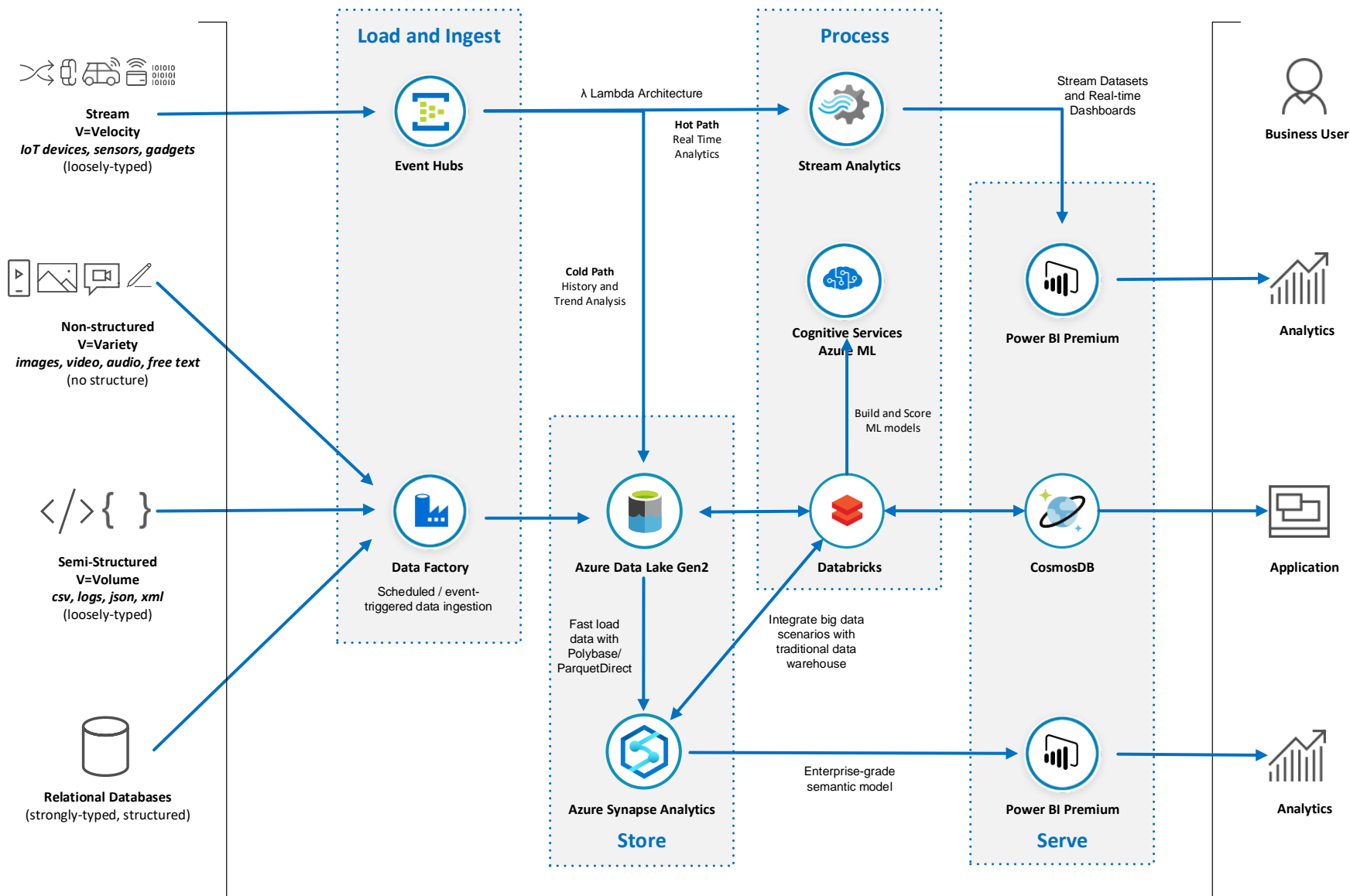


# Analytics on Azure

# Modern Data Platform Reference Architecture





Big Data and  
Analytics with  
Azure Data Lake

---

# 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



# Azure Analysis Services

```
for object to mirror...
mirror_mod.mirror_object

operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True

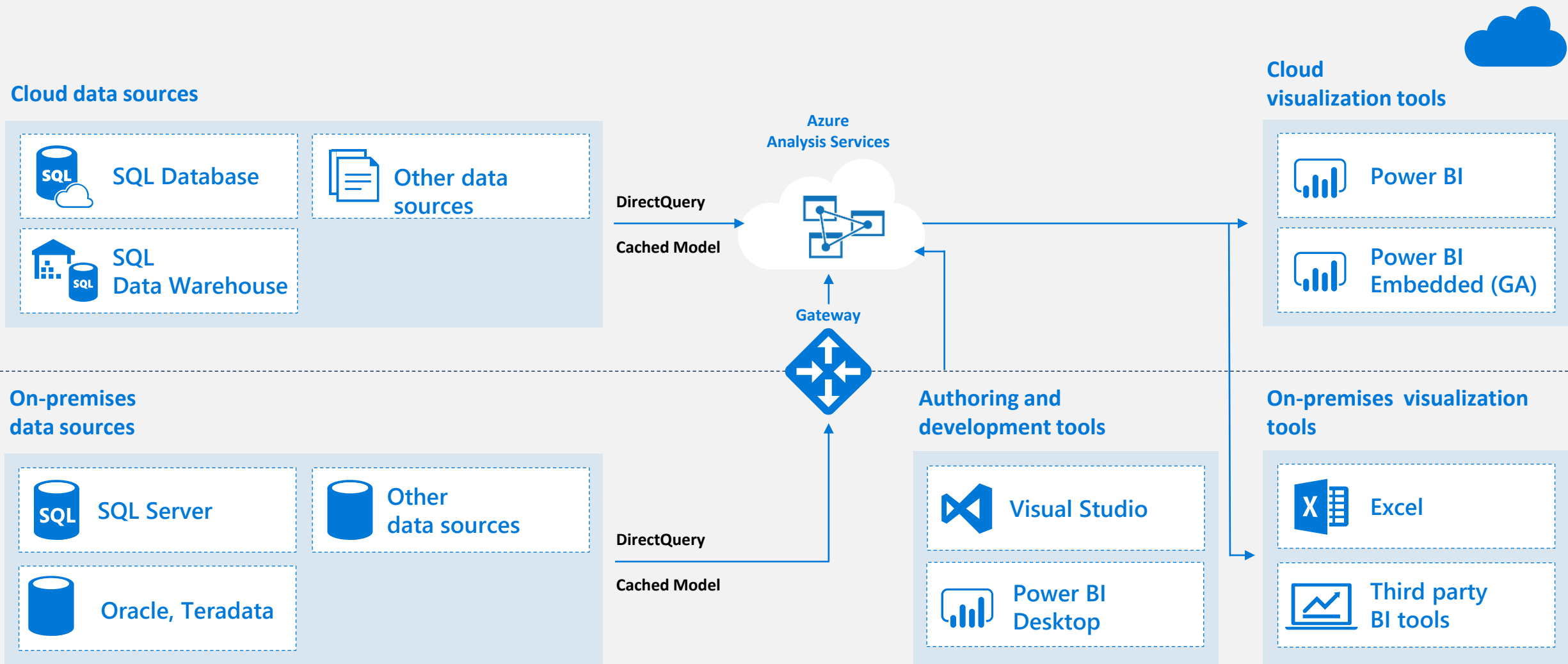
selection at the end -add
mirror_ob.select= 1
mirror_ob.select=1
context.scene.objects[one.name]
"selected" + str(modifier)
mirror_ob.select = 0
bpy.context.selected_objects
data.objects[one.name].select

print("please select exactly

-- OPERATOR CLASSES -----

types.Operator):
X mirror to the selected
object.mirror_mirror_x"
mirror X"
```

