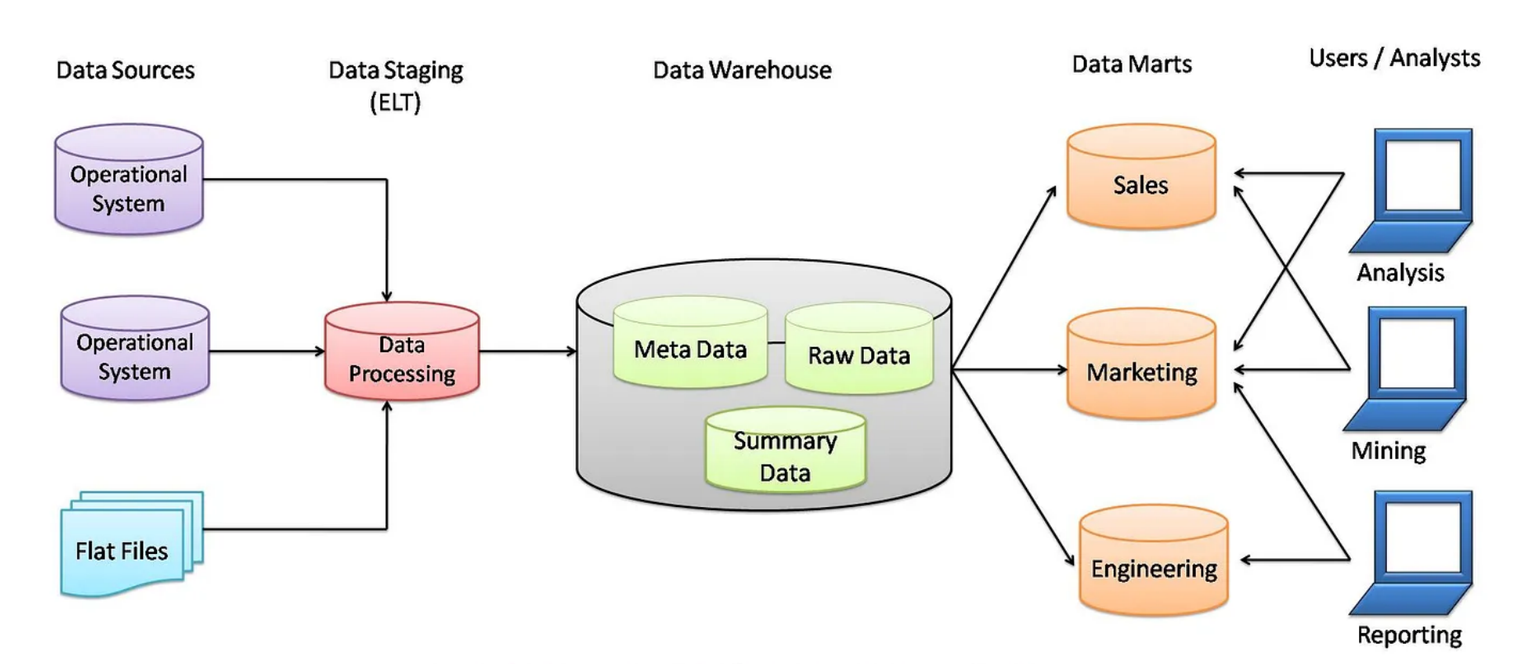
Data Warehousing Concepts for Beginners — Data Engineers

Rahul Sounder · Follow
5 min read · Jan 8, 2024



> *Data warehousing is a process of collecting, storing, and managing data from different sources to support business decision-making. It involves the integration of data from various operational systems into a central repository, known as a data warehouse.*

Data Warehouse

- A centralized repository that stores integrated, historical data from various sources.
- It is optimized for reporting and analysis rather than transaction processing.

ETL (Extract, Transform, Load)

- The process of extracting data from source systems, transforming it to meet the data warehouse's requirements, and loading it into the data warehouse.
- ETL ensures that data is cleansed, standardized, and integrated before being stored in the data warehouse.

