Create a Dataflow (Gen2) in Microsoft Fabric

In Microsoft Fabric, Dataflows (Gen2) connect to various data sources and perform transformations in Power Query Online. They can then be used in Data Pipelines to ingest data into a lakehouse or other analytical store, or to define a dataset for a Power BI report.

This lab is designed to introduce the different elements of Dataflows (Gen2), and not create a complex solution that may exist in an enterprise. This lab takes **approximately 30 minutes** to complete.

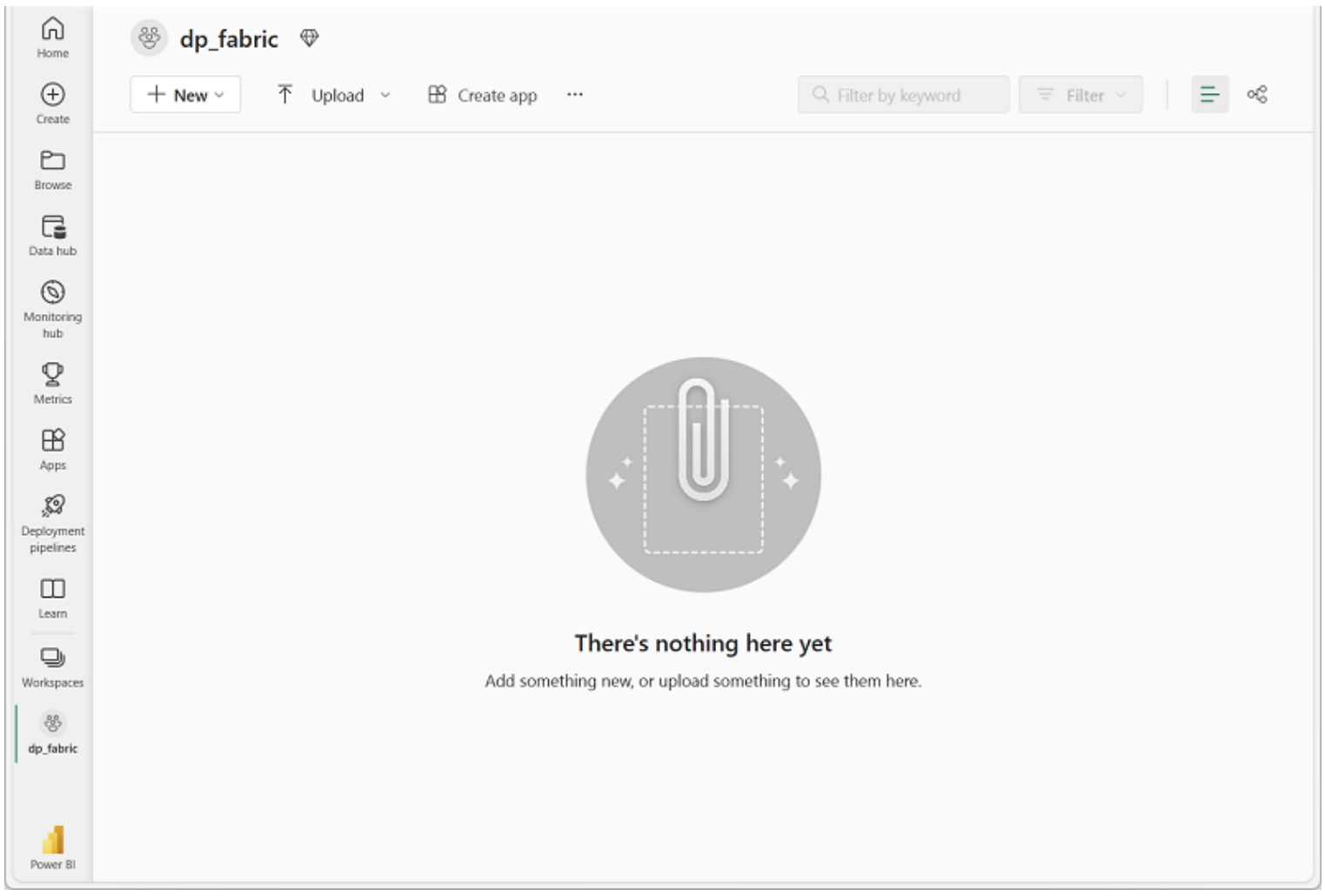
> **Note**: You need a Microsoft *school* or *work* account to complete this exercise. If you don't have one, you can sign up for a trial of Microsoft Office 365 E3 or higher.

Create a workspace

Before working with data in Fabric, create a workspace with the Fabric trial enabled.

1. On the Microsoft Fabric home page, select **Synapse Data Engineering**.
2. In the menu bar on the left, select **Workspaces** (the icon looks similar to).
3. Create a new workspace with a name of your choice, selecting a licensing mode that includes Fabric capacity (*Trial*, *Premium*, or *Fabric*).

