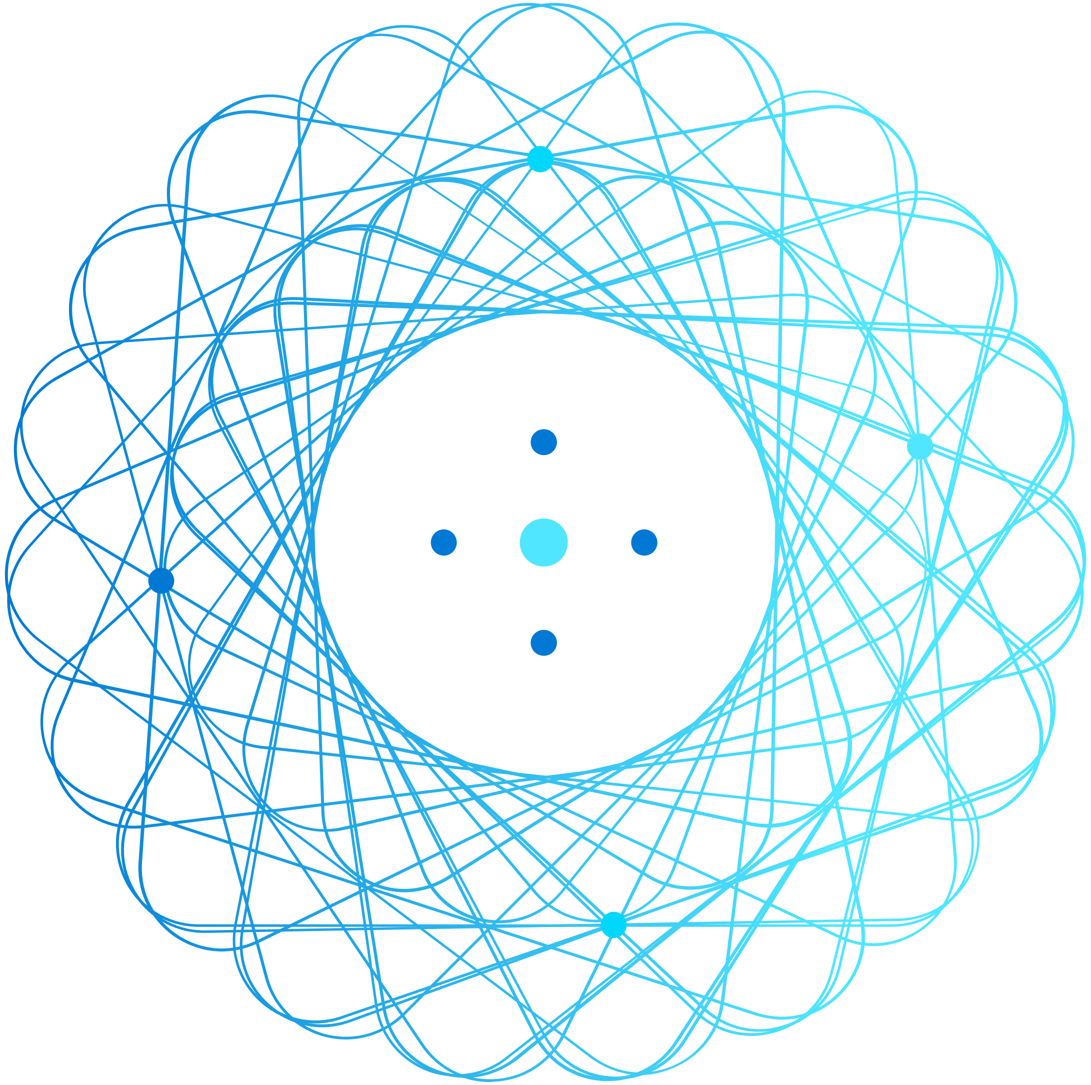
Microsoft Azure

Get started with data engineering on Azure



Agenda

Introduction to data engineering on Azure

Introduction to Azure Data Lake Storage Gen2

Introduction to Azure Synapse Analytics

Introduction to data engineering on Azure

What is data engineering?

Data engineers work with multiple types of data to perform a variety of data operations using a range of tools and scripting languages

Types of data

- Structured
- Semi-structured
- Unstructured

Data operations

- Integration
- Transformation
- Consolidation

Languages

- SQL

`SELECT…`

- Python

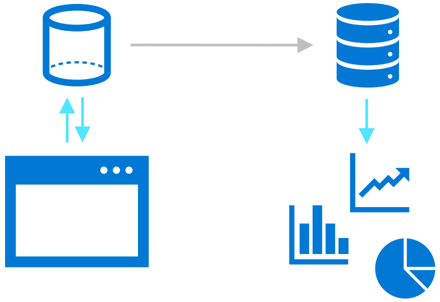
`df=spark.read(…)`

- Others

R Java .NET Scala

Important data engineering concepts

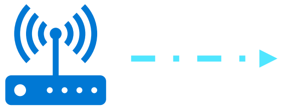
Operational and analytical data



Operational: Transactional data used by applications

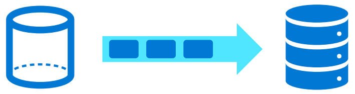
Analytical: Optimized for analysis and reporting

Streaming data



Perpetual, real-time data feeds

Data pipeline



Orchestrated activities to transfer and transform data.

Used to implement *extract, transform, and load* (ETL) or *extract, load, and transform* (ELT) operations.

