Write Query using Synapse Serverless SQL pool

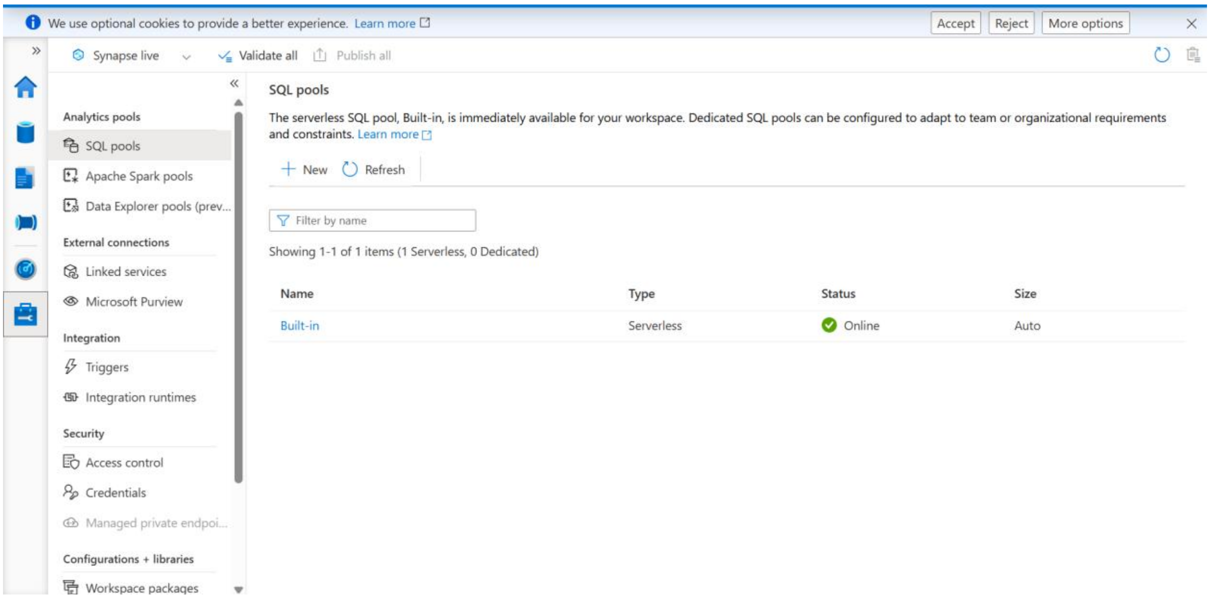
Name: Deepthi

Project overview

This project aims to optimize data ingestion and query performance on Azure Synapse Serverless SQL Pool by effectively handling various file formats (CSV,, Parquet, and Delta) from Azure Data Lake Storage Gen2. By implementing strategies like partitioning, indexing, and query tuning, we aim to improve data processing efficiency and reduce query execution times.

