
Microsoft Azure Solutions Business
Welcome to
Azure Synapse Technical Boot Camp
Day 3

# A look into Day 3

| | | | |
|---|---|---|---|
| Model | 8:00-8:05 | Welcome | |
| | 8:05-8:45 | Spark for Data Science | Main Call |
| | 8:45-9:00 | Break | |
| | 9:00-10:00 | Build Hands-on: Machine Learning | Table Group Call |
| | 10:00-11:00 | Activity: Model Implementation (Predict) | |
| | 11:00-11:15 | Break | |
| | 11:15-12:00 | Build Hands-on: Spark Machine Learning | |
| | 12:00-1:00 | Break | |
| Monitor | 1:00-1:30 | Monitor & Manage | Main Call |
| | 1:30-2:00 | Activity: Monitor & Manage | |
| | 2:00-3:00 | Build Hands-on: Monitoring | Table Group Call |
| | 3:00-3:15 | Closing | Main Call |

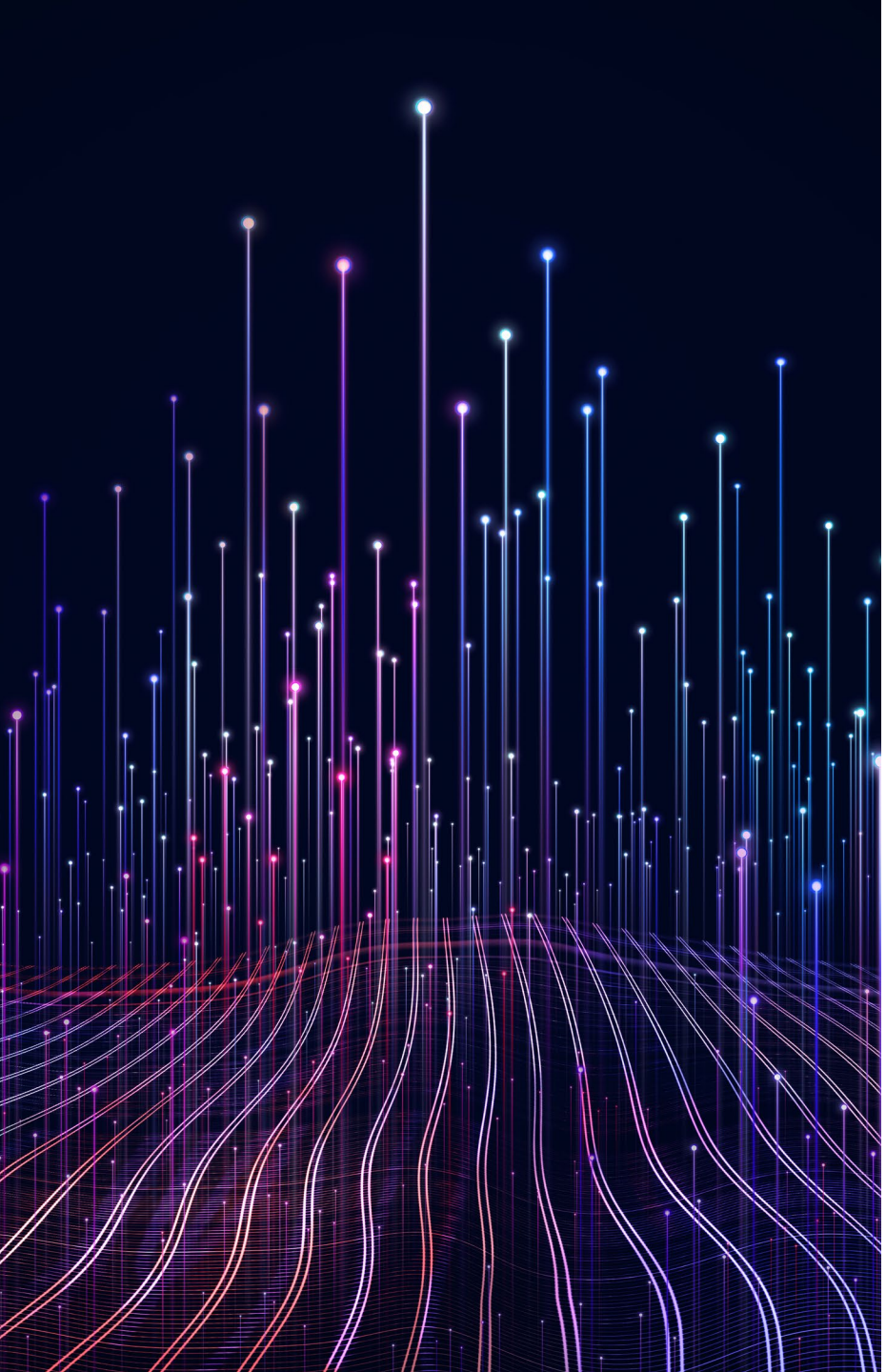
**Today we will be learning and collaborating across three spaces:**

- Main call (this meeting)
- Table Group channel within the event Team
- CloudLabs Learner Portal & Synapse environment

Presentation/ Whole Group | Lab | Activity/ Discussion/ Group Work | Announcements

Microsoft Azure Solutions Business

# Spark for Data Science



# Agenda

1 **Apache Spark**

Overview of Apache Spark.

2 **Apache Spark in Synapse**

How Spark is utilized within Synapse.

3 **Notebooks**

Understand the notebook paradigm and how to use them in Synapse.

4 **Machine Learning**

How to train and deploy models with Synapse.

# Apache Spark Overview

# Question...

How have you seen Spark used in your customer engagements?

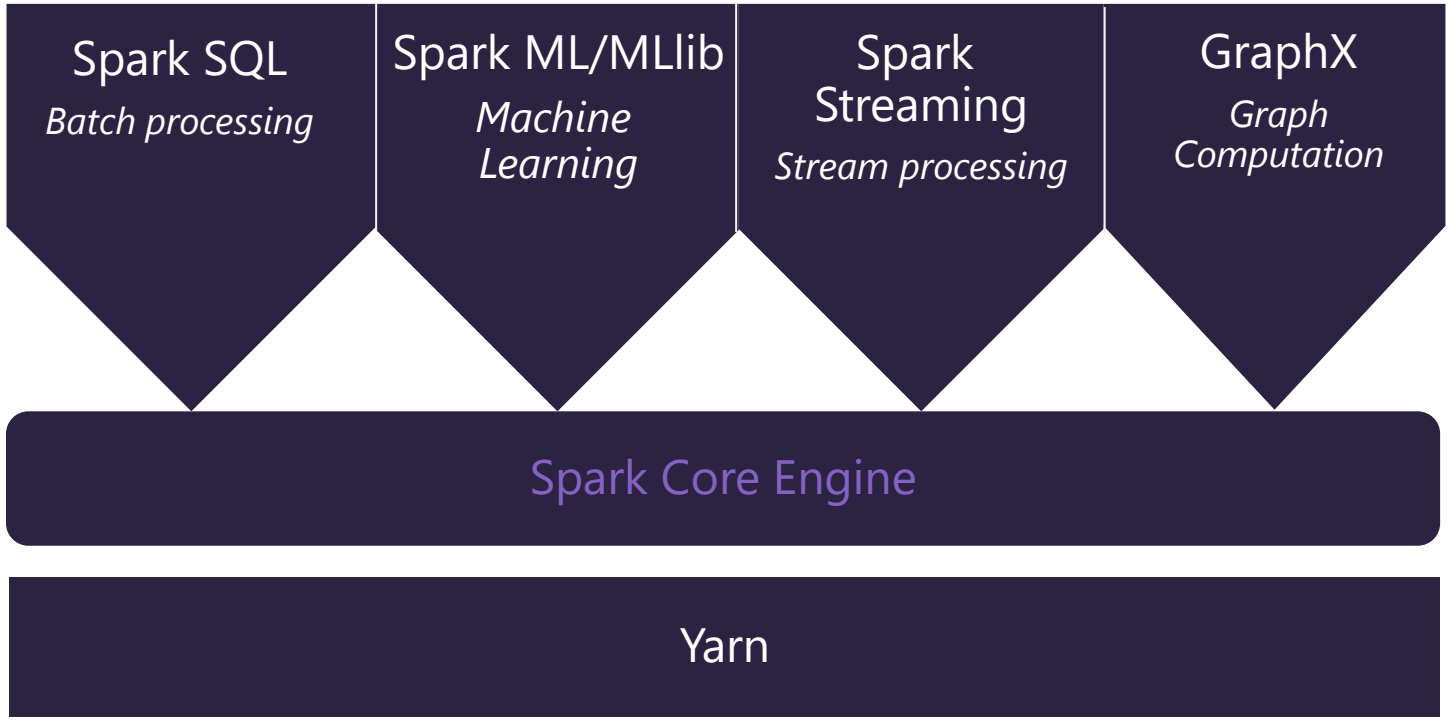
What use-cases does it excel in?

# Apache Spark

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

**Spark Unifies:**

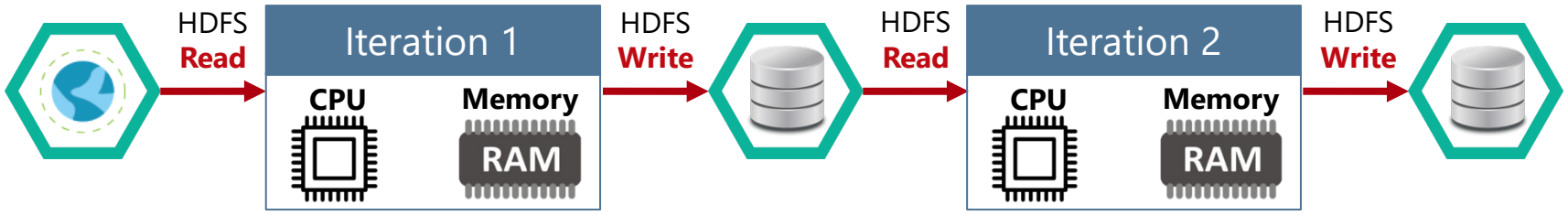
- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing




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

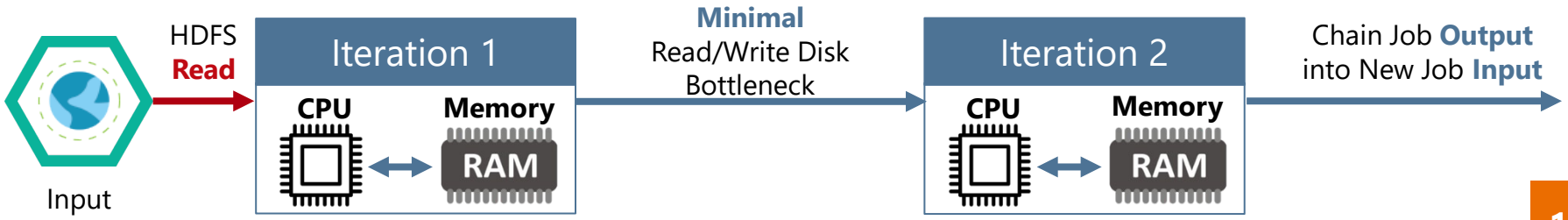


# Motivation for Apache Spark

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



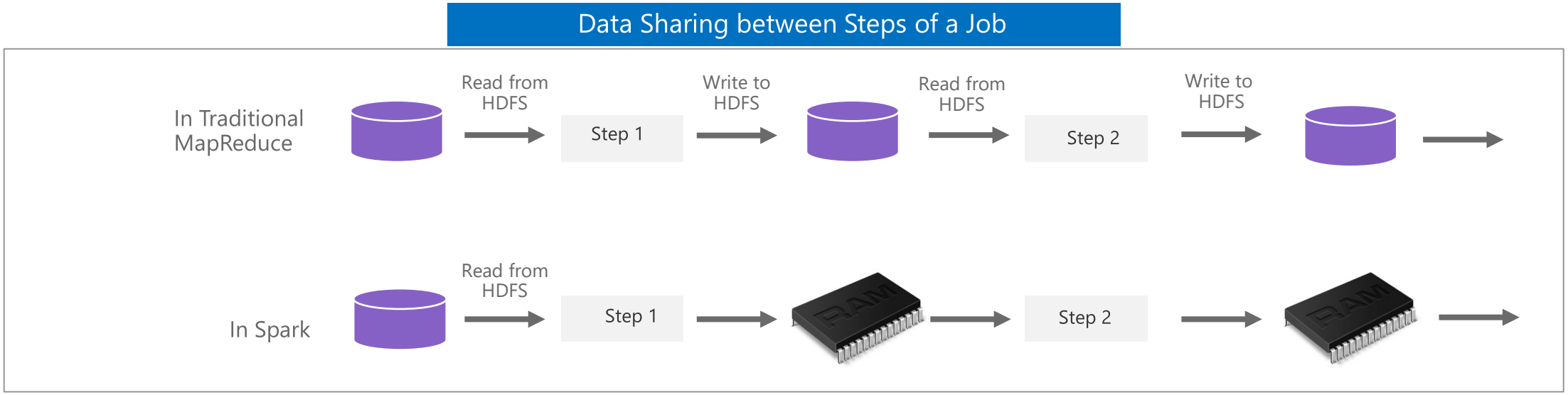

Solution: Keep data in-memory with a new distributed execution engine




**10–100x** faster than network & disk

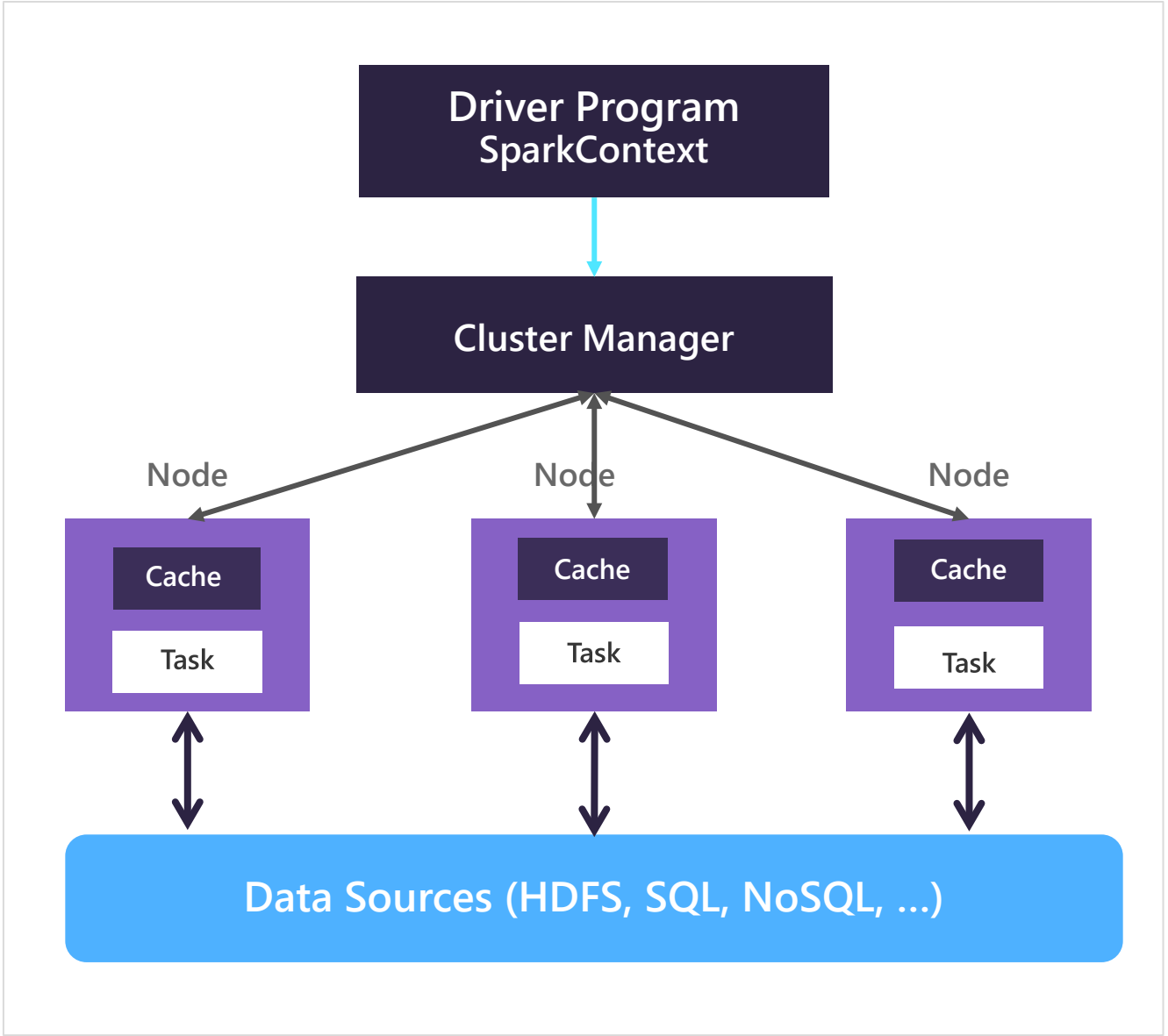
# What makes Spark fast

- **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 Distributed Datasets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).