Data Lake

Analytical data stored in files

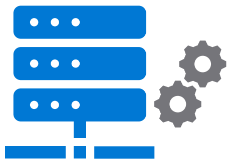
Distributed storage for massive scalability

Data Warehouse

Analytical data stored in a relational database

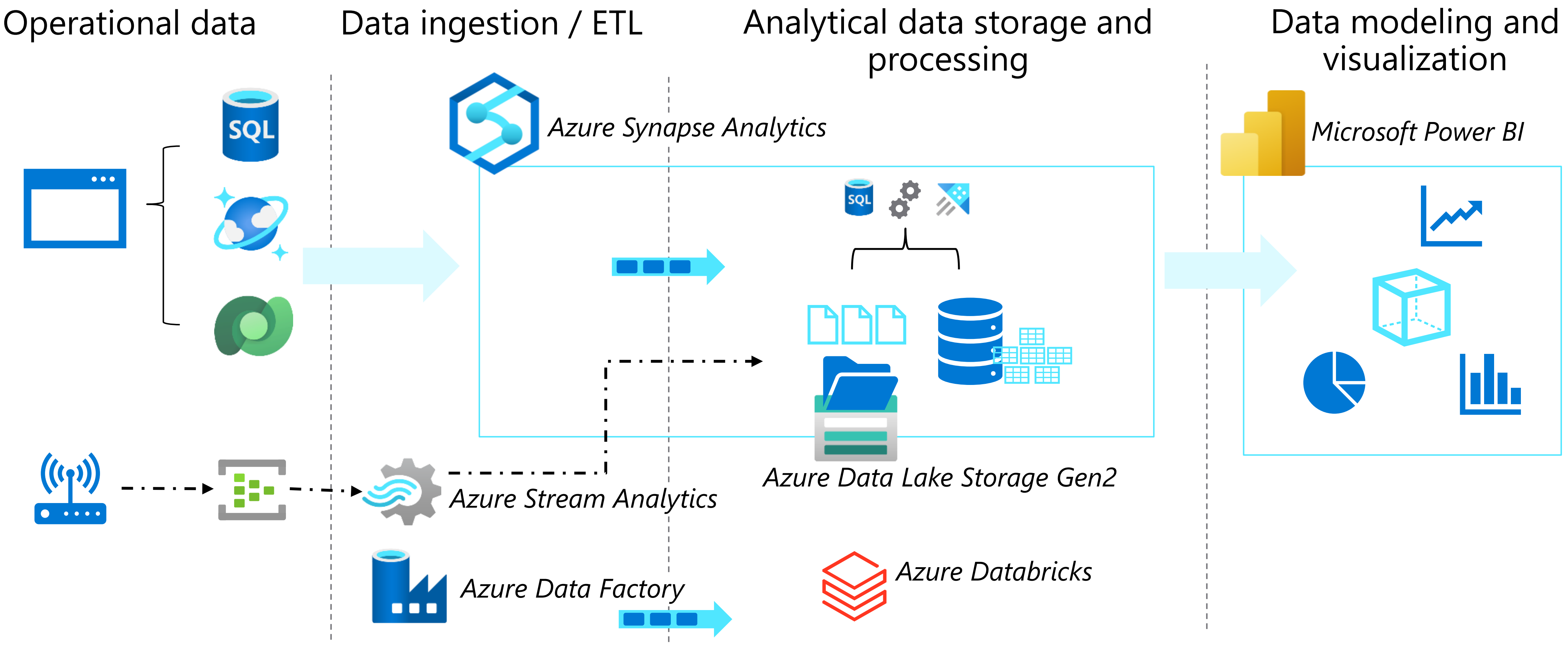
Typically modeled as a *star schema* to optimize summary analysis

Apache Spark



Open-source engine for distributed data processing

Data engineering in Azure



Knowledge check

Data in a relational database table is…

- ☑ Structured
- □ Semi-structured
- □ Unstructured

In a data lake, data is stored in…

- □ Relational tables
- ☑ Files
- □ A single JSON document

Which of the following Azure services provides capabilities for running data pipelines AND managing analytical data in a data lake or relational data warehouse?

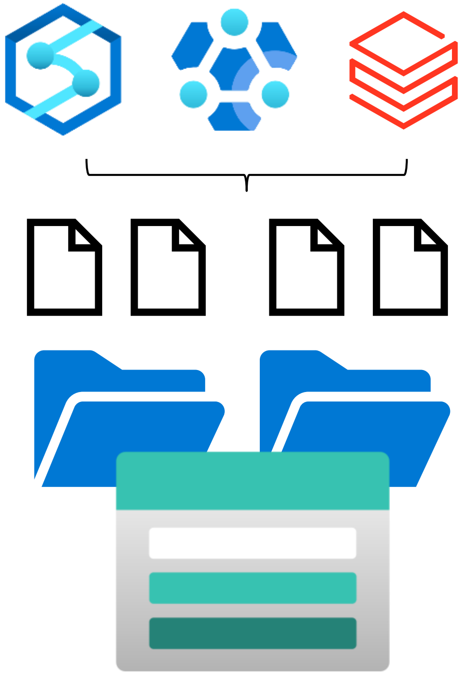
- □ Azure Stream Analytics
- ☑ Azure Synapse Analytics
- □ Azure Databricks

