Percipient
intelligent data engineering

ScalETL for Maximum Connectivity & Scalability

Overview

ETL software automates the flow of data from an operational source to a target storage system. In order for data to be stored in an accessible manner for later use, large volumes of data across many source systems needs to undergo a series of processes. Depending on the tool, this can include extraction, re-arrangement, conversion, cleaning, consolidation, integration, transportation and loading to a target system, typically a database.

Notably, a number of ETL functions are sometimes offered within data warehousing environments, but there are also many independent tools offering native functionality for ETL processes. In both cases, data unification and scalability are primary concerns, as there are usually limits to their unification capabilities and processing power.

Staging Areas for Traditional ETL Jobs

While ETL tools are able to connect to different data sources, this is executed by connecting to each data source individually. The data is then pulled into a staging area, that is, a temporary location where the data is copied before being loaded into the central warehouse. There are several reasons why a staging area is traditionally required.

Source systems are usually only available for extraction during fixed times in the day because the time taken and process needed for extraction can disrupt their normal functioning. These times can differ by system. Under such circumstances, a staging area is used to temporarily hold extracted data until it is ready to be loaded.

Additionally, data is often extracted from several physically distinct databases in order to join common information stored in these databases. However, a different language is often needed to query the information contained in different databases. The staging area enables analysts to use the same language to query multiple databases.

The other problem faced by ETL platforms is the unevenness of data flows within an organisation. This could be because a network or software has crashed, or data from a source is too big, in an unacceptable format, or flowing too quickly. Alternatively, the systems within the organisation may have gradually become mis-aligned such that the existing ETL tool can no longer cope.