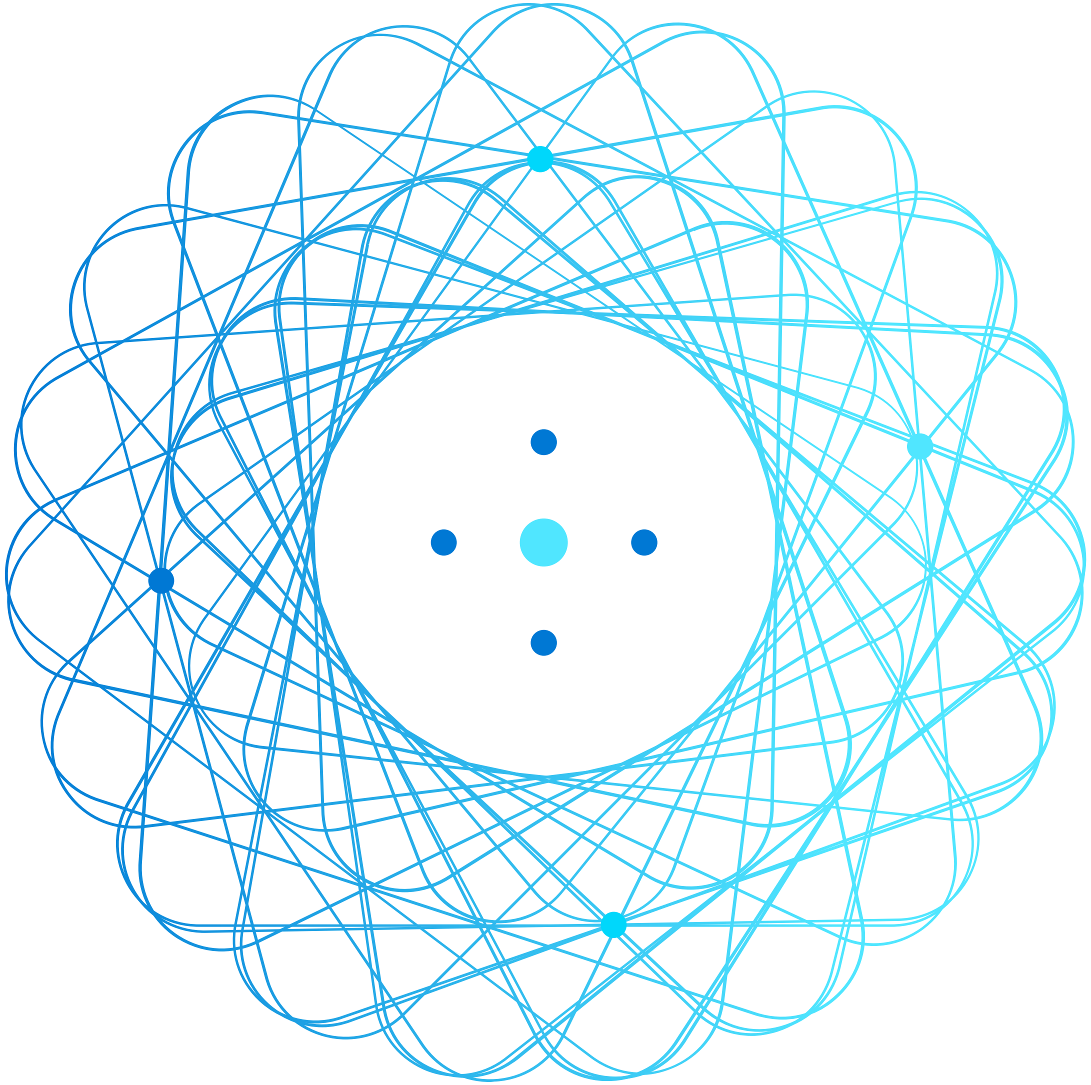
Microsoft Azure

Build data analytics solutions using Azure Synapse Analytics serverless SQL pools



Agenda

- Use a serverless SQL pool to query files in a data lake
- Use a serverless SQL pool to transform data
- Create a lake database

Use a serverless SQL pool to query files in a data lake

SQL Pools in Azure Synapse Analytics

Azure Synapse Analytics

Serverless SQL Pool

- On-demand SQL query processing
- Data stored as files in a data lake
- Typical use cases:
  - Data exploration
  - Data transformation
  - Logical data warehouse

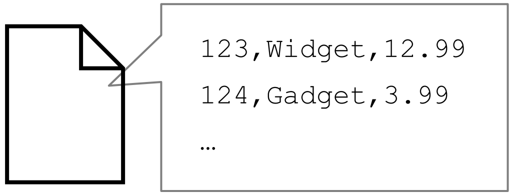
Dedicated SQL Pools

- Cloud-scale relational database
- Data stored in relational tables
- Typical use cases:
  - Relational data warehouse
  - Enterprise business intelligence