# Azure Analysis Services Architecture

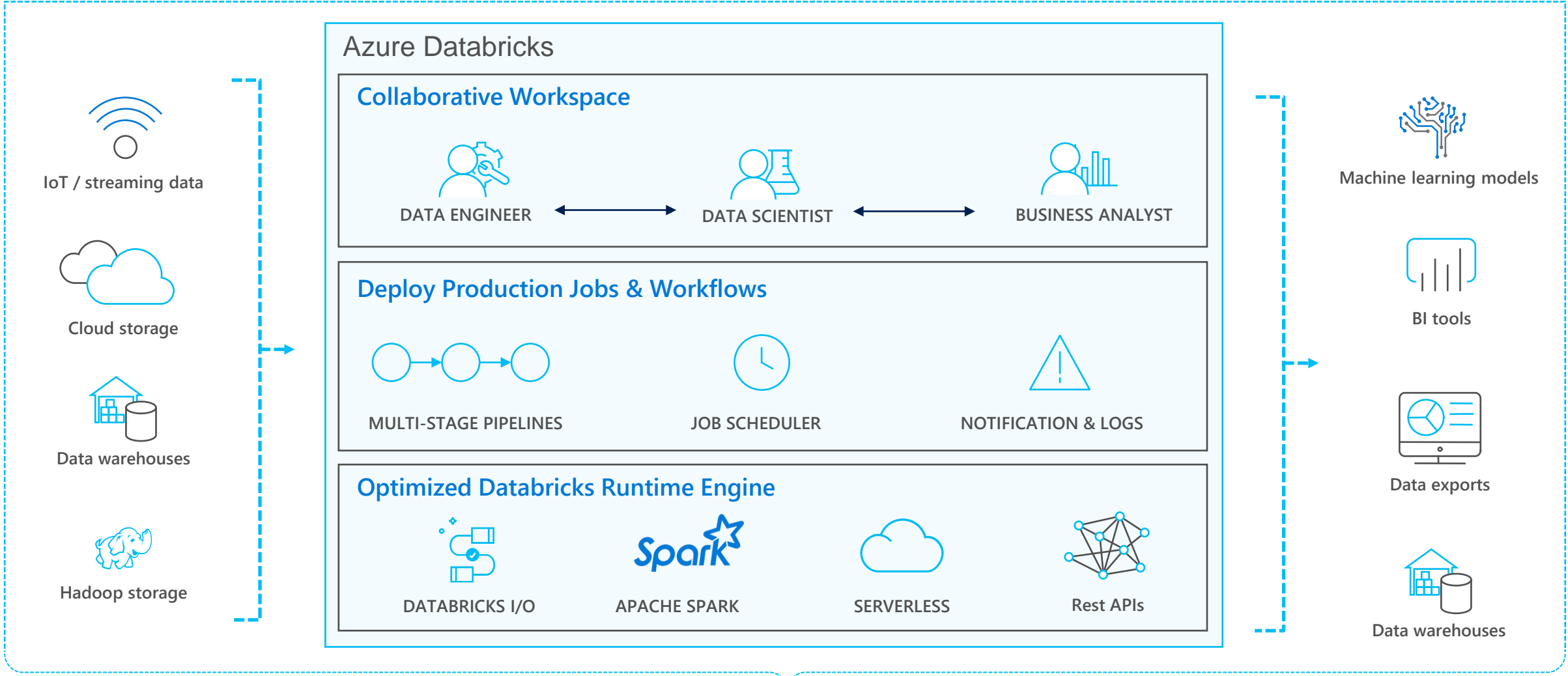


Note: not all capabilities available at public preview



# Azure Databricks

# Azure Databricks



Enhance Productivity

Build on secure & trusted cloud

Scale without limits



# Azure Synapse Analytics



Azure Synapse Analytics



Power BI



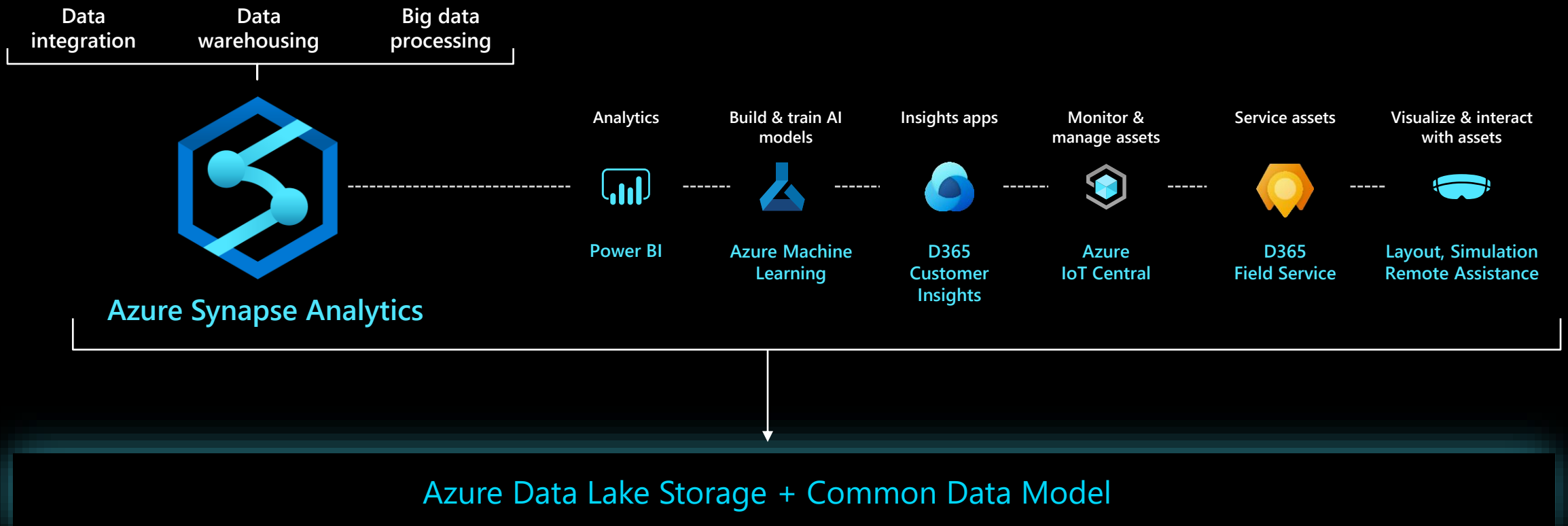
Azure Machine Learning



# Azure Synapse Analytics

Limitless analytics service with unmatched time to insight

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





# 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	<b>Languages</b> SQL Python .NET Java Scala R
SECURITY	<b>Form Factors</b> PROVISIONED ON-DEMAND
MONITORING	<b>Analytics Runtimes</b> SQL 
METASTORE	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

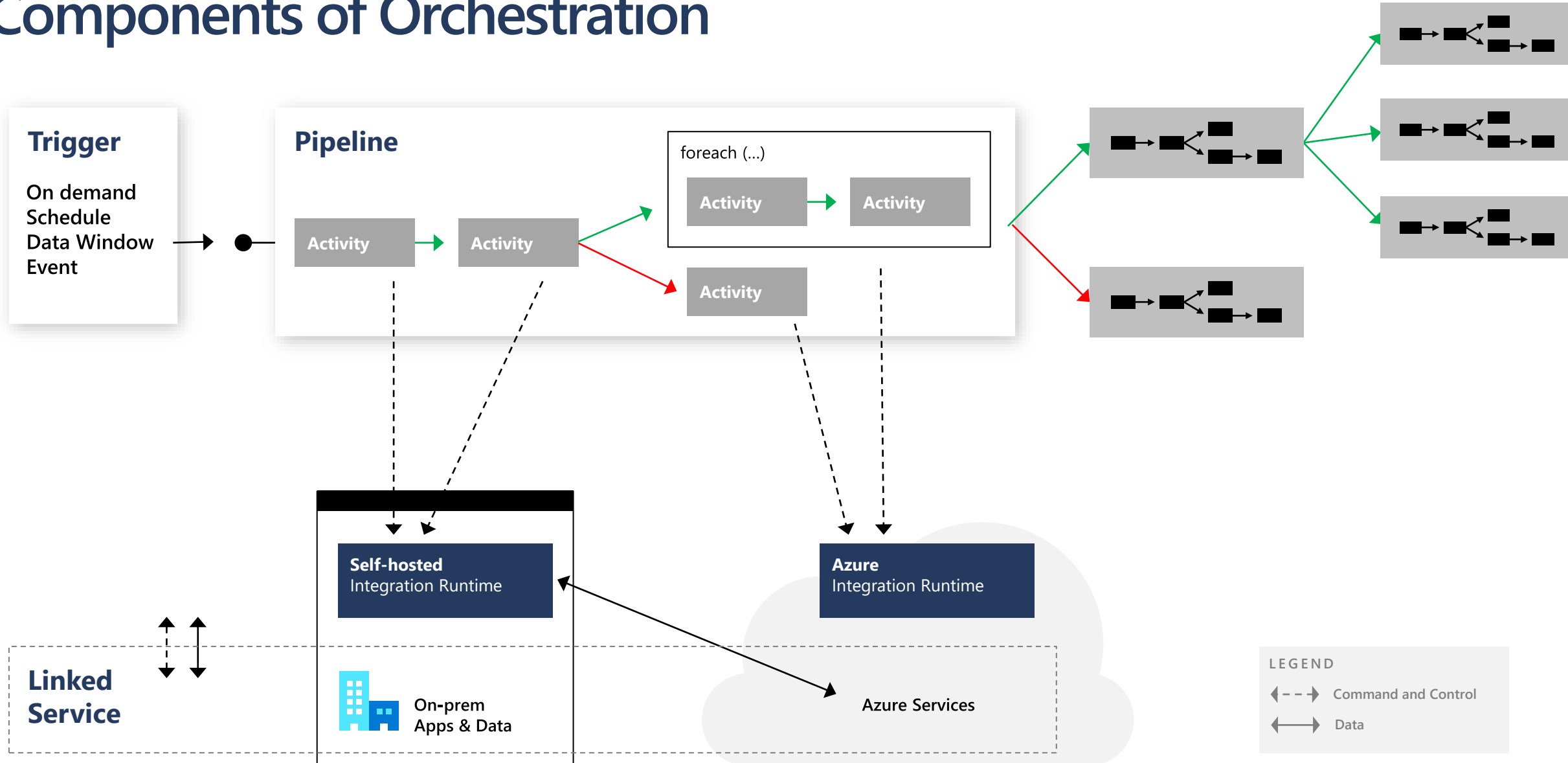
**SQL Analytics** offering T-SQL for batch, streaming and interactive processing

