



SQLBITS

Overview of Azure Synapse Analytics

Insert coins to start





BLUE ROCKET
DATA CONSULTING & SOLUTIONS

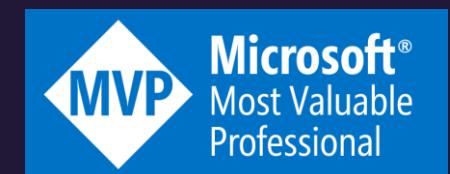
Dave Ruijter

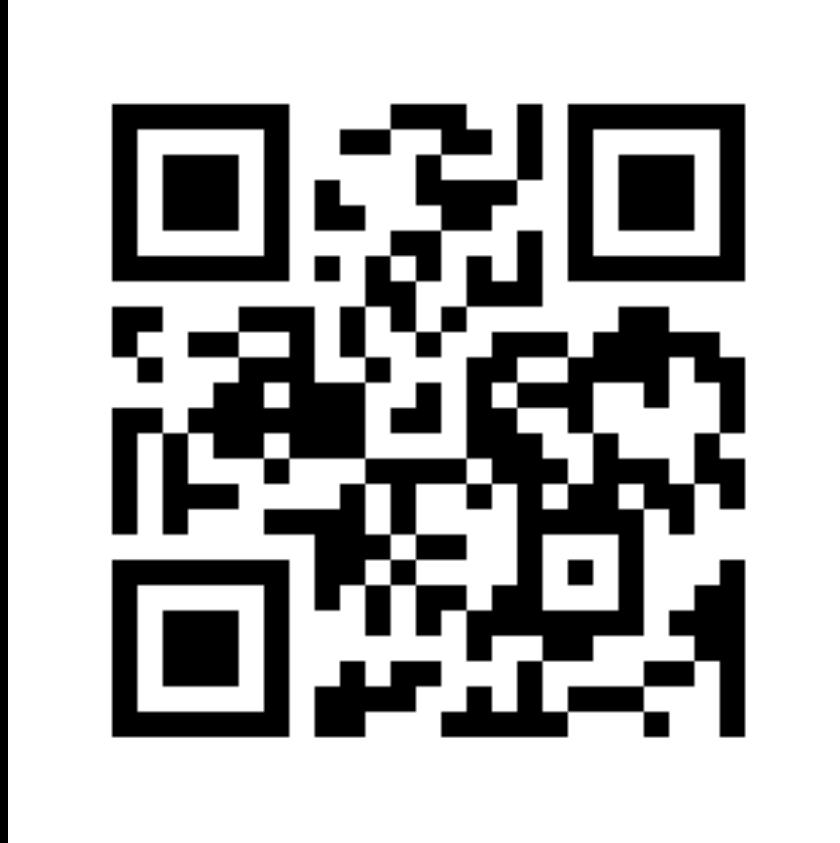
Data Platform MVP | Azure Solution Architect Data & Analytics Consultant

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





Feedback
Feedback
Feedback



<https://sqlb.it/?6952>

Azure Synapse Analytics

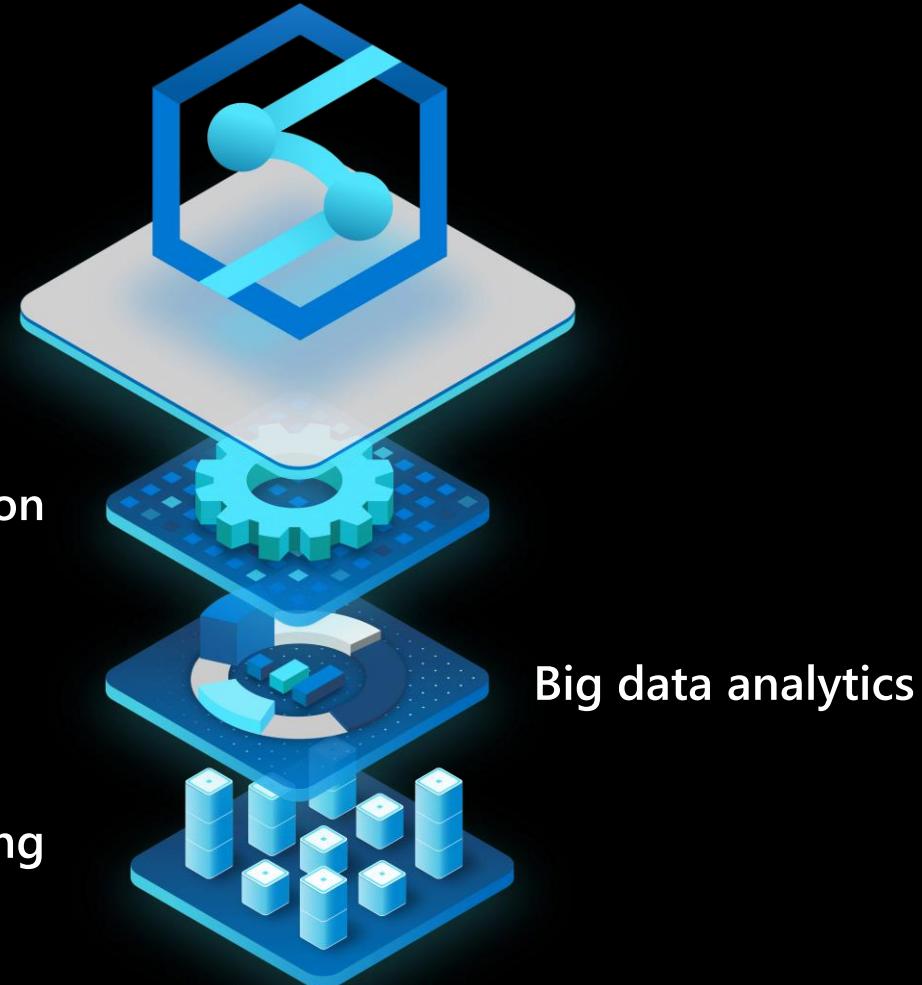
The first unified, cloud native platform for converged analytics

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.

