Microsoft

Azure Synapse Analytics

Marek Chmel

Sr. Cloud Solutions Architect | Data @ Microsoft CEE

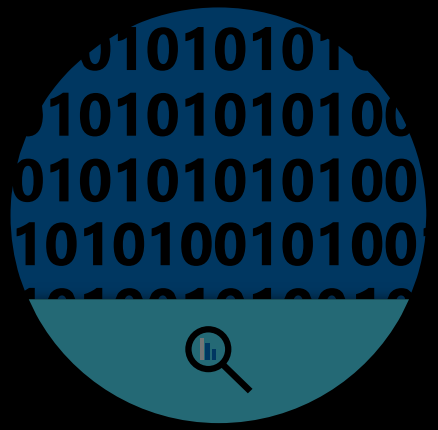
Agenda

Topics covered during presentation

- Introduction
- Azure Synapse Concepts and Terminologies
- Data Ingestion
  - Dedicated SQL Pool
  - Serverless Pool
  - Spark Pool
- Querying Data
- Reporting based on Azure Synapse Analytics

While data grows 400%, less than 30% gets analyzed

2020



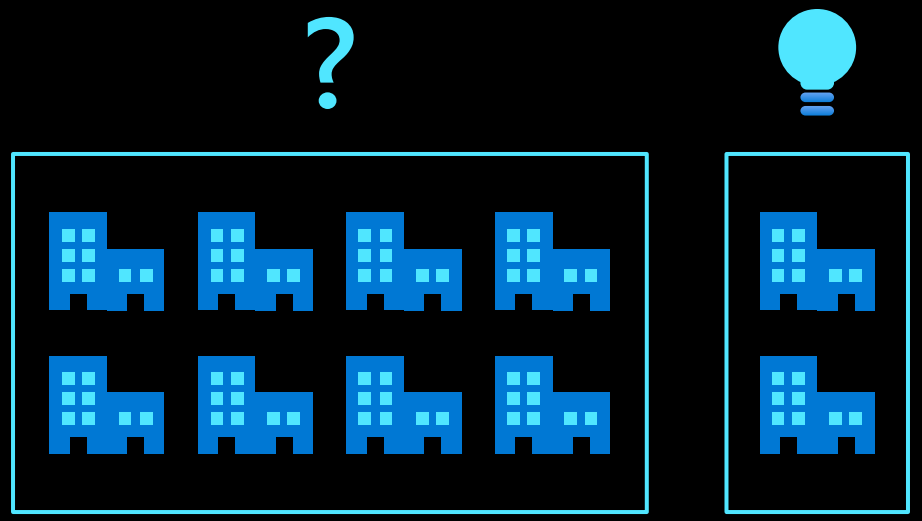
4x

2025

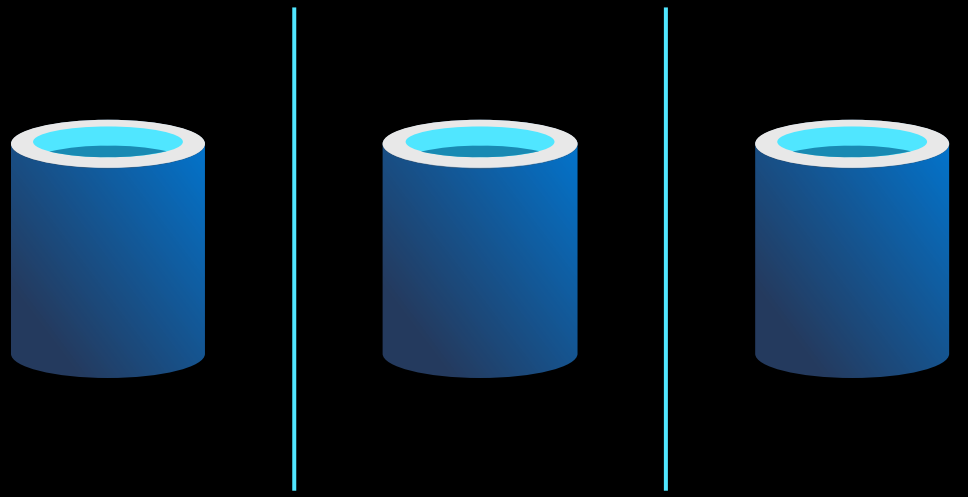
44ZB Data collected worldwide

175ZB Data collected worldwide

Analytics & AI is the #1 investment for business leaders, however they struggle to maximize ROI



80%



55%

* Harvard Business Review (2019), Understanding why analytics strategies fall short for some, but not for others