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

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

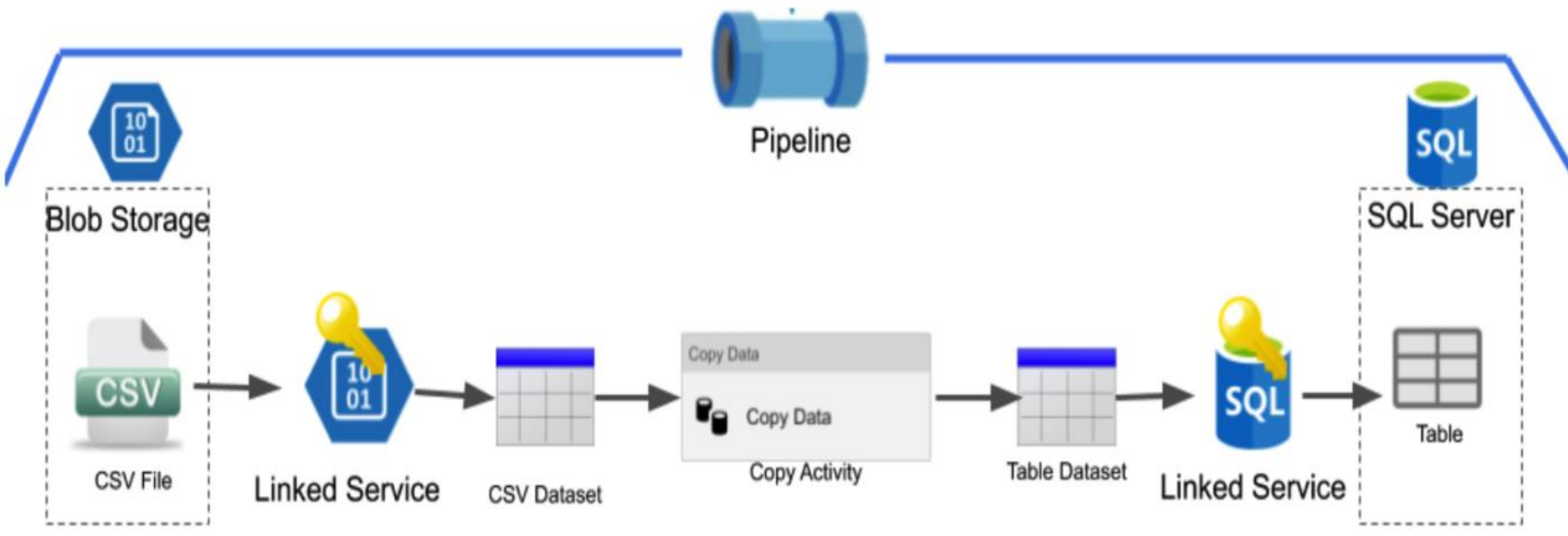
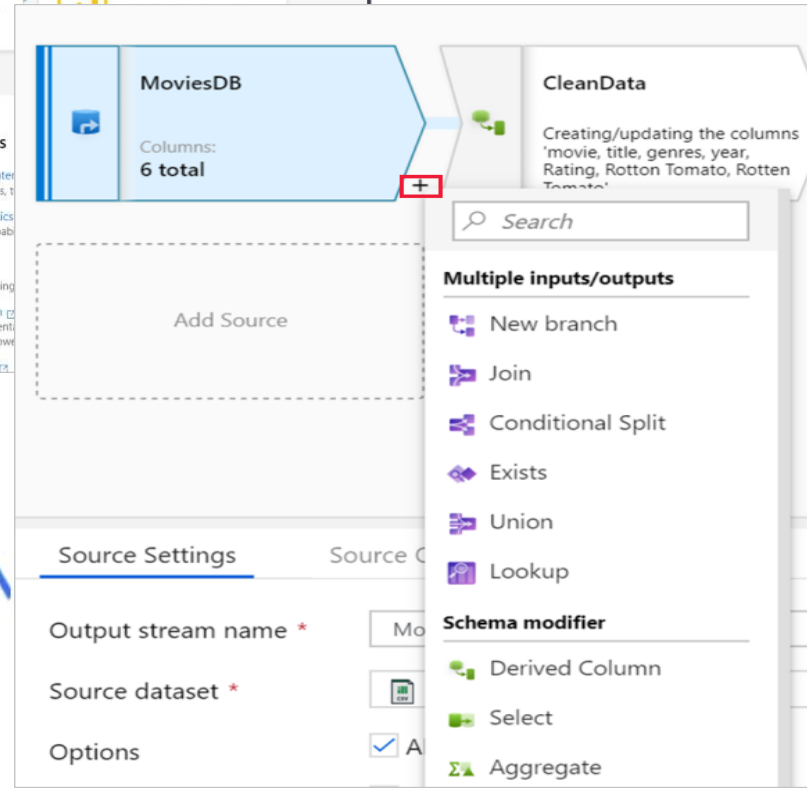
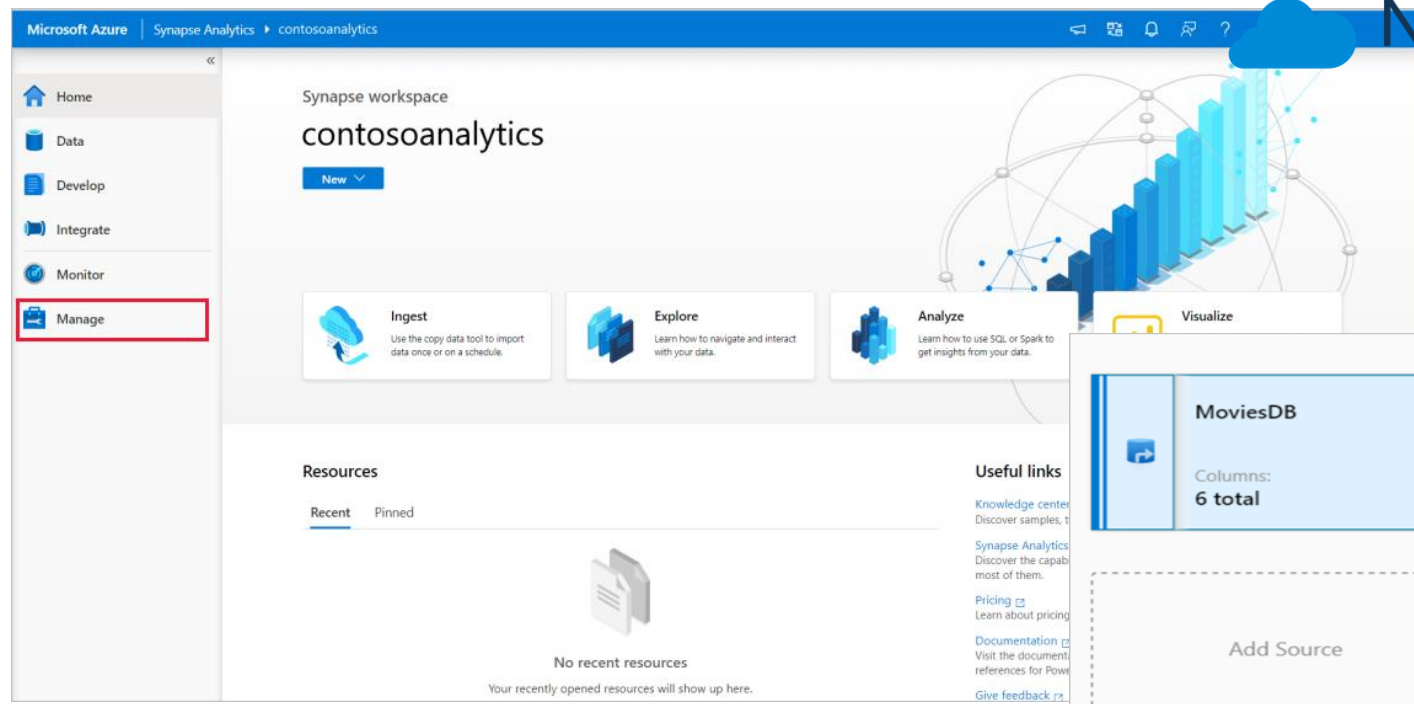
Data **lake integrated** and Common Data Model aware

