Microsoft

Microsoft Azure Virtual Training Day: Deliver Integrated Analytics with Azure Synapse

Microsoft

Realize Integrated Analytics with Azure Synapse Analytics

Agenda

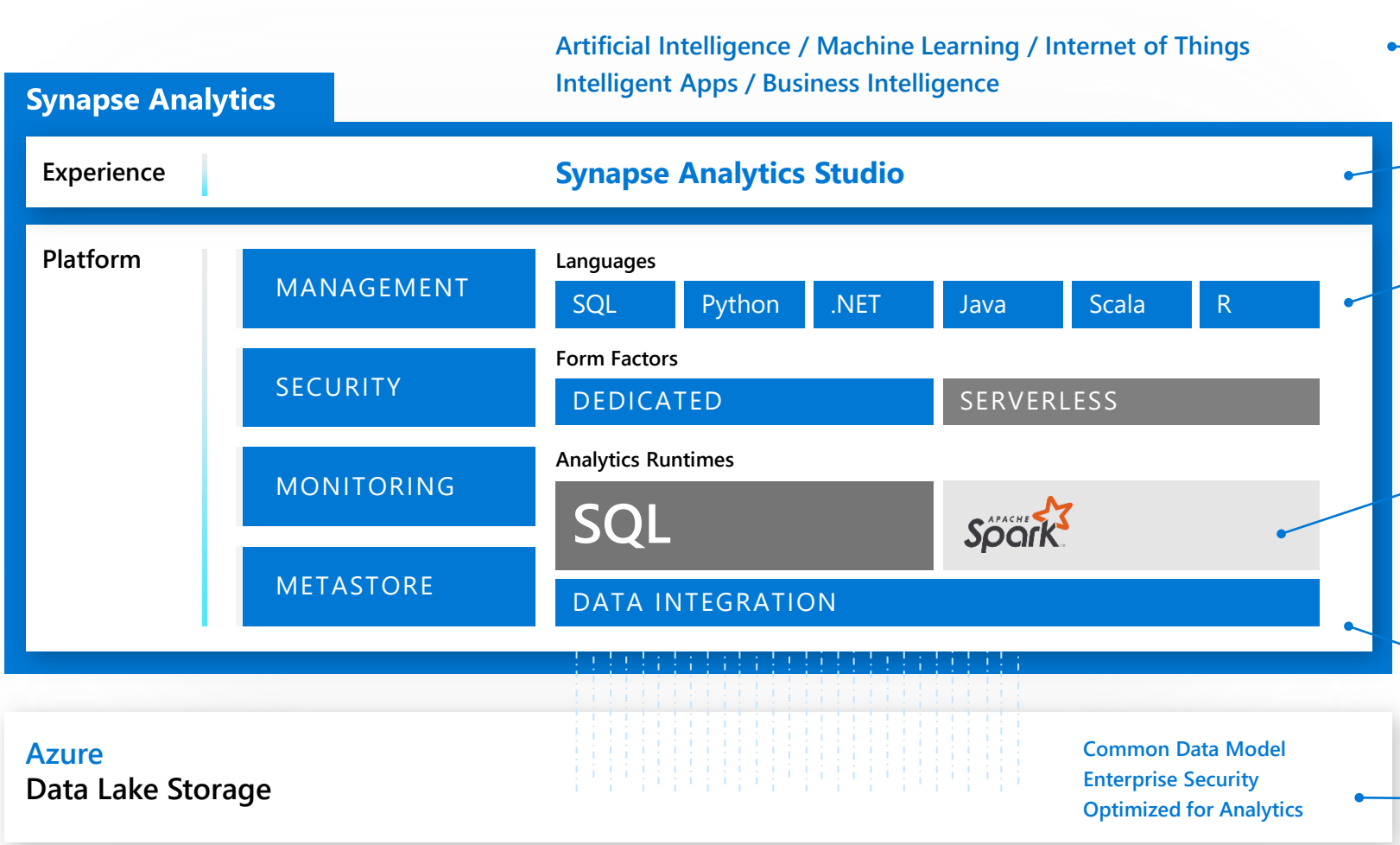
- Describe Azure Synapse Analytics
- Surveying the components of Azure Synapse Analytics
- Explore Azure Synapse Studio
- Designing a Modern Data Warehouse with Azure Synapse Analytics

Microsoft

Describe Azure Synapse Analytics

Azure Synapse Analytics

Limitless analytics service with unmatched time to insight



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 as a dedicated and serverless offering

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

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

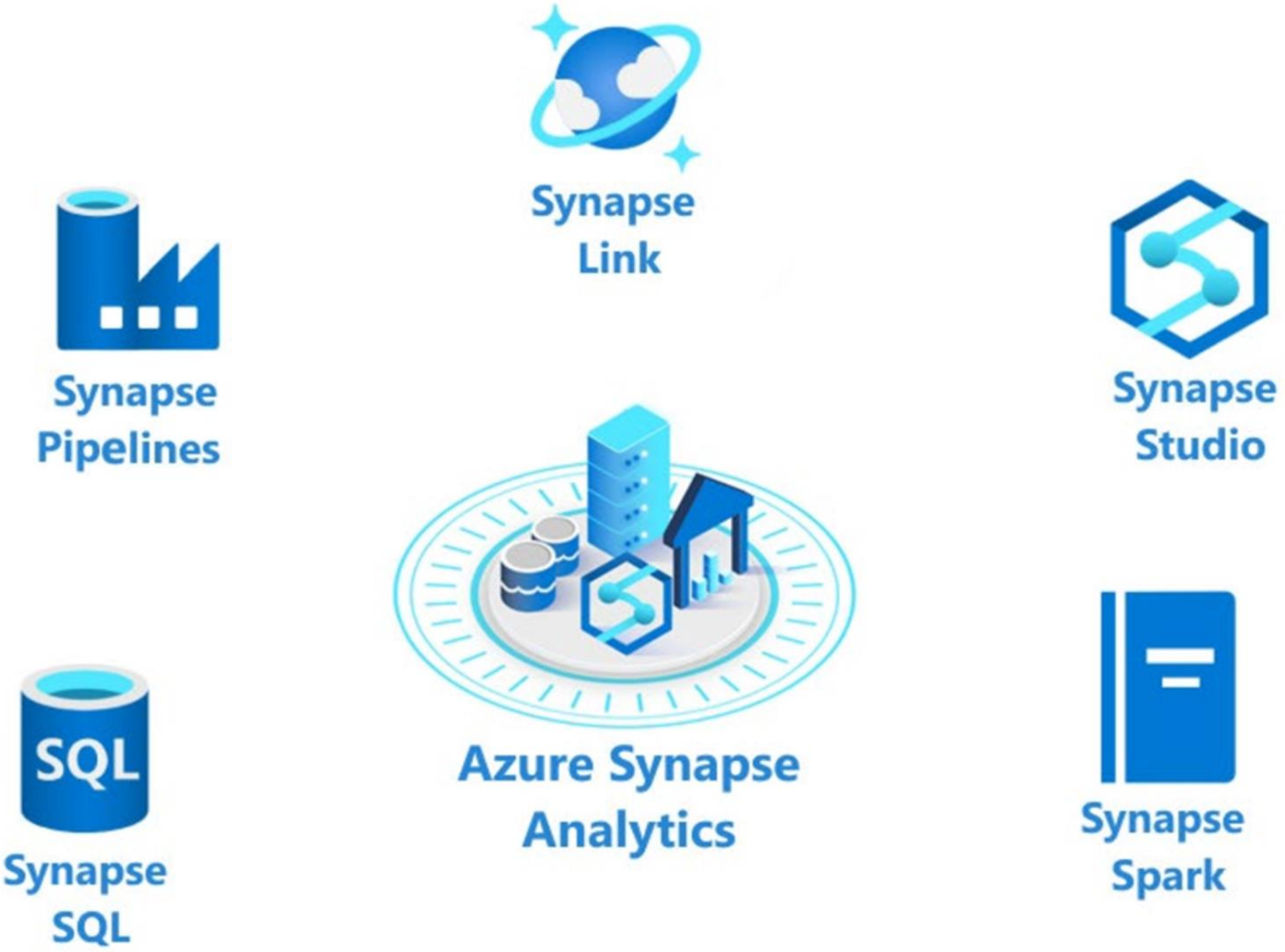
Integrated **platform services** for, management, security, monitoring, and metastore

