



# GETTING STARTED WITH BUILDING YOUR AZURE SYNAPSE ENVIRONMENT



**Erwin de Kreuk**

Principal Consultant – Lead Data & AI  
InSpark



@erwindekreuk



[linkedin.com/in/erwindekreuk](https://linkedin.com/in/erwindekreuk)



[erwindekreuk.com](http://erwindekreuk.com)



[github.com/edkreuk](https://github.com/edkreuk)

..

# Erwin de Kreuk

Principal Consultant – Lead Data & AI  
InSpark



**Erwin de Kreuk**

Principal Consultant - Lead Data & AI  
InSpark

 @erwindekreuk

 linkedin.com/in/erwindekreuk

 erwindekreuk.com

 github.com/edkreuk

 <https://sessionize.com/erwin-de-kreuk/>



Let's  
connect



# We Are InSpark

We help organizations  
**accelerating their digital  
transformation with impactful  
Microsoft solutions & expertise**

# What is Synapse Analytics

1. Unified Analytics Platform
2. Analytic Runtimes
3. Synapse Studio
4. Networking
5. Pipelines
6. Load Data



# Each new technology creates another siloed operation

- Big data
- Data integration
- Machine learning
- Business intelligence
- Data governance
- Security



# Azure Synapse Analytics

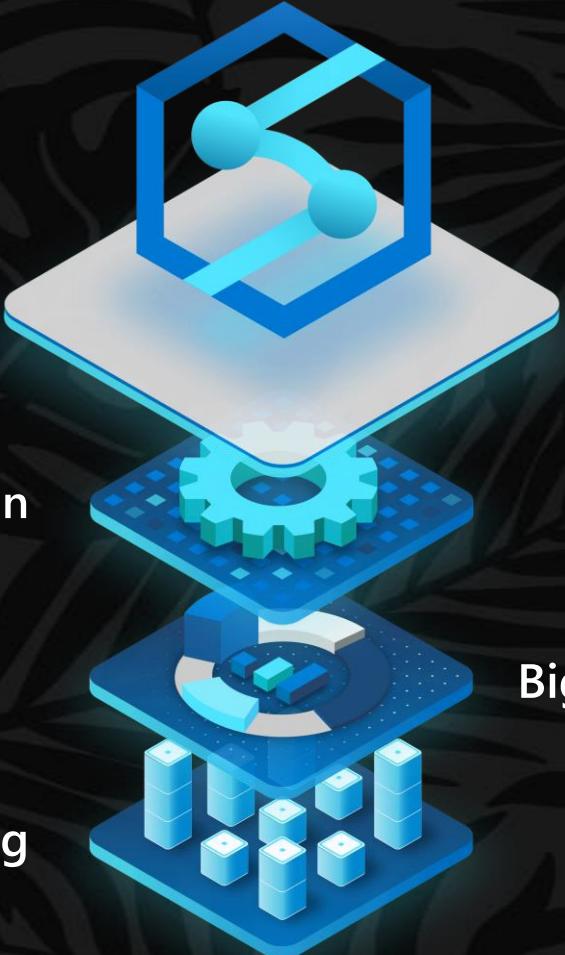
The first unified, cloud native platform for converged analytics

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

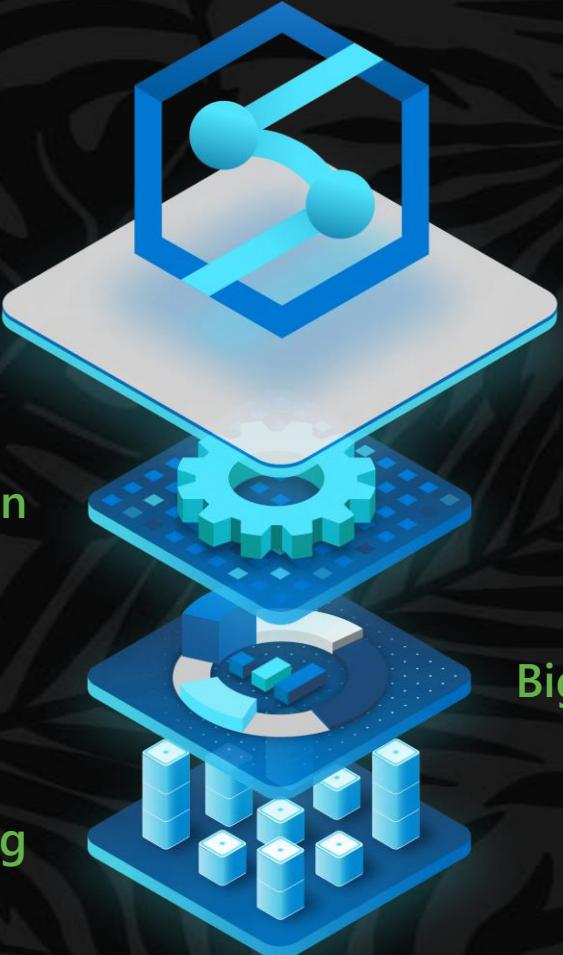
The first unified, cloud native platform for converged analytics

Azure Synapse 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.