Query delimited text files using a serverless SQL pool

Use the OPENROWSET function

- Use the BULK parameter specifies file path(s)
  - Include wildcards as required
- Use the FORMAT parameter to specify 'csv'
- Use additional parameters for:
  - Header row
  - Delimiter characters
  - Parser version
  - others...
- Use the WITH clause to specify column names and types



```
123,Widget,12.99
124,Gadget,3.99
...
```

```
SELECT *
FROM OPENROWSET(
    BULK 'https://.../data/files/*.csv',
    FORMAT = 'csv',
    PARSER_VERSION = '2.0')
WITH (
    product_id INT,
    product_name VARCHAR(20),
    list_price DECIMAL(5,2)
) AS rows
```

product_id	**product_name**	**list_price**
123	Widget	12.99
124	Gadget	3.99
...	...	...

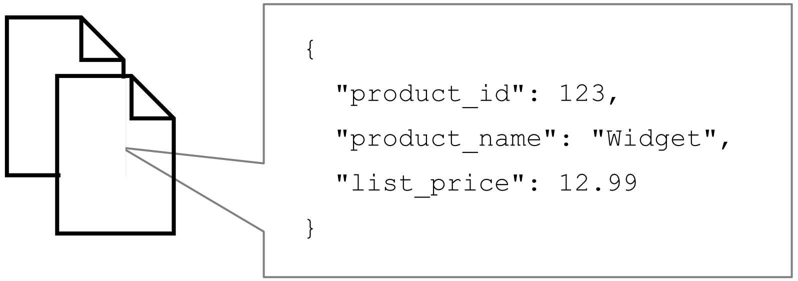
Query JSON files using a serverless SQL pool

Use the OPENROWSET function

- Use the BULK parameter specifies file path(s)
  - Include wildcards as required
- Use the FORMAT parameter to specify 'csv'
- Set terminators to '0x0b'
- Use the WITH clause to specify a single NVARCHAR column

Use JSON_VALUE function to specify JSON properties

- Specify attribute path based on JSON in the NVARCHAR column



```
{
  "product_id": 123,
  "product_name": "Widget",
  "list_price": 12.99
}
```

```
SELECT JSON_VALUE(doc, '$.product_name') AS product,
          JSON_VALUE(doc, '$.list_price') AS price
FROM
   OPENROWSET(
      BULK 'https://.../data/files/*.json',
      FORMAT = 'csv',
      FIELDTERMINATOR ='0x0b',
      FIELDQUOTE = '0x0b',
      ROWTERMINATOR = '0x0b'
   ) WITH (doc NVARCHAR(MAX)) as rows
```

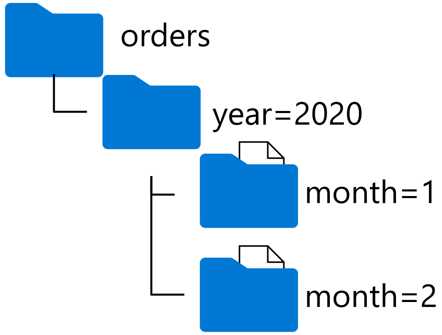
product	**price**
Widget	12.99
Gadget	3.99
…	…

Query parquet files using a serverless SQL pool

Use the OPENROWSET function

- Use the BULK parameter specifies file path(s)
  - Include wildcards as required
- Use the FORMAT parameter to specify 'parquet'

