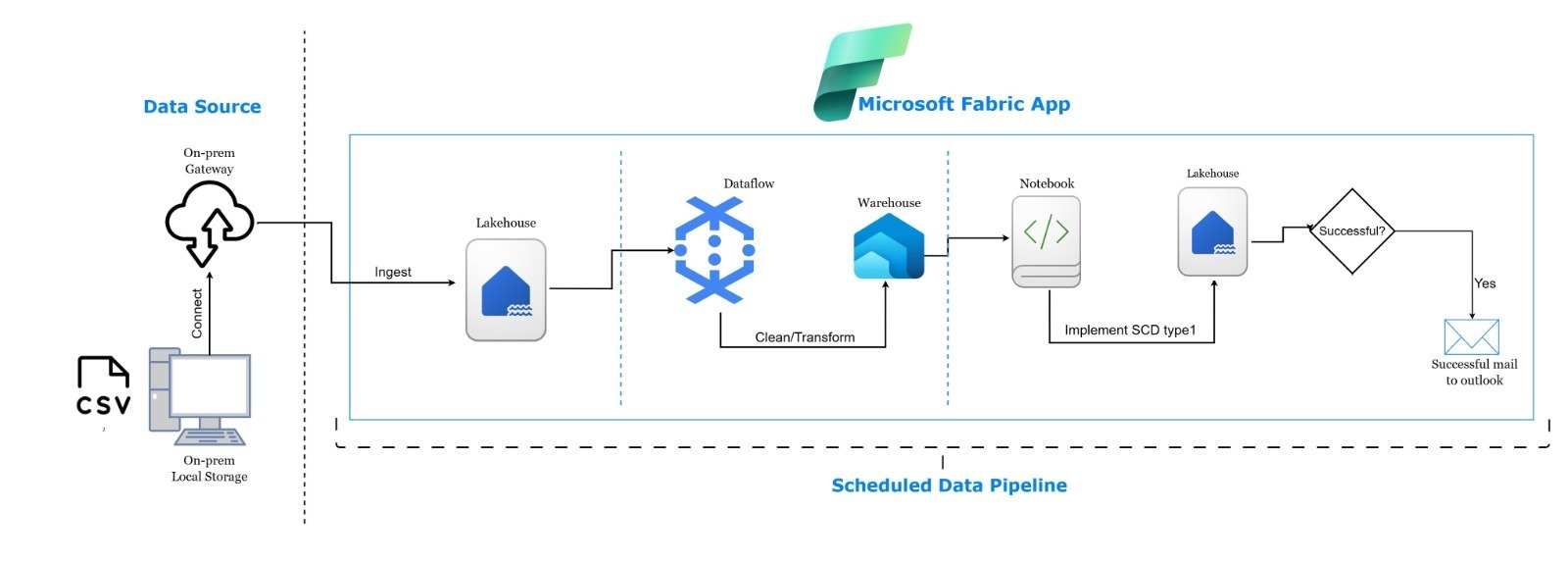
**Incremental Data Loading and Automated Notifications using Microsoft Fabric**

# **Project Overview:**

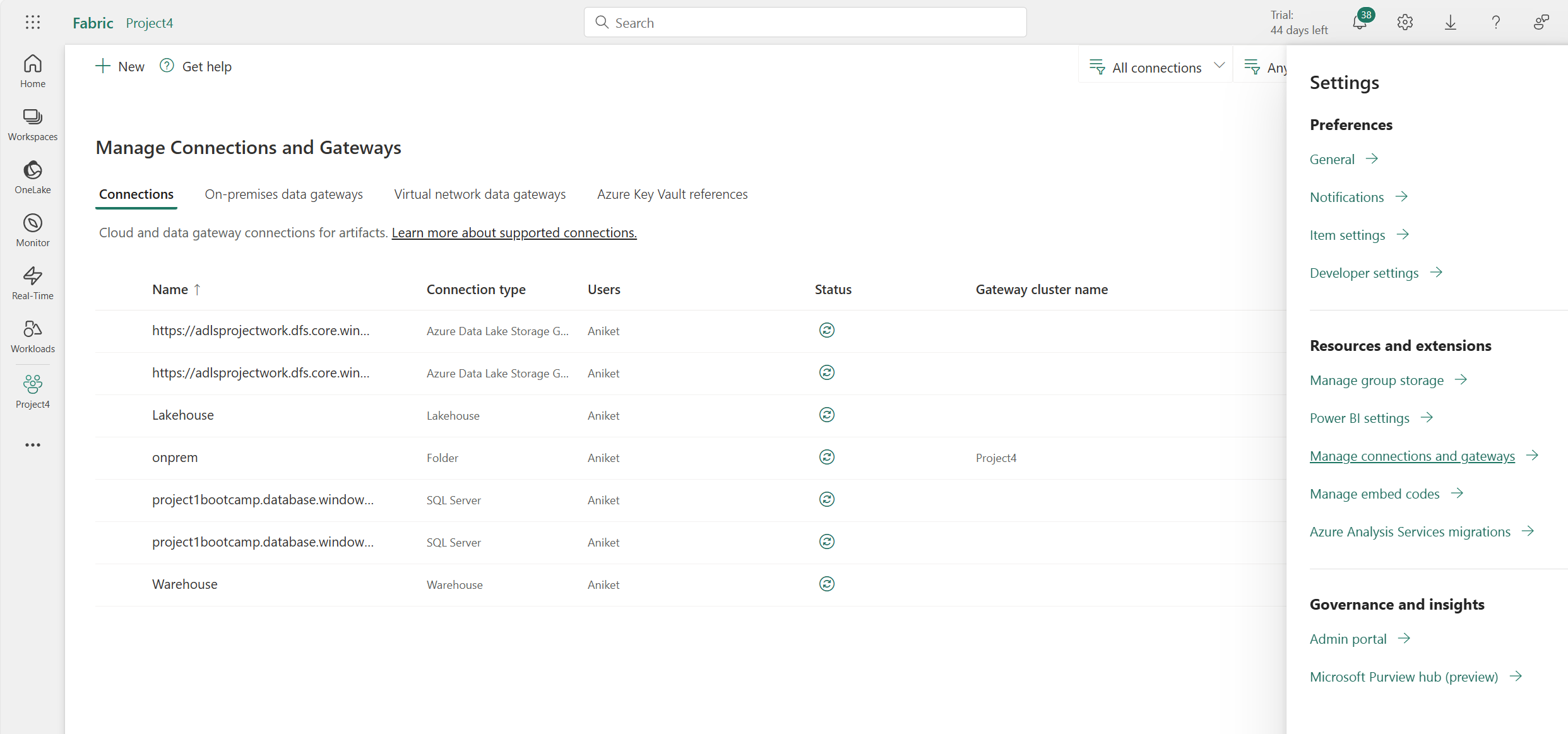


1. **Set up On-Premises Data Gateway** to access local folders.
2. **Copy data** from on-prem to **Lakehouse** using wildcard filtering and last modified time.
3. **Clean data** using **Dataflow Gen1**, then sink it to a **Warehouse**.
4. **Create SCD1 Delta tables** in Lakehouse for all entities.
5. **Develop reusable notebook functions** to read, hash, and merge data into Lakehouse tables using DeltaTable.merge().
6. **Trigger notebooks** from a pipeline and link it with previous data activities.
7. **Send email notifications** using Outlook connector on success.

