Creating the Data Warehouse for GreenLink Supply Chain Footprint Analyzer

Objective

To establish a star-schema data warehouse for the GreenLink Supply Chain Footprint Analyzer project. This warehouse will act as the centralized data repository, integrating all supply chain-related data for downstream analysis and machine learning.

Step-by-Step Implementation

1. **Workspace Creation**:
    - A new workspace titled **"GreenLink Supply Chain Footprint Analyzer"** was created in **Microsoft Fabric** to serve as the project's central environment.
2. **Lakehouse Setup**:
    - A **Lakehouse** named `SCM_LH` (Supply Chain Management Lakehouse) was created to act as the staging area for data ingestion.
    - The following CSV tables were imported into the Lakehouse:
        - **Production Table**: Contains information about materials, energy sources, and CO_2 emissions per unit of production.
        - **Transportation Table**: Details about transport modes, fuel types, and average CO_2 emissions per kilometer.
        - **Supplier Table**: Includes supplier information, locations, and average CO_2 emissions per delivery.
        - **Company Table (Fact Table)**: The central fact table containing metrics like distance traveled, units produced, and delivery counts, along with foreign keys linking to dimension tables.
3. **Data Warehouse Creation**:
    - A **Data Warehouse** named `SCM` was created to structure and manage the data.
    - Using **Dataflow Gen2**, the four tables were transferred from the `SCM_LH` Lakehouse to the `SCM` Data Warehouse for relational data management.
4. **SQL Table Creation**:
    - The imported tables were converted into SQL tables within the `SCM` Data Warehouse.
    - These tables maintain a structured format, facilitating queries and transformations required for analytics.
5. **Star Schema Design**:
    - The tables were linked to create a **star schema**:
        - The **Company Table** acts as the central **fact table**.
        - The **Production Table**, **Transportation Table**, and **Supplier Table** serve as **dimension tables**.
    - Relationships between the tables were established via foreign keys:
        - `ProductionID` in the **Fact Table** links to the **Production Table**.
        - `TransportID` in the **Fact Table** links to the **Transportation Table**.
        - `SupplierID` in the **Fact Table** links to the **Supplier Table**.

- The resulting schema enables efficient querying and analysis for business intelligence and machine learning workflows.