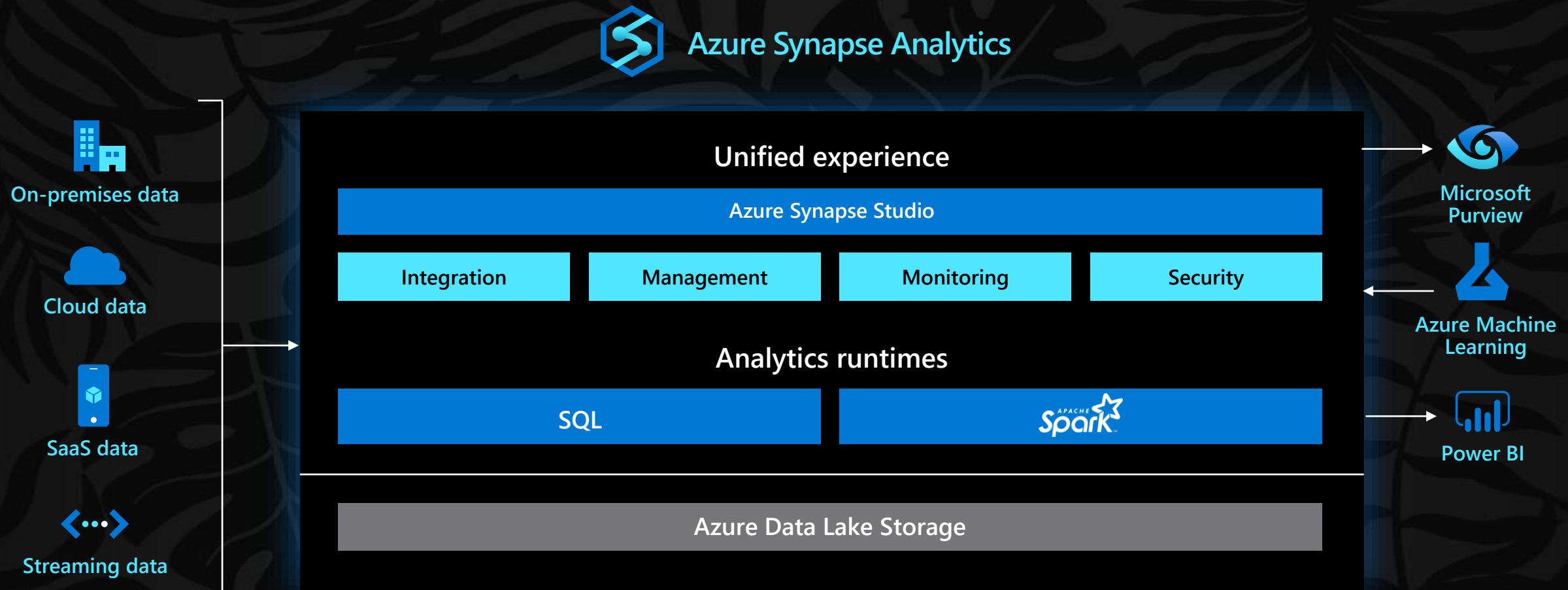
Data integration

Data warehousing

Big data analytics



# Azure Synapse lies at the heart of business, AI, and BI



# Azure Synapse Analytics



Synapse SQL



Synapse Spark



Synapse Pipelines



Synapse Studio



# Synapse SQL

- Serverless pay-per-query ideal for ad-hoc data lake exploration and transformation
- Dedicated clusters optimized mission-critical data warehouse workloads



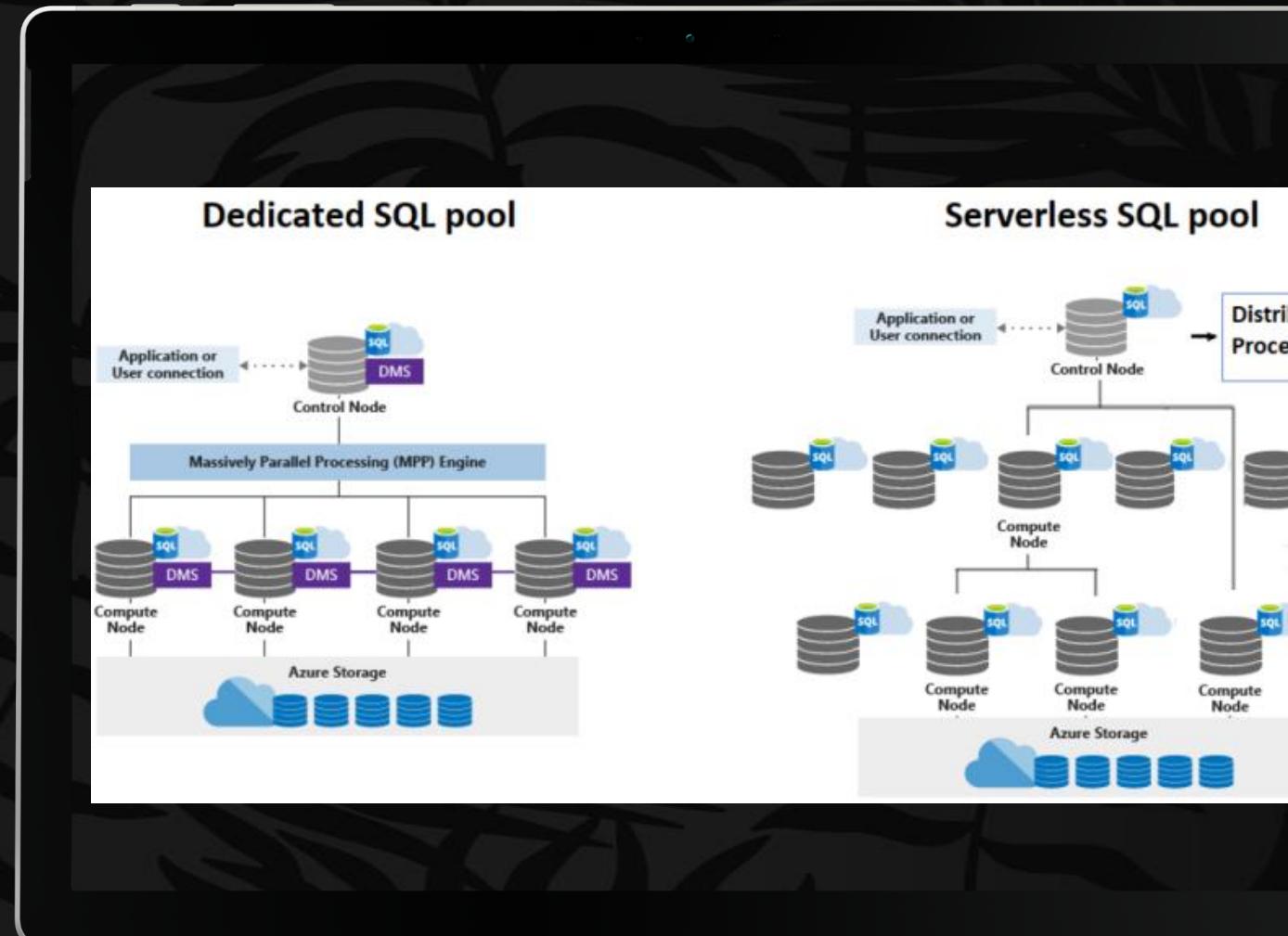
Serverless

Dedicated



# Synapse SQL

- Serverless pay-per-query ideal for ad-hoc data lake exploration and transformation
- Dedicated clusters optimized mission-critical data warehouse workloads
  - Minimum 1 TB, 60 million rows
  - Consider serverless
  - No small updates





