CASE STUDY

ROOM-09

PRESENTED BY

- SUMA VUNDAVALLI
- ADARSH
- ISHITA SAHA

Problem Statement

TechRetail, a mid-sized retail company, wants to create a data pipeline to collect retail data from various sources, process it using advanced analytics, and visualize the results in a dashboard. The goal is to gain insights into sales trends and improve decision-making. The company wants to leverage Azure Databricks for data processing and Microsoft Fabric for data integration and visualization.

SOLUTION PROPOSED

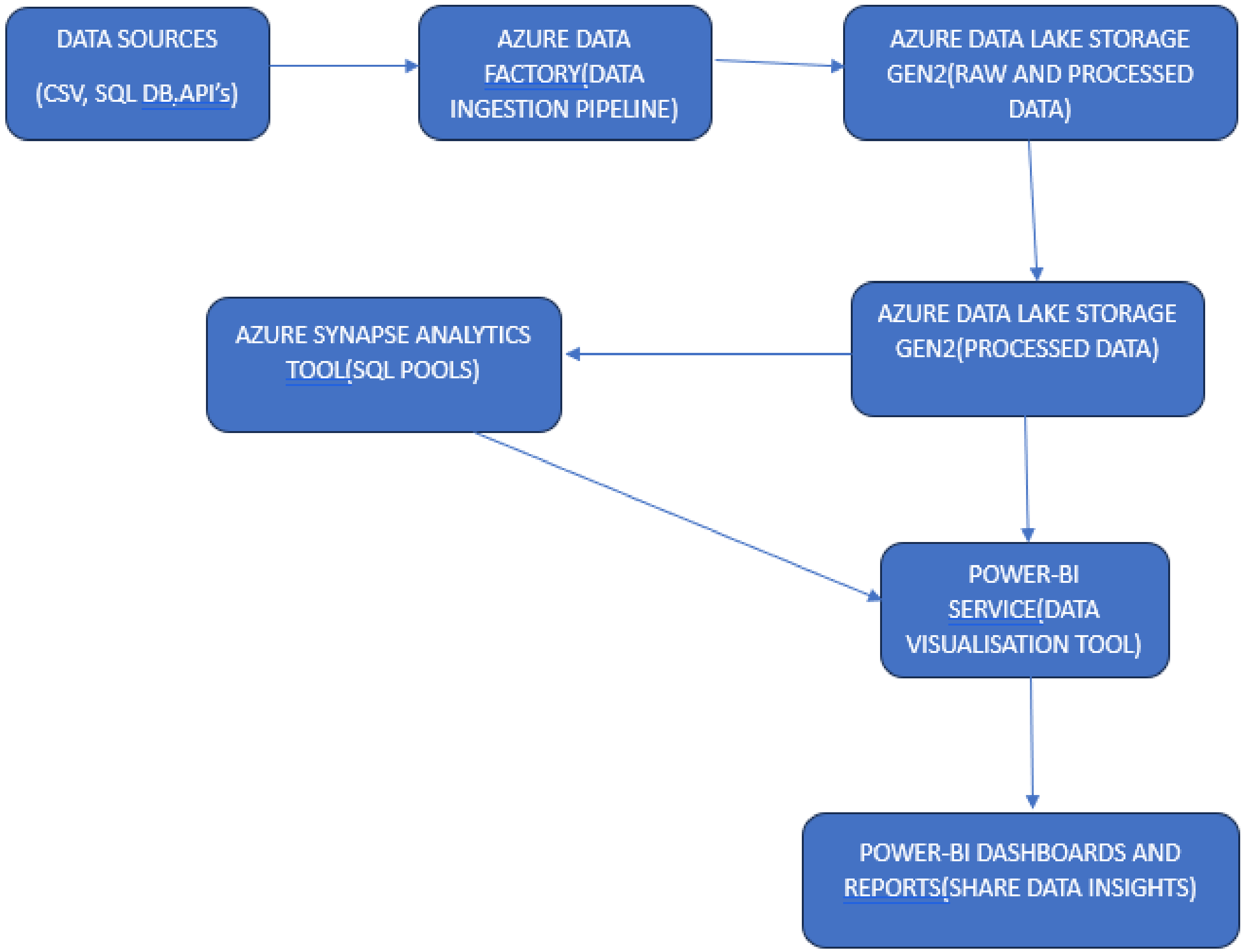
DATA INGESTION – AZURE DATA FACTORY(ADF)

DATA PROCESSING – AZURE DATABRICKS AZURE SYNAPSE ANALYTICS(SQL POOLS)

