



Azure Synapse Analytics

December 2021



From Data to Intelligence

Drive a data culture and power a new
class of data first applications

everyone | every decision | at any scale

Azure Synapse and Power BI

Data
Integration



Analytics



Business
Intelligence



The only vendor with leadership in all MQ's

Data Integration



Analytics



Business Intelligence



Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose. This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from Microsoft.

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.



Synapse



Power BI

Data
Integration

Data
Engineering

Data
Warehouse

Data
Science

Observation
Analytics

Business
Intelligence

Governance



Data Integration



Data Integration

Over 100 connectors to ingest
data from a variety of platforms

Integrate from On-Premise, PaaS, and SaaS

Batch and Real-time data integration

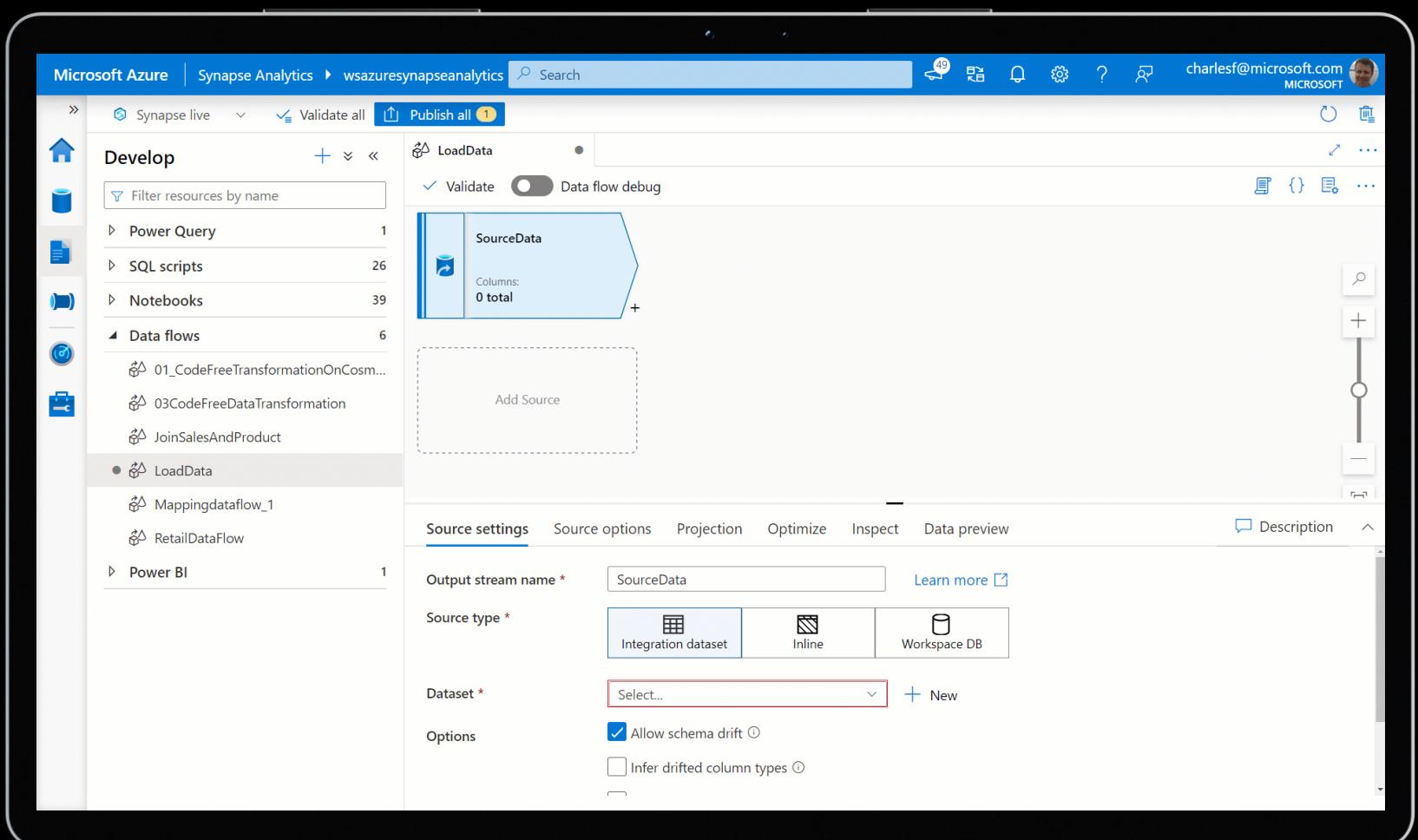
Secure hybrid connectivity

Code-free development environment

Generally Available

100+ Connectors

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

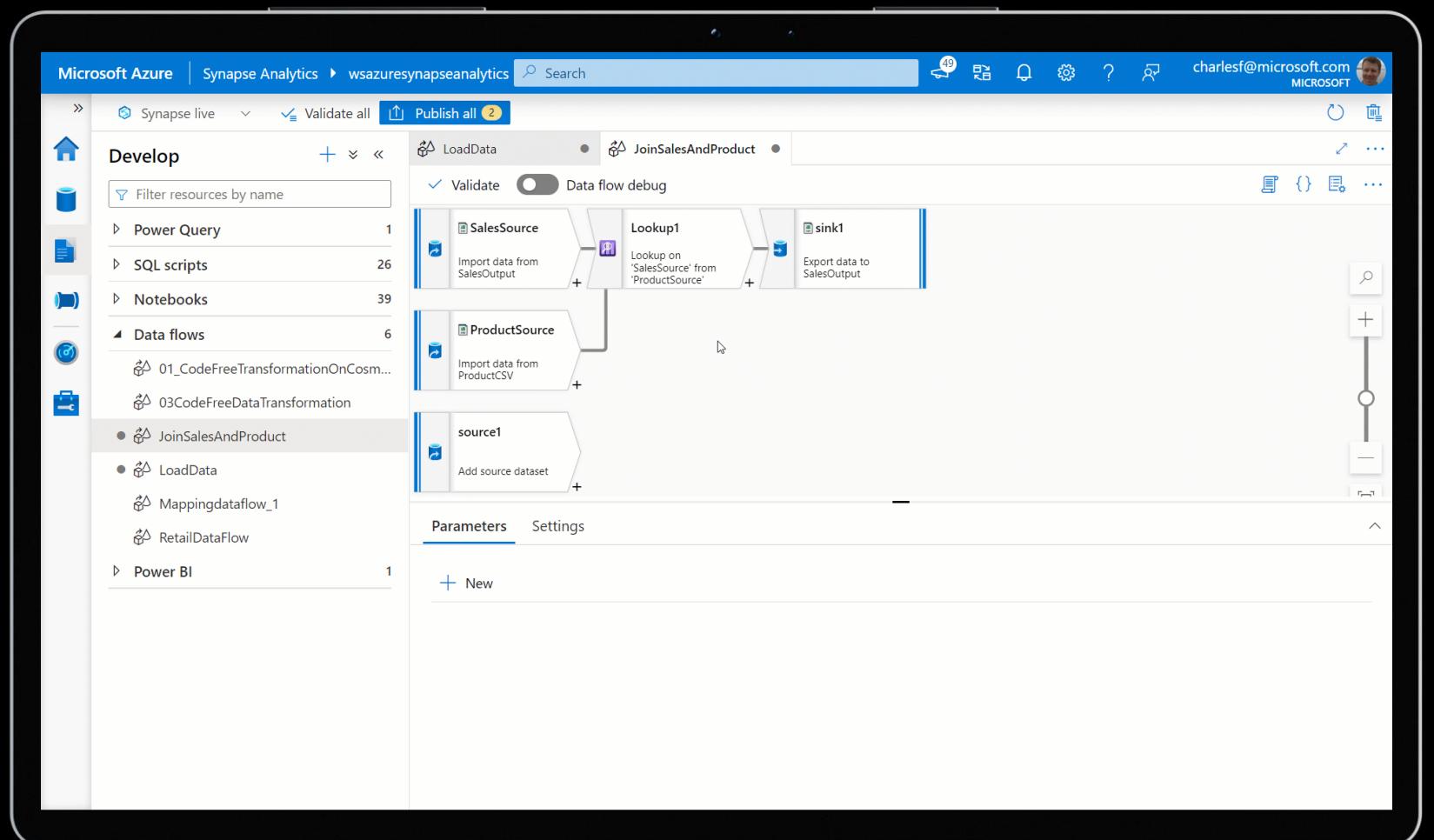


Generally Available

Code-free Data Flows

Enables developers to rapidly integrate data from a variety of sources

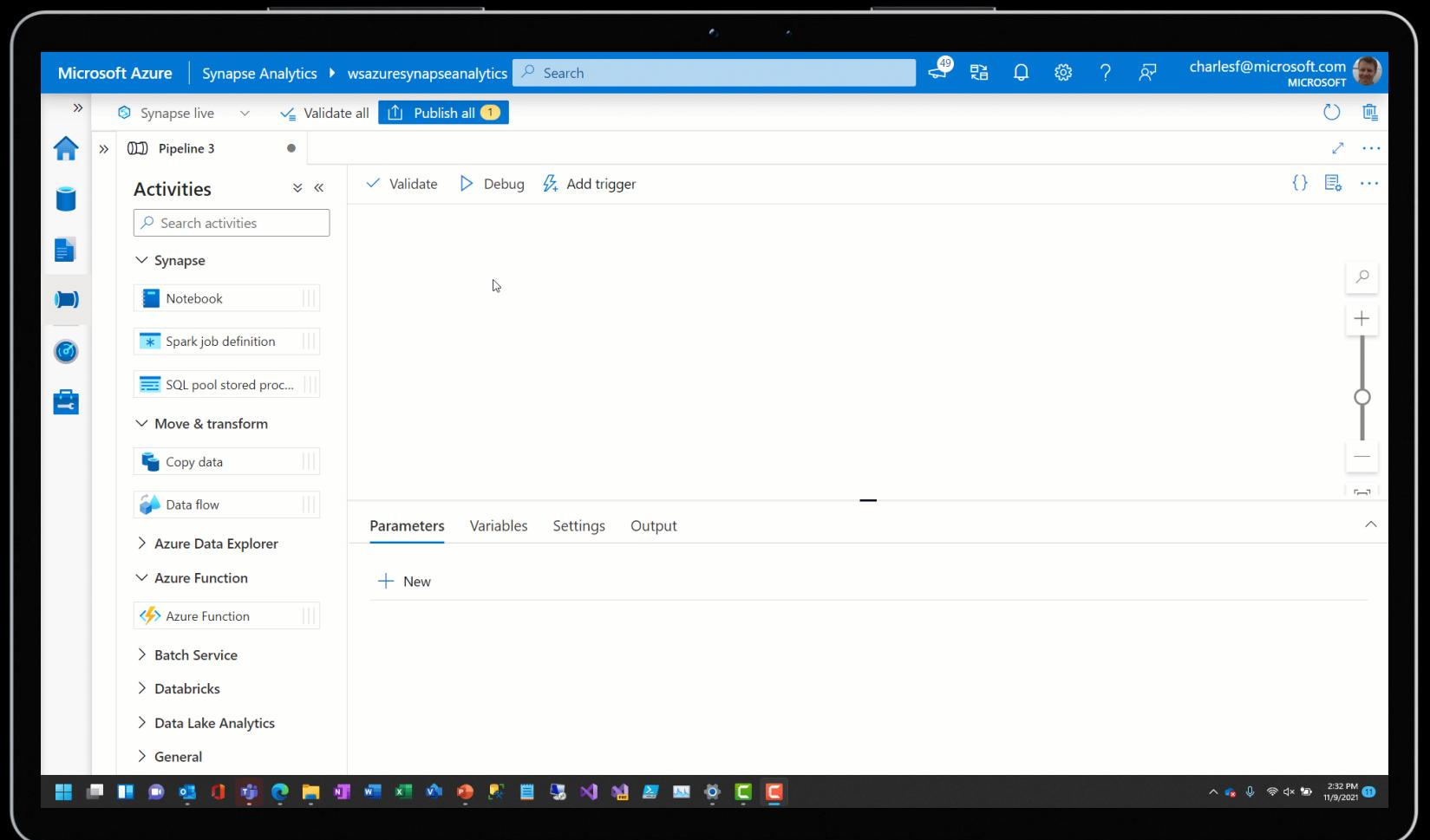
Execute on Spark for large scale processing