This leads to the paradox of analytics

Each new technology creates another siloed operation

Big data

Data integration

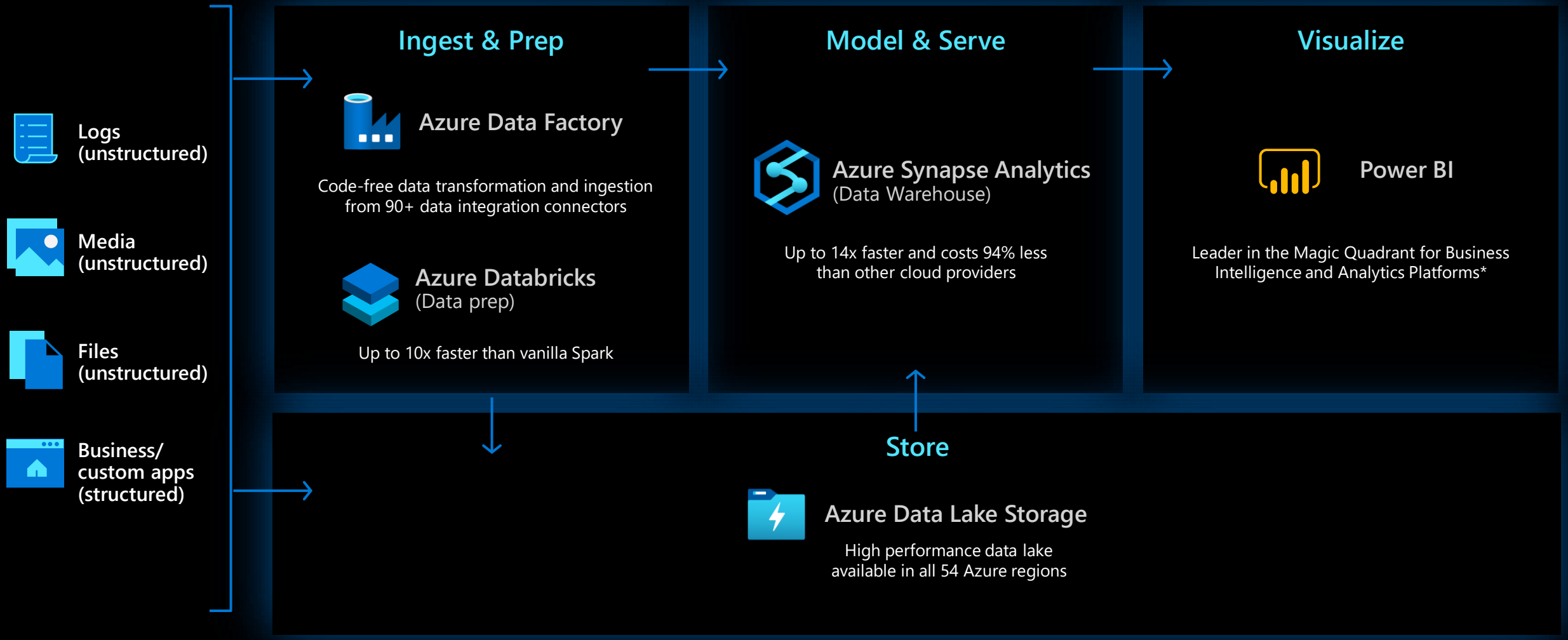
Machine learning

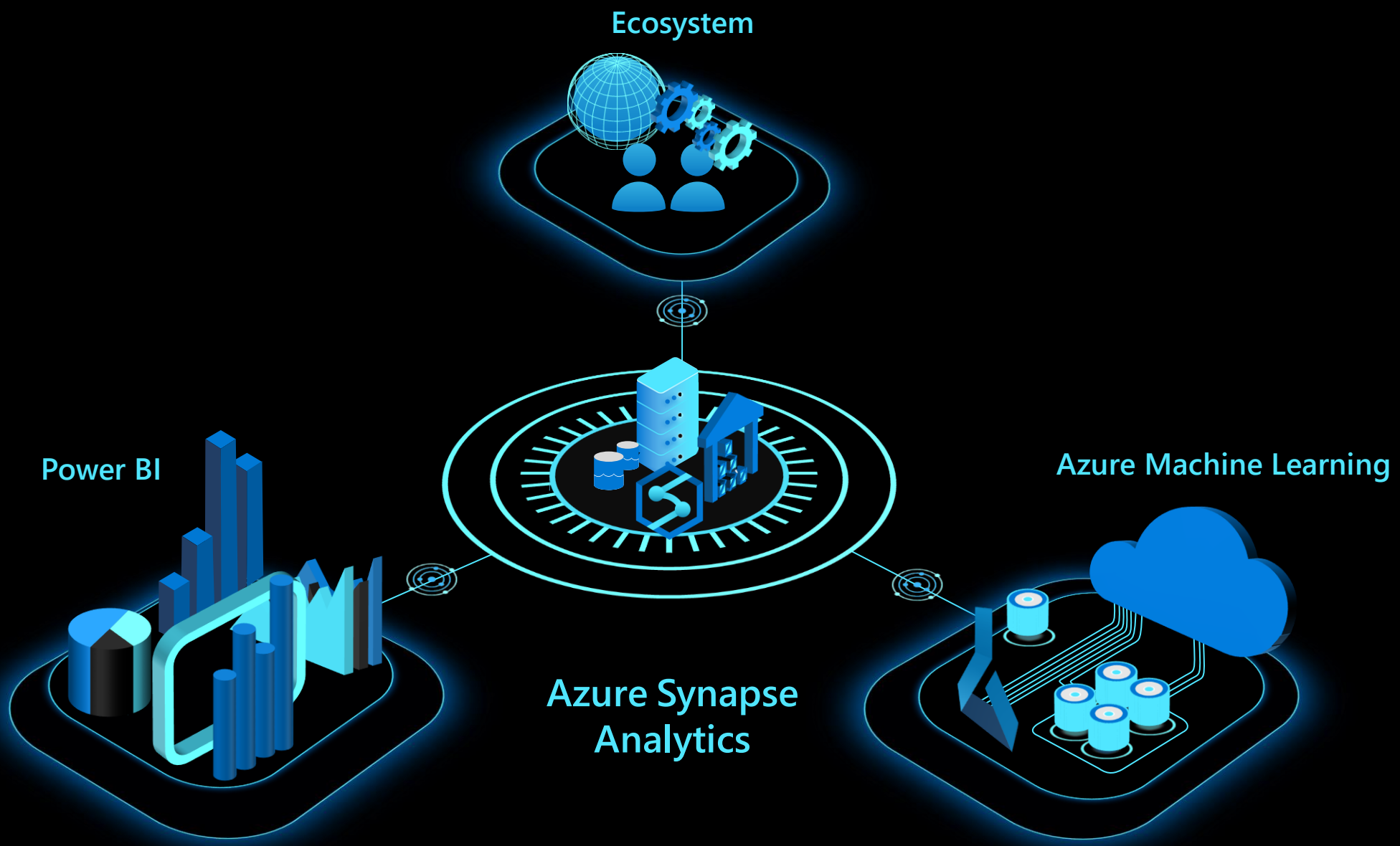
Business intelligence

Data governance

Security

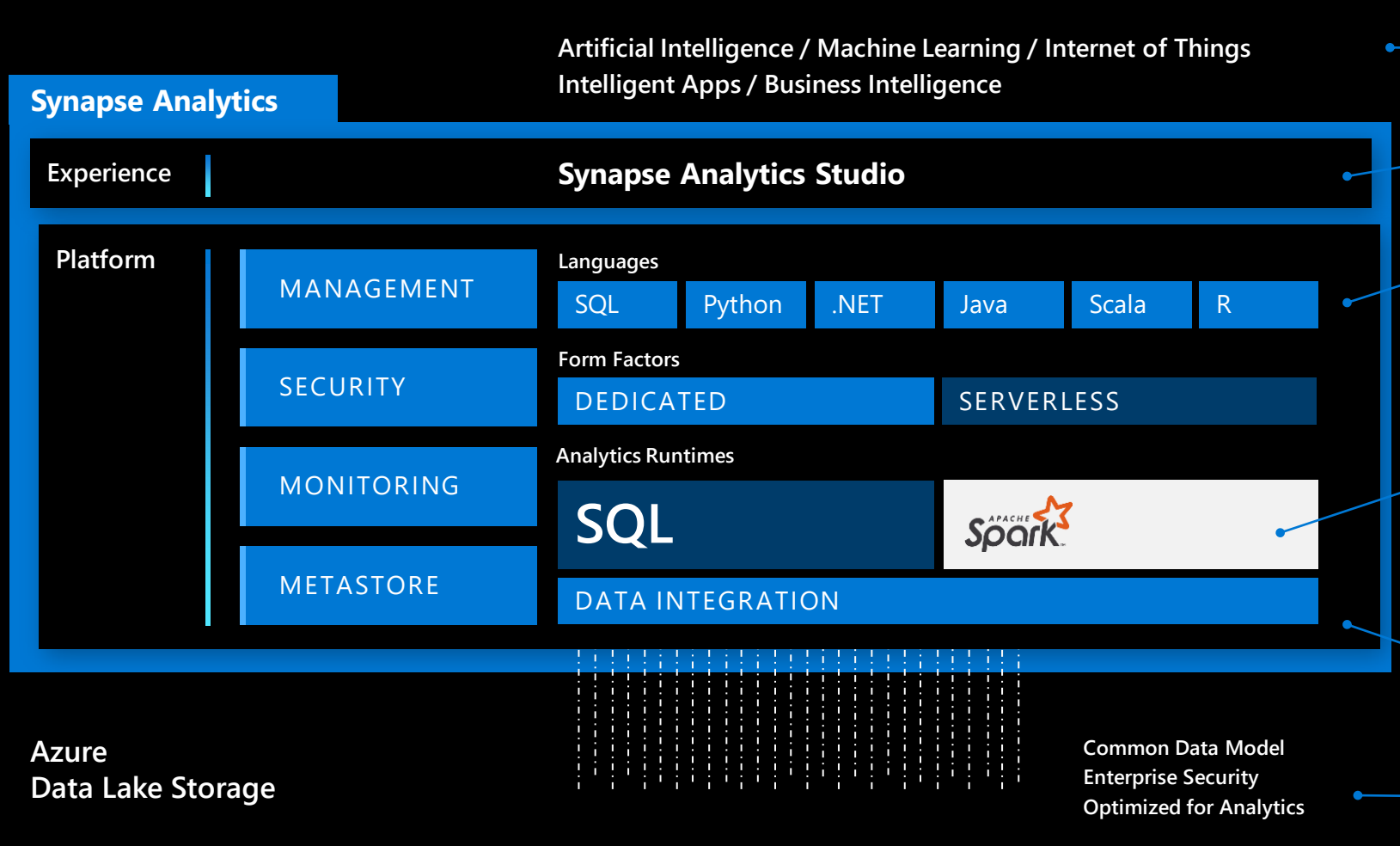
Modern Data Warehouse




Ecosystem
Power BI
Azure Machine Learning
Azure Synapse Analytics

Azure Synapse Analytics

Integrated data platform for BI, AI and continuous intelligence



Designed for analytics **workloads at any scale**

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

Multiple **languages** suited to different analytics workloads

Integrated analytics runtimes available provisioned and serverless on-demand

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

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

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

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

Azure Synapse Analytics

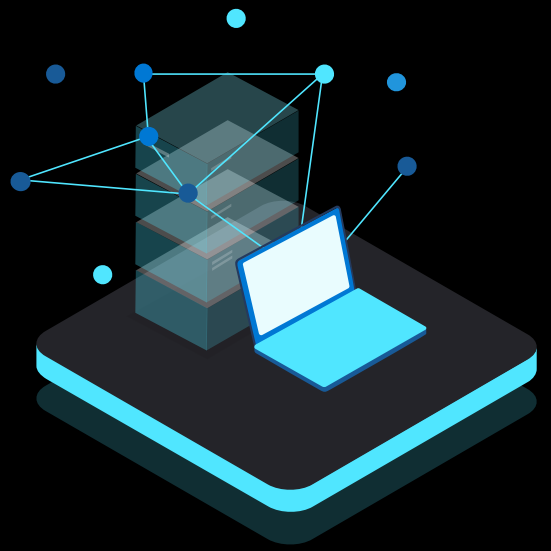
Powered by a new cloud native distributed SQL engine

Serverless + dedicated SQL

Flexible consumption models

Serverless pay-per-query ideal for ad-hoc data lake exploration and transformation

Dedicated clusters optimized mission-critical data warehouse workloads



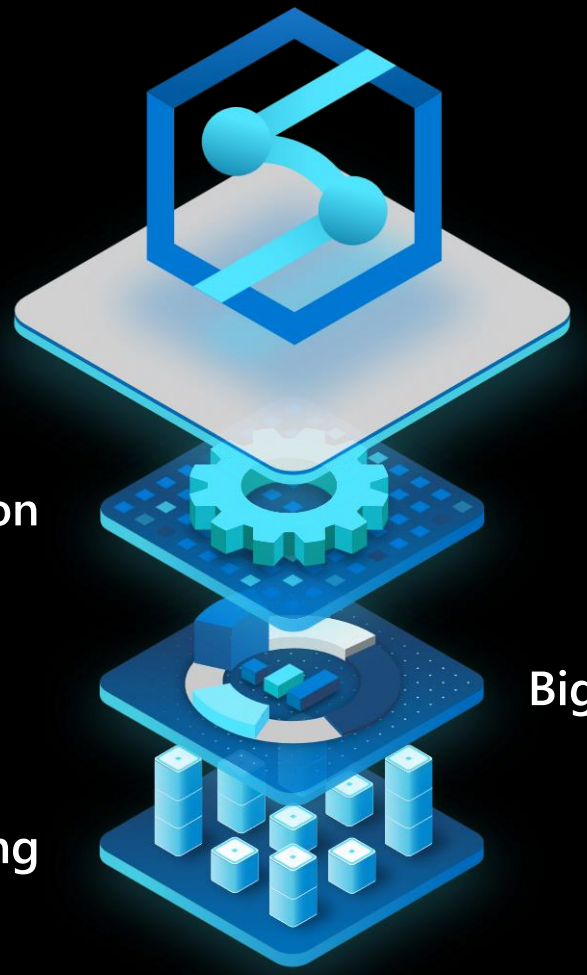
Serverless

Dedicated

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 Eco System

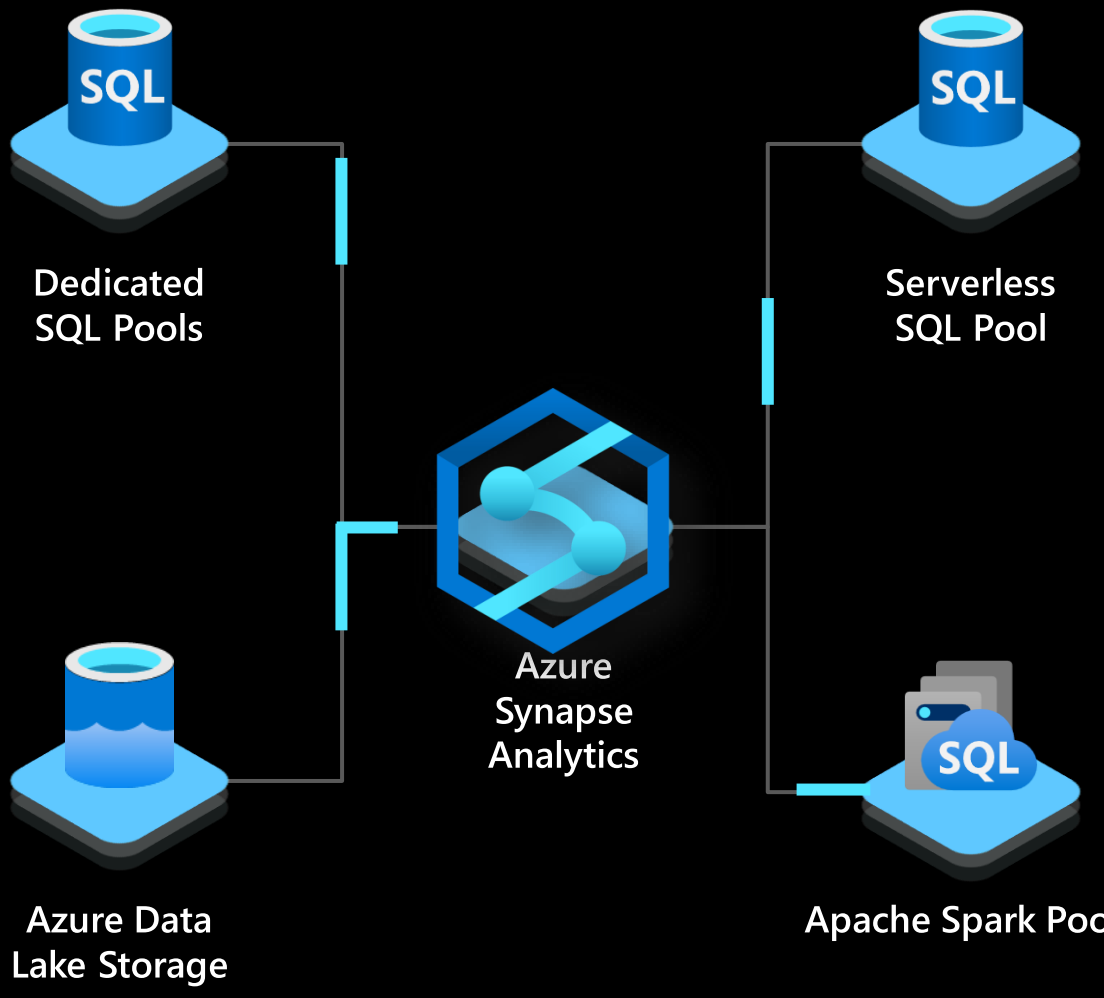
Integrated data platform

Dedicate SQL Pools

Serverless SQL Pool

Apache Spark Pool

Azure Data Lake Storage



Microsoft

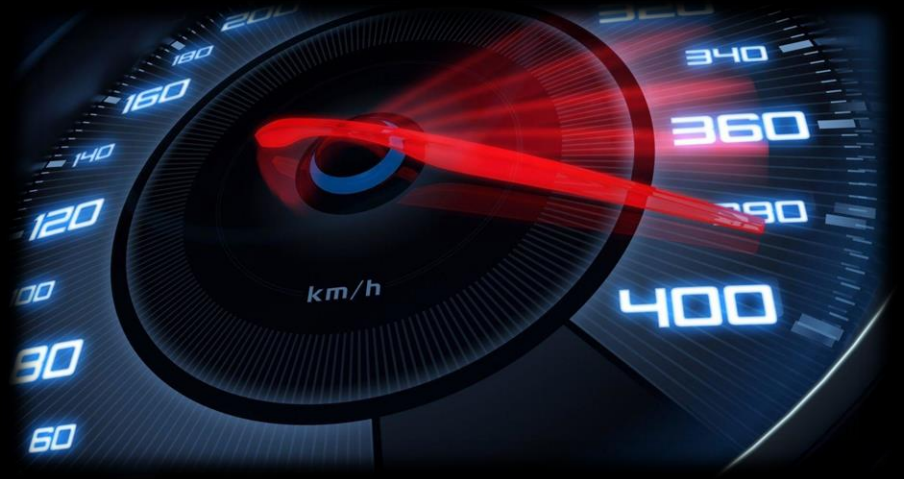
Let’s get the concepts clear

- Massively Parallel Processing (MPP)
- Control Node
- Compute Nodes
- Distributions
- Compute Datawarehouse Units (cDWU’s)

Microsoft

Massively Parallel Processing (MPP)

