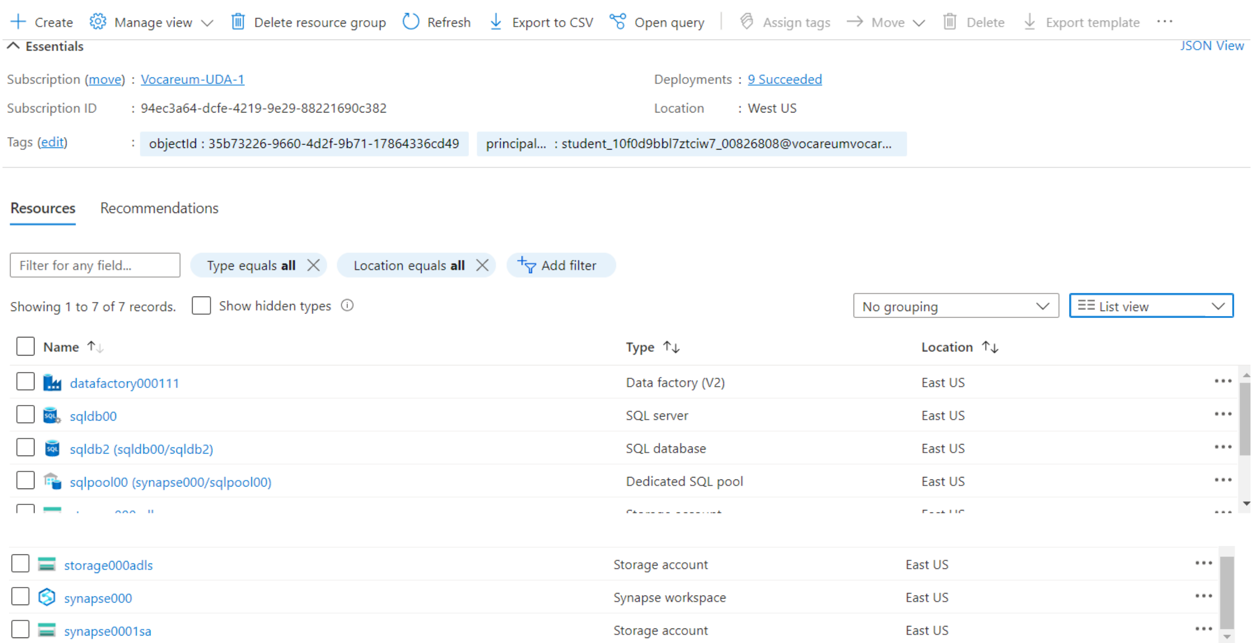
Contents

Azure resource

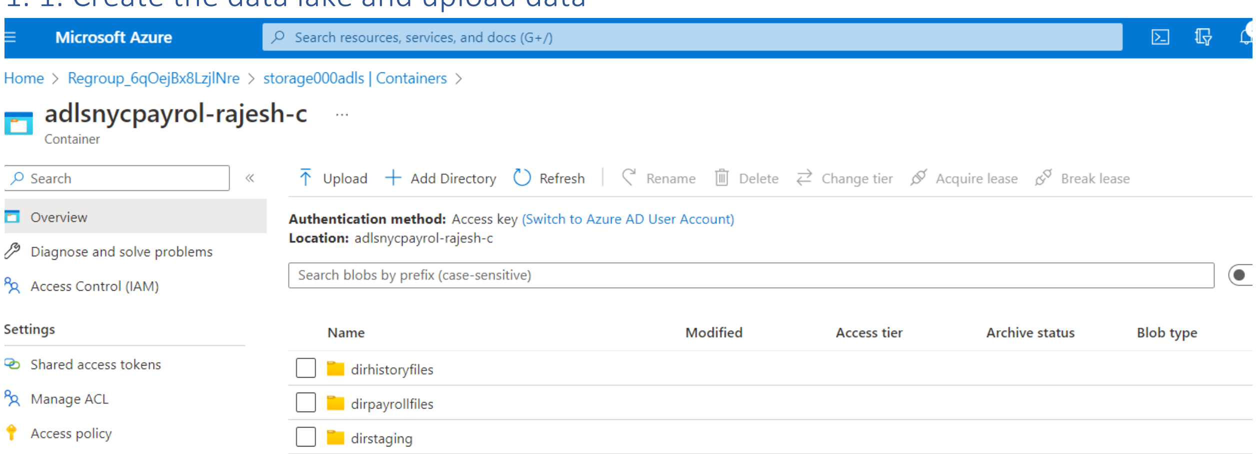
- Azure Data Lake Gen2 (Storage account with Hierarchical Namespaces checkbox checked when creating)
- Azure SQL DB
- Azure Data Factory
- Azure Synapse Analytics



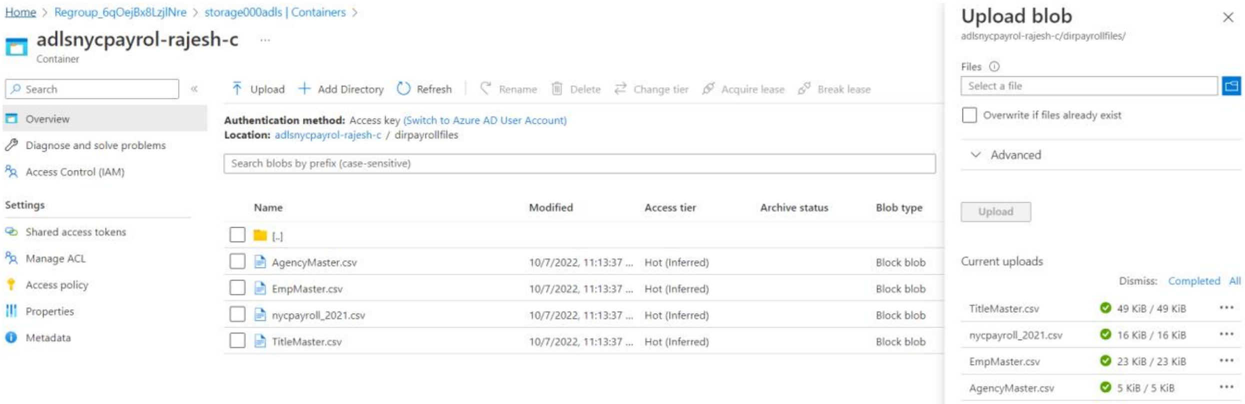
Step 1: Prepare the Data Infrastructure

Setup Data and Resources in Azure

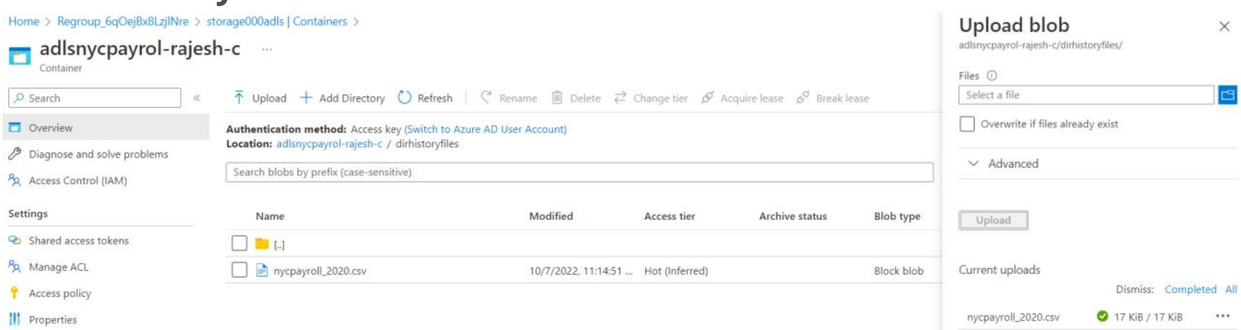
1. 1. Create the data lake and upload data



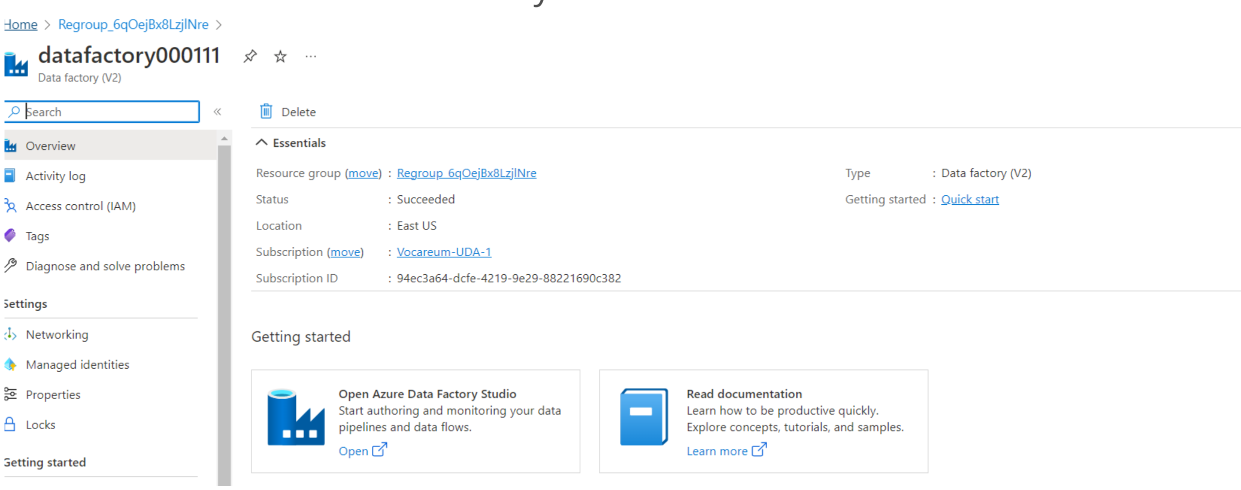
1.2. Upload these files from the [project data]() to the **dirpayrollfiles** folder



1.3. Upload this file (historical data) from the [project data]() to the **dirhistoryfiles** folder

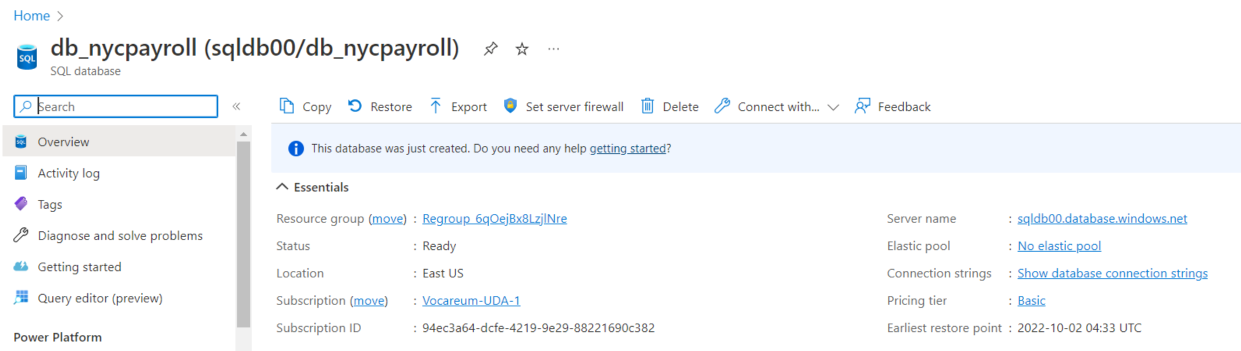


1. 2. Create an Azure Data Factory Resource

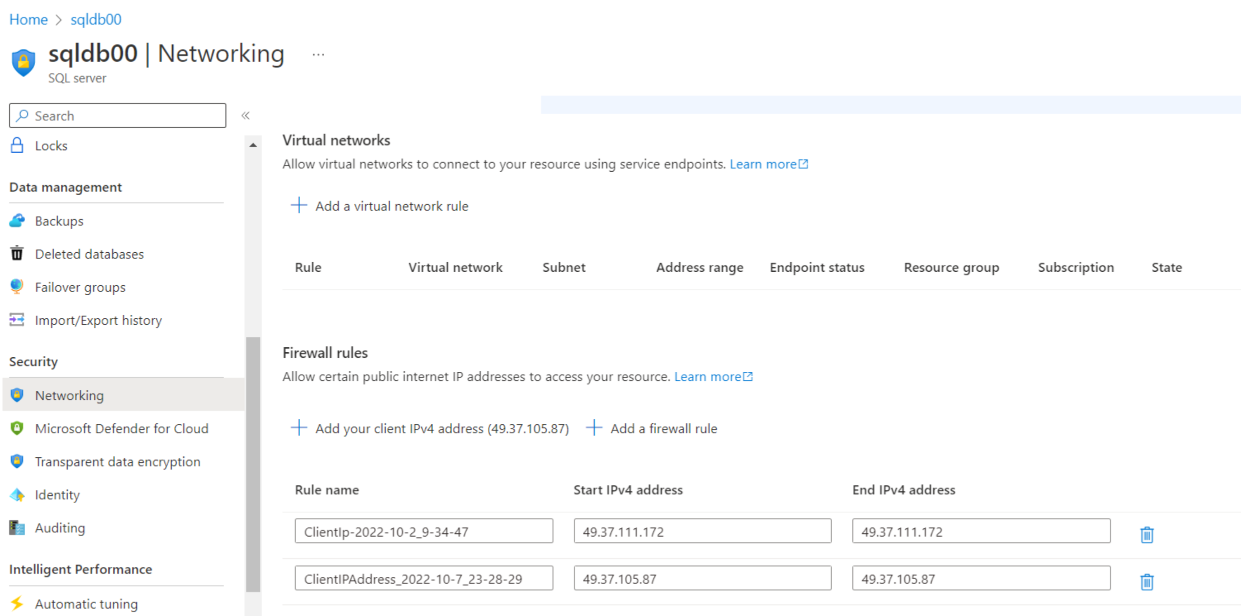


1. 3. Create a SQL Database to store the current year of the payroll data

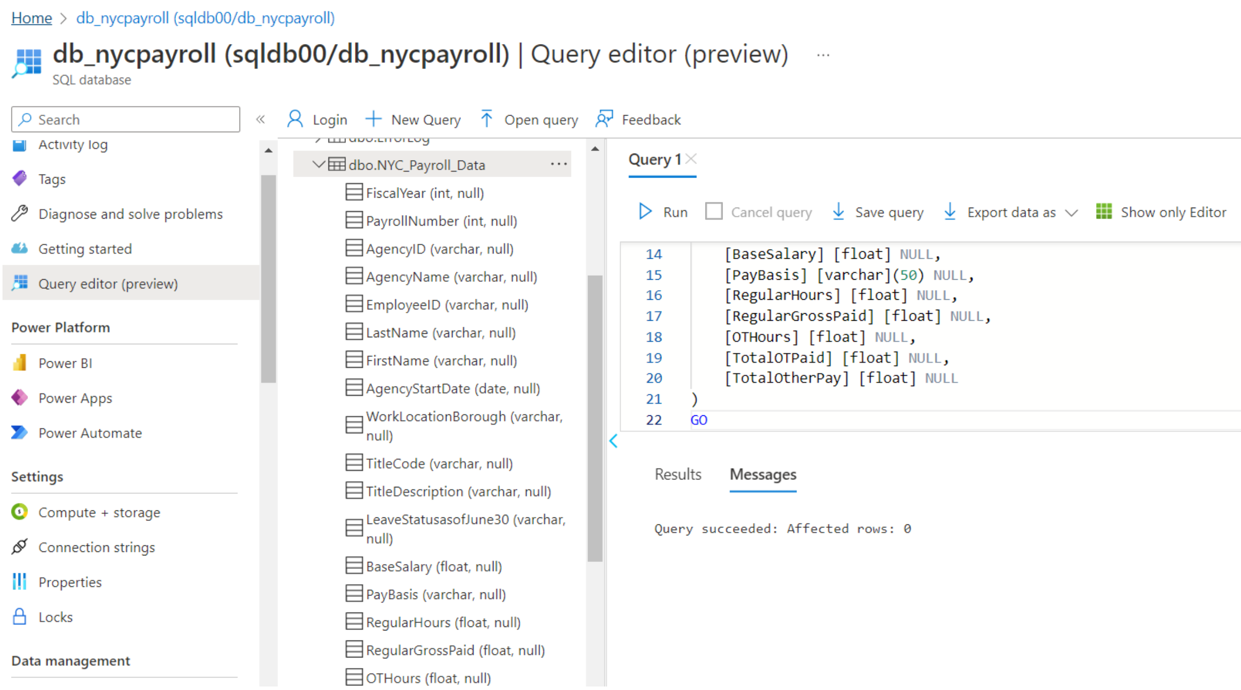
- Create a SQL Database resource named **db_nycpayroll**



