

Step 1: Prepare the Data Infrastructure

1. Create the data lake and upload data

The screenshot displays the Microsoft Azure portal interface for a storage account named 'nycpayrollstorageacct18'. The left sidebar shows the navigation menu with options like Overview, Activity log, Tags, and Data storage. The main content area is divided into sections: Essentials, Properties, Monitoring, Capabilities, Recommendations, Tutorials, and Tools + SDKs. The Essentials section provides key information about the storage account, including its resource group, location, primary/secondary locations, subscription ID, disk state, and tags. The Properties section shows various settings for Data Lake Storage, such as Hierarchical namespace, Default access tier, Blob anonymous access, Blob soft delete, Container soft delete, Versioning, and Change feed. The Security section lists settings like Require secure transfer for REST API operations, Storage account key access, Minimum TLS version, and Infrastructure encryption. The Networking section shows Public network access is enabled.

Property	Value
Resource group	ODL-DataEng-287671
Location	eastus
Primary/Secondary Location	Primary: East US, Secondary: West US
Subscription	UdacityDS_55
Subscription ID	850c12f5-152f-4692-a8e9-2a5d3b9f99db
Disk state	Primary: Available, Secondary: Available
Tags	Add tags
Performance	Standard
Replication	Read-access geo-redundant storage (RA-GRS)
Account kind	StorageV2 (general purpose v2)
Provisioning state	Succeeded
Created	9/25/2025, 9:55:20 AM

Property	Value
Hierarchical namespace	Enabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)
Container soft delete	Enabled (7 days)
Versioning	Disabled
Change feed	Disabled
Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled
Public network access	Enabled

Containers:

The screenshot displays the Microsoft Azure portal interface for the 'Containers' section of the storage account 'nycpayrollstorageacct18'. The left sidebar shows the navigation menu with options like Overview, Activity log, Tags, and Data storage. The main content area shows a table of containers. The table has columns for Name, Last modified, Anonymous access level, and Lease state. Two containers are listed: 'slogs' and 'nyccontainer', both created on 9/25/2025 at 9:55:44 AM and 9:57:05 AM respectively, with Private anonymous access level and Available lease state.

Name	Last modified	Anonymous access level	Lease state
slogs	9/25/2025, 9:55:44 AM	Private	Available
nyccontainer	9/25/2025, 9:57:05 AM	Private	Available

Directory:

Pre Joiner Batch - Nov 08 | Pro...

Proctor sent a message

Data Integration Pipelines for h...

nyccontainer - Microsoft Azure

+

portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/_/overview/storageAccountid/%2Fsubscriptions%2F850c12f5-152f-4692-a8e9-2a5d3b9f39db%2Ffr...

Microsoft Azure

Search resources, services, and docs (G+J)

Copilot

odl user_267671@udaci...
UDACITY - DS (UDACITYHOLON...

Home > nycpayrollstorageacct18_1758774298706 | Overview > nycpayrollstorageacct18 | Containers >

nyccontainer

Container

Search

«

+ Add Directory

↑ Upload

↻ Refresh

🗑 Delete

📄 Copy

📄 Paste

🔄 Rename

🔒 Acquire lease

🔒 Break lease

🔧 Edit columns

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

nyccontainer

Authentication method: Access key (Switch to Microsoft Entra user account)

Search blobs by prefix (case-sensitive)

Only show active objects

Showing all 3 items

<input type="checkbox"/>	Name	Last modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>	dirhistoryfiles	9/25/2025, 9:57:35 AM				...
<input type="checkbox"/>	dirpayrollfiles	9/25/2025, 9:57:21 AM				...
<input type="checkbox"/>	dirstaging	9/25/2025, 9:57:43 AM				...

Add or remove favorites by pressing Ctrl+Shift+F

27°C
Mostly cloudy

Search

📄 📄 📄 📄 📄 📄

ENG
IN

09:58
25-09-2025

- Dirpayrollfiles:

Pre Joiner Batch - Nov 08 | Pro...