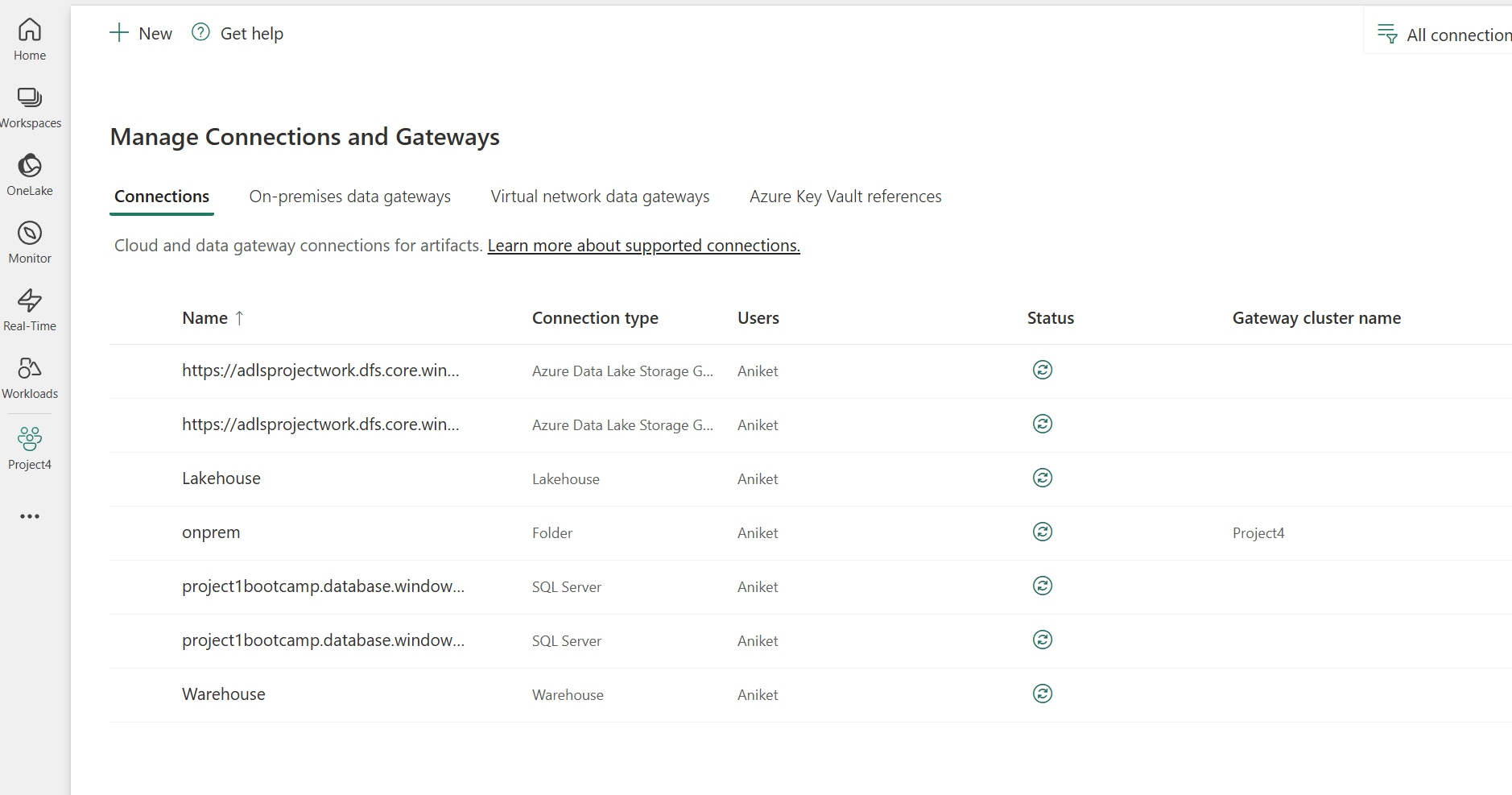
# **Steps:**

## **Create On-premise Data Gateway**

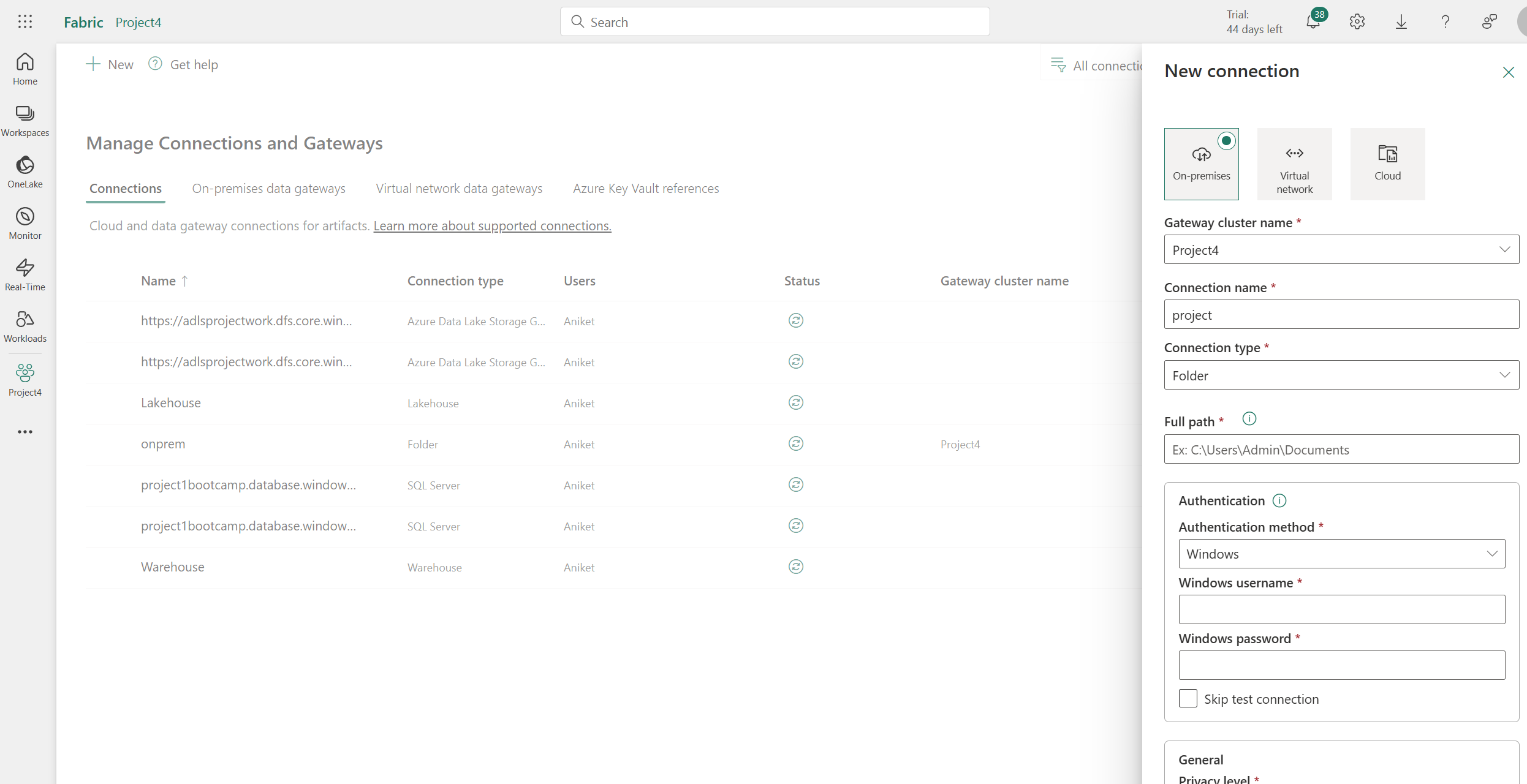
1. Install on-premise data gateway on your on-prem system.
2. Login using your fabric account.
3. Open fabric->Settings->Manage connections and gateways



1. Click on new



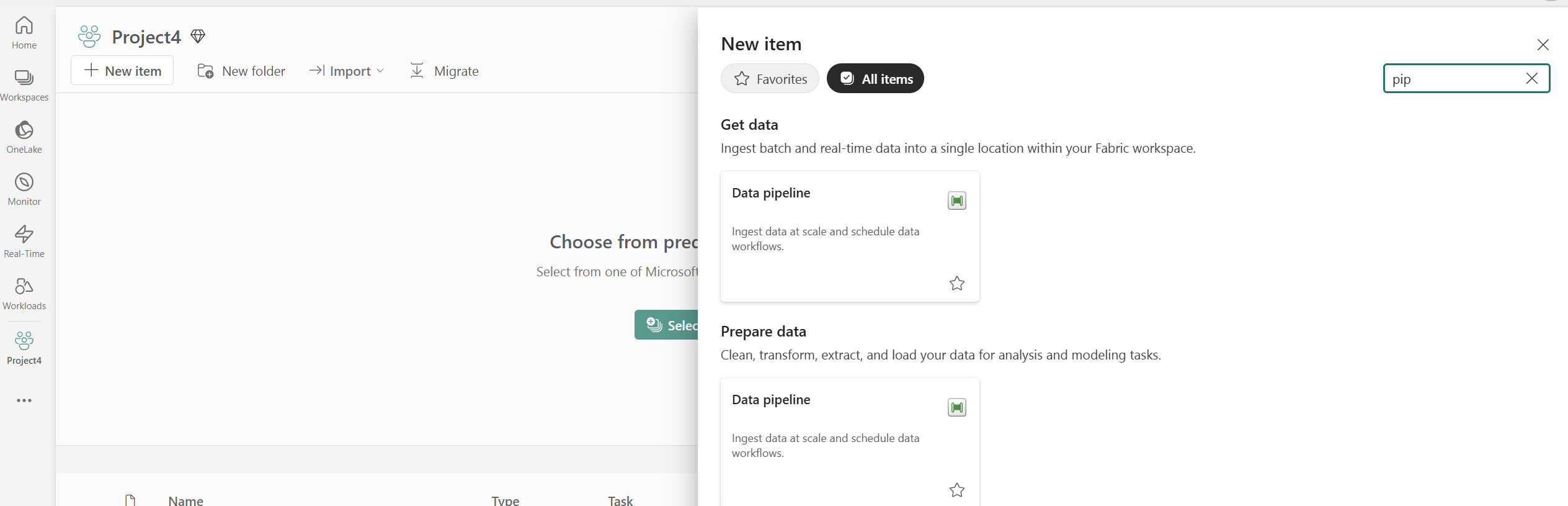
1. Select the Gateway cluster name from dropdown. Give connection name and choose folder in connection type.



