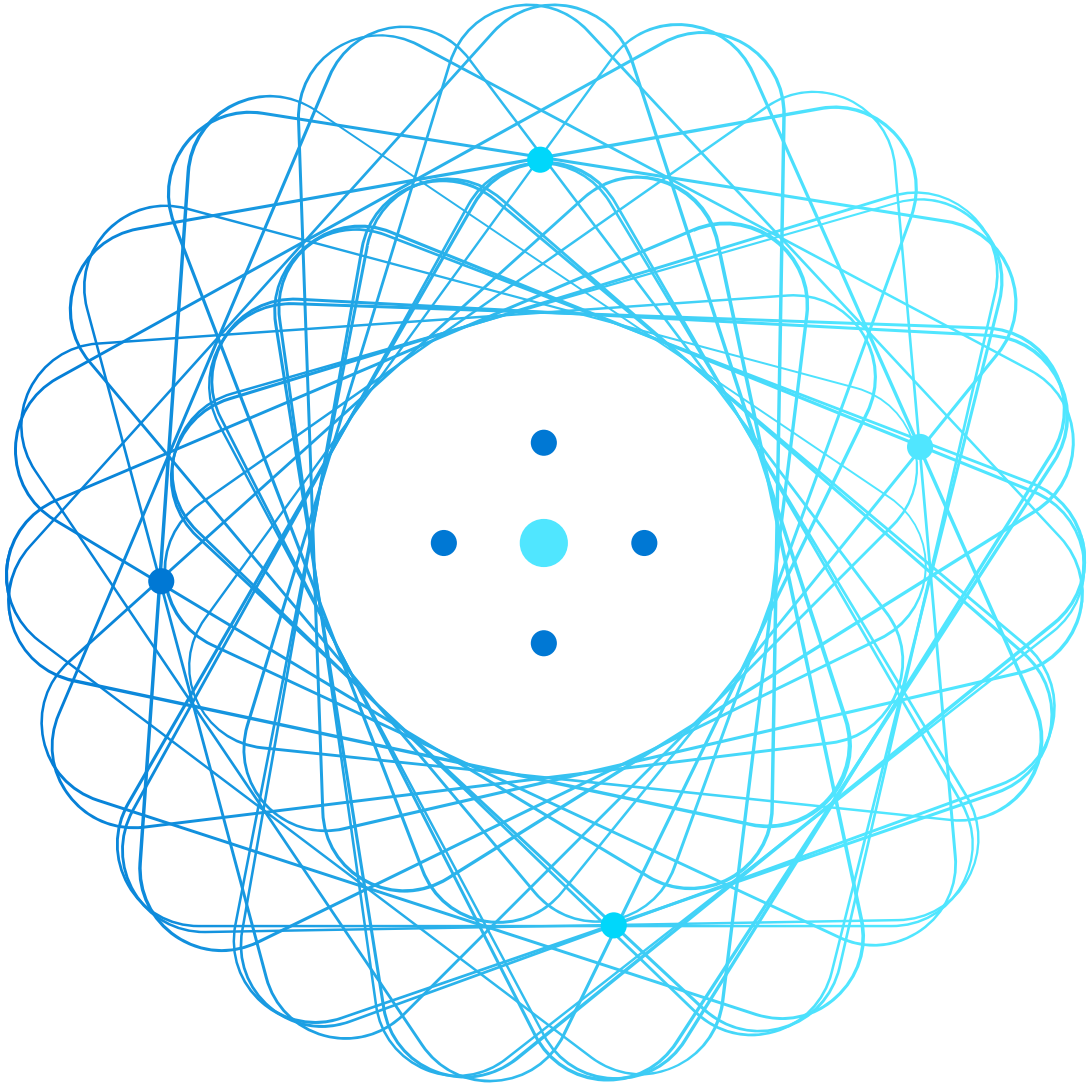
Microsoft Azure

Build data analytics solutions using Azure Synapse Analytics serverless SQL pools



