

DATA INTEGRATION PIPELINES FOR NYC PAYROLL DATA ANALYTICS

Project Overview

Project Introduction

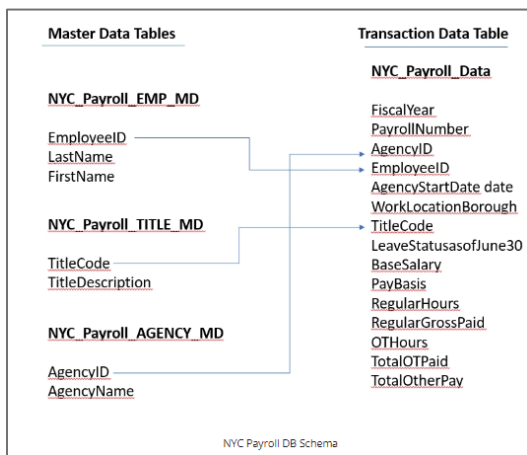
The City of New York would like to develop a Data Analytics platform on Azure Synapse Analytics to accomplish two primary objectives:

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**

Home > datapipenycstorageecc_1665706161620 | Overview > datapipenycstorageecc | Containers >

adlsyncpayroll-kellwyne

Container

Search

«

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: adlsyncpayroll-kellwyne

Search blobs by prefix (case-sensitive)

Name	Modified
<input type="checkbox"/> dirhistoryfiles	
<input type="checkbox"/> dirpayrollfiles	
<input type="checkbox"/> dirstaging	

Upload the files to the **dirpayrollfiles** folder

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

adlsyncpayroll-kellwyne

Container

Search

«

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: adlsyncpayroll-kellwyne / dirpayrollfiles

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> [-]						
<input type="checkbox"/> AgencyMaster.csv	10/14/2022, 1:16:48 PM	Hot (Inferred)		Block blob	4.9 KiB	Available
<input type="checkbox"/> EmpMaster.csv	10/14/2022, 1:16:48 PM	Hot (Inferred)		Block blob	23.04 KiB	Available
<input type="checkbox"/> nycpayroll_2021.csv	10/14/2022, 1:16:48 PM	Hot (Inferred)		Block blob	16.45 KiB	Available
<input type="checkbox"/> TitleMaster.csv	10/14/2022, 1:16:48 PM	Hot (Inferred)		Block blob	49.04 KiB	Available

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

- nycpayroll_2020.csv

adlsyncpayroll-kellwyne

Container

Search

«

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: adlsyncpayroll-kellwyne / dirhistoryfiles

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/> [-]				
<input type="checkbox"/> nycpayroll_2020.csv	10/14/2022, 1:17:51 PM	Hot (Inferred)		Block blob

2. Create an Azure Data Factory Resource

Home >

Microsoft.DataFactory-20221014132333 | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name: Microsoft.DataFactory-20221014132333
Subscription: Vocareum-UDA-1
Resource group: Regroup_4oyg3C

Start time: 10/14/2022, 1:43:04 PM
Correlation ID: 548606e3-58e6-4c4d-931a-58f923dc1cbb

Deployment details

Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

Deployment succeeded

Deployment "Microsoft.DataFactory-20221014132333" to resource group "Regroup_4oyg3C" was successful.

Pin to dashboard Go to resource group

Cost Management

Get notified to stay within your budget and prevent unexpected charges on your bill.

Set up cost alerts >

Microsoft Defender for Cloud

Secure your apps and infrastructure

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

Basics Networking Security Additional settings Tags Review + create

Product details

SQL database
by Microsoft
[Terms of use](#) | [Privacy policy](#)

Estimated cost per month
5.39 USD

Terms

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms of Use](#).

Basics

Subscription

Vocareum-UDA-1

Resource group

Regroup_4oyg3C

Region

Australia East

Database name

db_nycpayroll

Server

(new) datapipenycsql

Authentication method

SQL authentication

Server admin login

nycadmin01

Compute + storage

Basic: 2 GB storage

Backup storage redundancy

Locally-redundant backup storage

Networking

Allow Azure services and resources to access this server

No

Private endpoint

None

Create < Previous Download a template for automation

SQL

Cost summary

Basic (Basic)

Cost per DTU (in USD)

1.08

DTUs selected

x 5

ESTIMATED COST / MONTH

5.39 USD

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

Microsoft.Azure.SynapseAnalytics-20221014155720

Deployment

Overview

Search

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name: Microsoft.Azure.SynapseAnalytics-20221014155... Start time: 10/14/2022, 4:13:16 PM
Subscription: Vocareum-UDA-1
Resource group: Regroup_4oyg3C Correlation ID: cb2d3045-fa15-487a-91c6-d56b79ea9c8f

Deployment details

Resource	Type	Status	Operation details
✓ datapipesynapsespace/allowAll	Microsoft.Synapse/workspaces/firewallrules	OK	Operation details
✓ StorageRoleDeploymentResource	Microsoft.Resources/deployments	OK	Operation details
✓ datapipesynapsespace	Microsoft.Synapse/workspaces	OK	Operation details
✓ datapipesynapsespace	Microsoft.Synapse/workspaces	OK	Operation details
✓ synapsedatalakefileys	Microsoft.Resources/deployments	OK	Operation details

Next steps

Go to resource group

New dedicated SQL pool

✓ Validation succeeded.

Basics * Additional settings * Tags Review + create

Product details

Azure Synapse Analytics dedicated SQL pool by Microsoft
[Terms of use](#) | [Privacy policy](#)

Est. cost per hour
1.69 USD
[View pricing details](#)

Terms

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#)

Data source

Dedicated SQL pool name

datapipesynapsesqlpool

Performance level

DW100c

Additional settings

Use existing data

Blank

Collation

SQL_Latin1_General_CP1_CI_AS

datapipesynapsespace

better experience. [Learn more](#)

Update all [Publish all](#)

SQL pools

The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured to adapt to team or organizational requirements and constraints.