Data integration

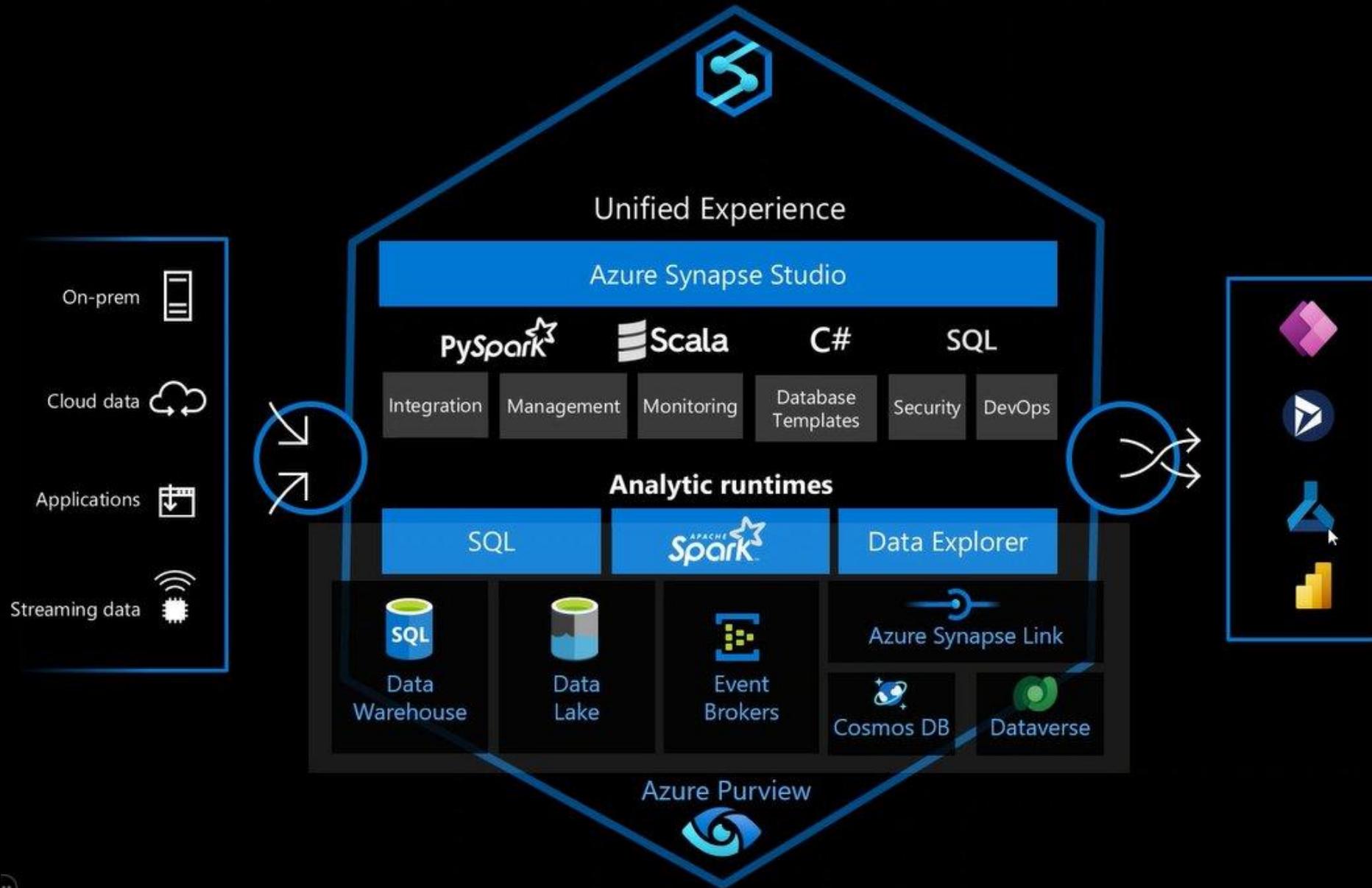
Data warehousing

Big data analytics



Azure Synapse Analytics

The first unified, cloud native platform for converged analytics



Azure Synapse Analytics

Powered by a new cloud native
distributed SQL engine

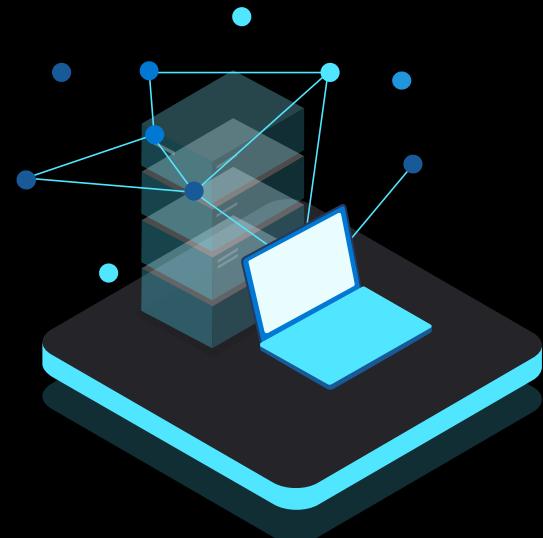


Serverless + dedicated SQL

Flexible consumption models

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

Dedicated clusters optimized mission-critical data warehouse workloads

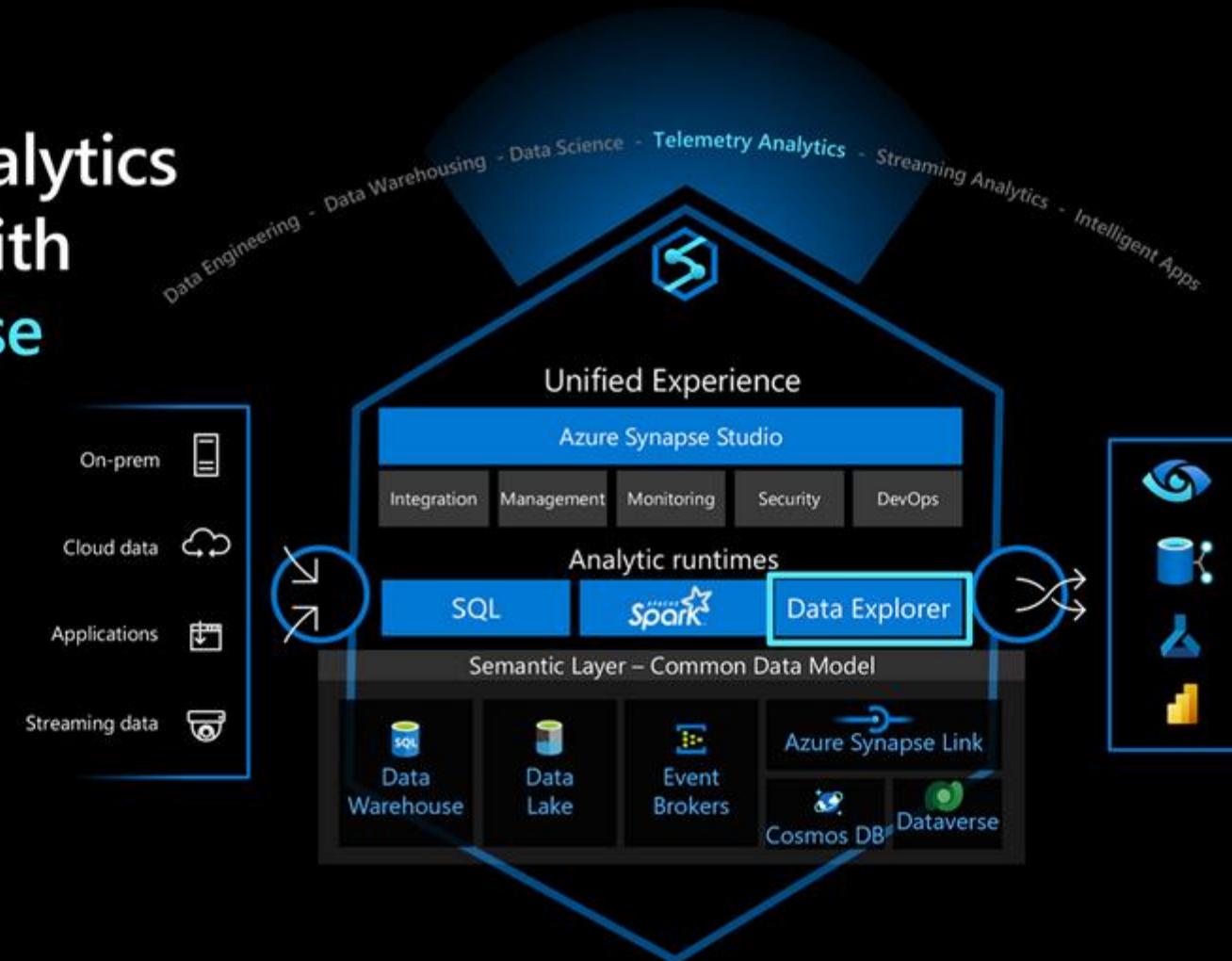


Serverless



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

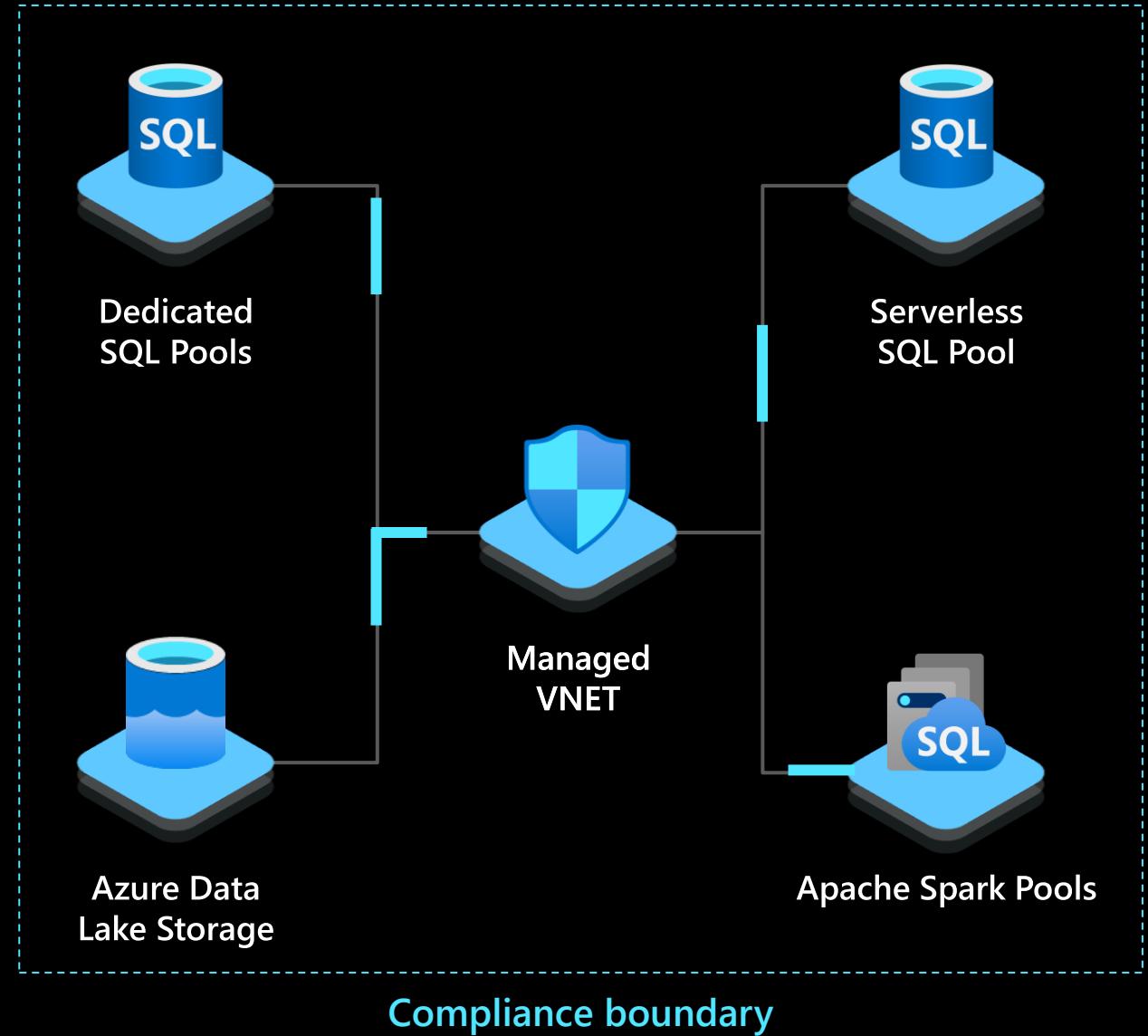
Eliminate network maintenance

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



95 connectors



All your data
(on-prem,
SaaS applications,
other clouds)

