Project Documentation: Synapse Warehouse

Transferring Processed Data to Synapse Warehouse for Analysis

Overview

This project involves transferring the processed.csv file from an Azure Blob Storage container (processed-data) to a Synapse Analytics warehouse for further data analysis. The entire process leverages **Azure Synapse Analytics**, **Azure Blob Storage**, and **Azure Key Vault** for secure management of connection strings and credentials.

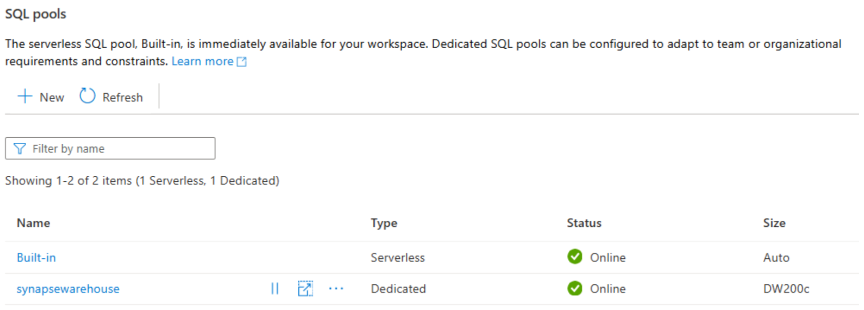
Components Used

1. **Azure Synapse Analytics** - Used to manage the dedicated SQL pool for data storage and analysis.
2. **Azure Blob Storage** - The source storage account (container: processed-data) where the data file is stored.
3. **Azure Key Vault** - Secures and stores sensitive connection information like account keys and secrets for the linked services.
4. **SQL Dedicated Pool** - A provisioned data warehouse in Synapse for analysis.

Steps Performed

1. Set Up Azure Synapse Analytics

- **Dedicated SQL Pool Creation**: A dedicated SQL pool (data warehouse) was created in Synapse Analytics for storing and analyzing the processed data.
  - **SQL Pool Name**: synapsewarehouse



SQL pools

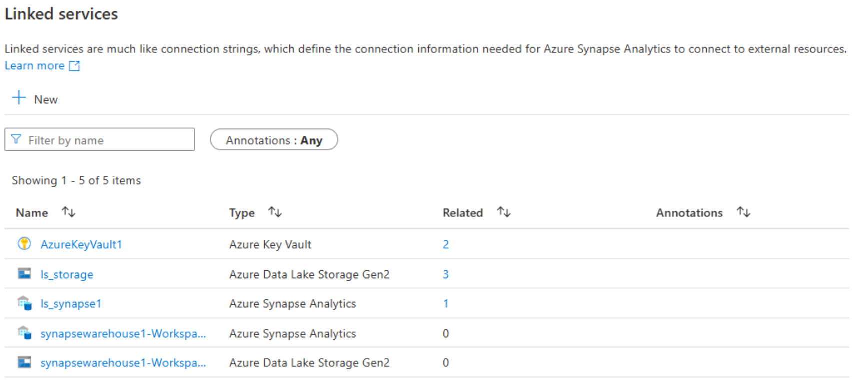
The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured to adapt to team or organizational requirements and constraints. Learn more

Showing 1-2 of 2 items (1 Serverless, 1 Dedicated)

Name	Type	Status	Size
Built-in	Serverless	✓ Online	Auto
synapsewarehouse	Dedicated	✓ Online	DW200c

2. Create Linked Services Using Azure Key Vault

- **Storage Account Linked Service**: Created a linked service for connecting to the Azure Blob Storage account using Azure Key Vault to securely manage connection strings.
  - **Linked Service Name**: ls_storage
  - **Key Vault Used**: AzureKeyVault1 (For managing storage account secrets and keys)
- **Synapse Analytics Linked Service**: Created a linked service for Synapse Analytics to connect to the data warehouse using Azure Key Vault.
  - **Linked Service Name**: ls_synapse1
  - **Key Vault Used**: AzureKeyVault1 (For securely managing SQL pool credentials)