DATA STORAGE – AZURE DATA LAKE STORAGE GEN-2

DATA VISUALISATION – POWER-BI

Architecture-Diagram



DATA SOURCES

(CSV, SQL DB.API’s)

AZURE DATA FACTORY(DATA INGESTION PIPELINE)

AZURE DATA LAKE STORAGE GEN2(RAW AND PROCESSED DATA)

AZURE DATA LAKE STORAGE GEN2(PROCESSED DATA)

AZURE SYNAPSE ANALYTICS TOOL(SQL POOLS)

POWER-BI SERVICE(DATA VISUALISATION TOOL)

POWER-BI DASHBOARDS AND REPORTS(SHARE DATA INSIGHTS)

Data Preparation Tasks – Data Cleaning

- Missing Data Handling

Problem: Some fields may have missing values (e.g., customer information or sales amounts).

Solution: Fill missing values with mean, median, or default values where applicable (e.g., missing Customer_Segment can be filled with "Regular").

Data Preparation Tasks – Data Cleaning

- Duplicate Records

Problem: Duplicate rows could exist, such as the same customer making multiple purchases within the same transaction.

Solution: Identify and remove duplicates by comparing the combination of fields like Customer_ID, Order_ID, and Date. Use Power Query or SQL queries in Synapse for this.

Data Preparation Tasks – Transformation Tasks

Aggregating Sales Data

Problem: The dataset includes individual transactions, but we need to aggregate sales data at a customer level or product category level.

Solution: Create new measures that calculate total sales, average order value, etc.

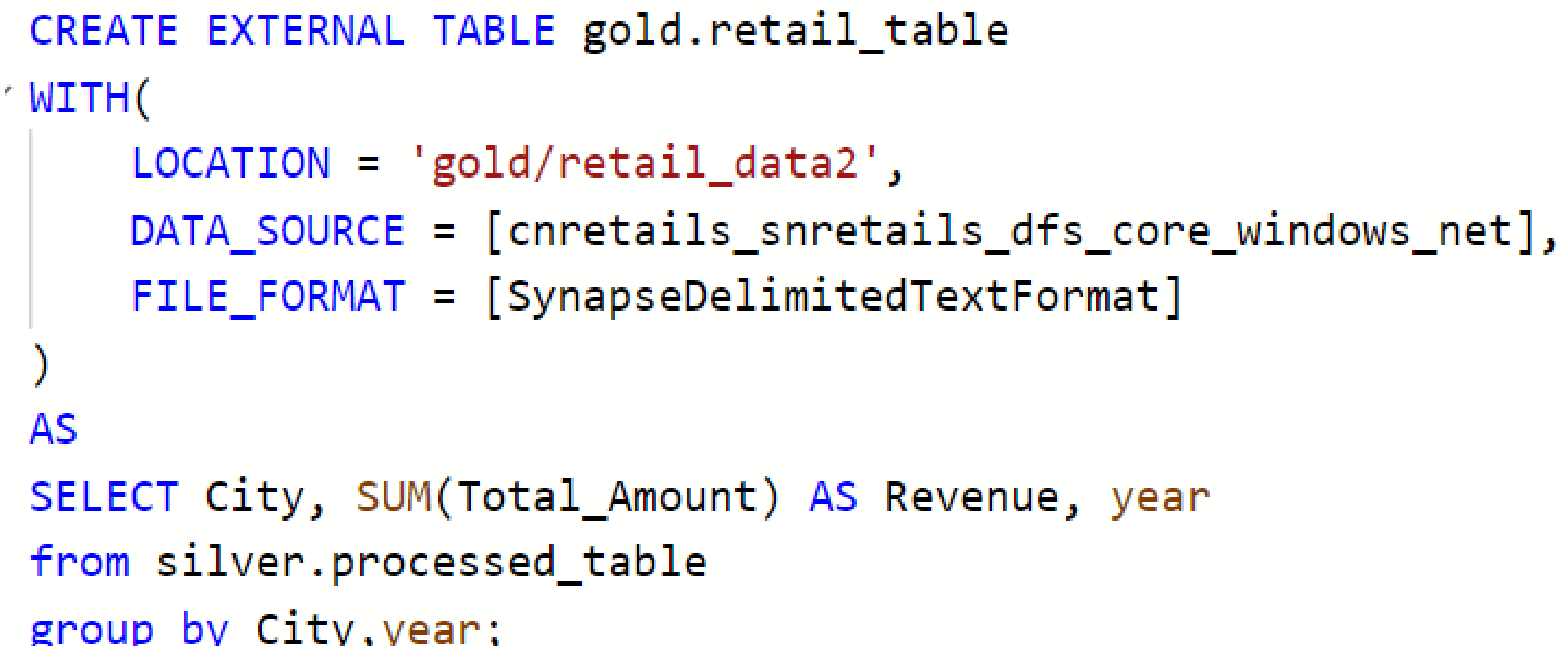
Data Preparation Tasks – Transformation Tasksata Preparation Tasks – Transformation Tasks

