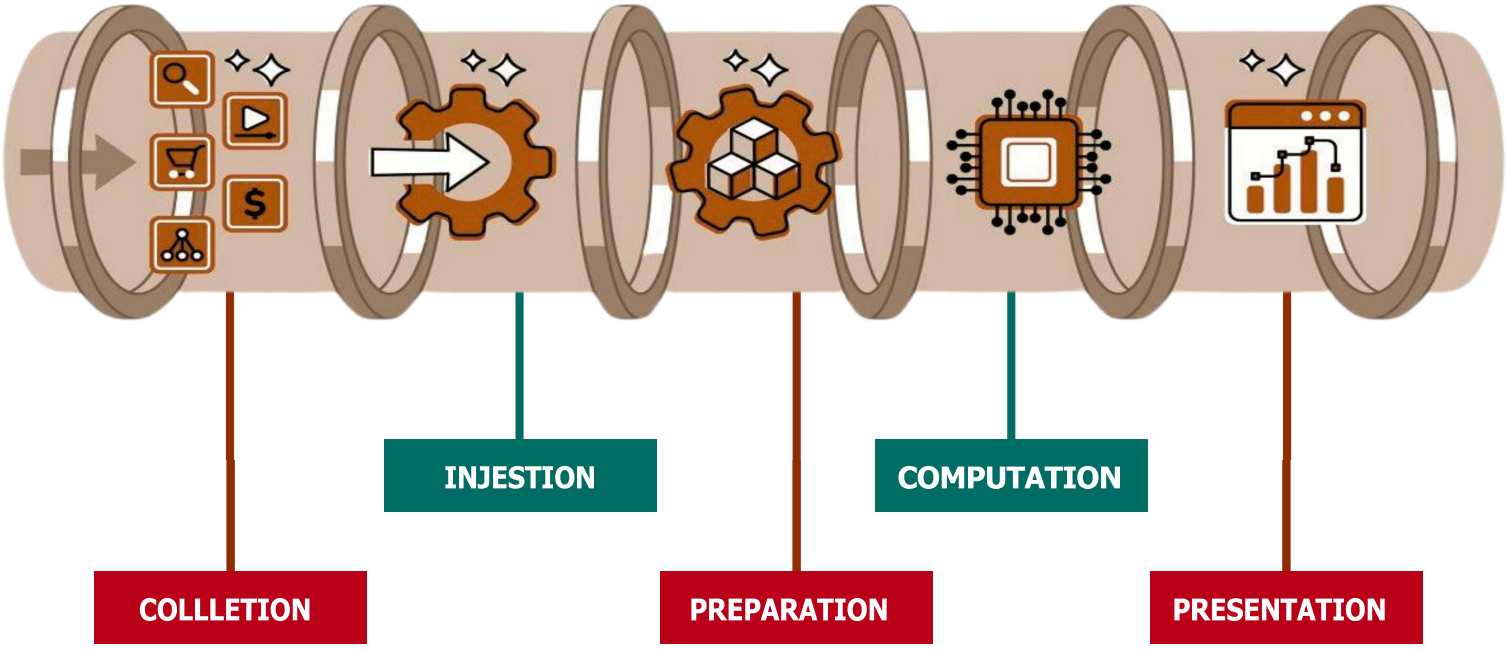
THE ULTIMATE 20-TERM CHECKLIST FOR EVERY DATA ENGINEER

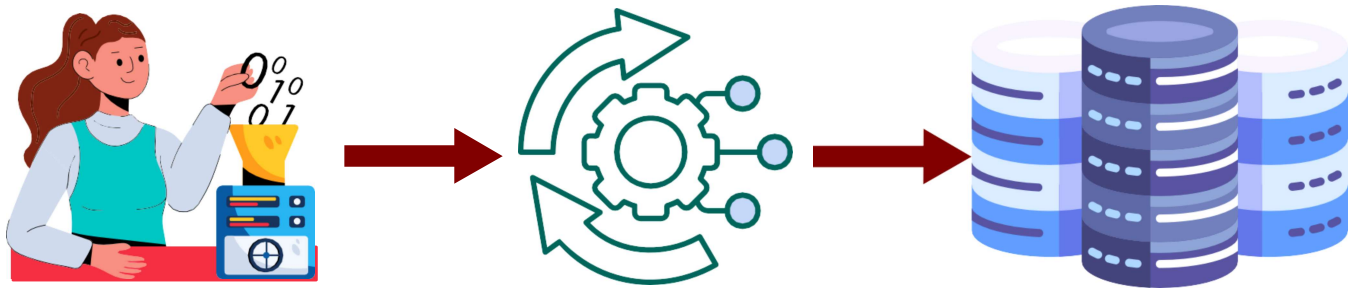
1. Data Pipeline



A data pipeline is a series of interconnected processes that automate the flow of data from various sources, through transformation steps, and into a destination system, such as a data warehouse, data lake, or analytics platform. Pipelines ensure that data is collected, processed, and made available for analysis with minimal manual intervention.