Linked services

Linked services are much like connection strings, which define the connection information needed for Azure Synapse Analytics to connect to external resources.
Learn more

+ New

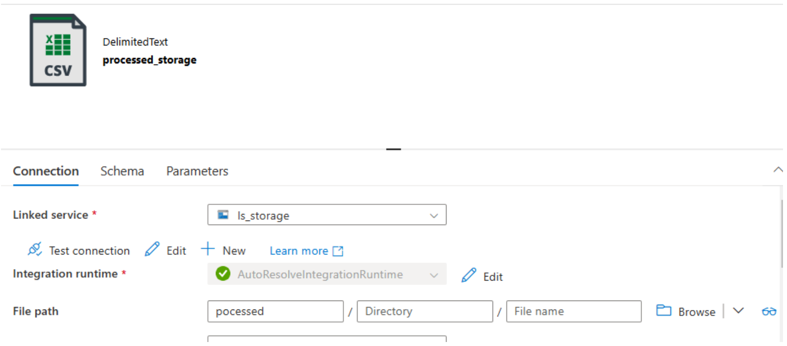
Filter by name   Annotations : **Any**

Showing 1 - 5 of 5 items

Name	Type	Related	Annotations
AzureKeyVault1	Azure Key Vault	2	
ls_storage	Azure Data Lake Storage Gen2	3	
ls_synapse1	Azure Synapse Analytics	1	
synapsewarehouse1-Workspa...	Azure Synapse Analytics	0	
synapsewarehouse1-Workspa...	Azure Data Lake Storage Gen2	0	

3. Data Pipeline Creation in Synapse Analytics

- **Source (Blob Storage)**: The processed.csv file from the processed-data container in Blob Storage is configured as the source in the pipeline within Synapse Analytics.
  - **Source Dataset**: processed_storage
  - **Source Linked Service**: ls_storage



DelimitedText
processed_storage

Connection   Schema   Parameters

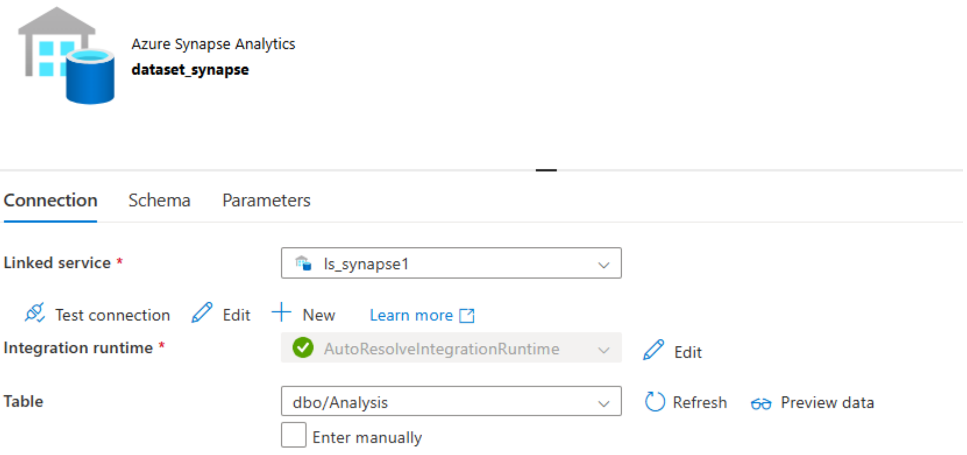
Linked service * : ls_storage

Test connection   Edit   + New   Learn more

Integration runtime * : AutoResolveIntegrationRuntime   Edit

File path: pocessed / Directory / File name   Browse

- **Sink (Synapse Analytics)**: Data from the source is transferred to the dedicated SQL pool in Synapse Analytics.
    - **Sink Dataset**: dataset_synapse
    - **Sink Linked Service**: ls_synapse1



