What is a Data Pipeline?

A data pipeline is a set of automated steps designed to move data from one system to another, often with processing or transformation along the way. Think of it like a factory assembly line raw data enters at one end, goes through a series of steps (like cleaning or formatting), and exits the other end as something useful, ready for analysis or storage.

Working of Data Pipeline:

1. **Data Sources**: These are where the raw data comes from. This could be databases, APIs, files, log files, or sensors in an IoT system. The pipeline starts by accessing or "ingesting" data from these sources.
2. **Ingestion**: This is the process of collecting or receiving the data from the sources. It might involve pulling data on a schedule (like once a day) or listening for new data in real time.
3. **Processing and Transformation**: Once the data is ingested, it usually needs to be cleaned, validated, transformed, or enriched. For example, missing values may be filled in, data types corrected, or new columns created by combining others. This is where the data becomes more usable and meaningful.
4. **Storage**: After processing, the data is stored somewhere for later use such as a data warehouse, data lake, or a database. This is where analysts, engineers, or machine learning systems can access it.
5. **Consumption**: Finally, the processed data is used by other systems or tools. It could feed dashboards, reports, machine learning models, or APIs.
6. **Orchestration and Monitoring**: Most pipelines are automated and run regularly. Tools are used to monitor whether they succeed or fail, retry on errors, and manage dependencies between tasks.

A data pipeline can be either batch-based, meaning it runs at specific intervals (like hourly or daily), or real-time/streaming, meaning it processes data as soon as it arrives.

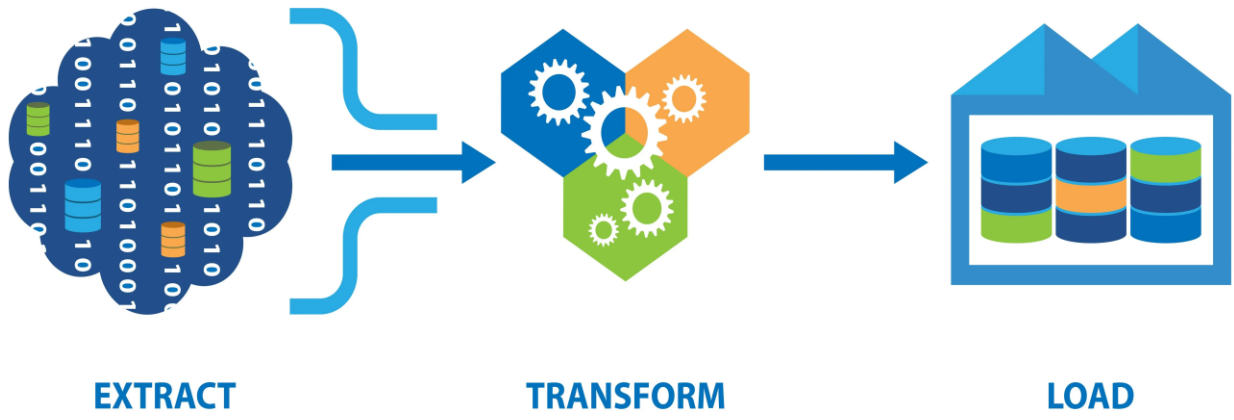
What is ETL?

ETL stands for Extract, Transform, Load. It is a specific kind of data pipeline where the focus is on preparing data before storing it in a central system like a data warehouse.

Let's break down what each of the three steps means:

1. Extract

This is the first step, where data is pulled or received from various source systems. These sources can be very different in type for example, a relational database (like MySQL), a flat file (like a CSV), or an API (like Twitter or Stripe). The goal is to gather all the raw data you need from these different places.



2. Transform

Once the raw data is extracted, it usually isn't ready to use. This step is where the data is processed and made usable. That includes cleaning (like removing duplicates or correcting formats), converting data types, merging

datasets together, filtering irrelevant rows, calculating new columns, or standardizing values (e.g., all dates in the same format).

The transformation step is often the most complex because it's where business rules and logic are applied to make the data meaningful.

3. Load

After transformation, the clean, structured data is loaded into a destination system. Most commonly, this is a data warehouse, which is optimized for analytics. Once in the warehouse, data can be accessed by analysts, reporting tools, or machine learning systems.