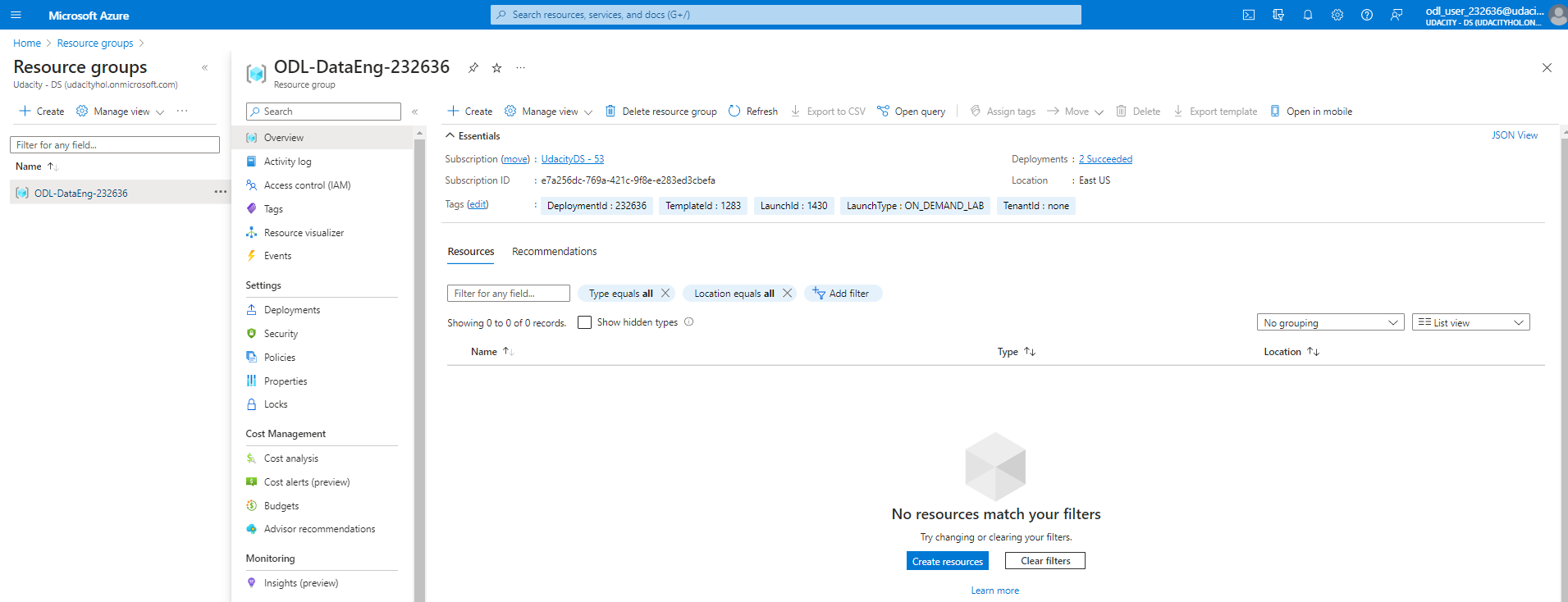
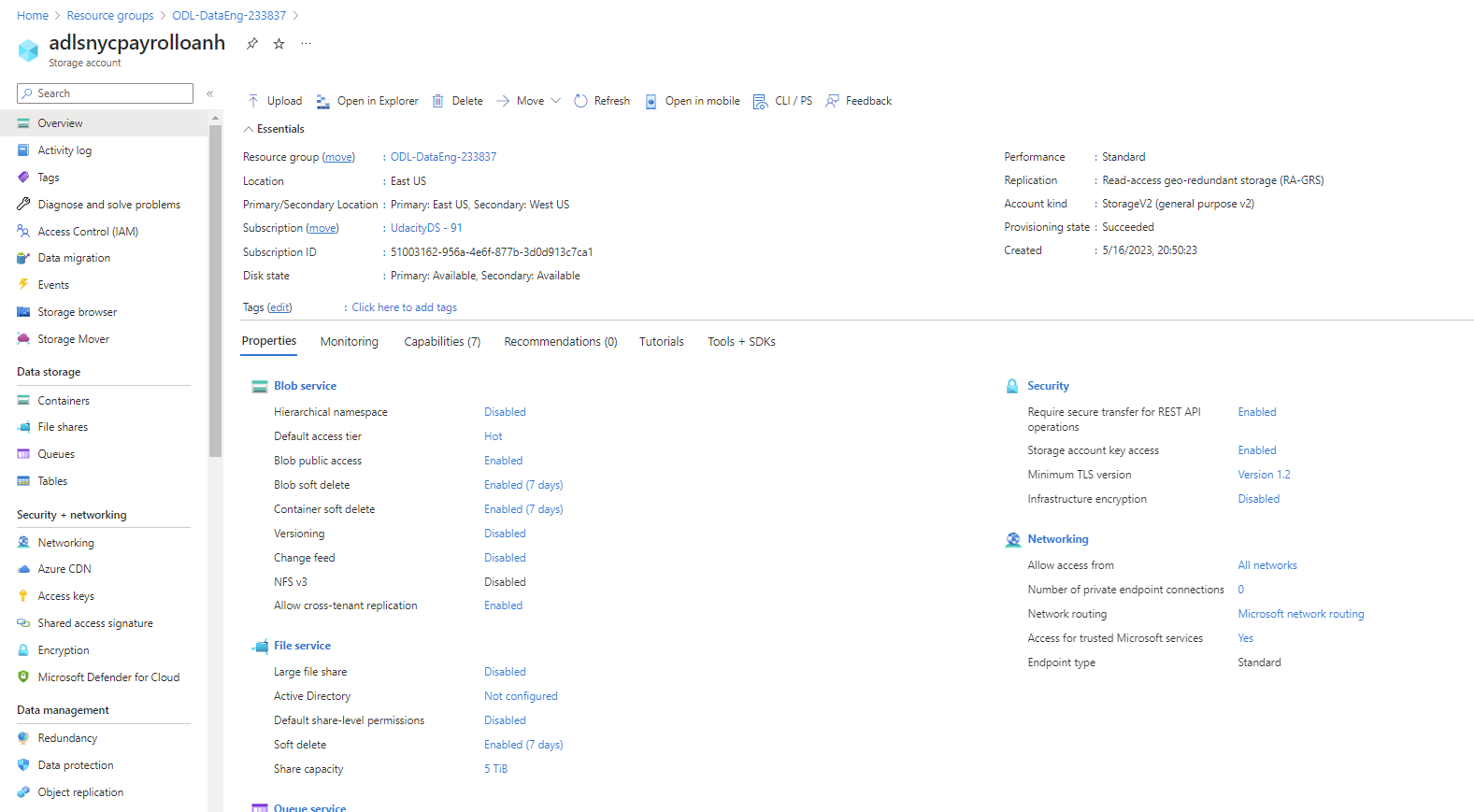
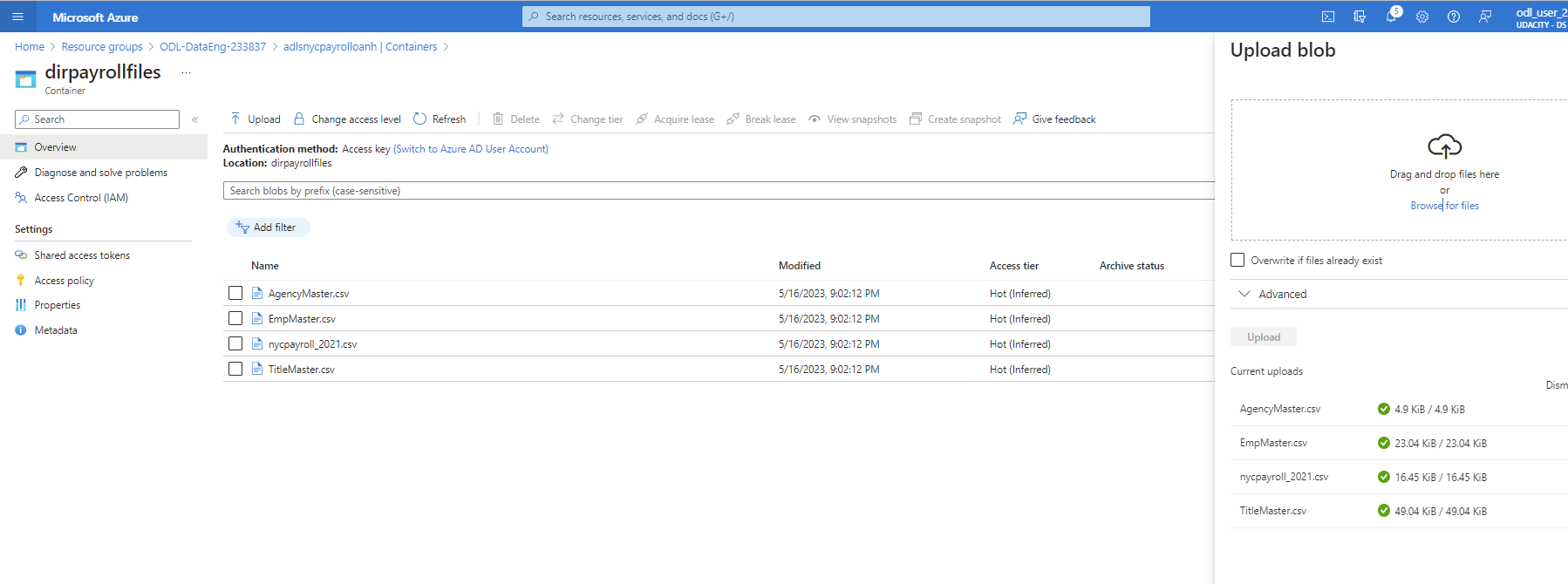
Step 1: Prepare the Data Infrastructure

**1.Create the data lake and upload data**



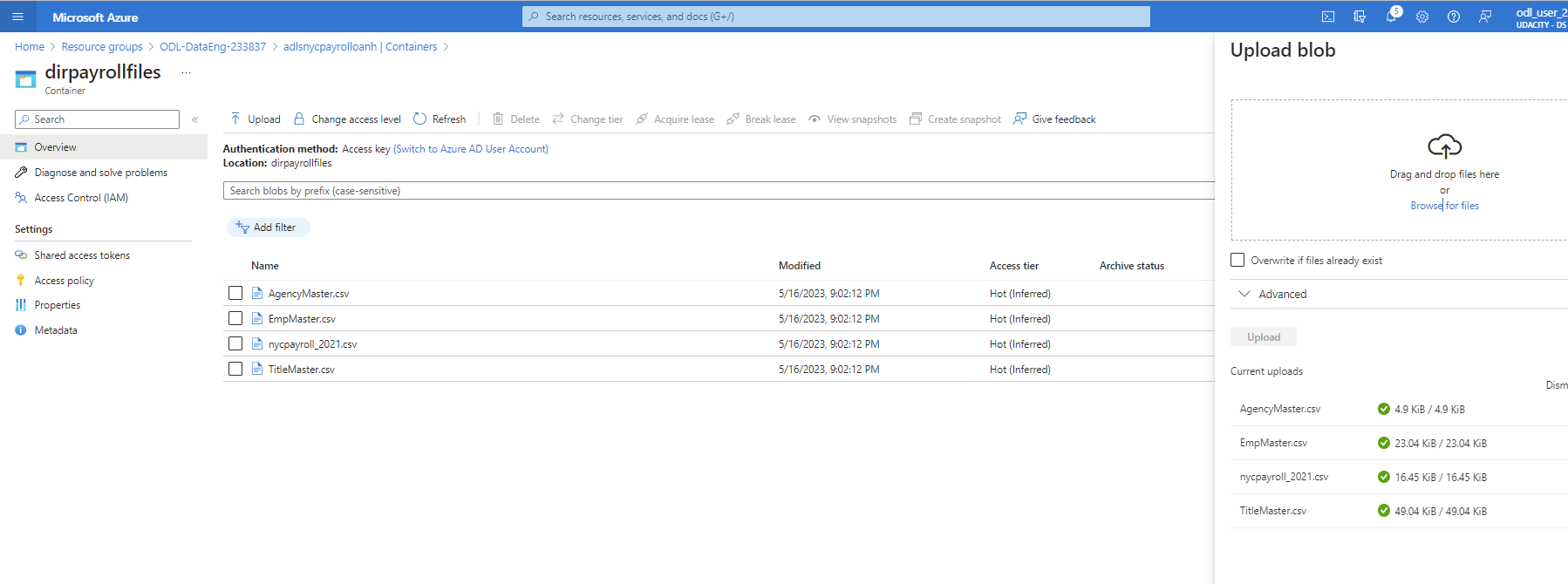
Create an Azure Data Lake Storage Gen2





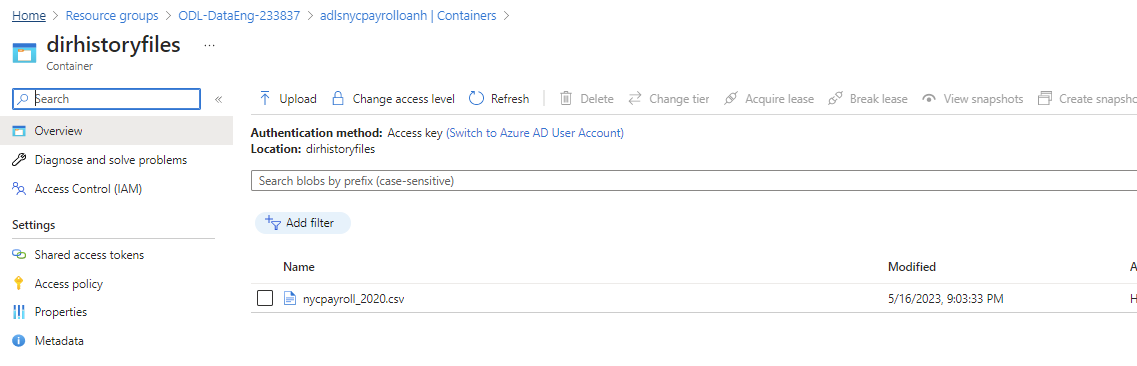
Upload these files from the [**project data**](https://video.udacity-data.com/topher/2022/May/6283aff5_data-nyc-payroll/data-nyc-payroll.zip) to the **dirpayrollfiles** folder

