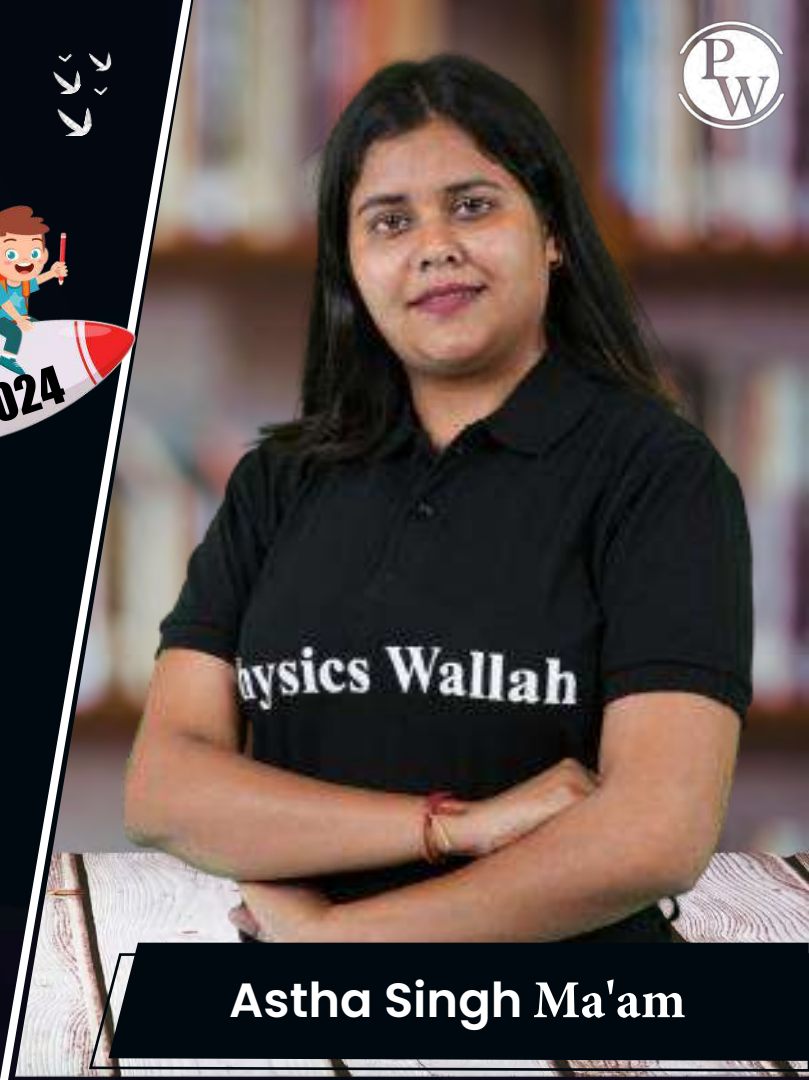
Data Science & Artificial Intelligence

Warehousing

Data Transformation

One Shot



Astha Singh Ma'am

Recap of Previous Lecture

What is Data Warehouse ?
How Does Data Warehouse Work?
Need for Data Warehouse
Characteristics
Function
Data Warehouse vs DBMS
Organizational trends Motivating data Warehouses
Comparison of Operational and Informational Systems
Data Warehouse Architecture
What is Data marts?
Data Warehouse Versus Data Mart
Derived data

Topics to be covered

- OLAP Technology
- Data Transformation
- Normalization (Standardization)
- Aggregation
- Discretization
- Sampling

Topic :OLAP IN DATAWAREHOUSE

Online Analytical Processing (OLAP) is a technology used in data warehousing to analyze large datasets and provide multidimensional views of data, enabling users to perform complex analytical queries and gain valuable insights.

Topic :Key Concepts in OLAP

1 Dimensions

Dimensions are the descriptive attributes of data, such as time, geography, or product categories, that provide context for analysis.

