SQLBITS

Overview of Azure Synapse Analytics

Insert coins to start

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

SQLBITS 2022

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Dave Ruijter

Data Platform MVP | Azure Solution Architect Data & Analytics Consultant

Data & analytics all-rounder with a strong technical focus. Enjoys facilitating workshops and training sessions to share knowledge and build other people's skills.

dave@blue-rocket.it

@DaveRuijter

linkedin.com/in/DaveRuijter

ModernData.ai

MVP Microsoft® Most Valuable Professional

Feedback

Feedback

Feedback

https://sqlb.it/?6952

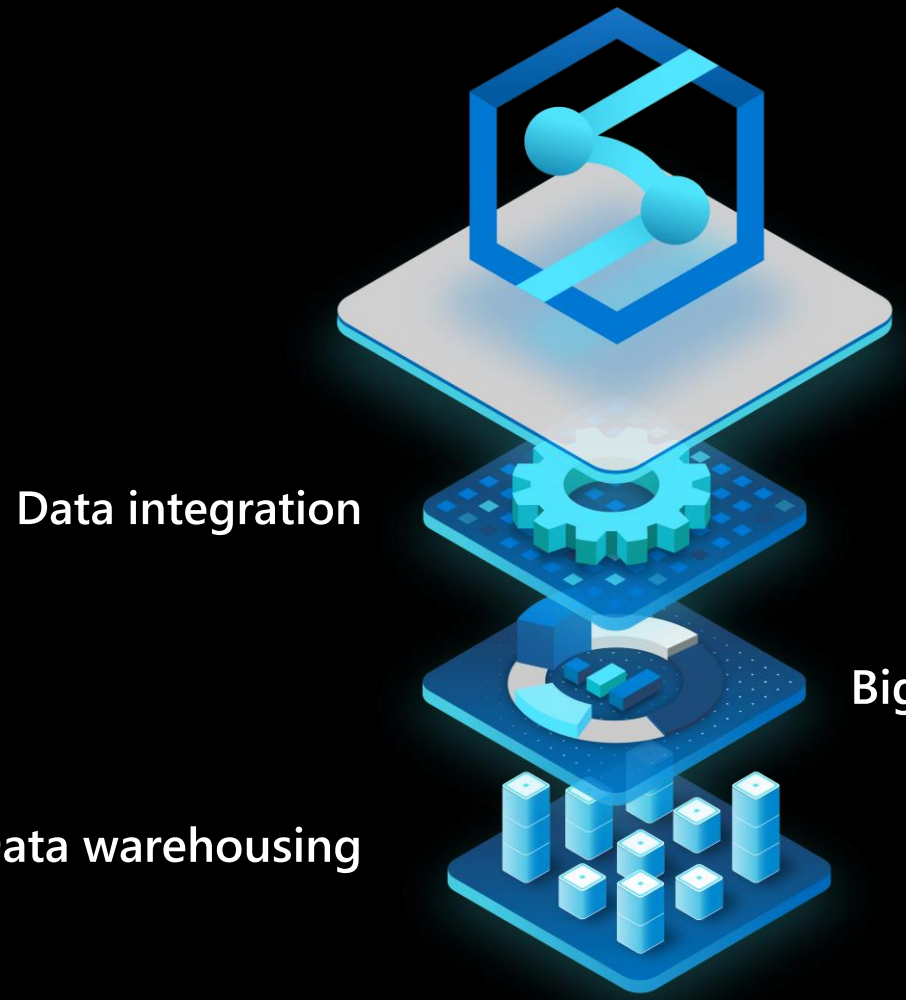
BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Data-Marc.com

Azure Synapse Analytics

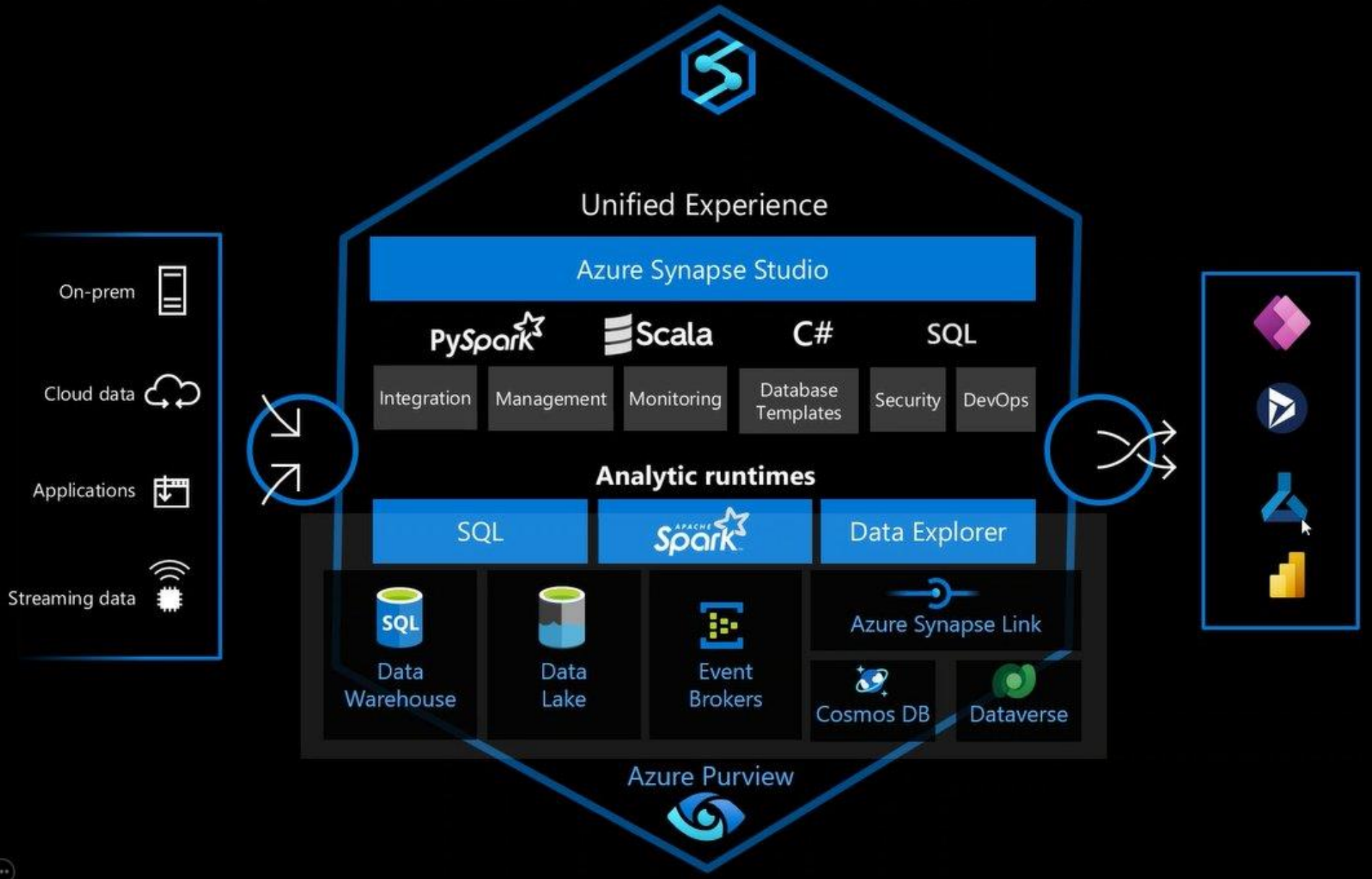
The first unified, cloud native platform for converged analytics

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



Azure Synapse Analytics

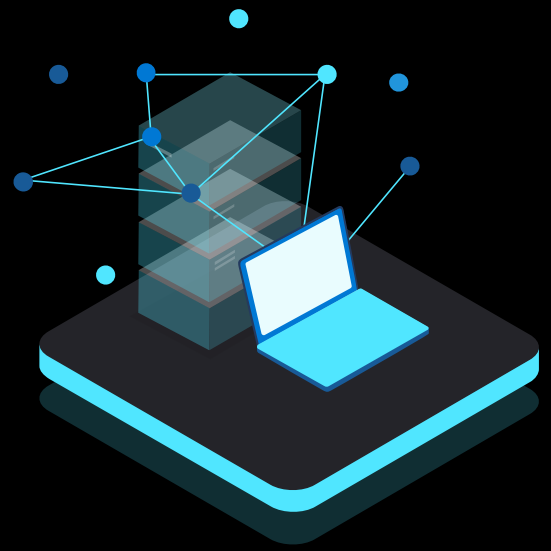
Powered by a new cloud native distributed SQL engine

