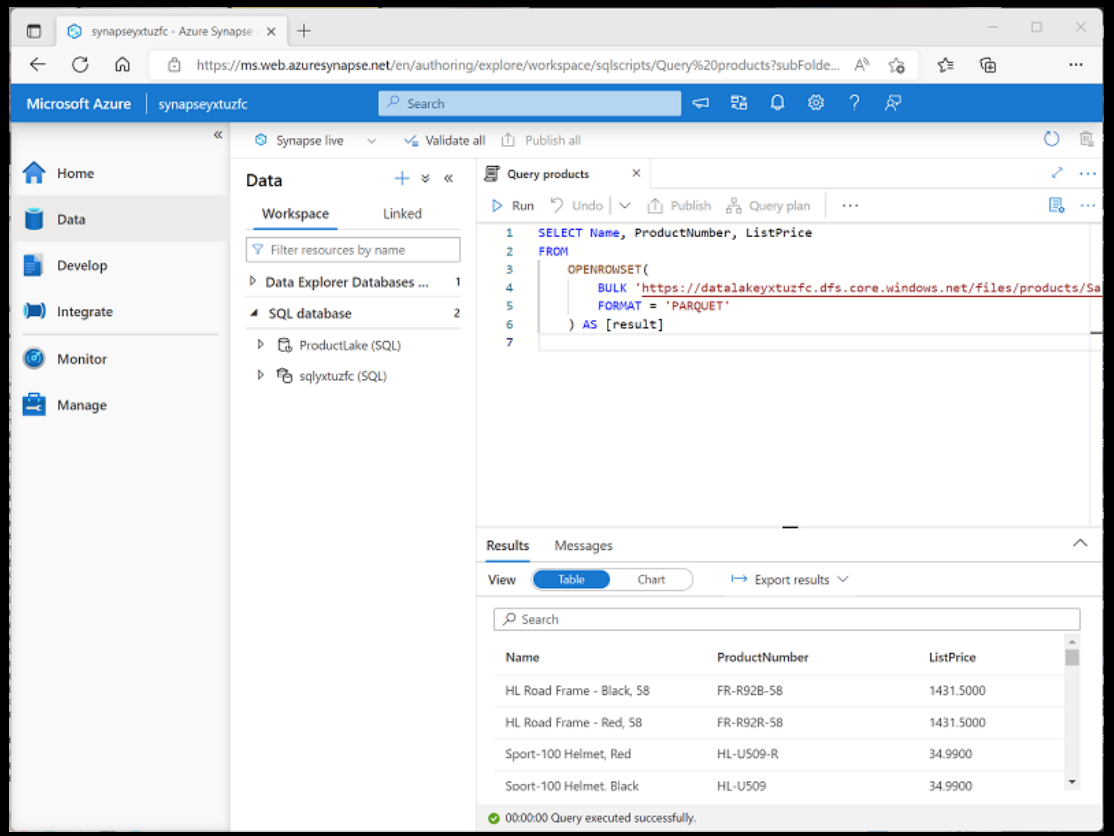
Support for Delta file system (Lakehouse) via OPENROWSET and External Tables

Supports DELTA File format

Persist common queries via CETAS







Data Engineering with Spark

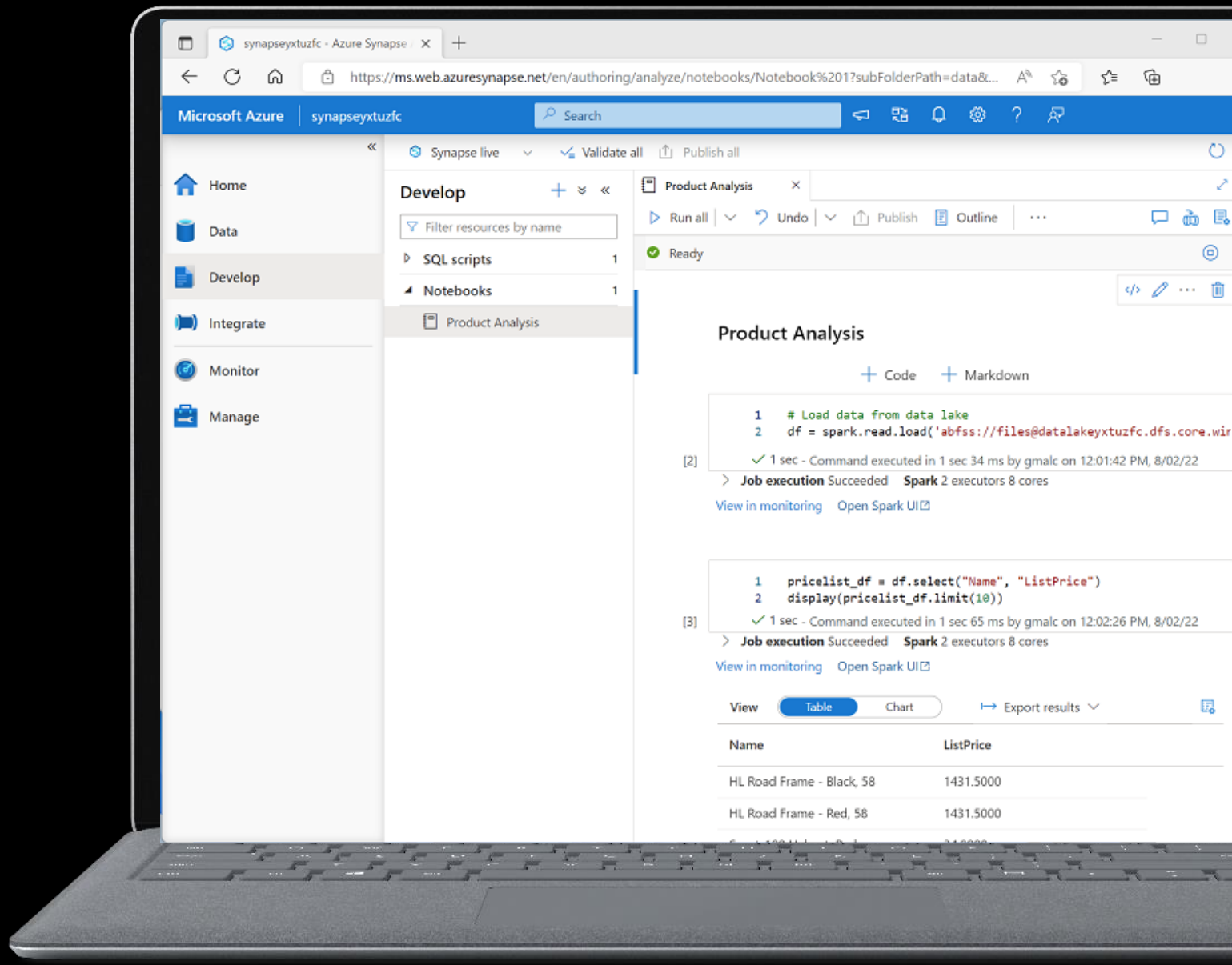
Code-first Data Engineering

PySpark, Scala, SQL and C# languages supported

Author multiple languages in a single notebook

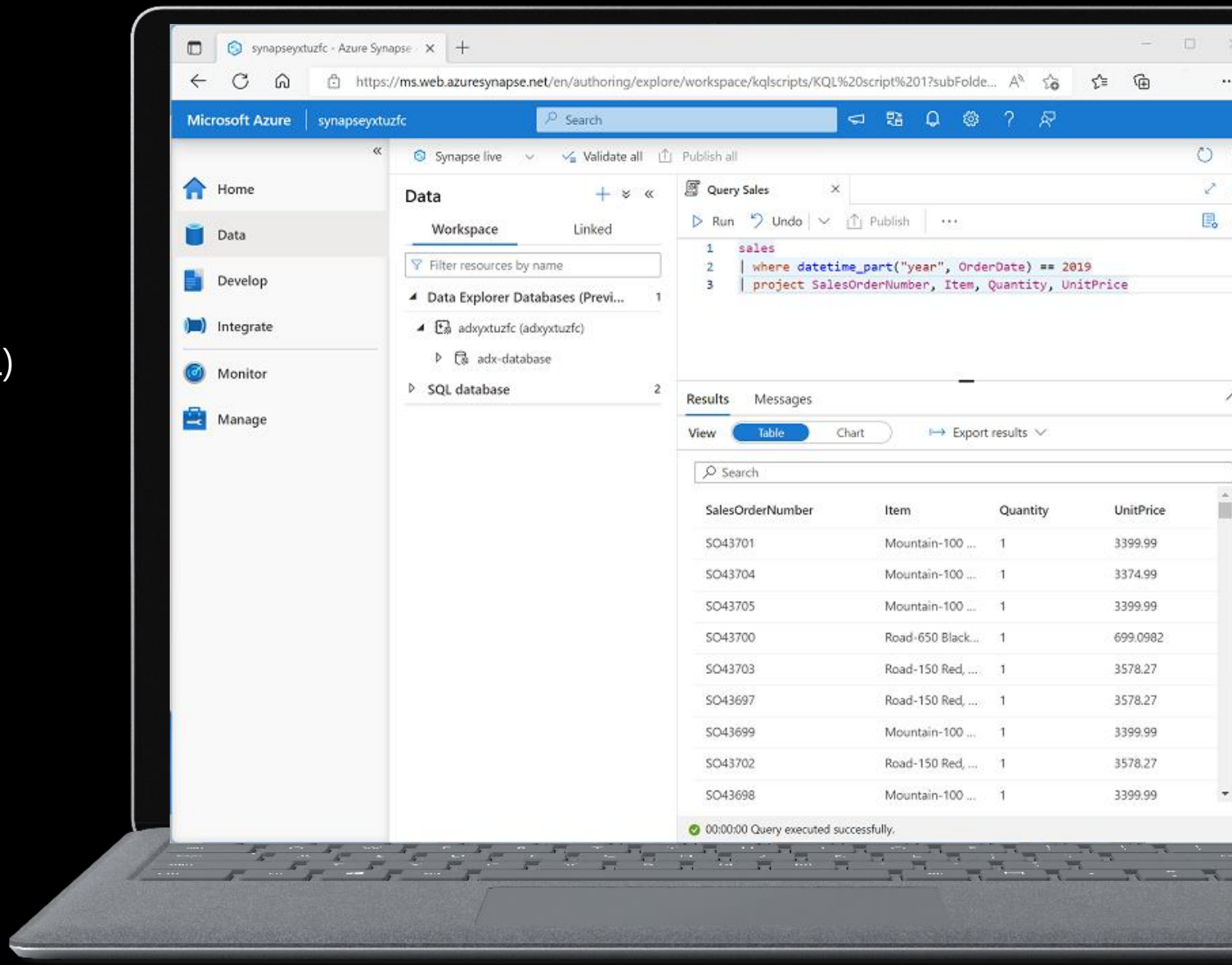
Analyze & transform data from the data warehouse, data lake, and real-time operational data from one place

Synapse Spark uses Spark 3.2 runtime, which includes Delta Lake 1.0

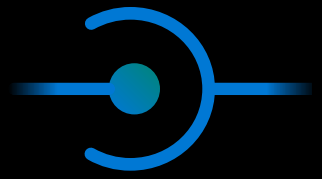


Exploring data with Data Explorer

Data Explorer uses an intuitive query syntax named Kusto Query Language (KQL) to enable high performance, low-latency analysis of batch and streaming data.



Integration with other Azure data services



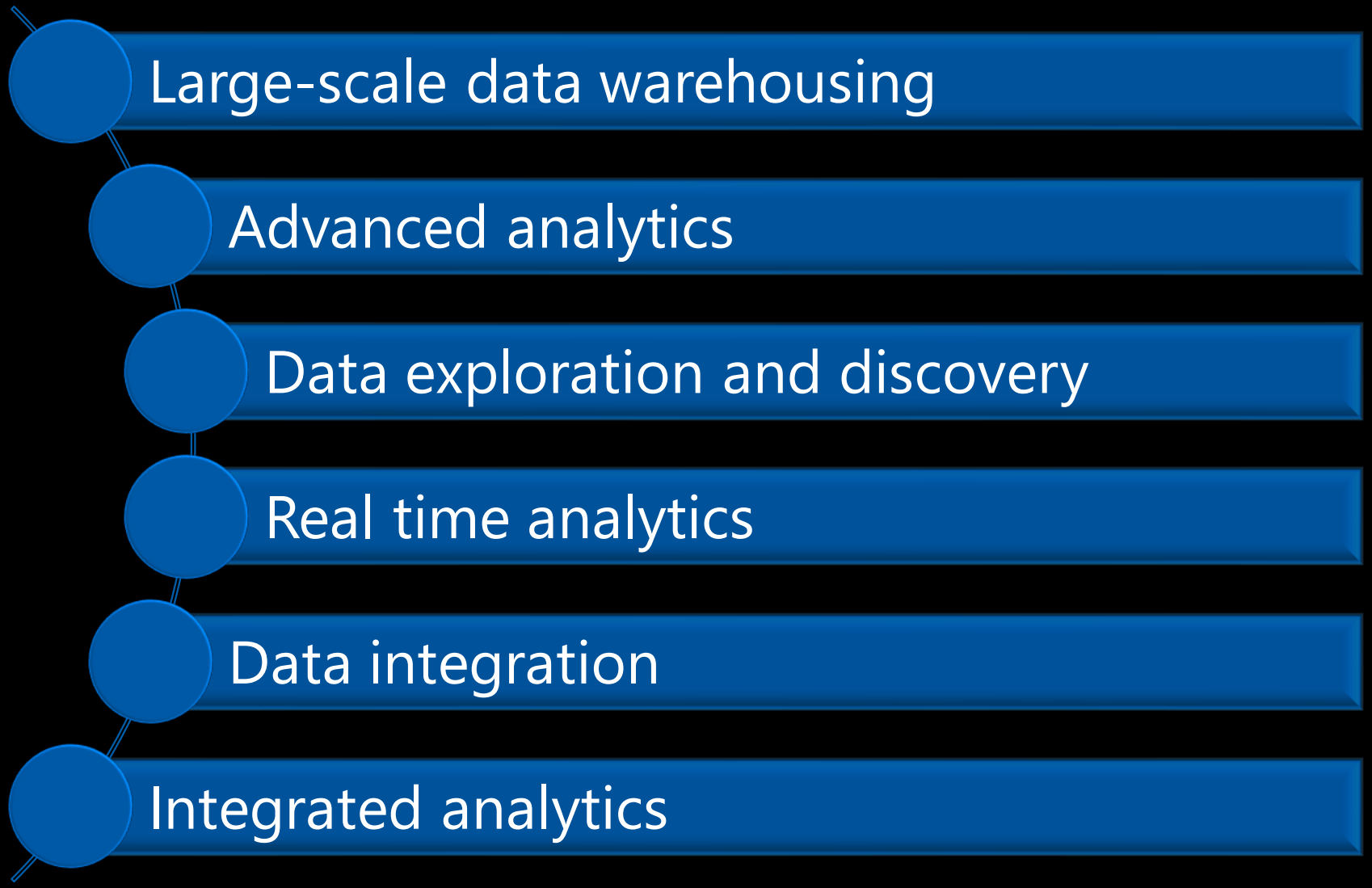
Synapse Link

Power BI

Purview

Azure ML

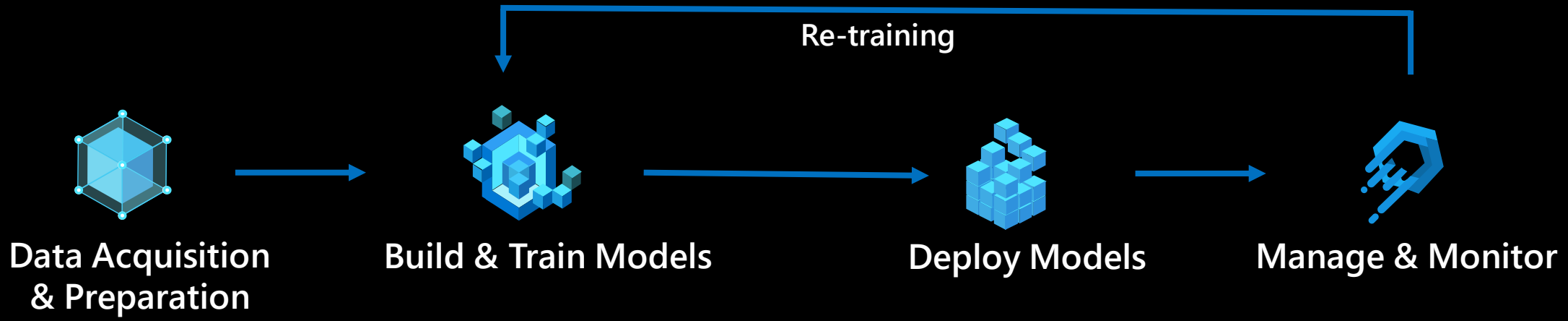
When to use Azure Synapse Analytics?



Demo:

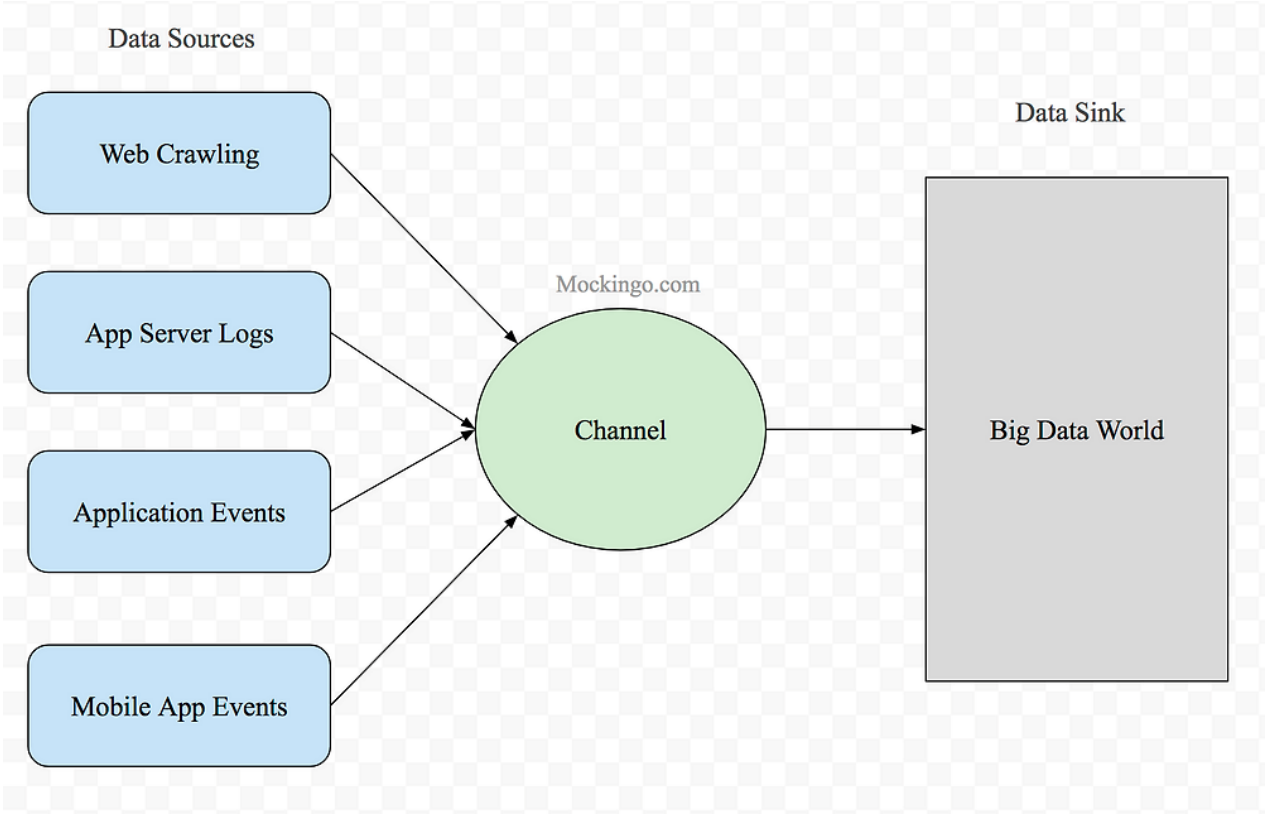
Azure Synapse Studio Tour

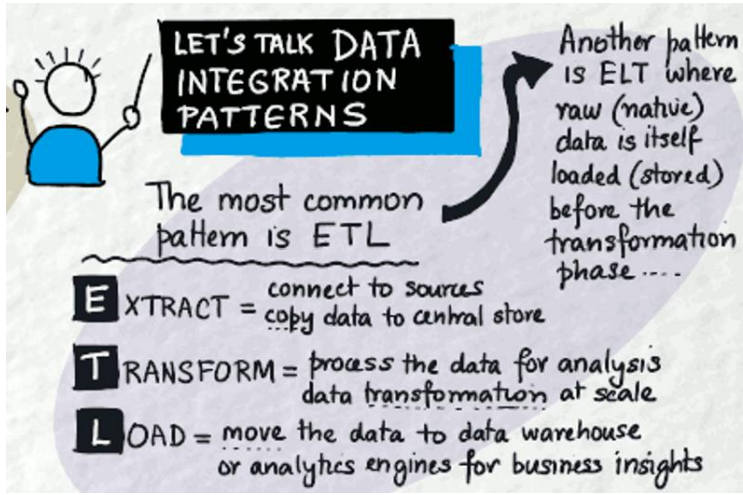
ML Lifecycle workflows

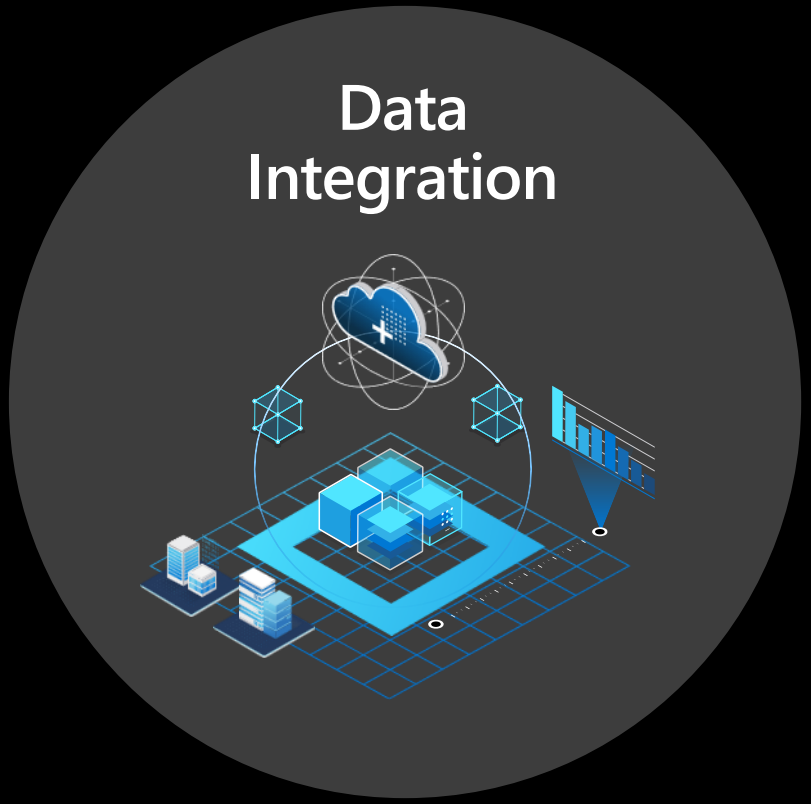


Data Acquisition & preparation

What is data ingestion?







Over 100 connectors to ingest data from a variety of platforms

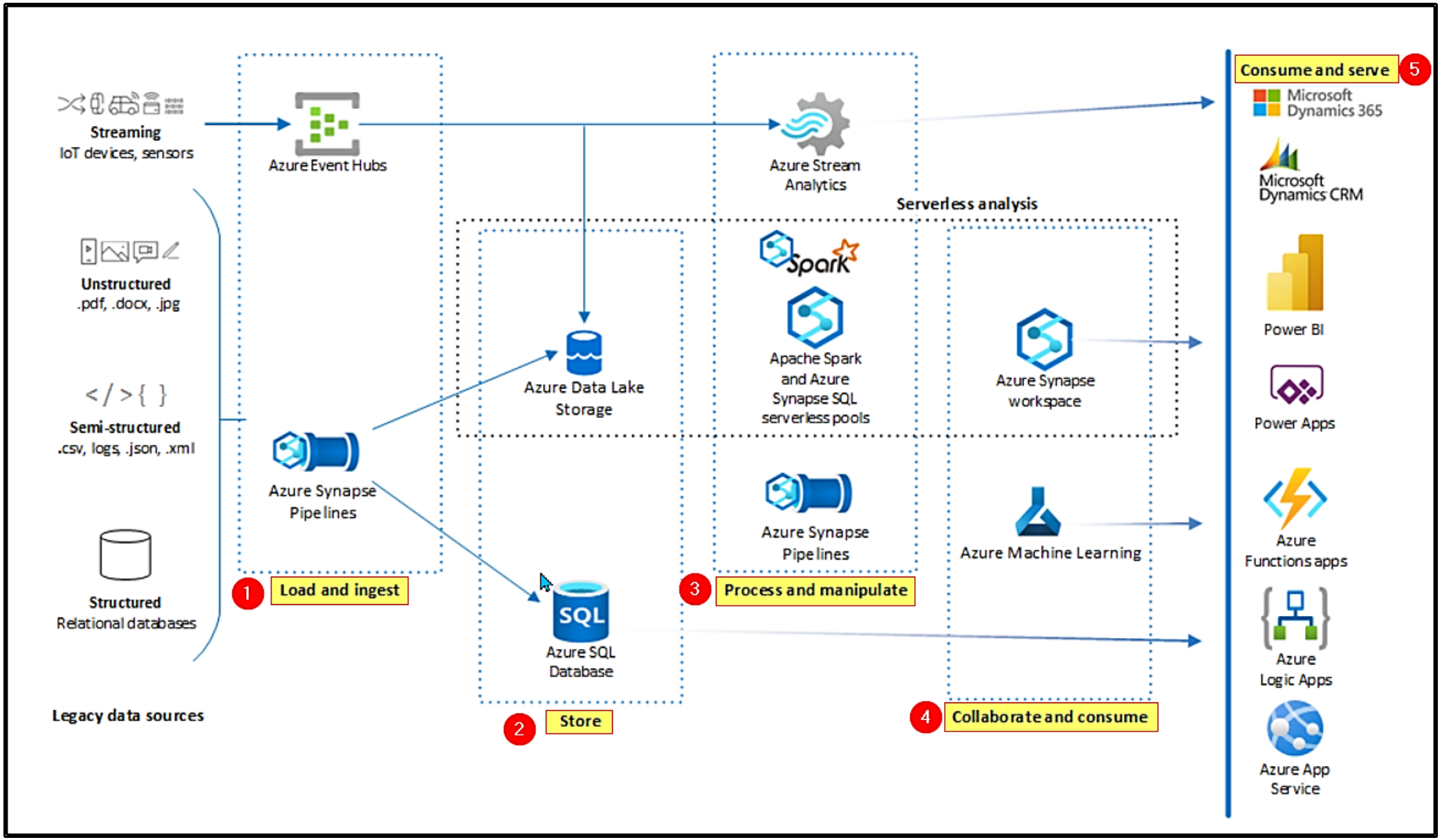
Integrate from On-Premise, PaaS, and SaaS

Batch and Real-time data integration

Secure hybrid connectivity

Code-free development environment

Batch vs streaming

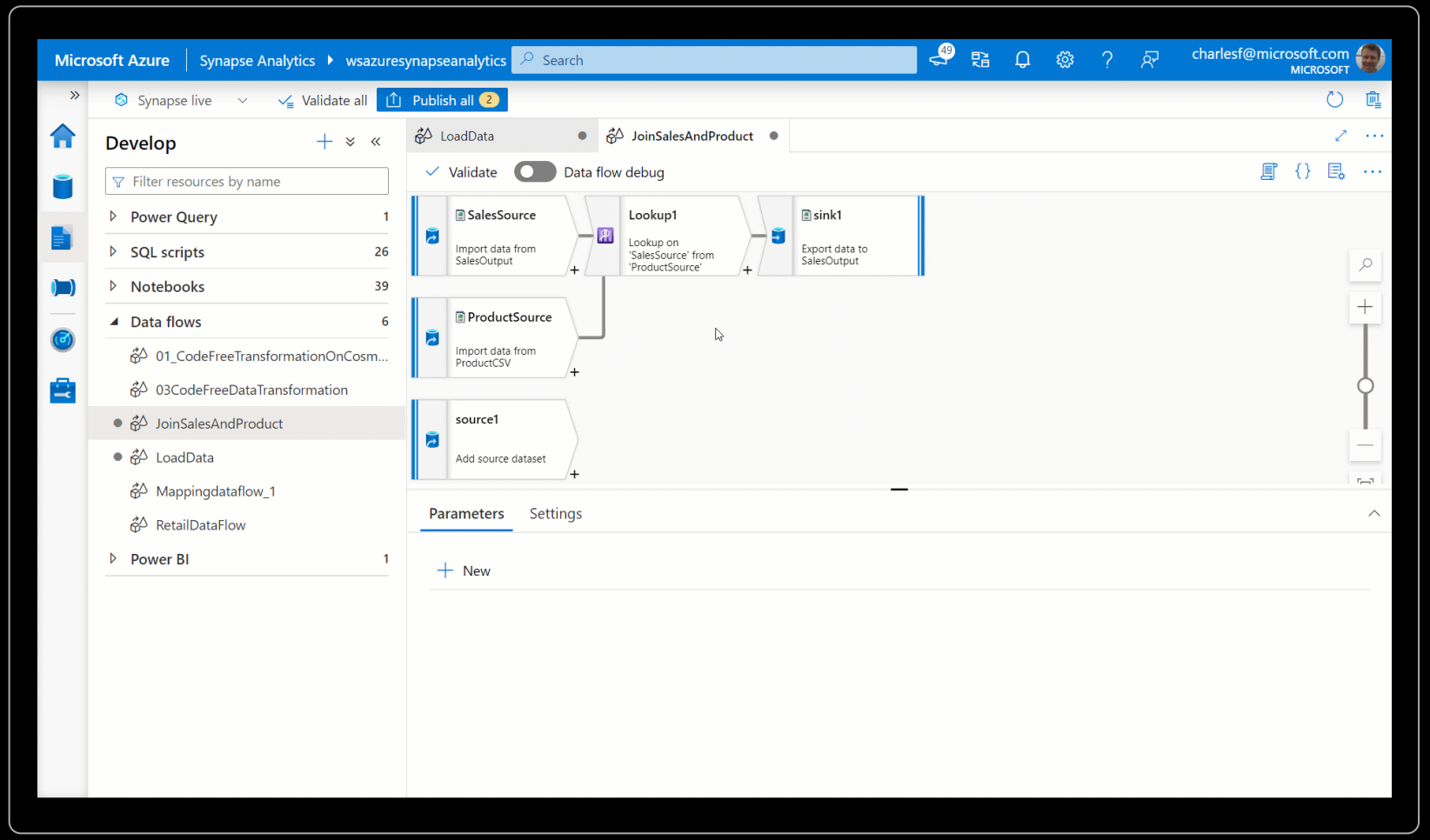


Generally Available

Code-free Data Flows

Enables developers to rapidly integrate data from a variety of sources

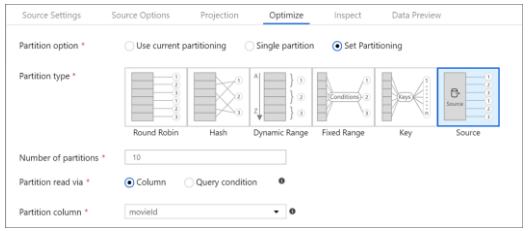
Execute on Spark for large scale processing



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)

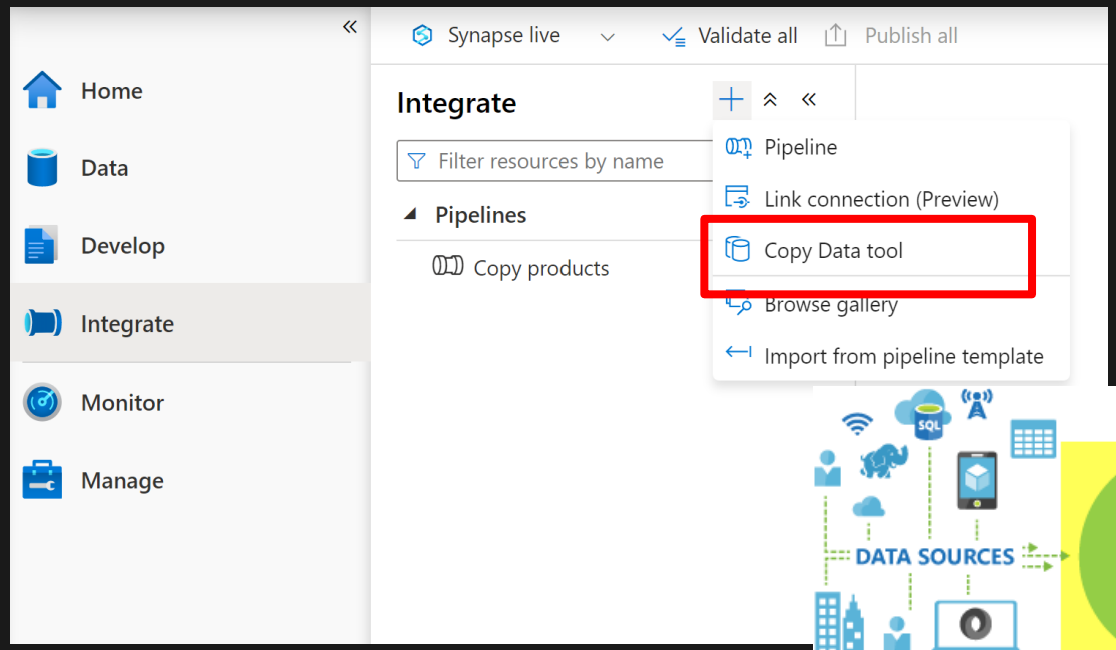
Commonly used ETL patterns(Sequence generator/Lookup transformation/SCD…)

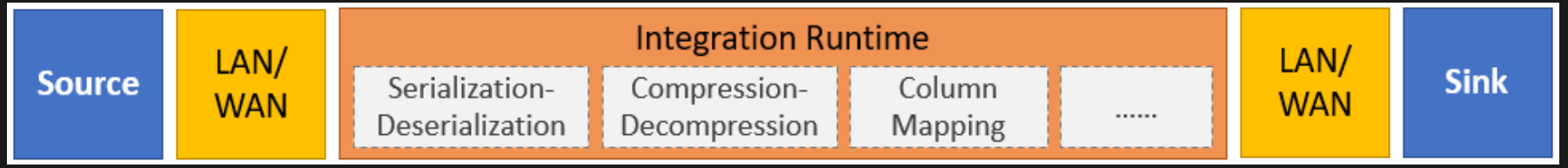
Data lineage – Capturing sink column lineage & impact analysis(invaluable if this is for enterprise deployment)

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

[Copy activity - Azure Data Factory & Azure Synapse | Microsoft Learn]

