

# Azure DevOps Coding Challenge

J Jatin DE120

## Problem Statement:

Build an ETL pipeline with azure synapse with dataflow running on it.

## Step 1: Set Up Azure Synapse Workspace

### 1. Create an Azure Synapse Workspace:

- Go to the Azure Portal.
- Create a new Synapse workspace or use an existing one.
- Configure the required storage account and data lake (ADLS Gen2).

### 2. Configure Synapse Studio:

- Access Synapse Studio via your Synapse Workspace.
- Verify that you have appropriate permissions to connect and manipulate data.

The screenshot displays the Microsoft Azure portal interface for a Synapse workspace named 'azsynapsejj'. The top navigation bar includes the 'Microsoft Azure' logo, a search bar, and a 'Copilot' 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, Settings, Analytics pools, Security, Monitoring, Automation, and Help. The main content area is titled 'Essentials' and provides key information about the workspace:

- Resource group:** rg-azuser2367\_mml.local-jeNlUz
- Status:** Succeeded
- Location:** South India
- Subscription:** MML Learners
- Subscription ID:** 2a3c6418-97b9-4d96-a24b-2c2d7633d375
- Managed virtual network:** No
- Managed Identity object ...:** 50ef2898-1c6d-48dc-837c-a7b25c3a7434
- Workspace web URL:** <https://web.azuresynapse.net?workspace=%2fsubscriptions%2f2a...>
- Tags:** [Add tags](#)

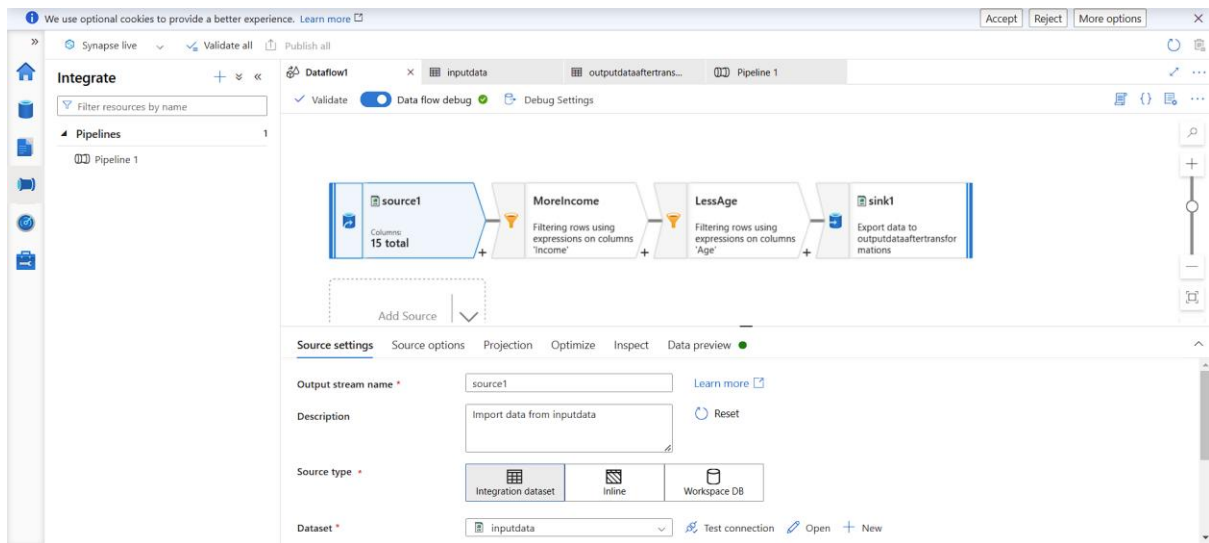
Below the Essentials section, there are two 'Getting started' cards:

- Open Synapse Studio:** Start building your fully-integrated analytics solution and unlock new insights. [Open in](#)
- Read documentation:** Learn how to be productive quickly. Explore concepts, tutorials, and samples. [Learn more in](#)

The right sidebar shows a 'JSON View' button and a list of endpoints:

- Networking:** [Show firewall settings](#)
- Primary ADLS Gen2 acco...:** <https://azdatalakejj.dfs.core.windows.net>
- Primary ADLS Gen2 file s...:** storejj
- SQL admin username:** sqladminuser
- SQL Microsoft Entra admin:** [azuser2367\\_mml.local@techademy.com](mailto:azuser2367_mml.local@techademy.com)
- Dedicated SQL endpoint:** azsynapsejj.sqlazuresynapse.net
- Serverless SQL endpoint:** azsynapsejj-ondemand.sqlazuresynapse.net
- Development endpoint:** <https://azsynapsejj.dev.azuresynapse.net>

## Open Synapse Studio:



## Step 2: Create the Data Flow

### 1. Set Up a Data Flow:

- In Synapse Studio, navigate to **Data > Data flows** and create a new Data Flow.