Outcome

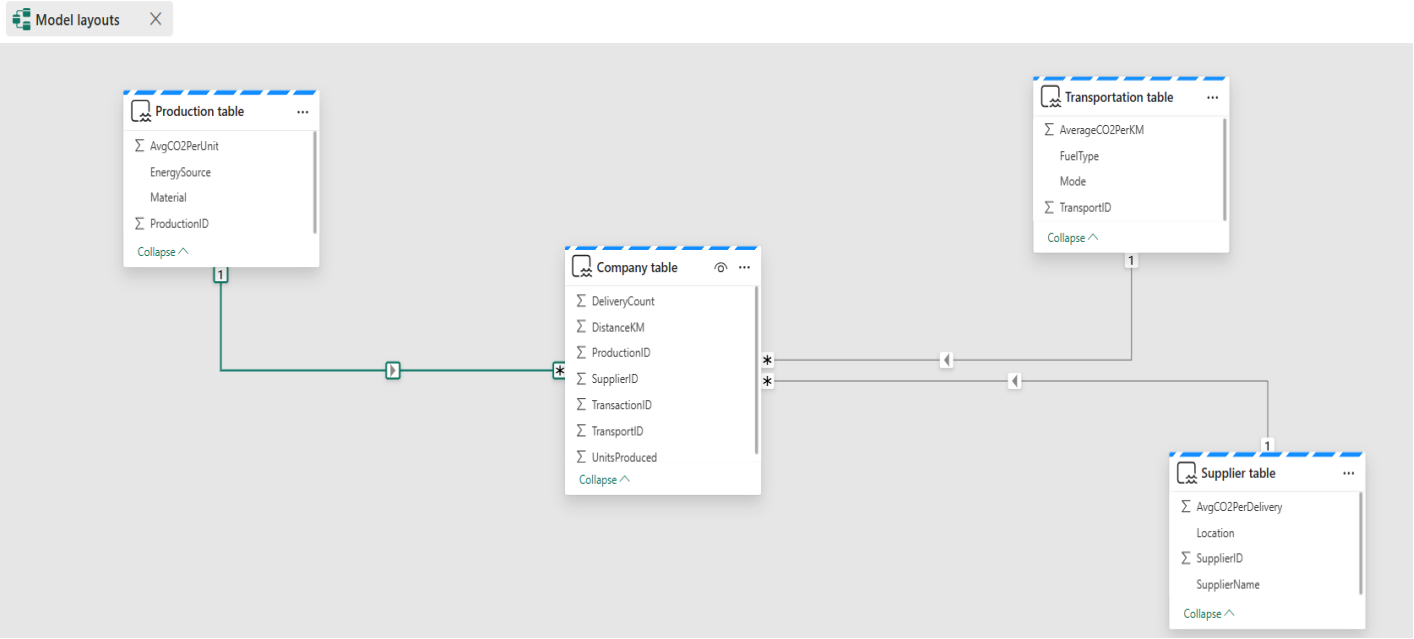

Figure 1 Star schema Datawarehouse

Merging Tables Using Data Factory

To facilitate advanced analytics and machine learning, the dimension and fact tables were merged into a single table, consolidating all key attributes required for the GreenLink Supply Chain Footprint Analyzer. This phase utilized **Dataflow Gen2** in Microsoft Fabric's Data Factory for seamless table merging.

Implementation Steps

1. **Setup in Data Factory**:
   - Navigated to **Dataflow Gen2** in **Microsoft Fabric Data Factory**.
   - Established a new data flow specifically for merging the dimension and fact tables.
2. **Table Import**:
   - Imported the four tables—**Production Dimension Table**, **Supplier Dimension Table**, **Transportation Dimension Table**, and **Fact Table**—from the **Lakehouse**.
   - Ensured that the column headers and data types were correctly interpreted during the import process.
3. **Merging Process**:

- Sequentially merged the tables to create a unified dataset:
  - Merged the **Fact Table** with the **Production Dimension Table** on the `ProductionID` key.
  - Joined the resulting table with the **Supplier Dimension Table** on the `SupplierID` key.
  - Finally, merged with the **Transportation Dimension Table** on the `TransportID` key.
- The resulting table was named **Company Merged Table**, encapsulating metrics such as:
  - Production details (e.g., material, energy source, CO_2 per unit).
  - Supplier information (e.g., location, CO_2 per delivery).
  - Transportation details (e.g., fuel type, CO_2 per kilometer).
  - Fact metrics like distance traveled and units produced.

4. **Data Validation and Transformation**:
   - Applied transformation steps such as renaming columns for consistency, standardizing data types, and handling missing or inconsistent data.
   - Verified the merged table for data integrity and alignment with the project requirements.
5. **Output Destination**:
   - Saved the **Company Merged Table** back into the **Lakehouse**, enabling downstream processes such as machine learning and visualization.