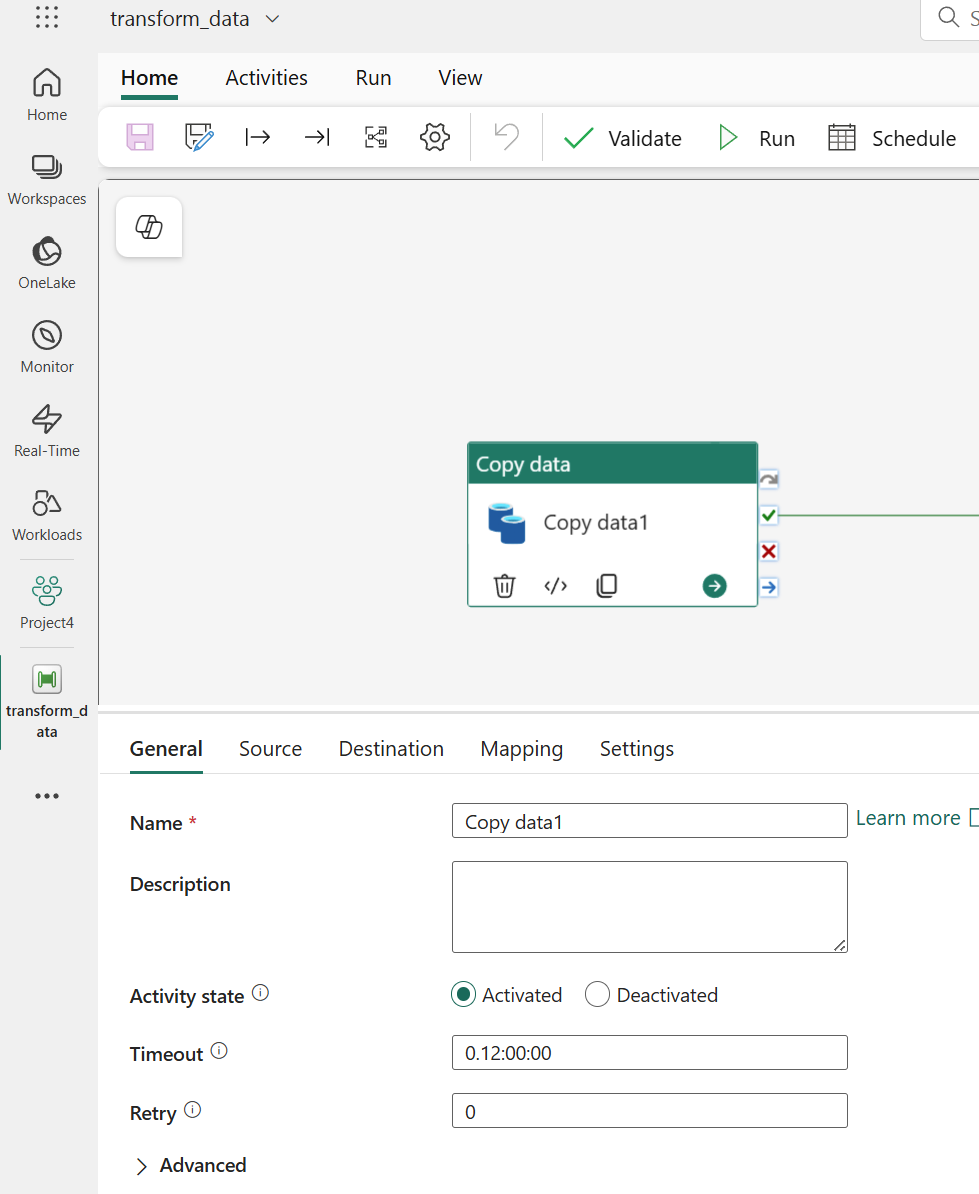
1. Provide path to the folder you want to share->Enter windows username and password and click on create.

## **Copy Data from On-Prem to Lakehouse files**

1. Create a lakehouse that will hold the files from on-prem
2. Create a warehouse that will act as source table for SCD1 transformations.
3. Create a new workspace->Data Pipeline



1. Once in data pipeline->click on Copy Data activity.



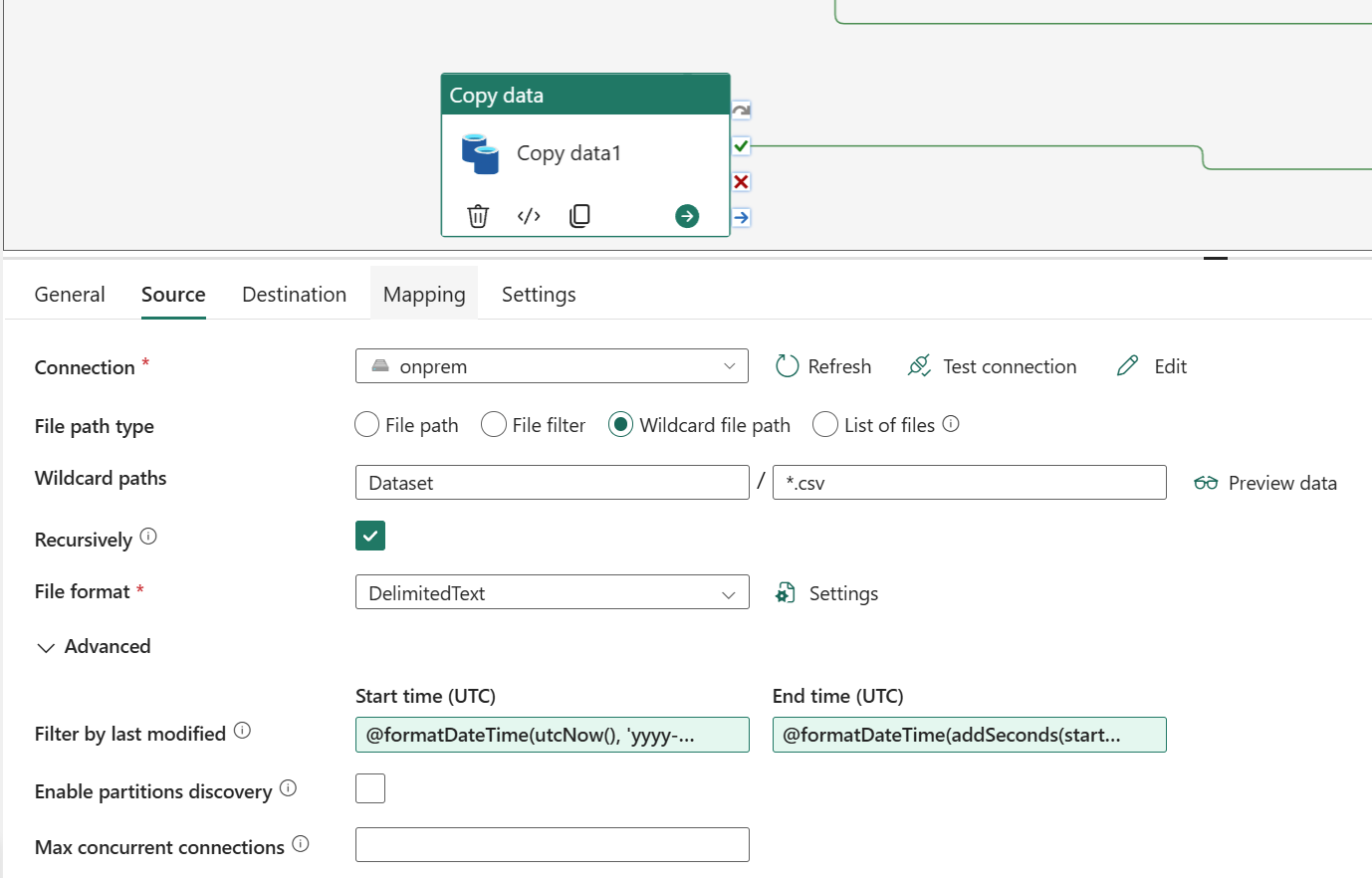
1. Go to source->Select Onprem from dropdown
2. Select wildpath file path and select \*.csv.
3. Click on the advanced tab

For start Time use below expression->

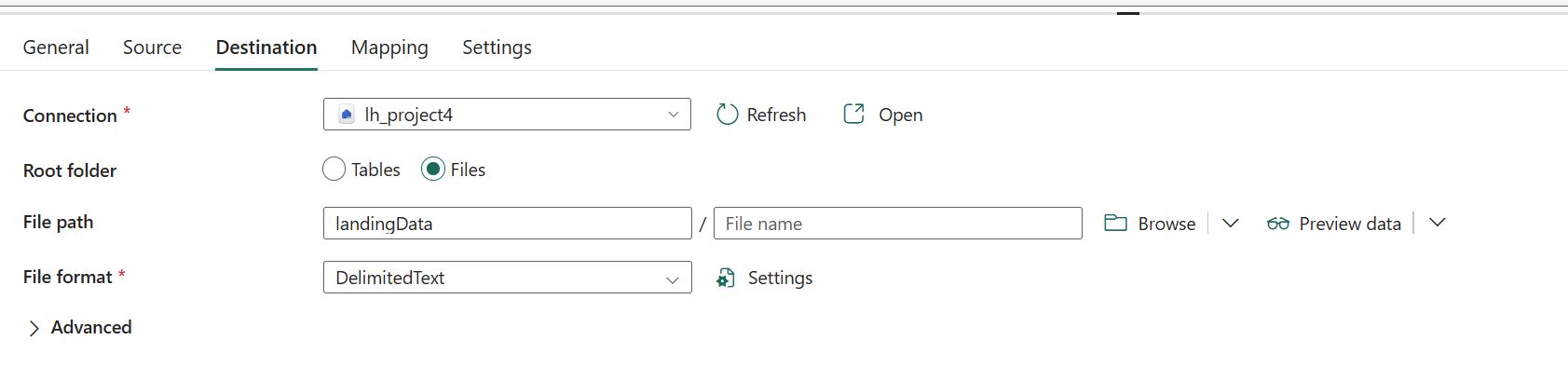
@formatDateTime(utcNow(), 'yyyy-MM-ddT00:00:00Z')