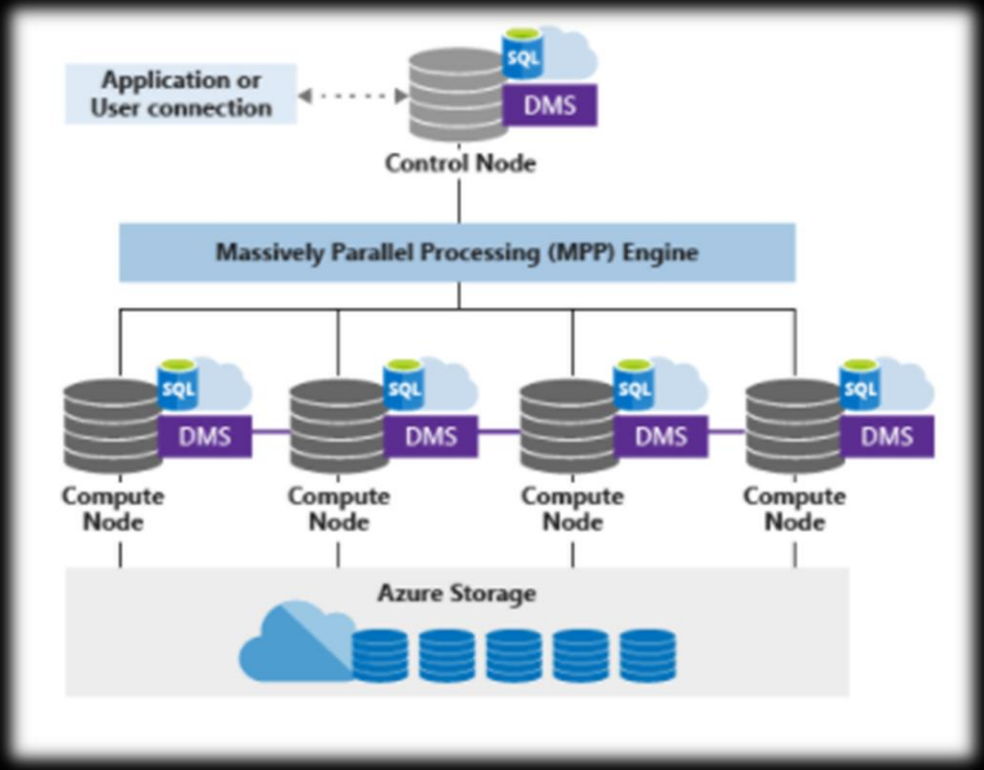
- "Divide and Conquer" Concept
- Uses many separate CPUs running in parallel to execute a single program
- Shared Nothing: Each CPU has its own memory and disk (scale-out)



Microsoft

Control Node

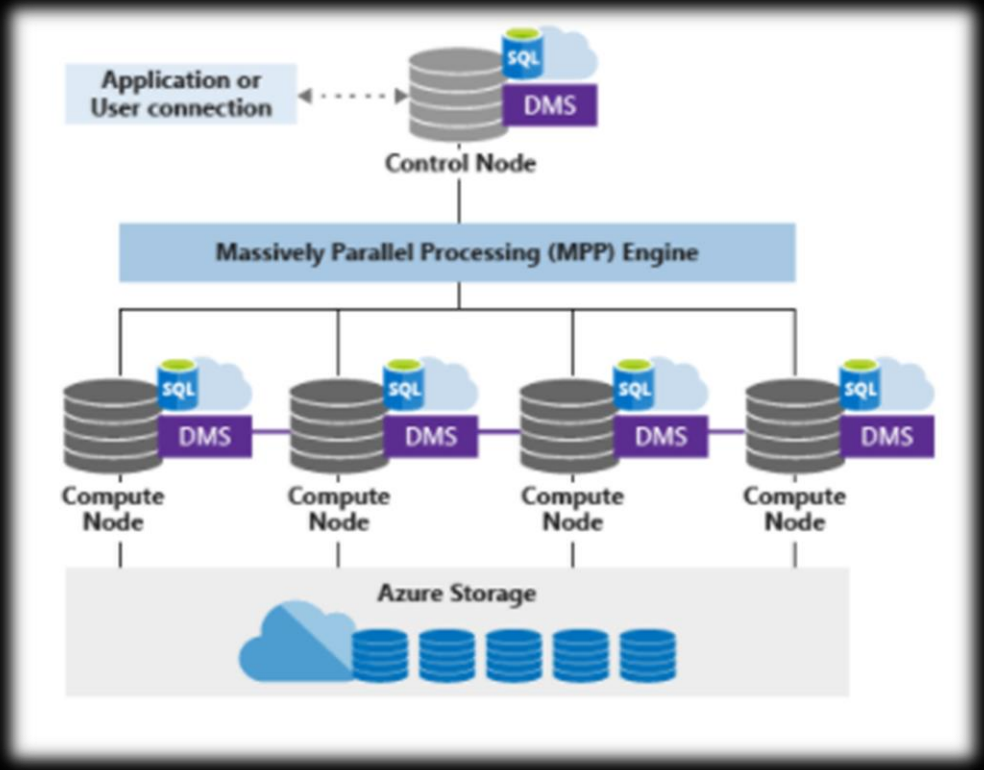
- The brain of the MPP architecture
- It is the front end to that interacts with applications and connections
- The distributed query engine runs on the control node
- Responsible for optimizing and coordinating queries



Microsoft

Compute Nodes

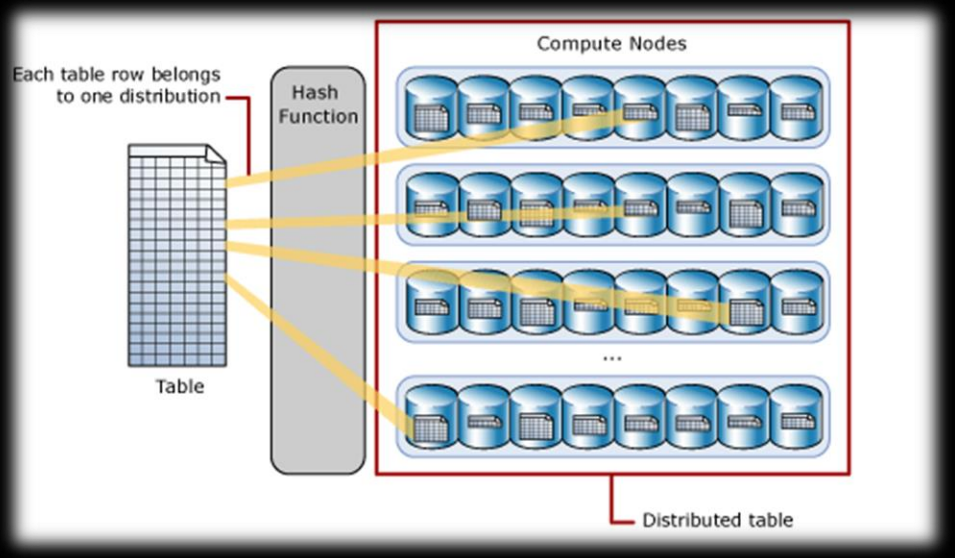
- The brawn of the MPP architecture
- Provides computational power
- Number of compute nodes can range from 1 to 60

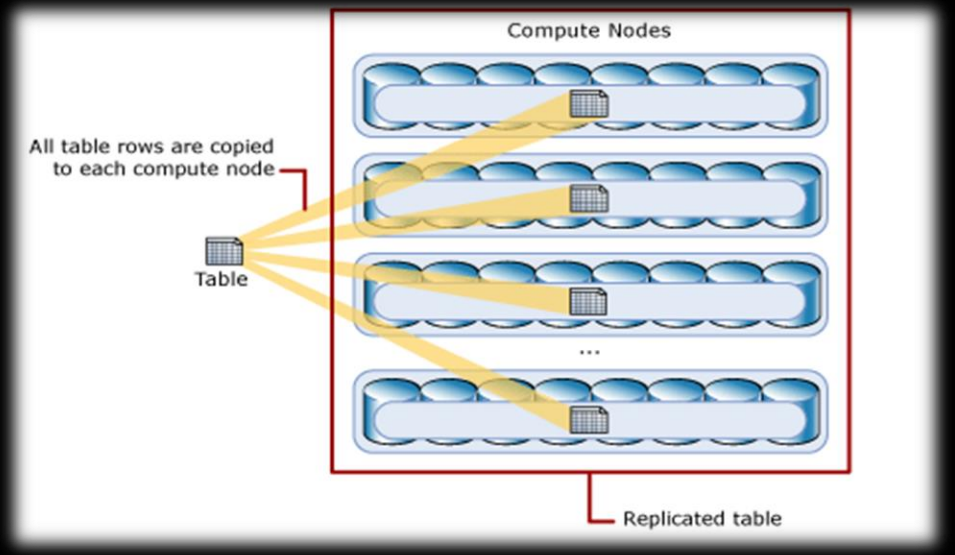


Microsoft

Distributions

- The work is divided into 60 distributions
- Various ways to distribute the data:
- Hash-distributed tables
- Round-robin distributed tables
- Replicated tables