### 2. Add Data Sources:

- Drag and drop a Source object onto the canvas.
- Configure the source to pull data from your source systems using Linked Services.

### 3. Perform Transformations:

- Use transformation activities such as Filter, Join, Aggregate, Derived Column, etc., to process the data.
- Leverage mapping and transformations to clean, enrich, or aggregate the data.

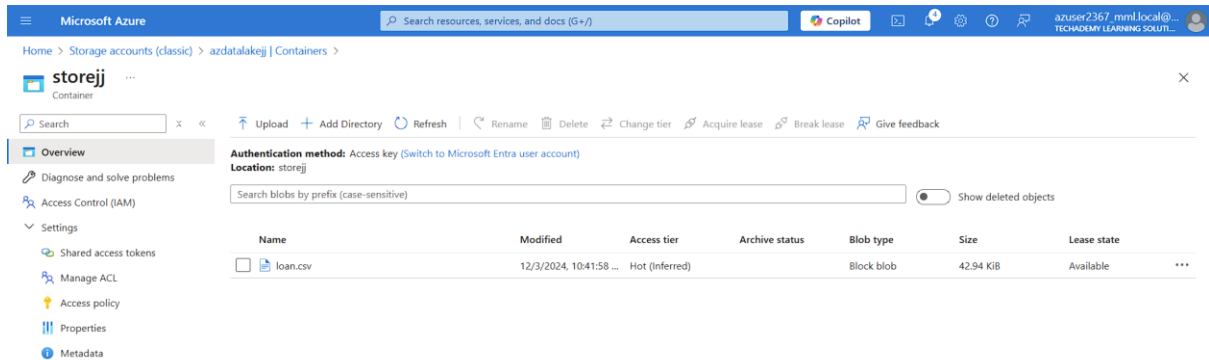
### 4. Configure Sink:

- Drag a Sink object onto the canvas.
- Configure it to write the transformed data to your destination (e.g., Synapse SQL Pool, ADLS).

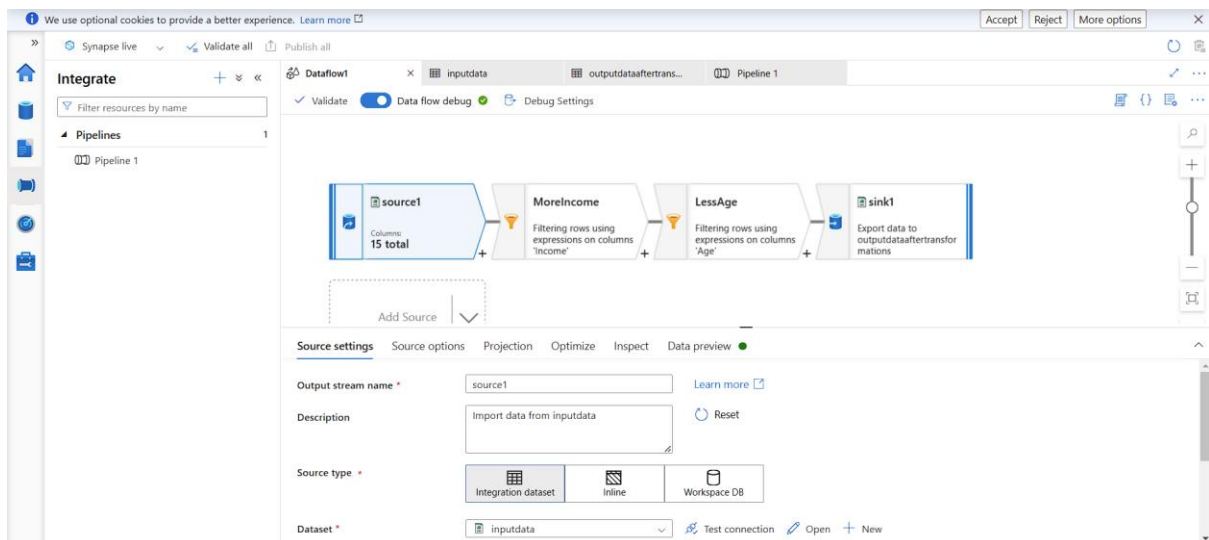
### 5. Validate Data Flow:

- Use the **Debug** feature to run and validate your Data Flow.