Serverless + dedicated SQL

Flexible consumption models

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

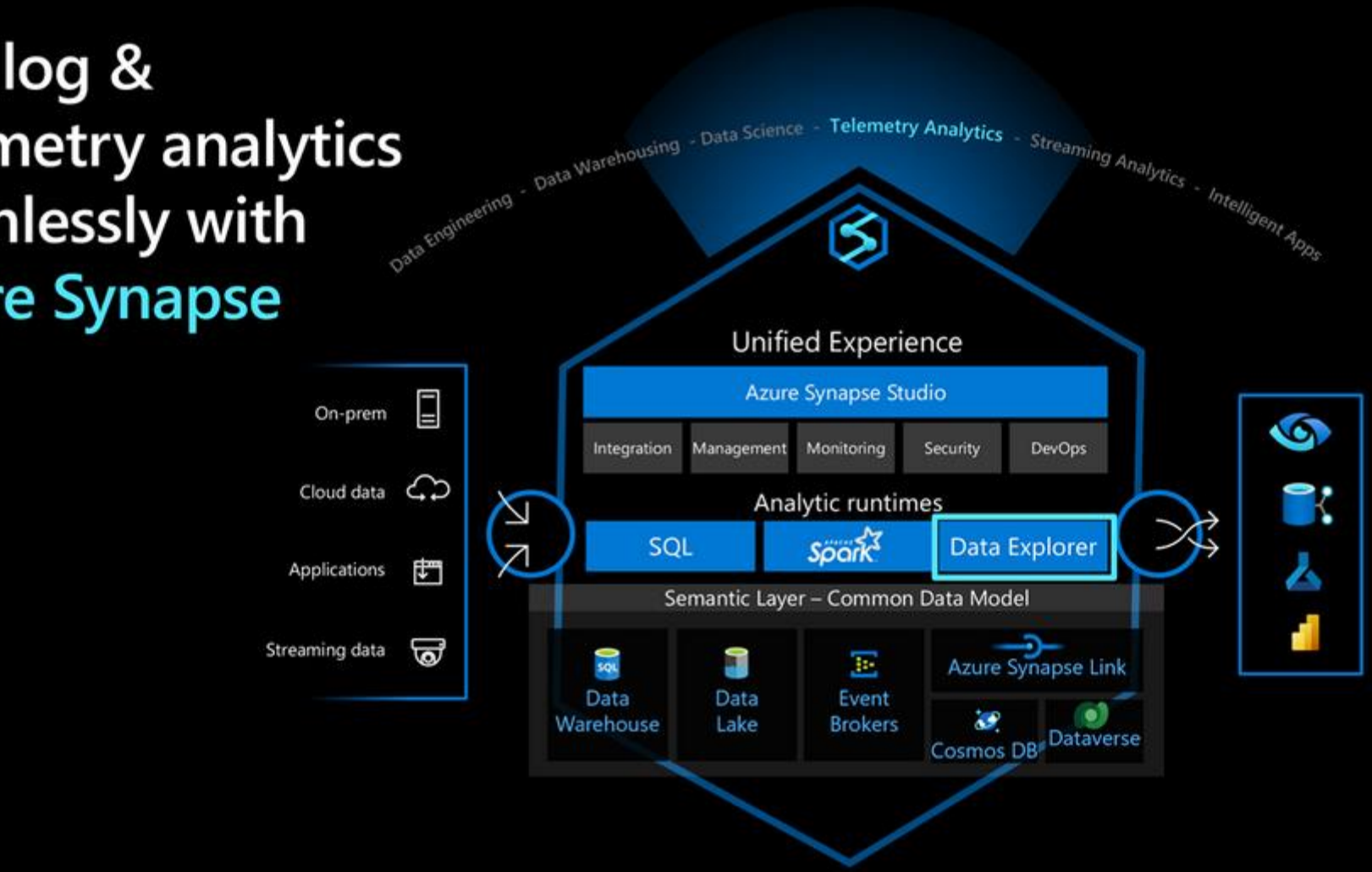
Dedicated clusters optimized mission-critical data warehouse workloads



Serverless

Dedicated

Run log & telemetry analytics seamlessly with Azure Synapse



Let’s dive deeper

1. Security and compliance
2. Data Integration
3. Flexible data warehousing
4. Azure Synapse Link
5. Synapse Data Explorer
6. Integrated machine learning
7. Purview integration
8. Lake Database templates
9. Power BI + Azure Synapse
10. Pricing
11. Best-practices
12. Comparison | what to use when

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	✓

Managed virtual networks

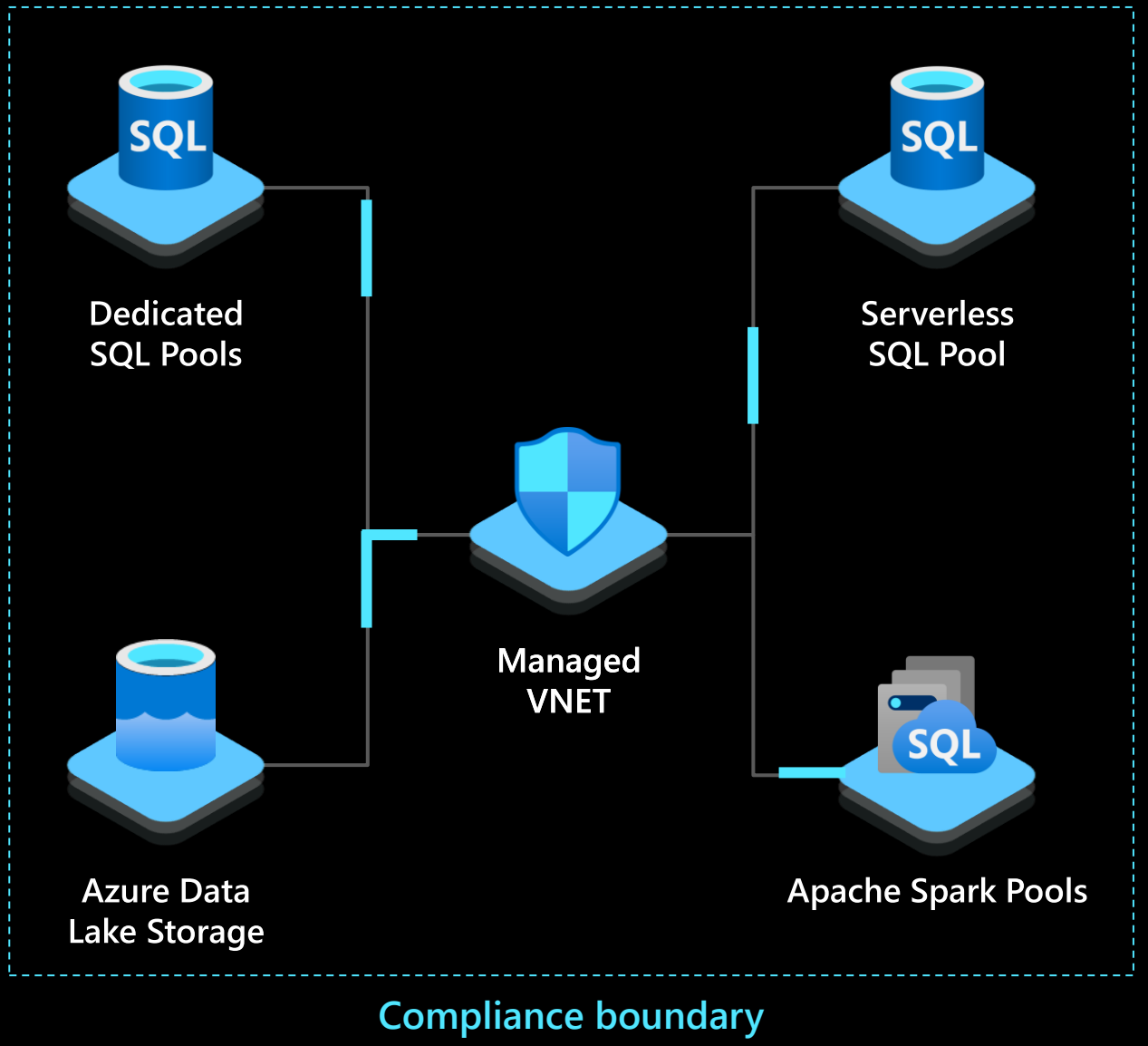
Eliminate network maintenance

One-click enables automated management of virtual networks between cluster endpoints

Synapse resources only ever interop with private endpoints

No management of subnets or IP Ranges

Prevents data exfiltration



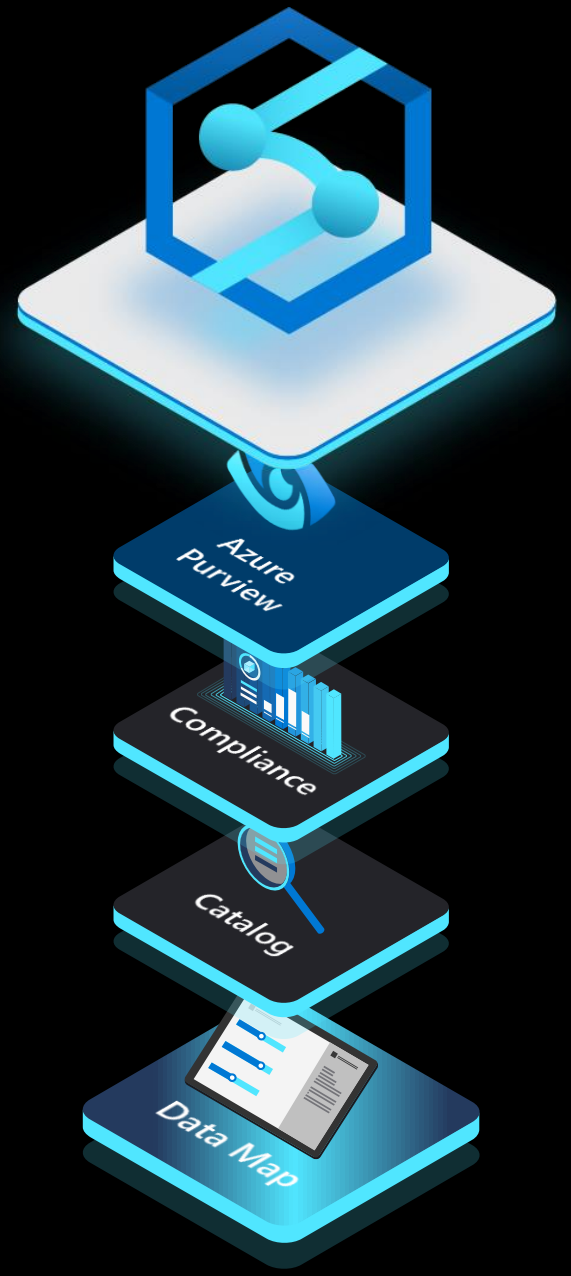
Integrated data governance

More than just data security

Native integration with Azure Purview

Automatically discover and classify data assets

End-to-end data lineage



Data integration

Hybrid data integration

Cloud native ETL/ELT

95+ connectors available

Secure connectivity to on-premise data sources, other clouds, and SaaS applications

