Project Overview

Project Introduction

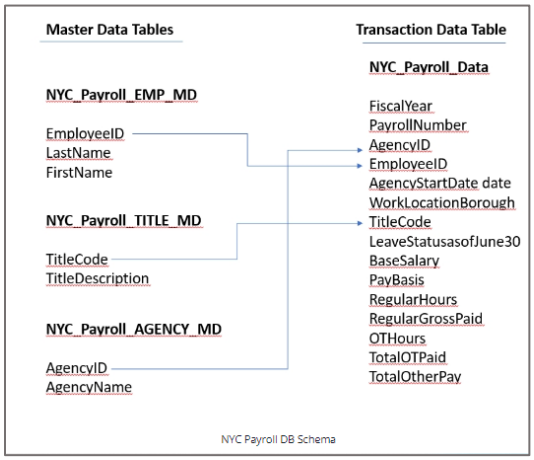
The City of New York would like to develop a Data Analytics platform on Azure Synapse Analytics to accomplish two primary objectives:

1. Analyze how the City's financial resources are allocated and how much of the City's budget is being devoted to overtime.
2. Make the data available to the interested public to show how the City's budget is being spent on salary and overtime pay for all municipal employees.

You have been hired as a Data Engineer to create high-quality data pipelines that are dynamic, can be automated, and monitored for efficient operation. The project team also includes the city's quality assurance experts who will test the pipelines to find any errors and improve overall data quality.

The source data resides in Azure Data Lake and needs to be processed in a NYC data warehouse in Azure Synapse Analytics. The source datasets consist of CSV files with Employee master data and monthly payroll data entered by various City agencies.

NYC Payroll DB Schema



Create and Configure Resources

Project Instructions

For this project, below Azure resources have been mainly utilized:

- Azure Data Lake Gen2
- Azure SQL DB
- Azure Data Factory
- Azure Synapse Analytics

Project Data

4 csv data files were provided for the project.

Name	Date modified	Type	Size
AgencyMaster	10/13/2022 9:06 PM	Microsoft Excel C...	5 KB
EmpMaster	10/13/2022 9:06 PM	Microsoft Excel C...	24 KB
nycpayroll_2020	10/13/2022 9:06 PM	Microsoft Excel C...	18 KB
nycpayroll_2021	10/13/2022 9:06 PM	Microsoft Excel C...	17 KB
TitleMaster	10/13/2022 9:06 PM	Microsoft Excel C...	50 KB

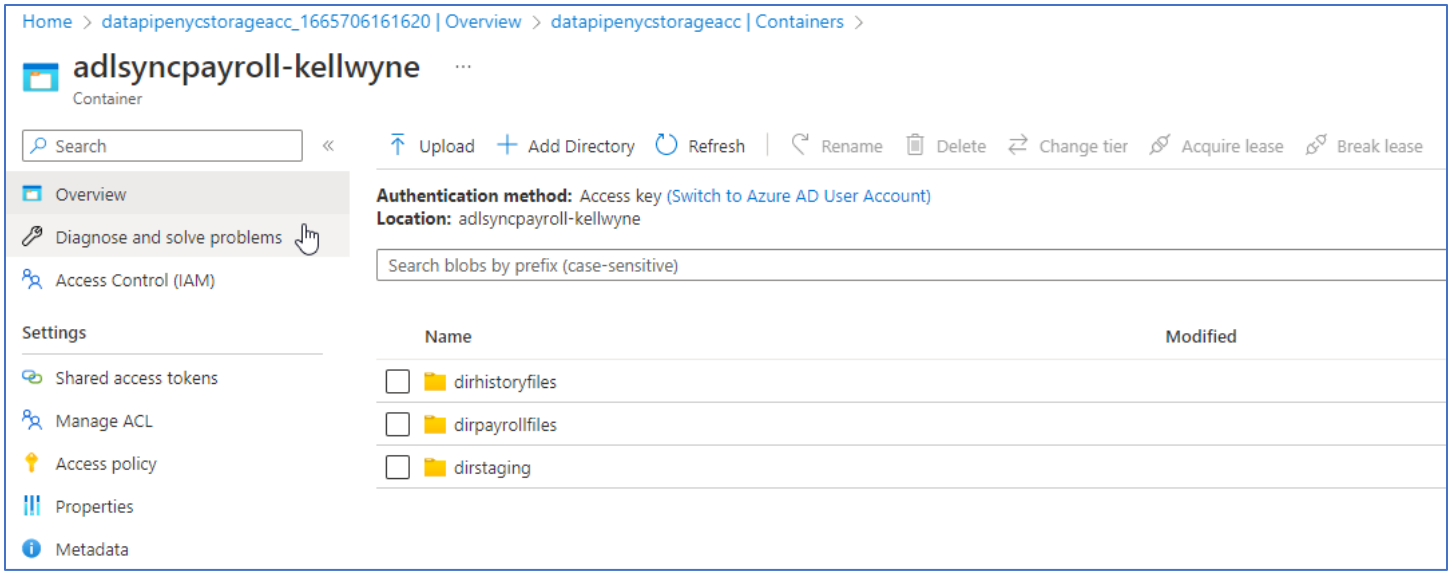
Step 1: Prepare the Data Infrastructure

Setup Data and Resources in Azure

1.Create the data lake and upload data

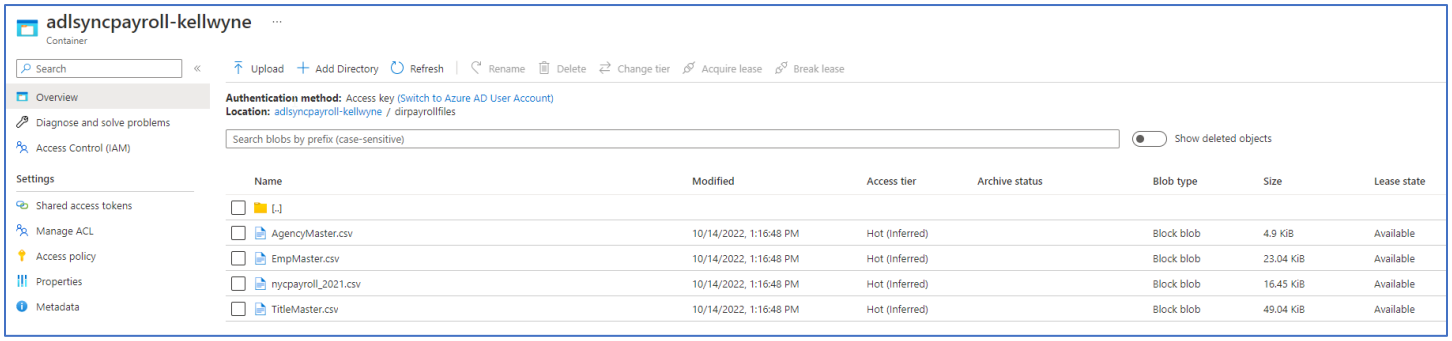
Created an Azure Data Lake Storage Gen2 (storage account) and associated storage container resource named **adlsnycpayroll-yourfirstname-lastintial**. Create three directories in this storage container named

- **dirpayrollfiles**
- **dirhistoryfiles**
- **dirstaging**



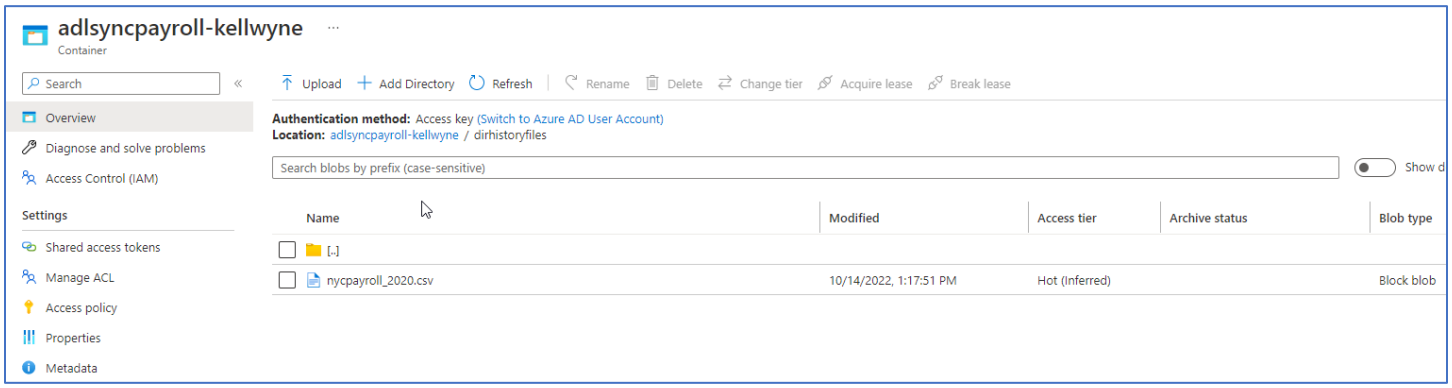
Upload the files to the **dirpayrollfiles** folder

- EmpMaster.csv
- AgencyMaster.csv
- TitleMaster.csv
- nycpayroll_2021.csv

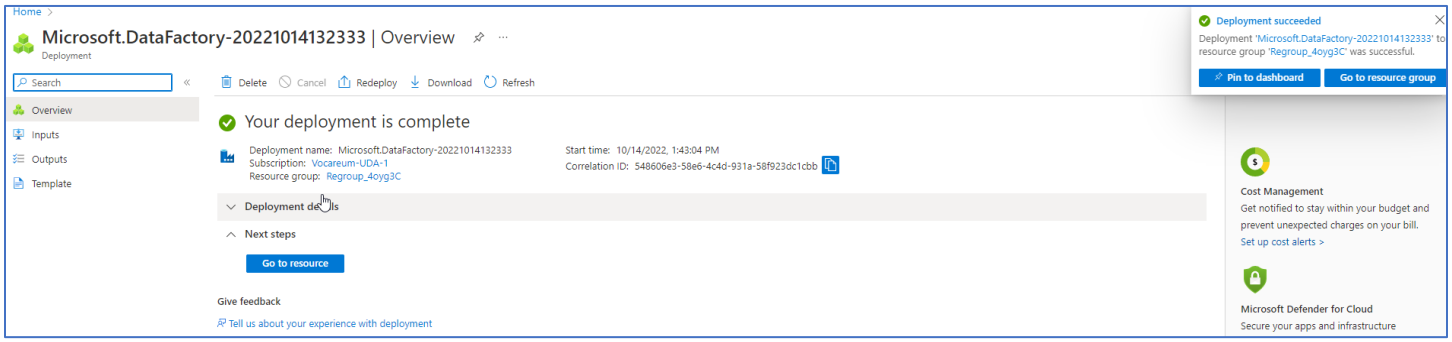


Upload this file (historical data) to the **dirhistoryfiles** folder

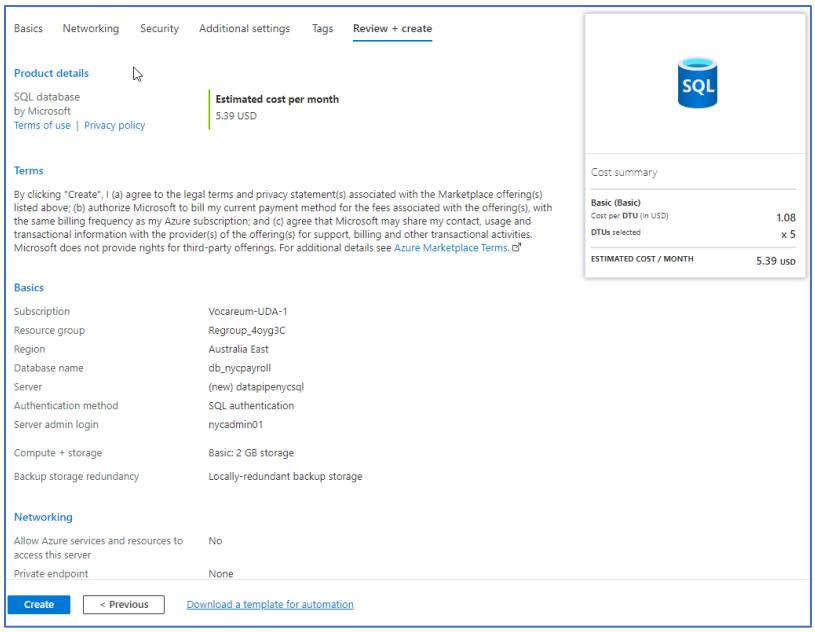
- nycpayroll_2020.csv



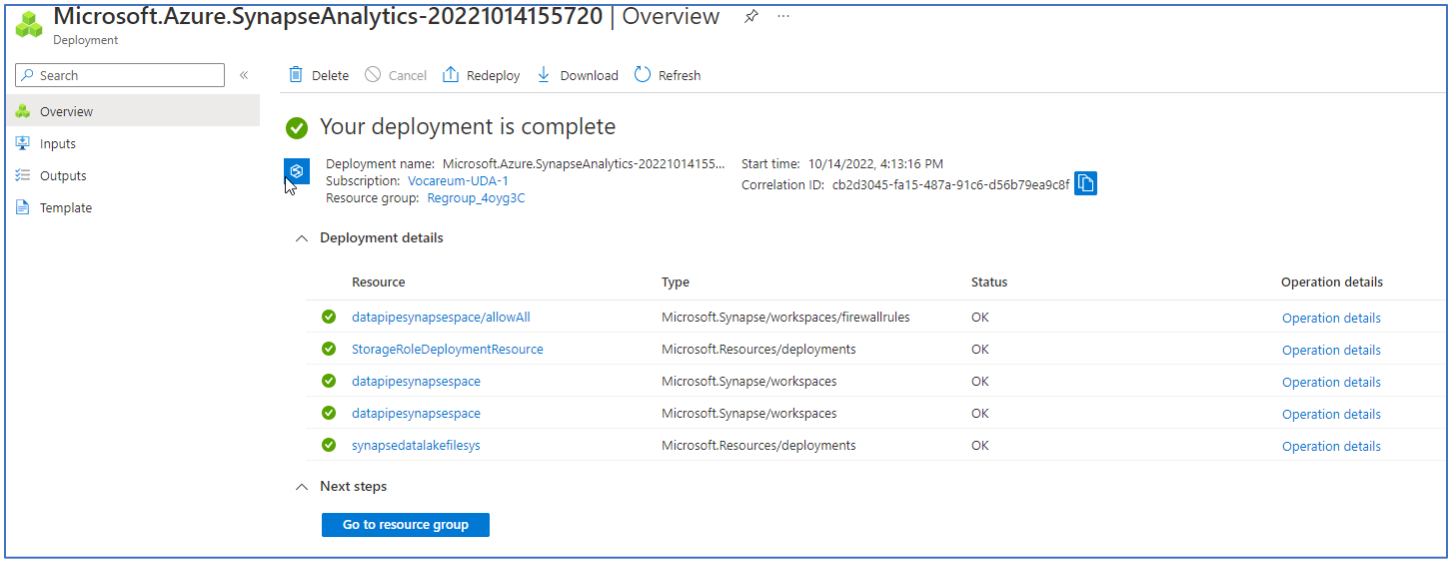
2. Create an Azure Data Factory Resource