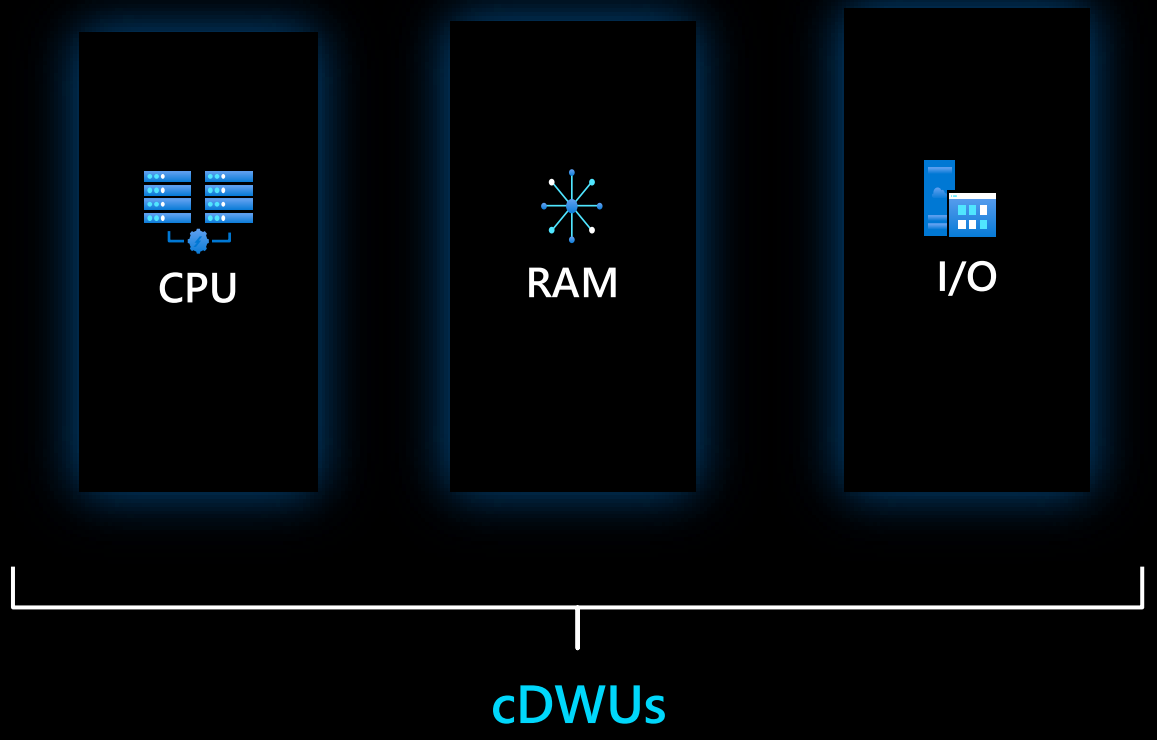




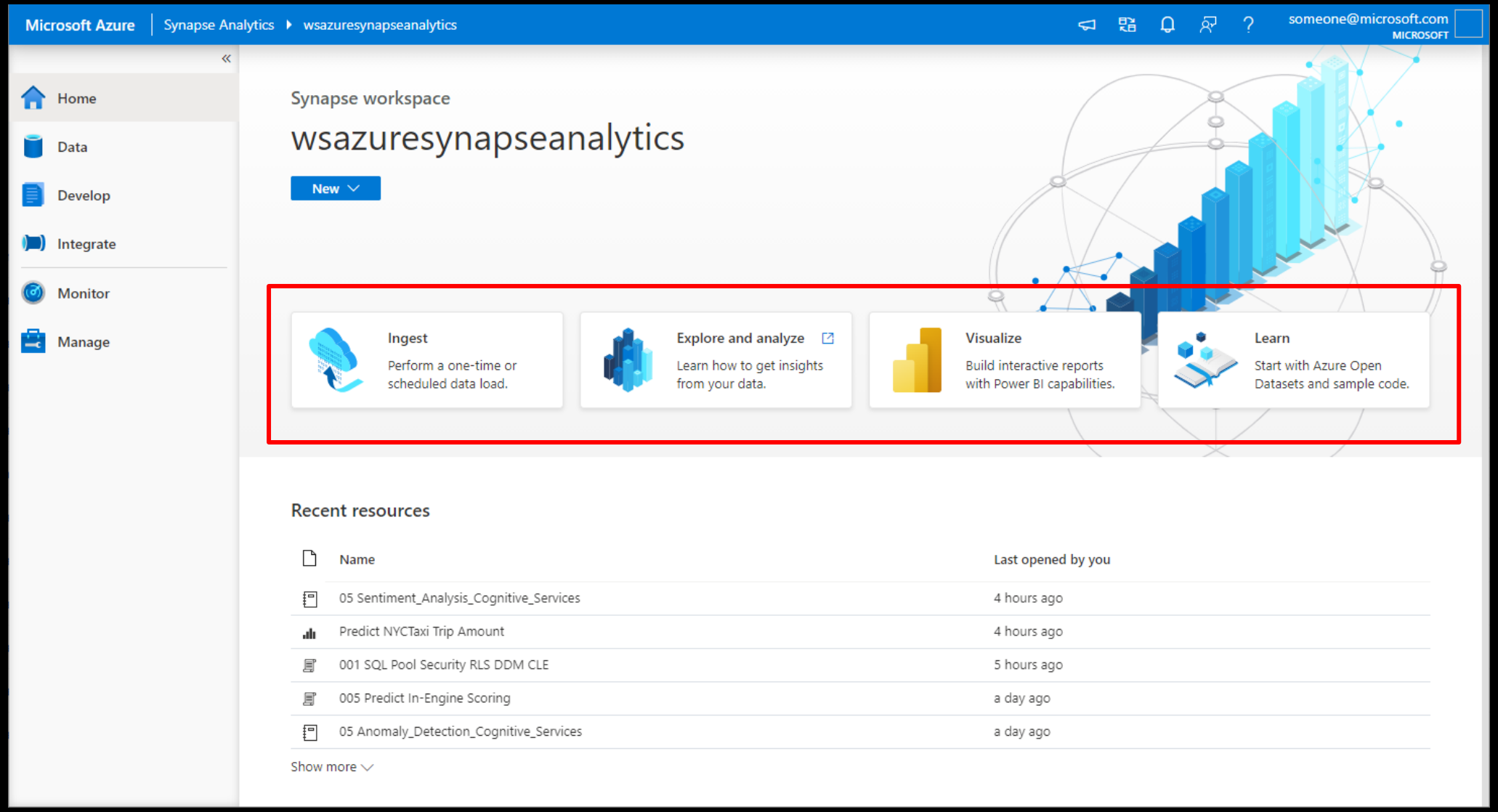
Microsoft

Compute Data Warehouse Units (cDWUs)

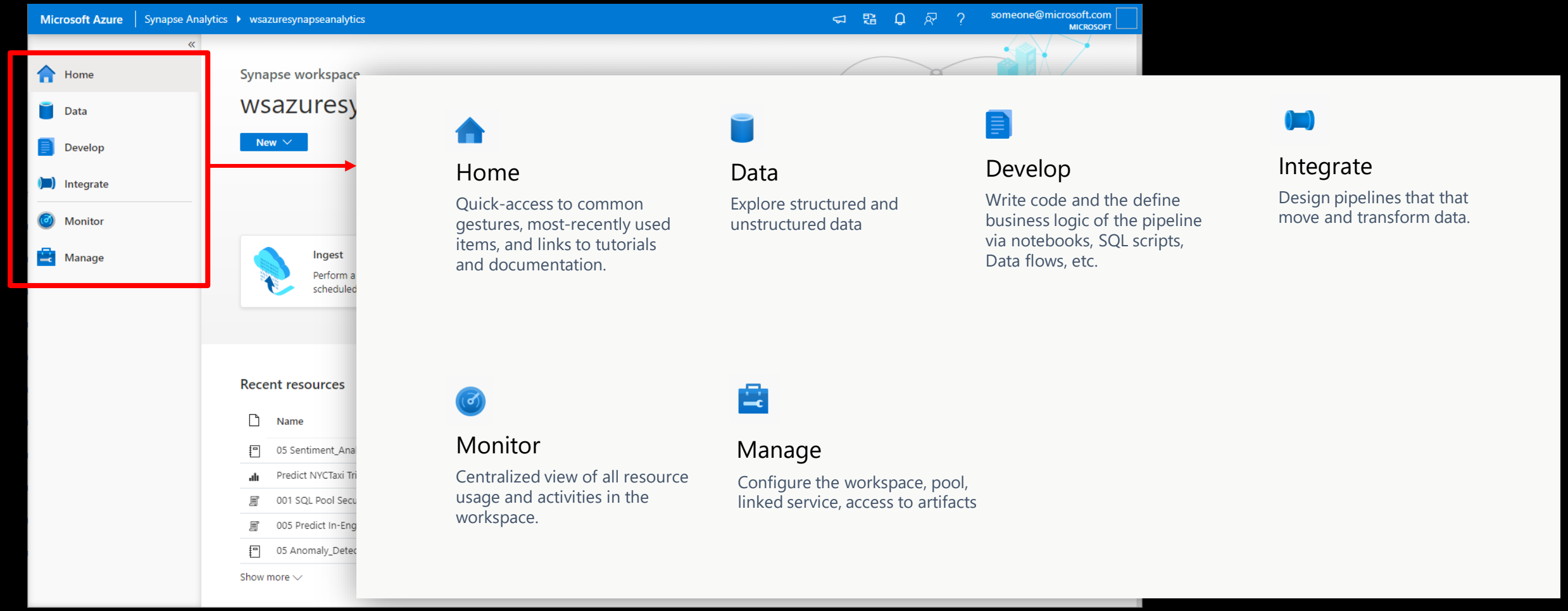
- cDWU is a combination of CPU, Memory and IO
- More cDWU you have, more the compute power
- cDWU ranges from DW100c to DW30,000c

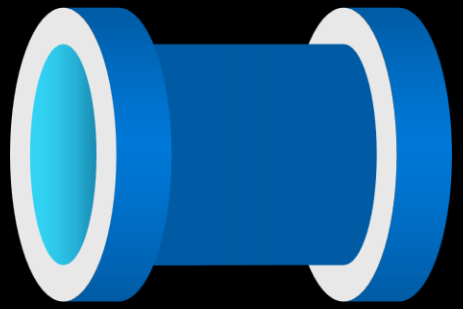


Home Hub


Microsoft Azure
Synapse Analytics
wsazuresynapseanalytics
someone@microsoft.com
MICROSOFT
Home
Data
Develop
Integrate
Monitor
Manage
Synapse workspace
wsazuresynapseanalytics
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Learn
Start with Azure Open Datasets and sample code.
Recent resources
Name
Last opened by you
05 Sentiment_Analysis_Cognitive_Services
4 hours ago
Predict NYCTaxi Trip Amount
4 hours ago
001 SQL Pool Security RLS DDM CLE
5 hours ago
005 Predict In-Engine Scoring
a day ago
05 Anomaly_Detection_Cognitive_Services
a day ago
Show more