Generally Available

Pipeline Orchestration

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



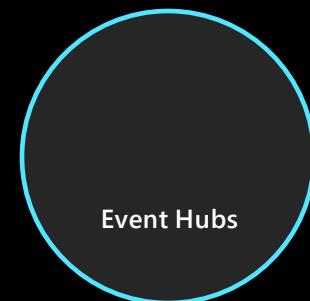


Generally Available

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



Event Hubs



Stream
Analytics

SQL Query
Language



Data
Warehouse



Data Lake

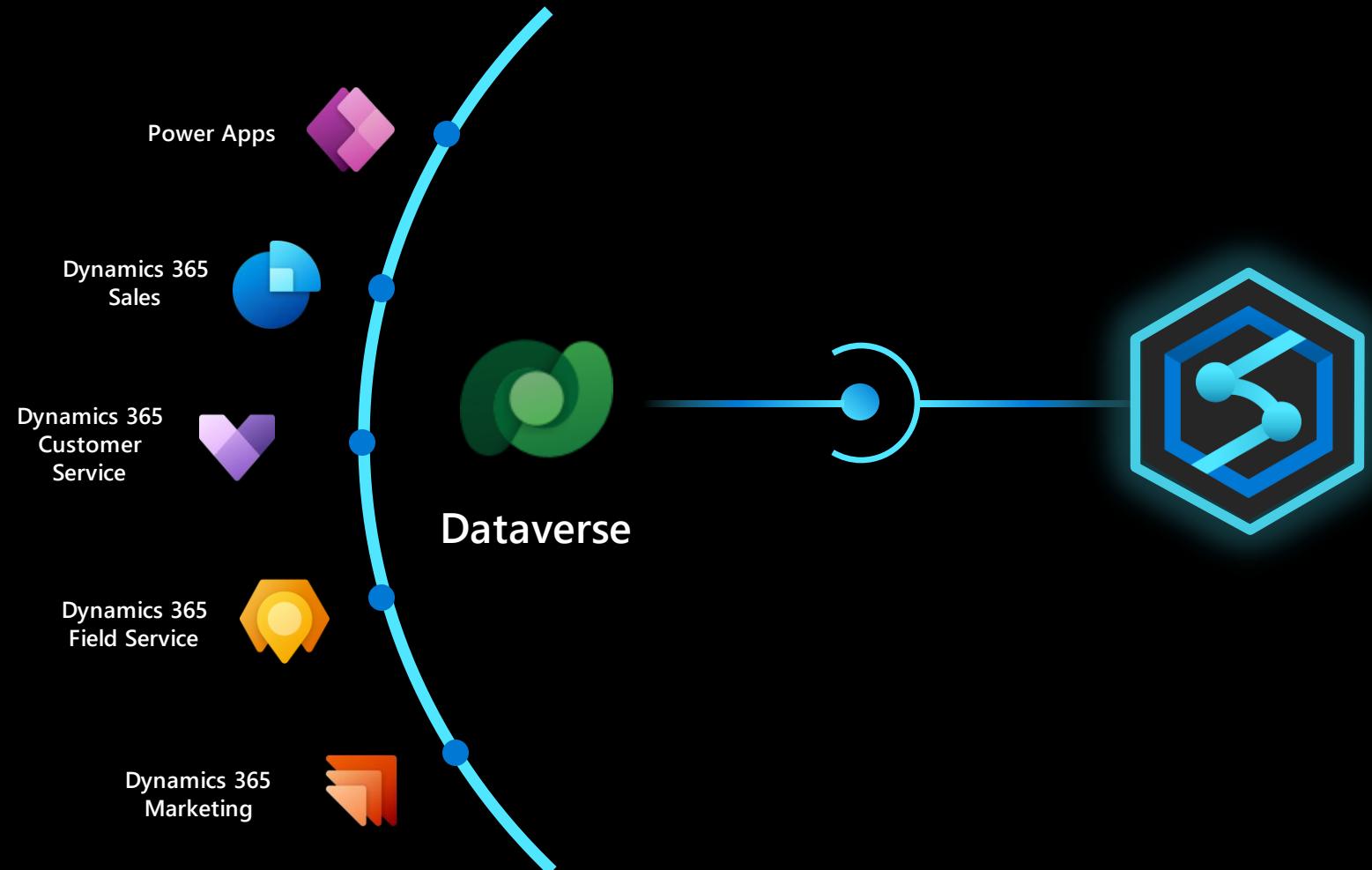
Generally Available

November 2021

Synapse Link for Dataverse

One-click integration of D365
data into Synapse for analytics

No data pipelines required



Public Preview

Q2 2022

Parquet & Virtual Network Support for Dataverse

Parquet columnar file format optimizes query performance for user queries

Enables customers to apply Virtual Network security to Dataverse connection



Public Preview

Q2 2022 (SQL Server 2022)

Synapse Link for Microsoft SQL

Near real-time operational analytics in Synapse

No data pipelines required

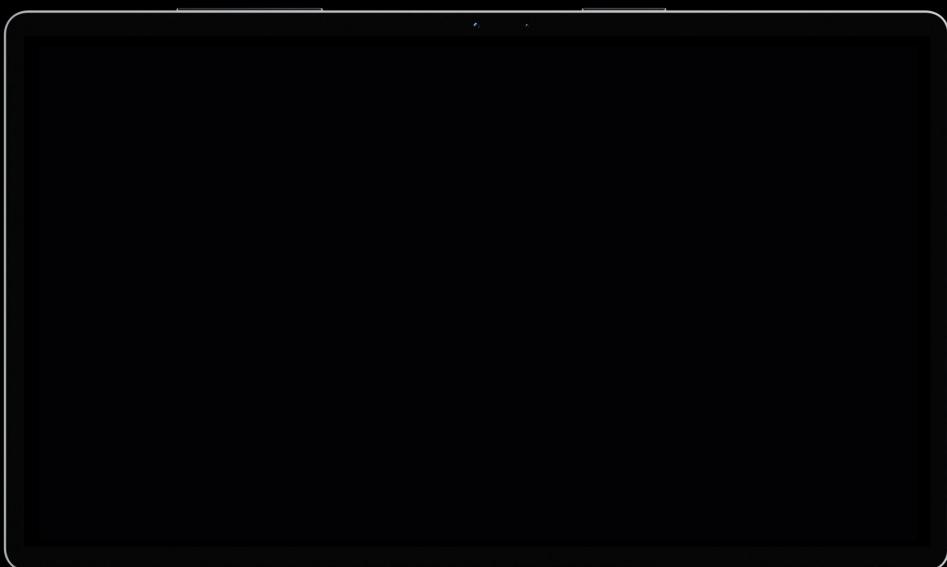
Hybrid integration for SQL Server running on-premise or other clouds

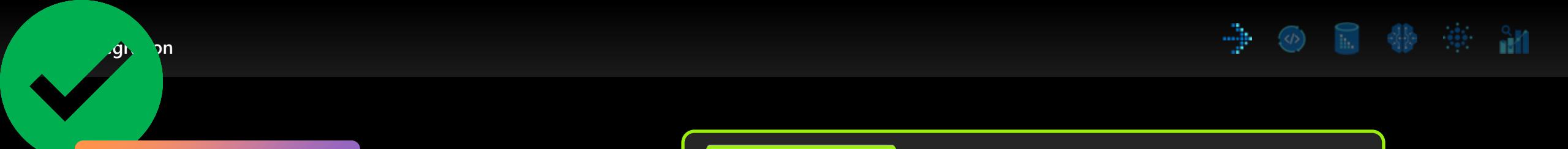


SQL Server 2022
Public Preview Q2 2022



Azure SQL Database
Public Preview Soon





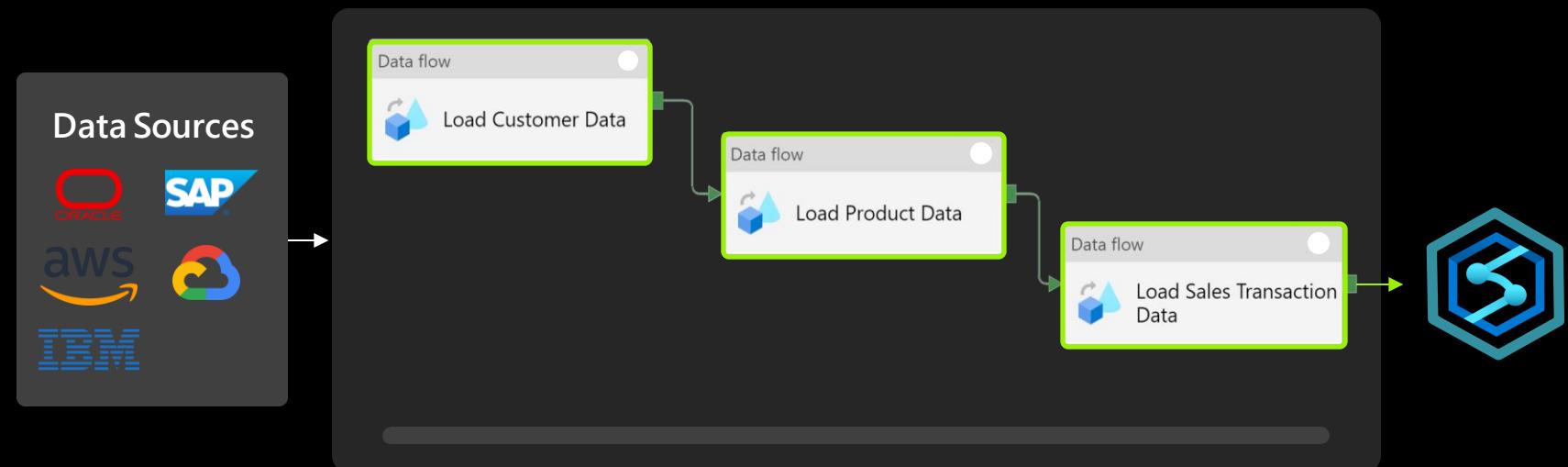
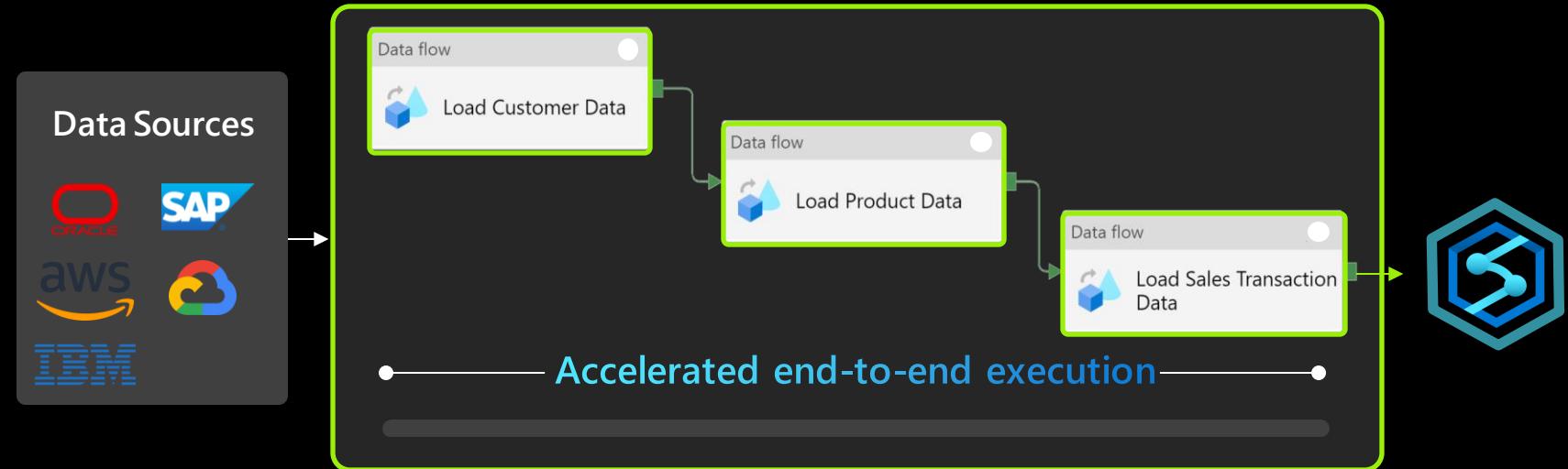
Generally Available