Swipe

2. ETL
(Extract, Transform, Load)



Extract Data | **Transform** | **Load in database**

ETL is a data integration process involving three key stages:

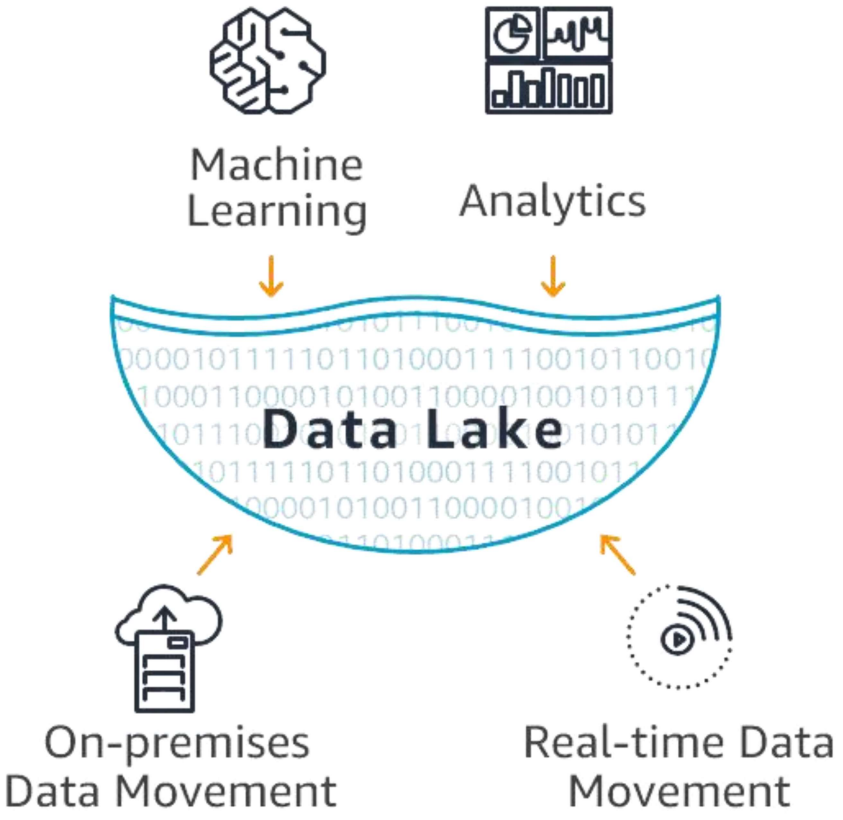
Extract: Gathering data from diverse sources, such as databases, APIs, or files.

Transform: Cleaning, validating, and converting the data into the desired format or structure.

Load: Moving the transformed data into a destination system like a data warehouse or database for further analysis.

Swipe

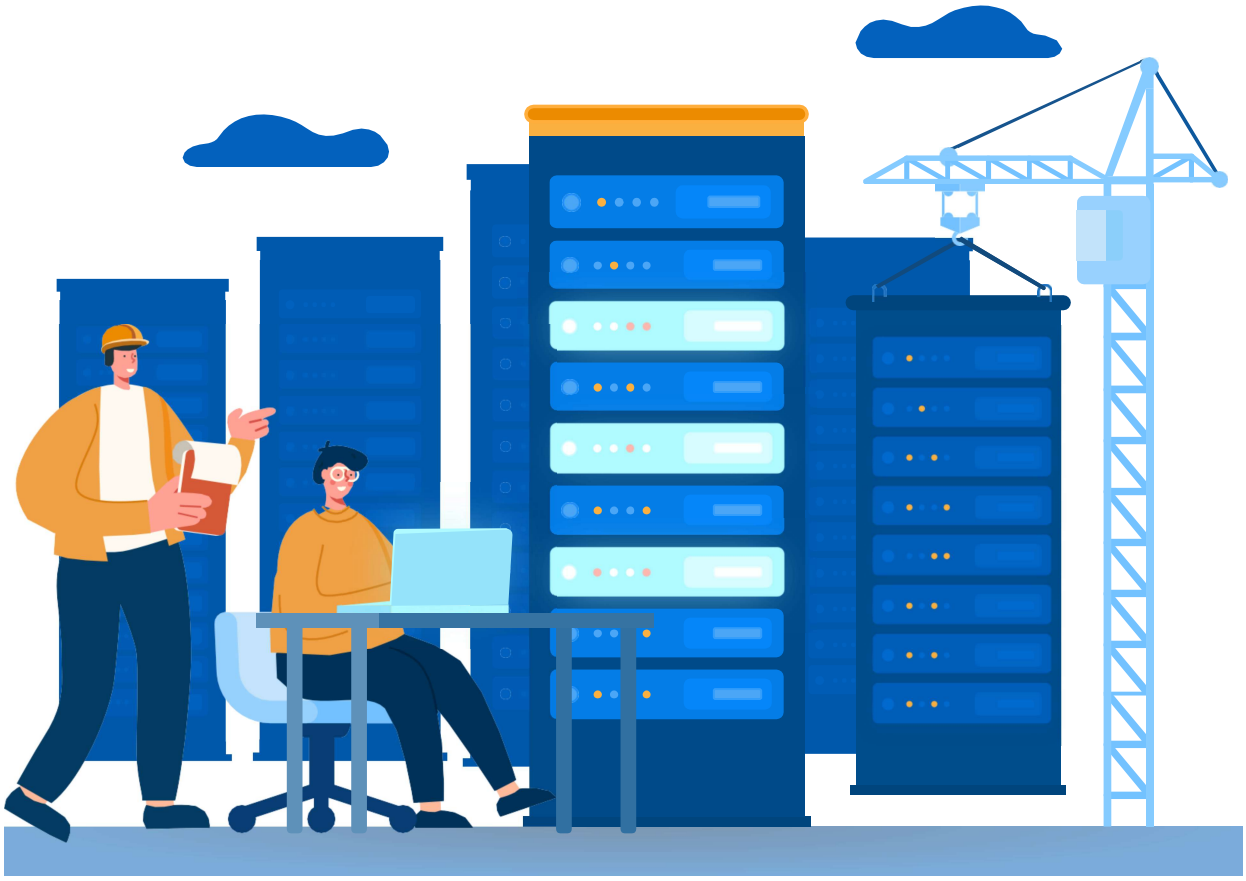
3. Data Lake



A data lake is a large, centralized repository that stores vast amounts of unstructured, semi-structured, and structured data at scale. Unlike traditional databases, data lakes hold raw data in its native form, enabling storage flexibility and easier access for machine learning, analytics, and future processing.