Proctor sent a message

Data Integration Pipelines for h...

nyccontainer - Microsoft Azure

+

portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/_/overview/storageAccountid/%2Fsubscriptions%2F850c12f5-152f-4692-a8e9-2a5d3b9f39db%2Ffr...

Microsoft Azure

Search resources, services, and docs (G+J)

Copilot

odl user_267671@udaci...
UDACITY - DS (UDACITYHOLON...

Home > nycpayrollstorageacct18_1758774298706 | Overview > nycpayrollstorageacct18 | Containers >

nyccontainer

Container

Search

«

+ Add Directory

↑ Upload

↻ Refresh

🗑 Delete

📄 Copy

📄 Paste

🔄 Rename

🔒 Acquire lease

🔒 Break lease

🔧 Edit columns

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

nyccontainer > dirpayrollfiles

Authentication method: Access key (Switch to Microsoft Entra user account)

Search blobs by prefix (case-sensitive)

Only show active objects

Showing all 4 items

<input type="checkbox"/>	Name	Last modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>	[-]					...
<input type="checkbox"/>	AgencyMaster.csv	9/25/2025, 9:58:48 AM	Hot (Inferred)	Block blob	4.9 KiB	Available
<input type="checkbox"/>	EmpMaster.csv	9/25/2025, 9:58:49 AM	Hot (Inferred)	Block blob	23.04 KiB	Available
<input type="checkbox"/>	TitleMaster.csv	9/25/2025, 9:58:49 AM	Hot (Inferred)	Block blob	49.04 KiB	Available
<input type="checkbox"/>	nycpayroll_2021.csv	9/25/2025, 9:58:49 AM	Hot (Inferred)	Block blob	16.45 KiB	Available

Add or remove favorites by pressing Ctrl+Shift+F

27°C
Mostly cloudy

Search

📄 📄 📄 📄 📄 📄

ENG
IN

09:58
25-09-2025

• Dirhistoryfiles:

Microsoft Azure portal interface showing the 'nyccontainer' storage account overview. The 'dirhistoryfiles' container is selected, displaying a table of blobs. A notification at the top right states 'Successfully uploaded blob(s)'. The table shows one item: 'nycpayroll_2020.csv'.

Name	Last modified	Access tier	Blob type	Size	Lease state
nycpayroll_2020.csv	9/25/2025, 9:59:18 AM	Hot (Inferred)	Block blob	17.07 KiB	Available

2. Create an Azure Data Factory Resource

Microsoft Azure portal interface showing the 'nycdatafactory12' Data Factory (V2) overview. The 'Essentials' section displays metadata: Resource group (ODL-DataEng-287671), Status (Succeeded), Location (East US), Subscription (UdacityDS-55), and Subscription ID (850c12f5-152f-4692-a8e9-2a5d3b9f39db). The 'Azure Data Factory Studio' logo is prominent, with a 'Launch studio' button. Below are links for Quick Starts, Tutorials, Template Gallery, and Training Modules.

Azure Data Factory Studio

Launch studio

Quick Starts | Tutorials | Template Gallery | Training Modules

3. Create a SQL Database

The screenshot shows the Microsoft Azure portal interface for a SQL database named 'nycdatabase (nycserver12/nycdatabase)'. The page is titled 'nycdatabase (nycserver12/nycdatabase)' and includes a search bar, a 'Copy' button, and a 'Restore' button. The left sidebar shows the 'Overview' tab selected, with a list of navigation options including 'Activity log', 'Tags', 'Diagnose and solve problems', 'Query editor (preview)', 'Mirror database in Fabric (preview)', 'Resource visualizer', 'Settings', 'Data management', 'Replicas', 'Sync to other databases', 'Integrations', 'Power Platform', 'Security', 'Intelligent performance', and 'Monitoring'. The main content area displays the 'Essentials' section with a table of properties:

Property	Value
Resource group	ODL-DataEng-287671
Status	Online
Location	East US
Subscription	UdacityDS-55
Subscription ID	850c12f5-152f-4692-a8e9-2a5d3b9f99db
Server name	nycserver12
Elastic pool	No elastic pool
Connection strings	Show database connection strings
Pricing tier	Basic
Earliest restore point	2025-09-25 05:08 UTC

Below the table, there are tabs for 'Getting started', 'Monitoring', 'Properties', 'Features', 'Notifications (0)', 'Integrations', and 'Tutorials'. The 'Getting started' tab is active, showing a 'Start working with your database' section with a 'Learn more' link. Below this, there are four cards: 'Configure access', 'Connect to application', 'Start developing', and 'Mirror database in Fabric'.

4. Create a Synapse Analytics workspace

The screenshot shows the Microsoft Azure portal interface for a Synapse workspace named 'nycpayrollworkspace12'. The page is titled 'nycpayrollworkspace12' and includes a search bar, a 'Suggest a workload for this Synapse workspace' button, and a 'Delete' button. The left sidebar shows the 'Overview' tab selected, with a list of navigation options including 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Resource visualizer', 'Settings', 'Microsoft Entra ID', 'Properties', 'Locks', 'Analytics pools', 'SQL pools', 'Apache Spark pools', 'Data Explorer pools (preview)', 'Security', and 'Monitoring'. The main content area displays the 'Essentials' section with a table of properties:

Property	Value
Resource group	ODL-DataEng-287671
Status	Succeeded
Location	East US
Subscription	UdacityDS-55
Subscription ID	850c12f5-152f-4692-a8e9-2a5d3b9f99db
Managed virtual network	No
Managed Identity object	96f3cca8-23a8-4c35-817b-5a447f9471d7
Workspace web URL	https://web.azure.synapse.net/workspace=%2fsubscriptions%2f850c12f5-152f-4692-a8e9-2a5d3b9f99db
Tags	Add tags
Networking	Show firewall settings
Primary ADLS Gen2 account	https://nycpayrollstorageacctn18.dfs.core.windows.net
Primary ADLS Gen2 file system	nyccontainer
SQL admin username	azureuser
SQL Microsoft Entra admin	odl_user_287671@udacityhol.onmicrosoft.com
Dedicated SQL endpoint	nycpayrollworkspace12.sqlazuresynapse.net
Serverless SQL endpoint	nycpayrollworkspace12-ondemand.sqlazuresynapse.net
Development endpoint	https://nycpayrollworkspace12.dev.azure.synapse.net

Below the table, there are tabs for 'Getting started', 'Monitoring', 'Properties', 'Features', 'Notifications (0)', 'Integrations', and 'Tutorials'. The 'Getting started' tab is active, showing a 'Start working with your workspace' section with a 'Learn more' link. Below this, there are two cards: 'Open Synapse Studio' and 'Read documentation'.

5. Create summary data external table in Synapse Analytics workspace

External table created in Synapse

The screenshot displays the Microsoft Azure Synapse Analytics workspace interface. The left sidebar shows the 'Data' section with 'Workspace' and 'Linked' tabs. Under 'Workspace', the 'SQL database' is expanded, showing 'unknowndatabase (SQL)' and 'External tables'. The 'External tables' section is further expanded, showing 'dbo.NYC_Payroll_Summary'. The main pane shows the 'SQL script 1' editor with the following T-SQL code:

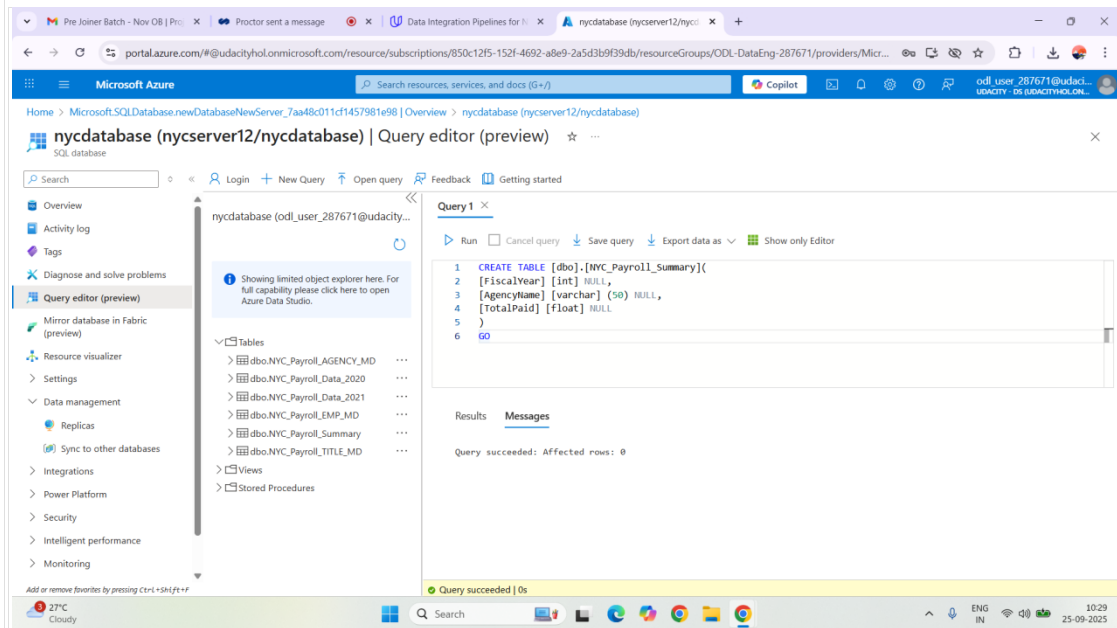
```
16  
17 CREATE EXTERNAL TABLE [dbo].[NYC_Payroll_Summary] (  
18     [FiscalYear] VARCHAR(50) NULL,  
19     [AgencyName] VARCHAR(50) NULL,  
20     [TotalPaid] VARCHAR(50) NULL  
21 )  
22 WITH (  
23     LOCATION = 'dirstaging/**',  
24     DATA_SOURCE = [nycontainer_nycpayrollstorageacct18_dfs_core_windows_net],  
25     FILE_FORMAT = [SynapseDelimitedTextFormat]  
26 )  
27 GO  
28  
29  
30 SELECT TOP 100 * FROM [dbo].[NYC_Payroll_Summary]  
31 GO
```

The 'Results' tab is selected, showing a table with the following columns: FiscalYear, AgencyName, and TotalPaid. The table is empty. The 'Properties' pane on the right shows the 'General' tab with the following details:

- Name: SQL script 1
- Description:
- Type: .sql script
- Size: 927 bytes
- Results settings per query: First 5000 rows (default)

The status bar at the bottom indicates '00:00:02 Query executed successfully.' and the system clock shows 10:57 on 25-09-2025.