October 2021

Accelerated Data Pipelines

Cluster time-to-live enables near instant start of data flow pipelines for faster data integration

Data is available to the business faster to enable more timely decision making





Data Engineering

Data Engineering

Scalable Spark engine

Industry standard languages

Delta Lake enabled

Azure DevOps integrated

Public Preview

Q1 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

RocksDB integration for managing state

Maintain streaming state for improved restartability

Spark Streaming Enhancements

Enables streaming jobs to run for virtually infinite duration



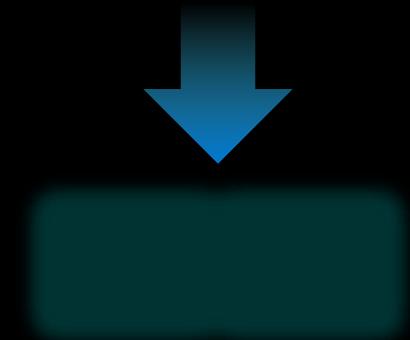
Public Preview

Q1 2022

Delta Lake Performance Enhancements

Improved performance and
reduced cost with support for
OPTIMIZE and Z-ORDER

OPTIMIZE



Improve query performance
by coalescing small files into
larger ones

Z-ORDER



Improves filter query performance
by ordering data for fast lookups on
large datasets

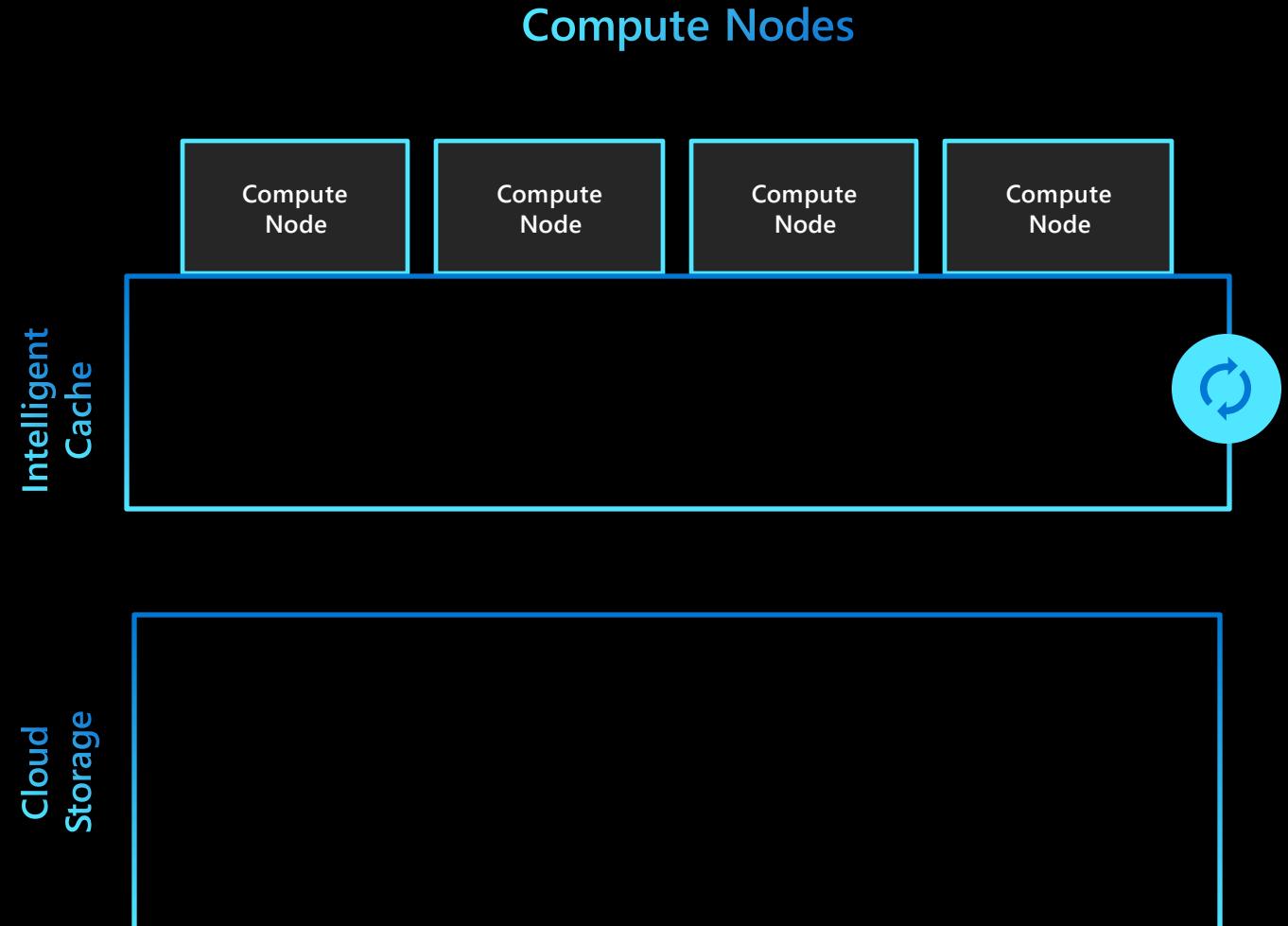


Public Preview

Q1 2022

Spark Caching Enhancements

Intelligent cache automatically detects changes in data to ensure data is fresh and results are accurate





Data Warehousing



Data Warehousing

Cornerstone of enterprise analytics for decades

Industry standard SQL language

Structured and semi-structured data

Broad ecosystem of applications

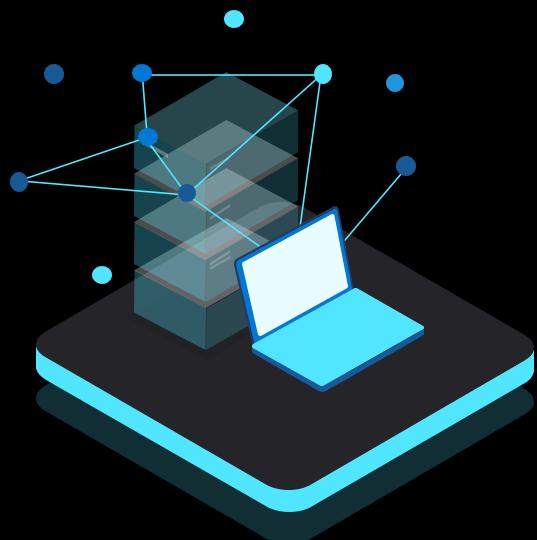
Fine-grained data security

Data models tailored to business consumption

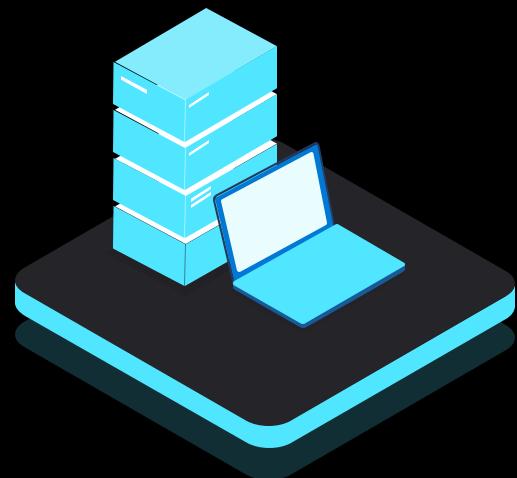
Generally Available

Dedicated & Serverless SQL

Elastic clusters with in-memory caching provide enterprise class performance combined with cloud economics



Serverless



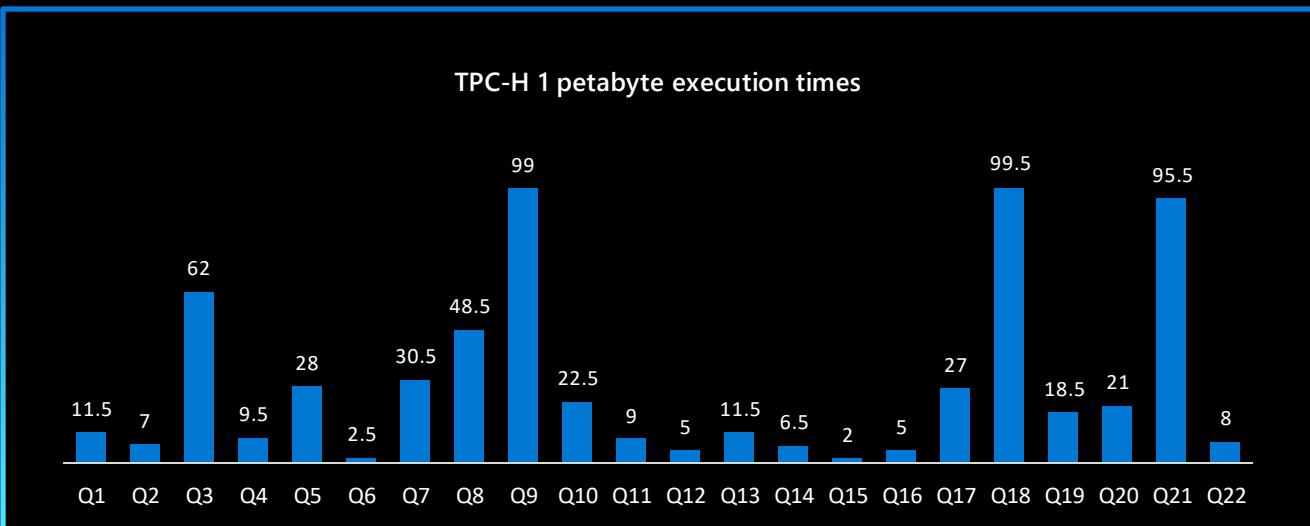
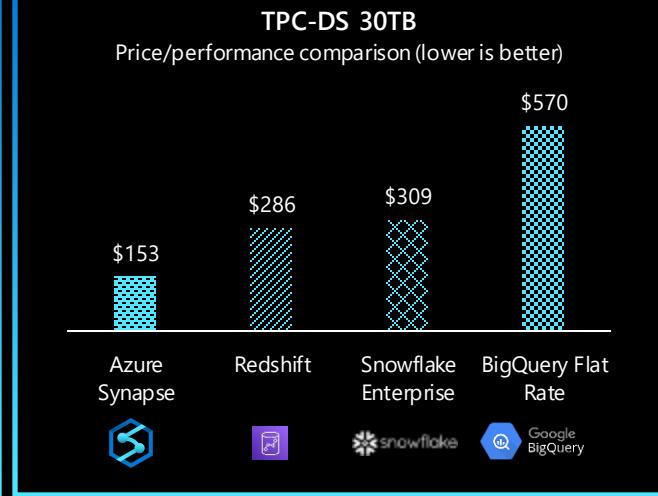
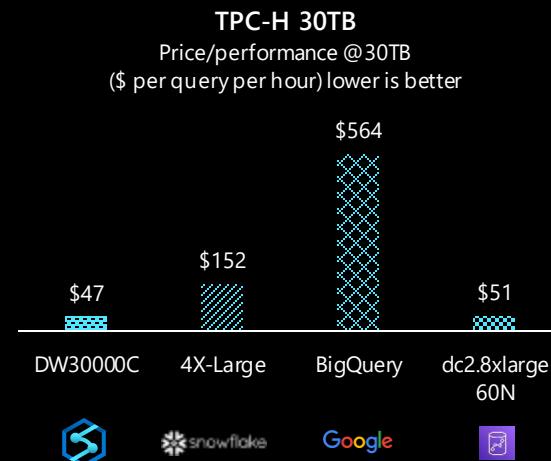
Dedicated

