

## Azure Datalake to SQL task 2

```
▶ 07:48 AM (17s) 3 Python

# Define storage account and container details
storage_account_name = "azurenewaccount"
container_name = "sqltodatalake"
file_path = "dbo.Orders.txt"

# Ensure the storage account access key is correctly set in the Spark configuration
spark.conf.set(
    f"fs.azure.account.key.{storage_account_name}.blob.core.windows.net",
    "XexJU0ApV6eLGoc/shWgIL341gskEG09mVVi50A0U6niJqdeNEh10AMyKAjovtrclt7vo2huX0qx+AST6Z66Dg=="
)

# Read data from Blob Storage using the correct URI scheme and path
df = spark.read.format("csv") \
    .option("header", "true") \
    .option("inferSchema", "true") \
    .load(f"wasbs://{container_name}@{storage_account_name}.blob.core.windows.net/{file_path}")

# Display the data
display(df)
```

▶ (3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame

```
ORDERNUMBER: integer
QUANTITYORDERED: integer
PRICEEACH: double
ORDERLINENUMBER: integer
SALES: double
ORDERDATE: timestamp
STATUS: string
QTR_ID: integer
MONTH_ID: integer
YEAR_ID: integer
PRODUCTLINE: string
MSRP: double
PRODUCTCODE: string
CUSTOMERNAME: string
PHONE: string
ADDRESSLINE1: string
ADDRESSLINE2: string
CITY: string
STATE: string
POSTALCODE: string
COUNTRY: string
TERRITORY: string
```

Table ▼ + 🔍 🔧 📄

|    | 1 <sup>2</sup> <sub>3</sub> ORDERNUMBER | 1 <sup>2</sup> <sub>3</sub> QUANTITYORDERED | 1.2 PRICEEACH | 1 <sup>2</sup> <sub>3</sub> ORDERLINENUMBER | 1.2 SALES |
|----|---|---|---------------|---|-----------|
| 1  | 10107                                   | 30  | 95.7          | 2   | 2871      |
| 2  | 10121                                   | 34  | 81.35         | 5   | 2765.9    |
| 3  | 10134                                   | 41  | 94.74         | 2   | 3884.34   |
| 4  | 10145                                   | 45  | 83.26         | 6   | 3746.7    |
| 5  | 10159                                   | 49  | 100           | 14  | 5205.27   |
| 6  | 10168                                   | 36  | 96.66         | 1   | 3479.76   |
| 7  | 10180                                   | 29  | 86.13         | 9   | 2497.77   |
| 8  | 10188                                   | 48  | 100           | 1   | 5512.32   |
| 9  | 10201                                   | 22  | 98.57         | 2   | 2168.54   |
| 10 | 10211                                   | 41  | 100           | 14  | 4708.44   |
| 11 | 10223                                   | 37  | 100           | 1   | 3965.66   |
| 12 | 10237                                   | 23  | 100           | 7   | 2333.12   |
| 13 | 10251                                   | 28  | 100           | 2   | 3188.64   |
| 14 | 10263                                   | 34  | 100           | 2   | 3676.76   |

⬅
⬆
8,870+ rows | Truncated data due to byte limit | 16.64 seconds runtime
Refreshed 13 minutes ago

```
▶ 07:57 AM (3s) 5 Python
```

```
# Define storage account and container details for ADLS
adls_storage_account_name = "azurenewaccount"
adls_container_name = "sqltodalake"
output_path = f"abfss://{adls_container_name}@{adls_storage_account_name}.dfs.core.windows.net/transformed_data"

spark.conf.set(
    f"fs.azure.account.key.{adls_storage_account_name}.dfs.core.windows.net",
    "XexJU0ApV6eLGoc/shWgIL341gskEG09mVVi50A0U6niJqdeNEh10AMyKAjovtrclt7vo2huX0qx+AS6Z66Dg=="
)

newdf.write.mode('overwrite').option("header", "true").csv(output_path)

print(f"Transformed data written to {output_path}")
```

▶ (1) Spark Jobs

Transformed data written to `abfss://sqltodalake@azurenewaccount.dfs.core.windows.net/transformed_data`

Microsoft Azure | Data Factory | azure-datafactory-demo | Search factory and documentation | vineshkumar7901@outlook.com

Factory Resources

- Pipelines (4)
  - datafactorypipeline
  - dalalake
  - DatalaketoSQLtask
  - Incrementalpipeline
- Change Data Capture (preview) (0)
- Datasets (8)
- Data flows (0)
- Power Query (0)

