Data Engineering in the Cloud

Petabyte Scale Ingestion

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Prerequisites
- Petabyte Scale Ingestion

Prerequisites

- The topic is a continuation of the last lecture (we skip this lecture since Azure Data Factory is better for pipelines)
- Prerequisites (10 minutes):
  - Create a datalake (storage account with hierarchical namespace enabled)
    - To lower the cost, you may choose `Redundancy` as `LRS`
    - Upload the parquet file `parquet.parquet` to your container
  - Create a container
  - Upload the data set `new_customers.csv` to the container
  - Create a Synapse workspace and a dedicated SQL pool
  - Open the Synapse Studio
  - Run the SQL code in the last lecture

Prerequisites

```
CREATE SCHEMA [product_staging];
GO

CREATE TABLE [product_staging].[ProductHeapp]
(
    [Id] int,
    [Correlationid] nvarchar(4000),
    [Operationname] nvarchar(4000),
    [Status] nvarchar(4000),
    [Eventcategory] nvarchar(4000),
    [Level] nvarchar(4000),
    [Time] datetime2(7),
    [Subscription] nvarchar(4000),
    [Eventinitiatedby] nvarchar(4000),
    [Resourcetype] nvarchar(4000),
    [Resourcegroup] nvarchar(4000)
    )
WITH
    (
    DISTRIBUTION = ROUND_ROBIN,
    HEAP
    )
GO
```

Introduction to Petabyte Scale Ingestion

- Petabyte-scale ingestion refers to the process of transferring, processing, and storing extremely large volumes of data, typically in the range of petabytes (1 petabyte = 1,024 terabytes or about 1 million gigabytes).
- For example, Telecom companies collect huge volumes of call records, network data, and usage statistics that need to be ingested for network optimization and customer insights.
- We now Perform petabyte-scale ingestion with the Azure Synapse pipeline.

Introduction to Petabyte Scale Ingestion

- Creating workload groups and classifiers as shown is a way to manage and prioritize resources in your Azure Synapse Analytics.
  - These steps are not strictly required for petabyte-scale ingestion but can be highly beneficial for ensuring that your data ingestion process gets the necessary resources and operates efficiently.
- The workload group `DemoLoad` is created with specific resource allocation parameters
- The workload classifier HeavyLoader is associated with the DemoLoad workload group and specifies that:
  - Requests from the `dbo` user will be routed to the `DemoLoad` workload group.
  - These requests are given `IMPORTANCE = HIGH`, meaning they will be prioritized higher than other requests.

Introduction to Petabyte Scale Ingestion

- We first configure the classification of the workload management. To do so, go to the synapse studio and Develop tab.
  - Then Click on + and add new SQL script
  - Connect to the dedicated SQL pool
- Create a workload group that will use the workload isolation by reserving a minimum of 50% resources with a cap of 100% using the command below:

```
IF NOT EXISTS (SELECT * FROM sys.workload_management_workload_classifiers
               WHERE group_name = 'DemoLoad')

BEGIN
    CREATE WORKLOAD GROUP DemoLoad WITH
    (
        MIN_PERCENTAGE_RESOURCE = 50
        ,REQUEST_MIN_RESOURCE_GRANT_PERCENT = 25
        ,CAP_PERCENTAGE_RESOURCE = 100
    );
END
```

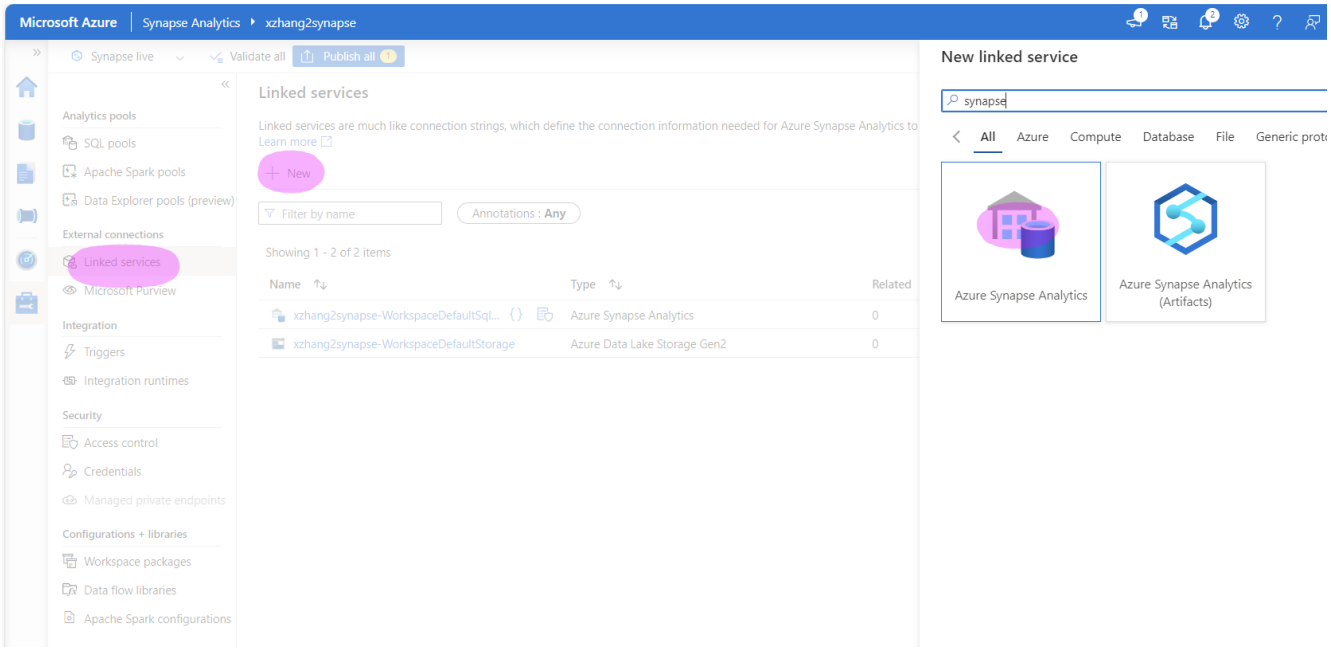
Introduction to Petabyte Scale Ingestion