Generally Available

Industry Leading Performance

Price/performance leader for data warehousing

The only platform to complete TPC-H at 1PB



Generally Available

Most Complete Workload Management

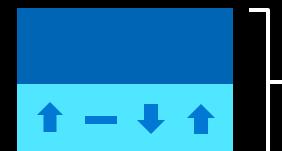
Scale-in to maximize output with the predictable cost

Scale-out to leverage cloud scale resources for spikes in demand

Scale-In

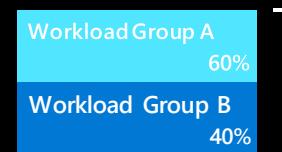
- Predictable cost
- Prioritize higher value work
- Prevents global contention

Workload Importance



Data Warehouse

Workload Isolation



Data Warehouse

Scale-Out

- Add compute for variable workloads
- Pause compute when idle

Elastic Cluster (Scale Up)



Data Warehouse

Generally Available

Complete Data Protection

Democratize data compliantly with fine-grained access controls and multi-level encryption

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	✓
Authentication	Column level encryption	✓
	SQL login	✓
	Azure active directory	✓
	Multi-factor authentication	✓
Network Security	Managed virtual network	✓
	Custom virtual network	✓
	Firewall	✓
	Azure ExpressRoute	✓
Threat protection	Azure Private Link	✓
	Threat detection	✓
	Auditing	✓
Isolation	Vulnerability assessment	✓
	Dedicated metadata store	✓
	Hosted in customer tenant	✓

Generally Available

Democratize ML predictions with SQL

In-engine ML scoring provides interactive query response times without any data leaving the system and no additional scoring cost



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

Synapse SQL



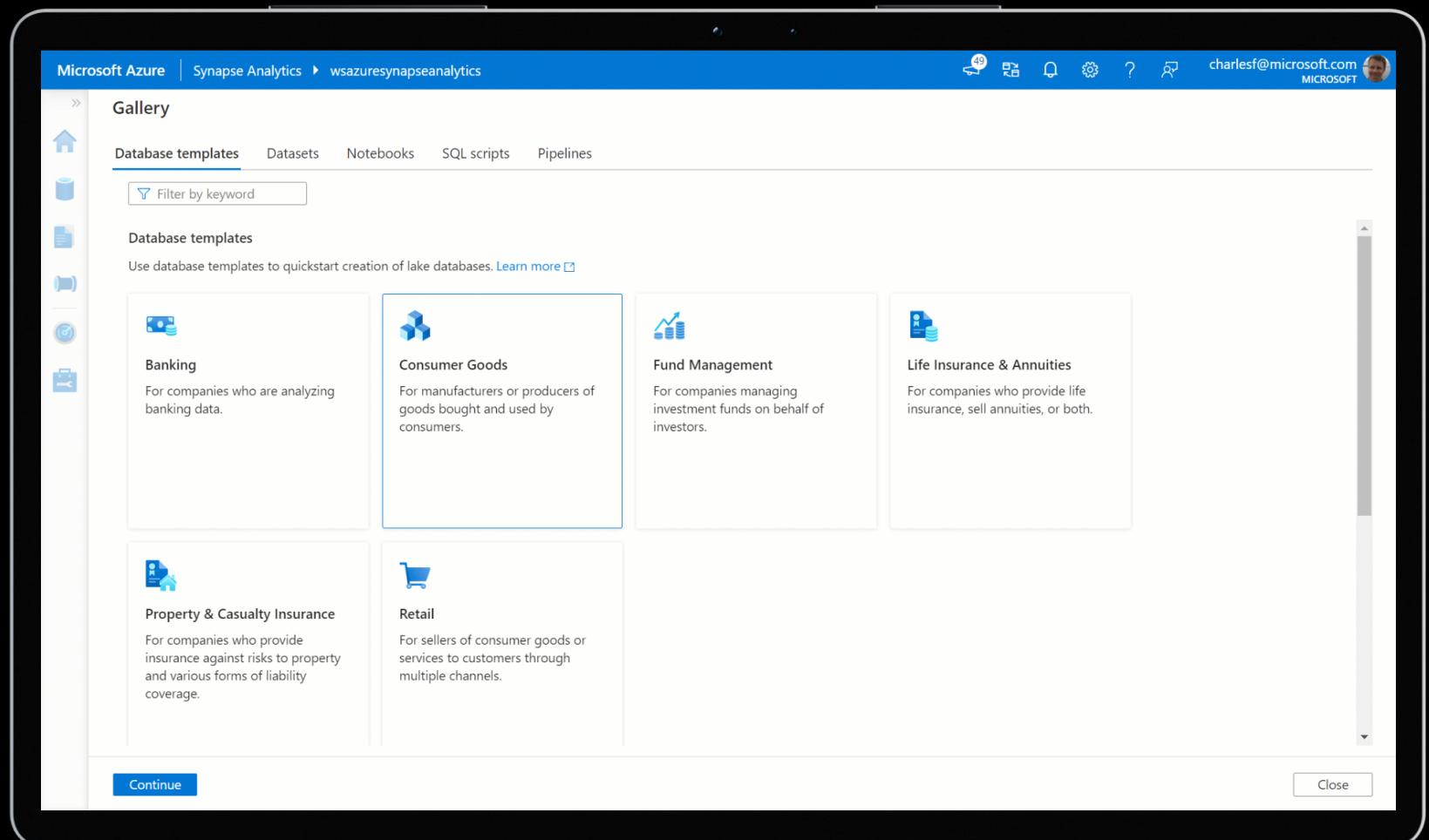
Public Preview

November 2021

Database Templates

Built-in database templates

Low code database designer



Public Preview

Q2 2022



Data Warehouse Restore SLA

Guaranteed SLA for
database restore
reduces disaster
recovery downtime

Microsoft Azure (Preview) Search resources, services, and docs (G+)

Home > demosynapsev3 > New dedicated SQL pool

*Basics *Additional settings Tags Review + create

Create a dedicated SQL pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

Dedicated SQL pool details

Name your dedicated SQL pool and choose its initial settings.

Dedicated SQL pool name *

Enter dedicated SQL pool name

Performance level ⓘ DW1000c

Estimated price ⓘ

Est. Cost Per Hour
12.00 USD
[View pricing details](#)

[Review + create](#) [Next: Additional settings >](#)

Public Preview

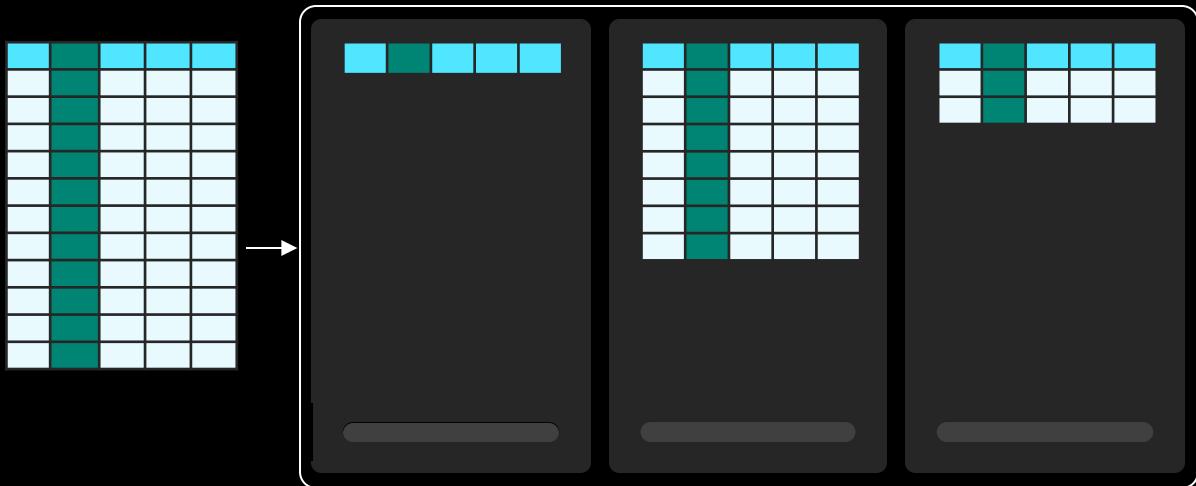
Q2 2022

Multi-column distributed table sharding

Improved query performance and easier migrations

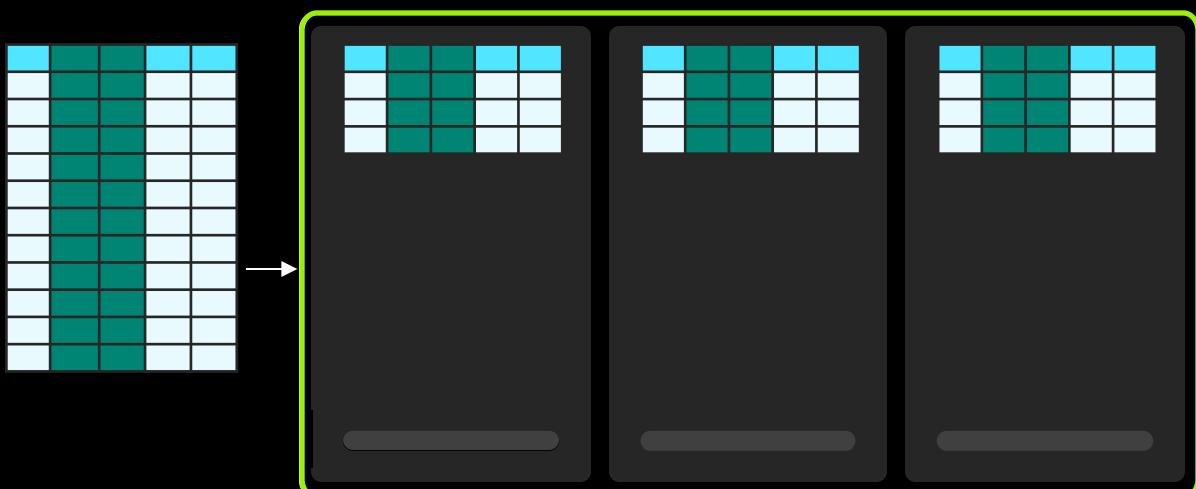
Uneven Distribution (Skew)

```
CREATE TABLE SalesTransactions (
    WITH DISTRIBUTION =
        (HASH(ProductKey))
```