Create Time-based Features (e.g., Year, Month, Day of Week)

Problem: The dataset has the Date, but time-based aggregation will be more efficient for analysis (e.g., trends by year or month)

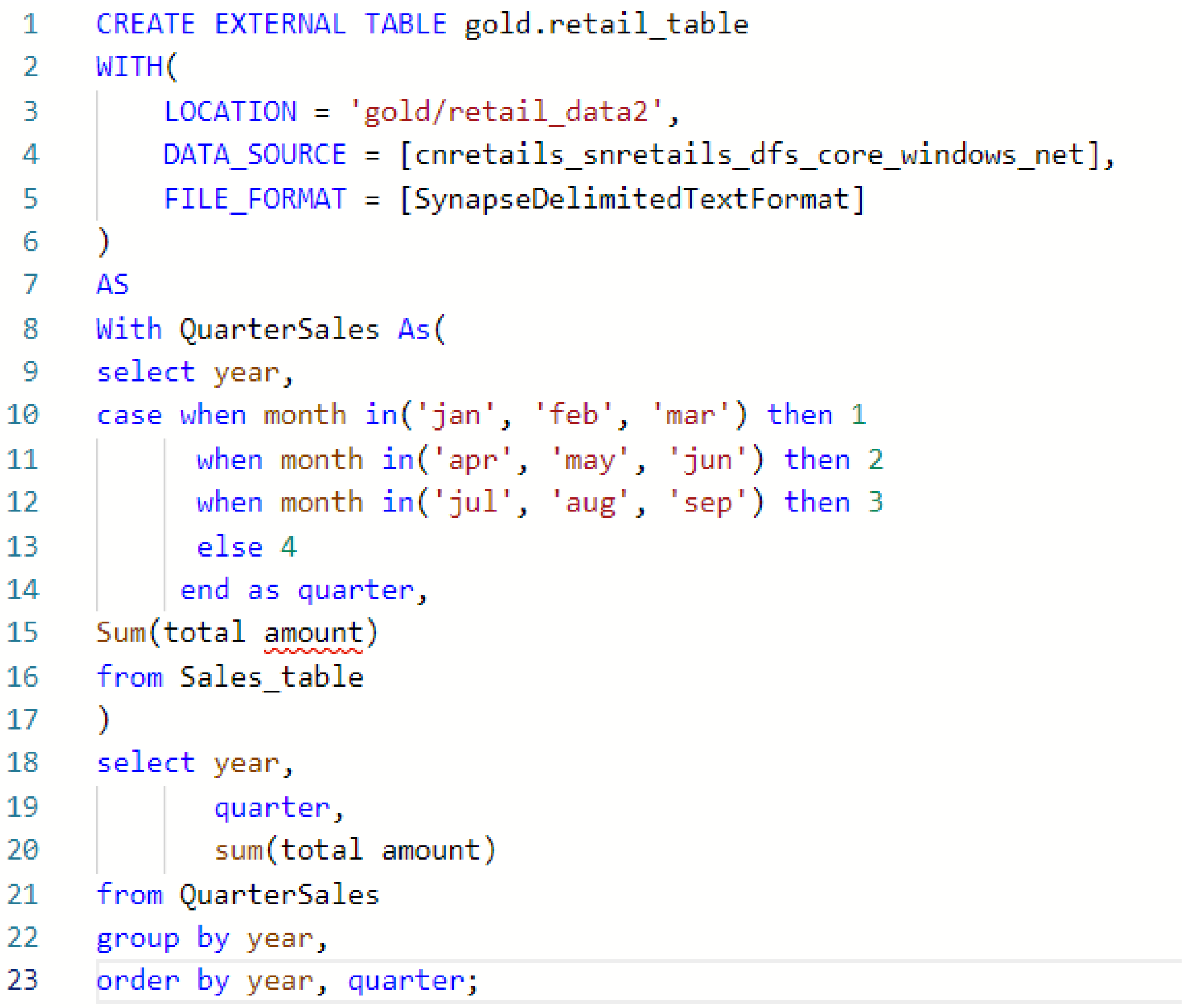
Solution: Extract year, month, and other time-based features like Day of Week for grouping and aggregation.