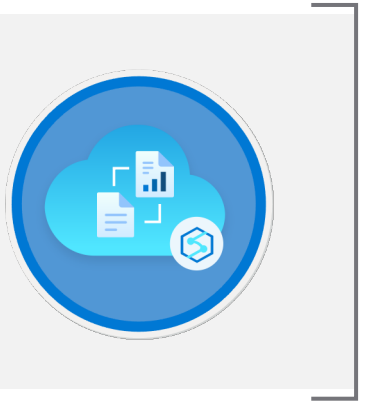
Agenda

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

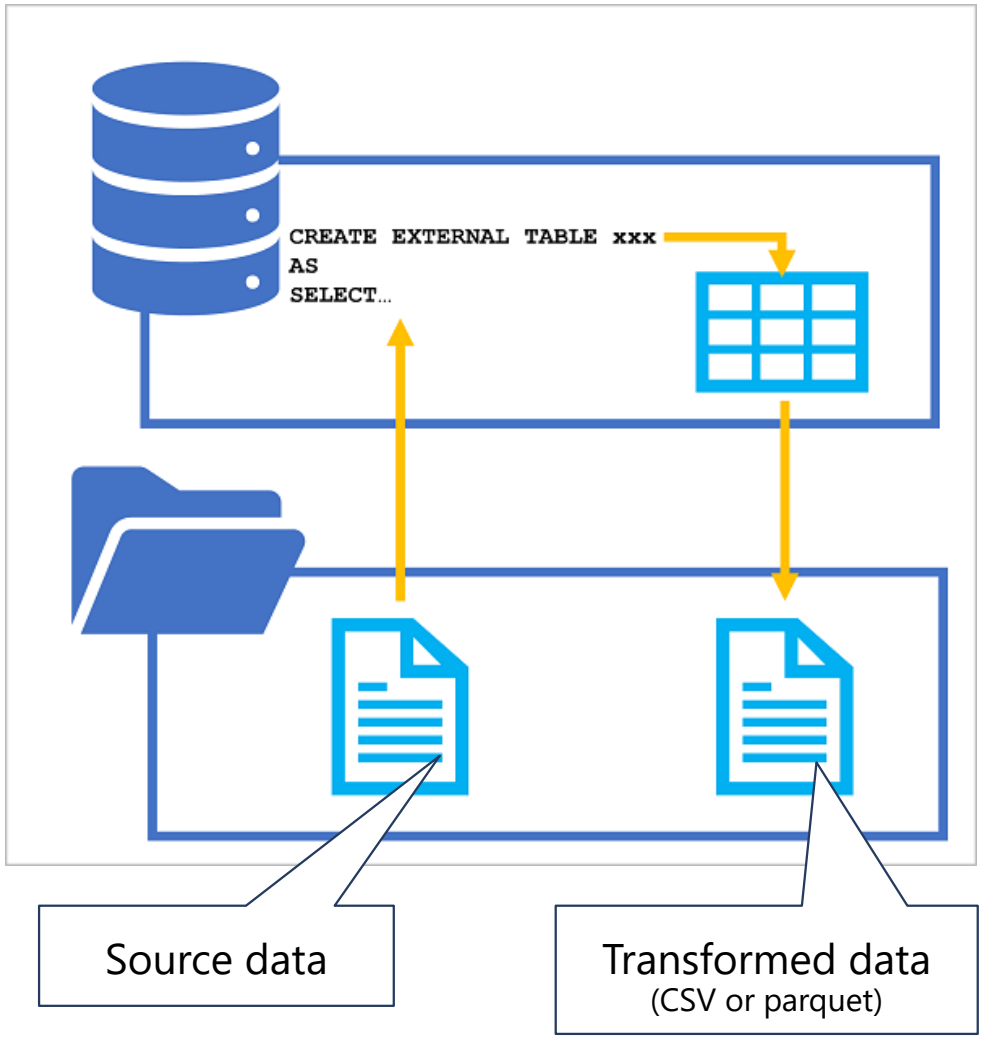
Use a serverless SQL pool to transform data