Balanced Distribution

```
CREATE TABLE SalesTransactions (
    WITH DISTRIBUTION =
        (HASH(ProductKey, RegionKey))
```



Balanced Distribution: Queries execute faster

Public Preview

Q2 2022

MERGE SQL Statement

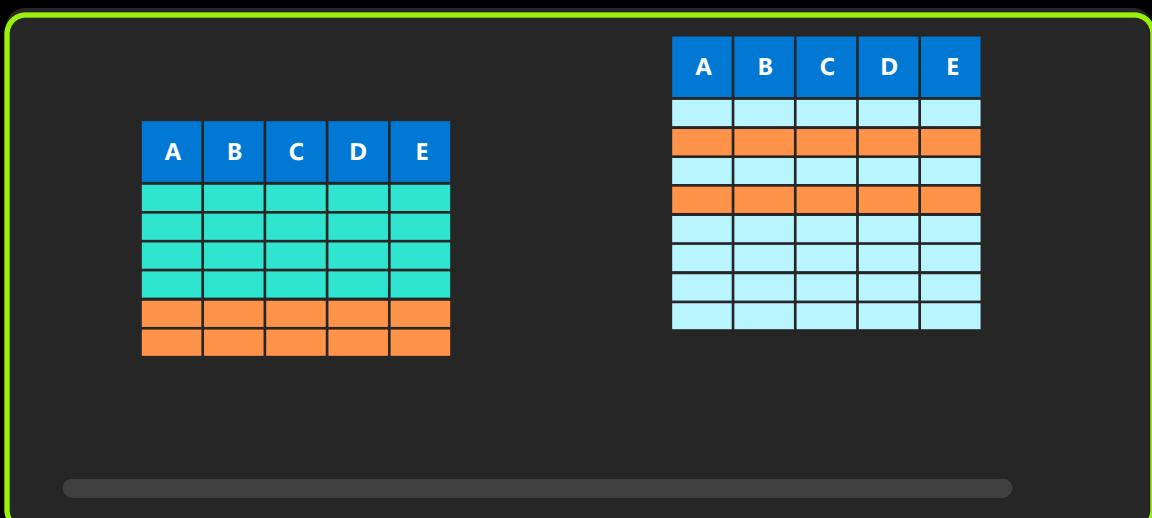
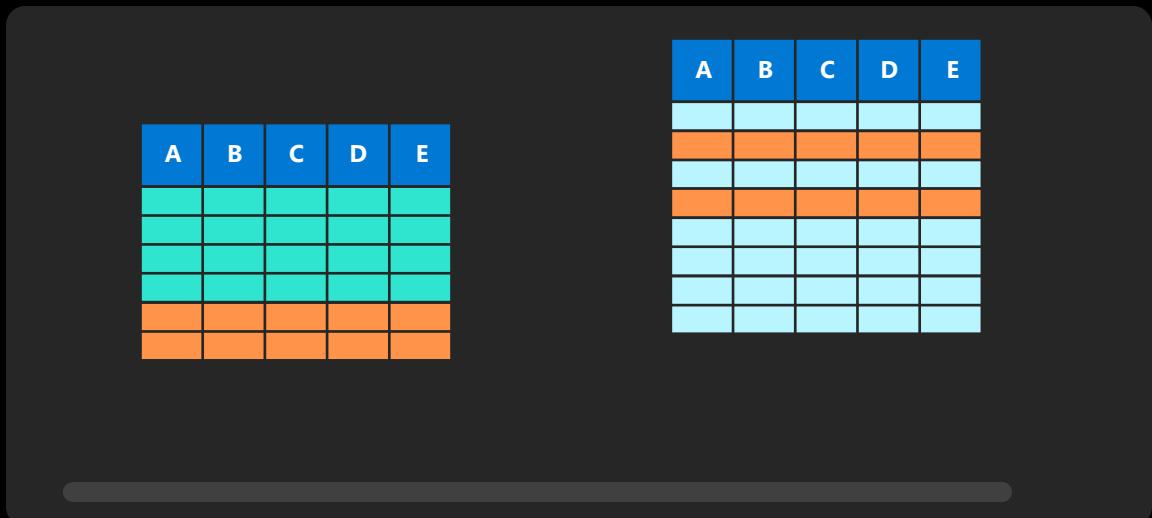
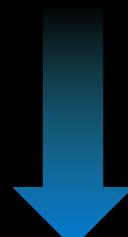
Improved performance and easier migration by executing INSERT, UPDATE, and DELETE functionality in a single statement

MERGE ...

INSERT ...

UPDATE ...

DELETE ...



Generally Available

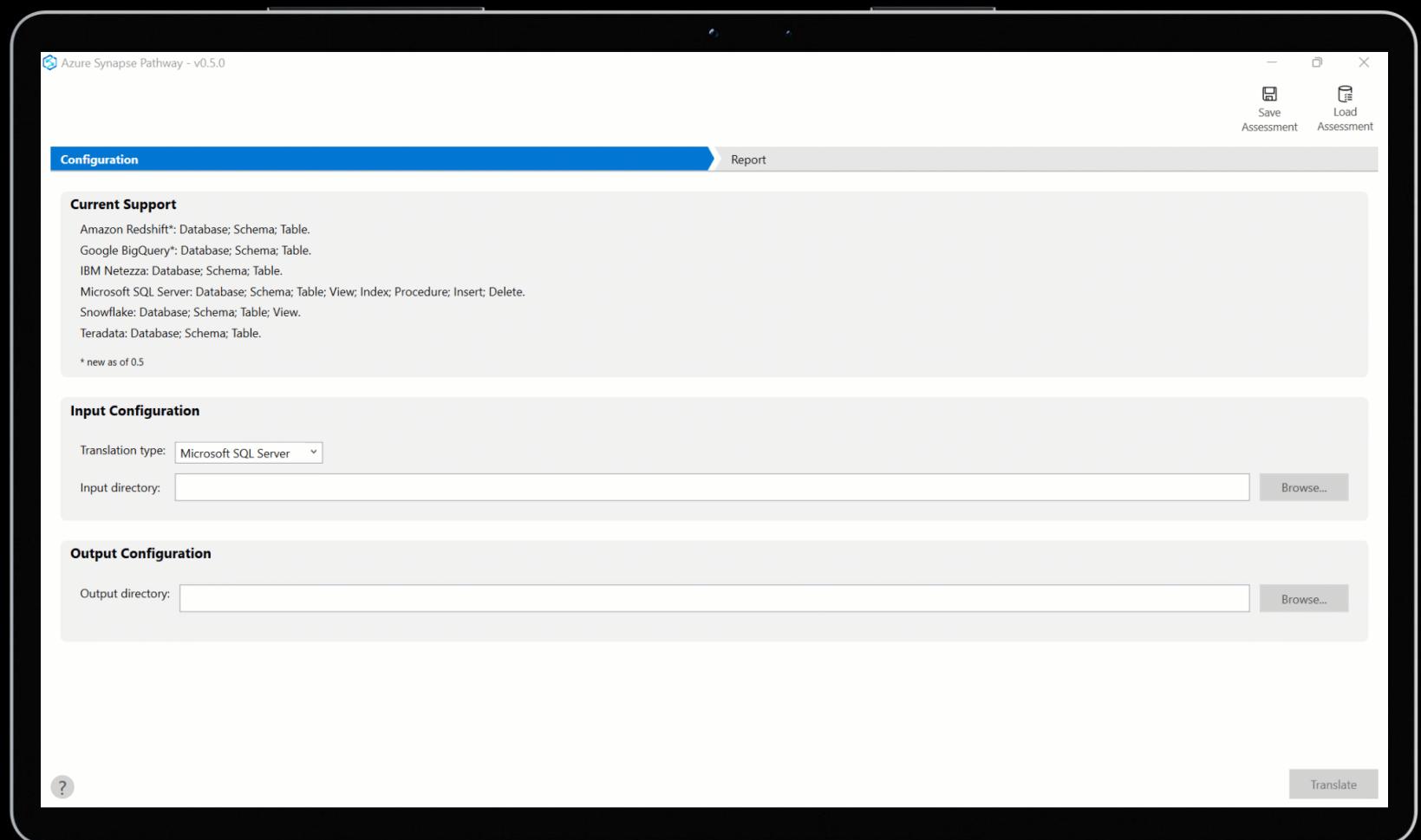
Updated Frequently

Free SQL Migration Utility

Automate SQL conversion
from existing systems

Supported Sources

- Teradata
- Redshift
- Netezza
- Snowflake
- SQL Server
- BigQuery





Data Science

Data Science

Industry standard languages such as PySpark

Code-first and Code-free Auto ML

Integrated ML Model Repository

Hardware accelerated GPU model training

Generally Available

Notebook Development Experience

Empower data scientists will a familiar Notebook based development interface

The screenshot shows the Microsoft Azure Synapse Analytics notebook interface. The left sidebar displays a tree view of notebooks, categorized into 'Develop', 'SQL scripts', 'Notebooks' (which includes 'Demo notebooks' and 'MS Conf notebooks'), and 'Test notebooks'. A specific notebook titled '020 Surface Sales Forecasting with Synapse' is selected. The main workspace shows a single code cell containing Python code for forecasting. The code imports various libraries like pandas, numpy, and matplotlib, defines a function 'align_outputs', and performs data manipulation and merging of datasets. The interface includes standard notebook controls like 'Cell', 'Run all', 'Publish', and 'Attach to'.

```
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[[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
```

Generally Available

Built-in Cognitive Services

Enables simple integration of pre-built machine learning models

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, the 'Data' section displays a 'Workspace' tab and a 'Linked' tab. Under 'Lake database', there is a 'default' folder containing several parquet tables: myparquettable, myparquettable2, myparquettable3, myparquettable5, myparquettable6, and retailsales. Below these are 'Views', 'surfacesalesdb', and 'SQL database'. In the center, a code editor window shows Python code for interacting with an Azure ML workspace:

```
1 import azureml.core
2 import pandas as pd
3 import numpy as np
4 import logging
5 from azureml.core.works
6 from azureml.core import
7 from azureml.core.exper
8 from azureml.train.auto
9 import os
10 subscription_id = os.get
11 resource_group = os.get
12 workspace_name = os.get
13 workspace_region = os.g
14
15 ws = Workspace(subscrip
16 ws.write_config()
17
18 experiment_name = 'auto
19 experiment = Experiment
20 output = {}
21 output['Subscription ID
22 output['Workspace'] = w
23 output['SKU'] = ws.sku
24 output['Resource Group'
25 output['Location'] = ws
26 output['Run History Nam
27 pd.set_option('display.
28 outputDF = pd.DataFrame
```

To the right, there are sections for 'Predict with a model' (checkbox for 'retailsales') and 'Choose a pre-trained model'. Under 'Azure Cognitive Services', it says: 'This experience allows you to enrich the selected dataset with pre-trained Azure Cognitive Services models.' It lists two options: 'Anomaly Detector' (described as identifying rare items) and 'Sentiment Analysis' (described as evaluating sentiment). At the bottom right are 'Continue' and 'Cancel' buttons.