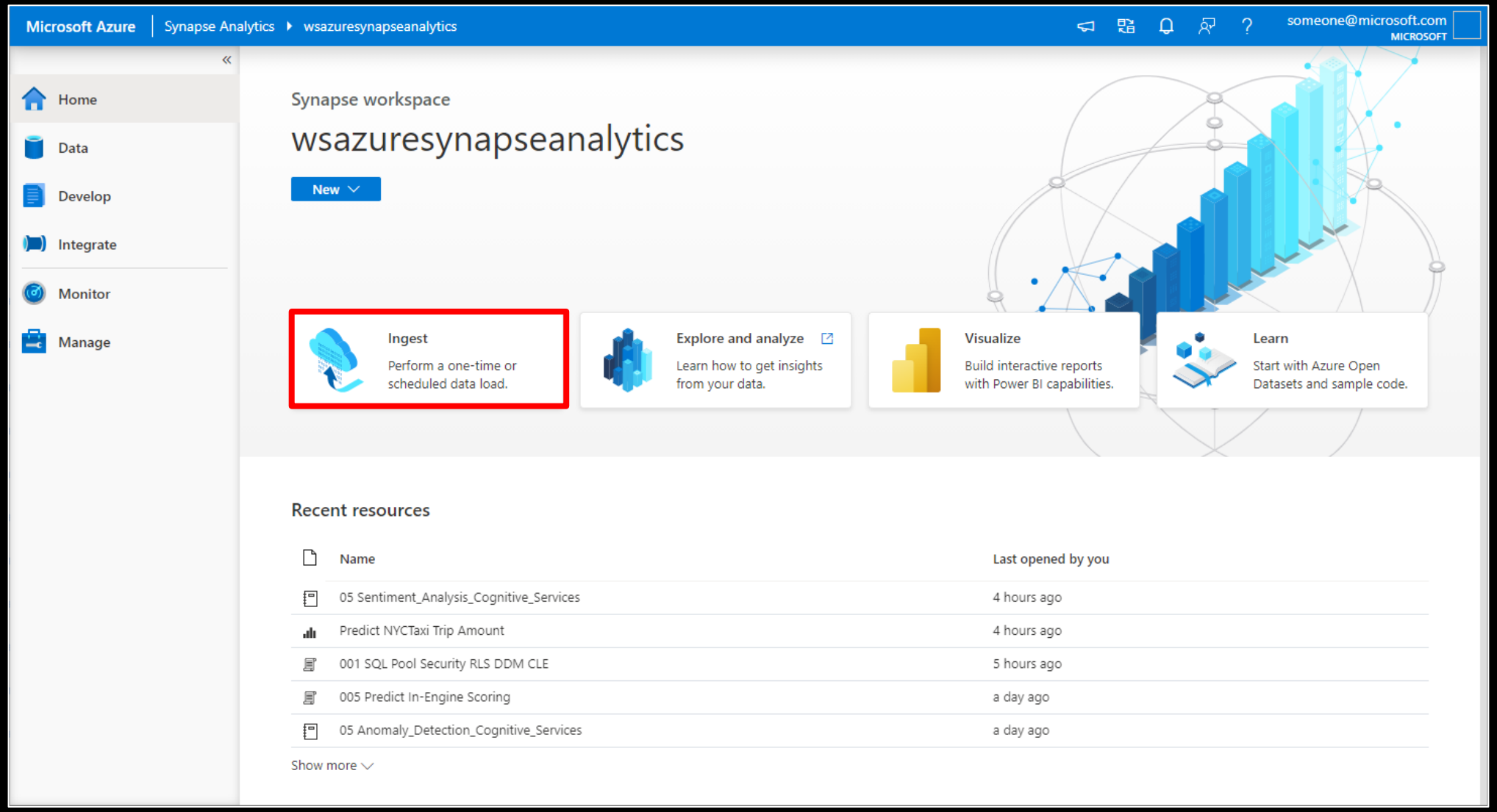
Synapse Studio



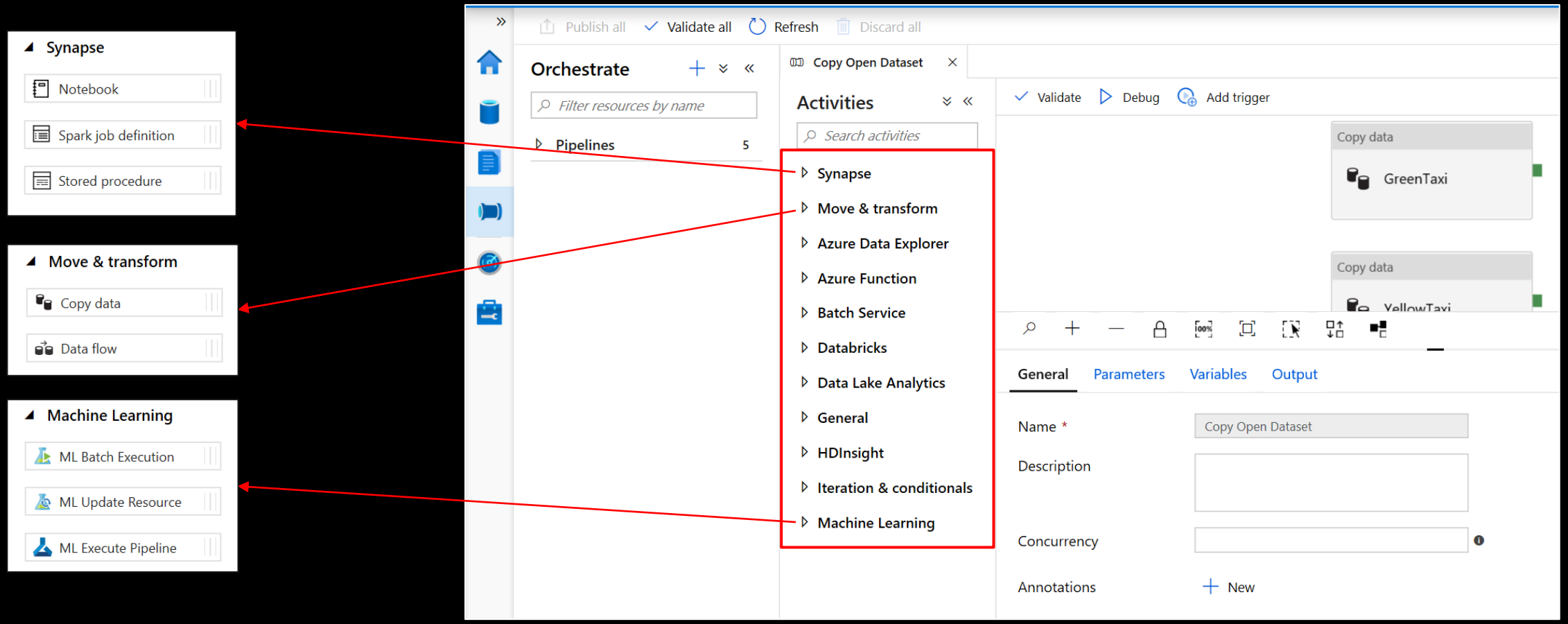


Synapse Studio
Data Ingestion

Studio


Microsoft Azure
Synapse Analytics
wsazuresynapseanalytics
someone@microsoft.com
MICROSOFT
Home
Data
Develop
Integrate
Monitor
Manage
Synapse workspace
wsazuresynapseanalytics
New
Ingest
Perform a one-time or scheduled data load.
Explore and analyze
Learn how to get insights from your data.
Visualize
Build interactive reports with Power BI capabilities.
Learn
Start with Azure Open Datasets and sample code.
Recent resources
Name
Last opened by you
05 Sentiment_Analysis_Cognitive_Services
4 hours ago
Predict NYCTaxi Trip Amount
4 hours ago
001 SQL Pool Security RLS DDM CLE
5 hours ago
005 Predict In-Engine Scoring
a day ago
05 Anomaly_Detection_Cognitive_Services
a day ago
Show more

Integrate Hub

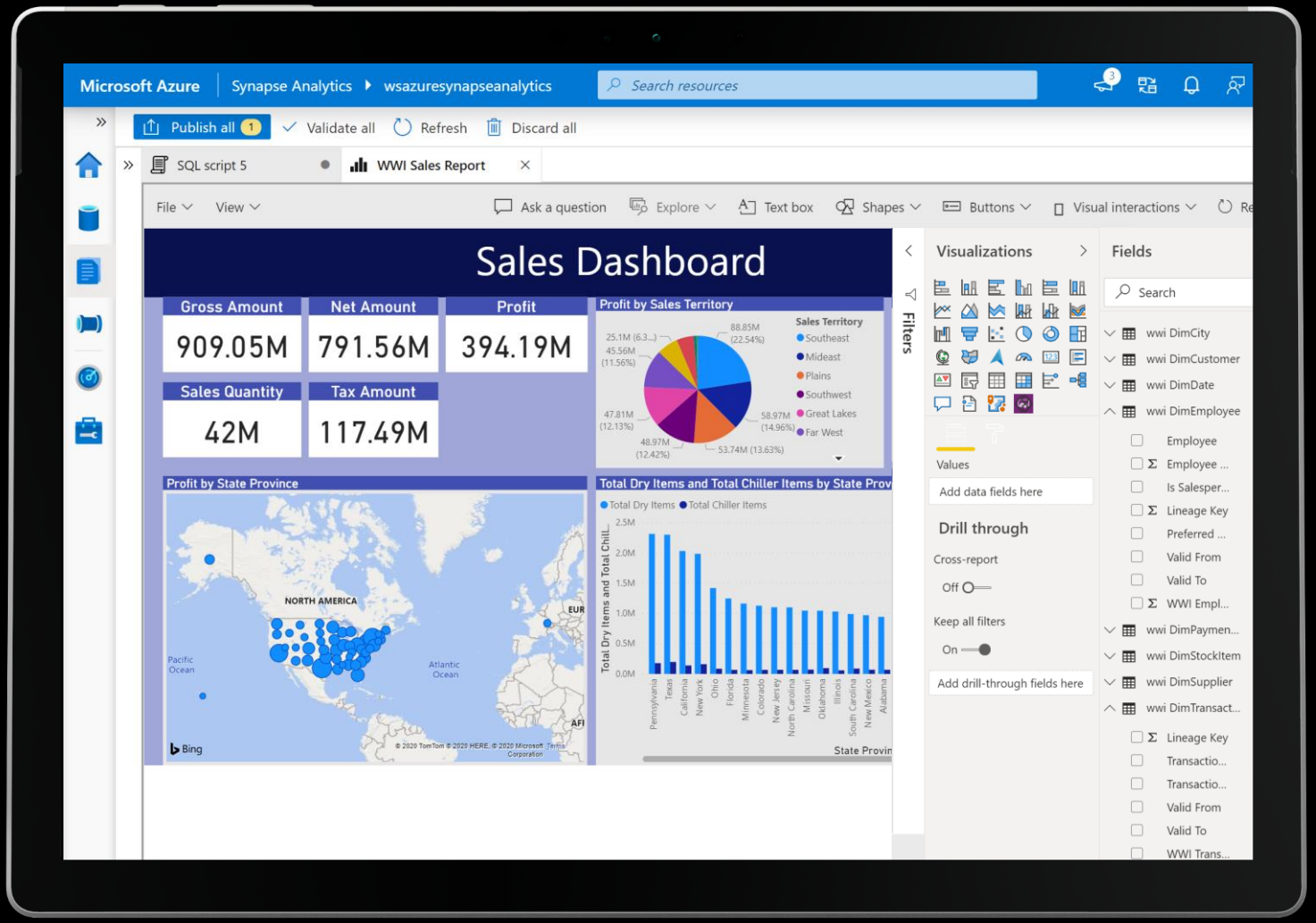


Power BI integration

Build dashboard in Synapse Studio

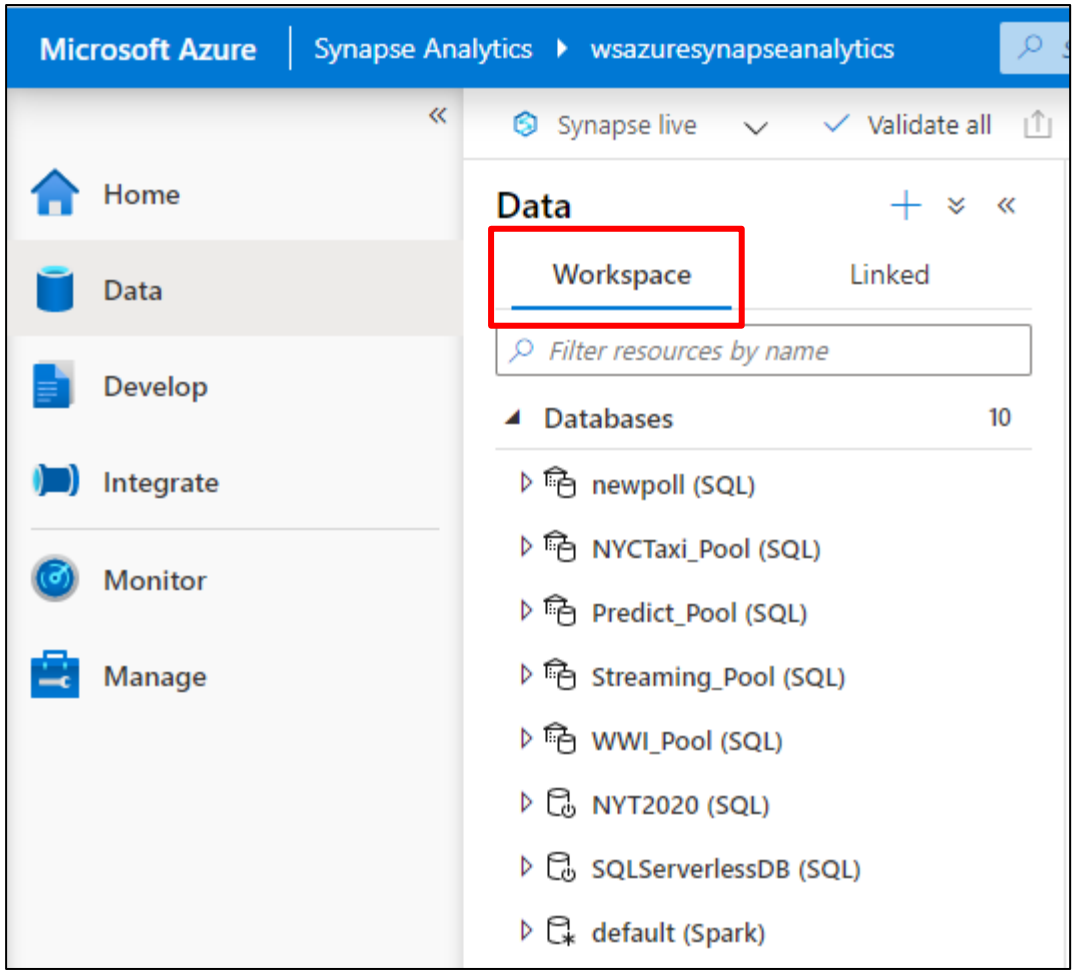
Code-free experience for development rich visualizations

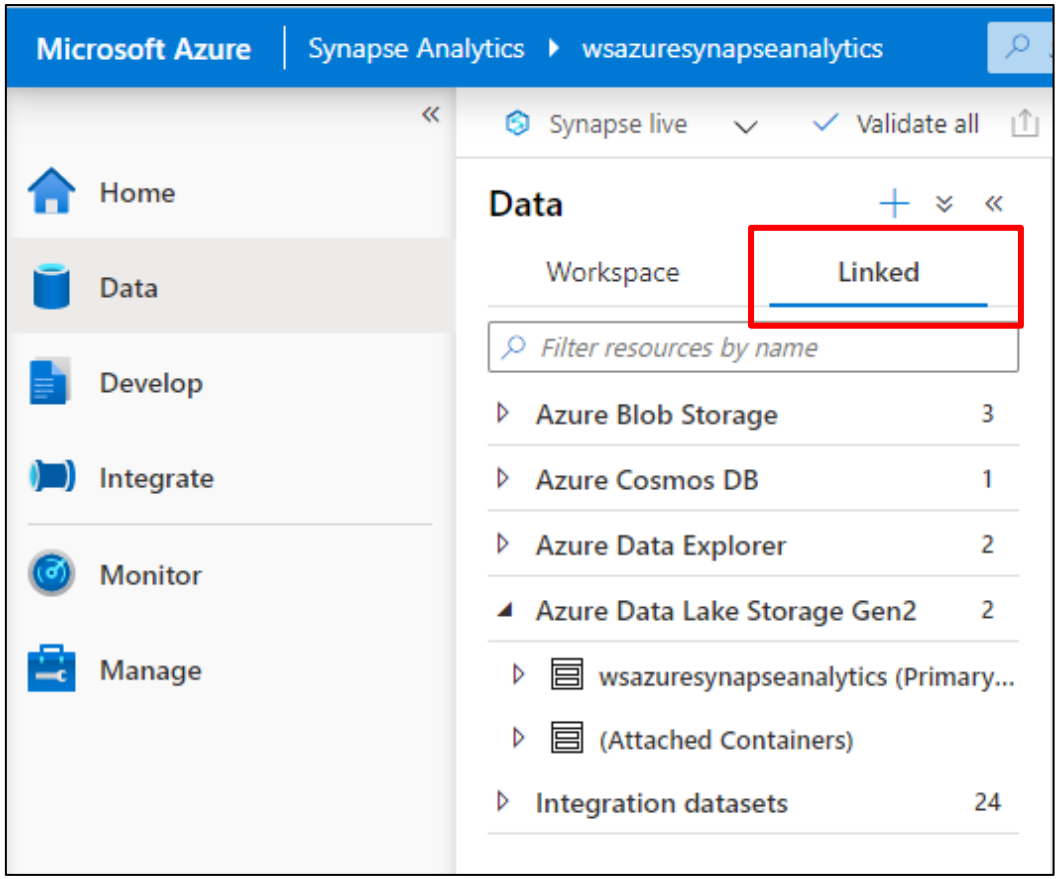
One-click publishing to for secure consumption across the enterprise