Introduction to Azure Data Lake Storage Gen2

Understand Azure Data Lake Storage Gen2

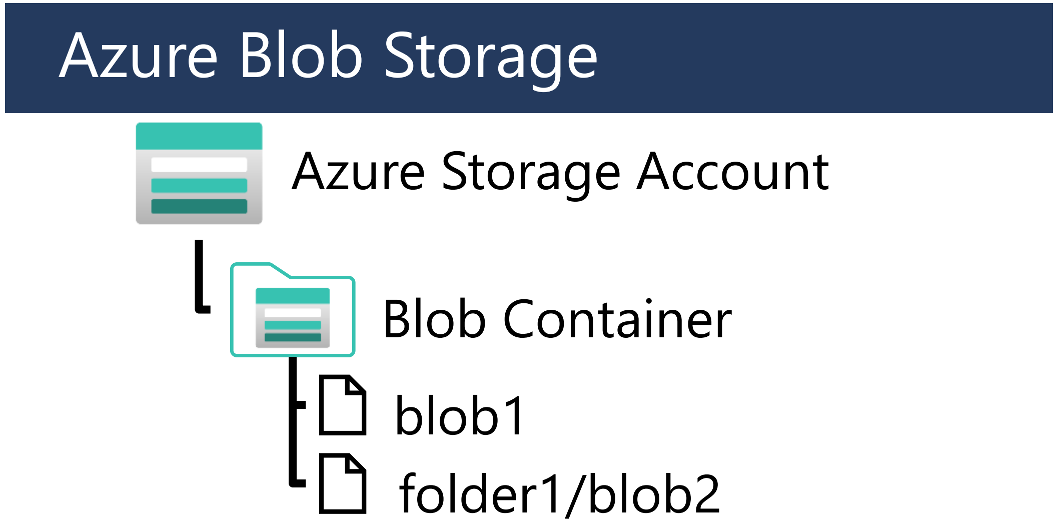
Distributed cloud storage for data lakes

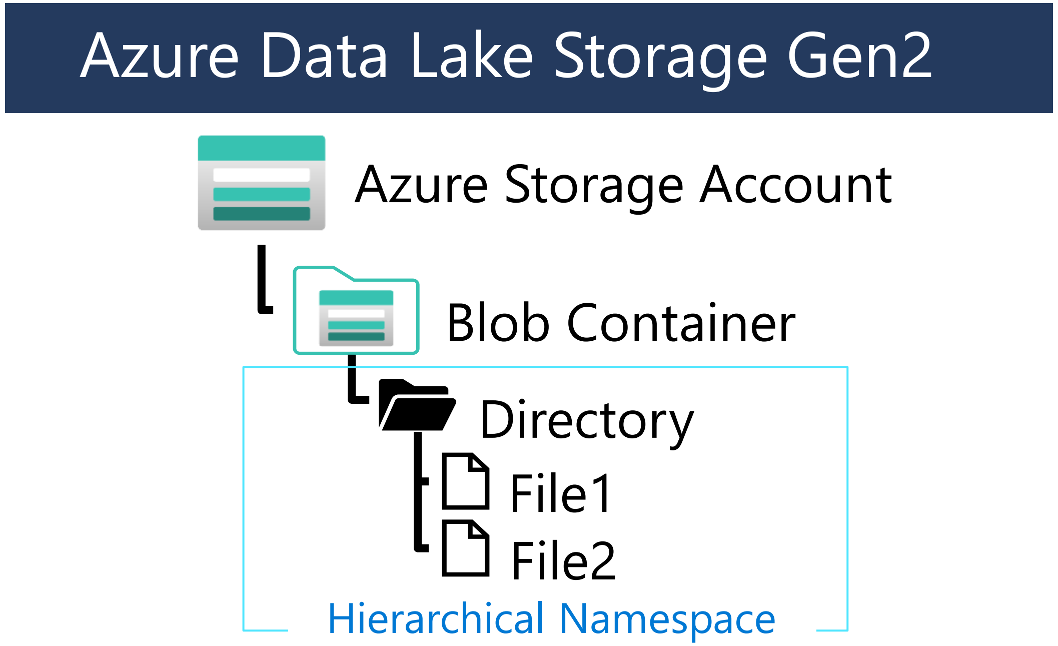
- HDFS-compatibility - common file system for Hadoop, Spark, and others
- Flexible security through folder and file level permissions
- Built on Azure Storage:
  - High performance and scalability
  - Data redundancy through built-in replication



Azure Data Lake Storage Gen 2 vs Azure Blob Storage

Enable *Hierarchical Namespace* in a blob container to use Azure Data Lake Storage Gen2





Blobs can be organized in virtual directories, but each path is considered a single blob in a flat namespace – folder level operations are not supported

File system includes directories and files, and is compatible with large scale data analytics systems like Hadoop, Databricks, and Azure Synapse Analytics

Knowledge check

Azure Data Lake Storage Gen2 stores data in…

- □ A document database hosted in Azure Cosmos DB
- ☑ An HDFS-compatible file system hosted in Azure Storage
- □ A relational data warehouse hosted in Azure Synapse Analytics

What option must you enable to use Azure Data Lake Storage Gen2?