4. When your new workspace opens, it should be empty.

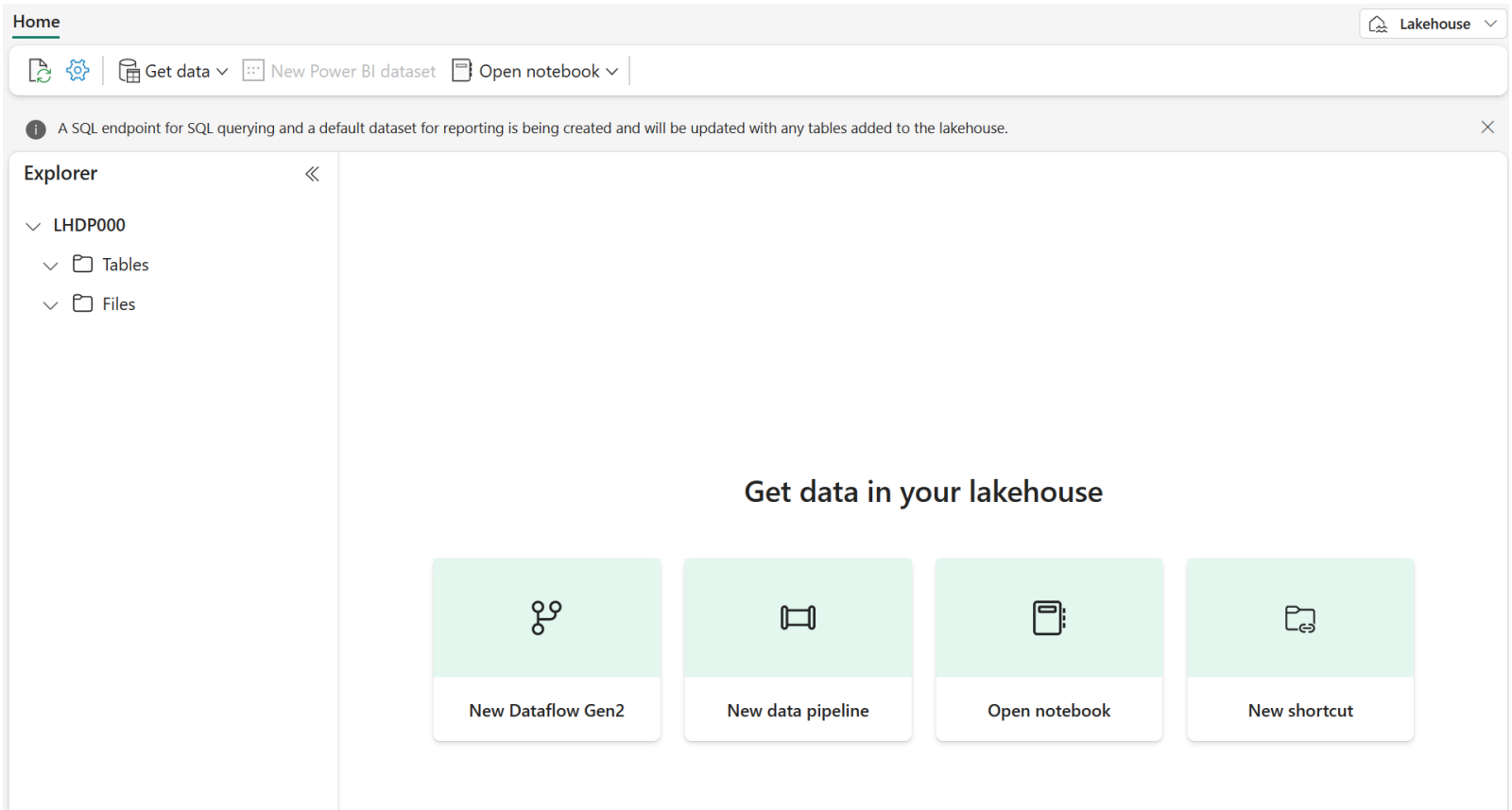


Create a lakehouse

Now that you have a workspace, it's time to create a data lakehouse into which you'll ingest data.

1. In the **Synapse Data Engineering** home page, create a new **Lakehouse** with a name of your choice.

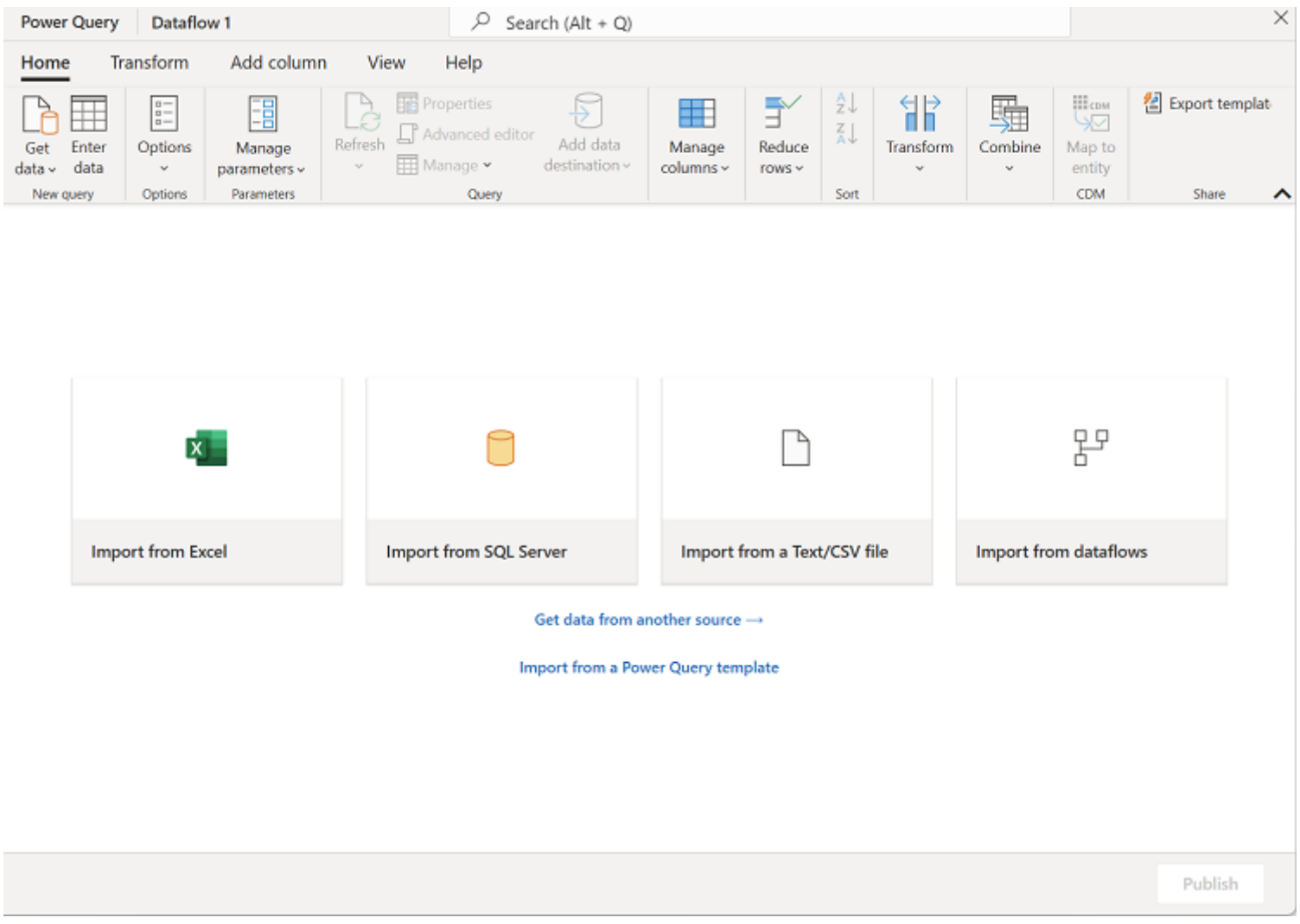
   After a minute or so, a new empty lakehouse will be created.



