

Contents

Azure resource	3
Step 1: Prepare the Data Infrastructure	3
1. 1. Create the data lake and upload data.....	3
1.2. Upload these files from the project data to the dirpayrollfiles folder.....	4
1.3. Upload this file (historical data) from the project data to the dirhistoryfiles folder....	4
1. 2. Create an Azure Data Factory Resource	4
1. 3. Create a SQL Database to store the current year of the payroll data.....	4
1. 4. Create A Synapse Analytics workspace, or use one you already have created.	6
Step 2: Create Linked Services	8
2.1. Create a Linked Service for Azure Data Lake	8
2. 2. Create a Linked Service to SQL Database that has the current (2021) data.....	9
2. 3. Create a Linked Service for Synapse Analytics	9
Step 3: Create Datasets in Azure Data Factory	10
3.1. Create the datasets for the 2021 Payroll file on Azure Data Lake Gen2	11
3.2. Repeat the same process to create datasets for the rest of the data files in the Data Lake	11
3.3. Create the dataset for transaction data table that should contain current (2021) data in SQL DB	14
3.4. Create the datasets for destination (target) tables in Synapse Analytics	14
Step 4: Create Data Flows	16
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).	16
4.2. Create Pipeline to load 2021 Payroll data into transaction table in the SQL DB	16
4.3. Create data flows to load the data from the data lake files into the Synapse Analytics data tables	17
4.4. Create a data flow to load 2021 data from SQL DB to Synapse Analytics	19
4.5. Create pipelines for Employee, Title, Agency, and year 2021 Payroll transaction data to Synapse Analytics containing the data flows.	19
4.6. Trigger and monitor the Pipelines	20
4.6.1. Screenshot of - Pipeline Load 2021 Payroll Into SQLDB	21
4.6.2. Screenshot of - Pipeline Load Agency Master To Synapse	22
4.6.3 Screenshot of - Pipeline Load Current Year Payroll Data from SQLDB To Synapse	23

4.6.4. Screenshot of - Pipeline Load Employee Master To Synapse	23
4.6.5. Screenshot of - Pipeline Load Title Master To Synapse	24
Step 5: Data Aggregation and Parameterization	25
5.1. Create a Summary table in Synapse with the given SQL script and create a dataset named table_synapse_nycpayroll_summary	25
5.2. Create a new dataset for the Azure Data Lake Gen2 folder that contains the historical files.	25
5.3. Create new data flow and name it Dataflow Aggregate Data	26
5.4. Create a new Union activity in the data flow and Union with history files	27
5.5. Add a Filter activity after Union.....	27
5.6. Derive a new TotalPaid column	27
5.7. Add an Aggregate activity to the data flow next to the TotalPaid activity.....	28

Azure resource

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

The screenshot shows the Azure portal interface with the following details:

- Header:** + Create, Manage view, Delete resource group, Refresh, Export to CSV, Open query, Assign tags, Move, Delete, Export template, JSON View.
- Subscription:** Subscription (move) : Vocareum-UDA-1, Subscription ID : 94ec3a64-dcfe-4219-9e29-88221690c382, Tags (edit) : objectid : 35b73226-9660-4d2f-9b71-17864336cd49, Deployments : 9_Succeeded, Location : West US.
- Resources:** Resources tab selected, Recommendations tab available. Filter buttons: Type equals all, Location equals all, Add filter.
- Table:** A list of resources with columns: Name, Type, Location. Resources listed include:
 - datafactory000111 (Data factory (V2), East US)
 - sqldb00 (SQL server, East US)
 - sqldb2 (sqldb00/sqldb2) (SQL database, East US)
 - sqlpool00 (synapse000/sqlpool00) (Dedicated SQL pool, East US)
 - storage000adls (Storage account, East US)
 - synapse000 (Synapse workspace, East US)
 - synapse0001sa (Storage account, East US)

Step 1: Prepare the Data Infrastructure

Setup Data and Resources in Azure

1. 1. Create the data lake and upload data

The screenshot shows the Microsoft Azure Storage Explorer interface for a blob container named "adlsnycpayrol-rajesh-c".

- Header:** Microsoft Azure, Search resources, services, and docs (G+).
- Breadcrumb:** Home > Regroup_6qOejBx8LzjNre > storage000adls | Containers > adlsnycpayrol-rajesh-c.
- Container Overview:** Container name: adlsnycpayrol-rajesh-c, Authentication method: Access key (Switch to Azure AD User Account), Location: adlsnycpayrol-rajesh-c.
- Actions:** Search, Upload, Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, Break lease.
- Search Bar:** Search blobs by prefix (case-sensitive).
- Table:** A list of blobs with columns: Name, Modified, Access tier, Archive status, Blob type.

Name	Modified	Access tier	Archive status	Blob type
dirhistoryfiles				
dirpayrollfiles				
dirstaging				
- Left Sidebar:** Settings, Shared access tokens, Manage ACL, Access policy, Firewall settings.

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

The screenshot shows the Azure Storage Explorer interface. On the left, there's a sidebar with options like Overview, Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and a Search bar. The main area shows a list of blobs in the 'dirpayrollfiles' folder. The blob names are AgencyMaster.csv, EmpMaster.csv, nycpayroll_2021.csv, and TitleMaster.csv. Each blob has a modified date of 10/7/2022, 11:13:37 ..., an access tier of Hot (Inferred), and a blob type of Block blob. On the right, there's an 'Upload blob' dialog with a 'Select a file' input field, an 'Overwrite if files already exist' checkbox, and an 'Upload' button. Below it, a 'Current uploads' section shows four completed uploads: TitleMaster.csv (49 KiB / 49 KiB), nycpayroll_2021.csv (16 KiB / 16 KiB), EmpMaster.csv (23 KiB / 23 KiB), and AgencyMaster.csv (5 KiB / 5 KiB).

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

This screenshot is similar to the previous one but shows a different folder, 'dirhistoryfiles'. It contains a single blob named 'nycpayroll_2020.csv' with a modified date of 10/7/2022, 11:14:51 ... and an access tier of Hot (Inferred). The rest of the interface is identical to the first screenshot, including the upload dialog and current uploads list.

1.2. Create an Azure Data Factory Resource

The screenshot shows the Azure Data Factory resource 'datafactory000111'. The left sidebar includes Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Networking, Managed identities, Properties, Locks, and Setting started. The main area displays the 'Essentials' section with the following details: Resource group (move) : Regroup_6qOejBx8LzjNre, Status : Succeeded, Location : East US, Subscription (move) : Vocareum-UDA-1, Subscription ID : 94ec3a64-dcfe-4219-9e29-88221690c382, Type : Data factory (V2), and Getting started : Quick start. Below this, there's a 'Getting started' section with two cards: 'Open Azure Data Factory Studio' (Start authoring and monitoring your data pipelines and data flows) and 'Read documentation' (Learn how to be productive quickly, Explore concepts, tutorials, and samples).

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

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

Home >

db_nycpayroll (sqldb00/db_nycpayroll)

SQL database

Search Copy Restore Export Set server firewall Delete Connect with... Feedback

Overview Activity log Tags Diagnose and solve problems Getting started Query editor (preview) Power Platform

This database was just created. Do you need any help getting started?