* EmpMaster.csv
* AgencyMaster.csv
* TitleMaster.csv
* nycpayroll\_2021.csv

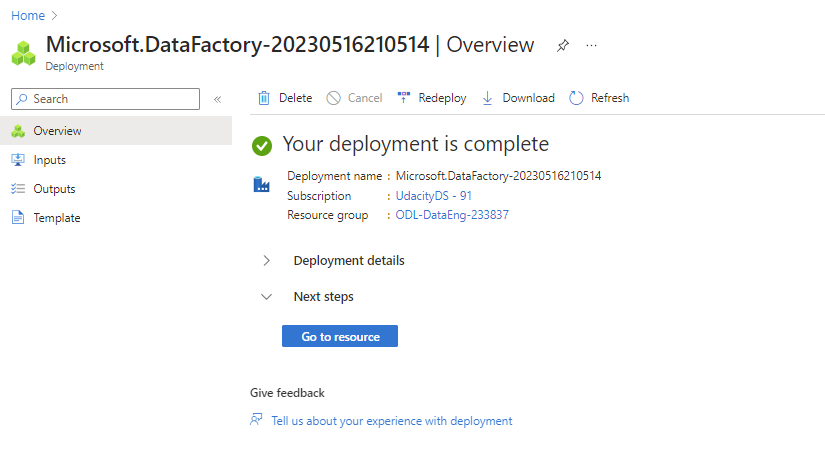


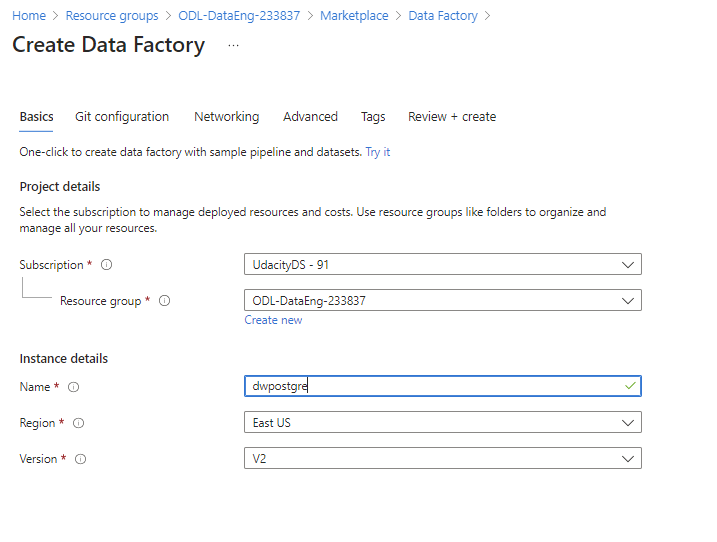
Upload this file (historical data) from the [**project data**](https://video.udacity-data.com/topher/2022/May/6283aff5_data-nyc-payroll/data-nyc-payroll.zip) to the **dirhistoryfiles** folder

* nycpayroll\_2020.csv

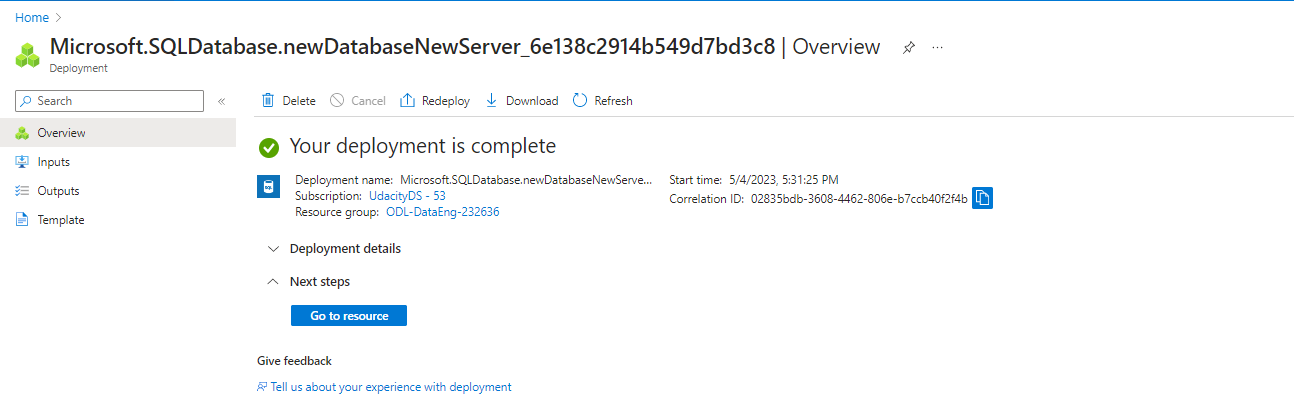


**2. Create an Azure Data Factory Resource**



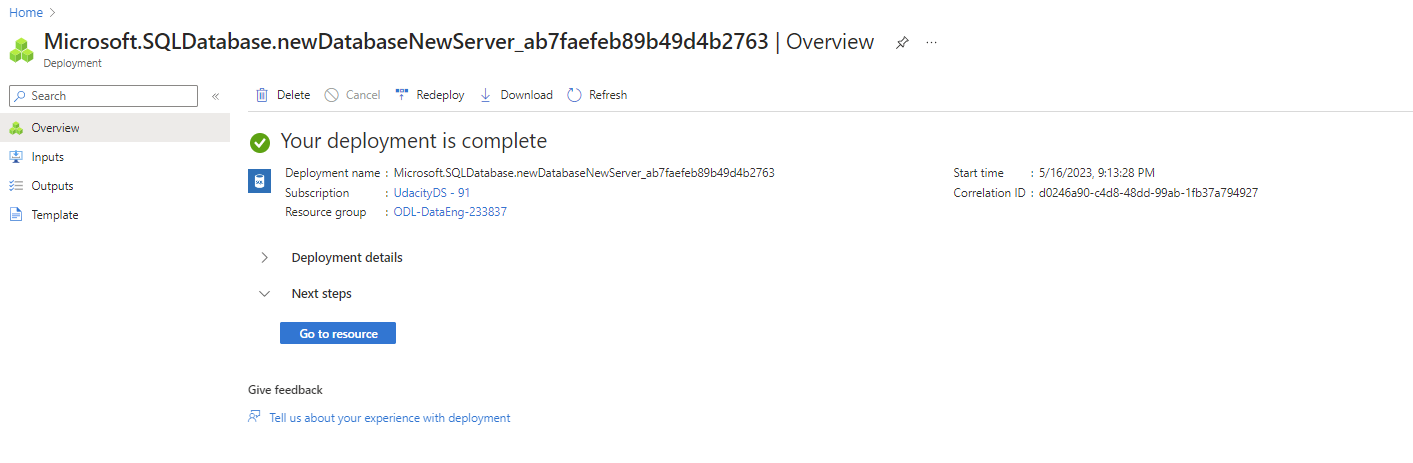


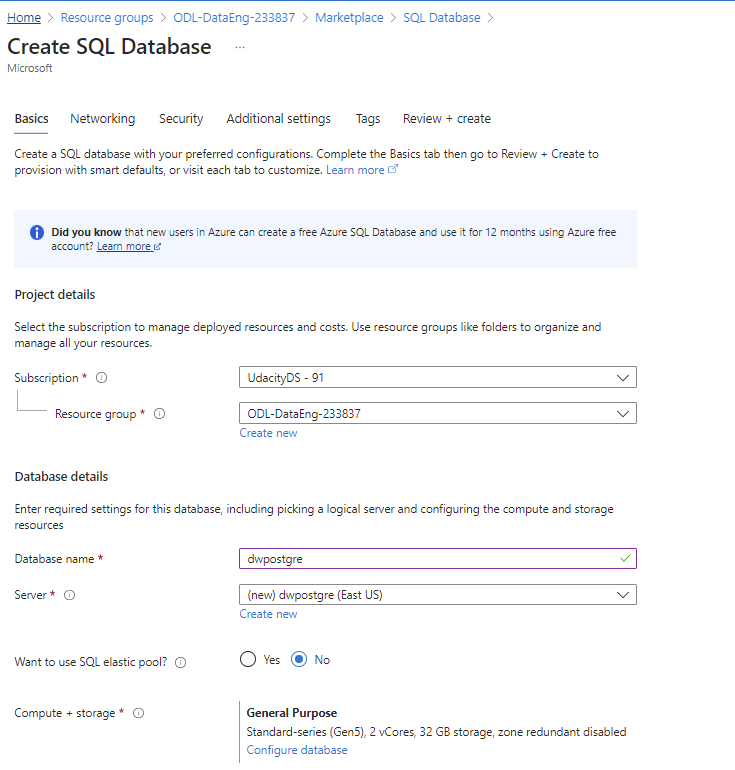
**3. Create a SQL Database to store the current year of the payroll data**

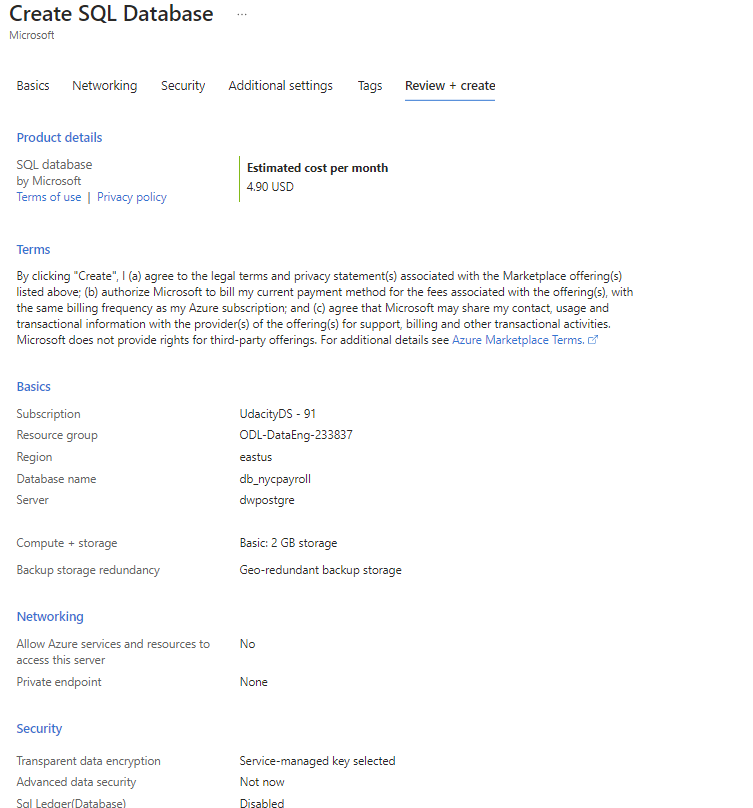


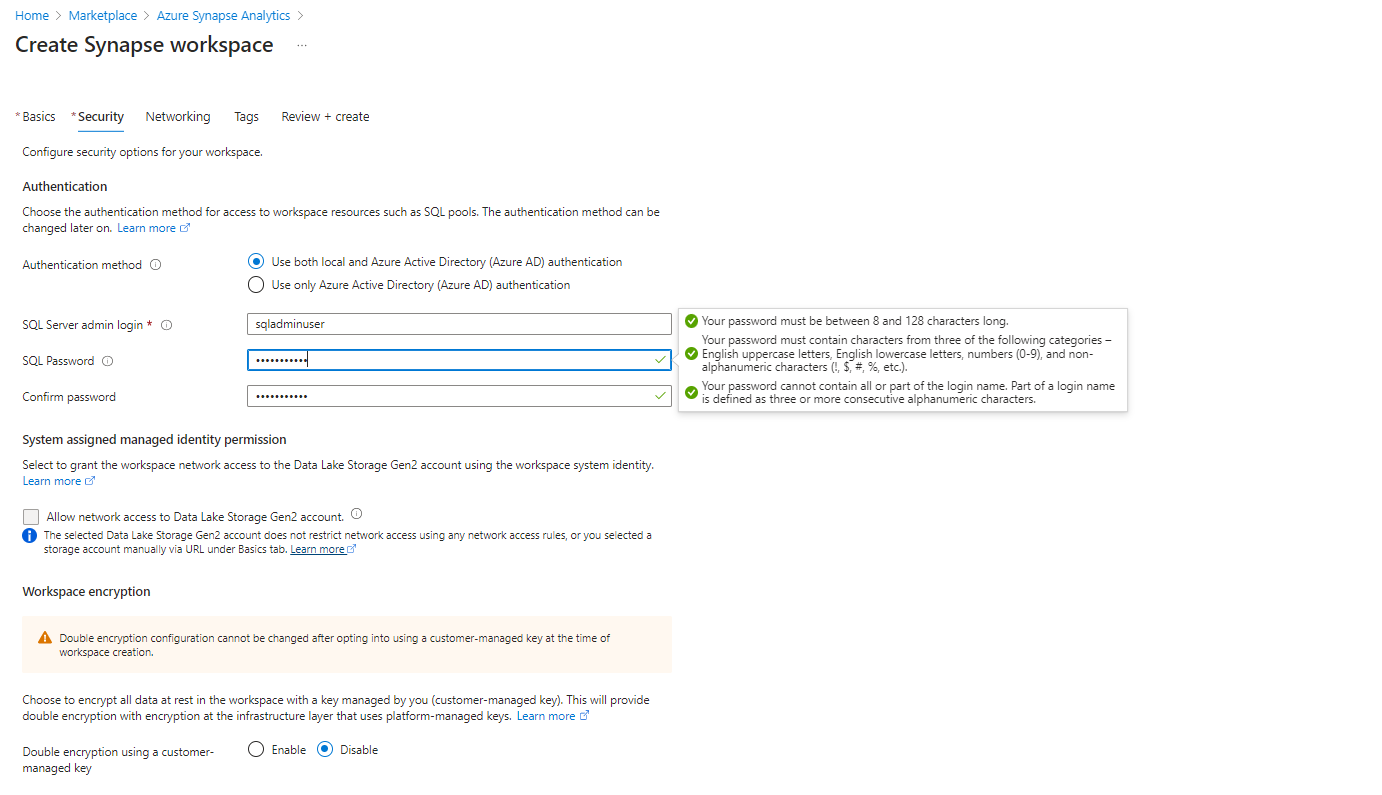
**4. Create A Synapse Analytics workspace, or use one you already have created.**

