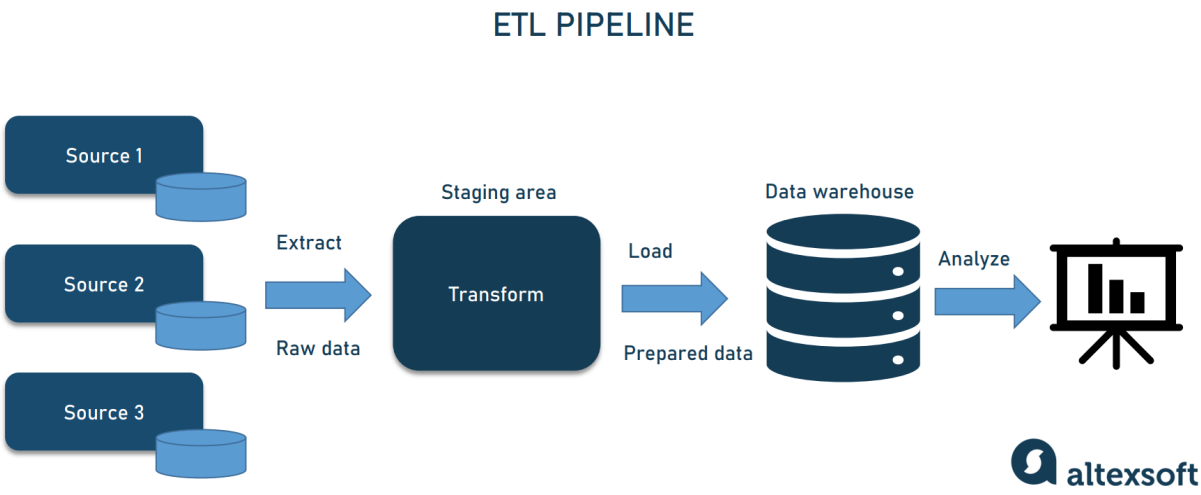
Data Pipelining



What is a Data Pipeline?

A **data pipeline** is a set of automated processes for moving, transforming, and managing data from one or more sources to a specific destination. It acts as the "circulatory system" for an organization's data, ensuring that raw data is efficiently and reliably processed into a refined, usable format for analysis, business intelligence, and other applications.

The goal of a data pipeline is to ensure that data is high-quality, consistent, and available for consumption.

Key Components & Stages

- **Data Sources**: The origin points of the data, which can be diverse. Examples include databases (SQL, NoSQL), APIs, log files, IoT devices, web applications, and external data streams.
- **Data Ingestion (Extract/Collect)**: The process of collecting or capturing the raw data from its sources and bringing it into the pipeline. This can happen in two main ways:
    - **Batch Processing**: Data is collected and moved in large volumes at scheduled intervals (e.g., hourly or nightly).
    - **Streaming/Real-time Processing**: Data is processed continuously as it is generated, often used for immediate insights (e.g., stock market data)
- **Data Processing/Transformation (Transform)**: Once ingested, the raw data undergoes operations to clean, standardize, enrich, and convert it into a usable format
- **Data Destination (Load/Store)**: The final target where the processed and transformed data is stored, ready for consumption. This is typically a centralized repository like a **data**