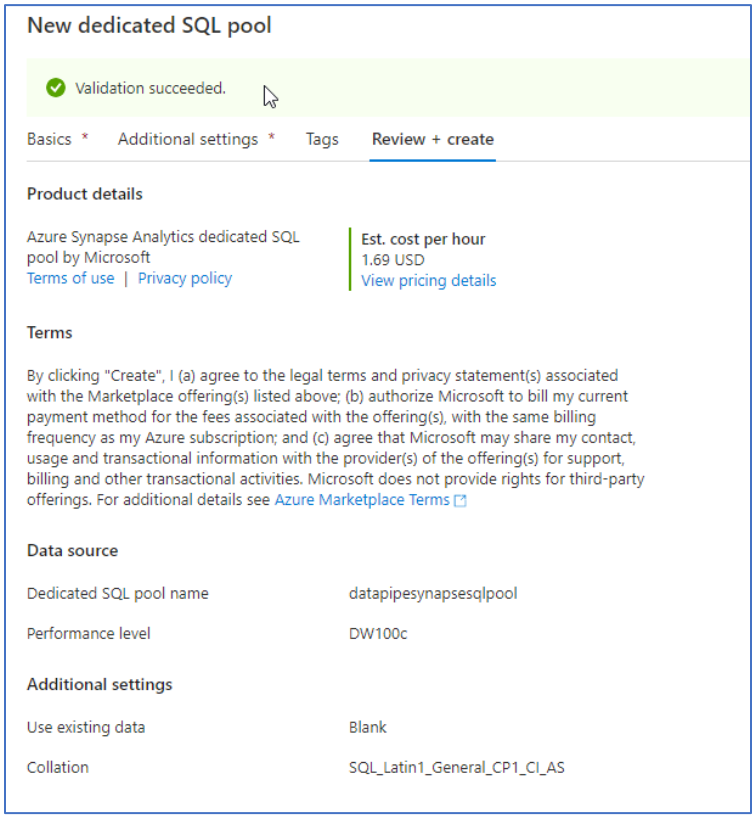


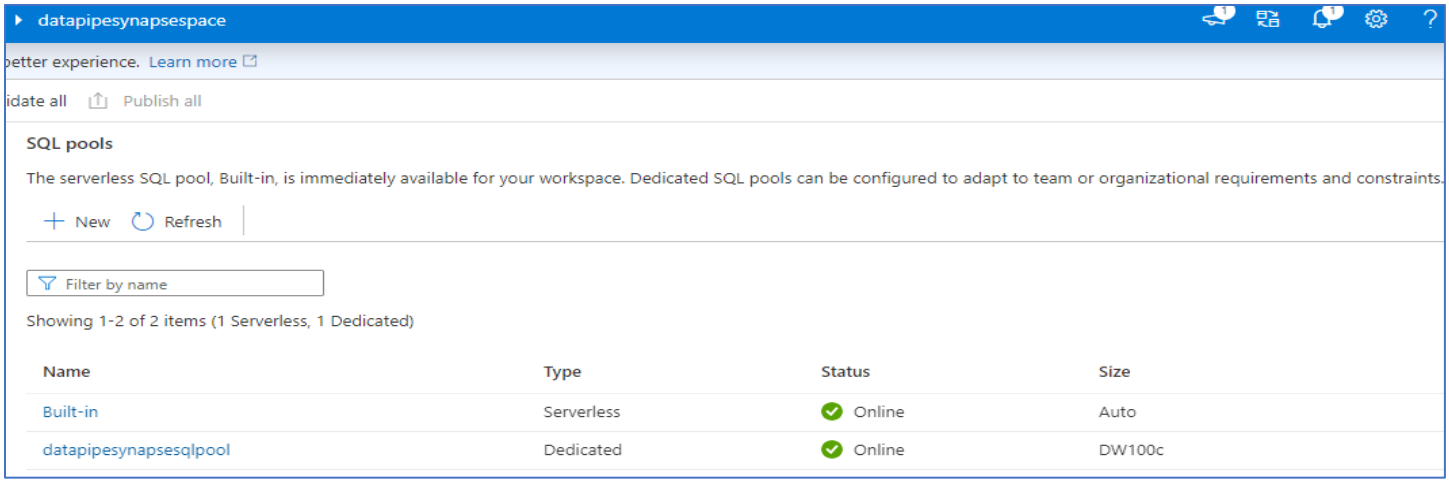
3. Create a SQL Database to store the current year of the payroll data



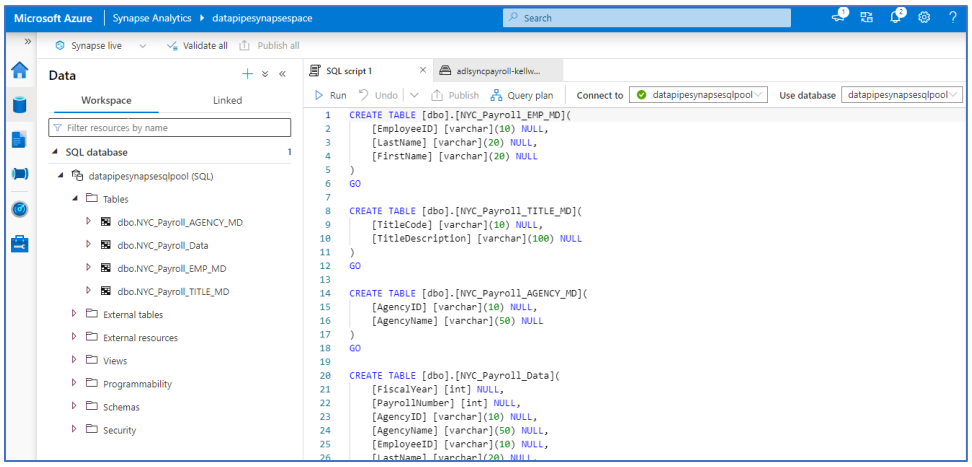
4. Create A Synapse Analytics workspace, SQL dedicated pool in the Synapse Analytics workspace.

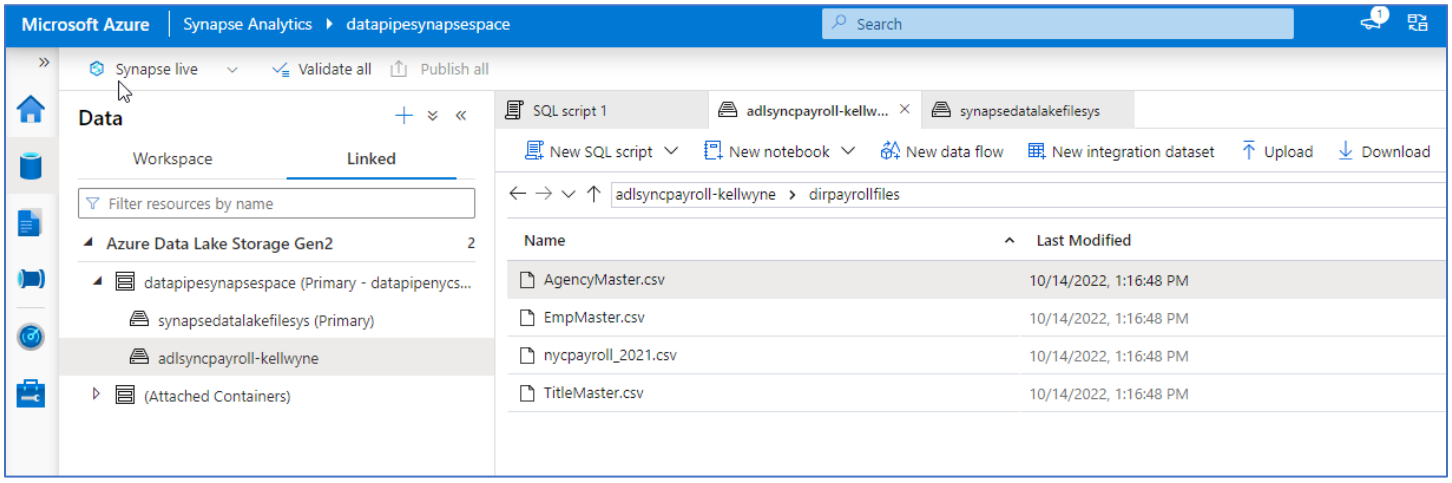




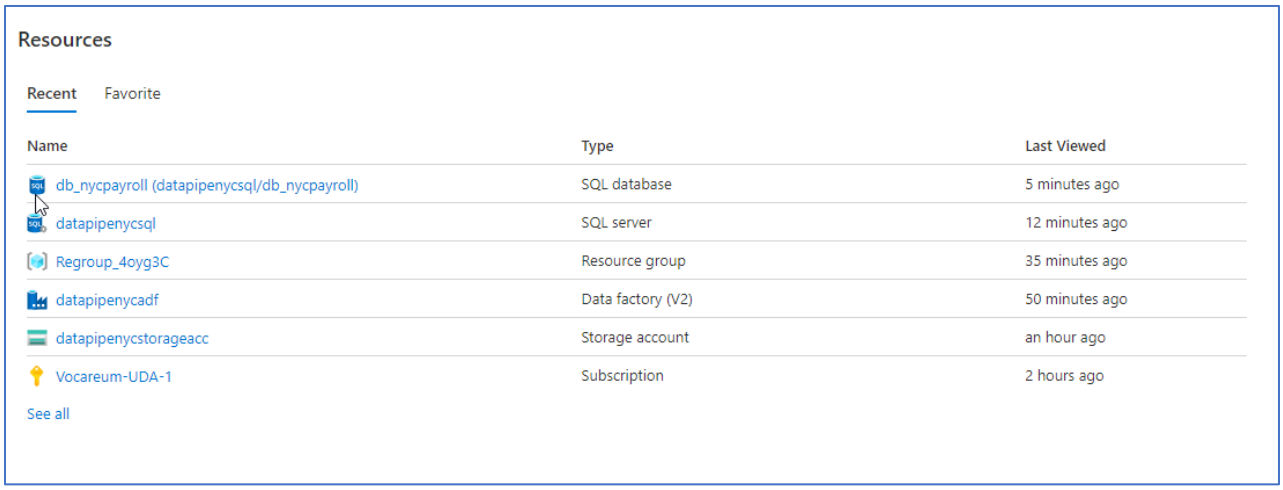


Create master data tables and payroll transaction tables:





Azure Resources Completed:

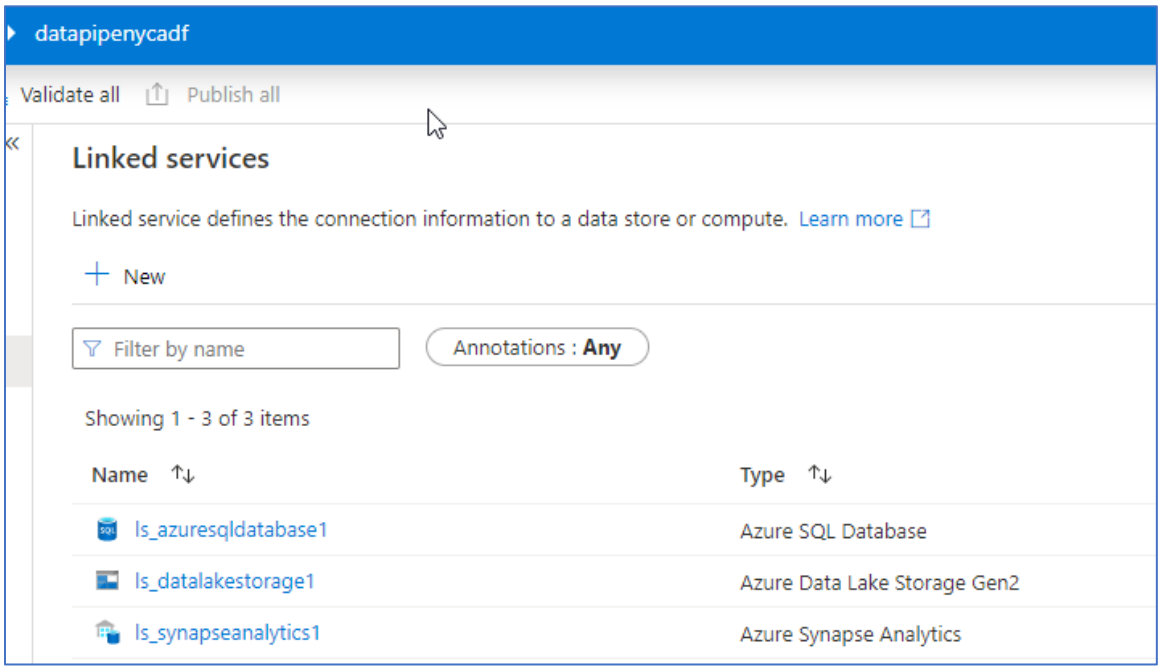


Create Linked Services

1.Create a Linked Service for Azure Data Lake

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

3. Create a Linked Service for Synapse Analytics



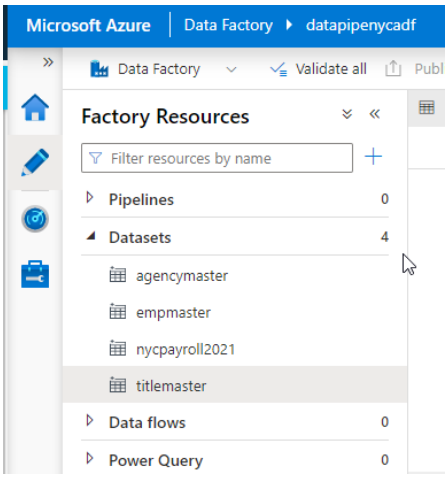
Create Datasets

Step 3: Create Datasets in Azure Data Factory

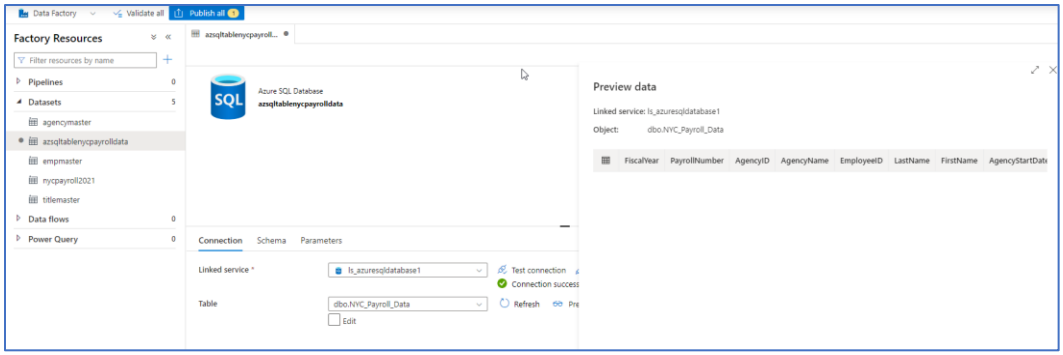
1.Create the datasets for the 2021 Payroll file on Azure Data Lake Gen2