Code-first and low/no code design interfaces

Schedule and Event based triggering

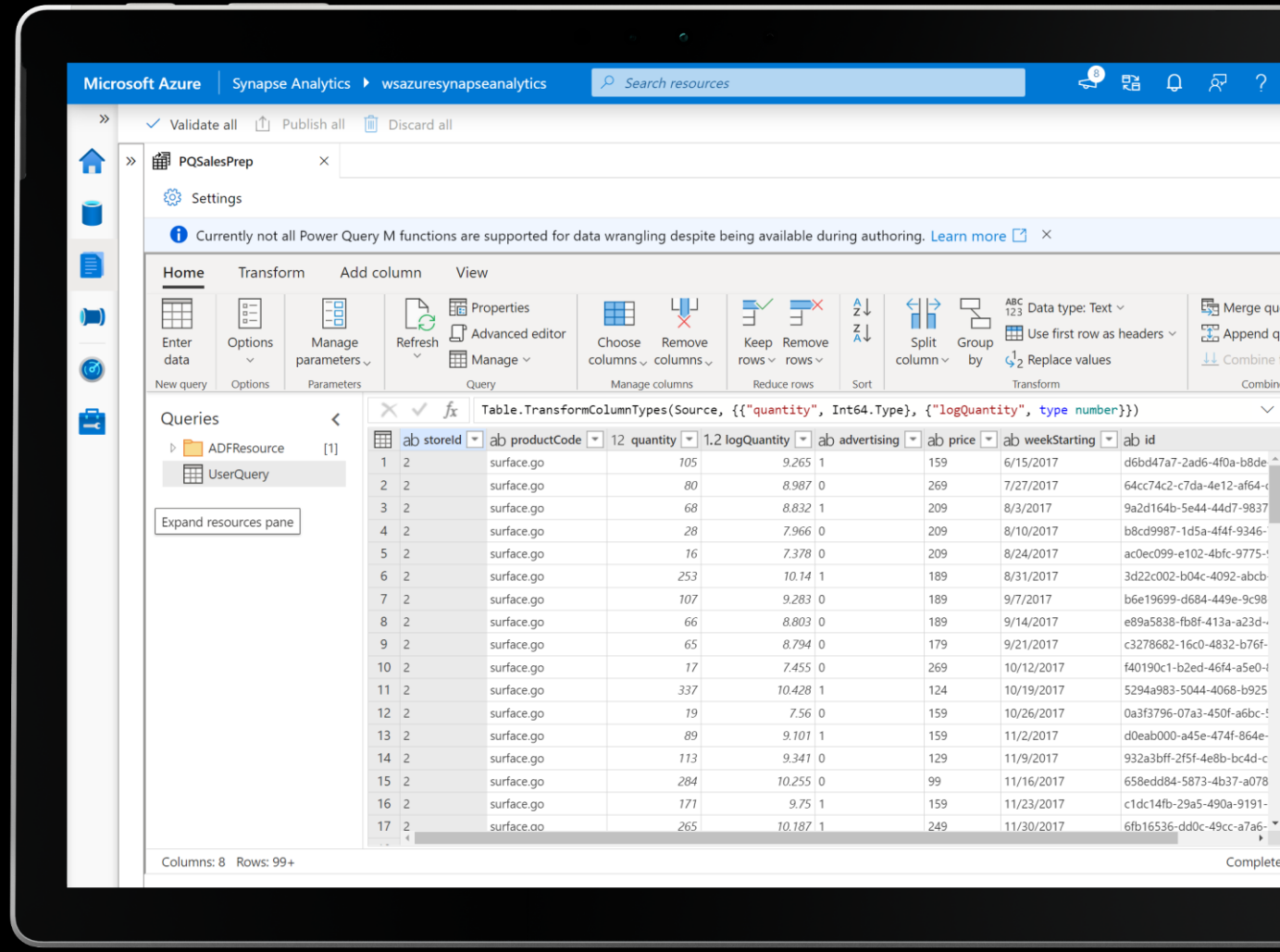
Code-free data wrangling

No/low-code data transformation

Excel-like interface is familiar to users

Transform data to desired shape completely visually

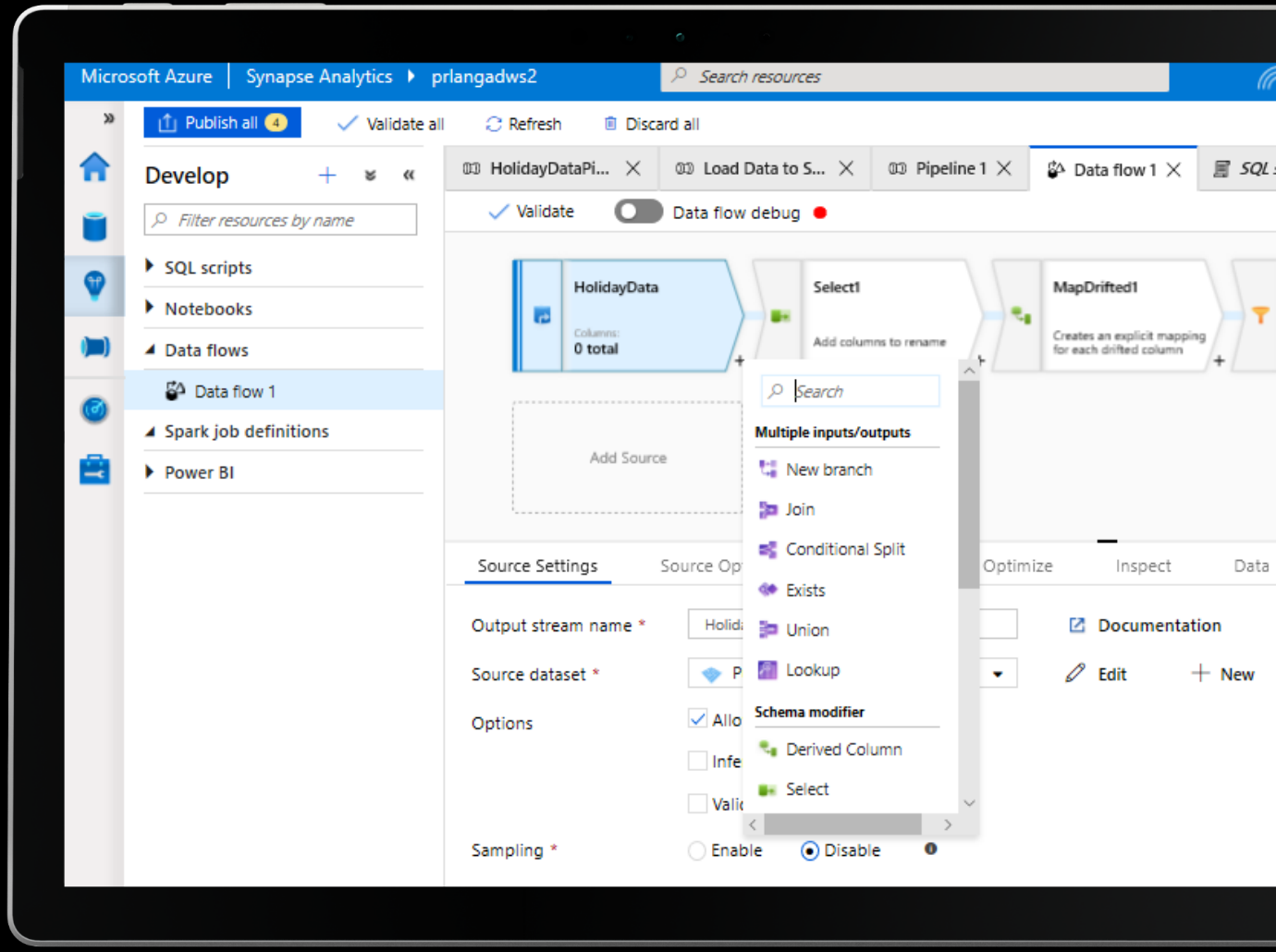
Operationalize into pipelines



BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Hybrid data integration

Data Flows



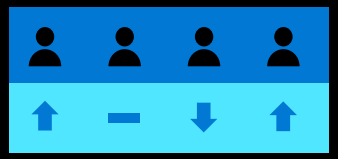
Data warehouse

Scalable and secure analytics platform for SQL workloads

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.

Synapse Studio

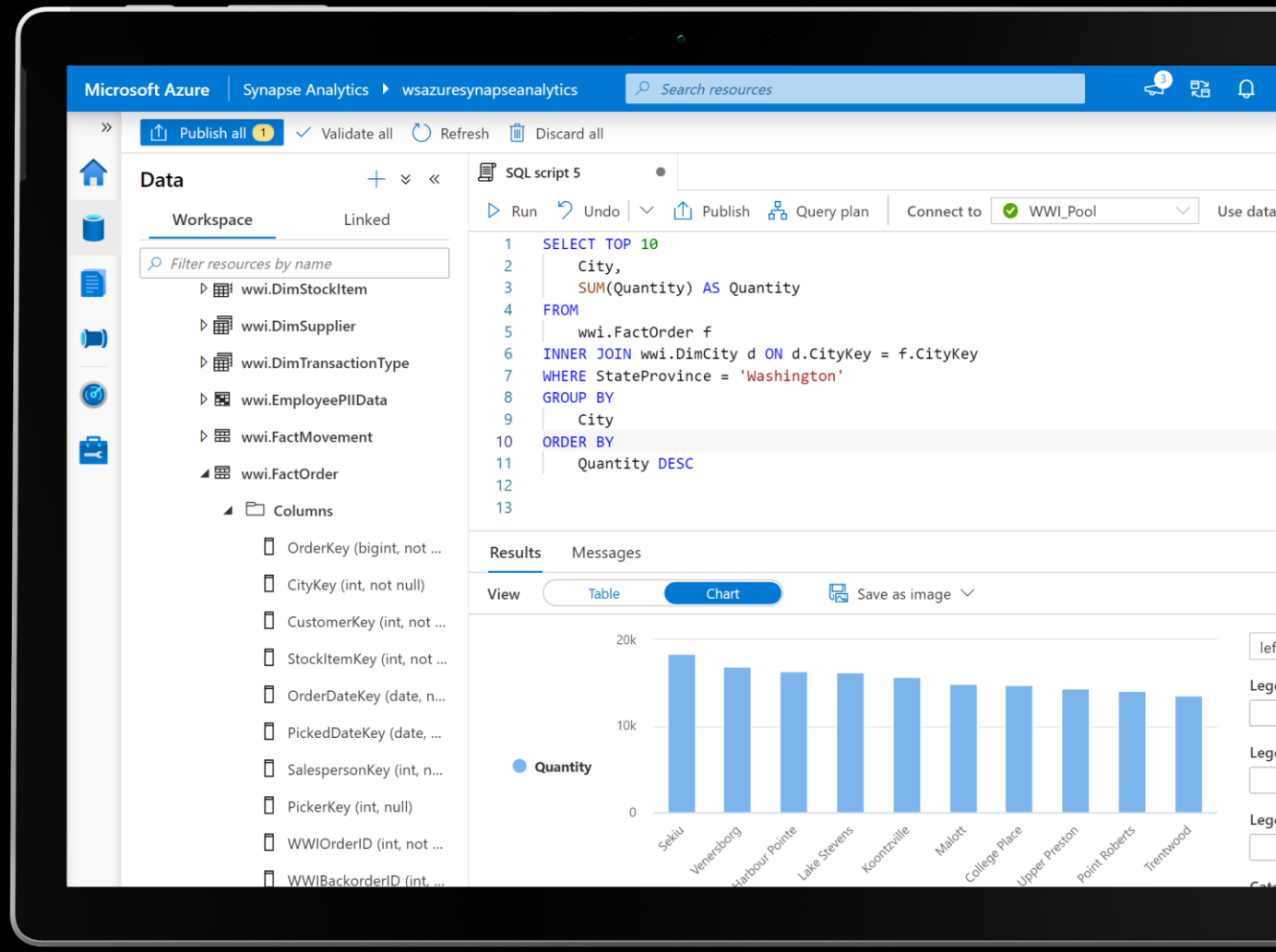
SQL Editor

Automatic code completion (Intellisense)