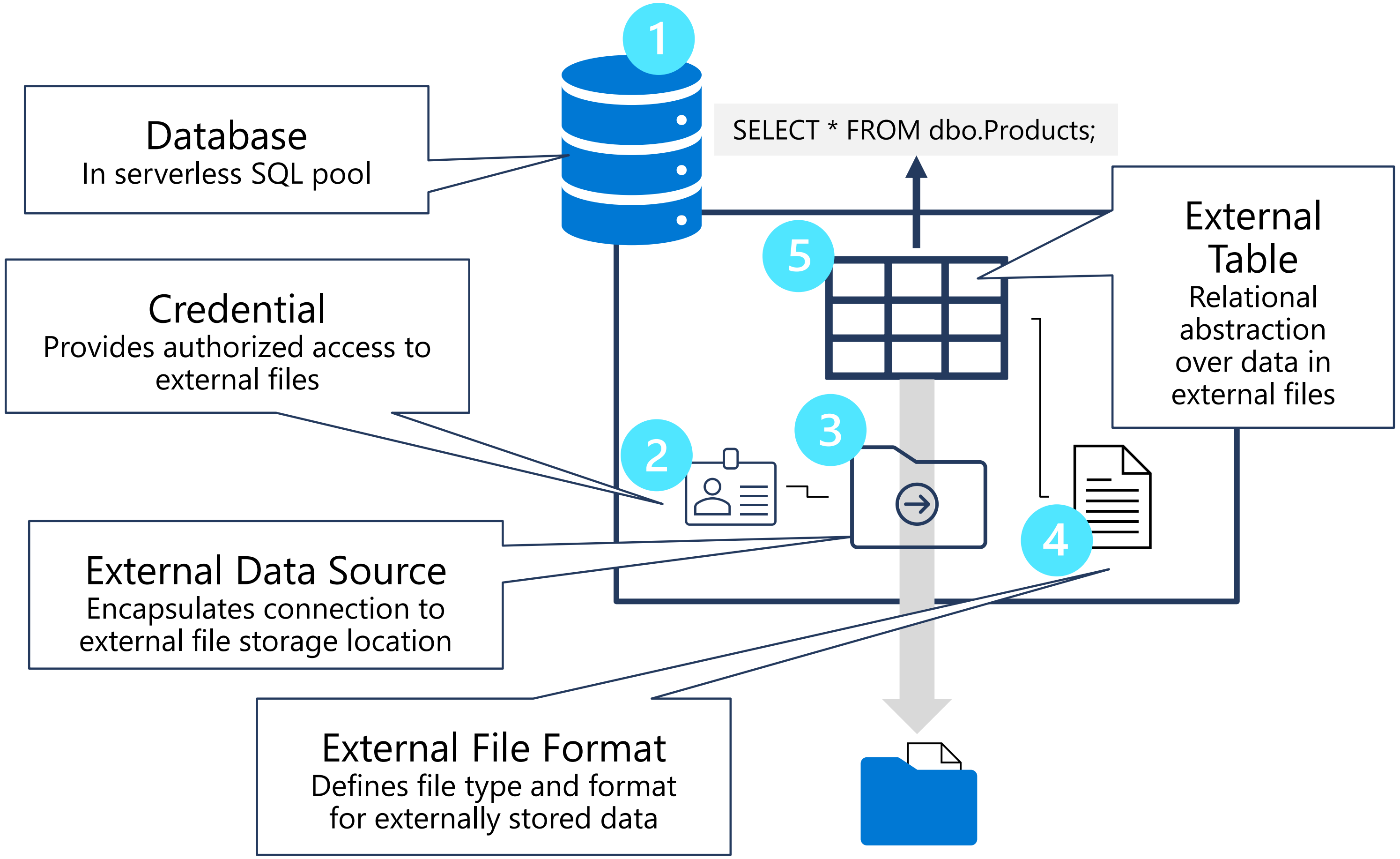
- **Use *filepath* property to filter by partitions**
  - Parameters reflect ordinal position of wildcards
  - Not specific to parquet, but commonly used to distribute data in parquet format



```
SELECT *
FROM OPENROWSET(
   BULK 'https://…/data/orders/year=*/month=*/*.*’,
   FORMAT = 'parquet’) AS orders
WHERE orders.filepath(1) = '2020’
   AND orders.filepath(2) IN ('1','2');
```

order_no	order_date	order_total
1001	2020-01-07	99.78
1002	2020-01-12	11.99
…	…	…

Create external database objects



Database
In serverless SQL pool

Credential
Provides authorized access to external files

External Data Source
Encapsulates connection to external file storage location

External File Format
Defines file type and format for externally stored data

External Table
Relational abstraction over data in external files

```
1  CREATE DATABASE SalesDB
      COLLATE Latin1_General_100_BIN2_UTF8;

   USE SalesDB;

2  CREATE DATABASE SCOPED CREDENTIAL sqlcred
      WITH
          IDENTITY='SHARED ACCESS SIGNATURE',
          SECRET = 'sv=xxx...';

3  CREATE EXTERNAL DATA SOURCE files
   WITH ( LOCATION =
   'https://mydatalake.blob.core.windows.net/data/files/',
            CREDENTIAL = sqlcred);

4  CREATE EXTERNAL FILE FORMAT CsvFormat
      WITH ( FORMAT_TYPE = DELIMITEDTEXT,
               FORMAT_OPTIONS(
                   FIELD_TERMINATOR = ',',
                   STRING_DELIMITER = '"'));

5  CREATE EXTERNAL TABLE dbo.products
   (
      product_id INT,
      product_name VARCHAR(20),
      list_price DECIMAL(5,2)
   )
   WITH
   (
      DATA_SOURCE = files,
      LOCATION = 'products/*.csv',
      FILE_FORMAT = CsvFormat
   );
```

Demo: Query files using a serverless SQL pool

You can try this for yourself later by following the instructions at the link below:

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

Knowledge check

What function is used to read the data in files stored in a data lake?