2. Create datasets for the rest of the data files in the Data Lake

- EmpMaster.csv
- TitleMaster.csv
- AgencyMaster.csv

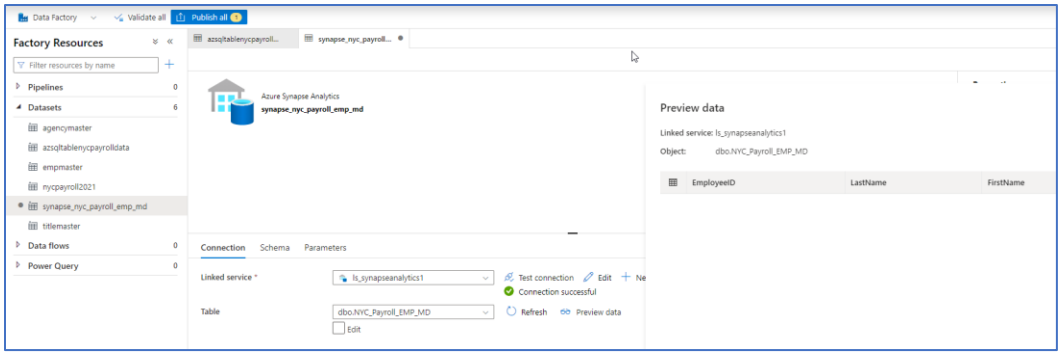


3. Create the dataset for transaction data table that should contain current (2021) data in SQL DB



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

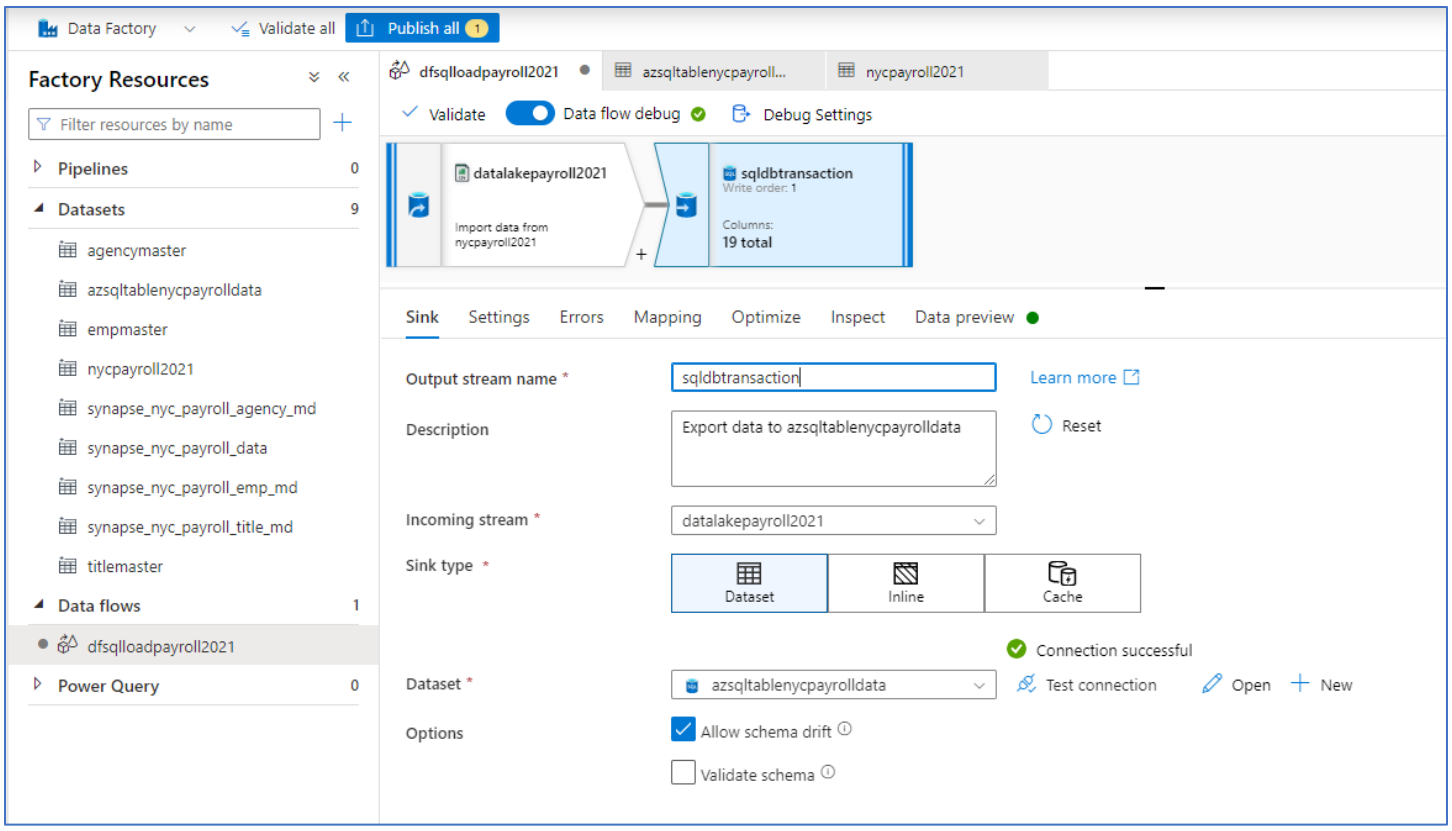
- dataset for NYC_Payroll_EMP_MD
- for NYC_Payroll_TITLE_MD
- for NYC_Payroll_AGENCY_MD
- for NYC_Payroll_Data

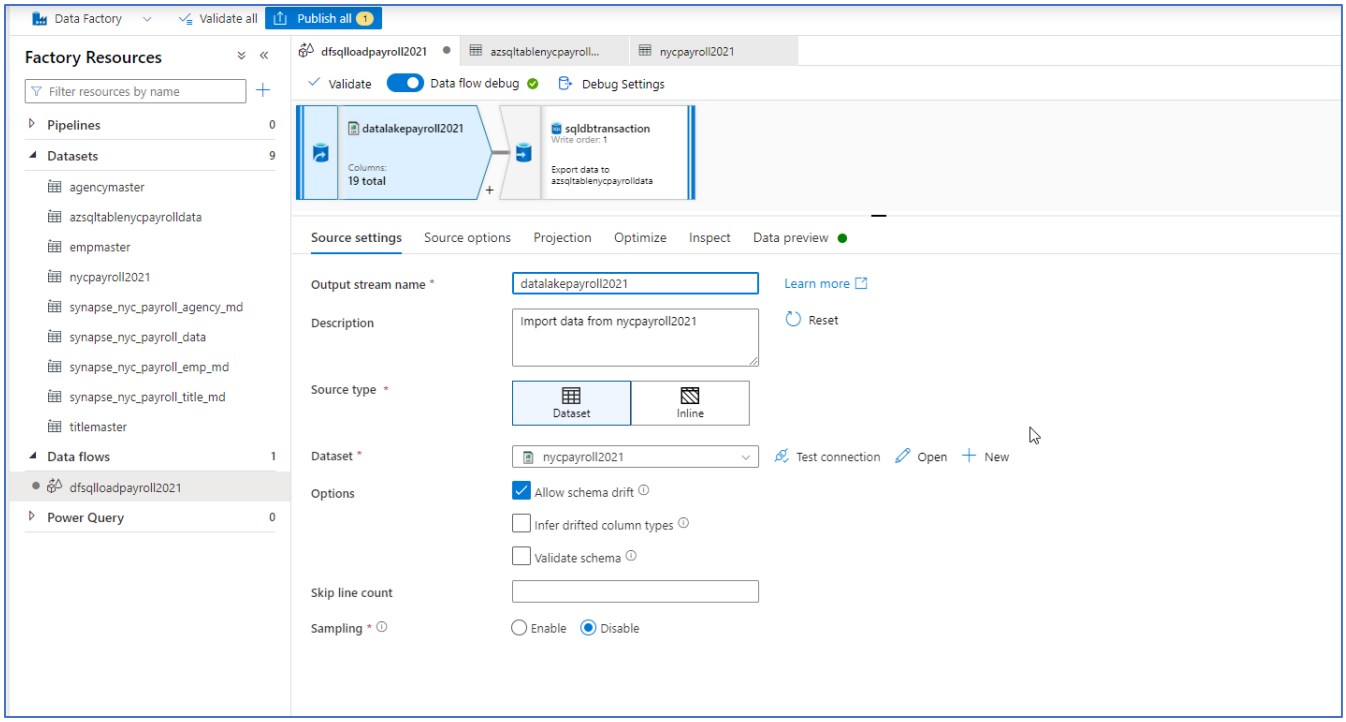


Create Dataflows and Pipelines

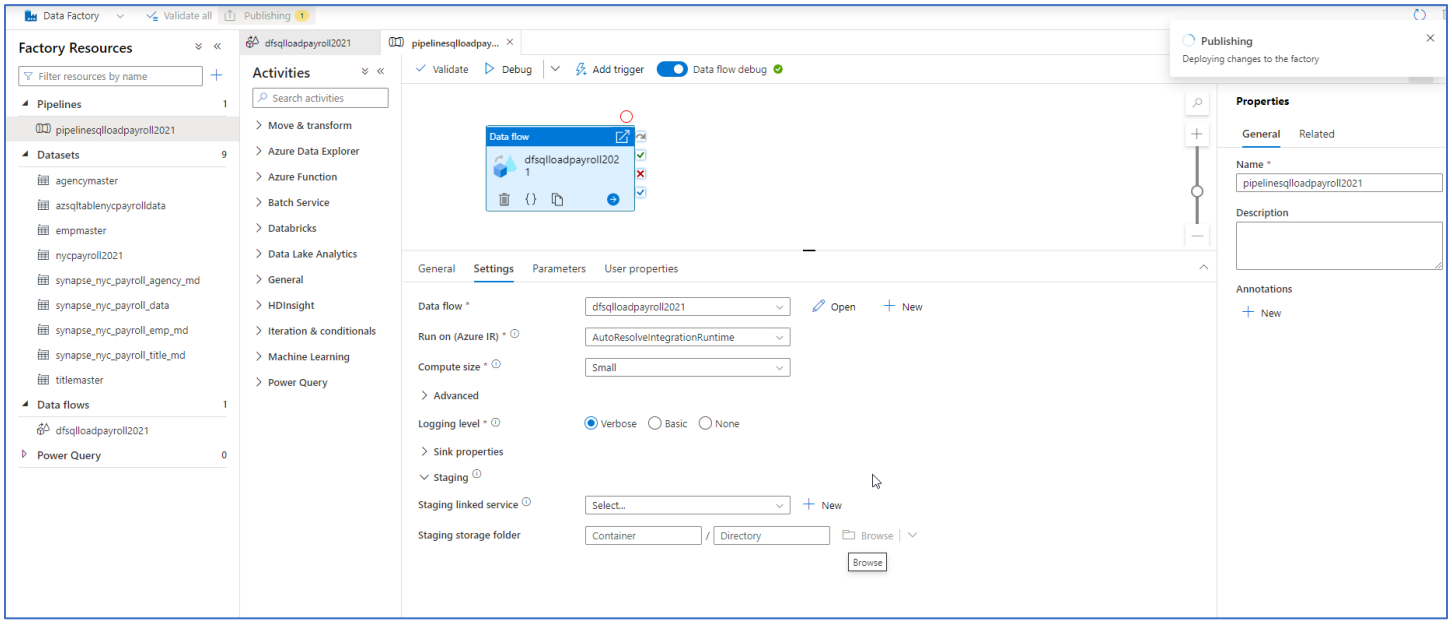
Step 4: Create Data Flows

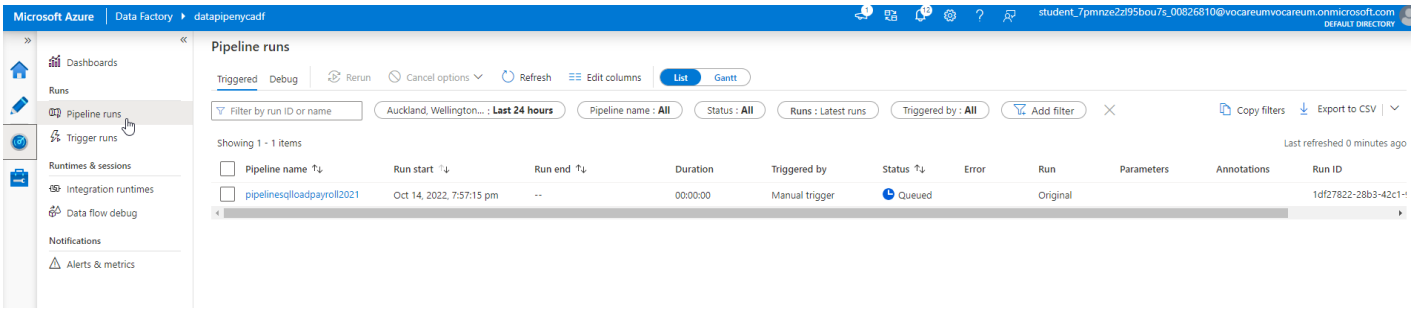
1.In Azure Data Factory, created the data flow to load 2021 Payroll Data to SQL DB transaction table (in the future NYC will load all the transaction data into this table).