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

Microsoft

Surveying the components of Azure Synapse Analytics

Synapse Analytics Components



DEMO

Surveying the components of Azure Synapse Analytics

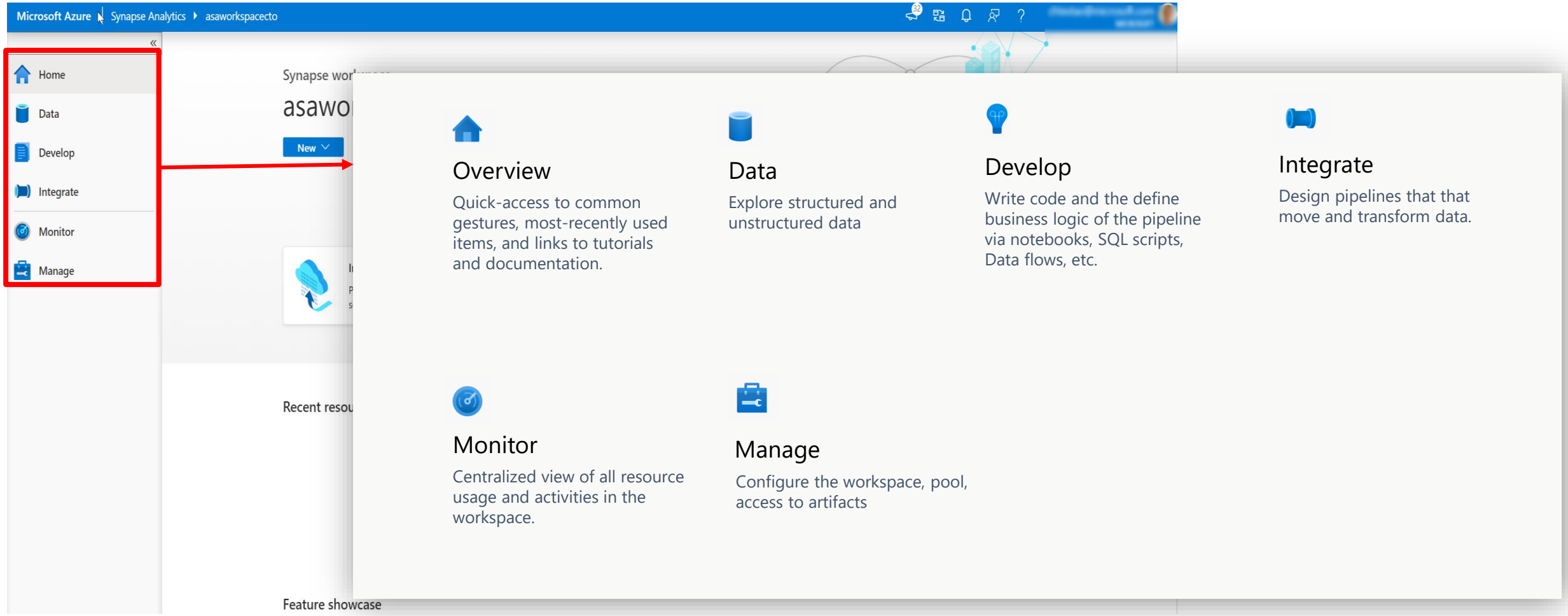
Microsoft

Exploring Azure Synapse Studio

Synapse Studio

Synapse Studio divided into **Activity hubs**.

These organize the tasks needed for building analytics solution.

