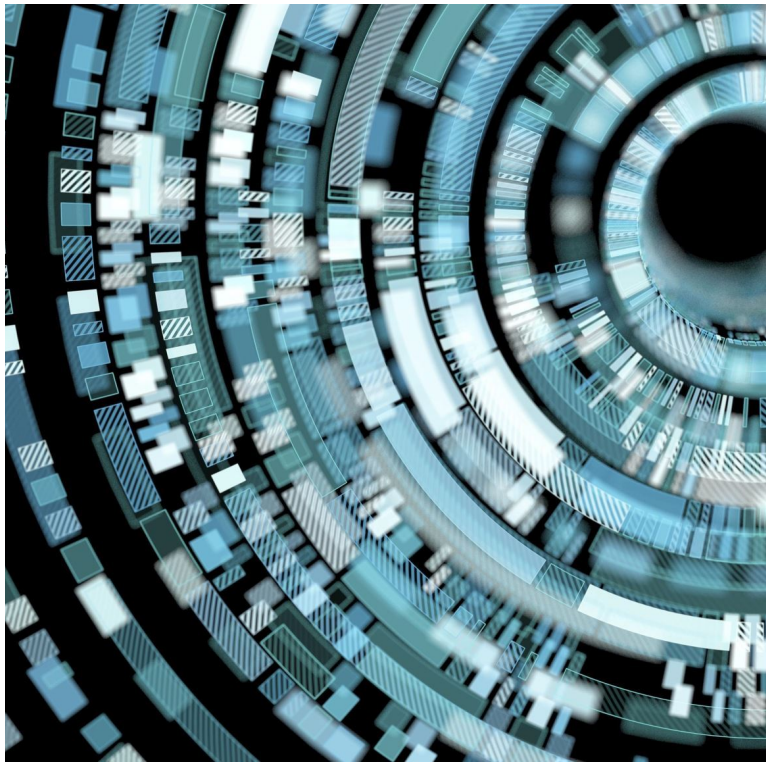
Introduction to Data Engineering

- Overview, Lifecycle, and Applications
- Pawan Kumar Sharma



What is Data Engineering ?

DEFINITION

IMPORTANCE

APPLICATIONS

Definition of Data Engineering

•The field of software engineering focused on the design, development, and management of systems that handle large volumes of data for Ensuring data is accessible, reliable, and ready for analysis and reporting.

Importance of Data Engineering

•Data engineering is crucial for ensuring that data is accessible, reliable, and ready for analysis. It supports data-driven decision-making, enhances business intelligence, and enables real-time analytics. By managing data pipelines, storage, and processing, data engineering allows organizations to extract valuable insights and maintain data quality and consistency.

Applications of Data Engineering

- Business Intelligence
- Machine Learning
- Data Warehousing
- Real-Time Analytics
- ETL (Extract, Transform, Load) Processes

Role of a Data Engineer

Responsibilities

Skills Required

Tools Used

Responsibilities of a Data Engineer

Data Pipeline Development

Data Integration

Database Management

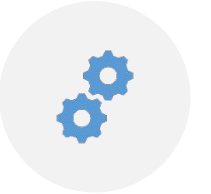
Data Quality Assurance

Collaboration

Skills Required for a Data Engineer

Proficiency in programming languages: Python, Java, Scala.

Knowledge of SQL and NoSQL databases.



Experience with ETL (Extract, Transform, Load) processes.

Tools Used by Data Engineers

- SQL Databases: MySQL, PostgreSQL, Oracle.
- NoSQL Databases: MongoDB, Cassandra, DynamoDB.
- ETL Tools(SSIS)
- Data pipeline tools(Airflow)