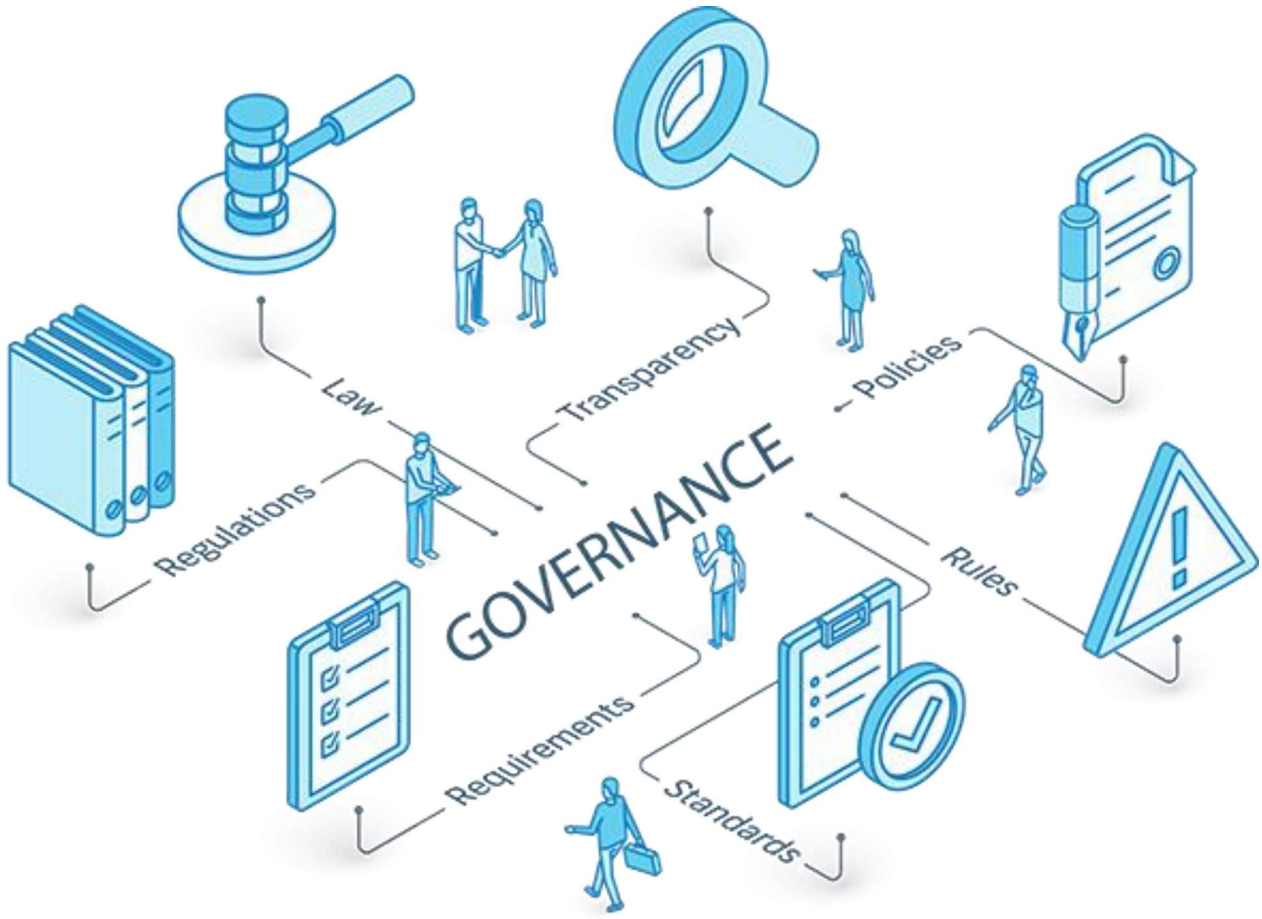
Swipe

4. Data Warehouse



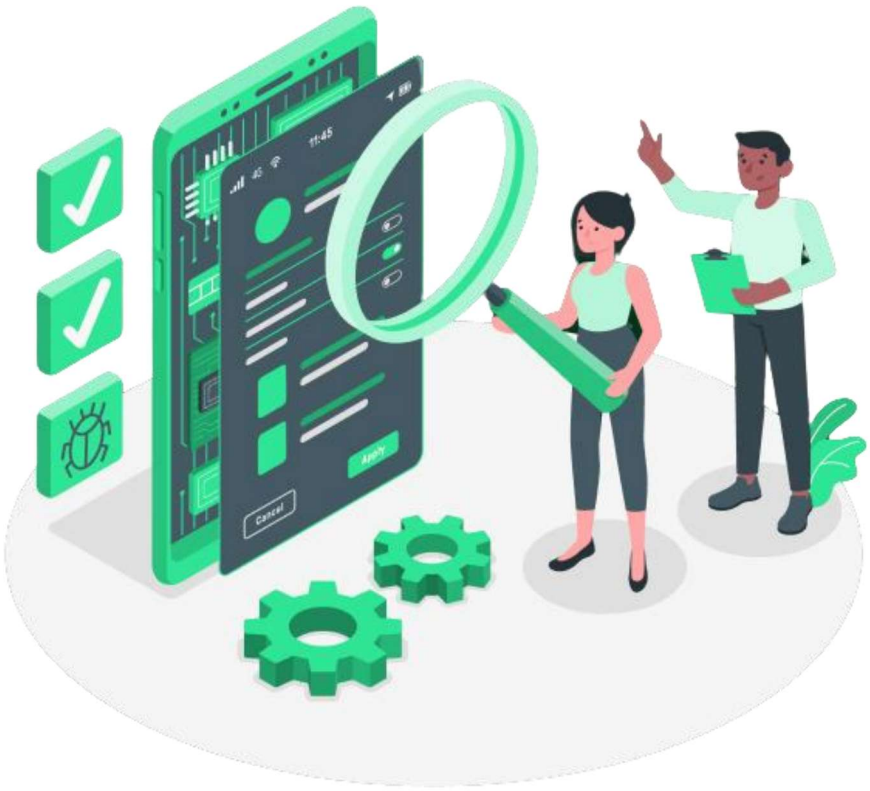
A data warehouse is a centralized system designed for storing structured data from different sources, optimized for querying and analysis. Data warehouses typically follow a schema-on-write approach, where data is pre-processed and structured for high-performance analytics, commonly used in business intelligence (BI).

5. Data Governance



Data governance refers to the processes, policies, and standards implemented to ensure that data is accurate, accessible, secure, and used responsibly across an organization. Effective governance includes data