- Now, create a new workload classifier that will be used to assign the user we created to the DemoLoad workload group

```
IF NOT EXISTS (SELECT * FROM sys.workload_management_workload_classifiers
               WHERE [name] = 'HeavyLoader')
BEGIN
    CREATE WORKLOAD Classifier HeavyLoader WITH
    (
        Workload_Group ='DemoLoad',
        MemberName='dbo',
        IMPORTANCE = HIGH
    );
END
```

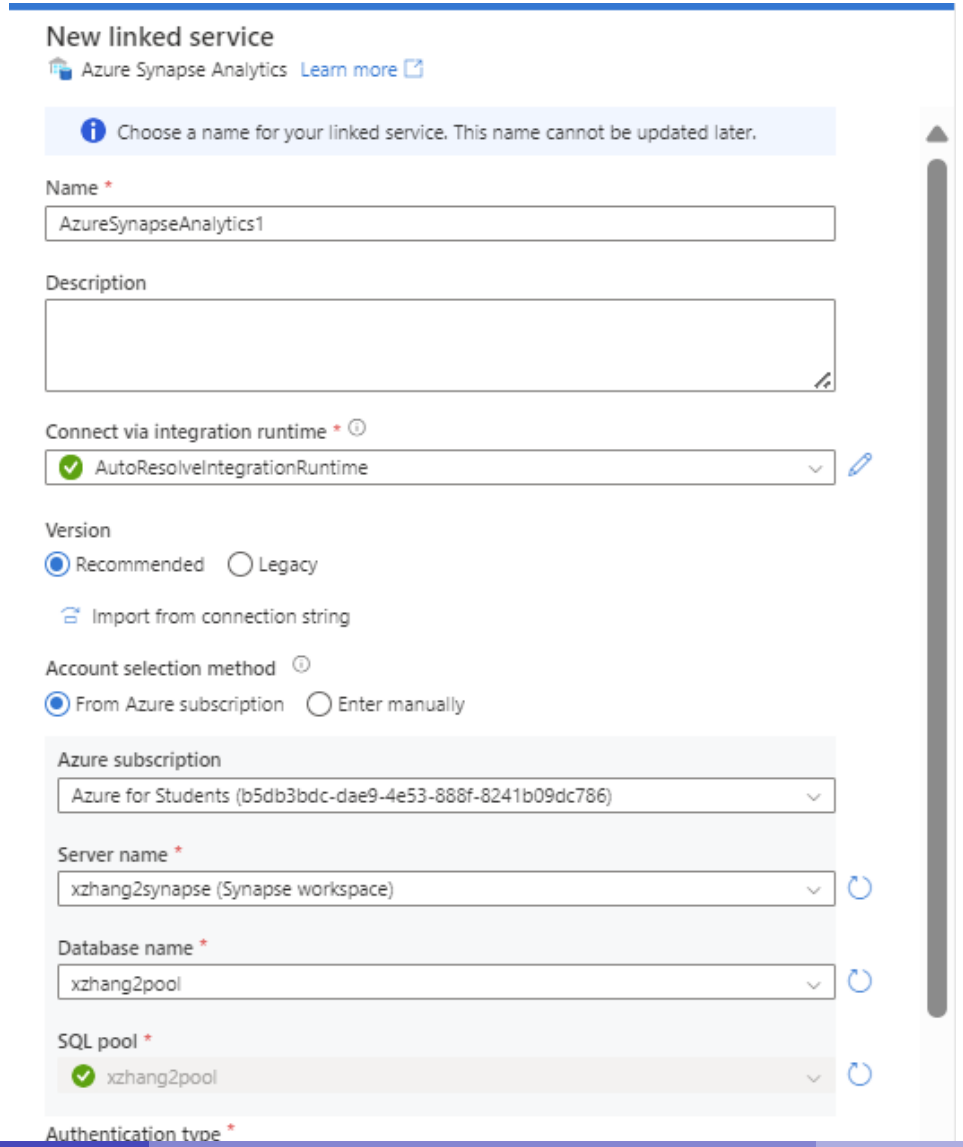
Introduction to Petabyte Scale Ingestion

- Go to the manage tab and if you don't see an auto generated linked service, click on `+New`