For end Time use below expression->

@formatDateTime(addSeconds(startOfDay(addDays(utcNow(), 1)), -1), 'yyyy-MM-ddTHH:mm:ssZ')

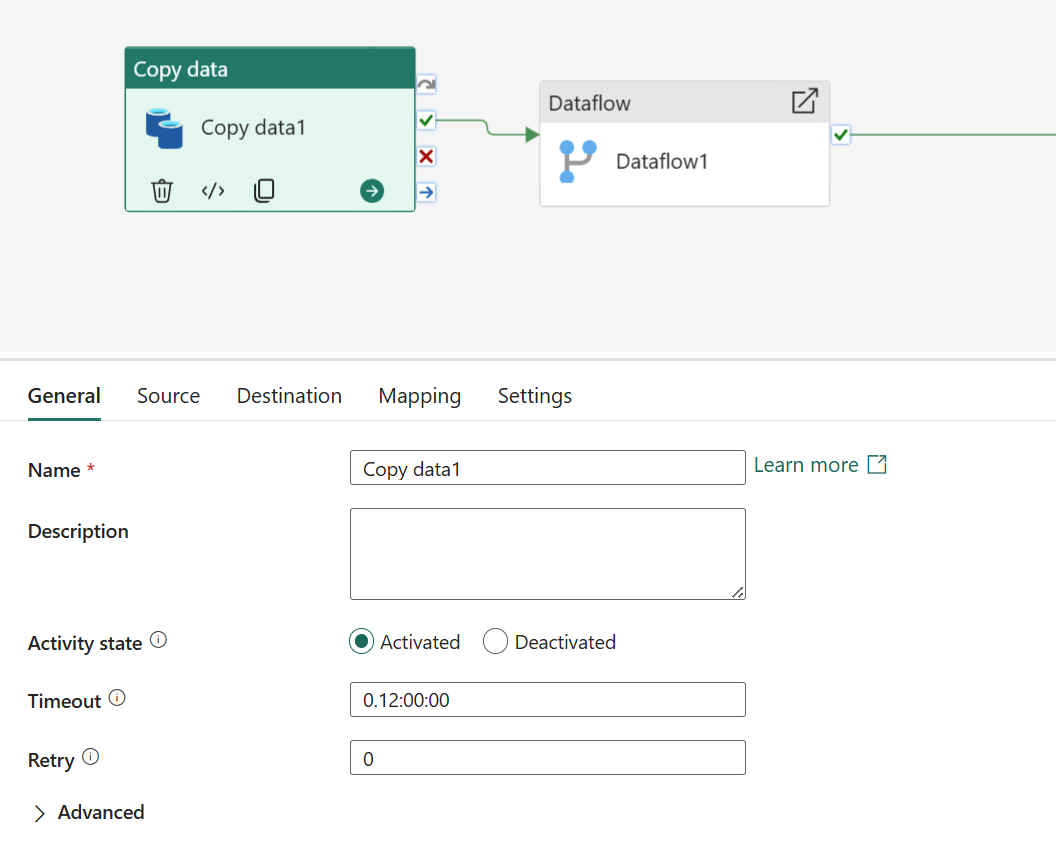


1. In destination setting->In connection select lakehouse.
2. Click on file option-> Browse the file path where we want to save it.

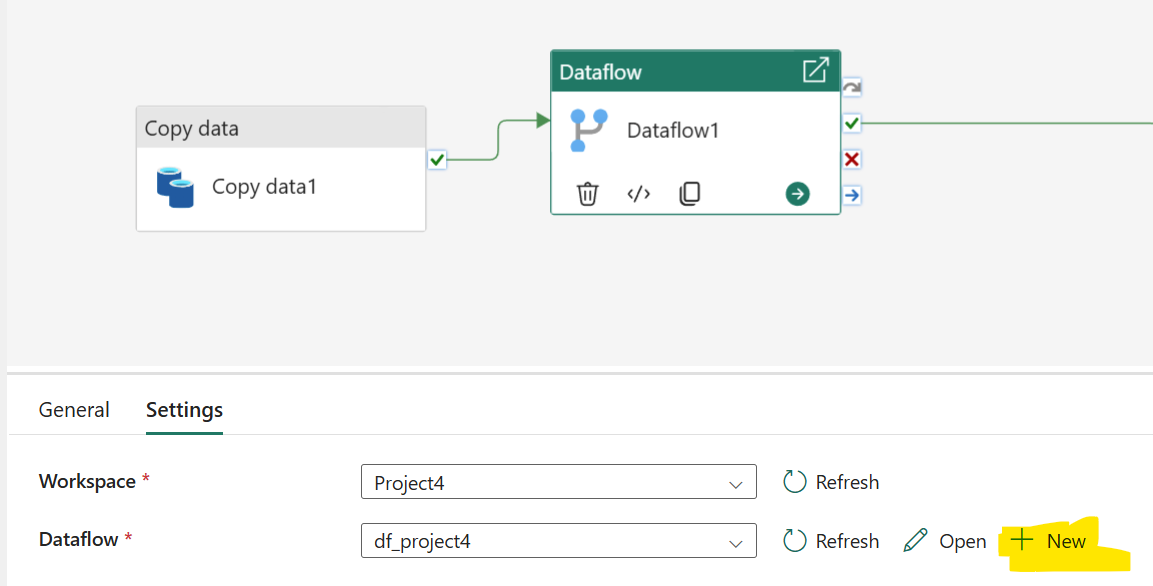


## **Clean Data and save to Warehouse**

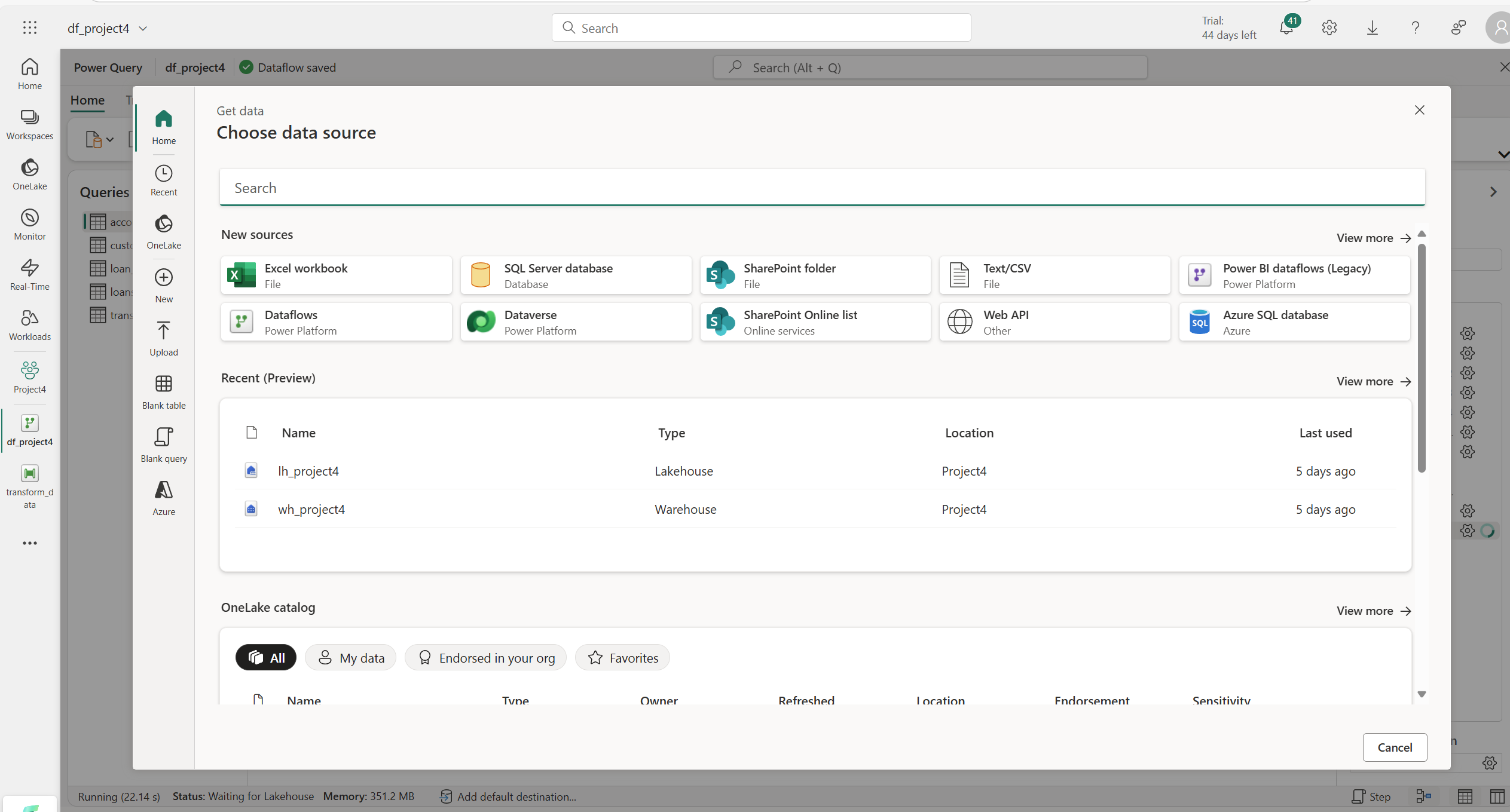
1. Drag a dataflow activity/dataflow gen 1 and connect on success with copy activity.



