Home / Topics / What is ETL (Extract, Transform, Load)?

What is ETL (extract, transform, load)?

Explore IBM's ETL solution →

Subscribe to AI topic updates →

Let’s talk

What is ETL?

Let’s talk

ETL—meaning extract, transform, load—is a data integration process that combines, cleans and organizes data from multiple sources into a single, consistent data set for storage in a data warehouse, data lake or other target system.

ETL data pipelines provide the foundation for data analytics and machine learning workstreams. Through a series of business rules, ETL cleanses and organizes data to address specific business intelligence needs, such as monthly reporting—but it can also tackle more advanced analytics, which can improve back-end processes and end-user experiences. 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

Let's talk

Related content

Register for 2023 Gartner Magic Quadrant for Data Integration Tools →

How ETL evolved

Businesses have been generating data since the age of the abacus, but modern analytics only became possible with the arrival of the digital computer and data storage.

A major step forward arrived in the 1970s, with a move to larger centralized databases. ETL was then introduced as a process for integrating and loading data for computation and analysis, eventually becoming the primary method to process data for data warehousing projects.

In the late 1980s, data warehouses and the move from transactional databases to relational databases that stored the information in relational data formats grew in popularity. Older transactional databases would store information transaction-by-transaction, with duplicate customer information stored with each transaction, so there was no easy way to access customer data in a unified way over time. With relational databases, analytics became the foundation of business intelligence (BI) and a significant tool in decision making.

Until the arrival of more sophisticated ETL software, early attempts were largely manual efforts by the IT team to extract data from various systems and connectors, transform the data into a common format, and then load it into interconnected tables. Still, the early ETL steps were worth the effort, as advanced algorithms, plus the rise of neural networks, produced ever-deeper opportunities for analytical insights.

The era of big data arrived in the 1990s as computing speeds and storage capacity continued to grow rapidly, with large volumes of data being pulled from new sources, such as social media and the Internet of Things (IoT). A limiting factor remained, with data often stored in on-premises data warehouses.

The next major step in both computing and ETL was cloud computing, which became popular in the late 1990s. Using data warehouses such as Amazon Web Services (AWS), Microsoft Azure and Snowflake, data can now be accessed from around the globe and quickly scaled to enable ETL solutions to deliver remarkable detailed insights and new-found competitive advantage.

The latest evolution is ETL solutions using streaming data to deliver up-to-the-second insights from huge amounts of data.

Let’s talk

ETL versus ELT

The most obvious difference between ETL and ELT—extract, load, transform—is the difference in order of operations. ELT copies or exports the data from the source locations, but instead of loading it to a staging area for transformation, it loads the raw data directly into the target data store to be transformed as needed.

While both processes leverage a variety of data repositories, such as databases, data warehouses, and data lakes, each process has its advantages and disadvantages. ELT is useful for ingesting high-volume, unstructured data sets as loading can occur directly from the source. ELT can be more ideal for big data management since it doesn't need much upfront planning for data extraction and storage.

The ETL process requires more definition at the onset. Specific data points need to be identified for extraction along with any potential "keys" to integrate across disparate source systems. The source of input data is often tracked by using metadata. Even after that work is completed, the business rules for data transformations need to be constructed. This work can usually have dependencies on the data requirements for a given type of data analysis, which will determine the level of summarization that the data needs to have.

While ELT pipelines have become increasingly popular with the adoption of cloud databases, ELT technology is still a developing process, meaning that best practices are still being established.

How ETL works

Let's talk

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 servers
- CRM and ERP systems
- JSON and XML
- Flat-file databases
- Email
- Web pages

Transform

In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase of the transformation process can include:

- 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.

Let’s talk

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 is often listed as a data integration method. In fact, it is most often used to create backups for 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.

Let’s talk

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.

Learn more about data integration →

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.

In addition, many ETL tools have evolved to include ELT capability and to support integration of real-time and streaming data for artificial intelligence (AI) applications.

Let’s talk

The future of integration—APIs using EAI

Application programming interfaces (APIs) using Enterprise Application Integration (EAI) can be used in place of ETL for a more flexible, scalable solution that includes workflow integration. While ETL is still the primary data integration resource, EAI is increasingly used with APIs in web-based settings.

Related solutions

Let’s talk

IBM DataStage

Supporting ETL and ELT patterns, IBM® DataStage® delivers flexible and near-real-time data integration both on premises and in the cloud.

Explore IBM DataStage →

IBM Data Replication

IBM Data Replication supports near real-time data synchronization across various sources and targets for improved business insight, continuous operations and the ability to react to changes in data.

Explore IBM Data Replication →

IBM Cloud Pak® for Data

IBM Cloud Pak for Data is an open, extensible data platform that provides a data fabric to make all data available for AI and analytics, on any cloud.

Explore IBM Cloud Pak for Data →

Data integration

Data integration enables you to transform structured and unstructured data, and deliver it to any system on a scalable big data platform.

Explore data integration →

Let’s talk

Resources

Tutorial

Hive as a tool for ETL or ELT

Learn how to extract, transform, and load OR extract, load, and then transform as you discover ways to process and analyze large datasets with ease using this tool.

Check out the tutorial →

Blog

ELT vs. ETL: What’s the Difference?

Learn the similarities and differences in the definitions, benefits and use cases of ELT and ETL.

Read the blog

Articles

Implementing ETL flows with Node-RED

Discover the power of ETL flows with Node-RED and learn how to streamline, implement and automate these critical processes and unlock the full potential of your data.

Learn more about ETL flows →

Take the next step

IBM DataStage is an industry-leading data integration tool that helps you design, develop and run jobs that move and transform data. At its core, DataStage supports extract, transform and load (ETL) and extract, load and transform (ELT) patterns.

Explore DataStage →

Try for free

Let’s talk

Let’s talk