

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
```

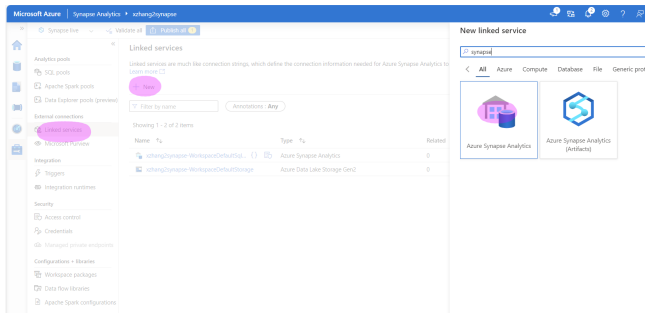
Introduction to Petabyte Scale Ingestion

- 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.

New linked service
Azure Synapse Analytics [Learn more](#)

Choose a name for your linked service. This name cannot be updated later.

Name *
AzureSynapseAnalytics1

Description

Connect via integration runtime *
☒ AutoResolveIntegrationRuntime

Version
☒ Recommended ☐ Legacy
[Import from connection string](#)

Account selection method
☒ From Azure subscription ☐ Enter manually

Azure subscription
Azure for Students (b5db3bdc-dae9-4e53-888f-8241b09dc786)

Server name *
xzhang2synapse (Synapse workspace)

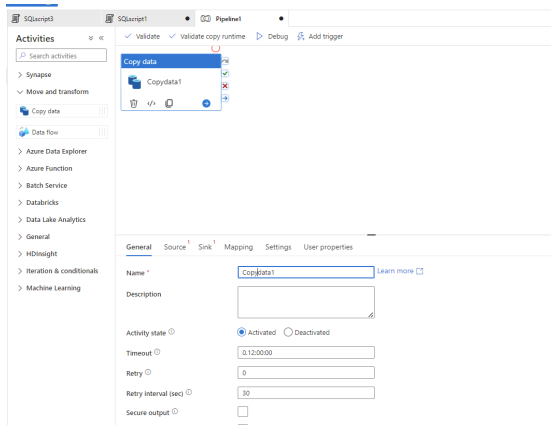
Database name *
xzhang2pool

SQL pool *
☒ xzhang2pool

Authentication type *

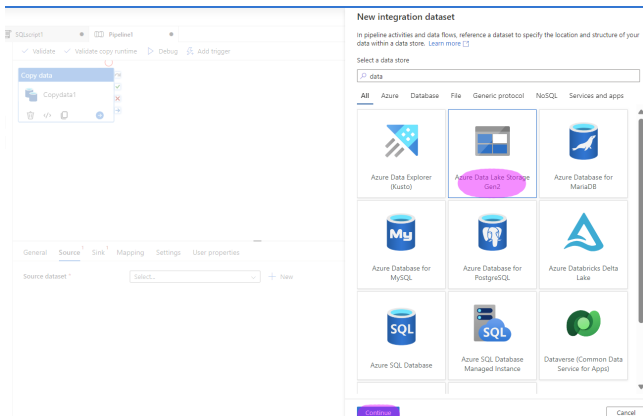
Introduction to Petabyte Scale Ingestion

- 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



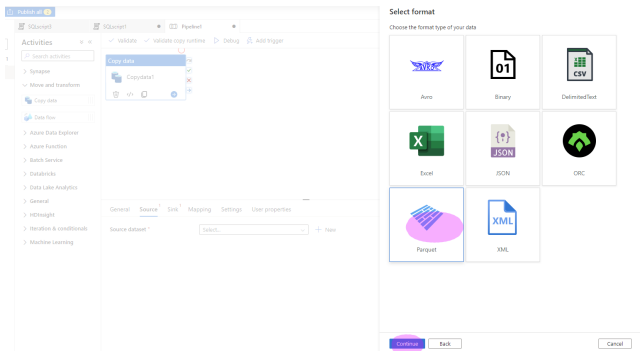
Introduction to Petabyte Scale Ingestion

- 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 Parquet 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!

Set properties

Name

Parquet1

Linked service *

xzhang2synapse-WorkspaceDefaultStorage

Connect via integration runtime * ⓘ

✓ AutoResolveIntegrationRuntime

File path

raw

/ Directory

/ parquet.parquet

Import schema

☐ From connection/store ☒ From sample file ☐ None

Select file

parquet.parquet

Browse

> Advanced

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.

