Data Science & Artificial Intelligence

Warehousing

Schema for multidimensional data models

Oneshot

Astha Singh Ma'am

Recap of Previous Lecture

Data transformation -

Topics to be covered

- Discretization
- Data Warehouse Modelling
- Schemas for Multidimensional data models

Topic : Discretization

- Data Generalization is the process of summarizing data by replacing relatively low level values with higher level concepts.

Topic : Method of Discretization

Equal-width Binning

Grouping data into fixed-width intervals to create equally sized bins.

Equal-frequency Binning

Dividing data into bins of equal frequency to ensure similar instances are grouped together.

Clustering-based Discretization

Utilizing clustering algorithms to group similar data points into discrete values.

Entropy-based Discretization

Using information theory to find the optimal cut-off points for data discretization.

Topic : Advantages of Discretization

1. **Reduction of Data Complexity**

   Simplifying complex continuous data into discrete categories for easier analysis.

2. **Improved Performance of Analysis Algorithms**

   Enhancing the efficiency and accuracy of certain algorithms designed for discrete data.

3. **Enhanced Interpretability of Results**

   Allowing for clearer insights and easier communication of findings to diverse audiences.

Topic : Challenges and Considerations in Discretization

Determining Appropriate Number of Bins

Strategically selecting the number of bins to ensure meaningful data granularity.

Handling Outliers and Missing Values

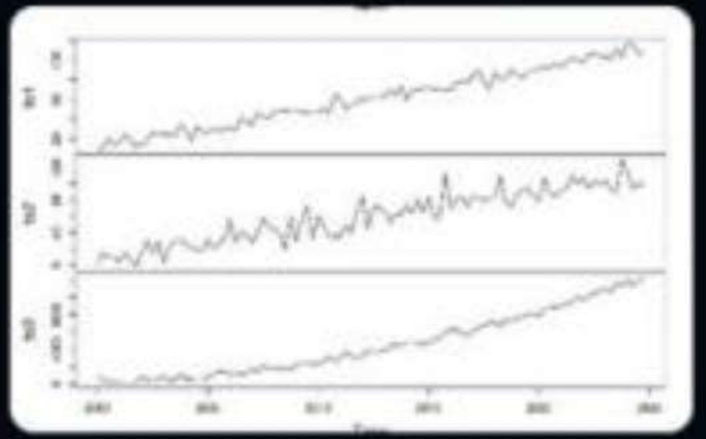
Dealing with extreme values or missing data points during the discretization process.

Selecting Suitable Discretization Method

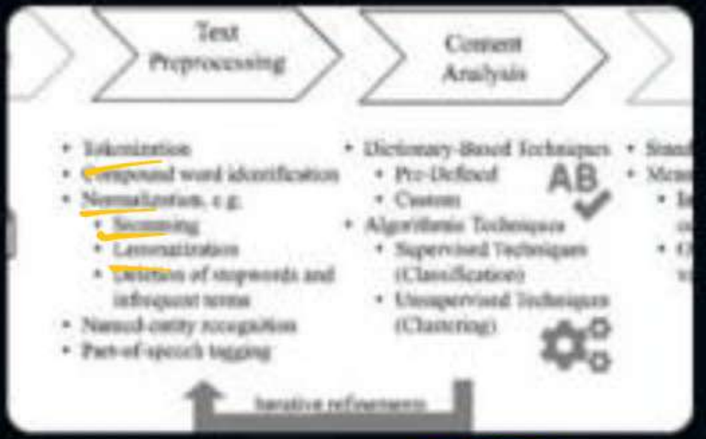
Choosing the most suitable technique based on the characteristics of the dataset and the specific analysis requirements.

Topic : Examples of Discretization in Practice

Feature Engineering for Machine Learning
How discretization contributes to effective feature engineering for better predictive models.



Discretization in Time Series Analysis
Unveiling hidden patterns and trends in timedependent data through discretization techniques.



Discretization in Text Mining and Sentiment Analysis
Unifying semantic information and unraveling sentiments through intelligent text discretization.

Topic : Numerical

- Calculate the bin width. If you decide to have four group and the age range is 18 to 70 the bin width would be Original data [21,35,28,42,50,18,70,30,40,60]

→ Original Data we can calculate in (4) bin

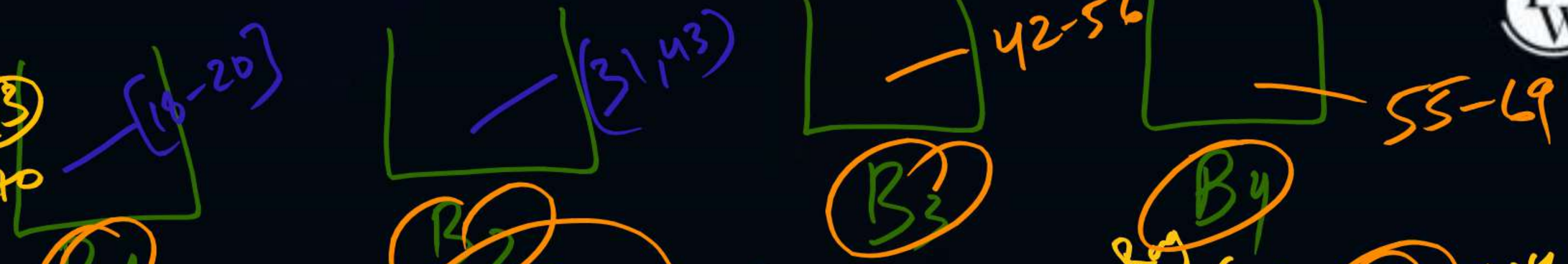
* Total no. of bin n $\boxed{n=4}$ ✓

Min = 18
Max = 70

Allocate $\boxed{\text{Width} = \frac{\text{Max} - \text{Min}}{n}}$

$\Rightarrow \frac{70-18}{4} = 13$

Min = 18
Width = 13
Max = 70



$B1 = Min - (Min + Width - 1)$ (Range)

$= 18 - ((18+13) - 1)$

$= [18 - 30]$

$B2 = Min + Width$

$= 18 + 13 - (18 + 2 \times 13 - 1)$

$= 31 - (43)$

$= [31 - 43)$ Range

$B4 = (Min + 2 \times Width - 1)$

$B_3 = \text{Min} + 2\,\text{width} - (\text{Min} + 3 \times \text{width} - 1)$

$= 18 + 2 \times 13 - (18 + 3 \times 13 - 1)$

$42 - (56)$

$[42 - 56]$ Range

$$B_4 = \text{Min} + 3 \times \text{Width} \;\underbrace{\quad}_{\text{Range}}\; (\text{Min} + 4 \times \text{width} - 1)$$

55

$$55 - (18 + 4 \times 13 - 1)$$

$$= 55 - 69 \quad \text{Range}$$

MCQ

#Q. What is discretization in the context of data analysis?

A. The process of converting continuous data into categorical or discrete intervals.

B. The process of transforming categorical data into numerical values.

C. The process of removing outliers from a dataset.

D. The process of standardizing data to have zero mean and unit variance.

MCQ

#Q. Which of the following is an advantage of discretization?

A. Increased interpretability of results.

B. Loss of information in the dataset.

C. Complexity in handling categorical data.

D. Difficulty in handling missing values.

MCQ

#Q. Which of the following techniques is commonly used for discretization?

A Principal Component Analysis (PCA).

B K-means clustering.

C Decision tree-based methods.

D Support Vector Machines (SVM).

MCQ

#Q. In supervised discretization, what does the process rely on?

A. Unlabeled data.

B. Expert knowledge or a target variable.

C. Random assignment of categories.

D. Dimensionality reduction techniques.

MCQ

#Q. What is Equal Frequency Discretization?

A Dividing data into intervals with an equal number of data points in each interval.

B Assigning equal weights to all data points.

C Dividing data based on the mean and standard deviation.

D Discretizing data without considering frequencies.

MCQ

#Q. Which of the following statements is true regarding discretization and decision trees?

A Decision trees are not affected by discretization.

B Discretization is unnecessary when using decision trees.

C Decision trees often benefit from discretization.

D Discretization makes decision trees less interpretable.

MCQ

#Q. What is the purpose of discretizing continuous features in machine learning?

A. To increase computational complexity.

B. To simplify the model and improve interpretability.

C. To introduce noise into the dataset.

D. To make the model more sensitive to outliers.

MCQ

#Q. Which method is used to handle missing values during discretization?

A. Removing the instances with missing values.

B. Assigning the mean value to missing instances.

C. Assigning a separate category for missing values.

D. Using clustering algorithms.

MCQ

#Q. Which type of discretization method uses statistical measures such as mean and standard deviation?

A. Equal Width Discretization.

B. Equal Frequency Discretization.

C. Supervised Discretization.

D. Unsupervised Discretization.

MCQ

#Q. What is the potential drawback of discretization?

A. Increased interpretability.

B. Loss of information.

C. Improved handling of outliers.

D. Decreased complexity.

Topic : Data Warehouse Modelling

- Dimensional modeling represents data with a cube operation, making more suitable logical data representation with OLAP data management. The perception of Dimensional Modeling was developed by Ralph Kimball and is consist of "fact" and "dimension" tables

Topic : Objectives of Dimensional Modeling

- To produce database architecture that is easy for end-clients to understand and write queries.
- To maximize the efficiency of queries. It achieves these goals by minimizing the number of tables and relationships between them.

Topic : Advantages Dimensional Modeling

Dimensional modeling is simple
Dimensional modeling promotes data quality:
Performance optimization is possible through aggregates

Topic : Disadvantages of Dimensional Modeling

- To maintain the integrity of fact and dimensions loading the data warehouses with a record from various operational systems is complicated.
- It is severe to modify the data warehouse operation if the organization adopting the dimensional technique changes the method in which it does business.

Topic : Elements of Dimensional Modeling

Fact

It is a collection of associated data items, consisting of measures and context data

Dimensions

It is a collection of data which describe one business dimension.

Measure

It is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions.

Star schema
Snowflake

Fact

It is a collection of associated data items, consisting of measures and context data. It typically represents business items or business transactions.

Dimensions

It is a collection of data which describe one business dimension. Dimensions decide the contextual background for the facts, and they are the framework over which OLAP is performed.

Measure

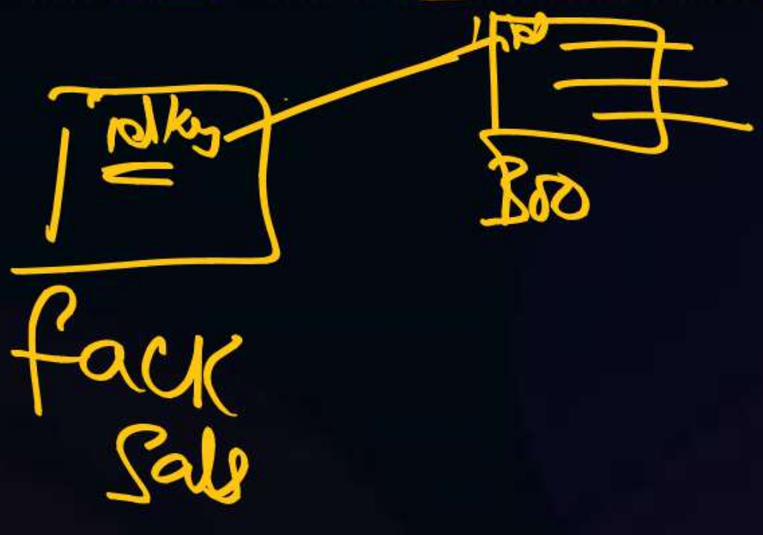
It is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions.

Fact Table

Fact tables are used to data facts or measures in the business. Facts are the numeric data elements that are of interest to the company.

Characteristics of the Fact table

The fact table includes numerical values of what we measure. For example, a fact value of 20 might means that 20 widgets have been sold. Each fact table includes the keys to associated dimension tables. These are known as foreign keys in the fact table.



Dimension Table

Dimension tables establish the context of the facts. Dimensional tables store fields that describe the facts.

Characteristics of the Dimension table

Dimension tables contain the details about the facts. That, as an example, enables the business analysts to understand the data and their reports better.

MCQ

#Q. What is the primary goal of dimensional modeling in a data warehouse?

A. Normalize data for efficient storage

B. Facilitate data entry and transaction processing

C. Provide a structure for easy and efficient querying

D. Implement complex business logic in the database

MCQ

#Q. In dimensional modeling, what is a "Fact Table"?

A. Stores descriptive attributes about business entities

B. Contains primary keys of dimension tables

C. Holds quantitative data for analysis

D. Manages access control for the database

MCQ

#Q. What is a "Dimension" in the context of dimensional modeling?

A. A measure of performance in the data warehouse

B. A set of descriptive attributes related to the business

C. The primary key of a fact table

D. A table containing metadata information

MCQ

#Q. Which schema is commonly used in dimensional modeling for organizing data in a star-like structure?

A Snowflake schema

B Star schema

C Relational schema

D Hierarchical schema

Topic : Schemas for Multidimensional data models

- STAR Schema
- SNOWflakes Schema

Topic : Star Schemas

A star schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions.

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model.

✓ fact, dimension
✓ Relational Schema

Fact Tables

A table in a star schema which contains facts and connected to dimensions. A fact table has two types of columns: those that include fact and those that are foreign keys to the dimension table. The primary key of the fact tables is generally a composite key that is made up of all of its foreign keys.

Dimension Tables

A dimension is an architecture usually composed of one or more hierarchies that categorize data. If a dimension has not got hierarchies and levels, it is called a flat dimension or list.

Characteristics of Star Schema

- The star schema is intensely suitable for data warehouse database design because of the following features:
- It creates a DE-normalized database that can quickly provide query responses.
- It provides a flexible design that can be changed easily or added to throughout the development cycle, and as the database grows.
- It provides a parallel in design to how end-users typically think of and use the data.
- It reduces the complexity of metadata for both developers and end-users.

Advantages of Star Schema

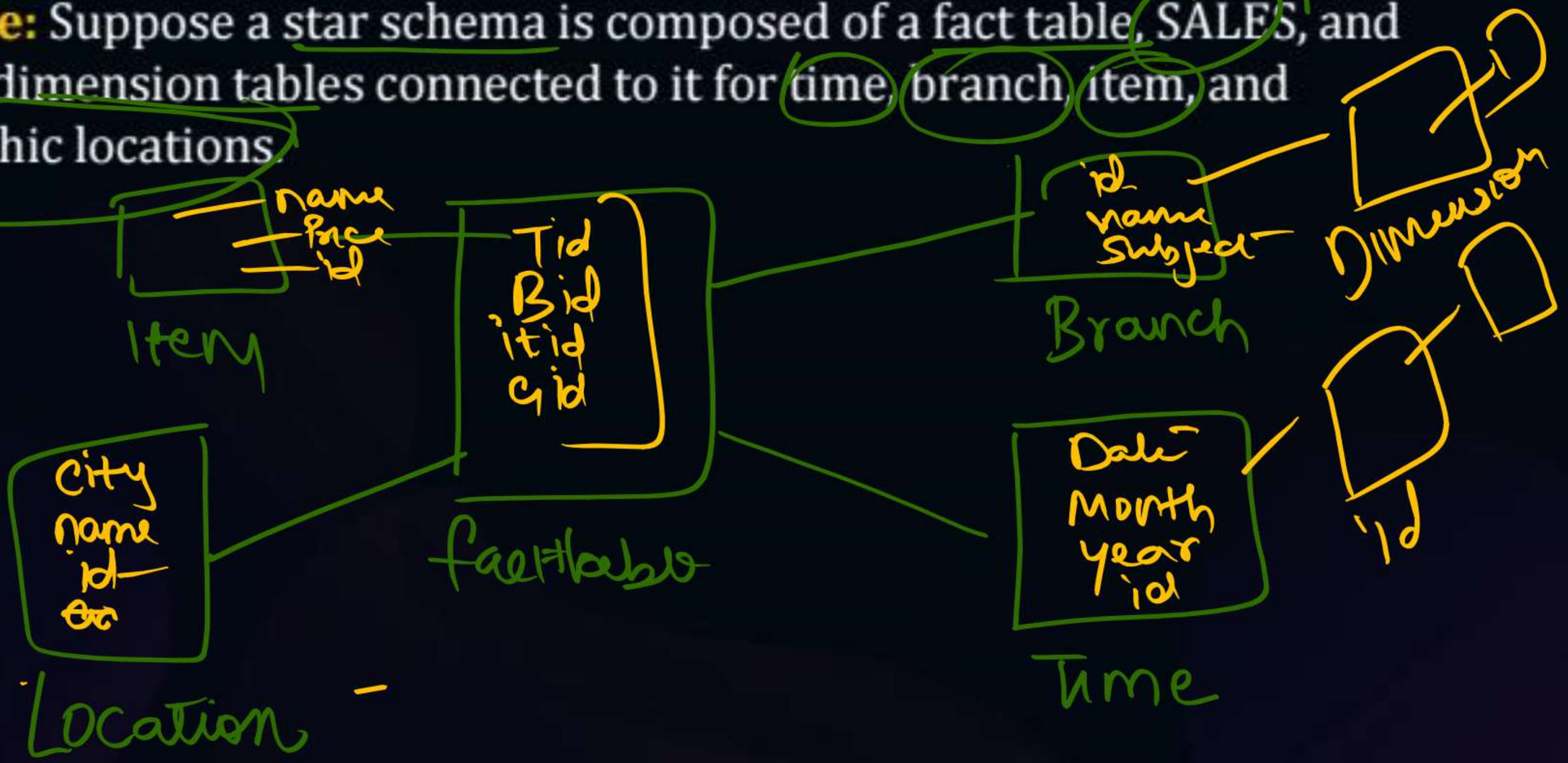
Star Schemas are easy for end-users and application to understand and navigate. With a well-designed schema, the customer can instantly analyze large, multidimensional data sets.

Advantages of Star Schema

Disadvantage of Star Schema

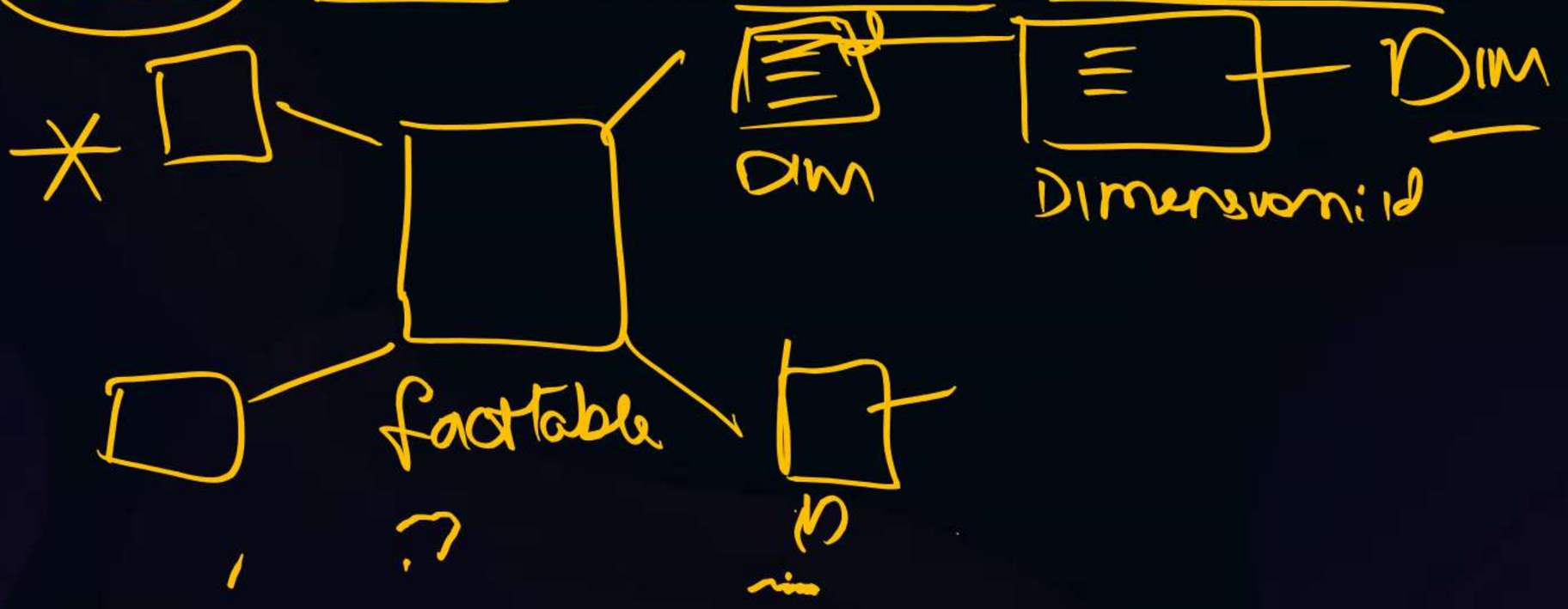
There is some condition which cannot be meet by star schemas like the relationship between the user, and bank account cannot describe as star schema as the relationship between them is many to many.

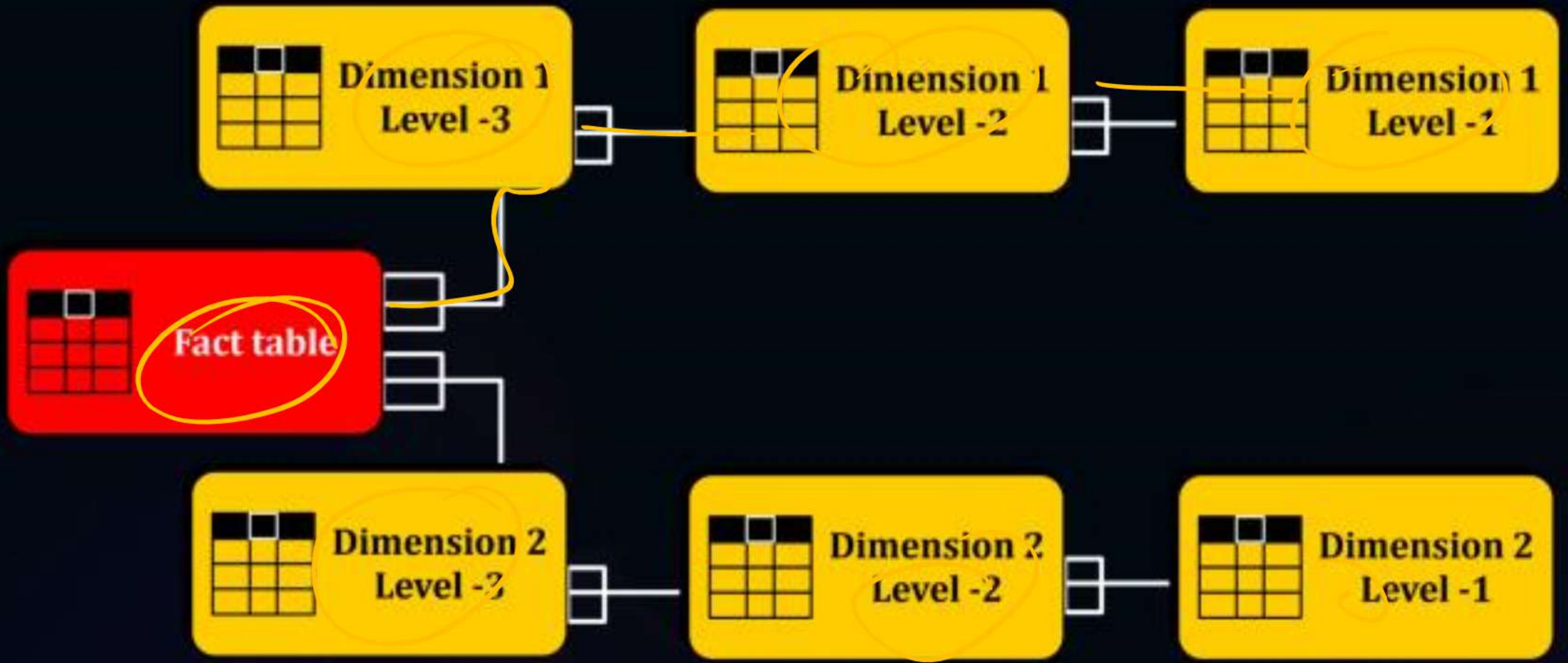
Example: Suppose a star schema is composed of a fact table, SALES, and several dimension tables connected to it for time, branch, item, and geographic locations.



What is Snowflake Schema?

A snowflake schema is equivalent to the star schema. "A schema is known as a snowflake if one or more dimension tables do not connect directly to the fact table but must join through other dimension tables."





Snowflake schema

Example: Figure shows a snowflake schema with a Sales fact table, with Store, Location, Time, Product, Line, and Family dimension tables. The Market dimension has two dimension tables with Store as the primary dimension table, and Location as the outrigger dimension table. The product dimension has three dimension tables with Product as the primary dimension table, and the Line and Family table are the outrigger dimension tables.

Advantage of Snowflake Schema

- The primary advantage of the snowflake schema is the development in query performance due to minimized disk storage requirements and joining smaller lookup tables.
- It provides greater scalability in the interrelationship between dimension levels and components.
- No redundancy, so it is easier to maintain.

Disadvantage of Snowflake Schema

- The primary disadvantage of the snowflake schema is the additional maintenance efforts required due to the increasing number of lookup tables. It is also known as a multi fact star schema.
- There are more complex queries and hence, difficult to understand.
- More tables more join so more query execution time.

S.No.	Star Schema	Snowflake Schema
1.	In star schema, The fact tables and the dimension tables are contained.	While in snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained.
2.	Star schema is a top-down model.	While it is a bottom-up model.
3.	Star schema uses more space.	While it uses less space.
4.	It takes less time for the execution of queries.	While it takes more time than star schema for the execution of queries.
5.	In star schema, Normalization is not used.	While in this, Both normalization and denormalization are used.
6.	It's design is very simple.	While it's design is complex.
7.	The query complexity of star schema is low.	While the query complexity of snowflake schema is higher than star schema
8.	It's understanding is very simple.	While it's understanding is difficult.
9.	It has less number of foreign keys.	While it has more number of foreign keys.
10.	It has high data redundancy.	While it has low data redundancy.

MCQ

#Q. What is the primary characteristic of a "Star Schema" in data warehousing?

A. It consists of a central fact table connected to dimension tables directly.

B. It normalizes data for efficient storage.

C. It uses a tree-like structure for organizing data.

D. It is designed for real-time data processing.

MCQ

#Q. In a Star Schema, what is the role of the central "Fact Table"?

A. Stores metadata information

B. Contains primary keys of dimension tables

C. Holds quantitative data for analysis

D. Manages access control for the database

MCQ

#Q. Which of the following best describes a "Snowflake Schema"?

A. It consists of a central fact table connected to dimension tables directly.

B. It normalizes dimension tables by breaking them into sub-dimensions.

C. It uses a hierarchical structure for organizing data.

D. It is designed for efficient storage of transactional data.

MCQ

#Q. What is a potential disadvantage of the Snowflake Schema compared to the Star Schema?

A. Improved query performance

B. Increased complexity in query formulation

C. Easier to maintain and update

D. Higher storage efficiency

MCQ

#Q. In a Snowflake Schema, how are dimensions typically represented?

A. In a star-like structure with a central fact table

B. In a normalized form with multiple related tables

C. In a denormalized form for efficient querying

D. In a tabular structure with no relationships

MCQ

#Q. Which schema is more suitable for scenarios where storage efficiency is a critical concern?

A Star Schema

B Snowflake Schema

C Hybrid Schema

D Constellation Schema

Storage Efficiency in Snowflake Schema:

In a Snowflake Schema, storage efficiency represents the reduction in storage achieved through normalization. The formula for storage efficiency at each level of normalization is:

Storage Efficiency = 1 − Normalization Factor Storage

where the Normalization Factor is the fraction of records retained after normalization.

Storage Efficiency in StarSchema:

In a Star Schema, the central concept is the fact table surrounded by dimension tables. The storage efficiency is determined by the denormalization process, which reduces the need for joins during queries.

The storage efficiency formula for a Star Schema is:

$$\text{Storage Efficiency} = \frac{\text{Size of De-Normalized Schema}}{\text{Size of Normalized Schema}} \times (100)$$

2 MIN Summary

THANKYOU