Code-free

Code-free data wrangling

No/low-code data transformation

Excel-like interface is familiar to users

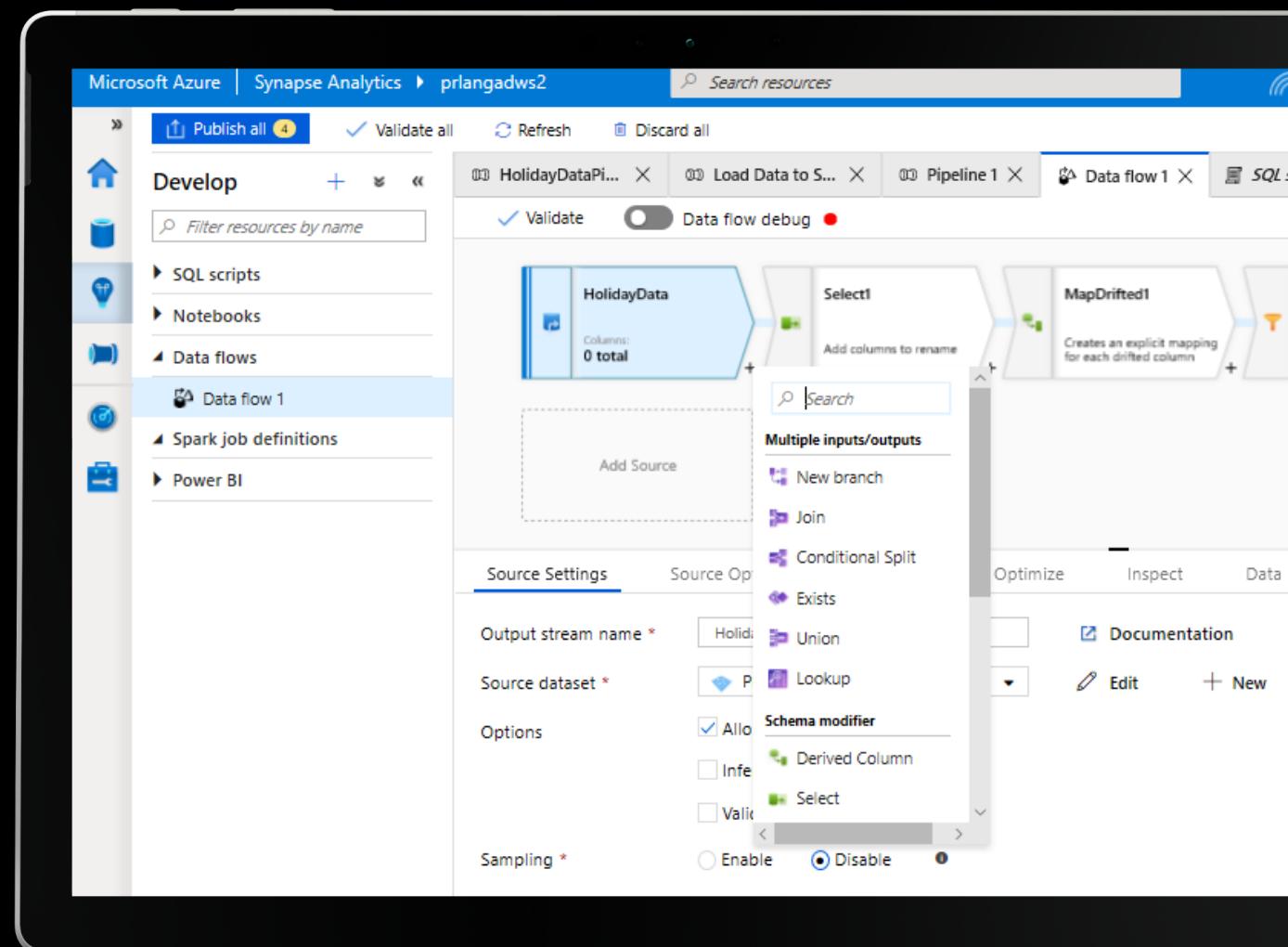
Transform data to desired shape completely visually

Operationalize into pipelines

The screenshot shows the Microsoft Azure Synapse Analytics Power Query Editor interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', 'wsazuresynapseanalytics', and a 'Search resources' bar. Below the navigation is a toolbar with various icons for validation, publishing, and discarding queries. The main workspace is titled 'PQSalesPrep' and shows a 'Settings' section with a note about Power Query M functions. The 'Home' tab is selected, displaying a ribbon of tools: Enter data, Options, Manage parameters, Refresh, Advanced editor, Properties, Manage columns, Choose columns, Remove columns, Keep rows, Remove rows, Reduce rows, Sort, Split column, Group by, Replace values, and Transform. To the right of the ribbon is a data preview pane showing a table with 17 rows and 8 columns. The columns are labeled: ab storeId, ab productCode, 12 quantity, 1.2 logQuantity, ab advertising, ab price, ab weekStarting, and ab id. The data consists of various surface.go entries with numerical values for quantity and price. At the bottom of the preview pane, it says 'Columns: 8 Rows: 99+'. The overall interface is clean and modern, designed for data wrangling and transformation.

Hybrid data integration

Data Flows



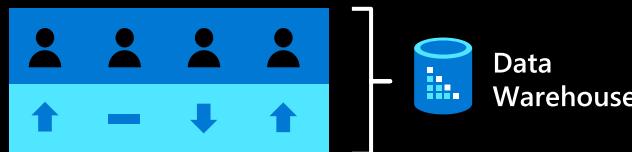
Data warehouse

Scalable and secure analytics platform for SQL workloads

Workload management

Scale in

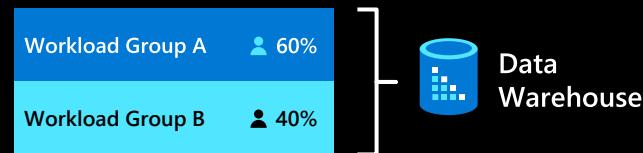
Workload Importance



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

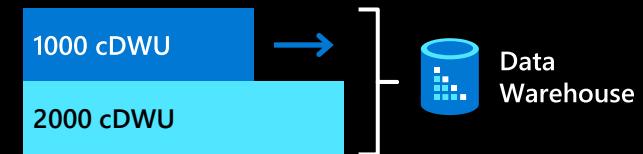


Benefits:

- 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)