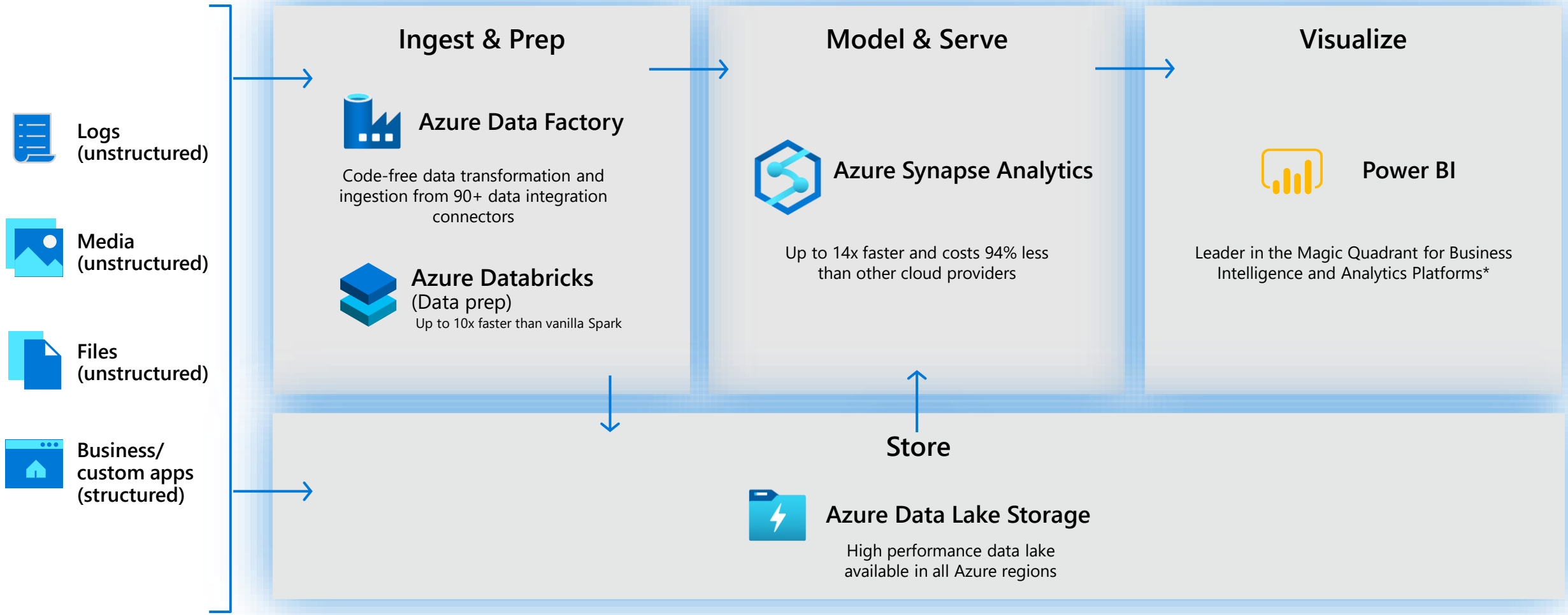


Demo: Exploring Synapse Studio

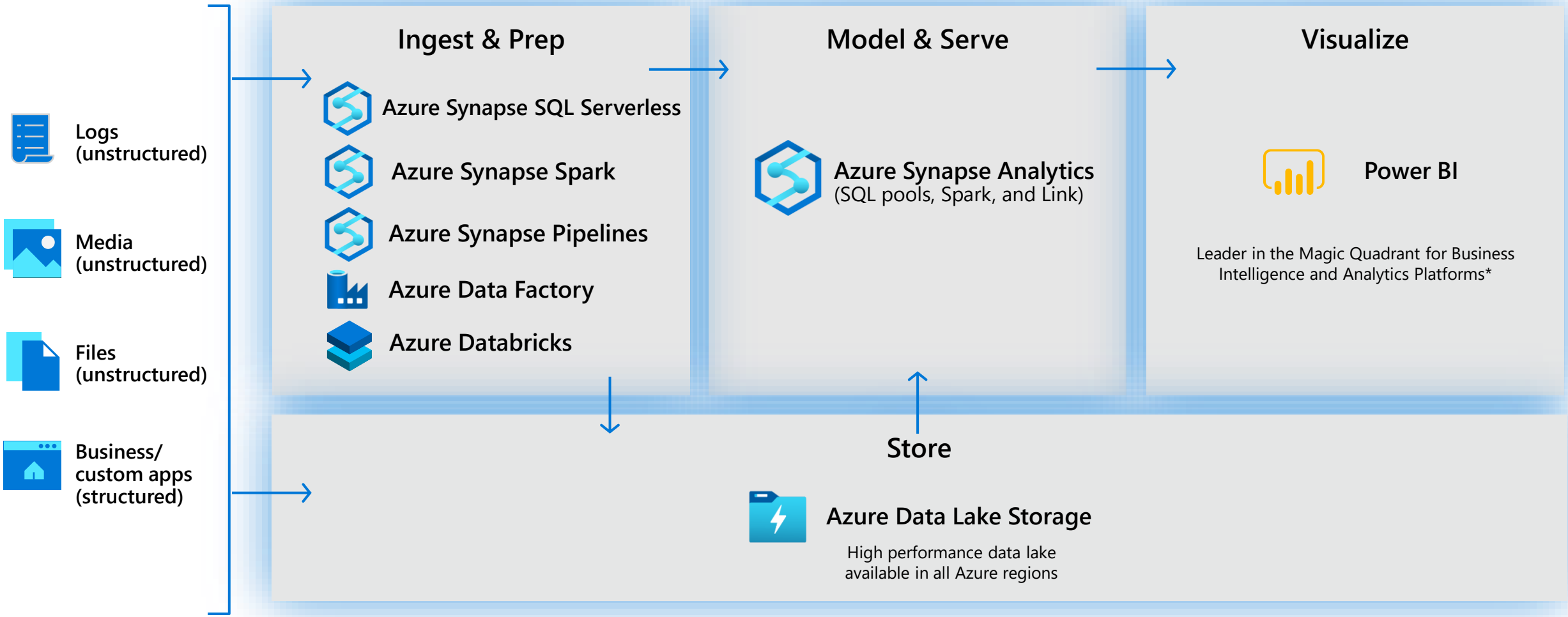
Microsoft

Designing a modern data warehouse with Azure Synapse Analytics

Modern data warehousing patterns



Modern data warehousing pattern with Azure Synapse Analytics



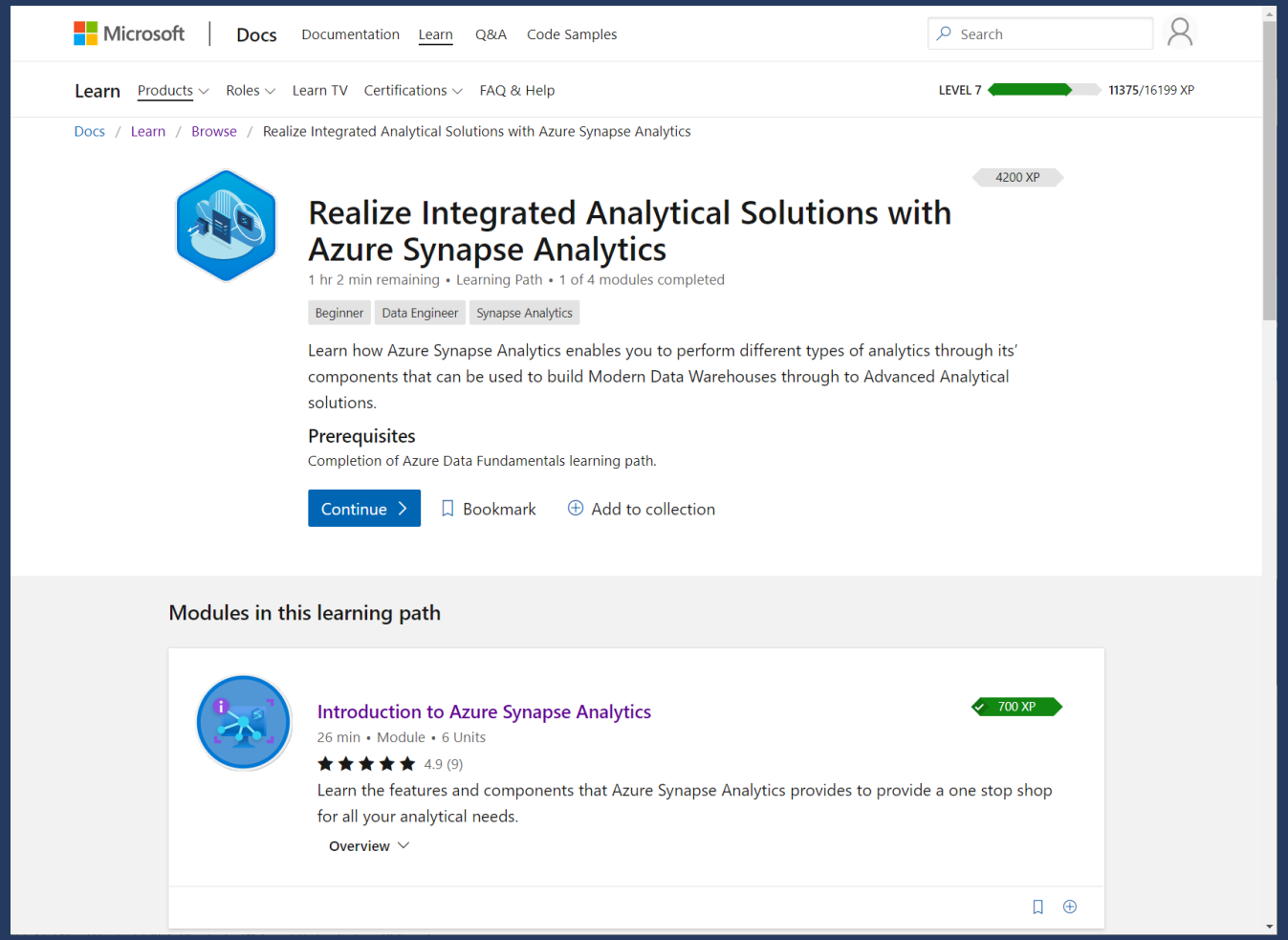
DEMO

Designing a modern data warehouse with Azure Synapse Analytics

/learn alert

Complete interactive learning exercises, watch videos, and practice and apply your new skills.

aka.ms/mslearnasalp1



Microsoft

Optimizing a Data Warehouse with Azure Synapse dedicated SQL Pools

Agenda

- Describe Azure Synapse dedicated SQL Pools
- Understand developer features of Azure Synapse dedicated SQL Pools
- Use data loading best practices in Azure Synapse dedicated SQL Pools
- Optimize data warehouse query performance in Azure Synapse dedicated SQL Pools

Microsoft

Describe Azure Synapse dedicated SQL Pools

Azure Synapse dedicated SQL Pools

Dedicated SQL Pools refers to the enterprise data warehousing feature available in Azure Synapse Analytics. Dedicated SQL Pools represent the dedicated compute resources required to process data loads and queries for relational big data workloads. The size of SQL pool is determined by Data Warehousing Units (DWU).

Comparing dedicated SQL Pools with serverless SQL pools in Azure Synapse Analytics

Dedicated SQL pools

- Used for Data Warehouse operations
- Provides predictable performance and costs
- Reserves processing power for data stored in SQL tables

Serverless SQL pools

- Used for data preparation or ad-hoc queries against unstructured data.
- Provides an always available SQL endpoint for unplanned workloads
- Enables interactive querying

Microsoft

Understand developer features of Azure Synapse dedicated SQL Pools

Customer Problem

"You need to make it easier for us to specify a set of rows within a query result set and compute a value for each row."

