Microsoft

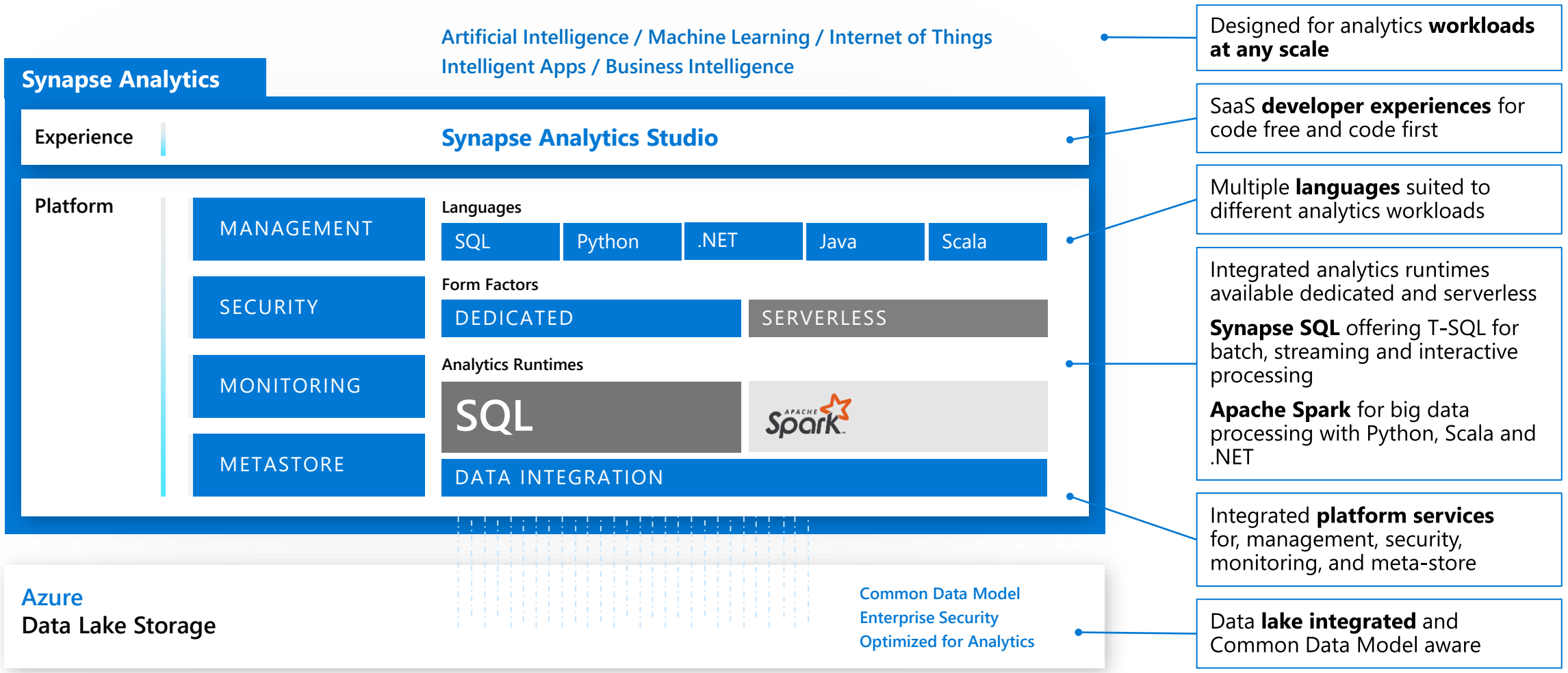
Azure Synapse Analytics

Gary Lee

Cloud Solution Architect (Data / AI / Blockchain)

Azure Synapse Analytics

Limitless analytics service with unmatched time to insight



Sections

- Studio
- Integrate
- Synapse SQL
  - Dedicated SQL Pool
  - Serverless SQL Pool
- Spark Analytics
- Security
- Monitor
- Management
- Metastore
- Linked services
- Synapse Link

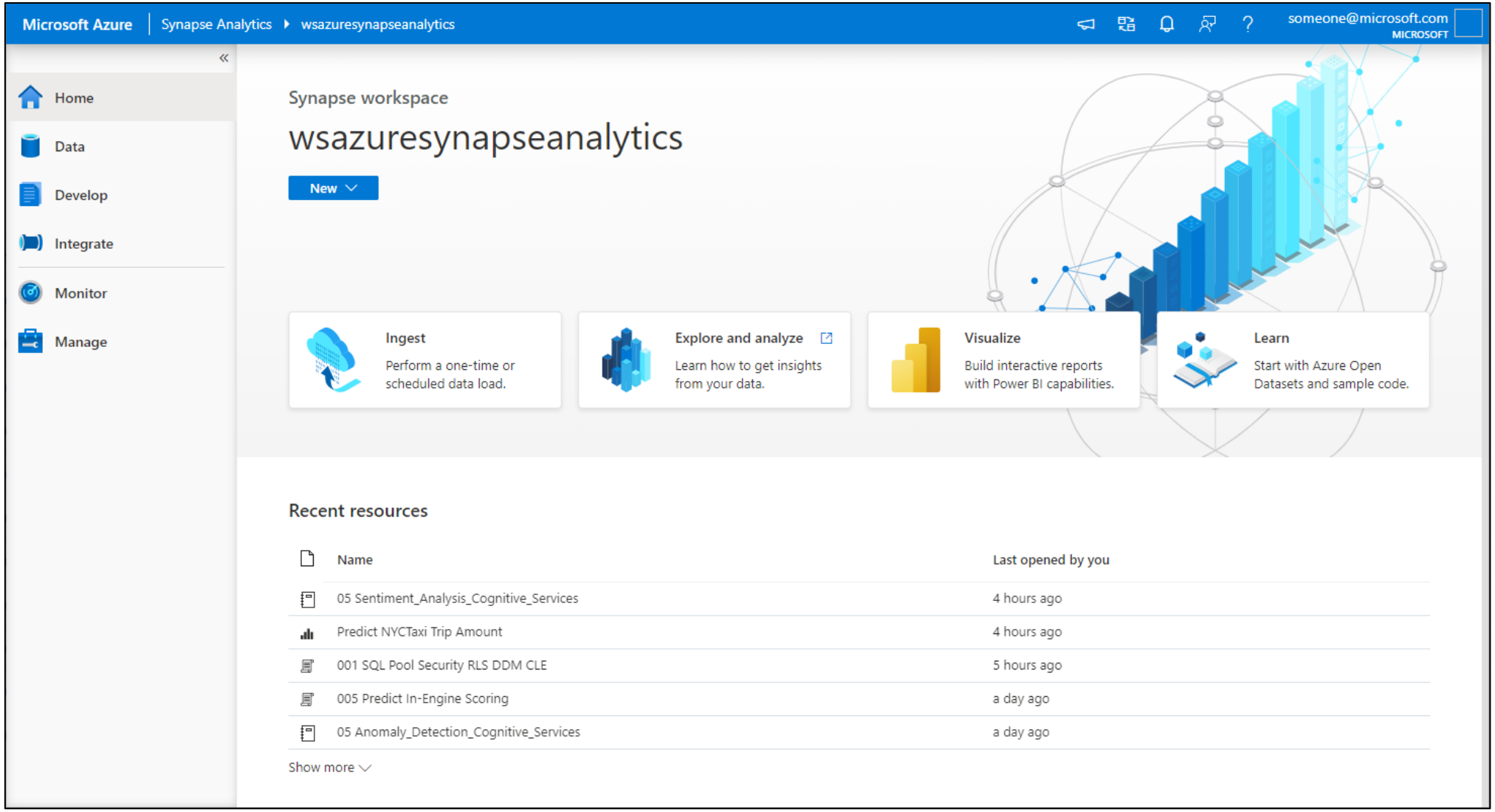
Microsoft

Azure Synapse Analytics

Studio

Studio

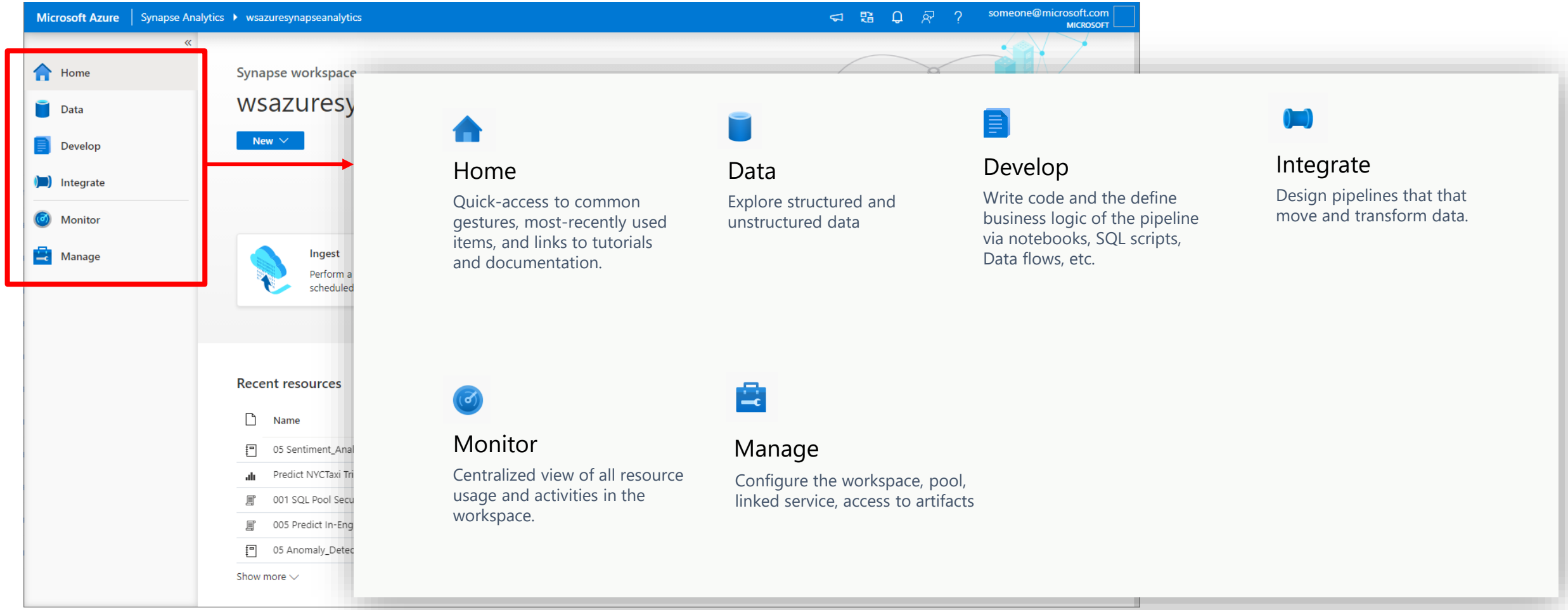
A single place for Data Engineers, Data Scientists, and IT Pros to collaborate on enterprise analytics



Synapse Studio

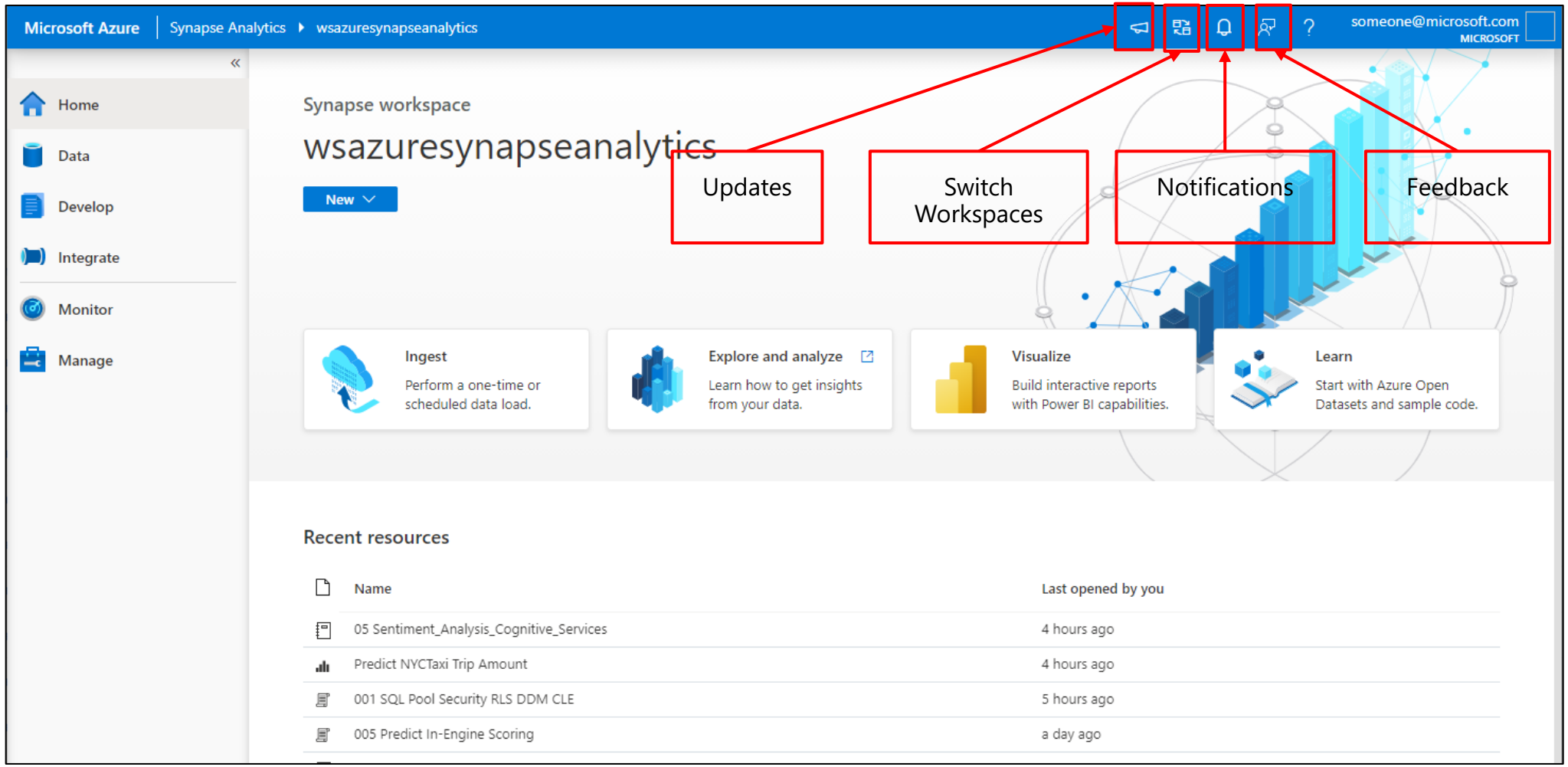
Synapse Studio divided into **Activity hubs**.

These organize the tasks needed for building analytics solution.



Home Hub

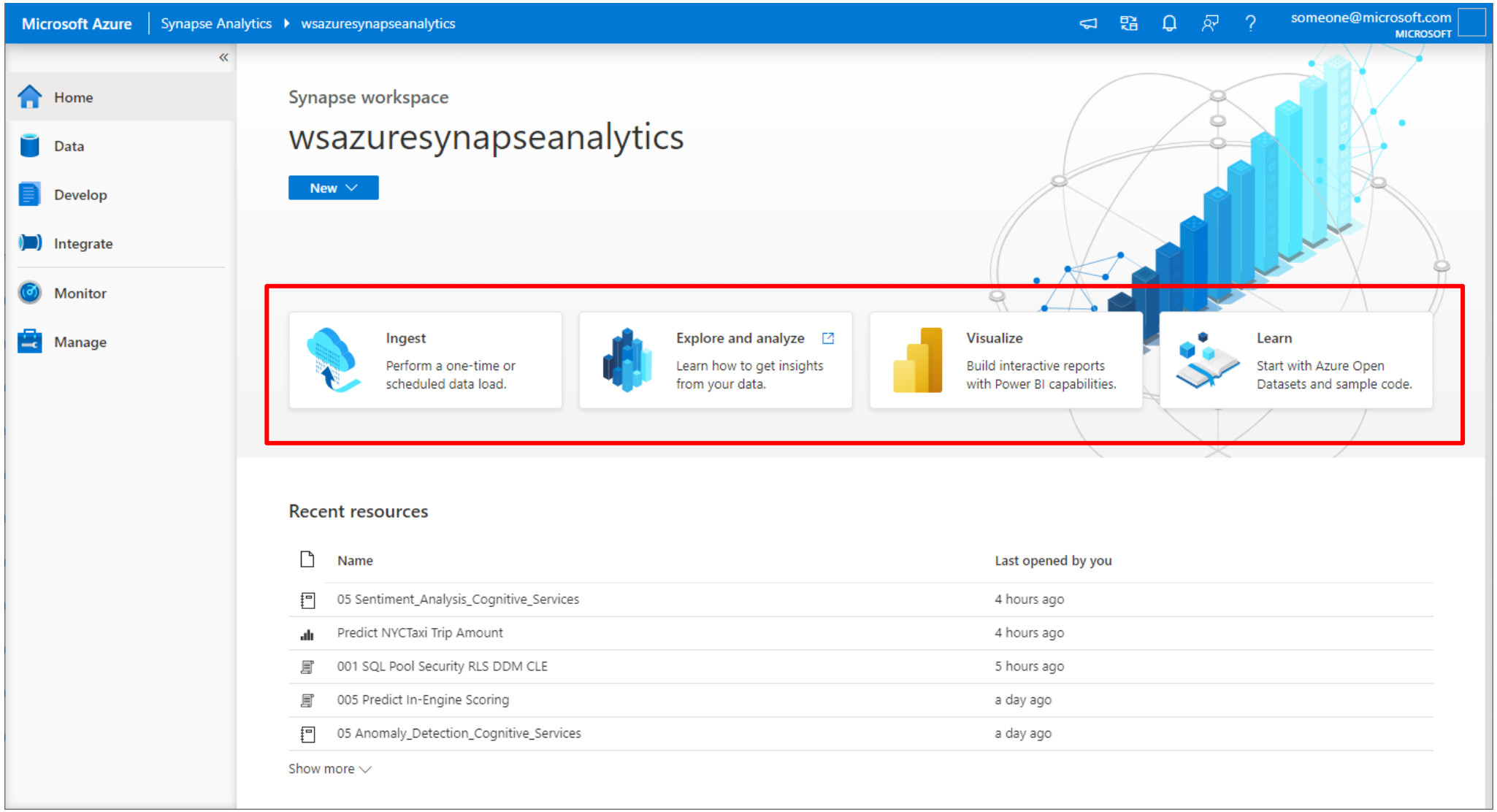
Ease of access to get updates, to switch workspace, to get notifications and to provide feedback



Synapse Studio
Home hub

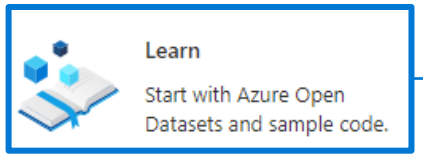
Home Hub

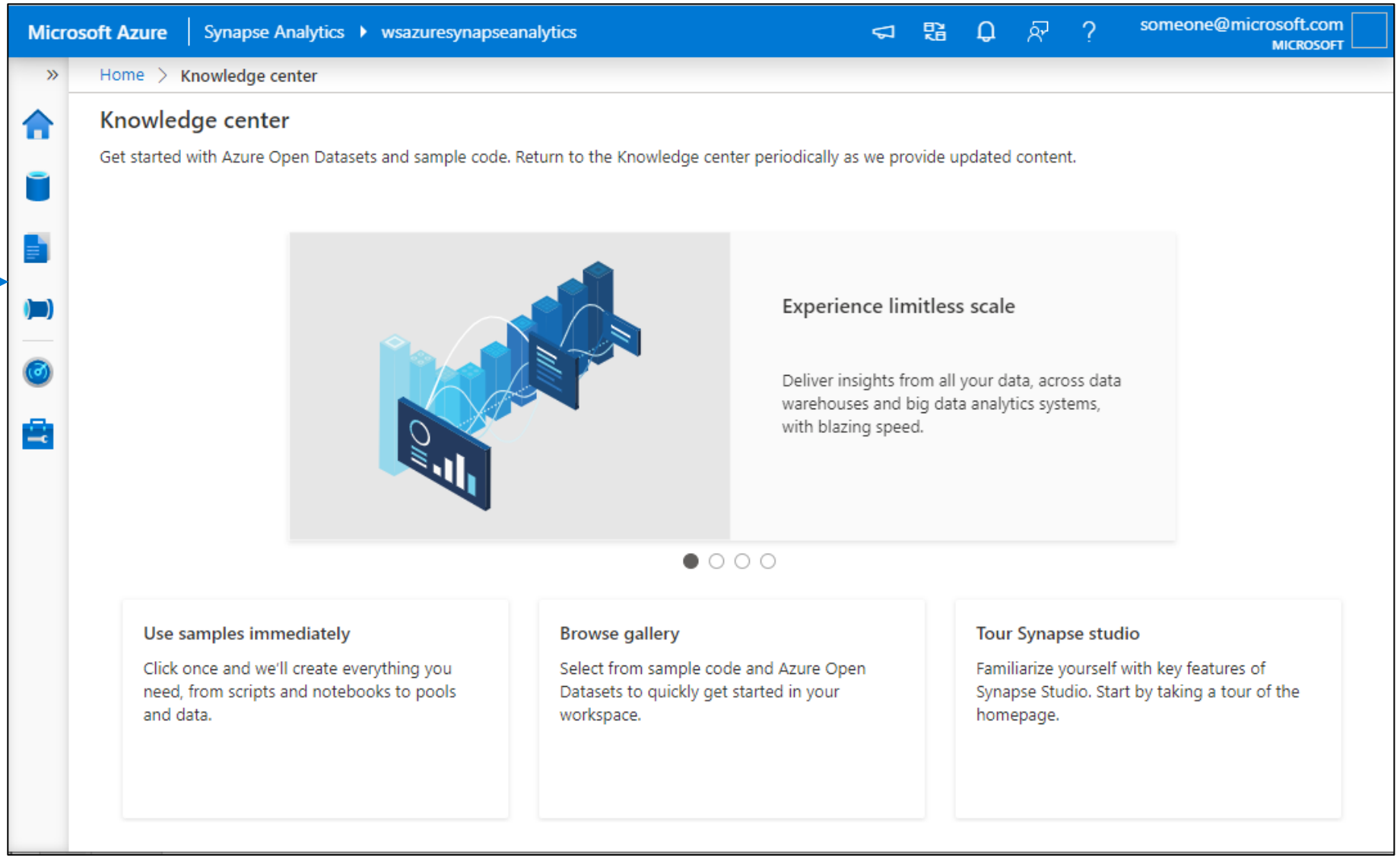
It is a starting point for the activities with key links to tasks, artifacts, documentation and sample artifacts for learning purpose



Home Hub - Learn

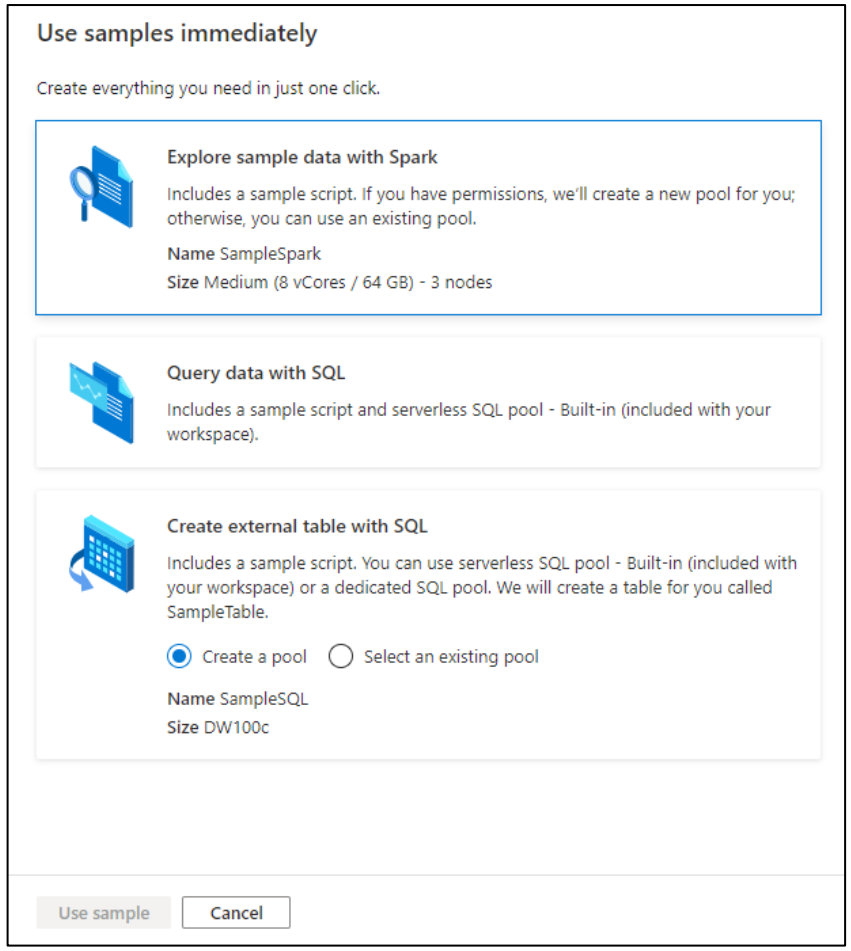
Knowledge center offers one stop for learning with samples, artifacts and Synapse studio tour

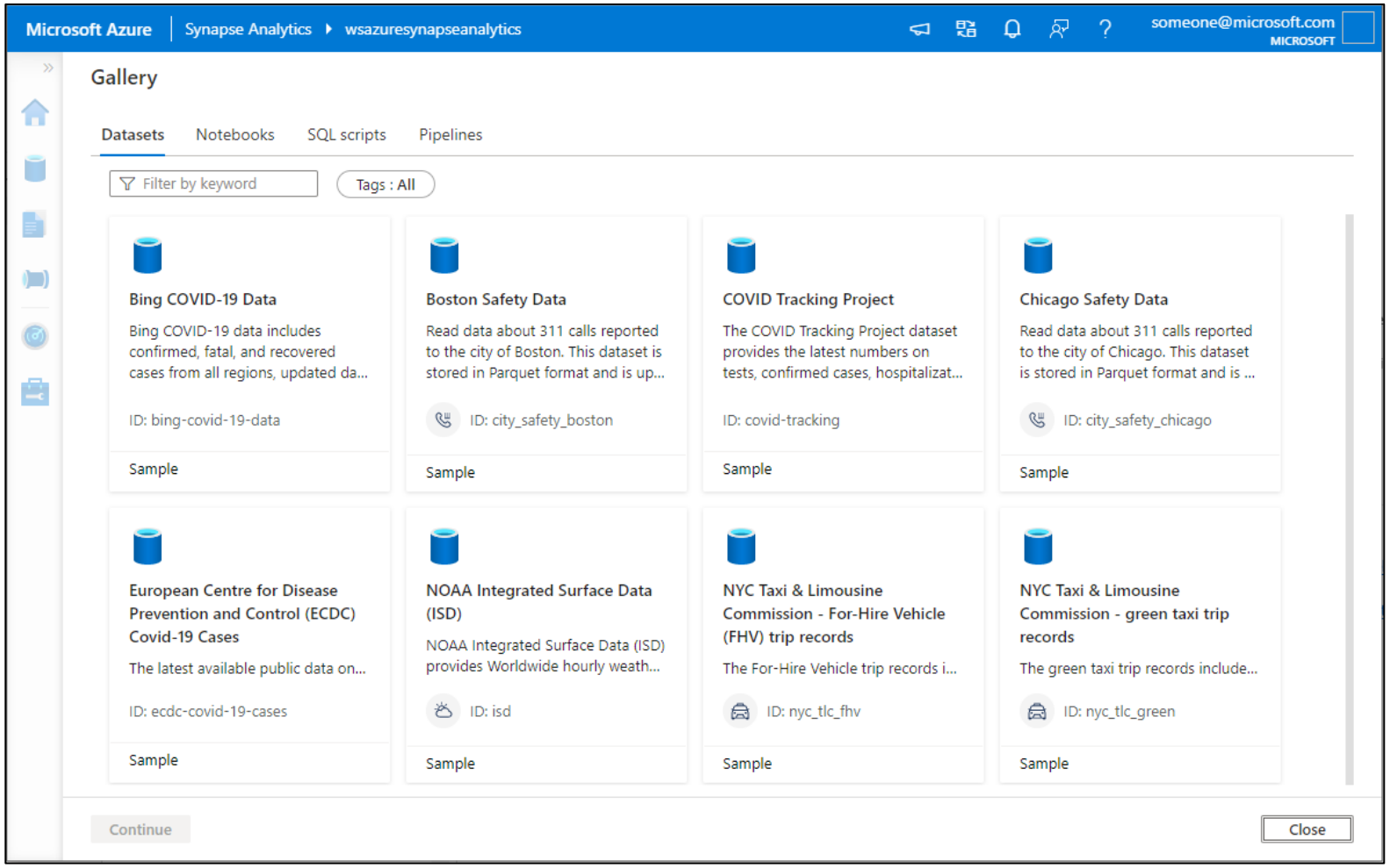




Knowledge center

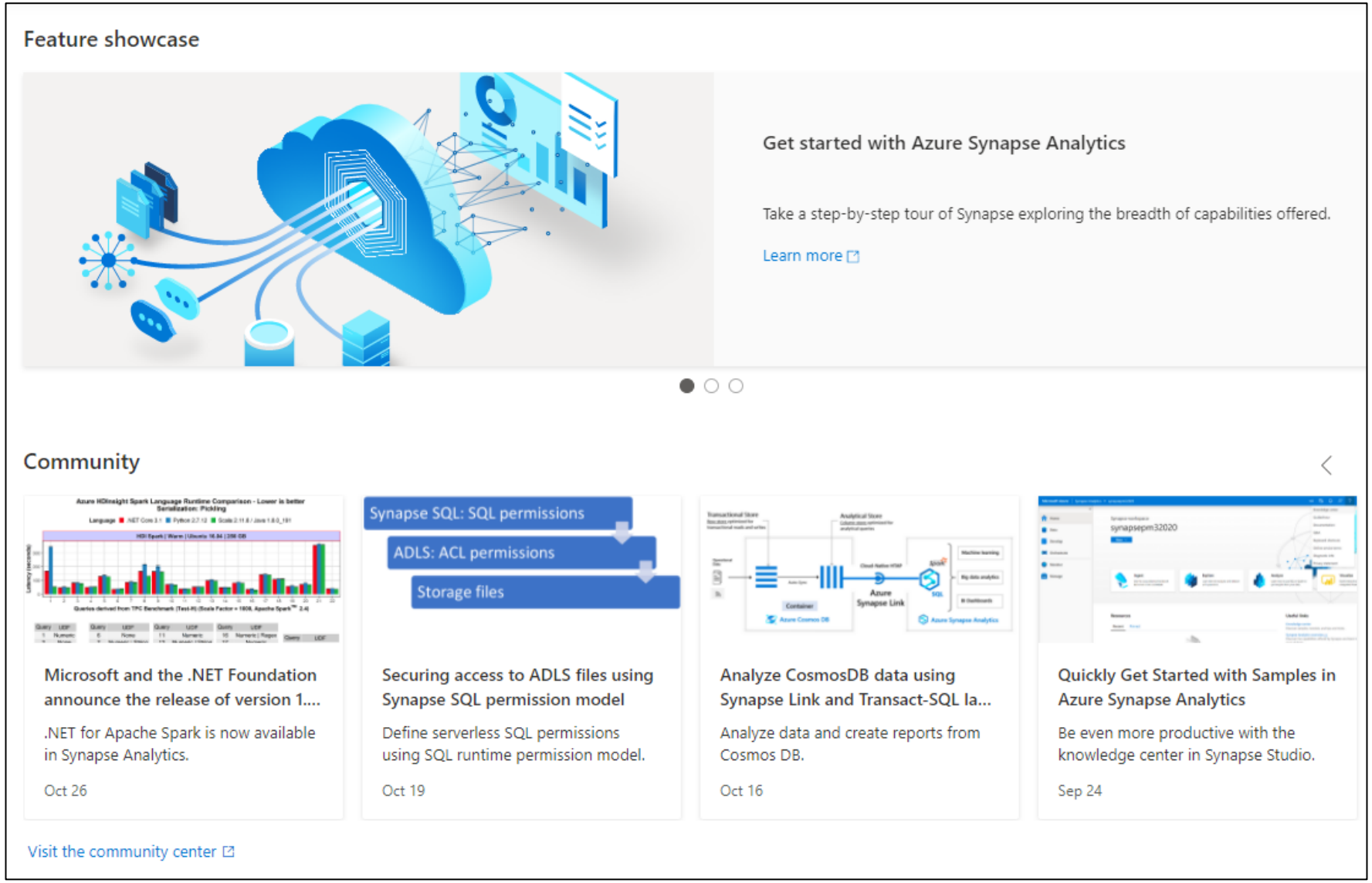
Knowledge center offers open datasets, sample notebooks, SQL scripts and pipeline templates for easy start and learning





Home Hub

It provides links to feature announcements, documents and community center for reference

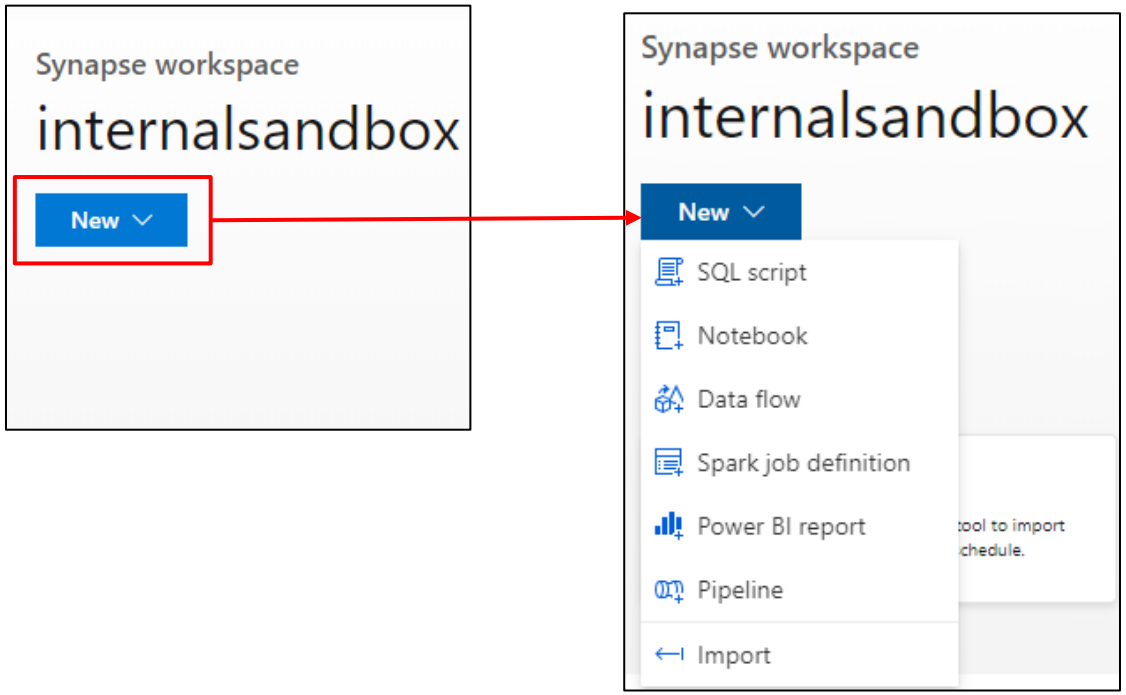


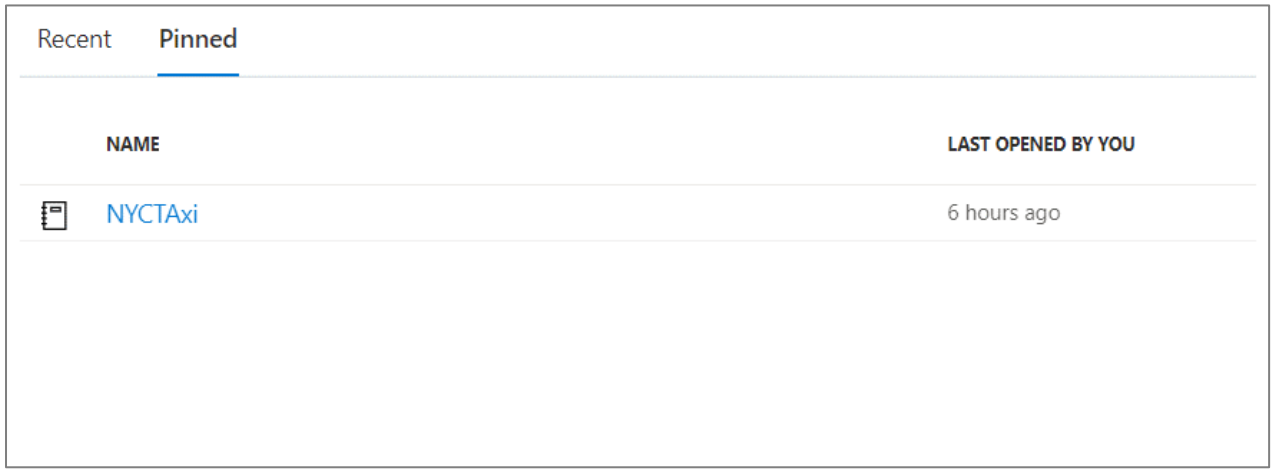
Home Hub

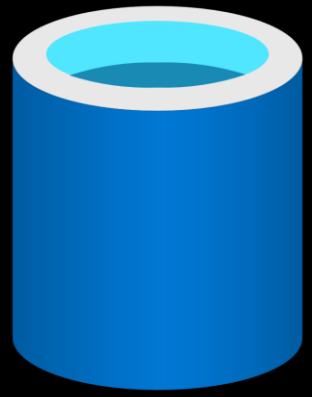
Overview

New dropdown – offers quickly start work item

Recent & Pinned – Lists recently opened code artifacts. Pin selected ones for quick access



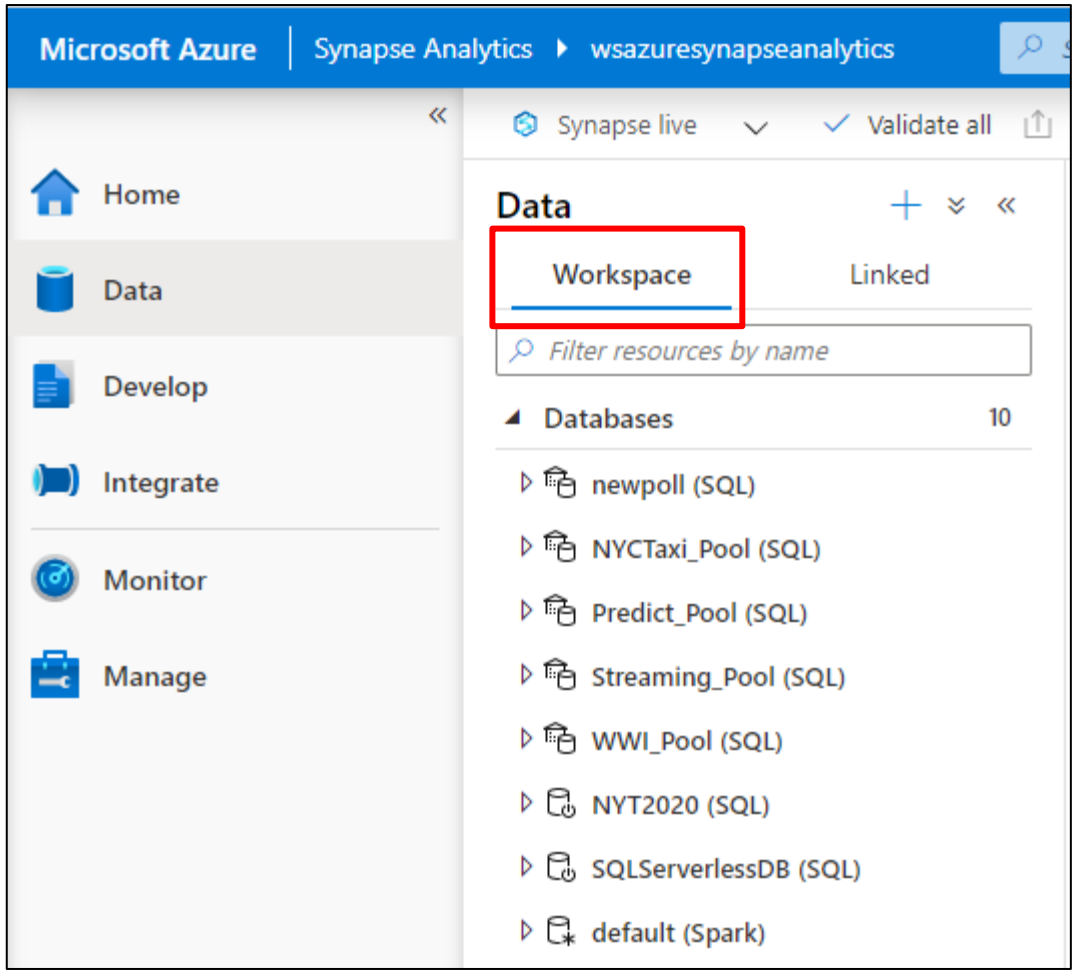


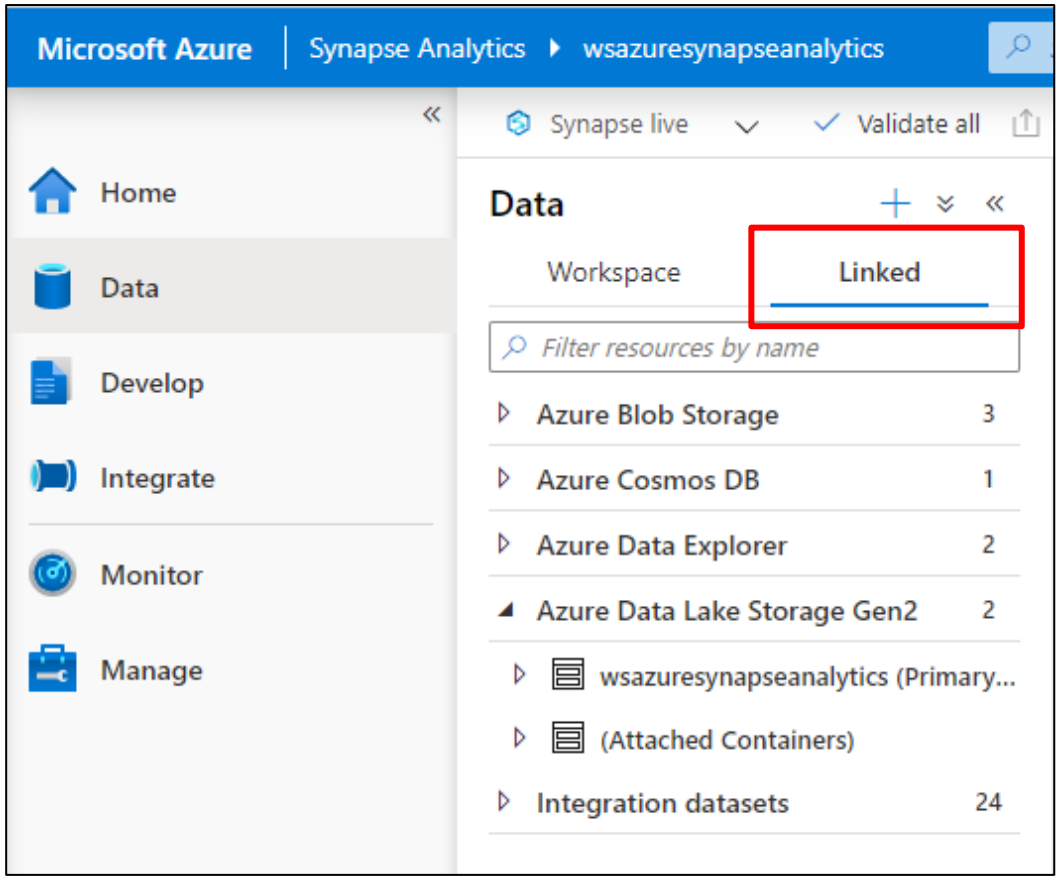


Synapse Studio
Data hub

Data Hub

Explore data inside the workspace and in linked storage accounts





Data Hub – Linked Storage

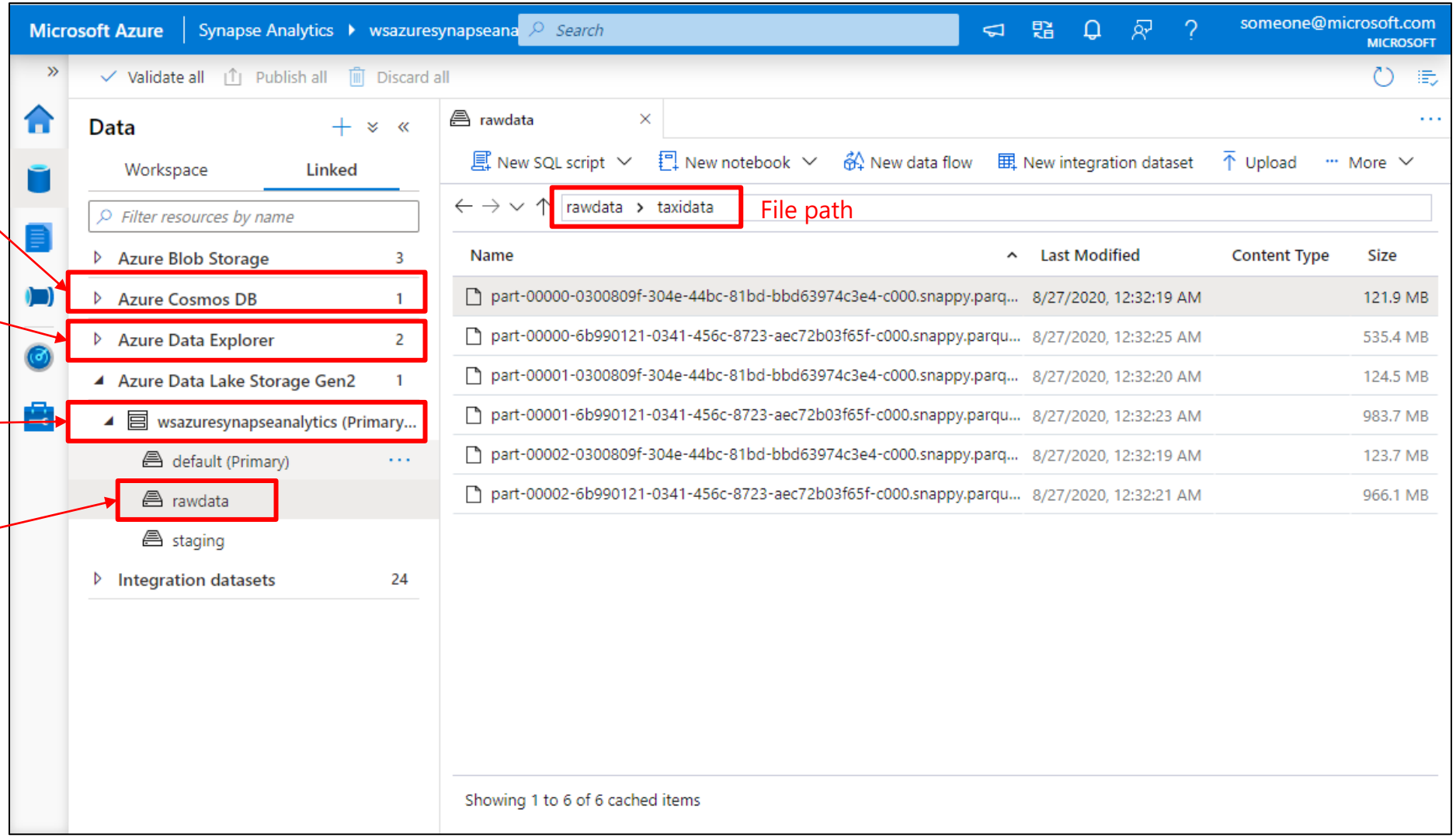
Browse Azure Data Lake Storage Gen2 accounts – filesystems, Azure Data Explorer – clusters, Azure Cosmos DB -containers

Linked Cosmos DB Analytical Store

Linked Azure Data Explorer

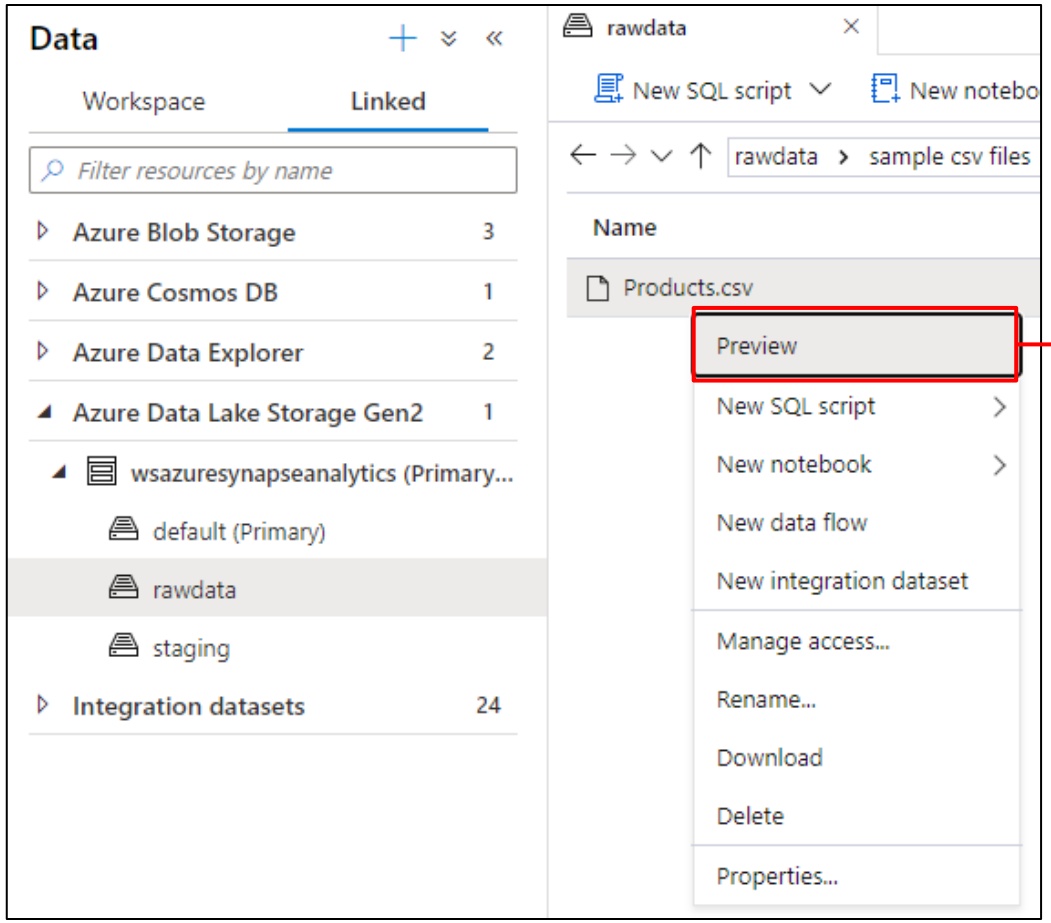
Linked ADLS Gen2 Account

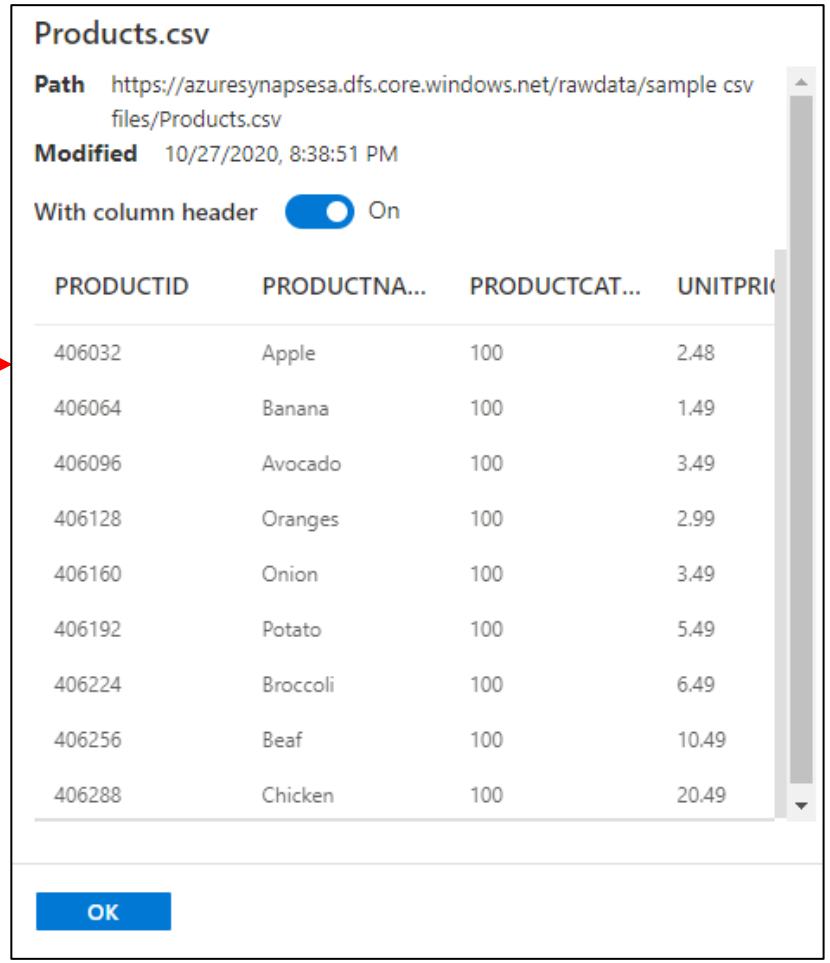
Container (filesystem)



Data Hub – Storage accounts

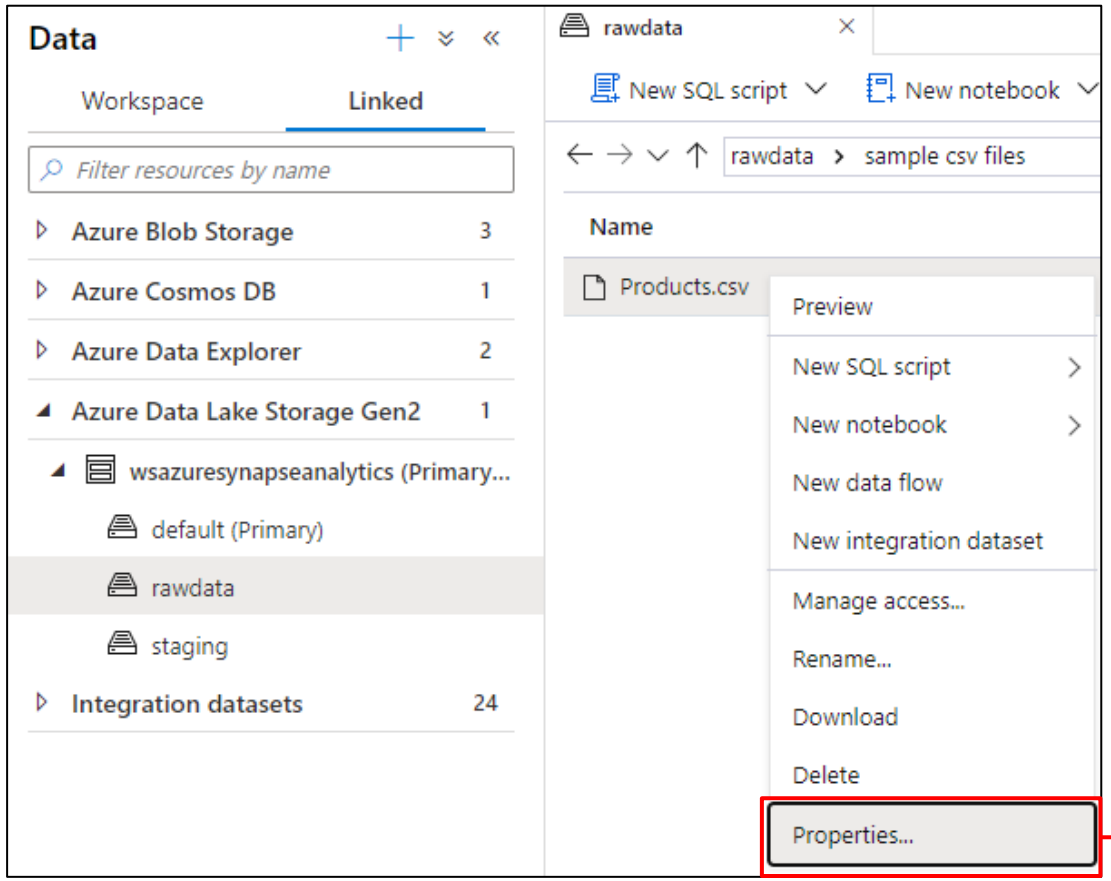
Preview a sample of your data

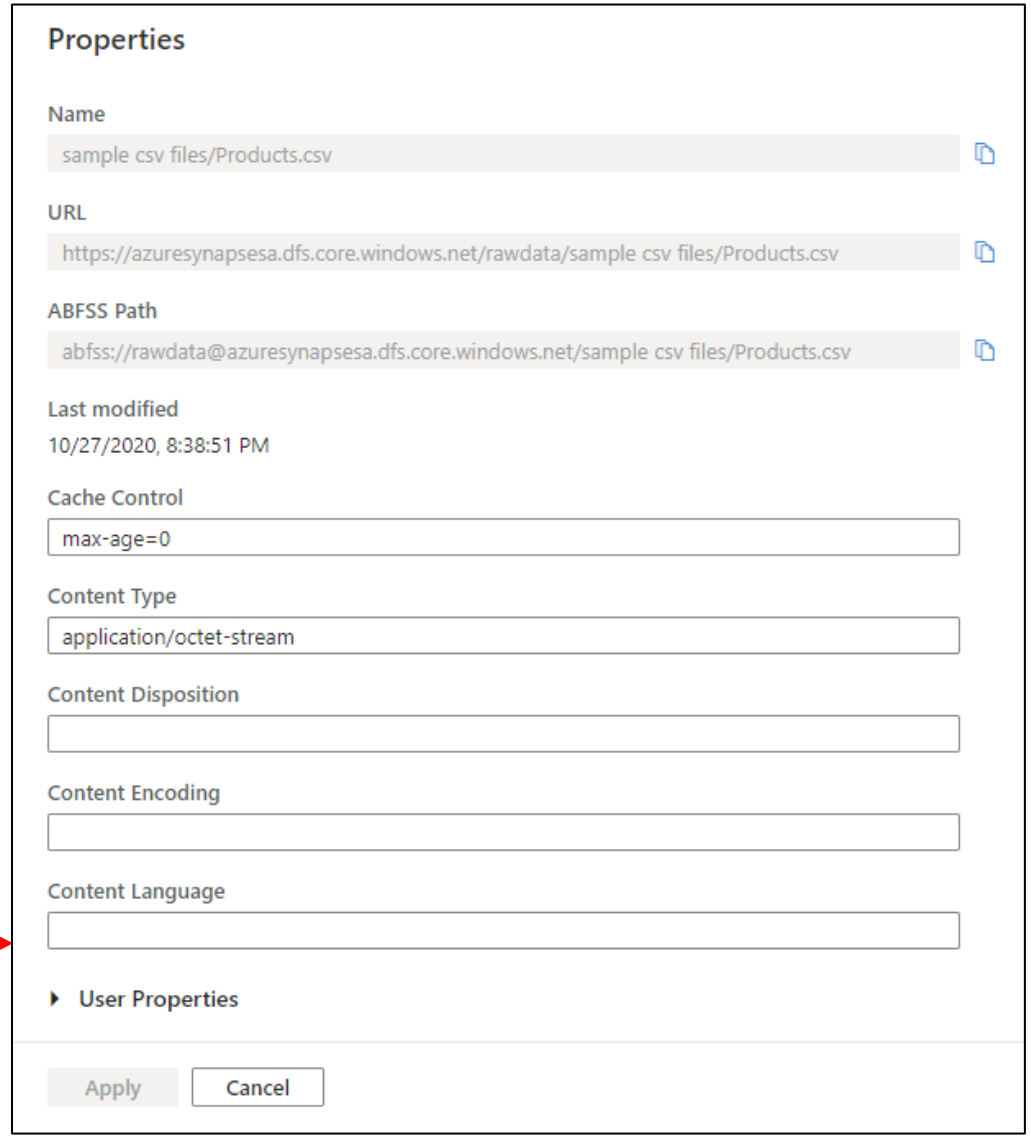




Data Hub – Storage accounts

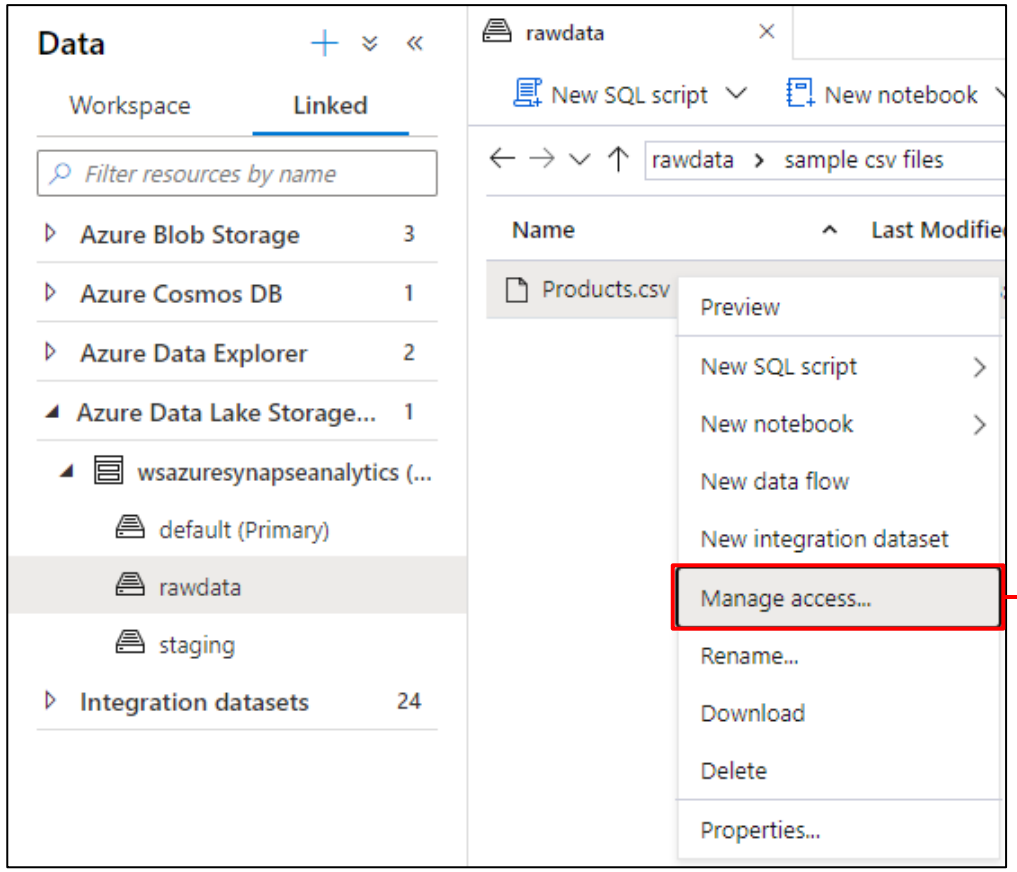
See basic file properties

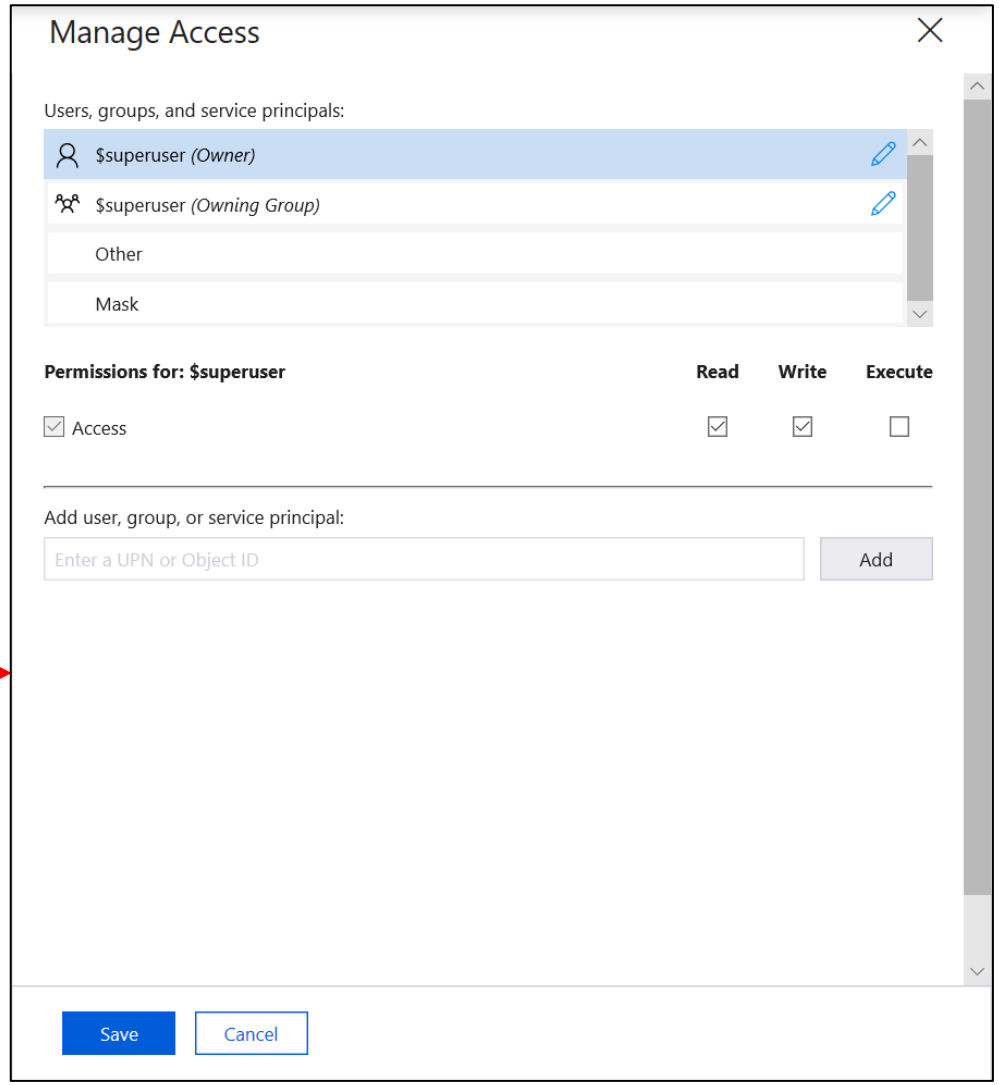




Data Hub – Storage accounts

Manage Access - Configure standard POSIX ACLs on files and folders

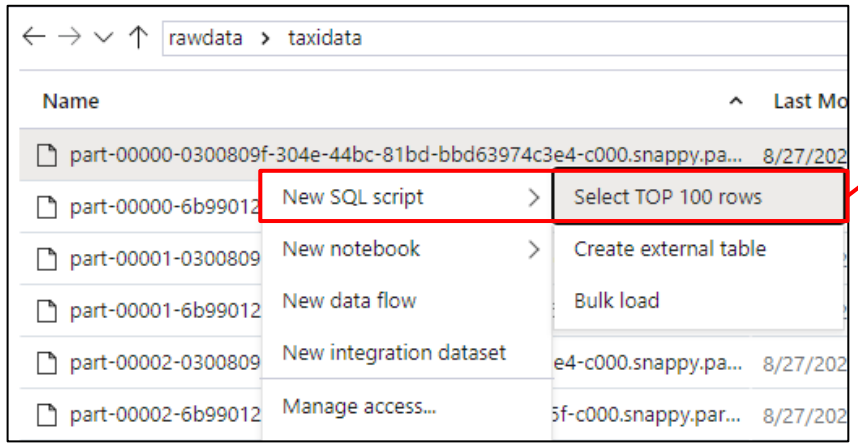


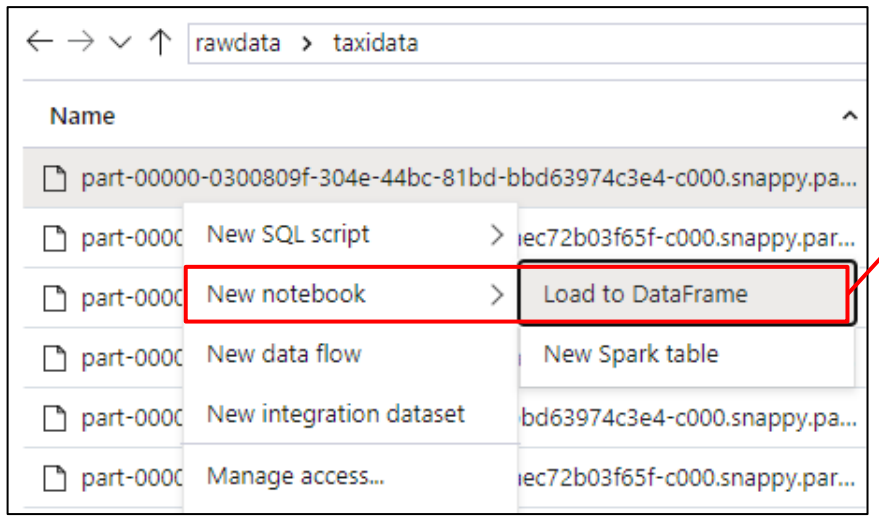


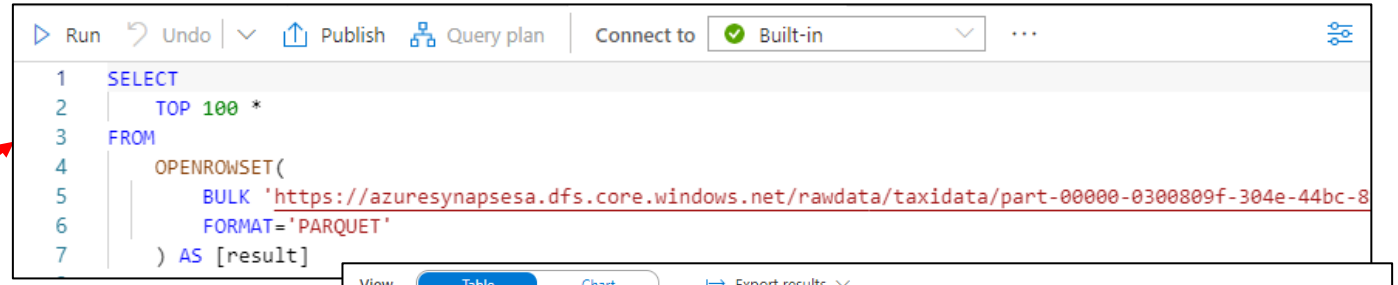
Data Hub – Storage accounts

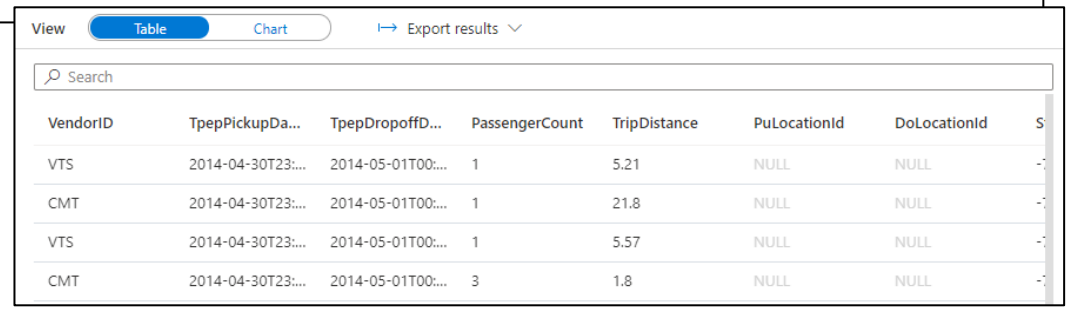
Two simple gestures to start analyzing with SQL scripts or with notebooks.

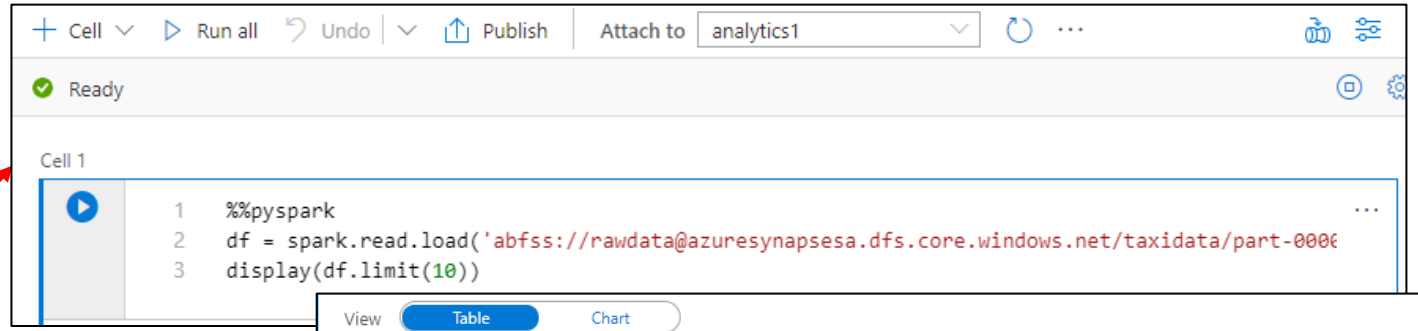
T-SQL or PySpark auto-generated.

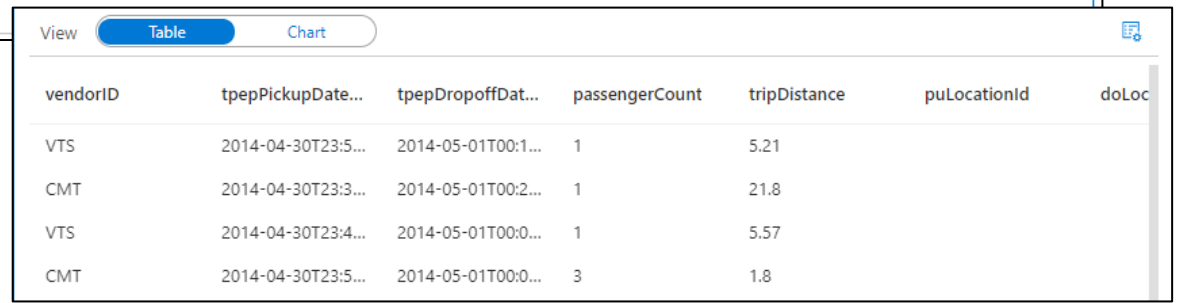










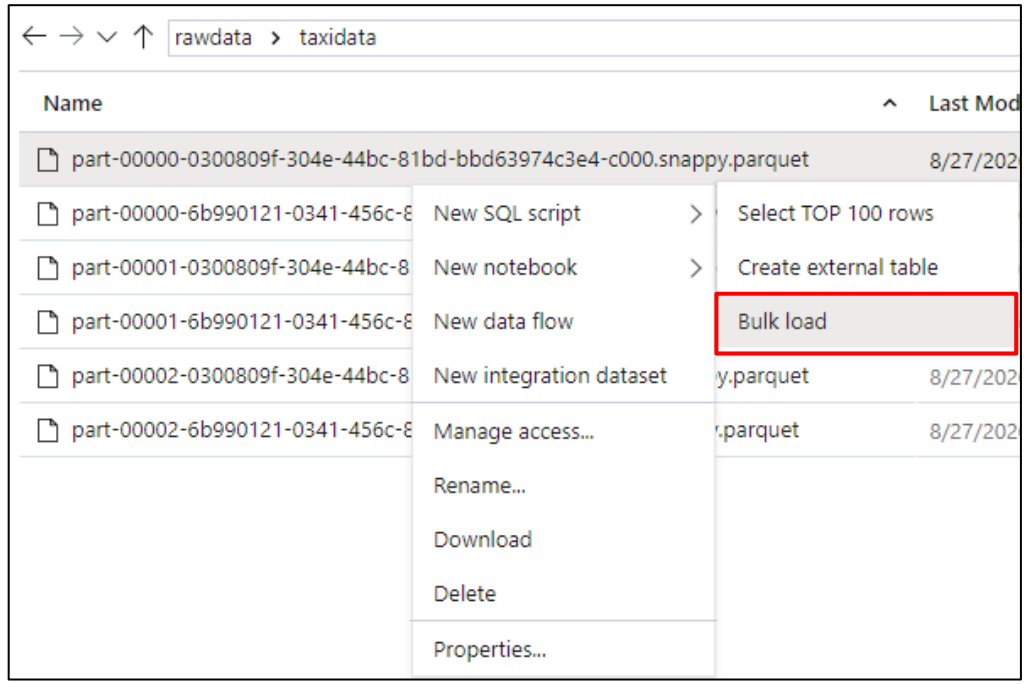


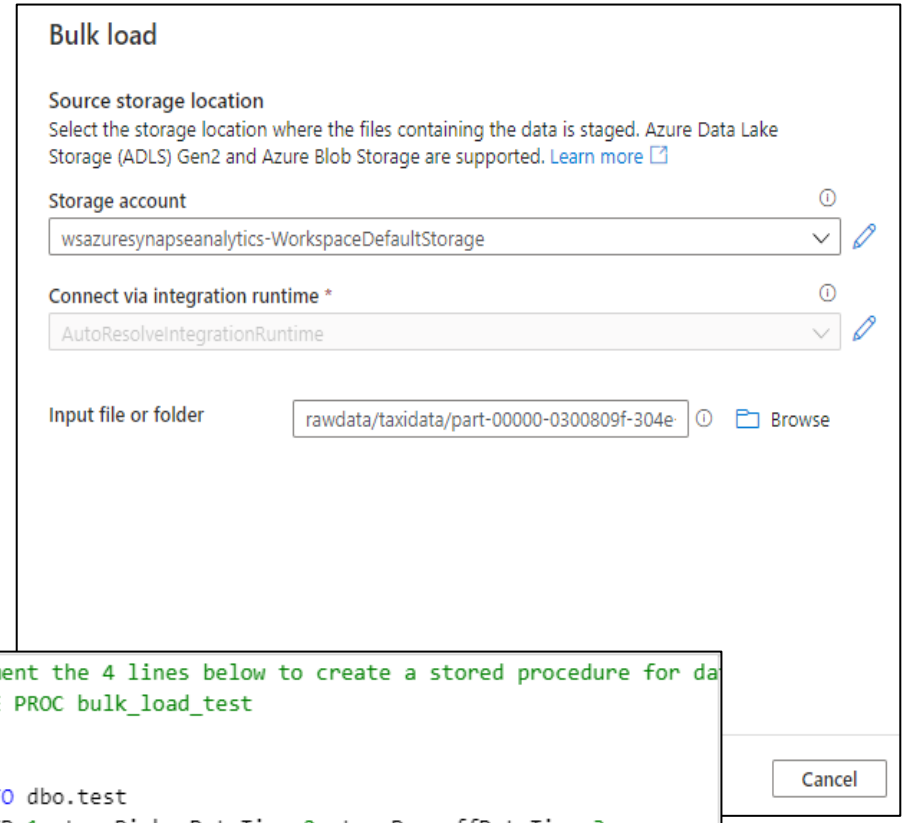
Data Hub – Bulk load wizard

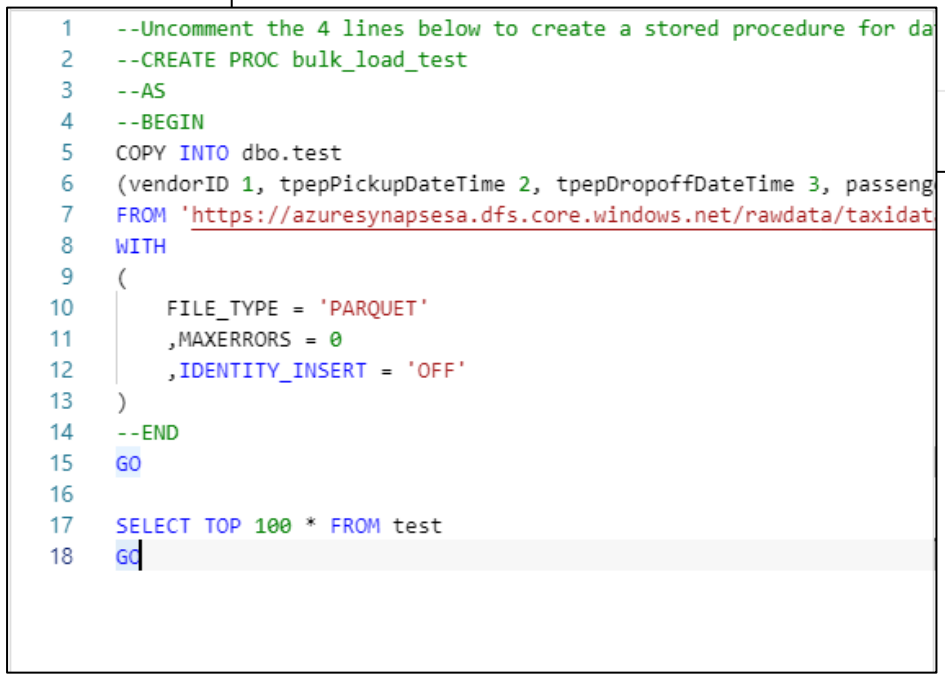
Bulk load data from the storage container or storage files with Bulk load wizard.

Auto create T-SQL scripts to load data

Offers creation of stored procedure to enable integration with pipeline





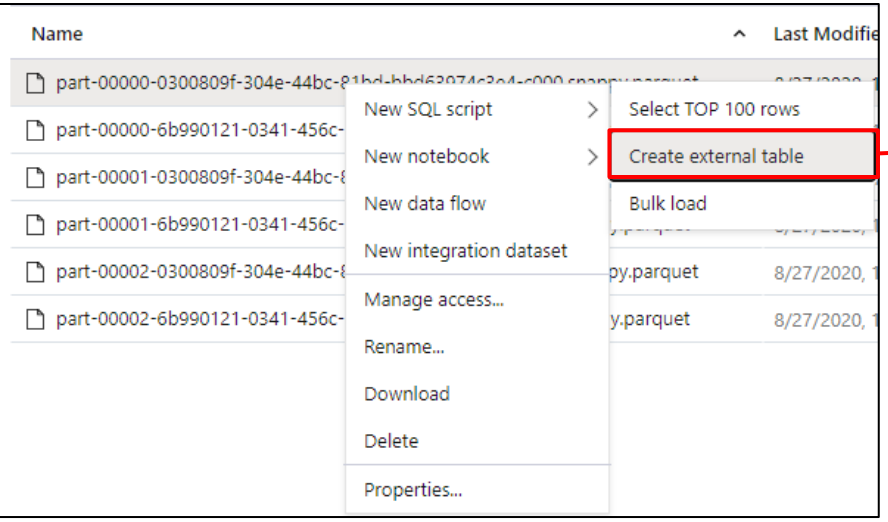


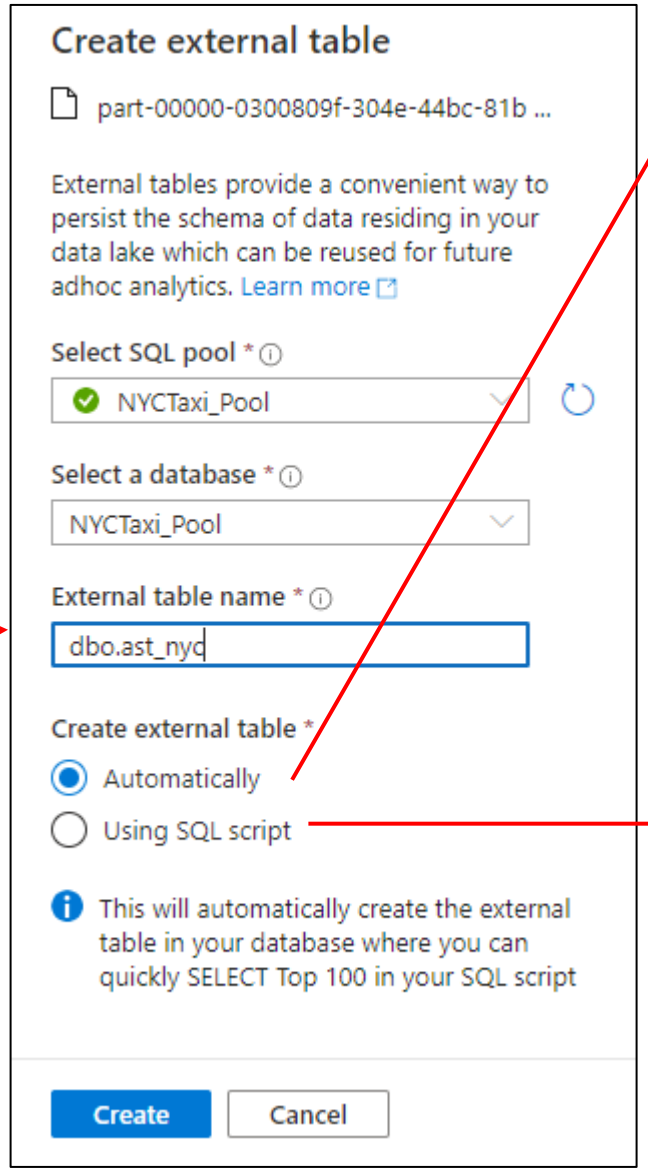
```
--Uncomment the 4 lines below to create a stored procedure for da
--CREATE PROC bulk_load_test
--AS
--BEGIN
COPY INTO dbo.test
(vendorID 1, tpepPickupDateTime 2, tpepDropoffDateTime 3, passeng
FROM 'https://azuresynapsesa.dfs.core.windows.net/rawdata/taxidat
WITH
(
    FILE_TYPE = 'PARQUET'
    ,MAXERRORS = 0
    ,IDENTITY_INSERT = 'OFF'
)
--END
GO

SELECT TOP 100 * FROM test
GO
```

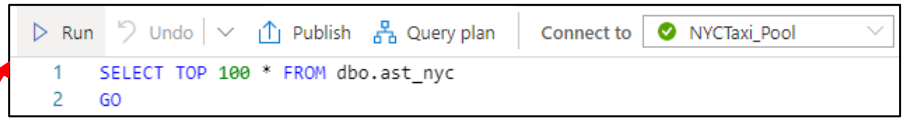
Data Hub – Storage accounts

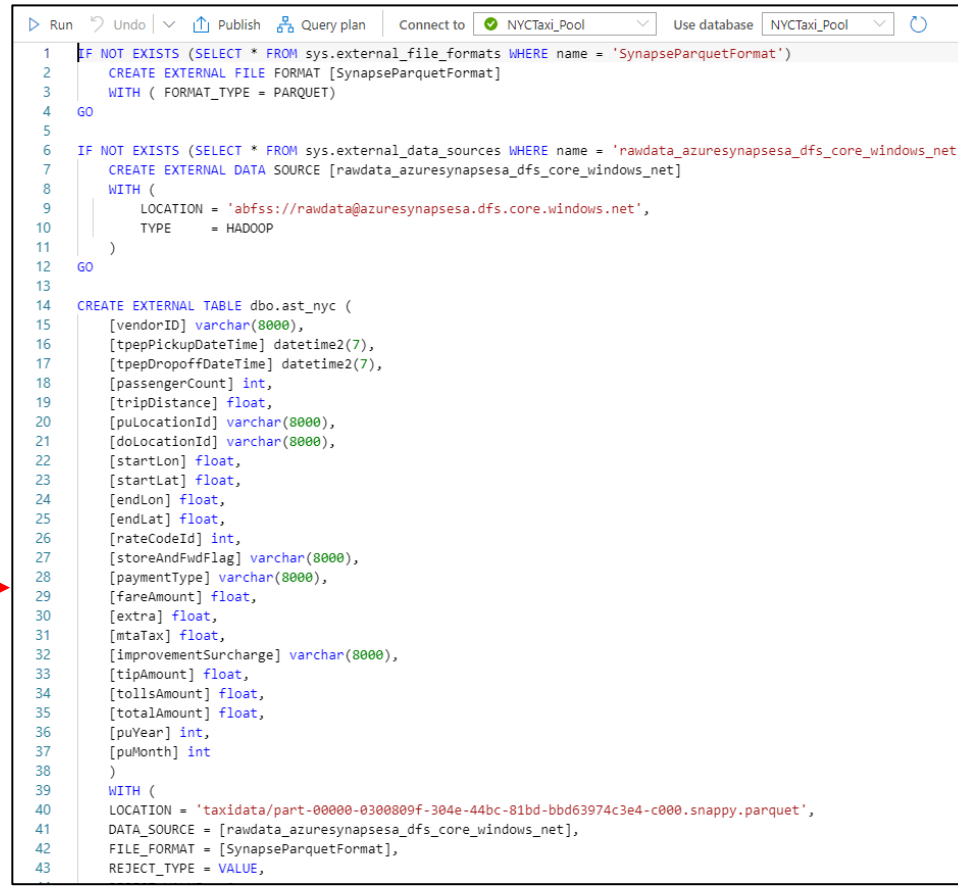
Increased productivity by automatic creation of external file format, external data source and external table.

It also offers generating scripts to create external file format, external data source and external table for manual execution.



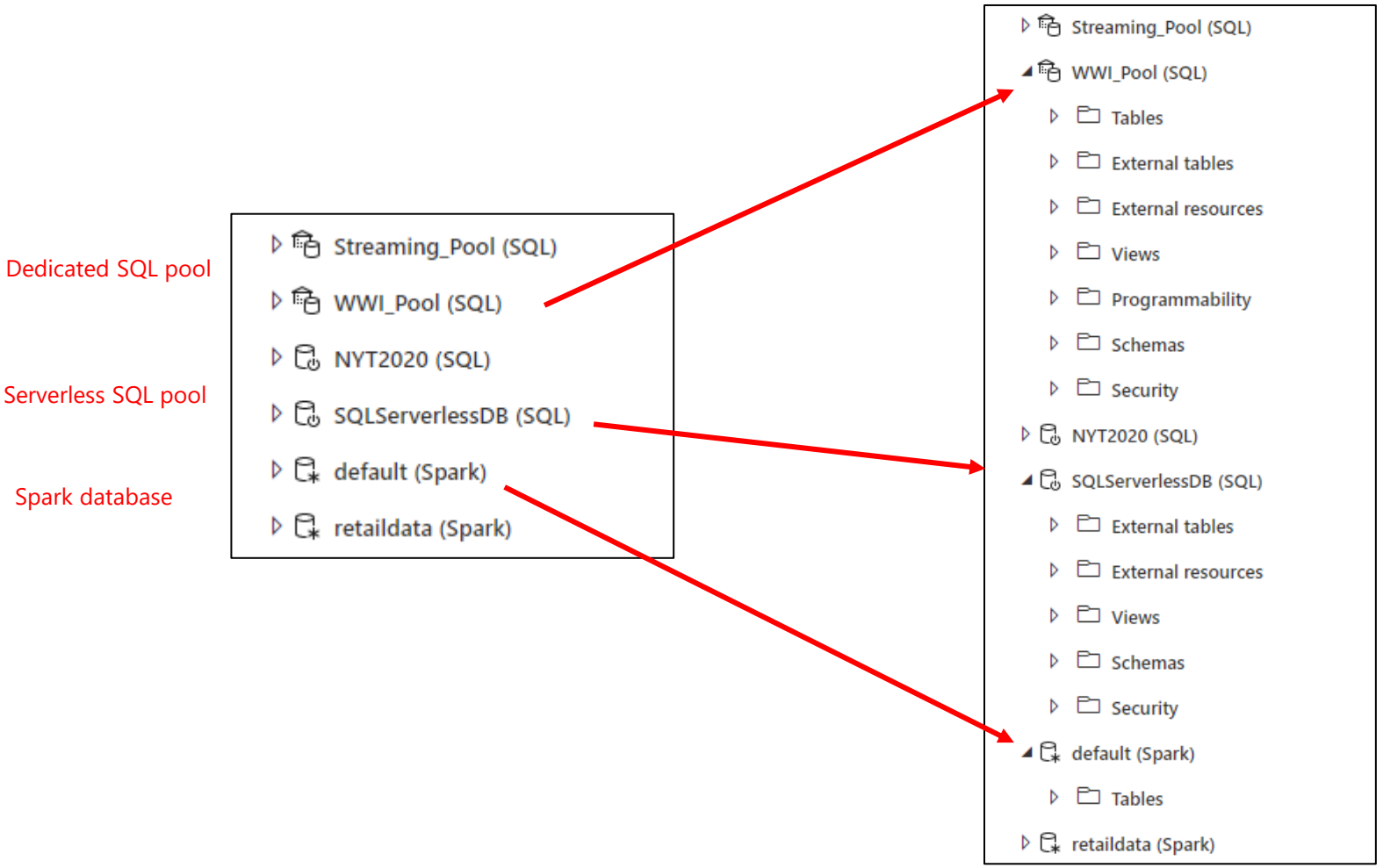






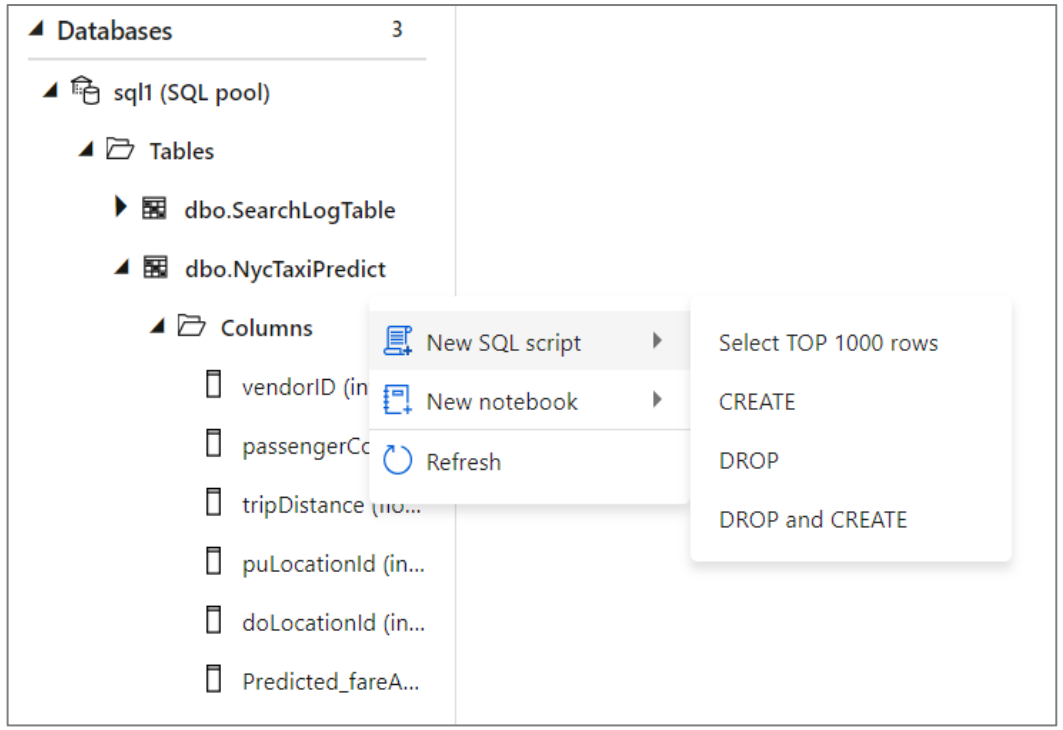
Data Hub – Databases

Explore the different kinds of databases that exist in a workspace.