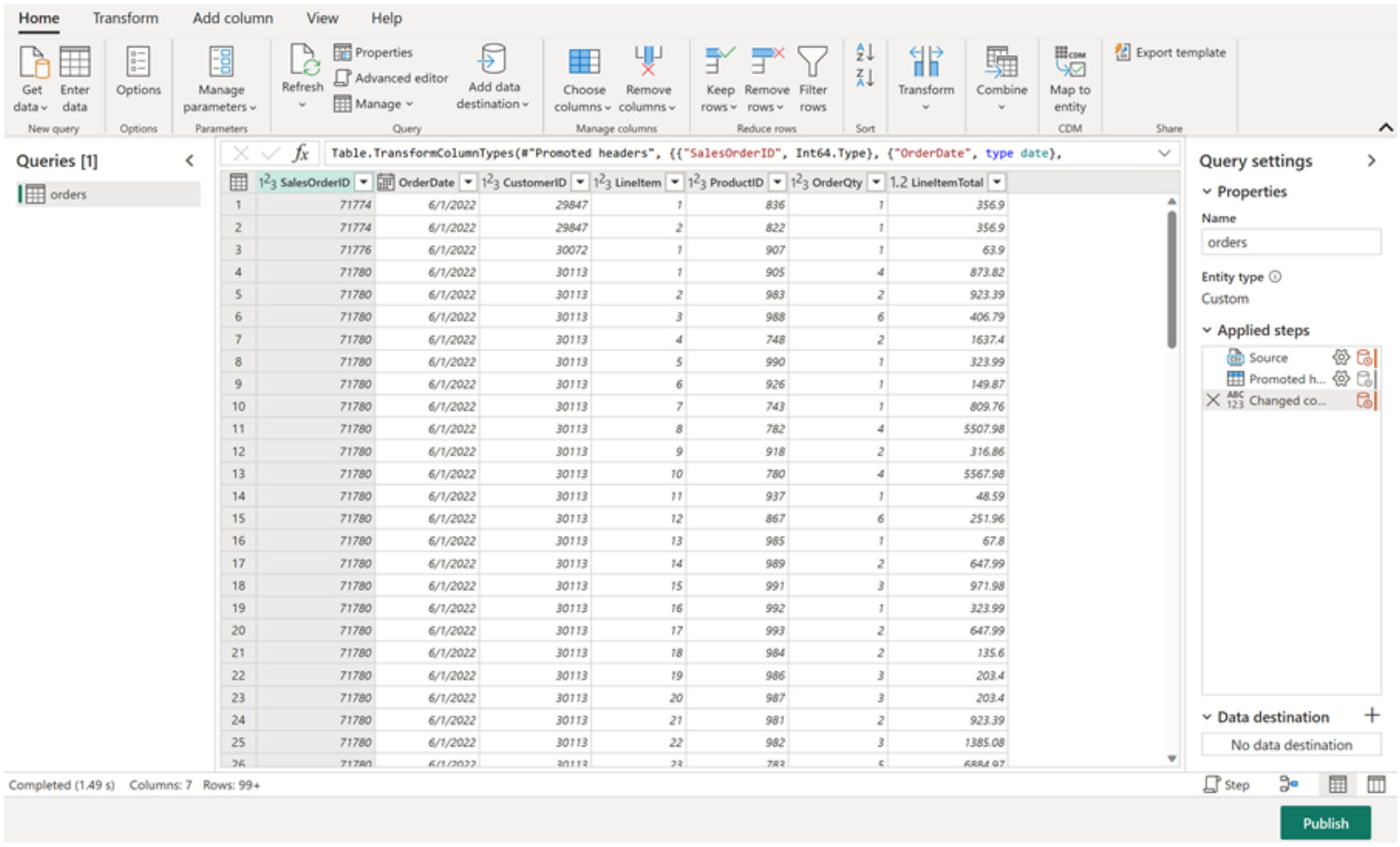
Create a Dataflow (Gen2) to ingest data

Now that you have a lakehouse, you need to ingest some data into it. One way to do this is to define a dataflow that encapsulates an *extract, transform, and load* (ETL) process.

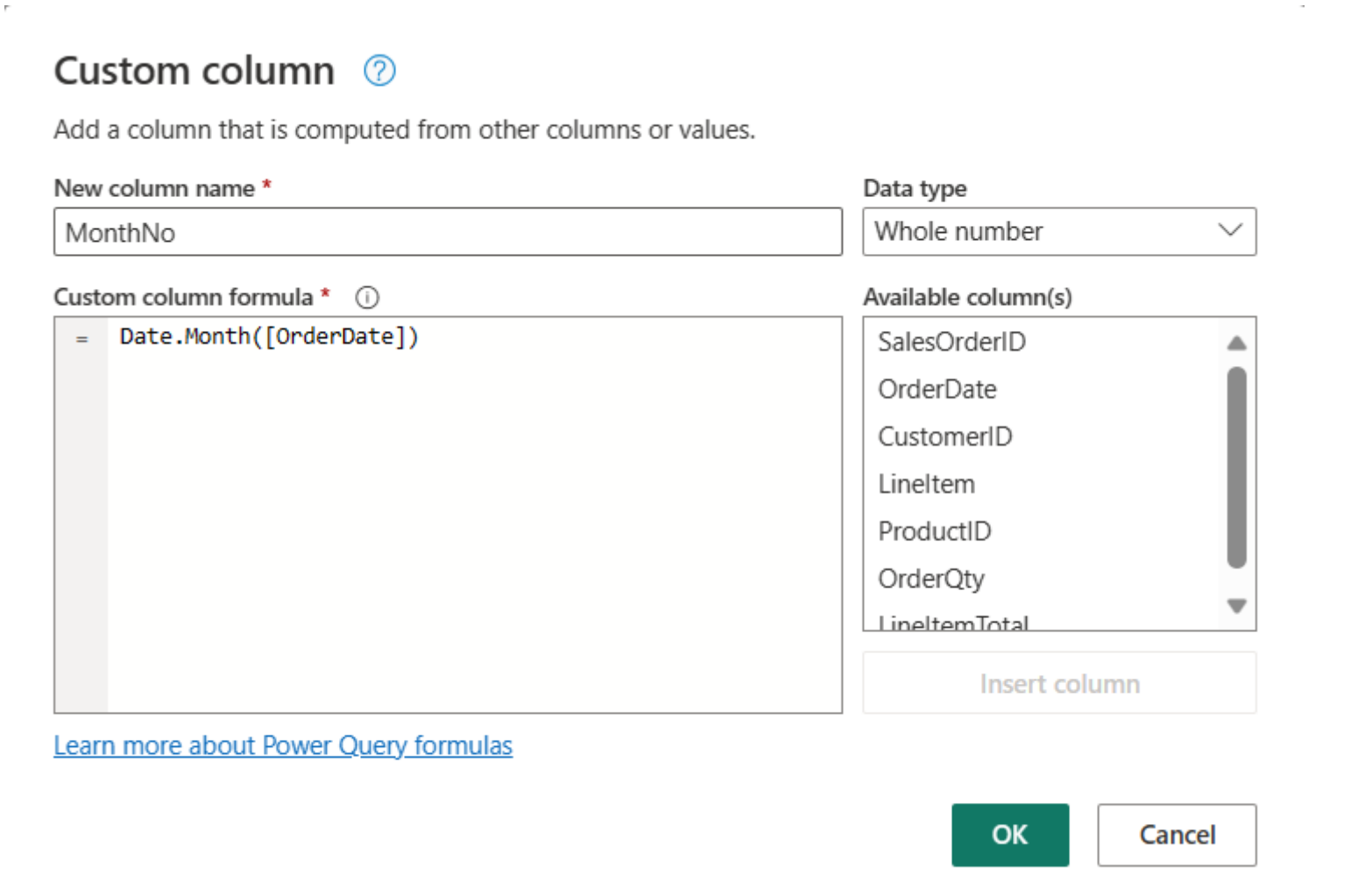
1. In the home page for your workspace, select **New Dataflow Gen2**. After a few seconds, the Power Query editor for your new dataflow opens as shown here.