# Components of Orchestration



Synapse Pipelines shares codebase with Azure Data Factory

# Synapse Workspace And Pipeline



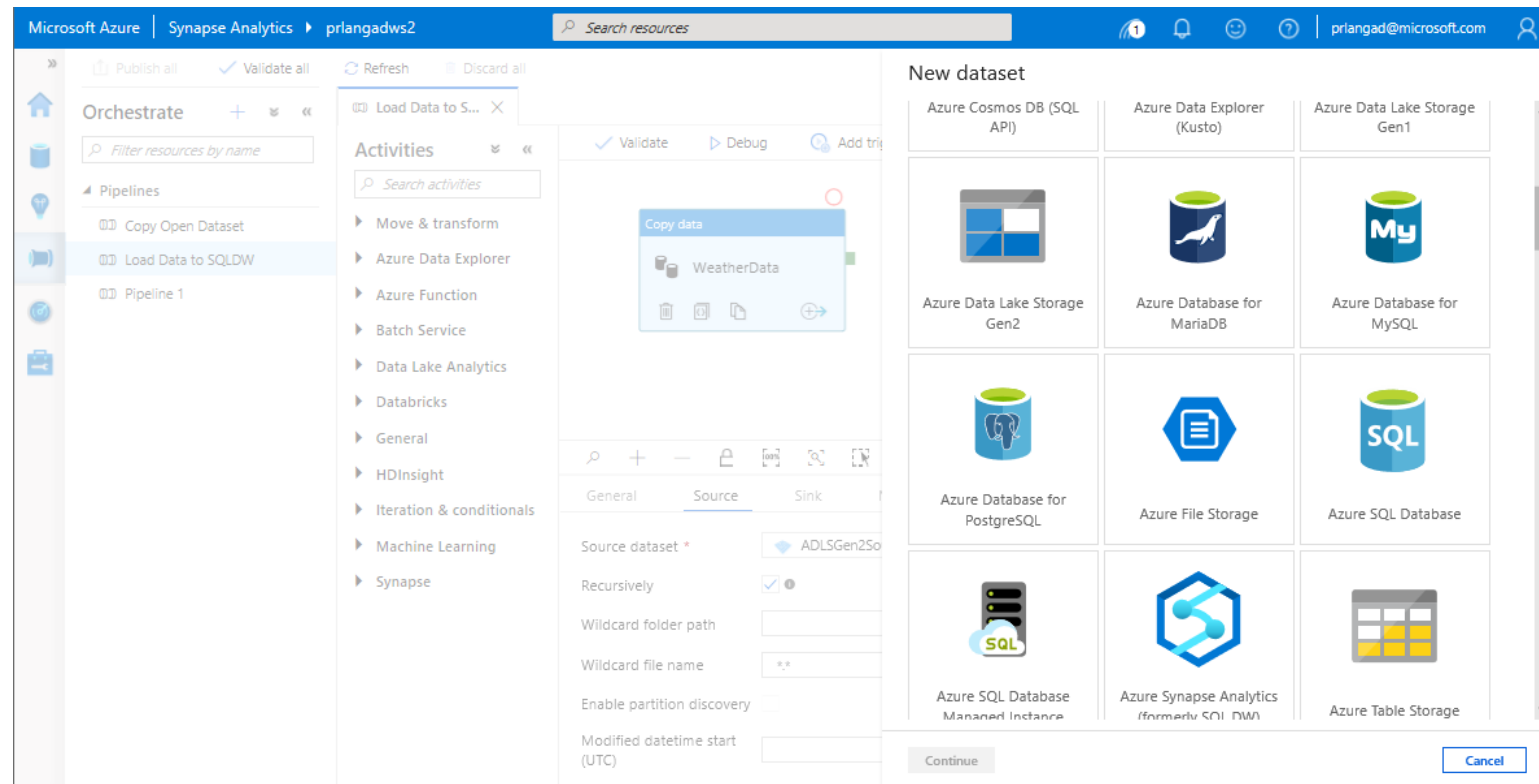
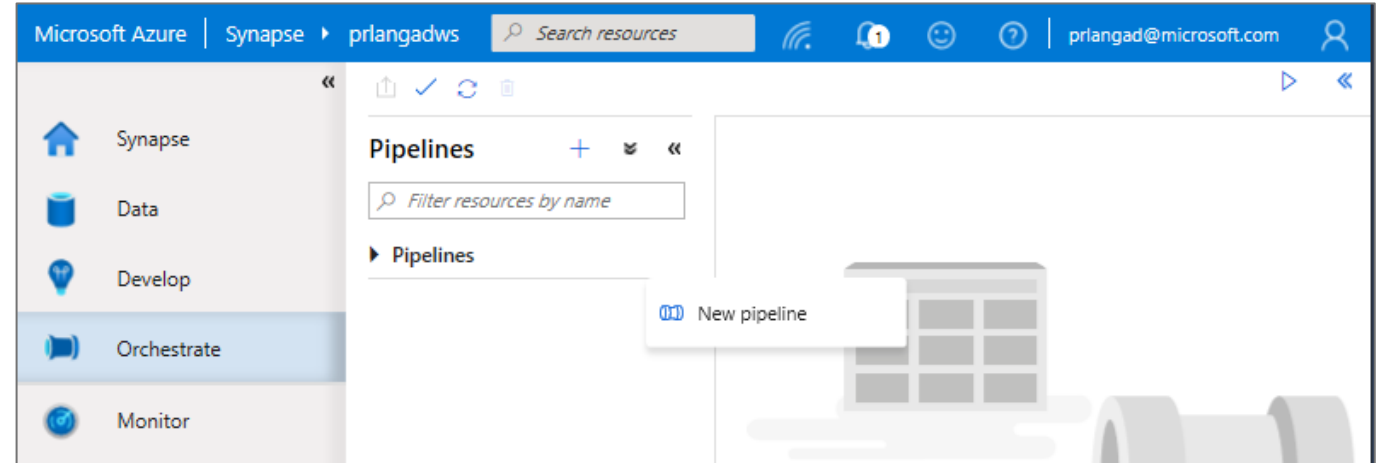
# Pipelines

## Overview

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

## Benefits

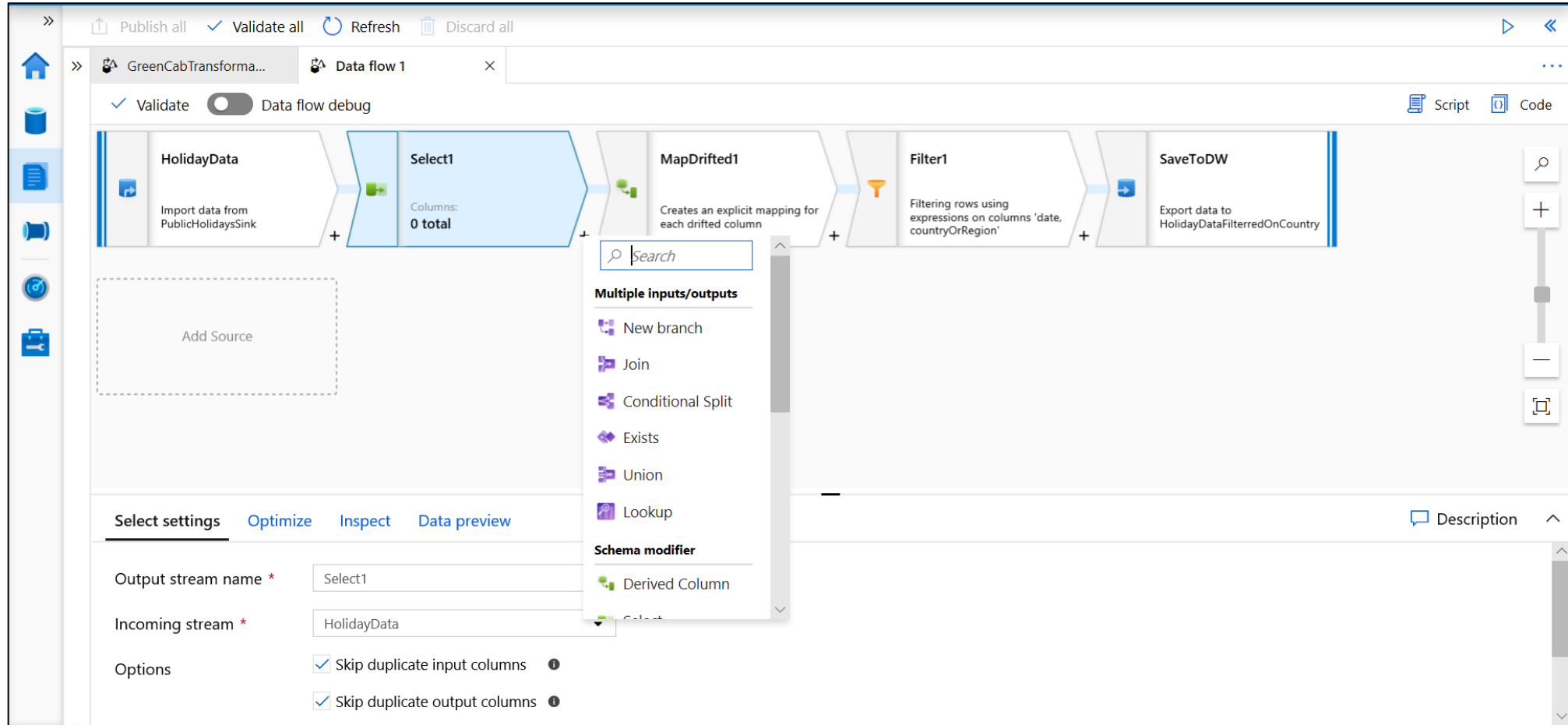
- Supports common loading patterns.
- Fully parallel loading into data lake or SQL tables.
- Graphical development experience.



# Develop Hub - Data Flows

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

Provides a code-free experience.

