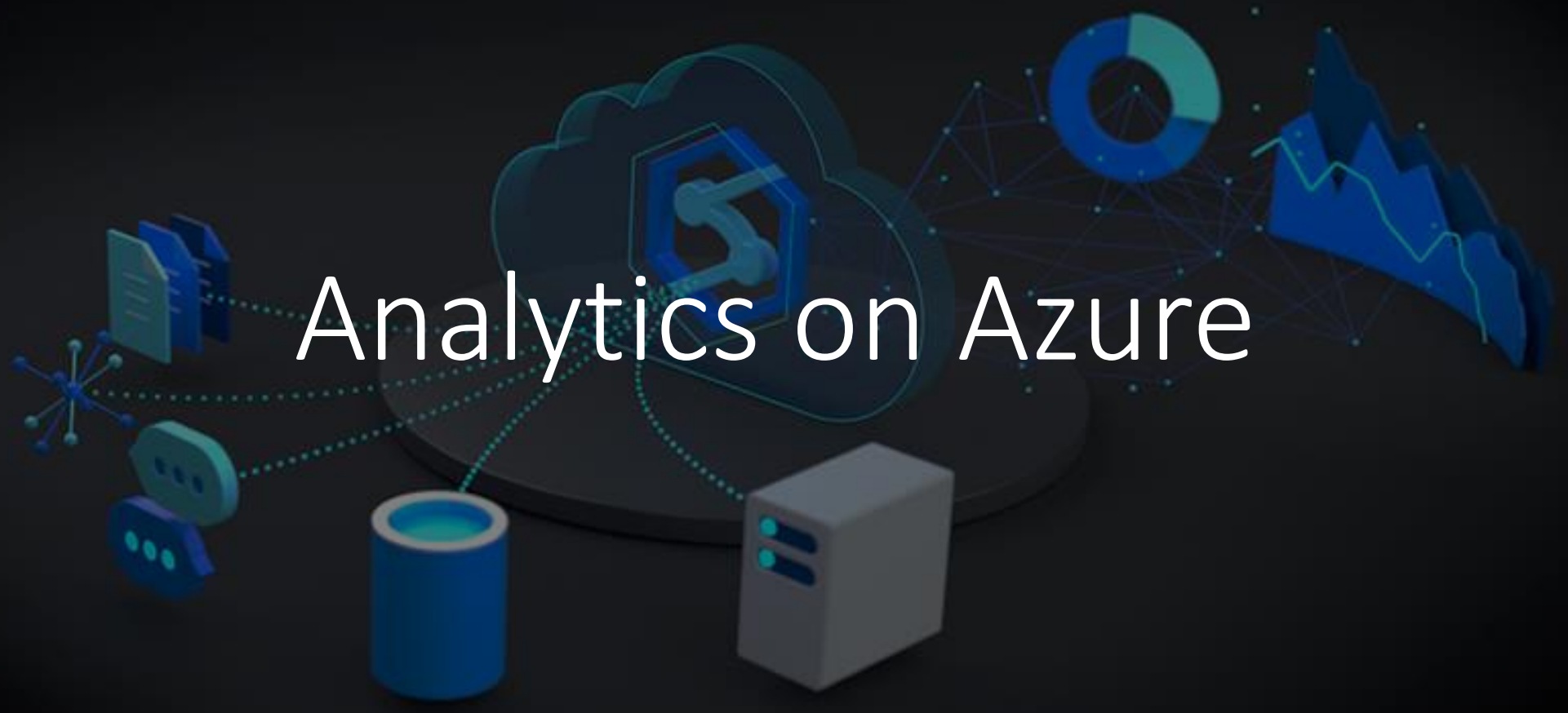
Microsoft Azure

NephOSystems
Cloud Computing for life

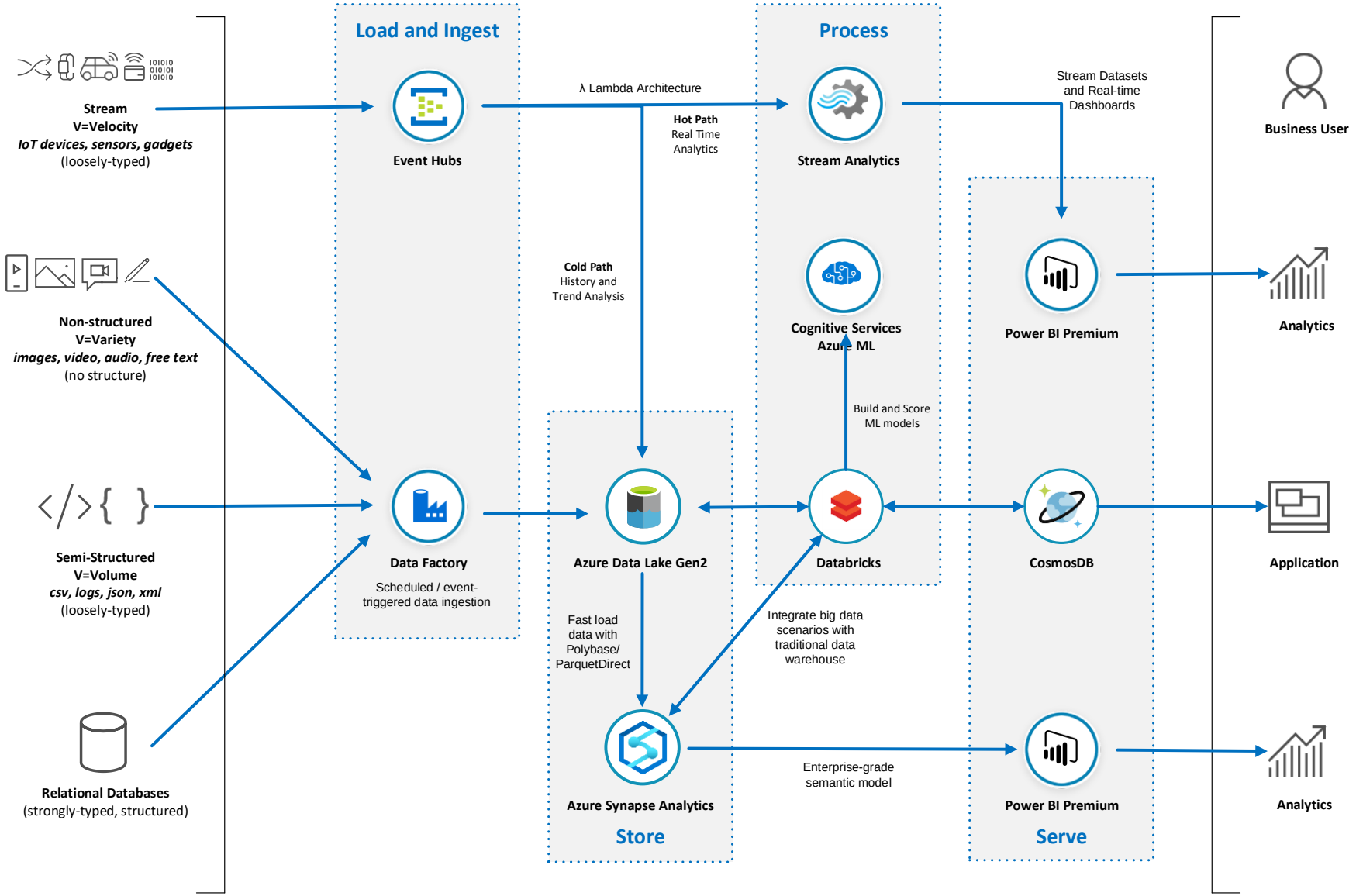
Analytics on Azure



Microsoft Azure

NephOSystems
Cloud Computing for life

Modern Data Platform Reference Architecture



NephOSystems
Cloud Computing for life

Big Data and Analytics with Azure Data Lake

Microsoft Azure

What is a Data Lake?

It is a central storage repository that holds data coming from many sources in a raw, granular format. It can store **structured, semi-structured, or unstructured data**, which means data ingested quickly and can be kept in a more flexible format for future use cases.

Characteristics

- Schema-on-read (ELT)
- Collection of data, not a platform
- Perfect place for evolving data

Benefits

- Quickly ingest high volumes of diverse data structures
- Enable advanced analytics and data exploration
- Scalability and storage cost reduction

Best Practices

- Data Governance needed to avoid Data Swamp
- Security considerations
- Design your Data Lake
- Metadata management

