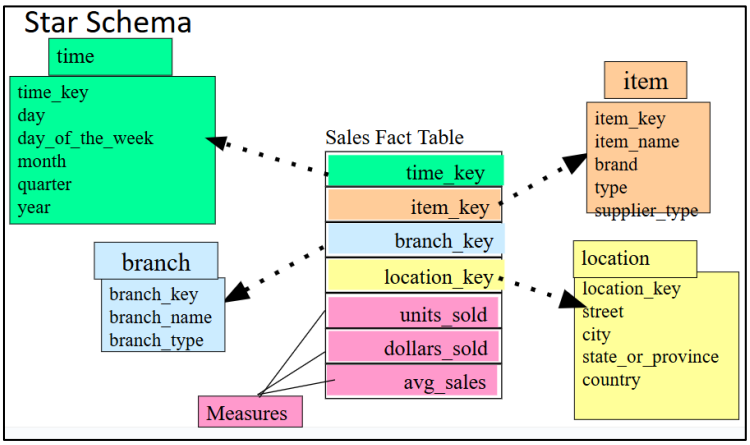
DMBI Experiment 1 – Design a Star and Snowflake Schema

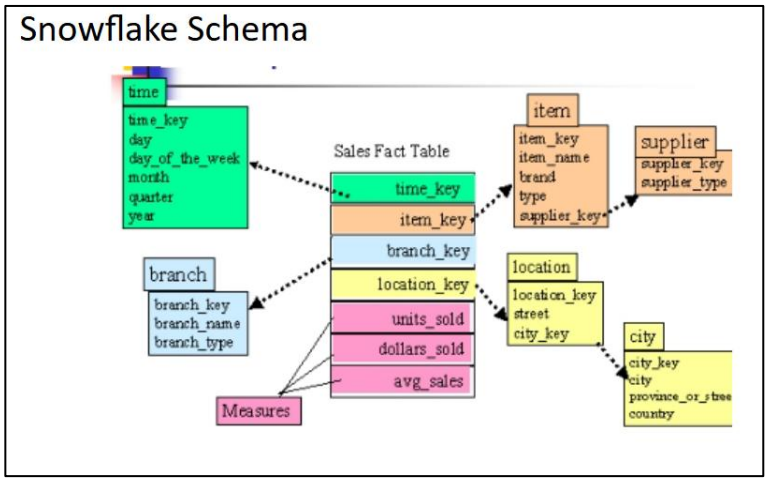
Name: Atharva Lotankar
Class: D15C; **Rnum:** 27
Batch B

Aim: To Design a Star and Snowflake Schema for the given system.

Theory:

A **Star Schema** is a widely used data warehouse model where a central fact table is connected to dimension tables. It is intuitive, simple, and allows fast query performance since most queries can be resolved with fewer joins. Its denormalized structure improves retrieval speed but can lead to redundancy and larger storage requirements.



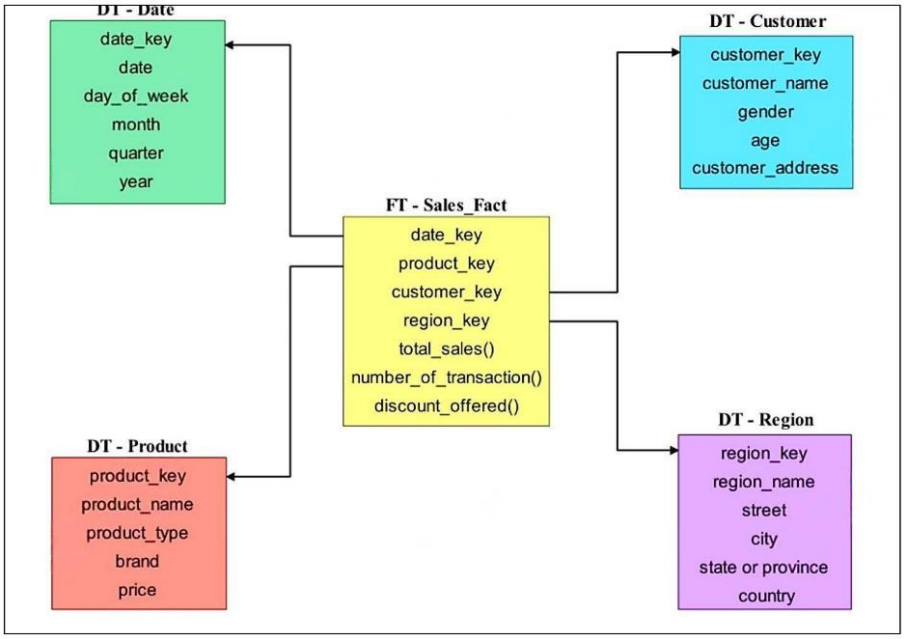


A **Snowflake Schema** is a normalized extension of the star schema in which dimension tables are further split into sub-dimensions. This reduces data redundancy and saves storage space but increases query complexity due to multiple joins, often leading to slower performance compared to the star schema.

