UNIT 1s

Questions:

1. Data types nominal ordinal / temporal or special data or compare data / info
2. Ide's feature of ide used for prog in python thoda case study types write about adv and disadvantage of ide and eda(exprolatory data analysis - handle missing values and data in ppt has given an example)
3. Decimal ka spacing two techniques or short note
4. Data mining and data warehouse application in respective fields or data science same kind of answer .
5. In data warehouse compare data warehouse and DBMS , architecture of data ware house adv and disadvantage create short note
6. Prac 1 basics of numpy and pandas so simple progs might come or he will give a SS of code spot errors and justify
7. Eg how np. Eye and np. I is diff

Purpose:

- Both np.eye and np.identity are used to create identity matrices in NumPy. An identity matrix is a square matrix with 1s on the main diagonal and 0s elsewhere.

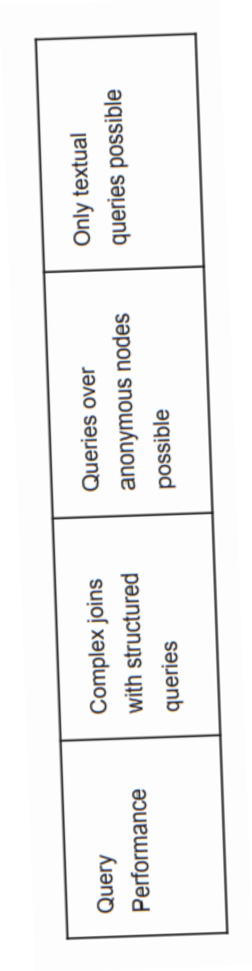
Key Differences:

1. Shape Flexibility:
   - np.eye(N): This function creates an N x N identity matrix. You can specify the size of the matrix by providing a single argument N (number of rows and columns).
   - np.identity(N): This function also creates an N x N identity matrix. However, it doesn't offer flexibility in the shape. You can only create square matrices.
2. Diagonal Offset (Optional Argument):
   - np.eye(N, M, k=0): This allows you to optionally specify the offset of the diagonal. By default (k=0), it creates an identity matrix with 1s on the main diagonal. A positive k value shifts the ones diagonally upwards by k positions, and a negative k value shifts them downwards.
   - np.identity:** This function only creates matrices with ones on the main diagonal (k=0). It doesn't allow specifying an offset value.

DATA

- data refers to raw, unprocessed information that can be collected, analysed, and used to extract knowledge, identify patterns, and make predictions.
- It's the foundation for all data science tasks.
- Data can come in various forms, each with its own characteristics and processing requirements.
- TYPES:
    - Structured Data
        - Structured data is highly organised and follows a predefined format.
        - It typically resides in relational databases or spreadsheets where each piece of data has a specific meaning and location.
        - It's easy for computers to process and analyse due to its well-definedstructure.
        - Examples:
            - Customer information in a database (customer ID, name, address, purchase history)
            - Financial data in spreadsheets (stock prices, transaction details)
            - Sensor data with timestamps and values
    - Unstructured Data
        - Unstructured data is information that lacks a consistent format or organisation.
        - It often contains text, images, audio, video, emails, social media posts, and other complex data types.
        - While valuable for insights, unstructured data requires additional processing and techniques for analysis.
        - Examples:
            - Text documents, emails, social media posts
            - Images, videos, audio files
            - Website content, sensor logs, network traffic dat

Property	Structured Data	Semi-structured Data	Unstructured Data
Technology	Relational database tables	XML, RDF (Resource Description Framework)	Character & binary data
Transaction Management	Matured, various concurrency techniques	Adapted from DBMS, less mature	No transaction management or concurrency
Version Management	Versioning over tuples, rows, tables	Versioning over tuples or graphs (possible)	Versioned as a whole
Flexibility	Schema-dependent, less flexible	More flexible than structured, less than unstructured	Most flexible, no schema
Scalability	Difficult to scale database schema	Simpler scaling than structured	Most scalable
Robustness	Very robust	Newer technology, less widespread	-
Query Performance	Complex joins with structured queries	Queries over anonymous nodes possible	Only textual queries possible



Data vs information

Data	Information
Variables for developing ideas/conclusions	Meaningful data
Text and numerical values	Refined form of actual data
Independent (doesn't rely on information)	Relies on data
Measured in bits and bytes	Measured in meaningful units (time, quantity)
Can be structured (tabular, graph, data tree)	Can be structured (language, ideas, thoughts)
No inherent meaning	Meaningful context
Low-level knowledge	Second level of knowledge
Indirectly helps in decision making	Directly helps in decision making
Collection of facts with no meaning	Puts facts into context
Example: Student test score	Example: Average score of class

Feature	Qualitative Data	Quantitative Data
Definition	Descriptive data; focuses on qualities or characteristics	Numerical data; focuses on quantities or measurements
Collection Methods	Interviews, Focus Groups, Observations, Open-ended Surveys	Surveys, Experiments, Sensors, Existing Records
Data Format	Textual; narratives, quotes, images, audio/video recordings	Numerical; numbers, statistics, ratings on scales
Analysis Methods	Thematic analysis, coding, identifying patterns	Statistical analysis, calculations, creating charts and graphs
Purpose	Understand experiences, motivations, opinions	Measure, quantify, compare, predict
Examples	Customer feedback in text, interview transcripts	Sales figures, website traffic data, survey responses with numerical scales

Nominal vs. Ordinal Data

Feature	Nominal Data	Ordinal Data
Definition	Categorical data with no inherent order or ranking	Categorical data with a specific order or ranking
Examples	Hair color (blonde, brown, black), Blood type (A, B, AB, O), Country of origin	Customer satisfaction (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied), Movie ratings (1 star, 2 stars, 3 stars, 4 stars, 5 stars)
Mathematical Operations	Limited; cannot perform calculations like addition, subtraction, or averaging	Possible; can determine "greater than," "less than," or "equal to," but not the difference between categories
Level of Measurement	Nominal (lowest level)	Ordinal (second level)
Analysis Methods	Frequency tables, Chi-square tests	Frequency tables, Chi-square tests, rankings, order statistics (median)
Insights	Identify groups, compare distributions	Identify groups, compare distributions, understand order or hierarchy

Discrete vs. Continuous Data

Feature	Discrete Data	Continuous Data
Definition	Data that can only take on whole, distinct values	Data that can take on any value within a specific range
Visualisation	Represented well with bar graphs or histograms (with bars)	Represented well with line graphs or histograms (without bars)
Examples	Number of people in a room (1, 2, 3, ...), Shoe size (5, 6, 7, ...), Exam scores (assuming whole numbers)	Temperature, Weight, Time, Distance
Measurement	Counted	Measured
Values Between Categories	Gaps or jumps exist between possible values	No gaps or jumps between possible values (theoretically infinite possibilities within the range)
Analysis Methods	Frequency tables, Chi-square tests, mode (most frequent value)	Statistical analysis (mean, median, standard deviation), regression analysis
Real-world Examples	Inventory counts, Number of website visitors per day, Number of correct answers on a multiple-choice test	Height, Weight, Speed, Temperature at a specific time

Further Classification of data

1. Demographic Data:

- Description: This data describes the characteristics of a population or individual, like age, gender, income, education level, family size, occupation, etc.
- Source: Collected through surveys, census data, customer profiles, loyalty programs.
- Use Cases: Segmenting customers, targeted marketing campaigns, understanding customer needs and preferences.