quality management, compliance with regulations, and ensuring the proper handling of sensitive data.

6. Data Quality



Data quality refers to the condition of data, focusing on factors such as accuracy, completeness, consistency, timeliness, and reliability. High-quality data is essential for ensuring accurate analysis, business intelligence, and decision-making.

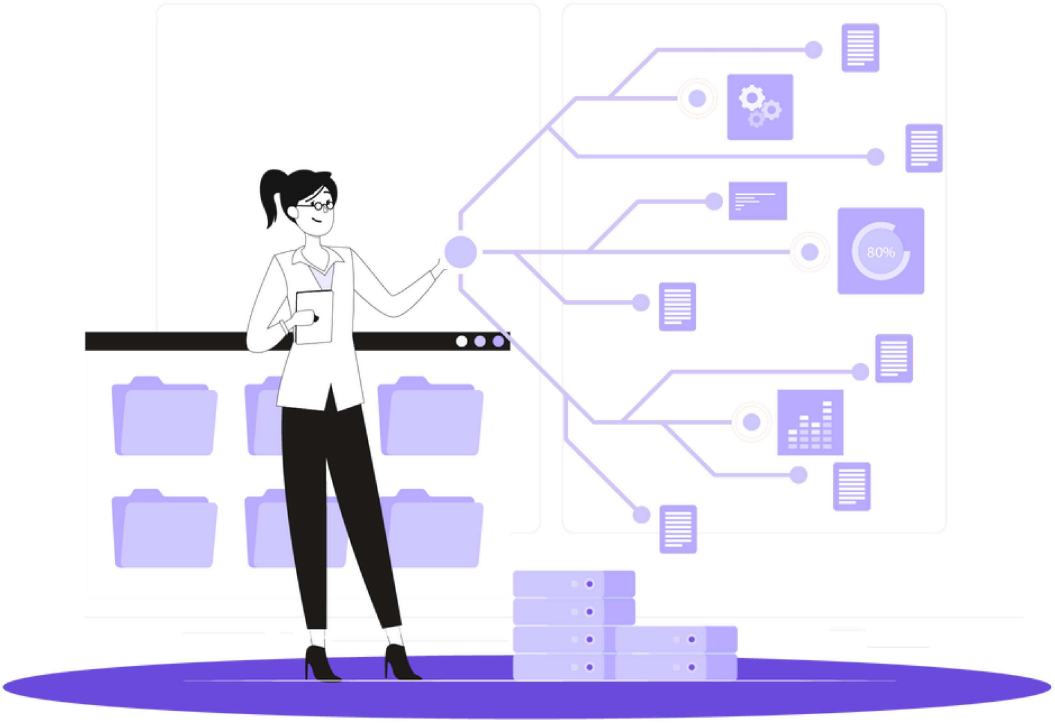
7. Data Cleansing

Data cleansing is the process of identifying and correcting or removing erroneous, incomplete, or inconsistent data from datasets. It helps improve the quality and reliability of data before it is used for analysis or decision-making.