Total Sales per City



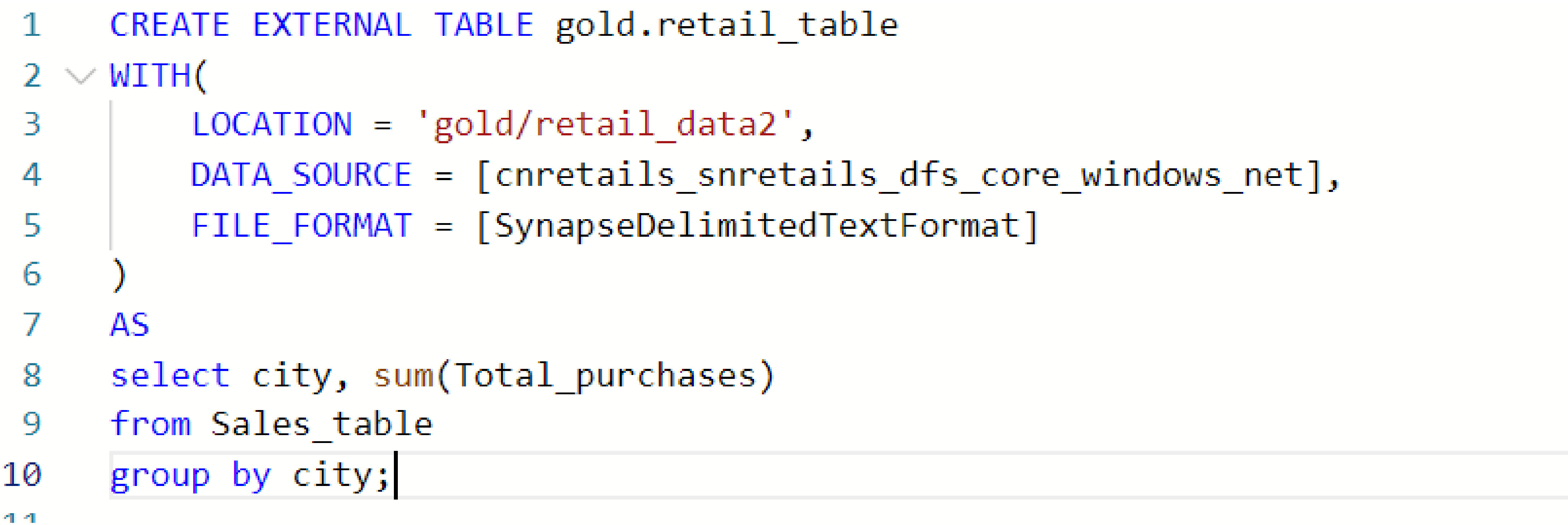
```
CREATE EXTERNAL TABLE gold.retail_table
WITH(
    LOCATION = 'gold/retail_data2',
    DATA_SOURCE = [cnretails_snretails_dfs_core_windows_net],
    FILE_FORMAT = [SynapseDelimitedTextFormat]
)
AS
SELECT City, SUM(Total_Amount) AS Revenue, year
from silver.processed_table
group by City,year;
```

Total Sales per quarter per year



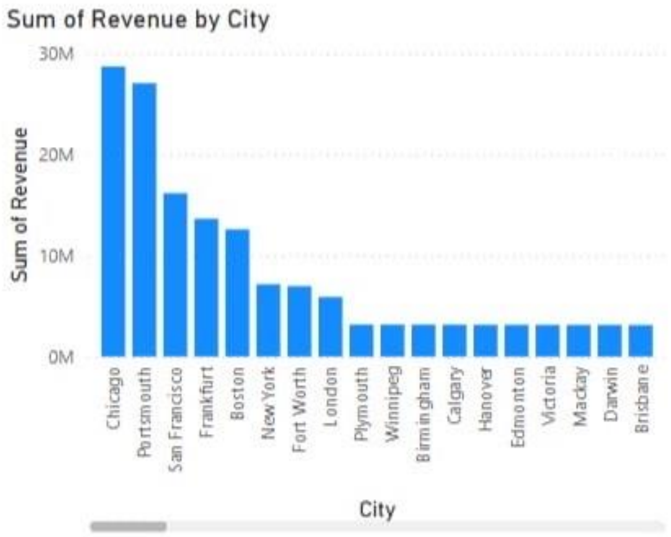
```sql
CREATE EXTERNAL TABLE gold.retail_table
WITH(
    LOCATION = 'gold/retail_data2',
    DATA_SOURCE = [cnretails_snretails_dfs_core_windows_net],
    FILE_FORMAT = [SynapseDelimitedTextFormat]
)
AS
With QuarterSales As(
select year,
case when month in('jan', 'feb', 'mar') then 1
      when month in('apr', 'may', 'jun') then 2
      when month in('jul', 'aug', 'sep') then 3
      else 4
     end as quarter,
Sum(total amount)
from Sales_table
)
select year,
       quarter,
       sum(total amount)
from QuarterSales
group by year,
order by year, quarter;
```