2. Transactional Data:

- Description: This data captures details about customer transactions, including products purchased, date and time of purchase, amount spent, payment method, etc.
- Source: Point-of-sale systems, e-commerce platforms, financial records.
- Use Cases: Analysing purchase patterns, identifying trends, recommending products, optimising pricing and promotions.

3. Web Behavior Data:

- Description: This data tracks how customers interact with a website or app, including pages visited, time spent on each page, products viewed, search queries, etc.
- Source: Web analytics tools (e.g., Google Analytics), clickstream data, user session recordings.
- Use Cases: Personalising user experience, website optimization, understanding customer journeys, identifying popular products.

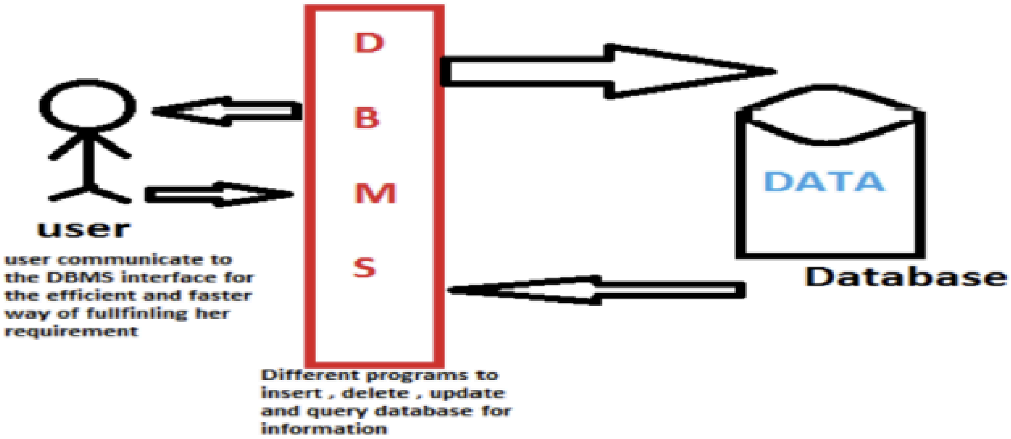
4. Data from Customer-Created Texts:

- Description: This data encompasses text reviews, feedback, comments, social media posts, and other forms of written communication from customers.
- Source: Online reviews, social media platforms, customer service interactions, chat logs.
- Use Cases: Sentiment analysis (understanding customer opinions), improving product/service offerings, identifying customer pain points, building brand loyalty.

DBMS

A Database Management System (DBMS) is software designed to efficiently store, retrieve, define, and manage data in a structured way. It acts as an intermediary between users and the database, providing a controlled and organized environment for data access and manipulation.

DBMS... way of data extraction



user communicate to the DBMS interface for the efficient and faster way of fullfinling her requirement

Different programs to insert , delete , update and query database for information

Problems faced by current DBMS

- large quantities of data is generated /processed.
    - data may get doubled in every say 3 months.
- Seeking knowledge from this massive data is most required.
    - Fast developing in computer science and engineering techniques generates new demands.
    - To fulfil those demands we require to analyse the data
- Data Rich , Information Poor..
    - Raw data by itself does not provide much information.
- In today's life we require only significant data from which we can judge the customer’s likings and strategies

DATA MINING

- Data mining is the multifaceted process of extracting valuable knowledge, patterns, and trends from massive datasets.
- APPLICATIONS
  a. Customer Relationship Management (CRM): Understanding customer behavior, preferences, and churn factors to personalize marketing campaigns, improve customer retention, and optimize product offerings.
  b. Fraud Detection: Identifying anomalies in transactions or credit card usage patterns to prevent fraudulent activity and protect financial systems.
  c. Recommendation Systems: Recommending products, services, or content to users based on their past purchases, browsing history, and similar user profiles.
  d. Market Research: Analyzing market trends, competitor strategies, and customer demographics to gain a competitive edge.
  e. Scientific Discovery: Uncovering hidden relationships and patterns in scientific data to advance research in areas like astronomy, medicine, and biology.
- STEPS:
  a. Data Integration:
     - Goal: Combine data from disparate sources (databases, spreadsheets, social media feeds) into a unified, consistent format.
     - Techniques: Data warehousing, entity resolution, data cleansing.

  b. Data selection
     - Goal: Select a relevant subset of data that aligns with the specific business problem or research question being addressed. Focusing on a targeted dataset improves processing efficiency and reduces computational costs.
     - Techniques: Statistical sampling, domain knowledge-based selection.

c. Data Cleaning:
    - Goal: Address data quality issues such as missing values, inconsistencies, and outliers. This step ensures the accuracy and reliability of the results.
    - Techniques: Imputation techniques for missing values, outlier detection and correction, data standardization.

d. Data Transformation:
    - Goal: Prepare the data for mining algorithms. This might involve feature scaling or engineering new features from existing ones to enhance the model's learning ability.
    - Techniques: normalization, feature selection, dimensionality reduction.

e. Data Mining:
    - Goal: Apply data mining algorithms to uncover patterns and relationships within the data.
    - Common algorithms include classification (categorizing data points), regression (predicting continuous values), clustering (grouping similar data points), and association rule learning (identifying frequently occurring itemsets).
    - Techniques: Classification trees, decision trees, support vector machines, neural networks, k-means clustering, Apriori algorithm.

f. Pattern Evaluation:
    - Goal: Assess the discovered patterns for validity, significance, and actionable insights. This ensures that the patterns are not merely coincidental and can be used to make informed decisions.
    - Techniques: Statistical tests, visualization techniques, domain expert revie

g. Knowledge Representation and Presentation:
    - Goal: Communicate the extracted knowledge effectively to stakeholders in a clear, concise, and actionable format. This might involve creating data visualizations, reports, or interactive dashboards.
    - Techniques: Data visualization tools (e.g., bar charts, scatter plots), reports, dashboards, presentations.

- EXAMPLES OF DATA MINING
  a. Retail: Analyzing customer purchase history to identify buying patterns and recommend complementary products for upselling.
  b. Finance: Detecting fraudulent credit card transactions by spotting anomalies in spending patterns.
  c. Healthcare: Predicting patient readmission risks using health records data to provide preventive care and reduce healthcare costs.
  d. Telecommunications: Identifying churn factors (reasons why customers leave) to design customer retention strategies and improve customer satisfaction.

BIG DATA

Big data refers to massive, complex datasets that traditional data processing tools struggle to handle. It's characterized by five key dimensions

- Volume: The sheer amount of data generated from various sources, often growing exponentially.

- Velocity: The speed at which data is created and needs to be processed, analyzed,or acted upon. Real-time or near-real-time processing is crucial.

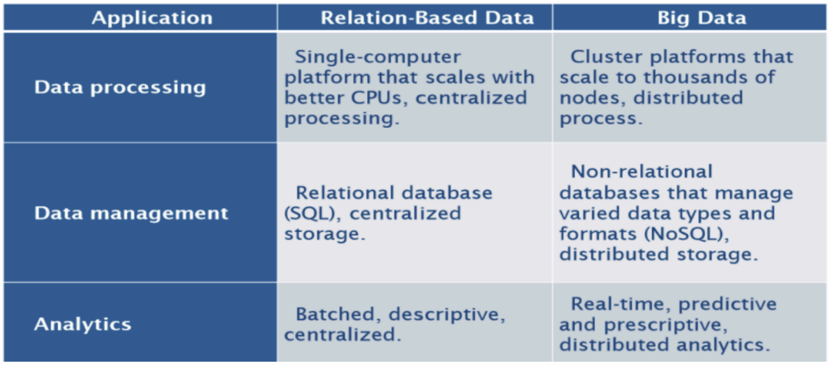
- Variety: The diverse nature of data, encompassing structured (databases), semi-structured (logs, emails), and unstructured (social media, text) formats.