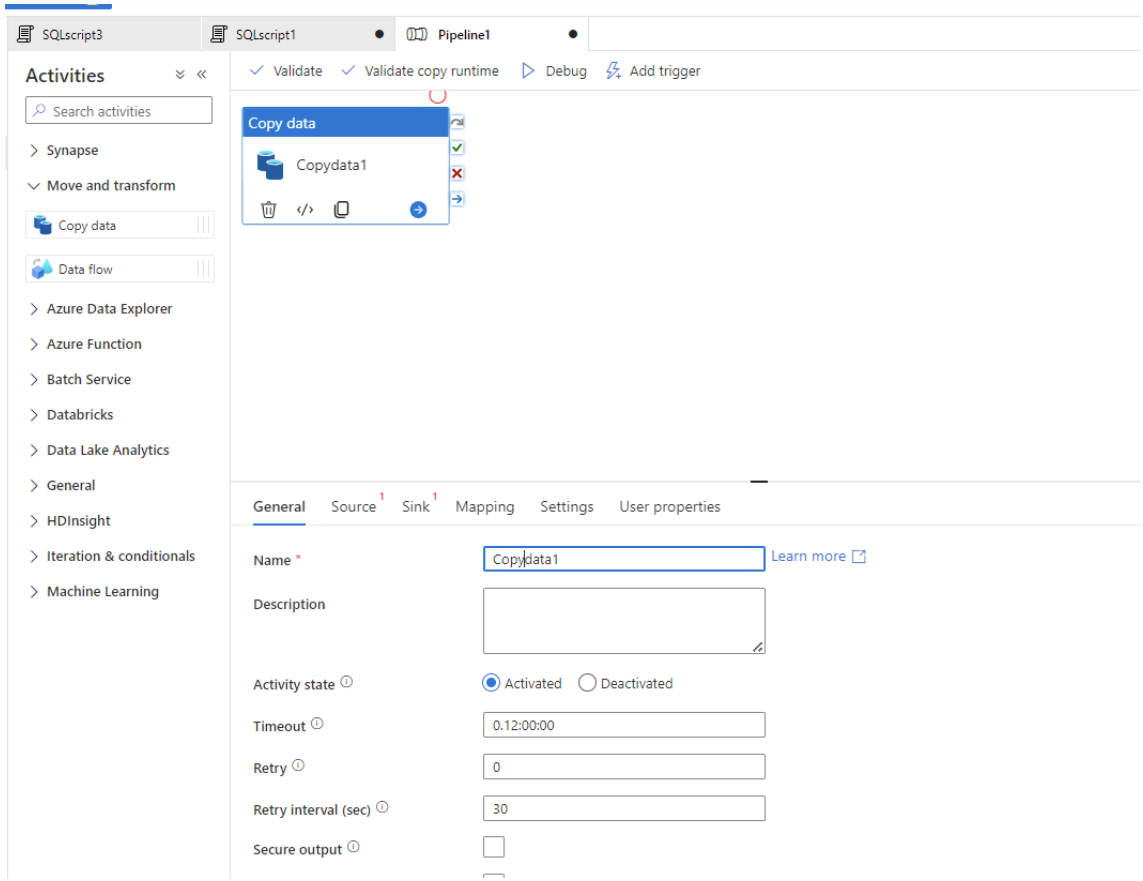
Introduction to Petabyte Scale Ingestion

- Create one and set the properties as shown below. This will be used to create a pipeline.