[+ New](#) [Refresh](#)

Filter by name

Showing 1-2 of 2 items (1 Serverless, 1 Dedicated)

Name	Type	Status	Size
Built-in	Serverless	✓ Online	Auto
datapipesynapsesqlpool	Dedicated	✓ Online	DW100c

Create master data tables and payroll transaction tables:

Microsoft Azure | Synapse Analytics | datapipesynapsespace

Synapse live [Validate all](#) [Publish all](#)

Data

Workspace Linked

Filter resources by name

SQL database 1

datapipesynapsesqlpool (SQL)

Tables

- dbo.NYC_Payroll_AGENCY_MD
- dbo.NYC_Payroll_Data
- dbo.NYC_Payroll_EMP_MD
- dbo.NYC_Payroll_TITLE_MD

External tables

External resources

Views

Programmability

Schemas

Security

SQL script 1

```

1 CREATE TABLE [dbo].[NYC_Payroll_EMP_MD](
2     [EmployeeID] [varchar](10) NULL,
3     [LastName] [varchar](20) NULL,
4     [FirstName] [varchar](20) NULL
5 )
6 GO
7
8 CREATE TABLE [dbo].[NYC_Payroll_TITLE_MD](
9     [TitleCode] [varchar](10) NULL,
10    [TitleDescription] [varchar](100) NULL
11 )
12 GO
13
14 CREATE TABLE [dbo].[NYC_Payroll_AGENCY_MD](
15     [AgencyID] [varchar](10) NULL,
16     [AgencyName] [varchar](50) NULL
17 )
18 GO
19
20 CREATE TABLE [dbo].[NYC_Payroll_Data](
21     [FiscalYear] [int] NULL,
22     [PayrollNumber] [int] NULL,
23     [AgencyID] [varchar](10) NULL,
24     [AgencyName] [varchar](50) NULL,
25     [EmployeeID] [varchar](10) NULL,
26     [LastName] [varchar](20) NULL,

```

Connect to datapipesynapsesqlpool Use database datapipesynapsesqlpool

Microsoft Azure | Synapse Analytics | datapipesynapsespace

Synapse live [Validate all](#) [Publish all](#)

Data

Workspace Linked

Filter resources by name

Azure Data Lake Storage Gen2 2

datapipesynapsespace (Primary - datapipenycs...)

synapsedatalakefilesys (Primary)

adlsyncpayroll-kellyne

(Attached Containers)

SQL script 1







adlsyncpayroll-kellyne x synapsedatalakefilesys

New SQL script New notebook New data flow New integration dataset Upload Download

adlsyncpayroll-kellyne > dirpayrollfiles

Name	Last Modified
AgencyMaster.csv	10/14/2022, 1:16:48 PM
EmpMaster.csv	10/14/2022, 1:16:48 PM
nycpayroll_2021.csv	10/14/2022, 1:16:48 PM
TitleMaster.csv	10/14/2022, 1:16:48 PM


Azure Resources Completed:

Resources		
<div>Recent Favorite</div>		
Name	Type	Last Viewed
 db_nycpayroll (datapipenycsql/db_nycpayroll)	SQL database	5 minutes ago
 datapipenycsql	SQL server	12 minutes ago
 Regroup_4oyg3C	Resource group	35 minutes ago
 datapipenycadf	Data factory (V2)	50 minutes ago
 datapipenycstorageacc	Storage account	an hour ago
 Vocareum-UDA-1	Subscription	2 hours ago
See all		

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


datapipenycadf


Validate all  Publish all

<<

Linked services






Linked service defines the connection information to a data store or compute. [Learn more](#) 

 New

 Filter by name

Annotations : **Any**

Showing 1 - 3 of 3 items

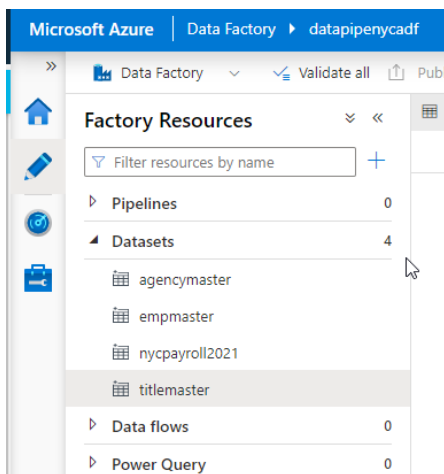
Name 	Type 
 ls_azuresqldatabase1	Azure SQL Database
 ls_datalakestorage1	Azure Data Lake Storage Gen2
 ls_synapseanalytics1	Azure Synapse Analytics

Step 3: Create Datasets in Azure Data Factory

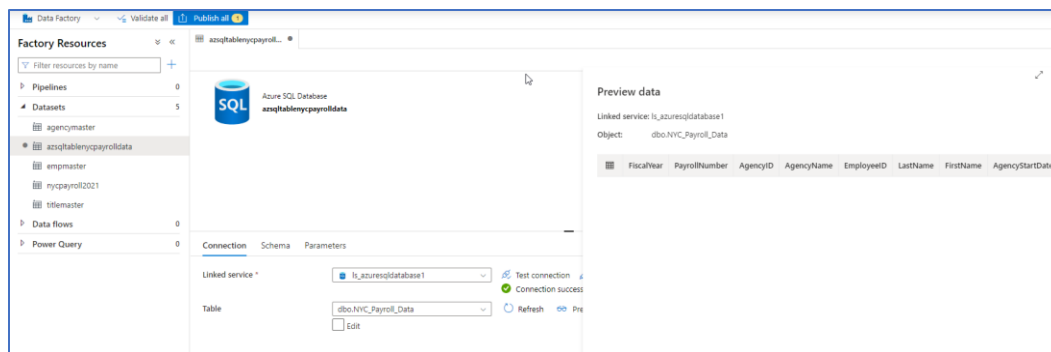
1. Create the datasets for the 2021 Payroll file on Azure Data Lake Gen2