6. Create master data tables and payroll transaction tables in SQL DB

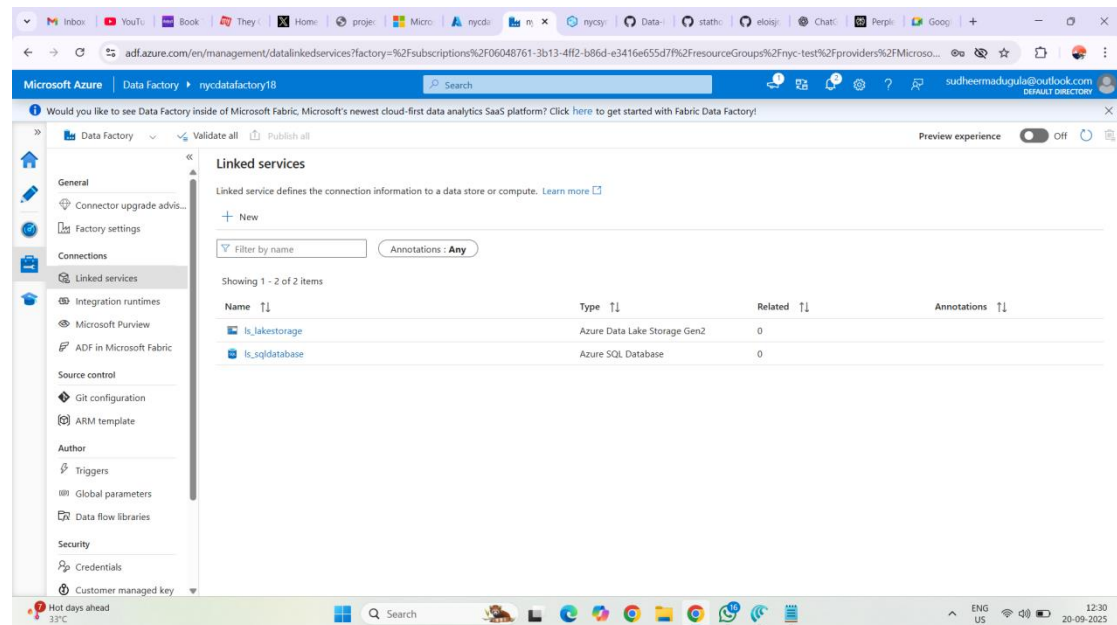


The screenshot shows the Microsoft Azure portal interface for a SQL database. The main window displays the Query editor (preview) for a database named 'nycdatabase (nycserver12/nycdatabase)'. The query editor shows a SQL query to create a table named [nyc_Payroll_Summary] with columns [fiscalYear], [AgencyName], and [TotalPaid]. The query is successful, showing 0 affected rows.

```
1 CREATE TABLE [dbo].[nyc_Payroll_Summary](
2 [fiscalYear] [int] NULL,
3 [AgencyName] [varchar] (50) NULL,
4 [TotalPaid] [float] NULL
5 )
6 GO
```

Query succeeded: Affected rows: 0

Step 2: Create Linked Services



The screenshot shows the Microsoft Azure portal interface for the Data Factory section. The main window displays the Linked services page, which lists two linked services: 'ls_lakestorage' (Azure Data Lake Storage Gen2) and 'ls_sqldatabase' (Azure SQL Database).

Name	Type	Related	Annotations
ls_lakestorage	Azure Data Lake Storage Gen2	0	
ls_sqldatabase	Azure SQL Database	0	

Step 3: Create Datasets in Azure Data Factory

The screenshot displays the Microsoft Azure Data Factory portal interface. The left sidebar shows the 'Factory Resources' tree with 'Datasets' expanded, listing various datasets including 'ds_nycpayroll_2020'. The main pane shows the configuration for the 'ds_nycpayroll_2020' dataset, which is an 'Azure SQL Database' type. The 'Connection' tab is active, showing the 'Linked service' as 'ads_linkedservice' and the 'Table' as 'dbo.NYC_Payroll_Data_2020'. The 'Properties' pane on the right shows the 'Name' as 'ds_nycpayroll_2020' and a 'Description' field. The bottom status bar indicates 'Hot days ahead 27°C'.

4: Create Data Flows

The screenshot displays the Microsoft Azure Data Factory portal interface. The left sidebar shows the 'Factory Resources' tree with 'Data flows' expanded, listing various data flows including 'dataflow_titlemaster'. The main pane shows the configuration for the 'dataflow_titlemaster' data flow. The 'Validate' tab is active, showing a data flow diagram with a source node 'source1' (Import data from ds_titlemaster) and a sink node 'sink1' (Export data to ds_titlemaster). The 'Parameters' pane on the right shows a 'New' button. The bottom status bar indicates '27°C Cloudy'.