- Many industries

Windowing function

```
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;
```

Customer Problem

"It takes too long to get basic information about the data in a large data set as we are exploring data of a big data set" – Data Engineers

APPROX_COUNT_DISTINCT

```
SELECT O_OrderStatus, APPROX_COUNT_DISTINCT(O_OrderKey) AS
Approx_Distinct_OrderKey
FROM dbo.Orders
GROUP BY O_OrderStatus
ORDER BY O_OrderStatus;
```

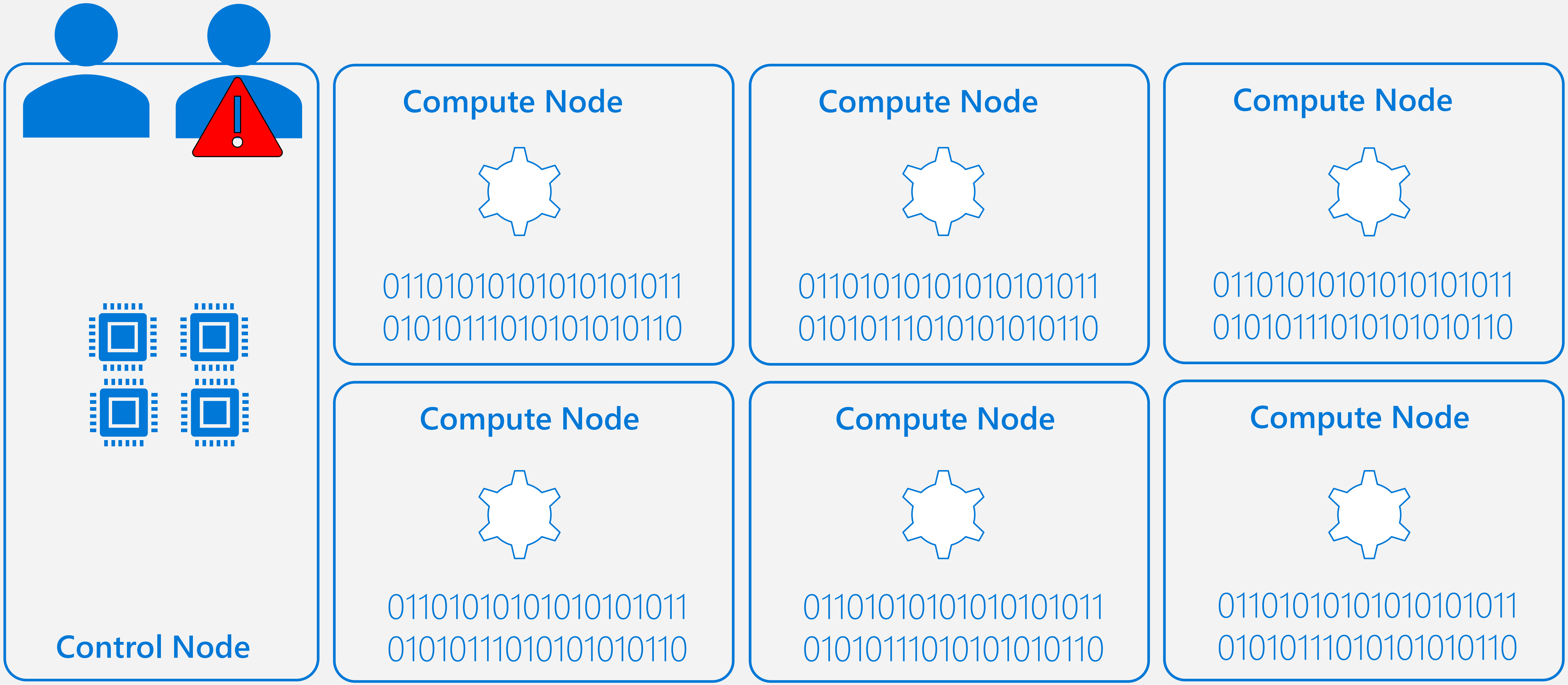
DEMO

Understand developer features of Azure Synapse dedicated SQL Pools

Microsoft

Using data loading best practices in Azure Synapse dedicated SQL Pools

Use workload management to prioritize workloads



DEMO

Using data loading best practices in Azure Synapse dedicated SQL Pools

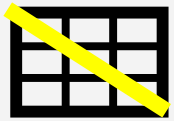
Microsoft

Optimize data warehouse query performance in Azure Synapse dedicated SQL Pools

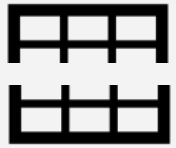
Maximize query performance

Table distribution

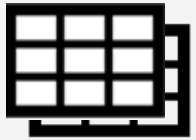
Round Robin Tables



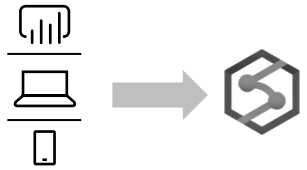
Hash Distributed Tables



Replicated Tables

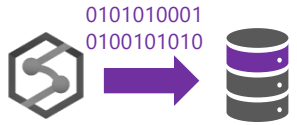


Result-set caching flow

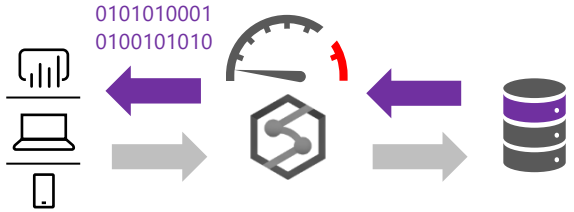


1 Client sends query to SQL pool

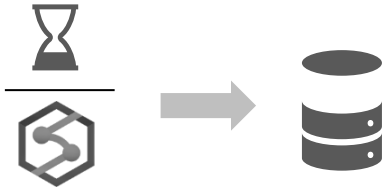
2 Query is processed using compute nodes which pull data from remote storage, process query and output back to client app



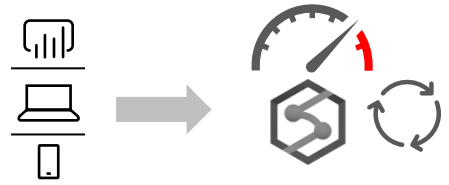
Query results are cached in remote storage so subsequent requests can be served immediately