Data Mart

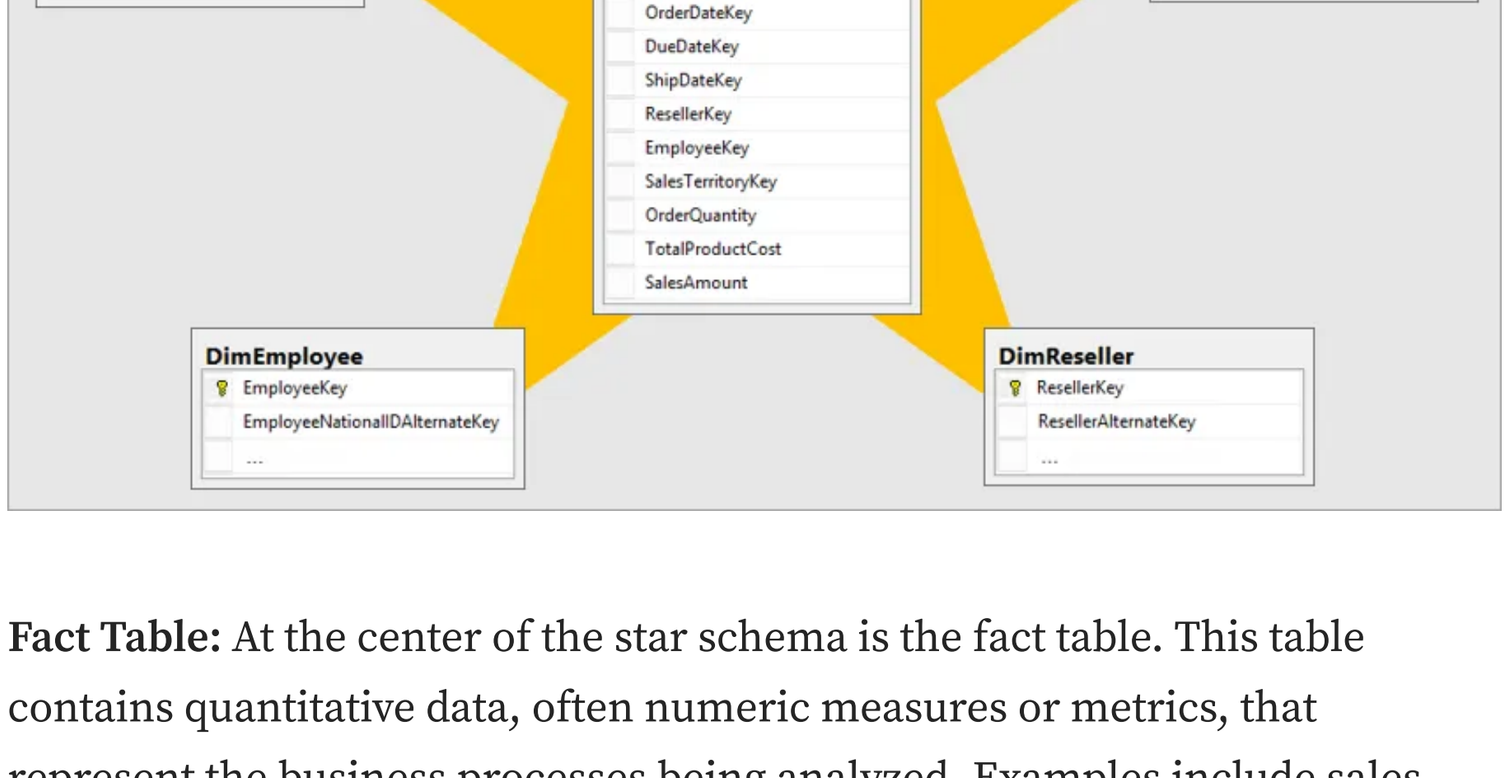
- A subset of a data warehouse that is designed for a specific business line, department, or function.
- Data marts can be independent or linked to the main data warehouse.