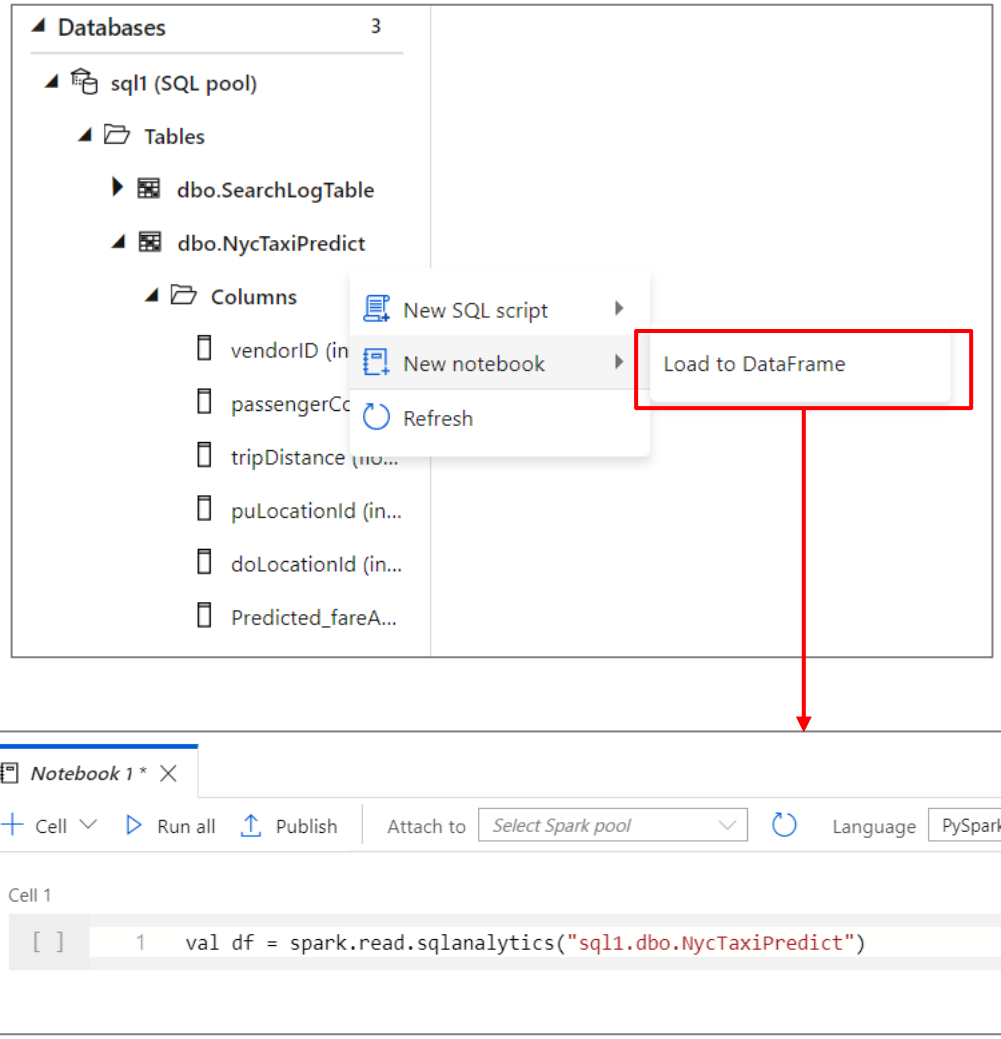


Data Hub – Databases

Familiar gesture to generate T-SQL scripts from SQL metadata objects such as tables.

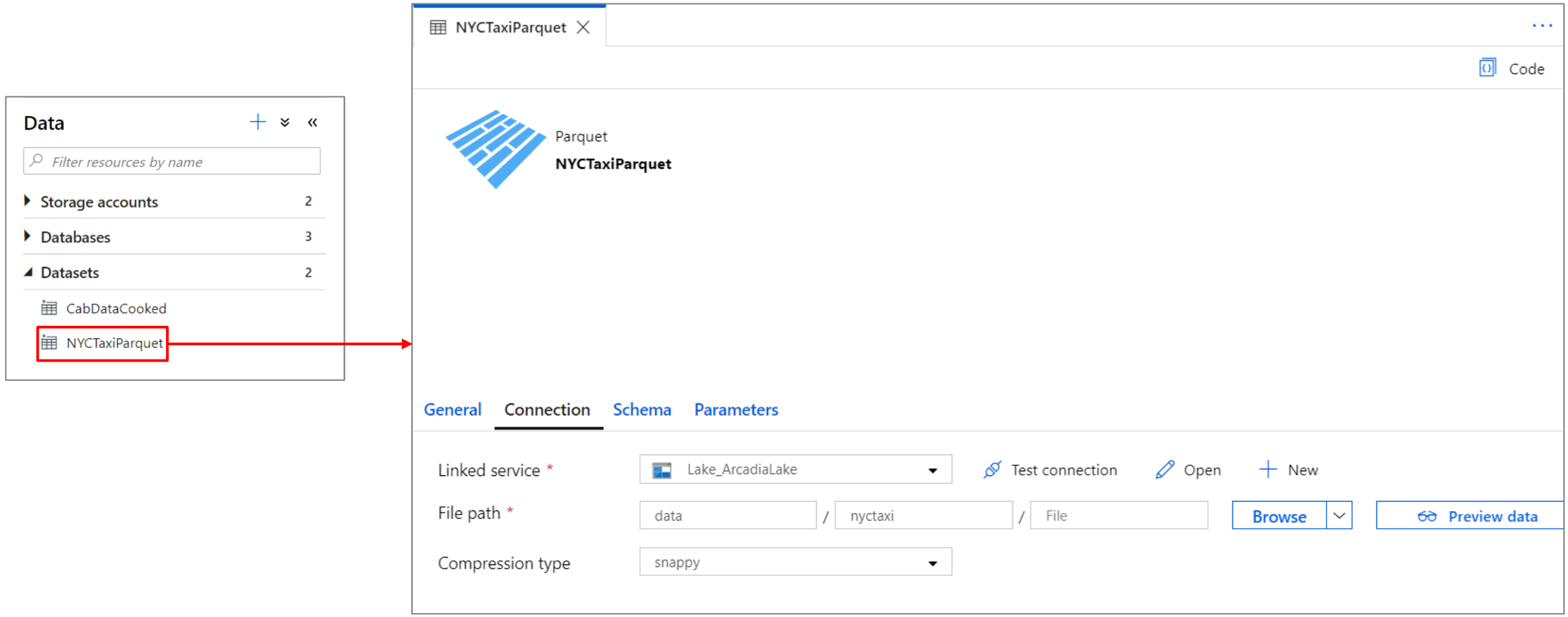


Starting from a table, auto-generate a single line of PySpark code that makes it easy to load a SQL table into a Spark dataframe



Data Hub – Datasets

Orchestration datasets describe data that is persisted. Once a dataset is defined, it can be used in pipelines and sources of data or as sinks of data.



Synapse Studio
Develop hub

Develop Hub

Overview

It provides development experience to query, analyze, model data

Benefits

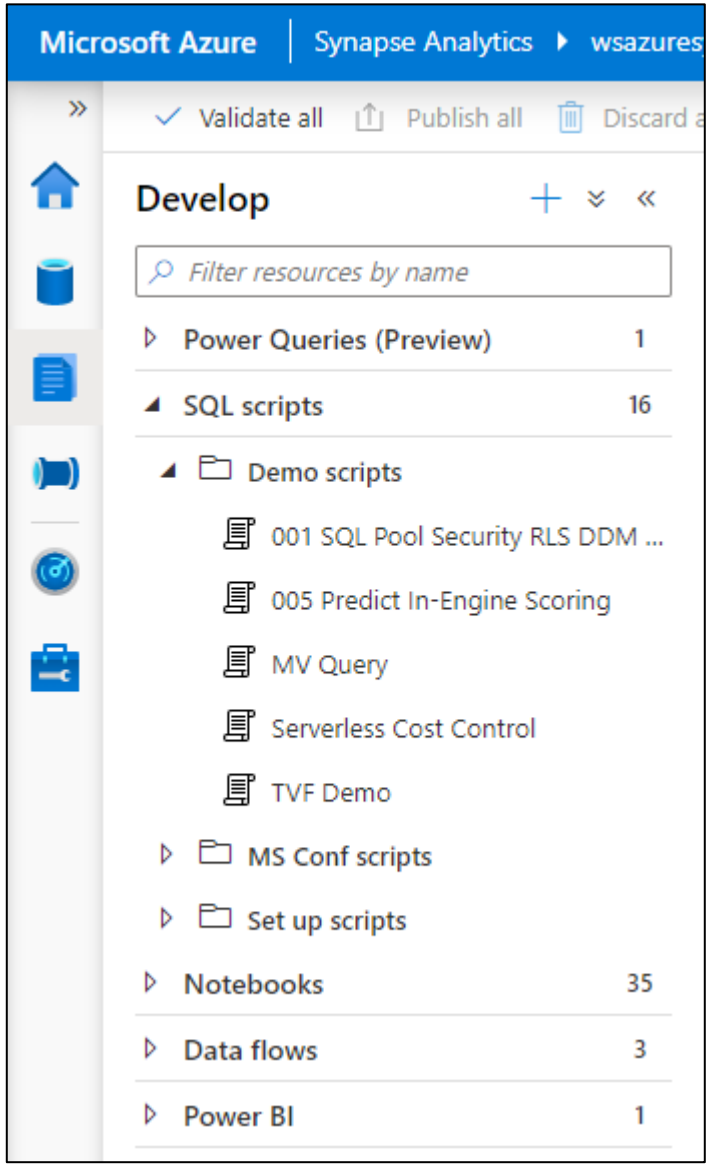
Multiple languages to analyze data under one umbrella

Switch over notebooks and scripts without loosing content

Code intellisense offers reliable code development

Create insightful visualizations

Organize artifacts in folders and sub-folders



Develop Hub - SQL scripts

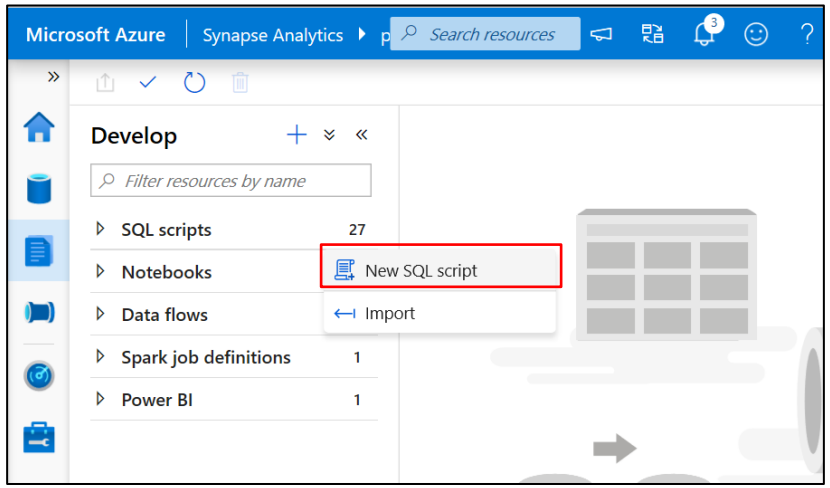
SQL Script

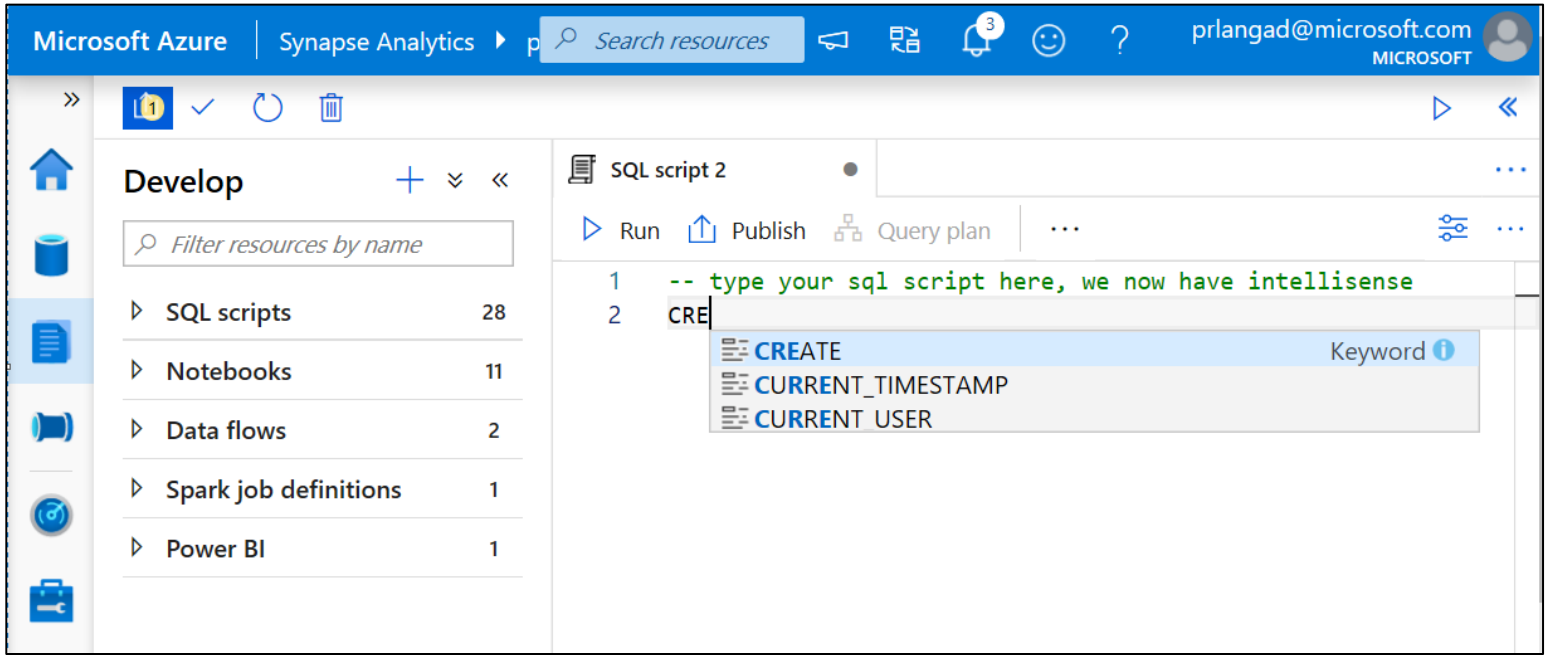
Authoring SQL Scripts

Execute SQL script on dedicated SQL pool or serverless SQL pool

Commit individual SQL script or multiple SQL scripts through Commit all feature

Language support and intellisense

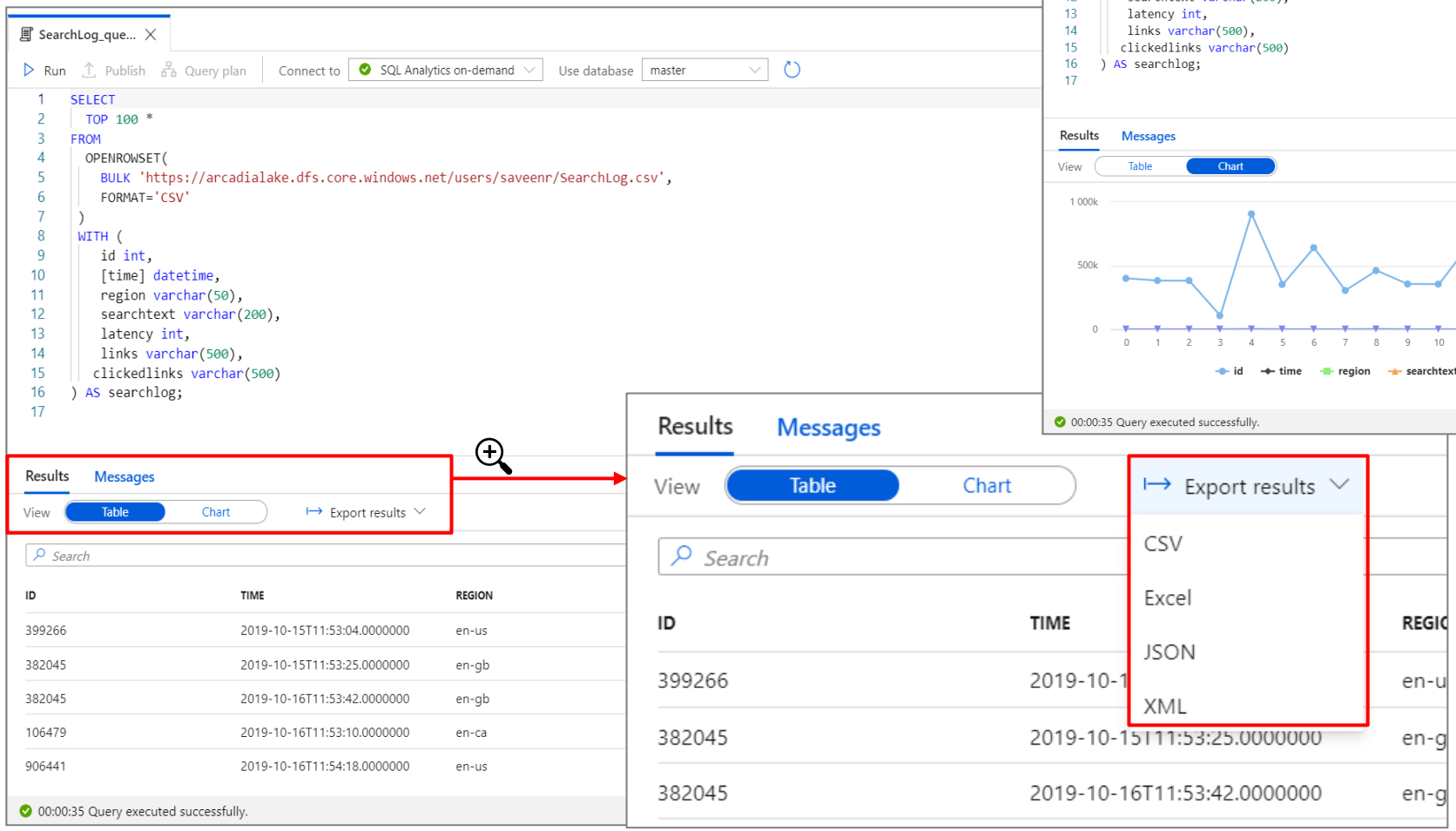


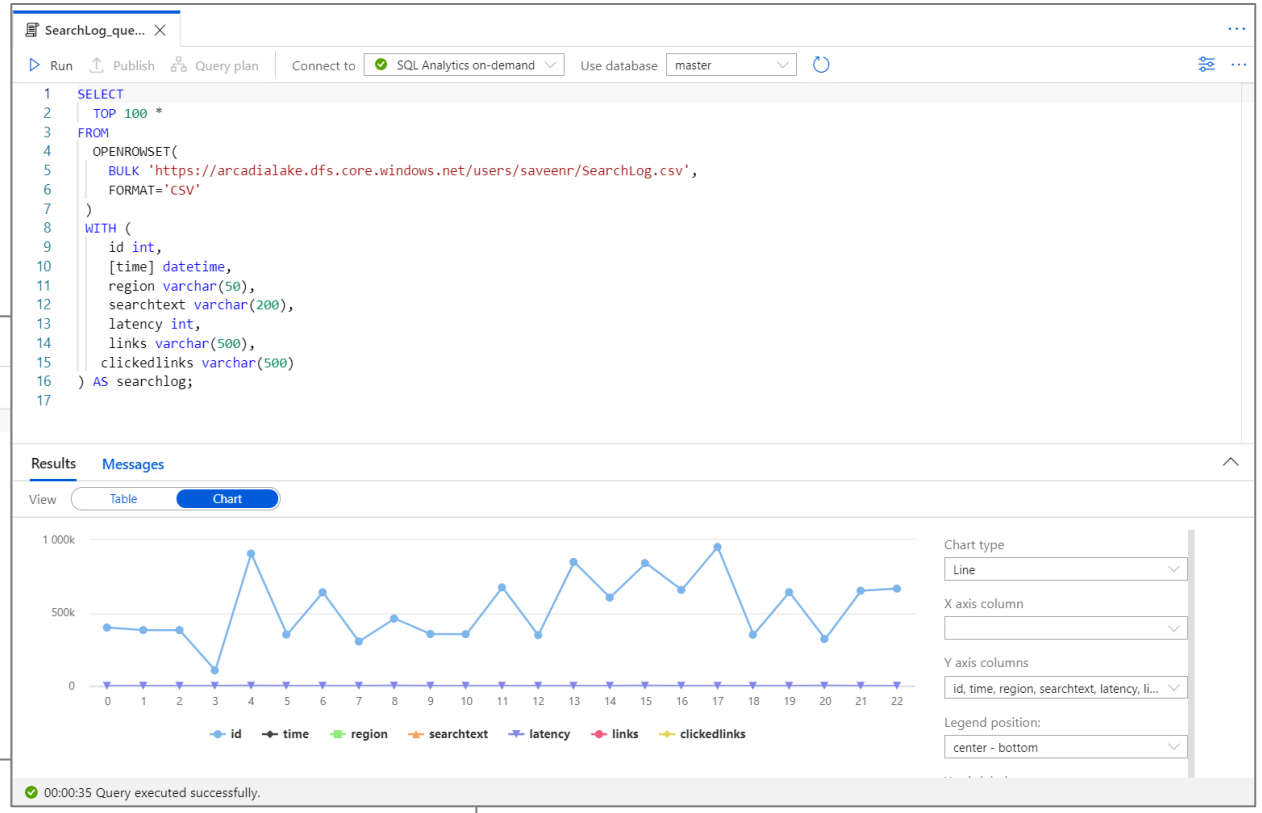


Develop Hub - SQL scripts

SQL Script

View results in Table or Chart form and export results in several popular formats





Develop Hub - Notebooks

Notebooks

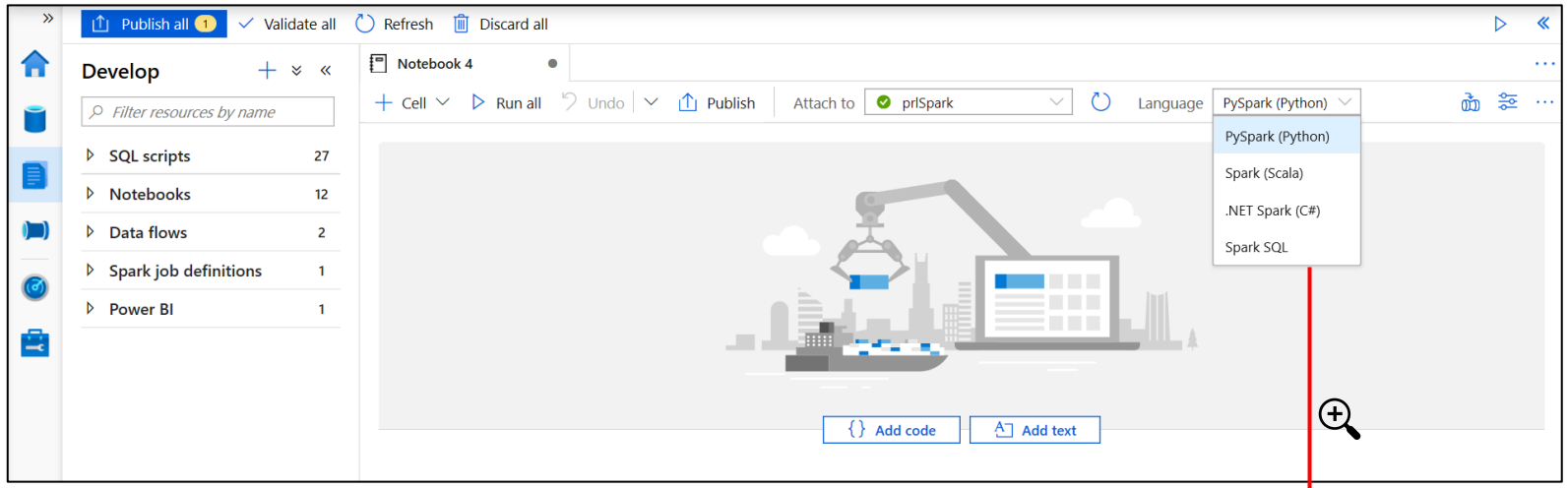
Allows to write multiple languages in one notebook

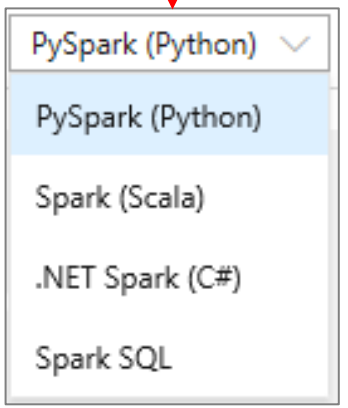
%%<Name of language>

Offers use of temporary tables across languages

Language support for Syntax highlight, syntax error, syntax code completion, smart indent, code folding

Export results

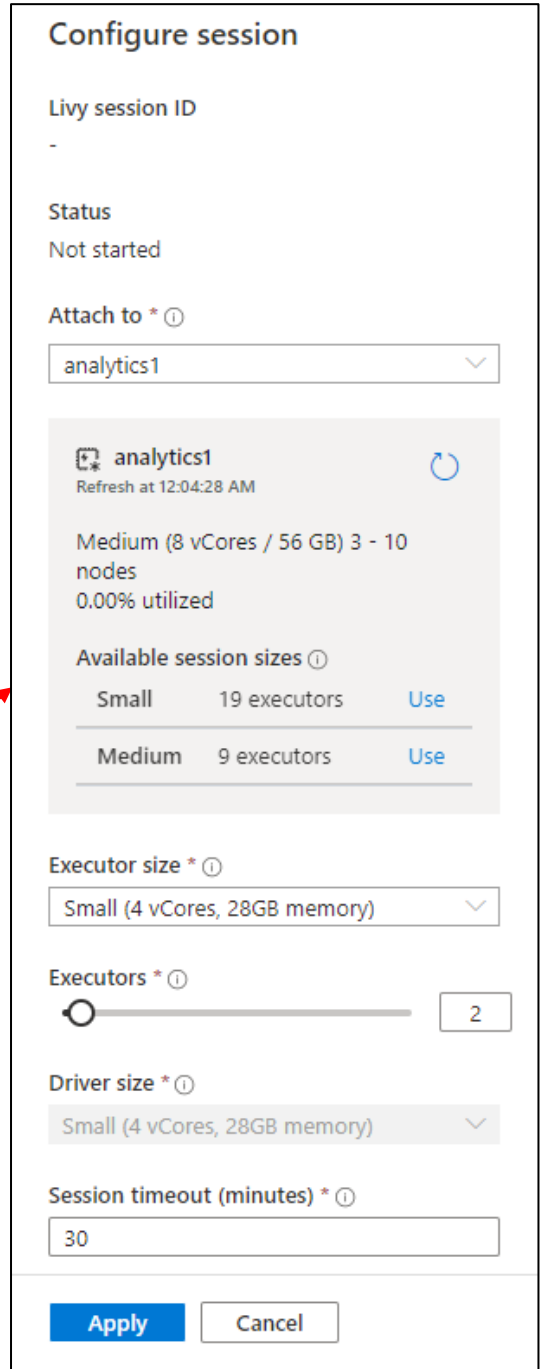




Develop Hub - Notebooks

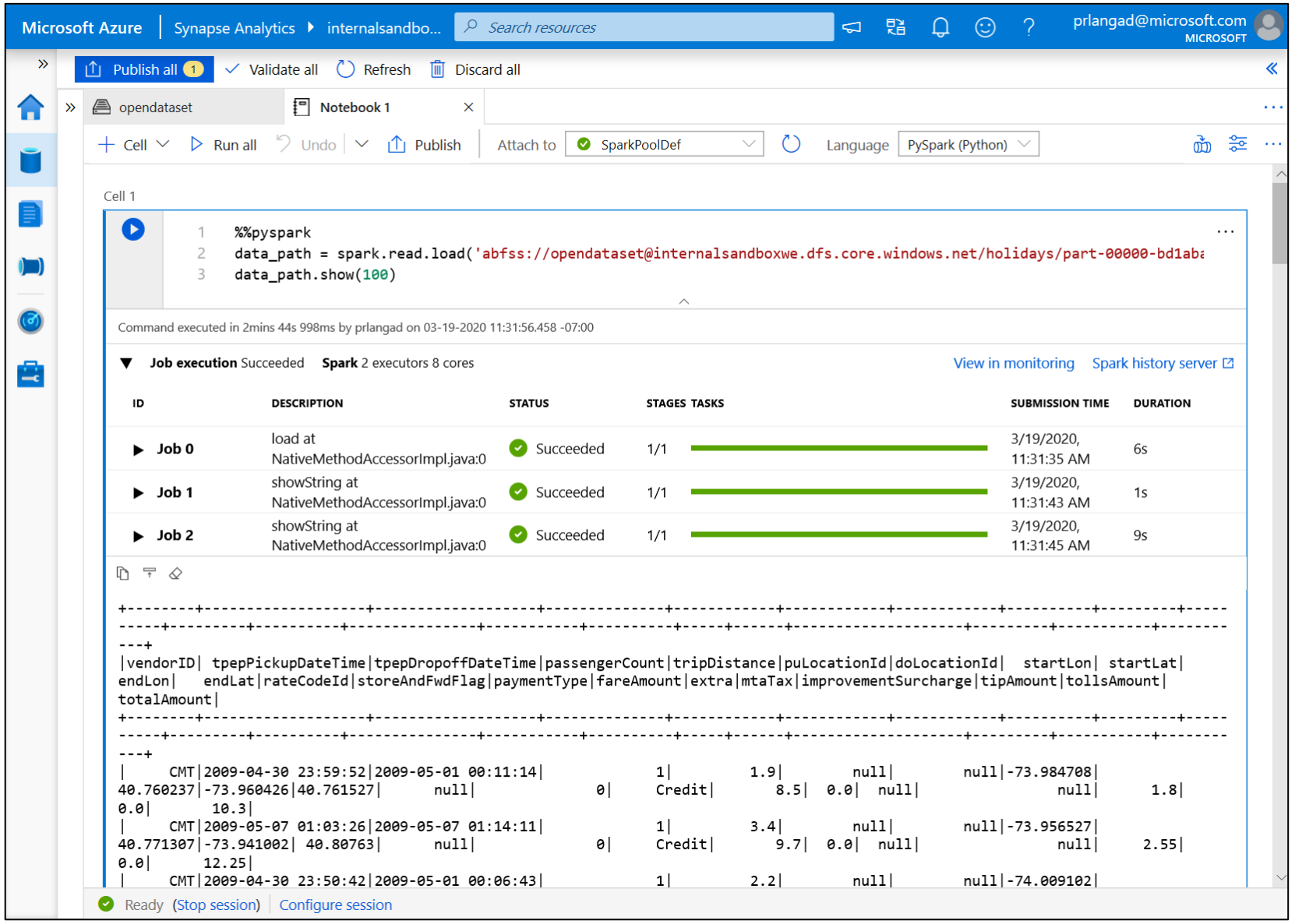
Configure session allows developers to control how many resources are devoted to running their notebook.

Provides quick links to monitor session and Spark history server



Develop Hub - Notebooks

As notebook cells run, the underlying Spark application status is shown. Providing immediate feedback and progress tracking.



Dataflow Capabilities

Handle upserts, updates, deletes on sql sinks

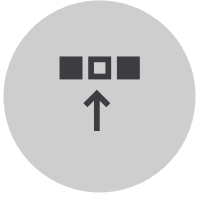
Add new partition methods

Add schema drift support

Add file handling (move files after read, write files to file names described in rows etc)

New inventory of functions (for e.g. Hash functions for row comparison)

Commonly used ETL patterns(Sequence generator/Lookup transformation/SCD…)



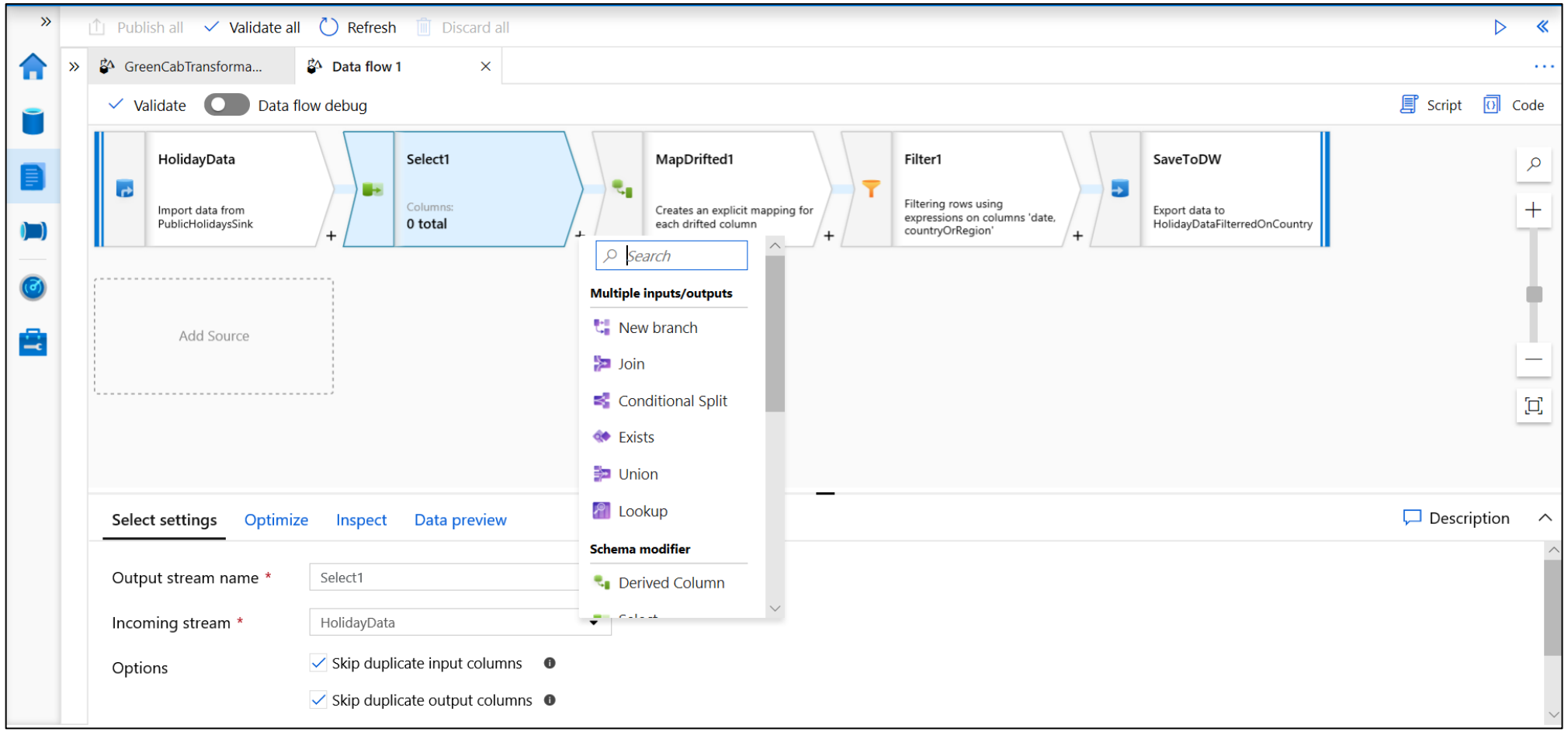
Data lineage – Capturing sink column lineage & impact analysis(invaluable if this is for enterprise deployment)

Implement commonly used ETL patterns as templates(SCD Type1, Type2, Data Vault)

Develop Hub - Data Flows

Data flows are a visual way of specifying how to transform data.

Provides a code-free experience.



Develop Hub – Power BI

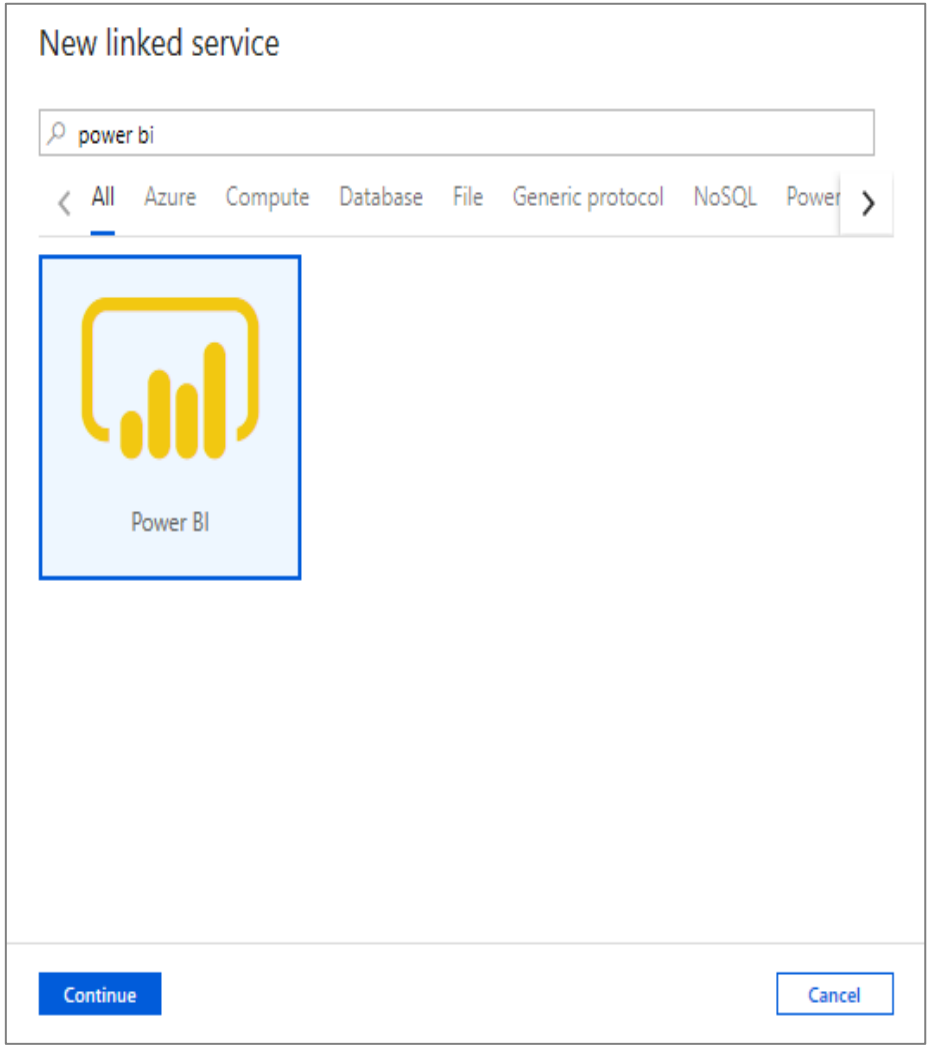
Overview

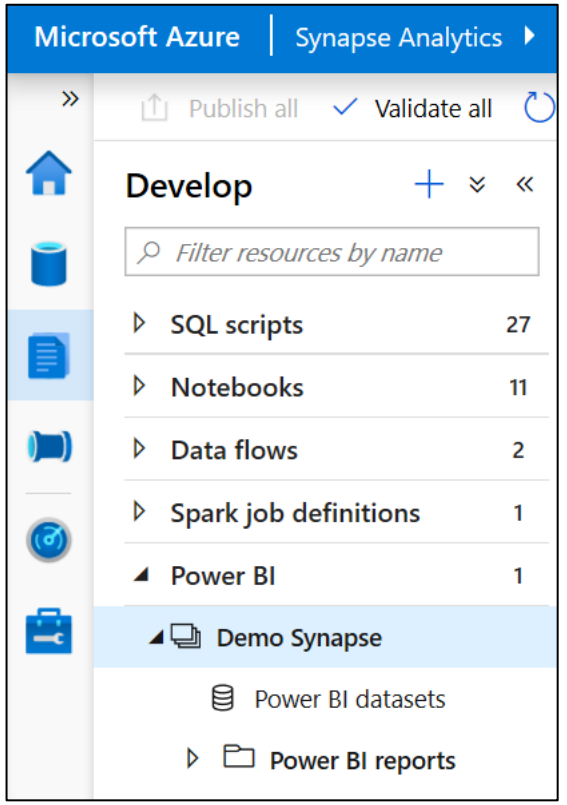
Create Power BI reports in the workspace

Provides access to published reports in the workspace

Update reports real time from Synapse workspace to get it reflected on Power BI service

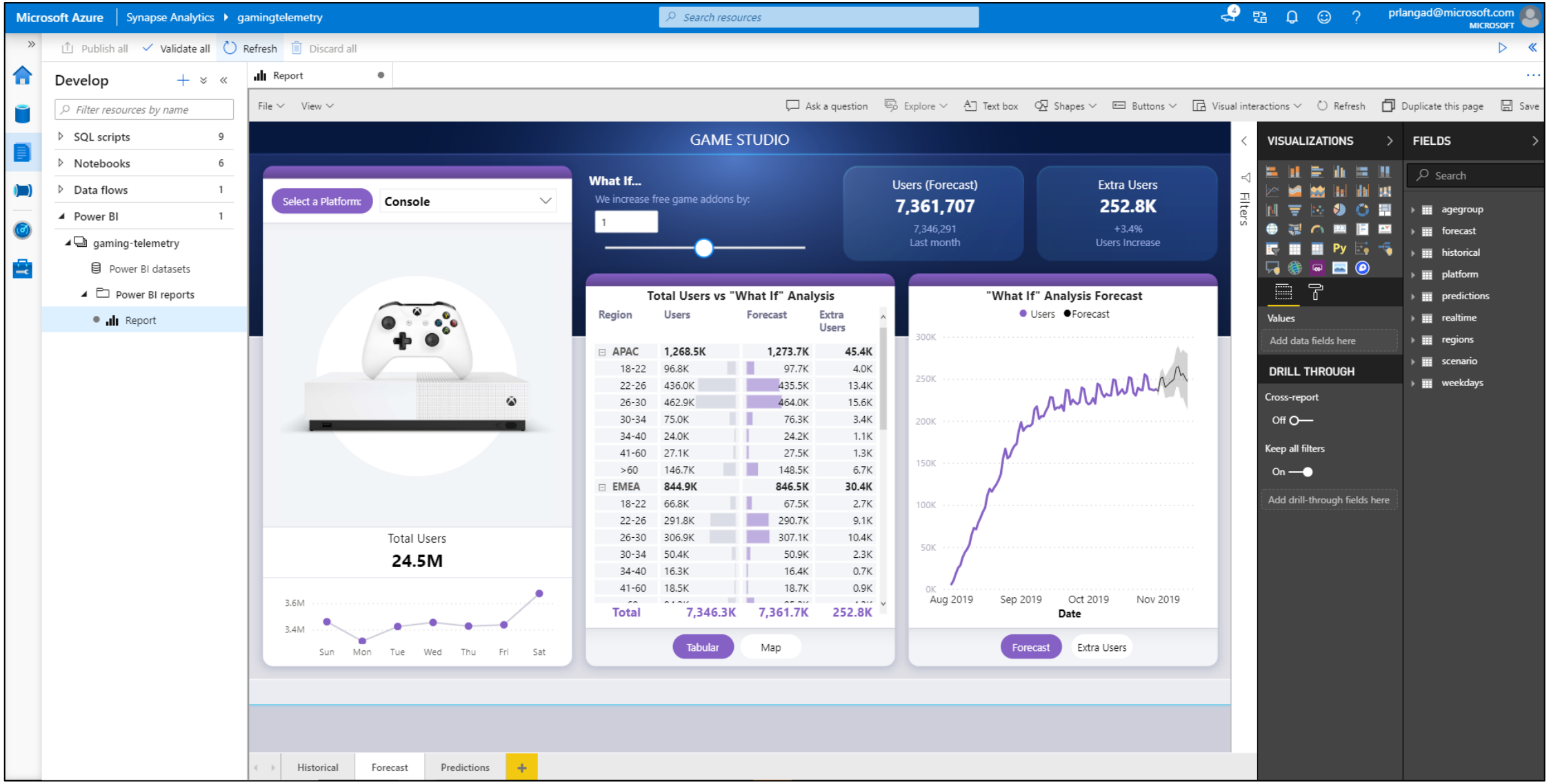
Visually explore and analyze data





Develop Hub – Power BI

View published reports in Power BI workspace

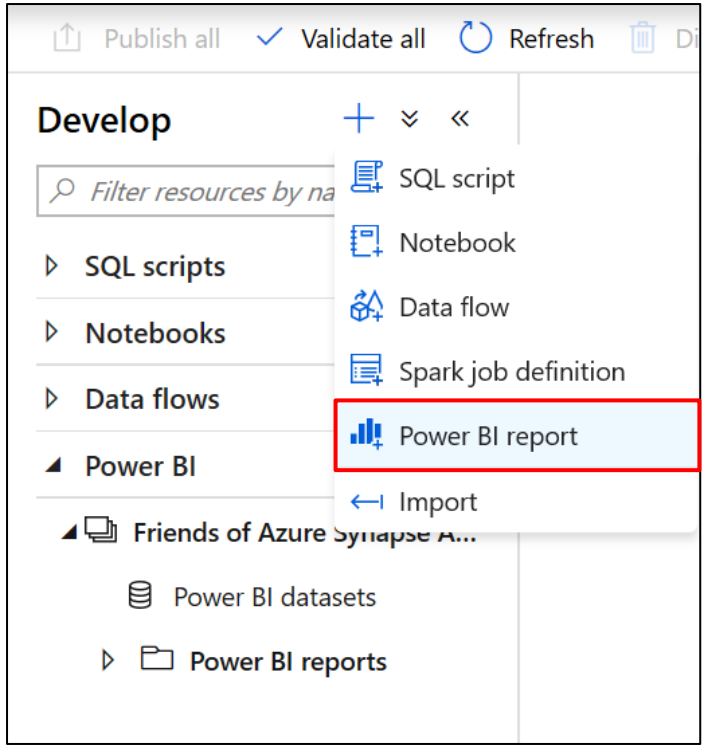


Develop Hub – Power BI

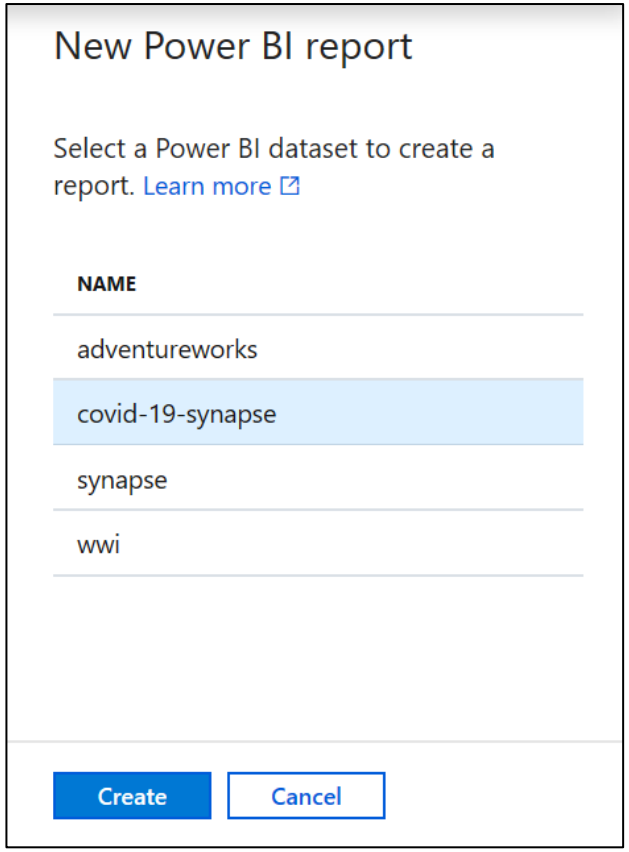
Create new reports from existing published Power BI datasets

Create new Power BI datasets

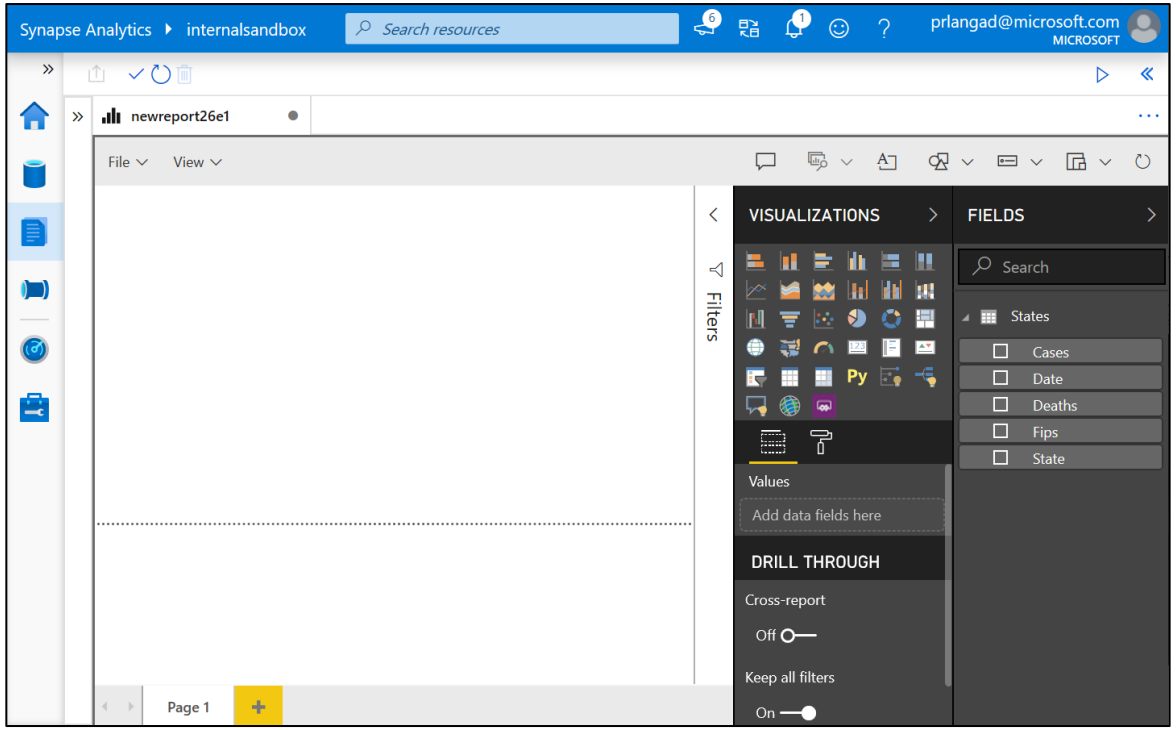
1



2

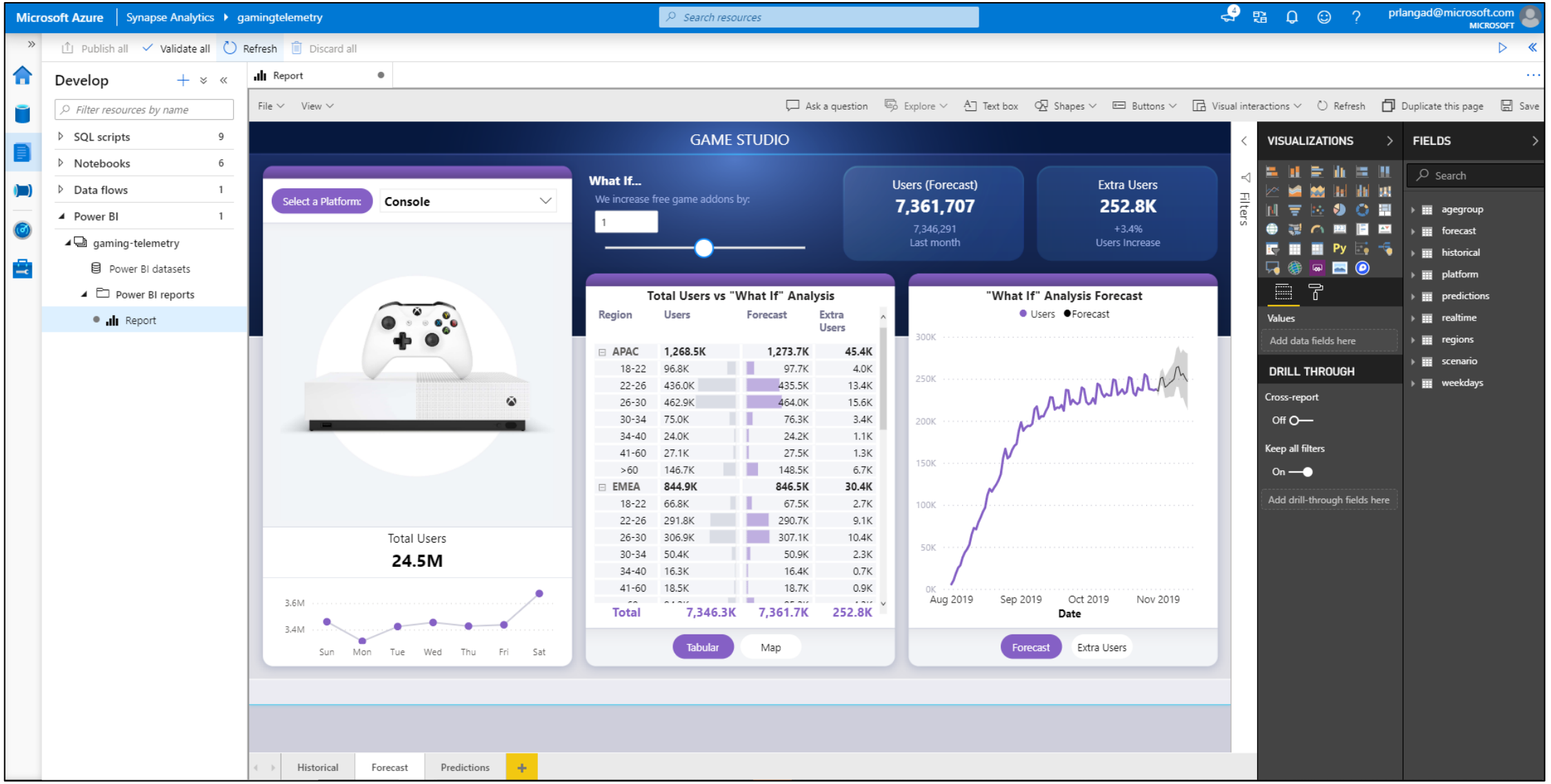


3



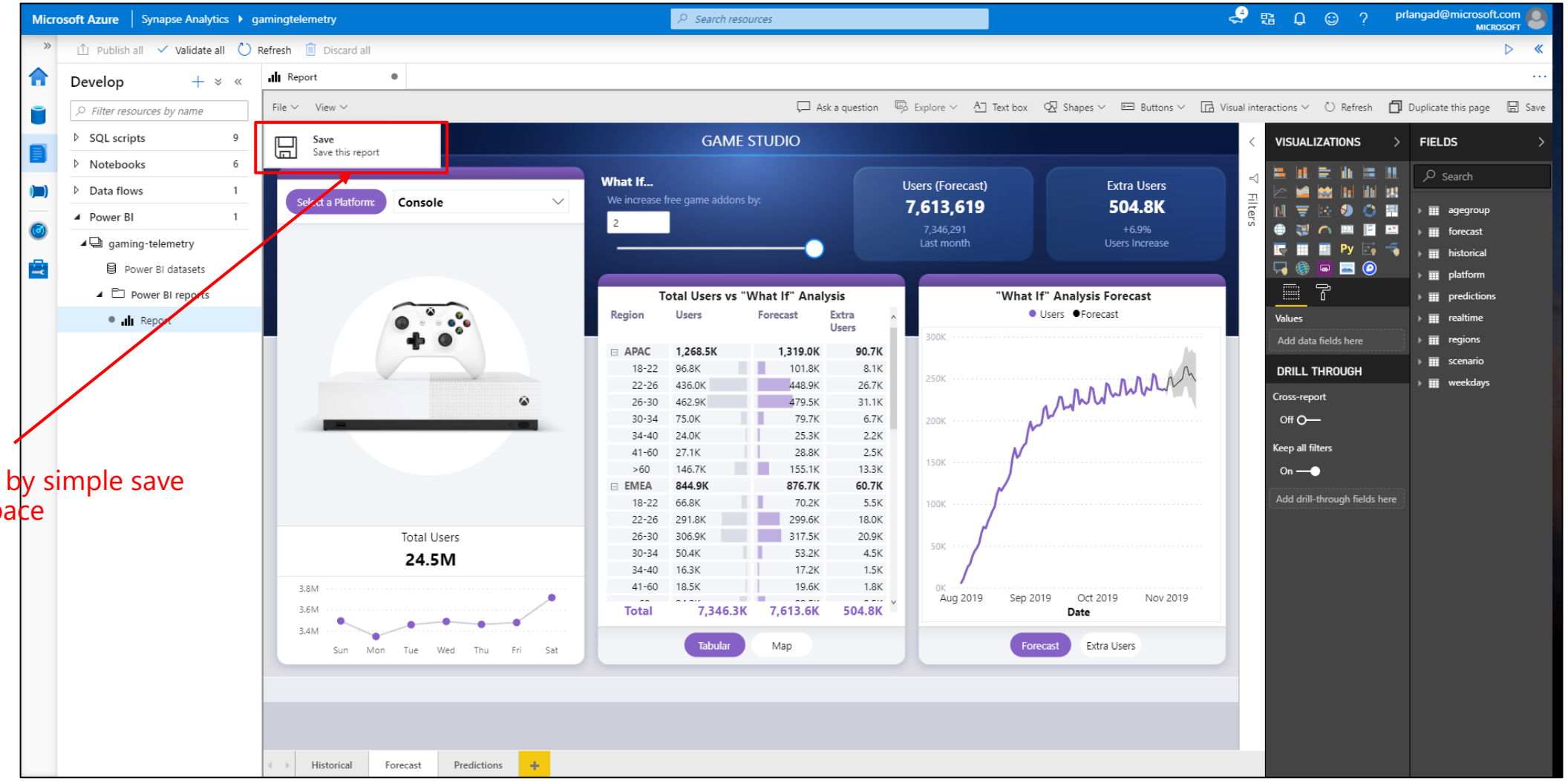
Develop Hub – Power BI

Edit reports in Synapse workspace



Develop Hub – Power BI

Publish edited reports in Synapse workspace to Power BI workspace



Publish changes by simple save report in workspace

Microsoft | Power BI | gaming-telemetry > Report

New look off | Search

File | View | Edit report | Explore | Refresh | Pin a live Page | Reset to default | Comments | Bookmarks | Usage metrics | View related | Favorite | Subscribe | Share

Home
Favorites
Recent
Apps
Shared with me
Workspaces
gaming-telemetry

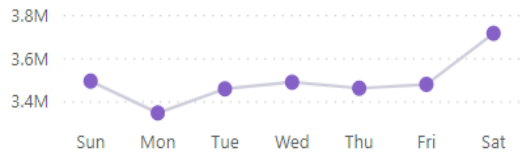
Real-time publish on save

Get data

GAME STUDIO

Select a Platform: **Console**

Total Users
24.5M



What If...
We increase free game addons by:
2

Users (Forecast)
7,613,619
7,346,291
Last month

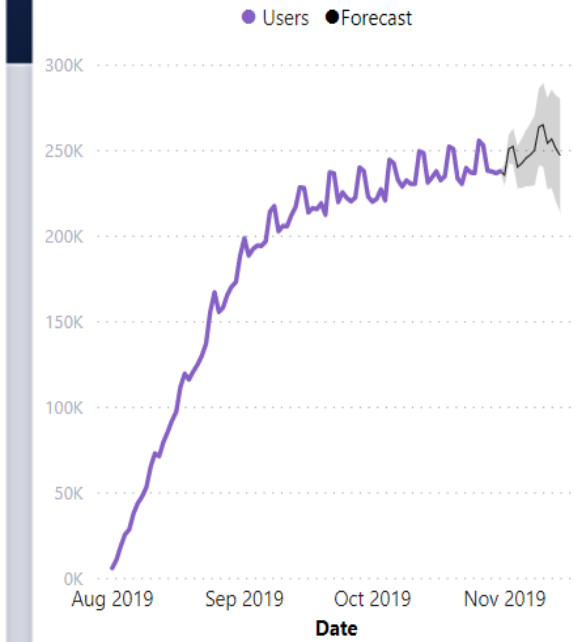
Extra Users
504.8K
+6.9%
Users Increase

Total Users vs "What If" Analysis

Region	Users	Forecast	Extra Users
APAC	1,268.5K	1,319.0K	90.7K
18-22	96.8K	101.8K	8.1K
22-26	436.0K	448.9K	26.7K
26-30	462.9K	479.5K	31.1K
30-34	75.0K	79.7K	6.7K
34-40	24.0K	25.3K	2.2K
41-60	27.1K	28.8K	2.5K
>60	146.7K	155.1K	13.3K
EMEA	844.9K	876.7K	60.7K
18-22	66.8K	70.2K	5.5K
22-26	291.8K	299.6K	18.0K
26-30	306.9K	317.5K	20.9K
30-34	50.4K	53.2K	4.5K
34-40	16.3K	17.2K	1.5K
41-60	18.5K	19.6K	1.8K
Total	7,346.3K	7,613.6K	504.8K

Tabular | Map

"What If" Analysis Forecast



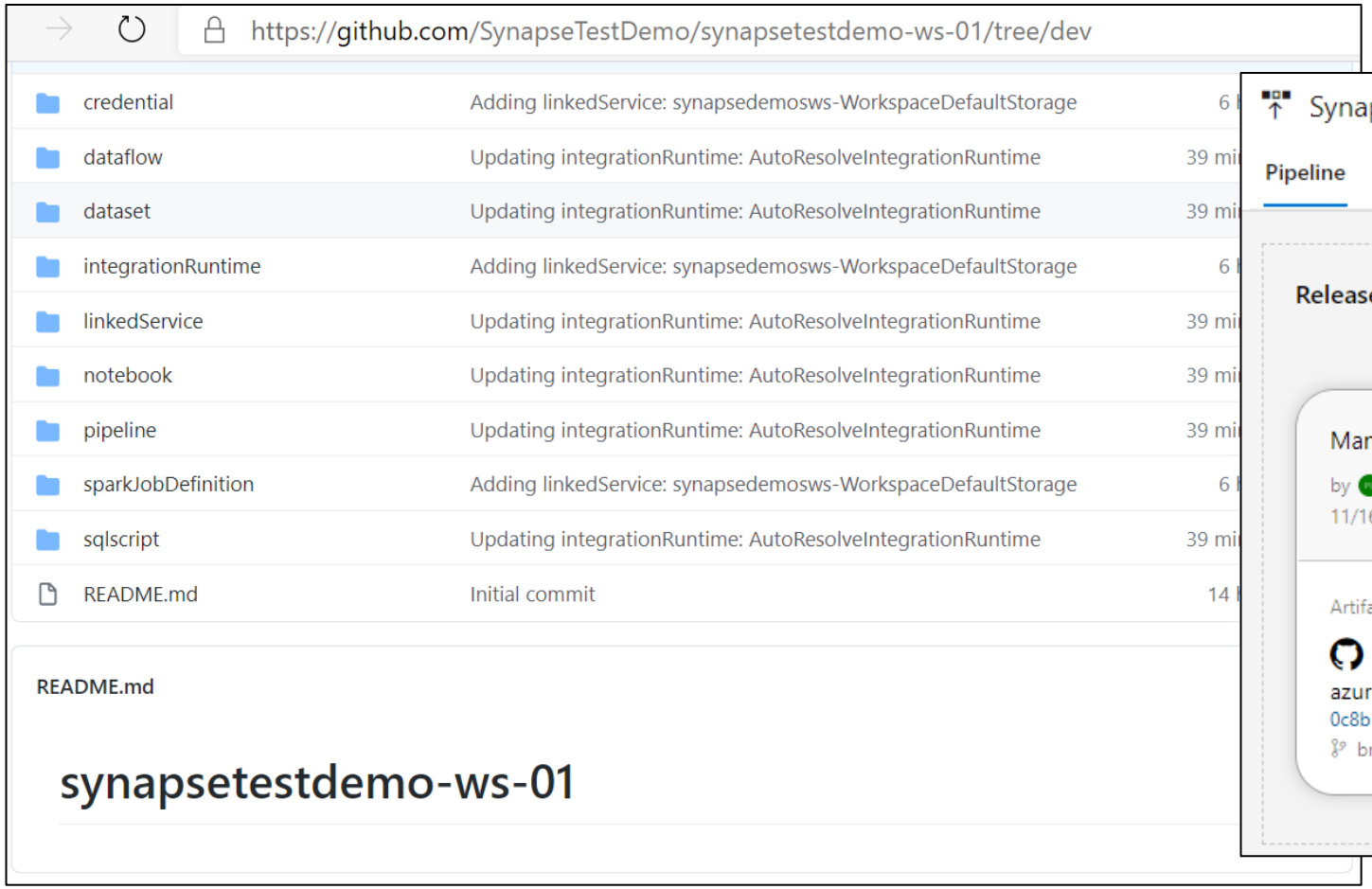
Forecast | Extra Users

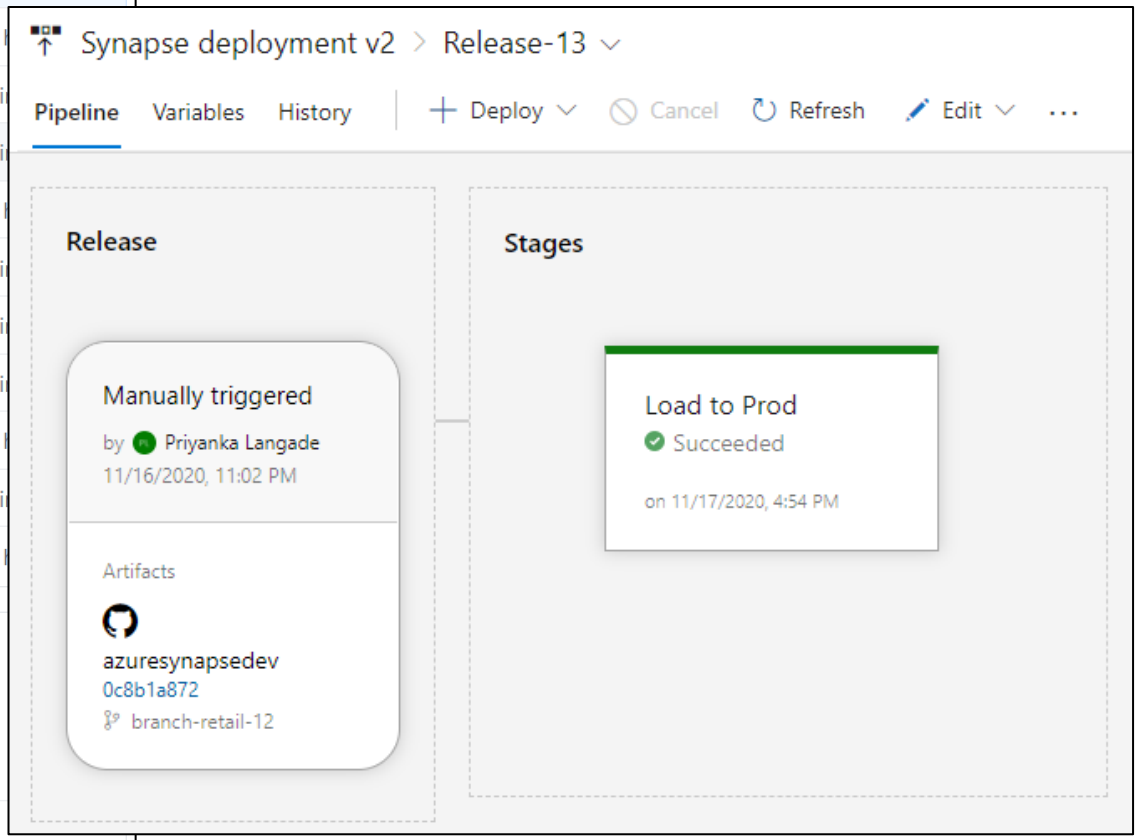
Historical | Forecast | Predictions

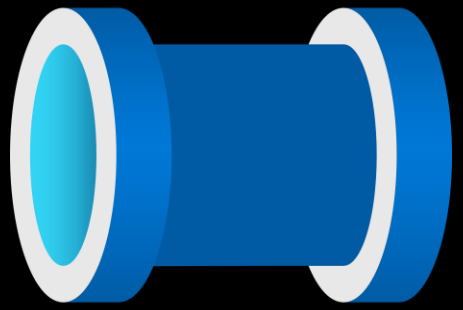
Filters

Develop – CI/CD

Commit artifacts to source-controlled repository and operationalize release pipelines with Synapse deployment task





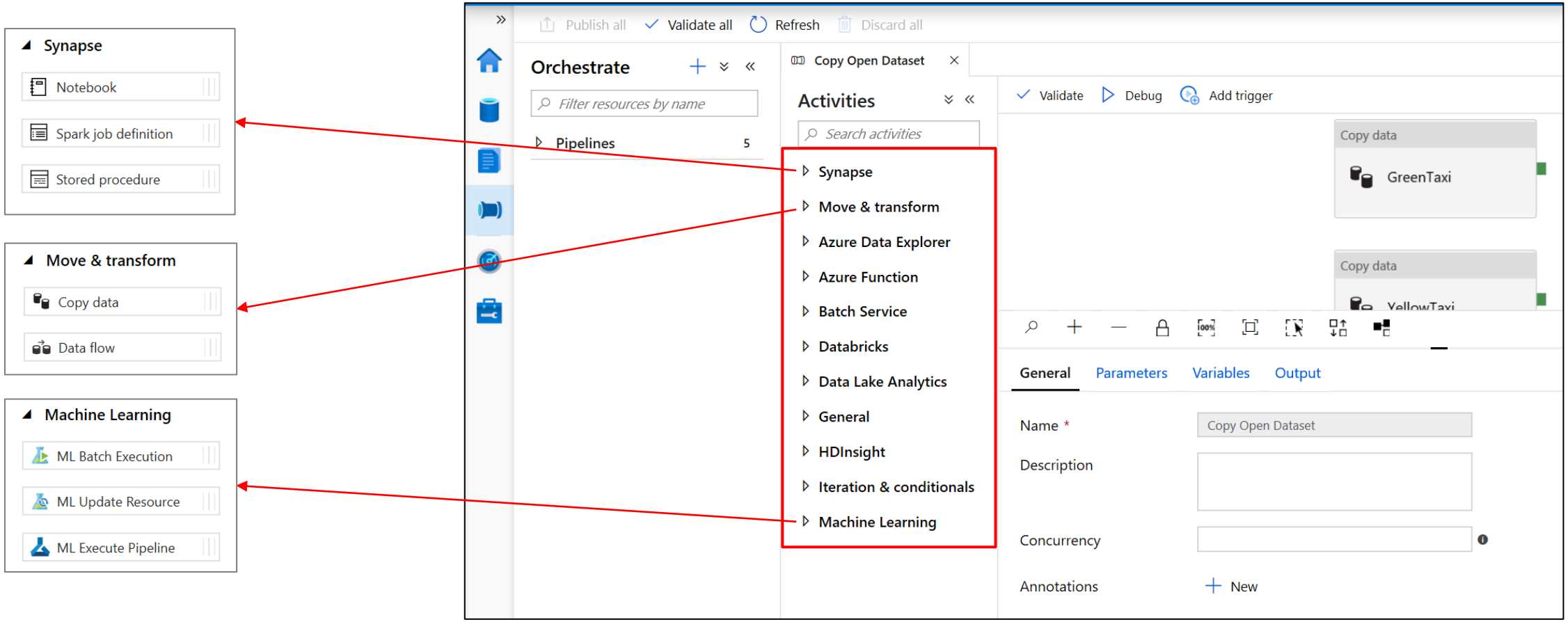


Synapse Studio
Integrate hub

Integrate Hub

It provides ability to create pipelines to ingest, transform and load data with 90+ inbuilt connectors.

Offers a wide range of activities that a pipeline can perform.



Synapse Studio
Monitor hub

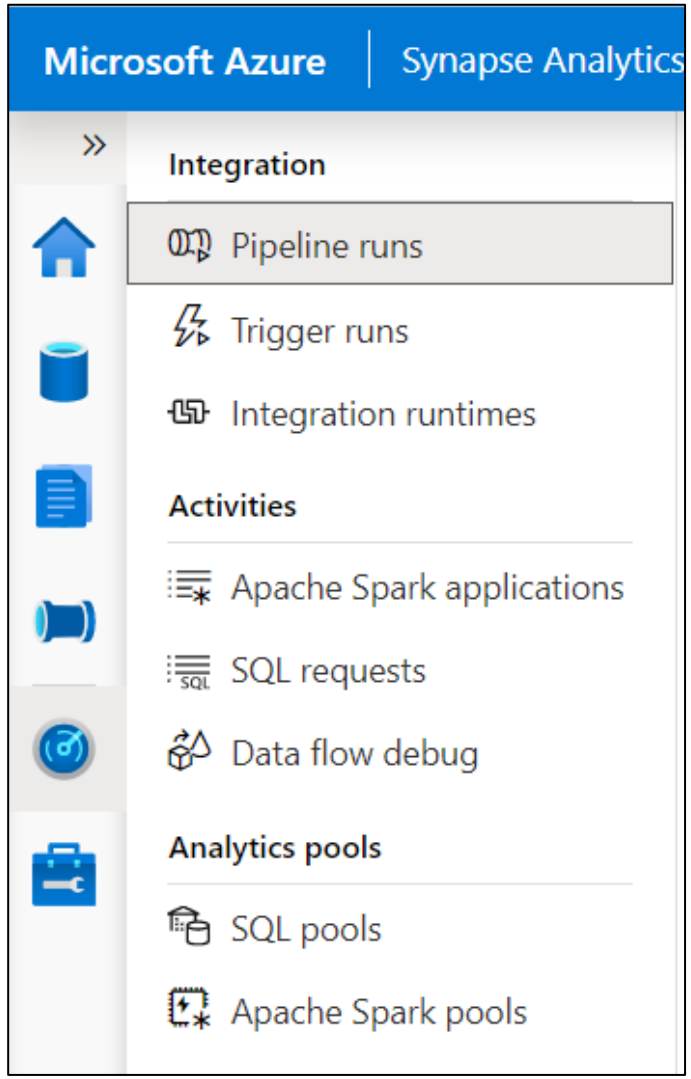
Monitor Hub

Overview

This feature provides single pane of glass to monitor orchestration, activities for Apache Spark Application and SQL requests.

Benefits

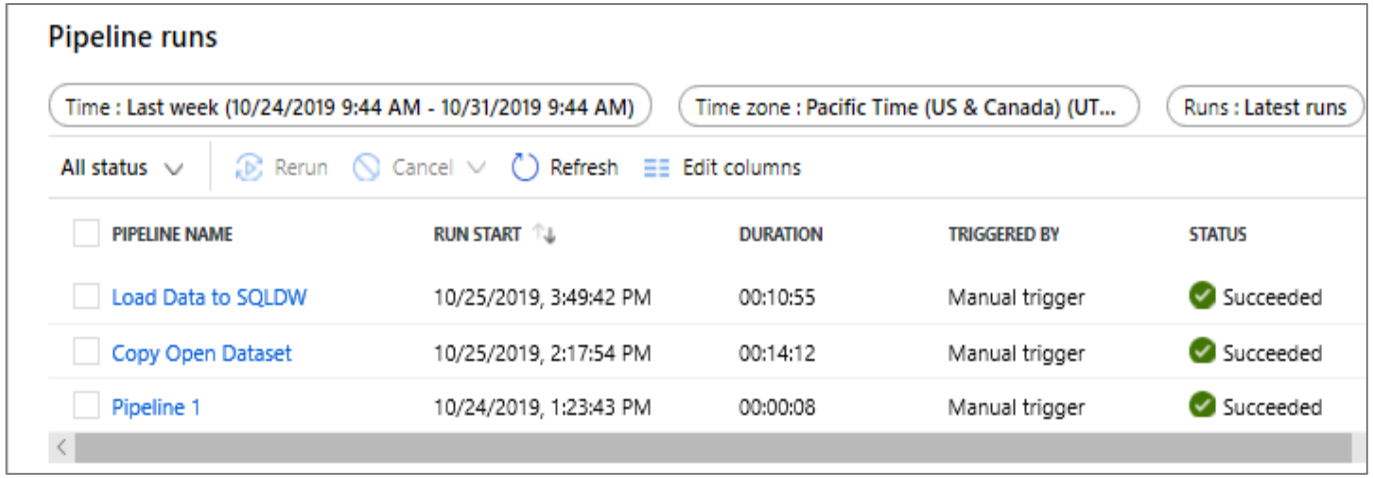
Offers additional filters to monitor specific activities or orchestration



Monitor Hub - Integration

Overview

Monitor orchestration in the Synapse workspace for the progress and status of pipeline

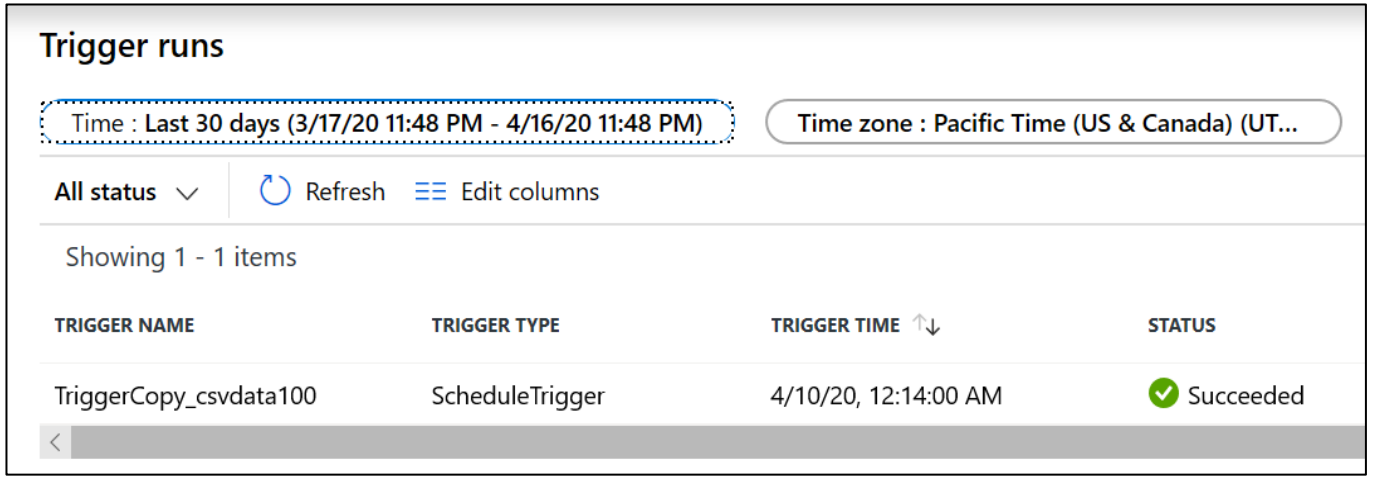


Benefits

Track all/specific pipelines

Monitor pipeline run and activity run details

Find the root cause of pipeline failure or activity failure



Monitor Hub - Spark applications

Overview

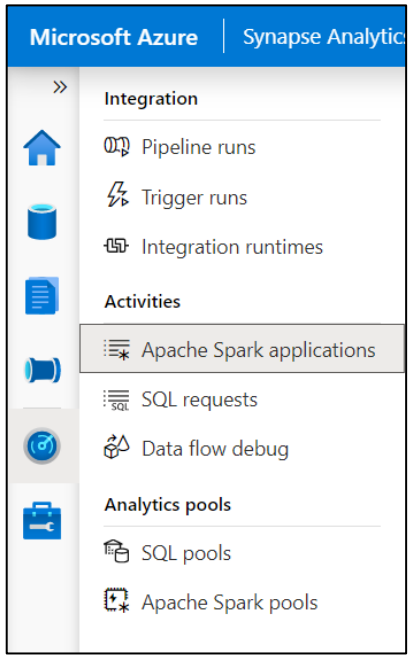
Monitor Spark pools, Spark applications for the status of activities

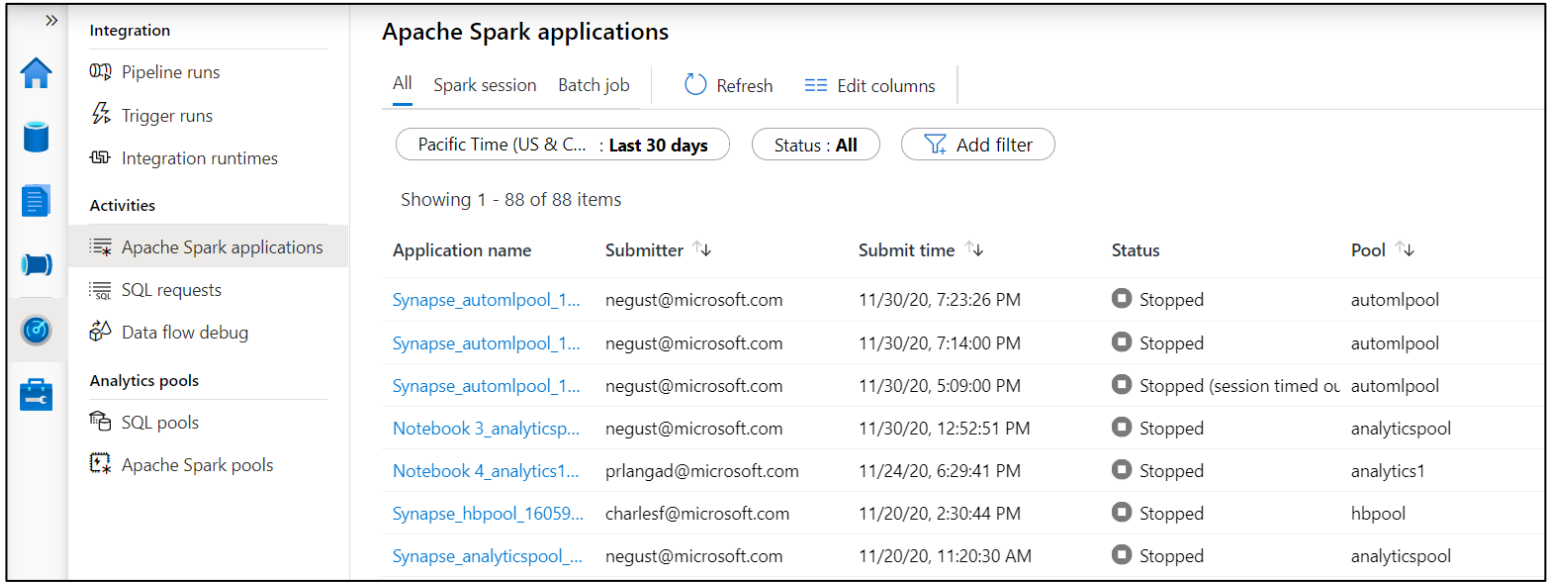
Benefits

Apply filter for pool to get Apache Spark Applications per pool

Additional available filters include

1. Application Name
2. Livy ID
3. Status
4. End time





Monitor Hub – SQL requests

Overview

Monitor SQL requests for the progress and status of activities

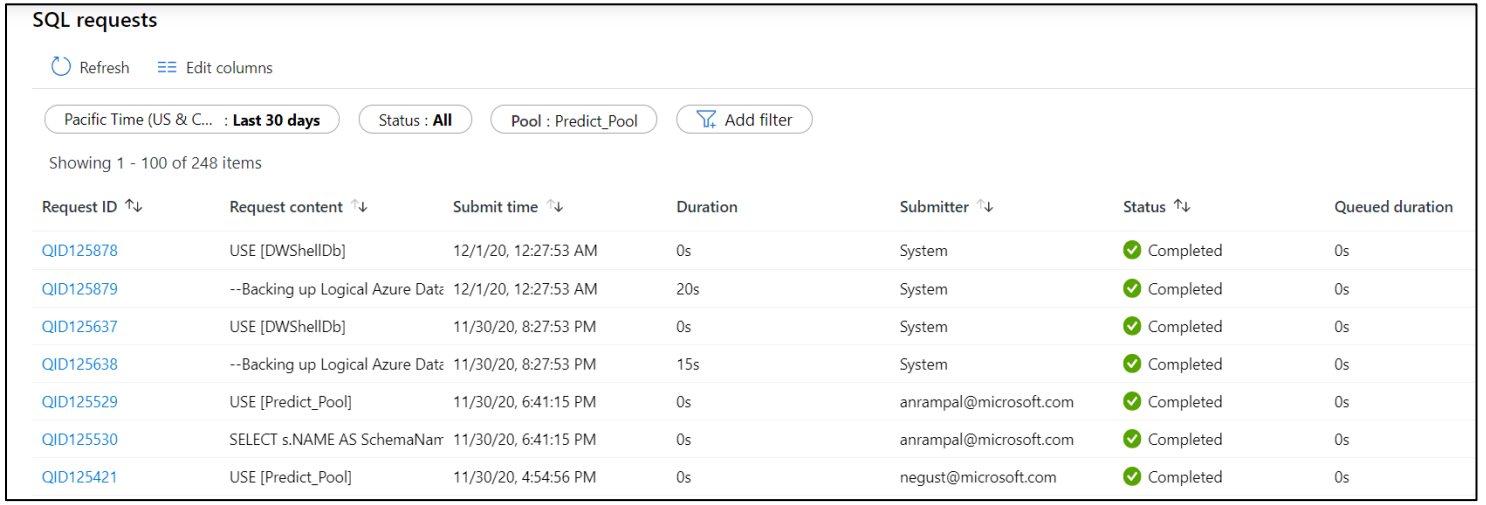
Benefits

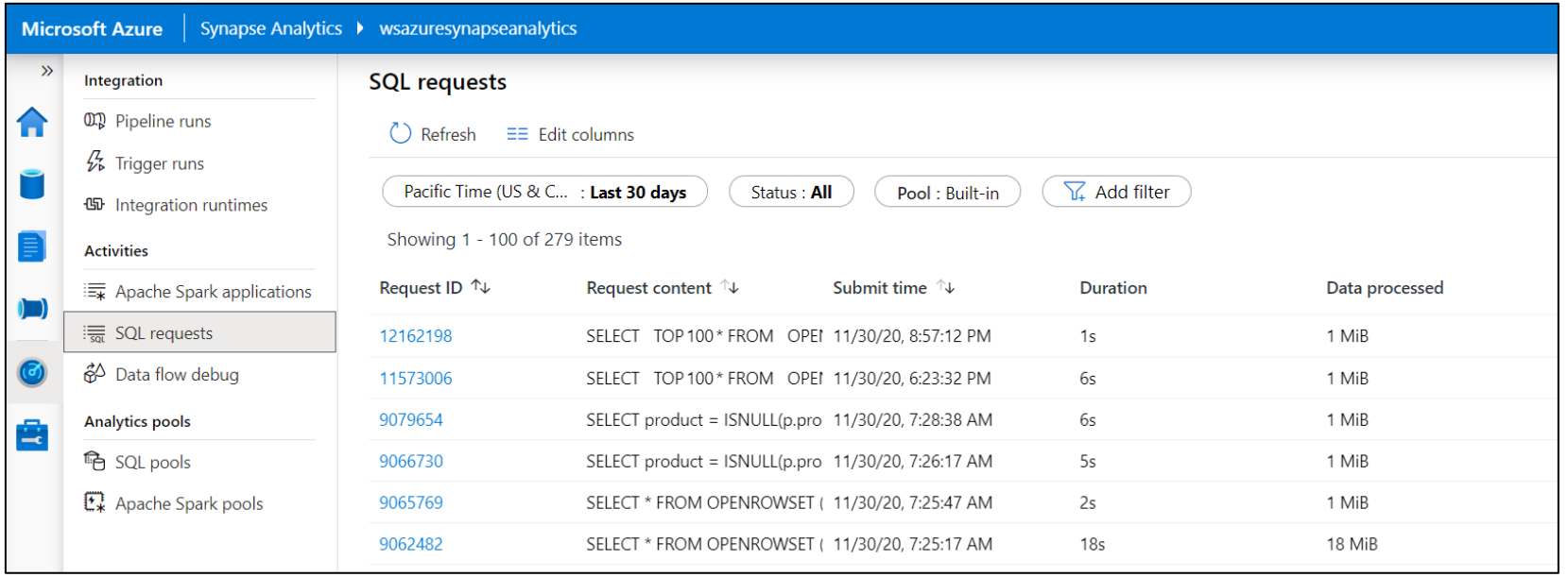
Apply filter for pool to get SQL requests per compute pool

Validate query text

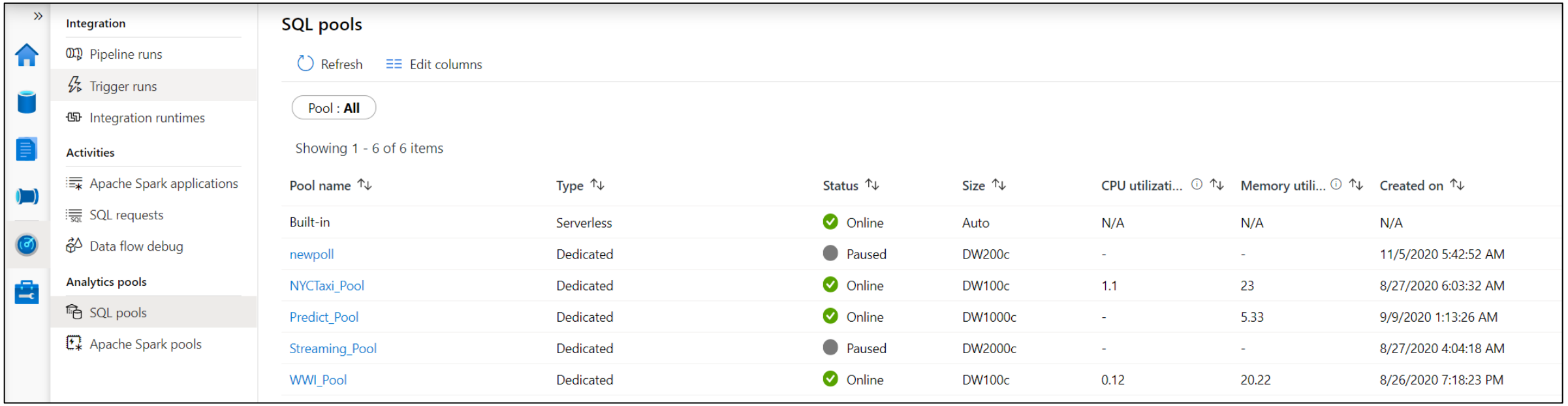
Additional available filters include

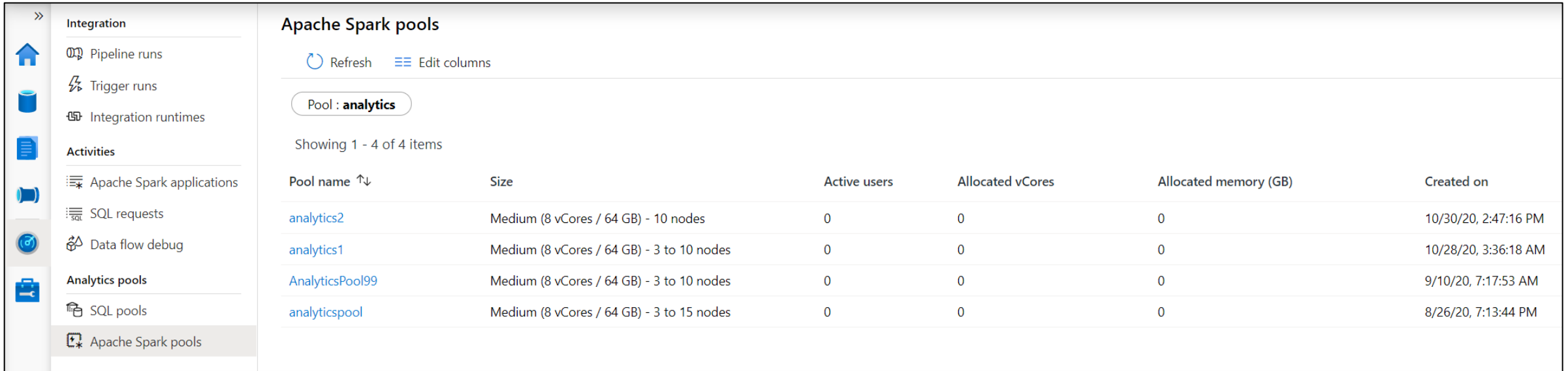
1. Start time
2. End time
3. Request ID
4. Session ID
5. Submitter
6. Workload group





Monitor Hub – Analytics pools



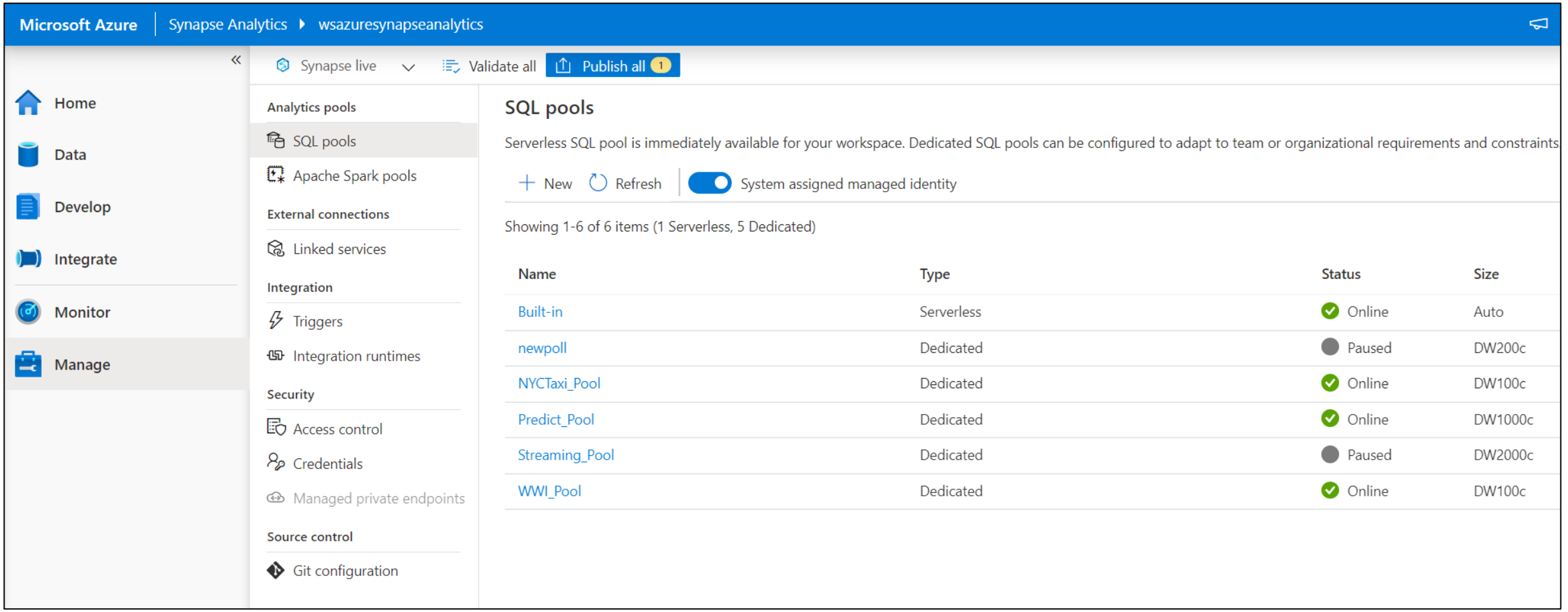


Synapse Studio
Manage hub

Manage Hub

Overview

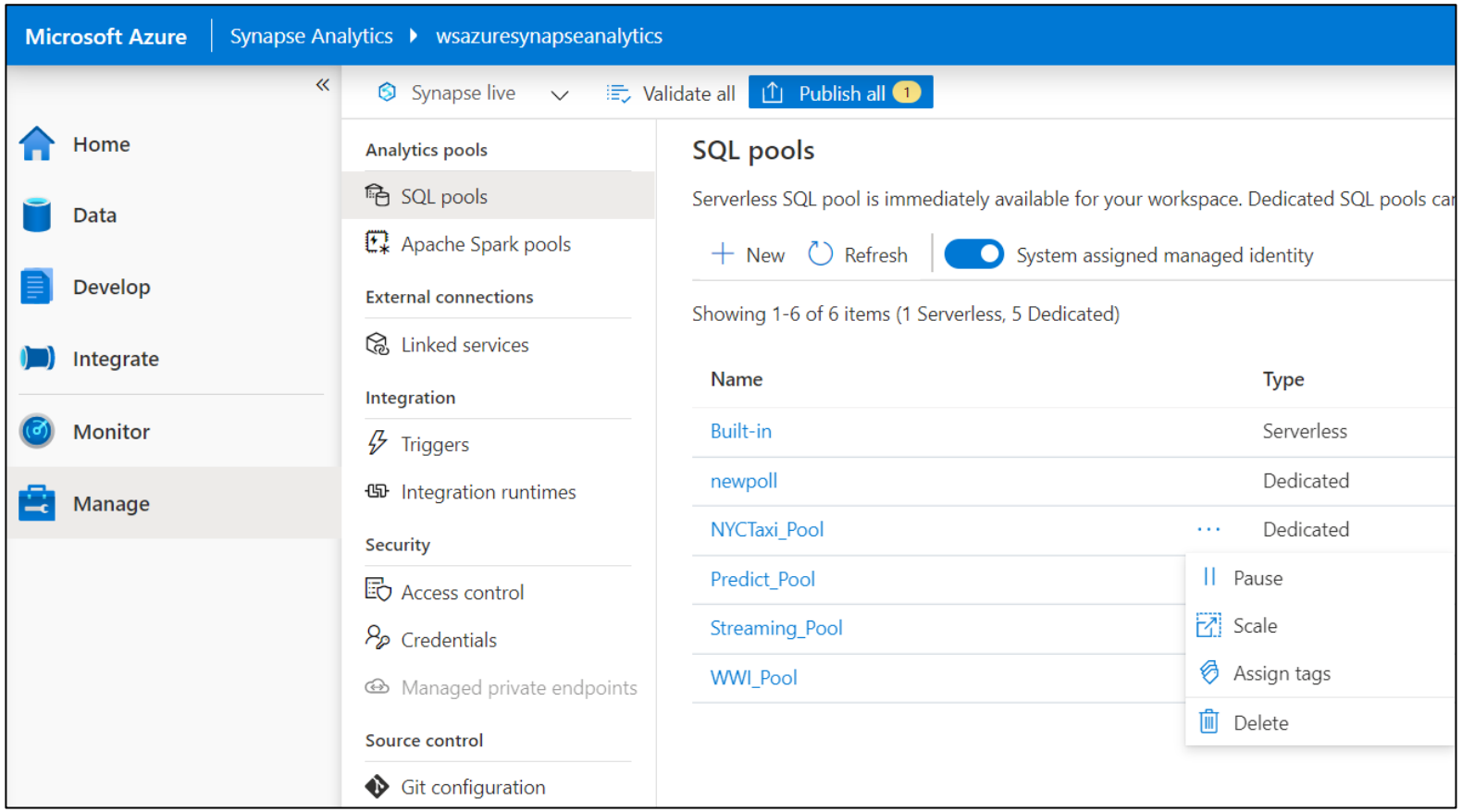
This feature provides ability to manage Analytics pools, Linked Services, Integration, Security and Source Control.