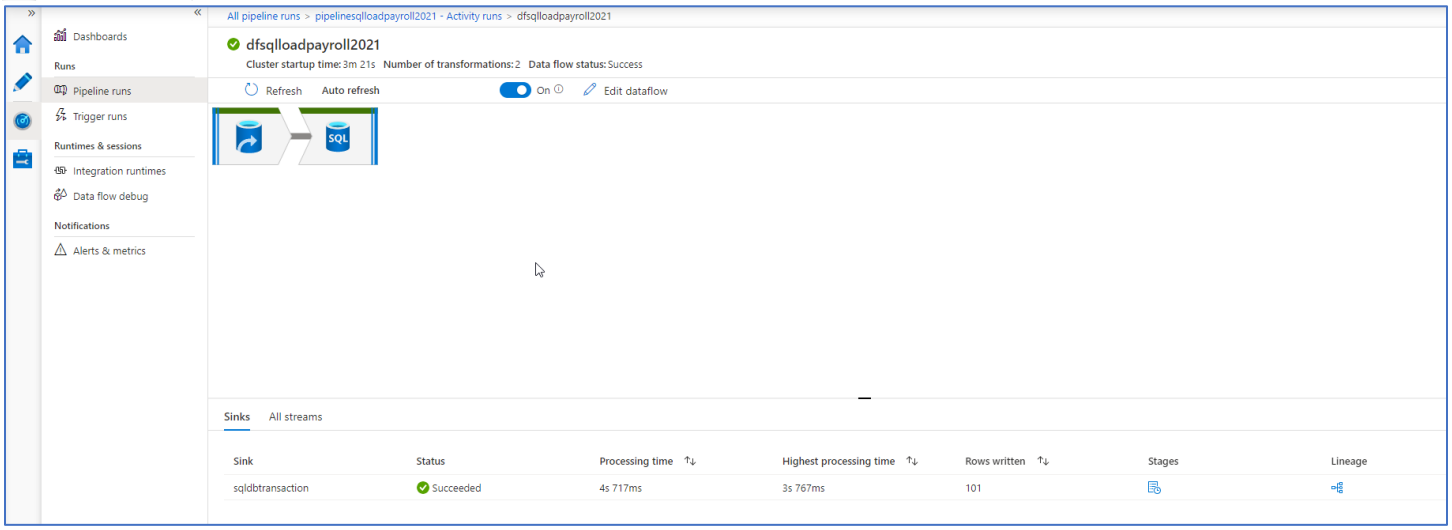


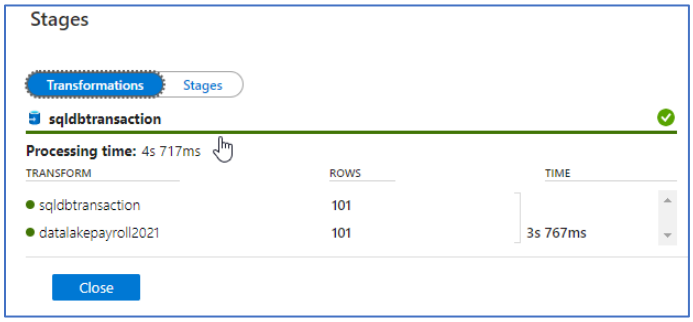


2.Create Pipeline to load 2021 Payroll data into transaction table in the SQL DB



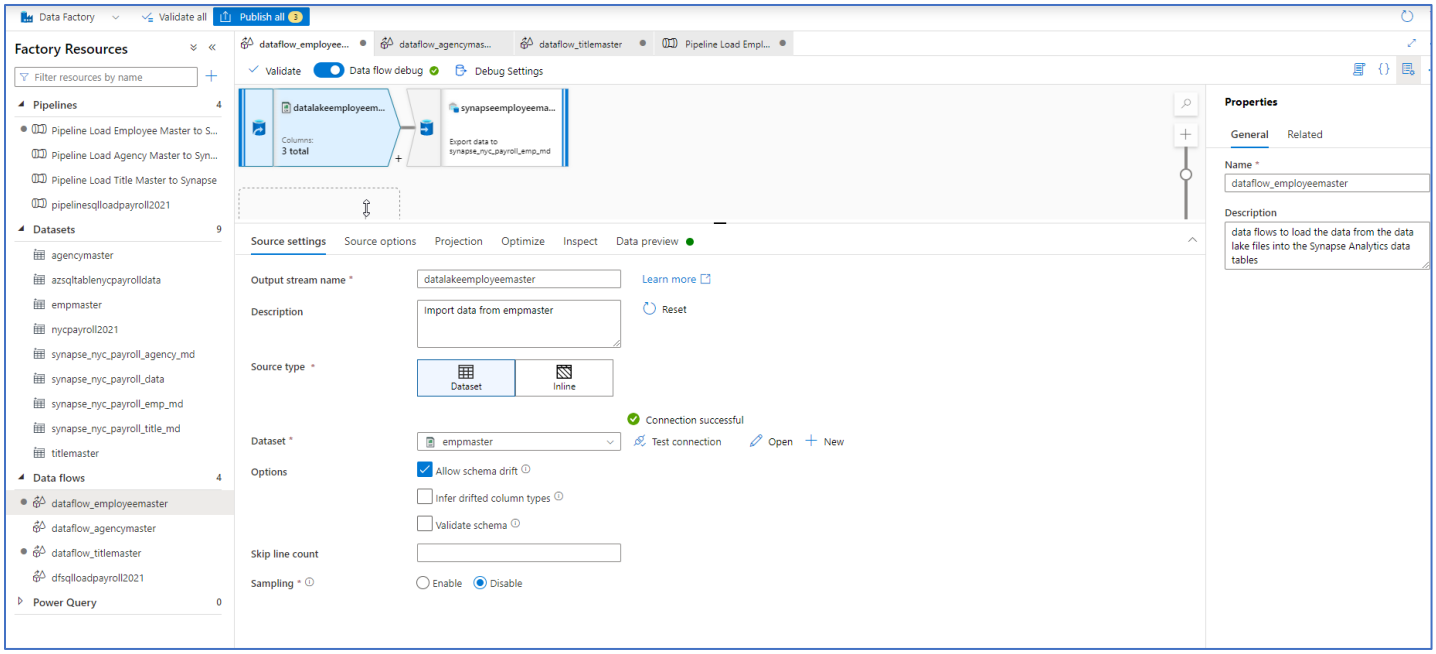


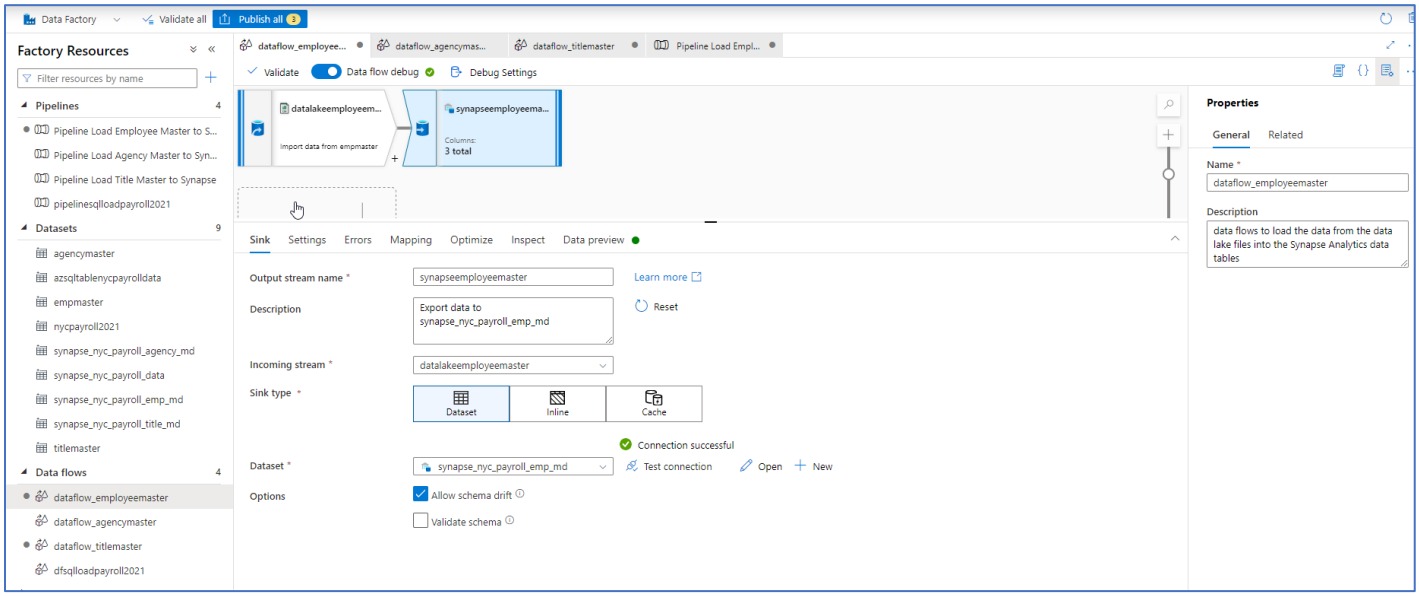




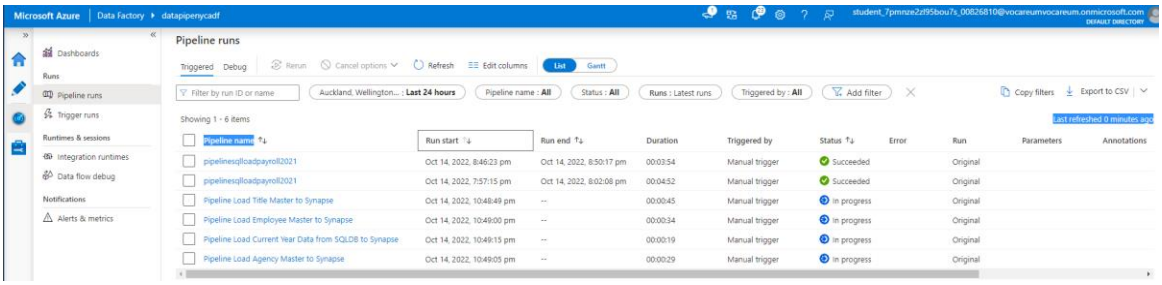
3. Create data flows to load the Employee, Title, and Agency data files from the data lake files into the Synapse Analytics data tables

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

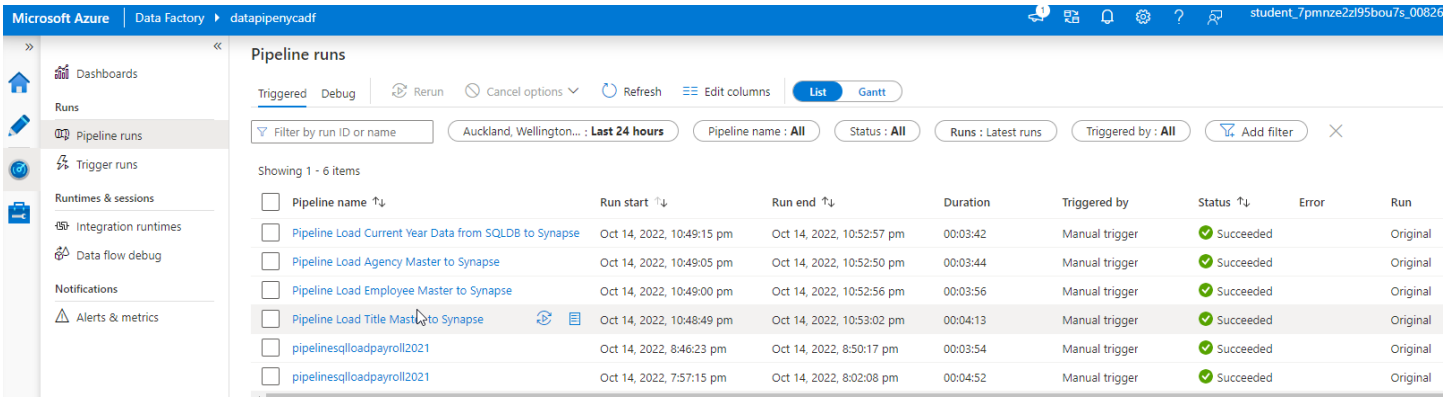


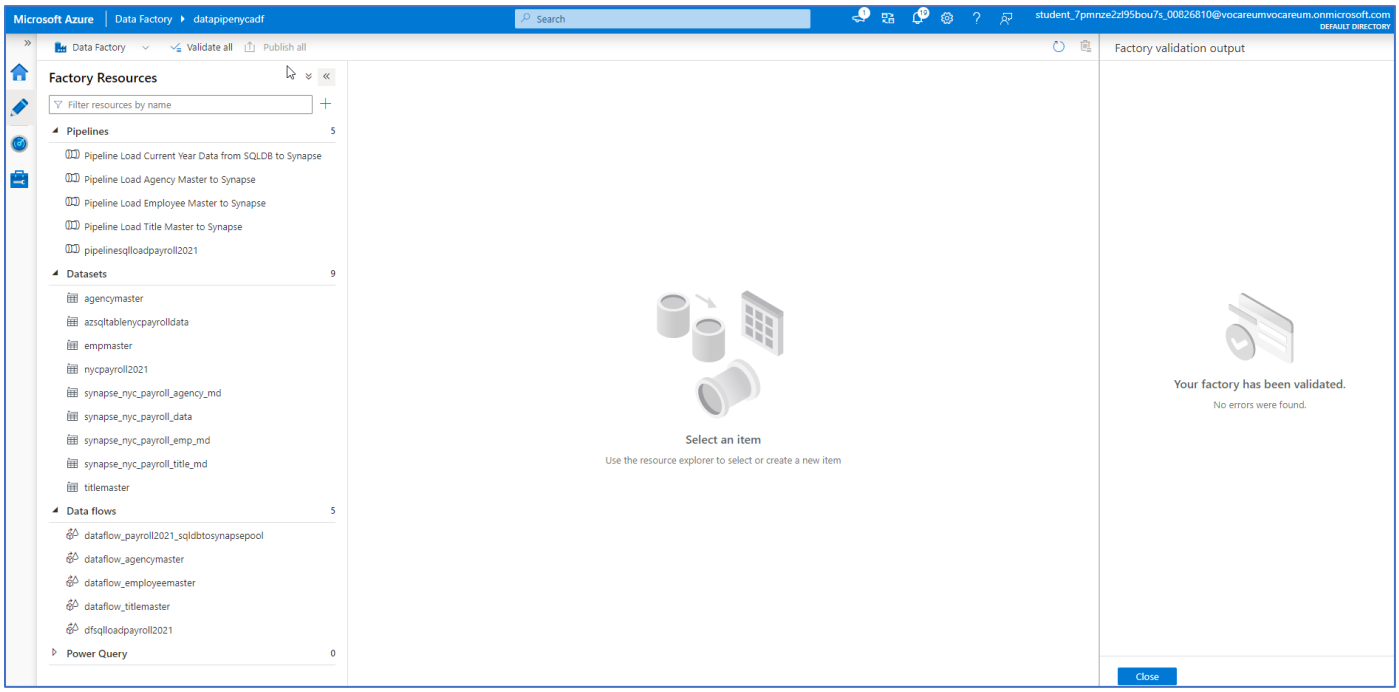


5. Create pipelines for Employee, Title, Agency, and year 2021 Payroll transaction data to Synapse Analytics containing the data flows.