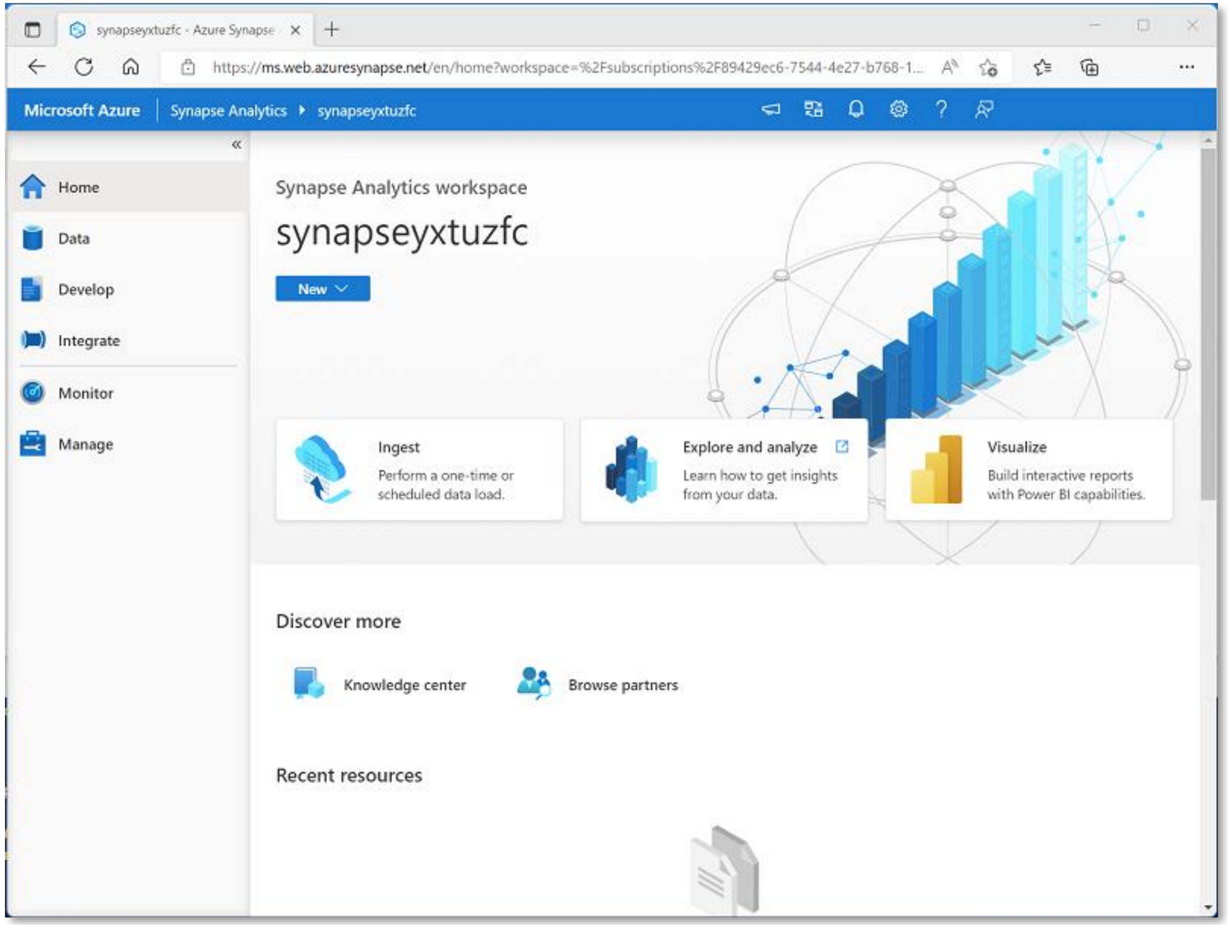
- □ Global replication
- □ Data encryption
- ☑ Hierarchical namespace

Introduction to Azure Synapse Analytics

What is Azure Synapse Analytics?

Cloud platform for data analytics

- Large-scale data warehousing
- Advanced analytics
- Data exploration and discovery
- Real time analytics
- Data integration
- Integrated analytics