Essentials

Resource group (move)	: Regroup_6qOejBx8LzjNre	Server name	: sqldb00.database.windows.net
Status	: Ready	Elastic pool	: No elastic pool
Location	: East US	Connection strings	: Show database connection strings
Subscription (move)	: Vocareum-UDA-1	Pricing tier	: Basic
Subscription ID	: 94ec3a64-dcfe-4219-9e29-88221690c382	Earliest restore point	: 2022-10-02 04:33 UTC

- Add client IP address to the SQL DB firewall

Home > sqldb00

sqldb00 | Networking

SQL server

Search Locks

Data management

- Backups
- Deleted databases
- Failover groups
- Import/Export history

Security

- Networking
- Microsoft Defender for Cloud
- Transparent data encryption
- Identity
- Auditing

Intelligent Performance

- Automatic tuning

Virtual networks

Allow virtual networks to connect to your resource using service endpoints. [Learn more](#)

+ Add a virtual network rule

Rule	Virtual network	Subnet	Address range	Endpoint status	Resource group	Subscription	State
------	-----------------	--------	---------------	-----------------	----------------	--------------	-------

Firewall rules

Allow certain public internet IP addresses to access your resource. [Learn more](#)

+ Add your client IPv4 address (49.37.105.87) + Add a firewall rule

Rule name	Start IPv4 address	End IPv4 address
ClientIp-2022-10-2_9-34-47	49.37.111.172	49.37.111.172
ClientIPAddress_2022-10-7_23-28-29	49.37.105.87	49.37.105.87

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

```

14: [BaseSalary] [float] NULL,
15: [PayBasis] [varchar](50) NULL,
16: [RegularHours] [float] NULL,
17: [RegularGrossPaid] [float] NULL,
18: [OTHours] [float] NULL,
19: [TotaloTPaid] [float] NULL,
20: [TotalotherPay] [float] NULL
21: )
22: GO

```

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

- Create a Synapse Analytics workspace in the Azure portal.

Essentials

Resource group (move)	: Regroup_6qQejBx8LzjNre	Networking	: Show firewall settings
Status	: Succeeded	Primary ADLS Gen2 acco...	: https://synapse0001sa.dfs.core.windows.net
Location	: East US	Primary ADLS Gen2 file ...	: synapse001fs
Subscription (move)	: Vocareum-UDA-1	SQL admin username	: sqldminuser
Subscription ID	: 94ec3a64-dcfe-4219-9e29-88221690c382	SQL Active Directory ad...	: student_10f0d9bb7ztciw7_00826808@vocareumvocareum.onmicrosoft.com
Managed virtual network	: No	Dedicated SQL endpoint	: synapse000.sql.azuresynapse.net
Managed identity object ...	: 1531c950-f7f1-4edd-8ccb-f64ef34afe6	Serverless SQL endpoint	: synapse000-ondemand.sql.azuresynapse.net
Workspace web URL	: https://web.azure-synapse.net/workspace=%2bsubscriptions%2f94...	Development endpoint	: https://synapse000.dev.azuresynapse.net
Tags (edit)	: Click here to add tags		

Getting started

Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
[Open](#)

Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
[Learn more](#)

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

Analytics pools		
<input type="text"/> Search to filter items...		
Name	Type	Size
SQL pools		
Built-in	Serverless	Auto
sqlpool00	Dedicated	DW100c

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

Microsoft Azure | Synapse Analytics > synapse000 Search

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace. On the left, there's a navigation sidebar with icons for Home, Databases, Workspaces, Pipelines, and Jobs. The 'Data' workspace is selected. In the center, there's a tree view under 'Workspace' showing a 'SQL database' named 'sqlpool00 (SQL)'. Under this database, there are 'Tables', 'External tables', 'External resources', and 'Views'. Four specific tables are highlighted with yellow boxes: 'dbo.NYC_Payroll_AGENCY...', 'dbo.NYC_Payroll_Data', 'dbo.NYC_Payroll_EMP_MD', and 'dbo.NYC_Payroll_TITLE_MD'. To the right of the tree view is a large text area titled 'SQL script 1' containing the following SQL code:

```

3     [PayrollNumber] [int] NULL,
4     [AgencyID] [varchar](10) NULL,
5     [AgencyName] [varchar](50) NULL,
6     [EmployeeID] [varchar](10) NULL,
7     [LastName] [varchar](20) NULL,
8     [FirstName] [varchar](20) NULL,
9     [AgencyStartDate] [date] NULL,
10    [WorkLocationBorough] [varchar](50) NULL,
11    [TitleCode] [varchar](10) NULL,
12    [TitleDescription] [varchar](100) NULL,
13    [LeaveStatusasofJune30] [varchar](50) NULL,
14    [BaseSalary] [float] NULL,
15    [PayBasis] [varchar](50) NULL,
16    [RegularHours] [float] NULL,
17    [RegularGrossPaid] [float] NULL,
18    [OTHours] [float] NULL,
19    [TotalOTPaid] [float] NULL,
20    [TotalOtherPay] [float] NULL
21 )
22 GO

```

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

 Azure Data Lake Storage Gen2 [Learn more](#)

Name *

AzureDataLakeStorageLinkService

Description

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime



Authentication type

Account key



Account selection method ⓘ

From Azure subscription Enter manually

Azure subscription ⓘ

Vocareum-UDA-1 (94ec3a64-dcfe-4219-9e29-88221690c382)



Storage account name *

storage000adls



Test connection ⓘ

To linked service To file path



Connection successful

Apply

Cancel



Test 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

 Azure SQL Database [Learn more](#)

AutoResolveIntegrationRuntime

Connection string **Azure Key Vault**

Account selection method ⓘ

From Azure subscription Enter manually

Fully qualified domain name *

sqldb00.database.windows.net

Database name *

db_nycpayroll

Authentication type *

SQL authentication

User name *

sqldb

Password **Azure Key Vault**

Password *

.....

Always encrypted ⓘ

Apply **Cancel**

 Connection successful

 Test connection

2. 3. Create a Linked Service for Synapse Analytics

- Create the linked service to the SQL pool.



AzureSynapseAnalyticsLinkServices

Description

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime

Connection string

Azure Key Vault

Account selection method ⓘ

From Azure subscription Enter manually

Azure subscription

Vocareum-UDA-1 (94ec3a64-dcfe-4219-9e29-88221690c382)

Server name *

synapse000 (Synapse workspace)



Database name *

sqlpool00



SQL pool *

sqlpool00



Connection successful

Test connection

Apply

Cancel

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

Properties

Name * nycpayroll_2021

Description

Annotations

Connection

Linked service * AzureDataLakeStorageLinkService Test connection Edit + New Learn more Connection successful

File path * adlsnycpayrol-rajesh-c / dirpayrollfiles / nycpayroll_2021.csv Browse Prev

Compression type None

Column delimiter ⓘ Comma (,) Edit

Row delimiter ⓘ Default (\r\n, or \n) Edit

Encoding ⓘ Default(UTF-8)

Preview data

Linked service: AzureDataLakeStorageLinkService
Object: nycpayroll_2021.csv

	FiscalYear	PayrollNumber	AgencyCode	AgencyName	EmployeeID	LastName	FirstName	AgencyStartDate	WorkLocationBorough	TitleCode	TitleDescription	LeaveStatusasofJune30	BaseSalary	Per
1	2021	996	2153	NYC HOUSING AUTHORITY	209184	MUSTACIUOLO	VITO	2/26/2018	MANHATTAN	40475	EXECUTIVE DIRECTOR	ACTIVE	258000	per An
2	2021	996	2153	NYC HOUSING AUTHORITY	302330	RUSS	GREGORY	8/12/2019	MANHATTAN	41143	CHAIR	ACTIVE	414707	per An
3	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	49788	HALLAHAN	PATRICK	2/26/2018	BROOKLYN	40782	STATIONARY ENGINEER	ACTIVE	508.8	per
4	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	251626	PETTIT	PATRICK	8/2/2010	MANHATTAN	40782	STATIONARY ENGINEER	ACTIVE	508.8	per

