Data Engineering in the Cloud

Hybrid Transactional Analytical Processing

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- In the lab, we Implement Configuration of Synapse Link with Azure Cosmos DB

Prererquisites

- Prerequisites (5 minutes):
  - ▸ Create a datalake (storage account with hierarchical namespace enabled)
    - ⋆ To lower the cost, you may choose Redundancy as LRS
  - ▸ Create a container in the storage account
  - ▸ Create an Azure Synapse workspace and
    - ⋆ create a Apache spark pool in the Manage tab

Lab

- ① Go to create "resource tab" and search for "Azure Cosmos DB".

Azure Cosmos DB

Lab

- ② Click on "Create" for `Azure Cosmos DB for NoSQL`.

Lab

- ③ Then, enter the configuration details, account name (eg. `esucosmos`), location, and resource group.

Lab

- ④ Add other configuration details as required. Then, go to "Review + create" and validate the configuration and click on "Create".

Lab

- ⑤ Go to the `Data Explorer`, create a new container called `container2`

Lab

- The settings of container2 are as the following

Lab

- ⑥ Now, to configure the Azure Synapse Link with Azure Cosmos DB, we need to enable Azure Synapse Link. For this, go to the "Azure Cosmos DB account" and then search for `Azure Synapse Link`.

Lab

- 7. Click on "Azure Synapse Link" and then click on "Enable". It may take 1-2 minutes to be enabled.

Lab

- ⑧ Now go back to "Azure Cosmos DB account" home and click the "Data Explorer" tab to create a another container, called `container1`. However, enable the analytical store on the container.

Lab

- ⑨ Click on "New container" and enter the details. Make sure the `analytical store` is on. Click on `OK`.

Lab

- ⑩ Go to "Items" under the container

Lab
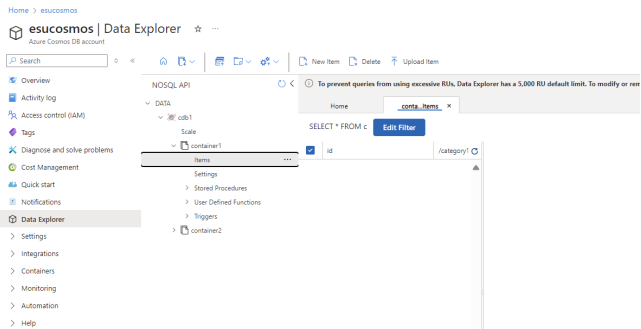
- ⓫ And click on "New Item" to add a new item. We create two new items and enter the random data as shown below and click on "Save".
  - ▸ Make sure the value of "id" is different in the two items. Write the code as given below in the script.

Lab

- ⑫ After you click on "Save", you will see

Lab

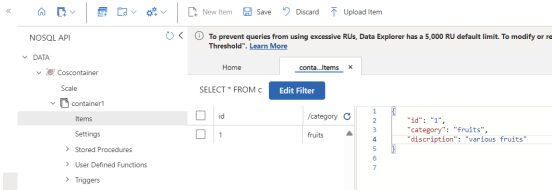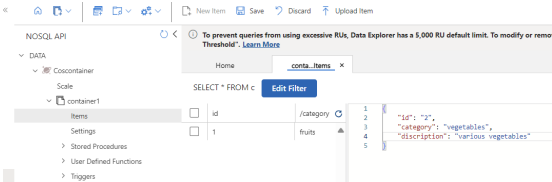- ⑬ Now, in the left pane, go to "Keys" under the `Settings` and note down the name of Azure Cosmos DB account and the primary key for later use.

Lab

- For example, mine is

URI
https://esucosmos.documents.azure.com:443/

PRIMARY KEY
Im1RN8JtTNleZOMNTWOjKEsanKaLwECeIoHmdC34KbcllywPy4GHn3wlo7RJp2vMjlw1eCxofOLYACDbKcNE

SECONDARY KEY
opO4T73DeWKGOzH3vmheSfaDISiP85VrMLR42FVqFUSDOQNZJ60tfwMvnHRMRwbEWv8bzUM8SDrEACDbOagE

PRIMARY CONNECTION STRING
AccountEndpoint=https://esucosmos.documents.azure.com:443/;AccountKey=Im1RN8JtTNleZO

SECONDARY CONNECTION STRING
AccountEndpoint=https://esucosmos.documents.azure.com:443/;AccountKey=opO4T73DeWKGOz

- Note: To create a linked service, there must be a synapse analytics workspace with an Apache spark pool. Create them if you do not have them yet.

Lab

- ⑭ Open "synapse studio" and go to the "Manage" section. Go to Linked services, and click on "+ New" to create a new linked service.

Lab

- ⑮ Select "Azure Cosmos DB for NoSQL" and click on "Continue".

Lab

- **16** Specify the details as shown below:

Lab

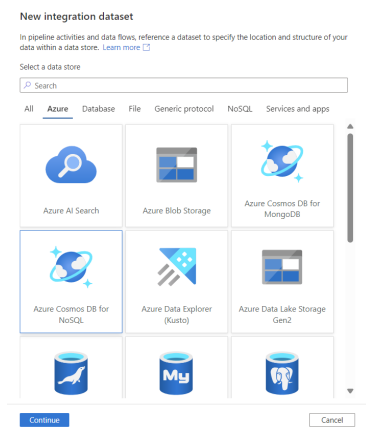- **17** Click on "Test connection" and click on "Create"

Lab

- ⓲ Now, go to the "Data" section of the synapse studio, click on "+" and choose "Integration dataset" to create a dataset.

