ETL

Extract - Transform – Load

Extract, Transform, and Load (ETL) is the process of combining data from multiple sources into a large, central repository called a data warehouse. ETL uses a set of business rules to clean and organize raw data and prepare it for storage, data analytics, and machine learning (ML).

Organizations today have both structured and unstructured data from various sources including:

- Customer data from online payment and customer relationship management (CRM) systems
- Inventory and operations data from vendor systems
- Sensor data from Internet of Things (IoT) devices
- Marketing data from social media and customer feedback
- Employee data from internal human resources systems

By applying the process of extract, transform, and load (ETL), individual raw datasets can be prepared in a format and structure that is more consumable for analytics purposes, resulting in more meaningful insights. For example, online retailers can analyze data from points of sale to forecast demand and manage inventory. Marketing teams can integrate CRM data with customer feedback on social media to study consumer behavior.

ELT

Extract – Load – Transform

ETL stands for “Extract, Load, and Transform” and describes the set of data integration processes to extract data from one system, load it into a target repository, and then transform it for downstream uses such as business intelligence (BI) and big data analytics.

ELT and cloud-based data warehouses and data lakes are the modern alternative to the traditional ETL pipeline and on-premises hardware approach to data integration. ELT and cloud-based repositories are more scalable, more flexible, and allow you to move faster.

The ELT process is broken out as follows:

- **Extract** A data extraction tool pulls data from a source or sources such as SQL or NoSQL databases, cloud platforms or XML files. This extracted data is often stored temporarily in a staging area in a database to confirm data integrity and to apply any necessary business rules.
- **Load** The second step involves placing the data into the target system, typically a cloud data warehouse, where it is ready to be analyzed by BI tools or data analytics tools.
- **Transform** Data transformation refers to converting the structure or format of a data set to match that of the target system. Examples of transformations include data mapping, replacing codes with values and applying concatenations or calculations.

3 Tier Architecture in DE

Three-tier architecture is a well-established software application architecture that organizes applications into three logical and physical computing tiers: the presentation tier, or user interface; the application tier, where data is processed; and the data tier, where the data associated with the application is stored and managed.

The 3-tier of DE are:

- **Presentation Tier** The presentation tier is the user interface and communication layer of the application, where the end user interacts with the application. Its main purpose is to display information to and collect information from the user. This top-level tier can run on a web browser, as desktop application, or a graphical user interface (GUI), for example. Web presentation tiers are usually developed using HTML, CSS and JavaScript. Desktop applications can be written in a variety of languages depending on the platform.

- **Application Tier** The application tier, also known as the logic tier or middle tier, is the heart of the application. In this tier, information collected in the presentation tier is processed - sometimes against other information in the data tier - using business logic, a specific set of business rules. The application tier can also add, delete or modify data in the data tier.

- **Data Tier** The data tier, sometimes called database tier, data access tier or back-end, is where the information processed by the application is stored and managed.

ETL TOOLS

1. **Informatica PowerCenter** Informatica PowerCenter is one of the best ETL tools on the market. It has a wide range of connectors for cloud data warehouses and lakes, including AWS, Azure, Google Cloud, and SalesForce. Its low- and no-code tools are designed to save time and simplify workflows.

Informatica PowerCenter includes several services that allow users to design, deploy, and monitor data pipelines. For example, the Repository Manager helps with user management, the Designer allows users to specify the flow of data from source to target, and the Workflow Manager defines the sequence of tasks.

2. **Apache Airflow** Apache Airflow is an open-source platform to programmatically author, schedule, and monitor workflows. The platform features a web-based user interface and a command-line interface for managing and triggering workflows.

Workflows are defined using directed acyclic graphs (DAGs), which allow for clear visualization and management of tasks and dependencies. Airflow also integrates with other tools commonly used in data engineering and data science, such as Apache Spark and Pandas.

Companies using Airflow can benefit from its ability to scale and manage complex workflows, as well as its active open-source community and extensive documentation. You can learn about Airflow in the following DataCamp course.

3. **Oracle Data Integrator** Oracle Data Integrator is an ETL tool that helps users build, deploy, and manage complex data warehouses. It comes with out-of-the-box connectors for many databases, including Hadoop, EREPs, CRMs, XML, JSON, LDAP, JDBC, and ODBC.

ODI includes Data Integrator Studio, which provides business users and developers with access to multiple artifacts through a graphical user interface. These artifacts offer all the elements of data integration, from data movement to synchronization, quality, and management.