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

- EmpMaster.csv

Connection

Linked service * AzureDataLakeStorageLinkService Test connection Edit + New Learn more Connection successful

File path * adlsnycpayrol-rajesh-c / dirpayrollfiles / EmpMaster.csv Browse Prev

Compression type None

Column delimiter ⓘ Comma (,) Edit

Row delimiter ⓘ Default (\r\n, or \n) Edit

Encoding ⓘ Default(UTF-8)

Preview data

Linked service: AzureDataLakeStorageLinkService
Object: EmpMaster.csv

	FiscalYear	PayrollNumber	AgencyCode	AgencyName	EmployeeID	LastName	FirstName	AgencyStartDate	WorkLocationBorough	TitleCode	TitleDescription	LeaveStatusasofJune30	BaseSalary	Per
1	2021	996	2153	NYC HOUSING AUTHORITY	209184	MUSTACIUOLO	VITO	2/26/2018	MANHATTAN	40475	EXECUTIVE DIRECTOR	ACTIVE	258000	per An
2	2021	996	2153	NYC HOUSING AUTHORITY	302330	RUSS	GREGORY	8/12/2019	MANHATTAN	41143	CHAIR	ACTIVE	414707	per An
3	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	49788	HALLAHAN	PATRICK	2/26/2018	BROOKLYN	40782	STATIONARY ENGINEER	ACTIVE	508.8	per
4	2021	816	2129	DEPT OF HEALTH/MENTAL HYGIENE	251626	PETTIT	PATRICK	8/2/2010	MANHATTAN	40782	STATIONARY ENGINEER	ACTIVE	508.8	per

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

The screenshot shows the Azure Data Factory interface for configuring a dataset named 'TitleMaster'. The top navigation bar includes tabs for 'the factory', 'Pipelines', 'EmpMaster', and 'TitleMaster'. The main area displays the dataset properties, connection settings, schema, and preview data.

Properties:

- General:** Name is set to 'TitleMaster'.
- Description:** An empty text field.
- Annotations:** A '+ New' button.

Connection: Linked service is set to 'AzureDataLakeStorageLinkService' (connection successful). File path is 'adlsnycpayrol-rajesh-c / dirpayrollfiles / TitleMaster.csv'. Compression type is 'None'. Column delimiter is 'Comma (,)'. Row delimiter is 'Default (\r\n, or \n)'.

Schema: The dataset has one column: 'TitleCode' (Type: String) and 'TitleDescription' (Type: String).

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

Properties

- General
- Related

Name *
AgencyMaster

Description

Annotations

+ New

Connection Schema Parameters

Linked service * AzureDataLakeStorageLinkService Test connection Edit + New Learn more Connection successful

File path * adlsnycpayroll-rajesh-c / dirpayrollfiles / AgencyMaster.csv Browse Prev

Compression type None

Column delimiter (.) Comma (.)

Preview data

Linked service: AzureDataLakeStorageLinkService

Object: AgencyMaster.csv

	AgencyID	AgencyName
1	2001	ADMIN FOR CHILDREN'S SVCS
2	2002	ADMIN TRIALS AND HEARINGS
3	2003	BOARD OF CORRECTION
4	2004	BOARD OF ELECTION
5	2005	BOARD OF ELECTION POLL WORKERS

- Remember to publish all the datasets

Factory Resources < Filter resources by name +

- Pipelines 0
- Datasets 4
 - AgencyMaster
 - EmpMaster
 - nycpayroll_2021
 - TitleMaster
- Data flows 0
- Power Query 0

Dismiss all

✓ Publishing completed Successfully published a few seconds ago

✓ Publishing completed Successfully published 2 minutes ago

✓ Publishing completed Successfully published 8 minutes ago

✓ Publishing completed Successfully published 12 minutes ago

Connection Schema Parameters

Linked service * AzureDataLakeStorageLinkService Test connection Connection successful

File path * adlsnycpayroll-rajesh-c / dirpayrollfiles / AgencyMaster.csv

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

The screenshot shows the Azure Data Factory studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Datasets', 'AzureSqlTable_NYC_Payroll_data' is selected. The main workspace displays the 'AzureSQLDatabase' icon and the table name 'dbo.NYC_Payroll_Data'. The 'Properties' panel on the right shows the dataset's name as 'AzureSqlTable_NYC_Payroll_data' and its connection status as 'Connection successful'. A green checkmark indicates 'Publishing completed'.

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

- dataset for NYC_Payroll_EMP_MD

The screenshot shows the Azure Data Factory studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Datasets', 'AzureSynapseAnalyticsTable_NYC_Payroll_EMP_MD' is selected. The main workspace displays the 'Azure Synapse Analytics' icon and the table name 'dbo.NYC_Payroll_EMP_MD'. The 'Properties' panel on the right shows the dataset's name as 'AzureSynapseAnalyticsTable_NYC_Payroll_EI' and its connection status as 'Connection successful'.

- dataset for NYC_Payroll_TITLE_MD

The screenshot shows the Azure Data Factory studio interface. On the left, the 'Factory Resources' sidebar lists Pipelines, Datasets, Data flows, and Power Query. Under 'Datasets', 'AzureSynapseAnalyticsTable_NYC_Payroll_TITLE_MD' is selected. The main workspace displays the 'Azure Synapse Analytics' icon and the table name 'dbo.NYC_Payroll_TITLE_MD'. The 'Properties' panel on the right shows the dataset's name as 'AzureSynapseAnalyticsTable_NYC_Pay' and its connection status as 'Connection successful'.

- dataset for NYC_Payroll_AGENCY_MD

The screenshot shows the Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various datasets, pipelines, and data flows. In the center, the properties of a specific dataset are displayed. The dataset is named 'AzureSynapseAnalyticsTable_NYC_Payroll_AGENCY_MD'. It is connected via a 'Linked service' to an 'AzureSynapseAnalyticsLinkServices' instance. The 'Table' dropdown is set to 'dbo.NYC_Payroll_AGENCY_MD'. The 'Properties' pane on the right shows the general settings, including the name 'AzureSynapseAnalyticsTable_NYC_Payroll_AGENCY_MD' and a successful connection status.

- dataset for NYC_Payroll_Data

This screenshot shows another dataset in the Azure Data Factory interface, named 'AzureSynapseAnalyticsTable_NYC_Payroll_Data'. It uses the same 'AzureSynapseAnalyticsLinkServices' linked service and connects to the 'dbo.NYC_Payroll_Data' table. The properties pane indicates a successful connection.

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).