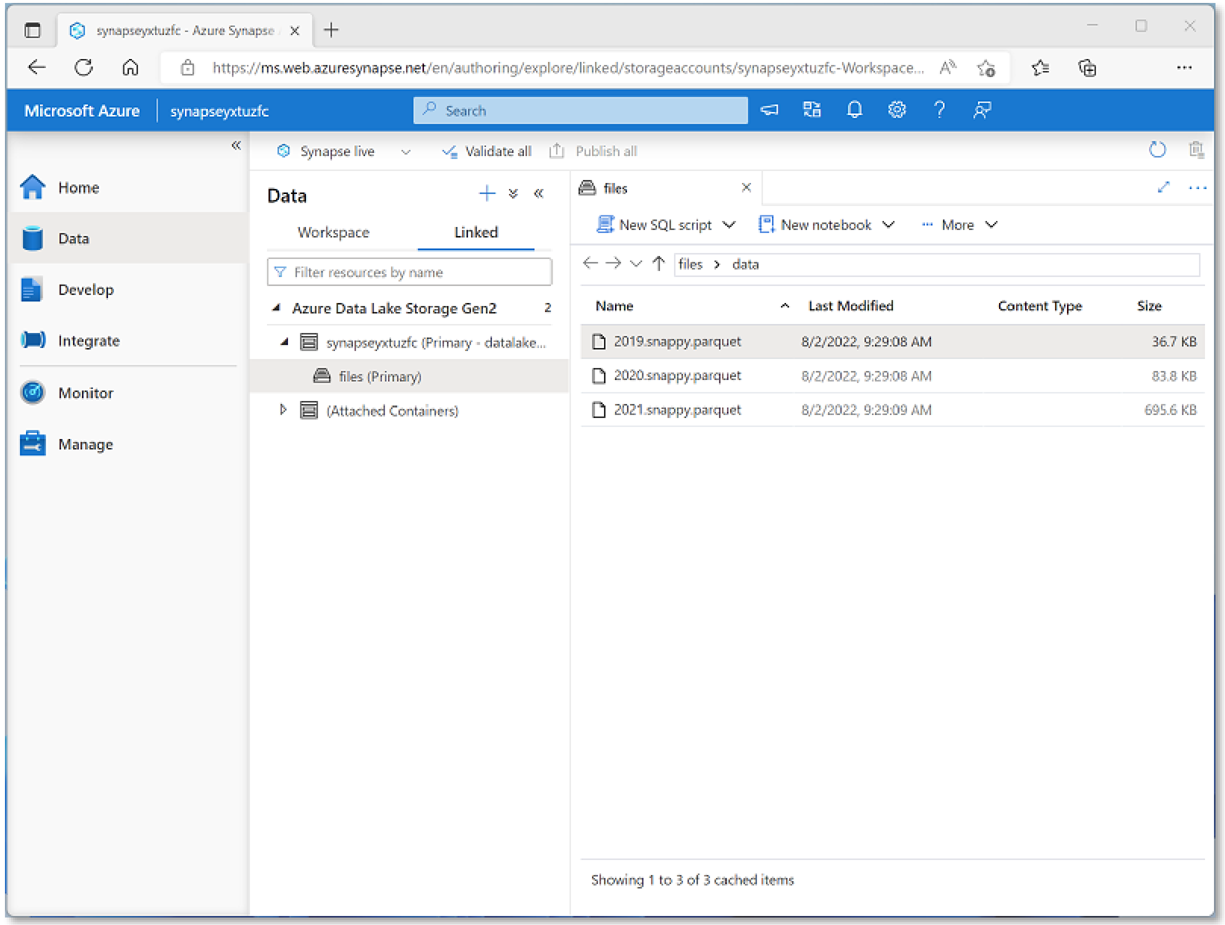
Ingest: Perform a one-time or scheduled data load.

Explore and analyze: Learn how to get insights from your data.

Visualize: Build interactive reports with Power BI capabilities.

Work with files in a data lake

- Connect to data lake storage using *linked services*
- Every Azure Synapse Analytics workspace has a default data lake



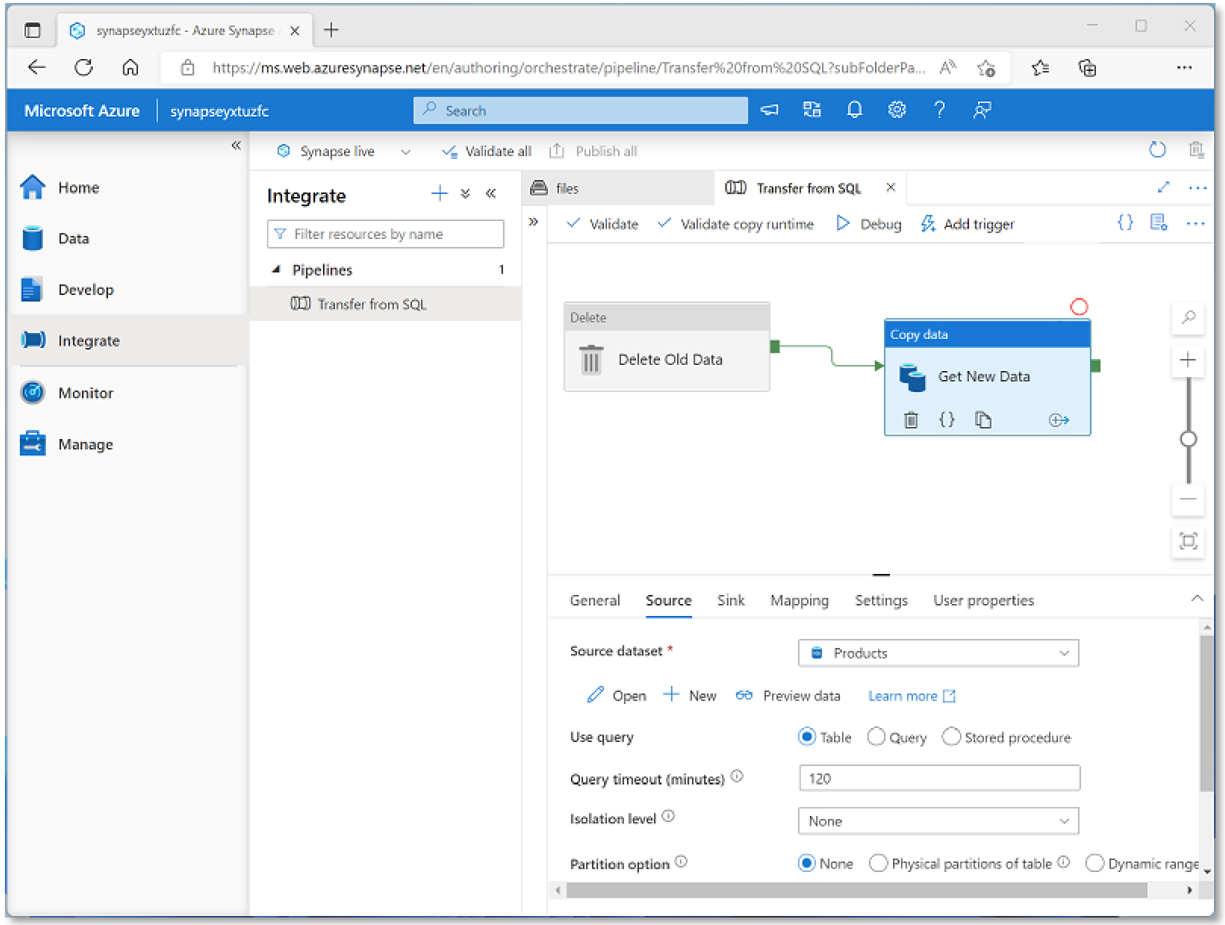
https://ms.web.azuresynapse.net/en/authoring/explore/linked/storageaccounts/synapseyxtuzfc-Workspace...

