**Step 1: Prepare the Data Infrastructure**

1.Create the data lake and upload data

A computer screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Storage Container

A screenshot of a computer

AI-generated content may be incorrect.

Three directories in storage container

A screenshot of a computer

AI-generated content may be incorrect.

Uploaded files in **dirpayrollfiles**

A screenshot of a computer

AI-generated content may be incorrect.

Uploaded file in **dirhistoryfiles**

A screenshot of a computer

AI-generated content may be incorrect.

2. Create an Azure Data Factory Resource

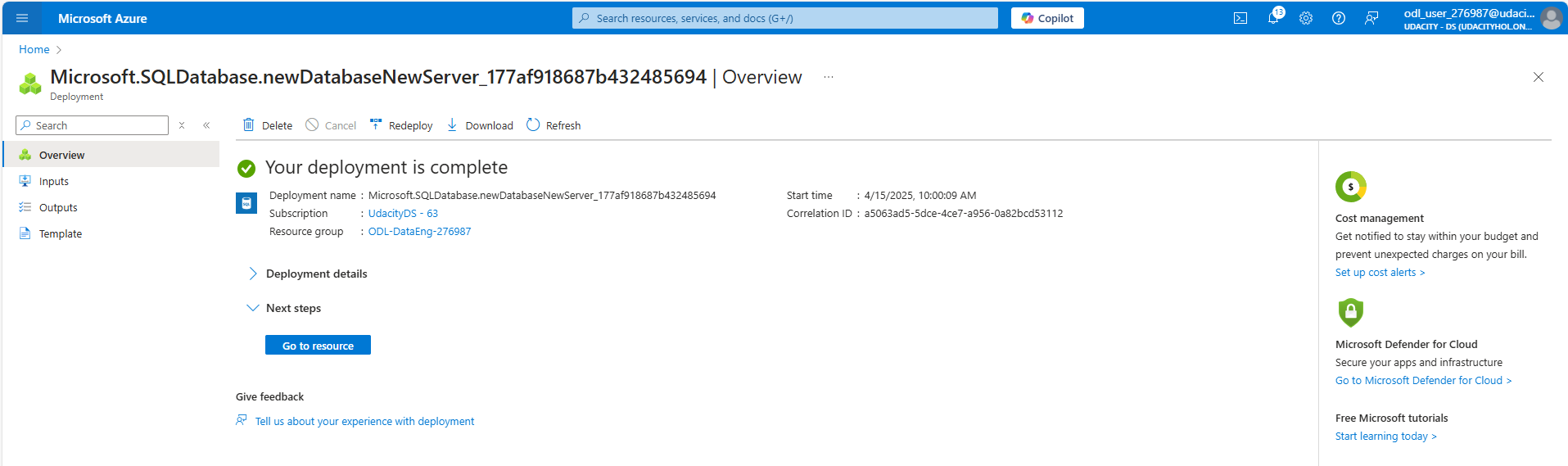
A screenshot of a computer

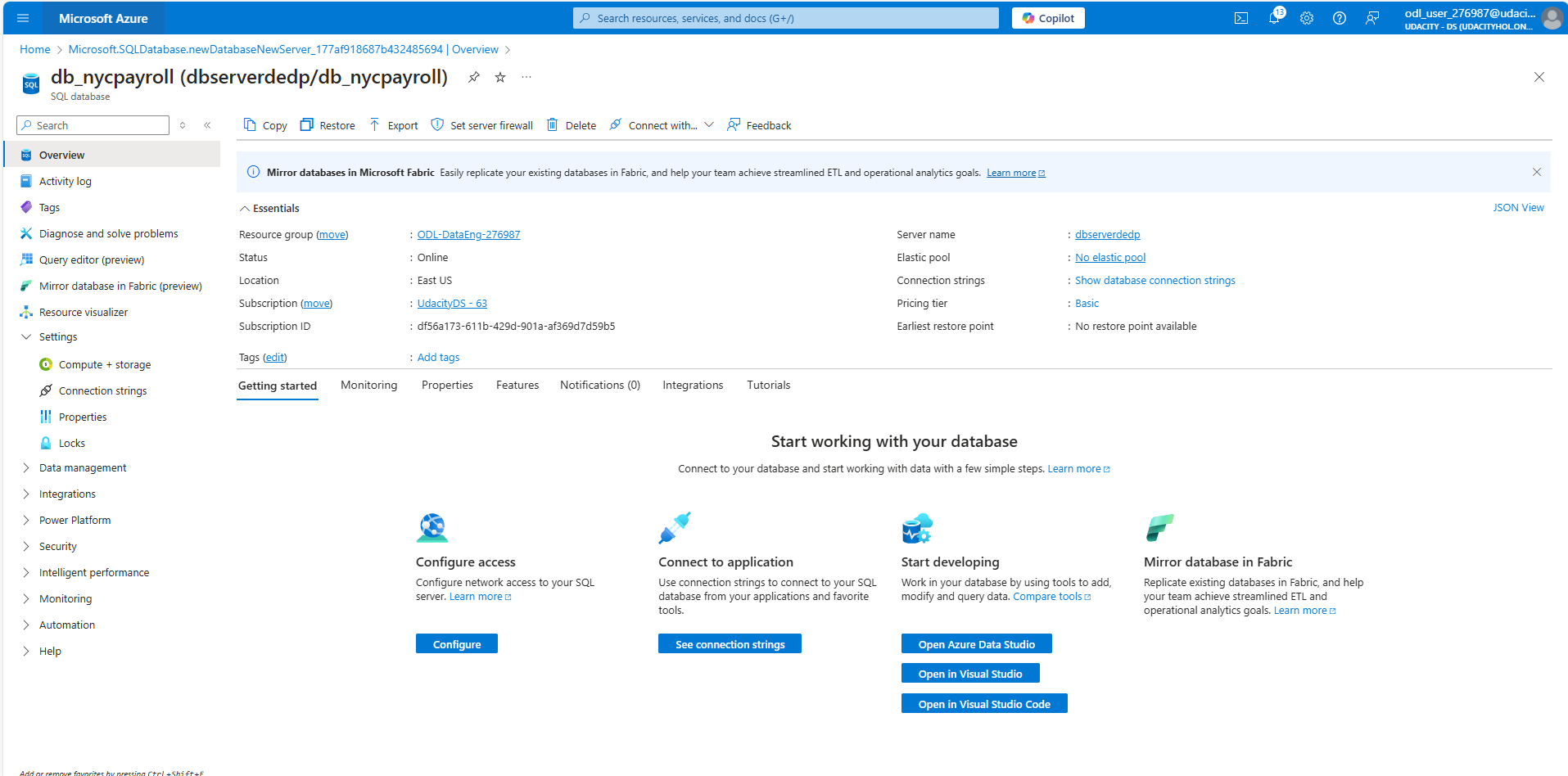
AI-generated content may be incorrect.

A screenshot of a computer

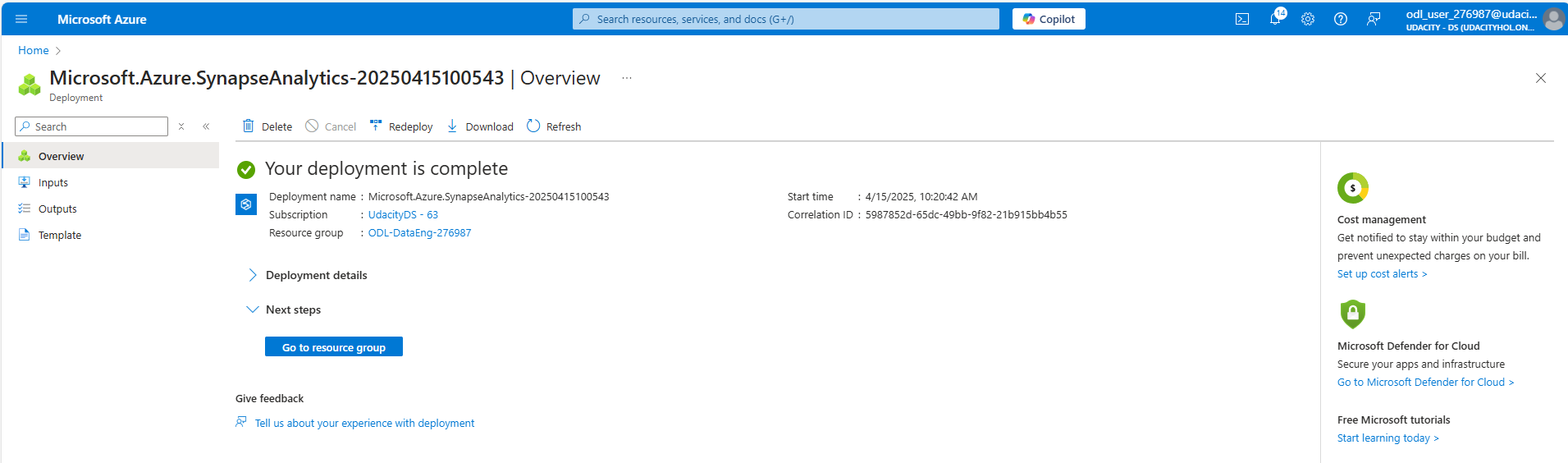
AI-generated content may be incorrect.