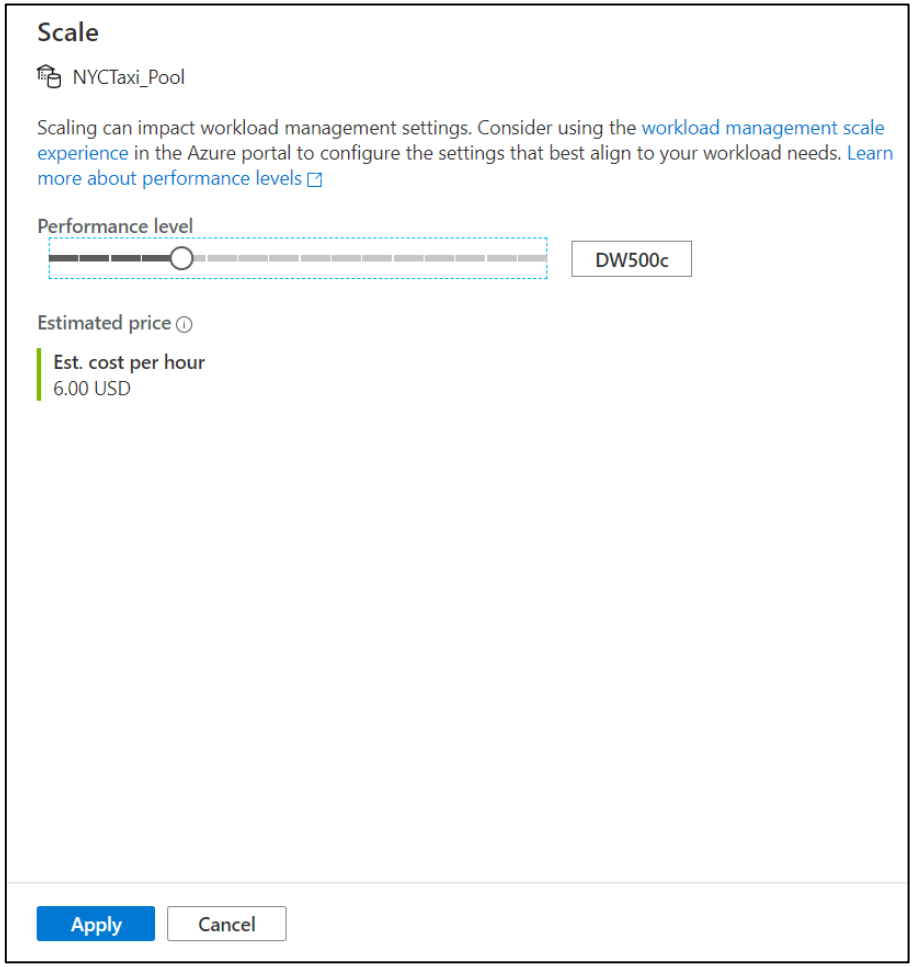


Manage – dedicated SQL pools

Overview

Provides ability to Pause and Resume, change Scale, Assign Tags from Studio.

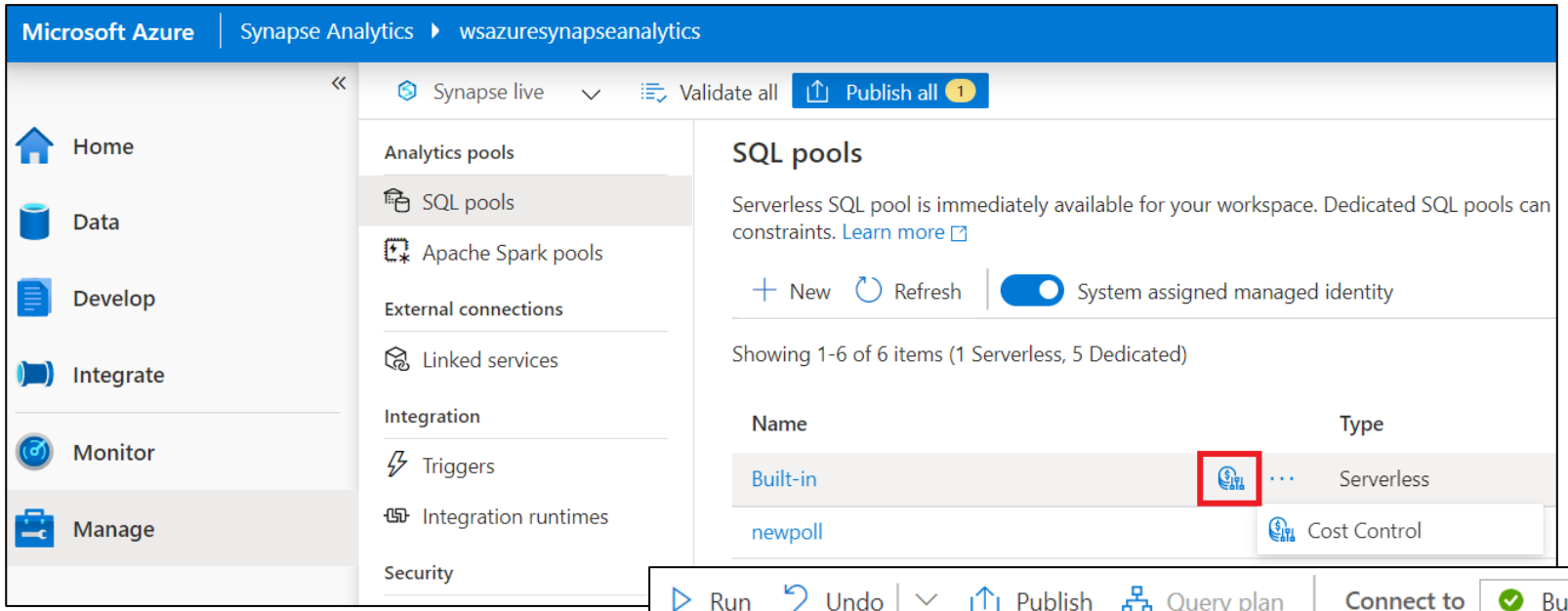


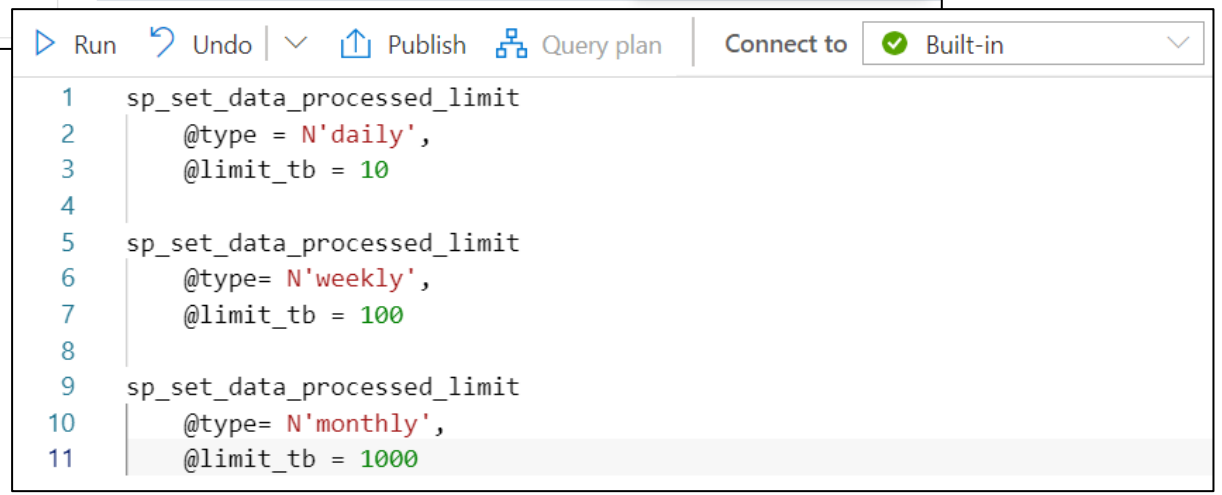


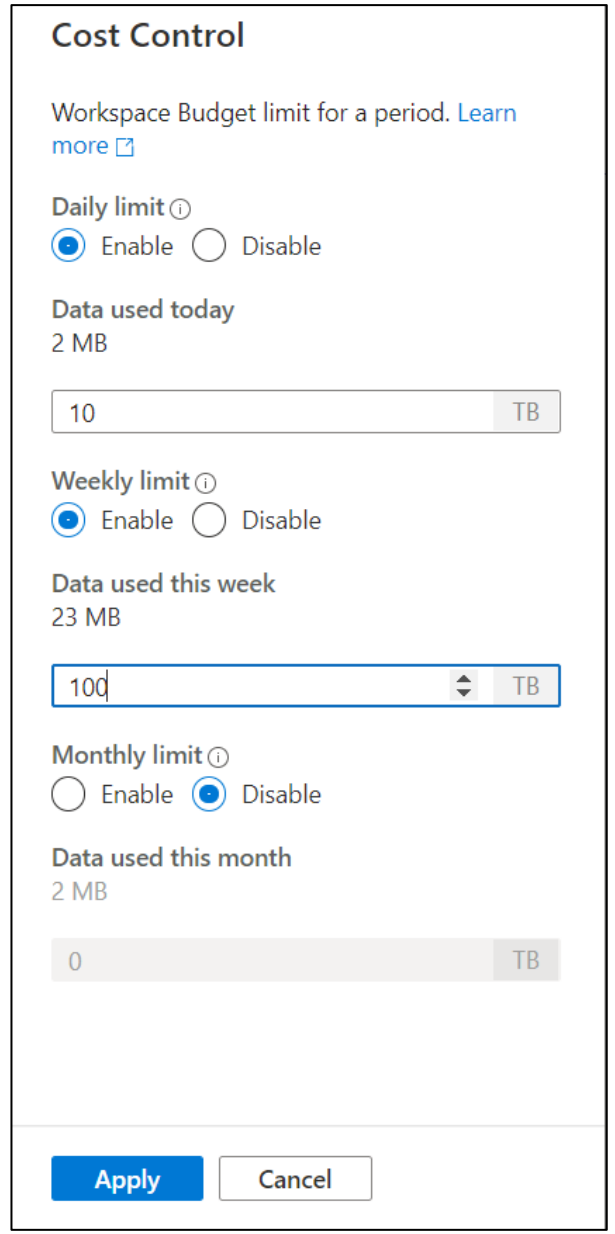
Manage – serverless SQL pools

Overview

It enables to set the budget for amount of data processed within Synapse Studio or T-SQL.



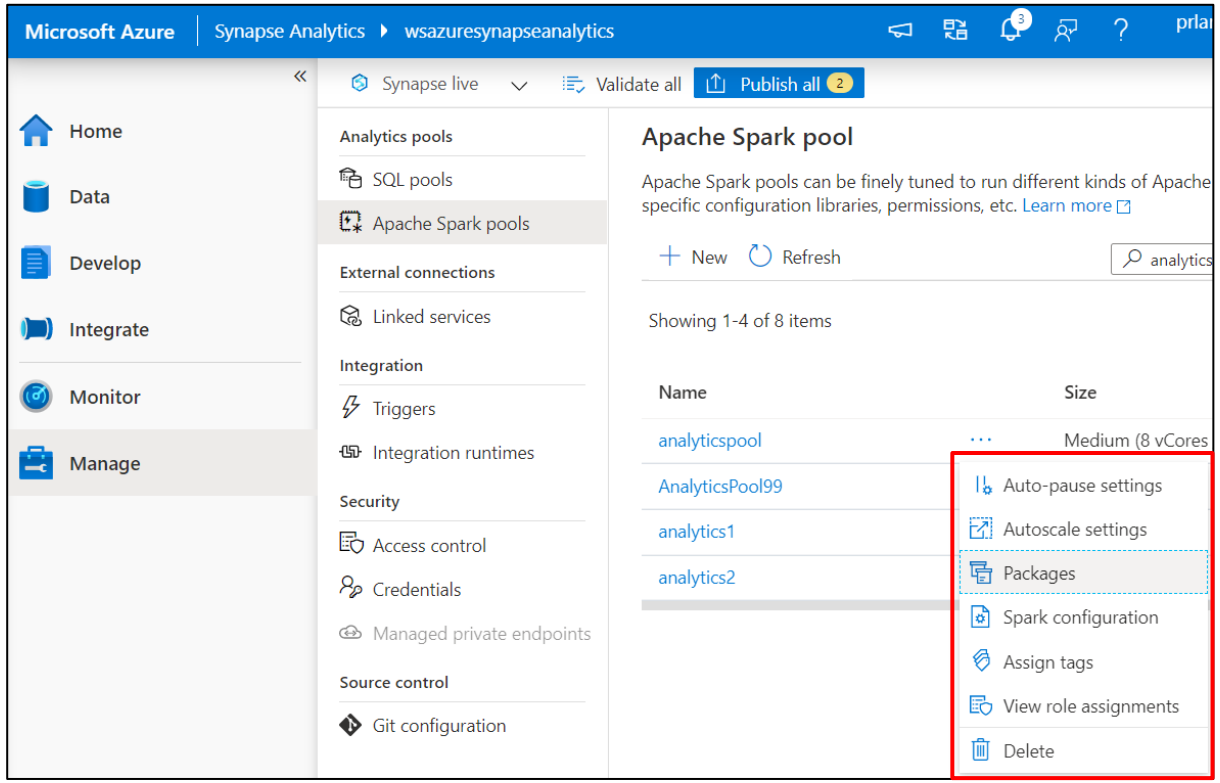


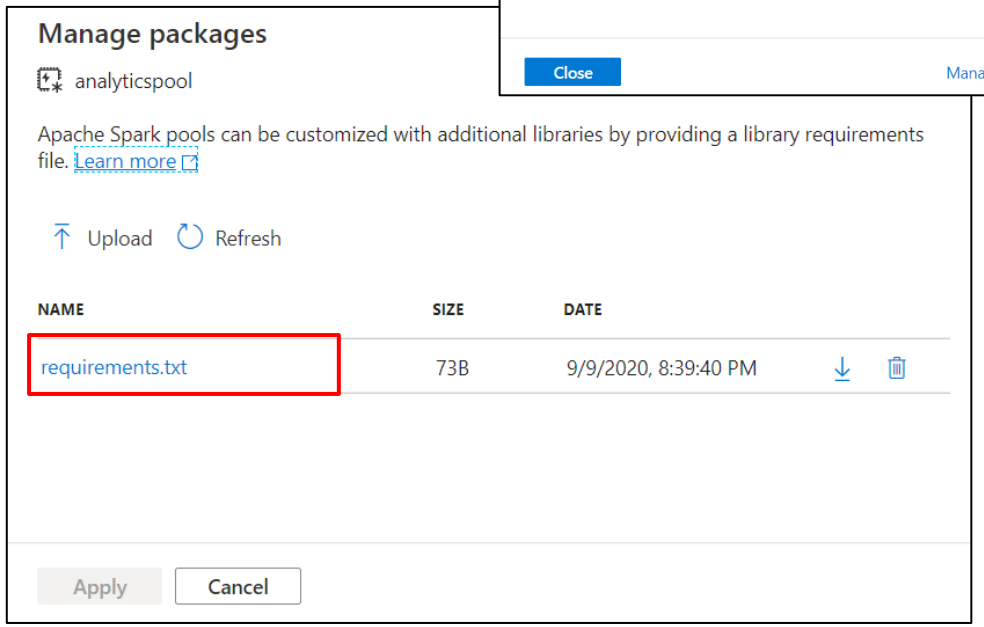


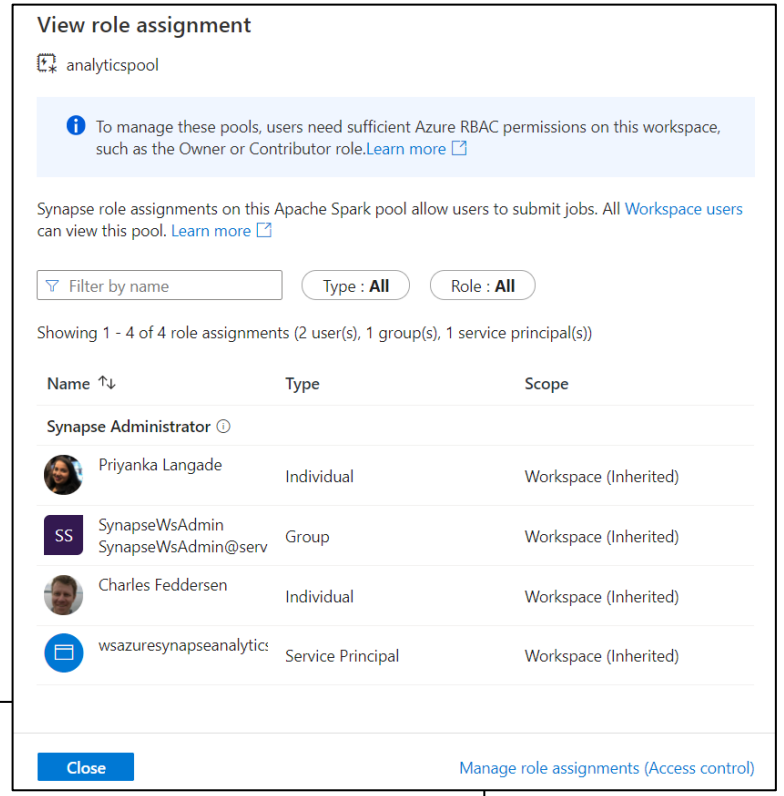
Manage – Apache Spark pools

Overview

Provides ability to Pause, Scale, Assign Tags, upload Spark configuration, packages from Studio and validate role assignments on respective Spark pool







Manage – Linked services

Overview

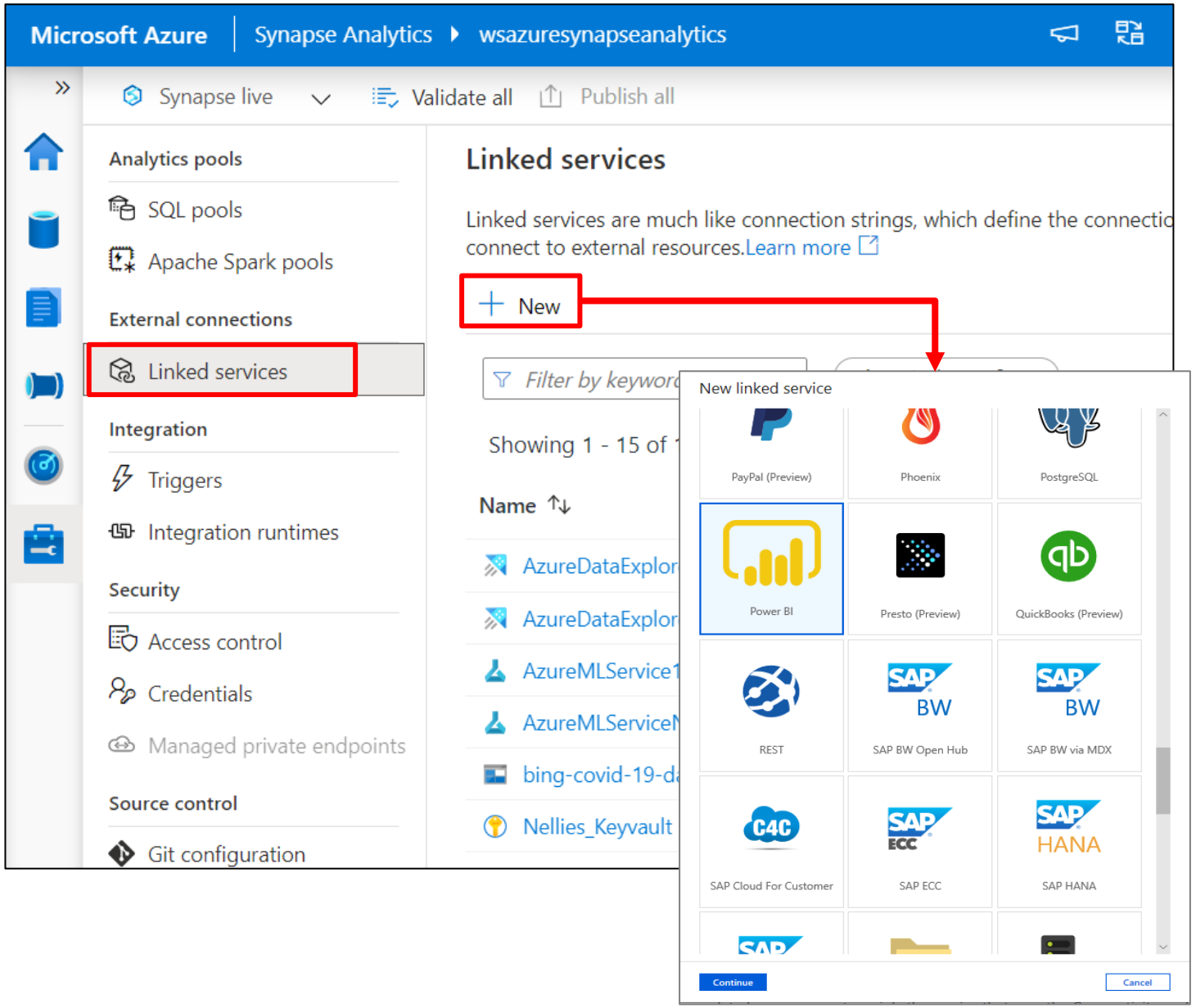
It defines the connection information needed to connect to external resources.

Benefits

Offers pre-build 90+ connectors

Easy cross platform data migration

Represents data store or compute resources



Manage – Triggers

Overview

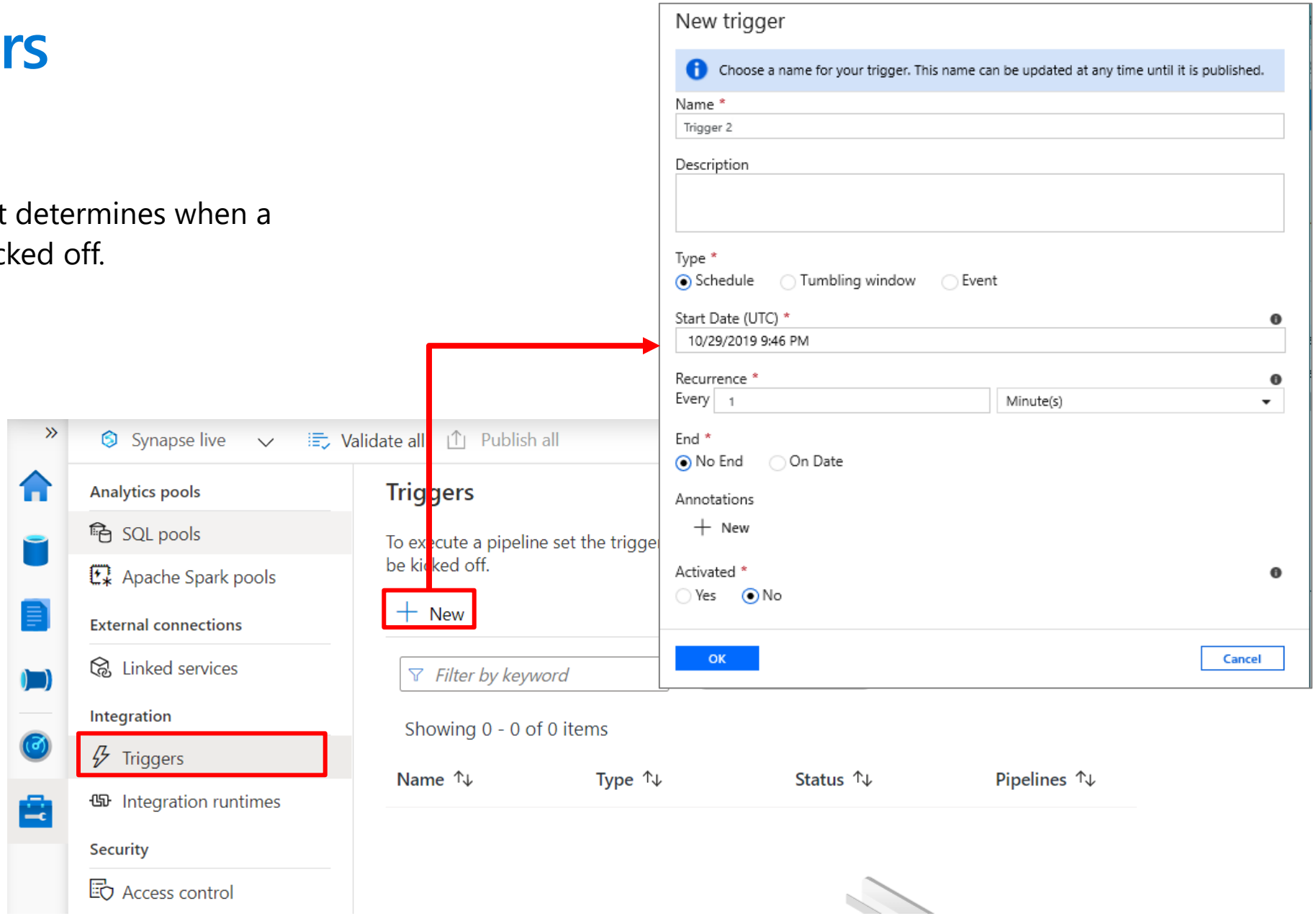
It defines a unit of processing that determines when a pipeline execution needs to be kicked off.

Benefits

Create and manage

- Schedule trigger
- Tumbling window trigger
- Event trigger

Control pipeline execution



Manage – Integration runtimes

Overview

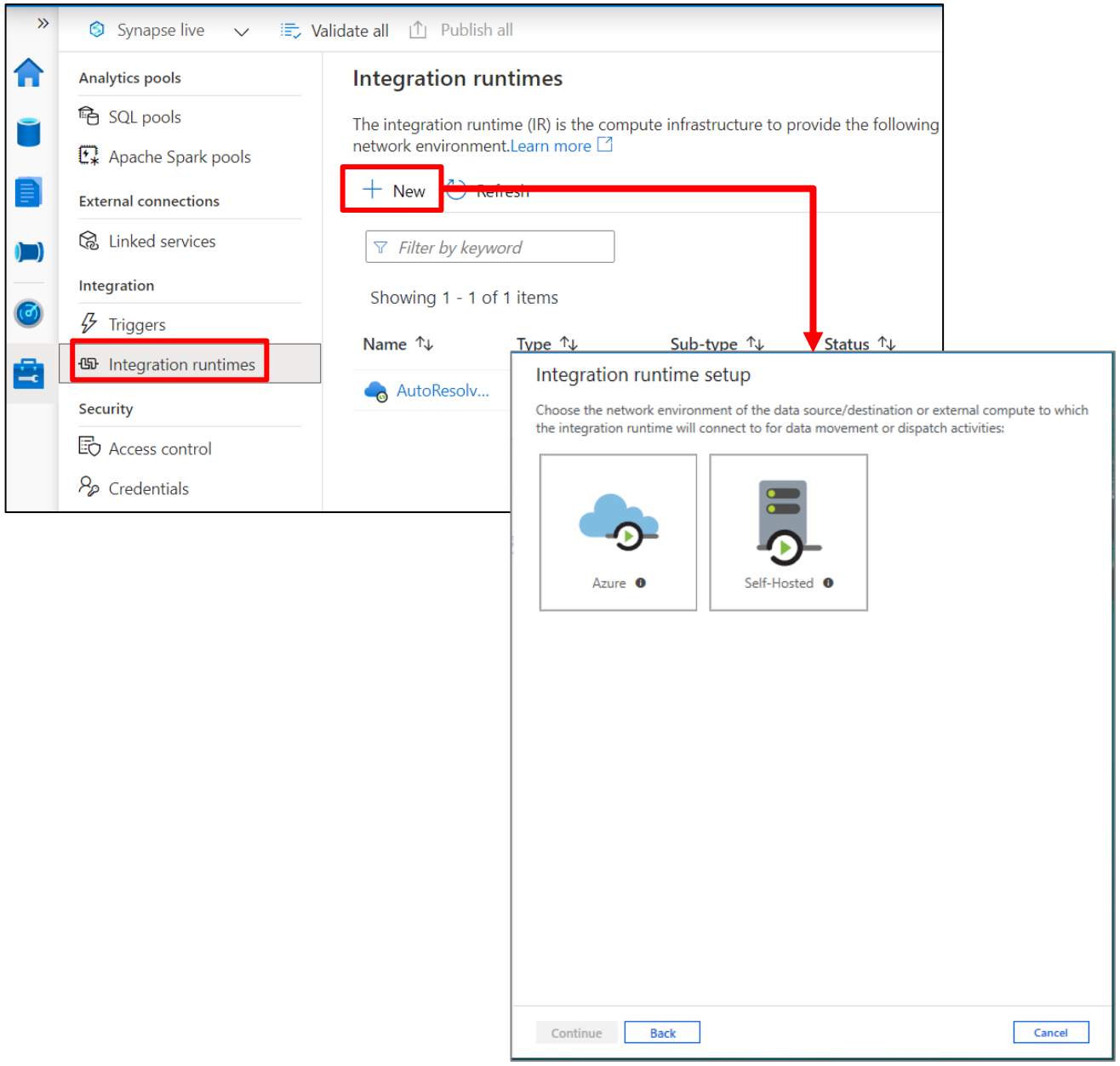
Integration runtimes are the compute infrastructure used by Pipelines to provide the data integration capabilities across different network environments. An integration runtime provides the bridge between the activity and linked services.

Benefits

Offers Azure Integration Runtime or Self-Hosted Integration Runtime

Azure Integration Runtime – provides fully managed, serverless compute in Azure

Self-Hosted Integration Runtime – use compute resources in on-premises machine or a VM inside private network



Manage – Access Control

Overview

It provides access control management to workspace resources and artifacts for admins

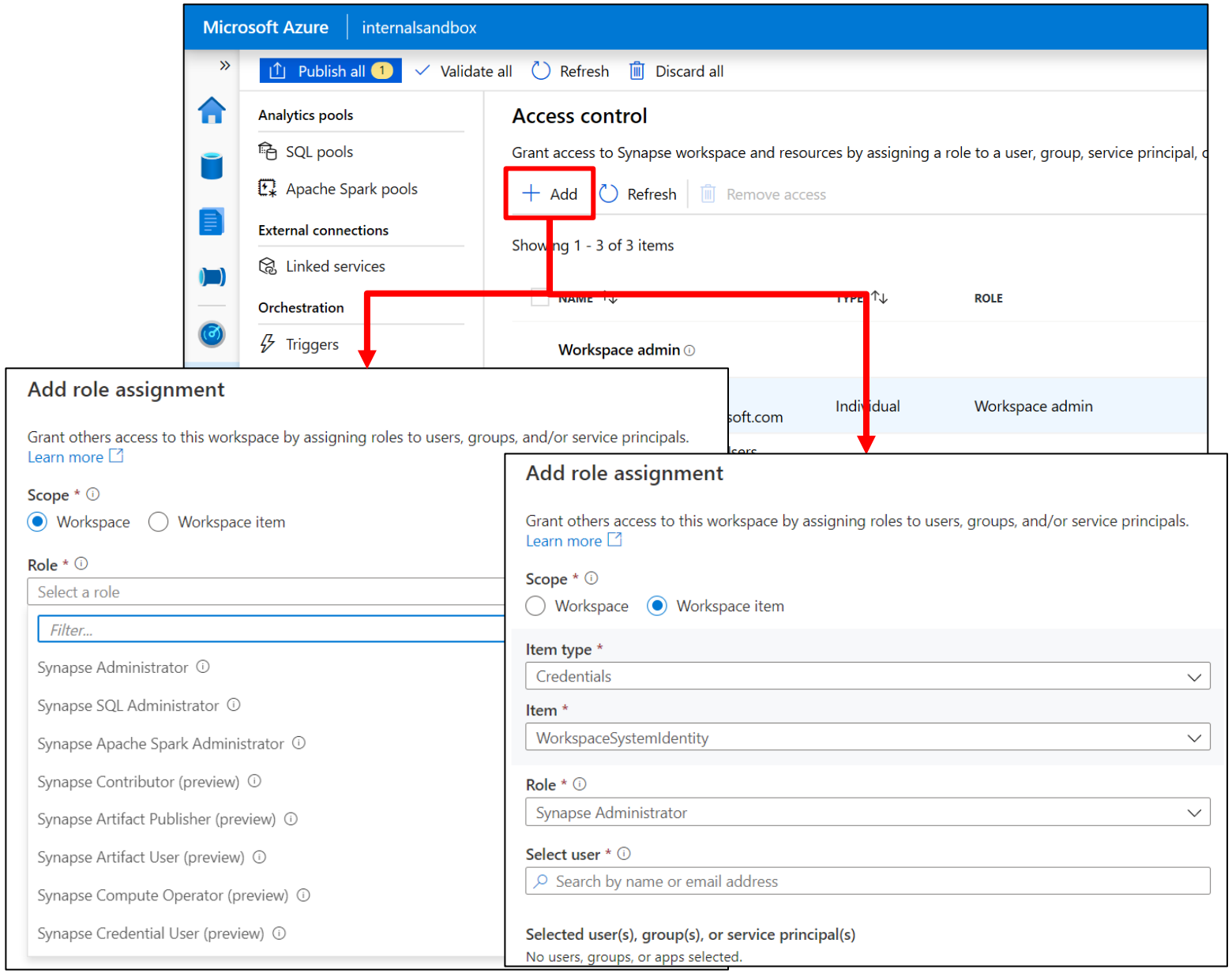
Benefits

Share workspace with the team

Increases productivity

Assign granular level permissions

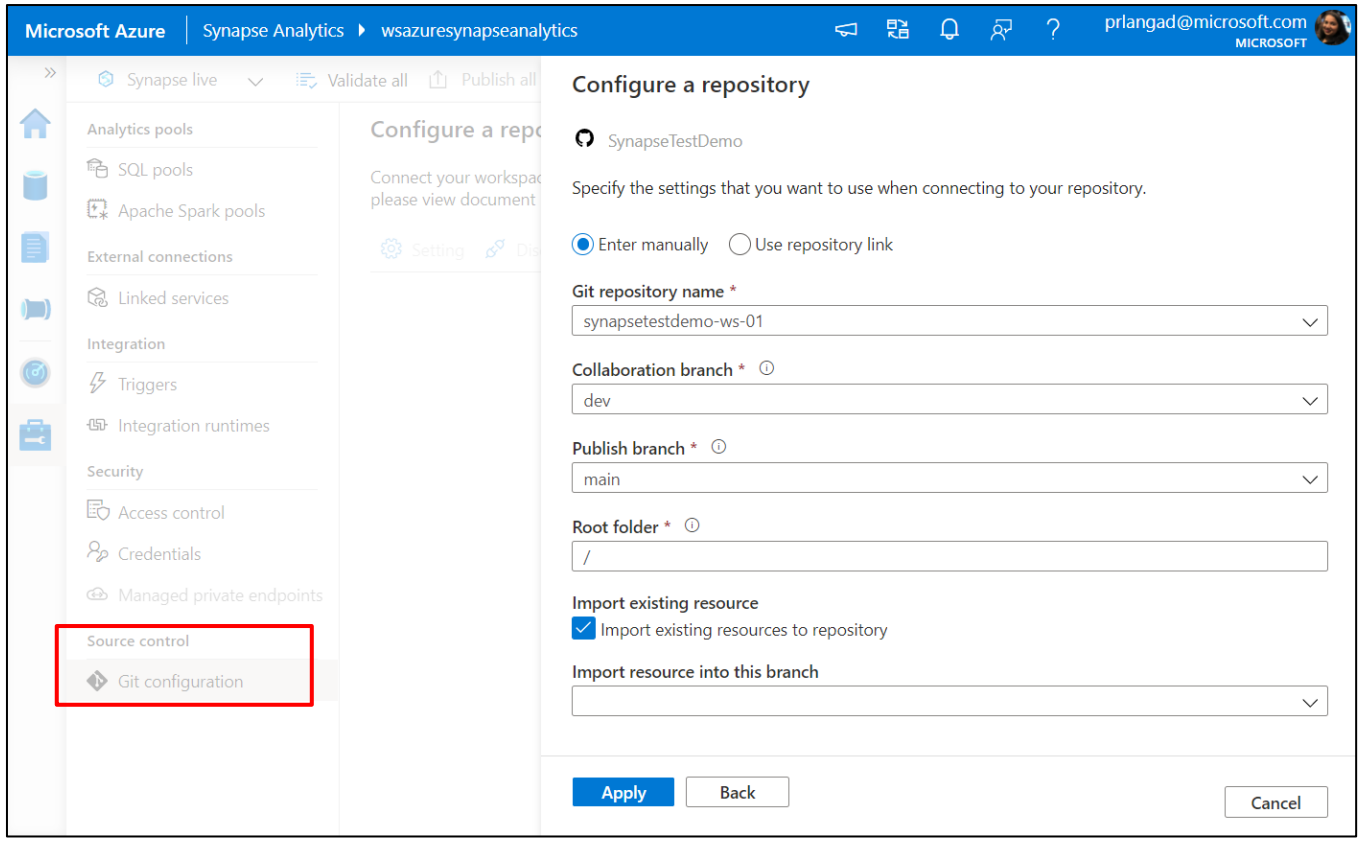
Manage permissions on Spark pools, Integration Runtimes, Linked services, Credentials

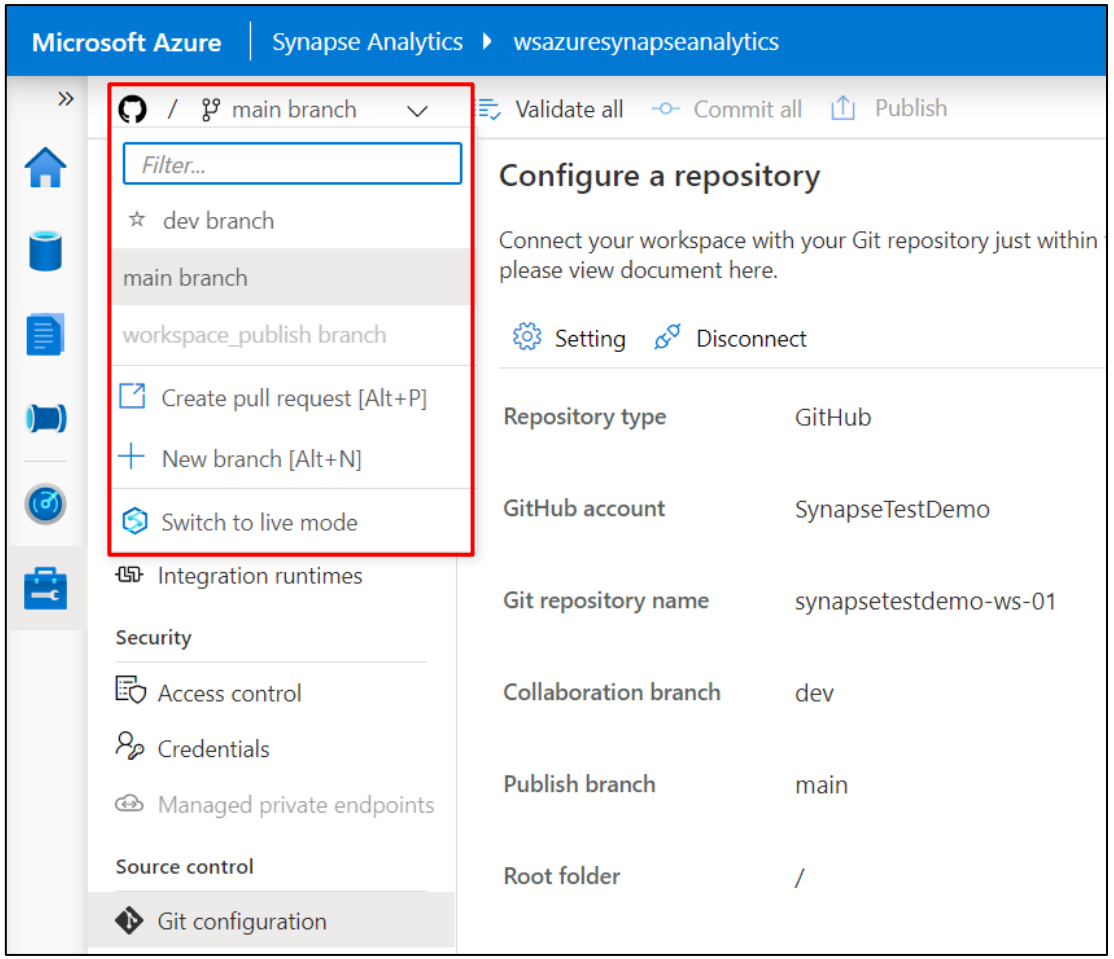


Manage – Source Control

Overview

Associate Synapse workspace with a Git repository, Azure DevOps, or GitHub



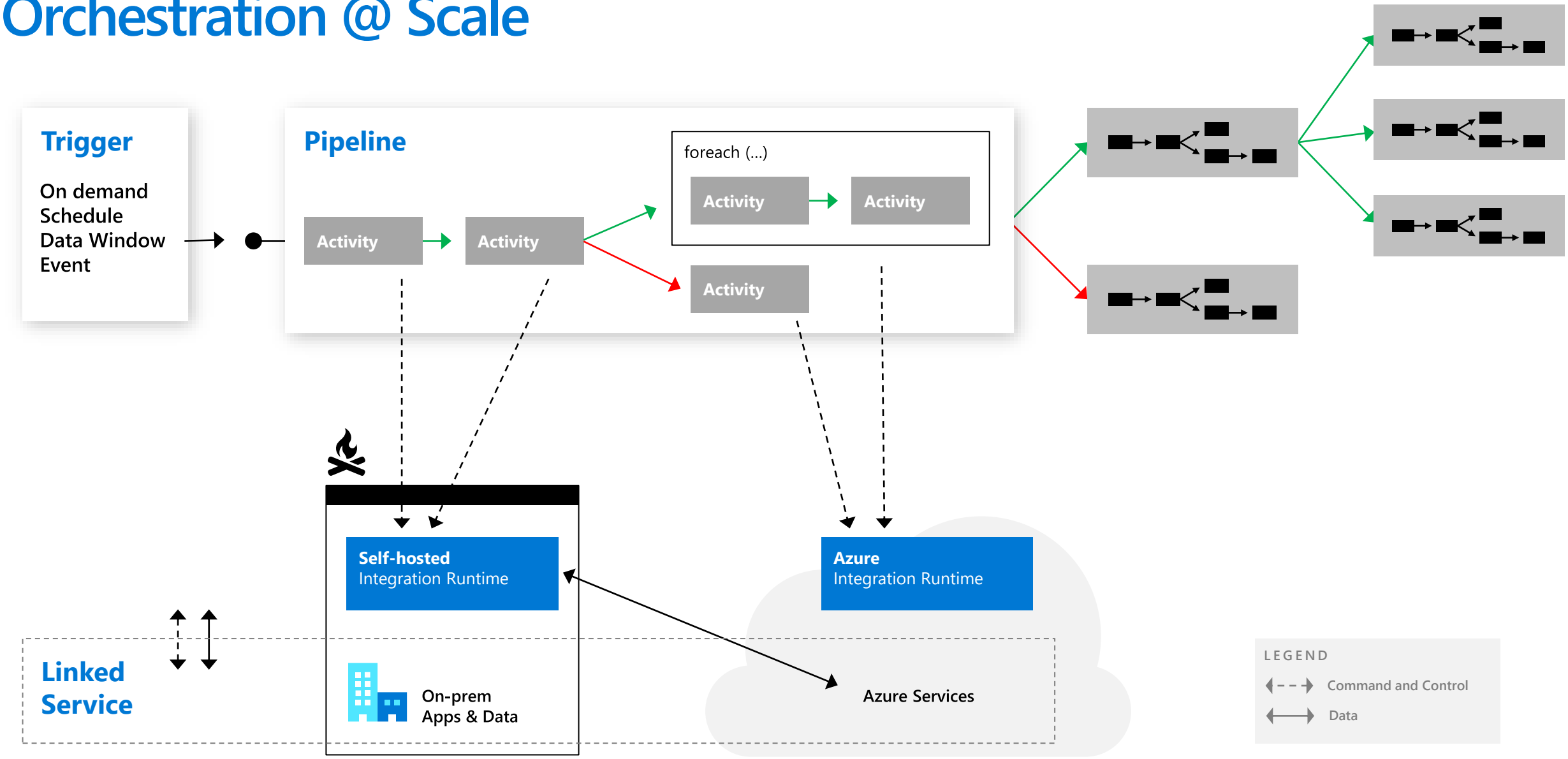


Microsoft

Azure Synapse Analytics

Integrate

Orchestration @ Scale



Data Movement

Scalable

per job elasticity

Up to 4 GB/s

Simple

Visually author or via code (Python, .NET, etc.)

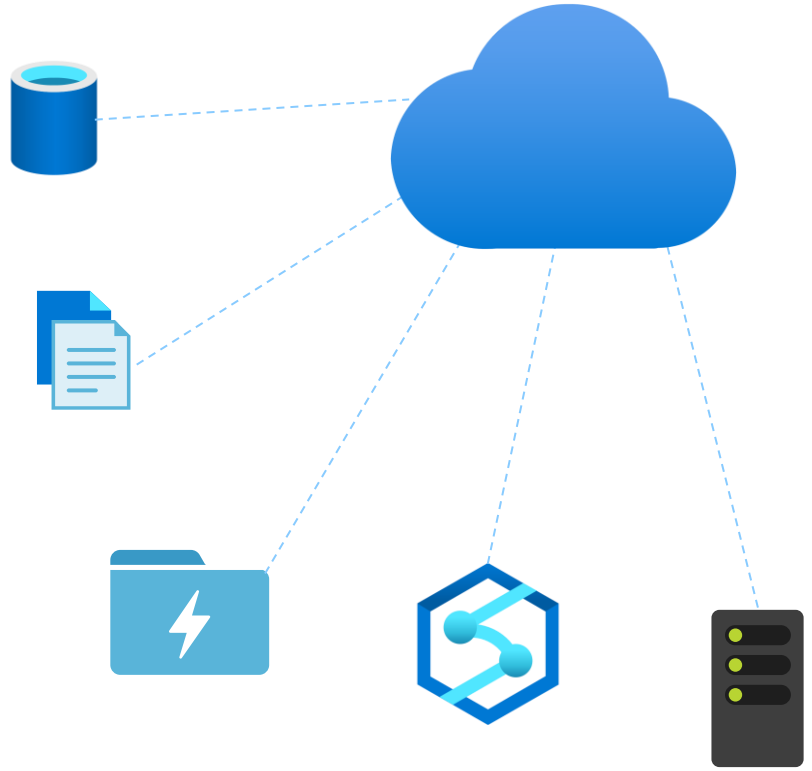
Serverless, no infrastructure to manage

Access all your data

90+ connectors provided and growing (cloud, on premises, SaaS)

Data Movement as a Service: 25 points of presence worldwide

Self-hostable Integration Runtime for hybrid movement



90+ Connectors out of the box

Azure (15)	Database & DW (26)		File Storage (6)	File Formats(6)	NoSQL (3)	Services and App (28)		Generic (4)
Blob storage	Amazon Redshift	Oracle	Amazon S3	AVRO	Cassandra	Amazon MWS	Oracle Service Cloud	Generic HTTP
Cosmos DB - SQL API	DB2	Phoenix	File system	Binary	Couchbase	CDS for Apps	PayPal	Generic OData
Cosmos DB - MongoDB API	Drill	PostgreSQL	FTP	Delimited Text	MongoDB	Concur	QuickBooks	Generic ODBC
Data Explorer	Google BigQuery	Presto	Google Cloud Storage	JSON		Dynamics 365	Salesforce	Generic REST
Data Lake Storage Gen1	Greenplum	SAP BW Open Hub	HDFS	ORC		Dynamics AX	SF Service Cloud	
Data Lake Storage Gen2	HBase	SAP BW via MDX	SFTP	Parquet		Dynamics CRM	SF Marketing Cloud	
Database for MariaDB	Hive	SAP HANA				Google AdWords	SAP C4C	
Database for MySQL	Apache Impala	SAP table				HubSpot	SAP ECC	
Database for PostgreSQL	Informix	Spark				Jira	ServiceNow	
File Storage	MariaDB	SQL Server				Magento	Shopify	
SQL Database	Microsoft Access	Sybase				Marketo	Square	
SQL Database MI	MySQL	Teradata				Office 365	Web table	
SQL Data Warehouse	Netezza	Vertica				Oracle Eloqua	Xero	
Search index						Oracle Responsys	Zoho	
Table storage								

Pipelines

Overview

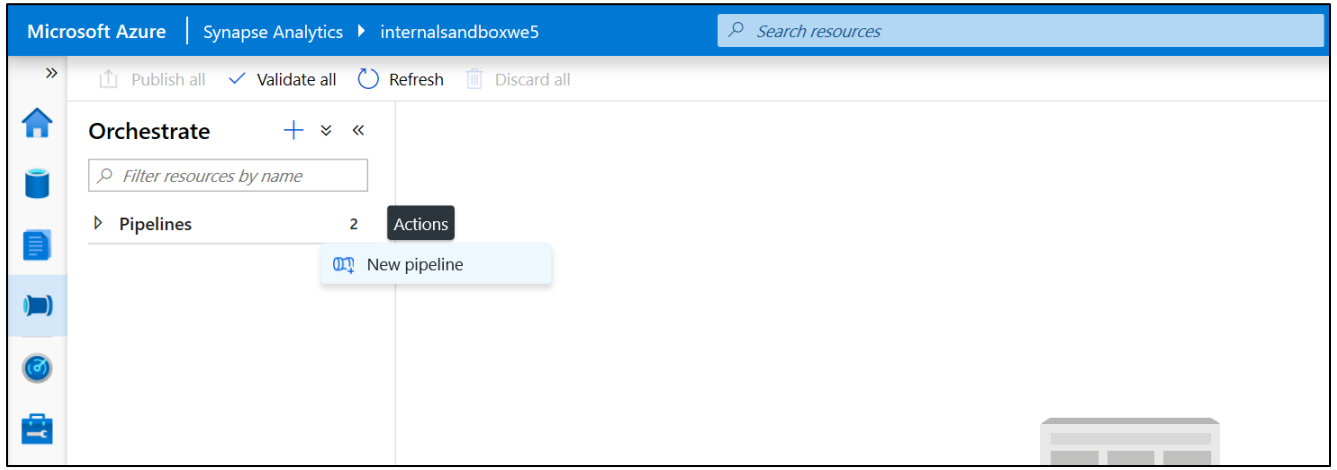
It provides ability to load data from storage account to desired linked service. Load data by manual execution of pipeline or by orchestration

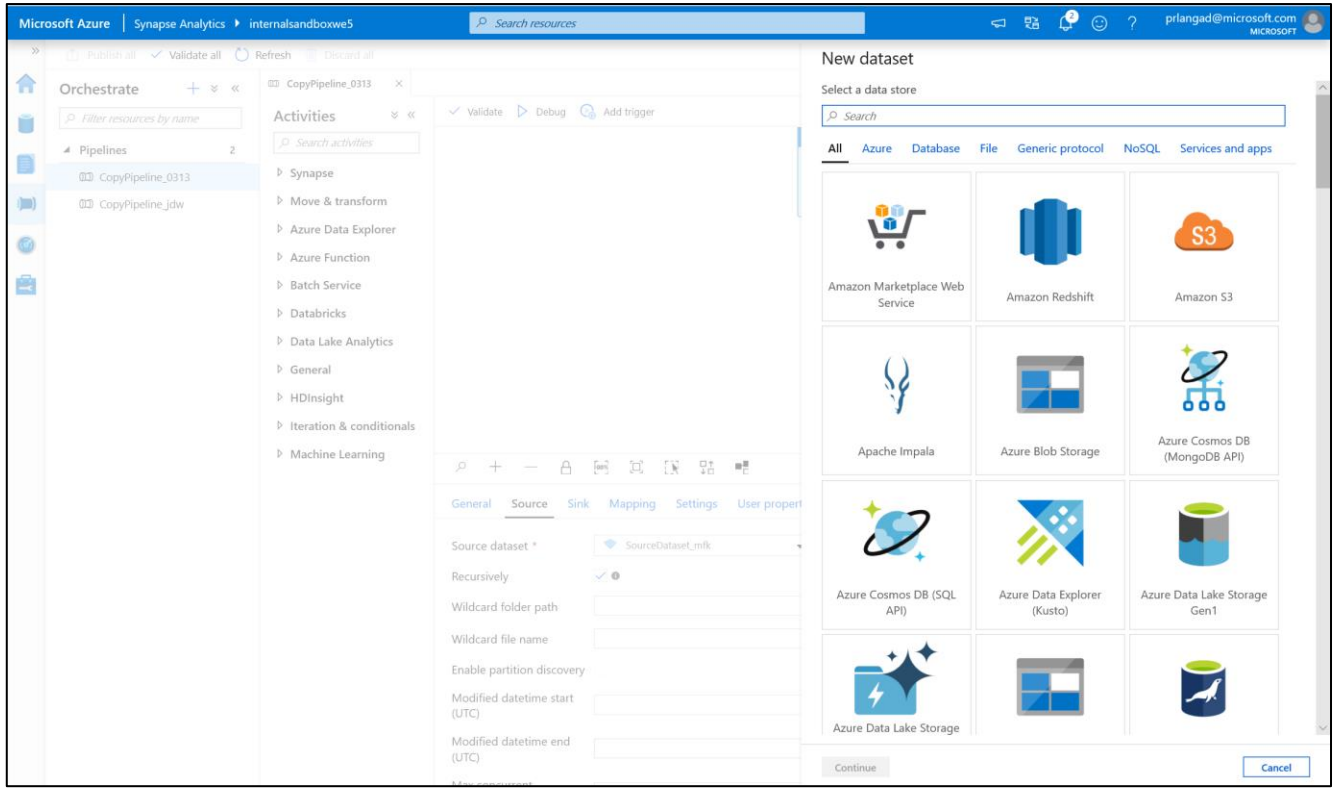
Benefits

Supports common loading patterns

Fully parallel loading into data lake or SQL tables

Graphical development experience





Prep & Transform Data

Overview

It offers data cleansing, transformation, aggregation, conversion, etc.

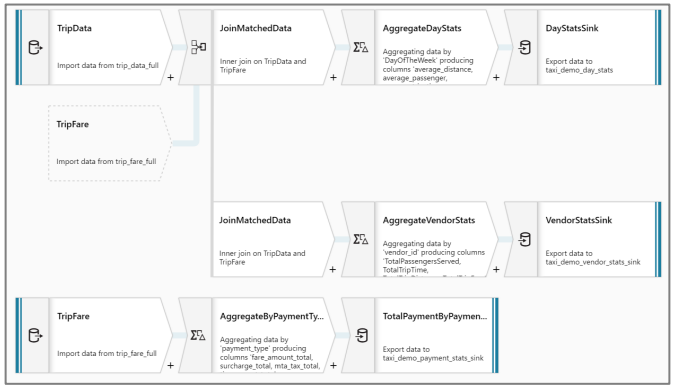
Benefits

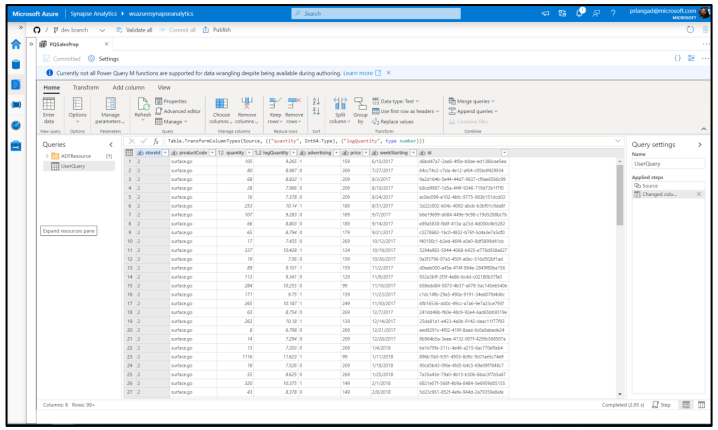
Cloud scale via Spark execution

Guided experience to easily build resilient data flows

Flexibility to transform data per user's comfort

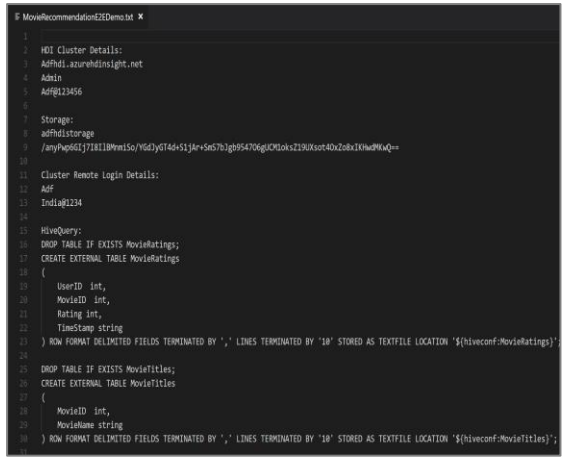
Monitor and manage dataflows from a single pane of glass





...

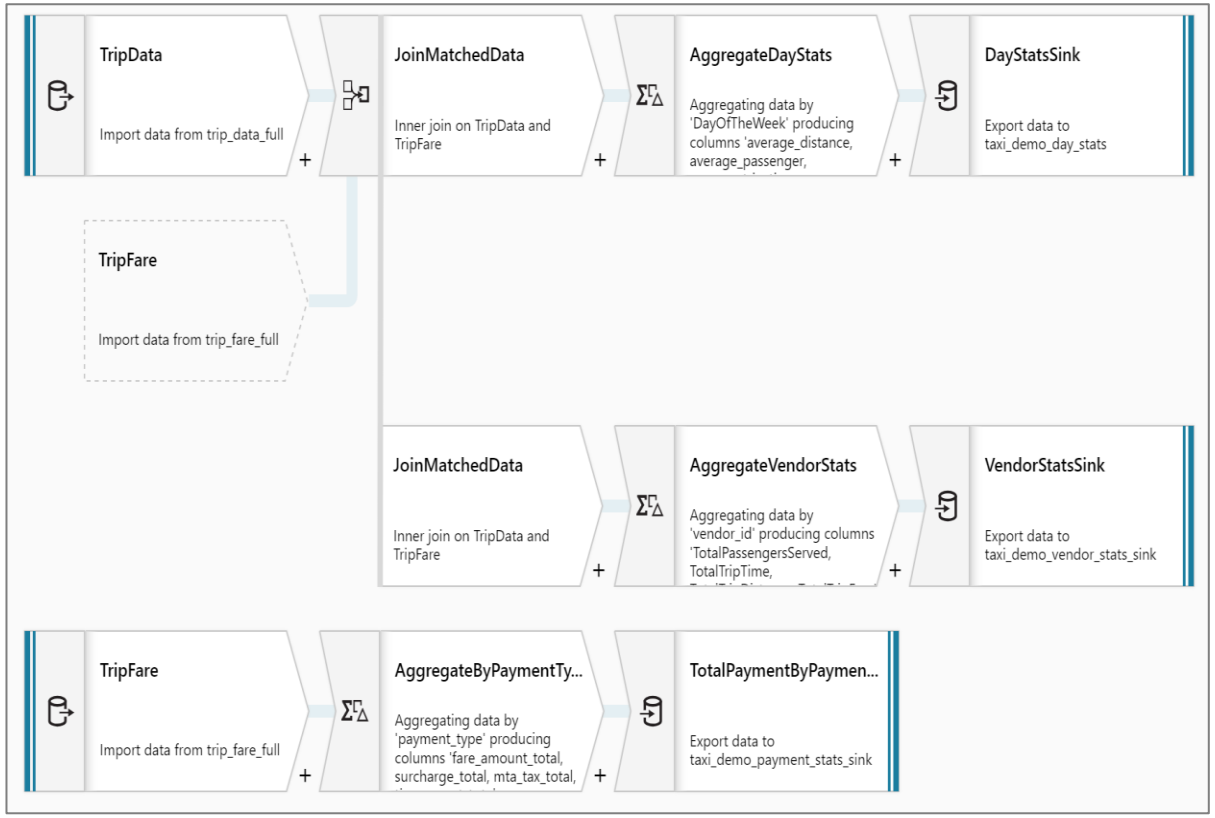
not



Prep & Transform Data

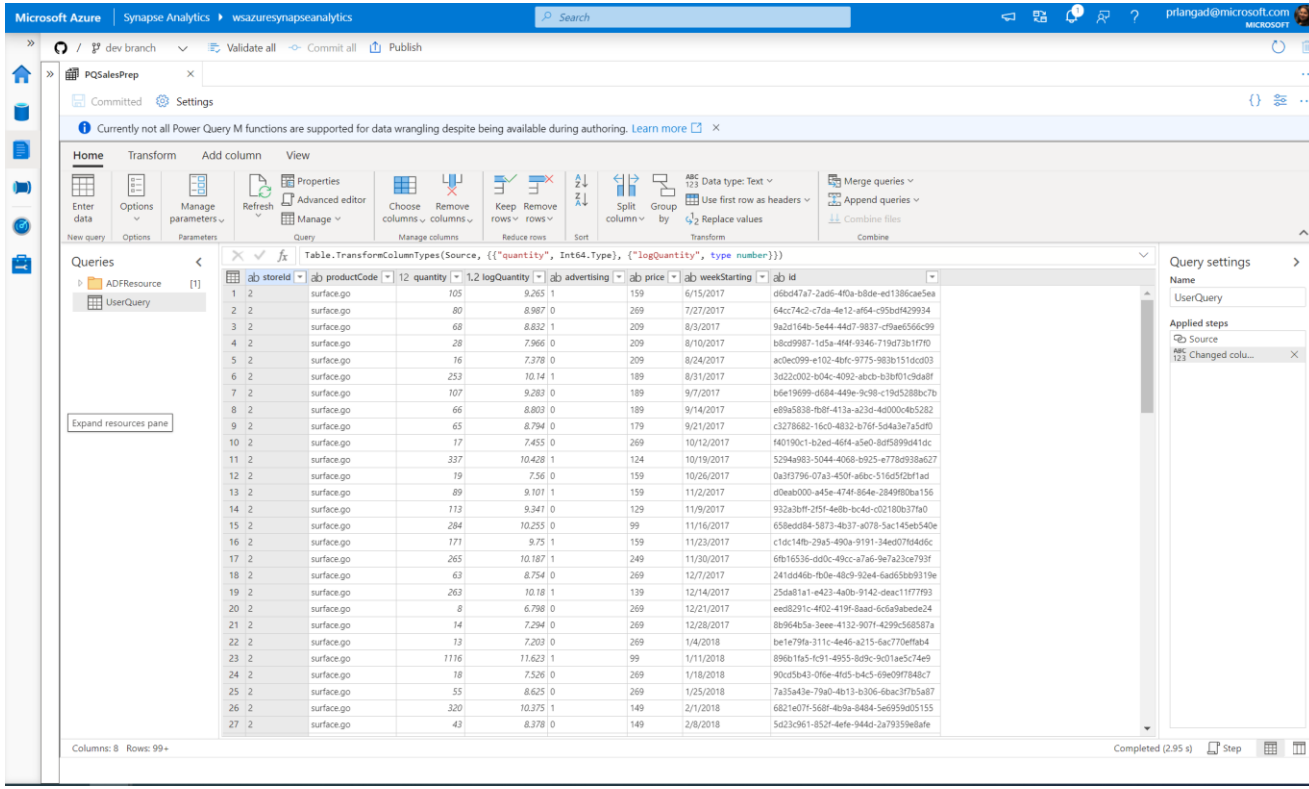
Mapping Dataflow

Code free data transformation @scale



Power Query

Code free data preparation @scale



Triggers

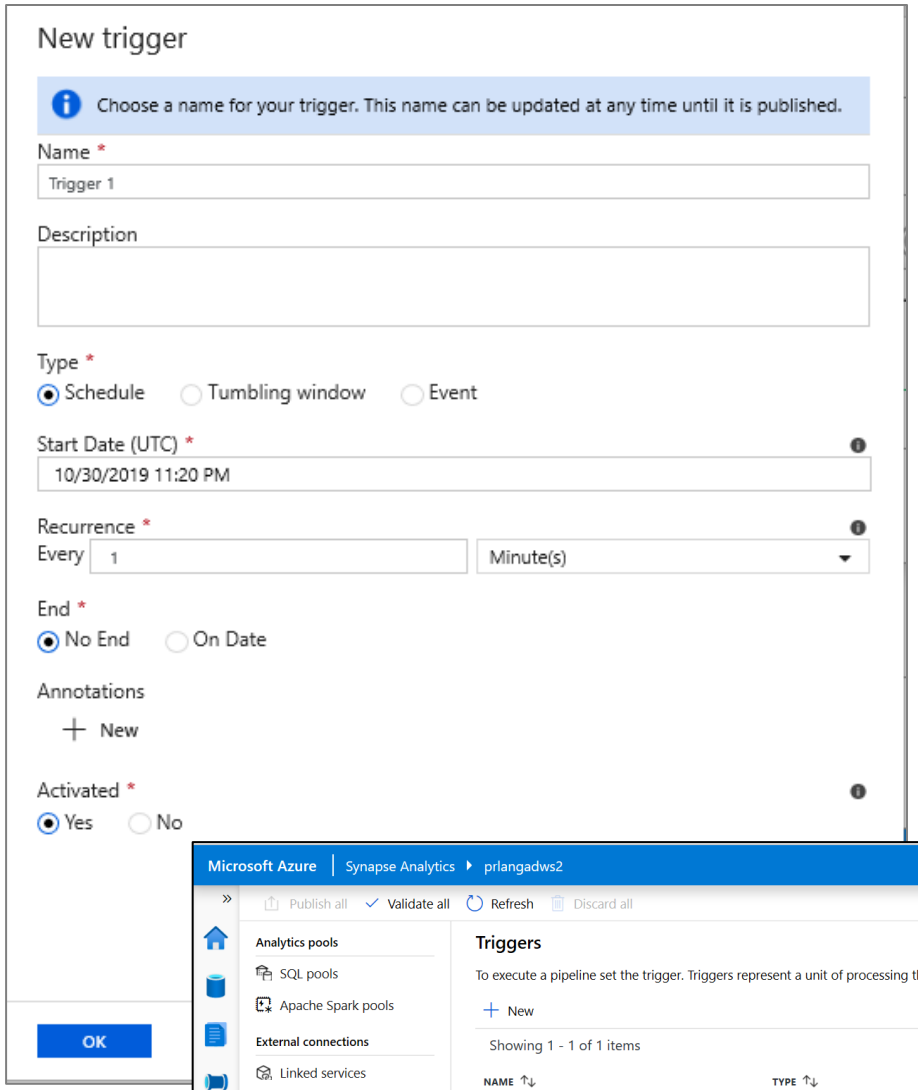
Overview

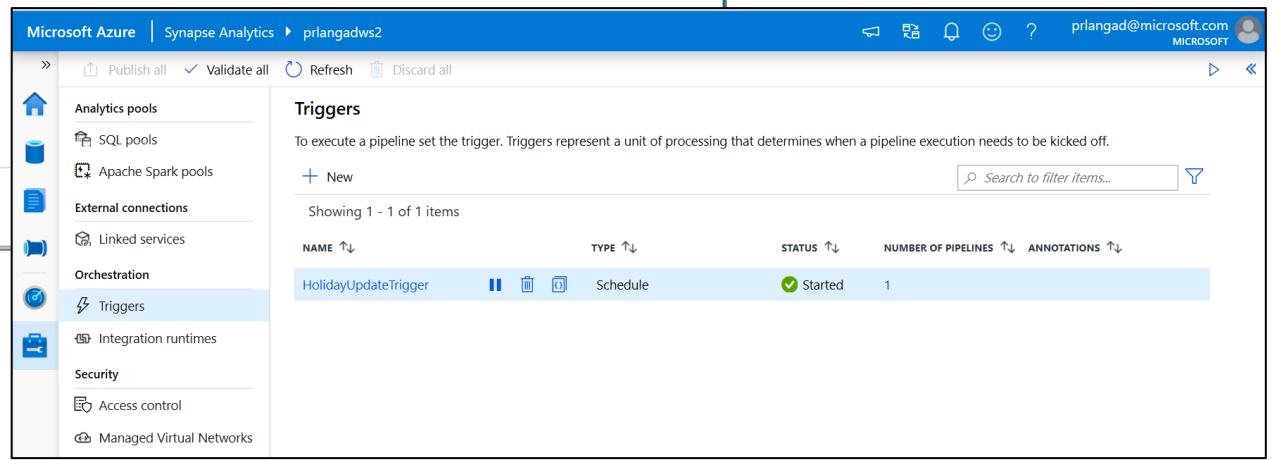
Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off.

Data Integration offers 3 trigger types as –

1. Schedule – gets fired at a schedule with information of start date, recurrence, end date
2. Event – gets fired on specified event
3. Tumbling window – gets fired at a periodic time interval from a specified start date, while retaining state

It also provides ability to monitor pipeline runs and control trigger execution.





Microsoft

Azure Synapse Analytics

Synapse SQL

Key features

Rich surface area

- T-SQL language for data analytics
- Supporting large number of languages and tools
- Enterprise-grade security

dedicated SQL pool

- Modern Data Warehouse
- Indexing and caching
- Import and query external data
- Workload management

serverless SQL pool

- Querying external data
- Model raw files as virtual tables and views
- Easy data transformation

Synapse SQL - clients and tools

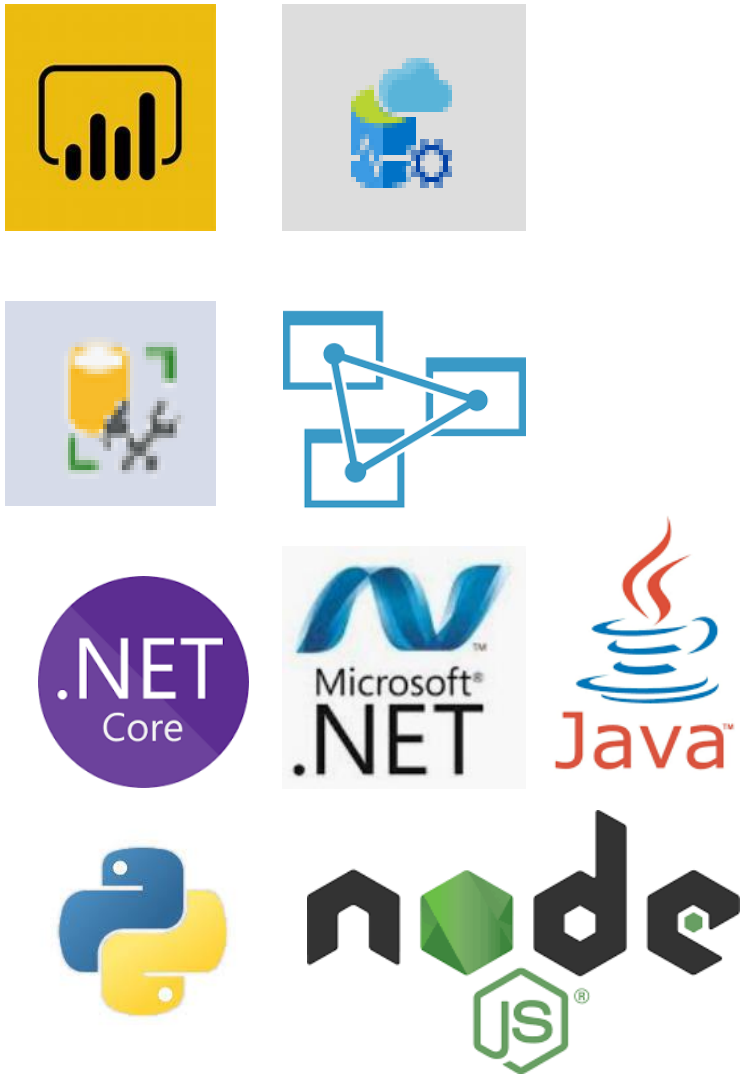
Tools

Web IDE - Synapse Studio

Client tools - Azure Data Studio, SSMS,

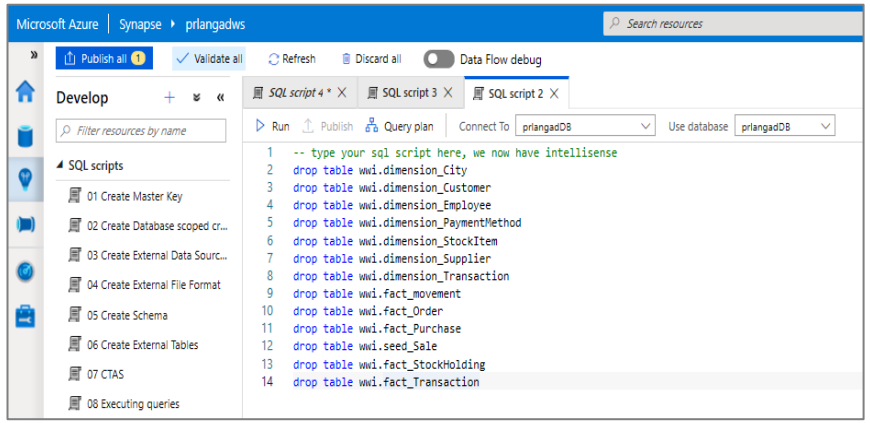
Any tool/library that uses standard SQL can access serverless SQL pool

- PowerBI
- Azure Analytic Services
- Client languages and drivers that works with Azure SQL can be used to access serverless SQL pool

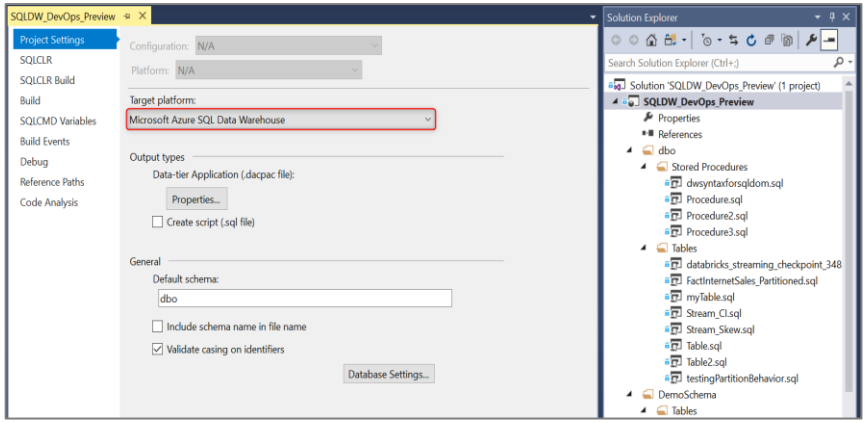


Developer Tools

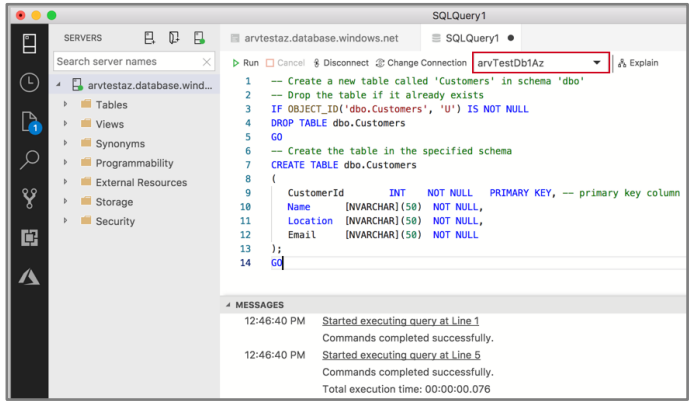
Azure Synapse Analytics



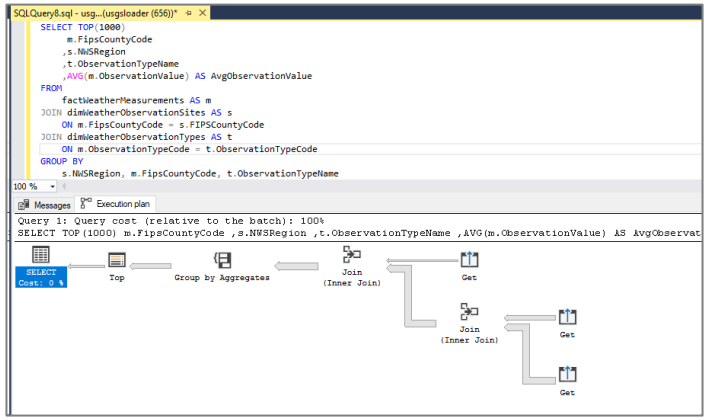
Visual Studio - SSDT database projects



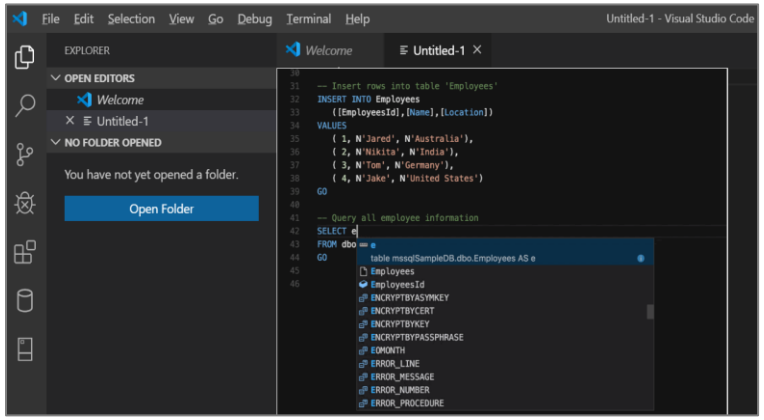
Azure Data Studio (queries, extensions etc.)



SQL Server Management Studio (queries, execution plans etc.)



Visual Studio Code

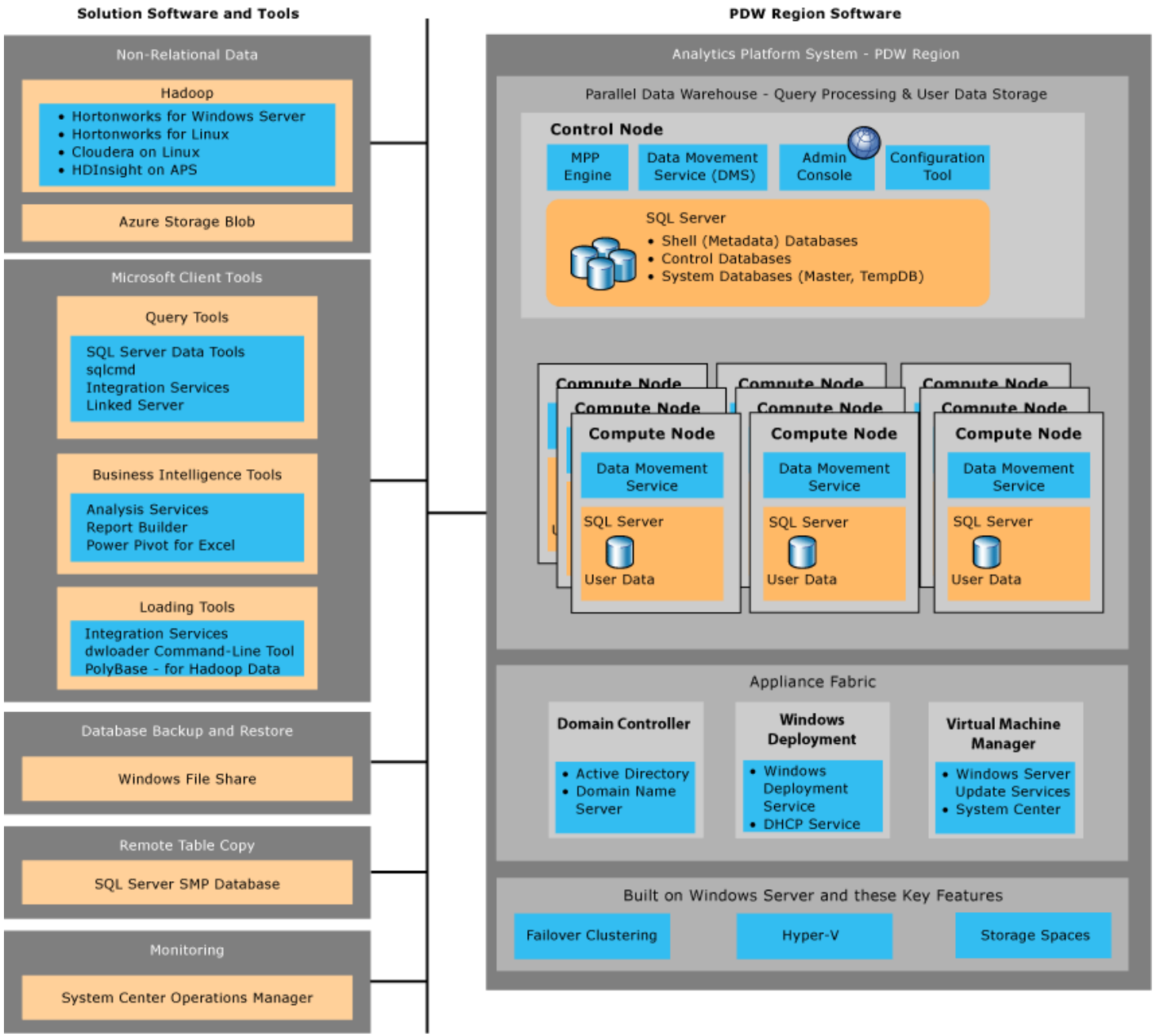


Microsoft

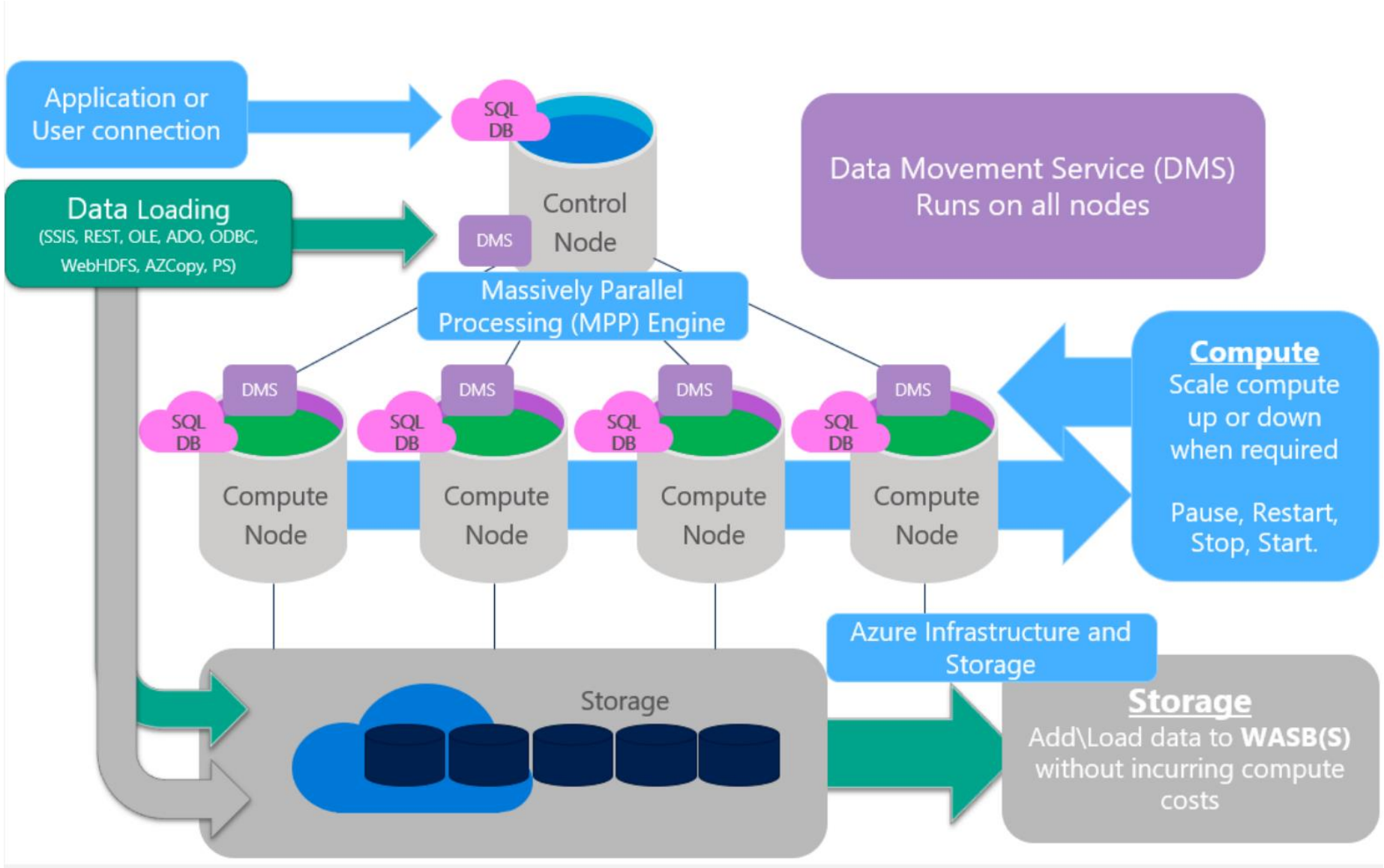
Azure Synapse Analytics

Synapse dedicated SQL pool

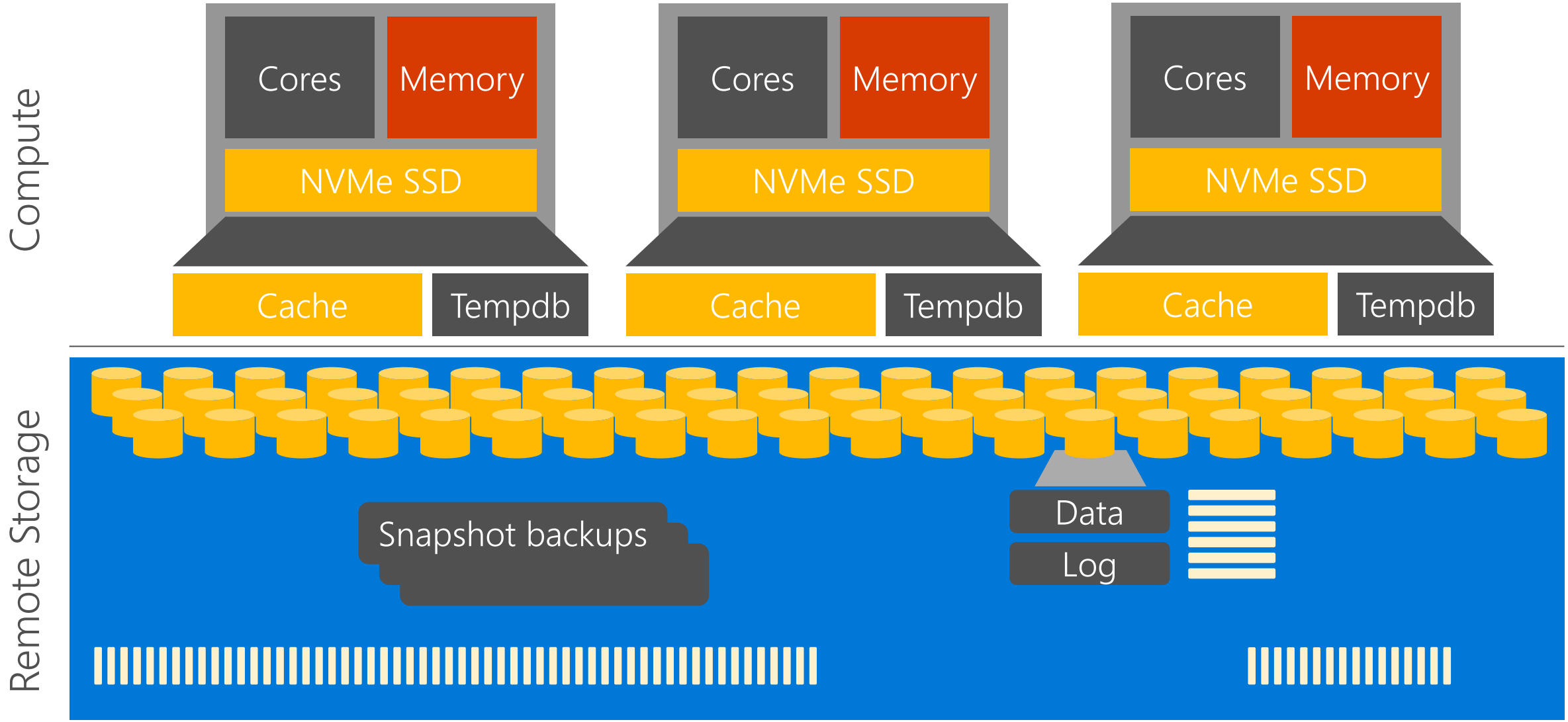
PDW (Parallel Data Warehouse) Architecture



Synapse SQL Pool Architecture

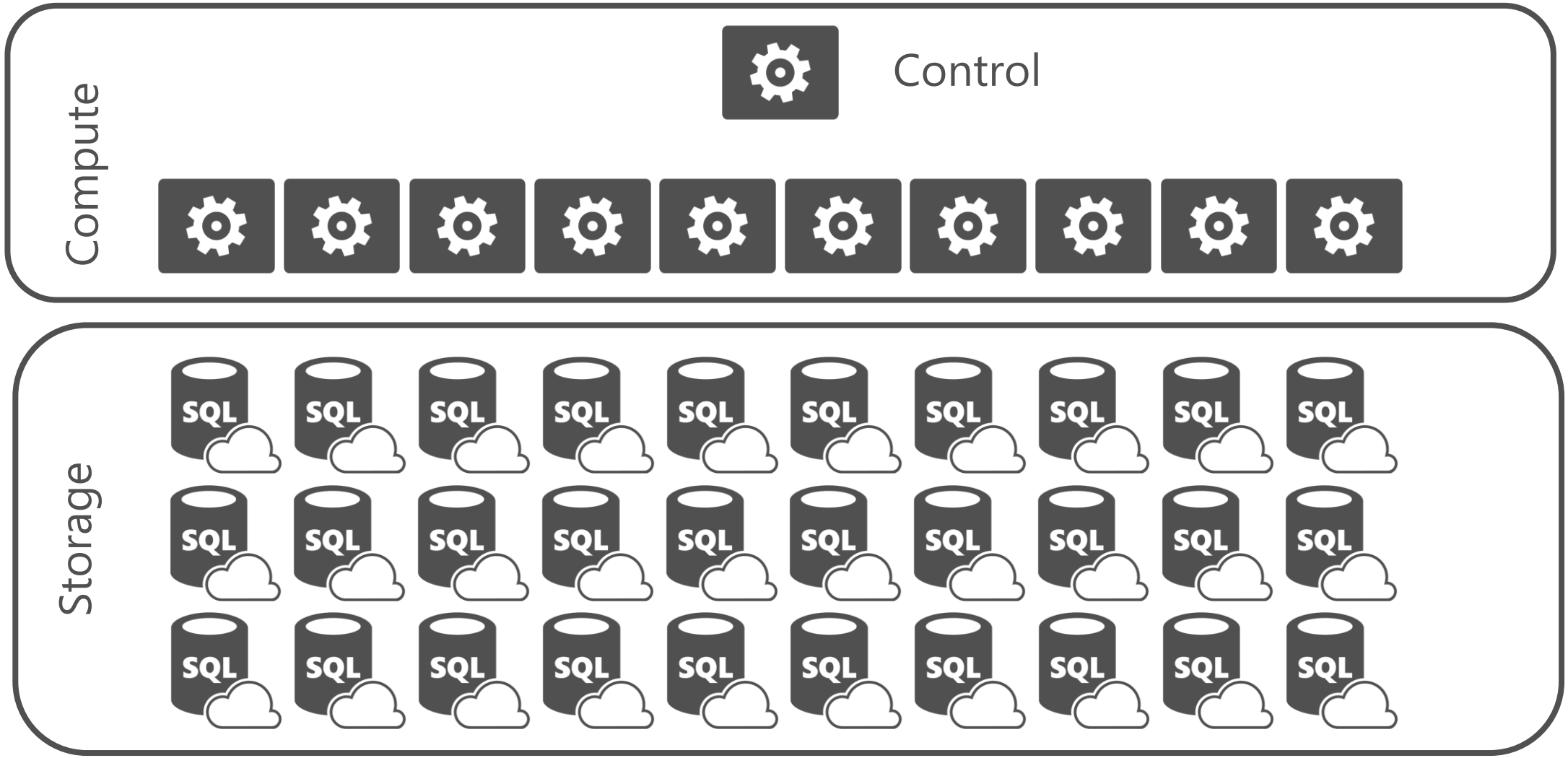


TIERED STORAGE MODEL AUTOMATES DATA TEMPERATURE

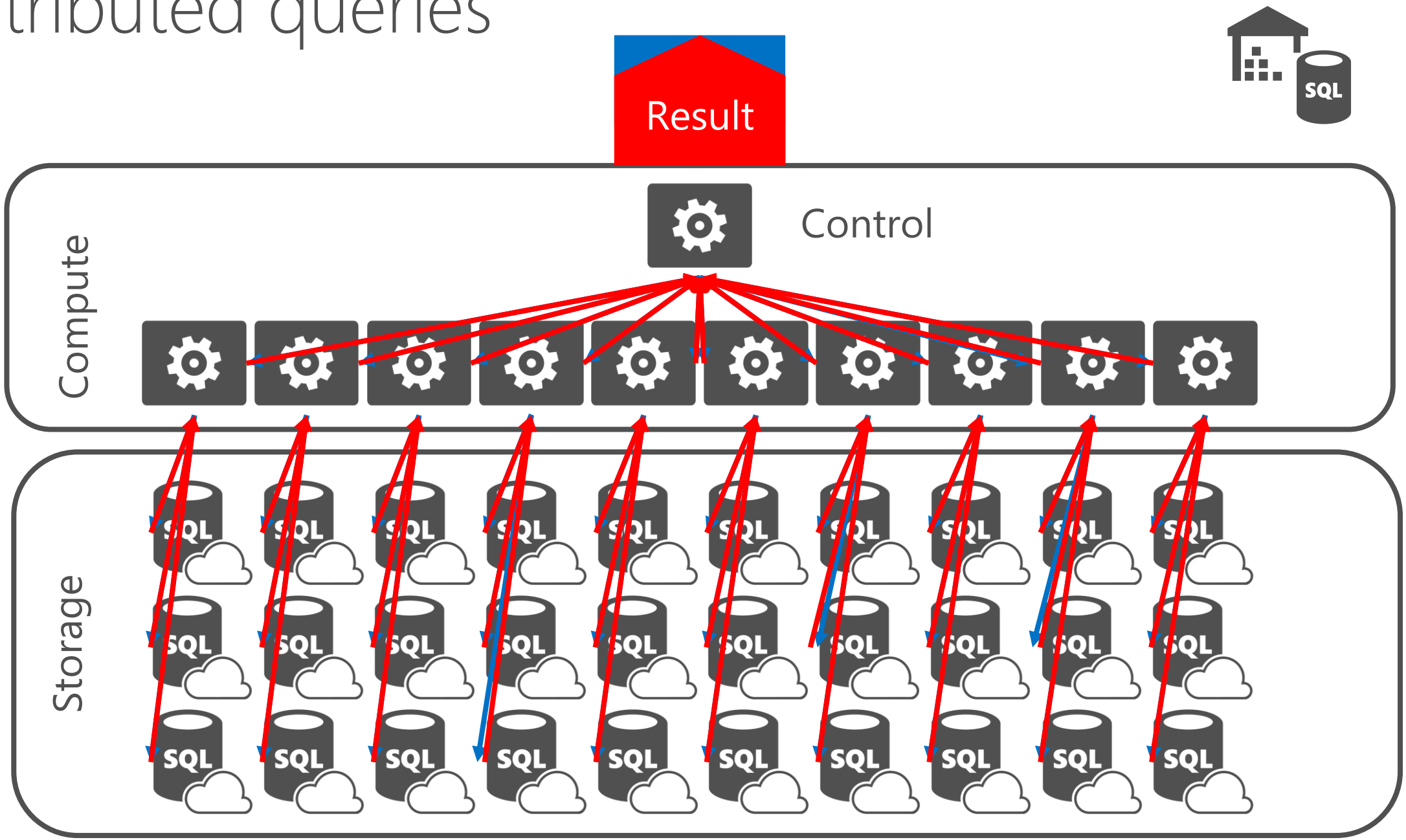


Synapse Analytics Architectural overview

Logical overview



Distributed queries



Simple example

```
SELECT  COUNT_BIG(*)
FROM    dbo.[FactInternetSales]
;
```

```
SELECT  SUM(*)
FROM    dbo.[FactInternetSales]
;
```

