Chapter 1

Data Analysis Pipeline (3hrs)

1. Introduction to Data Analysis

Data Analysis is the process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making. It helps convert raw data into meaningful insights.

In simple terms, data analysis means *examining data to find patterns, trends, or relationships that can help solve problems or make better decisions.*

In today's data-driven world, data analysis is the backbone of intelligent decision-making. It helps organizations, governments, and individuals transform raw data into actionable knowledge, enabling smarter strategies, better policies, and innovation.

Significance of Data Analysis

Data analysis plays a crucial role in almost every field today. Its significance can be understood as follows:

a) *Informed Decision Making*

- Helps organizations make data-driven decisions rather than relying on intuition.
- Example: A company can analyze sales data to decide which product to promote.

b) Identification of Patterns and Trends

- Reveals hidden patterns, correlations, or anomalies in data.
- Example: Detecting customer behavior trends in marketing data.

c) Problem Solving

- Helps identify root causes of issues and evaluate the effectiveness of solutions.

d) Prediction and Forecasting

- Using statistical and machine learning methods, data analysis helps predict future events (e.g., demand forecasting, risk assessment).

ER SHIVA RAM DAM

e) Efficiency and Optimization

- Enables better resource allocation and process improvement by identifying inefficiencies.

Applications of Data Analysis

Data analysis is widely applied across many domains:

a) Business and Marketing

- Customer segmentation, market trend analysis, and demand forecasting.

b) Healthcare

- Analyzing patient data for disease prediction, treatment effectiveness, and health policy planning.

c) Education

- Tracking student performance, predicting dropouts, and improving teaching methods.

d) Government and Public Policy

- Population surveys, poverty estimation, and infrastructure planning (e.g., *Nepal MICS Report 2019*).

e) Environment and Ecology

- Predicting habitat suitability and assessing environmental changes (e.g., logistic regression in habitat modeling).

f) Social Sciences

- Understanding social trends, behaviors, and economic indicators.

Tools and Techniques Used

- **Statistical Methods:** Mean, correlation, regression, hypothesis testing.
- **Software/Tools:** Excel, SPSS, R, Python, SQL, Tableau, Power BI.
- **Machine Learning Algorithms:** Classification, clustering, prediction models.

ER. SHIVA RAM DAM

2. The knowledge Discovery from Database Process

In today's world, organizations collect massive amounts of data—student records, bank transactions, medical reports, web clicks, etc. However, raw data alone is not useful unless it is processed and analyzed to extract meaningful knowledge. This systematic process of turning data into knowledge is called Knowledge Discovery in Databases (KDD).

KDD includes data cleaning, integration, selection, transformation, data mining, pattern evaluation, and knowledge presentation. It is broader than data mining, which is just one step of the entire KDD process.

Steps in KDD:

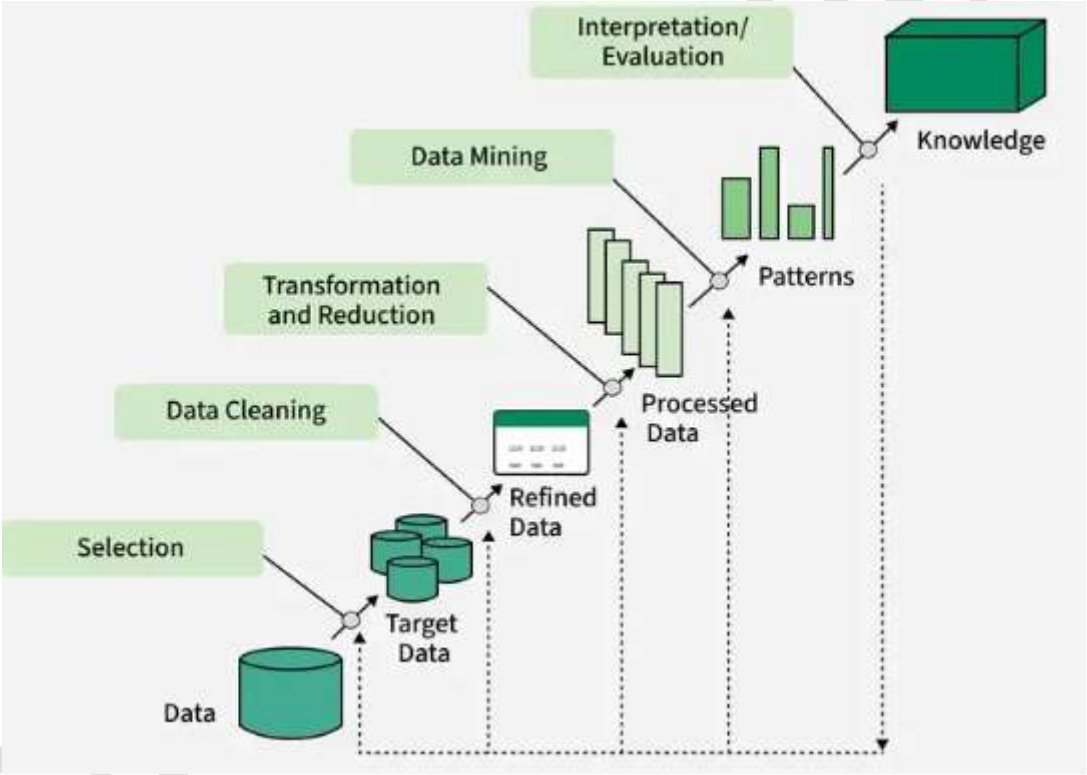

Figure: The KDD process

Source: https://media.geeksforgeeks.org/wp-content/uploads/20250128111402276995/kdd_process-768.webp

Step 1: Data Selection

- Involves selecting a dataset or focusing on specific variables, samples or subsets of data.

- Example: In a university database, selecting only *students of Computer Science department* for grade analysis.
- Tools: SQL queries, sampling, and filtering.

Step 2: Data Preprocessing (Cleaning and Integration)

- Corrects errors and inconsistencies in data.
- Deals with missing values, duplicate records, and outliers.
- Example: If attendance data and academic records are stored separately, integrate them for combined analysis.

Step 3: Data Transformation

- Converts data into formats suitable for mining.
- Involves dimensionality reduction.
- Includes normalization, aggregation, and feature generation.
- Example: Converting numerical grades into letter grades.

Step 4: Data Mining

- The core step: applying algorithms to identify patterns and models.
- Techniques: classification, clustering, association, regression.
- Example: Predicting student performance based on study hours and attendance.

Step 5: Pattern Evaluation

- Measures interestingness and validity of patterns.
- Example: A rule "Students with >80% attendance have 90% pass rate" is meaningful.

Step 6: Knowledge Presentation

- Uses visualization and reports to make the discovered knowledge understandable to decision-makers.
- Tools: graphs, dashboards, reports.

ER. SHIVA RAM DAM

3. Structured and Unstructured data and their examples

Data in the real world comes in different formats and levels of organization. Broadly, it can be categorized into Structured Data and Unstructured Data based on how easily it can be stored, accessed, and analyzed.

a) Structured Data

Structured data refers to data that is organized in a fixed format or schema — usually stored in rows and columns (like in databases or spreadsheets). Each field has a defined data type, such as integer, text, or date.

Characteristics:

- Stored in **tabular form** (tables, records, attributes).
- **Highly organized** and easy to enter, query, and analyze.
- Uses a **data model or schema** (e.g., relational model).
- Can be easily managed using **SQL** or traditional database systems.

Examples:

D	Name	Age	Salary
1	Sita	25	30,000
2	Ram	30	40,000

Other examples:

- Student records
- Bank transactions
- Census or survey data
- Sales data in Excel sheets

Storage Systems:

- Relational Databases (MySQL, Oracle, PostgreSQL)
- Data Warehouses

Advantages:

- Easy to store, retrieve, and analyze.
- Highly efficient for structured queries.
- Suitable for statistical and quantitative analysis.

ER. SHIVA RAM DAM

Limitations:

- Cannot handle complex data types like images, videos, or social media posts.
- Fixed schema — less flexibility.

b) Unstructured Data

Unstructured data refers to data that has no predefined format or organization. It doesn't fit neatly into tables — the structure is irregular or undefined.

Characteristics:

- Does not follow a fixed data model.
- Often text-heavy, but may contain numbers, images, audio, or video.
- Harder to store, search, and analyze.
- Requires special tools and techniques (like Natural Language Processing or Machine Learning).

Examples:

- Text documents and emails
- Social media posts (Facebook, Twitter, etc.)
- Images, audio, and video files
- Web pages and online reviews
- Sensor data and logs

Storage Systems:

- NoSQL Databases (MongoDB, Cassandra)
- Data Lakes (Hadoop, AWS S3)

Advantages:

- Can capture rich and detailed information.
- Reflects real-world data more accurately (e.g., customer opinions, behaviors).

Limitations:

- Difficult to process and analyze directly.
- Requires advanced analytical techniques (AI, NLP, deep learning).

ER SHIVA RAM DAM

- No fixed schema or predefined structure.
- Includes text, images, videos, social media posts, etc.
- Requires special tools for processing (e.g., NLP for text, CNNs for images).
- Examples:
    - YouTube videos, Tweets, E-mails, Audio files.