Dimensional Modeling

- A modeling technique used in data warehousing to organize and structure data for easy querying and reporting.
- It involves defining dimensions (descriptive attributes) and facts (numeric measures) to create a star or snowflake schema.

Star Schema

> *A dimensional model where a central fact table is connected to dimension tables, forming a star-like structure.*



Fact Table: At the center of the star schema is the fact table. This table contains quantitative data, often numeric measures or metrics, that represent the business processes being analyzed. Examples include sales revenue, quantity sold, or profit.

Dimension Tables: Surrounding the fact table are dimension tables. Each dimension table represents a specific aspect or attribute related to the business process. Dimensions are descriptive and provide context to the data in the fact table. Examples of dimensions could be time, geography, product, or customer.

Relationships: The fact table and dimension tables are connected through relationships established by keys. A primary key in a dimension table is linked to a foreign key in the fact table. These relationships allow for the integration of data across different dimensions.

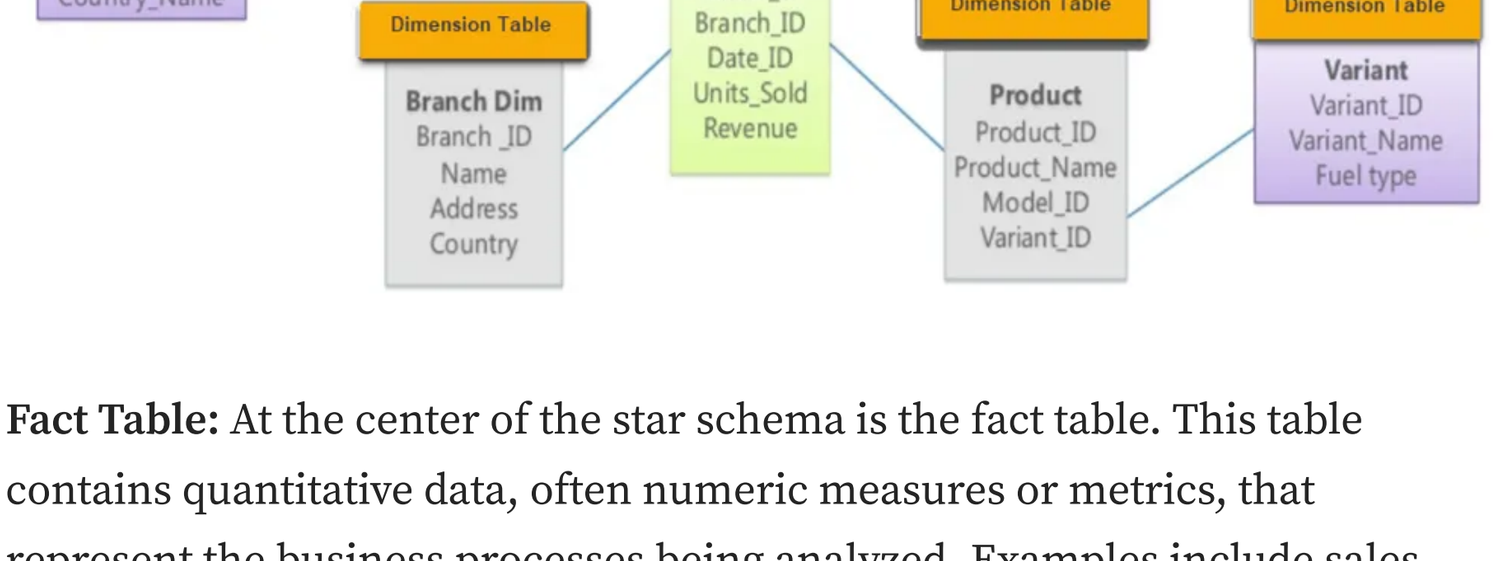
Attributes: Each dimension table contains attributes that provide details about the dimension. For example, a time dimension might have attributes such as year, quarter, month, and day. Attributes in dimension tables are used for filtering, grouping, and labeling data in the fact table.