# Synapse SQL

- Serverless pay-per-query ideal for ad-hoc data lake exploration and transformation
- Dedicated clusters optimized mission-critical data warehouse workloads
  - Minimum 1 TB, 60 million rows
  - Consider serverless
  - No small updates

Service Level	DWU	Pay as you go	1 year reserved
DW100c	100	€1,048.661/month	€660.6287/mo ~37% savings
DW200c	200	€2,097.322/month	€1,321.2573/mo ~37% savings
DW300c	300	€3,145.983/month	€1,981.8860/mo ~37% savings
DW400c	400	€4,194.644/month	€2,642.5146/mo ~37% savings
DW500c	500	€5,243.305/month	€3,303.1433/mo ~37% savings
DW1000c	1000	€10,486.610/month	€6,606.2865/mo ~37% savings

Serverless = **€4.757** per TB of data processed



# Synapse Spark

- New Cloud native Engine
- Integration with Spark
- Allows multiple languages in one notebook
- Offers use of temporary tables across languages
- Support for syntax highlight, syntax error, syntax code completion, smart indent, and code folding





# Synapse Data Explorer (PREVIEW)

- Building near Realtime Big Data Analytics on custom event/log data
- Security log analytics
- IOT Solutions analytics

Near real-time insights on big data

Optimized engine for interactive query exploration

Agile telemetry service for time series data and IoT devices

Automated Scaling from GBS to Petabytes



# 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 in the navigation pane, displaying a list of database objects like 'wwi.DimStockItem', 'wwi.DimSupplier', etc. In the center, an 'SQL script 5' tab is open with the following query:

```
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
```

On the right, the results are displayed as a bar chart titled 'Quantity' versus 'City'. The chart shows the total quantity for various cities, with 'Sekiu' having the highest value (approximately 19k) and 'Trentwood' having the lowest (approximately 13k). The chart has a 'Chart' tab selected.



# Synapse Studio

## Notebook IDE code authoring

- 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 Studio interface. The left sidebar displays a tree view of resources under 'Develop', including 'SQL scripts' (16 items), 'Notebooks' (35 items, expanded to show 'Demo notebooks' with 11 sub-items like '000 Data Exploration on Cosmos...', '000 Retail-scoring-onnx-HB-AML...', etc., and 'MS Conf notebooks' with 3 items), and 'Test notebooks' (1 item). The right pane shows a code editor for a notebook titled '020 Surface Sales Fo...'. The code is written in Python and performs operations like importing pandas, aligning outputs, and merging datasets. The interface includes a toolbar with 'Cell', 'Run all', 'Publish', 'Attach to', 'Language' (set to 'PySpark (Python)'), and a status bar with '020 Surface Sales Forecasting with Synapse Analytics'.

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



# Synapse Studio

## Automatic machine learning

- 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 workspace interface. On the left, the 'Data' section displays various databases and tables. A Jupyter notebook cell is open in the center, containing the following Python code:

```
from pandas import *
from azureml import *
from matplotlib import *
from automl import *
def align_(df):
    if (horizon == 'short-term'):
        df = df[['retail_sales']]
    else:
        df = df[['retail_sales']]
    # align X_train, X_test, X_test_
    X_train, X_test, X_test_ = df['X'].iloc[:train_size], df['X'].iloc[train_size:-test_size], df['X'].iloc[-test_size:]
    # drop clean :
    return X_train, X_test, X_test_
X_test[time_index] = df_all['retail_sales'].values[time_index]
# use autoencoder
```

To the right, there's a sidebar titled 'Choose a model type' with three options: 'Classification', 'Regression', and 'Time series forecasting'. Each option has a description and an example.