Data Engineering Lifecycle

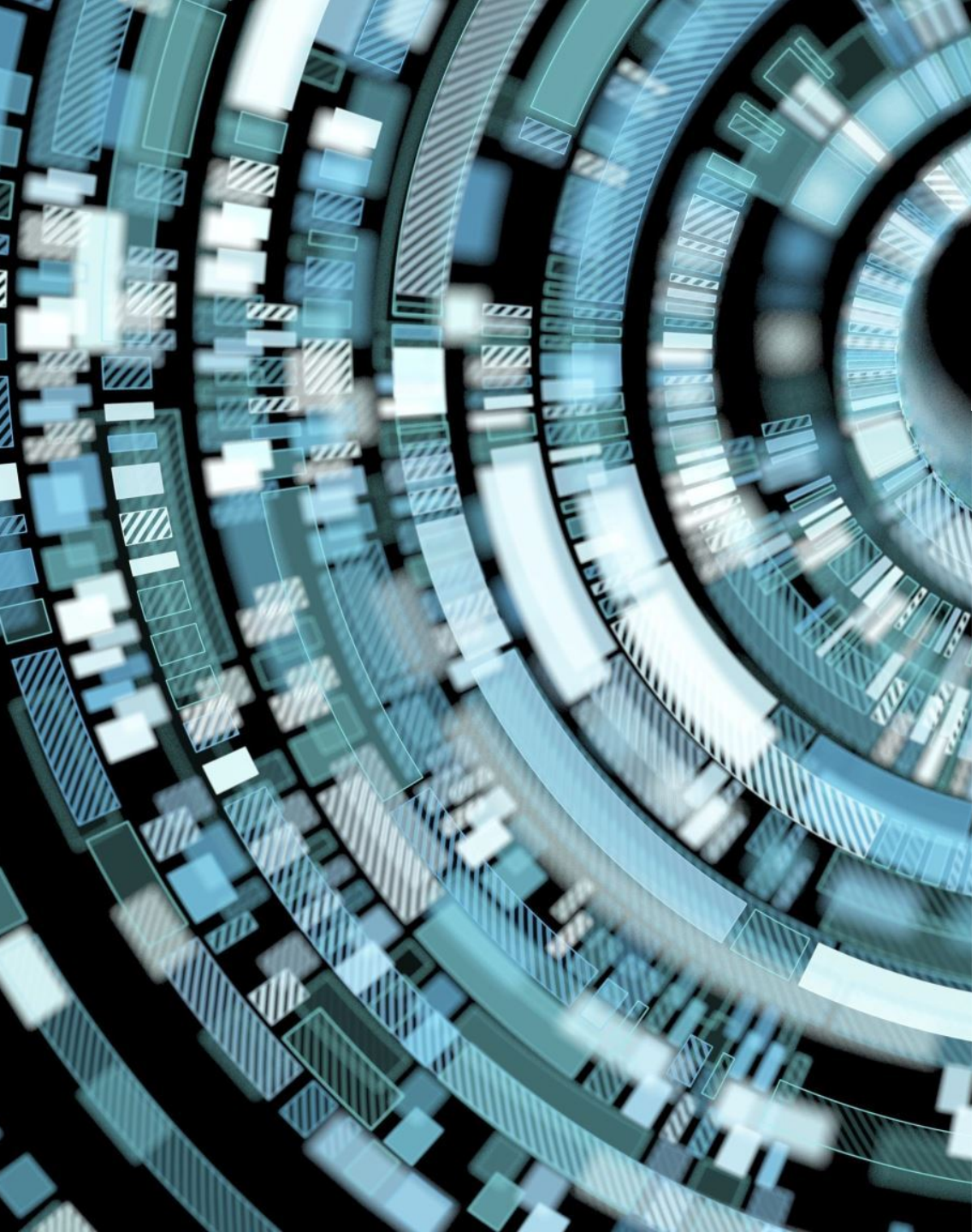
Data Generation

Data Collection

Data Storage

Data Processing

Data Analysis



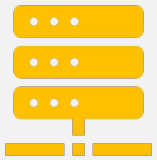
What is Data Generation ?

- Data generation is the process of creating data from various sources.
- Examples of data sources: Sensors, user interactions, social media, transactions, etc.

Types of Data Generated

Structured data: Tables, databases.

Unstructured data: Text, images, videos.



Semi-structured data: JSON, XML,HTML.

Importance of Data Generation

- Basis for informed decision-making.
- Enhances the capability to analyze trends and patterns.

What is Data Collection & Data Collection Techniques ?

The process of gathering and measuring information on variables of interest.

Surveys and questionnaires.

