DATA PIPELINE

A data pipeline is an automated series of steps that ingest raw data from various sources, process and transform it to make it usable and meaningful, and then deliver it to a destination like a data warehouse for analysis, reporting, or other applications. These pipelines automate the movement, cleaning, filtering, and reformatting of data, removing silos and ensuring data quality and reliability to generate valuable insights for business intelligence and decision-making.

Key Stages of a Data Pipeline:

A typical data pipeline involves the following stages:

1. Ingestion: Data is collected from multiple sources, such as databases, applications, IoT devices, and websites.
2. Processing/Transformation: The raw data is cleaned, filtered, aggregated, and reformatted according to predefined rules to improve its quality and prepare it for its intended use.
3. Storage: The processed and transformed data is loaded into a suitable repository, like a data lake or data warehouse, for centralized storage and easy access.
4. Analysis: Analytical tools and algorithms are applied to the stored data to extract insights, identify trends, and support business intelligence and machine learning tasks.

Why Data Pipelines Are Important

- Automation: Data pipelines automate complex processes, making data movement and preparation efficient and reliable.

- Data Quality: By cleaning and standardizing data, pipelines ensure accuracy and consistency across the organization.

- Integration: They integrate data from disparate sources, breaking down data silos and creating a comprehensive view of information.

- Actionable Insights: They transform raw, often unusable data into meaningful information that can drive informed business decisions.

- Support for Big Data: Pipelines are essential for managing and analyzing the massive volumes of data generated today, powering various data projects.

What is ETL?

ETL stands for **Extract, Transform, Load** — it's a **type of data pipeline** specifically used for preparing data for analysis.

1. Extract

- Data is **collected** from various sources.
- Example: Pulling data from APIs, databases, or CSV files.

2. Transform

- Data is **cleaned, reformatted, and enriched** to make it consistent and usable.
- Examples:
    - Removing duplicates
    - Changing date formats
    - Combining datasets
    - Applying business rules (e.g., converting prices to a single currency)

3. Load

- The final data is **loaded** into a target system — often a **data warehouse** (like Snowflake, BigQuery, or Redshift).
- From here, analysts or dashboards can use the clean data for reporting or machine learning.

Modern Variation — ELT

Modern data systems often use **ELT (Extract → Load → Transform)** instead of ETL.

- Data is **first loaded** into a powerful data warehouse.
- Then **transformation happens inside** the warehouse (using SQL or tools like dbt).
- This is common in cloud-based analytics because it's faster and more scalable.

Example in Real Life

Use case: An e-commerce company

- **Extract:** Get order data from MySQL, customer data from CRM, and marketing data from Google Ads.
- **Transform:** Clean up missing values, unify customer IDs, and calculate total revenue per region.
- **Load:** Store the final dataset in BigQuery for reporting dashboards (e.g., Tableau or Power BI).