

ETL VS ELT

What is ETL?

ETL (Extract, Transform, Load) is a data integration process involving three key steps:

- **Extract:** Data is extracted from various source systems.
- **Transform:** Data is transformed (cleaned, aggregated, enriched) according to business rules.
- **Load:** Transformed data is loaded into a target data warehouse or database.

What is ELT?

ELT (Extract, Load, Transform) is another data integration process, but the steps occur in a different order than ETL:

- **Extract:** Data is extracted from various source systems.
- **Load:** Raw data is loaded into a target data warehouse or database.
- **Transform:** Data is transformed within the data warehouse using its processing power.

The main difference between both is that in **ETL**, transformation takes place outside the data warehouse while in **ELT**, the transformation takes place inside the data warehouse.

When to choose ETL or ELT?

Choose ETL:

- When you need to perform complex transformations before loading data into the data warehouse, working with traditional on-premises data environments, and data consistency and cleaning are crucial before storage

Choose ELT:

- When working with large datasets in a cloud-based data warehouse, the data warehouse can efficiently handle transformations, and you want to load raw data quickly and perform transformations on demand.

For example, A retail company wants to integrate sales data from multiple sources (e.g., POS systems, online sales platforms) into a centralized data warehouse for reporting and analysis.

Reason for choosing ETL:

- The company needs to perform complex data transformations, including cleaning data inconsistencies, deduplicating records, and aggregating sales data by region and product category before loading it into the data warehouse.
- Ensures that only cleaned and processed data is stored in the data warehouse, maintaining data quality and consistency.

Batch Pipelines vs. Stream Pipelines

Batch Pipelines process data in large, scheduled chunks. They are ideal for historical analysis and reporting, offering higher latency. **Streaming Pipelines** handle data continuously, in real-time, providing crucial immediate insights and actions with low latency.

The main differences between batch and stream pipelines are in processing time, volume handling, complexity, resource utilization, and state management. Batch pipelines are suitable for large, scheduled data processing tasks with higher latency, while stream pipelines are ideal for real-time processing requiring immediate insights and low latency.

Cloud platforms like AWS, GCP, and Azure offer specific services tailored to each type of pipeline, allowing organizations to choose the best tools for their data processing needs.

Batch Pipeline Scenario:

A retail company generates daily sales reports. Sales data from various sources, such as POS systems and e-commerce platforms, is collected throughout the day and processed overnight to produce a comprehensive sales report for the previous day.

Stream Pipeline Scenario:

A financial services company monitors and analyzes transaction data in real-time to detect and prevent fraudulent activities. Transaction data is processed as it is generated, allowing for immediate detection of suspicious patterns and triggering alerts for further investigation.