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

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

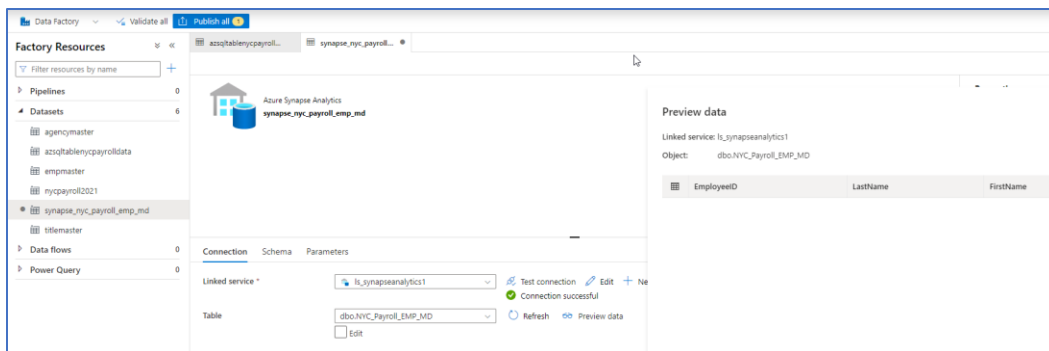


3. Create the dataset for transaction data table that should contain current (2021) data in SQL DB



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

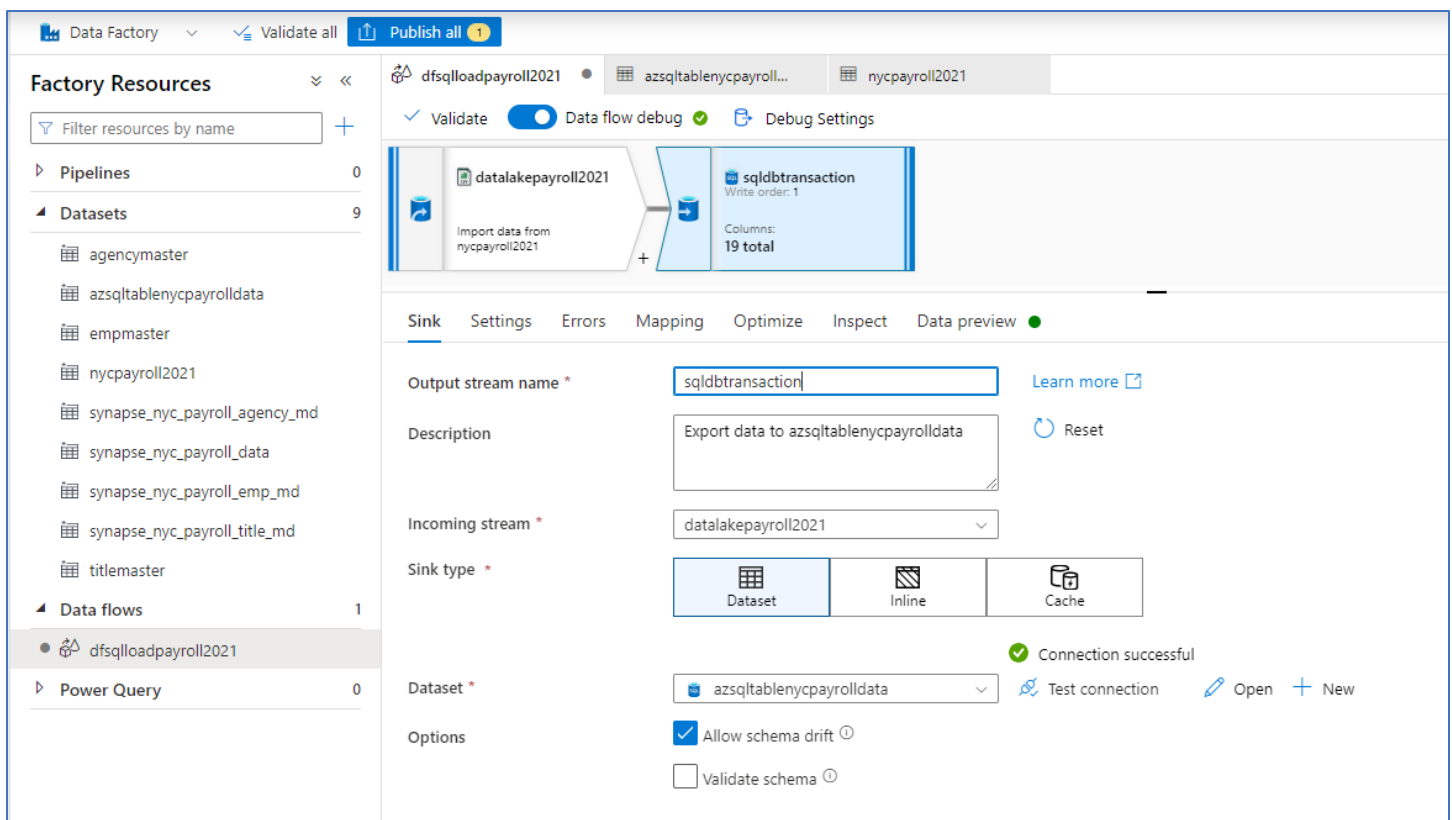
- dataset for NYC_Payroll_EMP_MD
- for NYC_Payroll_TITLE_MD
- for NYC_Payroll_AGENCY_MD
- for NYC_Payroll_Data



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).



Factory Resources

Filter resources by name

- Pipelines: 0
- Datasets: 9
 - agencymaster
 - azsqltablenycpayrolldata
 - empmaster
 - nycpayroll2021
 - synapse_nyc_payroll_agency_md
 - synapse_nyc_payroll_data
 - synapse_nyc_payroll_emp_md
 - synapse_nyc_payroll_title_md
 - titlemaster
- Data flows: 1
 - dfsqloadpayroll2021
- Power Query: 0

Source settings | Source options | Projection | Optimize | Inspect | Data preview

Output stream name: [Learn more](#)

Description: [Reset](#)

Source type: ☒ Dataset ☐ Inline

Dataset: [Test connection](#) [Open](#) [New](#)

Options:

- ☒ Allow schema drift
- ☐ Infer drifted column types
- ☐ Validate schema

Skip line count:

Sampling: ☐ Enable ☒ Disable

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

Factory Resources

Filter resources by name

- Pipelines: 1
 - pipelinesqloadpayroll2021
- Datasets: 9
 - agencymaster
 - azsqltablenycpayrolldata
 - empmaster
 - nycpayroll2021
 - synapse_nyc_payroll_agency_md
 - synapse_nyc_payroll_data
 - synapse_nyc_payroll_emp_md
 - synapse_nyc_payroll_title_md
 - titlemaster
- Data flows: 1
 - dfsqloadpayroll2021
- Power Query: 0

Activities

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Settings

Data flow: [Open](#) [New](#)

Run on (Azure IR):

Compute size:

Advanced

Logging level: ☒ Verbose ☐ Basic ☐ None

Sink properties

Staging: [New](#)

Staging linked service: [New](#)

Staging storage folder: / [Browse](#)

Properties

General

Name:

Description:

Annotations: [New](#)

Microsoft Azure | Data Factory | datapipenycadf

Pipeline runs

Triggered Debug Rerun Cancel options Refresh Edit columns List Gantt

Filter by run ID or name Auckland, Wellington...: Last 24 hours Pipeline name: All Status: All Runs: Latest runs Triggered by: All Add filter Copy filters Export to CSV

Showing 1 - 1 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Error	Run	Parameters	Annotations	Run ID
pipeline\$sqlloadpayroll2021	Oct 14, 2022, 7:57:15 pm	--	00:00:00	Manual trigger	Queued		Original			1df27822-28b3-42c1-