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

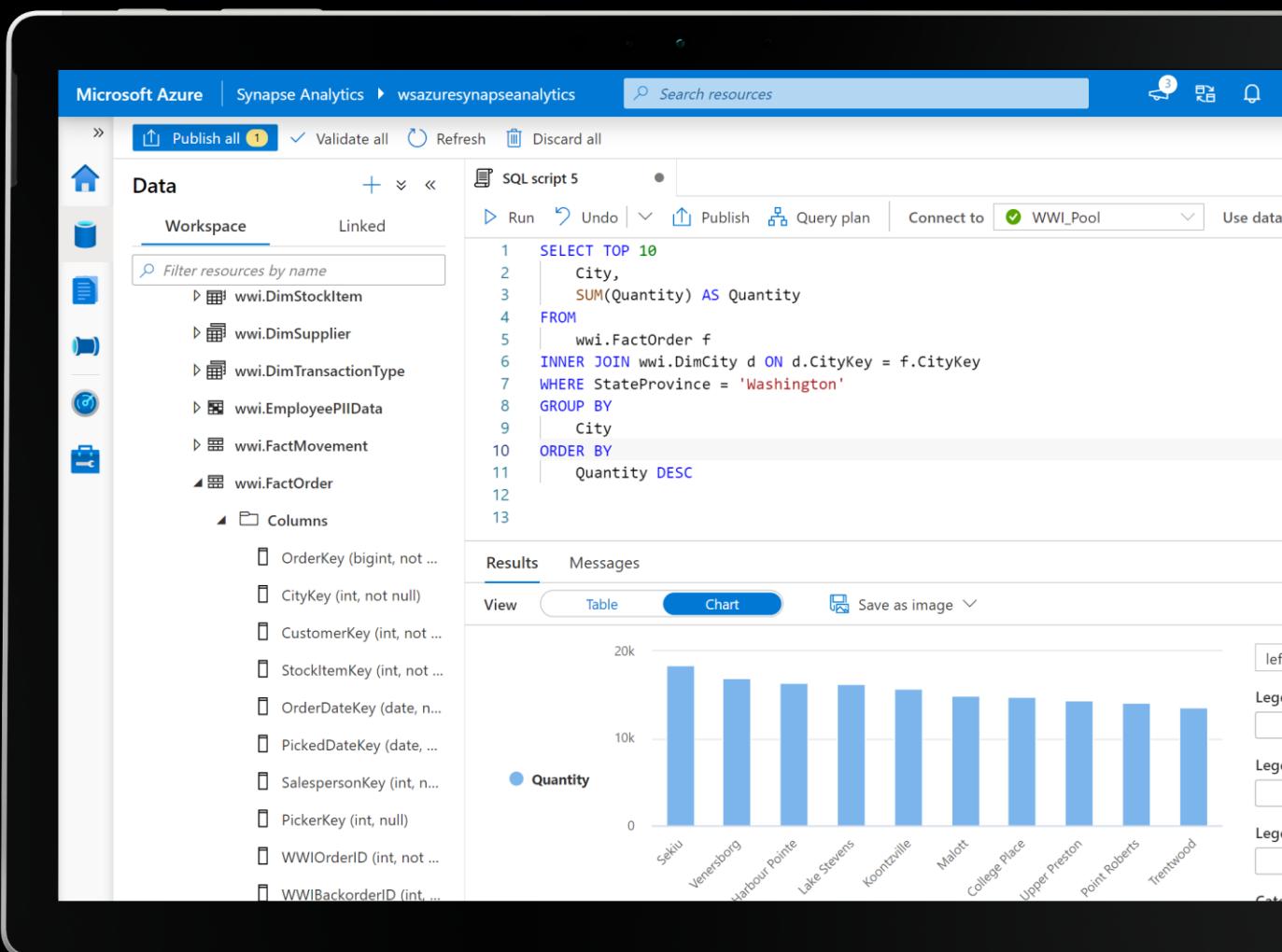
SQL Editor

Automatic code completion (Intellisense)

Script collaboration within the Workspace

Built-in visualizations

Easily switch between clusters



The screenshot shows the Microsoft Azure Synapse Analytics Studio interface. On the left, the 'Data' workspace is selected, displaying a list of datasets and tables from a schema named 'wwi'. A specific table, 'FactOrder', is expanded to show its columns. On the right, an SQL script editor displays a query to select top 10 cities with their total quantity, grouped by city and ordered by quantity in descending order. Below the script, a bar chart titled 'Quantity' visualizes the data, showing the total quantity for each city. The chart has bars for Sekiu, Venerborg, Harbour Pointe, Lake Stevens, Koontzville, Malott, College Place, Upper Preston, Point Roberts, and Trentwood.

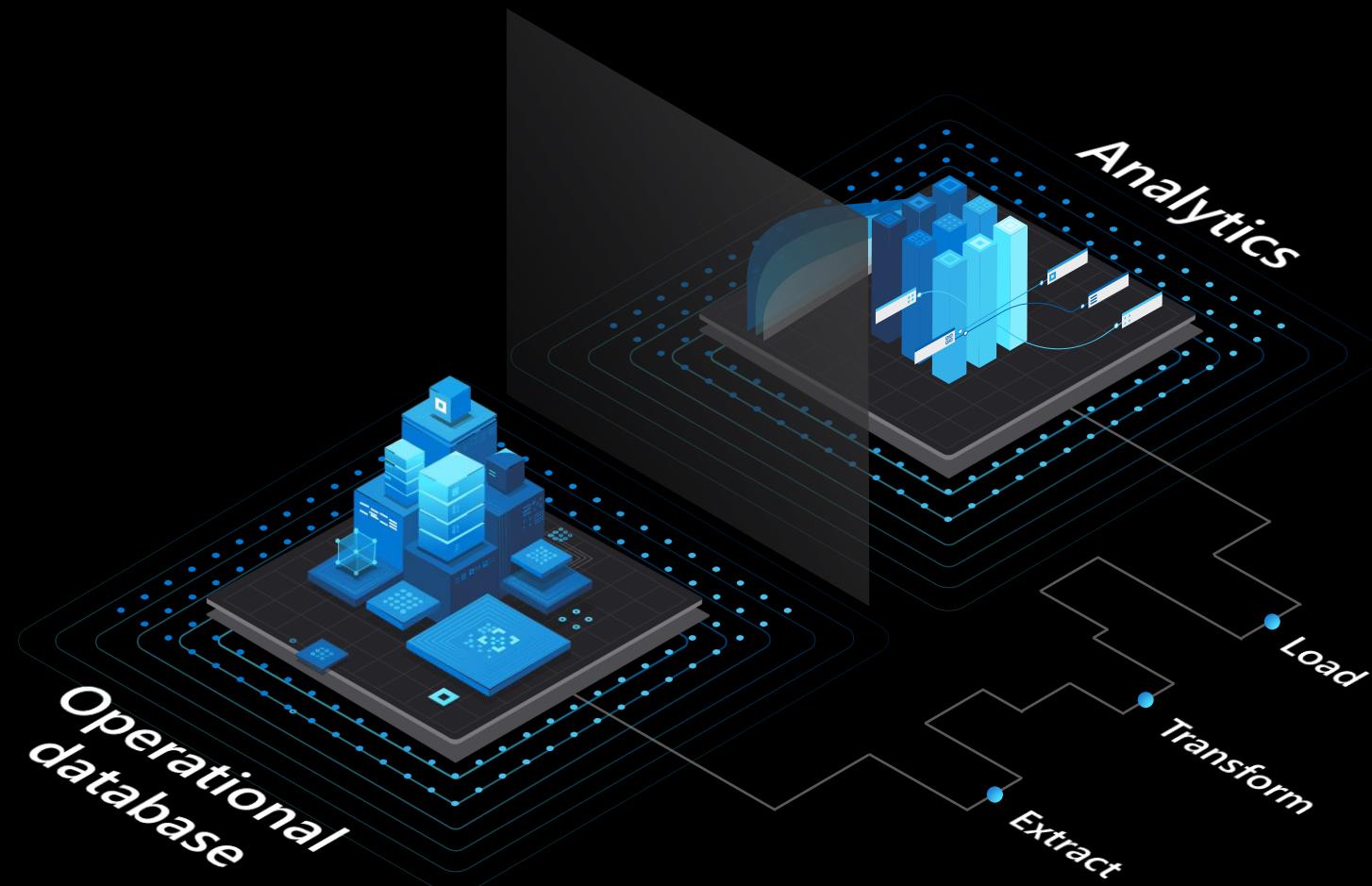
```
1 SELECT TOP 10
2     City,
3         SUM(Quantity) AS Quantity
4 FROM
5     wwi.FactOrder f
6 INNER JOIN wwi.DimCity d ON d.CityKey = f.CityKey
7 WHERE StateProvince = 'Washington'
8 GROUP BY
9     City
10 ORDER BY
11     Quantity DESC
12
13
```

| City | Quantity |
|----------------|----------|
| Sekiu | ~19k |
| Venerborg | ~17k |
| Harbour Pointe | ~16.5k |
| Lake Stevens | ~16.5k |
| Koontzville | ~15.5k |
| Malott | ~14.5k |
| College Place | ~14k |
| Upper Preston | ~13.5k |
| Point Roberts | ~13.5k |
| Trentwood | ~12.5k |

