Azure DevOps Coding Challenge

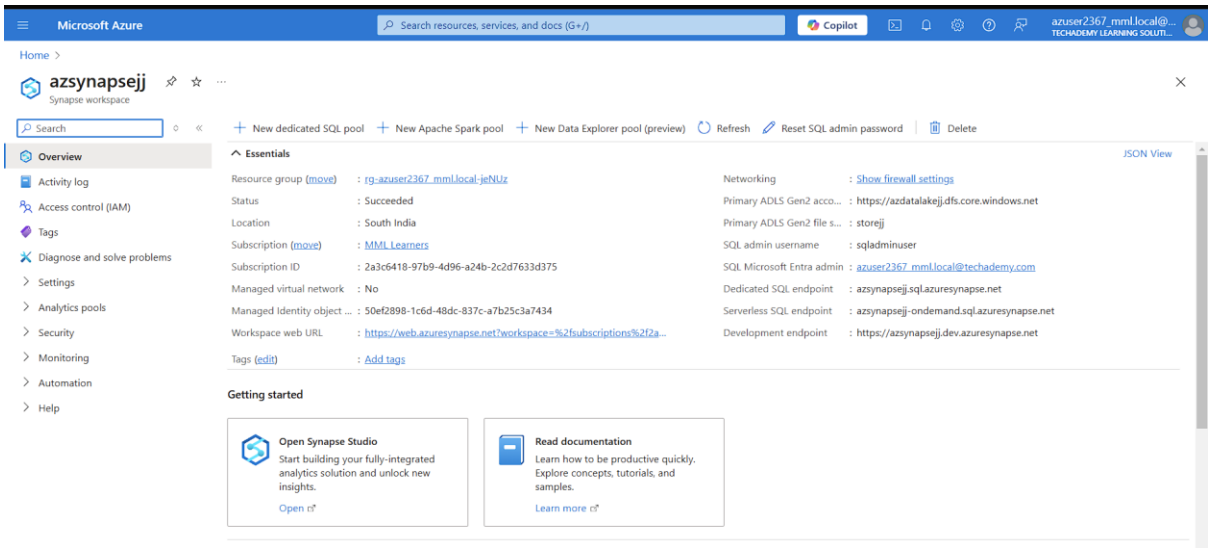
J Jatin DE120

Problem Statement:

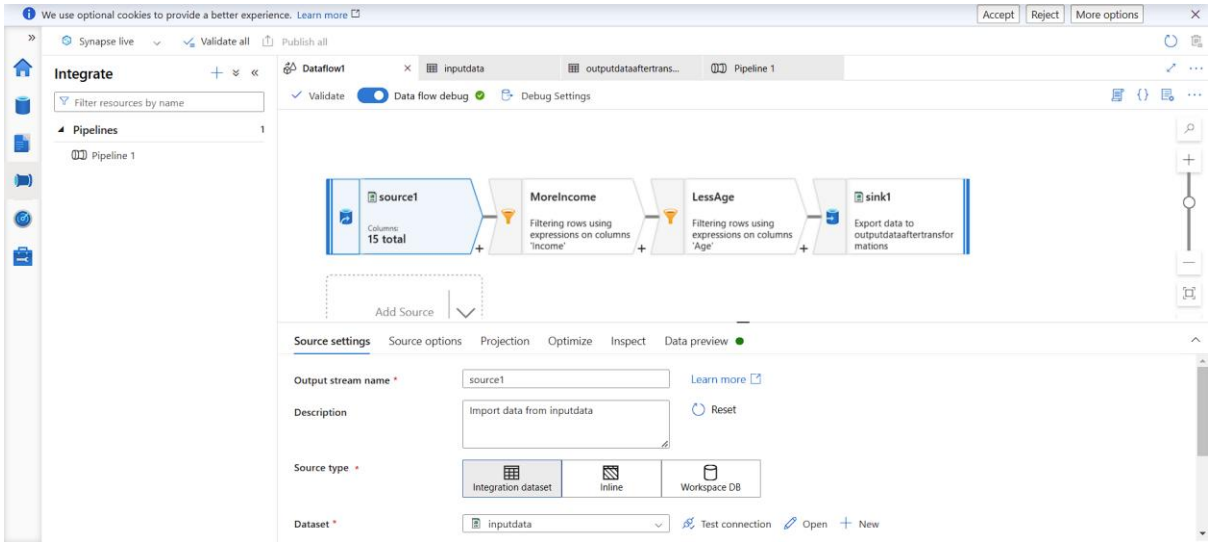
Build an ETL pipline with azure synapse with dataflow running on it.