- Value: Extracting meaningful insights and patterns from the data to drive informeddecision-making, improve efficiency, and create new opportunities.

- Veracity: Ensuring the accuracy, reliability, and completeness of the data to avoid misleading results.

Sources of data in Big Data

- Social Media
- Online cloud platforms
- Internet of things
- Online Web pages
- Search Engine Data
- Stock Exchange Data



Application	Relation-Based Data	Big Data
Data processing	Single-computer platform that scales with better CPUs, centralized processing.	Cluster platforms that scale to thousands of nodes, distributed process.
Data management	Relational database (SQL), centralized storage.	Non-relational databases that manage varied data types and formats (NoSQL), distributed storage.
Analytics	Batched, descriptive, centralized.	Real-time, predictive and prescriptive, distributed analytics.

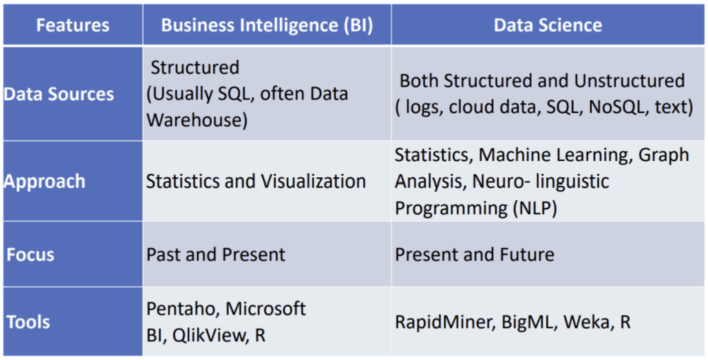
Big Data vs. Data Science

Feature	Big Data	Data Science
Focus	Managing and processing large datasets	Extracting knowledge and insights from data
Data Size	Primarily deals with **very large** datasets	Can handle **various data sizes**, including big data
Skills	Data engineering, distributed systems, storage	Statistics, machine learning, programming
Tools	Hadoop, Spark, NoSQL databases	Python, R, SQL, machine learning libraries
Goal	Make massive datasets usable and accessible	Uncover patterns, trends, and actionable insights
Output	Cleaned, organized, and accessible data	Models, predictions, visualizations, reports
Applications	Real-time analytics, fraud detection, log analysis	Customer segmentation, product recommendation, risk assessment

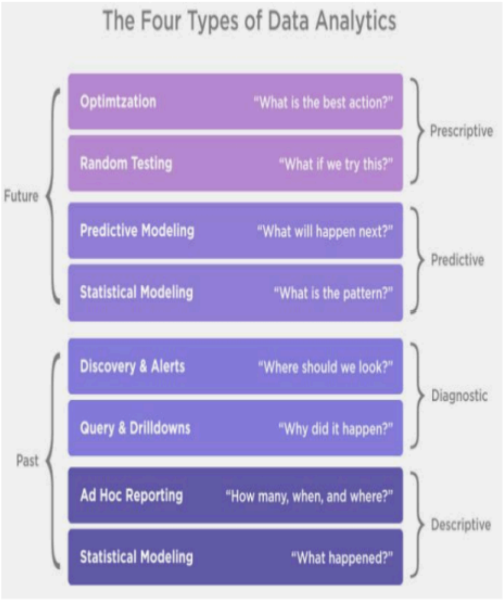
DATA ANALYSIS

Data analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

1. **Descriptive Analytics: Understanding What Happened**
    a. Focus: Summarises past data to provide a clear picture of what has transpired.
    b. Techniques: Uses basic statistics (averages, medians, frequencies), data visualisation (charts, graphs).
    c. Example: A retail store analyses its sales data for the past year, identifying top-selling products, seasonal trends, and customer demographics.
2. **Diagnostic Analytics: Uncovering the "Why"**
    a. Focus: Drills down into past data to pinpoint the root causes of events or trends.
    b. Techniques: Data mining, correlation analysis, anomaly detection.
    c. Example: A hospital examines patient readmission rates, identifying factors like specific diagnoses or lack of follow-up care that contribute to readmissions.
3. **Predictive Analytics: Forecasting the Future**
    a. Focus: Leverages historical data and statistical models to predict future outcomes or trends.
    b. Techniques: Regression analysis, machine learning algorithms, forecasting models.
    c. Example: A bank assesses a loan applicant's creditworthiness using a model trained on past loan repayment data, predicting the probability of on-time payments for the new applicant.
4. **Prescriptive Analytics: Taking Action**
    a. Focus: Goes beyond prediction to recommend optimal courses of action based on anticipated outcomes.
    b. Techniques: Optimization algorithms, simulation modelling, decision rules.
    c. Example: A ride-sharing service uses real-time traffic data and predictive analytics to suggest the most efficient route for drivers, optimising passenger wait times and driver earnings.



Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R



Analytics: Understanding What Happened

LIFECYCLE OF DATA SCIENCE

● Phase 1—Discovery various specifications, requirements, priorities , budget, business problem and hypothesis.

● Phase 2—Data preparation loading, cleaning, transformation, and visualization

● Phase 3—Model planning methods and techniques to draw relationship between variables

● Phase 4—Model building develop datasets for training and testing purposes; classify them

● Phase 5—Operationalize reports, briefings, code and technical documents

● Phase 6—Communicate results you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1

introduction to high level programming language + Integrated Development environment

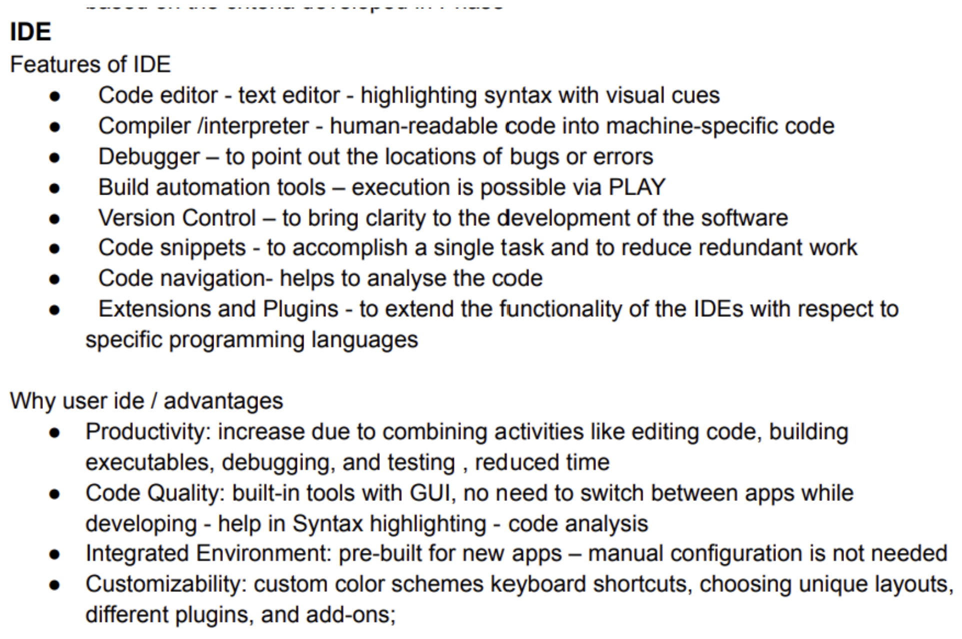
High-level programming languages are designed to be more human-readable and easier to write than machine code (the language computers understand directly). They use instructions similar to natural languages with a defined syntax (set of rules for structuring code).