Step 5: Data Aggregation and Parameterization

The screenshot displays the Microsoft Azure Data Factory interface for a data flow named 'dataflow_nycpayroll_summary'. The left sidebar shows the 'Factory Resources' tree with 'Data flows' expanded, listing several data flows including 'dataflow_nycpayroll_summary'. The main canvas shows a data flow diagram with the following components:

- source1**: Import data from 'dl_nycpayroll_2020'.
- union1**: Combining rows from transformation 'source1' and 'source2'.
- filter1**: Filtering rows using expressions on columns 'FiscalYear'.
- derivedColumn1**: Creating/updating the columns 'FiscalYear', 'AgencyName', 'AgencyID', 'AgencyPhone', 'EmployeeID', 'LastName'.
- aggregate1**: Aggregating data by 'FiscalYear', 'AgencyName' producing columns 'TotalPaid'.
- sink1**: Export data to 'dl_nycpayroll_summary'.
- source2**: Import data from 'dl_nycpayroll_2021'.
- aggregate1**: Aggregating data by 'FiscalYear', 'AgencyName' producing columns 'TotalPaid'.
- sink2**: Export data to 'dl_nycpayroll_summary'.

The 'Parameters' tab is active, showing a table with the following data:

Name	Type	Default value
dataflow_param_fiscalyear	int integer	2020

Step 6: Pipeline Creation

The screenshot displays the Microsoft Azure Data Factory interface for a pipeline named 'pipeline_nycpayroll'. The left sidebar shows the 'Factory Resources' tree with 'Pipelines' expanded, listing 'pipeline_nycpayroll'. The main canvas shows a pipeline diagram with the following components:

- Data flow**: Data flow_agency
- Data flow**: Data flow_emp
- Data flow**: Data flow_title
- Data flow**: Data flow_2021
- Data flow**: Data flow_2020
- Data flow**: Data flow_summary

The 'Parameters' tab is active, showing a table with the following data:

Name	Type	Default value
dataflow_param_fiscalyear	int integer	2020

Microsoft Azure | Data Factory | nycdatafactory12

All pipeline runs > pipeline_nycpayroll - Activity runs

Run Cancel Refresh Update pipeline List Gantt

Data flow Data flow_agency Data flow Data flow_2021 Data flow Data flow_2020 Data flow Data flow_title Data flow summary

Activity runs

Pipeline run ID 632c09dd-cde5-4bb9-a4d9-ec5c49e90768

All status Monitor in Azure Metrics Export to CSV

Showing 1 - 6 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Data flow summary	Succeeded	Data flow	9/25/2025, 11:45:32 AM	55s	AutoResolveIntegrationRuntime (East US)		0ceec778-53cd-4f...
Data flow_2021	Succeeded	Data flow	9/25/2025, 11:45:09 AM	22s	AutoResolveIntegrationRuntime (East US)		c3abd551-b87f-43...
Data flow_2020	Succeeded	Data flow	9/25/2025, 11:45:09 AM	20s	AutoResolveIntegrationRuntime (East US)		05c5849d-54cb-49c...
Data flow_agency	Succeeded	Data flow	9/25/2025, 11:39:38 AM	5m 29s	AutoResolveIntegrationRuntime (East US)		4705a2fa-3cc4-4f35...
Data flow_title	Succeeded	Data flow	9/25/2025, 11:39:38 AM	5m 14s	AutoResolveIntegrationRuntime (East US)		acea7cc5-5b20-40b...
Data flow_emp	Succeeded	Data flow	9/25/2025, 11:39:38 AM	5m 15s	AutoResolveIntegrationRuntime (East US)		d1727528-7487-41c...