- □ FORMAT
- □ ROWSET
- ☑ OPENROWSET

What character in file path can be used to select all the file/folders that match rest of the path?

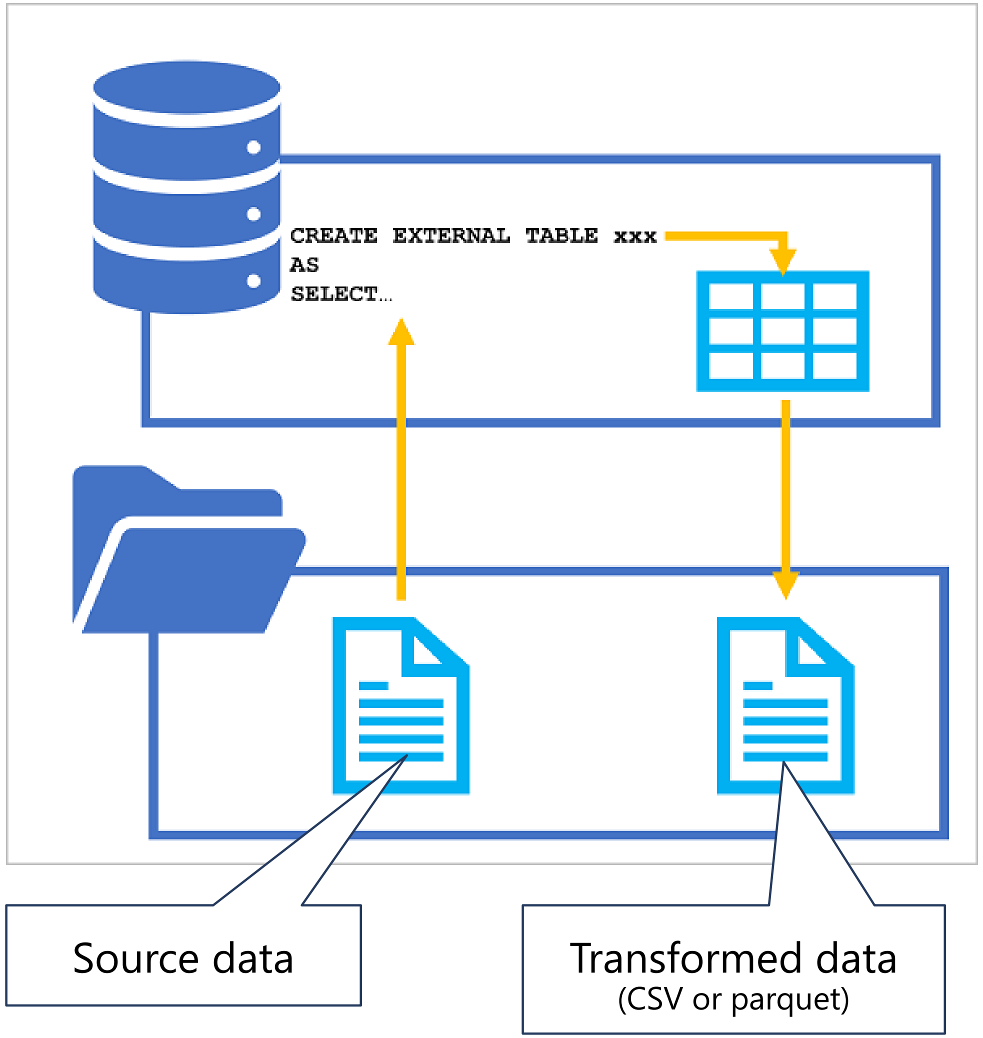
- □ &
- ☑ *
- □ /

Which external database object encapsulates the connection information to a file location in a data lake store?

- □ FILE FORMAT
- ☑ DATA SOURCE
- □ EXTERNAL TABLE

Use a serverless SQL pool to transform data

The CREATE EXTERNAL TABLE AS SELECT (CETAS) statement



```
CREATE EXTERNAL TABLE xxx
AS
SELECT…
```

Source data

Transformed data
(CSV or parquet)

```
CREATE EXTERNAL TABLE SpecialOrders
   WITH (
      -- details for storing results
      LOCATION = 'special_orders/',
      DATA_SOURCE = files,
      FILE_FORMAT = ParquetFormat
   )
AS
SELECT OrderID, CustomerName, OrderTotal
FROM
   OPENROWSET (
      -- details for reading source files
      BULK 'sales_orders/*.csv',
      DATA_SOURCE = 'files',
      FORMAT = 'CSV',
      PARSER_VERSION = '2.0',
      HEADER_ROW = TRUE
   ) AS source_data
WHERE OrderType = 'Special Order';
```