8. Data Modeling

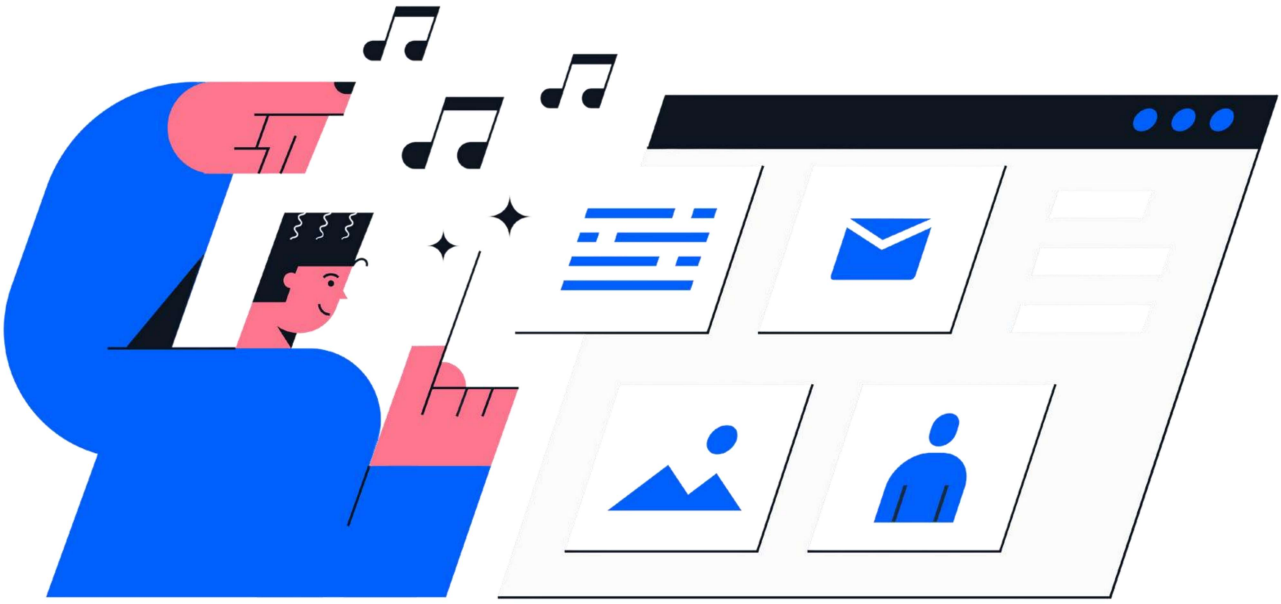
Data modeling involves creating a conceptual representation of data and its relationships. It helps in organizing data into structured formats (e.g., tables, entities, and attributes) that align with business needs, making it easier to store, query, and analyze data effectively.

9. Data Integration



Data integration is the process of combining data from different sources, both internal and external, into a unified view. It ensures that all data is accessible and can be analyzed holistically, improving decision-making and insights across the organization.

10. Data Orchestration

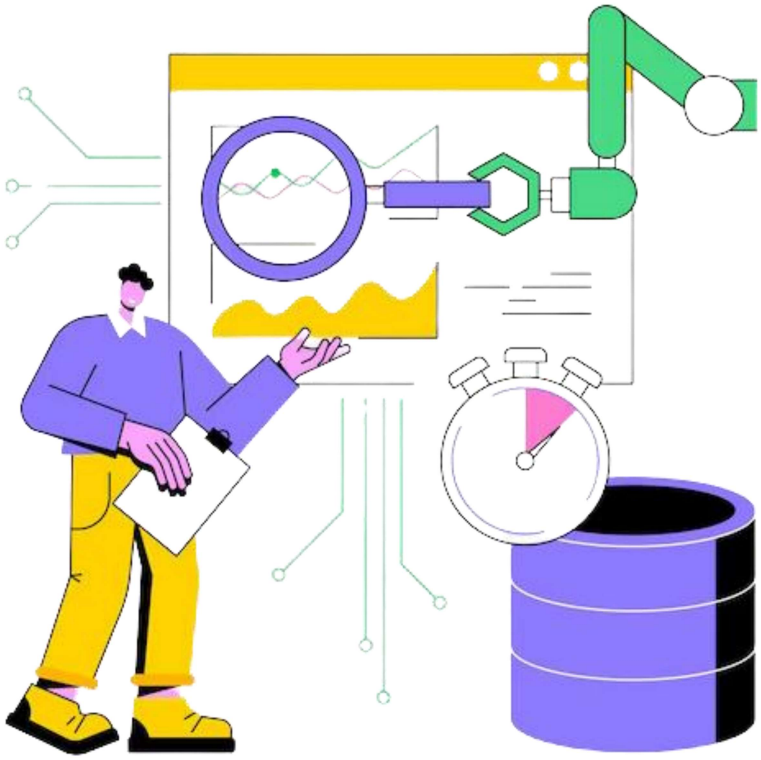


Data orchestration is the automated management and coordination of data workflows across multiple systems. It ensures that data flows smoothly between different stages of a pipeline, including extraction, transformation, and loading, and handles scheduling, dependencies, and error management.

11. Data Transformation

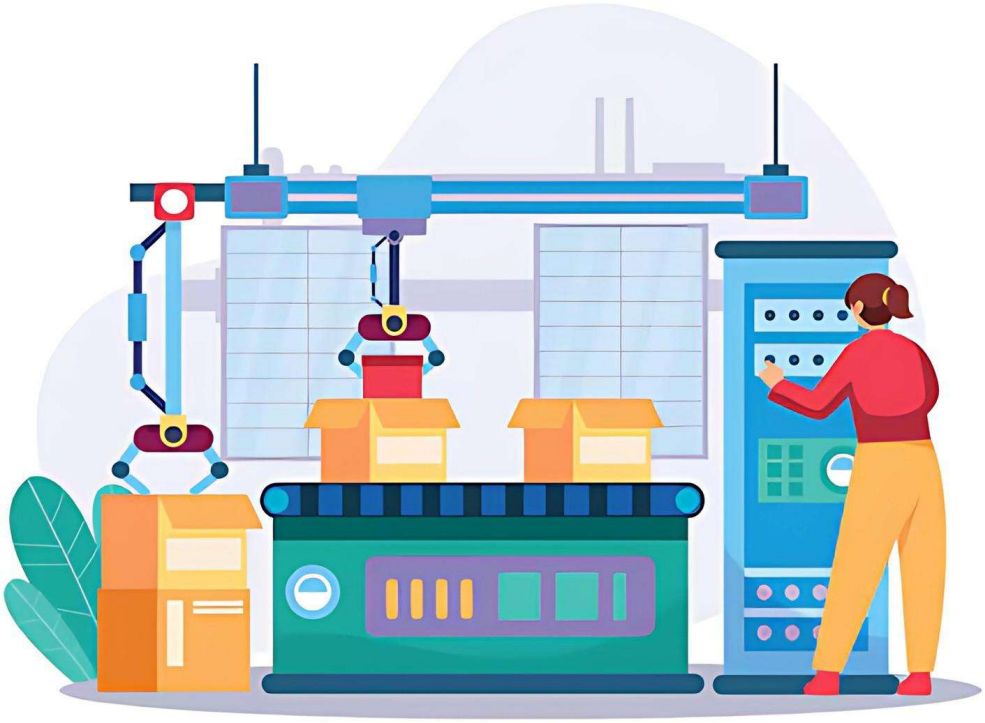
Data transformation refers to modifying data into a desired format or structure for analysis, reporting, or integration with other systems. This can include actions such as aggregating data, filtering, sorting, and converting data types.