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

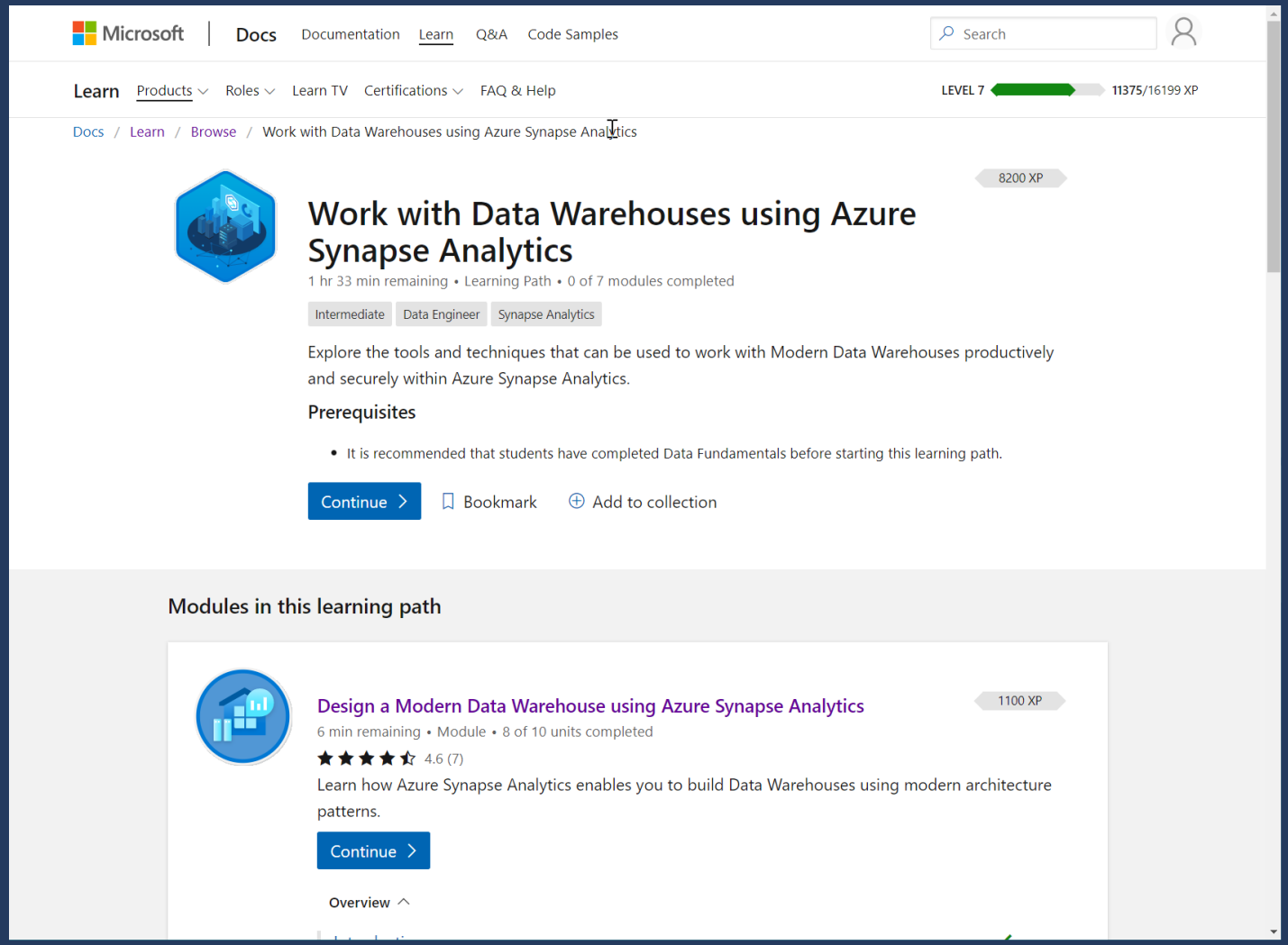
DEMO

Optimize data warehouse query performance in Azure Synapse dedicated SQL Pools

/learn alert

Complete interactive learning exercises, watch videos, and practice and apply your new skills.

aka.ms/mslearnasadedicated



Microsoft

Perform Data Engineering with Azure Synapse Spark pools

Agenda

- Understanding big data engineering with Apache Spark in Azure Synapse Analytics
- Ingesting data with Apache Spark notebooks in Azure Synapse Analytics
- Transforming data with DataFrames in Spark pools in Azure Synapse Analytics
- Integrating SQL and Spark pools in Azure Synapse Analytics

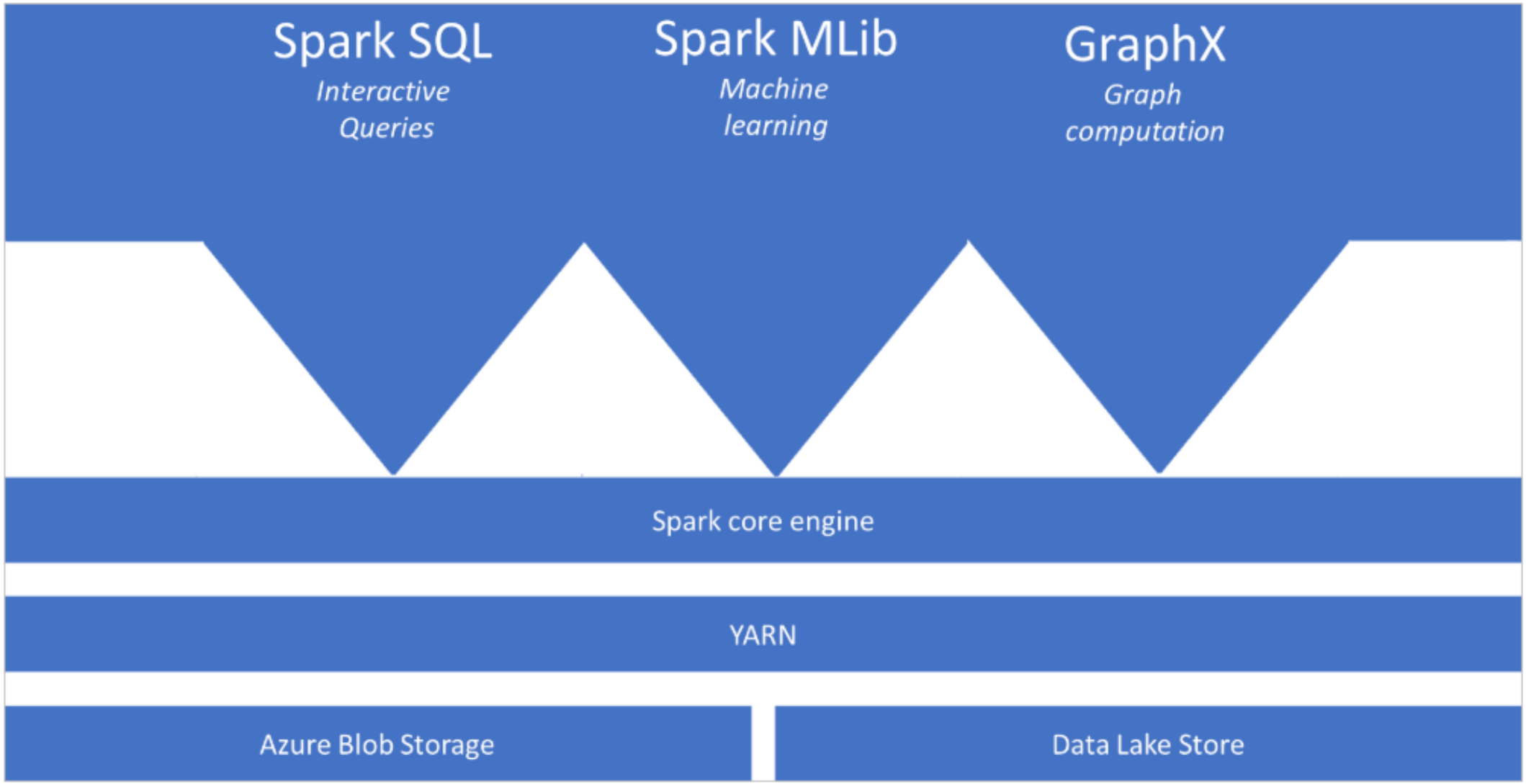
Microsoft

Understanding big data engineering with Apache Spark in Azure Synapse Analytics

What to use when and where

	Apache Spark	HDInsight	Azure Databricks	Synapse Spark
WHAT	Is an Open Source memory optimized system for managing big data workloads	Microsoft implementation of Open Source Spark managed within the realms of Azure	A managed Spark as a Service solution	Embedded Spark capability within Azure Synapse Analytics
WHEN	When you want to benefits of spark for big data processing and/or data science work without the Service Level Agreements of a provider	When you want to benefits of OSS spark with the Service Level Agreement of a provider	Provides end to end data engineering and data science solution and management platform	Enables organizations without existing Spark implementations to fire up a Spark cluster to meet data engineering needs without the overheads of the other Spark platforms listed
WHO	Open Source Professionals	Open Source Professionals wanting SLA's and Microsoft Data Platform experts	Data Engineers and Data Scientists working on big data projects every day	Data Engineers, Data Scientists, Data Platform experts and Data Analysts
WHY	To overcome the limitations of SMP systems imposed on big data workloads	To take advantage of the OSS Big Data Analytics platform with SLA's in place to ensure business continuity	It provides the ability to create and manage an end-to-end big data/data science project using one platform	It provides the ability to scale efficiently with spark clusters within a one stop shop Data Warehousing platform of Synapse.

