Data Warehouse Concepts

Anil Reddy Chenchu

www.linkedin.com/in/chenchuanil

Data Warehouse

Data warehouse is a process for collecting the data and managing data from different sources to provide meaningful business insights,which is used for data analysis and reporting.Data warehouse holds only transformed data and it represents data only in structured format.

Ex : Hive
AWS Redshift
Snowflake
Azure Synapse Analytics

Data Lake : A data lake is a centralized repository that allows you to store all your structured, semi-structured and unstructured data at any scale. Data lakes are designed to handle large volumes of diverse data efficiently. Ex : HDFS, AWS S3, Azure Blob, GCP Storage .

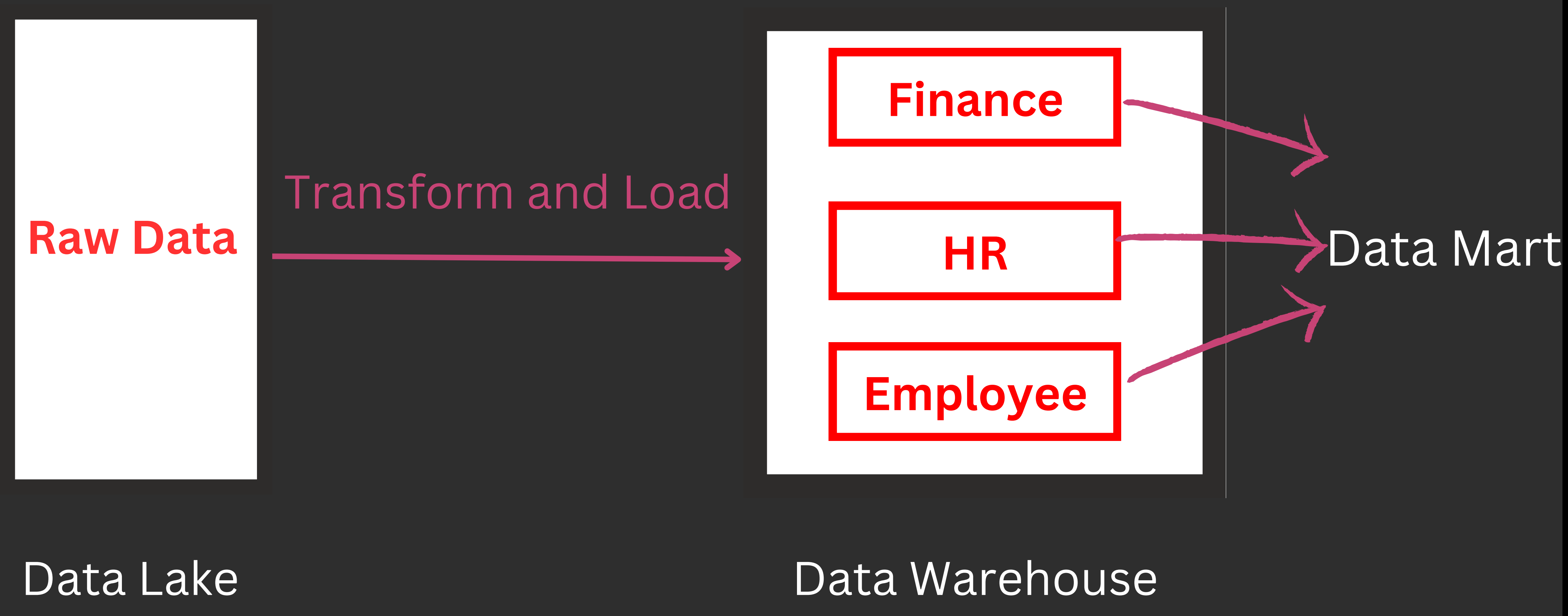
Granularity : A fact table is usually designed at a low level of granularity. This means that we need to find the lowest level of information that can be stored in a fact table.
The depth of the data level is known as granularity.
Ex: In date dimension, the level could be year, month, quarter, period, week, and day of granularity.