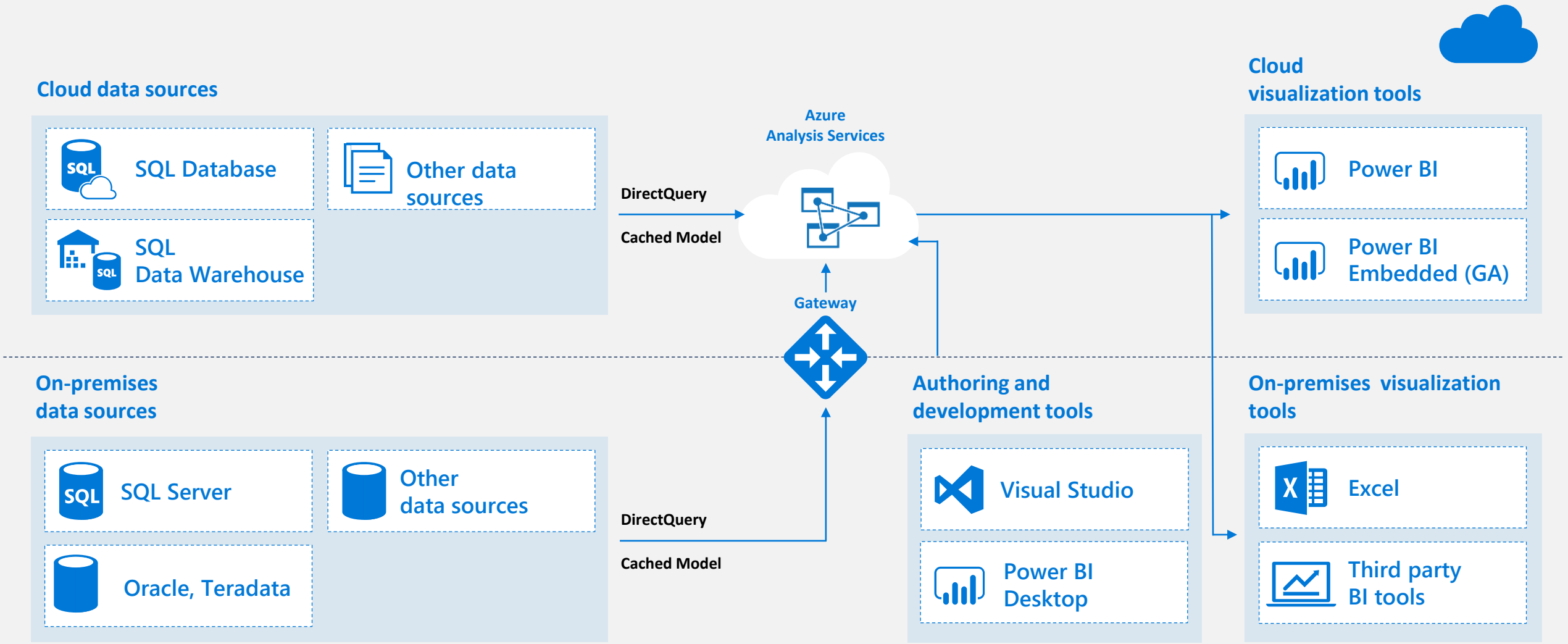
NephOSystems
Cloud Computing for life

Azure Analysis Services

Microsoft Azure

NephOSystems
Cloud Computing for life

Azure Analysis Services Architecture



Note: not all capabilities available at public preview

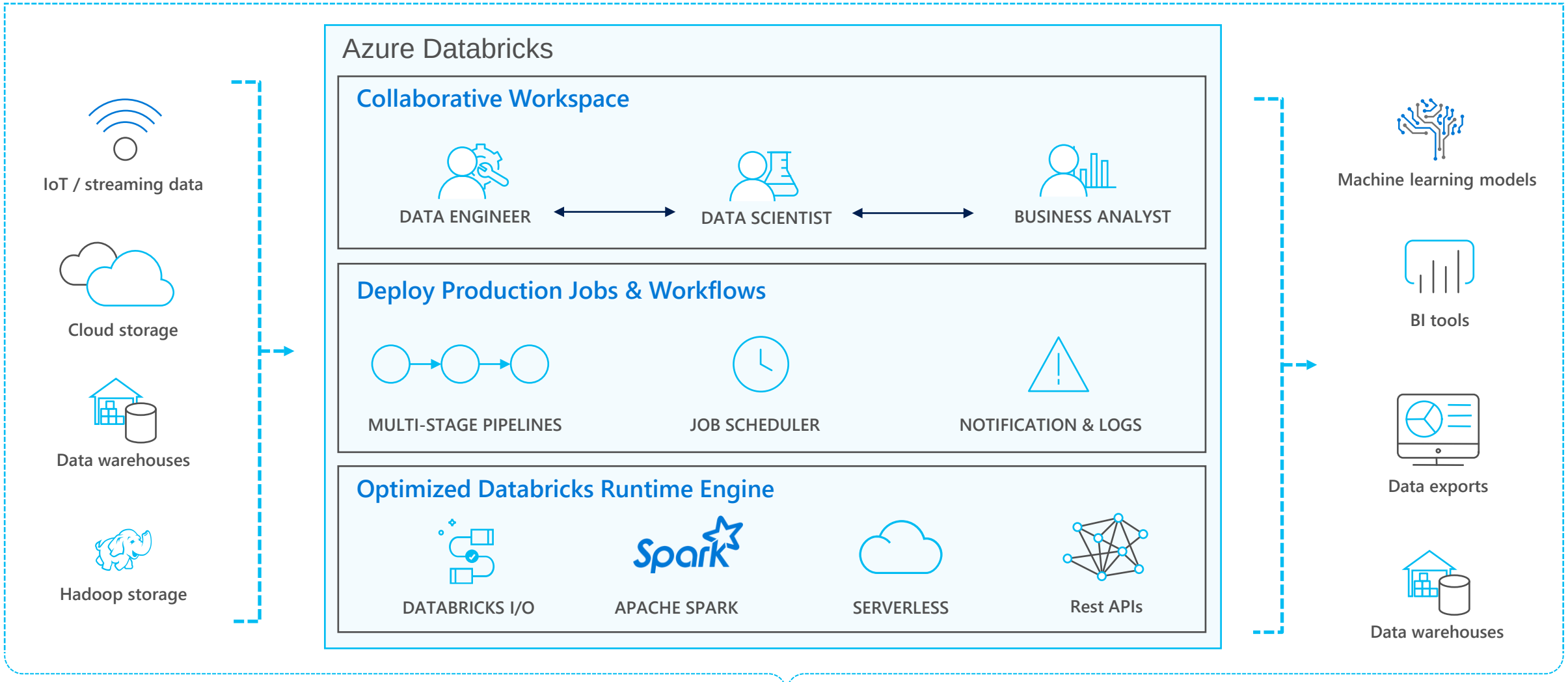
NephOSystems
Cloud Computing for life

Azure Databricks

Microsoft Azure

NephOSystems
Cloud Computing for life

Azure Databricks



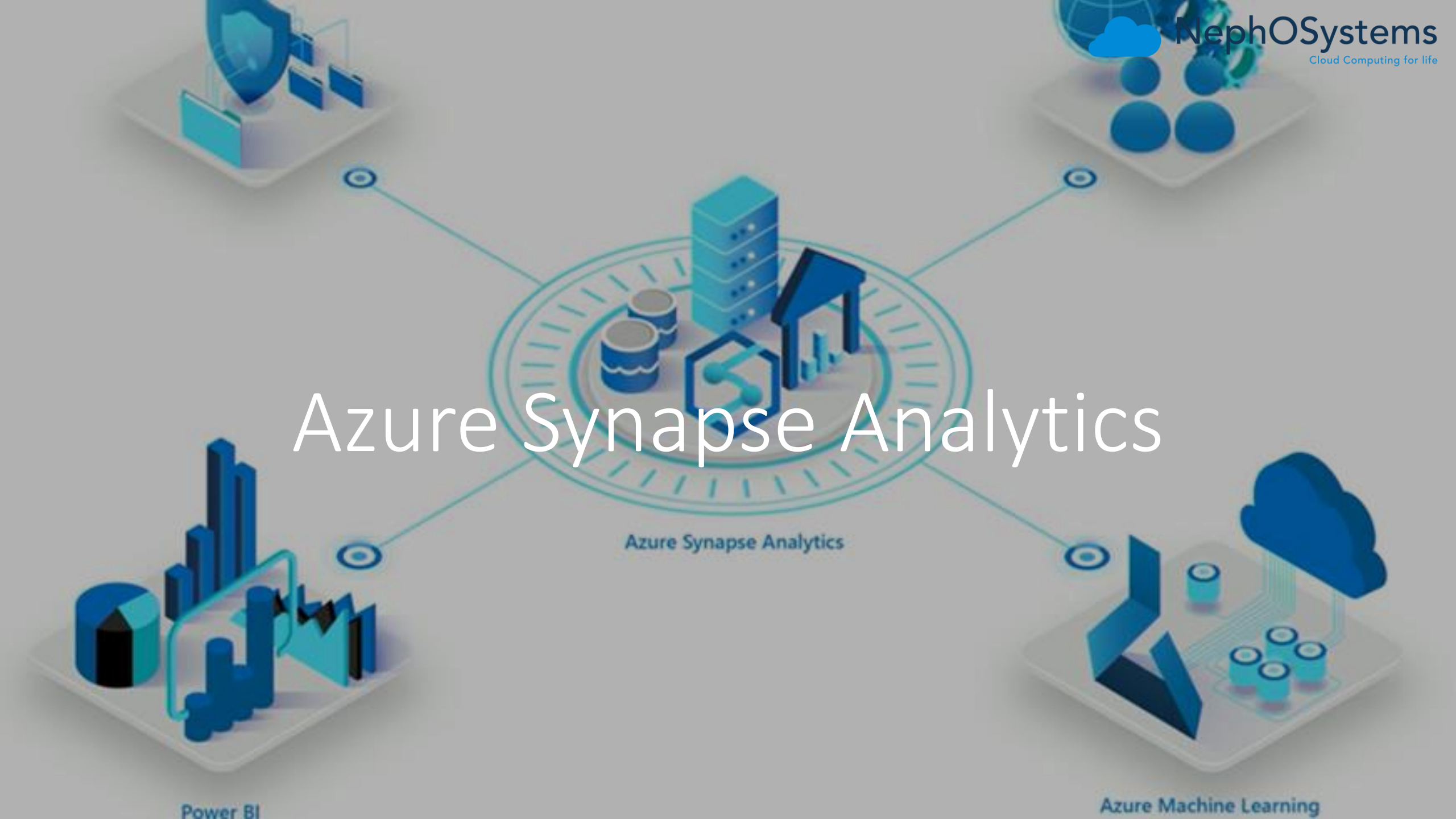
Enhance Productivity

Build on secure & trusted cloud

Scale without limits

NephOSystems
Cloud Computing for life

Azure Synapse Analytics

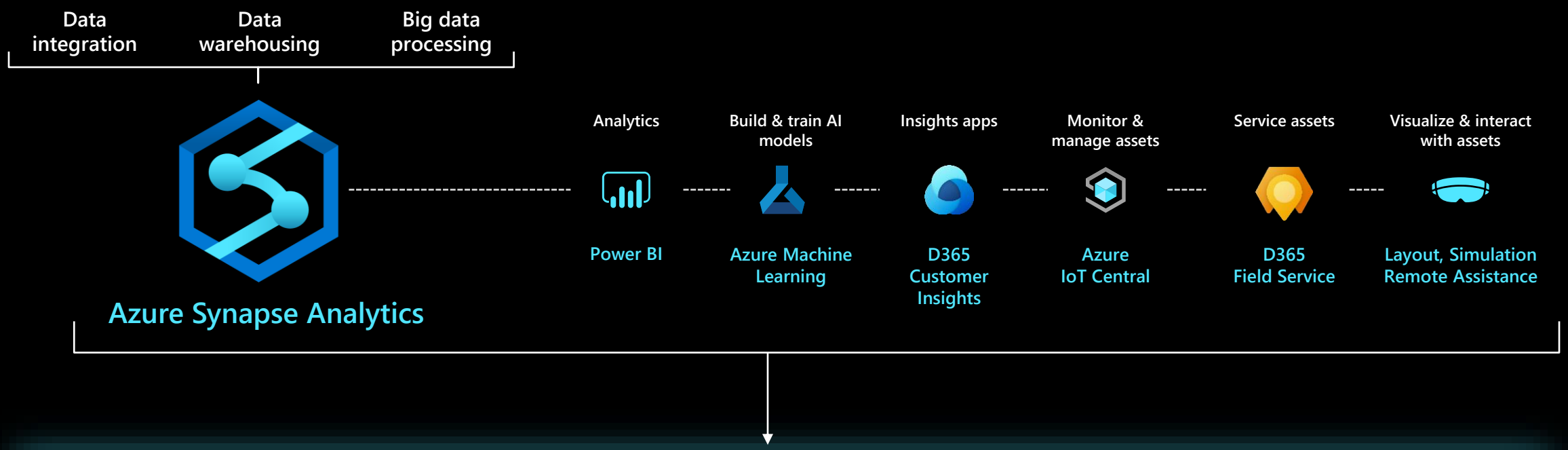


NephOSystems
Cloud Computing for life

Azure Synapse Analytics

Limitless analytics service with unmatched time to insight

Limitless scale | Powerful insights | Unified experience | Instant clarity | Unmatched security



Azure Data Lake Storage + Common Data Model

Microsoft Azure

Microsoft Azure

NephOSystems
Cloud Computing for life

Azure Synapse Analytics

Integrated data platform for BI, AI and continuous intelligence

Artificial Intelligence / Machine Learning / Internet of Things
Intelligent Apps / Business Intelligence