Control

Compute

```
SELECT  COUNT_BIG(*)
FROM    dbo.[FactInternetSales]
;
```

```
SELECT  COUNT_BIG(*)
FROM    dbo.[FactInternetSales]
;
```

```
SELECT  COUNT_BIG(*)
FROM    dbo.[FactInternetSales]
;
```

```
SELECT  COUNT_BIG(*)
FROM    dbo.[FactInternetSales]
;
```

Scale operations

Data Warehouse Units

Normalized amount of compute

Converts to billing units i.e. what you pay

DWUc
100
200
300
400
500
600
1000
1200
1500
2000
3000
6000
...
30000

Management Interfaces

	Portal	PowerShell	T-SQL
Create	✓	✓	⊘
Pause and resume	✓	✓	⊘
Scale	✓	✓	✓

Scaling via T-SQL and PowerShell

T-SQL

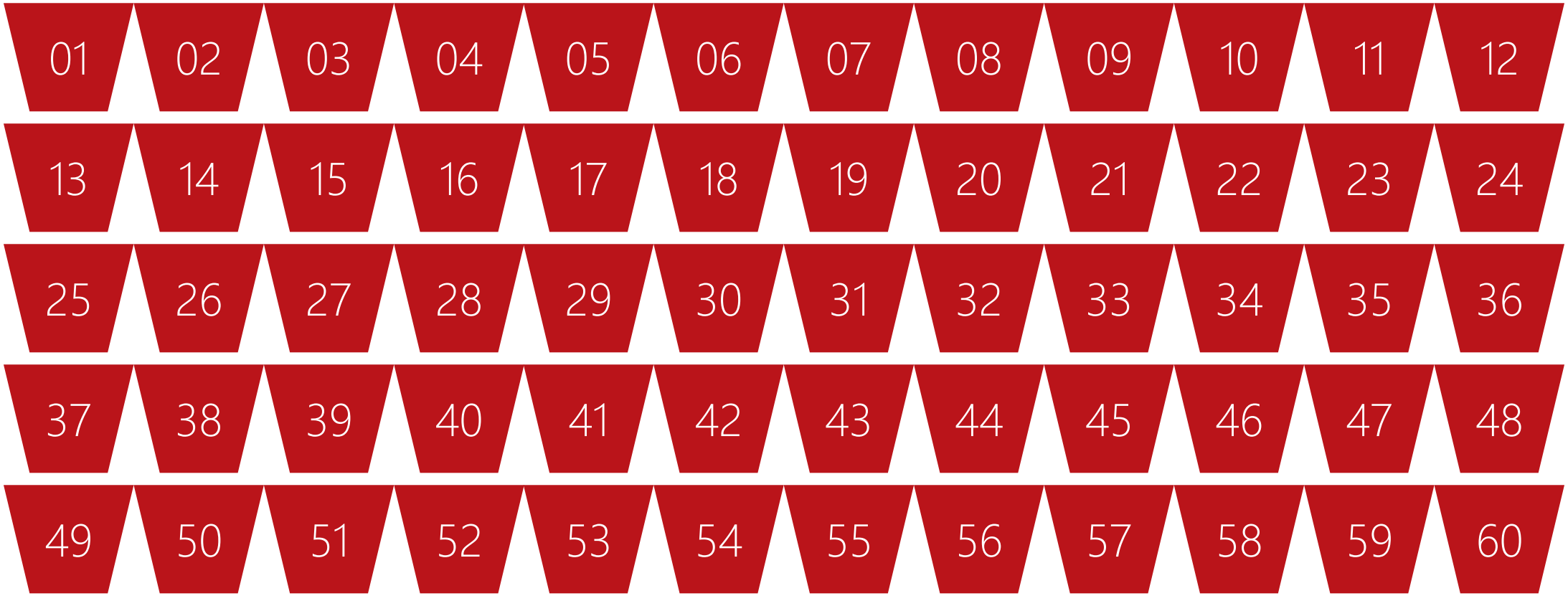
```
ALTER DATABASE ContosoRetailDW
MODIFY (service_objective = 'DW100');
```

PowerShell

```
Set-AzureRmSqlDatabase `
    -ResourceGroupName "RG_name" `
    -ServerName "SRV_name" `
    -DatabaseName "DB_name" `
    -RequestedServiceObjectiveName "DW1000"
```

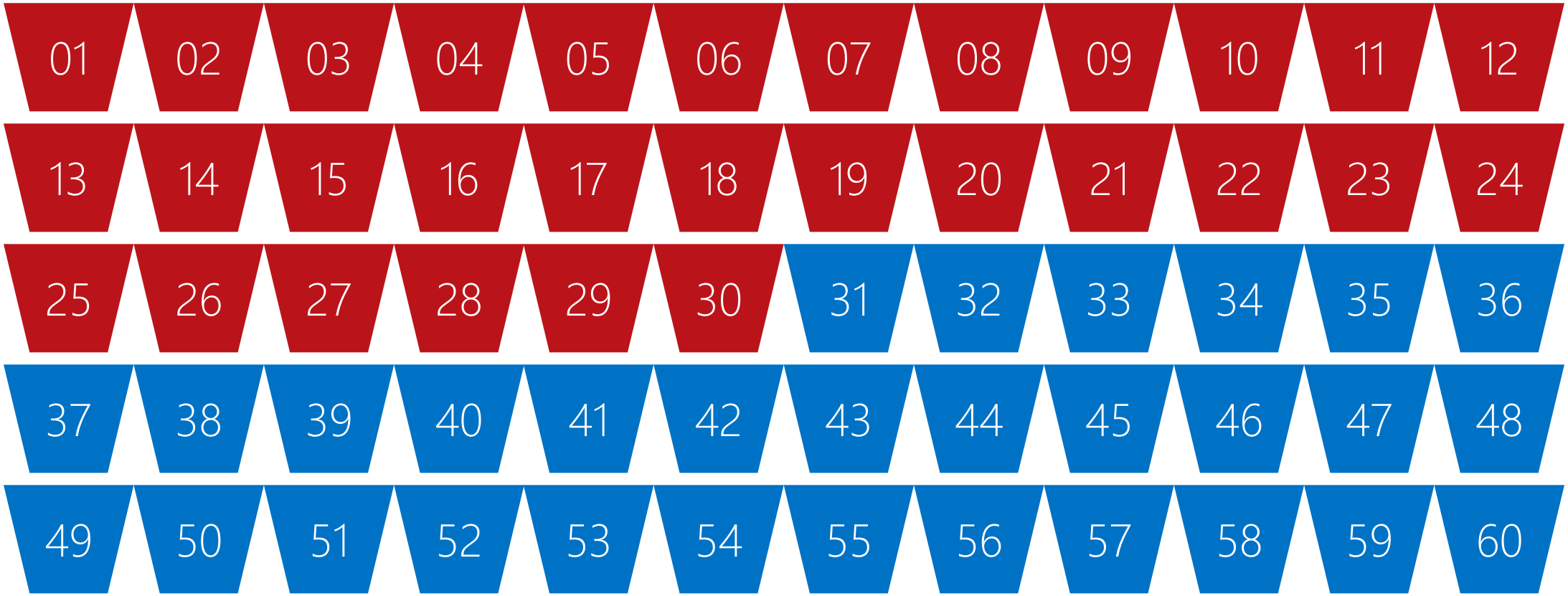
Mapping Compute in SQLDW

DW100



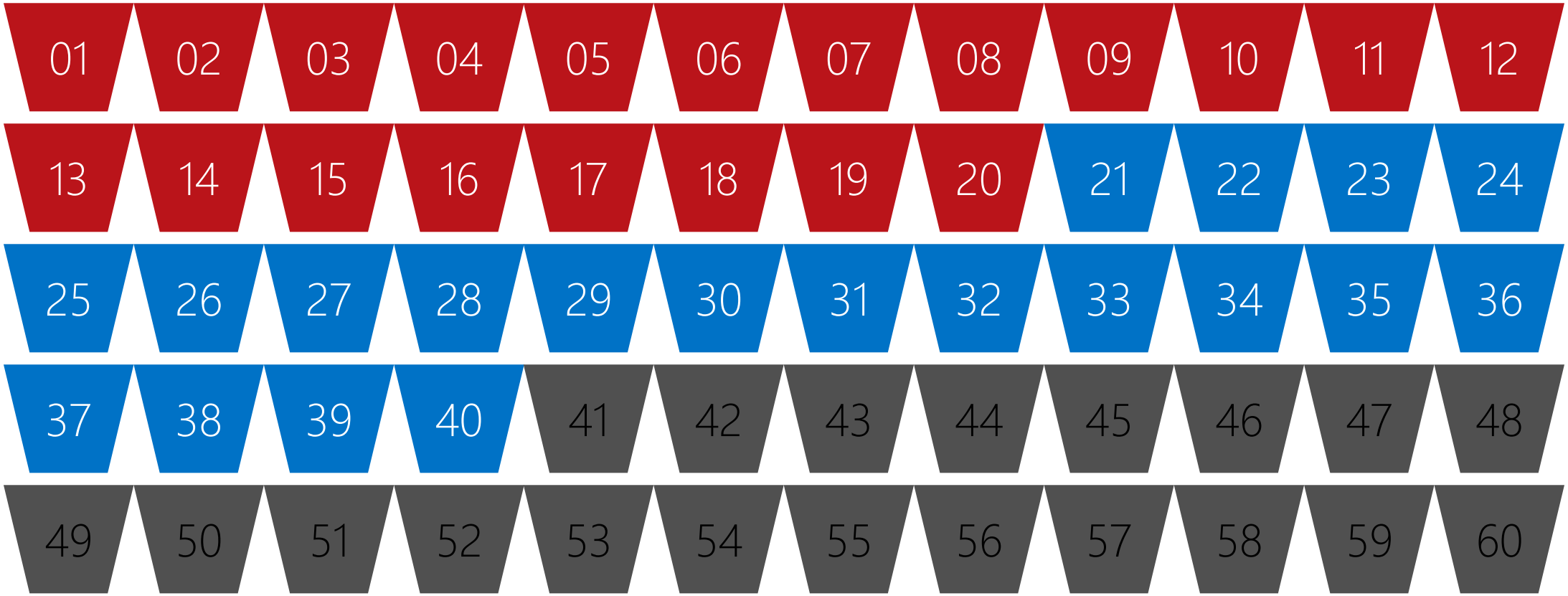
Mapping Compute in SQLDW

DW200



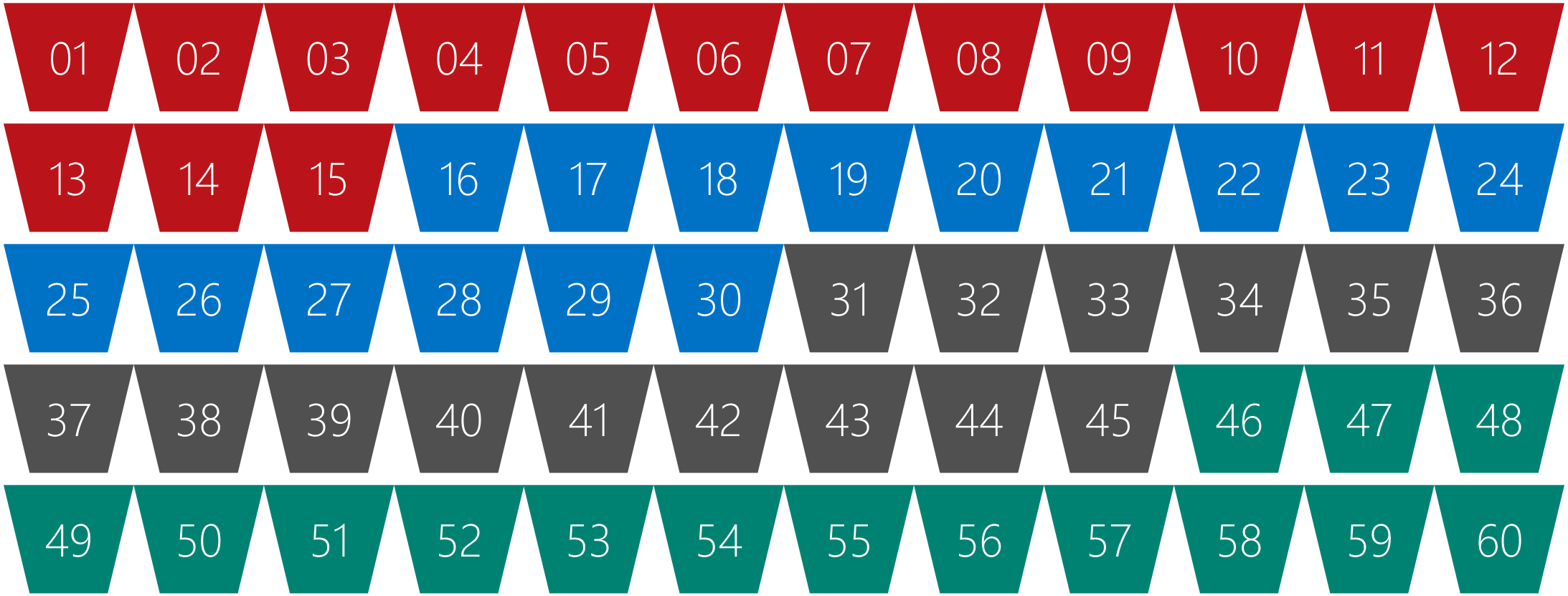
Mapping Compute in SQLDW

DW300



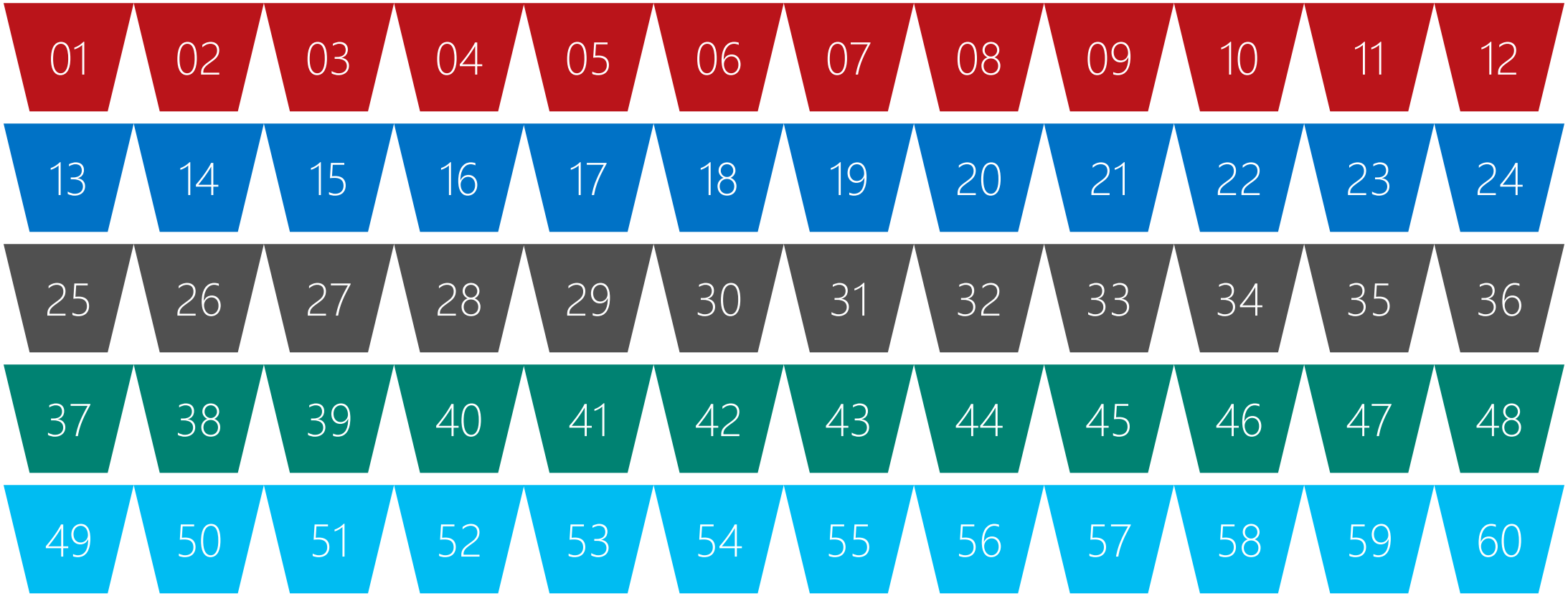
Mapping Compute in SQLDW

DW400



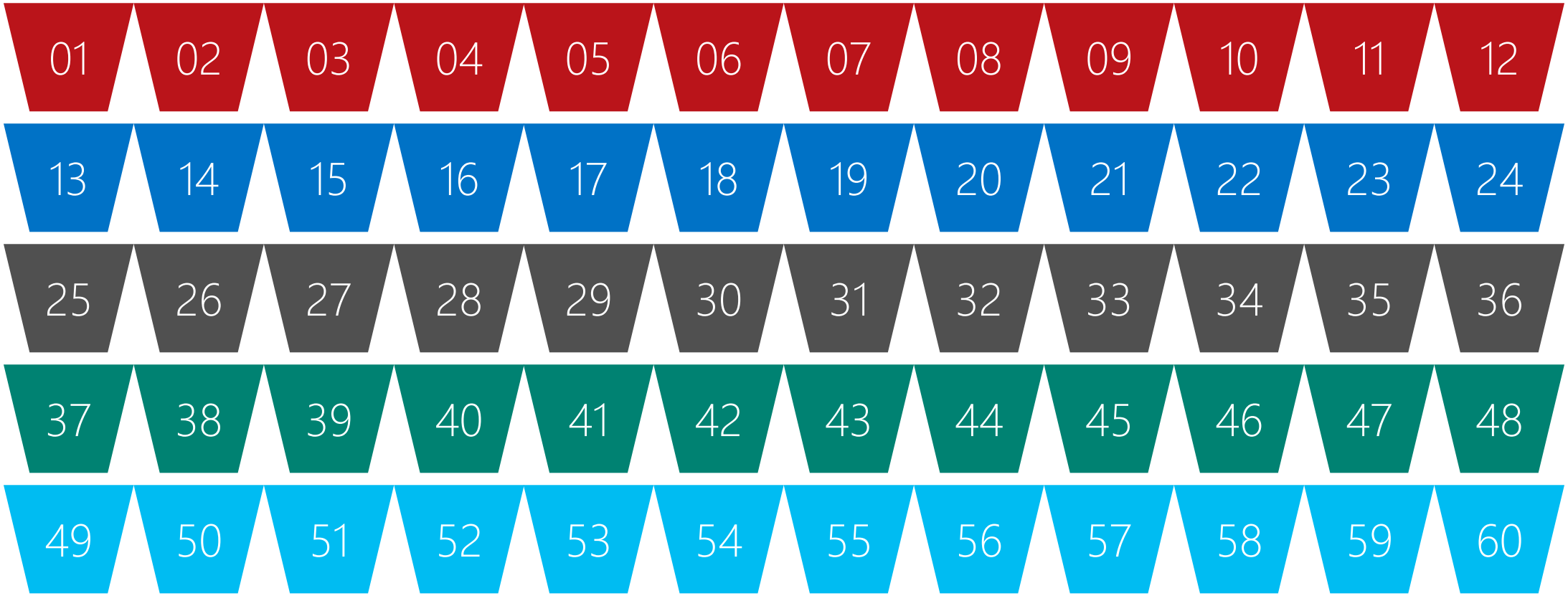
Mapping Compute in SQLDW

DW500



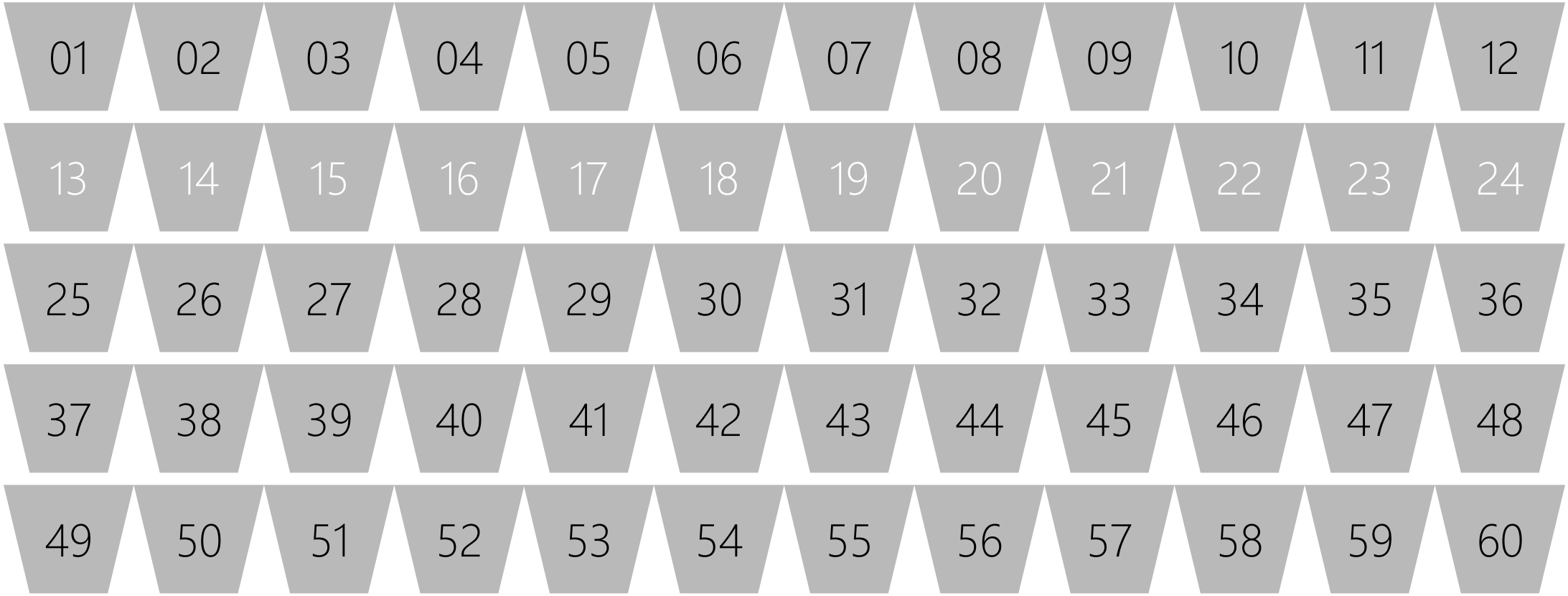
Pausing compute in SQLDW

DW500



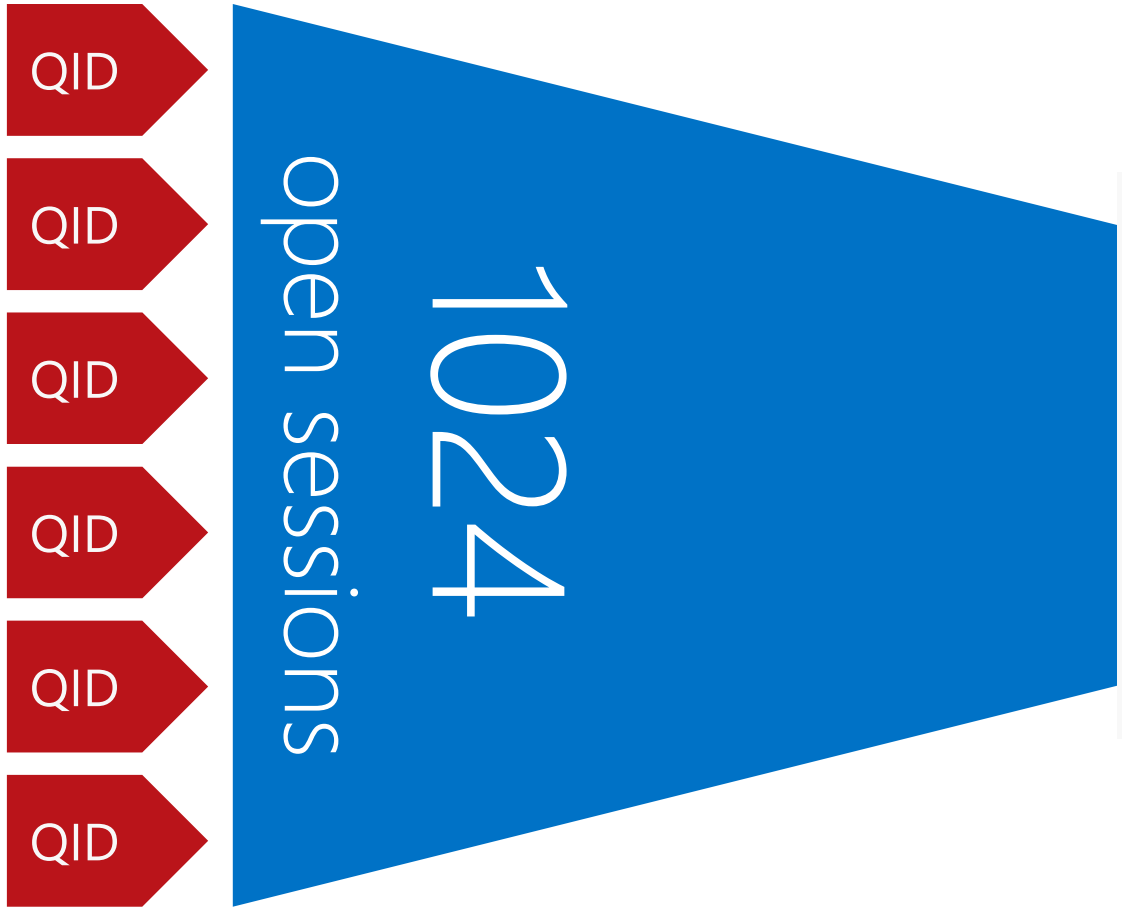
Resuming compute in SQLDW

DW500



Concurrency

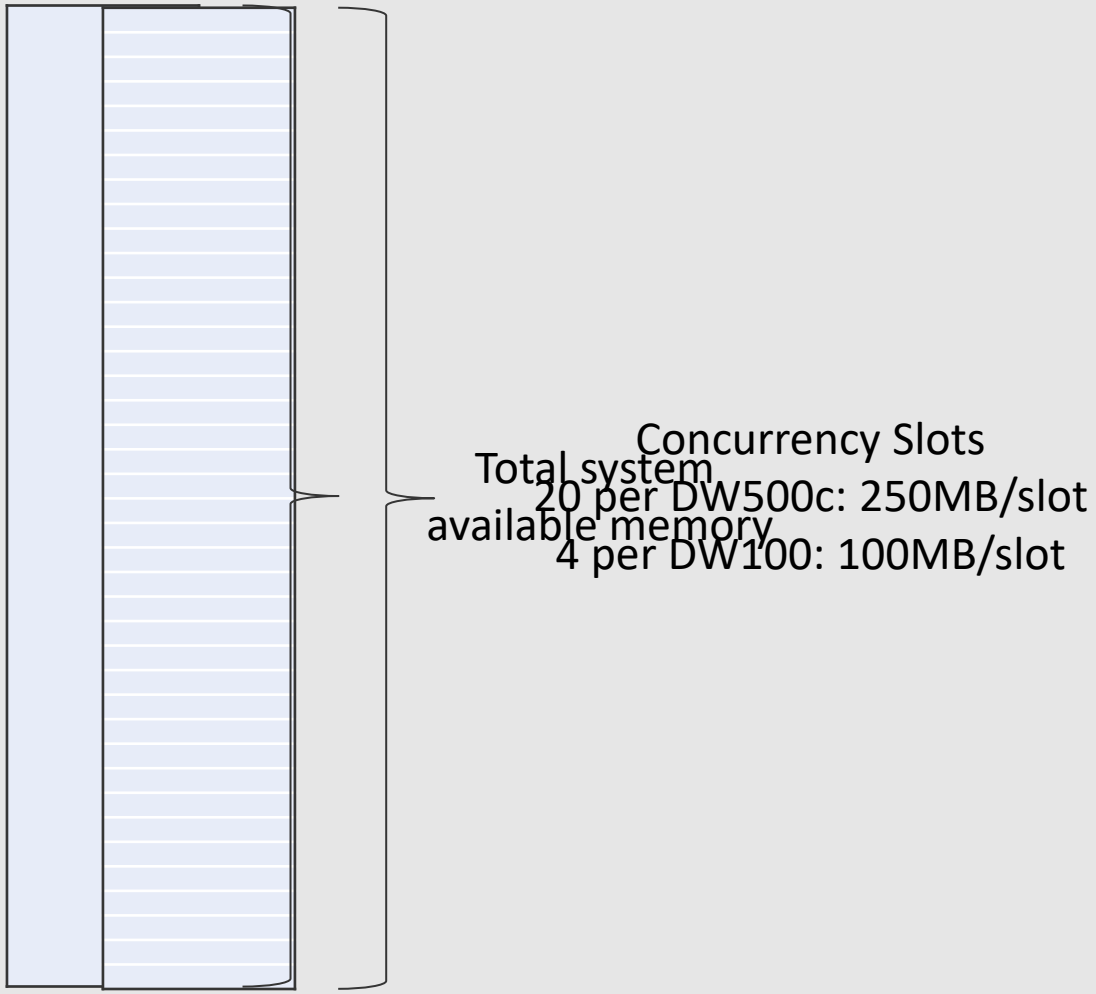
Concurrency: queries



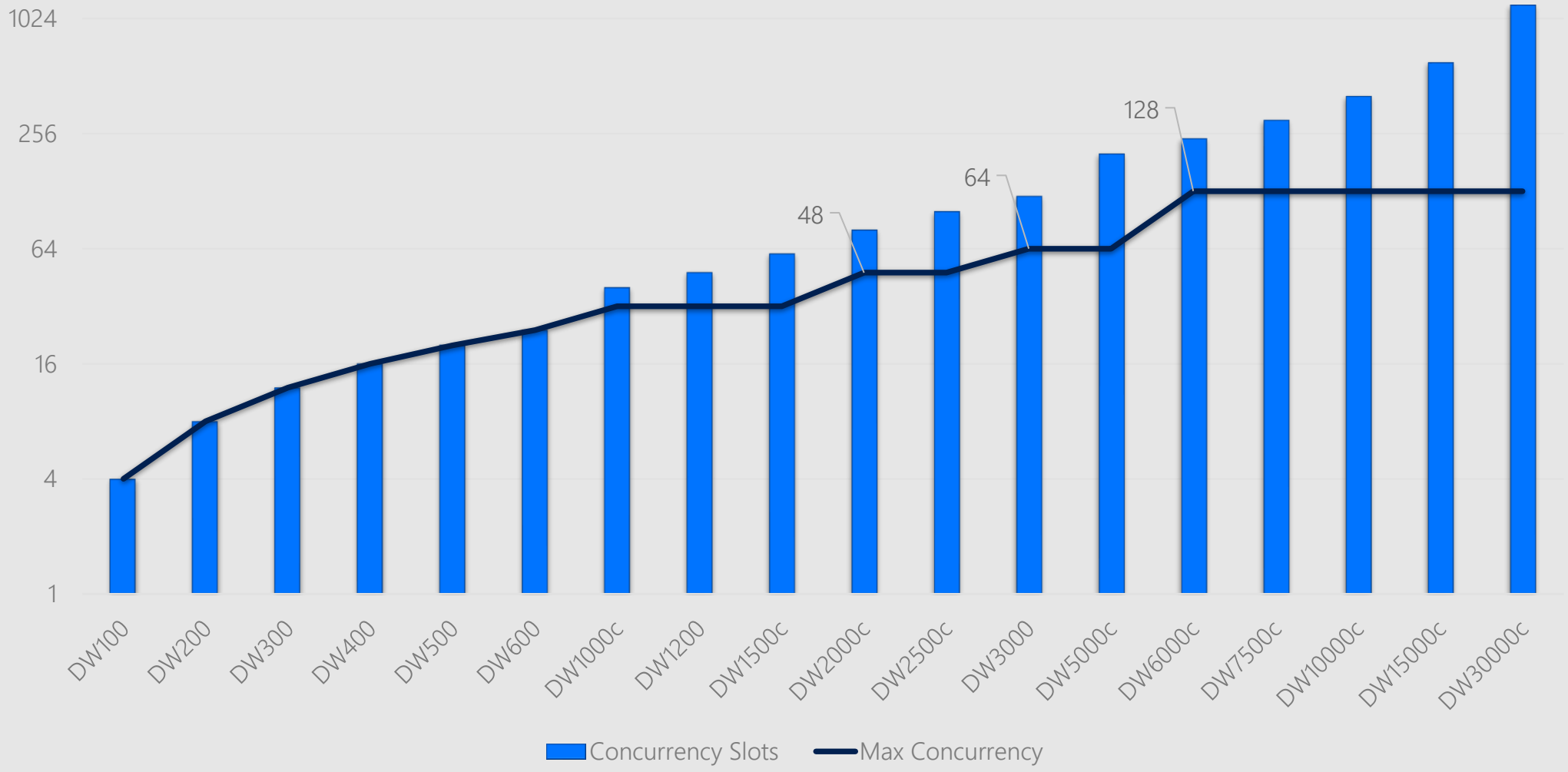
128

active queries

Concurrency slots



Concurrent Query and Slots



Memory & Concurrency Limits

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/memory-concurrency-limits

Concurrency maximums for workload groups

With the introduction of workload groups, the concept of concurrency slots no longer applies. Resources per request are allocated on a percentage basis and specified in the workload group definition. However, even with the removal of concurrency slots, there are minimum amounts of resources needed per queries based on the service level. The below table defined the minimum amount of resources needed per query across service levels and the associated concurrency that can be achieved.

Service Level	Maximum concurrent queries	Min % supported for REQUEST_MIN_RESOURCE_GRANT_PERCENT
DW100c	4	25%
DW200c	8	12.5%
DW300c	12	8%
DW400c	16	6.25%
DW500c	20	5%
DW1000c	32	3%
DW1500c	32	3%
DW2000c	48	2%
DW2500c	48	2%
DW3000c	64	1.5%
DW5000c	64	1.5%
DW6000c	128	0.75%
DW7500c	128	0.75%
DW10000c	128	0.75%
DW15000c	128	0.75%
DW30000c	128	0.75%

Resource Class

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management

Static resource classes

Static resource classes allocate the same amount of memory regardless of the current performance level, which is measured in data warehouse units. Since queries get the same memory allocation regardless of the performance level, scaling out the data warehouse allows more queries to run within a resource class. Static resource classes are ideal if the data volume is known and constant.

The static resource classes are implemented with these pre-defined database roles:

- staticrc10
- staticrc20
- staticrc30
- staticrc40
- staticrc50
- staticrc60
- staticrc70
- staticrc80

Dynamic resource classes

Dynamic Resource Classes allocate a variable amount of memory depending on the current service level. While static resource classes are beneficial for higher concurrency and static data volumes, dynamic resource classes are better suited for a growing or variable amount of data. When you scale up to a larger service level, your queries automatically get more memory.

The dynamic resource classes are implemented with these pre-defined database roles:

- smallrc
- mediumrc
- largerc
- xlargerc

The memory allocation for each resource class is as follows.

Service Level	smallrc	mediumrc	largerc	xlargerc
DW100c	25%	25%	25%	70%
DW200c	12.5%	12.5%	22%	70%
DW300c	8%	10%	22%	70%
DW400c	6.25%	10%	22%	70%
DW500c	5%	10%	22%	70%
DW1000c to DW30000c	3%	10%	22%	70%

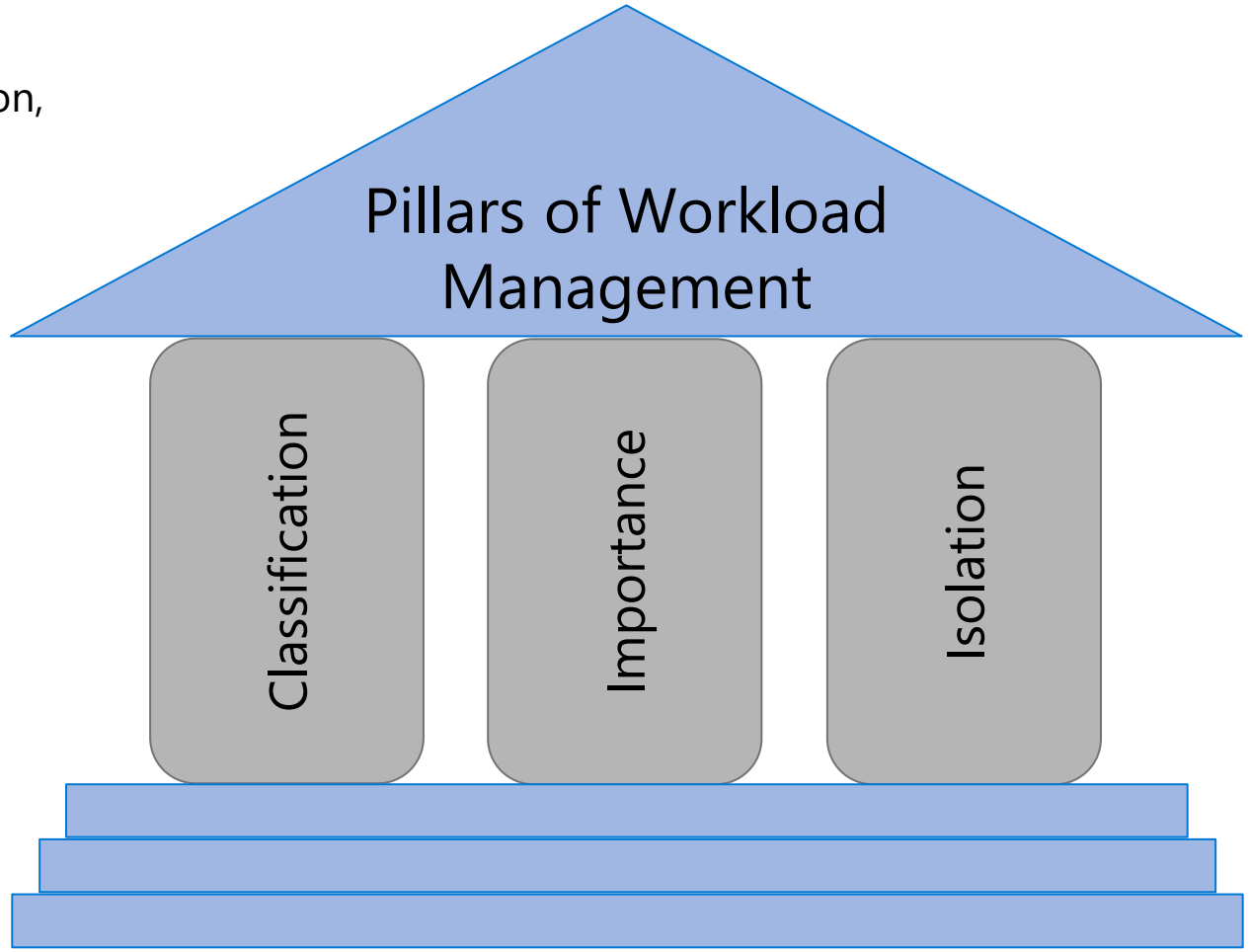
Workload Management

Overview

It manages resources, ensures highly efficient resource utilization, and maximizes return on investment (ROI).

The three pillars of workload management are

1. Workload Classification – To assign a request to a workload group and setting importance levels.
2. Workload Importance – To influence the order in which a request gets access to resources.
3. Workload Isolation – To reserve resources for a workload group.



Workload classification

Overview

Map queries to allocations of resources via pre-determined rules.

Use with workload importance to effectively share resources across different workload types.

If a query request is not matched to a classifier, it is assigned to the default workload group.

Benefits

Map queries to both Resource Management and Workload Isolation concepts.

Monitoring DMVs

sys.workload_management_workload_classifiers
sys.workload_management_workload_classifier_details

Query DMVs to view details about all active workload classifiers.

```
CREATE WORKLOAD CLASSIFIER classifier_name
WITH
(
      WORKLOAD_GROUP = 'name'
   ,  MEMBERNAME     = 'security_account'
 [ [ , ] IMPORTANCE     = {LOW|BELOW_NORMAL|NORMAL|ABOVE_NORMAL|HIGH} ] )
 [ [ , ] WLM_LABEL      = 'label' ]
 [ [ , ] WLM_CONTEXT    = 'name' ]
 [ [ , ] START_TIME     = 'start_time' ]
 [ [ , ] END_TIME       = 'end_time' ]
)[ ; ]
```

***WORKLOAD_GROUP**: maps to an existing resource class*
***IMPORTANCE**: specifies relative importance of request*
***MEMBERNAME**: database user, role, AAD login or AAD group*

Workload importance

Overview

Queries past the concurrency limit enter a FiFo queue

By default, queries are released from the queue on a first-in, first-out basis as resources become available

Workload importance allows higher priority queries to receive resources immediately regardless of queue

Example Video

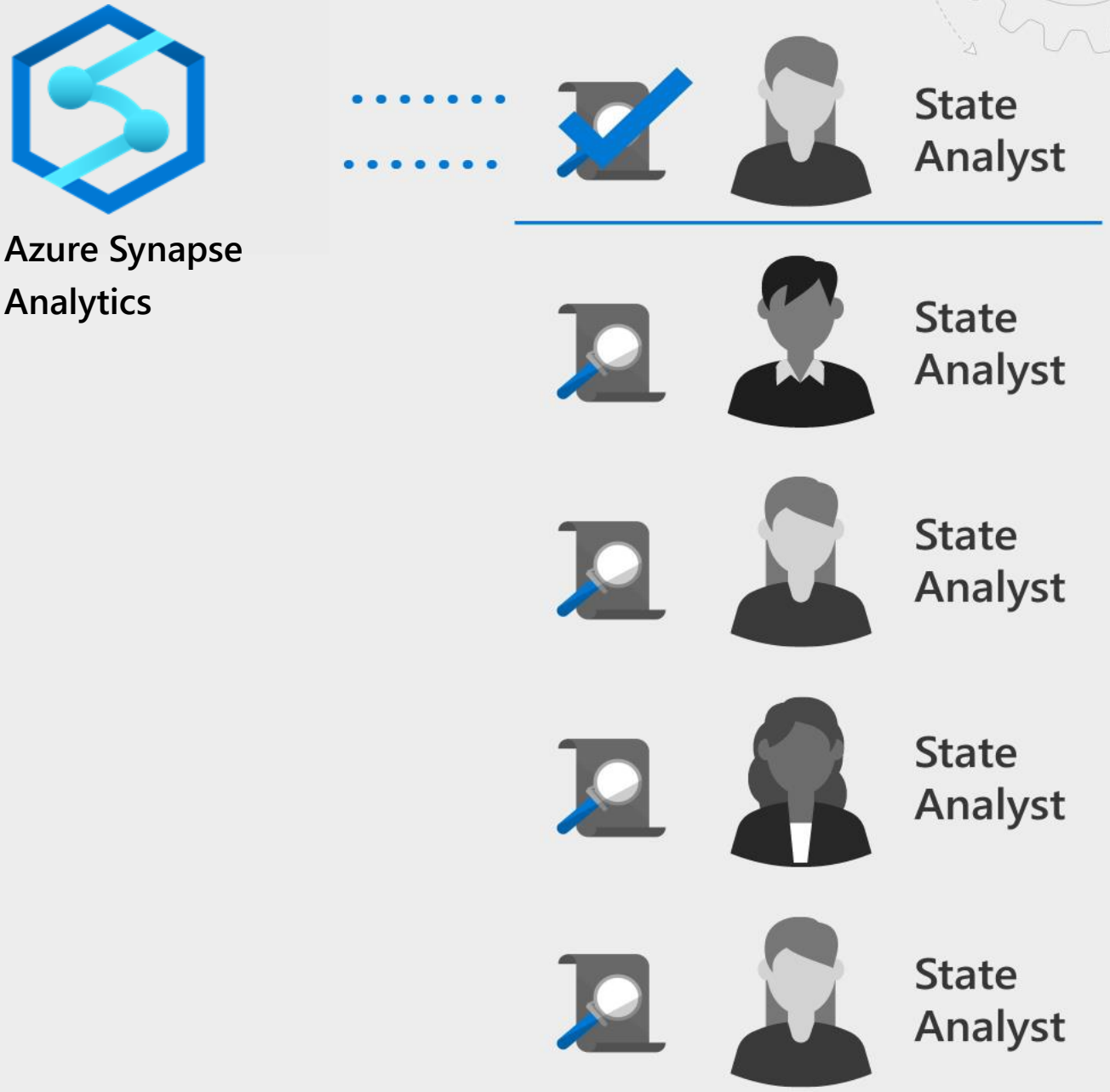
State analysts have normal importance.

National analyst is assigned high importance.

State analyst queries execute in order of arrival

When the national analyst's query arrives, it jumps to the top of the queue

```
CREATE WORKLOAD CLASSIFIER National_Analyst
WITH
(
  WORKLOAD_GROUP = 'analyst'
  ,IMPORTANCE    = HIGH
  ,MEMBERNAME    = 'National_Analyst_Login')
```



Workload Isolation

Overview

Allocate fixed resources to workload group.

Assign maximum and minimum usage for varying resources under load. These adjustments can be done live without having to Synapse SQL (provisioned) offline.

Benefits

Reserve resources for a group of requests

Limit the amount of resources a group of requests can consume

Shared resources accessed based on importance level

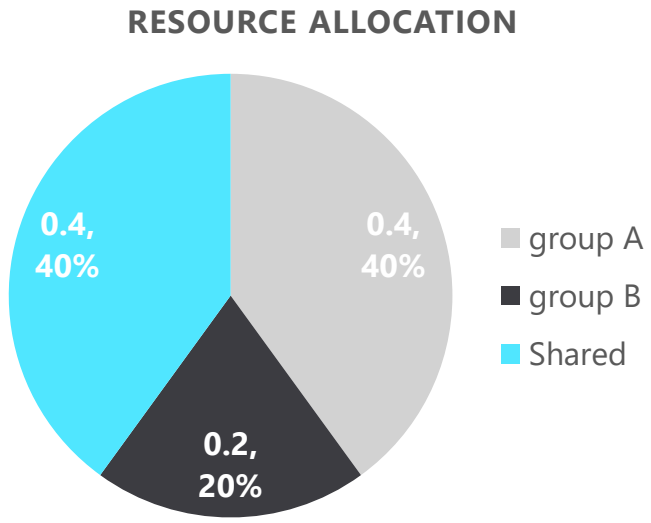
Set Query timeout value. Get DBAs out of the business of killing runaway queries

Monitoring DMVs

sys.workload_management_workload_groups

Query to view configured workload group.

```
CREATE WORKLOAD GROUP group_name
WITH
(
   MIN_PERCENTAGE_RESOURCE = value
 , CAP_PERCENTAGE_RESOURCE = value
 , REQUEST_MIN_RESOURCE_GRANT_PERCENT = value
 [ [ , ] REQUEST_MAX_RESOURCE_GRANT_PERCENT = value ]
 [ [ , ] IMPORTANCE = {LOW | BELOW_NORMAL | NORMAL | ABOVE_NORMAL | HIGH} ]
 [ [ , ] QUERY_EXECUTION_TIMEOUT_SEC = value ]
)[ ; ]
```

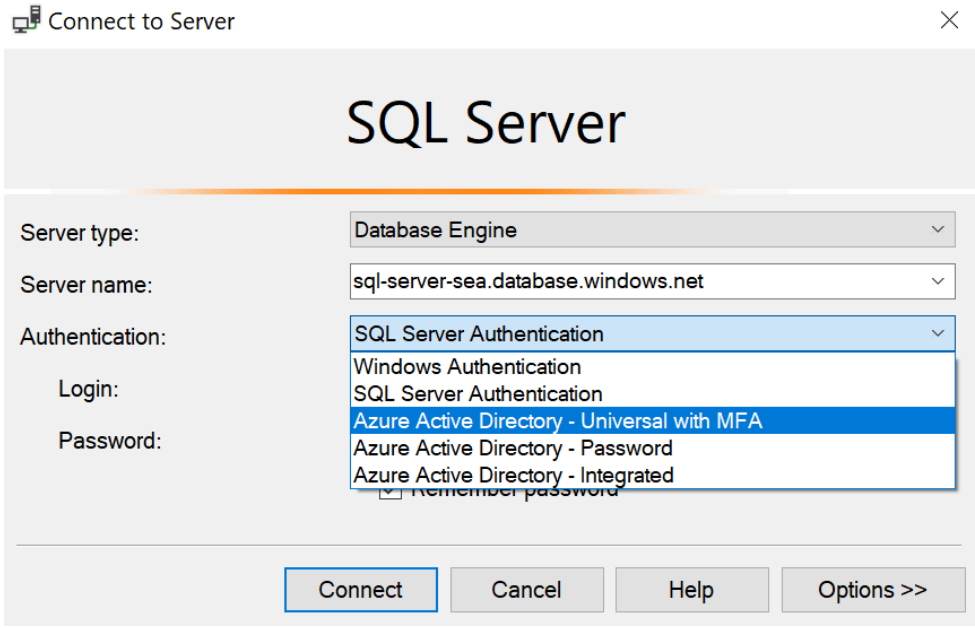


User & Authentication

```
-- Master Database
CREATE LOGIN MedRCLogin WITH PASSWORD = 'rladydtjs$1971';

CREATE USER LoadingUser FOR LOGIN MedRCLogin;
```

```
-- Master Database
CREATE USER LoadingUser FOR LOGIN MedRCLogin;

GRANT CONTROL ON DATABASE::[yongdw] to LoadingUser;

EXEC sp_addrolemember 'mediumrc', 'LoadingUser';
```

```
Create Login Mary@domainname.net From EXTERNAL PROVIDER;
Create User Mary From Login Mary@domainname.net;
```



Tables & Distribution

Table Distribution Options

Hash Distributed	Round Robin (Default)	Replicated
Data divided across nodes based on hashing algorithm Same value will always hash to same distribution Single column only	Data distributed evenly across nodes Easy place to start, don't need to know anything about the data Simplicity at a cost	Data repeated on every node Simplifies many query plans and reduces data movement Best with joining hash table
Check for Data Skew, NULLS, -1	Will incur more data movement at query time	Consumes more space Joining two Replicated Table runs on one node

Creating tables

```
CREATE TABLE [build].[FactOnlineSales]
(
    [OnlineSalesKey]        int           NOT NULL
,   [DateKey]               datetime      NOT NULL
,   [StoreKey]              int           NOT NULL
,   [ProductKey]            int           NOT NULL
,   [PromotionKey]          int           NOT NULL
,   [CurrencyKey]           int           NOT NULL
,   [CustomerKey]           int           NOT NULL
,   [SalesOrderNumber]      nvarchar(20)  NOT NULL
,   [SalesOrderLineNumber]  int               NULL
,   [SalesQuantity]         int           NOT NULL
,   [SalesAmount]           money         NOT NULL
)
WITH
(   CLUSTERED COLUMNSTORE INDEX
,   DISTRIBUTION = ROUND_ROBIN
)
;
```

```
CREATE TABLE [build].[FactOnlineSales]
(
    [OnlineSalesKey]        int           NOT NULL
,   [DateKey]               datetime      NOT NULL
,   [StoreKey]              int           NOT NULL
,   [ProductKey]            int           NOT NULL
,   [PromotionKey]          int           NOT NULL
,   [CurrencyKey]           int           NOT NULL
,   [CustomerKey]           int           NOT NULL
,   [SalesOrderNumber]      nvarchar(20)  NOT NULL
,   [SalesOrderLineNumber]  int               NULL
,   [SalesQuantity]         int           NOT NULL
,   [SalesAmount]           money         NOT NULL
)
WITH
(   CLUSTERED COLUMNSTORE INDEX
,   DISTRIBUTION = HASH([ProductKey])
)
;
```

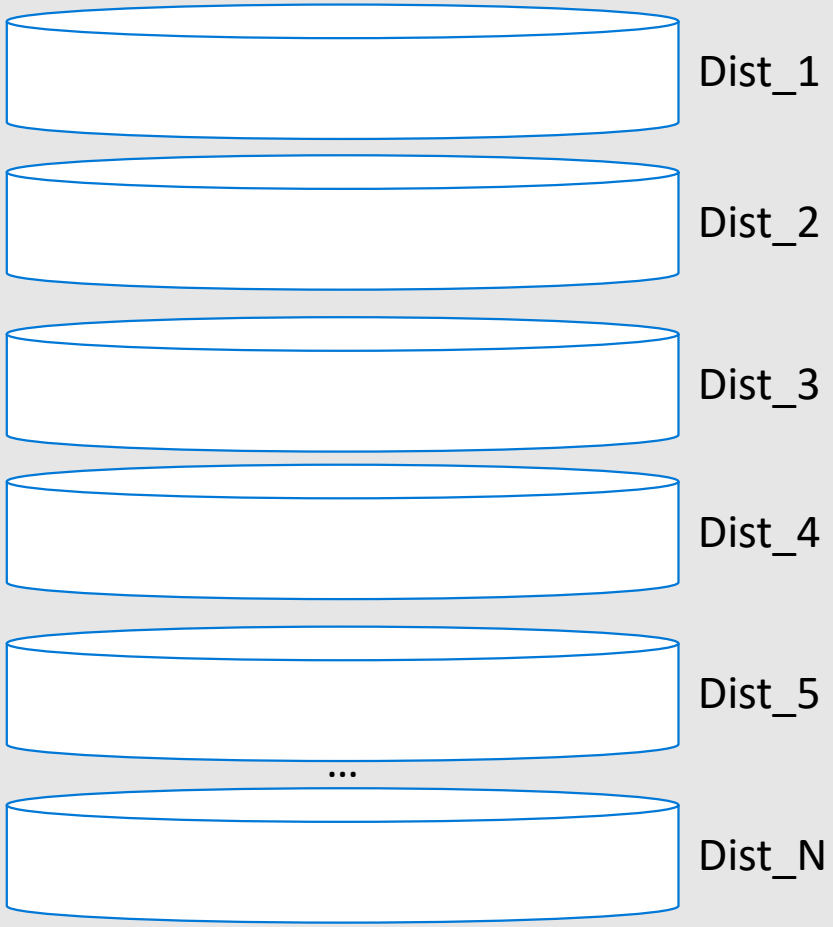
Hash Distributed

```
CREATE TABLE ProductSales
WITH (DISTRIBUTION=HASH(AccountID))
AS...
```

ProductSales – Raw Data

AccountI	SalesAmt	…
47	$1,234.36	…
36	$2,345.47	…
14	$3,456.58	…
25	$4,567.69	…
48	$5,678.70	…
37	$6,789.81	…
…	…	…

Hash(37)

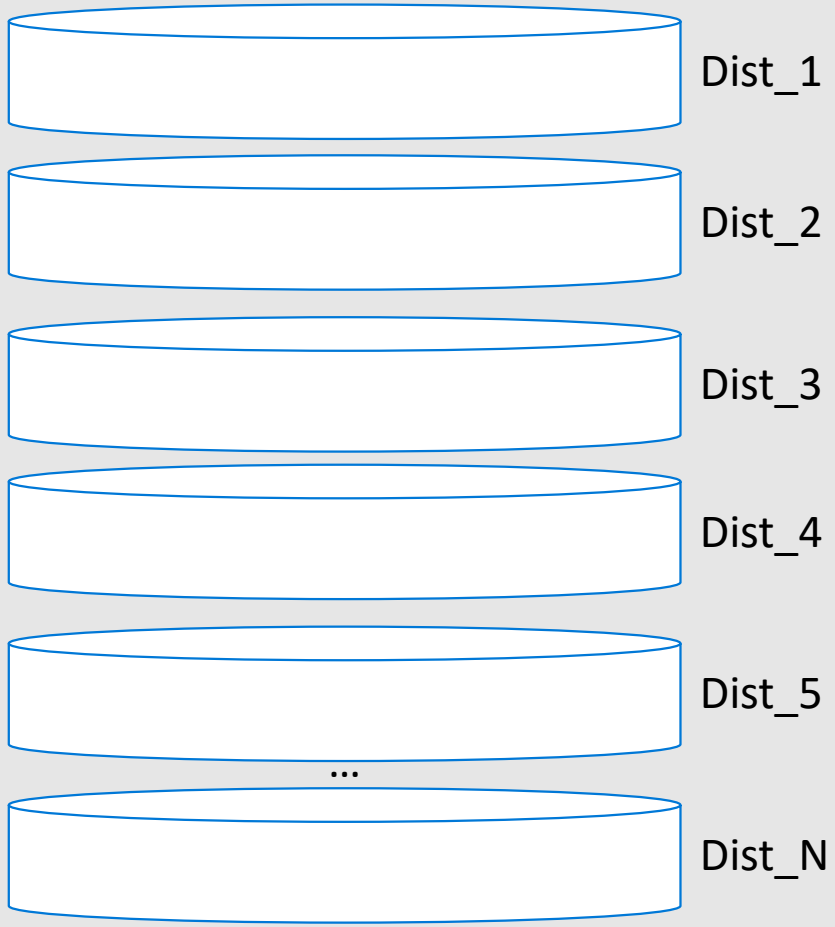


Round Robin Distributed

```
CREATE TABLE ProductSales
WITH (DISTRIBUTION = ROUND_ROBIN)
AS ...
```

ProductSales – Raw Data

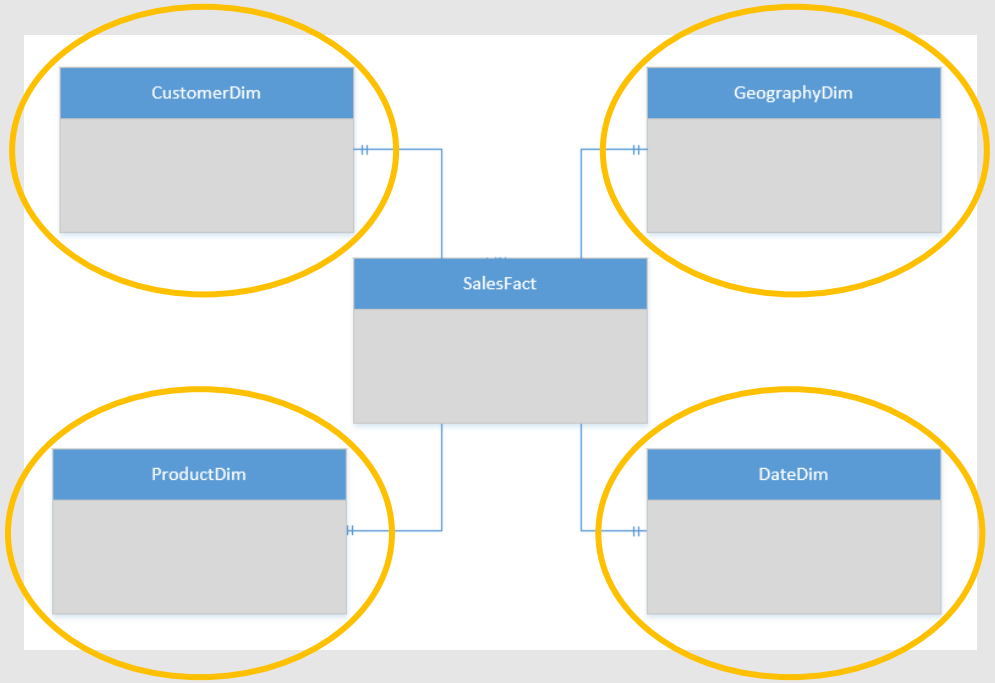
AccountID	SalesAmt	…
47	$1,234.36	…
36	$2,345.47	…
14	$3,456.58	…
25	$4,567.69	…
47	$5,678.70	…
37	$6,789.81	…
42	$1,632.25	…
88	$4,453.21	…
52	$7,892.81	…
91	$9,549.64	…
88	$2,498.14	…
42	$3,145.99	…
23	$3,145.99	…



Replicated Table Scenarios

Scenarios to consider using Replicated tables:

- Star schema reporting
  - Dimensions – descriptive entities about fact data
- ETL master data, common domain data used during transaction loading



T-SQL: Create a Replicated table

```
CREATE TABLE dbo.DimCustomer
(    CustomerKey            int             NOT NULL
,    GeographyKey           int             NULL
,    CustomerAlternateKey   nvarchar(15)    NOT NULL
,    Title                  nvarchar(8)     NULL
,    FirstName              nvarchar(50)    NULL
,    LastName               nvarchar(50)    NULL
,    BirthDate              date            NULL
,    Gender                 nvarchar(1)     NULL
,    EmailAddress           nvarchar(50)    NULL
,    YearlyIncome           money           NULL
,    DateFirstPurchase      date            NULL
)
WITH
(    CLUSTERED COLUMNSTORE INDEX
,    DISTRIBUTION = REPLICATED
)
```

Why data moves

Data has to be co-located to operated on…

Common reasons:

- Incompatible join
- Incompatible aggregation

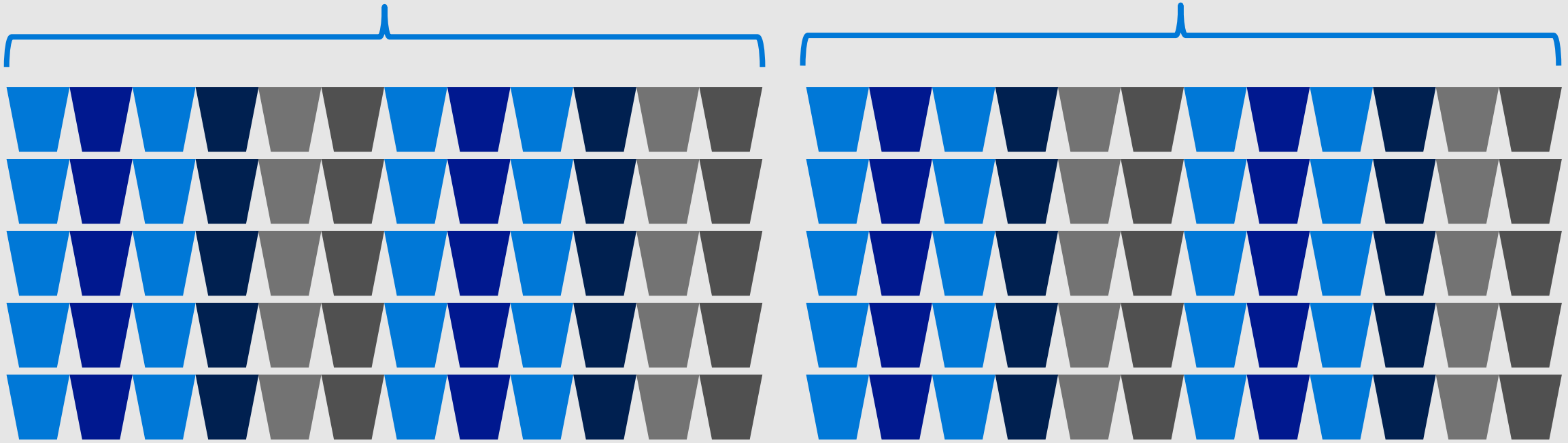
Data Movement Types for a Query

DMS Operation	Description
ShuffleMoveOperation	Distribution → Hash algorithm → New distribution Changing the distribution column in preparation for join.
PartitionMoveOperation	Distribution → Control Node Aggregations - count(*) is count on nodes, sum of count
BroadcastMoveOperation	Distribution → Copy to all distributions Changes distributed table to replicated table for join.
TrimMoveOperation	Replicated table → Hash algorithm → Distribution When a replicated table needs to become distributed. Needed for outer joins.
MoveOperation	Control Node → Copy to all distributions Data moved from Control Node back to Compute Nodes resulting in a replicated table for further processing.

Joining HASH tables

Store_Sales HASH([ProductKey])
[ProductKey] **INT** NULL

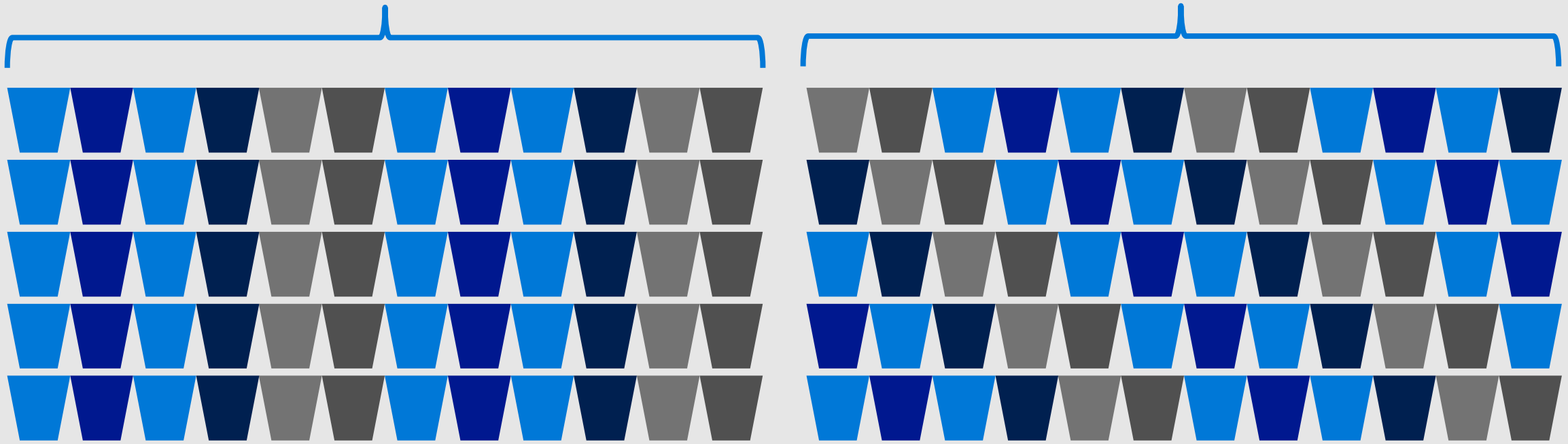
Web_Sales HASH([ProductKey])
[ProductKey] **INT** NULL



Joining HASH tables

Store_Sales HASH([ProductKey])
[ProductKey] **INT** NULL

Web_Sales HASH([ProductKey])
[ProductKey] **BIGINT** NULL



Aggregation - compatible

Resolved completely on each compute node

No Data Movement

1. Hash Distribution Key is contained in the group by keys
2. Count Distinct on distribution key

```
-- FactOnlineSales distributed on ProductKey

SELECT      COUNT_BIG(*)
FROM        [cso].[FactOnlineSales]
GROUP BY    [ProductKey]

SELECT  COUNT_BIG(DISTINCT ([ProductKey]))
FROM    [cso].[FactOnlineSales]
```

Aggregation - Incompatible

Partially aggregated on each node

Shuffle move co-locates rows with same group by key

1. Table is round robin distributed
2. Hash Distribution key is not contained in group by keys
3. Count Distinct on non-distribution key or on round robin table

```
-- FactOnlineSales distributed on ProductKey

SELECT      COUNT_BIG(*)
FROM        [cso].[FactOnlineSales]
GROUP BY    [StoreKey]

SELECT  COUNT_BIG(DISTINCT [DateKey])
FROM    [cso].[FactOnlineSales]
```

Common table distribution methods

Table Category	Recommended Distribution Option
Fact	Use hash-distribution with clustered columnstore index. Performance improves because hashing enables the platform to localize certain operations within the node itself during query execution. Operations that benefit: COUNT(DISTINCT(<hashed_key>)) OVER PARTITION BY <hashed_key> most JOIN <table_name> ON <hashed_key> GROUP BY <hashed_key>
Dimension	Use replicated for smaller tables. If tables are too large to store on each Compute node, use hash-distributed.
Staging	Use round-robin for the staging table. The load with CTAS is faster. Once the data is in the staging table, use INSERT…SELECT to move the data to production tables.

Distribution Guidance

For large fact tables, best option is to Hash Distribute

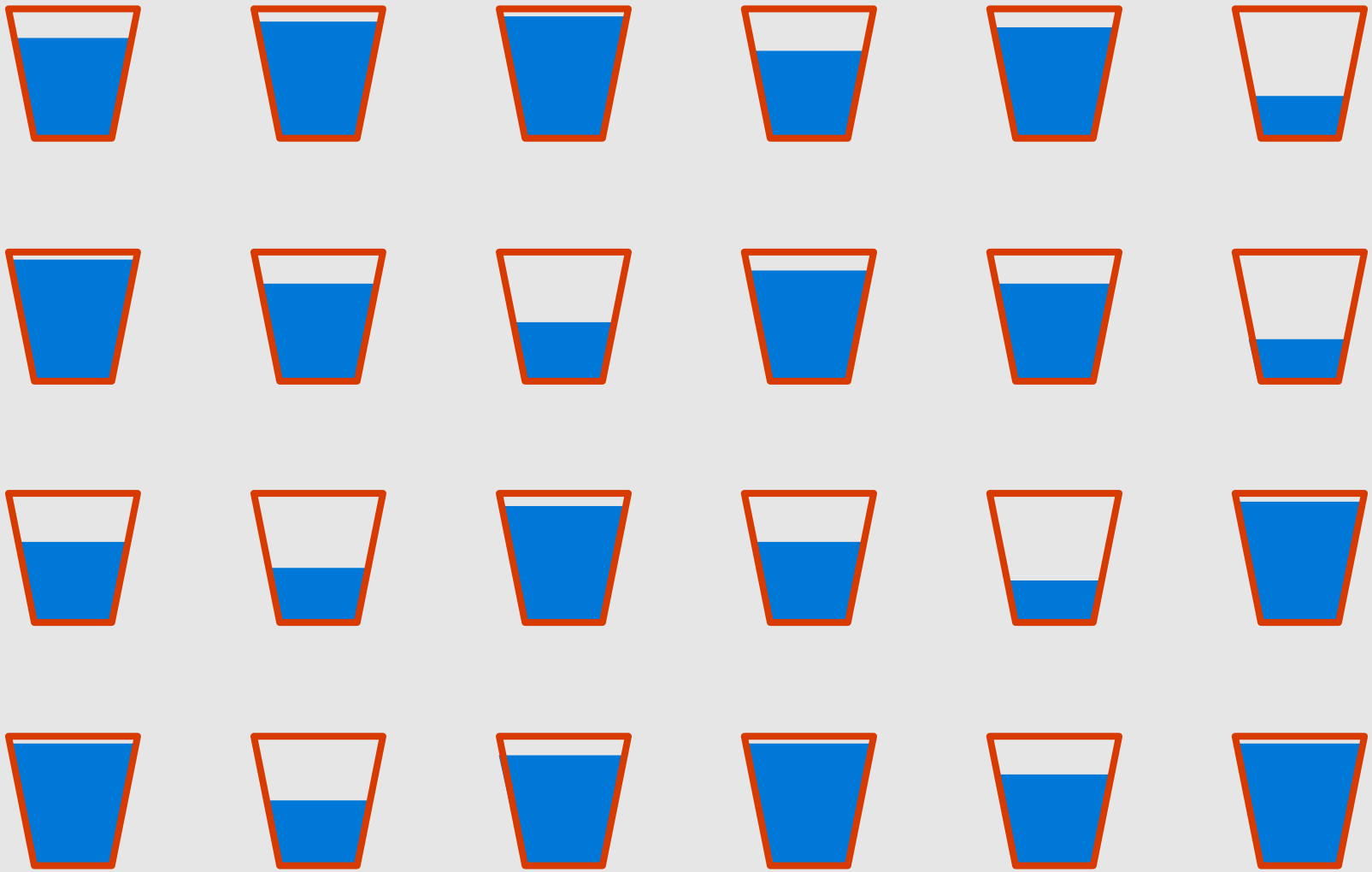
- Clustered Columnstore
- Distribute on column that is joined to other fact tables or large dimensions
- Primary or surrogate key maybe a good choice for distribution

However, be mindful of …

- Hash column should have highly distinct values (Minimum 600 distinct values)
- Avoid distributing on a date column
- Avoid distributing on column with high frequency of NULLs and default values (e.g. -1)
- Distribution column is NOT updatable
- For compatible joins use the same data types for two distributed tables

If there are no distribution columns that make sense, then use Round Robin as last resort

Skewed Distribution



Finding Skew

DBCC PDW_SHOWSPACEUSED

- Not a programmatic interface

DMV gives more detail and control

- sys.dm_pdw_nodes_db_partition_stats
- Refer to the view dbo.vTableSizes provided in docs (Table Size Queries)

```
select distribution_id,SUM(row_count) as total_distribution_row_count
from dbo.vTableSizes
where schema_name = 'Fact' and table_name = 'Flights'
group by distribution_id
order by total_distribution_row_count;
```

Tables – Partitions

Overview

Table partitions divide data into smaller groups

In most cases, partitions are created on a date column

Supported on all table types

RANGE RIGHT – Used for time partitions

RANGE LEFT – Used for number partitions

Benefits

Improves efficiency and performance of loading and querying by limiting the scope to subset of data.

Offers significant query performance enhancements where filtering on the partition key can eliminate unnecessary scans and eliminate IO.

```
CREATE TABLE partitionedOrderTable
(
    OrderId  INT NOT NULL,
    Date     DATE NOT NULL,
    Name     VARCHAR(2),
    Country  VARCHAR(2)
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX,
    DISTRIBUTION = HASH([OrderId]),
    PARTITION (
    [Date] RANGE RIGHT FOR VALUES (
    '2000-01-01', '2001-01-01', '2002-01-01',
    '2003-01-01', '2004-01-01', '2005-01-01'
    )
  )
);
```

Tables – DW Distributions & Partitions Illustrated

Logical table structure

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

Physical data distribution

(Hash distribution (OrderId), Date partitions)

Distribution1
(OrderId 80,000 – 100,000)

11-2-2018 partition

OrderId	Date	Name	Country
85016	11-2-2018	V	UK
85018	11-2-2018	Q	SP
85216	11-2-2018	Q	DE
85395	11-2-2018	V	NL
82147	11-2-2018	Q	FR
86881	11-2-2018	D	UK
...	...	...	...

11-3-2018 partition

OrderId	Date	Name	Country
93080	11-3-2018	R	UK
94156	11-3-2018	S	FR
96250	11-3-2018	Q	NL
98799	11-3-2018	R	NL
98015	11-3-2018	T	UK
98310	11-3-2018	D	DE
98979	11-3-2018	Z	DE
98137	11-3-2018	T	FR
...	...	...	...

• • •

x 60 distributions (shards)

- Each shard is partitioned with the same date partitions
- A minimum of 1 million rows per distribution and partition is needed for optimal compression and performance of clustered Columnstore tables

Tables – Partitions

	logical_table_name	row_group_id	state	state_desc	total_rows	trim_reason_desc	physical_name
1	FactSalesQuotaCCI	3	3	COMPRESSED	514925	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_1
2	FactSalesQuotaCCI	3	3	COMPRESSED	512875	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_2
3	FactSalesQuotaCCI	3	3	COMPRESSED	511350	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_3
4	FactSalesQuotaCCI	3	3	COMPRESSED	509161	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_4
5	FactSalesQuotaCCI	3	3	COMPRESSED	512640	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_5
6	FactSalesQuotaCCI	3	3	COMPRESSED	512640	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_6
7	FactSalesQuotaCCI	3	3	COMPRESSED	512638	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_7
8	FactSalesQuotaCCI	3	3	COMPRESSED	511216	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_8
9	FactSalesQuotaCCI	3	3	COMPRESSED	511216	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_9
10	FactSalesQuotaCCI	3	3	COMPRESSED	512360	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_10
11	FactSalesQuotaCCI	3	3	COMPRESSED	511216	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_11
12	FactSalesQuotaCCI	3	3	COMPRESSED	510102	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_12
13	FactSalesQuotaCCI	3	3	COMPRESSED	509792	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_13
14	FactSalesQuotaCCI	3	3	COMPRESSED	512640	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_14
15	FactSalesQuotaCCI	3	3	COMPRESSED	512388	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_15
16	FactSalesQuotaCCI	3	3	COMPRESSED	513046	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_16
17	FactSalesQuotaCCI	3	3	COMPRESSED	513105	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_17
18	FactSalesQuotaCCI	3	3	COMPRESSED	512640	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_18
19	FactSalesQuotaCCI	3	3	COMPRESSED	512607	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_19
20	FactSalesQuotaCCI	3	3	COMPRESSED	511583	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_20
21	FactSalesQuotaCCI	3	3	COMPRESSED	512156	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_21
22	FactSalesQuotaCCI	3	3	COMPRESSED	512479	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_22
23	FactSalesQuotaCCI	3	3	COMPRESSED	511480	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_23
24	FactSalesQuotaCCI	3	3	COMPRESSED	512171	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_24
25	FactSalesQuotaCCI	3	3	COMPRESSED	512753	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_25
26	FactSalesQuotaCCI	3	3	COMPRESSED	511216	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_26
27	FactSalesQuotaCCI	3	3	COMPRESSED	509843	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_27
28	FactSalesQuotaCCI	3	3	COMPRESSED	511065	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_28
29	FactSalesQuotaCCI	3	3	COMPRESSED	510639	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_29
30	FactSalesQuotaCCI	3	3	COMPRESSED	510366	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_30
31	FactSalesQuotaCCI	3	3	COMPRESSED	511247	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_31
32	FactSalesQuotaCCI	3	3	COMPRESSED	511917	BULKLOAD	Table_0916c48a65e14db285dfed9c57386023_32

Tables – Indexes

Clustered Columnstore index (Default Primary)

Highest level of data compression

Best overall query performance

Clustered index (Primary)

Performant for looking up a single to few rows

Heap (Primary)

Faster loading and landing temporary data

Best for small lookup tables

Nonclustered indexes (Secondary)

Enable ordering of multiple columns in a table

Allows multiple nonclustered on a single table

Can be created on any of the above primary indexes

More performant lookup queries

```
-- Create table with index
CREATE TABLE orderTable
(
   OrderId  INT NOT NULL,
   Date     DATE NOT NULL,
   Name     VARCHAR(2),
   Country  VARCHAR(2)
)
WITH
(
CLUSTERED COLUMNSTORE INDEX |
   HEAP |
   CLUSTERED INDEX (OrderId)
);


-- Add non-clustered index to table
CREATE INDEX NameIndex ON orderTable (Name);
```

Creating tables

```
CREATE TABLE [dbo].[DimStore]
(
    [StoreKey]          int             NOT NULL
,   [GeographyKey]      int             NOT NULL
,   [StoreName]         nvarchar(100)   NOT NULL
,   [StoreType]         nvarchar(15)        NULL
,   [StoreDescription]  nvarchar(300)   NOT NULL
,   [Status]            nvarchar(20)    NOT NULL
,   [OpenDate]          datetime        NOT NULL
,   [CloseDate]         datetime            NULL
,   [ETLLoadID]         int                 NULL
,   [LoadDate]          datetime            NULL
,   [UpdateDate]        datetime            NULL
)
WITH
(   CLUSTERED INDEX([StoreKey])
,   DISTRIBUTION = ROUND_ROBIN
)
;
```

Row

```
CREATE TABLE [dbo].[FactOnlineSales]
(
    [OnlineSalesKey]        int             NOT NULL
,   [DateKey]               datetime        NOT NULL
,   [StoreKey]              int             NOT NULL
,   [ProductKey]            int             NOT NULL
,   [PromotionKey]          int             NOT NULL
,   [CurrencyKey]           int             NOT NULL
,   [CustomerKey]           int             NOT NULL
,   [SalesOrderNumber]      nvarchar(20)    NOT NULL
,   [SalesOrderLineNumber]  int                 NULL
,   [SalesQuantity]         int             NOT NULL
,   [SalesAmount]           money           NOT NULL
)
WITH
(   CLUSTERED COLUMNSTORE INDEX
,   DISTRIBUTION = HASH([ProductKey])
)
;
```

Column

Distribution

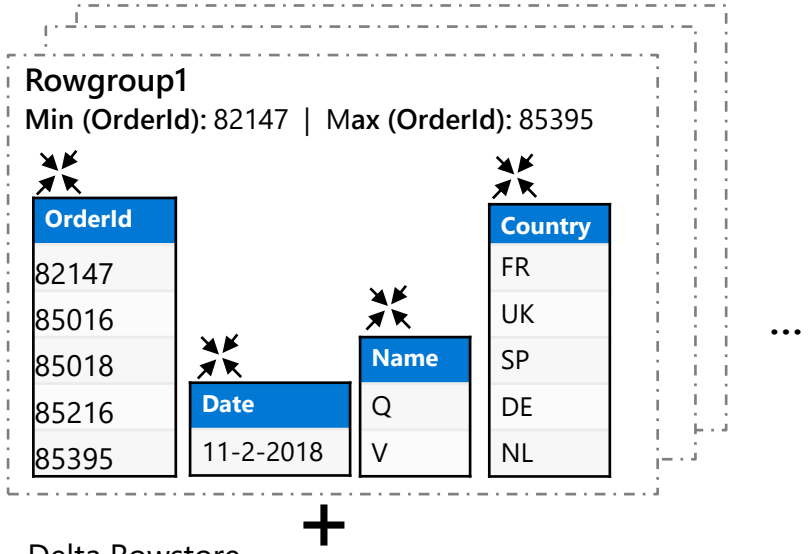
Synapse SQL (provisioned) Columnstore Tables

Logical table structure

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

Clustered columnstore index

(OrderId)



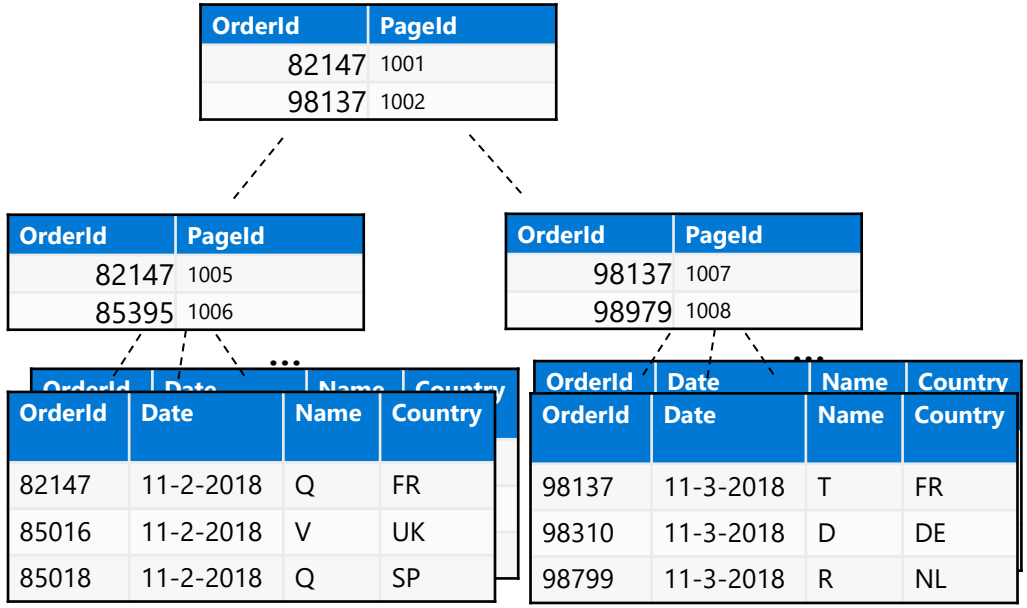
Delta Rowstore

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

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

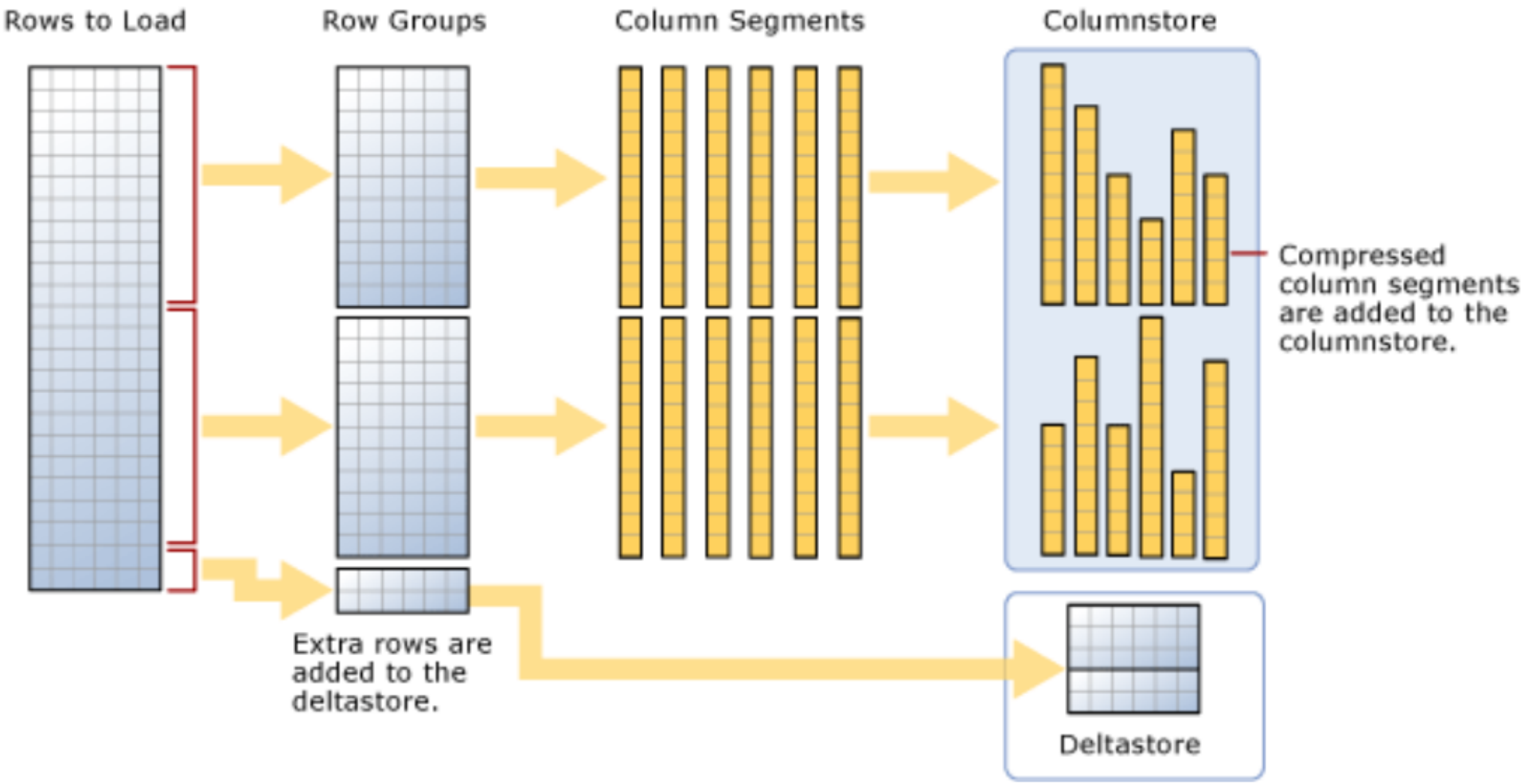
Clustered/Non-clustered rowstore index

(OrderId)



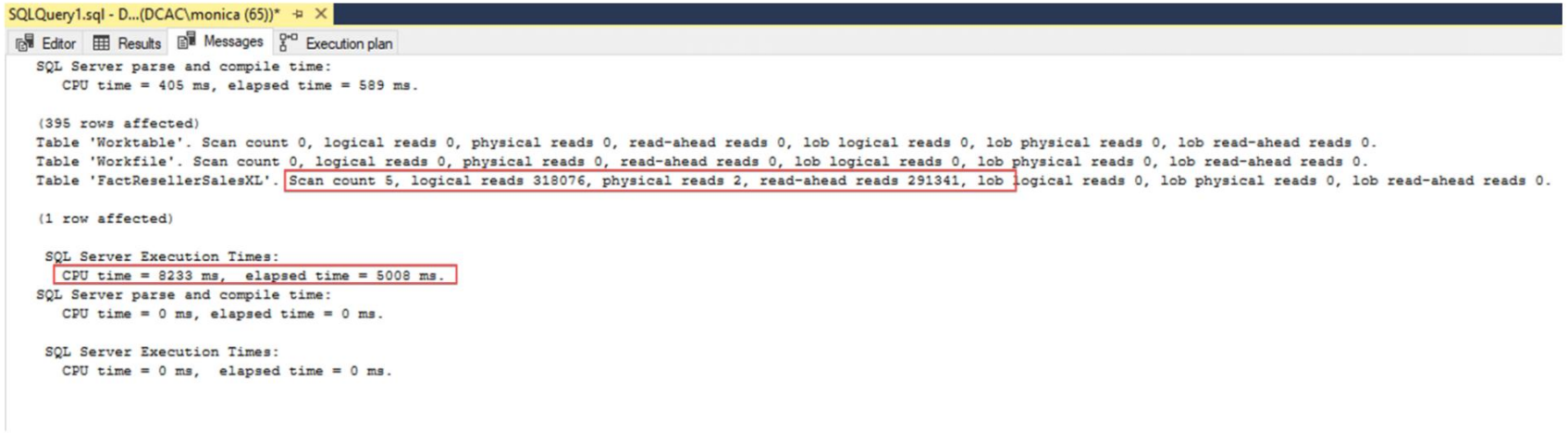
- Data is stored in a B-tree index structure for performant lookup queries for particular rows.
- Clustered rowstore index: The leaf nodes in the structure store the data values in a row (as pictured above)
- Non-clustered (secondary) rowstore index: The leaf nodes store pointers to the data values,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

Columnstore Storage Model

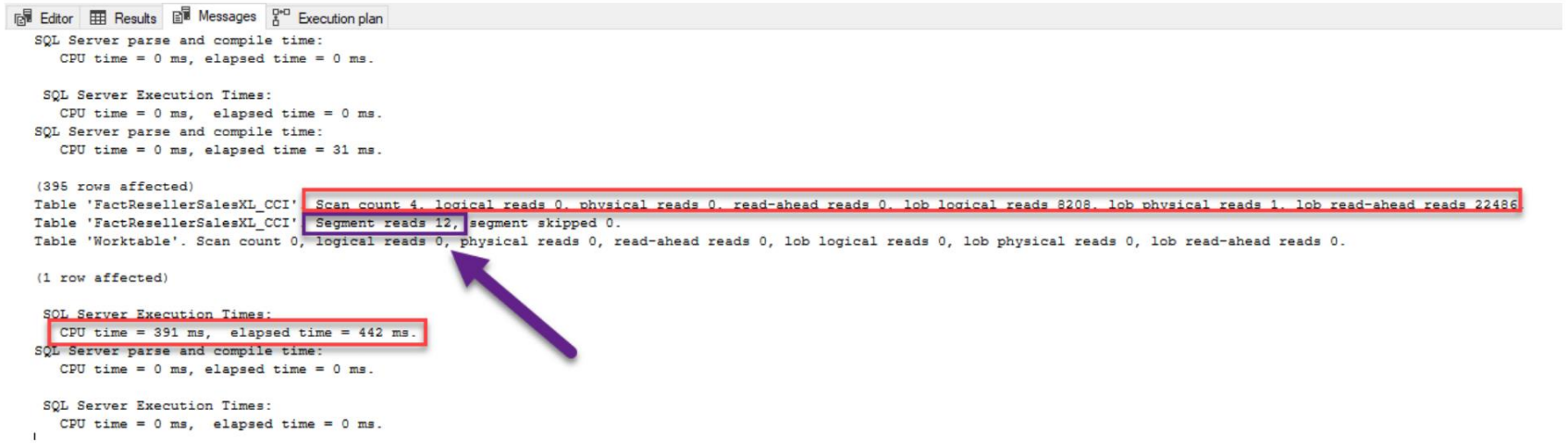


Columnstore Storage Model

ROW STORE



COLUMNSTORE



Columnstore Storage Model

Clustered Index Scan (Clustered)

Scanning a clustered index, entirely or only a range.

Physical Operation	Clustered Index Scan
Logical Operation	Clustered Index Scan
Actual Execution Mode	Row
Estimated Execution Mode	Row
Storage	RowStore
Number of Rows Read	11669638
Actual Number of Rows	11669638
Actual Number of Batches	0
Estimated Operator Cost	240.048 (90%)
Estimated I/O Cost	233.63
Estimated CPU Cost	6.41838
Estimated Subtree Cost	240.048
Number of Executions	4
Estimated Number of Executions	1
Estimated Number of Rows to be Read	11669600
Estimated Number of Rows	11669600
Estimated Row Size	21 B
Actual Rebinds	0
Actual Rewinds	0
Ordered	False
Node ID	7

Object

[AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL].[PK_FactResellerSalesXL_SalesOrderNumber_SalesOrderLineNumber]

Output List

[AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL].ProductKey, [AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL].OrderQuantity, [AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL].SalesAmount

Columnstore Index Scan (Clustered)

Scan a columnstore index, entirely or only a range.

Physical Operation	Columnstore Index Scan
Logical Operation	Clustered Index Scan
Actual Execution Mode	Batch
Estimated Execution Mode	Batch
Storage	ColumnStore
Actual Number of Rows	0
Actual Number of Batches	0
Estimated Operator Cost	1.58719 (36%)
Estimated I/O Cost	0.945347
Estimated CPU Cost	0.641838
Estimated Subtree Cost	1.58719
Number of Executions	4
Estimated Number of Executions	1
Estimated Number of Rows	11669600
Estimated Number of Rows to be Read	11669600
Estimated Row Size	21 B
Actual Rebinds	0
Actual Rewinds	0
Ordered	False
Actual Number of Locally Aggregated Rows	11669638
Node ID	3

Object

[AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL_CCI].[IndFactResellerSalesXL_CCI]

Output List

[AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL_CCI].ProductKey, [AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL_CCI].OrderQuantity, [AdventureworksDW2016CTP3].[dbo].[FactResellerSalesXL_CCI].SalesAmount

Ordered Clustered Columnstore Indexes

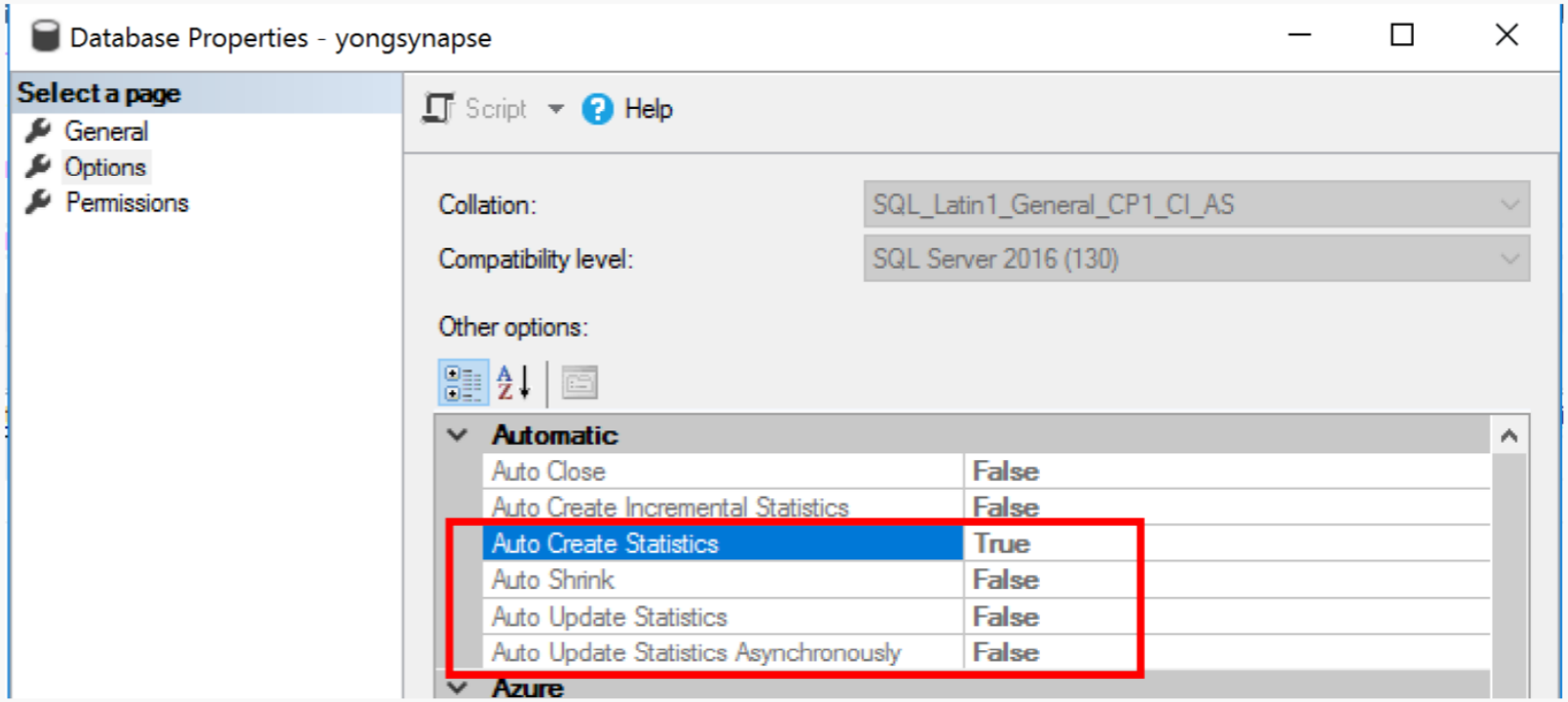
Overview

Queries against tables with ordered columnstore segments can take advantage of improved segment elimination to drastically reduce the time needed to service a query.

```
-- Create Table with Ordered Columnstore Index
CREATE TABLE sortedOrderTable
(
   OrderId  INT NOT NULL,
   Date     DATE NOT NULL,
   Name     VARCHAR(2),
   Country  VARCHAR(2)
)
WITH
(
 CLUSTERED COLUMNSTORE INDEX ORDER (OrderId)
)
-- Create Clustered Columnstore Index on existing table
CREATE CLUSTERED COLUMNSTORE INDEX cciOrderId
ON dbo.OrderTable ORDER (OrderId)
```

```
-- Insert data into table with ordered columnstore index
INSERT INTO sortedOrderTable
VALUES (1, '01-01-2019','Dave', 'UK')
```

Statistics



CREATE STATISTICS [statistics_name] ON [schema_name].[table_name]([column_name]) WITH FULLSCAN;

CREATE STATISTICS col1_stats ON dbo.table1 (col1) WITH SAMPLE = 50 PERCENT;

UPDATE STATISTICS [dbo].[table1] ([stats_col1]);

‘Optimized for Elasticity’ Sizing & Storage tiers

Sizing factors

Database capacity
Tempdb
Concurrency & Memory
Load
Transaction size
Memory management

DWU

Sizing Node Count

As Is System's Node Count

What is Hadoop Worker Node Count?

What is Greenplum DB Total Node Count?

Performance level	Compute nodes
DW100c	1
DW200c	1
DW300c	1
DW400c	1
DW500c	1
DW1000c	2
DW1500c	3
DW2000c	4
DW2500c	5
DW3000c	6
DW5000c	10
DW6000c	12
DW7500c	15
DW10000c	20
DW15000c	30
DW30000c	60

Sizing CPU & Memory

CPU (Hyperthreaded vCore)

Convert to DTU = DWU *9

vCore = DTU / 100 (DW6000 * 9 = 54,000 DTU) = 540 vCore

Memory

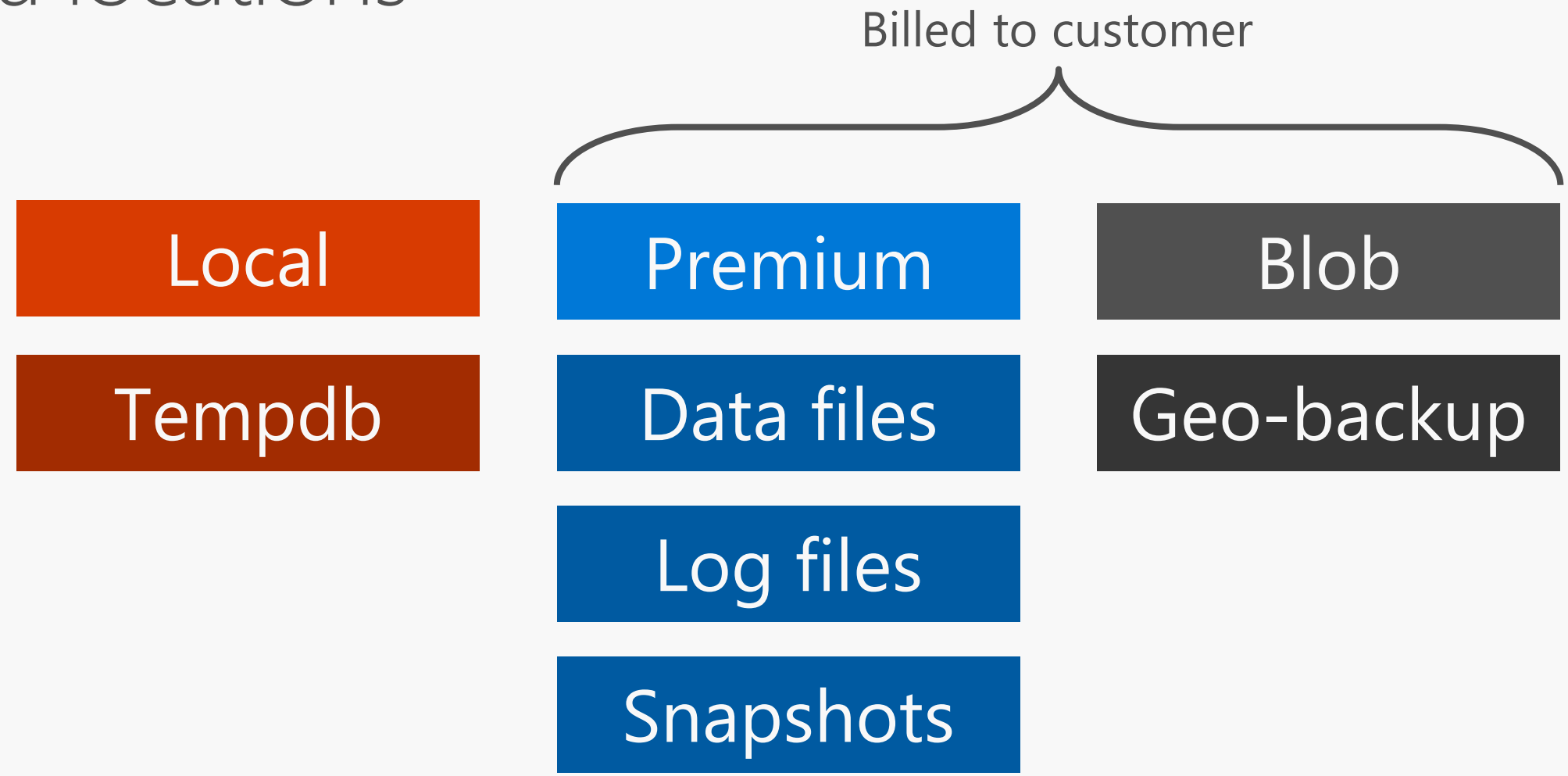
DWU * 60

(DW6000c * 60 = 3.6 Tb)

Performance level	Memory per data warehouse (GB)
DW100c	60
DW200c	120
DW300c	180
DW400c	240
DW500c	300
DW1000c	600
DW1500c	900
DW2000c	1,200
DW2500c	1,500
DW3000c	1,800
DW5000c	3,000
DW6000c	3,600
DW7500c	4,500
DW10000c	6,000
DW15000c	9,000
DW30000c	18,000

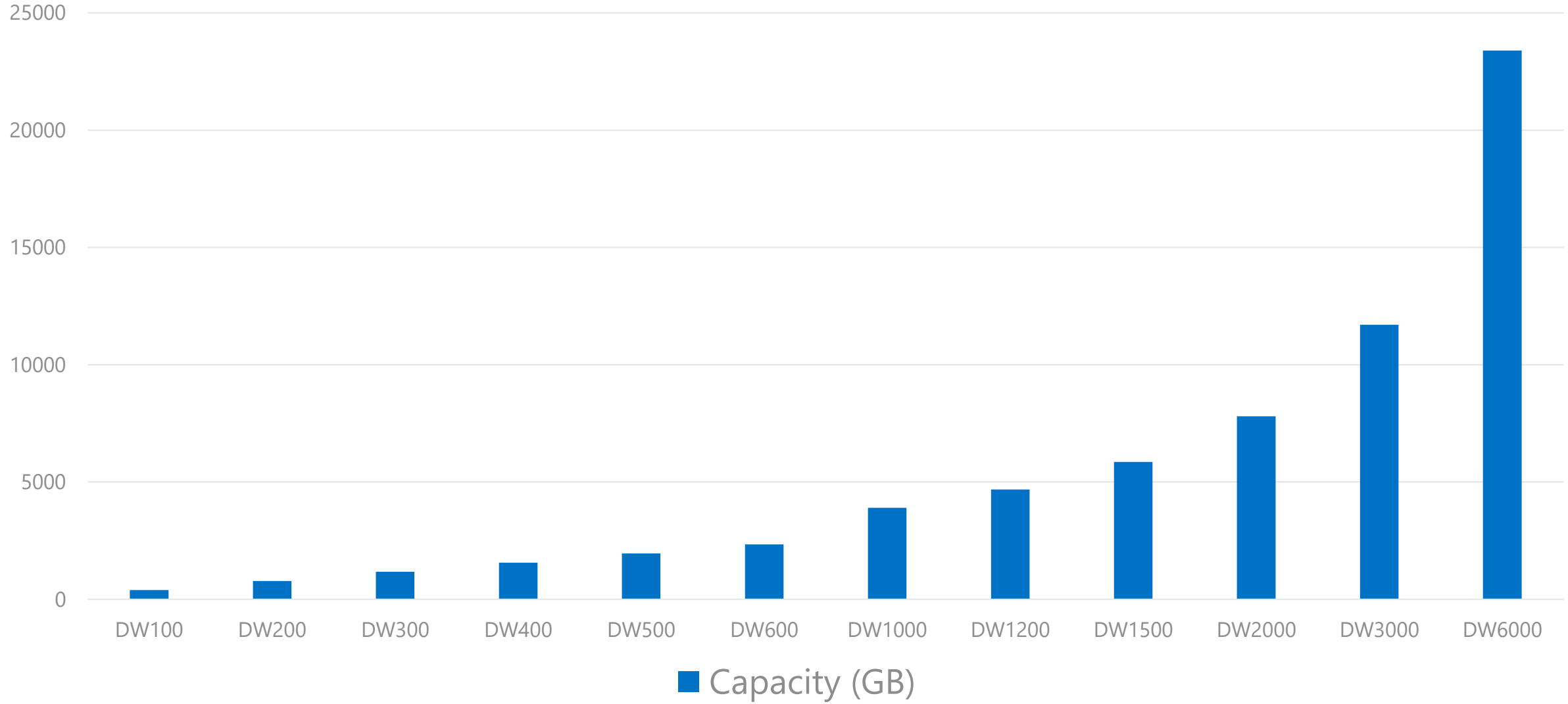
Storage tiers

Data locations

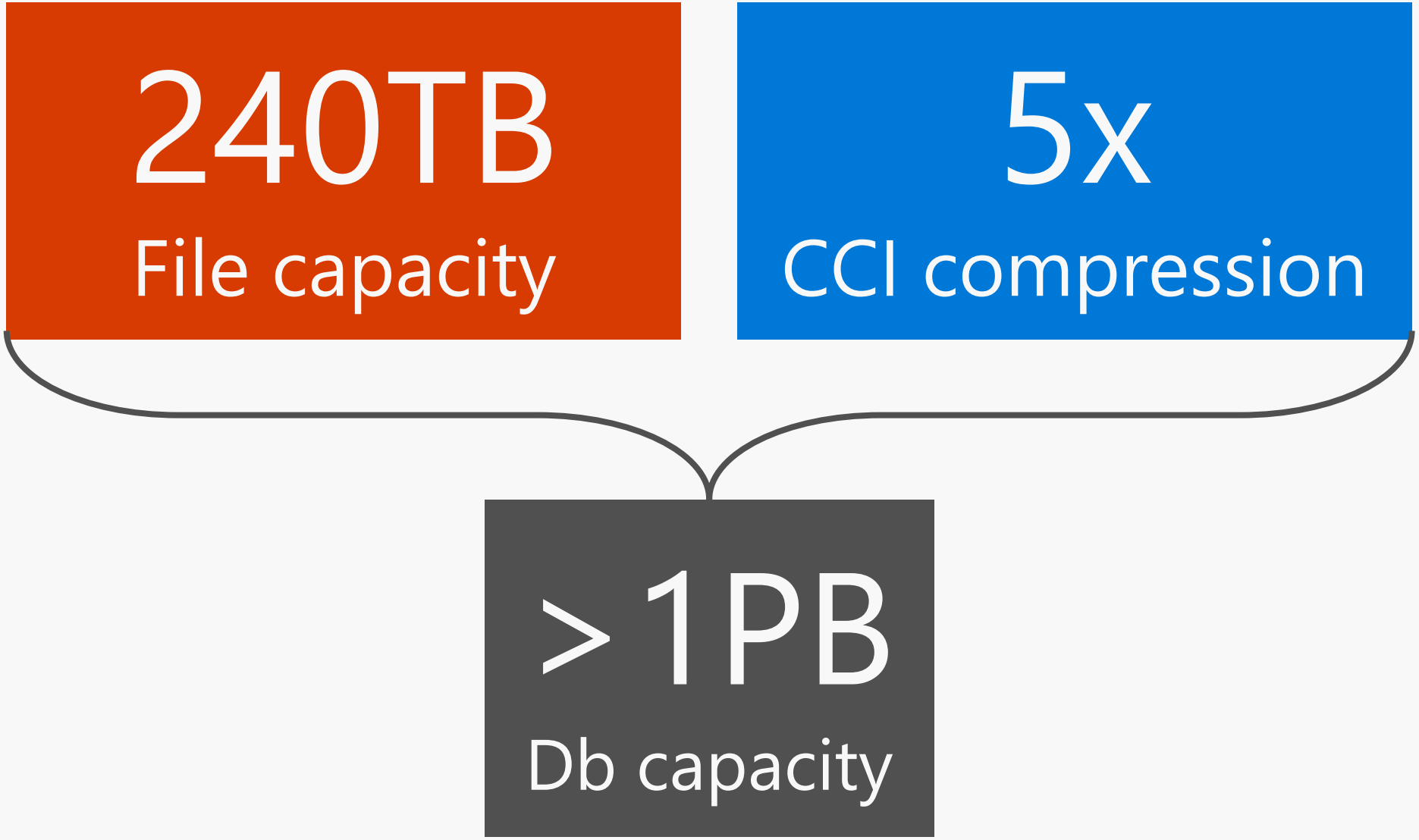


Local Storage: Tempdb sizing

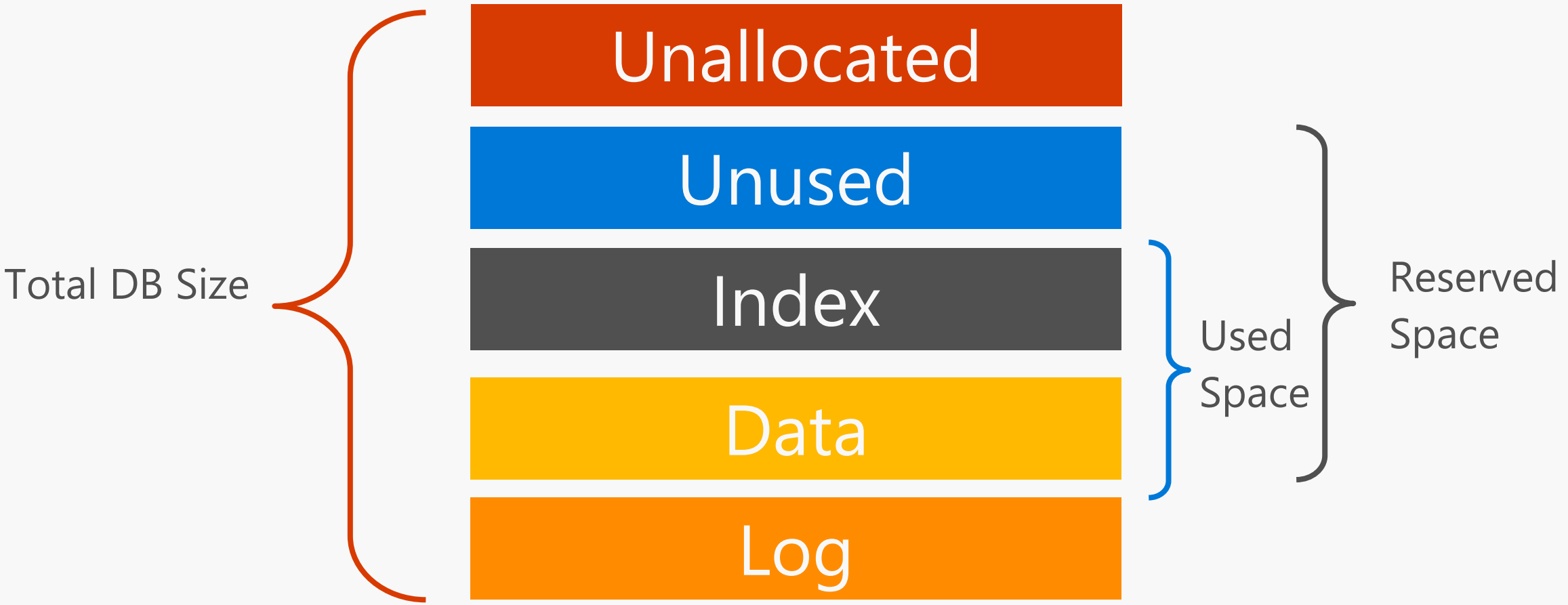
~399GB
per DW100



Premium Storage: Capacity limits



Premium Storage: Database Size



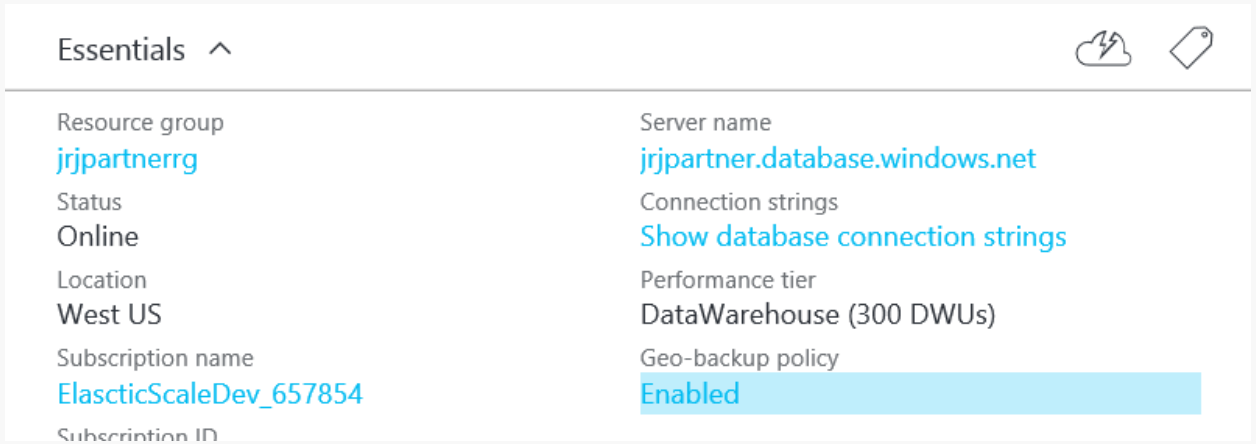
Premium Storage: Snapshots

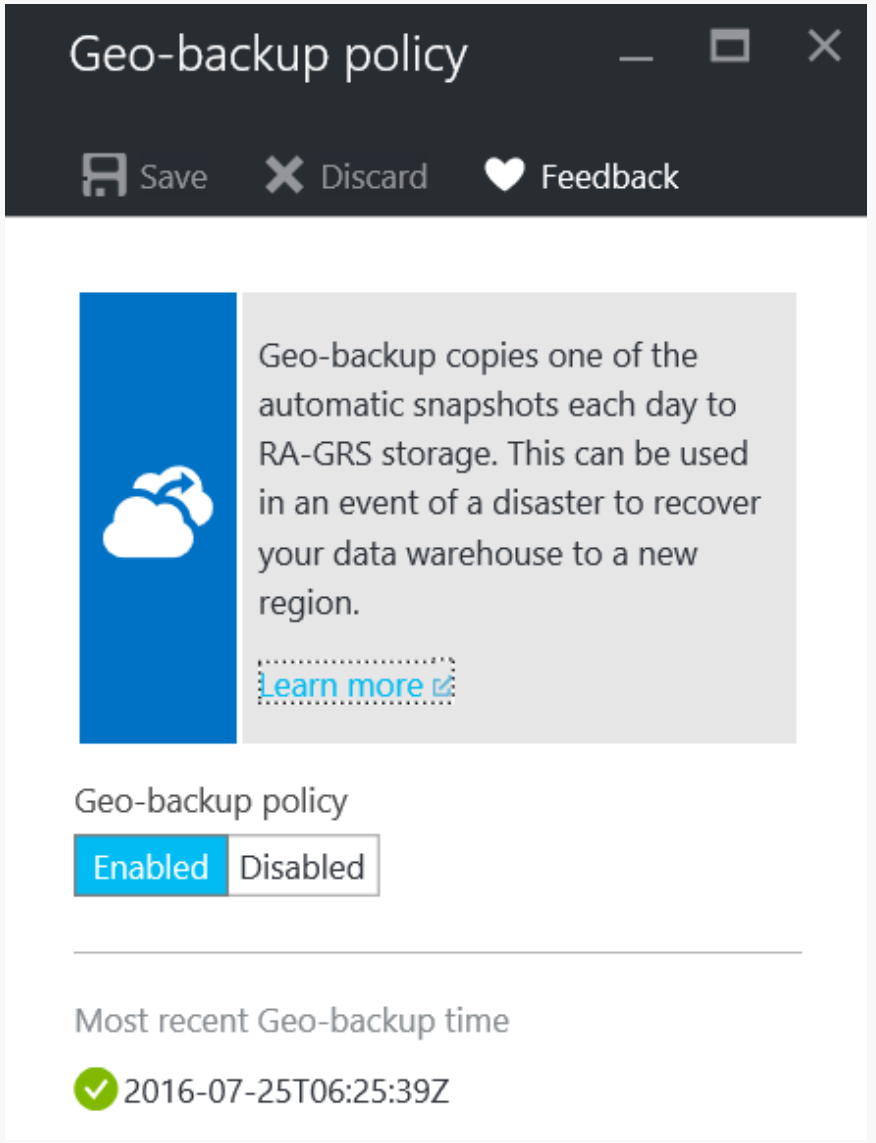
Blob Storage: Geo-redundant backups

Frequency and Retention

1
Geo-backup

24hr
RPO





Loading

Sizing for the data load

Delimited text guidance.

- Evenly split the data into multiple files.
- One file per reader.
- Delimited text is the fastest.

Compressed text limits concurrent access to text files

Split data across files
OR
Use different file format

DWU	Readers	Writers
DW100	8	60
DW200	16	60
DW300	24	60
DW400	32	60
DW500	40	60
DW600	48	60
DW1000	60	60

Data loading

Exception

Target Table = CI or NCI

Load user is defaultrc

60 Writers

TAKE AWAY

Use mediumrc+ for high DWU loads

DWU	Max External Readers	Max Writers
DW100	8	60
DW200	16	60
DW300	24	60
DW400	32	60
DW500	40	60
DW600	48	60
DW1000	80	80
DW1200	96	96
DW1500	120	120
DW2000	160	160
DW3000	240	240
DW6000	480	480

Loading Option

PolyBase	BCP	SSIS	ADF
Fastest and preferred load option. Use CTAS for initial load. Use INSERT/INTO for incremental load or CTAS into stage table and partition switch into final table.	Use only for small files **< 10 GB.** Limited retry logic. Does not scale as you increase DWU (single thread, single CPU on client). Increase parallel threads to improve performance.	Increase client timeout at least 10 min, default 30 sec. Increase parallel threads to improve performance. Slight performance improvement & greater reliability if run on VM.	A simpler way to use PolyBase. Quick to configure since you don't need to define the T-SQL objects.

Polybase Sample

```
CREATE MASTER KEY;

CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential
WITH
    IDENTITY = 'user',
    SECRET =
'dIzoQupXULTk/Dy09gXWbCfP2hD6TJ8H9Lf0ypOT+ubQ3uy//PB9tNxbt04WN47dT0strIHBw83zAkB1bm+Xiw==‘
;

CREATE EXTERNAL DATA SOURCE AzureStorage
WITH (
    TYPE = HADOOP,
    LOCATION = 'wasbs://datacontainer@yongstorage.blob.core.windows.net',
    CREDENTIAL = AzureStorageCredential
);
CREATE EXTERNAL FILE FORMAT TextFile
WITH (
    FORMAT_TYPE = DelimitedText,
    FORMAT_OPTIONS (FIELD_TERMINATOR = ',')
);
```

Polybase Sample

```
CREATE EXTERNAL TABLE dbo.DimDate2External (
    DateId INT NOT NULL,
    CalendarQuarter TINYINT NOT NULL,
    FiscalQuarter TINYINT NOT NULL
)
WITH (
    LOCATION='/datedimension/',
    DATA_SOURCE=AzureStorage,
    FILE_FORMAT=TextFile
);

CREATE TABLE dbo.DimDate2
WITH
(
    CLUSTERED COLUMNSTORE INDEX,
    DISTRIBUTION = ROUND_ROBIN
)
AS
SELECT * FROM [dbo].[DimDate2External];
```

Polybase Sample

```
CREATE EXTERNAL TABLE dbo.DimDate2External (
    DateId INT NOT NULL,
    CalendarQuarter TINYINT NOT NULL,
    FiscalQuarter TINYINT NOT NULL
)
WITH (
    LOCATION='/datedimension/',
    DATA_SOURCE=AzureStorage,
    FILE_FORMAT=TextFile
);

CREATE TABLE dbo.DimDate2
WITH
(
    CLUSTERED COLUMNSTORE INDEX,
    DISTRIBUTION = ROUND_ROBIN
)
AS
SELECT * FROM [dbo].[DimDate2External];
```

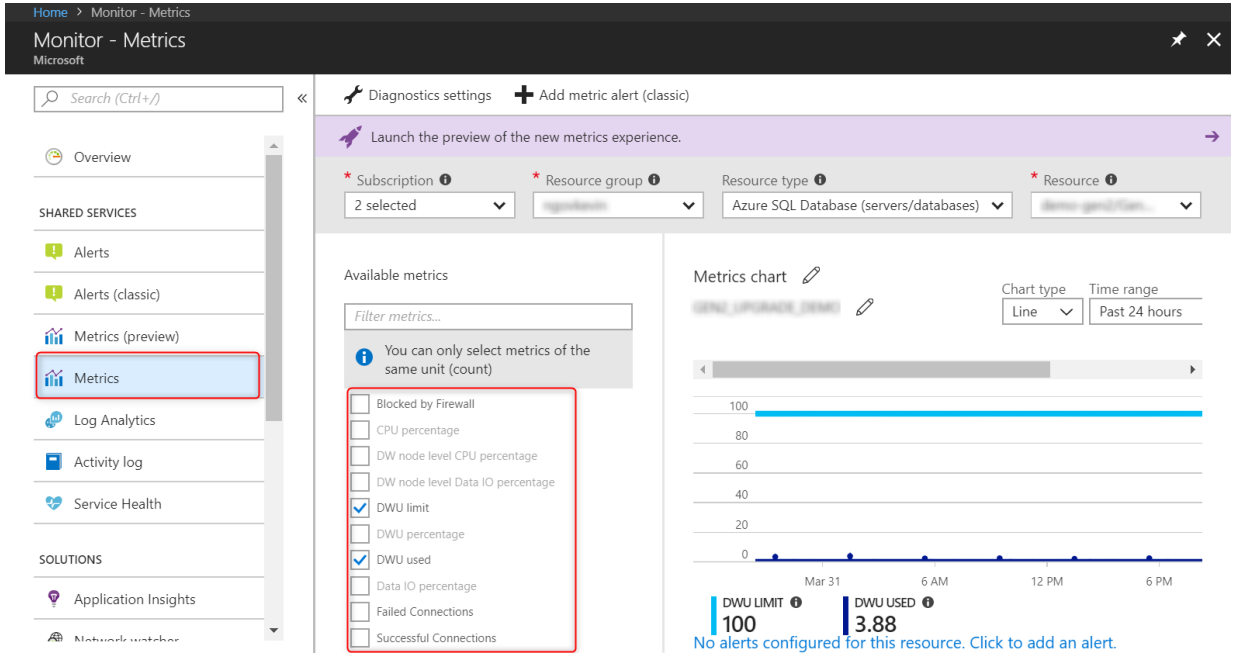
Monitoring

Azure Synapse Monitoring Tools

- Azure Portal
- TSQL and Data Management Views (DMVs)
- Azure Monitor
- Azure Advisors
- SQL Server Management Studio
- Azure Data Studio
- Visual Studio Data Tools

Monitoring using the Portal

- **Overview**
  - Metrics surfaced through Azure Monitor
  - Tracks metrics related to utilization and security
  - Export to Log Analytics for further analysis
  - No cost for first 31 days of each GB of data
- **Benefits**
  - Holistic monitoring across Azure analytics platform
  - Customize charts and create enriched dashboards
  - Identify under or over-utilization
  - Make informed scaling decisions



Metrics

CPU Percentage

Data IO Percentage

Successful Connections

Failed Connections

Blocked Connections

DWU (Compute) Limit

DWU Percentage (Used)

Cache Hit Percentage

Cache Used Percentage

Local tempdb Percentage

Azure Monitor - Alerts

Overview

Set alerts on metrics or activity logs.

Define metric thresholds.

Track individual or count of log events.

Send emails or call webhooks on triggers.

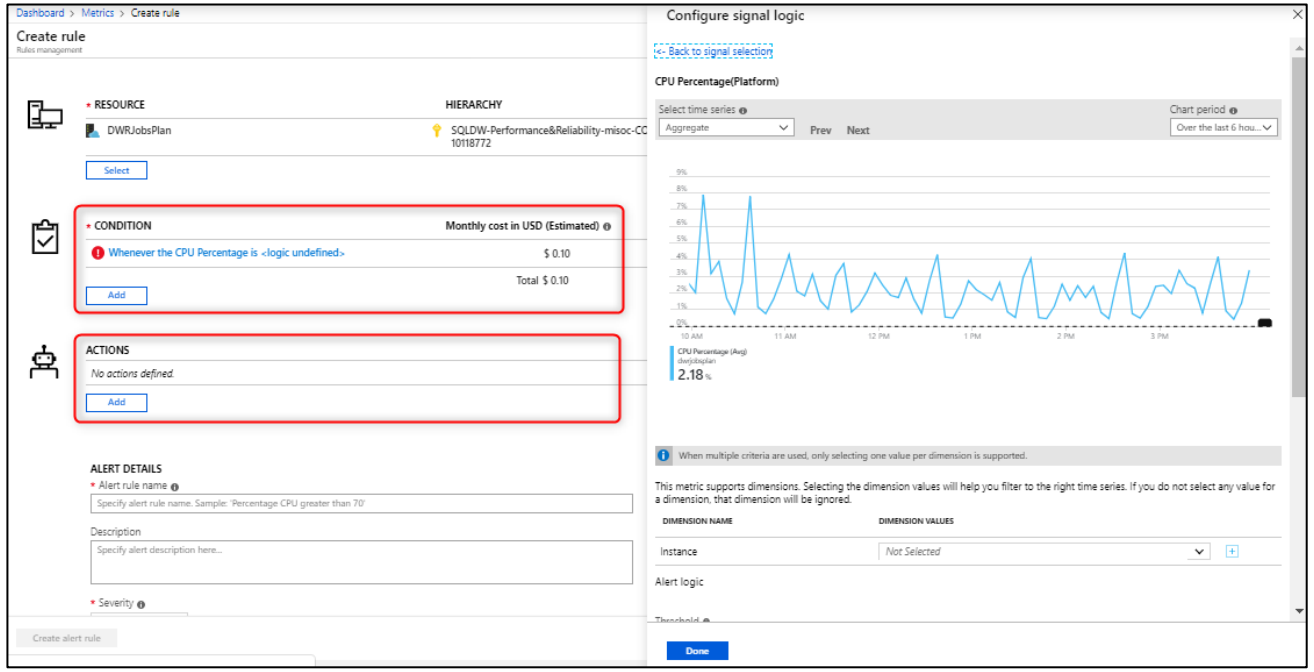
Manage with Azure portal, PowerShell, CLI, and REST APIs.

Benefits

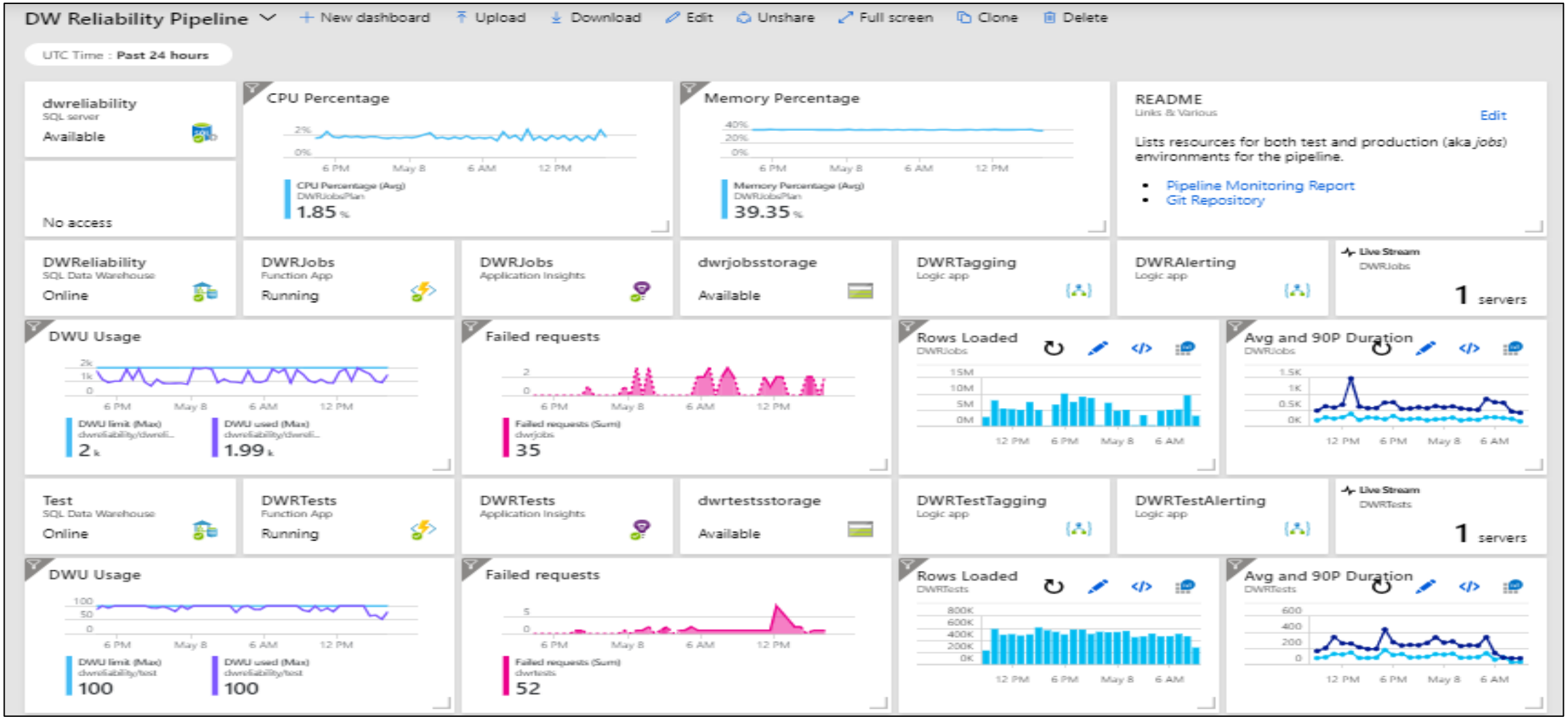
Automatically track data warehouse state.

Customizable triggers for immediate action.

Integration with SQL Data Warehouse monitoring.



Azure Dashboards



SQL Pool – Database Auditing

Configure Auditing

- Server-level vs. database-level auditing
- Default auditing policy includes all actions plus

  BATCH_COMPLETED_GROUP

  SUCCESSFUL_DATABASE_AUTHENTICATION_GROUP

  FAILED_DATABASE_AUTHENTICATION_GROUP
- Use PowerShell or RestAPI to customize the audited events

Analyze Audit Logs and reports

- **Azure Monitor Logs**
  - Auditing blade – View audit logs
  - Open in OMS
  - Log Analytics blade
- **Event Hub**
- **Azure Storage Account**
  - Azure Storage Explorer
  - Auditing blade – View audit logs
  - System function sys.fn_get_audit_file
  - SSMS
  - PowerBI
  - PowerShell (query extended events files)

Dynamic Management Views (DMVs)

Overview

Dynamic Management Views (DMV) are queries that return information about model objects, server operations, and server health.

Benefits:

Simple SQL syntax

Returns result in table format

Easier to read and copy result

SQL Monitor with DMVs

Overview

Offers monitoring of

-all open, closed sessions

-count sessions by user

-count completed queries by user

-all active, complete queries

-longest running queries

-memory consumption

Count sessions by user

```
--count sessions by user
SELECT login_name, COUNT(*) as session_count FROM sys.dm_pdw_exec_sessions
where status = 'Closed' and session_id <> session_id() GROUP BY login_name;
```

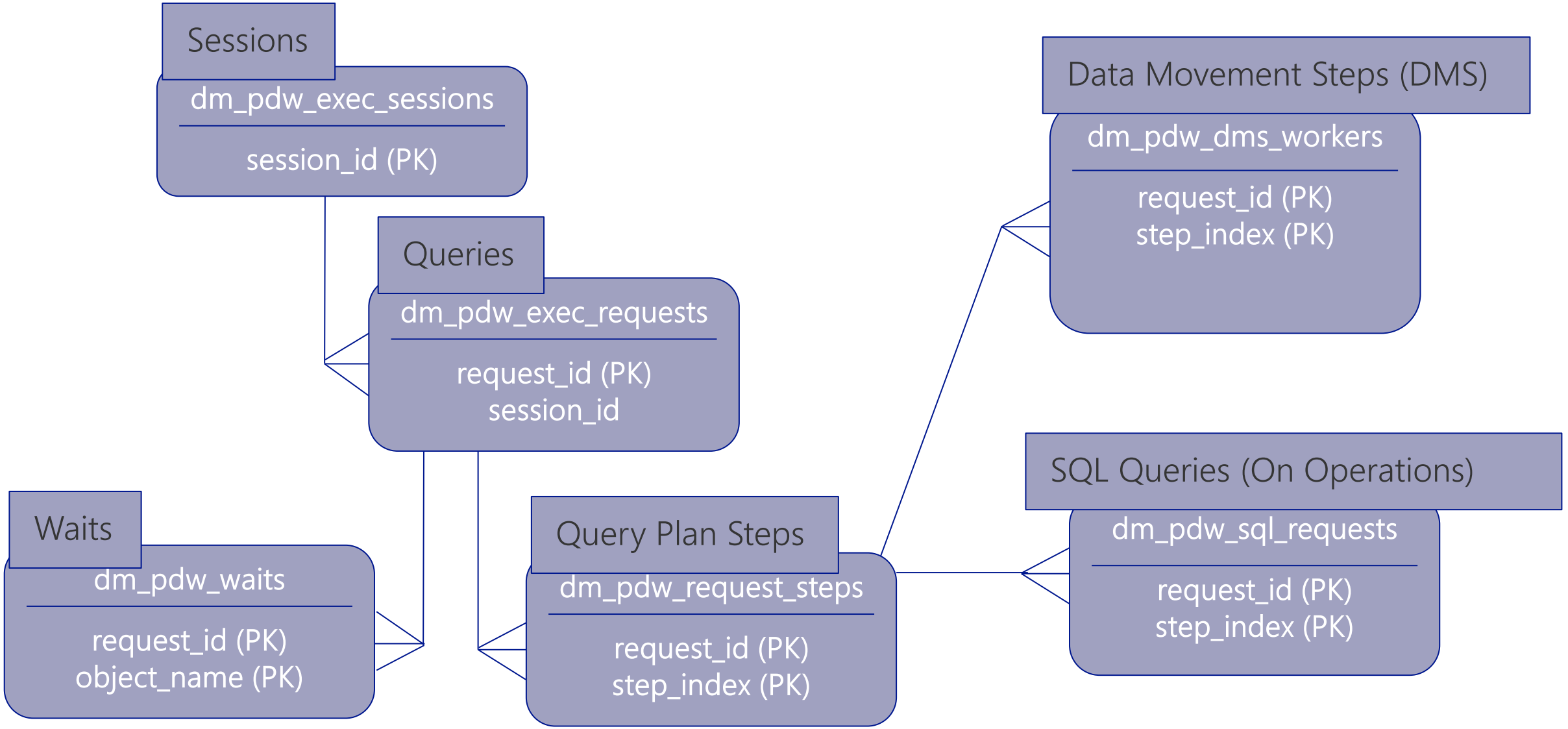
List all open sessions

```
-- List all open sessions
SELECT * FROM sys.dm_pdw_exec_sessions where status <> 'Closed' and session_id <>
session_id();
```

List all active queries

```
-- List all active queries
SELECT * FROM sys.dm_pdw_exec_requests WHERE status not in
('Completed','Failed','Cancelled') AND session_id <> session_id() ORDER BY submit_time
DESC;
```

Monitoring using DMVs



Query Tuning Point Eexplorations (Session)

```
SELECT * FROM sys.dm_pdw_exec_sessions where status <> 'Closed' and session_id <> session_id();
```

	session_id	status	request_id	security_id	login_name	login_time	query_count
1	SID198	Idle	NULL	NULL	MedRCLogin	2017-11-03 08:02:57.973	1
2	SID196	Idle	NULL	NULL	yongkim	2017-11-03 08:01:24.893	1
3	SID158	Idle	NULL	NULL	yongkim	2017-11-03 06:18:31.310	20
4	SID178	Idle	NULL	NULL	yongkim	2017-11-03 07:15:56.457	12

is_transactional	client_id	app_name	sql_spid
0	220.88.36.134:6329	Microsoft SQL Server Management Studio	123
0	220.88.36.134:47048	Microsoft SQL Server Management Studio	116
0	220.88.36.134:6776	Microsoft SQL Server Management Studio - Query	134
0	220.88.36.134:30120	Microsoft SQL Server Management Studio - Query	183

Query Tuning Point Eexploration (Long Running Query)

```
SELECT TOP 10 * FROM sys.dm_pdw_exec_requests  ORDER BY total_elapsed_time DESC;
```

	request_id	session_id	status	submit_time	start_time	end_compile_time	end_time
1	QID1505	SID178	Completed	2017-11-03 07:19:35.193	2017-11-03 07:19:35.380	2017-11-03 07:19:35.380	2017-11-03 07:19:43.270
2	QID1507	SID178	Completed	2017-11-03 07:20:23.543	2017-11-03 07:20:24.107	2017-11-03 07:20:24.090	2017-11-03 07:20:27.923
3	QID1439	SID157	Completed	2017-11-03 06:18:14.437	2017-11-03 06:18:14.873	2017-11-03 06:18:14.873	2017-11-03 06:18:17.437
4	QID1468	SID158	Completed	2017-11-03 06:30:51.940	2017-11-03 06:30:54.110	2017-11-03 06:30:54.110	2017-11-03 06:30:54.203
5	QID1440	SID157	Completed	2017-11-03 06:18:14.967	2017-11-03 06:18:16.703	2017-11-03 06:18:16.670	2017-11-03 06:18:17.093
6	QID1560	SID194	Failed	2017-11-03 08:08:48.967	NULL	NULL	2017-11-03 08:08:49.720
7	QID1459	SID158	Completed	2017-11-03 06:18:35.593	2017-11-03 06:18:36.030	2017-11-03 06:18:36.030	2017-11-03 06:18:36.280
8	QID1553	SID194	Completed	2017-11-03 08:05:06.300	2017-11-03 08:05:06.377	2017-11-03 08:05:06.377	2017-11-03 08:05:06.753
9	QID1453	SID159	Completed	2017-11-03 06:18:32.390	2017-11-03 06:18:32.390	2017-11-03 06:18:32.390	2017-11-03 06:18:32.810
10	QID1542	SID194	Completed	2017-11-03 08:02:17.863	2017-11-03 08:02:17.910	2017-11-03 08:02:17.910	2017-11-03 08:02:18.223

total_elap...	label	error_id	database_id	command	resource_class
8078	NULL	NULL	9	CREATE TABLE myTable (id int NOT NULL, ...	NULL
4379	NULL	NULL	9	CREATE INDEX IX_myTable ON myTAble (zipCode)	smallrc
2999	NULL	NULL	9	DECLARE @edition sysname; SET @edition = cast(SERV...	NULL
2265	NULL	NULL	9	DBCC SHOW_STATISTICS ('dbo.DimDate2','DateId')	NULL
2125	NULL	NULL	9	SELECT (SERVERPROPERTY(N'EDITION'))	NULL
750	NULL	9c9deb58-aad...	9	status not in ('Completed','Failed','Cancelled') AND	NULL
687	NULL	NULL	9	SELECT sm.[name] AS [schema_name], tb.[name] A...	NULL
453	NULL	NULL	9	-- Other Active Connections SELECT * FROM sys.dm_pd...	NULL
421	NULL	NULL	9	DECLARE @edition sysname; SET @edition = cast(SERV...	NULL
359	NULL	NULL	9	CREATE USER viewuser FOR LOGIN viewlogin; GRANT...	NULL

Query Tuning Point Eexploration (Request Steps and SQL Request)

SELECT * FROM sys.**dm_pdw_request_steps** WHERE request_id = 'QID1592' ORDER BY step_index;

	request_id	step_index	operation_type	distribution_type	location_type	status	error_id	start_time	end_time
1	QID1592	0	ReturnOperation	AllDistributions	Compute	Complete	NULL	2017-11-03 08:17:36.883	2017-11-03 08:17:36.960

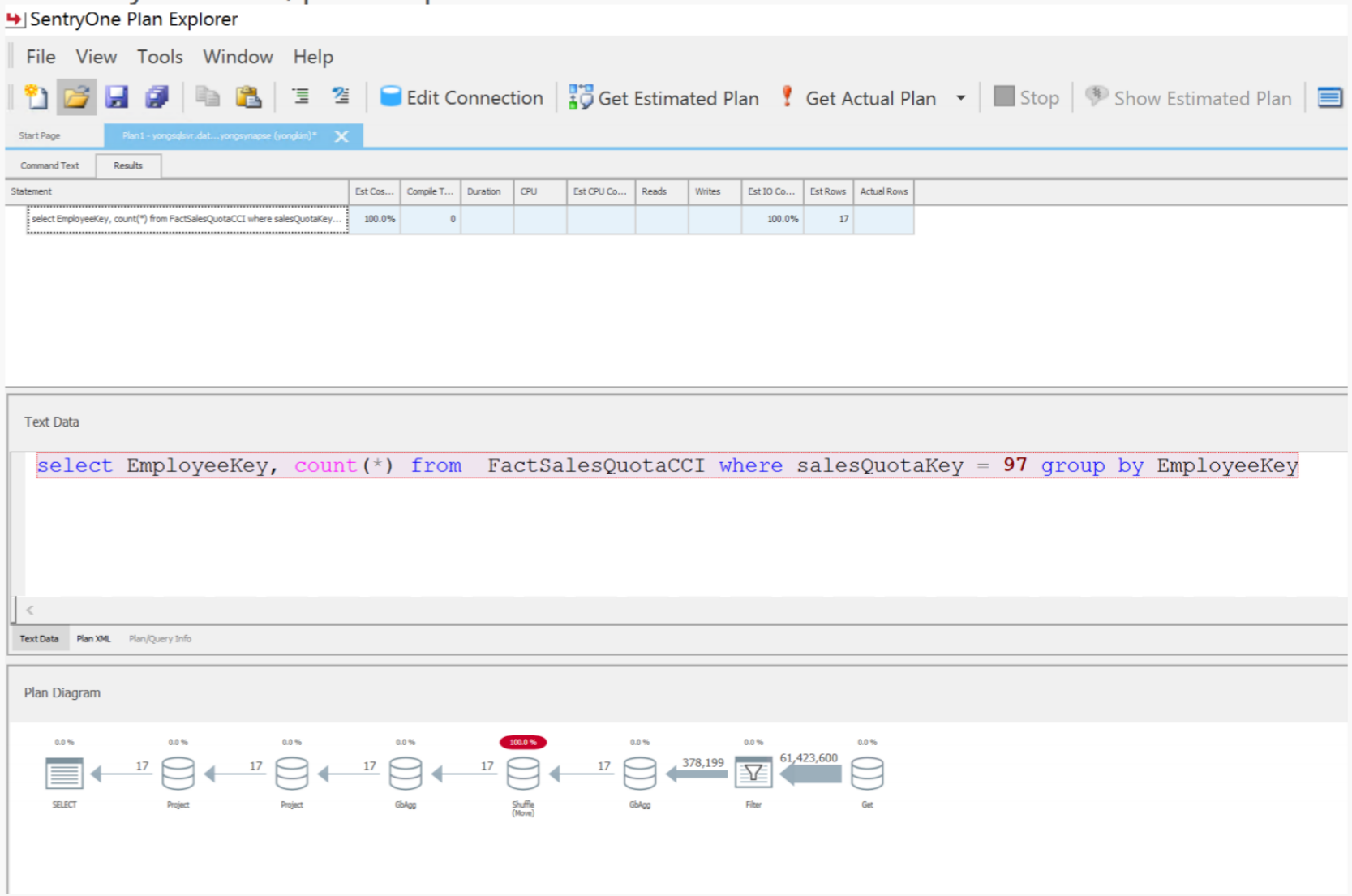
total_elapsed_time	row_count	command
78	-1	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter],...

SELECT * FROM sys.**dm_pdw_sql_requests** WHERE request_id = 'QID1592' AND step_index = **0**;

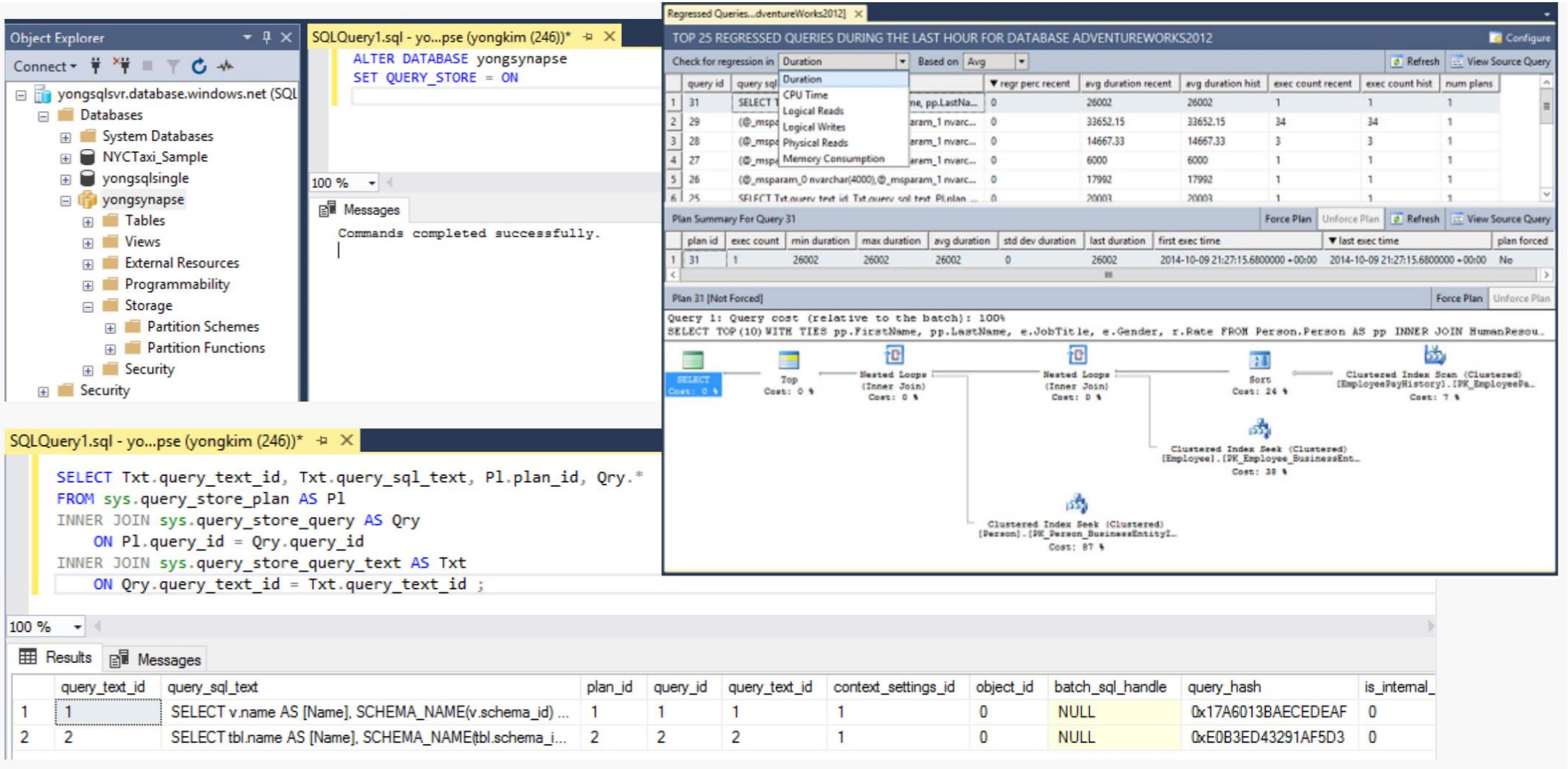
	request_id	step_index	pdw_node_id	distribution_id	status	error_id	start_time	end_time	total_elapsed_time	row_count	spid	command
1	QID1592	0	1	1	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2540	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
2	QID1592	0	1	2	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2632	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
3	QID1592	0	1	3	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2539	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
4	QID1592	0	1	4	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2618	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
5	QID1592	0	1	5	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2536	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
6	QID1592	0	1	6	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2627	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
7	QID1592	0	1	7	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2616	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
8	QID1592	0	1	8	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2615	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
9	QID1592	0	1	9	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2591	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
10	QID1592	0	1	10	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2624	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
11	QID1592	0	1	11	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2458	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
12	QID1592	0	1	12	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2629	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
13	QID1592	0	1	13	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2538	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
14	QID1592	0	1	14	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2502	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
15	QID1592	0	1	15	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2554	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
16	QID1592	0	1	16	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2563	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
17	QID1592	0	1	17	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2619	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
18	QID1592	0	1	18	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2427	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
19	QID1592	0	1	19	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2557	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...
20	QID1592	0	1	20	Complete	NULL	2017-11-03 08:17:36.947	2017-11-03 08:17:36.947	0	-1	2621	SELECT [T1_1].[DateId] AS [DateId], [T1_1].[CalendarQuarter] AS [CalendarQuarter], [T1...

Query Plan

https://www.sentryone.com/plan-explorer



Query Store Enable



Grafana

Grafana Support

Grafana Labs

Graf

All dashboards » Synapse

Synapse by fg

DASHBOARD

Synapse dashboard from https://github.com/matrix-org/synapse/blob/master/contrib/grafana/synapse.json

Last updated: a year ago

Overview Revisions Reviews

Window functions

OVER clause

Defines a window or specified set of rows within a query result set

Computes a value for each row in the window

Aggregate functions

COUNT, MAX, AVG, SUM, APPROX_COUNT_DISTINCT, MIN, STDEV, STDEVP, STRING_AGG, VAR, VARP, GROUPING, GROUPING_ID, COUNT_BIG, CHECKSUM_AGG

Ranking functions

RANK, NTILE, DENSE_RANK, ROW_NUMBER

Analytical functions

LAG, LEAD, FIRST_VALUE, LAST_VALUE, CUME_DIST, PERCENTILE_CONT, PERCENTILE_DISC, PERCENT_RANK

ROWS | RANGE

PRECEDING, UNBOUNDING PRECEDING, CURRENT ROW, BETWEEN, FOLLOWING, UNBOUNDED FOLLOWING

```
SELECT
  ROW_NUMBER() OVER(PARTITION BY PostalCode ORDER BY SalesYTD DESC) AS "Row Number",
  LastName,
  SalesYTD,
  PostalCode
FROM Sales
WHERE SalesYTD <> 0
ORDER BY PostalCode;
```

Row Number	LastName	SalesYTD	PostalCode
1	Mitchell	4251368.5497	98027
2	Blythe	3763178.1787	98027
3	Carson	3189418.3662	98027
4	Reiter	2315185.611	98027
5	Vargas	1453719.4653	98027
6	Ansman-Wolfe	1352577.1325	98027
1	Pak	4116870.2277	98055
2	Varkey Chudukaktil	3121616.3202	98055
3	Saraiva	2604540.7172	98055
4	Ito	2458535.6169	98055
5	Valdez	1827066.7118	98055
6	Mensa-Annan	1576562.1966	98055
7	Campbell	1573012.9383	98055
8	Tsoflias	1421810.9242	98055

Window Functions (continued)

Analytical functions

LAG, LEAD, FIRST_VALUE, LAST_VALUE, CUME_DIST, PERCENTILE_CONT, PERCENTILE_DISC, PERCENT_RANK

```
-- PERCENTILE_CONT, PERCENTILE_DISC
 SELECT DISTINCT Name AS DepartmentName
,PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY ph.Rate)
                OVER (PARTITION BY Name) AS MedianCont
,PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY ph.Rate)
                OVER (PARTITION BY Name) AS MedianDisc
FROM HumanResources.Department AS d
INNER JOIN HumanResources.EmployeeDepartmentHistory AS dh
   ON dh.DepartmentID = d.DepartmentID
INNER JOIN HumanResources.EmployeePayHistory AS ph
   ON ph.BusinessEntityID = dh.BusinessEntityID
WHERE dh.EndDate IS NULL;
```

DepartmentName	MedianCont	MedianDisc
Document Control	16.8269	16.8269
Engineering	34.375	32.6923
Executive	54.32695	48.5577
Human Resources	17.427850	16.5865

```
--LAG Function
SELECT BusinessEntityID,
YEAR(QuotaDate) AS SalesYear,
SalesQuota AS CurrentQuota,
LAG(SalesQuota, 1,0) OVER (ORDER BY YEAR(QuotaDate)) AS PreviousQuota
FROM Sales.SalesPersonQuotaHistory
WHERE BusinessEntityID = 275 and YEAR(QuotaDate) IN ('2005','2006');
```

BusinessEntityID	SalesYear	CurrentQuota	PreviousQuota
275	2005	367000.00	0.00
275	2005	556000.00	367000.00
275	2006	502000.00	556000.00
275	2006	550000.00	502000.00
275	2006	1429000.00	550000.00
275	2006	1324000.00	1429000.00

Window Functions (continued)

ROWS | RANGE

PRECEDING, UNBOUNDING PRECEDING, CURRENT ROW, BETWEEN, FOLLOWING, UNBOUNDED FOLLOWING

```
-- First_Value
SELECT JobTitle, LastName, VacationHours AS VacHours,
FIRST_VALUE(LastName) OVER (PARTITION BY JobTitle
ORDER BY VacationHours ASC ROWS UNBOUNDED PRECEDING ) AS FewestVacHours
FROM HumanResources.Employee AS e
INNER JOIN Person.Person AS p
ON e.BusinessEntityID = p.BusinessEntityID
ORDER BY JobTitle;
```

JobTitle	LastName	VacHours	FewestVacHours
Accountant	Moreland	58	Moreland
Accountant	Seamans	59	Moreland
Accounts Manager	Liu	57	Liu
Accounts Payable Specialist	Tomic	63	Tomic
Accounts Payable Specialist	Sheperdigian	64	Tomic
Accounts Receivable Specialist	Poe	60	Poe
Accounts Receivable Specialist	Spoon	61	Poe
Accounts Receivable Specialist	Walton	62	Poe

Group by options

Group by with rollup

Creates a group for each combination of column expressions.

Rolls up the results into subtotals and grand totals

Calculate the aggregates of hierarchical data

Grouping sets

Combine multiple GROUP BY clauses into one GROUP BY CLAUSE.

Equivalent of UNION ALL of specified groups.

```
-- GROUP BY SETS Example --
SELECT Country,
SUM(Sales) AS TotalSales
FROM Sales
GROUP BY GROUPING SETS ( Country, () );
```

```
-- GROUP BY ROLLUP Example --
SELECT Country,
Region,
SUM(Sales) AS TotalSales
FROM Sales
GROUP BY ROLLUP (Country, Region);
-- Results --
```


Country	Region	TotalSales
Canada	Alberta	100
Canada	British Columbia	500
Canada	NULL	600
United States	Montana	100
United States	NULL	100
NULL	NULL	700

JSON data support – read JSON data

Overview

Read JSON data stored in a string column with the following:

- ISJSON – verify if text is valid JSON
- JSON_VALUE – extract a scalar value from a JSON string
- JSON_QUERY – extract a JSON object or array from a JSON string

Benefits

Ability to get standard columns as well as JSON column

Perform aggregation and filter on JSON values

```
-- Return all rows with valid JSON data
SELECT CustomerId, OrderDetails
FROM   CustomerOrders
WHERE  ISJSON(OrderDetails) > 0;
```

CustomerId	OrderDetails
101	N'[{ StoreId": "AW73565", "Order": { "Number":"SO43659", "Date":"2011-05-31T00:00:00“ }, "Item": { "Price":2024.40, "Quantity":1 }}]'

```
-- Extract values from JSON string
SELECT CustomerId,
      Country,
      JSON_VALUE(OrderDetails,'$.StoreId') AS StoreId,
      JSON_QUERY(OrderDetails,'$.Item') AS ItemDetails
FROM   CustomerOrders;
```

CustomerId	Country	StoreId	ItemDetails
101	Bahrain	AW73565	{ "Price":2024.40, "Quantity":1 }

Approximate execution

HyperLogLog accuracy

Will return a result with a 2% accuracy of true cardinality on average.

e.g. COUNT (DISTINCT) returns 1,000,000, HyperLogLog will return a value in the range of 999,736 to 1,016,234.

APPROX_COUNT_DISTINCT

Returns the approximate number of unique non-null values in a group.

Use Case: Approximating web usage trend behavior

```
-- Syntax
APPROX_COUNT_DISTINCT ( expression )

-- The approximate number of different order keys by order status from the orders table.
SELECT O_OrderStatus, APPROX_COUNT_DISTINCT(O_OrderKey) AS Approx_Distinct_OrderKey
FROM dbo.Orders
GROUP BY O_OrderStatus
ORDER BY O_OrderStatus;
```

Approximate execution

APPROX_COUNT_DISTINCT

```
SELECT APPROX_COUNT_DISTINCT([SalesOrderDetailID]) AS Approx_Distinct_OrderKey
FROM [SalesLT].[SalesOrderDetail]
```

100 %

Results | Messages

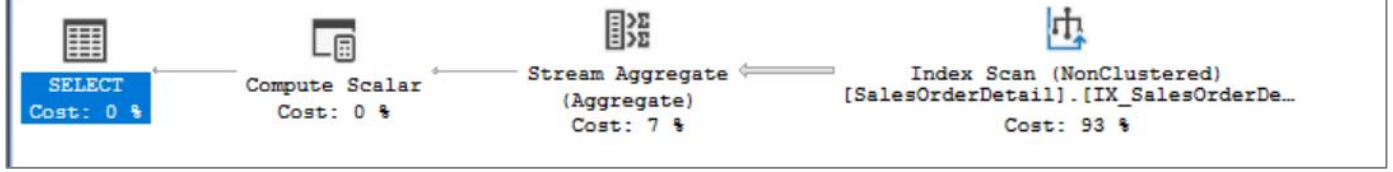
	Approx_Distinct_OrderKey
1	540

```
SELECT APPROX_COUNT_DISTINCT([SalesOrderDetailID]) AS Approx_Distinct_OrderKey
FROM [SalesLT].[SalesOrderDetail]
```

100 %

Messages | Execution plan

```
Query 1: Query cost (relative to the batch): 100%
SELECT APPROX_COUNT_DISTINCT([SalesOrderDetailID]) AS Approx_Distinct_OrderKey FROM [S
```



COUNT DISTINCT

```
SELECT COUNT (DISTINCT [SalesOrderDetailID]) AS Distinct_OrderKey
FROM [SalesLT].[SalesOrderDetail]
```

100 %

Results | Messages

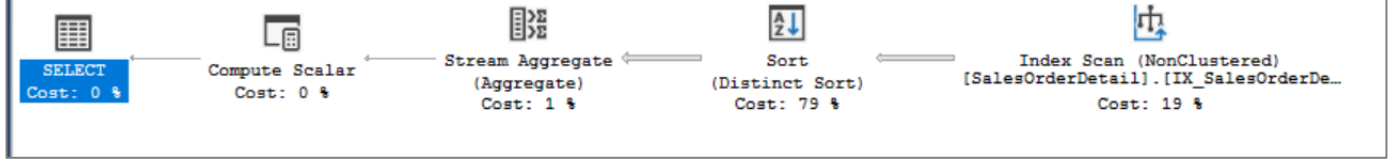
	Distinct_OrderKey
1	542

```
SELECT COUNT (DISTINCT [SalesOrderDetailID]) AS Distinct_OrderKey
FROM [SalesLT].[SalesOrderDetail]
```

100 %

Messages | Execution plan

```
Query 1: Query cost (relative to the batch): 100%
SELECT COUNT (DISTINCT [SalesOrderDetailID]) AS Distinct_OrderKey FROM [SalesLT].[SalesOrderDetail]
```



Snapshot isolation

Overview

Specifies that statements cannot read data that has been modified but not committed by other transactions.

This prevents dirty reads.

```
ALTER DATABASE MyDatabase
SET ALLOW_SNAPSHOT_ISOLATION ON
```

```
ALTER DATABASE MyDatabase SET
READ_COMMITTED_SNAPSHOT ON
```

Isolation level

- READ COMMITTED
- REPEATABLE READ
- SNAPSHOT
- READ UNCOMMITTED
- SERIALIZABLE

READ_COMMITTED_SNAPSHOT

OFF (Default) – Uses shared locks to prevent other transactions from modifying rows while running a read operation

ON – Uses row versioning to present each statement with a transactionally consistent snapshot of the data as it existed at the start of the statement. Locks are not used to protect the data from updates.

JSON data support – insert JSON data

Overview

The JSON format enables representation of complex or hierarchical data structures in tables.

JSON data is stored using standard NVARCHAR table columns.

Benefits

Transform arrays of JSON objects into table format

Performance optimization using clustered columnstore indexes and memory optimized tables

```
-- Create Table with column for JSON string
CREATE TABLE CustomerOrders
(
 CustomerId   BIGINT NOT NULL,
 Country      NVARCHAR(150) NOT NULL,
 OrderDetails NVARCHAR(3000) NOT NULL –- NVARCHAR column for JSON
) WITH (DISTRIBUTION = ROUND_ROBIN)

-- Populate table with semi-structured data
INSERT INTO CustomerOrders
VALUES
( 101, -- CustomerId
 'Bahrain', -- Country
 N'[{ StoreId": "AW73565",
     "Order": { "Number":"SO43659",
                "Date":"2011-05-31T00:00:00"
              },
     "Item":  { "Price":2024.40, "Quantity":1 }
    }]' -- OrderDetails
)
```

JSON data support – modify and operate on JSON data

Overview

Use standard table columns and values from JSON text in the same analytical query.

Modify JSON data with the following:

- JSON_MODIFY – modifies a value in a JSON string
- OPENJSON – convert JSON collection to a set of rows and columns

Benefits

Flexibility to update JSON string using T-SQL

Convert hierarchical data into flat tabular structure

```
-- Modify Item Quantity value
UPDATE CustomerOrders SET OrderDetails =
JSON_MODIFY(OrderDetails, '$.OrderDetails.Item.Quantity',2)
```

OrderDetails
N'[{ StoreId": "AW73565", "Order": { "Number":"SO43659", "Date":"2011-05-31T00:00:00“ }, "Item": { "Price":2024.40, "Quantity": 2}}]'

```
-- Convert JSON collection to rows and columns
SELECT CustomerId,
      StoreId,
      OrderDetails.OrderDate,
      OrderDetails.OrderPrice
FROM   CustomerOrders
CROSS APPLY OPENJSON (CustomerOrders.OrderDetails)
WITH ( StoreId      VARCHAR(50)  '$.StoreId',
      OrderNumber   VARCHAR(100) '$.Order.Date',
      OrderDate     DATETIME     '$.Order.Date',
      OrderPrice    DECIMAL      ‘$.Item.Price',
      OrderQuantity INT          '$.Item.Quantity'
     ) AS OrderDetails
```

CustomerId	StoreId	OrderDate	OrderPrice
101	AW73565	2011-05-31T00:00:00	2024.40

Stored Procedures

Overview

It is a group of one or more SQL statements or a reference to a Microsoft .NET Framework common language runtime (CLR) method.

Promotes flexibility and modularity.

Supports parameters and nesting.

Benefits

Reduced server/client network traffic, improved performance

Stronger security

Easy maintenance

```
CREATE PROCEDURE HumanResources.uspGetAllEmployees
AS
   SET NOCOUNT ON;
   SELECT LastName, FirstName, JobTitle, Department
   FROM HumanResources.vEmployeeDepartment;
GO

-- Execute a stored procedures
EXECUTE HumanResources.uspGetAllEmployees;
GO
-- Or
EXEC HumanResources.uspGetAllEmployees;
GO
-- Or, if this procedure is the first statement within a batch:
HumanResources.uspGetAllEmployees;
```

Materialized views

Overview

A materialized view pre-computes, stores, and maintains its data like a table.

Materialized views are automatically updated when data in underlying tables are changed. This is a synchronous operation that occurs as soon as the data is changed.

The auto caching functionality allows Azure Synapse Analytics Query Optimizer to consider using indexed view even if the view is not referenced in the query.

Supported aggregations: MAX, MIN, AVG, COUNT, COUNT_BIG, SUM, VAR, STDEV

Benefits

Automatic and synchronous data refresh with data changes in base tables. No user action is required.

High availability and resiliency as regular tables

```
-- Create indexed view
CREATE MATERIALIZED VIEW Sales.vw_Orders
WITH
(
  DISTRIBUTION = ROUND_ROBIN |
  HASH(ProductID)
)
AS
   SELECT SUM(UnitPrice*OrderQty) AS Revenue,
        OrderDate,
        ProductID,
        COUNT_BIG(*) AS OrderCount
   FROM   Sales.SalesOrderDetail
   GROUP  BY OrderDate, ProductID;
GO

-- Disable index view and put it in suspended mode
ALTER INDEX ALL ON Sales.vw_Orders DISABLE;

-- Re-enable index view by rebuilding it
ALTER INDEX ALL ON Sales.vw_Orders REBUILD;
```

Materialized views - example

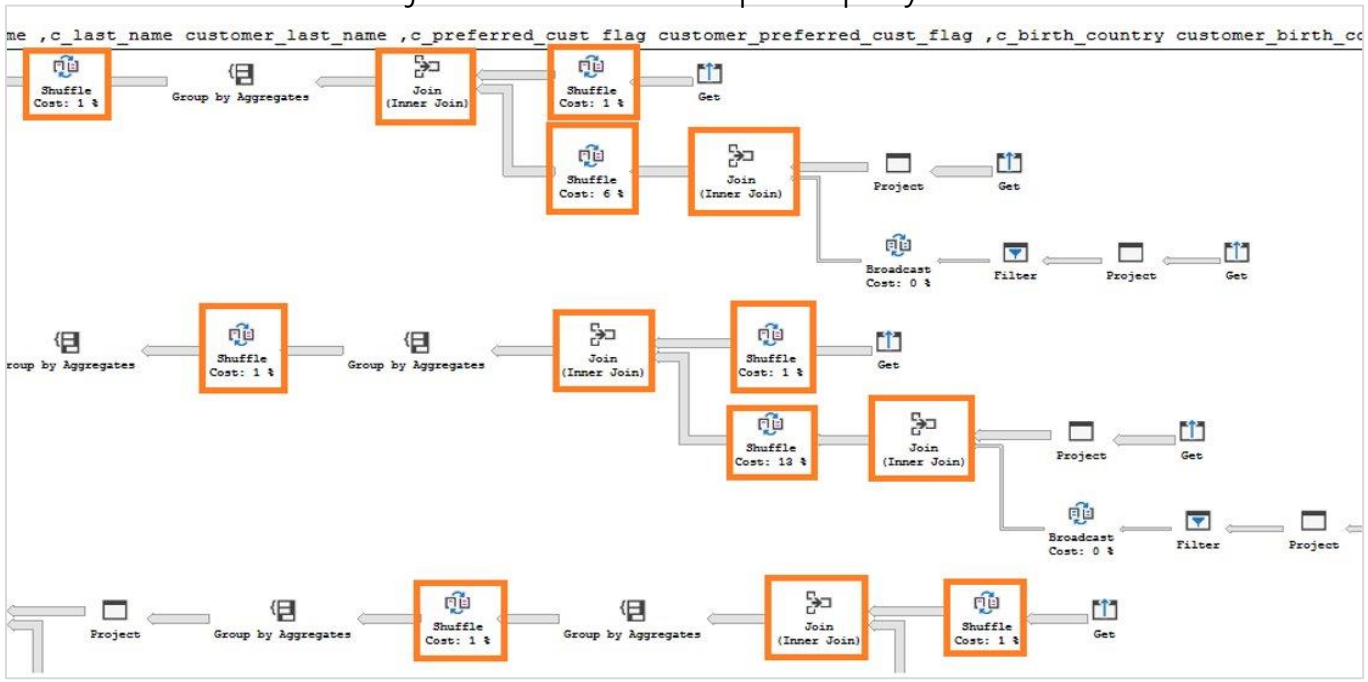
In this example, a query to get the year total sales per customer is shown to have a lot of data shuffles and joins that contribute to slow performance:

No relevant indexed views created on the data warehouse

```
-- Get year total sales per customer
(WITH year_total AS
        SELECT customer_id,
             first_name,
             last_name,
             birth_country,
             login,
             email_address,
             d_year,
             SUM(ISNULL(list_price – wholesale_cost –
             discount_amt + sales_price, 0)/2)year_total
        FROM   customer cust
        JOIN   catalog_sales sales ON cust.sk = sales.sk
        JOIN   date_dim ON sales.sold_date = date_dim.date
        GROUP  BY customer_id, first_name,
            last_name,birth_country,
            login,email_address ,d_year
)
SELECT TOP 100 …
FROM   year_total …
WHERE  …
ORDER BY …
```

Execution time: 103 seconds

Lots of data shuffles and joins needed to complete query



Materialized views - example

Now, we add an indexed view to the data warehouse to increase the performance of the previous query. This view can be leveraged by the query even though it is not directly referenced.

Original query – get year total sales per customer

```
-- Get year total sales per customer
(WITH year_total AS
        SELECT customer_id,
             first_name,
             last_name,
             birth_country,
             login,
             email_address,
             d_year,
             SUM(ISNULL(list_price – wholesale_cost –
             discount_amt + sales_price, 0)/2)year_total
        FROM   customer cust
        JOIN   catalog_sales sales ON cust.sk = sales.sk
        JOIN   date_dim ON sales.sold_date = date_dim.date
        GROUP  BY customer_id, first_name,
            last_name,birth_country,
            login,email_address ,d_year
)
SELECT TOP 100 …
FROM   year_total …
WHERE  …
ORDER BY …
```

Create indexed view with hash distribution on customer_id column

```
-- Create indexed view for query
CREATE INDEXED VIEW nbViewCS WITH (DISTRIBUTION=HASH(customer_id)) AS
SELECT customer_id,
        first_name,
        last_name,
        birth_country,
        login,
        email_address,
        d_year,
        SUM(ISNULL(list_price – wholesale_cost – discount_amt +
        sales_price, 0)/2) AS year_total
FROM   customer cust
JOIN   catalog_sales sales ON cust.sk = sales.sk
JOIN   date_dim ON sales.sold_date = date_dim.date
GROUP  BY customer_id, first_name,
    last_name,birth_country,
    login, email_address, d_year
```

Indexed (materialized) views - example

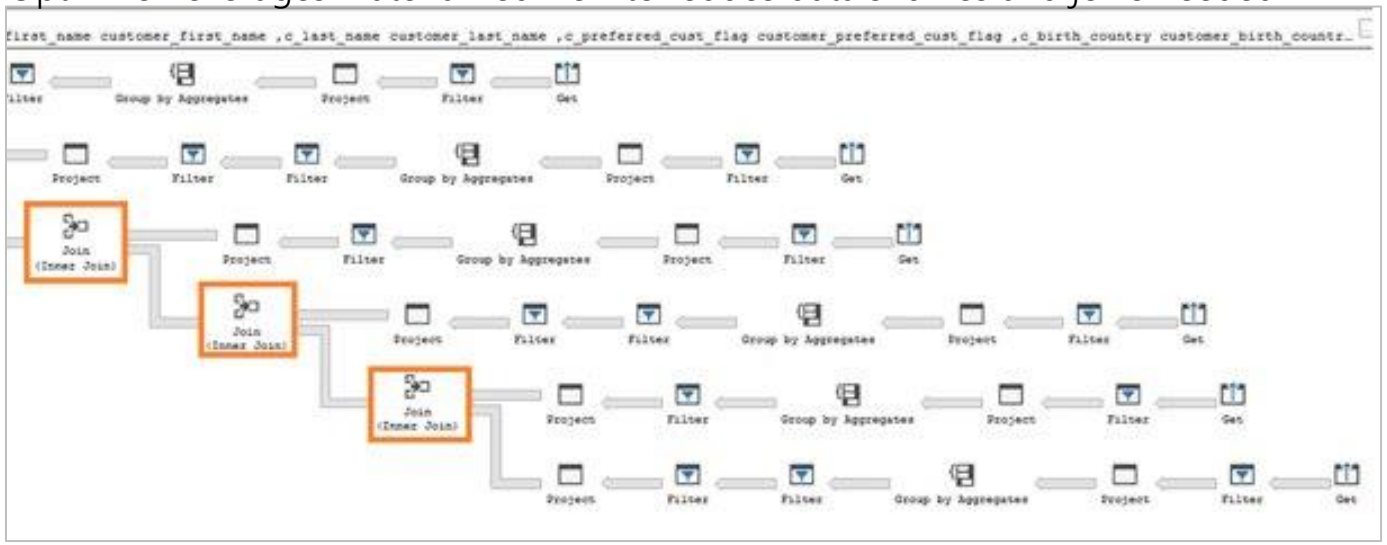
Synapse SQL (provisioned) query optimizer automatically leverages the indexed view to speed up the same query.
Notice that the query does not need to reference the view directly

Original query – no changes have been made to query

```
-- Get year total sales per customer
(WITH year_total AS
        SELECT customer_id,
             first_name,
             last_name,
             birth_country,
             login,
             email_address,
             d_year,
             SUM(ISNULL(list_price – wholesale_cost –
             discount_amt + sales_price, 0)/2)year_total
        FROM   customer cust
        JOIN   catalog_sales sales ON cust.sk = sales.sk
        JOIN   date_dim ON sales.sold_date = date_dim.date
        GROUP  BY customer_id, first_name,
            last_name,birth_country,
            login,email_address ,d_year
)
SELECT TOP 100 …
FROM   year_total …
WHERE  …
ORDER BY …
```

Execution time: 6 seconds

Optimizer leverages materialized view to reduce data shuffles and joins needed



Materialized views- Recommendations

EXPLAIN - provides query plan for SQL statement without running the statement; view estimated cost of the query operations.

EXPLAIN WITH_RECOMMENDATIONS - provides query plan with recommendations to optimize the SQL statement performance.

```
EXPLAIN WITH_RECOMMENDATIONS
select count(*)
from (
        (select distinct c_last_name, c_first_name, d_date
        from store_sales, date_dim, customer
        where store_sales.ss_sold_date_sk = date_dim.d_date_sk
        and store_sales.ss_customer_sk = customer.c_customer_sk
        and d_month_seq between 1194 and 1194+11)
except
        (select distinct c_last_name, c_first_name, d_date
        from catalog_sales, date_dim, customer
        where catalog_sales.cs_sold_date_sk = date_dim.d_date_sk
        and catalog_sales.cs_bill_customer_sk = customer.c_customer_sk and
        d_month_seq between 1194 and 1194+11)
) top_customers
```

COPY command

Overview

Copies data from source to destination

Benefits

Retrieves data from all files from the folder and all its subfolders.

Supports multiple locations from the same storage account, separated by comma

Supports Azure Data Lake Storage (ADLS) Gen 2 and Azure Blob Storage.

Supports CSV, PARQUET, ORC file formats

```
COPY INTO test_1
FROM 'https://XXX.blob.core.windows.net/customerdatasets/test_1.txt'
WITH (
   FILE_TYPE = 'CSV',
   CREDENTIAL=(IDENTITY= 'Shared Access Signature',
SECRET='<Your_SAS_Token>'),
   FIELDQUOTE = '"',
   FIELDTERMINATOR=';',
   ROWTERMINATOR='0X0A',
   ENCODING = 'UTF8',
   DATEFORMAT = 'ymd',
   MAXERRORS = 10,
   ERRORFILE = '/errorsfolder/'--path starting from the storage container,
   IDENTITY_INSERT
   )
```

```
COPY INTO test_parquet
FROM 'https://XXX.blob.core.windows.net/customerdatasets/test.parquet'
WITH (
   FILE_FORMAT = myFileFormat
   CREDENTIAL=(IDENTITY= 'Shared Access Signature',
SECRET='<Your_SAS_Token>')
)
```

Create External Table As Select

Overview

Creates an external table and then exports results of the Select statement. These operations will import data into the database for the duration of the query

Steps:

1. Create Master Key
2. Create Credentials
3. Create External Data Source
4. Create External Data Format
5. Create External Table

```
-- Create a database master key if one does not already exist
CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'S0me!nfo'
;
-- Create a database scoped credential with Azure storage account key as the secret.
CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential
WITH
    IDENTITY   = '<my_account>'
,   SECRET     = '<azure_storage_account_key>'
;
-- Create an external data source with CREDENTIAL option.
CREATE EXTERNAL DATA SOURCE MyAzureStorage
WITH
(   LOCATION   = 'wasbs://daily@logs.blob.core.windows.net/'
,   CREDENTIAL = AzureStorageCredential
,   TYPE       = HADOOP
)
-- Create an external file format
CREATE EXTERNAL FILE FORMAT MyAzureCSVFormat
WITH (FORMAT_TYPE = DELIMITEDTEXT,
    FORMAT_OPTIONS(
       FIELD_TERMINATOR = ',',
       FIRST_ROW = 2)
--Create an external table
CREATE EXTERNAL TABLE dbo.FactInternetSalesNew
WITH(
      LOCATION = '/files/Customer',
      DATA_SOURCE = MyAzureStorage,
      FILE_FORMAT = MyAzureCSVFormat
  )
AS SELECT T1.* FROM dbo.FactInternetSales T1 JOIN dbo.DimCustomer T2
ON ( T1.CustomerKey = T2.CustomerKey )
OPTION ( HASH JOIN );
```

Predict

Overview

It provides ability to import existing machine learning models and score them within provisioned SQL. It takes ONNX (Open Neural Network Exchange) and data as inputs and generates prediction based on model.

Benefits

1. It empowers data engineers to successfully deploy machine learning models with the familiar T-SQL interface

2. It offers seamless collaboration with data scientists

3. It generates new columns, but the number of columns and their data types depends on the type of model that was used for prediction.

Syntax:

```
PREDICT
(
MODEL = @model | model_literal,
DATA = object AS <table_alias>
)
WITH ( <result_set_definition> )
<result_set_definition> ::=
{
{ column_name
data_type
}
[,...n ]
}
MODEL = @model | model_literal
```

Example:

```
DECLARE @model varbinary(max) = (SELECT Model FROM Models WHERE Id = <>);
SELECT d.*, p.Score
FROM PREDICT(MODEL = @model,
DATA = dbo.mytable AS d) WITH (Score float) AS p;
```

Result-set caching

Overview

Cache the results of a query in provisioned SQL storage. This enables interactive response times for repetitive queries against tables with infrequent data changes.

The result-set cache persists even if dedicated SQL pool is paused and resumed later.

Query cache is invalidated and refreshed when underlying table data or query code changes.

Result cache is evicted regularly based on a time-aware least recently used algorithm (TLRU).

Benefits

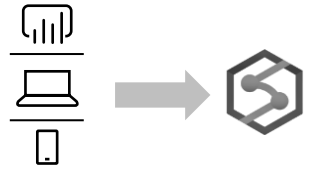
Enhances performance when same result is requested repetitively

Reduced load on server for repeated queries

Offers monitoring of query execution with a result cache hit or miss

```
-- Turn on/off result-set caching for a database
-- Must be run on the MASTER database
ALTER DATABASE {database_name}
SET RESULT_SET_CACHING { ON | OFF }

-- Turn on/off result-set caching for a client session
-- Run on target Azure Synapse Analytics
SET RESULT_SET_CACHING {ON | OFF}

-- Check result-set caching setting for a database
-- Run on target Azure Synapse Analytics
SELECT is_result_set_caching_on
FROM   sys.databases
WHERE  name = {database_name}

-- Return all query requests with cache hits
-- Run on target data warehouse
SELECT *
FROM   sys.dm_pdw_request_steps
WHERE  command like '%DWResultCacheDb%'
       AND step_index = 0
```

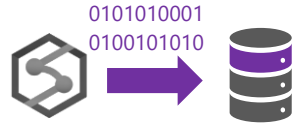
Result-set caching flow



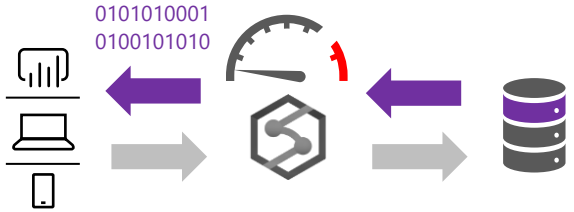
1. Client sends query to dedicated SQL pool

2. Query is processed using compute nodes which pull data from remote storage, process query and output back to client app

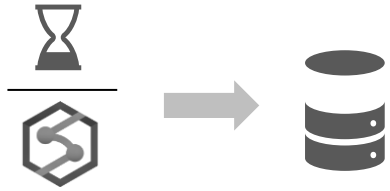
+



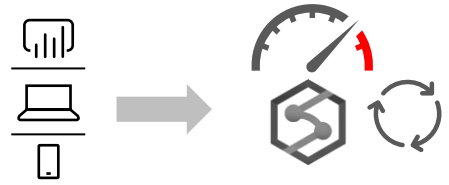
Query results are cached in remote storage so subsequent requests can be served immediately



3. Subsequent executions for the same query bypass compute nodes and can be fetched instantly from persistent cache in remote storage



4. Remote storage cache is evicted regularly based on time, cache usage, and any modifications to underlying table data.



5. Cache will need to be regenerated if query results have been evicted from cache

Continuous integration and delivery (CI/CD)

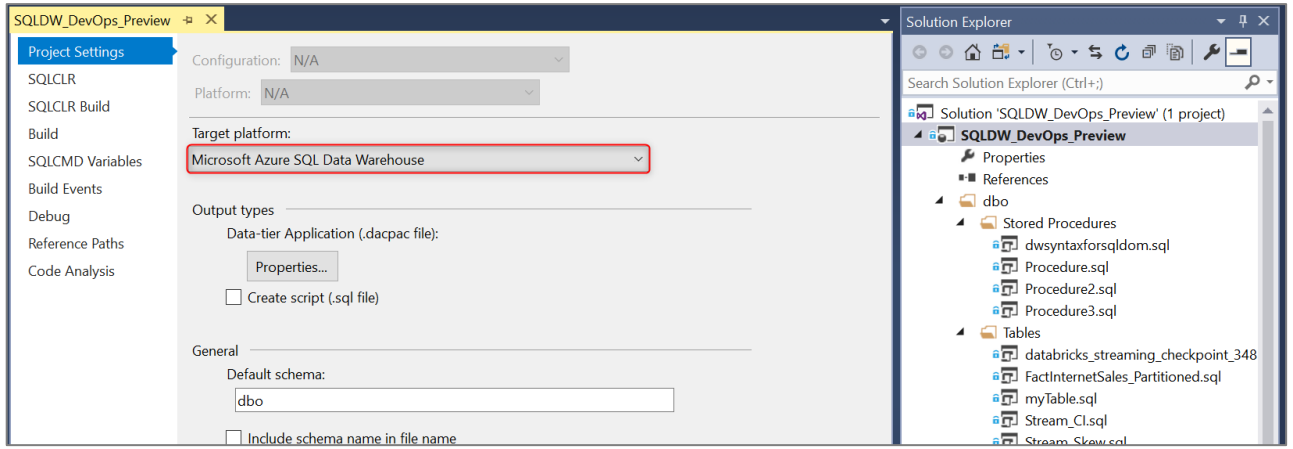
Overview

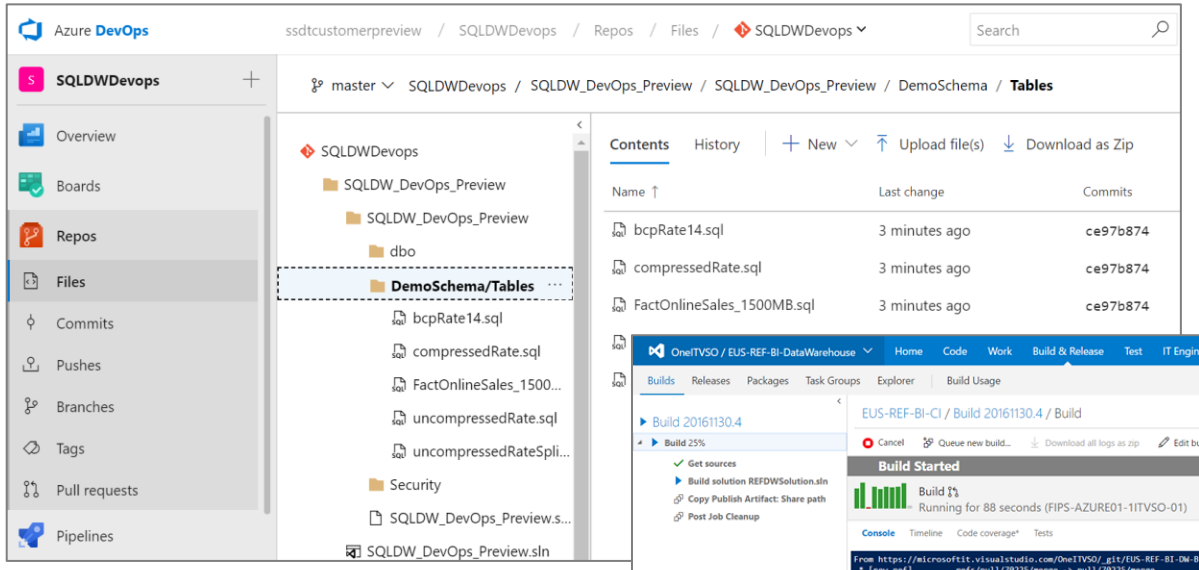
Database project support in SQL Server Data Tools (SSDT) allows teams of developers to collaborate over a version-controlled Azure Synapse Analytics, and track, deploy and test schema changes.

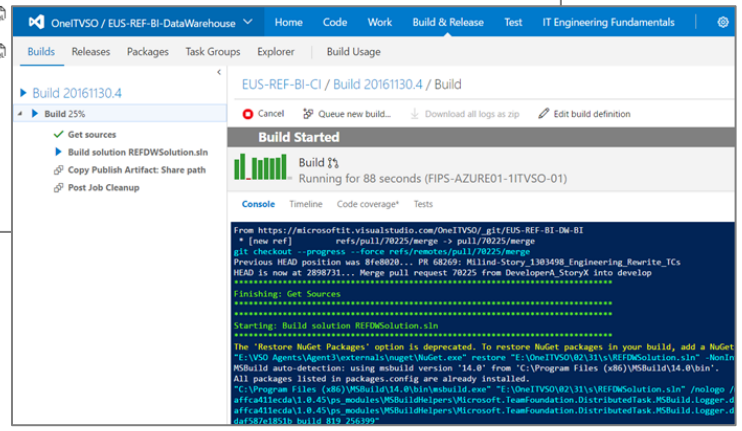
Benefits

Database project support includes first-class integration with Azure DevOps. This adds support for:

- **Azure Pipelines** to run CI/CD workflows for any platform (Linux, macOS, and Windows)
- **Azure Repos** to store project files in source control
- **Azure Test Plans** to run automated check-in tests to verify schema updates and modifications
- Growing ecosystem of third-party integrations that can be used to complement existing workflows (Timetracker, Microsoft Teams, Slack, Jenkins, etc.)







Azure Advisor recommendations

Suboptimal Table Distribution

Reduce data movement by replicating tables

Data Skew

Choose new hash-distribution key

Slowest distribution limits performance

Cache Misses

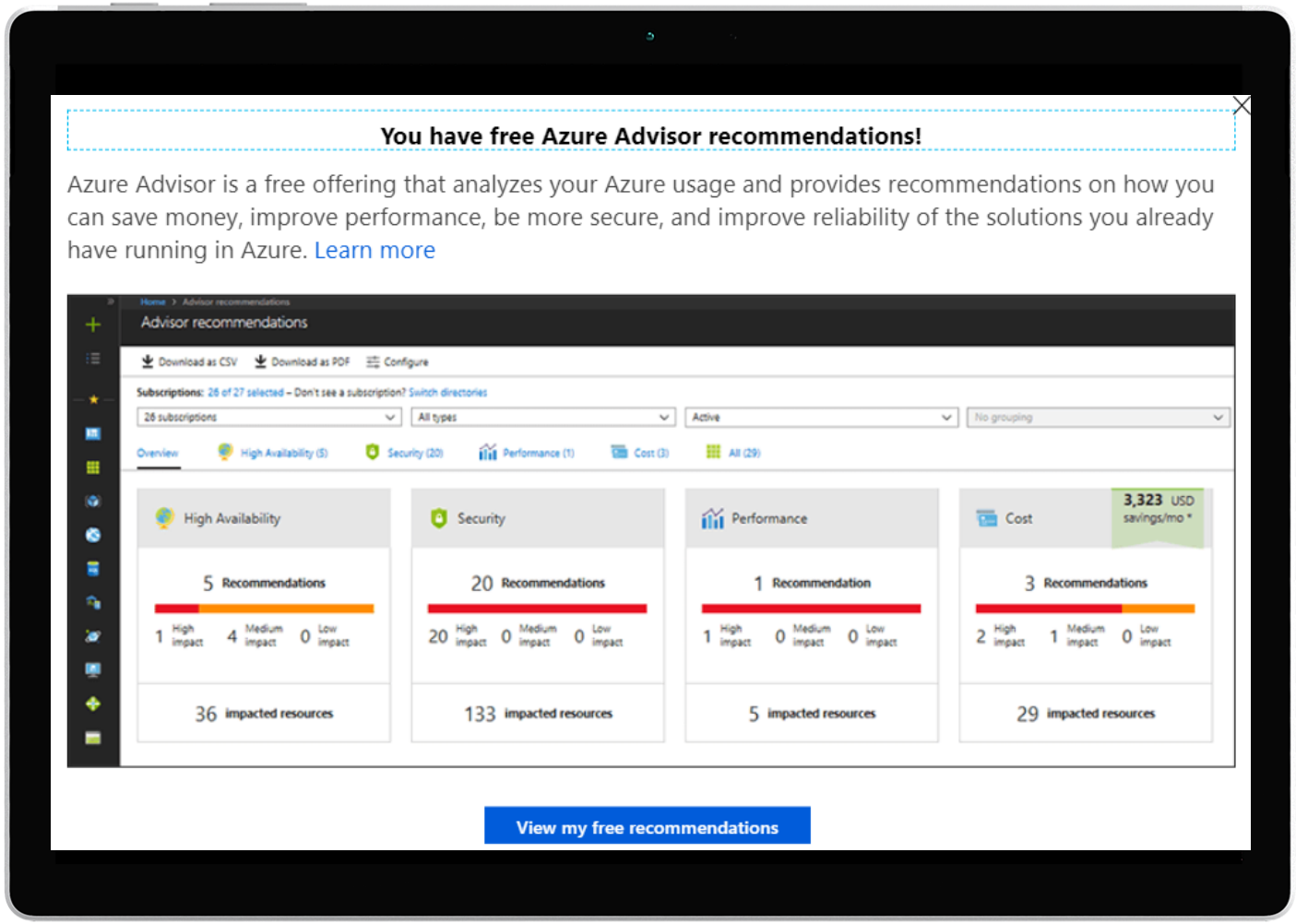
Provision additional capacity

Tempdb Contention

Scale or update user resource class

Suboptimal Plan Selection

Create or update table statistics



Maintenance windows

Overview

Choose a time window for your upgrades.