All pipeline runs > pipelinesqlloadpayroll2021 - Activity runs > dfsqlloadpayroll2021

dfsqlloadpayroll2021

Cluster startup time: 3m 21s Number of transformations: 2 Data flow status: Success

Refresh Auto refresh On Edit dataflow

Sinks All streams

Sink	Status	Processing time	Highest processing time	Rows written	Stages	Lineage
sqldbtransaction	Succeeded	4s 717ms	3s 767ms	101		

Stages

Transformations Stages

sqldbtransaction

Processing time: 4s 717ms

TRANSFORM	ROWS	TIME
sqldbtransaction	101	
datalakepayroll2021	101	3s 767ms

Close

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

Factory Resources

- Pipelines**
 - Pipeline Load Employee Master to S...
 - Pipeline Load Agency Master to Syn...
 - Pipeline Load Title Master to Synapse
 - pipelinesqlloadpayroll2021
- Datasets**
 - agencymaster
 - azsqltablencypayrolldata
 - empmaster
 - nycpayroll2021
 - synapse_nyc_payroll_agency_md
 - synapse_nyc_payroll_data
 - synapse_nyc_payroll_emp_md
 - synapse_nyc_payroll_title_md
 - titlemaster
- Data flows**
 - dataflow_employeeemaster
 - dataflow_agencyemaster
 - dataflow_titleemaster
 - dfsloadpayroll2021
- Power Query**
 - 0

Source settings | Source options | Projection | Optimize | Inspect | Data preview

Output stream name * datalakeemployeeemaster [Learn more](#)

Description Import data from empmaster [Reset](#)

Source type * Dataset | Inline

Dataset * empmaster [Test connection](#) [Open](#) [New](#)

Options

- ☒ Allow schema drift
- ☐ Infer drifted column types
- ☐ Validate schema

Skip line count

Sampling * ☐ Enable ☒ Disable

Properties

General | Related

Name * dataflow_employeeemaster

Description data flows to load the data from the data lake files into the Synapse Analytics data tables

Factory Resources

- Pipelines**
 - Pipeline Load Employee Master to S...
 - Pipeline Load Agency Master to Syn...
 - Pipeline Load Title Master to Synapse
 - pipelinesqlloadpayroll2021
- Datasets**
 - agencymaster
 - azsqltablencypayrolldata
 - empmaster
 - nycpayroll2021
 - synapse_nyc_payroll_agency_md
 - synapse_nyc_payroll_data
 - synapse_nyc_payroll_emp_md
 - synapse_nyc_payroll_title_md
 - titlemaster
- Data flows**
 - dataflow_employeeemaster
 - dataflow_agencyemaster
 - dataflow_titleemaster
 - dfsloadpayroll2021

Sink | Settings | Errors | Mapping | Optimize | Inspect | Data preview

Output stream name * synapseemployeeemaster [Learn more](#)

Description Export data to synapse_nyc_payroll_emp_md [Reset](#)

Incoming stream * datalakeemployeeemaster

Sink type * Dataset | Inline | Cache

Dataset * synapse_nyc_payroll_emp_md [Test connection](#) [Open](#) [New](#)

Options

- ☒ Allow schema drift
- ☐ Validate schema

Properties

General | Related

Name * dataflow_employeeemaster

Description data flows to load the data from the data lake files into the Synapse Analytics data tables

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

Microsoft Azure | Data Factory | datapipenycadl

Pipeline runs

Triggered | Debug | Run | Cancel options | Refresh | Edit columns | List | Get

Filter by run ID or name | Auckland, Wellington... | Last 24 hours | Pipeline name: All | Status: All | Runs: Latest runs | Triggered by: All | Add filter | Copy filters | Export to CSV | Last refreshed 0 minutes ago

	Run start	Run end	Duration	Triggered by	Status	Error	Run	Parameters	Annotations
<input type="checkbox"/> pipelinesqlloadpayroll2021	Oct 14, 2022, 8:46:23 pm	Oct 14, 2022, 8:50:17 pm	00:03:54	Manual trigger	Succeeded		Original		
<input type="checkbox"/> pipelinesqlloadpayroll2021	Oct 14, 2022, 7:57:15 pm	Oct 14, 2022, 8:02:06 pm	00:04:52	Manual trigger	Succeeded		Original		
<input type="checkbox"/> Pipeline Load Title Master to Synapse	Oct 14, 2022, 10:48:49 pm	...	00:00:45	Manual trigger	In progress		Original		
<input type="checkbox"/> Pipeline Load Employee Master to Synapse	Oct 14, 2022, 10:49:00 pm	...	00:00:34	Manual trigger	In progress		Original		
<input type="checkbox"/> Pipeline Load Current Year Data from SQLDB to Synapse	Oct 14, 2022, 10:49:15 pm	...	00:00:19	Manual trigger	In progress		Original		
<input type="checkbox"/> Pipeline Load Agency Master to Synapse	Oct 14, 2022, 10:49:05 pm	...	00:00:29	Manual trigger	In progress		Original		

6. Trigger and monitor the Pipelines

Microsoft Azure

Data Factory > datapipenycadf

Home

Dashboard

Runs

Pipeline runs

Trigger runs

Runtimes & sessions

Integration runtimes

Data flow debug

Notifications

Alerts & metrics

Pipeline runs

Triggered Debug Rerun Cancel options Refresh Edit columns List Gantt

Filter by run ID or name Auckland, Wellington... Last 24 hours Pipeline name : All Status : All Runs : Latest runs Triggered by : All Add filter

Showing 1 - 6 items

<input type="checkbox"/>	Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓	Error	Run
<input type="checkbox"/>	Pipeline Load Current Year Data from SQLDB to Synapse	Oct 14, 2022, 10:49:15 pm	Oct 14, 2022, 10:52:57 pm	00:03:42	Manual trigger	✓ Succeeded		Original
<input type="checkbox"/>	Pipeline Load Agency Master to Synapse	Oct 14, 2022, 10:49:05 pm	Oct 14, 2022, 10:52:50 pm	00:03:44	Manual trigger	✓ Succeeded		Original
<input type="checkbox"/>	Pipeline Load Employee Master to Synapse	Oct 14, 2022, 10:49:00 pm	Oct 14, 2022, 10:52:56 pm	00:03:56	Manual trigger	✓ Succeeded		Original
<input type="checkbox"/>	Pipeline Load Title Master to Synapse	Oct 14, 2022, 10:48:49 pm	Oct 14, 2022, 10:53:02 pm	00:04:13	Manual trigger	✓ Succeeded		Original
<input type="checkbox"/>	pipelinesqlloadpayroll2021	Oct 14, 2022, 8:46:23 pm	Oct 14, 2022, 8:50:17 pm	00:03:54	Manual trigger	✓ Succeeded		Original
<input type="checkbox"/>	pipelinesqlloadpayroll2021	Oct 14, 2022, 7:57:15 pm	Oct 14, 2022, 8:02:08 pm	00:04:52	Manual trigger	✓ Succeeded		Original