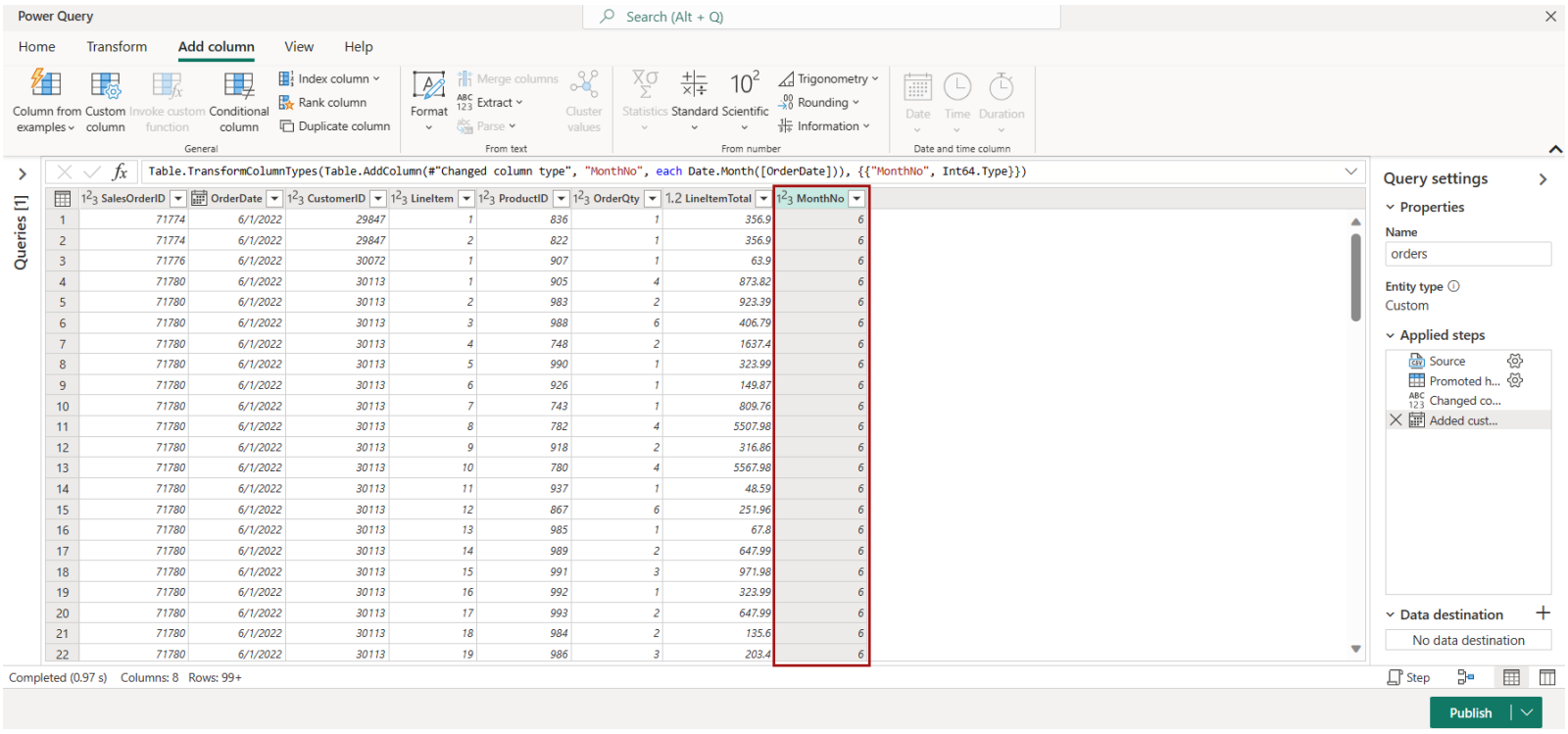
1. Select **Import from a Text/CSV file**, and create a new data source with the following settings:
    - **Link to file**: *Selected*
    - **File path or URL**: `https://raw.githubusercontent.com/MicrosoftLearning/dp-data/main/orders.csv`
    - **Connection**: Create new connection
    - **data gateway**: (none)
    - **Authentication kind**: Anonymous
2. Select **Next** to preview the file data, and then **Create** the data source. The Power Query editor shows the data source and an initial set of query steps to format the data, as shown here:



1. On the toolbar ribbon, select the **Add column** tab. Then select **Custom column** and create a new column.
2. Set the *New column name* to `MonthNo`, set the *Data type* to **Whole Number** and then add the following formula: `Date.Month([OrderDate])` - as shown here:



1. Select **OK** to create the column and notice how the step to add the custom column is added to the query. The resulting column is displayed in the data pane:



> **Tip:** In the Query Settings pane on the right side, notice the **Applied Steps** include each transformation step. At the bottom, you can also toggle the **Diagram flow** button to turn on the Visual Diagram of the steps. Steps can be moved up or down, edited by selecting the gear icon, and you can select each step to see the transformations apply in the preview pane.

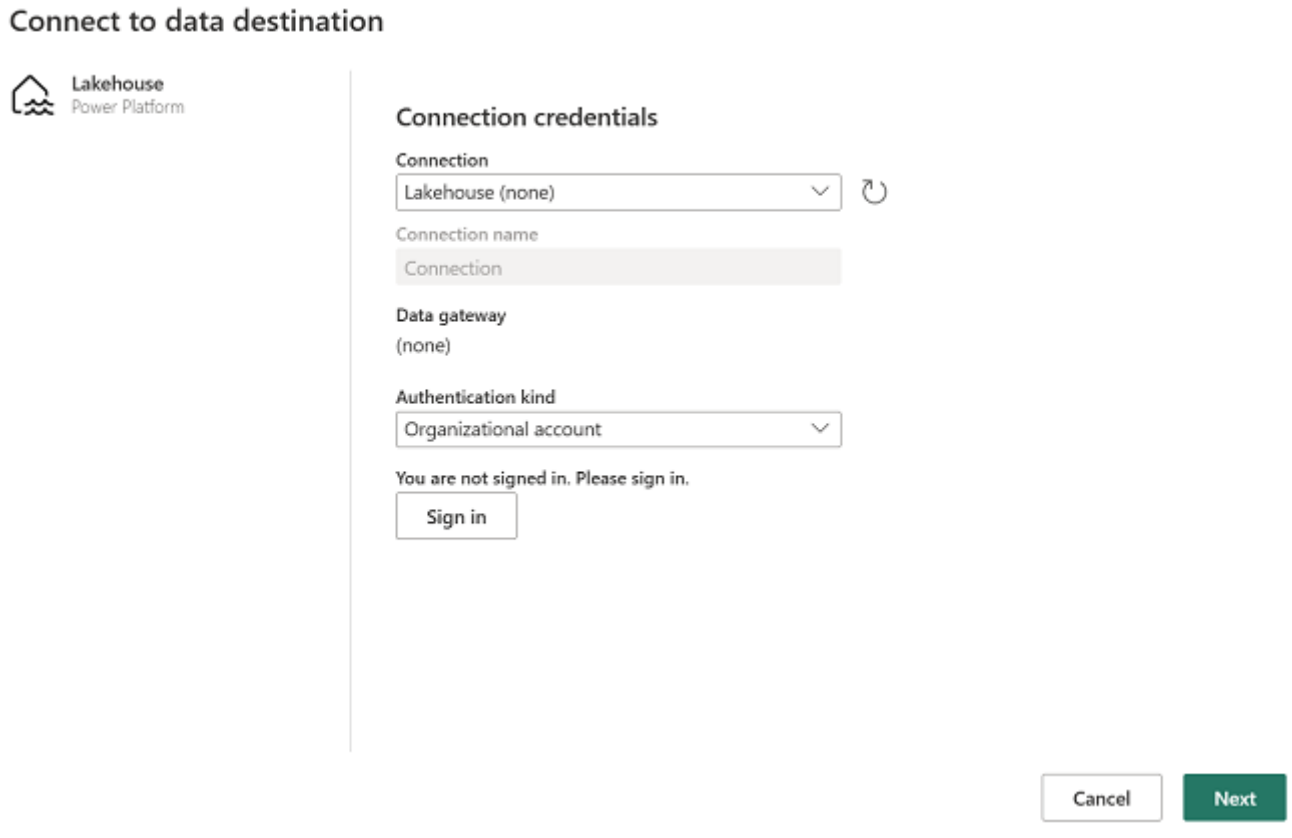
1. Check and confirm that the data type for the **OrderDate** column is set to **Date** and the data type for the newly created column **MonthNo** is set to **Whole Number**.