- Add client IP address to the SQL DB firewall

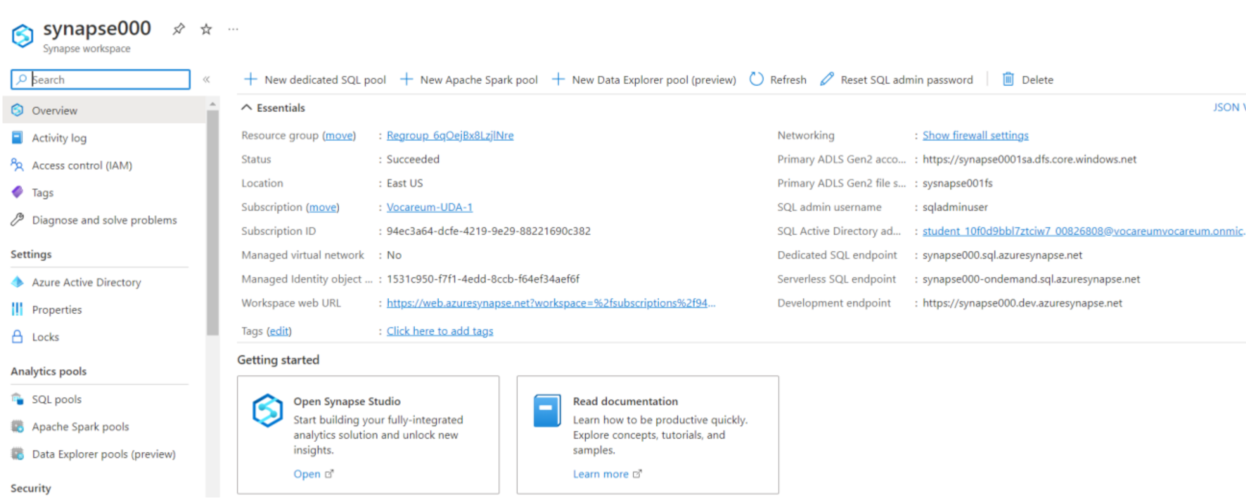


- Create a table called **NYC_Payroll_Data** in **db_nycpayroll** in the Azure Query Editor with given SQL



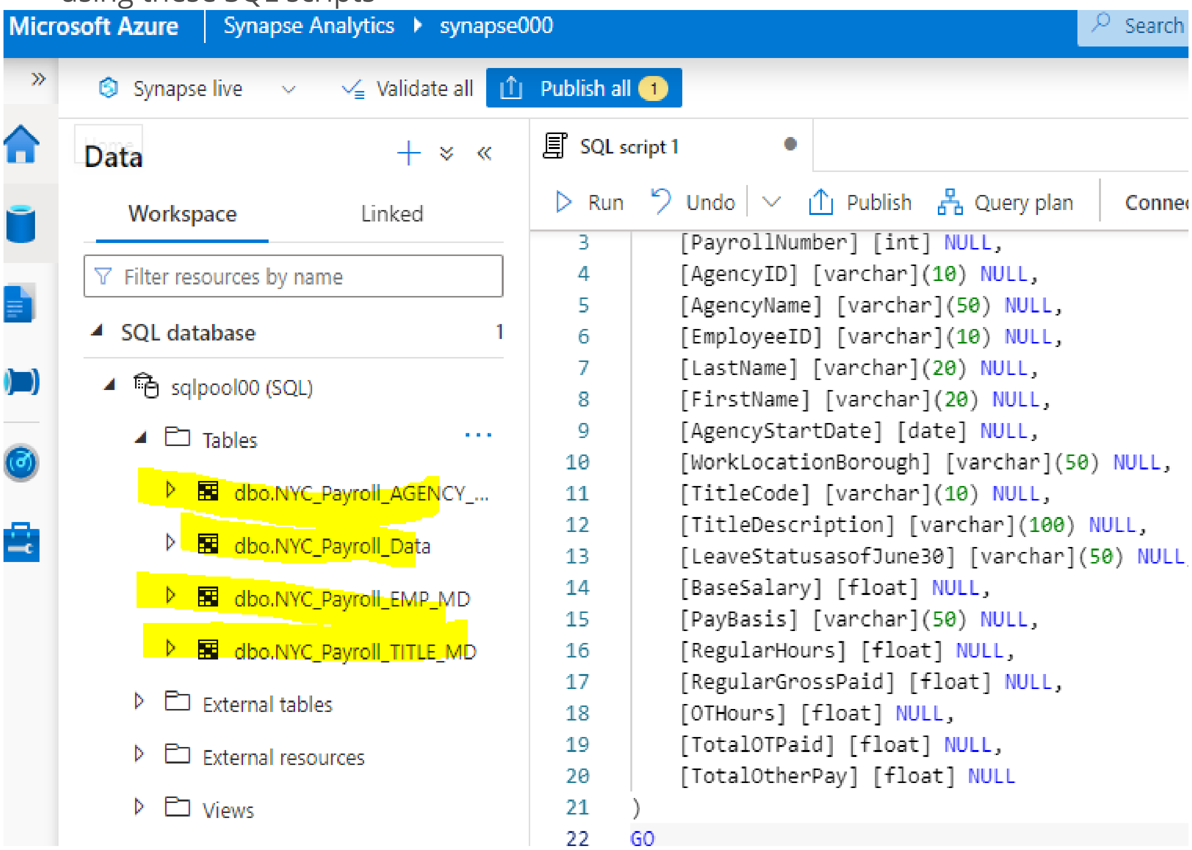
1. 4. Create A Synapse Analytics workspace, or use one you already have created.

- Create a Synapse Analytics workspace in the Azure portal.



- Create a SQL dedicated pool in the Synapse Analytics workspace. Select DW100c as performance level. Keep defaults for other settings.

- In the SQL dedicated pool, created master data tables and payroll transaction tables using these SQL scripts

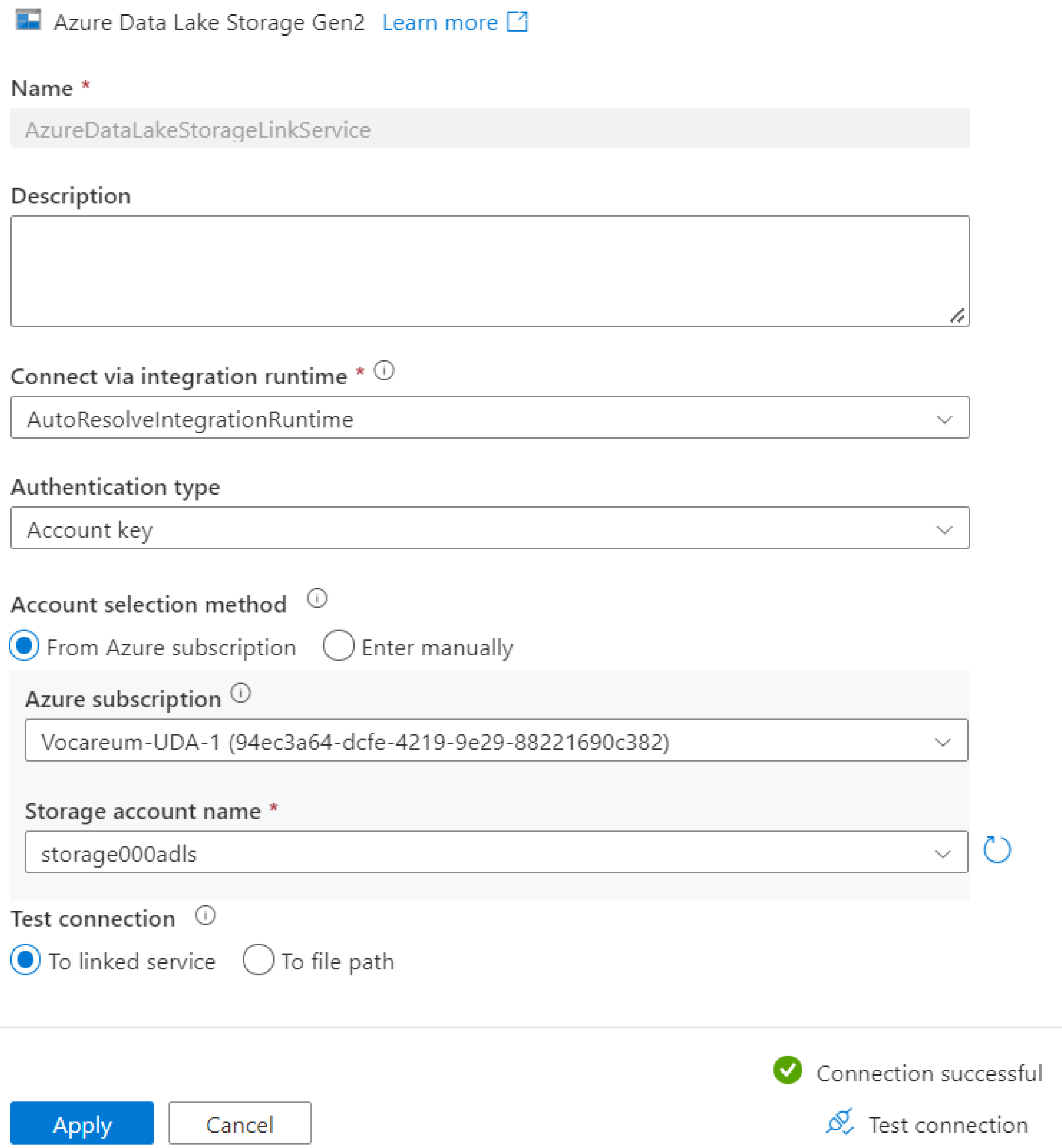


Step 2: Create Linked Services

2.1. Create a Linked Service for Azure Data Lake

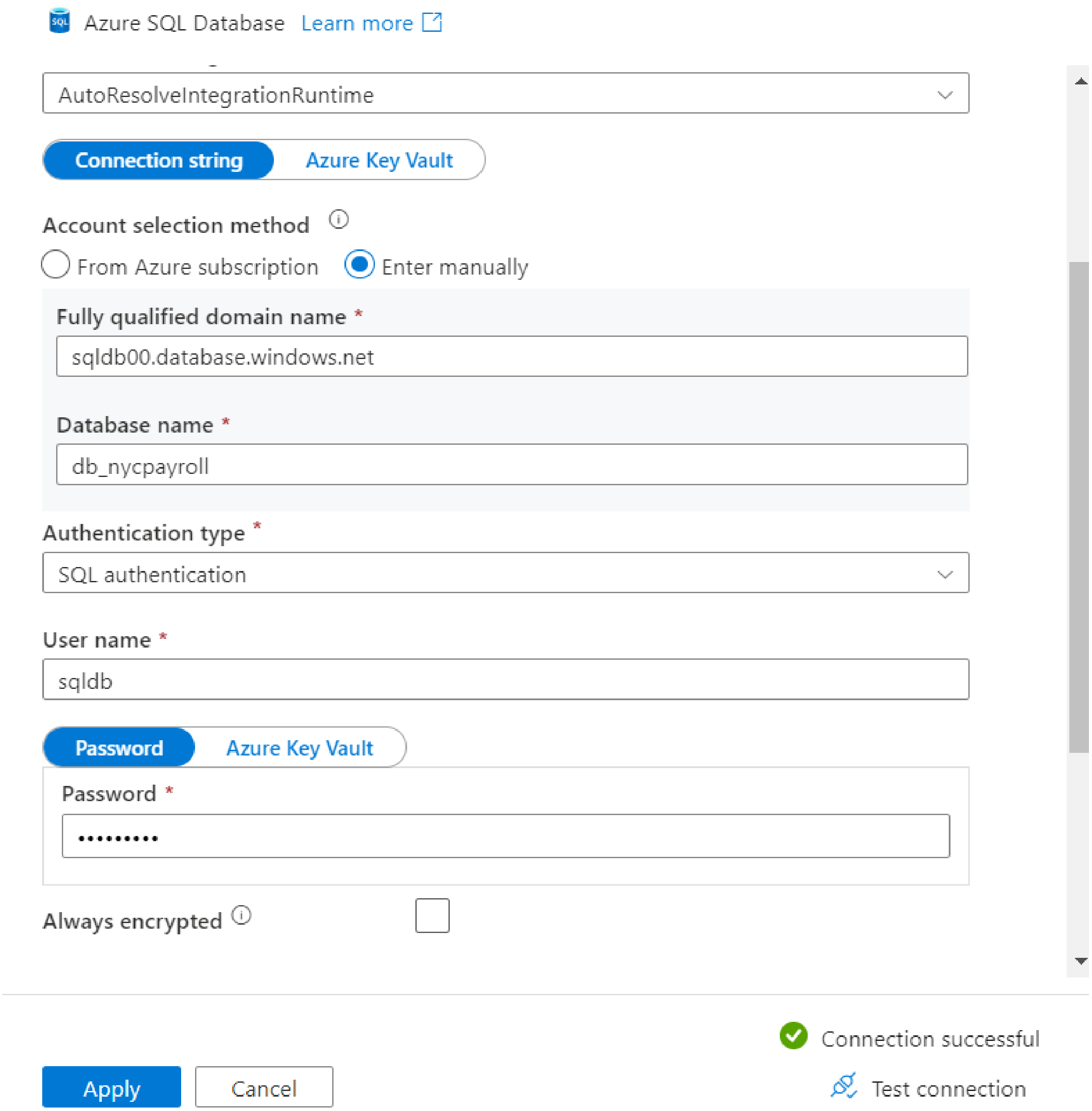
In Azure Data Factory, create a linked service to the data lake that contains the data files