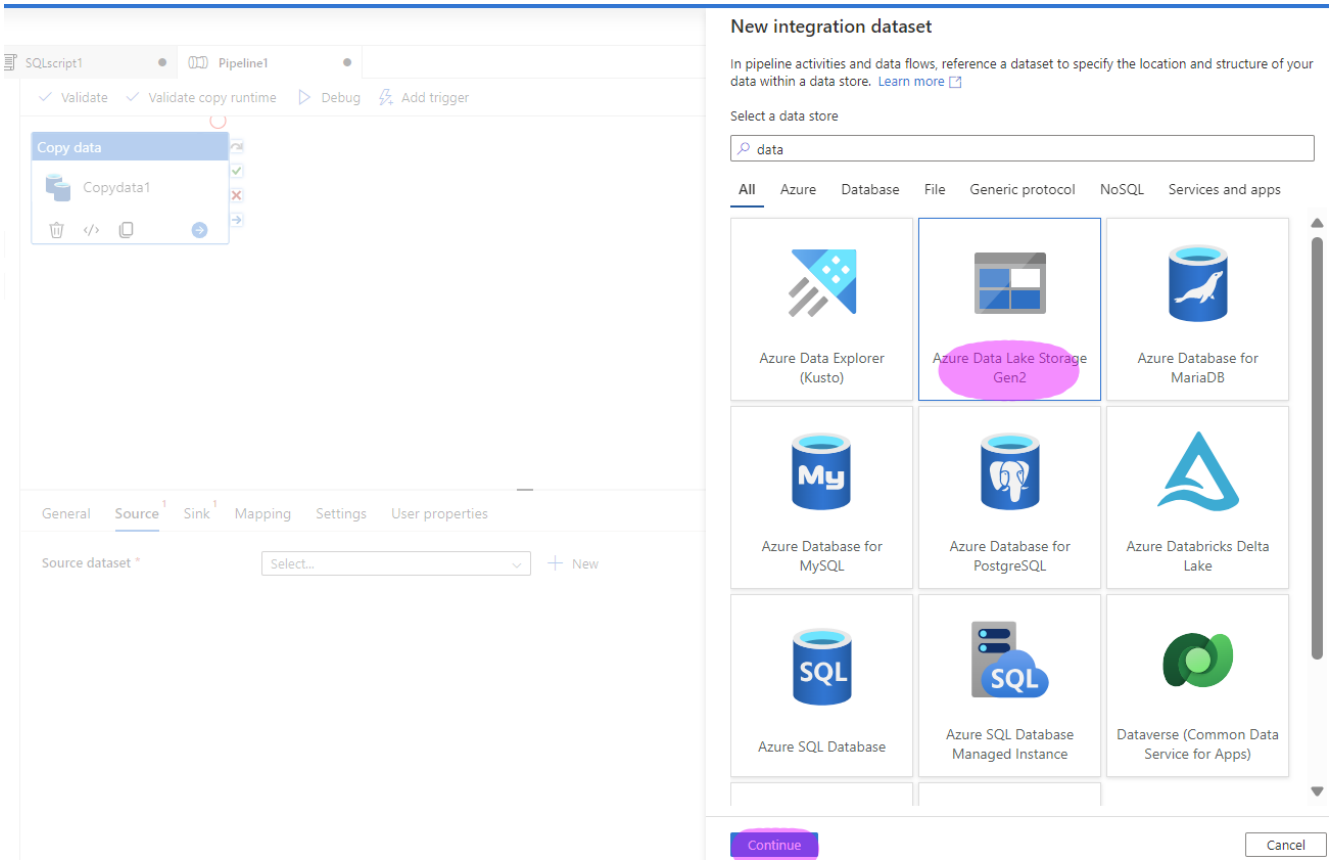
Introduction to Petabyte Scale Ingestion

- Now go to the Integrate tab. Click on + and `Pipeline`
  - Name the pipeline
  - From the `Activities` section, under `Move and transform`, drag and drop the copy data
  - Rename the `copy data 1` from the general tab



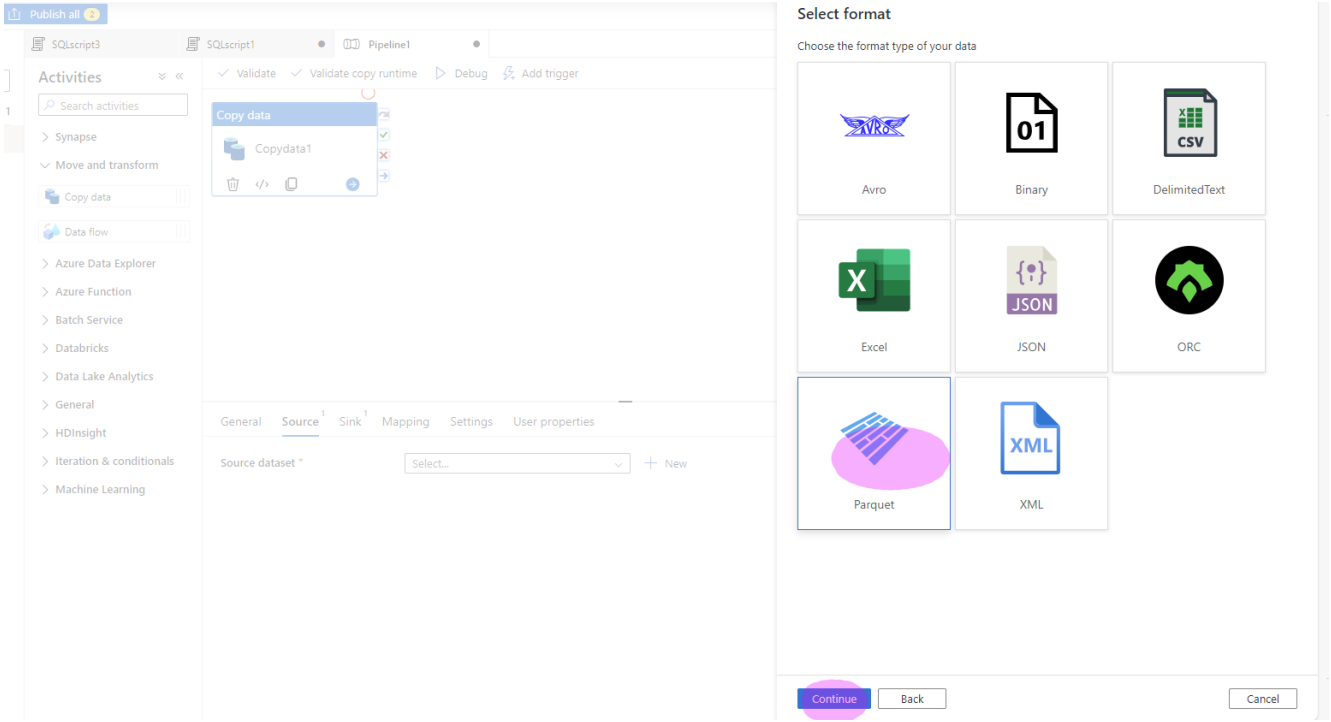
Introduction to Petabyte Scale Ingestion

- Go to Source and click on `+New` button.
  - Select Azure Data Lake Storage Gen-2 and click on `Continue`.