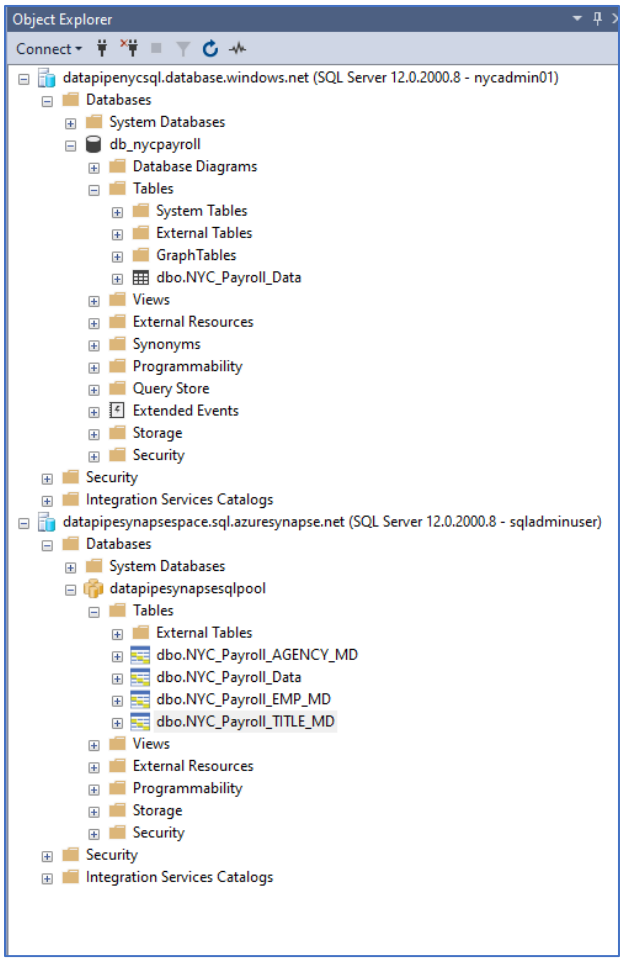


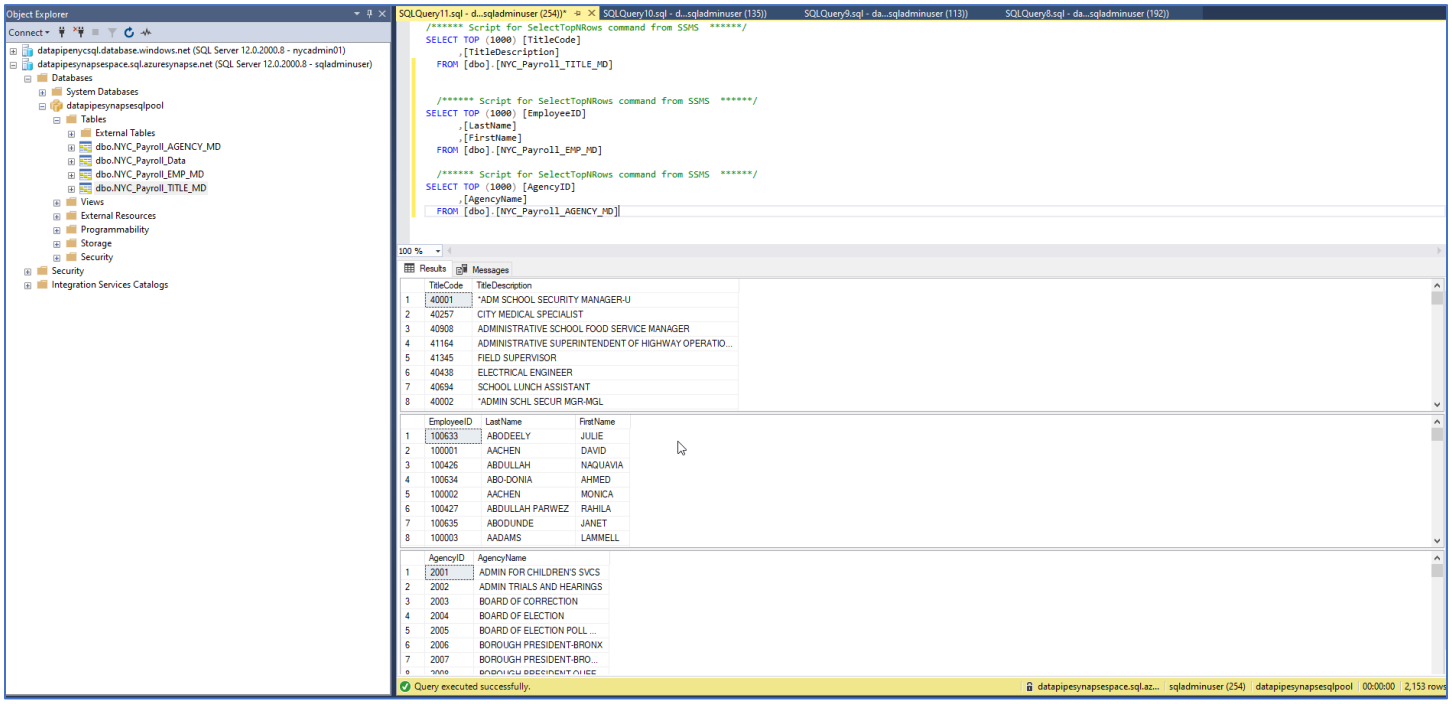
6. Trigger and monitor the Pipelines

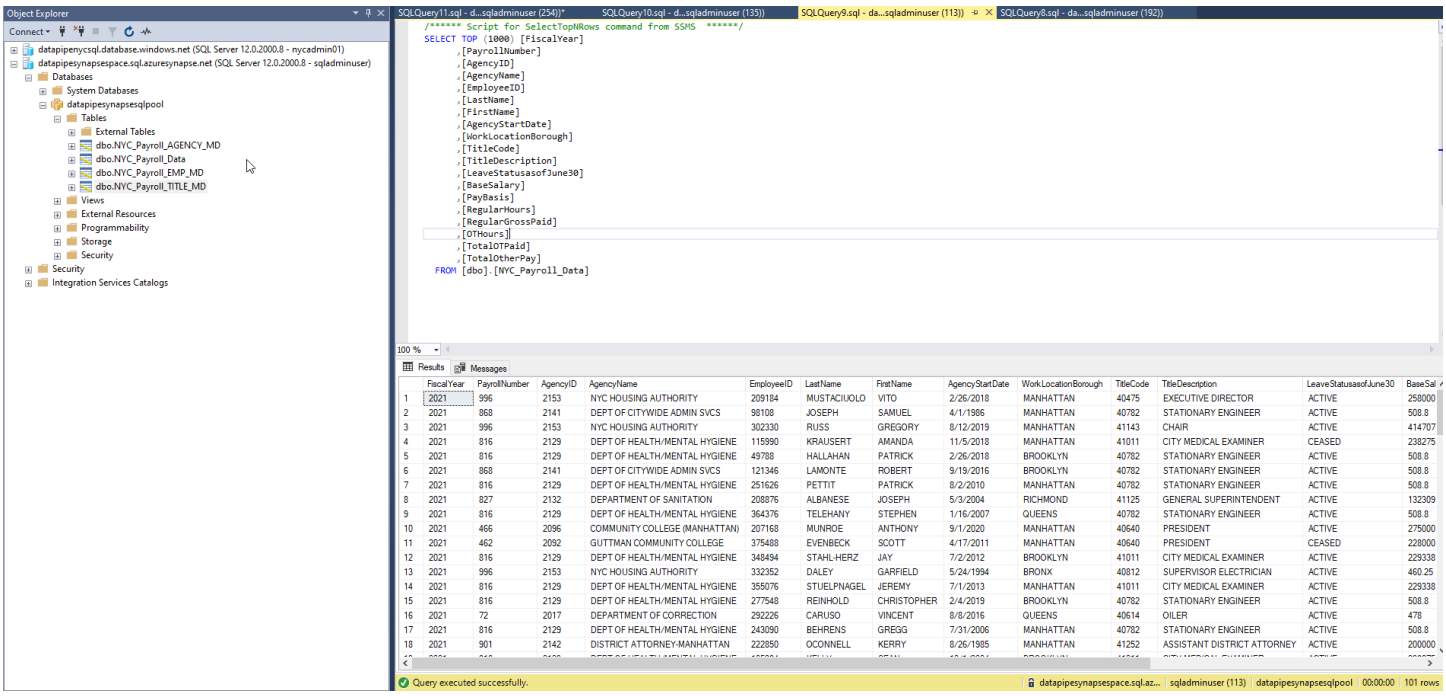




Connected to SQL resources below to confirm data has been copied after pipeline runs.







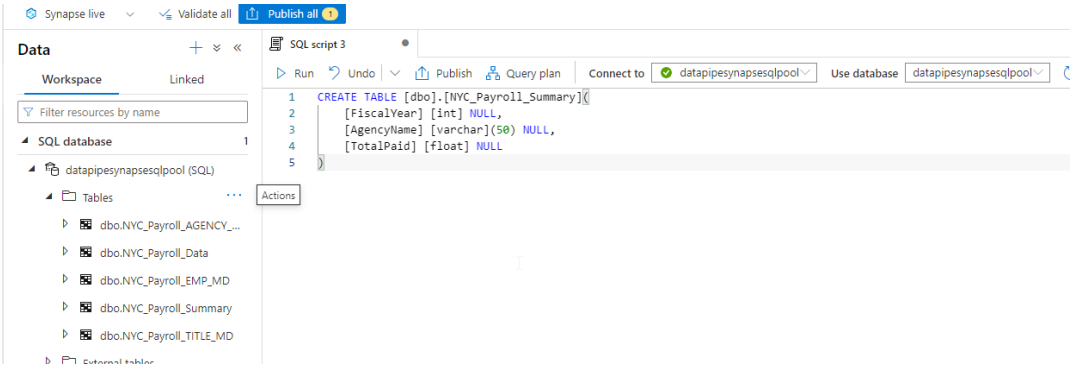
	FiscalYear	PayrollNumber	AgencyID	AgencyName	EmployeeID	LastName	FirstName	AgencyStartDate	WorkLocationBorough	TitleCode	TitleDescription	LeaveStatusasofJune30	BaseSal
1	2021	996	2153	NYC HOUSING AUTHORITY	209184	MUSTACIUOLO	VITO	2/26/2018	MANHATTAN	40475	EXECUTIVE DIRECTOR	ACTIVE	258000
2	2021	868	2141	DEPT OF CITYWIDE ADMIN SVCS	98108	JOSEPH	SAMUEL	4/1/1986	MANHATTAN	40782	STATIONARY ENGINEER	ACTIVE	508.8
3	2021	996	2153	NYC HOUSING AUTHORITY	302330	RUSS	GREGORY	8/12/2019	MANHATTAN	41143	CHAIR	ACTIVE	414707
4	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	115990	KRAUSERT	AMANDA	11/5/2018	MANHATTAN	41011	CITY MEDICAL EXAMINER	CEASED	238275
5	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	49788	HALLAHAN	PATRICK	2/26/2018	BROOKLYN	40782	STATIONARY ENGINEER	ACTIVE	508.8
6	2021	868	2141	DEPT OF CITYWIDE ADMIN SVCS	121346	LAMONTE	ROBERT	9/19/2016	BROOKLYN	40782	STATIONARY ENGINEER	ACTIVE	508.8
7	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	251626	PETTIT	PATRICK	8/2/2010	MANHATTAN	40782	STATIONARY ENGINEER	ACTIVE	508.8
8	2021	827	2132	DEPARTMENT OF SANITATION	208876	ALBANESE	JOSEPH	5/3/2004	RICHMOND	41125	GENERAL SUPERINTENDENT	ACTIVE	132309
9	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	364376	TELEHANY	STEPHEN	1/16/2007	QUEENS	40782	STATIONARY ENGINEER	ACTIVE	508.8
10	2021	466	2096	COMMUNITY COLLEGE (MANHATTAN)	207168	MUNROE	ANTHONY	9/1/2020	MANHATTAN	40640	PRESIDENT	ACTIVE	275000
11	2021	462	2092	GUTTMAN COMMUNITY COLLEGE	375488	EVENBECK	SCOTT	4/17/2011	MANHATTAN	40640	PRESIDENT	CEASED	228000
12	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	348494	STAHL-HERZ	JAY	7/2/2012	BROOKLYN	41011	CITY MEDICAL EXAMINER	ACTIVE	229338
13	2021	996	2153	NYC HOUSING AUTHORITY	332352	DALEY	GARFIELD	5/24/1994	BRONX	40812	SUPERVISOR ELECTRICIAN	ACTIVE	460.25
14	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	355076	STUELPNAGEL	JEREMY	7/1/2013	MANHATTAN	41011	CITY MEDICAL EXAMINER	ACTIVE	229338
15	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	277548	REINHOLD	CHRISTOPHER	2/4/2019	BROOKLYN	40782	STATIONARY ENGINEER	ACTIVE	508.8
16	2021	72	2017	DEPARTMENT OF CORRECTION	292226	CARUSO	VINCENT	8/8/2016	QUEENS	40614	OILER	ACTIVE	478
17	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	243090	BEHRENS	GREGG	7/31/2006	MANHATTAN	40782	STATIONARY ENGINEER	ACTIVE	508.8
18	2021	901	2142	DISTRICT ATTORNEY-MANHATTAN	222850	OCONNELL	KERRY	8/26/1985	MANHATTAN	41252	ASSISTANT DISTRICT ATTORNEY	ACTIVE	200000

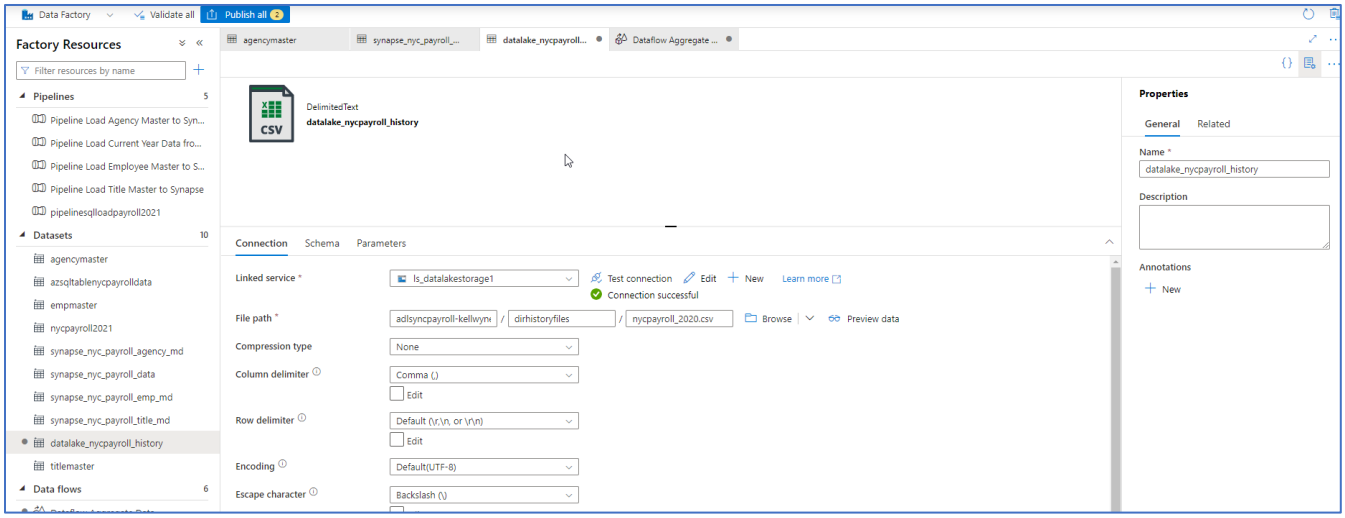
Aggregate Data Flow

Step 5: Data Aggregation and Parameterization

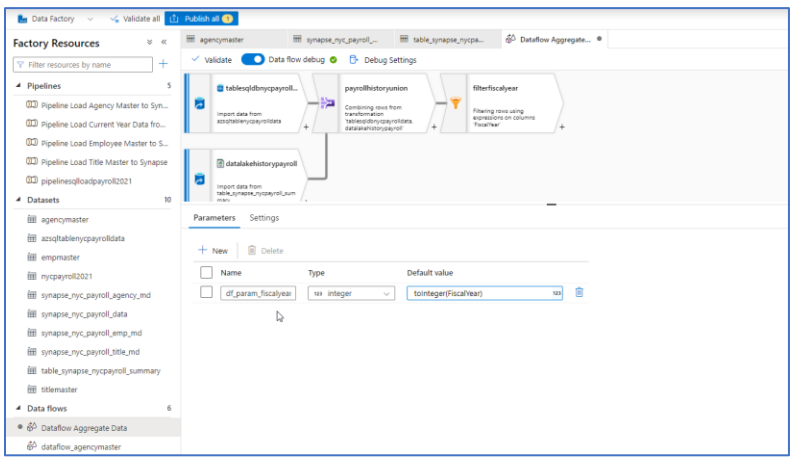
extract the 2021 year data and historical data, merge, aggregate and store it in Synapse Analytics. The aggregation will be on Agency Name, Fiscal Year and TotalPaid.