Prerequisites

Azure synapse workspace with **Serverless SQL pool**



We use optional cookies to provide a better experience. Learn more

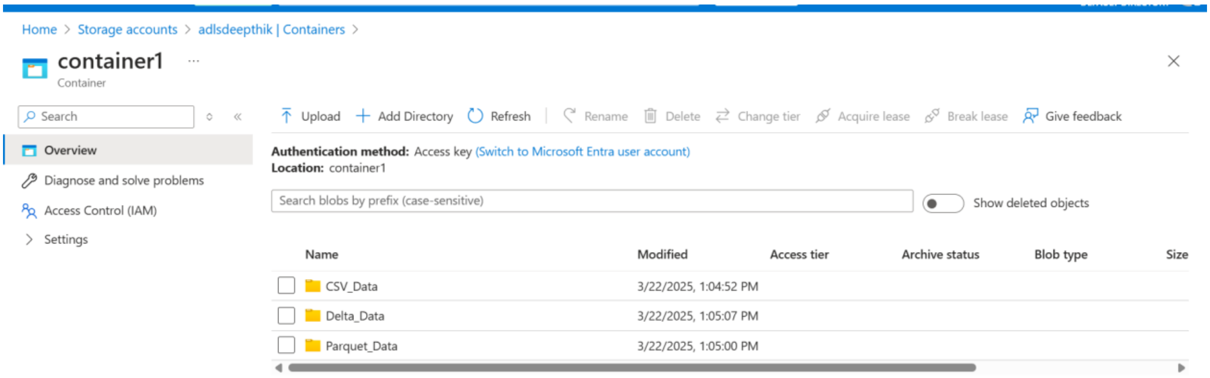
SQL pools

The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured to adapt to team or organizational requirements and constraints. Learn more

Showing 1-1 of 1 items (1 Serverless, 0 Dedicated)

Name	Type	Status	Size
Built-in	Serverless	Online	Auto

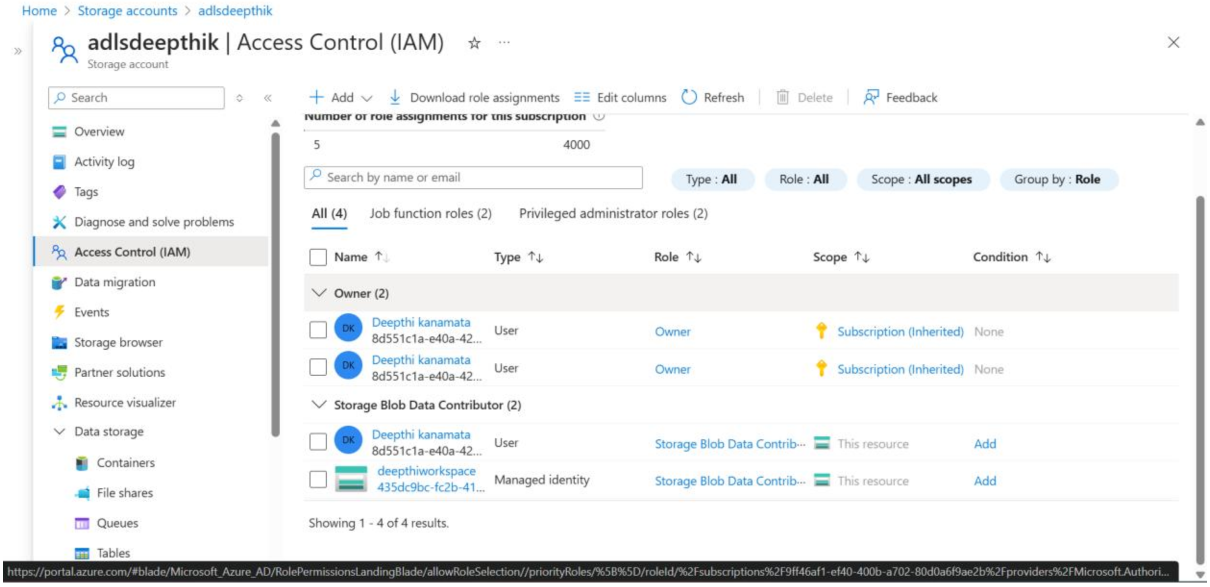
Azure Data Lake Storage Gen 2 account along with **CSV**, **Parquet** and **Delta** files



Authentication method: Access key (Switch to Microsoft Entra user account)
Location: container1

Name	Modified	Access tier	Archive status	Blob type	Size
CSV_Data	3/22/2025, 1:04:52 PM				
Delta_Data	3/22/2025, 1:05:07 PM				
Parquet_Data	3/22/2025, 1:05:00 PM				

Storage Access Permissions (Synapse needs the **Storage Blob Data Reader** role)