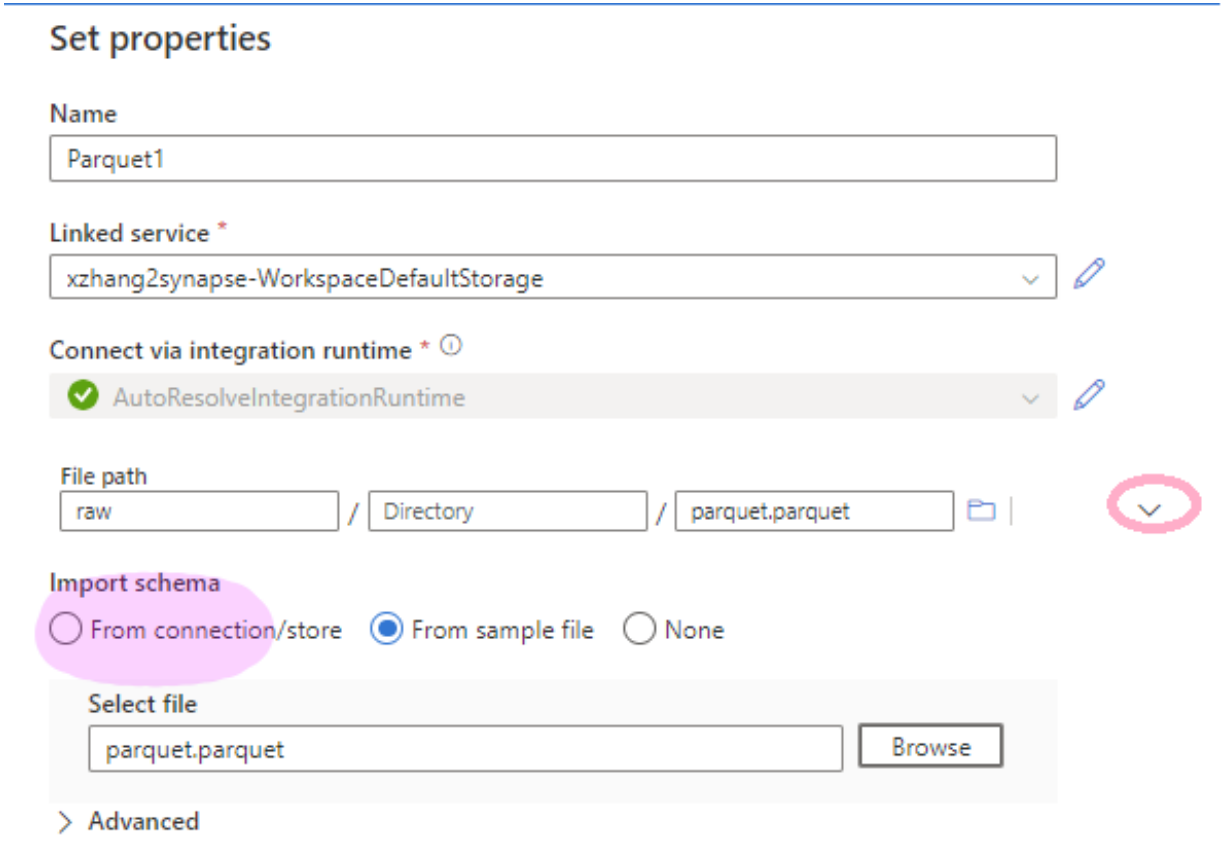
Introduction to Petabyte Scale Ingestion

- Select the Paraquet format and click on `Continue`.



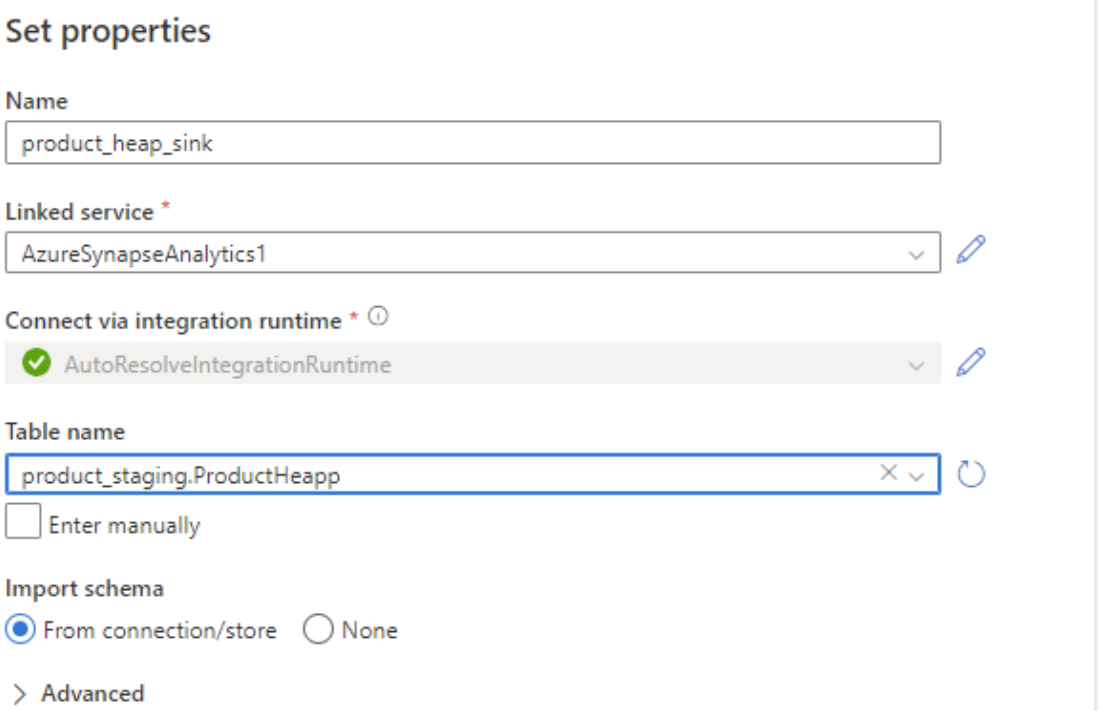
Introduction to Petabyte Scale Ingestion