Encapsulate data transformations in a stored procedure

Using a stored procedure:

- Reduces client to server network traffic
- Provides a security boundary
- Eases maintenance
- Improved performance

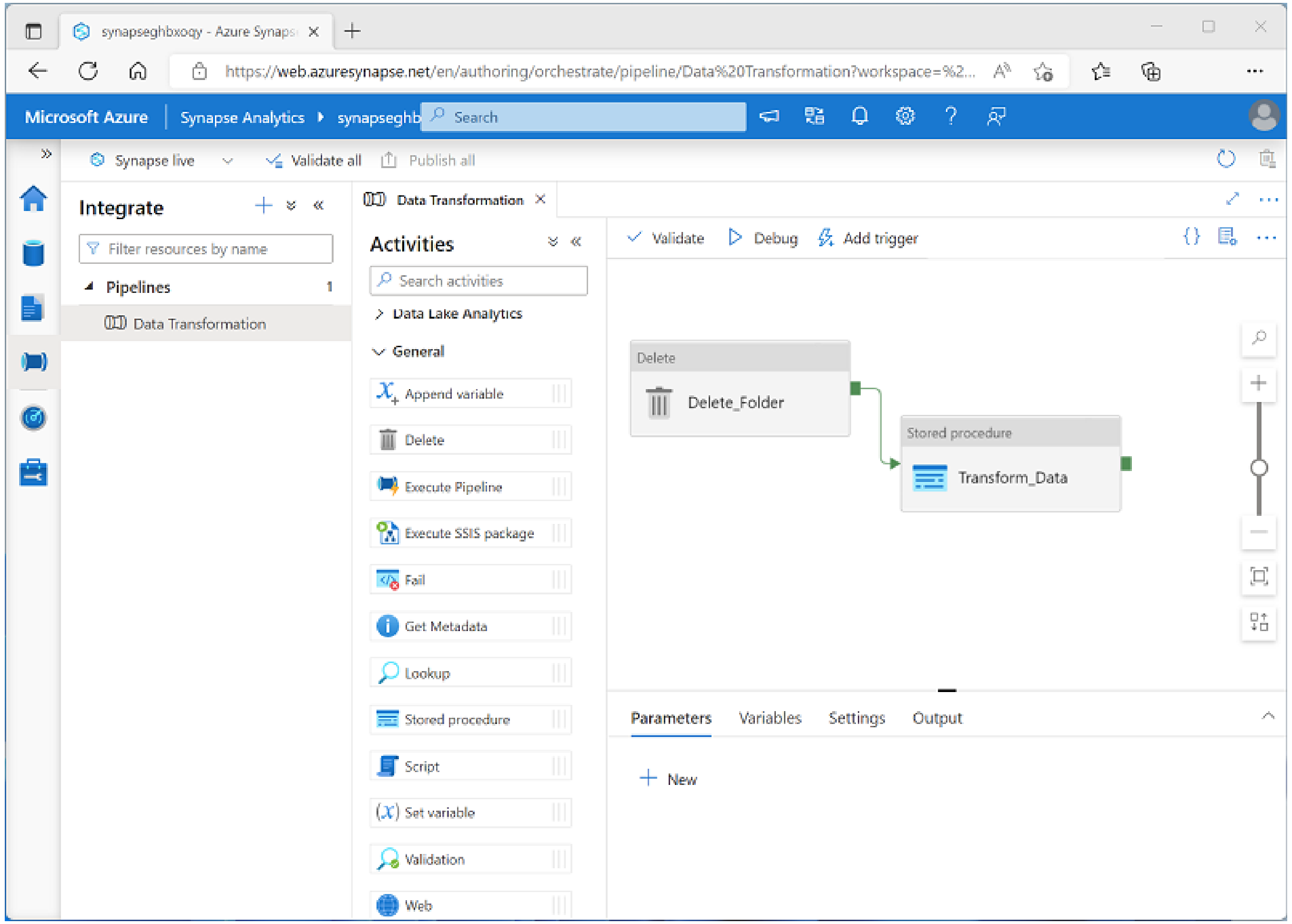
```
CREATE PROCEDURE Transform_Data @order_year INT
AS
BEGIN

  -- Drop the table if it already exists
  IF EXISTS (
            SELECT * FROM sys.external_tables
            WHERE name = 'SpecialOrders'
           )
     DROP EXTERNAL TABLE SpecialOrders

  -- Create external table
  CREATE EXTERNAL TABLE SpecialOrders
  WITH (
         ...
       )
END
```

⚠ Dropping the table doesn’t delete the underlying files

Include a data transformation stored procedure in a pipeline

Create a pipeline with the following activities:

- A **Delete** activity that deletes the target folder for the transformed data in the data lake if it already exists.
- A **Stored procedure** activity that connects to your serverless SQL pool and runs the stored procedure that encapsulates your CETAS operation.