Number of role assignments for this subscription

5 4000

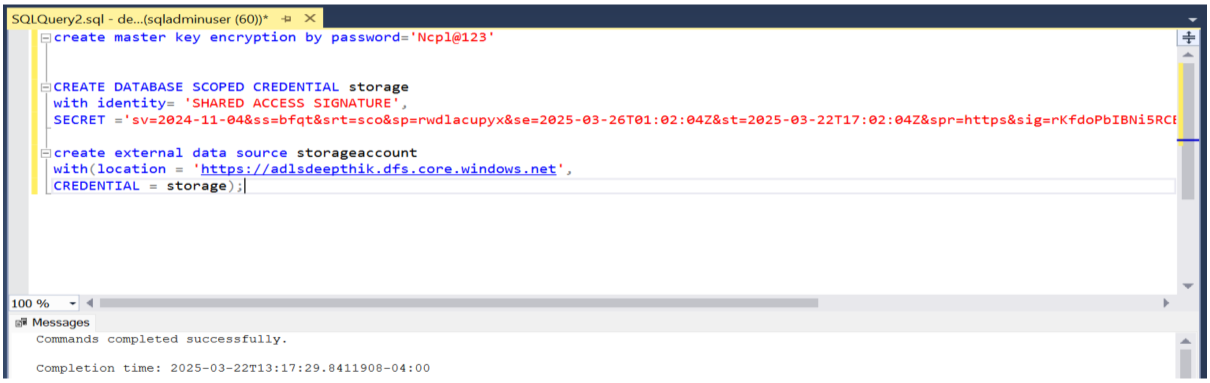
Name	Type	Role	Scope	Condition
Owner (2)				
Deepthi kanamata 8d551c1a-e40a-42...	User	Owner	Subscription (Inherited)	None
Deepthi kanamata 8d551c1a-e40a-42...	User	Owner	Subscription (Inherited)	None
Storage Blob Data Contributor (2)				
Deepthi kanamata 8d551c1a-e40a-42...	User	Storage Blob Data Contrib...	This resource	Add
deepthiworkspace 435dc9bc-fc2b-41...	Managed identity	Storage Blob Data Contrib...	This resource	Add

Showing 1 - 4 of 4 results.

https://portal.azure.com/#blade/Microsoft_Azure_AD/RolePermissionsLandingBlade/allowRoleSelection//priorityRoles/%5B%5D/roleId/%2Fsubscriptions%2F9ff46af1-ef40-400b-a702-80d0a6f9ae2b%2Fproviders%2FMicrosoft.Authori...

Implementing the Queries

1. Create an External Data Source



```
create master key encryption by password='Ncpl@123'


CREATE DATABASE SCOPED CREDENTIAL storage
with identity= 'SHARED ACCESS SIGNATURE',
SECRET ='sv=2024-11-04&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2025-03-26T01:02:04Z&st=2025-03-22T17:02:04Z&spr=https&sig=rKfdoPbIBNi5RCE

create external data source storageaccount
with(location = 'https://adlsdeepthik.dfs.core.windows.net',
CREDENTIAL = storage);
```