Add data destination for Dataflow

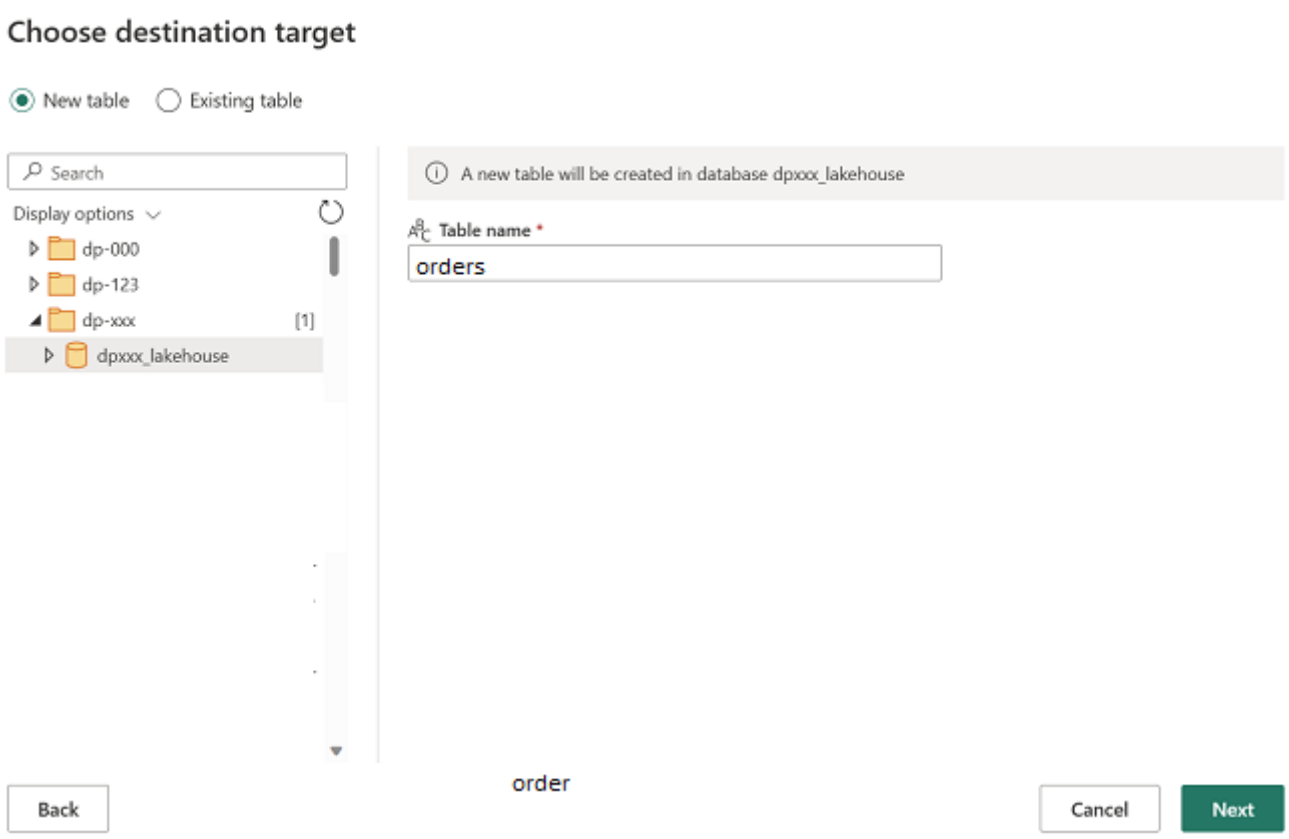
1. On the toolbar ribbon, select the **Home** tab. Then in the **Add data destination** drop-down menu, select **Lakehouse**.

> **Note:** If this option is grayed out, you may already have a data destination set. Check the data destination at the bottom of the Query settings pane on the right side of the Power Query editor. If a destination is already set, you can change it using the gear.

2. In the **Connect to data destination** dialog box, edit the connection and sign in using your Power BI organizational account to set the identity that the dataflow uses to access the lakehouse.