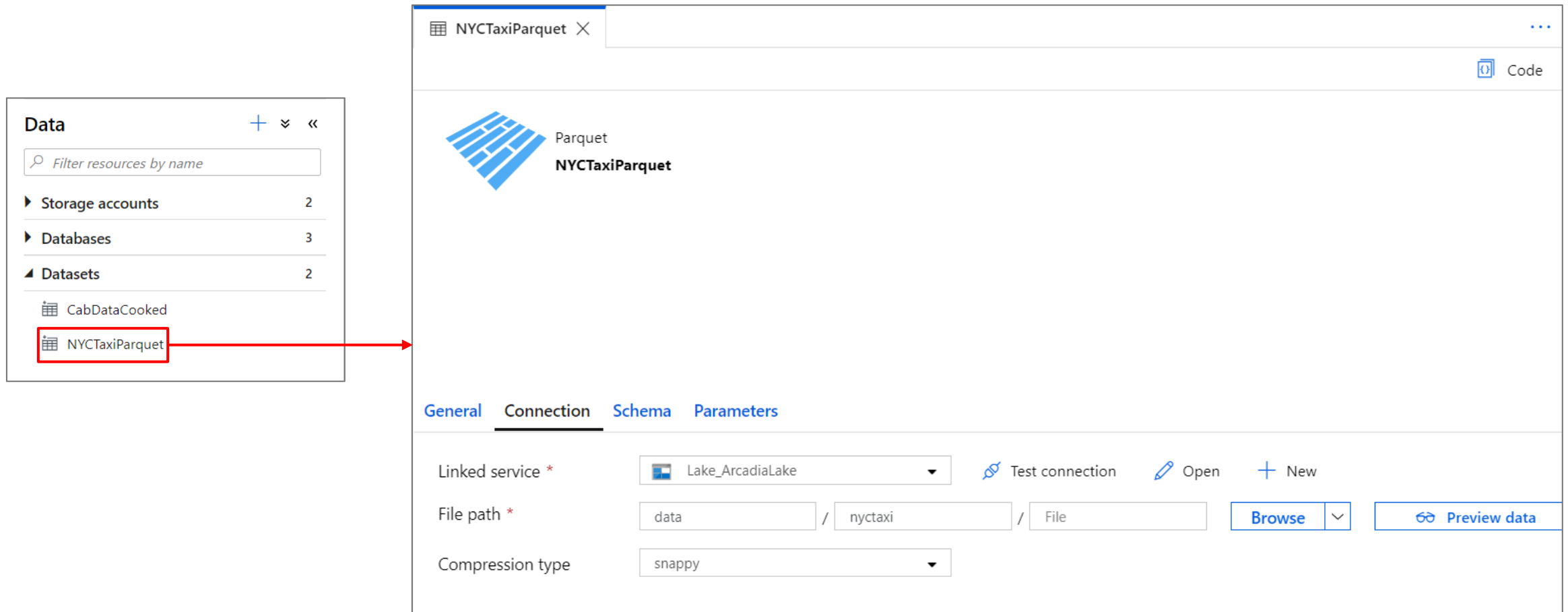


The screenshot displays the NephOSystems Data Flow Designer interface. At the top, there are action buttons: Publish all, Validate all, Refresh, and Discard all. Below this, the current data flow is named "Data flow 1" and is in a "Validate" state. The main workspace shows a sequence of data flow components: "HolidayData" (Import data from PublicHolidaysSink), "Select1" (Columns: 0 total), "MapDrifted1" (Creates an explicit mapping for each drifted column), "Filter1" (Filtering rows using expressions on columns 'date, countryOrRegion'), and "SaveToDW" (Export data to HolidayDataFilteredOnCountry). A search bar and a menu are open over the "Select1" component. The menu includes options for "Multiple inputs/outputs" (New branch, Join, Conditional Split, Exists, Union, Lookup) and "Schema modifier" (Derived Column). Below the workspace, there are tabs for "Select settings", "Optimize", "Inspect", and "Data preview". The "Select settings" tab is active, showing fields for "Output stream name" (Select1), "Incoming stream" (HolidayData), and "Options" (Skip duplicate input columns, Skip duplicate output columns).

# 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.



The screenshot displays the NephOSystems interface for configuring a dataset. On the left, a sidebar titled "Data" shows a list of resources: "Storage accounts" (2), "Databases" (3), and "Datasets" (2). Under "Datasets", "CabDataCooked" and "NYCTaxiParquet" are listed. A red box highlights "NYCTaxiParquet", with a red arrow pointing to the main configuration panel.

The main panel shows the configuration for the "NYCTaxiParquet" dataset. It is identified as a "Parquet" dataset. The configuration is divided into four tabs: "General", "Connection", "Schema", and "Parameters". The "Connection" tab is active, showing the following settings:

- Linked service: Lake\_ArcadiaLake
- File path: data / nyctaxi / File
- Compression type: snappy

Additional controls include "Test connection", "Open", and "New" buttons. A "Browse" button is present next to the file path, and a "Preview data" button is located at the bottom right.