Microsoft Azure

Data Factory > datapipenycadf

Search

Validate all Publish all

Factory validation output

Home

Dashboard

Runs

Pipeline runs

Trigger runs

Runtimes & sessions

Integration runtimes

Data flow debug

Notifications

Alerts & metrics

Factory Resources

Filter resources by name

Pipelines 5

- Pipeline Load Current Year Data from SQLDB to Synapse
- Pipeline Load Agency Master to Synapse
- Pipeline Load Employee Master to Synapse
- Pipeline Load Title Master to Synapse
- pipelinesqlloadpayroll2021

Datasets 9

- agencymaster
- azsqltablenycpayrolldata
- empmaster
- nycpayroll2021
- synapse_nyc_payroll_agency_md
- synapse_nyc_payroll_data
- synapse_nyc_payroll_emp_md
- synapse_nyc_payroll_title_md
- titlemaster

Data flows 5

- dataflow_payroll2021_sqldbtoynapsepool
- dataflow_agencymaster
- dataflow_employeeemaster
- dataflow_titlemaster
- dfsloadpayroll2021

Power Query 0

Select an item

Use the resource explorer to select or create a new item

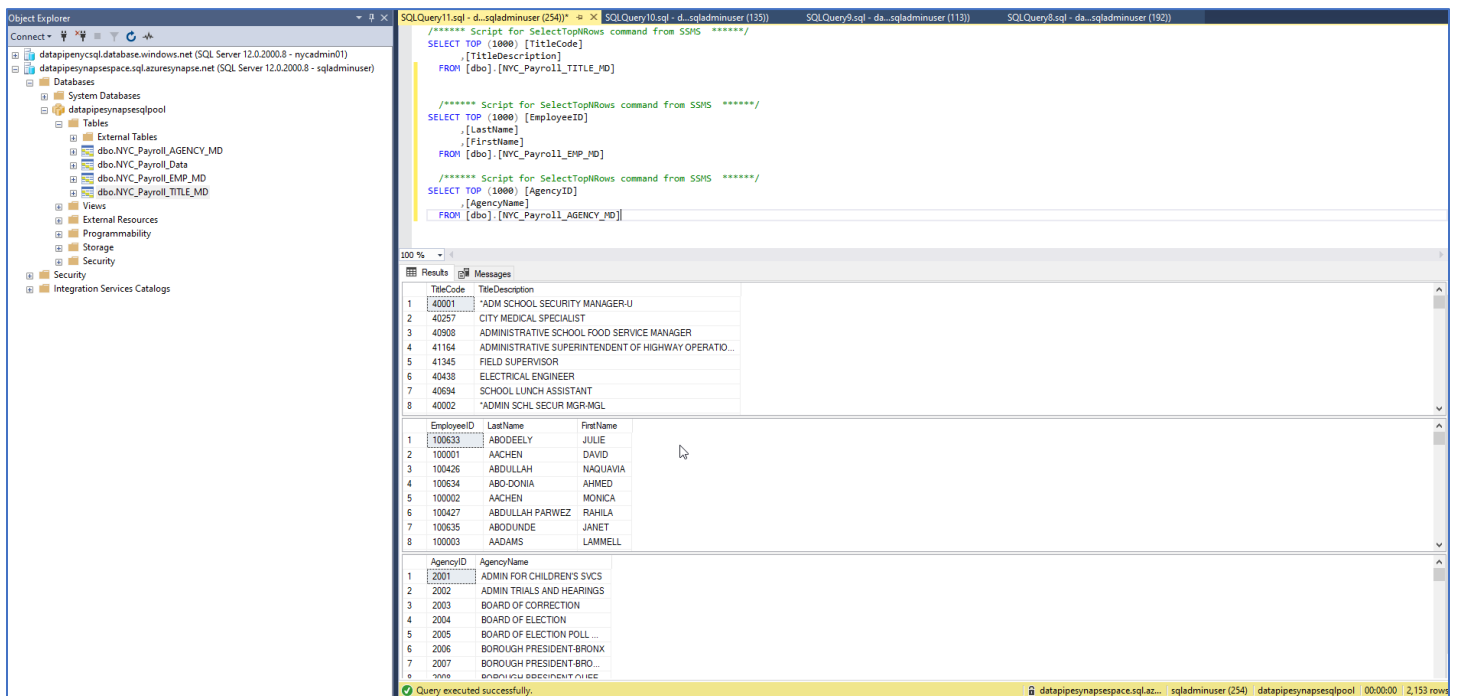
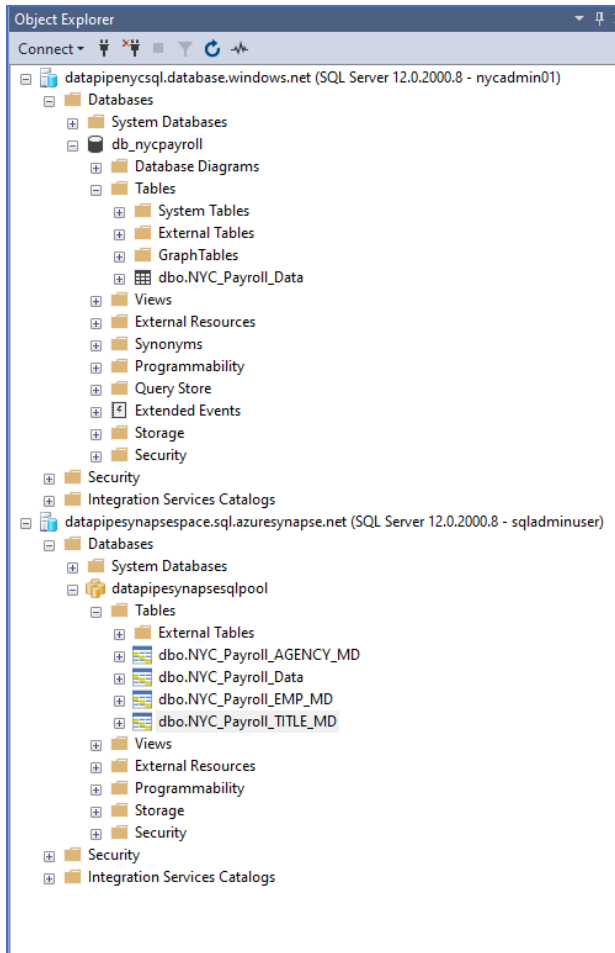
Factory validation output

Your factory has been validated.