2 Measures

Measures are the numerical values or data that can be analyzed, aggregated, and displayed in OLAP cubes or reports.

3 Hierarchies

Hierarchies represent the relationships between different levels of data, allowing drill-down and roll-up operations for more granular or summarized analysis.

Topic : Types of OLAP Operations

Drill-down

Allows users to navigate from summarized data to more detailed levels, revealing underlying data points.

Roll-up

Enables users to aggregate data from detailed levels to higher-level summaries, facilitating trend analysis.

Slice and dice

Allows users to select specific dimensions and measures to focus the analysis on particular subsets of data.

Topic : Types of OLAP

- MOLAP:
  - Informational Retrival is fast
  - Incorporate use
  - Uses * separate data base
- ROLAP
  - Informational Retrival Slow
  - experiences users
  - Strong
  - Does not Require any other Space rather DW
- HOLAP

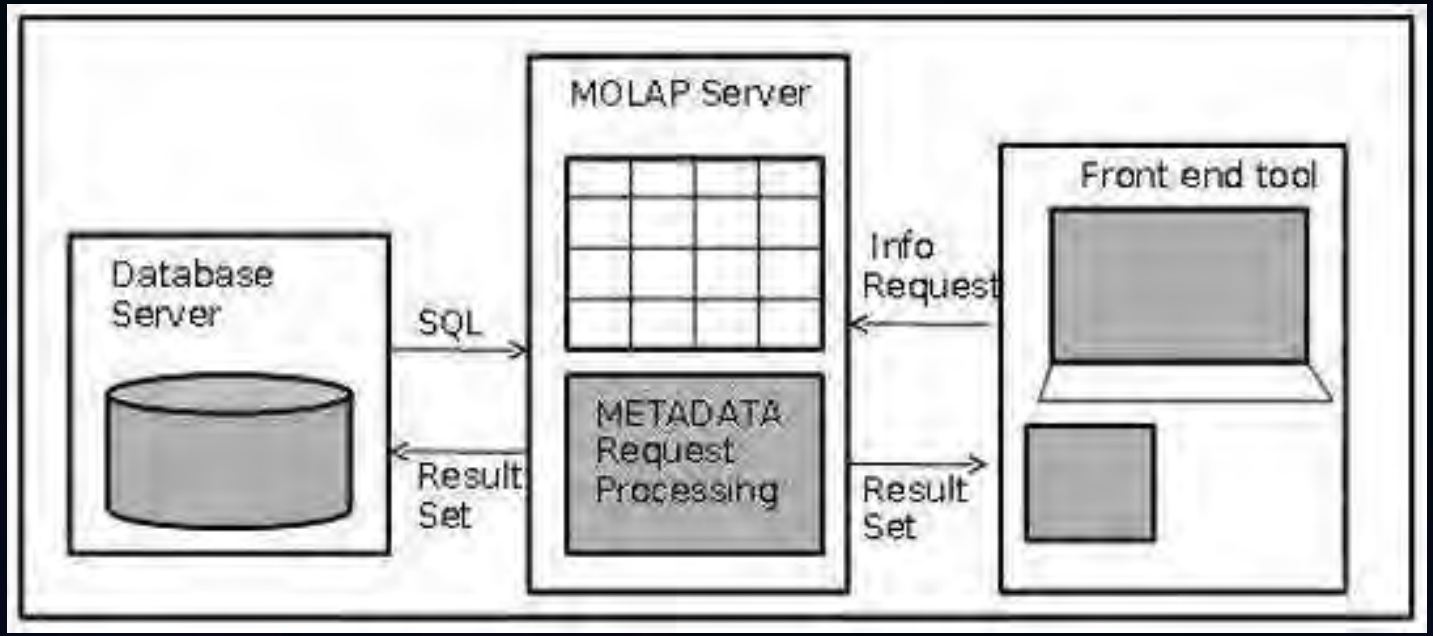
Topic :Multidimensional OLAP (MOLAP)

Definition and characteristics

MOLAP stores data in a specialized multidimensional cube, enabling fast analytical processing and efficient use of storage space.