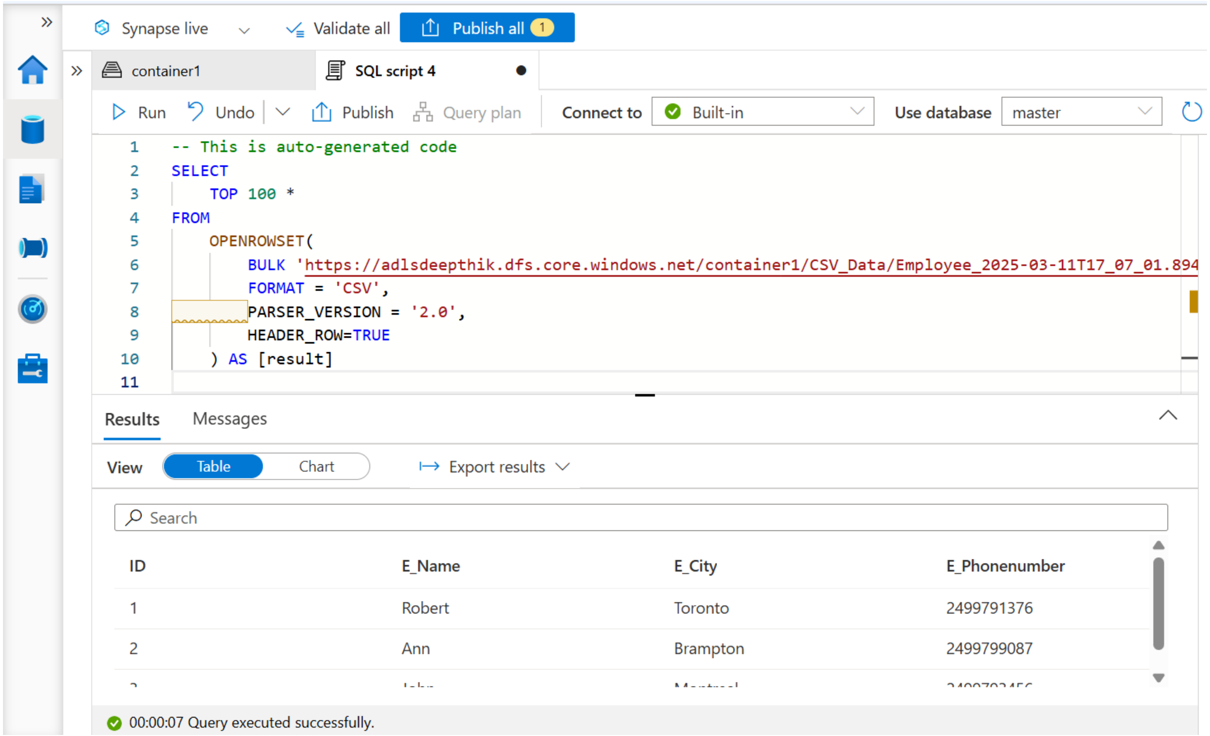
```
Commands completed successfully.

Completion time: 2025-03-22T13:17:29.8411908-04:00
```

Query a Specific File

External File format CSV

```
SELECT
   TOP 100 *
FROM
   OPENROWSET(
      BULK 'https://adlsdeepthik.dfs.core.windows.net/container1/CSV_Data/Employee_2025-03-11T17_07_01.8945923Z (1).csv',
      FORMAT = 'CSV',
      PARSER_VERSION = '2.0',
      HEADER_ROW=TRUE
   ) AS [result]
```



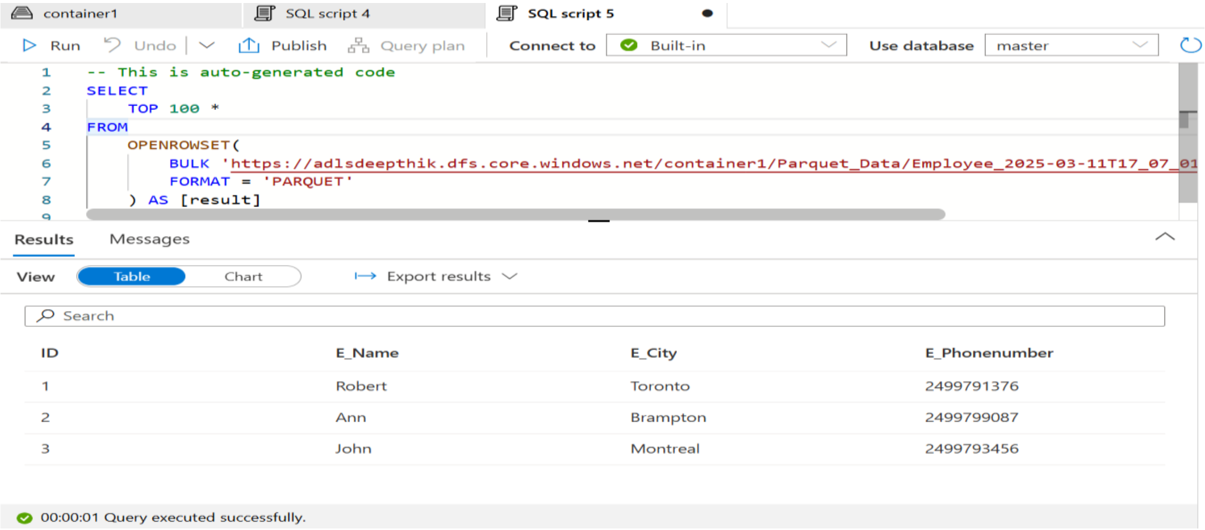
```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://adlsdeepthik.dfs.core.windows.net/container1/CSV_Data/Employee_2025-03-11T17_07_01.894
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        HEADER_ROW=TRUE
    ) AS [result]
```