Benefits:

Abstraction: High-level languages hide the low-level details of computer hardware, allowing programmers to focus on the problem they're solving rather than the specifics of the machine.

Portability: Code written in a high-level language can often be run on different computer systems with minimal modifications, unlike machine code which is specific to a particular architecture.

Maintainability: High-level languages typically promote code readability and maintainability due to their use of clear syntax and constructs.



based on the criteria developed in Phase

IDE

Features of IDE

- Code editor - text editor - highlighting syntax with visual cues
- Compiler /interpreter - human-readable code into machine-specific code
- Debugger – to point out the locations of bugs or errors
- Build automation tools – execution is possible via PLAY
- Version Control – to bring clarity to the development of the software
- Code snippets - to accomplish a single task and to reduce redundant work
- Code navigation- helps to analyse the code
- Extensions and Plugins - to extend the functionality of the IDEs with respect to specific programming languages

Why user ide / advantages

- Productivity: increase due to combining activities like editing code, building executables, debugging, and testing , reduced time
- Code Quality: built-in tools with GUI, no need to switch between apps while developing - help in Syntax highlighting - code analysis
- Integrated Environment: pre-built for new apps – manual configuration is not needed
- Customizability: custom color schemes keyboard shortcuts, choosing unique layouts, different plugins, and add-ons;

Exploratory Data Analysis (EDA) Techniques

It involves investigating, summarizing, and visualizing data to understand its characteristics, identify patterns, and uncover potential relationships between variables.

- This process helps data scientists gain insights into the data before diving into more complex modeling or analysis techniques.

Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis process. It involves investigating, summarizing, and visualizing your data to understand its characteristics, patterns, and relationships. Here are some key EDA techniques:

1. Summarizing Data:
   - Central Tendency: Measures like mean, median, and mode provide a general idea of the "center" of your data distribution (i.e., where most of the data points lie).
   - Dispersion: Measures like variance and standard deviation quantify how spread out the data is around the central tendency.
   - Quantiles: Dividing the data into equal-sized portions (quartiles) can reveal potential skewness or outliers.
2. Visualization Techniques:
   - Histograms: Bar charts that display the frequency distribution of a continuous variable, illustrating the shape of the data (e.g., normal, skewed).
   - Scatter Plots: Visualize the relationship between two continuous variables, showing potential correlations or trends.
   - Box Plots: Summarize the distribution of a continuous variable with quartiles (boxes) and whiskers (extremes) to identify outliers.
   - Bar Charts: Suitable for comparing categorical variables, showing the frequency of each category.
   - Pie Charts: Another way to visualize categorical data proportions, but can be less effective for many categories.
3. Identifying Patterns and Trends:
   - Correlation: Measures the strength and direction of the linear relationship between two variables (positive, negative, or no correlation).
   - Outliers: Identifying and investigating data points that fall significantly outside the overall distribution. They might indicate errors or interesting insights.

- Grouping: Splitting data into categories (e.g., by demographic groups) to compare distributions or relationships within those groups.

4. Data Cleaning Techniques (often done alongside EDA):

- Handling Missing Values: Techniques like imputation (filling in missing values based on other data) or deletion may be necessary.
- Dealing with Outliers: Investigate outliers – they could be errors or valuable insights. Decide whether to keep, remove, or transform them.
- Encoding Categorical Variables: Convert non-numeric categories into a format suitable for analysis (e.g., one-hot encoding).

Additional Tips for Effective EDA:

- Start with understanding the context and goals of your analysis. What are you trying to learn from the data?
- Choose appropriate visualization techniques based on your data types and questions.
- Don't be afraid to experiment with different visualizations and see what reveals the most insights.
- Document your findings and observations throughout the EDA process.

By effectively using EDA techniques, you can gain a deeper understanding of your data, uncover hidden patterns and relationships, and prepare your data for further analysis tasks like modeling or hypothesis testing.

Data Cleaning

- Data cleansing
- Process of removing inaccurate records
- Removing dirty records
- Non-relevant part of data
- Unfinished data
- Useless data
- Errors may occur while
- collecting data
- Transforming data
- Analyzing
- Submission of draft

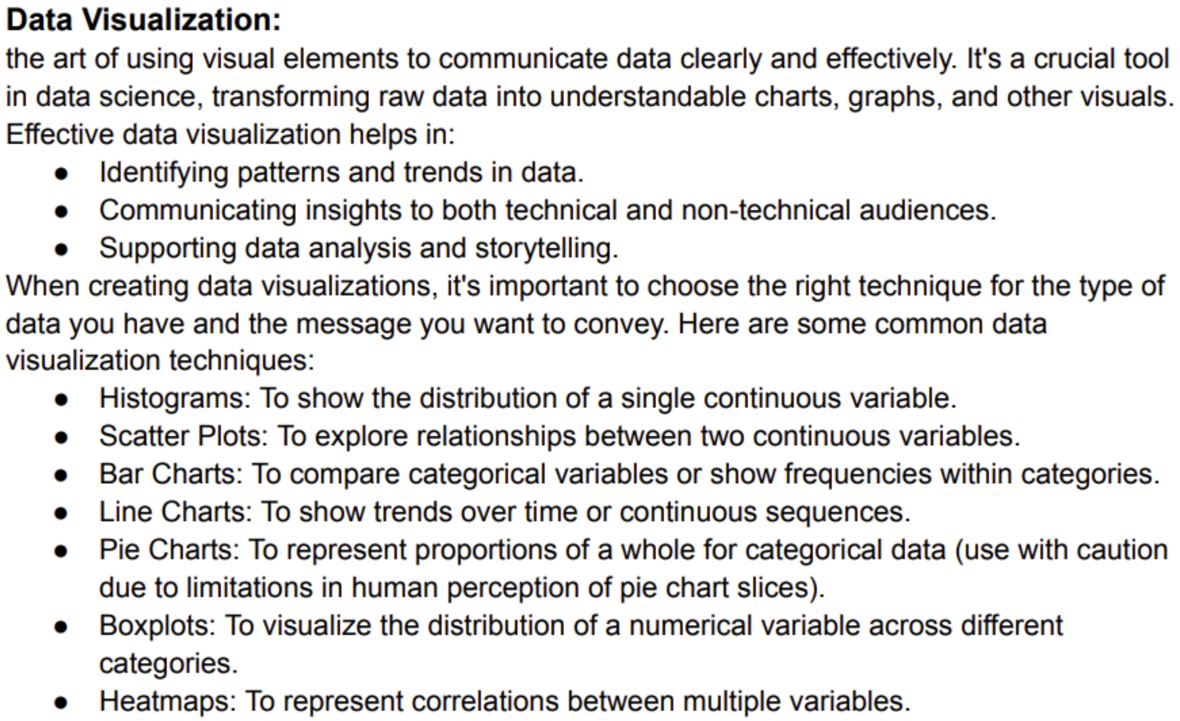
data scientists spend 80% of their time cleaning and manipulating data and only 20% of their
time actually analyzing it.

Techniques for handling these issues include:

- Missing Values
- Inconsistencies
- Outliers

Techniques :

- Descriptive Statistics:Calculating summary statistics for numerical variables
- Central Tendency: Mean, median, mode to understand the "center" of the data distribution.
- Spread: Standard deviation, variance, range to quantify how spread out the data is.
- Distribution: Measures like skewness and kurtosis to assess if the data is symmetrical or skewed (lopsided) and how peaked or flat the distribution is compared to a normal distribution.
- Calculating frequency distributions for categorical variables to understand the distribution of categories and identify potential biases.



Data Visualization:

the art of using visual elements to communicate data clearly and effectively. It's a crucial tool in data science, transforming raw data into understandable charts, graphs, and other visuals. Effective data visualization helps in:

- Identifying patterns and trends in data.
- Communicating insights to both technical and non-technical audiences.
- Supporting data analysis and storytelling.

When creating data visualizations, it's important to choose the right technique for the type of data you have and the message you want to convey. Here are some common data visualization techniques:

- Histograms: To show the distribution of a single continuous variable.
- Scatter Plots: To explore relationships between two continuous variables.
- Bar Charts: To compare categorical variables or show frequencies within categories.
- Line Charts: To show trends over time or continuous sequences.
- Pie Charts: To represent proportions of a whole for categorical data (use with caution due to limitations in human perception of pie chart slices).
- Boxplots: To visualize the distribution of a numerical variable across different categories.
- Heatmaps: To represent correlations between multiple variables.

Data Sources

1. Relational Databases:

Structured data storage organized in tables with rows and columns. Each table represents a specific entity and rows represent individual records within that entity. Columns represent attributes of those records

Access: Structured Query Language (SQL) is used to retrieve, manipulate, and manage data stored in relational databases.

Advantages:

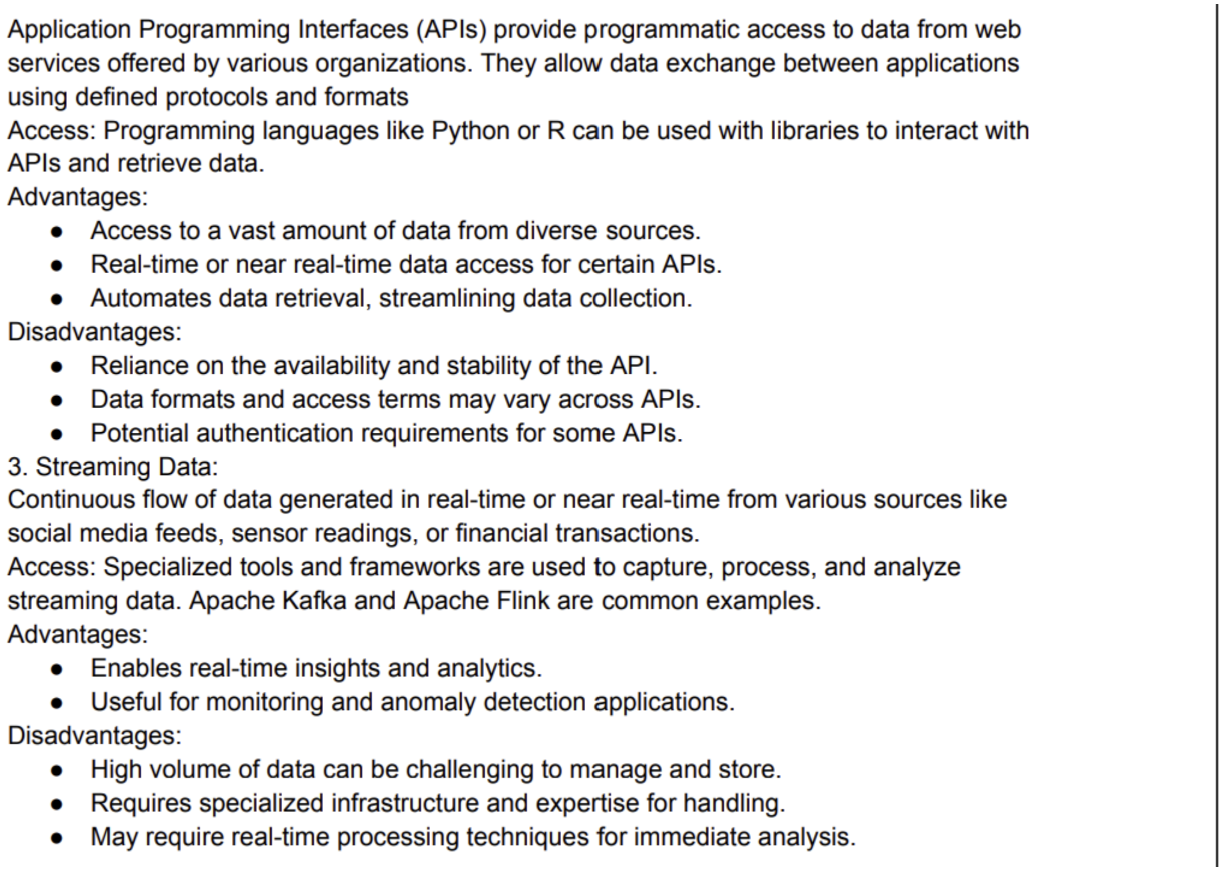
Efficient storage and retrieval of structured data.

- Strong data integrity through constraints and relationships between tables.
- Widely used and familiar to many data professionals.

Disadvantages:

- Less flexible for unstructured data (e.g., text, images).
- Can be complex to manage for very large datasets.

2. Web/API Access:



Application Programming Interfaces (APIs) provide programmatic access to data from web services offered by various organizations. They allow data exchange between applications using defined protocols and formats

Access: Programming languages like Python or R can be used with libraries to interact with APIs and retrieve data.

Advantages:

- Access to a vast amount of data from diverse sources.
- Real-time or near real-time data access for certain APIs.
- Automates data retrieval, streamlining data collection.

Disadvantages:

- Reliance on the availability and stability of the API.
- Data formats and access terms may vary across APIs.
- Potential authentication requirements for some APIs.

3. Streaming Data:

Continuous flow of data generated in real-time or near real-time from various sources like social media feeds, sensor readings, or financial transactions.

Access: Specialized tools and frameworks are used to capture, process, and analyze streaming data. Apache Kafka and Apache Flink are common examples.

Advantages:

- Enables real-time insights and analytics.
- Useful for monitoring and anomaly detection applications.

Disadvantages:

- High volume of data can be challenging to manage and store.
- Requires specialized infrastructure and expertise for handling.
- May require real-time processing techniques for immediate analysis.

Data Collection

Data collection is the first step in any data science project. It involves gathering the raw data that will be used for analysis and modeling.

Methods of Data Collection:

1.Sampling

- Random Sampling: Each member of the population has an equal chance of being selected.

  This is the gold standard for ensuring unbiased data and generalizable results.
- Techniques include:
  - ○ Simple Random Sampling: Every individual has an equal chance of being chosen, often achieved through random number generation.
  - ○ Systematic Random Sampling: The population is ordered in some way, and a random starting point is chosen. Then, every nth individual is selected.
  - ○ Stratified Random Sampling: The population is divided into subgroups (strata) based on relevant characteristics. Random sampling is then performed within each stratum to ensure representation of all subgroups.
- Non-Random Sampling: These techniques can be quicker or cheaper but may introduce bias:
  - Convenience Sampling: Selecting readily available individuals, often leading to biased representation.
  - Judgment Sampling: The researcher selects individuals based on their judgment of who is most appropriate, potentially introducing subjectivity.
  - Quota Sampling: Setting quotas for different subgroups within the sample to ensure representation, but may not be truly random.

Impact on Data Visualization, Modeling, and Generalizability:

- Data Visualization:

  Sampling bias can lead to misleading visualizations. A well-chosen sample should represent the population accurately to ensure visualizations reflect true patterns and trends.

- Modeling:

Biased samples can lead to models that perform well on the specific data used but fail to generalize to the broader population. Random sampling techniques help mitigate this risk.

- Generalizability:

  The generalizability of results refers to how well your findings from a sample apply to the entire population. Random sampling techniques increase the likelihood that your results can be generalized to the population you're interested in.

2. Study Design (Observational vs. Experimental)

Observational Studies: Analyze existing data without manipulating variables. They can identify associations between variables but cannot definitively establish cause-and-effect relationships. Here are some common types of observational studies:

- Cohort Studies: Follow a group of individuals (the cohort) over time, often for years or decades. Researchers observe how exposure to a factor (e.g., smoking) affects outcomes (e.g., lung cancer) within the cohort.

  Example: A study following a group of healthcare workers from different departments (exposed/not exposed to radiation) over 20 years to investigate the potential link between radiation exposure and cancer rates.

- Case-Control Studies: Compare individuals with a specific outcome (cases) to those without it (controls). Researchers then examine past exposures or characteristics to identify potential risk factors associated with the outcome.

  Example: A study comparing women with breast cancer (cases) to women without breast cancer (controls) to investigate past lifestyle habits and potential risk factors.
- Cross-Sectional Studies: Examine relationships between variables at a single point in time. They can identify associations but cannot determine the direction of causality (which variable came first).

  Example: A survey at a grocery store to investigate the relationship between customer age and preferred type of milk (whole, low-fat, etc.).

- Experimental Studies: Manipulate variables to observe their effect on an outcome. They can establish cause-and-effect relationships by controlling for other factors that might influence the outcome.

- Randomized Controlled Trials (RCTs): Considered the gold standard for establishing causality. Participants are randomly assigned to either a treatment group (receives the intervention) or a control group (receives a placebo or no intervention). Researchers then measure the outcome to see if the treatment has a causal effect.

  Example: A clinical trial where participants with high blood pressure are randomly assigned to receive a new medication (treatment group) or a placebo (control group) to assess the medication's effectiveness in lowering blood pressure.

- Pre-test/Post-test Designs: Measure an outcome before and after an intervention to assess its impact. However, external factors that occur between the pre-test and post-test might influence the results, making it challenging to establish a clear cause-and-effect relationship.

  Example: A pre-test/post-test design might be used to evaluate the effectiveness of a neweducational program. Students' test scores are measured before and after the program to assess learning gains. However, other factors like individual study habits could also influence the post-test scores.

Impact on Data Visualization, Modeling, and Generalizability:

The study design, whether observational or experimental, affects how you interpret data and the conclusions you can draw:

- Data Visualization:

  Observational studies may reveal correlations between variables, but visualizations cannot definitively establish cause-and-effect relationships.

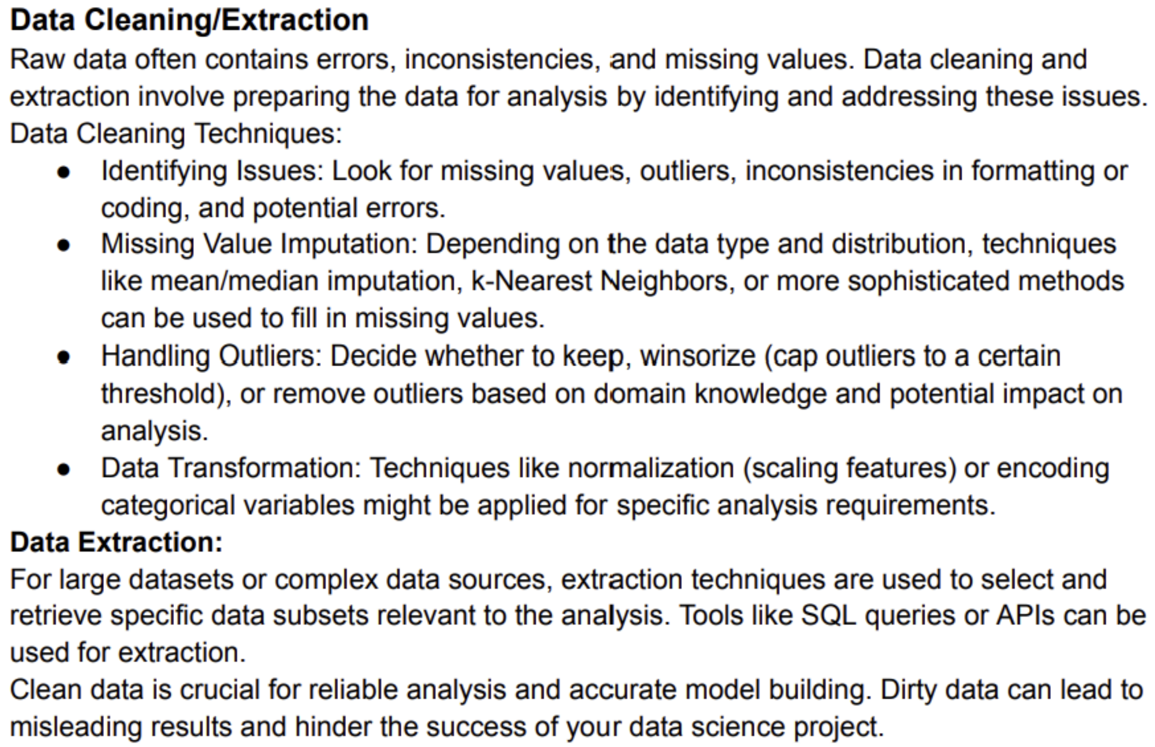
- Modeling:

Models built on observational data can predict future outcomes based on existing relationships, but they cannot isolate the impact of specific variables due to mpotential confounding factors.

- Generalizability:

The generalizability of results depends on the study design.

Experimental studies can provide stronger evidence of causality, but their generalizability can be limited if the experiment doesn't reflect real-world conditions.



Data Cleaning/Extraction

Raw data often contains errors, inconsistencies, and missing values. Data cleaning and extraction involve preparing the data for analysis by identifying and addressing these issues.

Data Cleaning Techniques:

- Identifying Issues: Look for missing values, outliers, inconsistencies in formatting or coding, and potential errors.
- Missing Value Imputation: Depending on the data type and distribution, techniques like mean/median imputation, k-Nearest Neighbors, or more sophisticated methods can be used to fill in missing values.
- Handling Outliers: Decide whether to keep, winsorize (cap outliers to a certain threshold), or remove outliers based on domain knowledge and potential impact on analysis.
- Data Transformation: Techniques like normalization (scaling features) or encoding categorical variables might be applied for specific analysis requirements.

Data Extraction:

For large datasets or complex data sources, extraction techniques are used to select and retrieve specific data subsets relevant to the analysis. Tools like SQL queries or APIs can be used for extraction.

Clean data is crucial for reliable analysis and accurate model building. Dirty data can lead to misleading results and hinder the success of your data science project.

Data Analysis & Modeling

Data analysis and modeling involve using statistical techniques and machine learning algorithms to uncover patterns, trends, and relationships within the data.

Data Analysis Techniques:

- Exploratory Data Analysis (EDA):

  This initial analysis involves summarizing, visualizing, and exploring the data to gain initial insights, identify patterns, and understand data distribution.
- Descriptive Statistics:

  Calculations like mean, median, standard deviation, and frequency distributions summarize the data and provide a foundational understanding of its central tendency, spread, and distribution.
- Correlation Analysis:

  Measures the strength and direction of linear relationship between numerical variables, helping identify potential relationships for further Exploration.

Data Modeling Techniques:

- Model Selection: Choose the appropriate modeling technique based on the data type (numerical, categorical) and the desired outcome (prediction, classification, etc.).

  Common models include linear regression, logistic regression, decision trees,random forests, etc.
- Model Training: Train the model on a portion of the data, allowing it to learn the patterns and relationships within the data.
- Model Evaluation: Evaluate the model's performance on a separate test dataset to assess its generalizability and effectiveness in making predictions or classifications on unseen data.

Data Warehouses

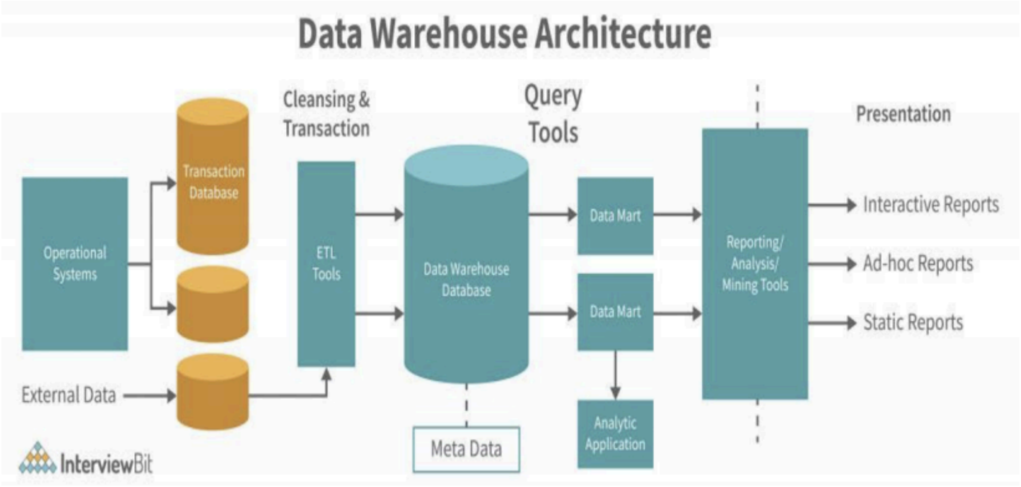
Data analysis and modeling are often iterative processes. Based on initial results, you may refine your analysis techniques, feature selection, or model training approach to improve model performance.

Data warehouse

- A Data Warehouse is separate from DBMS, it stores a huge amount of data
- collected from multiple sources like files, DBMS ,….
- to produce statistical results
- Strategic decision making

Features of Data Warehousing

- Centralized Data Repository
- Data Integration
- Historical Data Storage
- Query and Analysis
- Data Transformation
- Data Mining
- Data Security – backup, encryption

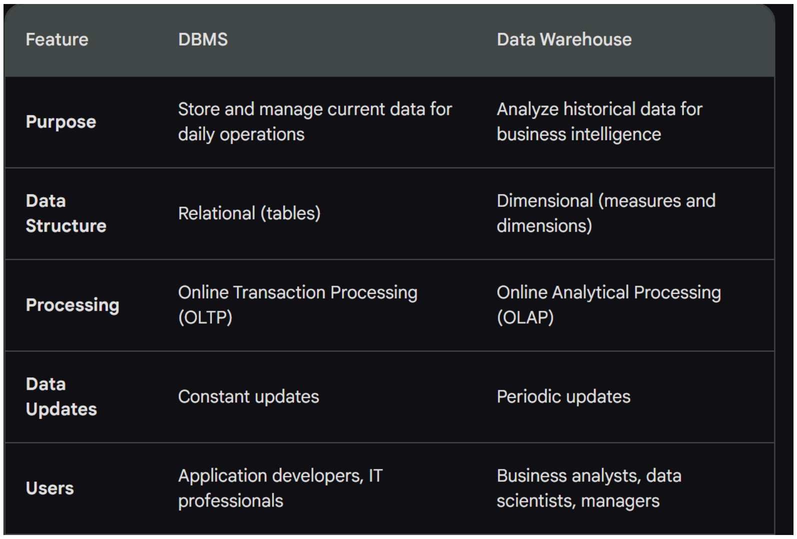


Benefits of warehouse

- Intelligent Decision-Making: With centralized data in warehouses, decisions may
- be made more quickly and intelligently.
- Business Intelligence: Provides strong operational insights through business intelligence.
- Historical Analysis: Predictions and trend analysis are made easier by storing past data.
- Data Quality: Guarantees data quality and consistency for trustworthy reporting.
- Scalability: Capable of managing massive data volumes and expanding to meet changing requirements.
- Effective Queries: Fast and effective data retrieval is made possible by an optimized structure.
- Cost reductions: Data warehousing can result in cost savings over time by reducing data management procedures and increasing overall efficiency, even when there are setup costs initially.
- Data security: Data warehouses employ security protocols to safeguard confidential information, guaranteeing that only authorized personnel are granted access to certain data.

Disadvantages of Data Warehousing

- Cost: Building a data warehouse can be expensive, requiring significant investments in hardware, software, and personnel.
- Complexity: Data warehousing can be complex, and businesses may need to hire specialized personnel to manage the system.
- Time-consuming: Building a data warehouse can take a significant amount of time, requiring businesses to be patient and committed to the process.
- Data integration challenges: Data from different sources can be challenging to integrate, requiring significant effort to ensure consistency and accuracy.
- Data security: Data warehousing can pose data security risks, and businesses must take measures to protect sensitive data from unauthorized access or breaches.

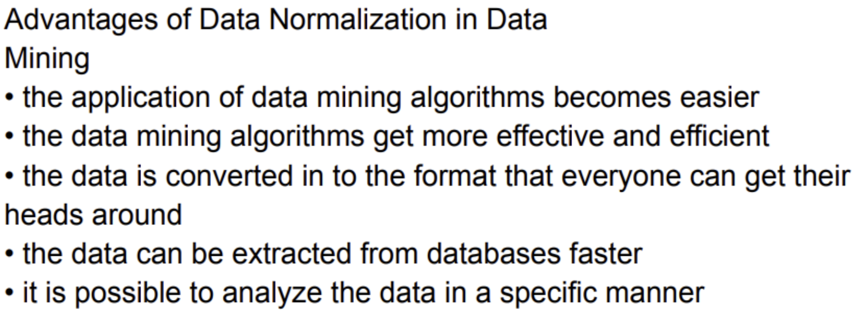


Feature	DBMS	Data Warehouse
Purpose	Store and manage current data for daily operations	Analyze historical data for business intelligence
Data Structure	Relational (tables)	Dimensional (measures and dimensions)
Processing	Online Transaction Processing (OLTP)	Online Analytical Processing (OLAP)
Data Updates	Constant updates	Periodic updates
Users	Application developers, IT professionals	Business analysts, data scientists, managers

DATA PRE PROCESSING

DATA NORMALIZATION

- Normalization is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0.
- It is generally useful for classification algorithms.