Multidimensional OLAP (MOLAP) Architecture

* Data Server tool
* HOLAP | ROLAP | MOLAP
* Frontend tool


MOLAP Server
Front end tool
Database Server
SQL
Info Request
METADATA Request Processing
Result Set
Result Set

KEY POINT

MOLAP tools process information with consistent response time regardless of level of summarizing or calculations selected.

MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.

MOLAP tools need fastest possible performance.

MOLAP server adopts two level of storage representation to handle dense and sparse data sets.

Denser sub-cubes are identified and stored as array structure.

Sparse sub-cubes employ compression technology

Advantages

MOLAP offers superior performance, advanced calculations, and robust support for complex hierarchies and aggregations.

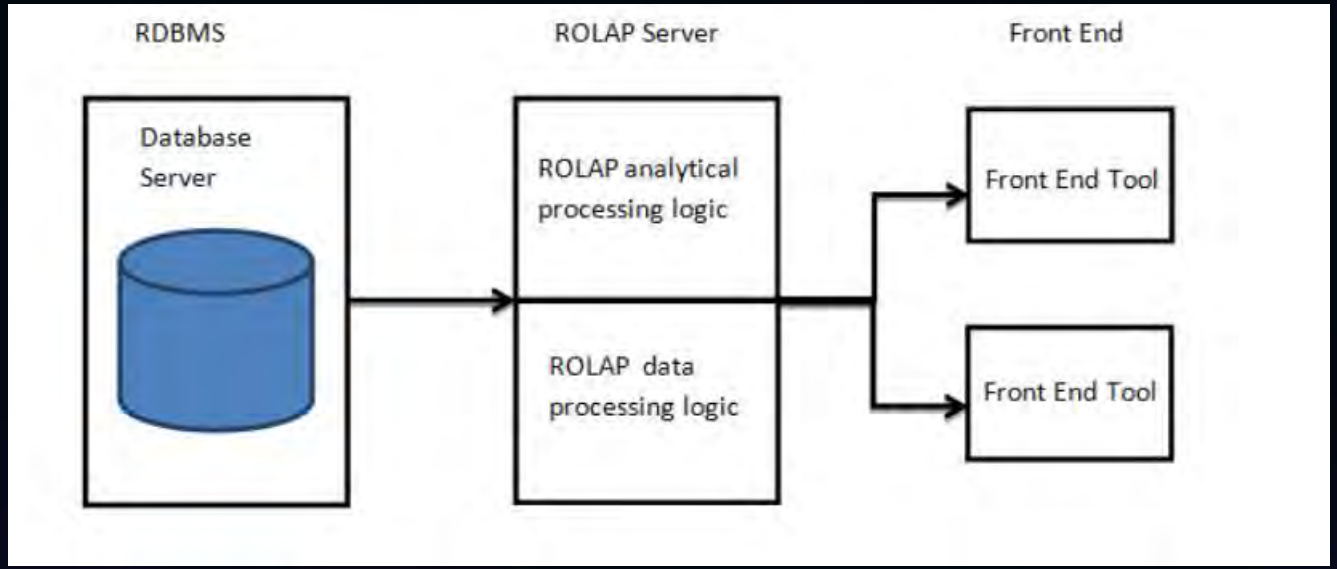
Disadvantages

MOLAP can be resource-intensive, requiring more storage space and often limited scalability compared to other OLAP types.

Relational OLAP (ROLAP)

Definition and characteristics

ROLAP stores data in a relational database, leveraging its querying capabilities and benefiting from existing infrastructure.