Script collaboration within the Workspace

Built-in visualizations

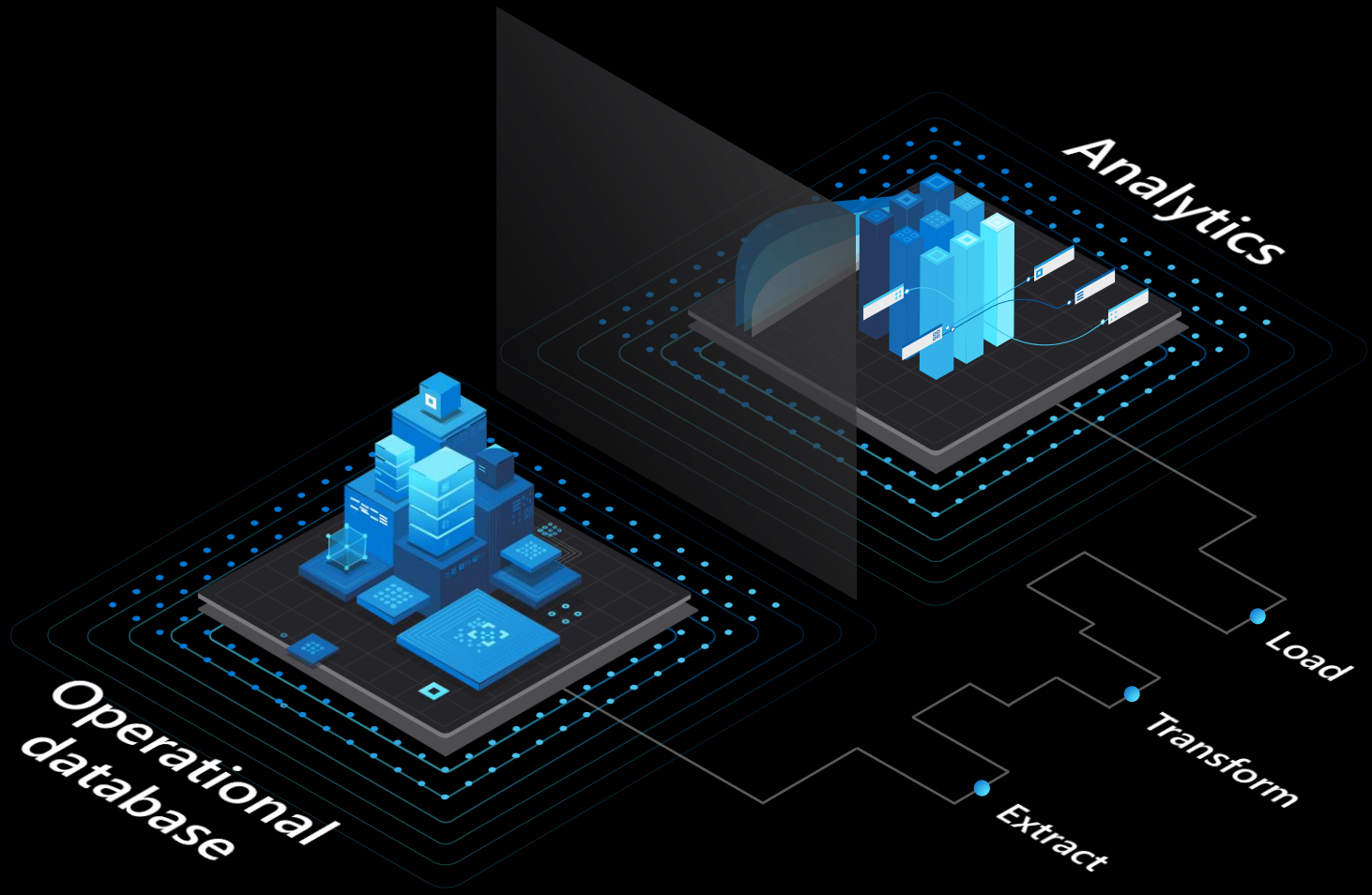
Easily switch between clusters



Real-time operational analytics

Eliminate latency and accelerate decision making

Integrating operational data with analytics systems



Integrating operational data with analytics systems

Azure Synapse Link

Real-time data analytics

No ETL required

No performance impact on transactions

Azure Synapse Link for Dataverse

General Availability

Azure Synapse Link for Azure Cosmos DB

Custom partitioning in Private Preview

Azure Synapse Link for SQL Server 2022

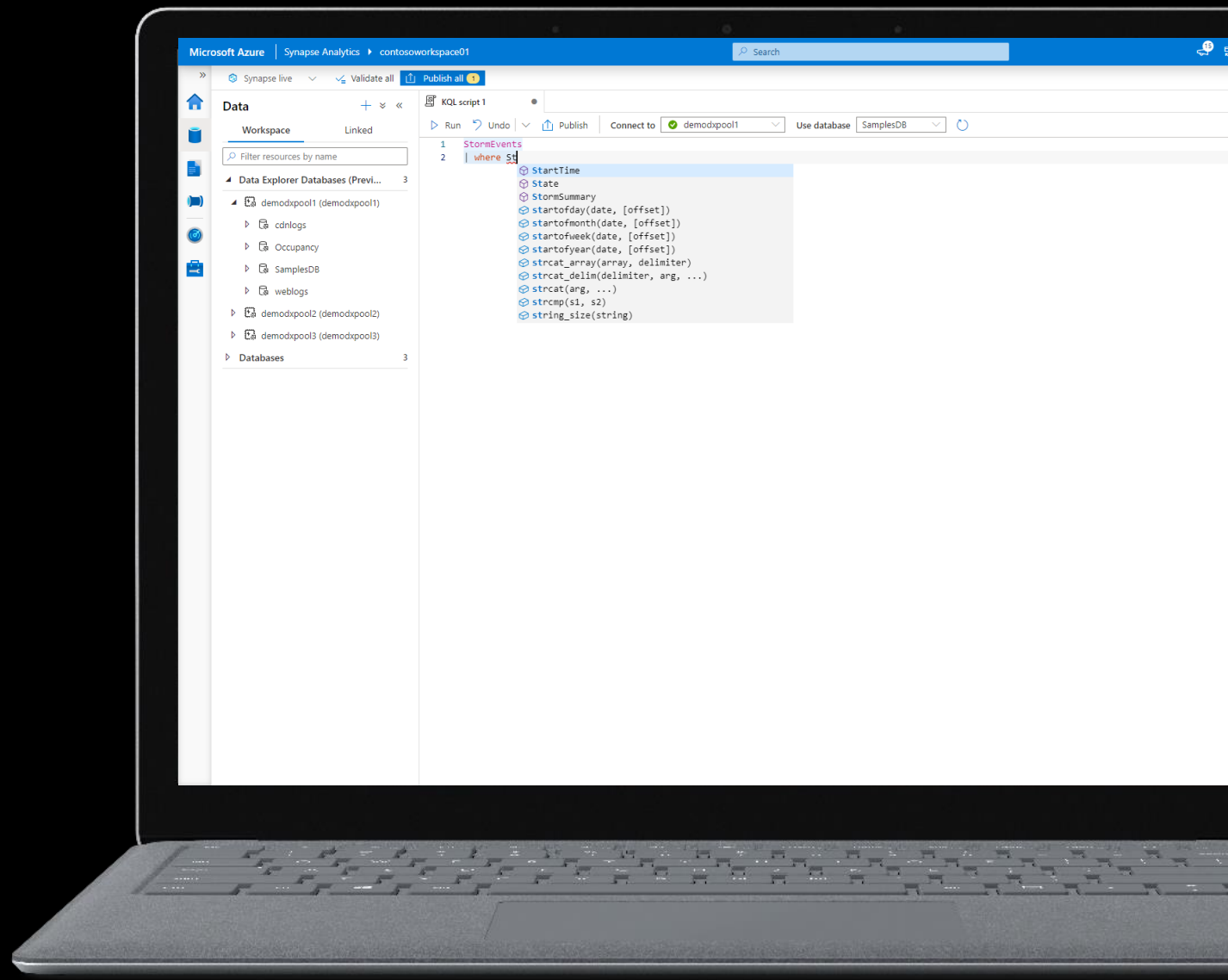
SQL Server 2022 CTP 1.0 – Coming Soon

New

Synapse data explorer

Build events, log, time series, and telemetry analytics solutions

- New analytics runtime engine tightly integrated in Synapse
- Industry leading free-text and semi-structured data indexing
- Kusto Query Language (KQL) optimized for telemetry workload