Azure Synapse Analytics

Experience

Azure Synapse Analytics Studio

Platform

- MANAGEMENT
- SECURITY
- MONITORING
- METASTORE

Languages: SQL | Python | .NET | Java | Scala | R

Form Factors: PROVISIONED | ON-DEMAND

Analytics Runtimes: SQL | Apache Spark

DATA INTEGRATION

Azure
Data Lake Storage

Common Data Model
Enterprise Security
Optimized for Analytics

Designed for analytics **workloads at any scale**

SaaS **developer experiences** for code free and code first

Multiple **languages** suited to different analytics workloads

Integrated analytics runtimes available provisioned and serverless on-demand

SQL Analytics offering T-SQL for batch, streaming and interactive processing

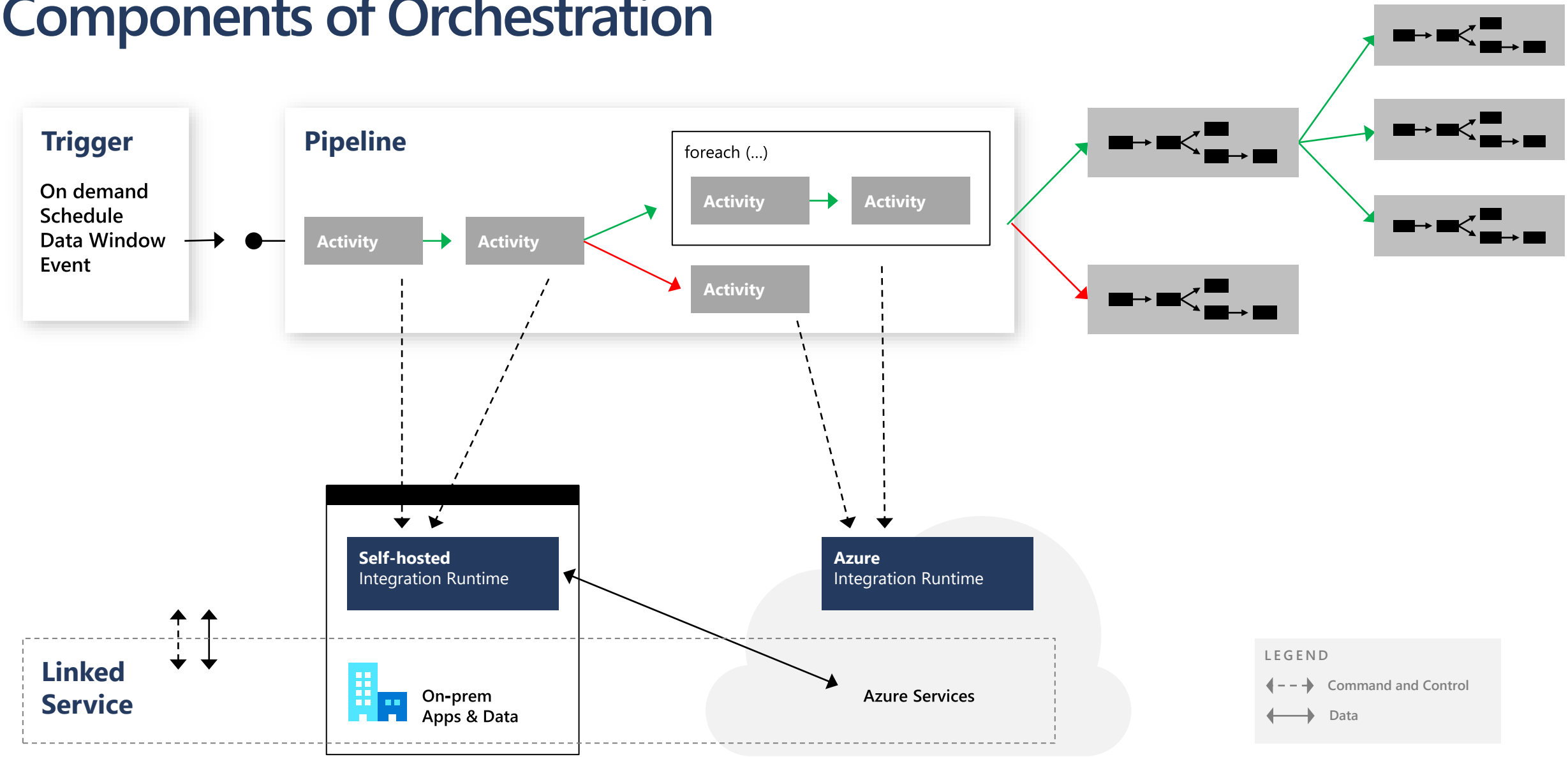
Spark for big data processing with Python, Scala, R and .NET

Integrated **platform services** for, management, security, monitoring, and metastore

Data **lake integrated** and Common Data Model aware

NephOSystems
Cloud Computing for life

Components of Orchestration

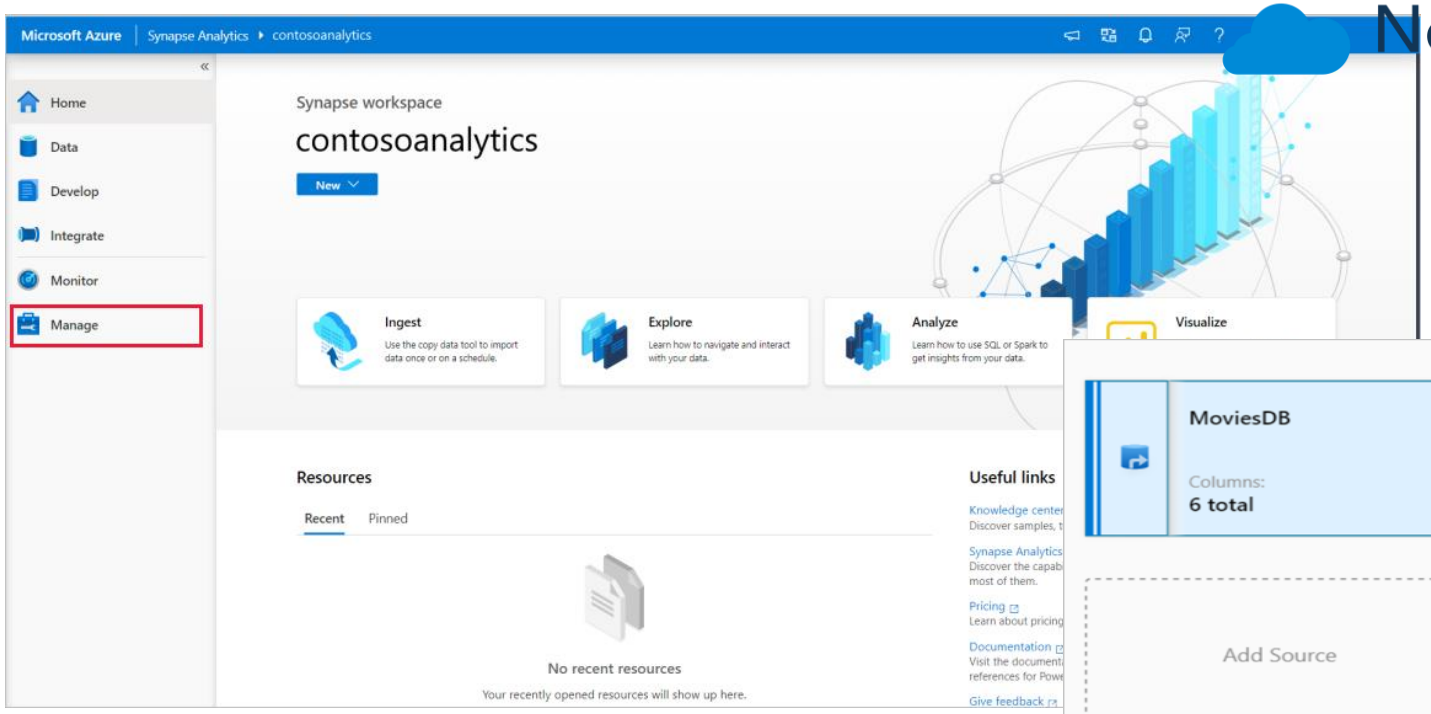


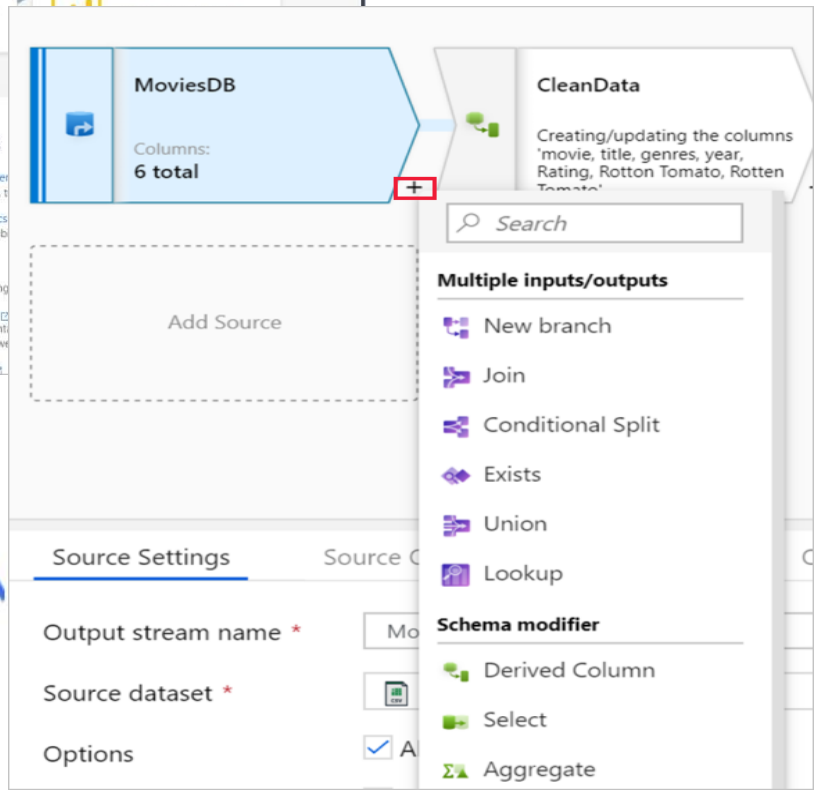
Synapse Pipelines shares codebase with Azure Data Factory