No errors were found.

Close

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

Aggregate Data Flow

Step 5: Data Aggregation and Parameterization

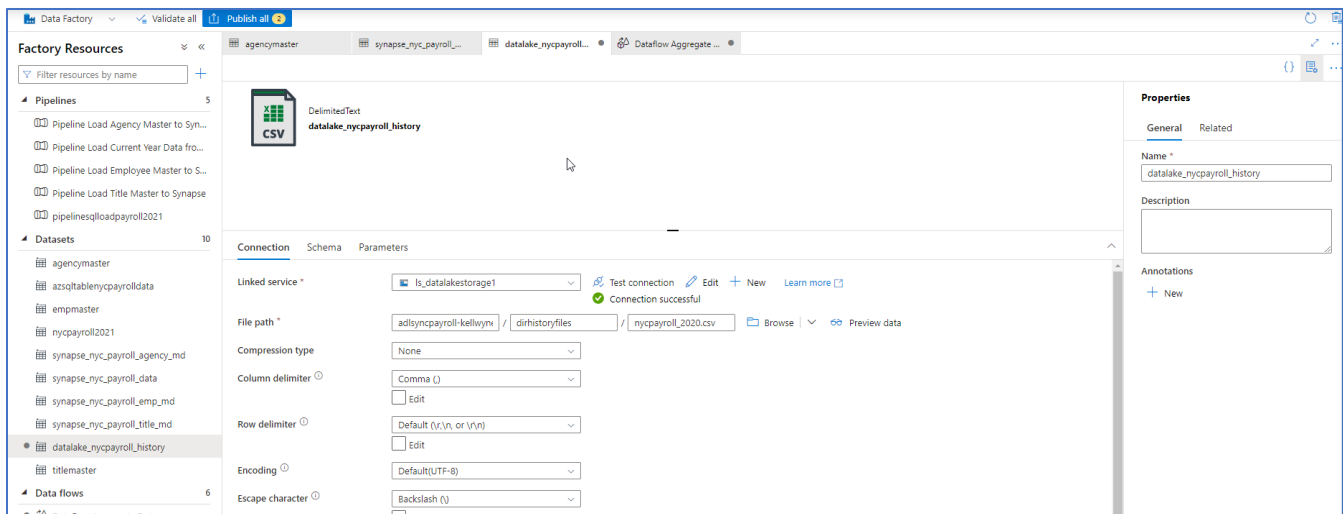
extract the 2021 year data and historical data, merge, aggregate and store it in Synapse Analytics. The aggregation will be on Agency Name, Fiscal Year and TotalPaid.

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

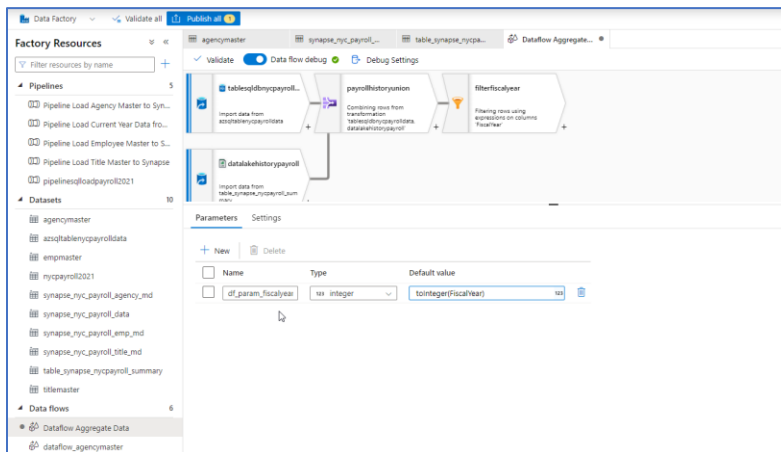
```

1 CREATE TABLE [dbo].[NYC_Payroll_Summary](
2     [FiscalYear] [int] NULL,
3     [AgencyName] [varchar](50) NULL,
4     [TotalPaid] [float] NULL
5 )

```



Setup dataflow parameter:

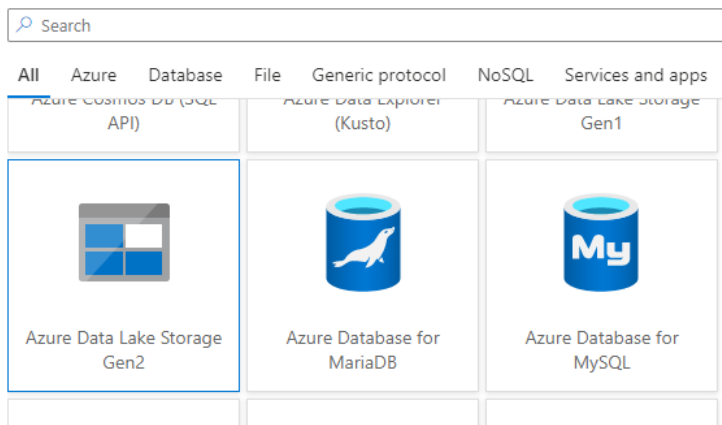


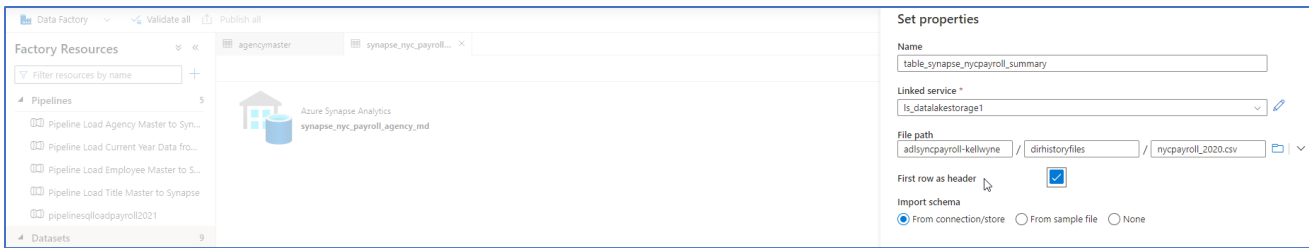
2.Create a new dataset for the Azure Data Lake Gen2 folder that contains the historical files.