12. Real-time Data Processing



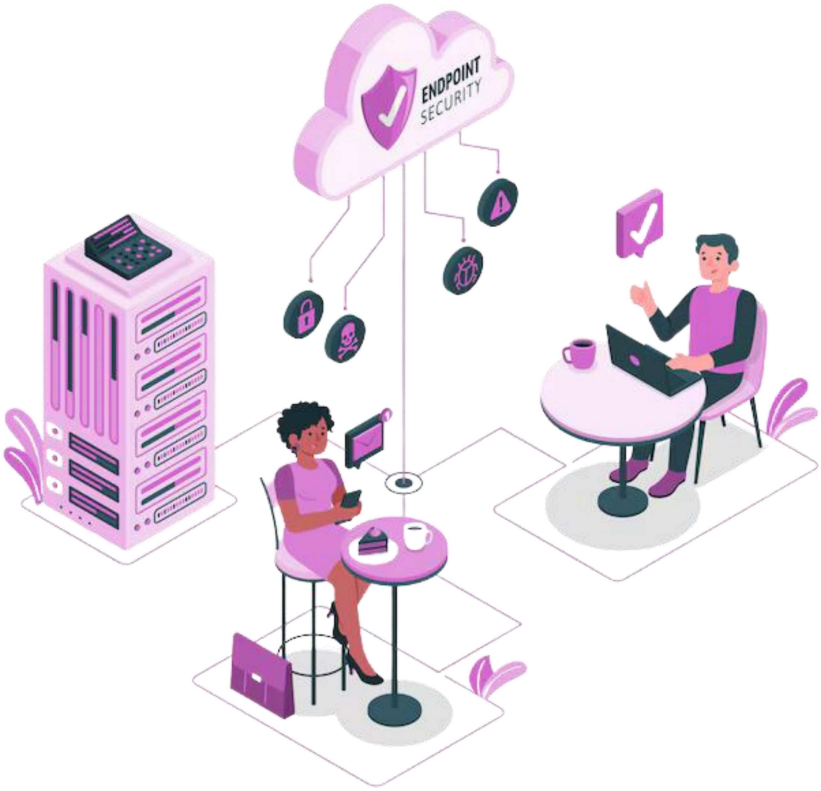
Real-time data processing involves collecting and analyzing data instantly as it is generated, often used for systems such as IoT, social media, or financial trading platforms. It enables quick decision-making and actions based on live data.

13. Batch Processing



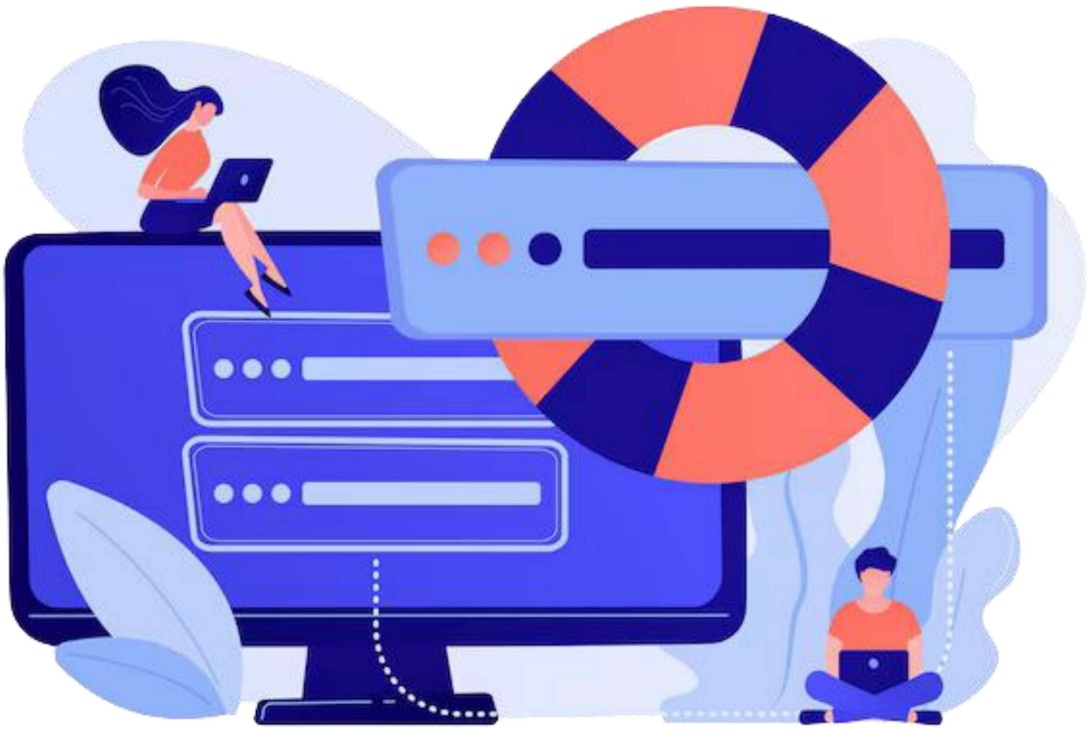
Batch processing refers to processing large volumes of data in chunks, typically at scheduled intervals. This approach is often used when real-time processing isn't required, and it allows for efficient handling of massive datasets, such as overnight data processing.