Databasename: dwpostgre

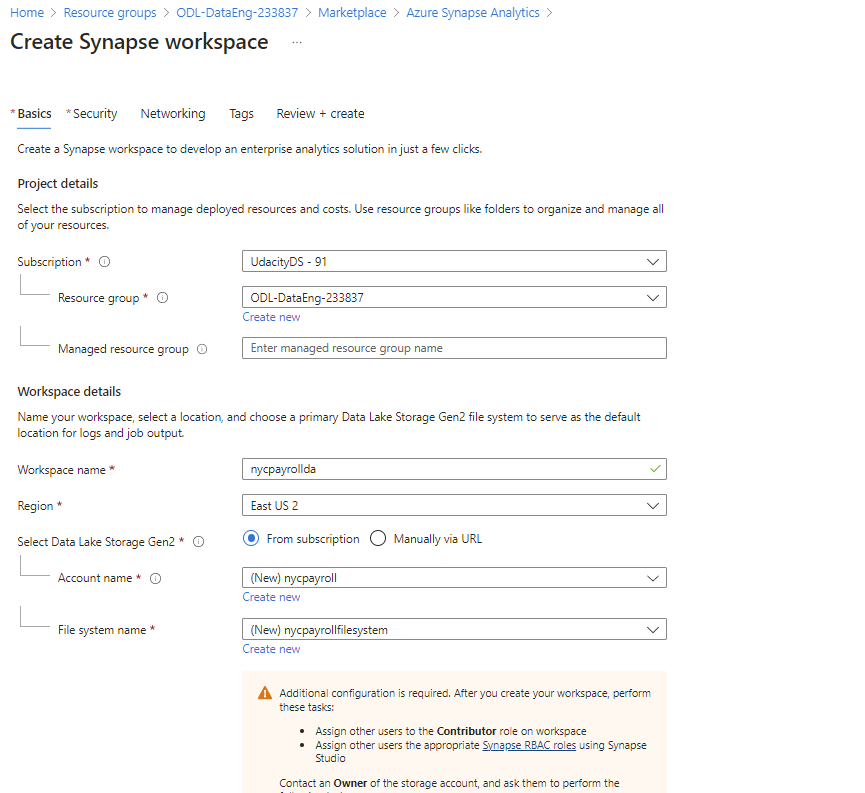






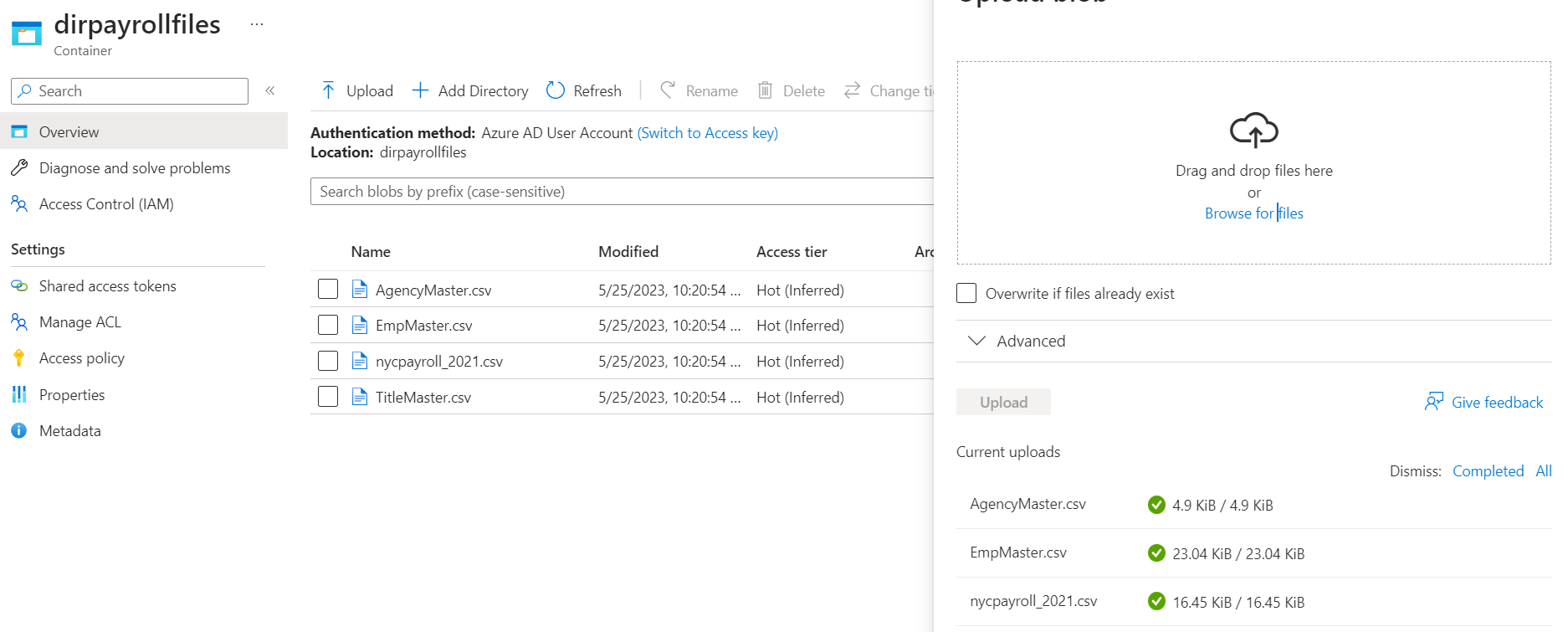


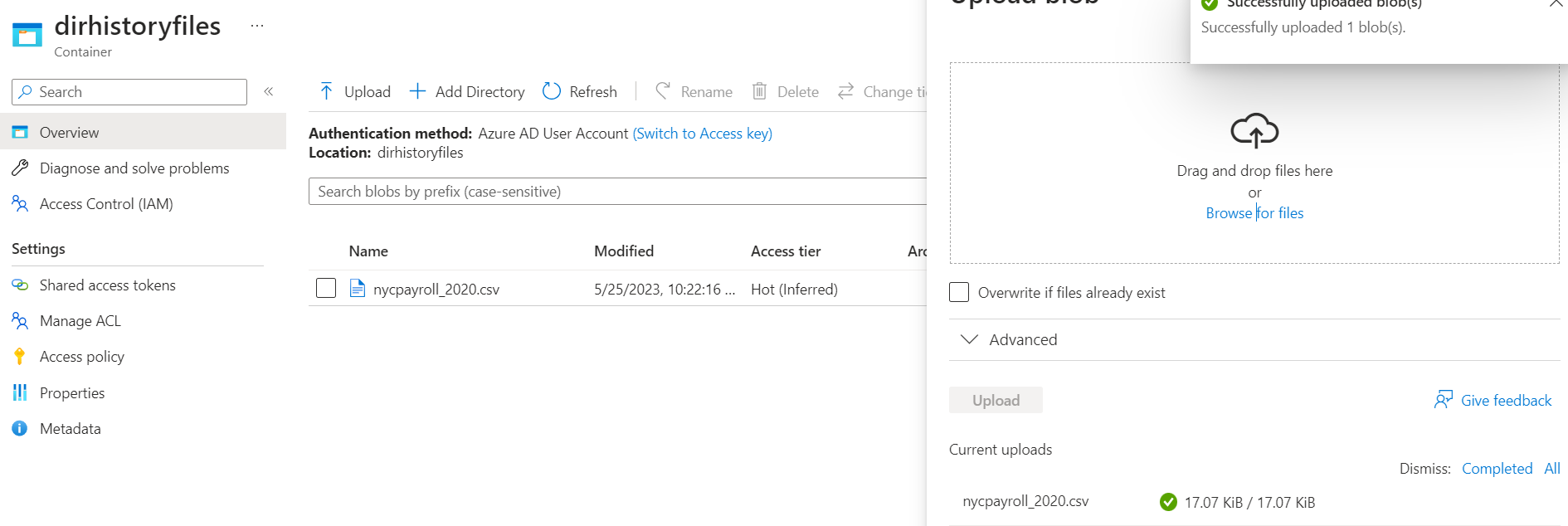
nycpayroll@1 : pw



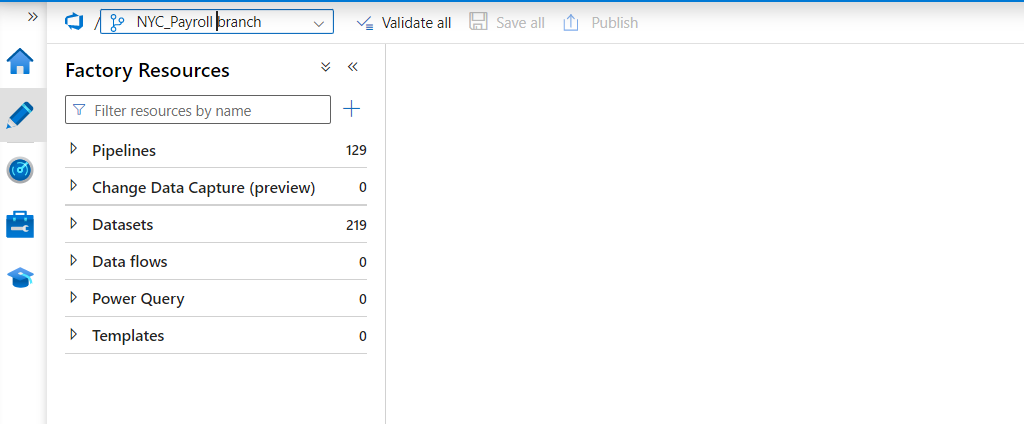
### Step 1: Prepare the Data Infrastructure

**1.Create the data lake and upload data**

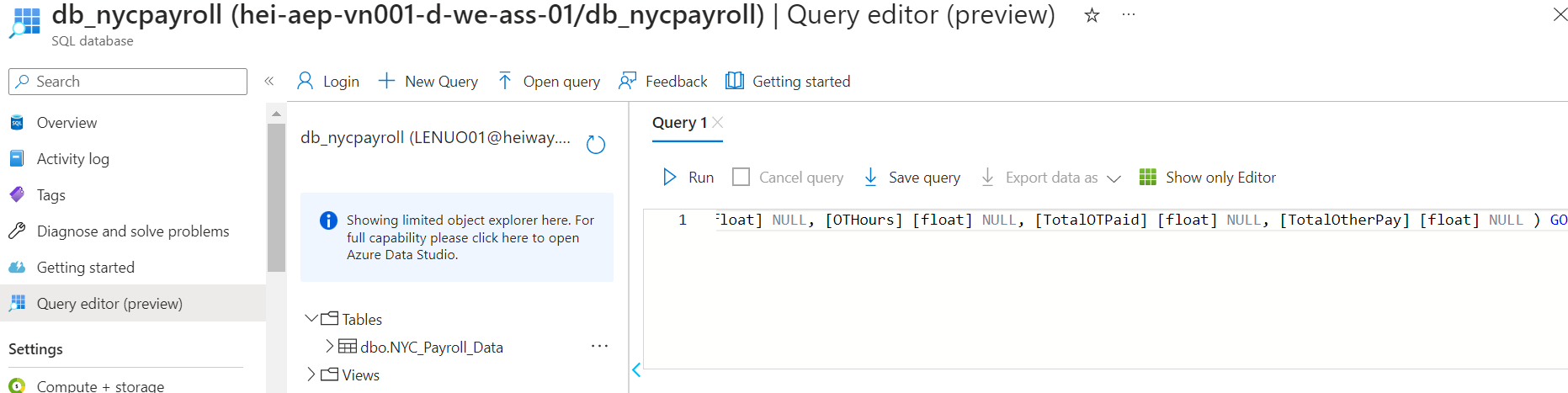


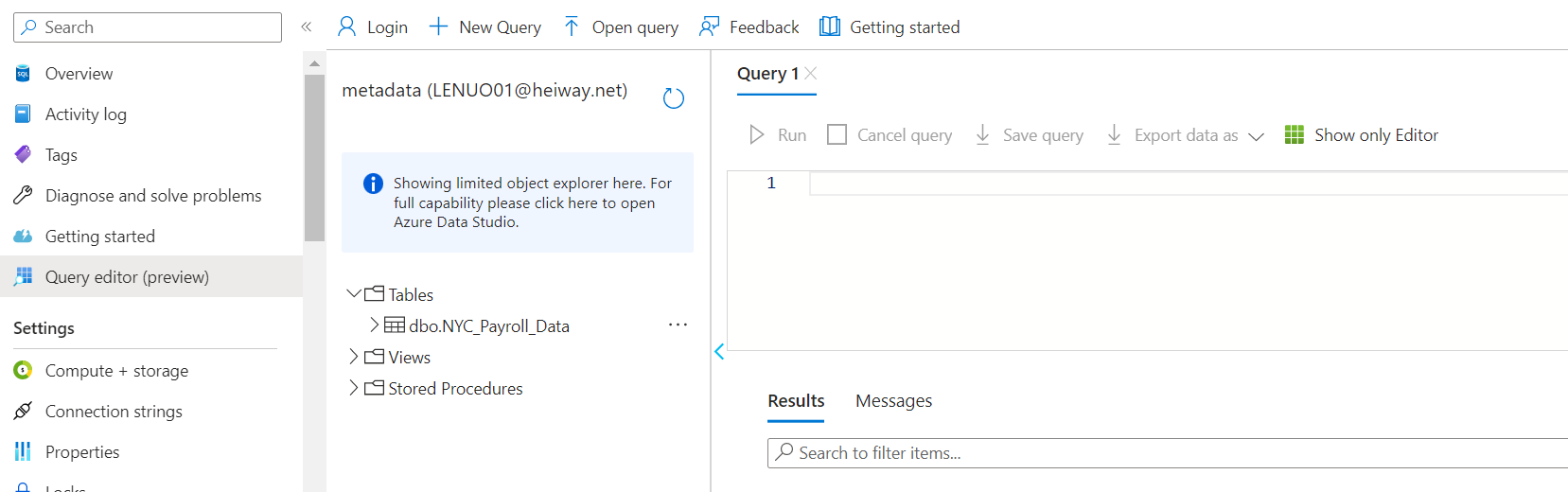


**2. Create an Azure Data Factory Resource**

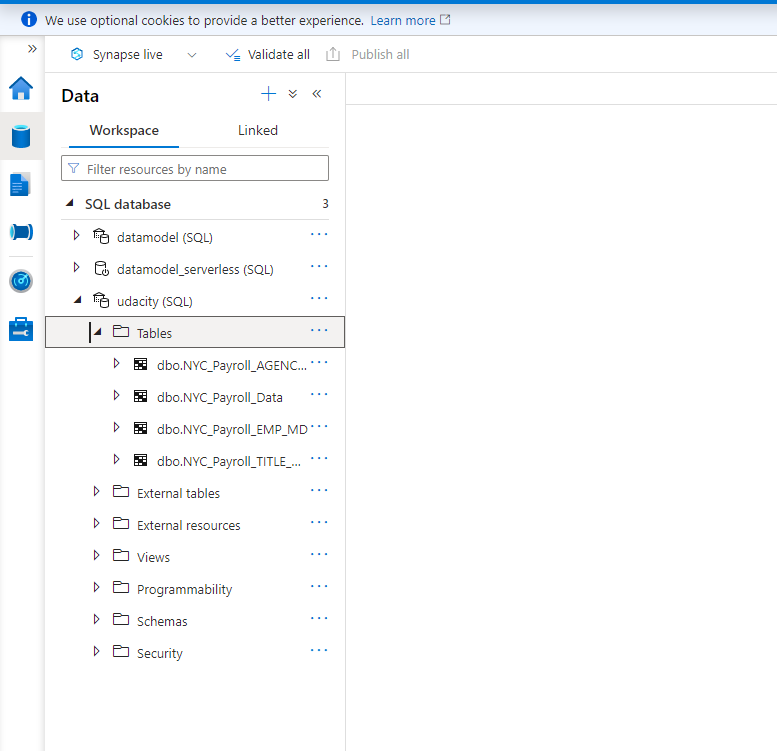


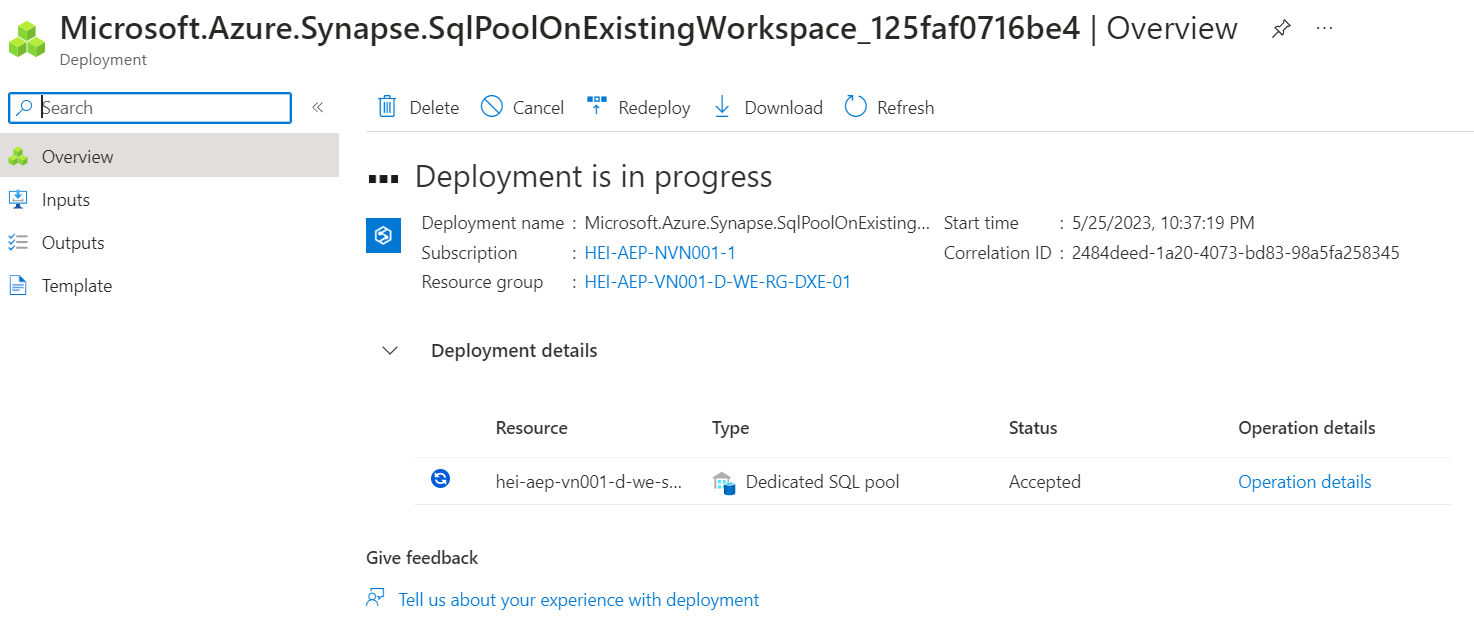
**3. Create a SQL Database to store the current year of the payroll data**

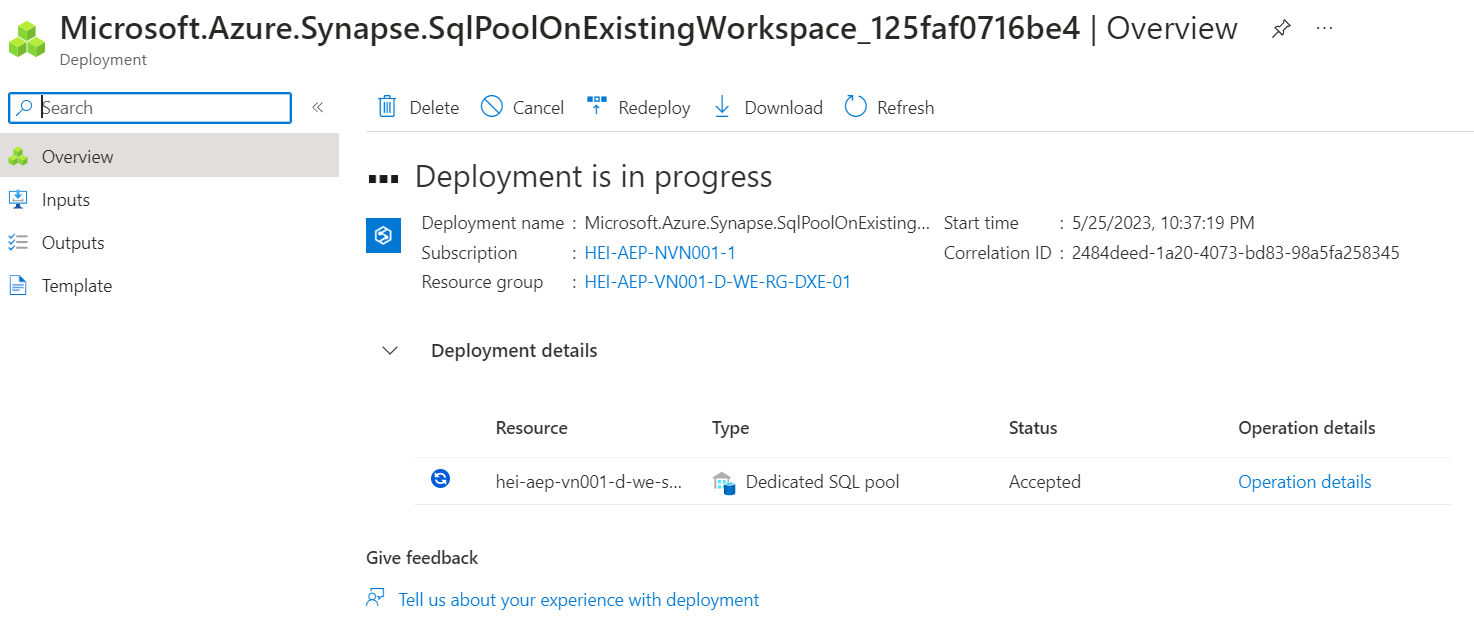




**4. Create A Synapse Analytics workspace, or use one you already have created.**

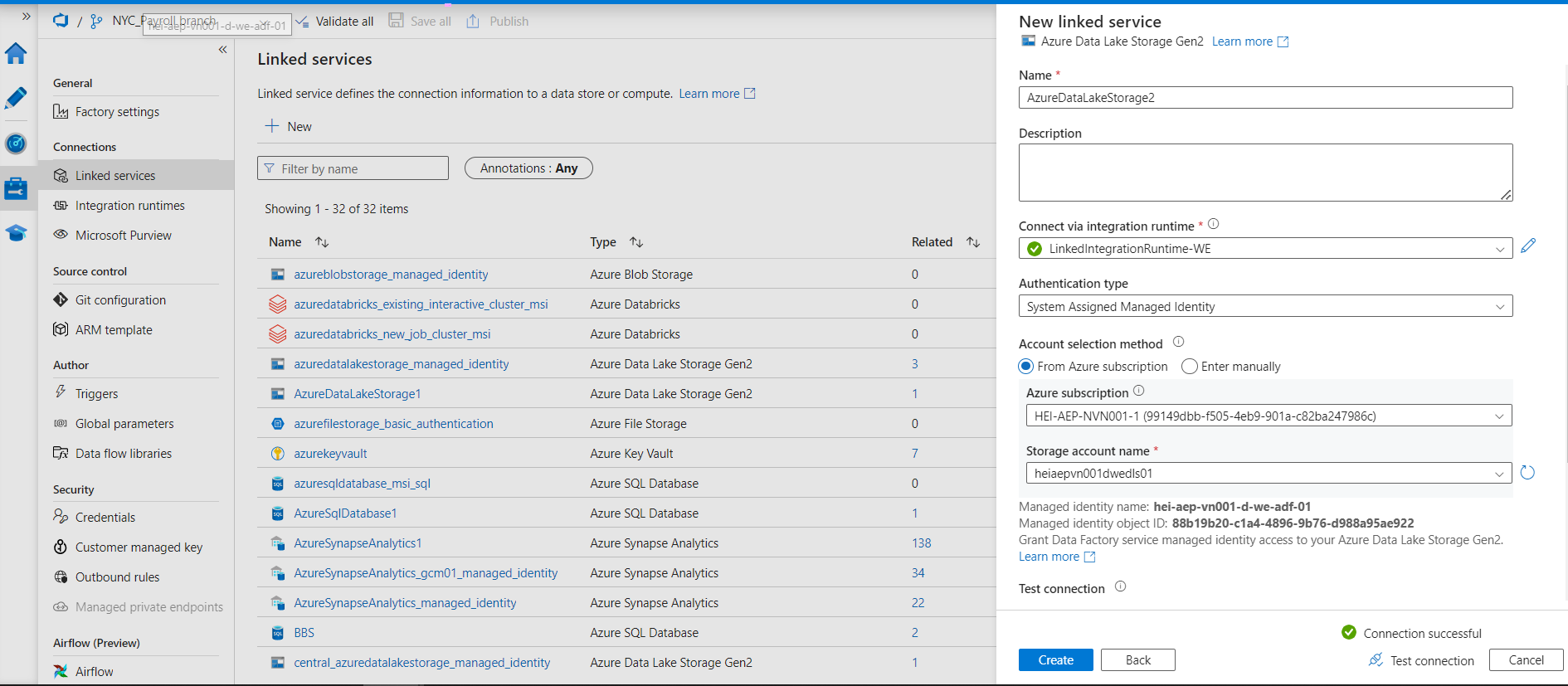




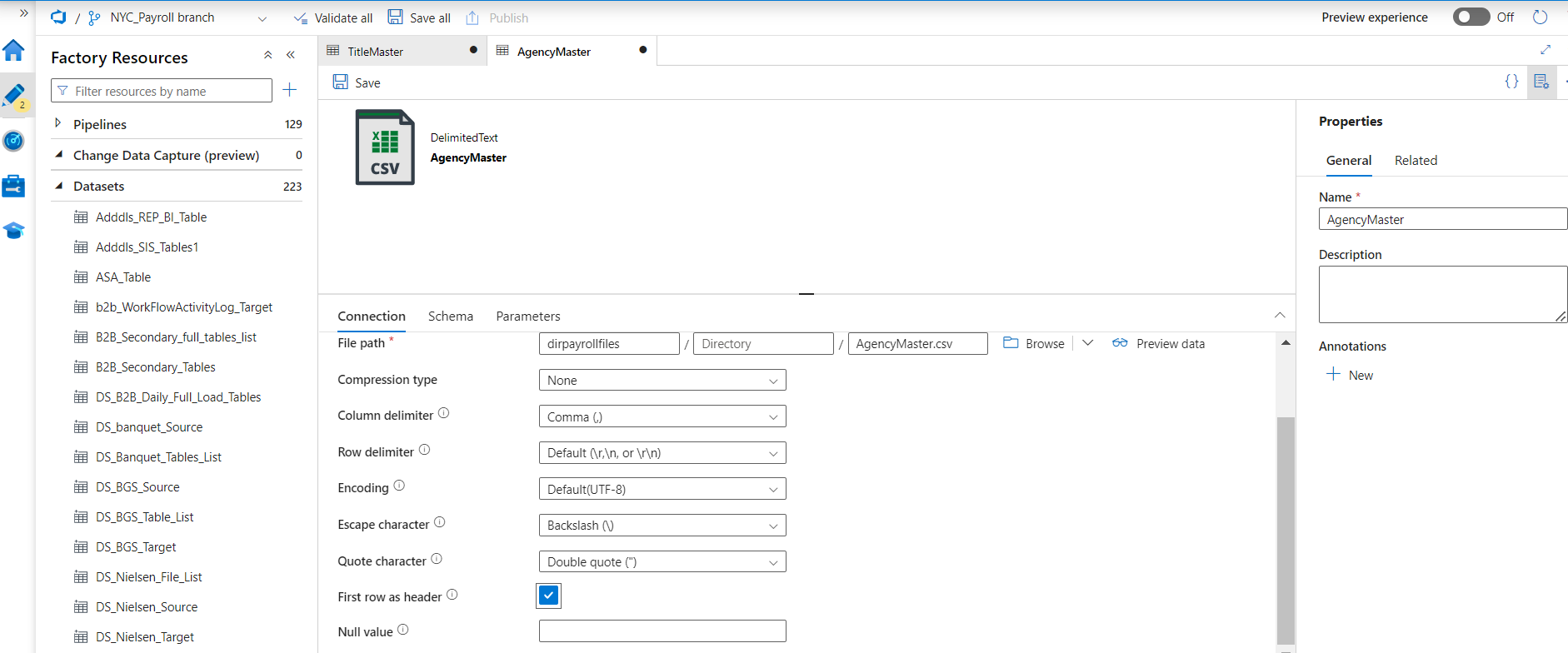


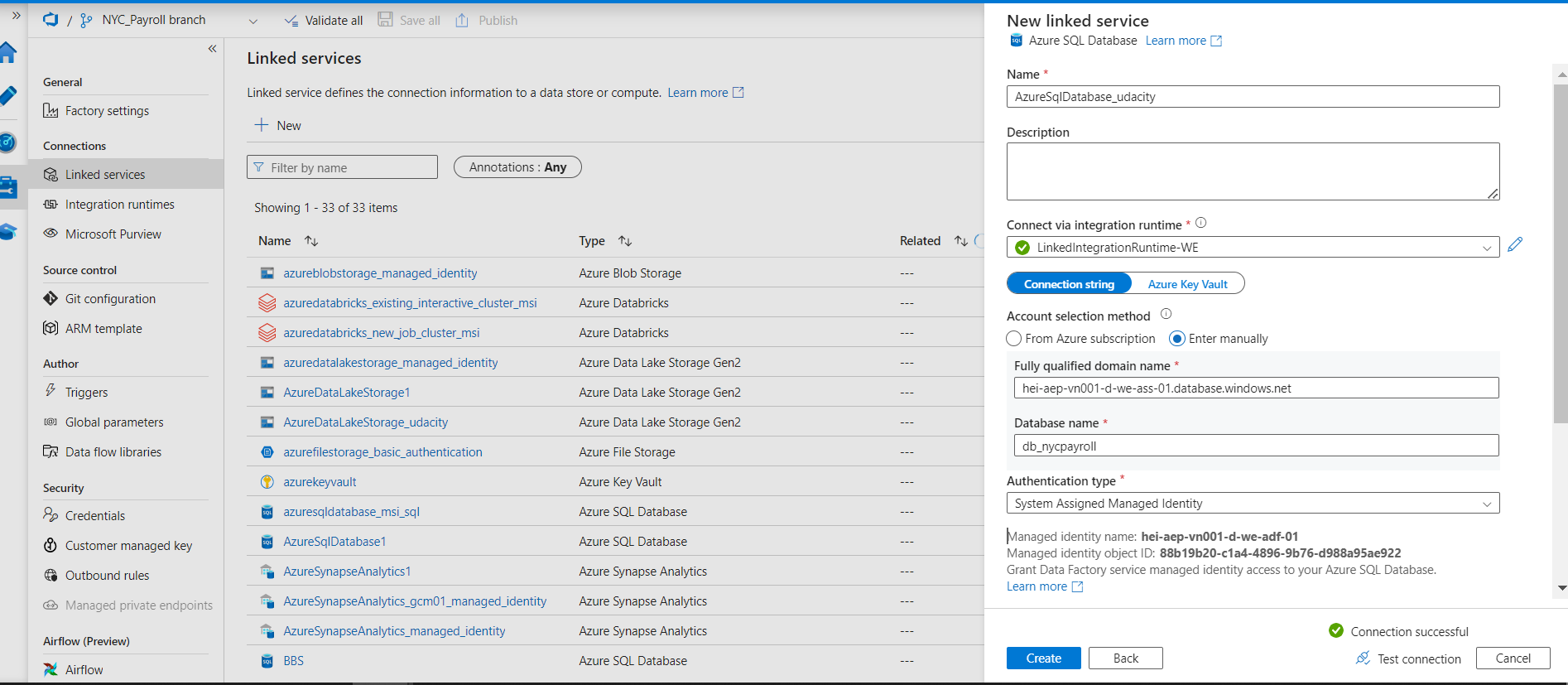
# **Instructions: Create Linked Services**

### Step 2: Create Linked Services

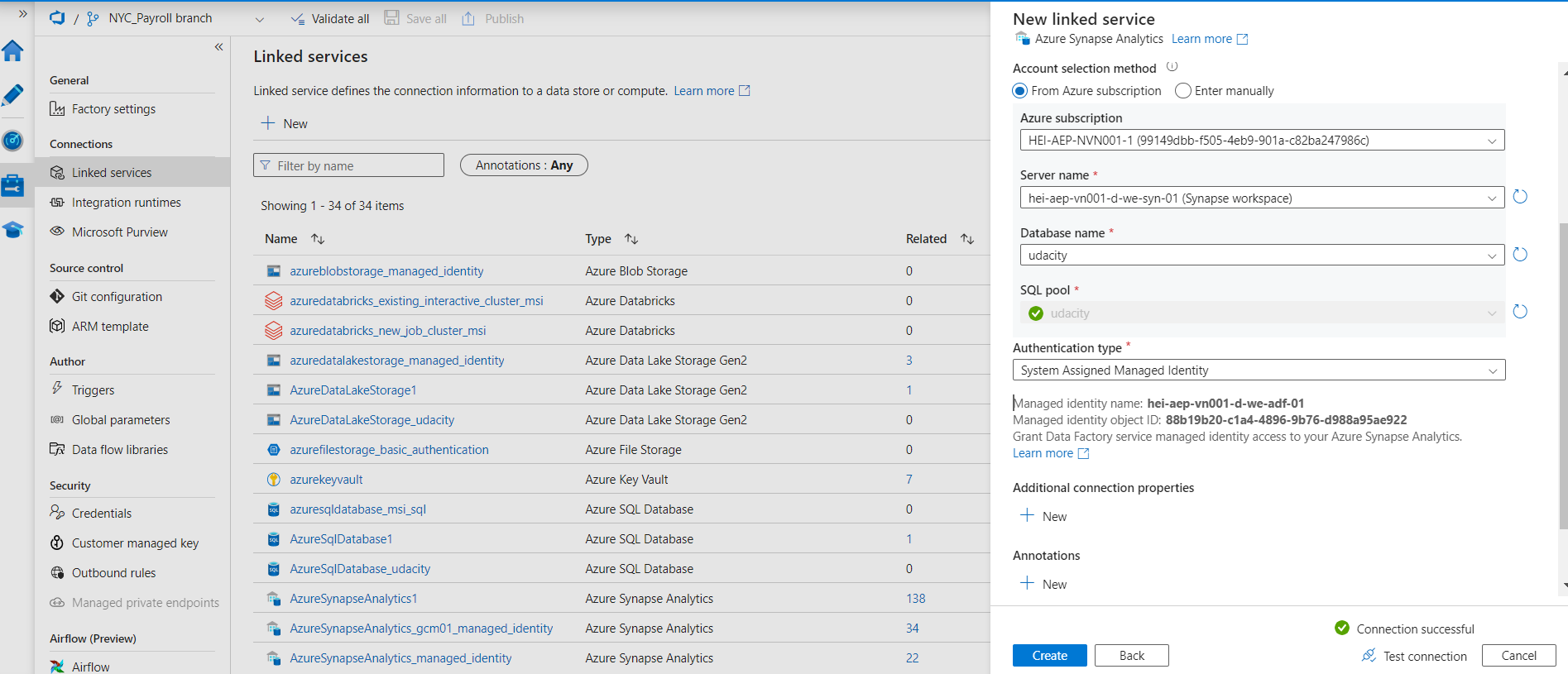


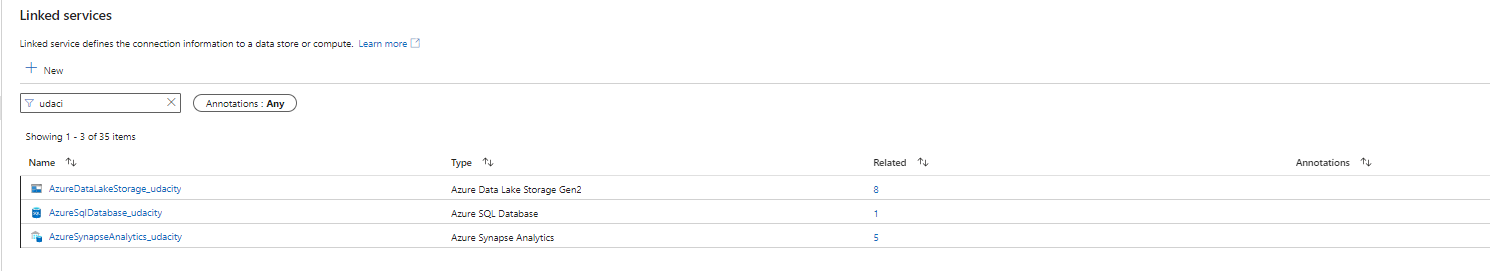
**2.Create a Linked Service to SQL Database that has the current (2021) data**





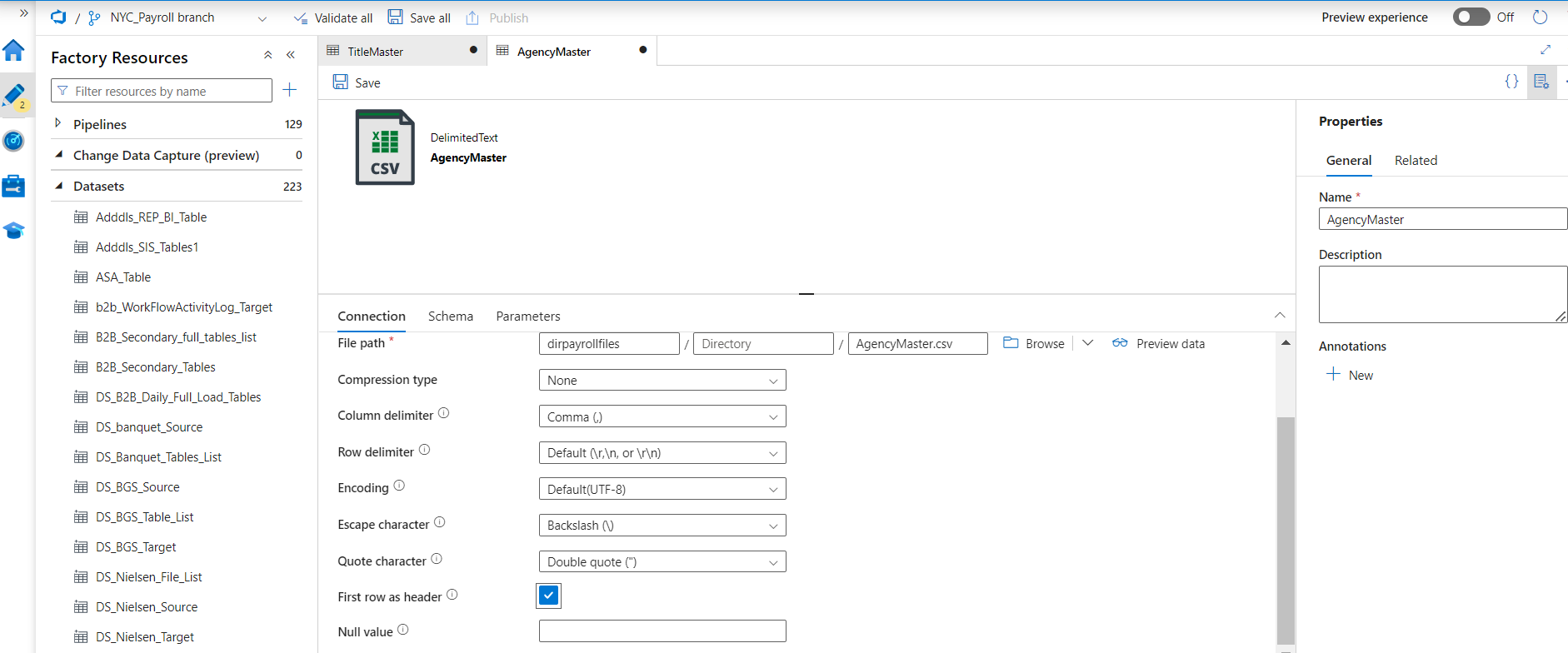
**3. Create a Linked Service for Synapse Analytics**