DatalaketoSQLtask

Copy data1 → Notebook1 → Copy data2

Parameters Variables Settings Output

+ New

DatalaketoSQLtask

Copy data1 → Notebook1

Source dataset: datalaketasksql

Use query: Table

Query timeout (minutes): 120

Isolation level: Serializable

DatalaketoSQLtask

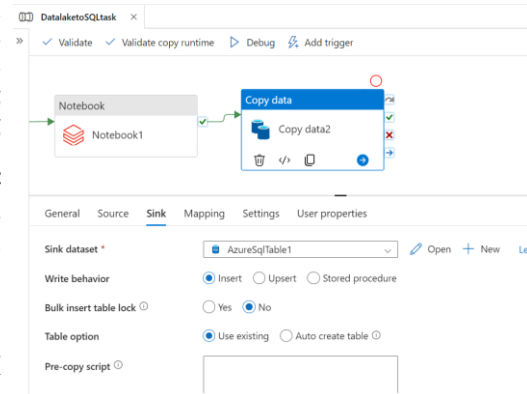
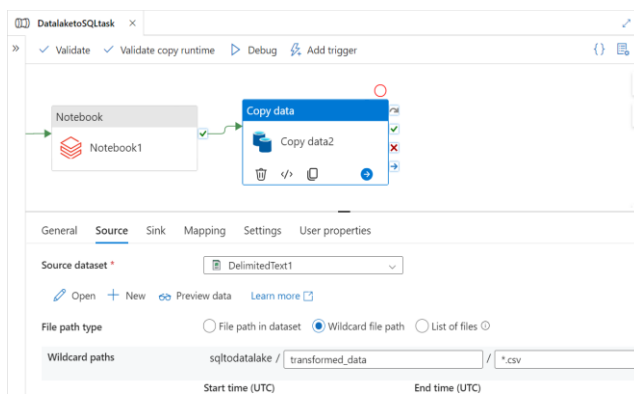
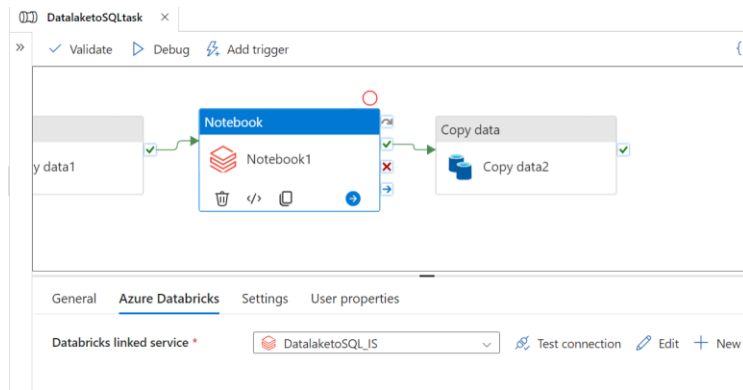
Copy data1 → Notebook1

Sink dataset: datalakeoutput1

Copy behavior: Select...

Max concurrent connections:

Block size (MB):



All pipeline runs > ✔ DatalaketoSQLtask - Activity runs

Refresh Update pipeline List Gantt

i This is a recent debug run. The local pipeline configuration is shown.

```
graph LR; Copydata1[Copy data1] --> Notebook1[Notebook1]; Notebook1 --> Copydata2[Copy data2];
```

### Activity runs

Pipeline run ID 08cce629-74bd-4ed8-9aab-2af191e79fc4


All status Monitor in Azure Metrics Export to CSV

Showing 1 - 3 of 3 items


| Activity name | Activity status                                | Activity type | Run start             | Duration | Integration runtime   |
|---------------|--|---------------|-----------------------|----------|-----------------------|
| Copy data2    | <span style="color: green;">✔</span> Succeeded | Copy data     | 7/29/2024, 6:00:23 PM | 15s      | AutoResolveIntegrator |
| Notebook1     | <span style="color: green;">✔</span> Succeeded | Notebook      | 7/29/2024, 6:00:00 PM | 22s      | AutoResolveIntegrator |
| Copy data1    | <span style="color: green;">✔</span> Succeeded | Copy data     | 7/29/2024, 5:59:47 PM | 13s      | AutoResolveIntegrator |

Details  Refresh

Activity run id: 9eec894e-3dd5-4401-9ebb-0a344cf4dc55

**Azure SQL Database**  
Region: East US 2

Succeeded →

**Azure Data Lake Storage Gen2**  
Region: Central US

Data read: ⓘ

Rows read: ⓘ

Peak connections: ⓘ

0 byte

0

1

Data written: ⓘ

Files written: ⓘ

Rows written: ⓘ

Peak connections: ⓘ

258 bytes

1


0


1

Copy duration 00:00:10  
Throughput: ⓘ 0 byte/s

▼ Azure SQL Database → Azure Data Lake Storage Gen2

|                        |                       |                |          |
|------------------------|-----------------------|----------------|----------|
| Start time             | 7/29/2024, 5:59:48 PM |                |          |
| Used DIUs ⓘ            | 4                     |                |          |
| Used parallel copies ⓘ | 1                     |                |          |
| ▼ Duration             | 00:00:10              |                |          |
| Details                | Working duration      | Total duration |          |
| ● Queue ⓘ              |                       | 00:00:05       |          |
| ● Transfer ⓘ           | Time to first byte ⓘ  | 00:00:01       | 00:00:02 |
|                        | Reading from source ⓘ | 00:00:00       |          |
|                        | Writing to sink ⓘ     | 00:00:00       |          |