- Set the properties as shown and click "OK".
  - Preview data to make sure source data is accessible!



Introduction to Petabyte Scale Ingestion

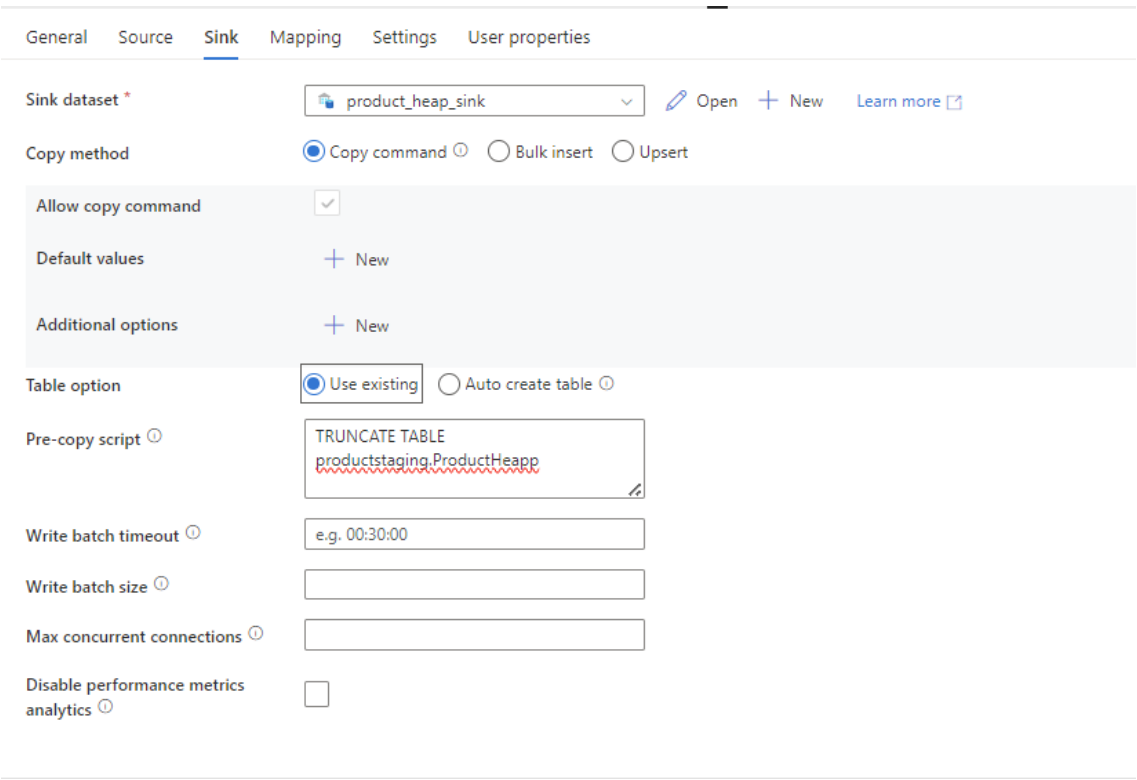
- Then go to the Sink option and click on `+New`
  - Select Azure Synapse Analytics and click on `Continue`.
  - Then Set the properties as shown and click `OK`.