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store





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

Publish all

You are about to publish all pending changes to the live environment. [Learn more](#)

Pending changes (4)

NAME	CHANGE	EXISTING
Pipelines		
Pipeline Load aggregate t...	(New)	-
Datasets		
datalake_nycpayroll_history	(Edited, Renamed)	table_synapse_nycpayroll_summary
table_synapse_nycpayroll_...	(New)	-
Data flows		
Dataflow Aggregate Data	(New)	-

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

Union settings

Output stream name *: payrollhistoryunion [Learn more](#)

Description: Combining rows from transformation 'tablesqldbnycpayrolldata, datalakehistorypayroll' [Reset](#)

Incoming stream *: tablesqldbnycpayrolldata

Union by *: ☒ Name ☐ Position

Union with *: datalakehistorypayroll

5.Add a Filter activity after Union

Factory Resources

Filter resources by name

Pipelines

- Pipeline Load Agency Master to Syn...
- Pipeline Load Current Year Data fro...
- Pipeline Load Employee Master to S...
- Pipeline Load Title Master to Synapse
- pipelinesqlloadpayroll2021

Datasets

- agencymaster
- azsqltablenycpayrolldata
- empmaster
- nycpayroll2021
- synapse_nyc_payroll_agency_md
- synapse_nyc_payroll_data
- synapse_nyc_payroll_emp_md
- synapse_nyc_payroll_title_md
- table_synapse_nycpayroll_summary
- titlemaster

Data flows

- Dataflow Aggregate Data
- dataflow_agencymaster
- dataflow_employeemaster
- dataflow_payroll2021_sqldbtosynaps...
- dataflow_titlemaster
- dfsloadpayroll2021

agencymaster synapse_nyc_payroll... table_synapse_nycpa... Dataflow Aggregate...

Validate Data flow debug Debug Settings

tablesqldbnycpayroll... payrollhistoryunion filterfiscalyear

Import data from azsqltablenycpayrolldata

Combining rows from transformation 'tablesqldbnycpayrolldata, datalakehistorypayroll'

Filtering rows using expressions on columns 'FiscalYear'

Columns: 19 total

Filter settings Optimize Inspect Data preview

Output stream name * filterfiscalyear

Description Filtering rows using expressions on columns 'FiscalYear'

Incoming stream * payrollhistoryunion

Filter on * toInteger(FiscalYear) >= \$df_param_fiscalyear

6. Derive a new TotalPaid column

agencymaster synapse_nyc_payroll... table_synapse_nycpa... Dataflow Aggregate...

Validate Data flow debug Debug Settings

tablesqldbnycpayroll... payrollhistoryunion filterfiscalyear derivedtotalpaid

Import data from azsqltablenycpayrolldata

Combining rows from transformation 'tablesqldbnycpayrolldata, datalakehistorypayroll'

Filtering rows using expressions on columns 'FiscalYear'

Columns: 20 total

Import data from table_synapse_nycpayroll_summary

Derived column's settings Optimize Inspect Data preview

Number of rows INSERT 100 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 ERROR 0 TOTAL

Refresh Typecast Modify Map drifted Statistics Remove Export to CSV

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

Dataflow expression builder Expression reference documentation

derivedtotalpaid

Derived Columns

+ Create new

=== TotalPaid

Column name *

TotalPaid

Expression

RegularGrossPaid + TotalOTPaid+TotalOtherPay

Expression elements

All

Functions

Input schema

Parameters

Cached lookup

Data flow library functions

Locals

Expression values

Filter by keyword

+ Create new

FiscalYear

PayrollNumber

AgencyID

AgencyName

EmployeeID

Data preview

Refresh

Output: TotalPaid	RegularGrossPaid	TotalOTPaid	TotalOtherPay
515260.3	257260.3	257260.3	258000.0
414018.05	413518.05	0.0	500.0
407532.25	132288.0	218628.18	56616.07
389594.77999999997	132288.0	218694.96	38611.82
375744.39	132288.0	192296.19	51160.2

Save and finish Cancel Clear contents

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

- Under Group By, Select AgencyName and Fiscal Year

Data Factory Validate all Publish all

Factory Resources

Filter resources by name

Pipelines

5

Pipeline Load Agency Master to Syn...

Pipeline Load Current Year Data fro...

Pipeline Load Employee Master to S...

Pipeline Load Title Master to Synapse

pipelinesqlloadpayroll2021

Datasets

11

agencymaster

azsqltablenycpayrolldata

datalake_nycpayroll_history

empmaster

nycpayroll2021

synapse_nyc_payroll_agency_md

synapse_nyc_payroll_data

synapse_nyc_payroll_emp_md

synapse_nyc_payroll_title_md

table_synapse_nycpayroll_summary_...

titlmaster

Data flows

6

Dataflow Aggregate Data

dataflow_agencymaster

dataflow_employeemaster

agencymaster

synapse_nyc_payroll_...

datalake_nycpayroll_...

Dataflow Aggregate...

table_synapse_nycpa...

Validate Data flow debug Debug Settings

tablesqldbnycpayroll...

payrollhistoryunion

filterfiscalyear

derivedtotalpaid

aggregatebyAgencyNa...

Import data from azsqltablenycpayrolldata

Combining rows from transformation Tablesqldbnycpayrolldata, datalakehistorypayroll

Filtering rows using expressions on columns 'FiscalYear'

Creating/updating the columns 'FiscalYear', 'PayrollNumber', 'AgencyID', 'AgencyName', 'EmployeeID', 'Lastname', 'EmployeeID'

Columns: 3 total

Aggregate settings Optimize Inspect Data preview

Output stream name *

aggregatebyAgencyNameFiscalYear

Description

Aggregating data by 'FiscalYear', 'AgencyName' producing columns 'TotalPaid'

Incoming stream *

derivedtotalpaid

Group by Aggregates

Columns	Name as
FiscalYear	FiscalYear
AgencyName	AgencyName

agencymaster | synapse_nyc_payroll... | datalake_nycpayroll... | Dataflow Aggregate... | table_synapse_nycpa...