Select a primary and secondary window within a seven-day period.

Windows can be from 3 to 8 hours.

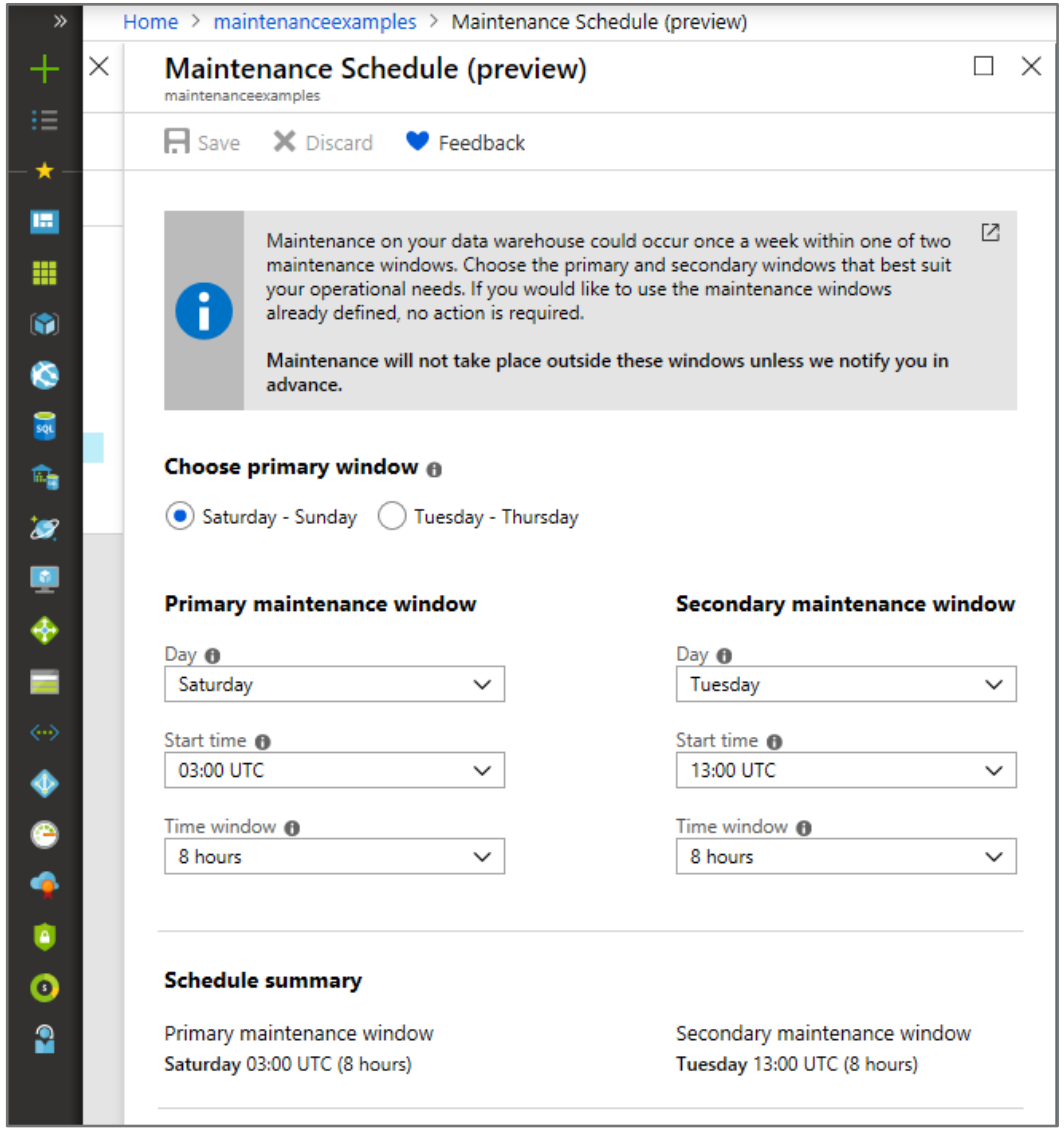
24-hour advance notification for maintenance events.

Benefits

Ensure upgrades happen on your schedule.

Predictable planning for long-running jobs.

Stay informed of start and end of maintenance.



Automatic statistics management

Overview

Statistics are automatically created and maintained for dedicated SQL pool. Incoming queries are analyzed, and individual column statistics are generated on the columns that improve cardinality estimates to enhance query performance.

Statistics are automatically updated as data modifications occur in underlying tables. By default, these updates are synchronous but can be configured to be asynchronous.

Statistics are considered out of date when:

- There was a data change on an empty table
- The number of rows in the table at time of statistics creation was 500 or less, and more than 500 rows have been updated
- The number of rows in the table at time of statistics creation was more than 500, and more than 500 + 20% of rows have been updated

```
-- Turn on/off auto-create statistics settings
ALTER DATABASE {database_name}
SET AUTO_CREATE_STATISTICS { ON | OFF }

-- Turn on/off auto-update statistics settings
ALTER DATABASE {database_name}
SET AUTO_UPDATE_STATISTICS { ON | OFF }

-- Configure synchronous/asynchronous update
ALTER DATABASE {database_name}
SET AUTO_UPDATE_STATISTICS_ASYNC { ON | OFF }

-- Check statistics settings for a database
SELECT      is_auto_create_stats_on,
            is_auto_update_stats_on,
            is_auto_update_stats_async_on
FROM        sys.databases
```

Azure Synapse Analytics

Synapse serverless SQL pool

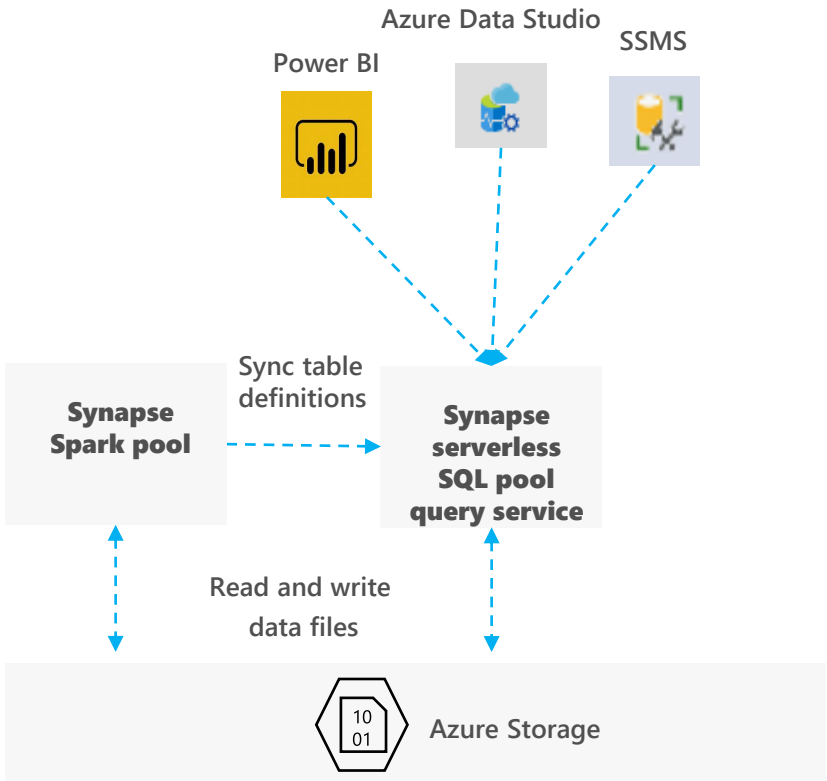
serverless SQL pool

Overview

An interactive query service that enables you to use standard T-SQL queries over files in Azure storage.

Benefits

- Use SQL to work with files on Azure storage
  - Directly query files on Azure storage using T-SQL
  - Logical Data Warehouse on top of Azure storage
  - Easy data transformation of Azure storage files
- Supports any tool or library that uses T-SQL to query data
- Automatically synchronize tables from Spark
- Serverless
  - No infrastructure, no upfront cost, no resource reservation
  - Pay only for query execution (per data processed)



Recommended usage scenarios

Quick data exploration

- Easily explore schema and data in files on Azure storage
- Supports various file formats (Parquet, CSV, JSON)
- Direct connector to Azure storage for large BI ecosystem

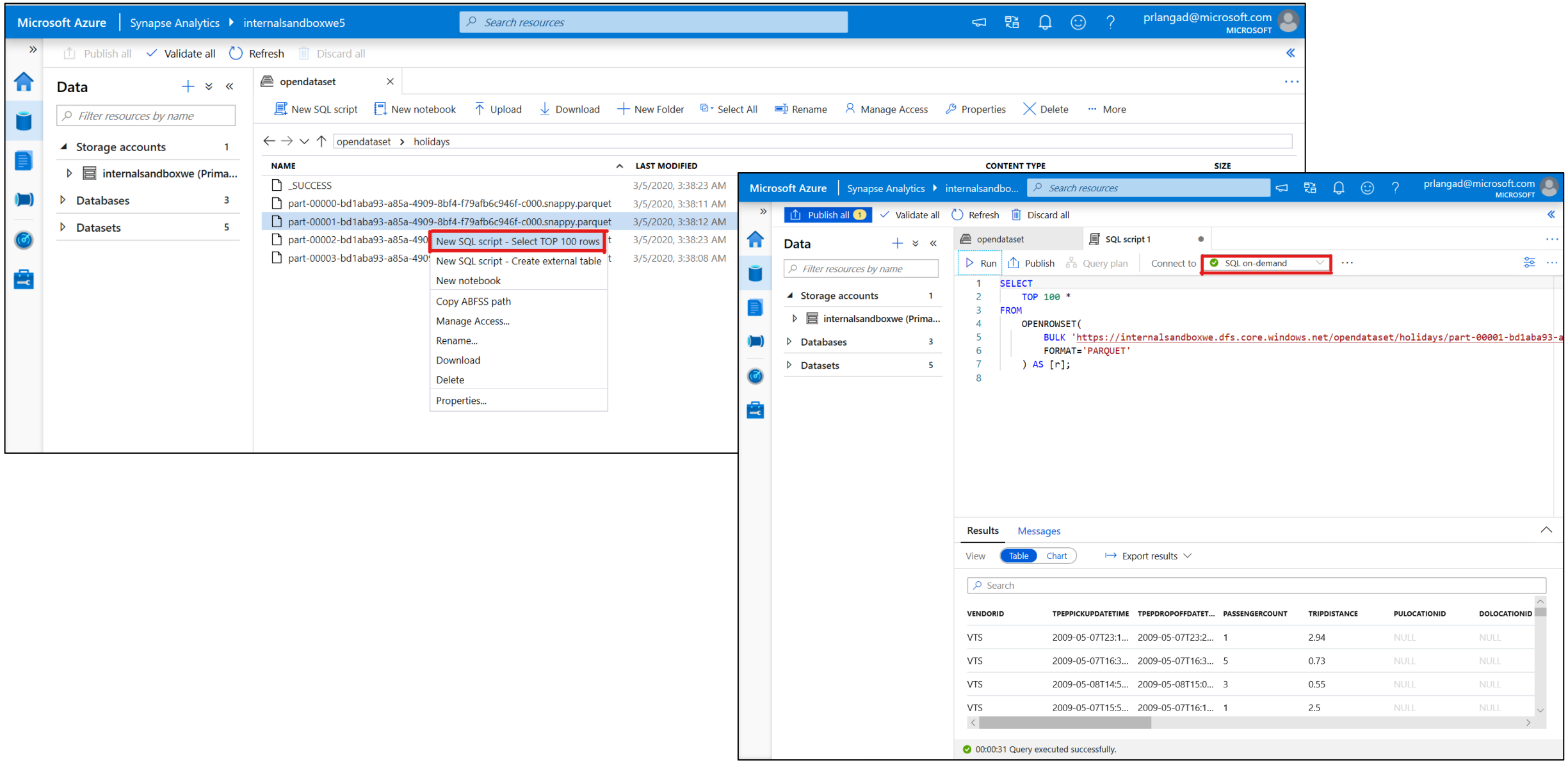
Logical Data Warehouse

- Model raw files as virtual tables and views
- Use any tool that works with SQL to analyze files
- Use enterprise-grade security model

Easy data transformation

- Transform CSV to parquet format
- Move data between containers and accounts
- Save the results of queries on external storage

Easily explore files on storage



Easily query files in various formats

Overview

Use OPENROWSET function to access data stored in various file formats

Benefits

Enables you to read CSV, parquet, and JSON files

Provides unified T-SQL interface for all file types

Use standard SQL language to transform and analyze returned data

- Use JSON functions to get the data from underlying files.
- Use JSON functions to get data from PARQUET nested types

```
SELECT TOP 10 *
FROM OPENROWSET(
      BULK 'https://XYZ.blob.core.windows.net/csv/taxi/*.csv',
      FORMAT = 'CSV')
WITH (
        country_code VARCHAR(4),
        country_name VARCHAR(50),
        year INT,
        population INT
        ) AS nyc
```

```
SELECT TOP 10 *
FROM OPENROWSET(
      BULK 'https://XYZ.blob.core.windows.net/csv/taxi/*.parquet',
      FORMAT = 'PARQUET') AS nyc
```

```
SELECT TOP 10 *
  JSON_VALUE(jsonContent, '$.countryCode') AS country_code,
  JSON_VALUE(jsonContent, '$.countryName') AS country_name,
  JSON_VALUE(jsonContent, '$.year')  AS year
  JSON_VALUE(jsonContent, '$.population')  AS population
FROM  OPENROWSET(
      BULK 'https://XYZ.blob.core.windows.net/json/taxi/*.json',
      FORMAT='CSV',
      FIELDTERMINATOR ='0x0b',
      FIELDQUOTE = '0x0b',
      ROWTERMINATOR = '0x0b'
   )
   WITH ( jsonContent varchar(MAX)   ) AS json_line
```

	country_code	country_name	year	population
1	LU	Luxembourg	2017	594130

Automatic schema inference

Overview

OPENROWSET will automatically determine columns and types of data stored in external file.

Benefits

No need to up-front analyze file structure to query the file

OPENROWSET identifies columns and their types based on underlying file metadata.

Perfect solution for data exploration where schema is unknown.

The functionality is available for both parquet & CSV files.

```
SELECT TOP 10 *
FROM OPENROWSET(
        BULK 'https://XYZ.blob.core.windows.net/csv/taxi/*.parquet',
        FORMAT = 'PARQUET') AS nyc
```

	country_code	country_name	year	population
1	LU	Luxembourg	2017	594130

```
SELECT
  TOP 100 *
FROM
  OPENROWSET(
    BULK 'https://azuresynapsesa.dfs.core.windows.net/default/RetailData/StoreDemoGraphics.csv',
    FORMAT = 'CSV',
    PARSER_VERSION='2.0',
HEADER_ROW = TRUE) AS [result]
```

StoreId	RatioAge60	CollegeRatio	Income	HighIncome15...	LargeHH	MinoritiesRatio	More1FullTime...	DistanceNeare...	SalesN
2	0.232864734	0.248934934	10.55320518	0.463887065	0.103953406	0.114279949	0.303585347	2.110122129	1.1428
5	0.117368032	0.32122573	10.92237097	0.535883355	0.103091585	0.053875277	0.410568032	3.801997814	0.6818

Defined the query result schema inline

Overview

Specify columns and types at query time.

Benefits

Define result schema at query time in WITH clause.

No need for external format files.

Explicitly define exact return types, their sizes, and collations.

Improve performance by column elimination in parquet files.

```
SELECT TOP 10 *
FROM OPENROWSET(
        BULK 'https://XYZ.blob.core.windows.net/csv/taxi/*.csv',
        FORMAT = 'CSV')
WITH (
        country_code VARCHAR(4),
        country_name VARCHAR(50),
        year INT,
        population INT
        ) AS nyc
```

	country_code	country_name	year	population
1	LU	Luxembourg	2017	594130

Customize the content parsing to fit your case

Overview

Uses OPENROWSET function to access data from various types of CSV files.

Benefits

Ability to read CSV files with custom format

- With or without header row
- Handle any new-line terminator (Windows or Unix style)
- Use custom field terminator and quote character
- Read UTF-8 and UTF-18 encoded files
- Use only a subset of columns by specifying column position after column types

```
SELECT *
FROM OPENROWSET(
      BULK 'https://XYZ.blob.core.windows.net/csv/population/population.csv',
      FORMAT = 'CSV',
      FIELDTERMINATOR =',',
      ROWTERMINATOR = '\n'
   )
WITH (
   [country_code] VARCHAR (5) 2,
   [country_name] VARCHAR (100) 4,
   [year] smallint 7,
   [population] bigint 9
) AS [r]
WHERE
   country_name = 'Luxembourg'
   AND year = 2017
```

Second, fourth, seventh and ninth columns are returned

	country_code	country_name	year	population
1	LU	Luxembourg	2017	594130

Easily query multiple files, with wildcards

Overview

Uses OPENROWSET function to access data from multiple files or folders using wildcards in path

Benefits

Offers reading multiple files/folders through usage of wildcards

Offers reading specific file/folder

Supports use of multiple wildcards

```
SELECT YEAR(pickup_datetime) as [year],
        SUM(passenger_count) AS passengers_total,
        COUNT(*) AS [rides_total]
FROM OPENROWSET(
       BULK 'https://XYZ.blob.core.windows.net/csv/taxi/year=*/month=1/*.parquet',
       FORMAT = 'PARQUET')  AS nyc
GROUP BY YEAR(pickup_datetime)
ORDER BY YEAR(pickup_datetime)
```

	year	passengers_total	rides_total
1	2001	14	10
2	2002	29	16
3	2003	22	16
4	2008	378	188
5	2009	594	353
6	2016	102093687	61758523
7	2017	184464988	113496932
8	2018	86272771	53925040
9	2019	37	29
…	2020	6	6

Query partitioned data, using the folder structure

Overview

Uses OPENROWSET function to access data partitioned in sub-folders

Benefits

Use filepath() function to access actual values from file paths.

Eliminate sub-folders/partitions before the query starts execution

Query Spark/Hive partitioned data sets

```
SELECT
    r.filepath(1) AS [year]
   ,r.filepath(2) AS [month]
   ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
      BULK 'https://XYZ.blob.core.windows.net/year=*/month=*/*.parquet',
      FORMAT = 'PARQUET') AS [r]
WHERE r.filepath(1) IN ('2017')
   AND r.filepath(2) IN ('10', '11', '12')

GROUP BY    r.filepath() ,r.filepath(1) ,r.filepath(2)
ORDER BY    filepath
```

year	month	rows
2017	10	9768815
2017	11	9284803
2017	12	9508276

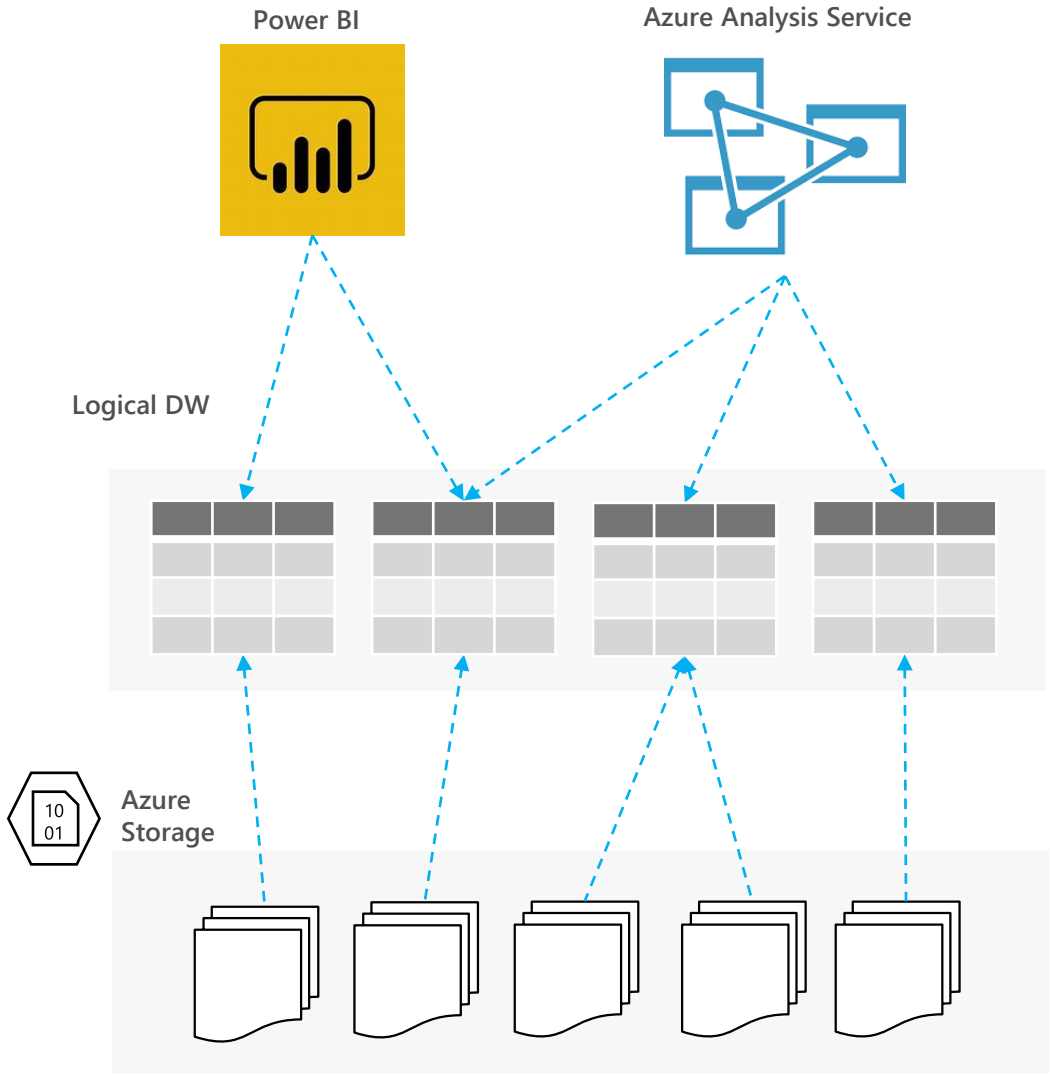
Synapse serverless SQL pool as a logical data warehouse

Overview

Logical relational layer on top of physical files in Azure Storage.

Benefits

- Abstract physical storage and file formats using well understandable relational concepts such as tables and views.
- Direct connector to Azure storage for large ecosystem of BI tools
- BI tools that use SQL can work with files on storage
  - Analytic tools use external tables that represent proxy to actual files.
  - No need for custom connectors in BI tools.
- Provides complex data processing (joining and aggregation) on top of raw files.
- Apply enterprise-ready security model and access control using battle-tested SQL Server permission model on top of Azure storage files



Logical Data Warehouse views

Overview

serverless SQL pool logical data warehouse views are created on external files placed in customer Azure storage

Benefits

Create SQL views on externally stored data

Access files using the view from various tools and language

Leverage rich T-SQL language to process and analyze data in external files exposed via views

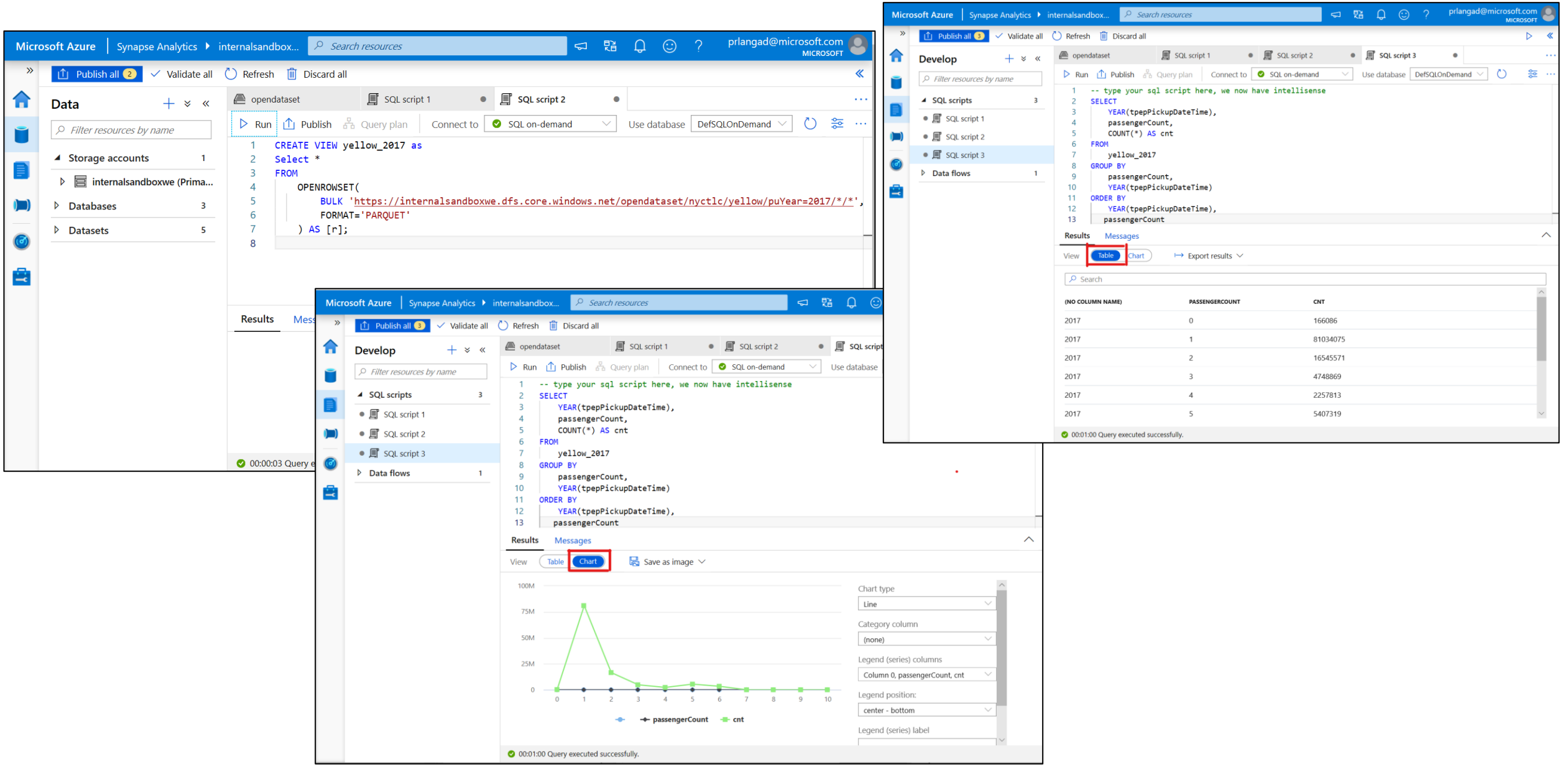
Create PowerBI reports on the views created on external data

```
USE [mydbname]
GO

DROP VIEW IF EXISTS  populationView
GO

CREATE VIEW populationView AS
SELECT *
FROM OPENROWSET(
      BULK 'https://XYZ.blob.core.windows.net/csv/population/*.csv',
      FORMAT = 'CSV',
      FIELDTERMINATOR =',',
      ROWTERMINATOR = '\n'
   )
WITH (
   [country_code] VARCHAR (5) ,
   [country_name] VARCHAR (100),
   [year] smallint,
   [population] bigint
) AS [r]
```

```
SELECT
   country_name, population
FROM populationView
WHERE
   [year] = 2019
ORDER BY
   [population] DESC
```

	country_name	population
1	China	1389618778
2	India	1311559204
3	United States	331883986
4	Indonesia	264935824
5	Pakistan	210797836
6	Brazil	210301591
7	Nigeria	208679114
8	Bangladesh	161062905
9	Russia	141944641
10	Mexico	127318112

Creating views



Logical Data Warehouse - tables

Overview

Create external tables that reference external files in your serverless SQL pool logical data warehouse

Benefits

Create external tables that reference set of files on Azure storage.

Join and transform multiple tables in the same query.

Enables you to analyze external files with the same experience that you have in classic databases.

Manage column statistics in external tables.

Manage access rights per table.

Create PowerBI reports on the views created on external data

```
USE [mydbname]
GO

DROP TABLE IF EXISTS dbo.Population
GO
CREATE EXTERNAL TABLE dbo.Population (
   country_code VARCHAR (5) COLLATE Latin1_General_BIN2,
   country_name VARCHAR (100) COLLATE Latin1_General_BIN2,
   year smallint,
   population bigint
)
WITH(
      LOCATION = '/csv/population/population-*.csv',
      DATA_SOURCE = MyAzureStorage,
      FILE_FORMAT = MyAzureCSVFormat
)
```

```
CREATE STATISTICS stat_country_name
ON dbo.Population(country_name);
```

```
SELECT
   country_name, population
FROM population
WHERE year = 2019
ORDER BY population DESC
```

	country_name	population
1	China	1389618778
2	India	1311559204
3	United States	331883986
4	Indonesia	264935824
5	Pakistan	210797836
6	Brazil	210301591
7	Nigeria	208679114
8	Bangladesh	161062905
9	Russia	141944641
10	Mexico	127318112

Easy data transformation

Overview

Easily perform data transformations of Azure Storage files using SQL queries

Optimize data pipeline - achieve more using serverless SQL pool

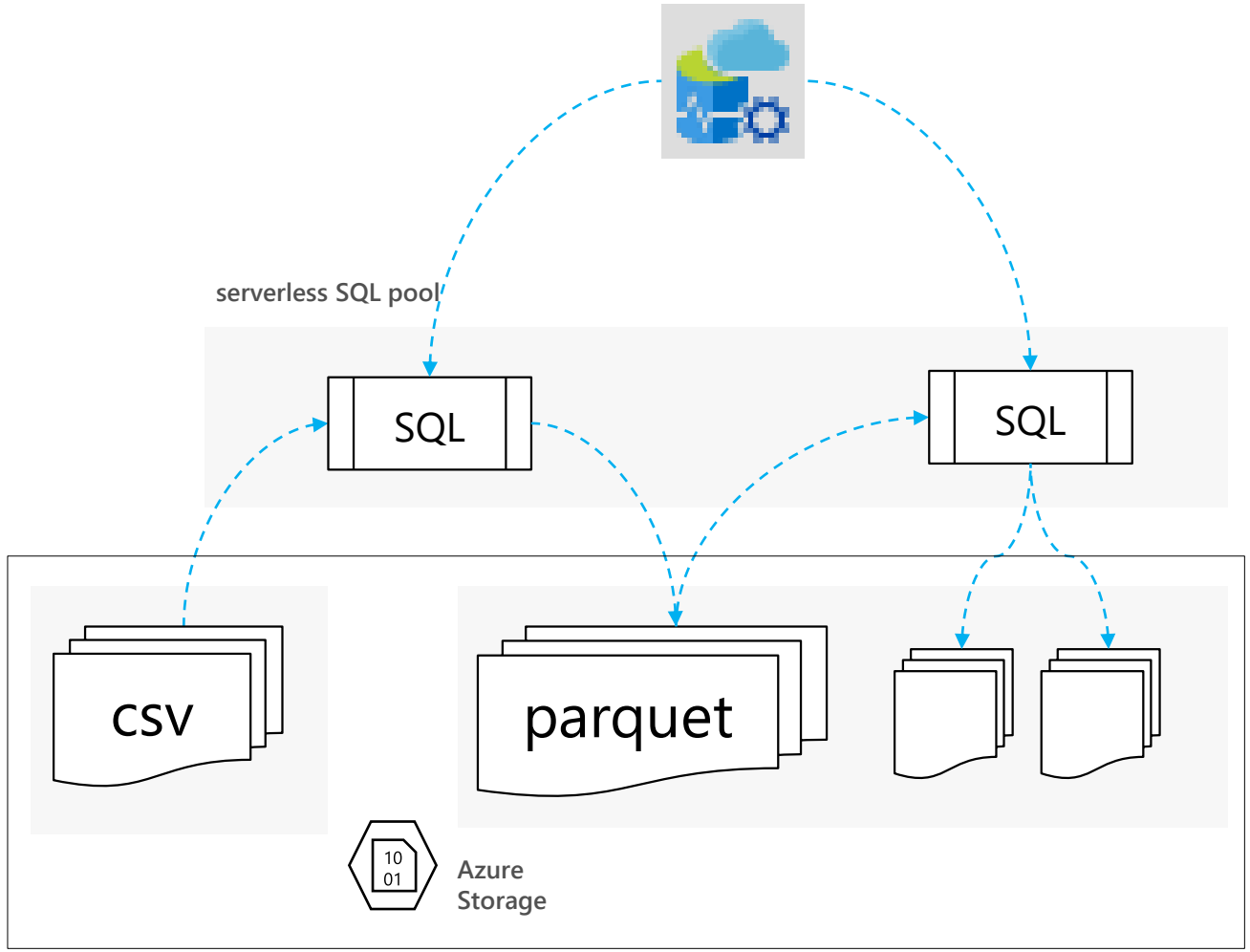
Benefits

Single statement transformations:

- convert CSV or JSON files to Parquet
- copy files from one storage account to another
- re-partition data to new location(s)
- store results of your query on Azure Storage

SQL ETL pipelines

- Use SQL commands to transform data
- Chain SQL statement for build ETL process
- Materialize reports created on the current snapshot of data



Easy data transformation with CETAS

Overview

Create external tables as select (CETAS) enables you to easily transform data and store the results of query on Azure storage

Benefits

Select any data set and store it in parquet format.

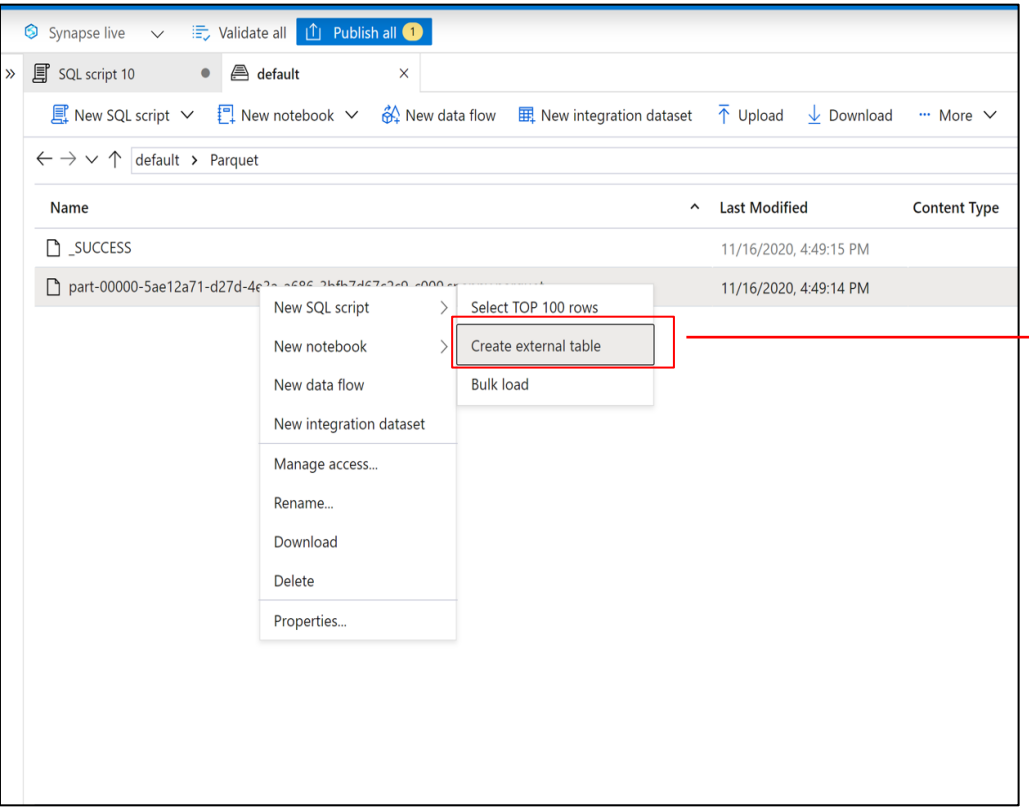
Pre-calculate and store results of query and store them permanently on Azure storage.

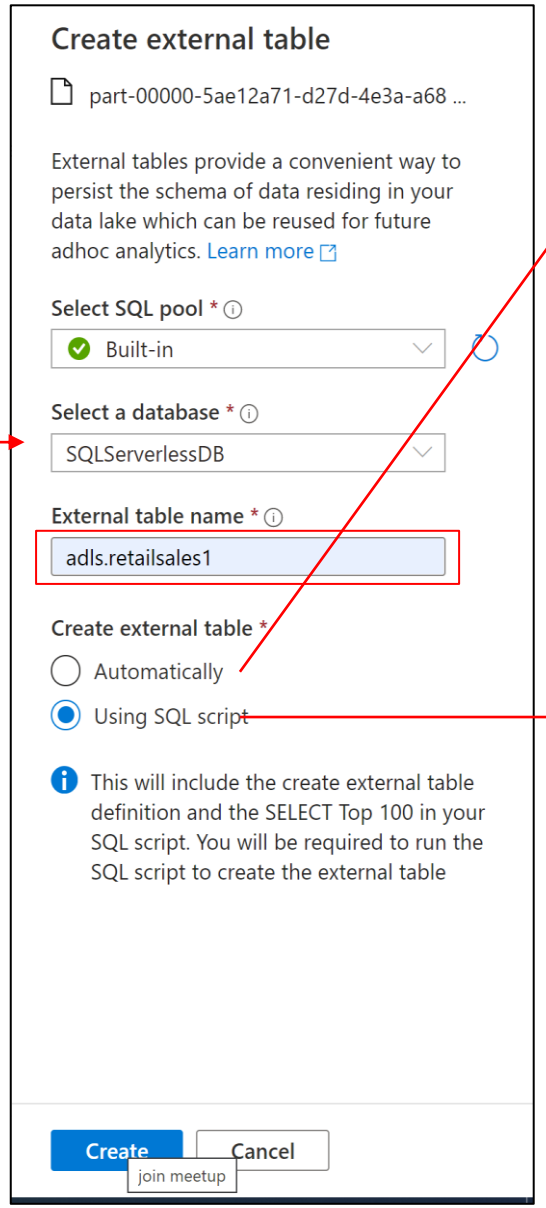
Use saved data using external table.

Improve performance of your reports by permanently storing the result based on current snapshot of data as parquet files.

```
-- copy CSV dataset into parquet data set
CREATE EXTERNAL TABLE parquet.Population
WITH(
      LOCATION = '/parquet/population',
      DATA_SOURCE = MyAzureStorage,
      FILE_FORMAT = MyAzureParquetFormat  )
AS
SELECT *
FROM csv.Population

-- pre-create report using new parquet data-set
CREATE EXTERNAL TABLE parquet.PopulationByMonth2017
WITH(
      LOCATION = '/parquet/population/bymonth/2017',
      DATA_SOURCE = MyAzureStorage,
      FILE_FORMAT = MyAzureParquetFormat  )
AS
SELECT month = p.month, population = COUNT ( p.population )
FROM parquet.Population p
WHERE p.year = 2017
GROUP BY p.month

-- Reporting tools can now directly read data from pre-created report
SELECT *
FROM parquet.PopulationByMonth2017
```

UI based data transformation






```
SELECT TOP 100 * FROM adls.retailsale
GO
```

```
IF NOT EXISTS (SELECT * FROM sys.external_file_formats WHERE name = 'SynapseParquetFormat')
    CREATE EXTERNAL FILE FORMAT [SynapseParquetFormat]
    WITH ( FORMAT_TYPE = PARQUET)
GO

IF NOT EXISTS (SELECT * FROM sys.external_data_sources WHERE name = 'default_azuresynapsesa_dfs_core_windows_net')
    CREATE EXTERNAL DATA SOURCE [default_azuresynapsesa_dfs_core_windows_net]
    WITH (
        LOCATION   = 'https://azuresynapsesa.dfs.core.windows.net/default',
    )
Go

CREATE EXTERNAL TABLE adls.retailsale (
    [storeId] varchar(8000),
    [productCode] varchar(8000),
    [quantity] varchar(8000),
    [logQuantity] varchar(8000),
    [advertising] varchar(8000),
    [price] varchar(8000),
    [weekStarting] varchar(8000),
    [id] varchar(8000)
    )
    WITH (
    LOCATION = 'Parquet/part-00000-5ae12a71-d27d-4e3a-a686-3bfb7d67c2c9-c000.snappy.parquet',
    DATA_SOURCE = [default_azuresynapsesa_dfs_core_windows_net],
    FILE_FORMAT = [SynapseParquetFormat]
    )
GO

SELECT TOP 100 * FROM adls.retailsale
```

Automatic syncing of Spark tables

Overview

Tables created in Spark pool are automatically created as external tables that reference external files in your serverless SQL pool logical data warehouse

Benefits

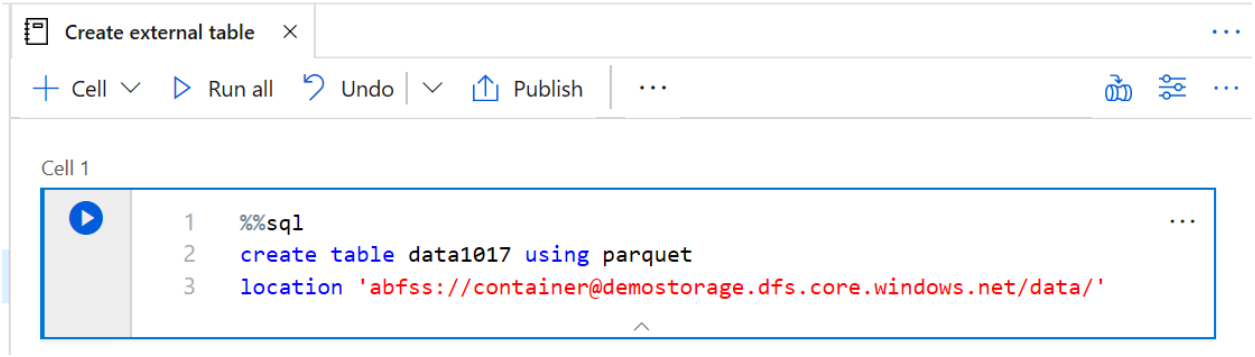
Tables designed using Spark languages are immediately available in serverless SQL pool.

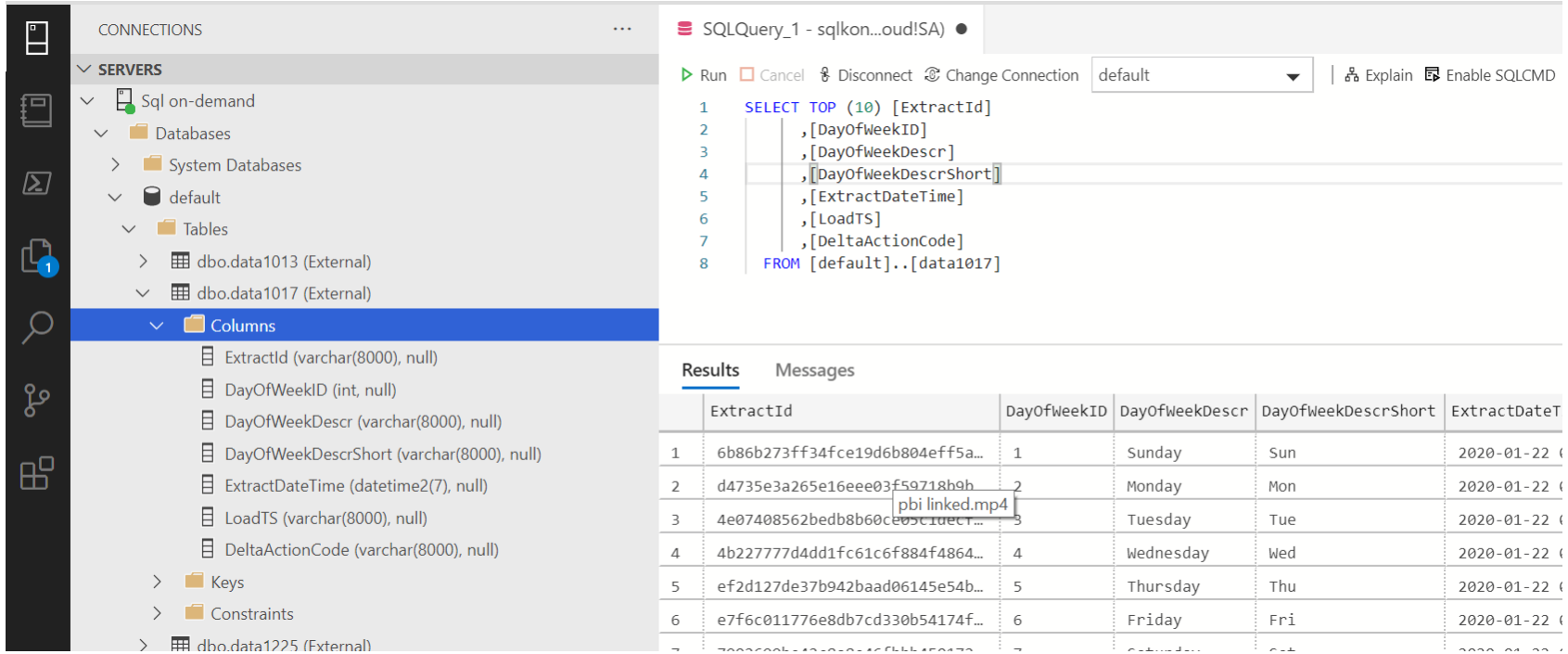
Schema definition matches original

Spark table updates are applied in serverless SQL pool

No need to manually create SQL tables that match Spark tables

Spark and serverless SQL pool tables references the same external files.





Microsoft

Azure Synapse Analytics

Apache Spark

Azure Synapse Apache Spark - Summary

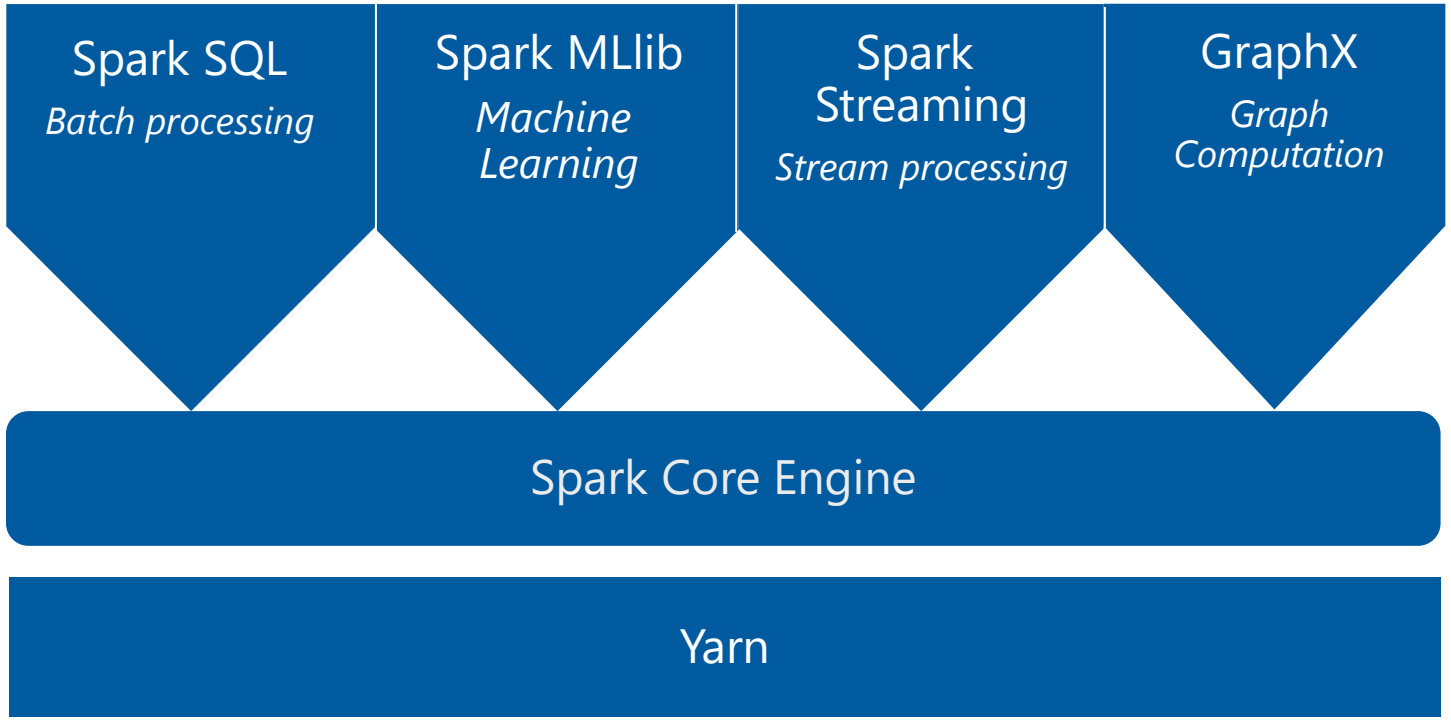
- **Apache Spark 2.4 derivation**
  - Linux Foundation Delta Lake 0.6 support
  - .Net Core 3.0 support
  - Python 3.6 + Anacondas support
- **Tightly coupled to other Azure Synapse services**
  - Integrated security and sign on
  - Integrated Metadata
  - Integrated and simplified provisioning
  - Integrated UX including nteract based notebooks
  - Fast load of Synapse SQL (provisioned) pools
- **Core scenarios**
  - Data Prep/Data Engineering/ETL
  - Machine Learning via Spark ML and Azure ML integration
  - Extensible through library management
- **Efficient resource utilization**
  - Fast Start
  - Auto scale (up and down)
  - Auto pause
  - Min cluster size of 3 nodes
- **Multi Language Support**
  - .Net (C#), PySpark, Scala, Spark SQL, Java

Apache Spark

An unified, open source, parallel, data processing framework for Big Data Analytics

Spark Unifies:

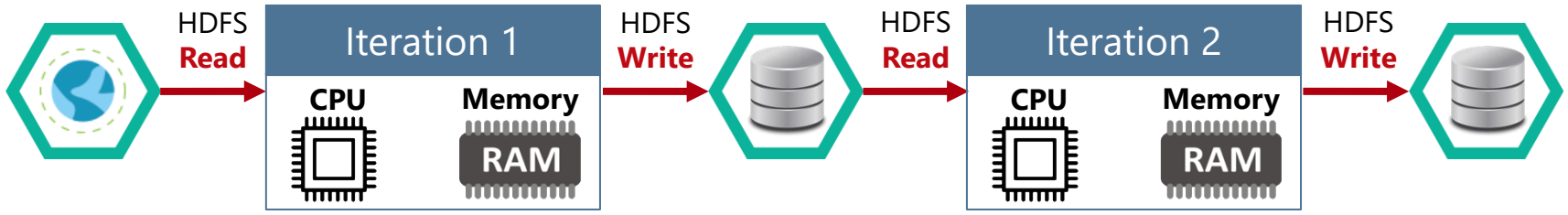
- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



http://spark.apache.org

Motivation for Apache Spark

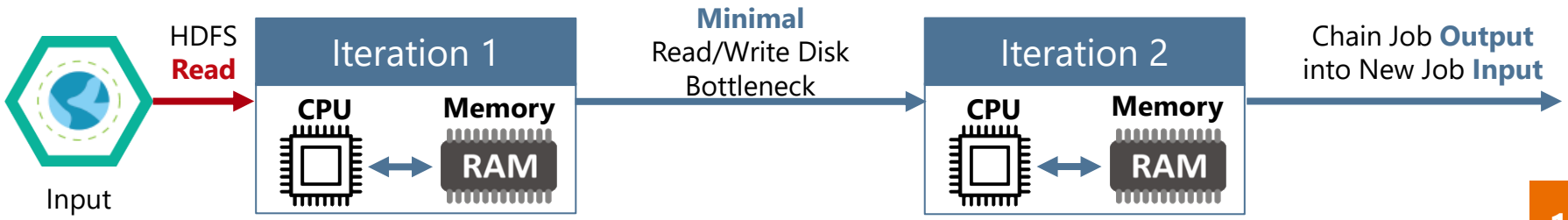
Traditional Approach: MapReduce jobs for complex jobs, interactive query, and online event-hub processing involves lots of (slow) disk I/O



Motivation for Apache Spark

Traditional Approach: MapReduce jobs for complex jobs, interactive query, and online event-hub processing involves lots of **(slow) disk I/O**

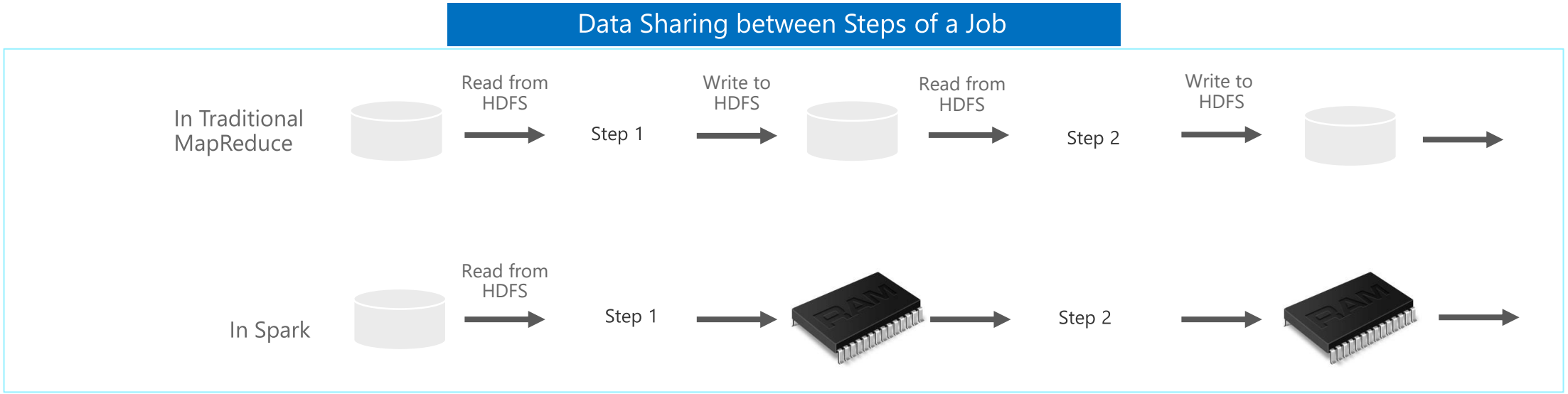
Solution: Keep data **in-memory** with a new distributed execution engine



10–100x faster than network & disk

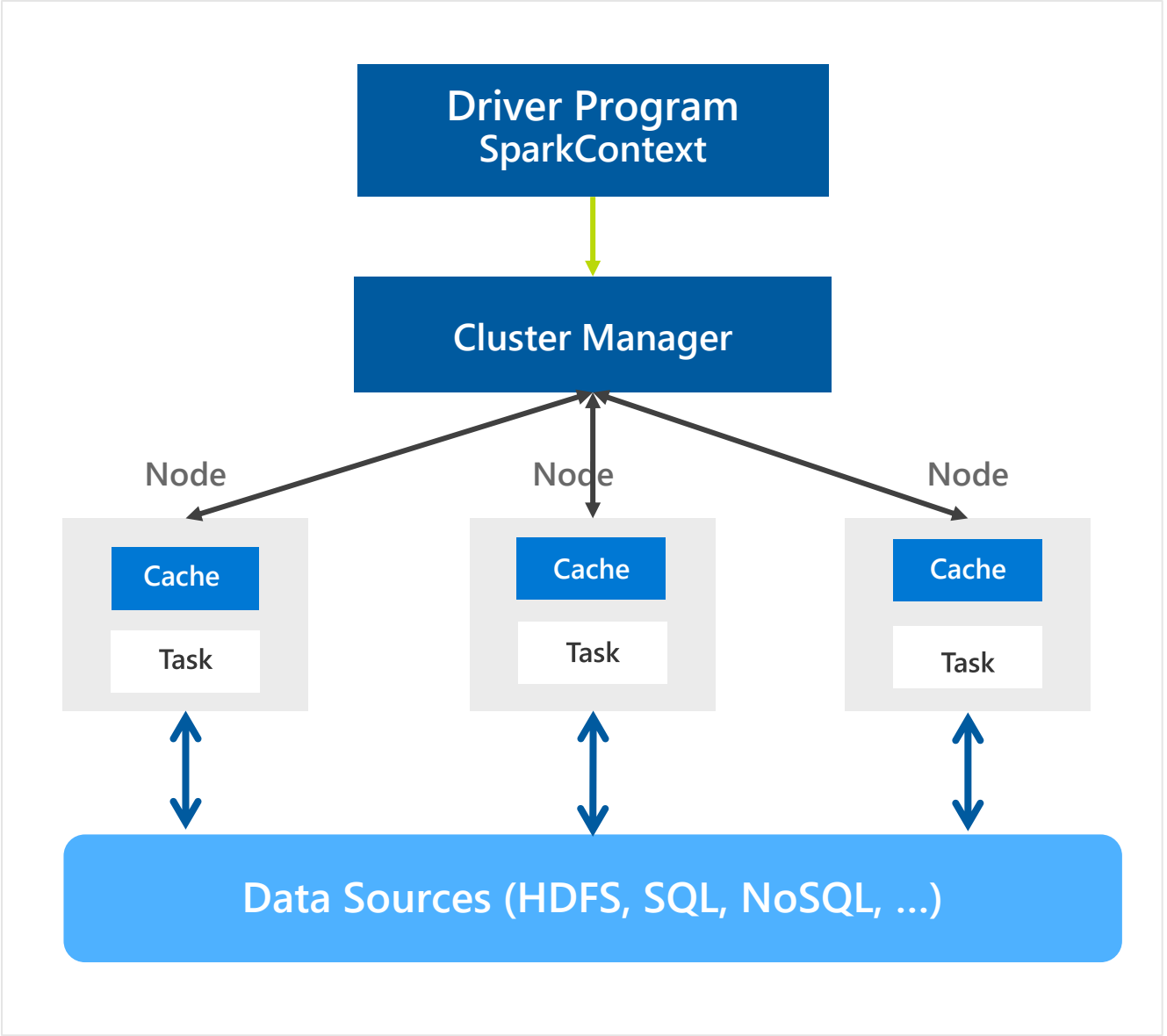
What makes Spark fast

- **In-memory cluster computing**: Spark provides primitives for *in-memory* cluster computing. A Spark job can *load and cache* data into memory and query it repeatedly (iteratively) much quicker than disk-based systems.
- **Scala Integration**: Spark integrates into the Scala programming language, letting you manipulate distributed datasets like local collections. No need to structure everything as map and reduce operations
- **Faster Data-sharing**: Data-sharing between operations is faster as data is in-memory:
  - In (traditional) Hadoop data is shared through HDFS which is expensive. HDFS maintains three replicas.
  - Spark stores data in-memory *without any replication*.



General Spark Cluster Architecture

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).



Spark Component Features

Spark SQL

- Unified data access: Query structured data sets with SQL or DataFrame APIs
- Fast, familiar query language across all your enterprise data
- Use BI tools to connect and query via JDBC or ODBC drivers

Mllib/SparkML

- Predictive and prescriptive analytics
- Machine learning algorithms for:
  - Clustering
  - Classification
  - Regression
  - etc.
- Smart application design from pre-built, out-of-the-box statistical and algorithmic models

Spark Streaming

- Micro-batch event processing for near-real time analytics
- e.g. Internet of Things (IoT) devices, Twitter feeds, Kafka (event hub), etc.
- Spark's engine drives some action or outputs data in batches to various data stores

GraphX

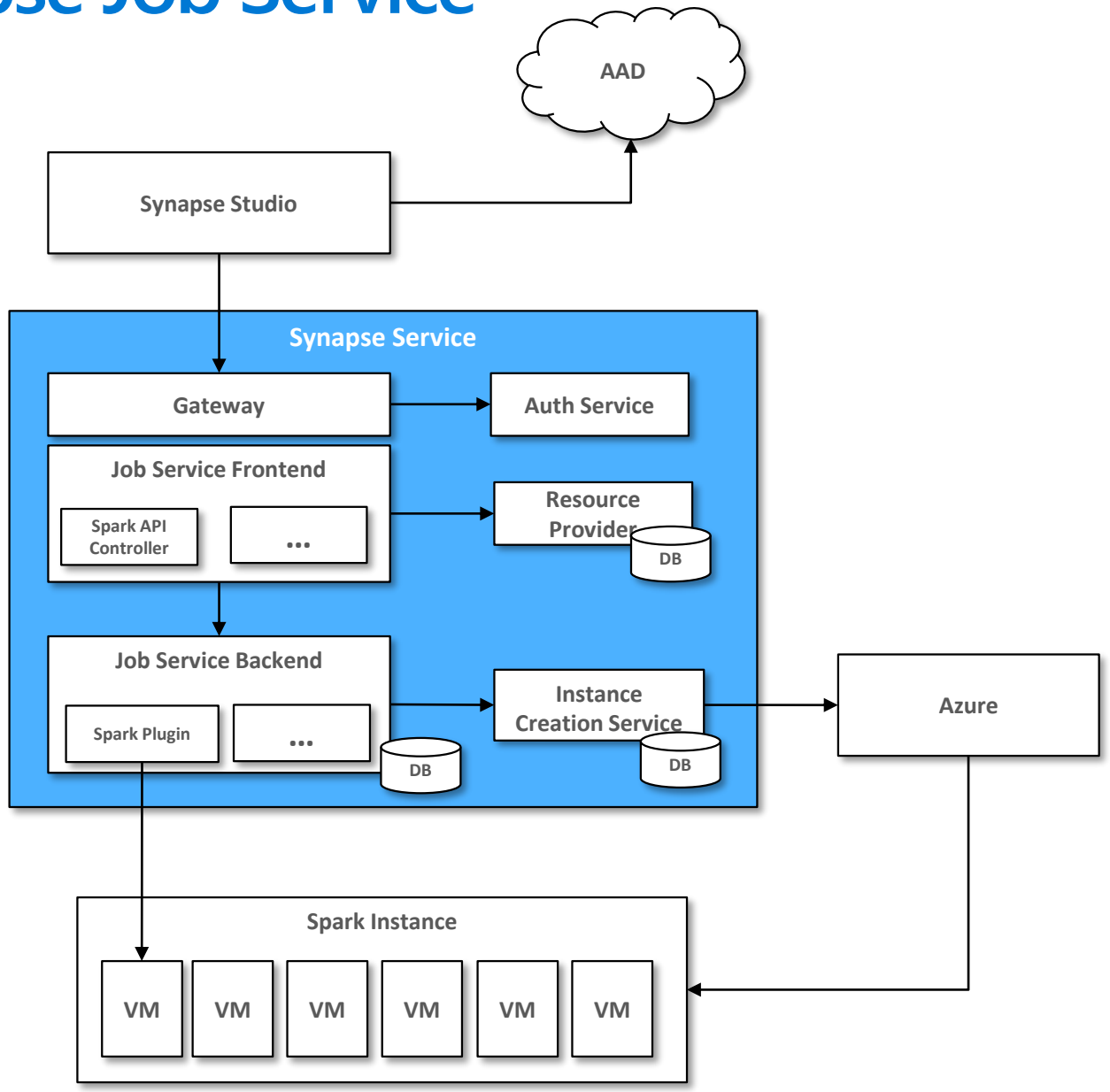
- Represent and analyze systems represented by graph nodes
- Trace interconnections between graph nodes
- Applicable to use cases in transportation, telecommunications, road networks, modeling personal relationships, social media, etc.

Microsoft

Azure Synapse Apache Spark

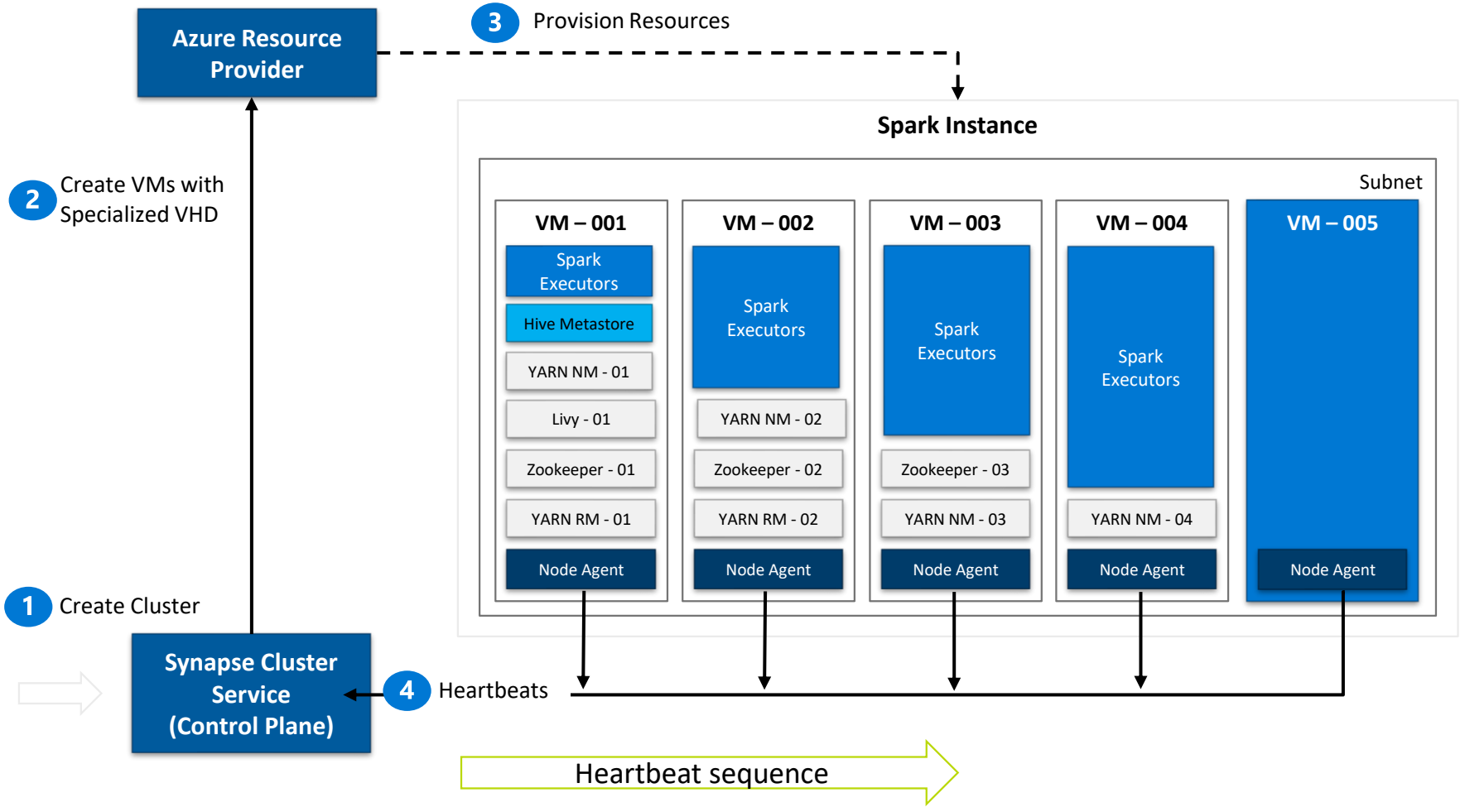
Architecture Overview

Synapse Job Service



- User creates Synapse Workspace and Spark pool and launches Synapse Studio.
- User attaches Notebook to Spark pool and enters one or more Spark statements (code blocks).
- The Notebook client gets user token from AAD and sends a Spark session create request to Synapse Gateway.
- Synapse Gateway authenticates the request and validates authorizations on the Workspace and Spark pool and forwards it to the Spark (Livy) controller hosted in Synapse Job Service frontend.
- The Job Service frontend forwards the request to Job Service backend that creates two jobs – one for creating the cluster and the other for creating the Spark session.
- The Job service backend contacts Synapse Resource Provider to obtain Workspace and Spark pool details and delegates the cluster creation request to Synapse Instance Service.
- Once the instance is created, the Job Service backend forwards the Spark session creation request to the Livy endpoint in the cluster.
- Once the Spark session is created the Notebook client sends Spark statements to the Job Service frontend.
- Job Service frontend obtains the actual Livy endpoint for the cluster created for the particular user from the backend and sends the statement directly to Livy for execution.

Synapse Spark Instances



1. Synapse Job Service sends request to Cluster Service for creating BBC clusters per the description in the associated Spark pool.
2. Cluster Service sends request to Azure using Azure SDK to create VMs (required plus additional) with specialized VHD.
3. The specialized VHD contains bits for all the services that are required by the Cluster type (for e.g. Spark) with prefetch instrumentation.
4. Once VM boots up, the Node Agent sends heartbeat to Cluster Service for getting node configuration.
5. The nodes are initialized and assigned roles based on their first heartbeat.
6. Extra nodes get deleted on first heartbeat.
7. After Cluster Service considers the cluster ready, it returns the Livy end-point to the Job Service.

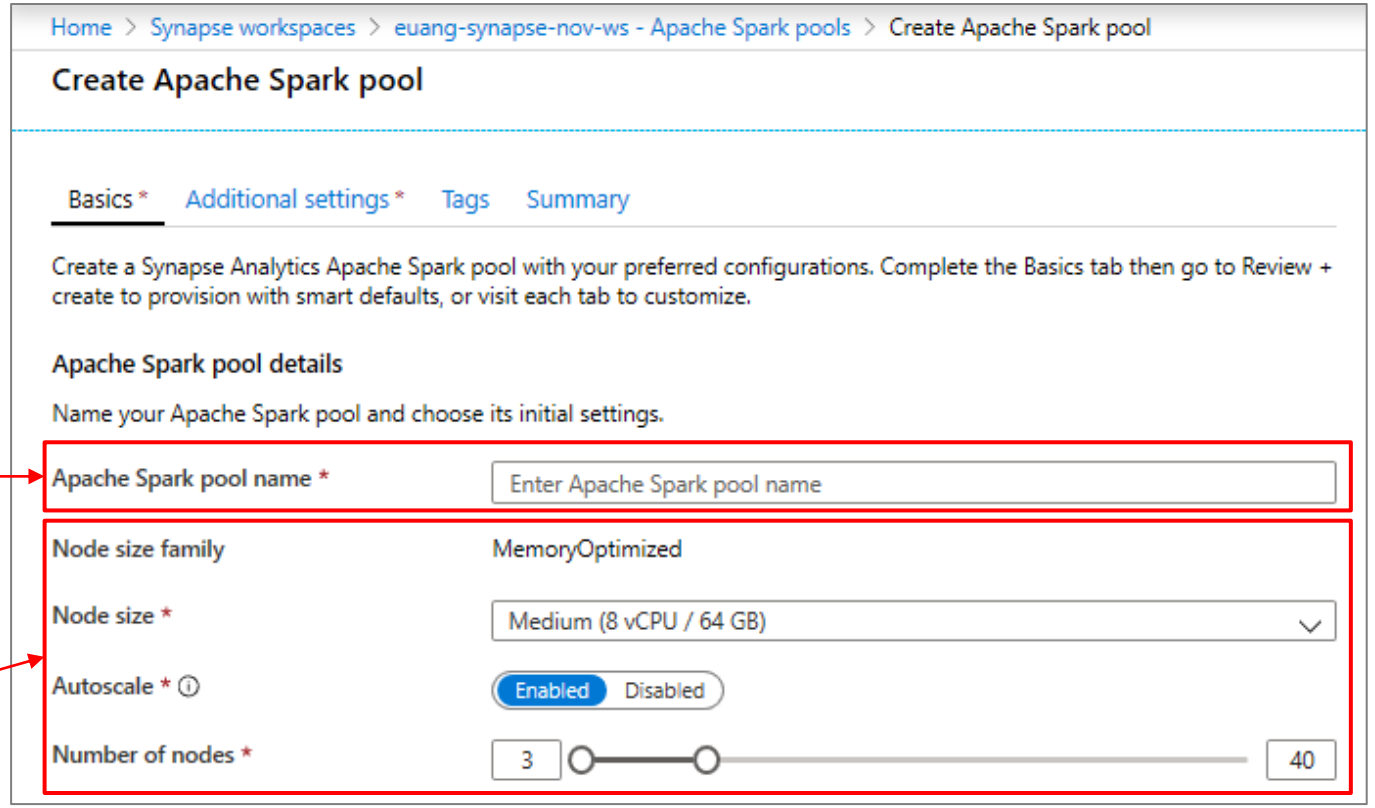
Creating a Spark pool (1 of 2)

Provision Spark Pool through Azure Portal with default settings or per requirements

Basic Settings – Minimum details required from user

Only required field from user

Default Settings

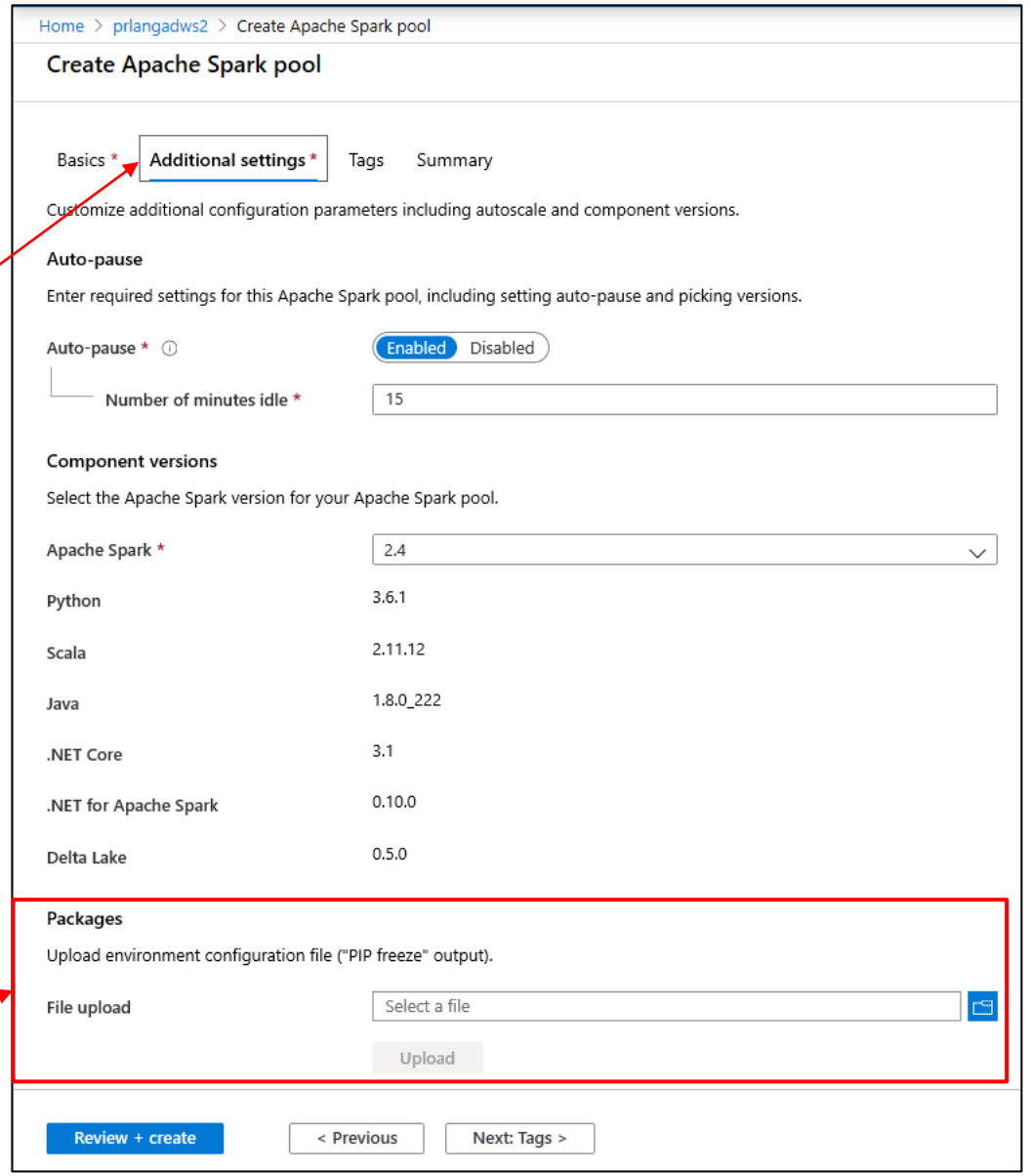


Creating a Spark pool (2 of 2) - optional

Additional Settings offer optional settings to customize Spark pool

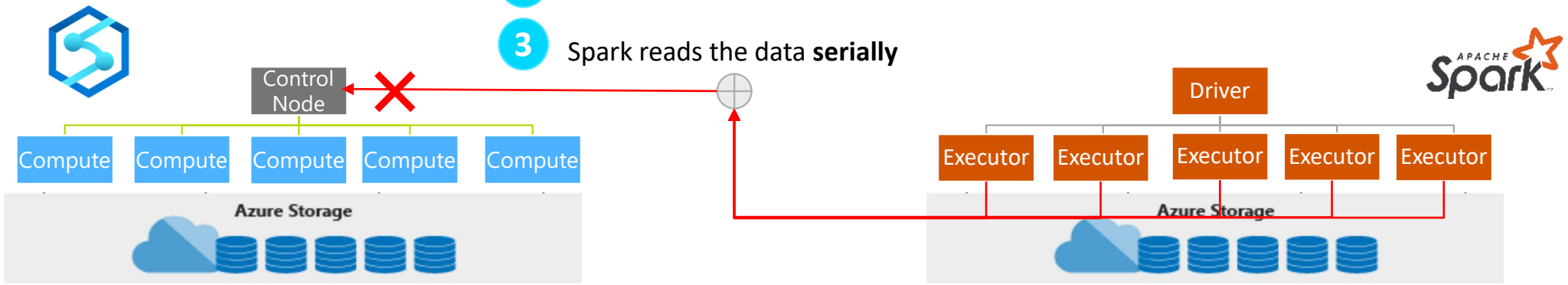
Customize component versions, auto-pause

Import libraries by providing text file containing library name and version



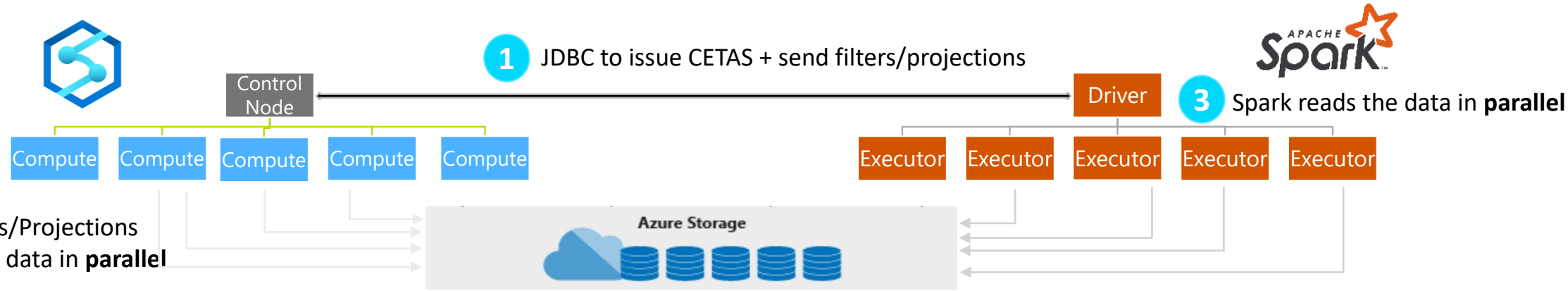
Existing Approach: JDBC

1. JDBC to open connection
2. Apply any Filters/Projections
3. Spark reads the data **serially**



New Approach: JDBC and Polybase

1. JDBC to issue CETAS + send filters/projections
2. Apply any Filters/Projections
   DW exports the data in **parallel**
3. Spark reads the data in **parallel**



User Provisioned Workspace-Default Data Lake

Code-Behind Experience

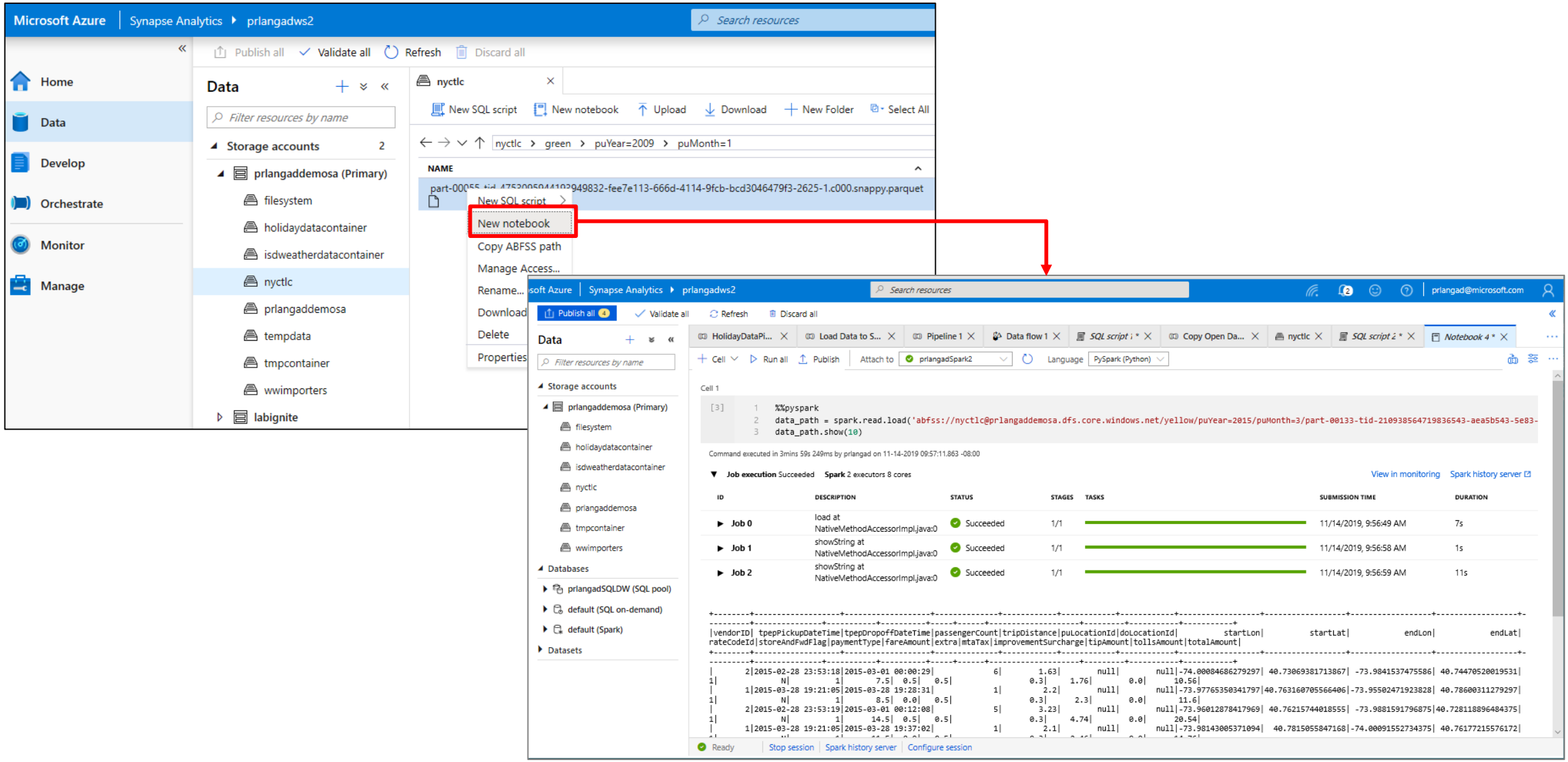
Existing Approach

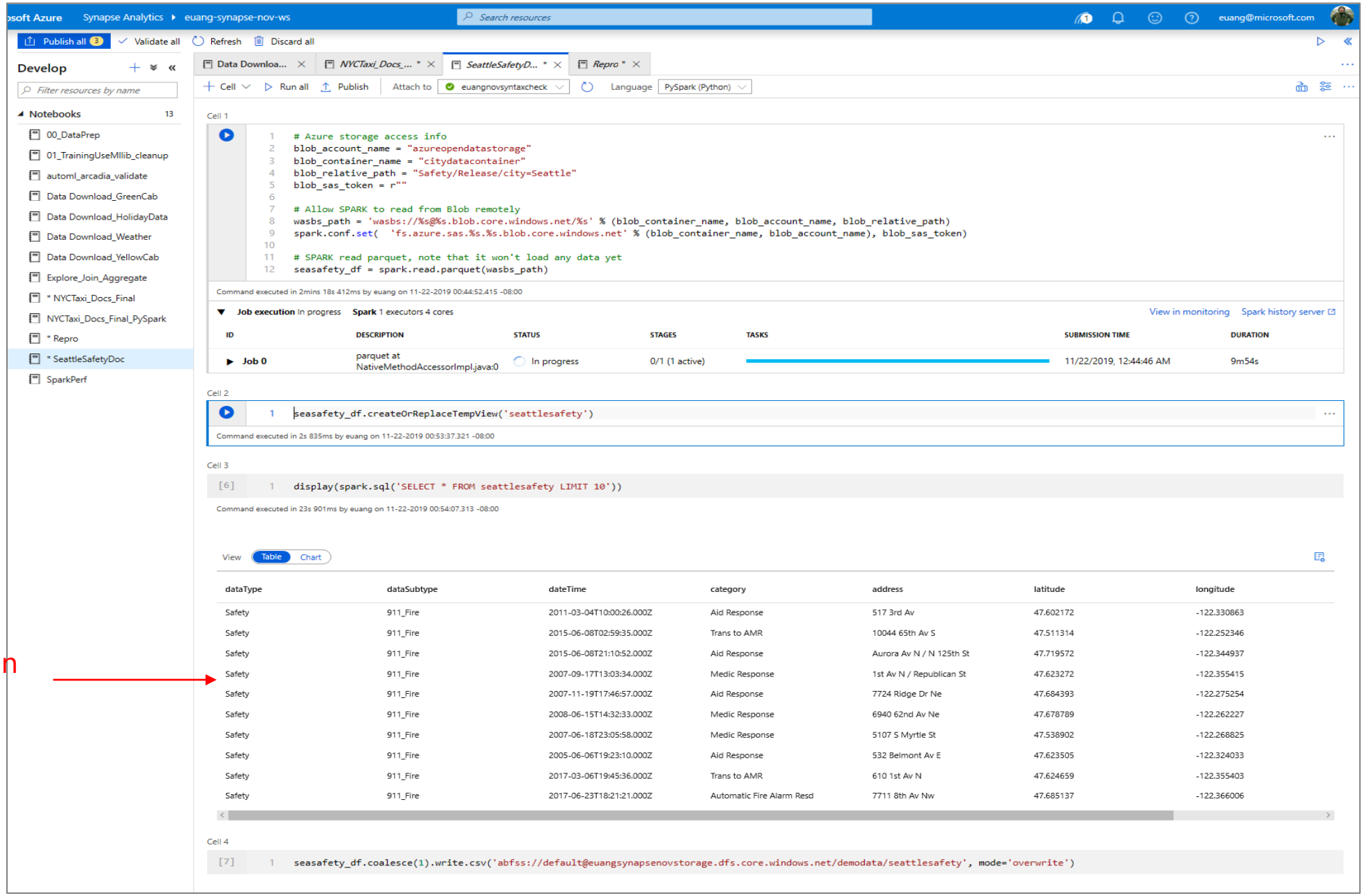
```
val jdbcUsername = "<SQL DB ADMIN USER>"
val jdbcPwd = "<SQL DB ADMIN PWD>"
val jdbcHostname = "servername.database.windows.net”
val jdbcPort = 1433
val jdbcDatabase ="<AZURE SQL DB NAME>“


val jdbc_url =
s"jdbc:sqlserver://${jdbcHostname}:${jdbcPort};database=${jdbcDatabase};
encrypt=true;trustServerCertificate=false;hostNameInCertificate=*.databas
e.windows.net;loginTimeout=60;“

val connectionProperties = new Properties()

connectionProperties.put("user", s"${jdbcUsername}")
connectionProperties.put("password", s"${jdbcPwd}")

val sqlTableDf = spark.read.jdbc(jdbc_url, “dbo.Tbl1", connectionProperties)
```

New Approach

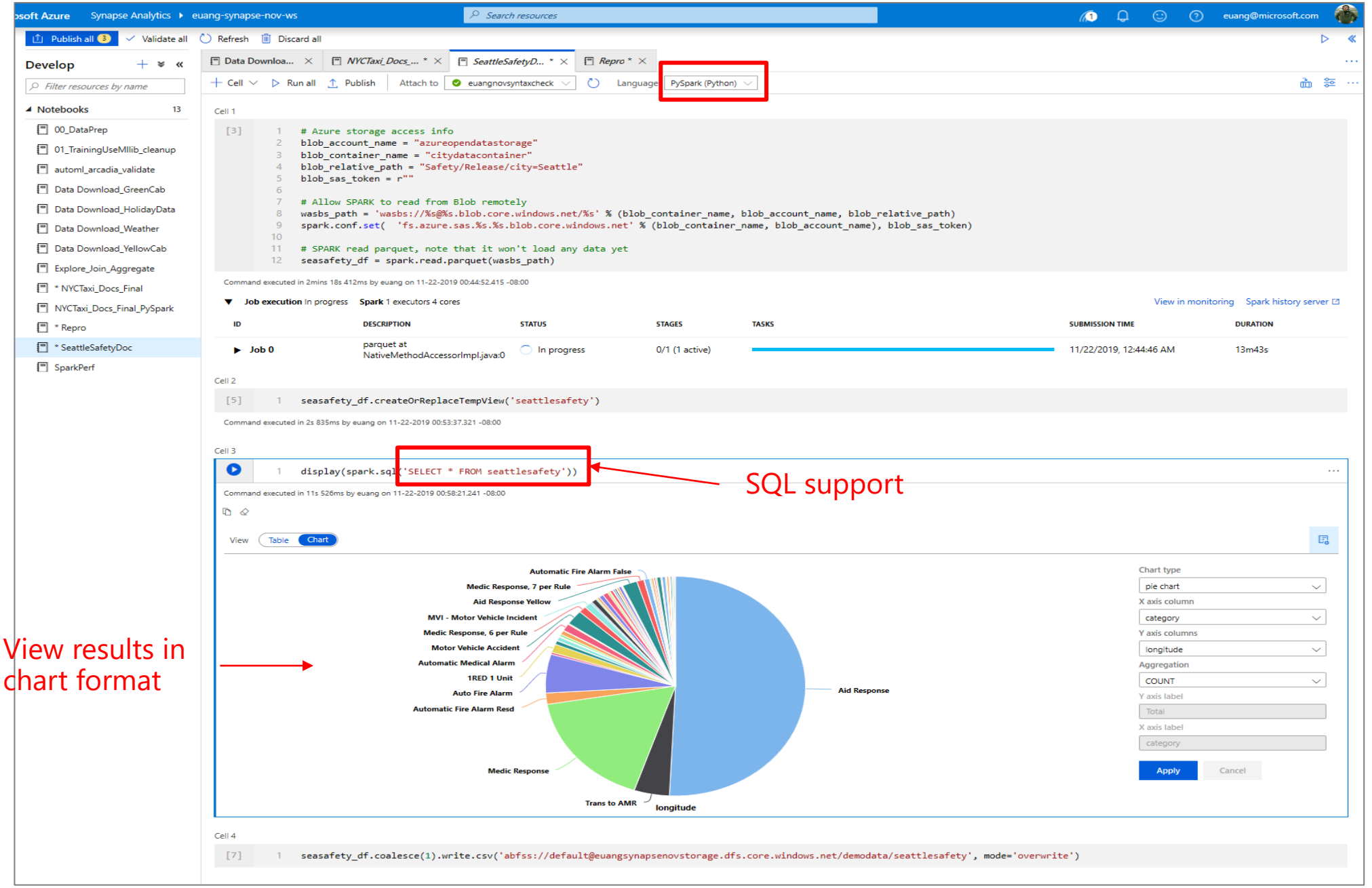
```
// Construct a Spark DataFrame from dedicated SQL pool
var df = spark.read.sqlanalytics("sql1.dbo.Tbl1")

// Write the Spark DataFrame into dedicated SQL pool
df.write.sqlanalytics(“sql1.dbo.Tbl2”)
```

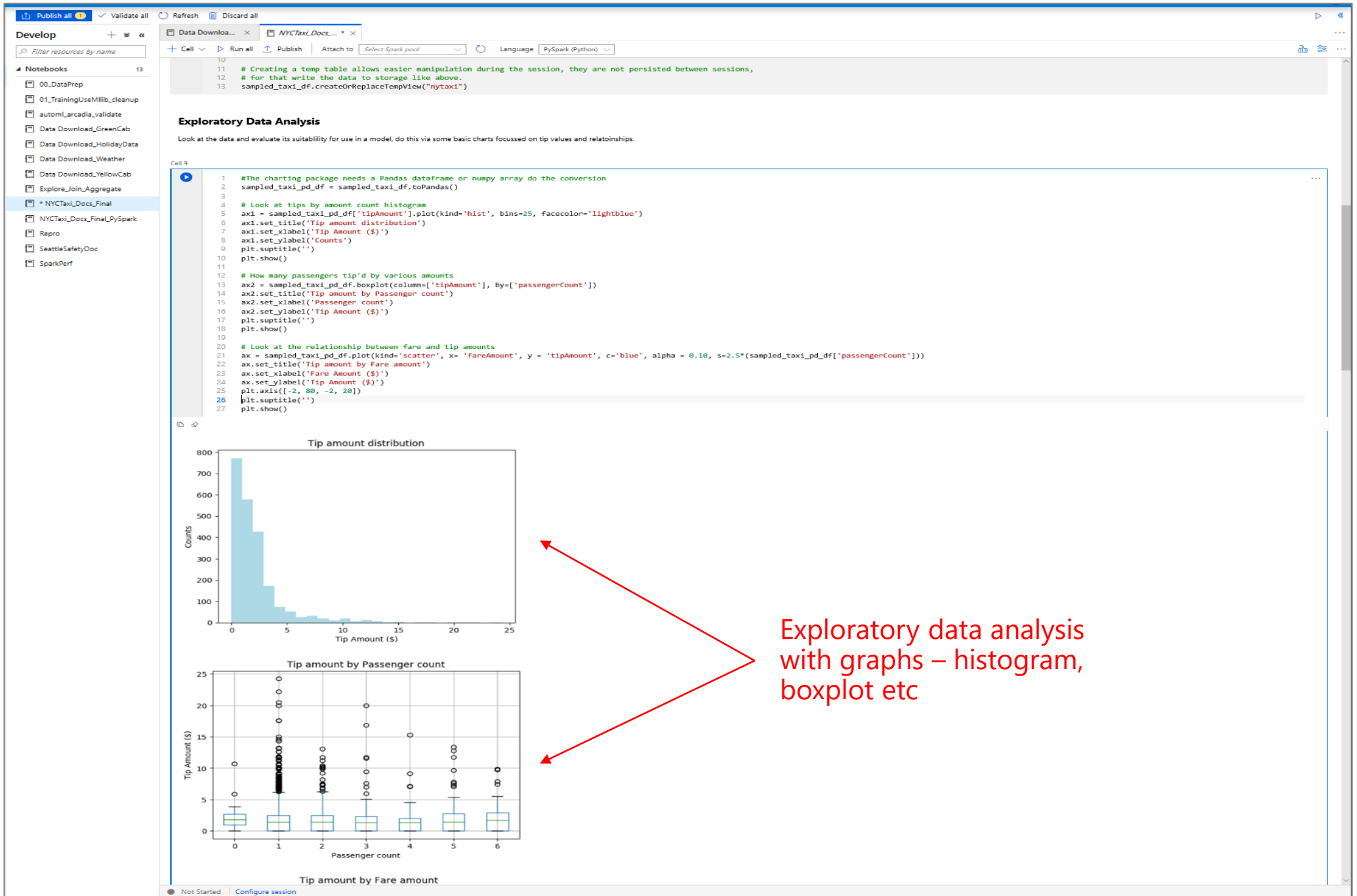
Create Notebook on files in storage





View results in table format


Synapse Analytics
euang-synapse-nov-ws
Search resources
euang@microsoft.com
Publish all
Validate all
Refresh
Discard all
Develop
Notebooks
00_DataPrep
01_TrainingUseMllib_cleanup
automl_arcadia_validate
Data Download_GreenCab
Data Download_HolidayData
Data Download_Weather
Data Download_YellowCab
Explore_Join_Aggregate
* NYCTaxi_Docs_Final
NYCTaxi_Docs_Final_PySpark
* Repro
* SeattleSafetyDoc
SparkPerf
Cell
Run all
Publish
Attach to
euangnovsyntaxcheck
Language
PySpark (Python)
Cell 1
# Azure storage access info
blob_account_name = "azureopendatastorage"
blob_container_name = "citydatacontainer"
blob_relative_path = "Safety/Release/city=Seattle"
blob_sas_token = r""
# Allow SPARK to read from Blob remotely
wasbs_path = 'wasbs://%s@%s.blob.core.windows.net/%s' % (blob_container_name, blob_account_name, blob_relative_path)
spark.conf.set('fs.azure.sas.%s.%s.blob.core.windows.net' % (blob_container_name, blob_account_name), blob_sas_token)
# SPARK read parquet, note that it won't load any data yet
seasafety_df = spark.read.parquet(wasbs_path)
Command executed in 2mins 18s 412ms by euang on 11-22-2019 00:44:52.415 -08:00
Job execution In progress
Spark 1 executors 4 cores
View in monitoring
Spark history server
ID
DESCRIPTION
STATUS
STAGES
TASKS
SUBMISSION TIME
DURATION
Job 0
parquet at NativeMethodAccessorImpl.java:0
In progress
0/1 (1 active)
11/22/2019, 12:44:46 AM
13m43s
Cell 2
seasafety_df.createOrReplaceTempView('seattlesafety')
Command executed in 2s 835ms by euang on 11-22-2019 00:53:37.321 -08:00
Cell 3
display(spark.sql('SELECT * FROM seattlesafety'))
Command executed in 11s 526ms by euang on 11-22-2019 00:58:21.241 -08:00
View
Table
Chart
Automatic Fire Alarm False
Medic Response, 7 per Rule
Aid Response Yellow
MVI - Motor Vehicle Incident
Medic Response, 6 per Rule
Motor Vehicle Accident
Automatic Medical Alarm
1RED 1 Unit
Auto Fire Alarm
Automatic Fire Alarm Resd
Medic Response
Trans to AMR
Aid Response
longitude
Chart type
pie chart
X axis column
category
Y axis columns
longitude
Aggregation
COUNT
Y axis label
Total
X axis label
category
Apply
Cancel
Cell 4
seasafety_df.coalesce(1).write.csv('abfss://default@euangsynapsenovstorage.dfs.core.windows.net/demodata/seattlesafety', mode='overwrite')
SQL support
View results in chart format



Exploratory data analysis with graphs – histogram, boxplot etc

Library Management - Python

Overview

Customers can add new python libraries at Spark pool level

Benefits

Input requirements.txt in simple pip freeze format

Add new libraries to your cluster

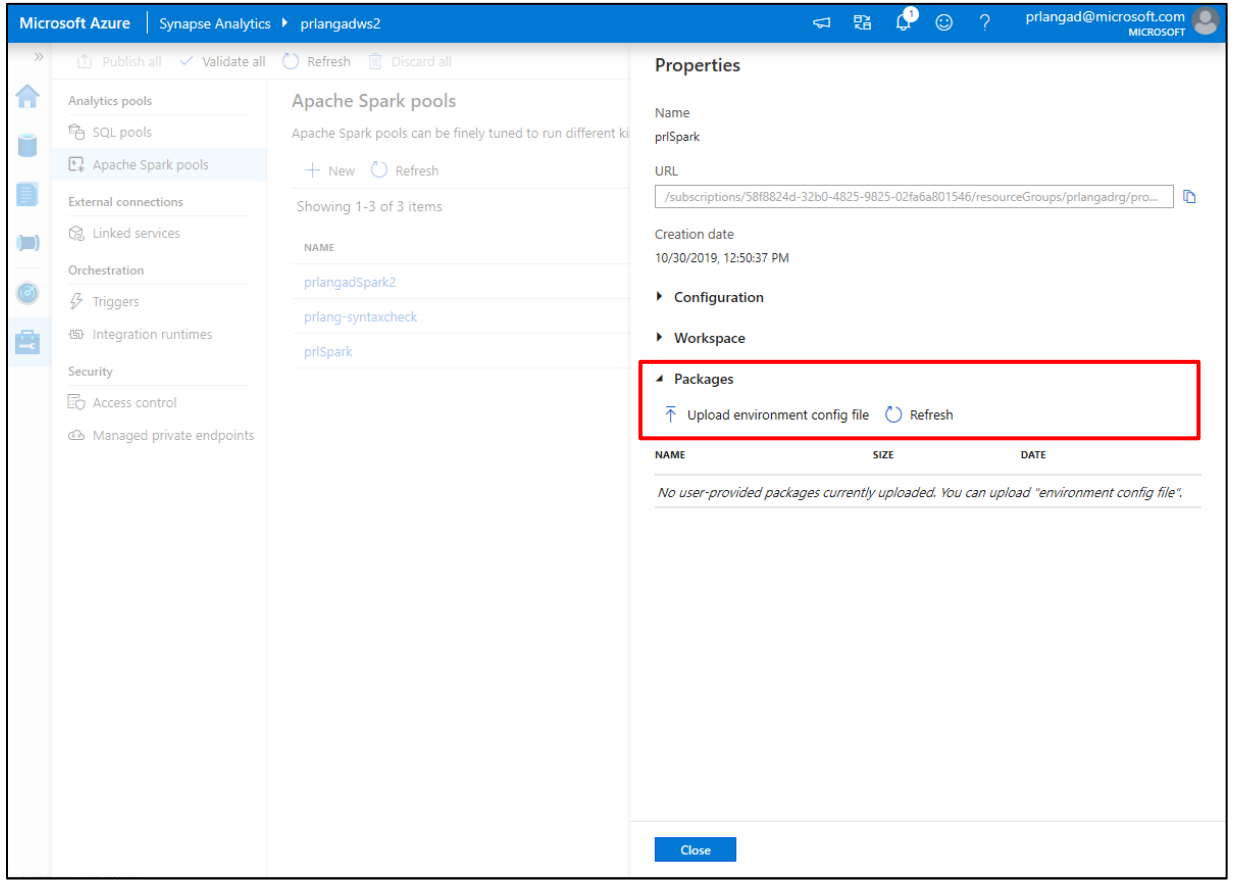
Update versions of existing libraries on your cluster

Ability to specify different requirements file for different pools within the same workspace

Constraints

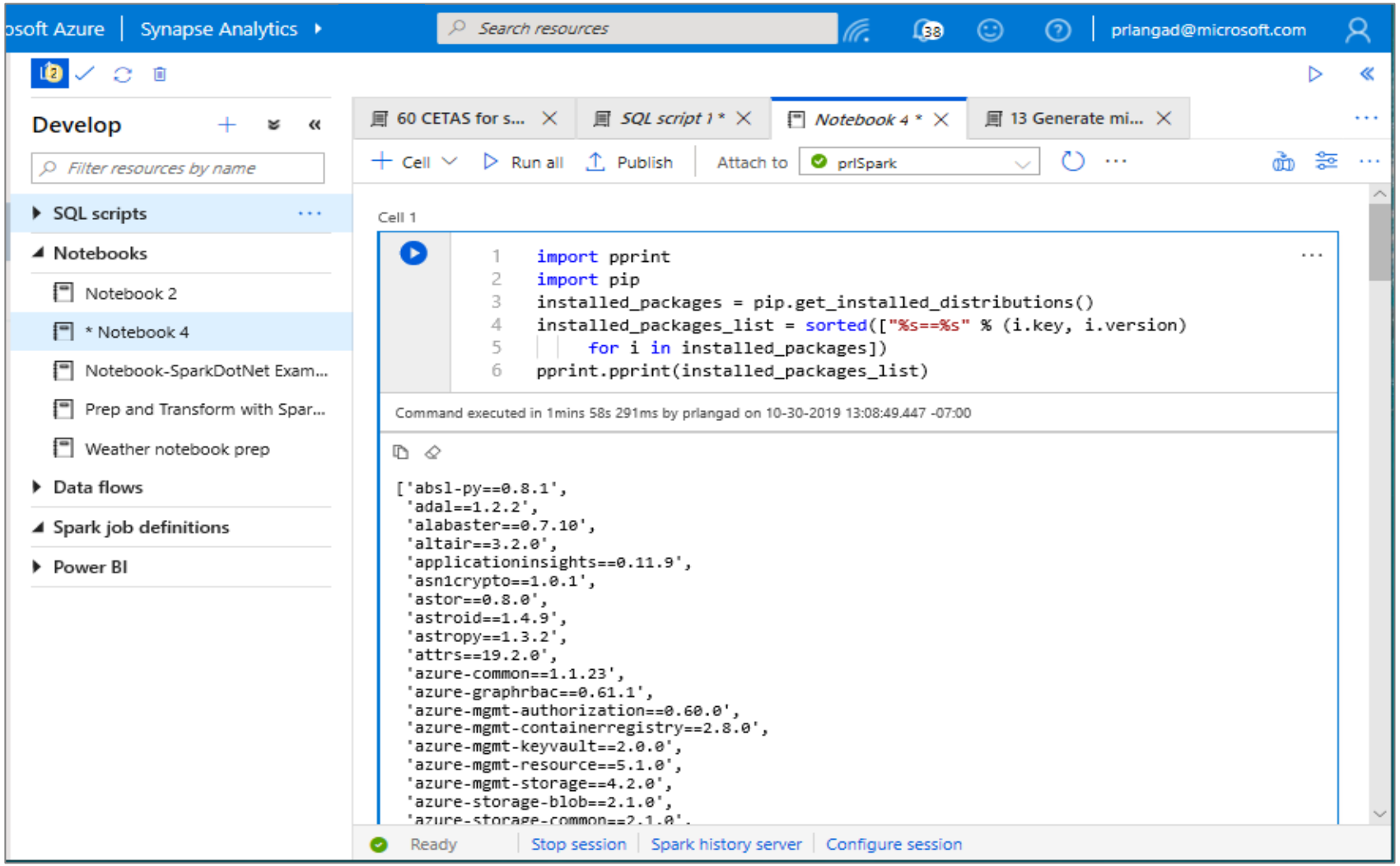
The library version must exist on PyPI repository

Version downgrade of an existing library not allowed



Library Management - Python

Get list of installed libraries with version information



Spark ML Algorithms

Classification and Regression	• Linear Models (SVMs, logistic regression, linear regression) • Naïve Bayes • Decision Trees • Ensembles of trees (Random Forest, Gradient-Boosted Trees) • Isotonic regression
Clustering	• k-means and streaming k-means • Gaussian mixture • Power iteration clustering (PIC) • Latent Dirichlet allocation (LDA)
Collaborative Filtering	• Alternating least squares (ALS)
Dimensionality Reduction	• SVD • PCA
Frequent Pattern Mining	• FP-growth • Association rules
Basic Statistics	• Summary statistics • Correlations • Stratified sampling • Hypothesis testing • Random data generation

Microsoft Machine Learning for Apache Spark

v1.0-rc

Microsoft's Open Source Contributions to Apache Spark

Distributed Machine Learning

Fast Model Deployment

Microservice Orchestration

Multilingual Binding Generation

www.aka.ms/spark

Azure/mmlspark

Microsoft Spark Utilities

Overview

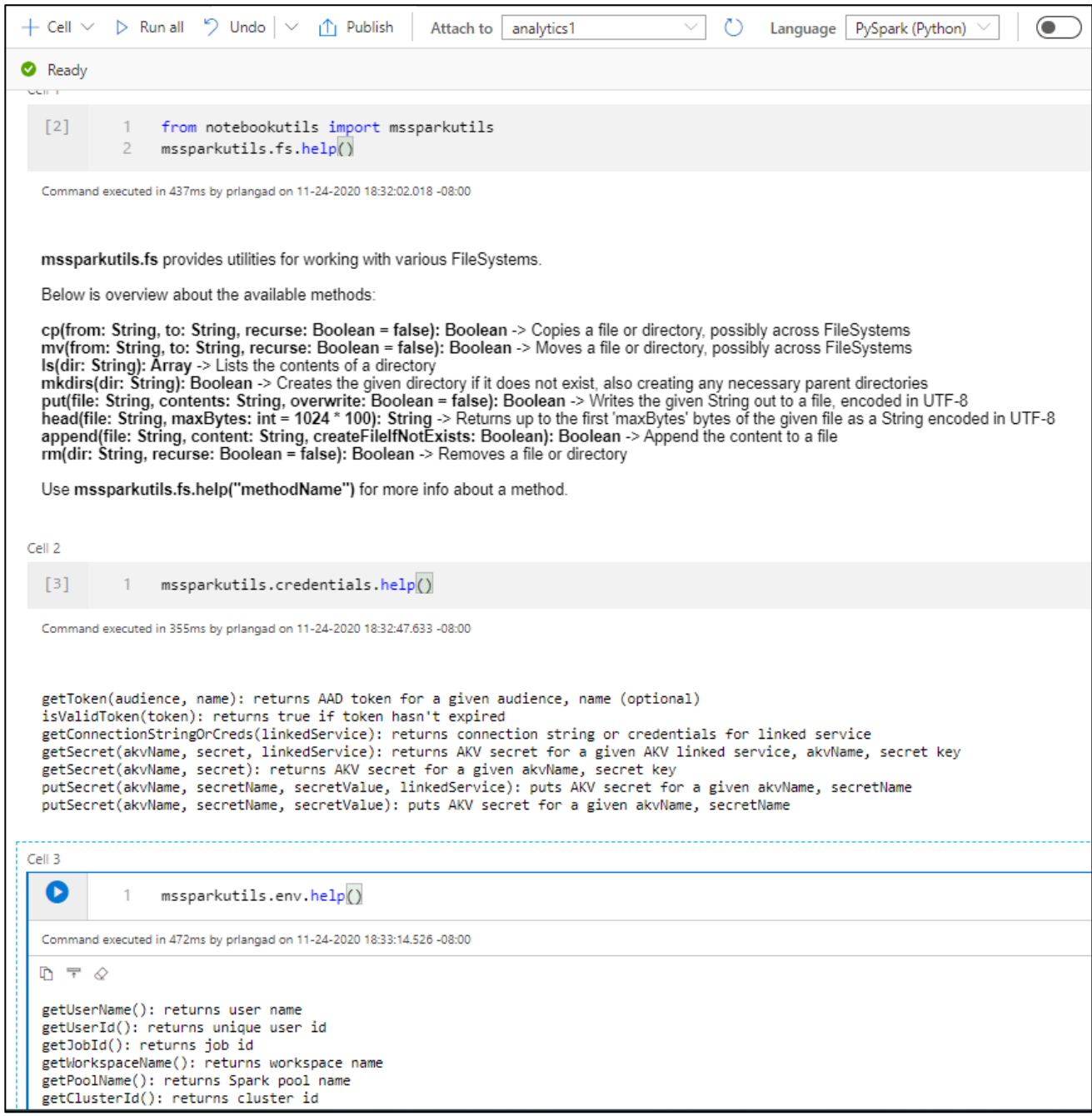
It provides utilities for working with file systems, including ADLS Gen2 and Azure Blob Storage.