Scenarios

Building near-real-time Big Data analytics on custom event / logs data

Centralized near-real-time observability solutions from all telemetry sources

Security log analytics

IoT analytics solutions

Machine learning

Empower everyone with predictive insights

Notebook IDE code authoring

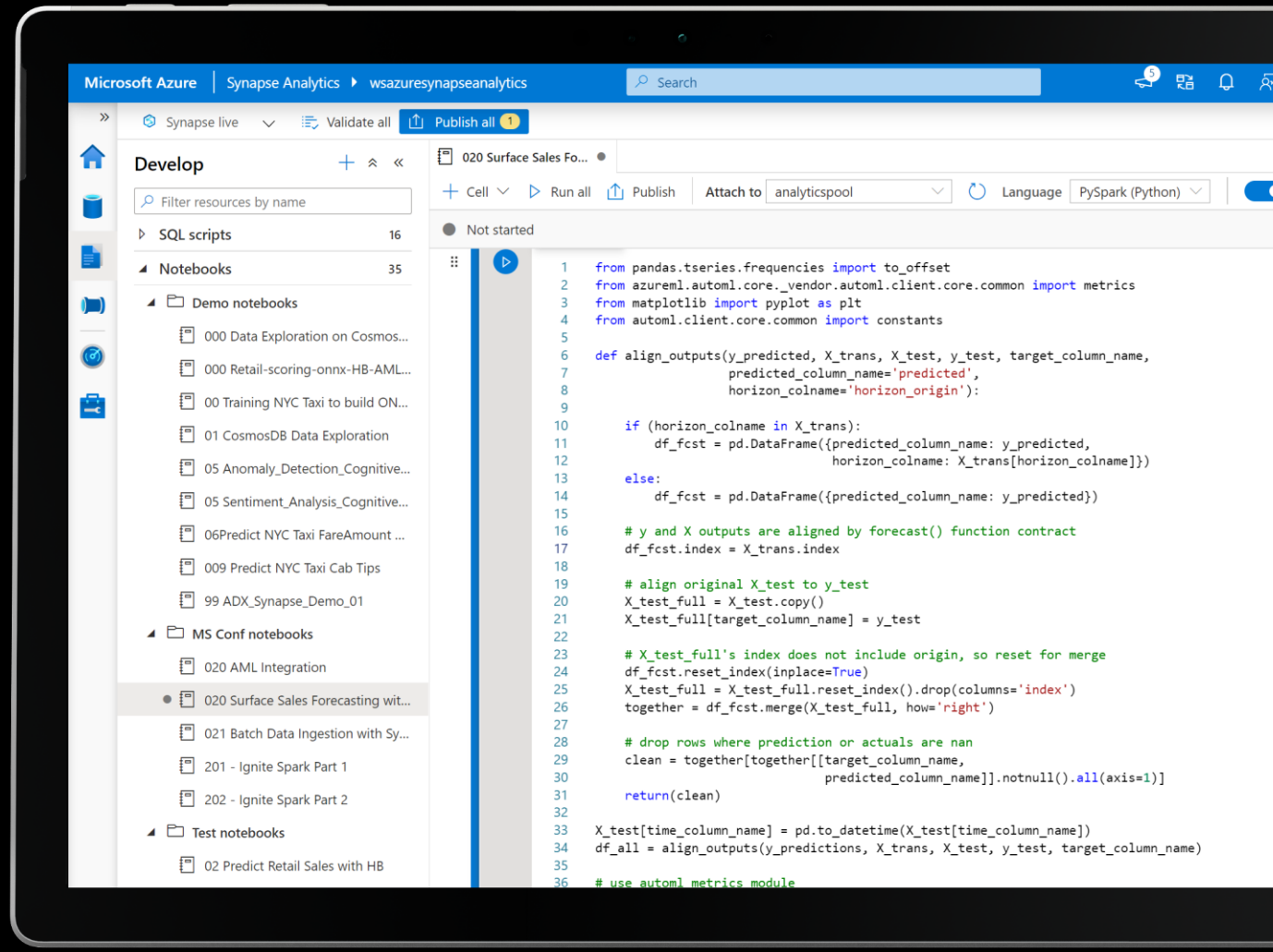
Code-first ML model development

PySpark, Scala, and C# languages supported

Automatic code completion (Intellisense)

Author multiple languages in a single notebook

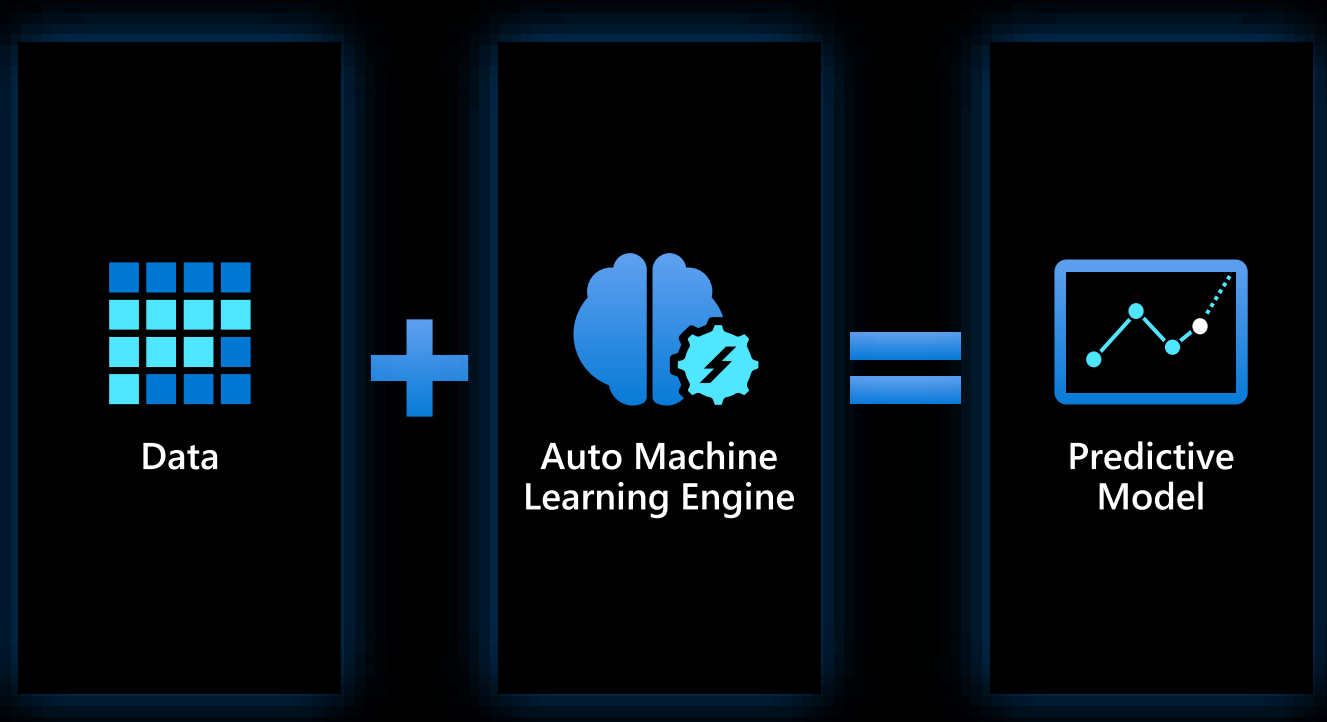
Analyze data from the data warehouse, data lake, and real-time operational data from one place