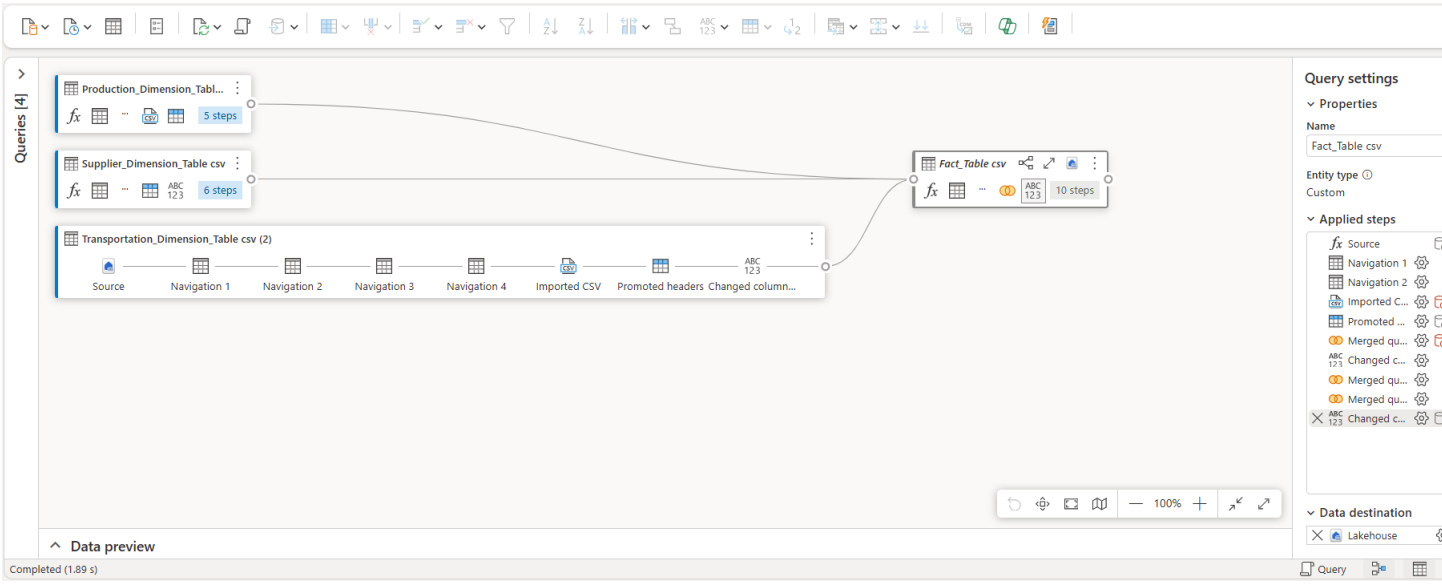

Figure 2 Dataflow model

Outcome

The **Company Merged Table** serves as a comprehensive dataset that integrates all supply chain dimensions. This unified structure simplifies querying, facilitates efficient feature engineering, and provides the foundation for machine learning models and dashboard development.

Analysis and Machine Learning in Fabric Data Science Notebook

Set Up the Data Science Notebook

The analysis was performed in the **Fabric Data Science Notebook**. The `MergedFactTable` from the data warehouse was connected and used as the primary dataset for this phase. This setup allowed seamless integration with Spark-based tools and machine learning libraries for preprocessing, analysis, and model training.

Data Preprocessing

1. **Exploratory Data Analysis (EDA)**:
    - **Distribution Analysis**: The dataset's key numerical features, including `DistanceKM`, `UnitsProduced`, and `DeliveryCount`, were analyzed for distributions. The statistics revealed:
        - Average distance traveled: 986.35 km.
        - Average units produced: 2632.3 units.
        - Average delivery count: 4.75.

```
+-------+------------------+------------------+------------------+
|summary|        DistanceKM|     UnitsProduced|     DeliveryCount|
+-------+------------------+------------------+------------------+
|  count|                20|                20|                20|
|   mean|            986.35|            2632.3|              4.75|
| stddev|681.3843706208088|1553.8462668657512|2.5313819816144116|
|    min|              53.0|             212.0|               1.0|
|    max|            1982.0|            4893.0|               9.0|
+-------+------------------+------------------+------------------+
```

Figure3 : Summary Statistics

    - **Correlation Matrix**: Relationships between features were studied, indicating weak correlations:
        - `DistanceKM` vs. `UnitsProduced`: -0.386.
        - `DistanceKM` vs. `DeliveryCount`: -0.198.
        - `UnitsProduced` vs. `DeliveryCount`: -0.242.
2. **Handling Missing and Outlier Values**:
    - Missing values in key columns were filled with defaults (e.g., `0` for numeric fields).
    - Outliers in `DistanceKM` were removed using z-scores with a threshold of 3.
3. **Feature Normalization**:
    - A **VectorAssembler** combined `DistanceKM`, `UnitsProduced`, and `DeliveryCount` into a feature vector.
    - Features were standardized using **StandardScaler** for improved model performance.

Train Machine Learning Model

1. **Model Selection and Training**:
    - A **Random Forest Regressor** was chosen for its ability to handle non-linear relationships and interpret feature importance.
    - The dataset was split into **80% training** and **20% testing** sets.
    - Hyperparameter tuning was performed using a parameter grid for the number of trees.
2. **Evaluation**:
    - The trained model was evaluated on the test set, yielding an RMSE (Root Mean Squared Error) of **0.3369**, indicating good predictive accuracy.