Benefits

It supports multiple methods for file systems such as List, Copy, Move, Write, Append, Delete file or directory, View file properties, Create new directory, Preview file content.

It supports environment utilities to get username, user id, job id, workspace name, pool name, cluster id.

It supports to get the access tokens of linked services and manage secrets in Azure Key Vault.



Hyperspace

Overview

Hyperspace introduces the ability for Apache Spark users to create indexes on their data

Benefits

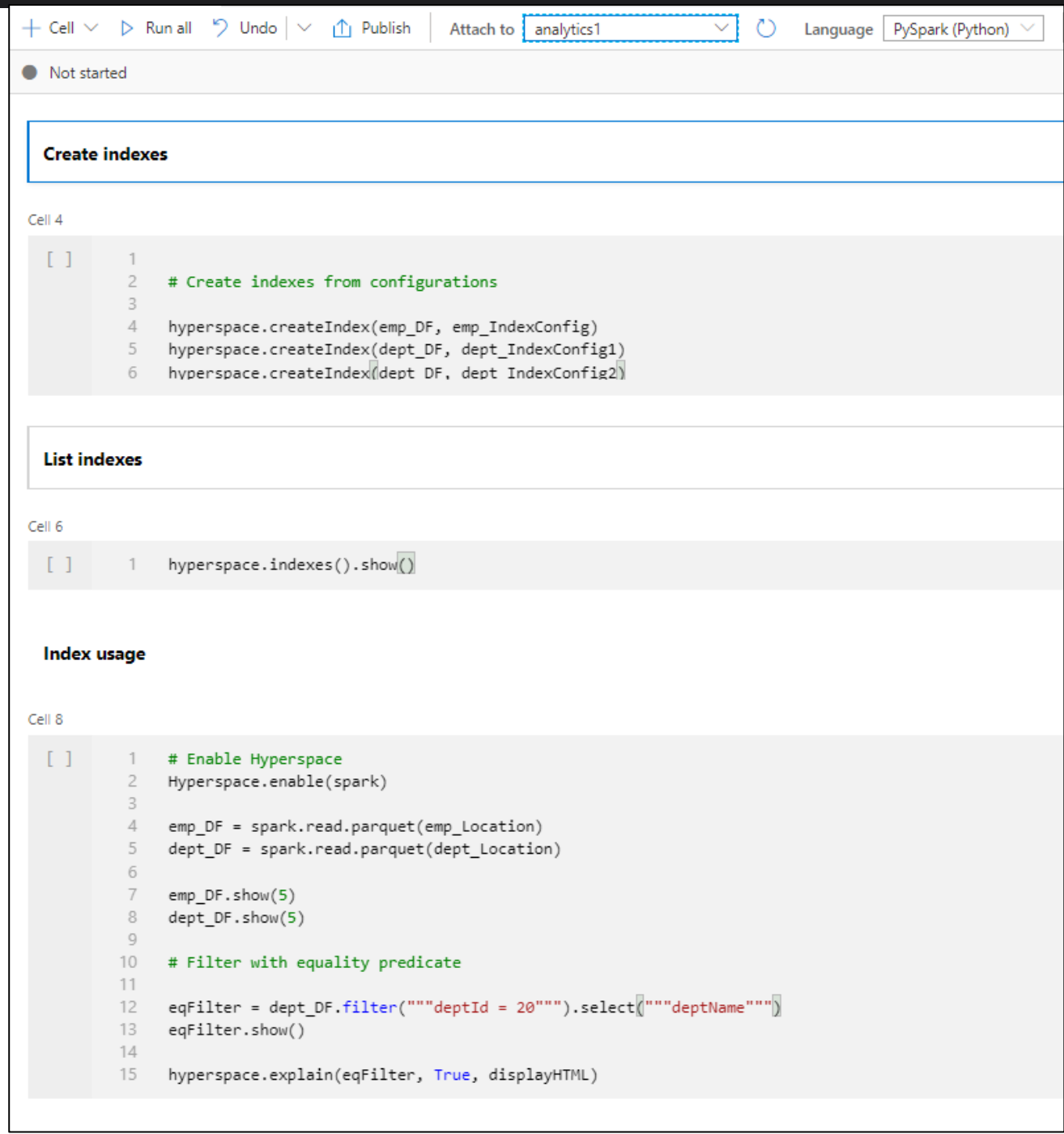
It helps accelerate your workloads or queries containing filters on predicates with high selectivity or a join that requires heavy shuffles.

Maintain the indexes through a multi-user concurrency model.

Leverage these indexes automatically, within your Spark workloads, without any changes to your application code for query/workload acceleration.

It supports index operations as create index, list index, restore index, delete index, vacuum index

Languages supported: Scala, Python, .NET



Spark CDM connector

Overview

It offers Spark dataframes to read and write entities in a CDM folder.

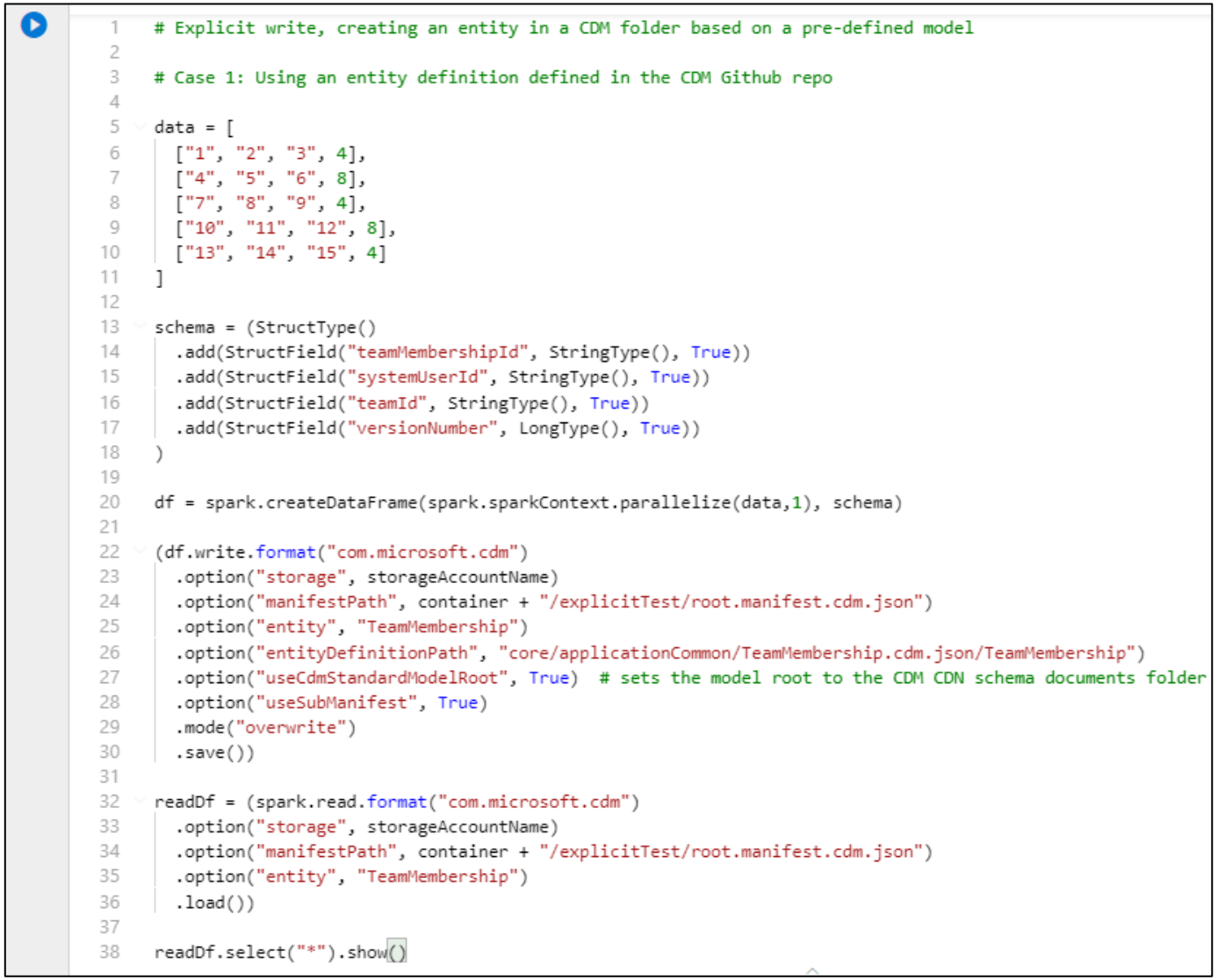
Benefits

It supports use of Managed Identities for Azure resources to mediate access to the Azure datalake storage.

Writes from a Spark dataframe to an entity in a CDM folder based on dataframe schema or CDM entity definition.

Supports data in Apache Parquet format and CSV format.

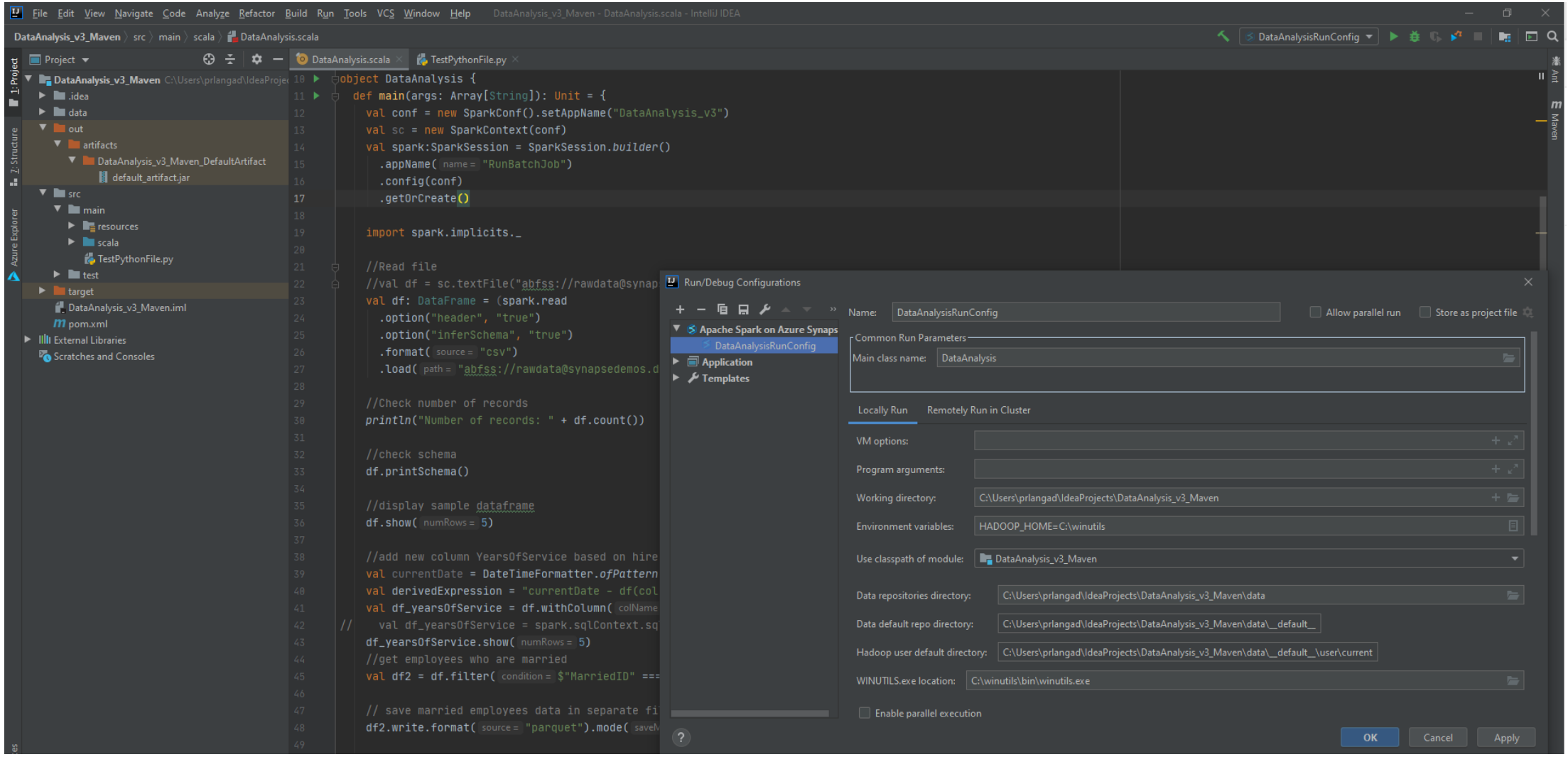
The CDM connector is pre-installed and supports languages: Python, Scala



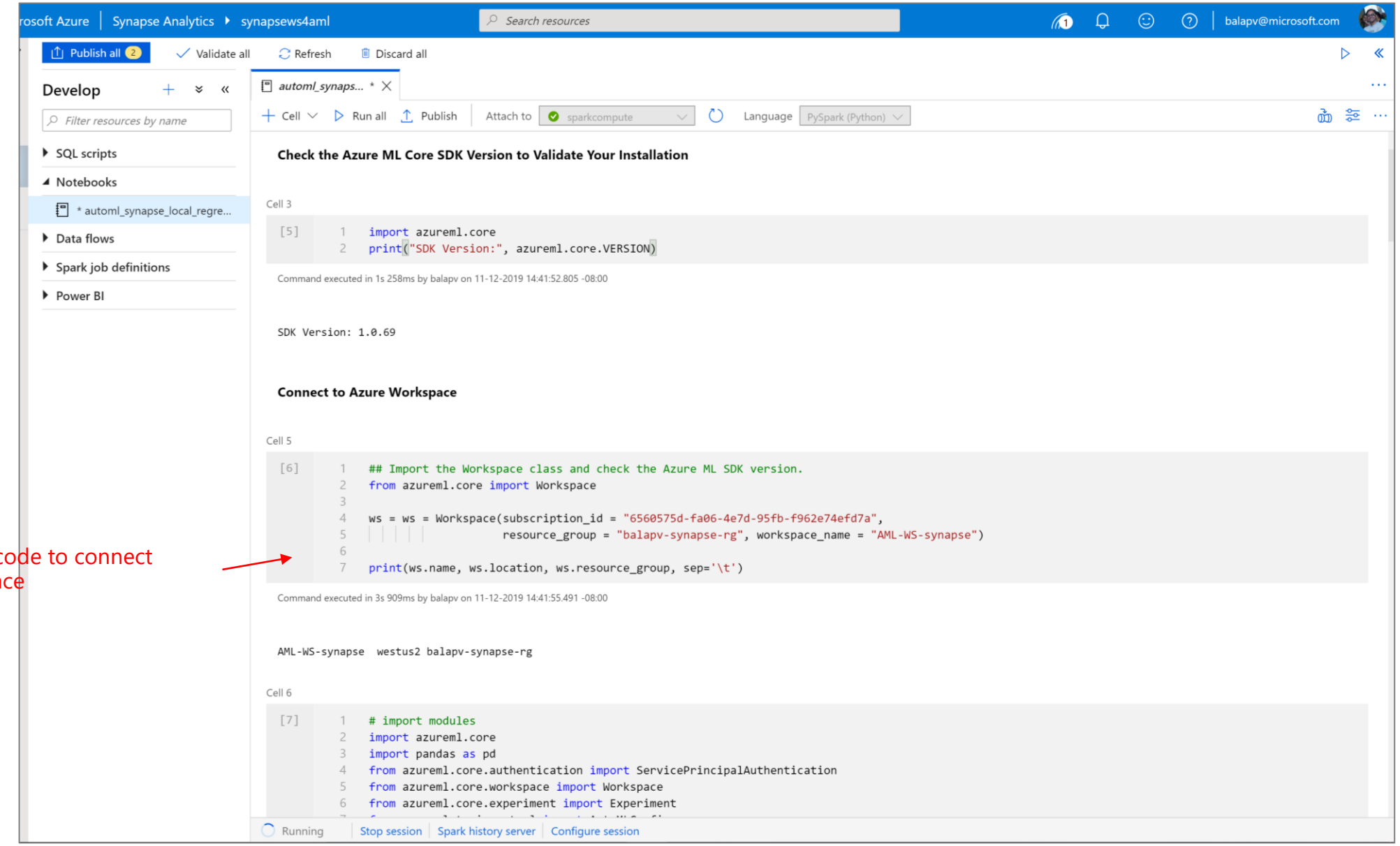
```python
# Explicit write, creating an entity in a CDM folder based on a pre-defined model

# Case 1: Using an entity definition defined in the CDM Github repo

data = [
  ["1", "2", "3", 4],
  ["4", "5", "6", 8],
  ["7", "8", "9", 4],
  ["10", "11", "12", 8],
  ["13", "14", "15", 4]
]

schema = (StructType()
  .add(StructField("teamMembershipId", StringType(), True))
  .add(StructField("systemUserId", StringType(), True))
  .add(StructField("teamId", StringType(), True))
  .add(StructField("versionNumber", LongType(), True))
)

df = spark.createDataFrame(spark.sparkContext.parallelize(data,1), schema)

(df.write.format("com.microsoft.cdm")
  .option("storage", storageAccountName)
  .option("manifestPath", container + "/explicitTest/root.manifest.cdm.json")
  .option("entity", "TeamMembership")
  .option("entityDefinitionPath", "core/applicationCommon/TeamMembership.cdm.json/TeamMembership")
  .option("useCdmStandardModelRoot", True)  # sets the model root to the CDM CDN schema documents folder
  .option("useSubManifest", True)
  .mode("overwrite")
  .save())

readDf = (spark.read.format("com.microsoft.cdm")
  .option("storage", storageAccountName)
  .option("manifestPath", container + "/explicitTest/root.manifest.cdm.json")
  .option("entity", "TeamMembership")
  .load())

readDf.select("*").show()
```

IntelliJ IDE

Use the Azure Toolkit for IntelliJ plug-in to develop Apache Spark applications and submit applications to Spark pools.

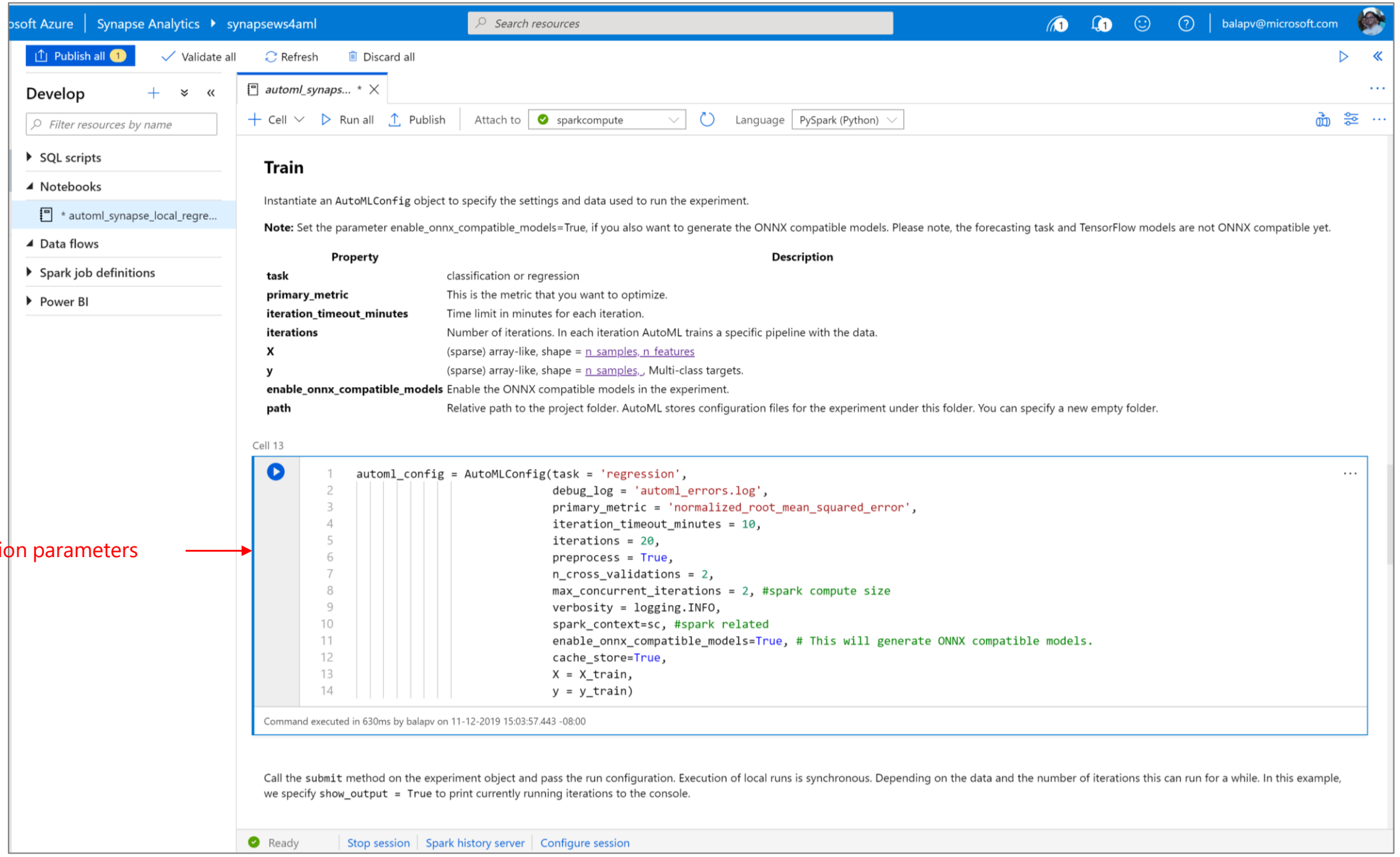


Synapse Notebook: Connect to AML workspace



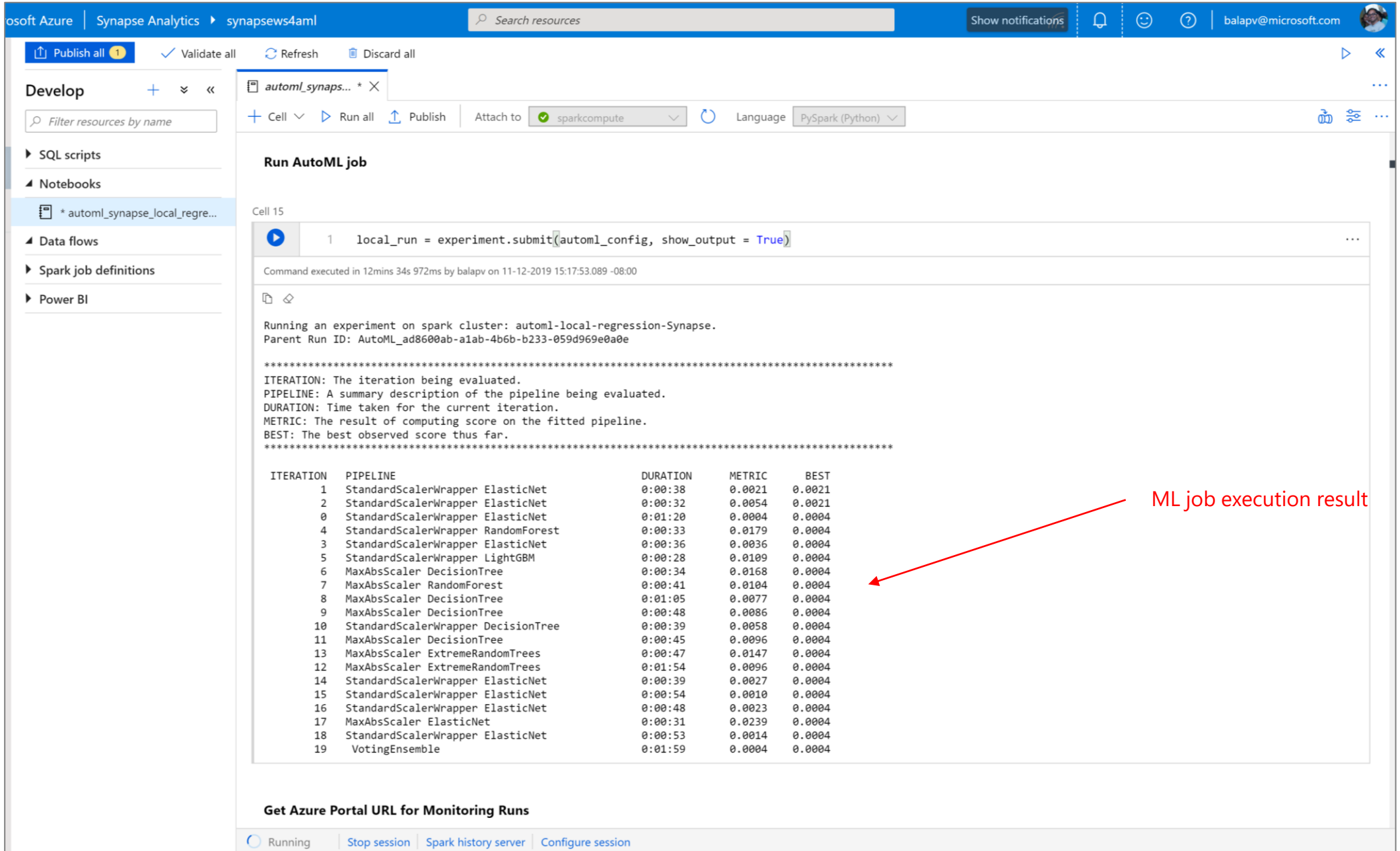
Simple code to connect workspace

Synapse Notebook: Configure AML job to run on Synapse



Configuration parameters

Synapse Notebook: Run AML job



Microsoft

Azure Synapse Analytics

Security

Securing with firewalls

Overview

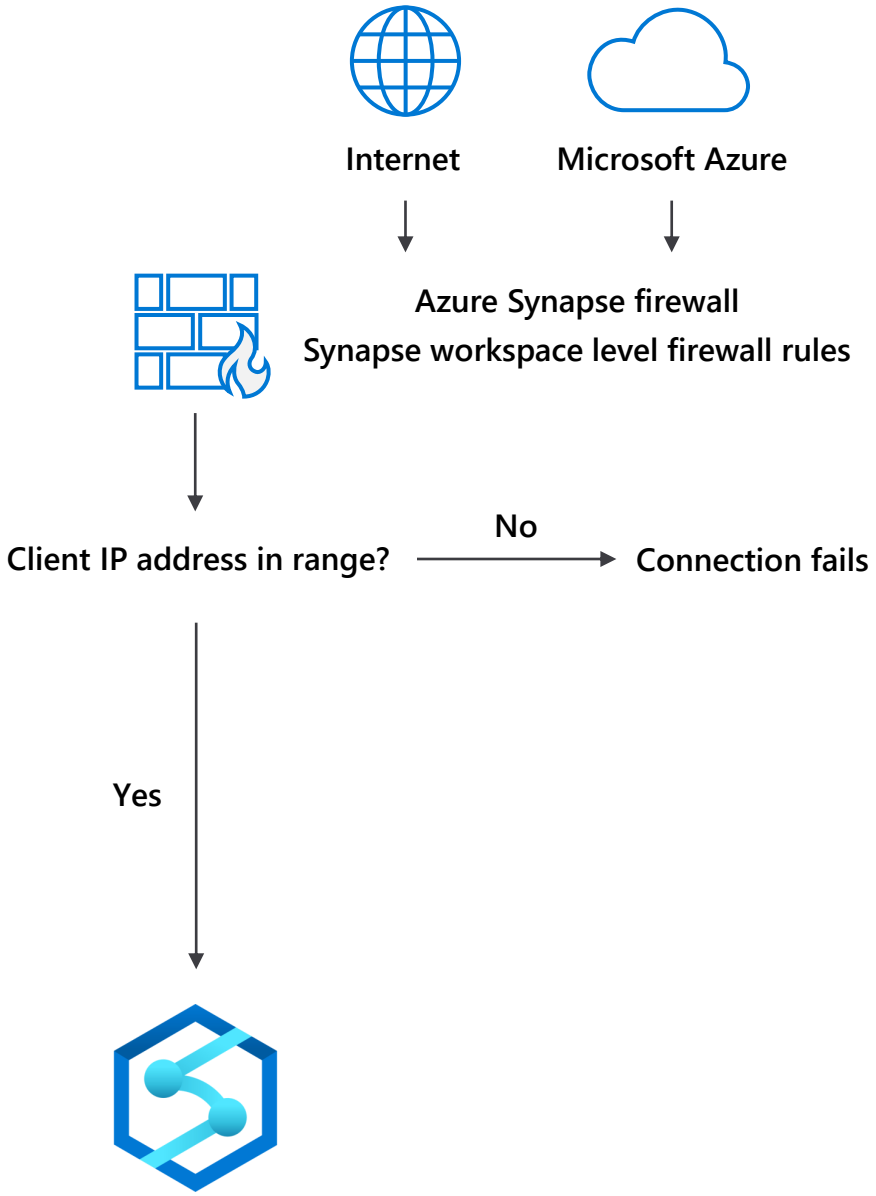
Access to your Azure Synapse Analytics is blocked by the firewall.

Firewall also manages virtual network rules that are based on virtual network service endpoints.

Rules

Allow specific or range of whitelisted IP addresses.

Allow Azure applications to connect.



IP Firewall Rules

Overview

IP firewall rules grant or deny access to user's Synapse workspace based on the originating IP address of each request.

IP firewall rules configured at the workspace level apply to all public endpoints of the workspace (dedicated SQL pool, serverless SQL pool, and Development).

Key Points

- Customers can also use SQL Server Management Studio (SSMS) to connect to the SQL resources (dedicated SQL pool and serverless SQL pool) in their workspace.
- Customers must ensure that the firewall on the network and local computer allow outgoing communication on TCP ports 80, 443 and 1443 for Synapse Studio.
- Customers must also allow outgoing communication on UDP port 53 for Synapse Studio.
- To connect using tools such as SSMS and Power BI, user must allow outgoing communication on TCP port 1433.

Managed VNet

Overview

Creating a workspace with a Managed workspace VNet associated with it ensures that user's workspace is network isolated from other workspaces. The VNet associated with your workspace is managed by Azure Synapse. This VNet is called a Managed workspace VNet.

Benefits

- With a Managed workspace customers can offload the burden of managing the VNet to Azure Synapse.
- Customers don't have to configure inbound NSG rules on their own VNets to allow Azure Synapse management traffic to enter their VNet.
- Customers don't need to create a subnet for your Spark clusters based on peak load.
- Managed workspace VNet along with Managed private endpoints protects against data exfiltration.

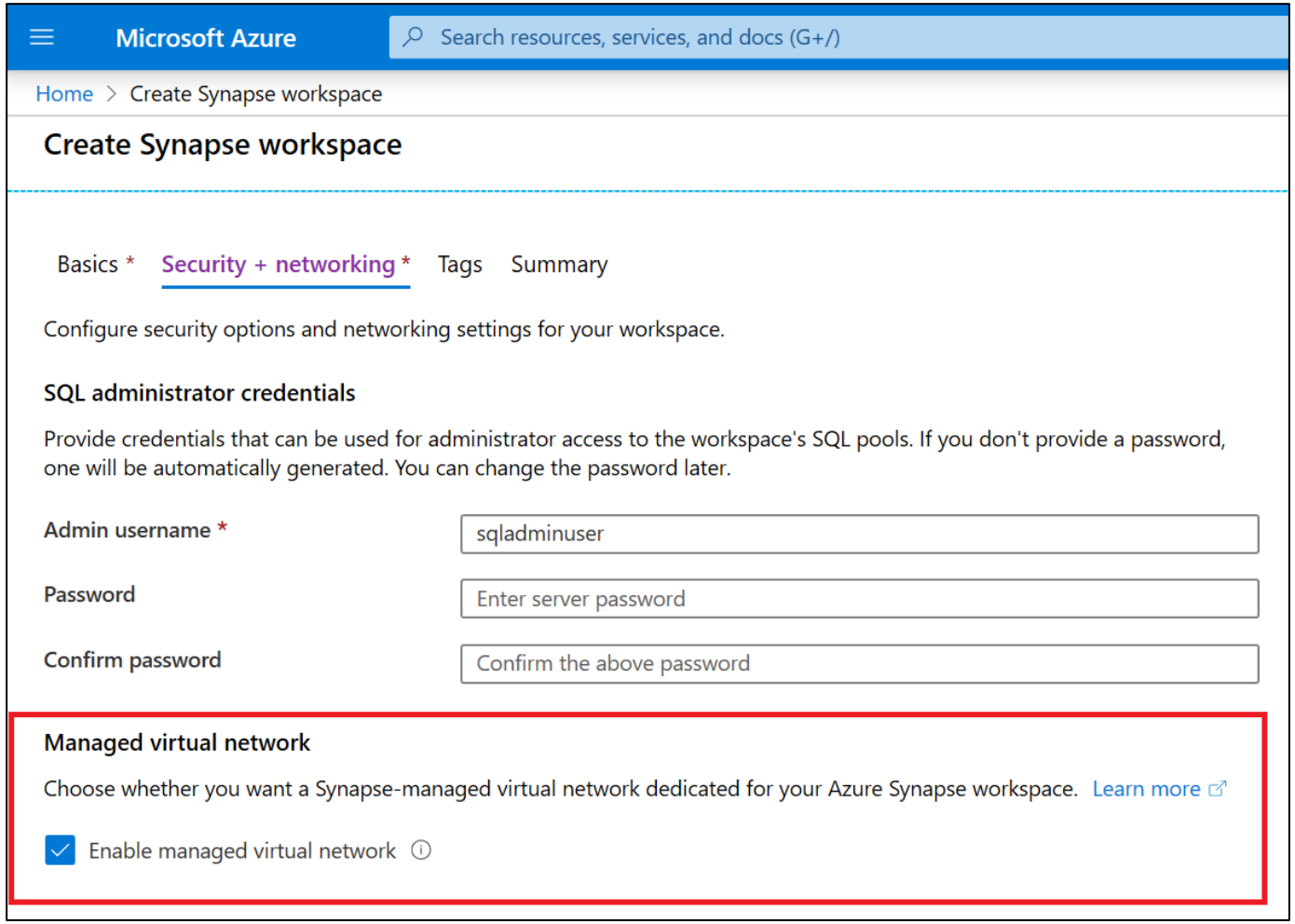
Managed VNet

Overview

During Azure Synapse workspace creation, user can choose to associate it to a managed VNet. User cannot change this workspace configuration after the workspace is created.

User cannot reconfigure a workspace that does not have a Managed workspace VNet associated with it and associate a VNet to it.

Private links are supported only in Synapse workspaces that have a managed VNet associated to it.



Private Endpoints

Overview

Managed private endpoints are private endpoints created in the Managed workspace VNet establishing a private link to Azure resources. Managed private endpoints are only supported in Azure Synapse workspaces with a Managed workspace VNet.

Benefits

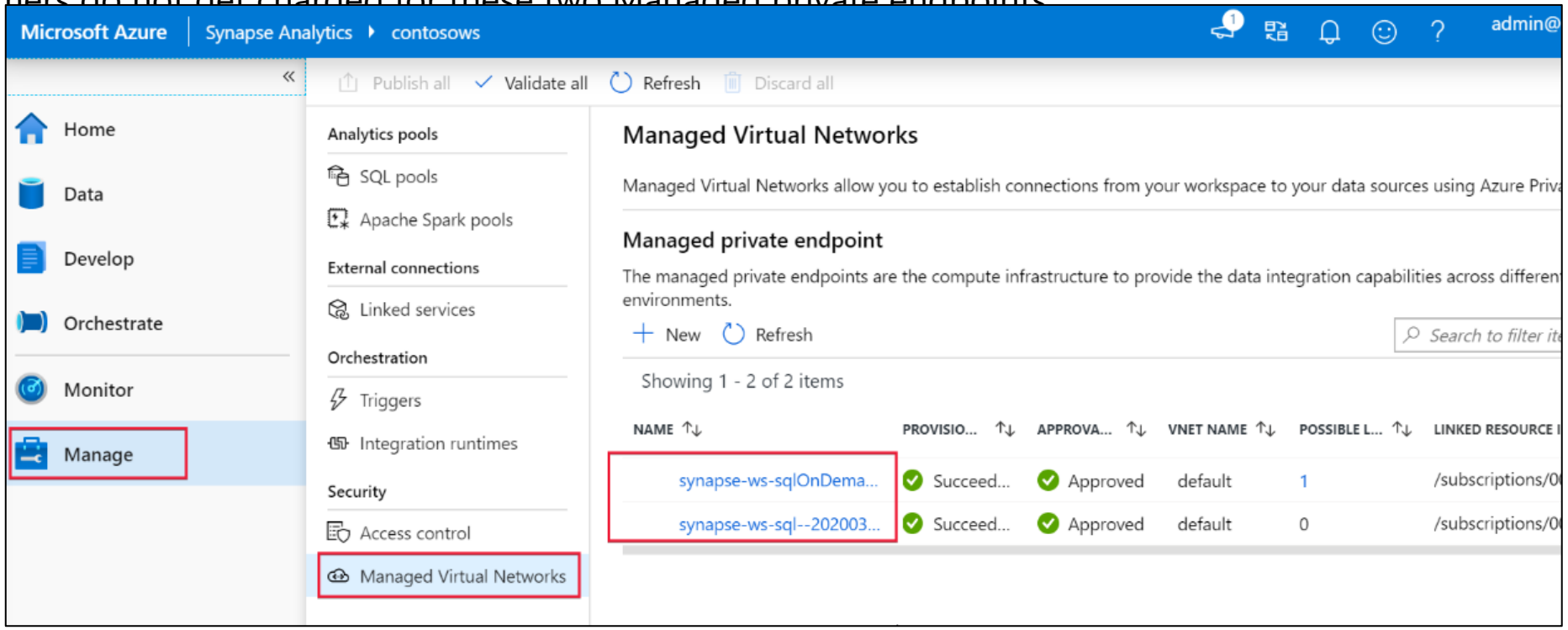
- Private link enables customers to access Azure services and Azure hosted customer/partner services from their Azure VNet securely.
- With use of private link, traffic between customer's VNet and workspace traverses entirely over the Microsoft backbone network.
- Private link protects against data exfiltration risks.
- Private endpoint uses a private IP address from customer's VNet to effectively bring the service into their VNet.
- Private endpoints are mapped to a specific resource in Azure and not the entire service.

Private Endpoints for Synapse SQL (provisioned & serverless)

Overview

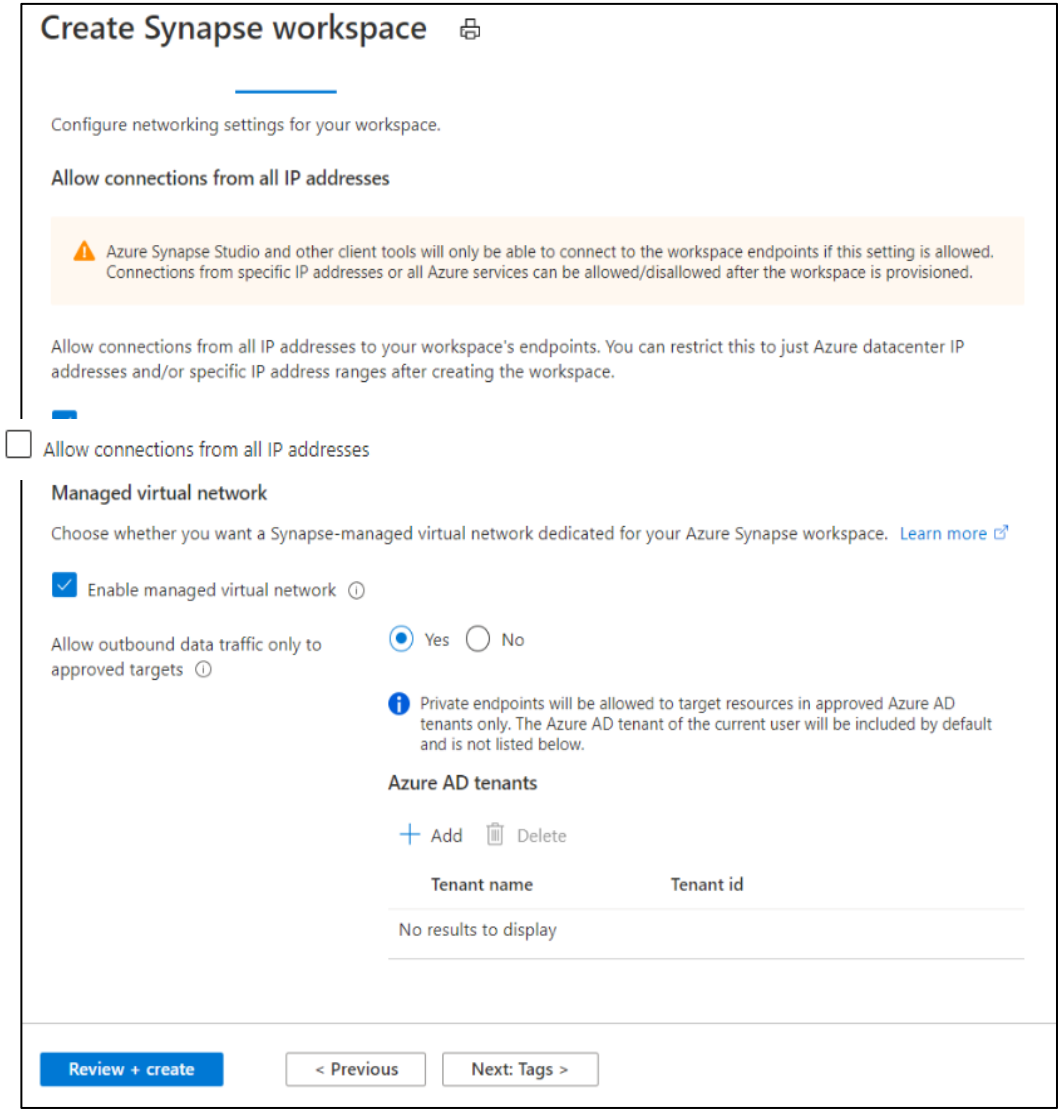
Dedicated SQL pool and serverless SQL pool use multi-tenant infrastructure that is not deployed into the Managed workspace VNet.

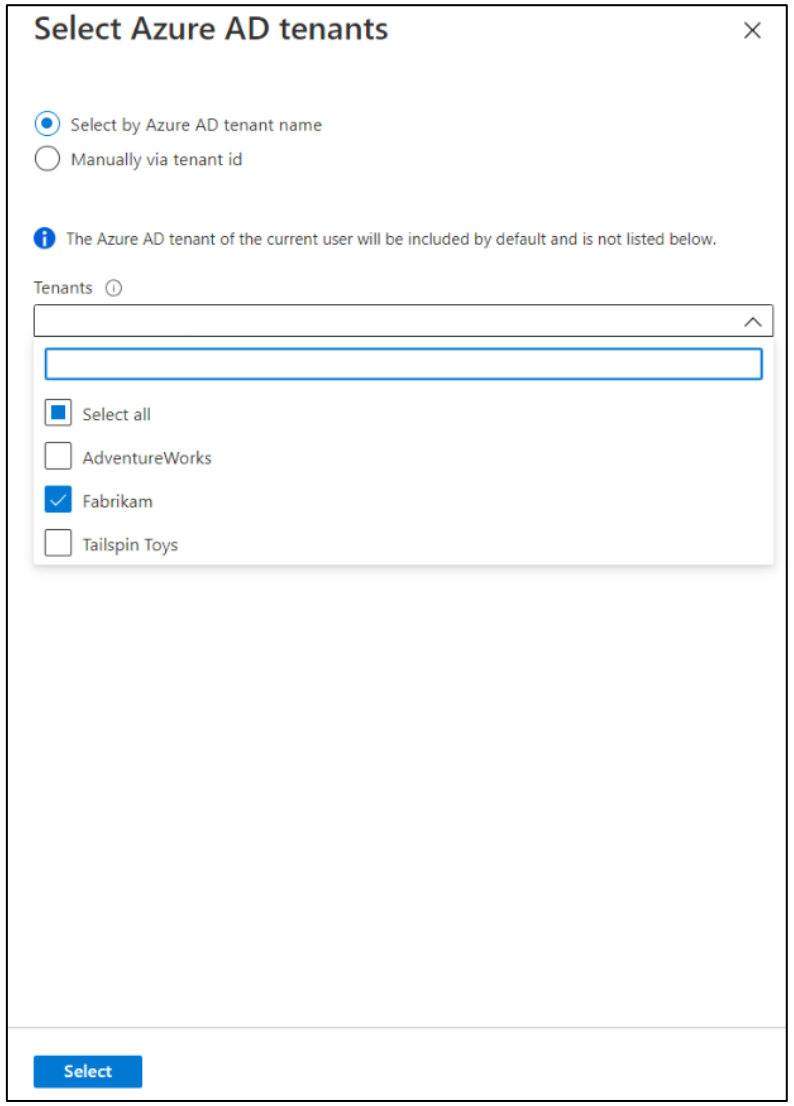
Azure Synapse creates two managed private endpoints to dedicated SQL pool and serverless SQL pool in that workspace. Customers do not get charged for these two Managed private endpoints



Data Exfiltration

Allow outbound data traffic over Synapse managed private endpoints to only approved tenants





Azure Active Directory authentication

Overview

Manage user identities in one location.

Enable access to Azure Synapse Analytics and other Microsoft services with Azure Active Directory user identities and groups.

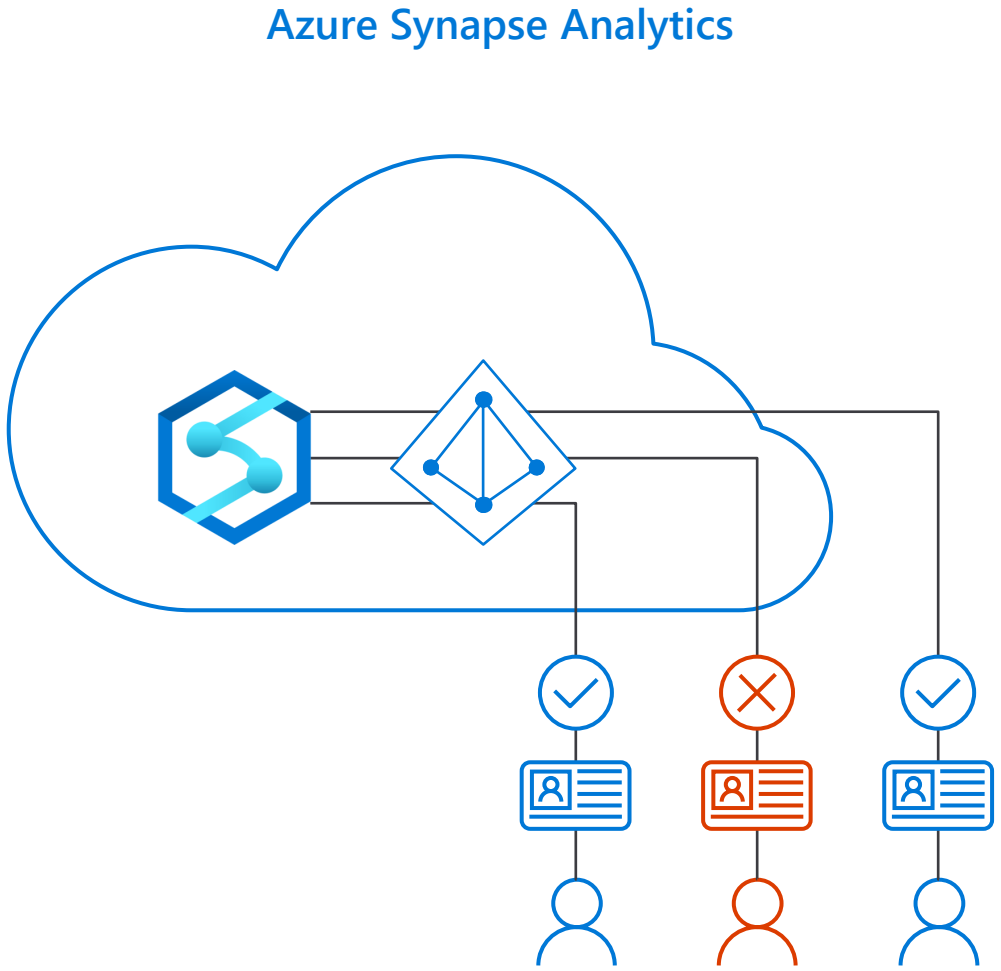
Benefits

Enables management of database permissions by using external Azure Active Directory groups

Allows password rotation in a single place

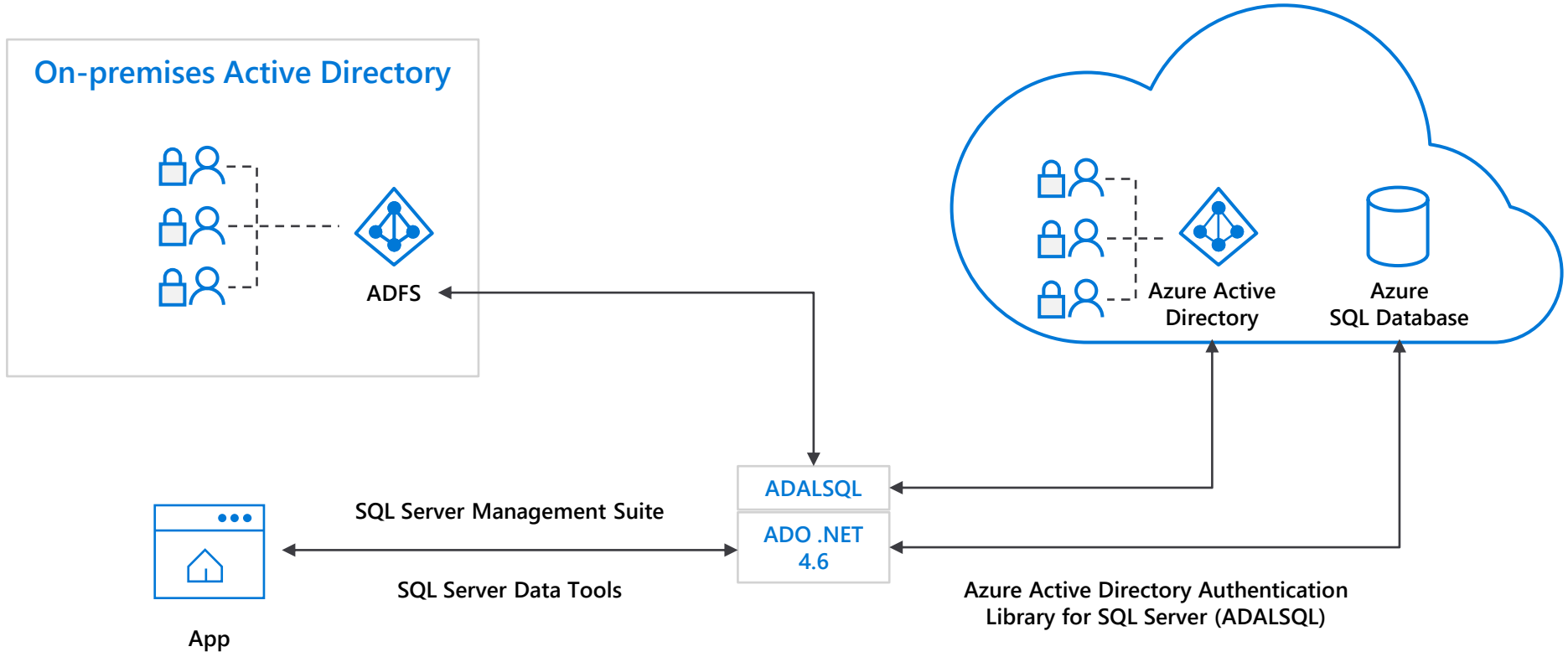
Alternative to SQL Server authentication

Eliminates the need to store passwords

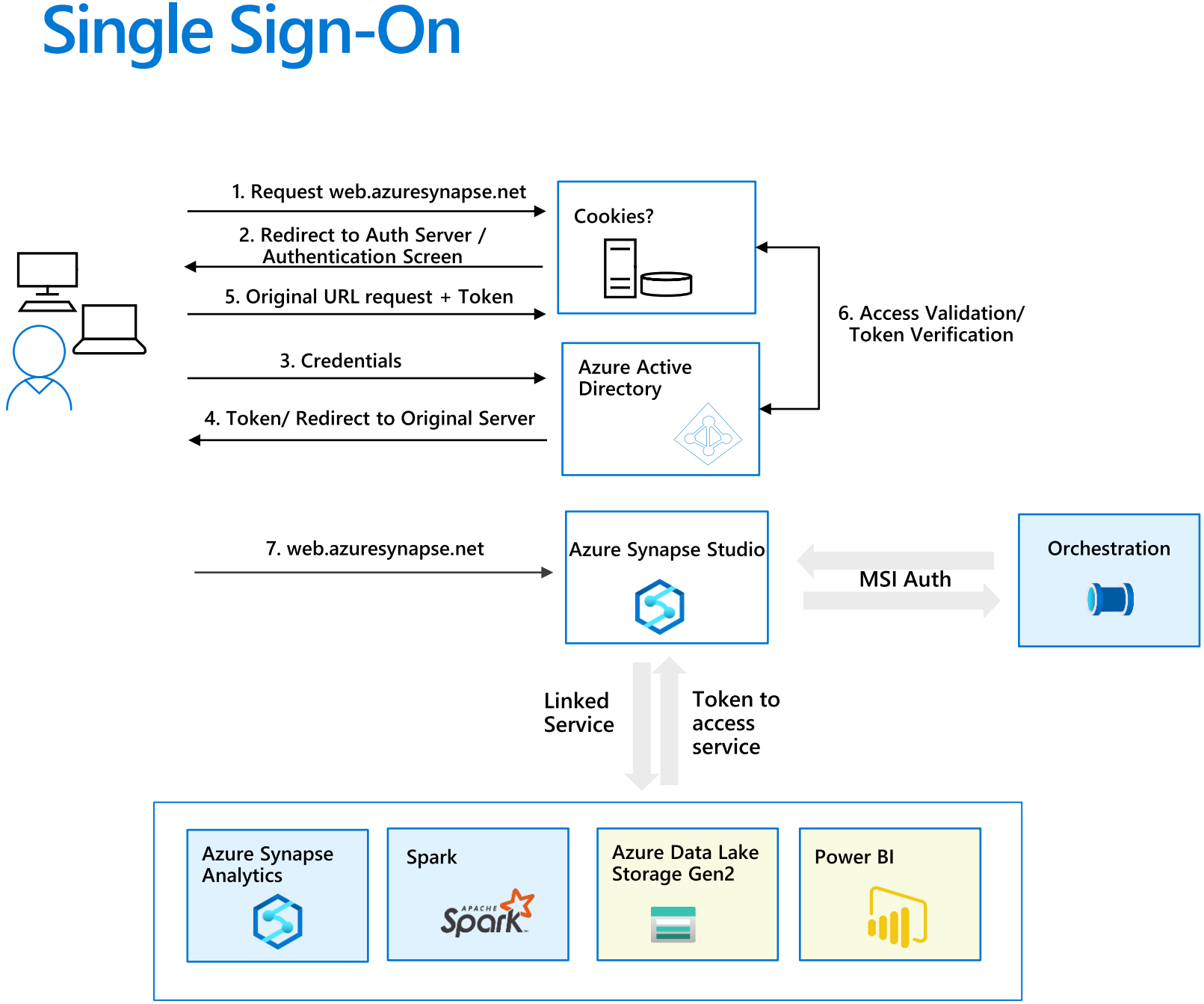


Azure Active Directory trust architecture

Azure Active Directory and Azure Synapse Analytics



Single Sign-On



Implicit authentication - User provides login credentials once to access Azure Synapse Workspace

AAD authentication - Azure Synapse Studio will request token to access each linked services as user. A separate token is acquired for each of the below services:

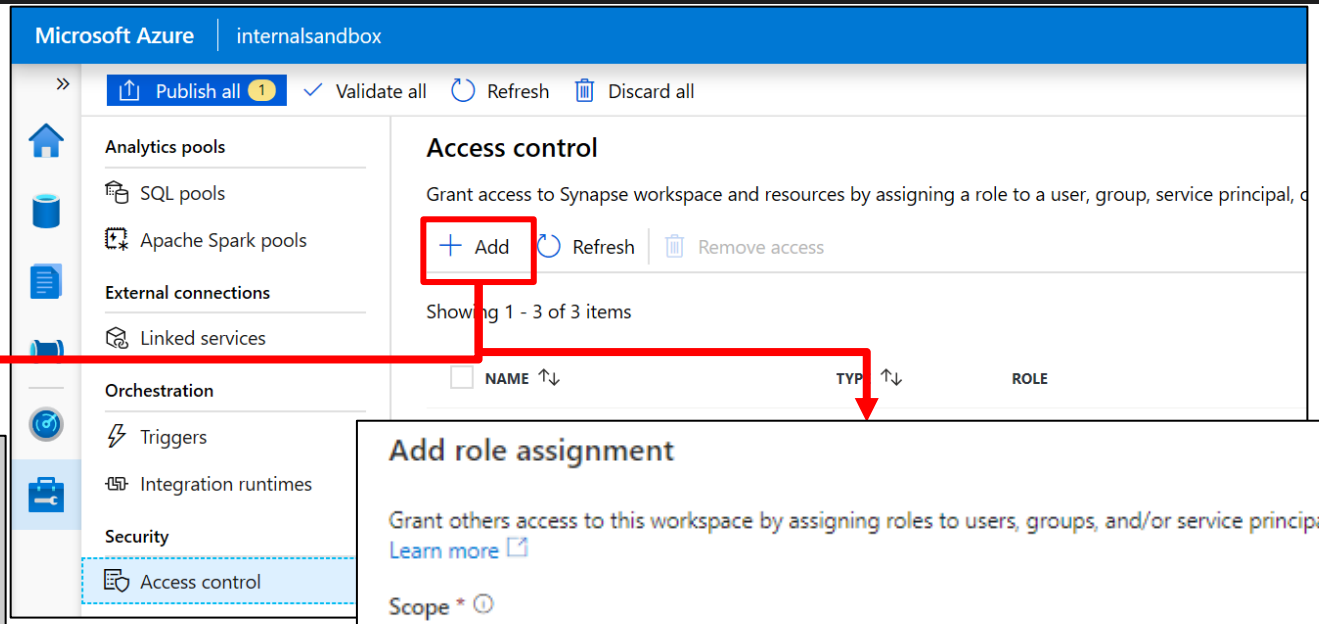
1. ADLS Gen2
2. Azure Synapse Analytics
3. Power BI
4. Spark – Spark Livy API
5. management.azure.com – resource provisioning
6. Develop artifacts – dev.workspace.net
7. Graph endpoints

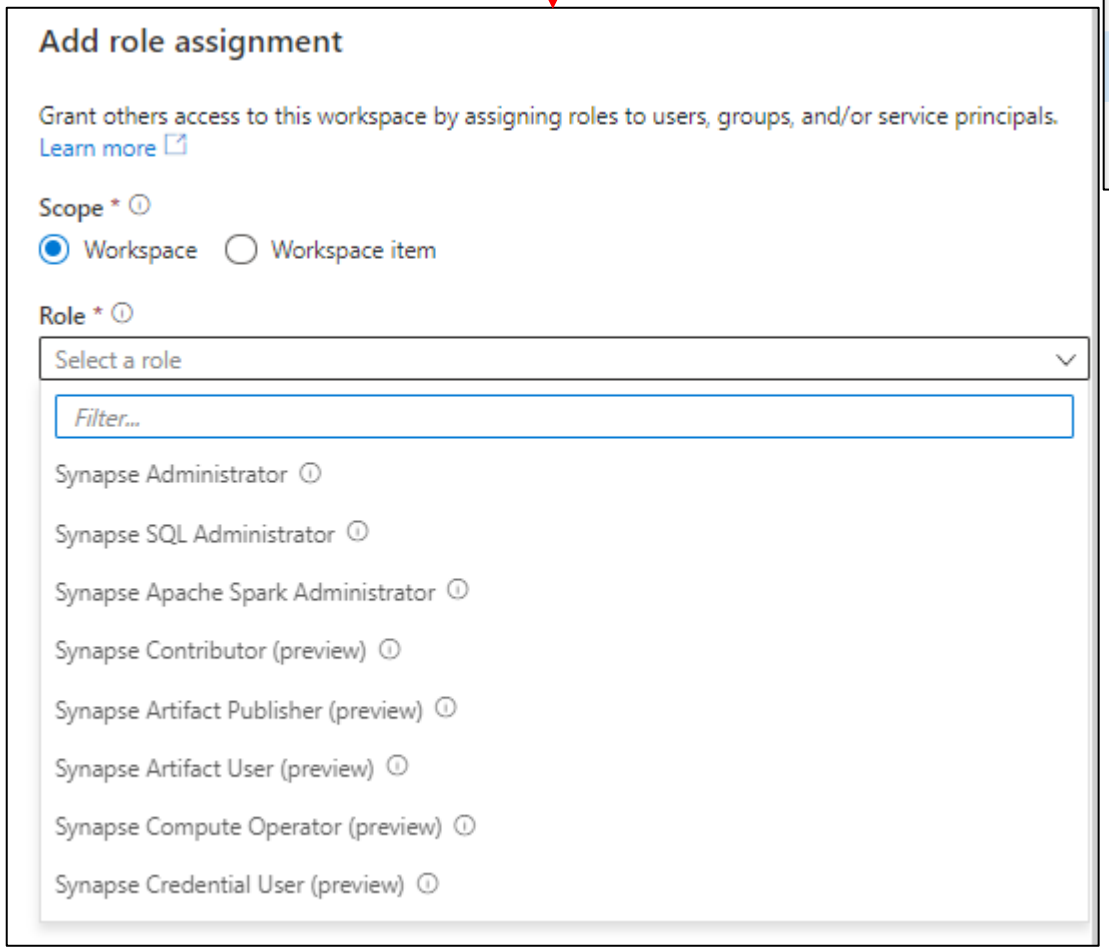
MSI authentication - Orchestration uses workspace MSI auth for automation

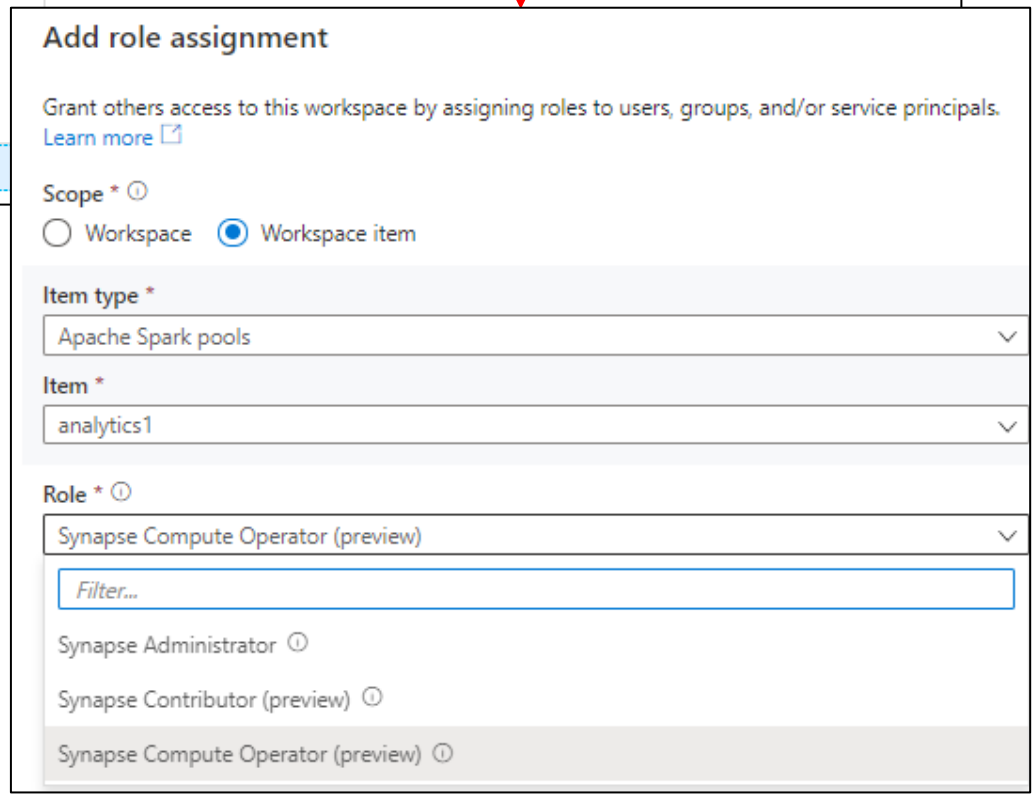
Access Control

Overview

It provides access control management to workspace resources and artifacts







Synapse RBAC Roles and permitted actions

Action	Synapse Administrator	Synapse Contributor	Synapse Artifact Author	Synapse Artifact Reader	Synapse Compute Manager	Synapse Credential User	Synapse Managed Private Endpoint Administrator	Synapse Reader
workspaces/read	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
workspaces/roleAssignments/write, delete	Yes							
workspaces/managedPrivateEndpoint/write, delete	Yes						Yes	
workspaces/bigDataPools/useCompute/action	Yes	Yes			Yes			
workspaces/bigDataPools/viewLogs/action	Yes	Yes			Yes			
workspaces/integrationRuntimes/useCompute/action	Yes	Yes			Yes			
workspaces/integrationRuntimes/viewLogs/action	Yes	Yes			Yes			
workspaces/artifacts/read	Yes	Yes	Yes	Yes				
workspaces/notebooks/write, delete	Yes	Yes	Yes					
workspaces/sparkJobDefinitions/write, delete	Yes	Yes	Yes					
workspaces/sqlScripts/write, delete	Yes	Yes	Yes					
workspaces/dataFlows/write, delete	Yes	Yes	Yes					
workspaces/pipelines/write, delete	Yes	Yes	Yes					
workspaces/triggers/write, delete	Yes	Yes	Yes					
workspaces/datasets/write, delete	Yes	Yes	Yes					
workspaces/libraries/write, delete	Yes	Yes	Yes					
workspaces/linkedServices/write, delete	Yes	Yes	Yes					
workspaces/credentials/write, delete	Yes	Yes	Yes					
workspaces/notebooks/viewOutputs/action	Yes	Yes						
workspaces/pipelines/viewOutputs/action	Yes	Yes						
workspaces/linkedServices/useSecret/action	Yes					Yes		
workspaces/credentials/useSecret/action	Yes					Yes		

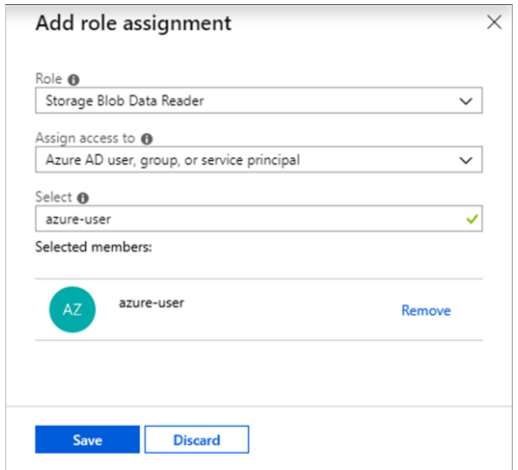
Access control – serverless SQL pool

Overview

Enterprise-grade security model enables you to control who can access data.

Benefits

- Use Azure Active Directory users or native SQL logins.
- SAS tokens, AAD or workspace identity access
- Specify access methods in credential
- Grant access to storage by referencing storage credential
- Enable some logins to access external tables
- Add AAD role assignments directly on Azure storage.




```
--Create Login to a a single serverless SQL pool database
CREATE LOGIN [alias@domain.com] FROM EXTERNAL PROVIDER;

-- create user under that login
use yourdb -- Use your DB name
go
CREATE USER alias FROM LOGIN [alias@domain.com];

To grant full access to a user to all serverless SQL pool databases
CREATE LOGIN [alias@domain.com] FROM EXTERNAL PROVIDER;
ALTER SERVER ROLE  sysadmin  ADD MEMBER [alias@domain.com];

-- enable impersonation using workspace Managed Identity
CREATE CREDENTIAL [ManagedIdentity]
WITH IDENTITY  = 'Managed Identity'

-- enable access to specified storage using SAS token
CREATE CREDENTIAL [https://XXX.blob.core.windows.net/csv]
WITH IDENTITY  = 'SHARED ACCESS SIGNATURE',
        SECRET  = 'sv=2014-02-
14&sr=b&si=TestPolicy&sig=o%2B5%2F0C%2BLm7tWWft'

-- grant login1 to use SAS token defined in credential for storage account
GRANT REFERENCES CREDENTIAL::[https://XXX.blob.core.windows.net/csv]
TO LOGIN  = 'login1'

-- grant login2 to use Managed Identity
GRANT REFERENCES CREDENTIAL::[ManagedIdentity]
TO LOGIN  = 'login2'


-- grant login2 to select external data via table
GRANT SELECT ON OBJECT::[dbo.population]  TO LOGIN  = 'login2'
```

Industry-leading security and compliance

Enterprise-grade security

Industry-leading compliance

ISO 27001

SOC 1 Type 2

SOC 2 Type 2

PCI DSS Level 1

Cloud Controls Matrix

ISO 27018

Content Delivery and Security Association

Shared Assessments

FedRAMP JAB P-ATO

HIPAA / HITECH

FIPS 140-2

21 CFR Part 11

FERPA

DISA Level 2

CJIS

IRS 1075

ITAR-ready

Section 508 VPAT

European Union Model Clauses

EU Safe Harbor

United Kingdom G-Cloud

China Multi Layer Protection Scheme

China GB 18030

China CCCPPF

Singapore MTCS Level 3

Australian Signals Directorate

New Zealand GCIO

Japan Financial Services

ENISA IAF

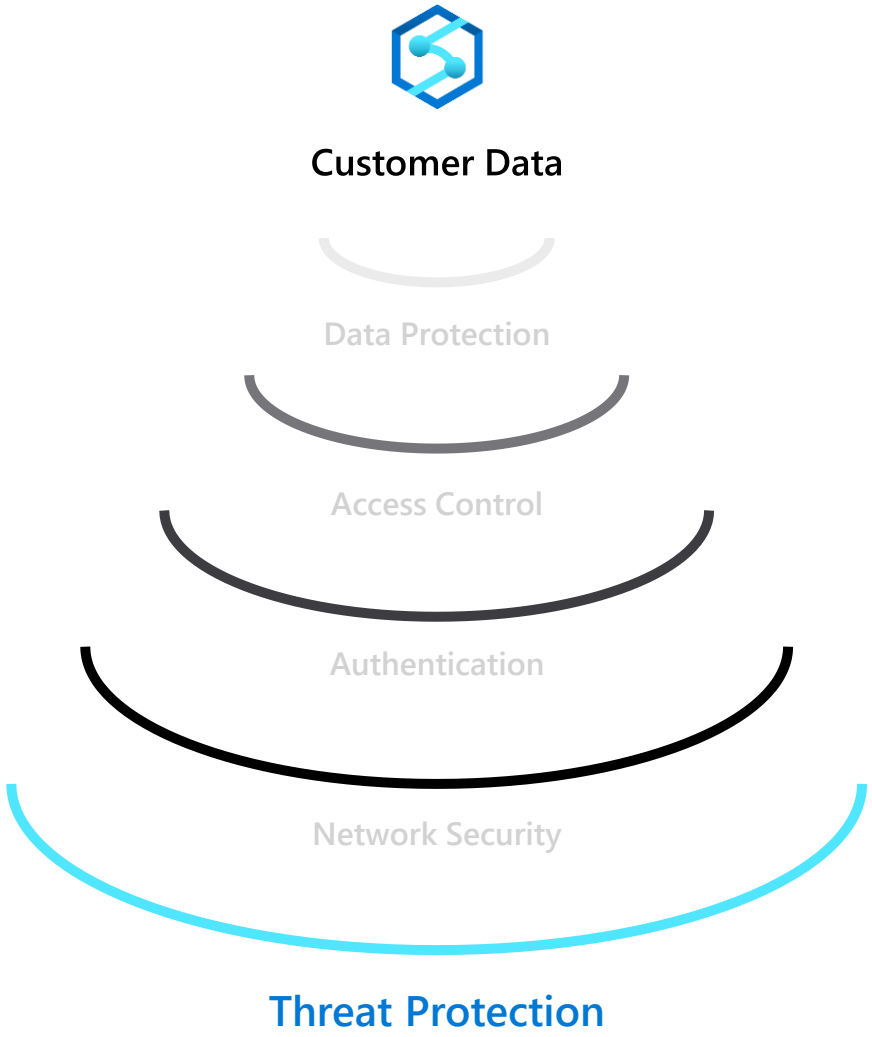
Threat Protection - Business requirements

How do we enumerate and track potential SQL vulnerabilities?

To mitigate any security misconfigurations before they become a serious issue.

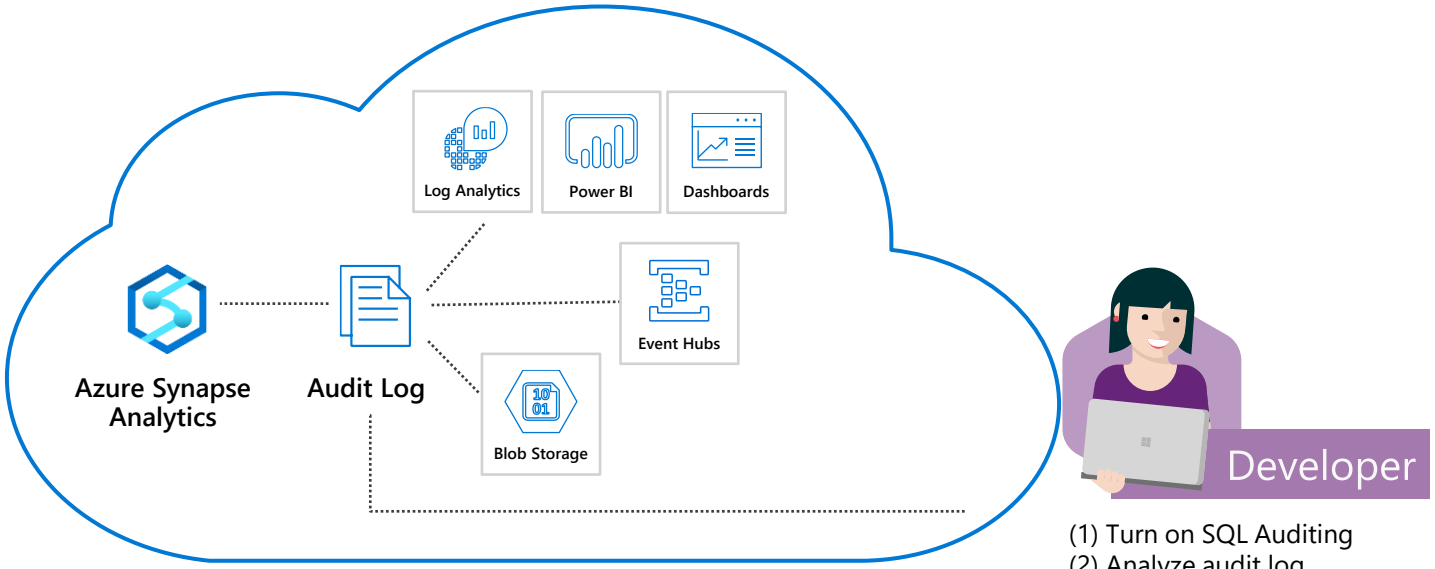
How do we discover and alert on suspicious database activity?

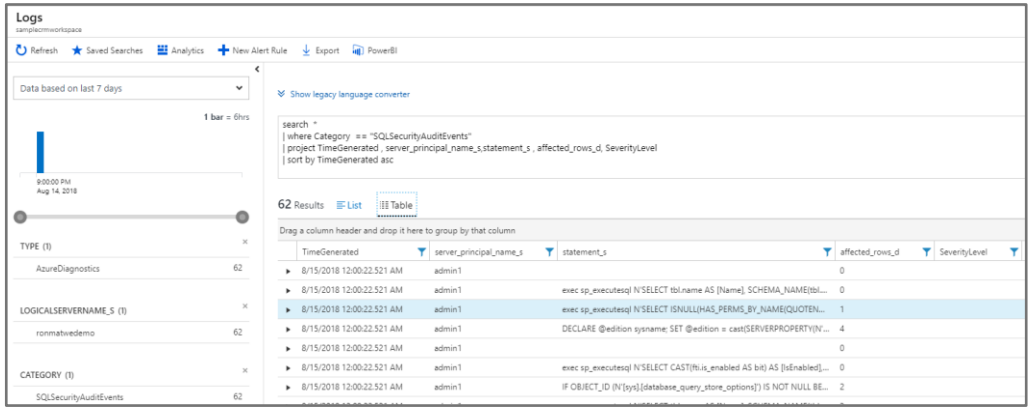
To detect and resolve any data exfiltration or SQL injection attacks.