Copy data tool

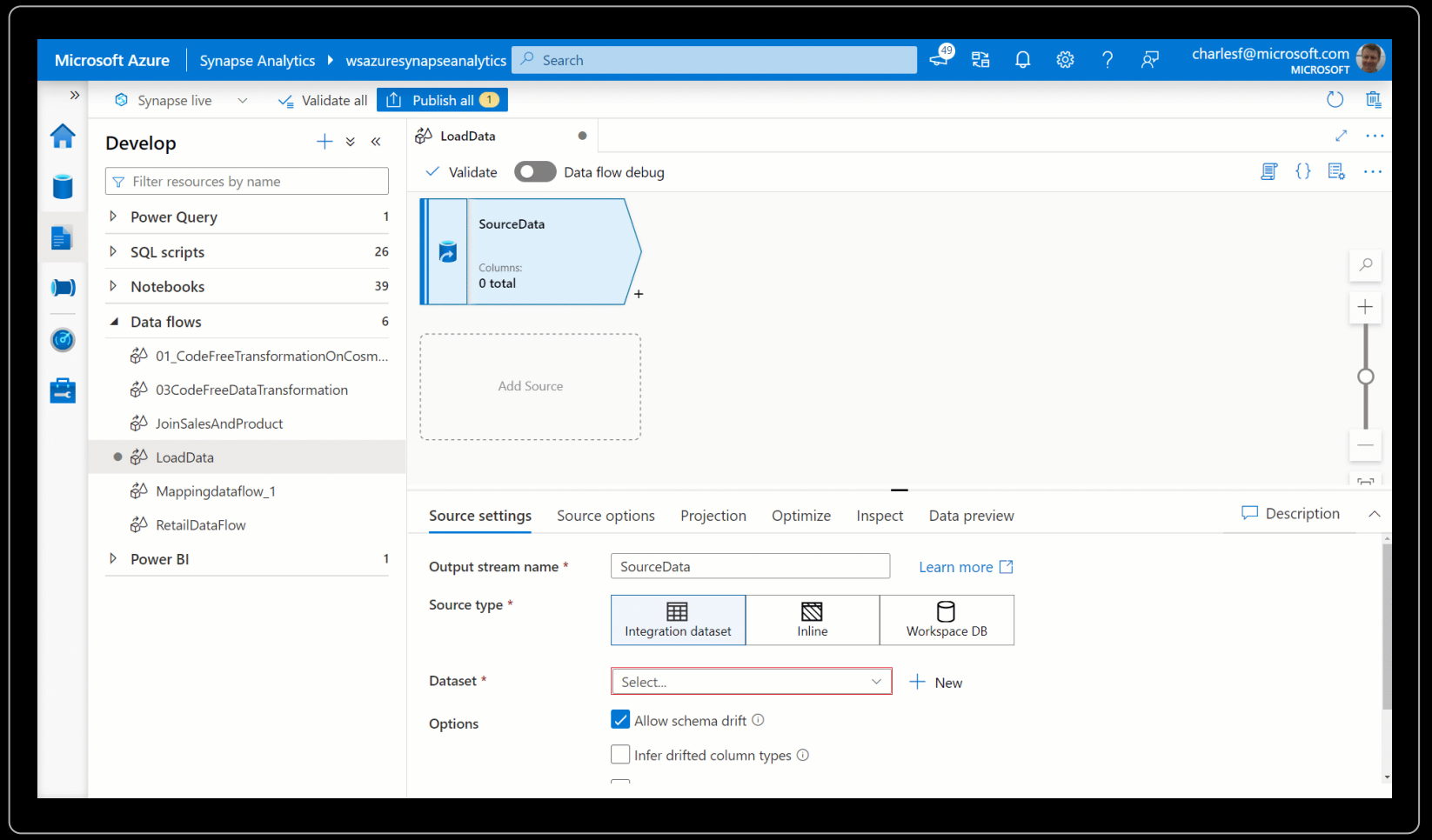




Generally Available

100+ Connectors

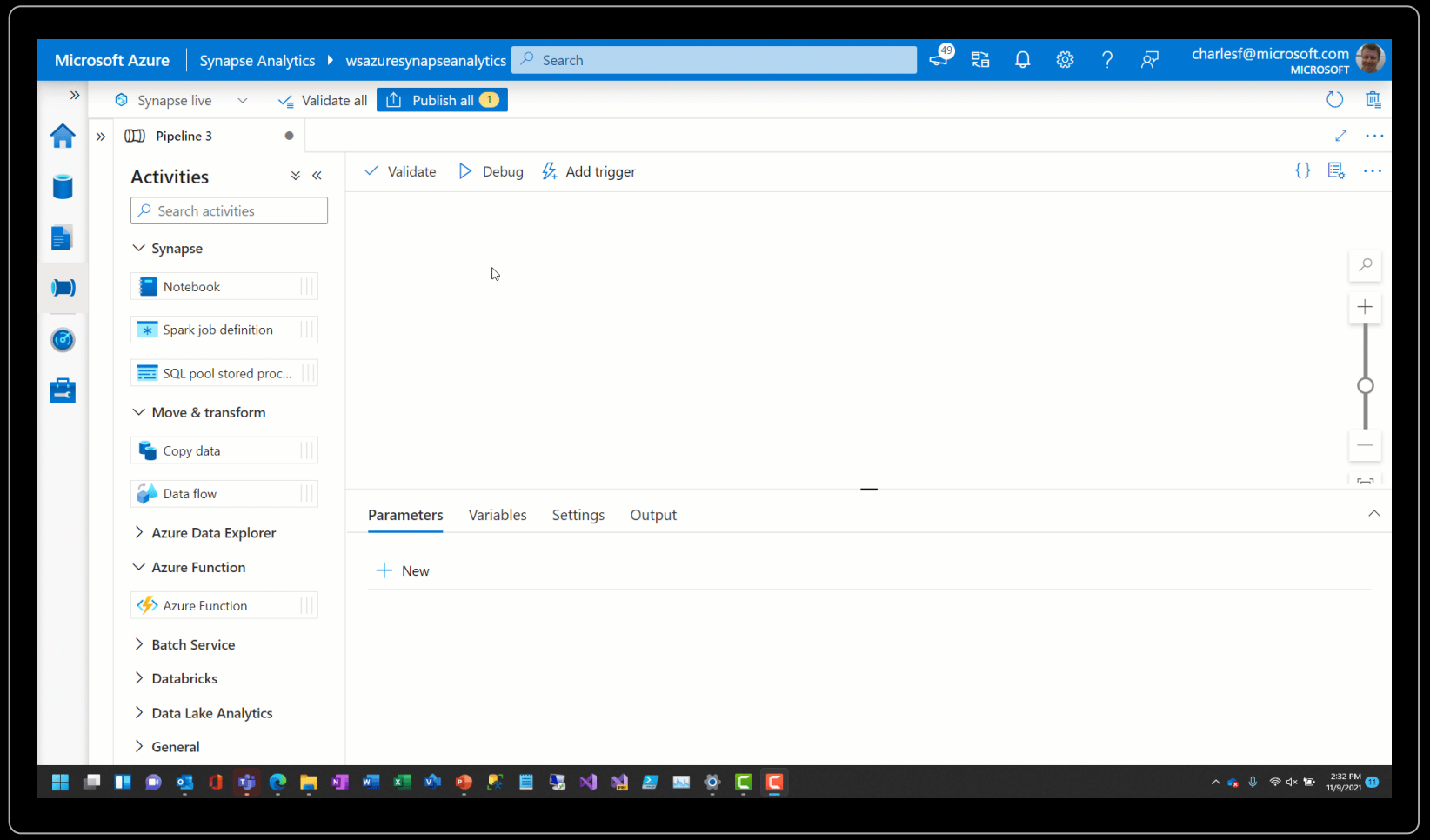
Connect on data sources in Azure, on-premise, other clouds, and SaaS applications



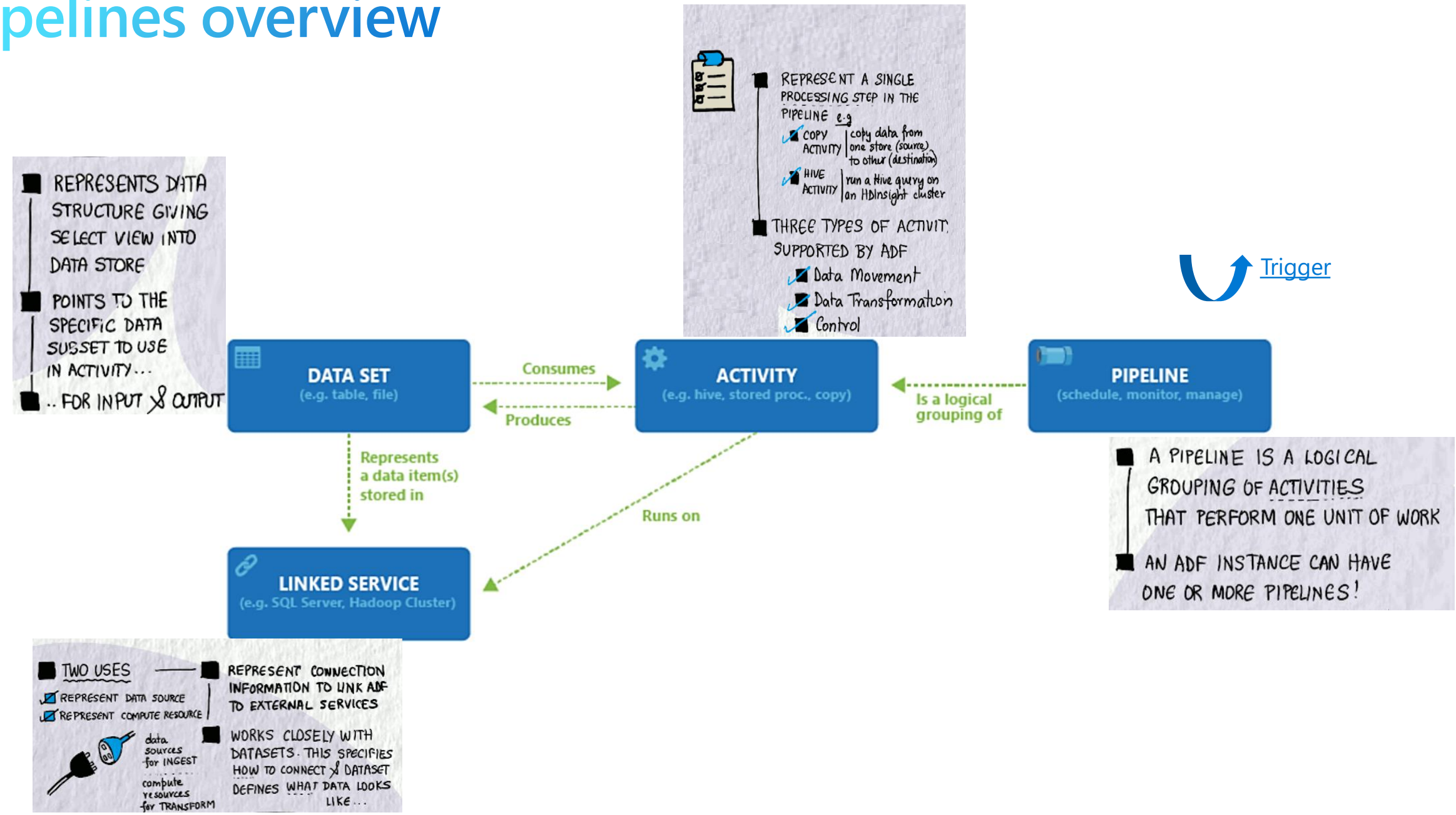
Generally Available

Azure Synapse Pipelines

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

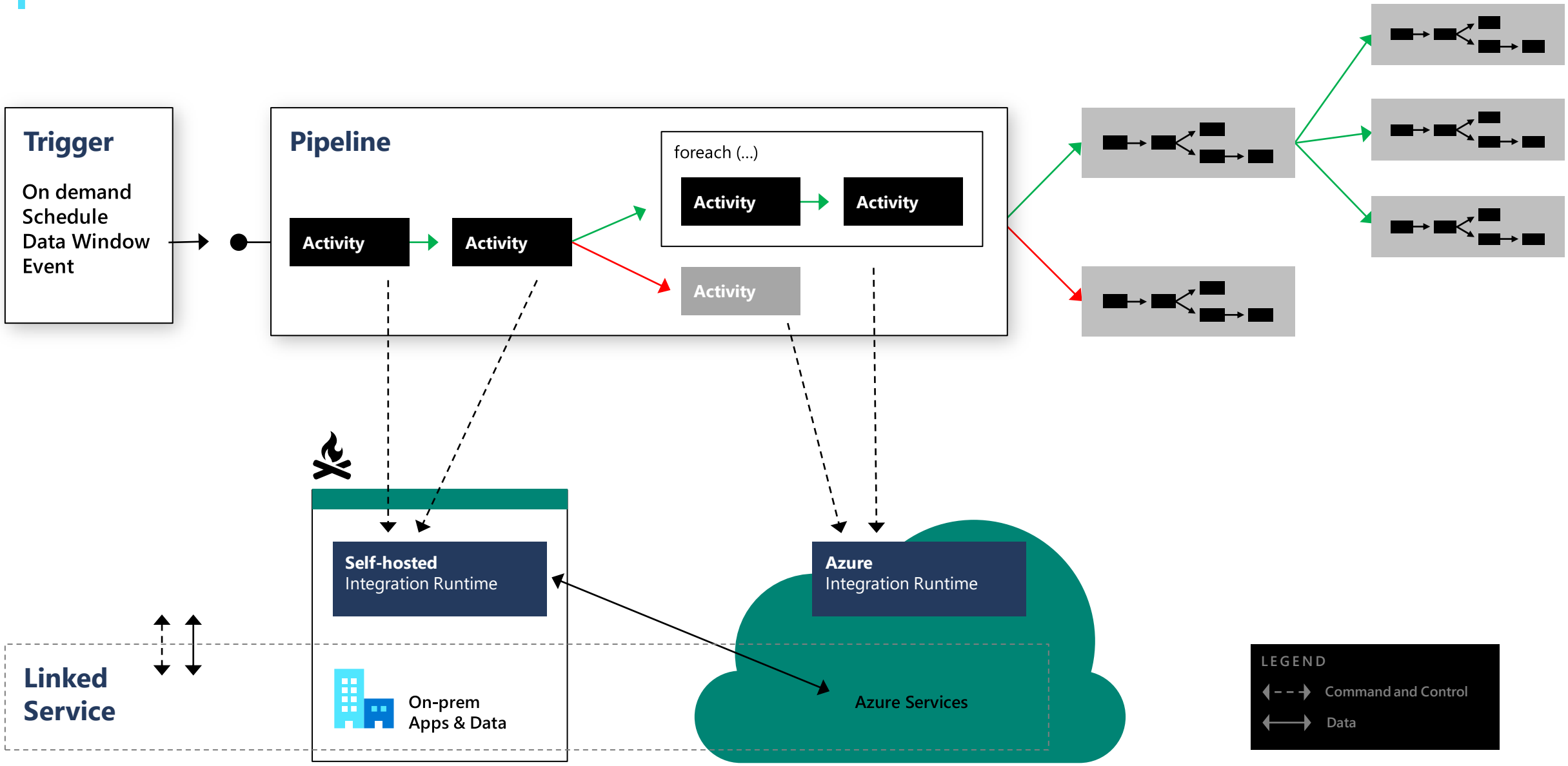


Pipelines overview



[Pipelines and activities - Azure Data Factory & Azure Synapse | Microsoft Learn]

Pipeline Orchestration

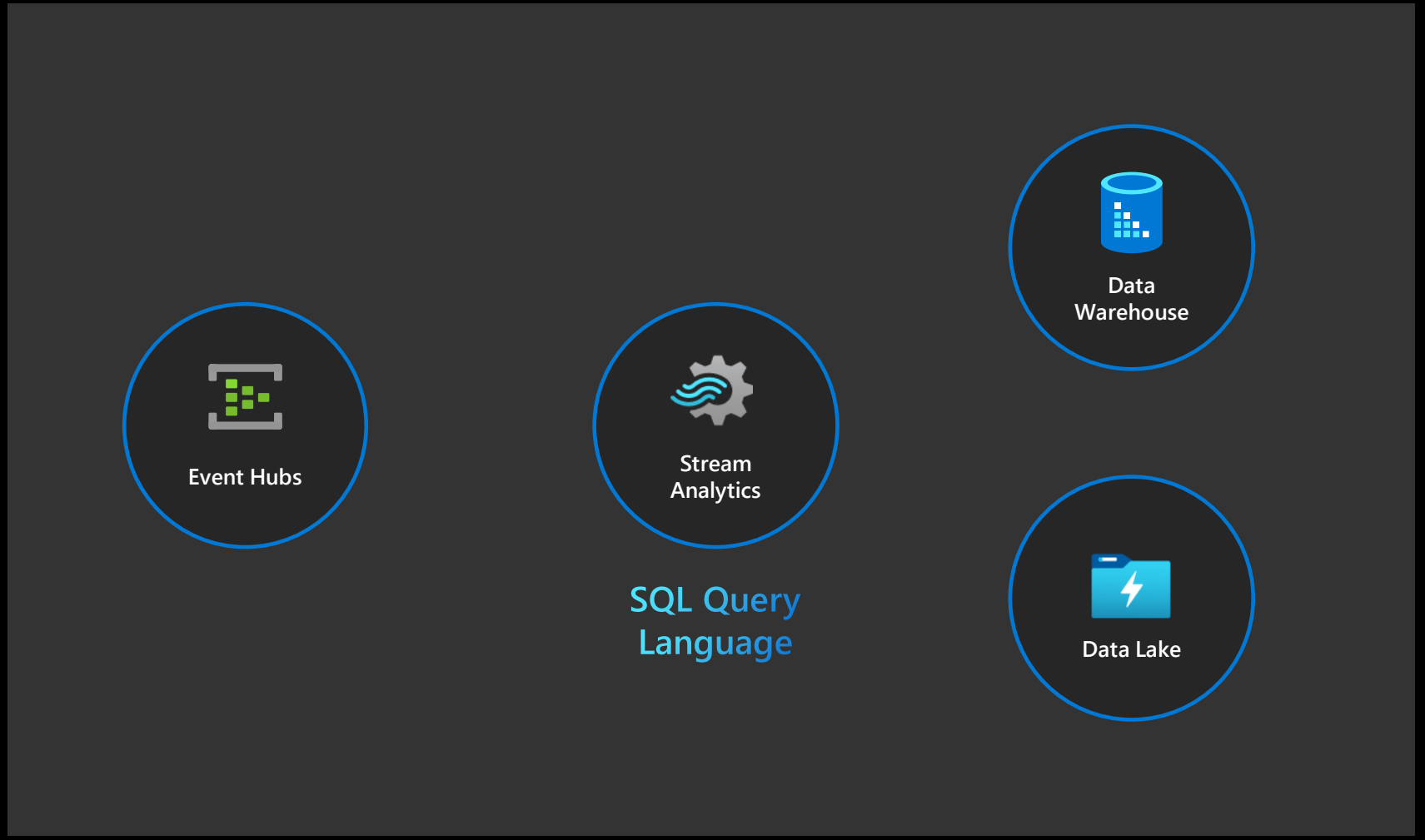


Generally Available

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



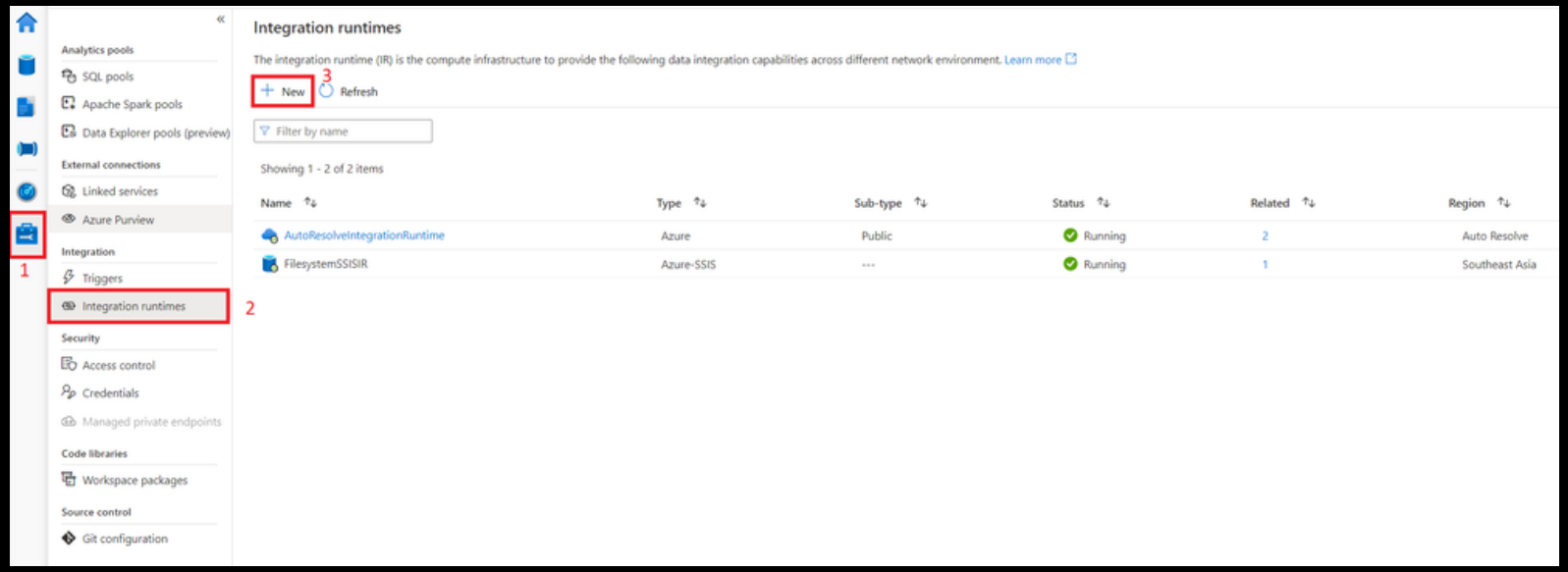
Public Preview

Q2 2022

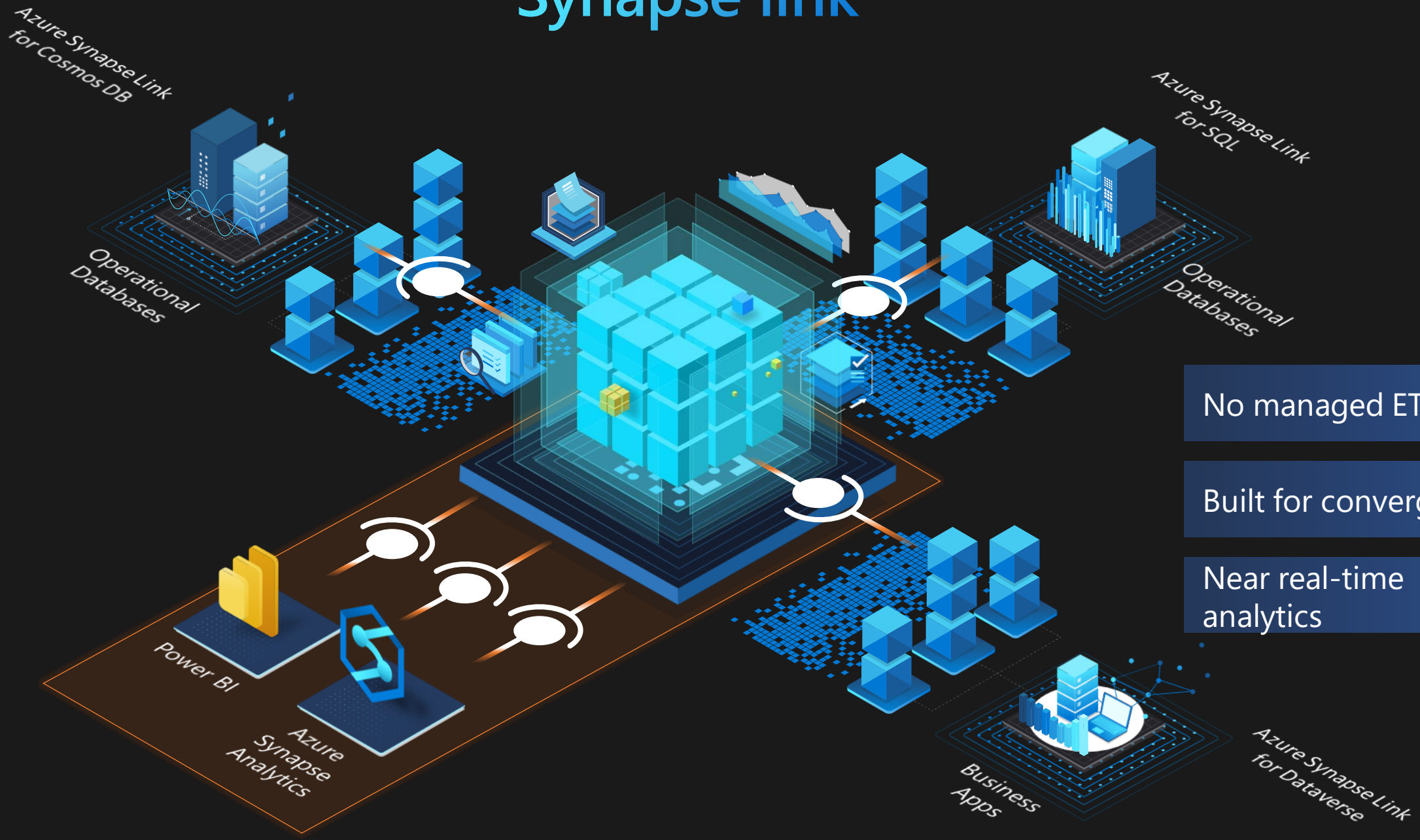
SSIS Integration Runtime for Azure Synapse