Apache Spark



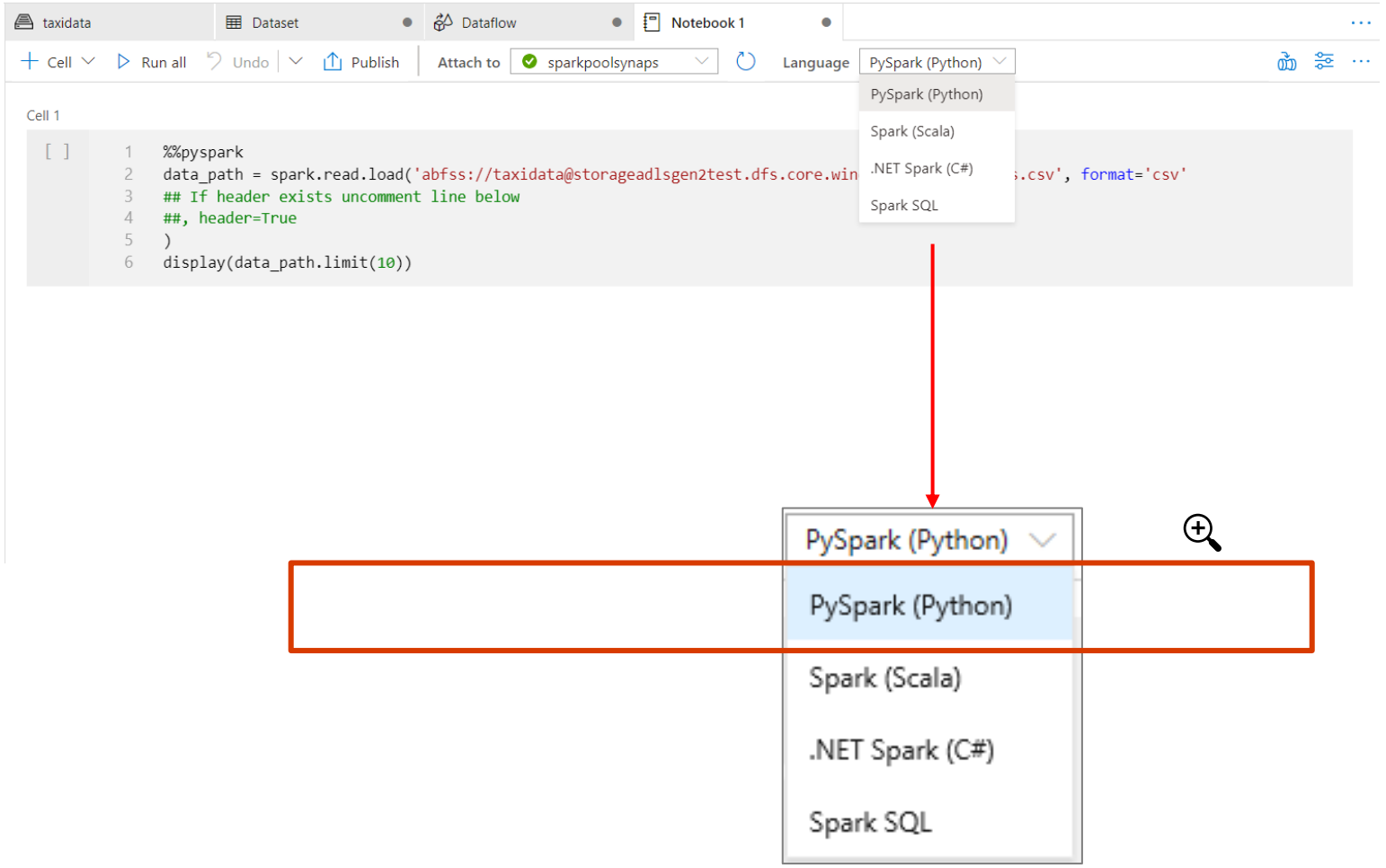
Microsoft

Ingesting data with Apache Spark notebooks in Azure Synapse Analytics

Develop Hub - Notebooks

Notebooks

- Access through workspace URL: https://web.azuresynapse.net/
- Examples Available through Knowledge Center
- Allows to write multiple languages in one notebook by using %%<Name of language>
- Support for Language Syntax highlight, syntax error, syntax code completion
- Offers temporary tables across languages
- Export results

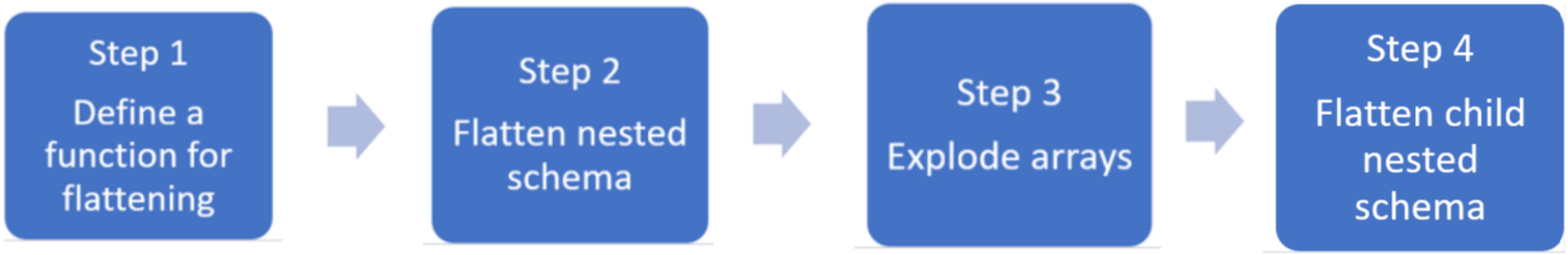


Demo: Ingesting data with Apache Spark notebooks in Azure Synapse Analytics

Microsoft

Transforming data with DataFrames in Spark pools in Azure Synapse Analytics

Analyze Complex Data types with Spark pools in Azure Synapse



Demo: Transforming data with Data Frames in Spark pools in Azure Synapse Analytics

