Portfolio Project:

Querying a Data Warehouse in Microsoft Fabric

Project Overview

This project focused on querying a data warehouse in Microsoft Fabric, which provides a relational database for large-scale analytics. The goal was to utilize SQL queries to extract insights from the data, ensuring data consistency and creating views for reporting purposes.

Objectives

1. **Create a Workspace:**
   - Set up a workspace with the Fabric trial enabled.
2. **Create a Sample Data Warehouse:**
   - Establish a sample data warehouse for analysis.
3. **Query the Data Warehouse:**
   - Execute various SQL queries to analyze data.
4. **Verify Data Consistency:**
   - Check for and handle inconsistent data.
5. **Save as View:**
   - Create a view for filtered data reporting.

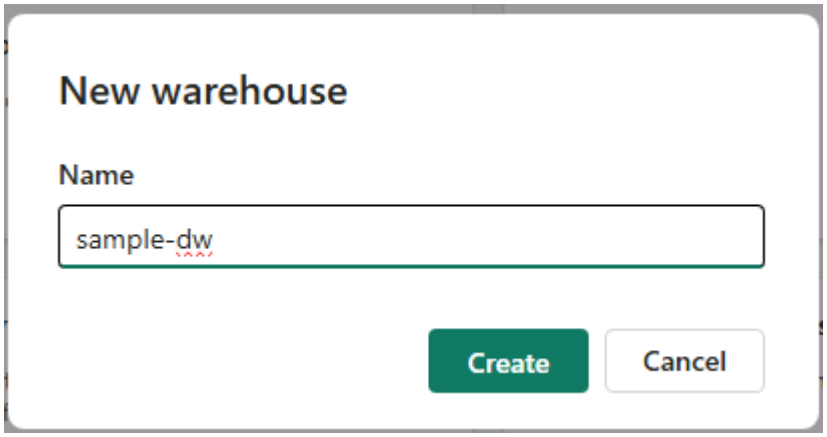
Experience

Create a Workspace

- Navigated to Microsoft Fabric Home and signed in.
- Selected Workspaces and created a new workspace.

Create a Sample Data Warehouse

- Selected Create, then Sample warehouse, and created a new data warehouse named sample-dw.



Query the Data Warehouse

- Opened a new SQL query and run the following code to get total trips and revenue by month:

```sql
SELECT
D.MonthName,
COUNT(*) AS TotalTrips,
SUM(T.TotalAmount) AS TotalRevenue
FROM dbo.Trip AS T
JOIN dbo.[Date] AS D
    ON T.[DateID]=D.[DateID]
GROUP BY D.MonthName;
```

Messages | Results | Search

	MonthName	TotalTrips	TotalRevenue
1	September	231384	3562548
2	December	233767	3540532
3	November	236513	3547993
4	June	238432	3627186
5	May	249340	3794240
6	January	237842	3401334
7	April	246153	3679261
8	July	228327	3417355

- Ran another query to get average trip duration and distance by day of the week:

```sql
SELECT
D.DayName,
AVG(T.TripDurationSeconds) AS AvgDuration,
AVG(T.TripDistanceMiles) AS AvgDistance
FROM dbo.Trip AS T
JOIN dbo.[Date] AS D
    ON T.[DateID]=D.[DateID]
GROUP BY D.DayName;
```

Messages Results Search

	DayName	AvgDuration	AvgDistance
1	Tuesday	763	2.81239639559921
2	Thursday	796	19.3689186988126
3	Friday	790	2.84746039283699
4	Sunday	1147	3.21572794977795
5	Saturday	835	33.0886460668962
6	Monday	742	6.82623660465939
7	Wednesday	780	2.8184765031265

- Queried the top 10 most popular pickup and dropoff locations:

```sql
SELECT TOP 10
    G.City,
    COUNT(*) AS TotalTrips
FROM dbo.Trip AS T
JOIN dbo.Geography AS G
    ON T.DropoffGeographyID=G.GeographyID
GROUP BY G.City
ORDER BY TotalTrips DESC;
```

Messages Results Search

	City	TotalTrips
1	Manhattan	1523186
2	New York	595638
3	Brooklyn	153569
4	Prince	92711
5	Flushing	80533
6	Queens	58181
7	Planetarium	52907
8	Brooklyn Heights	47909

Verify Data Consistency

- Checked for trips with unusually long duration:

```sql
-- Check for trips with unusually long duration
SELECT COUNT(*) FROM dbo.Trip WHERE TripDurationSeconds > 86400; -- 24 hours
```

	untitled1
1	49

- Checked for trips with negative trip duration:

```sql
-- Check for trips with negative trip duration
SELECT COUNT(*) FROM dbo.Trip WHERE TripDurationSeconds < 0;
```

	untitled1
1	4

- Removed trips with negative trip duration:

```sql
-- Remove trips with negative trip duration
DELETE FROM dbo.Trip WHERE TripDurationSeconds < 0;
```

Save as View

- Created a view for January trips:

```sql
SELECT
    D.DayName,
    AVG(T.TripDurationSeconds) AS AvgDuration,
    AVG(T.TripDistanceMiles) AS AvgDistance
FROM dbo.Trip AS T
JOIN dbo.[Date] AS D
    ON T.[DateID]=D.[DateID]
    WHERE D.Month = 1
GROUP BY D.DayName;
```