Pipeline Running successfully without Errors:

Microsoft Azure | Data Factory | nycdatafactory12

All pipeline runs > pipeline_nycpayroll - Activity runs

Run Cancel Refresh Update pipeline List Gantt

Data flow Data flow_agency Data flow Data flow_2021 Data flow Data flow_2020 Data flow Data flow_title Data flow summary

Activity runs

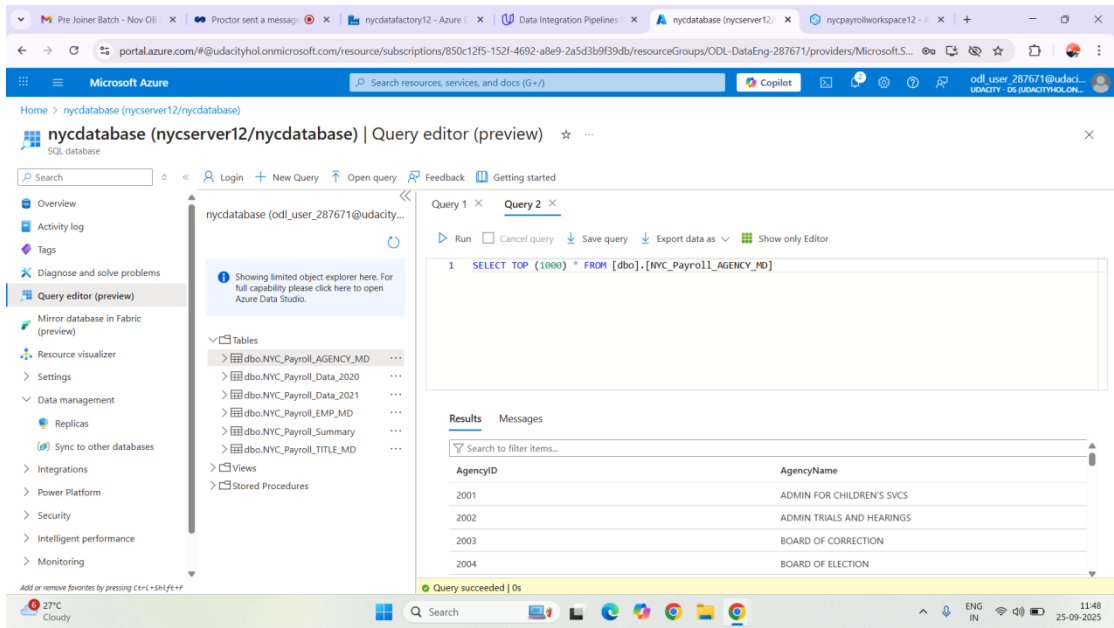
Pipeline run ID 632c09dd-cde5-4bb9-a4d9-ec5c49e90768

