EXTRACT, TRANSFORM, and LOAD(ETL)

INTRODUCTION

ETL extracts the data from different data sources, transforms the data, and at last, loads the data into the destination place such as Data Warehouse, repository. You know about ML pipeline, and similar to that, ETL is also pipeline, known as Data pipeline, which includes three steps, i.e., Extract, Transform, and Load. The diagram below shows the working process pipeline of ETL.

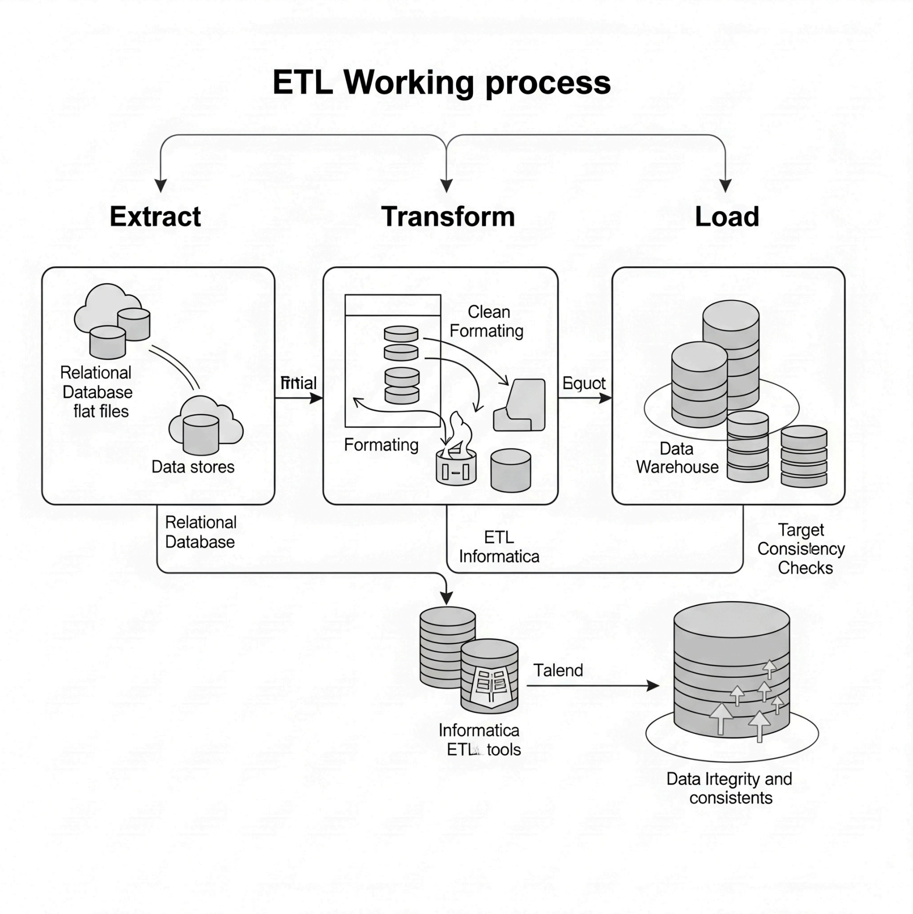

Figure 1 ETL working process Pipeline

WHY ETL?

ETL provides a lot of benefits to both organizations and data engineers/scientists. It provides quality data sets for ML projects. It also provides quality data so that companies/organizations can make critical decisions. It merges the data systems from different departments or companies in a single, reliable repository.

As ETL comes from Data Warehousing, it provides a method of moving from various sources into a data warehouse(a common repository). It allows sample data comparison between the source and the destination. It also performs complex transformations and hence, requires the extra area to store the data. It also offers a deep historical context for the business.

ETL PROCESS

As shown in the above diagram, the ETL process consists of three steps, i.e., Extract, Transform, and Load. Now, you learn what these steps are and how they work.

Rojesh Shikhrakar ✓
Apr 17, 2020

Need Graphical presentation

EXTRACT

This step covers the data extraction from different sources like databases, applications, files with as little resource utilization as possible. While implementing this step, we should also consider that it does not negatively affect the source in terms of performance, response time, or any kind of locking.

You can perform this step using different methods. Some of them are:

1. Full Extraction
2. Partial Extraction- without update notification
3. Partial Extraction- with update notification

You also need to do certain validation during extraction such as remove all types of duplicate/fragmented data, make sure that no spam/unwanted data loaded, data type check.

TRANSFORM

In this step, you transform the extracted data obtained using the extract process. This step transforms using transformation processes the data into the format required in the destination or target place. Transformation processes applied during this step to ensure the quality and integrity of data are:

- Basic transformations:

  Some of the basic transformations are:

  - Cleaning
  - Format revision
  - Data threshold validation checks
  - Restructuring
  - Deduplication
- Advanced Transformations:

  Some of the advanced transformations are:

  - Filtering
  - Merging
  - Splitting
  - Derivation
  - Summarization
  - Integration
  - Aggregation
  - Complex data validation

LOAD

This step is the last process in the ETL process. Here, you load huge transformed data into the final destination place. It is necessary to ensure that we load data correctly. In the case of load failure, we use the recovery mechanisms from the point of failure without data integrity loss. You can load data in different ways, and some of them are:

- Initial Load
- Incremental Load
- Full Refresh

ETL TOOLS

Some of the ETL tools available in the market are:

- Improvado
- MarkLogic
- Oracle
- Amazon RedShift

REFERENCES

1. ETL
2. ETL (Extract-Transform-Load)
3. Extract Transform Load [ETL]

Rojesh Shikhrakar ✓
Apr 17, 2020

remove references

add more details in the current content above.