Online tracking tools.

Logs and event data.

Challenges in Data Collection

- Ensuring data accuracy.
- Managing large volumes of data.
- Privacy and security concerns.

What is Data Storage?

- The process of saving data in a systematic way for future use.

Types of Data Storage

- Relational databases: SQL, PostgreSQL.
- NoSQL databases: MongoDB, Cassandra.
- Data warehouses: Snowflake, Amazon Redshift.

What is Data Processing ?

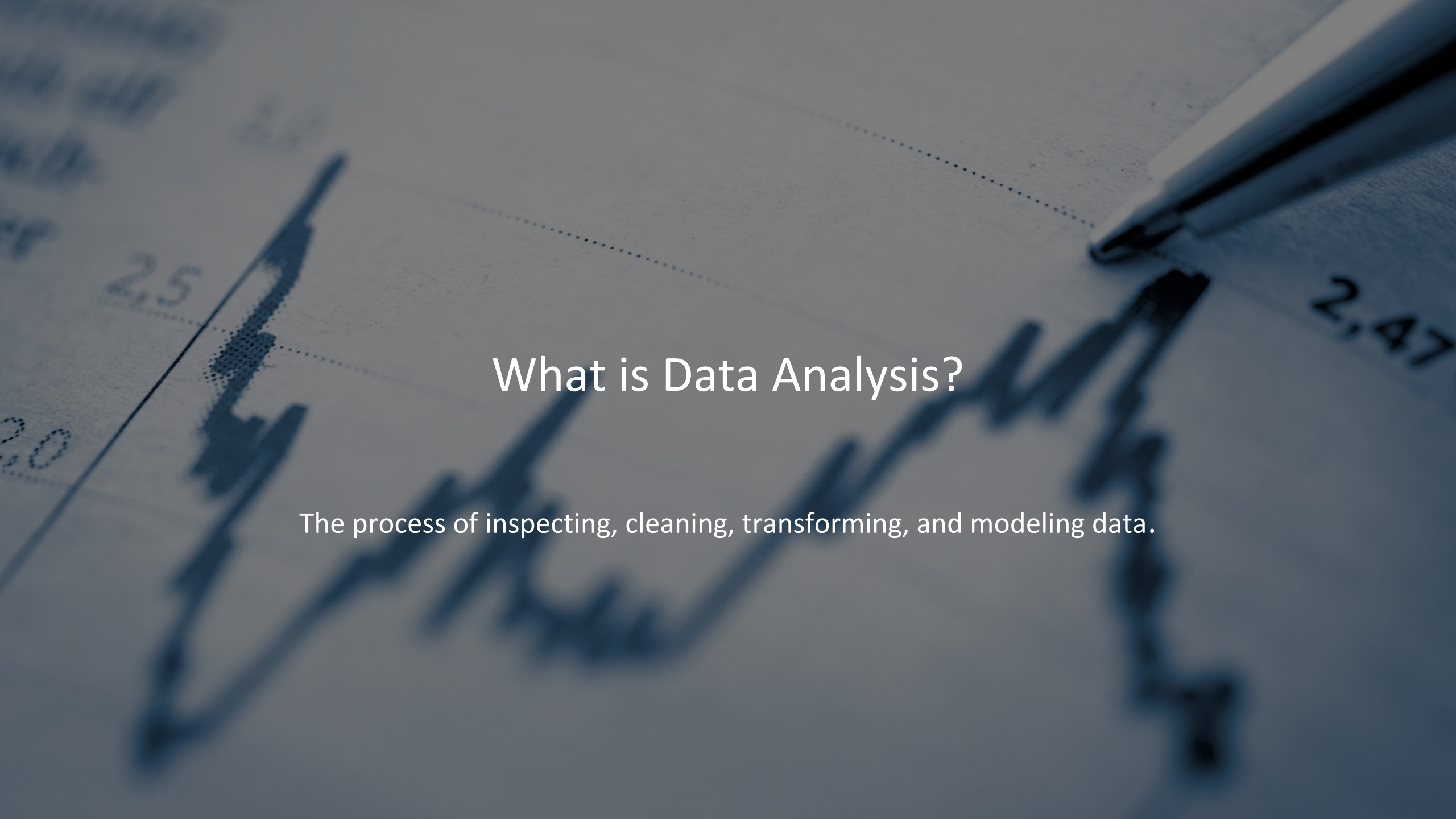
- The act of converting raw data into meaningful information.

Techniques :

- Batch processing.
- Stream processing.
- Real-time processing.

Challenges in Data Processing

- Handling big data volumes.
- Ensuring data quality.
- Maintaining data integrity.

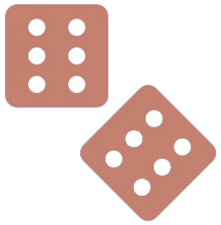


What is Data Analysis?

The process of inspecting, cleaning, transforming, and modeling data.

Types of Data Analysis

Descriptive analysis.



Predictive analysis.

Prescriptive analysis.

Data Generation and Collection

SOURCES OF DATA

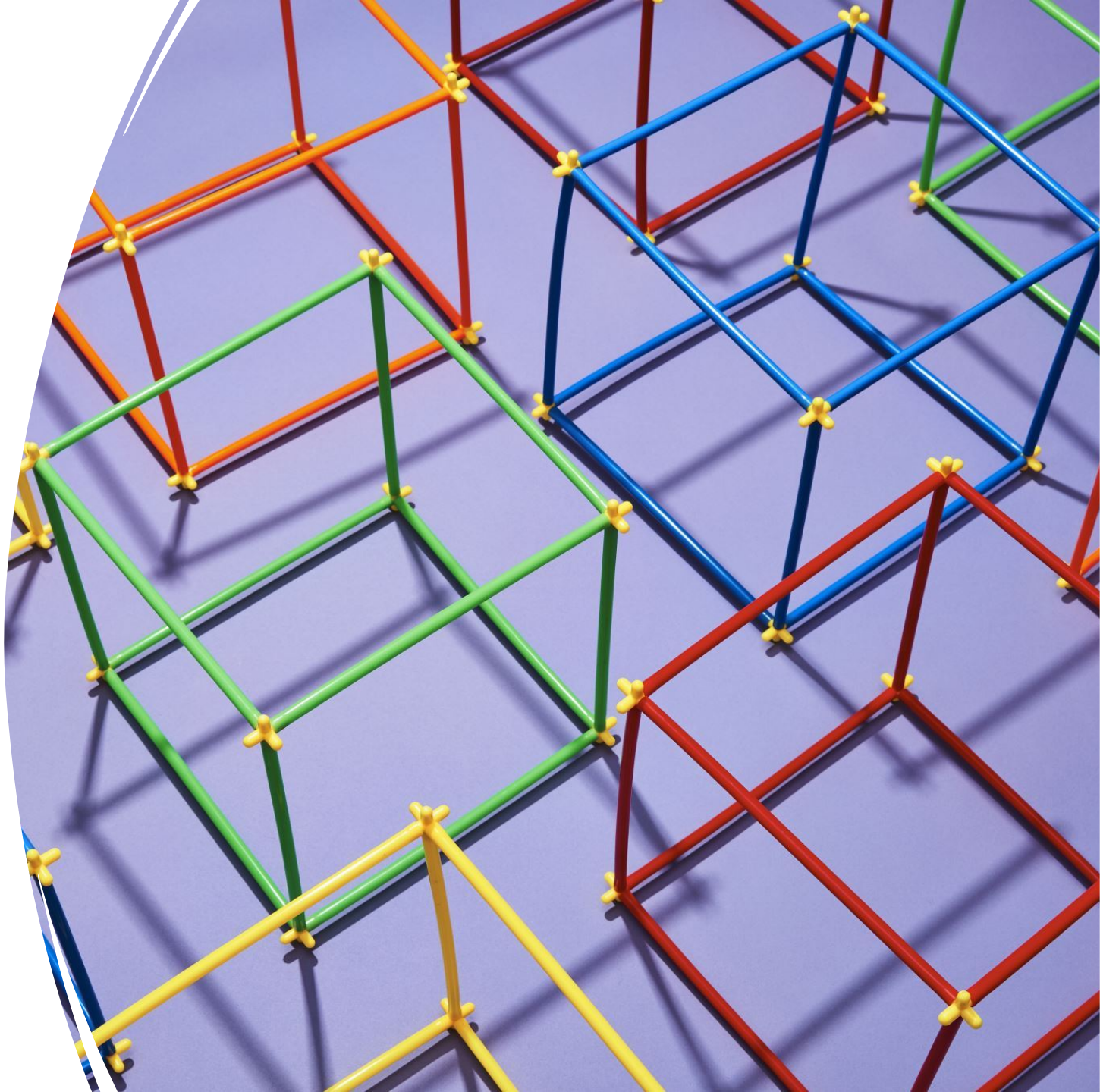
DATA LAKE

Data Sources

- Sources from which data is generated and collected for analysis.

Types of Data Sources

- **Internal Data Sources**: Data generated within an organization (e.g., transactional data, CRM data).
- **External Data Sources**: Data collected from outside the organization (e.g., social media, third-party data).



Data Lake

- A centralized repository that allows you to store all your structured and unstructured data at any scale.