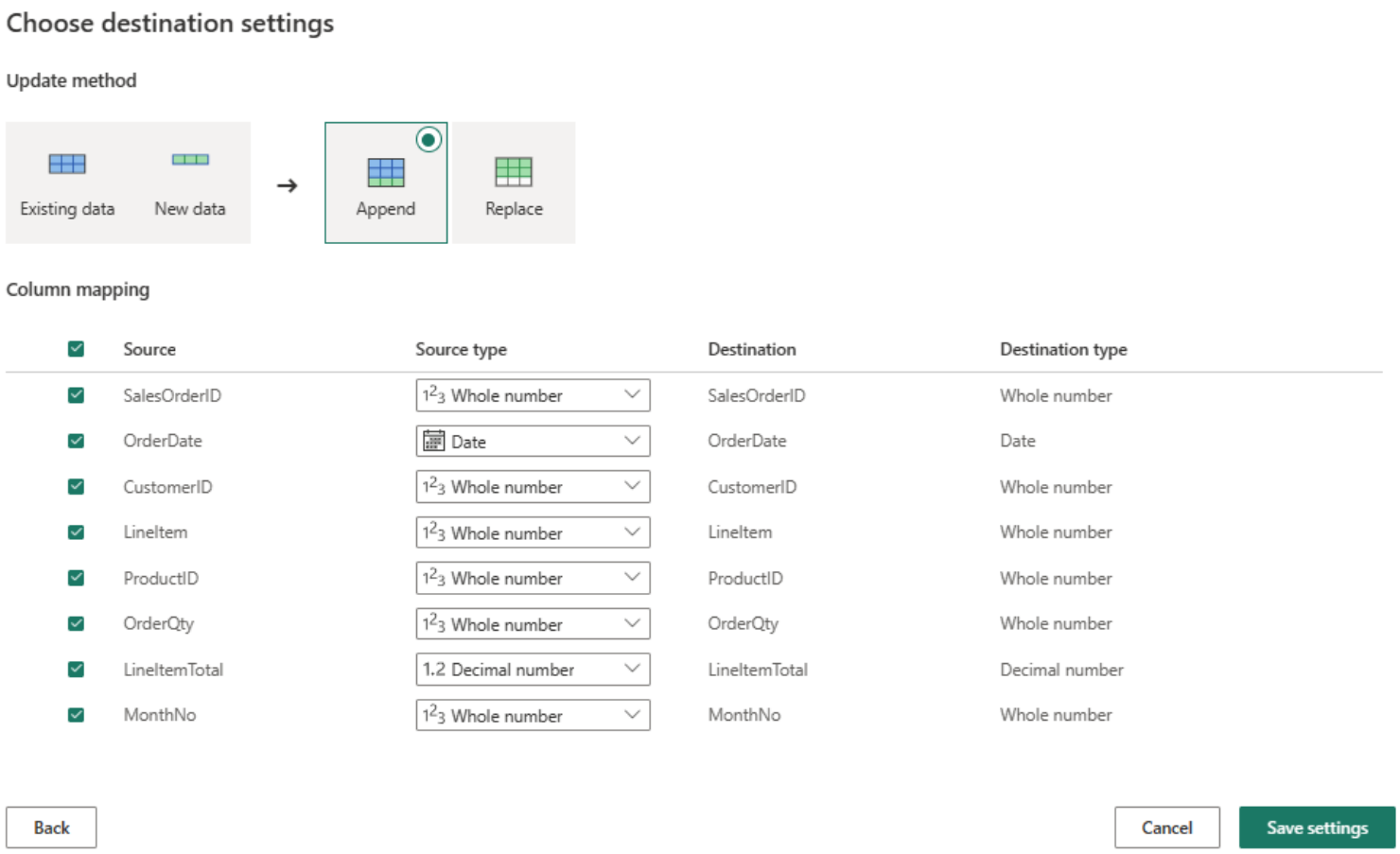


1. Select **Next** and in the list of available workspaces, find your workspace and select the lakehouse you created in it at the start of this exercise. Then specify a new table named **orders**:

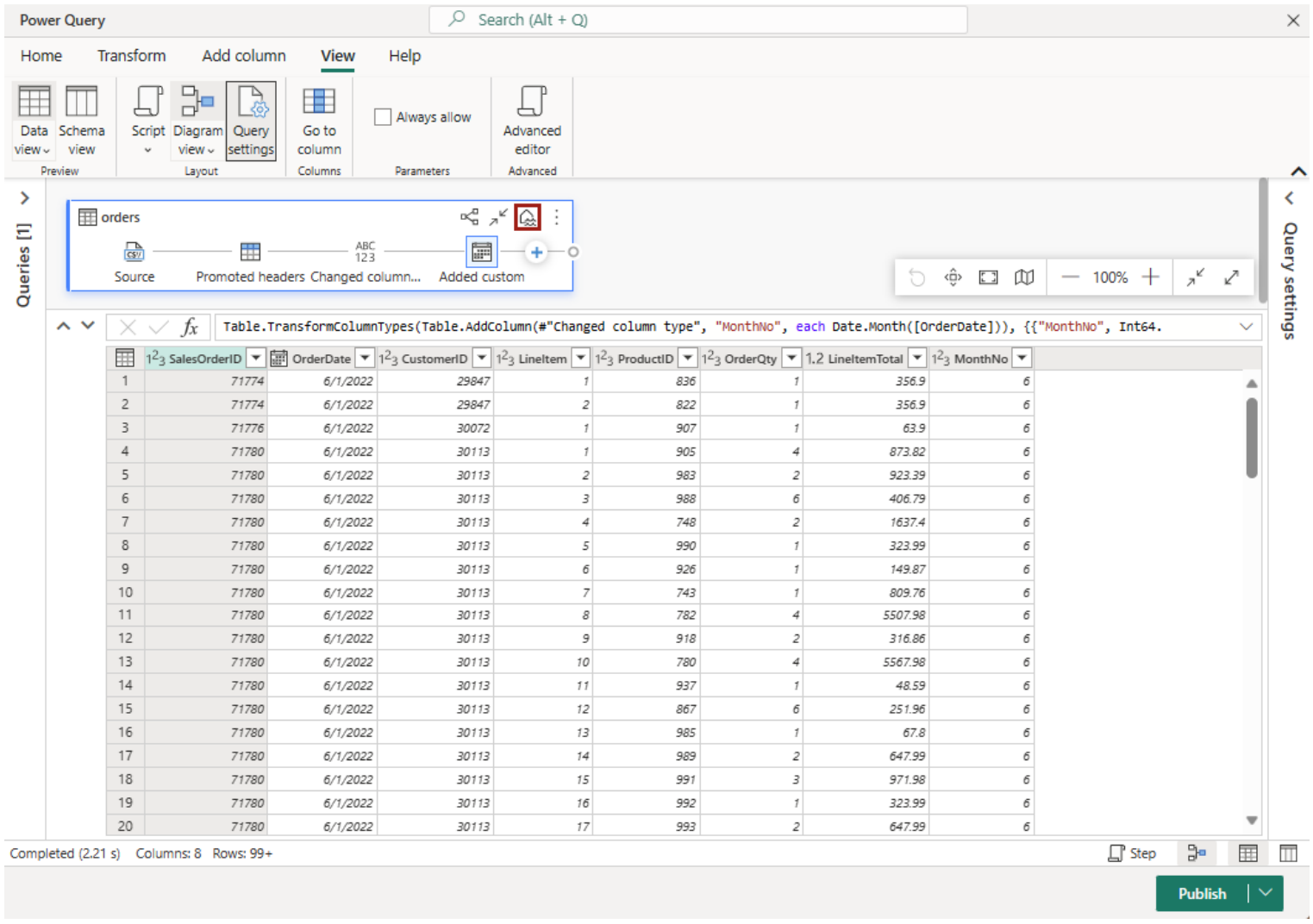


2. On the **Choose destination settings** page, select **Append** and then **Save settings**.

> **Note:** We suggest using the *Power query* editor for updating data types, but you can also do so from this page, if you prefer.



3. On the Menu bar, open **View** and select **Diagram view**. Notice the **Lakehouse** destination is indicated as an icon in the query in the Power Query editor.