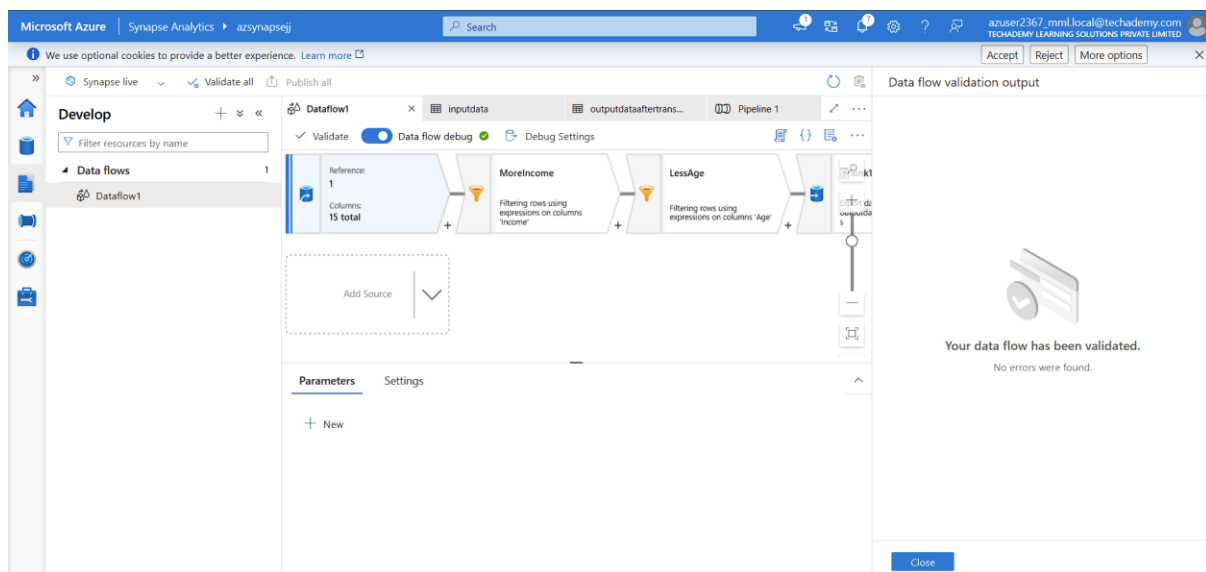
## Input Data Lake Storage Container:



## Creating Dataflow:



## Validating And Debugging It:



### Step 4: Integrate Data Flow into a Pipeline

#### 1. Create a Pipeline:

- Navigate to **Integrate > Pipelines** in Synapse Studio.
- Create a new pipeline to orchestrate the Data Flow.

#### 2. Add a Data Flow Activity:

- Drag the Data Flow activity onto the pipeline canvas.
- Link the Data Flow created in Step 3.

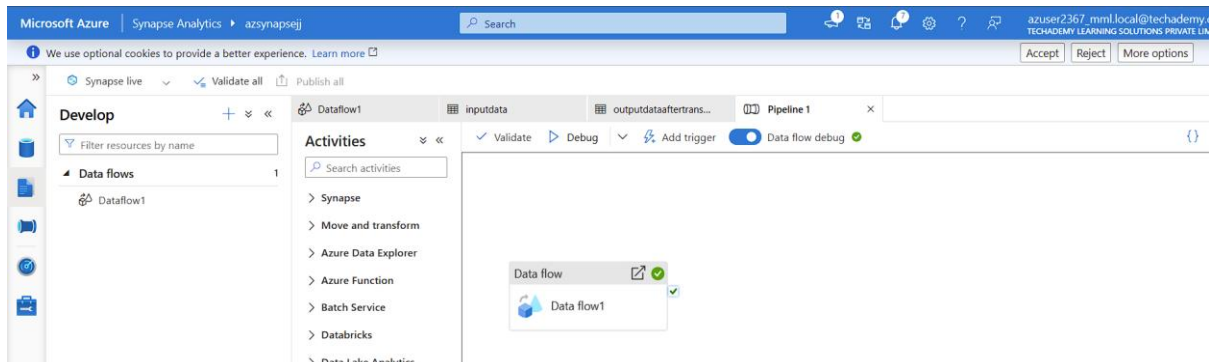
#### 3. Schedule Pipeline Runs:

- Add triggers to schedule the pipeline, such as:
  - Timer triggers for periodic execution.
  - Event-based triggers for real-time execution.