In **OLAP (Online Analytical Processing)**, both schemas play a critical role. The star schema is preferred for rapid, straightforward query execution and ease of understanding by business users. On the other hand, the snowflake schema is useful where storage efficiency and data consistency are prioritized. Choosing the schema depends on the specific OLAP requirements—whether performance speed or normalized structure is more important. Together, they provide the foundation for effective multidimensional analysis, enabling businesses to derive insights and make informed decisions.

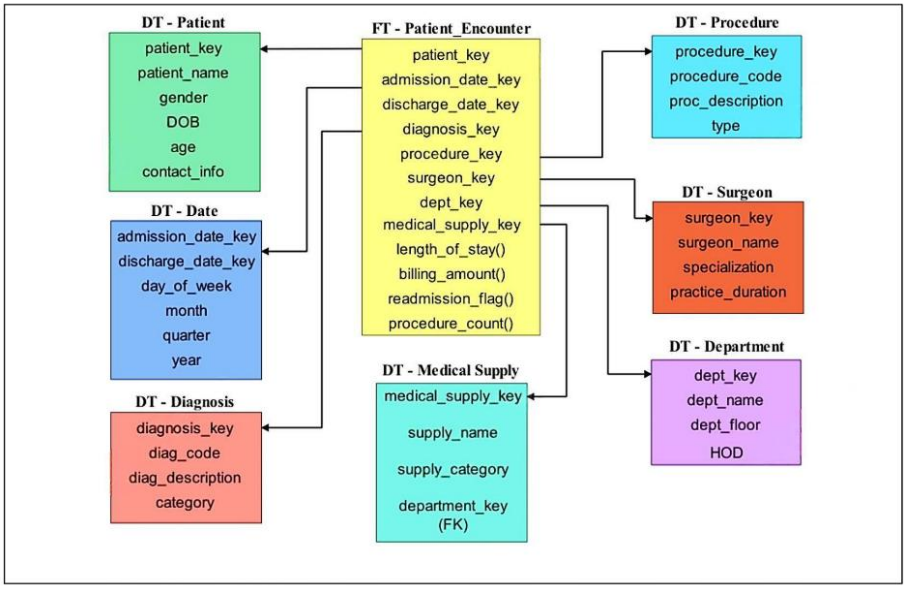
Design Implementation based on Questions:

Q) *A retail company wants to analyse its sales performance across different regions, time periods, products, and customer segments. The company wants to track total sales, number of transactions, and discount offered.*



Q) *A hospital management wants to create a data warehouse to analyze patient admissions, procedures, and billing information. The goal is to improve operational efficiency and patient care by answering questions such as:*

- *What is the average length of stay for patients with a specific diagnosis?*
- *How many surgical procedures were performed by each surgeon last month?*
- *What is the total revenue generated by a particular department (e.g., Cardiology, Orthopedics) per quarter?*
- *Which medical supplies are most frequently used in the emergency department?*
- *What is the readmission rate for patients who had a certain procedure?*



Based on Hospital Management Design:

- The Average length of stay for patients with specific diagnosis, one can use patient_encounter with diagnosis_key and length_of_stay.
- Surgical Procedures that were performed by each surgeon last month, could be figured out by filtering – procedure_key with surgeon_key and admission_date_key.
- Total Revenue generated by particular a department per quarter is by sum() billing_amount() grouping() by department_key and quarter (in date table).
- Medical supplies that are most frequently used in the emergency department by aggregating usage counts of medical_supply_key within department_dim filtered for emergency.
- Readmission rate for patients who had a certain procedure can be done by using readmission_flag() with procedure_key.

Inference:

Based on the hospital management design and the retail company analysis scenarios, the following inferences can be made regarding star and snowflake schemas:

1. The star schema suits both use cases due to its simple, denormalized structure enabling fast and straightforward queries for performance metrics like sales totals or patient stay lengths.
2. The snowflake schema, with normalized dimension tables, provides better data integrity and storage optimization, especially useful in complex contexts like hospital procedures with many keys and hierarchical relationships.
3. For operational efficiency and detailed hierarchical data (e.g., department and procedure keys), snowflake schema helps maintain consistent relationships, while star schema benefits quick ad-hoc reports and real-time analytics in retail sales and patient billing.

Thus, schema choice depends on the trade-off between query speed and data normalization/management complexity for the specific analytical tasks.

Results & Conclusion:

The experiment demonstrates that star schema offers faster query performance and simpler design, ideal for quick analytics in retail and hospital management. Snowflake schema provides optimized storage and better data integrity but involves complex queries and maintenance. Thus, choosing between them depends on balancing speed and storage needs, concluding that hybrid approaches may best suit evolving business requirements.