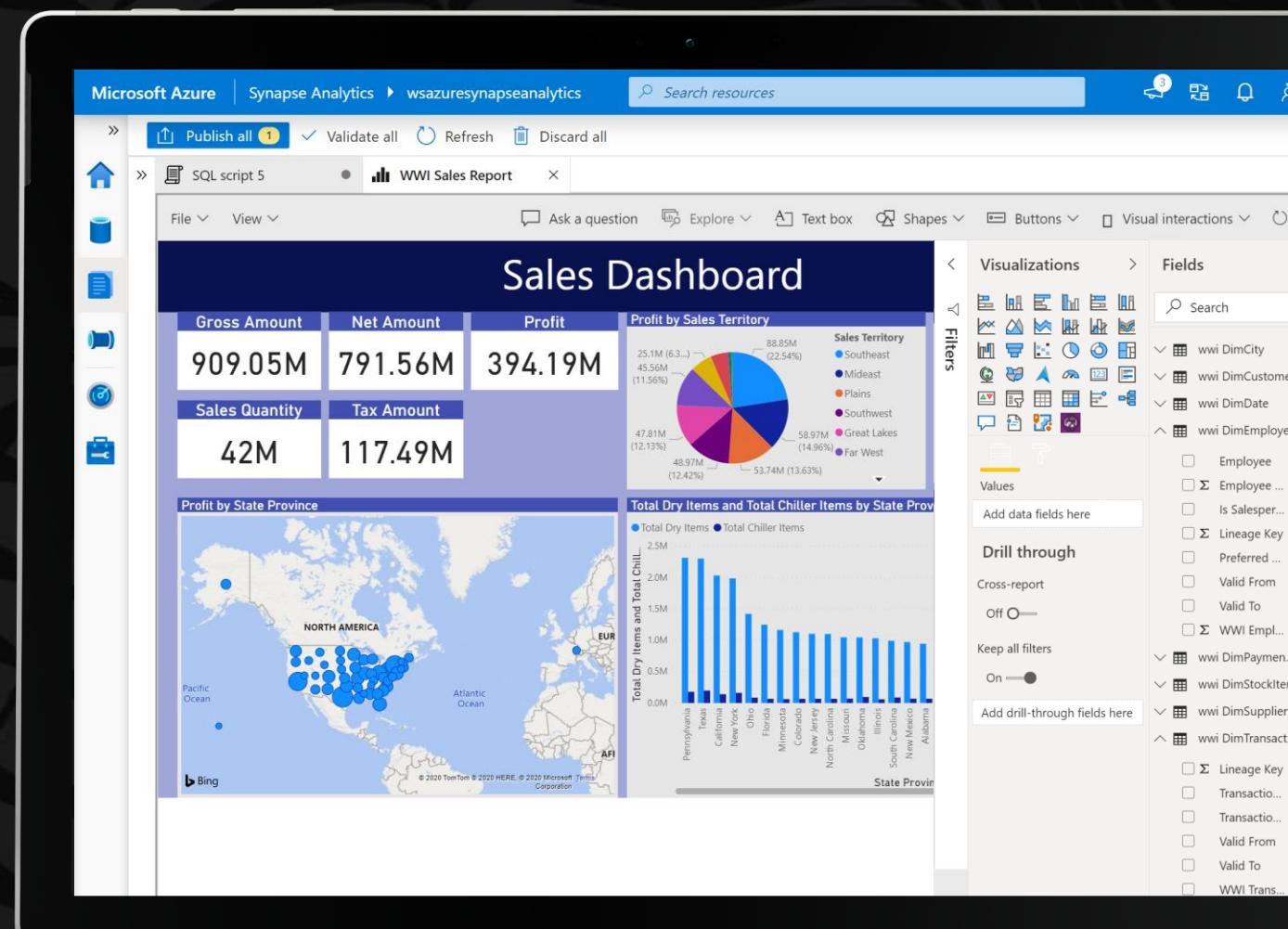
- Classification**: Determine the likelihood of a specific outcome being achieved (binary classification) or identify the category an attribute belongs to (multiclass classification). Example: Predict if a customer will renew or cancel their subscription.
- Regression**: Estimate a numeric value based on input variables. Example: Predict housing prices based on house size.
- Time series forecasting**: Estimate values and trends based on historical data. Example: Predict stock market trends over the next year.