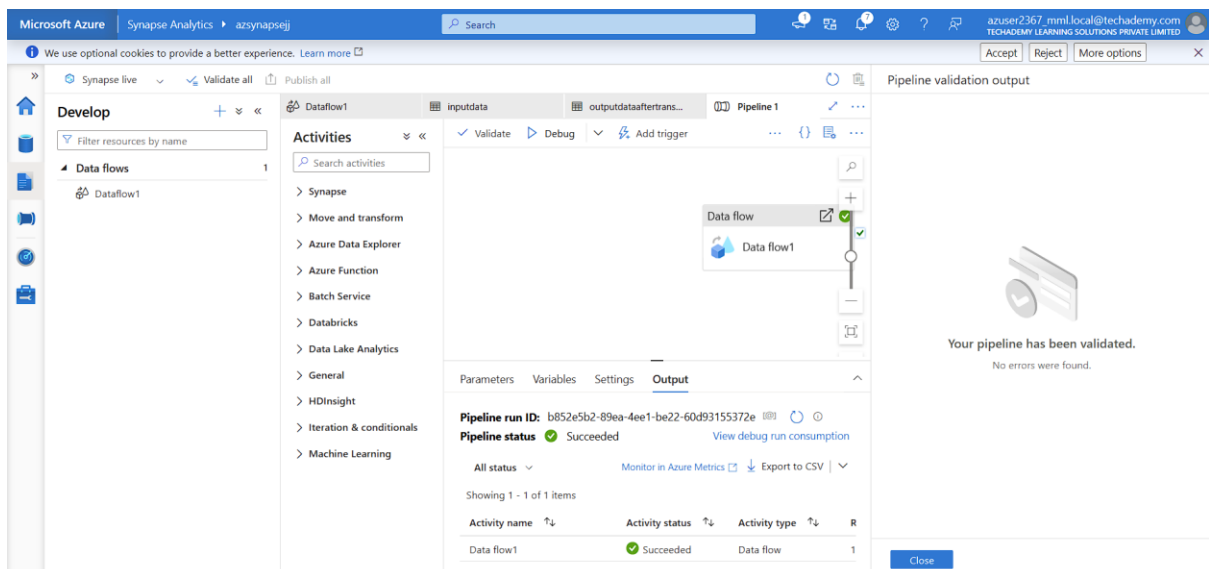
## 4. Test the Pipeline:

- Run the pipeline manually and review logs to ensure it functions as expected.

## Setting Up Pipeline:



## Validating Pipeline:



# Debugging Pipeline:

The screenshot displays the Microsoft Azure Synapse Analytics interface. The left sidebar shows the 'Develop' section with 'Data flows' expanded, listing 'Dataflow1'. The main area shows the 'Pipeline 1' validation output. The pipeline status is 'Succeeded' with a green checkmark. The output section shows 'Your pipeline has been validated. No errors were found.' Below this, there is a table with one row: 'Data flow1' with status 'Succeeded' and activity type 'Data flow'.

Activity name	Activity status	Activity type	R
Data flow1	Succeeded	Data flow	1

# Pipeline Runs:

The screenshot displays the Microsoft Azure Synapse Analytics interface, specifically the 'Pipeline runs' section. The left sidebar shows the 'Integration' section with 'Pipeline runs' expanded. The main area shows the 'Pipeline runs' table. The table has columns: Pipeline name, Run start, Run end, Duration, Status, Triggered by, Run ID, and Parameters. There is one row: 'Pipeline 1' with status 'Succeeded'.

Pipeline name	Run start	Run end	Duration	Status	Triggered by	Run ID	Parameters
Pipeline 1	12/19/2024, 5:18:33 PM	12/19/2024, 5:19:35 PM	1m 3s	Succeeded	Manual trigger	b852e5b2-89ea-4ee1-b...	

## Output Storage Account (Verifications if file is saved here):

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

azuser2367.mnt.local@...  
TECHADAM LEARNING SOLUTIONS

Home > Storage accounts (classic) > azdatalakejj | Containers >

storejj

Containers

Search

UploadAdd DirectoryRefreshRenameDeleteChange tierAcquire leaseBreak leaseGive feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: storejj

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
loan.csv	12/3/2024, 10:41:58 ...	Hot (Inferred)		Block blob	42.94 KiB	Available

## Conclusion:

Building an ETL pipeline with Azure Synapse Analytics provides a scalable and efficient solution for handling large-scale data transformations and integrations. By leveraging Synapse Data Flows and Pipelines, businesses can automate the data extraction, transformation, and loading processes while maintaining flexibility in handling diverse data sources. Continuous monitoring and optimization using Synapse's integrated tools ensure reliability, performance, and cost-effectiveness. This approach empowers organizations to derive actionable insights from their data, enhancing decision-making and driving business success.