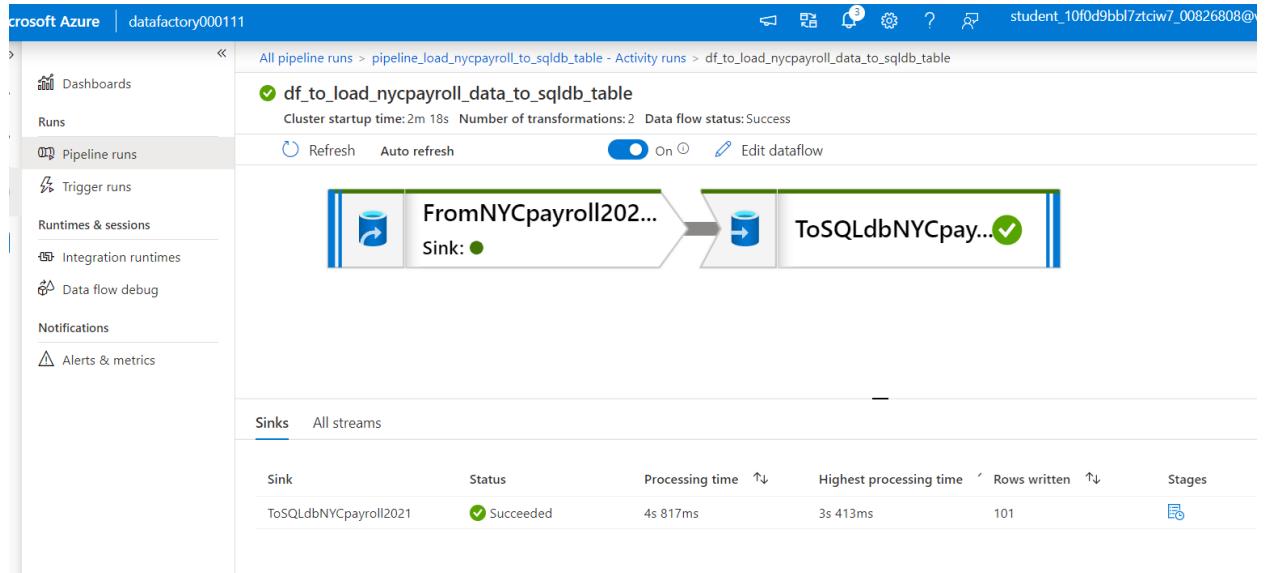
- Create a new data flow
- Select the dataset for the 2021 payroll file as the source
- Select the sink dataset as the payroll table on SQL DB
- Make sure to reassign any missing source to target mappings

The screenshot shows the Azure Data Factory Data Flow blade. On the left, there's a navigation sidebar with 'Pipelines' (0), 'Datasets' (1), 'Data flows' (1), and 'Power Query' (0). The main area displays a data flow with one reference and one sink. The sink is named 'ToSQLdbNYCPayroll...' and has 19 columns. Below the data flow, there's a 'Data preview' section showing a sample of 101 rows from the 2021 payroll file. The columns listed are: FiscalYear, PayrollNu..., AgencyID, AgencyNa..., EmployeeID, and LastName. The data preview table contains several rows of payroll information.

FiscalYear	PayrollNu...	AgencyID	AgencyNa...	EmployeeID	LastName
2021	996	2153	NYC HOUS...	209184	MUSTF...
2021	996	2153	NYC HOUS...	302330	RUSS
2021	816	2129	DEPT OF H...	49788	HALLA
2021	816	2129	DEPT OF H...	251626	PETTIT
2021	816	2129	DEPT OF H...	364376	TELEH/
2021	462	2092	GUTTMAN ...	375488	EVENB
2021	996	2153	NYC HOUS...	332352	DALEY

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

Query 2

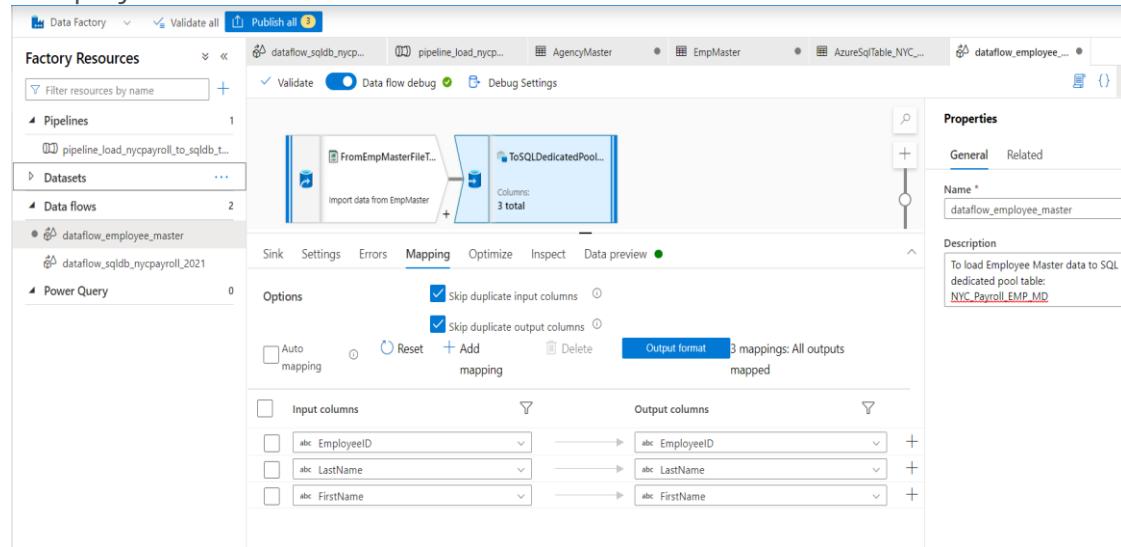
```
1. SELECT TOP (100) * FROM [dbo].[NYC_Payroll_Data]
```

FiscalYear	PayrollNumber	AgencyID	AgencyName	EmployeeID
2021	996	2153	NYC HOUSING AUTHORITY	209184
2021	996	2153	NYC HOUSING AUTHORITY	302330
2021	816	2129	DEPT OF HEALTH/MENTAL HYG...	49788
2021	816	2129	DEPT OF HEALTH/MENTAL HYG...	49788

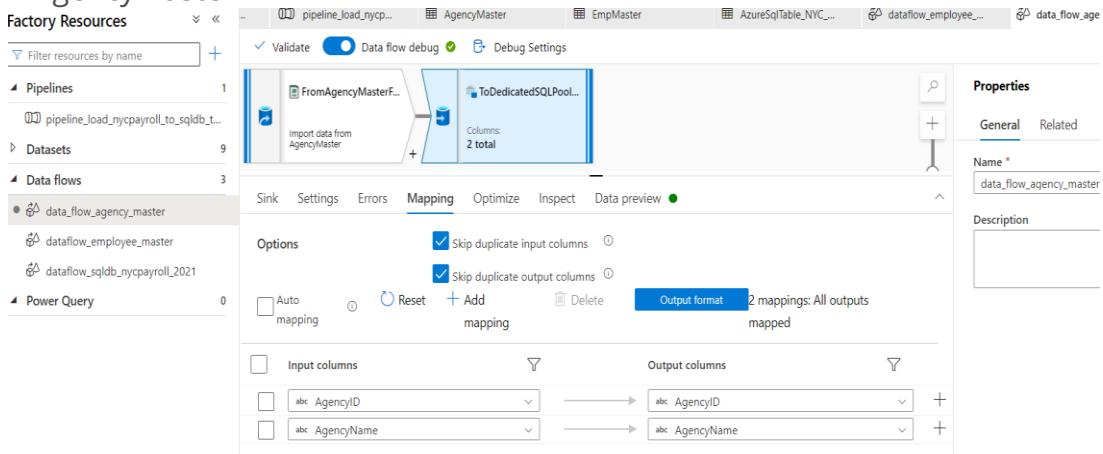
4.3. Create data flows to load the data from the data lake files into the Synapse Analytics data tables