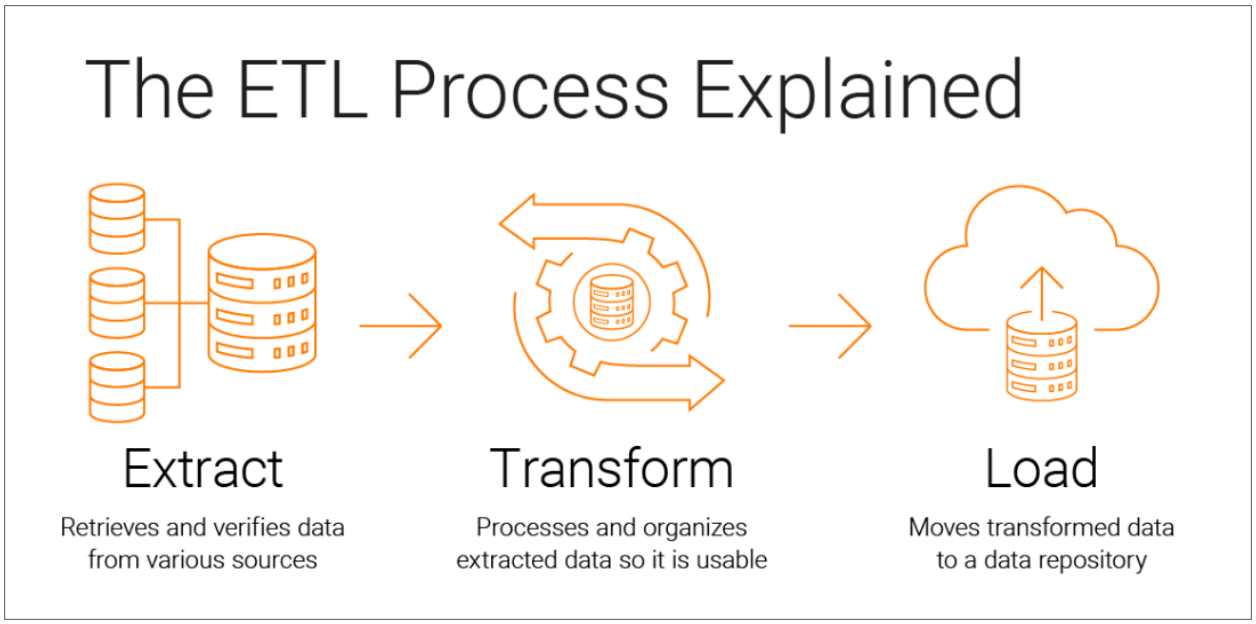
warehouse (optimized for analysis) or a **data lake** (for storing large volumes of raw and processed data).
- **Orchestration and Monitoring**: The management layer that schedules, coordinates, and manages the data flow, ensuring all steps run in the correct order and handling dependencies, errors, and performance tracking.

Common Types of Data Pipeline

Data pipelines are often categorized based on the order of their main processes:

- **ETL (Extract, Transform, Load)**: This traditional pattern extracts data, transforms it in a staging area, and then loads the refined data into the destination (typically a data warehouse).
- **ELT (Extract, Load, Transform)**: This pattern extracts data and loads the raw data directly into a powerful cloud-based data lake or warehouse, where the transformation step is then performed. This is often preferred in modern cloud environments because cloud storage and compute resources make it faster and cheaper to load first.

How ETL works?



The **Extract, Transform, Load (ETL)** process is a traditional data integration method that prepares data for analysis and storage, typically in a data warehouse.

Extract: This is the first step of ETL where the data is extracted from various sources.
The goal is to retrieve all the required data and move it to a temporary **staging area** (a safe, intermediate storage location) without altering the raw data. Data sources include relational databases, applications (like CRM or ERP), APIs, log files, and flat files. Extraction can be a **full**

extraction (retrieving all data) or an **incremental extraction** (retrieving only the data that has changed since the last run), with the latter being more efficient for continuous updates.

Transform: In this phase, the data is modified, cleansed, and converted into a consistent, analysis-ready format that matches the schema of the target destination. This occurs in the staging area **before** the data is loaded. The goal is to apply business rules, ensure data quality, and prepare the data for efficient querying.
Key data transformation include:

- **Cleansing:** Fixing errors, filling in missing values (nulls), and removing duplicates.
- **Standardization:** Converting data formats (e.g., date formats, currency units) to be consistent across all sources.
- **Derivation:** Calculating new values from existing data (e.g., calculating total profit).
- **Joining/Integration:** Combining data from multiple sources (e.g., matching customer records from a sales database and a marketing application).
- **Filtering:** Removing irrelevant rows or columns.

Load: In the final step, the transformed and validated data is **loaded** into the final target system, usually a **data warehouse** or **data lake**. The goal is to make the refined data available for business intelligence (BI) tools, reporting, and analysis.
Methods:

- **Full Load:** The entire transformed dataset is loaded into empty tables or overwrites existing data (typically for the first load).
- **Incremental Load:** Only the new or changed data records are added to the existing tables, which is the standard approach for continuous updates.