Automatic machine learning

All you need is data

Fully automated feature exploration



Automatic machine learning

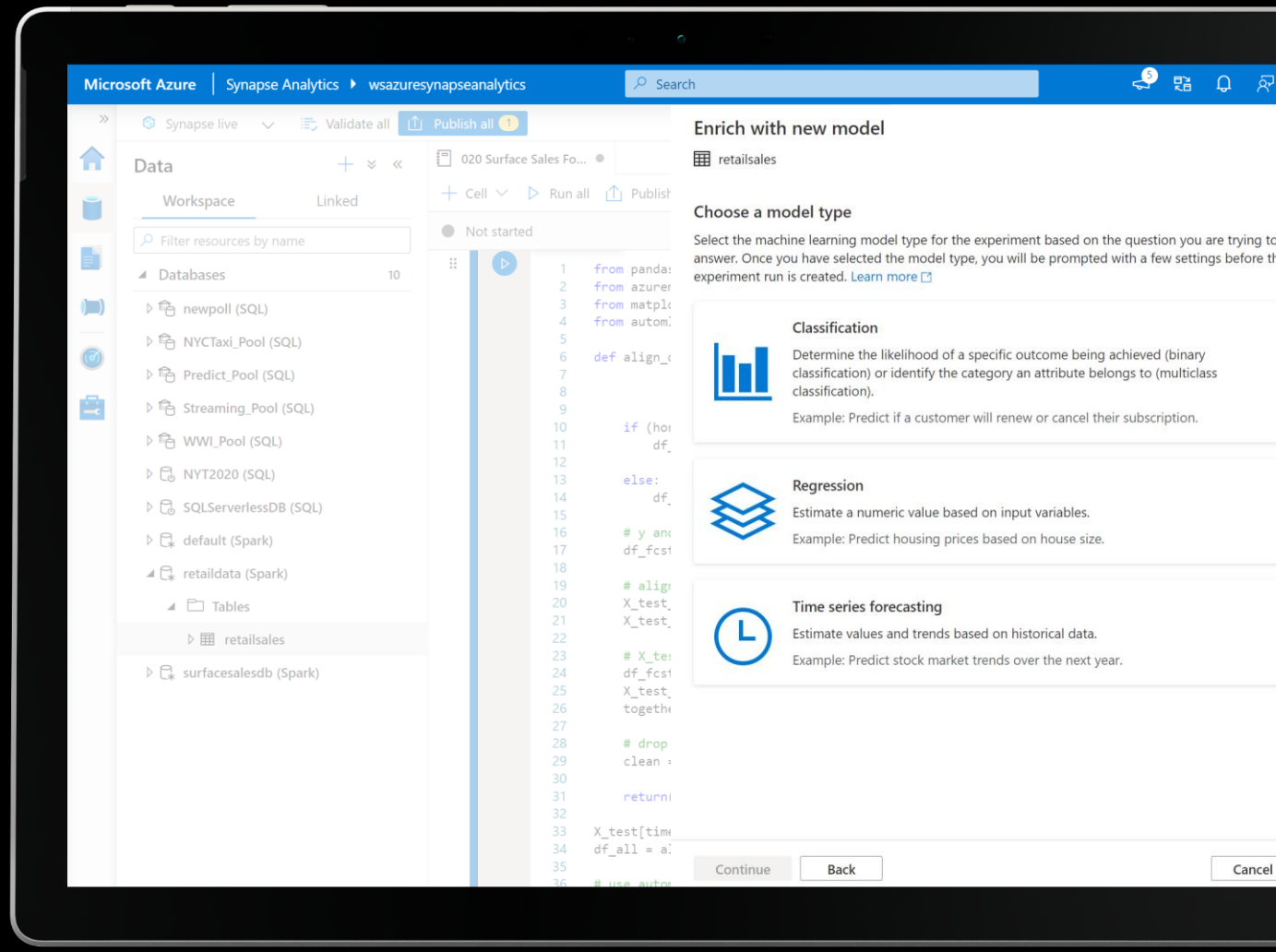
Code-free in Synapse Studio

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



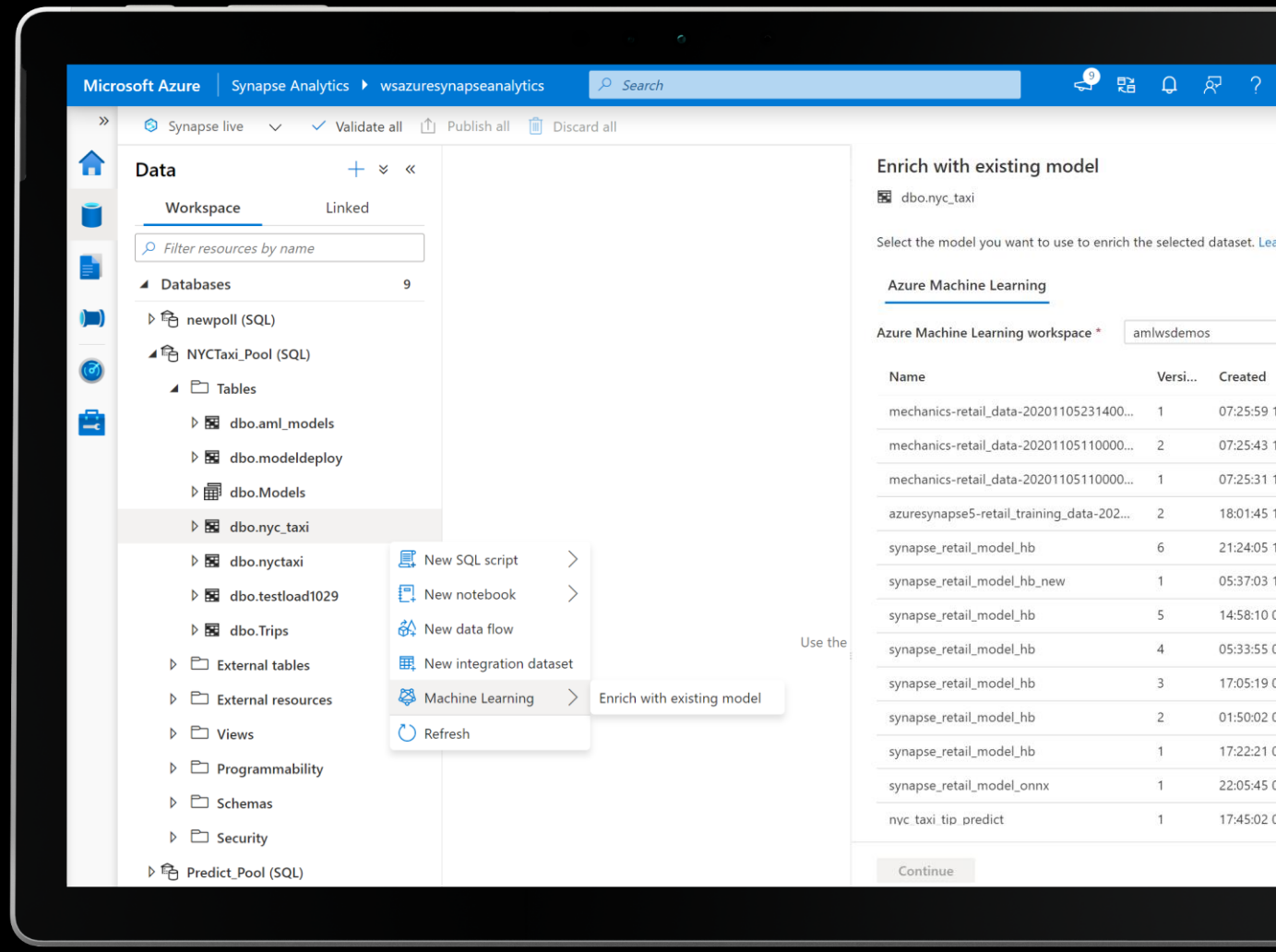
Code-free machine learning scoring

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



BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Democratize predictions to all