- 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 Synapse table

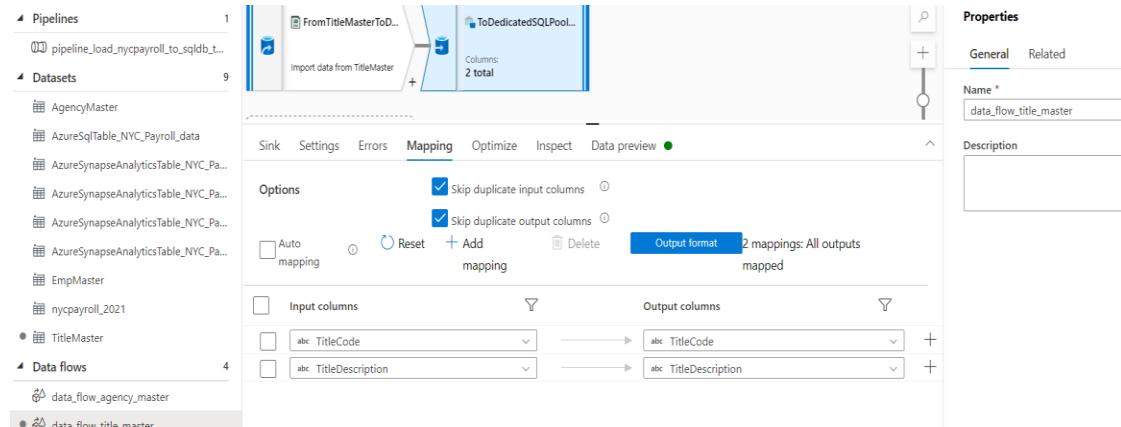
- EmployeeMaster



- AgencyMaster



- TitleMaster



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

The screenshot shows the Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various pipelines and datasets. In the center, a data flow named 'data_flow_sqldb_2021_data_to_synapse' is selected. The data flow diagram shows a 'FromSQLDBTable...' source connected to a 'ToSynapse' sink. The source is configured to import data from 'AzureSqlTable_NYC_Payroll_data'. The sink has 19 columns. Below the diagram, the 'Data preview' tab is active, displaying a table with three rows of data:

FiscalYear	PayrollNu...	AgencyID	AgencyNa...	EmployeeID
2021	996	2153	NYC HOUS...	209184
2021	996	2153	NYC HOUS...	302330
2021	816	2129	DEPT OF H...	49788

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**

The screenshot shows the Azure Data Factory interface. On the left, the 'Factory Resources' sidebar lists various pipelines and datasets. In the center, a pipeline named 'Pipeline Load Agency Master To Synapse' is selected. The 'Settings' tab is active. The 'linked service' dropdown is set to 'AzureDataLakeStorageLinkService'. The 'storage folder' input field contains 'adlsnycpayroll-rajesh-c / dirstaging'.

- **Title Pipeline**

Factory Resources

- Pipeline Load 2021 Payroll Into SQLDB
- Pipeline Load Agency Master To Synapse
- Pipeline Load Current Year Data from SQLDB To Sy...
- Pipeline Load Employee Master To Synapse
- Pipeline Load Title Master To Synapse
- Datasets 9
- Data flows 5
 - data_flow_agency_master
 - data_flow_sqldb_2021_data_to_synapse
 - data_flow_title_master
 - data_flow_employee_master
 - data_flow_nycpayroll_2021_to_sqldb

Properties

General

Name * Pipeline Load Title Master To Synapse

Description To load Title master data to staging

Annotations

+ New

• Agency Pipeline

Factory Resources

- Pipeline Load 2021 Payroll Into SQLDB
- Pipeline Load Agency Master To Synapse
- Pipeline Load Current Year Data from SQLDB To Sy...
- Pipeline Load Employee Master To Synapse
- Pipeline Load Title Master To Synapse
- Datasets 9
- Data flows 5
 - data_flow_agency_master
 - data_flow_sqldb_2021_data_to_synapse
 - data_flow_title_master
 - data_flow_employee_master
 - data_flow_nycpayroll_2021_to_sqldb

Properties

General

Name * Pipeline Load Agency Master To Synapse

Description To load Agency master data to staging

Annotations

+ New

• Year 2021 Payroll

Factory Resources

- Pipeline Load 2021 Payroll Into SQLDB
- Pipeline Load Agency Master To Synapse
- Pipeline Load Current Year Data from SQLDB To Sy...
- Pipeline Load Employee Master To Synapse
- Pipeline Load Title Master To Synapse
- Datasets 9
- Data flows 5
 - data_flow_agency_master
 - data_flow_sqldb_2021_data_to_synapse
 - data_flow_title_master
 - data_flow_employee_master
 - data_flow_nycpayroll_2021_to_sqldb

Properties

General

Name * Pipeline Load Current Year Data from SQLD

Description To load Current Year Data from SQLDB To Synapse

Annotations

+ New

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

Data Factory ▾ **Validate all** **Publishing** 2

Factory Resources

Filter resources by name +

Pipelines

- Pipeline Load 2021 Payroll Into SQLDB
- Pipeline Load Current Year Payroll Data From SQLDB To Synapse**
- Pipeline Load Agency Master To Synapse
- Pipeline Load All Data Synapse
- Pipeline Load Employee Master To Synapse
- Pipeline Load Title Master To Synapse

Datasets

9

Activities

Search activities

- > Move & transform
- > Azure Data Explorer
- > Azure Function
- > Batch Service
- > Databricks
- > Data Lake Analytics
- > General

4.6.1. Screenshot of - Pipeline Load 2021 Payroll Into SQLDB

All pipeline runs > Pipeline Load 2021 Payroll Into SQLDB - Activity runs > Data Flow Load NYCpayroll Data To SQLDB

Data Flow Load NYCpayroll Data To SQLDB
Cluster startup time: 2m 16s Number of transformations: 2 Data flow status: Success

Refresh Auto refresh On Edit dataflow

Sink	Status	Processing time	Highest processing	Rows written	Stages	Lineage
ToSQLdbNYCpayroll2021	Succeeded	4s 865ms	3s 811ms	101		

Data loaded into SQLDB table

db_nycpayroll (sqldb00/db_nycpayroll) | Query editor (preview)

Showing limited object explorer here. For full capability please open SSDT.

Tables

- dbo.BuildVersion
- dbo.ErrorLog
- dbo.NYC_Payroll_Data**
- SalesLT.Address
- SalesLT.Customer
- SalesLT.CustomerAddress
- SalesLT.Product
- SalesLT.ProductCategory
- SalesLT.ProductDescription
- SalesLT.ProductModel
- SalesLT.ProductModelProductDetail
- SalesLT.SalesOrderDetail

Query 1 × **Query 2 ×**

Run Cancel query Save query Export data as Show only Editor

```
1 SELECT TOP (1000) * FROM [dbo].[NYC_Payroll_Data]
```

Results Messages

FiscalYear PayrollNumber AgencyID AgencyName EmployeeID

2021	996	2153	NYC HOUSING AUTHORITY	209184
2021	996	2153	NYC HOUSING AUTHORITY	302330
2021	816	2129	DEPT OF HEALTH/MENTAL HYG...	49788

4.6.2. Screenshot of - Pipeline Load Agency Master To Synapse

All pipeline runs > Pipeline Load Agency Master To Synapse - Activity runs > Data Flow Load Agency MD

✓ Data Flow Load Agency MD

Cluster startup time: 2m 33s Number of transformations: 2 Data flow status: Success

Refresh Auto refresh On Edit dataflow

Sinks All streams

Sink	Status	Processing time ↑	Highest processing	Rows written ↑↓	Stages	Lineage
ToDedicatedSQLPoolNYCPayrollA	✓ Succeeded	10s	3s 690ms	153		

Data loaded to synapse table

Microsoft Azure | Synapse Analytics > synapse000

Synapse live Validate all Publish all

Data + ×

Workspace Linked

Filter resources by name

SQL database

sqlpool00 (SQL)