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

Running packages deployed into Azure Files (Package Deployment Model)



Synapse link

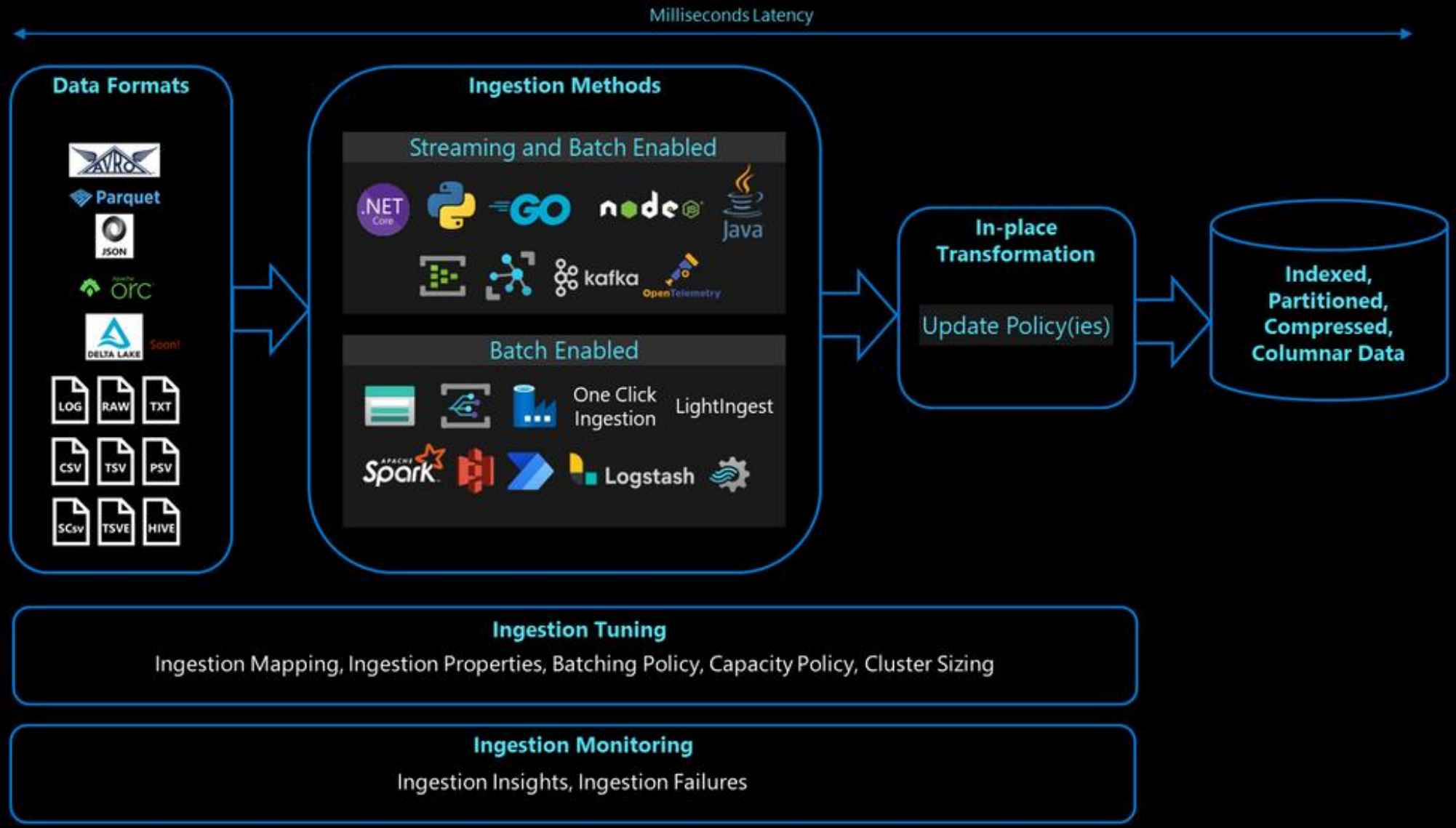


No managed ETL

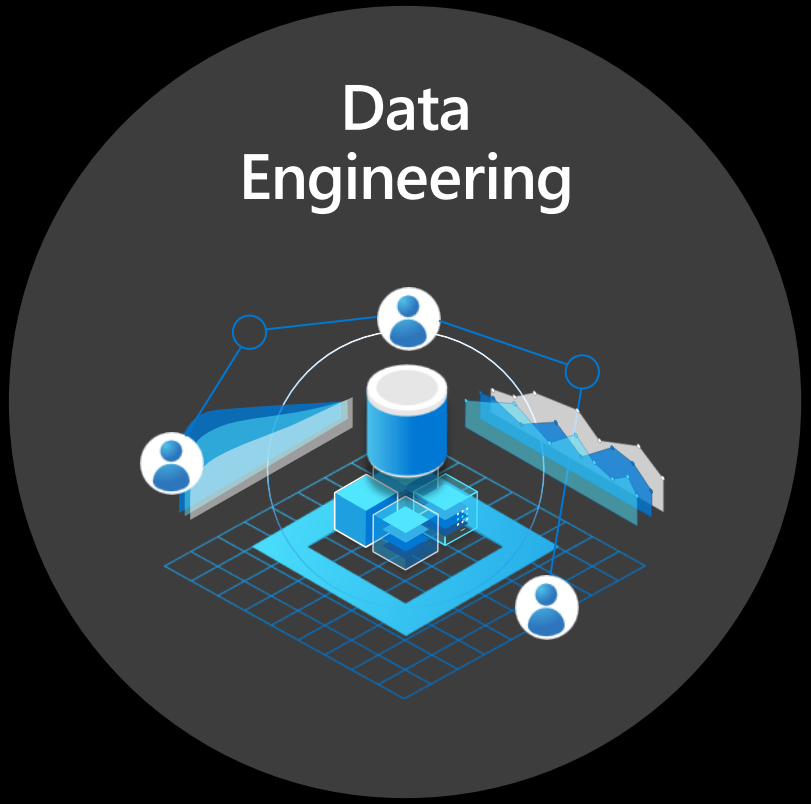
Built for convergence

Near real-time analytics

Azure Synapse data explorer - Preview



https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/demystifying-data-ingestion-in-azure-synapse-data-explorer/ba-p/3661133



Scalable Spark engine

Industry standard languages

Delta Lake Enabled

Azure DevOps integration

Data Engineering with Spark

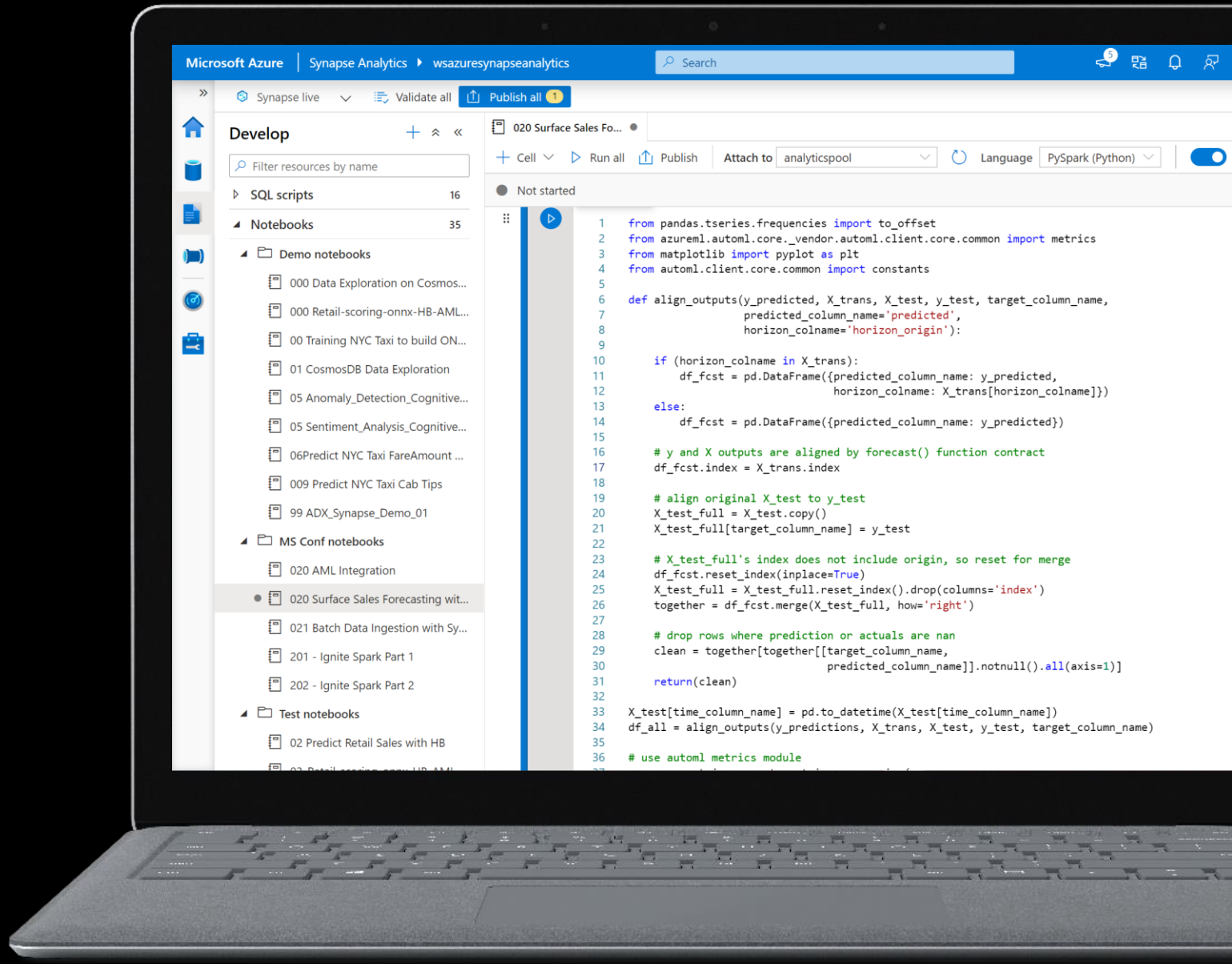
Code-first Data Engineering

PySpark, Scala, SQL and C# languages supported

Author multiple languages in a single notebook

Analyze & transform data from the data warehouse, data lake, and real-time operational data from one place

Synapse Spark uses Spark 3.2 runtime, which includes Delta Lake 1.0



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

What is Apache Spark?

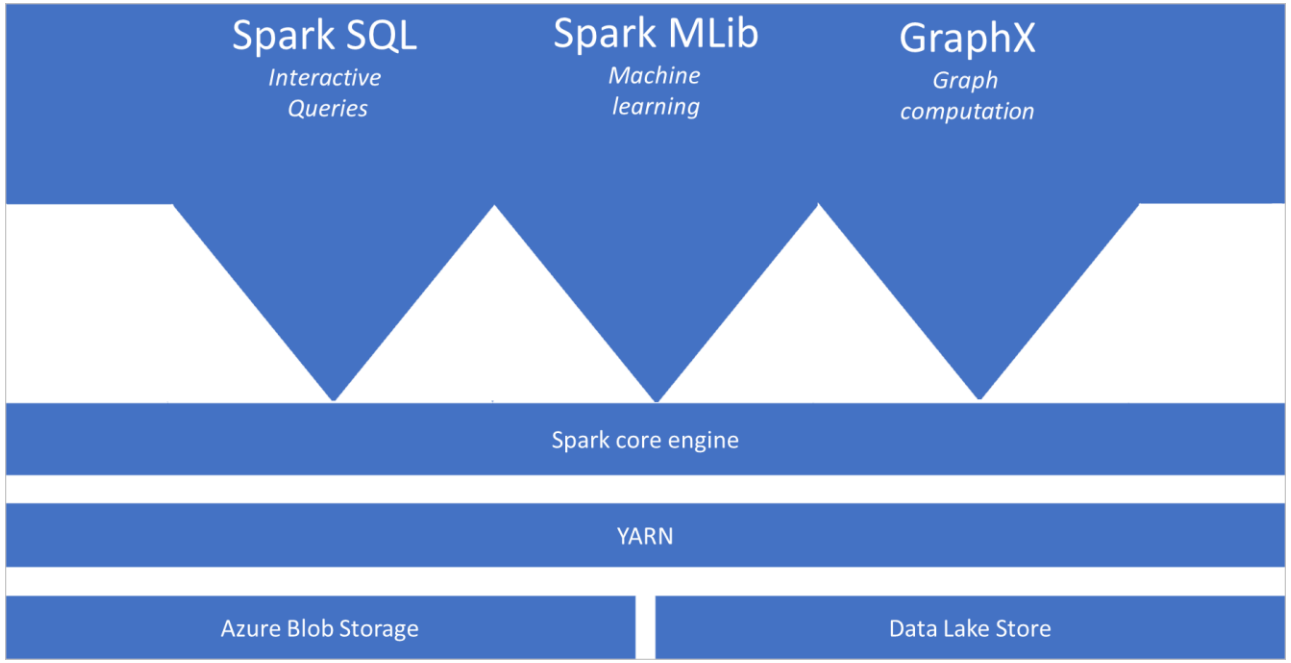
Parallel processing framework

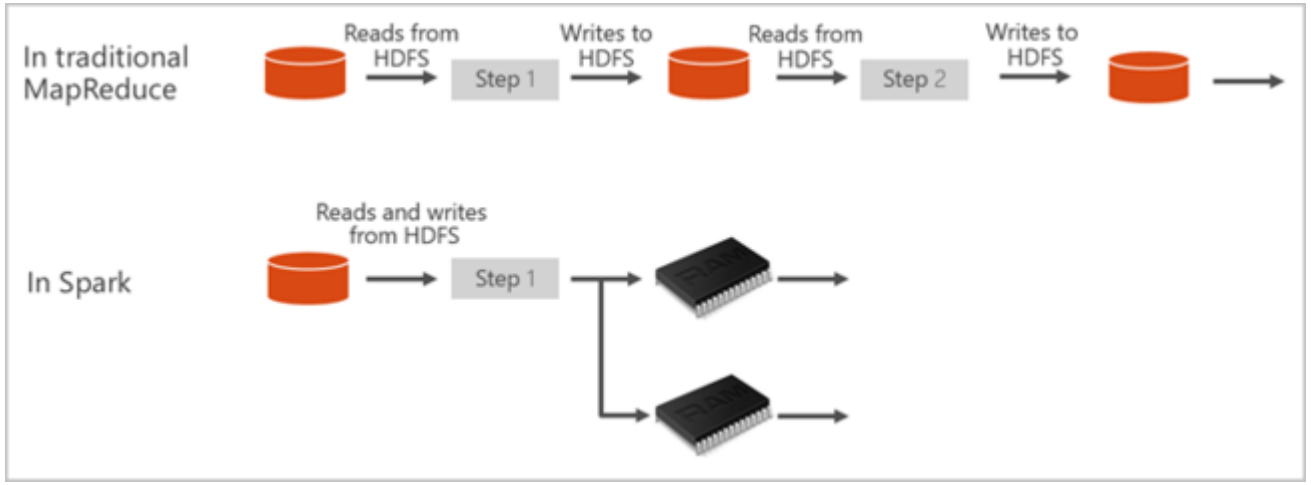
In-memory processing engine to boost the performance of big-data analytic applications

Much faster than disk-based applications

Integrates with multiple programming languages

Supports many workloads: Data Engineering, SQL, ML, Graphs., etc.





Motivation for Spark Pool in Synapse

Speed and efficiency	• Spark instances start in approximately 2 minutes for fewer than 60 nodes and approximately 5 minutes for more than 60 nodes. • The instance shuts down, by default, 5 minutes after the last job executed unless it is kept alive by a notebook connection.
Ease of creation	Quickly create a new Spark pool in Azure Synapse using: • Azure portal • Azure PowerShell • Synapse Analytics .NET SDK.
Ease of use	Synapse Analytics includes a custom notebook derived from Nteract. You can use these notebooks for interactive data processing and visualization.

Motivation for Spark Pool in Synapse

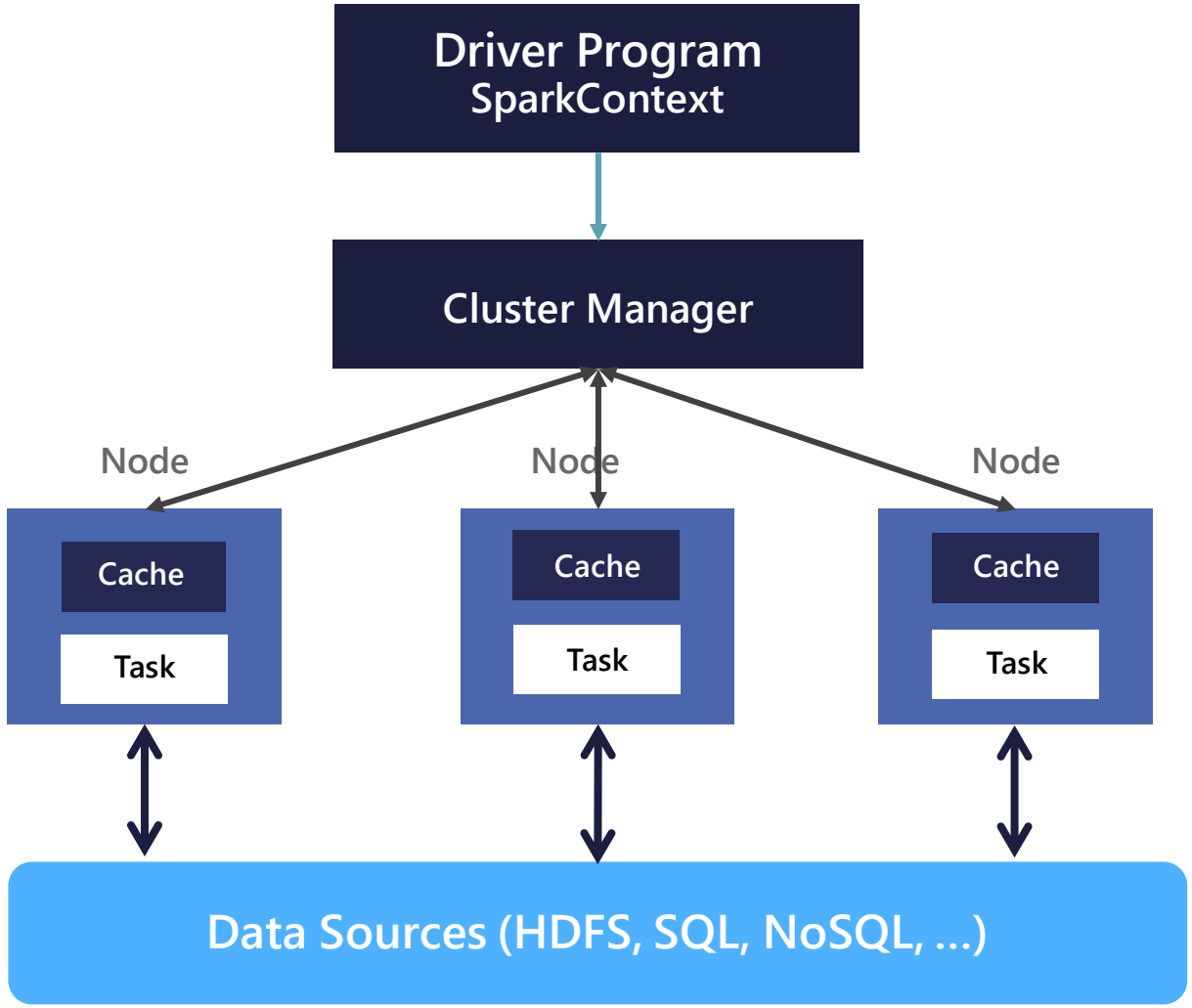
REST APIs	Spark in Azure Synapse Analytics includes Apache Livy, a REST API-based Spark job server to remotely submit and monitor jobs.
Support for Azure Data Lake Storage Generation 2	Spark pools in Azure Synapse can use: • Azure Data Lake Storage Generation 2 • BLOB storage.
Integration with third-party IDEs	Azure Synapse provides an IDE plugin for JetBrains' IntelliJ IDEA that is useful to create and submit applications to a Spark pool.