14. Cloud Data Platform



A cloud data platform is a data storage and analytics solution hosted on cloud services like AWS, Azure, or Google Cloud. These platforms offer scalability, flexibility, and reduced infrastructure management overhead, enabling organizations to store, process, and analyze data from anywhere in the world.

15. Data Sharding



Data sharding is the practice of partitioning a large database into smaller, more manageable pieces called "shards," which can be stored on different servers. Sharding improves performance by distributing the load across multiple systems, allowing for more efficient querying and scaling.

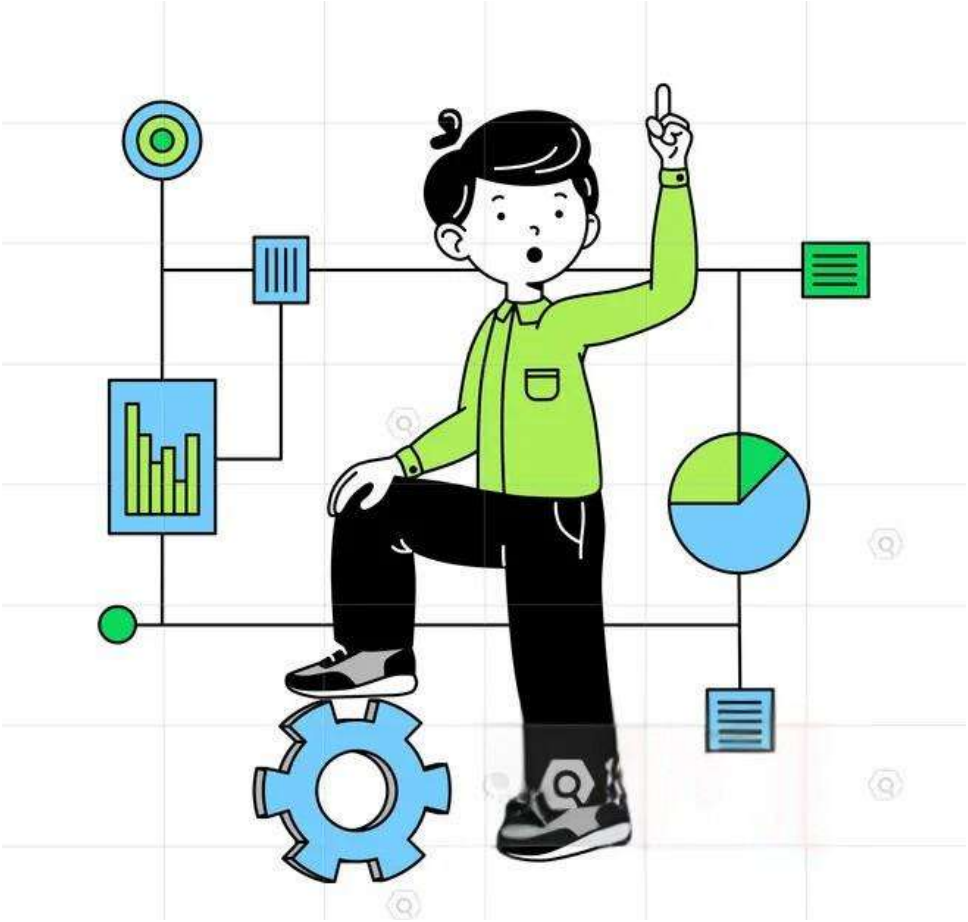
16. Data Partitioning

Data partitioning involves dividing large datasets into smaller, logically separated sections, or partitions. This can enhance performance and manageability by allowing data to be processed and queried in parallel, improving scalability and reducing latency.