Introduction to Petabyte Scale Ingestion

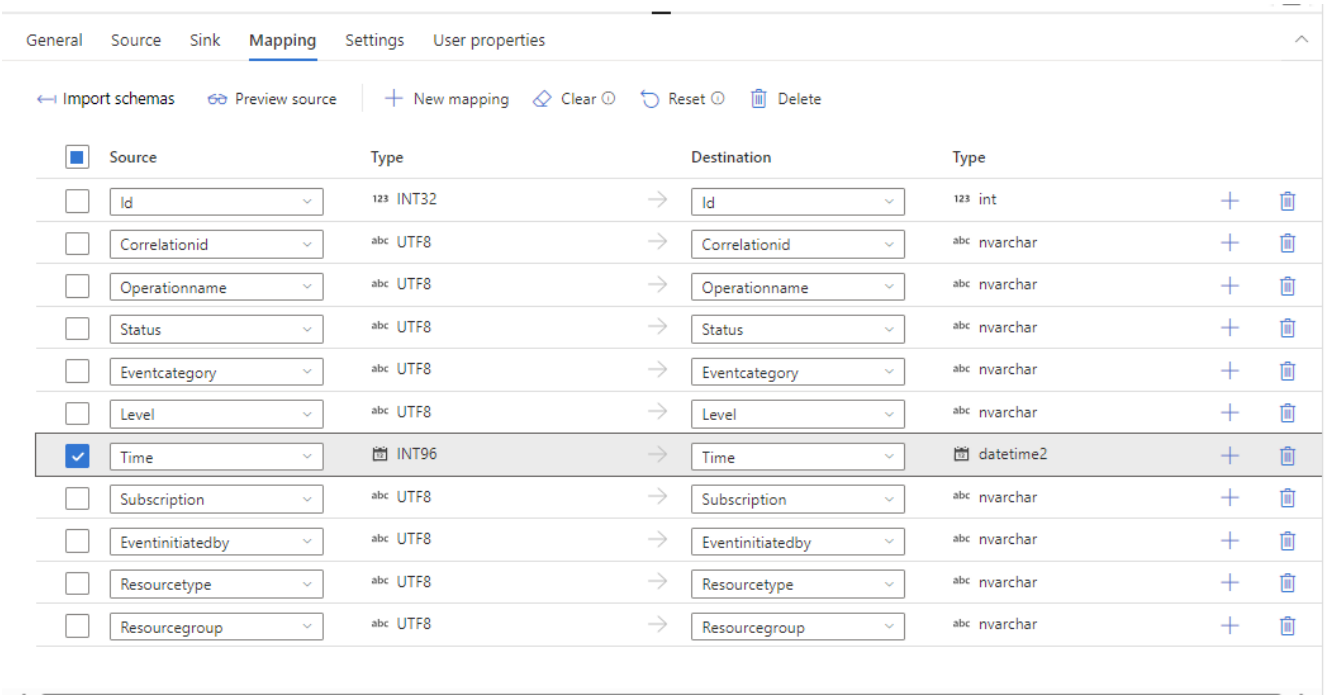
- In the `Sink` section, tick the **Copy command** option and mention **pre-copy script** to empty the table in case of any database. Use the command below:

```
TRUNCATE TABLE productstaging.ProductHeapp
```



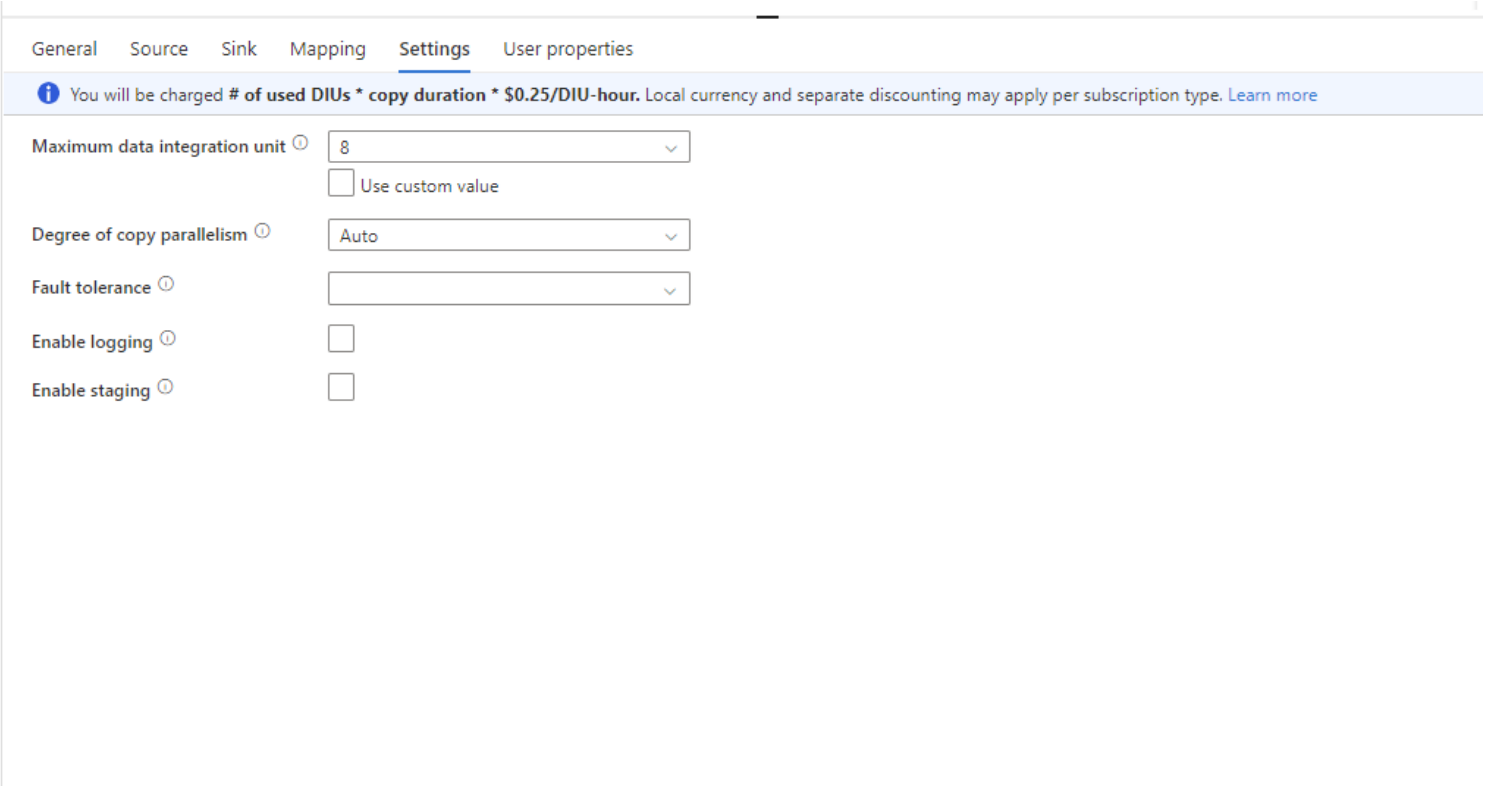
Introduction to Petabyte Scale Ingestion