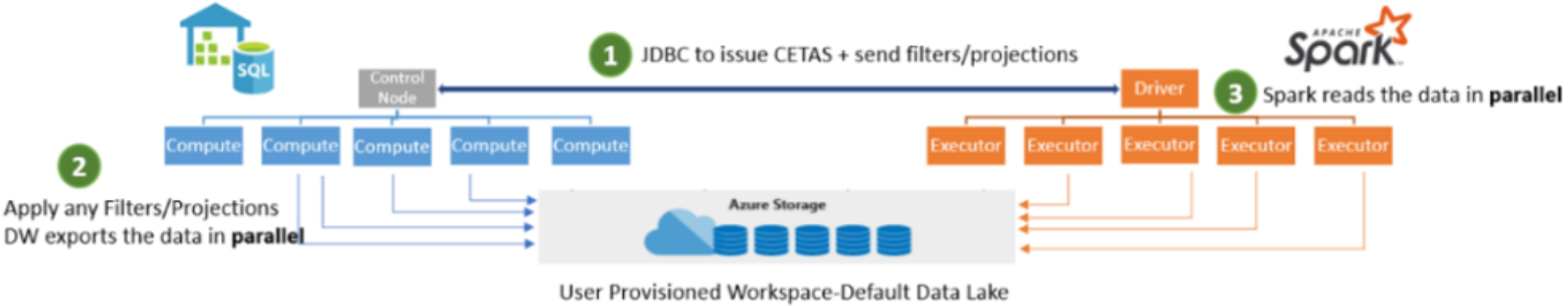
Microsoft

Integrating SQL and Spark pools in Azure Synapse Analytics

Design integrating Data Transfer between Spark and SQL pools

Existing Approach: JDBC

New Approach: JDBC and Polybase

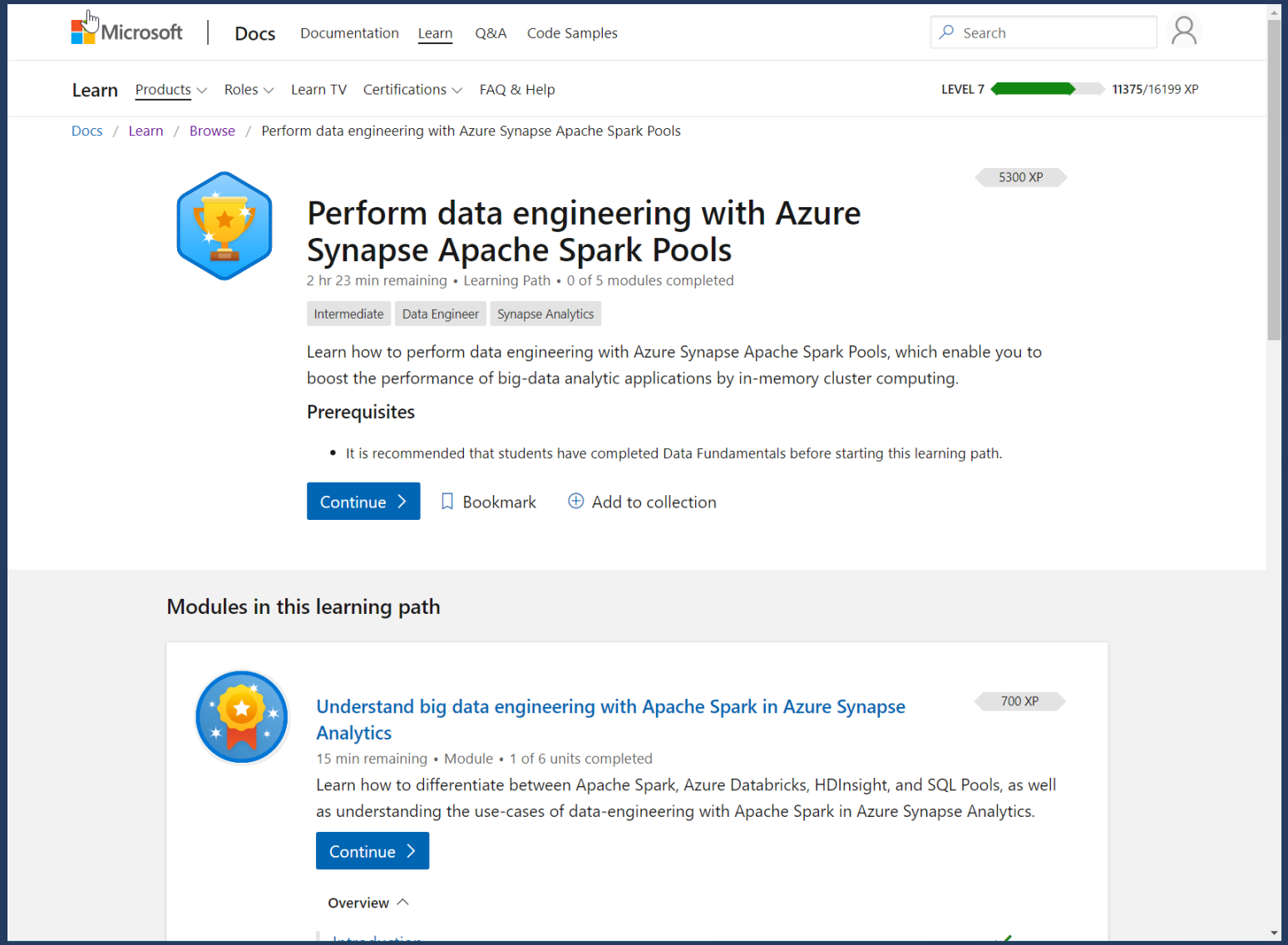


Demo: Integrating SQL and Spark pools in Azure Synapse Analytics

/learn alert

Complete interactive learning exercises, watch videos, and practice and apply your new skills.

aka.ms/mslearnasaspark



Microsoft

Building automated data integration pipelines with Azure Synapse Pipelines

Agenda

- Describe Azure Synapse Pipelines
- Perform petabyte-scale ingestion with Azure Synapse Pipelines
- Perform code-free transformation at scale with Azure Synapse Pipelines
- Orchestrate data movement and transformation in Azure Synapse Pipelines

Microsoft

Describe Azure Synapse Pipelines

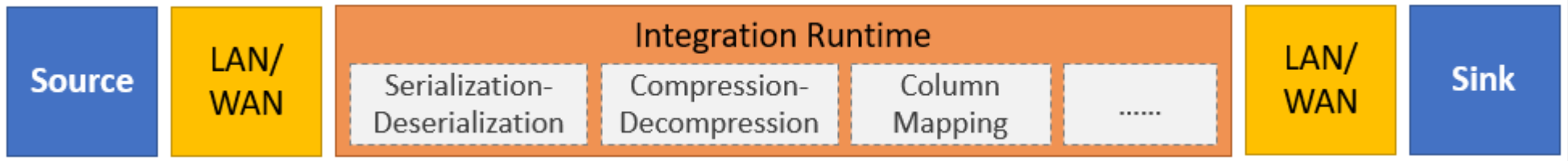
Azure Synapse Pipelines

Embeds the capabilities of Azure Data Factory within Azure Synapse Analytics for a cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.

Microsoft

Perform petabyte-scale ingestion with Azure Synapse Pipelines

Copy files with the Copy Activity



Supported file formats:

- Text
- JSON
- Avro
- ORC
- Parquet

Copy activity can compress and decompress files with The following codecs:

- Gzip
- Deflate
- Bzip2
- ZipDeflate

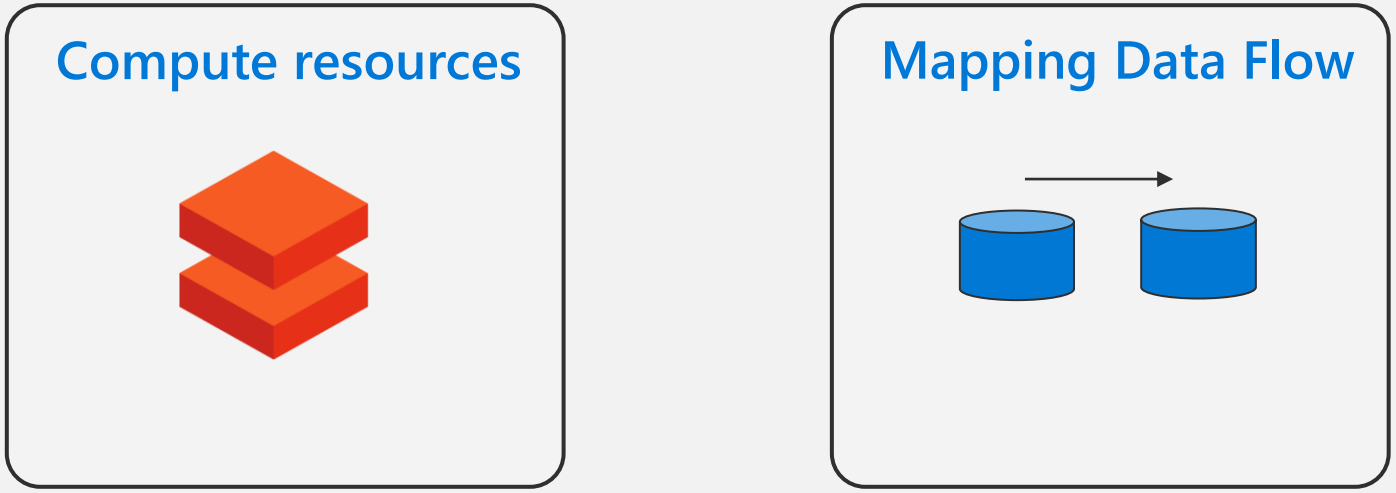
DEMO

Ingesting data with Azure Synapse Pipelines

Microsoft

Perform code-free transformation at scale with Azure Synapse Pipelines

Methods for transforming in Azure Synapse Pipelines



Methods for transforming in Azure Synapse Pipelines

Code free data transformation at scale

SPEAKER NAME

Demo: Transforming your data with Azure Synapse Pipelines

Microsoft

Orchestrate data movement and transformations with Azure Synapse Pipelines

Orchestrating Azure Synapse Analytic Spark Notebooks

1. Use Storage Account

2. Use Azure Synapse Pipeline

3. Create data workflow pipeline

4. Add Azure Synapse Analytics Spark Notebook to pipeline

5. Perform analysis on the data

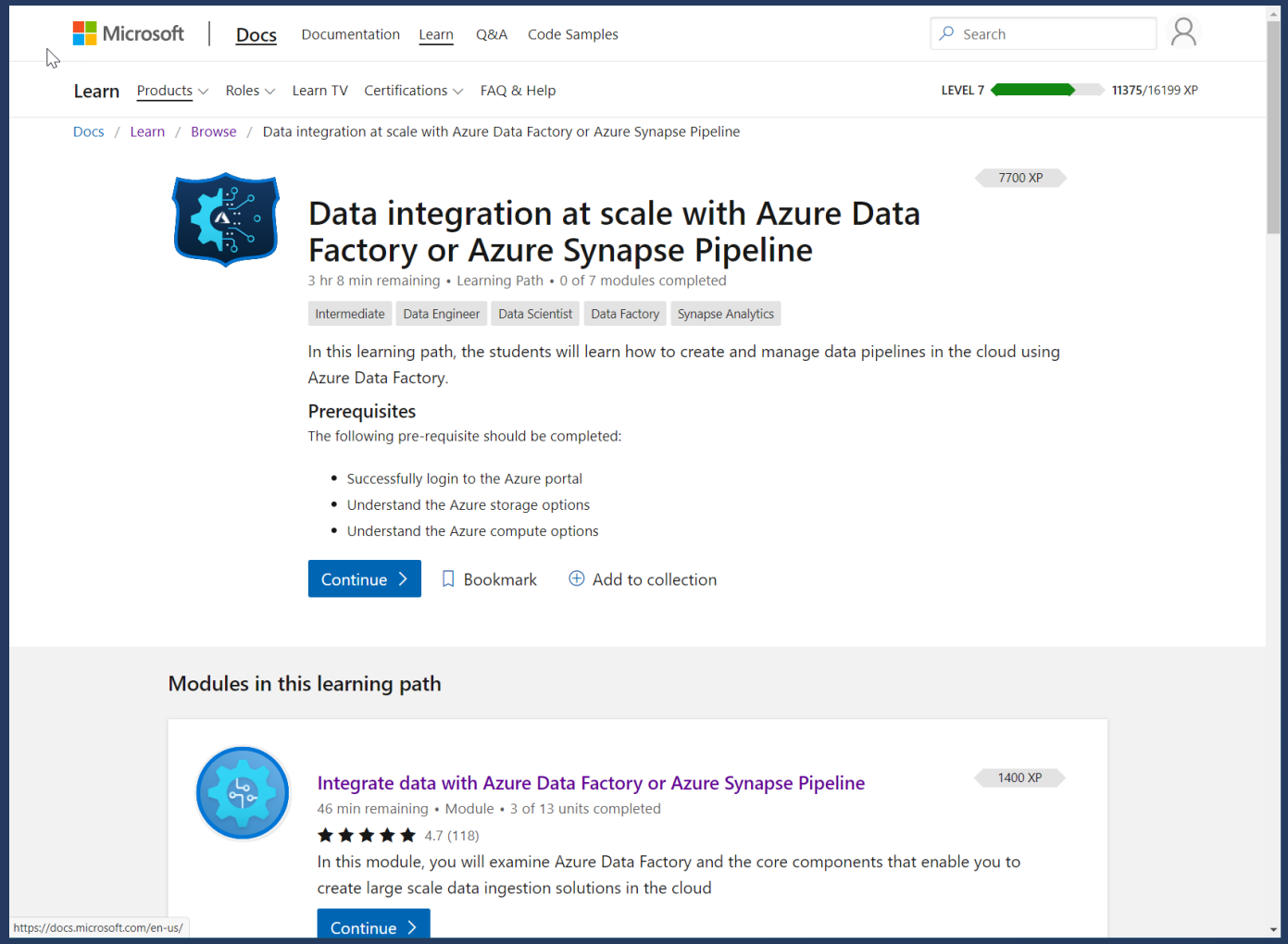
DEMO

Orchestrate data movement and transformations with Azure Synapse Pipelines

/learn alert

Complete interactive learning exercises, watch videos, and practice and apply your new skills.

aka.ms/mslearnasapipelines



Microsoft

Run interactive queries using Azure Synapse serverless SQL pools

Agenda

- Describe Azure Synapse serverless SQL pools
- Querying a data lake store using serverless SQL pools in Azure Synapse Analytics
- Securing access to data through using serverless SQL pools in Azure Synapse Analytics

Microsoft

Describe Azure Synapse serverless SQL pools

Azure Synapse serverless SQL Pools

Every Azure Synapse Analytics workspace comes with serverless SQL pool endpoints so you can start querying data in seconds to minutes in a data lake as soon as the workspace is created. There's no infrastructure to setup or clusters to maintain.

Comparing dedicated SQL Pools with serverless SQL pools in Azure Synapse Analytics

Dedicated SQL pools

- Used for Data Warehouse operations
- Provides predictable performance and costs
- Reserves processing power for data stored in SQL tables

Serverless SQL pools

- Used for data preparation or ad-hoc queries against unstructured data.
- Provides an always available SQL endpoint for unplanned workloads
- Enables interactive querying

Ensure **Built-in** is selected **(1)** in the Connect to dropdown list above the query window, then run the query **(2)**. Data is loaded by the serverless SQL endpoint and processed as if was coming from any regular relational database.

Microsoft

Querying a data lake store using serverless SQL pools in Azure Synapse Analytics

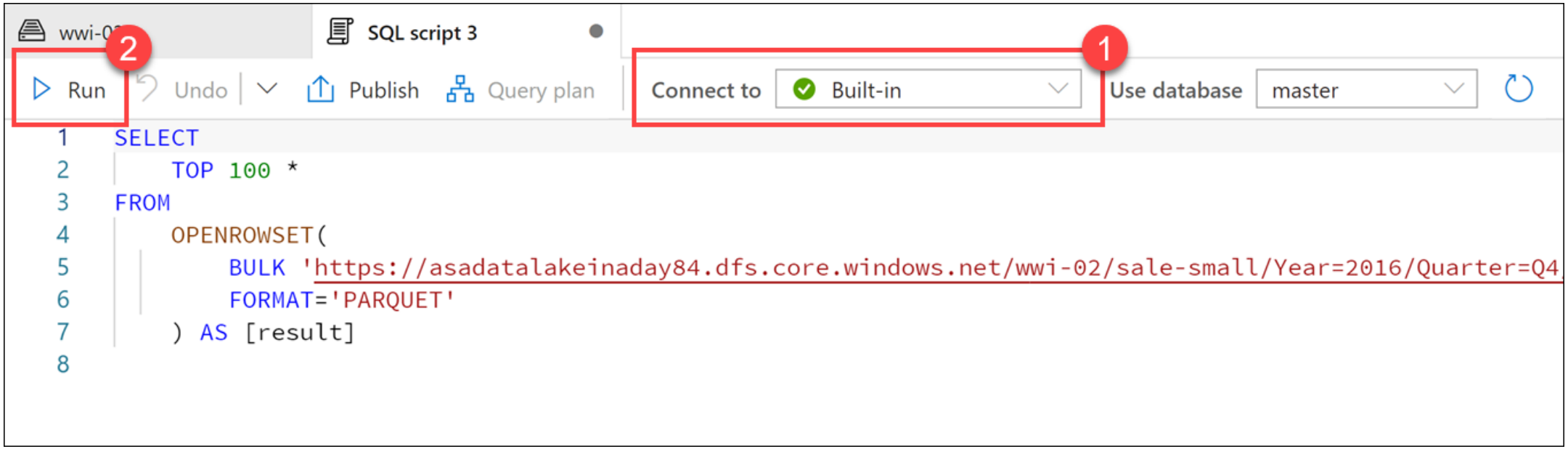
Common files to query

Parquet

Json

DelimitedText

Querying parquet files in a data lake



Microsoft

Securing access to data through using SQL serverless in Azure Synapse Analytics

Securing access to data in a data lake when using Azure Synapse Analytics

Demo: Run interactive queries using Azure Synapse serverless SQL pools

/learn alert

Complete interactive learning exercises, watch videos, and practice and apply your new skills.

aka.ms/mslearnasaserverless