Exercise: Transform files using a serverless SQL pool

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

https://aka.ms/mslearn-synapse-transform-sql

Knowledge check

You need to store the results of a query in a serverless SQL pool as files in a data lake. Which SQL statement should you use?

- ☐ BULK INSERT
- ☑ CREATE EXTERNAL TABLE AS SELECT
- ☐ COPY

Which of the following file formats can you use to persist the results of a query?

- ☐ CSV only
- ☐ Parquet only
- ☑ CSV and parquet

You drop an existing external table from a database in a serverless SQL pool. What else must you do before recreating an external table with the same location?

- ☑ Delete the folder containing the data files for dropped table
- ☐ Drop and recreate the database
- ☐ Create an Apache Spark pool

Create a lake database

Lake database concepts

Lake database schema:

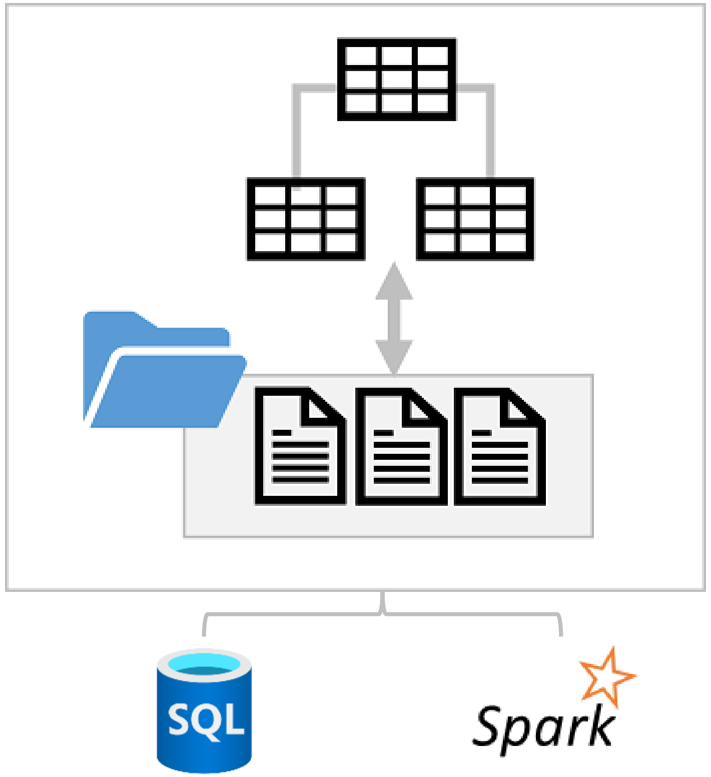
- Relational tables
- Proven data modeling principles
- Consistent naming conventions

Lake database storage:

- Parquet or CSV files in a data lake
- Managed independently of database
- Simplified data ingestion