Azure Synapse Analytics
dataset_synapse

Linked service * ls_synapse1

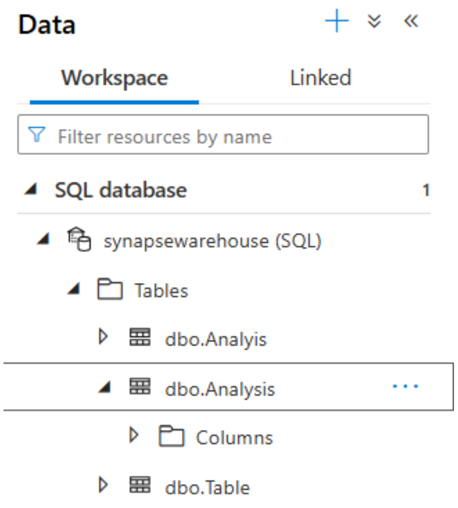
Integration runtime * AutoResolveIntegrationRuntime

Table dbo/Analysis

□ Enter manually

4. Create Analysis Table in Synapse

- An analysis table was created in the dedicated SQL pool to store the data transferred from the processed.csv file. The table structure is designed to match the schema of the CSV data.
    - **Table Name**: Analysis
    - **SQL Table Schema**: dbo



SQL database 1
- synapsewarehouse (SQL)
  - Tables
    - dbo.Analyis
    - dbo.Analysis
      - Columns
    - dbo.Table

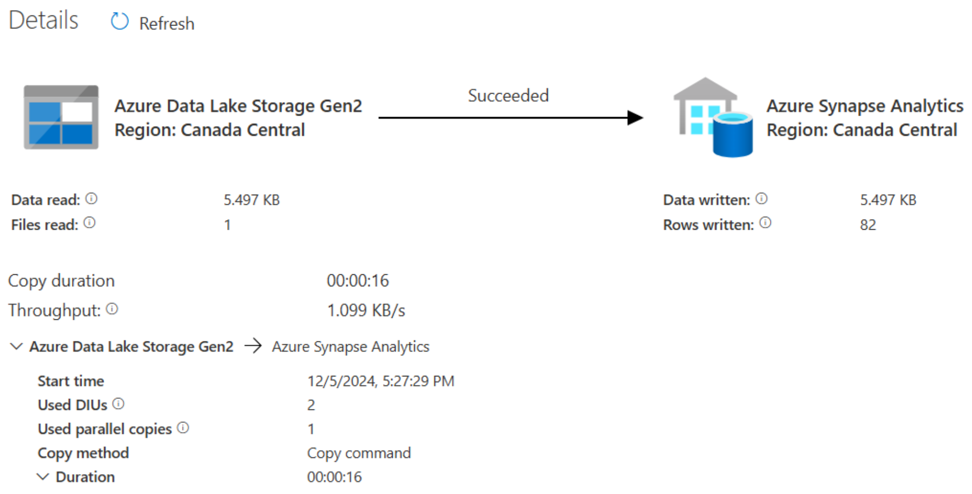
5. Debugging and Testing

- After configuring the pipeline, it was debugged using **Azure Synapse Studio**'s debugging tools to ensure that the data flow works as expected.
  - **Debugging**: Ensured that the file transfer from Blob Storage to Synapse SQL pool was successful without any issues.

- **Manual Trigger**: The pipeline was manually triggered to ensure the entire process from source to sink worked seamlessly.
  - **Trigger Type**: Manual
  - **Status**: Success

6. Publish the Pipeline

- Once debugging was successful, the pipeline was published for production use.
  - **Pipeline Name**: DataTransfer_Pipeline
  - **Status**: Published and ready for automation or future manual triggers



Azure Data Lake Storage Gen2
Region: Canada Central

Succeeded

Azure Synapse Analytics
Region: Canada Central

Data read:	5.497 KB
Files read:	1
Data written:	5.497 KB
Rows written:	82
Copy duration	00:00:16
Throughput:	1.099 KB/s

Azure Data Lake Storage Gen2 → Azure Synapse Analytics

Start time	12/5/2024, 5:27:29 PM
Used DIUs	2
Used parallel copies	1
Copy method	Copy command
Duration	00:00:16