- From the data stores, select Azure Data Lake Gen 2
- Test the connection



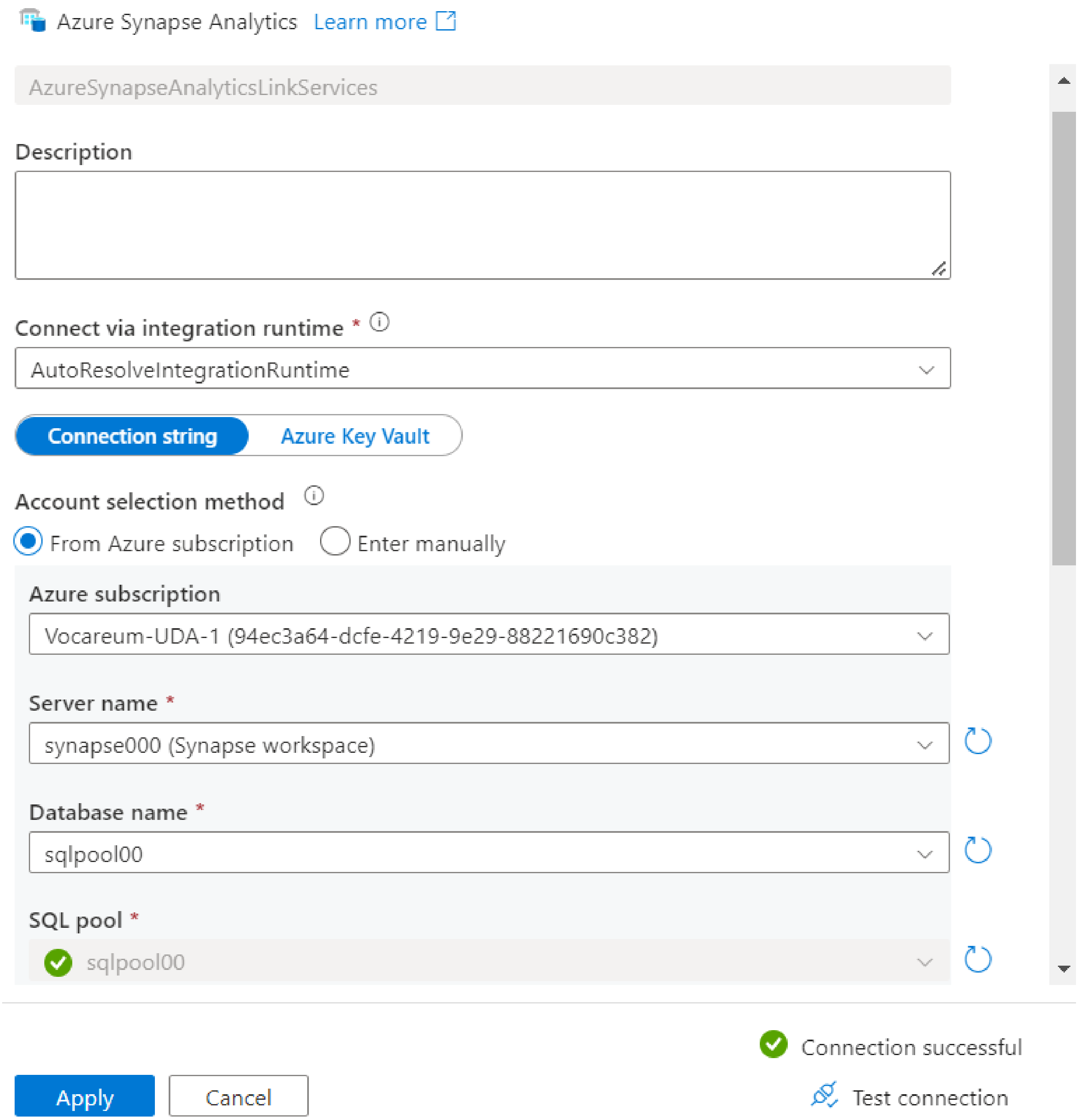
2. 2. Create a Linked Service to SQL Database that has the current (2021) data

- If you get a connection error, remember to add the IP address to the firewall settings in SQL DB in the Azure Portal



2. 3. Create a Linked Service for Synapse Analytics

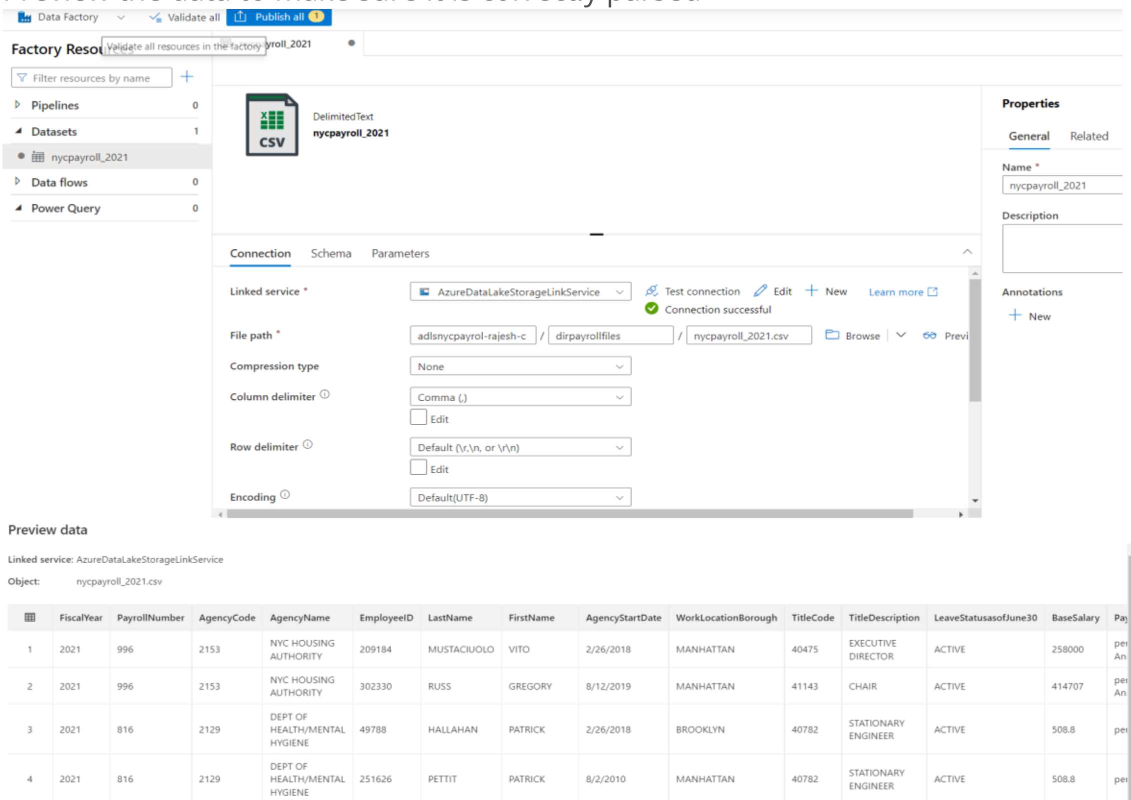
- Create the linked service to the SQL pool.



Step 3: Create Datasets in Azure Data Factory

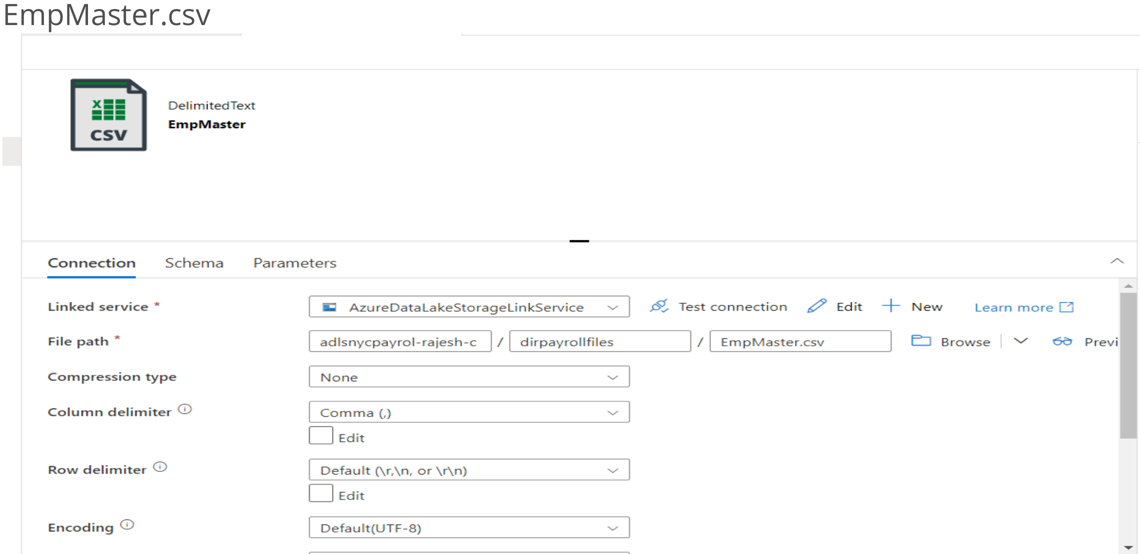
3.1. Create the datasets for the 2021 Payroll file on Azure Data Lake Gen2

- Select DelimitedText
- Set the path to the nycpayroll_2021.csv in the Data Lake
- Preview the data to make sure it is correctly parsed



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

- EmpMaster.csv



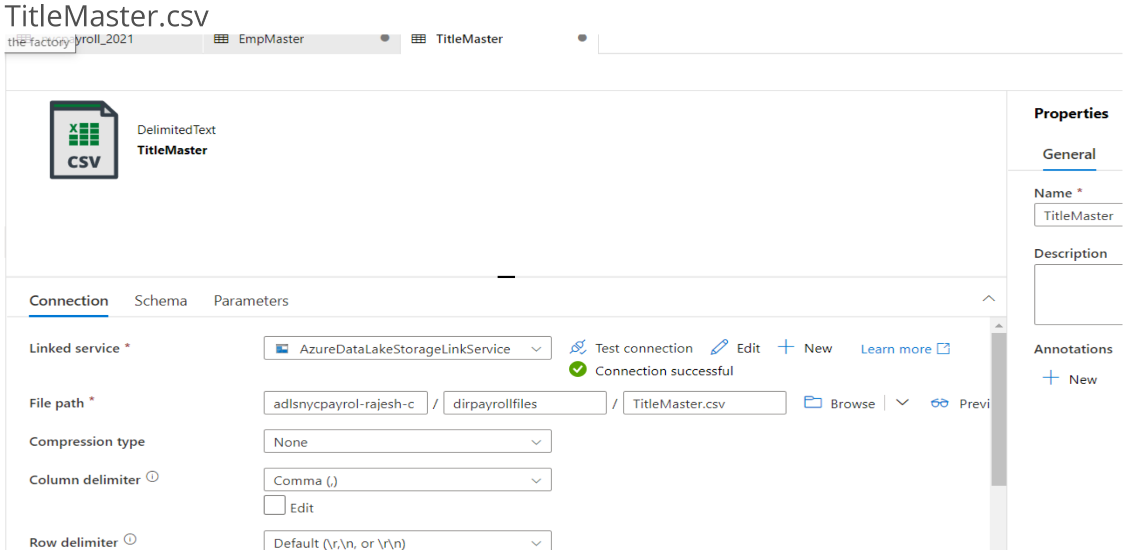
Preview data

Linked service: AzureDataLakeStorageLinkService

Object: EmpMaster.csv

	EmployeeID	LastName	FirstName
1	100001	AACHEN	DAVID
2	100002	AACHEN	MONICA
3	100003	AADAMS	LAMMELL
4	100004	AADIL	IRIS
5	100005	AALAAM	AMIR

- TitleMaster.csv



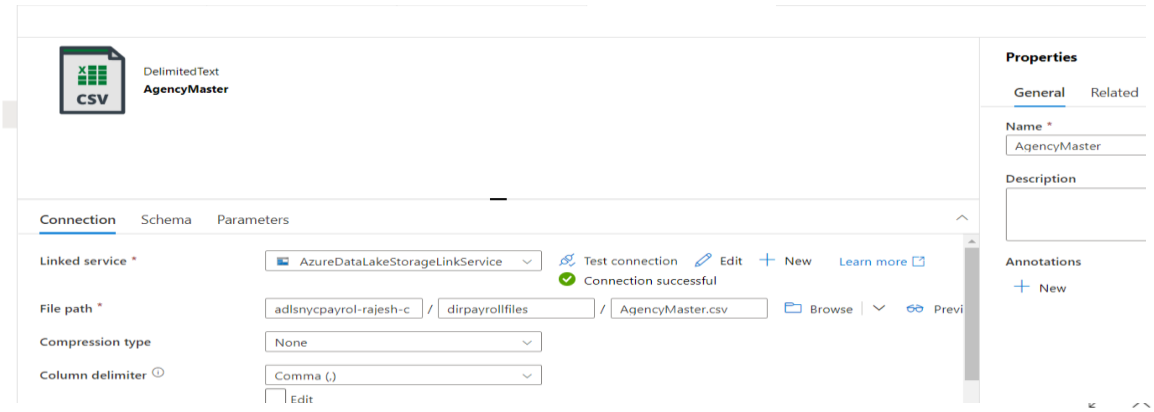
Preview data

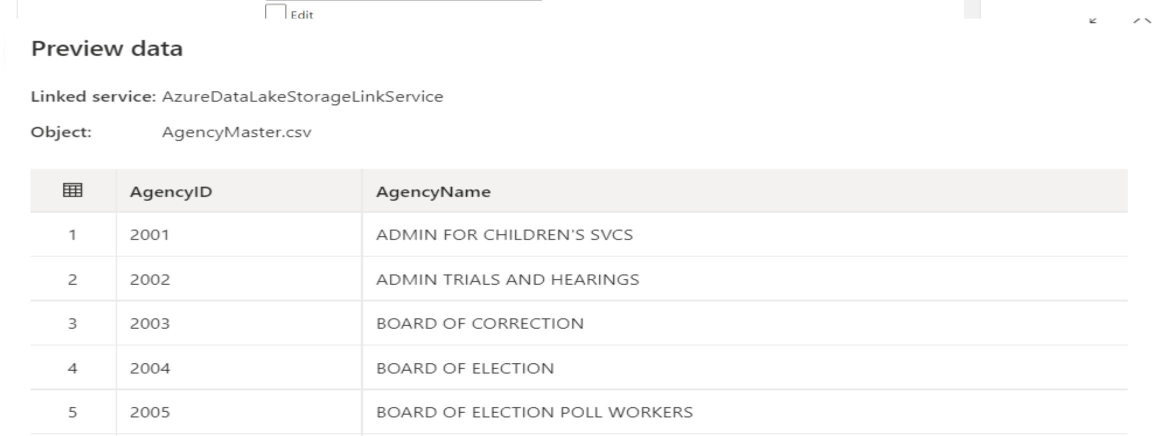
Linked service: AzureDataLakeStorageLinkService

Object: TitleMaster.csv

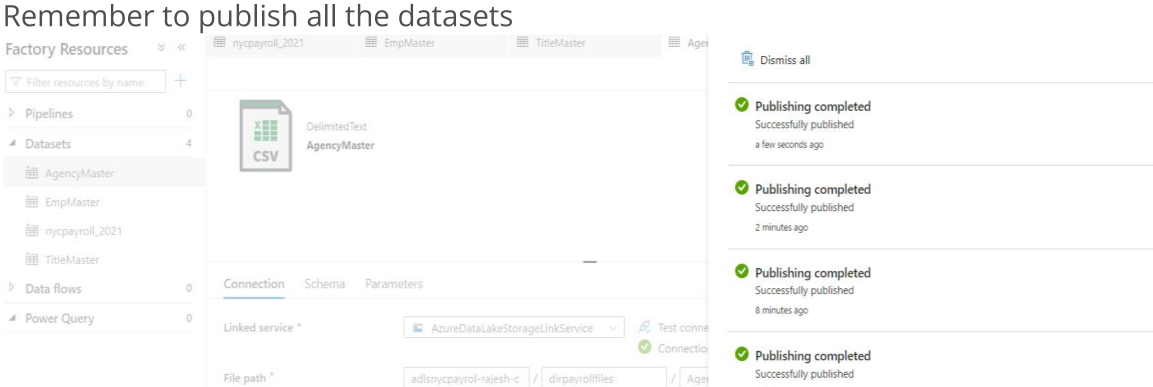
	TitleCode	TitleDescription
1	40001	*ADM SCHOOL SECURITY MANAGER-U
2	40002	*ADMIN SCHL SECUR MGR-MGL
3	40003	*AGENCY ATTORNEY
4	40004	*ASSISTANT ADVOCATE-PD
5	40005	*ASSOCIATE EDUCATION OFFICER
6	40006	*ATTORNEY AT LAW