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.



Open Synapse Studio:

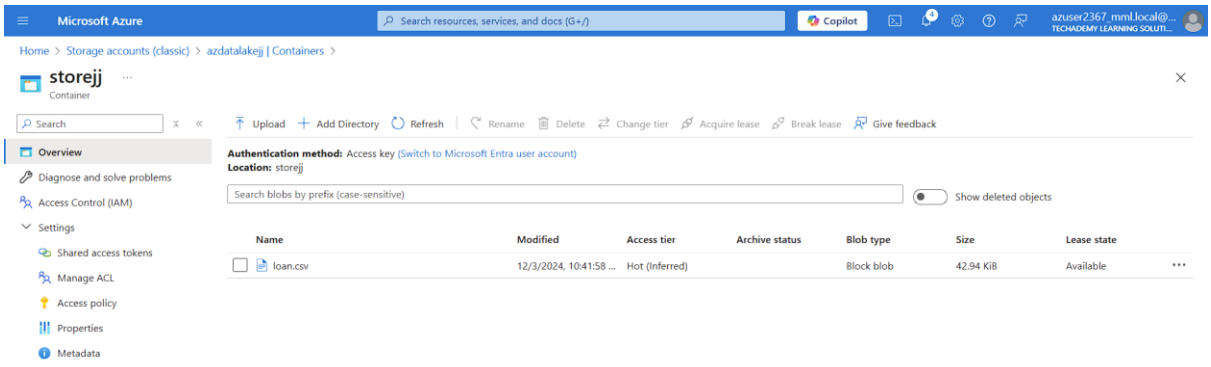


Step 2: Create the Data Flow

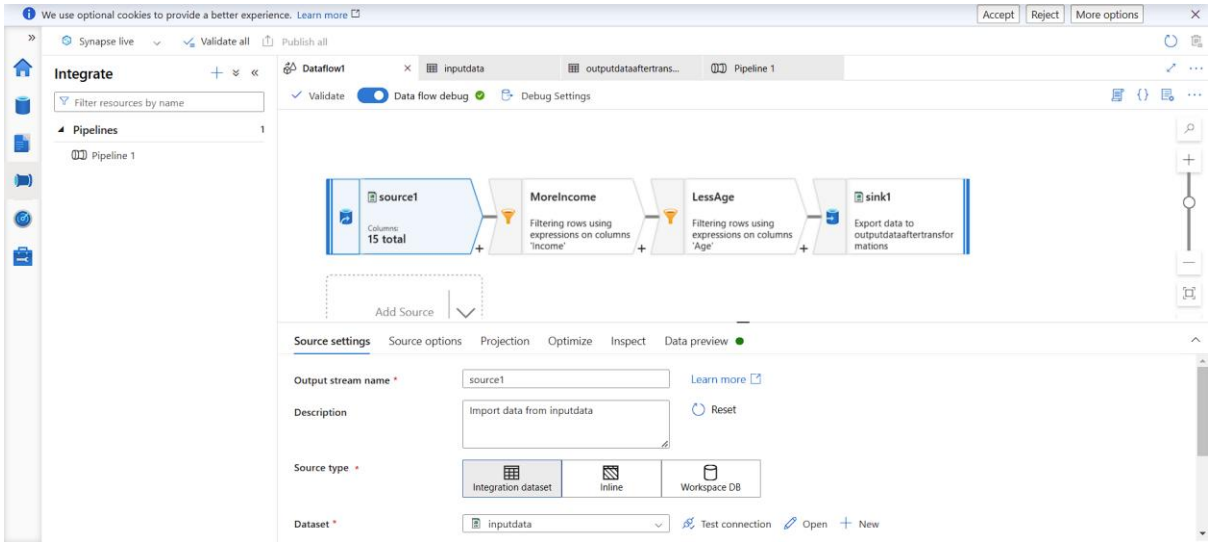
1. **Set Up a Data Flow**:
    - In Synapse Studio, navigate to **Data > Data flows** and create a new Data Flow.
2. **Add Data Sources**:
    - Drag and drop a Source object onto the canvas.
    - Configure the source to pull data from your source systems using Linked Services.
3. **Perform Transformations**:
    - 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.