Advantages of Data Normalization in Data Mining

• the application of data mining algorithms becomes easier

• the data mining algorithms get more effective and efficient

• the data is converted in to the format that everyone can get their heads around

• the data can be extracted from databases faster

• it is possible to analyze the data in a specific manner

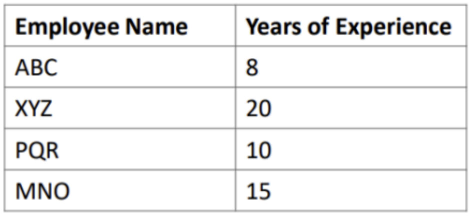
Min-max normalization

A technique in data science used to scale numeric data points into a fixed range, typically between 0 and 1. This is accomplished through a linear transformation, which adjusts the values based on the original dataset's minimum and maximum values.

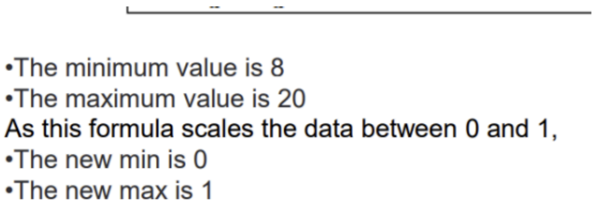
Formula = $v' = (v - min) / (max - min) * (new\_max - new\_min) + new\_min$

$v' = (10 - 8) / (20 - 8) * (1 - 0) + 0$

$v' = 0.16$ //for picc below sum can come



Employee Name	Years of Experience
ABC	8
XYZ	20
PQR	10
MNO	15



•The minimum value is 8

•The maximum value is 20

As this formula scales the data between 0 and 1,

•The new min is 0

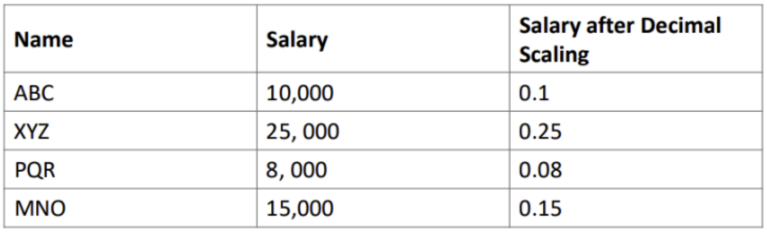
•The new max is 1

Decimal Scaling

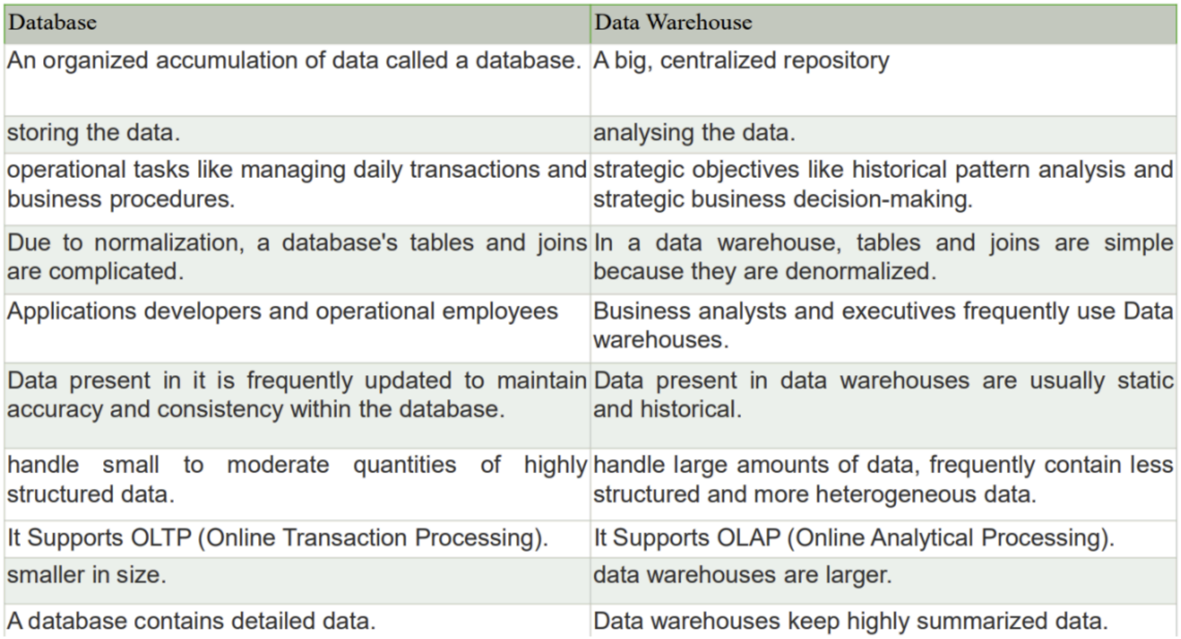
- Decimal scaling scales the data points relative to the maximum absolute value, so it doesn't necessarily guarantee a specific range like min-max normalization.
- It's a simpler approach compared to min-max normalization and can be useful when you know the data falls within a specific range.
- However, it can be sensitive to outliers, which can significantly alter the scaling factor and lead to unintended consequences.

v′ = v / (10^j)
v represents the original data value you want to normalize.
v′ represents the normalized value.
j represents the number of decimal places in the maximum absolute value.

The maximum absolute value is ₹25,000.
There are four decimal places (0.00025).
Apply the formula:
v′ = ₹10,000 / (10 ^ 4) = ₹0.10



Name	Salary	Salary after Decimal Scaling
ABC	10,000	0.1
XYZ	25, 000	0.25
PQR	8, 000	0.08
MNO	15,000	0.15



Database	Data Warehouse
An organized accumulation of data called a database.	A big, centralized repository
storing the data.	analysing the data.
operational tasks like managing daily transactions and business procedures.	strategic objectives like historical pattern analysis and strategic business decision-making.
Due to normalization, a database's tables and joins are complicated.	In a data warehouse, tables and joins are simple because they are denormalized.
Applications developers and operational employees	Business analysts and executives frequently use Data warehouses.
Data present in it is frequently updated to maintain accuracy and consistency within the database.	Data present in data warehouses are usually static and historical.
handle small to moderate quantities of highly structured data.	handle large amounts of data, frequently contain less structured and more heterogeneous data.
It Supports OLTP (Online Transaction Processing).	It Supports OLAP (Online Analytical Processing).
smaller in size.	data warehouses are larger.
A database contains detailed data.	Data warehouses keep highly summarized data.

Basic terms

- Independent variables: Data that can be controlled directly.
- Dependent variables: Data that cannot be controlled directly.

Experiment 1: You want to figure out which brand of microwave popcorn pops the mosY kernels so you can get the most value for your money. You test different brands of popcorn to see which bag pops the most popcorn kernels.

- Independent Variable: Brand of popcorn bag (It's the independent variable because you are actually deciding the popcorn bag brands)

- Dependent Variable: Number of kernels popped (This is the dependent variable because it's what youmeasure for each popcorn brand)

Experiment 2: You want to see which type of fertilizer helps plants grow fastest, so you add a different brand of fertilizer to each plant and see how tall they grow.

- Independent Variable: Type of fertilizer given to the plant
- Dependent Variable: Plant height