Create a lake database

Use a serverless SQL pool to transform data



The CREATE EXTERNAL TABLE AS SELECT (CETAS) statement



The CETAS statement provides a convenient way to use the results of a SQL query as the basis for a new external table, which is persisted as one or more data files in the data lake. You can use the SQL query to transform the data, and then query the resulting table or use the underlying data files in a downstream process.

```
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

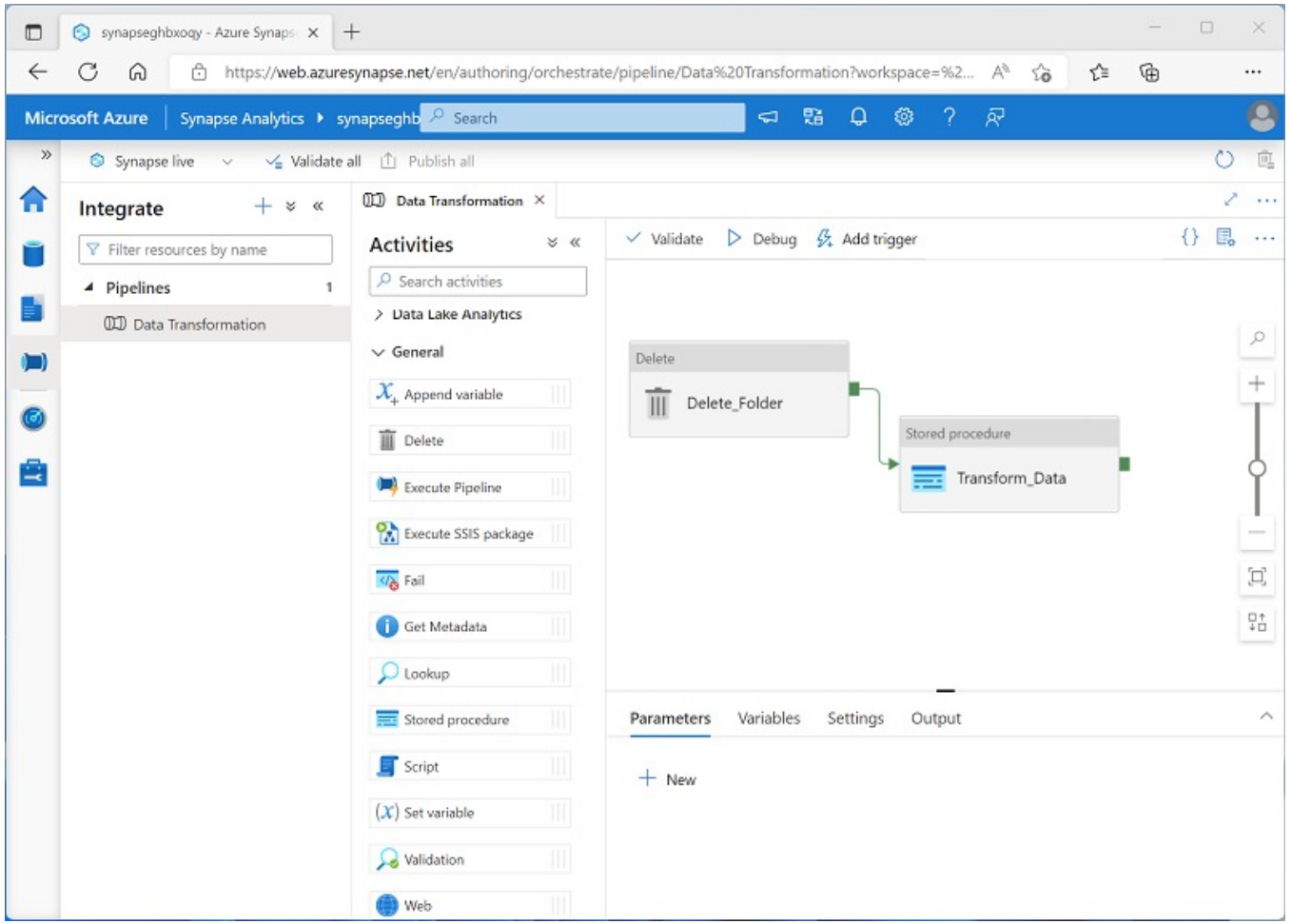
```
CREATE PROCEDURE Transform_Data @order_year INT
AS
BEGIN

  -- 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

Further reading

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