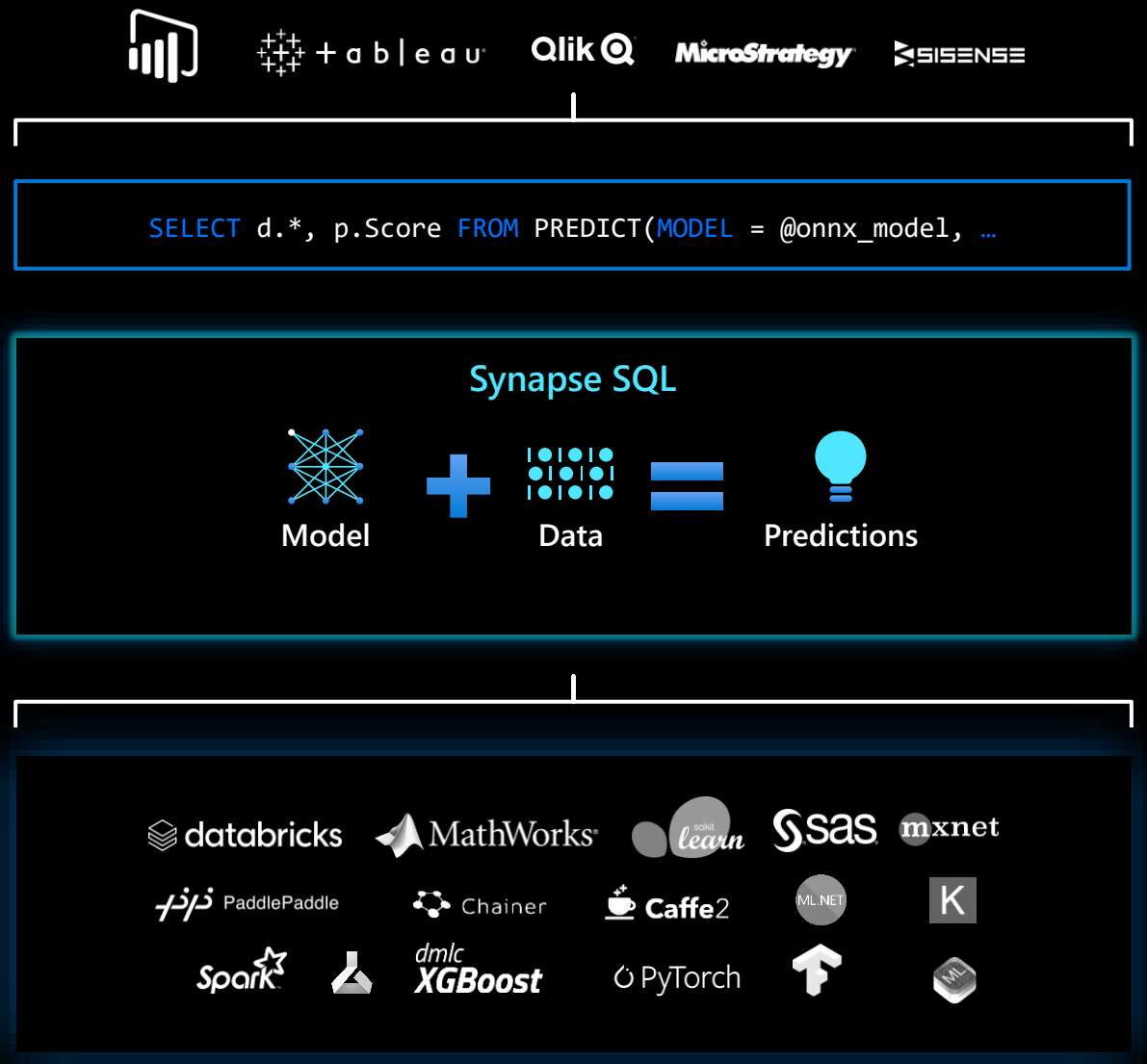
In-engine ML scoring

Machine learning models executed using SQL

"In-engine" for performance and scalability

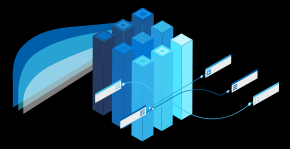
No data leaves the platform for scoring

No additional cost for scoring



Azure Purview

Unified data governance



Create a unified map of your data across hybrid sources

Discover data that powers business insights

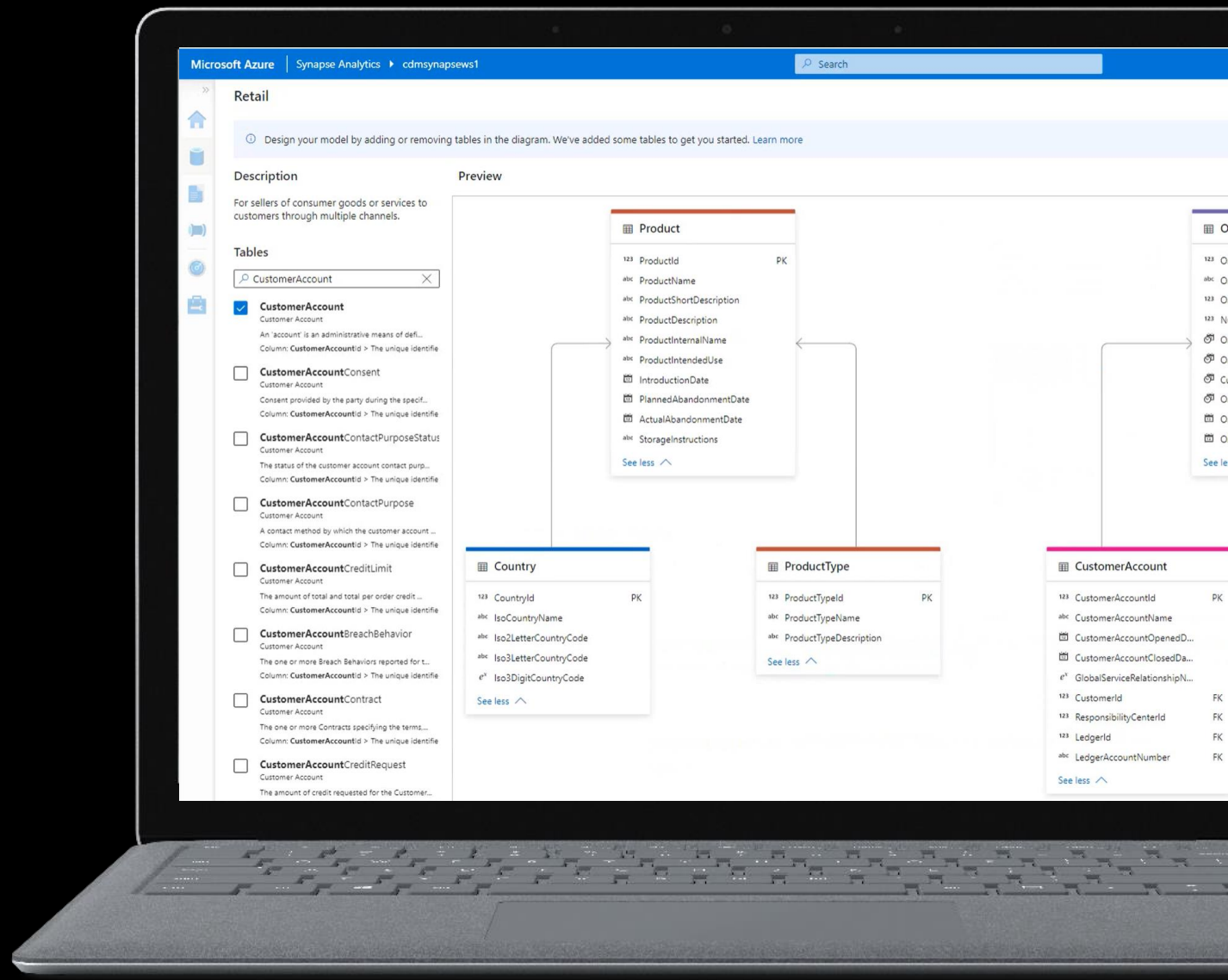
Gain insight into sensitive data across your organization

Announcing

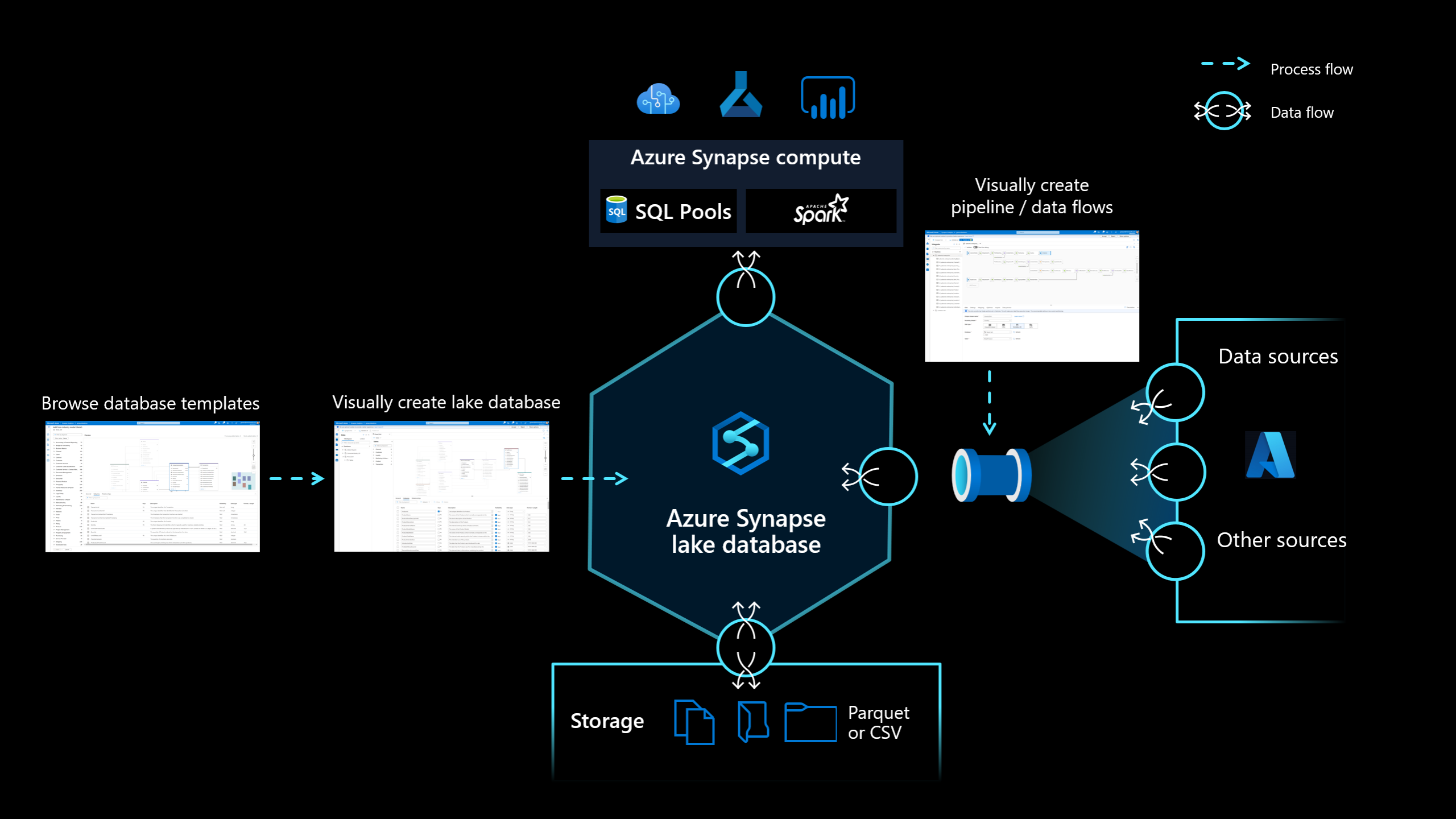
Database templates

Built-in database templates and low code database designer

- Accelerate solution development using industry data templates
- Native support for data integration pipelines
- Pre-built solutions for industry ML models



BLUE ROCKET
DATA CONSULTING & SOLUTIONS


Process flow
Data flow
Azure Synapse compute
SQL Pools
Apache Spark
Visually create pipeline / data flows
Browse database templates
Visually create lake database
Azure Synapse lake database
Data sources
Other sources
Storage
Parquet or CSV