- AgencyMaster.csv

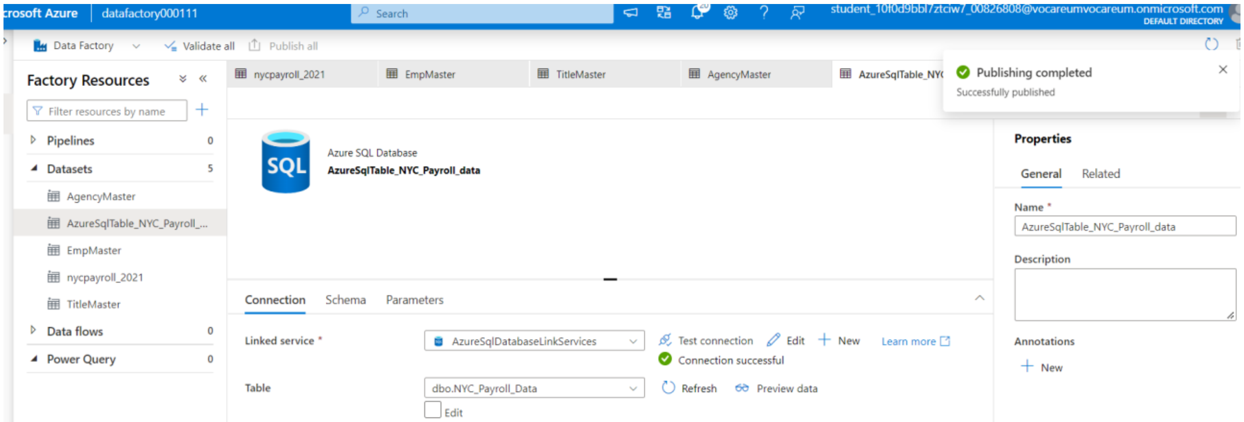




- Remember to publish all the datasets

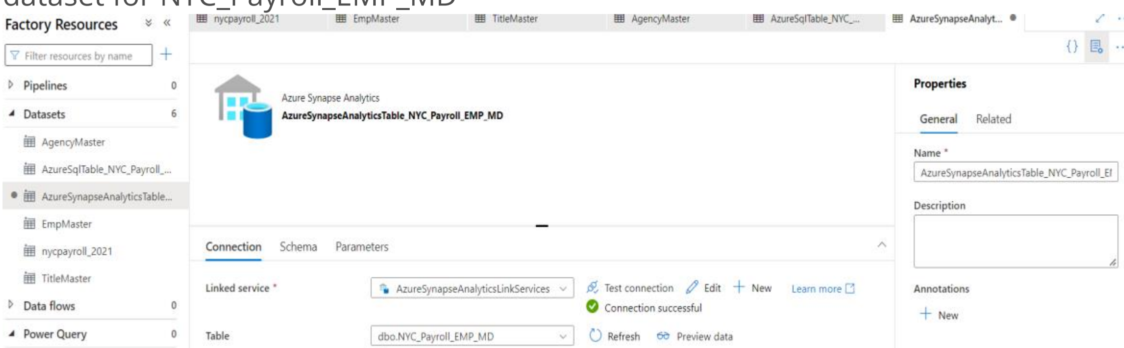


3.3. Create the dataset for transaction data table that should contain current (2021) data in SQL DB

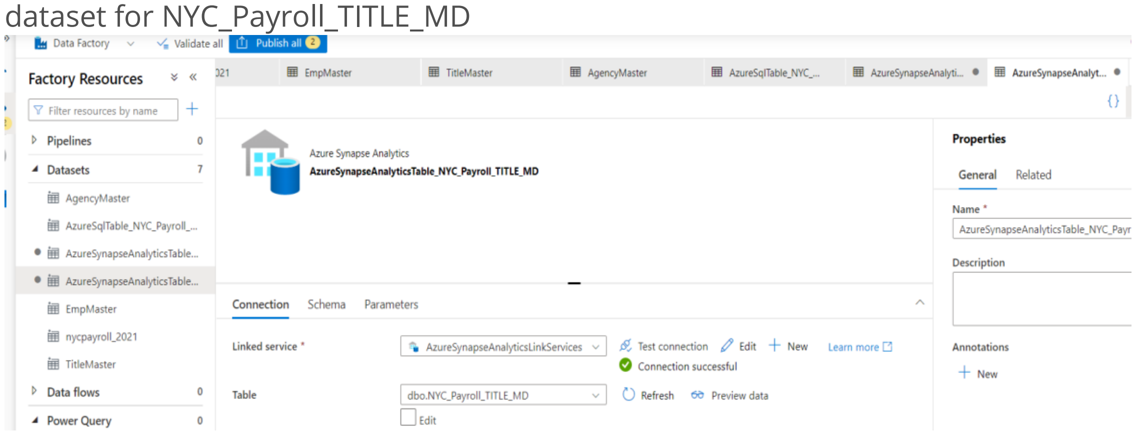


3.4. Create the datasets for destination (target) tables in Synapse Analytics

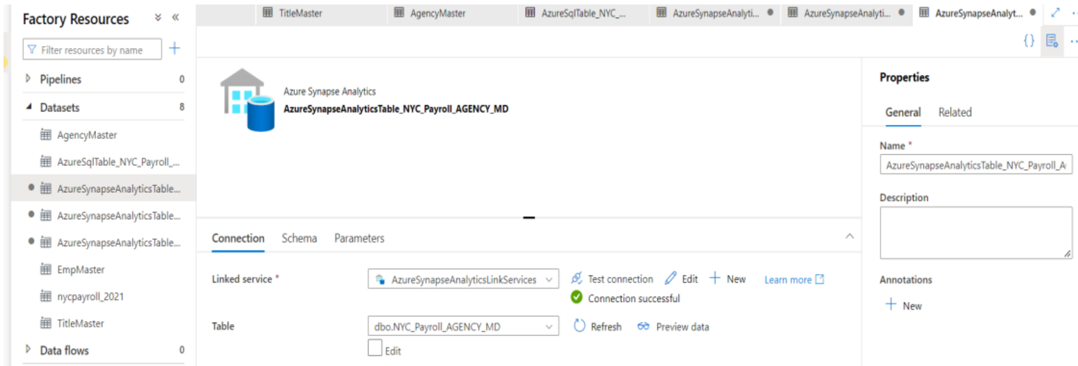
- dataset for NYC_Payroll_EMP_MD



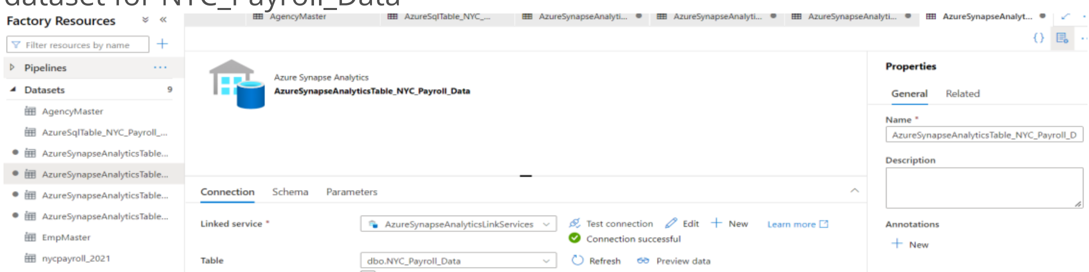
- dataset for NYC_Payroll_TITLE_MD



- dataset for NYC_Payroll_AGENCY_MD



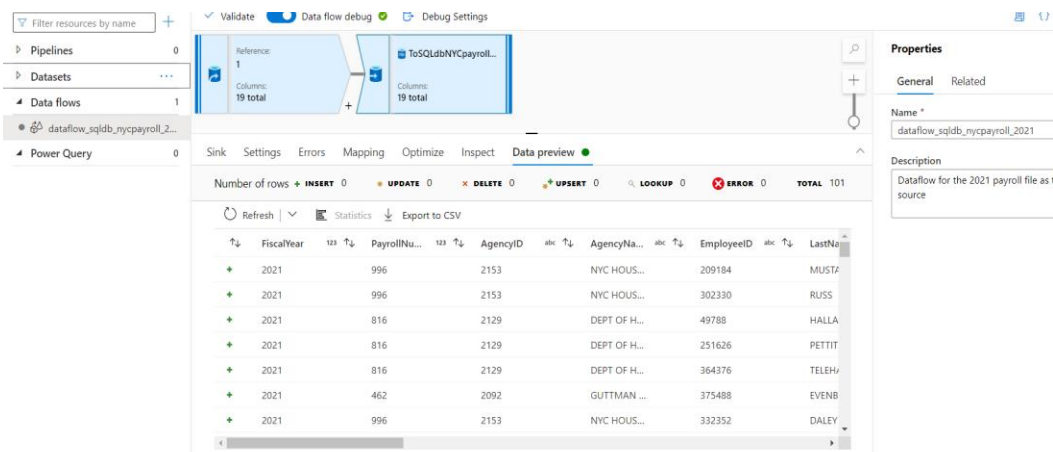
- dataset for NYC_Payroll_Data



Step 4: Create Data Flows

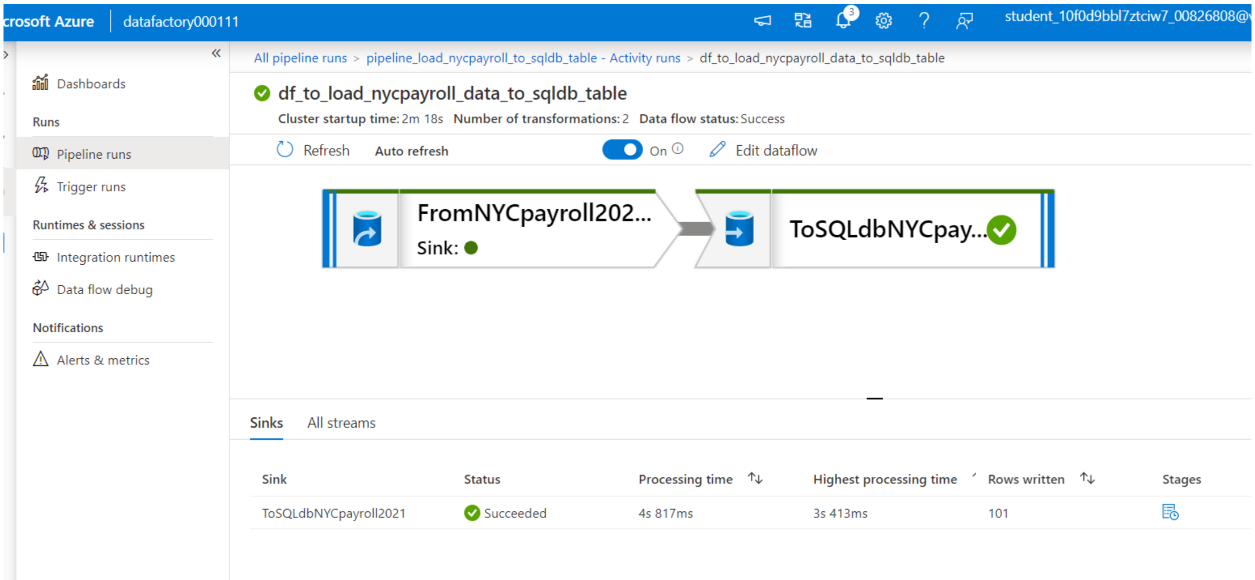
4.1. In Azure Data Factory, create the data flow to load 2021 Payroll Data to SQL DB transaction table (in the future NYC will load all the transaction data into this table).

- Create a new data flow
- Select the dataset for the 2021 payroll file as the source
- Select the sink dataset as the payroll table on SQL DB
- Make sure to reassign any missing source to target mappings

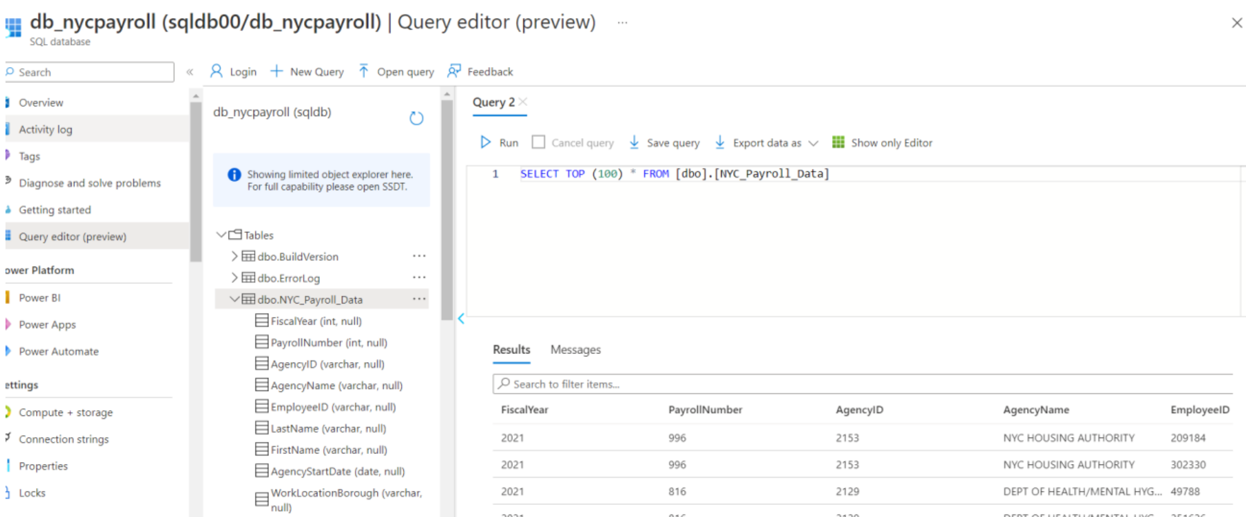


4.2. Create Pipeline to load 2021 Payroll data into transaction table in the SQL DB

- Create a new pipeline
- Select the data flow to load the 2021 file into SQLDB
- Trigger the pipeline
- Monitor the pipeline
- Take a screenshot of the Azure Data Factory screen pipeline run after it has finished.