Data Hub

Explore data inside the workspace and in linked storage accounts





Data Hub – Linked Storage

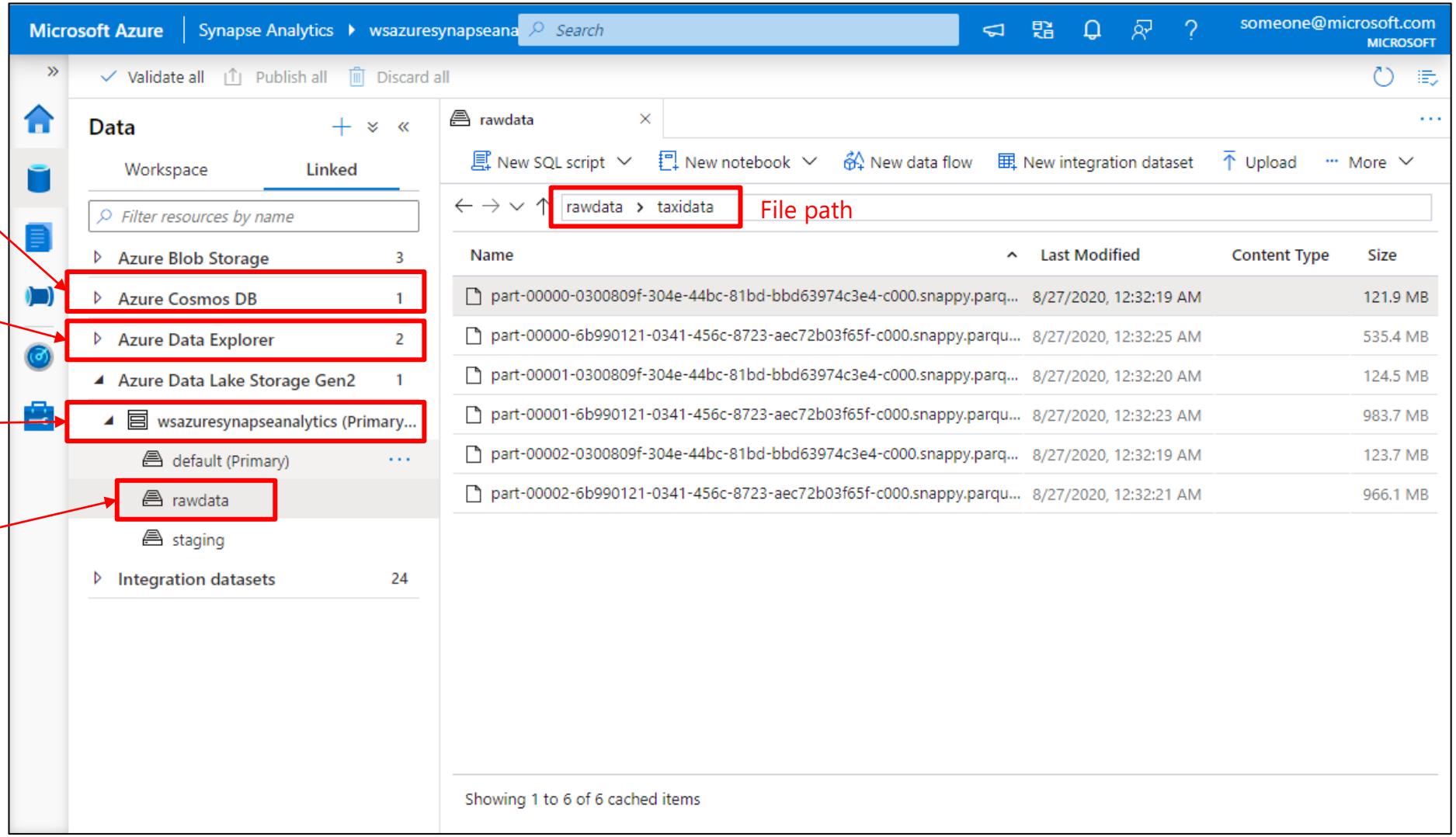
Browse Azure Data Lake Storage Gen2 accounts – filesystems, Azure Data Explorer – clusters, Azure Cosmos DB -containers

Linked Cosmos DB Analytical Store

Linked Azure Data Explorer

Linked ADLS Gen2 Account

Container (filesystem)



Comprehensive SQL functionality

Advanced storage system

- Columnstore Indexes
- Table partitions
- Distributed tables
- Isolation modes
- Materialized Views
- Nonclustered Indexes
- Result-set caching

T-SQL Querying

- Windowing aggregates
- Approximate execution (Hyperloglog)
- JSON data support
- Score machine learning models in ONNX format

Complete SQL object model

- Tables
- Views
- Stored procedures
- Functions

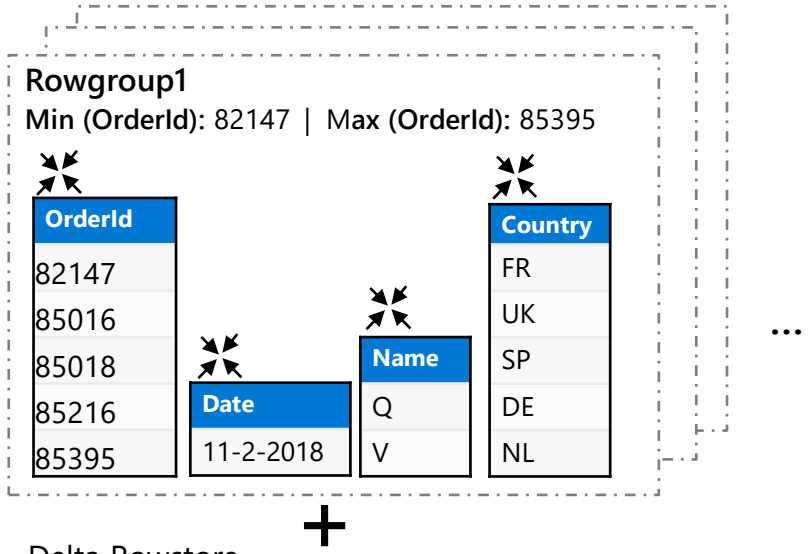
Synapse SQL (provisioned) Columnstore Tables

Logical table structure

OrderId	Date	Name	Country
85016	11-2-2018	V	UK
85018	11-2-2018	Q	SP
85216	11-2-2018	Q	DE
85395	11-2-2018	V	NL
82147	11-2-2018	Q	FR
86881	11-2-2018	D	UK
93080	11-3-2018	R	UK
94156	11-3-2018	S	FR
96250	11-3-2018	Q	NL
98799	11-3-2018	R	NL
98015	11-3-2018	T	UK
98310	11-3-2018	D	DE
98979	11-3-2018	Z	DE
98137	11-3-2018	T	FR
...	...	...	...

Clustered columnstore index

(OrderId)



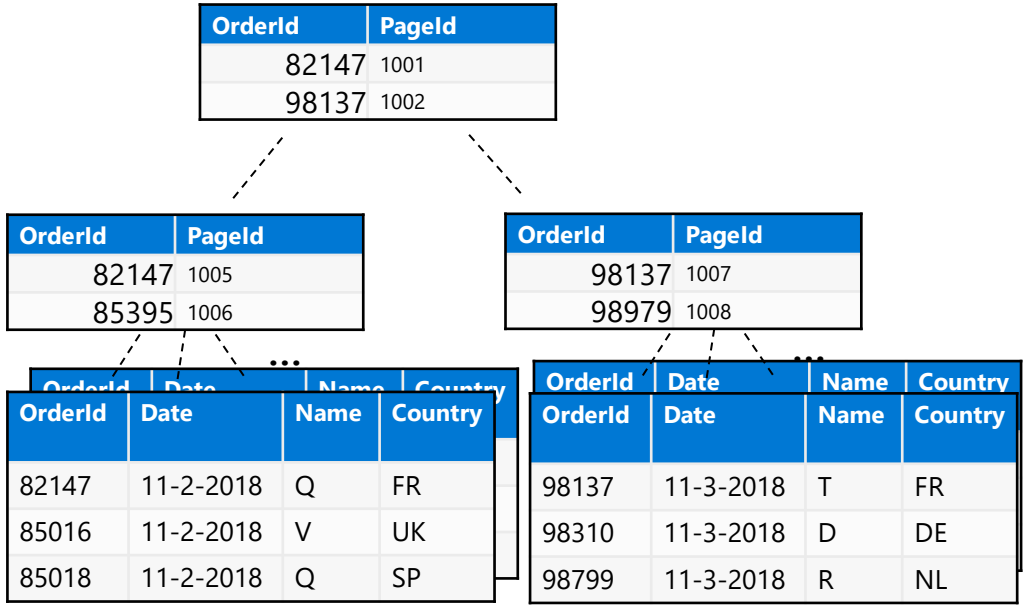
Delta Rowstore

OrderId	Date	Name	Country
98137	11-3-2018	T	FR
98310	11-3-2018	D	DE
98799	11-3-2018	R	NL
98979	11-3-2018	Z	DE

- Data stored in compressed columnstore segments after being sliced into groups of rows (rowgroups/micro-partitions) for maximum compression
- Rows are stored in the delta rowstore until the number of rows is large enough to be compressed into a columnstore

Clustered/Non-clustered rowstore index

(OrderId)



- Data is stored in a B-tree index structure for performant lookup queries for particular rows.
- Clustered rowstore index: The leaf nodes in the structure store the data values in a row (as pictured above)
- Non-clustered (secondary) rowstore index: The leaf nodes store pointers to the data values, not the values themselves

Security and compliance

Unified governance controls

Complete data protection

Best-in-class security

Customer & System Managed Keys

All data encrypted by default

Up to 3x levels of data encryption at rest

Democratize data at scale with fine-grained ACL

Proactive protection

Comprehensive Compliance

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	✓

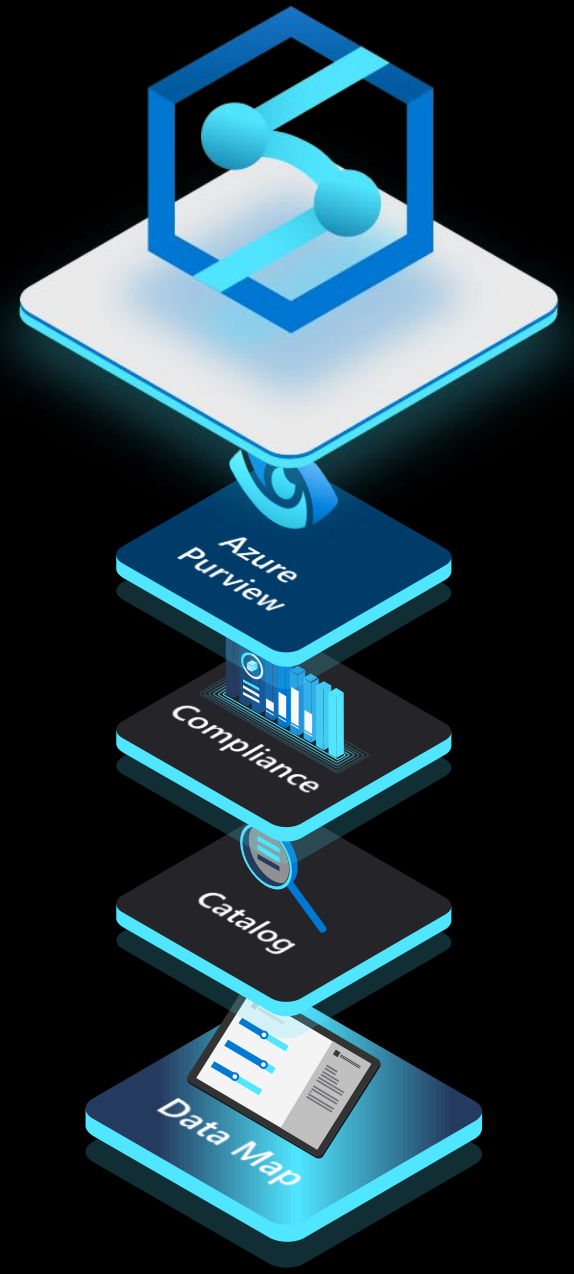
Integrated data governance

More than just data security

Native integration with Azure Purview

Automatically discover and classify data assets

End-to-end data lineage



Data warehouse

Scalable and secure analytics platform for SQL workloads

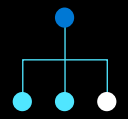
Suitable for any scale

Fully-managed elastic platform

Elastic compute that can be easily optimized to different classes of workload

All features available in a single tier

Infinite cost effective PAYG storage



Department

Gigabytes of data

10’s of users

Weekday usage

Organization

Terabytes of data

100’s of users

24/7 usage

Enterprise

Petabytes of data

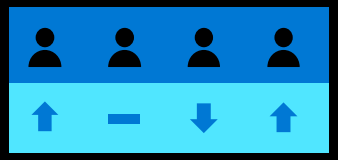
1000’s of users

Tier-0 Availability

Workload management

Scale in

Workload Importance

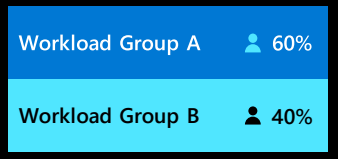


Data Warehouse

Benefits:

- Predictable cost
- Bias to high-value workloads within fixed/predictable budget
- Enables customers to easily deprioritize queries which don't need to be run immediately
- Built-in starvation prevention

Workload Isolation



Data Warehouse

Benefits:

- Predictable cost
- Efficient for unpredictable workloads since compute can overcommit
- No cache eviction for scaling (no performance cliff at restart)
- Single connection string for isolation (unlike Snowflake virtual warehouses)

Scale out

Elastic Cluster (Scale Up)

Data Warehouse

Benefits:

- Incremental add compute
- Increase large query performance
- Single cache for heterogeneous workloads
- Single endpoint

Azure Synapse supports a more diverse set of workload management tools through workload importance, intra-cluster isolation, and elastic clusters.

Machine learning

Empower everyone with predictive insights

Machine learning

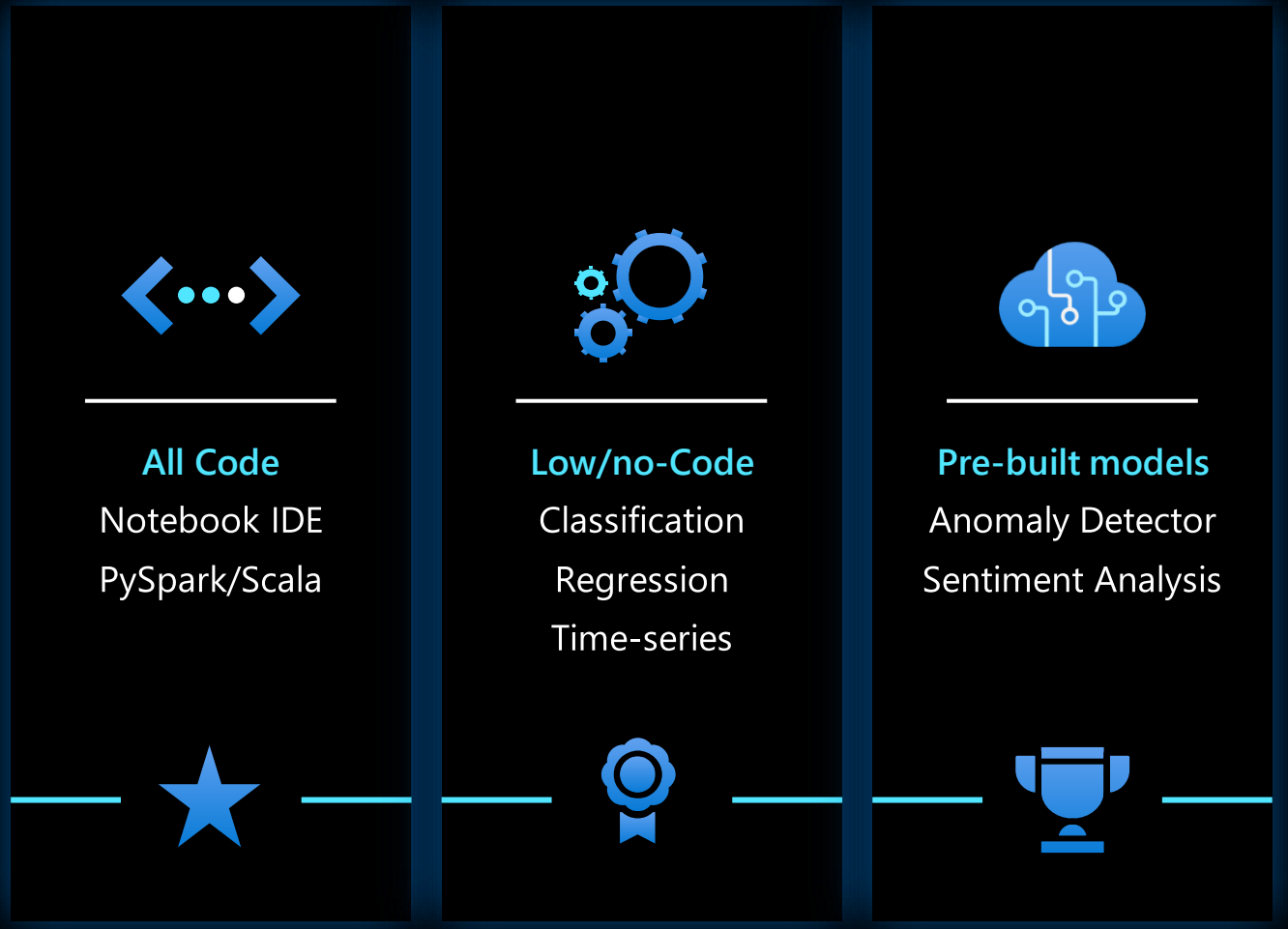
Democratize predictive power

Synapse makes predictive analytics accessible to all

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



Notebook IDE code authoring

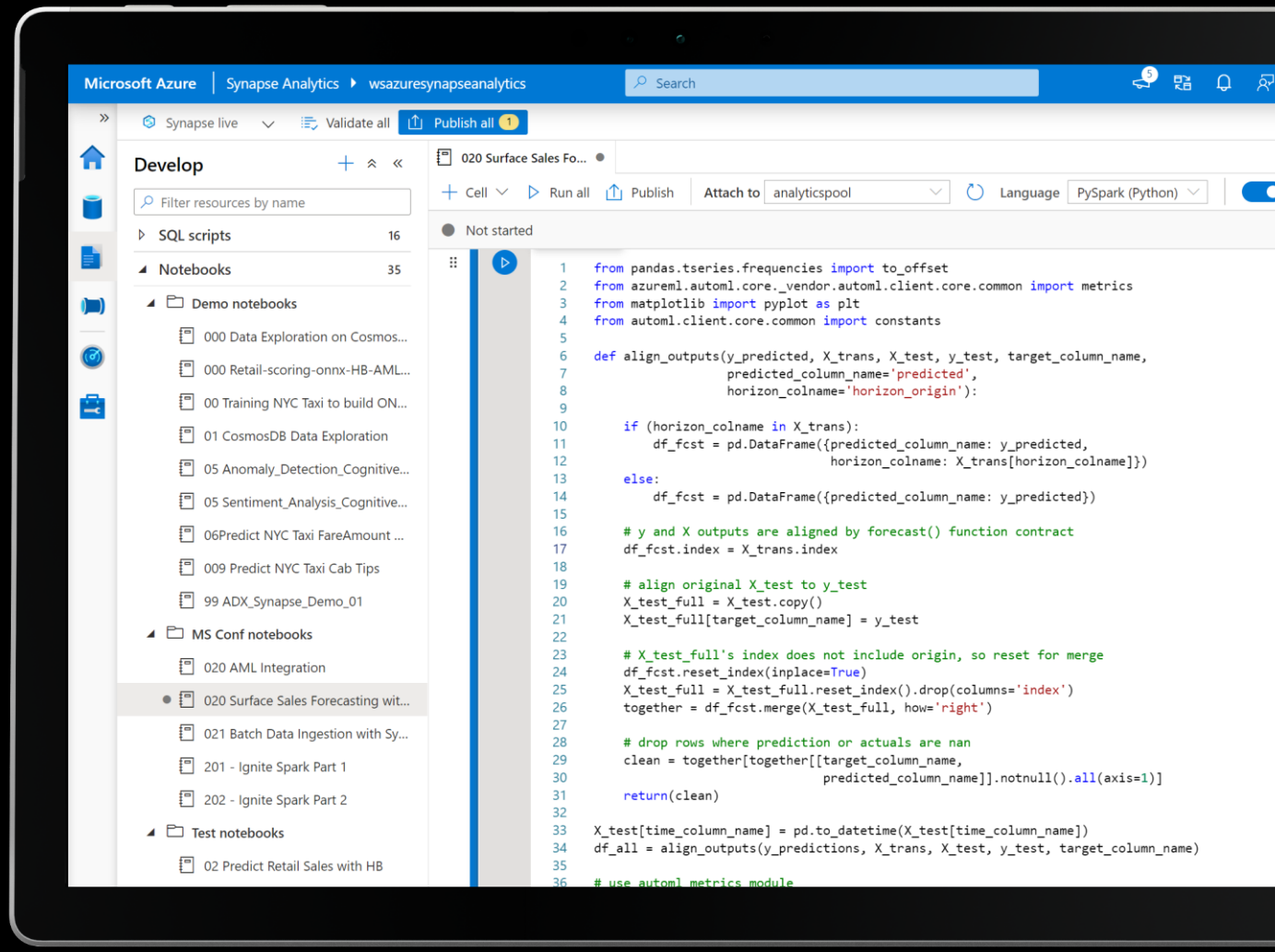
Code-first ML model development

PySpark, Scala, and C# languages supported

Automatic code completion (Intellisense)

Author multiple languages in a single notebook

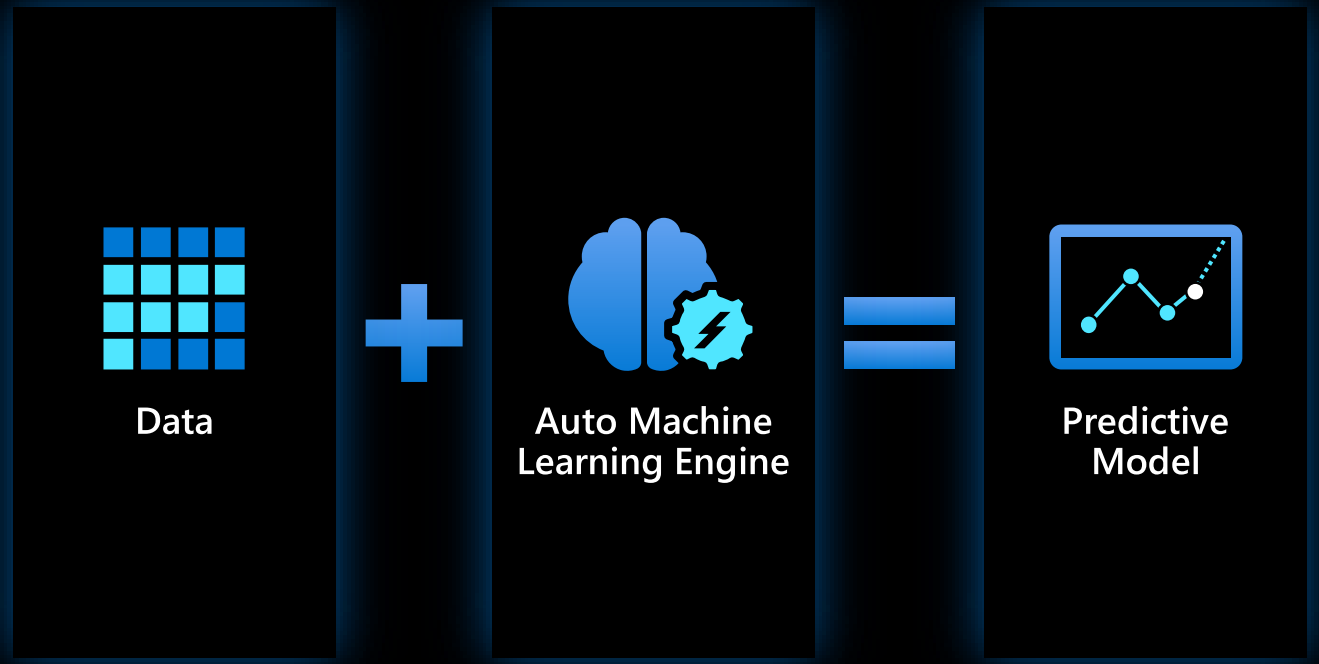
Analyze data from the data warehouse, data lake, and real-time operational data from one place



Automatic machine learning

All you need is data

Fully automated feature exploration



Automatic machine learning

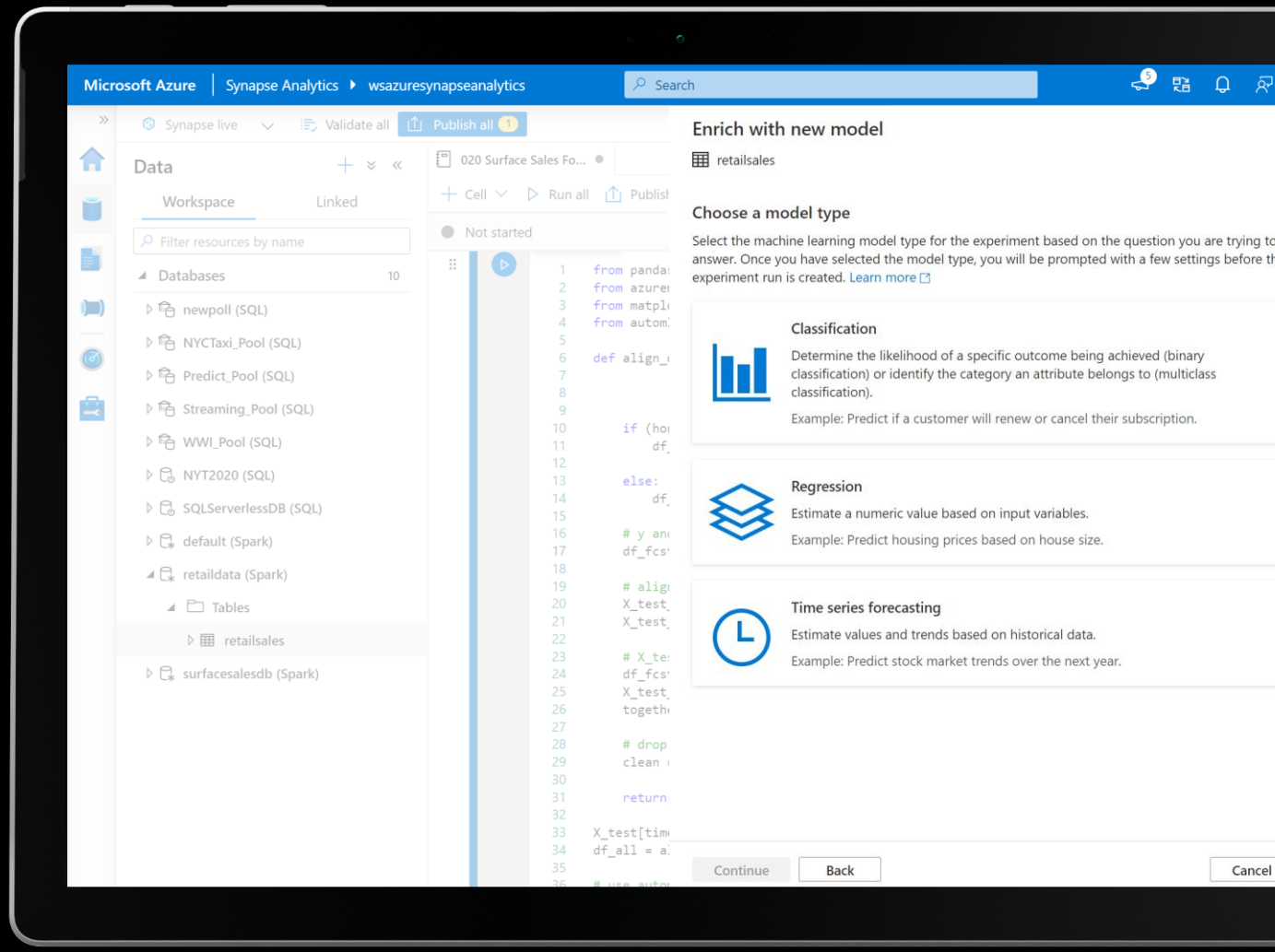
Code-free in Synapse Studio

No-code creation on Machine Learning models

Democratize ML to everyone since no data science domain knowledge required

Support for ensemble models

Supports classification, regression, and time-series forecasting



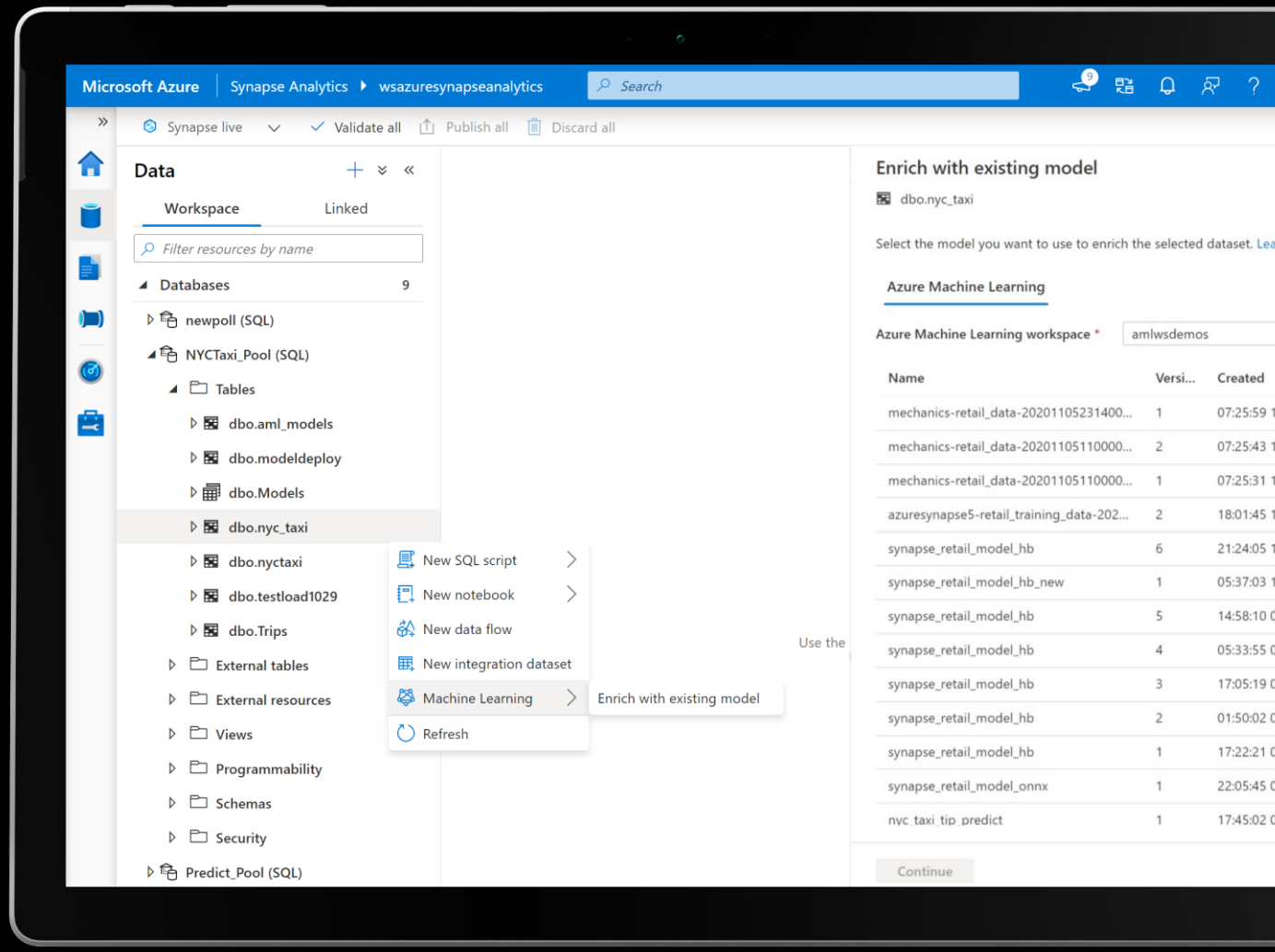
Code-free machine learning scoring

Code-free in Synapse Studio

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



Democratize predictions to all

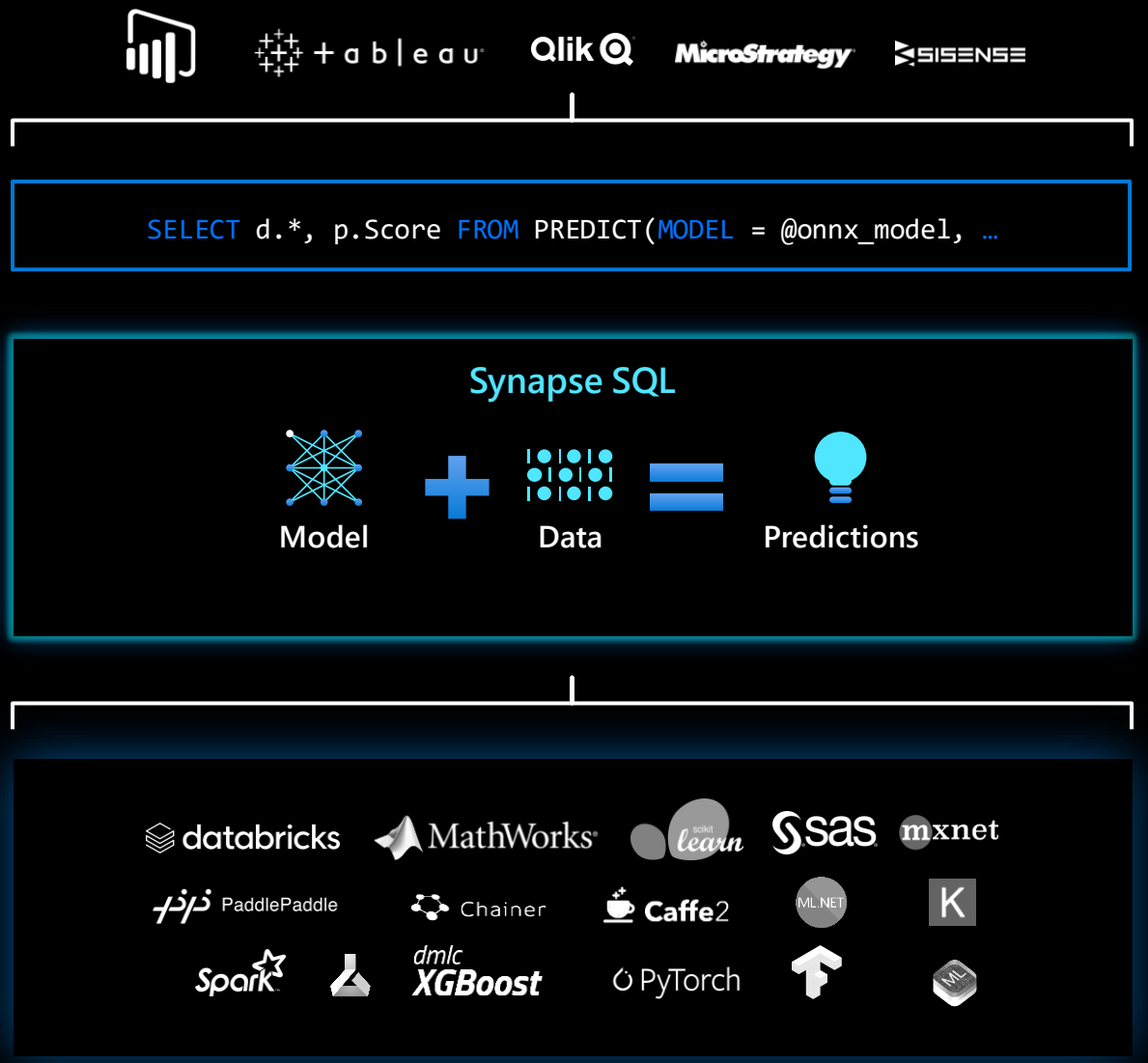
In-engine ML scoring

Machine learning models executed using SQL

"In-engine" for performance and scalability

No data leaves the platform for scoring

No additional cost for scoring



Get started today

Create a free Azure account and get started with Azure Synapse Analytics:
https://azure.microsoft.com/en-us/free/synapse-analytics/

Submit info for free proof of value package for modernizing SAP workloads:
https://aka.ms/synapse-qlik

Submit info for free proof of value package for migrating on-prem data warehouse:
http://aka.ms/synapse-informatica

Learn more:
https://aka.ms/synapse

© Copyright Microsoft Corporation. All rights reserved.