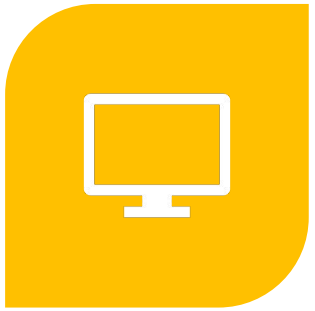
Features of a Data Lake

- **Scalability**: Can handle large volumes of data.
- **Flexibility**: Supports all data types (structured, semi-structured, unstructured).
- **Accessibility**: Data is easily accessible for processing and analysis.

Data Collection Methods

BATCH PROCESSING

STREAMING



WEB SCRAPING

APIS AND DATA EXTRACTION

Batch Processing

The collection and processing of data in large volumes at scheduled intervals.

Characteristics of Batch Processing

Scheduled Intervals: Data is processed at specific times (e.g., nightly, weekly).

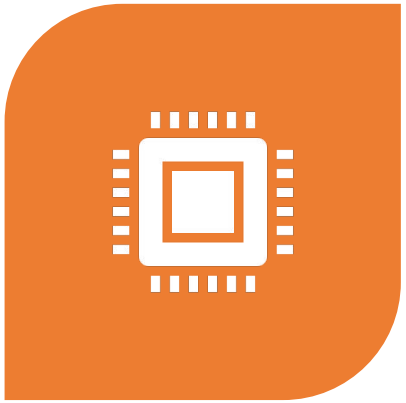
Large Volumes: Suitable for processing large datasets.

Non-Real-Time: Data is not processed in real-time, leading to some delay.

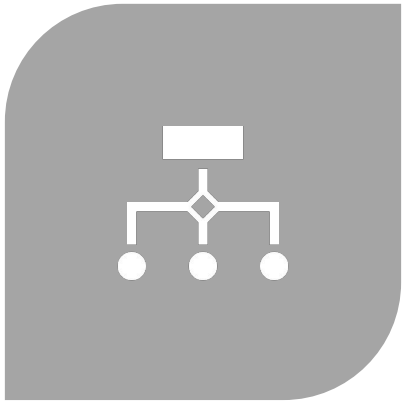
Streaming

The real-time collection and processing of data as it is generated.

Characteristics of Streaming



REAL-TIME PROCESSING: DATA IS PROCESSED AS SOON AS IT IS GENERATED.