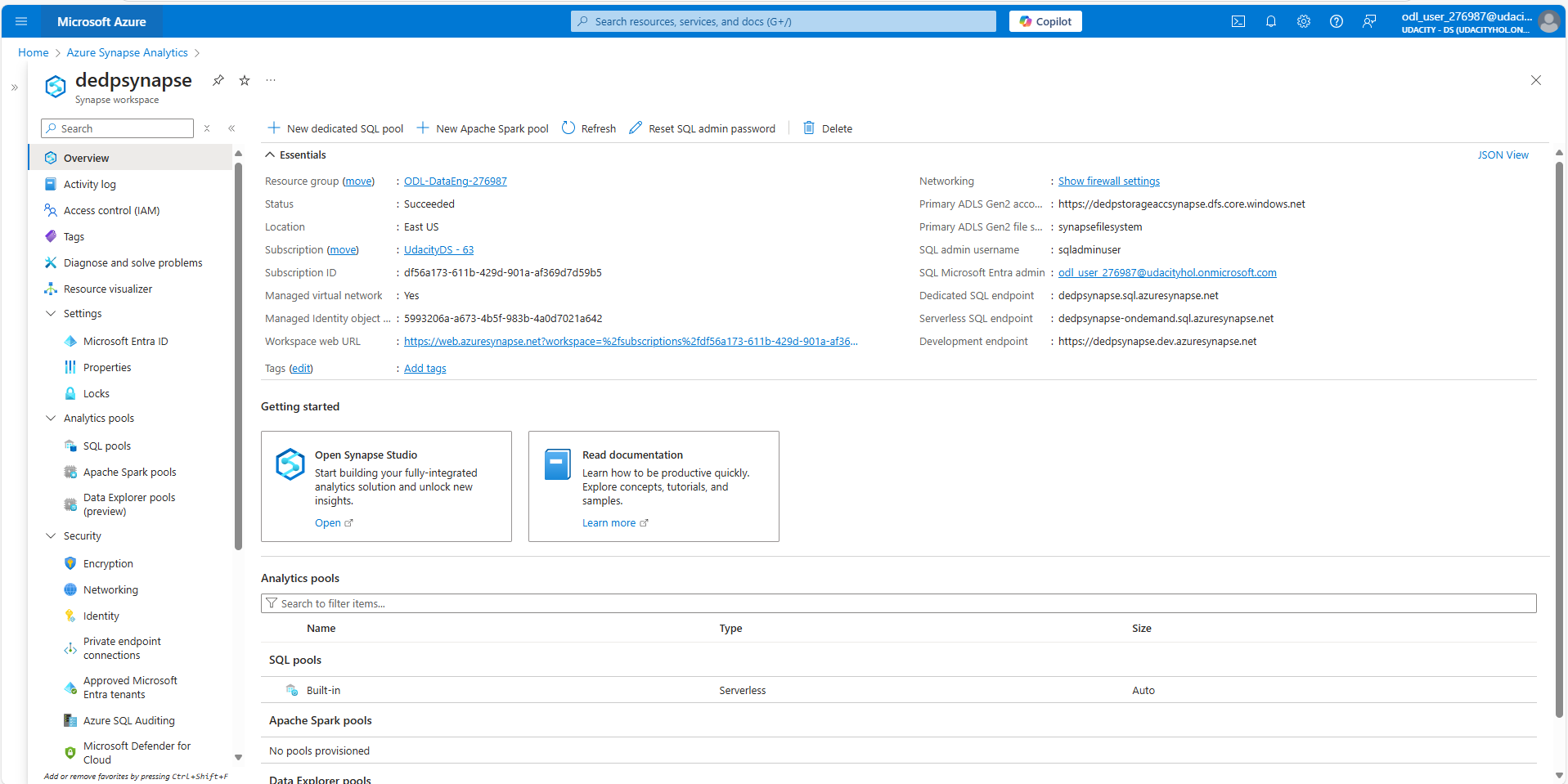
3. Create a SQL Database

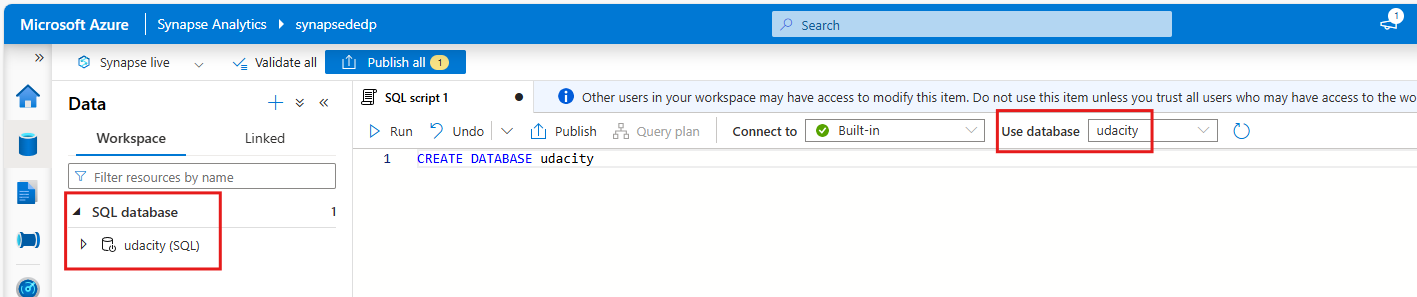




4. Create a Synapse Analytics workspace, or use one you already have created

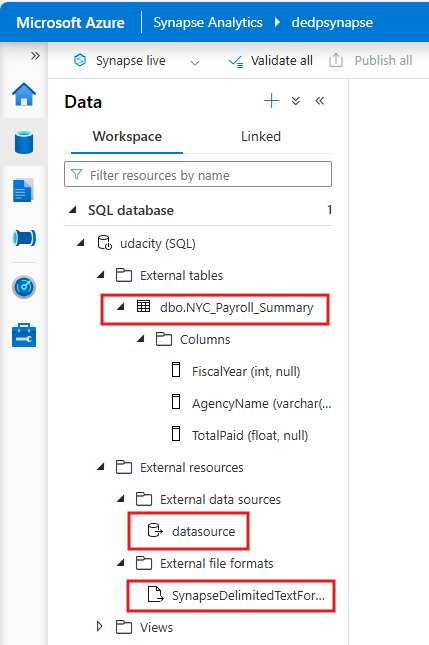




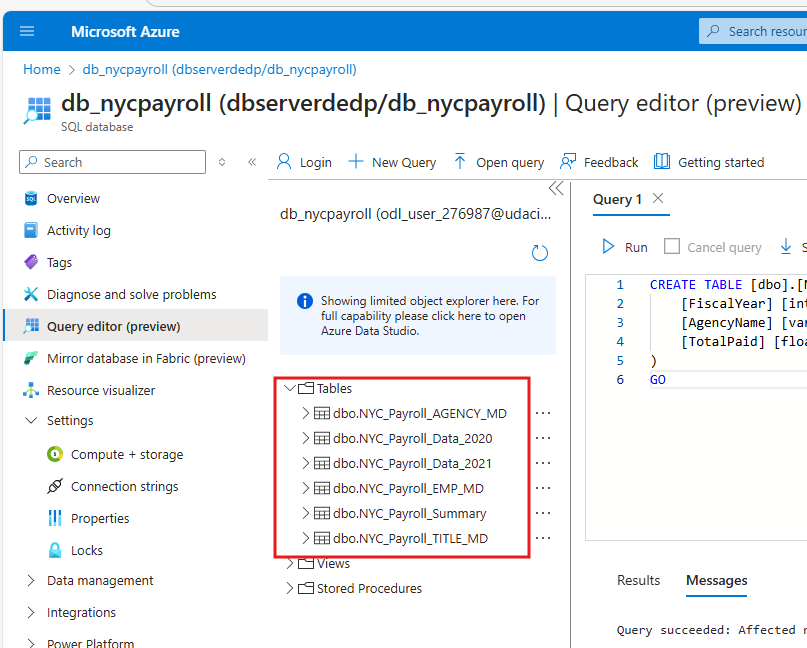


5. Create summary data external table in Synapse Analytics workspace

Created Data Source, File format and External tables

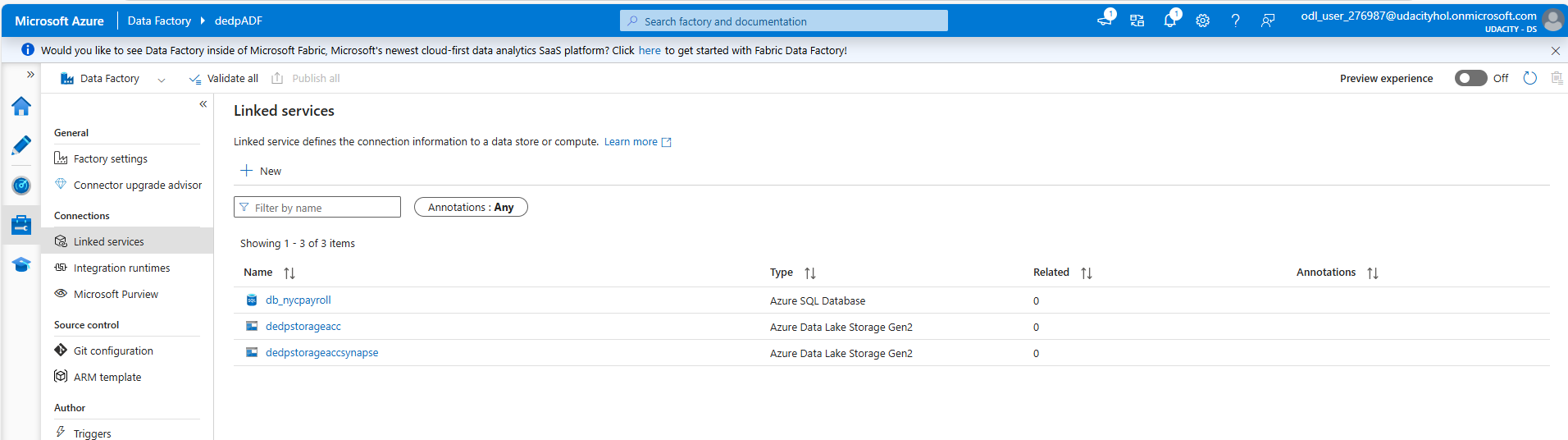


6. Create master data tables and payroll transaction tables in SQL DB



**Step 2: Create Linked Services**

1.Create a Linked Service for Azure Data Lake



**Step 3: Create Datasets in Azure Data Factory**

1.Create the datasets for the 2021 Payroll file on Azure Data Lake Gen2

2. Repeat the same process to create datasets for the rest of the data files in the Data Lake

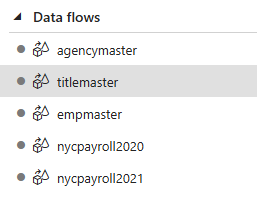
3. Create the dataset for all the data tables in SQL DB

4. Create the datasets for destination (target) table in Synapse Analytics

A screenshot of a computer

AI-generated content may be incorrect.

**Step 4: Create Data Flows**



**agencymaster**

**A diagram of a diagram

AI-generated content may be incorrect.**

**titlemaster**

**A screenshot of a computer

AI-generated content may be incorrect.**

**empmaster**

**A diagram of a diagram

AI-generated content may be incorrect.**

**nycpayroll2020**

****

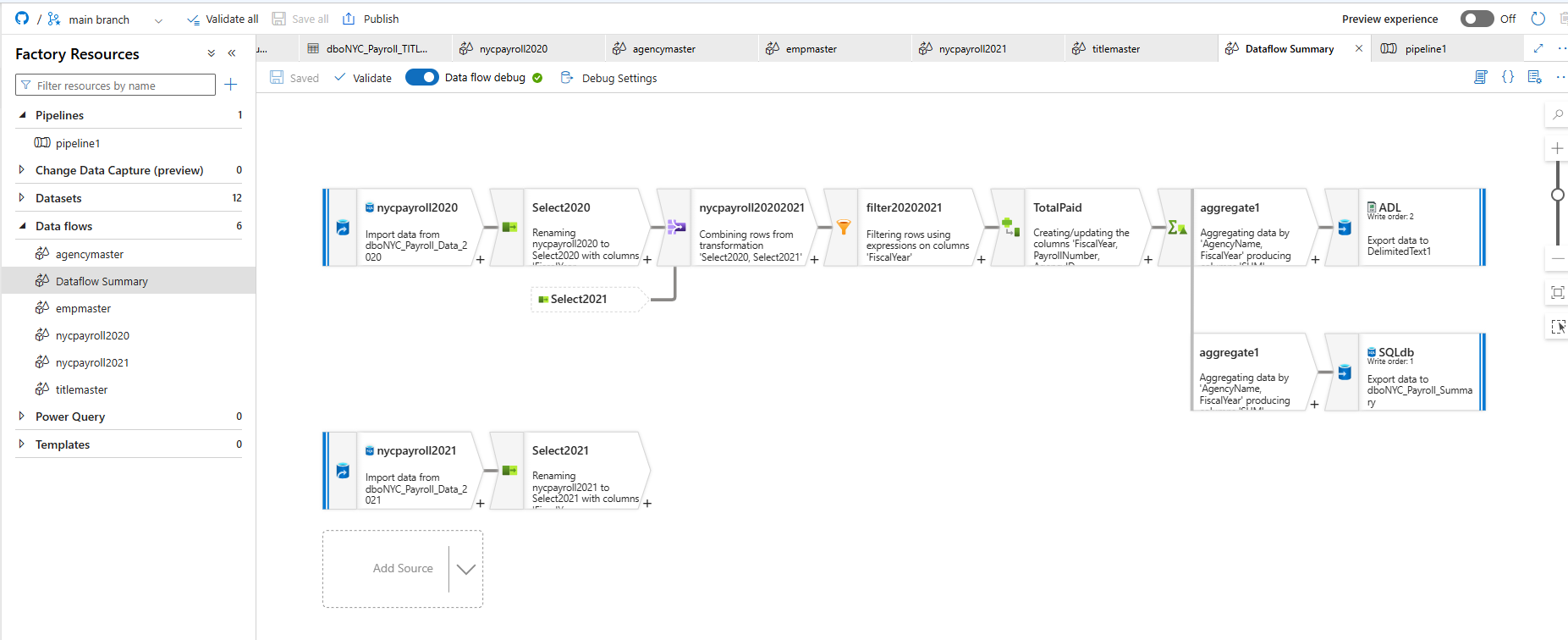
**nycpayroll2021**

**A diagram with a blue cylinder and a blue arrow

AI-generated content may be incorrect.**

**Step 5: Data Aggregation and Parameterization**

**Dataflow Summary**

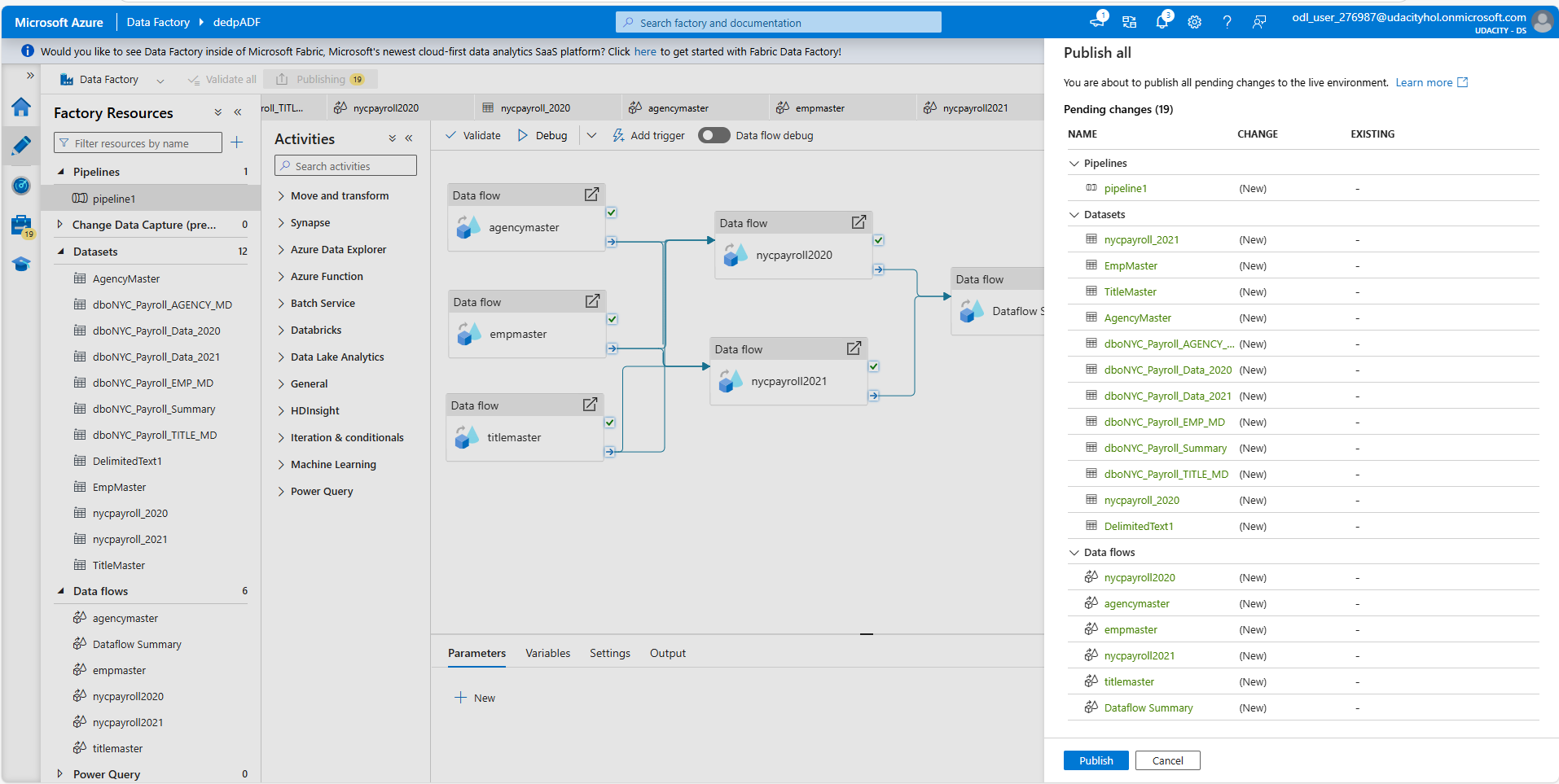


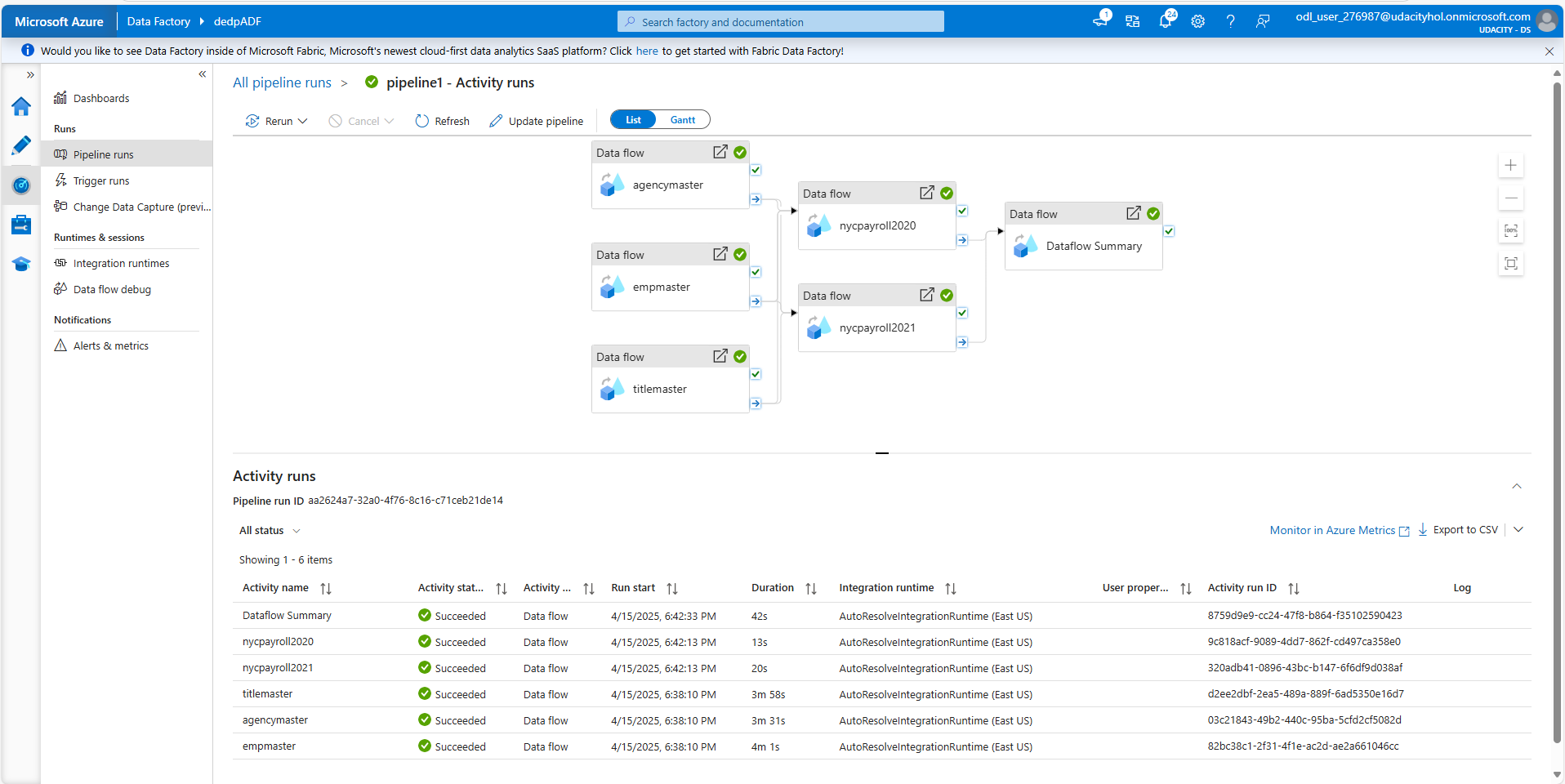
**Step 6: Pipeline Creation**

A screenshot of a computer

AI-generated content may be incorrect.

**Step 7: Trigger and Monitor Pipeline**





**Step 8: Verify Pipeline run artifacts**

dirstaging directory listing in Datalake

A screenshot of a computer

AI-generated content may be incorrect.

SQL DB summary table

A screenshot of a computer

AI-generated content may be incorrect.

Synapse summary external table

A screenshot of a computer

AI-generated content may be incorrect.