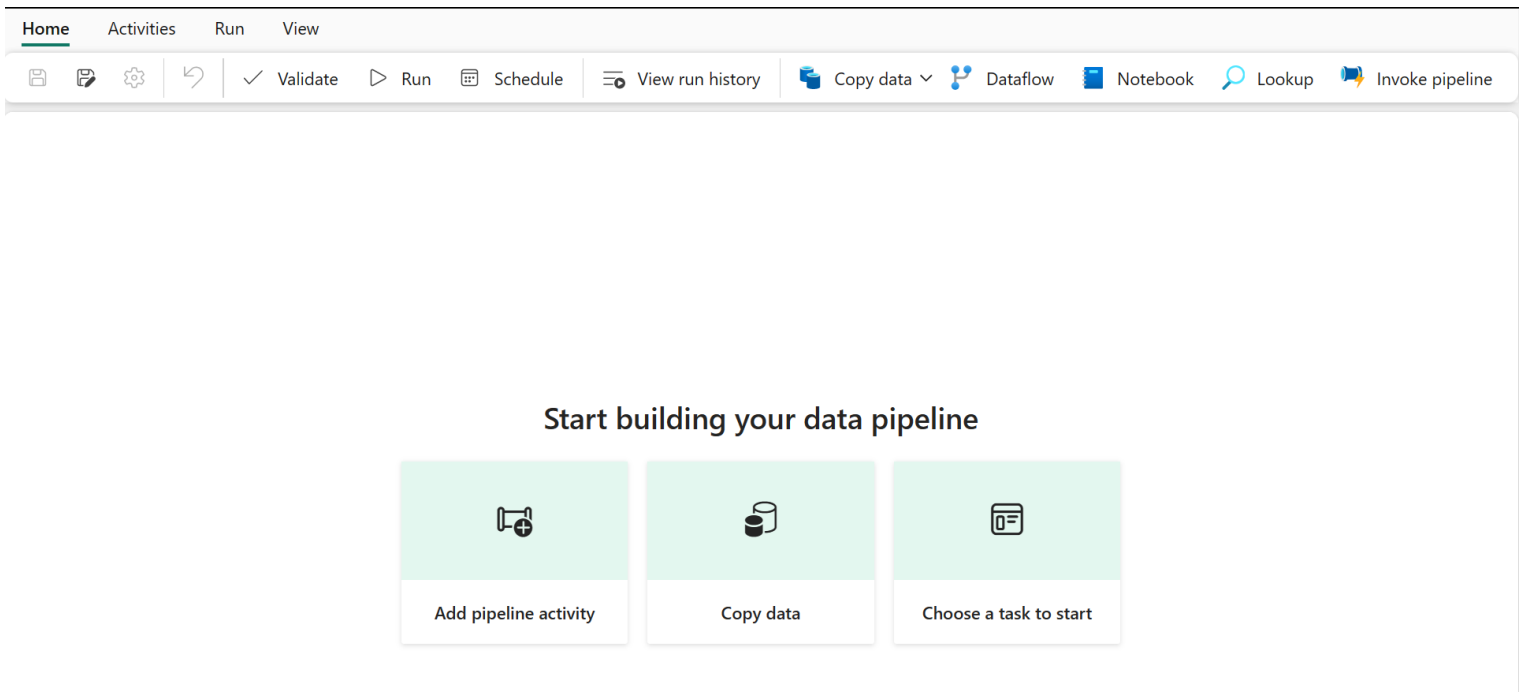
4. Select **Publish** to publish the dataflow. Then wait for the **Dataflow 1** dataflow to be created in your workspace.
5. Once published, you can right-click on the dataflow in your workspace, select **Properties**, and rename your dataflow.

Add a dataflow to a pipeline

You can include a dataflow as an activity in a pipeline. Pipelines are used to orchestrate data ingestion and processing activities, enabling you to combine dataflows with other kinds of operation in a single, scheduled process. Pipelines can be created in a few different experiences, including Data Factory experience.

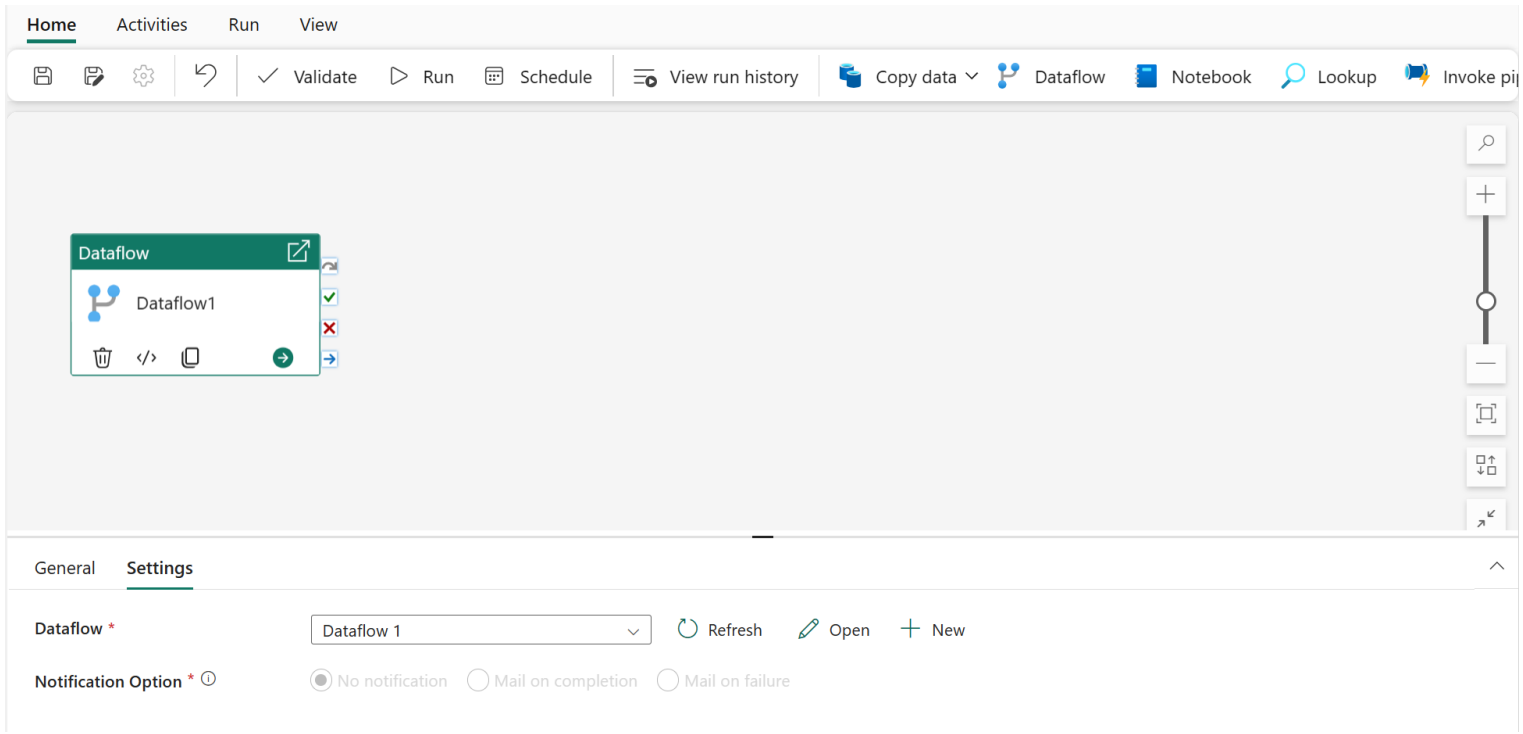
1. From your Fabric-enabled workspace, make sure you're still in the **Data Engineering** experience. Select **New**, **Data pipeline**, then when prompted, create a new pipeline named **Load data**.

   The pipeline editor opens.



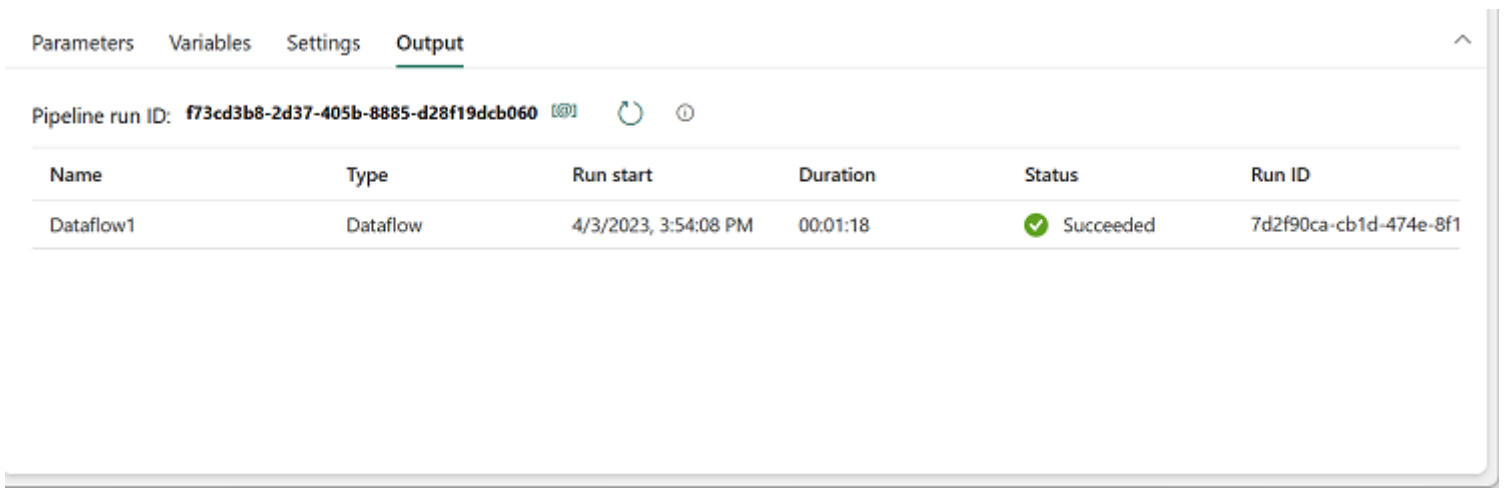
> **Tip**: If the Copy Data wizard opens automatically, close it!

