A close-up of a person smiling

Description automatically generated

# Lab 04: Dataflow Gen 2 in action

### 

### Introduction

In this lab, you will use Dataflow Gen2 to execute some advanced Data Preparation activities. You will first consume a complex JSON file with 2 methods, then you will extract and transform data from an Open Data environment based on REST API.

### Objectives

After completing this lab, you will be better able to:

1. Import and Transform Data from a complex data structure
2. Understand how to implement advanced M code
3. Import and Transform Data from a REST API call

**Estimated time to complete this lab**

60 minutes

Contents

[Lab 04: Dataflow Gen 2 in action 1](#_Toc148134563)

[Introduction 1](#_Toc148134564)

[Objectives 1](#_Toc148134565)

[Task 1: Import and Transform Data from a complex data structure 3](#_Toc148134566)

[Task 2: Execute advanced M code 14](#_Toc148134567)

[Task 3: Extract data from OData 19](#_Toc148134568)

**Lab Prerequisites**

* Workspace: Fabric, Power Premium or Fabric trial
* Individual license: Power Pro or Premium Per User account

**Information provided by your training provider**

* Trial tenant (if applicable): login & password, workspace to use for the lab.
* Azure Data Lake Gen2 (containing data sources): account name & shared access signature.

### Task 1: Import and Transform Data from a complex data structure

In this task, you will import data from a JSON file then manipulate the structure before loading it to the Lakehouse.

The file to be loaded is name **Contoso.json** and is stored in the labdate/json folder from the Azure Data Lake Store Gen 2 account used in the previous labs.

A screenshot of a computer

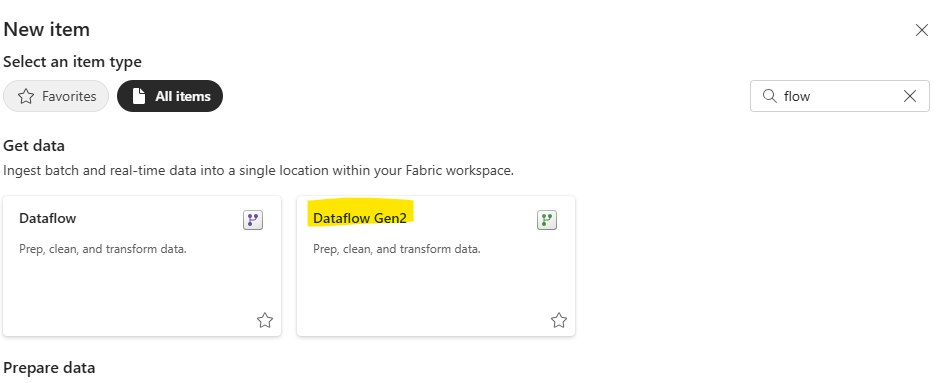
Description automatically generated

This file contains a list of contacts, with several attributes (age, gender…) and some multiple value attributes (tags, friends).

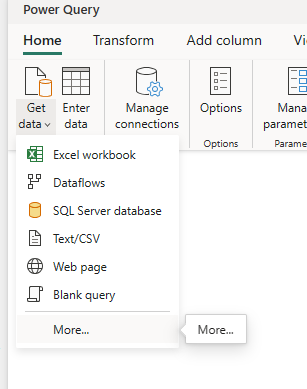
A screenshot of a computer program

Description automatically generated

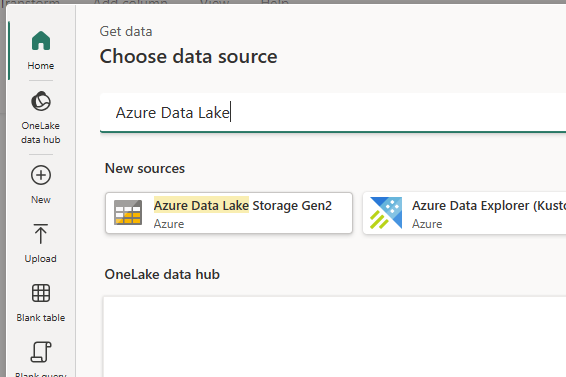
Create a **Dataflow Gen2** artifact in your workspace.



Go to the **Get Data** Menu and click on **More.**

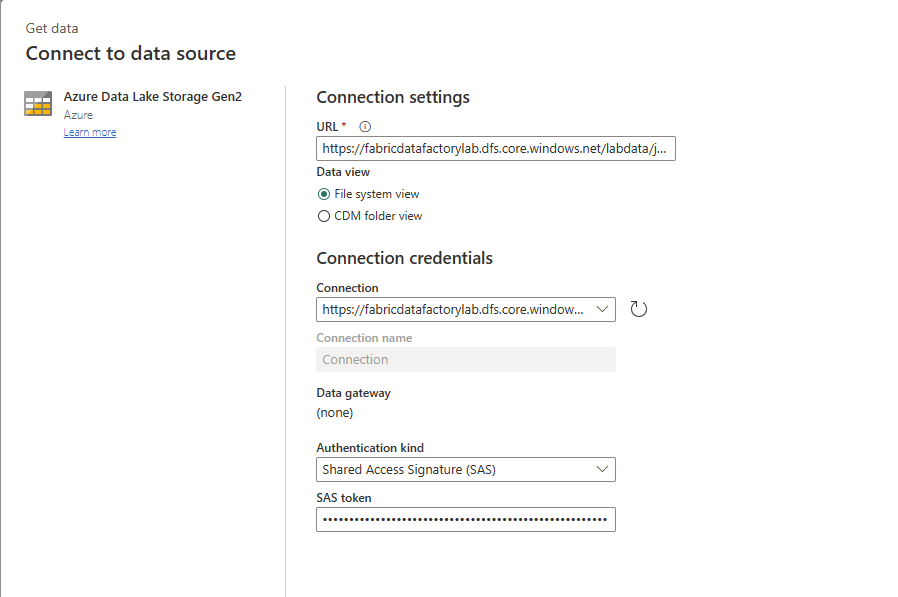
****

On the Choose data source window, type **Azure Data Lake** and select **Azure Data Lake Storage Gen2.**

****

Configure the connection to the Contoso json file stored in the ADLS Gen2 account using a Shared Access Signature.

* **URL**: https://<storage account>.dfs.core.windows.net/labdata/json/Contoso.json
* **Authentication kind**: Shared Access Signature (SAS)
* **SAS Token**: value token provided by your trainer

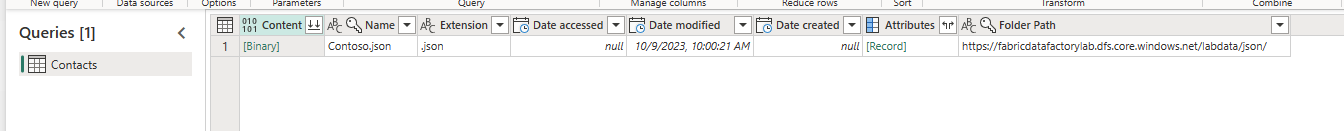
****

Click on **Create** to edit the query and extract data from the Contoso.json file.

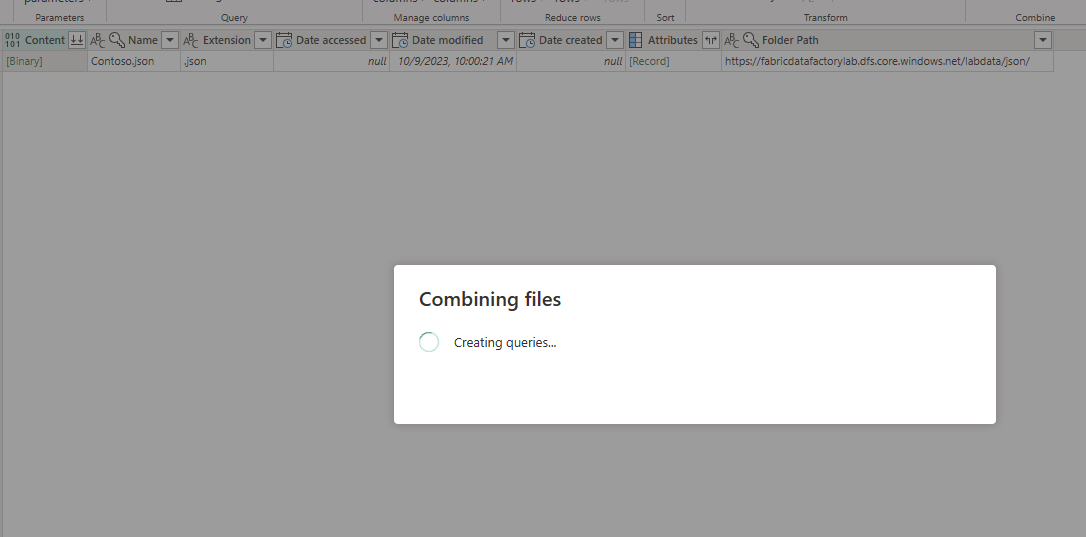
A screenshot of a computer

Description automatically generated

Rename the query named “Query” as “Contacts” in the Queries panel, and click on the icon in the **Content** column header to ask the Dataflow Gen2 to explorer and extract the data from JSON file.



The Dataflow engine will automatically analyze the file structure and display the query result.



The query result is displayed on the Data Grid, and collection of resources have been automatically added.

A screenshot of a computer

Description automatically generated

Select the **Contacts** query and clone it to create the new query **ContactTags** (to clone a query, right click, and select “Duplicate”).

A screenshot of a computer

Description automatically generated

Select the **Contacts** query, and as indicated in the screenshot hereafter, select the following group of columns: age, eyeColor, name, gender, company, email and phone. Keep the selection active, right on the columns header and select **Remove other columns**.

A screenshot of a computer

Description automatically generated

The **Contacts** query should only contain the selected columns.

A screenshot of a computer

Description automatically generated

Repeat the same operation for the **ContactTags** query, to keep the following columns: name and tags.

Then click on the small button located on the **tags** column header to expand the column values (which is a list).

A screenshot of a computer

Description automatically generated

As the **ContactTags** query has been expanded, the same name appears who each corresponding tag value.

A screenshot of a computer

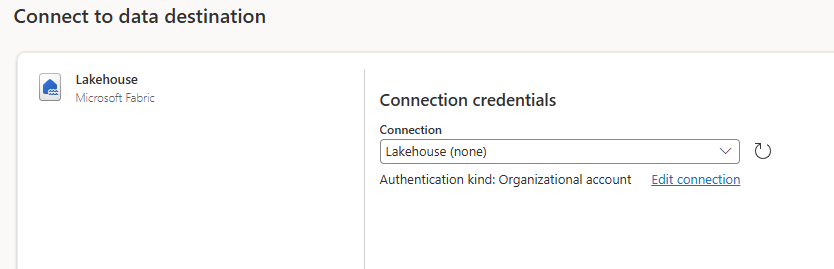
Description automatically generated

Click on the + icon in the **Data destination** area and select **Lakehouse**.

A screenshot of a computer

Description automatically generated

Keep the default configuration and click **Next**.



Select the existing **Contoso** lakehouse located in the same workspace, keep the destination table name as **ContactTags**, and click on **Next**.

A screenshot of a computer

Description automatically generated

Disable the “Use Automatic Settings” option, and configure the mapping as indicated hereafter, and click on **Save settings**.

A screenshot of a computer

Description automatically generated

Repeat the same operation for the **Contacts** query to load data in the **Contoso** Lakehouse using **Contacts** as the destination table name.

**A screenshot of a computer

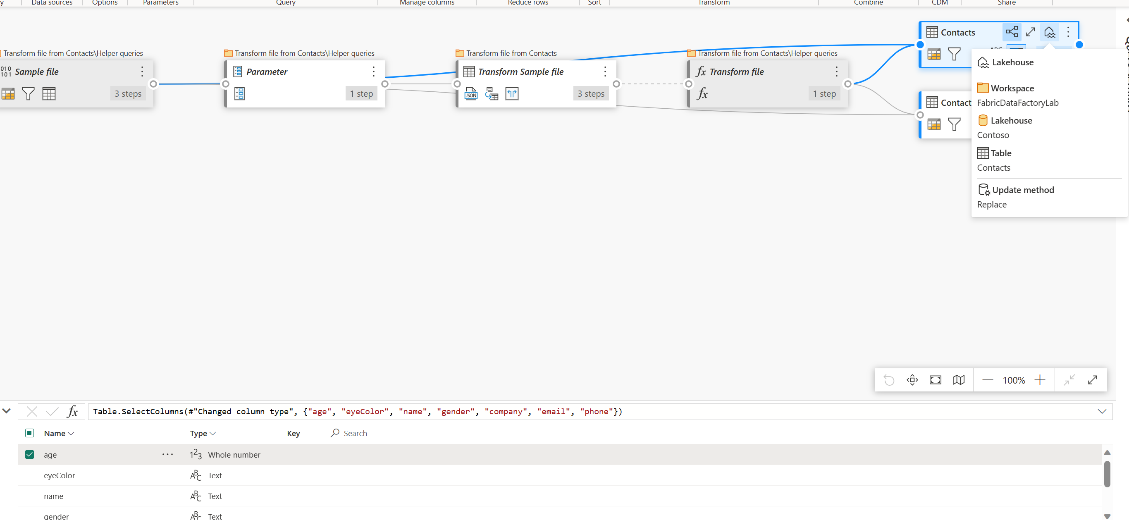
Description automatically generated**

Click on **Diagram view** to see graphically how the 2 queries are structured.

A screenshot of a computer

Description automatically generated

The **Contacts** query display the Lakehouse icon displaying where and how the data will be loaded.



Expand the **Contacts** query to display the sequence of transformation step.

A screenshot of a computer

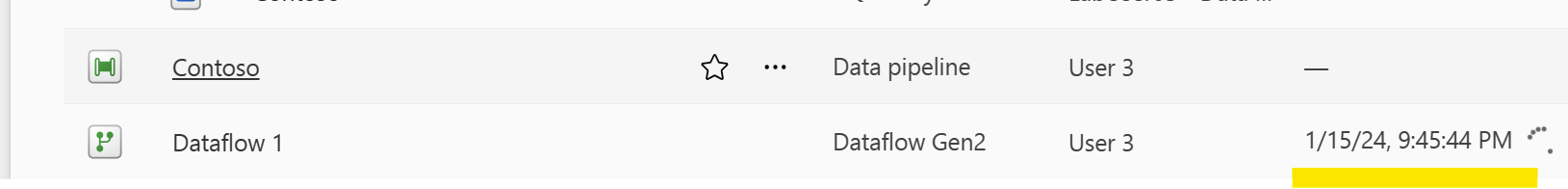
Description automatically generated

Click on **Publish** to save the Dataflow definition and start its first execution.

A screenshot of a computer

Description automatically generated

Wait until the end of the first Dataflow execution, by monitoring the column **Refreshed.**



When the dataflow has been successfully executed, click on the Dataflow and select **Properties.**

A screenshot of a computer

Description automatically generated

Rename the dataflow as **ContactsLoading.**

A screenshot of a computer

Description automatically generated

Open the Contoso Lakehouse, and open the 2 tables Contacts and ContactTags, to make sure they contain data loaded from the CSV file.

A screenshot of a computer

Description automatically generated

### Task 2: Execute advanced M code

In this task, you will process the same **Contoso.json** file using an advanced M (mashup) code snippet which:

* Directly extracts JSON structure based on a schema definition provided in the code
* Combines 2 instances of the same table object (with a single read of the json file)
* Expose detailed information about the main contact and the friend

A screenshot of a computer code

Description automatically generated

Edit the dataflow named **ContactsLoading** (right click and select **Edit**).

A close-up of a computer screen

Description automatically generated

Click on **GetData** and select **Blank query**

A screen shot of a computer

Description automatically generated

Paste the M code from the M\_snippet.txt file, and replace <storage account> with the account name already used in the previous task.

Take the time to analyze the M code which is document with comments.

A close-up of a computer screen

Description automatically generated

After few seconds, the query result should appear in the data grid.

Rename the query as **ContactsWithFriends**.

A screenshot of a computer

Description automatically generated

In the Query settings panel, select the first step named **Source**, and look and the displayed properties : the lightning indicates that the activity is evaluated by the data source.

A screenshot of a computer

Description automatically generated

Select the last step **PersonWithFriendInfo**, right click and select **View query plan**.

A screenshot of a computer

Description automatically generated

Review the Query Plan displayed with:

* The buffered tasks
* The streamed tasks

**A screenshot of a computer screen

Description automatically generated**

Close the query plan, and load the query to the Lakehouse using **ContactsWithFriends** as the table name.

**A screenshot of a test

Description automatically generated**

Configure the source type as defined hereafter, and click on **Save Settings**.

**A screenshot of a computer

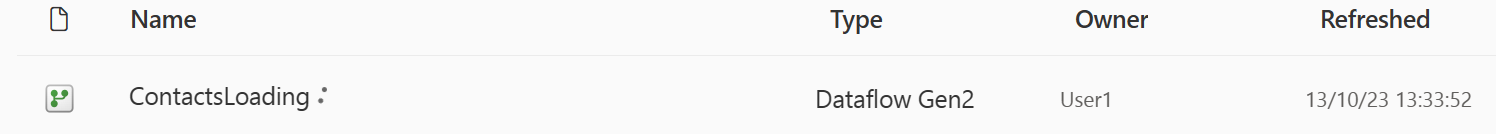
Description automatically generated**

Click on **Publish**

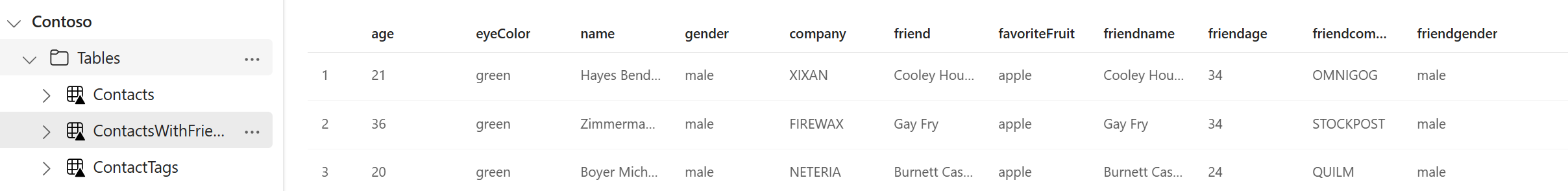
A screenshot of a computer

Description automatically generated

Wait until the end of the dataflow execution.

****

Control the table content in the Lakehouse.

****

### Task 3: Extract data from OData

In this task, you will extract data from an OData API endpoint: <https://services.odata.org/V4/Northwind/Northwind.svc/>

A white background with text

Description automatically generated

Create a **Dataflow Gen2** artifact in your workspace.

A screenshot of a computer

Description automatically generated

Click on **Get Data** and choose the **OData** connector.

A screenshot of a computer

Description automatically generated

Define the URL and use the **Anonymous** authentication kind.

A screenshot of a computer

Description automatically generated

Select the table **Invoices** and **Products** and click on Create.

A screenshot of a computer

Description automatically generated

Select the **Products** query, expand the **Category** table by selecting the **CategoryName** column.

A screenshot of a computer

Description automatically generated

Display the execution plan for the **Products** table which contains 2 possible execution plans, one containing the “Expand” predicate to extract the Category Name from the Products entity.

A screenshot of a computer

Description automatically generated

Select the **Invoices** query, then select the Merge queries from the Home ribbon.

A screenshot of a computer

Description automatically generated

Define the **Merge** transformation using the **ProductID** for the **Inner** join and validate.

A screenshot of a computer

Description automatically generated

Expand the created table object **Products** and select the columns **UnitsInStock**, **CategoryName** and **Supplie**r.

A screenshot of a computer

Description automatically generated

Configure the Lakehouse as the query destination, using **Invoices** as table name.

A screenshot of a computer

Description automatically generated

**Publish** the Dataflow.

Explore the **Invoices** table in the Lakehouse.

A screenshot of a computer

Description automatically generated

Wait until the end of the dataflow execution and rename it as **Invoices**.

A white rectangular object with black lines

Description automatically generated

Open the **Refresh Schedule** option for the **Invoices** dataflow.

A screenshot of a computer

Description automatically generated

Select the **Refresh history** link.

A screenshot of a computer

Description automatically generated

**Refresh history** gives details about the recent dataflow execution.

A screenshot of a computer

Description automatically generated

Configure and **Daily refresh** and click on **Apply**.

A screenshot of a computer

Description automatically generated