

Data Pipeline

A data pipeline is a method in which raw data is ingested from various data sources, transformed and then ported to a data store, such as a data lake or data warehouse, for analysis. Before data flows into a data repository, it usually undergoes some data processing.

It is a series of processes that move data from one system to another — from data sources (like databases, APIs, sensors, or logs) to destinations such as data warehouses, data lakes, or analytical tools.

It ensures that data flows smoothly, reliably, and efficiently through various stages of collection, processing, and storage.

Key Components

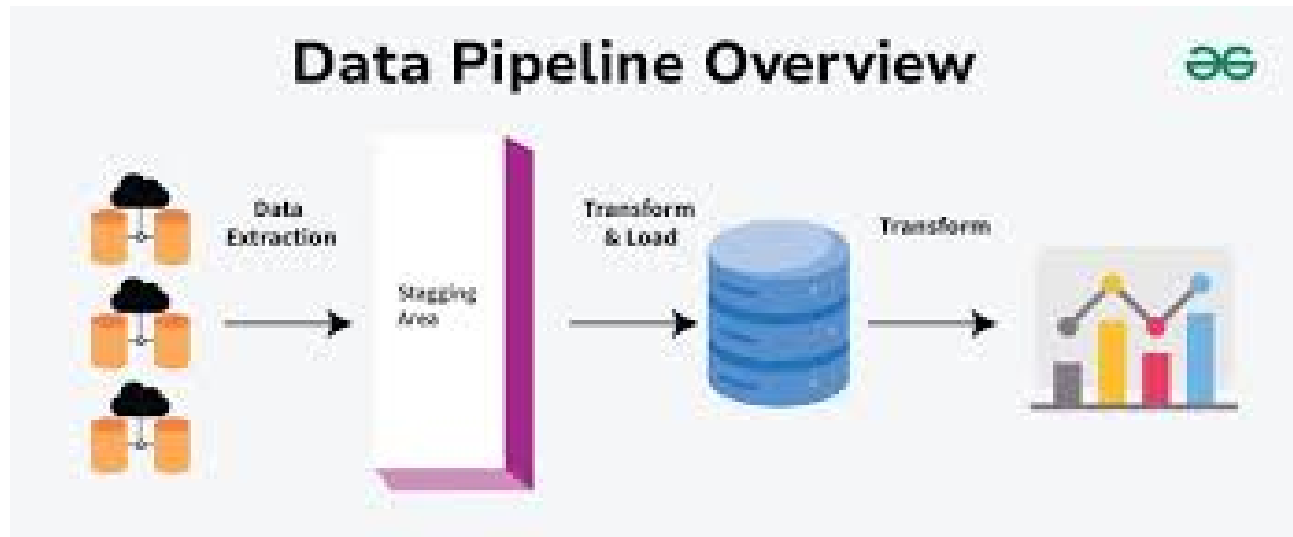
1. Data Source – Origin of data (e.g., SQL databases, web APIs, IoT devices).
2. Ingestion Layer – Collects and imports raw data into the system.
3. Processing Layer – Cleans, transforms, or enriches data for analysis.
4. Storage Layer – Stores processed data in a warehouse or lake.
5. Analytics Layer – Enables reporting, dashboards, or machine learning.

Purpose

- Automates the flow of data across systems.
- Reduces manual data handling and errors.
- Supports real-time and batch data processing.
- Ensures data consistency and quality for business insights.

Example

A company might use a data pipeline to collect sales data from multiple stores, process it daily, and load it into a data warehouse like Snowflake for dashboard reporting.



ETL (Extract, Transform, Load)

It is a key process in data warehousing that prepares data for analysis. It involves:

- Extracting data from multiple sources
- Transforming it into a consistent format
- Loading it into a central data warehouse or data lake

ETL helps businesses unify and clean data, making it reliable and ready for analysis. It improves data quality, security, and accessibility, enabling better insights and faster decision-making in a world of diverse data sources.

ETL (Extract, Transform, Load)

ETL is a specific type of data pipeline that focuses on preparing and loading data into a storage system. It has three main steps:

1. Extract – Retrieve data from various sources (databases, APIs, files).
2. Transform – Clean, filter, join, or aggregate the data to match business rules and schema requirements.
3. Load – Store the processed data into a target system (e.g., data warehouse).

Why ETL is Important

- Ensures data quality and consistency before analysis.
- Enables centralized reporting by integrating data from multiple sources.
- Supports data-driven decision-making through structured and reliable datasets.

ETL Tools

- Open Source: Apache Airflow, Talend, Luigi.
- Cloud-Based: AWS Glue, Google Dataflow, Azure Data Factory.
- Commercial: Informatica, Fivetran, Matillion.