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

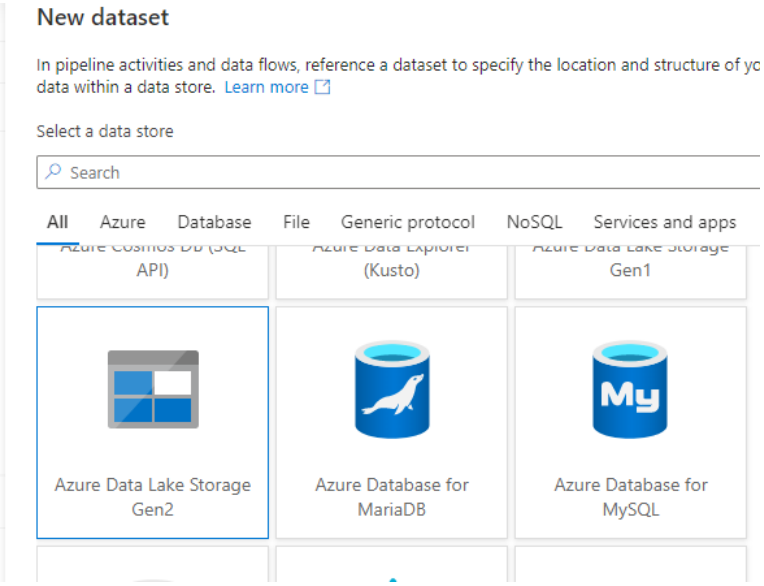


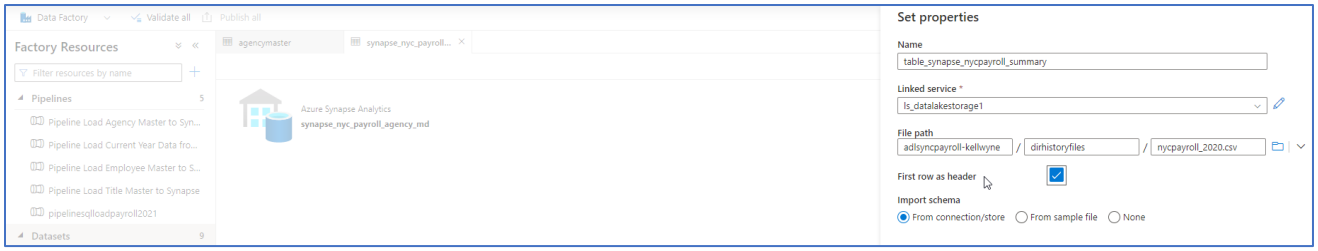


Setup dataflow parameter:

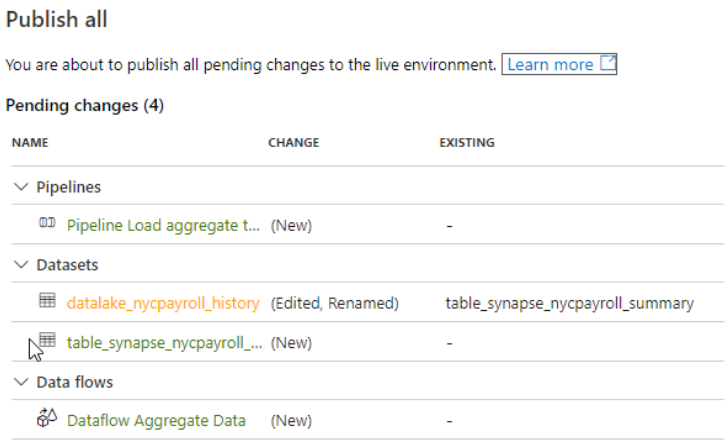


2.Create a new dataset for the Azure Data Lake Gen2 folder that contains the historical files.



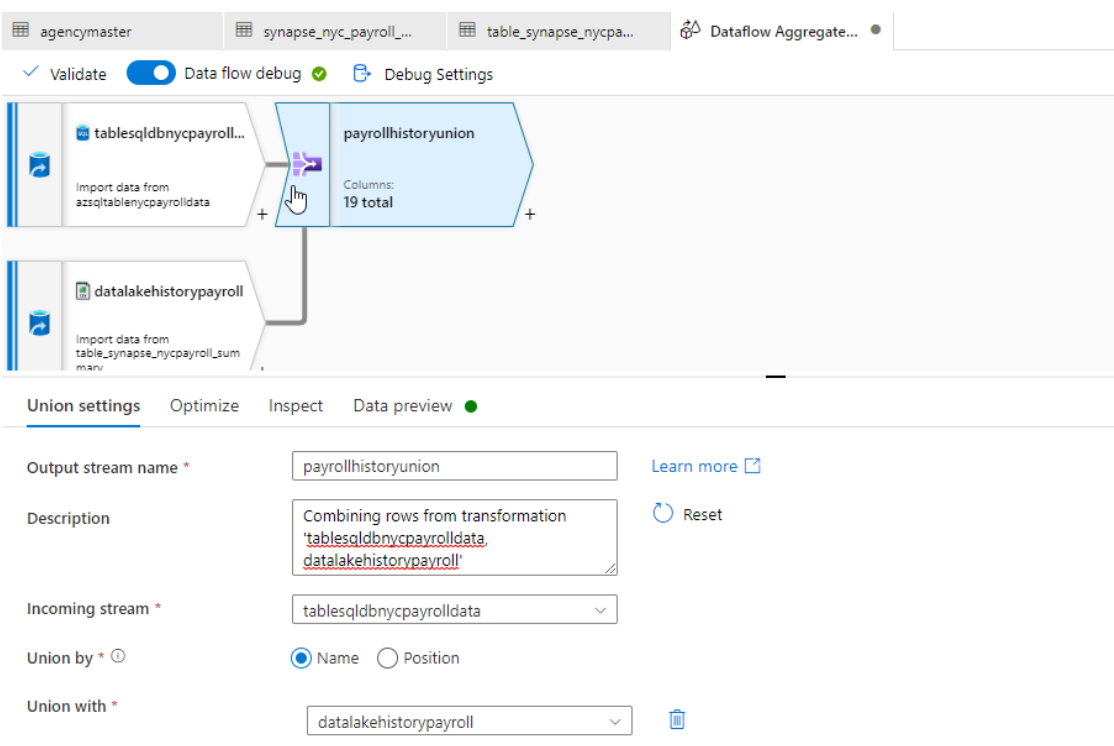


Correction had to rename this dataset as below,since there was a synapse dataset required with that name as per point 1 :

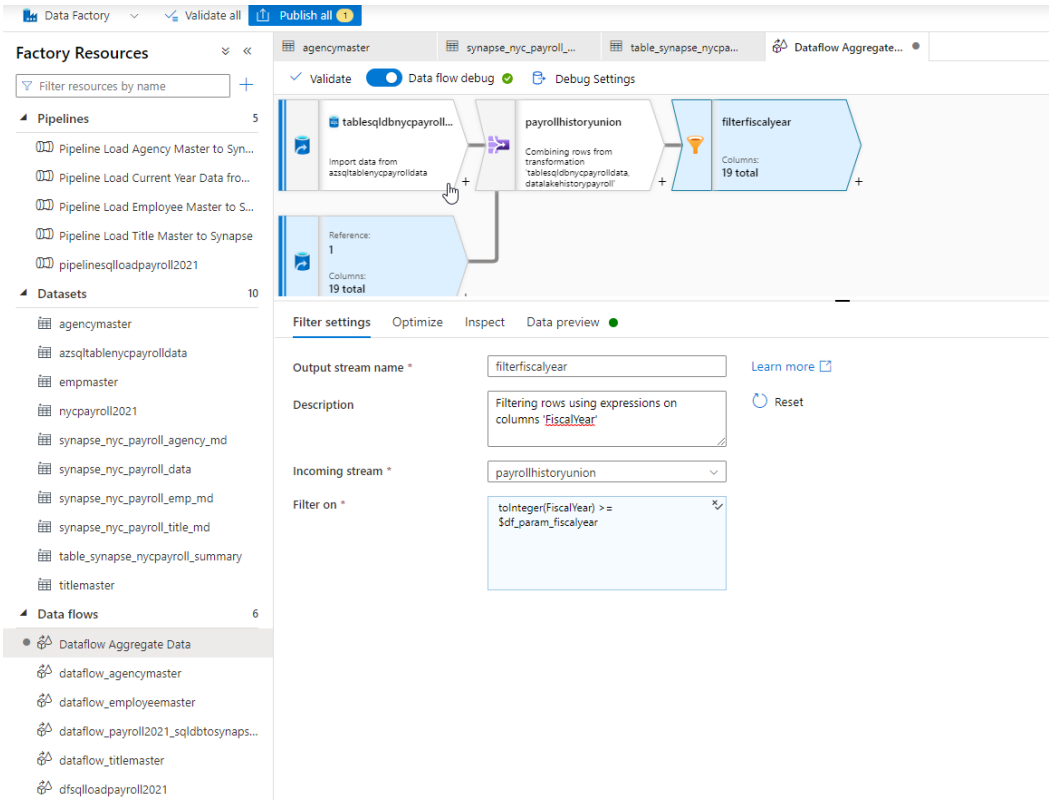


3.Create new data flow for the activity - Dataflow Aggregate Data

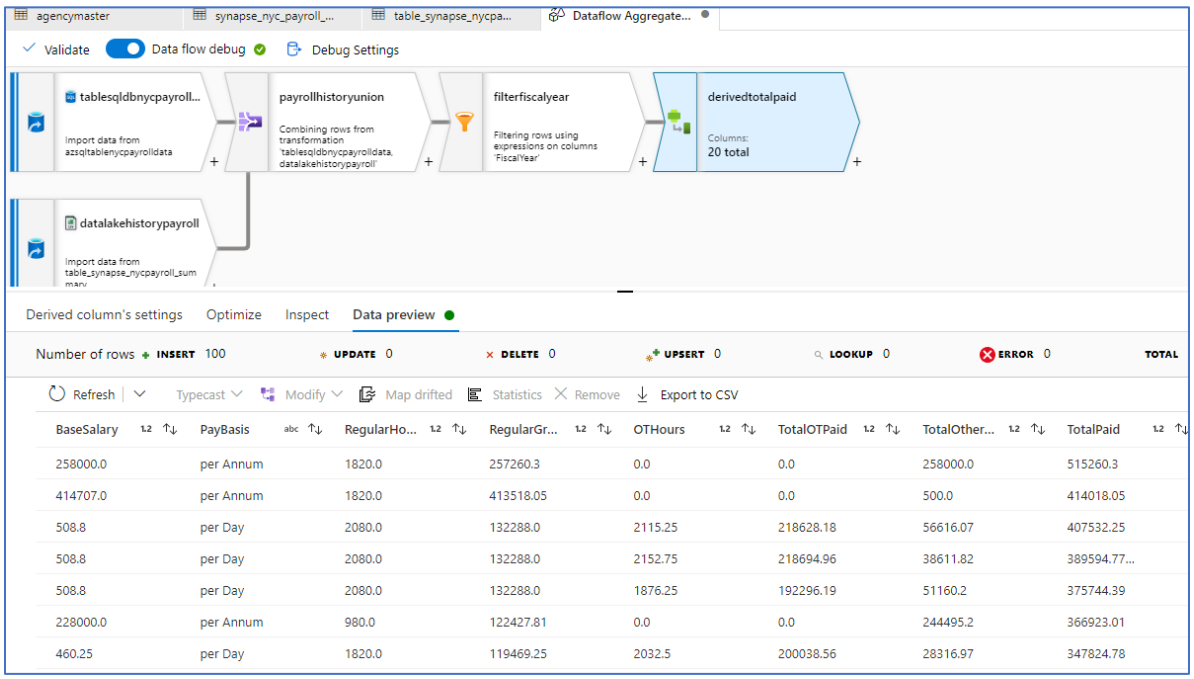
4.Create a new Union activity in the data flow and Union with history files