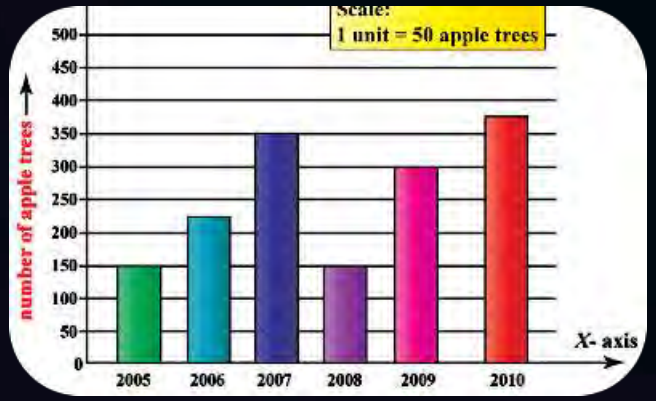

RDBMS
ROLAP Server
Front End
Database Server
ROLAP analytical processing logic
ROLAP data processing logic
Front End Tool
Front End Tool

Advantages

ROLAP offers scalability, flexibility, and compatibility with existing systems, making it easier to integrate with other applications.

Disadvantages

ROLAP may experience slower query response times for complex analyses and often requires manual query optimization.

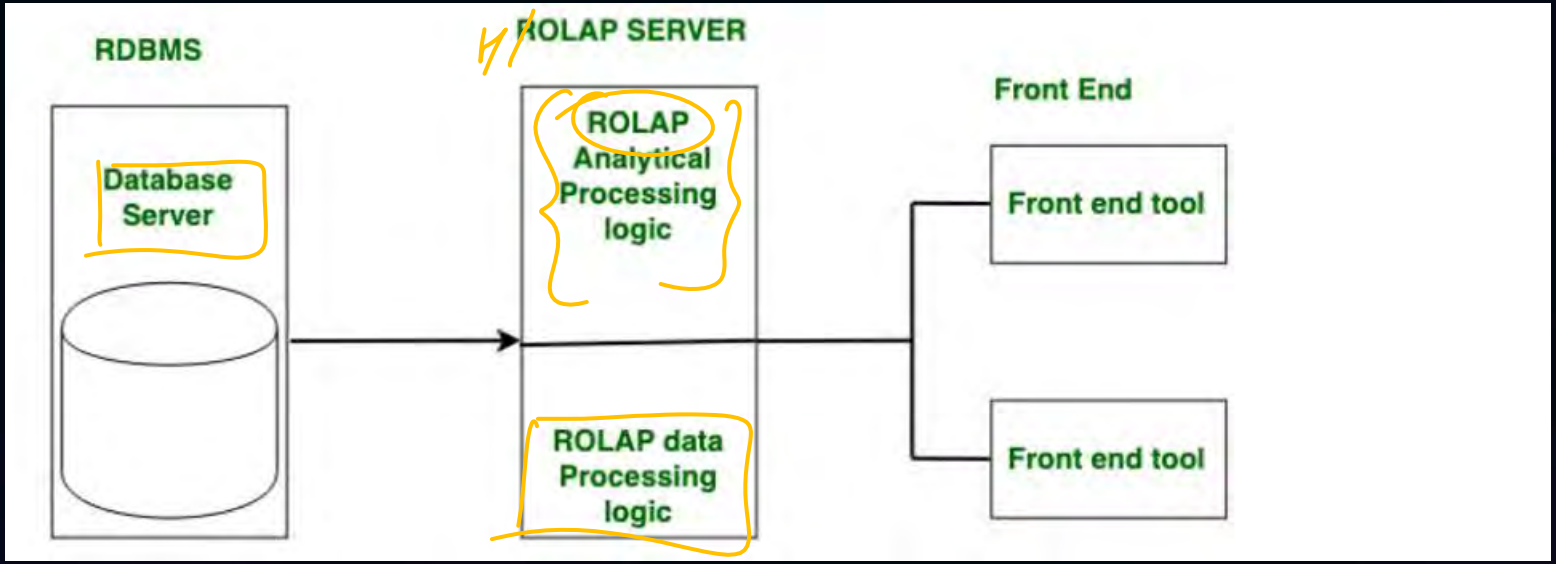


Slower

Hybrid OLAP (HOLAP)

Definition and characteristics

HOLAP combines the strengths of MOLAP and ROLAP, allowing for both multidimensional and relational data storage and analysis.


RDBMS
ROLAP SERVER
Front End
Database Server
ROLAP Analytical Processing logic
Front end tool
ROLAP data Processing logic
Front end tool

Advantages

HOLAP provides the flexibility to handle large datasets while leveraging the power of multidimensional analytical capabilities.

Disadvantages

HOLAP implementations may require complex integration and careful consideration of storage and performance trade-offs.

Comparison of OLAP Types

MOLAP	ROLAP	HOLAP
Specialized multidimensional cube	Relational database	Combination of MOLAP and ROLAP
Superior performance, advanced calculations	Scalability, flexibility, compatibility	Flexibility, multidimensional capabilities
More storage space, limited scalability	Slower query response times, manual optimization	Complex integration, storage and performance trade-offs

Benefits of OLAP in Data Warehousing

Fast and interactive analysis

OLAP enables users to explore data in real-time, empowering them to make informed decisions more efficiently.

Multi-dimensional insights

OLAP provides a holistic view of data, allowing users to analyze data from various perspectives and uncover trends and patterns.

Ad-hoc reporting

OLAP allows users to create customized reports on the fly, answering specific questions and addressing unique business requirements.

OLAP Implementation

1. **Data modeling**

   Design a flexible and scalable data model that aligns with business requirements and facilitates efficient analysis.

2. **Indexing and aggregation**

   Optimize query performance by implementing appropriate indexes and pre-aggregating data where applicable.

3. **Regular maintenance**

   Perform routine maintenance tasks like data cleansing, indexing, and backup to ensure the accuracy and reliability of OLAP data.

Challenges in Implementing OLAP

1. **Data integration**

   Integrating data from different sources and ensuring consistency can be complex and time-consuming.

2. **Performance optimization**

   OLAP processing can be resource-intensive, requiring efficient indexing, caching, and query optimization techniques.

3. **User training**

   Effectively utilizing OLAP tools and understanding its capabilities may require training and familiarization.

Key Point

OLAP is a powerful tool for data analysis in data warehousing, providing interactive, multidimensional insights for informed decision-making. However, it requires careful planning, implementation, and ongoing maintenance to maximize its benefits and overcome associated challenges.

MCQ

In OLAP, what does a dimension represent?

A) Numeric data points
B) Business perspectives for analysis
C) Aggregated values
D) None of the above

Which OLAP type is based on the concept of storing data in cubes?

A) MOLAP
B) ROLAP — Tabular
C) HOLAP — Tab + Cube
D) DOLAP

What is the purpose of the "drill-down" operation in OLAP?

A) Moving from a lower level to a higher level of abstraction
B) Extracting a 2D subset of a cube
C) Performing aggregations
D) None of the above

In OLAP, what does a measure represent?

A) Business perspectives for analysis
B) Numeric data points
C) Aggregated values
D) None of the above

Which OLAP type combines features of both MOLAP and ROLAP?

A) MOLAP
B) ROLAP
C) HOLAP
D) DOLAP

MOLAP + ROLAP = HOLAP
Cube + Row = CR

What operation involves viewing a 3D subset of an OLAP cube?

1. A) Drilling down
2. B) Rolling up
3. C) Slicing
4. D) Dicing

Which OLAP type stores data in a relational database?

- A) MOLAP
- B) ROLAP
- C) HOLAP
- D) DOLAP

1. **What is a key characteristic of OLAP systems?**

- A) Designed for Online Transaction Processing (OLTP)
- B) Facilitates quick and interactive analysis of multidimensional data
- C) Primarily used for data entry and updating
- D) Suitable for small-scale data storage

Topic : Data Transformation

Data transformation is the process of converting raw data into a format that is suitable for analysis and modeling. It involves manipulating, cleaning, and organizing data to improve its quality and make it more meaningful.

Types of Data Transformation

Scaling Data

Adjusting data values to a specific range, preserving relative differences between them.

Normalizing Data

Rescaling data to have a mean of zero and a standard deviation of one.

Standardizing Data

Bringing data within a similar range by subtracting the mean and dividing by the standard deviation.

Encoding Categorical Data

Converting categorical variables into numerical representations for analysis.

Data Transformation Objectives

Cleaning Data:

- Handling Missing Values
- Outlier Detection and Treatment

Integration:

- Combining Datasets
- Resolving Redundancy

Normalization:

- Min-Max Scaling
- Z-Score Normalization
- Decimal Scaling Normalization

Techniques and Methods

Handling Missing Values:

- Imputation Techniques
- Impact on Data Analysis
- Outlier Detection:
- Visualization Approaches
- Statistical Methods

Data Integration:

- Concatenation and Merging
- Dealing with Different Granularities

Normalization Techniques:

- Min-Max Scaling Example
- Z-Score Normalization Example

Handling Missing Values

1. **Imputation**

   Filling in missing values using statistical methods like mean, median, or regression.

2. **Deletion**

   Removing rows or columns with missing values if they don't significantly impact analysis.

3. **Flagging**

   Creating a new variable to indicate the presence of missing data for analysis purposes.

Handling Outliers

1. **Detection**

   Identifying outliers using statistical methods or visualization techniques.

2. **Transformation**

   Adjusting the value of outliers to minimize their impact on analysis.

3. **Removal**

   Removing outliers from the dataset if they are extreme and affect the analysis significantly.

Feature Selection and Extraction

Feature Selection

Selecting relevant features that have a significant impact on the outcome of the analysis.

Outcome

Feature Extraction

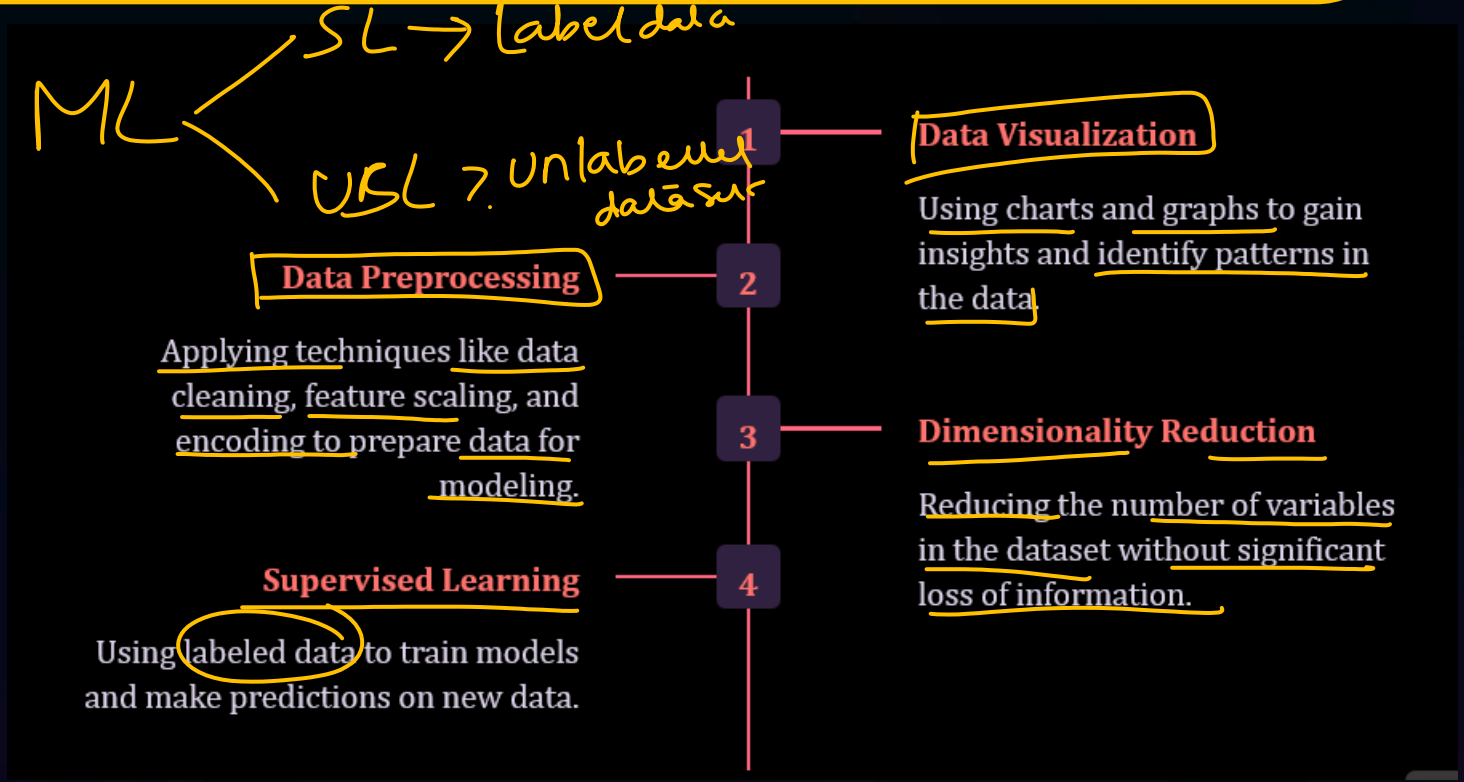
Creating new features by combining or transforming existing features to capture additional information.

* STRESS, Anxiety
Depression

Dimensionality Reduction

- Need for Dimensionality Reduction
- Principal Component Analysis (PCA)
  - Concept and Application
- Singular Value Decomposition (SVD)
  - Benefits and Use Cases

Effective Techniques for Data Transformation



Benefits of Data Transformation in Data Science

Data Science

Improved Accuracy

By transforming data, models can make more accurate predictions and classifications.

Enhanced Interpretability

Data transformation simplifies complex relationships, making it easier to understand and interpret the results.

Better Model Performance

Transformed data can lead to improved model performance and better decision-making.

Implementation Challenges of Data Transformation

1. **Data Quality**

   Poor-quality data can lead to inaccurate transformations and biased results.

2. **Data Compatibility**

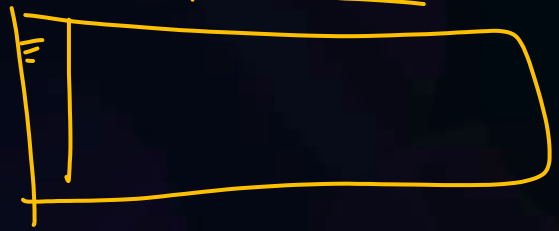
   Data from different sources may have varying formats and structures, requiring careful integration.

3. **Computational Complexity**

   Data transformation techniques can be computationally expensive, especially for large datasets.

Data Transformation in Python

- Libraries for Data Transformation:
  - Pandas for Cleaning and Preprocessing
  - Scikit-learn for Normalization and Scaling
- Code Snippets and Examples



Summary and Key Points

Data transformation is a critical step in data science.

It involves various techniques like scaling, normalization, encoding, handling missing values, outliers, and feature selection.

Effective data transformation techniques improve accuracy and model performance.

They enhance interpretability and lead to better decision-making.

However, challenges like data quality, compatibility, and computational complexity need to be addressed during implementation.

MCQ

Sample

What is the primary objective of data transformation in data science?

- A) Increase data complexity
- B) Enhance model performance
- C) Introduce noise to the data
- D) Decrease data quality

Which of the following is a common technique for handling missing values during data transformation?

