What is ETL?

ETL—meaning extract, transform, load—is a [data integration](https://www.ibm.com/topics/data-integration) process that combines, cleans and organizes data from multiple sources into a single, consistent data set for storage in a [data warehouse](https://www.ibm.com/topics/data-warehouse), [data lake](https://www.ibm.com/topics/data-lake) or other target system.

ETL data pipelines provide the foundation for data analytics and [machine learning](https://www.ibm.com/topics/machine-learning) workstreams. ETL pipelines are often used by organizations to:

* Extract data from legacy systems
* Cleanse the data to improve data quality and establish consistency
* Load data into a target database

**How ETL works**

The easiest way to understand how ETL works is to understand what happens in each step of the process.

Extract

During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of different sources, which can be structured or unstructured. Those data types include, but are not limited to:

* SQL or [NoSQL](https://www.ibm.com/topics/nosql-databases) servers
* CRM and ERP systems
* JSON and XML
* Flat-file databases
* Email
* Web pages

**Transform**

* Filtering, cleansing, aggregating, de-duplicating, validating and authenticating the data.
* Performing calculations, translations or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings and more.
* Conducting audits to ensure data quality and compliance, and computing metrics.
* Removing, encrypting or protecting data governed by industry or governmental regulators.
* Formatting the data into tables or joined tables to match the schema of the target data warehouse.

**Load**

In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous and batch-driven. Typically, the ETL load process takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest.

**ETL and other data integration methods**

ETL and ELT are just two data integration methods, and there are other approaches that are also used to facilitate data integration workflows. Some of these include:

* **Change Data Capture (CDC)**identifies and captures only the source data that has changed and moves that data to the target system. CDC can be used to reduce the resources required during the ETL “extract” step; it can also be used independently to move data that has been transformed into a data lake or other repository in real-time.
* **Data replication**copies changes in data sources in real-time or in batches to a central database. [Data replication](https://www.ibm.com/products/data-replication) is often listed as a data integration method. In fact, it is most often used to create backups for [disaster recovery](https://www.ibm.com/topics/disaster-recovery).
* **Data virtualization** uses a software abstraction layer to create a unified, integrated, fully usable *view* of data—without physically copying, transforming or loading the source data to a target system. Data virtualization functions enable an organization to create virtual data warehouses, data lakes and data marts from the same source data for data storage without the expense and complexity of building and managing separate platforms for each. While data virtualization can be used alongside ETL, it is increasingly seen as an alternative to ETL and to other physical data integration methods.
* **Stream Data Integration (SDI)** is just what it sounds like—it continuously consumes data streams in real time, transforms them, and loads them to a target system for analysis. The keyword here is *continuously*. Instead of integrating snapshots of data extracted from various sources at a given time, SDI integrates data constantly as it becomes available. SDI enables a data store for powering analytics, machine learning and real-time applications for improving customer experience, fraud detection and more.

**The benefits and challenges of ETL**

ETL solutions improve quality by performing data cleansing before loading the data to a different repository. A time-consuming batch operation, ETL is recommended more often for creating smaller target data repositories that require less frequent updating, while other data integration methods—including ELT (extract, load, transform), change data capture (CDC) and data virtualization—are used to integrate increasingly larger volumes of data that changes or real-time data streams.

**ETL tools**

In the past, organizations wrote their own ETL code. There are now many open source and commercial ETL tools and cloud-based services to choose from. Typical capabilities of these products include:

* **Comprehensive automation and ease of use**: Leading ETL tools automate the entire data flow, from data sources to the target data warehouse. This saves data engineers from the tedious tasks of moving and formatting data—for faster results and more efficient operations.
* **A visual, drag-and-drop interface**: This functionality can be used for specifying rules and data flows.
* **Support for complex data management**: This includes assistance with complex calculations, data integrations and string manipulations.
* **Security and compliance**: The best ETL tools encrypt data both in motion and at rest, and are certified compliant with industry or government regulations, including HIPAA and GDPR.