# Apache Spark in Azure Synapse Analytics

# 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

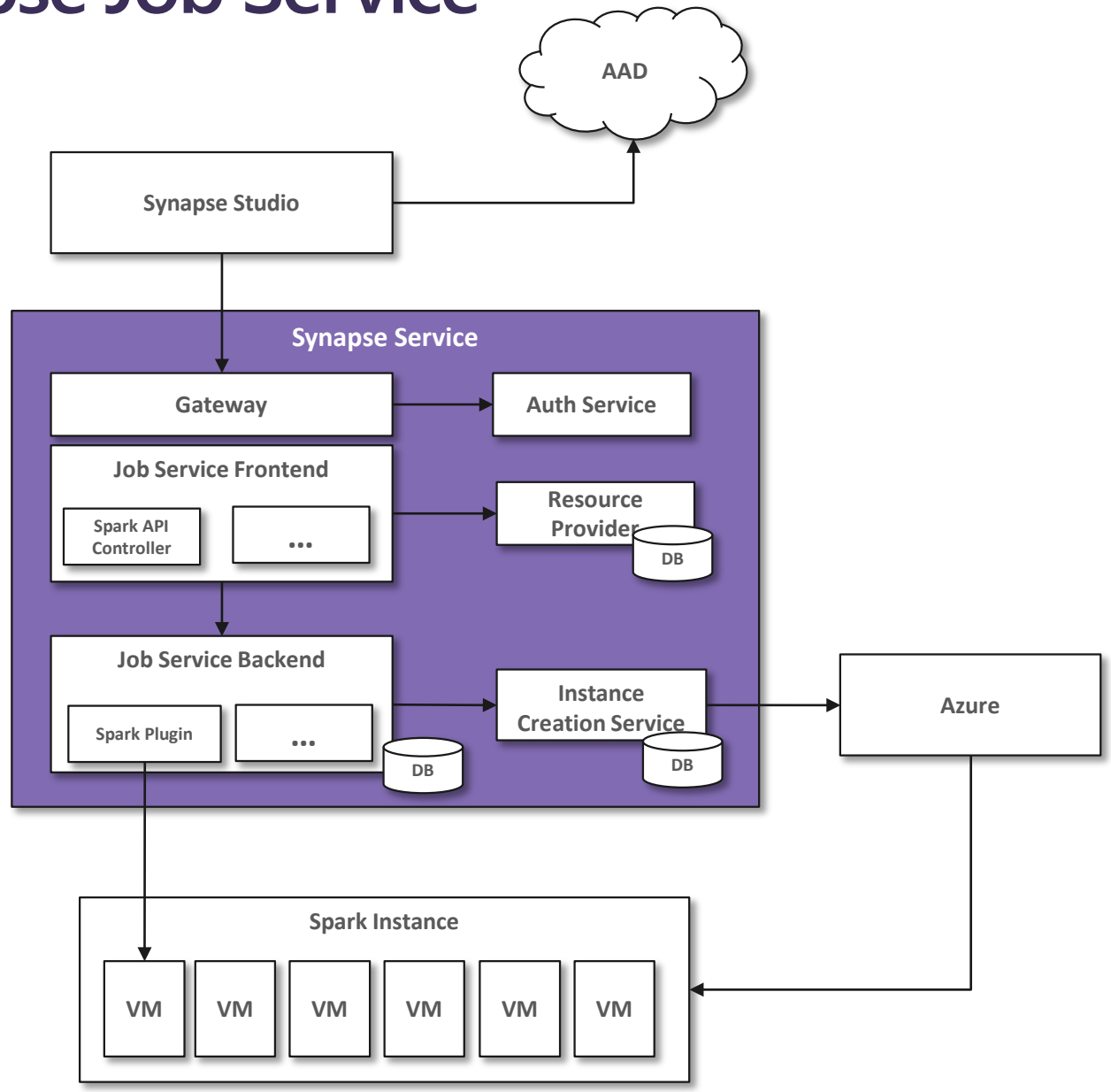
**Efficient resource utilization**

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

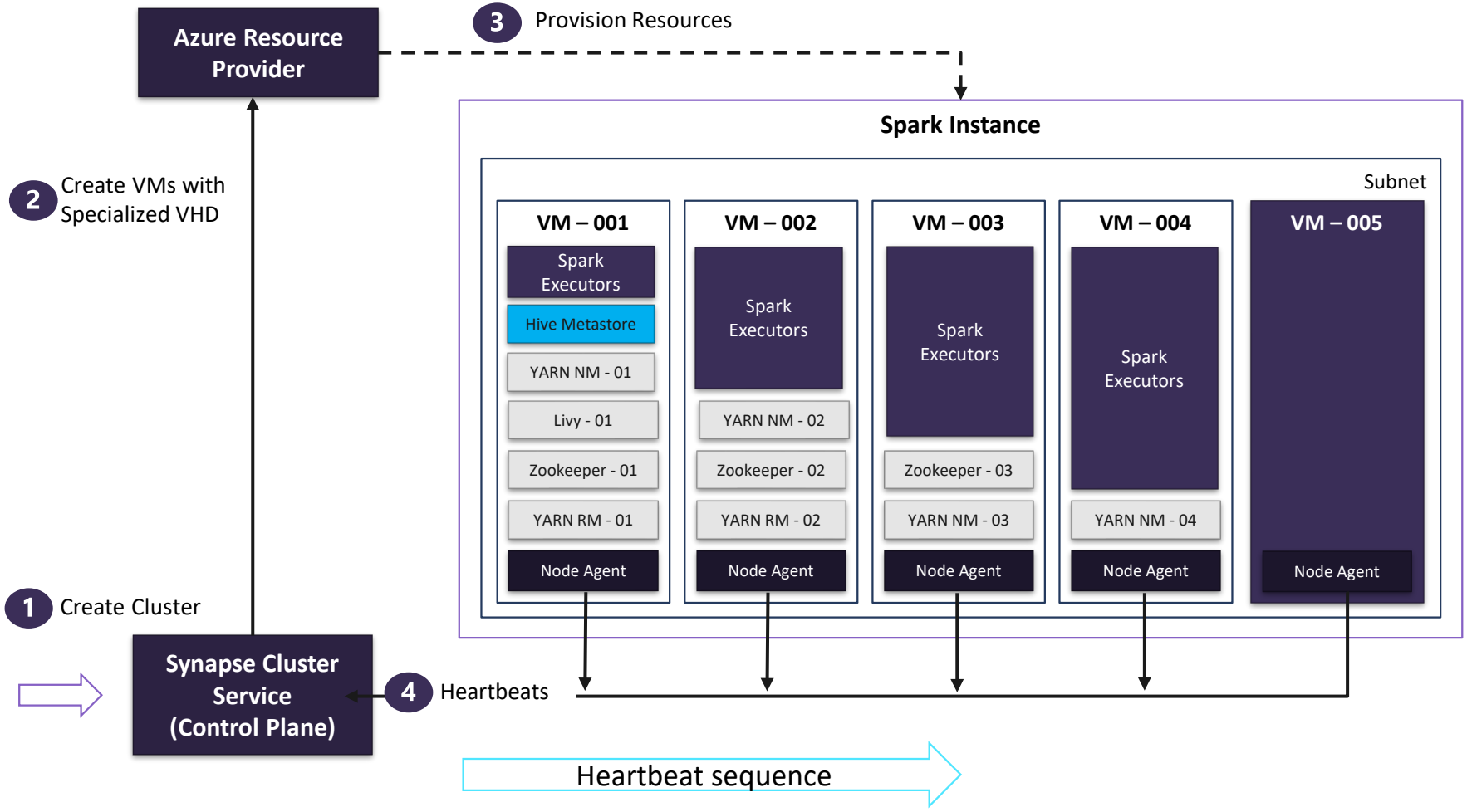
**Multi Language Support**

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

# Synapse Job Service



# Spark Instance Creation



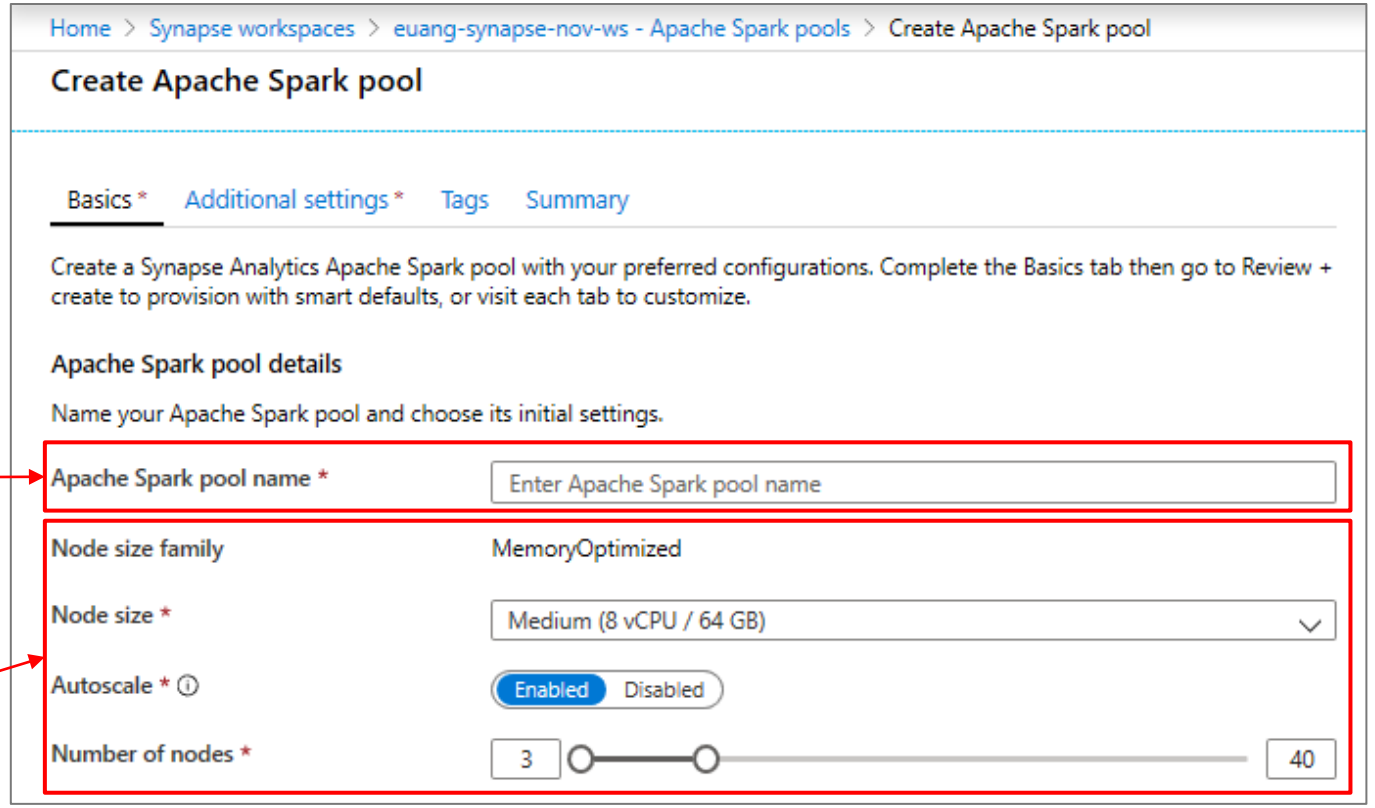
# Creating a Spark pool (1 of 2)

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

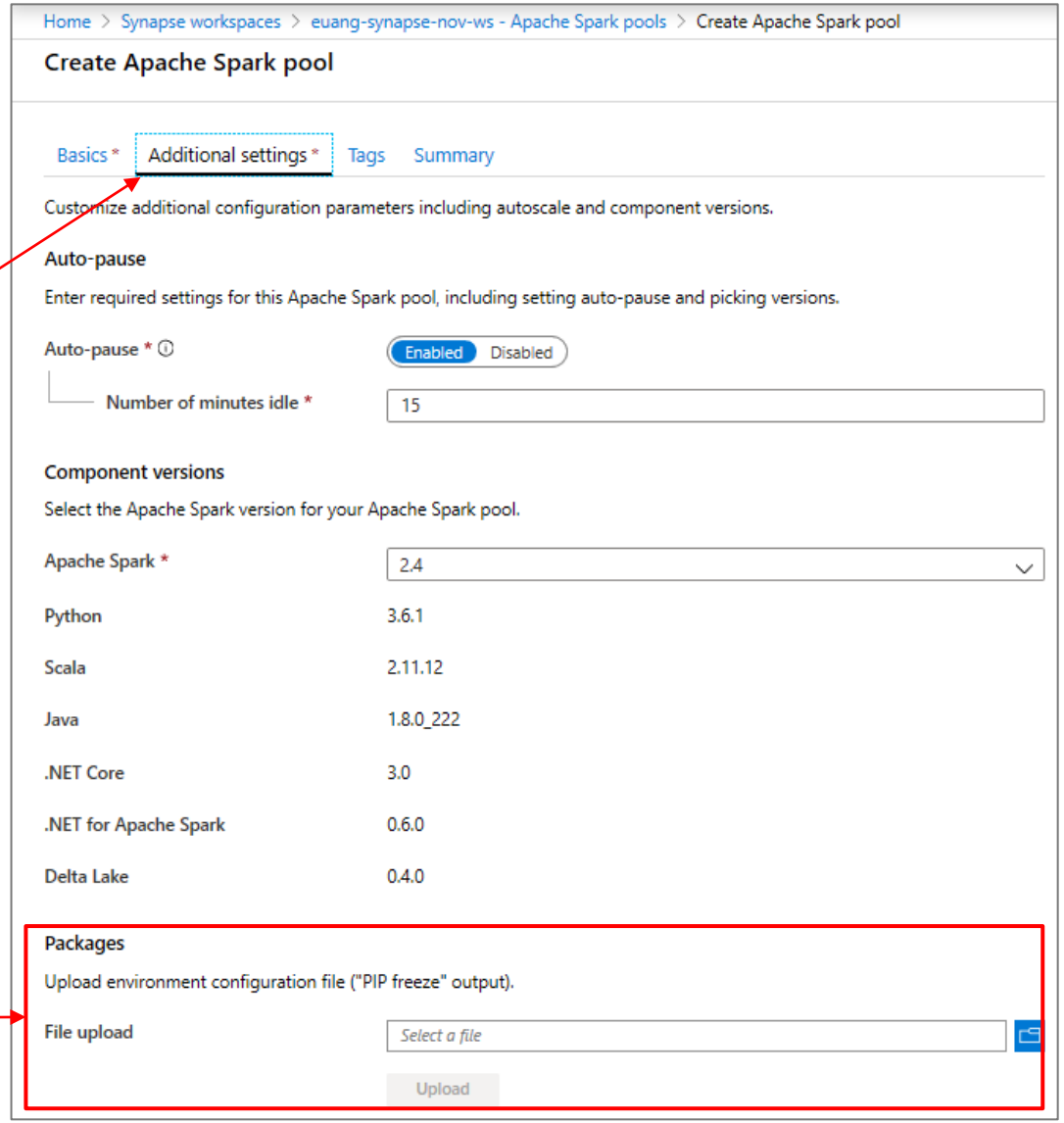


# Creating a Spark pool (2 of 2) - optional

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



# 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

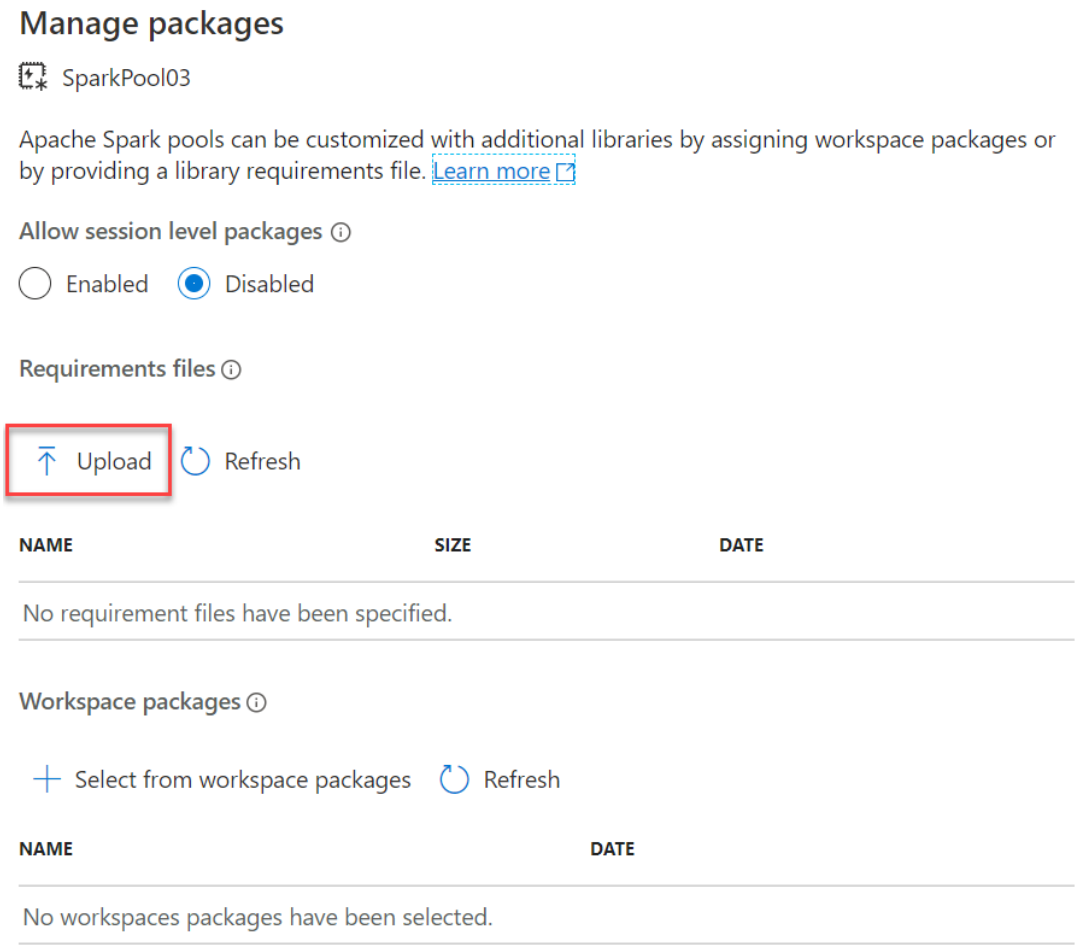
Support for custom Wheel files when uploaded to the workspace default file system.

## Constraints

The library version must exist on PyPI repository

Version downgrade of an existing library not allowed

Wheel files must be installed at a specific path:
abfss://<file_system>@<account_name>.dfs.core.windows.net/synapse/workspaces/<workspace_name>/sparkpools/<pool_name>/libraries/python/



# Python Library Requirements Format

## Overview

The requirements text file has specific format requirements. These files are traditionally created in Python environments by running the `pip freeze` command, but are also easy to create by hand.

File must be named **requirements.txt**.

Each line contains a requirement specifier.

- This must contain at least the package name. You can also specify optional extras, an operator and version numbers according to the PIP requirements file standard.

The basic operators used by version specifiers are:

- An exact version (using ==).
- A minimum version (using > or >=).
- A maximum version (using < or <=).

The extras listed in square brackets ([]) allow you to install additional components or variants of the library.

libraryname ==1.3.0

# Spark Pools - Configuring

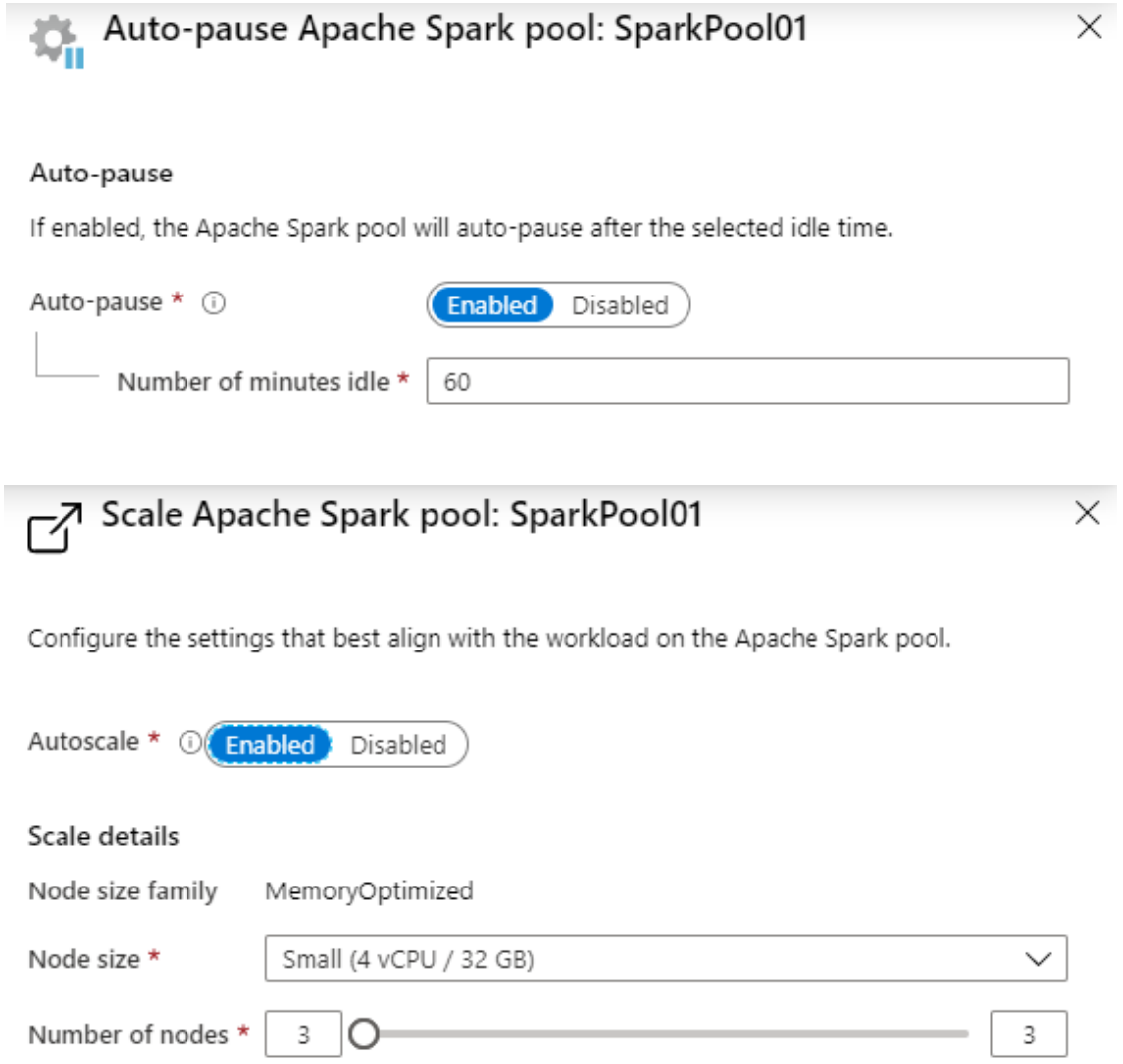
## Auto-pause

Enables the pool to automatically stop running when no Spark applications (e.g., Notebooks) are running.

## Autoscale

Enables the pool to automatically adjust the number of running nodes (VM's) supporting the Spark pool.

Note: Just because a pool is running, does not mean all the nodes allowed to run, according to the scale configuration, are actually running. They will be allocated on demand as new Notebooks are executed, up to the limit configured in autoscale.

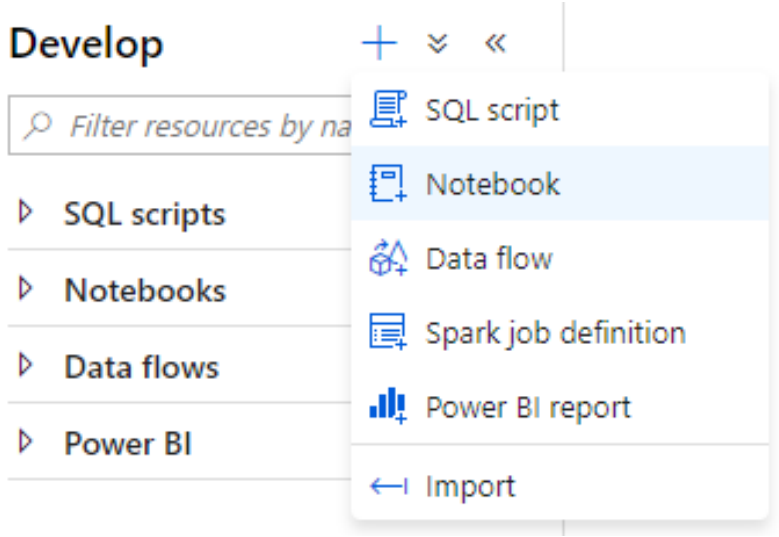


# Notebooks in Azure Synapse Analytics

# Creating a Notebook

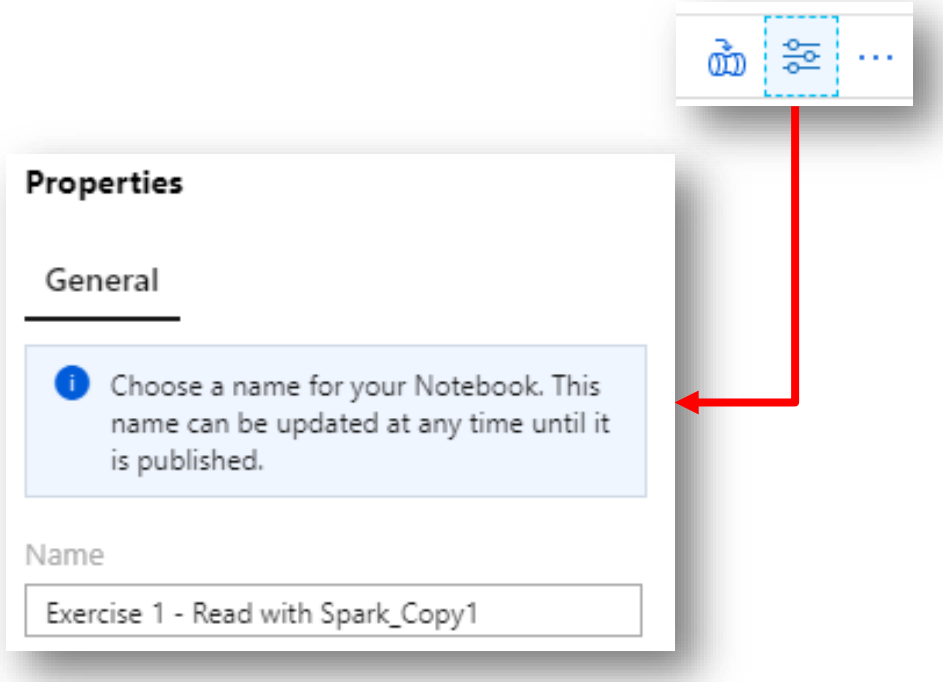
## Notebook creating

Create a new notebook within the Develop hub of Synapse Studio.




## Notebook naming

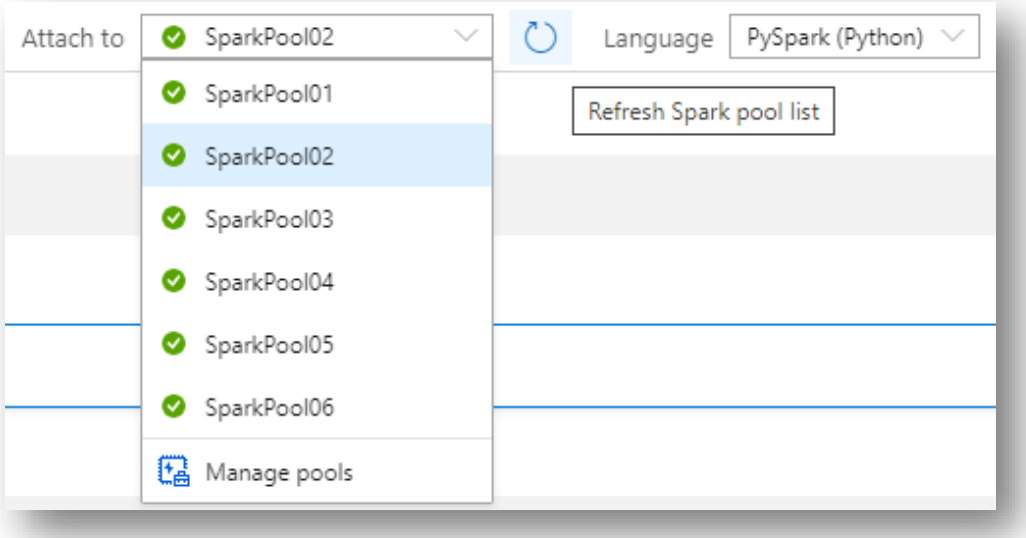
The name of the Jupyter notebook can be set in the Properties dialog. You can now re-name published notebooks!



# Notebook setup

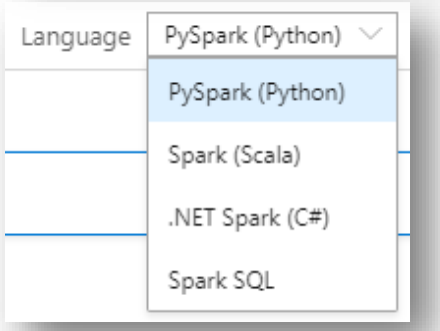
## Selecting the Spark pool to attach to

In the notebook toolbar, select the Spark pool to attach to before running any cells in the notebook.




## Notebook default language

The default language used for all cells in the notebook is selected in the Language drop down.




You can override this for any particular cell to use a different language by using the appropriate cell magic (covered later).