Visualization

Below is a visualization of the model evaluation metric:

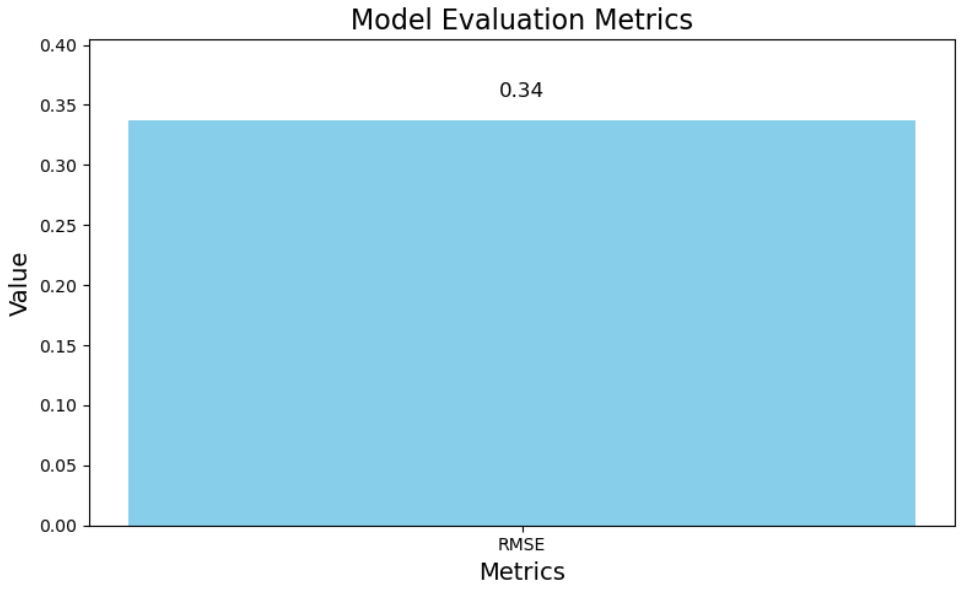

Figure 3 Evaluation plot

Outcome

The analysis revealed a reliable model for predicting CO_2 emissions based on features such as distance traveled, units produced, and delivery count. The trained **Random Forest Regression Model** with optimized parameters demonstrated strong performance with low error rates.

This structured approach ensures robust predictions, enabling actionable insights for sustainability improvements in supply chains.

GreenLink Supply Chain Footprint Analyzer Dashboard Description

The **GreenLink Supply Chain Footprint Analyzer** dashboard provides a comprehensive overview of key metrics and insights related to the sustainability of supply chain activities. It integrates descriptive and predictive analytics, enabling stakeholders to monitor and optimize carbon emissions across various dimensions of the supply chain.

1. Key Metrics Overview

- **Delivery Count (Top Left)**: Displays the total number of deliveries made across the supply chain (e.g., 95).
- **Distance Traveled (Top Center)**: Highlights the total kilometers traveled by all transportation modes (e.g., 20K km).
- **Units Produced (Top Right)**: Indicates the total units produced by manufacturing activities (e.g., 53K units).
- **Average CO_2 Emissions Per Unit (Top Far Right)**: Shows the average carbon emissions generated per unit of production (e.g., 1.98 CO_2/unit).

2. CO_2 Delivery Under Supplier

- **Visualization Type**: Horizontal Bar Chart
- **Purpose**: Shows the contribution of different suppliers to CO_2 emissions per delivery.
- **Insight**: Helps identify suppliers with the highest carbon footprint, allowing focused interventions.

3. Sum of Average CO_2 Per KM by Transportation Mode

- **Visualization Type**: Bar Chart
- **Purpose**: Displays the average CO_2 emissions per kilometer for different transportation modes such as Plane, Truck, Ship, Train, and Drone.
- **Insight**: Enables analysis of the most eco-friendly and least eco-friendly transportation options.

4. Geospatial Distribution of CO_2 Emissions

- **Visualization Type**: Map
- **Purpose**: Displays the geographical locations of suppliers and their respective CO_2 delivery contributions.
- **Insight**: Helps identify emission hotspots across different continents (e.g., North America, Europe, Asia).

5. Production Overview

- **Visualization Type**: Pie Chart
- **Purpose**: Displays the proportion of production under different IDs.
- **Insight**: Provides a high-level breakdown of production activities.

6. CO_2 Per Delivery by Supplier

- **Visualization Type**: Line Chart
- **Purpose**: Displays the CO_2 emissions per delivery across different suppliers.
- **Insight**: Highlights how different suppliers compare in terms of carbon efficiency, allowing optimization.

Conclusion

This dashboard enables users to analyze and interpret sustainability metrics across the supply chain. By combining geospatial, categorical, and predictive visualizations, it provides actionable insights for reducing the carbon footprint of supply chain operations.