 How satisfied or dissatisfied are you with the performance of this copy activity?  
★★★★★

Details  Refresh

Monitor real-time execution in Azure Databricks

Duration

00:00:17

Run page URL

<https://adb-2853020664689476.16.azuredatabricks.net/?o=2853020664689476#job/138226354201464/run/6472089>

d8-9aab-Zaf191e79fc4

Details  Refresh

**Azure Data Lake Storage Gen2**  
Region: Central US

Succeeded →

**Azure SQL Database**  
Region: East US 2

Data read: ⓘ

Files read: ⓘ

Rows read: ⓘ

Peak connections: ⓘ

42,768 KB

1

216

1

Data written: ⓘ

Rows written: ⓘ

Peak connections: ⓘ

62.84 KB

216

2

Copy duration 00:00:12  
Throughput: ⓘ 14,256 KB/s

▼ Azure Data Lake Storage Gen2 → Azure SQL Database

|                        |                       |                |          |
|------------------------|-----------------------|----------------|----------|
| Start time             | 7/29/2024, 6:00:24 PM |                |          |
| Used DIUs ⓘ            | 4                     |                |          |
| Used parallel copies ⓘ | 1                     |                |          |
| ▼ Duration             | 00:00:12              |                |          |
| Details                | Working duration      | Total duration |          |
| ● Queue ⓘ              |                       | 00:00:06       |          |
| ● Transfer ⓘ           | Listing source ⓘ      | 00:00:00       | 00:00:03 |
|                        | Reading from source ⓘ | 00:00:00       |          |
|                        | Writing to sink ⓘ     | 00:00:00       |          |

Data consistency verification ⓘ Not verified

## sqltodalake

Container

 x «[↑ Upload](#) [+ Add Directory](#) [↻ Refresh](#) | [🔄 Rename](#) [🗑 Delete](#) [↔ Change tier](#) [🔒 Acquire lease](#) [🔓 Break lease](#) [🗨 Give feedback](#)

### Overview

[🔧 Diagnose and solve problems](#)[👤 Access Control \(IAM\)](#)[⌵ Settings](#)[🔑 Shared access tokens](#)[👤 Manage ACL](#)[🔑 Access policy](#)[📄 Properties](#)[📘 Metadata](#)**Authentication method:** Access key ([Switch to Microsoft Entra user account](#))**Location:** sqltodalake ☒ Show deleted objects

|                          | Name               | Modified              | Access tier    | Archive status | Blob type  |
|--------------------------|--------------------|-----------------------|----------------|----------------|------------|
| <input type="checkbox"/> | 📁 transformed_data |                       |                |                |            |
| <input type="checkbox"/> | 📄 dbo.Orders.txt   | 7/27/2024, 2:36:52 PM | Hot (Inferred) |                | Block blob |

## sqltodalake

Container

 x «[↑ Upload](#) [+ Add Directory](#) [↻ Refresh](#) | [🔄 Rename](#) [🗑 Delete](#) [↔ Change tier](#) [🔒 Acquire lease](#) [🔓 Break lease](#) [🗨 Give feedback](#)

### Overview

[🔧 Diagnose and solve problems](#)[👤 Access Control \(IAM\)](#)[⌵ Settings](#)[🔑 Shared access tokens](#)[👤 Manage ACL](#) [★](#)[🔑 Access policy](#)[📄 Properties](#)[📘 Metadata](#)**Authentication method:** Access key ([Switch to Microsoft Entra user account](#))**Location:** [sqltodalake](#) / transformed\_data ☒ Show deleted objects

|                          | Name   | Modified              | Access tier    | Archive status | Blob type  |
|--------------------------|--|-----------------------|----------------|----------------|------------|
| <input type="checkbox"/> | 📁 [.]  |                       |                |                |            |
| <input type="checkbox"/> | 📄 _committed_7936958435667088038                   | 7/29/2024, 7:57:32 AM | Hot (Inferred) |                | Block blob |
| <input type="checkbox"/> | 📄 _started_7936958435667088038                     | 7/29/2024, 7:57:31 AM | Hot (Inferred) |                | Block blob |
| <input type="checkbox"/> | 📄 _SUCCESS   | 7/29/2024, 7:57:32 AM | Hot (Inferred) |                | Block blob |
| <input type="checkbox"/> | 📄 part-00000-tid-7936958435667088038-7246fcd8-e... | 7/29/2024, 7:57:32 AM | Hot (Inferred) |                | Block blob |