All status Monitor in Azure Metrics Export to CSV

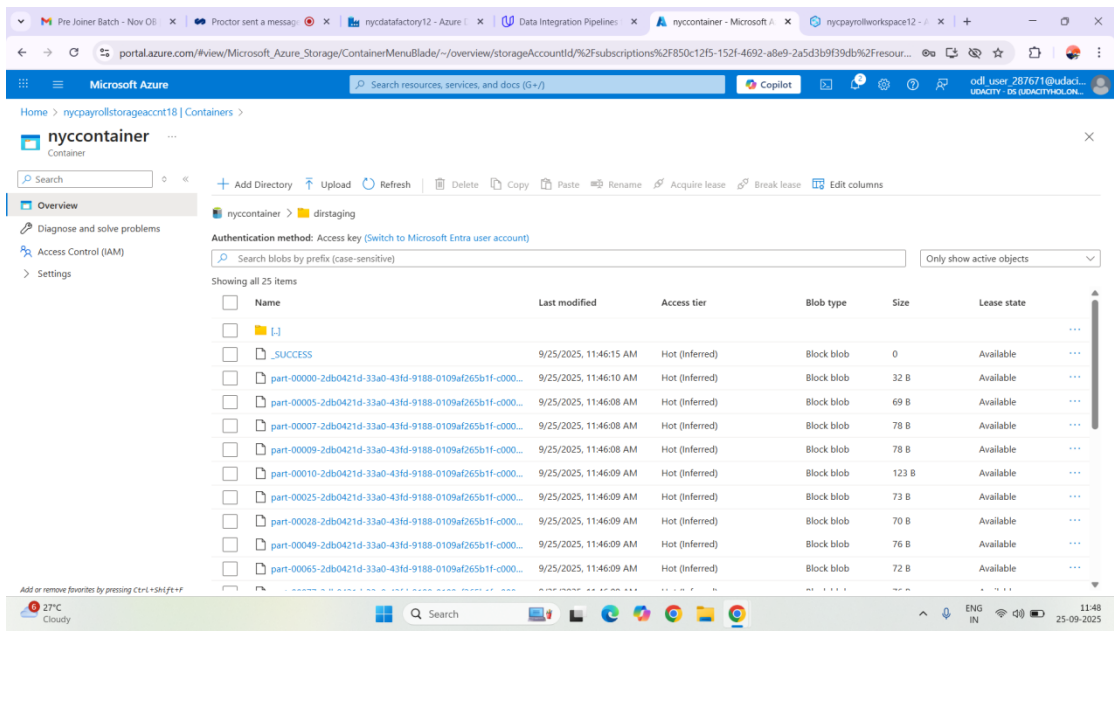
Showing 1 - 6 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Data flow summary	Succeeded	Data flow	9/25/2025, 11:45:32 AM	55s	AutoResolveIntegrationRuntime (East US)		0ceec778-53cd-409...
Data flow_2020	Succeeded	Data flow	9/25/2025, 11:45:09 AM	22s	AutoResolveIntegrationRuntime (East US)		c3abd551-b87f-433...
Data flow_2021	Succeeded	Data flow	9/25/2025, 11:45:09 AM	20s	AutoResolveIntegrationRuntime (East US)		05c5849d-54cb-49c...
Data flow_agency	Succeeded	Data flow	9/25/2025, 11:39:38 AM	5m 29s	AutoResolveIntegrationRuntime (East US)		4705a2fa-3cc4-4f35...
Data flow_title	Succeeded	Data flow	9/25/2025, 11:39:38 AM	5m 14s	AutoResolveIntegrationRuntime (East US)		acea7cc5-5b20-40b...
Data flow_emp	Succeeded	Data flow	9/25/2025, 11:39:38 AM	5m 15s	AutoResolveIntegrationRuntime (East US)		d1727528-7487-41c...

Capture screenshot of query from SQL DB summary table



dirstaging directory listing in Datalake



Synapse Summary External Table:

The screenshot displays the Microsoft Azure Synapse Analytics web interface. The left sidebar shows the workspace structure with 'nyccontainer' selected. The main pane shows a SQL script titled 'SQL script 1' with the following content:

```
16
17 CREATE EXTERNAL TABLE [dbo].[NYC_Payroll_Summary] (
18 [FiscalYear] VARCHAR(50) NULL,
```

The 'Results' tab is active, showing a table with 10 rows and 3 columns: FiscalYear, AgencyName, and TotalPaid. The data is as follows:

FiscalYear	AgencyName	TotalPaid
2021	COMMUNITY COLLEGE (QUEENS...	297484.08
2021	COMMUNITY COLLEGE (MANHA...	275457.88
2021	OFFICE OF THE COMPTROLLER	827663.74
2021	DEPT OF ENVIRONMENT PROTE...	858802.4
2021	COMMUNITY COLLEGE (BRONX)	281046.17000000004

The right sidebar shows the 'Properties' panel for the script, with the name 'SQL script 1' and a description field. The status bar at the bottom indicates '00:00:16 Query executed successfully.'