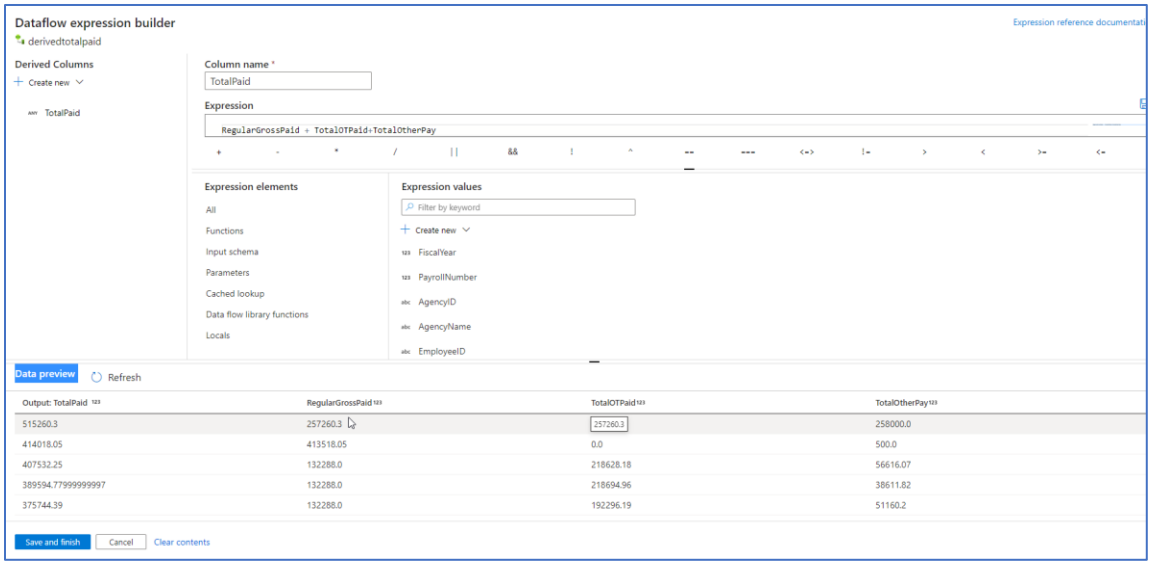
5.Add a Filter activity after Union



6.Derive a new TotalPaid column

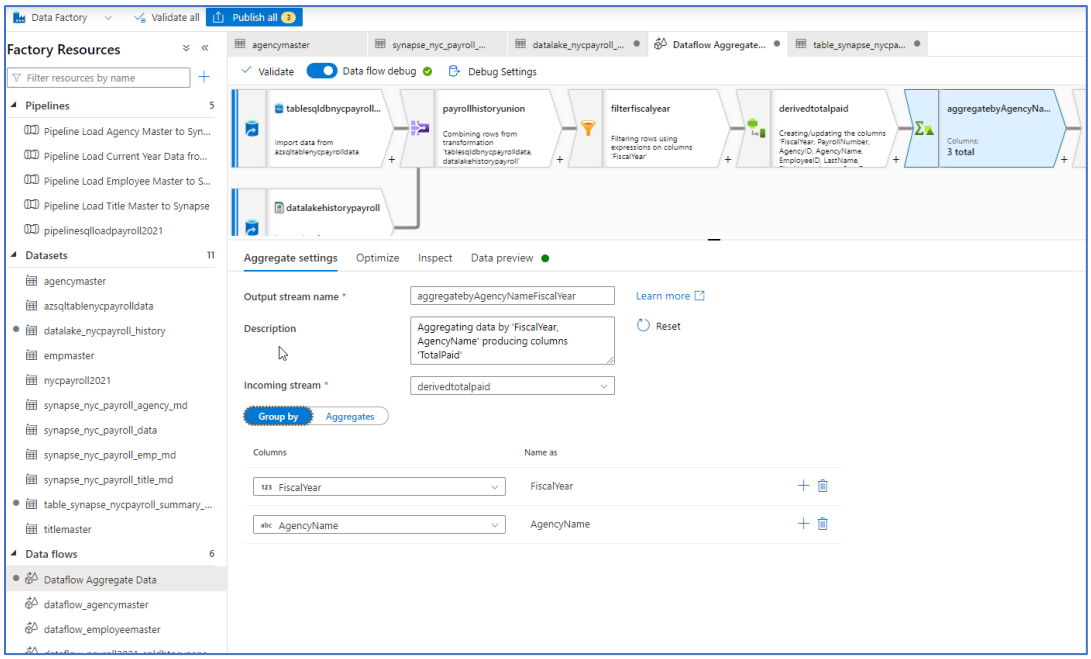


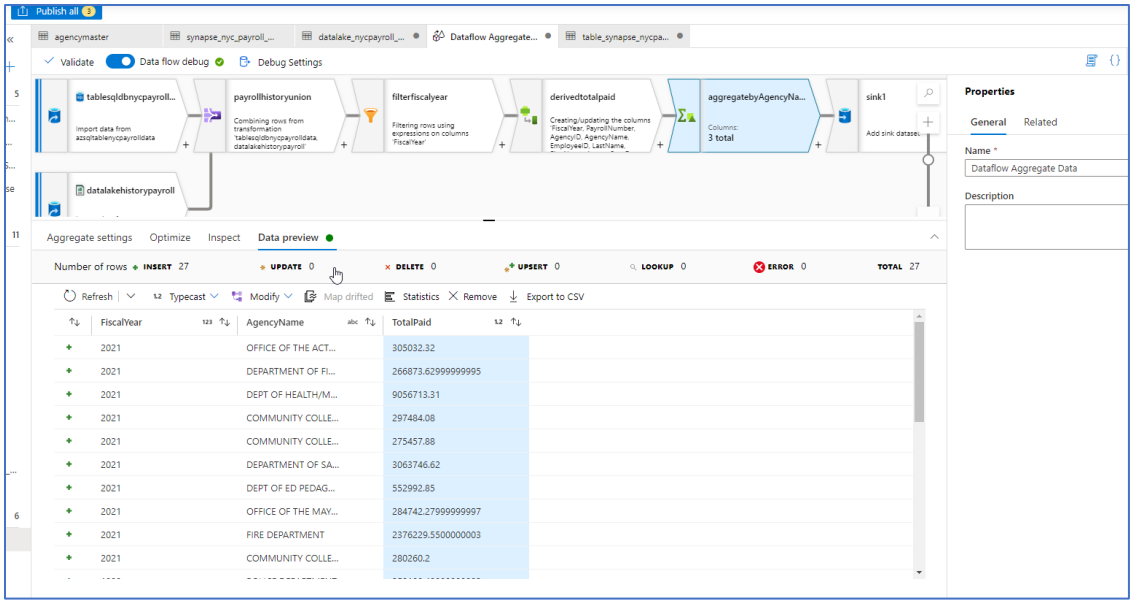
BaseSalary	PayBasis	RegularHo...	RegularGr...	OTHours	TotalOTPaid	TotalOther...	TotalPaid
258000.0	per Annum	1820.0	257260.3	0.0	0.0	258000.0	515260.3
414707.0	per Annum	1820.0	413518.05	0.0	0.0	500.0	414018.05
508.8	per Day	2080.0	132288.0	2115.25	218628.18	56616.07	407532.25
508.8	per Day	2080.0	132288.0	2152.75	218694.96	38611.82	389594.77...
508.8	per Day	2080.0	132288.0	1876.25	192296.19	51160.2	375744.39
228000.0	per Annum	980.0	122427.81	0.0	0.0	244495.2	366923.01
460.25	per Day	1820.0	119469.25	2032.5	200038.56	28316.97	347824.78



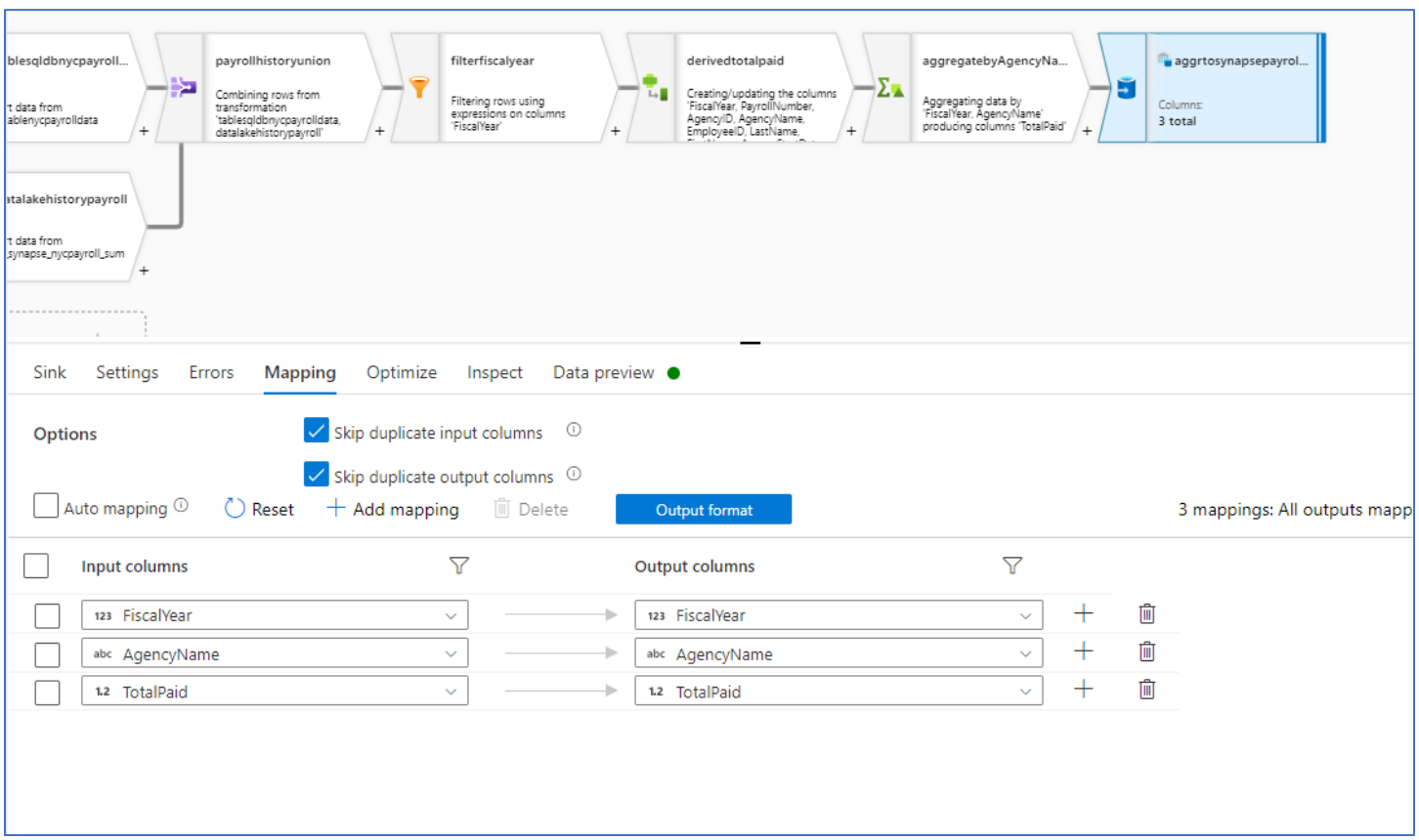
7.Add an Aggregate activity to the data flow next to the TotalPaid activity