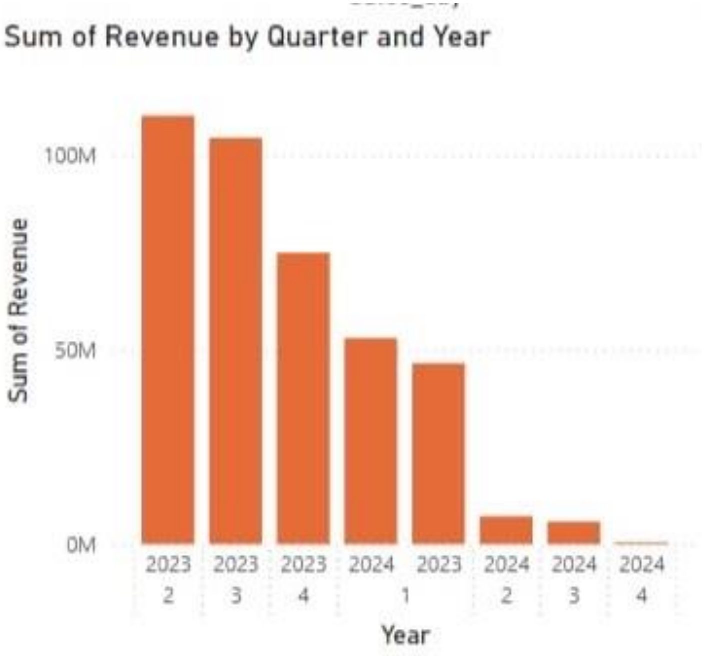
Total products sold per city

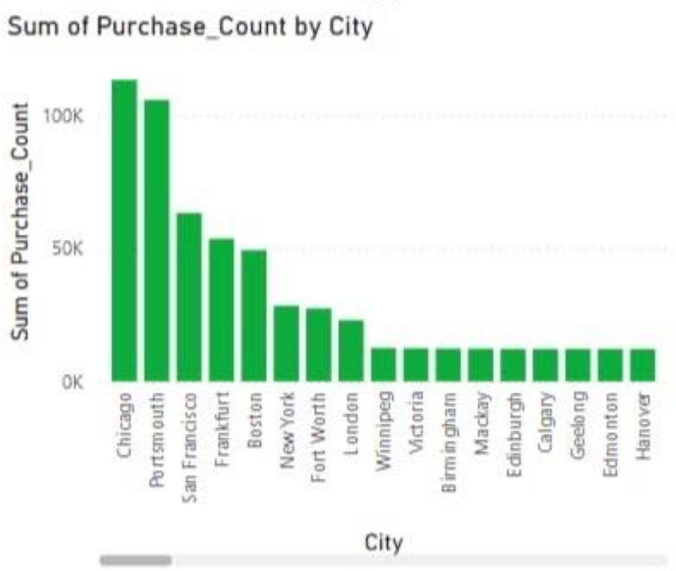


```sql
CREATE EXTERNAL TABLE gold.retail_table
WITH(
    LOCATION = 'gold/retail_data2',
    DATA_SOURCE = [cnretails_snretails_dfs_core_windows_net],
    FILE_FORMAT = [SynapseDelimitedTextFormat]
)
AS
select city, sum(Total_purchases)
from Sales_table
group by city;
```

Data Visualisation


Sum of Revenue by City
Sum of Revenue
30M
20M
10M
0M
Chicago
Portsmouth
San Francisco
Frankfurt
Boston
New York
Fort Worth
London
Plymouth
Winnipeg
Birmingham
Calgary
Hanover
Edmonton
Victoria
Mackay
Darwin
Brisbane
City


Sum of Revenue by Quarter and Year
Sum of Revenue
100M
50M
0M
Year


Sum of Purchase_Count by City
Sum of Purchase_Count
100K
50K
0K
Chicago
Portsmouth
San Francisco
Frankfurt
Boston
New York
Fort Worth
London
Winnipeg
Victoria
Birmingham
Mackay
Edinburgh
Calgary
Geelong
Edmonton
Hanover
City