**What is a data pipeline?**

A data pipeline is an automated workflow that extracts data from source systems, transforms it into a usable format, and then loads it into a target data store like a data warehouse. It's a specific type of data pipeline designed for data integration, ensuring data is accurate, consistent, and ready for analysis by moving it through the sequence of Extract, Transform, and Load.

**Data Source:**

This refers to the origin of the raw data. Data sources can be diverse and include:

* + - Databases: Relational databases (e.g., MySQL, PostgreSQL), NoSQL databases (e.g., MongoDB, Cassandra)
    - APIs and Web Services: Data from external applications or services accessed via REST APIs.
    - Files and Logs: CSV files, JSON files, log streams from applications or servers.
    - Event Streams: Real-time data feeds from sources like Apache Kafka or Amazon Kinesis.
    - IoT Devices and Sensors: Data generated by smart devices and sensors.

**Data Ingestion:**

Data ingestion is the process of collecting raw data from the various data sources and feeding it into the pipeline or a staging area. This can occur in two main ways:

* + - Batch Ingestion: Data is collected and processed in batches at scheduled intervals (e.g., daily, hourly). This is suitable for data that doesn't require real-time processing.
    - Streaming Ingestion: Data is collected and processed continuously as it is generated, providing near real-time data availability. This is crucial for applications requiring immediate insights or responses.
  + Ingestion often involves initial validation checks and pre-processing to ensure data quality before further transformation.

**Data Storage:**

After ingestion and often after transformation, the processed data is loaded into a target storage system. The choice of storage depends on the data type, volume, and intended use:

* Data Warehouses: Optimized for storing structured, historical data for analytical queries and business intelligence (e.g., Snowflake, Google BigQuery, Amazon Redshift).
* Data Lakes: Designed to store large volumes of raw, unstructured, and semi-structured data for future analysis and machine learning (e.g., Amazon S3, Azure Data Lake Storage).
* NoSQL Databases: Suitable for specific data models like document, key-value, or graph data, offering flexibility and scalability.
* Relational Databases: Used for structured data where strong consistency and transactional integrity are required.

**The Three Stages of an ETL Data Pipeline**

* 1. **Extract**: This is the initial step where raw data is pulled from various source systems. These sources can include databases (SQL, NoSQL), applications (like CRM systems), files, or APIs.
* 2. **Transform**: In this phase, the extracted data is cleaned, standardized, and structured to match the format and requirements of the target system. This can involve data cleaning (removing duplicates, handling missing values), data aggregation, or applying business rules.
* 3. **Load**: The final step is to load the transformed data into a target repository, such as a data warehouse, data lake, or another database. This centralized location makes the data readily available for reporting, analytics, and business intelligence.