Set properties

Name

product_heap_sink

Linked service *

AzureSynapseAnalytics1

Connect via integration runtime * ⓘ

✓ AutoResolveIntegrationRuntime

Table name

product_staging.ProductHeapp

☐ Enter manually

Import schema

☒ From connection/store ☐ None

> Advanced

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:

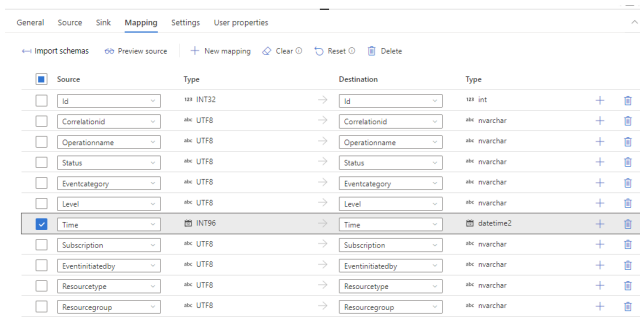
```
TRUNCATE TABLE productstaging.ProductHeapp
```

The screenshot shows the 'Sink' configuration tab in a Databricks interface. The 'Sink dataset' is set to 'product_heap_sink'. Under 'Copy method', 'Copy command' is selected. The 'Allow copy command' checkbox is checked. The 'Table option' is set to 'Use existing'. The 'Pre-copy script' field contains the SQL command 'TRUNCATE TABLE productstaging.ProductHeapp'. Other fields like 'Write batch timeout', 'Write batch size', 'Max concurrent connections', and 'Disable performance metrics analytics' are present but empty.

General	Source	Sink	Mapping	Settings	User properties
Sink dataset [*] <input type="text" value="product_heap_sink"/> Open + New Learn more					
Copy method <input checked="" type="radio"/> Copy command <input type="radio"/> Bulk insert <input type="radio"/> Upsert					
Allow copy command <input checked="" type="checkbox"/>					
Default values + New					
Additional options + New					
Table option <input checked="" type="radio"/> Use existing <input type="radio"/> Auto create table					
Pre-copy script <input type="text" value="TRUNCATE TABLE productstaging.ProductHeapp"/>					
Write batch timeout <input type="text" value="e.g. 00:30:00"/>					
Write batch size <input type="text"/>					
Max concurrent connections <input type="text"/>					
Disable performance metrics analytics <input type="checkbox"/>					

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.



The screenshot shows the 'Mapping' tab of a data ingestion tool. At the top, there are tabs for 'General', 'Source', 'Sink', 'Mapping' (selected), 'Settings', and 'User properties'. Below the tabs, there are buttons for 'Import schemas', 'Preview source', 'New mapping', 'Clear', 'Reset', and 'Delete'. The main area contains a table with columns: 'Source', 'Type', 'Destination', and 'Type'. The 'Time' row is selected, showing a mapping from 'INT96' to 'datetime2'. Other rows show mappings for 'Id', 'Correlationid', 'Operationname', 'Status', 'Eventcategory', 'Level', 'Subscription', 'Eventinitiatedby', 'Resource type', and 'Resourcegroup'.

Source	Type	Destination	Type
<input type="checkbox"/> Id	int INT32	Id	int
<input type="checkbox"/> Correlationid	str UTF8	Correlationid	str nvarchar
<input type="checkbox"/> Operationname	str UTF8	Operationname	str nvarchar
<input type="checkbox"/> Status	str UTF8	Status	str nvarchar
<input type="checkbox"/> Eventcategory	str UTF8	Eventcategory	str nvarchar
<input type="checkbox"/> Level	str UTF8	Level	str nvarchar
<input checked="" type="checkbox"/> Time	int INT96	Time	datetime2
<input type="checkbox"/> Subscription	str UTF8	Subscription	str nvarchar
<input type="checkbox"/> Eventinitiatedby	str UTF8	Eventinitiatedby	str nvarchar
<input type="checkbox"/> Resource type	str UTF8	Resource type	str nvarchar
<input type="checkbox"/> Resourcegroup	str UTF8	Resourcegroup	str nvarchar

Introduction to Petabyte Scale Ingestion

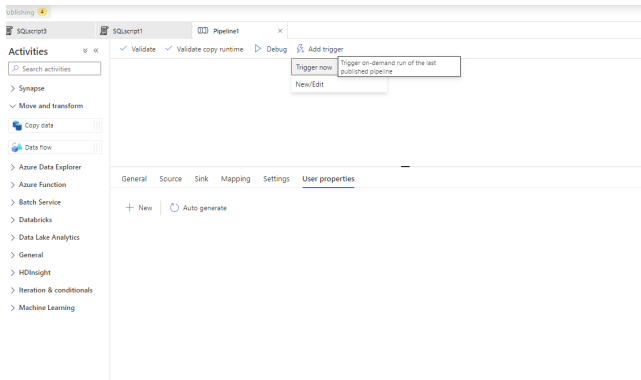
- Go to the Settings tab and change the maximum data integration unit value to 8

The screenshot shows a settings interface with tabs: General, Source, Sink, Mapping, Settings (selected), and User properties. A blue banner at the top of the Settings tab contains a warning icon and the text: "You will be charged # of used DIUs * copy duration * \$0.25/DIU-hour. Local currency and separate discounting may apply per subscription type. [Learn more](#)". Below the banner, the following settings are visible:

- Maximum data integration unit** (with a help icon): A dropdown menu is set to "8". Below it is a checkbox labeled "Use custom value" which is unchecked.
- Degree of copy parallelism** (with a help icon): A dropdown menu is set to "Auto".
- Fault tolerance** (with a help icon): A dropdown menu is currently empty.
- Enable logging** (with a help icon): An unchecked checkbox.
- Enable staging** (with a help icon): An unchecked checkbox.

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



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).