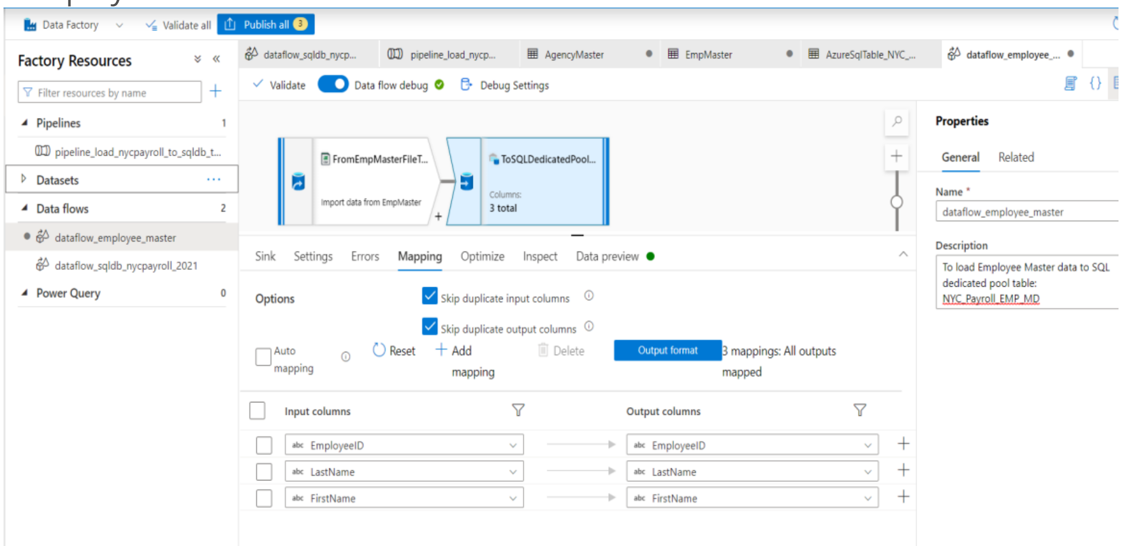
- Make sure the data is successfully loaded into the SQL DB table



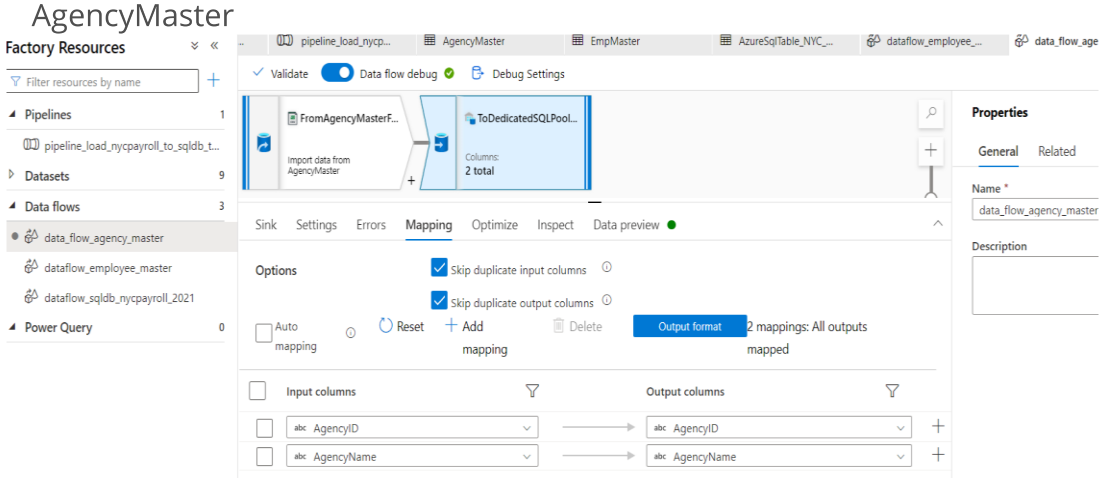
4.3. Create data flows to load the data from the data lake files into the Synapse Analytics data tables

- Create the data flows for loading Employee, Title, and Agency files into corresponding SQL pool tables on Synapse Analytics
- For each Employee, Title, and Agency file data flow, sink the data into each target Synaspe table

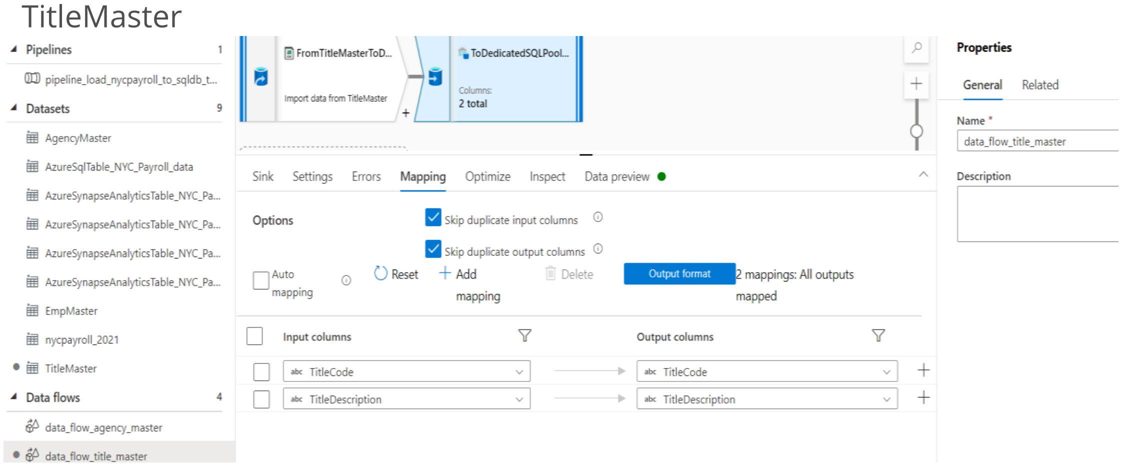
- EmployeeMaster



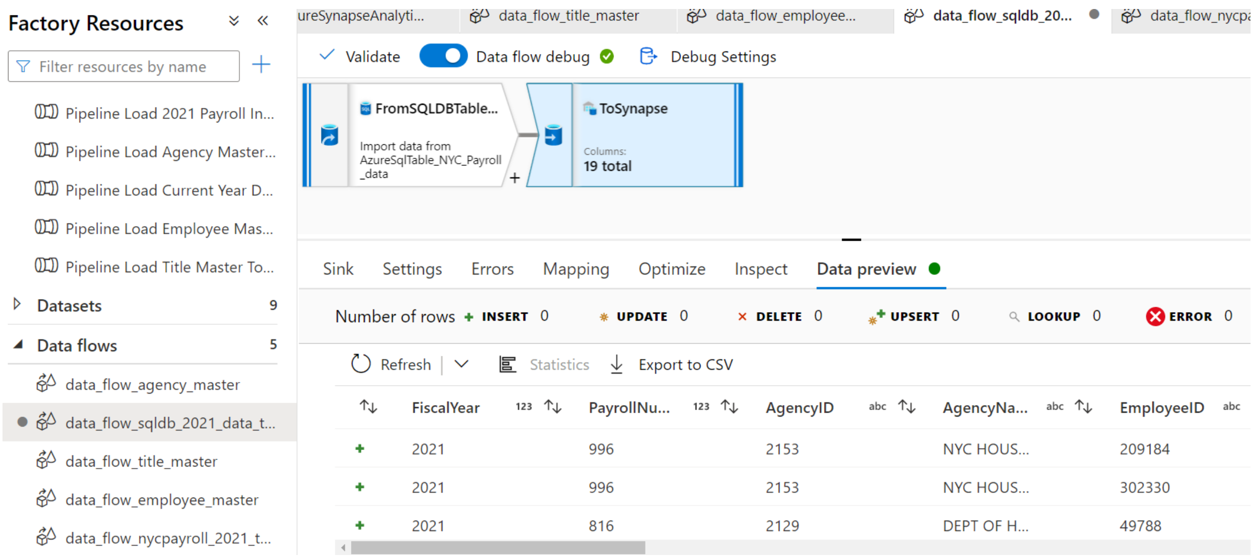
- AgencyMaster



- TitleMaster



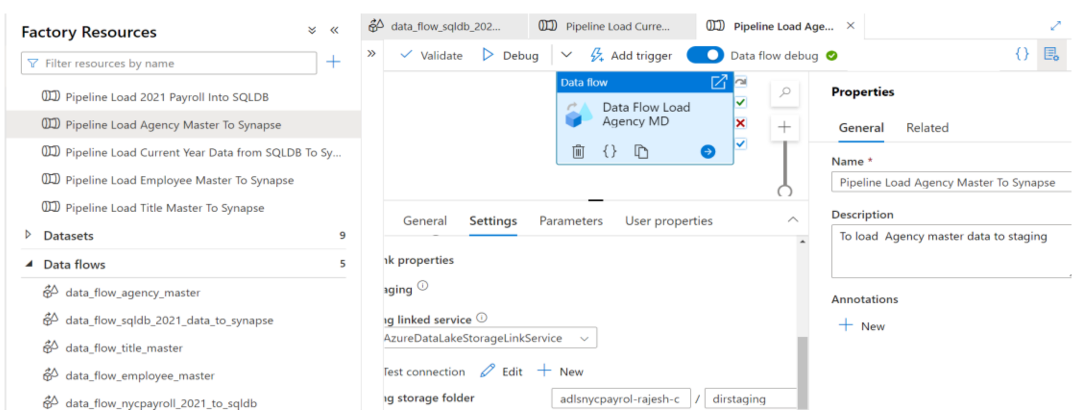
4.4. Create a data flow to load 2021 data from SQL DB to Synapse Analytics



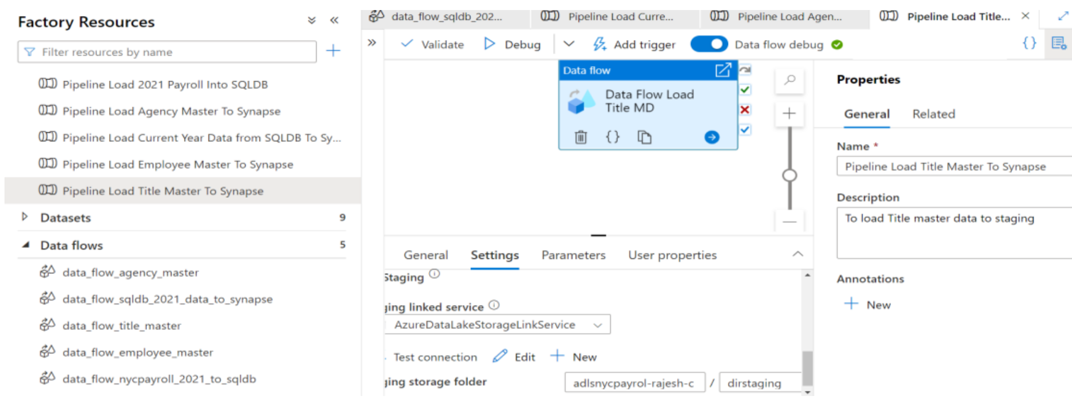
4.5. Create pipelines for Employee, Title, Agency, and year 2021 Payroll transaction data to Synapse Analytics containing the data flows.

- Select the dirstaging folder in the data lake storage for staging
- Optionally you can also create one master pipeline to invoke all the Data Flows
- Validate and publish the pipelines

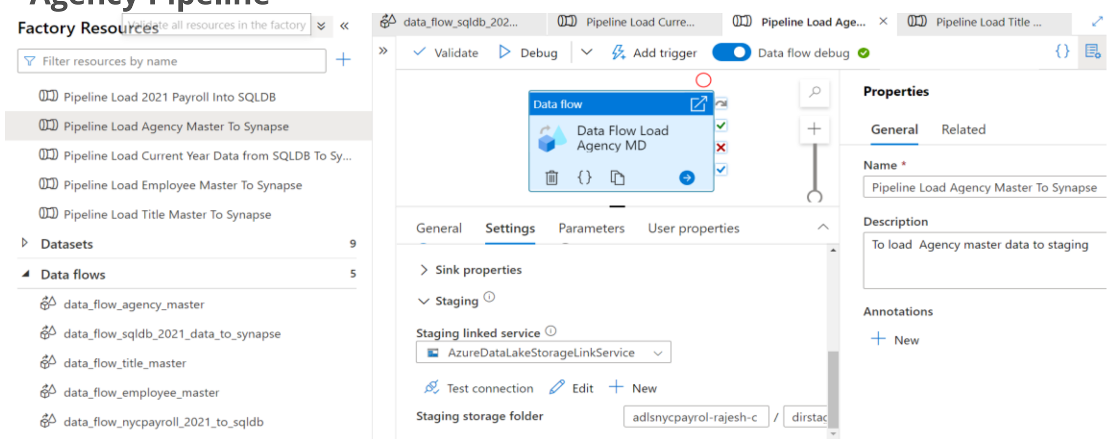
- **Employee pipeline**



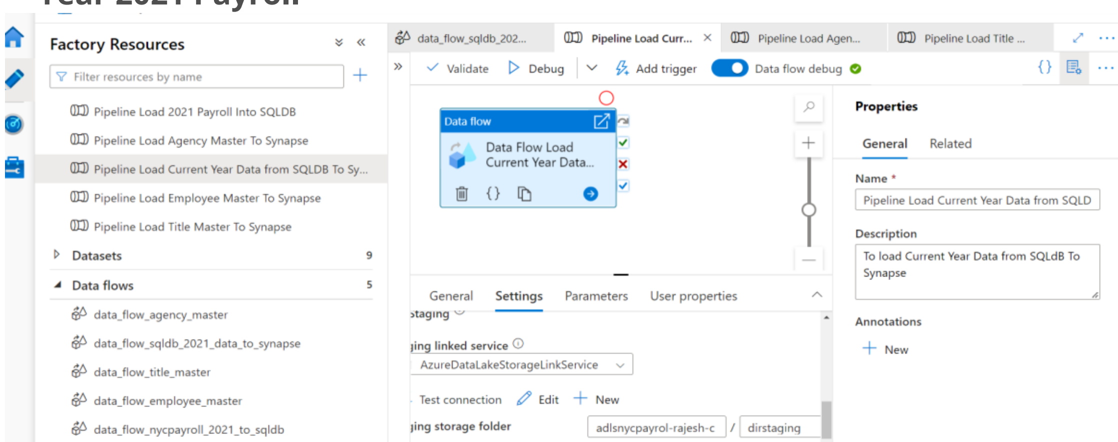
- **Title Pipeline**



- ## Agency Pipeline



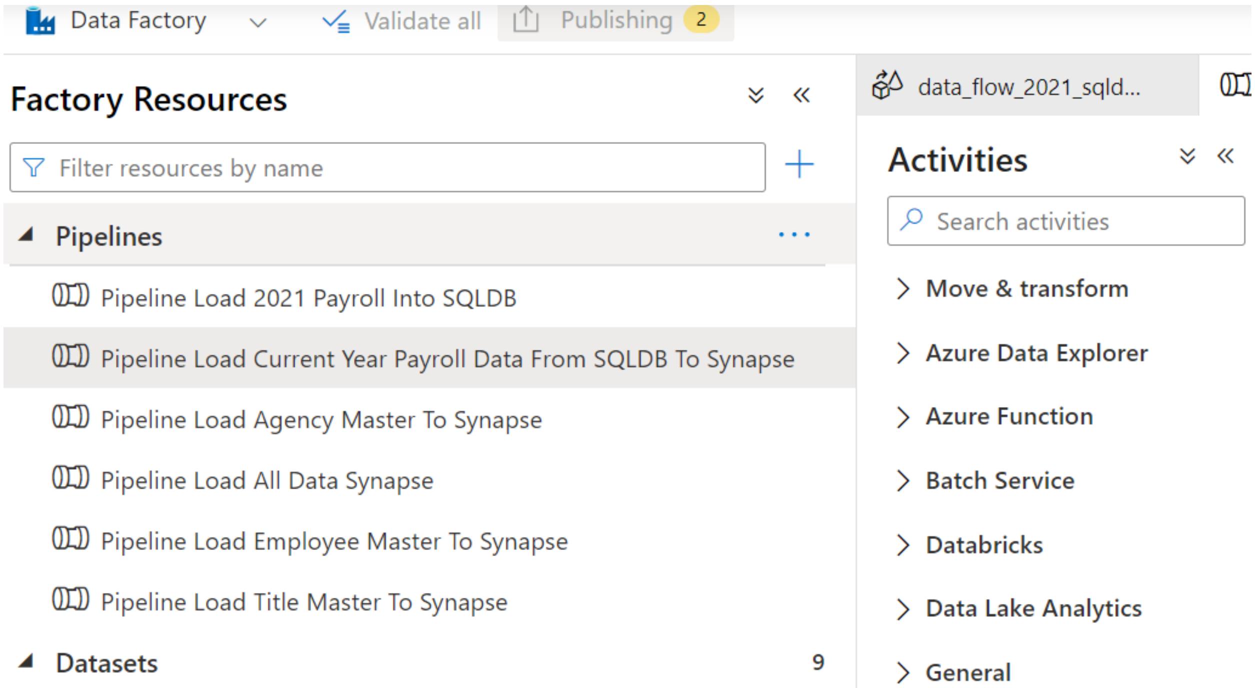
- ## Year 2021 Payroll



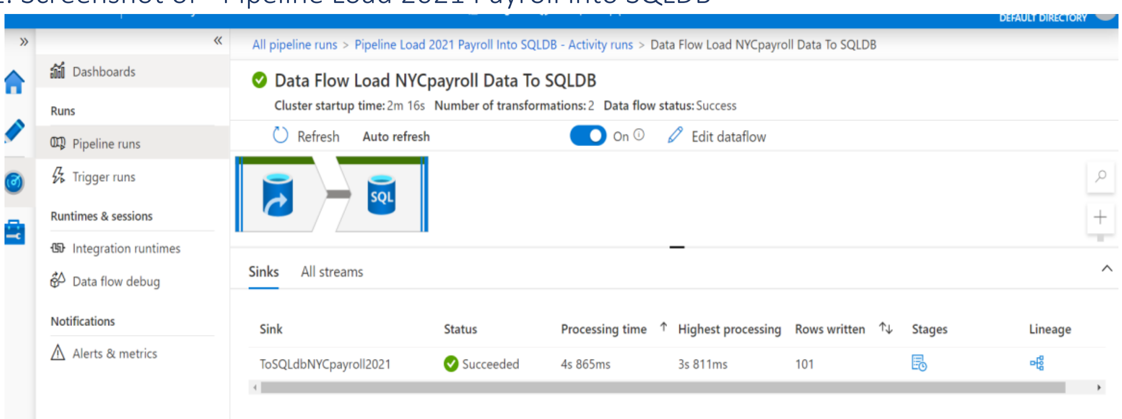
4.6. Trigger and monitor the Pipelines

- Take a screenshot of each pipeline run after it has finished, or one after your master pipeline run has finished.