ID	E_Name	E_City	E_Phonenumber
1	Robert	Toronto	2499791376
2	Ann	Brampton	2499799087
3	John	Montreal	2499703456

00:00:07 Query executed successfully.

External File format Parquet

```
SELECT
   TOP 100 *
FROM
   OPENROWSET(
      BULK 'https://adlsdeepthik.dfs.core.windows.net/container1/Parquet_Data/Employee_2025-03-11T17_07_01.8945923Z (1).parquet',
      FORMAT = 'PARQUET'
   ) AS [result]
```



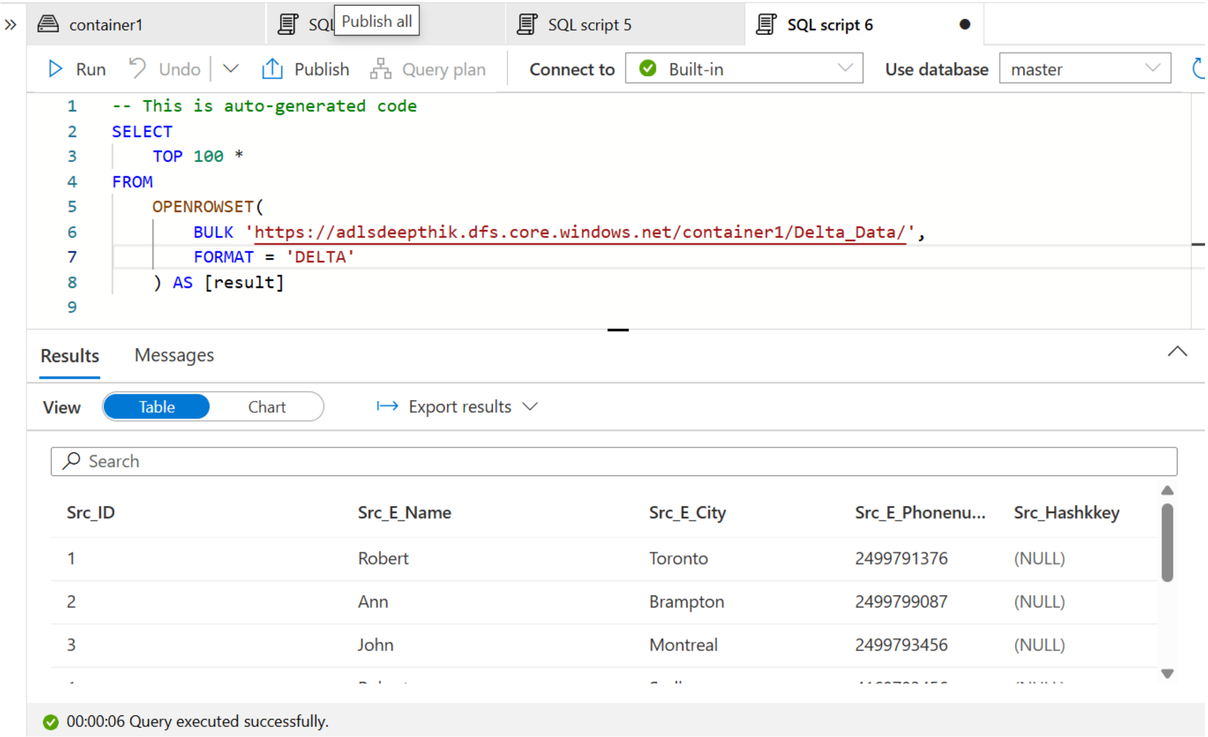
```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://adlsdeepthik.dfs.core.windows.net/container1/Parquet_Data/Employee_2025-03-11T17_07_01
        FORMAT = 'PARQUET'
    ) AS [result]
```