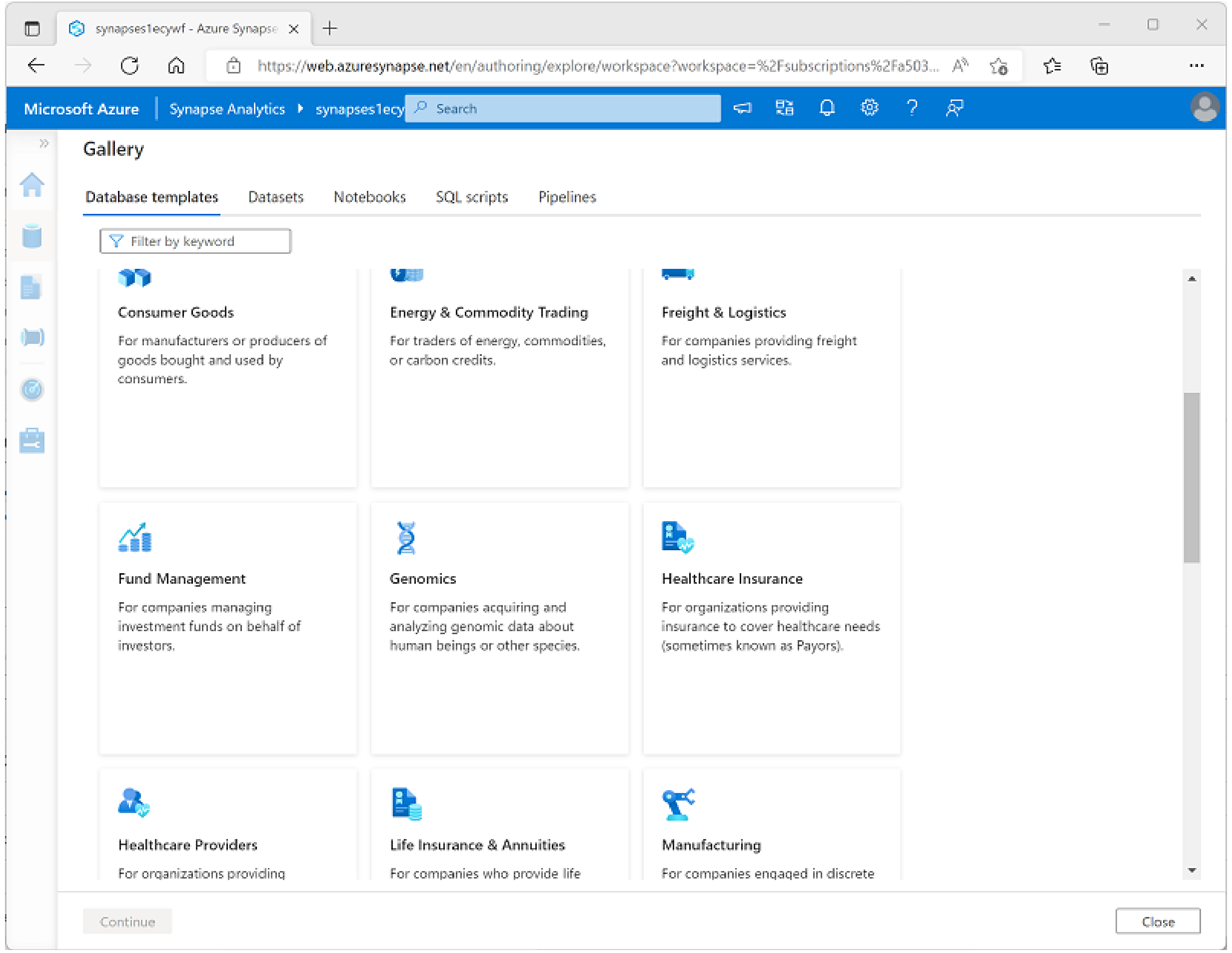
Lake database compute:

- Serverless SQL pool
- Apache Spark pool



Database templates

- Pre-defined relational schema based on common business scenarios
- Use as a template for a new database or start with a blank schema and add pre-defined table definitions



Gallery

Consumer Goods

For manufacturers or producers of goods bought and used by consumers.

Energy & Commodity Trading

For traders of energy, commodities, or carbon credits.

Freight & Logistics

For companies providing freight and logistics services.

Fund Management

For companies managing investment funds on behalf of investors.

Genomics

For companies acquiring and analyzing genomic data about human beings or other species.

Healthcare Insurance

For organizations providing insurance to cover healthcare needs (sometimes known as Payors).