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



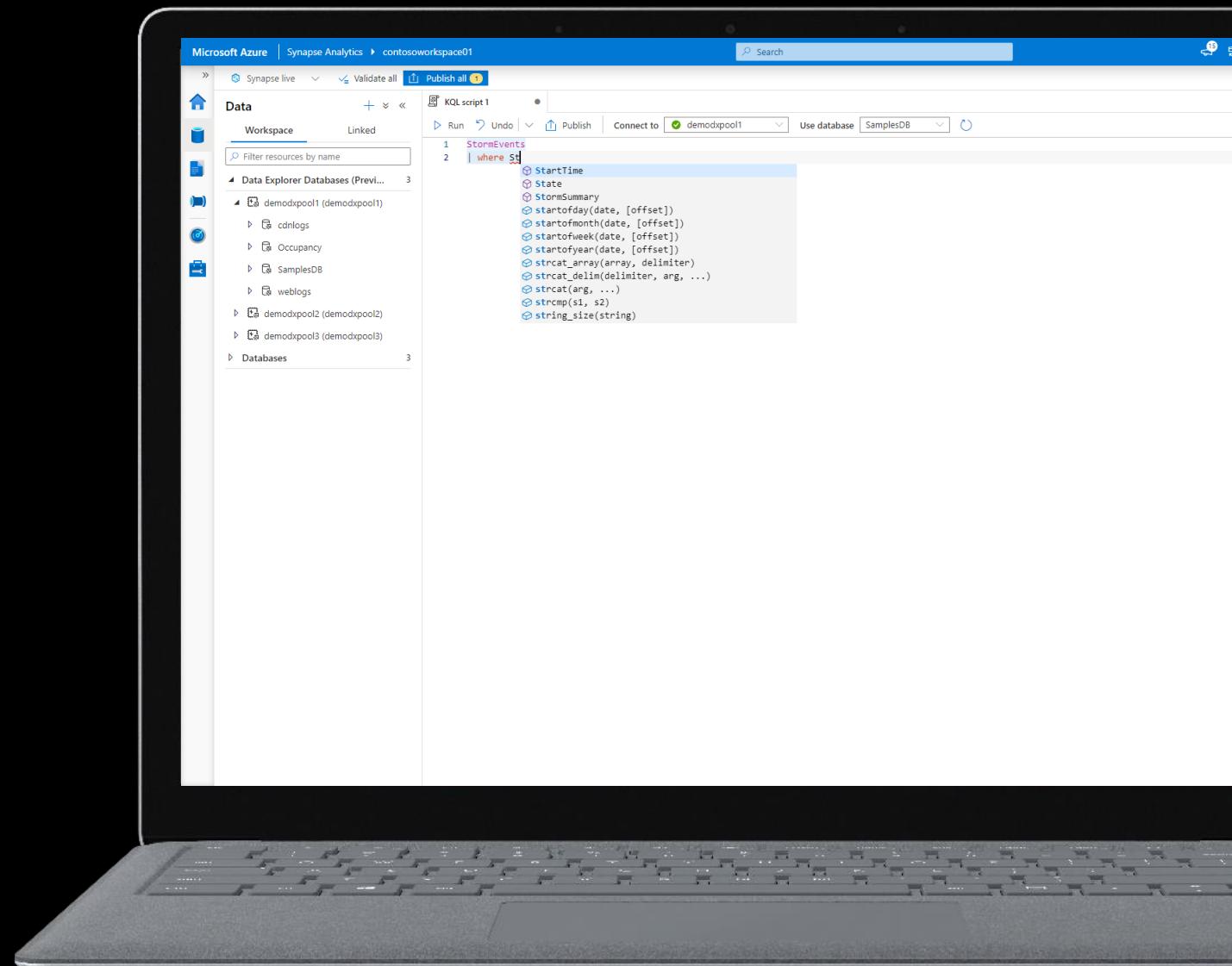
Azure Synapse Link for SQL Server 2022 SQL Server 2022 CTP 1.0 – Coming Soon