ID	E_Name	E_City	E_Phonenumber
1	Robert	Toronto	2499791376
2	Ann	Brampton	2499799087
3	John	Montreal	2499793456

00:00:01 Query executed successfully.

External File format Delta

```
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://adlsdeepthik.dfs.core.windows.net/container1/Delta_Data/',
        FORMAT = 'DELTA'
    ) AS [result]
```



```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://adlsdeepthik.dfs.core.windows.net/container1/Delta_Data/',
        FORMAT = 'DELTA'
    ) AS [result]
```

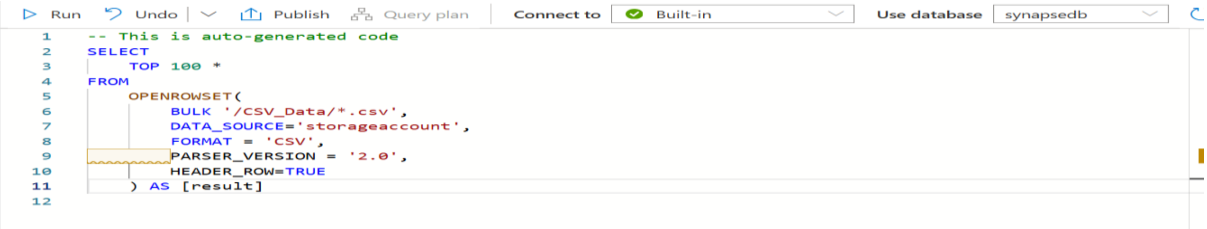
Src_ID	Src_E_Name	Src_E_City	Src_E_Phonenu...	Src_Hashkkey
1	Robert	Toronto	2499791376	(NULL)
2	Ann	Brampton	2499799087	(NULL)
3	John	Montreal	2499793456	(NULL)

00:00:06 Query executed successfully.

Querying Data with Wildcards

Wildcard for CSV file:

```
SELECT
   TOP 100 *
FROM
   OPENROWSET(
      BULK '/CSV_Data/*.csv',
      DATA_SOURCE='storageaccount',
      FORMAT = 'CSV',
      PARSER_VERSION = '2.0',
      HEADER_ROW=TRUE
   ) AS [result]
```