17. Data Source

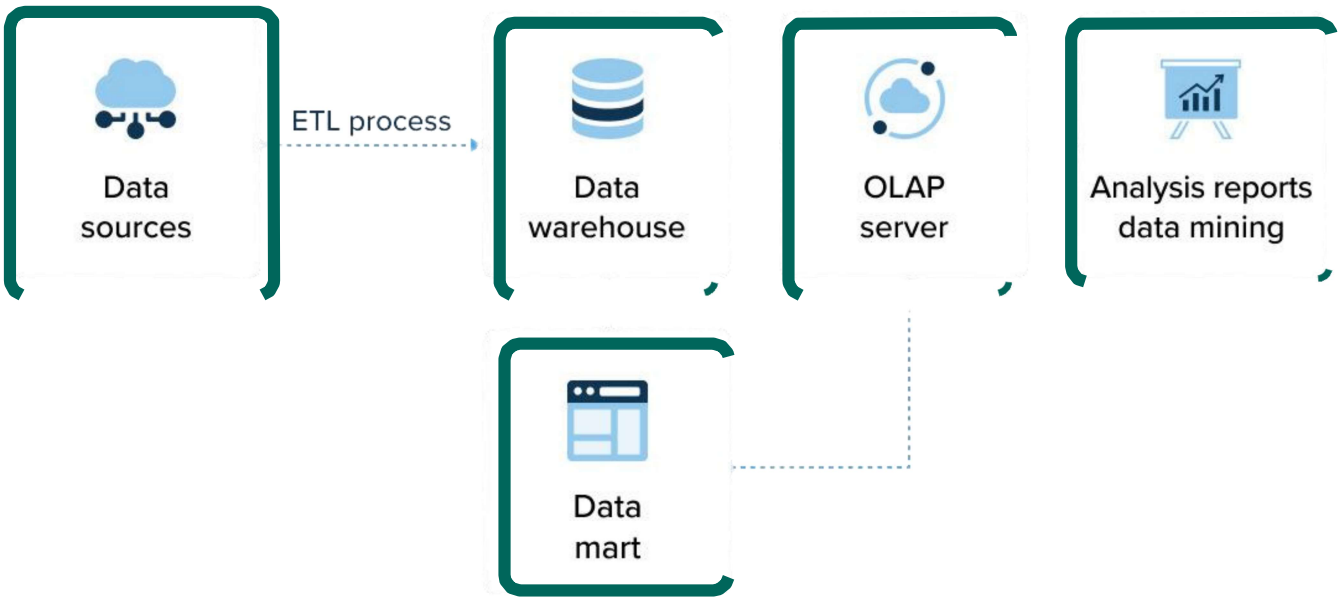
A data source refers to any origin from which data is collected. This can include databases, applications, websites, APIs, or external data providers, each of which contributes raw data for processing and analysis.

18. Data Schema



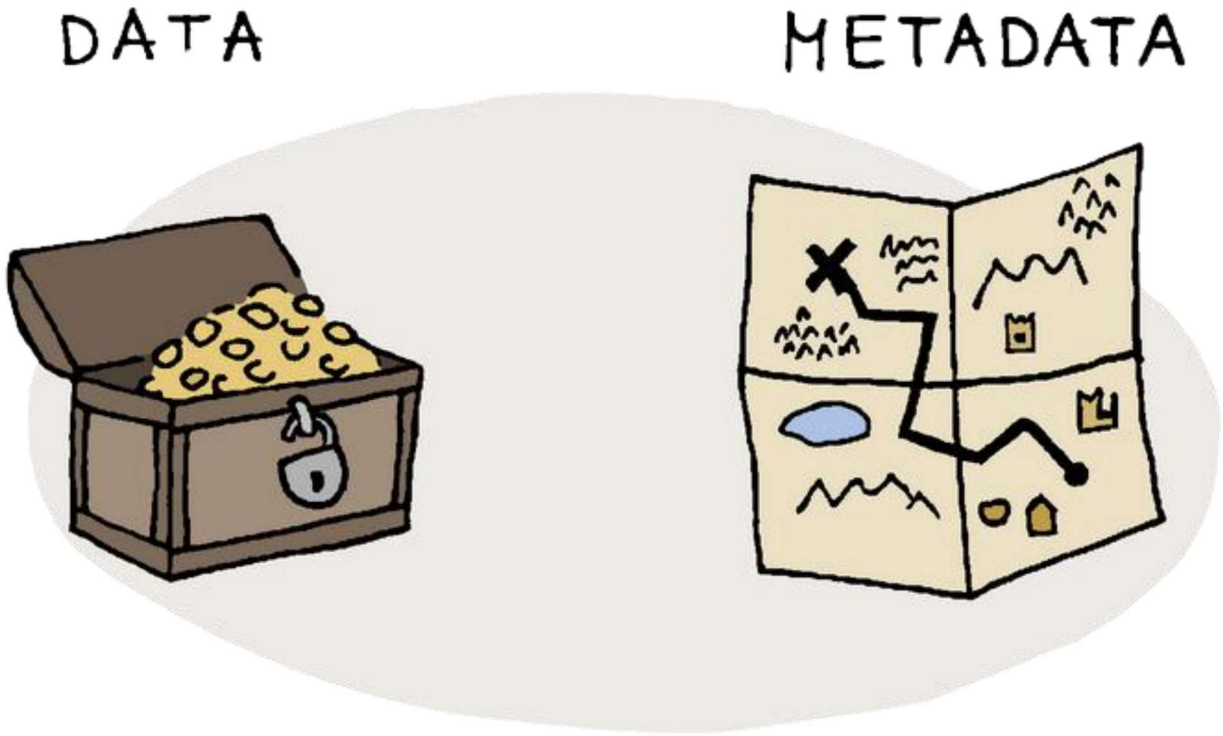
A data schema is a blueprint that defines the structure of a database or data model. It includes specifications such as tables, columns, relationships, constraints, and data types. A schema ensures that data is organized in a logical and efficient manner for querying and storage.

19. DataWarehouse Automation (DWA)



Data Warehouse Automation (DWA) refers to tools and technologies that streamline the creation, management, and maintenance of data warehouses. Automation reduces the need for manual intervention and increases the speed and consistency of data management processes.

20. Metadata



Metadata is "data about data." It provides context and additional information about the structure, source, quality, and relationships of data. Examples include data types, table descriptions, and field names, which help data engineers and analysts understand how to use and interpret the data.