Advantages:

- Rich in information (contextual, emotional, behavioral).
- Valuable for modern analytics (e.g., sentiment analysis).

Disadvantages:

- Difficult to store, process, and mine directly.
- Requires preprocessing and feature extraction.

c) Semi-Structured Data

Between the two types, there is also semi-structured data, which has *some organizational elements* but not a strict schema.

Examples:

- XML, JSON, HTML, log files, sensor data.
- They contain tags or markers that separate data elements but allow flexibility.

Comparison between Structured and Unstructured Data

Basis	Structured Data	Unstructured Data
Format	Fixed schema (rows and columns)	No predefined structure
Storage	RDBMS, data warehouses	NoSQL, data lakes
Ease of Analysis	Easy (SQL, statistics)	Difficult (AI/ML, NLP)
Examples	Student records, sales data	Emails, videos, tweets
Volume	Smaller, well-defined	Very large (Big Data)
Flexibility	Less flexible	Highly flexible

ER. SHIVA RAM DAM

4. Overview of Data Preprocessing

Raw data collected from various sources is often incomplete, inconsistent, noisy, and unformatted. Before data can be analyzed or mined for patterns, it must be properly preprocessed**.**

Data Preprocessing refers to the series of steps used to prepare raw data into a suitable format for analysis or data mining. It improves data quality, reduces errors, and enhances the accuracy of analytical results.

Importance of Data Preprocessing

- Increases data quality and reliability
- Enhances model performance
- Reduces processing time and errors
- Essential for accurate, meaningful data mining results

a) Data Cleaning

Data cleaning is the process of detecting, correcting, or removing errors and inconsistencies from data to improve its quality.

In real-world data collection, data often comes from different sources — surveys, sensors, social media, or databases — and may contain missing values, duplicates, inaccurate entries, or inconsistent formats. If these issues are not fixed, they can lead to misleading insights, biased models, and wrong decisions.

Common Problems in Raw Data:

1. **Missing Data** – Some values are absent (e.g., age not recorded).
2. **Inconsistent Data** – Different formats or standards (e.g., “Male” vs “M”).
3. **Duplicate Records** – The same observation appears more than once.
4. **Outliers** – Unusual values that deviate from the normal pattern.
5. **Typographical Errors** – Mistakes in data entry (e.g., “Nepaal” instead of “Nepal”).
6. **Irrelevant Data** – Unnecessary attributes that do not contribute to analysis.

ER. SHIVA RAM DAM

Steps in Data Cleaning

1. **Data Inspection**
    - Examine datasets using statistical summaries or visualization tools to find anomalies.
    - Example: Using boxplots to detect outliers.
2. **Handling Missing Values**
    - **Remove** rows or columns with too many missing values.
    - **Impute** values using mean, median, mode, or prediction methods.
3. **Removing Duplicates**
    - Identify and delete duplicate rows using unique keys or IDs.
4. **Correcting Data Types and Formats**
    - Ensure numeric fields are stored as numbers, dates in correct format, etc.
5. **Standardizing Data**
    - Convert data into a consistent format (e.g., "Yes/No" instead of "Y/N/yes").
    - Unify units (e.g., converting all weights to kilograms).
6. **Handling Outliers**
    - Detect using statistical tests or visualization (boxplots).
    - Decide whether to remove or transform them.
7. **Validation and Verification**
    - Recheck data accuracy after cleaning using summary statistics or domain knowledge.

b) Data Integration

Data integration is the process of combining data from multiple sources into a single, consistent data store, such as a data warehouse.

It ensures that data collected from different systems (such as databases, files, APIs, or sensors) can be accessed and analyzed together in a consistent and meaningful way**.**

In data science, data often comes from diverse sources — such as sales records, social media, surveys, or government databases — each using different formats and structures. To perform analysis, all this data must be merged, standardized, and harmonized**.**

ER. SHIVA RAM DAM

Example:

A company may store:

- Customer data in an Excel sheet
- Sales data in an SQL database
- Marketing data in a cloud platform

To analyze customer buying patterns, these data sources must be integrated.

Example:

Source	Data
Database 1	Customer ID, Name, Phone
Database 2	Customer ID, Purchase History
Excel Sheet	Customer ID, Feedback

Integrated Dataset:
| Customer ID | Name | Phone | Purchase History | Feedback |

Objectives of Data Integration

1. **Combine data** from multiple sources into a unified dataset.
2. **Eliminate redundancy** and inconsistency among sources.
3. **Provide a complete view** of information for analysis or reporting.
4. **Enable efficient data access** for business intelligence and data mining.