- Tables
 - dbo.NYC_Payroll_AGENCY_M...
 - dbo.NYC_Payroll_Data
 - Columns
 - dbo.NYC_Payroll_EMP_MD
 - dbo.NYC_Payroll_TITLE_MD
- External tables
- External resources
- Views
- Programmability
- Schemas
- Security

SQL script 1 SQL script 2

Run Undo Publish Query plan Connect to sqlpool00 Use database sqlpool00

```

1 SELECT TOP (100) [AgencyID]
2 ,[AgencyName]
3 FROM [dbo].[NYC_Payroll_AGENCY_MD]

```

Results Messages

View Table Chart Export results

Search

AgencyID	AgencyName
2009	BOROUGH PRESIDENT-STATEN IS
2010	BRONX COMMUNITY BOARD #10
2011	BRONX COMMUNITY BOARD #11
2012	BRONX COMMUNITY BOARD #12
2013	BRONX COMMUNITY BOARD #2
2014	BRONX COMMUNITY BOARD #3
2015	BRONX COMMUNITY BOARD #4

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

The screenshot shows the Azure Data Factory interface. The left sidebar navigation includes 'Dashboards', 'Runs', 'Pipeline runs' (selected), 'Trigger runs', 'Runtimes & sessions', 'Integration runtimes', 'Data flow debug', 'Notifications', and 'Alerts & metrics'. The main content area displays a pipeline named 'Data Flow Load SQLDB Table To Synapse Table'. It shows a cluster startup time of 2m 31s, 2 transformations, and a success status. A data flow diagram shows a source (SQLDB) connected to a sink (Synapse Table). Below the diagram, under 'Sinks', is a table with one row: 'sinkSynapseNYCPayrollData' with a status of 'Succeeded'. The table includes columns for Sink, Status, Processing time, Highest processing, Rows written, Stages, and Lineage.

Data loaded to synapse Table

The screenshot shows the Azure SQL Database blade. On the left, the database structure is visible, including 'Tables' containing 'dbo.NYC_Payroll_AGENCY_MD', 'dbo.NYC_Payroll_Data', and 'dbo.NYC_Payroll_EMP_MD'. The 'dbo.NYC_Payroll_Data' table is selected. The right pane shows the results of a query:

```

8 , [AgencyStartDate]
9 , [WorkLocationBorough]
10 , [TitleCode]
11 , [TitleDescription]
12 , [LeaveStatusasofJune30]
13 , [BaseSalary]
14 , [PayBasis]
15 , [RegularHours]
16 , [RegularGrossPaid]
17 , [OTHours]
18 , [TotalOTPaid]
19 , [TotalOtherPay]
20 , FROM [dbo].[NYC_Payroll_Data]

```

The results table shows data for four employees across four fiscal years (2021-2022).

FiscalYear	PayrollNumber	AgencyID	AgencyName	EmployeeID	LastName	FirstName	AgencyStartDa...	WorkLocation...	TitleCode	TitleDescrij...
2021	466	2096	COMMUNITY C...	207168	MUNROE	ANTHONY	(NULL)	MANHATTAN	40640	PRESIDENT
2021	816	2129	DEPT OF HEALT...	105284	KELLY	SEAN	(NULL)	BROOKLYN	41011	CITY MEDIC
2021	816	2129	DEPT OF HEALT...	328990	SHERROCK	MICHAEL	(NULL)	MANHATTAN	40614	OILER
2021	868	2141	DEPT OF CITY...	98108	JOSEPH	SAMUEL	(NULL)	MANHATTAN	40782	STATIONAR

4.6.4. Screenshot of - Pipeline Load Employee Master To Synapse

The screenshot shows the Azure Data Factory interface. The left sidebar navigation includes 'Dashboards', 'Runs', 'Pipeline runs' (selected), 'Trigger runs', 'Runtimes & sessions', 'Integration runtimes', 'Data flow debug', 'Notifications', and 'Alerts & metrics'. The main content area displays a pipeline named 'Data Flow Load Employee MD'. It shows a cluster startup time of 2m 19s, 2 transformations, and a success status. A data flow diagram shows a source (SQLDB) connected to a sink (Synapse Table). Below the diagram, under 'Sinks', is a table with one row: 'ToSQLDedicatedPoolNYCPayrollE' with a status of 'Succeeded'. The table includes columns for Sink, Status, Processing time, Highest processing, Rows written, Stages, and Lineage.

Data loaded to synapse table

The screenshot shows the Azure Data Studio interface. On the left, there's a sidebar with icons for Home, Data, Develop, Integrate, Monitor, and Manage. The main area has tabs for Data, Workspace, and Linked. Under Data, it shows a workspace named 'sqlpool00 (SQL)' containing tables like 'dbo.NYC_Payroll_AGENCY_MD', 'dbo.NYC_Payroll_Data', 'dbo.NYC_Payroll_EMP_MD', and 'dbo.NYC_Payroll_TITLE_MD'. Below these are External tables, External resources, Views, Programmability, Schemas, and Security. A SQL script editor at the top contains the following code:

```

1 SELECT TOP (100) [EmployeeID]
2 ,[LastName]
3 ,[FirstName]
4 | FROM [dbo].[NYC_Payroll_EMP_MD]

```

The Results tab shows a table with the following data:

EmployeeID	LastName	FirstName
100001	AACHEN	DAVID
100426	ABDULLAH	NAQUAVIA
100633	ABODEELY	JULIE
100002	AACHEN	MONICA
100427	ABDULLAH PARWEZ	RAHILA

4.6.5. Screenshot of - Pipeline Load Title Master To Synapse

The screenshot shows the Azure Data Flow pipeline run status page. The left sidebar includes Dashboards, Runs, Pipeline runs, Trigger runs, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area displays a pipeline named 'Data Flow Load Title MD' with a green checkmark indicating success. It shows a cluster startup time of 2m 15s, 2 transformations, and a data flow status of Success. A 'Refresh' button and an 'Edit dataflow' link are available. Below this, the 'Sinks' tab is selected, showing a single sink named 'ToDedicatedSQLPoolNYCPayrollIT' with a status of 'Succeeded', processing time of 5s, highest processing of 1s 692ms, 1446 rows written, and 2 stages.

Data loaded to synapse table

The screenshot shows the Microsoft Azure Synapse Analytics workspace for 'synapse000'. The left sidebar is identical to the previous screenshot. The main area shows a workspace named 'sqlpool00 (SQL)' with tables like 'dbo.NYC_Payroll_AGENCY_MD', 'dbo.NYC_Payroll_Data', 'dbo.NYC_Payroll_EMP_MD', and 'dbo.NYC_Payroll_TITLE_MD'. A SQL script editor at the top contains the following code:

```

1 SELECT TOP (100) [TitleCode]
2 ,[TitleDescription]
3 | FROM [dbo].[NYC_Payroll_TITLE_MD]

```

The Results tab shows a table with the following data:

TitleCode	TitleDescription
40908	ADMINISTRATIVE SCHOOL FOOD SERVICE MANAGER
41164	ADMINISTRATIVE SUPERINTENDENT OF HIGHWAY OPERATIONS
40001	ADM SCHOOL SECURITY MANAGER-U
40257	CITY MEDICAL SPECIALIST
41345	FIELD SUPERVISOR
40438	ELECTRICAL ENGINEER