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



Security log analytics



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



IoT analytics solutions



Machine learning

Empower everyone with predictive insights

Notebook IDE code authoring

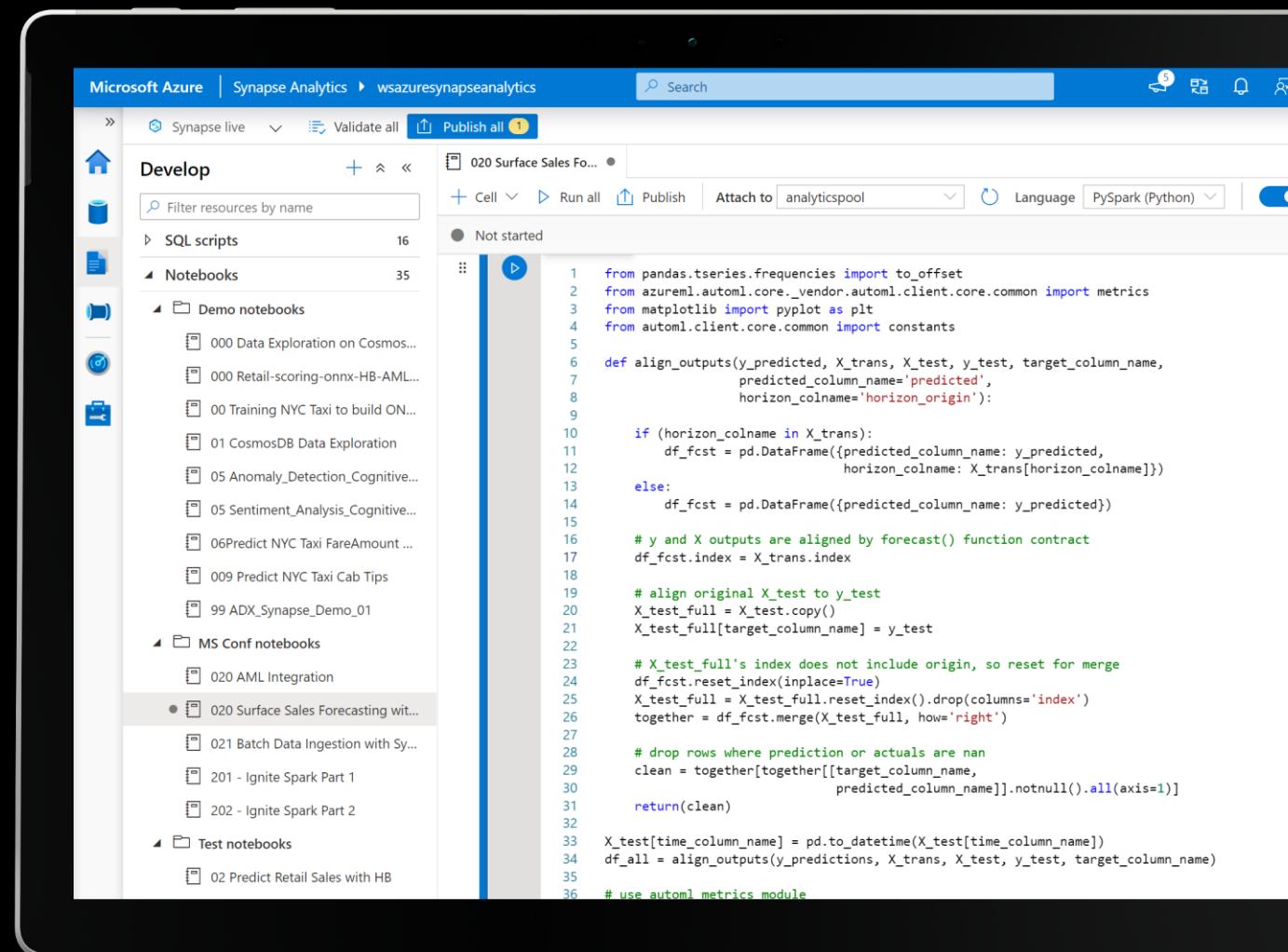
Code-first ML model development

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



The screenshot shows the Microsoft Azure Synapse Analytics Notebook IDE interface. The left sidebar displays a tree view of notebooks, SQL scripts, and MS Conf notebooks. The main area shows a Python script titled '020 Surface Sales Forecasting with Synapse'. The script imports pandas, azurerm, matplotlib, and automl modules, defines a function to align outputs, and performs data processing and merging operations.

```
from pandas.tseries.frequencies import to_offset
from azurerm.core._vendor.automl.client.core.common import metrics
from matplotlib import pyplot as plt
from automl.client.core.common import constants

def align_outputs(y_predicted, X_trans, X_test, y_test, target_column_name,
                  predicted_column_name='predicted',
                  horizon_colname='horizon_origin'):

    if (horizon_colname in X_trans):
        df_fcst = pd.DataFrame({predicted_column_name: y_predicted,
                                horizon_colname: X_trans[horizon_colname]})

    else:
        df_fcst = pd.DataFrame({predicted_column_name: y_predicted})

    # y and X outputs are aligned by forecast() function contract
    df_fcst.index = X_trans.index

    # align original X_test to y_test
    X_test_full = X_test.copy()
    X_test_full[target_column_name] = y_test

    # X_test_full's index does not include origin, so reset for merge
    df_fcst.reset_index(inplace=True)
    X_test_full = X_test_full.reset_index().drop(columns='index')
    together = df_fcst.merge(X_test_full, how='right')

    # drop rows where prediction or actuals are nan
    clean = together[[target_column_name,
                      predicted_column_name]].notnull().all(axis=1)

    return(clean)

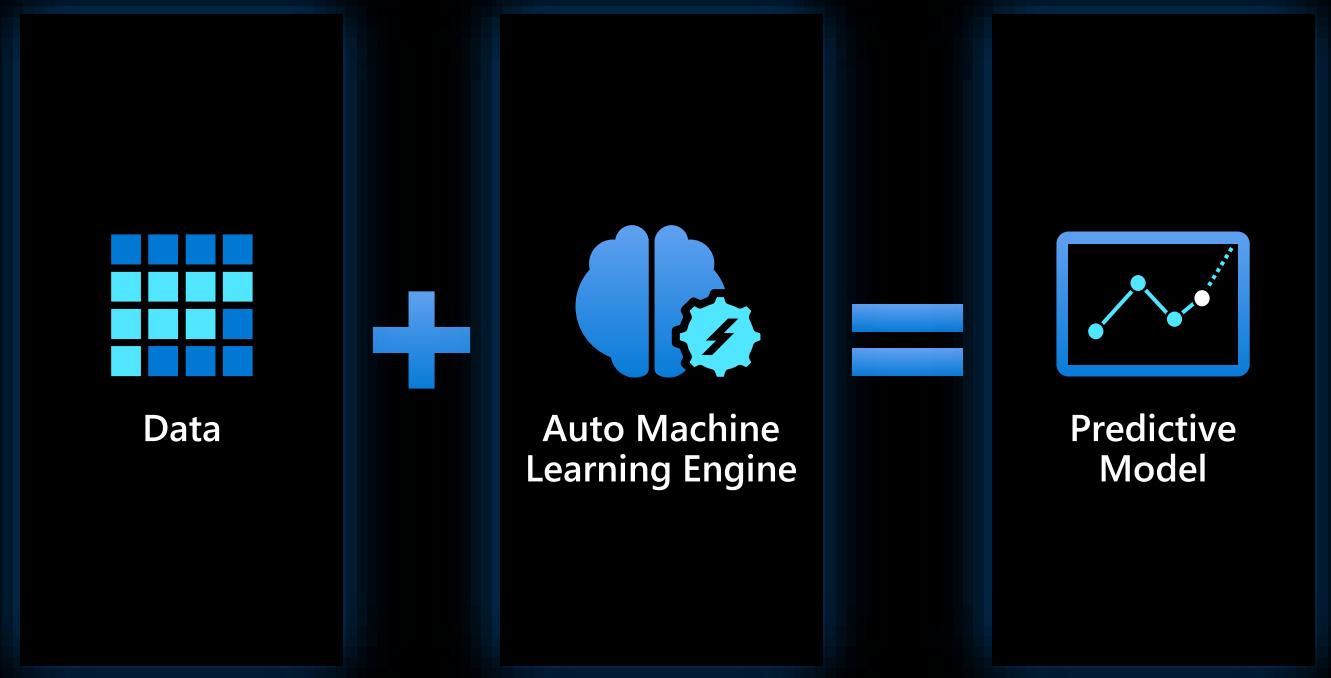
X_test[time_column_name] = pd.to_datetime(X_test[time_column_name])
df_all = align_outputs(y_predictions, X_trans, X_test, y_test, target_column_name)

# use automl.metrics module
```

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

The screenshot shows the Microsoft Azure Synapse Analytics Studio interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and 'wsazuresynapseanalytics'. A search bar and various icons are also present. On the left, a sidebar titled 'Data' shows a 'Workspace' section with a 'Linked' tab and a 'Databases' section containing several database connections like 'newpoll (SQL)', 'NYCTaxi_Pool (SQL)', etc. Below this is a 'Tables' section with 'retailsales' and 'surfacesalesdb (Spark)'. The main area is titled 'Not started' and contains a code editor with the following Python script:

```
1 from pandas import *
2 from azureml import *
3 from matplotlib import *
4 from automl import *
5 def align_(df):
6     if (horizon == 'short-term'):
7         df = df[['date', 'retailsales']]
8     else:
9         df = df[['date', 'retailsales']]
10    # y and df_fcst
11    # align X-test
12    X_test = df[['date', 'retailsales']]
13    X_test = pd.get_dummies(X_test)
14    X_test['date'] = pd.DatetimeIndex(X_test['date'])
15    X_test['date'] = X_test['date'].dt.date
16    X_test['date'] = X_test['date'].dt.strftime('%Y-%m-%d')
17    X_test['date'] = pd.to_datetime(X_test['date'])
18    X_test['date'] = X_test['date'].dt.date
19    X_test['date'] = X_test['date'].dt.strftime('%Y-%m-%d')
20    X_test['date'] = pd.DatetimeIndex(X_test['date'])
21    X_test['date'] = X_test['date'].dt.date
22    X_test['date'] = X_test['date'].dt.strftime('%Y-%m-%d')
23    X_test['date'] = pd.DatetimeIndex(X_test['date'])
24    X_test['date'] = X_test['date'].dt.date
25    X_test['date'] = X_test['date'].dt.strftime('%Y-%m-%d')
26    X_test['date'] = pd.DatetimeIndex(X_test['date'])
27    X_test['date'] = X_test['date'].dt.date
28    X_test['date'] = X_test['date'].dt.strftime('%Y-%m-%d')
29    X_test['date'] = pd.DatetimeIndex(X_test['date'])
30    X_test['date'] = X_test['date'].dt.date
31    X_test['date'] = X_test['date'].dt.strftime('%Y-%m-%d')
32    X_test['date'] = pd.DatetimeIndex(X_test['date'])
33    df_all = df[['date', 'retailsales']]
34    df_all['date'] = pd.DatetimeIndex(df_all['date'])
35    df_all['date'] = df_all['date'].dt.date
36    df_all['date'] = df_all['date'].dt.strftime('%Y-%m-%d')
```

To the right of the code editor, there's a 'Choose a model type' section with three options: 'Classification', 'Regression', and 'Time series forecasting', each with a brief description and example.

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

The screenshot shows the Microsoft Azure Synapse Analytics Data studio interface. The left sidebar shows a navigation tree with 'Data' selected, under 'Workspace'. The main area displays a list of databases: 'newpoll (SQL)', 'NYCTaxi_Pool (SQL)', and 'dbo.nyc_taxi'. Under 'dbo.nyc_taxi', there are tables like 'dbo.aml_models', 'dbo.modeldeploy', 'dbo.Models', and 'dbo.nyctaxi'. A context menu is open over the 'dbo.nyctaxi' table, with the 'Machine Learning' option highlighted. To the right of the menu, a tooltip says 'Enrich with existing model'. The top right corner of the interface shows a 'Machine Learning workspace' dropdown set to 'amlwsdemos'.

Enrich with existing model

dbo.nyc_taxi

Select the model you want to use to enrich the selected dataset. [Learn more](#)

Azure Machine Learning

Azure Machine Learning workspace *

amlwsdemos

| Name | Version | Created |
|---|---------|---------------|
| mechanics-retail_data-20201105231400... | 1 | 07:25:59 1... |
| mechanics-retail_data-20201105110000... | 2 | 07:25:43 1... |
| mechanics-retail_data-20201105110000... | 1 | 07:25:31 1... |
| azuresynapse5-retail_training_data-202... | 2 | 18:01:45 1... |
| synapse_retail_model_tb | 6 | 21:24:05 1... |
| synapse_retail_model_tb_new | 1 | 05:37:03 1... |
| synapse_retail_model_tb | 5 | 14:58:10 1... |
| synapse_retail_model_tb | 4 | 05:33:55 1... |
| synapse_retail_model_tb | 3 | 17:05:19 1... |
| synapse_retail_model_tb | 2 | 01:50:02 1... |
| synapse_retail_model_tb | 1 | 17:22:10 1... |
| synapse_retail_model_onnx | 1 | 22:05:45 1... |
| nyc_taxi_tip_predict | 1 | 17:45:02 1... |