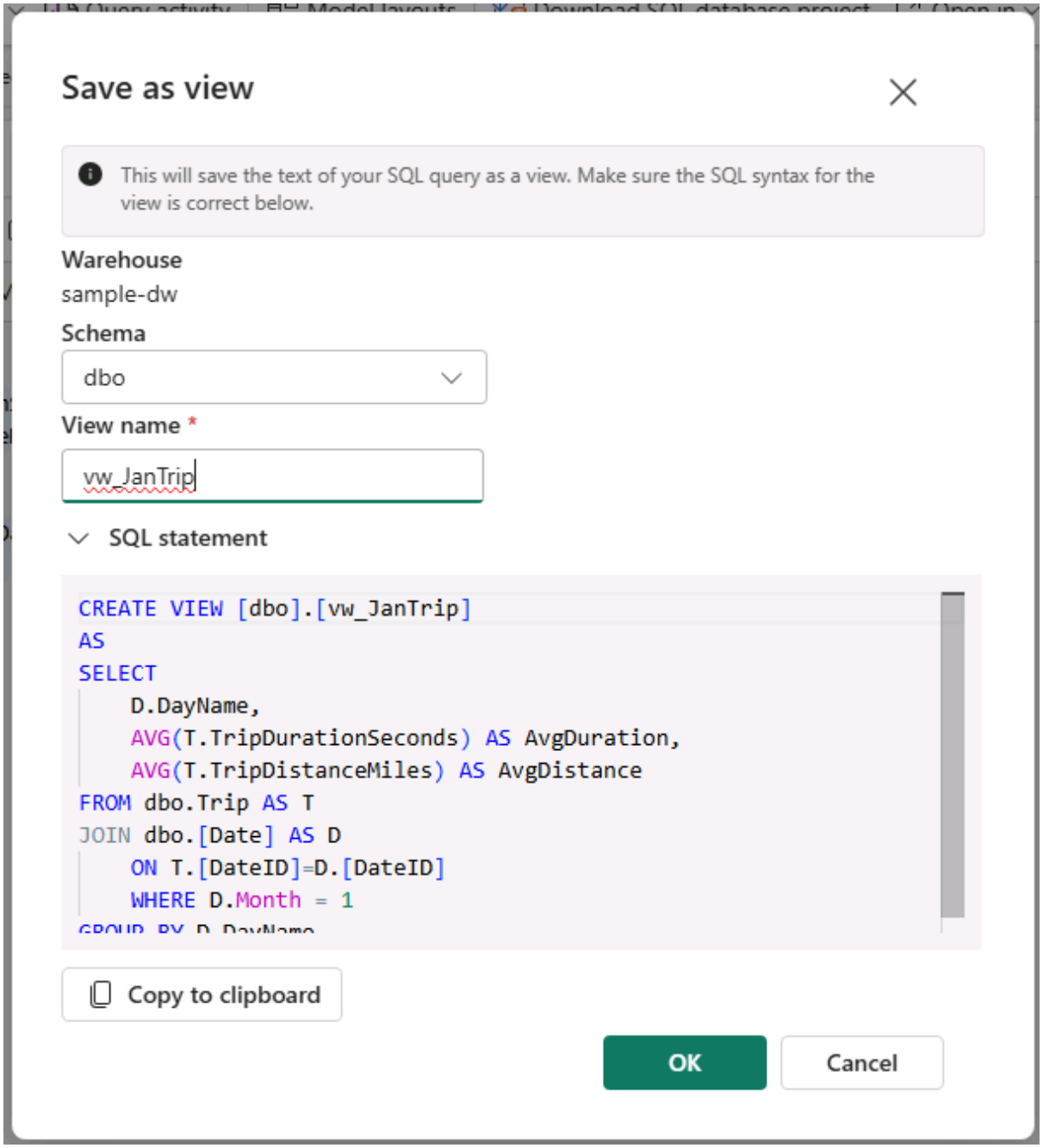
Reserved Keyword

Messages | Results

Search

	DayName	AvgDuration	AvgDistance
1	Tuesday	691	2.85065705430534
2	Thursday	715	2.79291961119472
3	Friday	713	2.62851462288349
4	Sunday	654	3.04907995357095
5	Saturday	663	2.69297393033589
6	Monday	667	2.87905740768423
7	Wednesday	716	2.86763962644205

- Saved the view as vw_JanTrip.



Results

- ✓ A workspace was successfully created in Microsoft Fabric.
- ✓ A sample data warehouse named sample-dw was established and populated with sample data for analysis.
- ✓ SQL queries were executed to analyze data, revealing:
  - Total trips and revenue by month.
  - Average trip duration and distance by day of the week.
  - The top 10 most popular pickup and dropoff locations.

- ✓ Data consistency checks were performed, identifying and handling trips with negative durations.
- ✓ A view named vw_JanTrip was created to filter and report on January trip data.

Conclusion

This project provided a practical introduction to querying a data warehouse in Microsoft Fabric. Key insights were gained into the use of SQL for data analysis and the importance of verifying data consistency. The ability to create views enhanced the usability of the data warehouse for reporting purposes. Overall, this exercise demonstrated the capabilities of Microsoft Fabric in managing and analyzing large-scale data efficiently.

Resources

GitHub: https://github.com/ThatoMTNG/Microsoft-Fabric-Analytics-Engineer-DP-600-

Mentions

Project Author: Thato Metsing (https://www.linkedin.com/in/thatometsing/)

Project Mentor: Maureen Direro (https://www.linkedin.com/in/maureen-direro-46a6b220/)