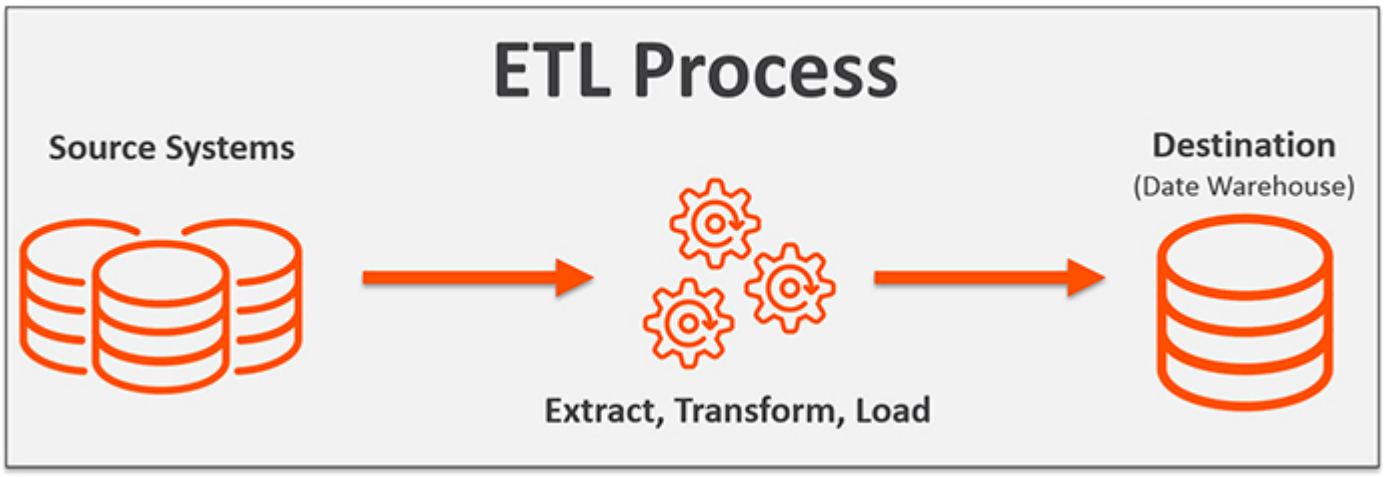
ETL BASICS: EXTRACT, TRANSFORM & LOAD

The process of extracting data from multiple source systems, transforming it to suit business needs, and loading it into a destination database is commonly called ETL, which stands for extraction, transformation, and loading. While ETL is usually explained as three distinct steps, this actually simplifies it too much as it is truly a broad process that requires a variety of actions.

ETL first saw a rise in popularity during the 1970s, when organizations began to use multiple databases to store their information. It quickly became the standard method for taking data from separate sources, transforming it, and loading it to a destination. A few decades later, data warehouses became the next big thing, providing a distinct database that integrated information from multiple systems. In order to accommodate our ever-changing world of digital technology in recent years, the number of data systems, sources, and formats has exponentially increased, but the need for ETL has remained just as important for an organization's broader data integration strategy.

ETL Basics

The following tasks are the main actions that happen in the ETL process:



Extraction of Data

The first step in ETL is extraction. During extraction, data is specifically identified and then taken from many different locations, referred to as the Source. The Source can be a variety of things, such as files, spreadsheets, database tables, a pipe, etc. It is not typically possible to pinpoint the exact subset of interest, so more data than necessary is extracted to ensure it covers everything needed. The volume of data extracted greatly varies and depends on business needs and requirements. Some extractions consist of hundreds of kilobytes all the way up to gigabytes. This is also the case for the timespan between two extractions; some may vary between days or hours to almost real-time.

Data extraction most typically occurs in one of three ways:

1. Update notification - the system notifies you when a record has been changed. This is typically referred to as the easiest method of extraction.
2. Incremental extraction - some systems cannot provide notifications for updates, so they identify when records have been modified and provide an extract on those specific records
3. Full extraction - some systems aren't able to identify when data has been changed at all, so the only way to get it out of the system is to reload it all. This is usually only recommended for small amounts of data as a last resort

Transformation of Data

The next step in the ETL process is transformation. After data is extracted, it must be physically transported to the target destination and converted into the appropriate format. This data transformation may include operations such as cleaning, joining, and validating data or generating calculated data based on existing values.

Whether the transformation takes place in the data warehouse or beforehand, there are both common and advanced transformation types that prepare data for analysis. Some of these include:

- Basic transformations:
    - Cleaning
    - Format revision
    - Restructuring
    - Deduplication

- Advanced transformations:
    - Filtering
    - Joining
    - Splitting
    - Derivation
    - Summarization
    - Integration

Loading Data

The final step in the ETL process involves loading the transformed data into the destination target. This target may be a database or a data warehouse. There are two primary methods for loading data into a warehouse: full load and incremental load. The full load method involves an entire data dump that occurs the first time the source is loaded into the warehouse. The incremental load, on the other hand, takes place at regular intervals. These intervals can be streaming increments (better for smaller data volumes) or batch increments (better for larger data volumes).

ETL in Data Warehouses

For a majority of companies, it is extremely likely that they will have years and years of data and information that needs to be stored. In order to consolidate all of this historical data, they will typically set up a data warehouse where all of their separate systems end up. Combining all of this information into one place allows easy reporting, planning, data mining, etc. Due to the fact that all of the data sources are different, as well as the specific format that the data is in may vary, their next step is to organize an ETL system that helps convert and manage the data flow.

In order to keep everything up-to-date for accurate business analysis, it is important that you load your data warehouse regularly. This means that all operational systems need to be extracted and copied into the data warehouse where they can be integrated, rearranged, and consolidated, creating a new type of unified information base for reports and reviews.

ETL Tools for Data Warehouses

While you can design and maintain your own ETL process, it is usually considered one of the most challenging and resource-intensive parts of the data warehouse project, requiring a lot of time and labor. Many organizations utilize ETL tools that assist with the process, providing capabilities and advantages unavailable if you were to complete it on your own. These tools can not only support with the extraction, transformation and loading process, but they can also help in designing the data warehouse and managing the data flow.

Who uses ETL?

ETL tools are often visual design tools that allow companies to build the program visually, versus just with programming techniques. For the most part, enterprises and companies that need to build and maintain complex data warehouses will invest in ETL and ETL tools, but other organizations may utilize them on a smaller scale, as well.

Why use ETL?

Since it was first introduced almost 50 years ago, businesses have relied on the ETL process to get a consolidated view of their data. ETL allows organizations to analyze data that resides in multiple locations in a variety of formats, streamlining the reviewing process and driving better business decisions.

Benefits of ETL

- Transforms data from multiple sources and loads it into various targets
- Provides deep historical context for businesses
- Allows organizations to analyze and report on data more efficiently and easily
- Increases productivity as it quickly moves data without requiring the technical skills of having to code it first
- Evolves and adapts to changing technology and integration guidelines