- Now go to Mapping and click on “Import schemas”.
- In Mapping, select a column to be mapped. Here, let’s select the `Time` column.



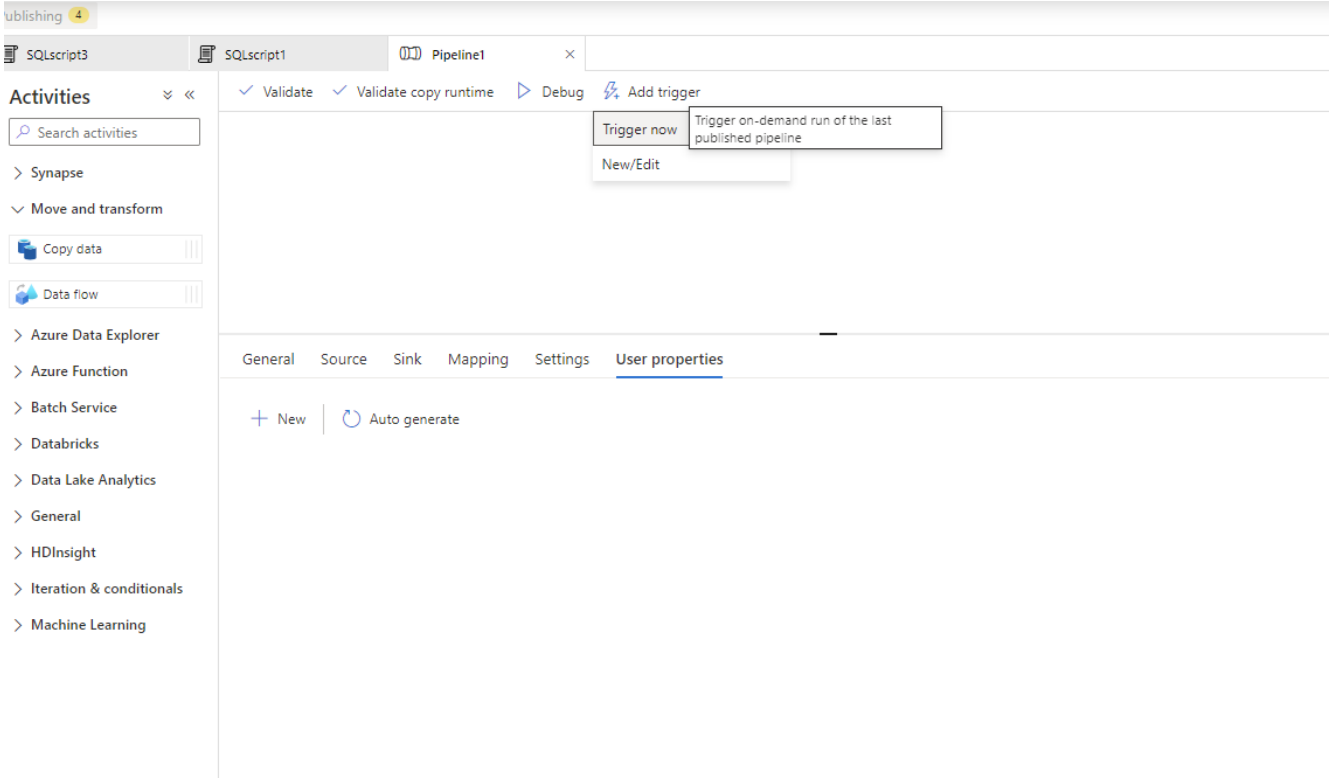
Introduction to Petabyte Scale Ingestion

- Go to the Settings tab and change the `maximum data integration unit` value to 8



Introduction to Petabyte Scale Ingestion

- Now click on the `Publish All` button
- Validate the pipeline
- Debug the pipeline to make sure it can run
- Then click on "Add trigger" and "Trigger now"



Introduction to Petabyte Scale Ingestion

- Go to the `Monitor` tab and check if the pipeline can run successfully.

License