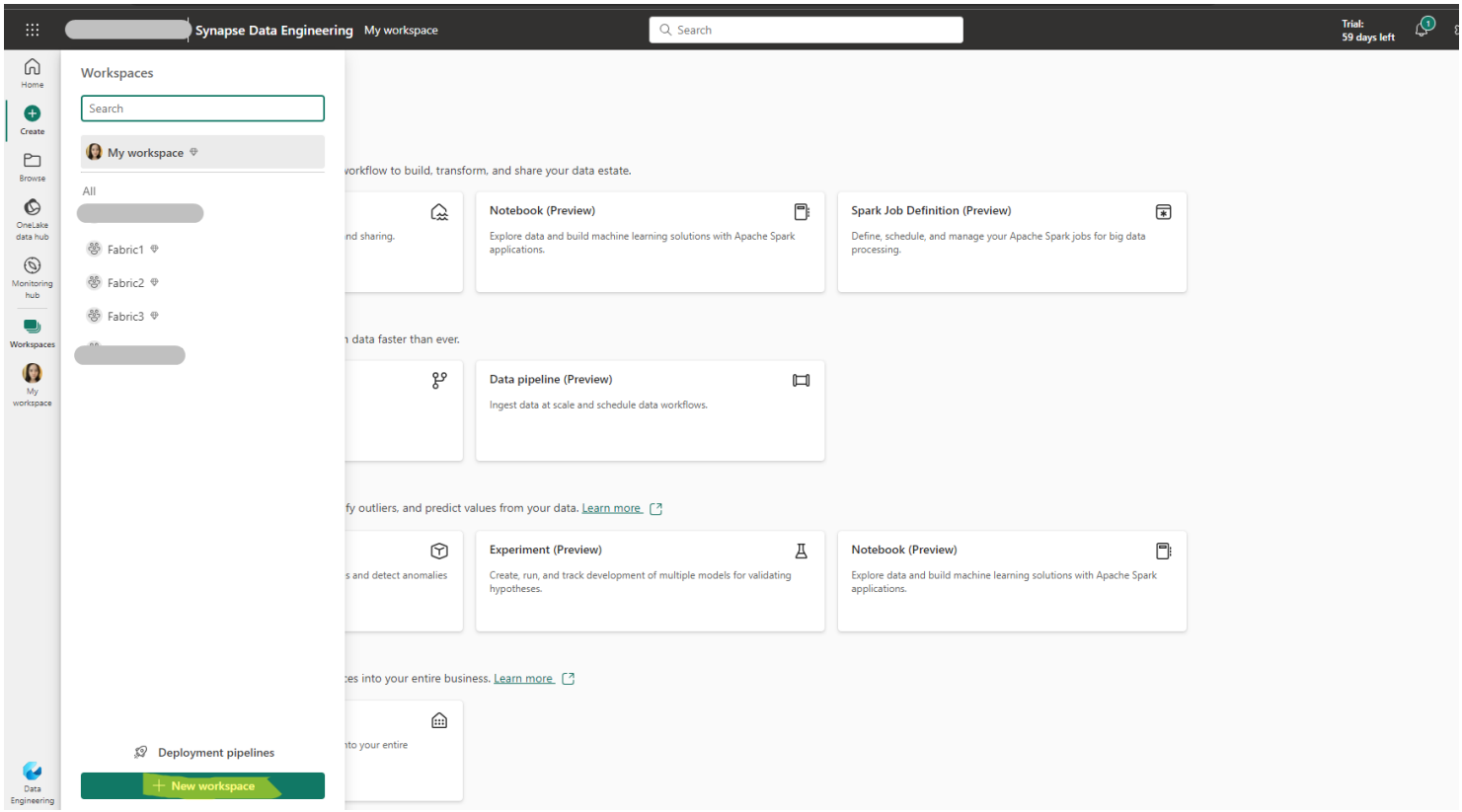
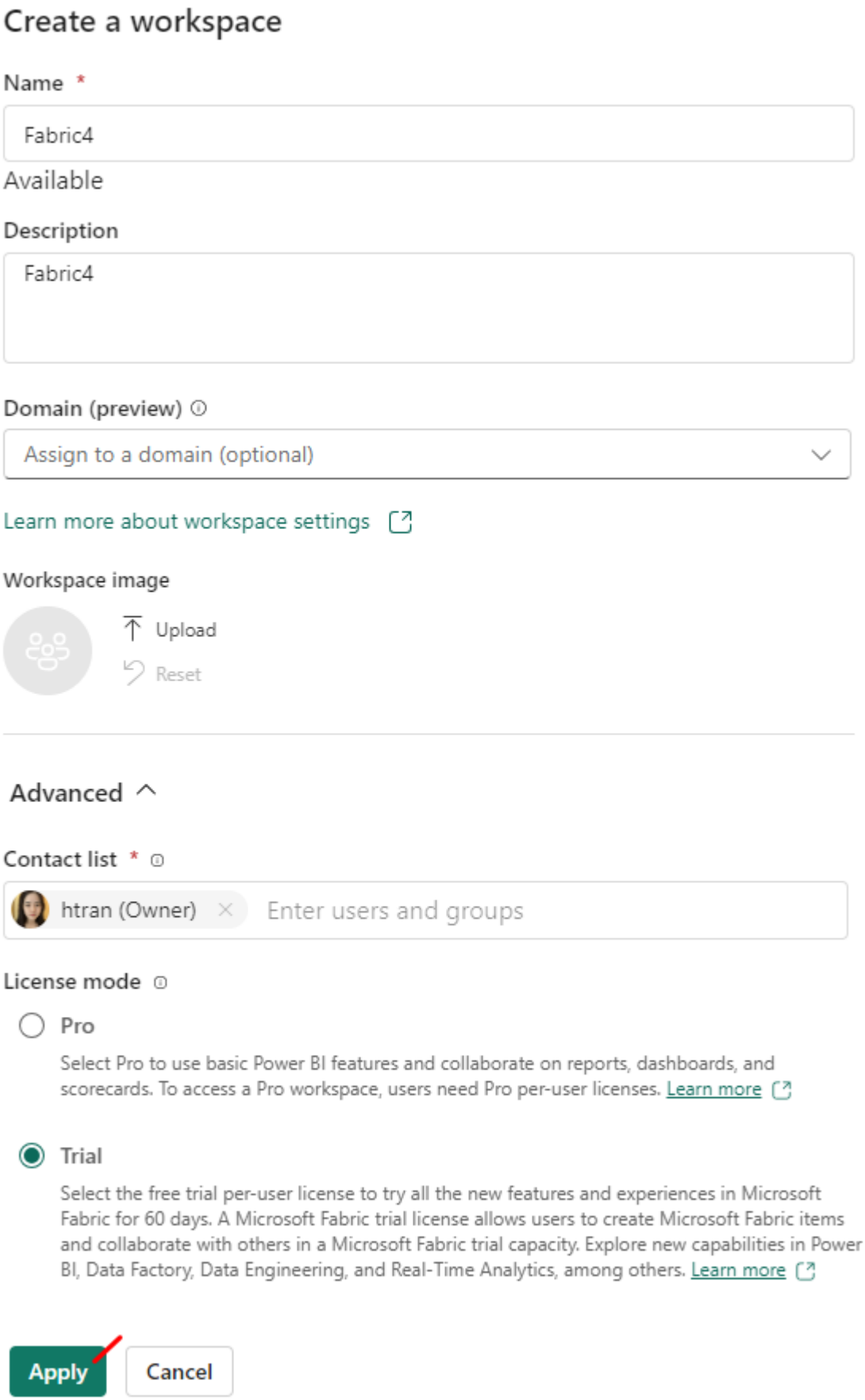
# 1. Lakehouse

A Lakehouse is a data architecture platform for storing, managing, and analyzing structured and unstructured data in a single location. It is a flexible and scalable solution that allows organizations to handle large volumes of data using a variety of tools and frameworks to process and analyze that data

## 1.1 Create Workspace

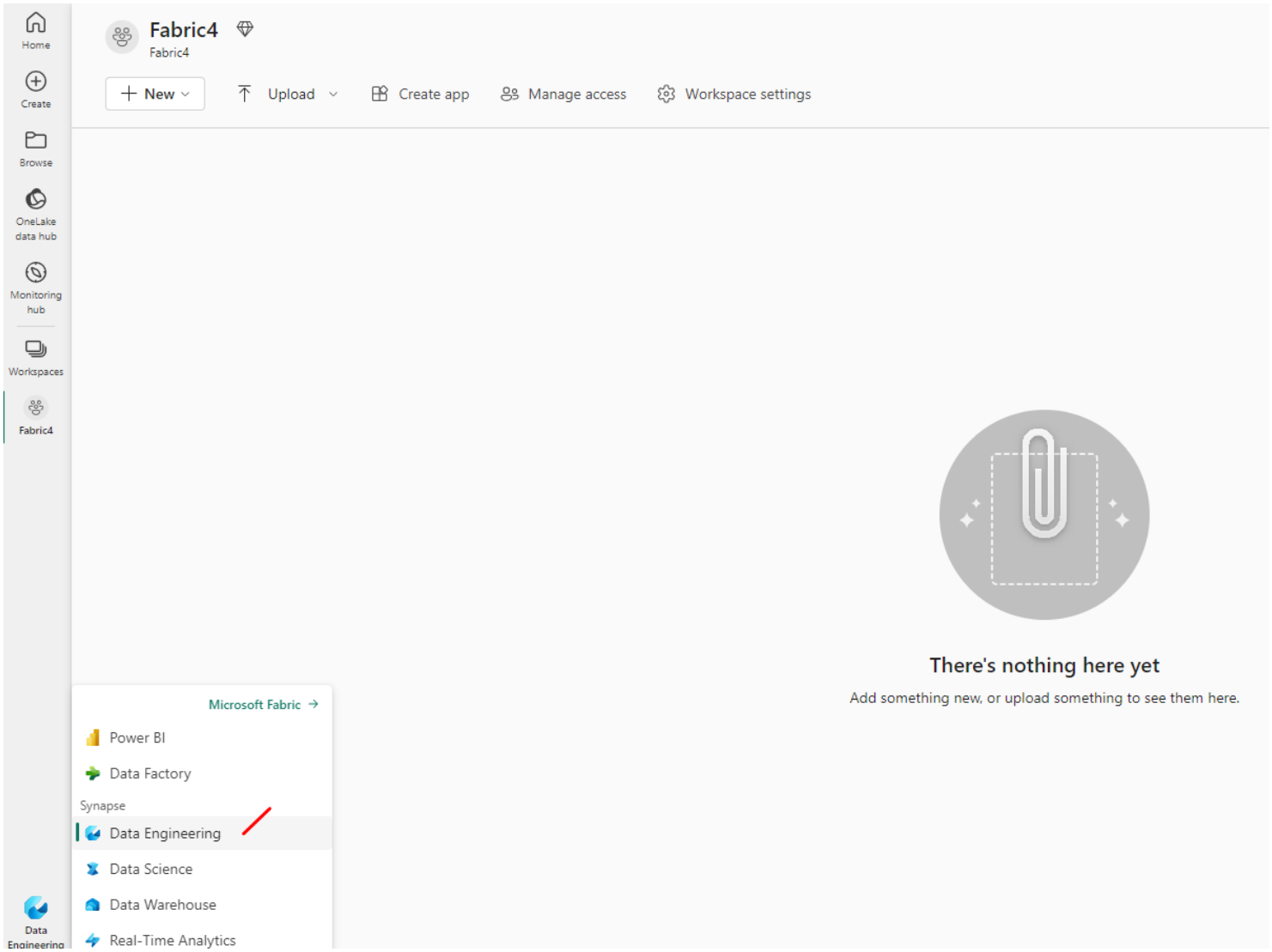


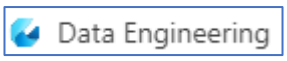

- Click 

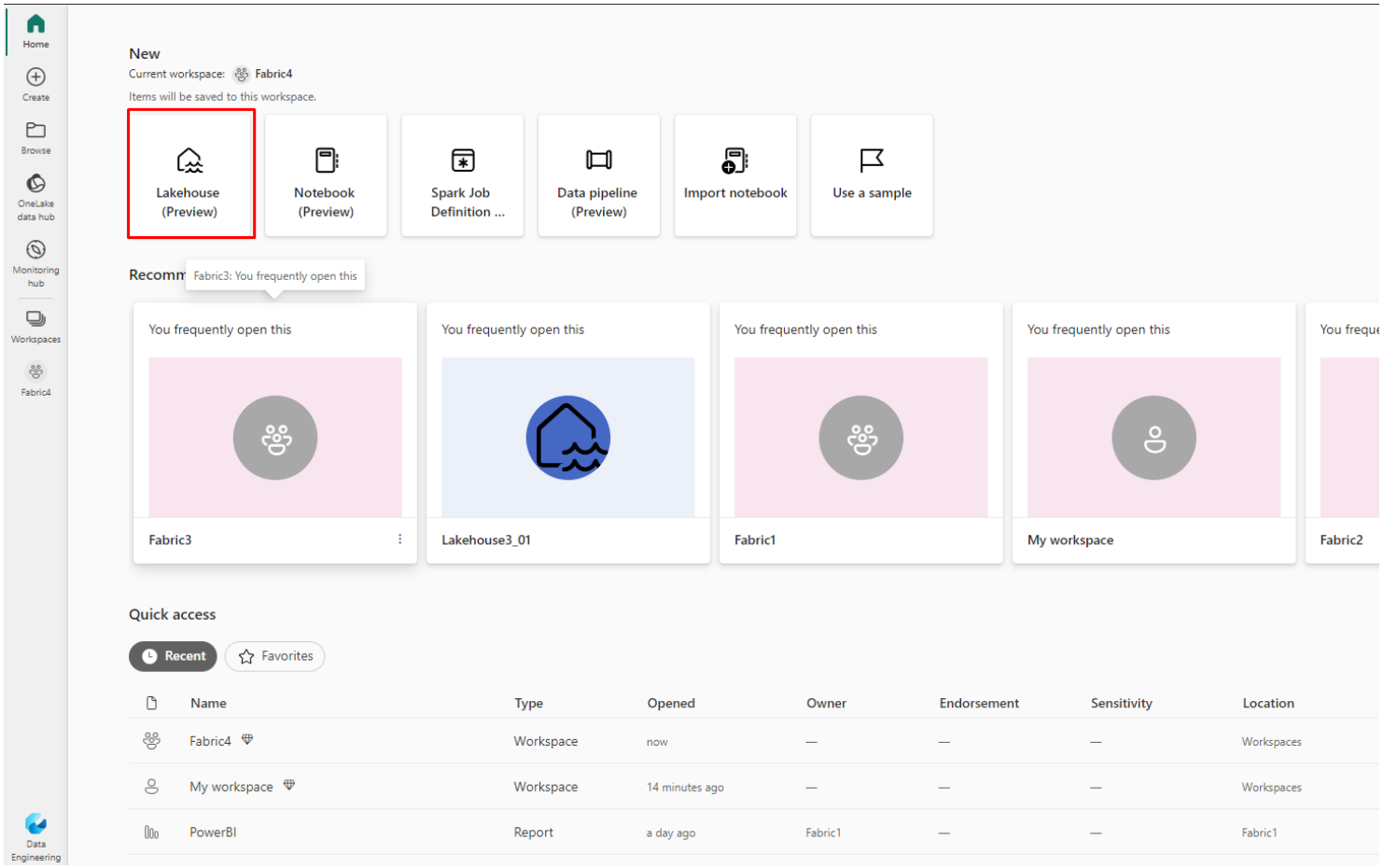



- After input workspace name
- Select **Trial**
- Click **Apply**

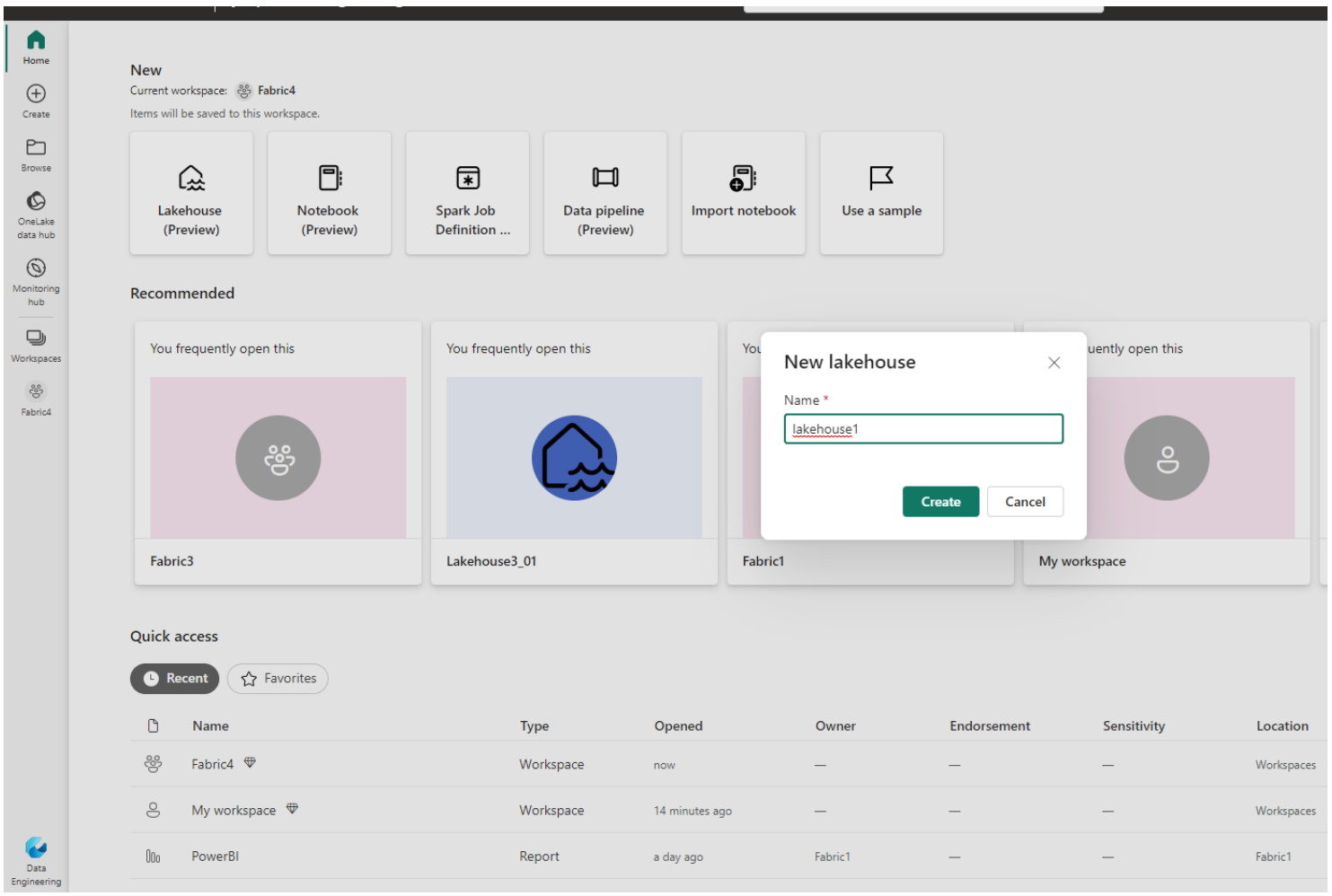
## 1.2 Adding Lakehouse




- Navigate to **Data Engineering** icon > click 



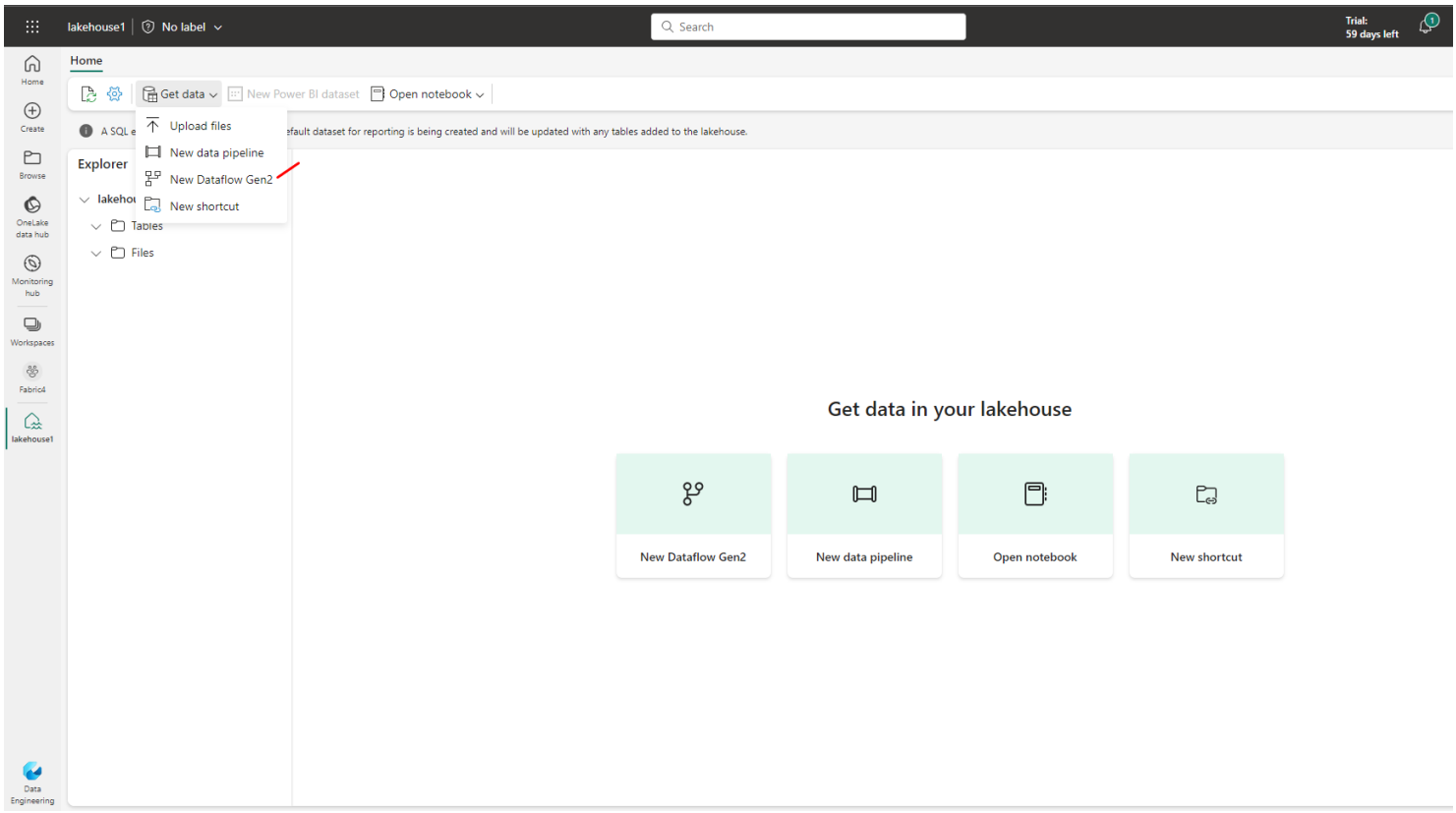

- Click in Lakehouse icon




- Input lakehouse name

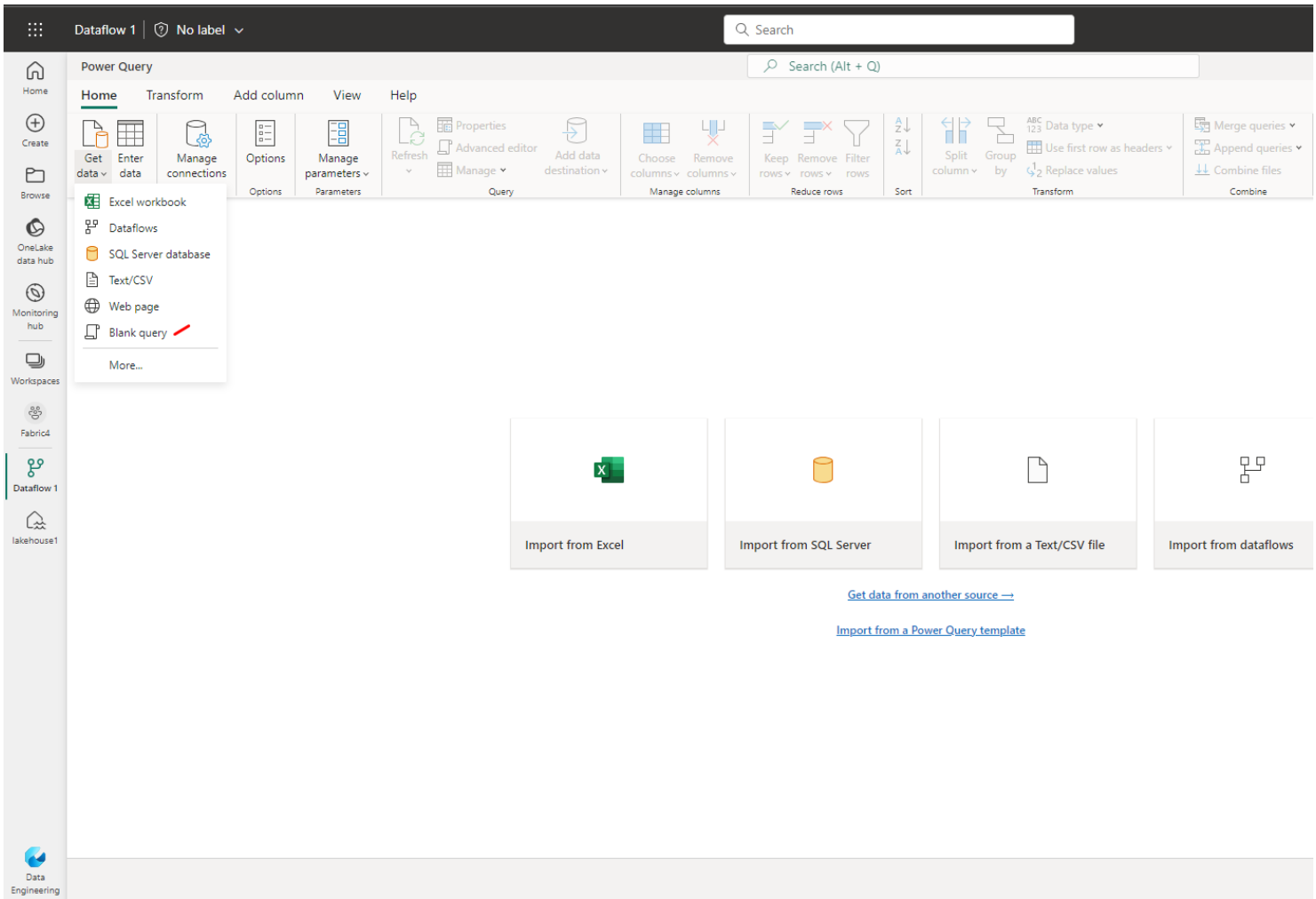
## 1.3 Dataflow Gen2

Dataflow Gen2 is for loading data from many sources. It looks like Power query in Power BI

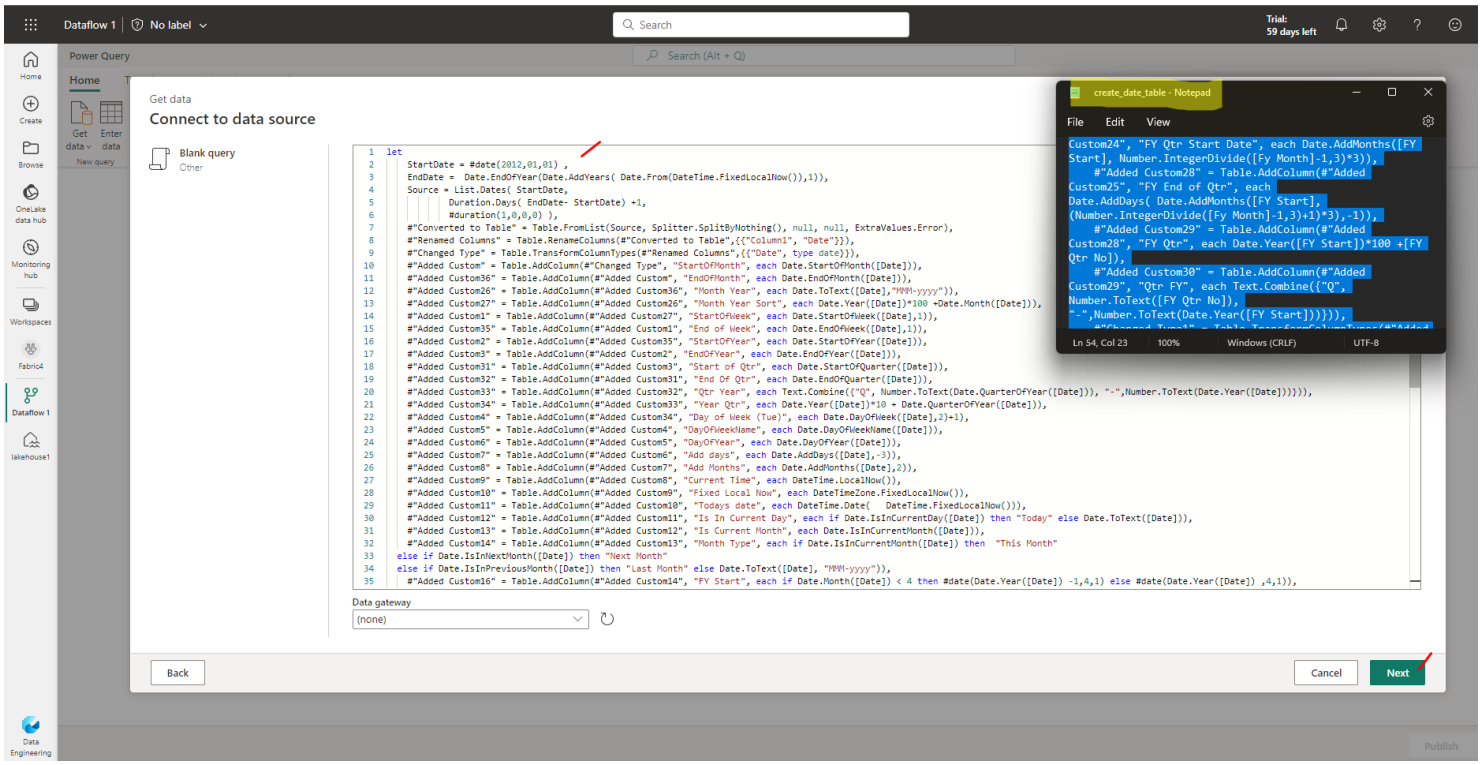



*Bank query*

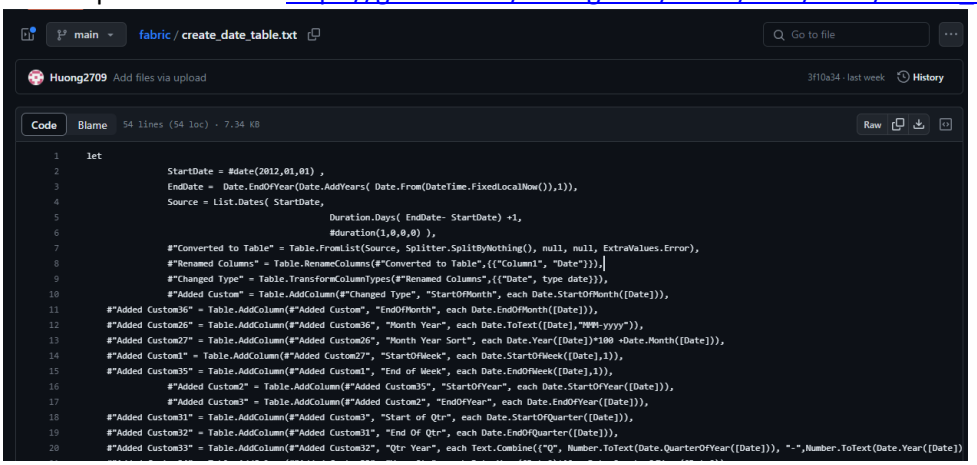
- Using Power Query M to add Date dimention table




- Select Blank query

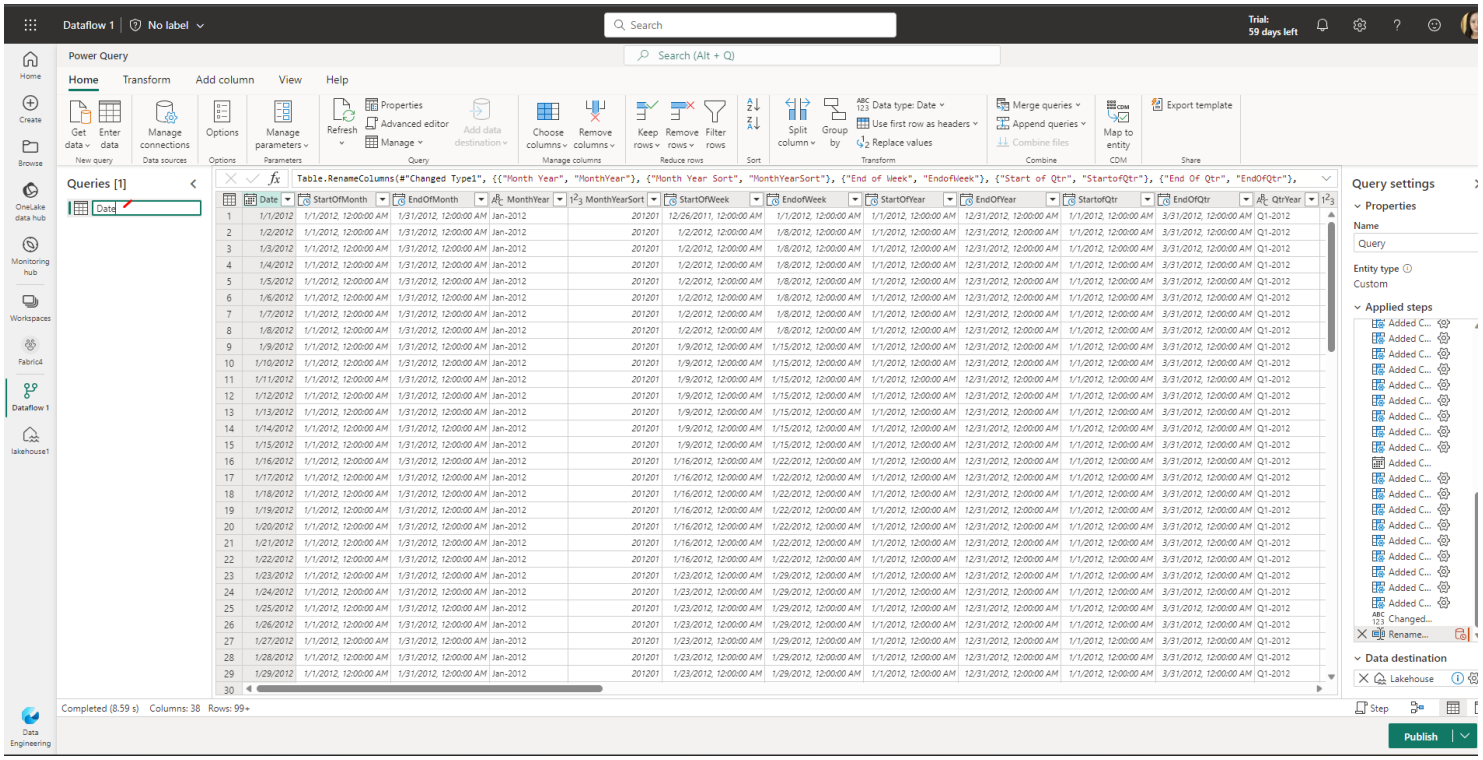



- Get Script from source: [https://github.com/Huong2709/fabric/blob/main/create_date_table.txt](https://github.com/Huong2709/fabric/blob/main/create_date_table.txt)

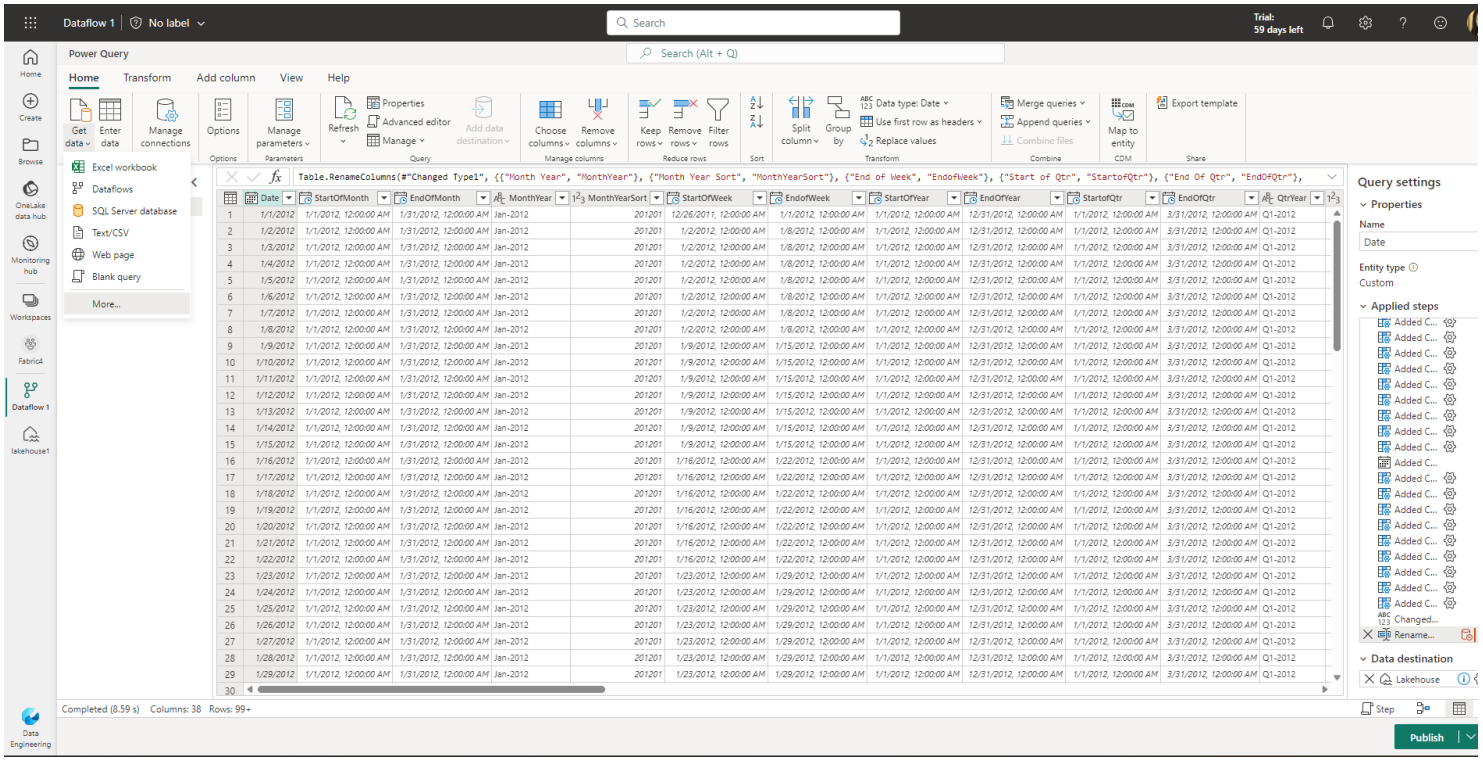



```
let
        StartDate = #date(2012,01,01) ,
        EndDate =  Date.EndOfYear(Date.AddYears( Date.From(DateTime.FixedLocalNow()),1)),
        Source = List.Dates( StartDate,
                             Duration.Days( EndDate- StartDate) +1,
                             #duration(1,0,0,0) ),
        #"Converted to Table" = Table.FromList(Source, Splitter.SplitByNothing(), null, null, ExtraValues.Error),
        #"Renamed Columns" = Table.RenameColumns(#"Converted to Table",{{"Column1", "Date"}}),
        #"Changed Type" = Table.TransformColumnTypes(#"Renamed Columns",{{"Date", type date}}),
        #"Added Custom" = Table.AddColumn(#"Changed Type", "StartOfMonth", each Date.StartOfMonth([Date])),
    #"Added Custom36" = Table.AddColumn(#"Added Custom", "EndOfMonth", each Date.EndOfMonth([Date])),
    #"Added Custom26" = Table.AddColumn(#"Added Custom36", "Month Year", each Date.ToText([Date],"MMM-yyyy")),
    #"Added Custom27" = Table.AddColumn(#"Added Custom26", "Month Year Sort", each Date.Year([Date])*100 +Date.Month([Date])),
    #"Added Custom1" = Table.AddColumn(#"Added Custom27", "StartOfWeek", each Date.StartOfWeek([Date],1)),
    #"Added Custom35" = Table.AddColumn(#"Added Custom1", "End of Week", each Date.EndOfWeek([Date],1)),
        #"Added Custom2" = Table.AddColumn(#"Added Custom35", "StartOfYear", each Date.StartOfYear([Date])),
        #"Added Custom3" = Table.AddColumn(#"Added Custom2", "EndOfYear", each Date.EndOfYear([Date])),
    #"Added Custom31" = Table.AddColumn(#"Added Custom3", "Start of Qtr", each Date.StartOfQuarter([Date])),
    #"Added Custom32" = Table.AddColumn(#"Added Custom31", "End Of Qtr", each Date.EndOfQuarter([Date])),
    #"Added Custom33" = Table.AddColumn(#"Added Custom32", "Qtr Year", each Text.Combine({"Q", Number.ToText(Date.QuarterOfYear([Date])), "-",Number.ToText(Date.Year([Date])
```

- Input script in M language

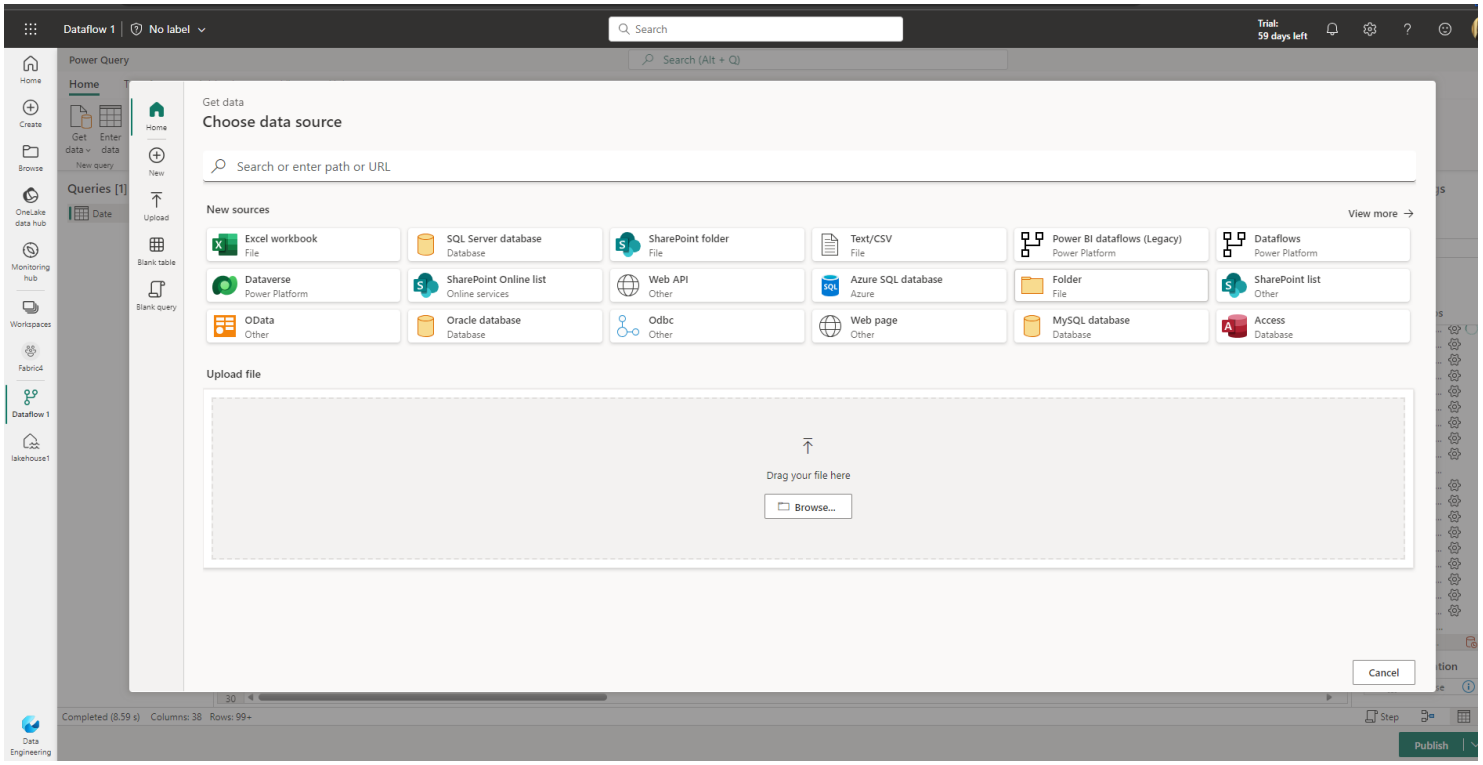


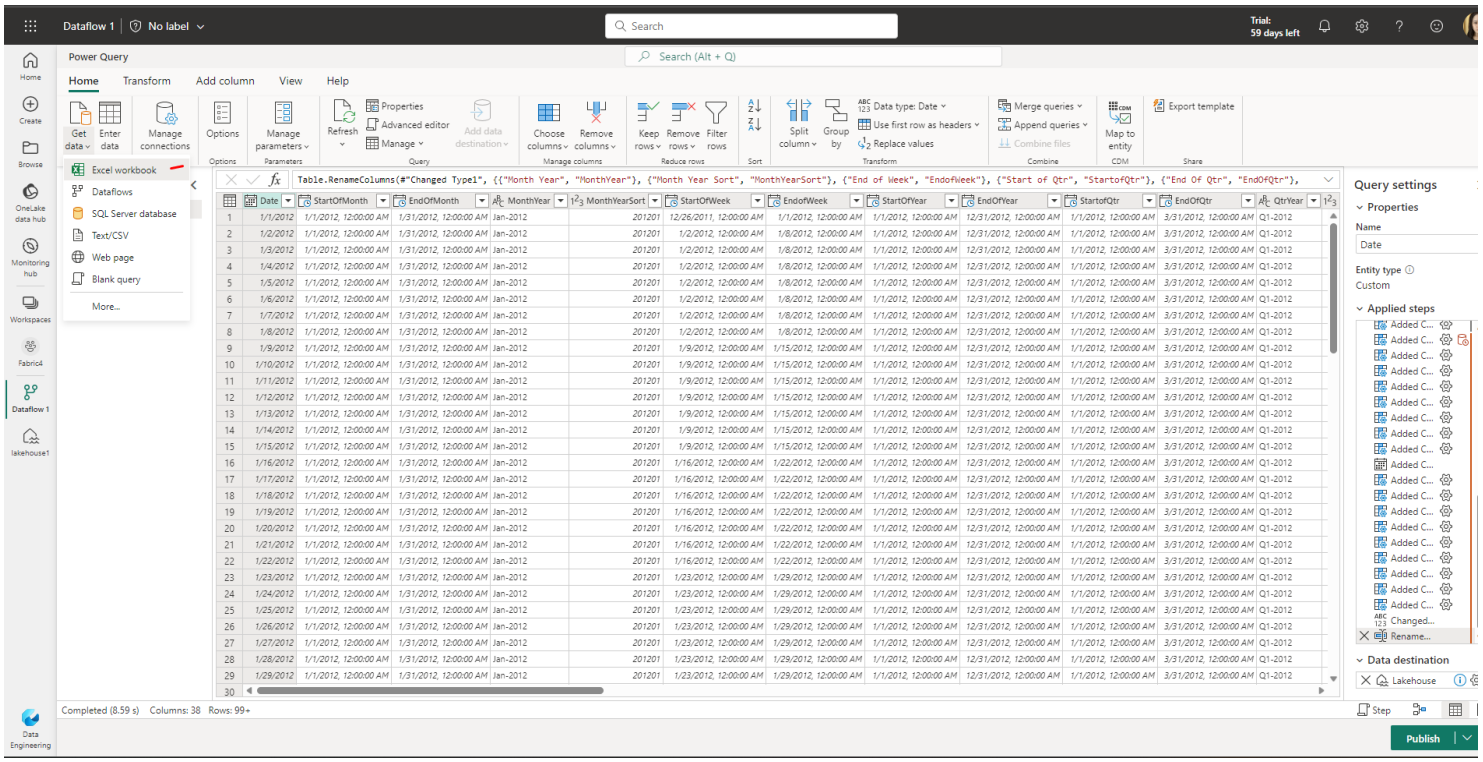
- Rename the table after creation if need



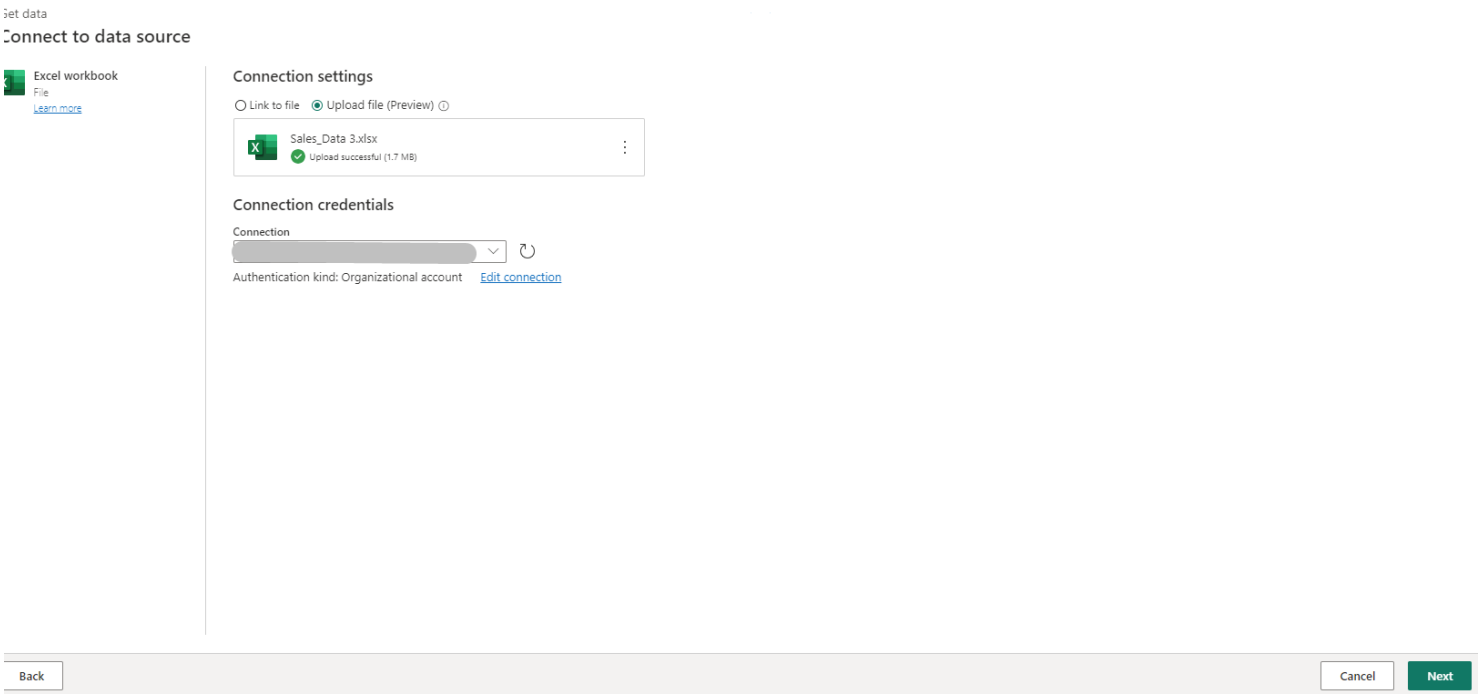
- Get data > More : to add data from other sources

## *Loading files from other sources*

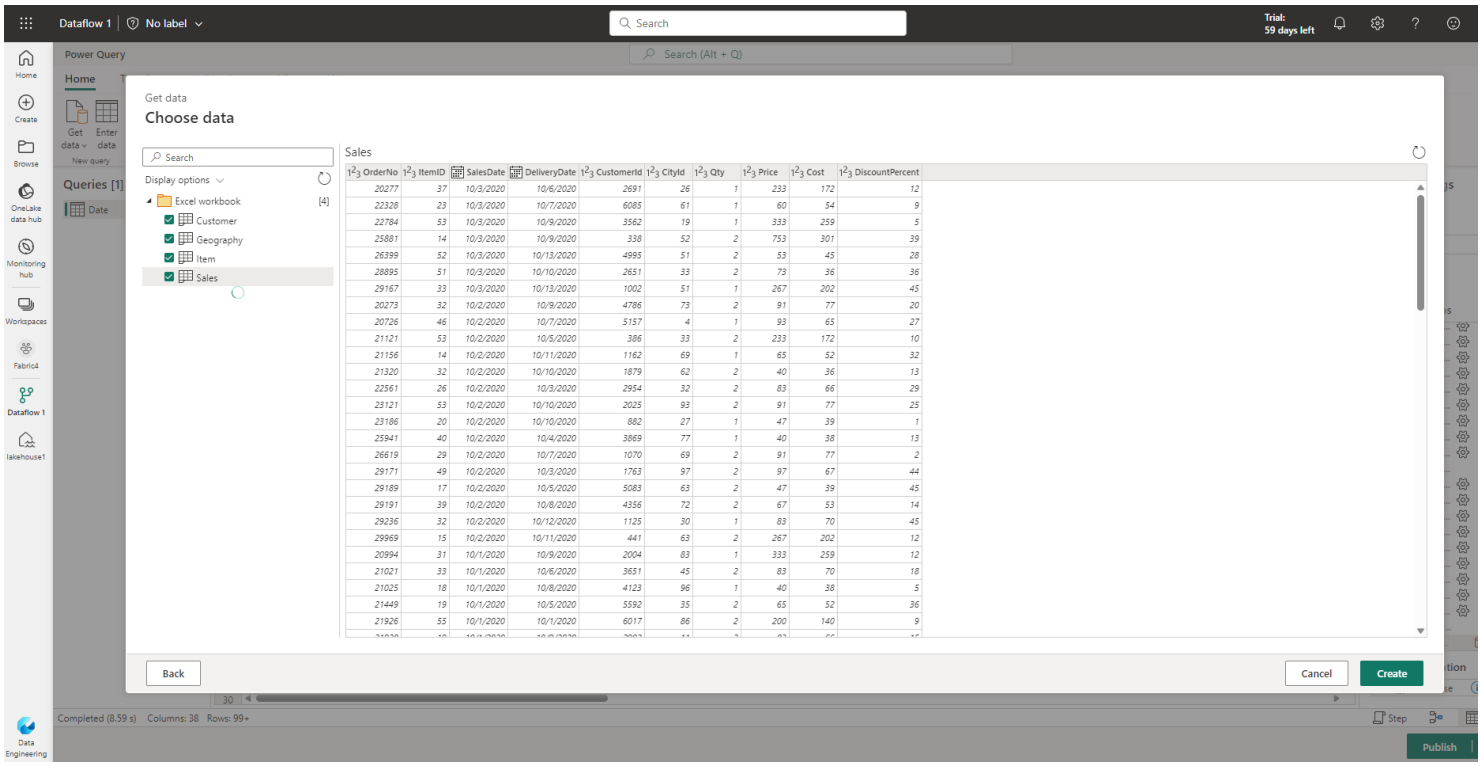


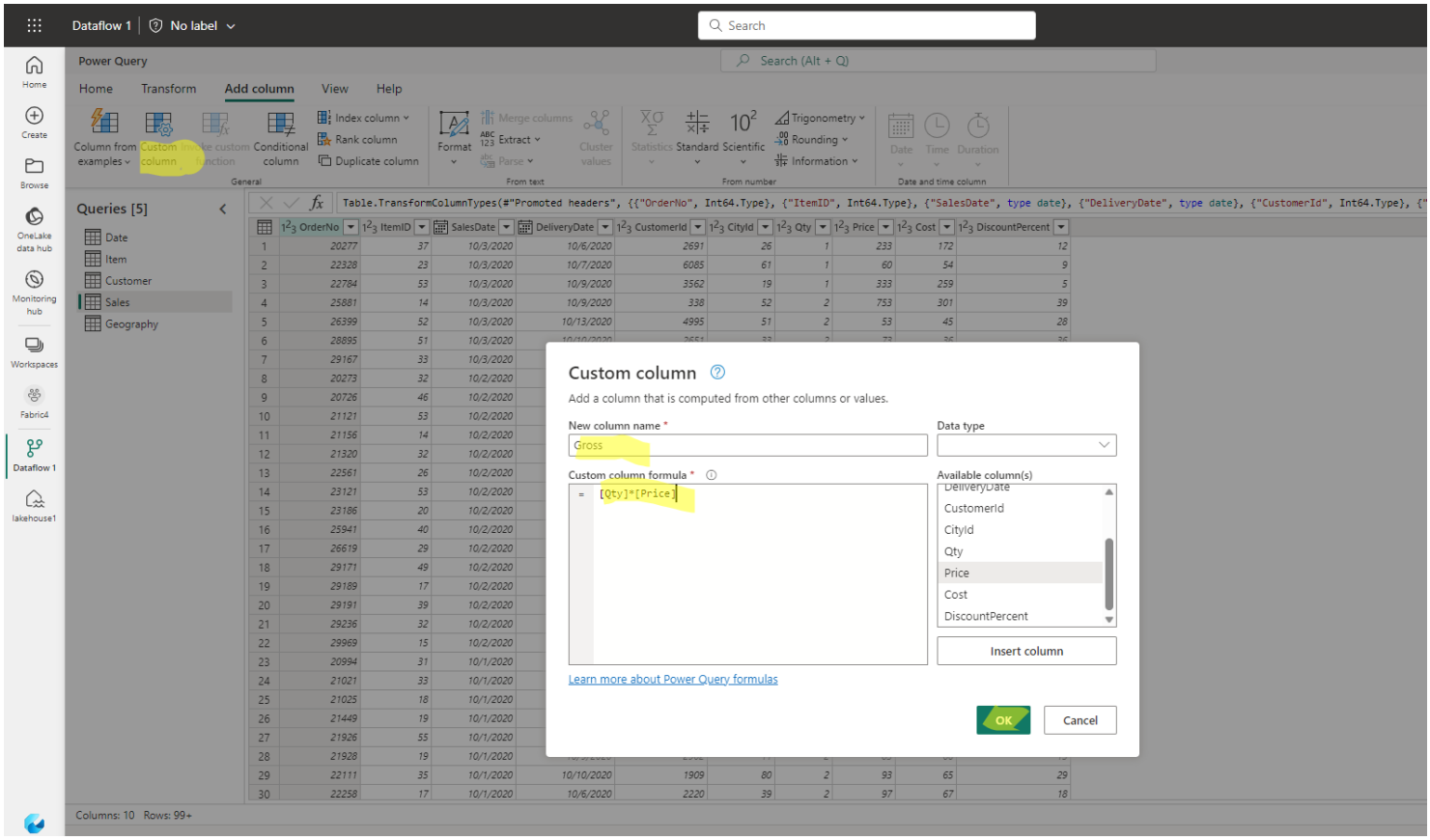

- Let's test with excel files
- Previously I can upload excel files from github, but recently I recently I have some issue with connection file by links. If you have same issue, please try to download the file to your computer and upload manually
- In this example, you can get the file from : https://github.com/Huong2709/fabric/blob/main/Sales_Data.xlsx




- We can upload files by using link or Upload from local file locations
- Select Connection settings
- Upload file or adding link to load files
- Click next



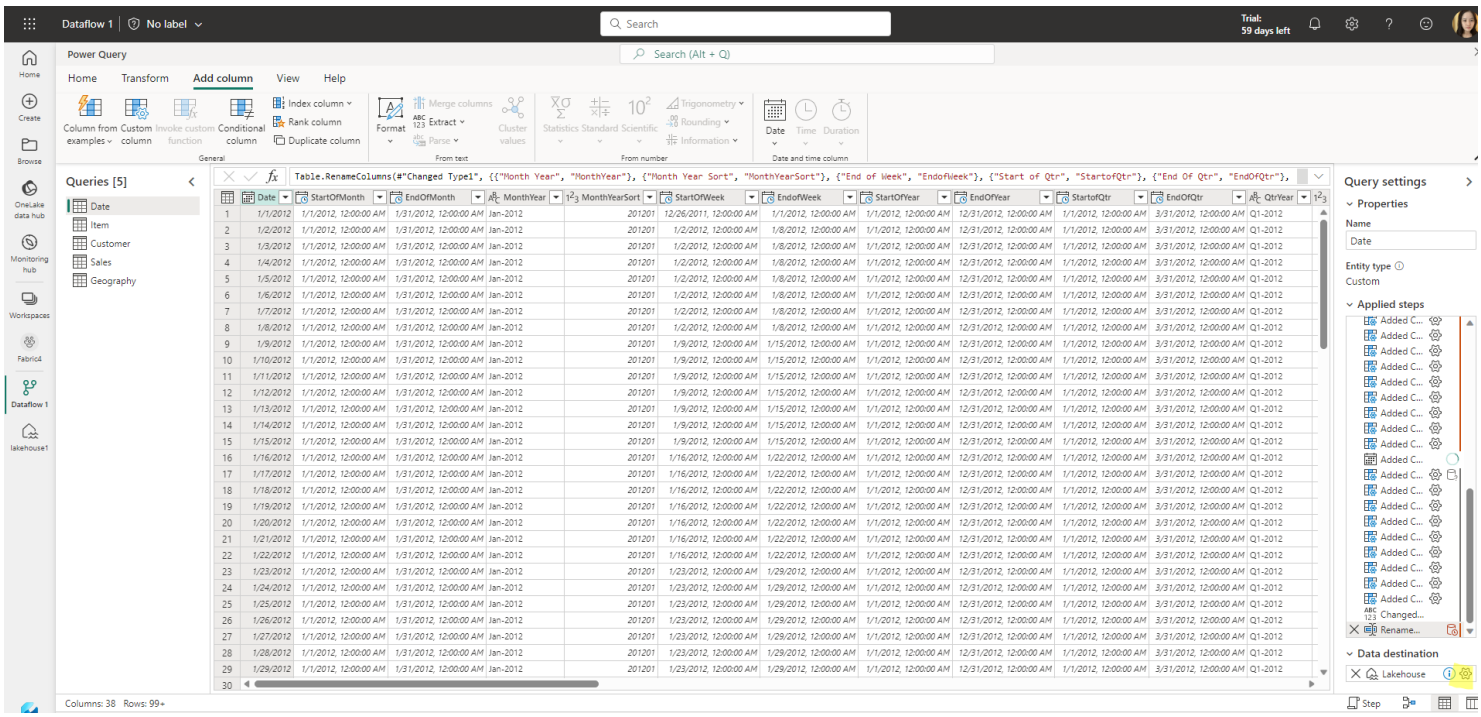

- Select related sheet in excel files what you want to upload as data source
- Click **Create**

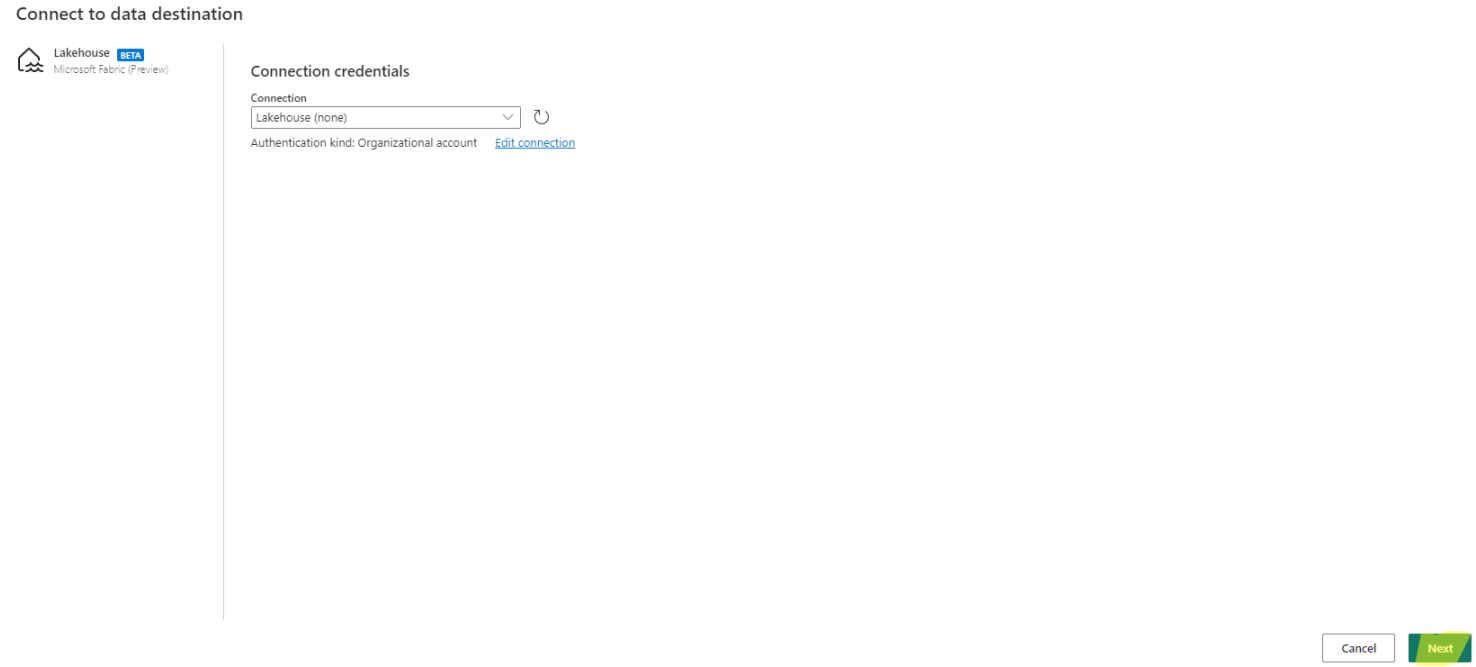



Similar as in Power BI, we can so similar activities in MS Fabric

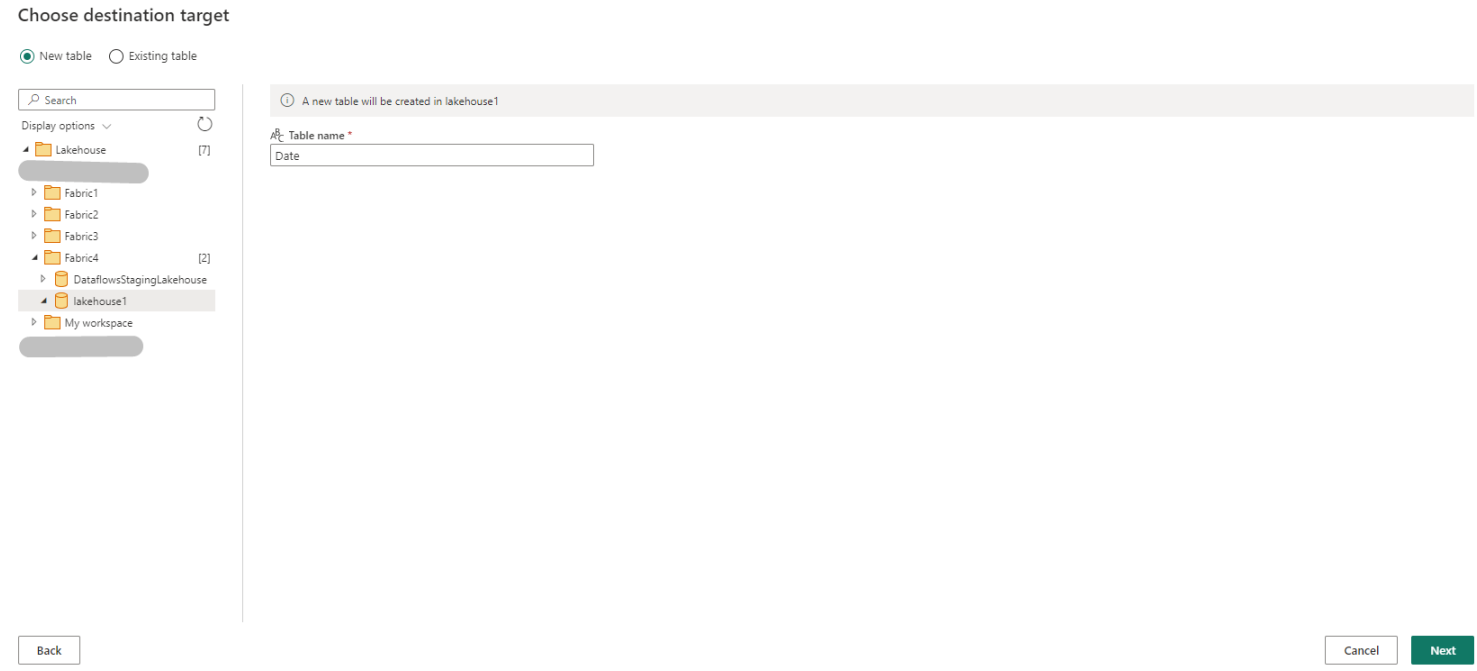
Let’s try to add 1 column as an example

- Click Customer column in action pane
- Input “New column name:
- Input Custom column formula
- Click OK

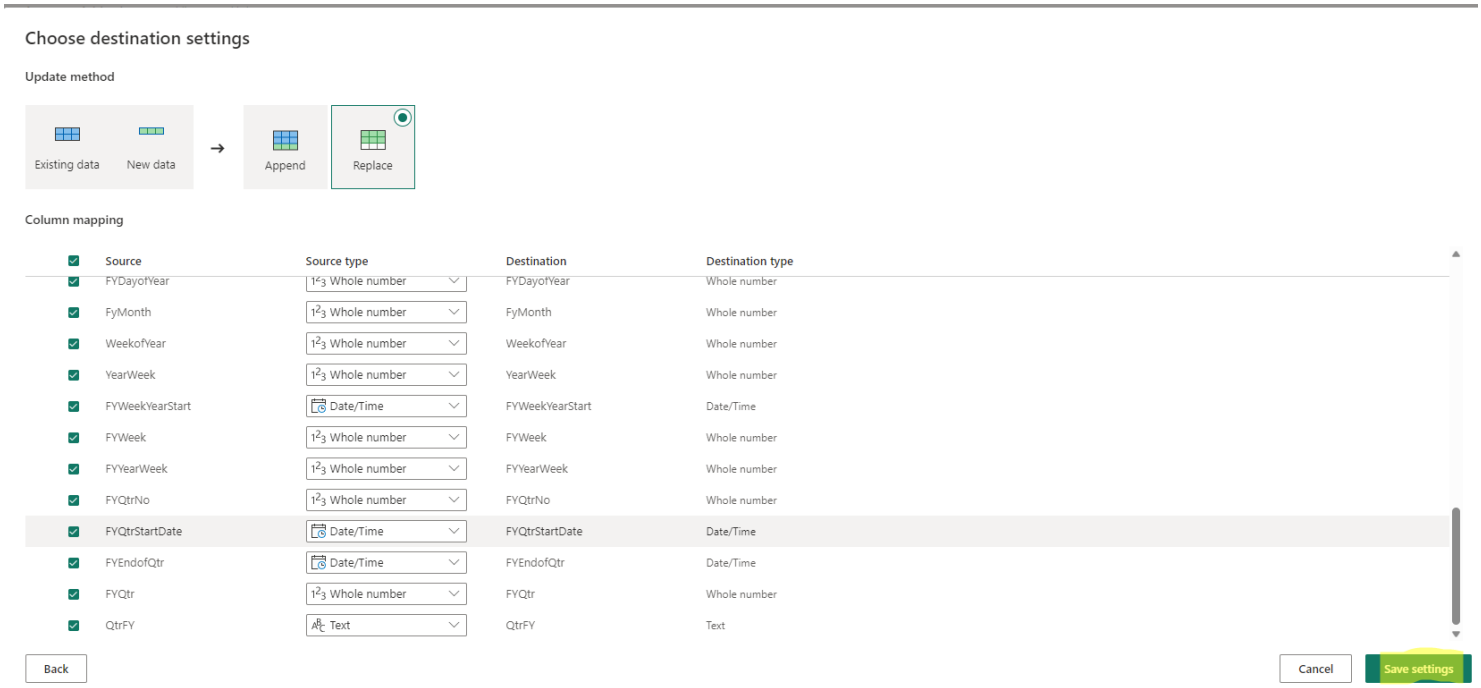




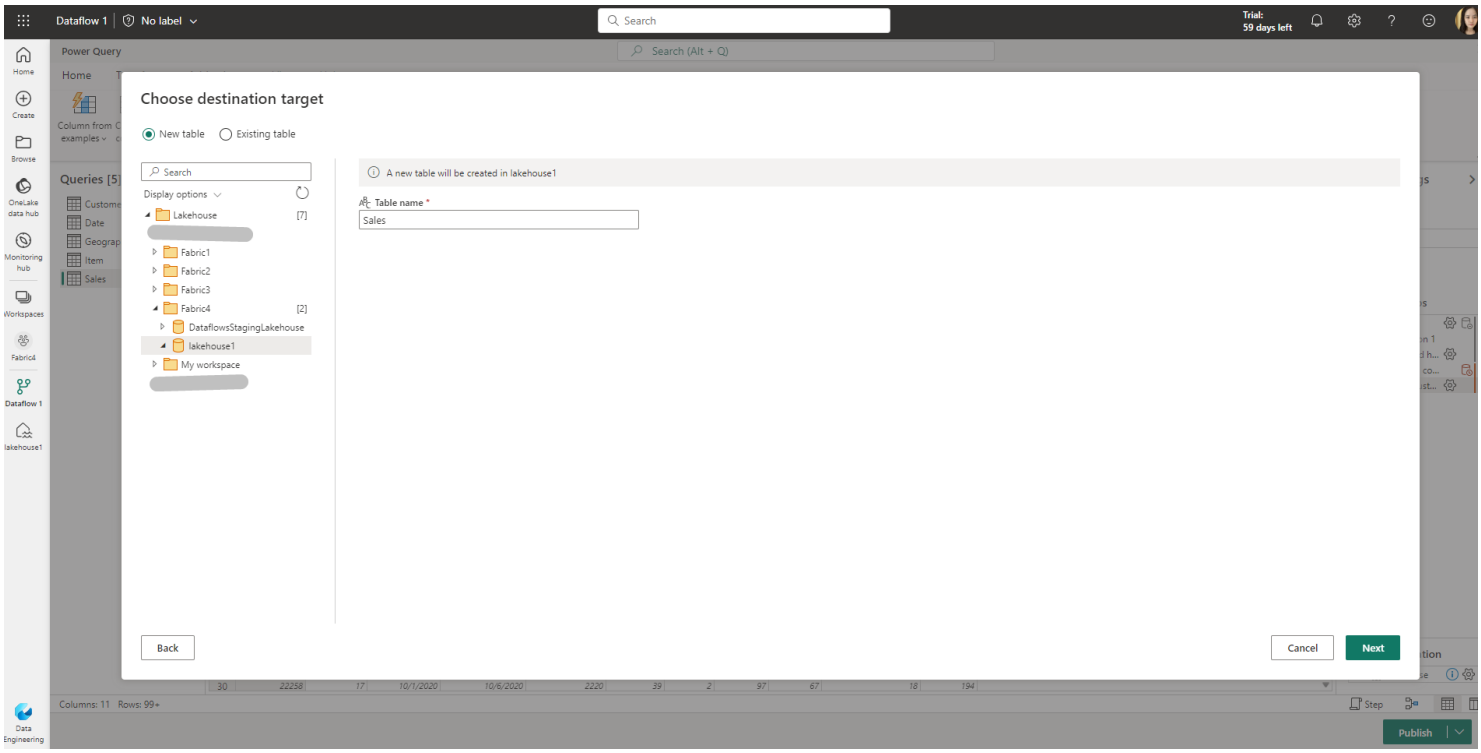

Select Lakehouse connection




- In Display option, navigate to Workspace > lakehouse that we want to load data into
- Click next

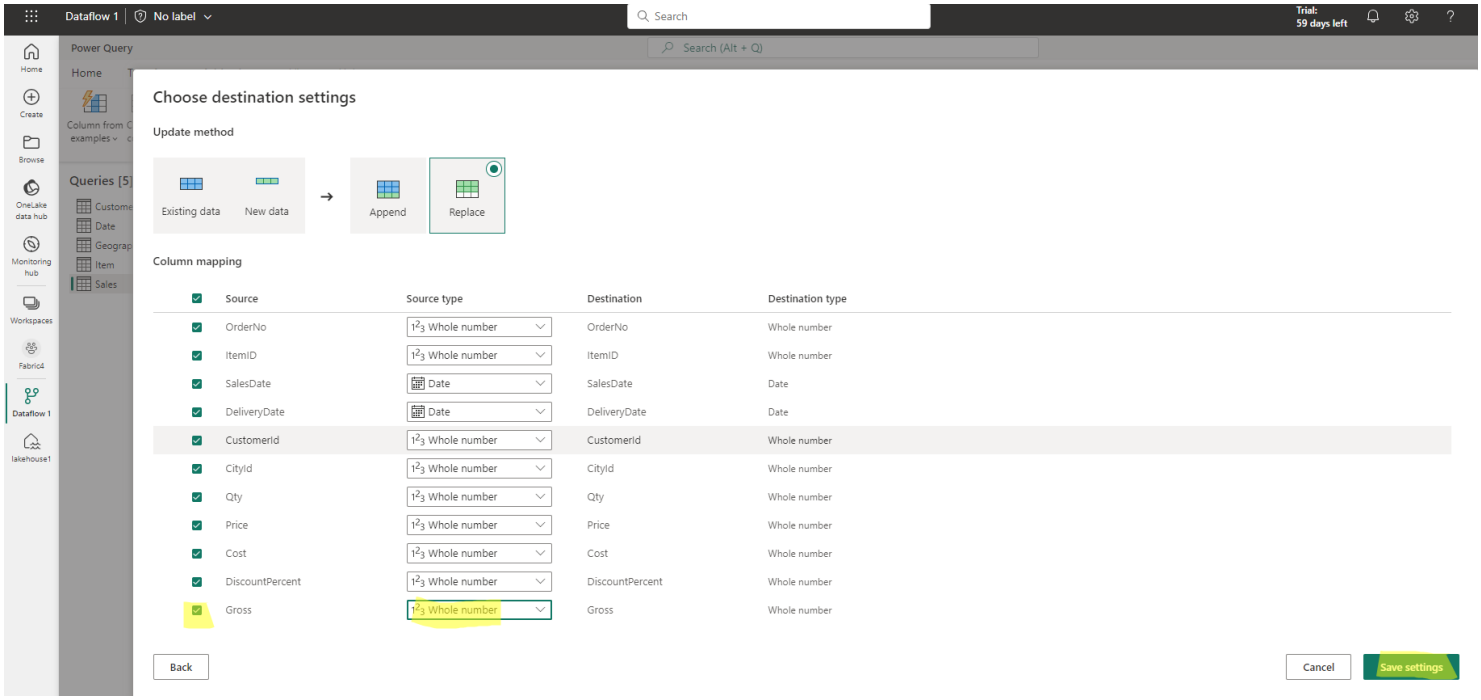



- Drill up and down to check Column mapping details, source type then **Save setting**

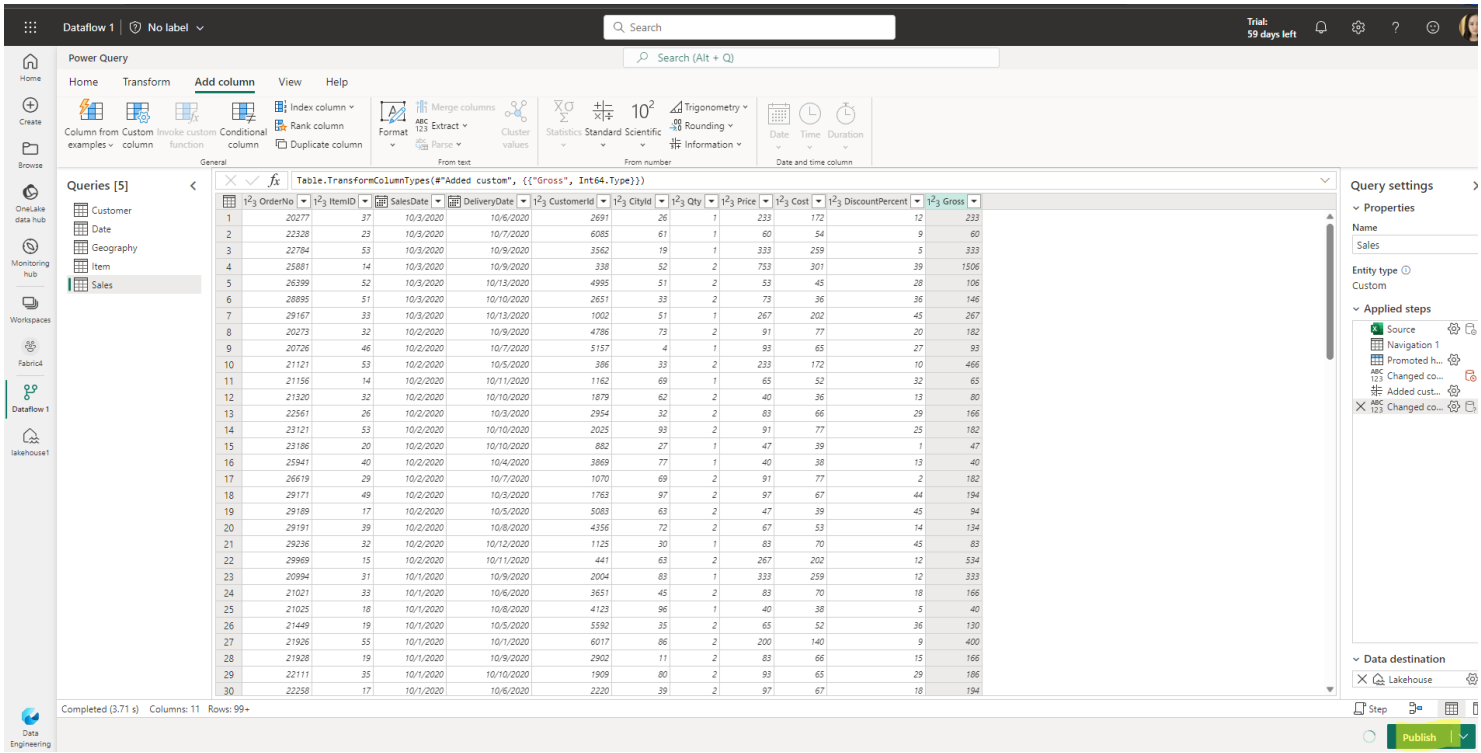



Continuing with sales table

- Navigate to Workspace > lakehouse that we want to load data into




- For custom column that we created in power query, It doesn't pop up data type and select column mapping. Make sure to double check before "Save setting" else you can't see it in Lakehouse after publish

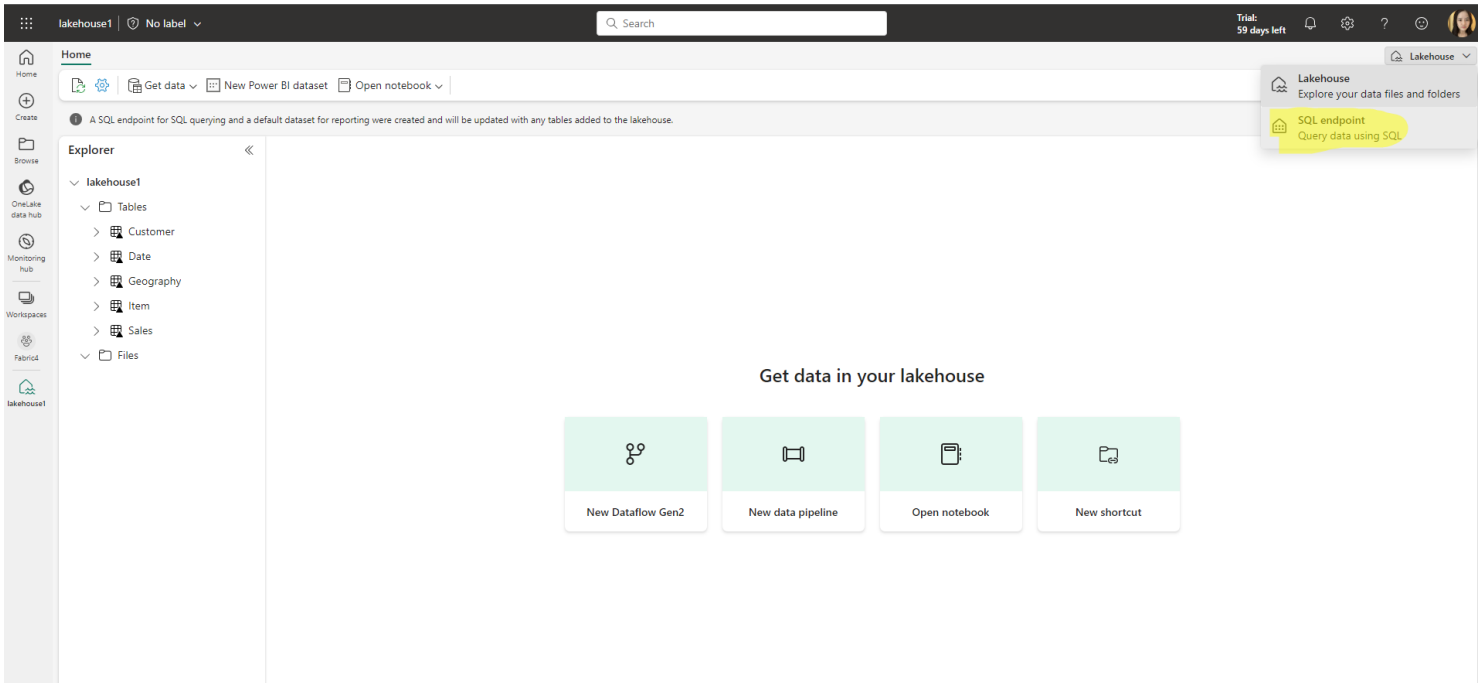



>> After publish for sales table, please continue to setting Lakehouse and publish for other tables (Customer, Geography, Item)

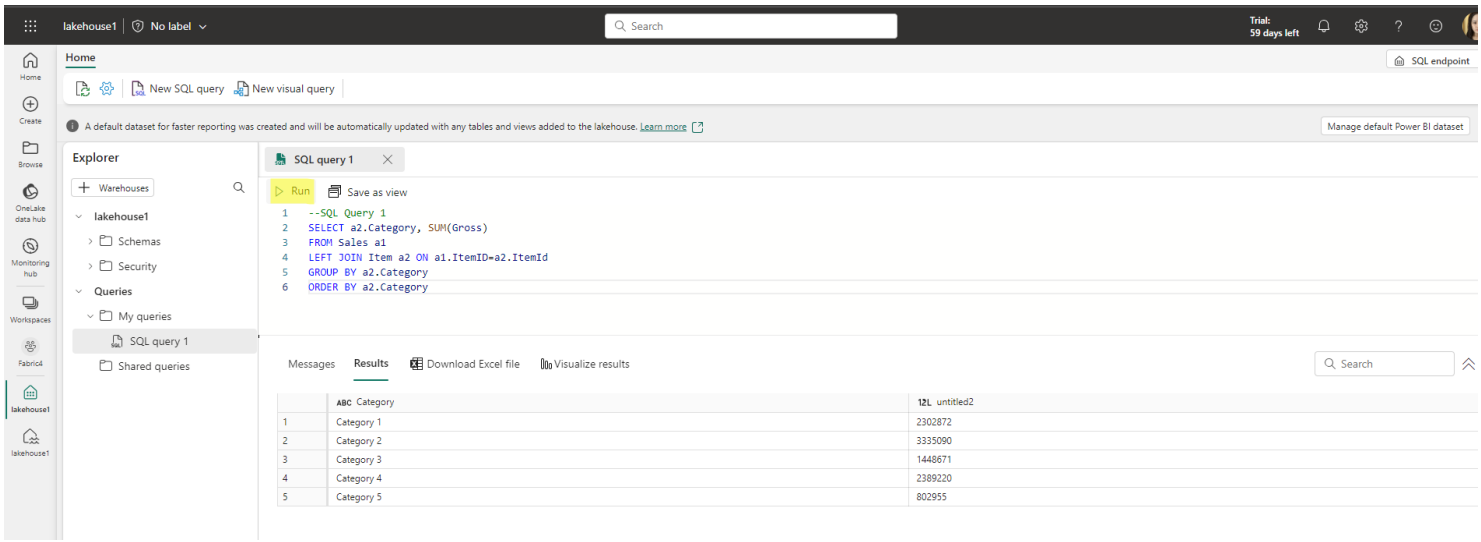
## 1.4 Using SQL endpoint

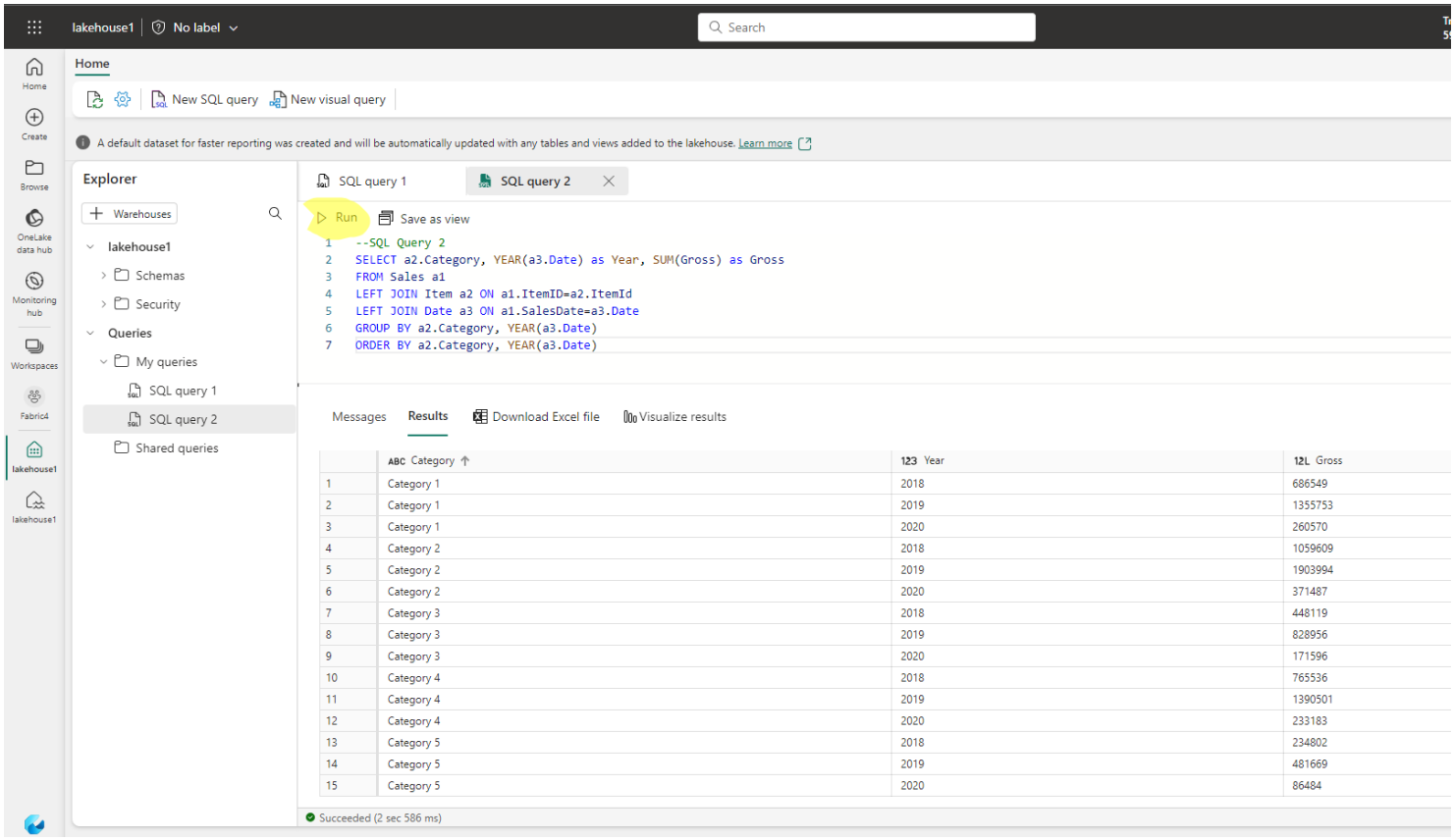
After adding at least 1 table in lakehouse, SQL is automatically enable for using

SQL endpoint operates in read-only mode over lakehouse delta tables




- From Lakehouse screen > Click **Lakehouse** icon> **Click SQL endpoint**

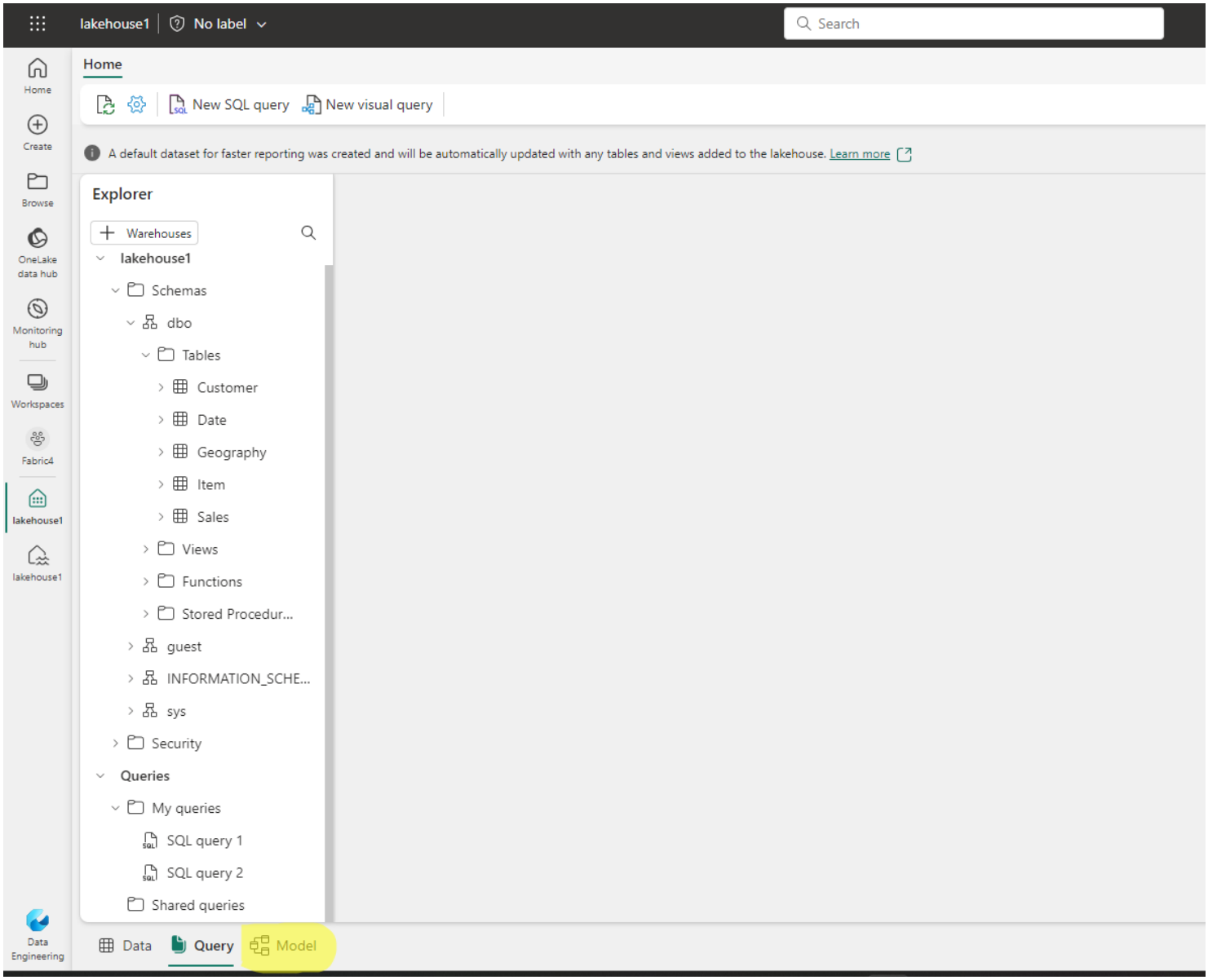




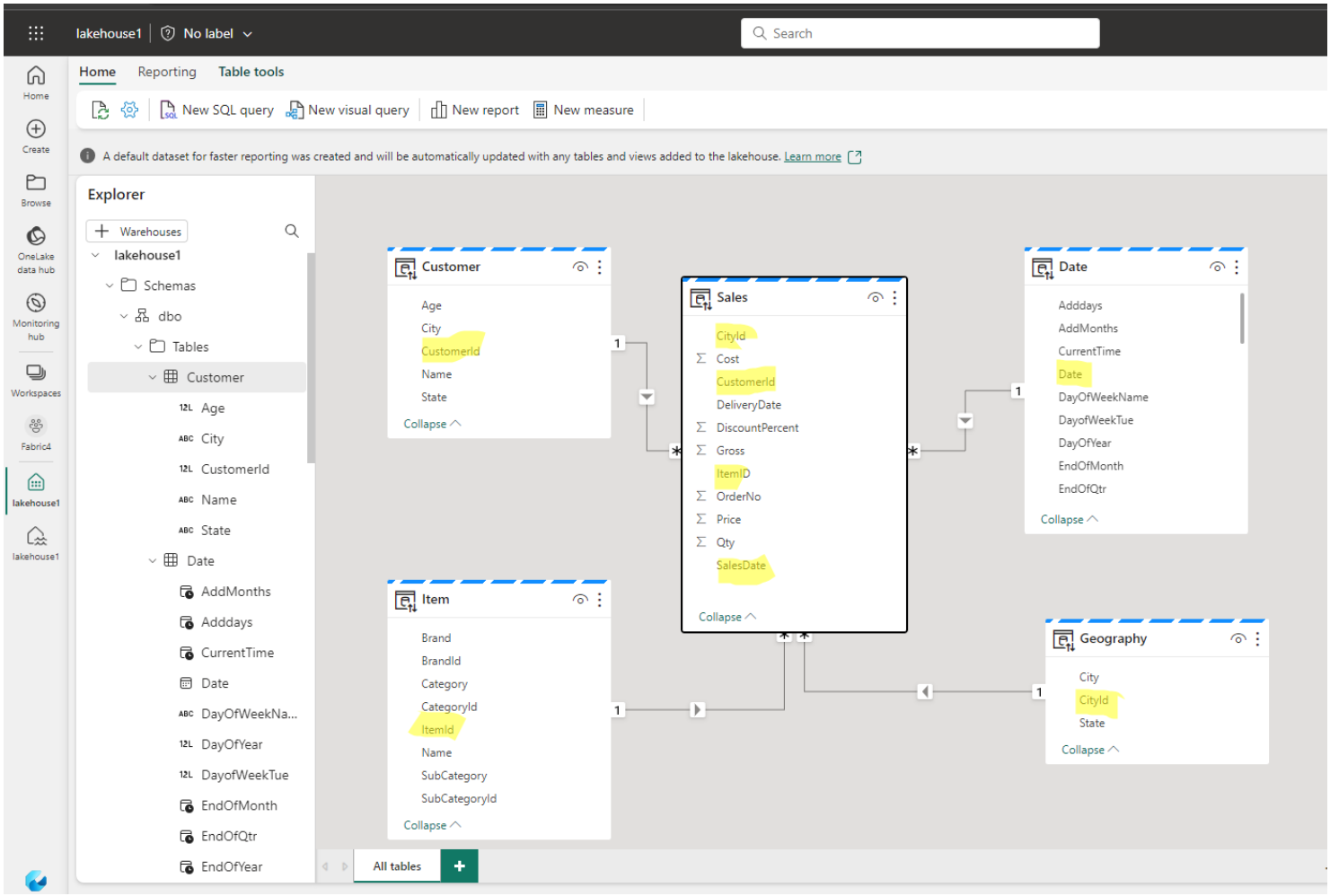

| | Category | Year | Gross |
|---|---|---|---|
| 1 | Category 1 | 2018 | 686549 |
| 2 | Category 1 | 2019 | 1355753 |
| 3 | Category 1 | 2020 | 260570 |
| 4 | Category 2 | 2018 | 1059609 |
| 5 | Category 2 | 2019 | 1903994 |
| 6 | Category 2 | 2020 | 371487 |
| 7 | Category 3 | 2018 | 448119 |
| 8 | Category 3 | 2019 | 828956 |
| 9 | Category 3 | 2020 | 171596 |
| 10 | Category 4 | 2018 | 765536 |
| 11 | Category 4 | 2019 | 1390501 |
| 12 | Category 4 | 2020 | 233183 |
| 13 | Category 5 | 2018 | 234802 |
| 14 | Category 5 | 2019 | 481669 |
| 15 | Category 5 | 2020 | 86484 |

## 1.5 Create model

Model is for creation connections between tables so we can create reports that getting data from many tables




Drag and drop between tables to create relationships base on key

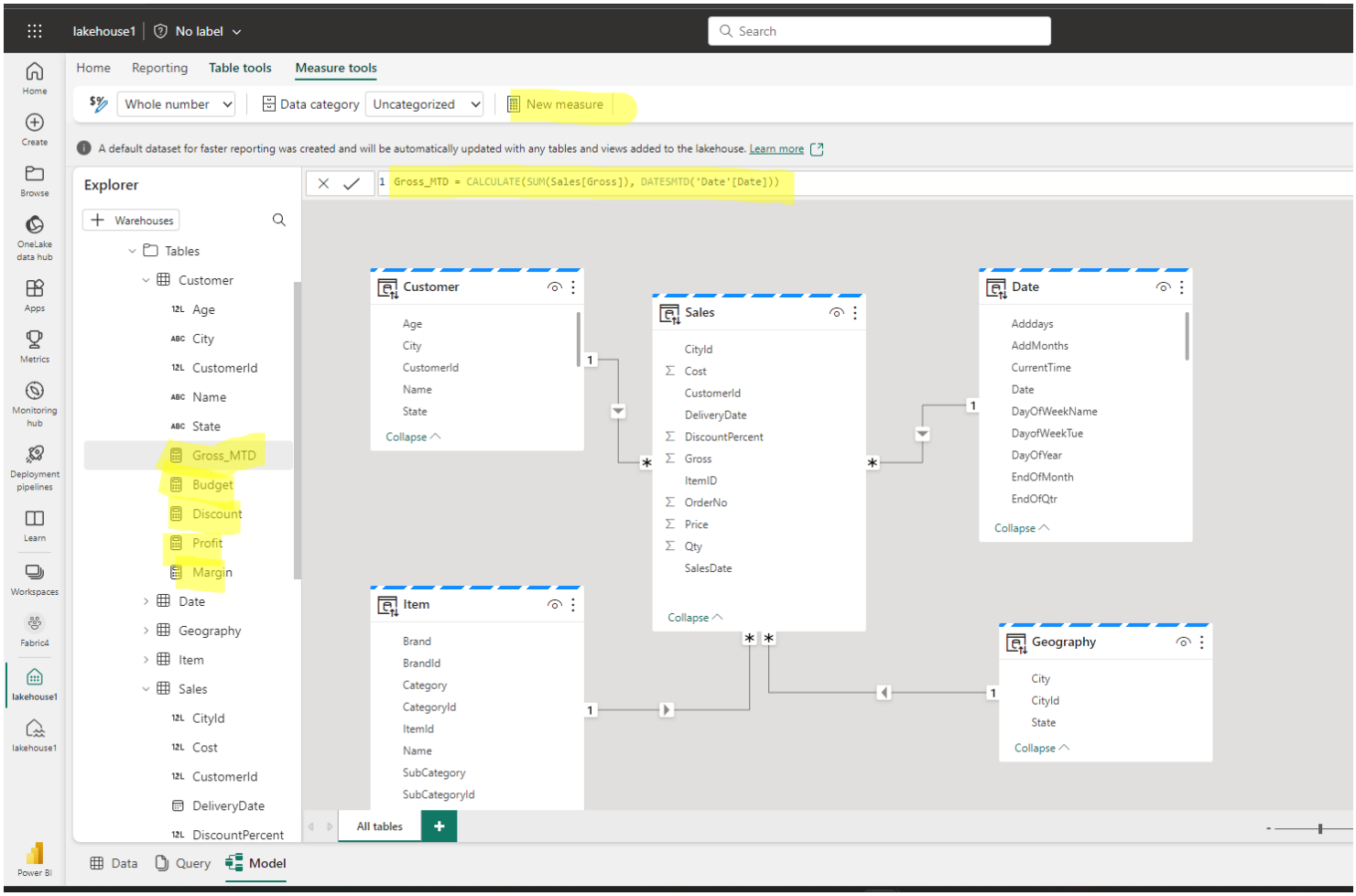



Sales table is including transaction. It's fact table, data is updated frequently. Let's put it in the central of model for easy viewing while creation relationships

Customer, item, date, Geography are dimension tables (Dim)

## 1.6 Create measure for reports

Measures are used in some of the most common data analyses. Simple summarizations such as sums, averages, minimum, maximum and counts can be set through the Fields well.



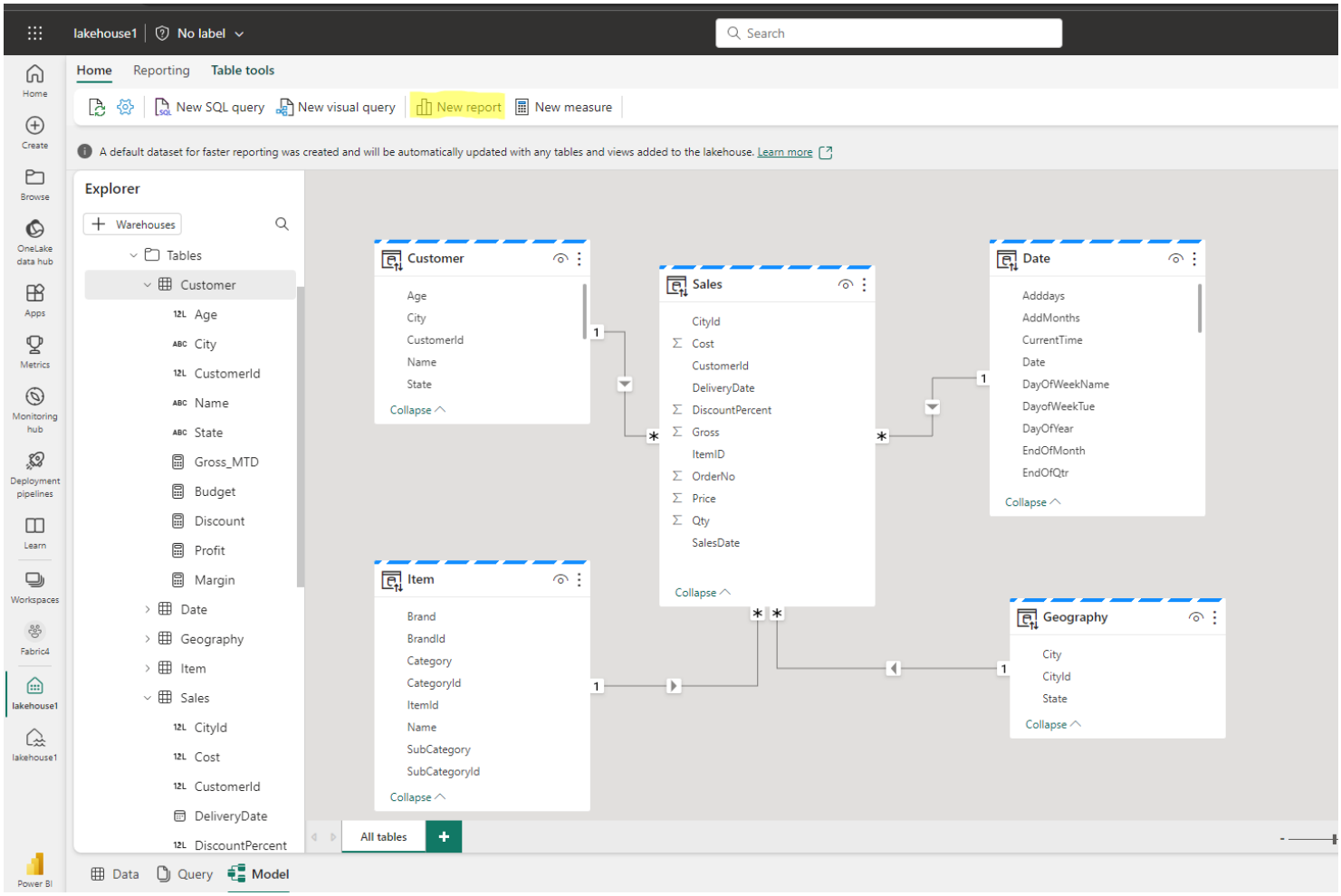

**Sample script for testing measure:**

```
Gross_MTD = CALCULATE(SUM(Sales[Gross]), DATESMTD('Date'[Date]))
Budget = SUMX(Sales, [Price] * [Qty])
Discount = SUMX(Sales, [Gross] * [DiscountPercent]/100)
Profit = SUMX(Sales, [Gross] - [Discount] - [Budget])
Margin = DIVIDE([Profit], SUMX(Sales, [Gross]))
```

- Click New measure
- Input formular for measure (wait 5-10 seconds for Fabric saving formular, else it will reloading the page)

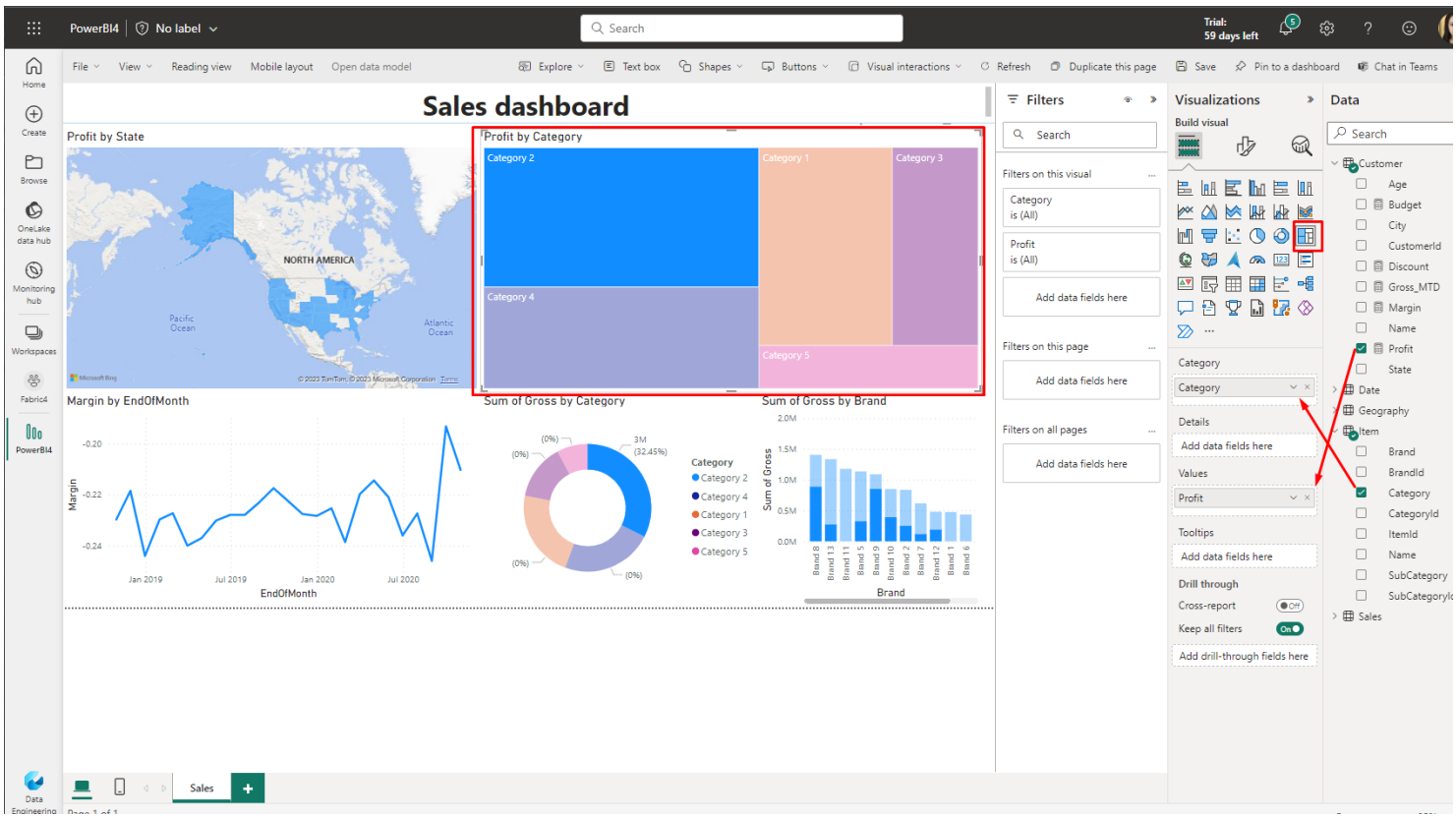
## 1.7 Create Power BI reports

We can open report design platform like power BI in MS fabric




- Click **New report** button

This is our result after drag and drop fields

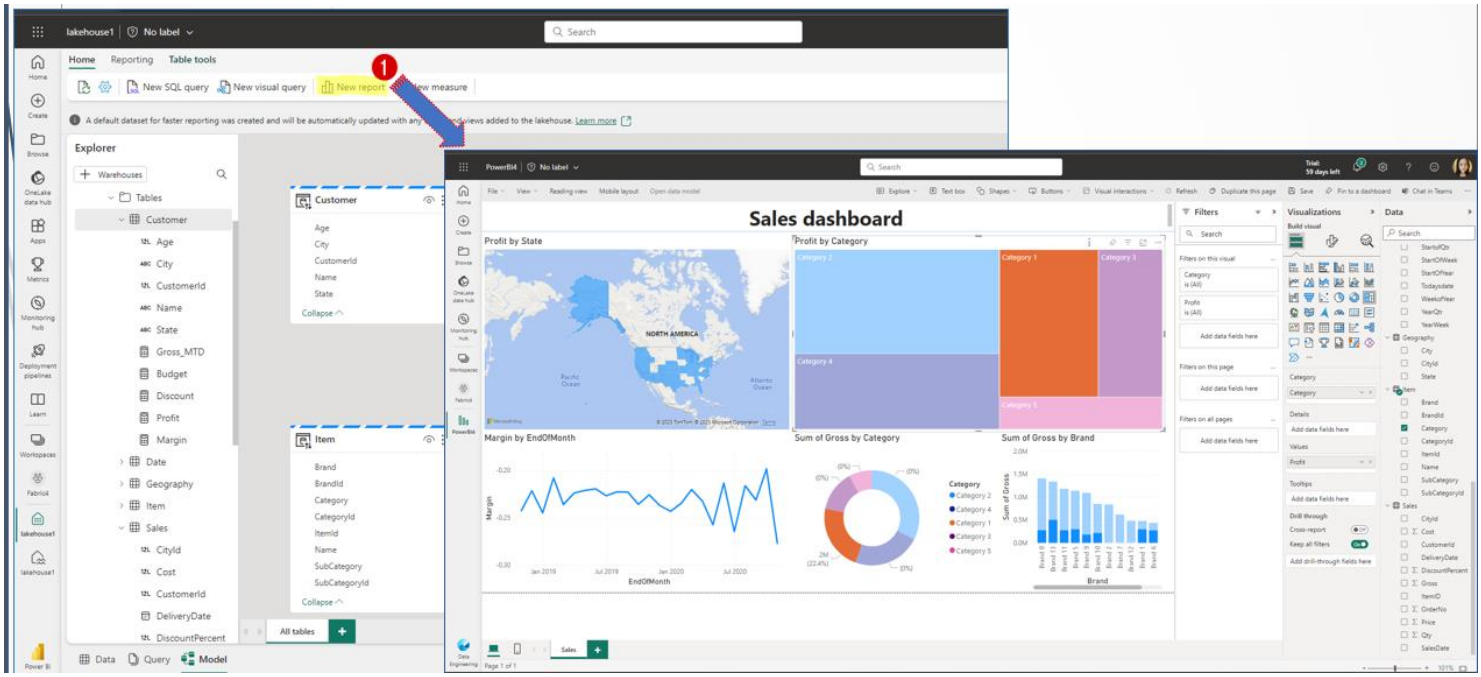



1. Select chart in Visualizations area

2. Drag and drop relate data (Category , Profit)

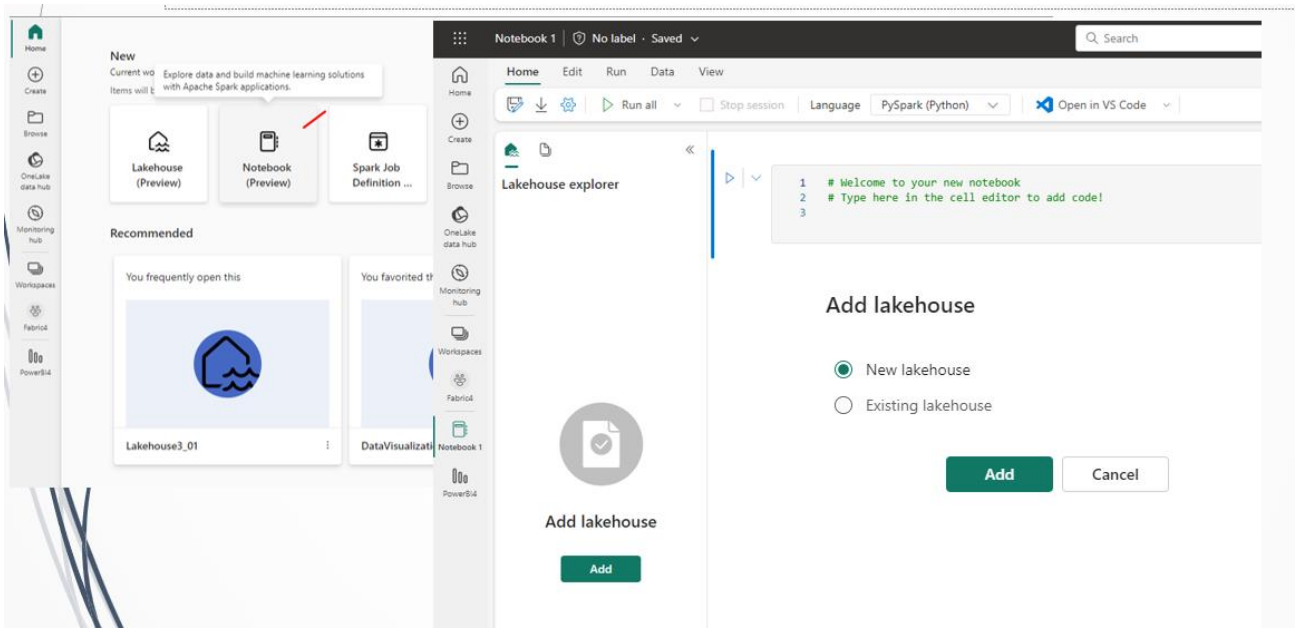
>> Yeah, we can a create report. This can be done easily even we haven't learnt power BI. You can try creating other reports for experience how cool MS fabric is

**Report**




## 2.Notebook – Warehouse- pipeline

### 2.1 Create Notebook and LakeHouse from Workspace




- Select workspace from MS Fabric home page

- Click Notebook icon

>> Wait some seconds till system open next from. I didn't know it so I keep clicking many times till more than 4 notebooks were created in my workspace

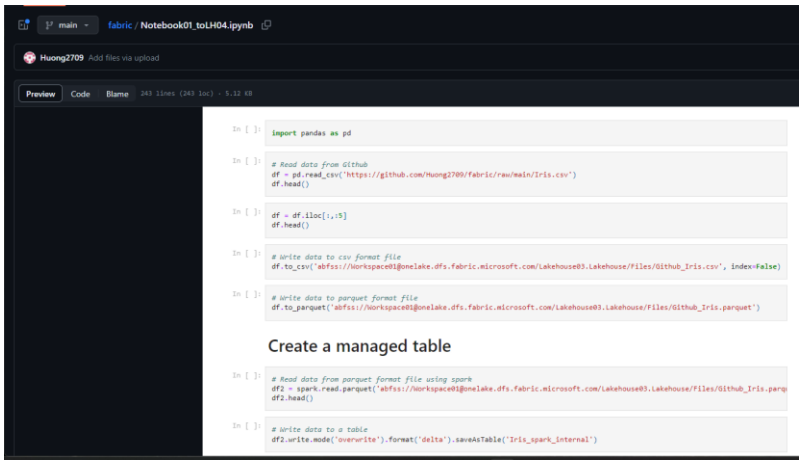
- After Welcome notebook screen is shown, click "Add" button to add Lakehouse. We can create new lakehouse or use existing lakehouse. In this example, let's create new lake house for testing

### 2.2 Adding data to LakeHouse from NoteBook

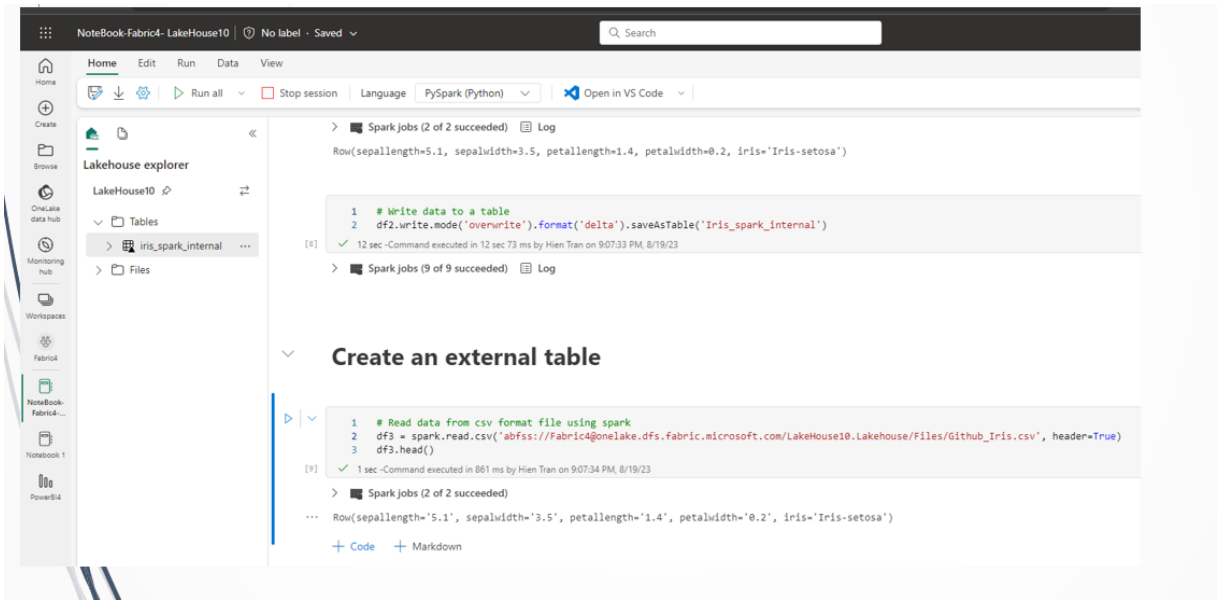
Now we can use notebook in MS Fabric

Check the below scrip source for Notebook experience

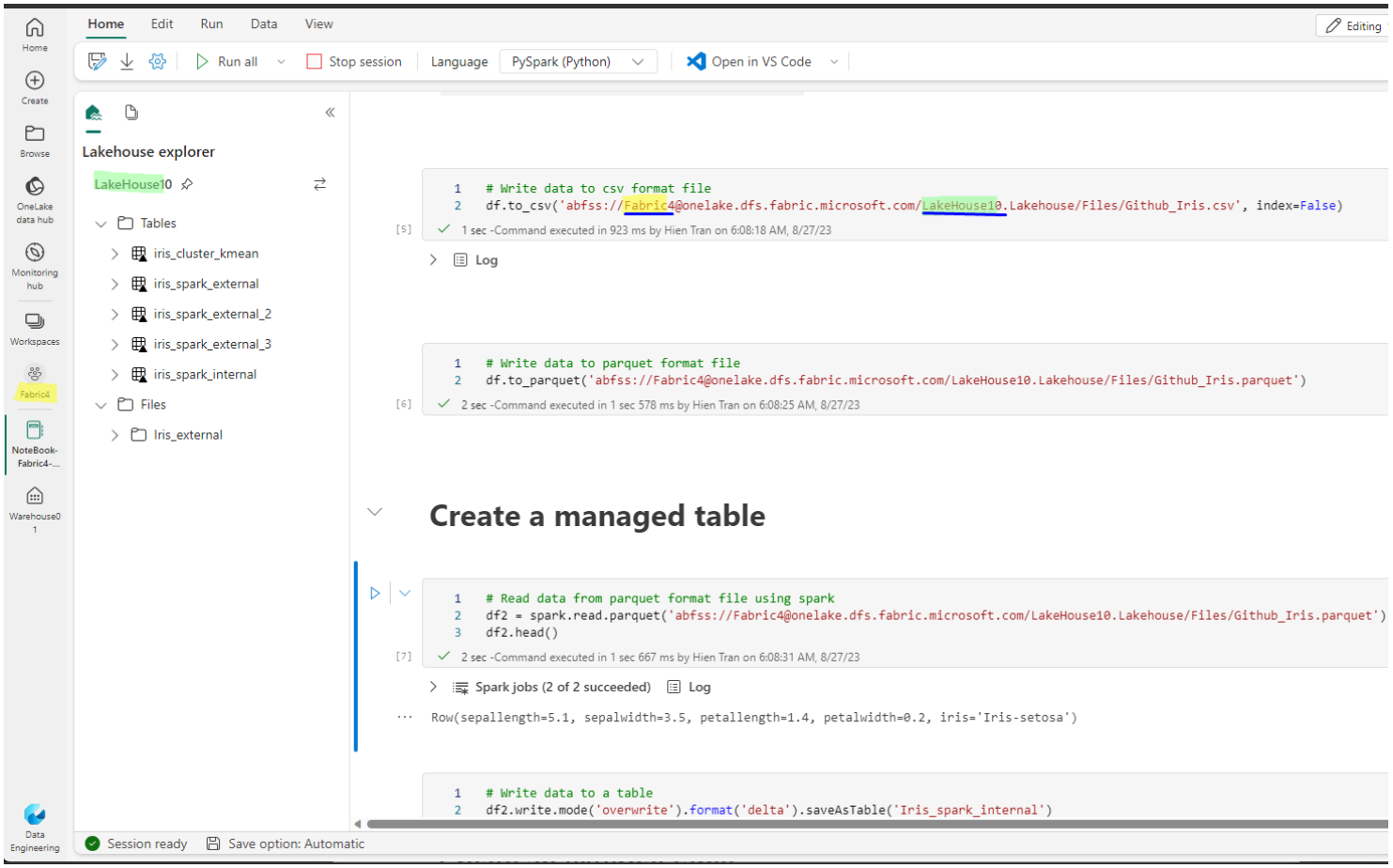
Link : https://github.com/Huong2709/fabric/blob/main/Notebook01_toLH04.ipynb



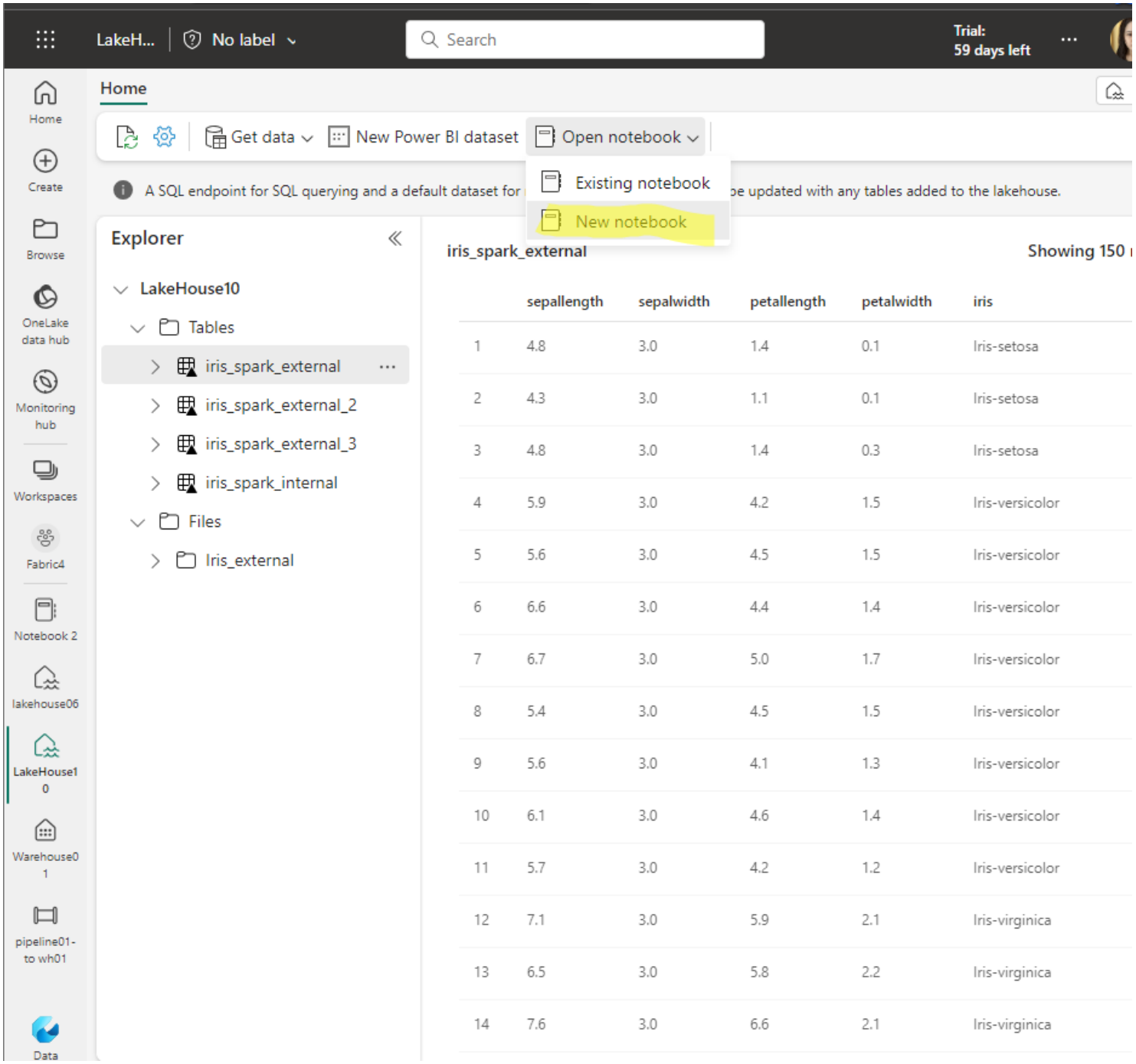

You can write Python code yourself for copy code for check experience




**Note**: Replace Your workspace and Lakehouse name in the script else it will throw error


Home
Edit
Run
Data
View
Editing
Run all
Stop session
Language
PySpark (Python)
Open in VS Code
Lakehouse explorer
LakeHouse10
Tables
iris_cluster_kmean
iris_spark_external
iris_spark_external_2
iris_spark_external_3
iris_spark_internal
Files
Iris_external
1 # Write data to csv format file
2 df.to_csv('abfss://Fabric4@onelake.dfs.fabric.microsoft.com/LakeHouse10.Lakehouse/Files/Github_Iris.csv', index=False)
[5] 1 sec -Command executed in 923 ms by Hien Tran on 6:08:18 AM, 8/27/23
Log
1 # Write data to parquet format file
2 df.to_parquet('abfss://Fabric4@onelake.dfs.fabric.microsoft.com/LakeHouse10.Lakehouse/Files/Github_Iris.parquet')
[6] 2 sec -Command executed in 1 sec 578 ms by Hien Tran on 6:08:25 AM, 8/27/23
Create a managed table
1 # Read data from parquet format file using spark
2 df2 = spark.read.parquet('abfss://Fabric4@onelake.dfs.fabric.microsoft.com/LakeHouse10.Lakehouse/Files/Github_Iris.parquet')
3 df2.head()
[7] 2 sec -Command executed in 1 sec 667 ms by Hien Tran on 6:08:31 AM, 8/27/23
Spark jobs (2 of 2 succeeded)
Log
Row(sepallength=5.1, sepalwidth=3.5, petallength=1.4, petalwidth=0.2, iris='Iris-setosa')
1 # Write data to a table
2 df2.write.mode('overwrite').format('delta').saveAsTable('Iris_spark_internal')
Session ready
Save option: Automatic

## 2.3 New NoteBook in Lakehouse




Source code

https://github.com/Huong2709/fabric/blob/main/Notebook02_Iris_KMean.ipynb

```
In [ ]:
import pandas as pd
from sklearn.cluster import KMeans
```

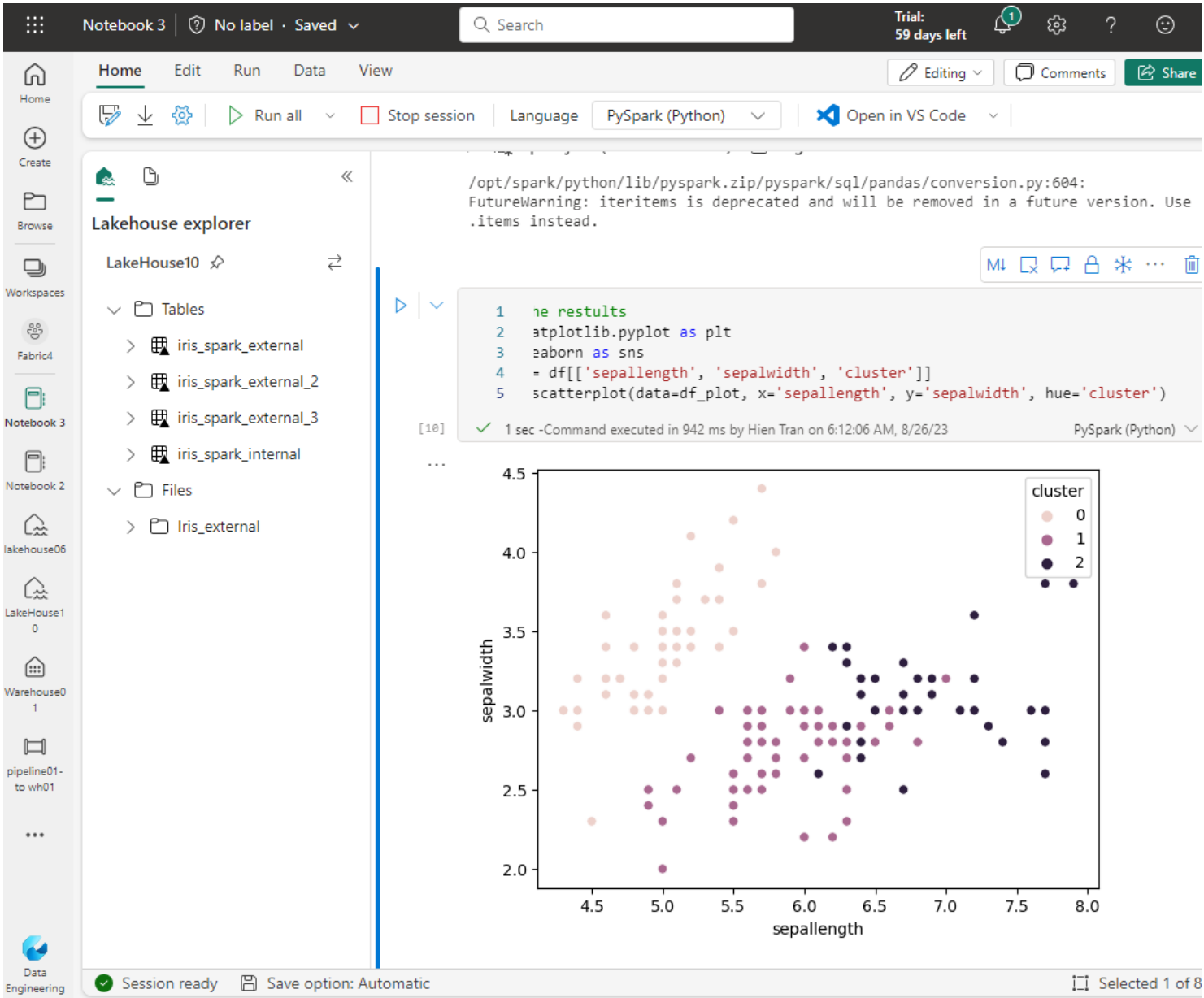
```
In [ ]:
# Read data from file Github_Iris.csv in Workspace03/Lakehouse04
df = pd.read_csv('abfss://Workspace03@onelake.dfs.fabric.microsoft.com/Lakehouse04.Lakehouse/Files/Github_Iris.csv')
df.head()
```

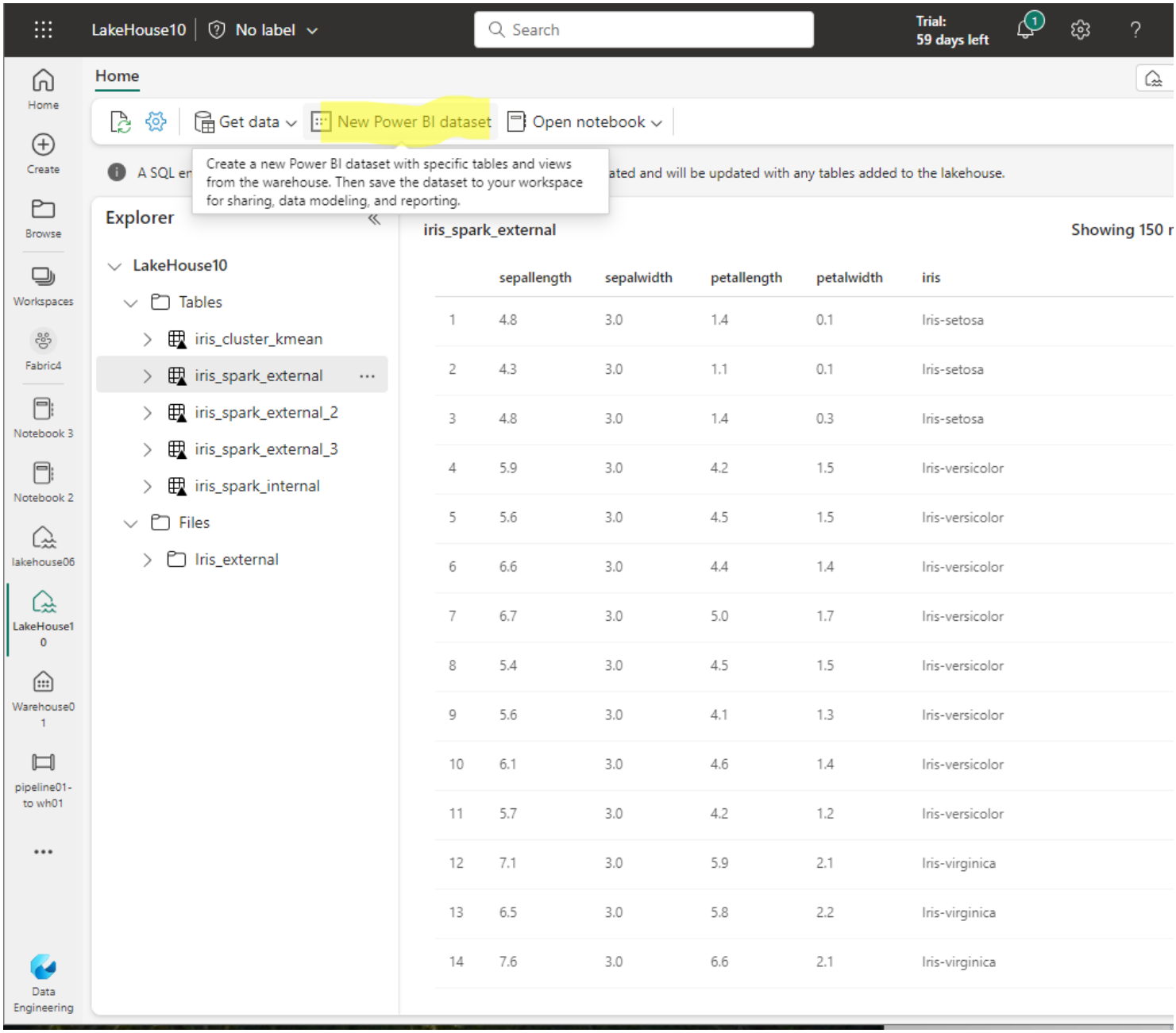
```
In [ ]:
# Choose features
column_for_cluster = ['sepallength', 'sepalwidth', 'petallength', 'petalwidth']
X = df[column_for_cluster]
X.head()
```

```
In [ ]:
# Build clustering model
model_km = KMeans(n_clusters=3)
model_km.fit(X)
```

```
In [ ]:
df['cluster'] = model_km.labels_
df.head()
```

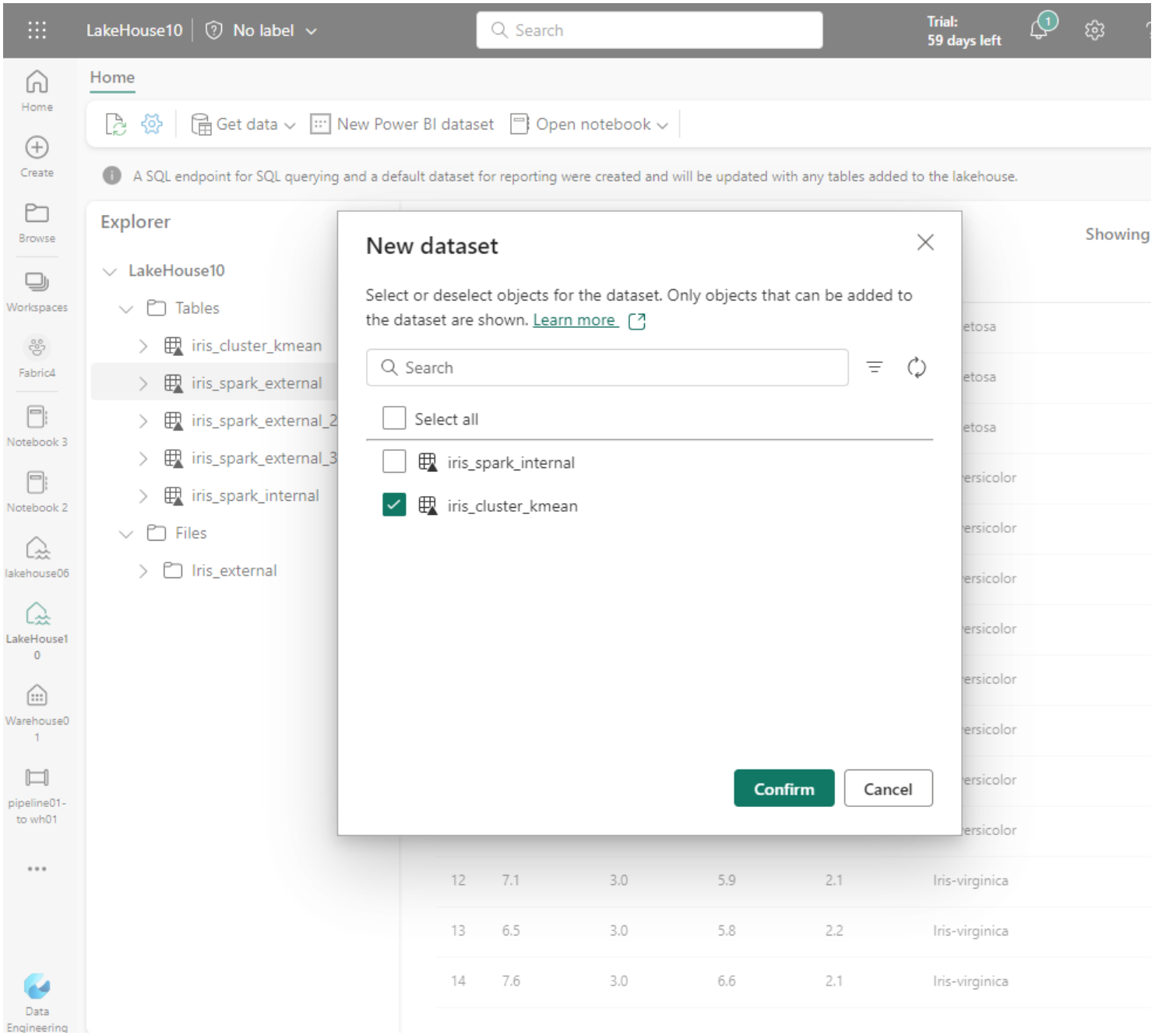
*Create Scatter chart by Notebook*

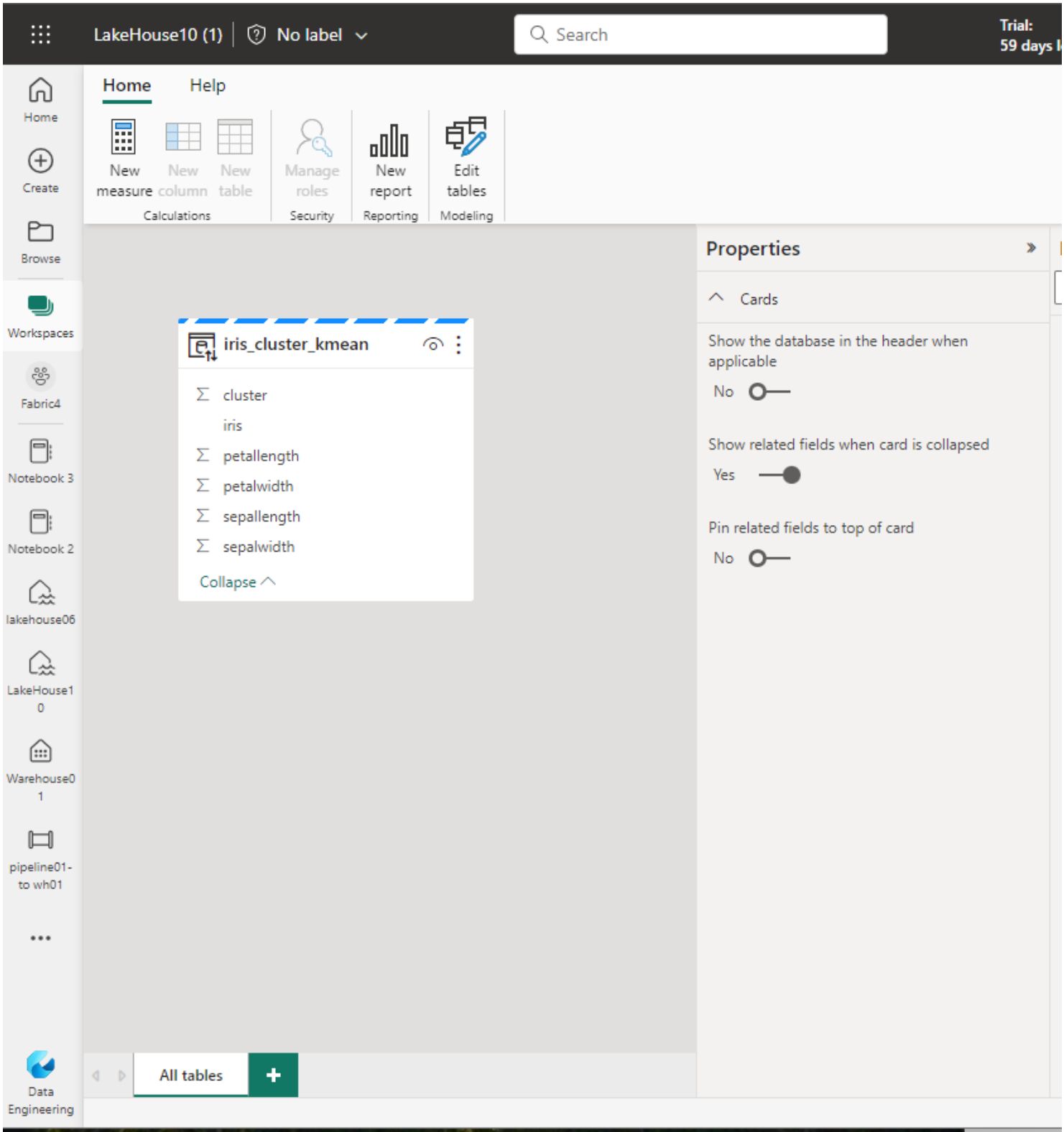


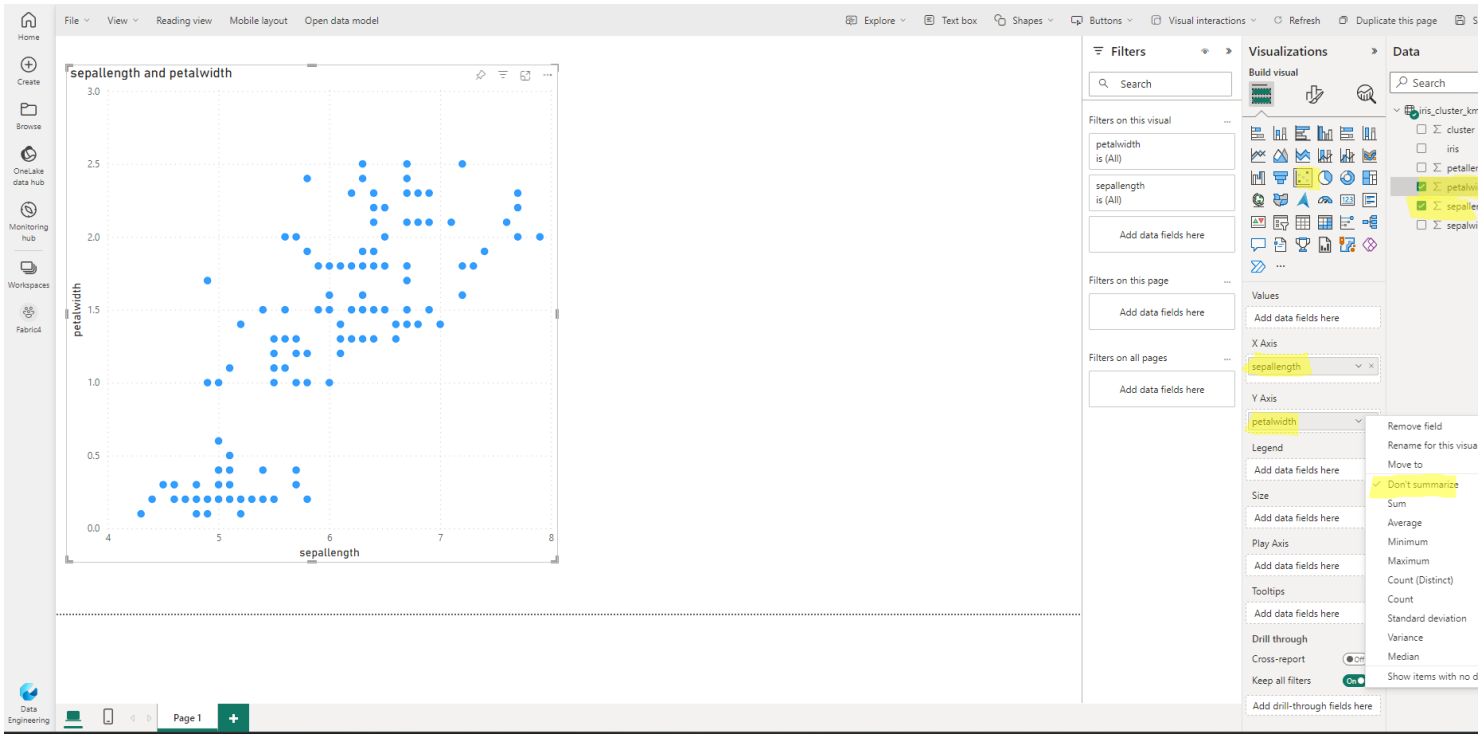


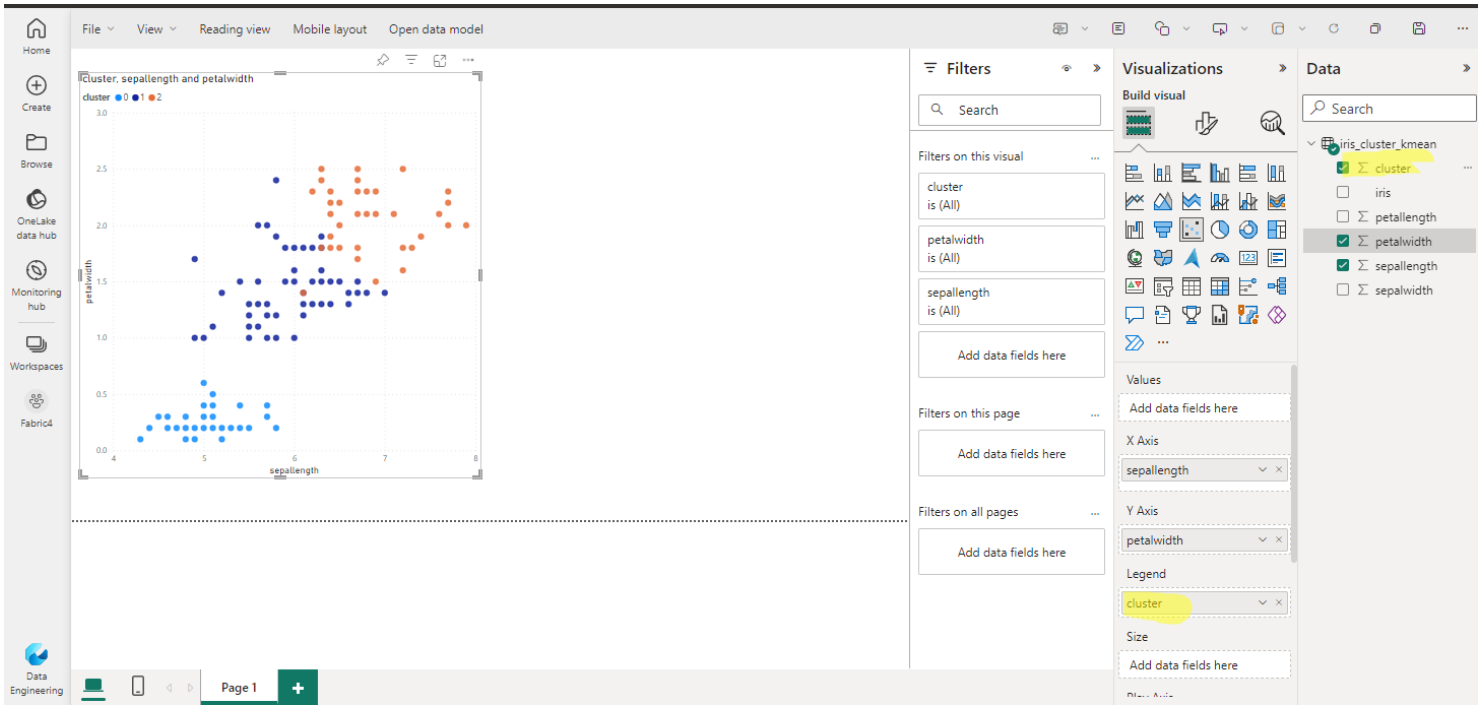

| | sepallength | sepalwidth | petallength | petalwidth | iris |
|---|---|---|---|---|---|
| 1 | 4.8 | 3.0 | 1.4 | 0.1 | Iris-setosa |
| 2 | 4.3 | 3.0 | 1.1 | 0.1 | Iris-setosa |
| 3 | 4.8 | 3.0 | 1.4 | 0.3 | Iris-setosa |
| 4 | 5.9 | 3.0 | 4.2 | 1.5 | Iris-versicolor |
| 5 | 5.6 | 3.0 | 4.5 | 1.5 | Iris-versicolor |
| 6 | 6.6 | 3.0 | 4.4 | 1.4 | Iris-versicolor |
| 7 | 6.7 | 3.0 | 5.0 | 1.7 | Iris-versicolor |
| 8 | 5.4 | 3.0 | 4.5 | 1.5 | Iris-versicolor |
| 9 | 5.6 | 3.0 | 4.1 | 1.3 | Iris-versicolor |
| 10 | 6.1 | 3.0 | 4.6 | 1.4 | Iris-versicolor |
| 11 | 5.7 | 3.0 | 4.2 | 1.2 | Iris-versicolor |
| 12 | 7.1 | 3.0 | 5.9 | 2.1 | Iris-virginica |
| 13 | 6.5 | 3.0 | 5.8 | 2.2 | Iris-virginica |
| 14 | 7.6 | 3.0 | 6.6 | 2.1 | Iris-virginica |

*Create Scatter chart by Power BI*

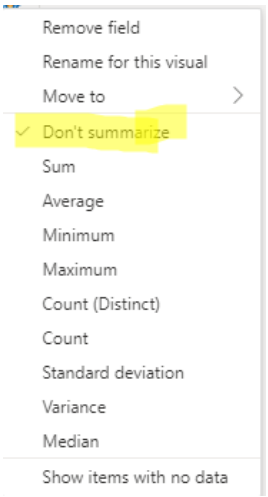



LakeHouse10 (1)
No label
Search
Trial:
59 days
Home
Help
Home
Create
Browse
Workspaces
Fabric4
Notebook 3
Notebook 2
lakehouse06
LakeHouse10
Warehouse01
pipeline01-to wh01
Data Engineering
New measure
New column
New table
Calculations
Manage roles
Security
New report
Reporting
Edit tables
Modeling
iris_cluster_kmean
cluster
iris
petallength
petalwidth
sepallength
sepalwidth
Collapse
Properties
Cards
Show the database in the header when applicable
No
Show related fields when card is collapsed
Yes
Pin related fields to top of card
No
All tables

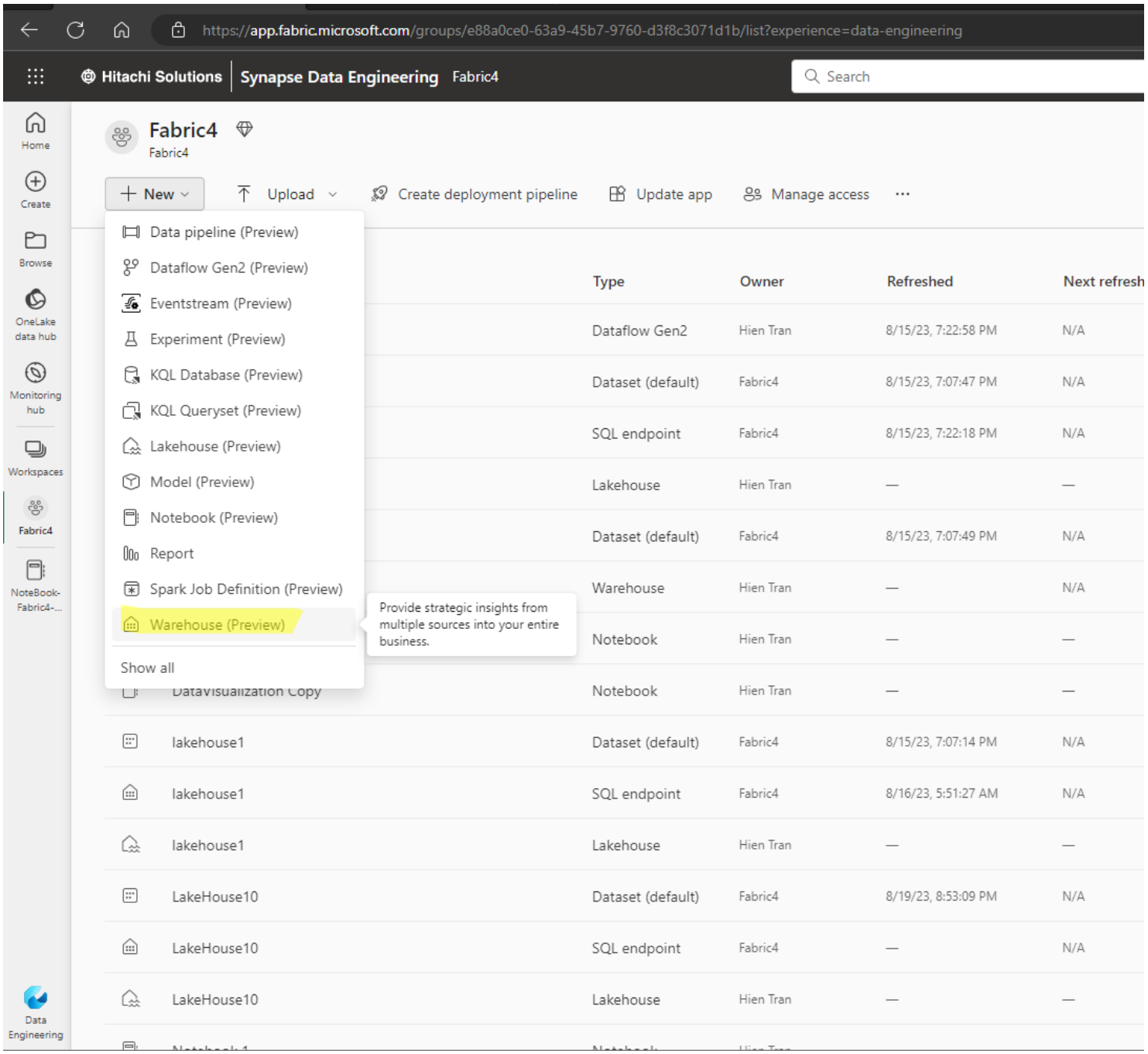

sepallength and petalwidth
petalwidth
sepallength
Filters
Search
Filters on this visual
petalwidth is (All)
sepallength is (All)
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
Values
Add data fields here
X Axis
sepallength
Y Axis
petalwidth
Legend
Add data fields here
Size
Add data fields here
Play Axis
Add data fields here
Tooltips
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Remove field
Rename for this visual
Move to
Don't summarize
Sum
Average
Minimum
Maximum
Count (Distinct)
Count
Standard deviation
Variance
Median
Show items with no d
Data
Search
iris_cluster_kmean
cluster
iris
petallength
petalwidth
sepallength
sepalwidth
File
View
Reading view
Mobile layout
Open data model
Explore
Text box
Shapes
Buttons
Visual interactions
Refresh
Duplicate this page
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
Data Engineering
Page 1


cluster, sepallength and petalwidth
cluster 0 1 2
petalwidth
sepallength
Filters
Search
Filters on this visual
cluster is (All)
petalwidth is (All)
sepallength is (All)
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
Values
Add data fields here
X Axis
sepallength
Y Axis
petalwidth
Legend
cluster
Size
Add data fields here
Data
Search
iris_cluster_kmean
cluster
iris
petallength
petalwidth
sepallength
sepalwidth
File
View
Reading view
Mobile layout
Open data model
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
Data Engineering
Page 1


**Note**: Click in X Axis, Y Axis and select " Don't summarize"

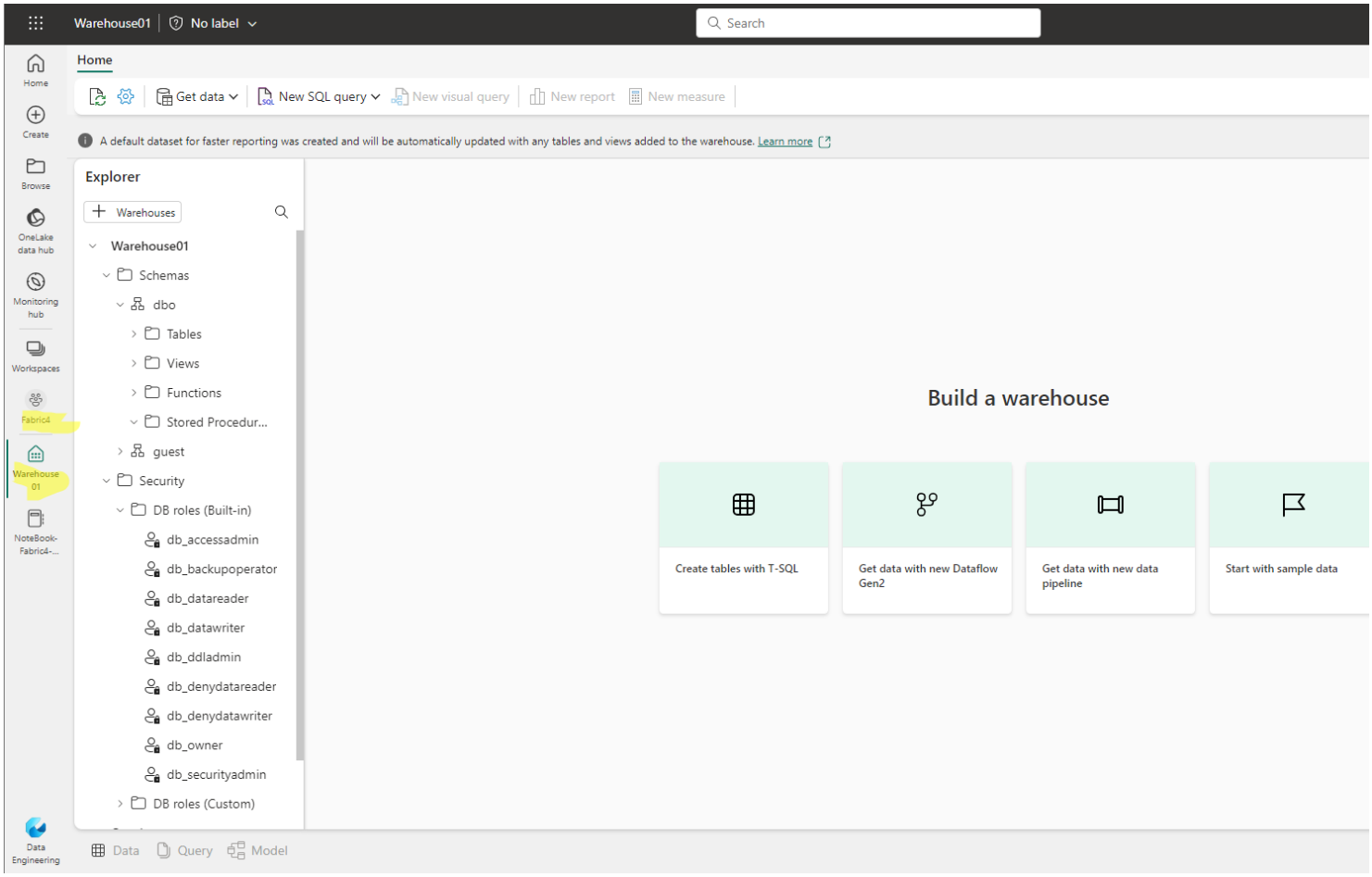


## 2.4 Create warehouse




- From Workspace screen > Click New > Select “Warehouse..”

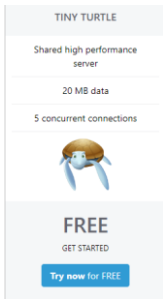

- Input Warehouse name




## 2.5 ElephantSQL

Create trial account in Elephant SQL for testing loading data from external database

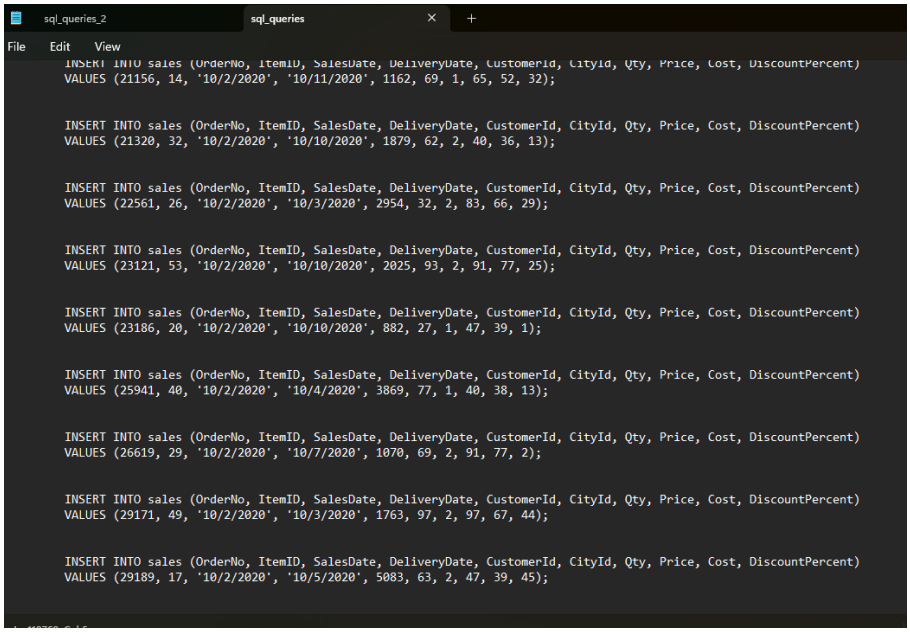
https://www.elephantsql.com/plans.html

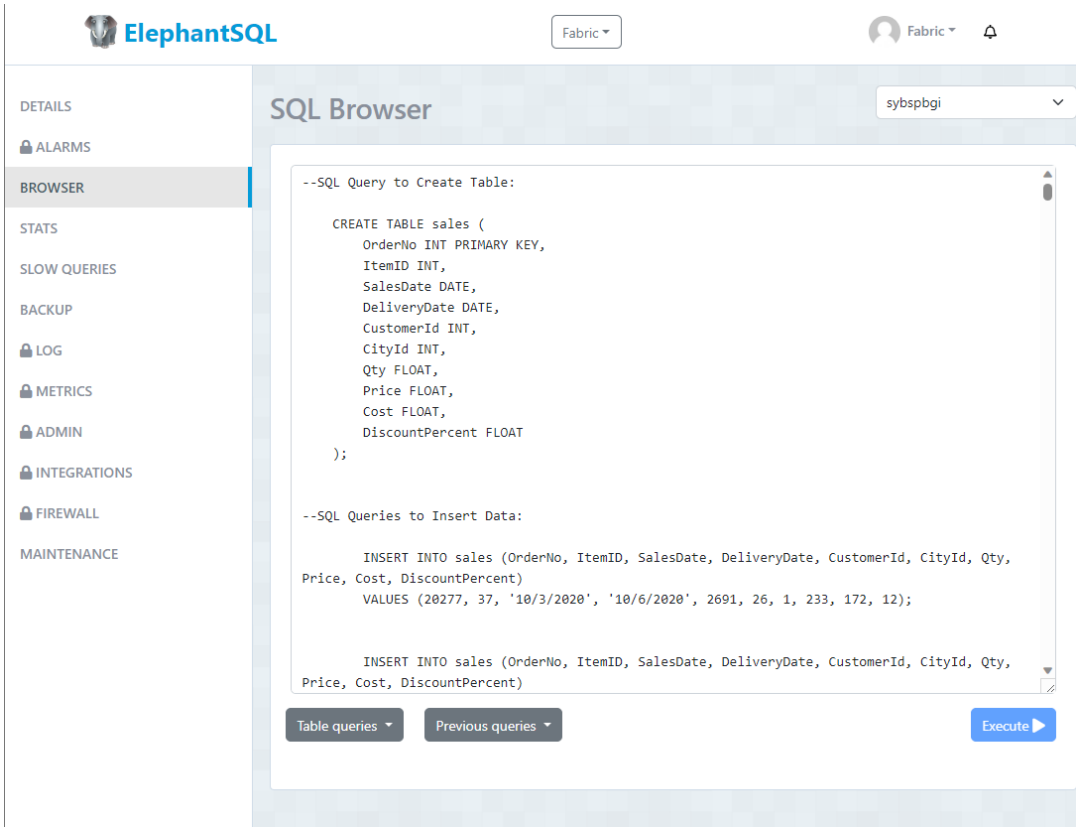



### *2.5.1 Create sales table in SQL*

SQL_Script

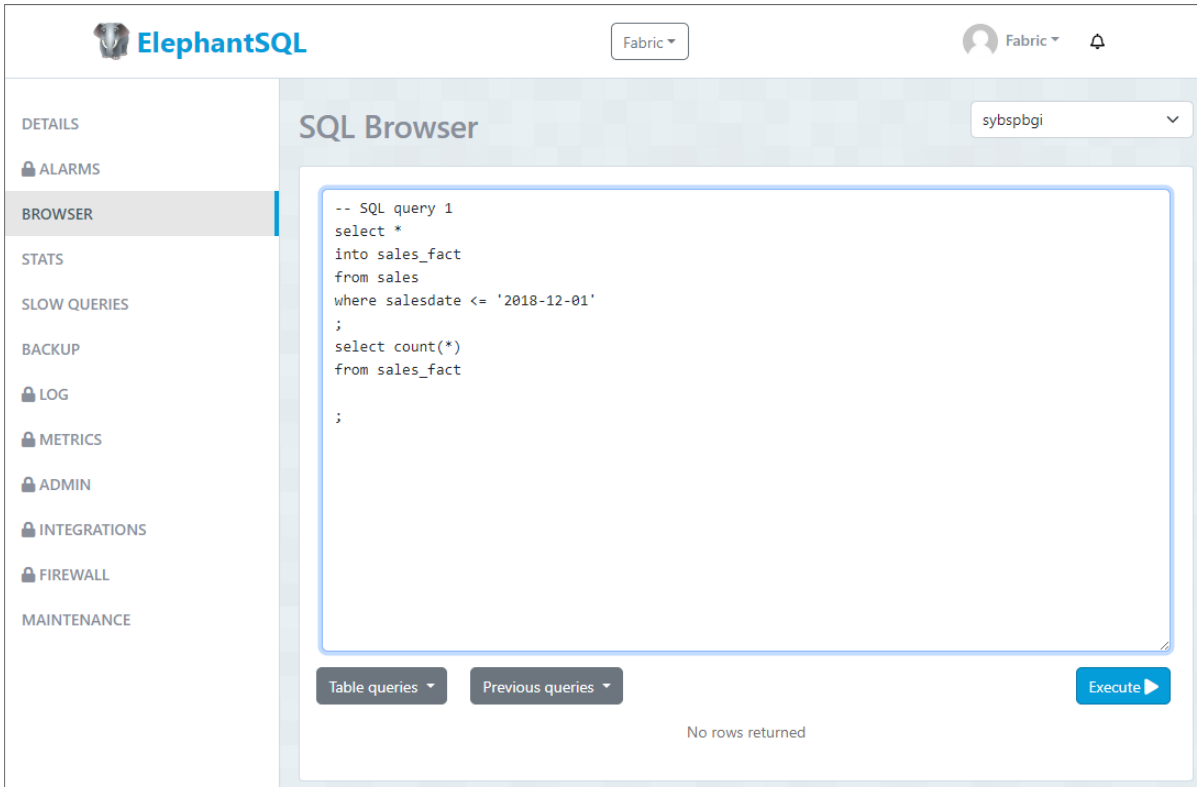
https://github.com/Huong2709/fabric/blob/main/sql_queries.txt



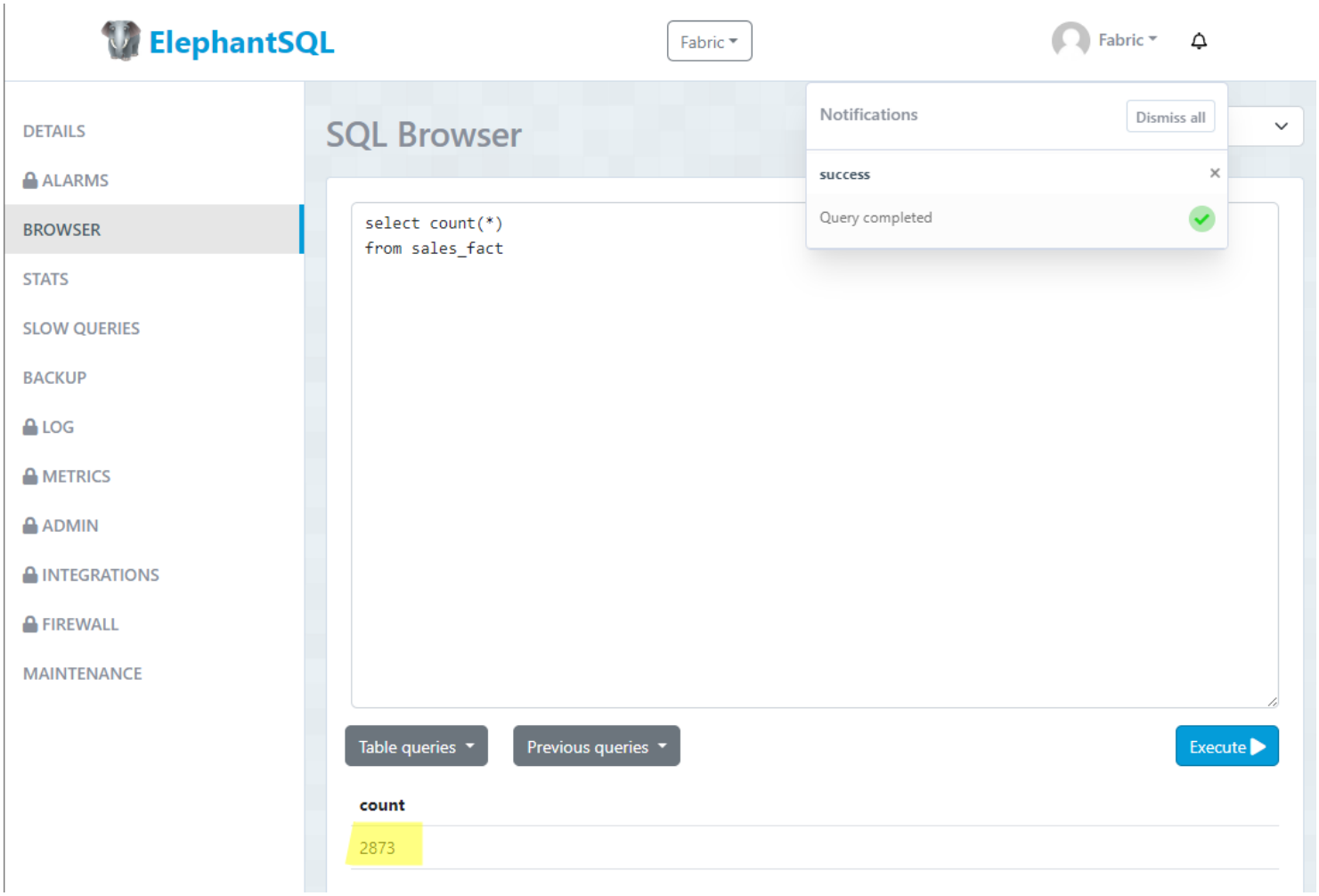




- Navigate to “Browser” input script for create your first table to test

*2.5.2 Create sales_fact from sales*



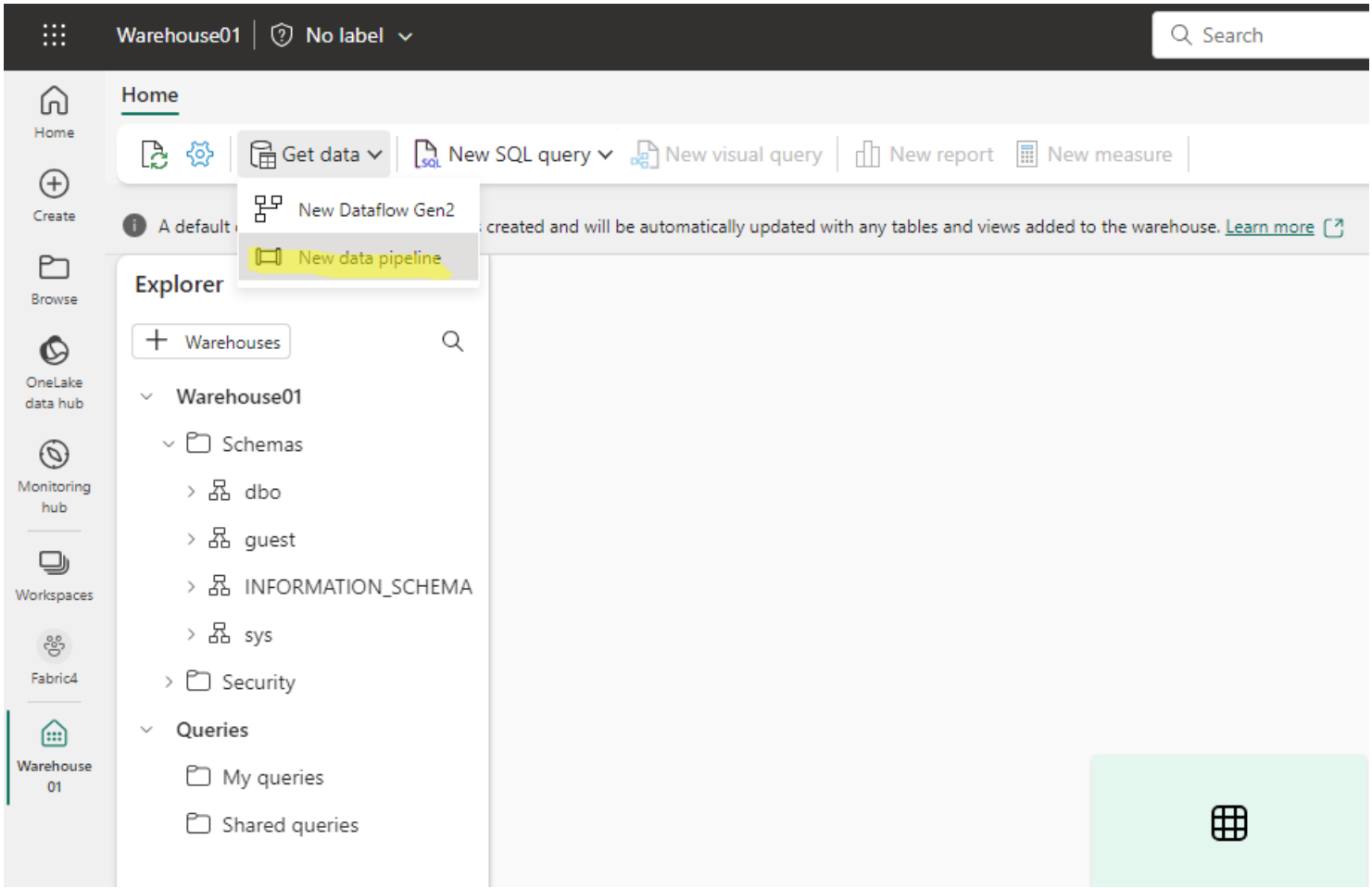

```
-- SQL query 1
select *
into sales_fact
from sales
where salesdate <= '2018-12-01'
;
select count(*)
from sales_fact


;
```



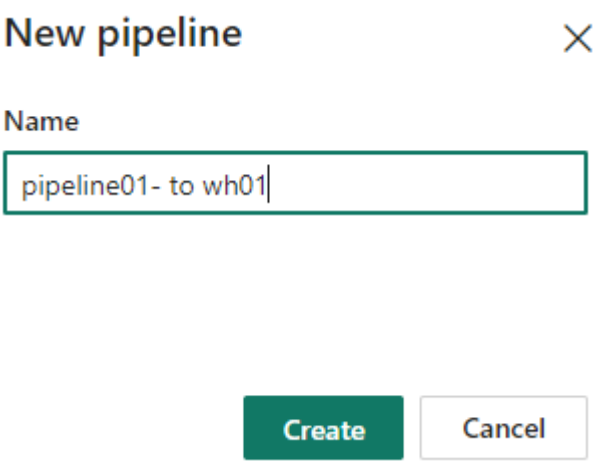

select count(*)

from sales_fact

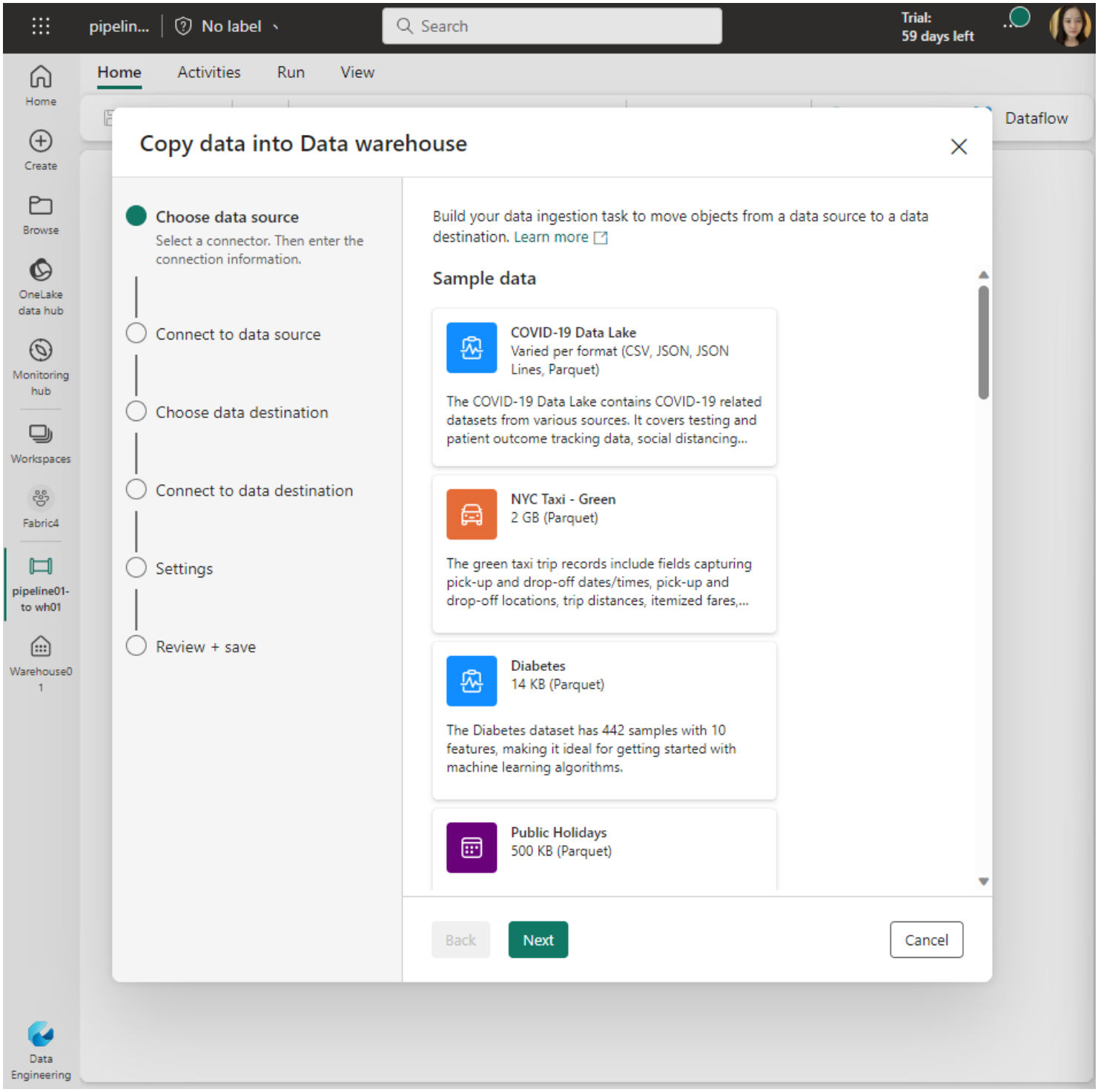
## 2.6 Create pipeline in warehouse

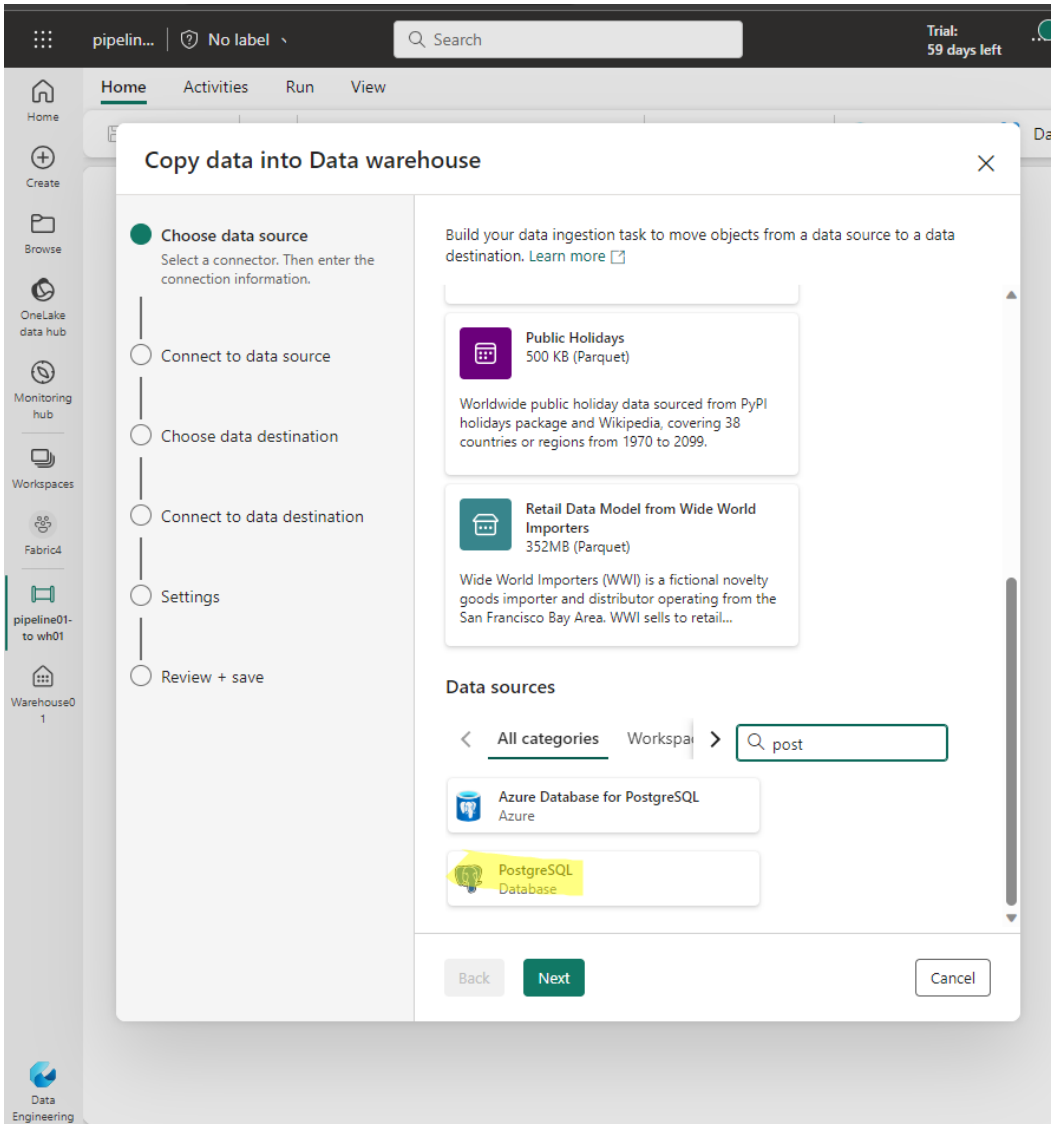



- From warehouse > click Get data > select New data pipeline

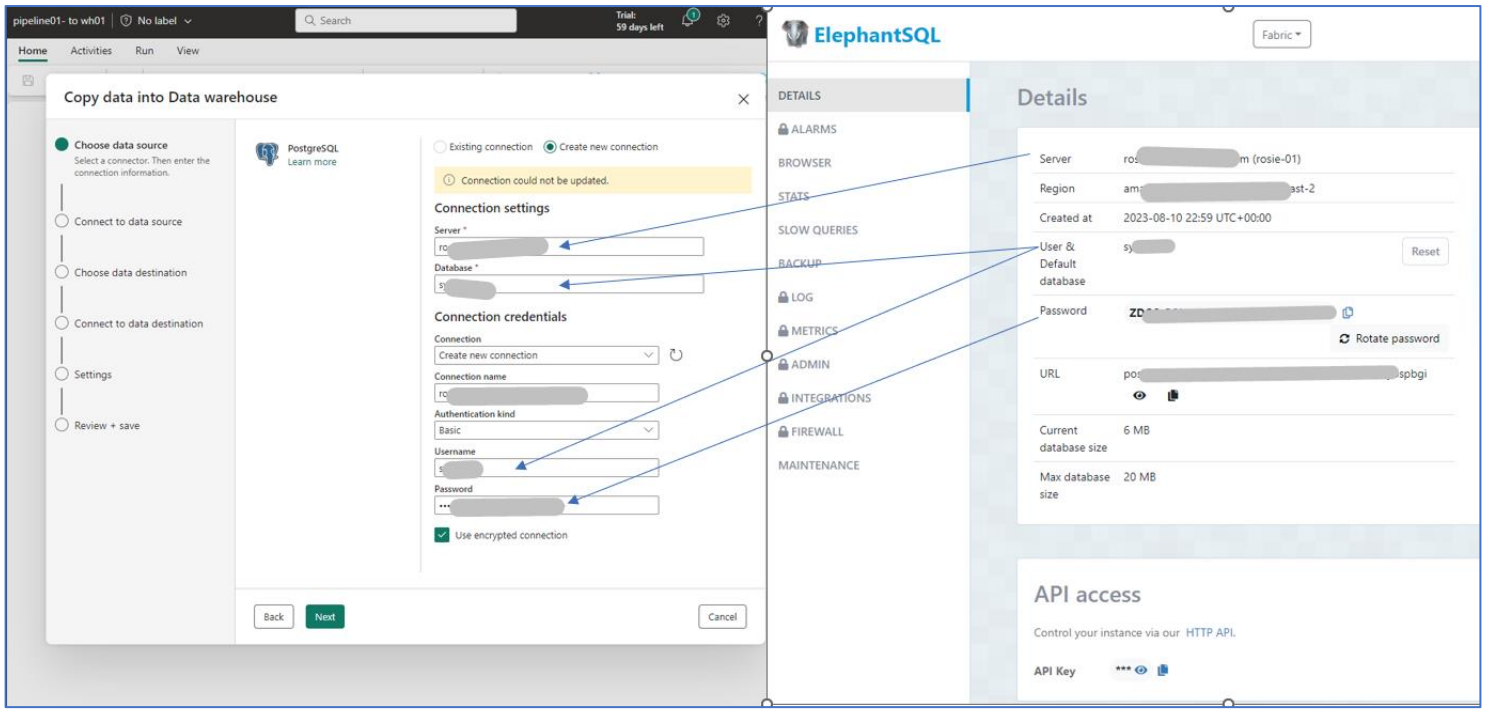



- Input New pipeline

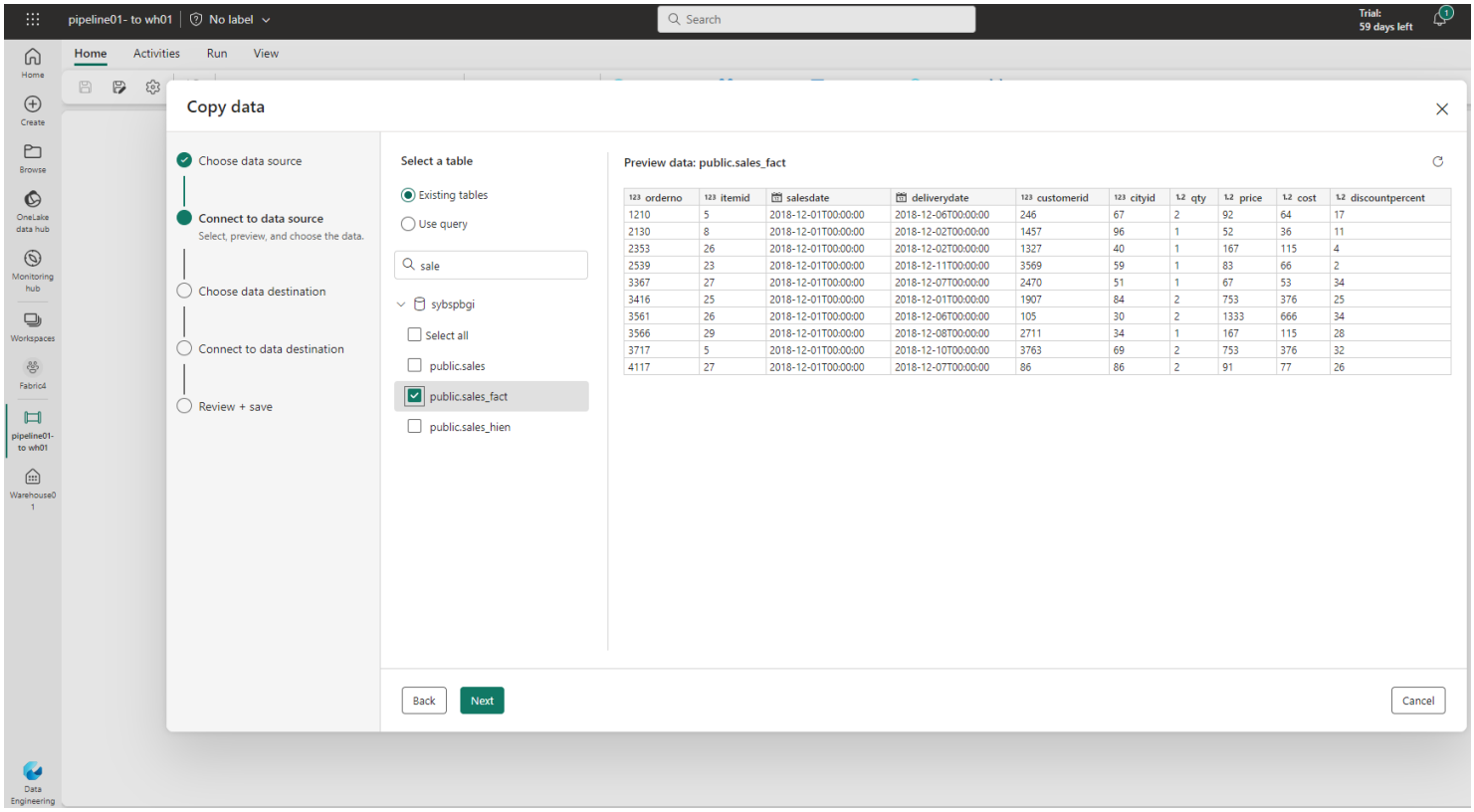

pipelin...
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Dataflow
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipeline01-to wh01
Warehouse01
Data Engineering
Copy data into Data warehouse
Choose data source
Select a connector. Then enter the connection information.
Connect to data source
Choose data destination
Connect to data destination
Settings
Review + save
Build your data ingestion task to move objects from a data source to a data destination. Learn more
Sample data
COVID-19 Data Lake
Varied per format (CSV, JSON, JSON Lines, Parquet)
The COVID-19 Data Lake contains COVID-19 related datasets from various sources. It covers testing and patient outcome tracking data, social distancing...
NYC Taxi - Green
2 GB (Parquet)
The green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares,...
Diabetes
14 KB (Parquet)
The Diabetes dataset has 442 samples with 10 features, making it ideal for getting started with machine learning algorithms.
Public Holidays
500 KB (Parquet)
Back
Next
Cancel



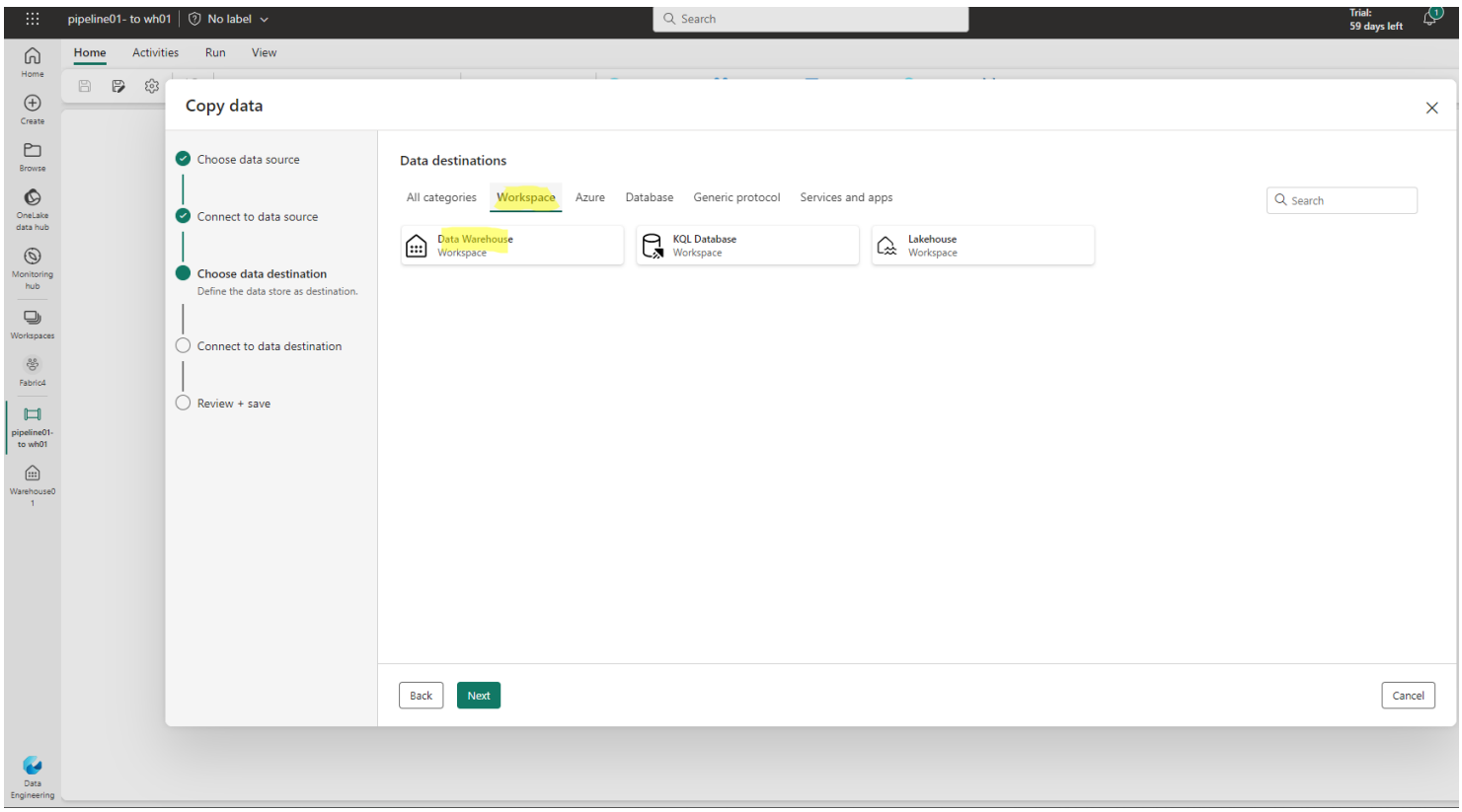

- Select PostgreSQL data source



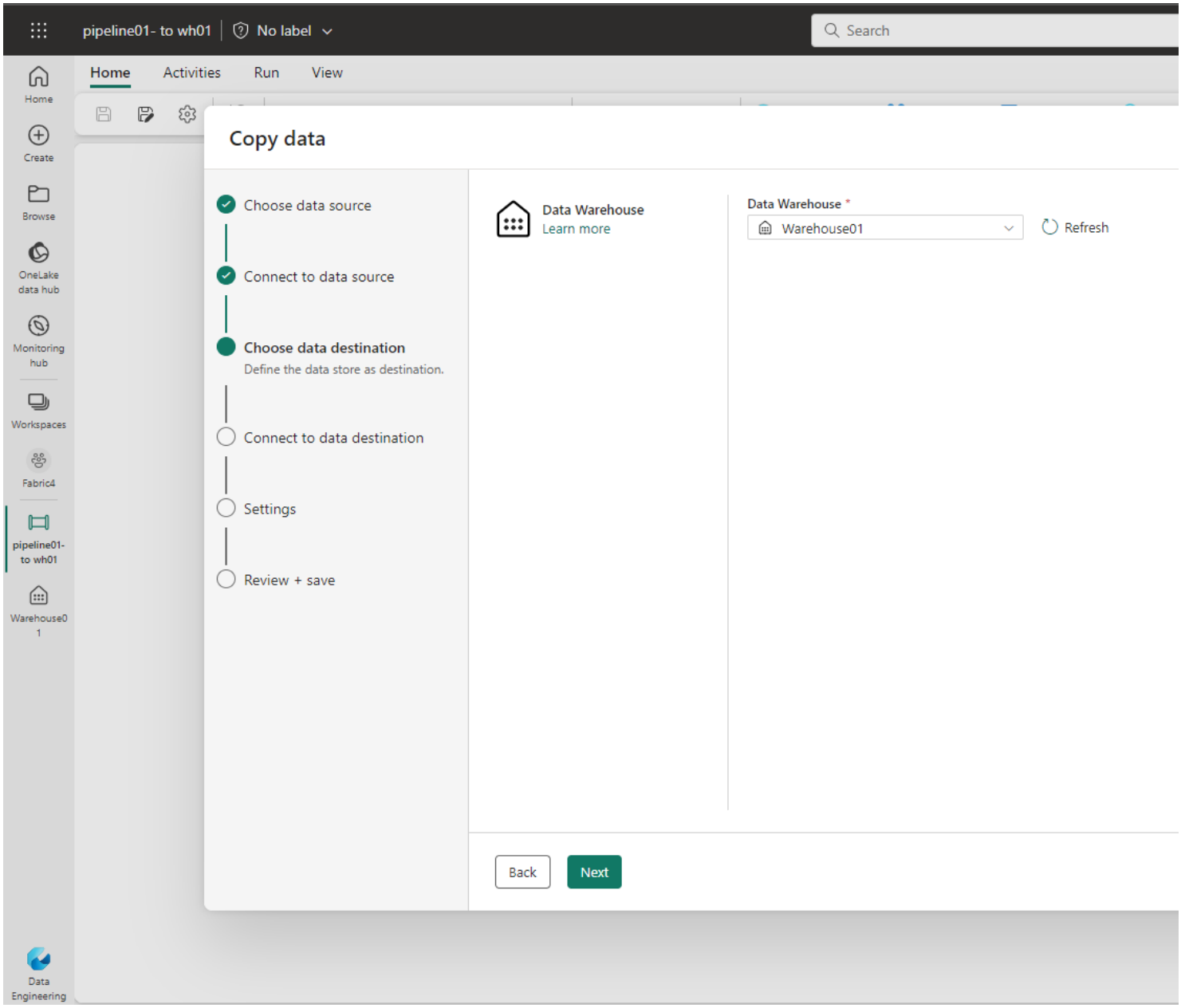

- *Input connection details*
- *Click Next*



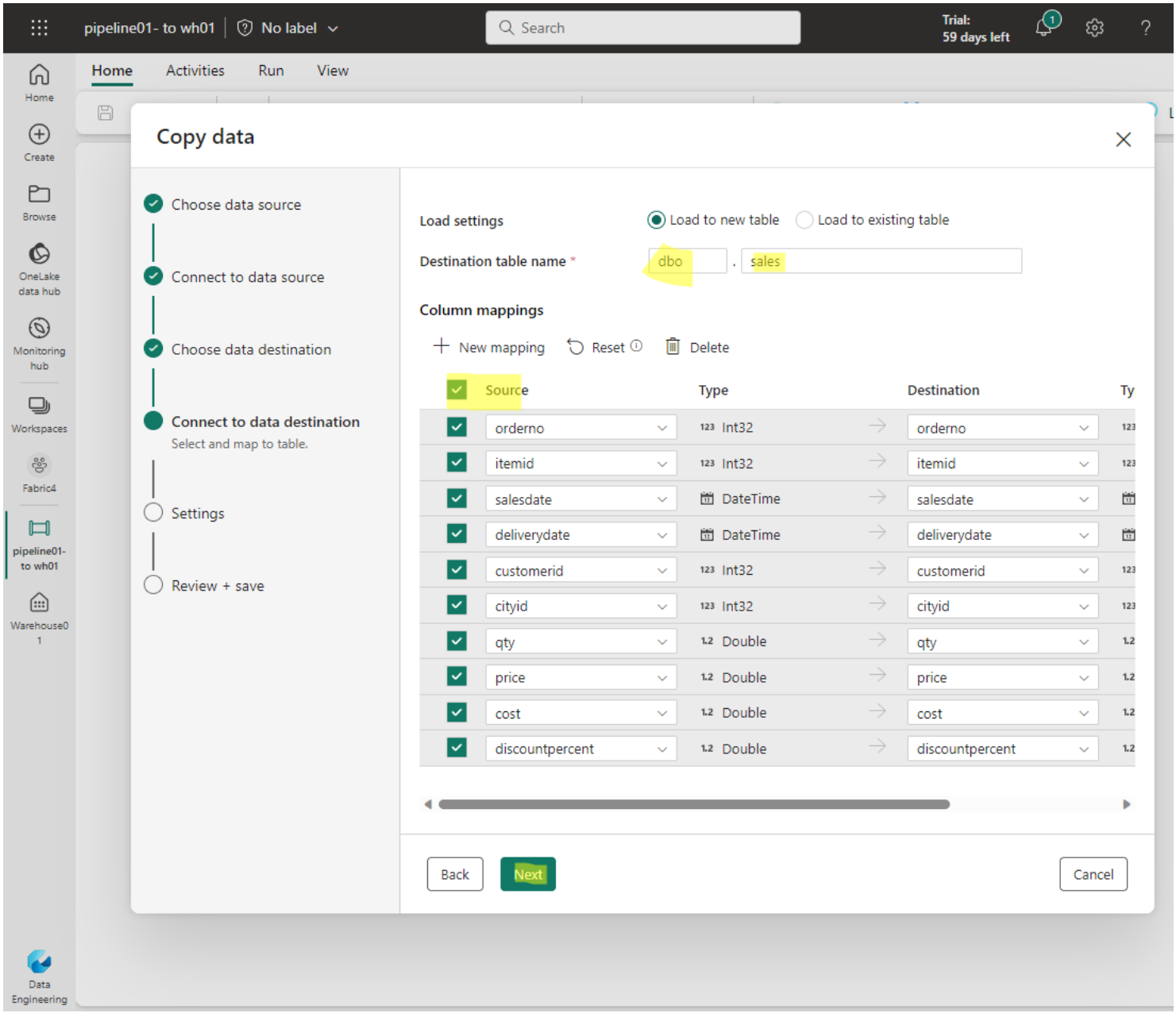

- From connect to data source step > Select Public.sales_fact > click Next



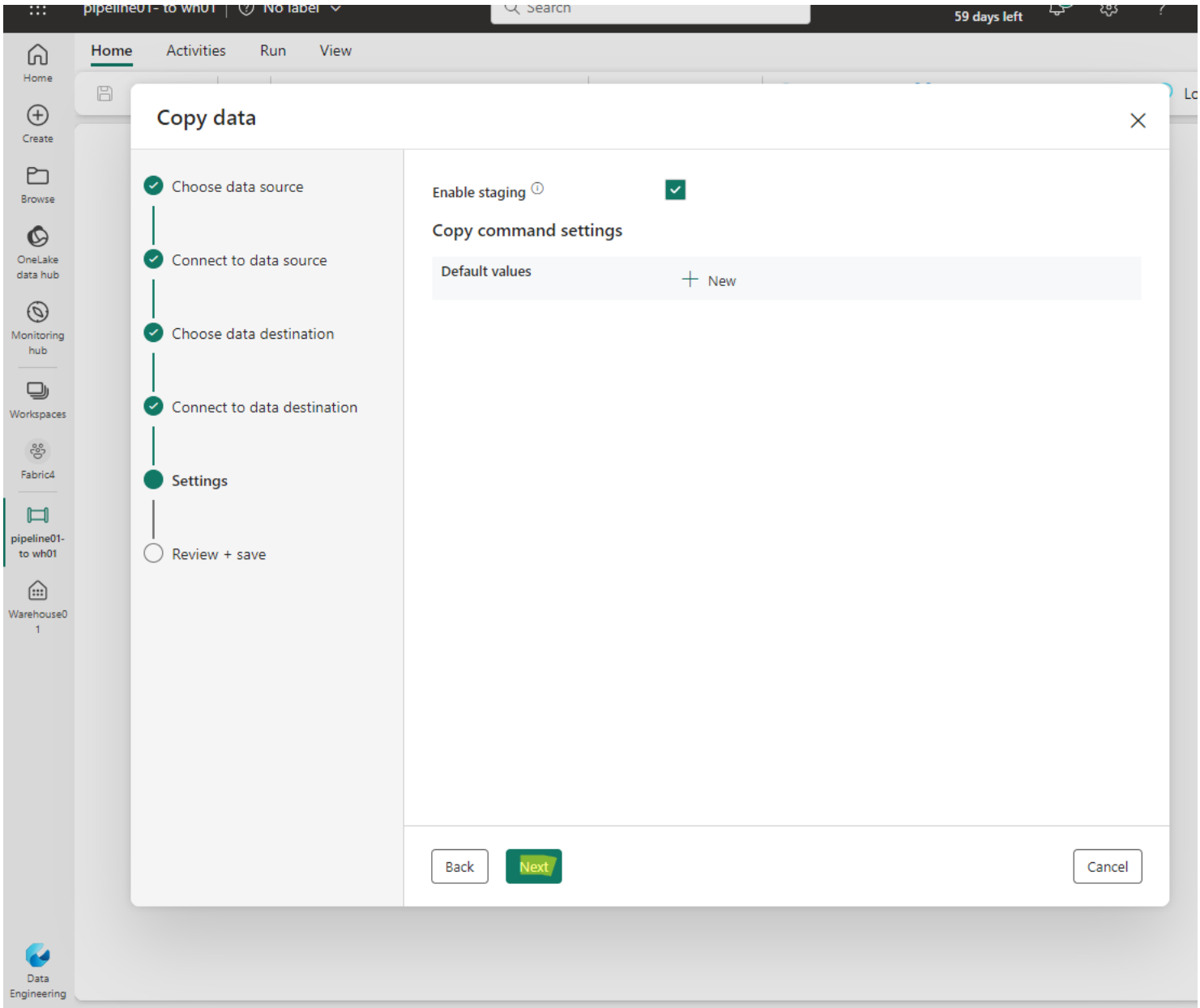

- From Choose data destination > select Workspace > Click Data warehouse > click Next



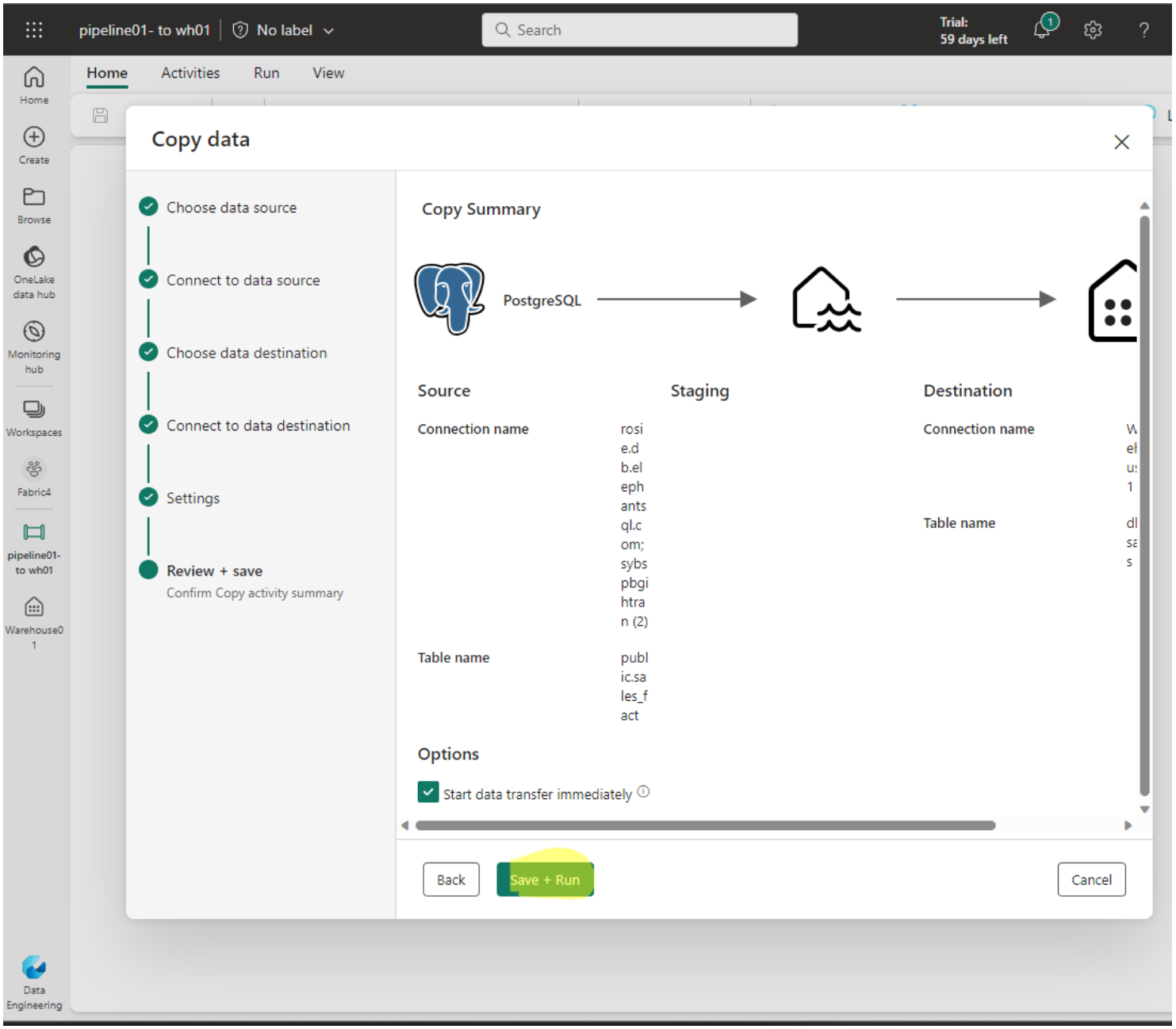

- Select Data warehouse




- Select “Load to new table”
- Destination table name: Provide table name in Fabric
- Select Source to load
- Click Next


pipeline01- to wh01
No label
Search
59 days left
Home
Activities
Run
View
Copy data
Choose data source
Connect to data source
Choose data destination
Connect to data destination
Settings
Review + save
Enable staging
Copy command settings
Default values
New
Back
Next
Cancel
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipeline01- to wh01
Warehouse01
Data Engineering

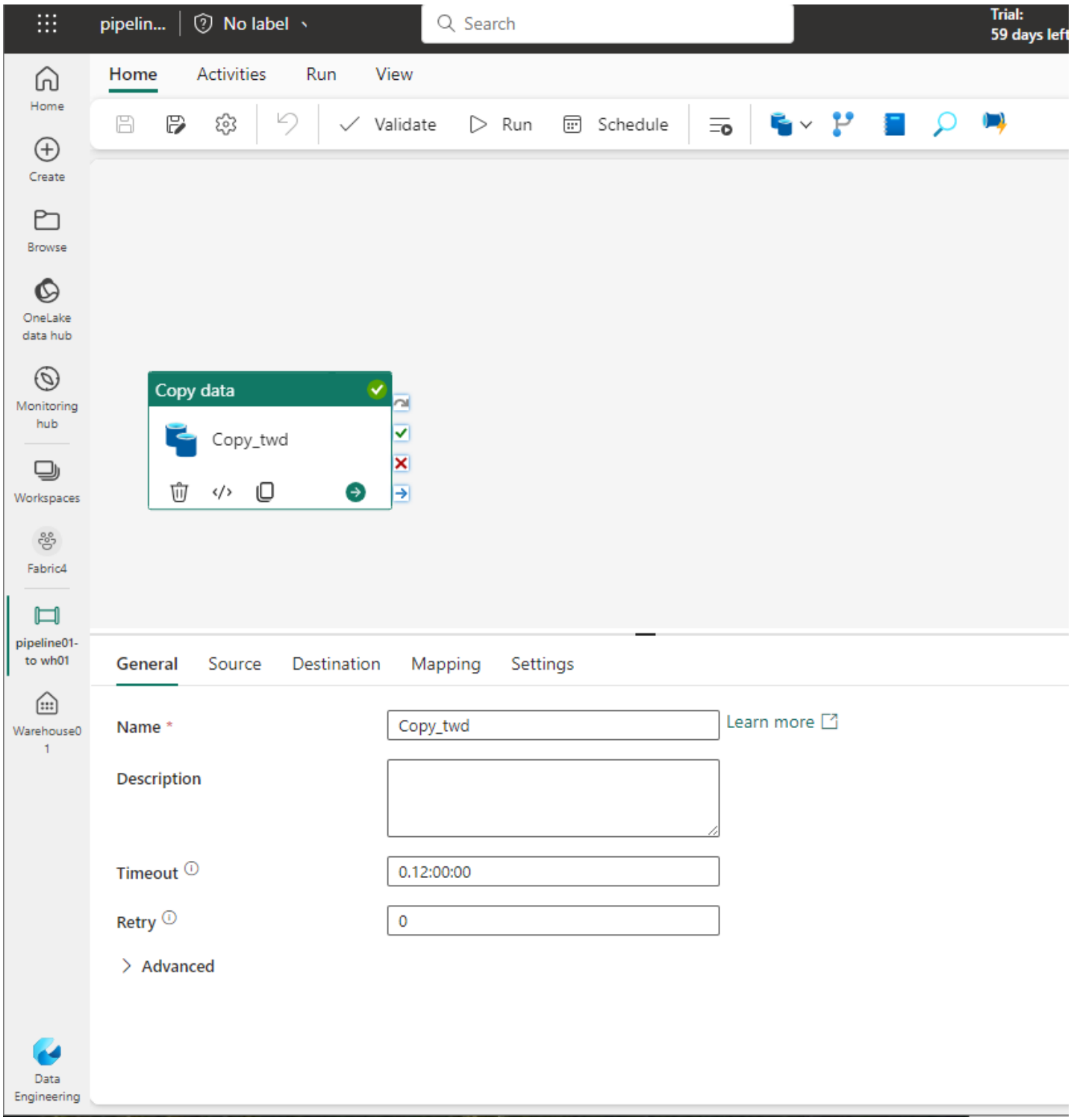
- Click Next



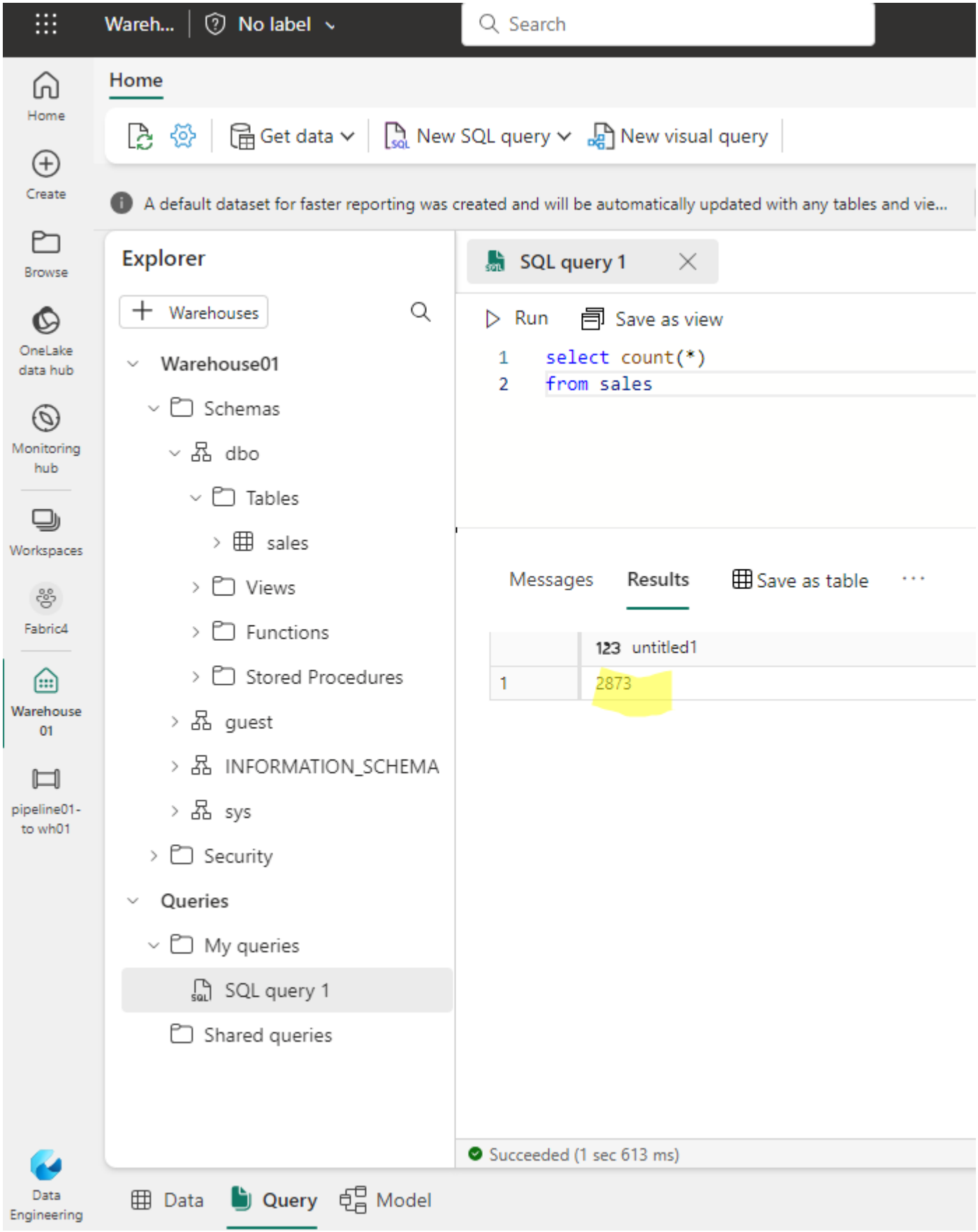

- Click “Save + Run”

*Pipeline*

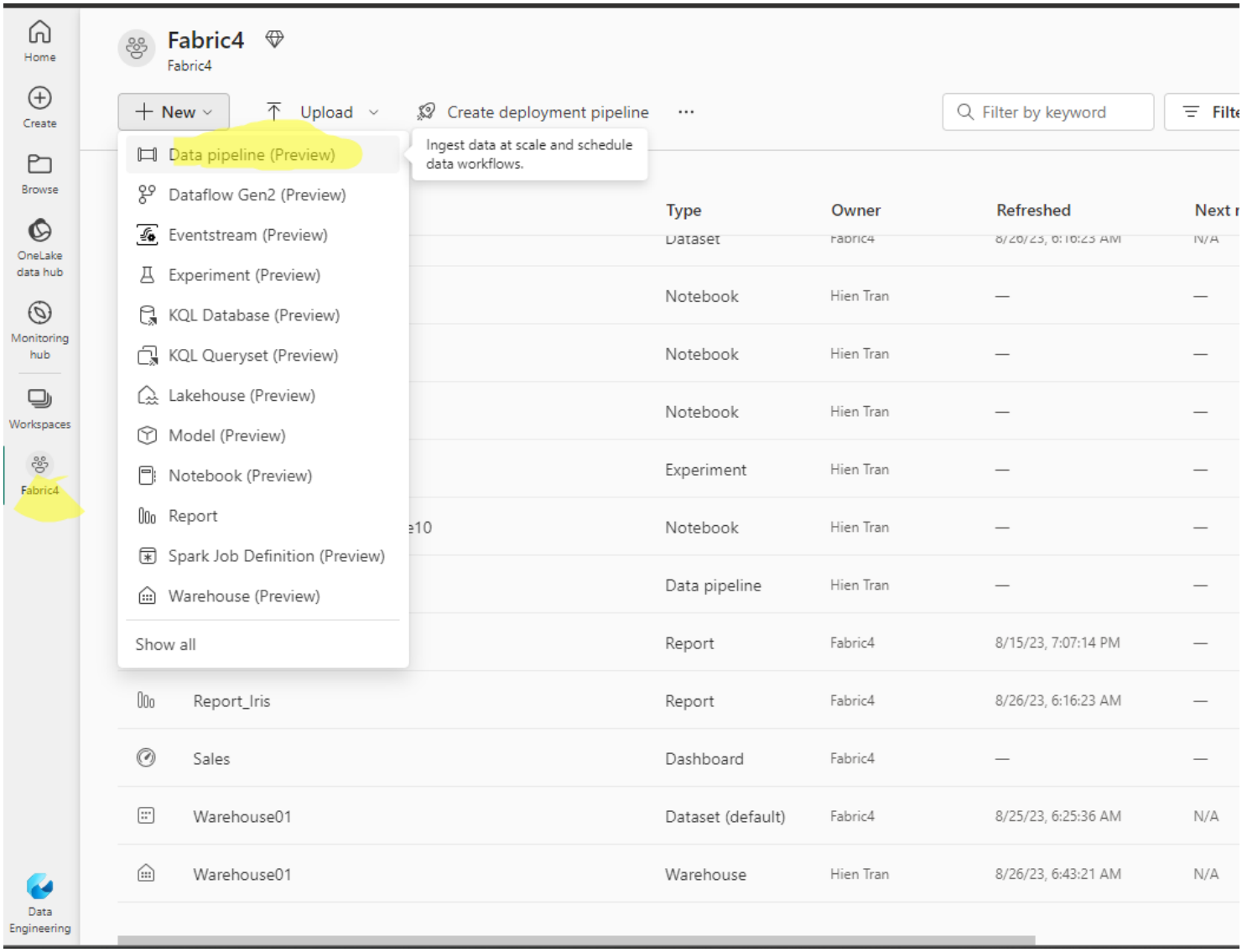
Copy data step is automatically creation after completion pipeline creation step in the  above

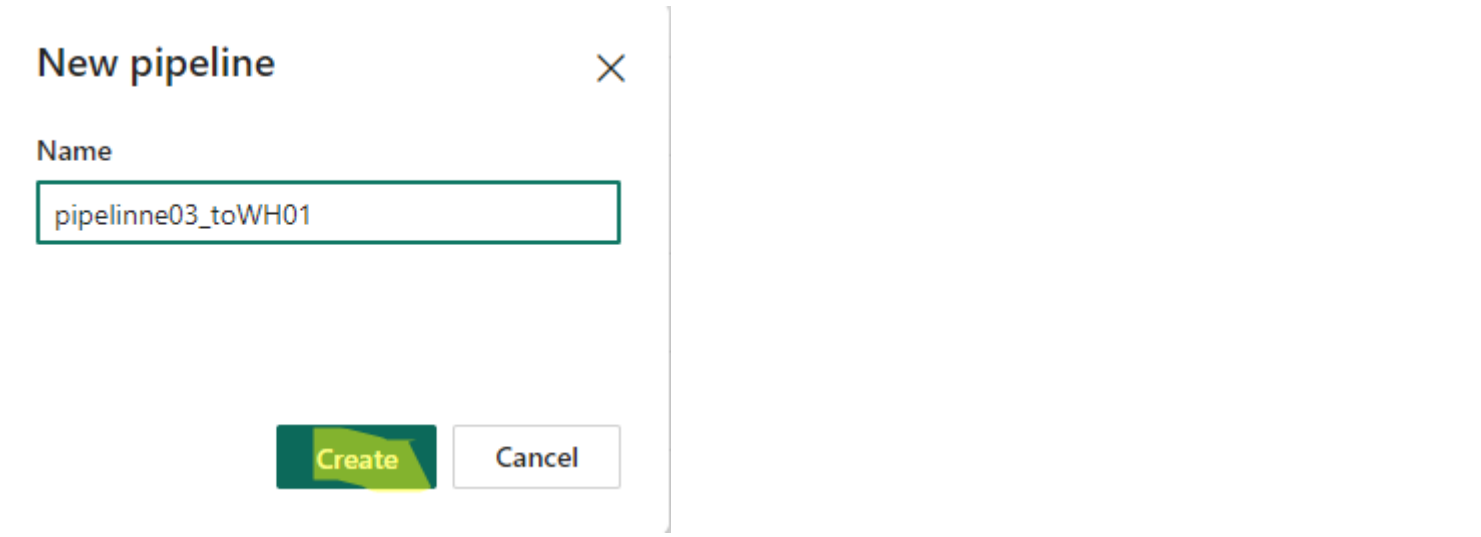

pipelin...
No label
Search
Trial:
59 days left
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipeline01-to wh01
Warehouse01
Data Engineering
Home
Activities
Run
View
Validate
Run
Schedule
Copy data
Copy_twd
General
Source
Destination
Mapping
Settings
Name *
Copy_twd
Learn more
Description
Timeout
0.12:00:00
Retry
0
Advanced

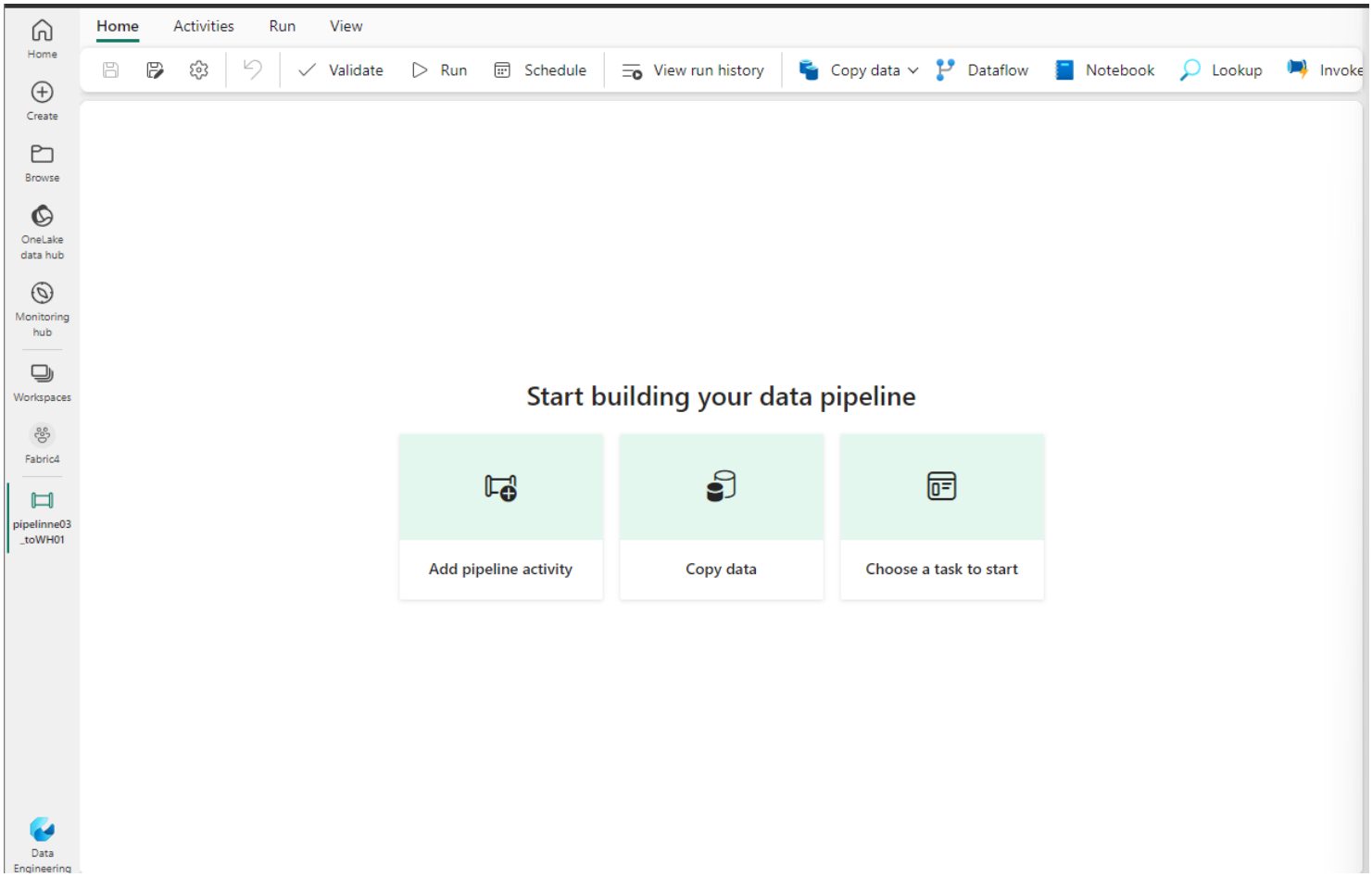
*Check data in Fabric after loading data by pipeline*



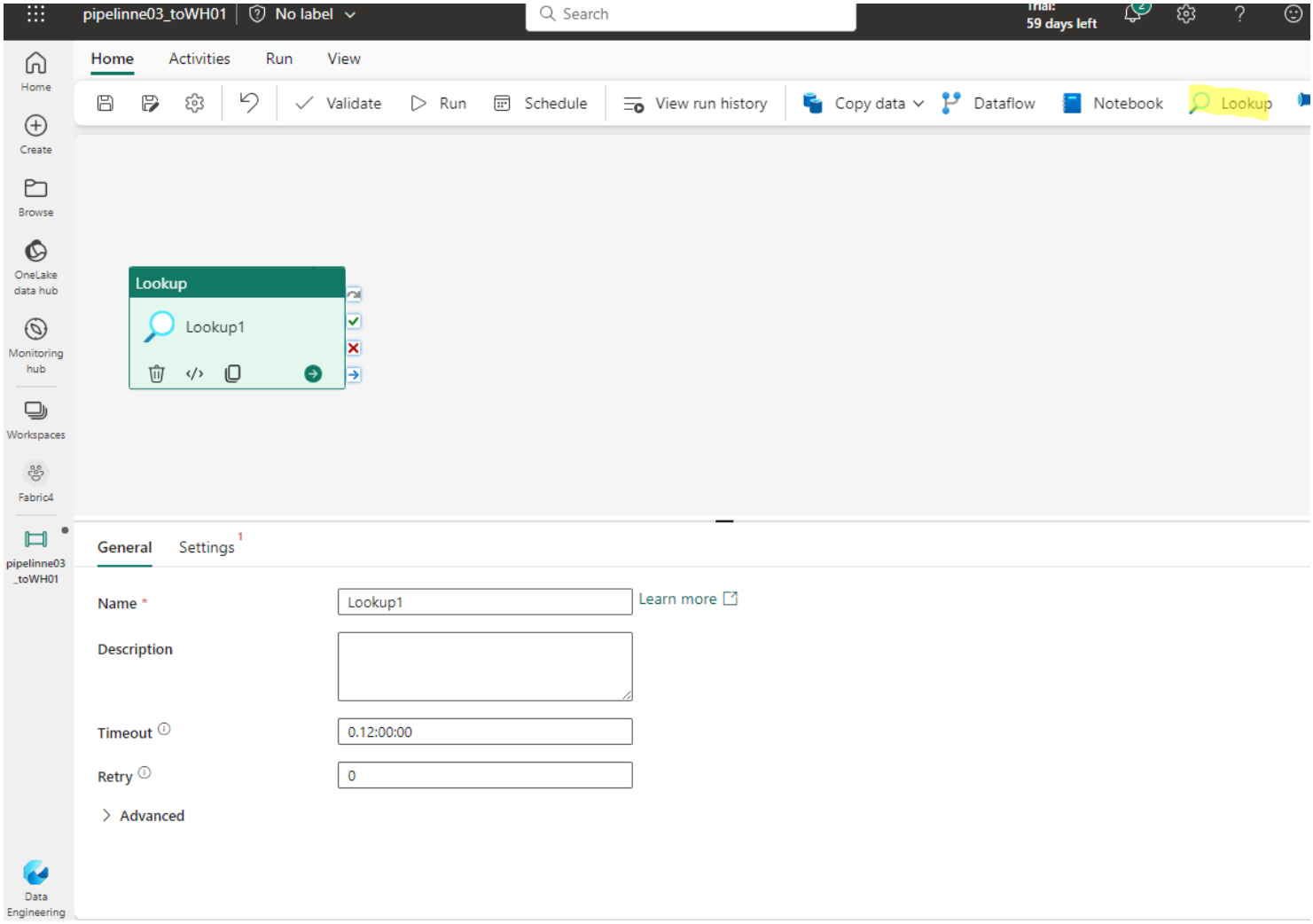
## 2.7 Create pipeline from Workspace



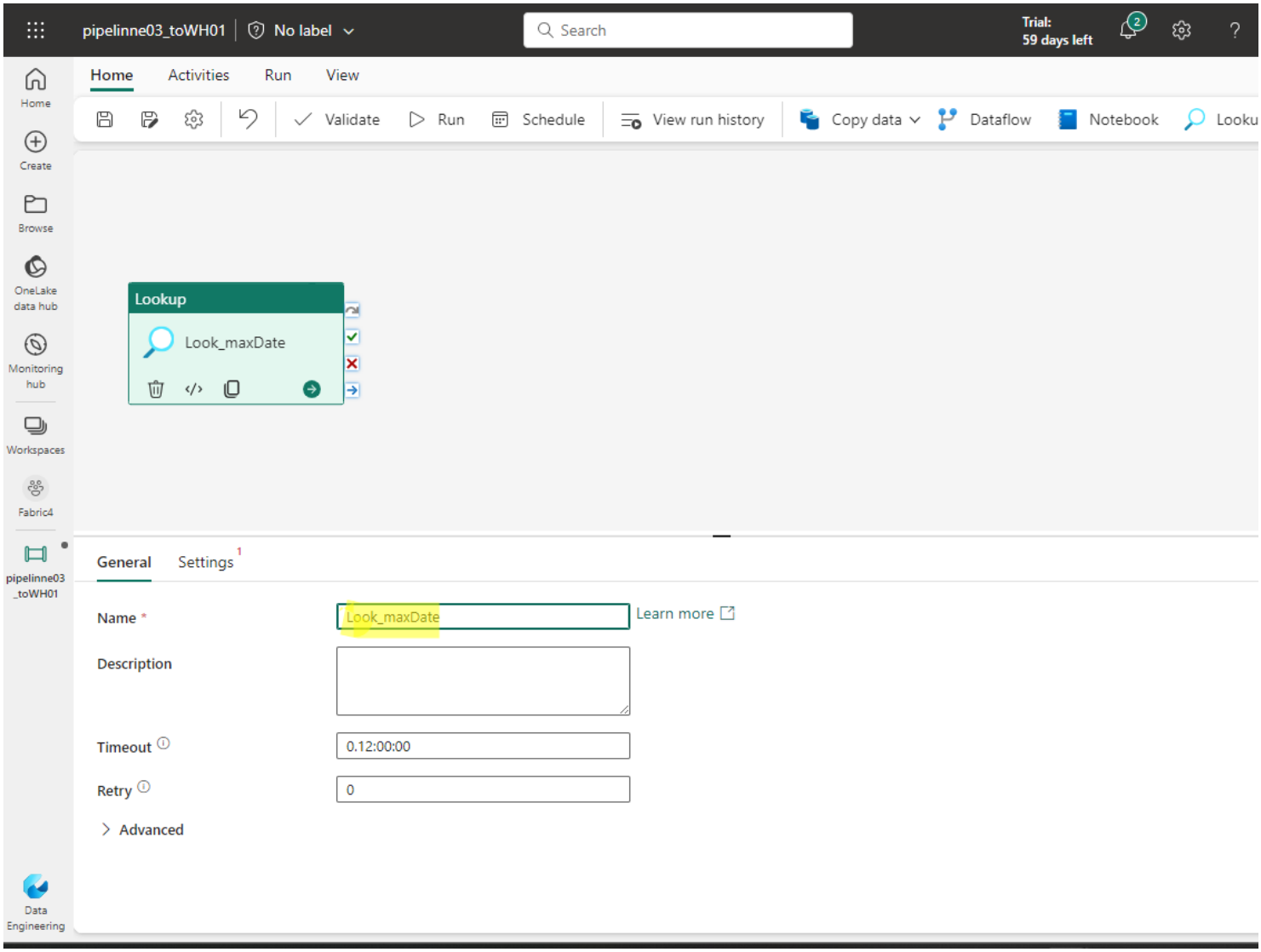



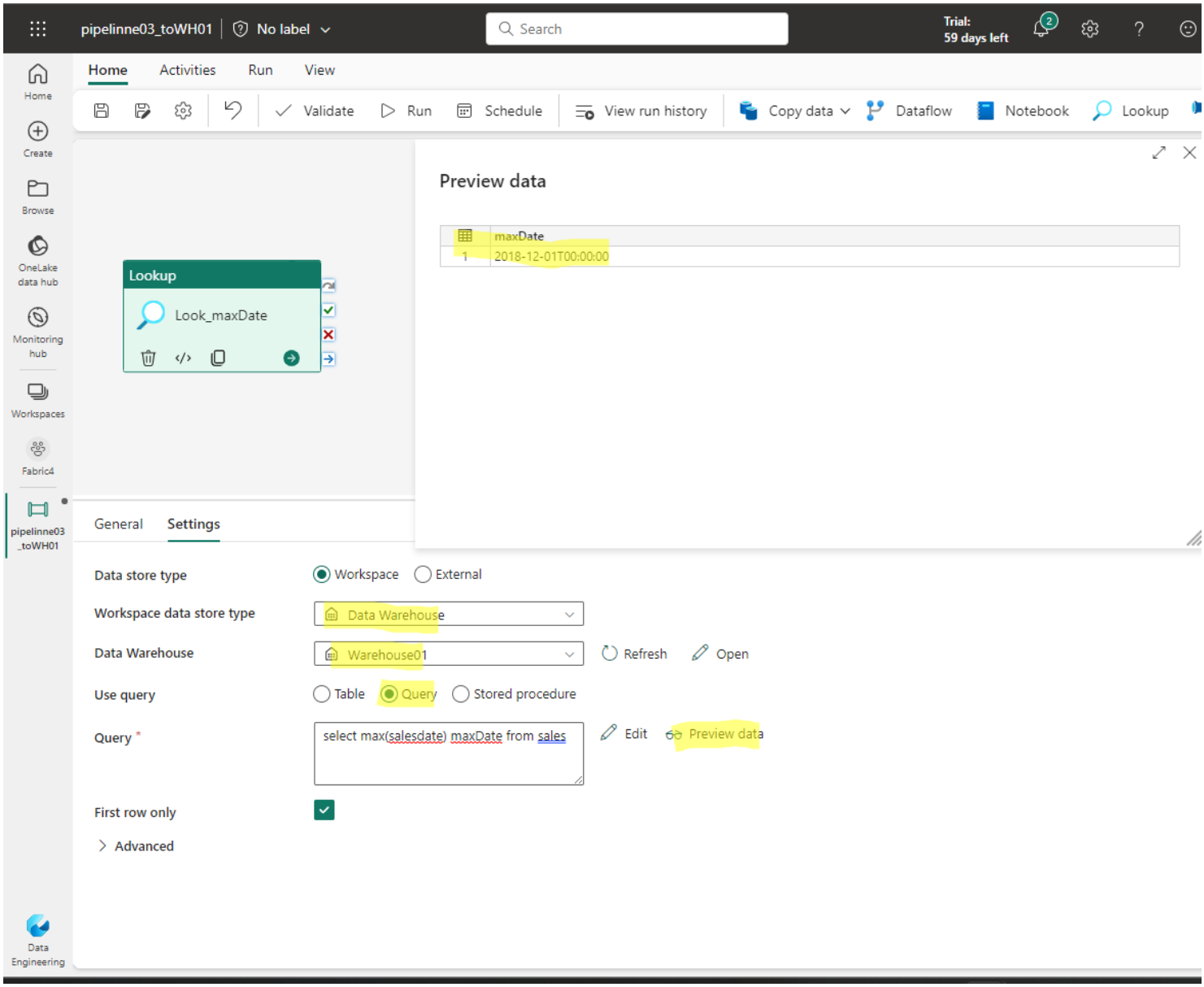



Activity 1: Lookup

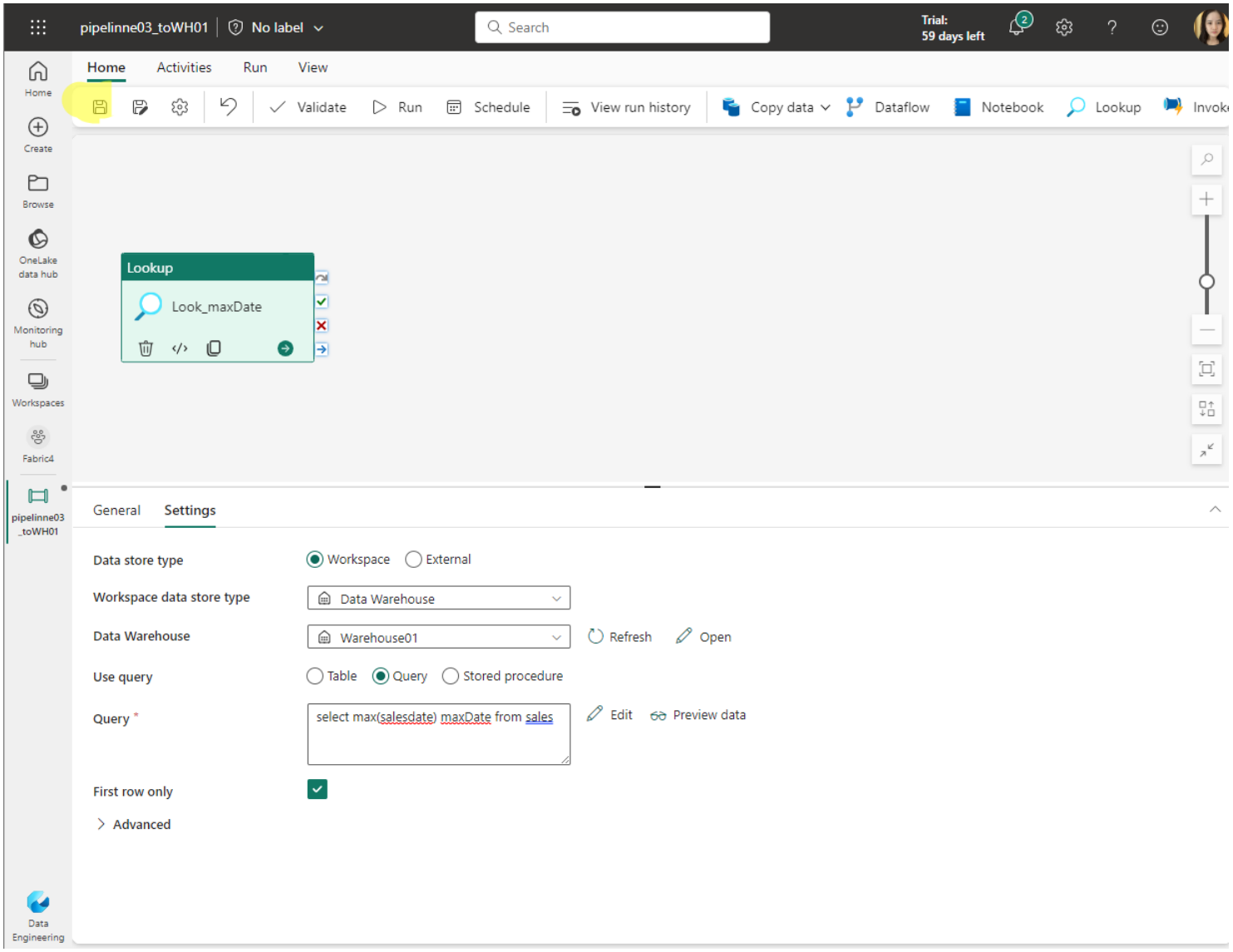



Name the activity

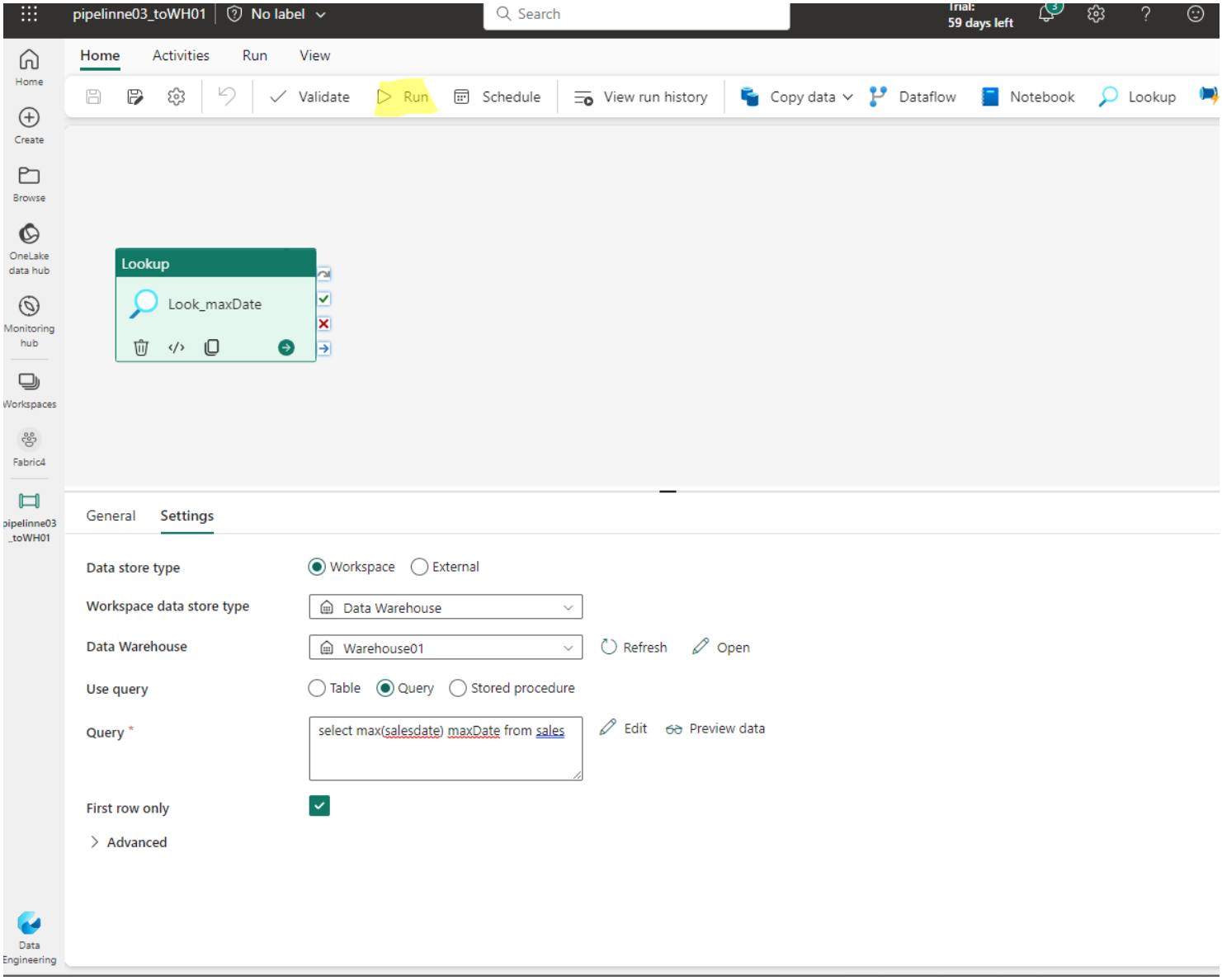

pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Looku
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03 _toWH01
Data Engineering
Lookup
Look_maxDate
General
Settings
Name *
Look_maxDate
Learn more
Description
Timeout
0.12:00:00
Retry
0
Advanced

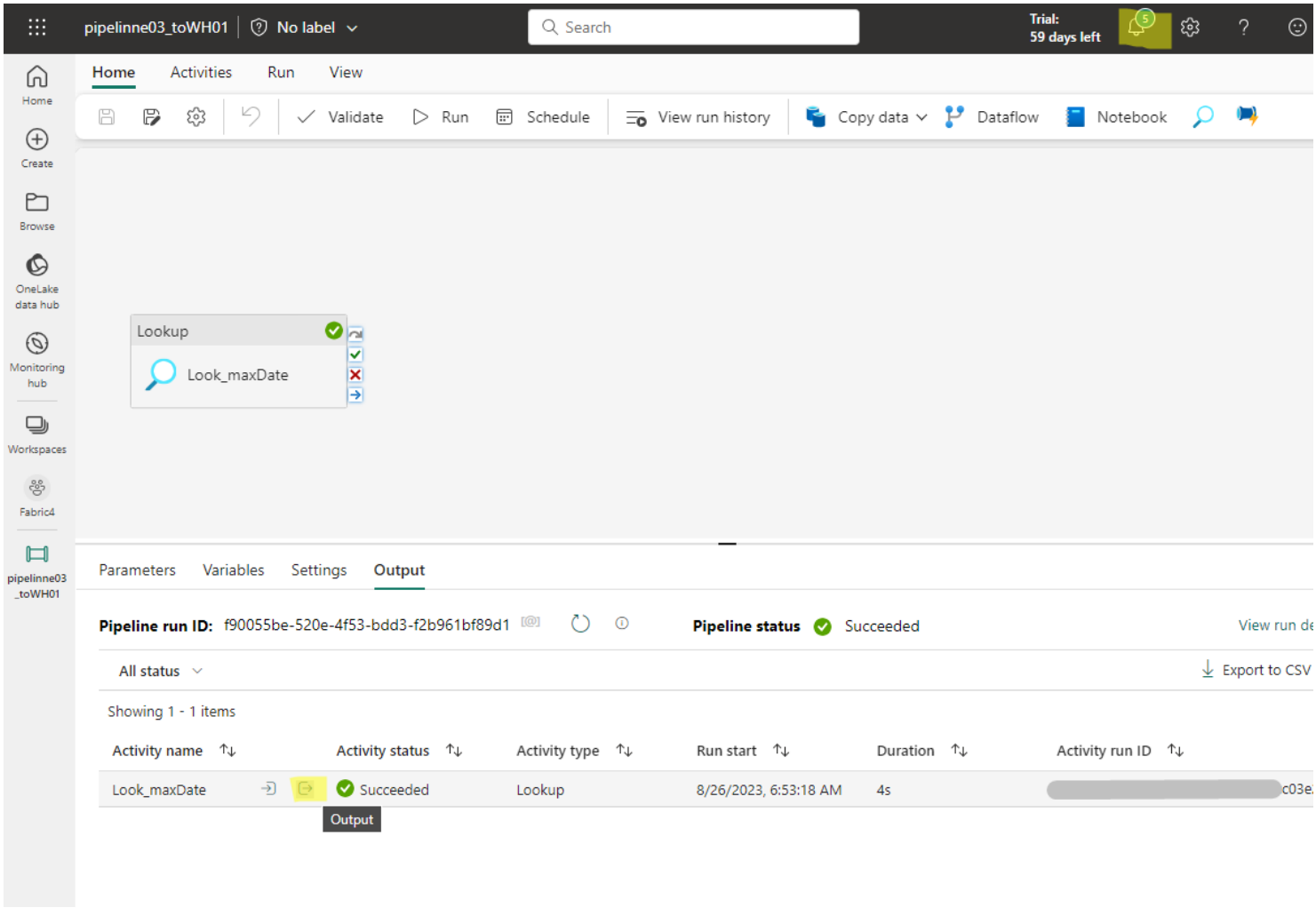



Save activity


pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Lookup
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03 _toWH01
Data Engineering
Lookup
Look_maxDate
General
Settings
Data store type
Workspace
External
Workspace data store type
Data Warehouse
Data Warehouse
Warehouse01
Refresh
Open
Use query
Table
Query
Stored procedure
Query *
select max(salesdate) maxDate from sales
Edit
Preview data
First row only
Advanced

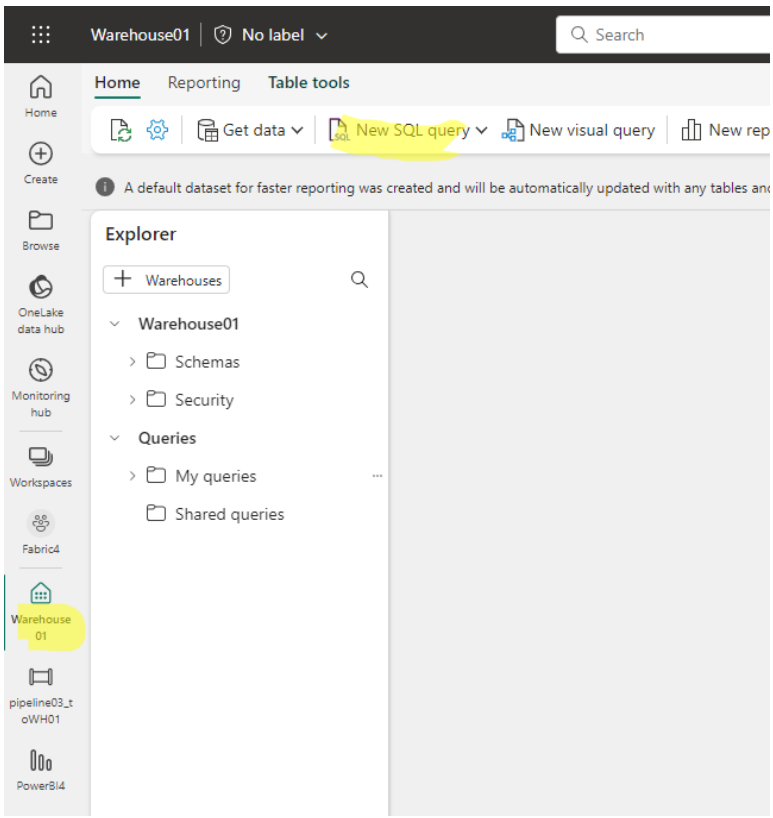
Run


pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Lookup
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03_toWH01
Data Engineering
Lookup
Look_maxDate
General
Settings
Data store type
Workspace
External
Workspace data store type
Data Warehouse
Data Warehouse
Warehouse01
Refresh
Open
Use query
Table
Query
Stored procedure
Query *
select max(salesdate) maxDate from sales
Edit
Preview data
First row only
Advanced



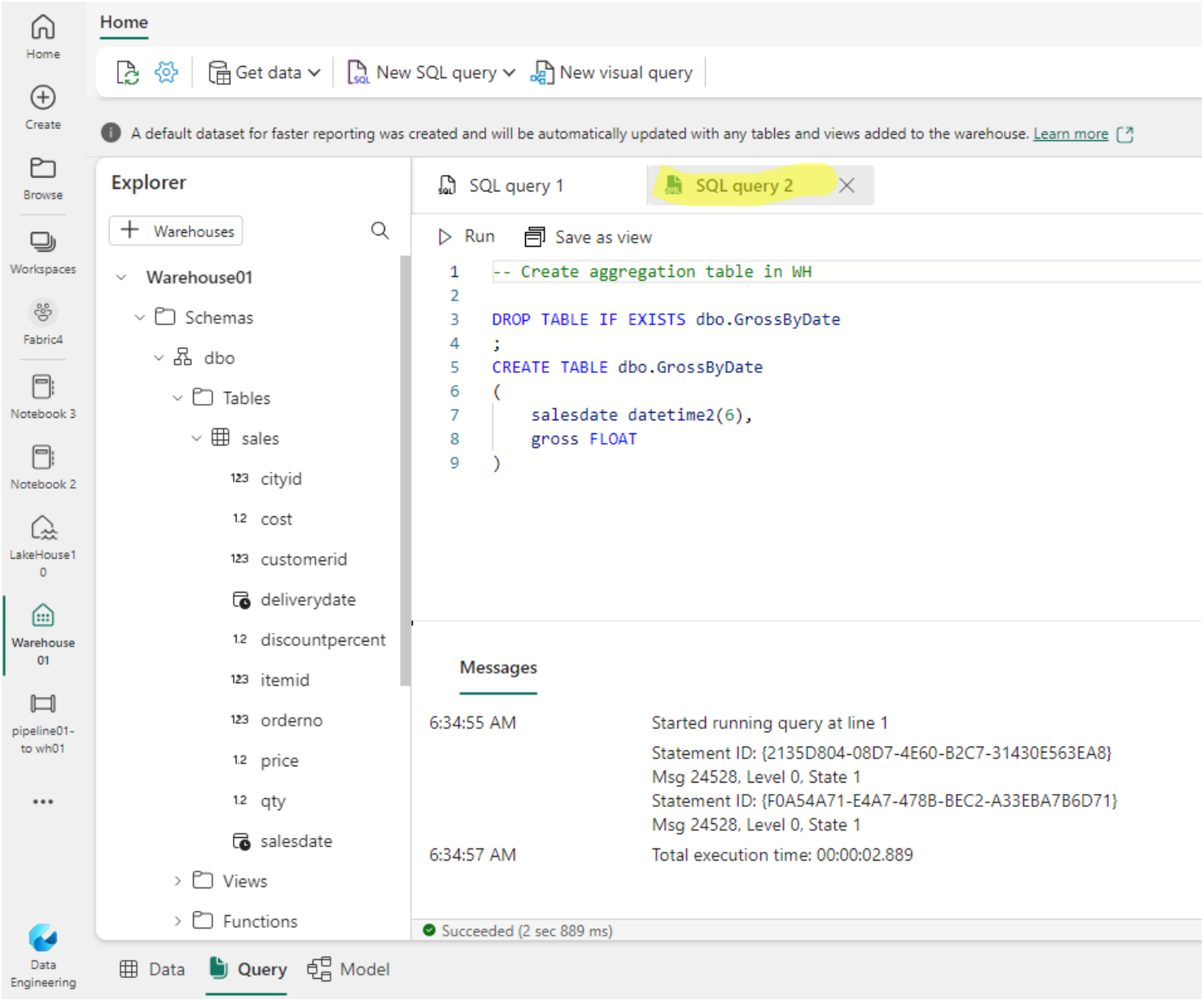

### *2.7.1 Preparation for create pipline which have many steps*

Open Warehouse




- Click New Query

## Create GrossByDate table

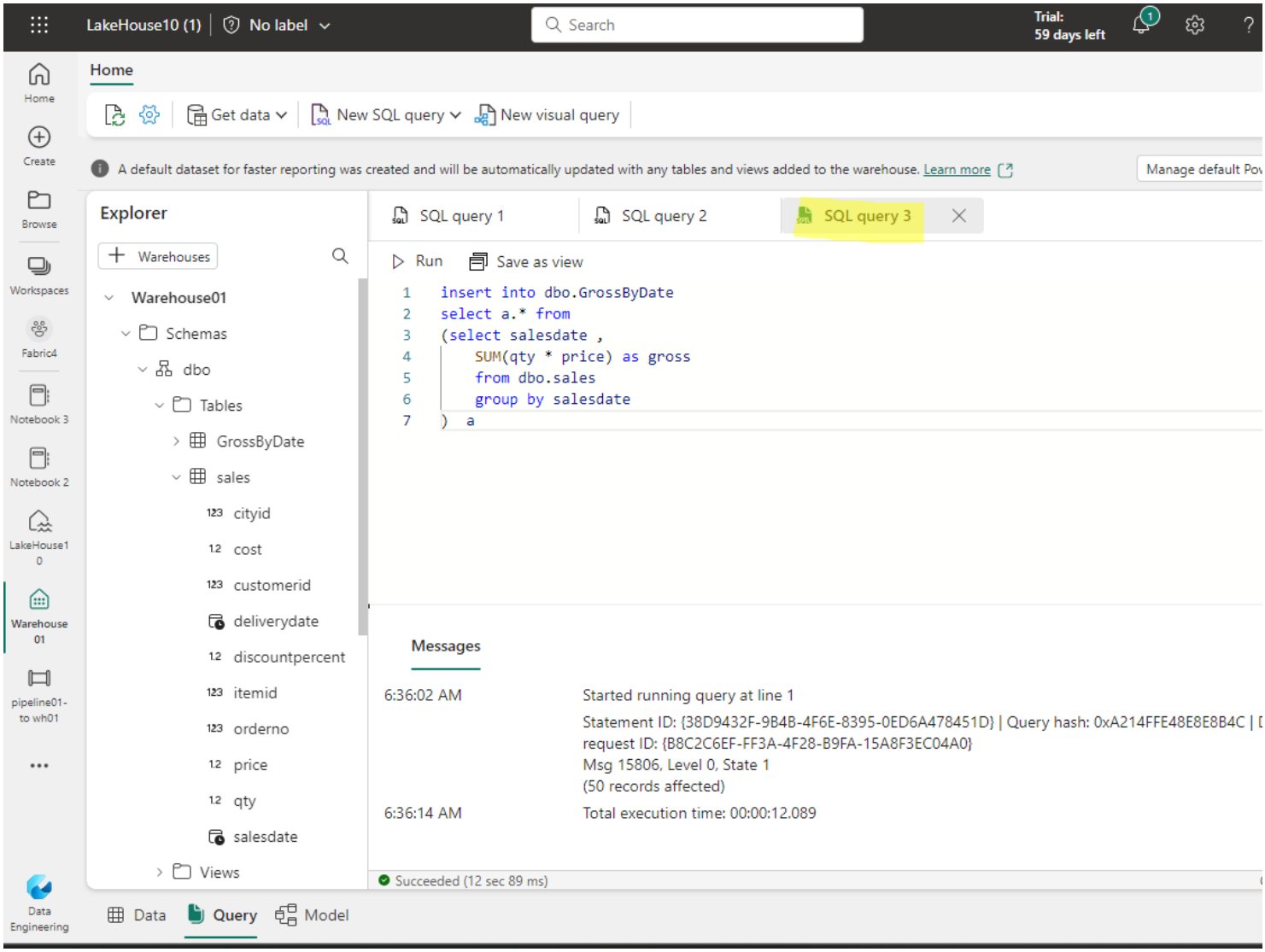



- **Input query**

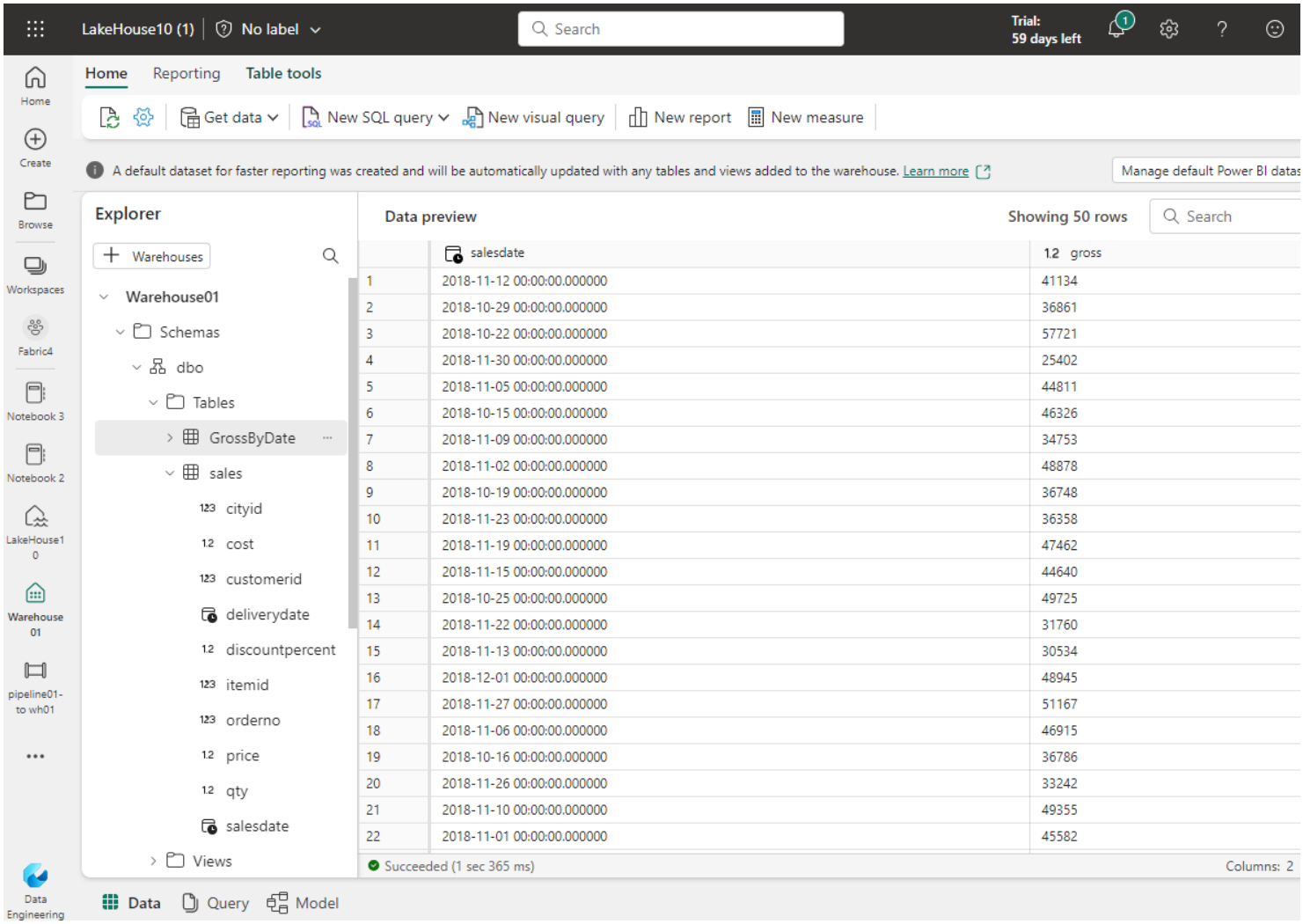
```
-- Create aggregation table in WH02


DROP TABLE IF EXISTS dbo.GrossByDate
;
CREATE TABLE dbo.GrossByDate
(
   salesdate datetime2(6),
   gross FLOAT
)
```

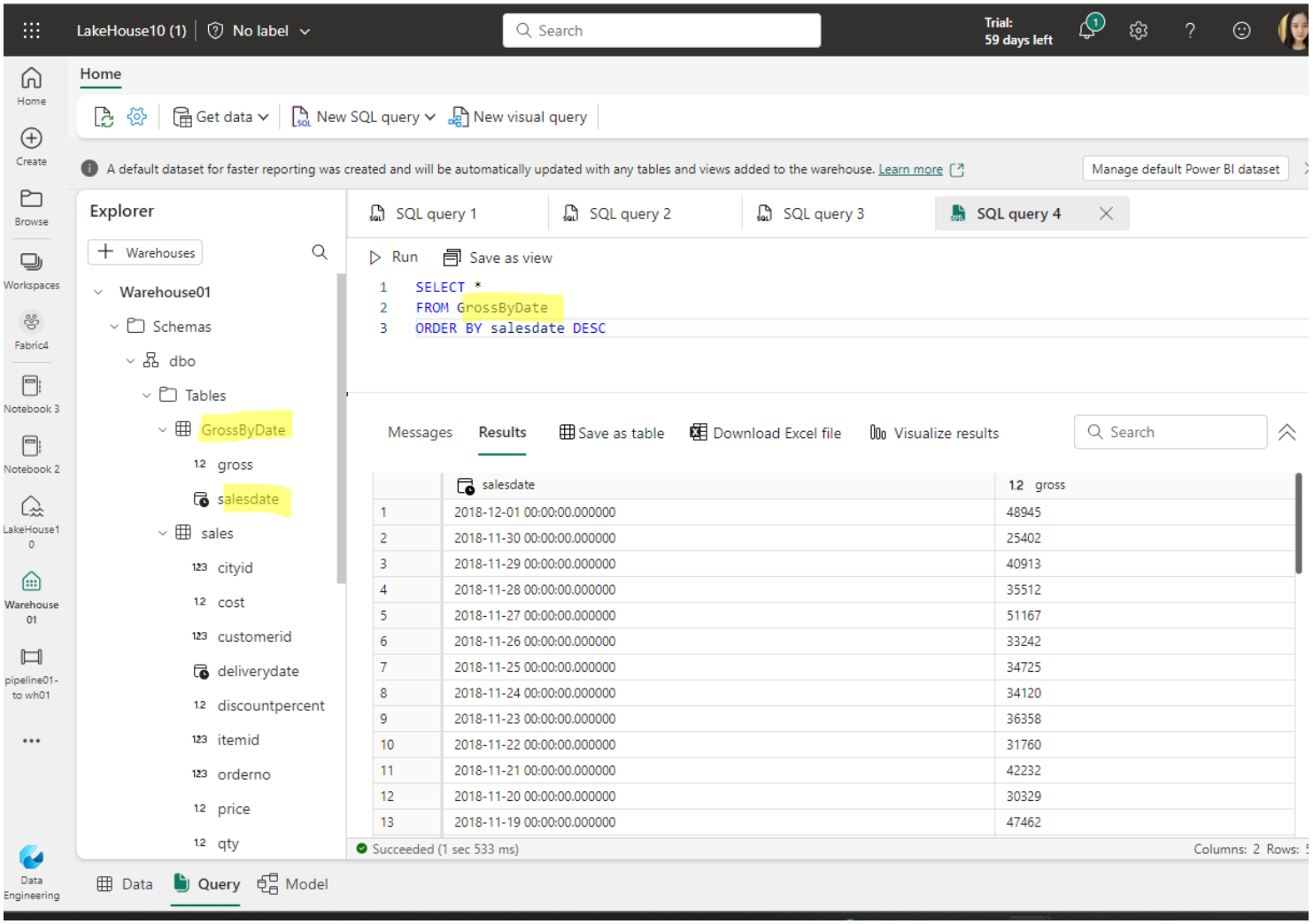
- **Run**

## Insert data

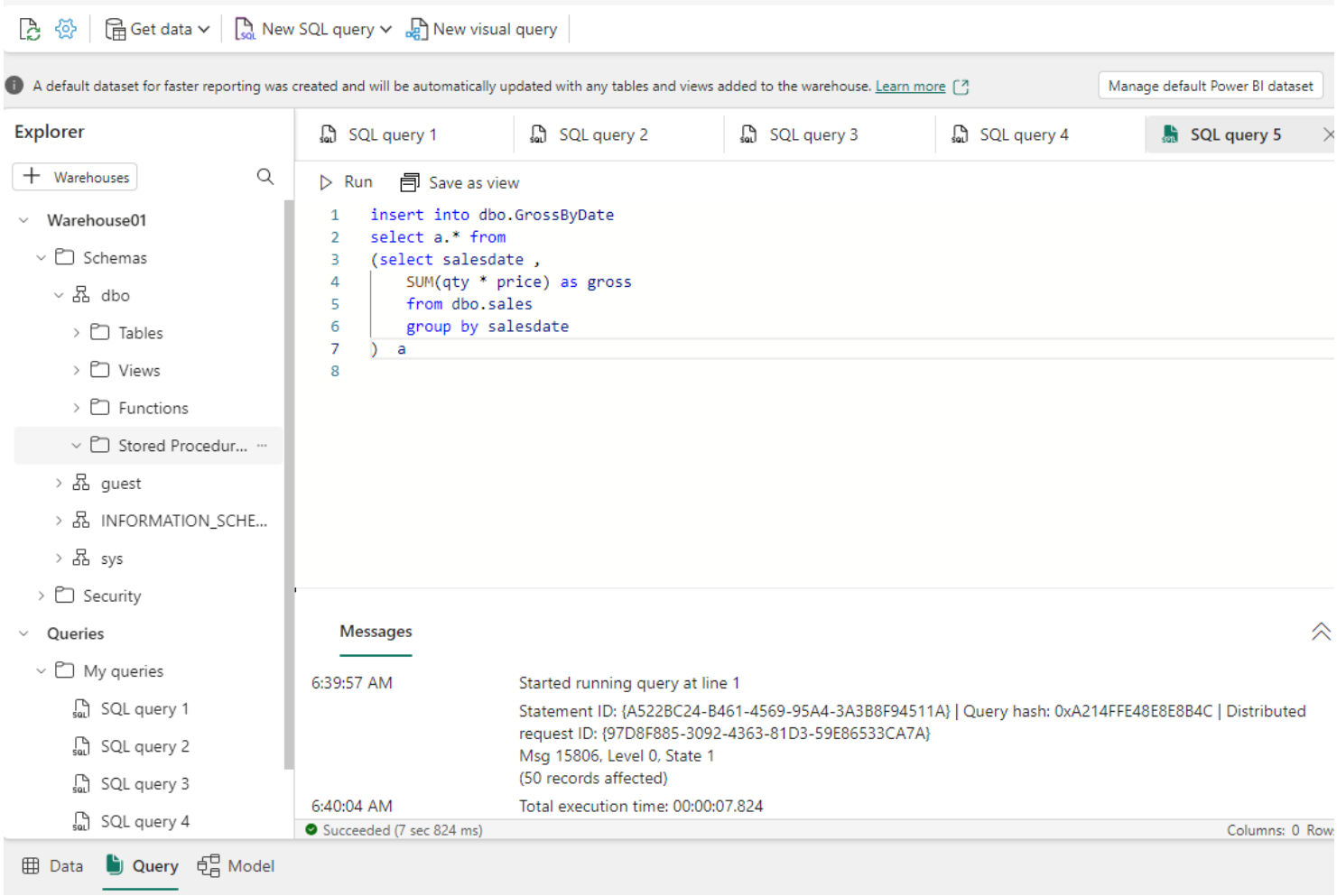


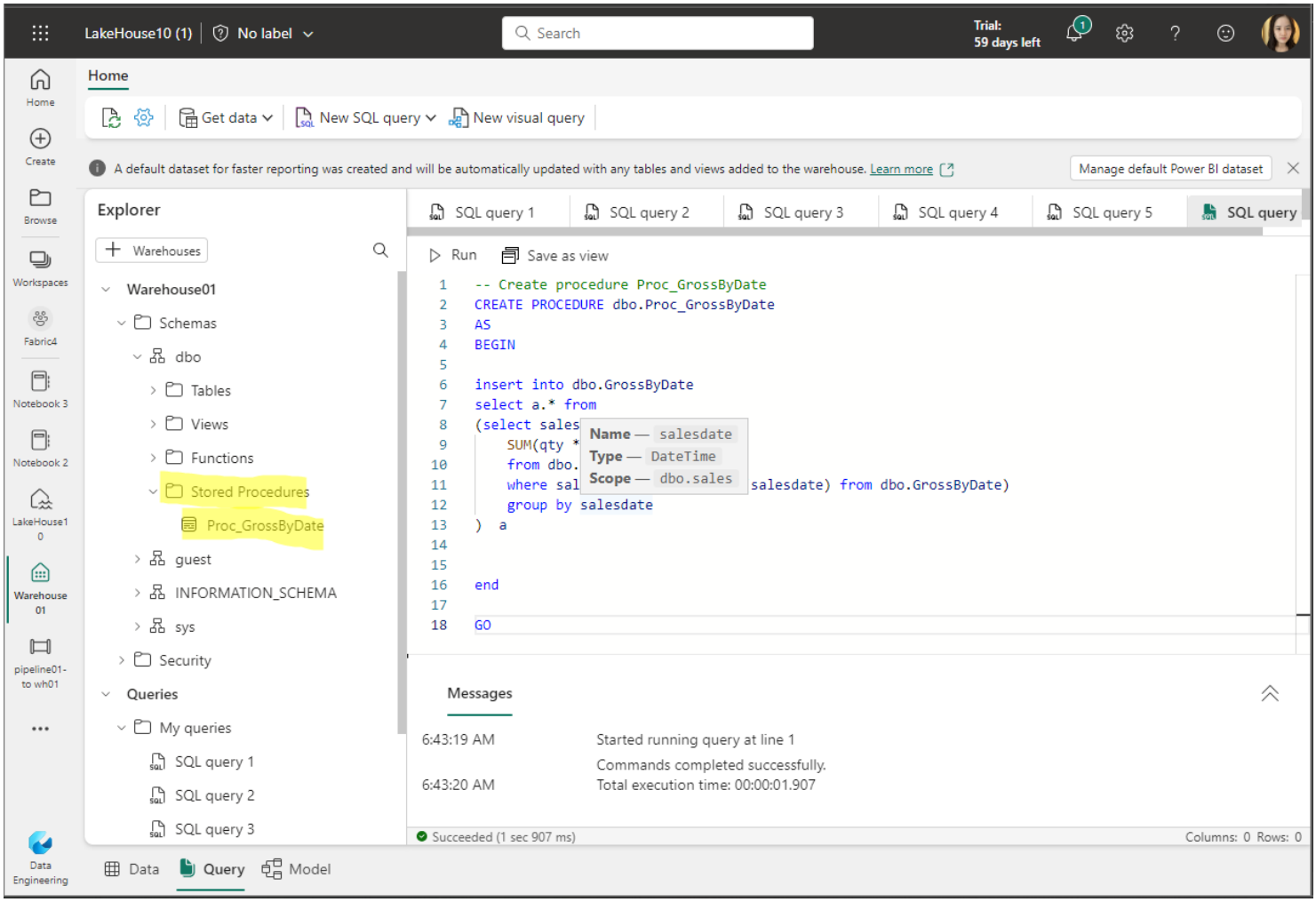

```
insert into dbo.GrossByDate
select a.* from
(select salesdate ,
   SUM(qty * price) as gross
   from dbo.sales
   group by salesdate
) a
```


LakeHouse10 (1)
No label
Search
Trial:
59 days left
Home
Reporting
Table tools
Get data
New SQL query
New visual query
New report
New measure
A default dataset for faster reporting was created and will be automatically updated with any tables and views added to the warehouse. Learn more
Manage default Power BI datas
Explorer
Warehouses
Warehouse01
Schemas
dbo
Tables
GrossByDate
sales
cityid
cost
customerid
deliverydate
discountpercent
itemid
orderno
price
qty
salesdate
Views
Data preview
Showing 50 rows
Search
salesdate
gross
1 2018-11-12 00:00:00.000000 41134
2 2018-10-29 00:00:00.000000 36861
3 2018-10-22 00:00:00.000000 57721
4 2018-11-30 00:00:00.000000 25402
5 2018-11-05 00:00:00.000000 44811
6 2018-10-15 00:00:00.000000 46326
7 2018-11-09 00:00:00.000000 34753
8 2018-11-02 00:00:00.000000 48878
9 2018-10-19 00:00:00.000000 36748
10 2018-11-23 00:00:00.000000 36358
11 2018-11-19 00:00:00.000000 47462
12 2018-11-15 00:00:00.000000 44640
13 2018-10-25 00:00:00.000000 49725
14 2018-11-22 00:00:00.000000 31760
15 2018-11-13 00:00:00.000000 30534
16 2018-12-01 00:00:00.000000 48945
17 2018-11-27 00:00:00.000000 51167
18 2018-11-06 00:00:00.000000 46915
19 2018-10-16 00:00:00.000000 36786
20 2018-11-26 00:00:00.000000 33242
21 2018-11-10 00:00:00.000000 49355
22 2018-11-01 00:00:00.000000 45582
Succeeded (1 sec 365 ms)
Columns: 2
Data
Query
Model
Home
Create
Browse
Workspaces
Fabric4
Notebook 3
Notebook 2
LakeHouse10
Warehouse01
pipeline01-to wh01
Data Engineering


LakeHouse10 (1)
No label
Search
Trial:
59 days left
Home
Get data
New SQL query
New visual query
A default dataset for faster reporting was created and will be automatically updated with any tables and views added to the warehouse. Learn more
Manage default Power BI dataset
Explorer
Warehouses
Warehouse01
Schemas
dbo
Tables
GrossByDate
gross
salesdate
sales
cityid
cost
customerid
deliverydate
discountpercent
itemid
orderno
price
qty
SQL query 1
SQL query 2
SQL query 3
SQL query 4
Run
Save as view
1 SELECT *
2 FROM GrossByDate
3 ORDER BY salesdate DESC
Messages
Results
Save as table
Download Excel file
Visualize results
Search
salesdate
gross
1 2018-12-01 00:00:00.000000 48945
2 2018-11-30 00:00:00.000000 25402
3 2018-11-29 00:00:00.000000 40913
4 2018-11-28 00:00:00.000000 35512
5 2018-11-27 00:00:00.000000 51167
6 2018-11-26 00:00:00.000000 33242
7 2018-11-25 00:00:00.000000 34725
8 2018-11-24 00:00:00.000000 34120
9 2018-11-23 00:00:00.000000 36358
10 2018-11-22 00:00:00.000000 31760
11 2018-11-21 00:00:00.000000 42232
12 2018-11-20 00:00:00.000000 30329
13 2018-11-19 00:00:00.000000 47462
Succeeded (1 sec 533 ms)
Columns: 2 Rows:
Data
Query
Model
Data Engineering

## Create storeprocedure

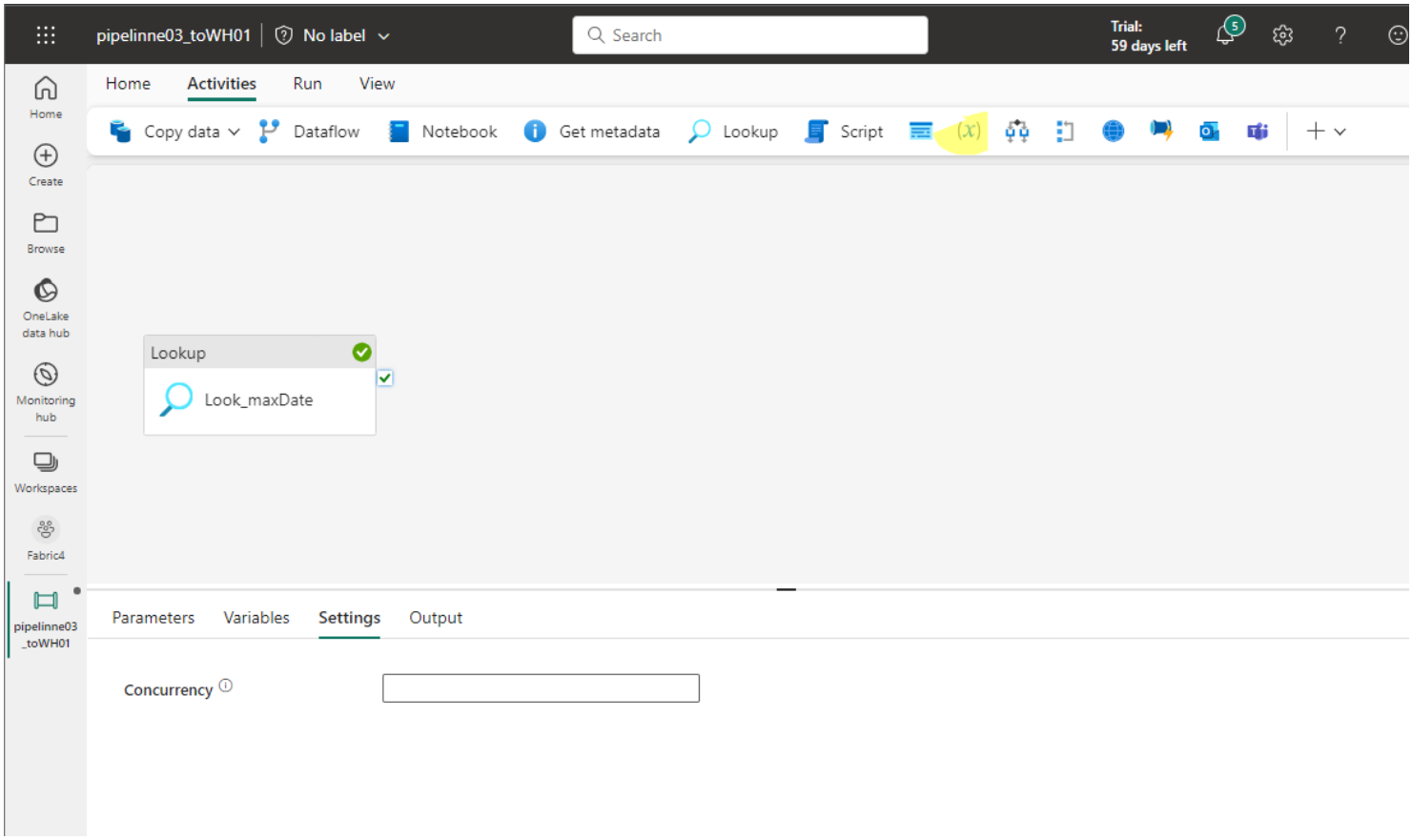


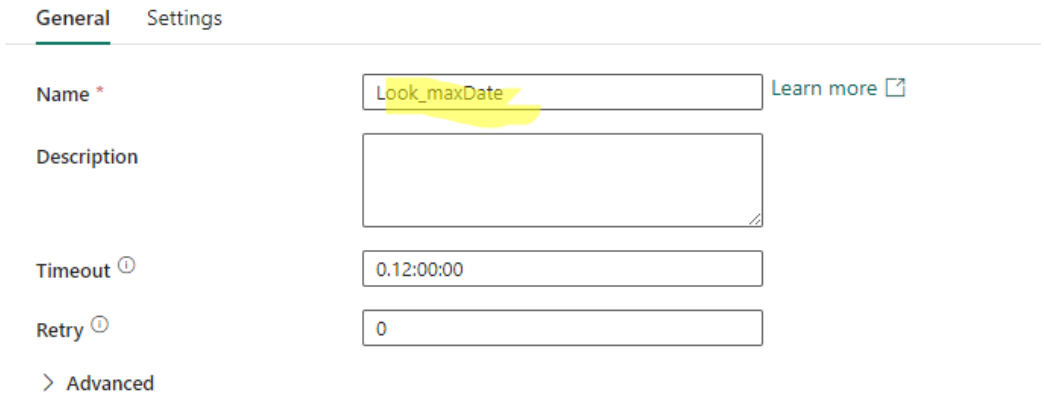



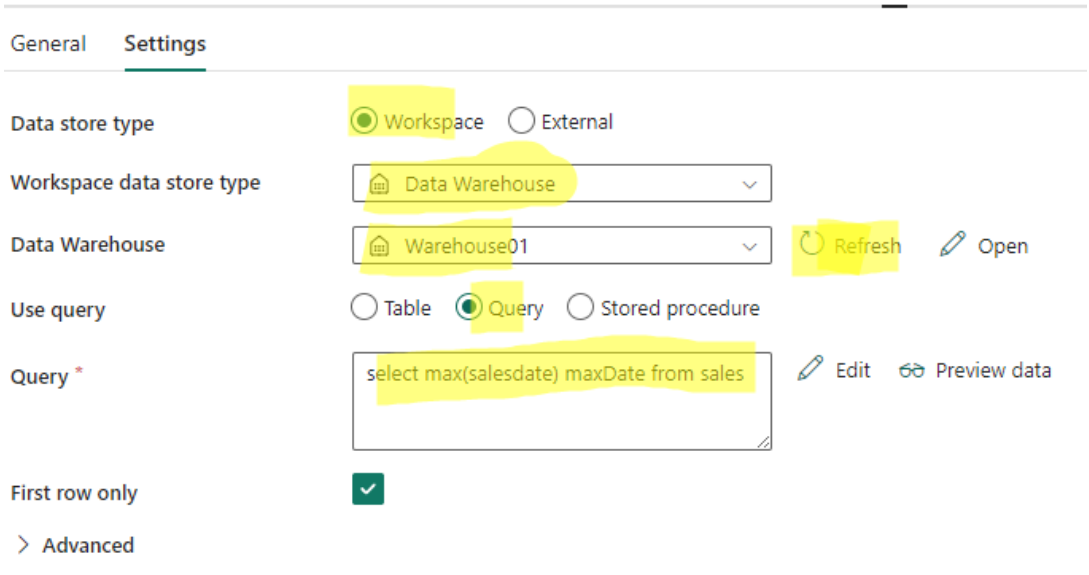
```
-- Create procedure Proc_GrossByDate
CREATE PROCEDURE dbo.Proc_GrossByDate
AS
BEGIN

insert into dbo.GrossByDate
select a.* from
(select salesdate ,
   SUM(qty * price) as gross
   from dbo.sales
   where salesdate > (select max(salesdate) from dbo.GrossByDate)
   group by salesdate
)  a


end

GO
```

*2.7.2 Set Pipeline*

Step 1: Lookup



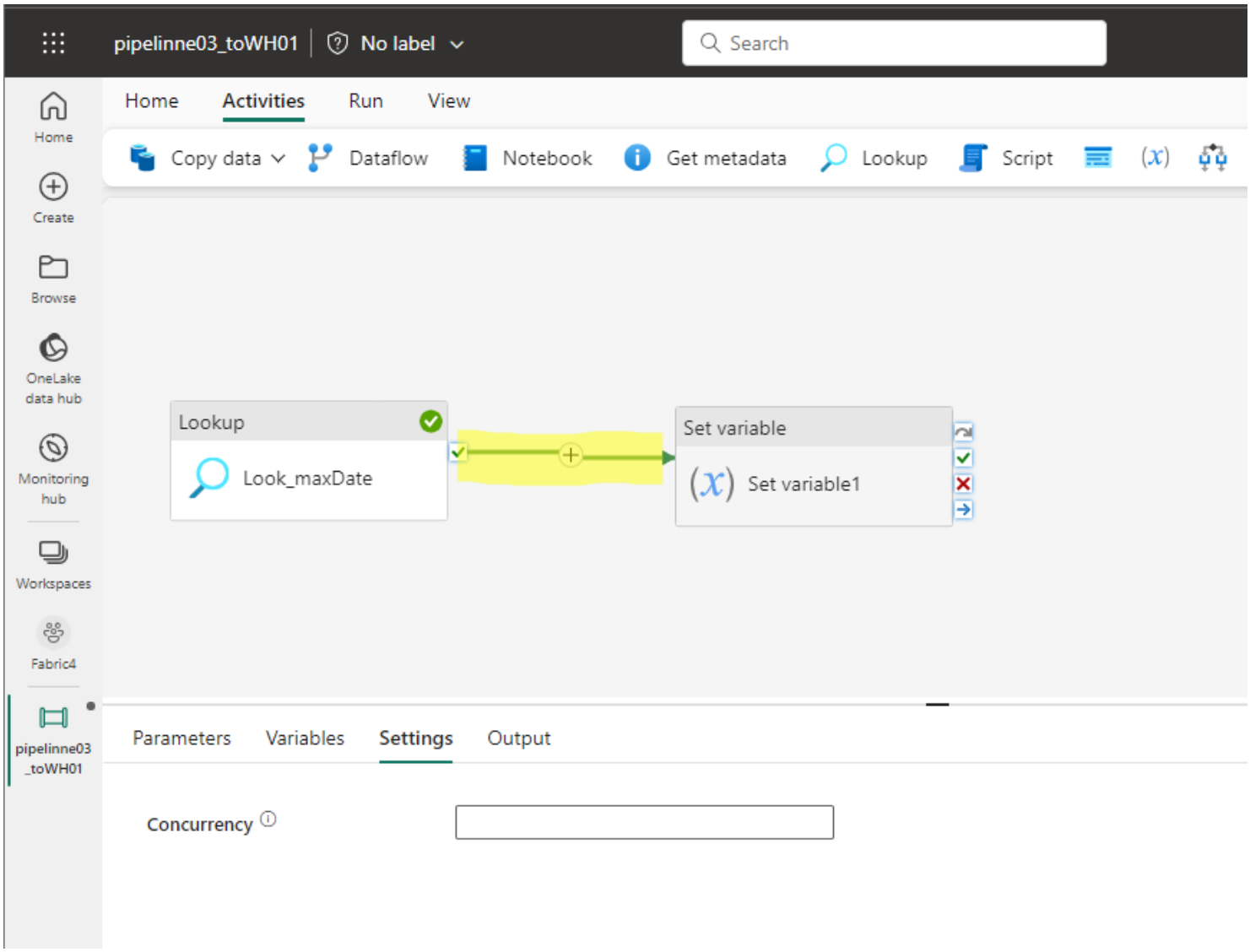



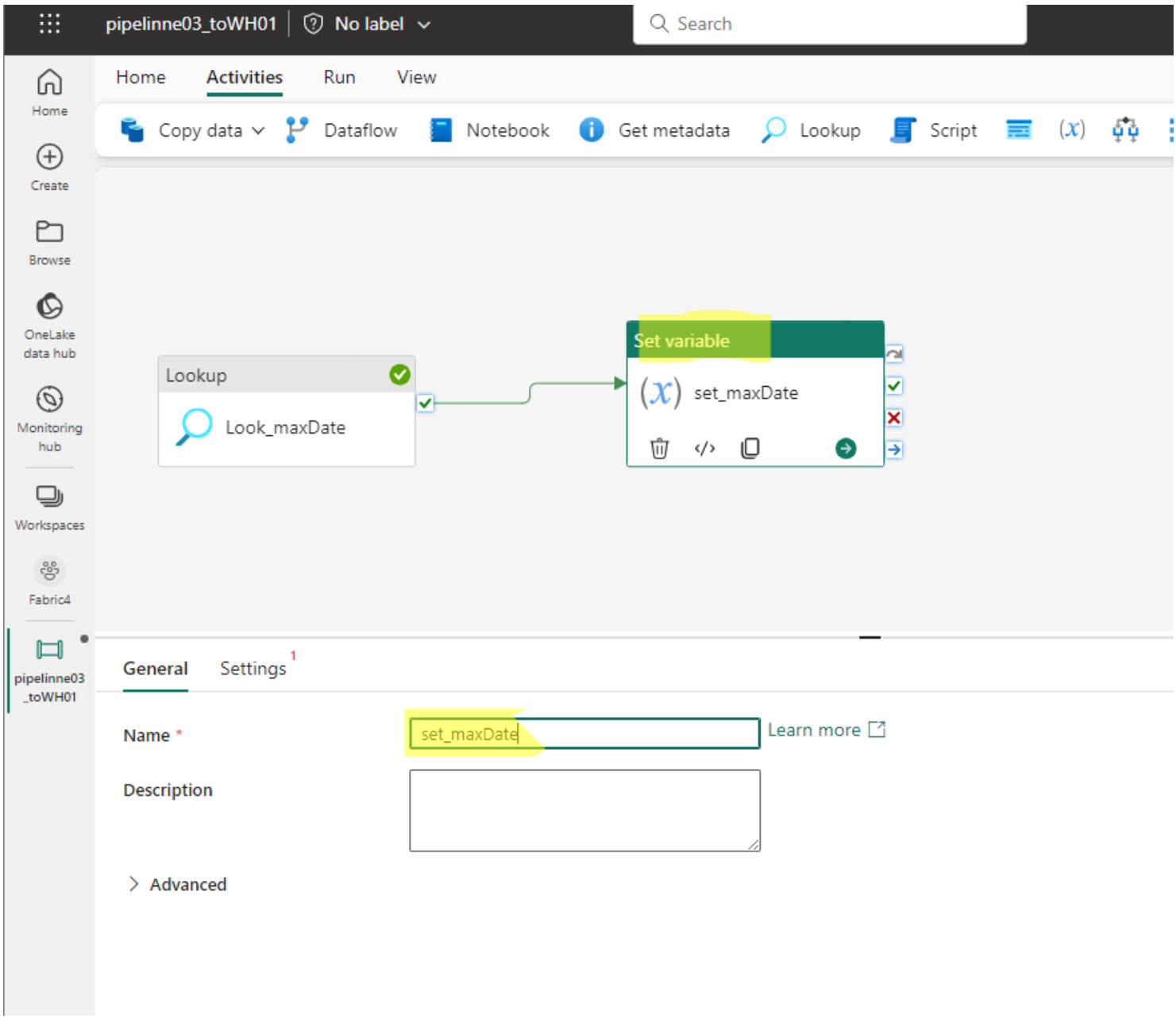



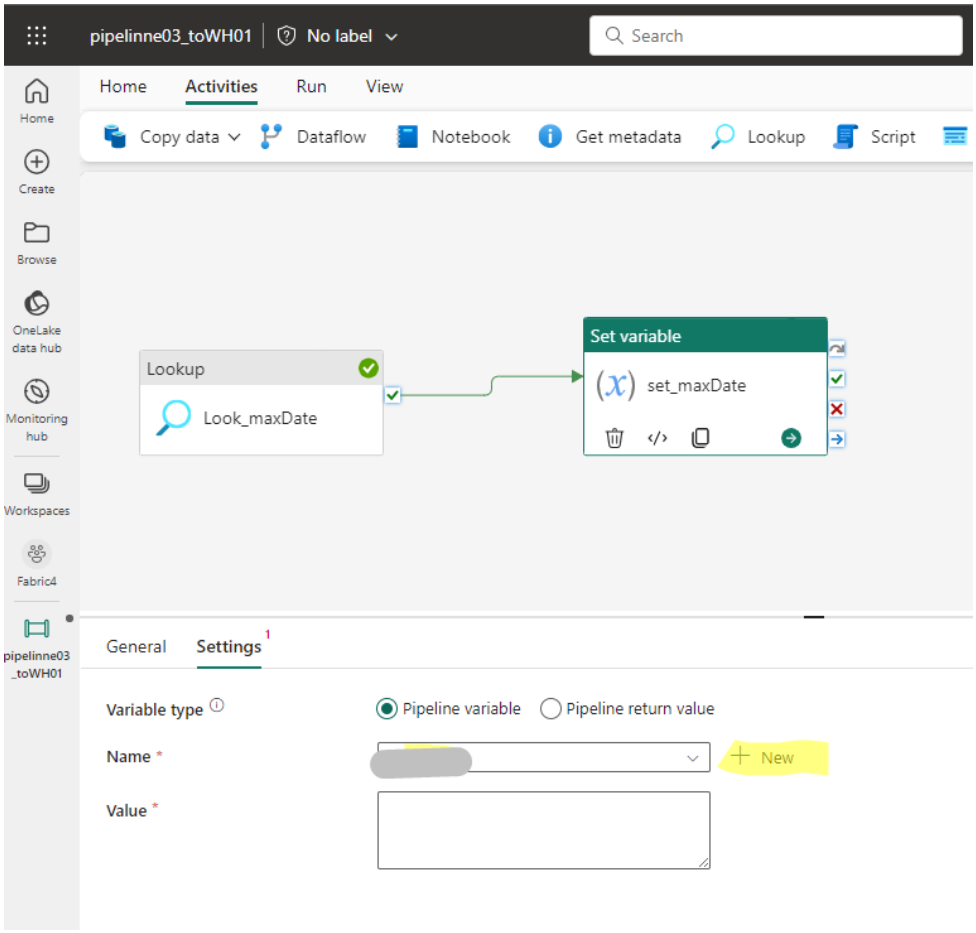

```
select max(salesdate) maxDate from sales
```

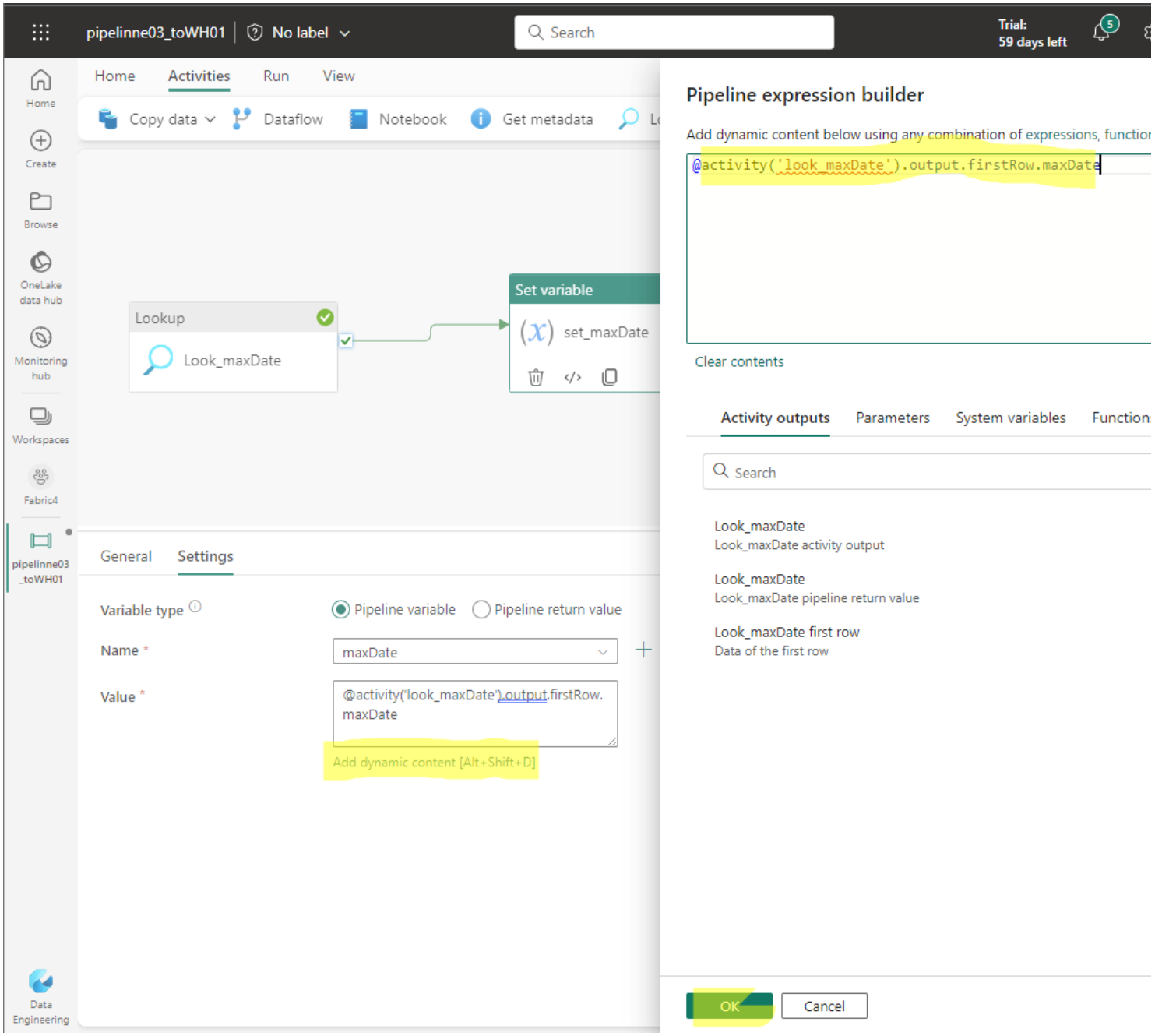
After complete the step, Save > validate  > run  to check if any error in the setup

## Step 2: Set variable

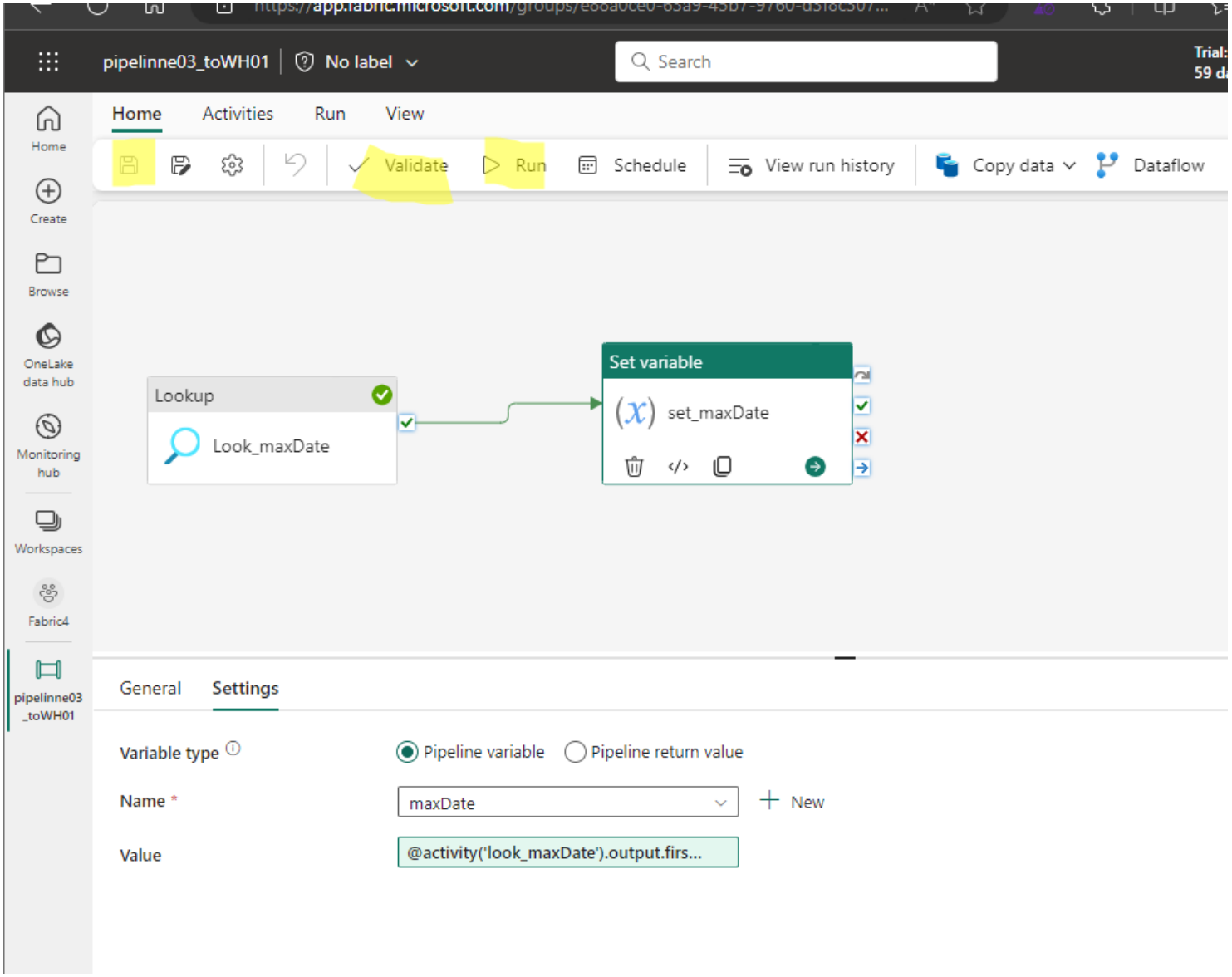



pipelinne03_toWH01
No label
Search
Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03_toWH01
Lookup
Look_maxDate
Set variable
set_maxDate
General
Settings
Name *
set_maxDate
Learn more
Description
Advanced

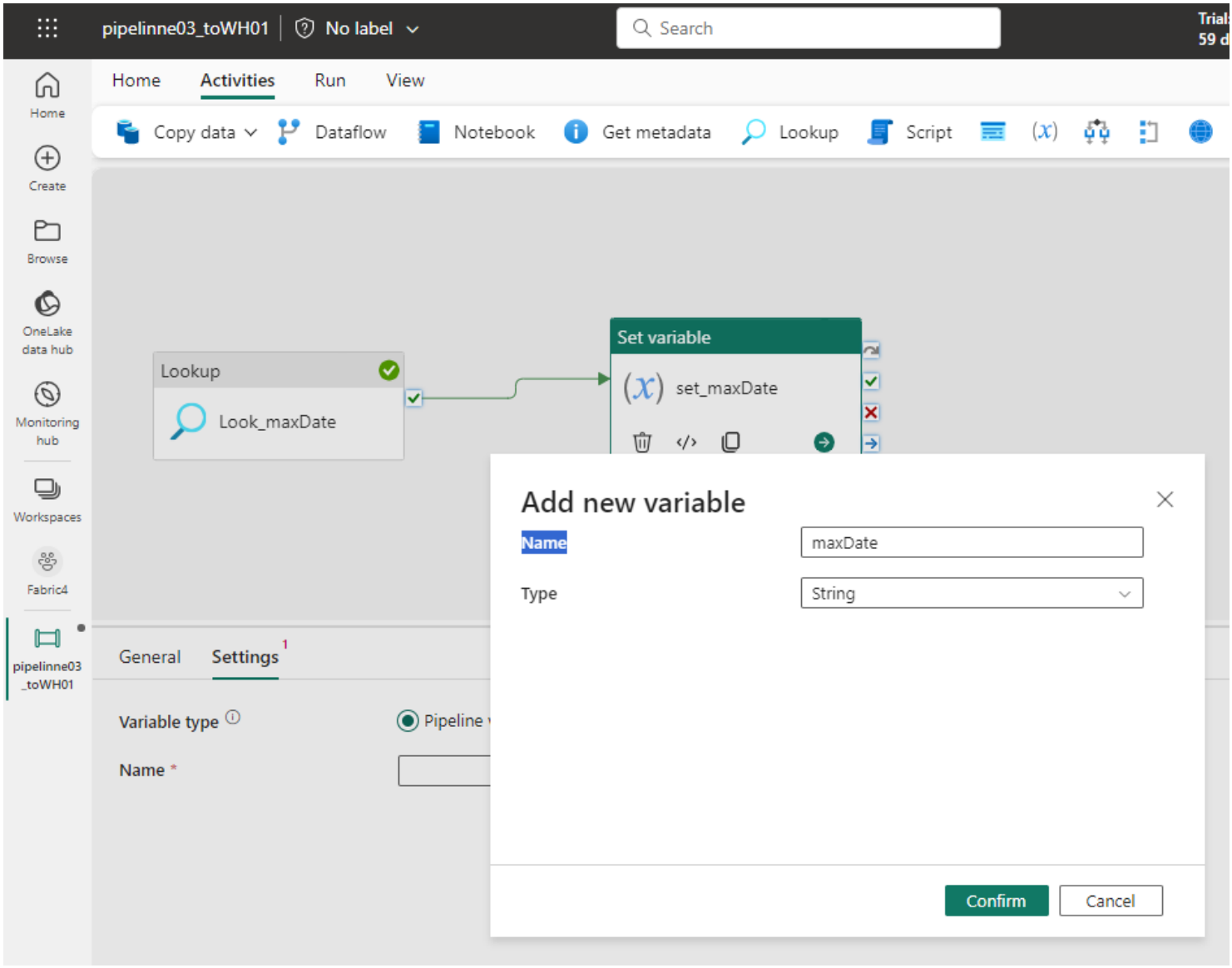

pipelinne03_toWH01
No label
Search
Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03 _toWH01
Lookup
Look_maxDate
Set variable
set_maxDate
General
Settings
Variable type
Pipeline variable
Pipeline return value
Name *
New
Value *



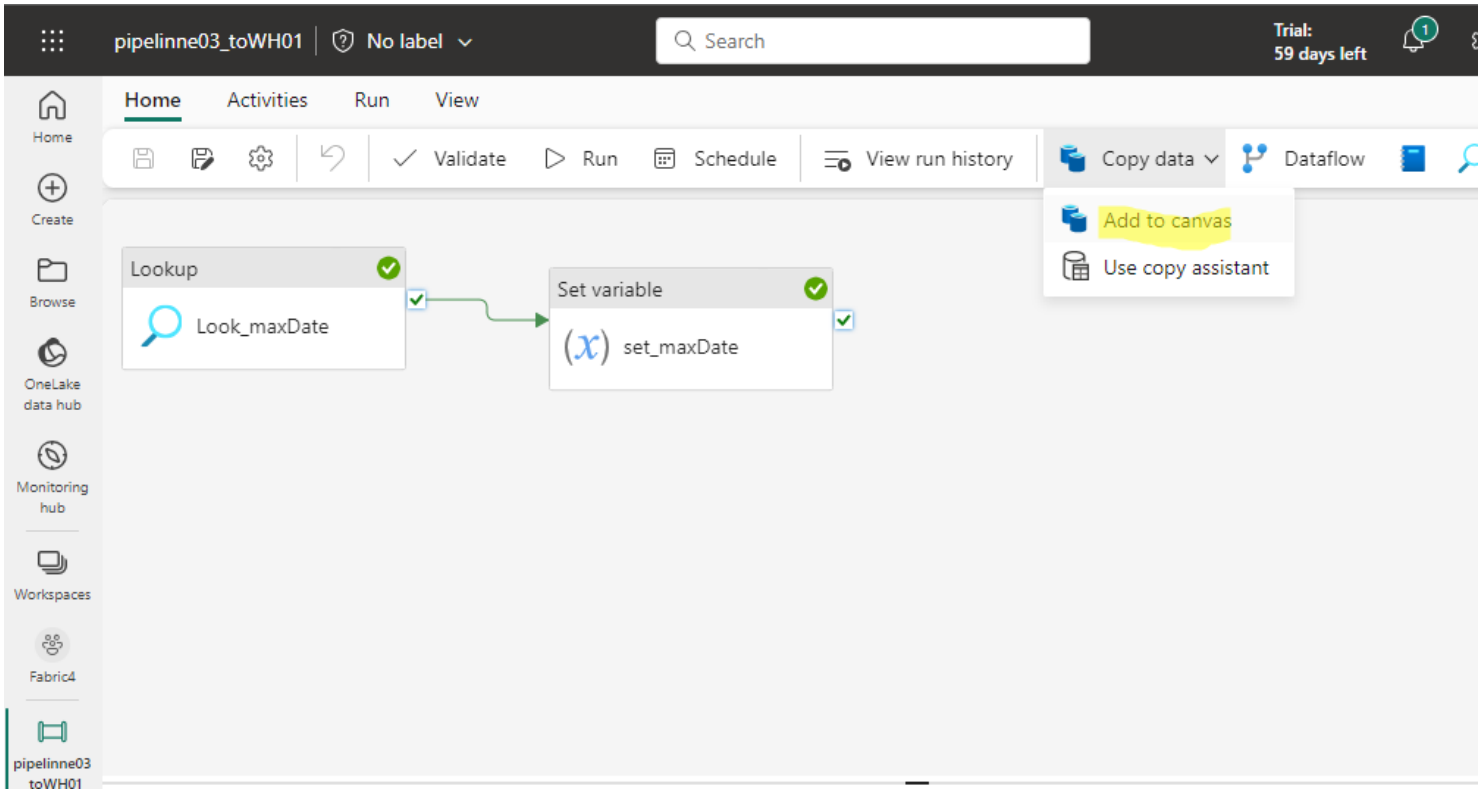

```
@activity('look_maxDate').output.firstRow.maxDate
```

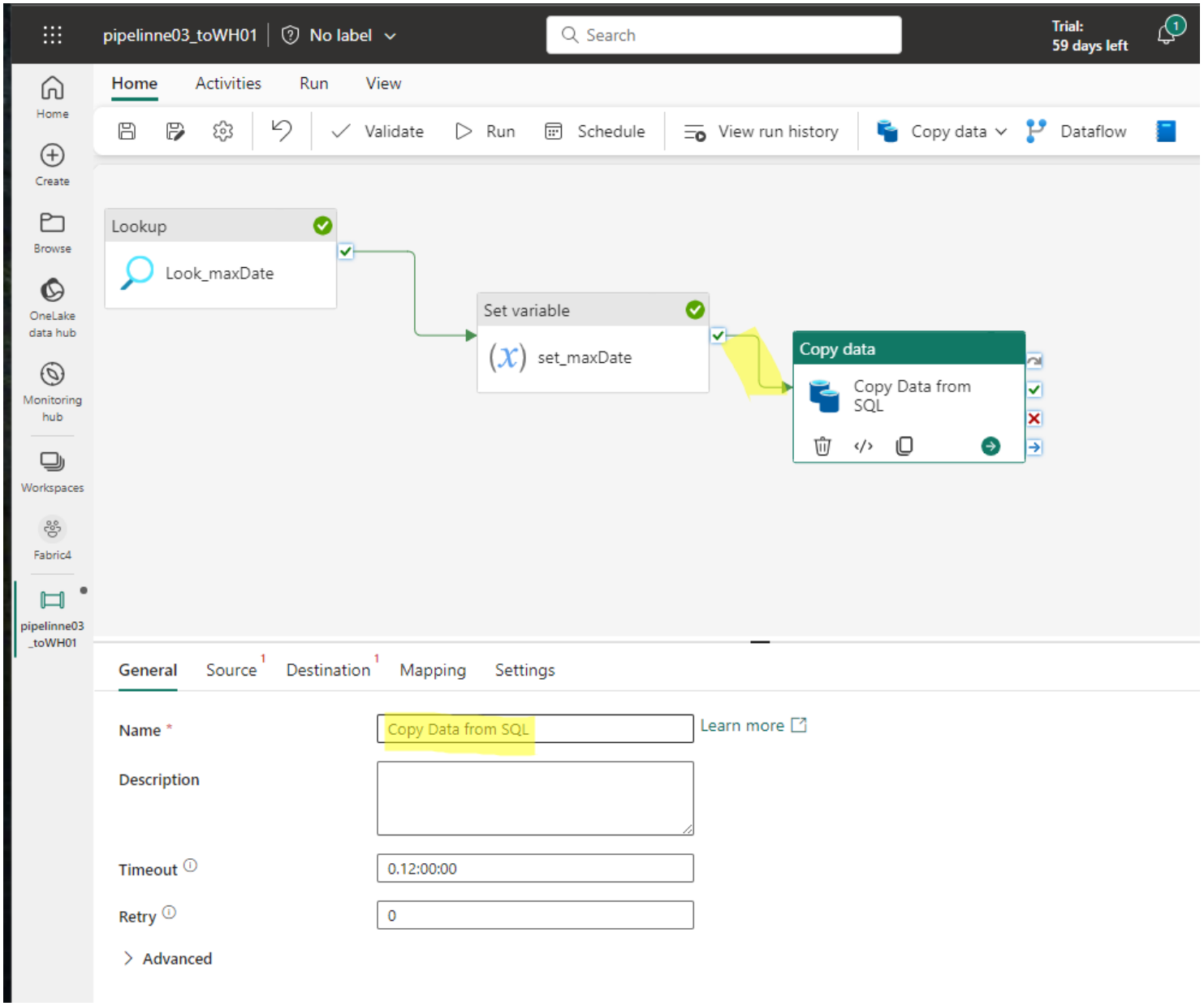


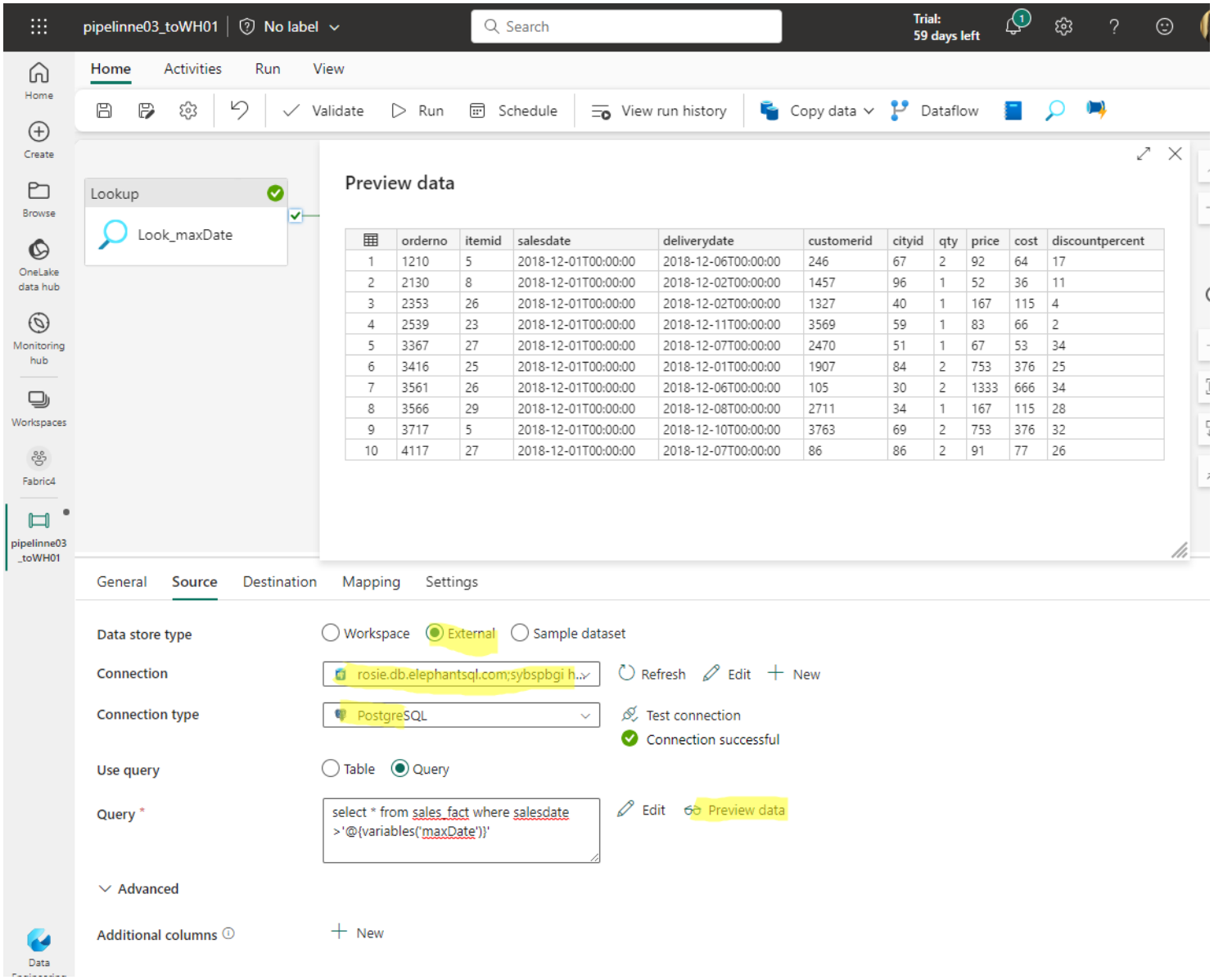

After complete the step, Save > validate > run to check if any error in the setup

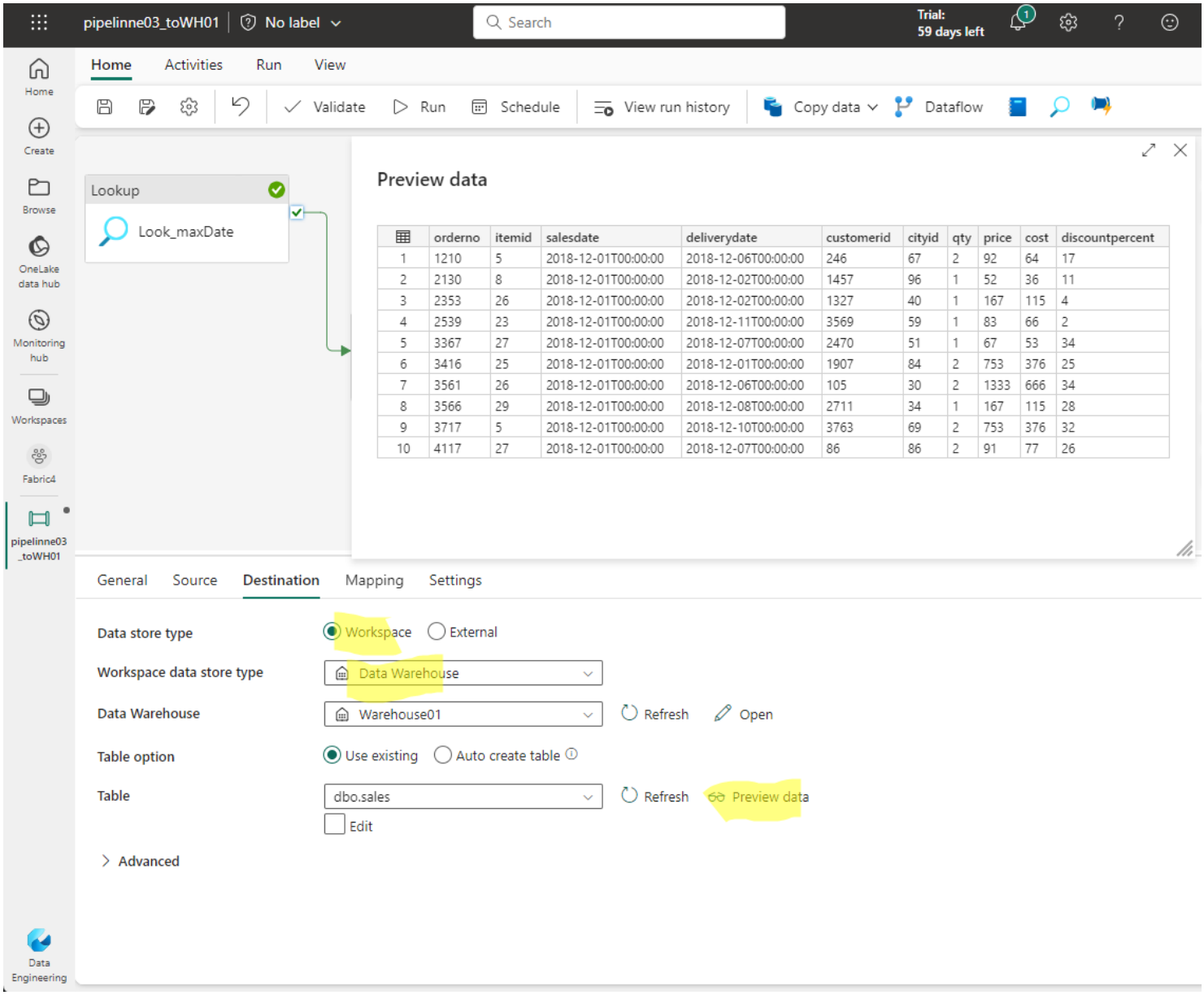


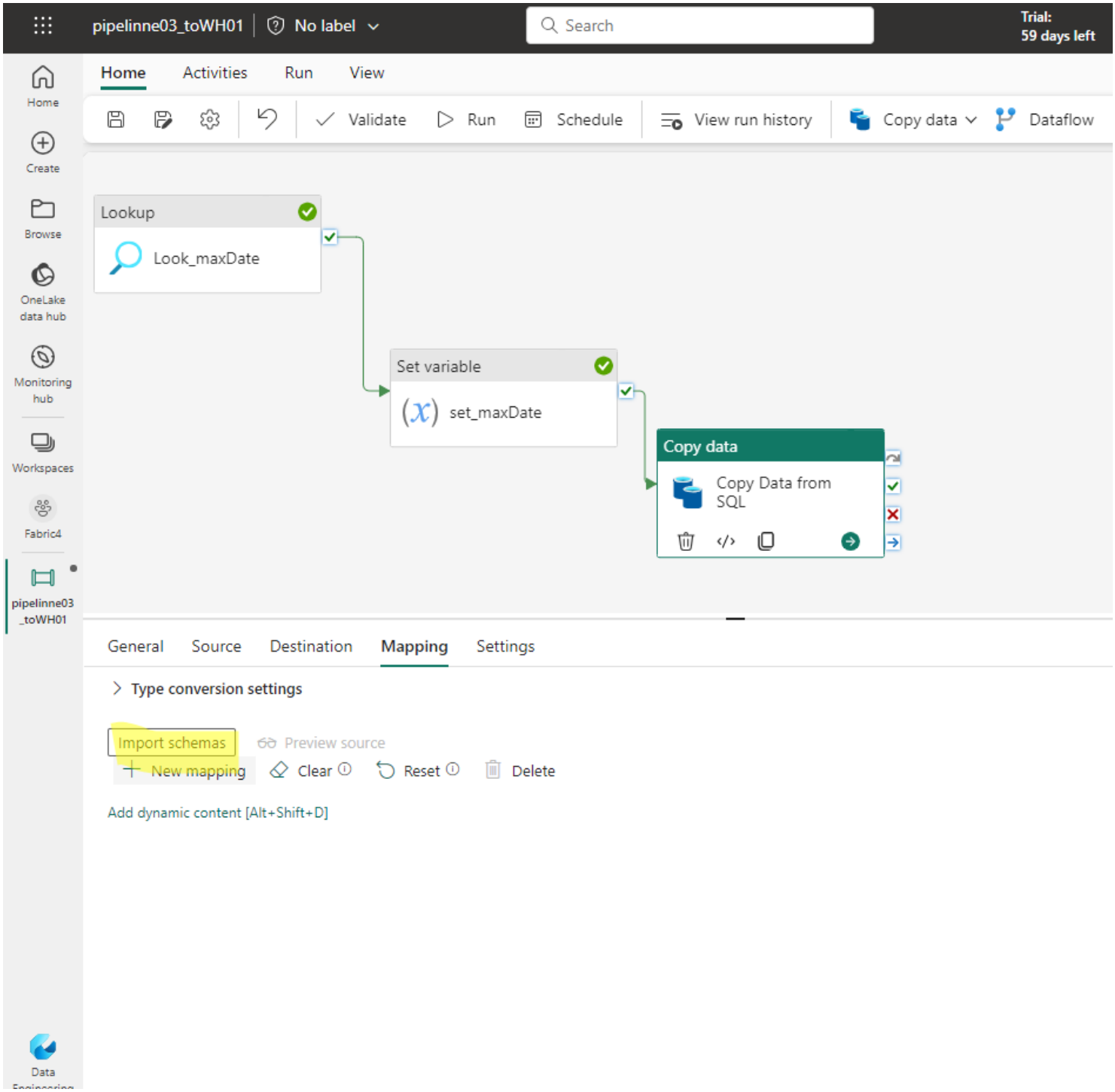

Step 3: Copy data

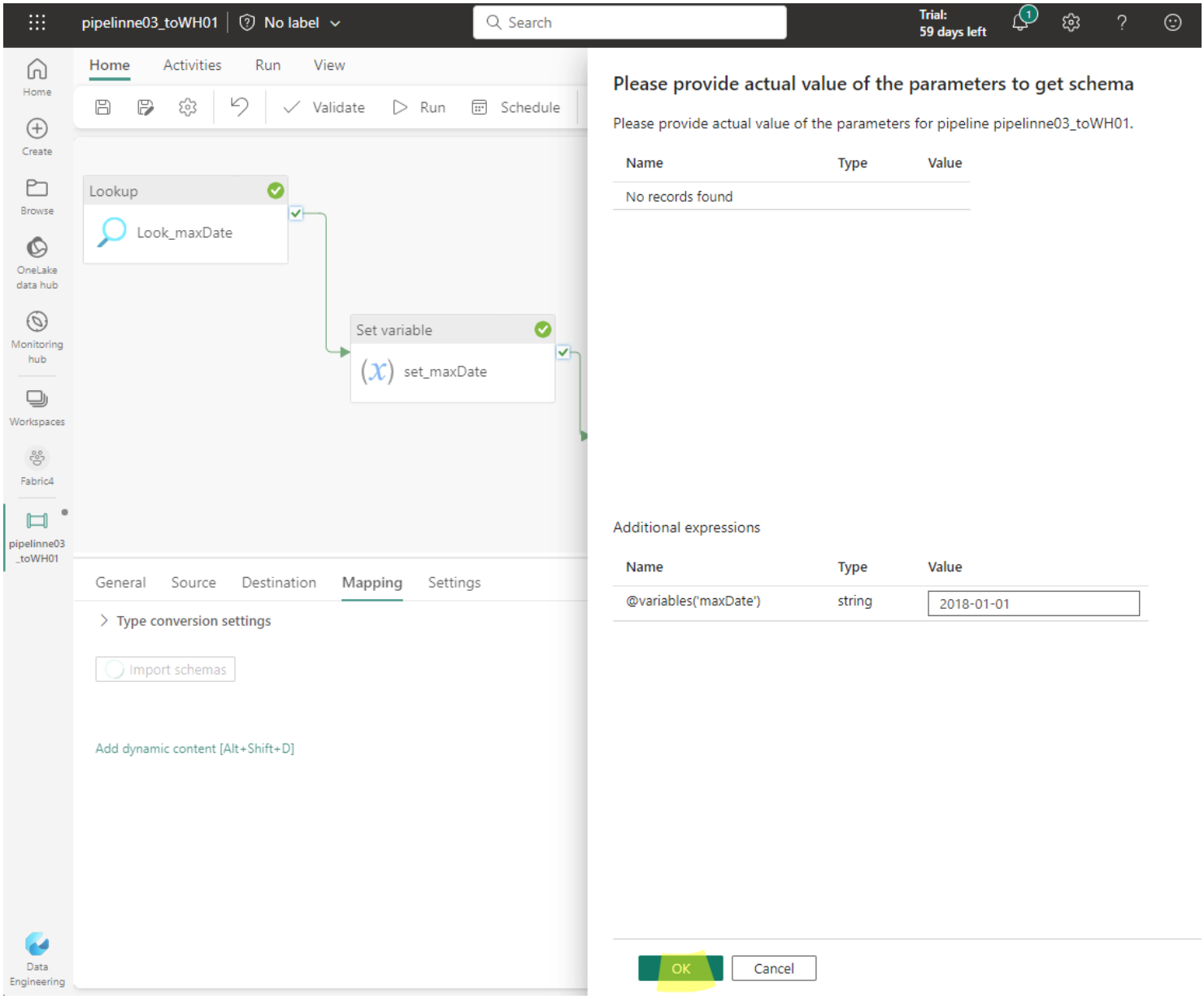

pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Add to canvas
Use copy assistant
Lookup
Look_maxDate
Set variable
set_maxDate
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03 toWH01


pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Lookup
Look_maxDate
Set variable
set_maxDate
Copy data
Copy Data from SQL
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03 _toWH01
General
Source 1
Destination 1
Mapping
Settings
Name *
Copy Data from SQL
Learn more
Description
Timeout
0.12:00:00
Retry
0
Advanced

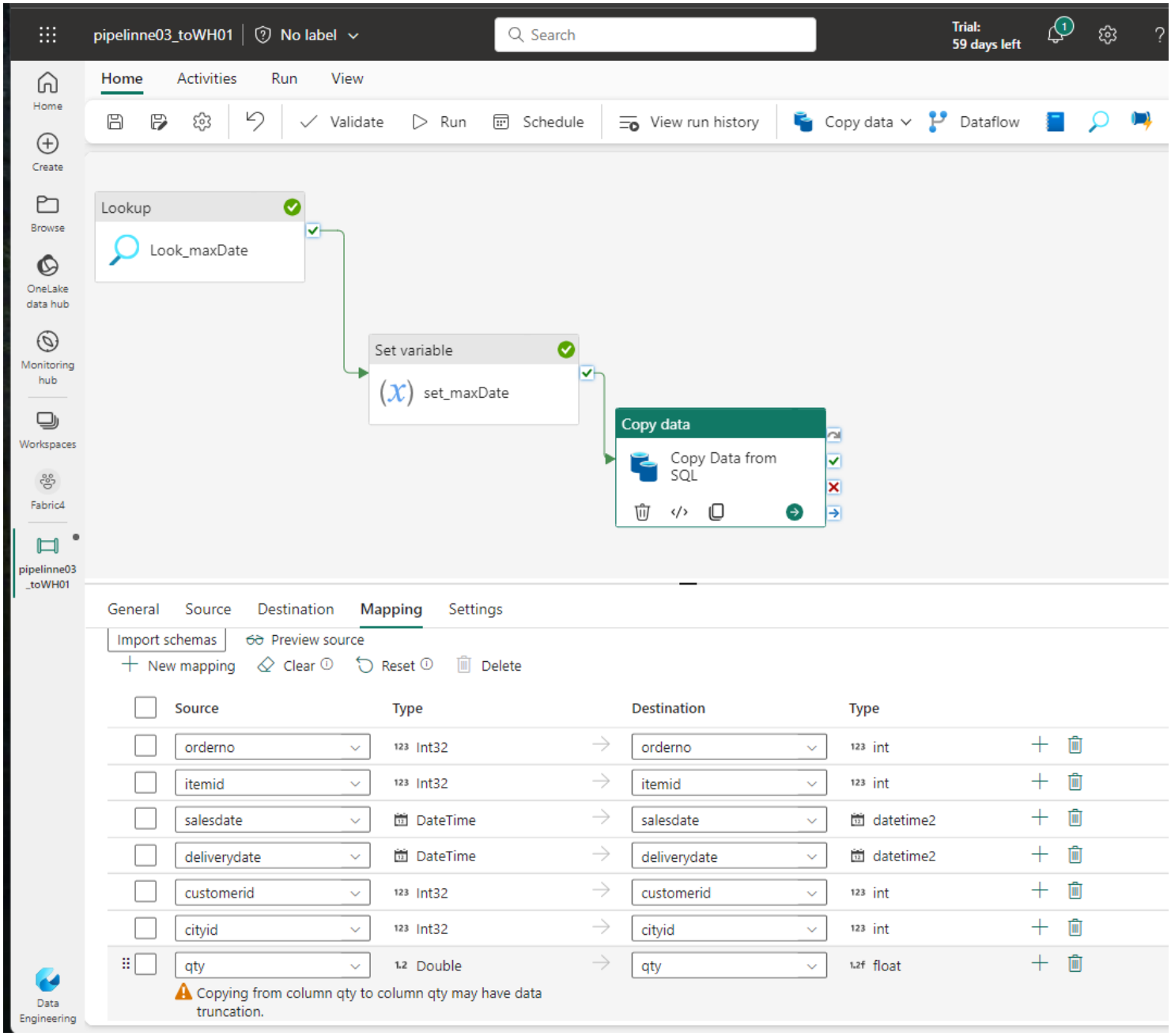

pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Lookup
Look_maxDate
Preview data
orderno itemid salesdate deliverydate customerid cityid qty price cost discountpercent
1 1210 5 2018-12-01T00:00:00 2018-12-06T00:00:00 246 67 2 92 64 17
2 2130 8 2018-12-01T00:00:00 2018-12-02T00:00:00 1457 96 1 52 36 11
3 2353 26 2018-12-01T00:00:00 2018-12-02T00:00:00 1327 40 1 167 115 4
4 2539 23 2018-12-01T00:00:00 2018-12-11T00:00:00 3569 59 1 83 66 2
5 3367 27 2018-12-01T00:00:00 2018-12-07T00:00:00 2470 51 1 67 53 34
6 3416 25 2018-12-01T00:00:00 2018-12-01T00:00:00 1907 84 2 753 376 25
7 3561 26 2018-12-01T00:00:00 2018-12-06T00:00:00 105 30 2 1333 666 34
8 3566 29 2018-12-01T00:00:00 2018-12-08T00:00:00 2711 34 1 167 115 28
9 3717 5 2018-12-01T00:00:00 2018-12-10T00:00:00 3763 69 2 753 376 32
10 4117 27 2018-12-01T00:00:00 2018-12-07T00:00:00 86 86 2 91 77 26
General
Source
Destination
Mapping
Settings
Data store type
Workspace
External
Sample dataset
Connection
rosie.db.elephantsql.com;sybspbgi h...
Refresh
Edit
New
Connection type
PostgreSQL
Test connection
Connection successful
Use query
Table
Query
Query *
select * from sales_fact where salesdate >'@{variables('maxDate')}'
Edit
Preview data
Advanced
Additional columns
New

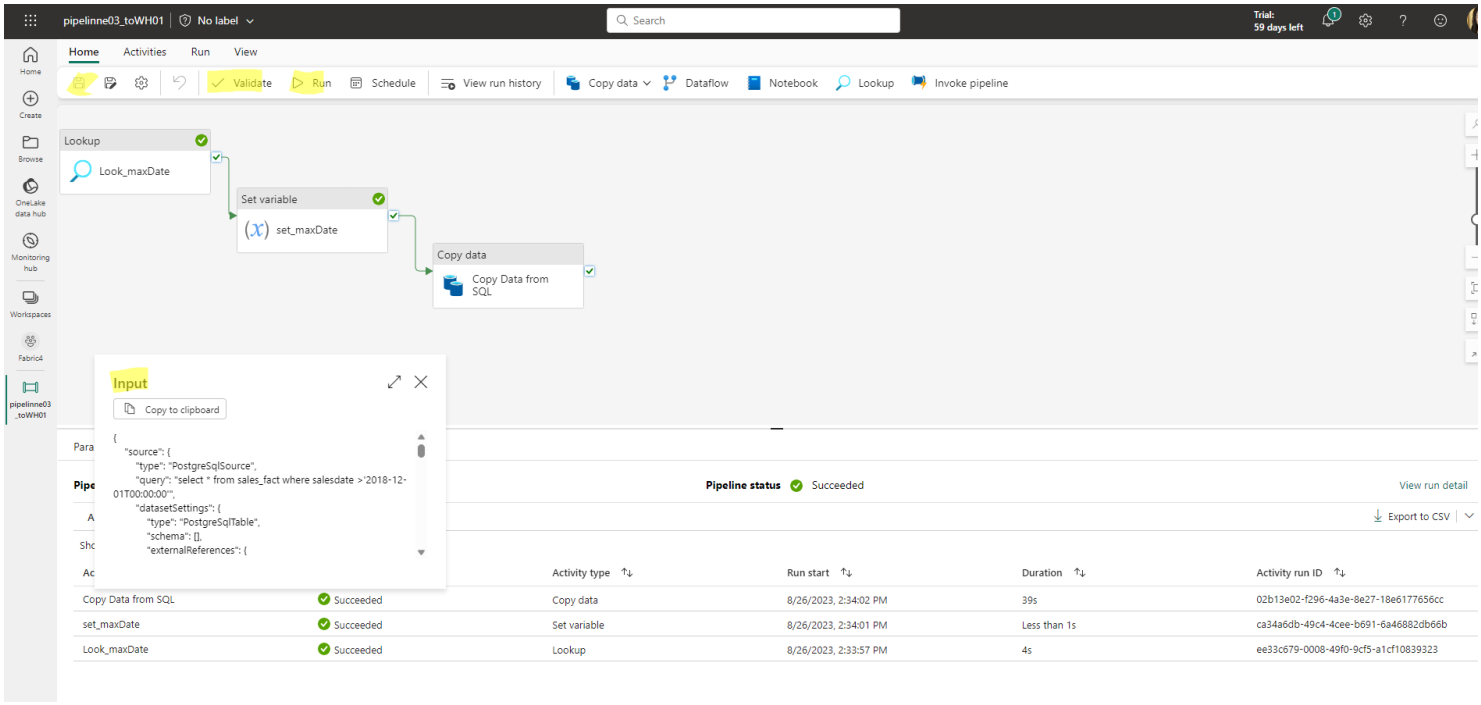

pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Lookup
Look_maxDate
Preview data
orderno itemid salesdate deliverydate customerid cityid qty price cost discountpercent
1 1210 5 2018-12-01T00:00:00 2018-12-06T00:00:00 246 67 2 92 64 17
2 2130 8 2018-12-01T00:00:00 2018-12-02T00:00:00 1457 96 1 52 36 11
3 2353 26 2018-12-01T00:00:00 2018-12-02T00:00:00 1327 40 1 167 115 4
4 2539 23 2018-12-01T00:00:00 2018-12-11T00:00:00 3569 59 1 83 66 2
5 3367 27 2018-12-01T00:00:00 2018-12-07T00:00:00 2470 51 1 67 53 34
6 3416 25 2018-12-01T00:00:00 2018-12-01T00:00:00 1907 84 2 753 376 25
7 3561 26 2018-12-01T00:00:00 2018-12-06T00:00:00 105 30 2 1333 666 34
8 3566 29 2018-12-01T00:00:00 2018-12-08T00:00:00 2711 34 1 167 115 28
9 3717 5 2018-12-01T00:00:00 2018-12-10T00:00:00 3763 69 2 753 376 32
10 4117 27 2018-12-01T00:00:00 2018-12-07T00:00:00 86 86 2 91 77 26
General
Source
Destination
Mapping
Settings
Data store type
Workspace
External
Workspace data store type
Data Warehouse
Data Warehouse
Warehouse01
Refresh
Open
Table option
Use existing
Auto create table
Table
dbo.sales
Refresh
Preview data
Edit
Advanced
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
pipelinne03_toWH01
Data Engineering


pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Lookup
Look_maxDate
Set variable
set_maxDate
Copy data
Copy Data from SQL
General
Source
Destination
Mapping
Settings
Type conversion settings
Import schemas
Preview source
New mapping
Clear
Reset
Delete
Add dynamic content [Alt+Shift+D]



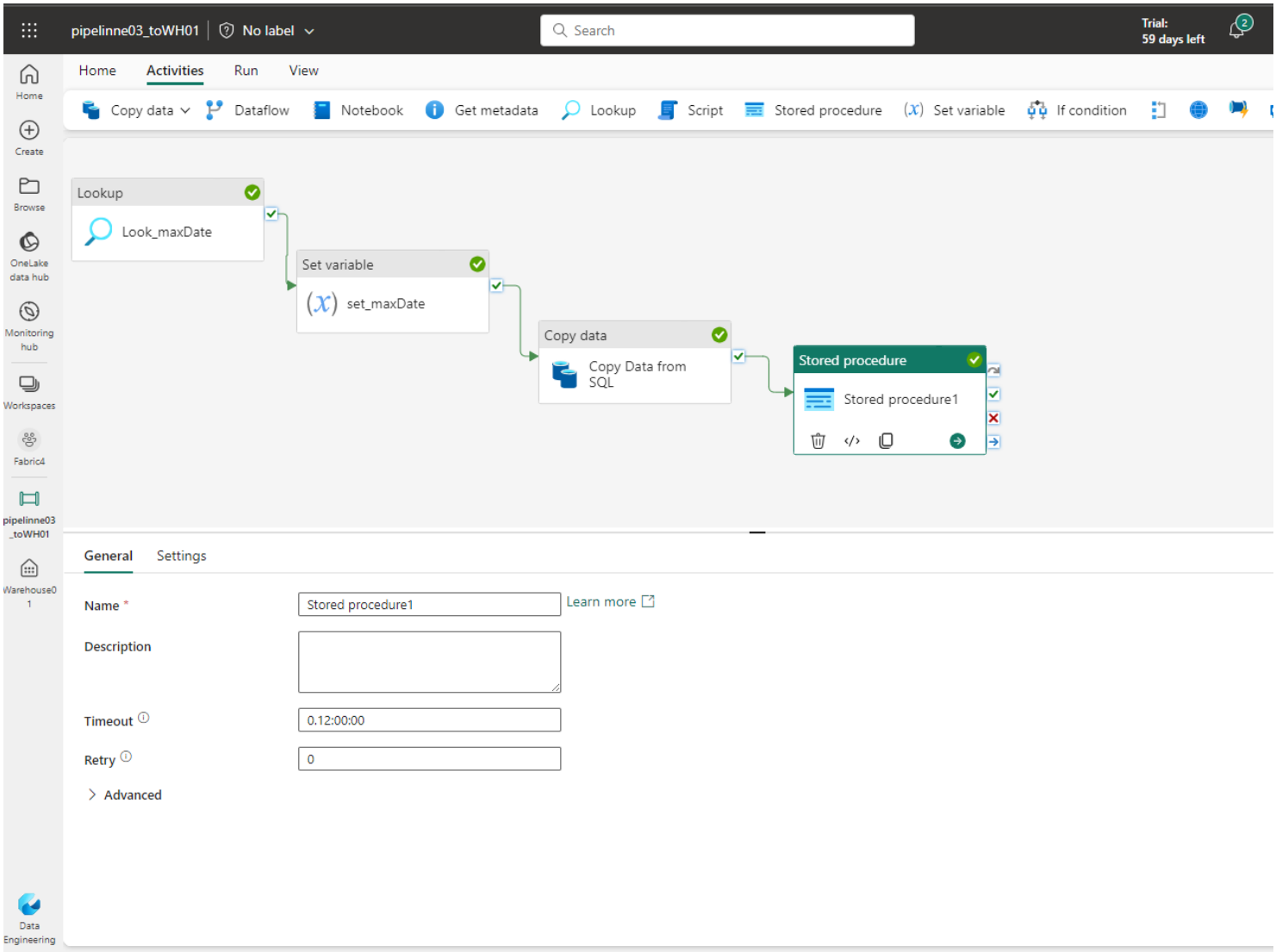

- Input sample date value to preview data

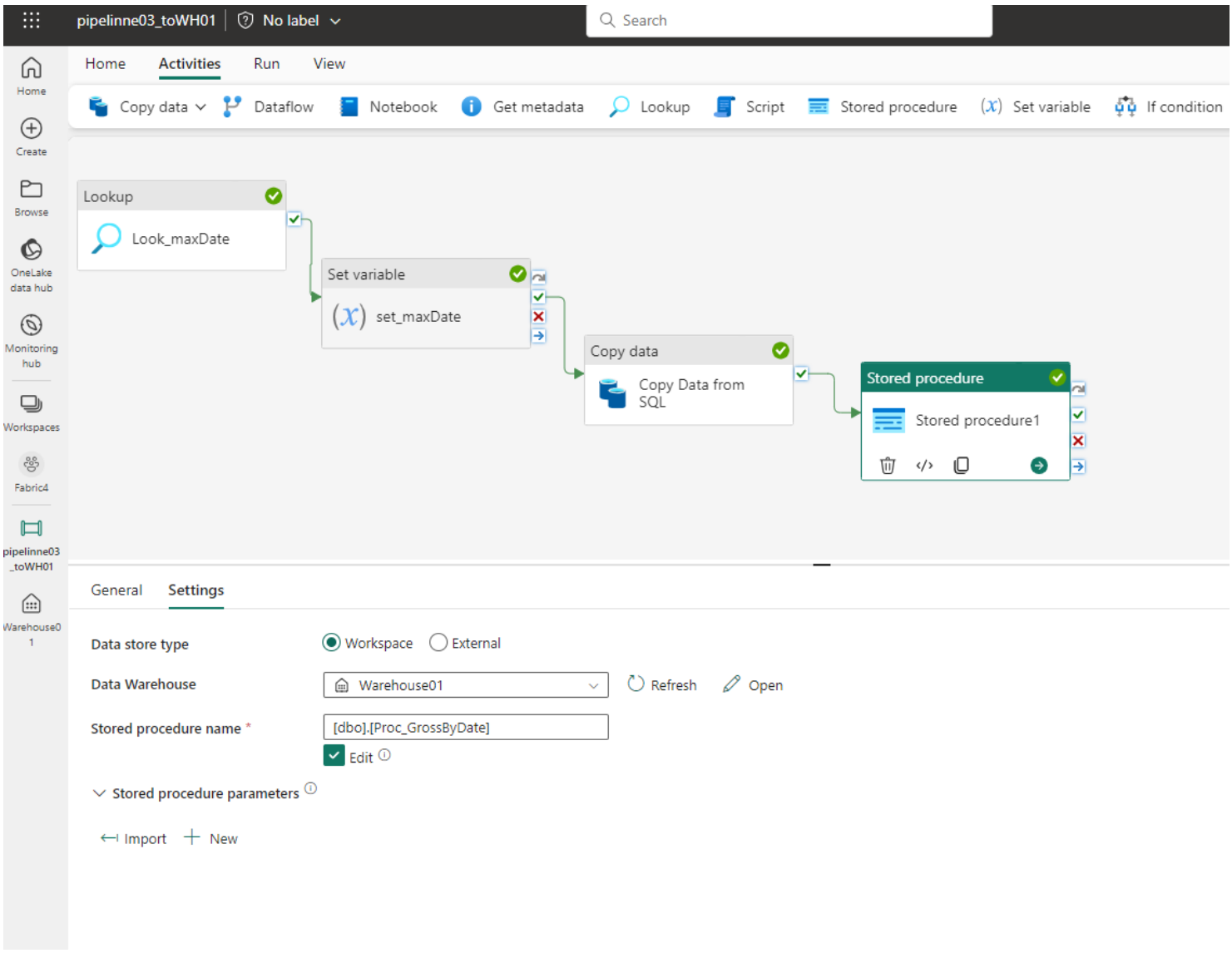

pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Lookup
Look_maxDate
Set variable
set_maxDate
Copy data
Copy Data from SQL
General
Source
Destination
Mapping
Settings
Import schemas
Preview source
New mapping
Clear
Reset
Delete
Source
Type
Destination
Type
orderno
Int32
orderno
int
itemid
Int32
itemid
int
salesdate
DateTime
salesdate
datetime2
deliverydate
DateTime
deliverydate
datetime2
customerid
Int32
customerid
int
cityid
Int32
cityid
int
qty
Double
qty
float
Copying from column qty to column qty may have data truncation.
Data Engineering

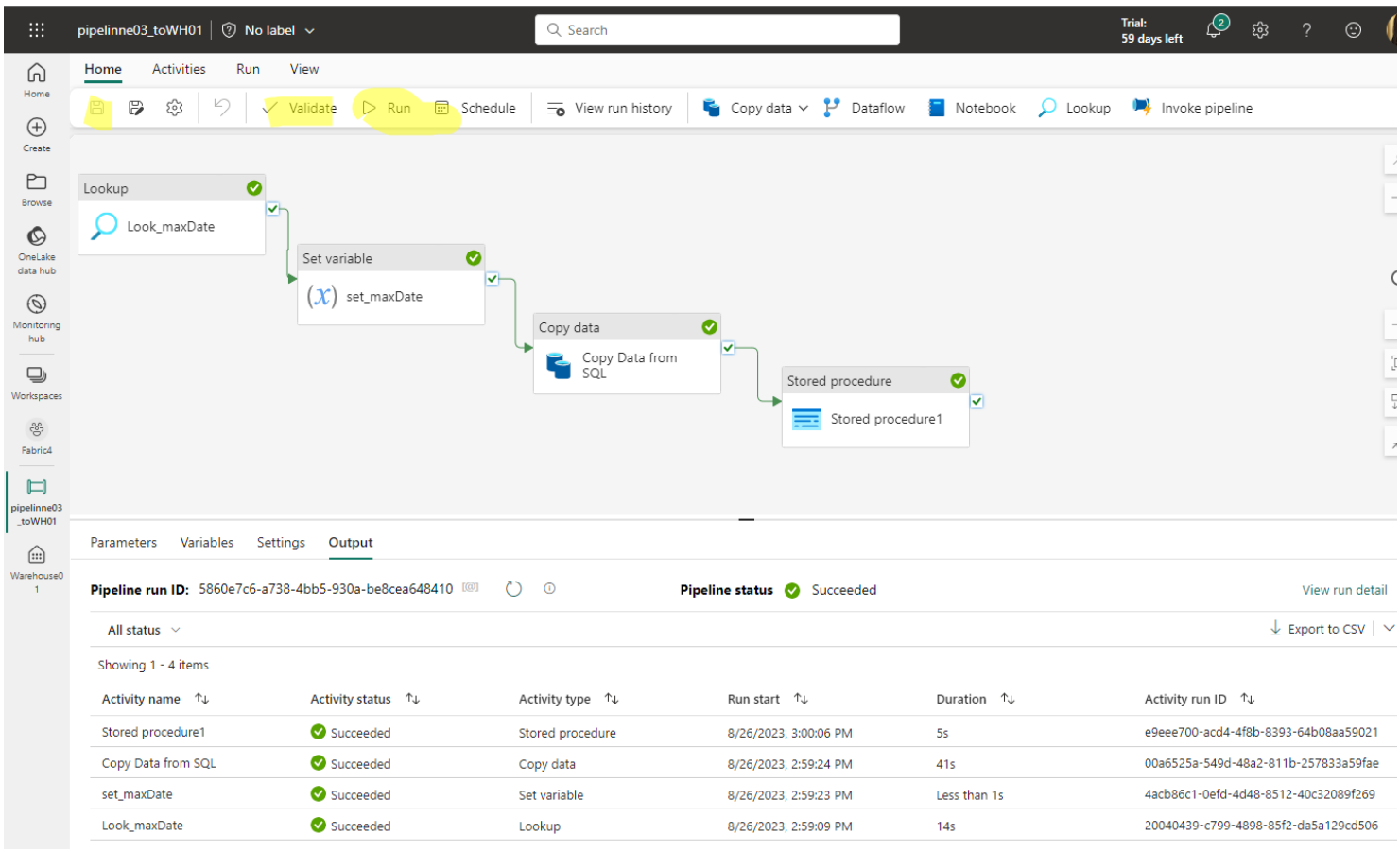

pipelinne03_toWH01
No label
Search
Trial:
59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Lookup
Invoke pipeline
Lookup
Look_maxDate
Set variable
set_maxDate
Copy data
Copy Data from SQL
Input
Copy to clipboard
{
"source": {
"type": "PostgreSqlSource",
"query": "select * from sales_fact where salesdate >'2018-12-01T00:00:00'",
"datasetSettings": {
"type": "PostgreSqlTable",
"schema": [],
"externalReferences": {
Pipeline status
Succeeded
View run detail
Export to CSV
Activity type
Run start
Duration
Activity run ID
Copy Data from SQL
Succeeded
Copy data
8/26/2023, 2:34:02 PM
39s
02b13e02-f296-4a3e-8e27-18e6177656cc
set_maxDate
Succeeded
Set variable
8/26/2023, 2:34:01 PM
Less than 1s
ca34a6db-49c4-4cee-b691-6a46882db66b
Look_maxDate
Succeeded
Lookup
8/26/2023, 2:33:57 PM
4s
ee33c679-0008-49f0-9cf5-a1cf10839323

After complete the step, Save > validate  > run  to check if any error in the setup

## Step 4: Stored procedure

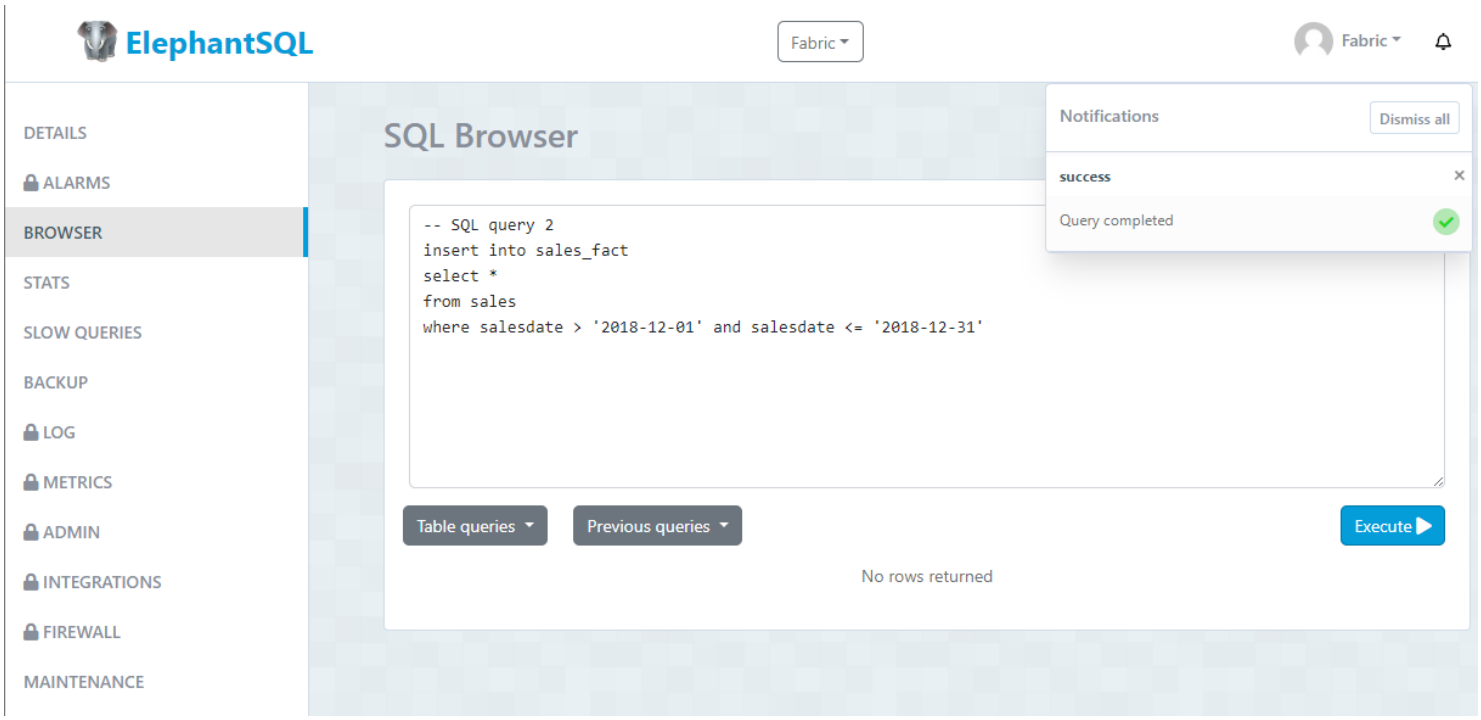


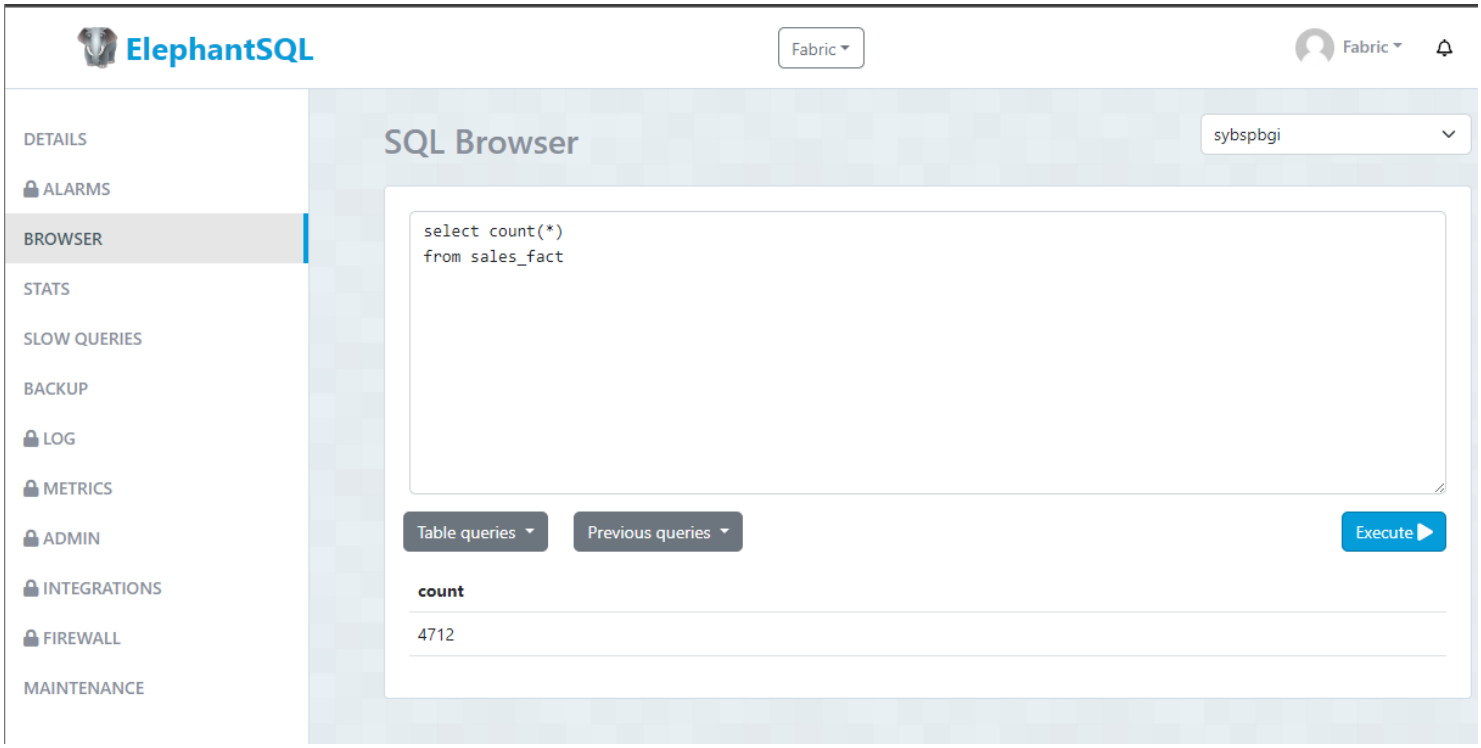

pipelinne03_toWH01
No label
Search
Home
Activities
Run
View
Copy data
Dataflow
Notebook
Get metadata
Lookup
Script
Stored procedure
Set variable
If condition
Lookup
Look_maxDate
Set variable
set_maxDate
Copy data
Copy Data from SQL
Stored procedure
Stored procedure1
General
Settings
Data store type
Workspace
External
Data Warehouse
Warehouse01
Refresh
Open
Stored procedure name *
[dbo].[Proc_GrossByDate]
Edit
Stored procedure parameters
Import
New


pipelinne03_toWH01
No label
Search
Trial: 59 days left
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Lookup
Invoke pipeline
Lookup
Look_maxDate
Set variable
set_maxDate
Copy data
Copy Data from SQL
Stored procedure
Stored procedure1
Parameters
Variables
Settings
Output
Pipeline run ID: 5860e7c6-a738-4bb5-930a-be8cea648410
Pipeline status Succeeded
View run detail
All status
Export to CSV
Showing 1 - 4 items
Activity name
Activity status
Activity type
Run start
Duration
Activity run ID
Stored procedure1 Succeeded Stored procedure 8/26/2023, 3:00:06 PM 5s e9eee700-acd4-4f8b-8393-64b08aa59021
Copy Data from SQL Succeeded Copy data 8/26/2023, 2:59:24 PM 41s 00a6525a-549d-48a2-811b-257833a59fae
set_maxDate Succeeded Set variable 8/26/2023, 2:59:23 PM Less than 1s 4acb86c1-0efd-4d48-8512-40c32089f269
Look_maxDate Succeeded Lookup 8/26/2023, 2:59:09 PM 14s 20040439-c799-4898-85f2-da5a129cd506

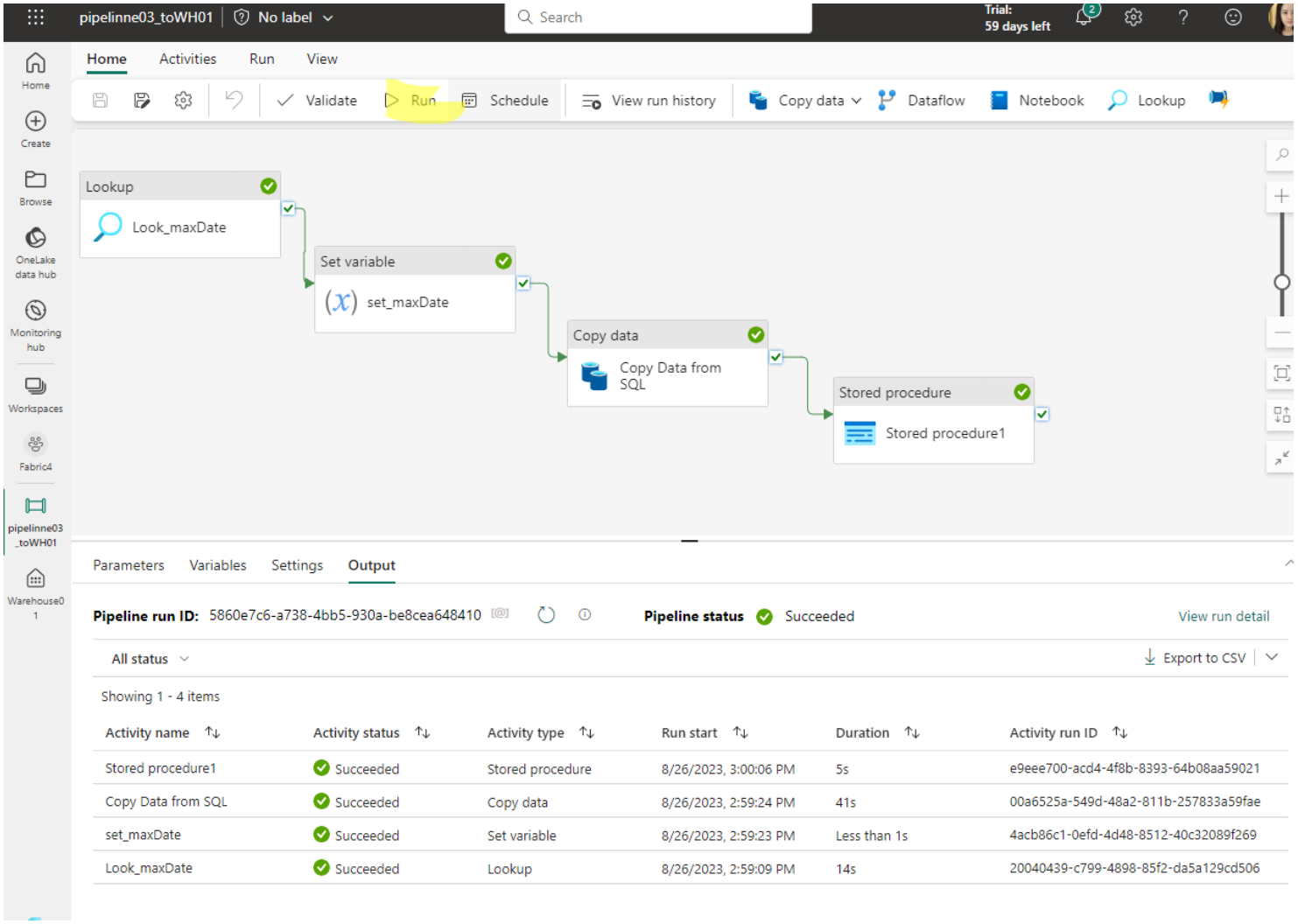
After complete the step, Save > validate  > run  to check if any error in the setup

### *2.7.3 Update data in SQL*

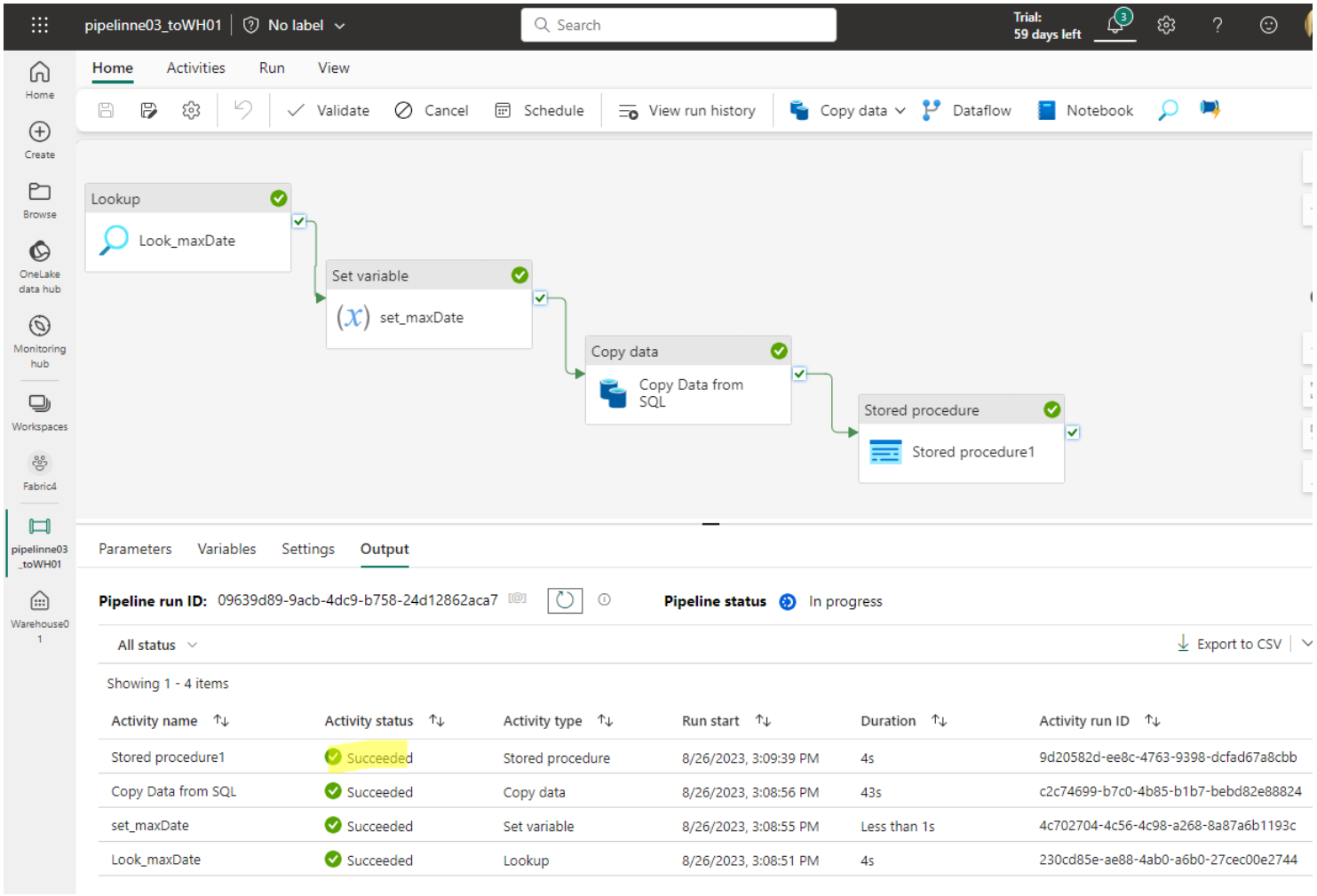


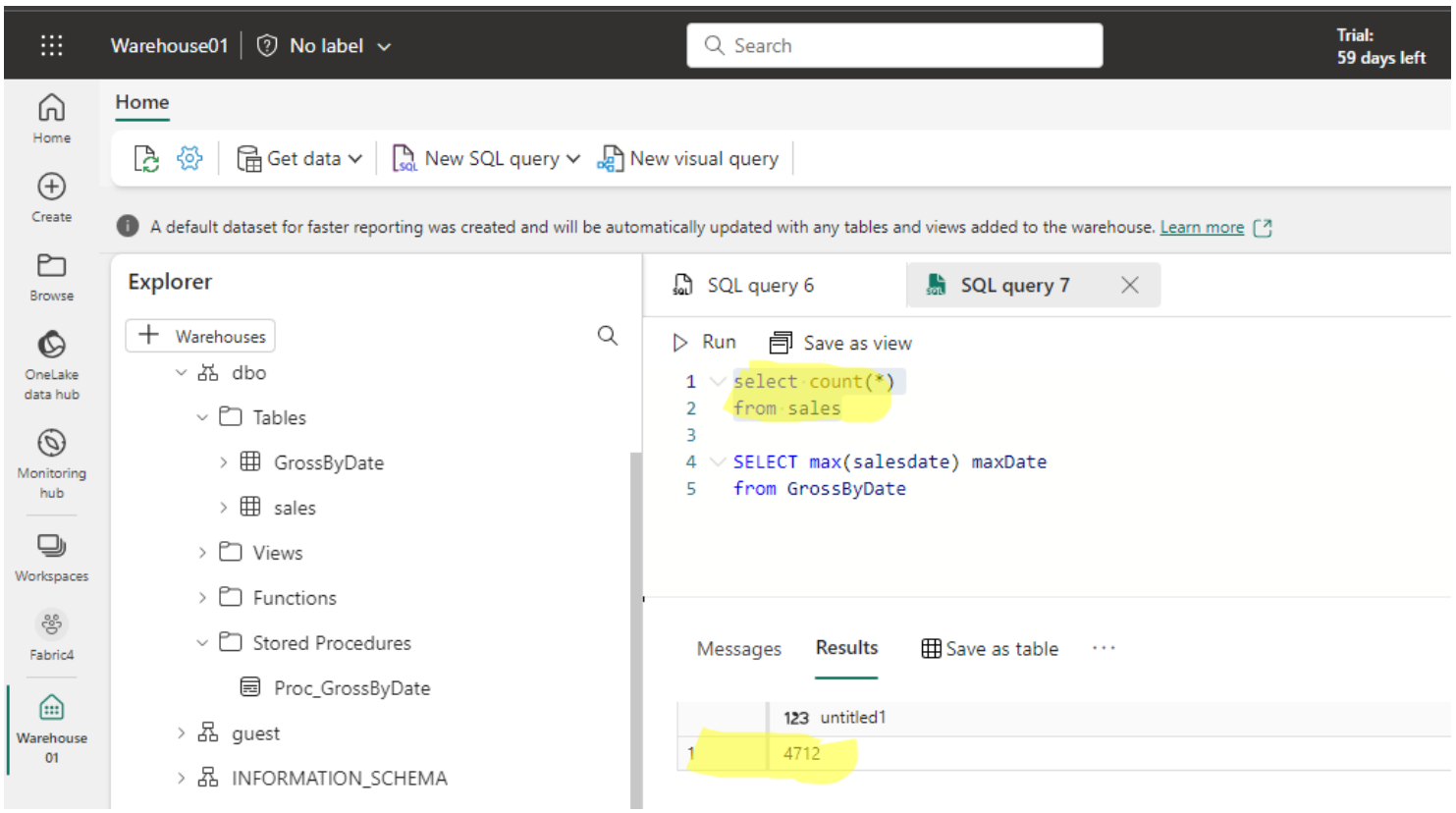
## Run pipeline manually to load data

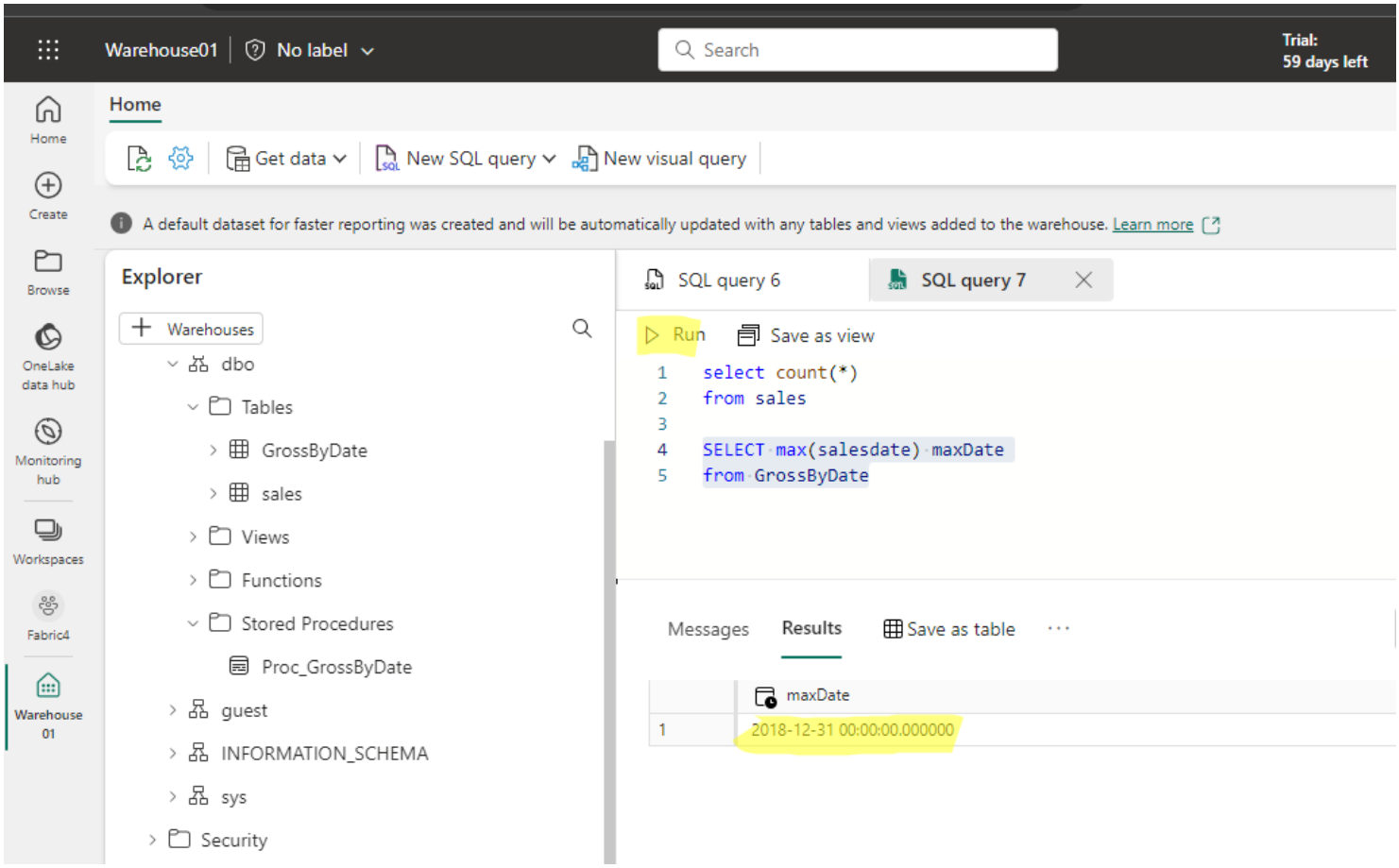



| Activity name | Activity status | Activity type | Run start | Duration | Activity run ID |
|---|---|---|---|---|---|
| Stored procedure1 | Succeeded | Stored procedure | 8/26/2023, 3:00:06 PM | 5s | e9eee700-acd4-4f8b-8393-64b08aa59021 |
| Copy Data from SQL | Succeeded | Copy data | 8/26/2023, 2:59:24 PM | 41s | 00a6525a-549d-48a2-811b-257833a59fae |
| set_maxDate | Succeeded | Set variable | 8/26/2023, 2:59:23 PM | Less than 1s | 4acb86c1-0efd-4d48-8512-40c32089f269 |
| Look_maxDate | Succeeded | Lookup | 8/26/2023, 2:59:09 PM | 14s | 20040439-c799-4898-85f2-da5a129cd506 |

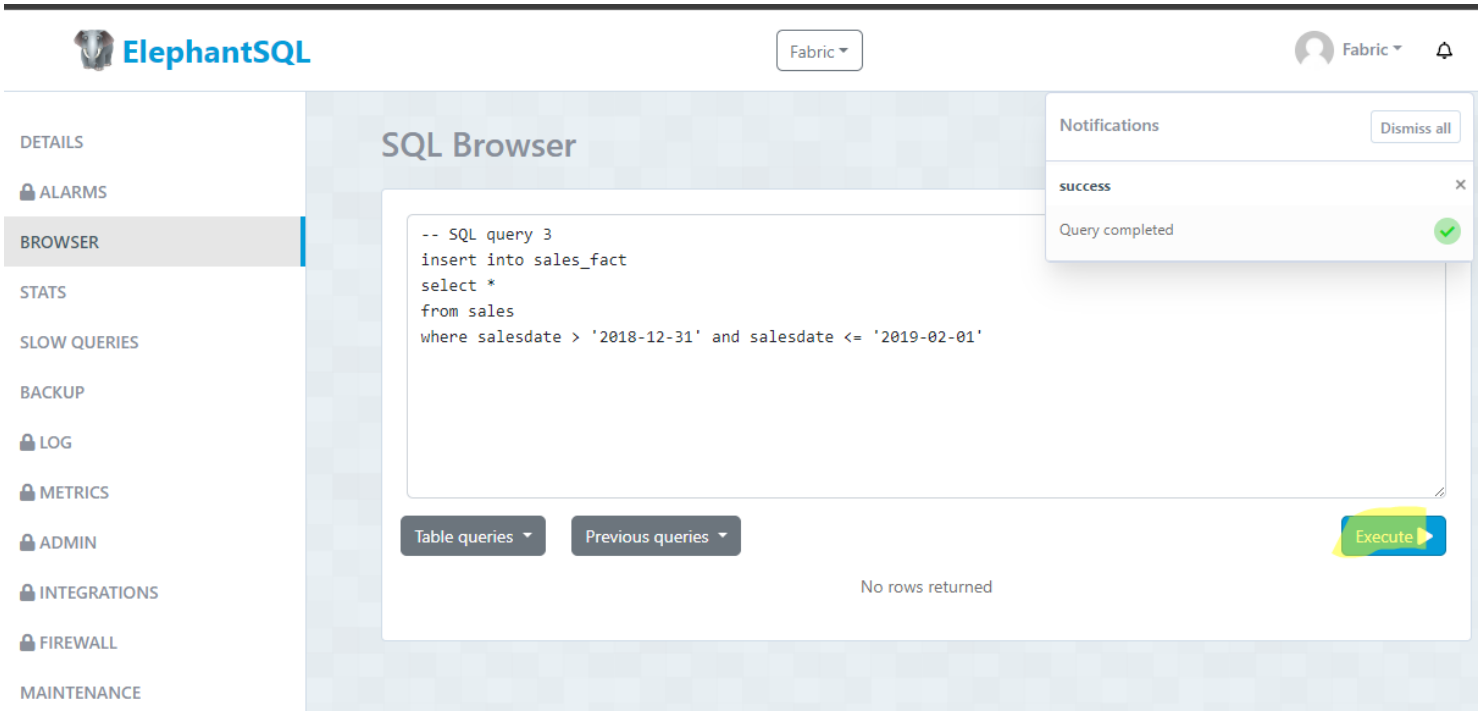


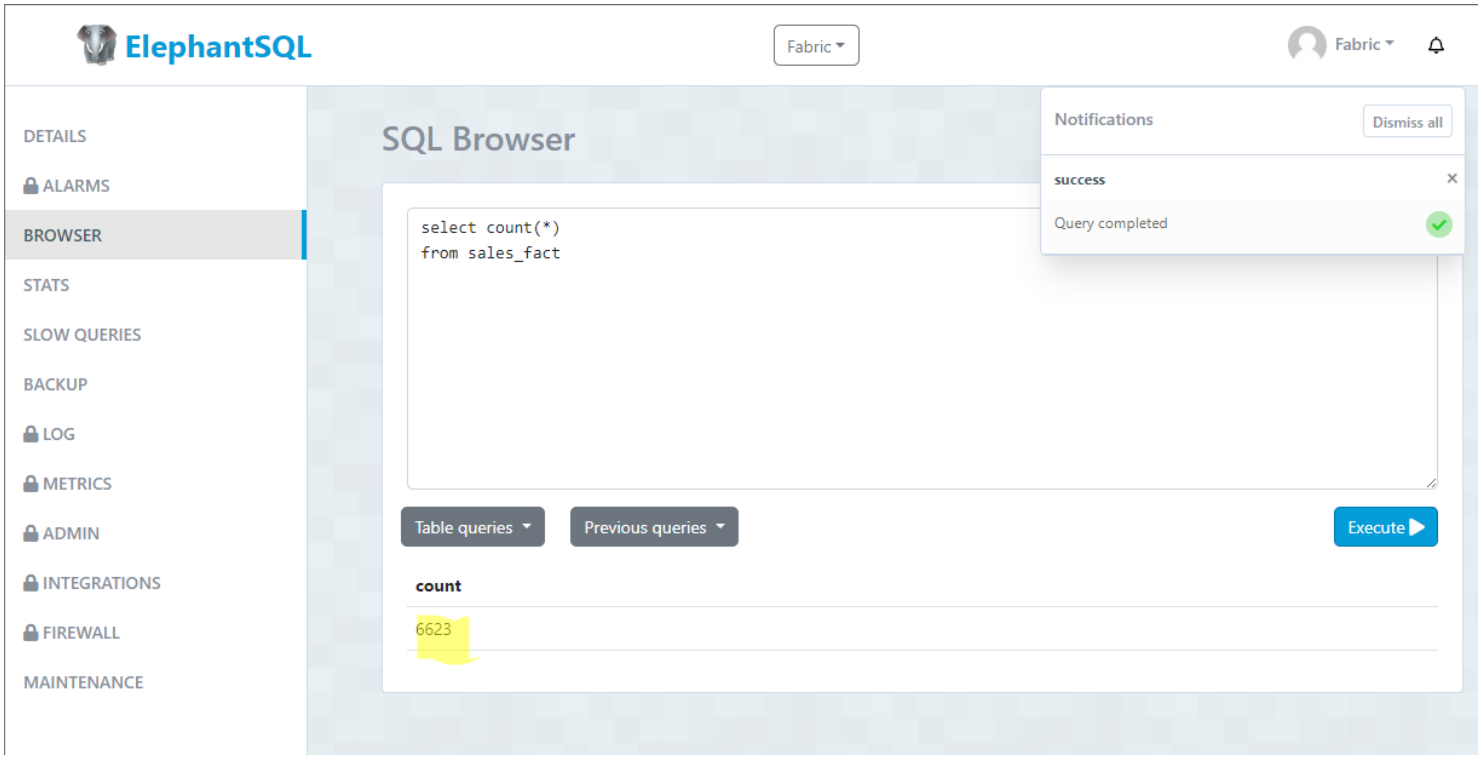

## Check data in WH



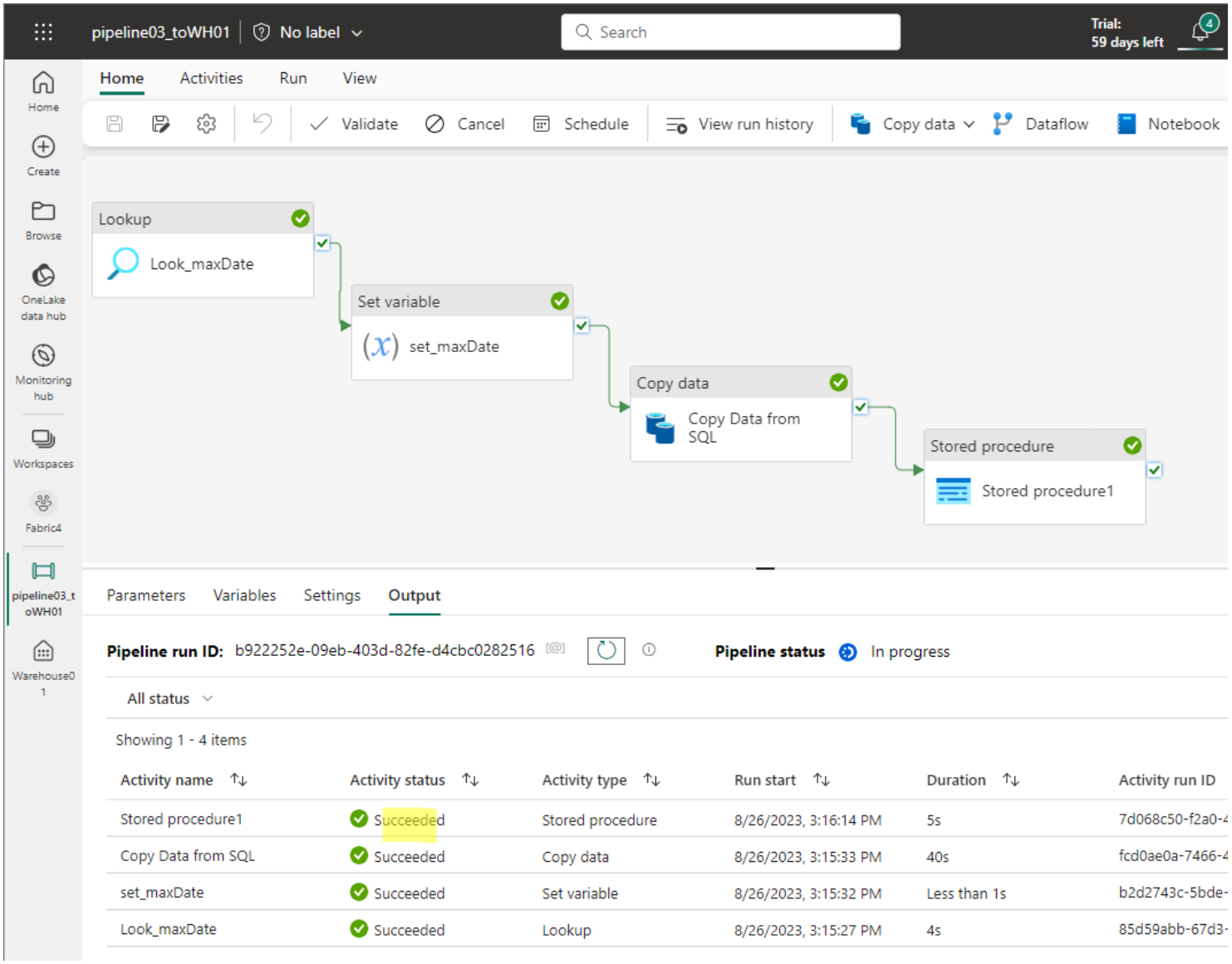


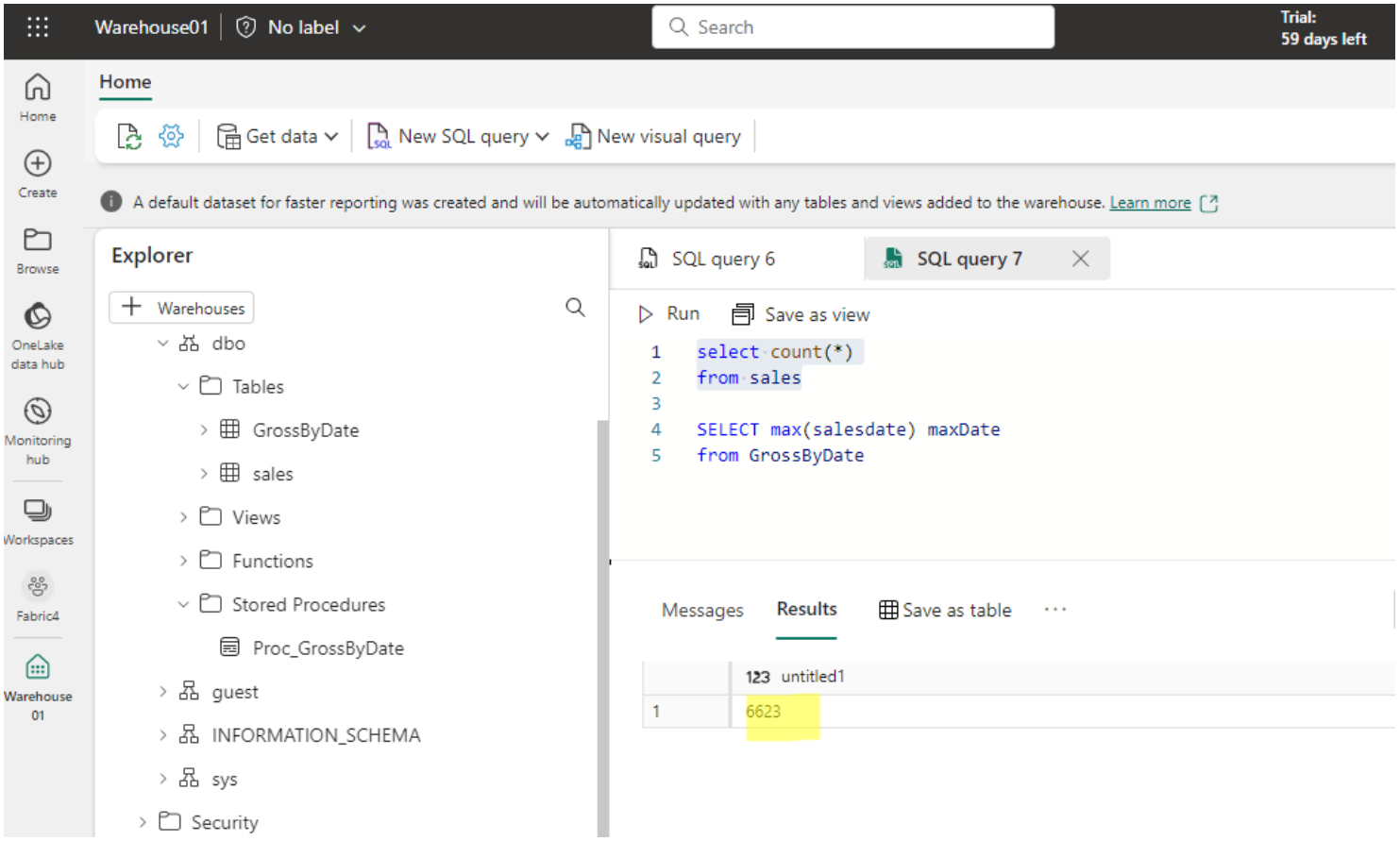

## Load new data to sales_fact table

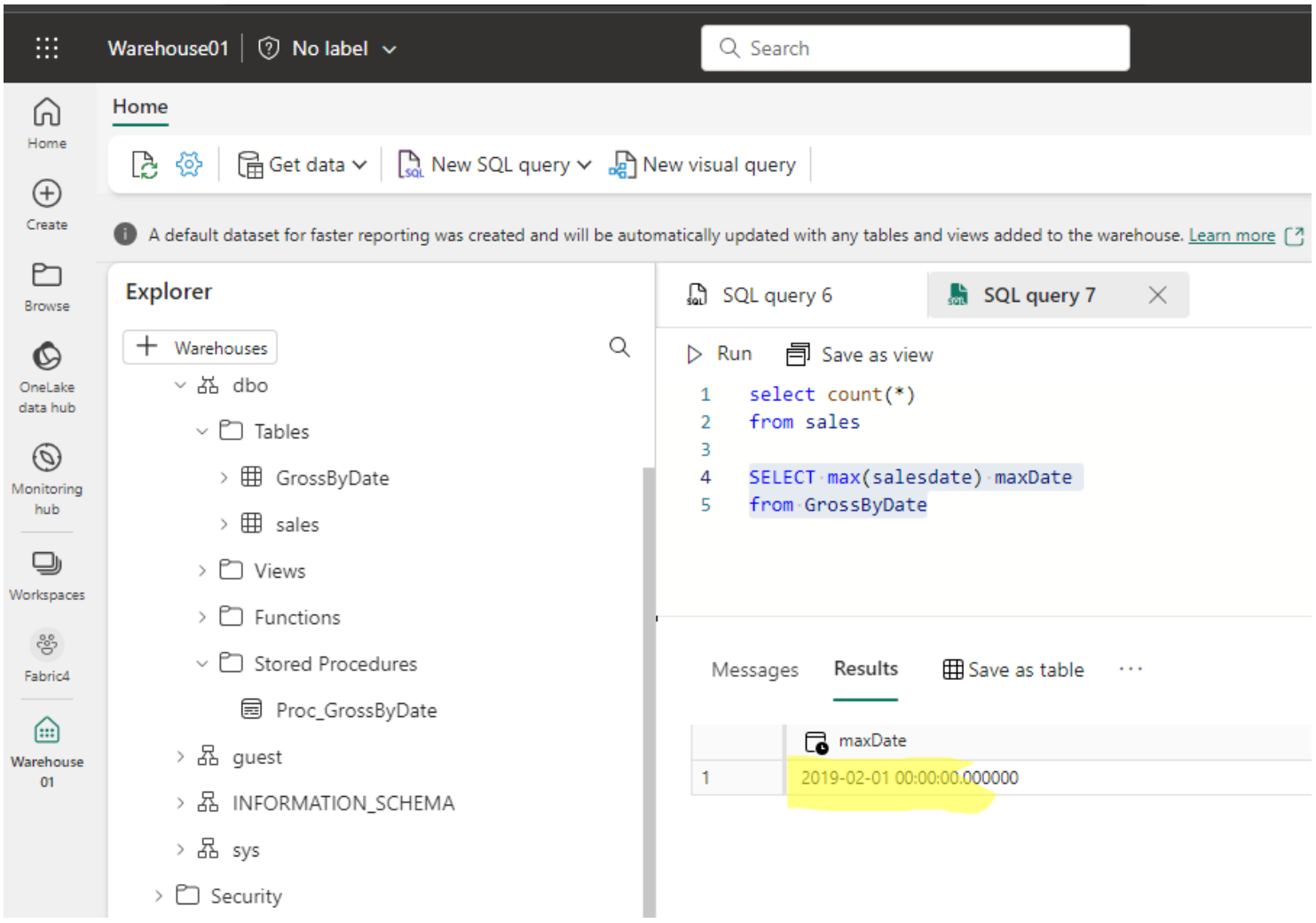





## Run manual pipeline

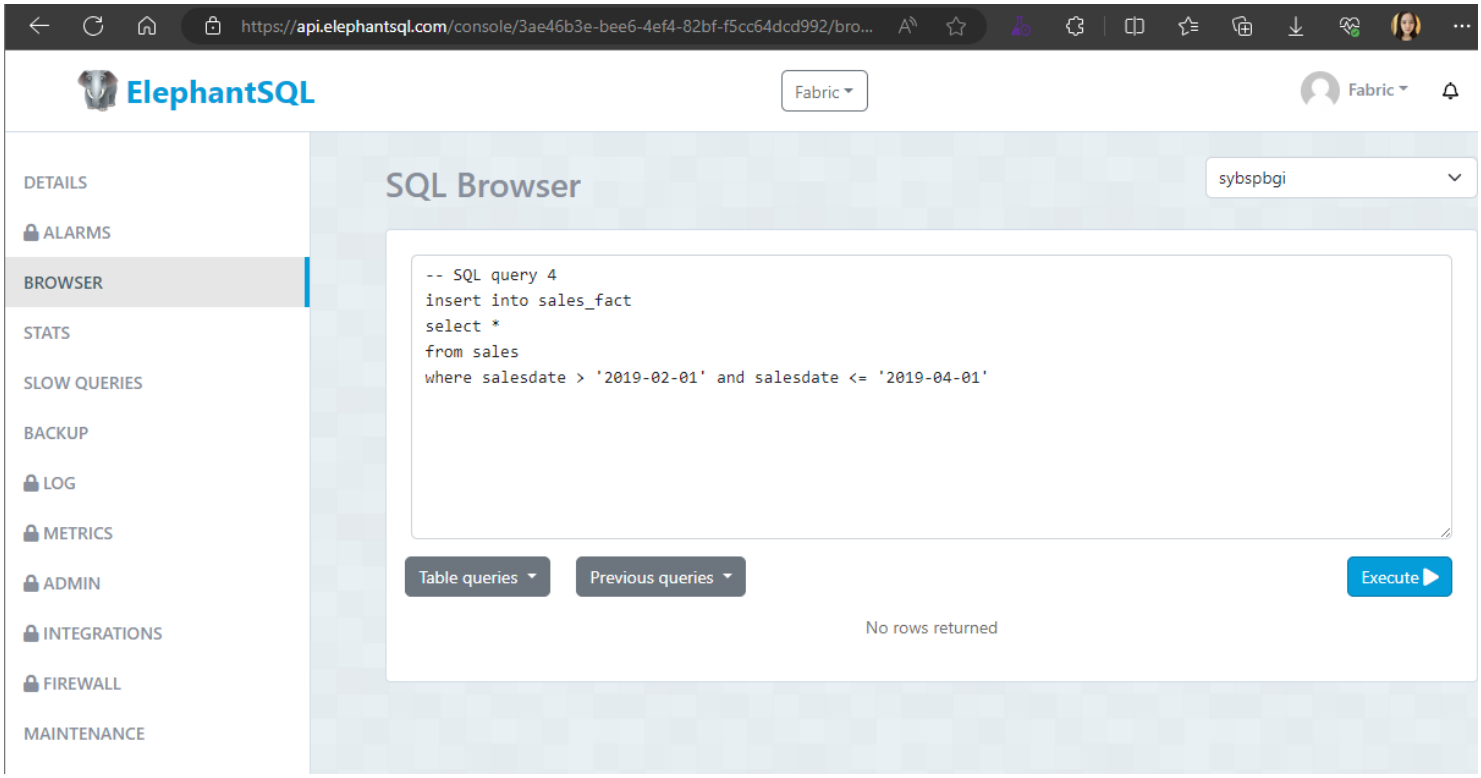


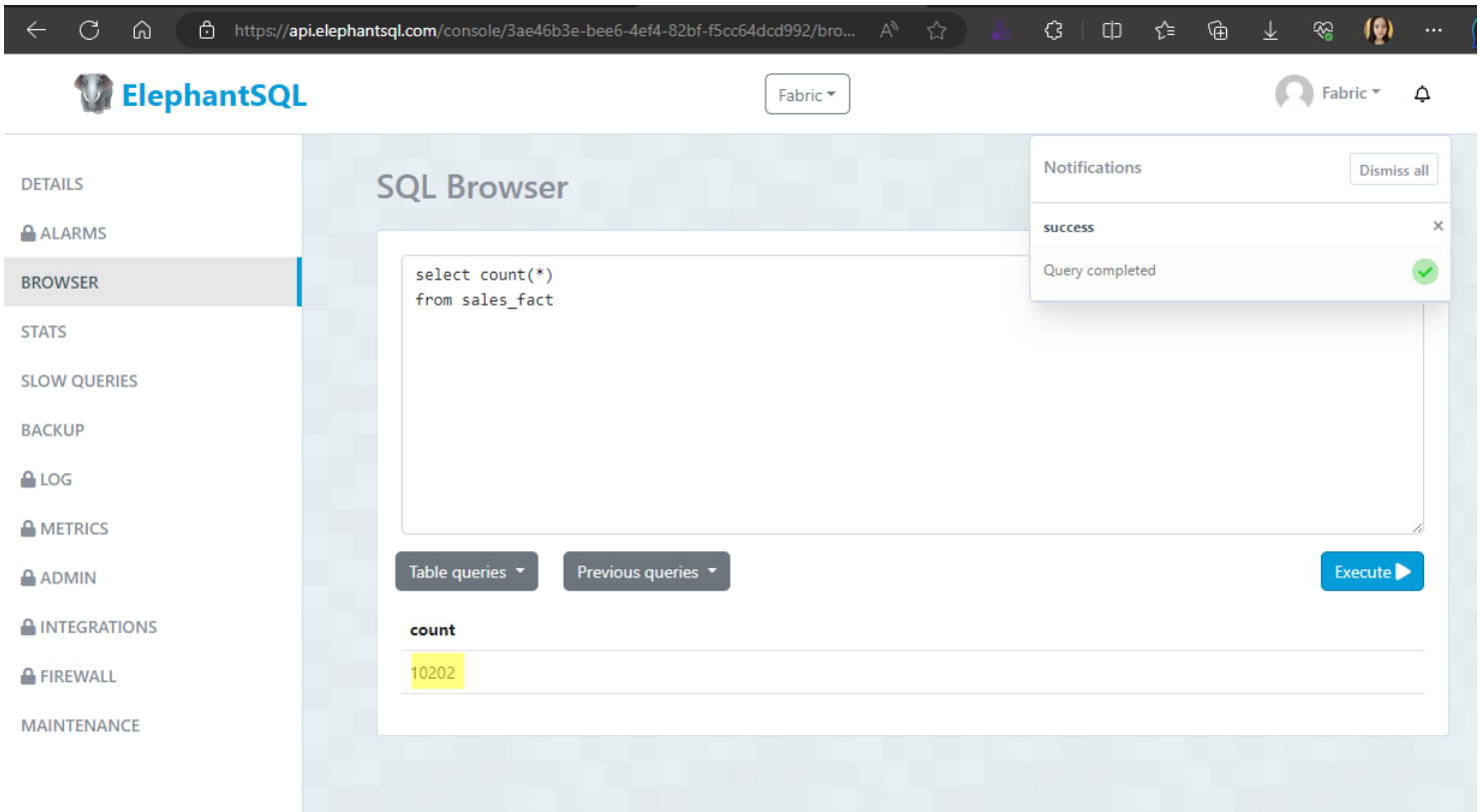


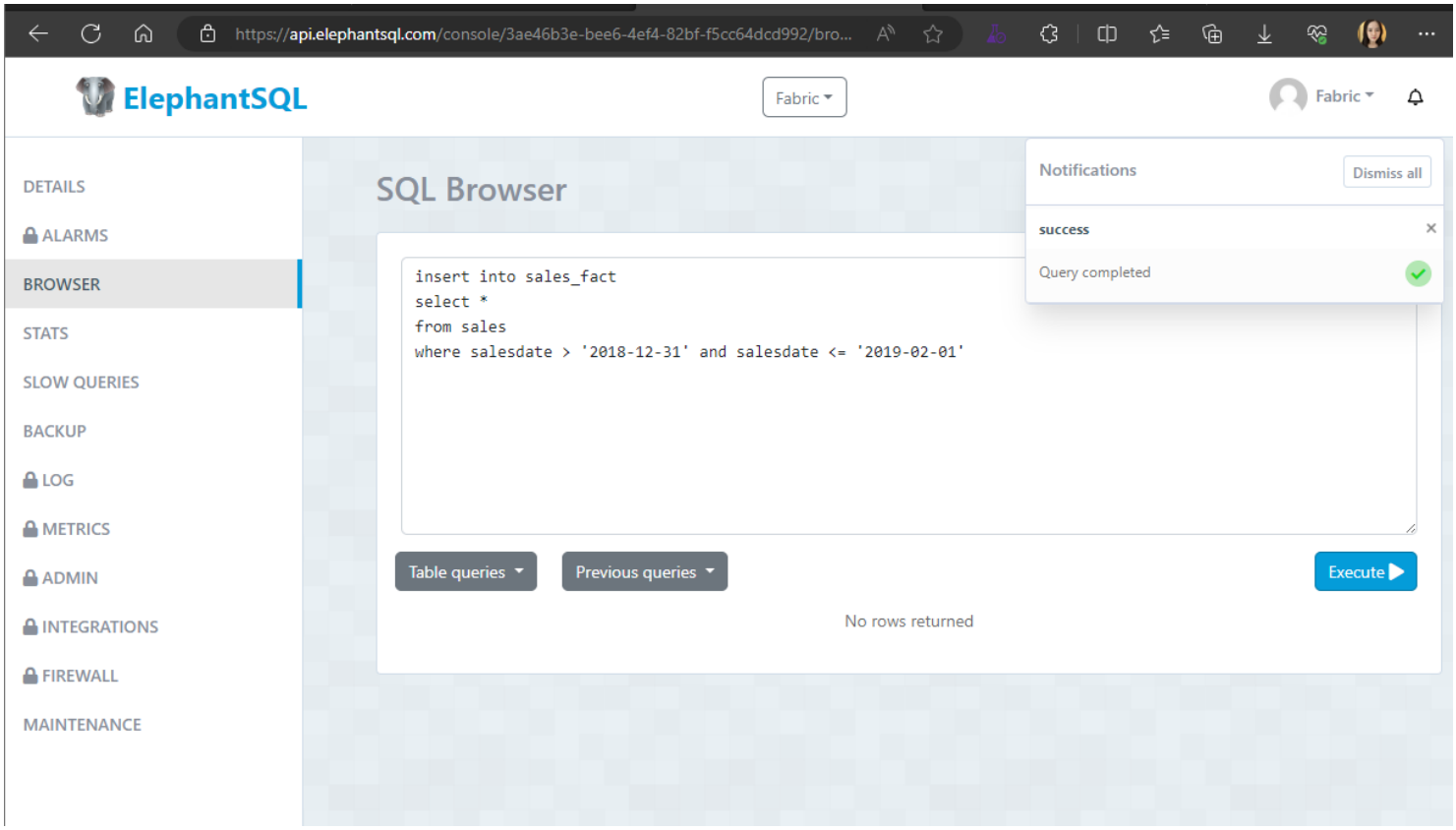

Make data double in SQL database

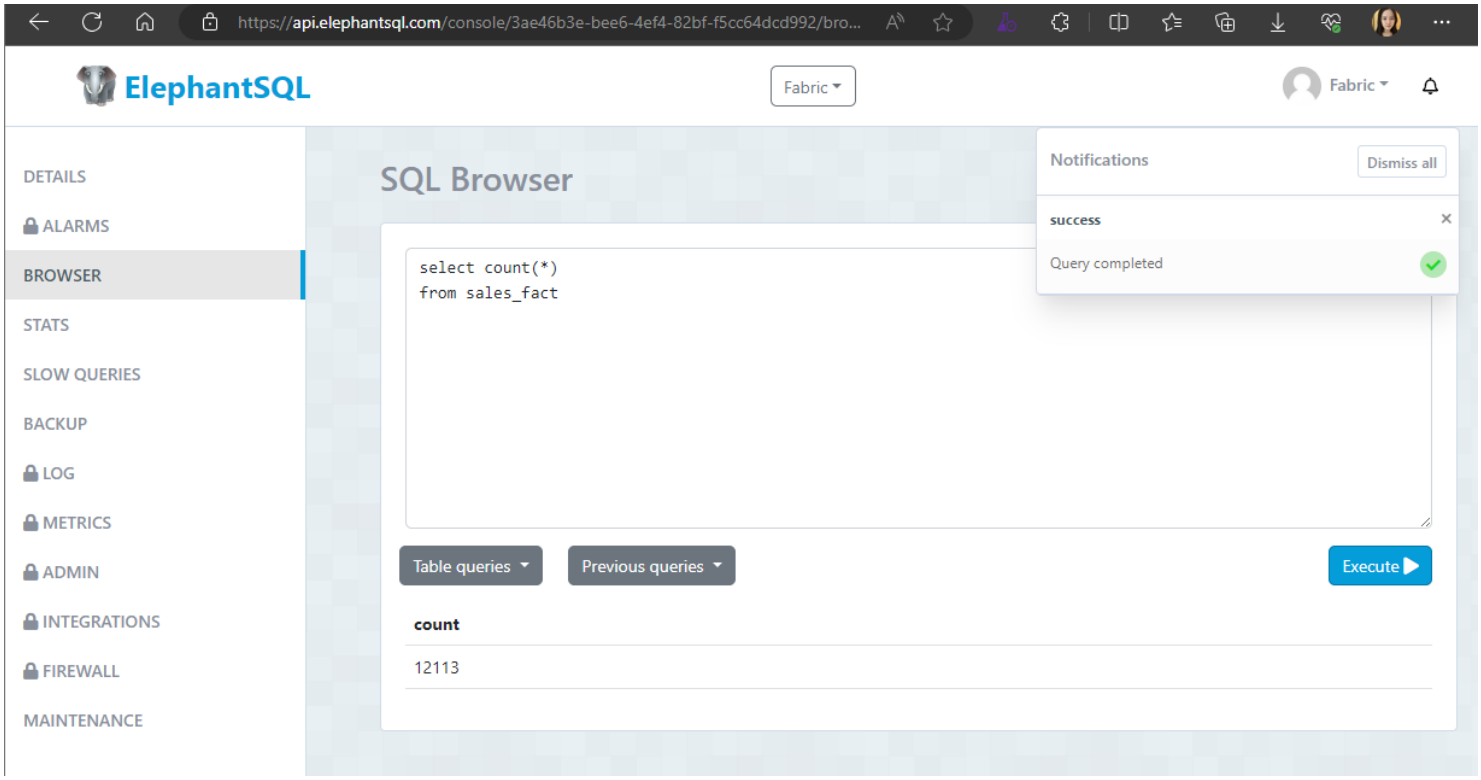
Create new data

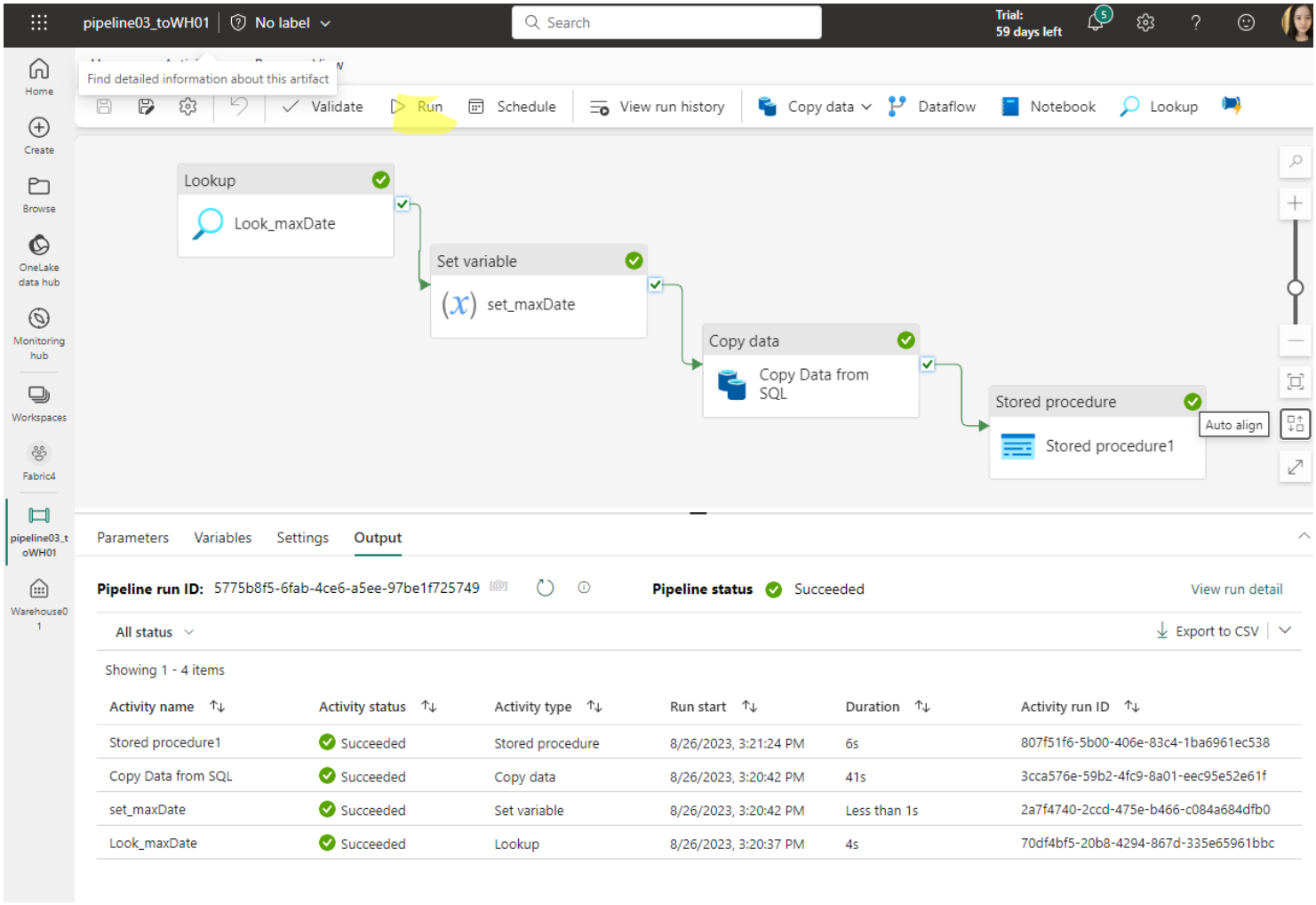


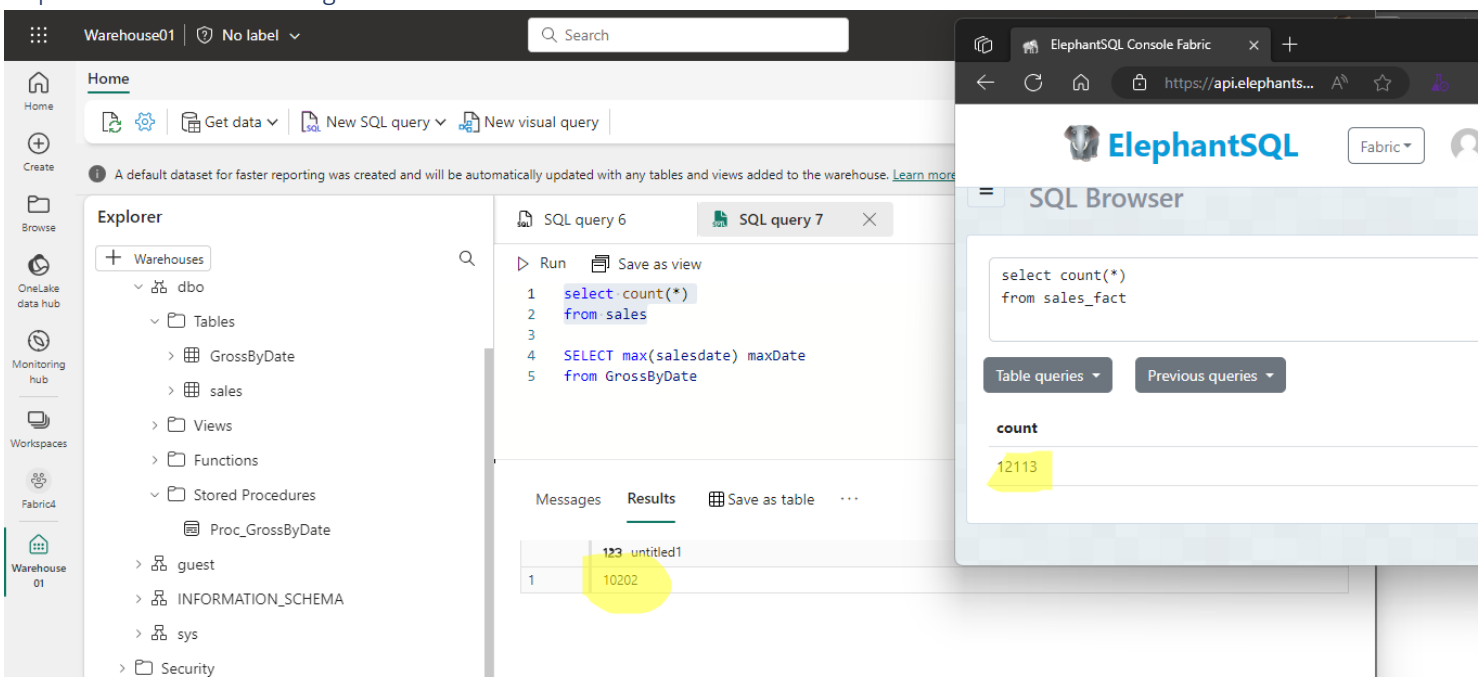
Input query to make duplicated data

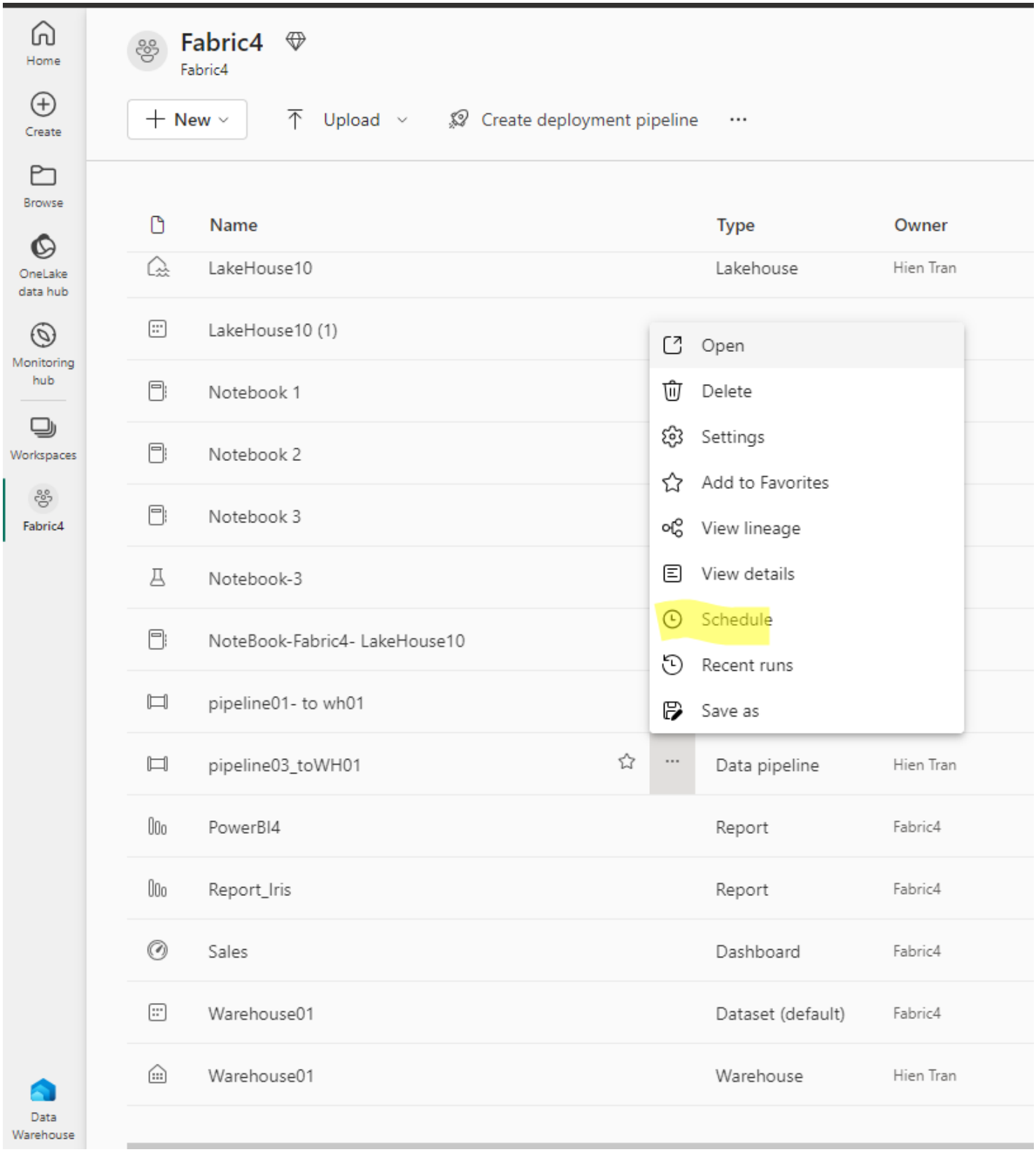


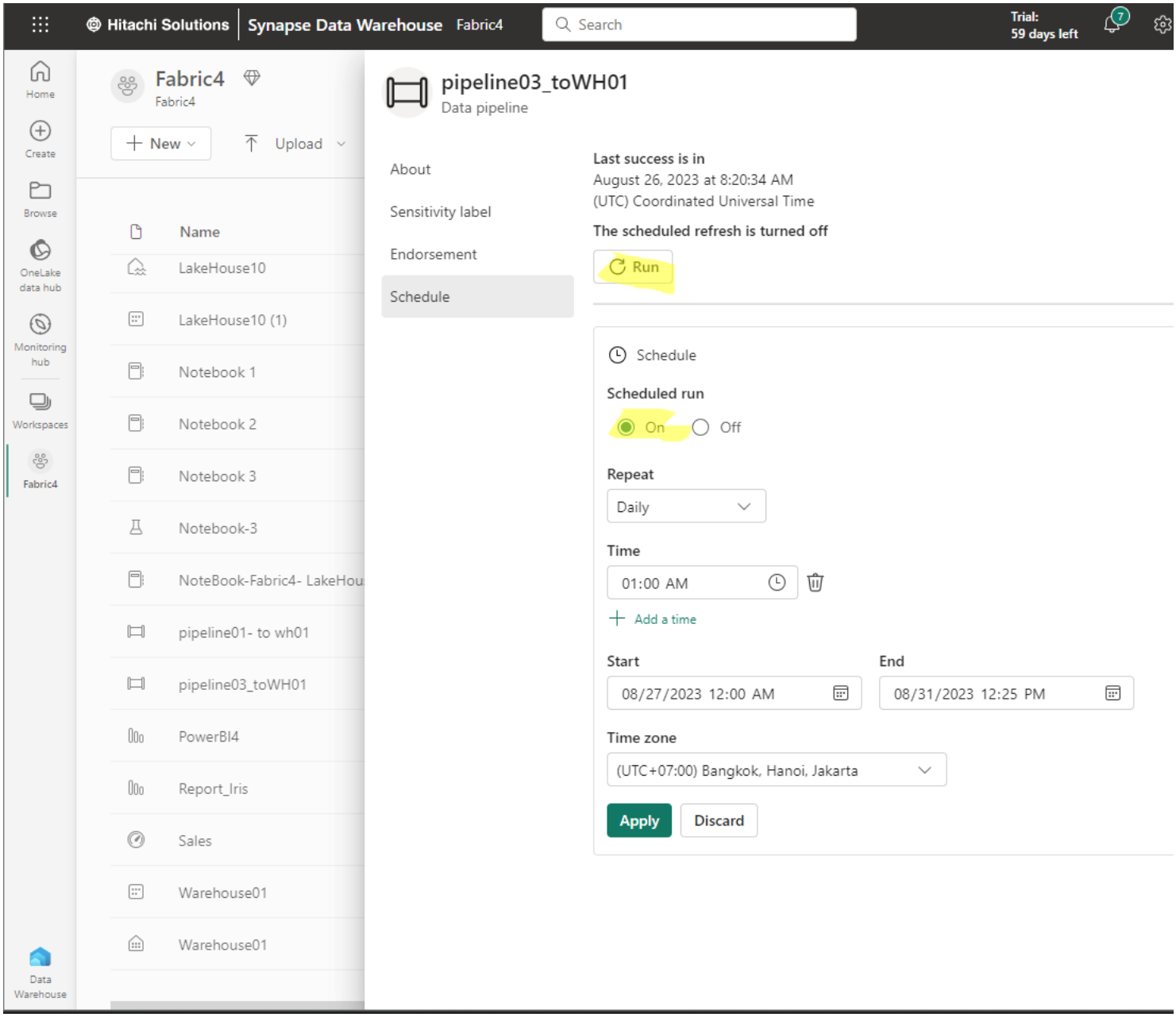






## Duplicated data is not brought to Fabric




Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
Data Warehouse
Fabric4
Fabric4
New
Upload
Create deployment pipeline
Name
Type
Owner
LakeHouse10
Lakehouse
Hien Tran
LakeHouse10 (1)
Notebook 1
Notebook 2
Notebook 3
Notebook-3
NoteBook-Fabric4- LakeHouse10
pipeline01- to wh01
pipeline03_toWH01
Data pipeline
Hien Tran
PowerBI4
Report
Fabric4
Report_Iris
Report
Fabric4
Sales
Dashboard
Fabric4
Warehouse01
Dataset (default)
Fabric4
Warehouse01
Warehouse
Hien Tran
Open
Delete
Settings
Add to Favorites
View lineage
View details
Schedule
Recent runs
Save as


Hitachi Solutions
Synapse Data Warehouse
Fabric4
Search
Trial:
59 days left
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric4
Data Warehouse
Fabric4
Fabric4
New
Upload
Name
LakeHouse10
LakeHouse10 (1)
Notebook 1
Notebook 2
Notebook 3
Notebook-3
NoteBook-Fabric4- LakeHou
pipeline01- to wh01
pipeline03_toWH01
PowerBI4
Report_Iris
Sales
Warehouse01
Warehouse01
pipeline03_toWH01
Data pipeline
About
Sensitivity label
Endorsement
Schedule
Last success is in
August 26, 2023 at 8:20:34 AM
(UTC) Coordinated Universal Time
The scheduled refresh is turned off
Run
Schedule
Scheduled run
On
Off
Repeat
Daily
Time
01:00 AM
Add a time
Start
08/27/2023 12:00 AM
End
08/31/2023 12:25 PM
Time zone
(UTC+07:00) Bangkok, Hanoi, Jakarta
Apply
Discard