# Azure Synapse Apache Spark - Summary

## 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

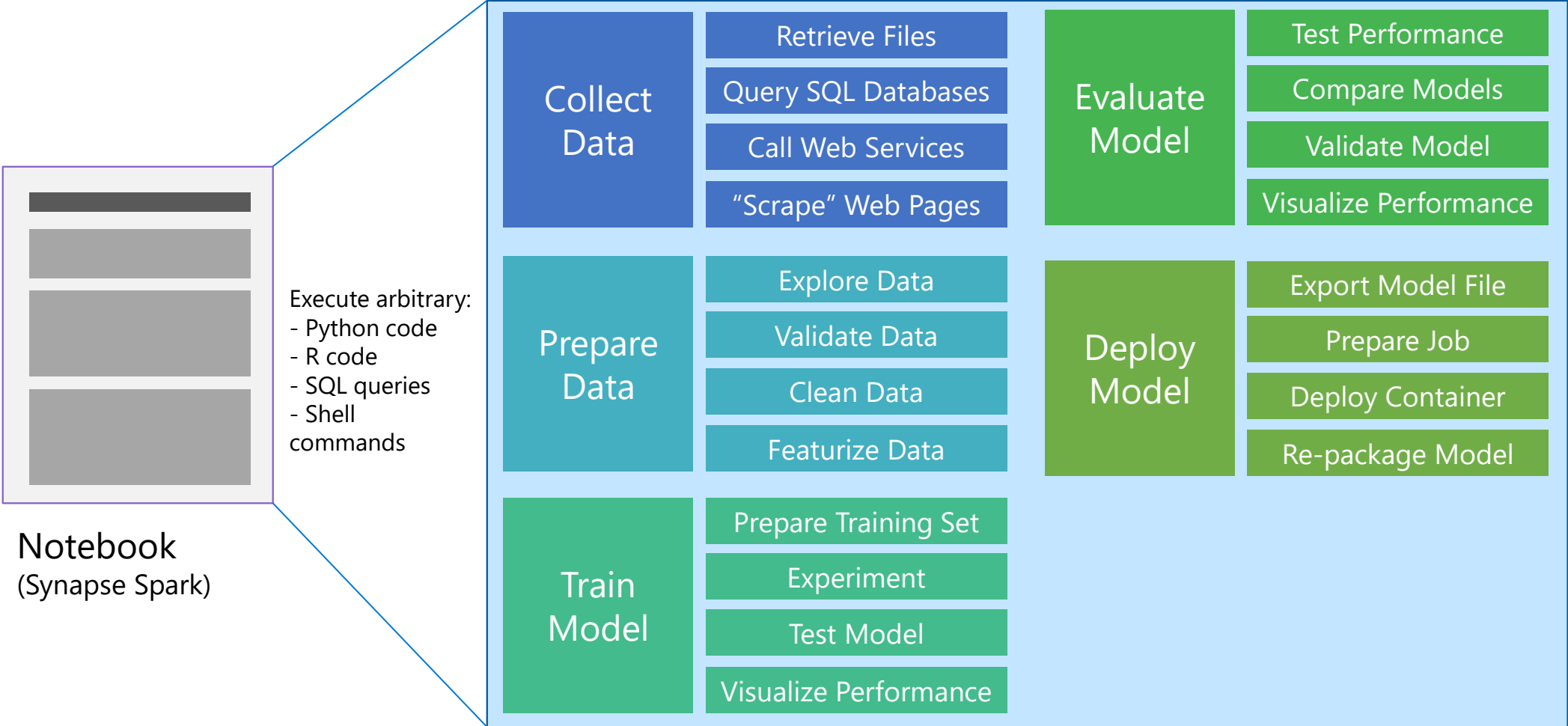
## Efficient resource utilization

- 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

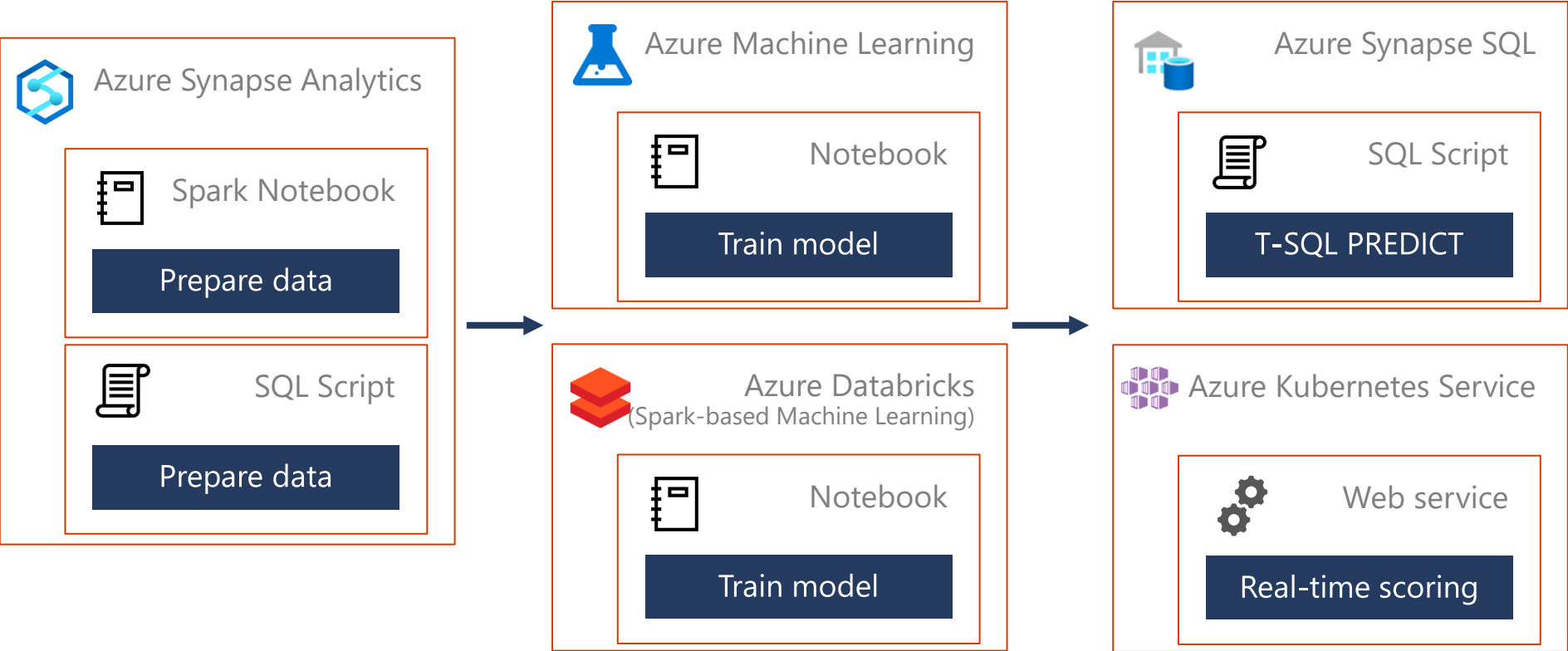


# Synapse Analytics and the Machine Learning Process

Prepare data

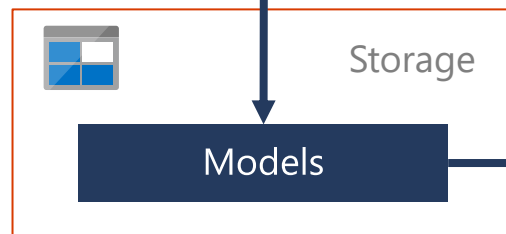
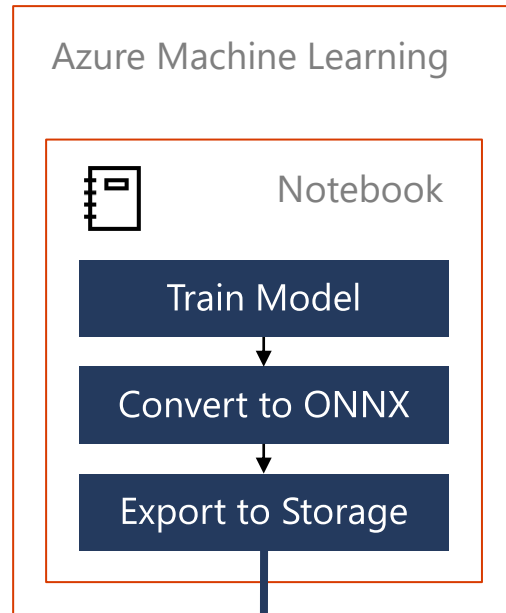
Train model

Use model

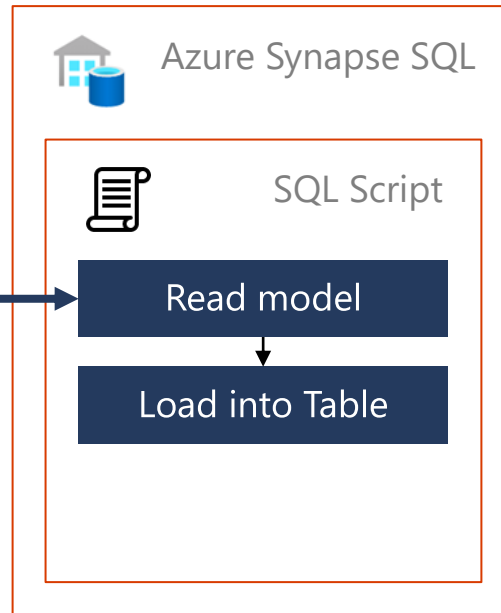


# Making predictions with T-SQL

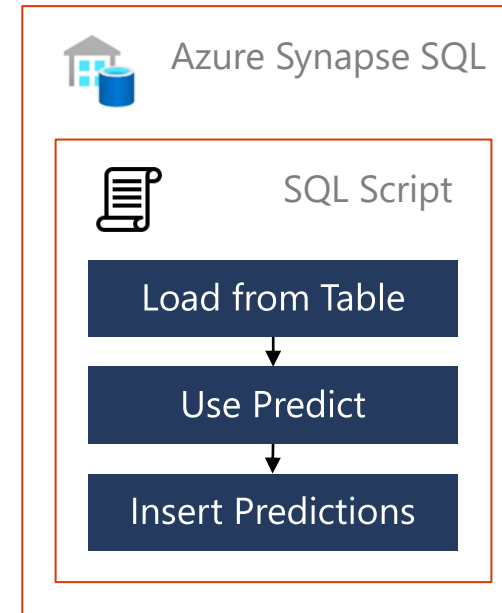
Create the model



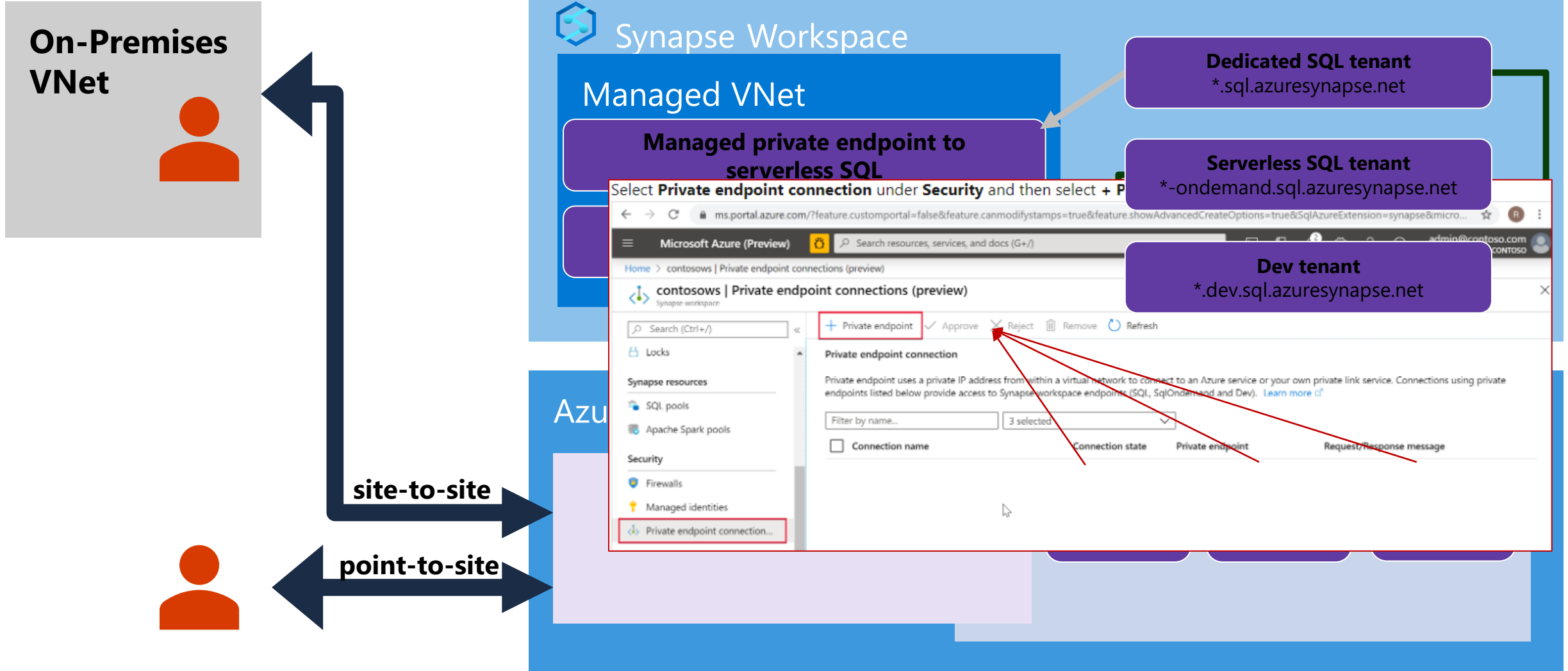
Register the model



Use the model



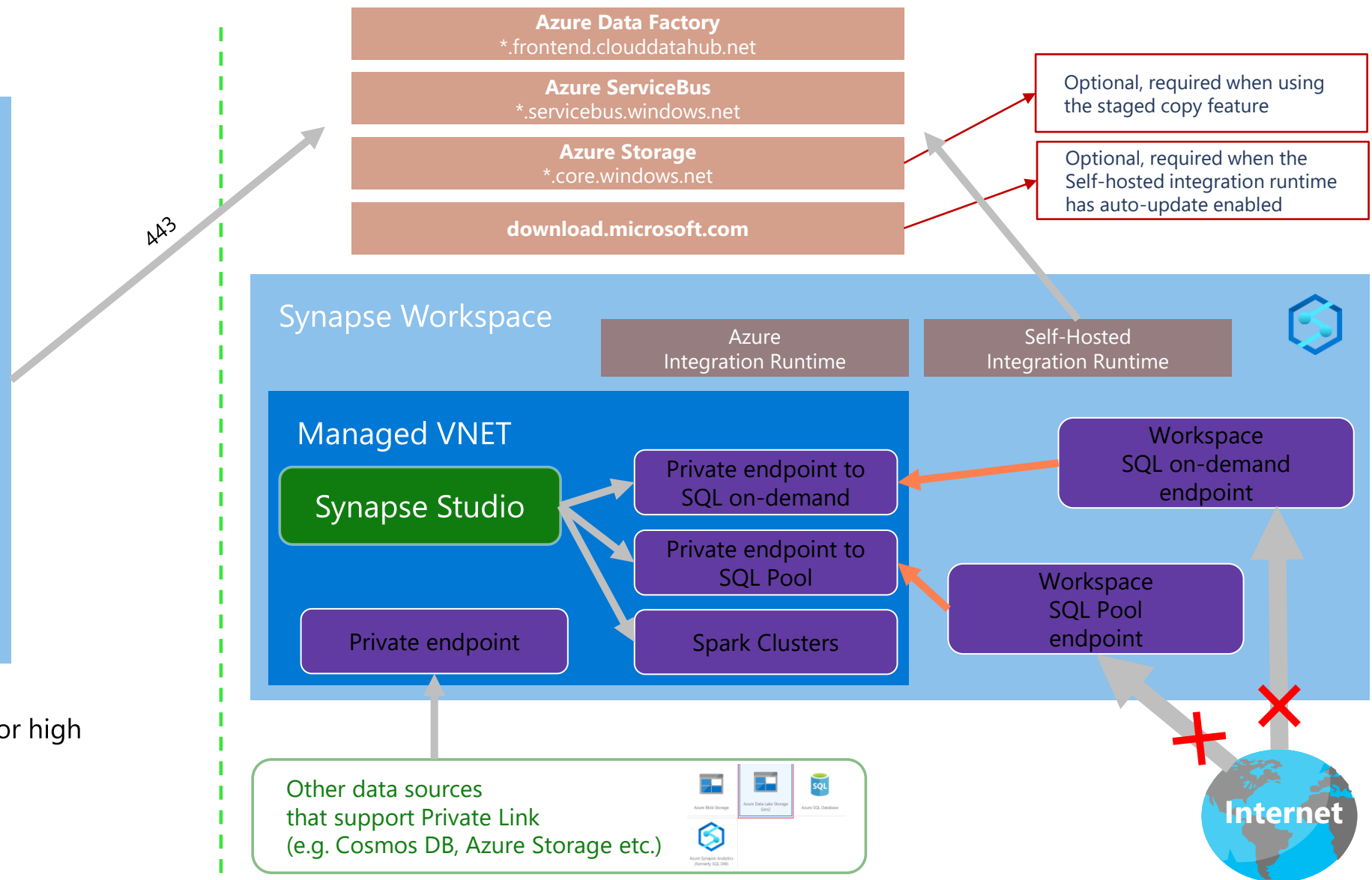
# Connecting to Synapse from On-Premises



# Self-hosted Integration Runtime in Synapse Workspace



Supports up to 4 nodes for high availability and scalability



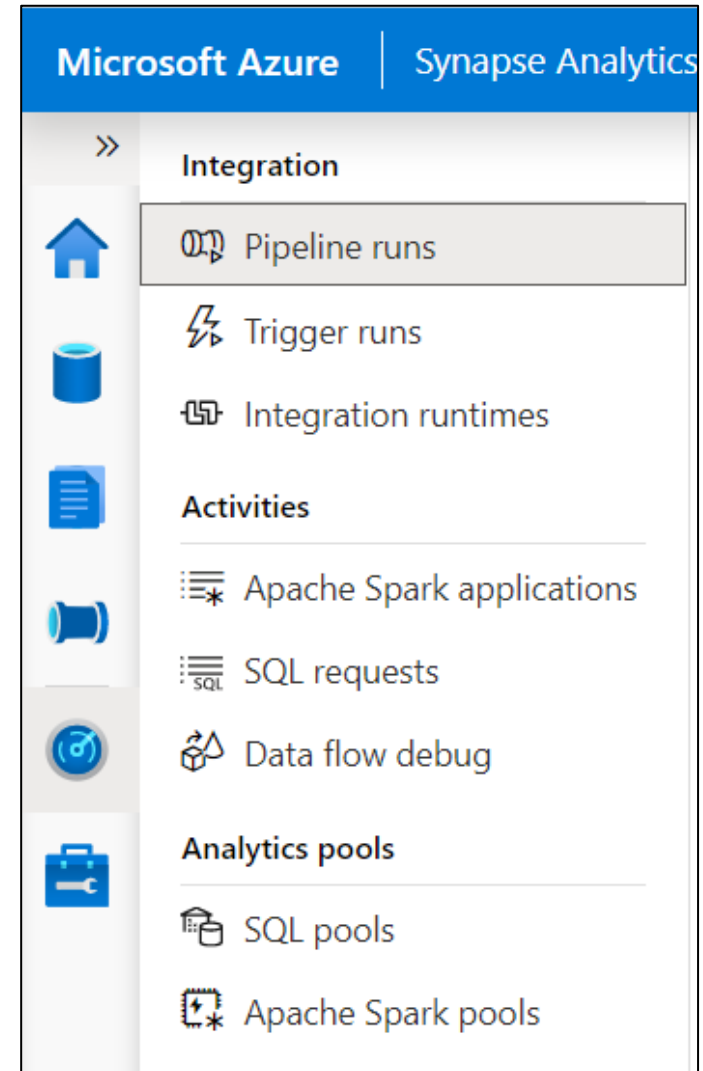
# Monitor Hub

## 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

## 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

**SQL requests**

Refresh Edit columns

Pacific Time (US & C... : **Last 30 days** Status : **All** Pool : Predict\_Pool Add filter

Showing 1 - 100 of 248 items

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 SchemaNarn	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

**Microsoft Azure** | Synapse Analytics | wsazuresynapseanalytics

**SQL requests**

Refresh Edit columns

Pacific Time (US & C... : **Last 30 days** Status : **All** Pool : Built-in Add filter

Showing 1 - 100 of 279 items

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

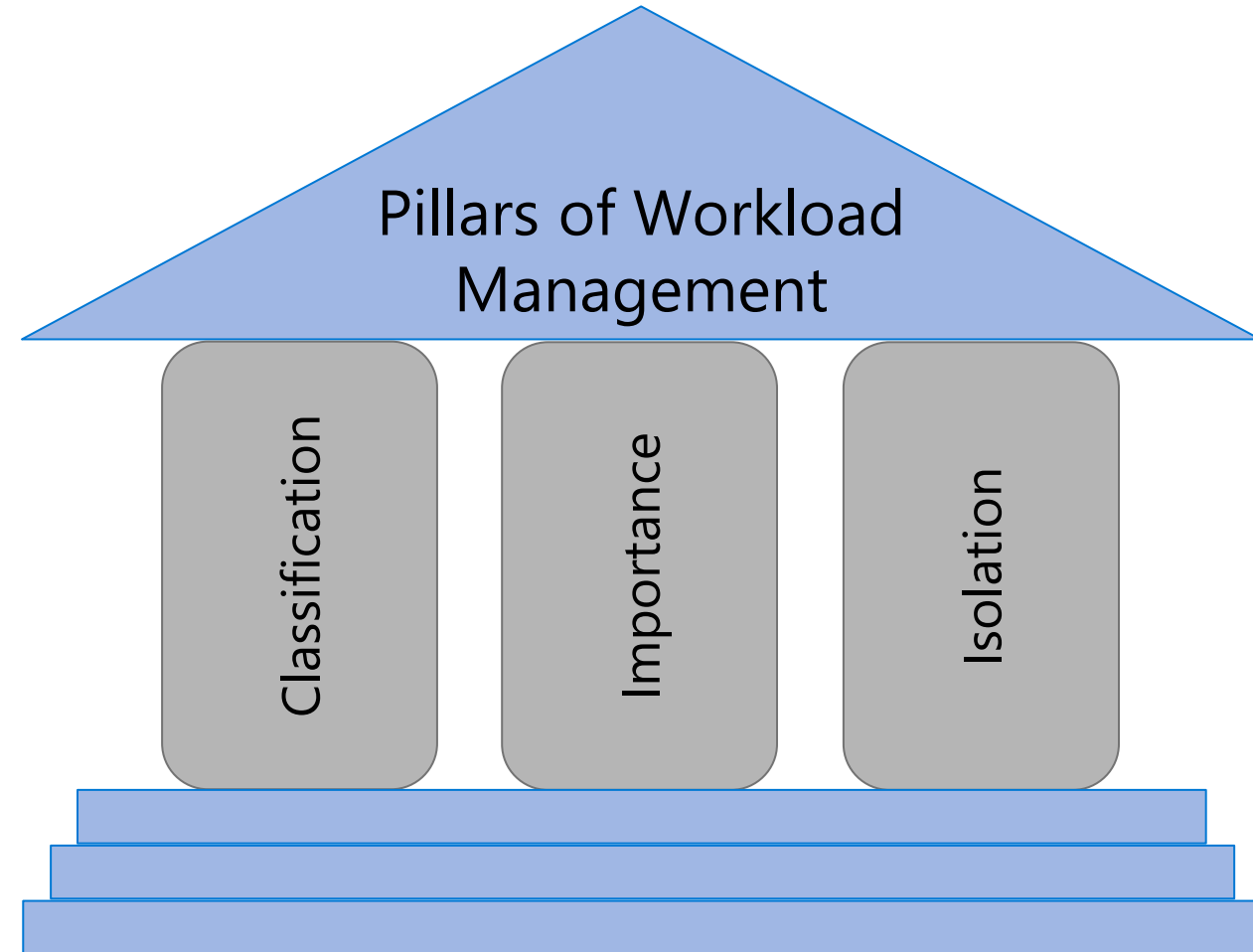
## Overview

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')
```



Azure Synapse  
Analytics



State  
Analyst



State  
Analyst



State  
Analyst



State  
Analyst



State  
Analyst



NephOSystems

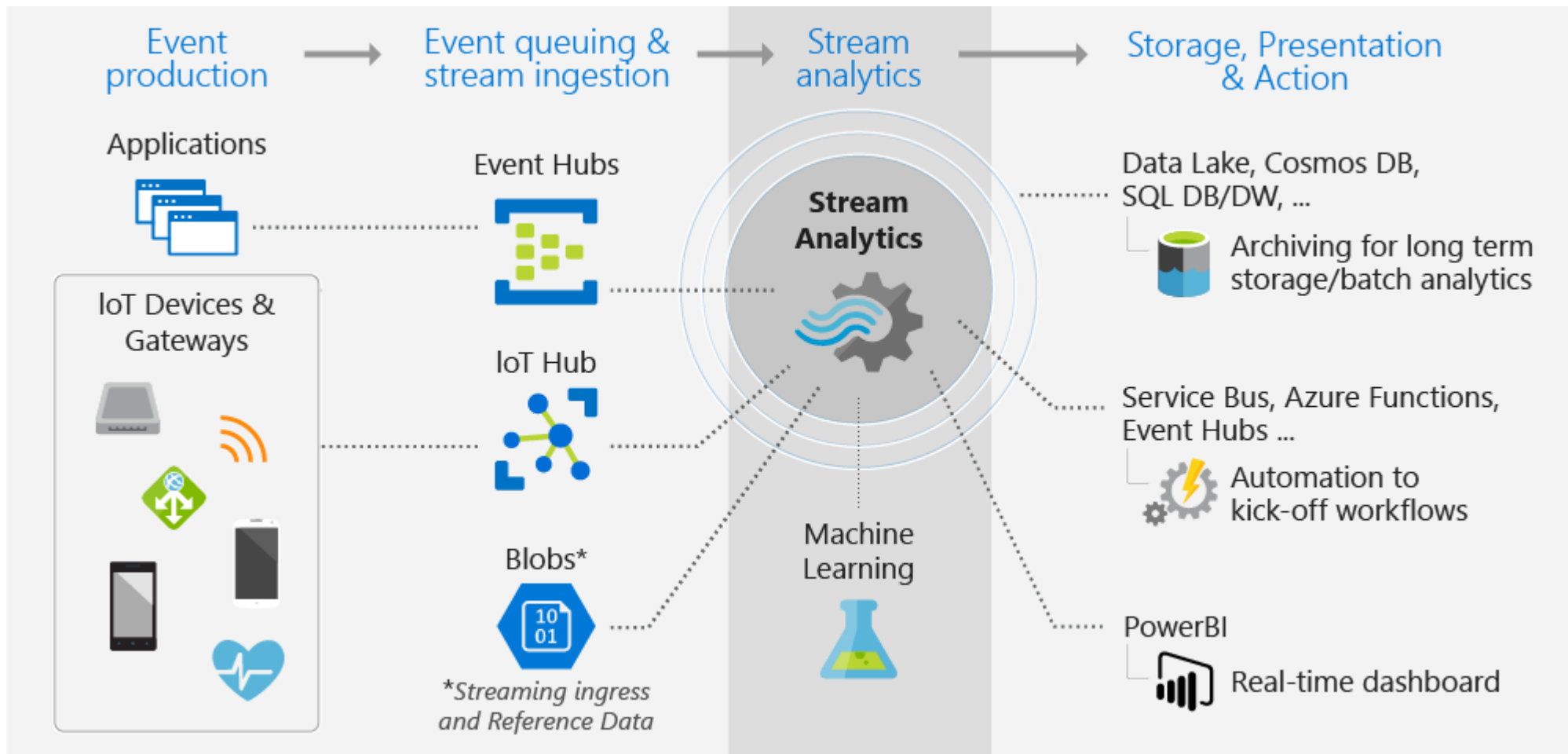
Cloud Computing for life

# Azure Stream Analytics



# Stream Analytics

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



Time Period: February 2017

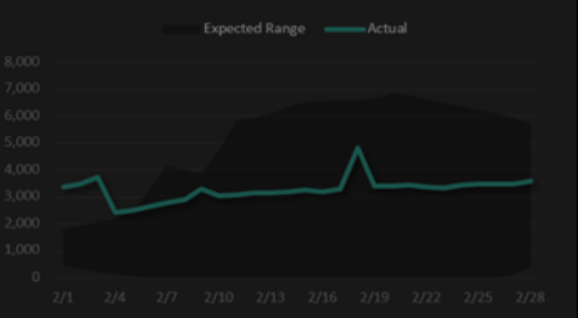
# Retail Demo KPI Dashboard

## TOP SITE METRICS

### Monthly Unique Visitors Trend

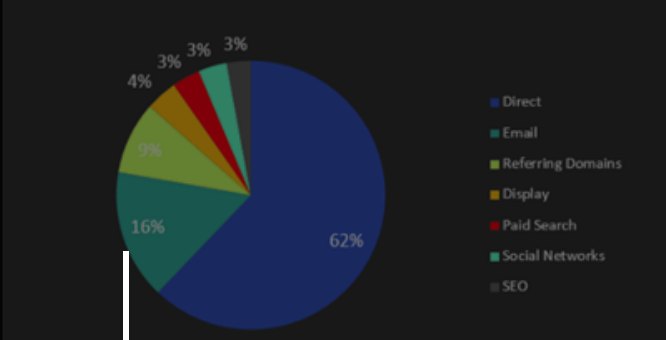


### Daily Unique Visitors Trend



## MARKETING CHANNELS

### Visits by Marketing Channel



### Top External Marketing Channels

Channel	6 Mo. Trend	Visits	MOM	Revenue	Conv. %
Direct	↑	67,924	55%	\$1,008,000	0.08%
Email	↑	17,000	19%	\$9,520	3.94%
Referring Domains	↑	1,540	15%	\$60,545	3.1%
Display	↑	4,160	257%	\$75,062	4.80%
Paid Search	↑	3,690	267%	\$47,684	5.38%
Social Networks	↑	3,750	233%	\$76,382	5.59%
SEO	↑	3,170	277%	\$54,797	4.76%

## PRODUCTS ORDERS & REVENUE

### Monthly Orders



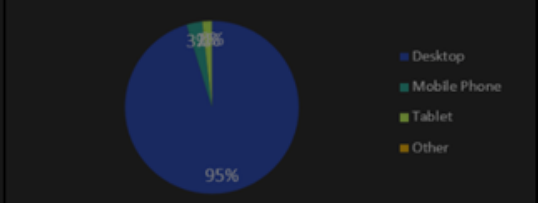
### Monthly Revenue



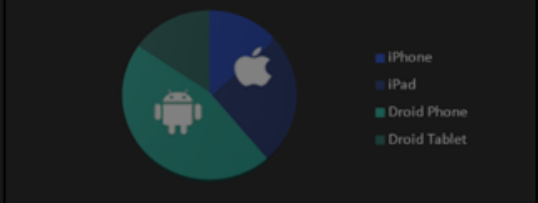
Products	6 Mo. Trend	Orders	Revenue
Timberline GTX Snowboard	↑	4	\$195,838
Vasatch Xtreme Parka	↑	3	\$163,754
Uintas TechX Snow Pants	↑	2	\$76,199
Bonneville Shore Swimsuit	↑	2	\$46,383
Timpanogos Scarf	↑	2	\$10,214
La Sal Sweatshirt	↑	2	\$21,878
Uintas Pro Ski Gloves	↑	1	\$20,579
Uintas TechX Parka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

## MOBILE DETAIL

### Mobile Visits



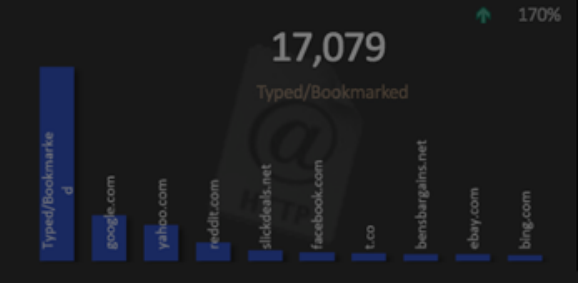
### Mobile OS



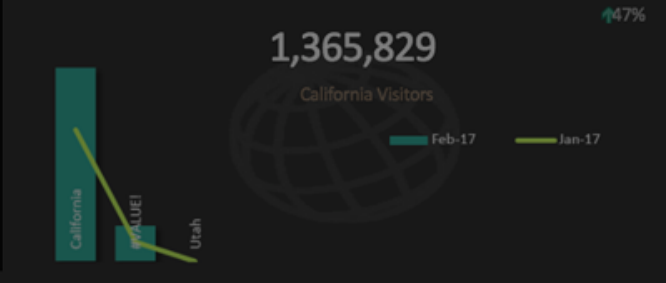
Device Name	Visits	MOM
Apple iPad	1,117	↑ 216%
Apple iPhone	61	↑ 219%
Samsung Galaxy S3 (GT-I9300)	21	↑ 451%
Samsung Galaxy S4 LTE (GT-I9505)	1	↑ 228%
Samsung Galaxy S2 Plus (GT-I9105P)	1	↔ 0%
Samsung Galaxy Tab 2 (GT-P5110)	1	↑ 131%
Google Nexus 7	1	↑ 21%
Samsung Galaxy Note 2 (GT-N7100)	1	↑ 1620%
Samsung Galaxy Tab 2 7.0 (GT-P3100)	1	↑ 240%

PowerBI

## Top Referring Domains



## Geography



## 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%
<b>Overall Conversion</b>			<b>0%</b>	<b>100%</b>

# 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.