Name	Last Modified	Content Type	Size
2019.snappy.parquet	8/2/2022, 9:29:08 AM		36.7 KB
2020.snappy.parquet	8/2/2022, 9:29:08 AM		83.8 KB
2021.snappy.parquet	8/2/2022, 9:29:09 AM		695.6 KB

Showing 1 to 3 of 3 cached items

Ingest and transform data with pipelines

- Native pipeline functionality built on Azure Data Factory
- Orchestrate activities to ingest, transform, and load data
- Integrate with other data services



Source dataset * Products

Use query ● Table ○ Query ○ Stored procedure

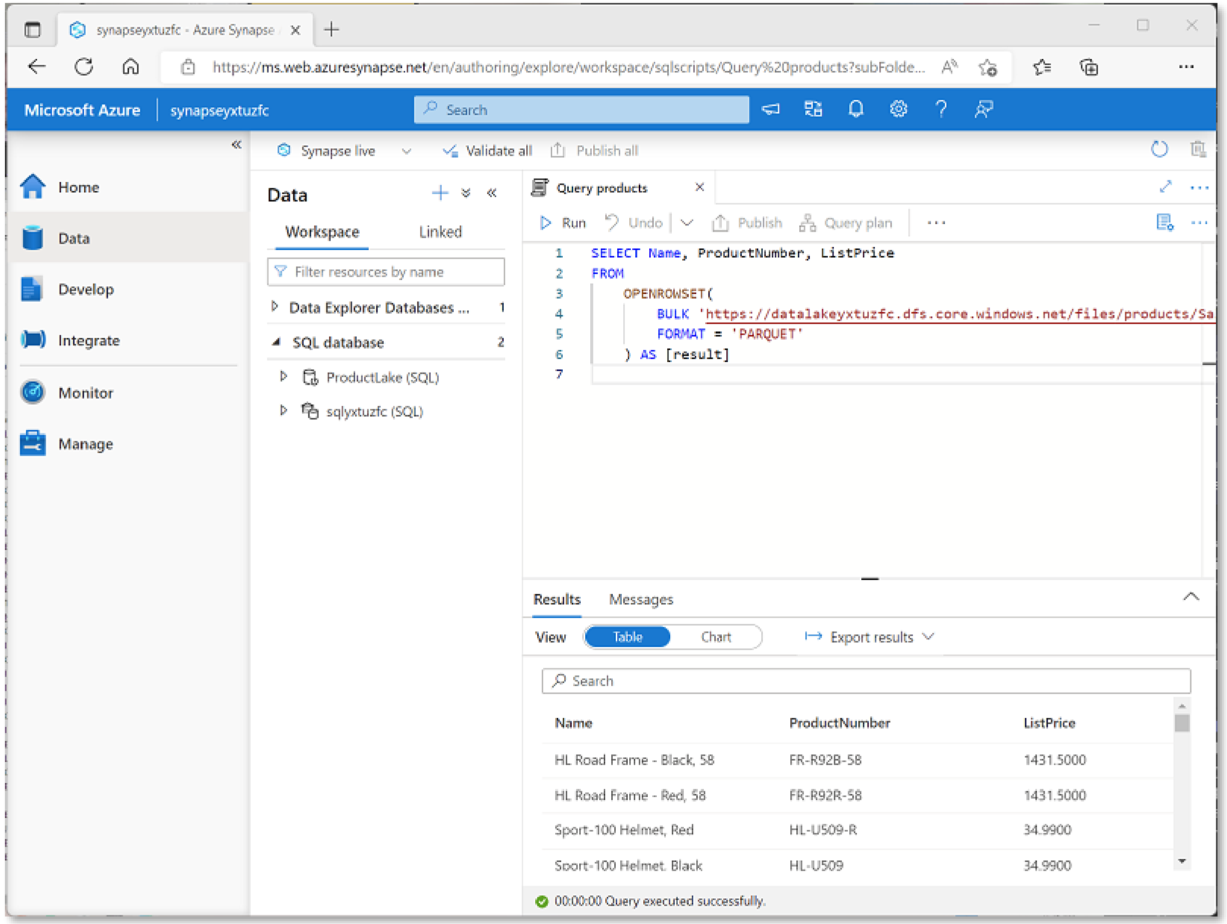
Query timeout (minutes) 120

Isolation level None

Partition option ● None ○ Physical partitions of table ○ Dynamic range

Query and manipulate data with SQL

- SQL Server based pools for scalable relational data processing:
    - Built-in *serverless* SQL pool for data exploration and analysis of files in the data lake
    - Custom *dedicated* SQL pools to host large-scale relational data warehouses