# Notebook setup – Session configuration

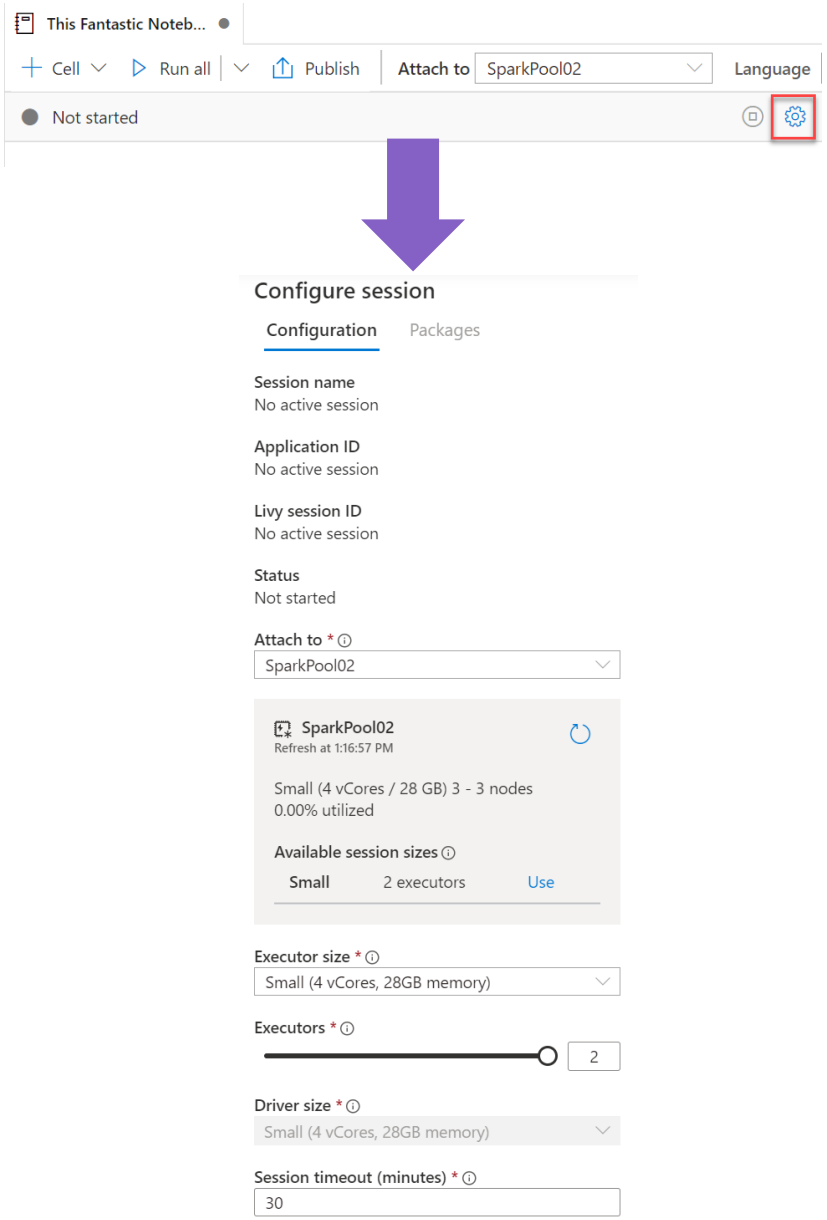
A Spark session is created upon first run of any cell in a notebook. The configuration for a Spark session can only be set *before* it is created.

The Spark session allocates one Spark application.

This Spark application consumes a fixed set of resources, as defined in the session configuration. These resources are reserved from the total resources available in the Spark pool upon session start.

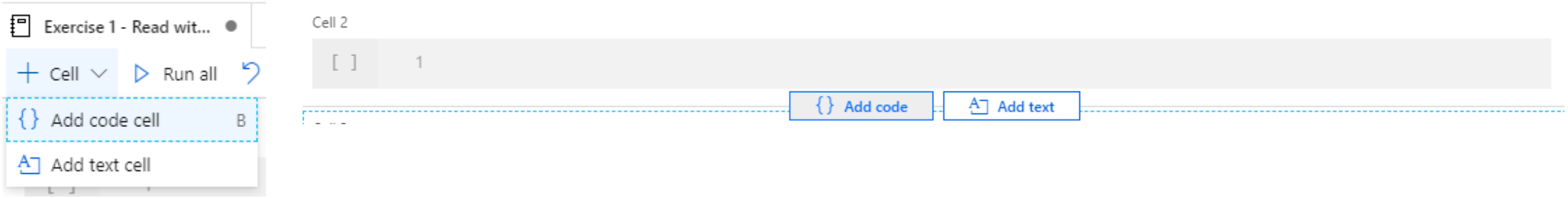
Each session reserves resources for at least one driver process and two executor processes. You can configure the size of executors, as well as the number of executors. The size of the driver is always the same as the executors.

A session ends when the notebook is closed or when the session timeout is reached without any activity in the notebook. You can configure the session timeout.

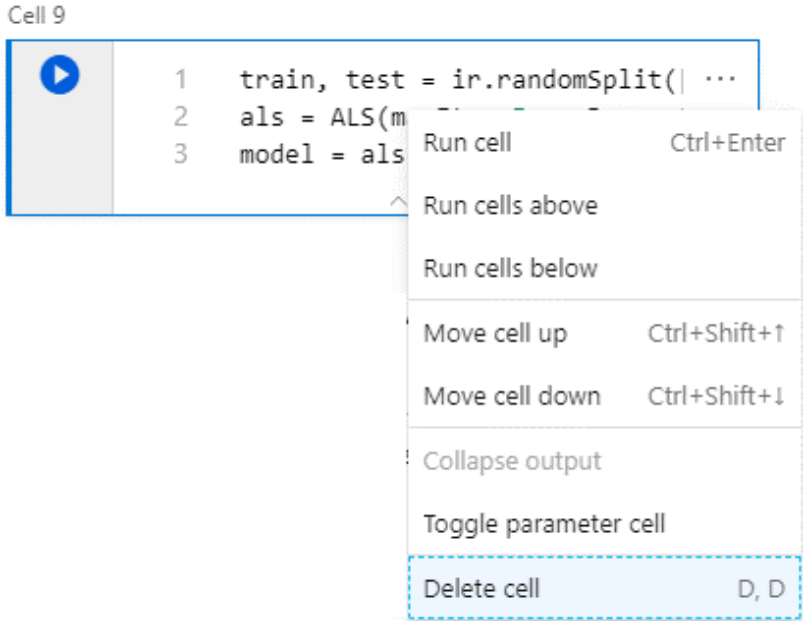


# Notebook capabilities – Cell management

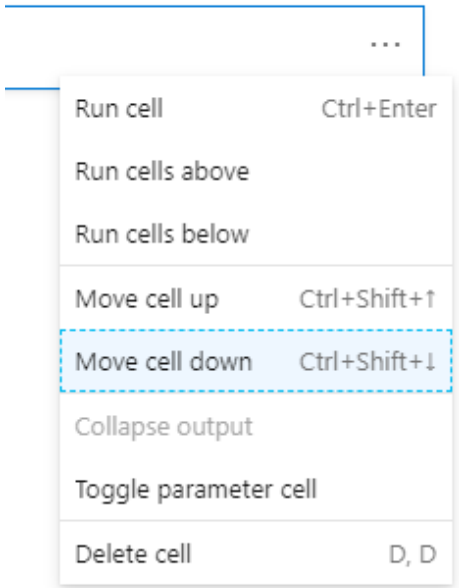
## Add code or text cells




## Remove code or text cells



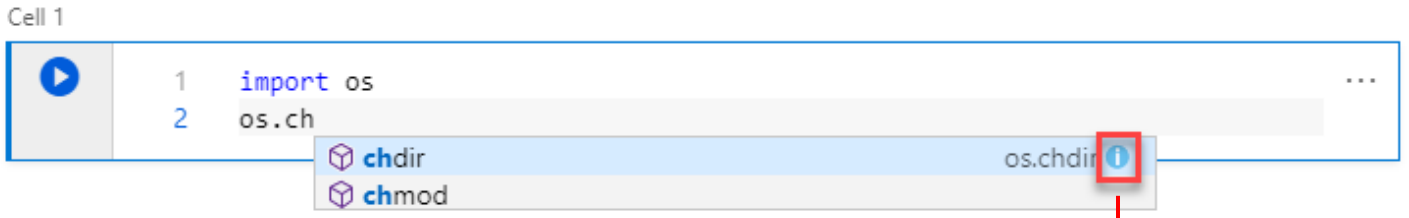

## Adjust cell order



# Notebook capabilities – Editing code

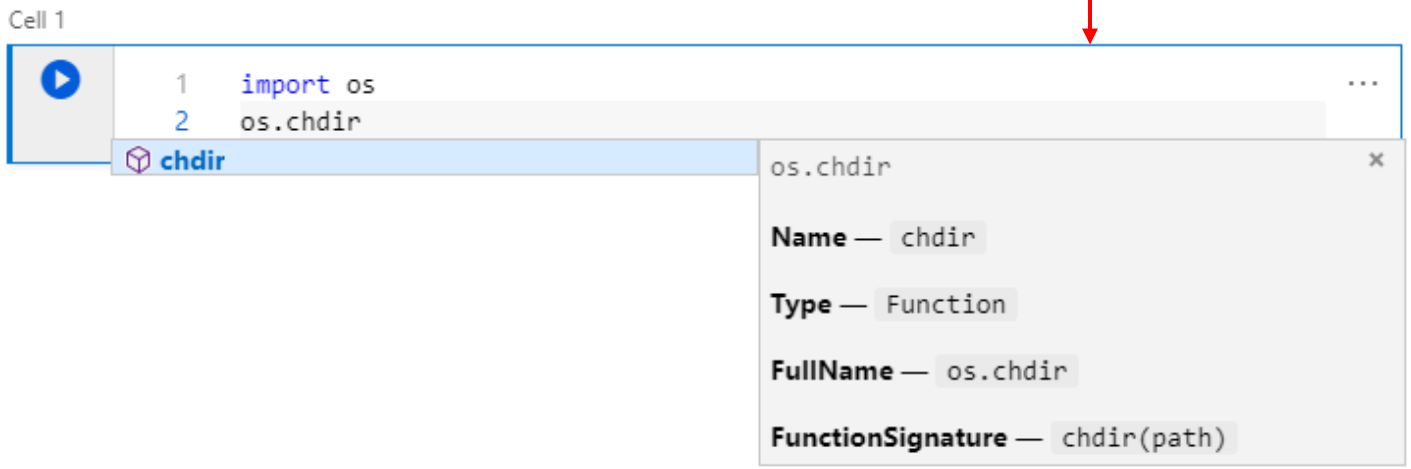
## Statement completion

Use Ctrl + Space to display




## Inline documentation

Select the information tooltip to display



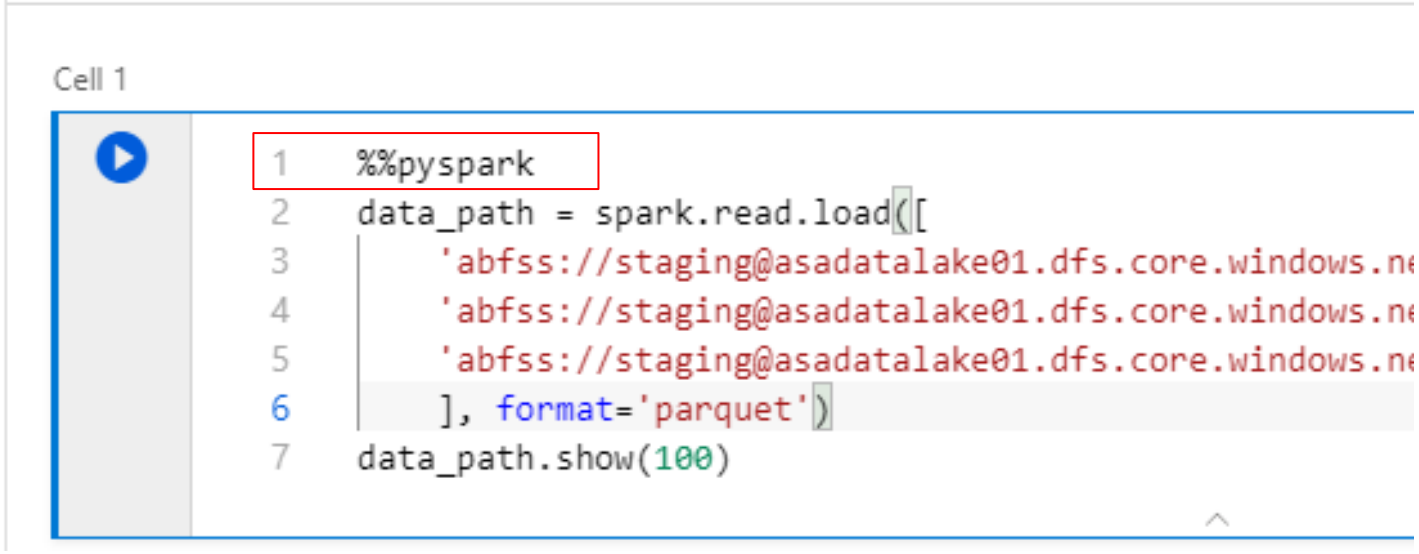
# Notebook capabilities – Editing code

## Cell magics

Use cell magics to change the language used when interpreting the cell.

These must be the first line in the cell.

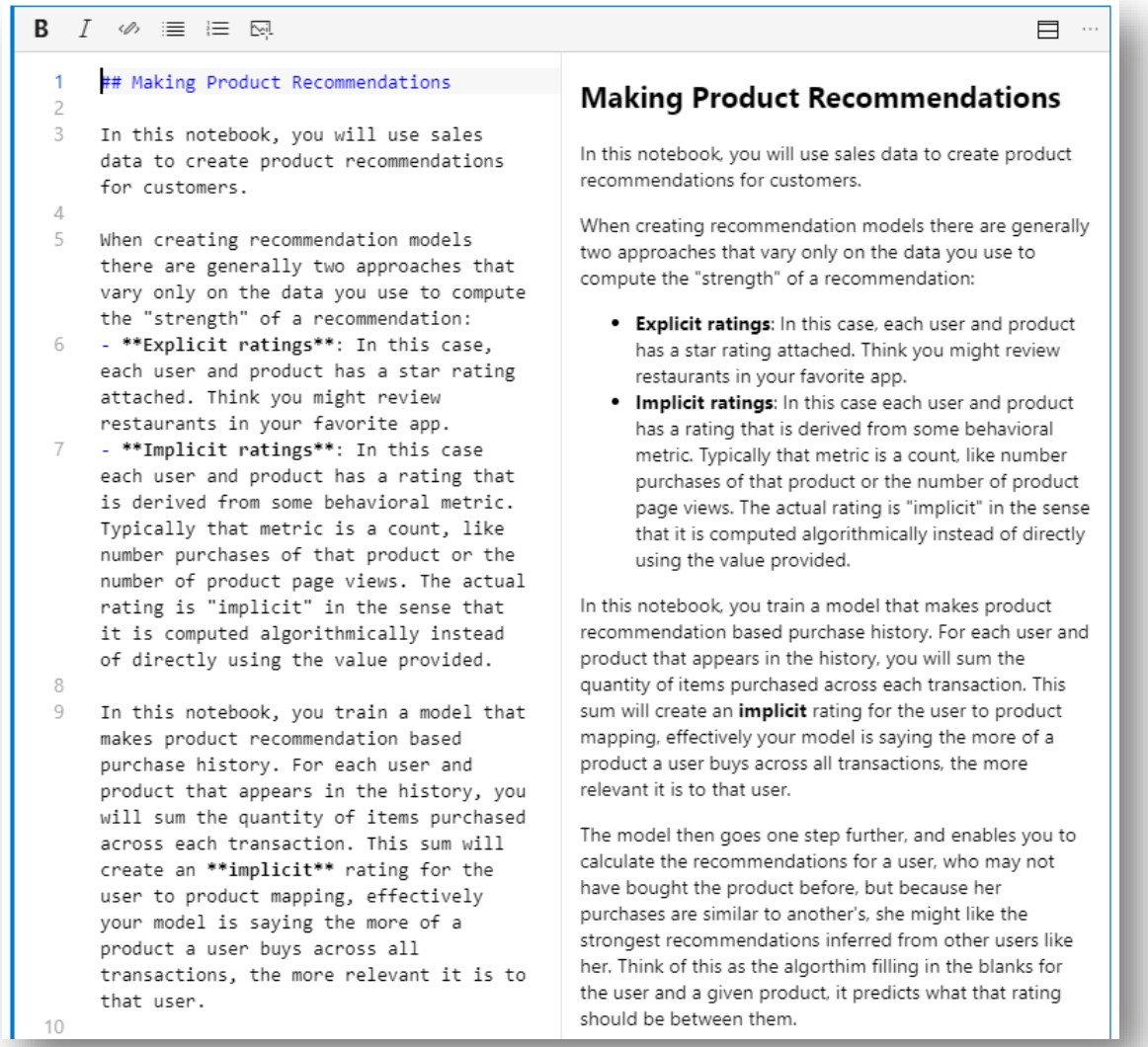
| Magic | Description |
|---|---|
| %%pyspark | Execute **Python** code against Spark Context. |
| %%spark | Execute **Scala** code against Spark Context. |
| %%sql | Execute **Spark SQL** query against Spark Context. |
| %%csharp | Execute **Spark.NET C#** code against Spark Context. |



# Notebook capabilities – Editing text cells

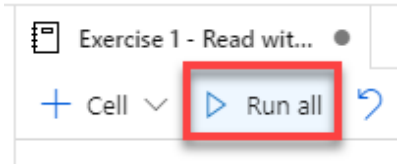
## Support for GitHub flavored markdown

Notebooks support GitHub flavored markdown, enabling limited text formatting, links, lists, code blocks and images.

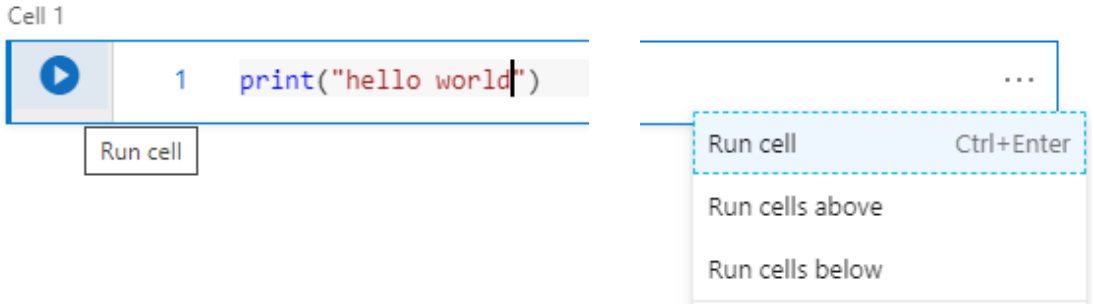


# Notebook capabilities – Running cells

## Run all cells from toolbar

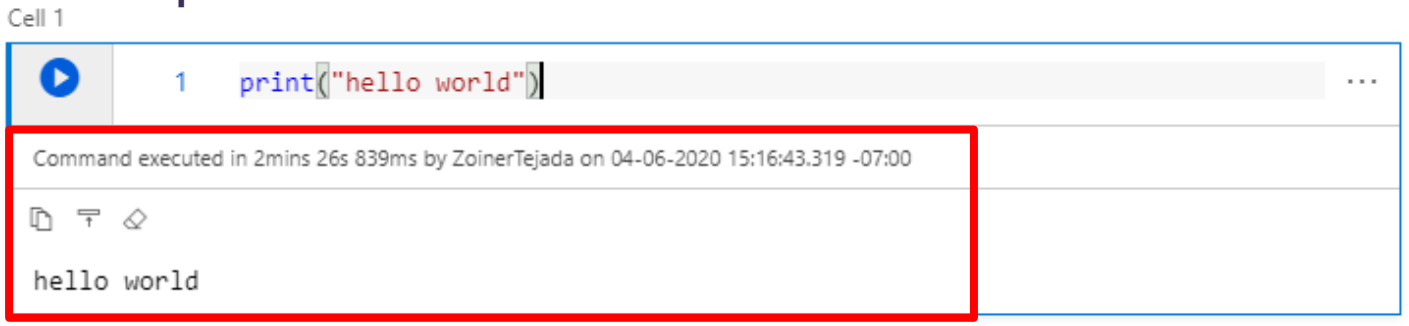



## Run current cell from toolbar from cell itself

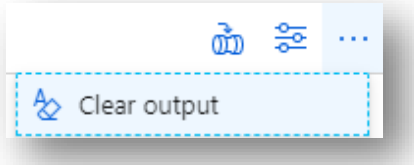


# Notebook capabilities – Cell output

## Cell output



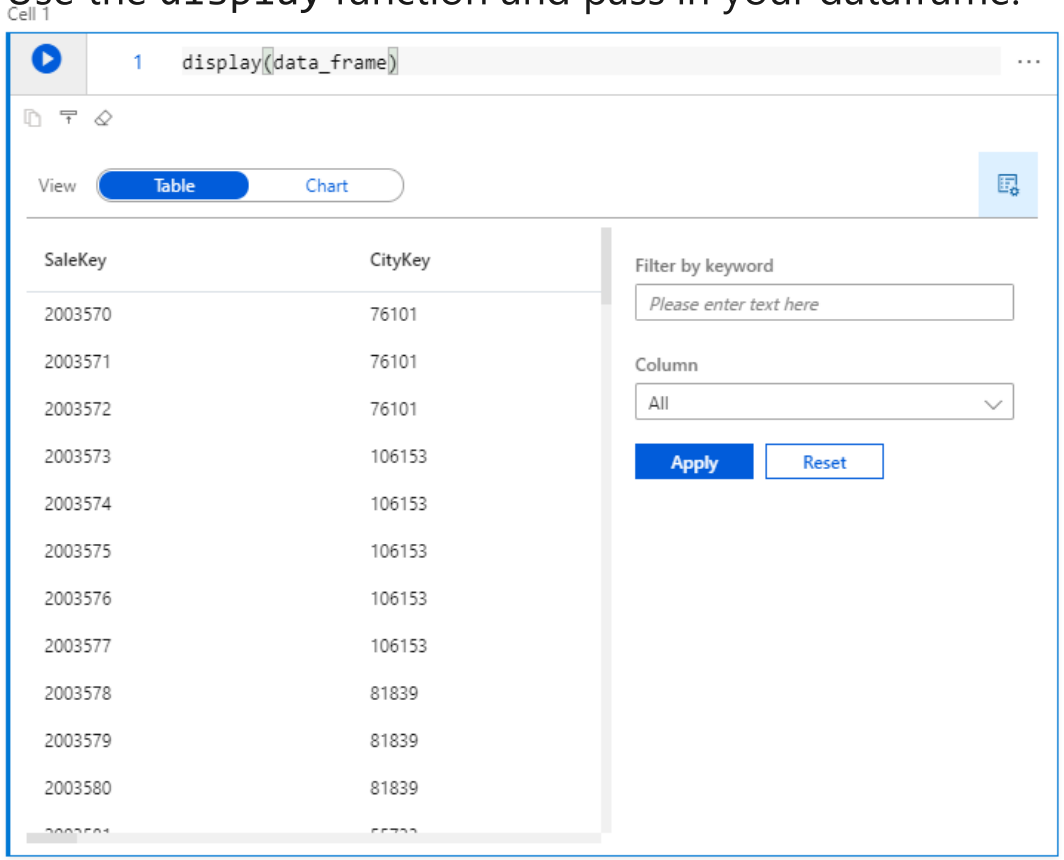

## Clear all cells output from toolbar

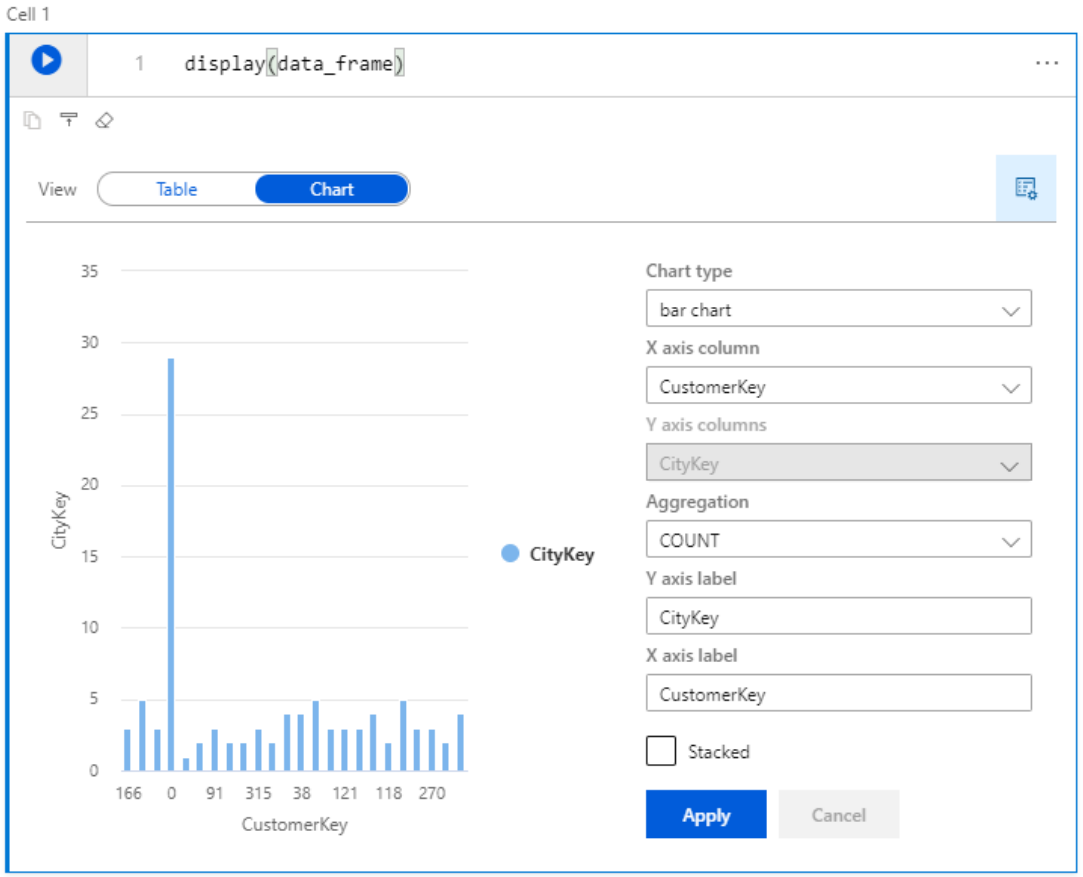


# Spark Pools – Notebook capabilities

## View DataFrame content as table or chart

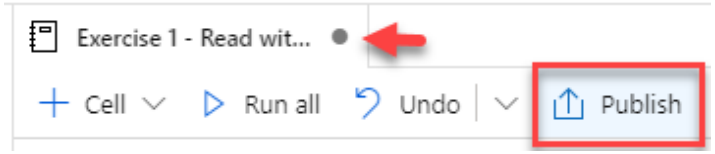
Use the `display` function and pass in your dataframe.