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

The screenshot shows the Azure Synapse Studio interface. On the left, the 'Workspace' navigation pane is open, showing 'SQL database' and 'sqlobject00 (SQL)'. Under 'Tables', there are four tables: 'dbo.NYC_Payroll_AGENCY_MD', 'dbo.NYC_Payroll_Data', 'dbo.NYC_Payroll_EMP_MD', and 'dbo.NYC_Payroll_Summary'. The 'dbo.NYC_Payroll_Summary' table is selected, and its 'Columns' are listed: 'FiscalYear (int, null)', 'AgencyName (varchar(50), null)', and 'TotalPaid (float, null)'. On the right, the 'SQL script 8' tab is active, displaying the following SQL code:

```
1 SELECT TOP (100) [FiscalYear]
2 , [AgencyName]
3 , [TotalPaid]
4 | FROM [dbo].[NYC_Payroll_Summary]
```

Dataset

The screenshot shows the Azure Data Factory interface. On the left, the 'Factory Resources' navigation pane is open, showing 'Datasets'. Under 'Datasets', there are several datasets: 'AgencyMaster', 'AzureSqlTable_NYC_Payroll_data', 'AzureSynapseAnalyticsTable_NYC_Pa...', 'AzureSynapseAnalyticsTable_NYC_Pa...', 'AzureSynapseAnalyticsTable_NYC_Pa...', 'AzureSynapseAnalyticsTable_NYC_Pa...', and 'EmpMaster'. A new dataset is being created, with the name 'table_synapse_nycpayroll_summary'. The 'Connection' tab is selected, showing 'Linked service *' set to 'AzureSynapseAnalyticsLinkServices'. The 'Properties' panel on the right shows the dataset's name as 'table_synapse_nycpayroll_summary'.

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

Name	Type	Value
FiscalYear	integer	2

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

The screenshot shows the 'Dataflow Aggregate...' interface. On the left, the 'actory Resources' sidebar lists various data flows and tables. In the main area, two source connectors ('sourceNYCpayrollData2020' and 'sourceSQLdbNYCpayrollData') merge into a 'Union' node. This union node then splits into '19 Columns'. Below the diagram, the 'Union settings' tab is selected, showing the output stream name 'union', a description of combining rows from the two sources, and options for 'Union by' (set to 'Name') and 'Union with'.

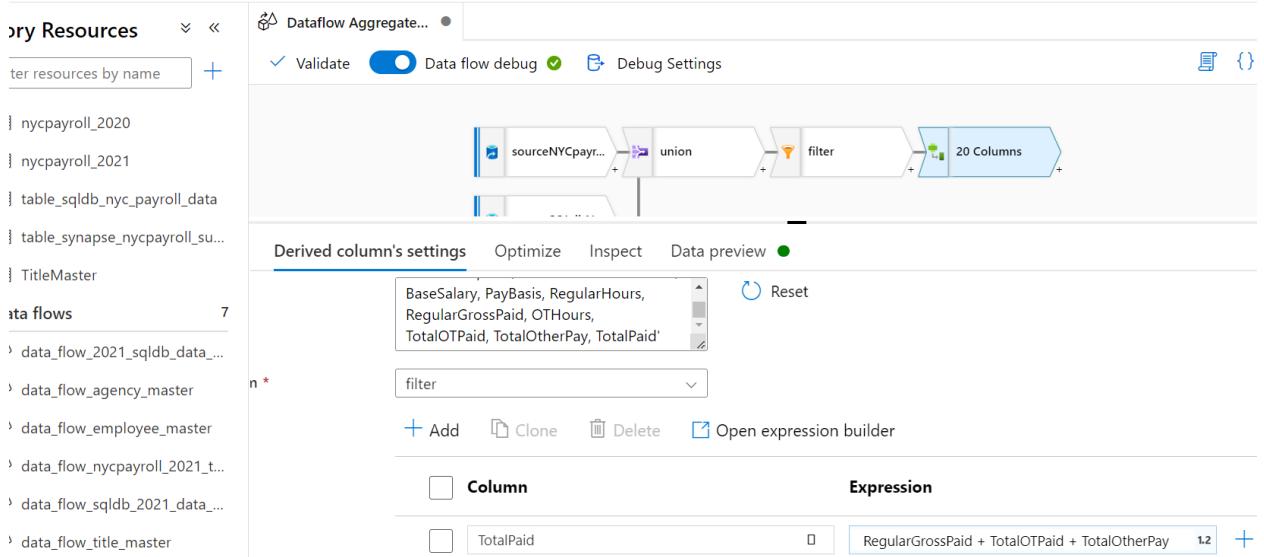
5.5. Add a Filter activity after Union

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

The screenshot shows the 'Dataflow Aggregate...' interface. The 'Filter settings' tab is selected. The 'Output stream name' is set to 'filter'. The 'Description' field contains the text 'Filtering rows using expressions on columns 'FiscalYear''. The 'Incoming stream' is set to 'union'. The 'Filter on' expression is set to `toInteger(FiscalYear) >= $dataflow_param_fiscalyear`.