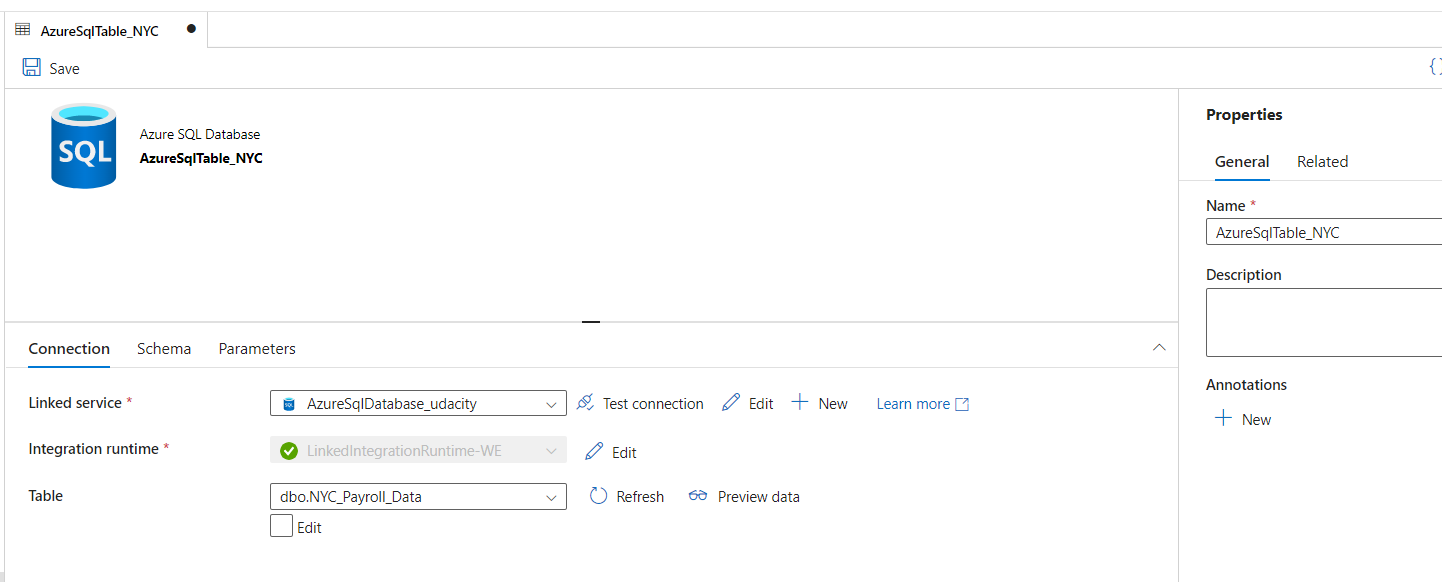


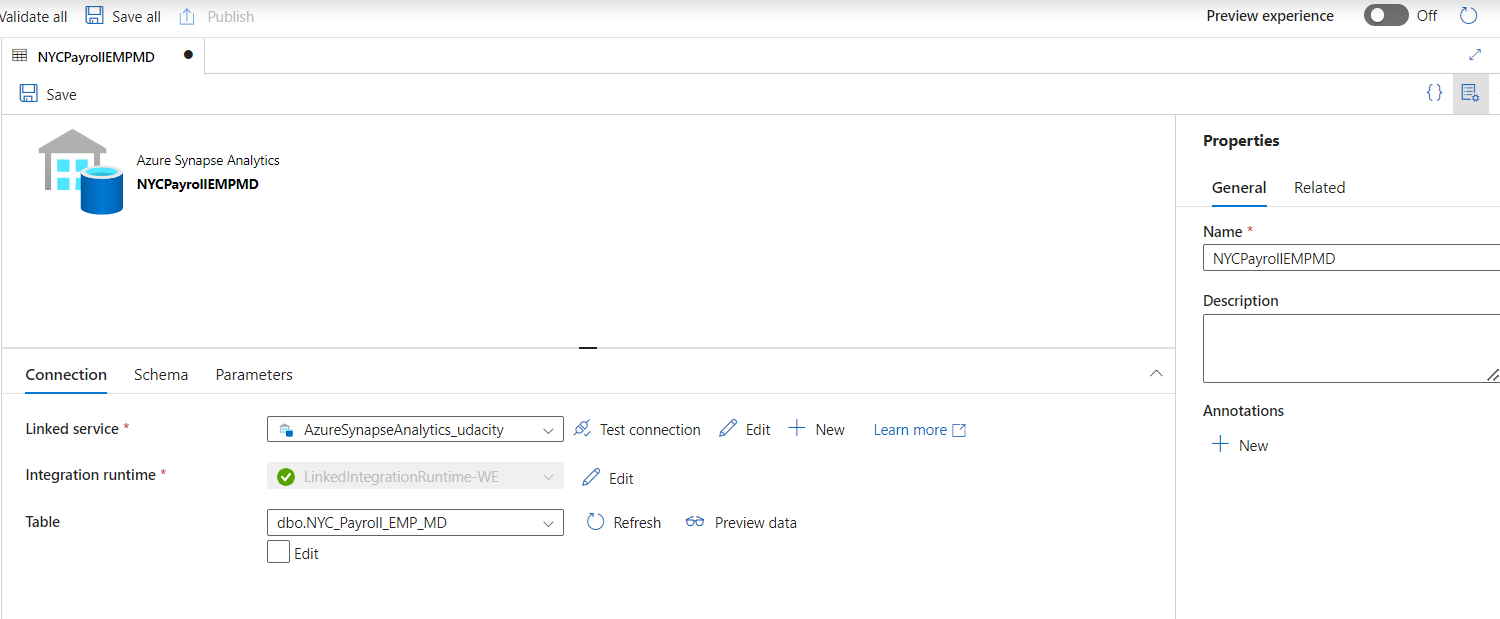


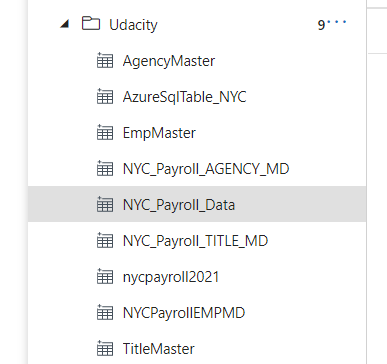
# **Instructions: Create Datasets**

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



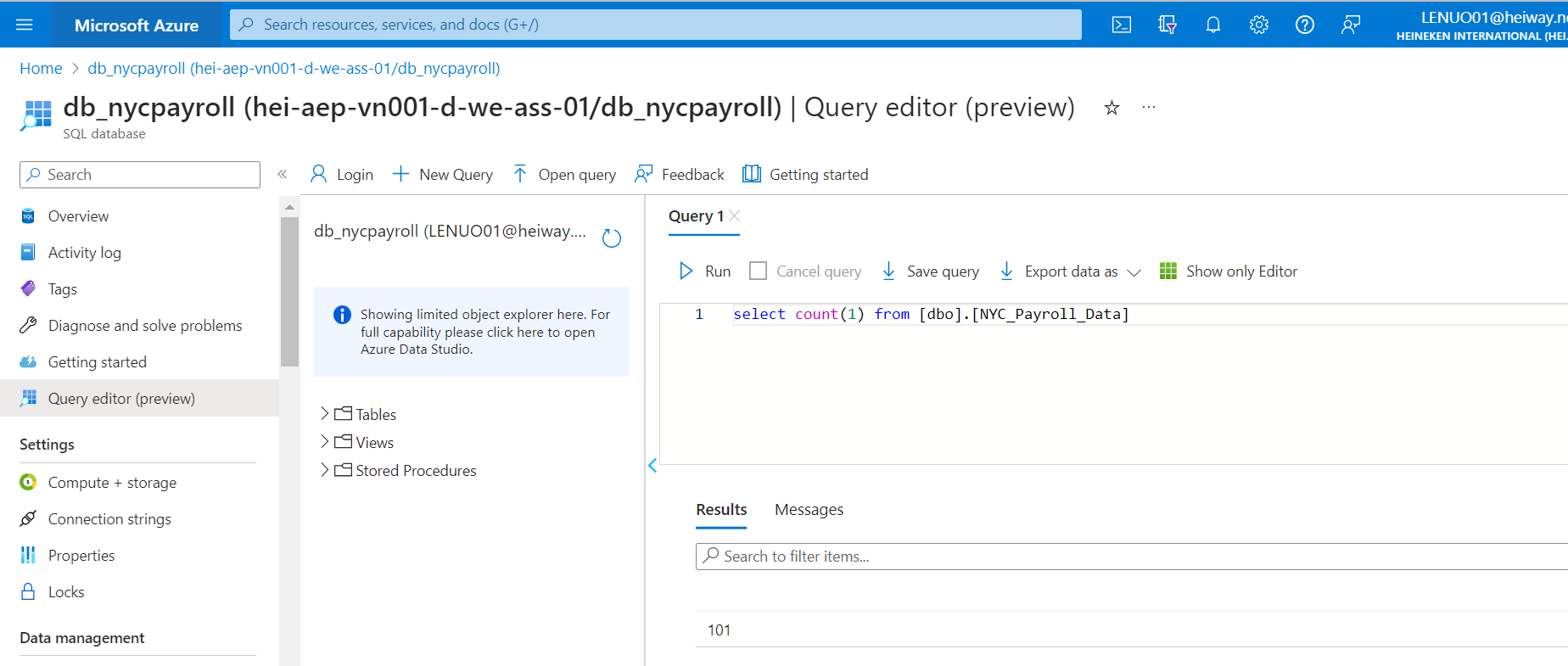




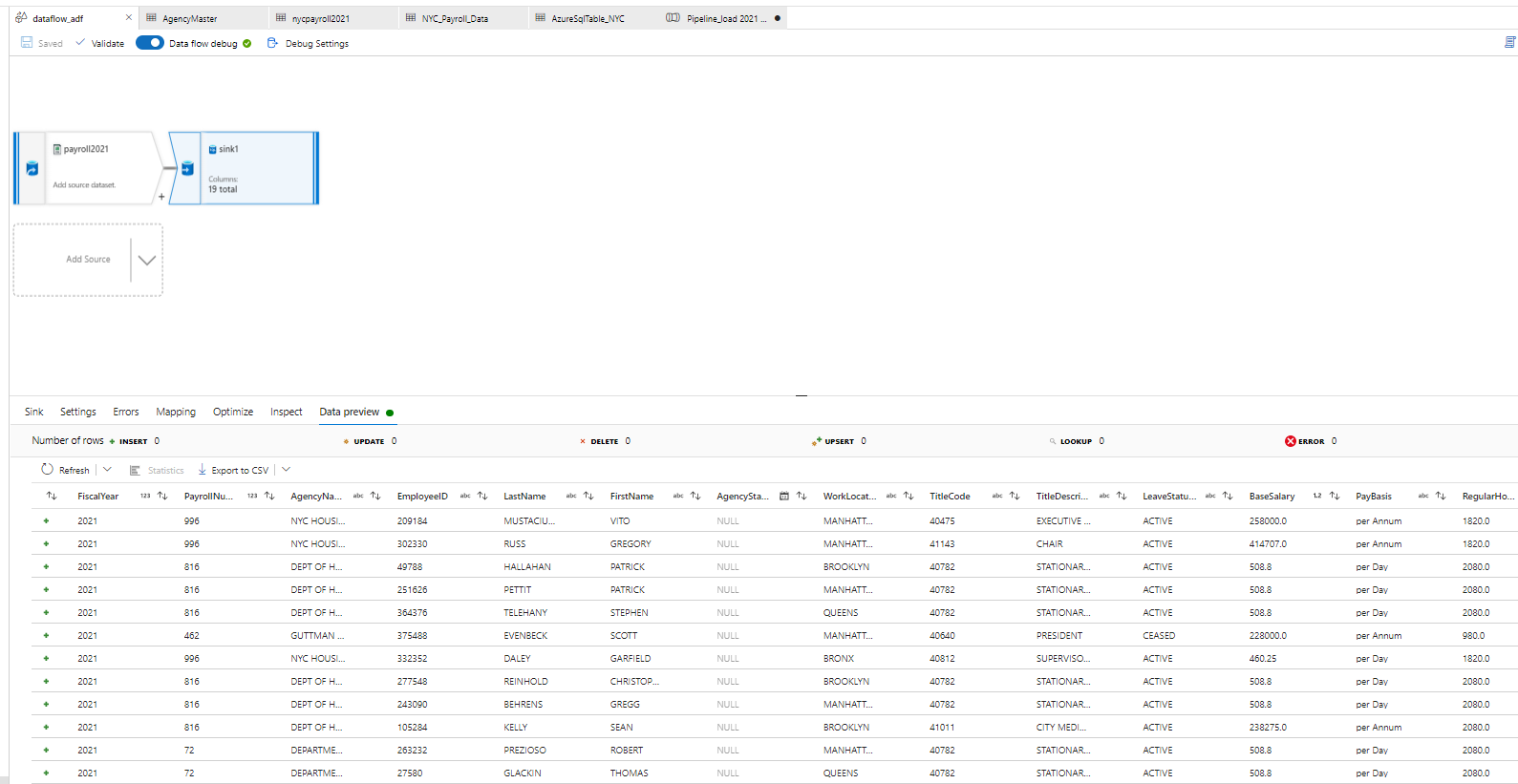


# **Instructions: Create Data Flows and Pipelines**

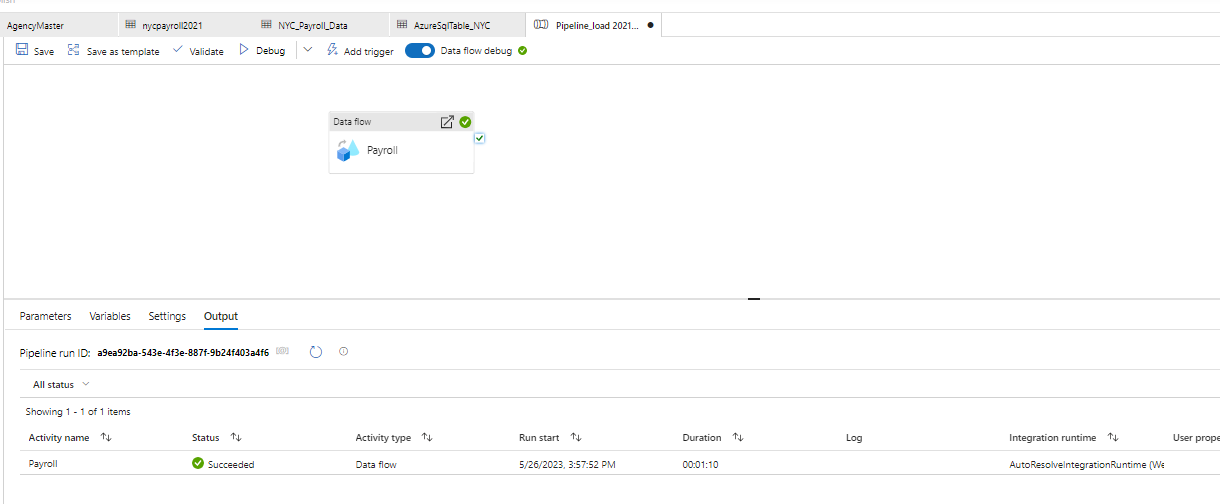
### Step 4: Create Data Flows



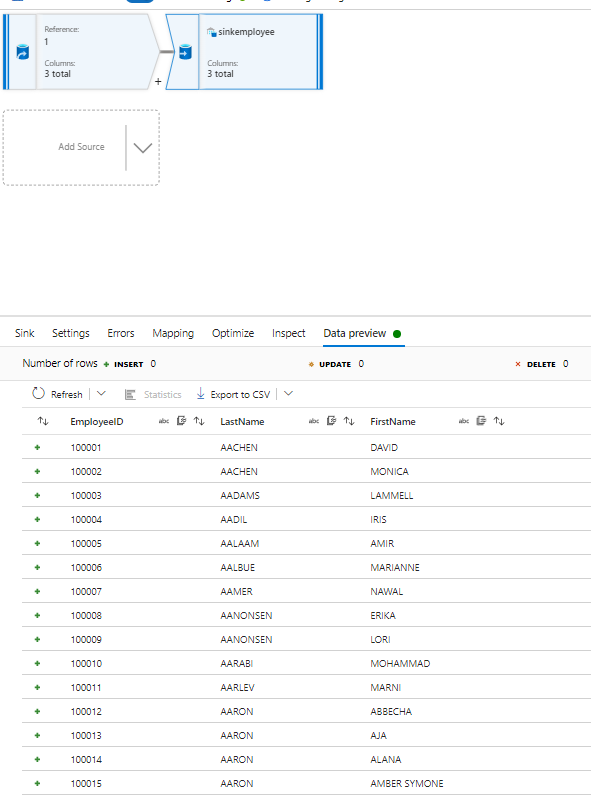
**1.In Azure Data Factory, create the data flow to load 2021 Payroll Data to SQL DB transaction table (in the future NYC will load all the transaction data into this table).**

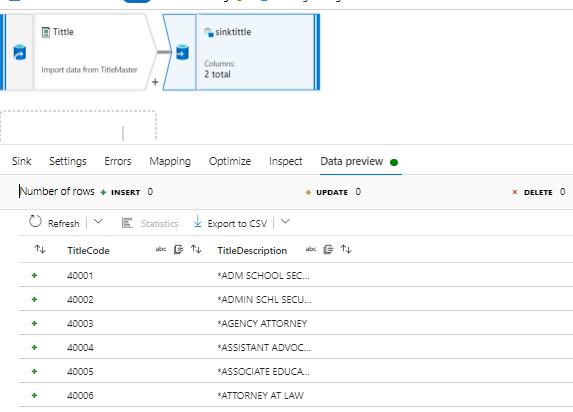


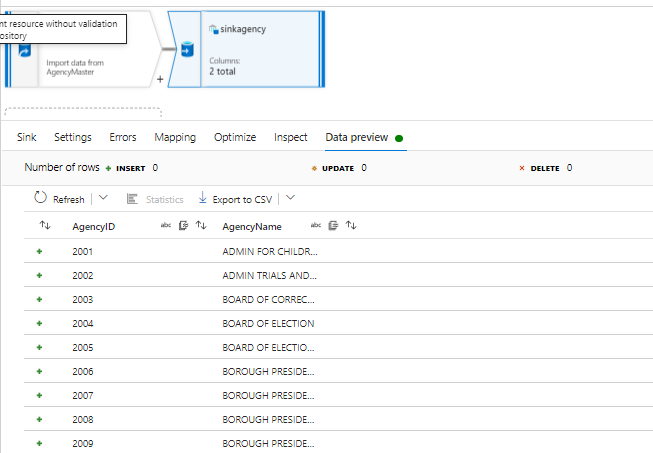
**2.Create Pipeline to load 2021 Payroll data into transaction table in the SQL DB**



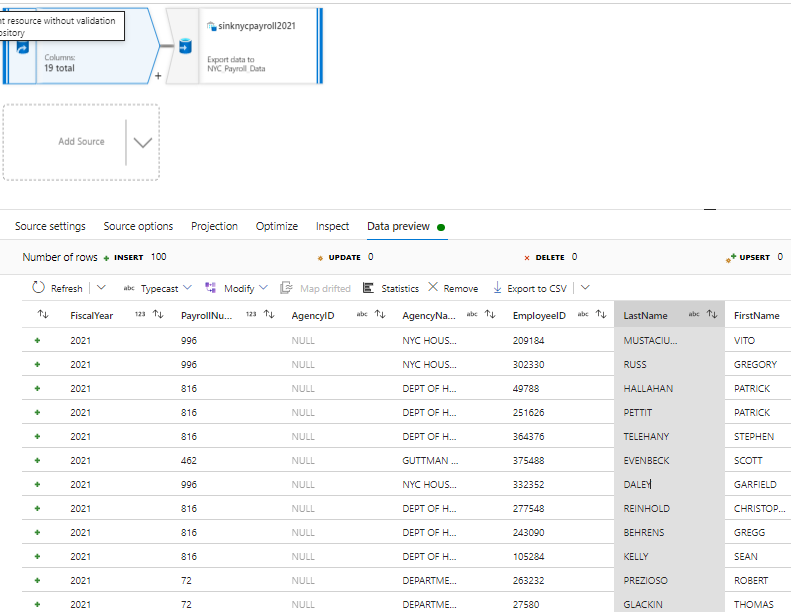
**3. Create data flows to load the data from the data lake files into the Synapse Analytics data tables**



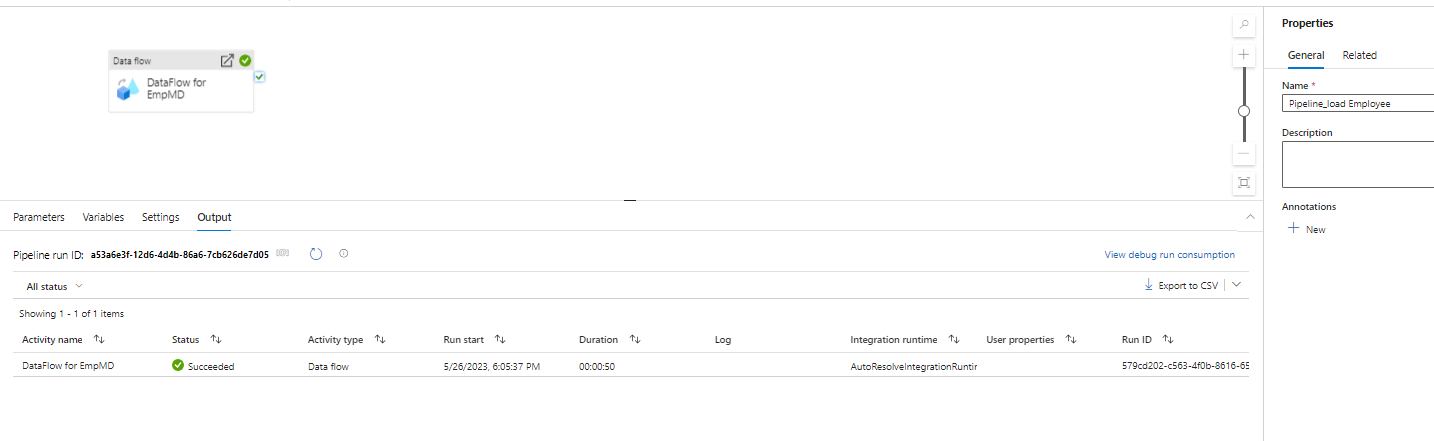


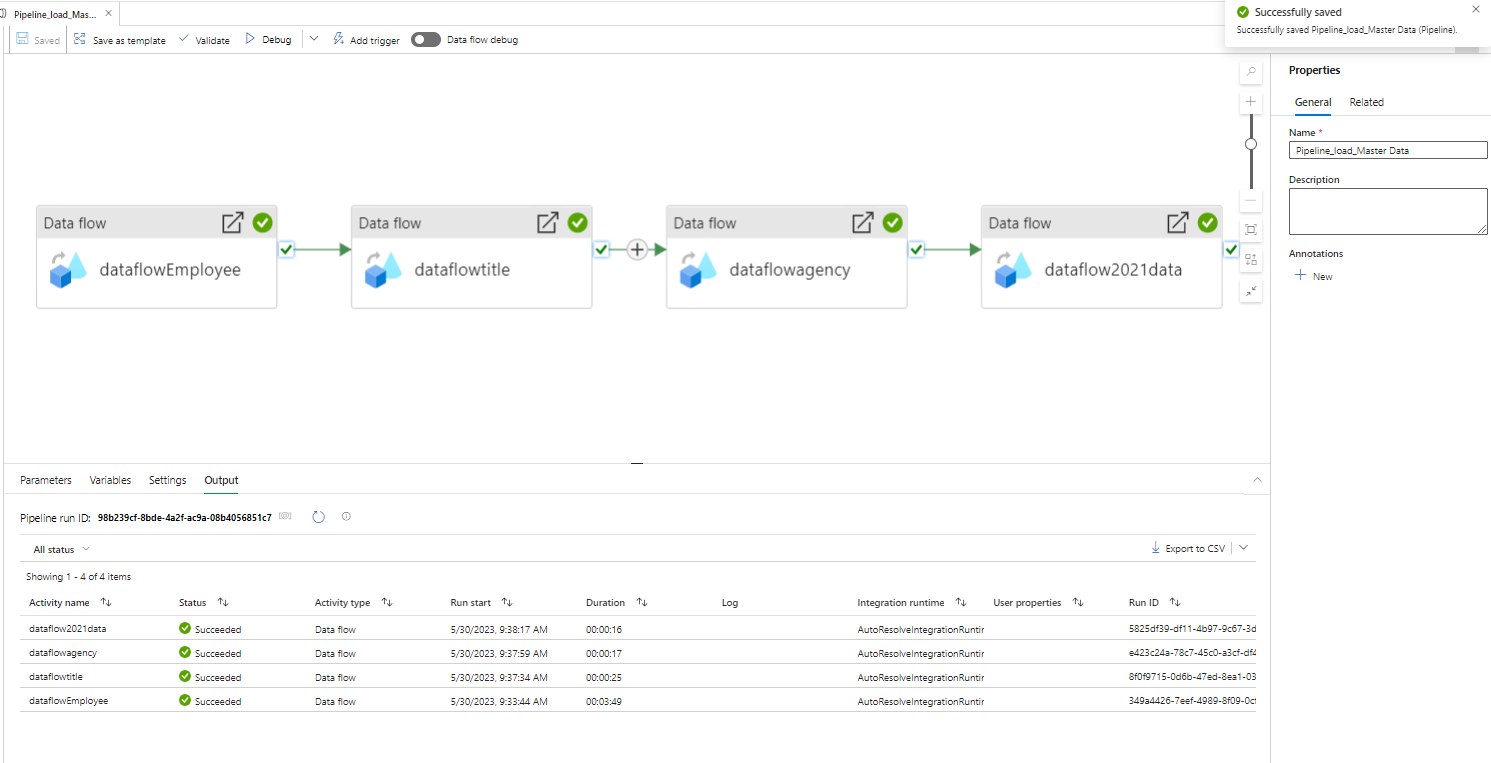


**4. Create a data flow to load 2021 data from SQL DB to Synapse Analytics**

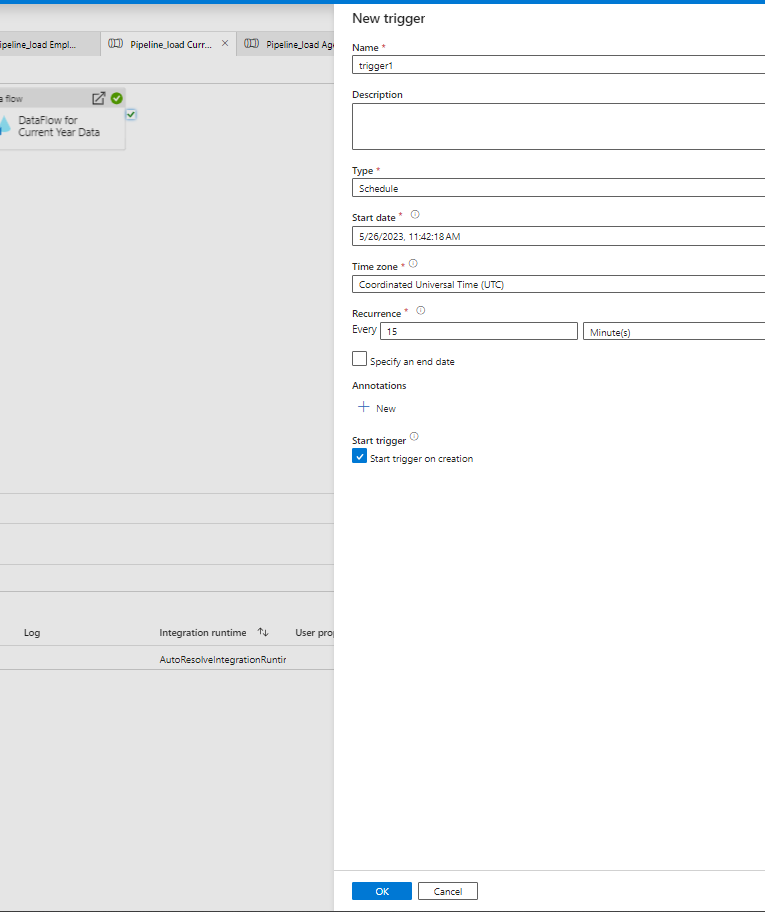


**5. Create pipelines for Employee, Title, Agency, and year 2021 Payroll transaction data to Synapse Analytics containing the data flows.**





**6. Trigger and monitor the Pipelines**



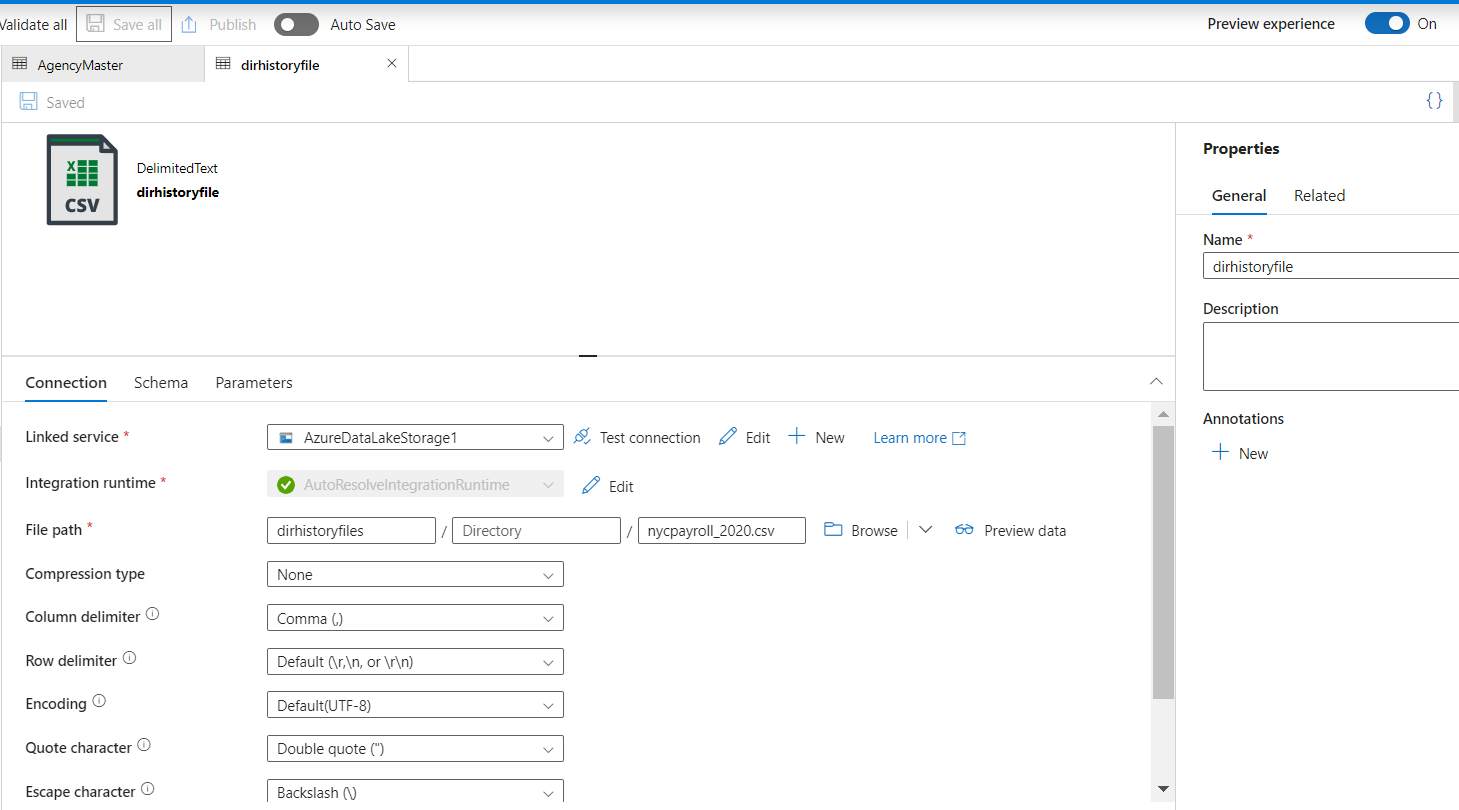
# **Instructions: Aggregate Data Flow**

### Step 5: Data Aggregation and Parameterization

**1.Create a Summary table in Synapse with the following SQL script and create a dataset named table\_synapse\_nycpayroll\_summary**

CREATE TABLE [dbo].[NYC\_Payroll\_Summary]( [FiscalYear] [int] NULL, [AgencyName] [varchar](50) NULL, [TotalPaid] [float] NULL )

**2.Create a new dataset for the Azure Data Lake Gen2 folder that contains the historical files.**



**3.Create new data flow and name it Dataflow Aggregate Data**

* Create a data flow level parameter for Fiscal Year
* Add first Source for table\_sqldb\_nyc\_payroll\_data table
* Add second Source for the Azure Data Lake history folder

**4.Create a new Union activity in the data flow and Union with history files**

**5.Add a Filter activity after Union**

* In Expression Builder, enter toInteger(FiscalYear) >= $dataflow\_param\_fiscalyear

**6.Derive a new TotalPaid column**

* In Expression Builder, enter RegularGrossPaid + TotalOTPaid+TotalOtherPay

**7.Add an Aggregate activity to the data flow next to the TotalPaid activity**

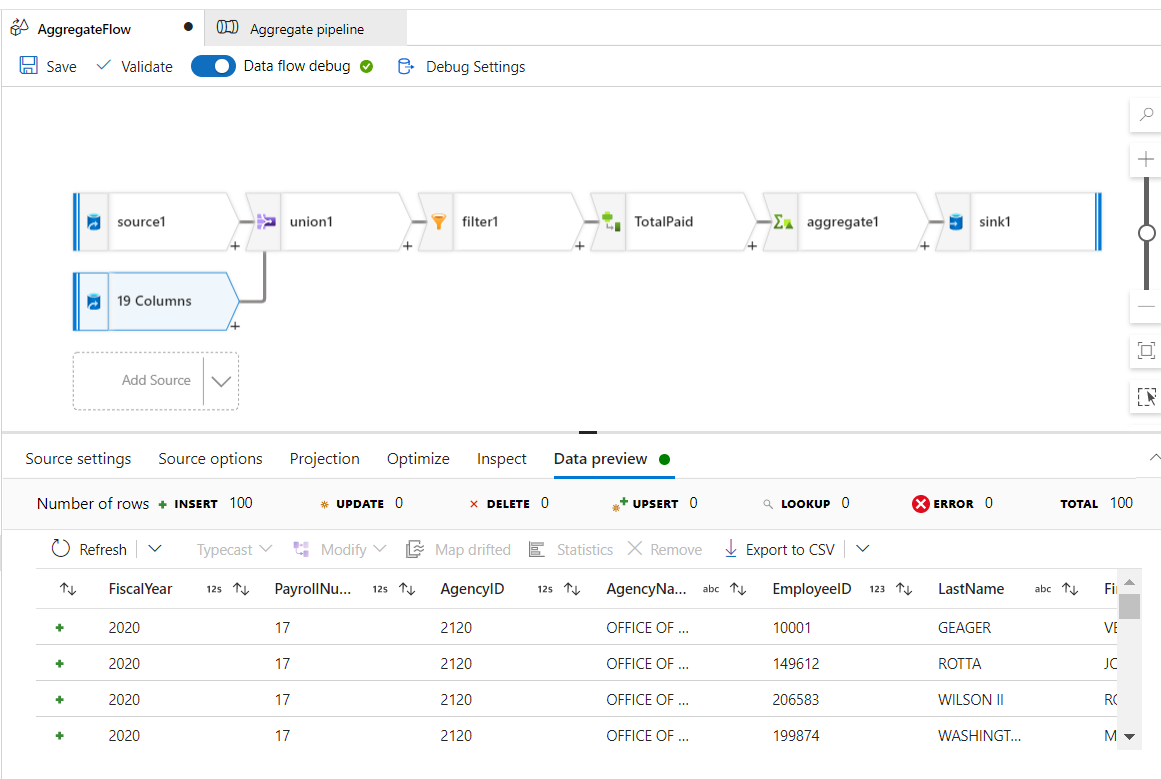
* Under Group By, Select AgencyName and Fiscal Year

**8.Add a Sink activity to the Data Flow**

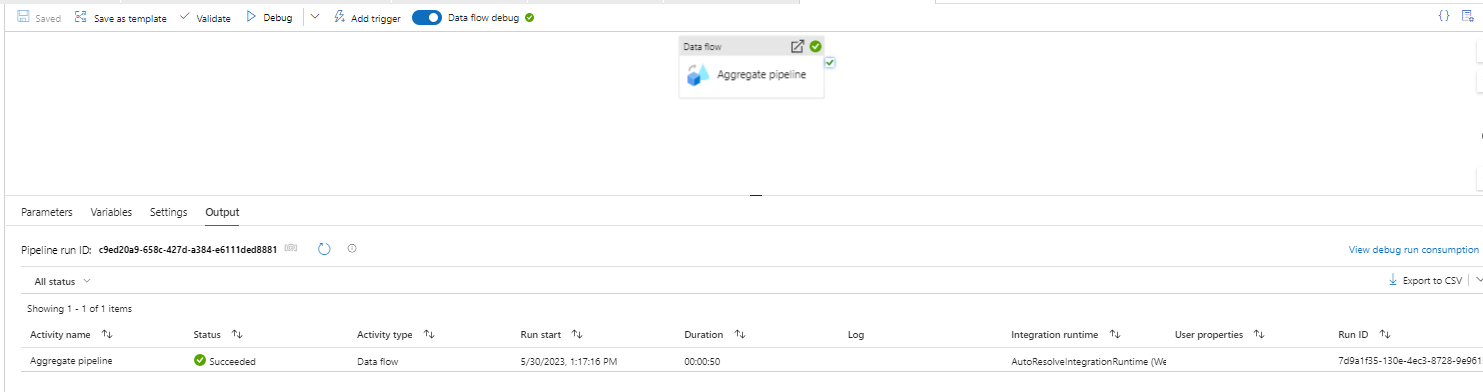
* Select the dataset to target (sink) the data into the Synapse Analytics Payroll Summary table.
* In Settings, select Truncate Table

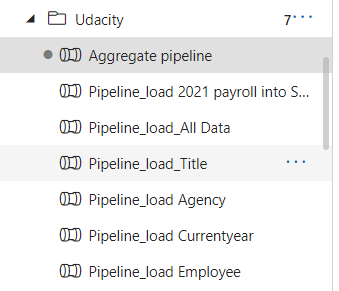
**9.Create a new Pipeline and add the Aggregate data flow**

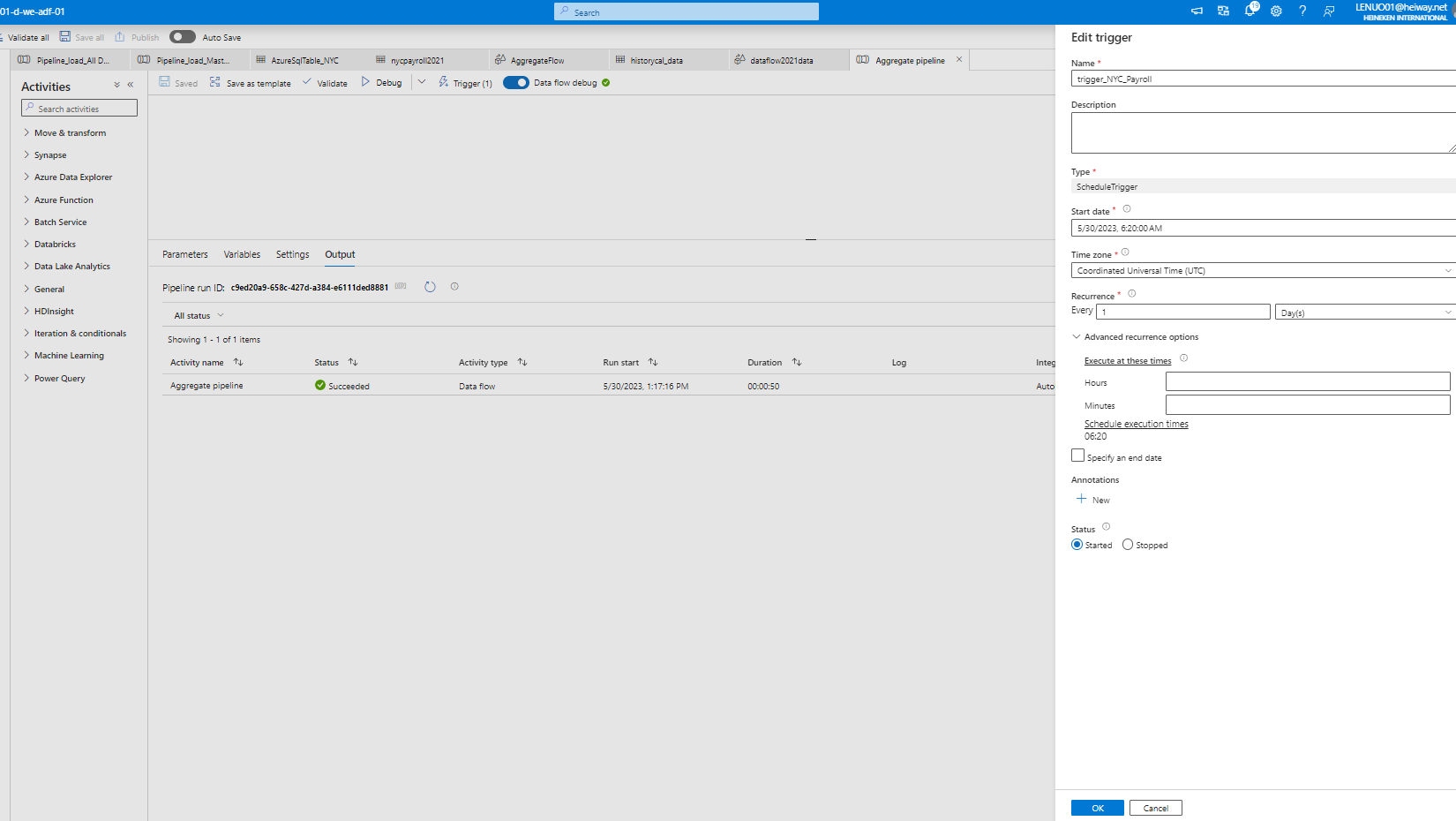
* Create a new Global Parameter (This will be the Parameter at the global pipeline level that will be passed on to the data flow
* In Parameters, select Pipeline Expression
* Choose the parameter created at the Pipeline level



**10.Validate, Publish and Trigger the pipeline. Enter the desired value for the parameter.**







# **Instructions: Connect Project to Github and Submit**

### Step 6: Connect your Project to Github

In this step, you'll connect Azure Data Factory to Github

* Login to your Github account and create a new Repo in Github
* Connect Azure Data Factory to Github
* Select your Github repository in Azure Data Factory
* Publish all objects to the repository in Azure Data Factory

Because Udacity accounts are not allowed to create all sources requitement and I can’t wait for long so I have use business account to do this project so I can’t clone Business Workspace to Git Hub due to conflict policy so I would like to send zip file only.

Could you please check and give me your feedback?