Data Flow Summary

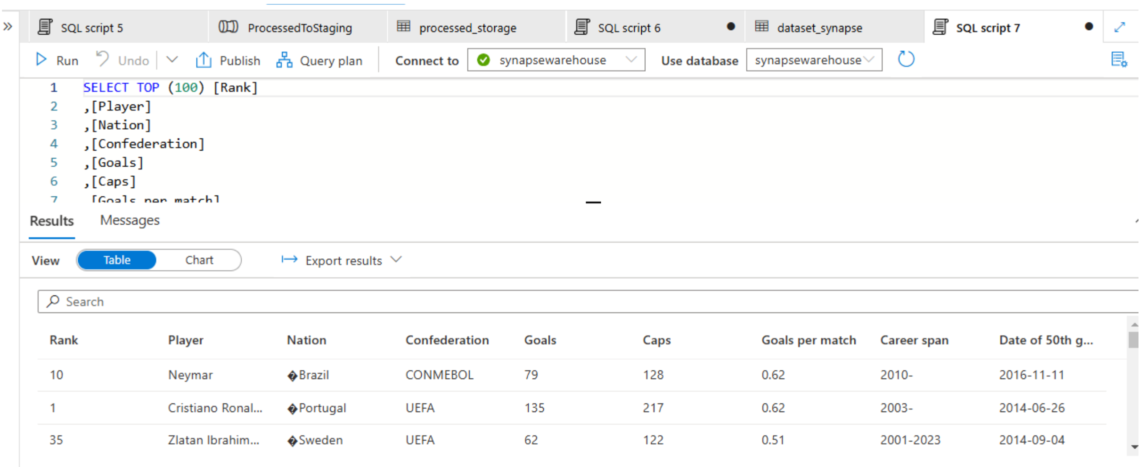
- **Source**: processed.csv file in the processed-data container in Azure Blob Storage
- **Destination**: Analysis table in Synapse Analytics dedicated SQL pool

- **Pipeline Orchestration**: Managed directly within **Azure Synapse Analytics** using the built-in pipeline features.
- **Security**: Managed securely using **Azure Key Vault** for credentials and secrets

---

Validation and Testing

- **Data Validation**: After the manual trigger, the data was validated in the dedicated SQL pool by querying the AnalysisTable to ensure the transferred data was correct.
- **Data Verification**: Verified that the data from the processed.csv file was transferred correctly into the SQL pool without loss or alteration.



Connect to: synapsewarehouse

Use database: synapsewarehouse

View: Table

```
SELECT TOP (100) [Rank]
,[Player]
,[Nation]
,[Confederation]
,[Goals]
,[Caps]
 [Goals per match]
```

Rank	Player	Nation	Confederation	Goals	Caps	Goals per match	Career span	Date of 50th g...
10	Neymar	�Brazil	CONMEBOL	79	128	0.62	2010-	2016-11-11
1	Cristiano Ronal...	�Portugal	UEFA	135	217	0.62	2003-	2014-06-26
35	Zlatan Ibrahim...	�Sweden	UEFA	62	122	0.51	2001-2023	2014-09-04

Challenges and Resolutions

- **Challenge**: Ensuring secure handling of credentials and connection information.
  - **Resolution**: Used **Azure Key Vault** to securely store and retrieve sensitive information for linked services.
- **Challenge**: Debugging the pipeline before publishing to ensure successful data transfer.
  - **Resolution**: Leveraged the debugging feature in **Azure Synapse Studio** and manually triggered the pipeline to validate the process.

Future Enhancements

- **Automation**: Set up triggers for automatic data transfers at scheduled intervals for future data ingestion.

- **Error Handling**: Implement enhanced error handling and logging within the pipeline for monitoring and alerting.

Conclusion

The transfer of processed.csv from Blob Storage to Synapse Analytics was successfully implemented using **Azure Synapse Analytics**, **Azure Blob Storage**, **Azure Key Vault**, and the native **Synapse pipeline** orchestration. This setup allows for scalable data analysis and can be further automated and enhanced for future use cases.