- A) Ignoring missing values
- B) Imputation
- C) Increasing feature dimension
- D) Outlier detection

What does Min-Max scaling aim to achieve in data normalization?

- A) Standardizing data with mean 0 and variance 1
- B) Scaling data to a specific range, typically [0, 1]
- C) Identifying outliers in the data
- D) Creating new features from existing ones

Feature engineering in data transformation involves:

- A) Reducing the number of features
- B) Enhancing model performance by creating new features
- C) Removing outliers from the dataset
- D) Normalizing the target variable

1. In data transformation, what does the process of dimensionality reduction aim to achieve?
    1. A) Increasing the number of dimensions in the data
    2. B) Enhancing interpretability of the data
    3. C) Reducing the number of features while preserving information
    4. D) Standardizing feature values
2. Which Python library is commonly used for data transformation and preprocessing in data science?
    1. A) TensorFlow
    2. B) Keras
    3. C) Scikit-learn
    4. D) PyTorch

1. What is a key benefit of principal component analysis (PCA) in dimensionality reduction?
    1. A) Creating new features
    2. B) Reducing feature redundancy
    3. C) Introducing noise to the data
    4. D) Ignoring missing values
2. Which of the following is an example of a normalization technique used in data transformation?
    1. A) Feature engineering
    2. B) Imputation
    3. C) Principal component analysis (PCA) → Dimensionality Reduction
    4. D) Z-score normalization → Range [0, 1]

Data Normalization

process of transforming data to a common scale, ensuring consistency and eliminating any bias caused by varying data ranges.

Types of Data Normalization

Range [0,1]

- Min-Max Normalization
- Z-score Normalization
- Decimal Scaling Normalization

Types of Data Normalization

1. **Min-Max Normalization**

   Standardizes data within a specific range, typically between 0 and 1.

2. **Z-Score Normalization**

   Transforms data to have a mean of 0 and a standard deviation of 1, enabling easy comparison across variables.

3. **Decimal Scaling Normalization**

   Shifts the decimal point of the values, making the largest value less than or equal to 1.

Min-Max Normalization

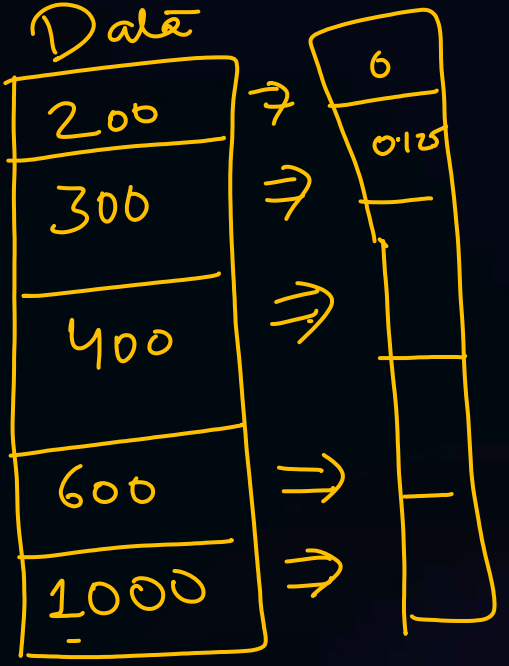
Q: Min = 200 Max = 1000

$$V = \frac{X - Min}{Max - Min}$$

* X = is a data point
  Min / Max = Value which identified in the table

I^st^ $V = \frac{200 - 200}{1000 - 200} = 0$

II^nd^ $V = \frac{300 - 200}{1000 - 200} = \frac{100}{800} = 0.125$



$$V = \frac{u - Min}{Max - Min}$$

III rd step $V = \frac{400 - 200}{1000 - 200} = ?\ 0.5$

$$V = \frac{1000 - 200}{1000 - 200} = ?\ 0$$

2 Min Summary

THANK - YOU