Validate | Data flow debug | Debug Settings

5 tablesqldbnycpayroll... payrollhistoryunion filterfiscalyear derivedtotalpaid aggregatebyAgencyNa... sink1

Import data from assttablenycpayrolldata + Combining rows from transformation 'tablesqldbnycpayrolldata, datalakehistorypayroll' + Filtering rows using expressions on columns 'FiscalYear' + Creating/updating the columns 'FiscalYear, PayrollNumber, AgencyID, AgencyName, EmployeeID, LastName, ...' + Columns: 3 total + Add sink dataset

11 Aggregate settings Optimize Inspect **Data preview**

Number of rows: INSERT 27 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 ERROR 0 TOTAL 27

Refresh Typecast Modify Map drifted Statistics Remove Export to CSV

FiscalYear	AgencyName	TotalPaid
2021	OFFICE OF THE ACT...	305032.32
2021	DEPARTMENT OF FI...	266873.62999999995
2021	DEPT OF HEALTH/M...	9056713.31
2021	COMMUNITY COLLE...	297484.08
2021	COMMUNITY COLLE...	275457.88
2021	DEPARTMENT OF SA...	3063746.62
2021	DEPT OF ED PEDAG...	552992.85
2021	OFFICE OF THE MAY...	284742.27999999997
2021	FIRE DEPARTMENT	2376229.5500000003
2021	COMMUNITY COLLE...	280260.2

Properties

General Related

Name * Dataflow Aggregate Data

Description

8.Add a Sink activity to the Data Flow

blesqldbnycpayroll... payrollhistoryunion filterfiscalyear derivedtotalpaid aggregatebyAgencyNa... aggrtosynapsepayrol...

t data from ablenypayrolldata + Combining rows from transformation 'tablesqldbnycpayrolldata, datalakehistorypayroll' + Filtering rows using expressions on columns 'FiscalYear' + Creating/updating the columns 'FiscalYear, PayrollNumber, AgencyID, AgencyName, EmployeeID, LastName, ...' + Aggregating data by 'FiscalYear, AgencyName' producing columns 'TotalPaid' + Columns: 3 total

6 datalakehistorypayroll t data from synapse_nycpayroll_sum

Sink Settings Errors **Mapping** Optimize Inspect Data preview

Options

☒ Skip duplicate input columns ⓘ

☒ Skip duplicate output columns ⓘ

☐ Auto mapping ⓘ Reset + Add mapping 🗑 Delete Output format 3 mappings: All outputs mapped

Input columns	Output columns
<input type="checkbox"/> 123 FiscalYear	<input type="checkbox"/> 123 FiscalYear + 🗑
<input type="checkbox"/> abc AgencyName	<input type="checkbox"/> abc AgencyName + 🗑
<input type="checkbox"/> 1.2 TotalPaid	<input type="checkbox"/> 1.2 TotalPaid + 🗑

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)

The screenshot shows the Azure Data Factory interface with the 'Dataflow Aggregate Data' activity selected. The 'Settings' tab is active, displaying the following configuration:

- Data flow ***: Dataflow Aggregate Data
- Run on (Azure IR) ***: AutoResolveIntegrationRuntime
- Compute size ***: Small
- Logging level ***: Verbose (selected), Basic, None
- Staging linked service**: ls_datalakestorage1 (Connection successful)
- Staging storage folder**: adlsyncpayroll-kellwyn / dirstaging

The 'Properties' pane on the right shows the activity name 'Pipeline Load aggregate to synapse' and a description field.

The screenshot shows the Azure Data Factory interface with the 'Dataflow Aggregate Data' activity selected. The 'Parameters' tab is active, displaying the following configuration:

Name	Value	Type	Expression
df_param_fiscalyear		integer	

A dropdown menu is open for the 'Value' field, showing options for 'Data flow expression' and 'Pipeline expression'.

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

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Activities' pane lists various activities, with 'Dataflow Aggregate Data' selected. The main canvas displays the 'Dataflow Aggregate Data' activity. Below the activity, the 'Parameters' tab is active, showing a table with the following data:

Name	Value
df_param_fiscalyear	@pipeline().parameters.pl_param_fiscalyear

On the right, the 'Pipeline expression builder' is open, showing the expression `@pipeline().parameters.pl_param_fiscalyear`. Below the expression builder, the 'Parameters' tab is active, showing a search bar and a list of parameters:

Name	Type	Value
pl_param_fiscalyear	Pipeline parameter	

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

The screenshot shows the Azure Data Factory pipeline editor. On the left, the 'Activities' pane lists various activities, with 'Dataflow Aggregate Data' selected. The main canvas displays the 'Dataflow Aggregate Data' activity. Below the activity, the 'Output' tab is active, showing a table with the following data:

Name	Type	Run start	Duration
Dataflow Aggregate Data	Data flow	2022-10-14T12:21:06.7668985	00:00:35

On the right, the 'Pipeline run' details are shown, including a warning message: 'Trigger pipeline now using last published configuration.' Below the warning, the 'Parameters' tab is active, showing a table with the following data:

Name	Type	Value
pl_param_fiscalyear	Int	2021

The screenshot shows the Microsoft Azure Data Factory console. On the left, the 'Factory Resources' pane is open, showing a list of pipelines and datasets. The 'Pipelines' section is expanded, showing a list of pipelines, including 'Pipeline Load aggregate to synapse'. The 'Datasets' section is also expanded, showing a list of datasets, including 'agencymaster'. On the right, the 'Pipeline run' details are shown, including a warning message: 'Trigger pipeline now using last published configuration.' Below the warning, the 'Parameters' tab is active, showing a table with the following data:

Name	Type	Run start	Duration	Status
Dataflow Aggregate Data	Data flow	2022-10-14T12:21:06.7668985	00:00:35	Succeeded

Microsoft Azure

Data Factory > datapipenycadf

»

«

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

filterfiscalyear

derivedtotalpaid

aggregatebyAgency...

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	3s 5ms
● filterfiscalyear	100	
● payrollhistoryunion	201	
● tablesqldbnycpayrolldata	101	
● datalakehistorypayroll	100	7s
● aggrtosynapsepayrollsu...	23	
● aggregatebyAgencyNam...	23	

Close

FACTORY RESOURCES:

Microsoft Azure | Data Factory > datapipenycadf

>> Data Factory Validate all Publish all

Factory Resources

Filter resources by name

Pipelines 6

- Pipeline Load aggregate to synapse
- Pipeline Load Agency Master to Synapse
- Pipeline Load Current Year Data from SQLDB to Synapse
- Pipeline Load Employee Master to Synapse
- Pipeline Load Title Master to Synapse
- pipelinesqlloadpayroll2021

Datasets 11

- agencymaster
- azsqltablenycpayrolldata
- datalake_nycpayroll_history
- empmaster
- nycpayroll2021
- synapse_nyc_payroll_agency_md
- synapse_nyc_payroll_data
- synapse_nyc_payroll_emp_md
- synapse_nyc_payroll_title_md
- synapse_nyc_payroll_title_md
- table_synapse_nycpayroll_summary_new
- titlemaster

Data flows 6

- Dataflow Aggregate Data
- dataflow_agencymaster
- dataflow_employeeemaster
- dataflow_payroll2021_sqldbtoynapsepool
- dataflow_titlemaster
- dfsloadpayroll2021