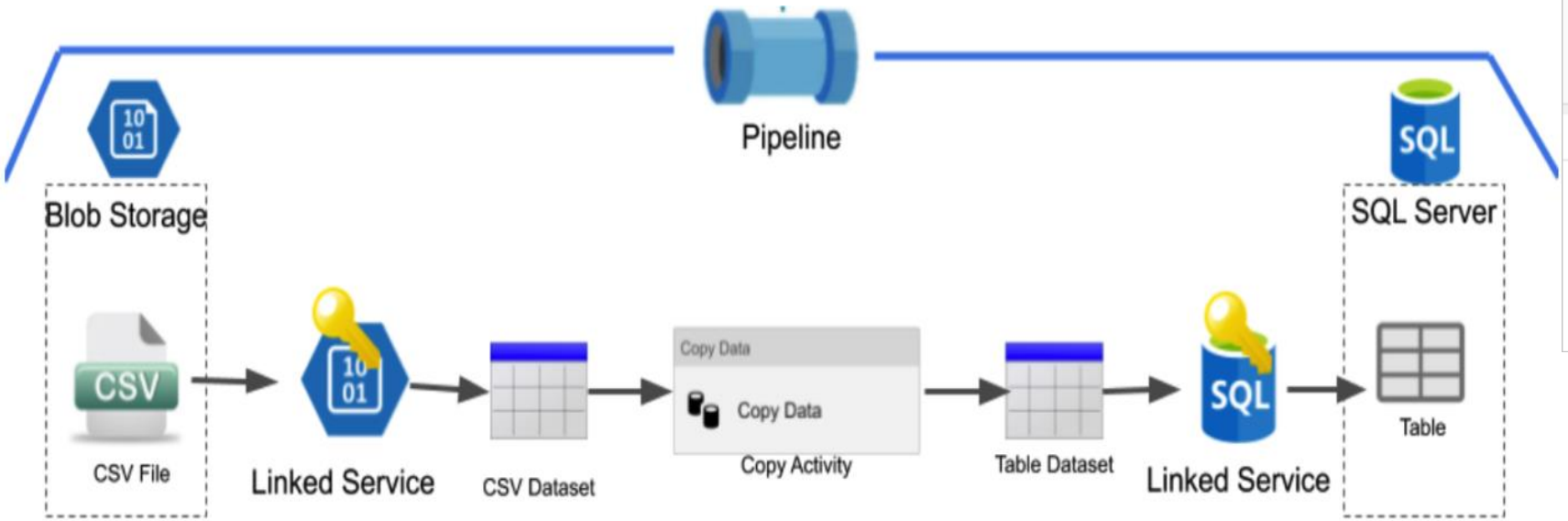
Microsoft Azure

NephOSystems
Cloud Computing for life

Synapse Workspace And Pipeline







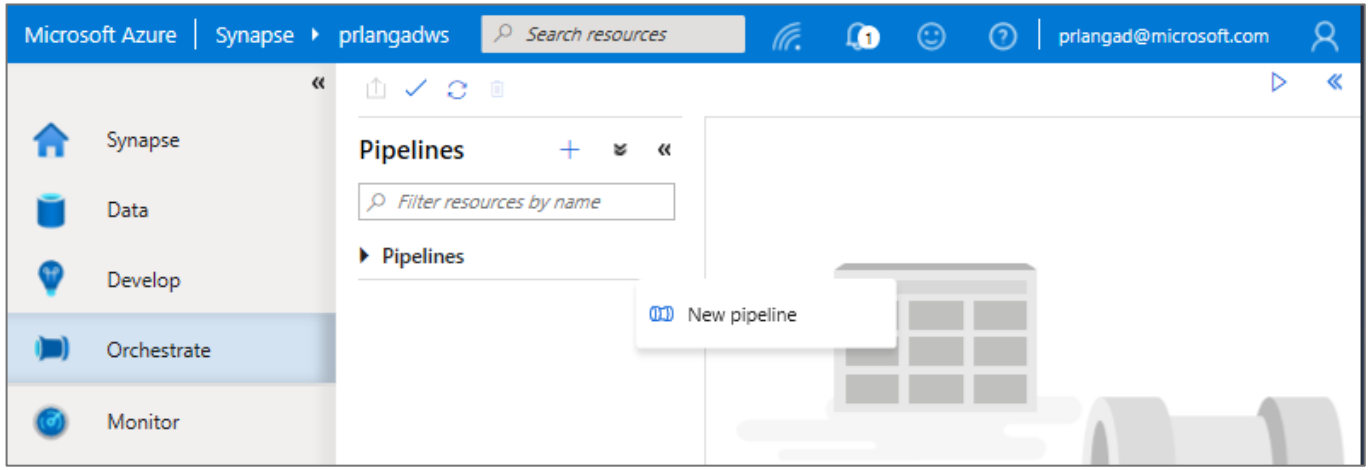
Pipelines

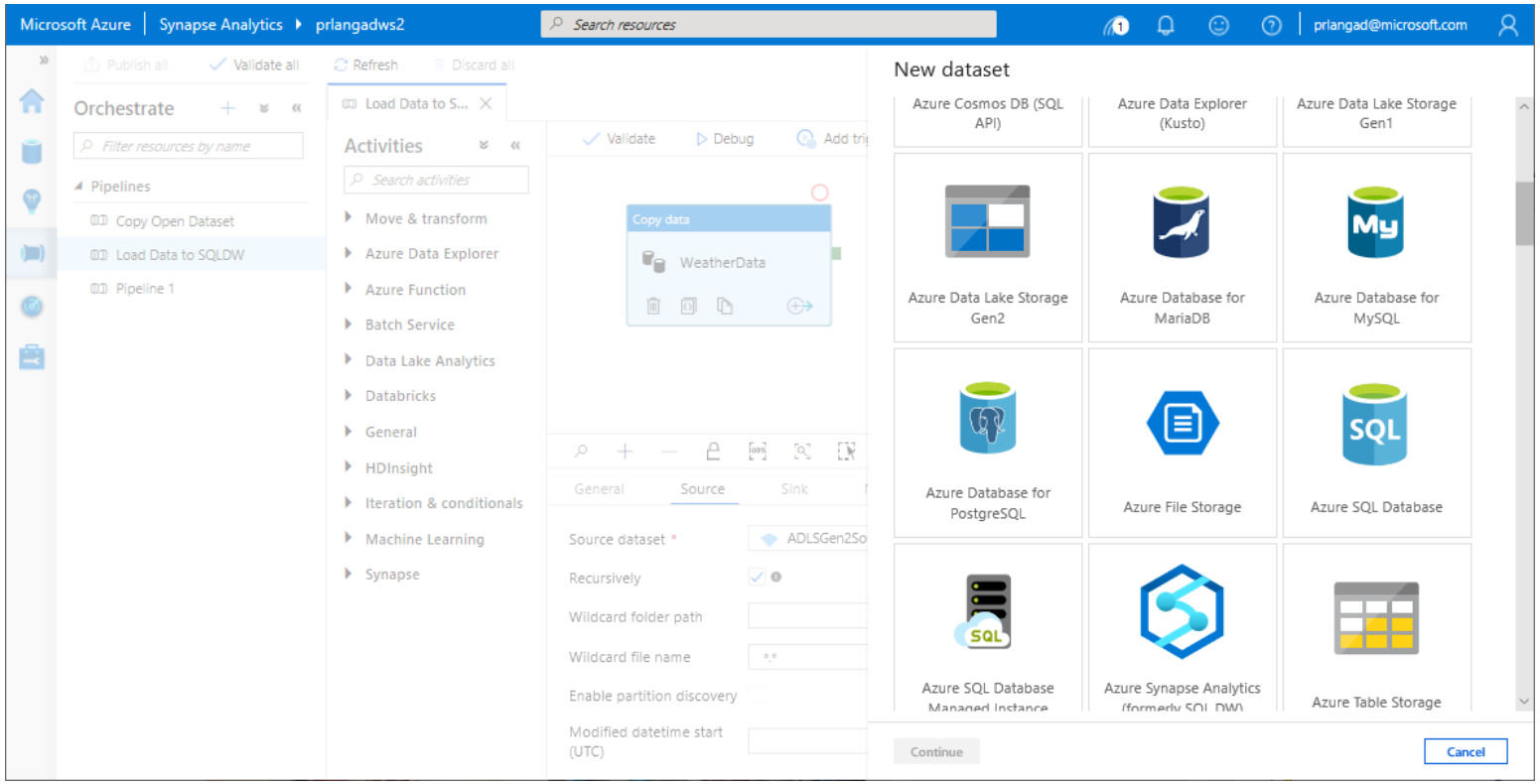
Overview

- Provide 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.

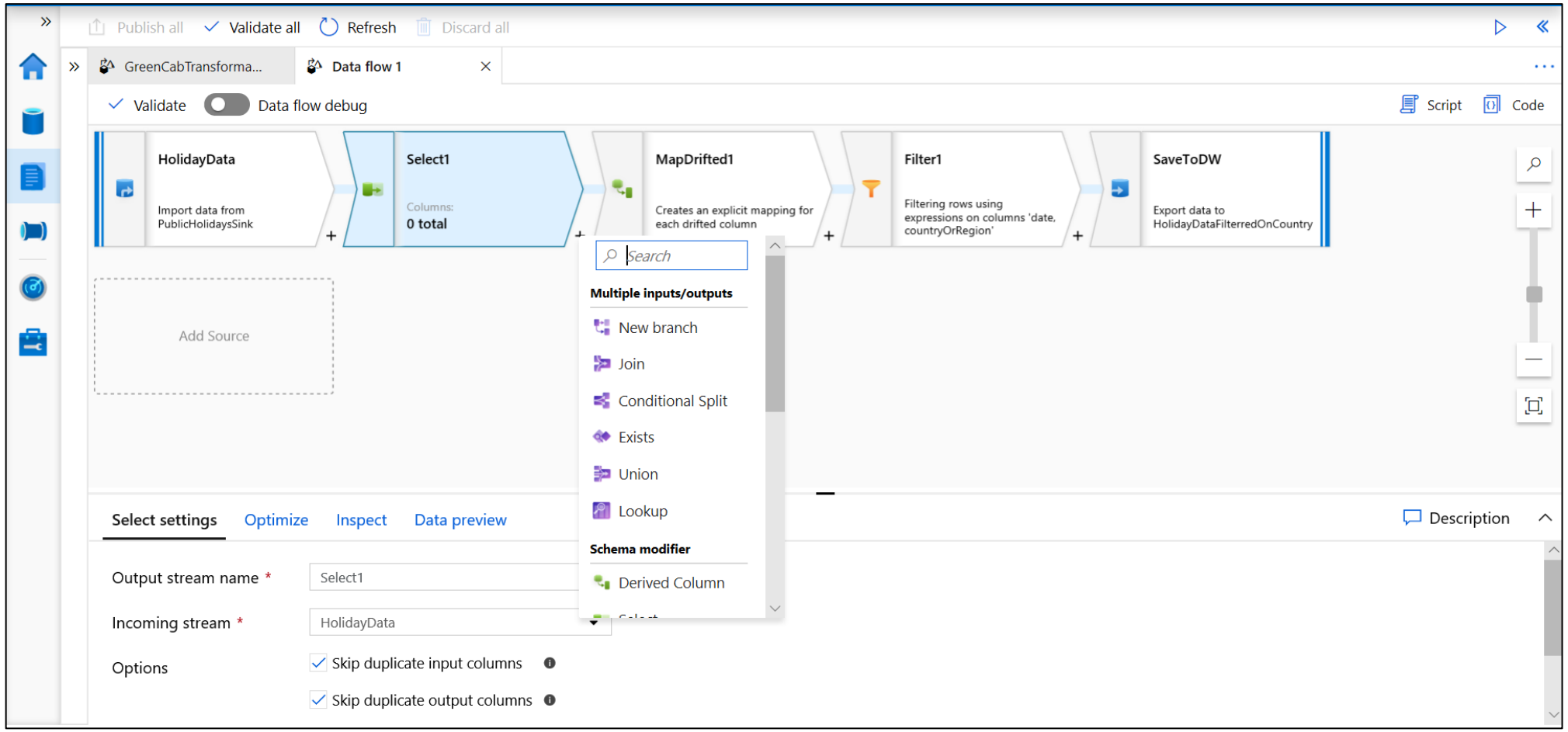




Develop Hub - Data Flows

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

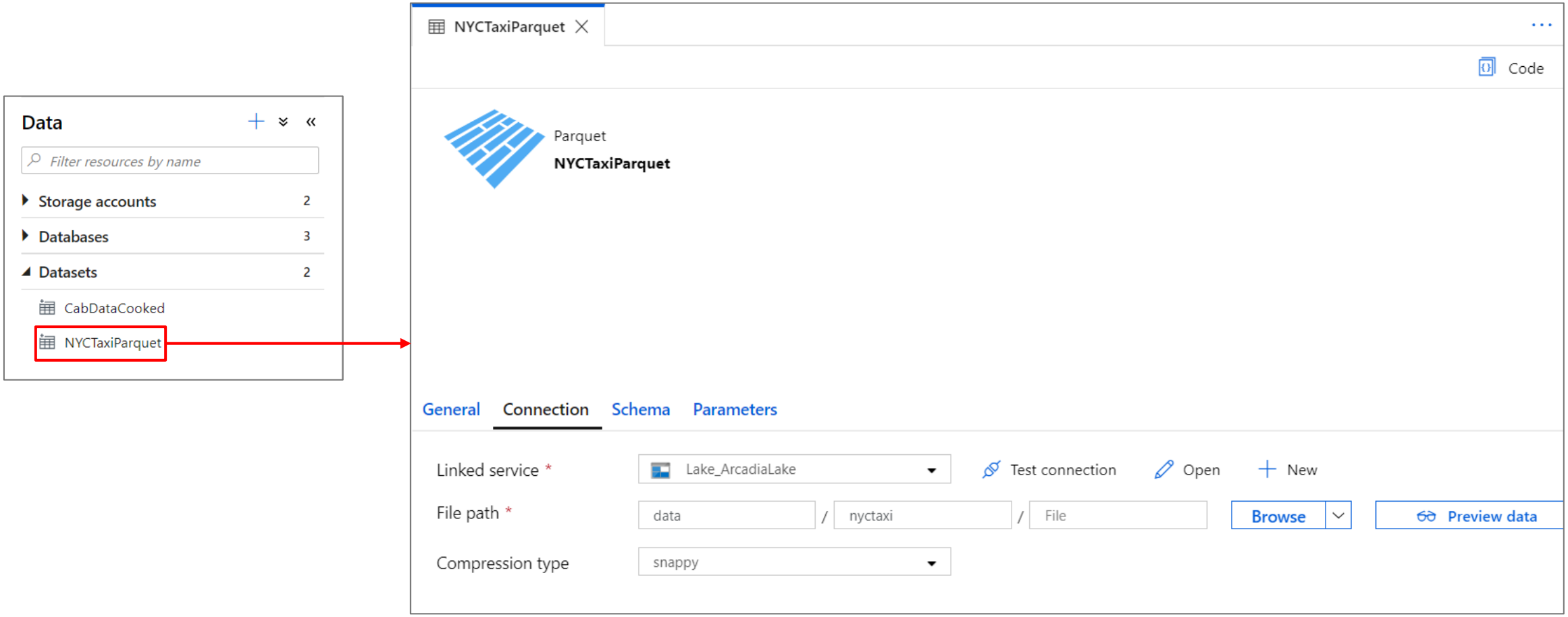
Provides a code-free experience.