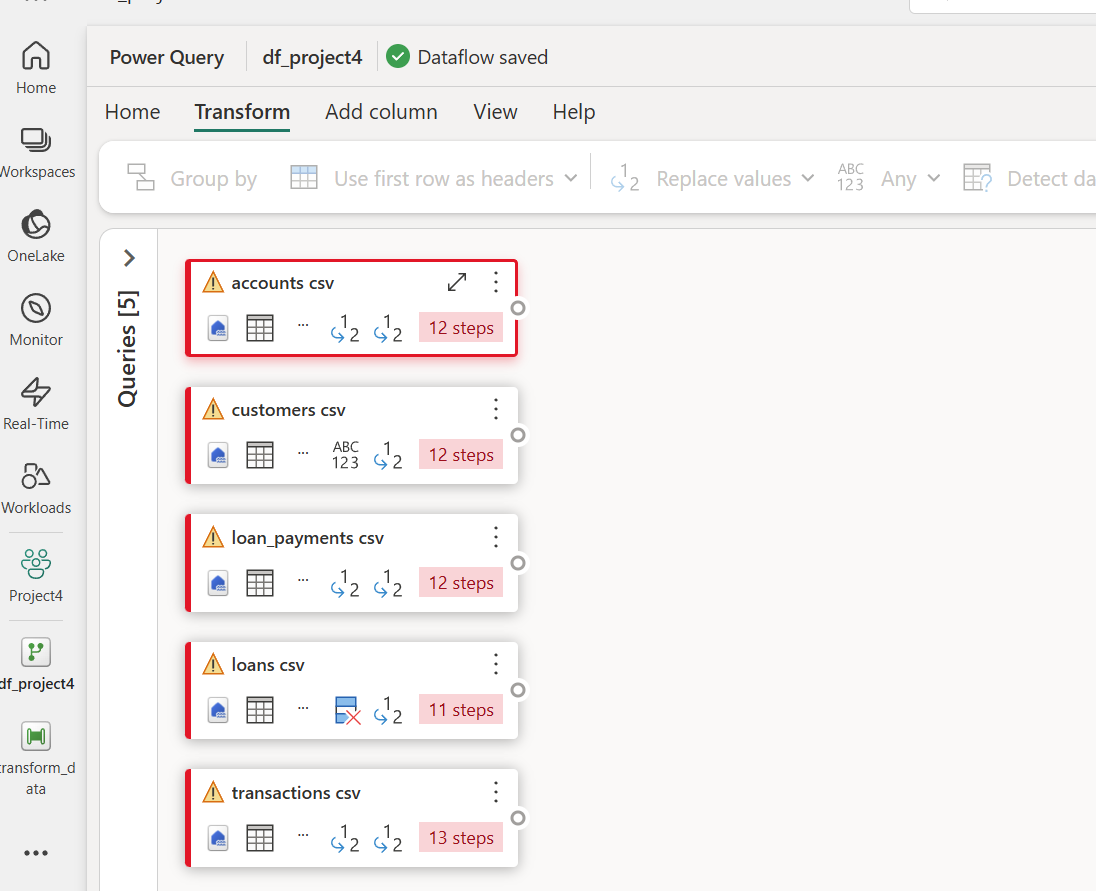
1. Go to settings-> Click on +New in Dataflow



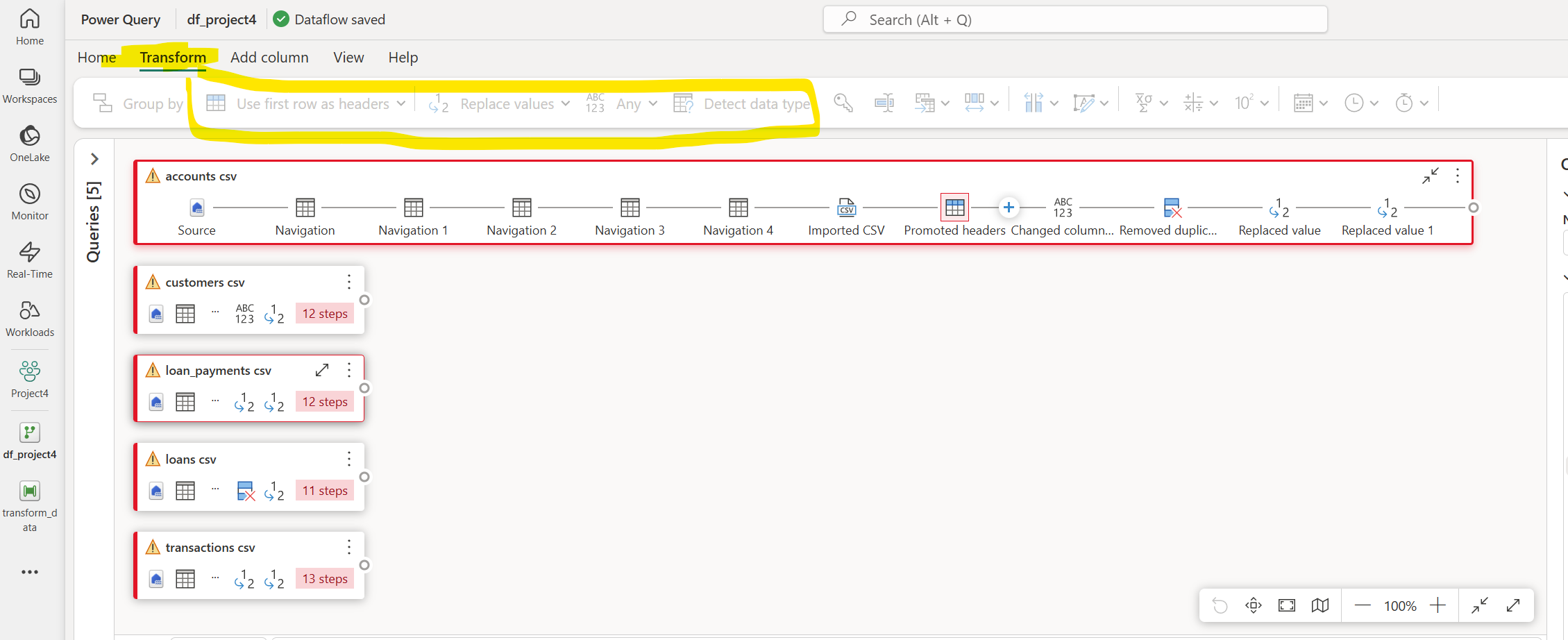
1. Once in Power Query->Click on Get source->Select Lakehouse



1. Click on files and from dropdown select all the files.
2. It will load all the tables and we can perform transformations on top of it.



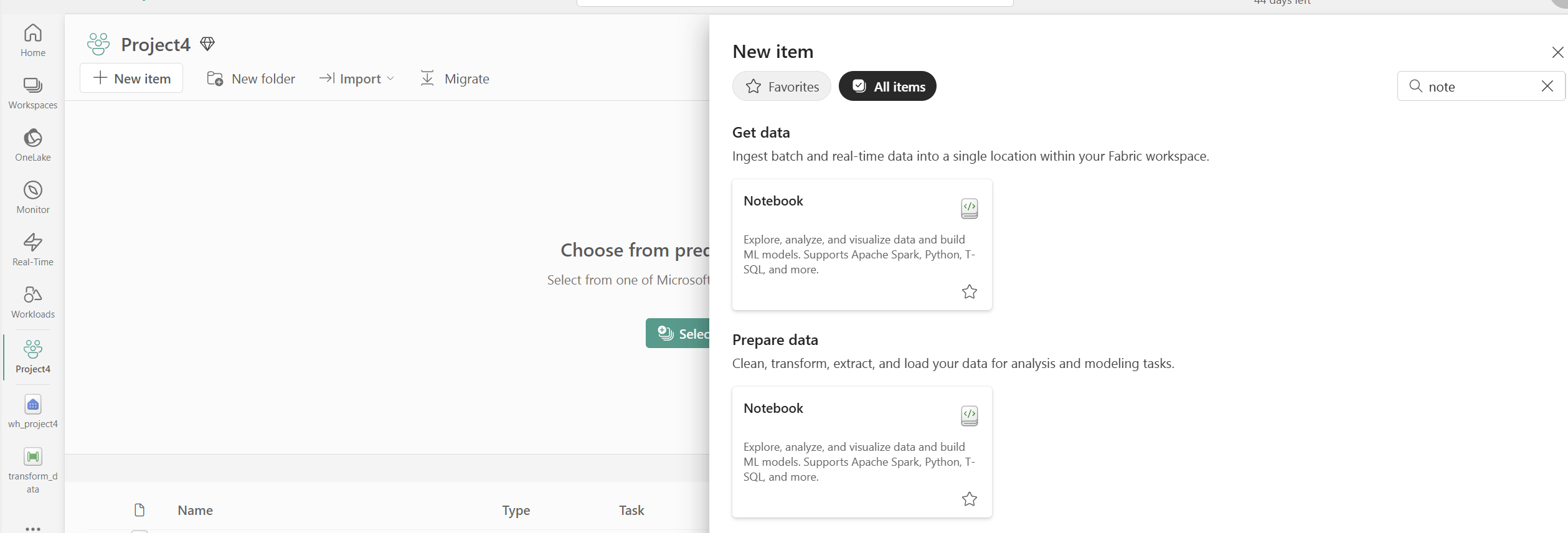
1. Go to transform tab-> Click on Detect Data type, Select columns and right click ->remove duplicate, Click on column->then replace value and replace null with default values.