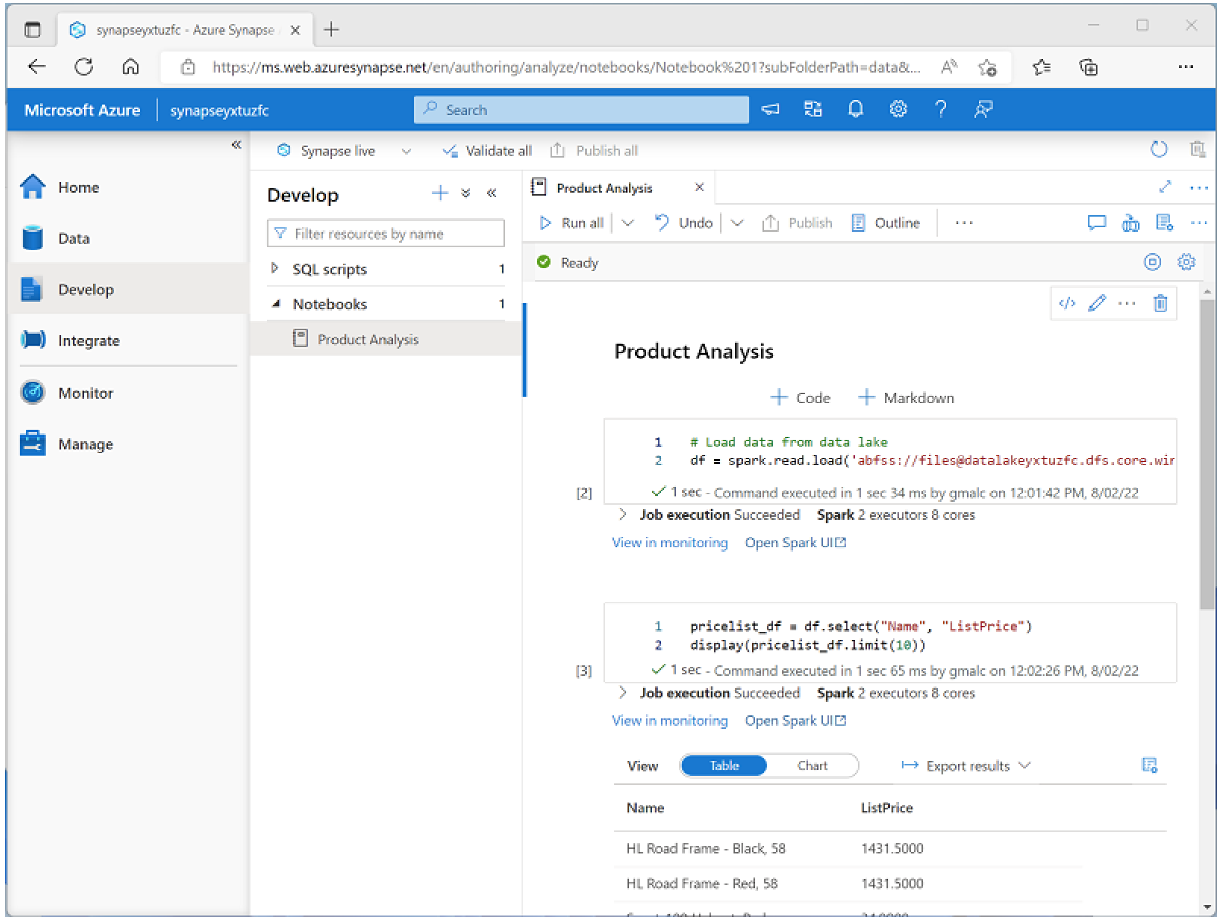
```sql
SELECT Name, ProductNumber, ListPrice
FROM
    OPENROWSET(
        BULK 'https://datalakeyxtuzfc.dfs.core.windows.net/files/products/Sa
        FORMAT = 'PARQUET'
    ) AS [result]
```

Name	ProductNumber	ListPrice
HL Road Frame - Black, 58	FR-R92B-58	1431.5000
HL Road Frame - Red, 58	FR-R92R-58	1431.5000
Sport-100 Helmet, Red	HL-U509-R	34.9900
Sport-100 Helmet, Black	HL-U509	34.9900

00:00:00 Query executed successfully.

Process and analyze data with Apache Spark

- Open-source Spark technology
  - Highly scalable, distributed processing
  - Common libraries and multiple programming languages
- Integrated notebook experience



```
# Load data from data lake
df = spark.read.load('abfss://files@datalakeyxtuzfc.dfs.core.wir
```

[2] ✓ 1 sec - Command executed in 1 sec 34 ms by gmalc on 12:01:42 PM, 8/02/22

Job execution Succeeded   Spark 2 executors 8 cores

View in monitoring   Open Spark UI

```
pricelist_df = df.select("Name", "ListPrice")
display(pricelist_df.limit(10))
```