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.



Azure Synapse Apache Spark - Summary

NephOSystems
Cloud Computing for life

Apache Spark 2.4 derivation

- Linux Foundation Delta Lake 0.6.1 support
- .Net Core 3.1 support
- Python 3.6 + Anaconda 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 SQL Analytics pools

Core scenarios

- Data Prep/Data Engineering/ETL
- Machine Learning via Spark ML and Azure ML integration
- Extensible through library management

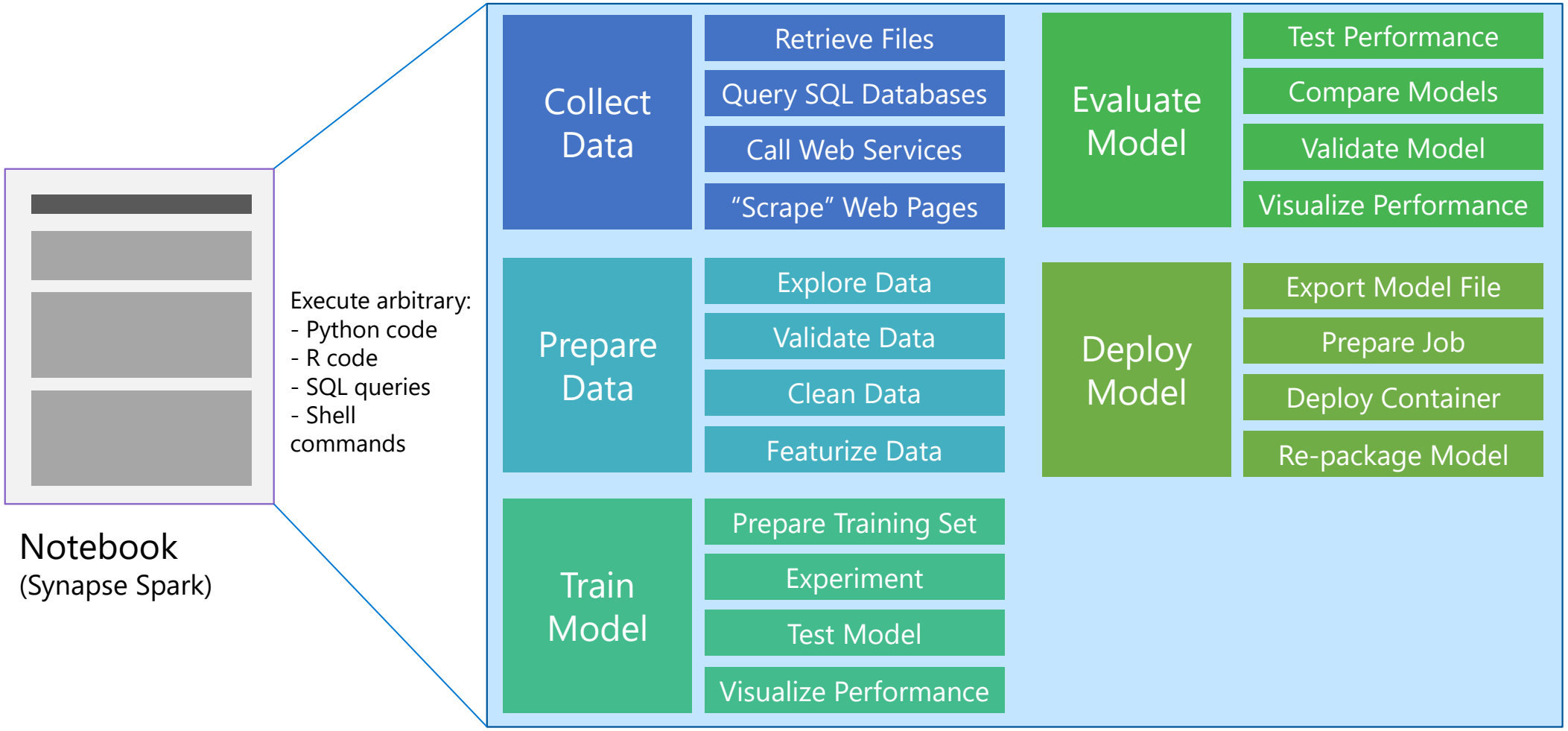
Efficient resource utilization

- Fast Start
- Auto scale (up and down)
- Auto pause
- Min cluster size of 3 nodes

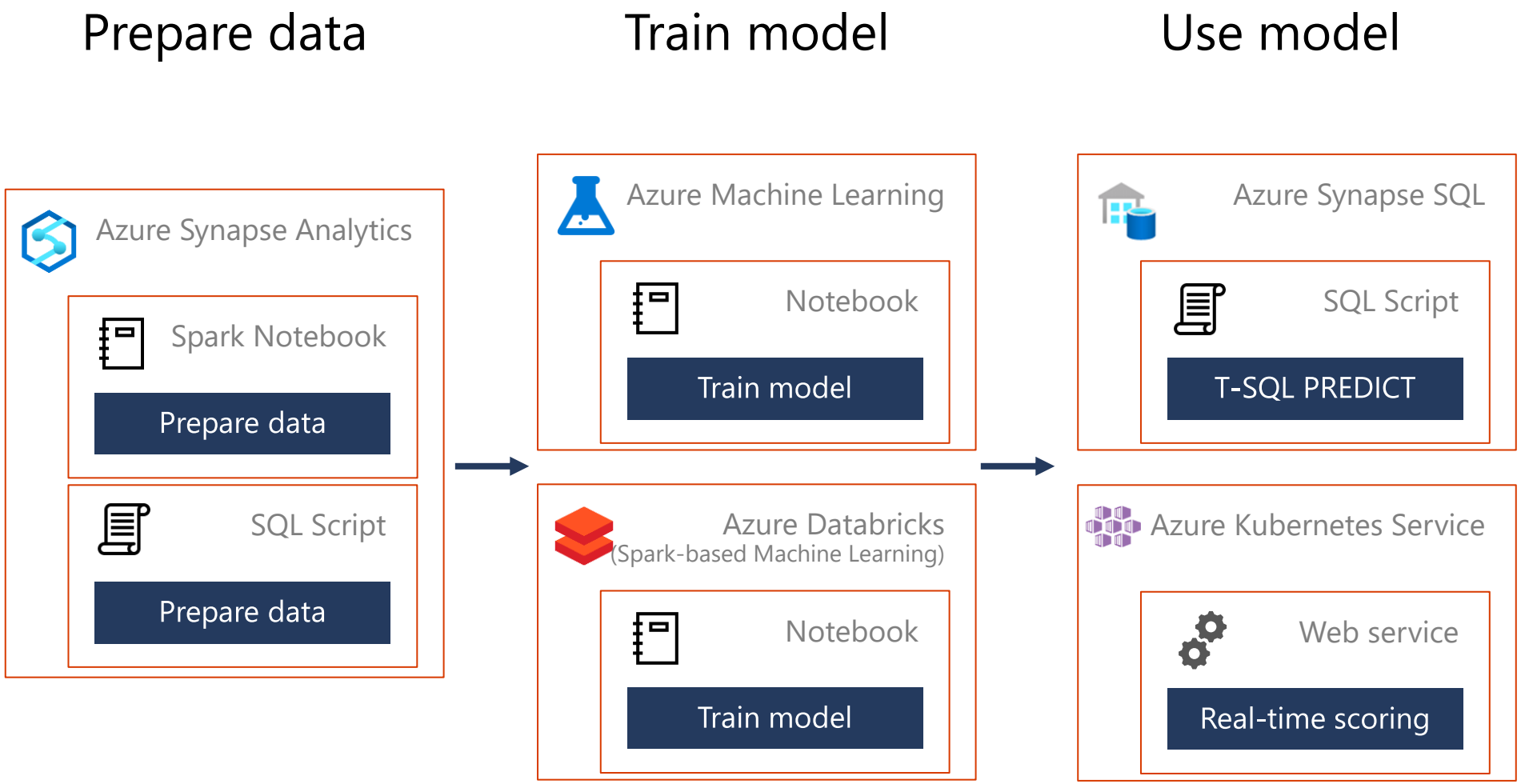
Multi Language Support

- .Net (C#), PySpark, Scala, Spark SQL, Java

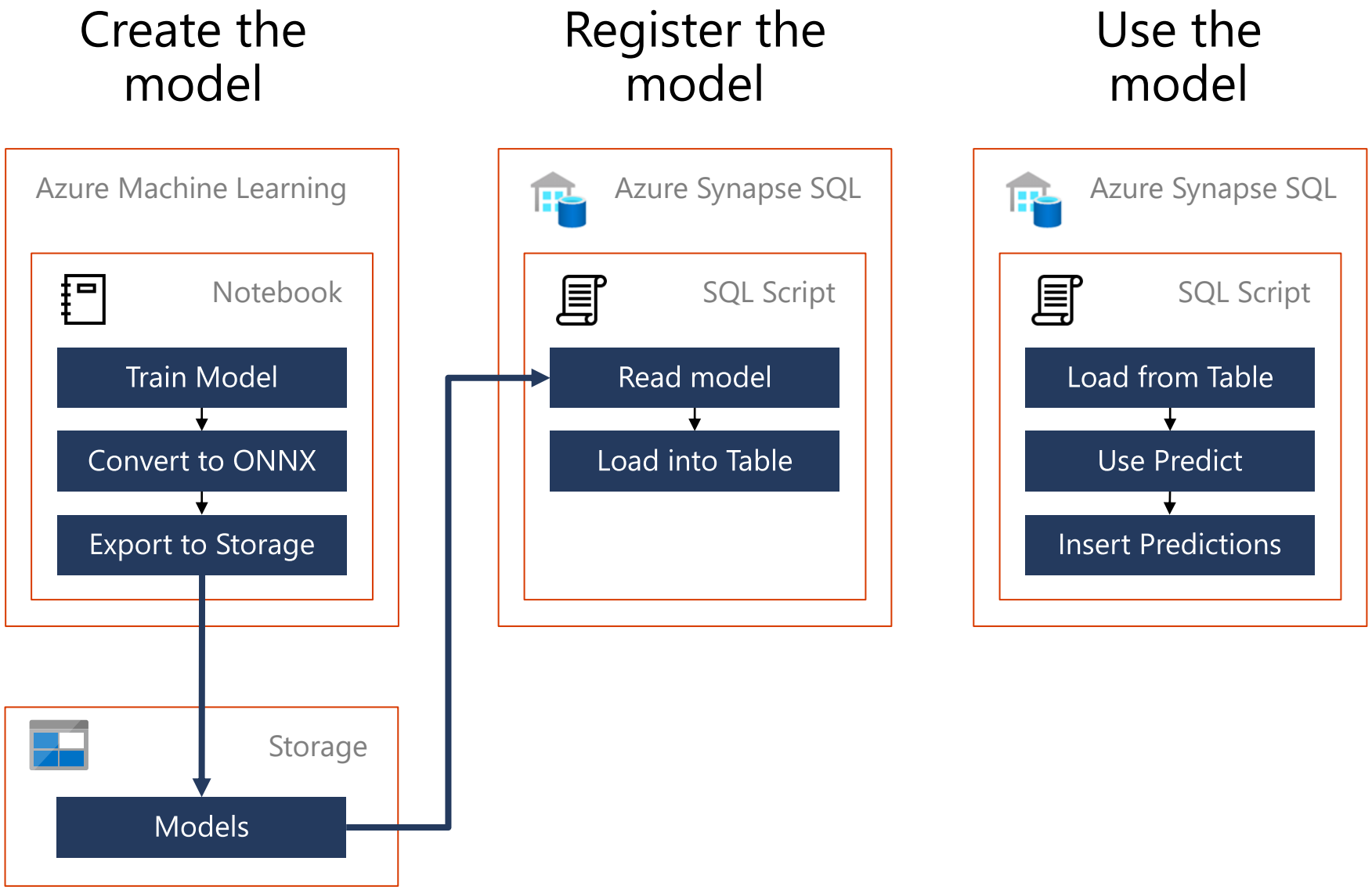
The Notebook Paradigm – one UI for data science