Motivation for Spark Pool in Synapse

Pre-loaded Anaconda libraries	• Spark pools in Azure Synapse come with Anaconda libraries pre-installed. • Providing close to 200 libraries for machine learning, data analysis, visualization, etc.
Scalability	• Can have Auto-Scale enabled, so that pools scale by adding or removing nodes as needed. • Spark pools can be shut down with no loss of data since all the data is stored in Azure Storage or Data Lake Storage.

General Spark Cluster Architecture

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (like Azure).
- These resources come in the form of *executors* on the worker nodes (logical blocks of CPU/memory). These executors are what read and write data for your application.
- The *SparkContext* breaks down the application code into *tasks* that are operated on in parallel by the executors.



Spark Pool - Configuration

- **Nodes**
  - Apache Spark pool instance consists of one head node and two or more worker nodes with a minimum of three nodes in a Spark instance.
  - The head node runs additional management services such as Livy, Yarn Resource Manager, Zookeeper, and the Spark driver.
  - All nodes run services such as Node Agent and Yarn Node Manager.
  - All worker nodes run the Spark Executor service.

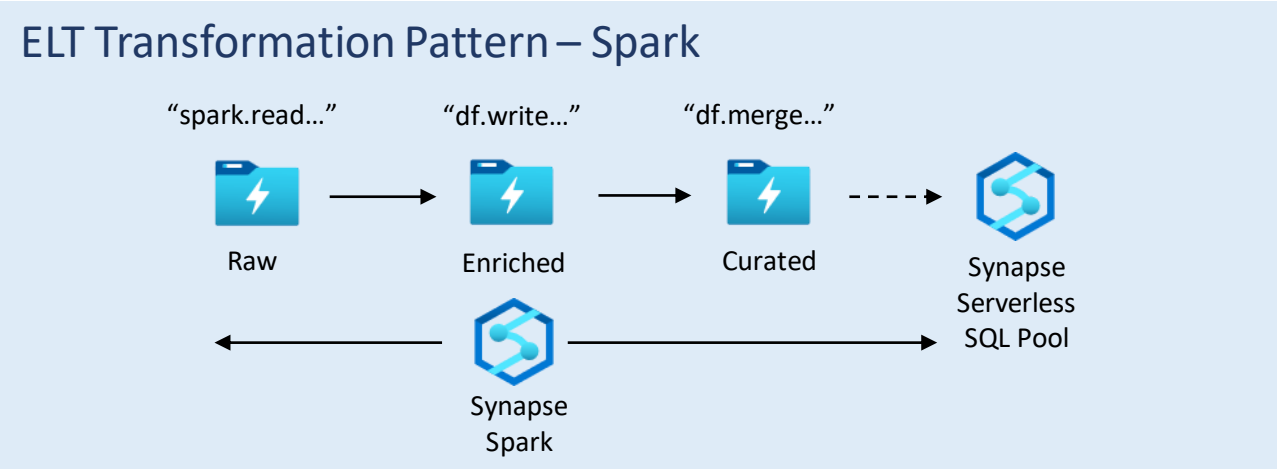
Spark Pool - Configuration

- **Node Sizes**
  - Node sizes can be altered after pool creation, instance may need to be restarted.

Size	vCore	Memory
Small	4	32 GB
Medium	8	64 GB
Large	16	128 GB
XLarge	32	256 GB
XXLarge	64	432 GB
XXX Large (Isolated Compute)	80	504 GB

Process & Transform (Synapse Spark)

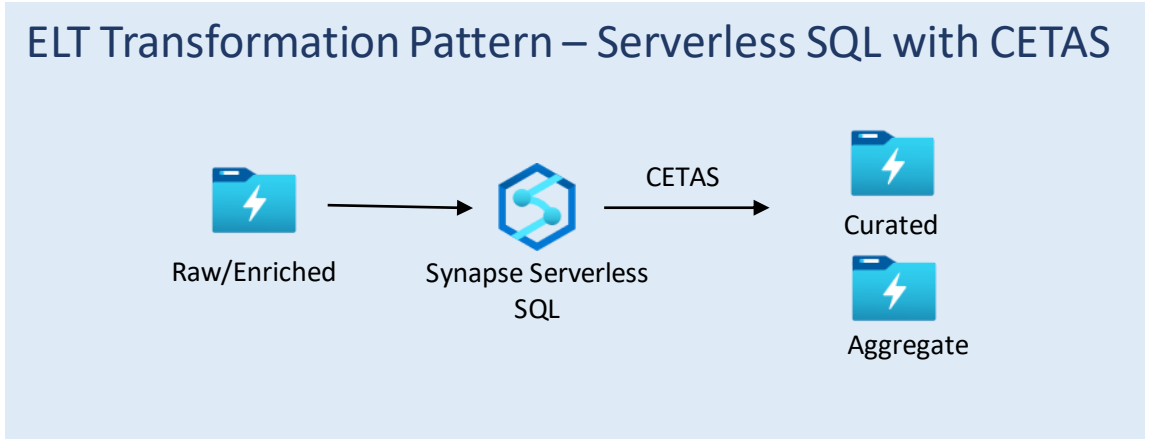
- Code-centric way of doing ETL/ELT
- Based on Synapse Spark notebooks
- Can use SQL, Scala, Python, C#
- Read raw files into dataframe
- Apply transforms/cleaning
- Write/Save back to data lake Delta files (parquet + delta_log)
- Apply DELTA optimisations (compact, partition) via code
- Parameterise notebooks and add to pipelines for orchestration
- Use **Delta** format for enriched/curated zones




```
4   from pyspark.sql.functions import col,when
5
6   #df = spark.read.load('abfss://data@techinsideradls.dfs.core.windows.net/raw/call_center/*.dat', format
7   df= spark.read.format("csv").schema(schema).load('abfss://data@techinsideradls.dfs.core.windows.net/raw
8
9
10  #replace nulls
11  df = df.na.fill(0)
12  #replace null strings
13  df = df.na.fill("Unknown")
14  df = df.distinct()
15
16  #write to enriched
17
18  df.write.mode('overwrite').format("parquet").partitionBy("cc_state").save('abfss://data@techinsideradls
19
20  #create view for merge
21  df.createOrReplaceTempView("vcallcentre")
22
23
```

Process & Transform (Serverless SQL)

- Create External Table as Select (CETAS)
- Creates new persisted table from select statement
- Good for commonly used aggregations & summary tables
- Outputted dataset is stored in data lake and referenced as external table
- Use Parquet (Delta not currently available) as storage format




```
CREATE EXTERNAL TABLE summary_metrics
WITH (
    LOCATION = 'aggregated_data/',
    DATA_SOURCE = CatalogSalesds,
    FILE_FORMAT = ParquetLakeFormat
)
AS
select top 100
  count(distinct cs_order_number) as "order count"
  ,sum(cs_ext_ship_cost) as "total shipping cost"
  ,sum(cs_net_profit) as "total net profit"
from
  catalog_sales cs1
  ,date_dim
  ,customer_address
  ,call_centre
where
    d_date between '1999-3-01' and
        dateadd(day, 60, cast('1999-3-01' as date))
and cs1.cs_ship_date_sk = d_date_sk
and cs1.cs_ship_addr_sk = ca_address_sk
and ca_state = 'TX'
and cs1.cs_call_center_sk = cc_call_center_sk
--and cc_county in ('Jefferson Davis Parish','Gage County','Sierra County','Pennington County',
  --                'Wadena County'
```

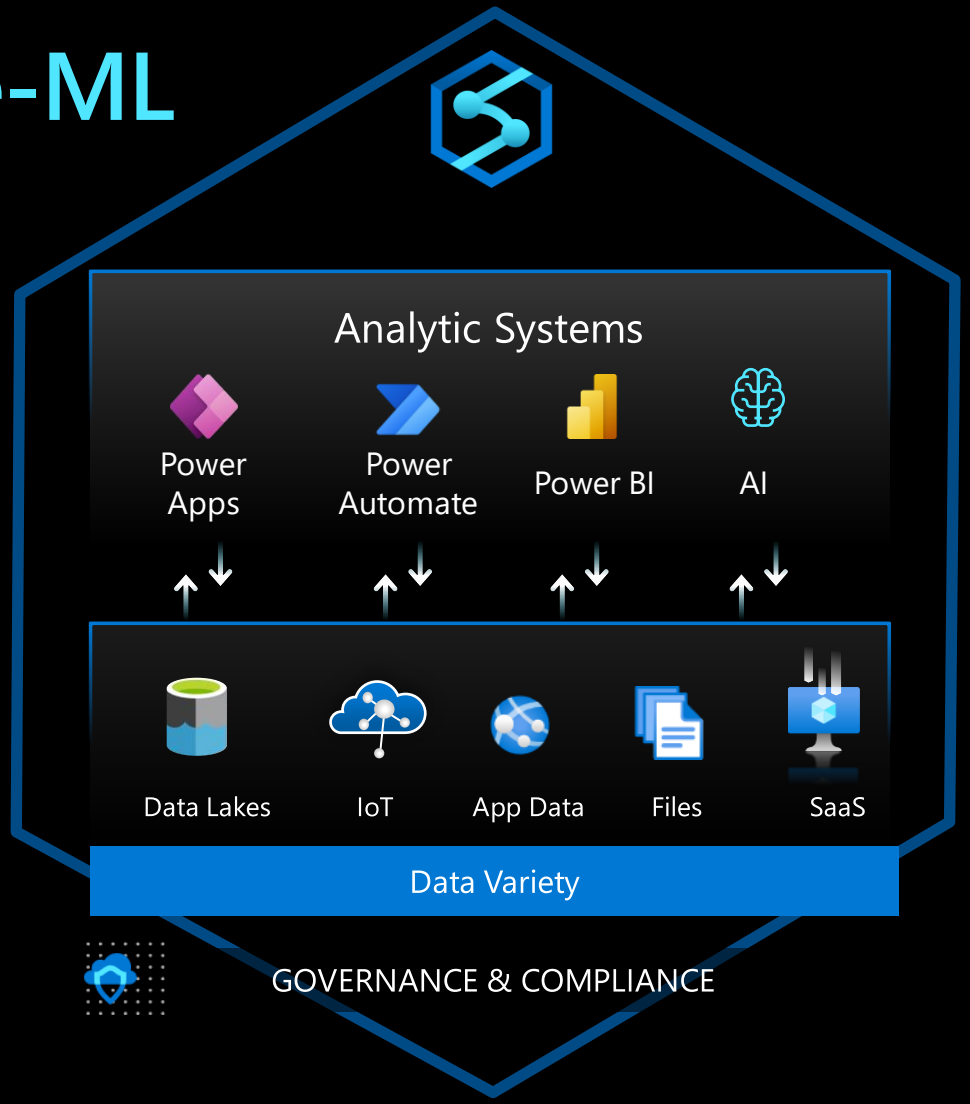
Demo :
Data exploration with Spark on Synapse

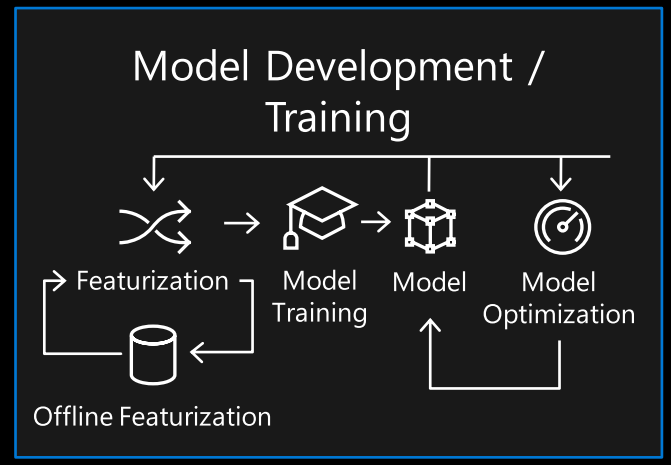
Modeling capabilities

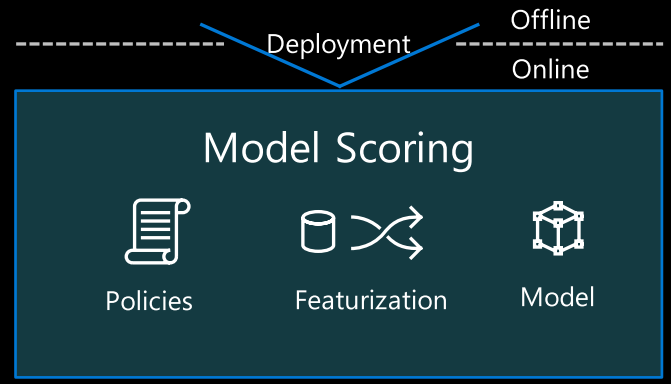
Why Synapse
Enterprise Grade-ML

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







Machine learning

Democratize predictive power

- Notebooks provide code authoring experiences
- Notebooks provides a code authoring experience for complex predictive models
- Automatic ML graphical interface provides a no-code experience for creating ML models
- Native integration with Azure Cognitive Search provides access to pre-built models

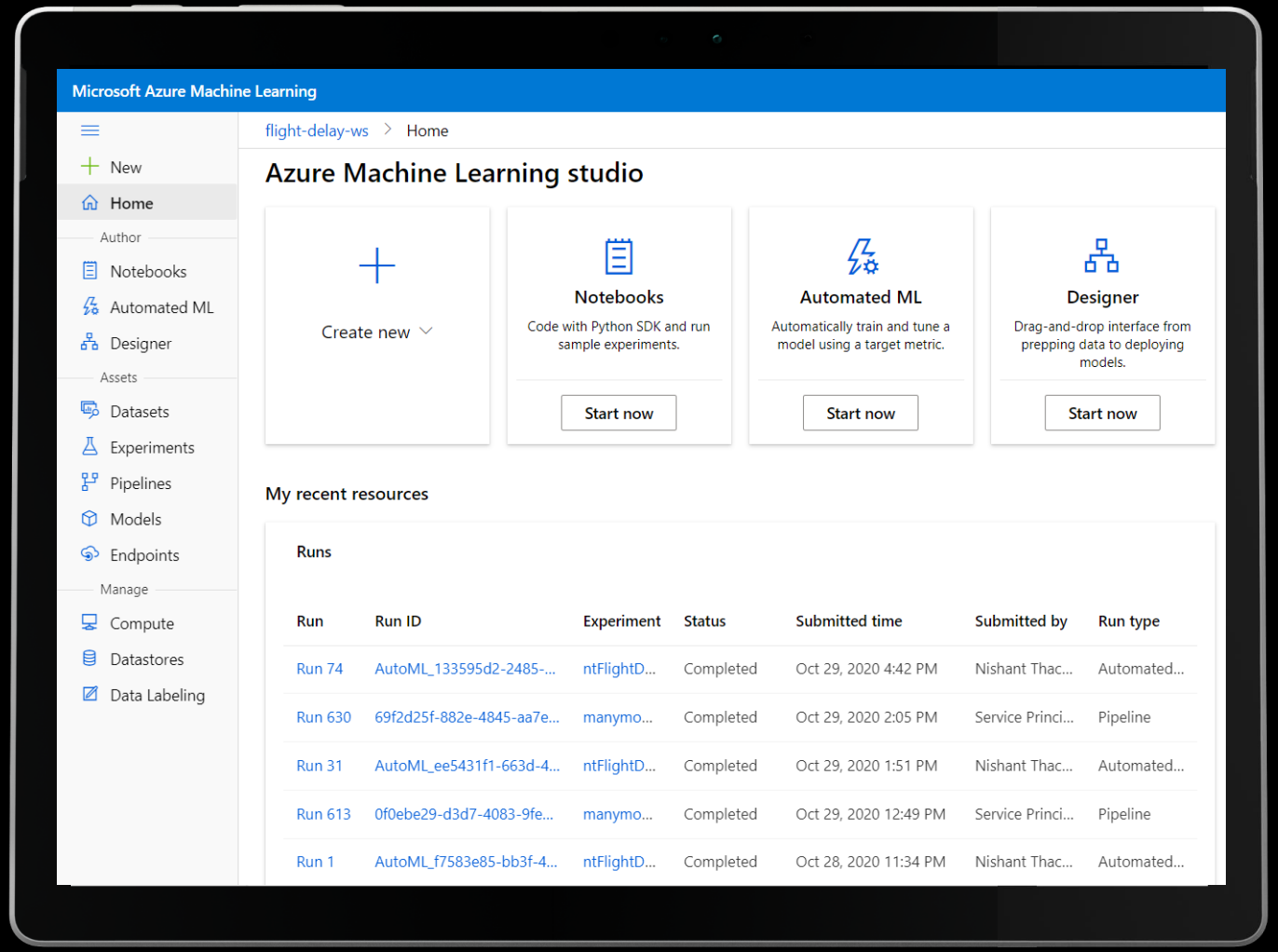
Synapse & Azure ML

Azure Machine Learning

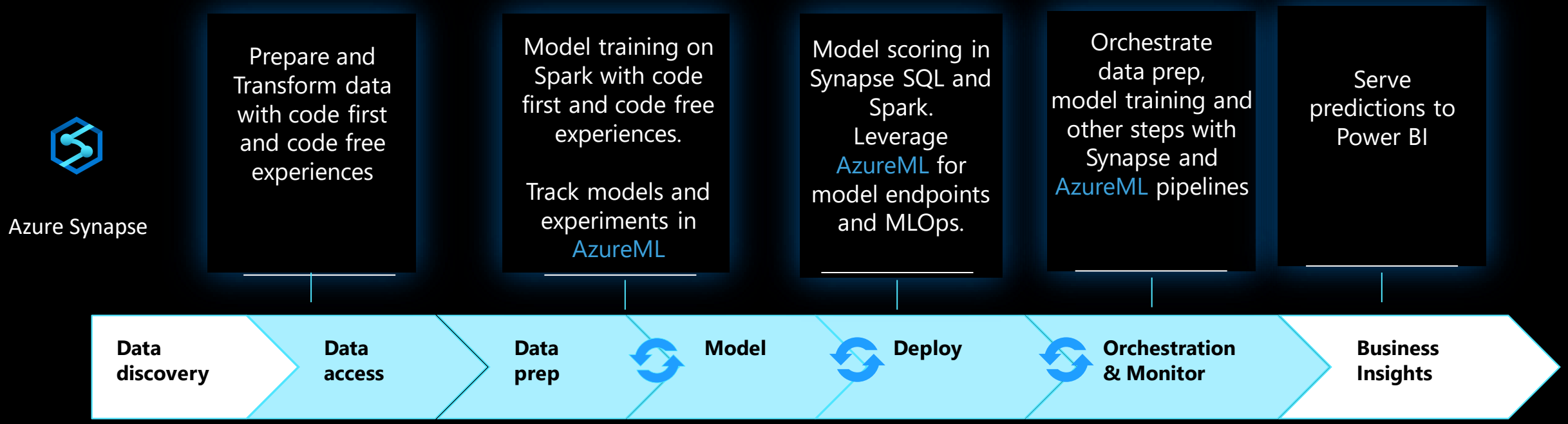
The one central hub for your data science team

Boosted collaboration

Integration with other Azure services



Synapse & Azure ML: Supporting the full Data & AI lifecycle



Data wrangling with Apache Spark pools (preview) - Azure Machine Learning | Microsoft Learn

MachineLearningNotebooks/how-to-use-azureml/azure-synapse at master · Azure/MachineLearningNotebooks (github.com)

Public Preview

Demo: Using Synapse Spark pools in AML

1. Link Synapse + AML workspaces
2. Attach Synapse Spark pool as a compute
3. Grant "Synapse Apache Spark Administrator" role of the synapse workspace to the generated MSI.
4. Use SynapseSparkStep in AML pipeline

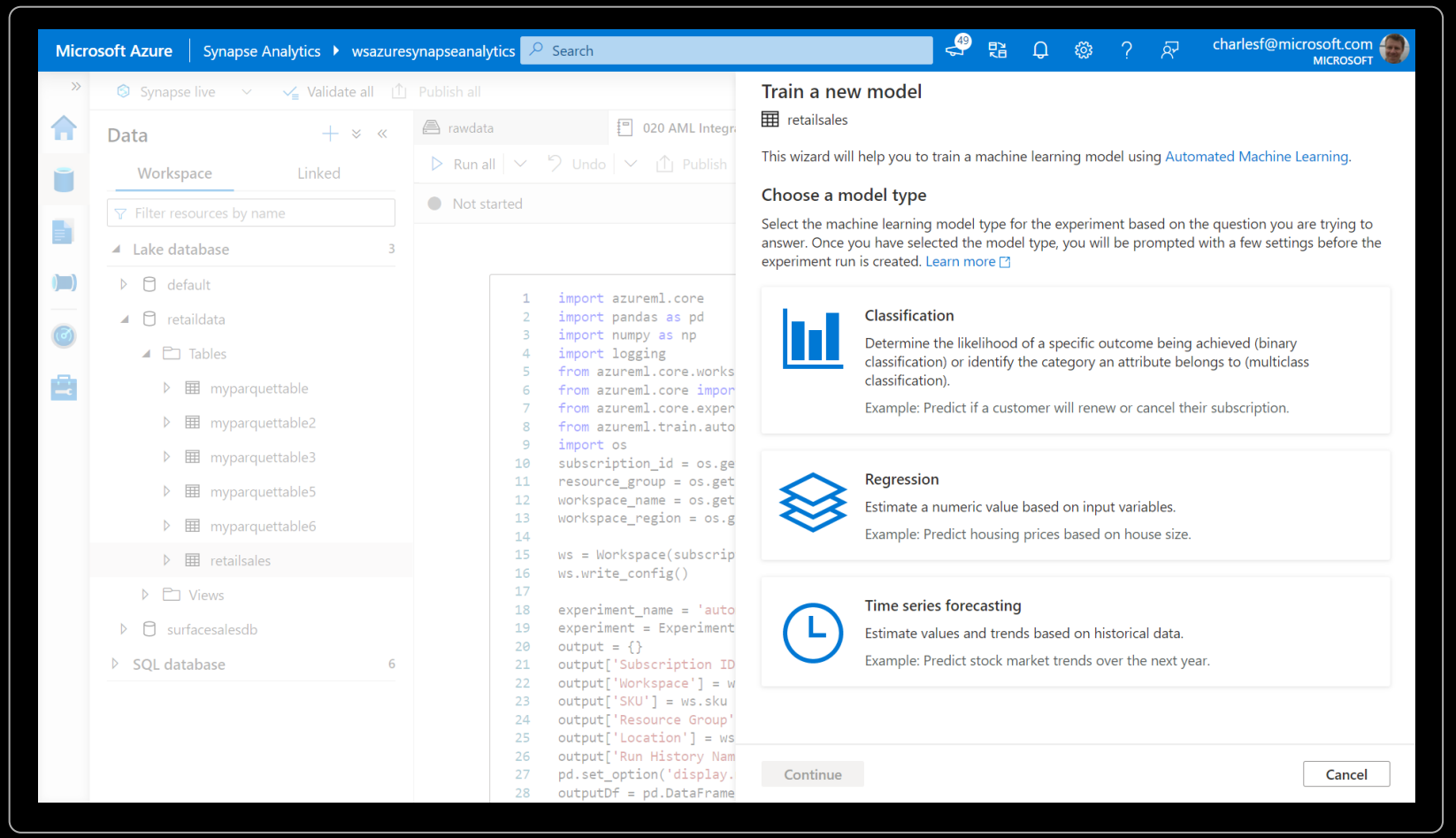
Create a linked service with Synapse and Azure Machine Learning workspaces (preview) - Azure Machine Learning | Microsoft Learn

Use Apache Spark in a machine learning pipeline (preview) - Azure Machine Learning | Microsoft Learn

Public Preview

Automated Machine Learning

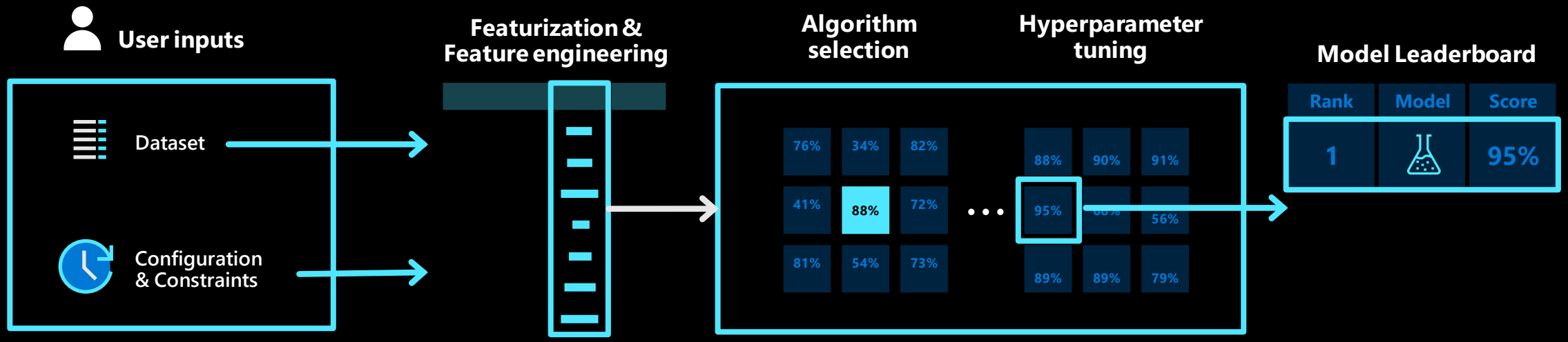
Code first + No-code training for ML models empowers everyone with data science



What is Automated ML?

Automated machine learning (automated ML) automates feature engineering, algorithm and hyperparameter selection to find the 'best model' for your data.

Loop until reaching **exit criteria**



Demo :
Synapse + Automated machine learning

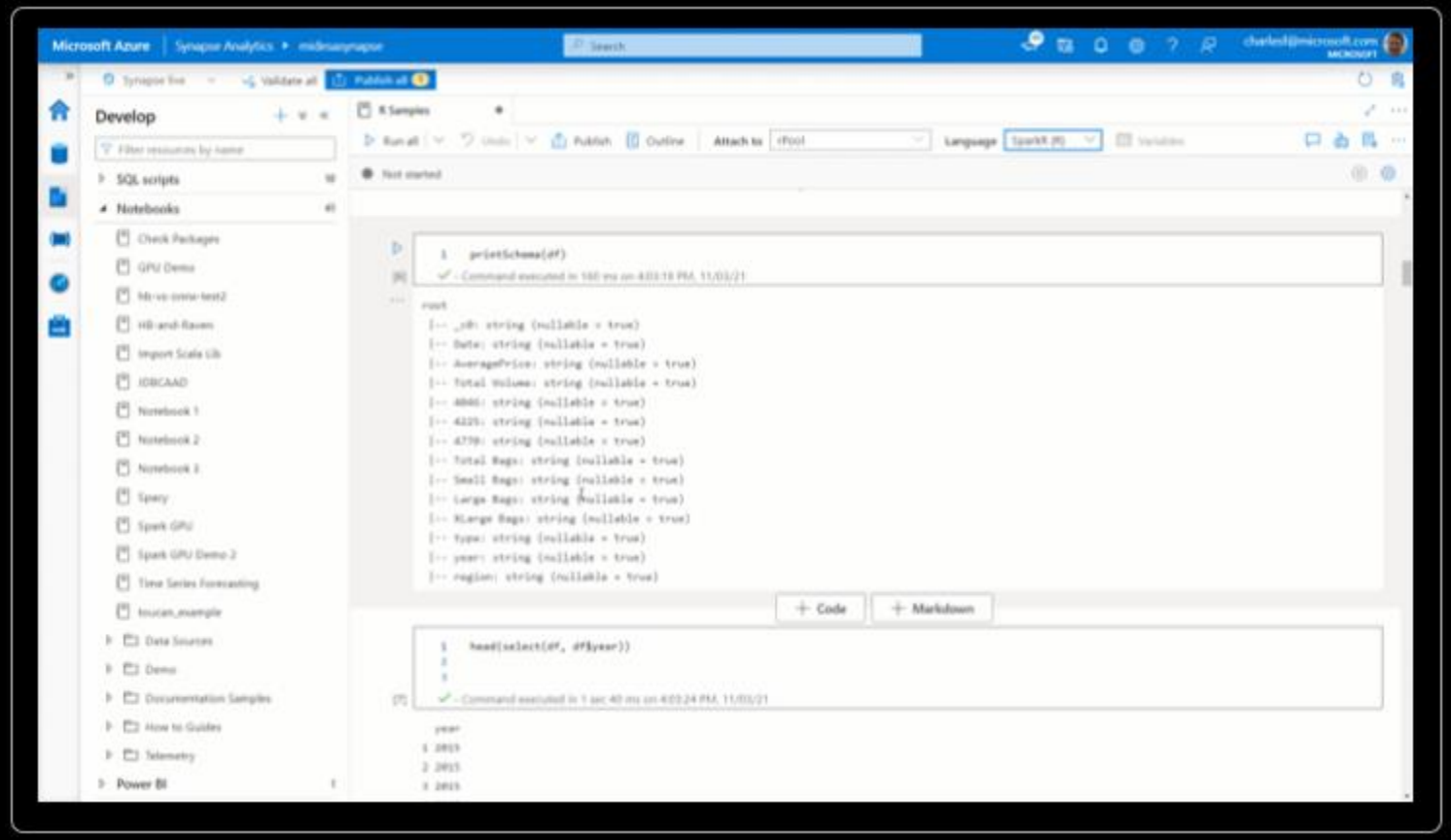
1. Create Synapse Spark pool
2. Create NYC taxi spark table
3. Link Synapse + AML workspaces
4. Use AutoML wizard to train the model

Private Preview

Q2 2022

R Language Support

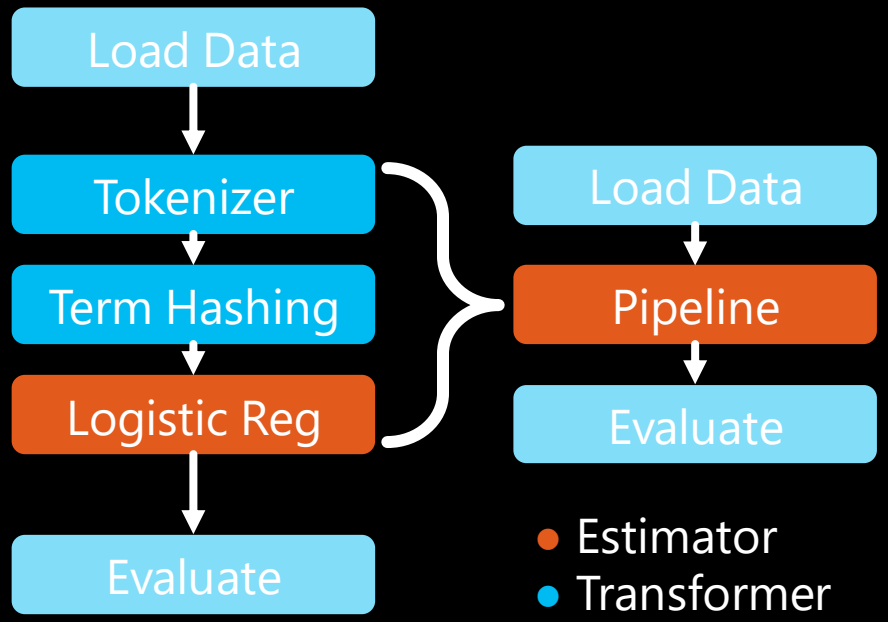
Enables data scientists to apply the industry standard R language to developing ML models



Spark ML

- High level library for distributed machine learning
- Inspired by scikit-learn
- All models have a uniform interface
  - Compose models into pipelines
  - Save, load, and transport models

```
data = spark.read.csv("hdfs://…")
train, test = data.randomSplit([.5,.5])
model = LogisticRegression().fit(train)
predictions = model.transform(test)
```



APACHE Spark ML

Demo:
Train a model + Spark MLlib

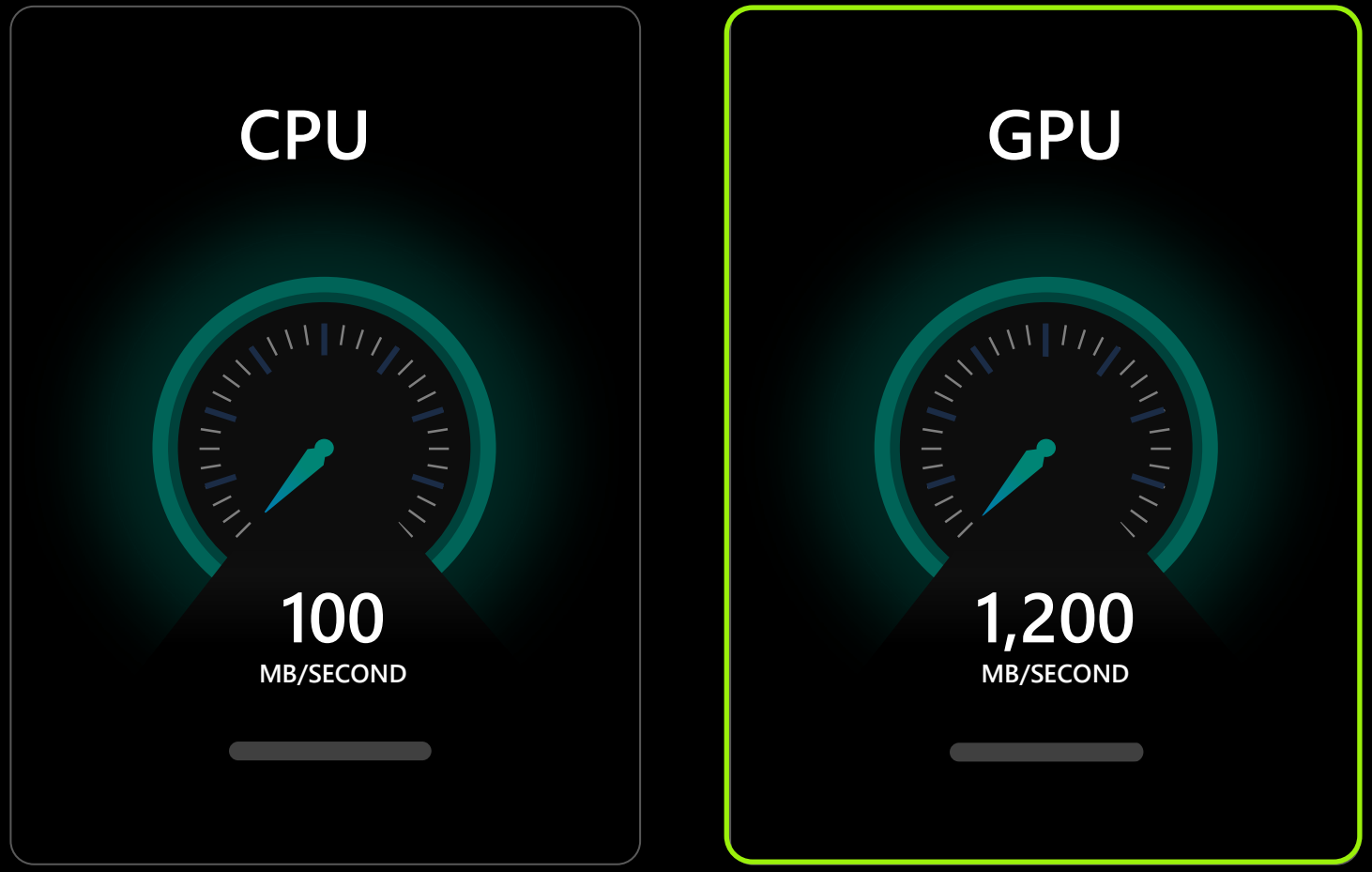
Train a logistic regression model for flight delays

Prediction on Synapse Spark pool using Spark MLlib

Public Preview

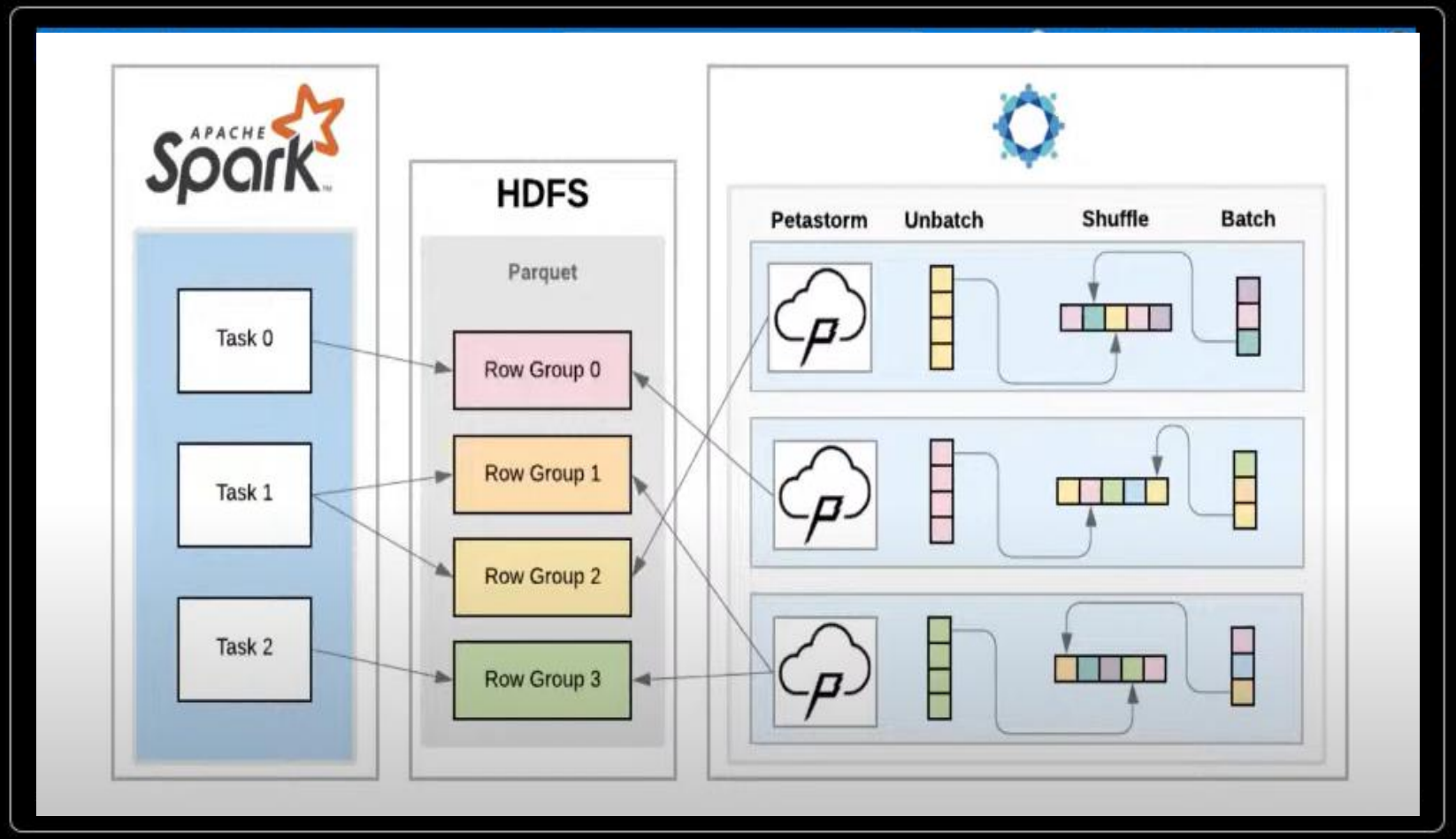
GPU Accelerated Workloads

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



Public Preview

Distributed Deep Neural Network Training with Horovod and Petastorm



https://jameskle.com/writes/spark-ai-2020

Generally Available

https://aka.ms/spark

SynapseML

A machine learning library that's

Simple

Quickly create, train, and use distributed machine learning tools in only a few lines of code.

Scalable

Scale ML workloads to hundreds of machines on your Apache Spark cluster.

Multilingual

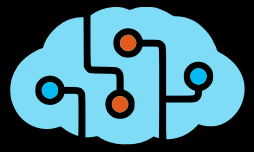
Use SynapseML from any Spark compatible language including Python, Scala, R, Java, .NET and C#.

Open

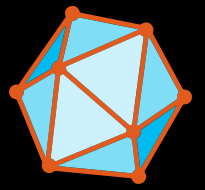
SynapseML is Open Source and can be installed and used on any Spark 3 infrastructure including your local machine, Databricks, Synapse Analytics, and others.

Synapse ML = Apache Spark ML +

Gradient Boosting



Cognitive Services



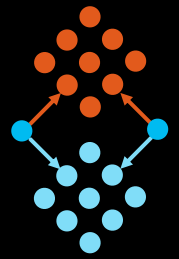
Deep learning

Reinforcement learning

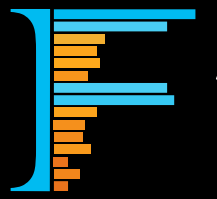
Responsible AI

Language Modeling

Search engine creation



Content retrieval



Anomaly detection

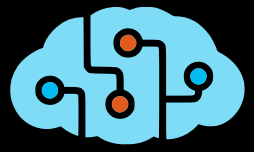
Cybersecurity

Explainable Models

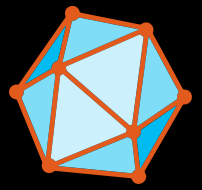
Image processing

Synapse ML = Apache Spark ML +

LightGBM



Cognitive Services



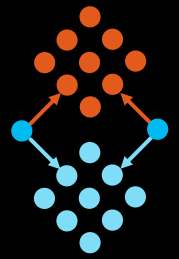
ONNX

Vowpal Wabbit

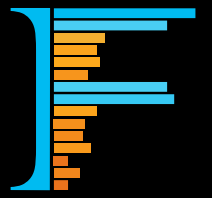
Responsible AI

PREDICT Keyword

Azure Search



Conditional KNNs



Isolation Forests

CyberML

Explainable Boosting Machines

OpenCV

In your language of choice

Python

R

Java and Scala

New in v0.10!

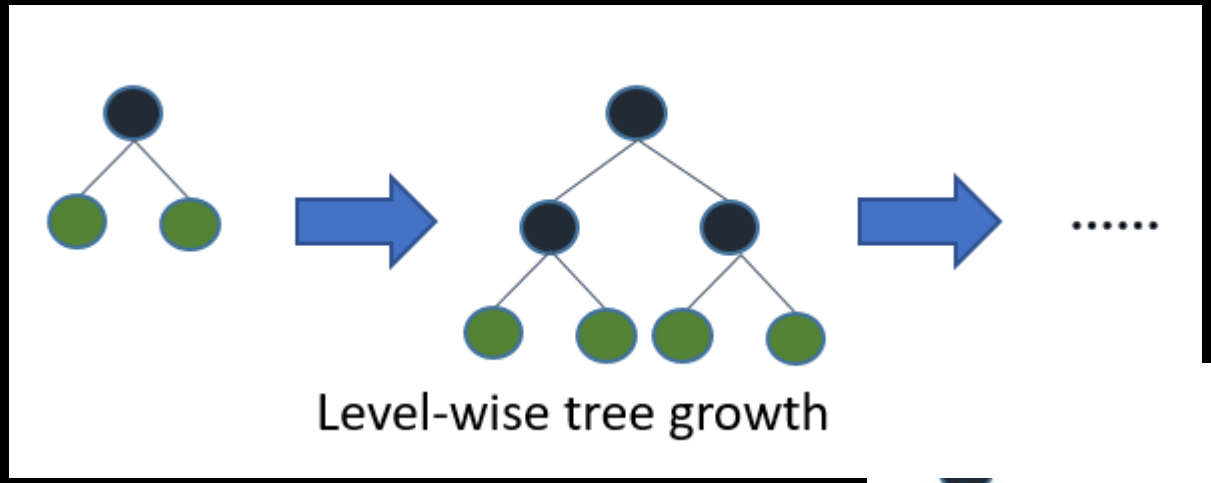
.NET, C#, and F#

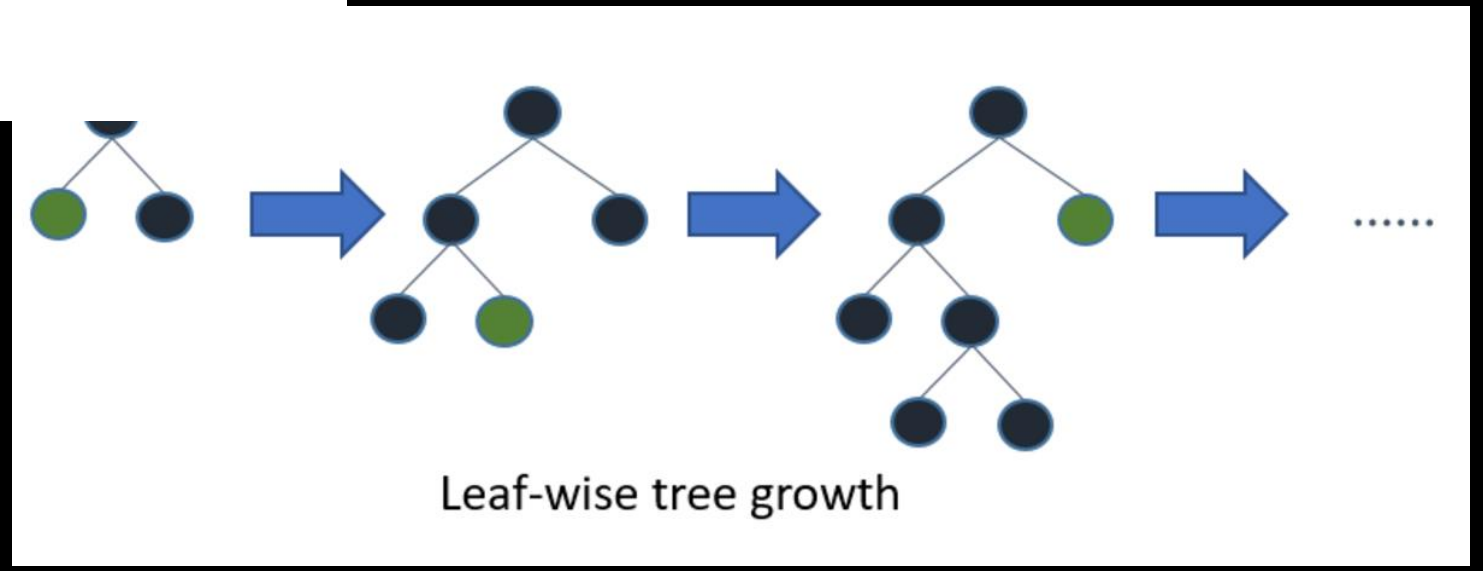
.NET

```
val model = new LightGBMRegressor()
    .setObjective("quantile")
    .setAlpha(0.2)
    .setLearningRate(0.3)

val results = model.fit(train).transform(test)
```

Demo:
Train a LightGBM model using SynapseML

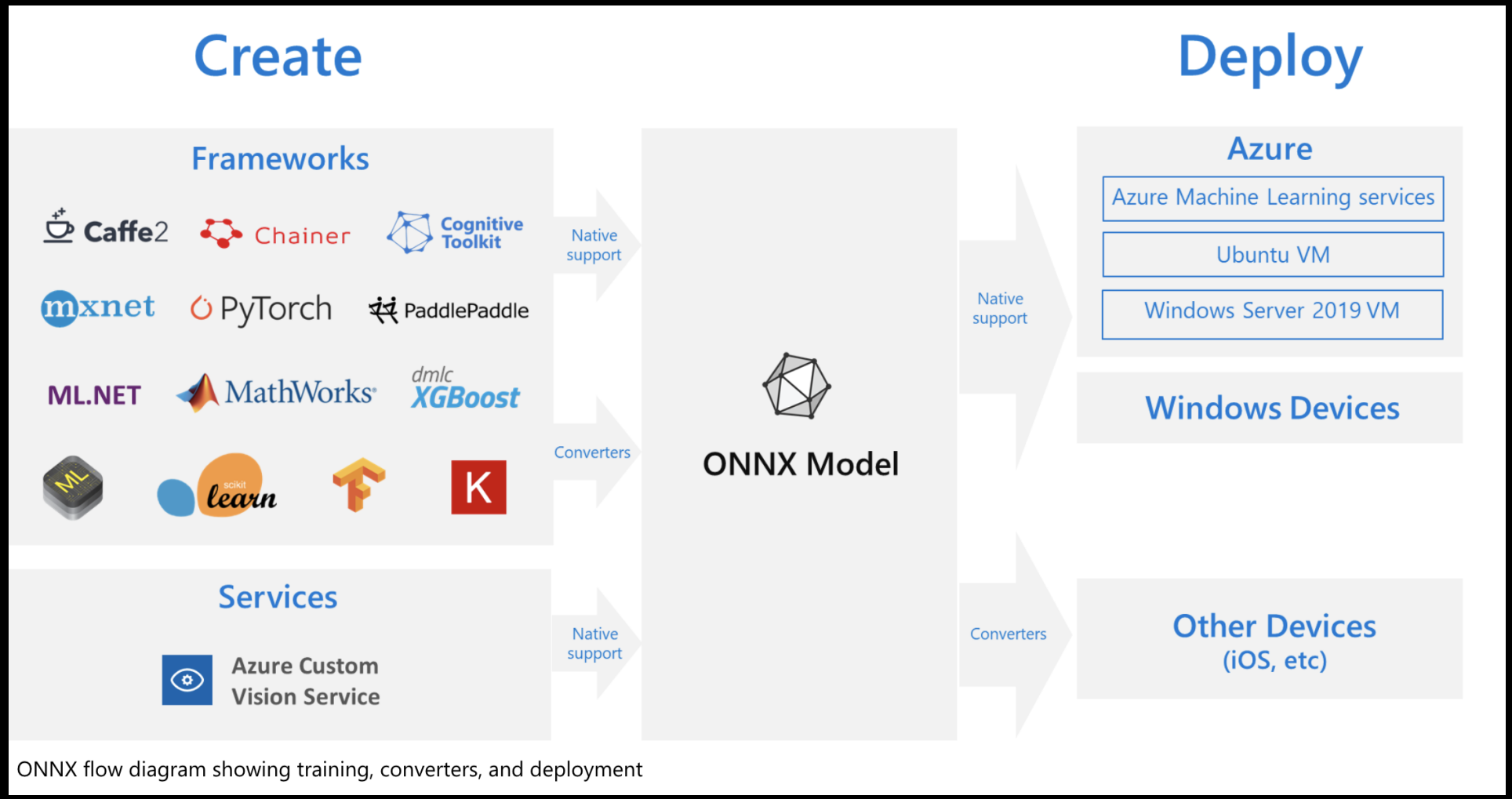




https://lightgbm.readthedocs.io/en/v3.3.2/

Scoring &
deployment

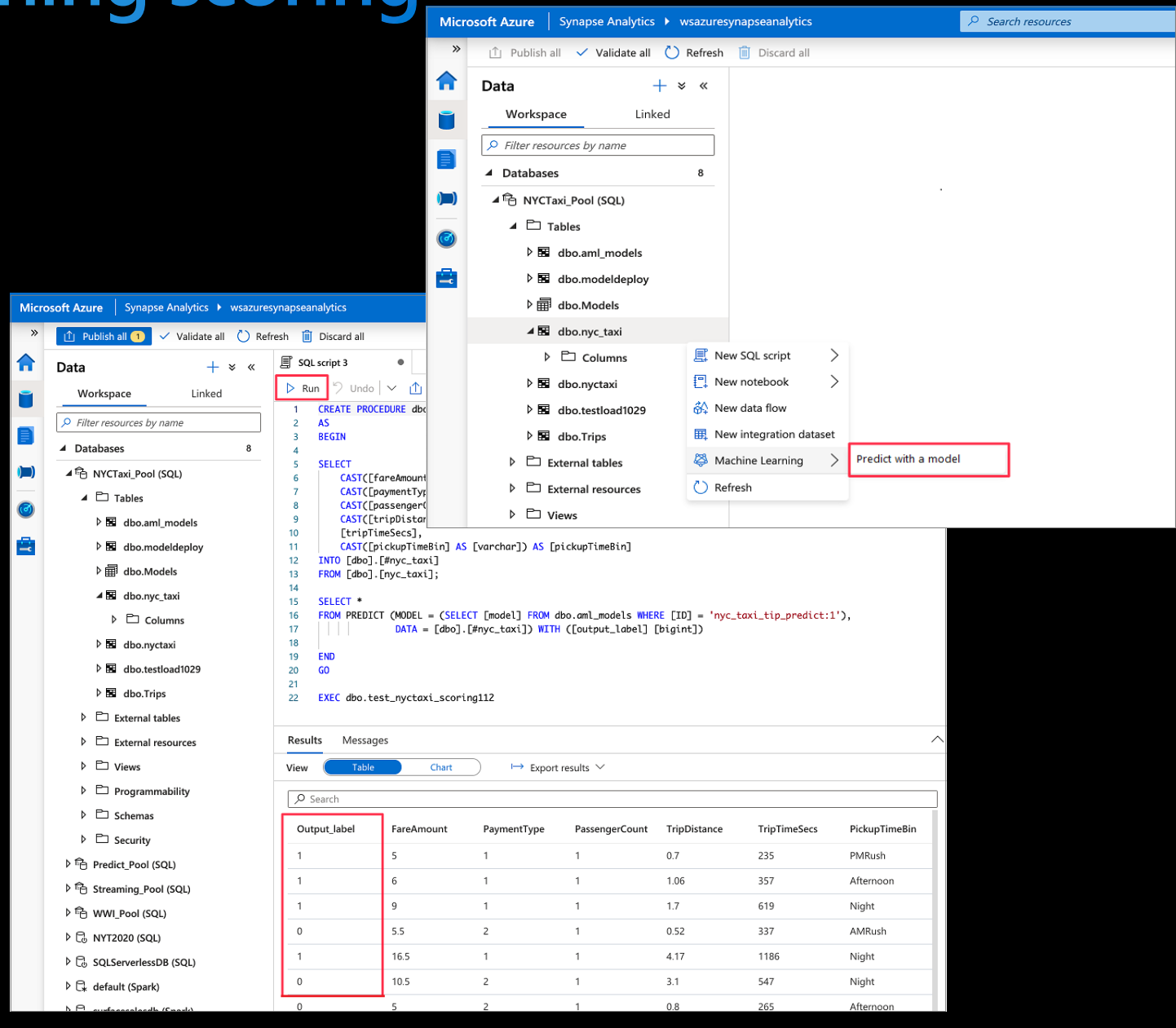
What is ONNX?



ONNX flow diagram showing training, converters, and deployment

Code-free machine learning scoring

- No-code **references** to machine learning models
- Democratize ML to everyone since no data science domain knowledge required
- Easily embed in SQL stored procedures for transformation of Views for reporting



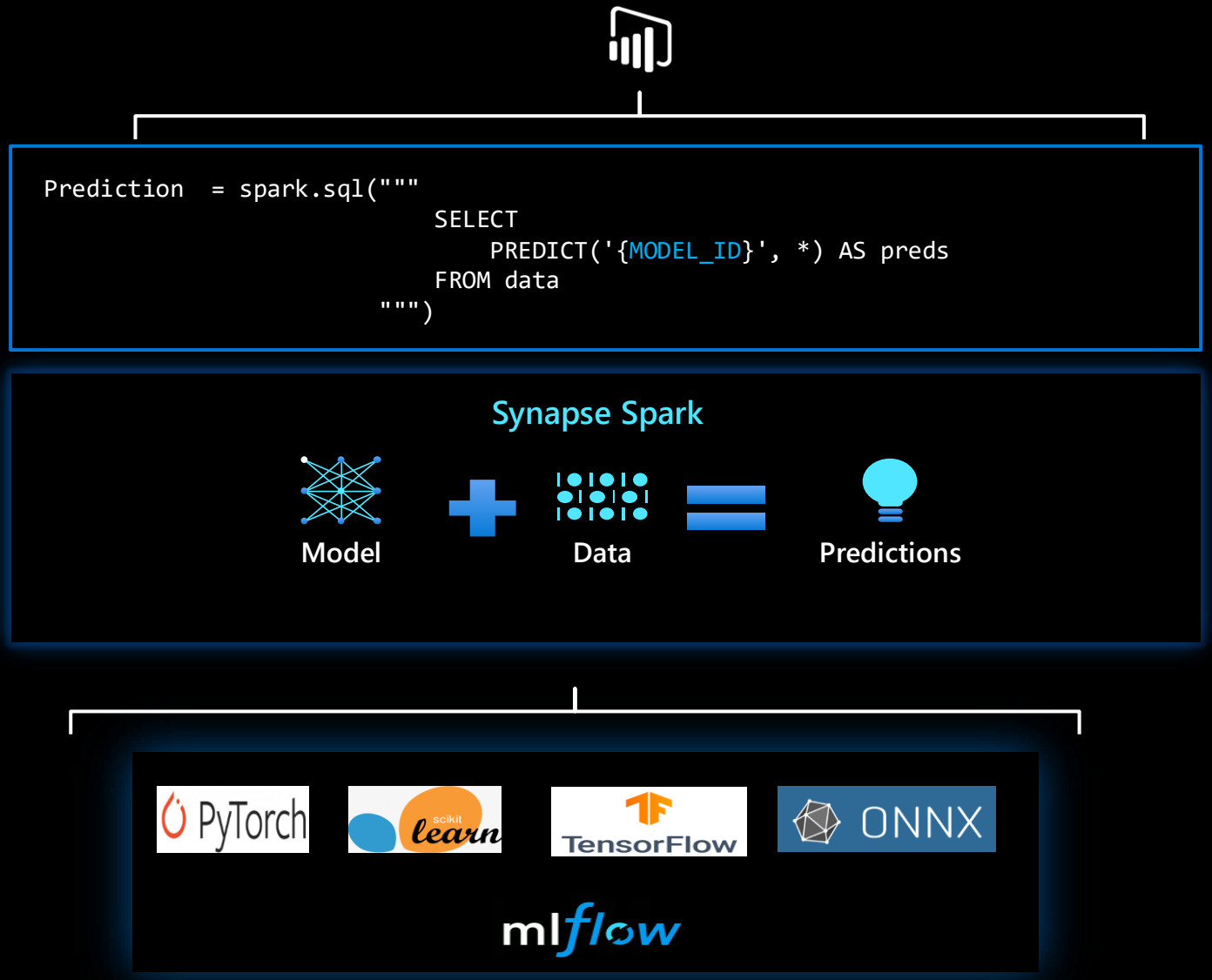
Model Prediction with Spark

Simplifying model scoring at scale on Spark

- Drastically simplifies handover of models from producer to consumer
- Machine Learning models from Azure ML, ADLS Gen2, REST
- "In-engine" for performance and scalability
- No data leaves the platform for scoring
- No additional cost for scoring

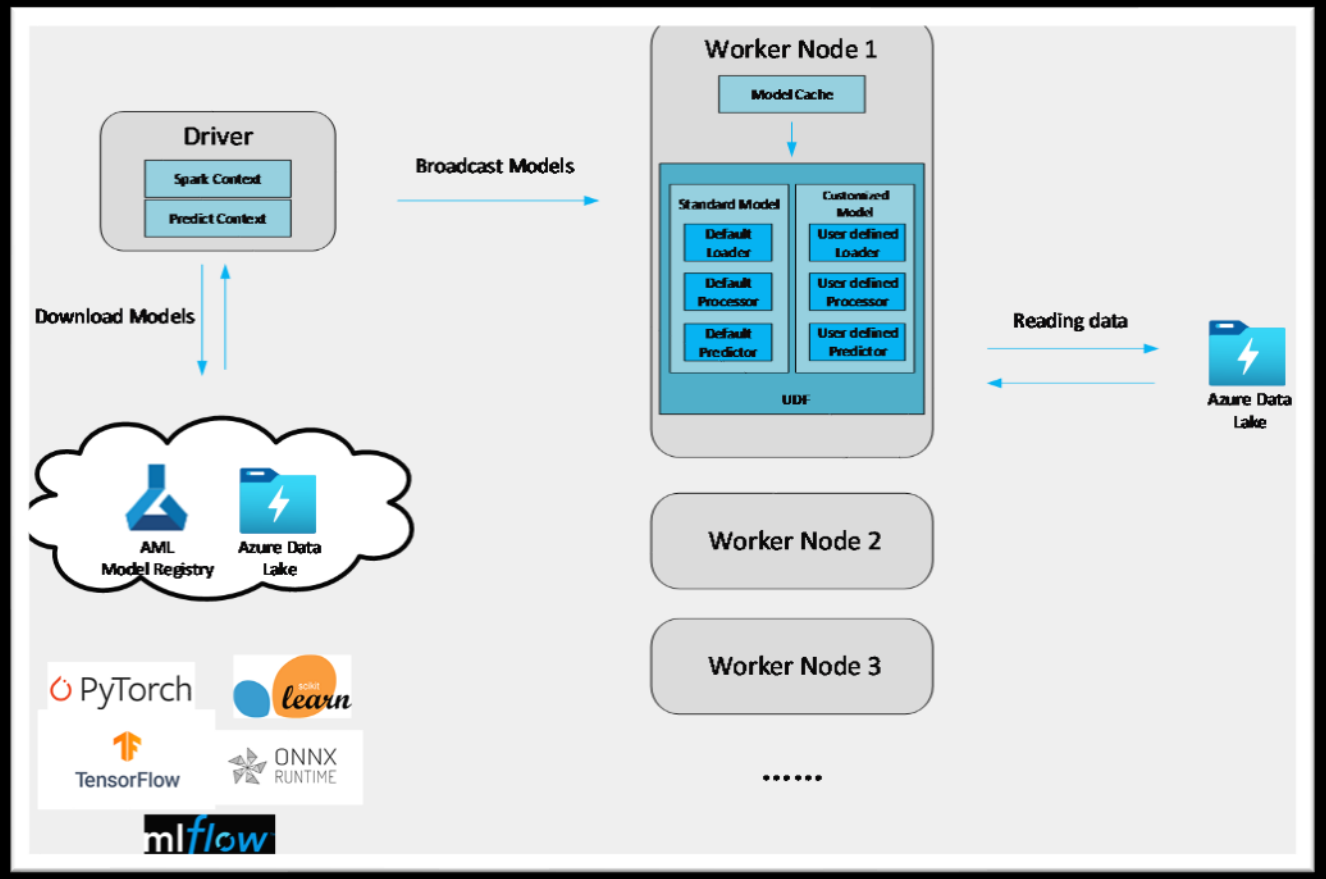
Scope public preview: Sklearn, Pytorch, TensorFlow, Onnx and PyFunc model flavors

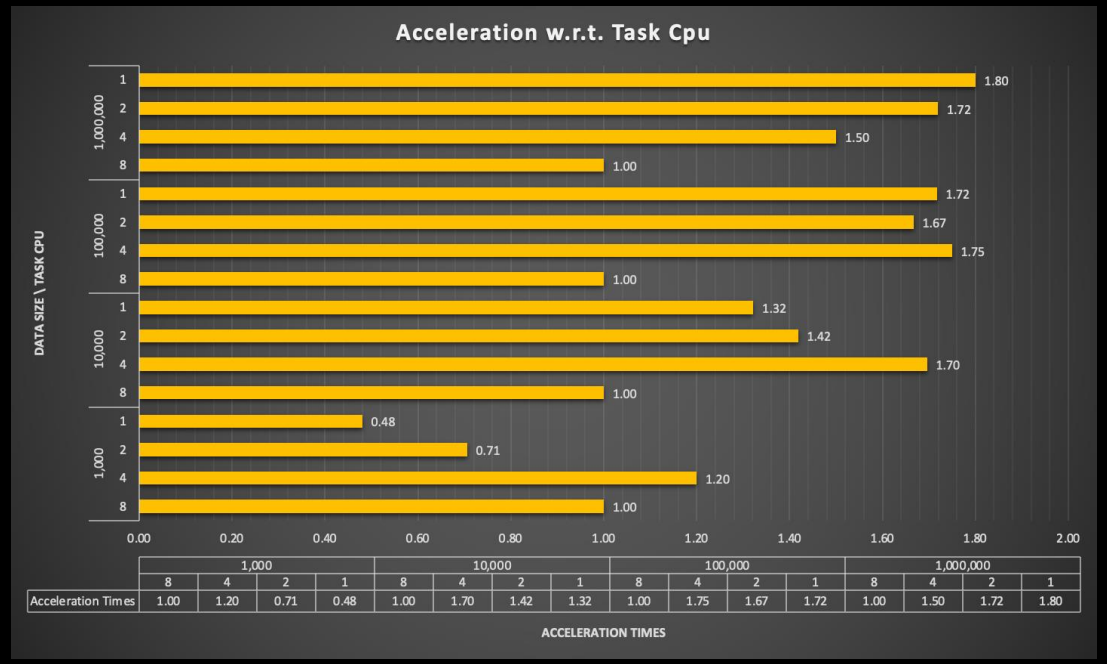
Scope GA: All mlflow model flavors

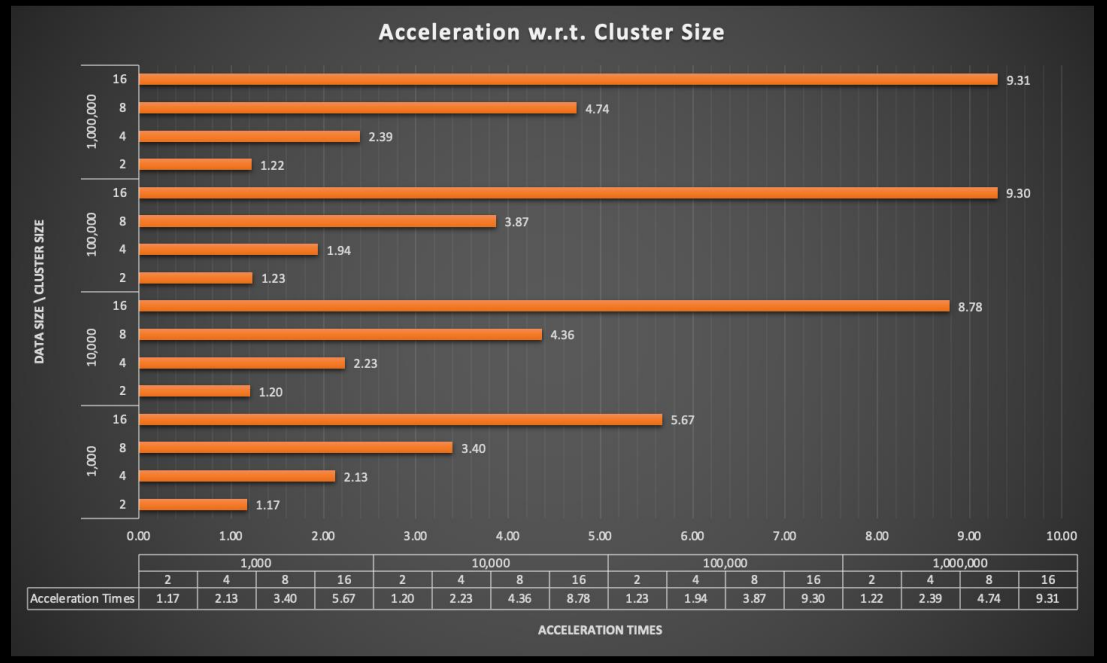


Public Preview

Model Prediction with Spark





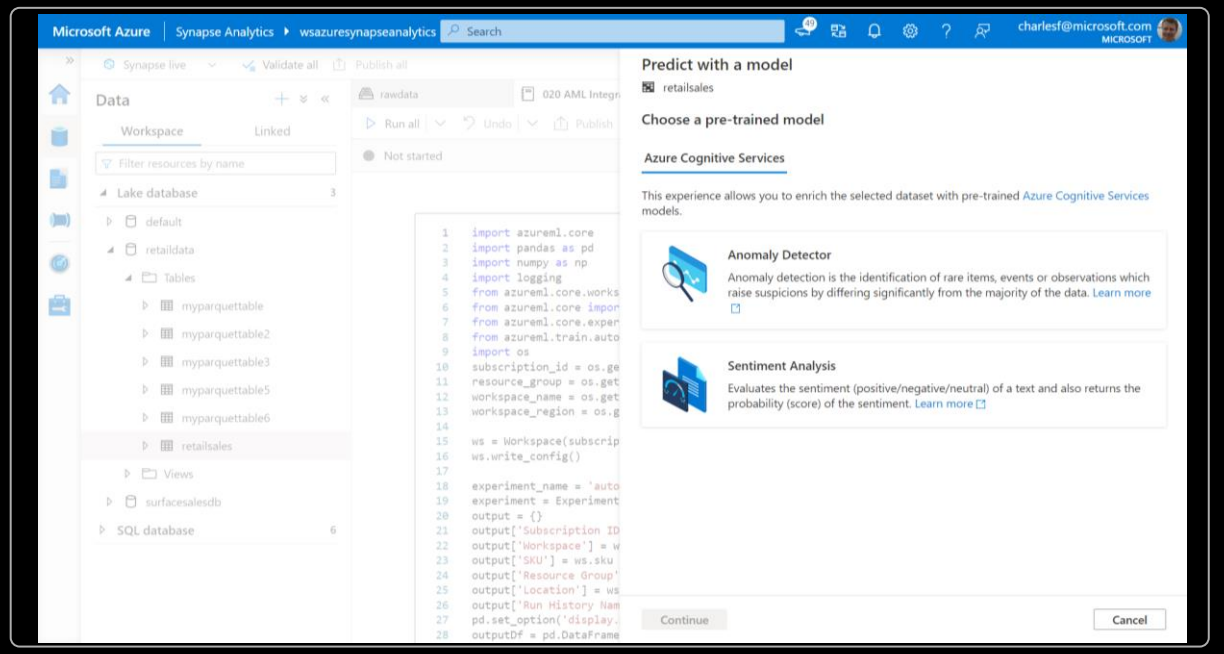


Generally Available

Built-in Cognitive Services

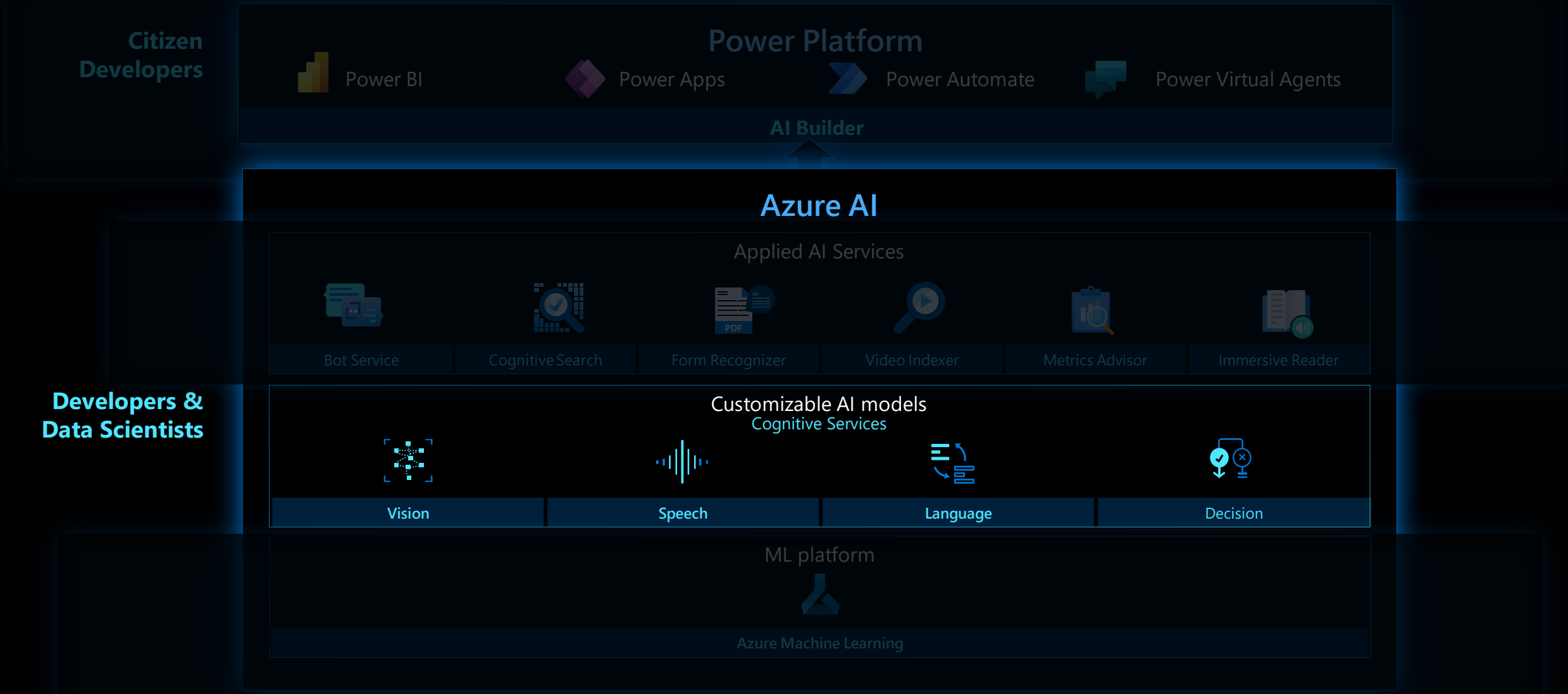
Enables simple integration of pre-built machine learning models.

- Guided UI experience for data enrichment
- SynpaseML
- Pyspark – code first



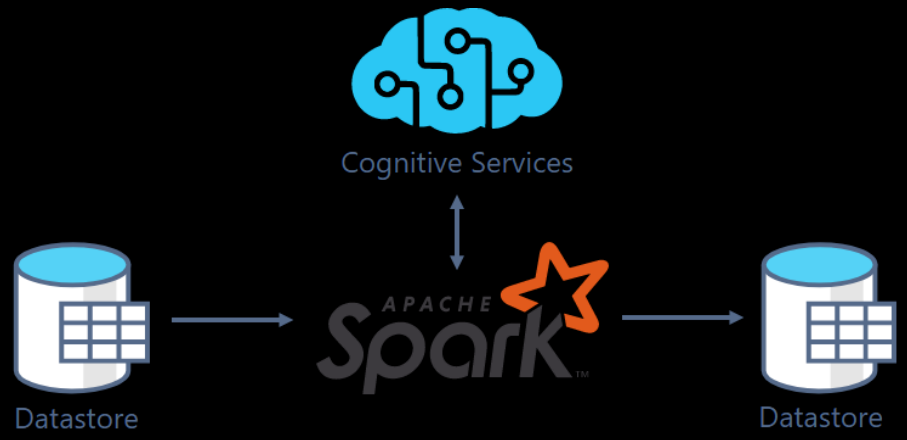
[Cognitive Services in Azure Synapse Analytics - Azure Synapse Analytics | Microsoft Learn]

What are cognitive services?



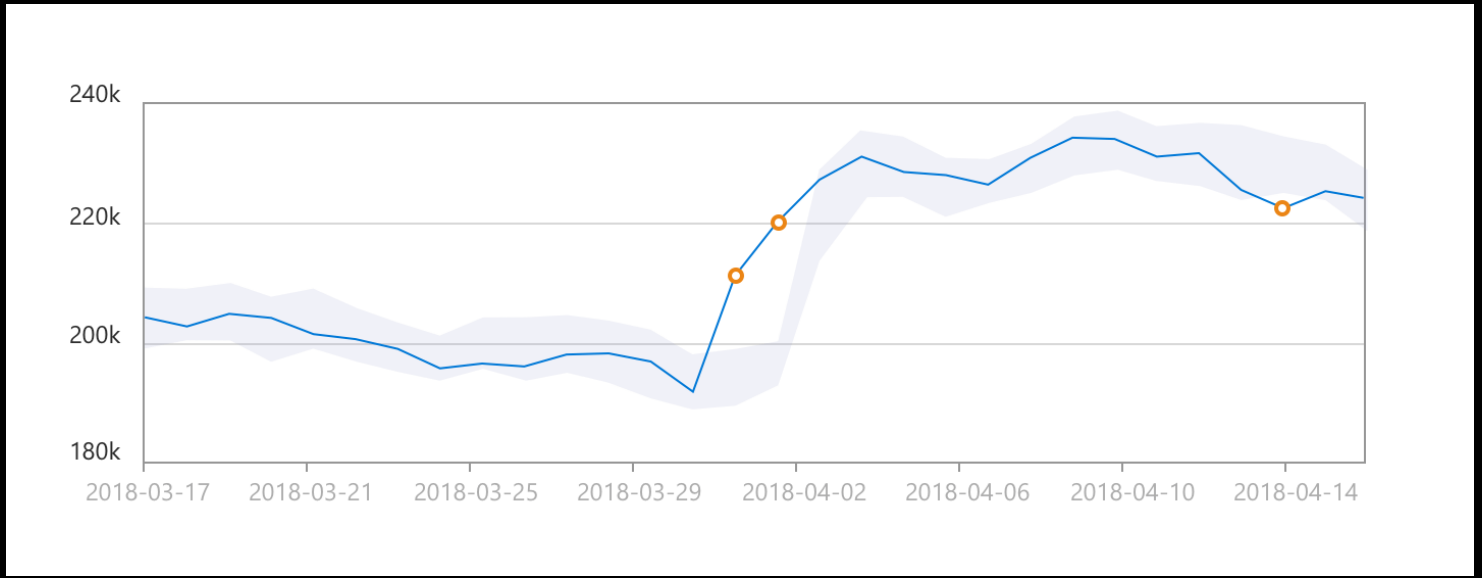
Demo: SynapseML + cognitive services

1. Create Synapse Spark pool
2. Create Azure cognitive services resource
3. Create keyvault + cog api key
4. Link key vault + cog service to Synapse workspace



Demo:
Anomaly detection wizard

1. Create Spark pool in Synapse
2. Generate data
3. Create Azure Anomaly detector
4. Add api key to keyvault
5. Link anomaly detector to Synapse workspace
6. Use wizard to predict



What is Anomaly Detector? - Azure Cognitive Services | Microsoft Learn

Q&A

Microsoft

Enregistrez vous dès maintenant au prochain Webinars Data AI

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