Use the

New SQL script

New notebook

New data flow

New integration dataset

Machine Learning

Enrich with existing model

Refresh

Continue

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



+ a b | e a u



MicroStrategy

SISENSE

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

Synapse SQL



Model



Data



Predictions

databricks

PaddlePaddle

Spark

MathWorks

Chainer

XGBoost

learn

Caffe2

PyTorch

Sas

ML.NET

F

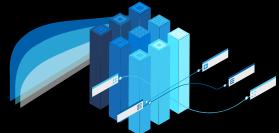
mxnet

K

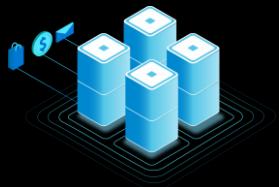
M

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

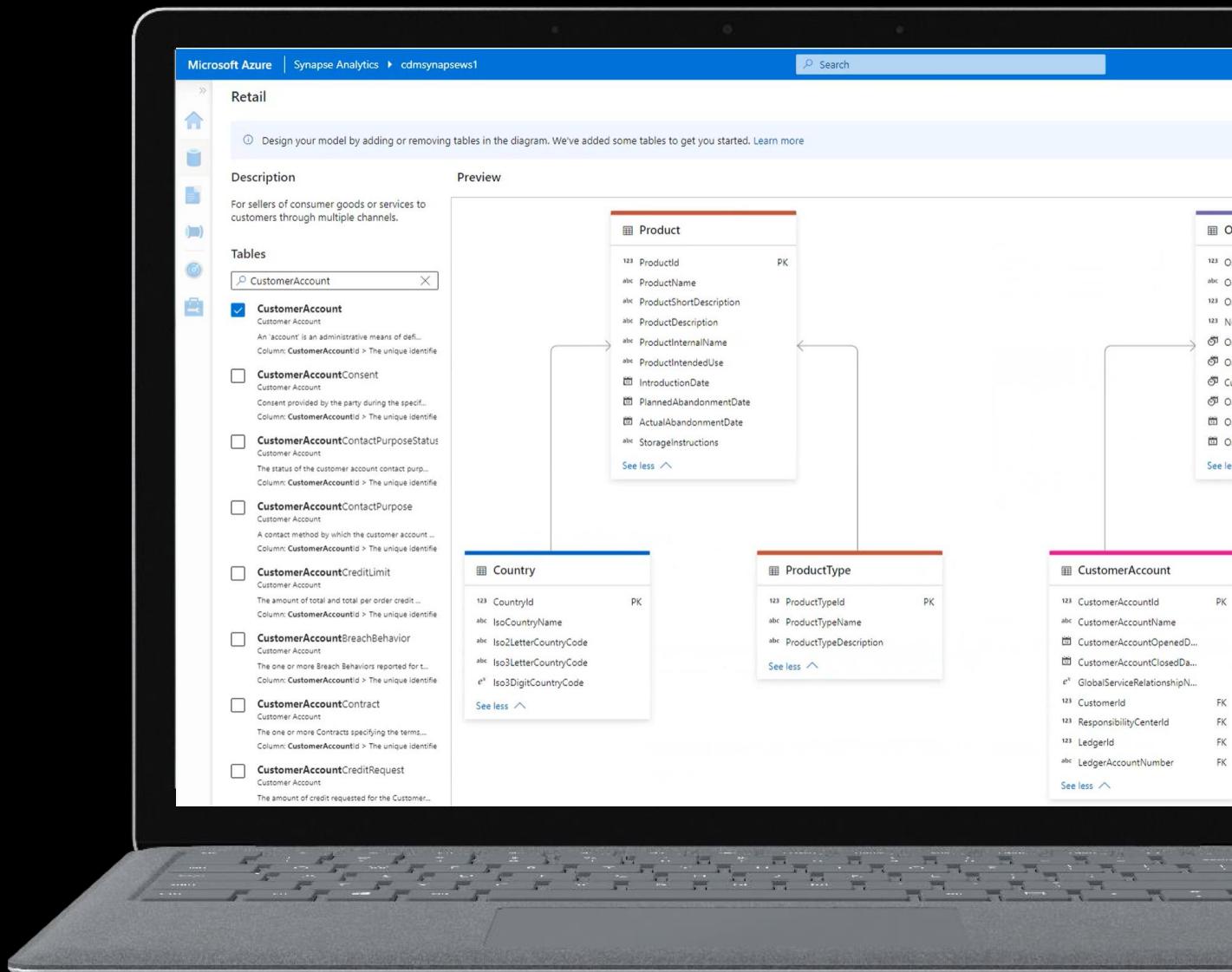


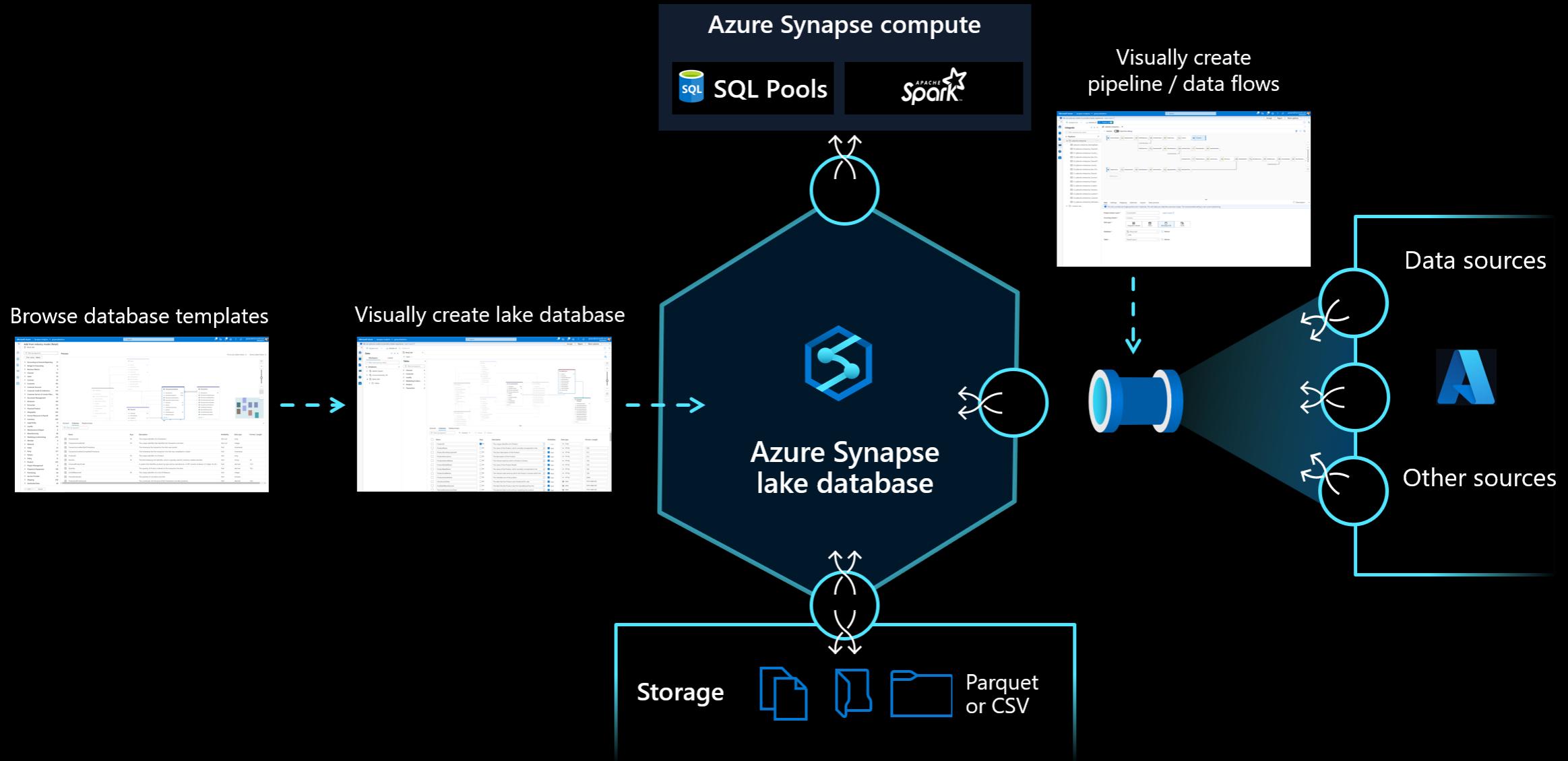
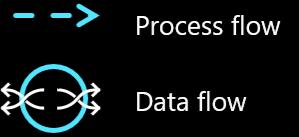
Announcing