# Synapse Studio

## Power BI Integration

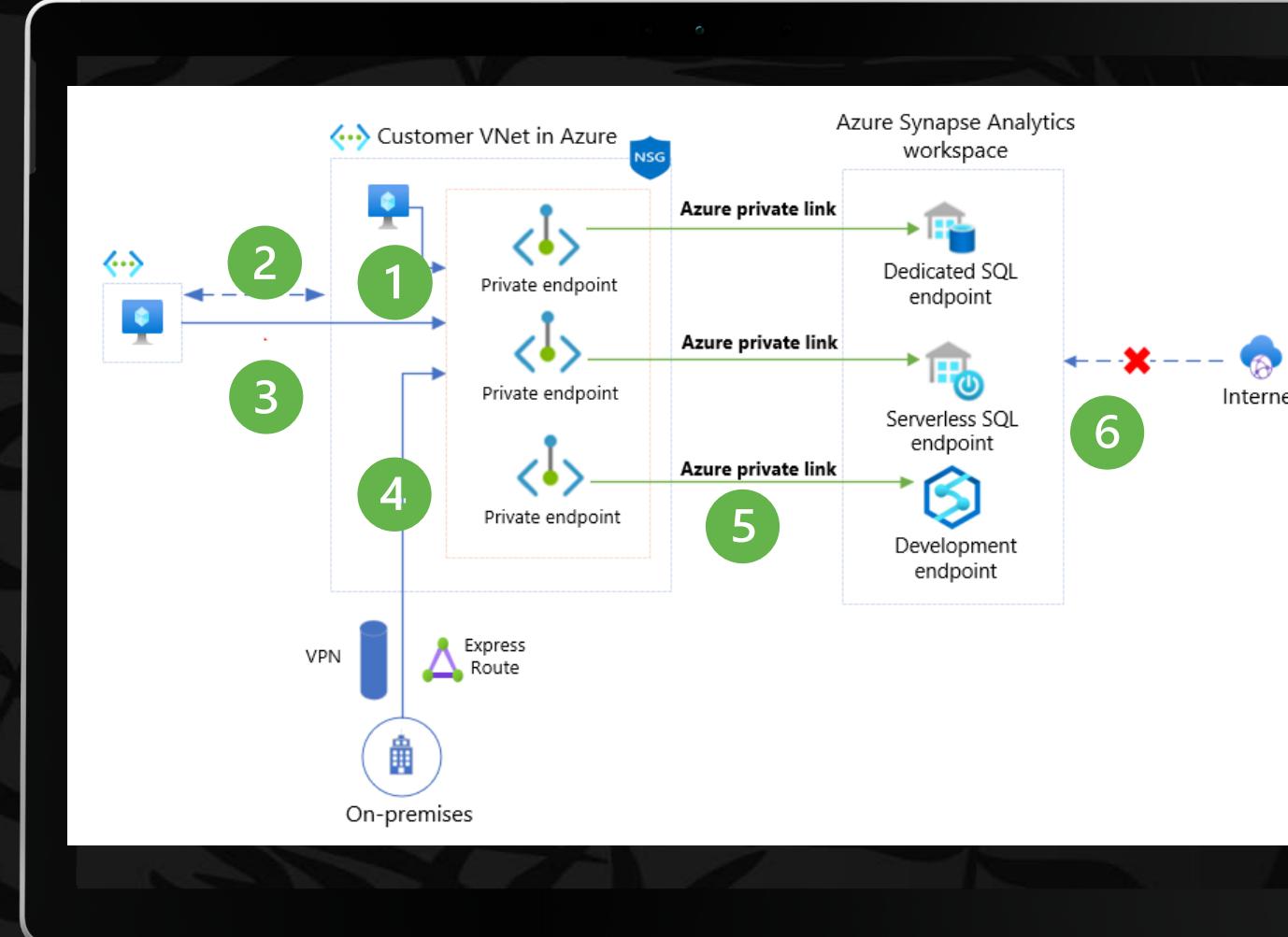
- Build dashboard in Synapse Studio





# Synapse Network Security

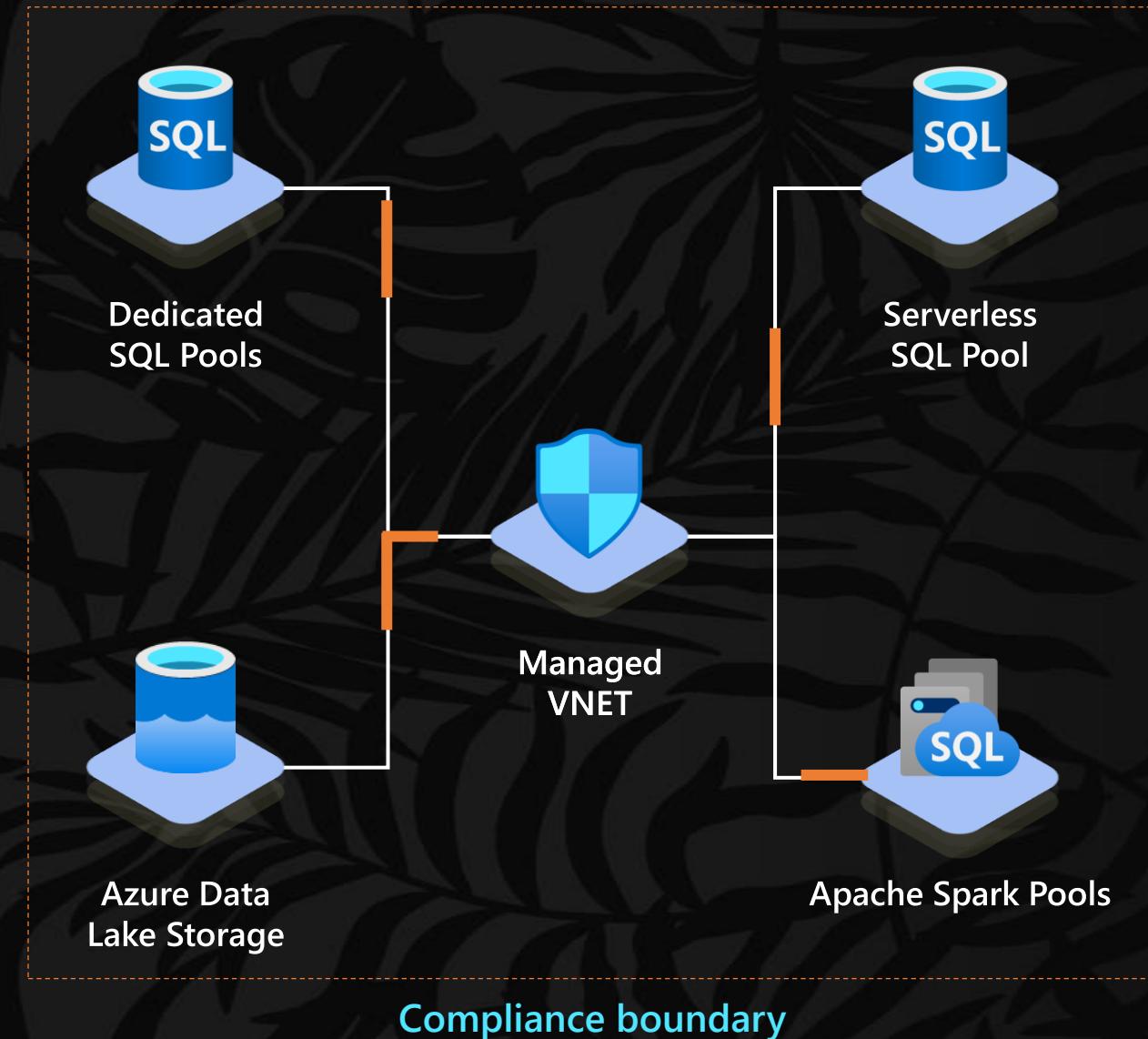
- Workstations from within the customer VNet access the Azure Synapse private endpoints.
- Peering between customer VNet and another VNet.
- Workstation from peered VNet access the Azure Synapse private endpoints.
- On-premises network access the Azure Synapse private endpoints through VPN or ExpressRoute.
- Workspace endpoints are mapped into customer's VNet through private endpoints using Azure Private Link service.
- Public access is disabled on the Synapse workspace.