2. Select **Add pipeline activity**, and add a **Dataflow** activity to the pipeline.
3. With the new **Dataflow1** activity selected, on the **Settings** tab, in the **Dataflow** drop-down list, select **Dataflow 1** (the data flow you created previously)



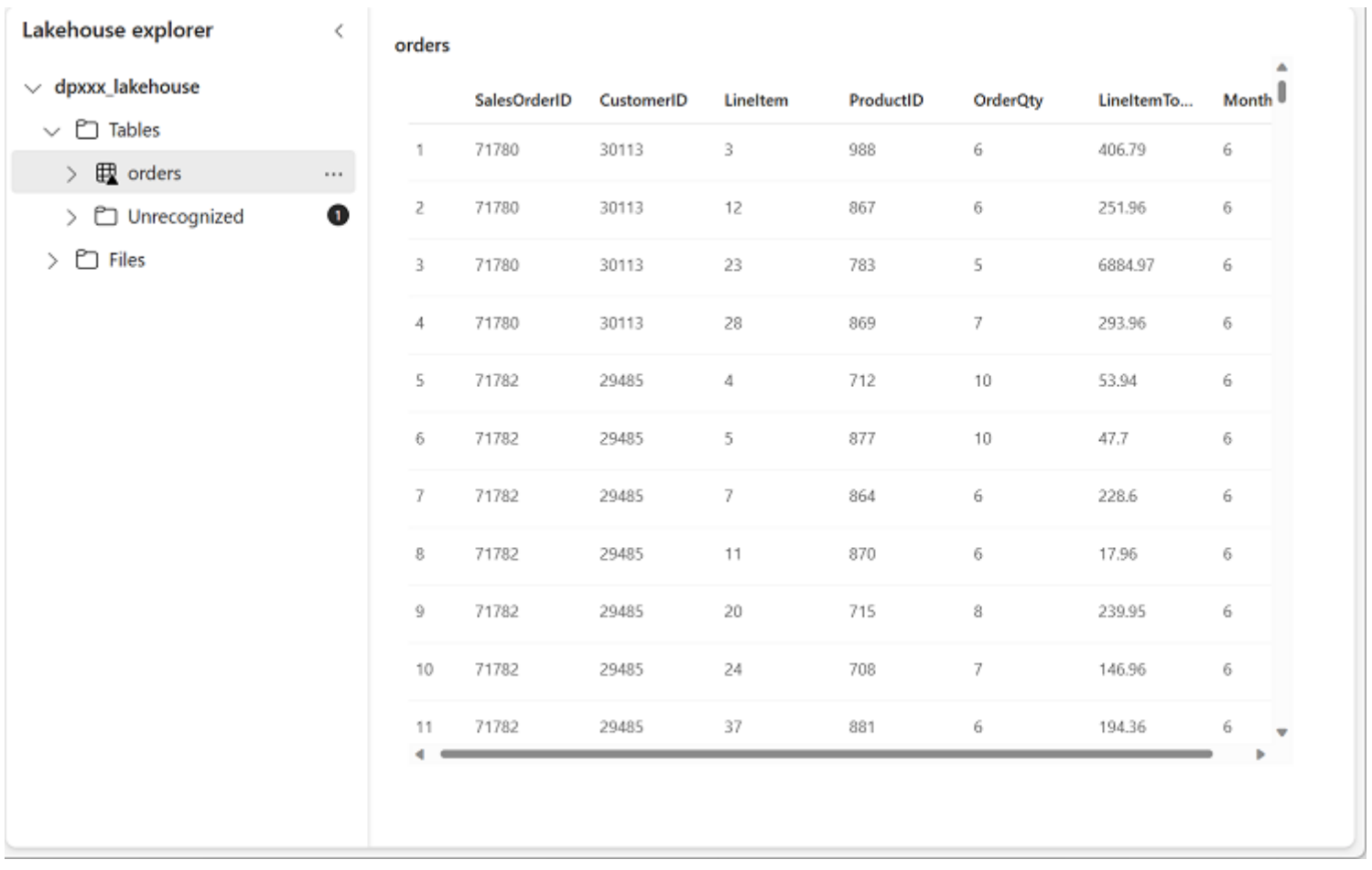
4. On the **Home** tab, save the pipeline using the 🖫 (*Save*) icon.

5. Use the ▷ **Run** button to run the pipeline, and wait for it to complete. It may take a few minutes.



6. In the menu bar on the left edge, select your lakehouse.

7. In the **...** menu for **Tables**, select **refresh**. Then expand **Tables** and select the **orders** table, which has been created by your dataflow.



Create a workspace

Create a lakehouse

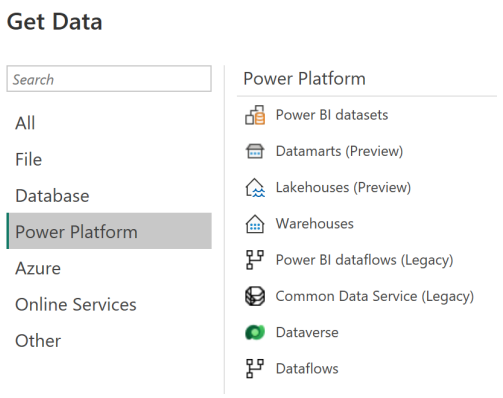
Create a Dataflow (Gen2) to ingest data

Add data destination for Dataflow

Add a dataflow to a pipeline

Clean up resources

> **Tip**: Use the Power BI Desktop *Dataflows connector* to connect directly to the data transformations done with your dataflow. You can also make additional transformations, publish as a new dataset, and distribute with intended audience for specialized datasets.



Clean up resources

If you've finished exploring dataflows in Microsoft Fabric, you can delete the workspace you created for this exercise.

1. Navigate to Microsoft Fabric in your browser.
2. In the bar on the left, select the icon for your workspace to view all of the items it contains.
3. In the **...** menu on the toolbar, select **Workspace settings**.
4. In the **Other** section, select **Remove this workspace**.
5. Don't save the changes to Power BI Desktop, or delete the .pbix file if already saved.