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





Power BI + Azure Synapse

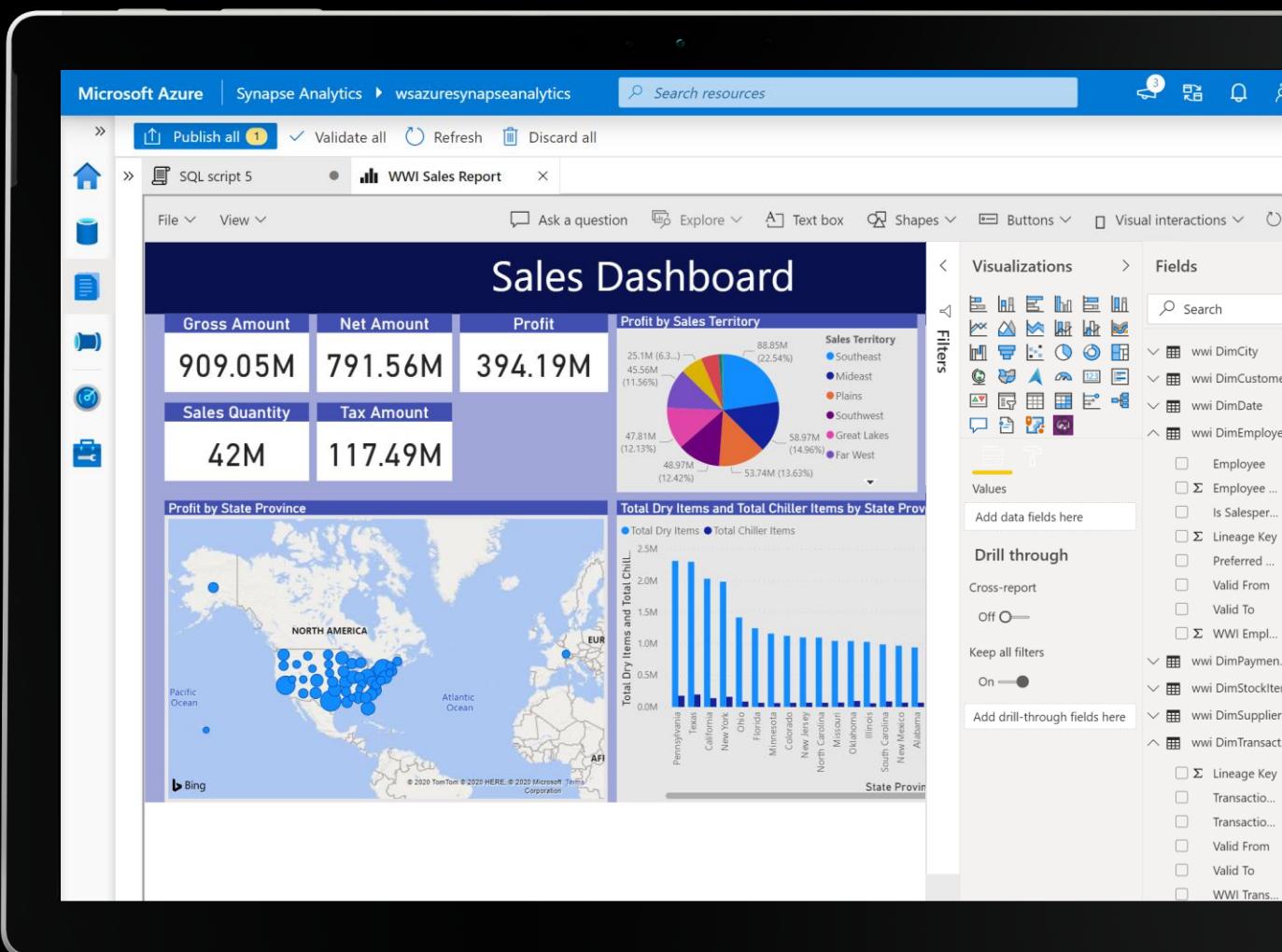
An unmatched combination

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



Knowledge Center

Accelerate time to solution

Azure Open Data sets

Pre-built samples to accelerate development

- SQL Scripts
- Notebooks
- Data Pipelines

The screenshot shows the Microsoft Azure Synapse Analytics Sample center interface. The top navigation bar includes 'Microsoft Azure' and 'Synapse Analytics' with a path to 'wsazuresynapseanalytics'. Below the navigation is a sidebar with icons for Home, Datasets, Notebooks, SQL scripts, and Pipelines. The main area is titled 'Sample center' and has tabs for 'Datasets', 'Notebooks', 'SQL scripts', and 'Pipelines'. A search bar says 'Filter by keyword' and a tag filter says 'Tags : All'. There are several dataset cards displayed in a grid:

| Dataset | Description | ID | Sample |
|--|---|-------------------------|--------|
| Bing COVID-19 Data | Bing COVID-19 data includes confirmed, fatal, and recovered cases from all regions, updated daily. | ID: bing-covid-19-data | Sample |
| Boston Safety Data | Read data about 311 calls reported to the city of Boston. This dataset is stored in Parquet format and is updated daily. | ID: city_safety_boston | Sample |
| COVID Tracking Project | The COVID Tracking Project dataset provides the latest numbers on tests, confirmed cases, hospitalizations, and deaths. | ID: covid-tracking | Sample |
| Chicago Safety Data | Read data about 311 calls reported to the city of Chicago. This dataset is stored in Parquet format and is updated daily. | ID: city_safety_chicago | Sample |
| European Centre for Disease Prevention and Control (ECDC) Covid-19 Cases | The latest available public data on COVID-19 cases and variants. | ID: ecdc-covid-19-cases | Sample |
| NOAA Integrated Surface Data (ISD) | NOAA Integrated Surface Data (ISD) provides Worldwide hourly weather observations. | ID: isd | Sample |
| NYC Taxi & Limousine Commission - For-Hire Vehicle (FHV) trip records | The For-Hire Vehicle trip records include pickup and drop-off locations, vehicle type, and fare information. | ID: nyc_tlc_fhv | Sample |
| NYC Taxi & Limousine Commission - green taxi trip records | The green taxi trip records include pickup and drop-off locations, vehicle type, and fare information. | ID: nyc_tlc_green | Sample |



'Recent' announcements

ignite

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

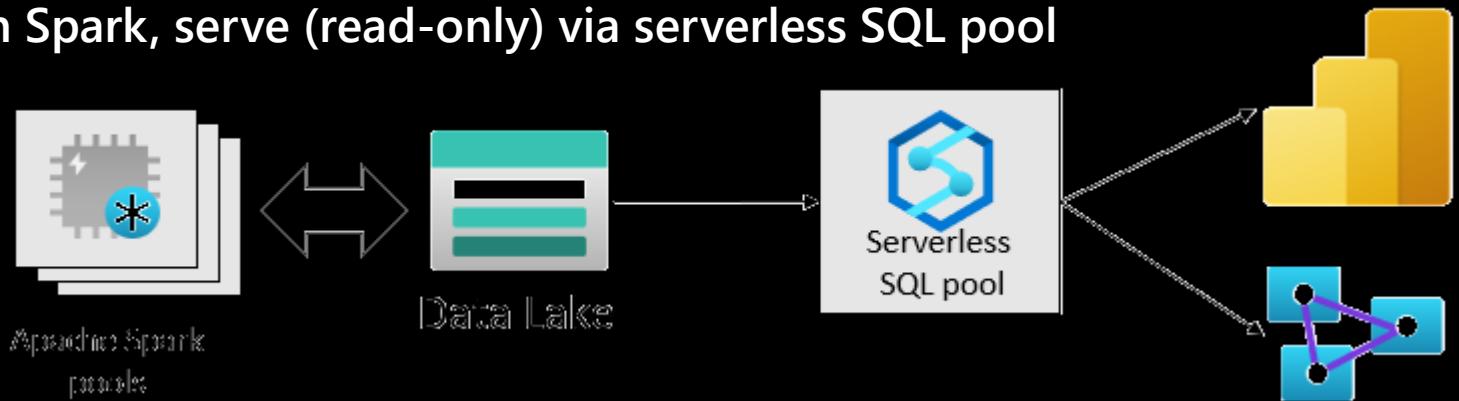


Azure Pricing Calculator:

<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

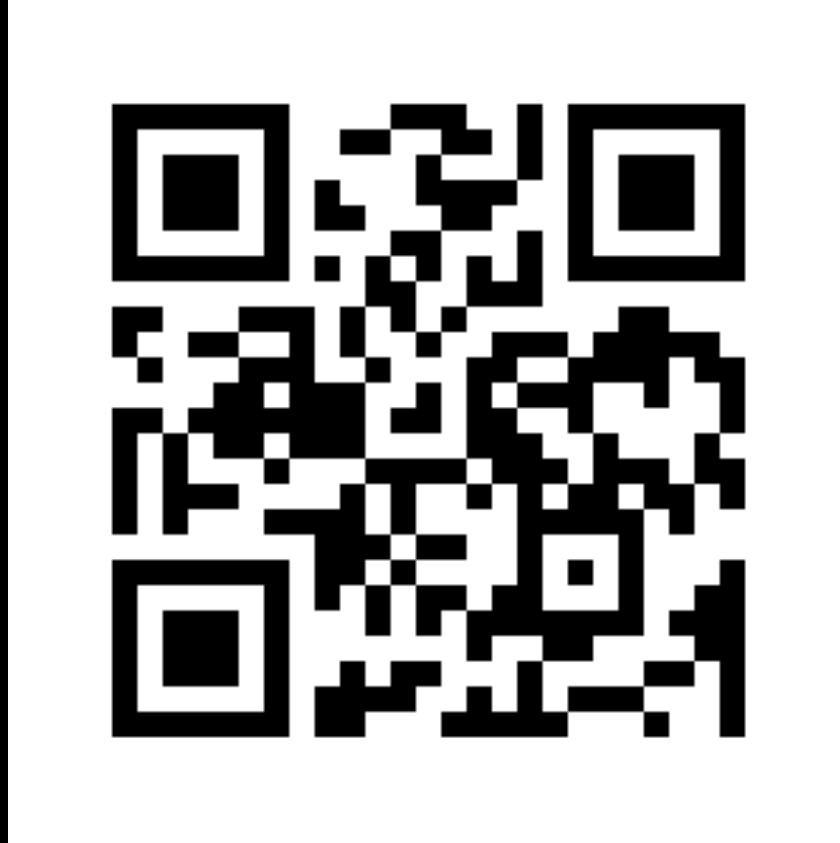
Near real-time transforms/analytics

Databricks:

- + Databricks Structured Streaming
- + Autoloader

Synapse Analytics:

- + Azure Stream Analytics in dedicated SQL pools



<https://sqlb.it/?6952>

Feedback
Feedback
Feedback



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