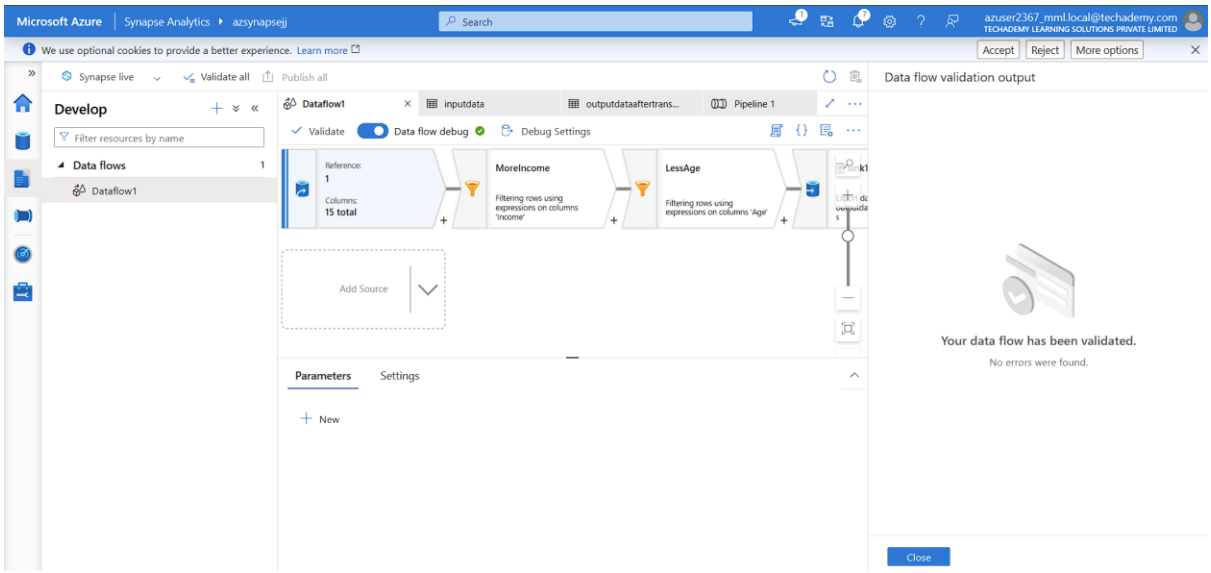
Input Data Lake Storage Container:



Creating Dataflow:



Validating And Debugging It:

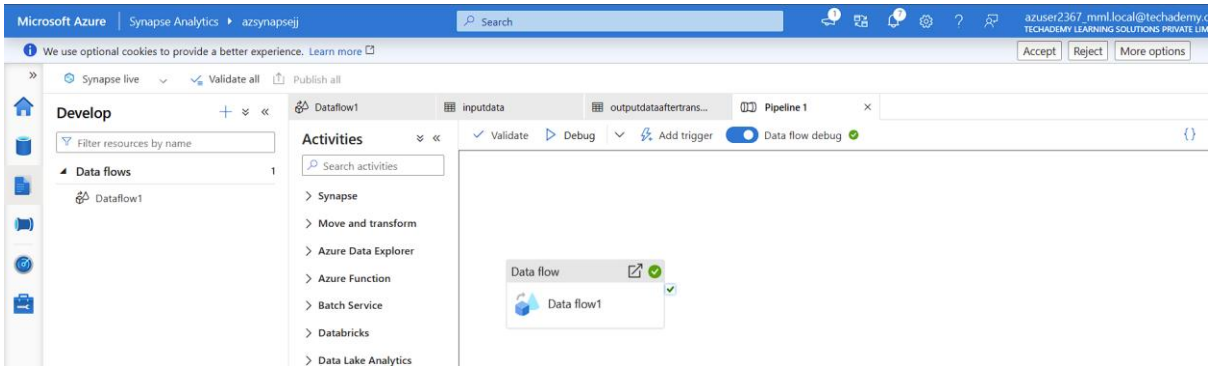


Step 4: Integrate Data Flow into a Pipeline

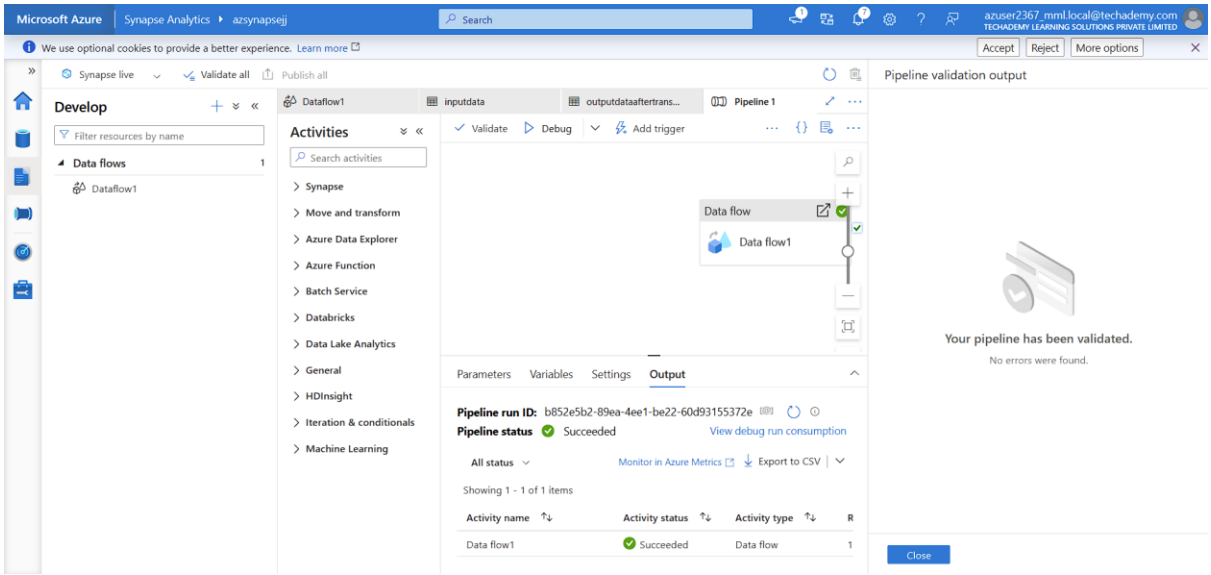
1. **Create a Pipeline**:
    - Navigate to **Integrate > Pipelines** in Synapse Studio.
    - Create a new pipeline to orchestrate the Data Flow.
2. **Add a Data Flow Activity**:
    - Drag the Data Flow activity onto the pipeline canvas.
    - Link the Data Flow created in Step 3.
3. **Schedule Pipeline 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.

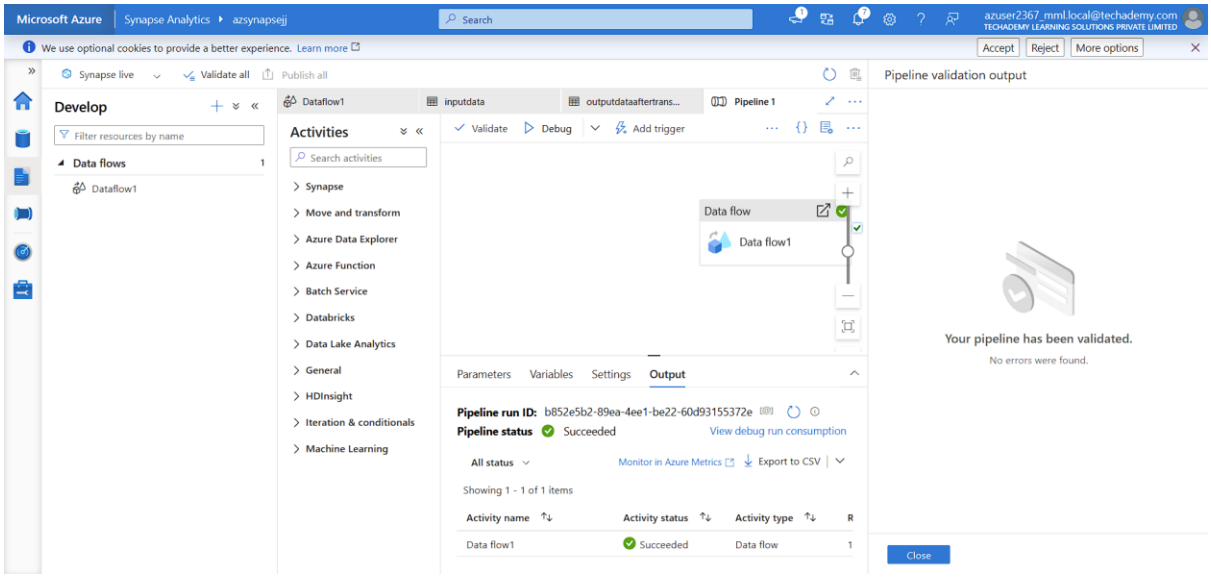
Setting Up Pipeline:



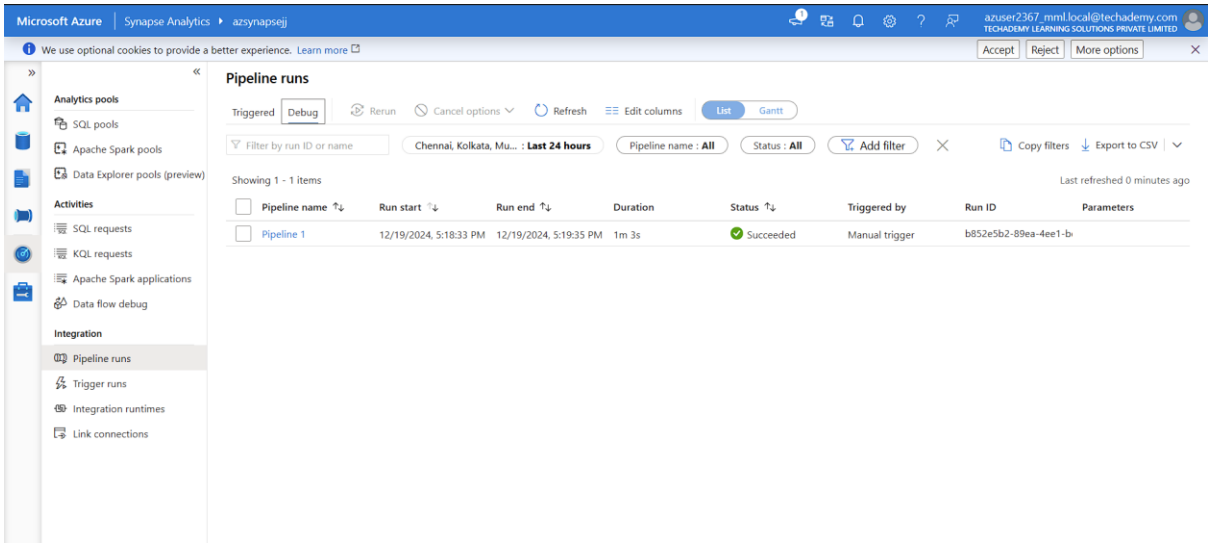
Validating Pipeline:



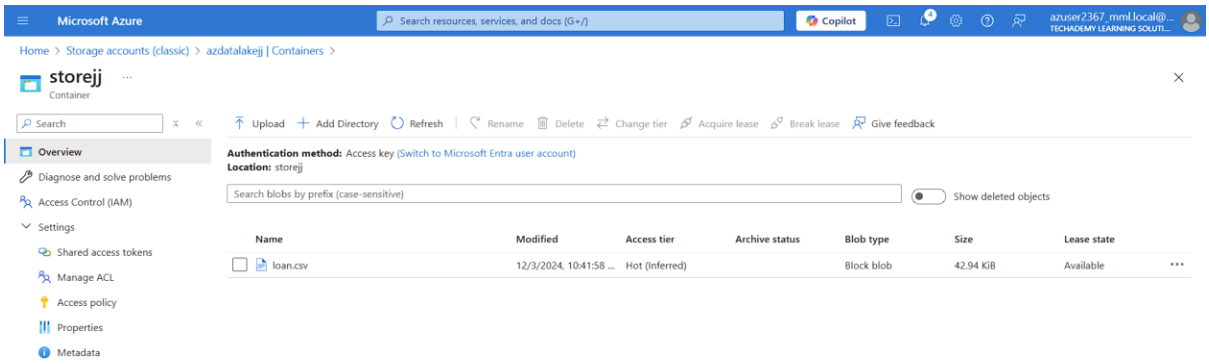
Debugging Pipeline:



Pipeline Runs:



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



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.