Healthcare Providers

For organizations providing

Life Insurance & Annuities

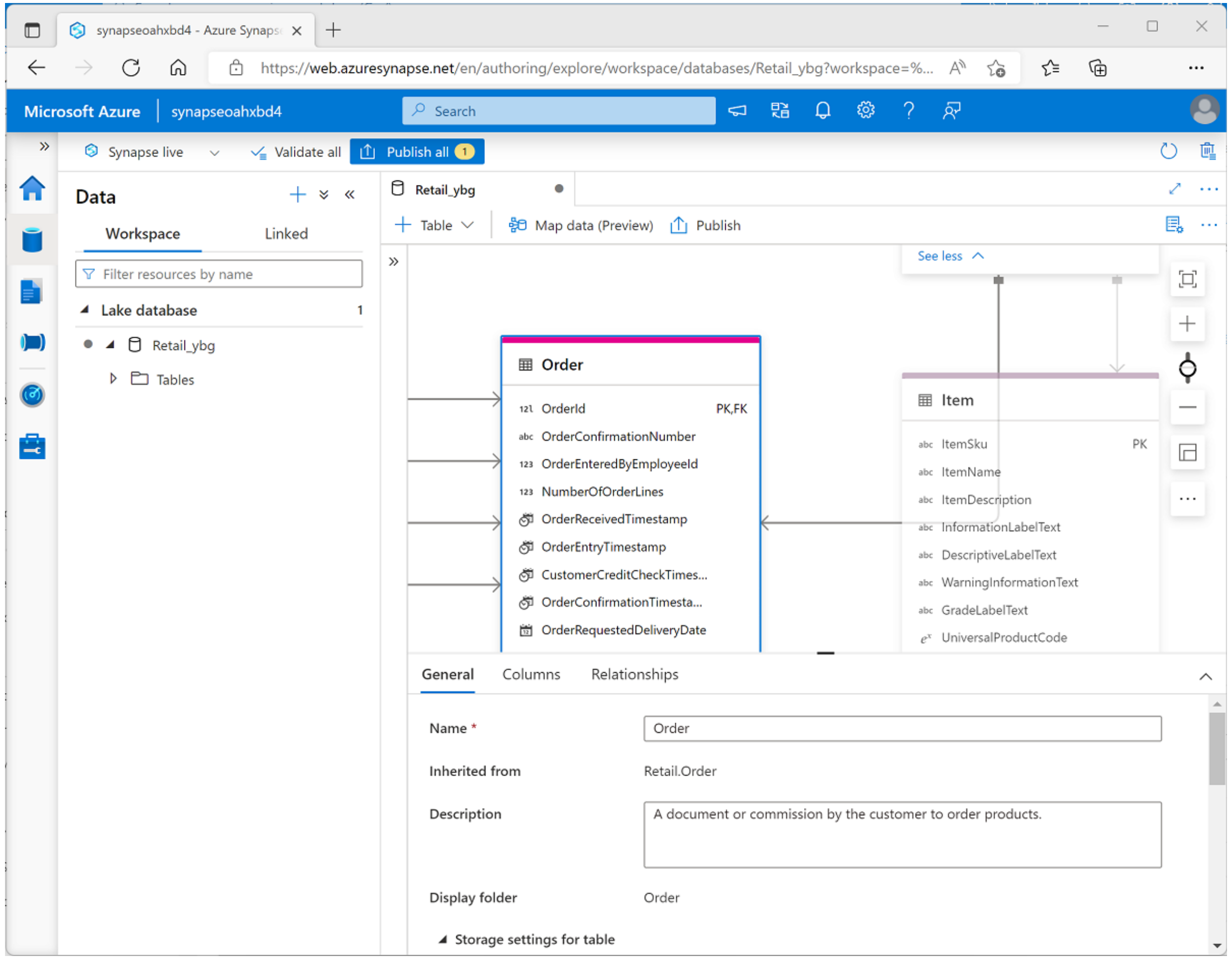
For companies who provide life

Manufacturing

For companies engaged in discrete

Database designer

- Visual tool for creating a database schema
- Add tables and specify:
  - Name and storage settings for the table
  - Names, key usage, nullability, and data types for each column
  - Relationships between key columns across tables



Order

OrderId	PK,FK
OrderConfirmationNumber	
OrderEnteredByEmployeeId	
NumberOfOrderLines	
OrderReceivedTimestamp	
OrderEntryTimestamp	
CustomerCreditCheckTimes...	
OrderConfirmationTimesta...	
OrderRequestedDeliveryDate	

Item

ItemSku	PK
ItemName	
ItemDescription	
InformationLabelText	
DescriptiveLabelText	
WarningInformationText	
GradeLabelText	
UniversalProductCode	

Name * Order

Inherited from Retail.Order

Description A document or commission by the customer to order products.

Display folder Order

Storage settings for table

Use a lake database

```
USE RetailDB;
GO

SELECT CustomerID, FirstName, LastName
FROM Customer
ORDER BY LastName;
```

```
%%sql

INSERT INTO `RetailDB`.`Customer` VALUES (123, 'John', 'Yang’)

SELECT * FROM `RetailDB`.`Customer` WHERE CustomerID = 123
```

Demo: Analyze data in a lake database

You can try this for yourself later by following the instructions at the link below:

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

Knowledge check

Which if the following statements is true of a lake database?

□ Data is stored in a relational database store and cannot be directly accessed in the data lake files

□ Data is stored in files that cannot be queried using SQL

☑ A relational schema is overlaid on the underlying files, and can be queried using a serverless SQL pool or a Spark pool

You need to create a new lake database for a retail solution. What's the most efficient way to do this?

□ Create a sample database in Azure SQL Database and export the SQL scripts to create the schema for the lake database

☑ Start with the *Retail* database template in Azure Synapse Studio, and adapt it as necessary

□ Start with an empty database and create a normalized schema

You have Parquet files in an existing data lake folder for which you want to create a table in a lake database. What should you do?

□ Use a CREATE EXTERNAL TABLE AS SELECT (CETAS) query to create the table

□ Convert the files in the folder to CSV format

☑ Use the database designer to create a table based on the existing folder

Further reading

Build data analytics solutions using Azure Synapse serverless SQL pools
https://aka.ms/mslearn-synapse-serverless-sql