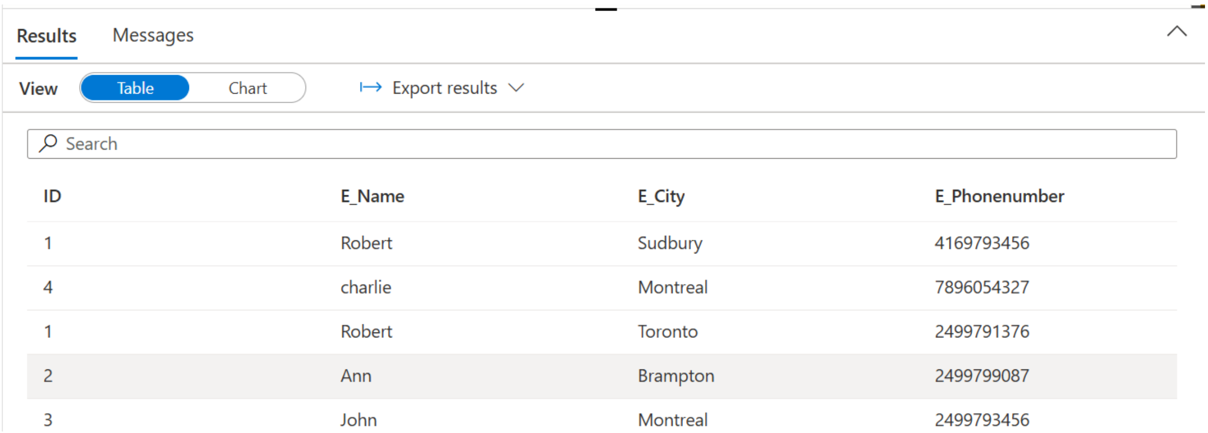


Connect to Built-in

Use database synapsedb

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK '/CSV_Data/*.csv',
        DATA_SOURCE='storageaccount',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        HEADER_ROW=TRUE
    ) AS [result]
```

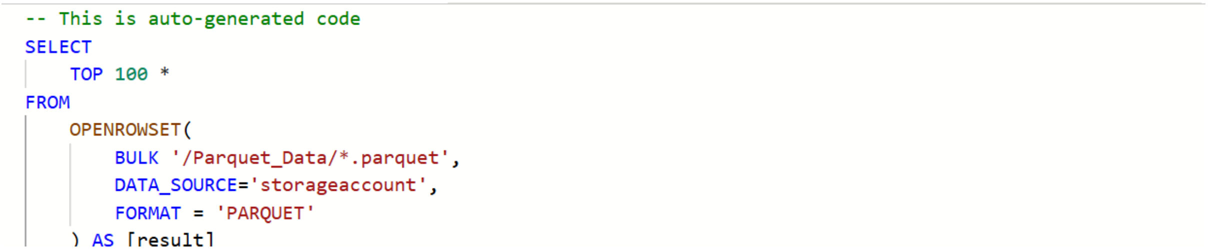
Result of all the files in that folder



ID	E_Name	E_City	E_Phonenumber
1	Robert	Sudbury	4169793456
4	charlie	Montreal	7896054327
1	Robert	Toronto	2499791376
2	Ann	Brampton	2499799087
3	John	Montreal	2499793456

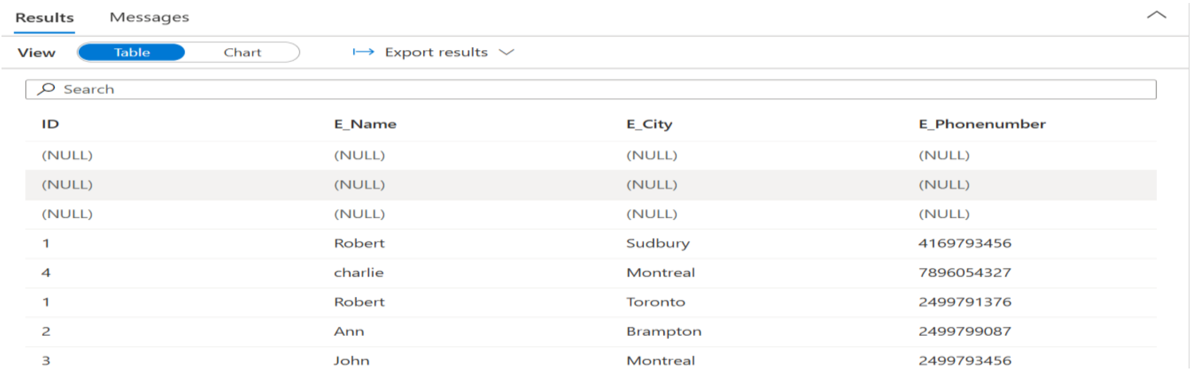
Wildcard for Parquet file:

```
SELECT
   TOP 100 *
FROM
   OPENROWSET(
      BULK '/Parquet_Data/*.parquet',
      DATA_SOURCE='storageaccount',
      FORMAT = 'PARQUET'
   ) AS [result]
```