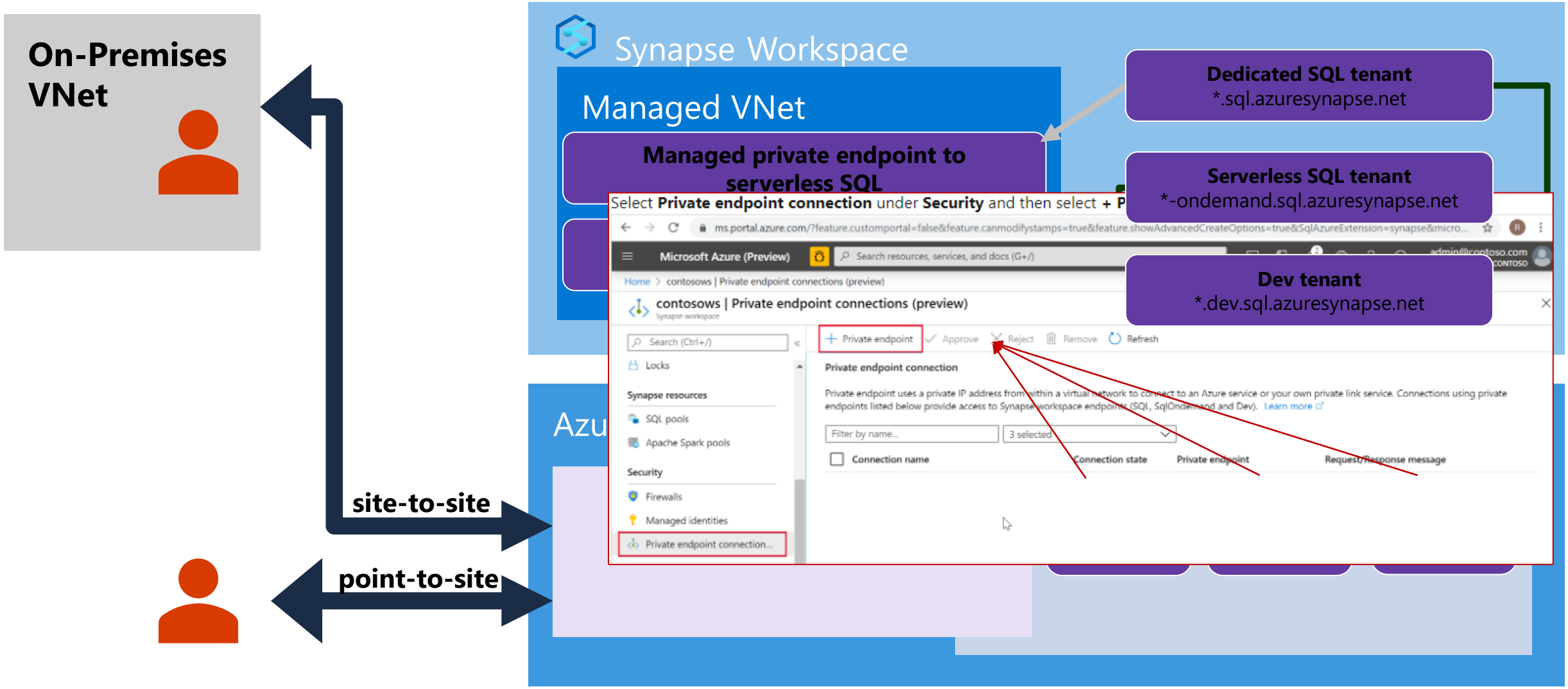
Synapse Analytics and the Machine Learning Process



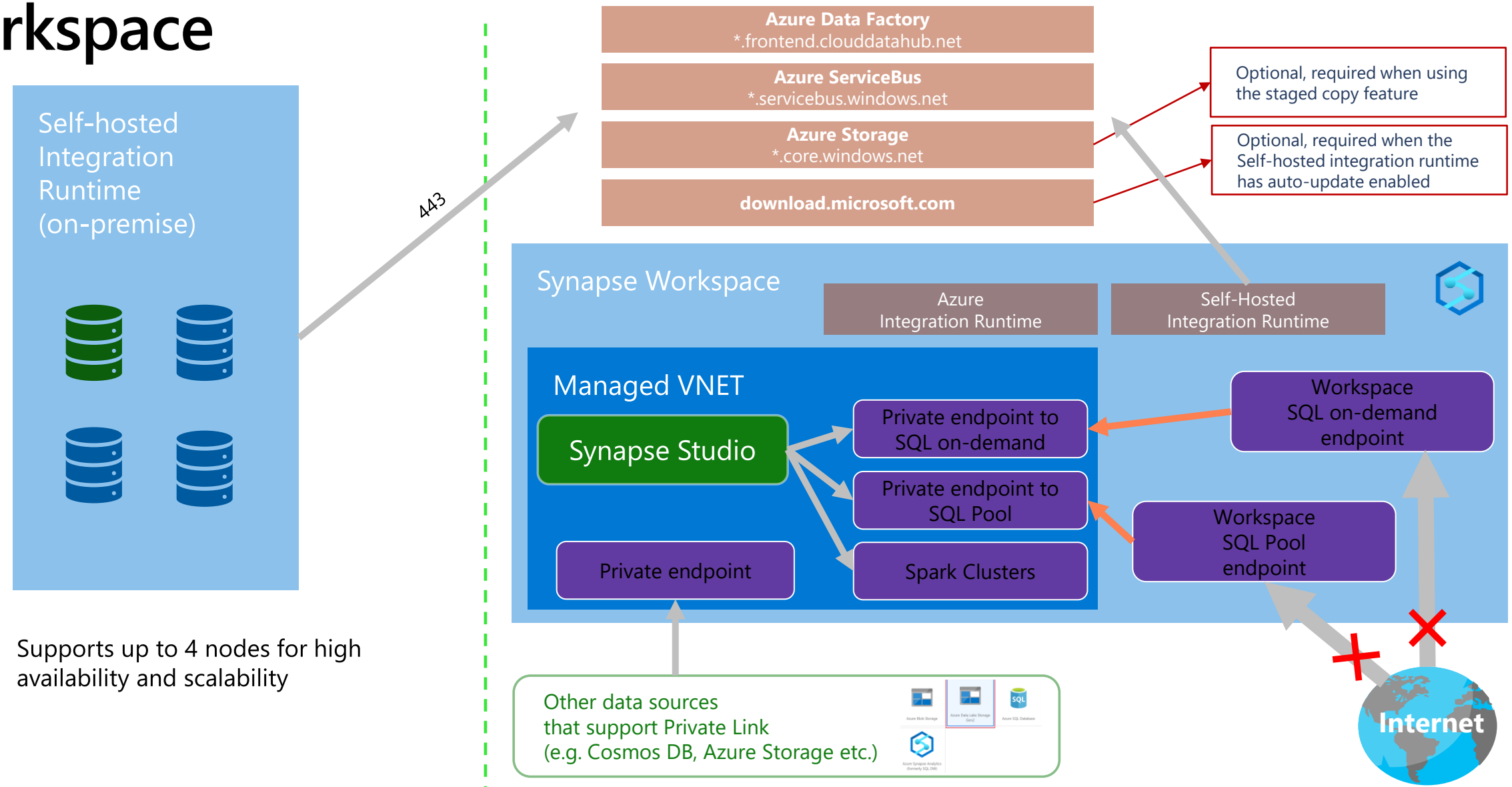
Making predictions with T-SQL



Connecting to Synapse from On-Premises



Self-hosted Integration Runtime in Synapse Workspace



Supports up to 4 nodes for high availability and scalability

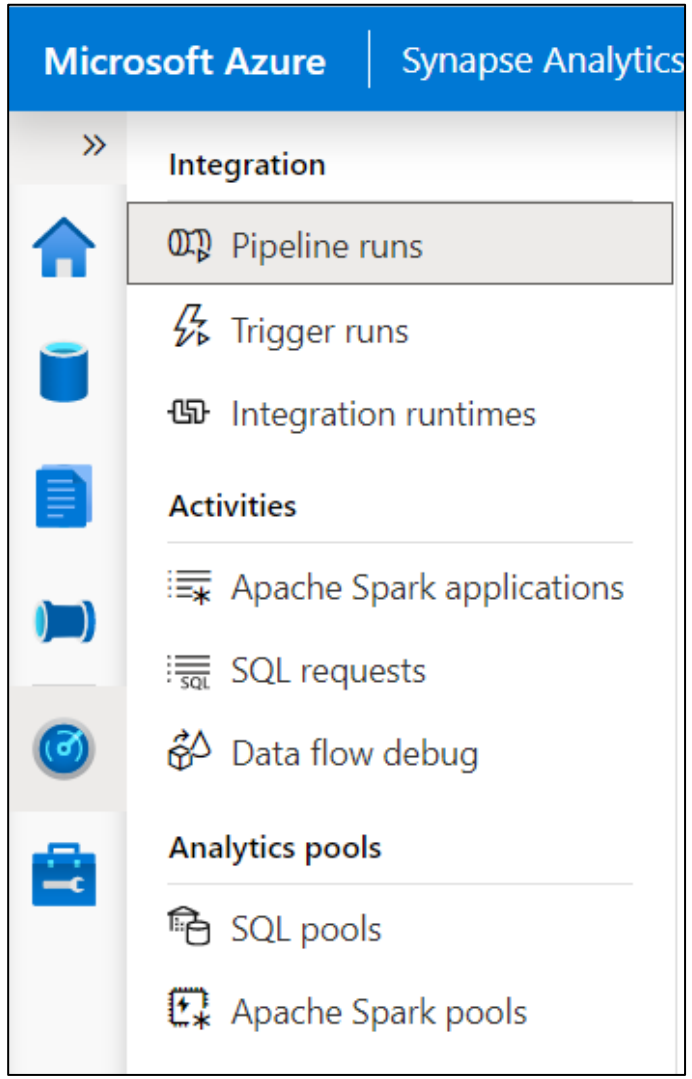
Monitor Hub

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 – SQL requests

NephOSystems
Cloud Computing for life

Overview

Monitor SQL requests for the progress and status of activities

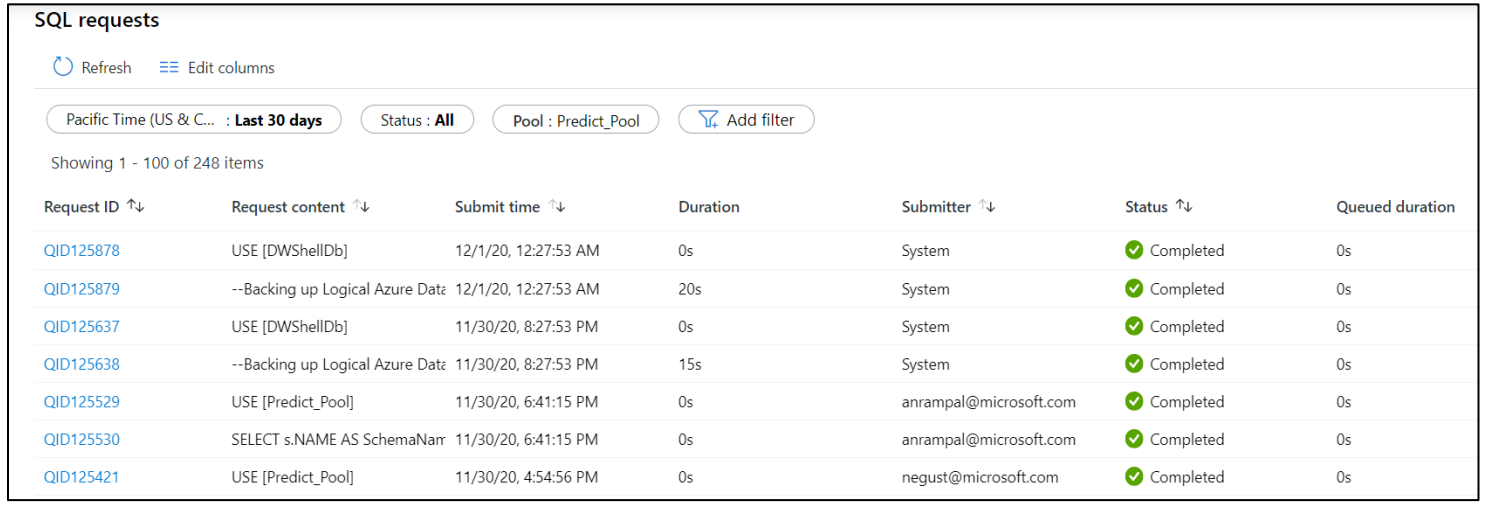
Benefits

Apply filter for pool to get SQL requests per compute pool

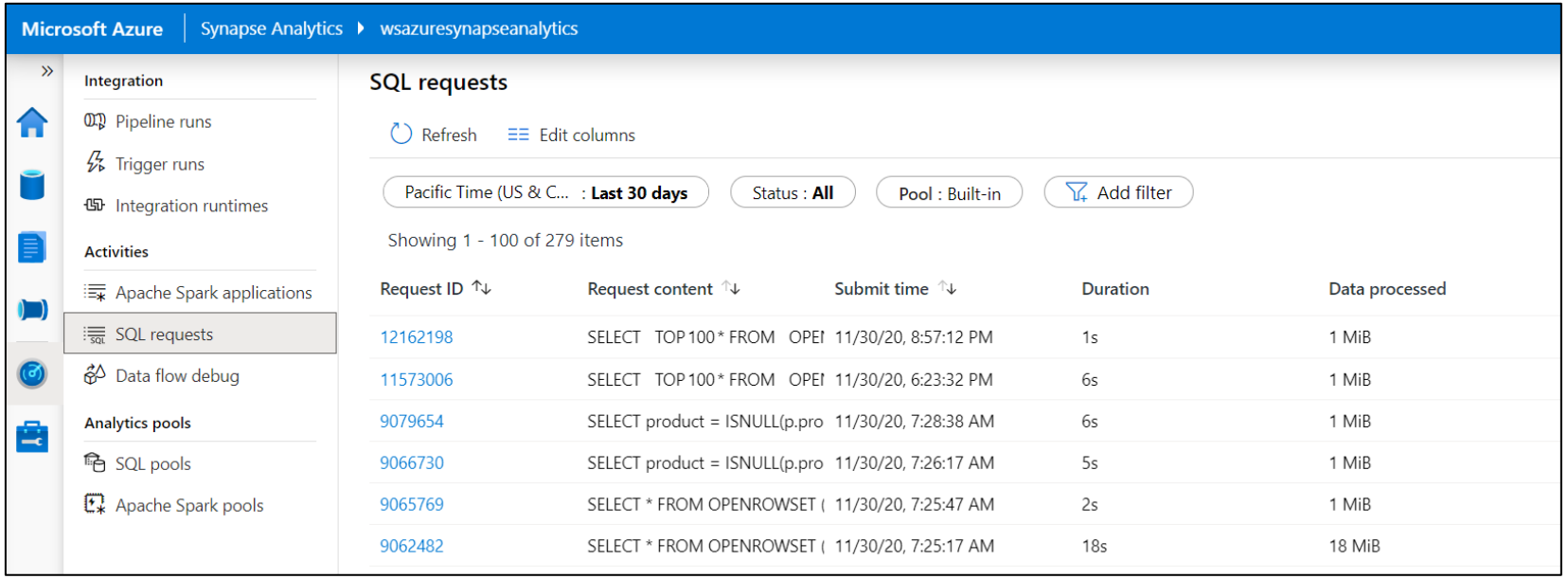
Validate query text

Additional available filters include

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



Request ID	Request content	Submit time	Duration	Submitter	Status	Queued duration
QID125878	USE [DWShellDb]	12/1/20, 12:27:53 AM	0s	System	Completed	0s
QID125879	--Backing up Logical Azure Data	12/1/20, 12:27:53 AM	20s	System	Completed	0s
QID125637	USE [DWShellDb]	11/30/20, 8:27:53 PM	0s	System	Completed	0s
QID125638	--Backing up Logical Azure Data	11/30/20, 8:27:53 PM	15s	System	Completed	0s
QID125529	USE [Predict_Pool]	11/30/20, 6:41:15 PM	0s	anrampal@microsoft.com	Completed	0s
QID125530	SELECT s.NAME AS SchemaNam	11/30/20, 6:41:15 PM	0s	anrampal@microsoft.com	Completed	0s
QID125421	USE [Predict_Pool]	11/30/20, 4:54:56 PM	0s	negust@microsoft.com	Completed	0s



Request ID	Request content	Submit time	Duration	Data processed
12162198	SELECT TOP 100 * FROM OPEI	11/30/20, 8:57:12 PM	1s	1 MiB
11573006	SELECT TOP 100 * FROM OPEI	11/30/20, 6:23:32 PM	6s	1 MiB
9079654	SELECT product = ISNULL(p.pro	11/30/20, 7:28:38 AM	6s	1 MiB
9066730	SELECT product = ISNULL(p.pro	11/30/20, 7:26:17 AM	5s	1 MiB
9065769	SELECT * FROM OPENROWSET (	11/30/20, 7:25:47 AM	2s	1 MiB
9062482	SELECT * FROM OPENROWSET (	11/30/20, 7:25:17 AM	18s	18 MiB

Workload Management

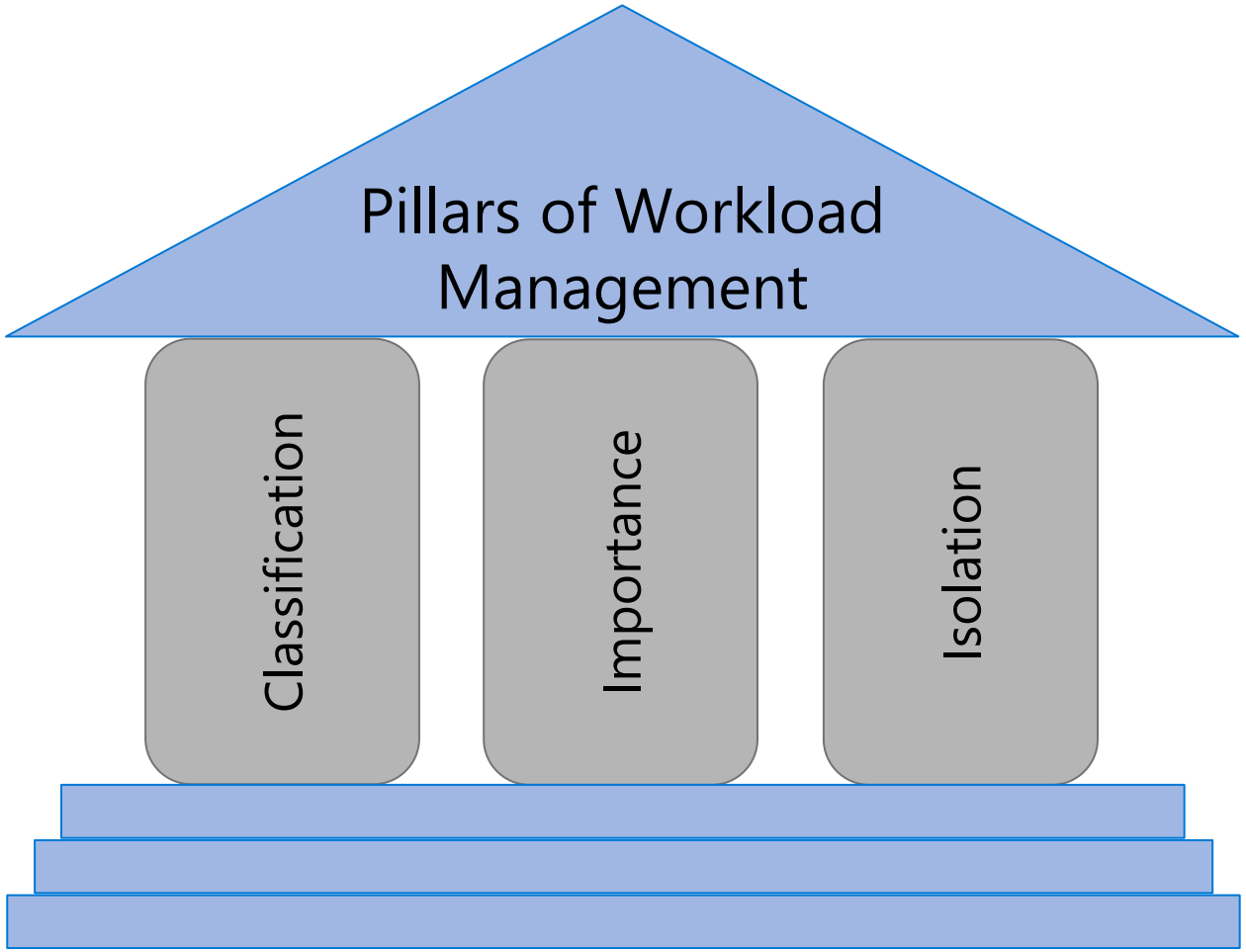
Overview

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

Synapse is moving away from Resource Class and Concurrency Slots to Workload Management.

The three pillars of workload management are

- Workload Classification – To assign a request to a workload group and setting importance levels.
- Workload Importance – To influence the order in which a request gets access to resources.
- 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

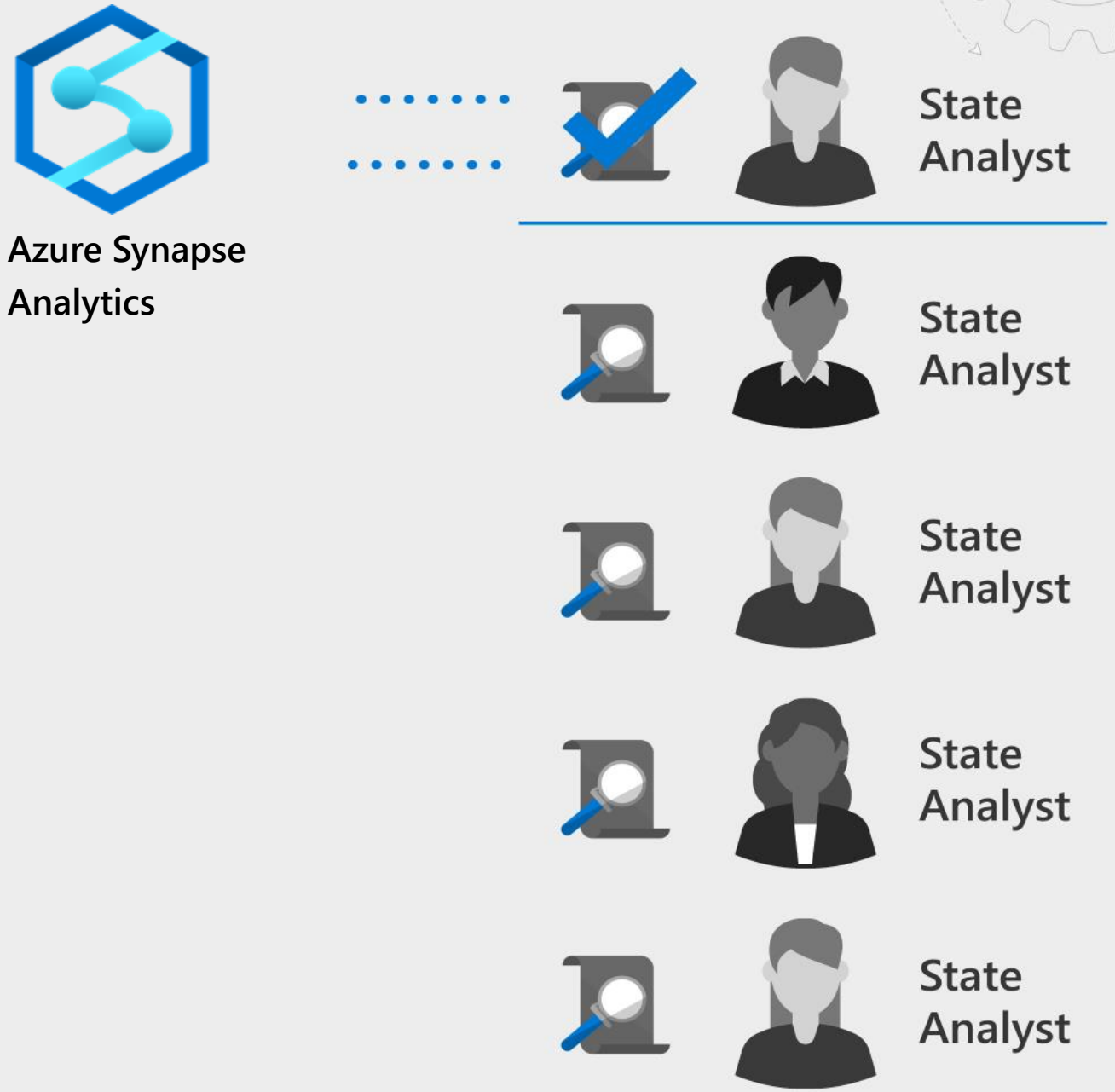
State analysts have normal importance.

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')
```



NephOSystems

Cloud Computing for life

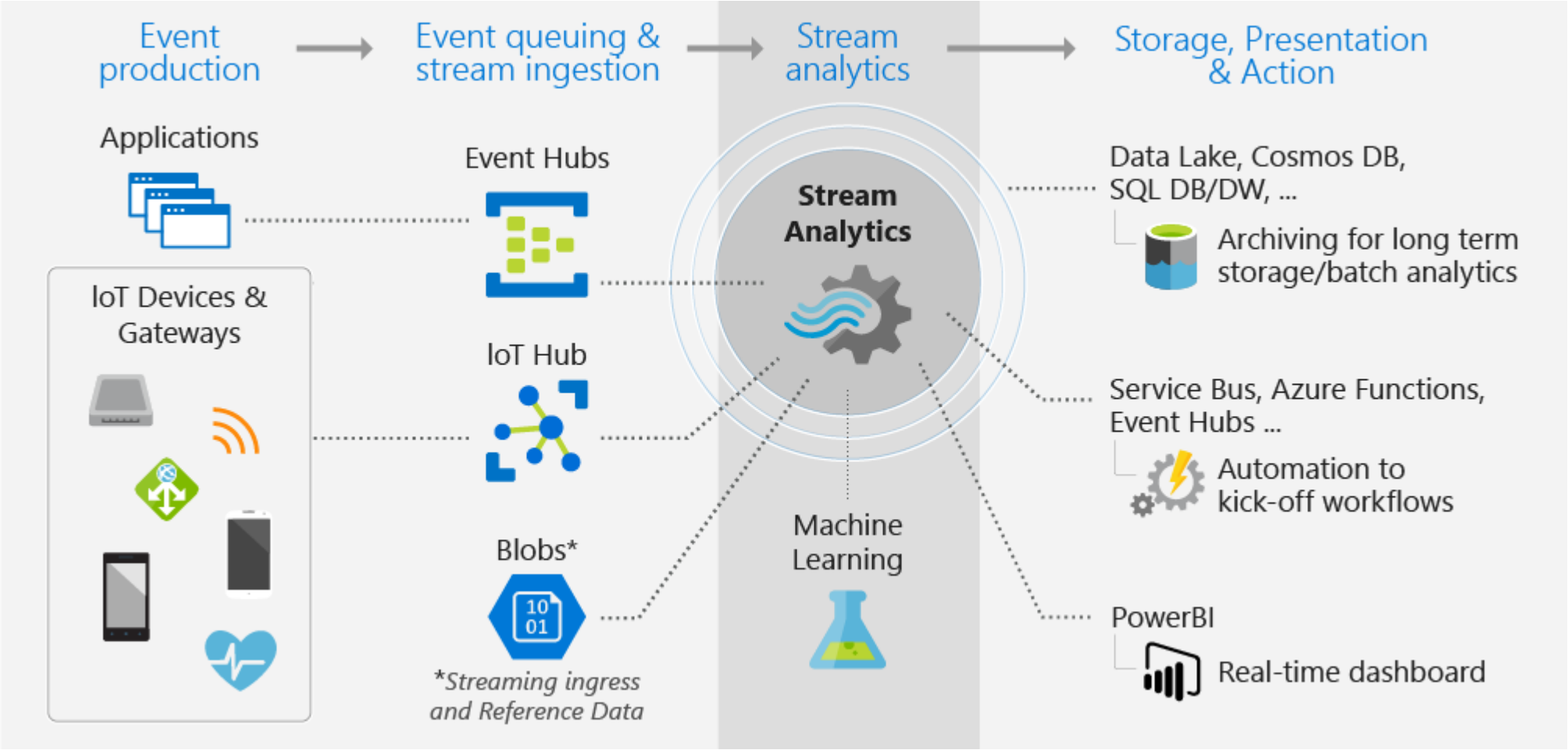
Azure Stream Analytics

Microsoft Azure

Stream Analytics

NephOSystems
Cloud Computing for life

Event-processing engine that allows you to examine high volumes of data streaming from devices



Retail Demo KPI Dashboard

Time Period: February 2017

NephOSystems
Cloud Computing for life

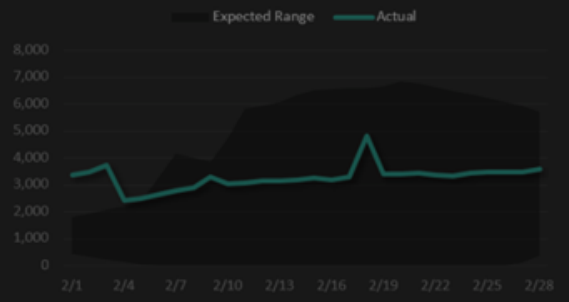
TOP SITE METRICS

Monthly Unique Visitors Trend

60,240

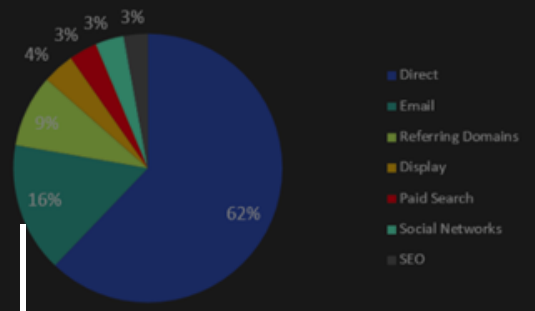
MOM 52.4%

Daily Unique Visitors Trend



MARKETING CHANNELS

Visits by Marketing Channel



Top External Marketing Channels

Channel	6 Mo. Trend	Visits	MOM	Revenue	CR
Direct		67,9	5	######	0.08%
Email		17,0	9%	9,52	3.94%
Referring Domains		54	25%	$60,545	[illegible]
Display		4,16	257%	$75,062	4.80%
Paid Search		3,69	267%	$47,684	5.38%
Social Networks		3,75	233%	$76,382	5.59%
SEO		3,17	277%	$54,797	4.76%

PRODUCTS ORDERS & REVENUE

Monthly Orders

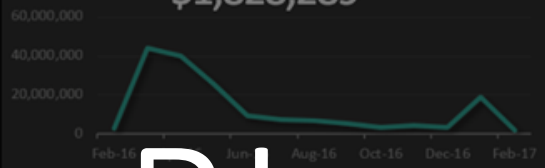
2,034

MOM 179%

Monthly Revenue

$1,828,289

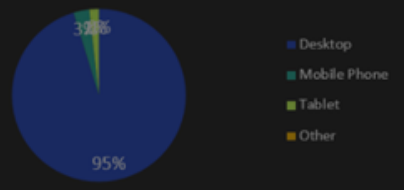
MOM -90%



Products	6 Mo. Trend	Orders	Revenue
Timberline GTX Boots		4	$195,838
Wasatch Xtreme Parka		3	$163,754
Uintas TechX Snow Pants		2	$76,199
Bonneville Shore Swimsuit		2	$46,383
Timpanogos Scarf		2	$10,214
La Sal Sweatshirt		2	$21,878
Uintas Pro Ski Gloves		1	$20,579
Uintas TechX Parka		1	$83,453
Amasa G2 Snow Goggles		1	$23,920

MOBILE DETAIL

Mobile Visits



Mobile OS

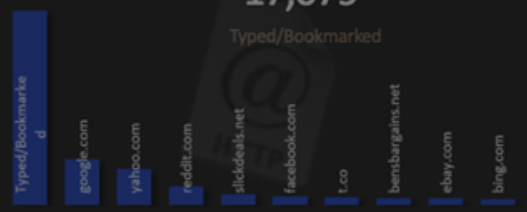
Device Name	Visits	MOM
Apple iPad	1,17	216%
Apple iPhone	6	219%
Samsung Galaxy S3 (GT-I930	2	451%
Samsung Galaxy S4 LTE (GT-I9505)	1	228%
Samsung Galaxy S2 Plus (GT-I9105P)	1	0%
Samsung Galaxy Tab 2 (GT-P5110)		131%
Google Nexus 7		21%
Samsung Galaxy Note 2 (GT-N7100)		1620%
Samsung Galaxy Tab 2 7.0 (GT-P3100)		240%

Top Referring Domains

MOM 170%

17,079

Typed/Bookmarked

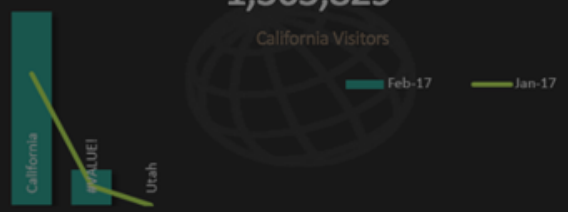


Geography

MOM 47%

1,365,829

California Visitors



Shopping Cart Fallout

	Feb-17 Visits	MOM	% Continued	% Fallout
Shopping Cart|Cart Details	0	0%	-	-
Shopping Cart|Shipping Information	0	0%	0%	100%
Shopping Cart|Billing Information	0	0%	0%	100%
Shopping Cart|Order Review	0	0%	0%	100%
Shopping Cart|Order Confirmation	0	0%	0%	100%
Overall Conversion			*0%*	*100%*

PowerBI

Instructions | Dashboard | Glossary

NephOSystems
Cloud Computing for life

Power BI

Power BI is a suite of business analytics tools to analyze data and share insights, with tools for business users to gain access to their most important metrics in a single location across all devices and platforms.