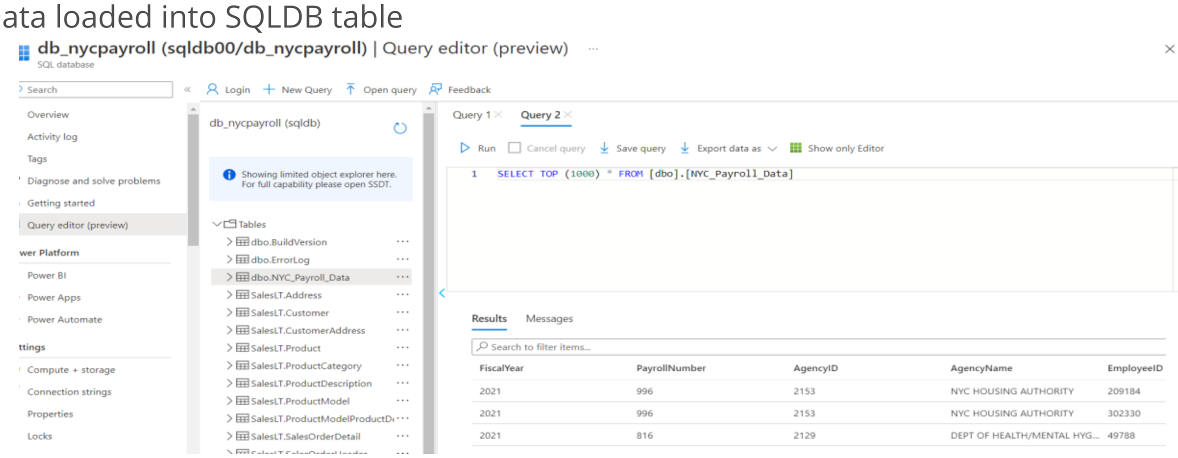
6 Pipelines name



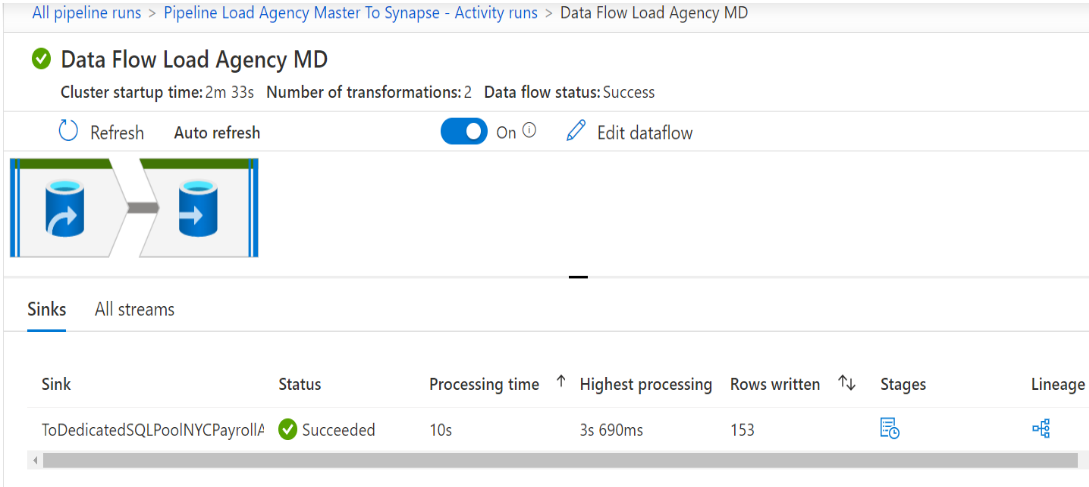
4.6.1. Screenshot of - Pipeline Load 2021 Payroll Into SQLDB



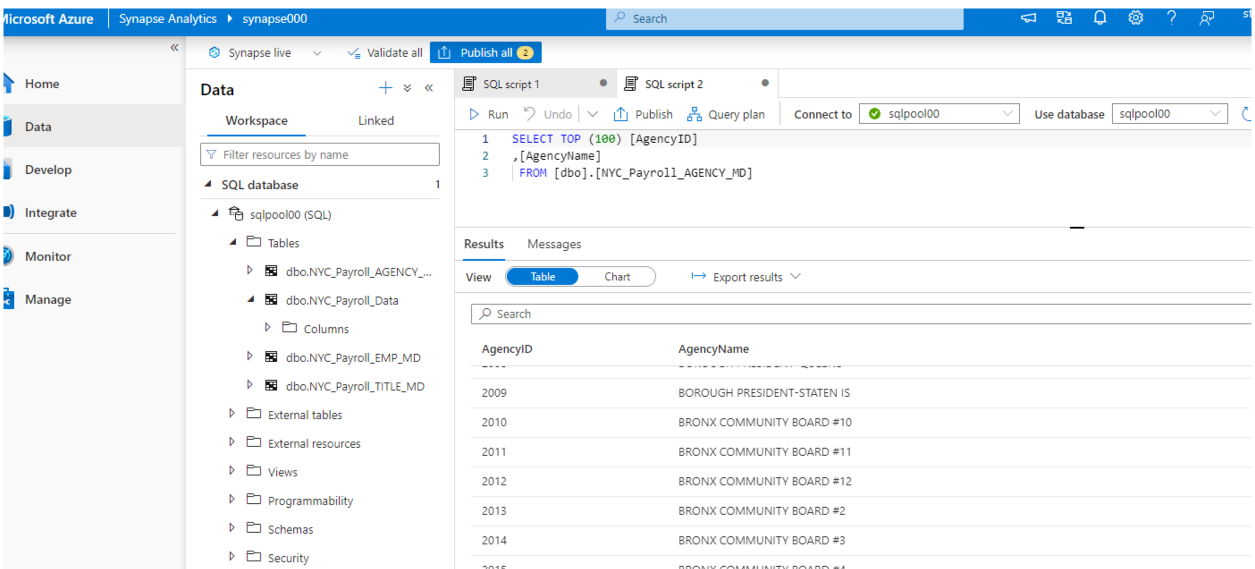
Data loaded into SQLDB table



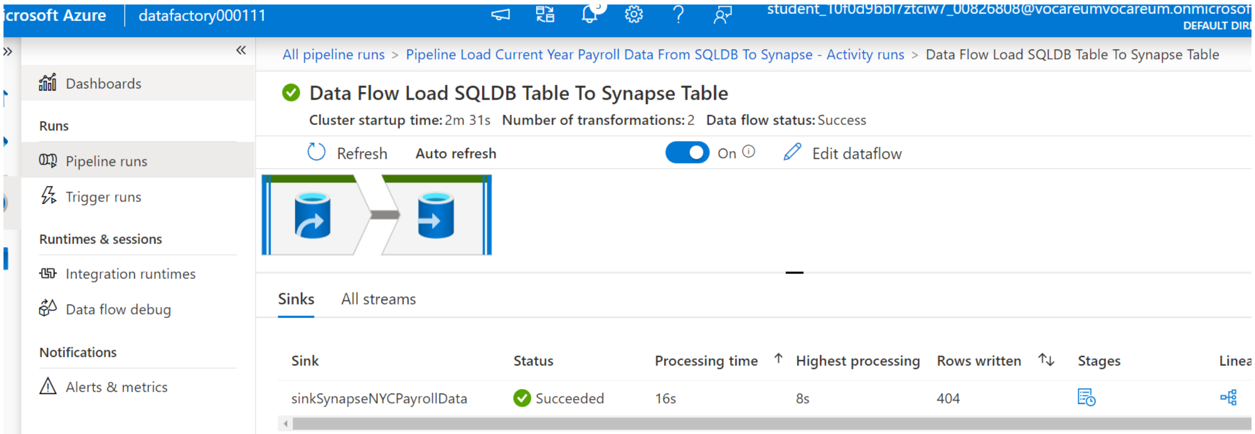
4.6.2. Screenshot of - Pipeline Load Agency Master To Synapse



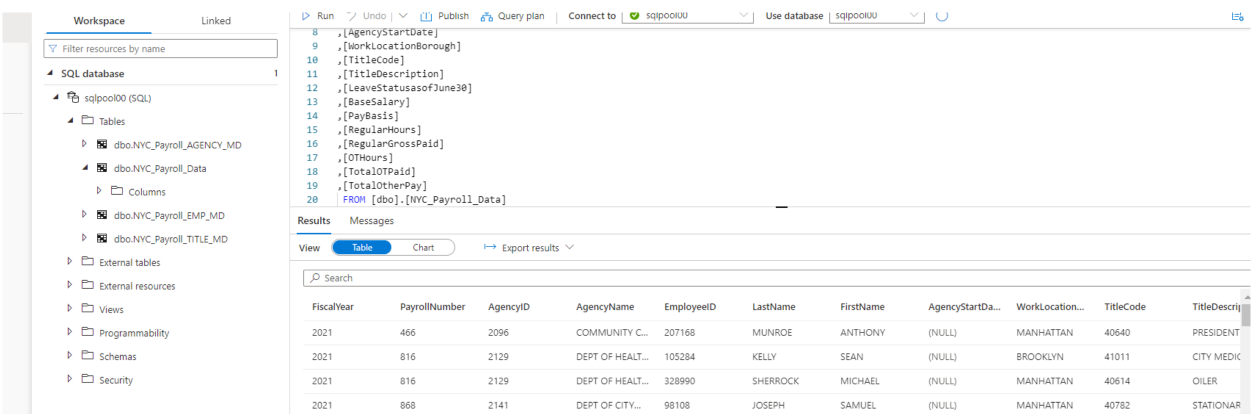
Data loaded to synapse table



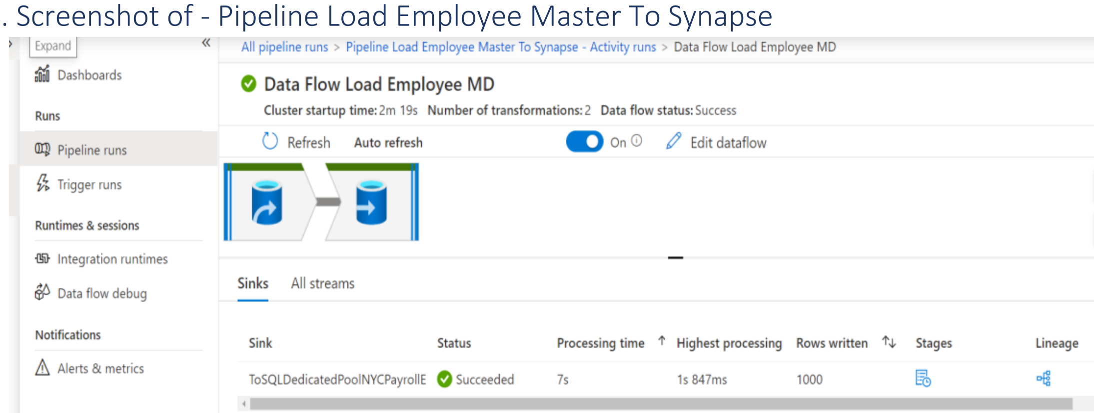
4.6.3 Screenshot of - Pipeline Load Current Year Payroll Data from SQLDB To Synapse



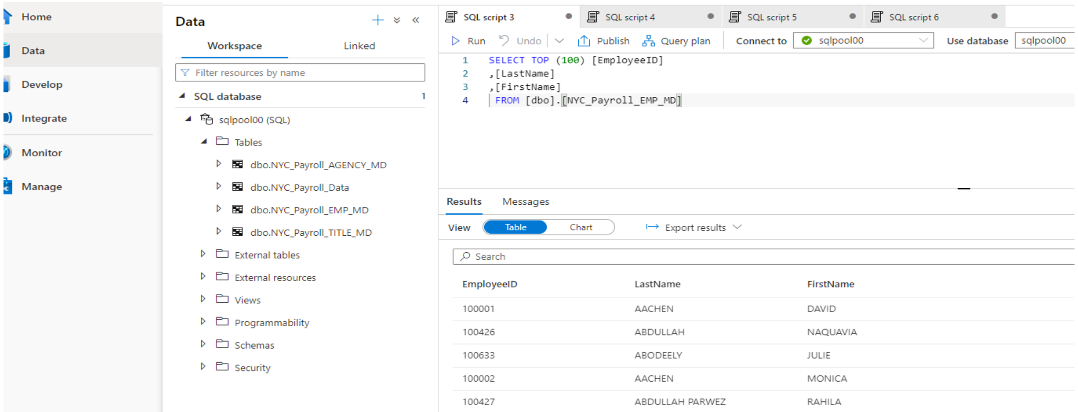
Data loaded to synapse Table



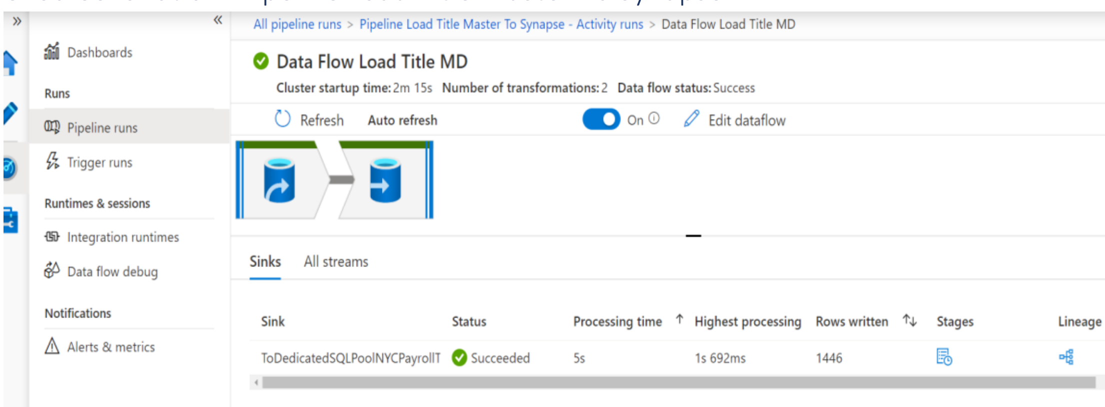
4.6.4. Screenshot of - Pipeline Load Employee Master To Synapse