- Under Group By, Select AgencyName and Fiscal Year



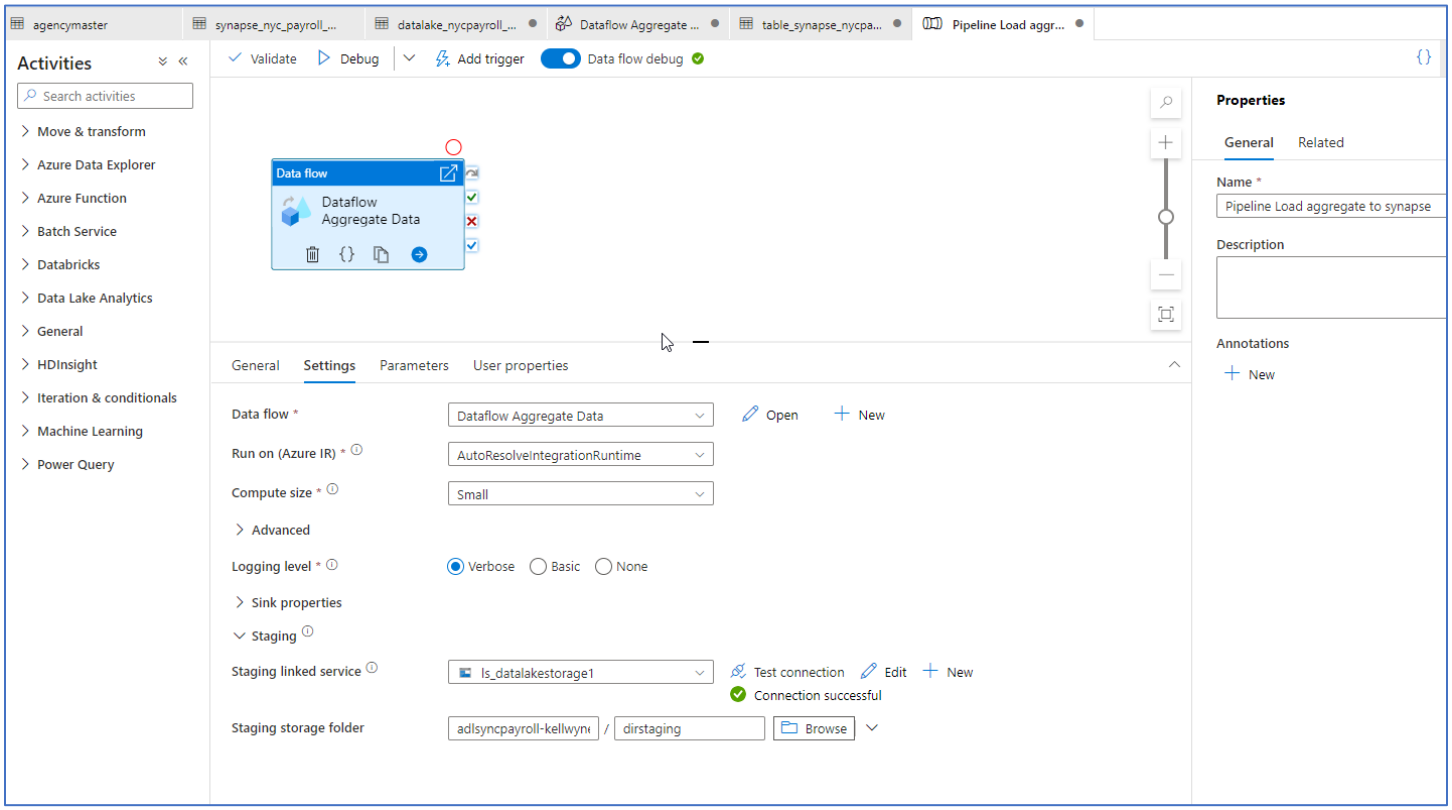


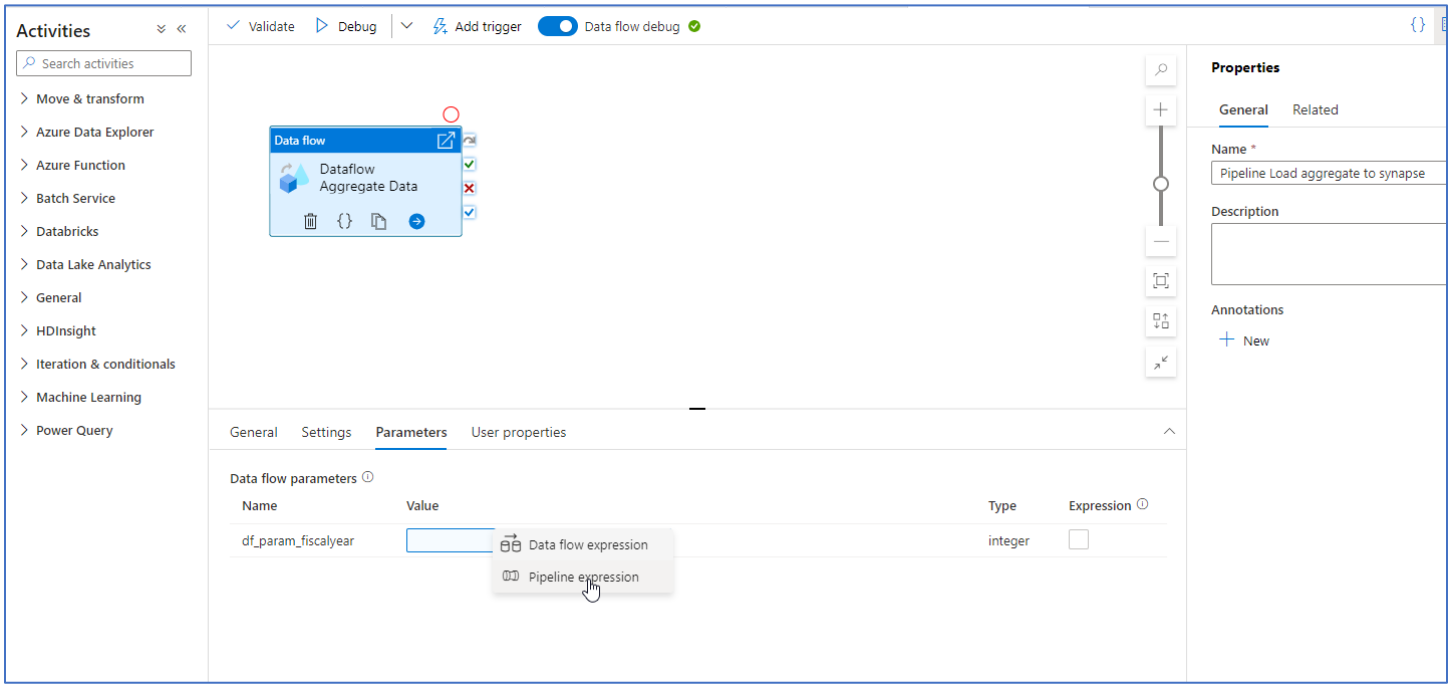
8.Add a Sink activity to the Data Flow



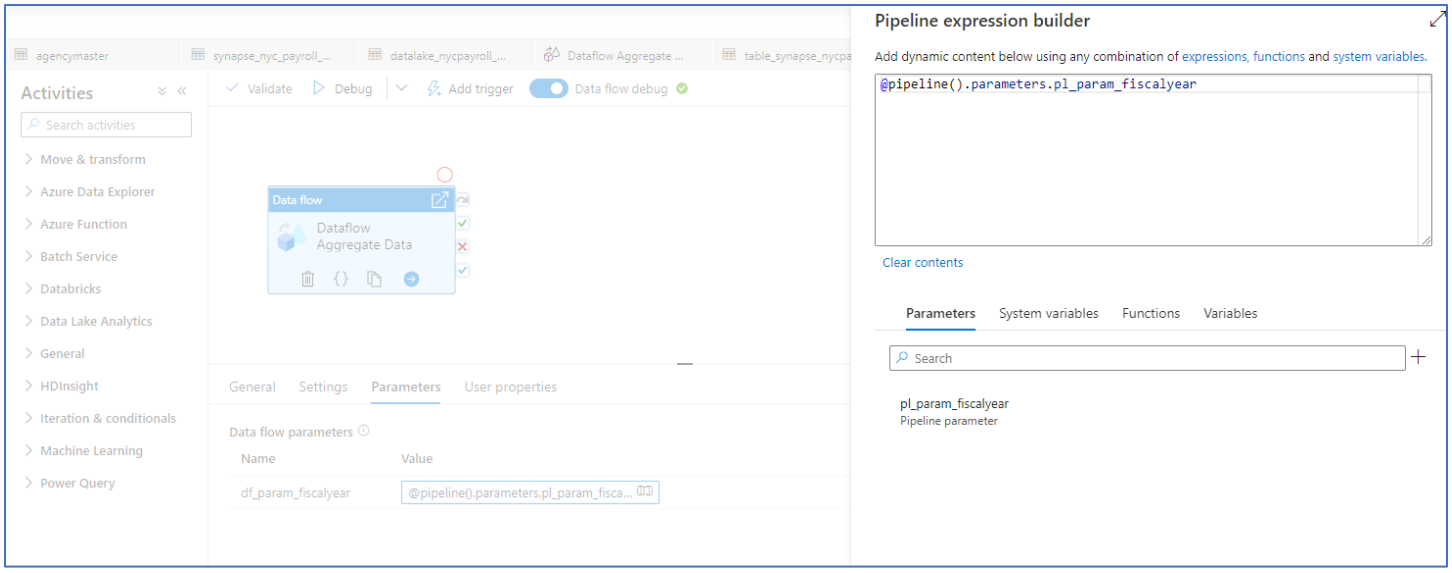
9.Create a new Pipeline and add the Aggregate data flow

- Create a new Global Parameter (This will be the Parameter at the global pipeline level that will be passed on to the data flow

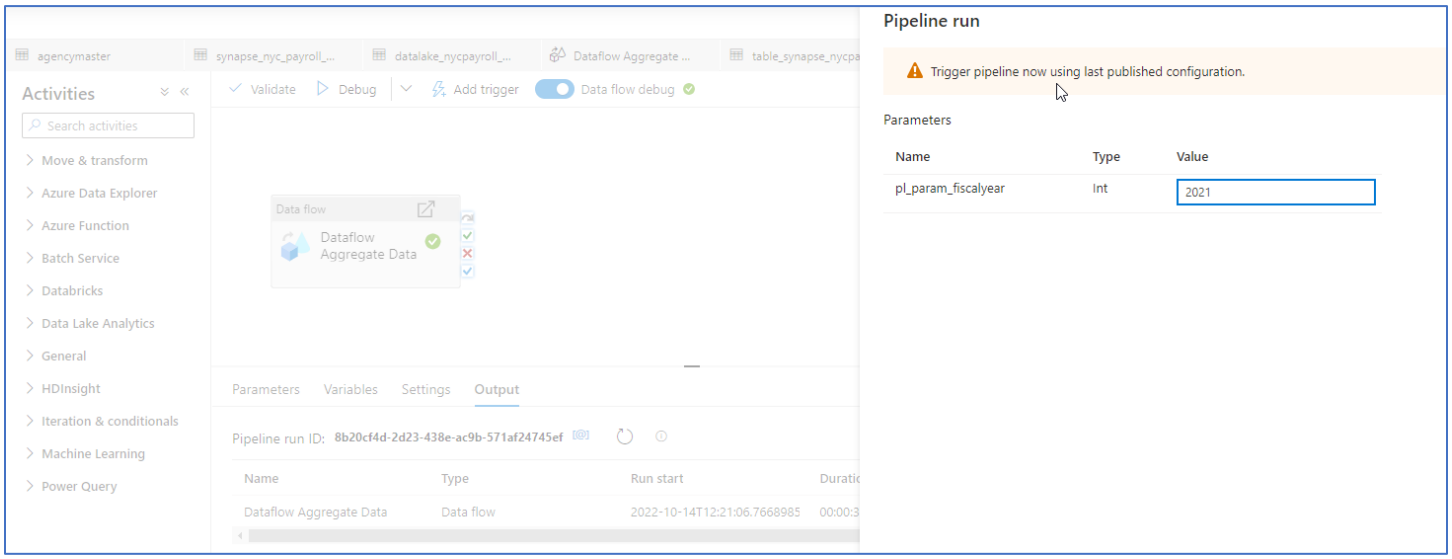


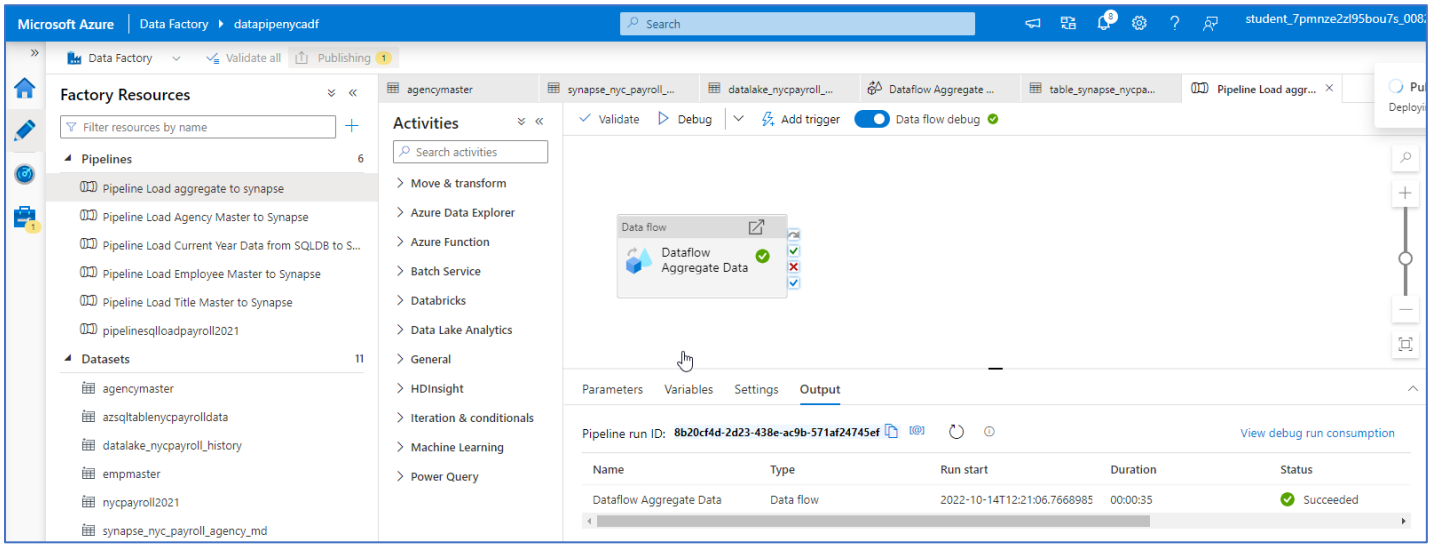


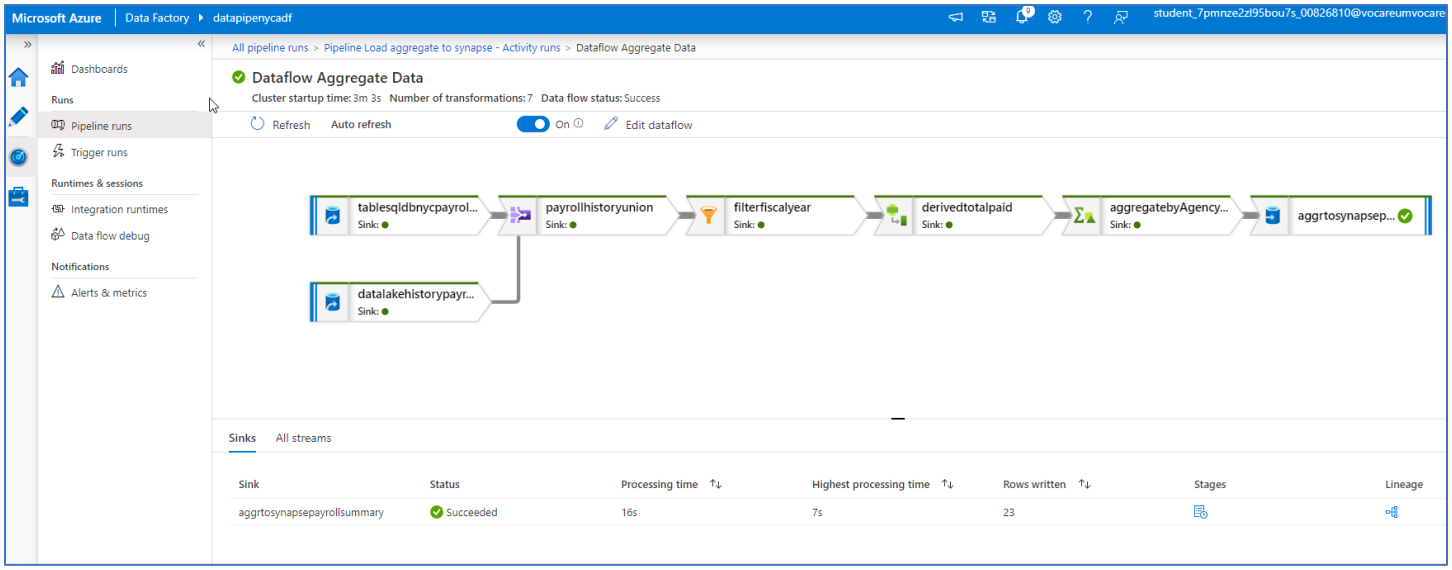
10.Validate, Publish and Trigger the pipeline. Enter the desired value for the parameter.

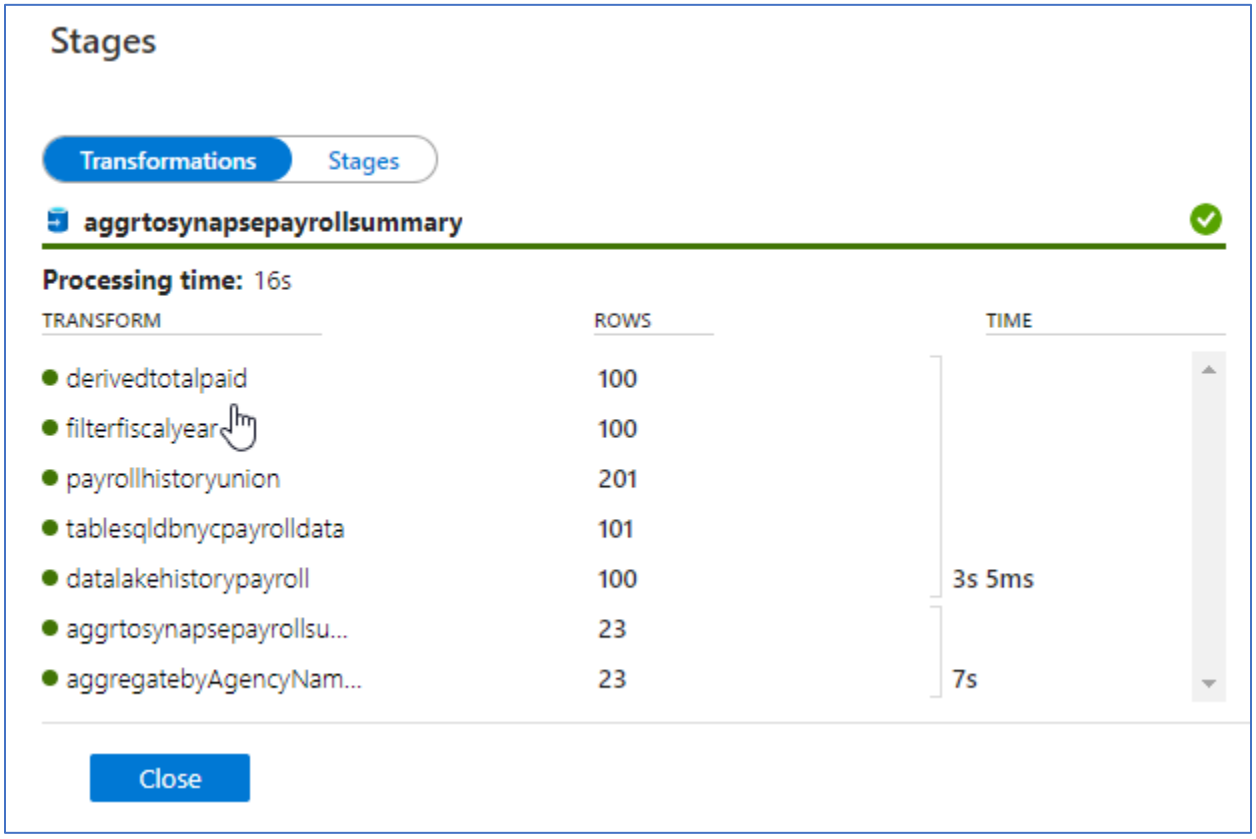


11.Monitor the Pipeline run and take a screenshot of the finished pipeline run.






Microsoft Azure
Data Factory
datapipenycadf
student_7pmnze2zl95bou7s_00826810@vocareumvocare
Dashboards
Runs
Pipeline runs
Trigger runs
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
All pipeline runs > Pipeline Load aggregate to synapse - Activity runs > Dataflow Aggregate Data
Dataflow Aggregate Data
Cluster startup time: 3m 3s Number of transformations: 7 Data flow status: Success
Refresh
Auto refresh
On
Edit dataflow
tablesqldbnycpayrol...
Sink:
payrollhistoryunion
Sink:
filterfiscalyear
Sink:
derivedtotalpaid
Sink:
aggregatebyAgency...
Sink:
aggrtosynapsep...
datalakehistorypayr...
Sink:
Sinks
All streams
Sink
Status
Processing time
Highest processing time
Rows written
Stages
Lineage
aggrtosynapsepayrollsummary
Succeeded
16s
7s
23


Stages
Transformations
Stages
aggrtosynapsepayrollsummary
Processing time: 16s
TRANSFORM
ROWS
TIME
derivedtotalpaid
100
filterfiscalyear
100
payrollhistoryunion
201
tablesqldbnycpayrolldata
101
datalakehistorypayroll
100
3s 5ms
aggrtosynapsepayrollsu...
23
aggregatebyAgencyNam...
23
7s
Close

FACTORY RESOURCES: