MICROSOFT FABRIC FOR ANALYTICS

Analytics use case for Microsoft Fabric

Table of Contents

1. Introduction

This training serves as introduction to end to end analytics in Microsoft Fabric.

Participants will work on ingesting data, transforming them, analyze them as a model following Medallion Architecture basis. For data ingestion, in this training we will work on multiple csv files for sales data and an excel for dimensional data. Participants will ingest the data using Dataflow and Notebook. The ingested data will be stored in lakehouse as Bronze/Silver Layer. Afterwards by using another Dataflow, data will be copied to Warehouse as Gold Layer. Power BI will then be used to create simple report from Gold Layer.

For this training participants will use sales data of 3 sample country (USA, Japan, Australia). Those data will then be combined with date, geography, products, and manufacturer as supporting dimensions.

Participants will work with basic objects and features on analytics in Microsoft Fabric like:

- Data Factory Features
  - Dataflow gen2
  - Data pipeline
- Data Engineering Features
  - Lakehouse
  - Notebook
- Data Warehouse Features
  - Warehouse
- Power BI Features
  - Power BI Report
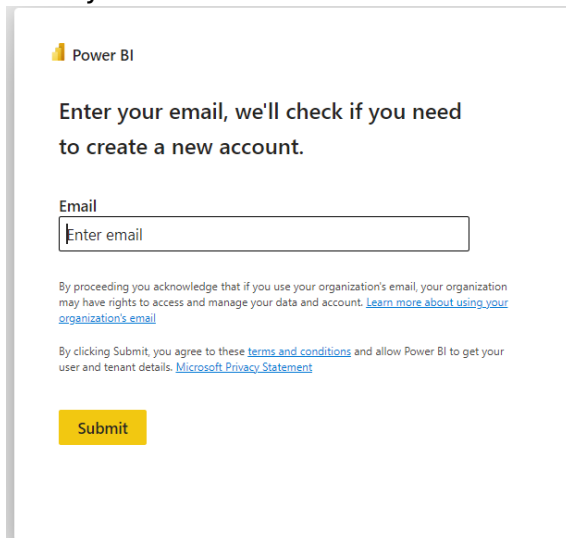
Solution Architecture
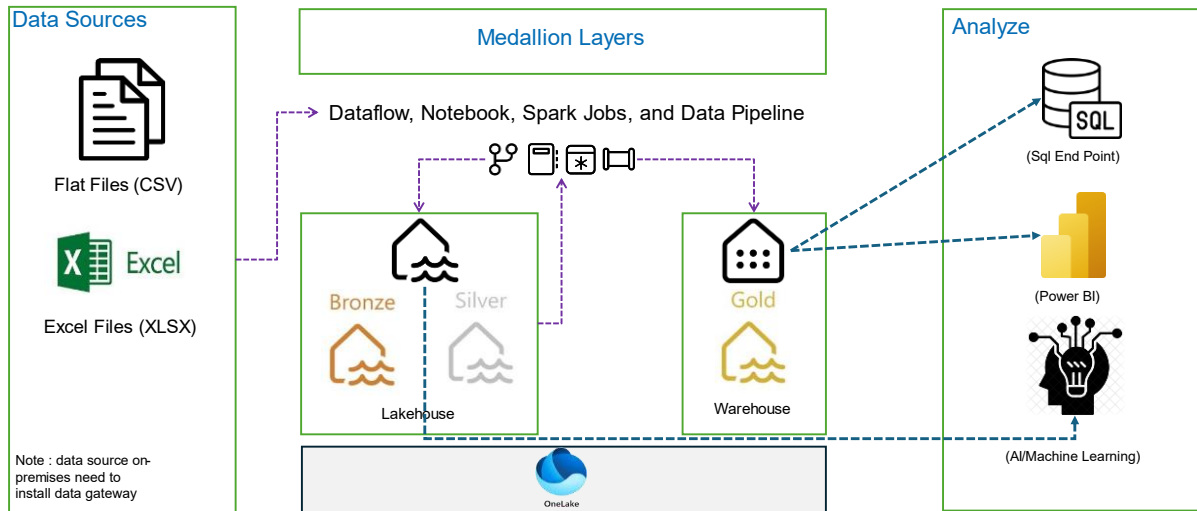
2. Prerequisite

Each user must have all items prepared before the lab:

1. Web Browser to access Power BI Service on https://app.powerbi.com/ .
   Google Chrome is recommended.
2. Each user has Power BI license with fabric trial activated or available for
   upgrade (lab setup will cover for this scenario), user that has previously upgrade
   to trial with expired license will need to get new user login.

3. Fabric Lab Setup

3.1.   Fabric License

1. Open the **browser** and navigate to https://app.powerbi.com/. You will be navigated to the login page.
2. Enter your **Email** account and click **Submit.**

3. You will be navigated to the **Password** screen or company sign in page. Enter your password.
4. Click **Sign in** and follow the prompts to sign into Fabric.
5. You will be navigated to the familiar **Power BI Service Home page**.
6. On the top right corner of the screen, select the **user icon**.
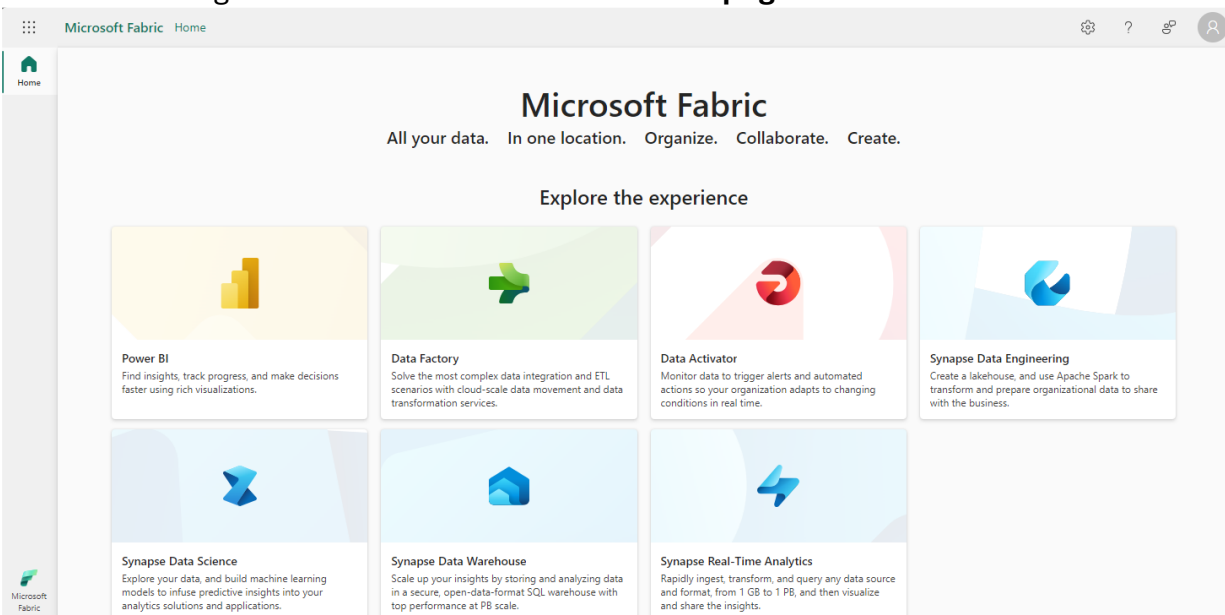7. Select **Start trial**.
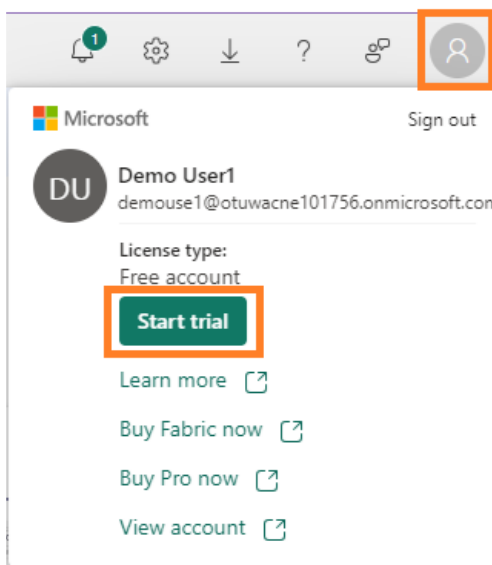
8. Upgrade to a free Microsoft Fabric trial dialog opens. Select **Start trial**.
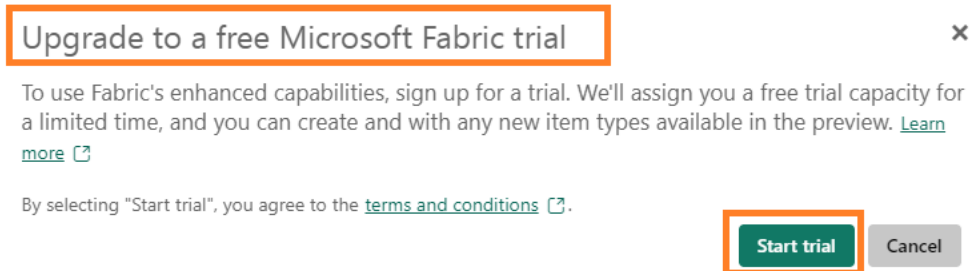
9. Successfully upgraded to a free Microsoft Fabric trial dialog opens. Select **Fabric Home Page**.
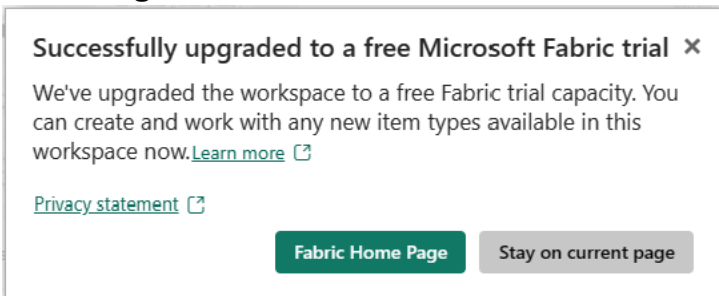
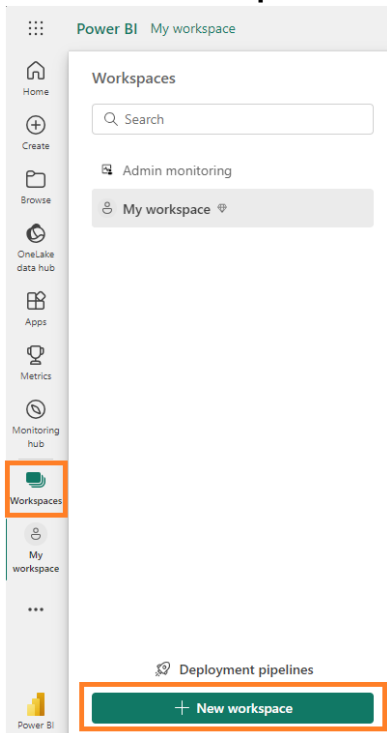10. You will be navigated to the **Microsoft Fabric Home page**.

3.2.   Resource Setup

3.2.1. Fabric Workspace

1. Now let's create a workspace with Fabric license. Select **Workspaces** from the left navigation bar. A dialog opens.

2. Select **New workspace**.

3. **Create a workspace** dialog opens on the right side of the browser.
4. In the **Name** field enter **FABRIC_<username>**
5. If you choose, you can enter a **Description** for the workspace. This is an optional field.
6. Click on **Advanced** to expand the section.

7. Under **License mode**, make sure **Trial** is selected. (It should be selected by default.)
8. Select **Apply** to create a new workspace.

3.2.2. Lakehouse (Bronze-Silver)

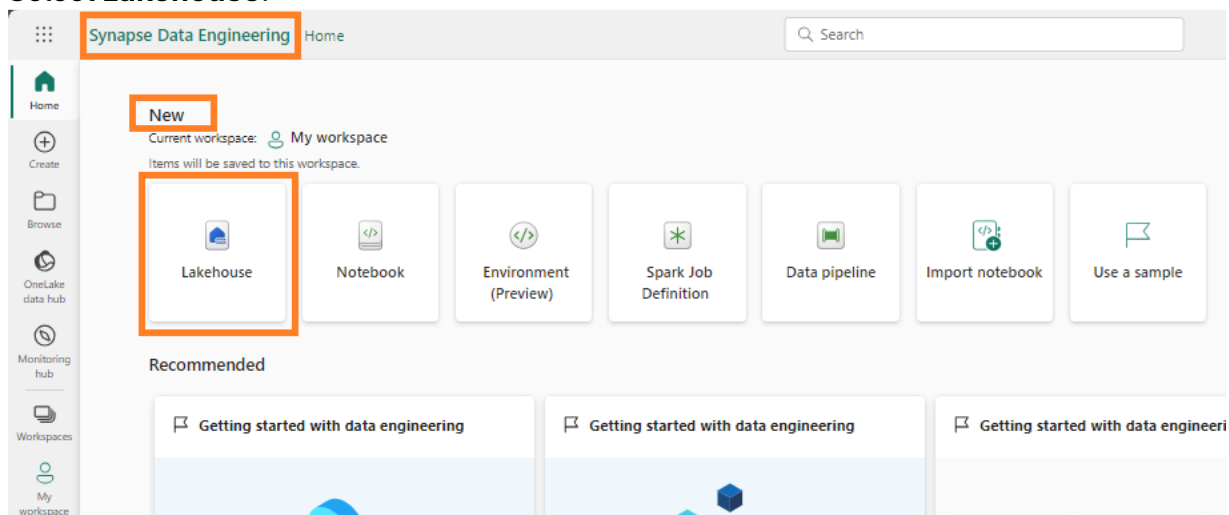1. Select **Data Engineering** to be navigated to Data Engineering Home page.
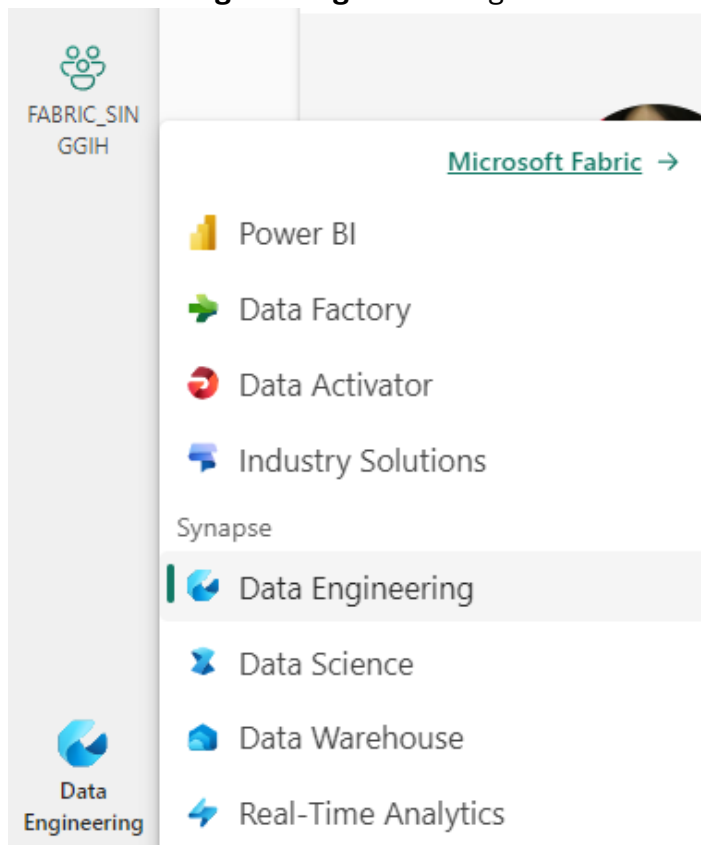
2. Select **Lakehouse**.

3. New lakehouse dialog opens. Type **lh_fabric** in the Name textbox.
4. Select **Create**.
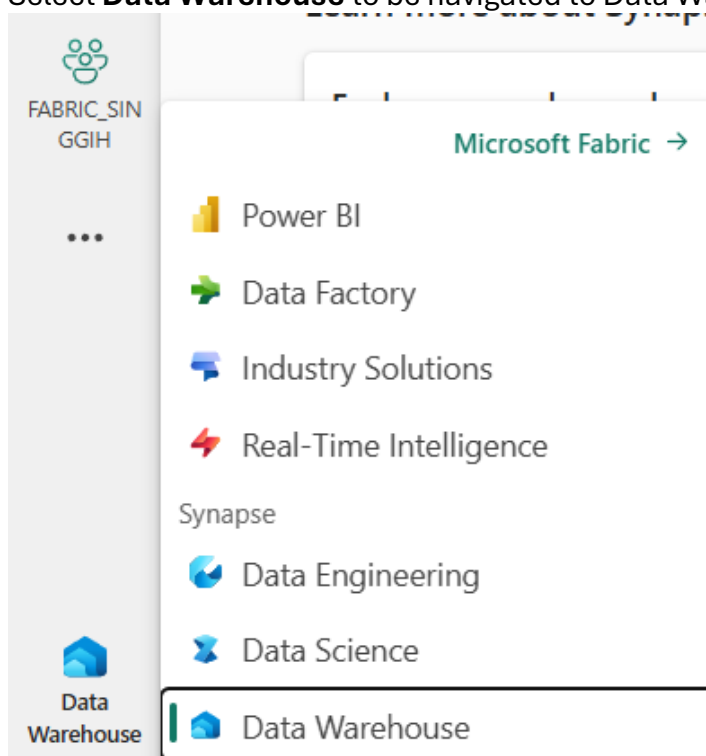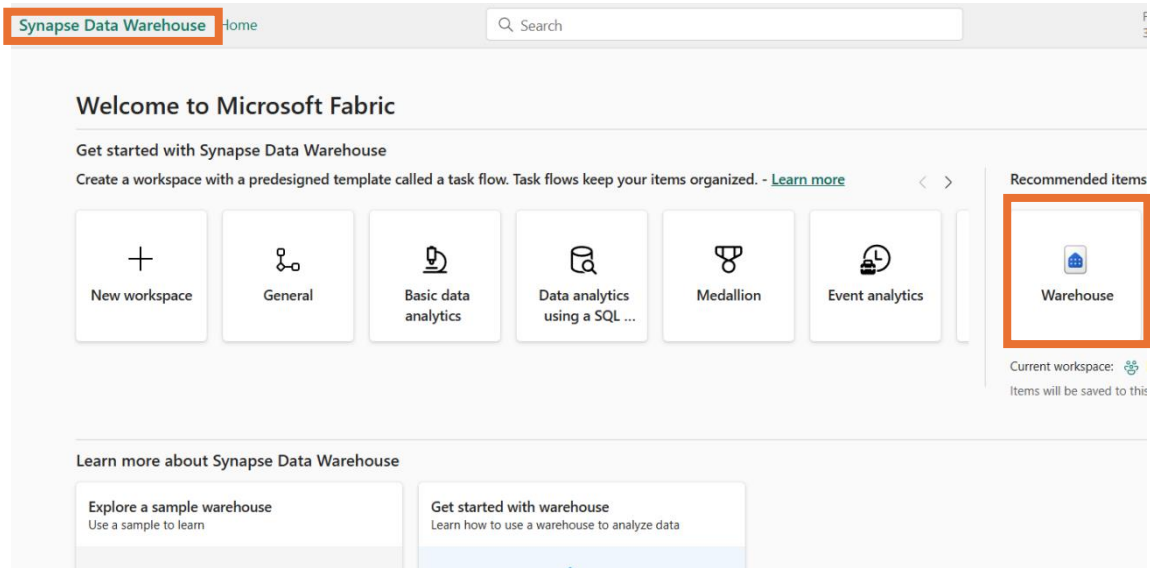
3.2.3. Warehouse (Gold)

1. Select **Data Warehouse** to be navigated to Data Warehouse Home page.

2. Select **Warehouse**.

3. When new warehouse dialog opens. Type **wh_fabric** in the Name textbox.
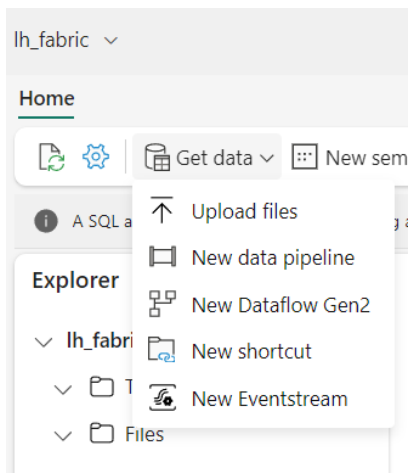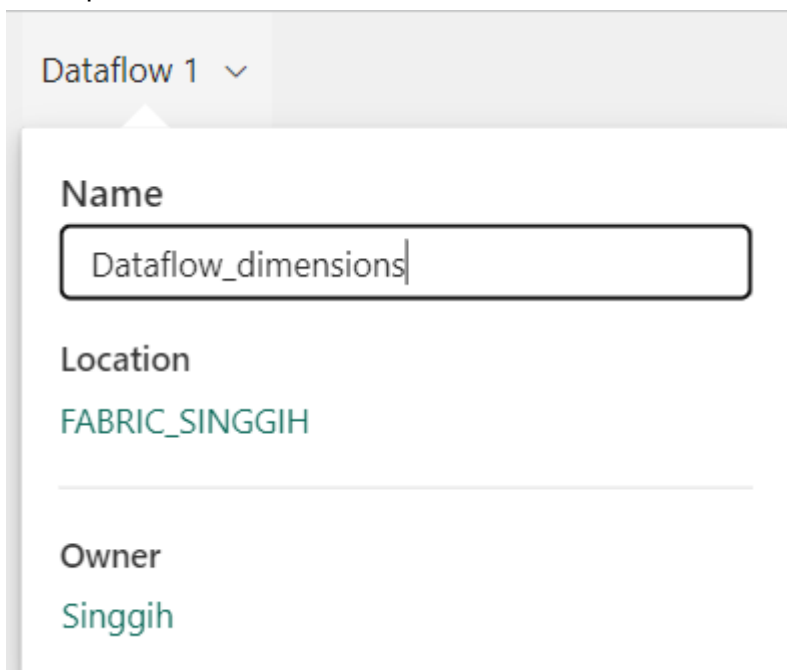4. Select **Create**.

4. Populate Medallion Architecture

4.1.   Get Dimensional Data from Excel (Bronze-Silver)

1. On the main page of the lakehouse, choose get data option and select New Dataflow Gen2

2. Change the created dataflow name by clicking on the dataflow default name on the top left.

3. Next choose Get Data and pick Excel Workbook as the new source
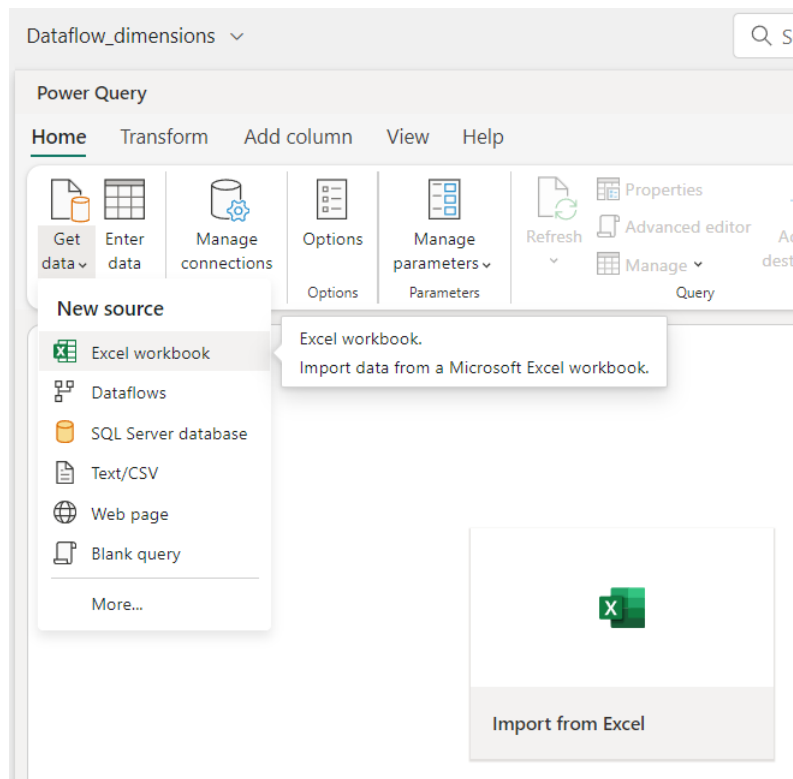
4. Select **Link to file** and under the **File path or URL** enter this link
   https://raw.githubusercontent.com/Singgoli/FabricTrainingResource/main/Data /DIM.xlsx

5. Under **Connection** choose Create new connection, under **Authentication kind** choose Anonymous.

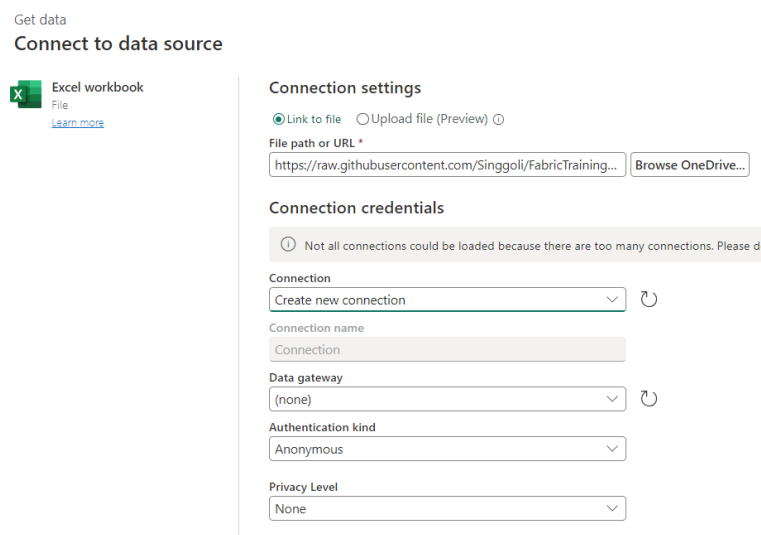6. On the **Choose Data** window, any sheets or tables in the excel will be listed. For this exercise, tick the 4 worksheets below.
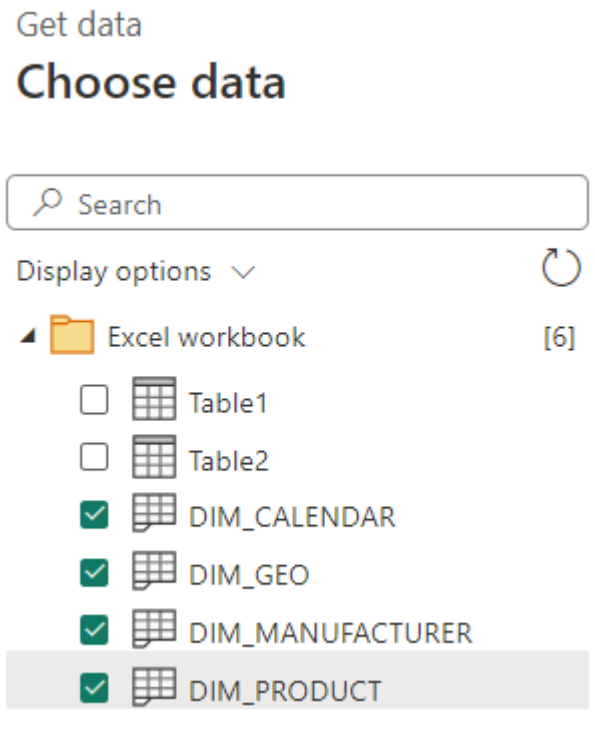
7. Then click **Create**.
8. This will result in 4 queries created for each worksheets
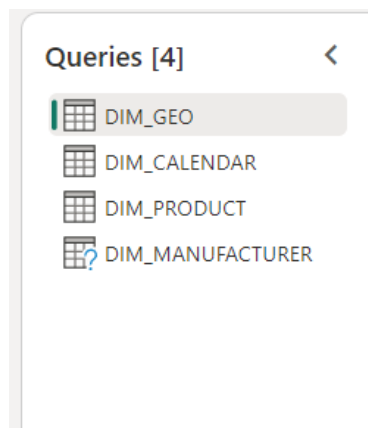
9. Select the query for DIM_GEO, Change the data type for **Zip** column to Text by clicking the 123 symbol
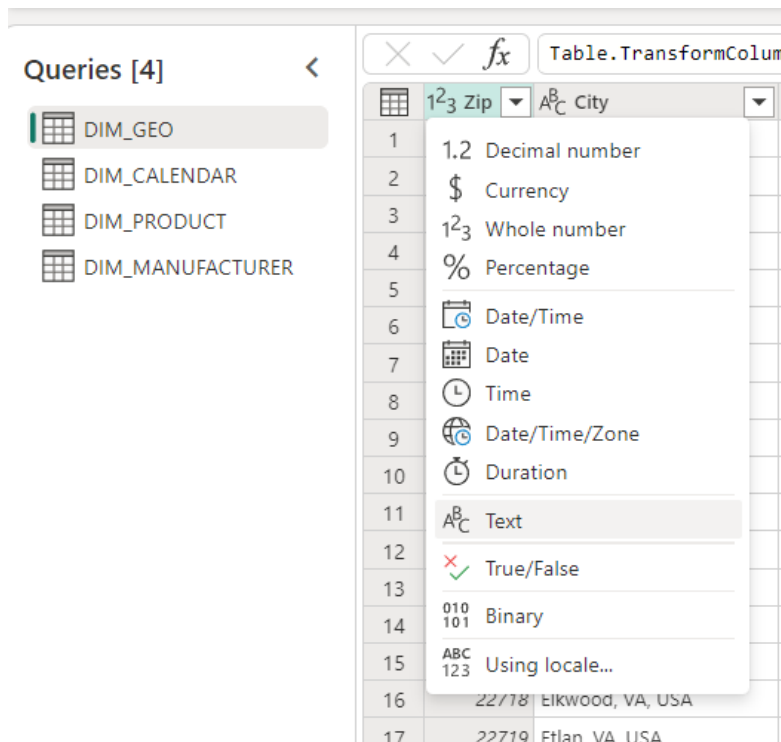
10. Choose Replace Current when prompted to replace the suggested automatic data type conversion.
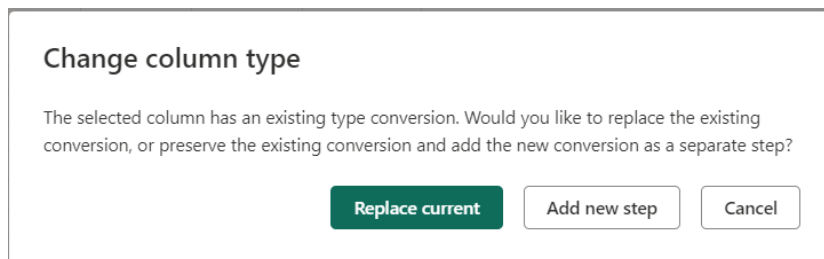
11. Next go to **DIM_PRODUCT** query and select the **Price** column. On the **Transform** toolbar, select Split Column and choose by positions.

12. Under the **Positions** enter 0, 4 to split the data into 2 column by 1st character and 5th character.
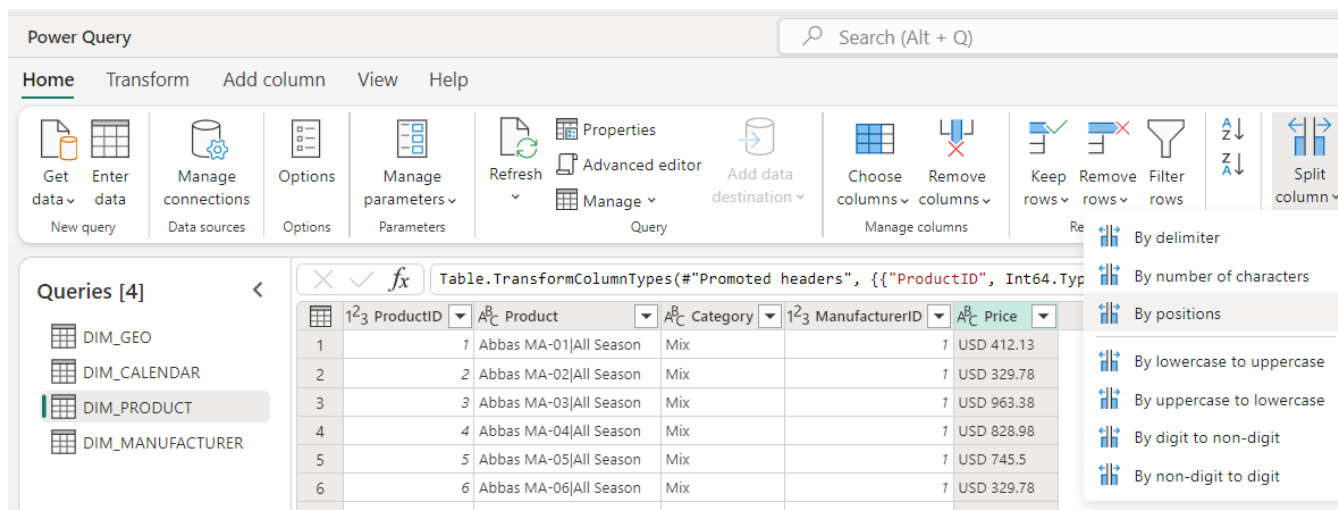

13. It will split the data to currency and amount columns

14. Rename the 2 result columns as **Currency** and **MSRP** by choosing **Rename** in dropdown menu after right clicking the column name or double clicking the column name.

15. Go back to DIM_GEO and select **Zip** and **Country** columns by holding **CTRL** when clicking the 2 columns. Then go to **Add column** toolbar and select

Column from examples>From selection.
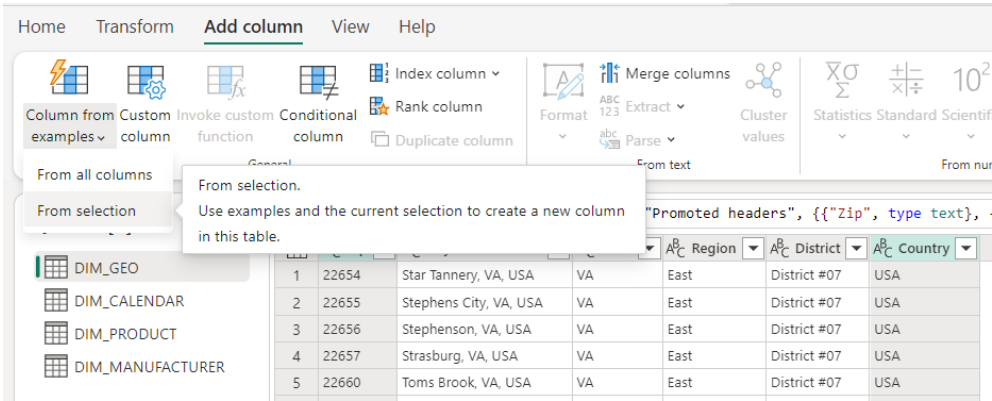
16. Input the value in **Zip** followed by " – " and then value from **Country** in the first row. It will then automatically generate the equivalent script from the example. Select **OK** if the sample shows the desired results.
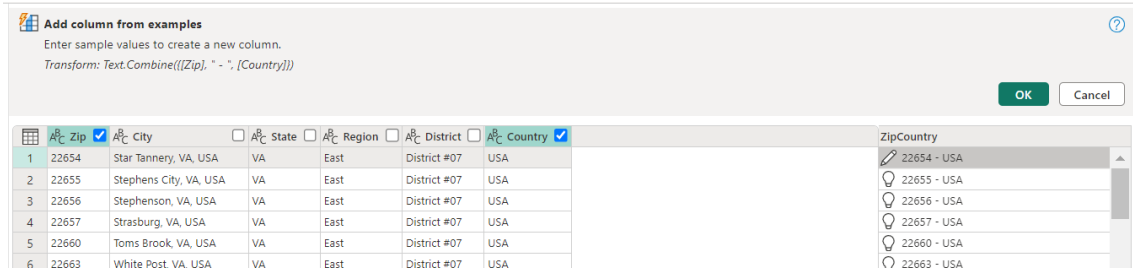
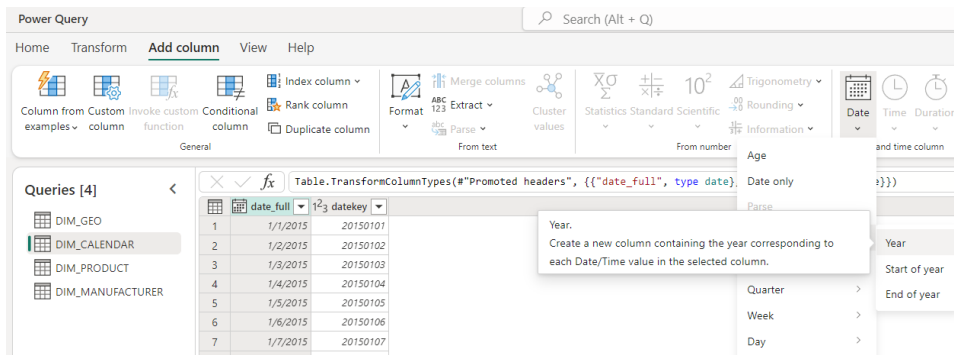17. Next select **DIM_CALENDAR** and column **date_full**, on the **Add column** toolbar choose the **Date > Year > Year** option. This will add a new column consisting of the year value of the date.

18. Repeat the process to add Month, Name of month and Day. Also change the data type of **Year**, **Month** and **Day** to whole number.
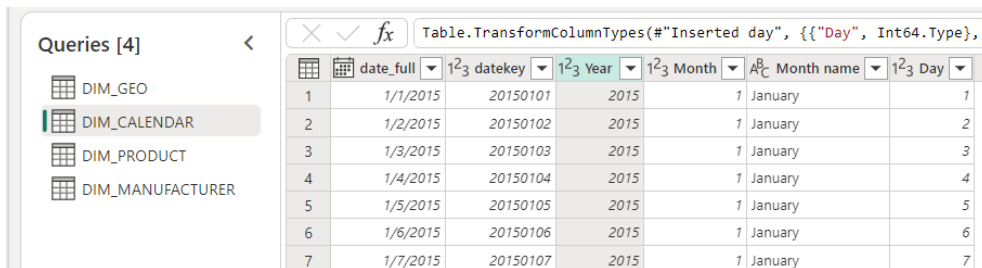
19. Select any query and check on the bottom right for **Data destination** option. Click on the cross to add data destination and select **Lakehouse**.
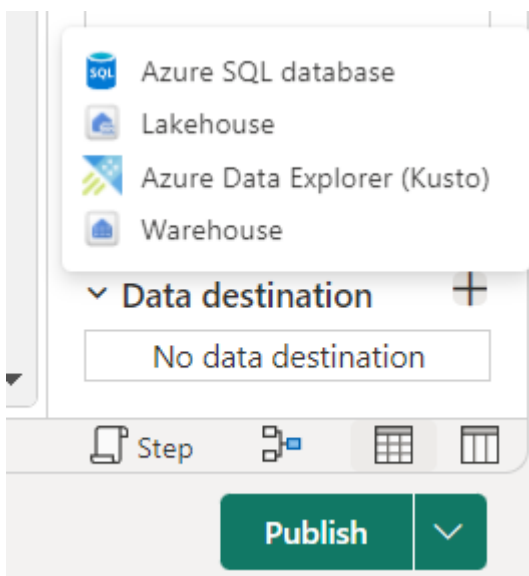
20. Choose to **Create a new connection** and set **Authentication kind** to **Organizational account**.

21. Next select the workspace and lakehouse created earlier during the lab and select **New table** instead of existing. Change the table name as needed then click **Next**.

22. Disable the **Use automatic settings** option and select **Replace** as **Update method** also select **Fixed schema** for **Schema options on publish**. Click **Save**

settings.
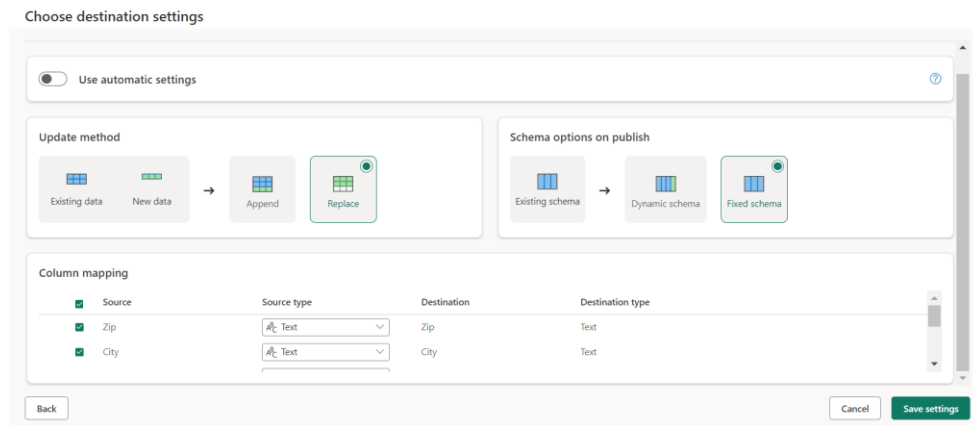
23. Repeat the process to add destination for the other 3 queries.
24. Once finished publish the dataflow using Publish button on bottom right.
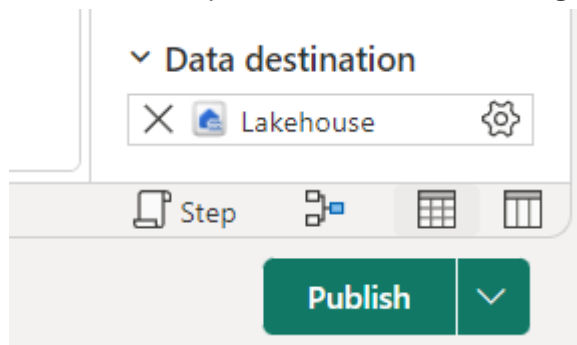
4.2.   Get Fact Data from Flat File (Bronze-Silver)

1. Open your lakehouse, from the ribbon on top choose Open notebook and choose New notebook.
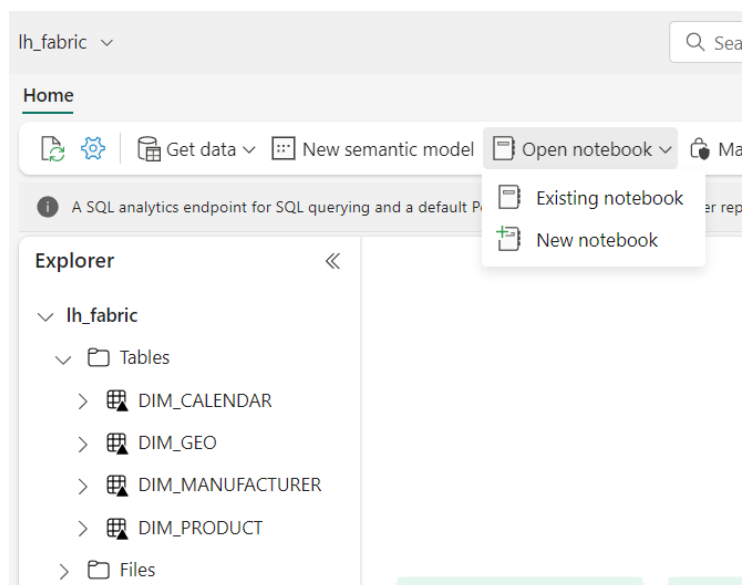
2. Once opened change the value text the existing cell and change the cell into a **Markdown cell.**
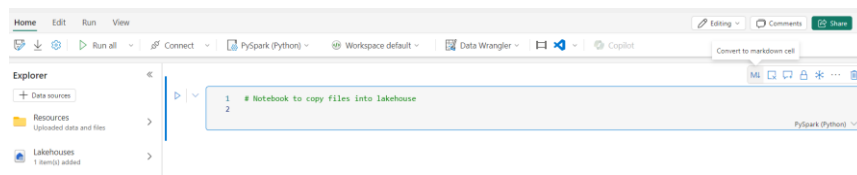
3. After that create a new Code cell by selecting **+ Code** below the existing cell.
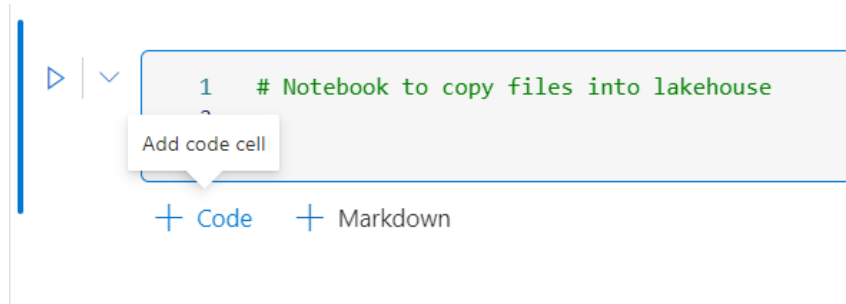
4. Copy the following code into the code cell

```python
import pandas as pd
df = pd.read_csv("https://raw.githubusercontent.com/Singgoli/FabricTrainingResource/main/Data/Sales/USA.csv")
df2 = pd.read_csv("https://raw.githubusercontent.com/Singgoli/FabricTrainingResource/main/Data/Sales/Japan.csv")
df3 = pd.read_csv("https://raw.githubusercontent.com/Singgoli/FabricTrainingResource/main/Data/Sales/Australia.csv")

def clean_data(df):
    # Created column 'Country' from formula
    df['Country'] = "USA"
    return df

df_clean = clean_data(df.copy())

df_sales = pd.concat([df_clean,df2,df3],axis = 0)

df_sales.to_csv("[PATH HERE]/salesdata.csv")
```
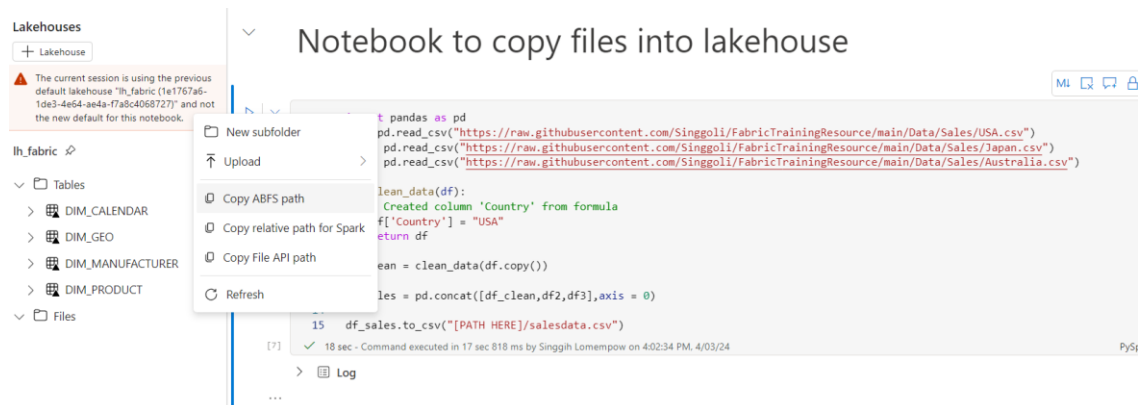
5. On the explorer tab to the left select your lakehouse and find the Files folder. Click the ellipsis to the right of it and choose **Copy ABFS path.**
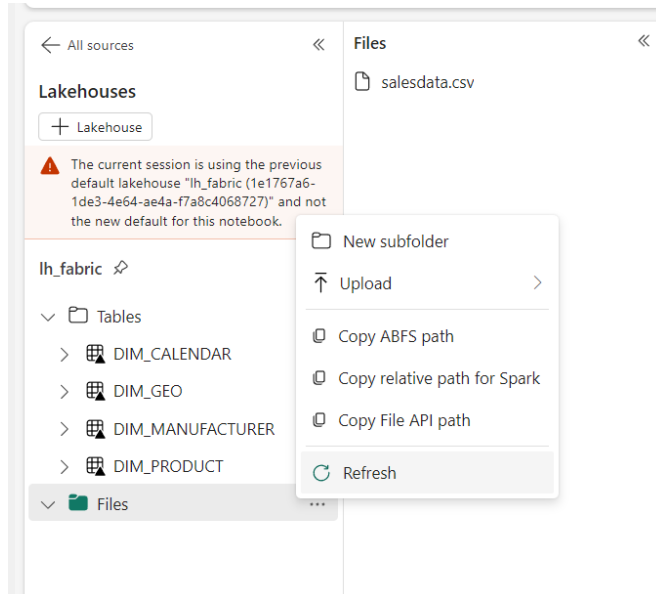

6. Replace the **[PATH HERE]** segment on the code with the ABFS path copied. It will look like this

```
df_sales.to_csv("abfss://FABRIC_SINGGIH@onelake.dfs.fabric.microsoft.com/lh_fabric.Lakehouse/Files/salesdata.csv")
```

7. Next run the notebook using **Run all** option
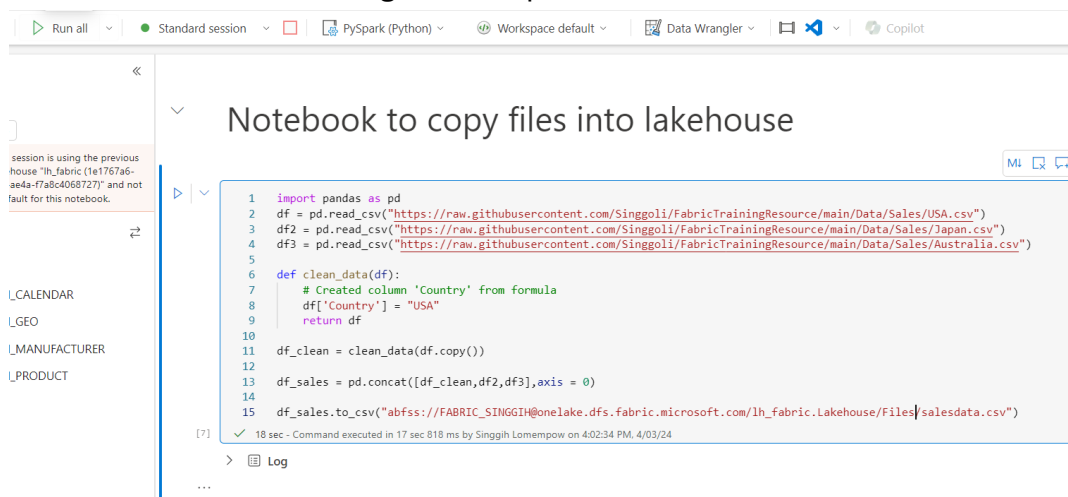
8. Once finished refresh the Files folder and click it, it will show the file salesdata.csv


9. Add a new code cell and copy this command there and then run the code cell.

```python
import pandas as pd
# Load data into pandas DataFrame from "/lakehouse/default/" +
"Files/salesdata.csv"
df4 = pd.read_csv("/lakehouse/default/" + "Files/salesdata.csv")
def clean_data(df4):
    df4 = df4.drop(columns=['Unnamed: 0'])
    df4['Date'] = pd.to_datetime(df4['Date'])
    df4['Zip'] = df4['Zip'].astype(str)
    df4['ZipCountry'] = df4.apply(lambda x : x['Zip'] + ' - ' +
x['Country'], axis = 1)
    return df4

df5 = clean_data(df4.copy())
dfspark = spark.createDataFrame(df5)

dfspark.write.format("delta").mode("overwrite").saveAsTable("FACT_SAL
ES")
```
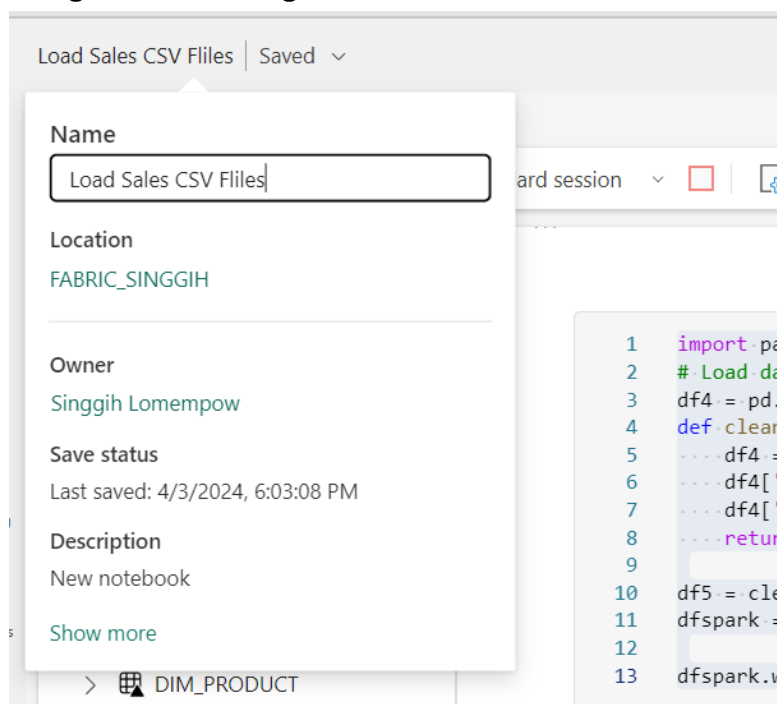
10. Once the cell finished running refresh the Tables folder. There will be a new table named FACT_SALES. To preview the data load it to spark and run the created code cell.
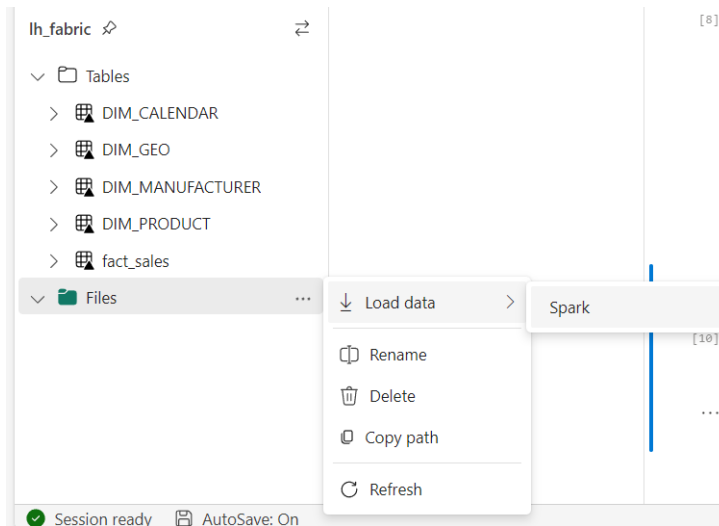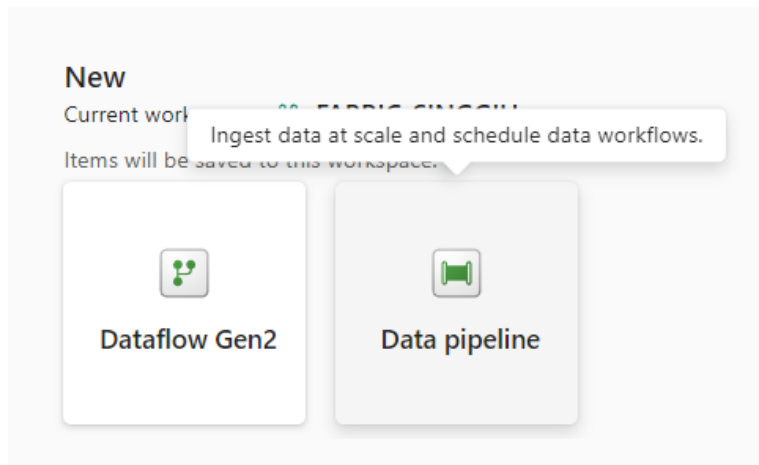

11. Finally rename the notebook from the top left menu. And stop the spark session using the red rectangular button.
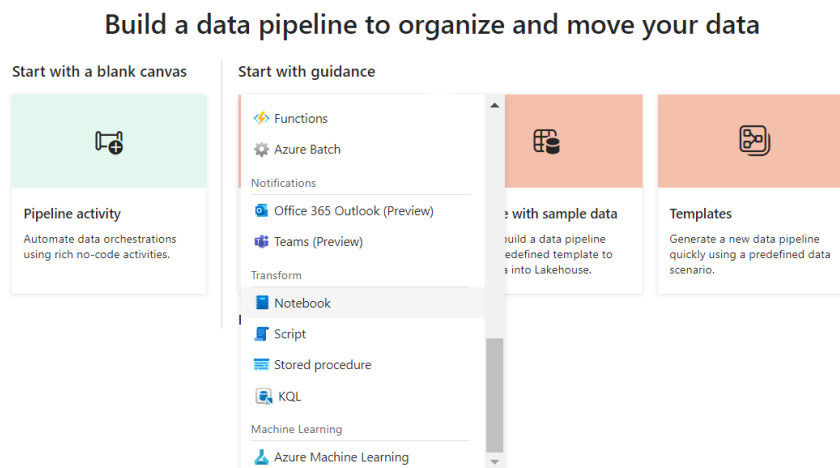
## 4.3. Automate the data ingestion using data pipeline

1. Go back to your created fabric workspace, click on the create new Data pipeline.
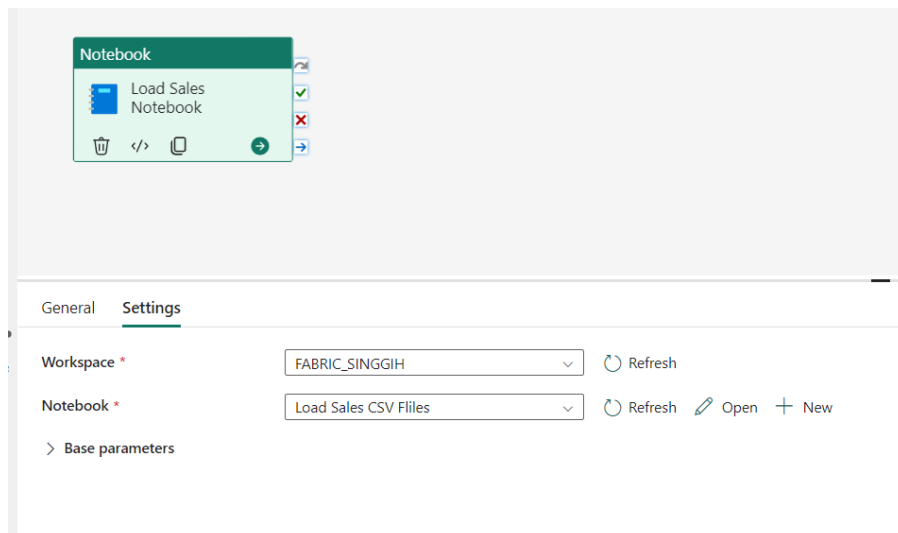

2. You will be prompted to name the pipeline, name it **pipeline_silver**.
3. Choose to start with a blank canvas and a pipeline activity, then on the selection choose Transform>Notebook.
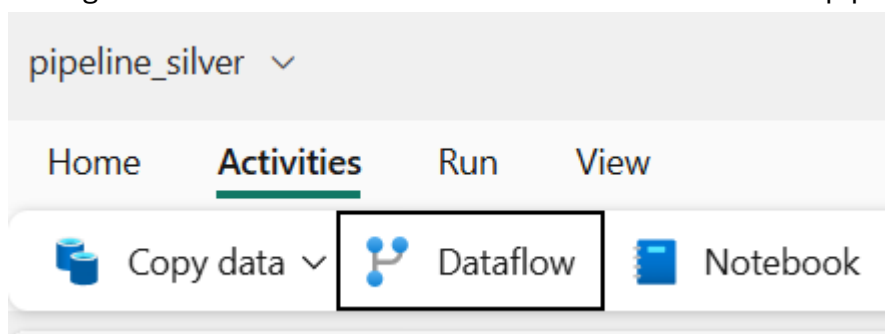
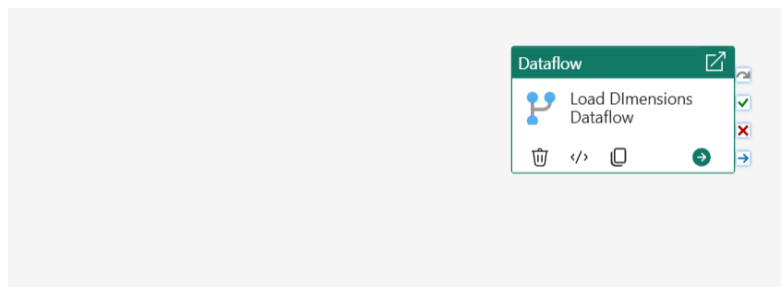4. Rename the pipeline object into **Load Sales Notebook** on the General tab.
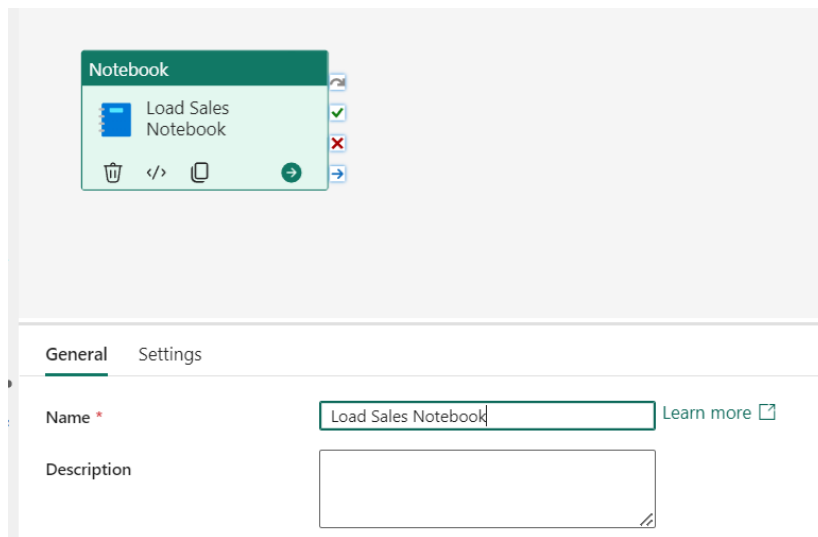
5. Next, go to the Settings tab and select the **Load Sales CSV Files** notebook.

6. Change tab to Activities and click Dataflow to add Dataflow pipeline object

7. Rename the object into **Load Dimensions Dataflow**

8. Change to Settings Tab and select the previously created dataflow in the dataflow dropdown menu
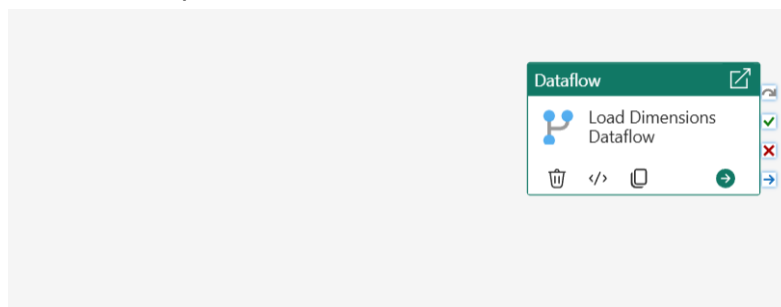
9. Next connect the 2 created objects by dragging the On Success condition from the dataflow object into the notebook object
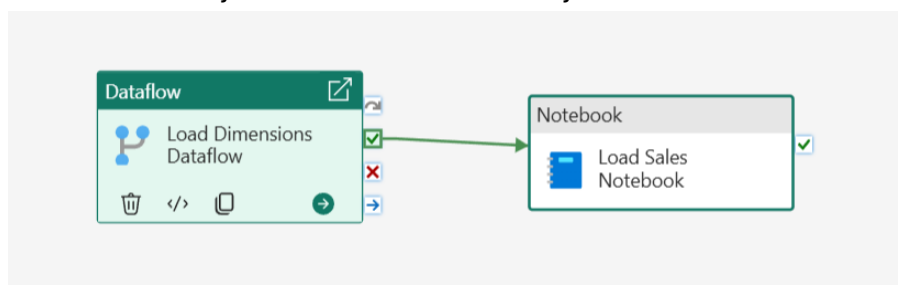
10. Go Back to Home tab and select Validate to check the pipeline
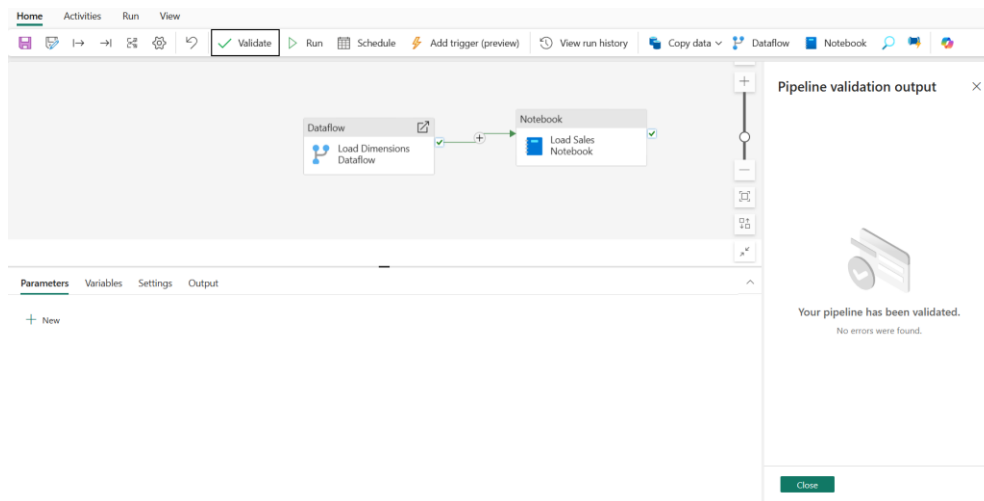
11. Test run the pipeline using the run button (You may be prompted to save the pipeline before running).
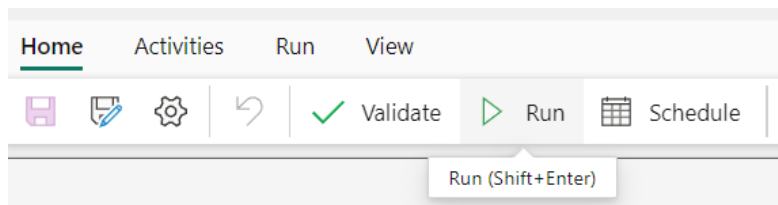
12. Wait until the run completed, you can see the execution result in the bottom pane

13. To create a scheduled automation, click the schedule button.

14. Configure the schedule as required, for this example the run is scheduled daily on 12.00 AM for 2 days. Once finished click Apply.

pipeline_fact
Data pipeline

About

Endorsement
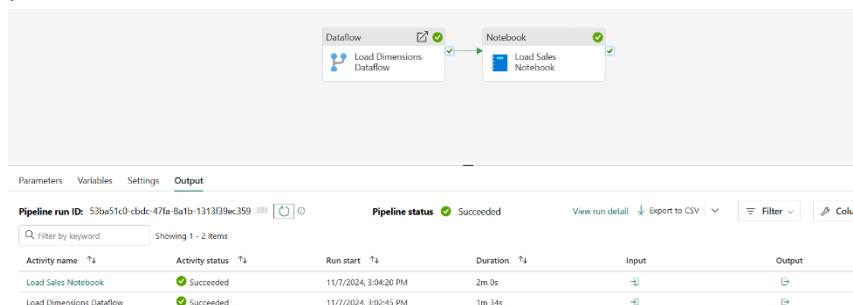
Schedule

The scheduled refresh is turned off

↻ Run

🕐 Schedule

Scheduled run
◉ On   ◯ Off

Repeat
Daily ⌄

Time
00:00   🕐   🗑

＋ Add a time

Start date and time   End date and time
05/03/2024   📅   05/04/2024   📅

Time zone
(UTC+07:00) Bangkok, Hanoi, Jakarta ⌄

Apply   Discard

15. Once finished you can see the next run schedule above.

Last success is in
May 2, 2024 at 9:53:36 AM
(UTC+07:00) Bangkok, Hanoi, Jakarta
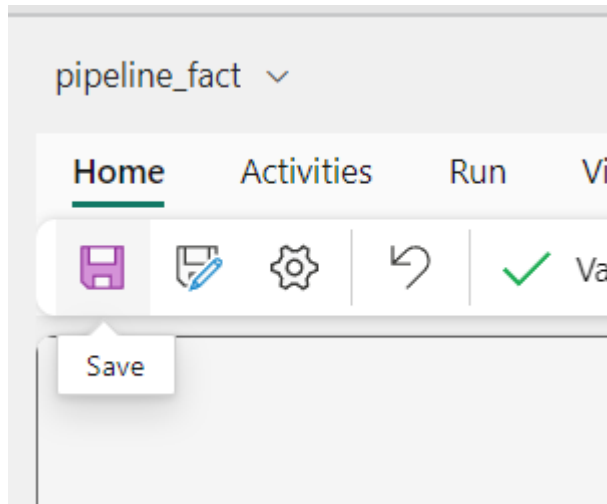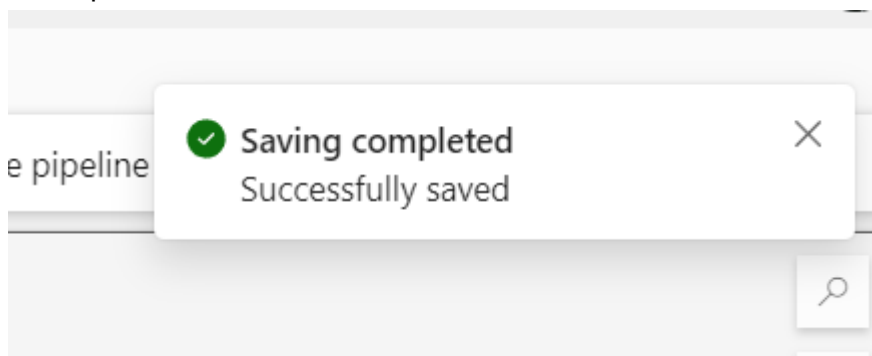
Next refresh in
13 hour(s) 41 minute(s)

↻ Run

16. Save your pipeline using the save button.

17. Once the save is completed, you can close the notebook and go back to fabric workspace.
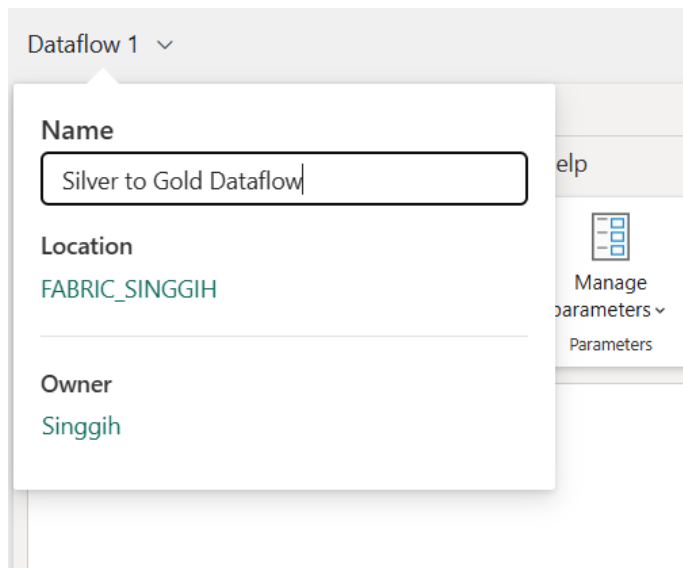

4.4.   Move processed data (Silver-Gold)

1. Open the fabric warehouse created while setting up for lab. Choose **New Dataflow Gen2** to make a new dataflow that will populate the warehouse.
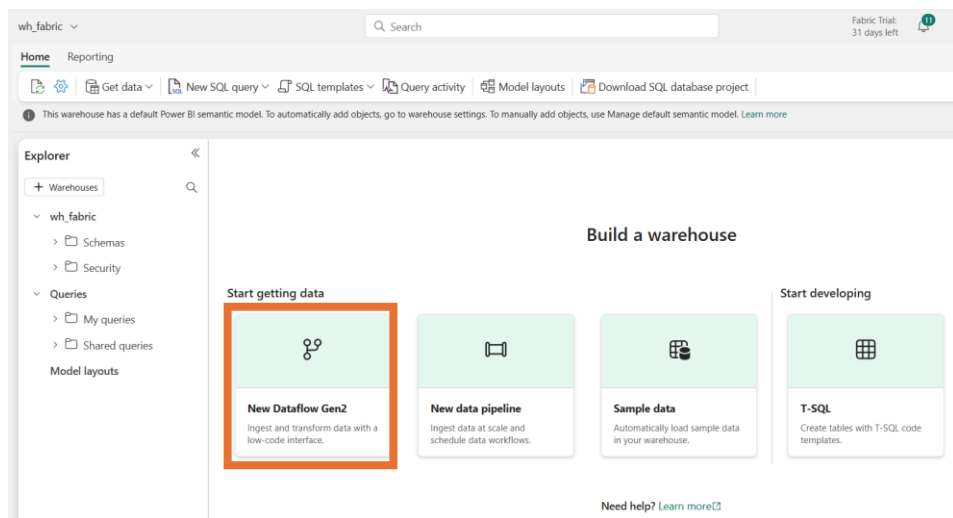
2. Rename the dataflow into **Silver to Gold Dataflow**

3. Click Get Data to open a new data source window
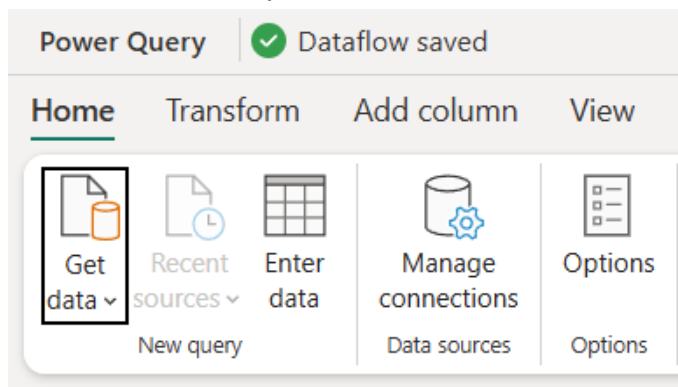
4. On the side choose OneLake Data Hub and select lh_fabric
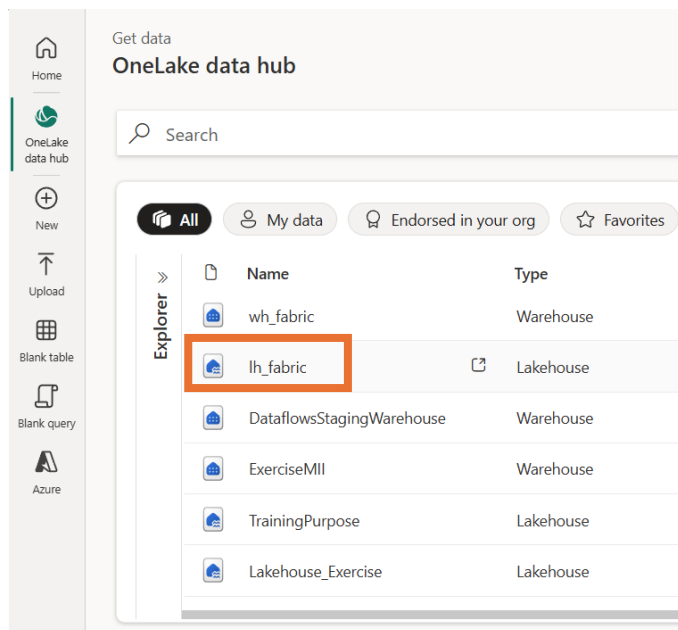
5. Select all the tables made so far and click Create button
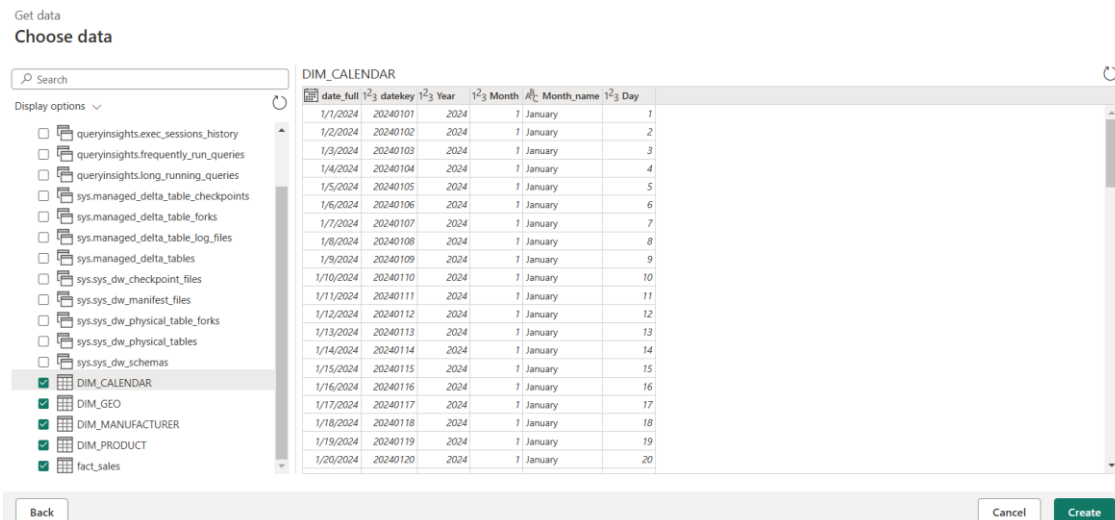
6. Once done you will have queries on the left for each respective tables selected, ON the bottom right it will show that the destination is set to the warehouse by default
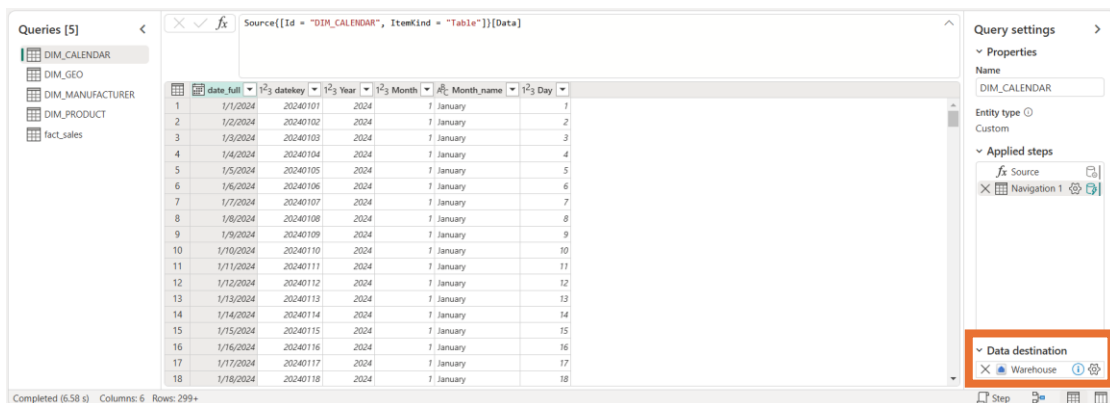
7. Press the gear button to setup the destination tables and update method
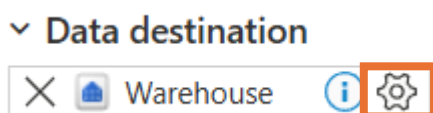
8. If the connection has been setup, click Next
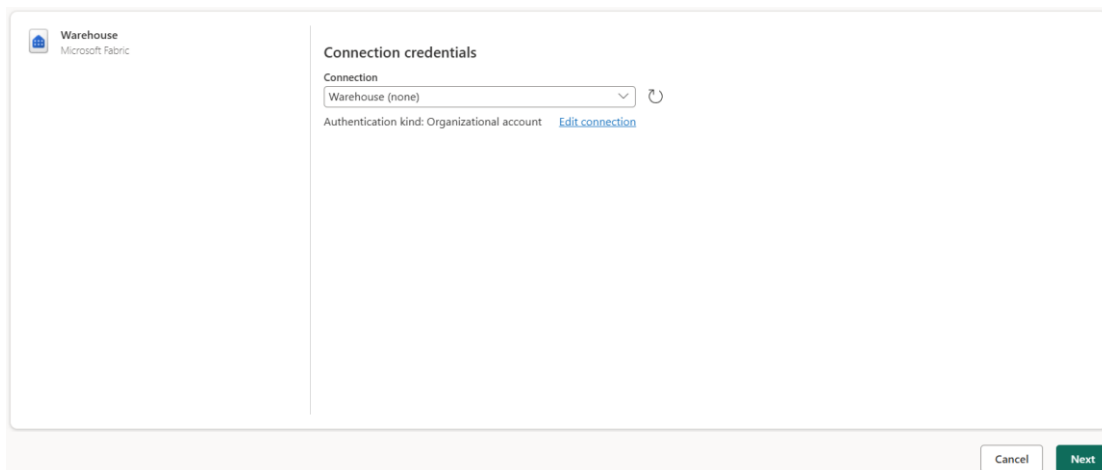
9. Make sure the warehouse selected and table name is correct before clicking Next

10. Next select **Replace** for Update Method and **Fixed schema** for Schema options

11. When done, click save settings

12. Repeat Step 7 to 11 for each queries, for FACT_SALES rename the table to all capitals

13. Once all queries destination set up, click Publish to publish and execute the dataflow
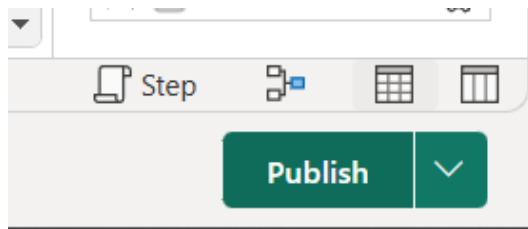
14. You will be taken back to the workspace, wait a bit until the dataflow deployment and execution completed and check the result in the refresh history

4.5. Add the new Dataflow into existing pipeline

1. Open the **pipeline_silver** pipeline and rename it into pipeline_silver_gold

pipeline_silver ⌄
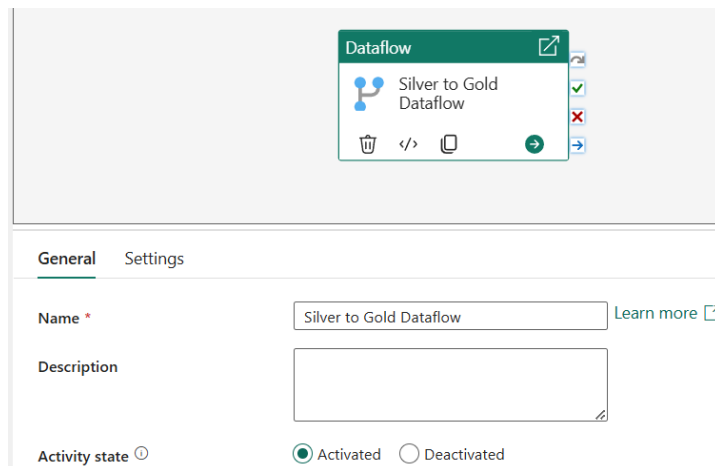
Name

pipeline_silver_gold

Location

FABRIC_SINGGIH

2. Create a new dataflow pipeline object and rename it into **Silver to Gold Dataflow**

Dataflow ⬀	
Silver to Gold Dataflow	

General **Settings**

Name *	Silver to Gold Dataflow	Learn more ⬀
Description		

Activity state ⓘ ⦿ Activated ◯ Deactivated

3. Set it to execute the newly created dataflow

General **Settings**

Workspace *	FABRIC_SINGGIH ⌄	↻ Refresh
Dataflow *	Silver to Gold Dataflow ⌄	↻ Refresh

4. Also set it to be executed after the notebook

Dataflow ⬀	Notebook	Dataflow ⬀
Load Dimensions Dataflow	Load Sales Notebook	Silver to Gold Dataflow
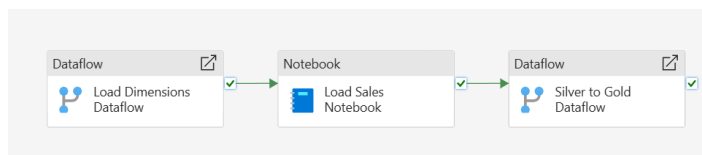
5. You can then save and execute the pipeline

5. Create Report

5.1. Create report from warehouse data

1. Access your wh_fabric warehouse.
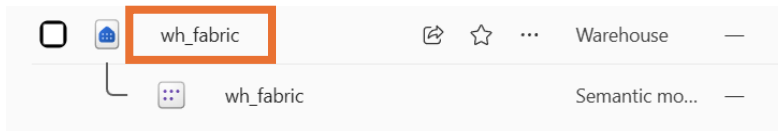
2. Once opened go to **Reporting** tab and click **New semantic model**.
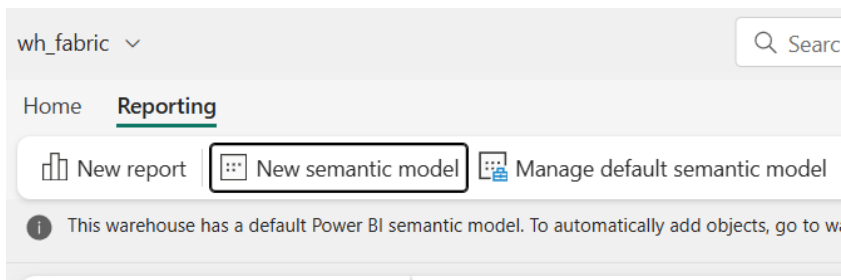
3. Name the semantic model **pbi_fabric_<your name>** and select all table we have created until now. Review your input and selection then click **Confirm**.
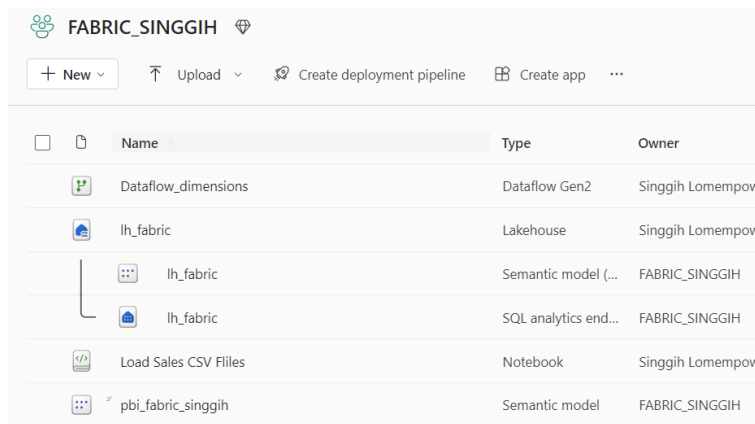
4. Go back to your fabric workspace and confirm the semantic model is created. Open the semantic model by clicking it.

5. Once opened, click Open data model on the menu
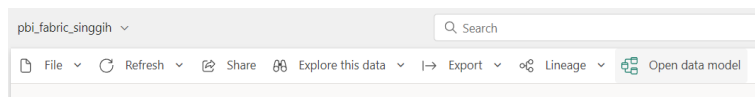
6. Rearrange the tables like the image below to make it easier to view the relationship that will be created later.
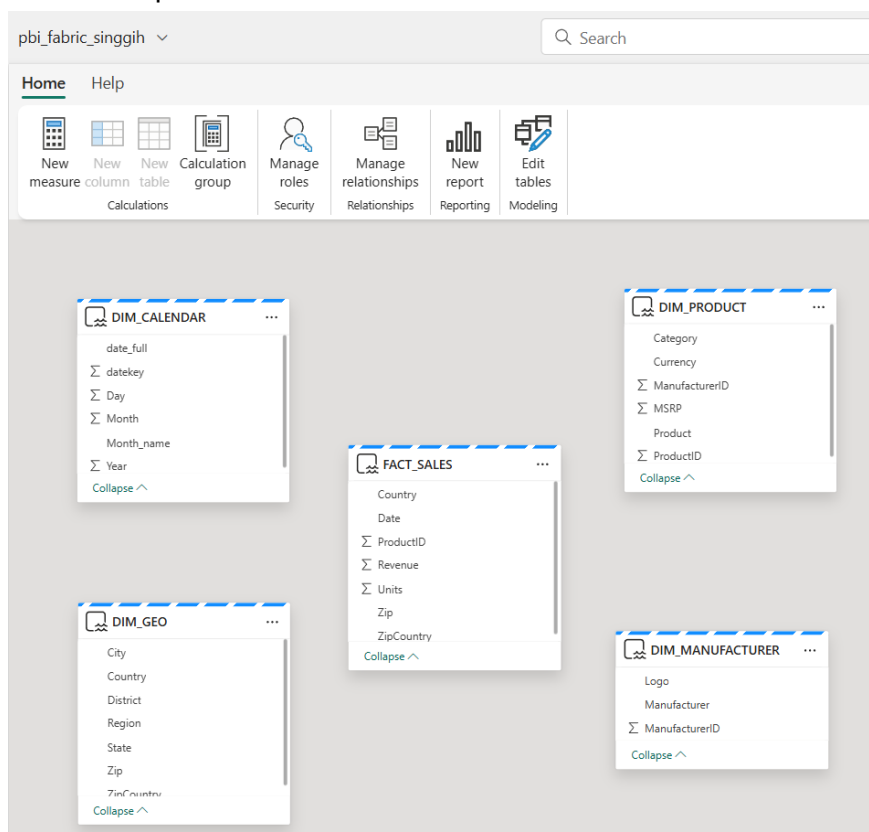
7. Create relationship by dragging foreign key column to the primary key, for example ZipCountry in FACT_SALES to DIM_GEO
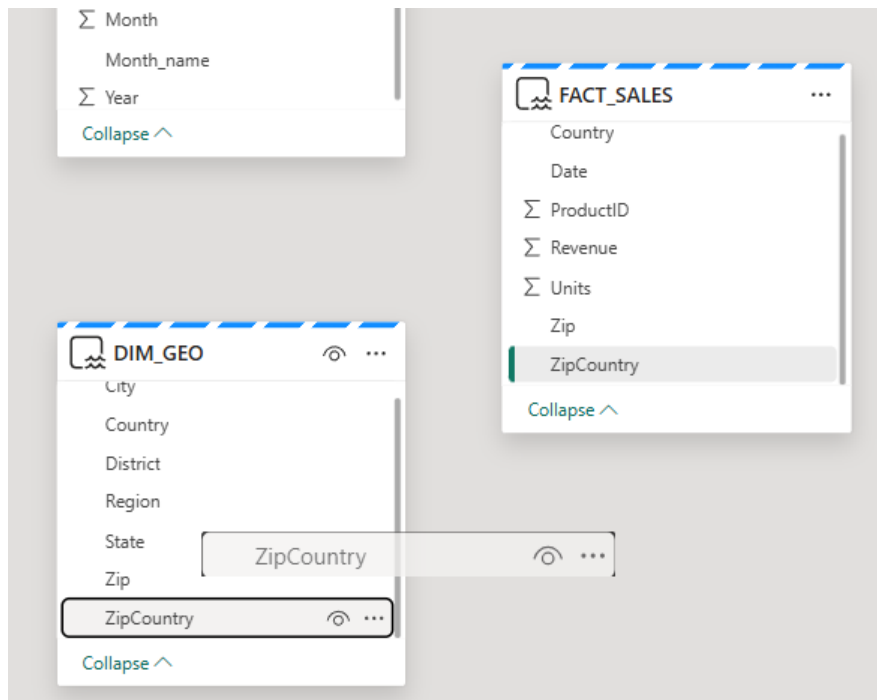
8. Review the **New relationship** dialog :
   - From table : FACT_SALES with ZipCountry selected
   - To table : DIM_GEO with ZipCountry selected
   - Cardinality : Many to one (*:1)
   - Cross-filter direction : Single
   - Make this relationship active : Selected

New relationship

Select tables and columns that are related.

From table

FACT_SALES

Country	Date	ProductID	Revenue	Units	Zip	ZipCountr

A preview of this table isn't available

To table

DIM_GEO

City	Country	District	Region	State	Zip	ZipCountr

A preview of this table isn't available

Cardinality

Many to one (*:1)

Cross-filter direction

Single

☑ Make this relationship active

☐ Apply security filter in both directions

☐ Assume referential integrity. Learn more ↗

[Save] [Cancel]

9. Once done click Ok and the relationship can be seen in form of link between both table and an arrow noting the filter direction.
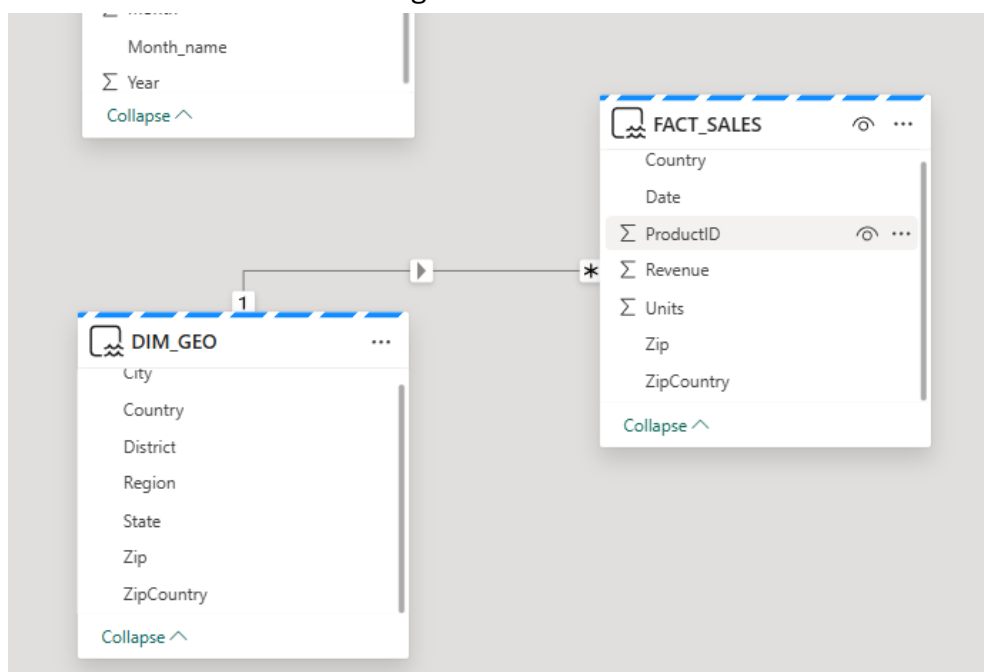
10. Repeat the relationship creation for the other table described below:
- FACT_SALES[Date] to DIM_DATE[date_full]
  note: You can change the data type of the column by changing the properties.

- FACT_SALES[ProductID] to DIM_PRODUCT[ProductId]
- DIM_PRODUCT[ManufacturerId] to DIM_MANUFACTURER[ManufacturerId]

11. Once finished, you will have the data model similar to example below.
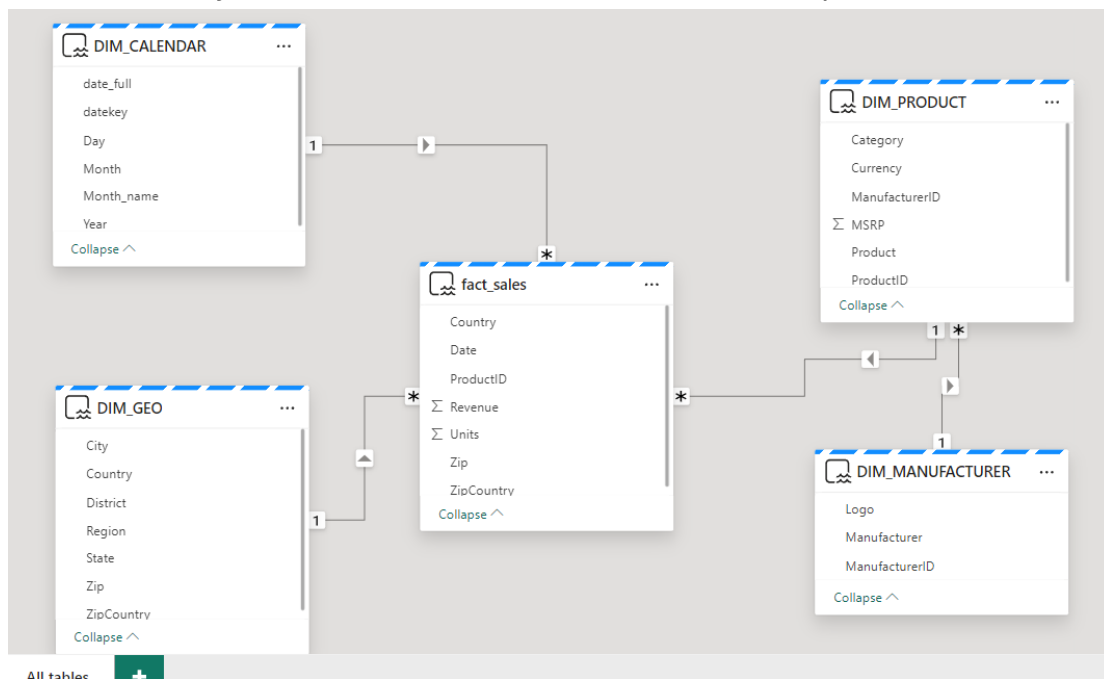
12. Next we will create measures using dax. Select the FACT_SALES table and click **New measure** button.
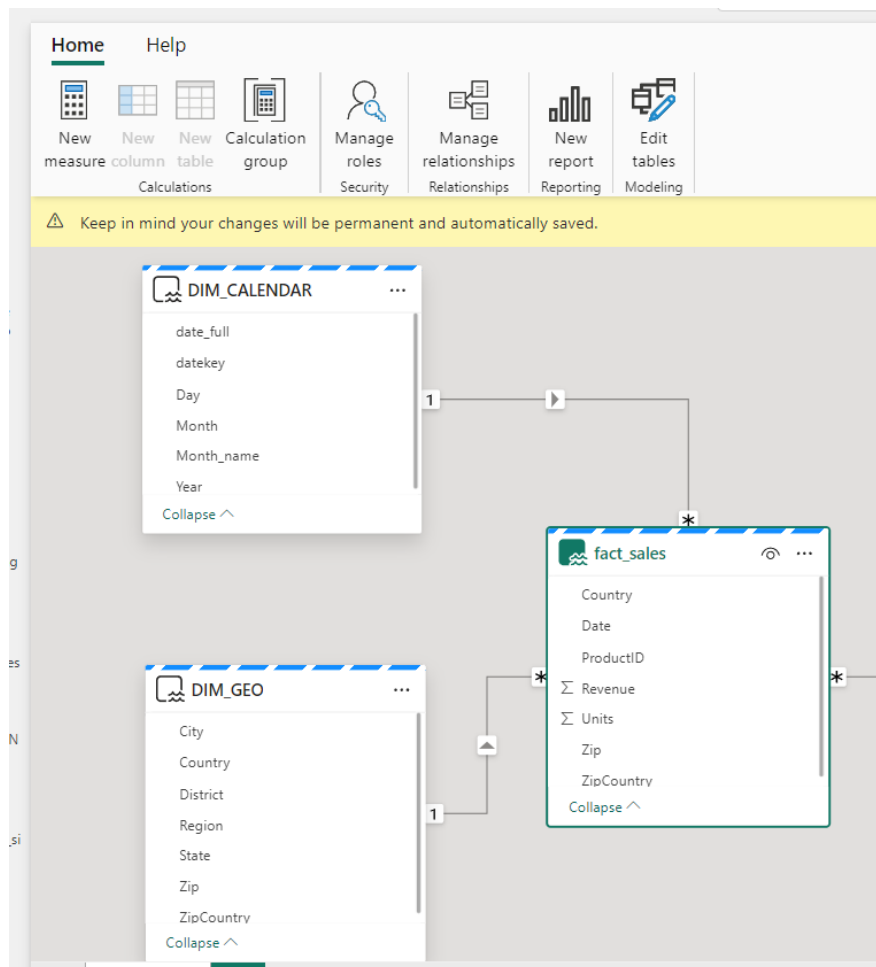
13. Type the following dax formula to the formula bar then click the **Checkmark** button to the left.

Sales Revenue = SUM(FACT_SALES[Revenue])
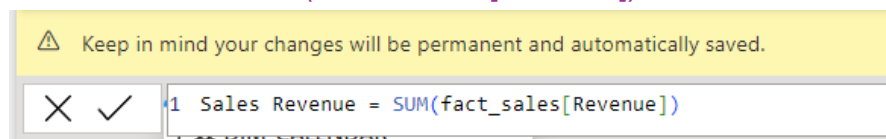
14. Select the created measure from the Data pane to the right, on the Properties pane change the data type to Decimal number, enable the Thousand separator

and set Decimal places to 2.
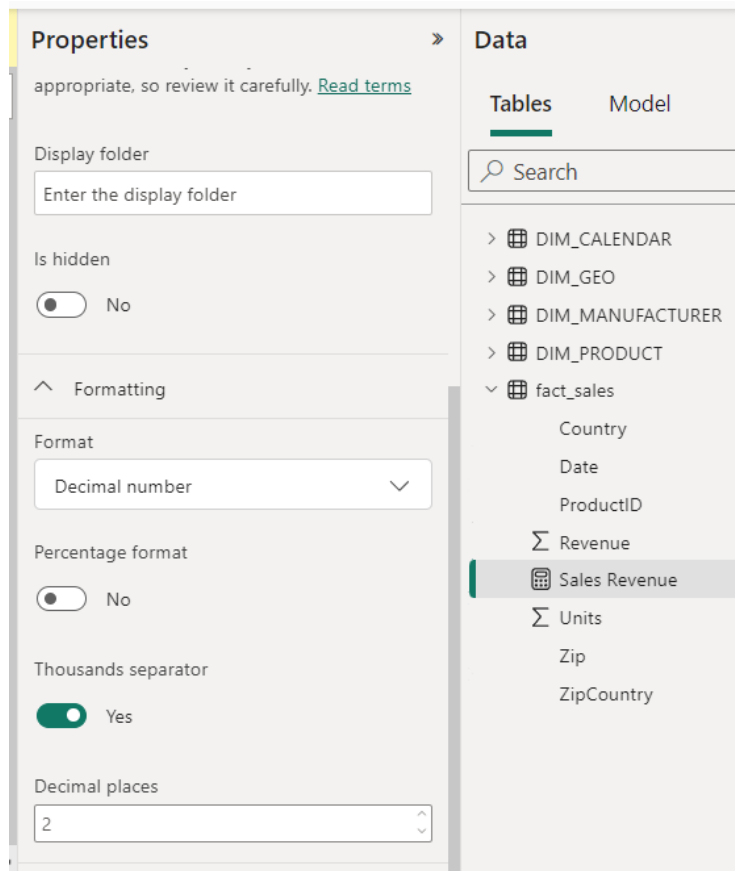
15. Create another measure on FACT_SALES with formula
Sales Qty = SUM(FACT_SALES[Units])
Change the result measure data type to Whole number (if not already).

16. Next, we will create calculation group to group similar calculations. Click
Calculation group button, change the **Calculation item** to Total in the formula
bar.
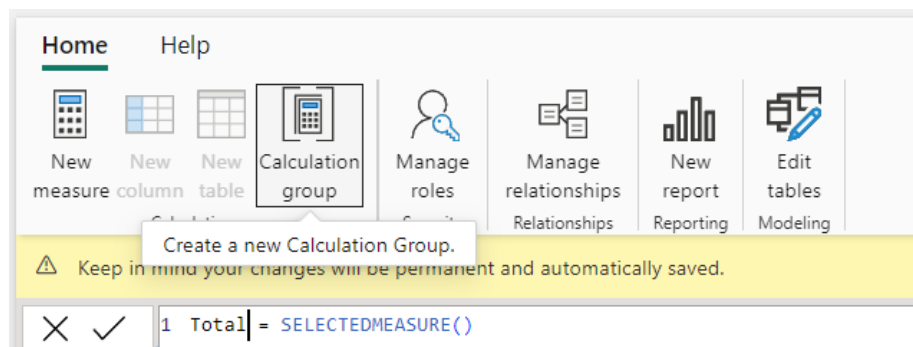
17. Double click on the Calculation group and rename it to Averages

18. Double click on the Calculation group column and rename it to Averages
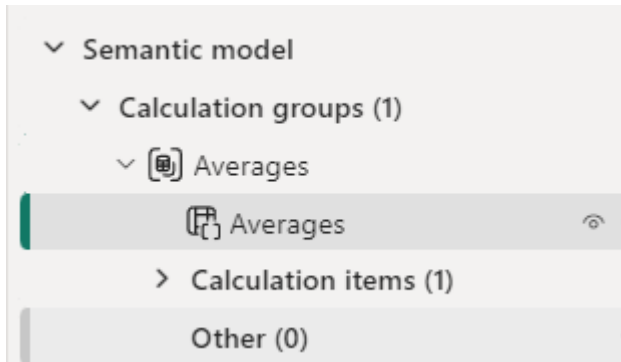
19. Next, add new calculation items by right clicking the Calculation items:
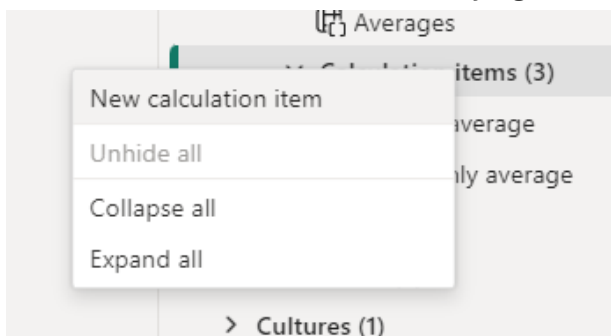
- *Daily average = AVERAGEX (VALUES (DIM_CALENDAR[date_full]), SELECTEDMEASURE())*
- *Monthly average = AVERAGEX (VALUES (DIM_CALENDAR[Month]), SELECTEDMEASURE())*

20. Next change the calculation item order with Total on top and Monthly average last.
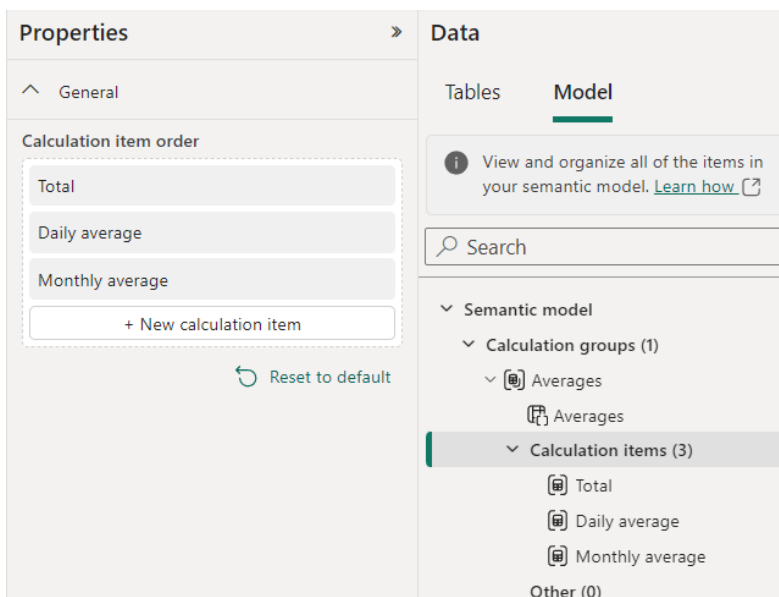
21. Create a new calculation group and name it **Time Intelligence**, also rename the Calculation group column to **Show as**. Add calculation items below and then reorder them as shown:

- *Current = SELECTEDMEASURE()*
- *MTD = CALCULATE(SELECTEDMEASURE(),DATESMTD(DIM_CALENDAR[date_full]))*
- *QTD = CALCULATE(SELECTEDMEASURE(),DATESQTD(DIM_CALENDAR[date_full]))*
- *YTD = CALCULATE(SELECTEDMEASURE(),DATESYTD(DIM_CALENDAR[date_full]))*
- *PY = CALCULATE(SELECTEDMEASURE(),SAMEPERIODLASTYEAR(DIM_CALENDAR[date_full]))*
- *YoY = SELECTEDMEASURE() - CALCULATE (SELECTEDMEASURE(), 'Time Intelligence'[Show as] = "PY")*
- *YoY% = DIVIDE (CALCULATE (SELECTEDMEASURE(),'Time Intelligence'[Show as] = "YoY"), CALCULATE (SELECTEDMEASURE(), 'Time Intelligence'[Show as] = "PY"))*
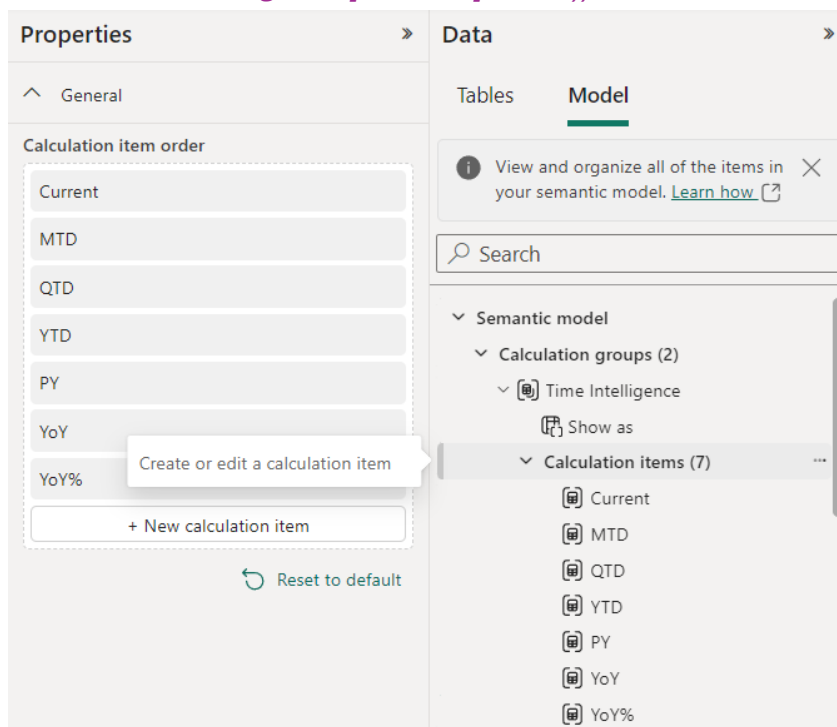
22. Select the calculation item YoY%, on the **Properties** pane enable the Dynamic format string and click **Edit in DAX formula bar**. Enter this in the formula bar

"#,##0.00%" . Then click the checkmark to the left of formula bar.
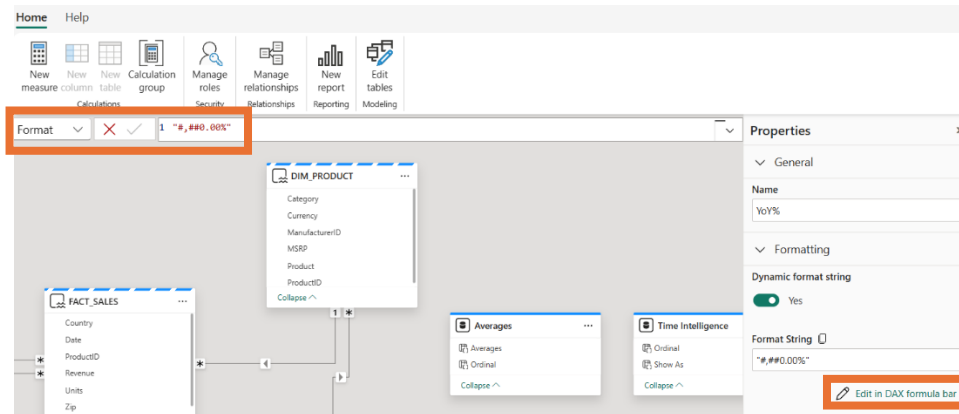
23. Access the **DIM_CALENDAR** table and select the **Month_name** column. Go to the properties pane and expand the **Advanced** options. Change the Sort by column to **Month**.
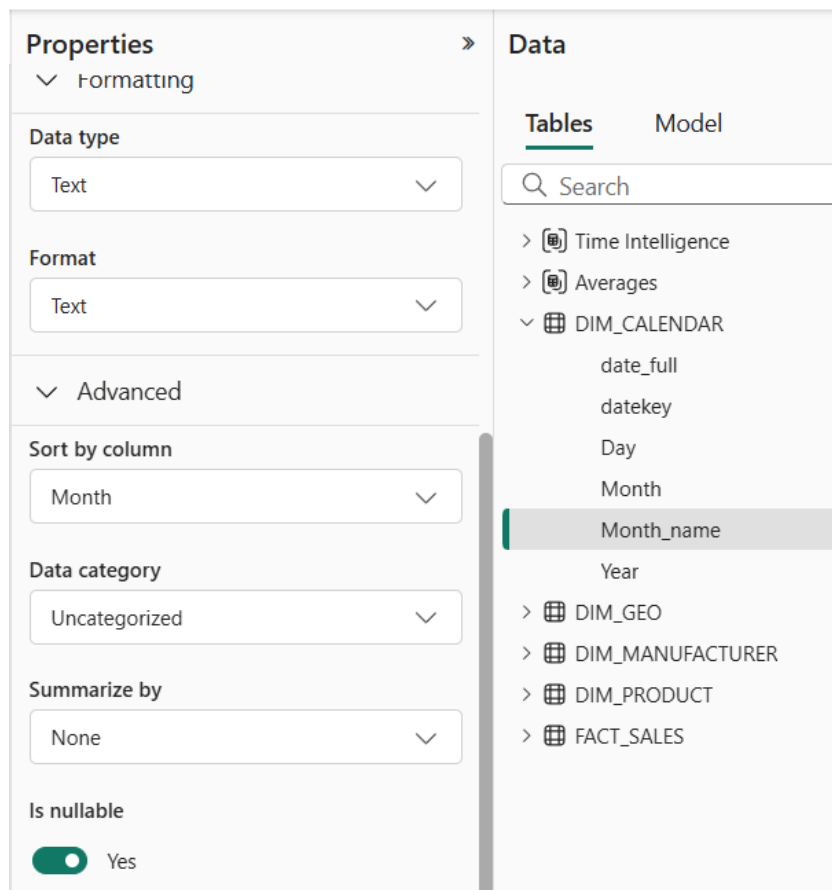
24. Then Access the **DIM_MANUFACTURER** table and select **Logo** column. On the properties pane expand the **Advanced** options and change the Data category into **Image URL**.
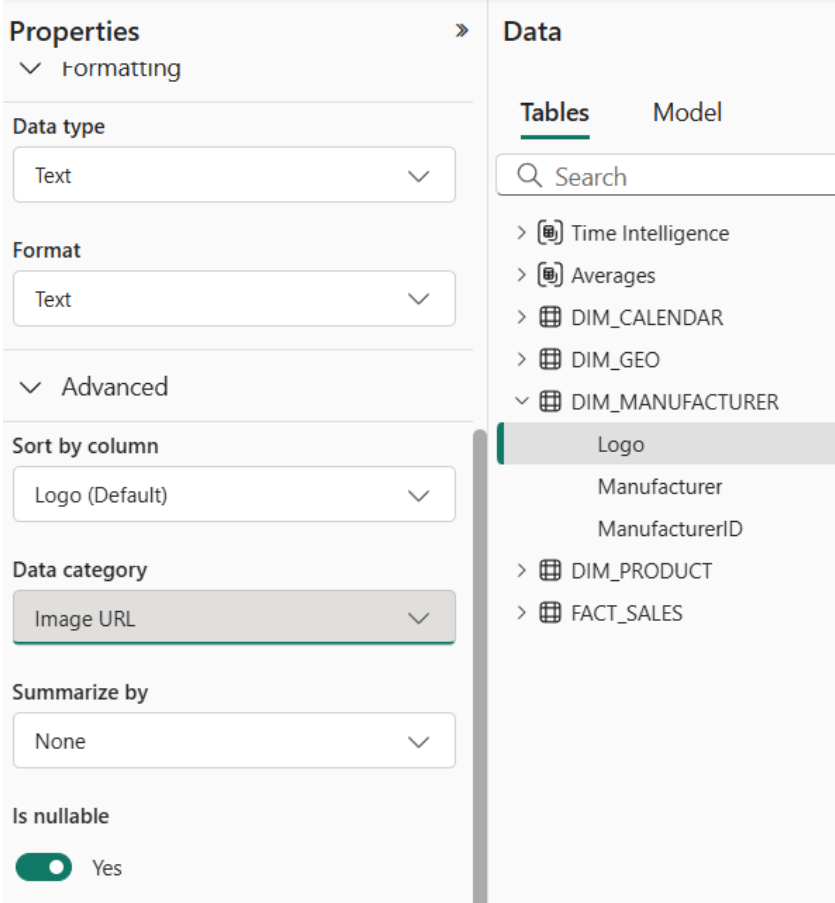
25. Next, we will create visualizations based on the semantic model. Click the New report button to open a new browser tab of power bi.
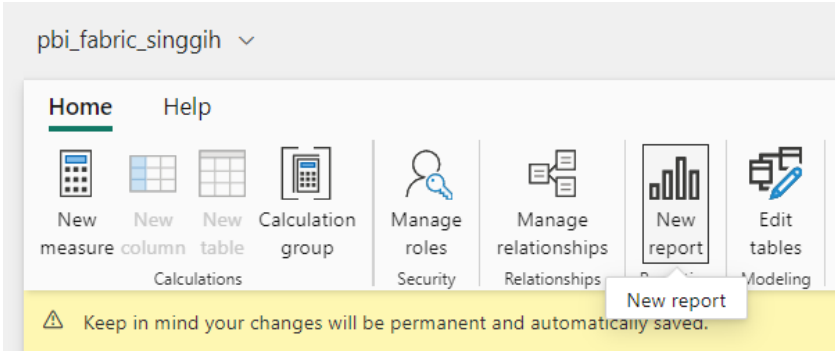
26. On the visualization pane, select matrix visual
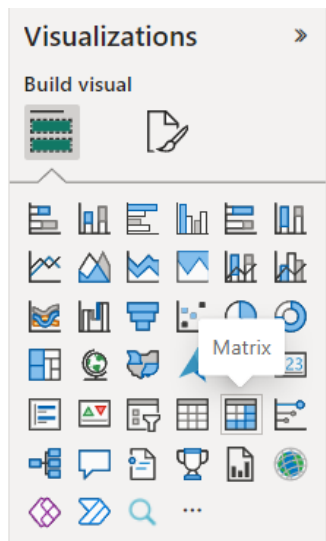
27. Move the created placeholder to the lower part of the canvas and resize it to fill half the page.

28. Drag data from the Data pane to the slot in visualizations pane following this example to make a simple matrix visual.

29. Once finished a matrix visual will be created on the canvas that looks like this

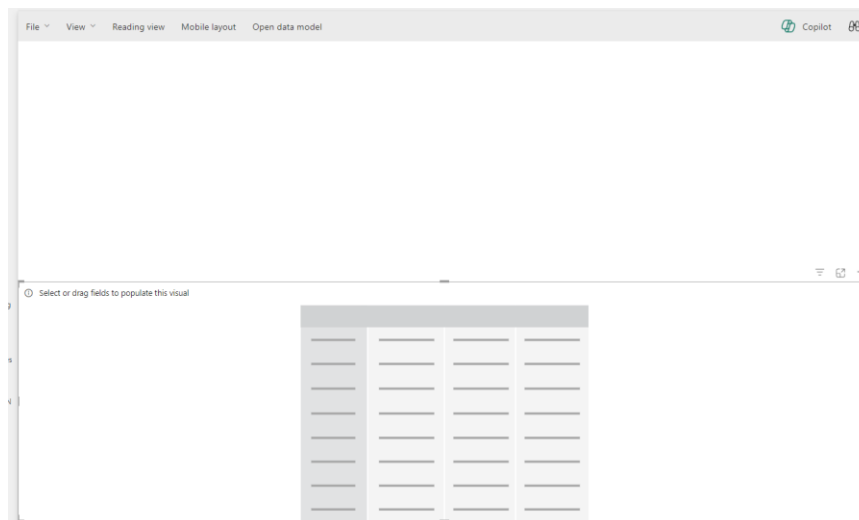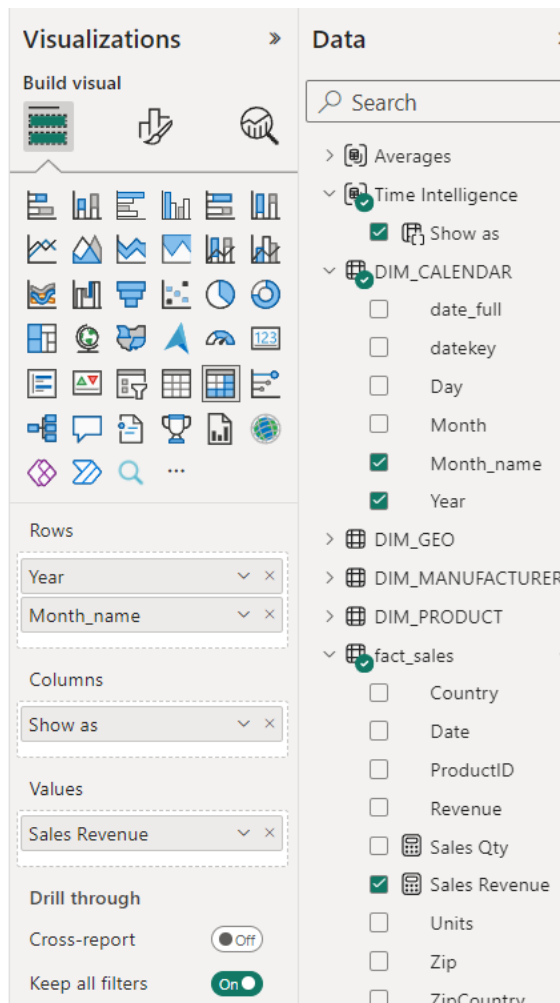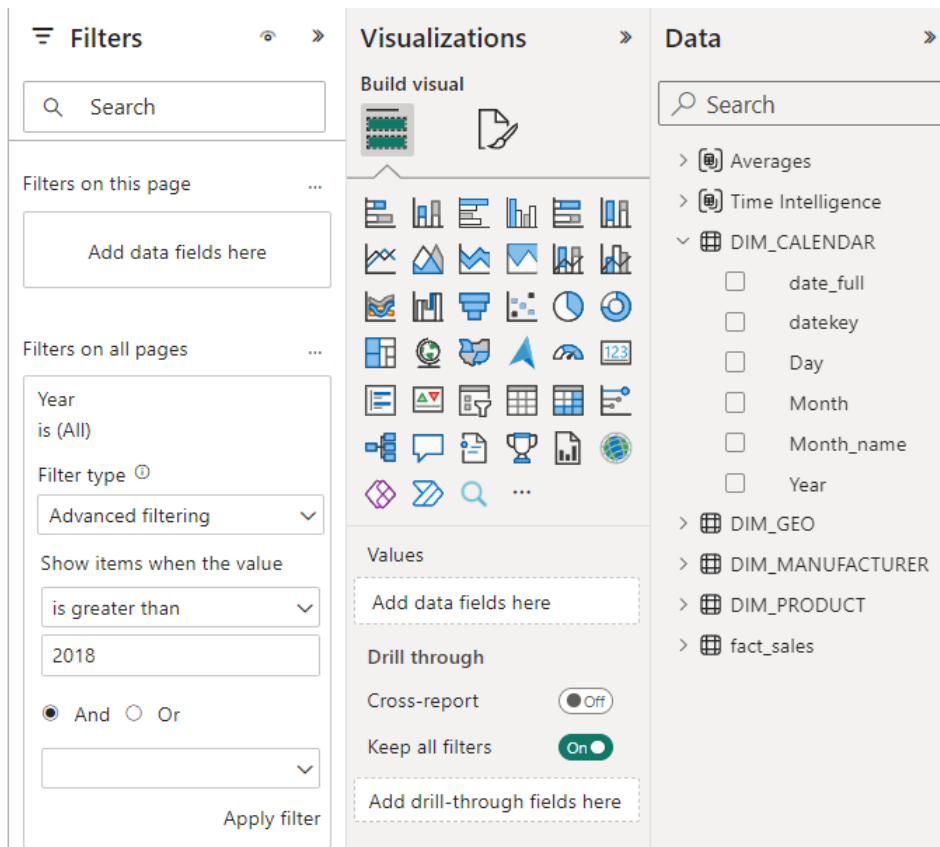| Year	Current	MTD	QTD	YTD	PY	YoY	YoY%
⊞ 2015	8,747.45	2,929.03	4,881.71	8,747.45		8,747.45	
⊞ 2016	113,566.16	3,099.34	3,099.34	113,566.16	8,747.45	104,818.72	1,198.28%
⊞ 2017	81,203.01	1,795.08	6,288.19	81,203.01	113,566.16	-32,363.15	-28.50%
⊞ 2018	258,584,218.67	22,707,401.96	68,527,825.12	258,584,218.67	81,203.01	258,503,015.66	318,341.67%
⊞ 2019	272,338,974.92	25,077,488.89	74,847,575.20	272,338,974.92	258,584,218.67	13,754,756.24	5.32%
⊞ 2020	283,044,878.29	24,957,780.33	77,509,662.93	283,044,878.29	272,338,974.92	10,705,903.37	3.93%
⊞ 2021	307,660,331.53	26,790,501.57	85,087,891.25	307,660,331.53	283,044,878.29	24,615,453.24	8.70%
⊞ 2022	182,147,803.39	15,942,041.64	48,962,712.04	182,147,803.39	307,660,331.53	-125,512,528.14	-40.80%
⊞ 2023	235,384,935.05	18,453,884.93	61,918,132.23	235,384,935.05	182,147,803.39	53,237,131.67	29.23%
⊞ 2024	349,159,497.05	31,705,073.71	99,347,884.67	349,159,497.05	235,384,935.05	113,774,561.99	48.34%
Total	1,888,524,155.51	31,705,073.71	99,347,884.67	349,159,497.05	1,539,364,658.46	349,159,497.05	22.68%

30. Notice that the value for year 2015-2017 is too small, we will now keep them from being shown on visuals.

31. On the Filter pane, drag Year column from DIM_CALENDAR and set the filter to only show data when year is greater than 2018 then click Apply filter.


32. Once filtered the matrix will change and only show data from 2019 and above.

Year	Current	MTD	QTD	YTD	PY	YoY	YoY%
2019	272,338,974.92	25,077,488.89	74,847,575.20	272,338,974.92	258,584,218.67	13,754,756.24	5.32%
2020	283,044,878.29	24,957,780.33	77,509,662.93	283,044,878.29	272,338,974.92	10,705,903.37	3.93%
2021	307,660,331.53	26,790,501.57	85,087,891.25	307,660,331.53	283,044,878.29	24,615,453.24	8.70%
2022	182,147,803.39	15,942,041.64	48,962,712.04	182,147,803.39	307,660,331.53	-125,512,528.14	-40.80%
2023	235,384,935.05	18,453,884.93	61,918,132.23	235,384,935.05	182,147,803.39	53,237,131.67	29.23%
2024	349,159,497.05	31,705,073.71	99,347,884.67	349,159,497.05	235,384,935.05	113,774,561.99	48.34%
Total	1,629,736,420.22	31,705,073.71	99,347,884.67	349,159,497.05	1,539,161,141.84	90,575,278.38	5.88%

33. Next is adding slicer to the canvas. On Visualization pane, click the slicer visual to add it to the canvas. Then drag Logo from DIM_MANUFACTURER to the field
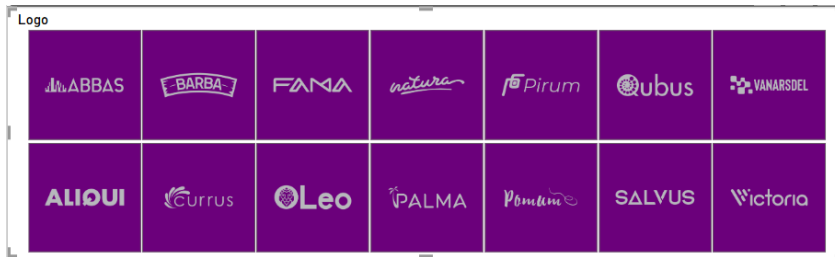
properties.
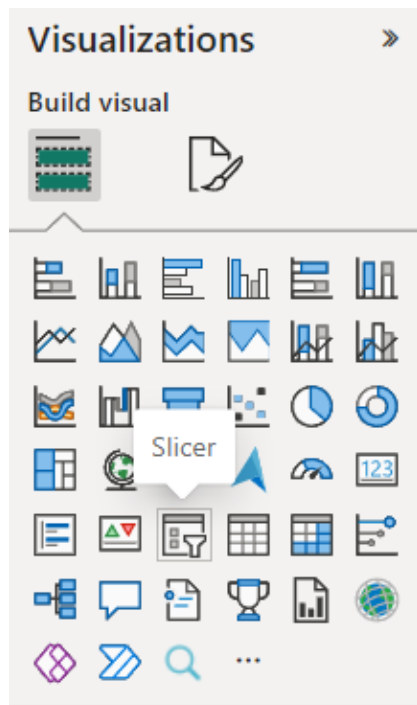
34. Next is changing the visual format, choose **Format visuals** on Visualization pane. On the Slicer settings > Options > Style, change it to Tile. Also check the Filters pane, add **Manufacturer** to the filters on this visual list and set it to display all values except (Blank).
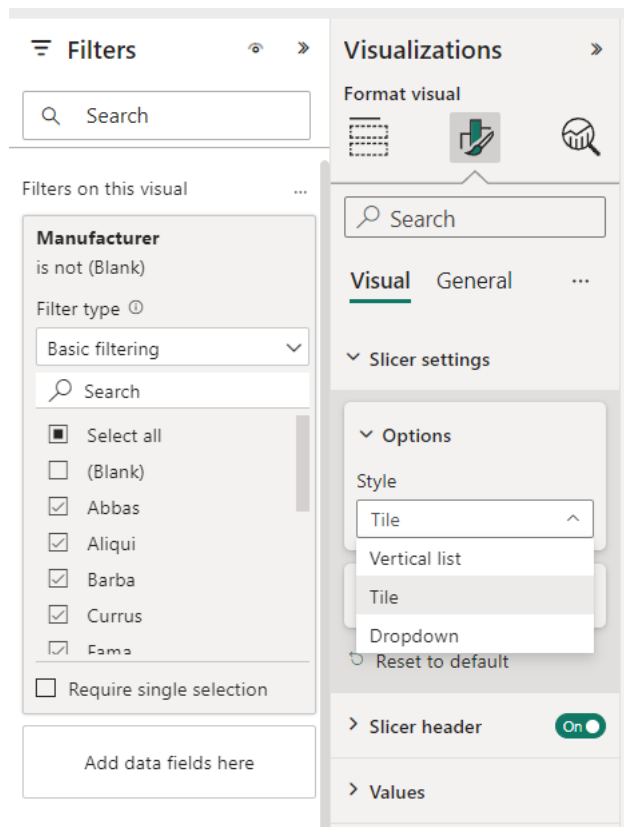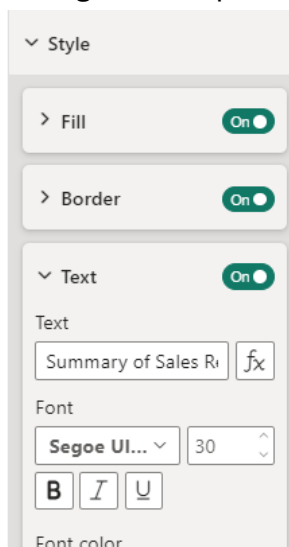
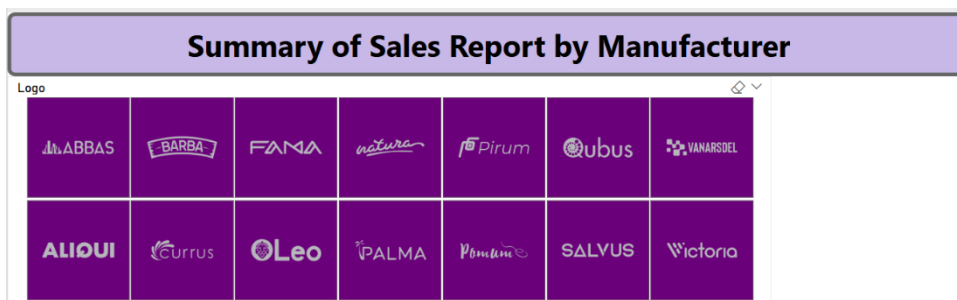35. Resize the visual on the canvas until it shows 2 rows of tiles.

36. Next, we will add a title using shapes, on the top right of toolbar. Click Shapes and choose Rectangle.
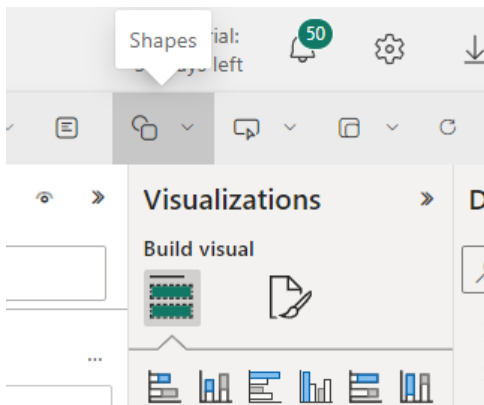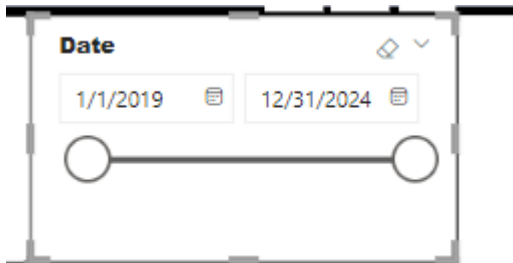

37. Change the shape format as you like it, Assign the title text in Style>Text.


38. Reposition and resize it to make it appear on the top part of the canvas.

39. Next add a date slicer using date_full column. Notice that the resulting slicer has a slider for date range selection. This happened because the data type of column is date/datetime.


40. Create a new visual of type Stacked bar chart. Use Country from DIM_GEO for Y-axis, Sales Qty for X-axis, and Category from DIM_PRODUCT as Legend.
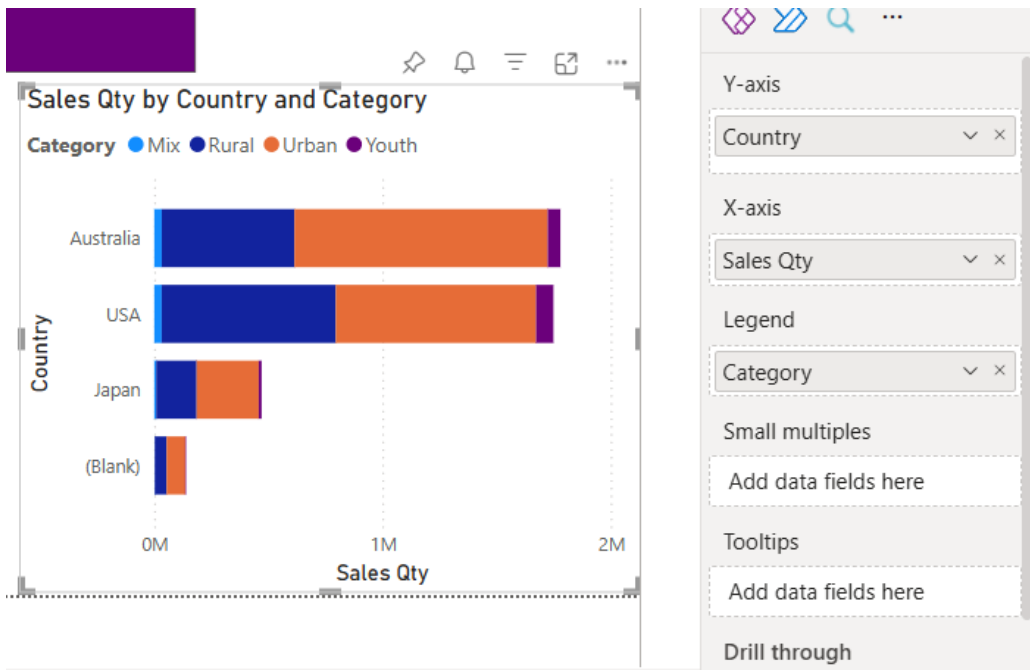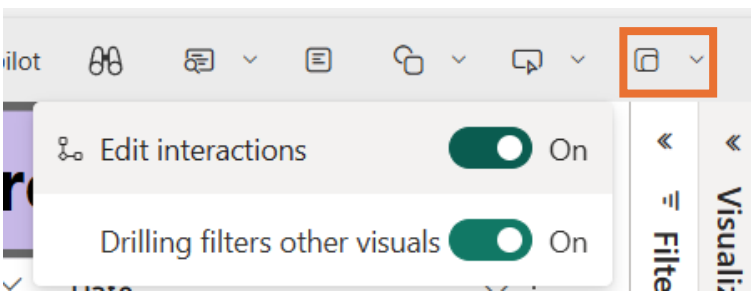
41. Notice the (Blank) country data, it seems we have data that needs to be cleansed. For now, we can add Filter on all pages where **Country** is not (Blank).
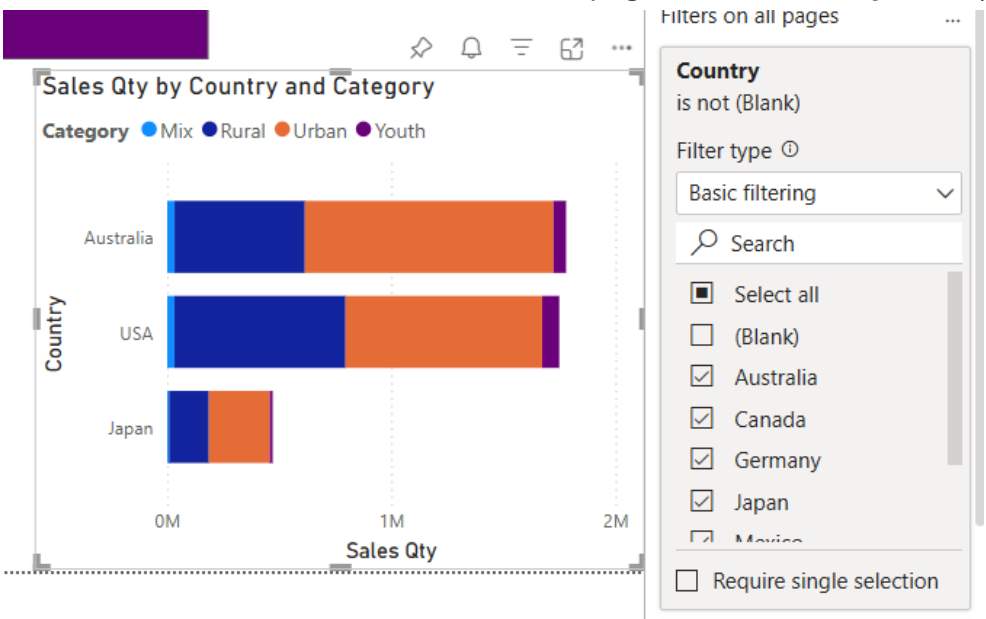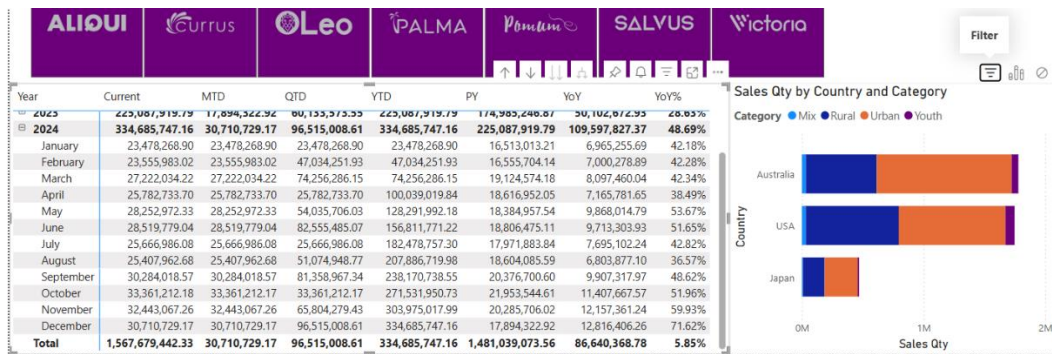

42. Try selecting a single row in the matrix, you will see the stacked bar chart got highlighted on a part of the bar.
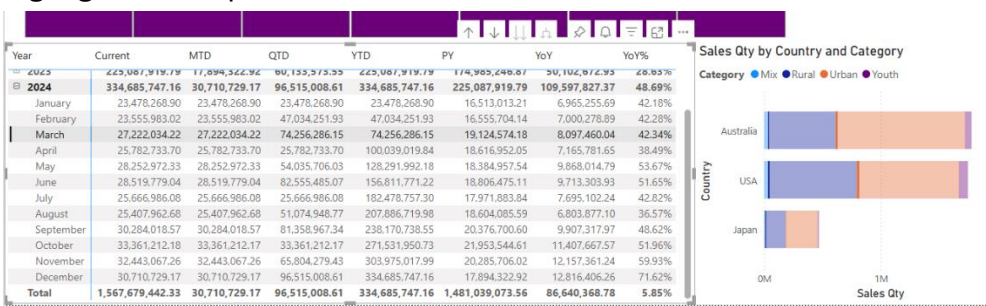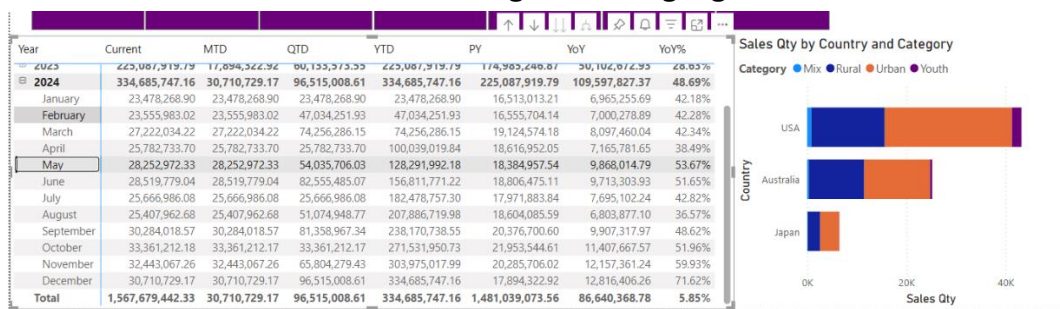

43. This is called intractions between visuals. To change it, enable Edit Interactions mode

44. Select the matrix and the icon for interations will show on top/bottom of other visuals. Change the Matrix to Filter interaction.
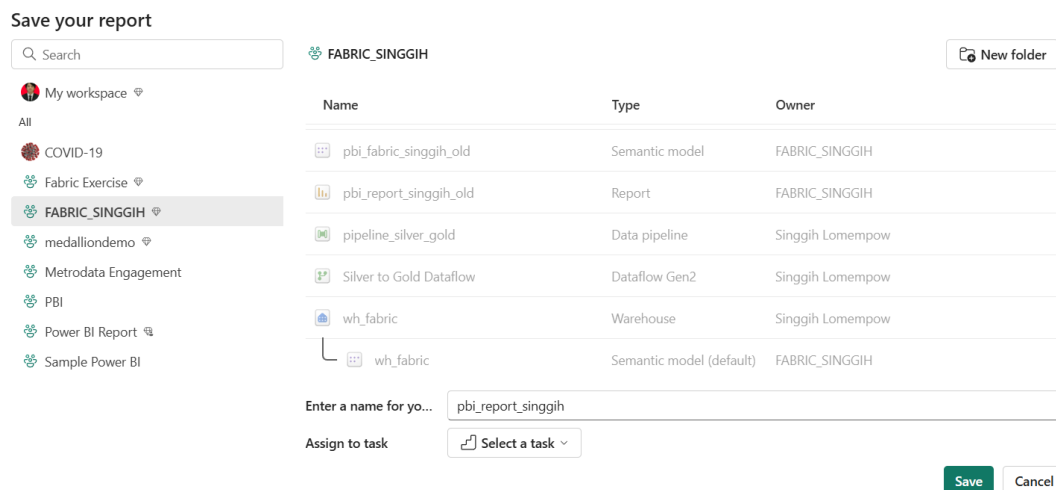

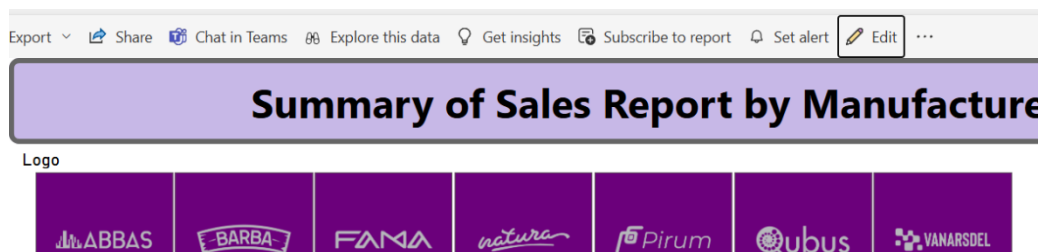45. Notice the interaction has been changed from highlight to filter.

46. Once finished disable the edit interaction and save the report by clicking File > Save/Save as. Save the report as **pbi_report_<your name>.** Once saved, click Reading view to exit editor and see the report as a viewer.
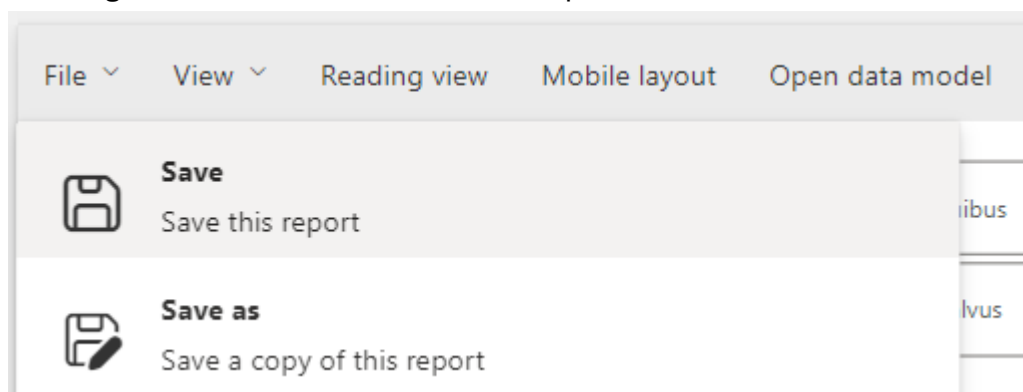
47. You can always go back to edit your report by using the Edit button.

6. Clean Up Resources

After you are finished in exploring fabric capabilities, you can delete the workspace by going into workspace settings, go to Others tab and click the Remove this workspace button.