Power BI + Azure Synapse

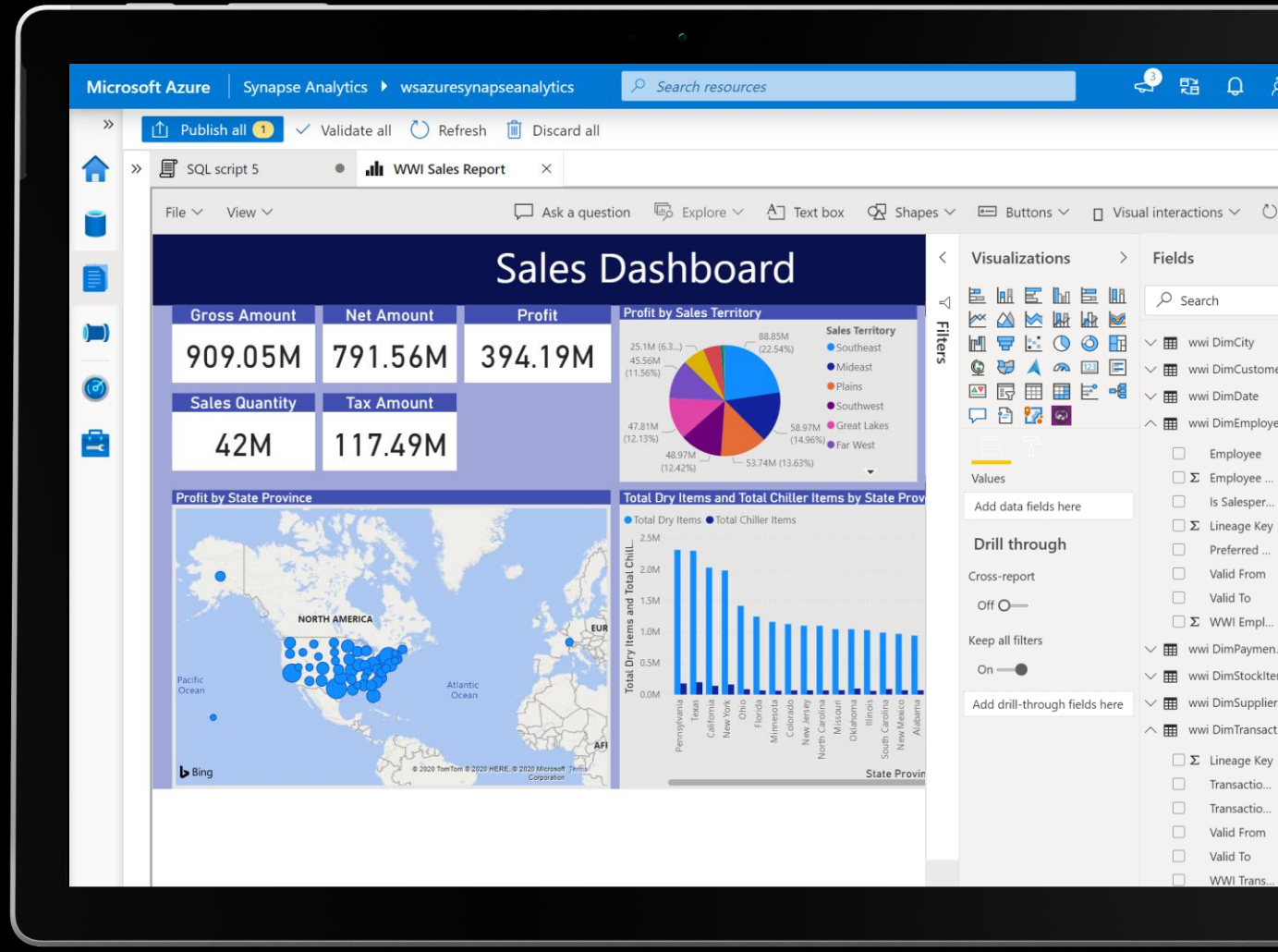
An unmatched combination

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



BLUE ROCKET
DATA CONSULTING & SOLUTIONS

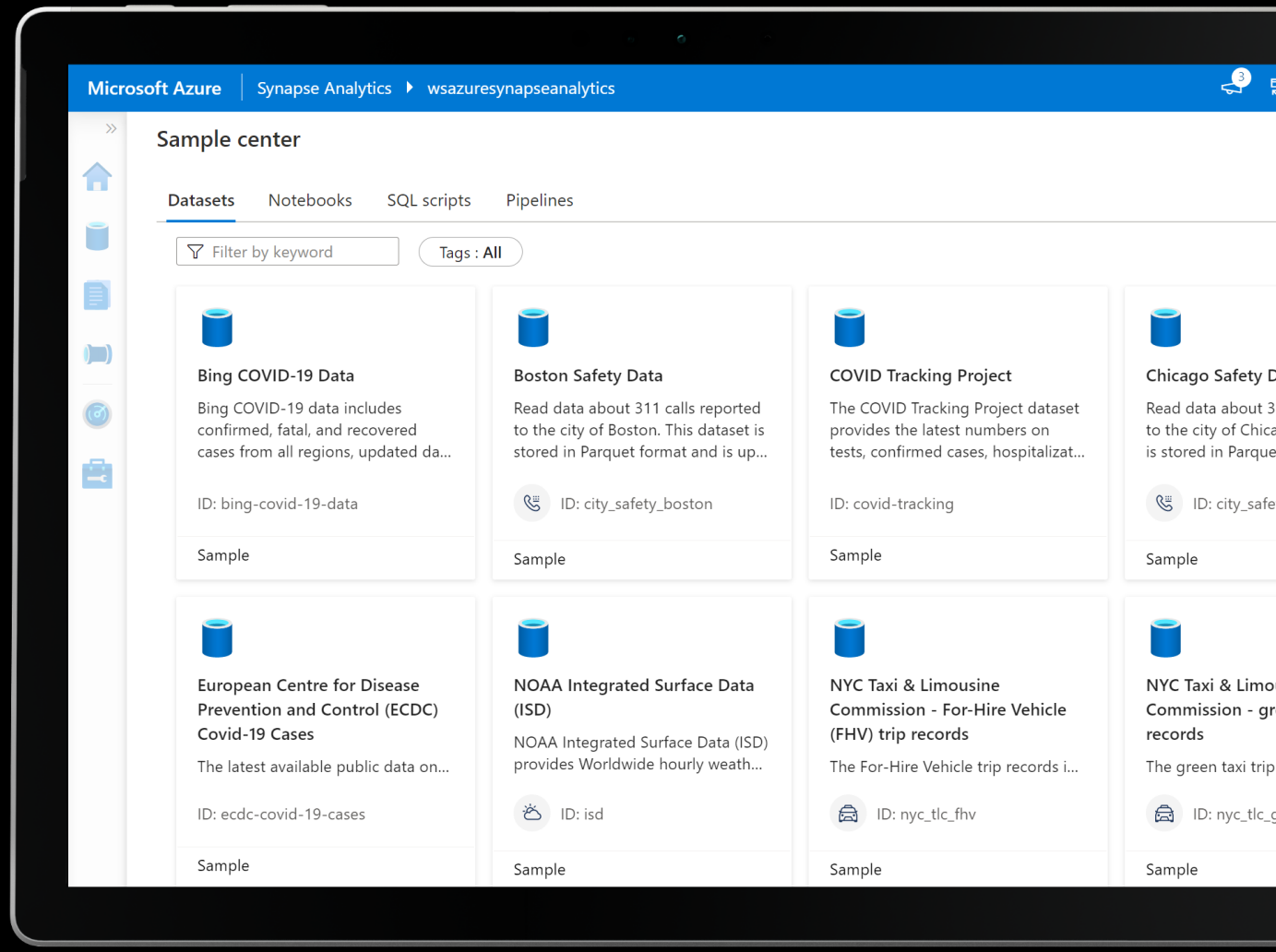
Knowledge Center

Accelerate time to solution

Azure Open Data sets

Pre-built samples to accelerate development

- SQL Scripts
- Notebooks
- Data Pipelines



Microsoft

‘Recent’ announcements

Ignite

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

GPU Support for Synapse Apache Spark

Simplified workloads with no need for separate clusters

Faster data prep using NVIDIA RAPIDS

Accelerated model scoring with Hummingbird

Announcing

Azure Synapse

Delta Lake Support

Azure Synapse Serverless SQL Pools support

Cognitive Services Integration

Native integration to pre-built AI models

Azure Synapse pre-purchase plan

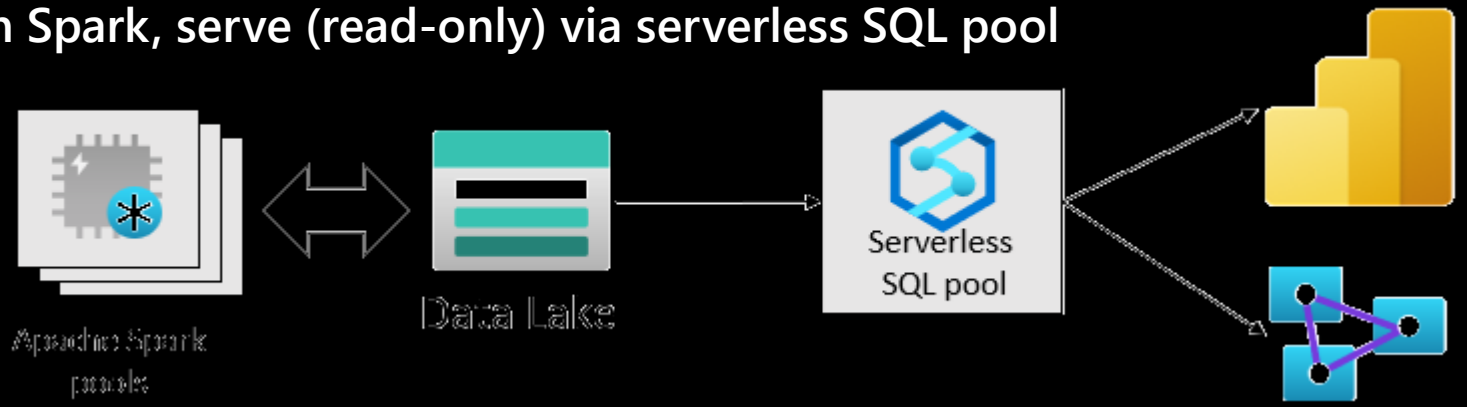
Simplified one-year term pricing option with immediate discounts

Pricing

Azure Pricing Calculator:
https://azure.microsoft.com/en-us/pricing/calculator/

Best practices

- Create/manage lakehouse database in Spark, serve (read-only) via serverless SQL pool



- Three-level hierarchy for Lakehouse (good for Spark)
  - Container (per workspace)
  - Lakehouse Database folders
  - Table folders (Delta)
- In case of multiple workspaces: separate data lake container for workspace-specific data

Comparison (general notes)

Databricks:
+ Better/more/advanced capabilities regarding machine learning / adv analytics. Mlflow.
+ Realtime co-authoring
+ versioning of notebook
+ Databricks Connect for local IDE
+ Photon engine

− Mounting a data lake before using it
− Not a data warehouse tool, rather a Spark-based notebook tool

Attention points:
÷ Networking setup can be difficult
÷ Security implementation can be hard to understand

Synapse Analytics:
+ SQL Serverless! Pay for data scanned. Cheap.
+ One suite to do everything.
+ Integration with Microsoft products (Purview!)
+ .NET support
+ Hyperspace engine (indexing)

− lagging in Delta, this is difficult to bridge.
− lagging in Spark 'power', but MSFT is working hard on this.
− sharing of cluster/session is not efficient
− no option to code together simultaneously in the same notebook
− deployment of Serverless objects not supported by DacPac
− no automatic versioning
− no local IDE connection

Attention points:
÷ default data type for SQL Serverless = varchar max
÷ Security implementation can be hard to understand
÷ Not all (parts of) your Databricks notebooks work, there are differences and gaps
÷ Power BI integration will be “amazing” (not so interesting right now though)

When to use Synapse and when Databricks

Machine Learning development

Databricks:

- ➕ More mature and advanced capabilities
- ➕ TensorFLow, PyTorch, Keras, MLflow, etc
- ➕ Better developer experience (use of IDEs)

Synapse Analytics:

- ➕ AzureML support
- ➕ Open source MLflow support (now renamed Synapse ML)

When to use Synapse and when Databricks

Data Warehousing | Lakehousing

Databricks:

- Databricks Delta (Live tables)
- Better developer experience (use of IDEs)
- Databricks Structured Streaming
- Autoloader

Synapse Analytics:

- Full width of SQL and DWH capabilities
- Both (t-)SQL as Spark
- Don’t need to have the cluster on to read data with Power BI

When to use Synapse and when Databricks

Near real-time transforms/analytics

Databricks:
- Databricks Structured Streaming
- Autoloader

Synapse Analytics:
- Azure Stream Analytics in dedicated SQL pools

Feedback

Feedback

Feedback

https://sqlb.it/?6952

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Resources

Azure Data Engineer Learning Path (PDF!)
https://aka.ms/30days2synapse

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

Whats new:
https://docs.microsoft.com/en-us/azure/synapse-analytics/whats-new

Blog
https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/

Release notes in the workspace