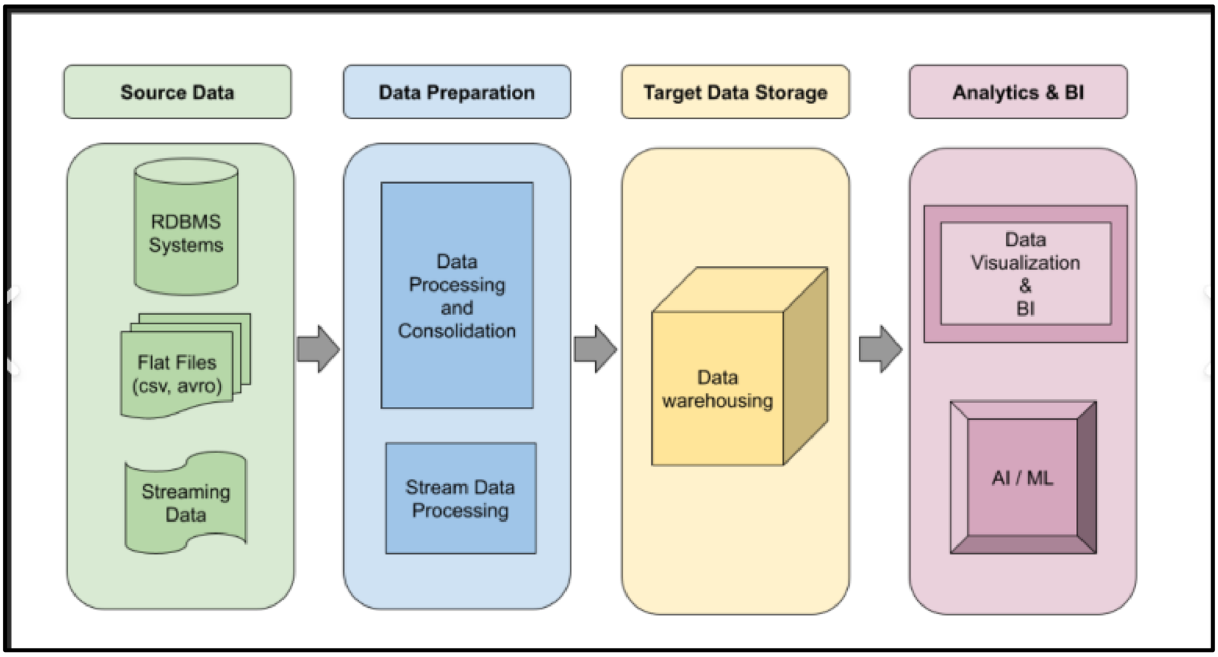
Data Pipelines and the ETL Process

What is a Data Pipeline?

A **data pipeline** is a series of steps that automate the flow of data from one system to another. It ensures that data is collected, processed, and delivered efficiently and reliably.



Key Components:

- **Source**: Where data originates (e.g., databases, APIs, files).
- **Processing**: Transforming, cleaning, or enriching the data.
- **Destination**: Where data is stored or consumed (e.g., data warehouse, dashboard).

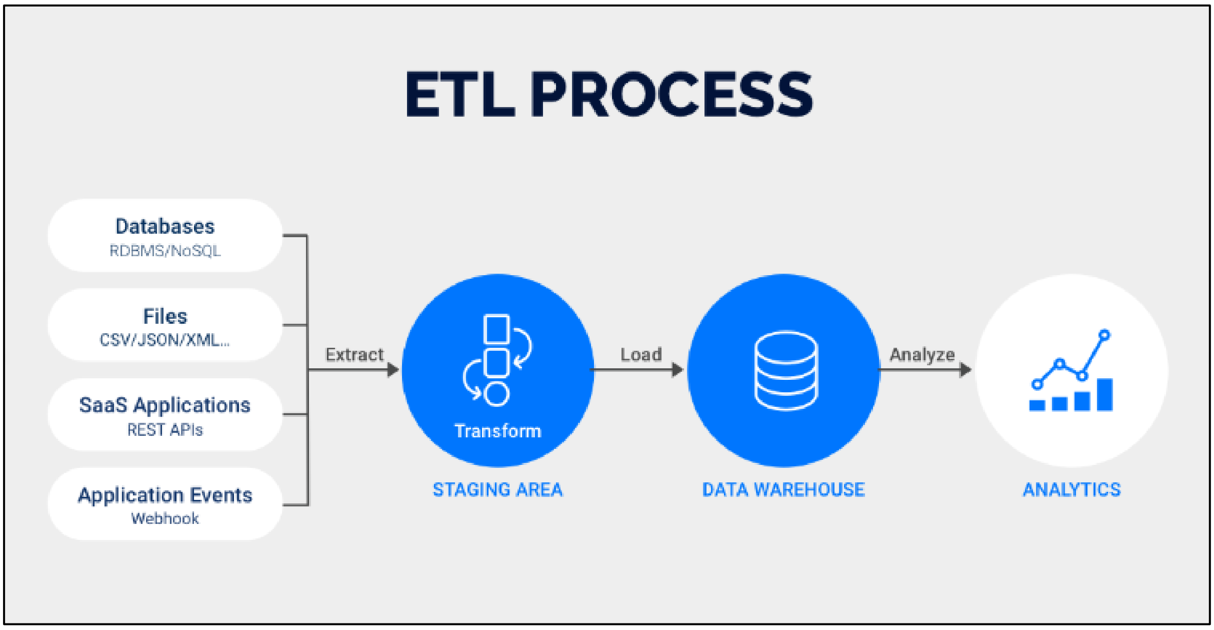
Why it matters:

Data pipelines are the backbone of modern data-driven systems. They enable real-time analytics, machine learning, and business intelligence.

s

ETL: Extract, Transform, Load

ETL is a classic data pipeline pattern used to move data from source to destination.



1. Extract

- Pull data from various sources.
- Examples: SQL databases, CSV files, REST APIs.

2. Transform

- Clean, format, and enrich the data.
- Examples: Removing duplicates, converting date formats, aggregating metrics.

3. Load

- Push the transformed data into a target system.
- Examples: Data warehouse (Snowflake, BigQuery), dashboards (Power BI, Tableau).

ETL vs ELT

Feature	ETL (Traditional)	ELT (Modern)
Transformation	Before loading	After loading
Speed	Slower for big data	Faster with cloud systems
Tools	Informatica, Talend	dbt, BigQuery, Snowflake

Real-World Example

Imagine an e-commerce company:

- **Extract**: Pulls customer orders from MySQL.
- **Transform**: Cleans data, calculates total spend.
- **Load**: Sends it to a dashboard for sales analysis.

Tools & Technologies

- **Apache Airflow** – Workflow orchestration
- **Kafka** – Real-time data streaming
- **dbt** – SQL-based transformation
- **Snowflake / BigQuery** – Cloud data warehouses

Final Thoughts

A well-designed data pipeline ensures:

- **Scalability**: Handles growing data volumes.
- **Reliability**: Minimizes data loss or corruption.
- **Efficiency**: Automates repetitive tasks.