# Managed Virtual Networks

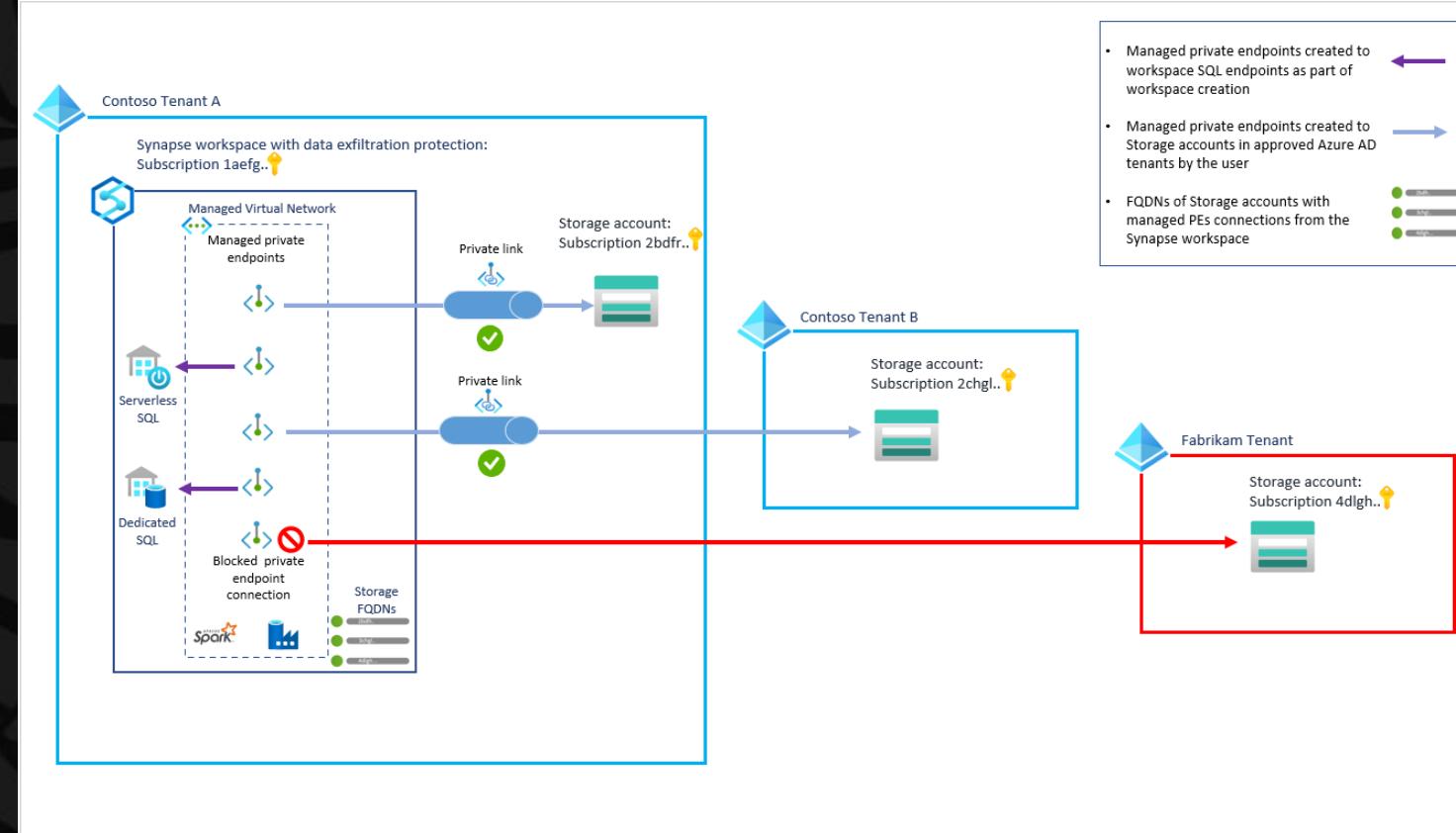
- Eliminate network maintenance
- One-click enables automated management of virtual networks between cluster endpoints
- Synapse resources only ever interop with private endpoints
- No management of subnets or IP Ranges
- Prevents data exfiltration





# Data exfiltration protection

- Managed Virtual Network
- Exfiltrating sensitive data to locations outside



# DEMO

## Creating a Synapse Analytics Environment

Azure Synapse Analytics ...

Microsoft

 Azure Synapse Analytics Add to Favorites

Microsoft | Azure Service

★ 4.2 (65 ratings)