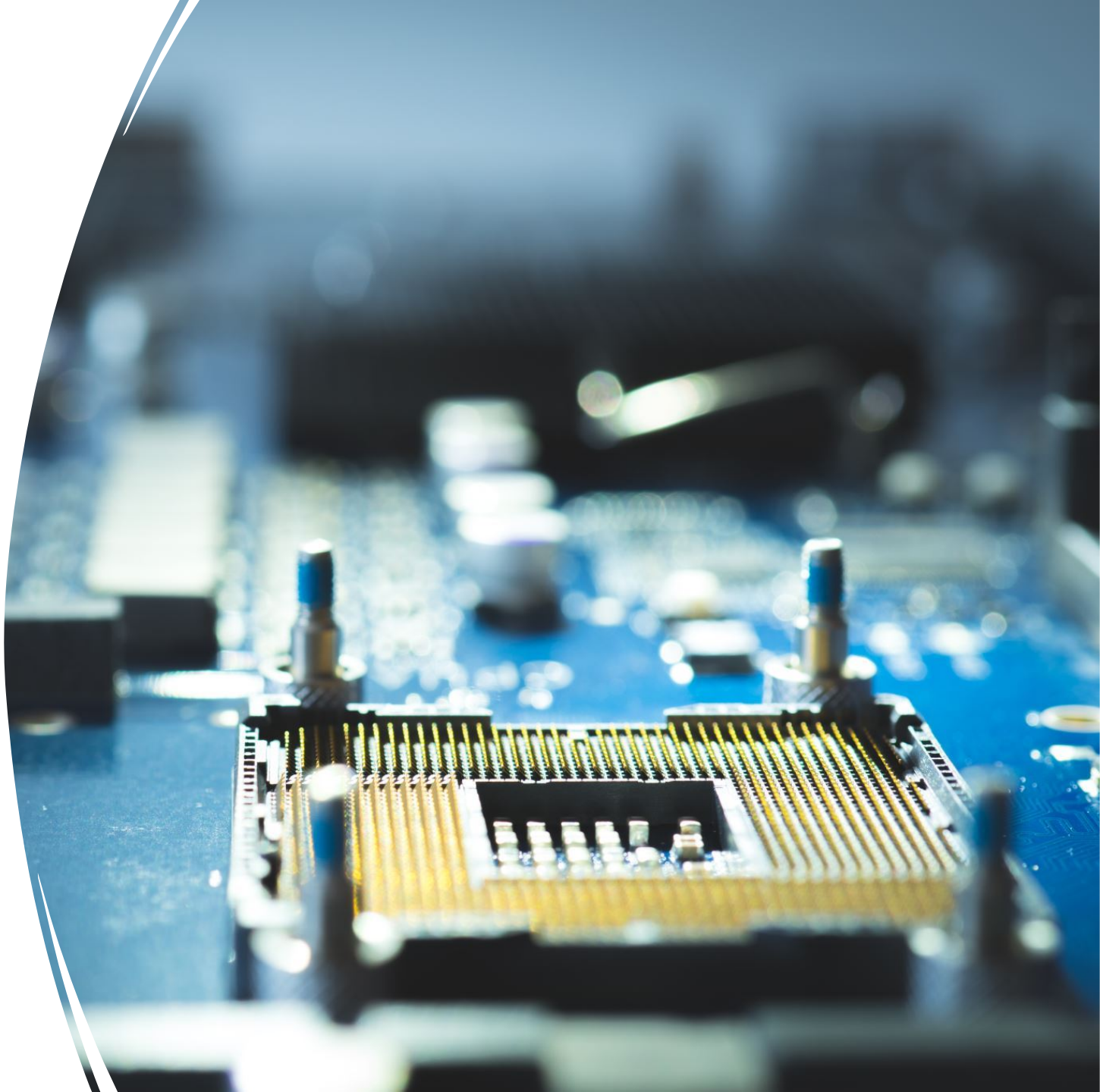


CONTINUOUS FLOW: DATA IS COLLECTED AND PROCESSED CONTINUOUSLY.

LOW LATENCY: IMMEDIATE INSIGHTS FROM DATA.

Web Scraping

- The automated extraction of data from websites.



Characteristics of Web Scraping

- **Automated Extraction**: Using bots or scripts to collect data from web pages.
- **Unstructured Data**: Often involves extracting data from unstructured or semi-structured sources.
- **Dynamic Content**: Can handle dynamic and frequently updated content.

APIs and Data Extraction

Using Application Programming Interfaces (APIs) to extract data from various sources.

Characteristics of APIs

- **Standardized Access**: Provides a standard way to access and retrieve data.
- **Real-Time or Batch**: Can be used for both real-time and batch data extraction.
- **Secure Access**: Often requires authentication and authorization.