[3] ✓ 1 sec - Command executed in 1 sec 65 ms by gmalc on 12:02:26 PM, 8/02/22

Job execution Succeeded   Spark 2 executors 8 cores

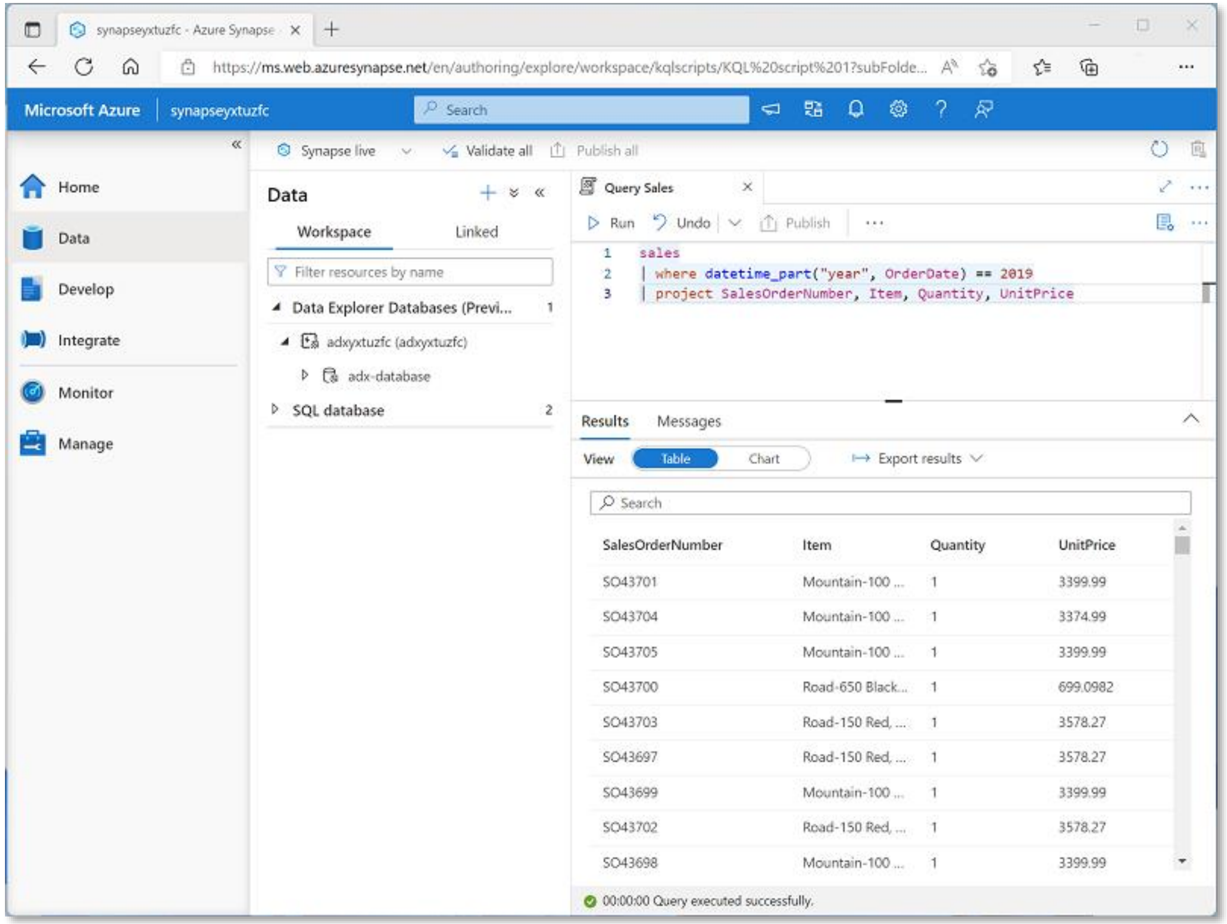
View in monitoring   Open Spark UI

Name	ListPrice
HL Road Frame - Black, 58	1431.5000
HL Road Frame - Red, 58	1431.5000

© Copyright Microsoft Corporation. All rights reserved.

Explore data with Data Explorer

- High-performance real-time data analytics
- Powerful, intuitive Kusto query language



```
sales
| where datetime_part("year", OrderDate) == 2019
| project SalesOrderNumber, Item, Quantity, UnitPrice
```

SalesOrderNumber	Item	Quantity	UnitPrice
SO43701	Mountain-100 ...	1	3399.99
SO43704	Mountain-100 ...	1	3374.99
SO43705	Mountain-100 ...	1	3399.99
SO43700	Road-650 Black...	1	699.0982
SO43703	Road-150 Red, ...	1	3578.27
SO43697	Road-150 Red, ...	1	3578.27
SO43699	Mountain-100 ...	1	3399.99
SO43702	Road-150 Red, ...	1	3578.27
SO43698	Mountain-100 ...	1	3399.99

00:00:00 Query executed successfully.

Exercise: Explore Azure Synapse Analytics

Use the hosted lab environment provided, or view the lab instructions at the link below:

https://aka.ms/mslearn-explore-synapse

Knowledge check

Which feature of Azure Synapse Analytics enables you to transfer data from one store to another and apply transformations to the data at scheduled intervals?

- □ Serverless SQL pool
- □ Apache Spark pool
- ☑ Pipelines

You want to create a data warehouse in Azure Synapse Analytics in which the data is stored and queried in a relational data store. What kind of pool should you create?

- □ Serverless SQL pool
- ☑ Dedicated SQL pool
- □ Apache Spark pool

A data analyst wants to analyze data by using Python code combined with text descriptions of the insights gained from the analysis. What should they use to perform the analysis?

- ☑ A notebook connected to an Apache Spark pool
- □ A SQL script connected to a serverless SQL pool
- □ A KQL script connected to a Data Explorer pool

Further reading

Get started with data engineering on Azure
https://aka.ms/mslearn-data-engineer