Plan

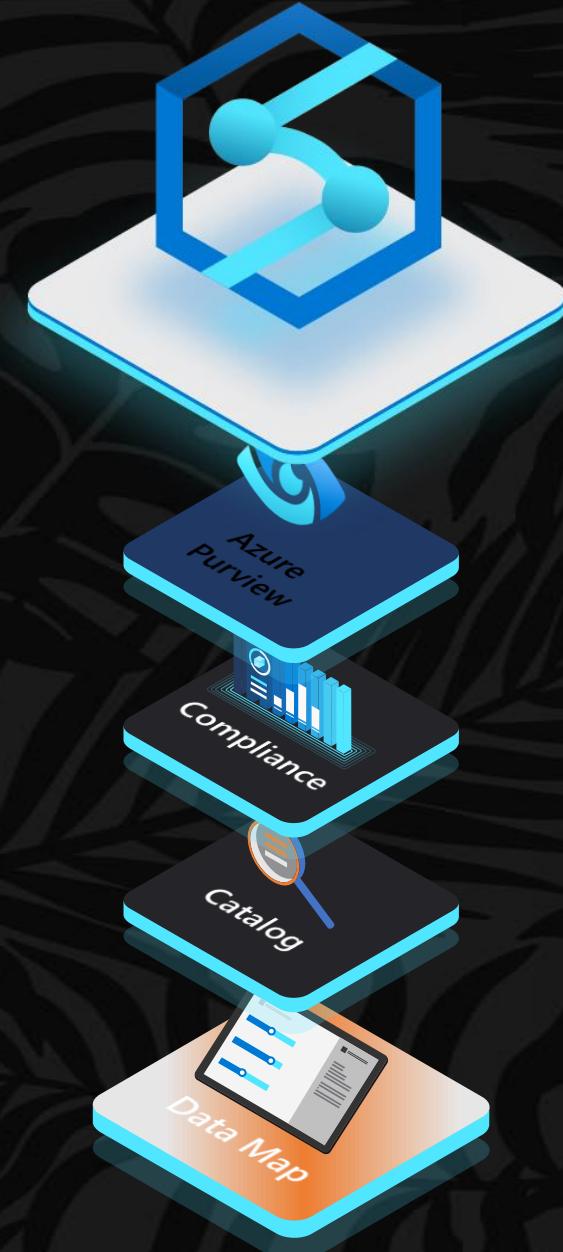
Azure Synapse Analytics ... Create





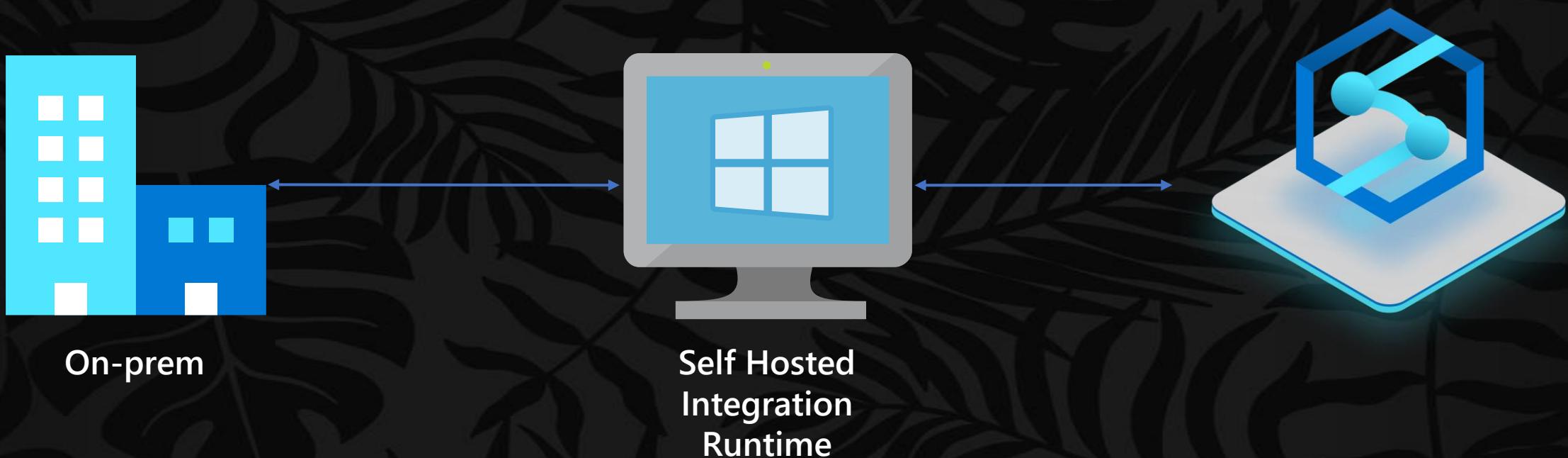
# Governance

- Native integration with Azure Purview
- Automatically discover and classify data assets
- End-to-end data lineage





# Connecting to On-Prem





# Connecting Power BI /Fabric



public access  
allowed/Denied



Power BI  
Gateway



public access allowed/Denied

Public network access

Choose whether to permit public network access to your workspace. You can modify the firewall rules after you enable this setting. [Learn more](#)

Enable  Disable

**i** You must use private endpoints to connect to your workspace when this setting is disabled. Selecting the **Disable** option will not apply any firewall rules that you may configure.

- ▷ Azure Private Link  
*Enabled for the entire organization*
- ▷ Block Public Internet Access  
*Enabled for the entire organization*

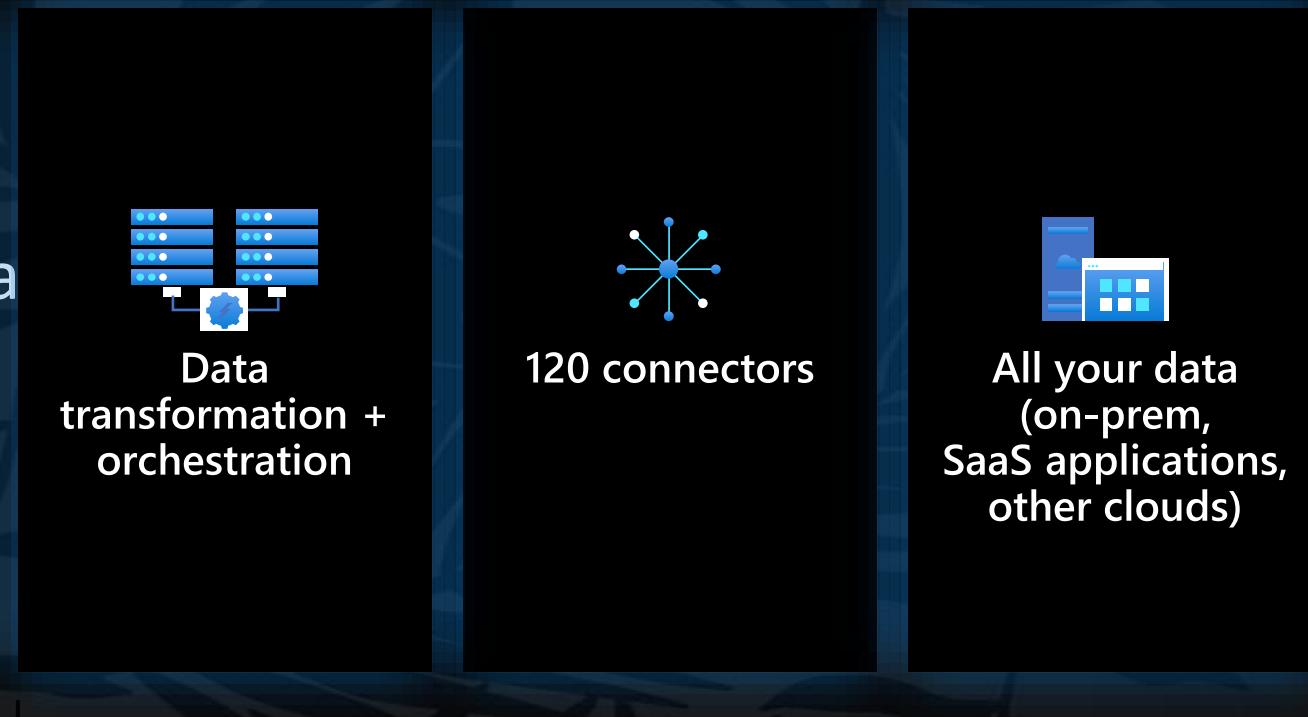
# Data Integration





# Synapse Pipelines

- Ingest Data from 120+ sources
- Cloud native ETL/ELT
- 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