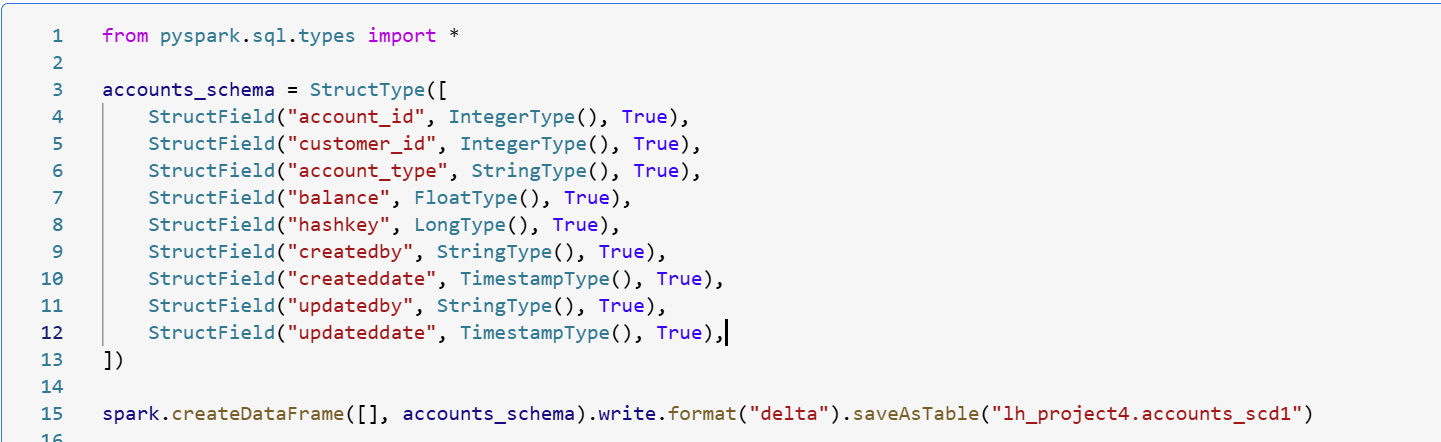
1. In destination->Select Warehouse->Give tablename
2. Repeat for other files as well and then click on publish later.

## **Create a new item Notebook from Workspace**

1. Create a notebook in which we will write SCD1 logic.



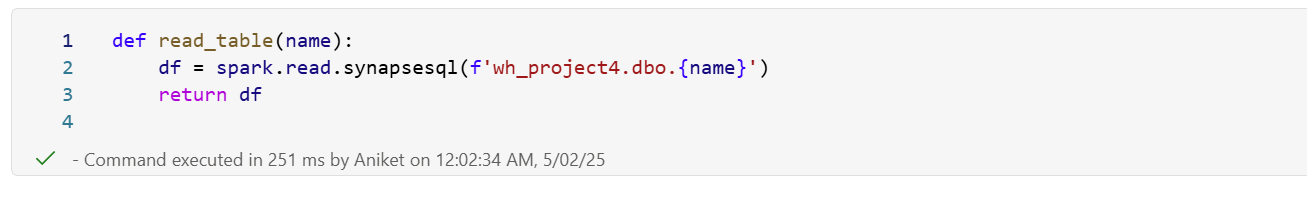
1. Name the notebook SCD1 functions.
2. We will create schema for SCD1 tables for Lakehouse.
3. Import Types module and create schema for each filetype.
4. Use the schema to create a dataframe and then write it back as table.



1. Repeat for the other 4 tables.
2. Import functions like col, crc32, concat from sql function.
3. Also, from Microsoft.spark.fabric import COnstanst that will help us to read data from warehouse.



1. Create a dynamic function to read tables into dataframe.



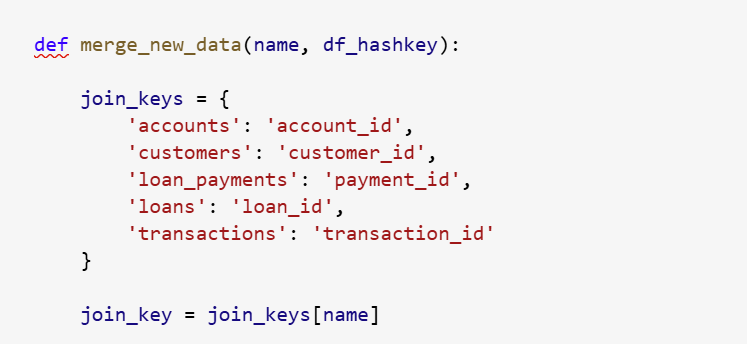
1. Once loaded into dataframe, we will create hashkey.



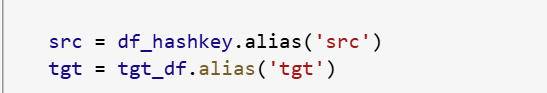
1. Now, import DeltaTable library from delta.tables to read Lakehouse tables into dataframe.



1. Now, for the merge logic we will create function that accepts name and hashkey dataframe.
2. We will create a dictionary and map the table name with id’s.



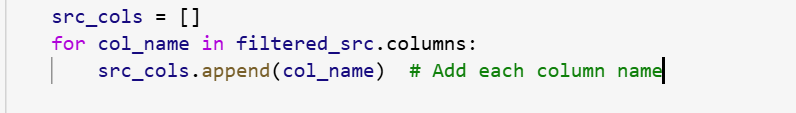
1. Now, we will simply create aliases for src and tgt dataframe.



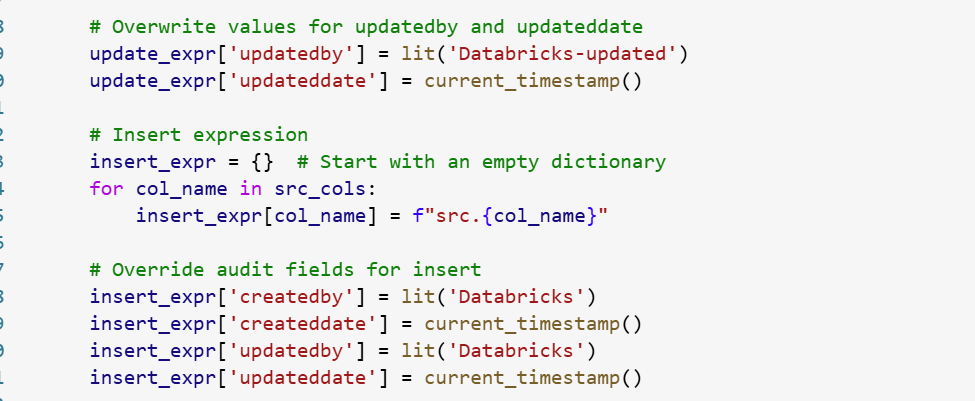
1. Now, we will apply left join on src and tgt on join keys, and filter if tgt join key is null or hashkey from src and tgt do not match then pick all records.



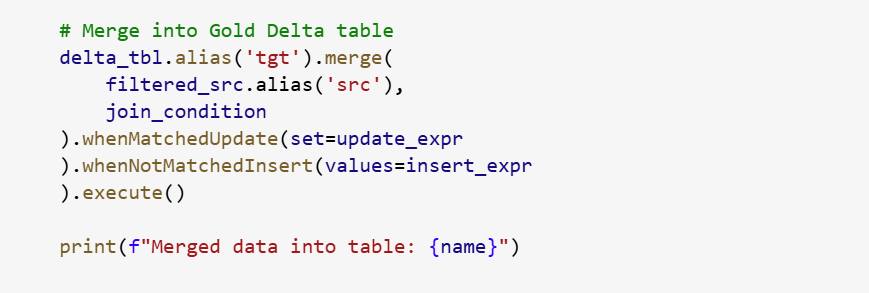
1. Now, we will append every column from filtered dataframe into a empty list.



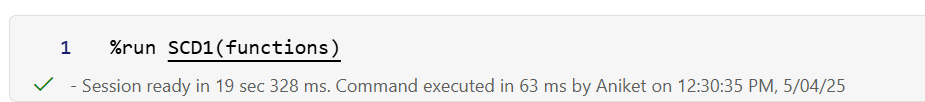
1. We will create two dictionaries update and insert and map values with src values.