The following pages will provide additional information on fact and dimension tables.

Data Mart

A data mart is a subset of a data warehouse oriented to a specific business line. in simple words category level seperation inside a data warehouse is known as data marts.
Data marts contain repositories of summarized data collected for analysis on a specific section or unit within an organization.

Ex: Finance department, HR department, Employee department



OLTP & OLAP

OLTP(Online transaction processing) : Systems designed to support transactional level operations like insert, update, delete etc. They support ACID properties and main goal is to capture data in real time. In OLTP data resides in normalization form.

Ex : Microsoft SQL Server
Oracle Database
MYSQL
PostgreSQL

OLAP(Online transaction processing) : It is designed to support analytical queries, it keeps data in denormalized form and does not support ACID properties, main goal is to help users to querry historical data and get insights out of it.

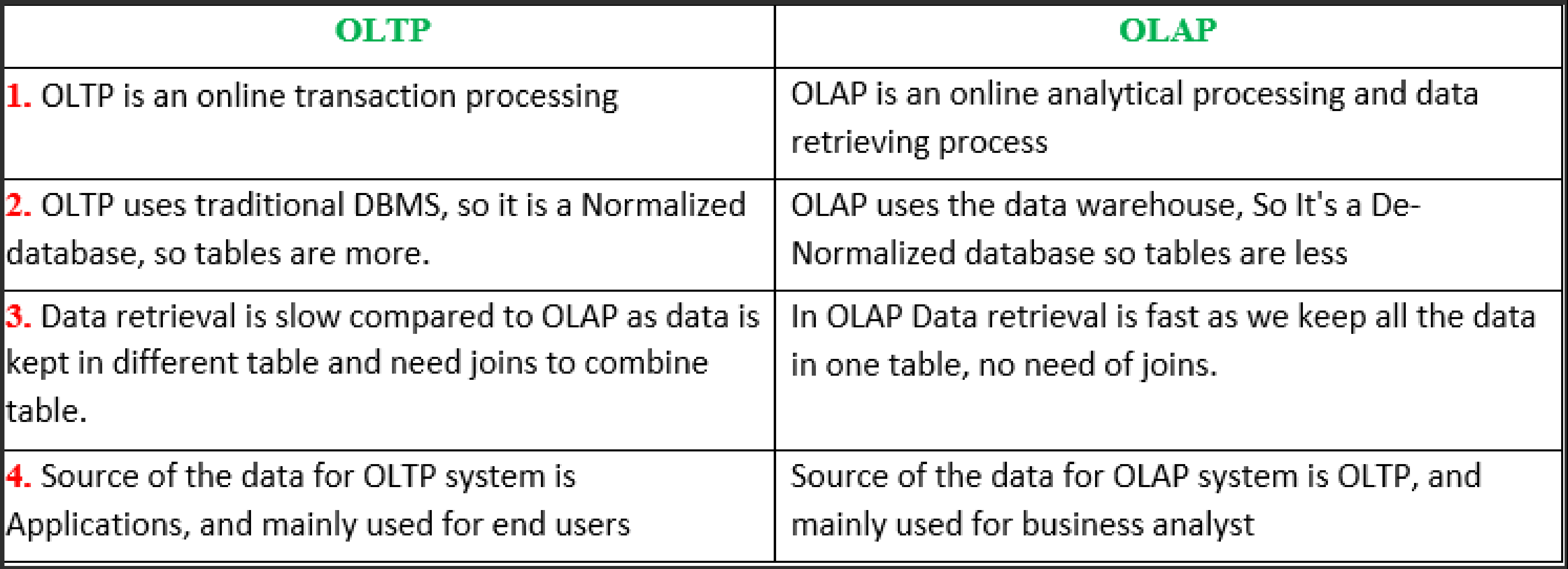
Ex : Redshift
Hive
No SQL
Snowflake

In my next post, I will provide a detailed discussion on the following topics.

- Normalization: The process of organizing data in a database to reduce redundancy and improve data integrity.

  Denormalization: The process of intentionally introducing redundancy into a database to improve read performance by reducing the number of joins needed in queries.t
- ACID Properties: A set of properties (Atomicity, Consistency, Isolation, Durability) that ensure reliable processing of database transactions.

Difference between OLTP and OLAP



OLTP	OLAP
1. OLTP is an online transaction processing	OLAP is an online analytical processing and data retrieving process
2. OLTP uses traditional DBMS, so it is a Normalized database, so tables are more.	OLAP uses the data warehouse, So It's a De-Normalized database so tables are less
3. Data retrieval is slow compared to OLAP as data is kept in different table and need joins to combine table.	In OLAP Data retrieval is fast as we keep all the data in one table, no need of joins.
4. Source of the data for OLTP system is Applications, and mainly used for end users	Source of the data for OLAP system is OLTP, and mainly used for business analyst

Fact Tables

A fact table is a central table in a star schema of a data warehouse. It stores quantitative data for analysis and is often used in conjunction with dimension tables.

Key Characteristics of Fact Tables:

1. Measures: Fact tables contain the measurable, quantitative data of a business process. Examples include sales amount, quantity sold, and transaction counts.
2. Foreign Keys: They include foreign keys that link to dimension tables, which provide context to the measures. For example, a sales fact table might link to dimensions like time, product, and store.
3. Granularity: The level of detail in a fact table is known as its granularity. It could be at the level of individual transactions, daily summaries, or monthly summaries.

4. Additive, Semi-Additive, and Non-Additive Facts:
   - Additive: Measures that can be summed across any dimension (e.g., sales amount).
   - Semi-Additive: Measures that can be summed across some dimensions but not all (e.g., account balances).
   - Non-Additive: Measures that cannot be summed (e.g., ratios, percentages).

Dimension Tables

Dimension tables are a crucial part of a data warehouse schema, often used in conjunction with fact tables. They provide the context and descriptive information needed to understand the measures stored in fact tables.

Key Characteristics of Dimension Tables:

1. Attributes: Dimension tables contain descriptive attributes (or fields) that provide context to the data in fact tables. For example, a product dimension might include attributes like product name, category, and brand.
2. Primary Keys: Each dimension table has a primary key that uniquely identifies each record. This primary key is used to link to the fact table.
3. Denormalization: Dimension tables are often denormalized to improve query performance. This means they might contain redundant data to avoid complex joins
4. Hierarchies: They often include hierarchical relationships, such as date dimensions having year, quarter, month, and day attributes.

STAR SCHEMA

In star schema one or more fact tables will be connected with any number of dimension tables. Star schema means Dimension tables are directly linked to the fact table. Creating a structure that resembles a star.

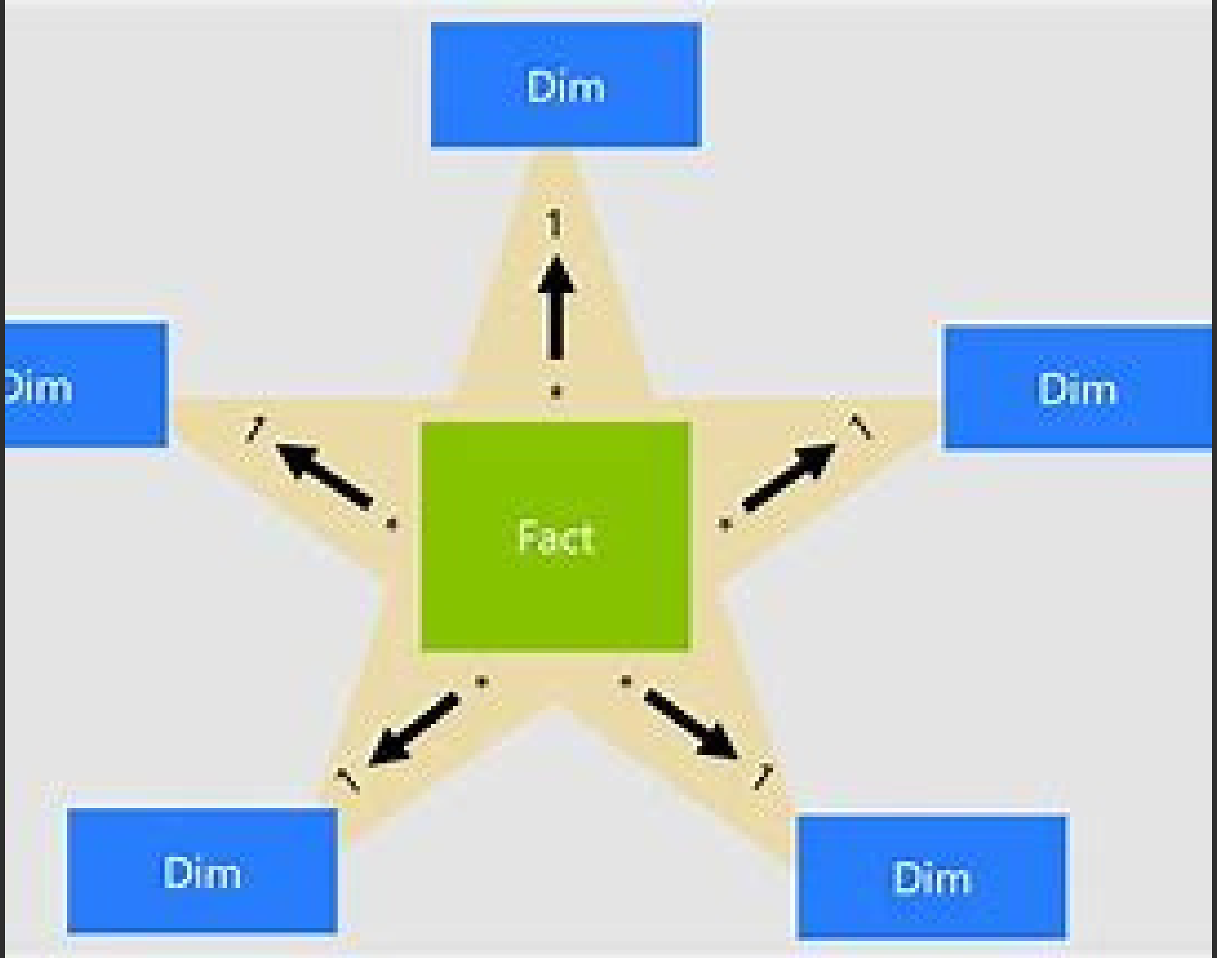
ADVANTAGES OF STAR SCHEMA

1. The Star schema is having de-normalized form and it tends to be better for performance.

2. Star schema uses less foreign keys so the query execution is fast.

3. The Star schema is easier for readability because its query structure is not as complex, the Star schema uses less joins and tends to have more data redundancy.

4. So for readability the schema to go with would be the star schema.



www.linkedin.com/in/chenchuanil

SNOWFLAKE SCHEMA

It is an extension of the star schema and is characterized by the normalization of dimension tables into multiple related tables, creating a structure that resembles a snowflake.

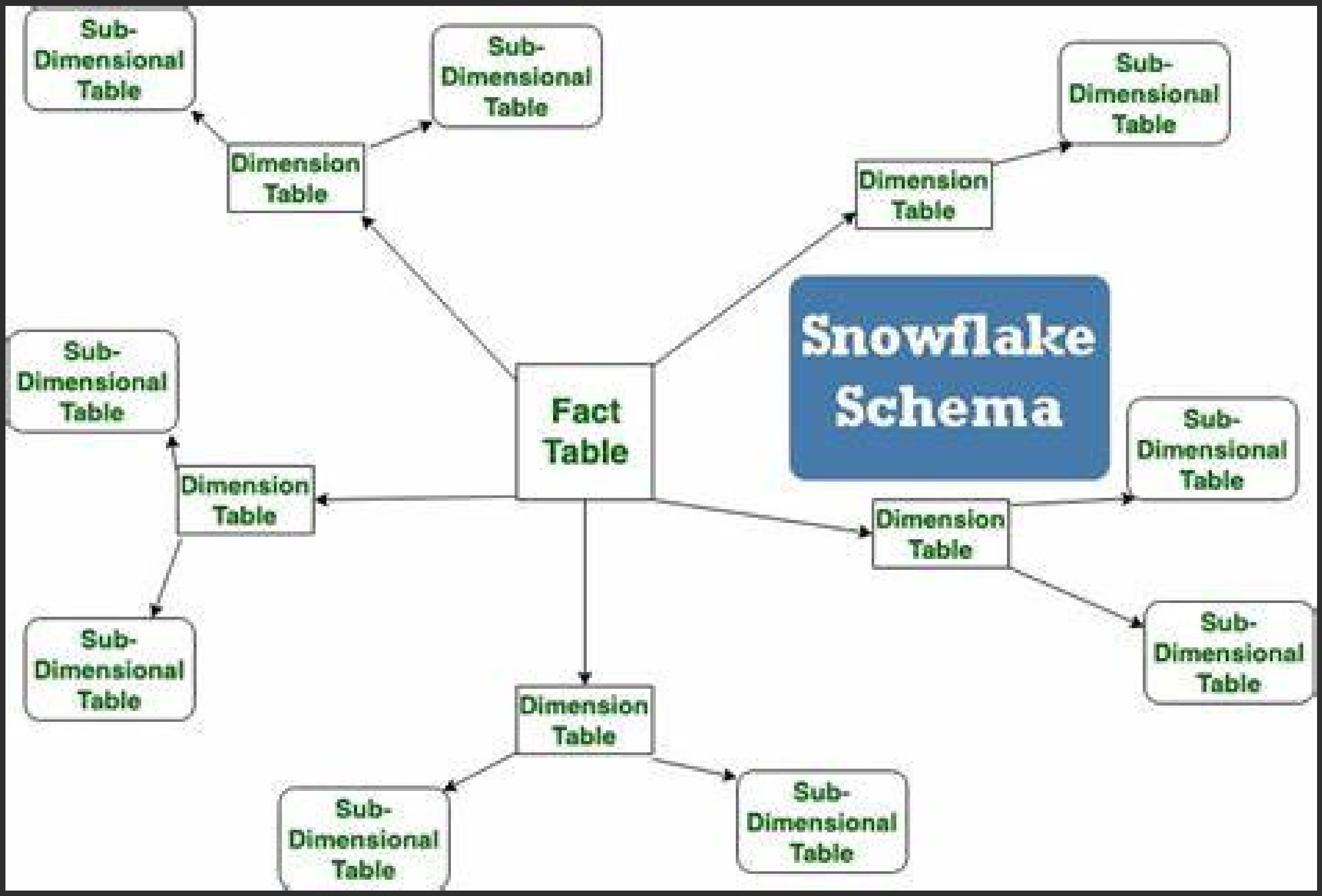
Advantages:

Reduced Data Redundancy: Normalization minimizes duplicate data.
Improved Data Integrity: Easier to maintain and update.

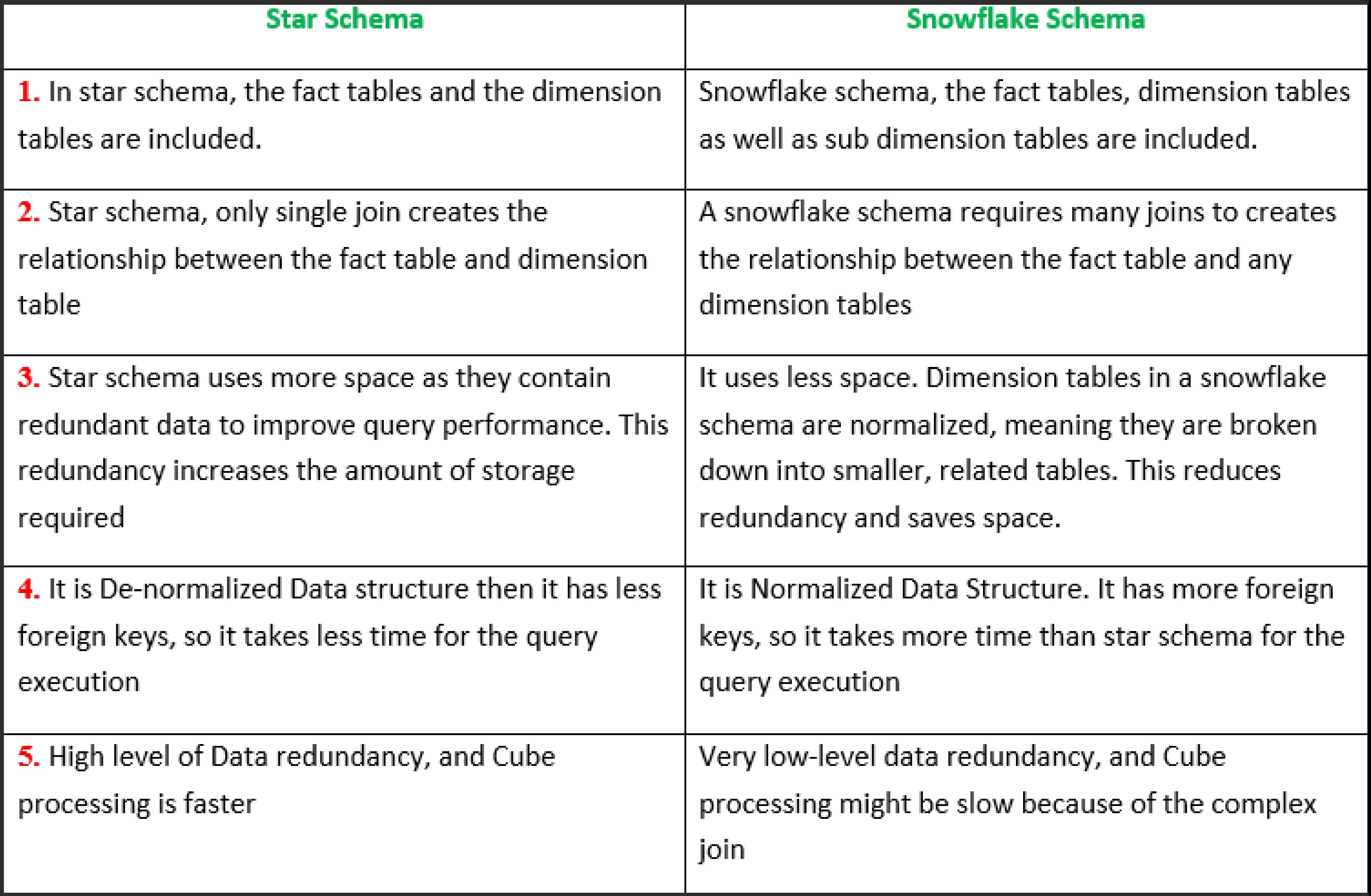
Disadvantages:

Complex Queries: More joins are required, which can slow down query performance.

Increased Complexity: More tables and relationships to manage.



DIFFERENCE BETWEEN STAR & SNOWFLAKE SCHEMAS



Star Schema	Snowflake Schema
1. In star schema, the fact tables and the dimension tables are included.	Snowflake schema, the fact tables, dimension tables as well as sub dimension tables are included.
2. Star schema, only single join creates the relationship between the fact table and dimension table	A snowflake schema requires many joins to creates the relationship between the fact table and any dimension tables
3. Star schema uses more space as they contain redundant data to improve query performance. This redundancy increases the amount of storage required	It uses less space. Dimension tables in a snowflake schema are normalized, meaning they are broken down into smaller, related tables. This reduces redundancy and saves space.
4. It is De-normalized Data structure then it has less foreign keys, so it takes less time for the query execution	It is Normalized Data Structure. It has more foreign keys, so it takes more time than star schema for the query execution
5. High level of Data redundancy, and Cube processing is faster	Very low-level data redundancy, and Cube processing might be slow because of the complex join