# Synapse Pipelines

- No/low-code data transformation
- Excel-like interface is familiar to users
- Transform data to desired shape completely visually
- Operationalize into pipelines

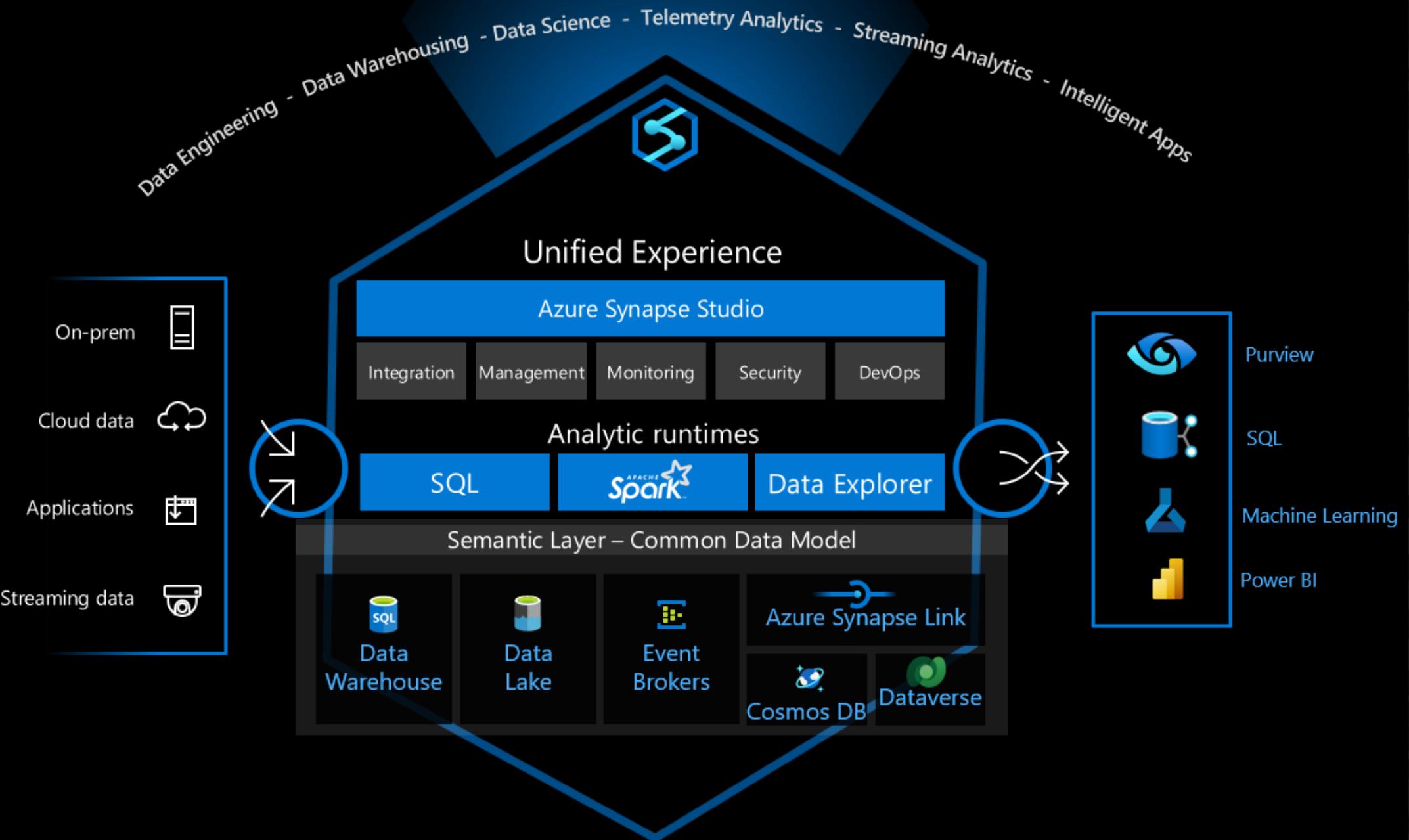
The screenshot shows the Microsoft Azure Synapse Analytics Power Query Editor interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', 'wsazuresynapseanalytics', and a search bar. Below the navigation is a toolbar with various icons for validation, publishing, and discarding changes. The main workspace is titled 'PQSalesPrep' and contains a 'Settings' section with a note about Power Query M functions. The 'Home' tab is selected, showing options like 'Enter data', 'Options', 'Manage parameters', and 'Refresh'. The 'Transform' tab is also visible. A 'Queries' pane on the left lists 'ADFResource' (1 item) and 'UserQuery'. The main area displays a table titled 'Table.TransformColumnTypes(Source, {"quantity", Int64.Type}, {"logQuantity", type number})'. The table has columns: ab storeId, ab productCode, 12 quantity, 1.2 logQuantity, ab advertising, ab price, ab weekStarting, ab id. The data shows rows of surface.go items with various quantities and dates. At the bottom, it says 'Columns: 8 Rows: 99+'. The status bar at the bottom right indicates 'Complete'.



# DEMO

Querie Data





# Final Words

- Azure Synapse Analytics is Unified Data Platform
- Blending Data Integration, Data Warehouse and Big Data Analytics
- SQL/Spark/Data Explorer analytical runtime's
- Easily query on Data Lake files
- Power BI Integrates
- Native integration with Microsoft Purview



# Thank You



**Erwin de Kreuk**

Principal Consultant – Lead Data & AI  
InSpark



@erwindekreuk



[linkedin.com/in/erwindekreuk](https://linkedin.com/in/erwindekreuk)



[erwindekreuk.com](http://erwindekreuk.com)



[github.com/edkreuk](https://github.com/edkreuk)



<https://sessionize.com/erwin-de-kreuk/>



**Let's connect**