Lab
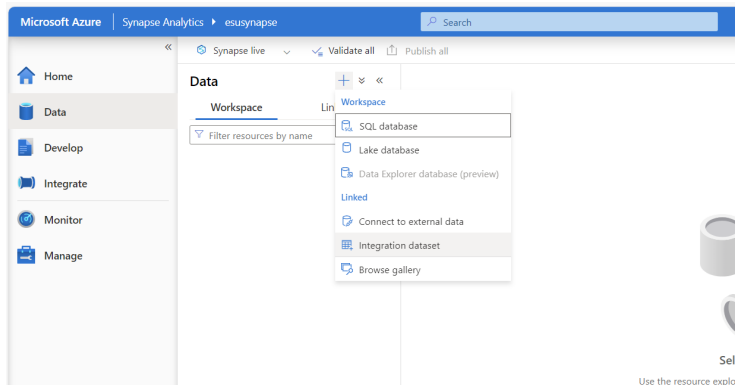
- ⑲ Select "Azure Cosmos DB for NoSQL" and click on "Continue".

Lab

- ⑳ Set the properties as shown and click on "OK".

Set properties

Name

CosmosDbNoSqlContainer1

Linked service *

CosmosDbNoSql1

Connect via integration runtime * ⓘ

✔ AutoResolveIntegrationRuntime

Container

container1

☐ Enter manually

Import schema

◉ From connection/store   ○ None

OK   Back   Cancel

Lab

- ㉑ In the below section, go to "Connections" and click on "Preview data" to view the data.

Lab

Lab

- ㉒ Now, click on "Publish All".
  - ► Then, next we create and run a pipeline.

Lab
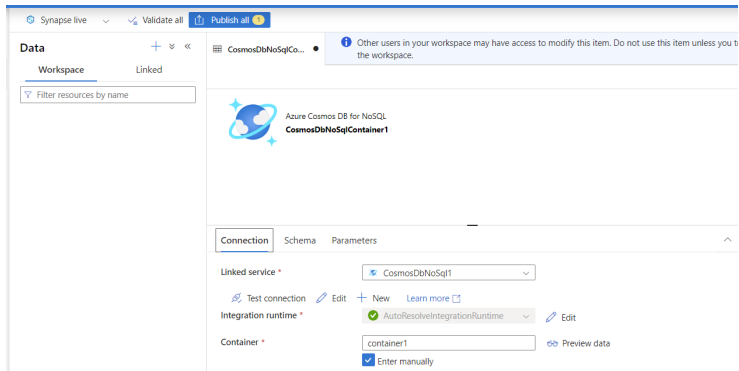
- ㉓ Go to the "Integrate" tab of the "synapse studio"
- ㉔ Click on "+" and choose "Pipeline".

Lab

- ㉕ In the left pane, under "Move & transform", drag and drop "Copy data" into the window.
- ㉖ Specify a name to the pipeline.

Lab

- ㉗ In the bottom pane, go to "Source" and select the "Source dataset" from the drop-down.

Lab

- 28 Go to "Sink" and click on "+ New".

Lab
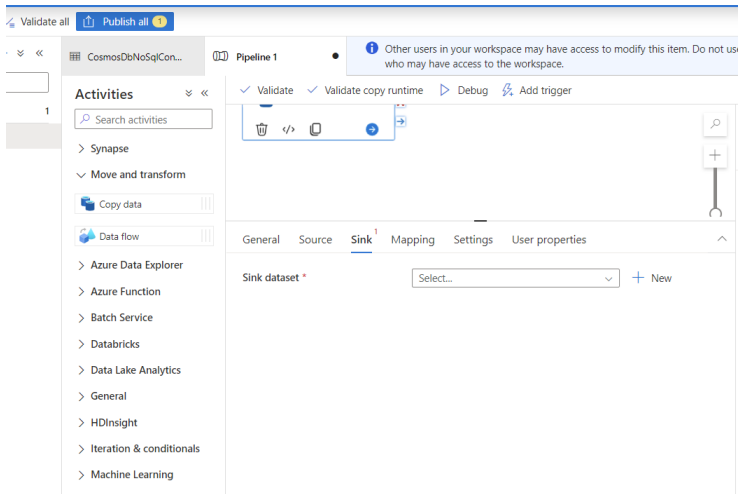
- **㉙** Select "Azure Cosmos DB for NoSQL" and click on "Continue".
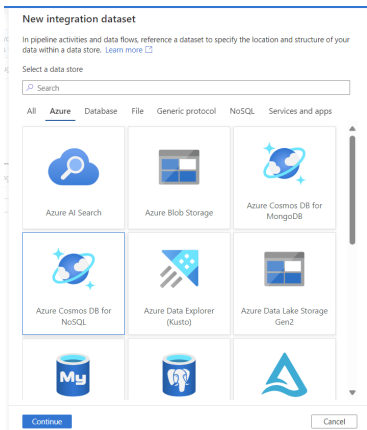
Lab

- **30** Specify the property details as shown below, and click on "OK".
  - ▶ We use the same container

Set properties

Name

CosmosDbNoSqlContainer2

Linked service *

CosmosDbNoSql1

Connect via integration runtime * ⓘ

✅ AutoResolveIntegrationRuntime

Container

container2

☐ Enter manually

Import schema

🔘 From connection/store   ◯ None

> Advanced

Lab

- **31** Select "Insert" in the below pane under `Sink`.

Lab

- **32** Click on "Publish All" to save the pipeline.
- **33** Once the publish is successful, your pipeline will be ready to run.

Lab

- **34** Check if the pipeline can be run successfully

Lab

- ㉟ After the pipeline is run, check `container2` in the Azure Cosmos DB database.

License