Advantages of the Star Schema:

- **Simplicity**: The star schema is straightforward and easy to understand. Its simplicity makes it user-friendly for both database administrators and end-users.
- **Query Performance:** The star schema is designed for optimal query performance. Queries can be executed quickly because the structure allows for efficient joins between the fact table and dimension tables.
- **Flexibility**: The star schema is flexible and adaptable to changes in business requirements. New dimensions can be added, and existing ones can be modified without significant impact on the overall structure.
- **Scalability**: The star schema is scalable, making it suitable for large data warehouses. As the volume of data grows, the star schema can handle it effectively, provided proper indexing and optimization are implemented.

Snowflake Schema

> *The snowflake schema is another type of dimensional data model used in data warehousing, similar to the star schema. Like the star schema, it organizes data for efficient querying and reporting, but the snowflake schema takes a more normalized approach to the structure. The term "snowflake" refers to the shape the schema takes on when visualized: a central fact table surrounded by dimension tables that are further normalized into a branching, snowflake-like pattern.*



Fact Table: At the center of the star schema is the fact table. This table contains quantitative data, often numeric measures or metrics, that represent the business processes being analyzed. Examples include sales revenue, quantity sold, or profit.

Dimension Tables: Surrounding the fact table are dimension tables. Each dimension table represents a specific aspect or attribute related to the business process. Dimensions are descriptive and provide context to the data in the fact table. Examples of dimensions could be time, geography, product, or customer.

Sub-Dimensions: Each dimension table in the snowflake schema can have sub-dimensions or related tables that store additional attributes. These sub-dimensions help to reduce redundancy by separating data into different tables.

Normalization: The snowflake schema employs normalization techniques by breaking down dimension tables into smaller, related tables. This reduces data redundancy and improves data integrity but can result in more complex queries due to additional joins.

Relationships: The fact table and dimension tables are connected through relationships established by keys. A primary key in a dimension table is linked to a foreign key in the fact table. These relationships allow for the integration of data across different dimensions.

Attributes: Each dimension table contains attributes that provide details about the dimension. For example, a time dimension might have attributes such as year, quarter, month, and day. Attributes in dimension tables are used for filtering, grouping, and labeling data in the fact table.

Advantages of the Snowflake Schema:

1. Reduced Redundancy: The snowflake schema reduces data redundancy by normalizing dimension tables. This can save storage space and improve data integrity.
2. Easier Maintenance: Because of the normalization, making changes to the schema, such as updating attributes or adding new ones, can be more straightforward and less prone to errors.
3. Improved Data Integrity: Normalization can enhance data integrity by reducing the risk of update anomalies that can occur when redundant data is stored in multiple places.
4. Suitability for Hierarchical Data: The snowflake schema is well-suited for representing hierarchical relationships within dimensions.

However, there are some trade-offs with the snowflake schema:

1. Query Performance: Due to the normalization and increased number of joins, query performance in a snowflake schema may be slightly slower compared to a star schema.
2. Complexity: The snowflake schema can be more complex to understand and work with, especially for users who are not familiar with the structure.

> ***The choice between a star schema and a snowflake schema depends on factors such as the nature of the data, the specific business requirements, and the balance between simplicity and normalization needs in the data warehouse design.***

Metadata

- Information about the data in the data warehouse, including its source, transformation rules, and usage.
- Metadata helps users understand and interpret the data in the warehouse.

Data Mining

- The process of discovering patterns and relationships in data stored in the data warehouse.
- Data mining techniques help uncover hidden insights and support predictive analysis.

Written by Rahul Sounder
208 Followers · 6 Following
Follow
Senior Engineering Manager - Data at Xiaomi Technology | Ex-Amazon, Merck | SAFe® 5 Agilist | Certified AWS Solutions Architect