Generally Available

Automatic Machine Learning

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

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a sidebar with icons for Home, Databases, Workspaces, and Jupyter notebooks. The main area shows a Jupyter notebook cell with the following Python code:

```
1 import azureml.core
2 import pandas as pd
3 import numpy as np
4 import logging
5 from azureml.core.workspace import Workspace
6 from azureml.core.experiment import Experiment
7 from azureml.train.automl import AutoMLConfig
8 import os
9 subscription_id = os.getenv('AZUREML_SUBSCRIPTION_ID')
10 resource_group = os.getenv('AZUREML_RESOURCE_GROUP')
11 workspace_name = os.getenv('AZUREML_WORKSPACE_NAME')
12 workspace_region = os.getenv('AZUREML_WORKSPACE_REGION')
13
14 ws = Workspace(subscription_id=subscription_id,
15                 resource_group=resource_group,
16                 workspace_name=workspace_name,
17                 workspace_region=workspace_region)
18 experiment_name = 'auto'
19 experiment = Experiment(ws, experiment_name)
20 output = {}
21 output['Subscription ID'] = ws.subscription_id
22 output['Workspace'] = ws.workspace_name
23 output['SKU'] = ws.sku.name
24 output['Resource Group'] = ws.resource_group
25 output['Location'] = ws.location
26 output['Run History Name'] = experiment.runs[-1].name
27 pd.set_option('display.max_rows', 10)
28 outputDf = pd.DataFrame([output])
```

To the right of the code cell, there's a 'Train a new model' wizard. It starts with a section titled 'Train a new model' with a 'retailsales' dataset selected. A note says: 'This wizard will help you to train a machine learning model using Automated Machine Learning.' Below that is a 'Choose a model type' section with three options: 'Classification', 'Regression', and 'Time series forecasting'. Each option has a brief description and an example. At the bottom right of the wizard are 'Continue' and 'Cancel' buttons.

Generally Available

Industry Standard Open Ecosystem

Open file formats enable easy integration with other data services

Industry standard languages make it easy for developers to get started

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a sidebar titled 'Develop' containing sections for Power Query, SQL scripts, Notebooks, Data flows, and Power BI. Under 'Notebooks', 'Notebook 5' is selected. The main area is a Jupyter Notebook titled 'rawdata'. It has three code cells:

```
1 import azureml.core
2 from azureml.core import Experiment, Workspace, Dataset, Datastore
3 from azureml.train.automl import AutoMLConfig
4 from azureml.data.dataset_factory import TabularDatasetFactory
```

```
1 subscription_id = "58f8824d-32b0-4825-9825-02fa6a801546"
2 resource_group = "plangadr"
3 workspace_name = "anlwsdemos"
4 experiment_name = "wsazuresynapseanalytics-retailsales-20210216065932"
5
6 ws = Workspace(subscription_id = subscription_id, resource_group = resource_group, workspace_name = workspace_name)
7 experiment = Experiment(ws, experiment_name)
```

```
1 df = spark.sql("SELECT * FROM retaildata.retailsales")
2 datastore = Datastore.get_default(ws)
3 dataset = TabularDatasetFactory.register_spark_dataframe(df, datastore, name = experiment_name + "-dataset")
```

```
1 automl_config = AutoMLConfig(spark_context = sc,
                                task = "regression",
```

Public Preview

Q2 2022



R Language Support

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

The screenshot shows the Microsoft Azure Synapse Analytics R Samples interface. The left sidebar lists various notebooks and scripts. The main area displays a code snippet for creating a DataFrame from a local R data.frame:

```
1 library(SparkR)
2 df <- createDataFrame(faithful)
3 
4 # Displays the content of the DataFrame to stdout
5 head(df)
```

The output shows the first six rows of the 'faithful' dataset:

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55

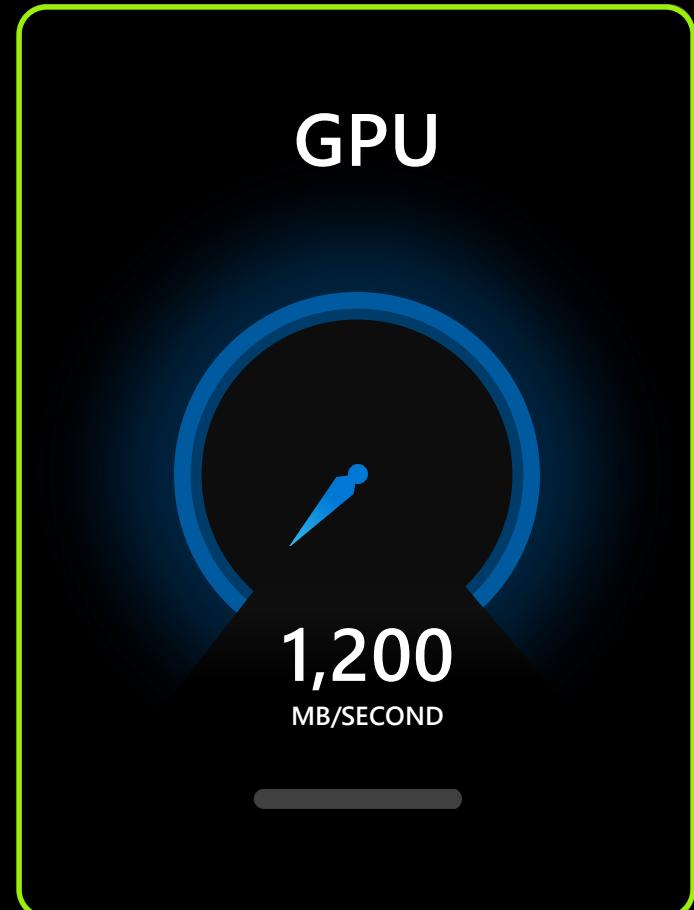
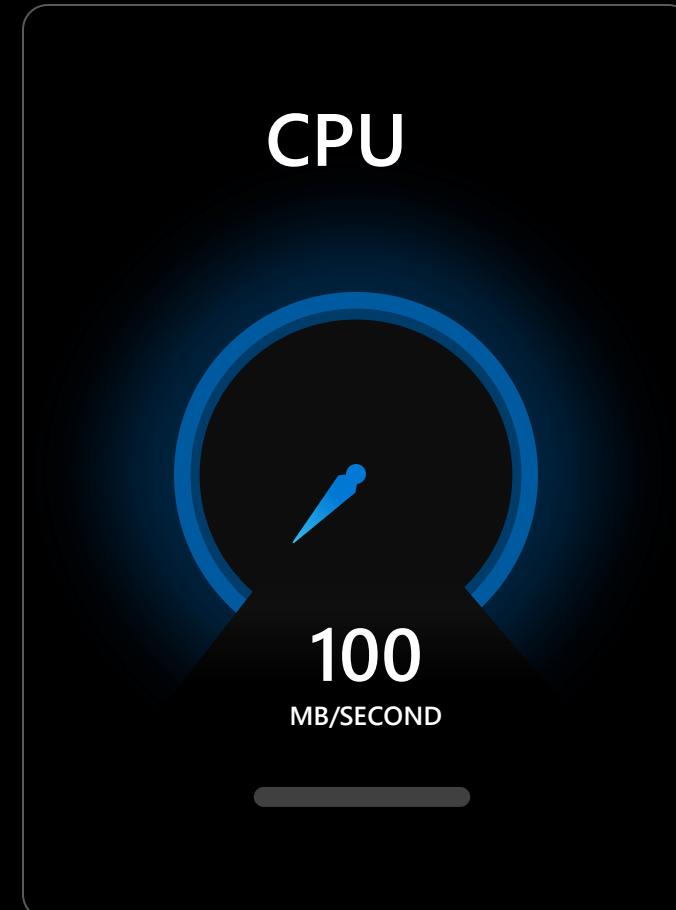
Below this, another section titled "From SQL Query" is visible.

Public Preview

November 2021

GPU Accelerated Workloads

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





Observational Analytics

Observational Data

The fastest growing data segment

50 BN
connected devices
by 2030



175 ZB
total amount of
data by 2025

Observational Data

What is it?

Semi-structured: text, json, time series

**Machine generated or machine recorded
human interactions**

Mass volume

High velocity

Few large fact tablesstreams

Observational Data

Why is it challenging to analyze?

Looking for unpredictable phenomena

Constantly changing schema

Near real time visibility required

Analytics systems costs are often prohibitive

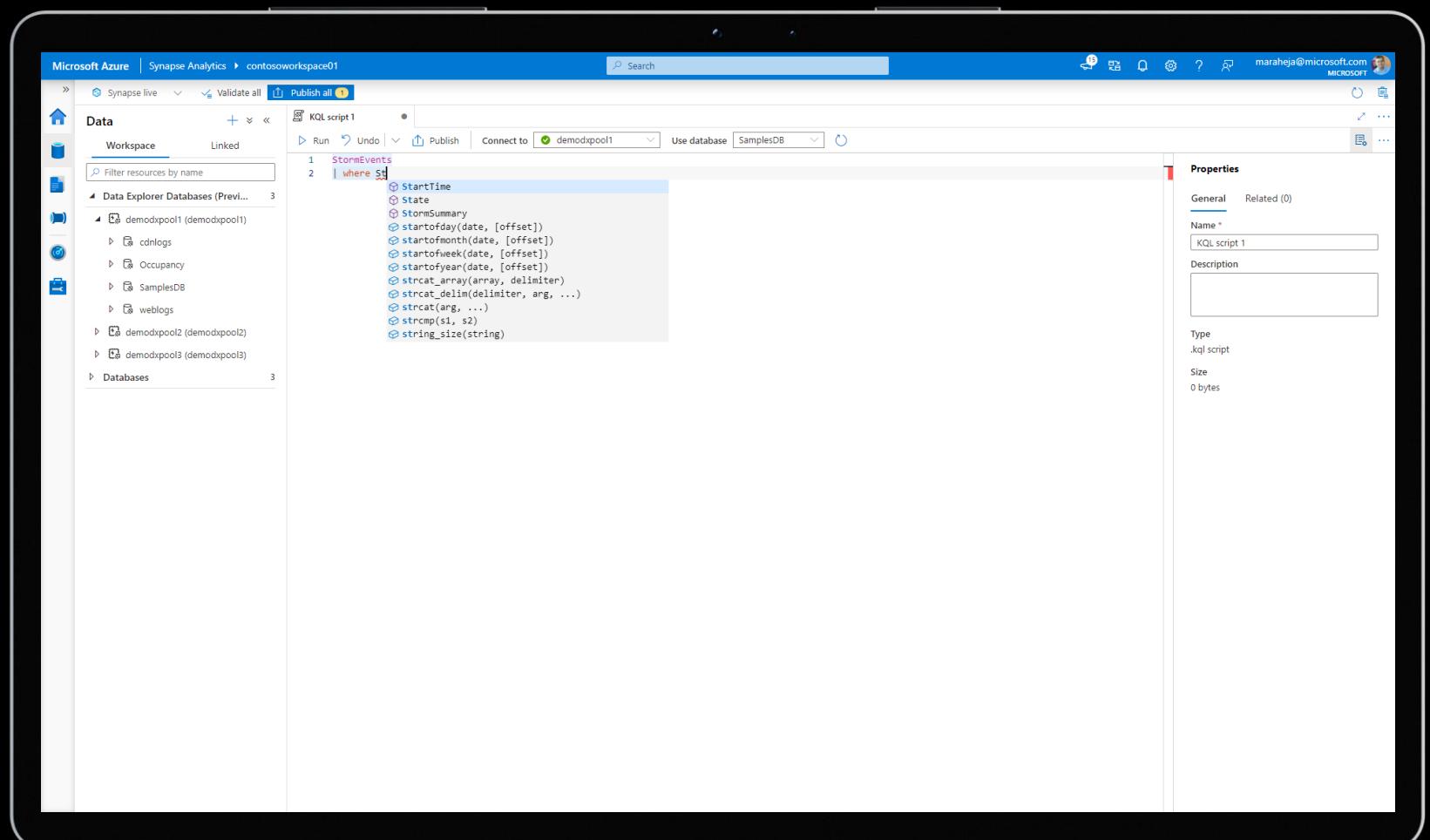
Frequently changing business questions

Near real-time analytics on
Observational data at petabyte scale

Public Preview

Synapse Data Explorer Engine

Industry leading
free-text and semi-
structured data
indexing for sub second
observational analytics



