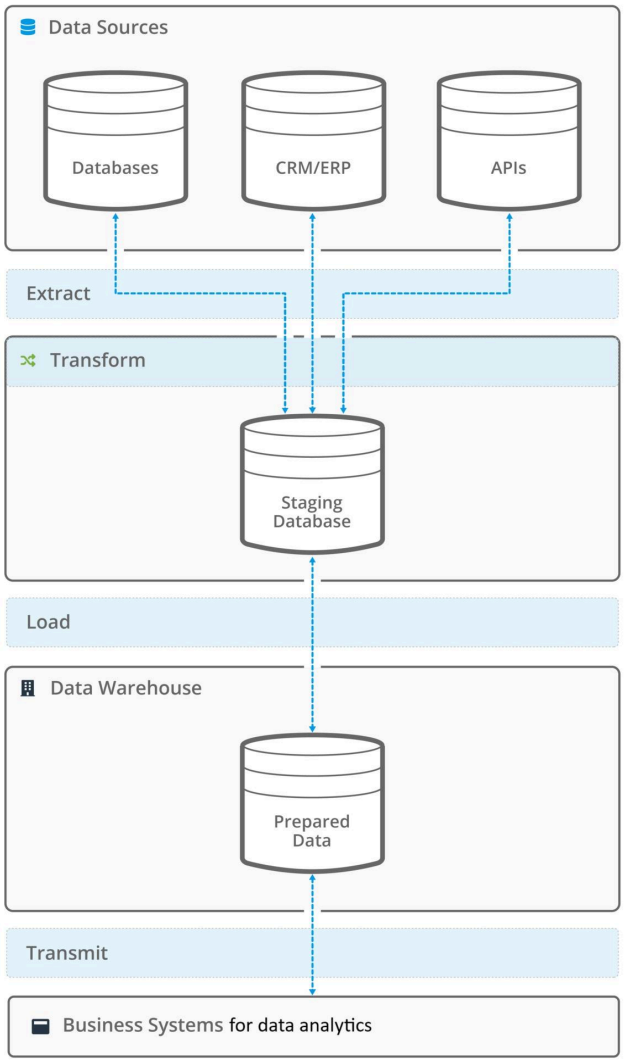
What is ETL?

As the abbreviation implies, there are three basic steps in this process:

1. Extraction: Identifying and capturing data from various sources
2. Transformation: The "cleaning" or standardization of data to meet the needs of the business or organization that wants to use it (e.g., converting date formats, removing duplicates)
3. Load: The loading of data into a new site or "data warehouse" that serves the business or organization's needs

Diagram



Prerequisites for ETL - Data Profiling:

Data profiling examines the quality, scope and context of the source data and enables the ETL team to build an effective ETL system.

Data profiling process reads the source data and generates a comprehensive report on-

1) Data types of each field

2) Natural keys

3) Relationships between tables

4) Data statistics like maximum values, minimum values, most occurred values, number of occurrences of each value etc...

5) Dates in non-date fields

6) Data anomalies like junk values, values outside a given range, missing values etc...

7) Null values

ETL details

1. Extract:
    a. The first step in the ETL process is data extraction.
    b. The data obtained can be structured, semi-structured or unstructured.
    c. This involves gathering raw data from various data sources using suitable APIs wherever needed.
    d. Examples of sources are flat files, databases, web-scraping, an enterprise resource planning (ERP) system or Customer Relationship Management (CRM) system or an Excel spreadsheet.
2. Transform:
    a. Data cleaning
        i. Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
    b. Data integration
    c. Data reduction
        i. Dimensionality reduction
        ii. Data compression

    d. Data transformation and data discretization
        i. Normalization
        ii. Concept hierarchy generation

3. Load:
    a. The final step of the ETL process is data loading.
    b. This is where the transformed data is loaded into the target business system
    c. Systems may include a data warehouse, data lake, or even a simple Excel spreadsheet.

Why is the ETL process important?

- The shorthand phrase “garbage in, garbage out” is used often in tech roles—and it speaks directly to why the ETL process is important.
- The validity and value of data analysis depend on having good inputs, and the ETL process is used to clear out as much of the noise in a collection of data as possible.

Examples of ETL systems:

- AWS Glue
- Talend Open Studio
- Oracle Data Integrator
- Infosphere Datastage
- Apache Airflow