```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK '/Parquet_Data/*.parquet',
        DATA_SOURCE='storageaccount',
        FORMAT = 'PARQUET'
    ) AS [result]
```

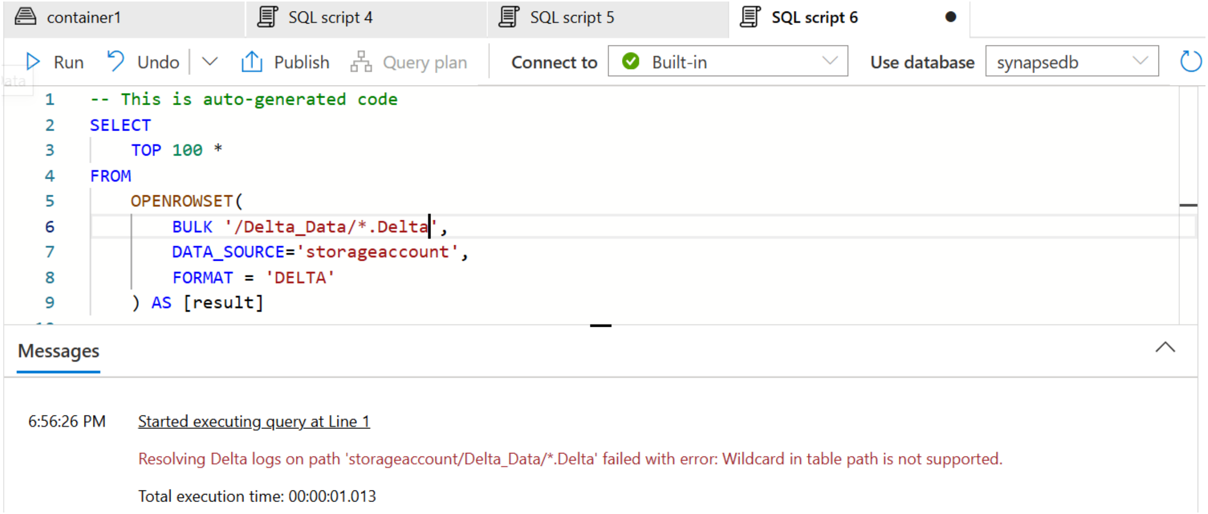
Results of all files in Parquet_Data folder



ID	E_Name	E_City	E_Phonenumber
(NULL)	(NULL)	(NULL)	(NULL)
(NULL)	(NULL)	(NULL)	(NULL)
(NULL)	(NULL)	(NULL)	(NULL)
1	Robert	Sudbury	4169793456
4	charlie	Montreal	7896054327
1	Robert	Toronto	2499791376
2	Ann	Brampton	2499799087
3	John	Montreal	2499793456

Wildcard for Delta file:

We cannot do a wild card on delta folder as this folder contains Delta logs



```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK '/Delta_Data/*.Delta',
        DATA_SOURCE='storageaccount',
        FORMAT = 'DELTA'
    ) AS [result]
```

```
6:56:26 PM    Started executing query at Line 1
              Resolving Delta logs on path 'storageaccount/Delta_Data/*.Delta' failed with error: Wildcard in table path is not supported.
              Total execution time: 00:00:01.013
```

This is because in delta folder files are stored in .parquet format

Performance Optimization

- Use partitioning (e.g., folders by date, region) to reduce scanned data.
- Query only required columns instead of SELECT * to lower costs.
- Prefer Parquet over CSV for better compression & faster queries.

Error Handling

- Check file existence with sys.external_files before querying.
- Use TRY_CAST() to handle data type mismatches & avoid query failures.
- Enable ERRORFILE to log corrupt data instead of failing queries.

Data Cleaning

- Remove **duplicates** using DISTINCT or window functions.
- Replace **NULL values** with defaults using COALESCE().
- Standardize **date formats** using TRY_CAST(OrderDate AS DATE).