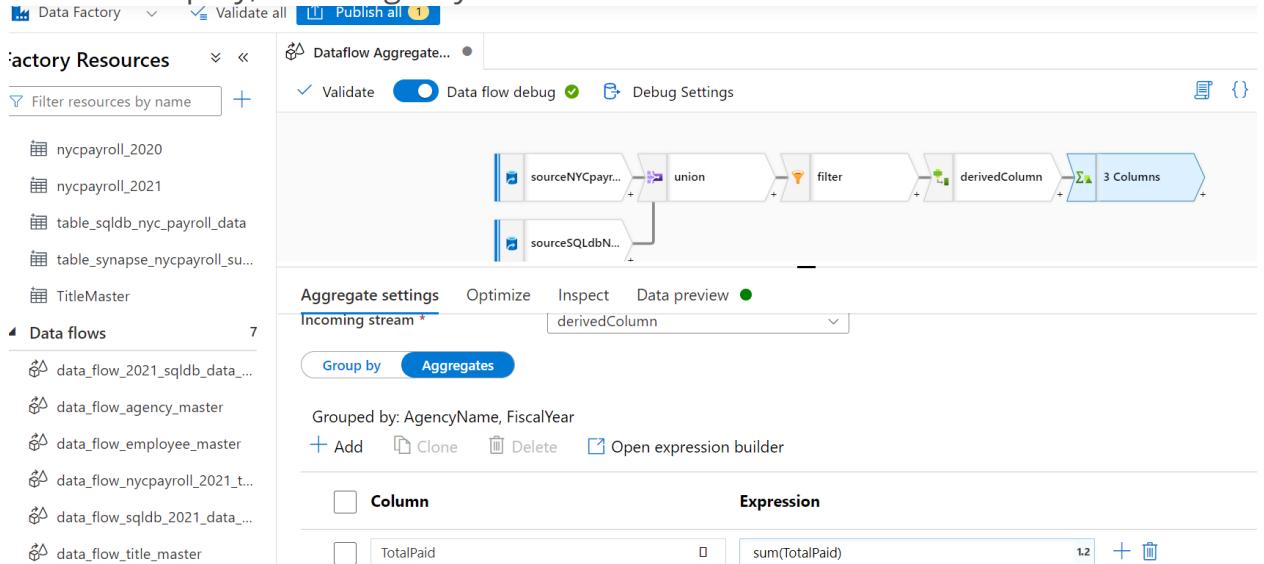
5.6. Derive a new TotalPaid column

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



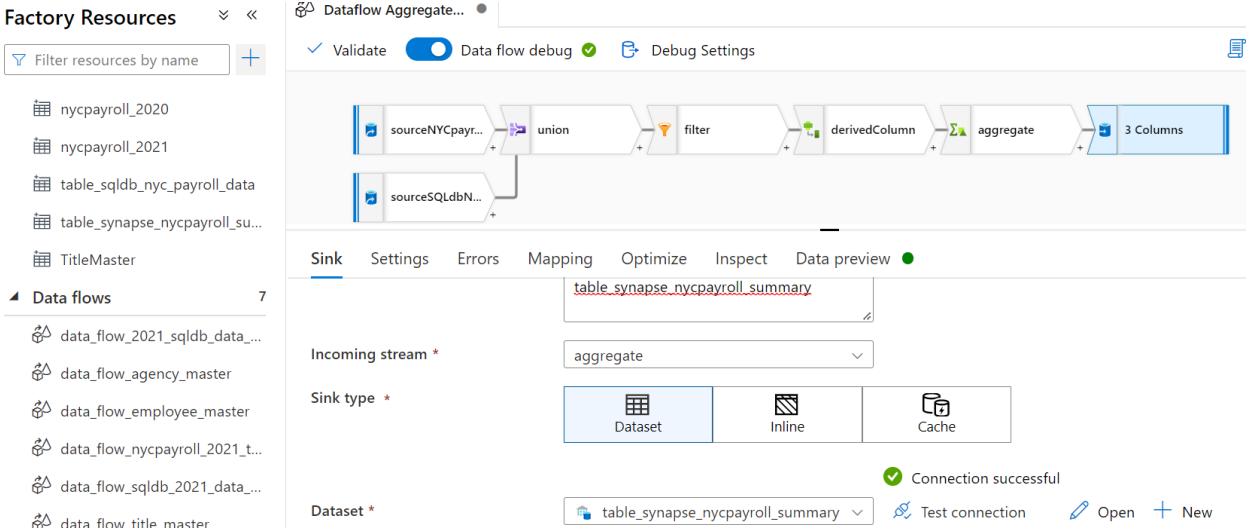
5.7. Add an Aggregate activity to the data flow next to the TotalPaid activity

- Under Group By, Select AgencyName and Fiscal Year

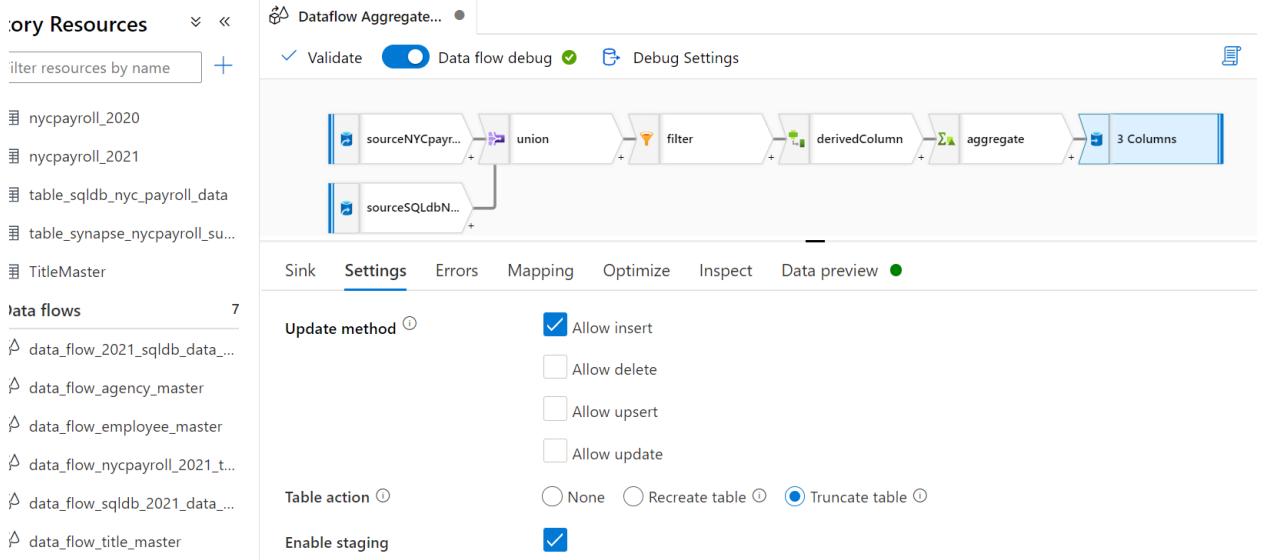


5.8. Add a Sink activity to the Data Flow

- 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

Factory Resources

- Pipelines (7)
- Datasets (12)

Data flow parameters

Name	Value
dataflow_param_fiscalyear	@pipeline().parameters.FiscalYear

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

Pipeline run

Trigger pipeline now using last published configuration.

Parameters

Name	Type	Value
FiscalYear	Int	2021

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

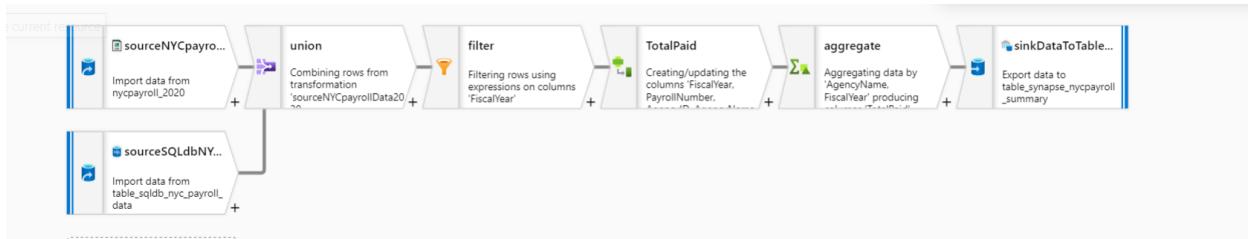
All pipeline runs > Pipeline Aggregate Paid - Activity runs > Data Flow Aggregate paid

Data Flow Aggregate paid

Cluster startup time: 2m 22s Number of transformations: 7 Data flow status: Success

Sinks All streams

Sink	Status	Processing time	Highest processing	Rows written	Stages	Lineage
sinkDataToTableSynapseNYCpayr	Succeeded	19s	6s	23		



Destination synapse table loaded with summary data

The screenshot shows the Synapse Analytics workspace interface:

- Workspace**: The current workspace is `synapse000`.
- Data**: The left sidebar shows the workspace structure, including a **SQL database** named `sqlpool00` containing tables like `dbo.NYC_Payroll_AGENCY_MD`, `dbo.NYC_Payroll_Data`, `dbo.NYC_Payroll_EMP_MD`, and `dbo.NYC_Payroll_Summary`.
- SQL script 8** and **SQL script 9** are displayed in the center, with **SQL script 9** containing the following query:

```

1 SELECT TOP (100) [FiscalYear]
2 ,[AgencyName]
3 ,[TotalPaid]
4 FROM [dbo].[NYC_Payroll_Summary]
    
```

- Results**: The results of the query are displayed in a table format:

FiscalYear	AgencyName	TotalPaid
2021	DEPT OF PARKS & RECREATION	1060979.32
2021	CAMPAIGN FINANCE BOARD	1077402.4
2021	COMMUNITY COLLEGE (BRONX)	1124184.68
2021	FIRE DEPARTMENT	9504918.2
2021	DEPT OF ED PEDAGOGICAL	2211971.4
2021	DEPT OF CITYWIDE ADMIN SVCS	7492881.56
2021	DEPT OF HEALTH/MENTAL HYGIE...	36226853.24
2021	GUTTMAN COMMUNITY COLLEGE	1467692.04
2021	COMMUNITY COLLEGE (LAGUAR...	1121040.8
2021	DEPT OF ENVIRONMENT PROTE...	3435209.6