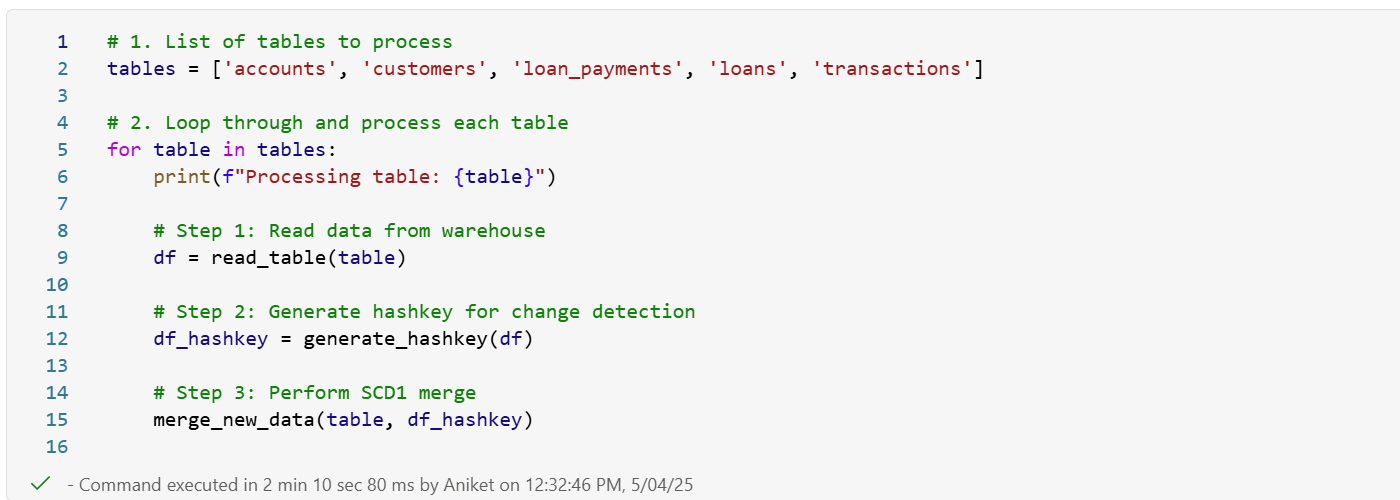
1. We will finally apply merge on delta object.



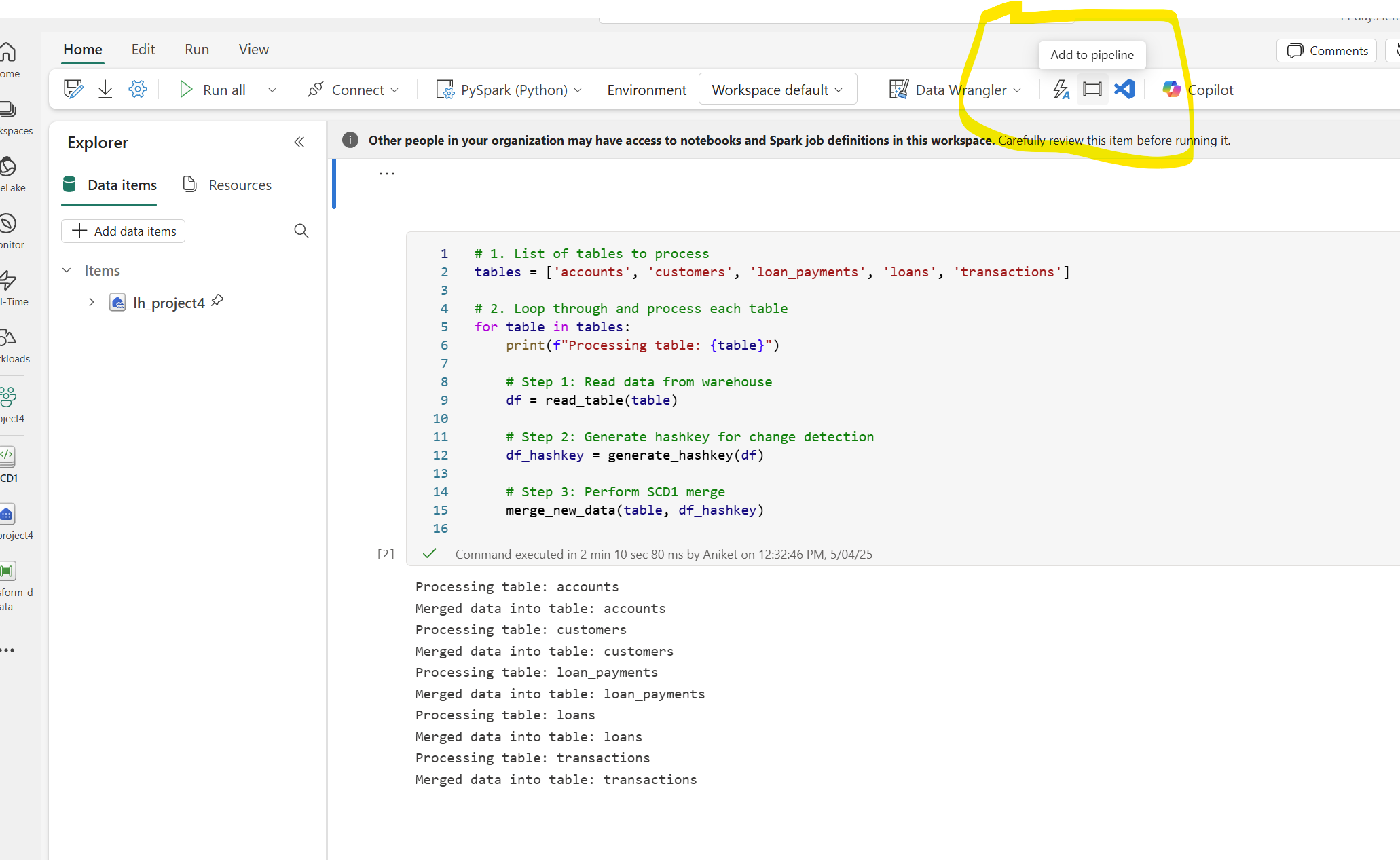
1. Create a new notebook and using magic command %run to import this notebook.