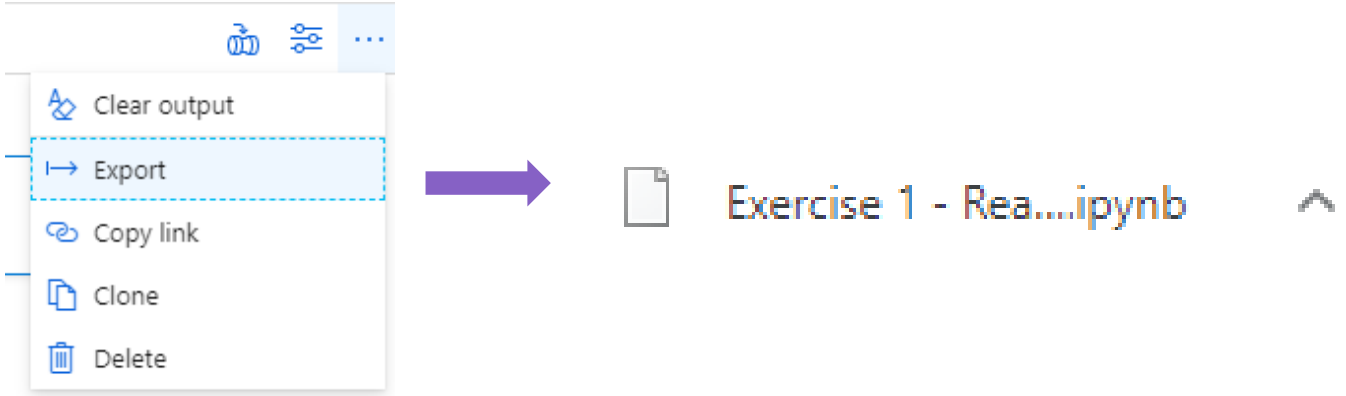



# Spark Pools – Notebook capabilities

## Saving (publishing)

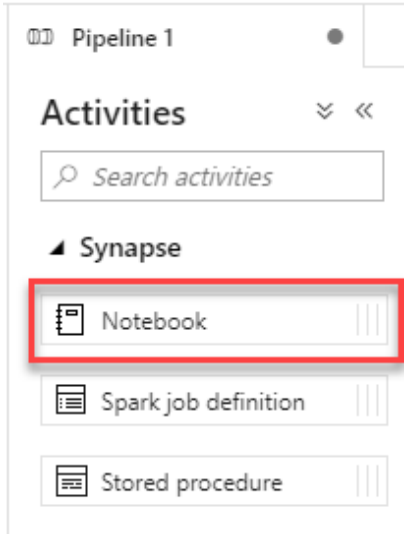



## Exporting (download the .ipynb)

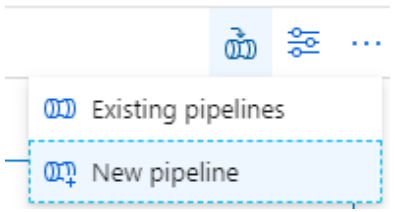


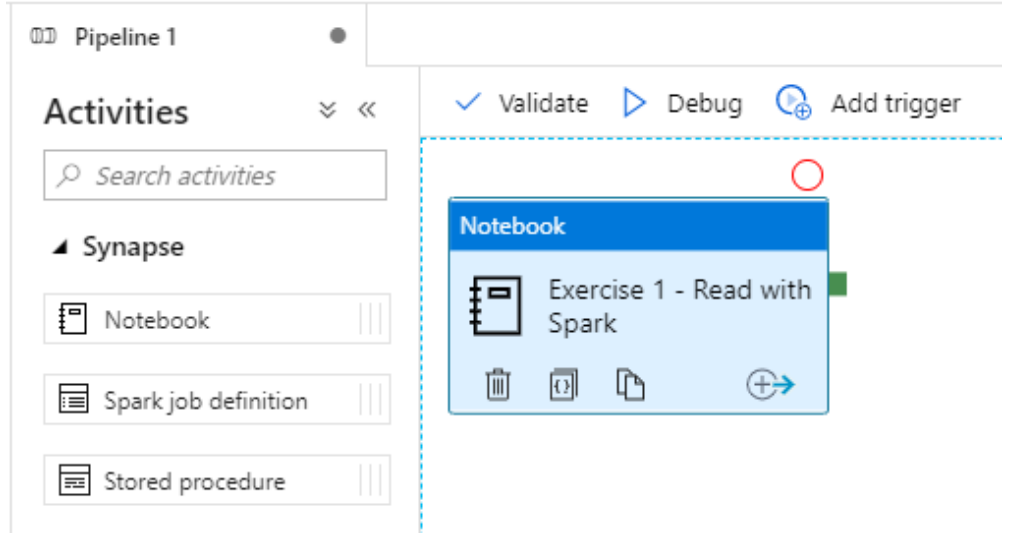
# Notebook capabilities – Use in pipelines

## Notebooks are pipeline activities



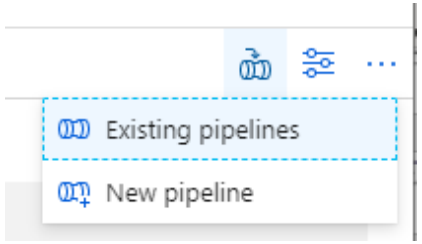

## Execute a notebook as a step in a new pipeline

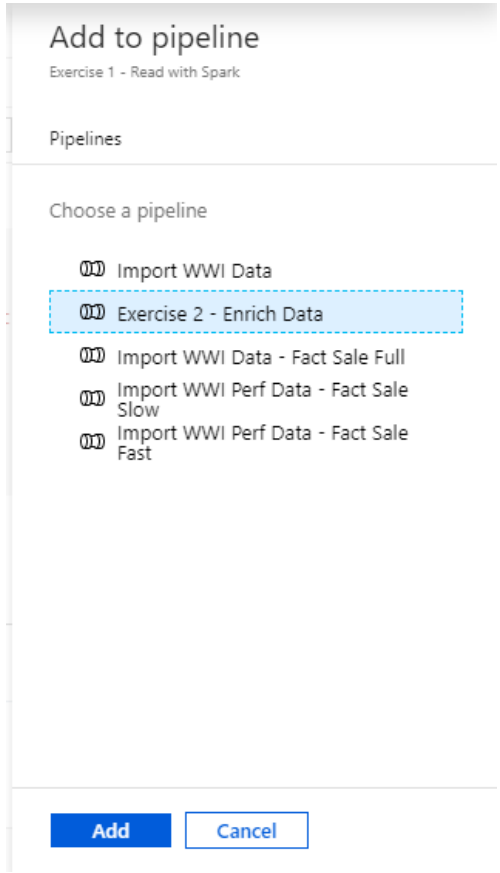


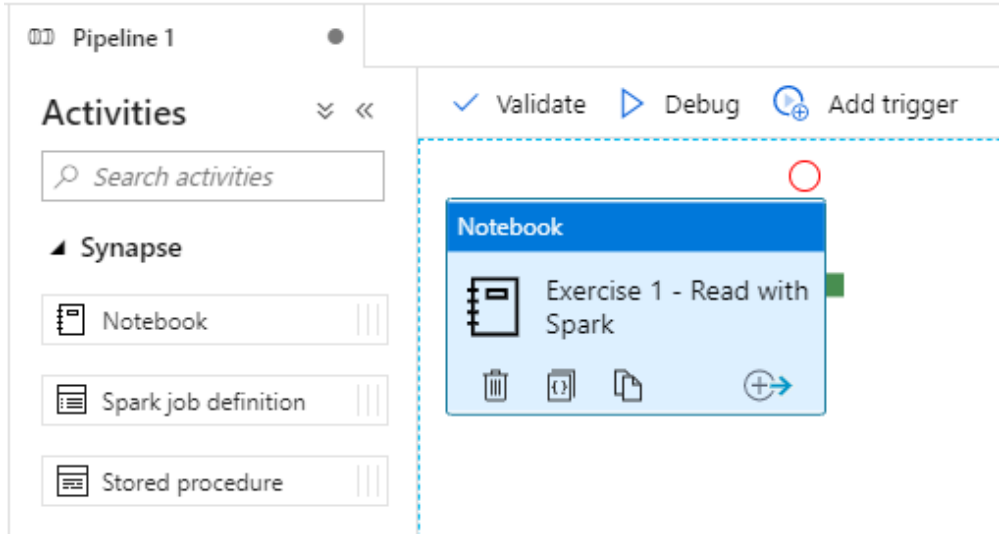



# Notebook capabilities – Use in pipelines (2)

## Execute a notebook as a step in an existing pipeline

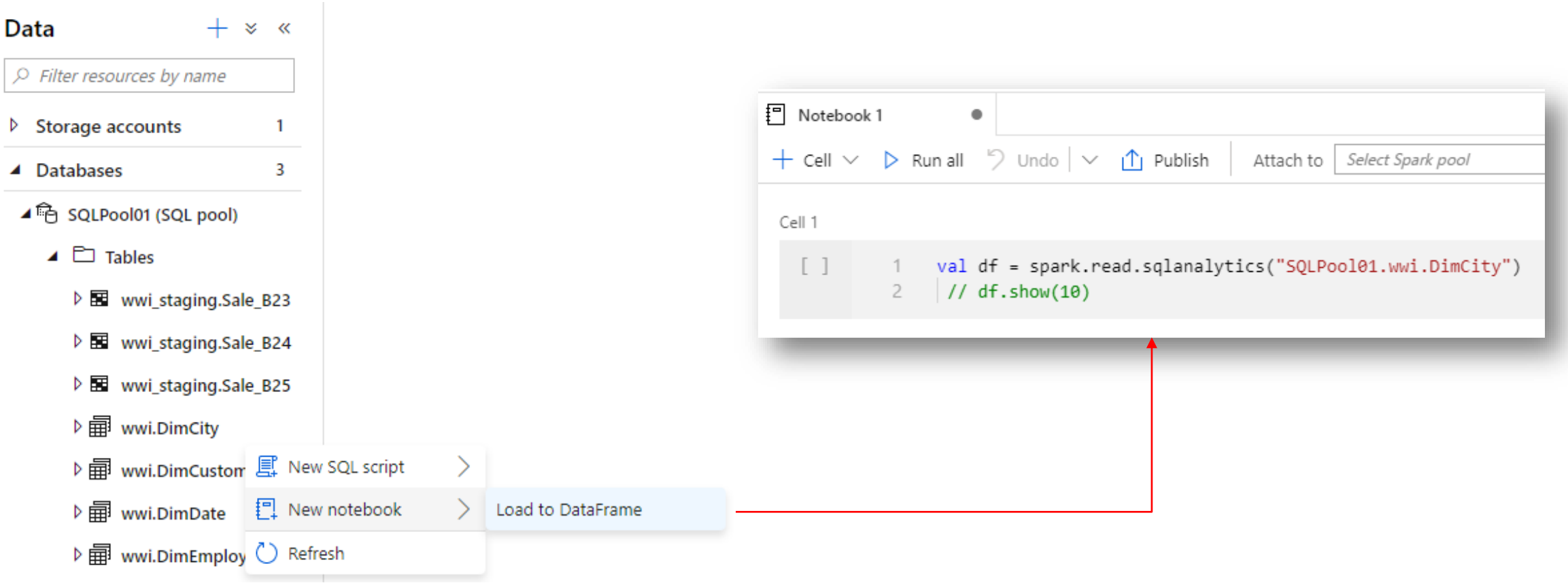


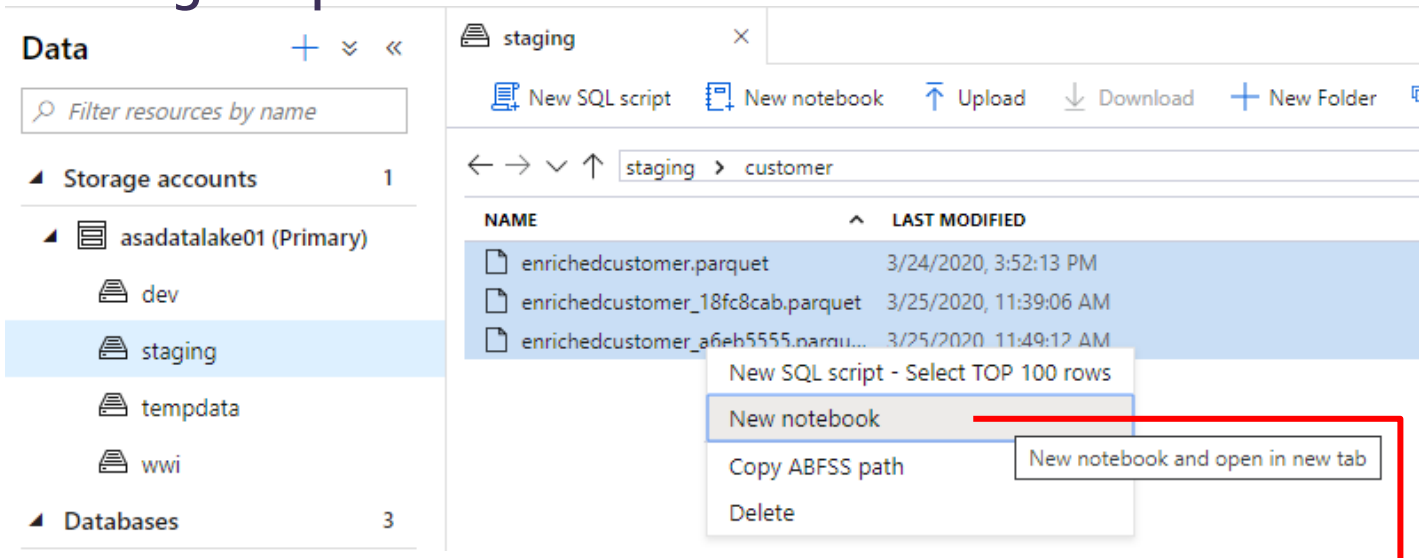
# Notebook capabilities – Querying SQL Pools

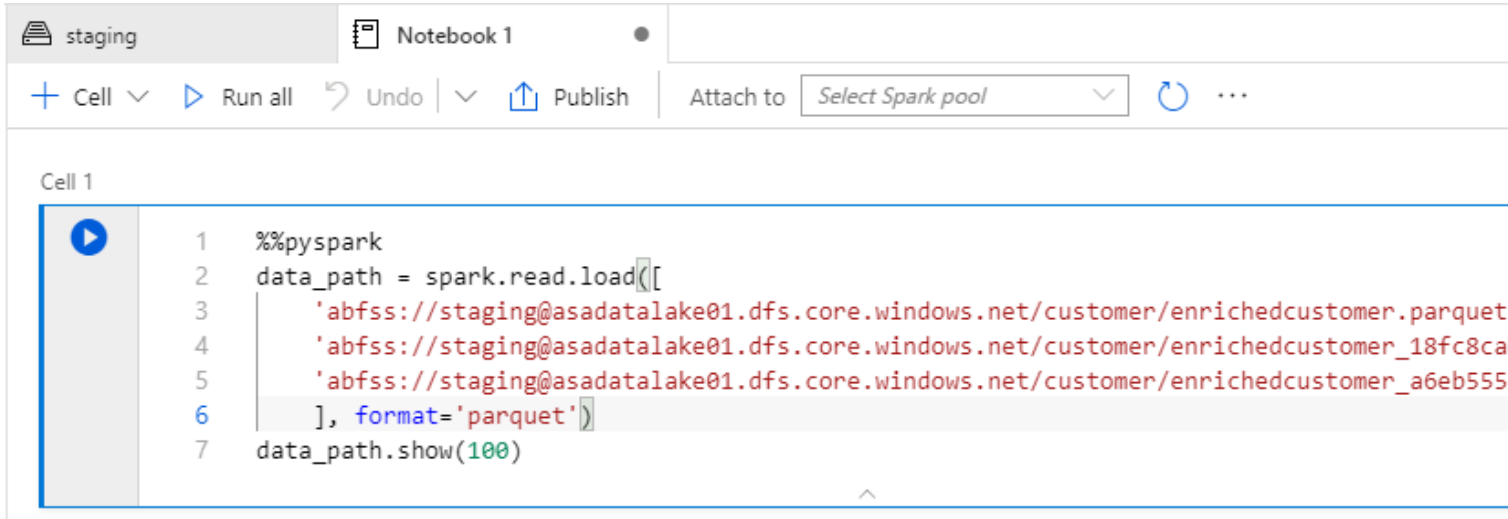
## Loading a SQL Pool table to a DataFrame



# Notebook capabilities – Querying Parquet Files

## Loading Parquet files to a DataFrame



# Pop Quiz 1

What types of library sources are supported in Apache Spark in Azure Synapse Analytics?

Maven

Spark

PyPi

# Pop Quiz 1

What types of library sources are supported in Apache Spark in Azure Synapse Analytics?

Maven

Spark

**PyPi**

# Machine Learning in Azure Synapse Analytics

# Microsoft Spark Utilities

## Overview

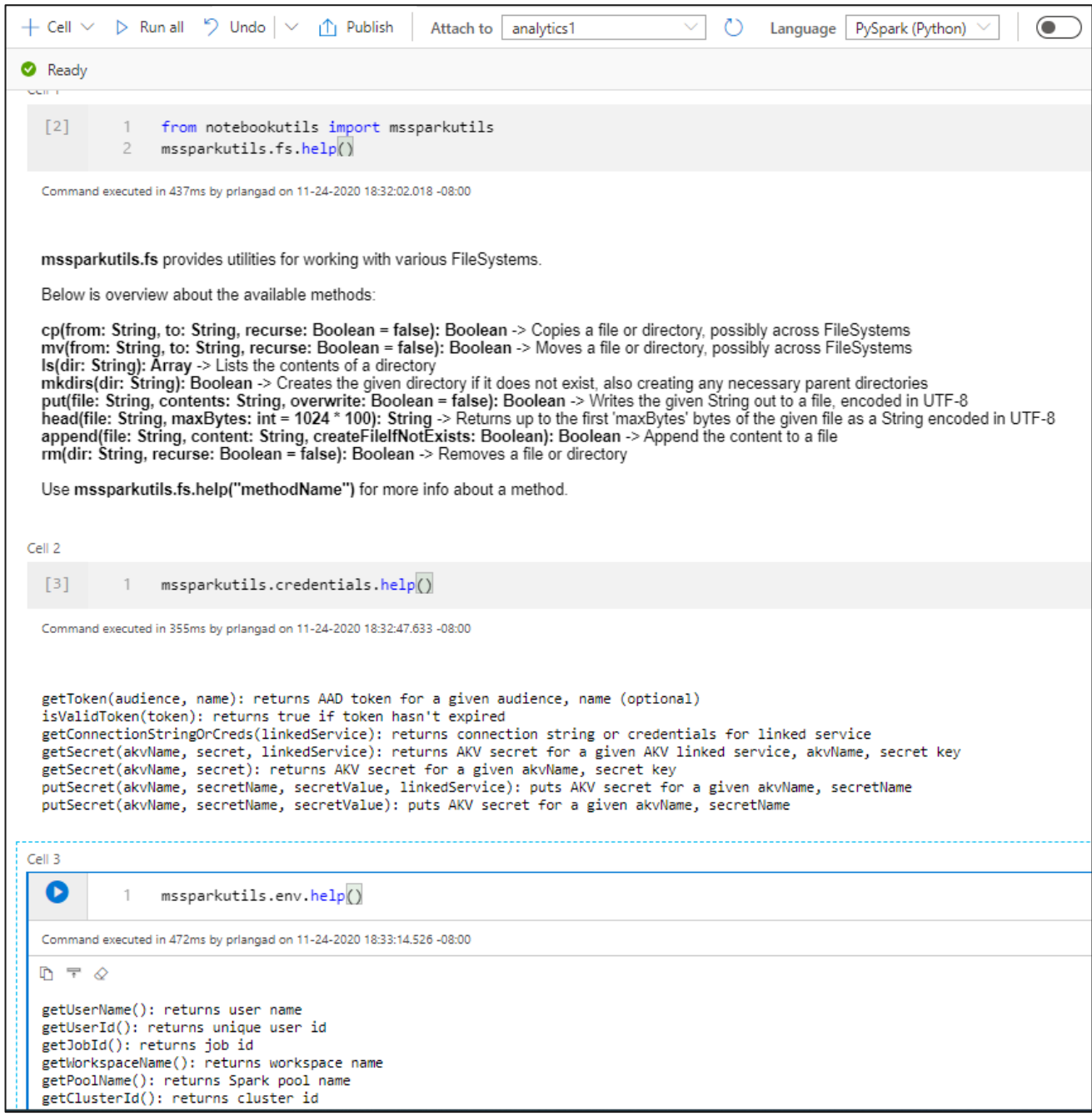
It provides utilities for working with file systems, including ADLS Gen2 and Azure Blob Storage.

## 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

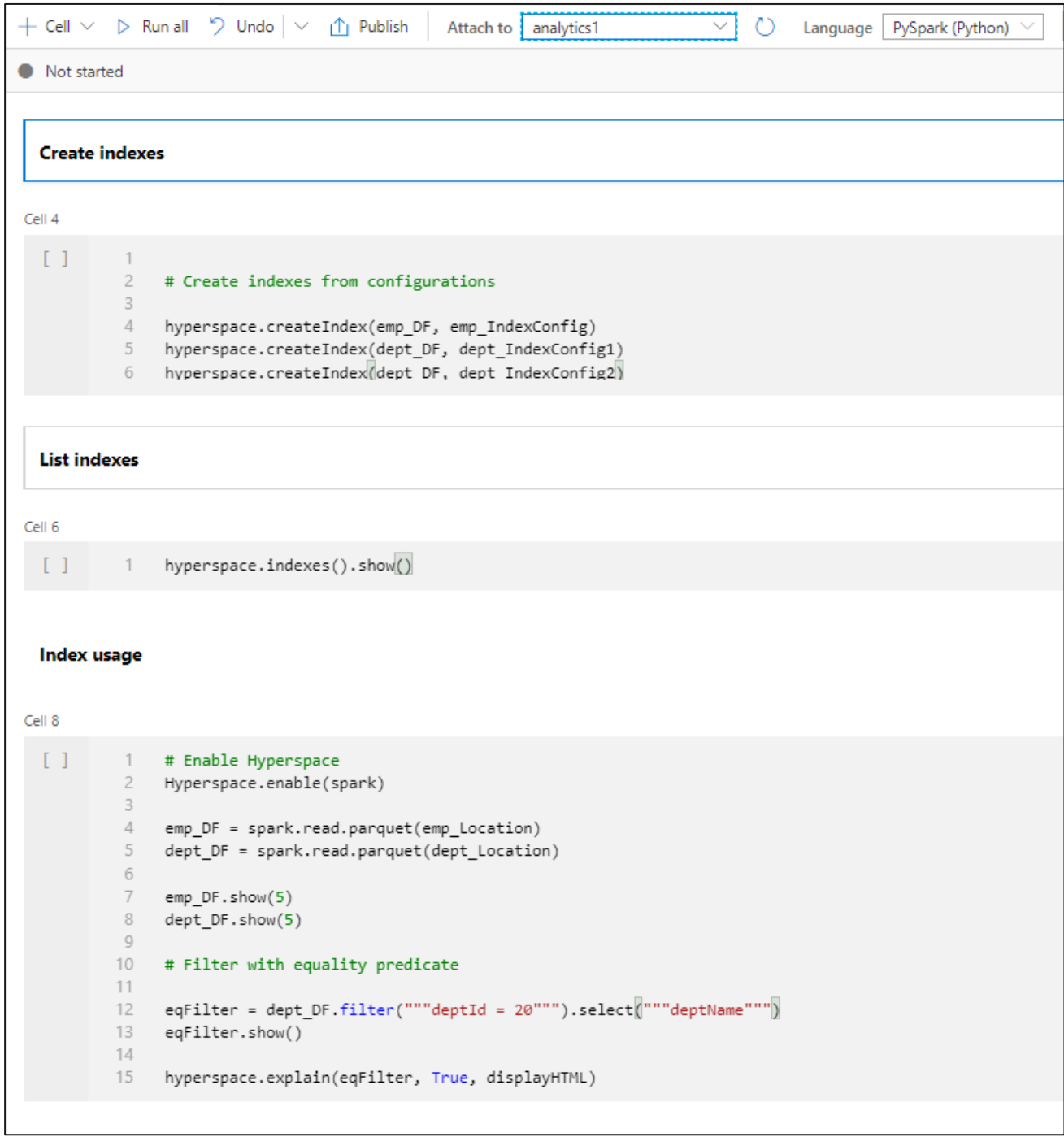
It helps accelerate your workloads or queries containing filters on predicates with high selectivity or a join that requires heavy shuffles.

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.

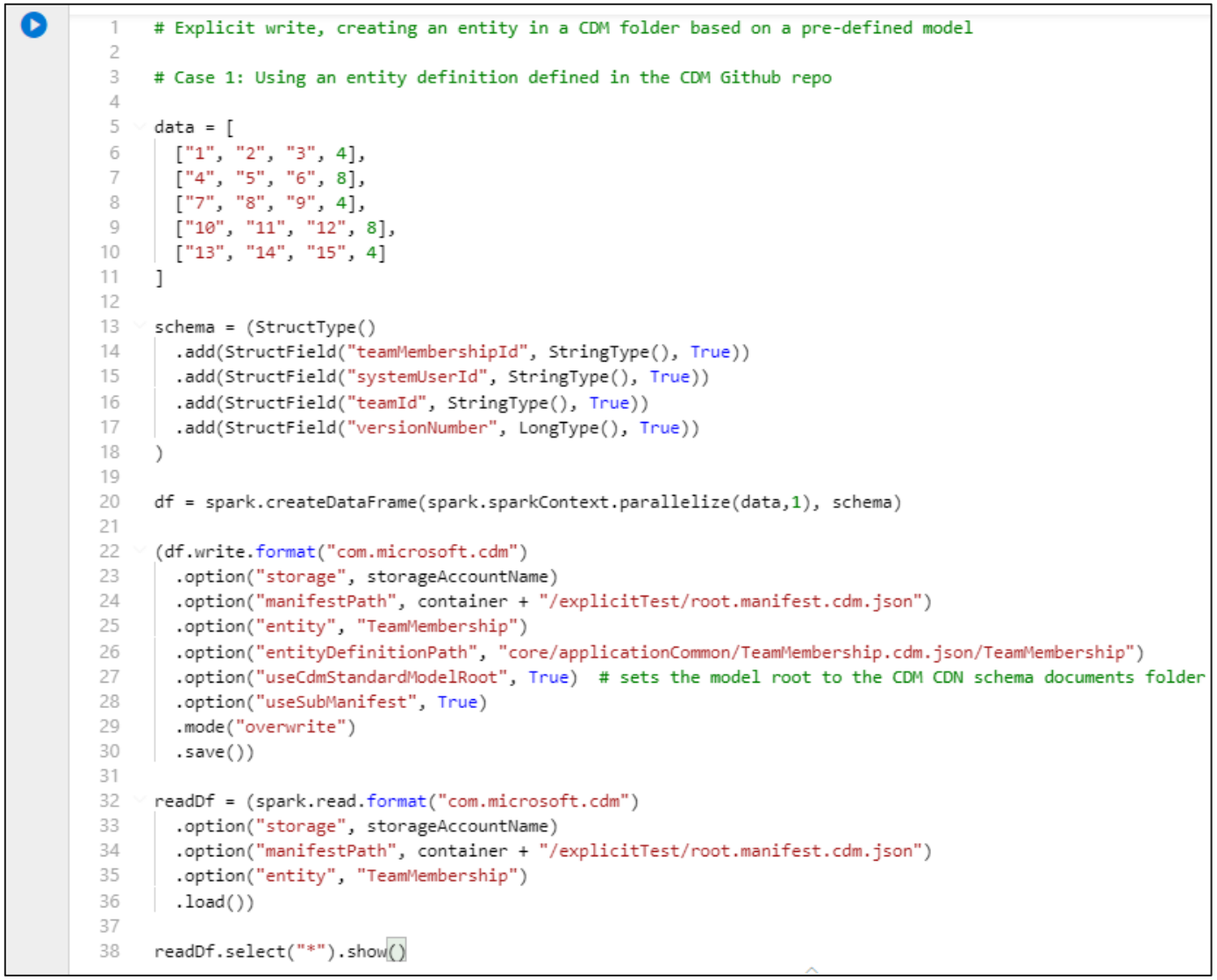
## Benefits

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()
```

# Examples of tools for data science used in Notebooks

**Core frameworks and tools**





Machine Learning & Deep Learning





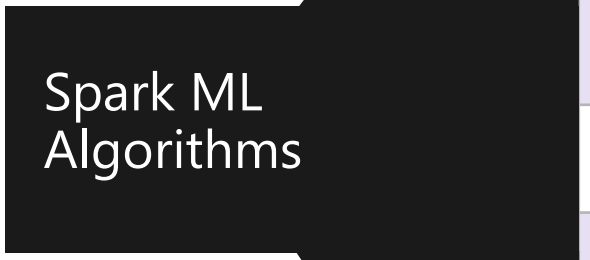
Keras

PyTorch

Visualization




# Spark ML Algorithms

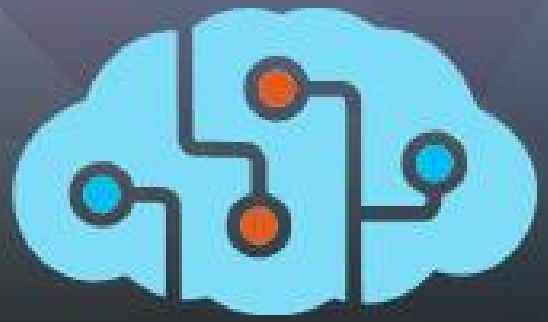



| | |
|---|---|
| Classification and Regression | • Linear Models (SVMs, logistic regression, linear regression)<br>• Naïve Bayes<br>• Decision Trees<br>• Ensembles of trees (Random Forest, Gradient-Boosted Trees)<br>• Isotonic regression |
| Clustering | • k-means and streaming k-means<br>• Gaussian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

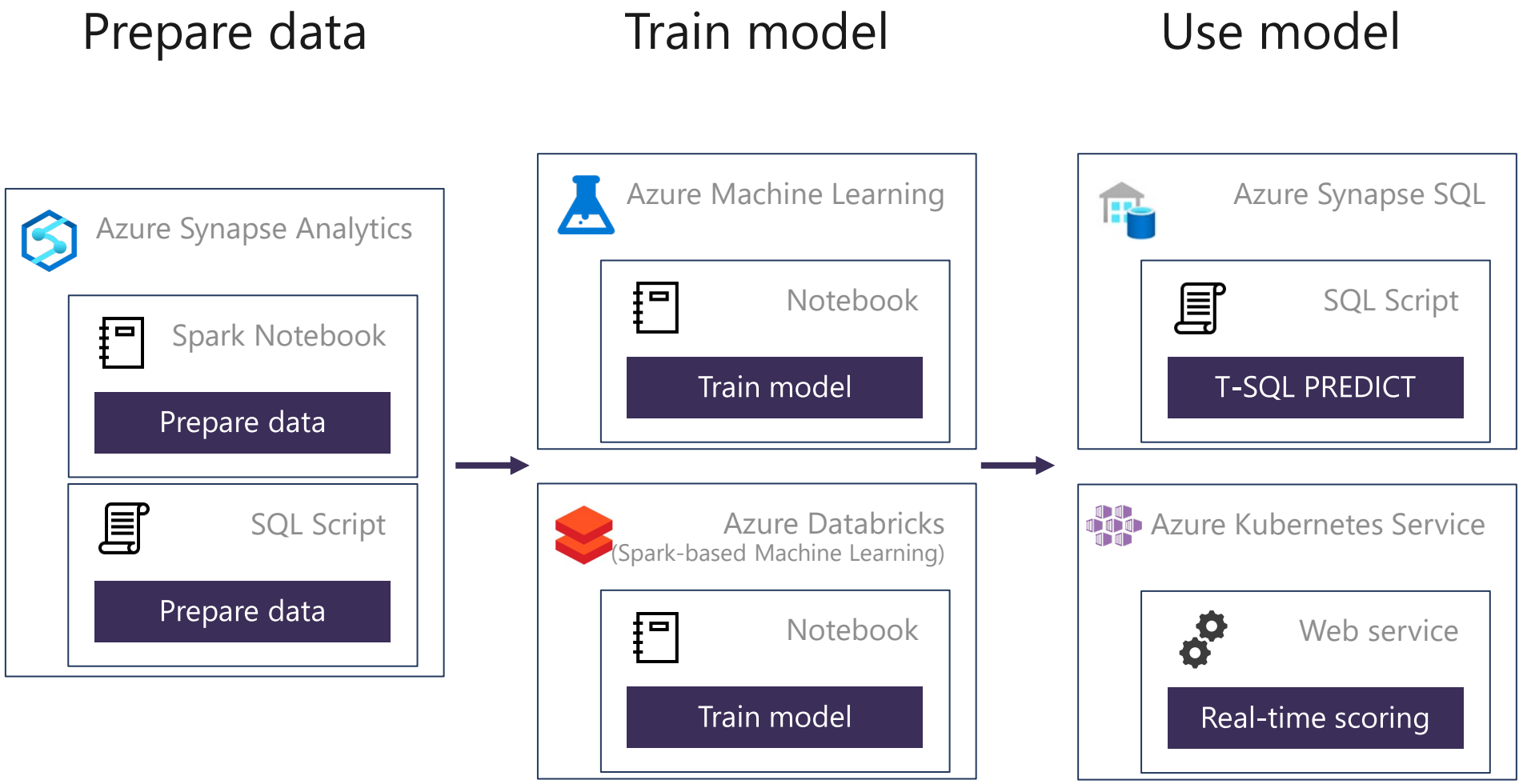
Microservice Orchestration

Multilingual Binding Generation

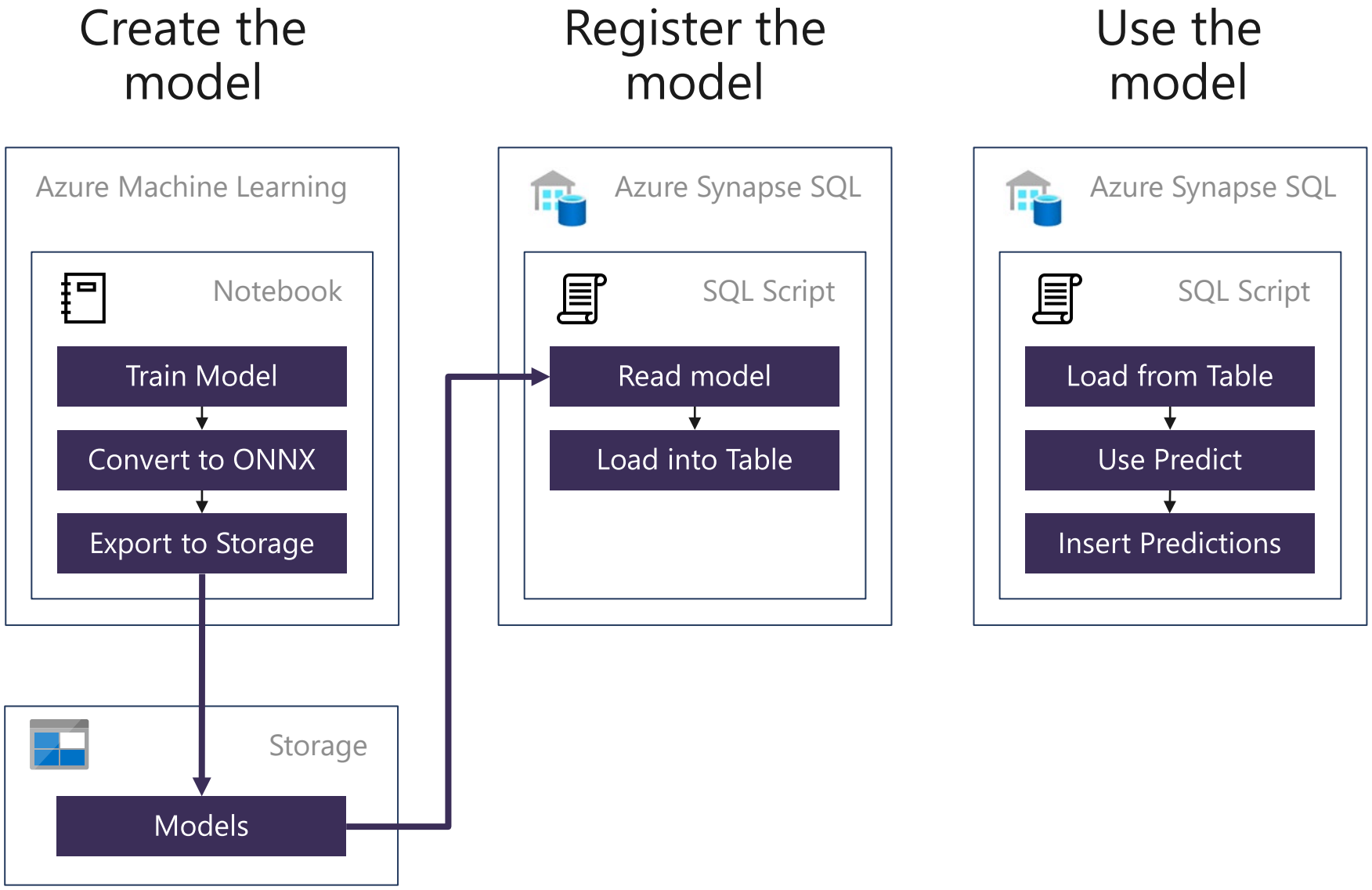
www.aka.ms/spark

Azure/mmlspark

# Synapse Analytics and the Machine Learning Process



# Making predictions with T-SQL



# 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.

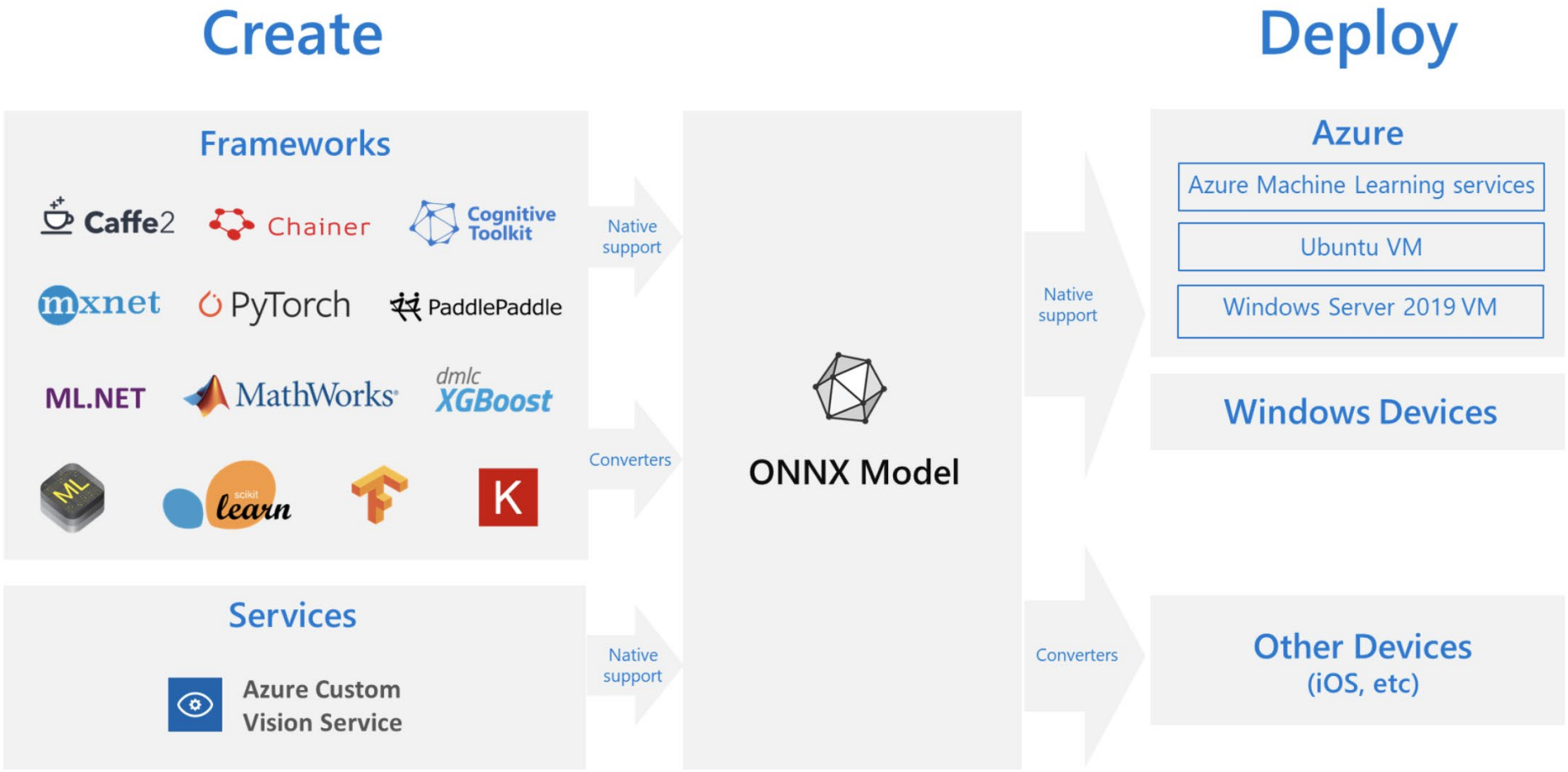
**Syntax**:

```
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**:

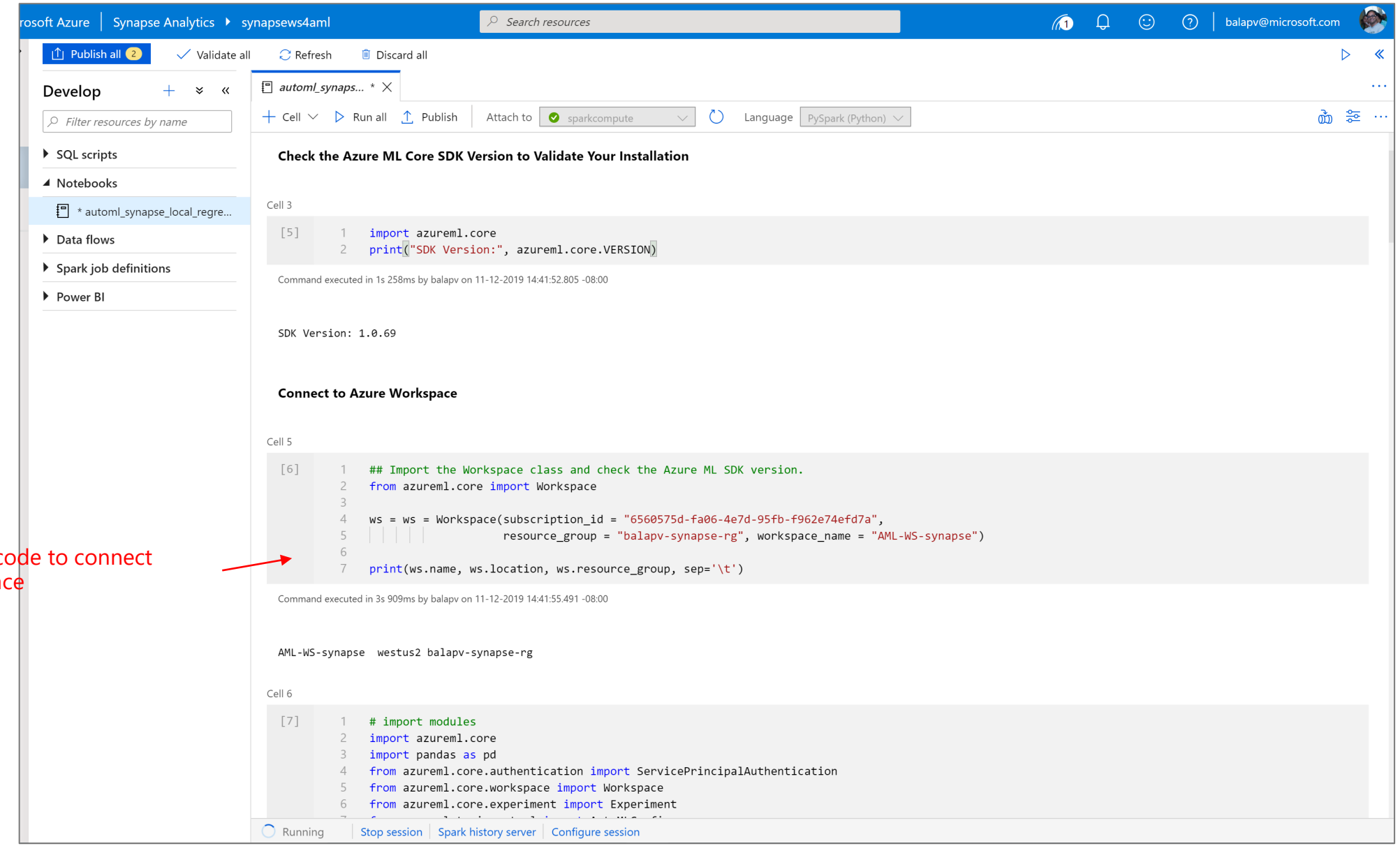
```
DECLARE @model varbinary(max) = (SELECT Model FROM Models WHERE Id = <>);
SELECT d.*, p.Score
FROM PREDICT(MODEL = @model,
DATA = dbo.mytable AS d) WITH (Score float) AS p;
```

# What is ONNX?



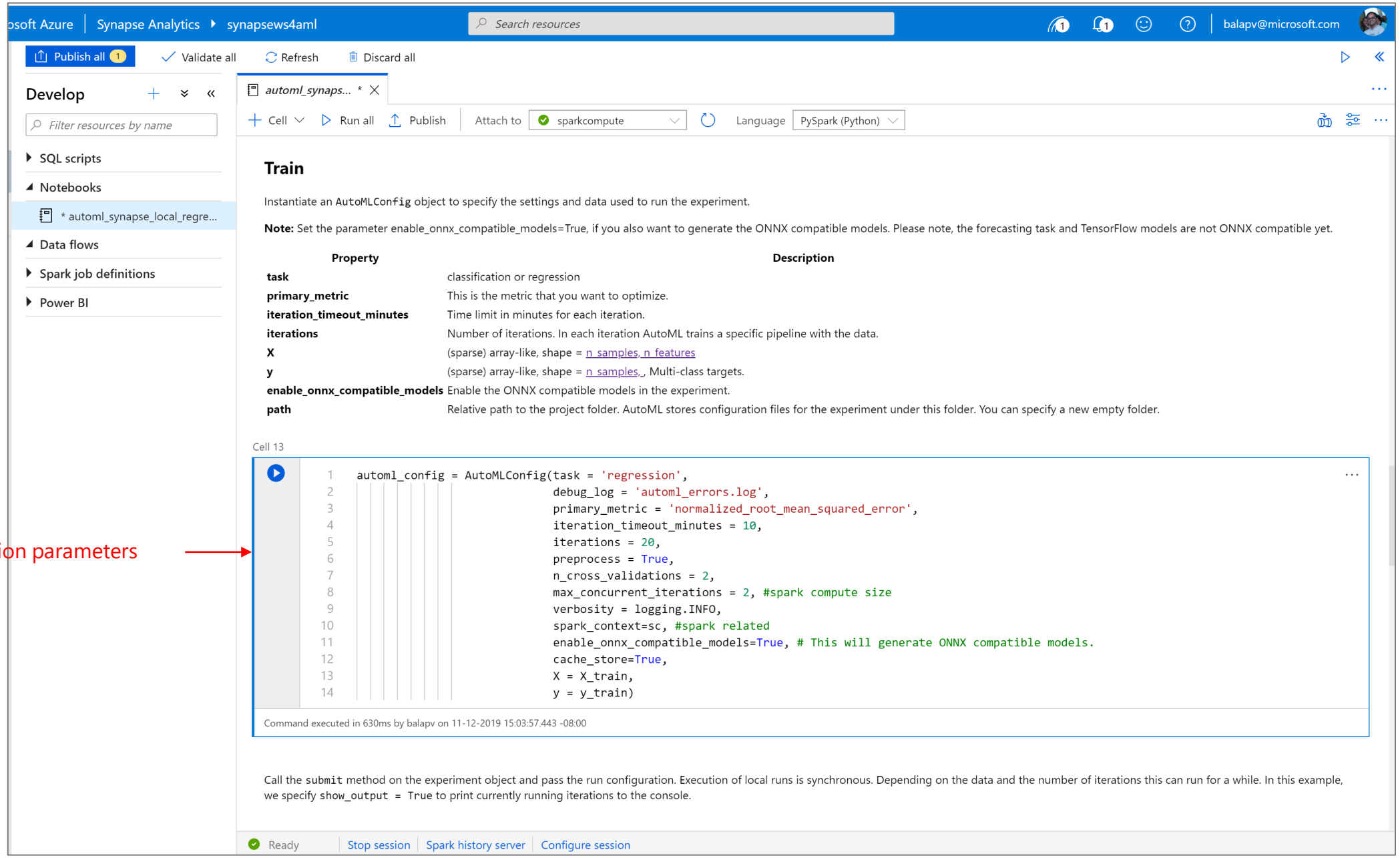

ONNX flow diagram showing training, converters, and deployment

# Synapse Notebook: Connect to AML workspace



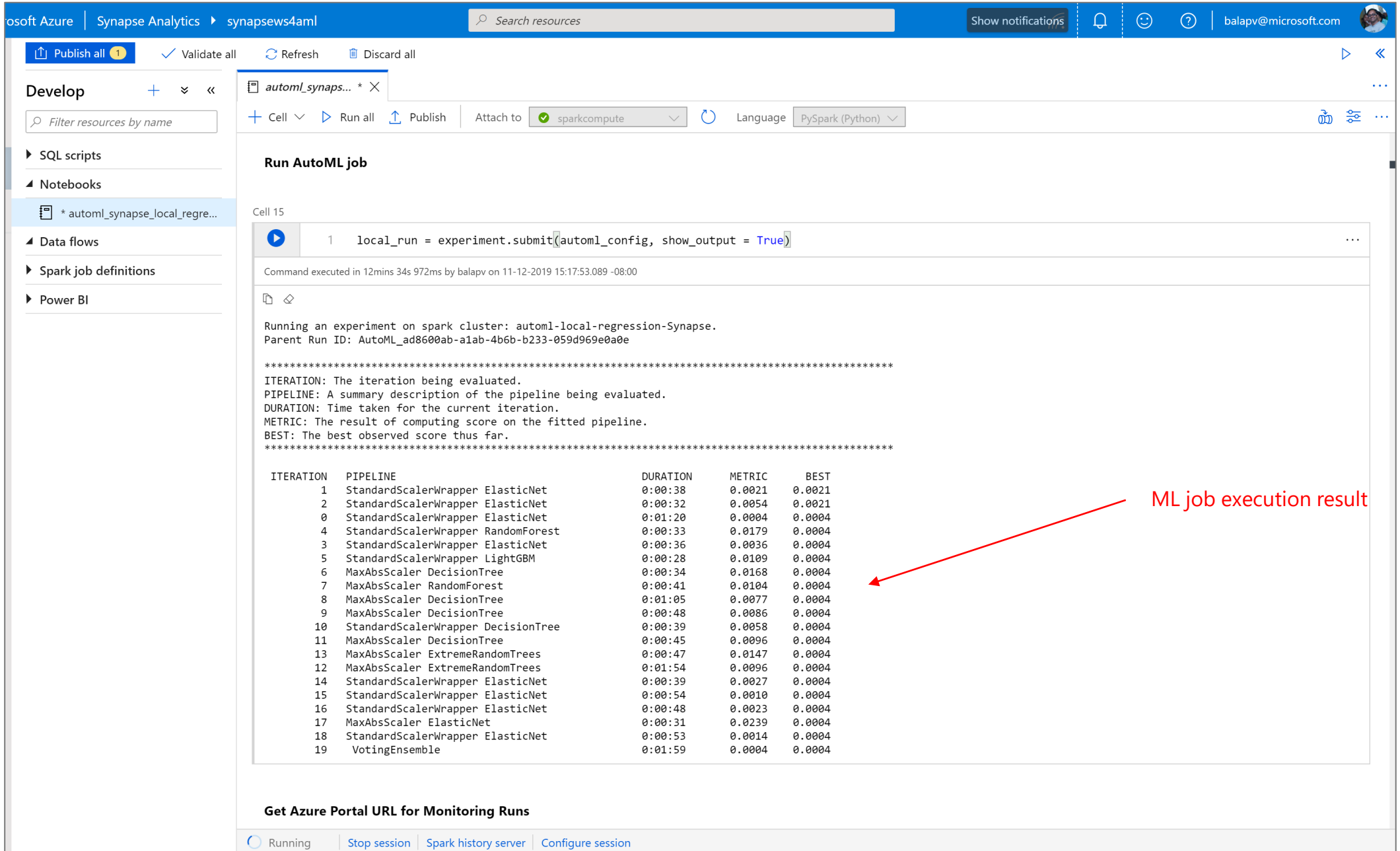

Simple code to connect workspace

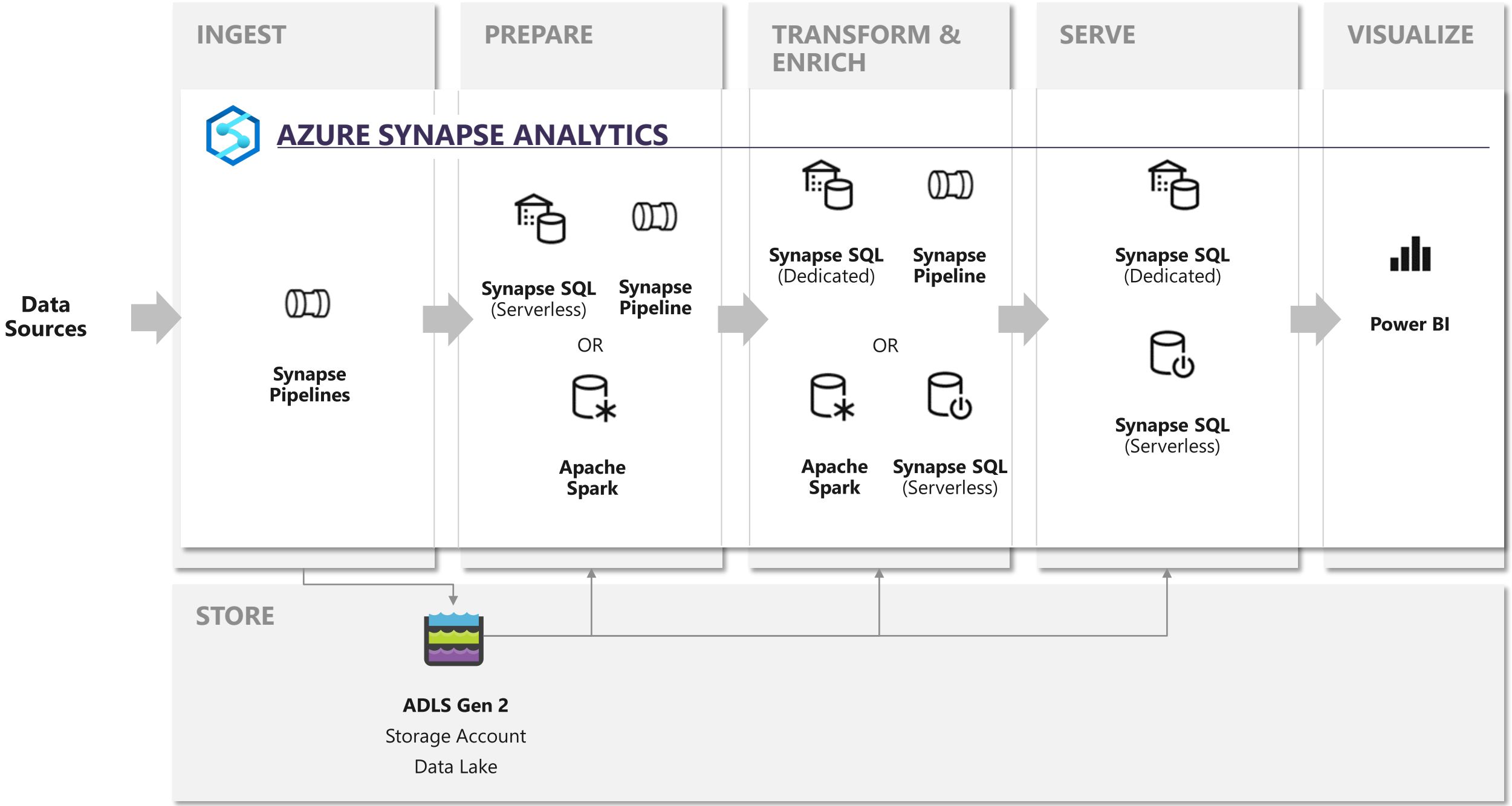
# Synapse Notebook: Configure AML job to run on Synapse




Configuration parameters

# Synapse Notebook: Run AML job




INGEST
PREPARE
TRANSFORM & ENRICH
SERVE
VISUALIZE
AZURE SYNAPSE ANALYTICS
Data Sources
Synapse Pipelines
Synapse SQL (Serverless)
Synapse Pipeline
OR
Apache Spark
Synapse SQL (Dedicated)
Synapse Pipeline
OR
Apache Spark
Synapse SQL (Serverless)
Synapse SQL (Dedicated)
Synapse SQL (Serverless)
Power BI
STORE
ADLS Gen 2
Storage Account
Data Lake

Microsoft Azure Solutions Business

# Thank you

# ⌛ Break Time!

We are at break for the next 15 minutes.

Please feel free to step away and we will see you back in your **Table Group call** when break ends.

| | |
|---|---|
| 7:00-7:05 | Welcome |
| 7:05-7:45 | Spark for Data Science |
| 7:45-8:00 | Break |
| 8:00-9:00 | Build Hands-on: Machine Learning |
| 9:00-10:00 | Activity: Model Implementation (Predict) |
| 10:00-10:15 | Break |
| 10:15-11:00 | Build Hands-on: Spark Machine Learning |
| 11:00-12:00 | Break |
| 12:00-12:30 | Monitor & Manage |
| 12:30-13:00 | Activity: Monitor & Manage |
| 13:00-14:00 | Build Hands-on: Monitoring |
| 14:00-14:15 | Closing |

*If at anytime you require assistance, please send a message to the "Need help – Ask Here" channel in the Microsoft Teams site for this event*

# Build Hands-on Lab: Machine Learning

OUTCOMES

As a result of participating in this lab, you will be better able to **create several machine learning models and use them to make predictions using the T-SQL *Predict* statement**.

ACTIVITY

60 minutes | Independent | Table Group Channel & CloudLabs Environment

Independently review and complete Exercises 1-2 in the Lab 6 Guide.

Login to your **CloudLabs environment** and work through a set of tasks you would typically follow in work associated with the model, manage & secure part of your process.

The Lab Guide is available in the 'Files' tab of the General channel. Engage your Table Group call for support and troubleshooting.

# Breakout Activity:
# Model Implementation (Predict)

OUTCOMES

As a result of participating in this activity, you will better be able to **implement a process that trains a machine learning model, registers it for use and utilize the model for scoring using the Predict statement**.

ACTIVITY

60 minutes | Learning Adviser | Table Group Channel

As a team review the challenges posed and decide how to address them.

Choose a "driver" who will open their **CloudLabs** environment and then complete the challenges as a team using the driver's Azure Synapse Analytics Studio.

The challenges cover model training, model registration, and scoring. Team results are collected in the event leaderboard.

# ⌛ Break Time!

We are at break for the next 15 minutes.

Please feel free to step away and we will see you back in your **Table Group call** when break ends.

| | |
|---|---|
| 7:00-7:05 | Welcome |
| 7:05-7:45 | Spark for Data Science |
| 7:45-8:00 | Break |
| 8:00-9:00 | Build Hands-on:<br>Machine Learning |
| 9:00-10:00 | Activity: Model Implementation (Predict) |
| 10:00-10:15 | Break |
| 10:15-11:00 | Build Hands-on:<br>Spark Machine Learning |
| 11:00-12:00 | Break |
| 12:00-12:30 | Monitor & Manage |
| 12:30-13:00 | Activity: Monitor & Manage |
| 13:00-14:00 | Build Hands-on:<br>Monitoring |
| 14:00-14:15 | Closing |

*If at anytime you require assistance, please send a message to the "Need help – Ask Here" channel in the Microsoft Teams site for this event*

# Build Hands-on Lab:
# Spark Machine Learning

OUTCOMES

As a result of participating in this lab, you will be better able to **train and deploy a recommender ML model**.

ACTIVITY

45 minutes | Independent | Table Group Channel & CloudLabs Environment

Independently review and complete the exercise in the Lab 7 Guide.

Login to your **CloudLabs environment** and work through a set of tasks you would typically follow in work associated with ingesting and integrating your customer's data.

The Lab Guide is available in the 'Files' tab of the General channel. Engage your Table Group call for support and troubleshooting.

# ⌛ Break Time!

We are at break for the next 60 minutes.

Please feel free to step away and we will see you back in the **main call** when break ends.

| | |
|---|---|
| 7:00-7:05 | Welcome |
| 7:05-7:45 | Spark for Data Science |
| 7:45-8:00 | Break |
| 8:00-9:00 | Build Hands-on: Machine Learning |
| 9:00-10:00 | Activity: Model Implementation (Predict) |
| 10:00-10:15 | Break |
| 10:15-11:00 | Build Hands-on: Spark Machine Learning |
| 11:00-12:00 | Break |
| 12:00-12:30 | Monitor & Manage |
| 12:30-13:00 | Activity: Monitor & Manage |
| 13:00-14:00 | Build Hands-on: Monitoring |
| 14:00-14:15 | Closing |

*If at anytime you require assistance, please send a message to the "Need help – Ask Here" channel in the Microsoft Teams site for this event*

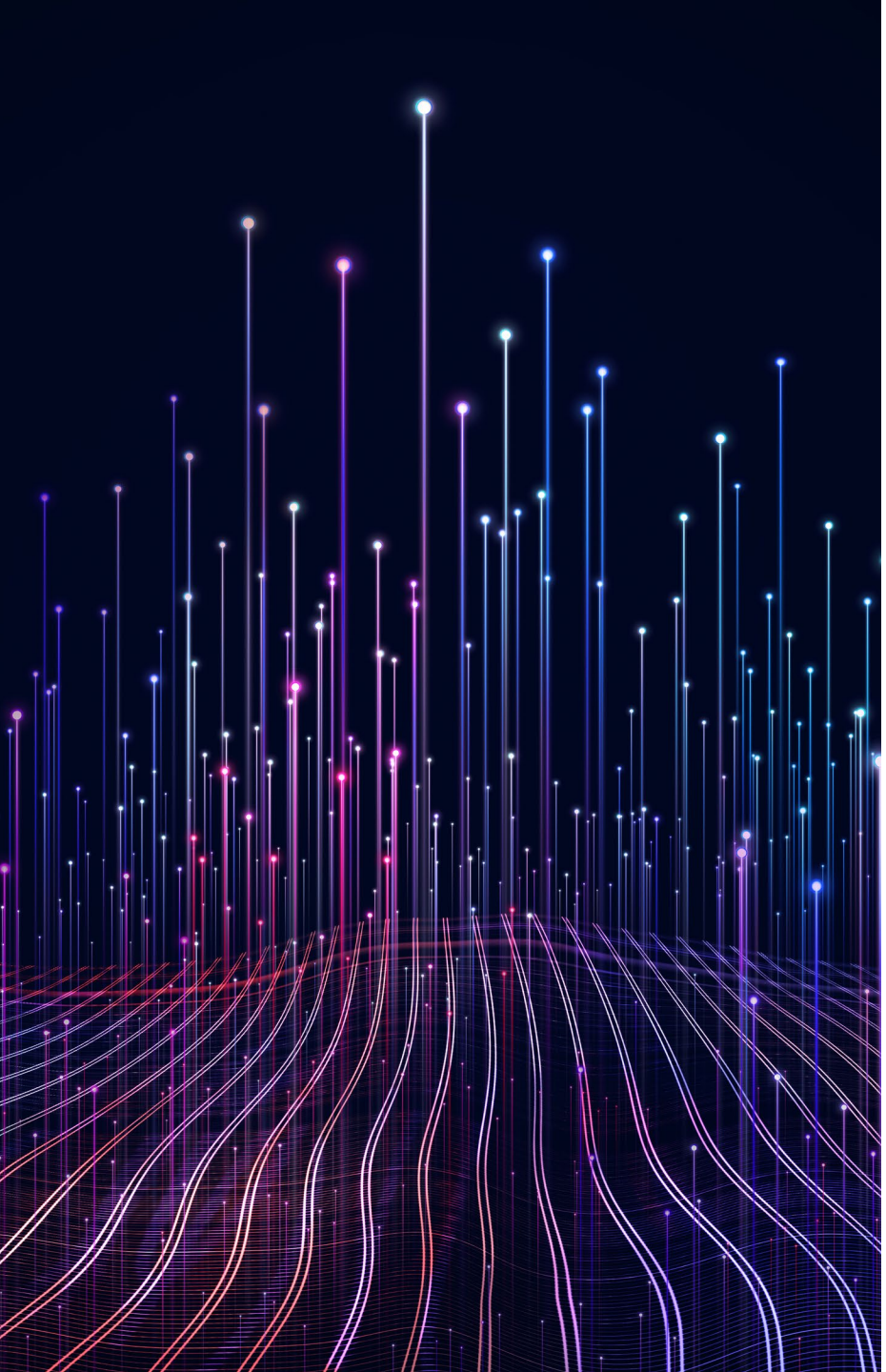
# Welcome Back!

| | | |
|---|---|---|
| 7:00-7:05 | Welcome | |
| 7:05-7:45 | Spark for Data Science | Main Call |
| 7:45-8:00 | Break | |
| 8:00-9:00 | Build Hands-on:<br>Machine Learning | Table Group Call |
| 9:00-10:00 | Activity: Model Implementation (Predict) | |
| 10:00-10:15 | Break | Main Call |
| 10:15-11:00 | Build Hands-on:<br>Spark Machine Learning | |
| 11:00-12:00 | Break | Table Group Call |
| 12:00-12:30 | Monitor & Manage | Main Call |
| 12:30-13:00 | Activity: Monitor & Manage | |
| 13:00-14:00 | Build Hands-on:<br>Monitoring | Table Group Call |
| 14:00-14:15 | Closing | Main Call |

- Presentation/ Whole Group
- Lab
- Activity/ Discussion/ Group Work
- Announcements

Microsoft Azure Solutions Business

# Monitor & Manage



# Agenda

1 **Monitoring**

Monitor resources within Synapse Studio and the Azure Portal.

2 **Manage**

Manage resources for optimal utilization and performance.

# Discussion..

How has the role of a DBA changed in the era of managed Data-Warehouses-as-a-Service?

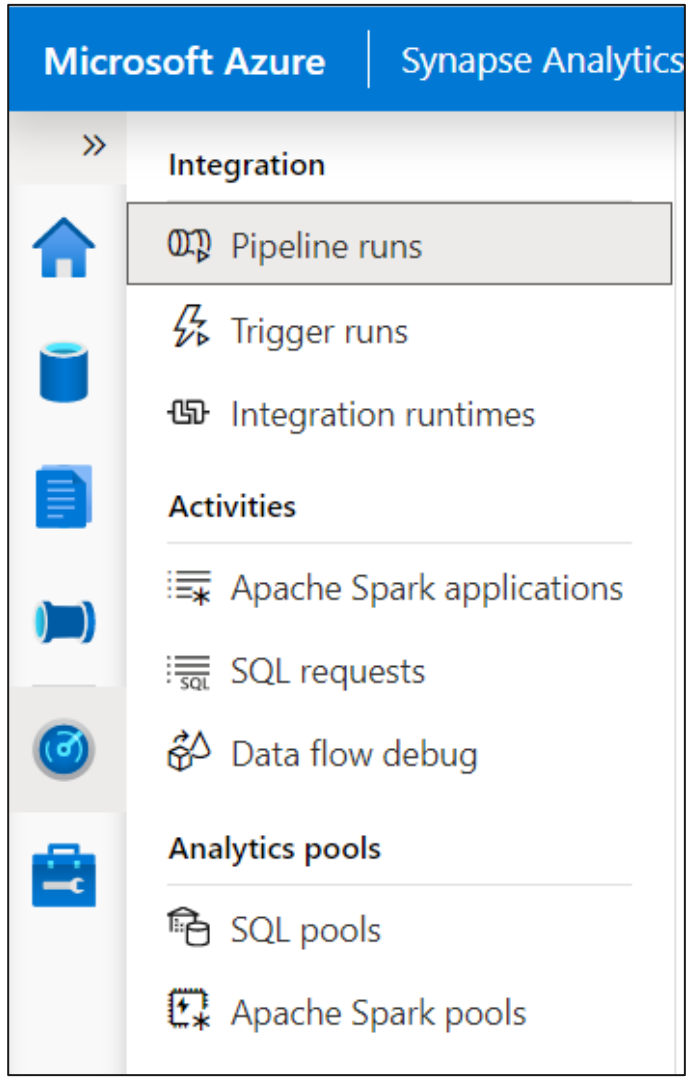
# Monitor Hub

## Overview

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

## Benefits

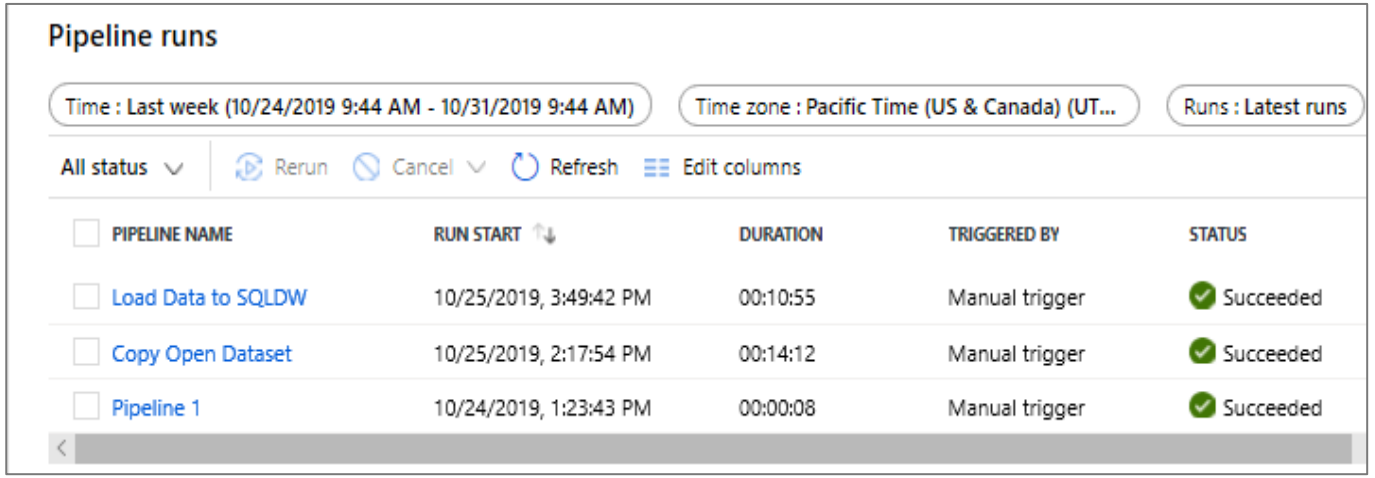
Offers additional filters to monitor specific activities or orchestration



# Monitor Hub - Integration

## Overview

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

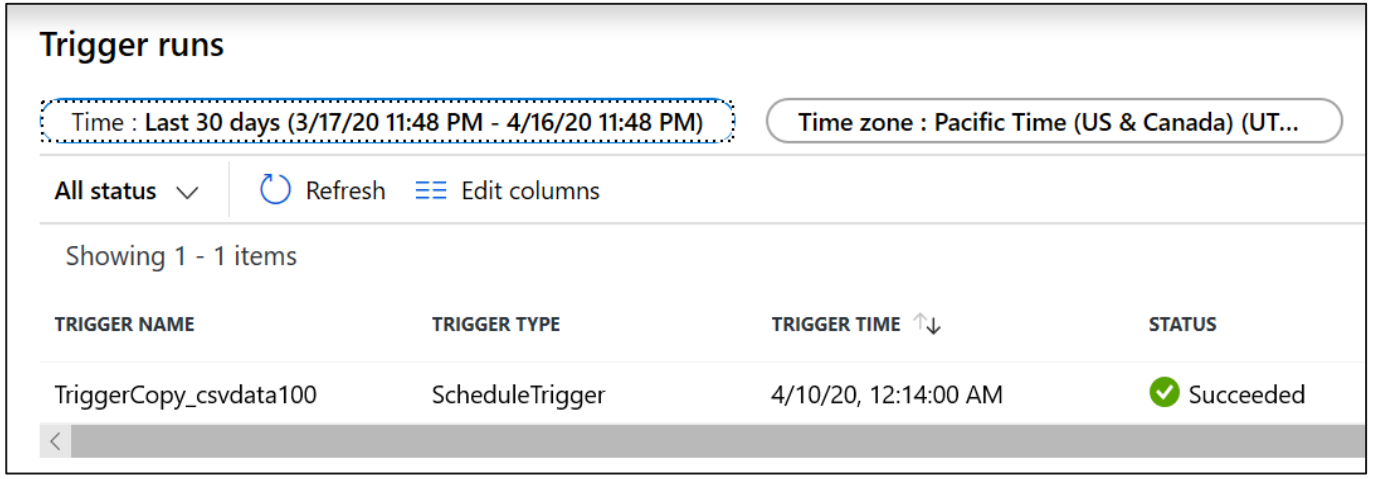



## Benefits

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

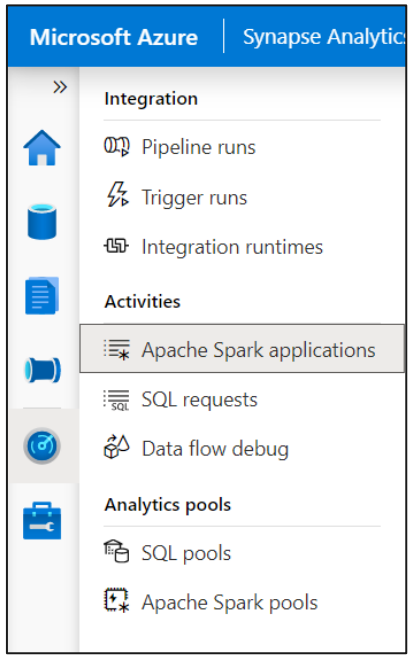
Monitor Spark pools, Spark applications for the status of activities

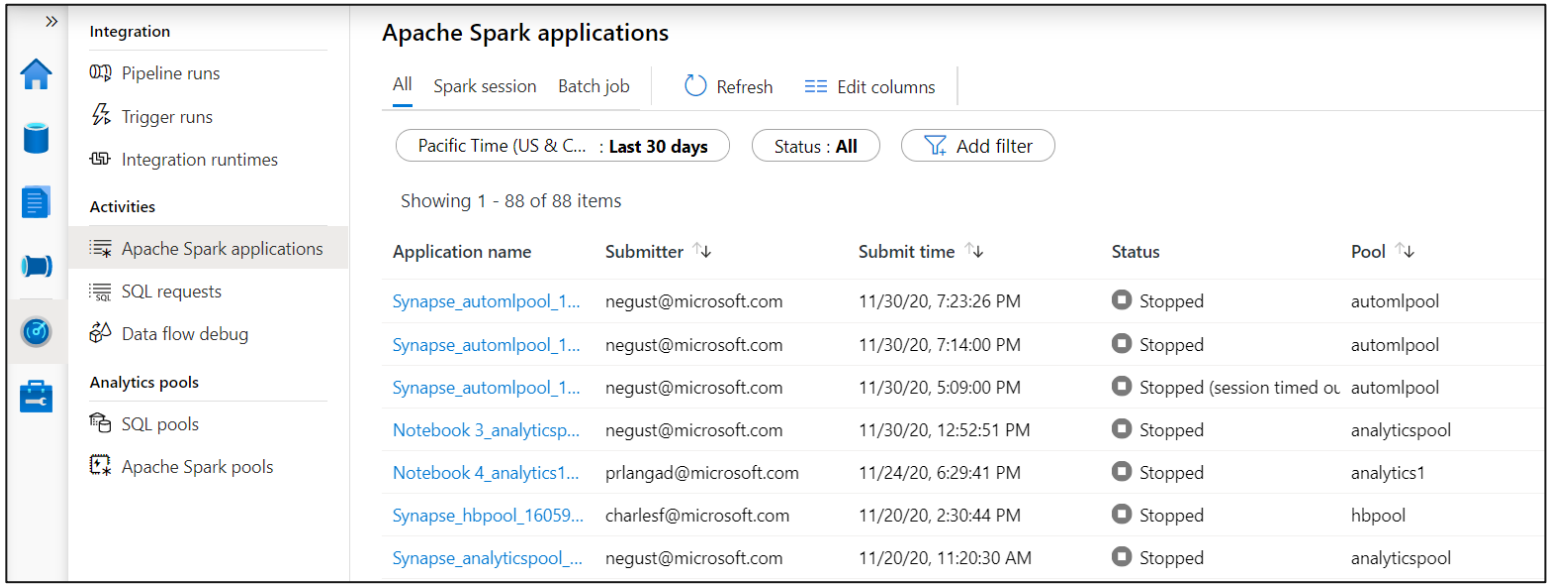
## Benefits

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

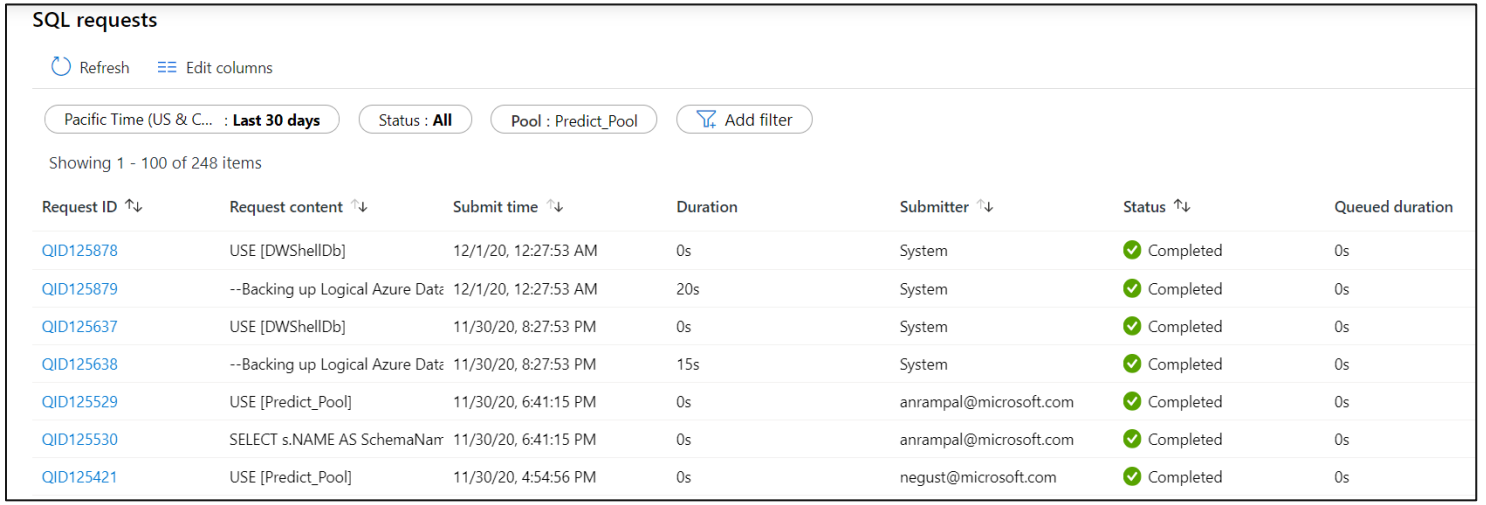
## Benefits

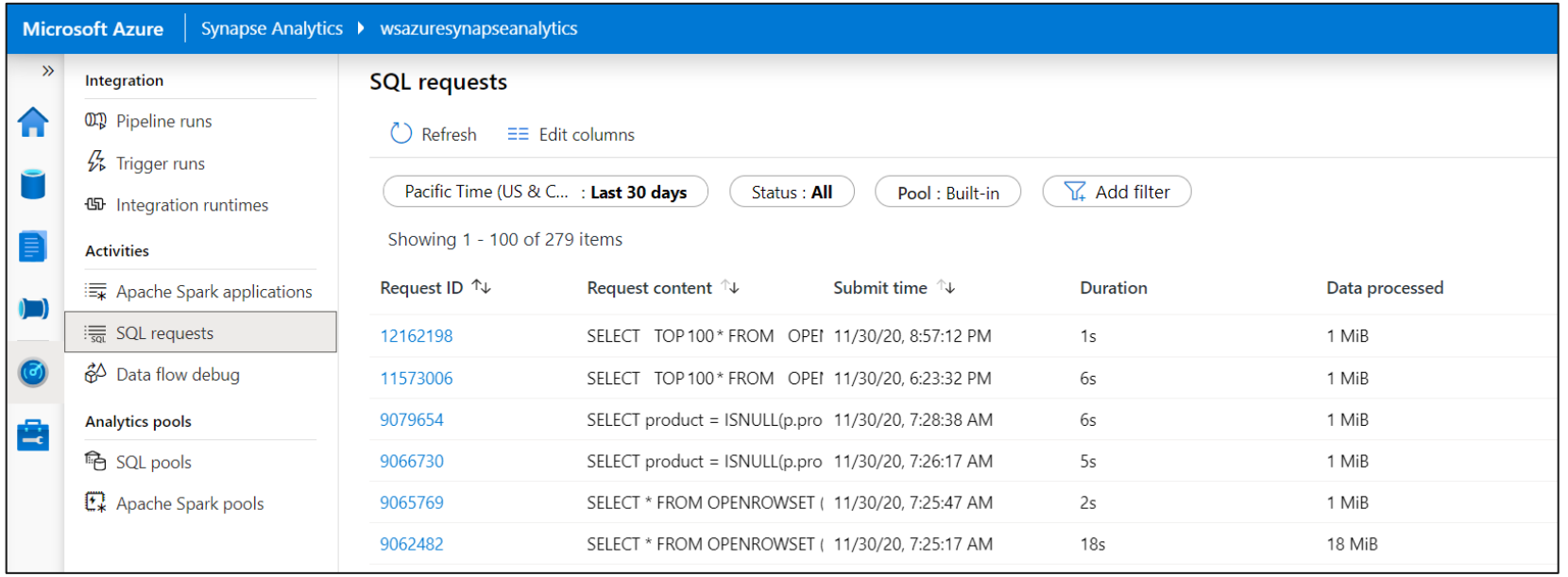
Apply filter for pool to get SQL requests per compute pool

Validate query text

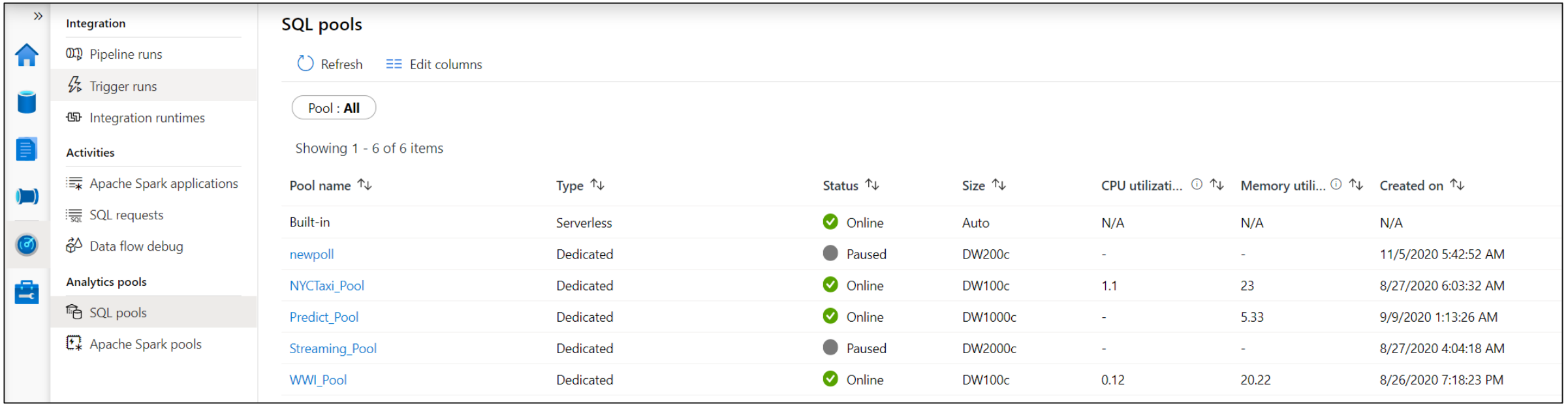
Additional available filters include

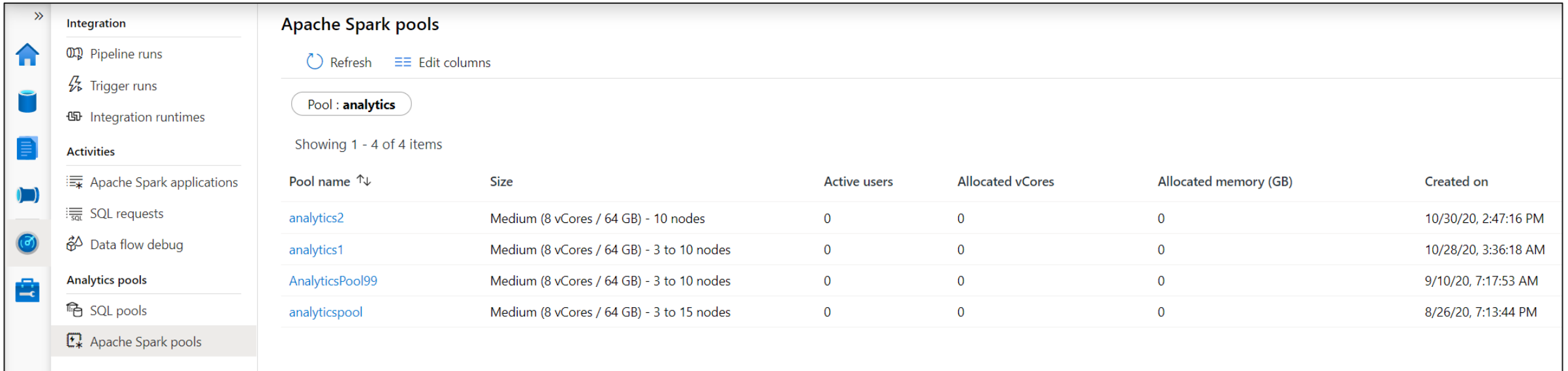
1. Start time
2. End time
3. Request ID
4. Session ID
5. Submitter
6. Workload group



# Monitor Hub – Analytics pools



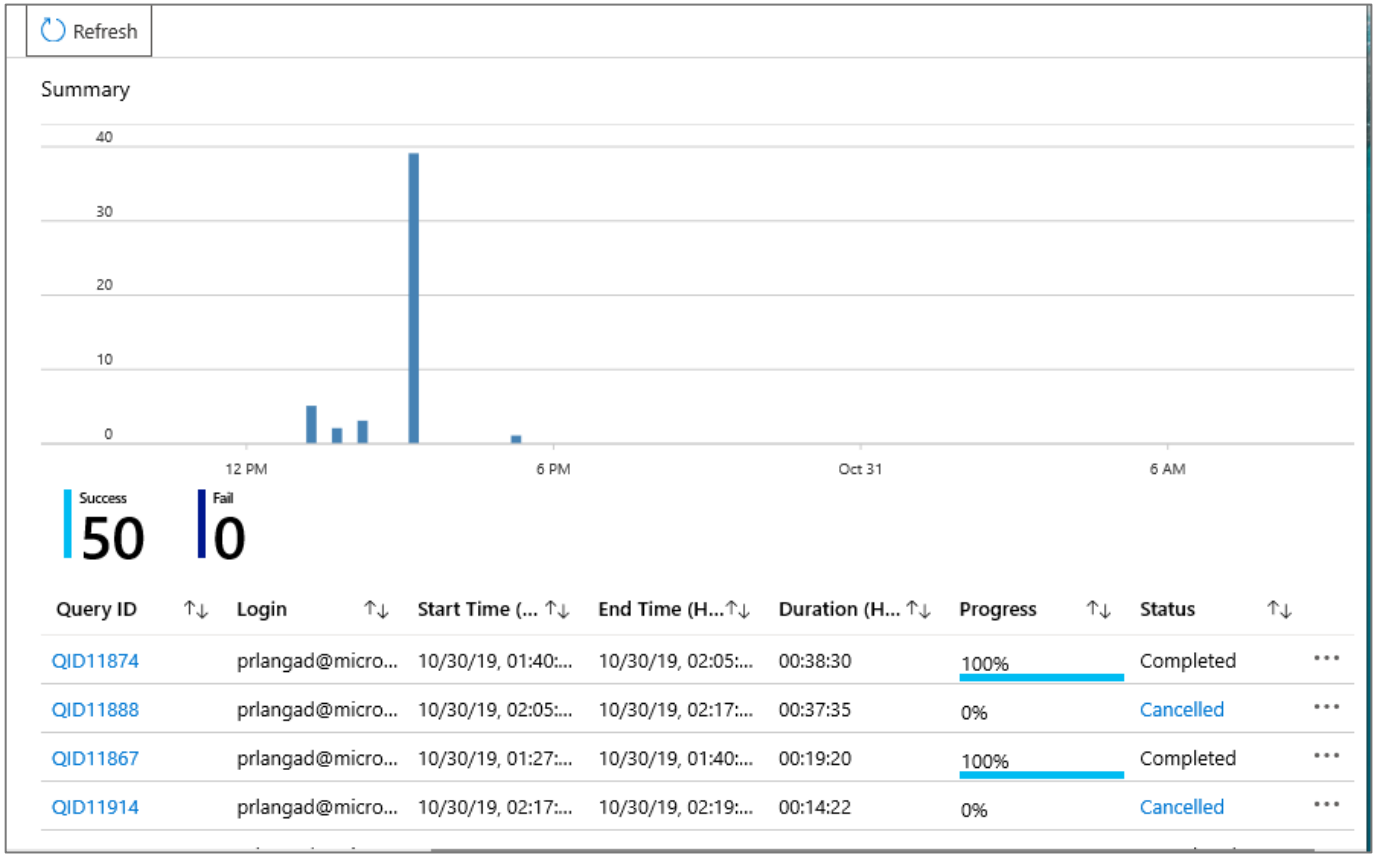
# Monitor SQL Pools

## Overview

Monitor SQL Pool in Azure Portal for overall usage and query activities.

## Benefits

- Access SQL Audit Logs for my SQL computes
- Monitor status and progress of all/specific activities
- Dashboard view to monitor performance
- Get to know scale of SQL compute resource



# Manage

# Workload Management

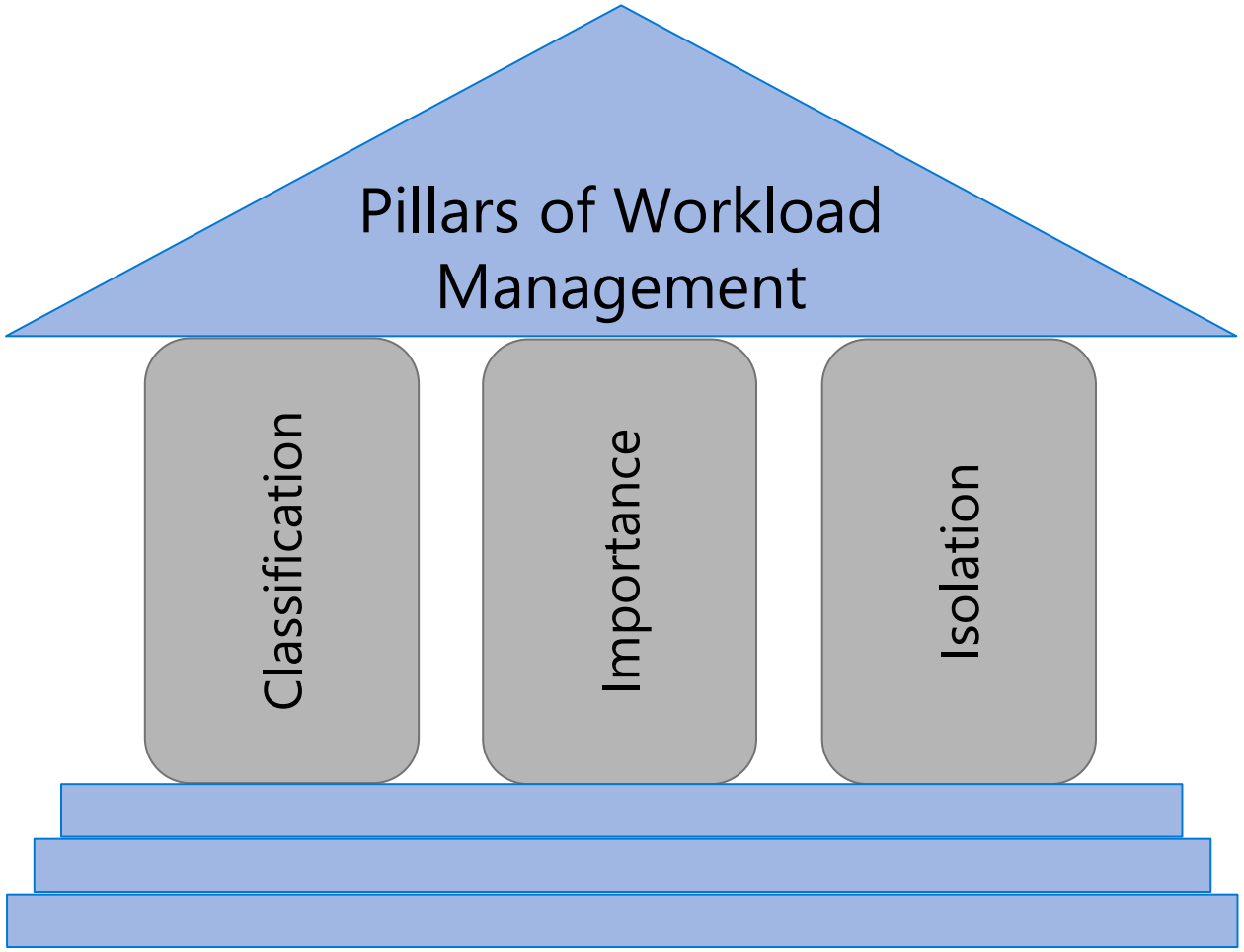
## Overview

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

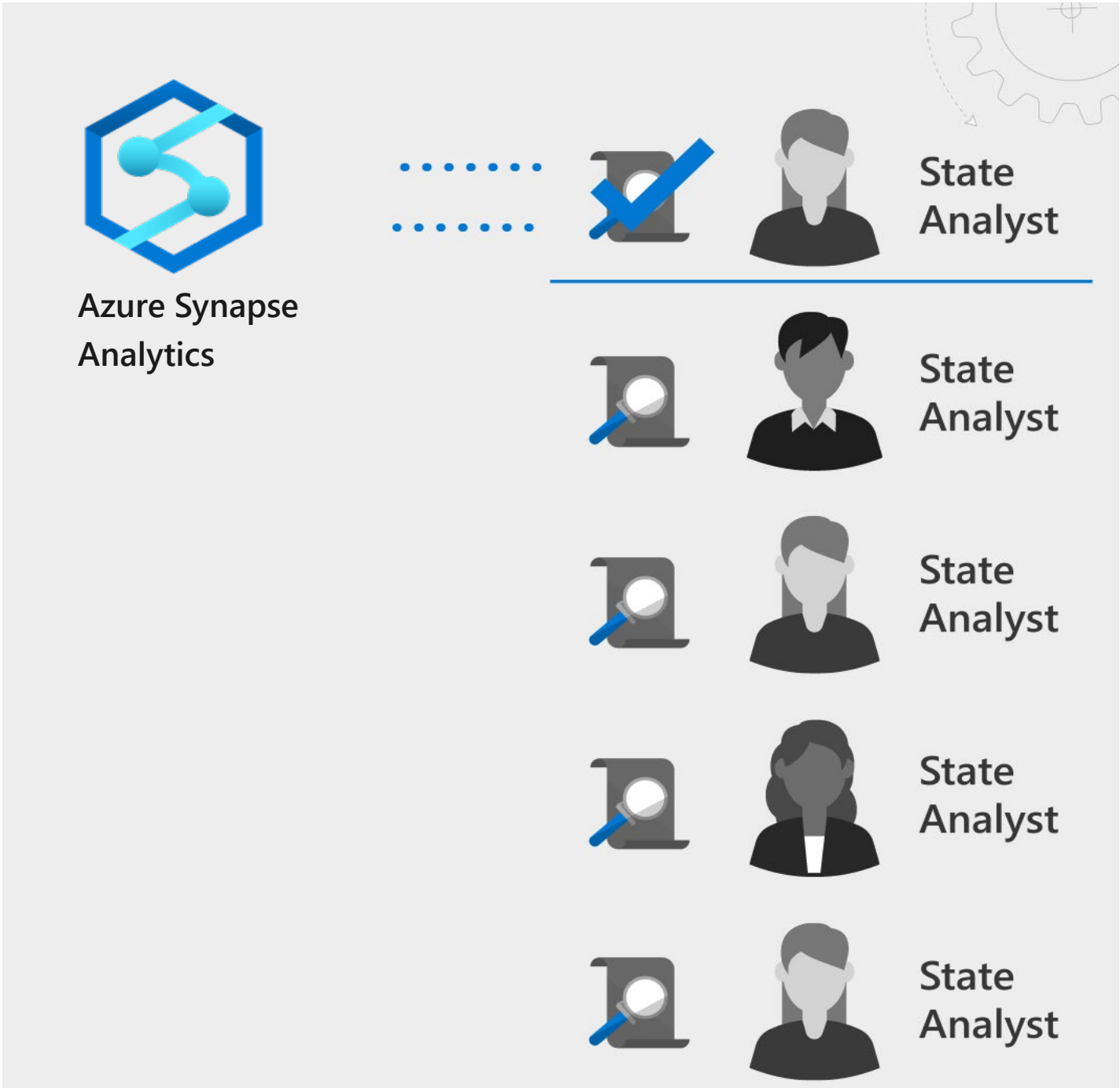
State analysts have normal importance.

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')
```



# 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 take SQL Analytics 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

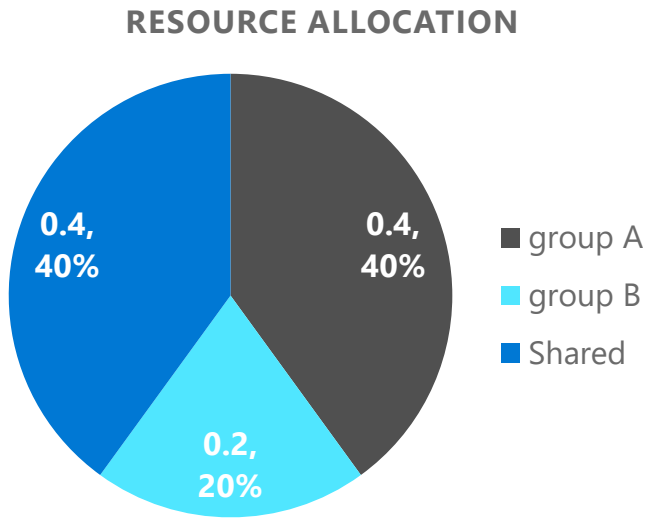
Set Query timeout value. Get DBAs out of the business of killing runaway queries

## 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 ]
)[ ; ]
```



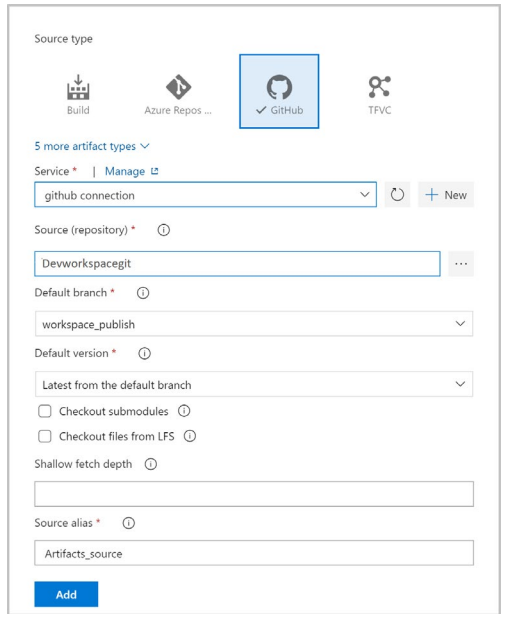
# Discussion..

What are some of the advantages and disadvantages of workload isolation versus SQL Pool scaling?

# Continuous integration and delivery (CI/CD)

## Overview

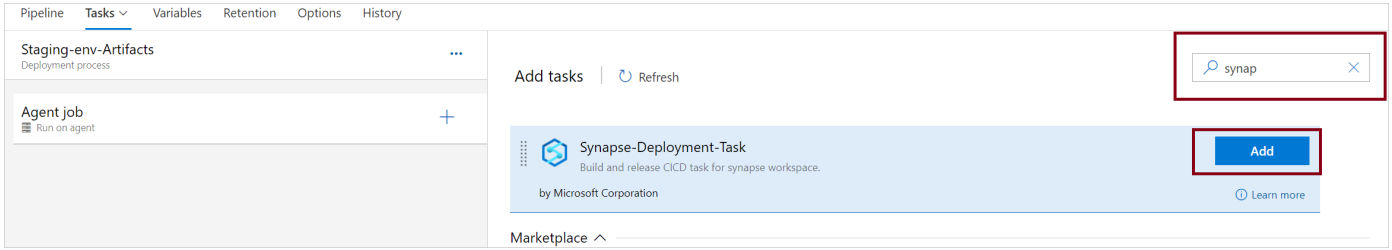
Database project support in SQL Server Data Tools (SSDT) allows teams of developers to collaborate over a version-controlled Azure Synapse Analytics, and track, deploy and test schema changes.

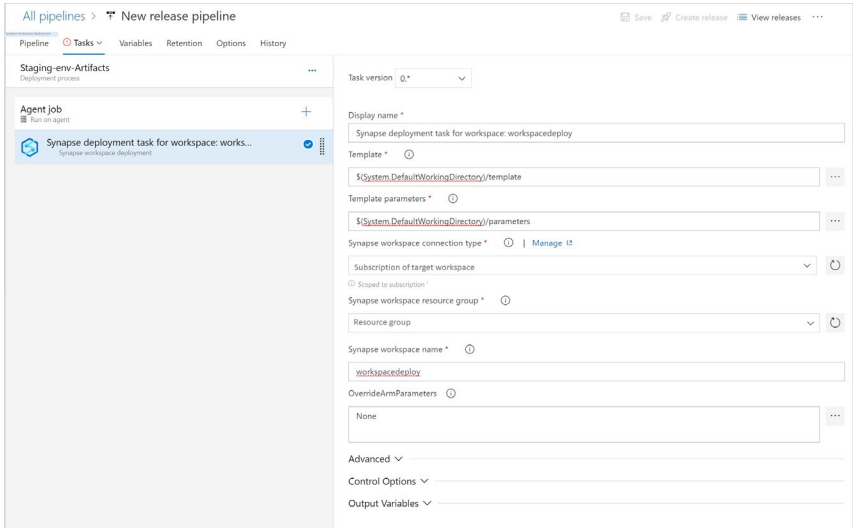



## Benefits

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

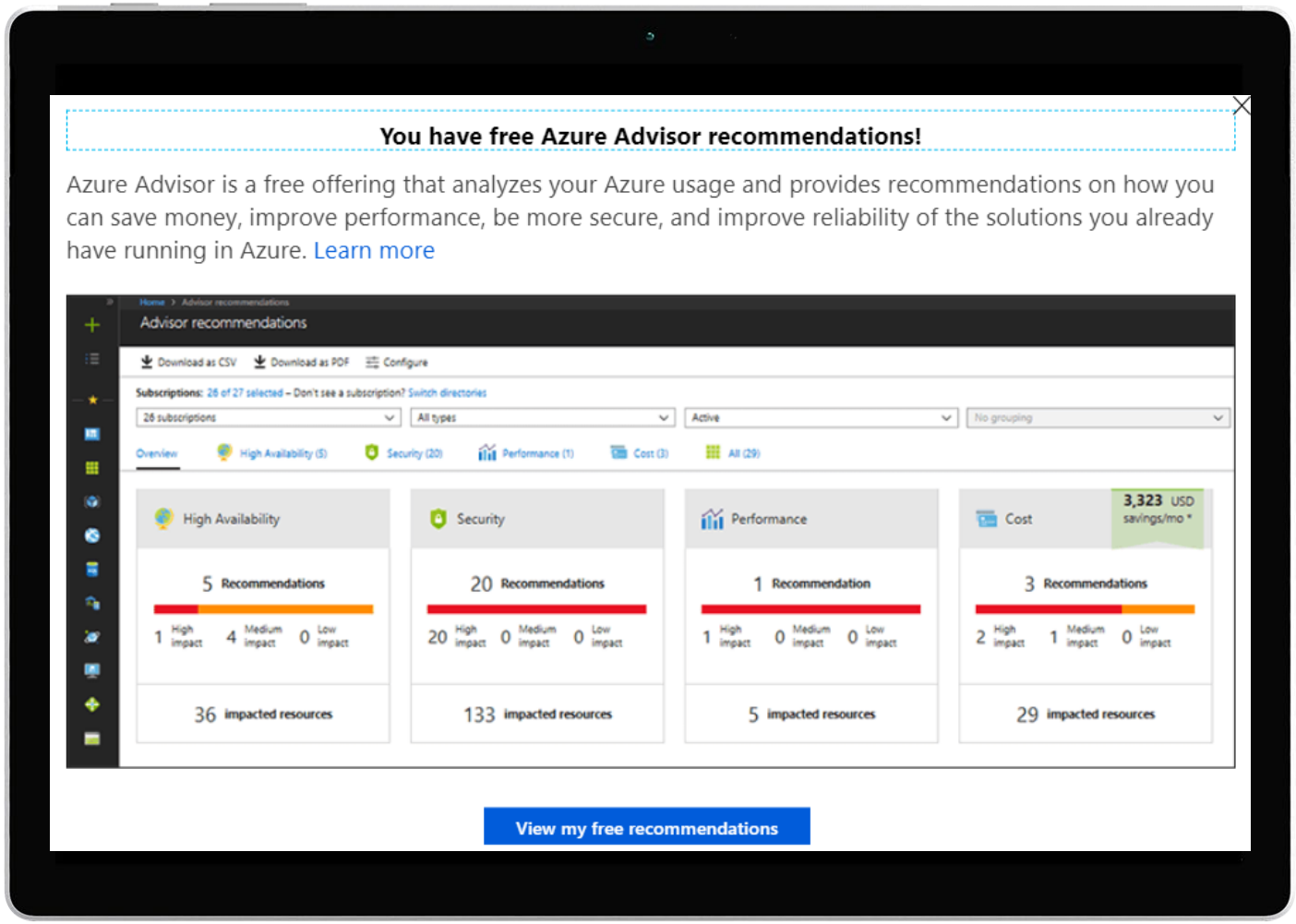
Provision additional capacity

## 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.

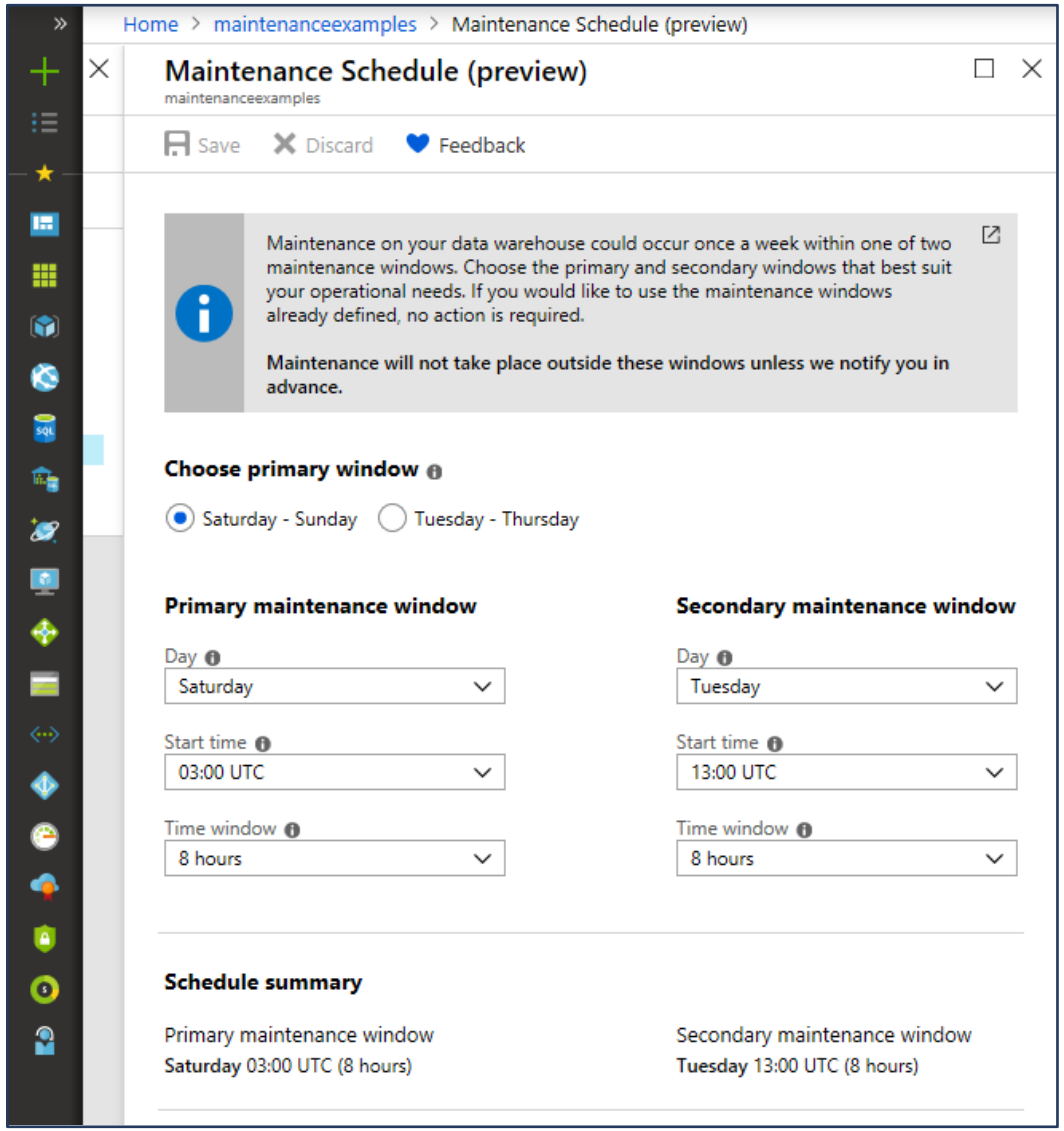
Select a primary and secondary window within a seven-day period.

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.



# Manage Summary

- Resource contention can be mitigated by applying Workload Management in Azure Synapse Analytics.
  - **Workload Classification:** Assign a request to a workload group and set importance levels.
  - **Workload Importance:** Influence the order in which a request gets access to resources.
  - **Workload Isolation:** Reserve resources for a workload group.

- Use Azure Advisor recommendations to identify suboptimal table distribution, data skew, cache misses, tempdb contention, and suboptimal plan selection.

- Avoid disruptive system downtime due to system upgrades by configuring maintenance windows and notifications in Azure Synapse Analytics.

# Whole Group Activity: Monitor & Manage

OUTCOMES

As a result of participating in this activity, you will better be able to **address scenarios that involve the design of how to monitor and manage a solution implemented with Azure Synapse Analytics**.

ACTIVITY

30 minutes | Session Speakers | Main Call

Read and respond to questions posed by the speakers.

Open the interactive tool **Mentimeter** using the QR code and answer the speakers' questions.

Microsoft Azure
Solutions Business

# Mentimeter Poll

Scan QR Code
or

Go to www.menti.com and use code

**2166 7832**

# Build Hands-on Lab: Monitoring

OUTCOMES

As a result of participating in this lab, you will be better able to **build and deploy monitoring solutions for Azure Synapse Analytics and associated Spark applications.**

ACTIVITY

60 minutes | Independent | Table Group Channel & CloudLabs Environment

Independently review and complete Exercises 1-2 in the Lab 8 Guide.

Login to your **CloudLabs environment** and work through a set of tasks you would typically follow in work associated with ingesting and integrating your customer's data.

The Lab Guide is available in the 'Files' tab of the General channel. Engage your Table Group call for support and troubleshooting.

# Closing

*Thank you for your participation in today's Azure Synapse Technical Boot Camp!*

## TODAY

We learned:

- ✓ Implement a process that trains a machine learning model, registers it for use and utilizes the model for scoring using the Predict statement.
- ✓ Address scenarios the involve the design of how to monitor and manage a solution implemented with Azure Synapse Analytics.

## TOMORROW

We will learn to:

- How to address real world challenges in creating a PoC
- Applying what you have learned throughout the week

Microsoft Azure
Solutions Business

# Mentimeter Poll

Scan QR Code
or

Go to www.menti.com and use code

**4438 6935**

# Microsoft Azure Solutions Business