SQL auditing in Azure Log Analytics and Event Hubs

Gain insight into database audit log

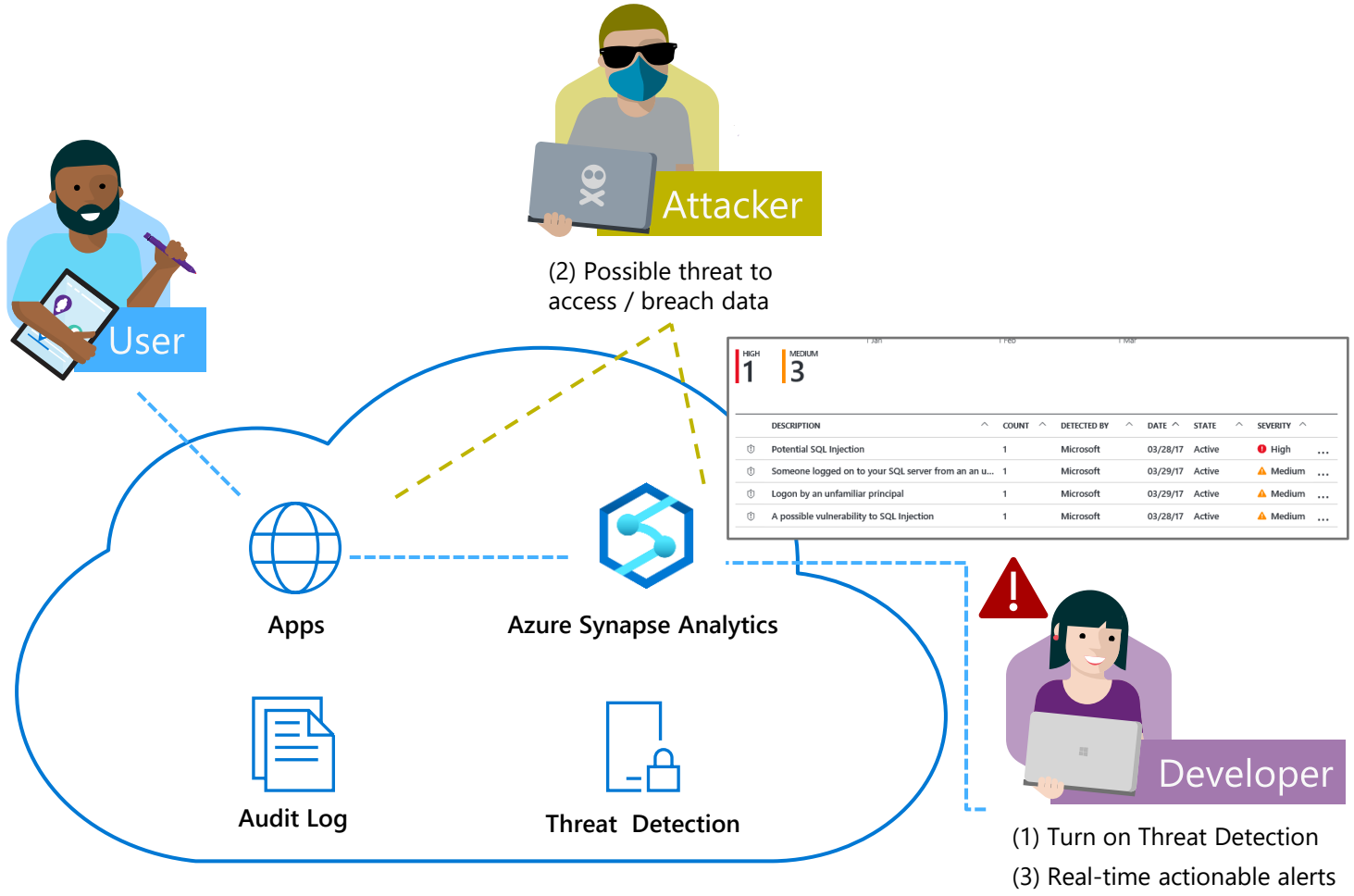




- ✓ **Configurable via audit policy**
- ✓ **SQL audit logs can reside in**
  - Azure Storage account
  - Azure Log Analytics
  - Azure Event Hubs
- ✓ **Rich set of tools for**
  - Investigating security alerts
  - Tracking access to sensitive data

SQL threat detection

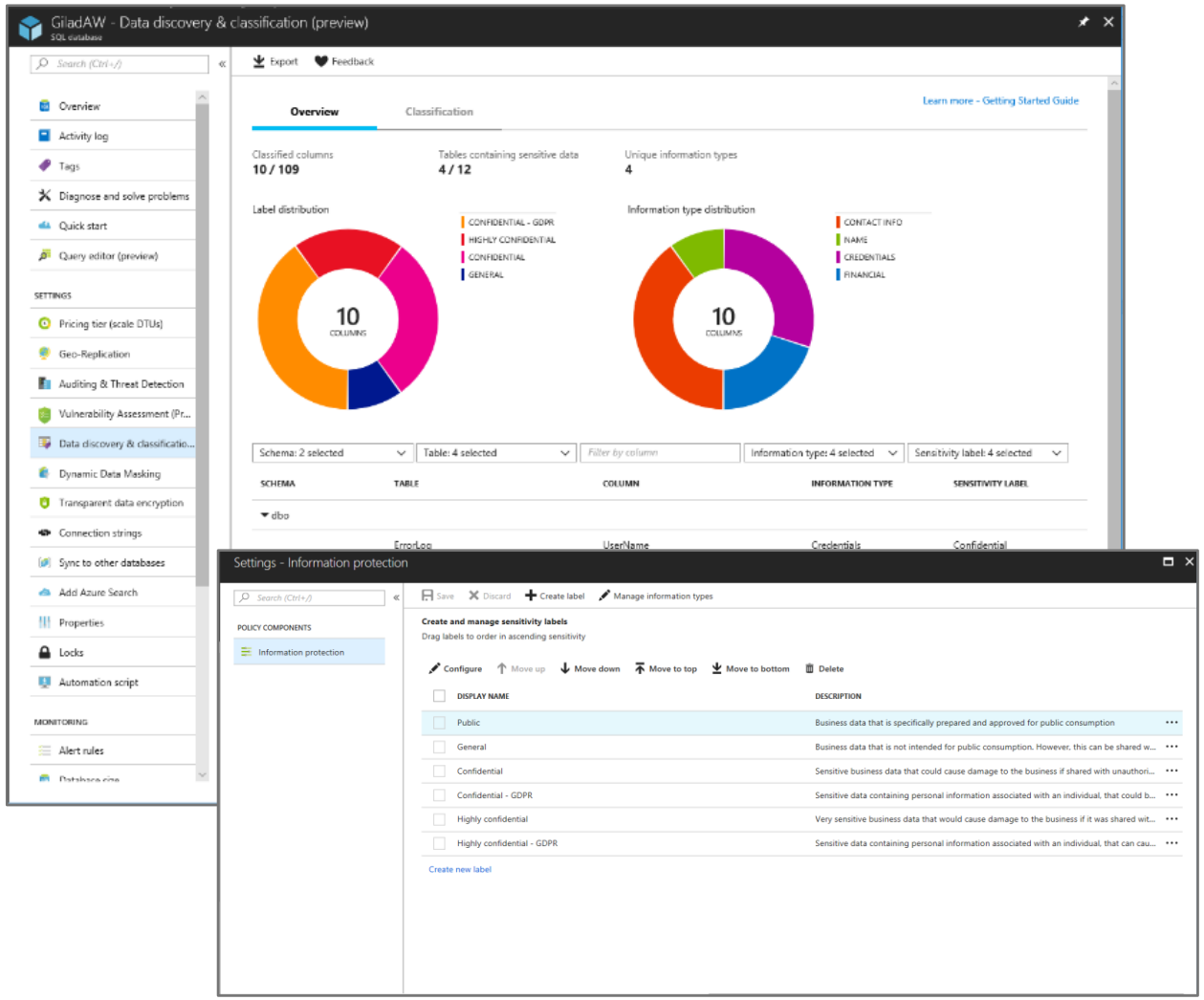
Detect and investigate anomalous database activity



- ✓ Detects potential SQL injection attacks
- ✓ Detects unusual access & data exfiltration activities
- ✓ Actionable alerts to investigate & remediate
- ✓ View alerts for your entire Azure tenant using Azure Security Center

SQL Data Discovery & Classification

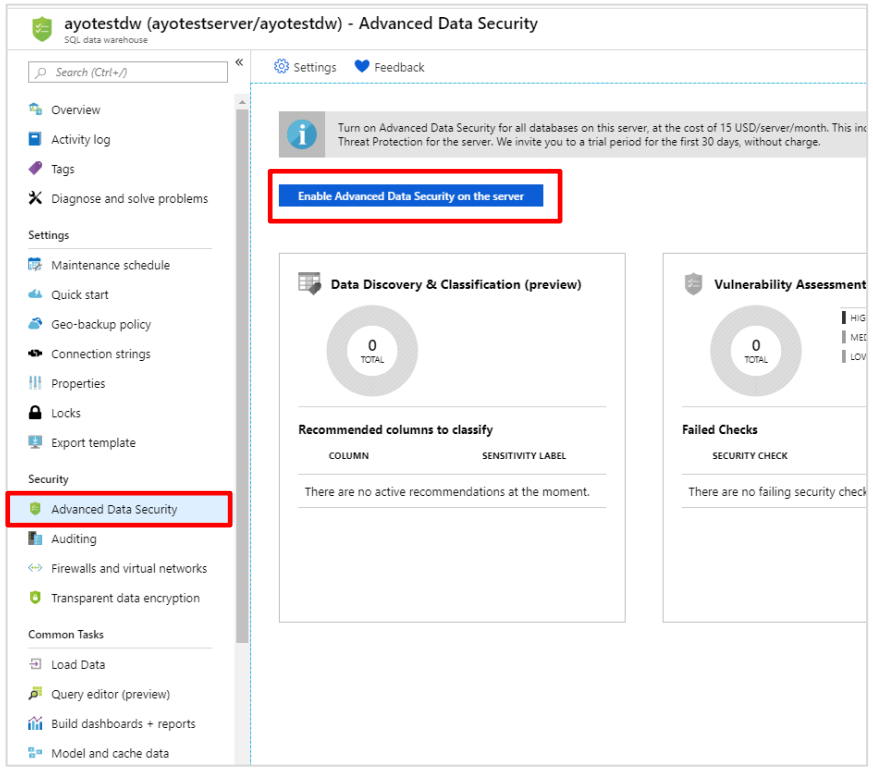
Discover, classify, protect and track access to sensitive data



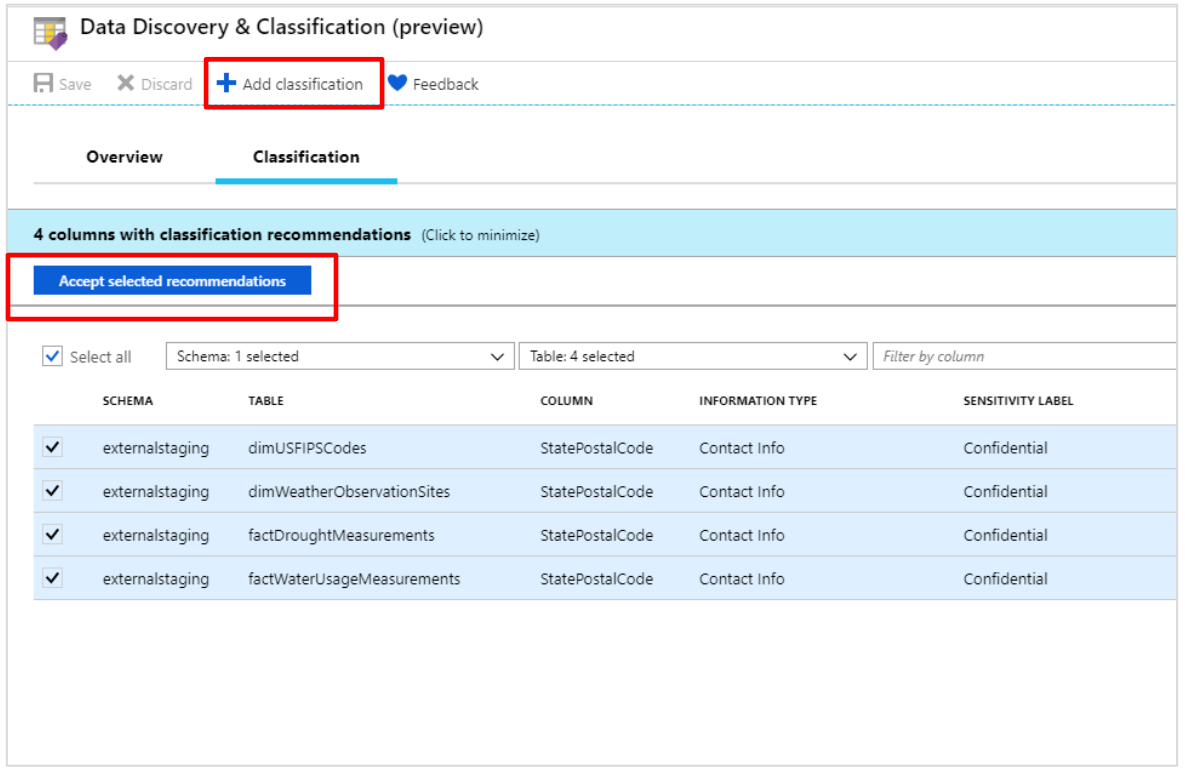
- ✓ Automatic discovery of columns with sensitive data
- ✓ Add persistent sensitive data labels
- ✓ Audit and detect access to the sensitive data
- ✓ Manage labels for your entire Azure tenant using Azure Security Center

SQL Data Discovery & Classification - setup

Step 1: Enable Advanced Data Security on the logical SQL Server

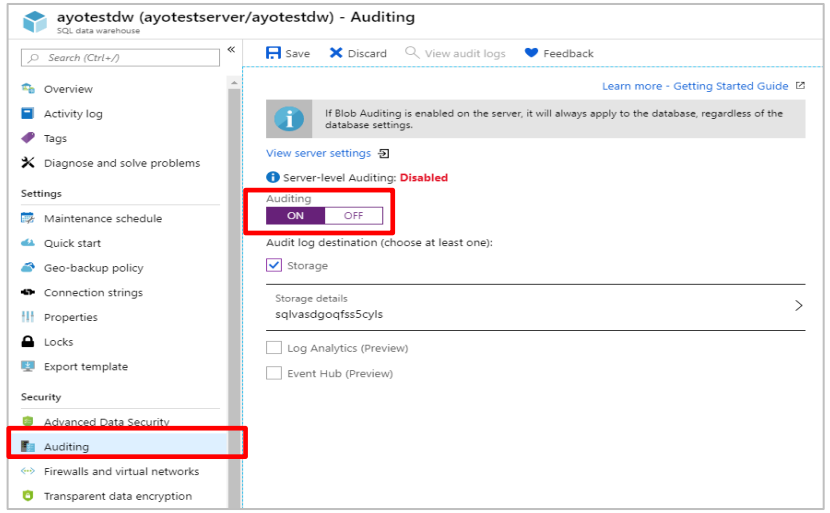


Step 2: Use recommendations and/or manual classification to classify all the sensitive columns in your tables

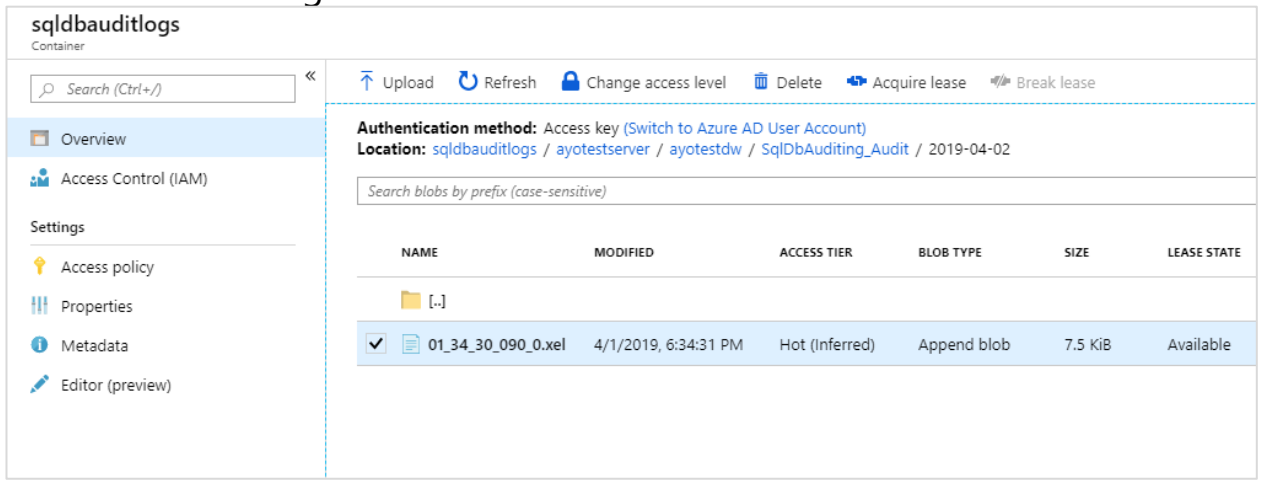


SQL Data Discovery & Classification – audit sensitive data access

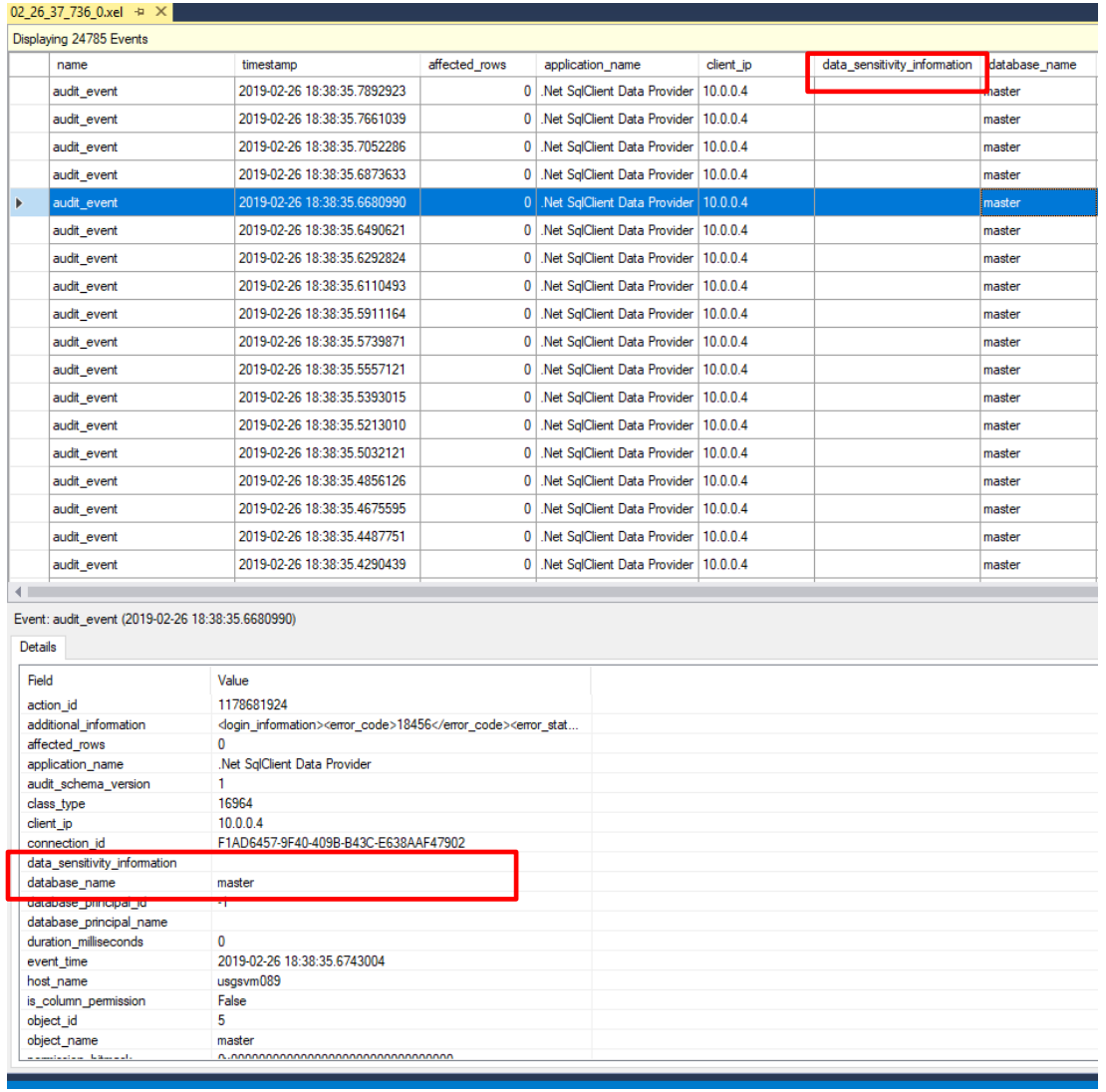
Step 1: Configure auditing for your target Data warehouse. This can be configured for just a single data warehouse or all databases on a server.



Step 2: Navigate to audit logs in storage account and download 'xel' log files to local machine.



Step 3: Open logs using extended events viewer in SSMS. Configure viewer to include 'data_sensitivity_information' column



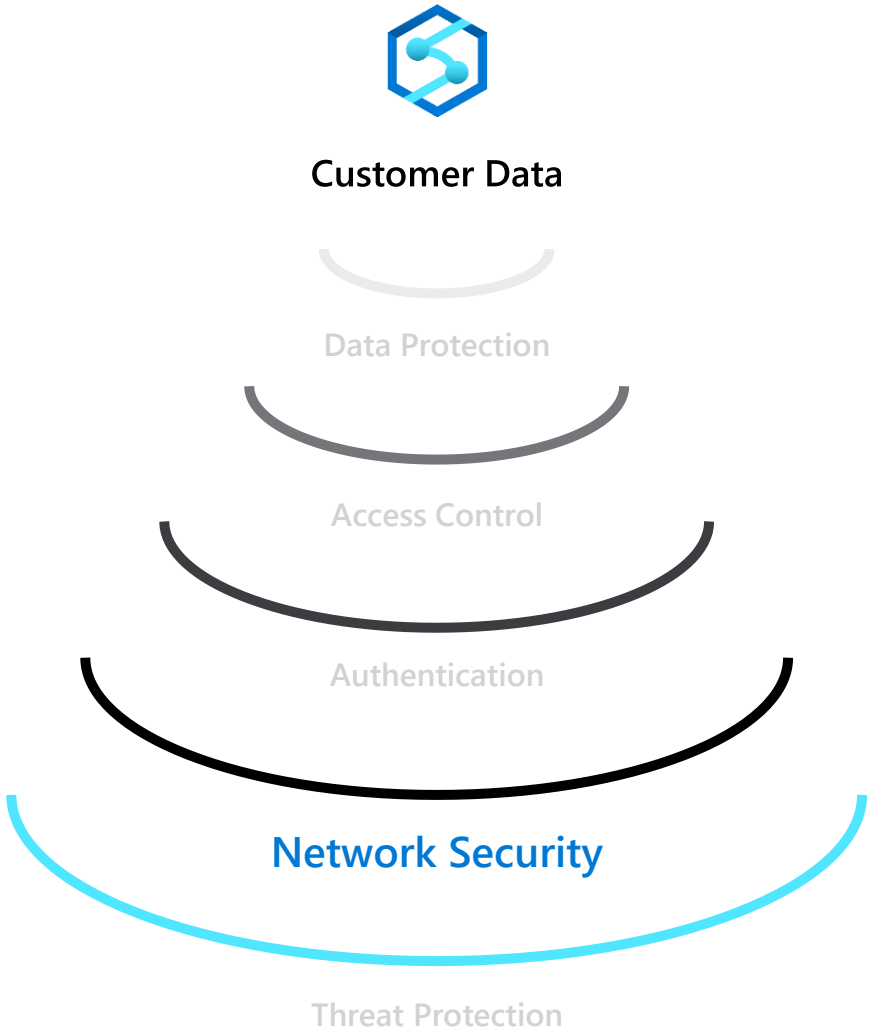
Network Security - Business requirements

How do we implement network isolation?

Data at different levels of security needs to be accessed from different locations.

How do we achieve separation?

Disallowing access to entities outside the company's network security boundary.

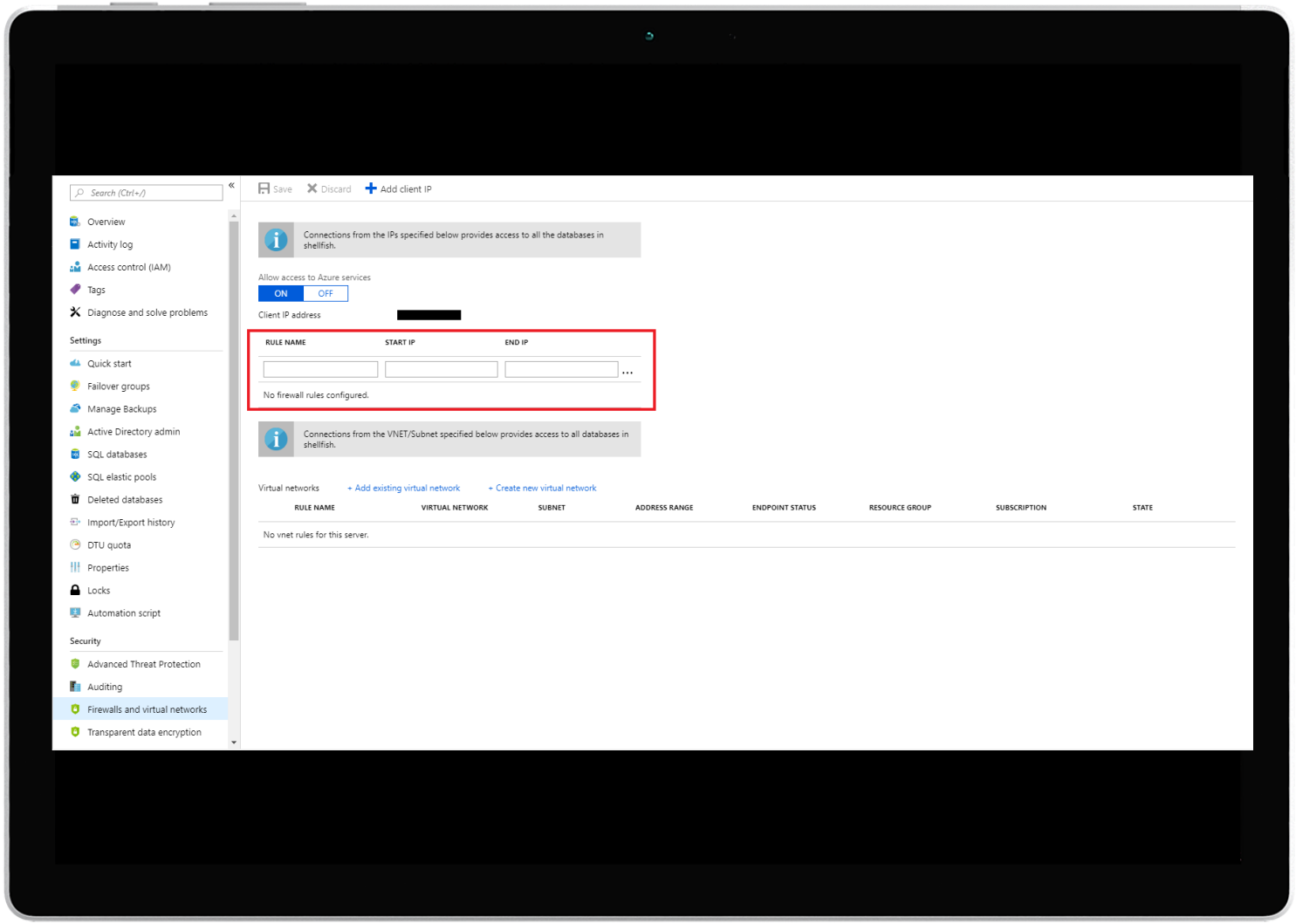


Firewall configuration on the portal

By default, Azure blocks all external connections to port 1433

Configure with the following steps:

Azure Synapse Analytics Resource:
Server name > Firewalls and virtual networks



Firewall configuration using REST API

Managing firewall rules through REST API must be authenticated.

For information, see Authenticating Service Management Requests.

Server-level rules can be created, updated, or deleted using REST API.

To create or update a server-level firewall rule, execute the PUT method.

To remove an existing server-level firewall rule, execute the DELETE method.

To list firewall rules, execute the GET.

```
PUT
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/firewallRules/{firewallRuleName}?api-version=2014-04-01REQUEST BODY
{
  "properties": {
    "startIpAddress": "0.0.0.3",
    "endIpAddress": "0.0.0.3"
  }
}

DELETE
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/firewallRules/{firewallRuleName}?api-version=2014-04-01

GET
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.Sql/servers/{serverName}/firewallRules/{firewallRuleName}?api-version=2014-04-01
```

Firewall configuration using PowerShell/T-SQL

Windows PowerShell Azure cmdlets

New-AzureRmSqlServerFirewallRule

Get-AzureRmSqlServerFirewallRule

Set-AzureRmSqlServerFirewallRule

Transact SQL

sp_set_firewall_rule

sp_delete_firewall_rule

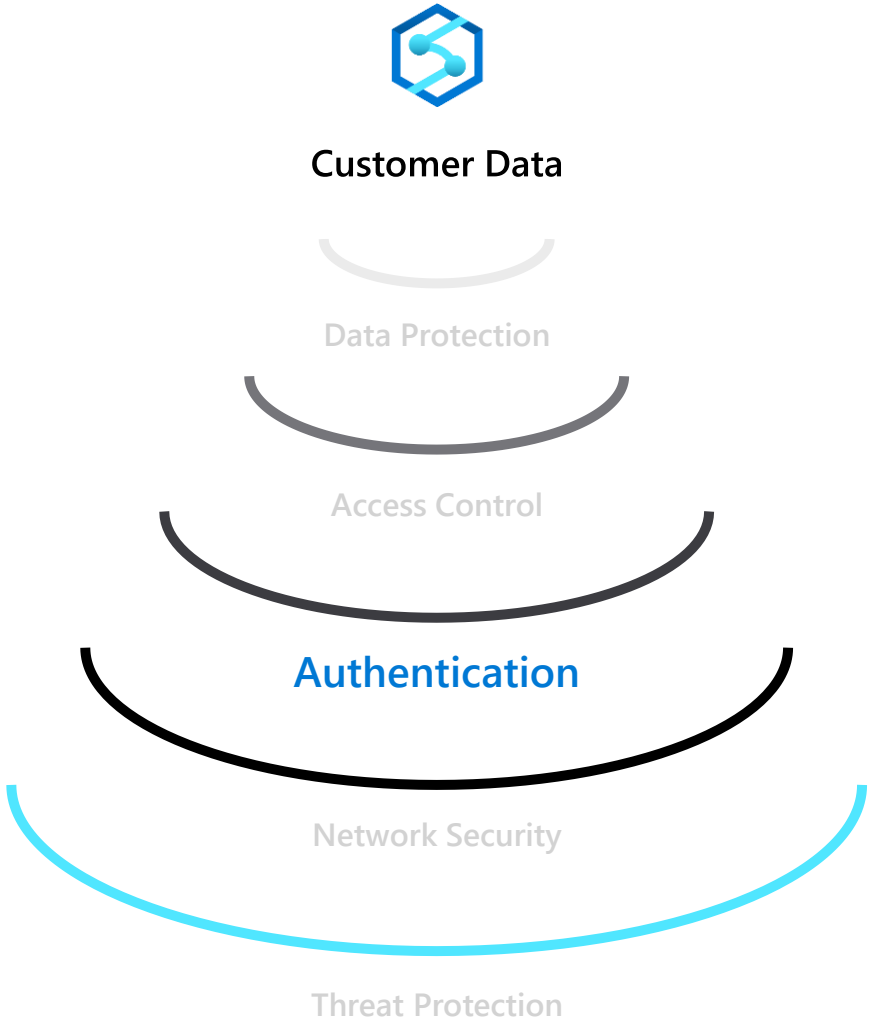
```
# PS Allow external IP access to SQL DW
PS C:\> New-AzureRmSqlServerFirewallRule
          -ResourceGroupName "myResourceGroup" `
          -ServerName $servername `
          -FirewallRuleName "AllowSome"
          -StartIpAddress "0.0.0.0"
          -EndIpAddress "0.0.0.0“

-- T-SQL Allow external IP access to SQL DW
EXECUTE sp_set_firewall_rule
          @name = N'ContosoFirewallRule’,
          @start_ip_address = '192.168.1.1’,
          @end_ip_address = '192.168.1.10'
```

Authentication - Business requirements

How do I configure Azure Active Directory with Azure Synapse Analytics?

I want additional control in the form of multi-factor authentication

How do I allow non-Microsoft accounts to be able to authenticate?

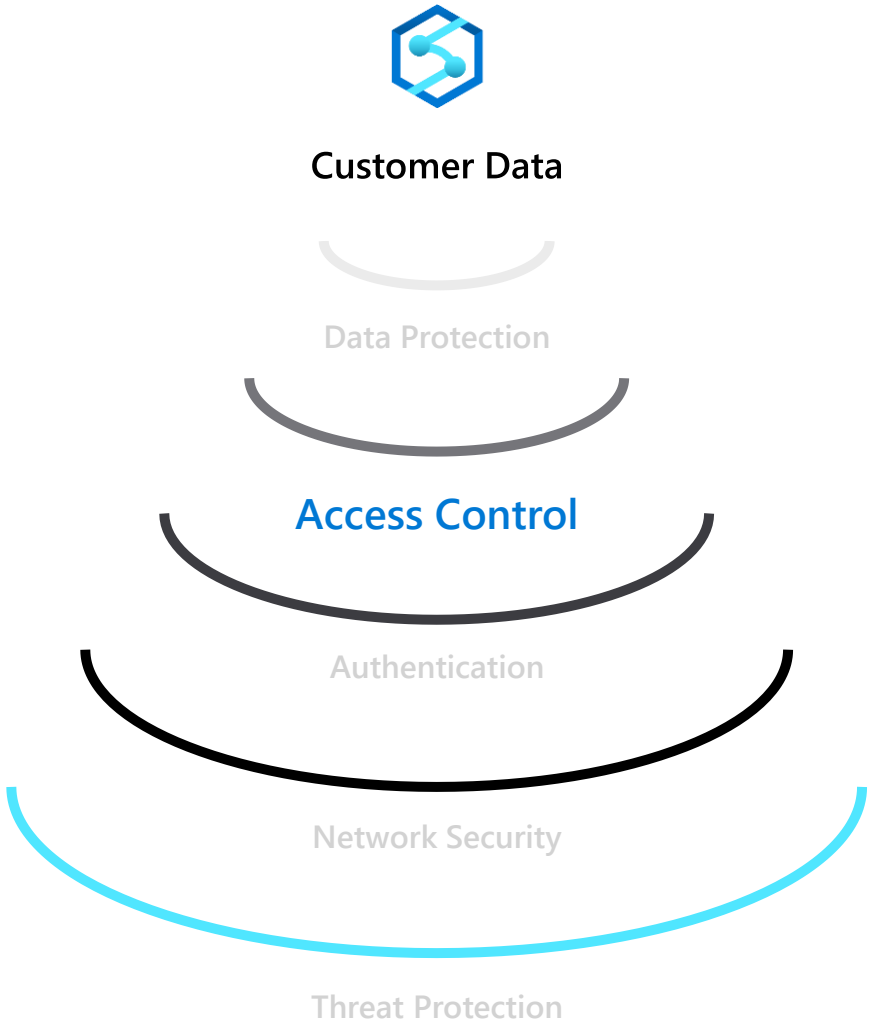


Access Control - Business requirements

How do I restrict access to sensitive data to specific database users?

How do I ensure users only have access to relevant data?

For example, in a hospital only medical staff should be allowed to see patient data that is relevant to them—and not every patient's data.



Object-level security (tables, views, and more)

Overview

GRANT controls permissions on designated tables, views, stored procedures, and functions.

Prevent unauthorized queries against certain tables.

Simplifies design and implementation of security at the database level as opposed to application level.

```
-- Grant SELECT permission to user RosaQdM on table Person.Address in the AdventureWorks2012 database
GRANT SELECT ON OBJECT::Person.Address TO RosaQdM;
GO
-- Grant REFERENCES permission on column BusinessEntityID in view HumanResources.vEmployee to user Wanida
GRANT REFERENCES(BusinessEntityID) ON OBJECT::HumanResources.vEmployee to Wanida with GRANT OPTION;
GO
-- Grant EXECUTE permission on stored procedure HumanResources.uspUpdateEmployeeHireInfo to an application role called Recruiting11
USE AdventureWorks2012;
GRANT EXECUTE ON OBJECT::HumanResources.uspUpdateEmployeeHireInfo TO RECRUITING 11;
GO
```

Row-level security (RLS)

Overview

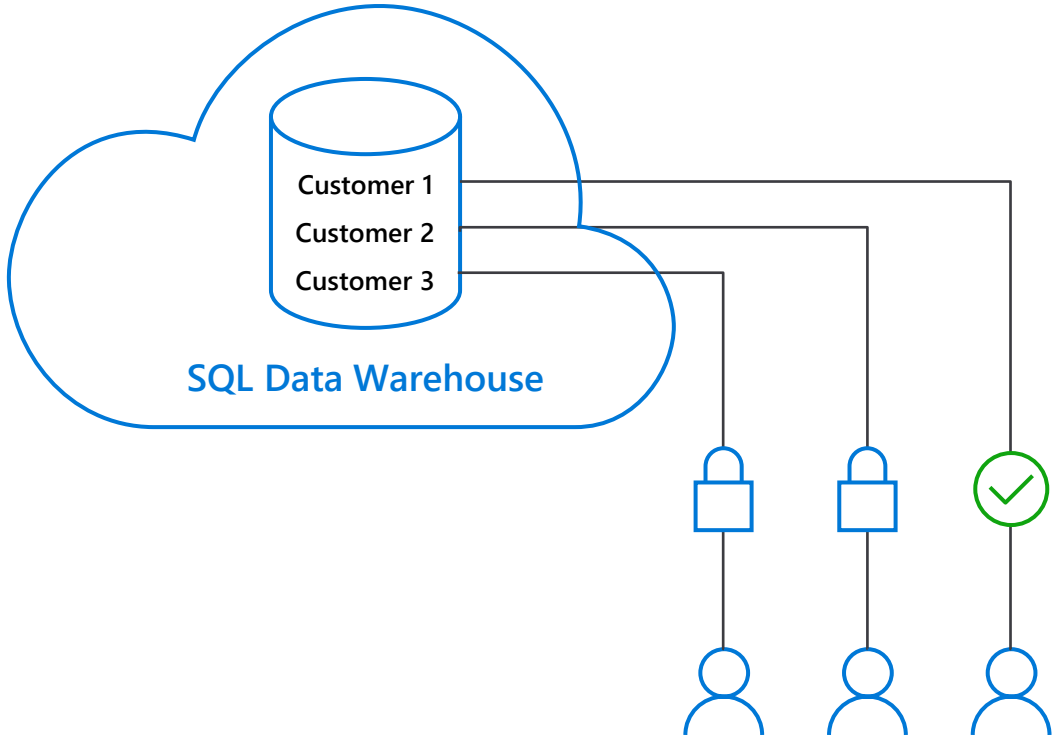
Fine grained access control of specific rows in a database table.

Help prevent unauthorized access when multiple users share the same tables.

Eliminates need to implement connection filtering in multi-tenant applications.

Administer via SQL Server Management Studio or SQL Server Data Tools.

Easily locate enforcement logic inside the database and schema bound to the table.



Row-level security

Creating policies

Filter predicates silently filter the rows available to read operations (SELECT, UPDATE, and DELETE).

The following examples demonstrate the use of the CREATE SECURITY POLICY syntax

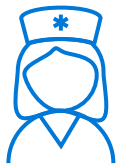
```
-- The following syntax creates a security policy with a filter predicate for the Customer table
CREATE SECURITY POLICY [FederatedSecurityPolicy]
ADD FILTER PREDICATE [rls].[fn_securitypredicate]([CustomerId])
ON [dbo].[Customer];


-- Create a new schema and predicate function, which will use the application user ID stored in CONTEXT_INFO to filter rows.
CREATE FUNCTION rls.fn_securitypredicate (@AppUserId int)
RETURNS TABLE
WITH SCHEMABINDING
AS
RETURN (
SELECT 1 AS fn_securitypredicate_result
WHERE
DATABASE_PRINCIPAL_ID() = DATABASE_PRINCIPAL_ID('dbo') -- application context
AND CONTEXT_INFO() = CONVERT(VARBINARY(128), @AppUserId));
GO
```

Row-level security

Three steps:

1. Policy manager creates filter predicate and security policy in T-SQL, binding the predicate to the patients table.
2. App user (e.g., nurse) selects from Patients table.
3. Security policy transparently rewrites query to apply filter predicate.

Policy manager



Nurse

Application

```
SELECT * FROM Patients
```

Database

Security policy

Patients

Filter Predicate: INNER JOIN...

```
CREATE FUNCTION dbo.fn_securitypredicate(@wing int)
    RETURNS TABLE WITH SCHEMABINDING AS
    return SELECT 1 as [fn_securitypredicate_result] FROM
        StaffDuties d INNER JOIN Employees e
        ON (d.EmpId = e.EmpId)
        WHERE e.UserSID = SUSER_SID() AND @wing = d.Wing;

CREATE SECURITY POLICY dbo.SecPol
    ADD FILTER PREDICATE dbo.fn_securitypredicate(Wing) ON Patients
    WITH (STATE = ON)
```

```
SELECT * FROM Patients
   SEMIJOIN APPLY dbo.fn_securitypredicate(patients.Wing);
```

```
SELECT Patients.* FROM Patients,
   StaffDuties d INNER JOIN Employees e ON (d.EmpId = e.EmpId)
   WHERE e.UserSID = SUSER_SID() AND Patients.wing = d.Wing;
```

Column-level security

Overview

Control access of specific columns in a database table based on customer's group membership or execution context.

Simplifies the design and implementation of security by putting restriction logic in database tier as opposed to application tier.

Administer via GRANT T-SQL statement.

Both Azure Active Directory (AAD) and SQL authentication are supported.

Column-level security

Three steps:

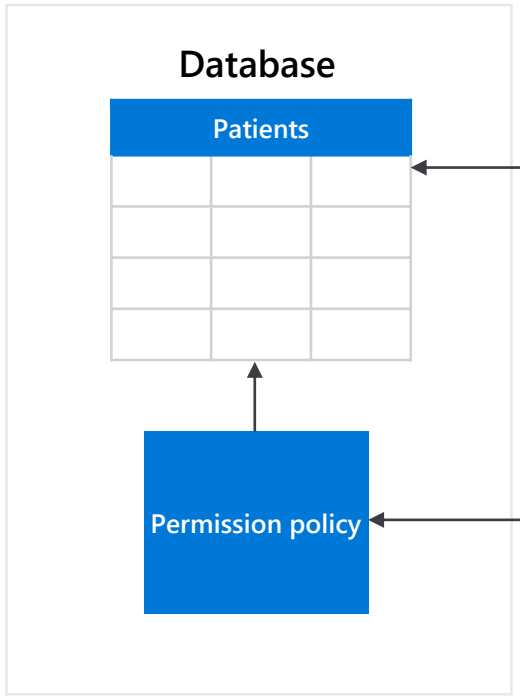
1. Policy manager creates permission policy in T-SQL, binding the policy to the Patients table on a specific group.
2. App user (for example, a nurse) selects from Patients table.
3. Permission policy prevents access on sensitive data.

Nurse

Application

```
SELECT * FROM Membership;

Msg 230, Level 14, State 1, Line 12
The SELECT permission was denied on the column
'SSN' of the object 'Membership', database
'CLS_TestDW', schema 'dbo'.
```

Queries executed as 'Nurse' will fail if they include the SSN column



Policy manager

```
CREATE TABLE Patients (
  PatientID int IDENTITY,
  FirstName varchar(100) NULL,
  SSN char(9) NOT NULL,
  LastName varchar(100) NOT NULL,
  Phone varchar(12) NULL,
  Email varchar(100) NULL
);
```

```
GRANT SELECT ON Patients (
  PatientID, FirstName, LastName, Phone, Email
) TO Nurse;
```

Allow 'Nurse' to access all columns except for sensitive SSN column

Data Protection - Business requirements

How do I protect sensitive data against unauthorized (high-privileged) users?

What key management options do I have?

Dynamic Data Masking

Overview

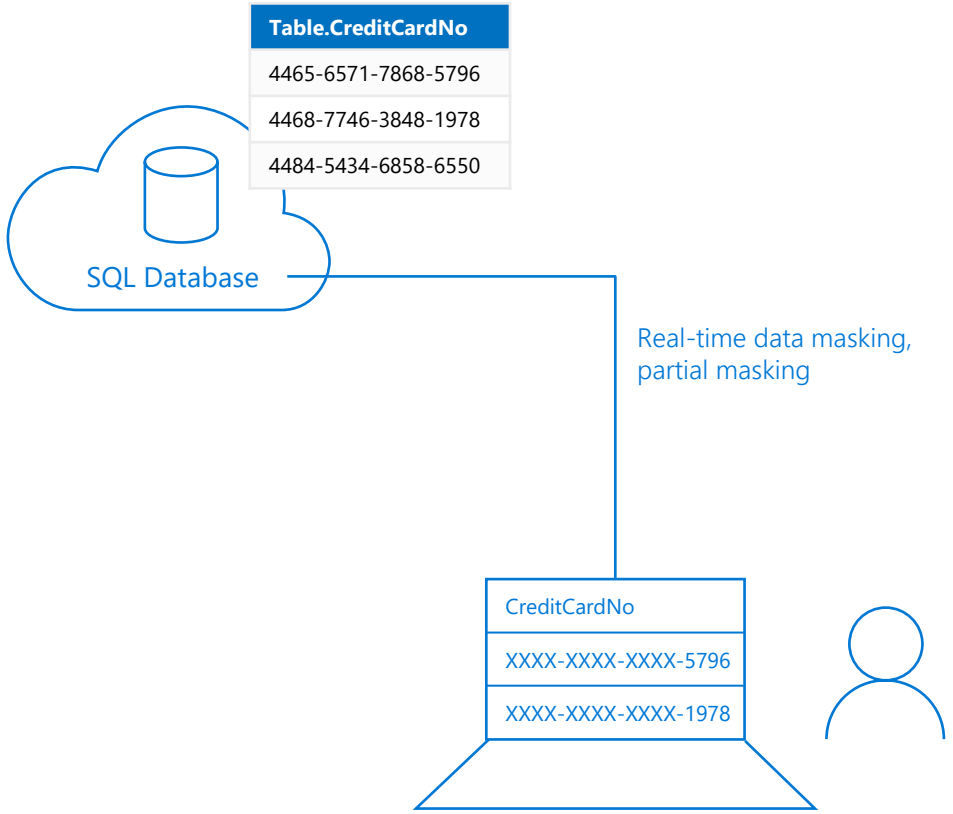
Prevent abuse of sensitive data by hiding it from users

Easy configuration in new Azure Portal

Policy-driven at table and column level, for a defined set of users

Data masking applied in real-time to query results based on policy

Multiple masking functions available, such as full or partial, for various sensitive data categories (credit card numbers, SSN, etc.)



Column Level Encryption

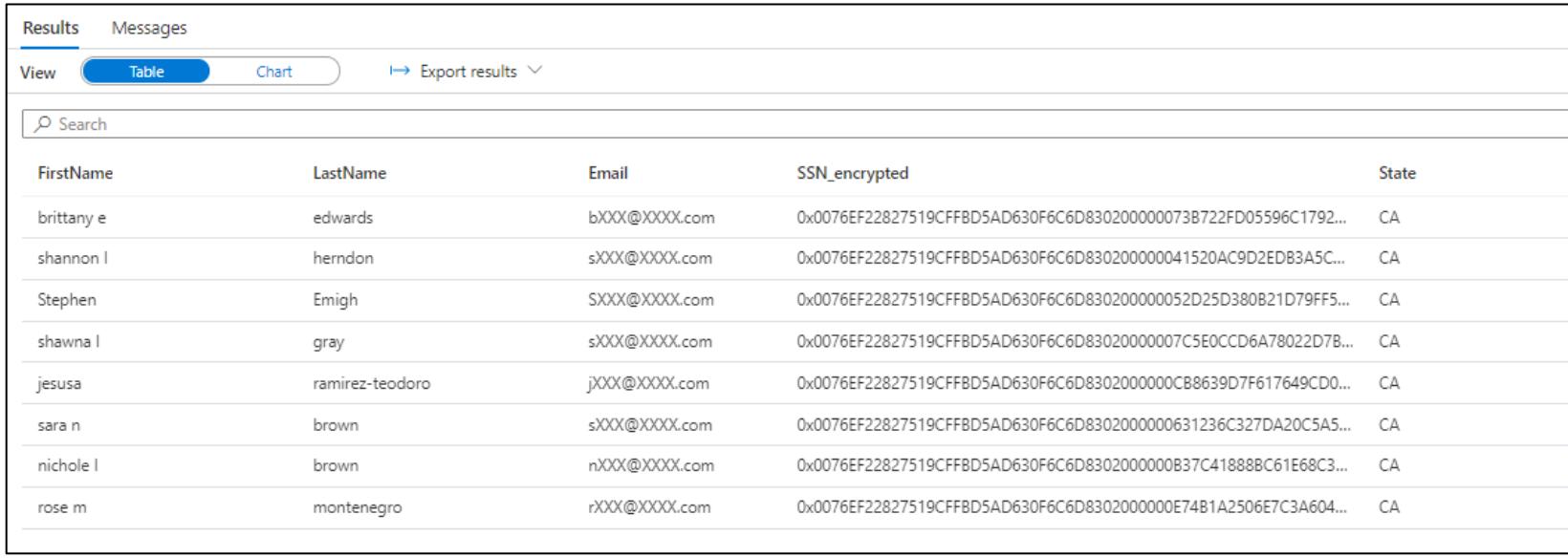
Overview

It helps to implement fine-grained protection of sensitive data within a table in dedicated SQL pool.

The data in CLE enforced columns is encrypted on disk. User need to use DECRYPTBYKEY function to decrypt it.

5 step process to set up CLE

1. Create master key
2. Create certificate
3. Configure symmetric key for encryption
4. Encrypt the column data
5. Close symmetric key



FirstName	LastName	Email	SSN_encrypted	State
brittany e	edwards	bXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D830200000073B722FD05596C1792...	CA
shannon l	herndon	sXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D830200000041520AC9D2EDB3A5C...	CA
Stephen	Emigh	SXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D830200000052D25D380B21D79FF5...	CA
shawna l	gray	sXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D83020000007C5E0CCD6A78022D7B...	CA
jesusa	ramirez-teodoro	jXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D8302000000CB8639D7F617649CD0...	CA
sara n	brown	sXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D8302000000631236C327DA20C5A5...	CA
nichole l	brown	nXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D8302000000B37C41888BC61E68C3...	CA
rose m	montenegro	rXXX@XXXX.com	0x0076EF22827519CFFBD5AD630F6C6D8302000000E74B1A2506E7C3A604...	CA

Dynamic Data Masking

Three steps

1. Security officer defines dynamic data masking policy in T-SQL over sensitive data in the Employee table. The security officer uses the built-in masking functions (default, email, random)

2. The app-user selects from the Employee table

3. The dynamic data masking policy obfuscates the sensitive data in the query results for non-privileged users

Security officer

1

```
ALTER TABLE [Employee]
ALTER COLUMN [SocialSecurityNumber]
ADD MASKED WITH (FUNCTION = 'DEFAULT()')

ALTER TABLE [Employee]
ALTER COLUMN [Email]
ADD MASKED WITH (FUNCTION = 'EMAIL()')

ALTER TABLE [Employee]
ALTER COLUMN [Salary]
ADD MASKED WITH (FUNCTION = 'RANDOM(1,20000)')

GRANT UNMASK to admin1
```

2

Business app

```
SELECT [First Name],
       [Social Security Number],
       [Email],
       [Salary]
FROM   [Employee]
```

3

Non-masked data (admin login)

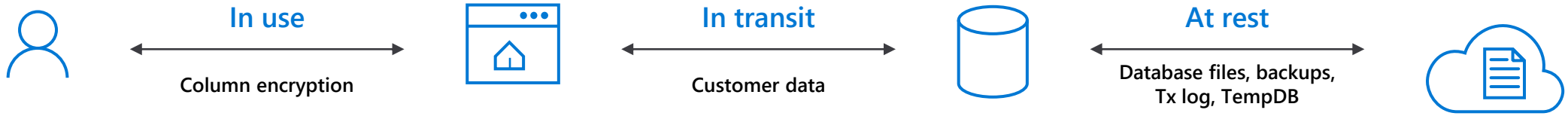
	First Name	Social Security Num...	Email	Salary
1	LILA	758-10-9637	lila.barnett@comcast.net	1012794
2	JAMIE	113-29-4314	jamie.brown@ntlworld.com	1025713
3	SHELLEY	550-72-2028	shelley.lynn@charter.net	1040131
4	MARCELLA	903-94-5665	marcella.estrada@comcast.net	1040753
5	GILBERT	376-79-4787	gilbert.juarez@verizon.net	1041308

Masked data (admin1 login)

	First Name	Social Security Number	Email	Salary
1	LILA	XXX-XX-XX37	lXX@XXXX.net	8940
2	JAMIE	XXX-XX-XX14	jXX@XXXX.com	19582
3	SHELLEY	XXX-XX-XX28	sXX@XXXX.net	3713
4	MARCELLA	XXX-XX-XX65	mXX@XXXX.net	11572
5	GILBERT	XXX-XX-XX87	gXX@XXXX.net	4487

Types of data encryption

Data Encryption	Encryption Technology	Customer Value
In transit	Transport Layer Security (TLS) from the client to the server TLS 1.2	Protects data between client and server against snooping and man-in-the-middle attacks
At rest	Transparent Data Encryption (TDE) for Azure Synapse Analytics	Protects data on the disk User or Service Managed key management is handled by Azure, which makes it easier to obtain compliance



Transparent data encryption (TDE)

Overview

All customer data encrypted at rest

TDE performs real-time I/O encryption and decryption of the data and log files.

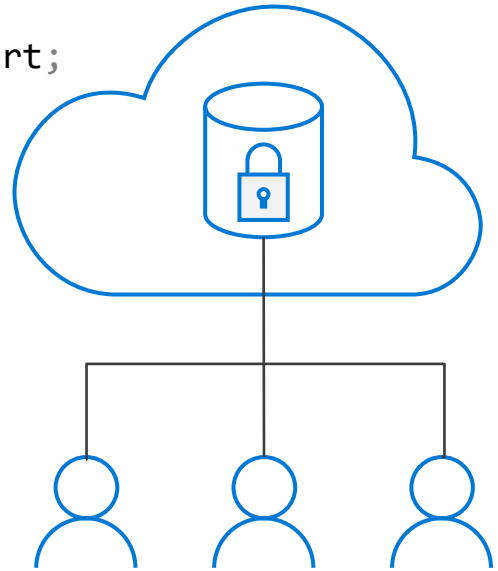
Service OR User managed keys.

Application changes kept to a minimum.

Transparent encryption/decryption of data in a TDE-enabled client driver.

Compliant with many laws, regulations, and guidelines established across various industries.

```
USE master;
GO
CREATE MASTER KEY ENCRYPTION BY PASSWORD = '<UseStrongPasswordHere>';
go
CREATE CERTIFICATE MyServerCert WITH SUBJECT = 'My DEK Certificate';
go
USE MyDatabase;
GO
CREATE DATABASE ENCRYPTION KEY
WITH ALGORITHM = AES_128
ENCRYPTION BY SERVER CERTIFICATE MyServerCert;
GO
ALTER DATABASE MyDatabase
SET ENCRYPTION ON;
GO
```



Transparent data encryption (TDE)

Key Vault

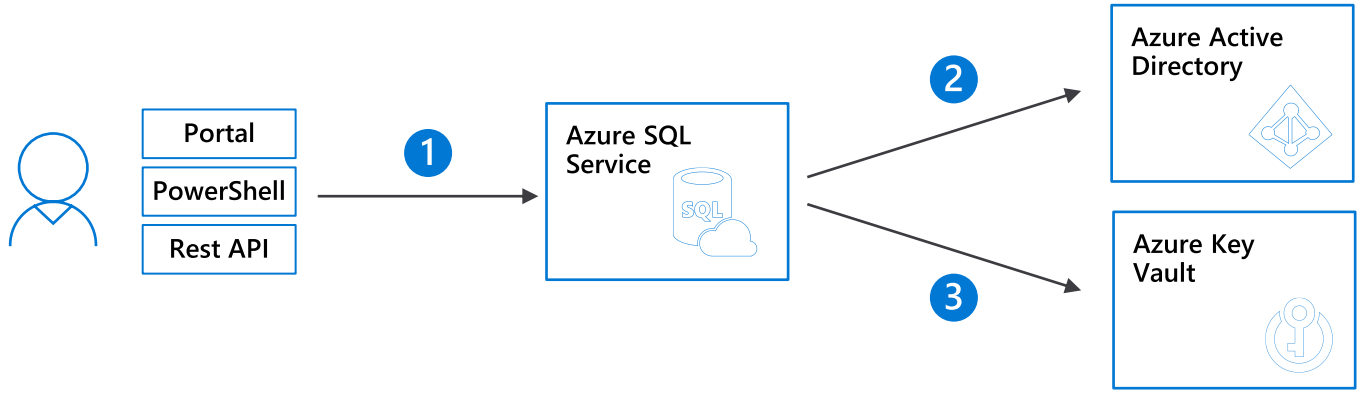
Benefits with User Managed Keys

Assume more control over who has access to your data and when.

Highly available and scalable cloud-based key store.

Central key management that allows separation of key management and data.

Configurable via Azure Portal, PowerShell, and REST API.



1. The Key Vault admin grants vault access to the SQL Database server using its unique Azure Active Directory (AD) identity
2. The server uses its Azure AD identity to authenticate with Azure AD for access to your Key Vault
3. The server sends get, wrap key, and unwrap key request to the asymmetric key in key Vault for database encryption key protection.

Microsoft

Azure Synapse Analytics

Metastore

Metastore

Overview

It offers the different computational engines of a workspace to share databases and Parquet-backed tables between its Apache Spark pools, serverless SQL pool, and dedicated SQL pool.

Benefits

- The shared metadata model supports the modern data warehouse pattern.
- The Spark created databases and all their tables become visible in any of the Azure Synapse workspace Spark pool instances and can be used from any of the Spark jobs provided necessary permissions are provided.
- Databases are created automatically in the serverless SQL pool metadata.
- The external and managed tables created by Spark job are made accessible as external tables in the serverless SQL pool metadata in the dbo schema of the corresponding database.
- Spark created databases and their Parquet-backed tables will be mapped into the SQL pools for which metadata synchronization enabled.

Microsoft

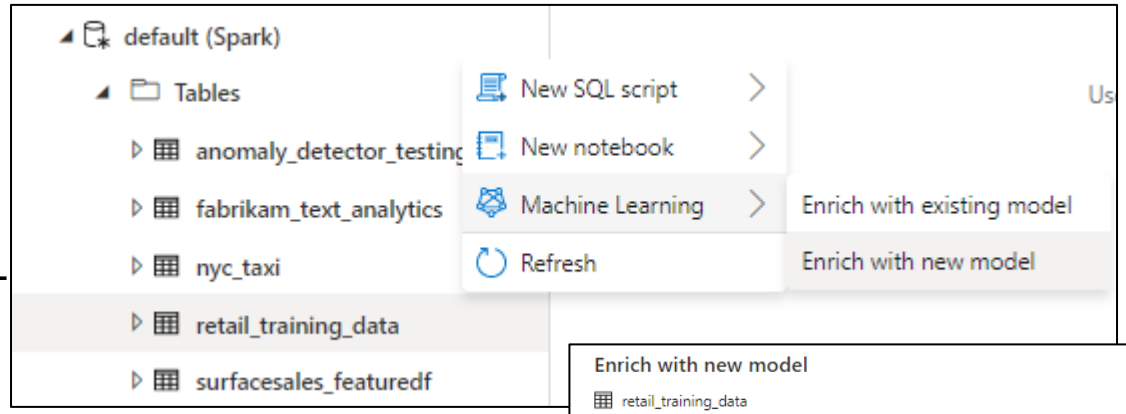
Azure Synapse Analytics

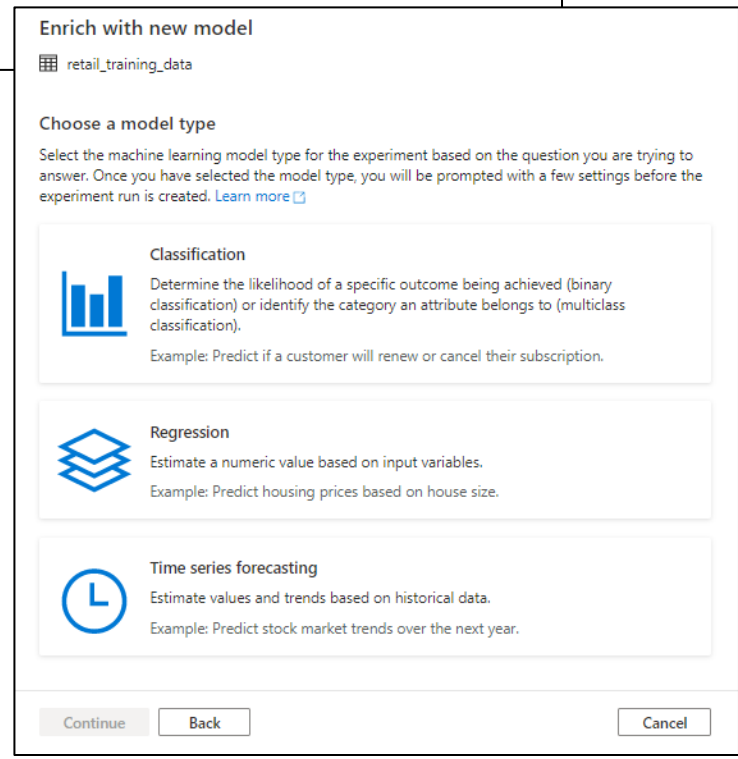
Linked Services

Azure Machine Learning

Overview

Native integration with Azure Machine Learning offers training model in code-first and code-free format. It enables ease of access to models from Machine Learning model registry for deployment in Synapse dedicated SQL pools and launch predictions to enrich the data.





Benefits

Train models in code free manner with AutoML capabilities from Synapse.

Register models in Machine Learning model registry.

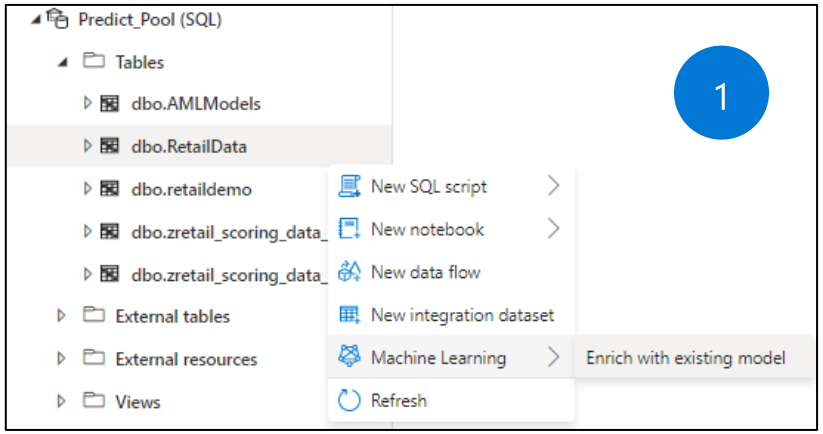
Deploy ONNX models in dedicated SQL pool to enrich data.

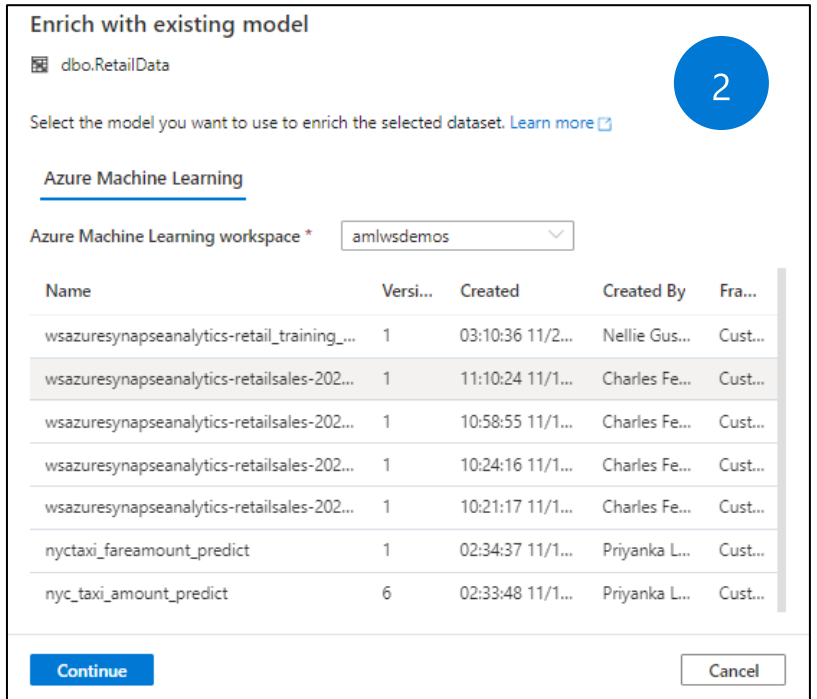
No data movement to get predictions.

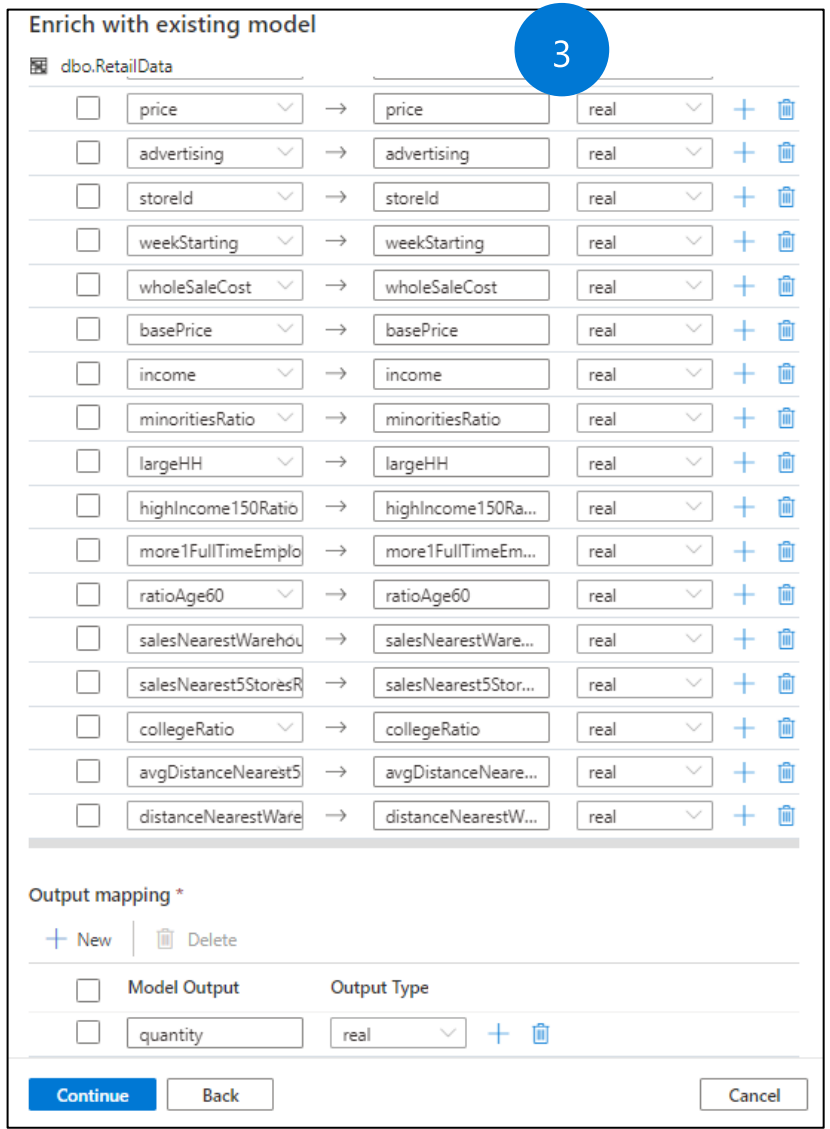
Enables in-engine scoring the data and operationalize predictions.

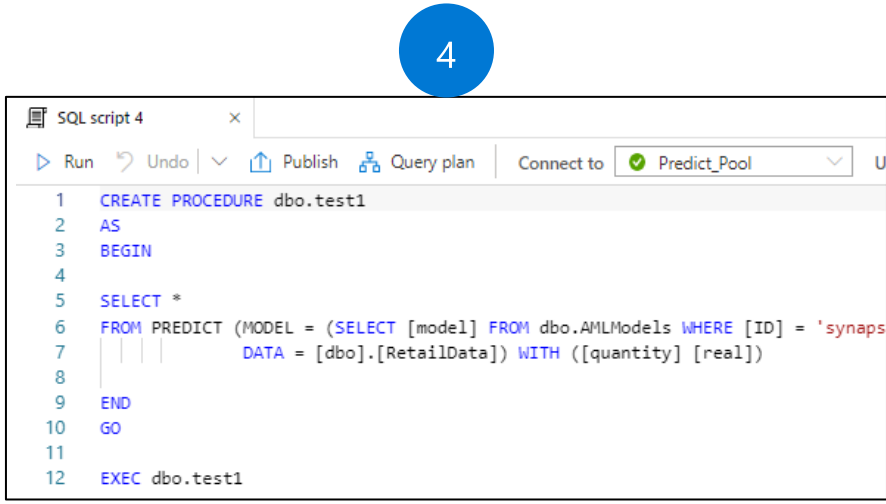
Azure Machine Learning (Continued)

Deploy ONNX models in dedicated SQL pool to enrich data and operationalize with stored procedure









Azure Cognitive Services

Overview

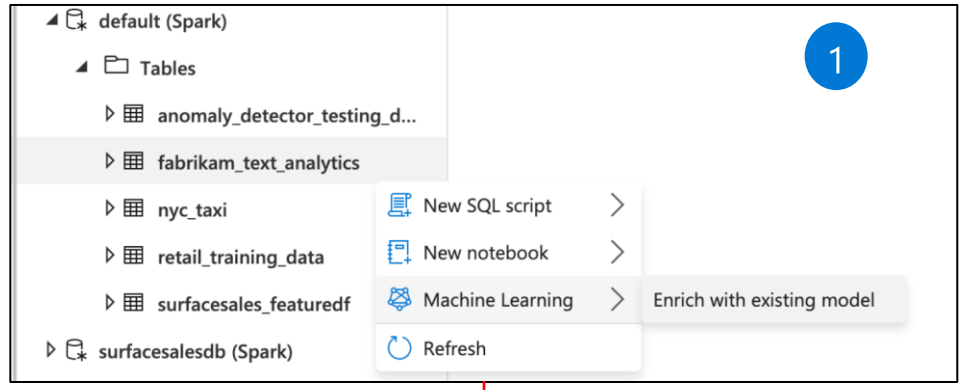
Azure Cognitive Services provides machine learning capabilities to solve general problems such as analyzing text for emotional sentiment or analyzing images to recognize objects or faces.

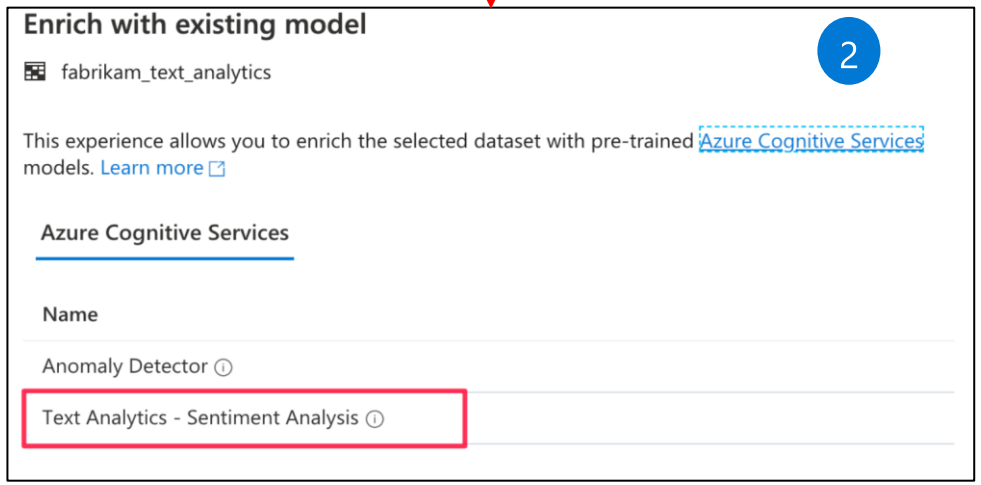
It enables adding cognitive features to analytics applications without having artificial intelligence (AI) or data science skills.

Benefits:

Add Azure Cognitive Services as a Linked service in Azure Synapse and start enriching data in Azure Synapse with capabilities like anomaly detection or sentiment analysis.

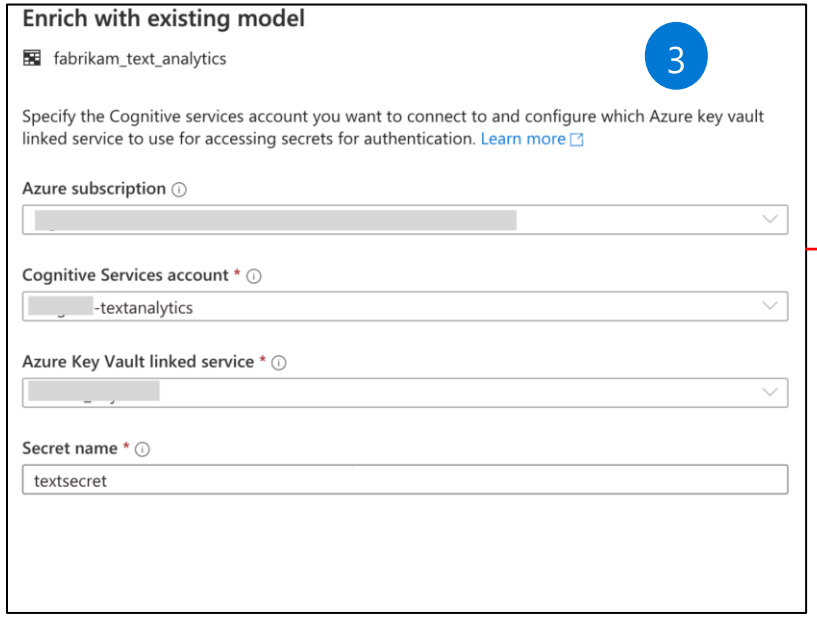
No data movement to get predictions.



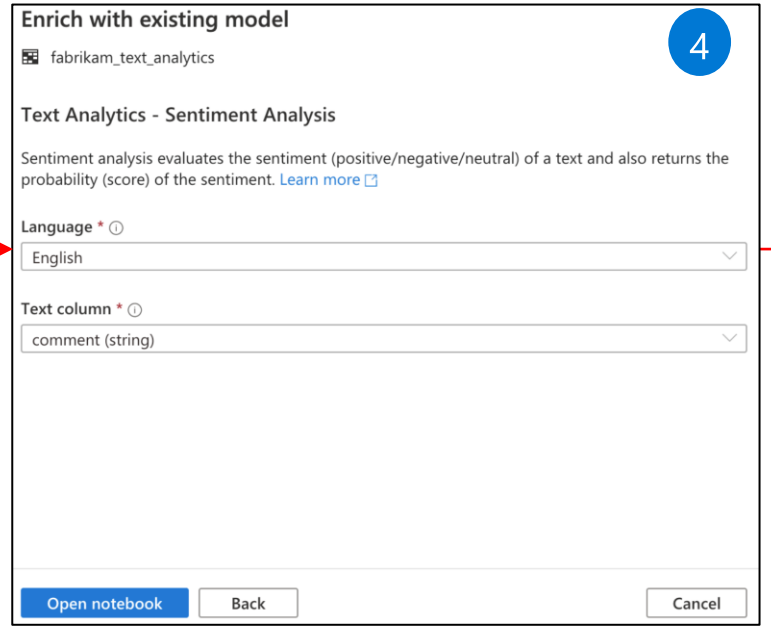


Azure Cognitive Services (continued)

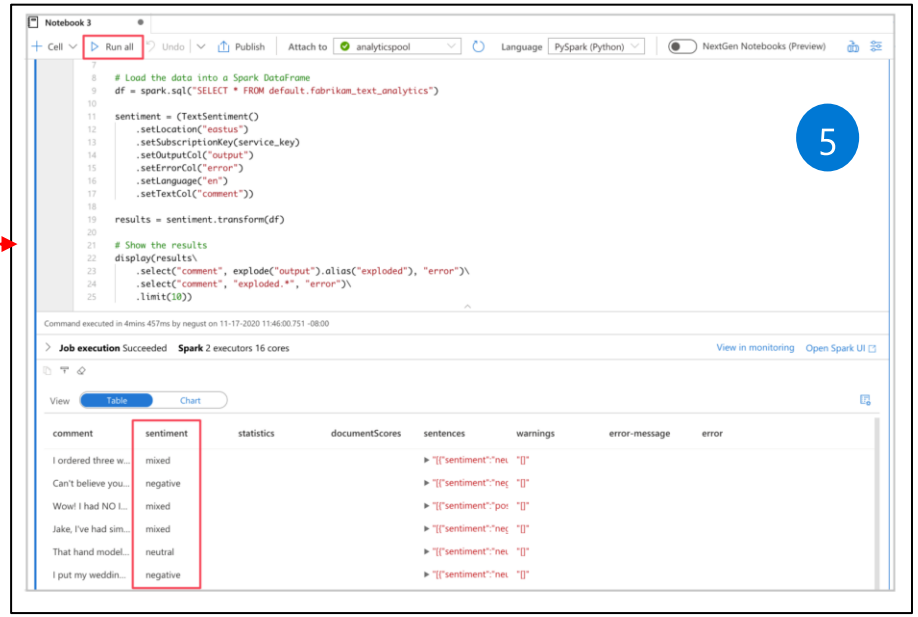
Provide authentication details



Configure sentiment analysis



Open Notebook and run



Azure Data Explorer

Overview

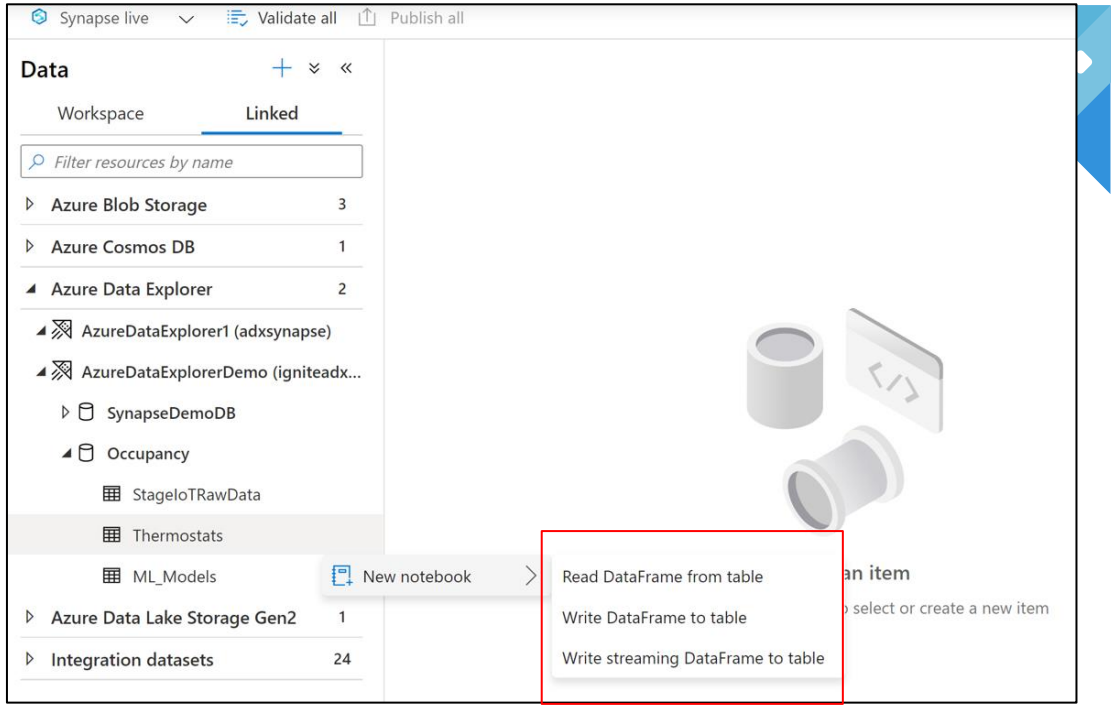
Azure Data Explorer is a fast, fully managed data analytics service for real-time analysis on large volumes of Telemetry, Logs, Time Series data streaming from applications, websites, IoT devices, and more

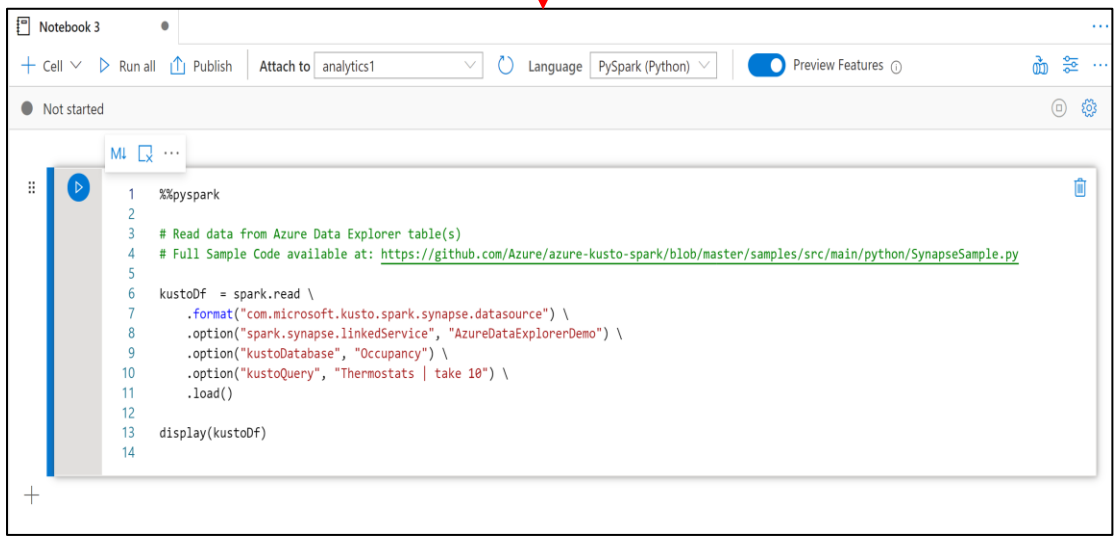
The Azure Data Explore data connector for Azure Synapse Workspace is an extension of the Azure Data Explorer Spark connector that is now natively integrated into Azure Synapse Apache Spark pools

Benefits

Use the continuous export feature to export data from Azure Data Explorer to Data Lake Store linked to Azure Synapse Workspace for archiving data in the data lake

Use the new ADX connector in Azure Synapse to query the cached and indexed data in Azure Data Explorer using Synapse Apache Spark



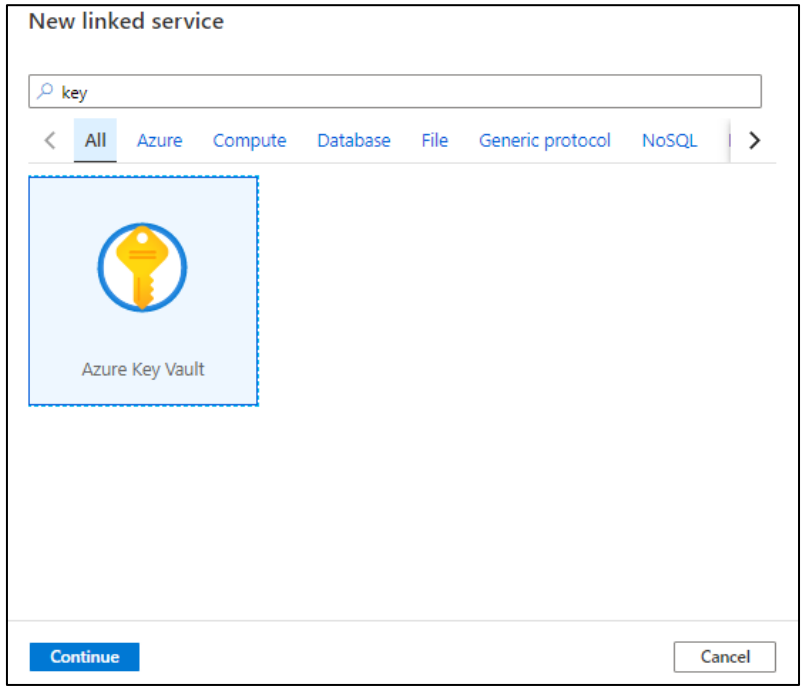


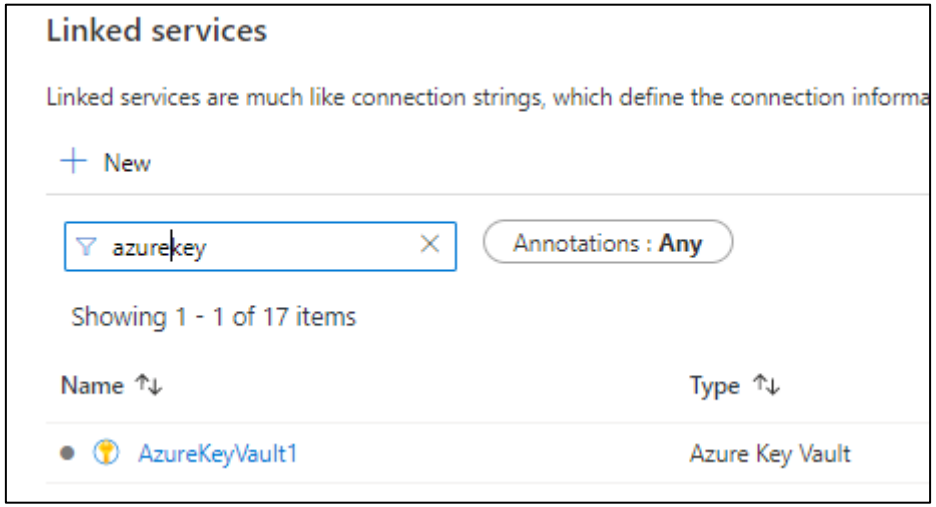
Azure Key Vault

Overview

Azure Key Vault integration offers increased security and control over keys and password.

It enables to securely store and tightly control access to tokens, passwords, secrets.





Azure Monitor

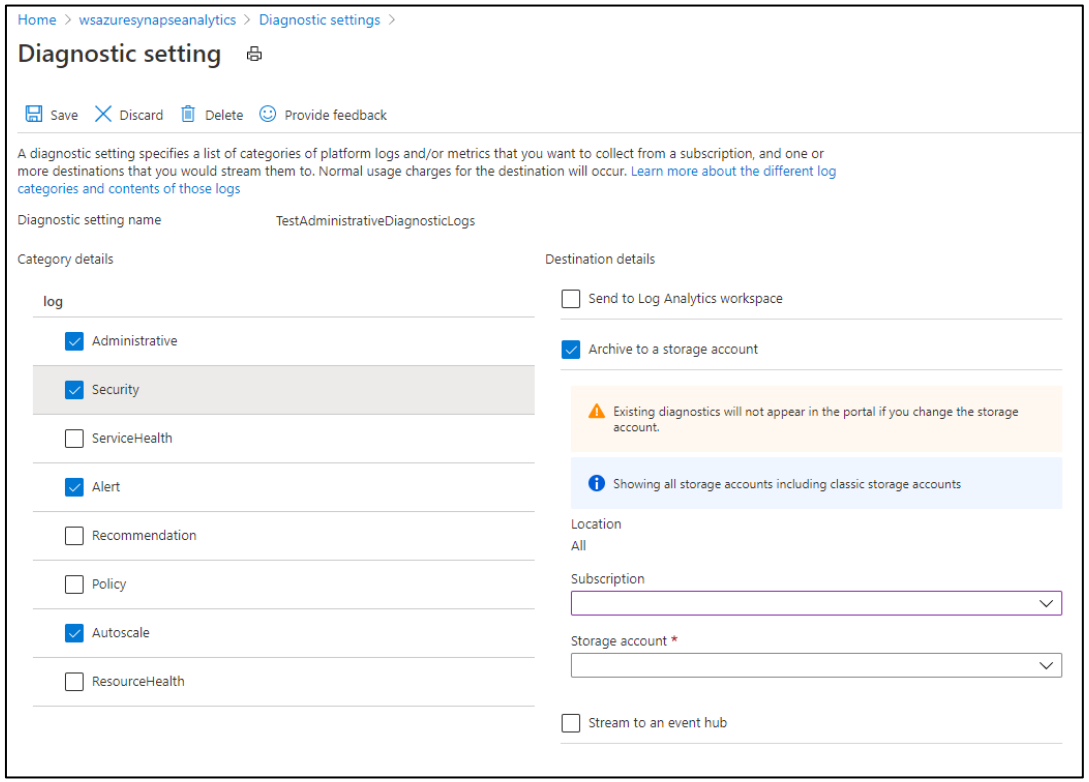
Overview

Leverage Azure Monitor to obtain metrics, alerts, and logs.

Azure Synapse Analytics can write diagnostic logs in Azure Monitor.

Workspace level metrics, dedicated SQL pool level metrics, Apache Spark pool metrics are supported.

With Azure Monitor diagnostic settings, you can route diagnostic logs for analysis to multiple different targets.



With Azure Monitor diagnostic settings, you can route diagnostic logs for analysis to multiple different targets

- Storage Account
- Log Analytics workspace
- Event Hub

Microsoft

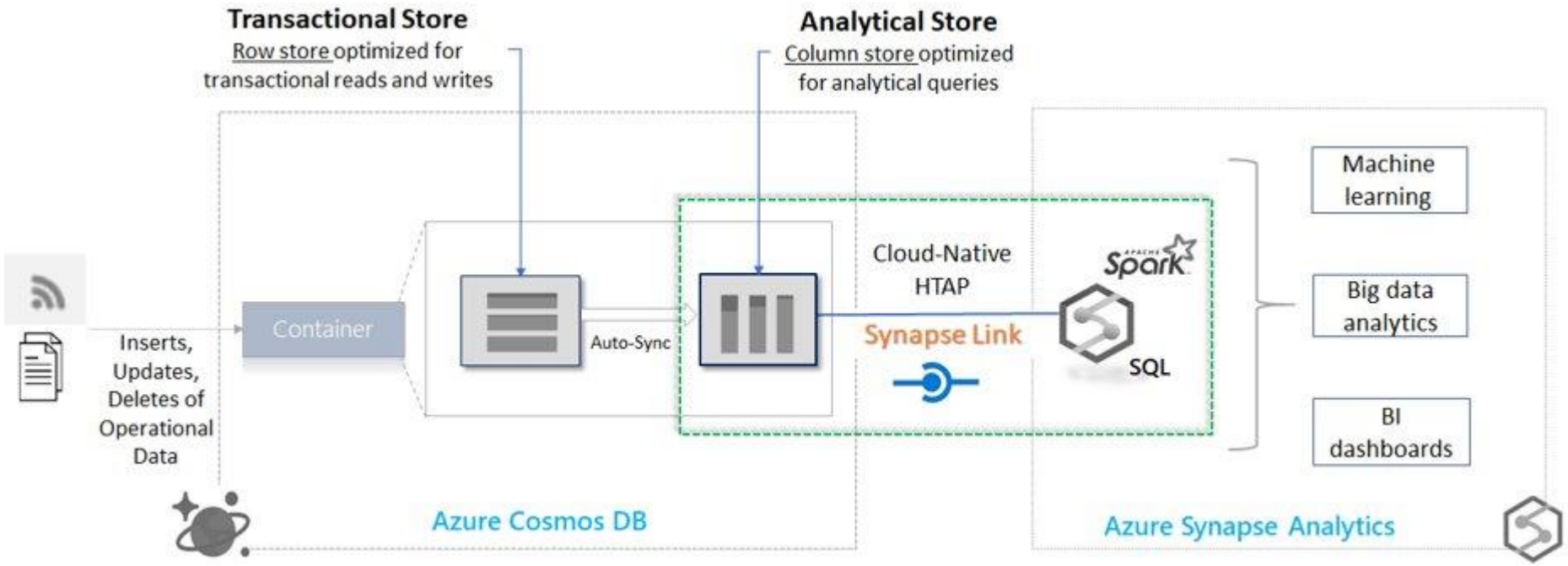
Azure Synapse Analytics

Synapse Link for Cosmos DB

Synapse Link for Cosmos DB integration

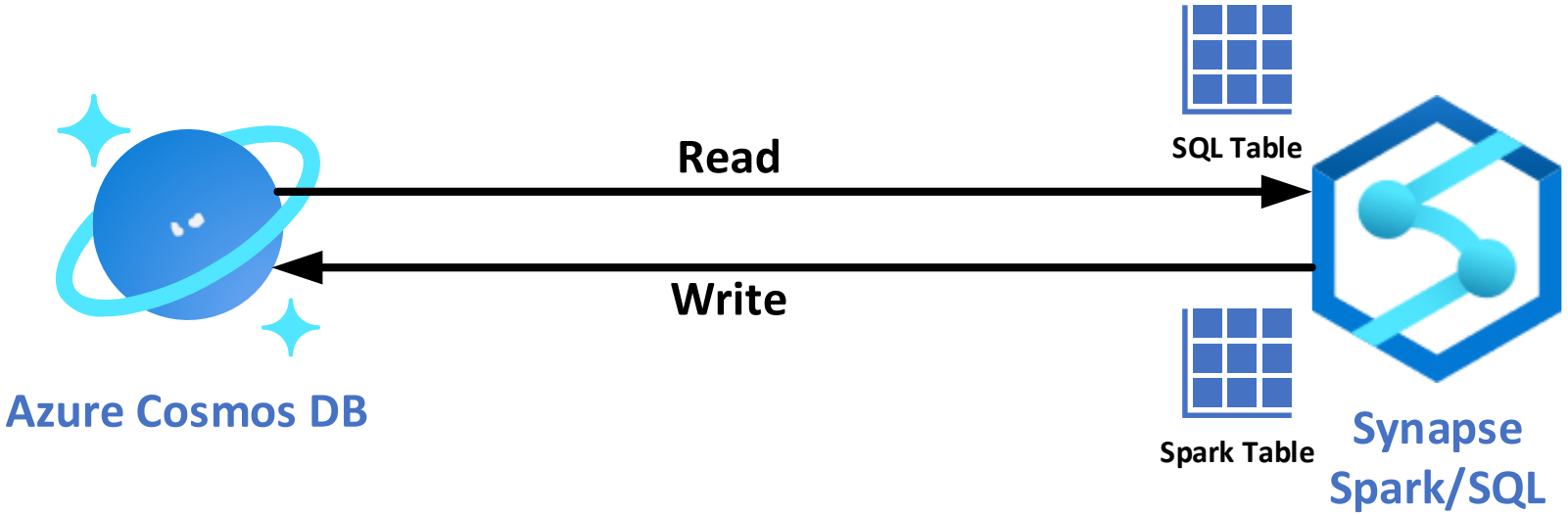
Cloud native hybrid transactional and analytical processing (HTAP) capability

In-place analytics over operational data in Azure Cosmos DB



Modern Data Warehouse Architecture

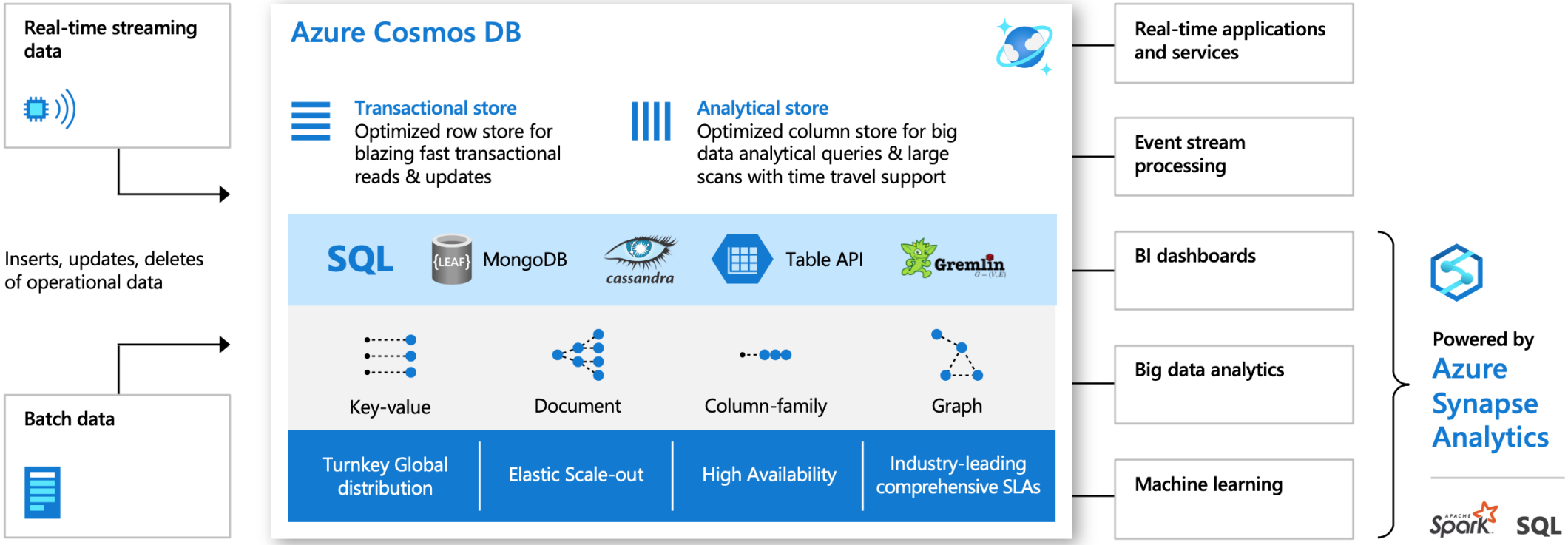
With Synapse Link for Cosmos DB



HTAP reduces architectural complexity by collapsing the stack to:

- Accelerate new insight to perform near-real time analytics at the source with no impact on the transactional workload performance
- Deliver analytics and operational insights through Spark or SQL

(Cosmos DB + Synapse) ♥ Analytics



Benefits of running HTAP in Cosmos DB and Synapse

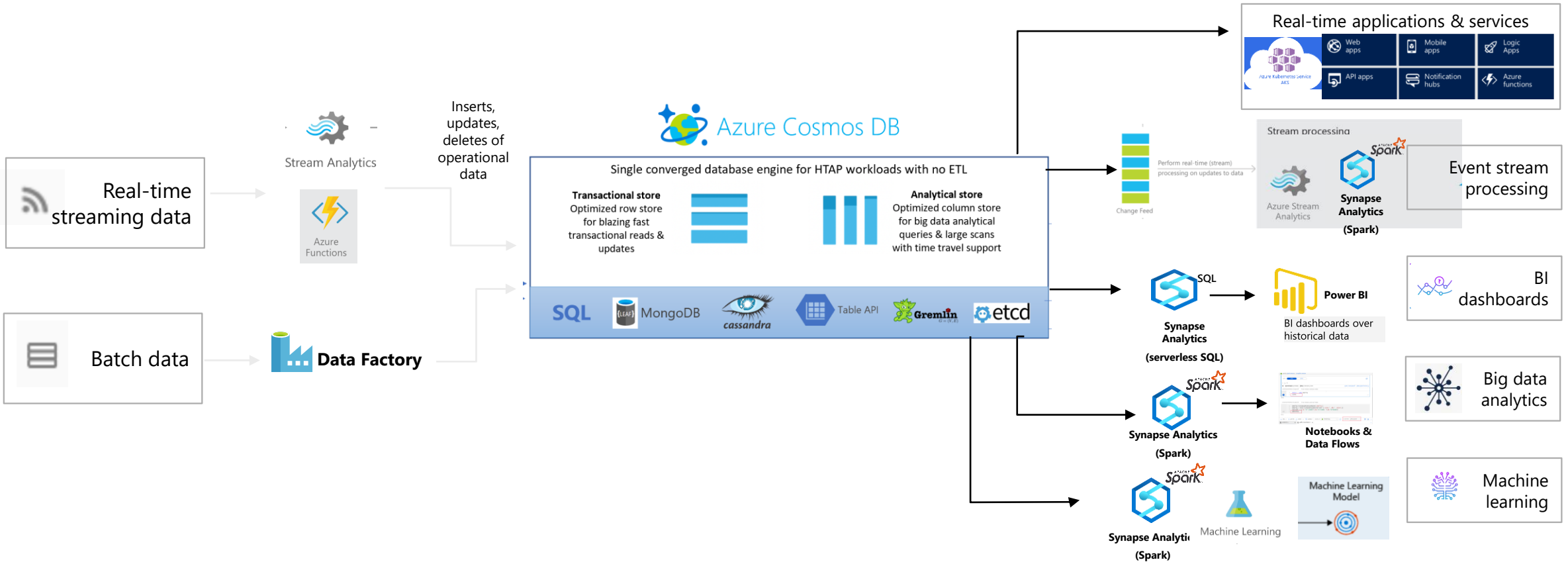
Operational Needs

- High throughput ingestion of both batch & streaming feeds with real-time indexing
- Global distribution with active-active setup
- Apps to perform real-time CRUD & queries using developer friendly NoSQL APIs
- No downtime elastic scaling of database
- SLAs for latency/availability/consistency/throughput

Analytics Needs

- Run analytics without impacting operational performance
- Low cost on storage and compute to run near-real time analytics workloads
- Analyze in-place data, eliminating ETL complexity, with Synapse Spark and SQL
- Rich BI ecosystem of SQL available for cheaper & faster queries over analytical storage
- Native integration of Azure Enterprise AI ecosystem for training and scoring ML models over analytical storage

High level reference architecture for HTAP workloads



End