Techniques:

- Schema matching and merging
- Entity resolution
- Correlation analysis to detect redundant attributes

ER SHIVA RAM DAM

Steps in Data Integration

1. **Data Collection**
    - Gather data from different sources such as databases, files, APIs, or web services.
2. **Data Cleaning and Transformation**
    - Standardize units, formats, and values (e.g., date formats, currency, naming conventions).
    - Remove duplicates and handle missing values.
3. **Schema Integration**
    - Combine tables or attributes with different names or structures but same meaning.
    - Example: “Emp_ID” in one dataset and “Employee_No” in another refer to the same field.
4. **Entity Identification (Record Linkage)**
    - Identify and merge records referring to the same entity across datasets.
    - Example: “A. Sharma” and “Anil Sharma” may represent the same person.
5. **Data Redundancy Handling**
    - Detect and remove overlapping or repeated data.
6. **Data Loading and Storage**
    - Load the integrated data into a central repository such as a **Data Warehouse** or **Data Lake**.

Challenges in Data Integration:

1. **Schema integration:** Aligning different data formats, names, or structures. Example: “Cust_ID” in one dataset and “CustomerNo” in another.
2. **Entity identification problem:** Identifying whether records from different sources refer to the same entity.
3. **Data redundancy:** Removing duplicate or overlapping data.

ER. SHIVA RAM DAM

4. **Data value conflicts:** Handling conflicting information (e.g., same person with two different birthdates).

Common Data Integration Techniques

Technique	Description
Manual Integration	Data combined manually by analysts (used in small projects).
ETL (Extract, Transform, Load)	Data extracted from sources, transformed for consistency, and loaded into a warehouse.
Data Warehousing	Central repository storing integrated, historical data for analysis.
Data Virtualization	Accessing and querying data from multiple sources without physically merging them.
Middleware Integration	Software that connects different systems and facilitates data flow.

Tools Used for Data Integration

- **ETL Tools:** Talend, Informatica, Apache Nifi
- **Programming:** Python (Pandas, PySpark)
- **Database Systems:** SQL Server Integration Services (SSIS)
- **Big Data Platforms:** Hadoop, Apache Kafka

c) Data Transformation and Discretization

Data Transformation:

Data Transformation is the process of converting data into an appropriate format or structure that is suitable for analysis or modeling. It helps make data consistent, comparable, and easier to interpret.

Why It's Needed:

- Raw data often comes in **different scales, units, or formats**.

ER. SHIVA RAM DAM

- Many data mining and machine learning algorithms require **normalized or standardized** input.
- Transformation makes data more meaningful and ready for statistical analysis.

Common Types of Data Transformation

1. **Normalization**
    - Adjusts the scale of numeric data so that values fall within a specific range, usually **[0, 1]**.
    - Useful when features have different units (e.g., income vs. age).
    - **Formula:**

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Example:

Age values (10, 20, 30, 40) → Normalized to (0, 0.33, 0.66, 1.0)

2. **Standardization (Z-score scaling)**
    - Converts data to have **mean = 0** and **standard deviation = 1**.
    - Useful when data follows a **normal distribution**.
    - **Formula:**

$$z = \frac{x - \mu}{\sigma}$$

Example:

Temperature data transformed so that deviations from the mean are comparable.

3. **Aggregation**
    - Combines multiple values into a single value.
    - Example: Average monthly sales from daily sales data.

ER. SHIVA RAM DAM

4. **Generalization**
    - Replaces low-level details with higher-level concepts.
    - Example: Replace “Kathmandu”, “Pokhara”, “Biratnagar” with “Nepal”.
5. **Attribute Construction**
    - Creates new attributes from existing ones to enhance analysis.
    - Example: From “Height” and “Weight”, construct “BMI”.
6. **Encoding (for categorical data)**
    - Converts categorical data into numeric form.
    - Example: “Gender” → Male = 0, Female = 1.

Data Discretization

Discretization is the process of converting continuous numerical data into discrete categories or intervals. It helps in reducing data size and improving interpretability.

Why Discretization?

- Reduces the impact of minor observation errors.
- Makes patterns **easier to detect**.
- Useful in models like **Naïve Bayes** and **Decision Trees** that prefer categorical input.

Methods of Discretization:

1. **Equal-width discretization:**
   Divides data range into *k* equal-sized intervals.
   Example:
   Marks range = 0–100
   Bins:
    - 0–33 → “Low”

ER. SHIVA RAM DAM

- 34–66 → "Medium"
- 67–100 → "High"

2. **Equal-frequency discretization:**
   Each interval has approximately the same number of data points.
   Example:
   12 students' marks divided into 3 bins → 4 students per bin.

3. **Supervised discretization:** Uses class information to create meaningful intervals (e.g., decision tree-based).
   Example: Creating income ranges that best separate "buy" vs "not buy" customers.

***

ER. SHIVA RAM DAM