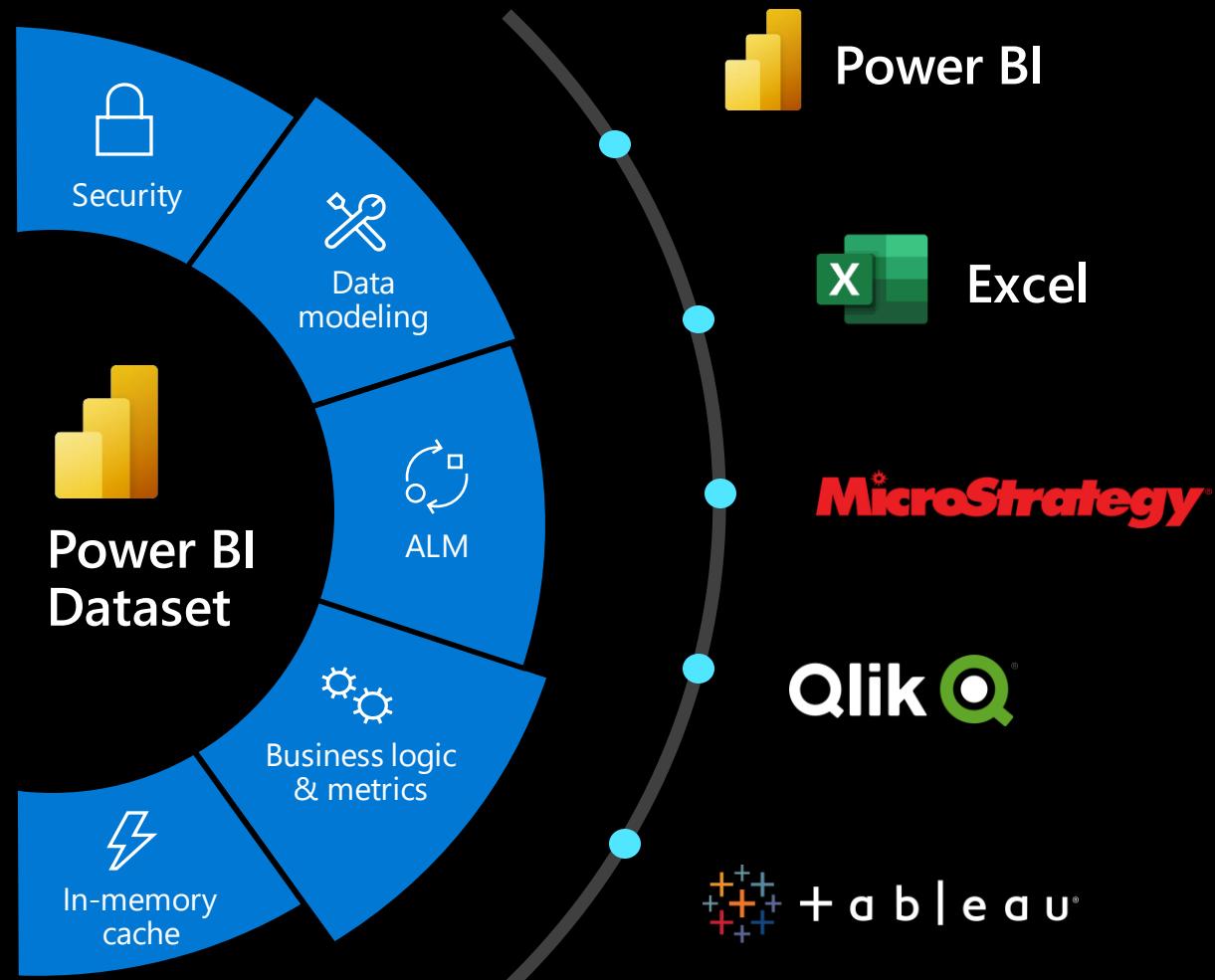


**Business
Intelligence**

Generally Available

Worlds leading
OLAP engine

Blazing fast
performance with
connectivity for a
variety of data
visualization
applications



Generally Available

Combine enterprise and local datasets

Seamless evolution
from self-service BI to
an enterprise semantic
model for company
wide adoption



Public Preview

Composite models

Seamlessly combine and extend self-service BI with corporate BI models

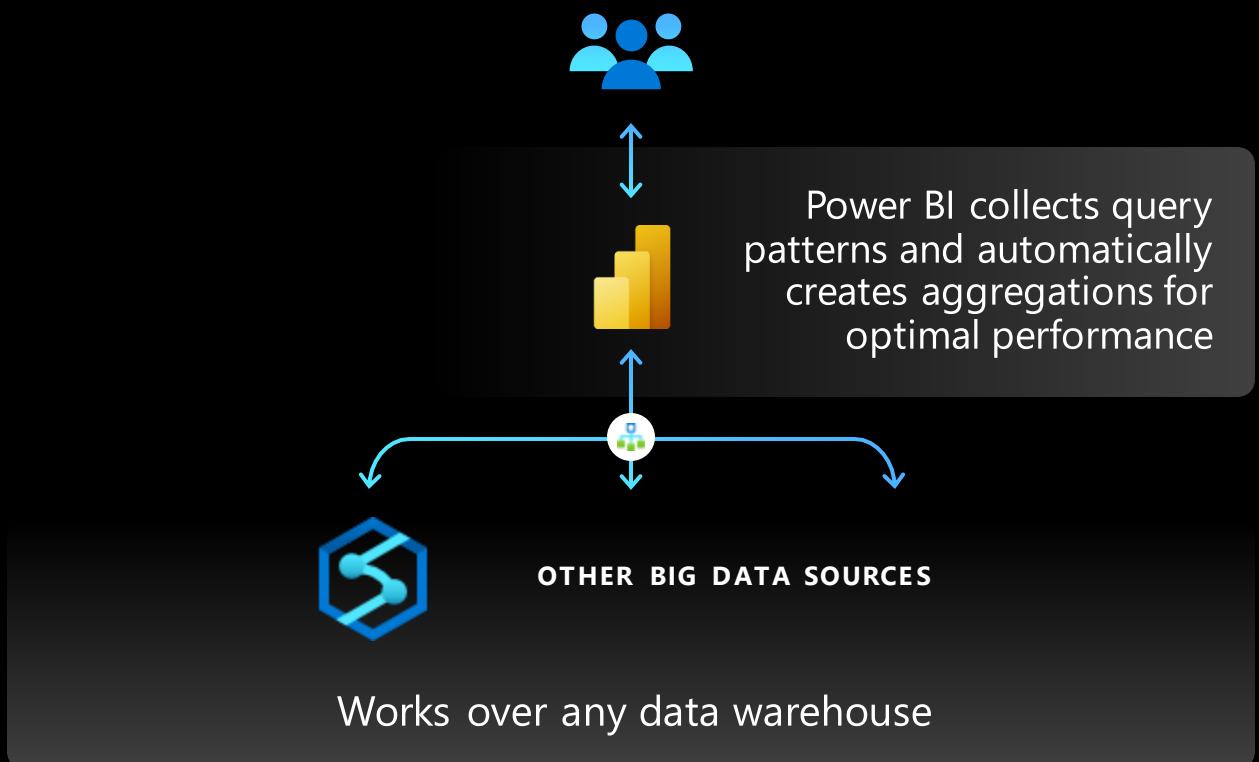
The screenshot shows the Power BI Desktop interface with a report titled "Sales Report". The report features three main summary measures at the top: "Recharge Amount" (\$8.6M), "Sales" (\$94.12M), and "Pipeline" (\$78.71M). Below these are two visualizations: a bar chart showing sales by date from April 24 to May 5, 2020, and a gauge chart showing the hour of sale (0 to 23) with a value of 0 highlighted. A table below the gauge chart lists daily sales figures. The Power BI desktop ribbon is visible at the top, and the "Filters" pane on the right shows a hierarchy of corporate BI models like Customer, Date, and Sales, alongside self-service BI fields like Geography and Lead Time.

Date	Sales
4/24/2020	\$7,092,263.04
4/25/2020	\$7,627,960.2
4/26/2020	\$7,832,140.77
4/27/2020	\$7,417,329.26
4/28/2020	\$6,768,925.96
4/29/2020	\$6,478,403.22
4/30/2020	\$6,762,418.12
5/1/2020	\$7,211,322.3
5/2/2020	\$7,593,083.56
5/3/2020	\$7,883,966.76
5/4/2020	\$7,480,390.01
5/5/2020	\$6,983,100.07

Public Preview

Automatic Aggregations

Automatically learns about customer usage patterns and create aggregates to optimize performance and reduce cost



Public Preview

Automatic aggregations

AI driven self-optimizing performance improvement

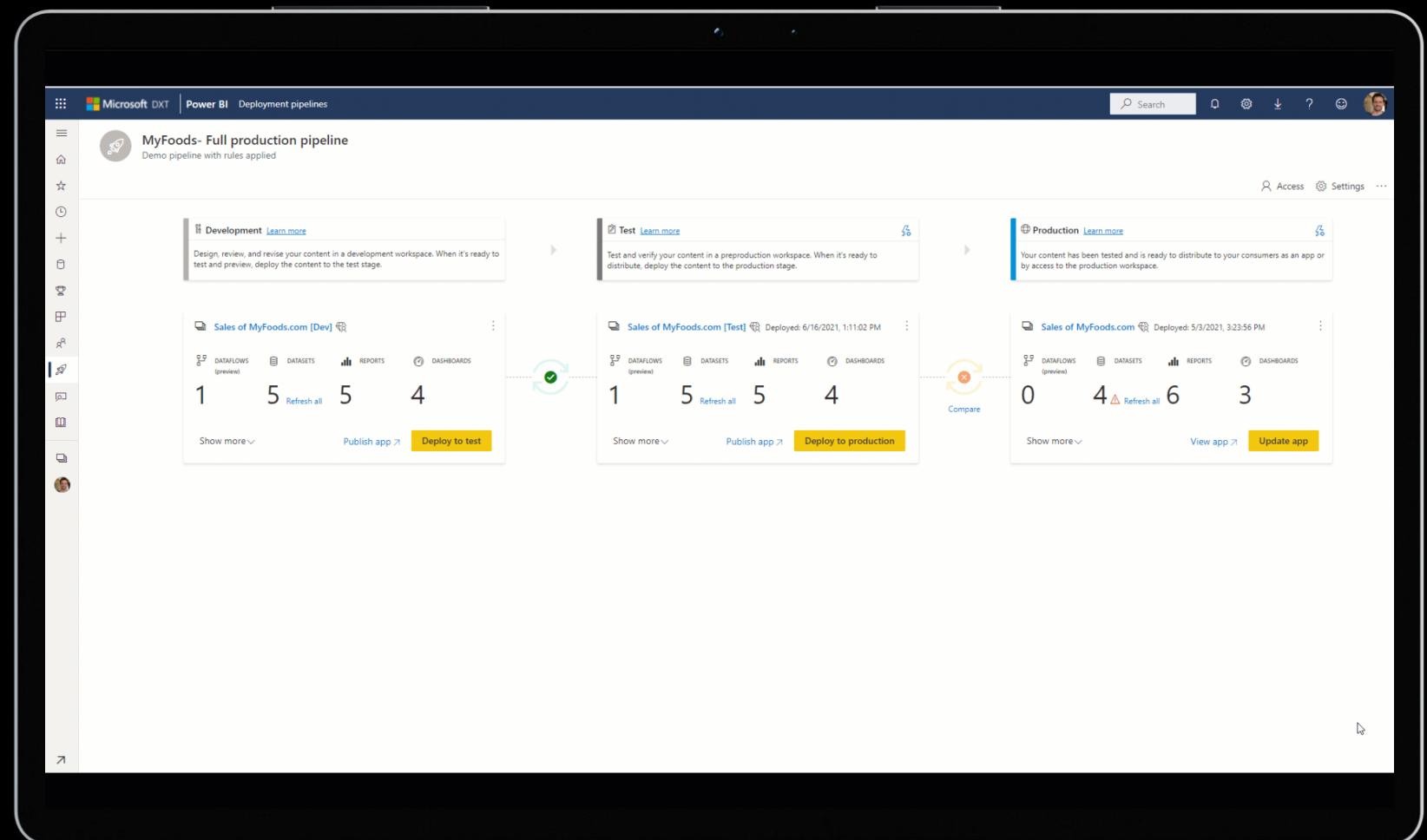
The screenshot shows the Power BI dataset management interface for the "NY Taxi Analysis" workspace. The left sidebar contains navigation icons for Home, Recent, Favorites, Pipelines, Datasets, Dataflows, Reports, Dashboards, and Help. The main area displays a list of datasets:

Name	Type	Owner	Refreshed	Next refresh	Endorsement
MS Finance1	Dataset	NY Taxi Analysis	4/11/21, 8:52:51 PM	N/A	—
MS Finance2	Dataset	NY Taxi Analysis	4/27/21, 5:21:16 AM	⚠️ N/A	—
Taxi1	Dataset	NY Taxi Analysis	4/12/21, 8:01:06 AM	6/30/21, 8:00:00 AM	—
Taxi2	Dataset	NY Taxi Analysis	4/27/21, 5:21:16 AM	⚠️ N/A	—

Generally Available

Controlled change management

Power BI deployment pipelines enable efficient and reusable release processes



Generally Available

Premium Gen2
Unmatched large-scale
analytics with simple low-
overhead administration

Over 50%

Premium Customer Nodes
running on Gen 2 two weeks
after release

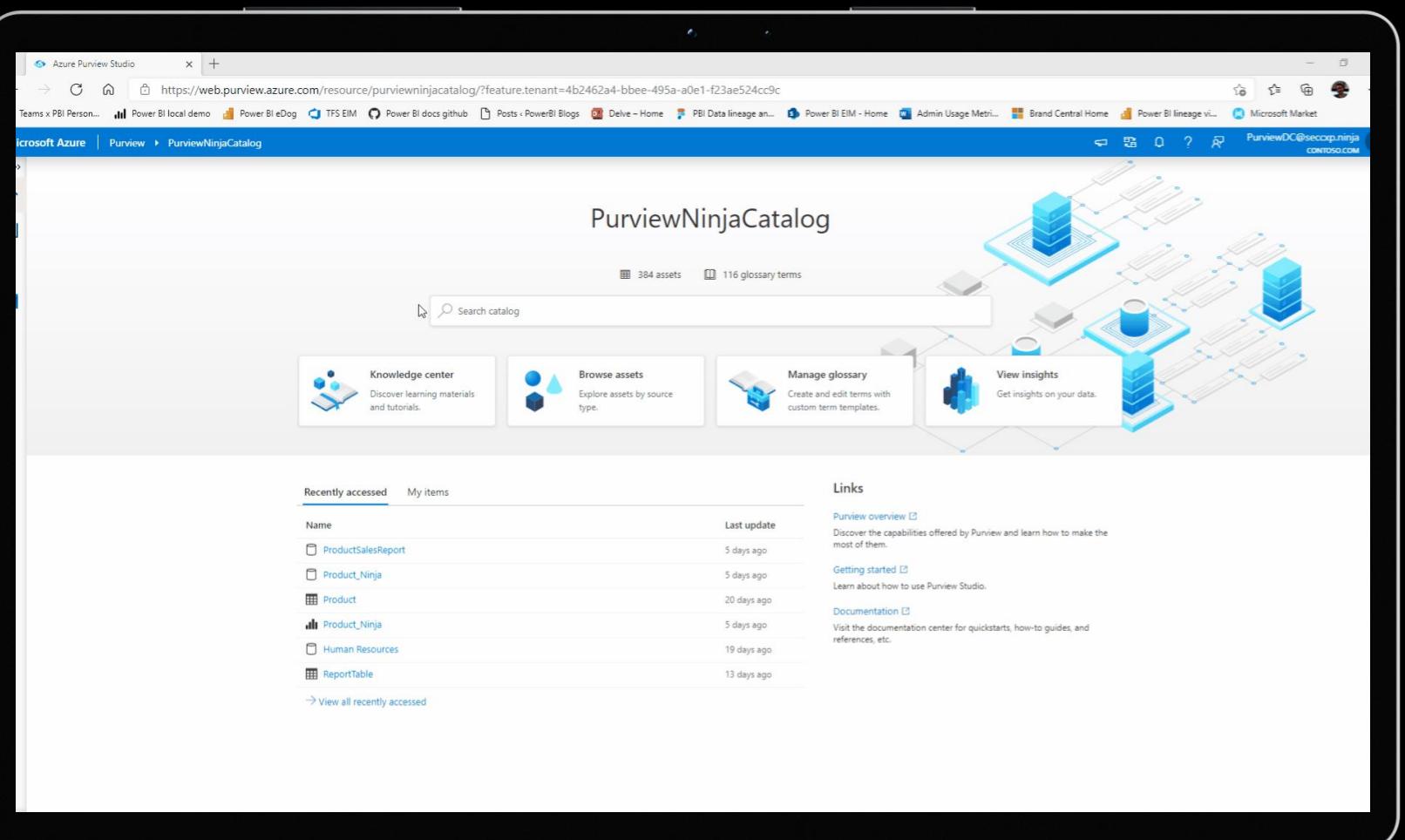


The image is a composite of several elements. On the left, there's a screenshot of the Microsoft Power BI service interface showing a workspace named "2 - Demo Gen2". The interface displays a chart titled "Premium capacity" with a bar chart showing "Core Count" over time (Feb 02 to Feb 03) and a line graph showing "Daily peak usage". Below this is a list of "Workspace ID" entries. On the right, there's a tweet from André Kamman (@AndreKamman) dated Oct 14. The tweet text reads: "Went from carefully scheduling refreshes and still getting memory errors regularly to REFRESH ALL THE THINGS! Power BI Premium users, move to Gen2, do it now! My biggest model went from a 70 minute processing time to 12!". Below the tweet is a small video thumbnail showing two people looking at a large circular screen displaying a complex, light-based visualization. At the bottom of the image, there are social media interaction metrics: 2 comments, 2 retweets, 27 likes, and a share icon.

Generally Available

Power BI + Azure Purview

Enhanced governance and
cataloging capabilities
integrated with Power BI

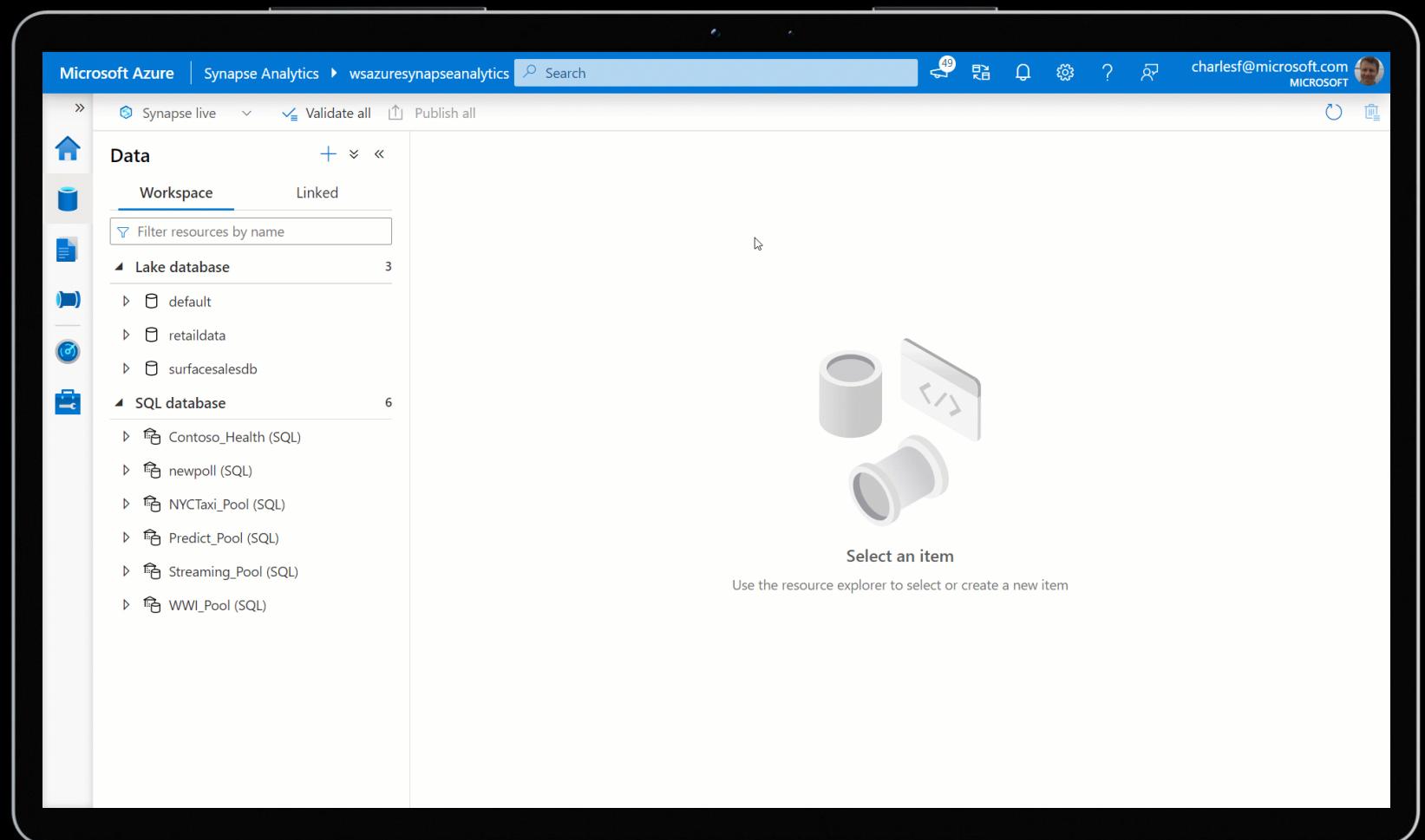


Governance

Generally Available

Integrated Catalog search in Synapse

Enables developers to search for data assets across the entire data estate to analyze with Synapse





Synapse



Power BI

Data
Integration

Data
Engineering

Data
Warehouse

Data
Science

Observation
Analytics

Business
Intelligence


Governance