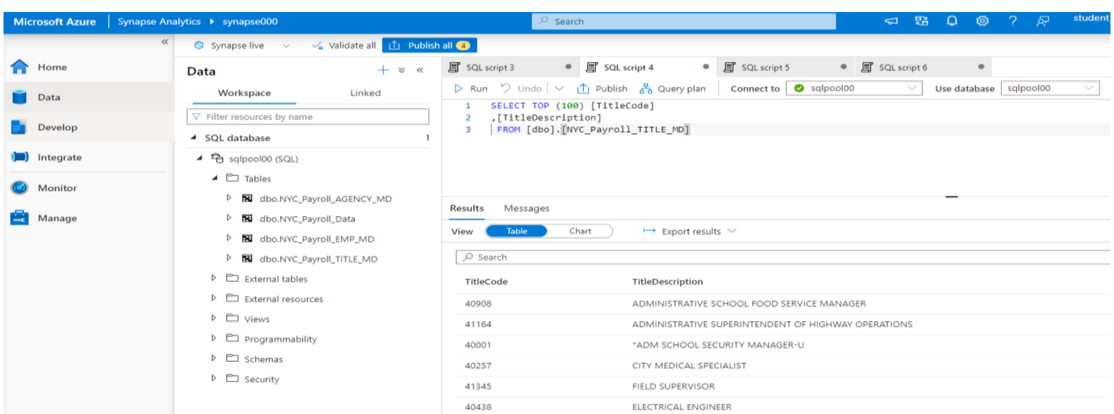
Data loaded to synapse table



4.6.5. Screenshot of - Pipeline Load Title Master To Synapse

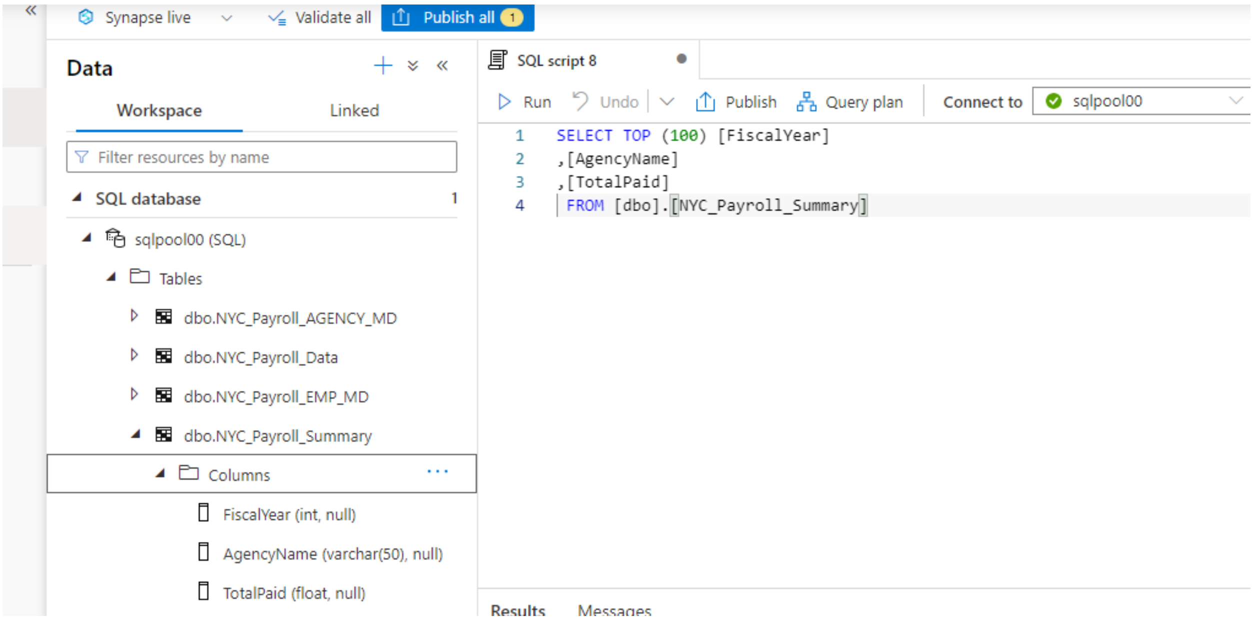


Data loaded to synapse table

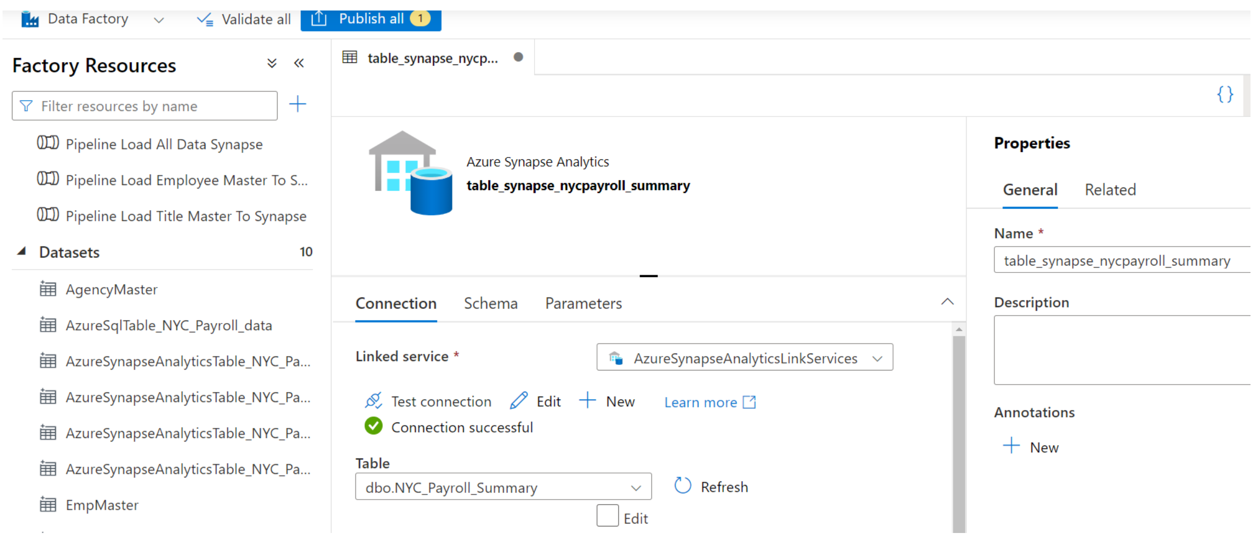


Step 5: Data Aggregation and Parameterization

5.1. Create a Summary table in Synapse with the given SQL script and create a dataset named table_synapse_nycpayroll_summary

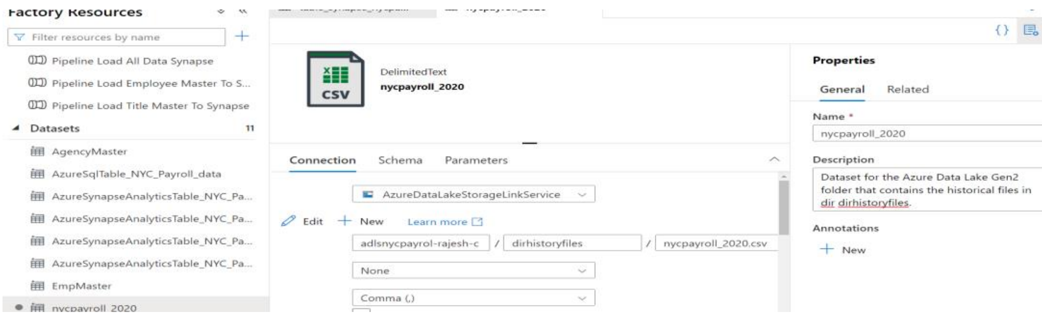


Dataset



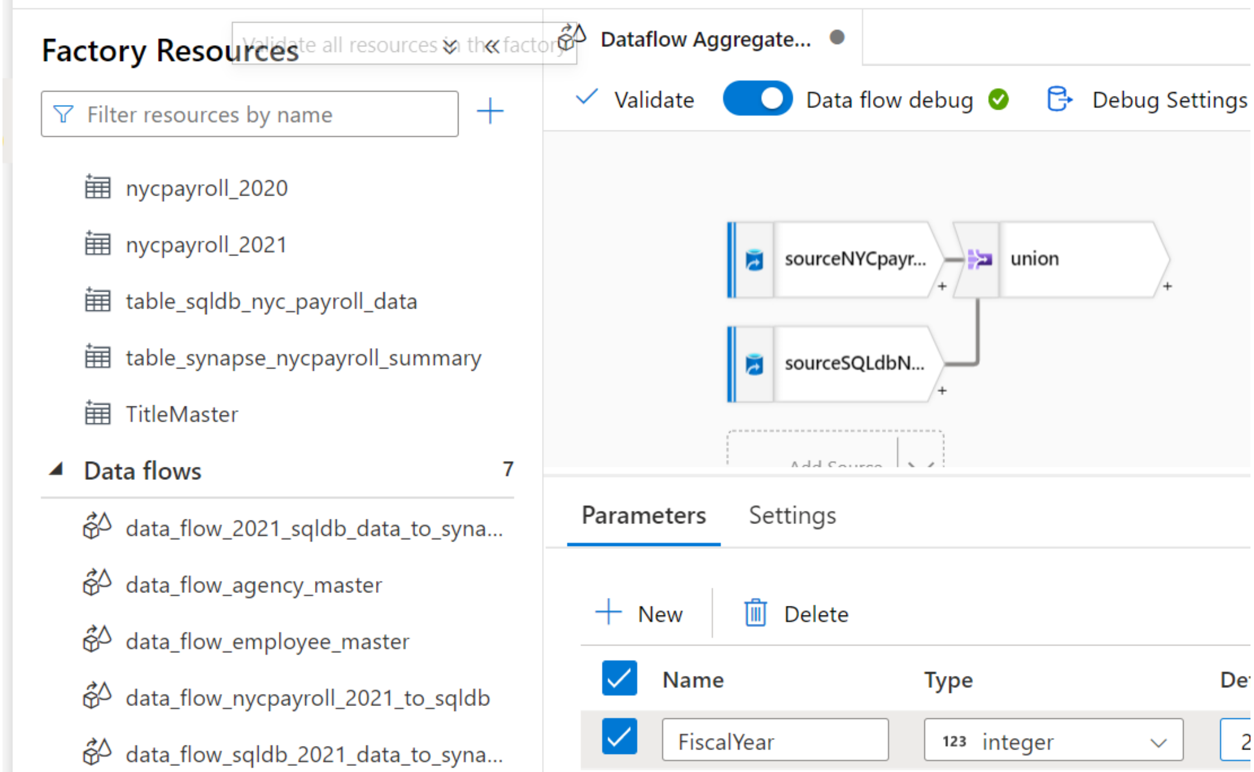
5.2. Create a new dataset for the Azure Data Lake Gen2 folder that contains the historical files.

- Select dirhistoryfiles in the data lake as the source

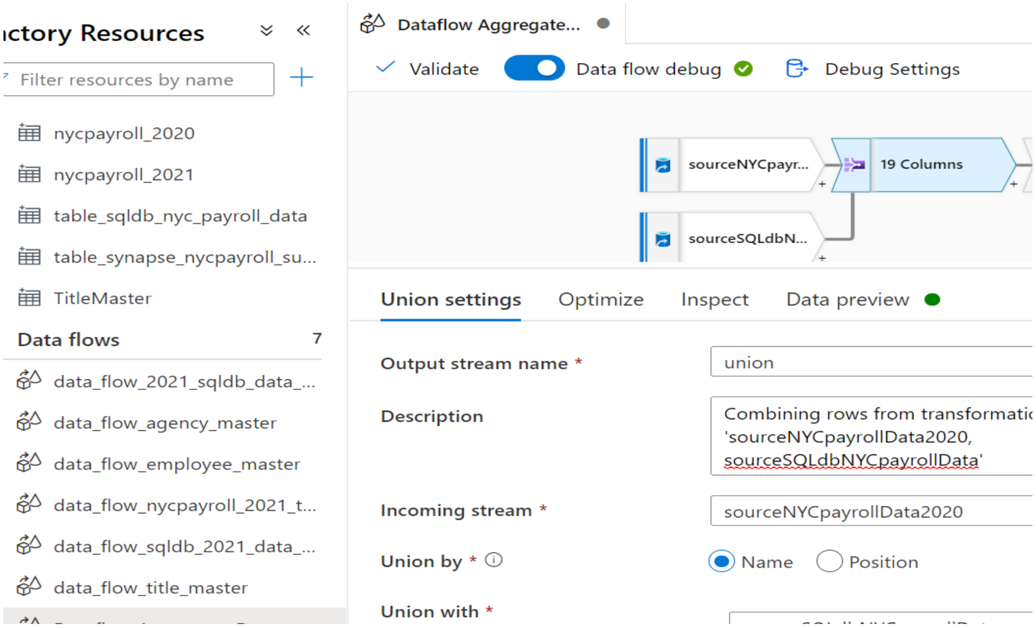


5.3. Create new data flow and name it Dataflow Aggregate Data

- 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

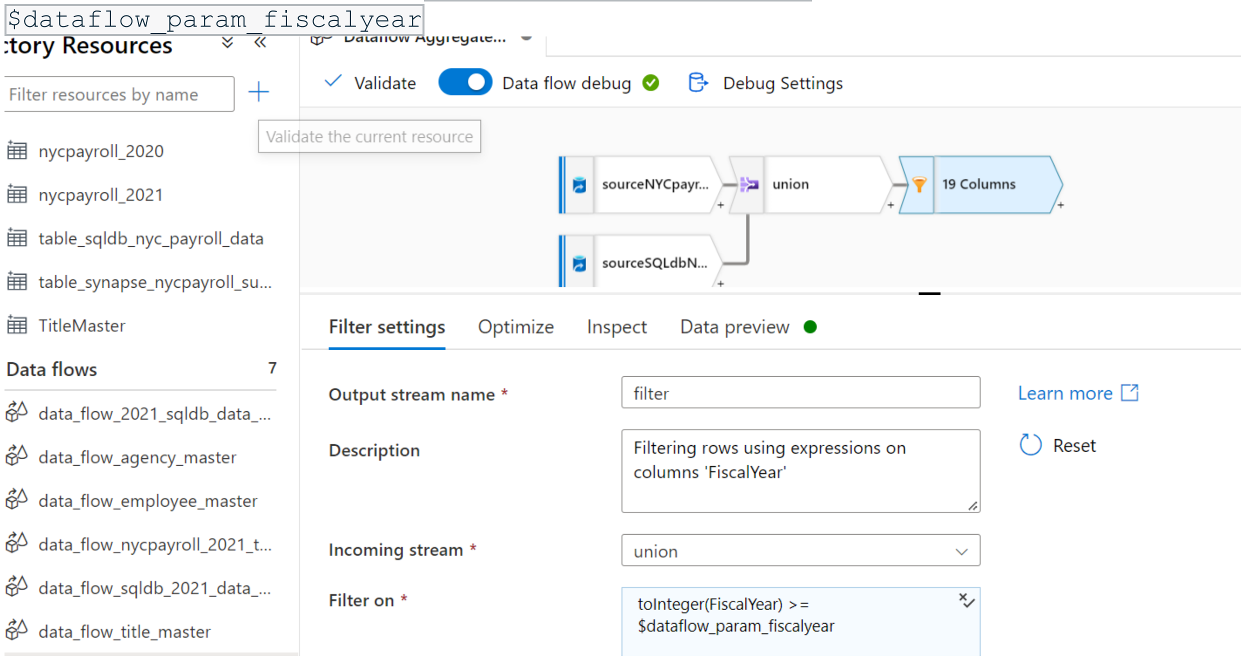


5.4. Create a new Union activity in the data flow and Union with history files



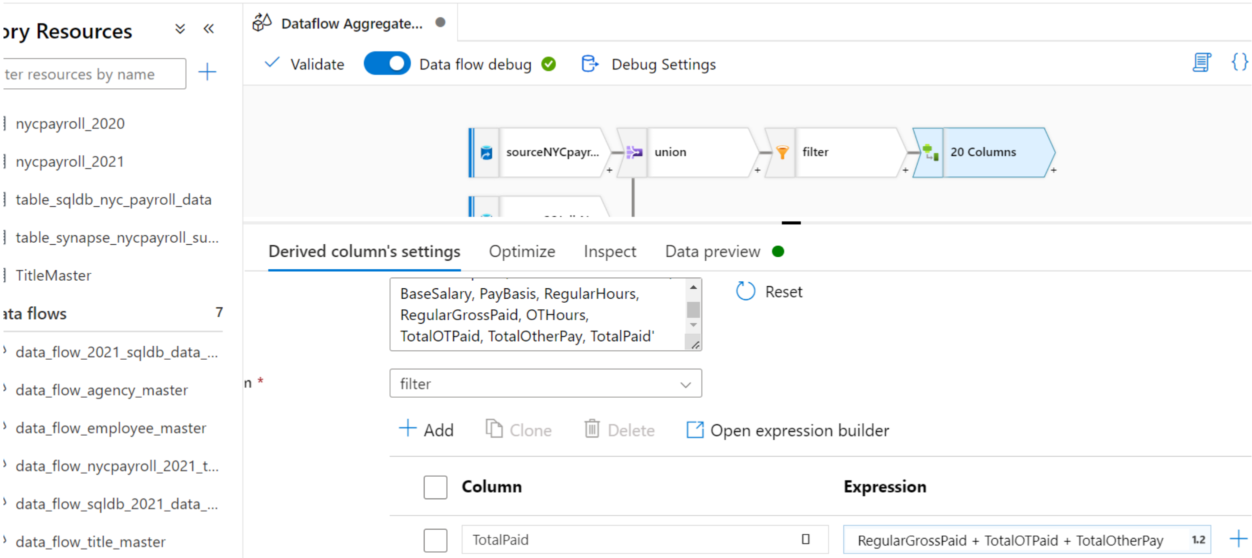
5.5. Add a Filter activity after Union

- In Expression Builder, enter `toInteger(FiscalYear) >= $dataflow_param_fiscalyear`



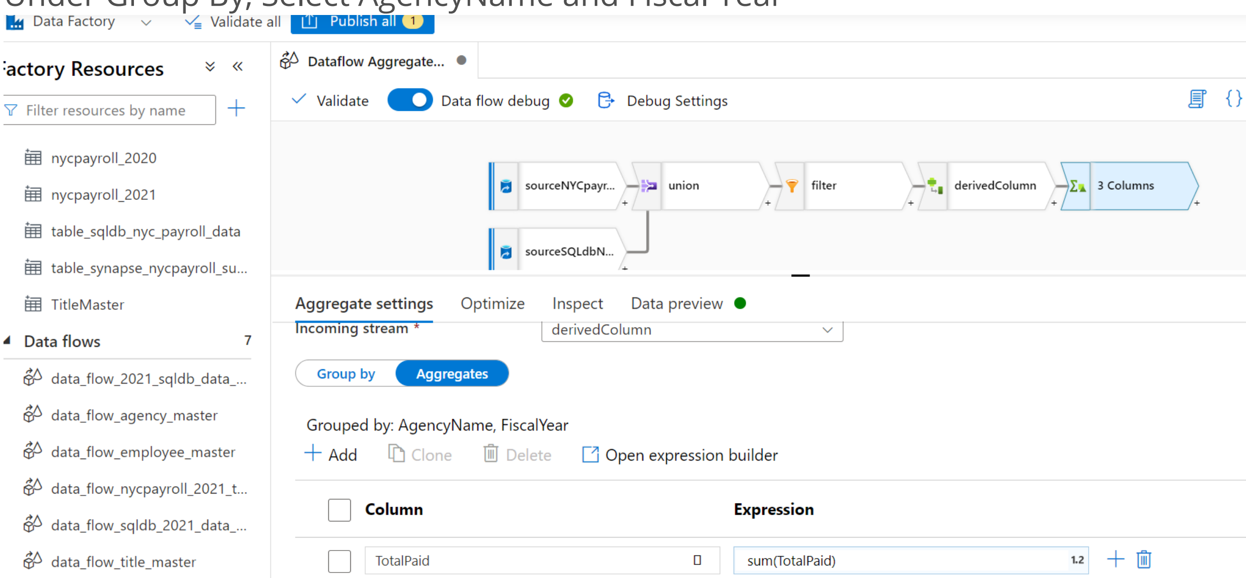
5.6. Derive a new TotalPaid column

- In Expression Builder, enter `RegularGrossPaid + TotalOTPaid+TotalOtherPay`



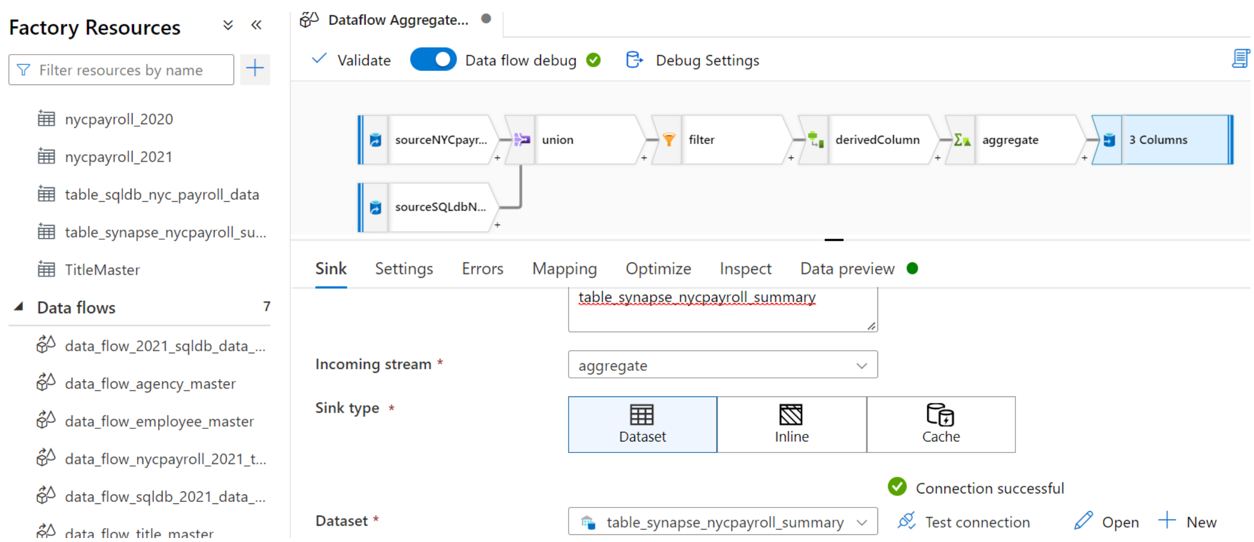
5.7. Add an Aggregate activity to the data flow next to the TotalPaid activity

- Under Group By, Select AgencyName and Fiscal Year

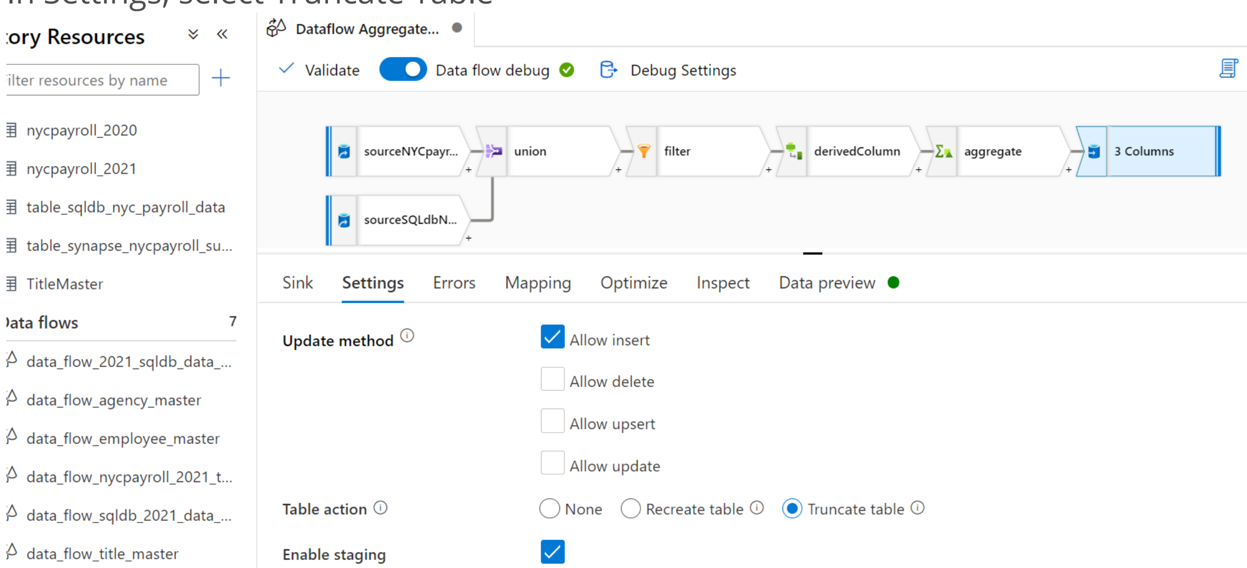


5.8.Add a Sink activity to the Data **Flow**

- Select the dataset to target (sink) the data into the Synapse Analytics Payroll Summary table.

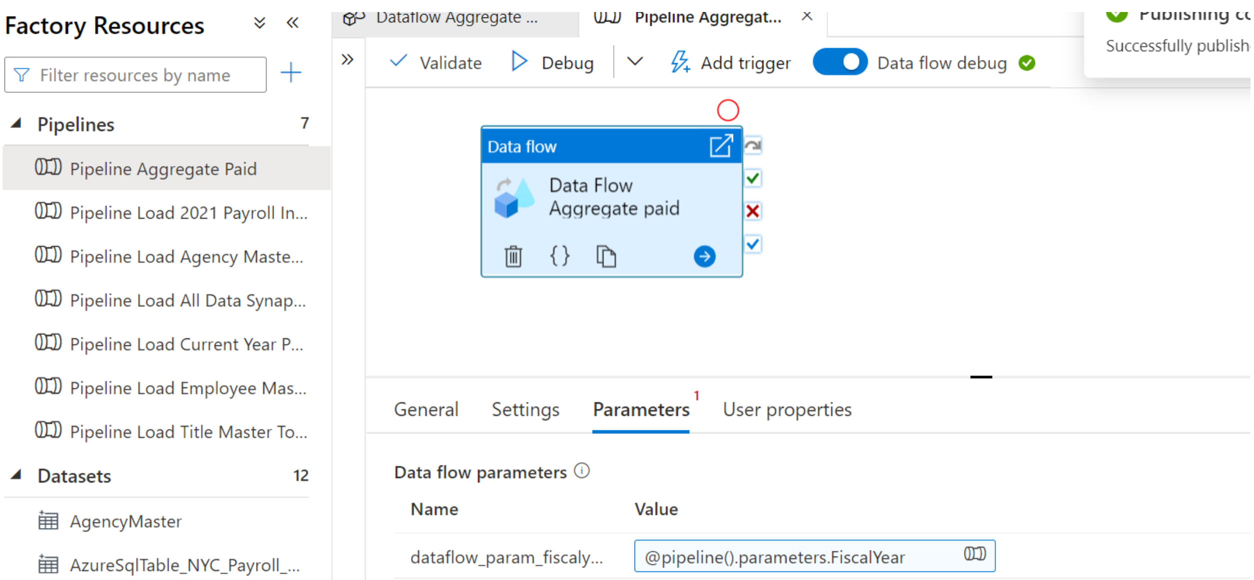


- In Settings, select Truncate Table

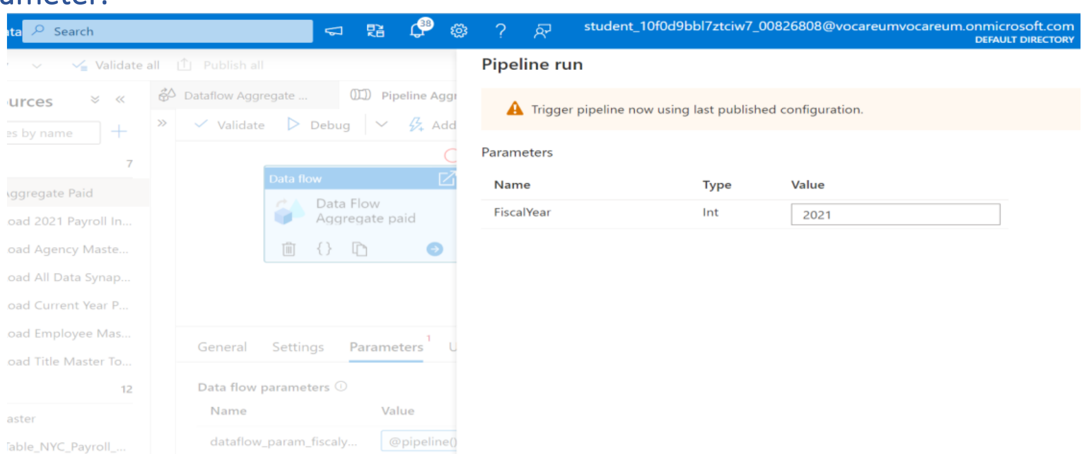


5.9.Create a new Pipeline and add the Aggregate data flow

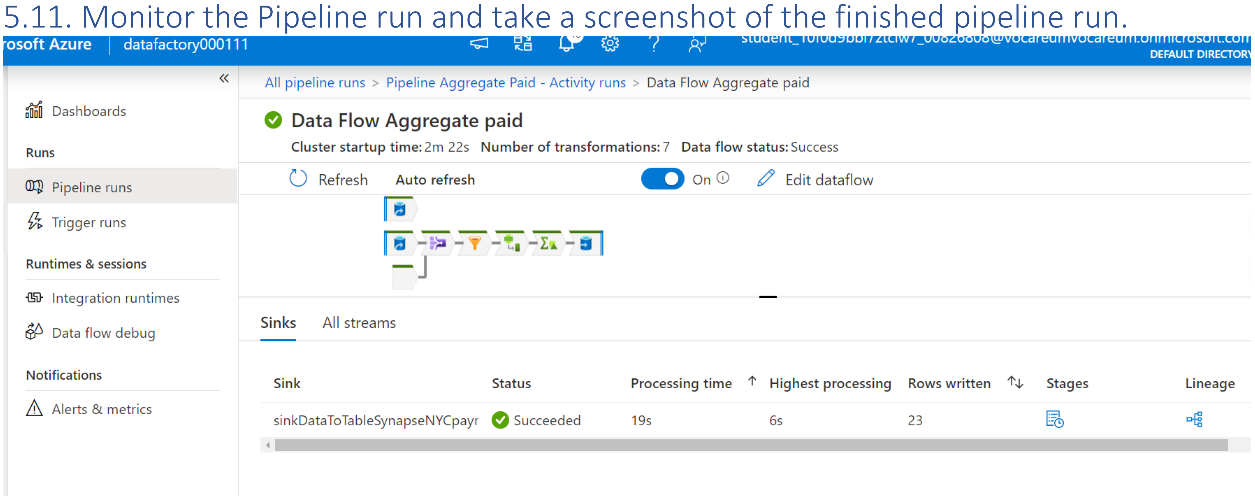
- Create a new Global Parameter (This will be the Parameter at the global pipeline level that will be passed on to the data flow
- In Parameters, select Pipeline Expression
- Choose the parameter created at the Pipeline level

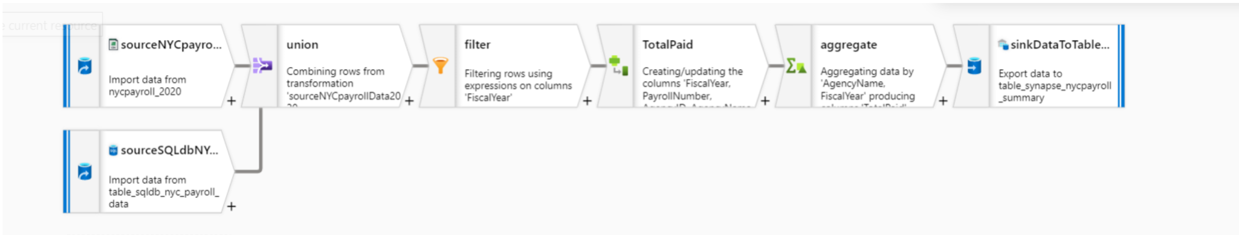


5.10. Validate, Publish and Trigger the pipeline. Enter the desired value for the parameter.



5.11. Monitor the Pipeline run and take a screenshot of the finished pipeline run.





Destination synapse table loaded with summary data