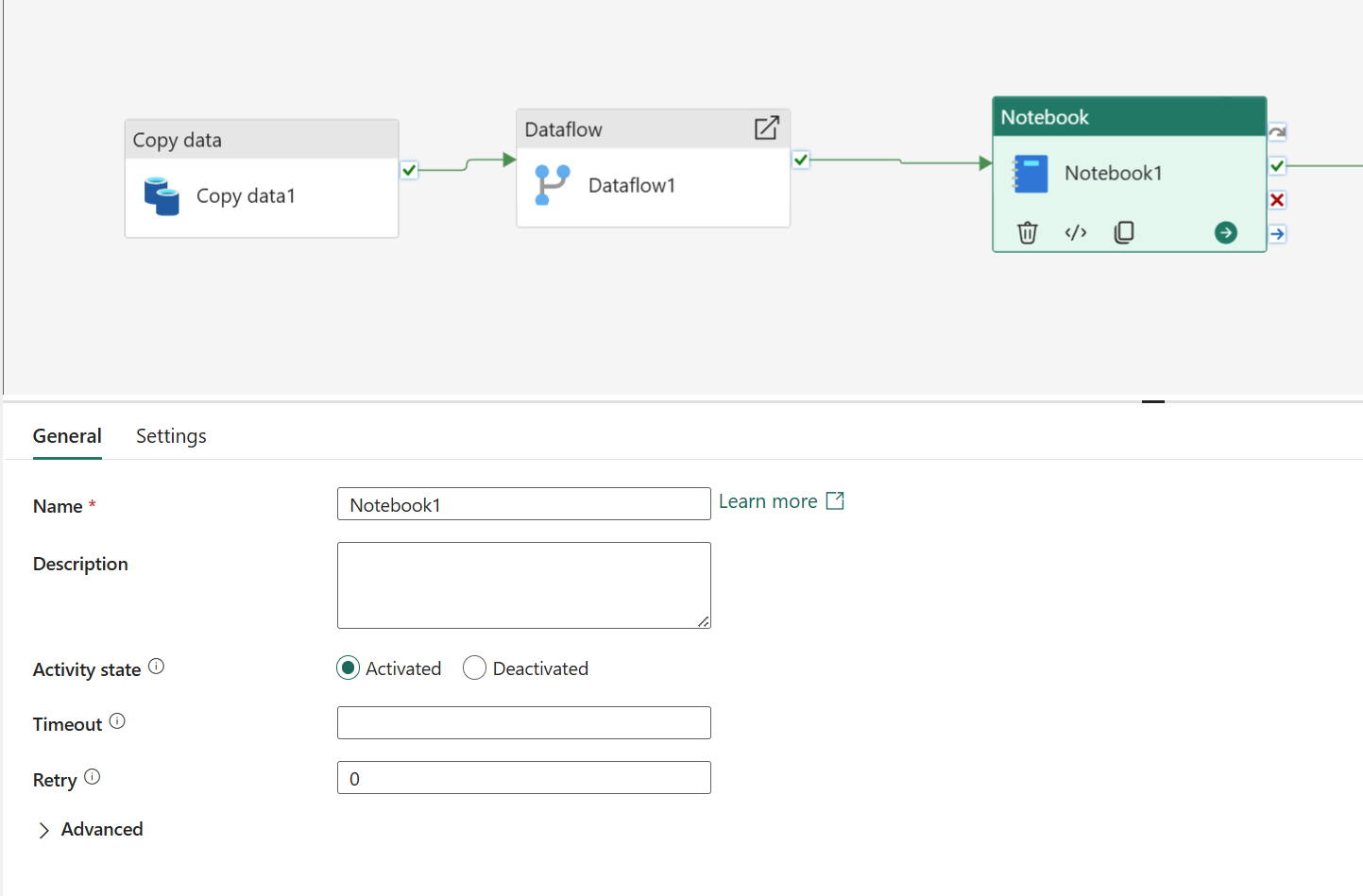


1. Call the predefined functions from the main notebook and pass table names for SCD1 processing.



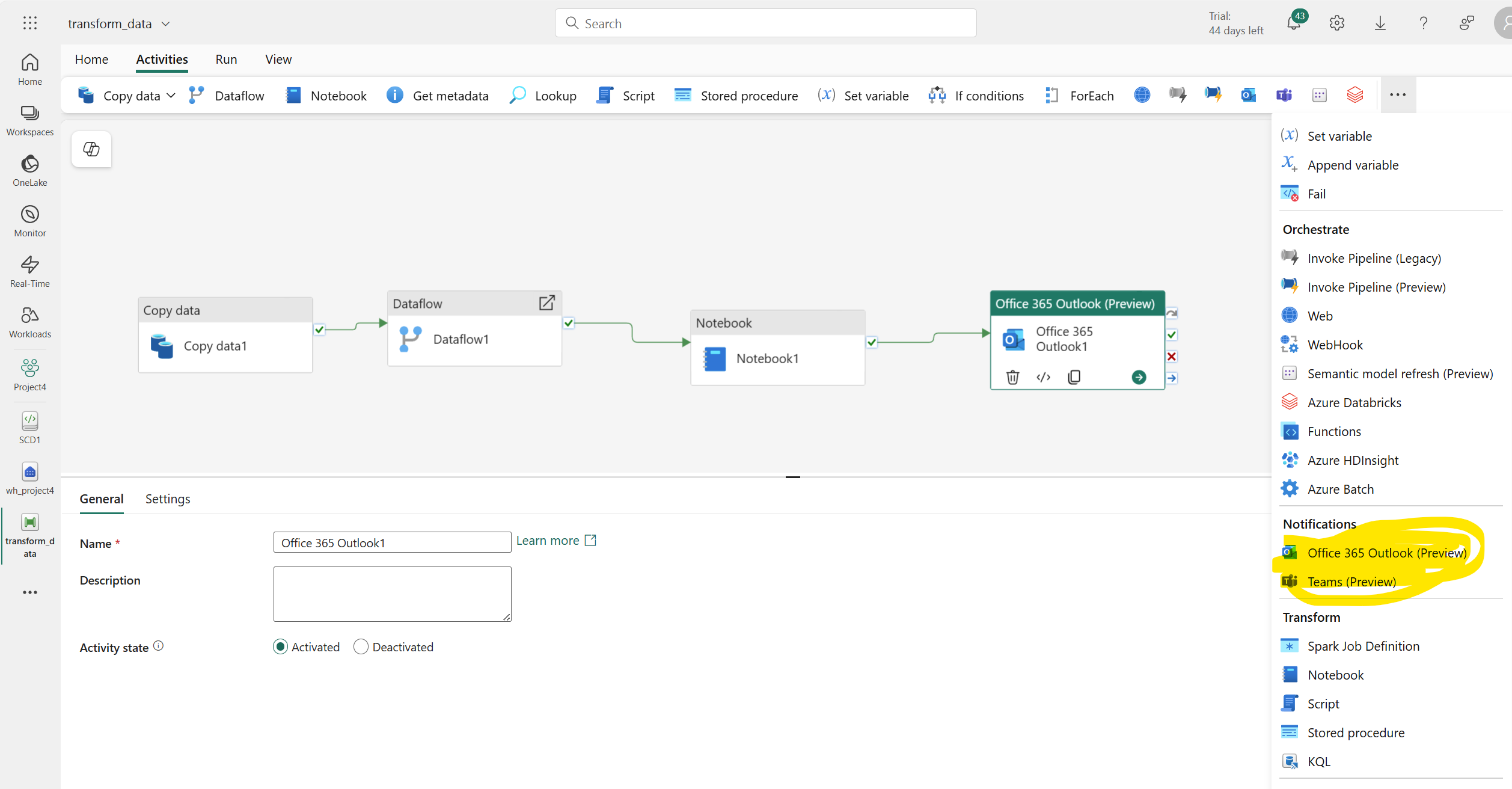
1. Click on add to pipeline and then connect success node from dataflow to the notebook.



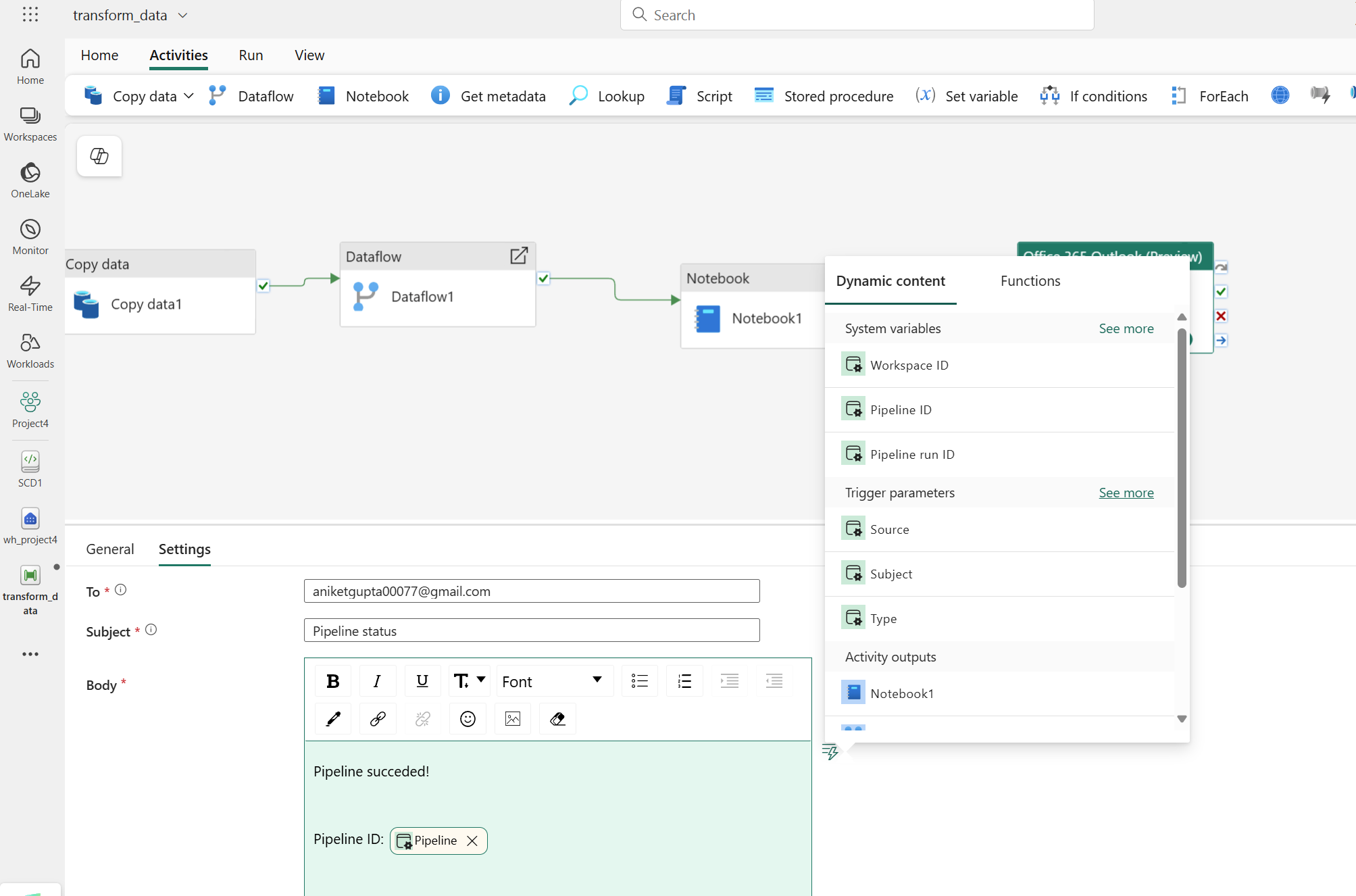


## **Add Notification Activity to notify pipeline successful run.**

1. Under notifications->click on Office 365 Outlook.
2. Connect notebook success node with outlook activity.



1. In settings-> we can add system variable in the body like Pipeline Id, Workspace Id etc.



1. Debug the pipeline.