Limitations of Extract, Transform, Load (ETL)

Extract, Transform, Load (ETL) is a process to move data from its original source to another target destination. The target destination is often a database or a data warehouse, but it could be a simple flat file. The commonly referred to acronym, ETL, consists of three distinct stages:

Extract

Read data from a source (e.g. database) and extract the desired subset of data. The purpose of this step is to retrieve all the required data from the source system with minimum resources. This step needs to be designed in a way that it does not affect the source system negatively in terms of performance or response time. Often this means that the regularly scheduled pull is performed at night, when the system is not under load.

Transform

The transform stage cleanses and prepares the extracted data using lookup tables or rules. Data from heterogeneous sources can be combined at this stage. The transform step also includes validation of records, rejection of data (if they are not acceptable). The commonly used processes for transformation are conversion, sorting, filtering, clearing the duplicates, standardizing, translating and looking up or verifying the consistency of data sources.

Load

Loading is the last stage of an ETL process. The load function writes the extracted and transformed data (all of the subset or just the changes) to a target data location. Often the target data are inserted as a record in a target database using SQL insert statement.

ETL has been around for a long time and it often implies the use of SQL databases. When ETL is described we often come up short covering all of common tasks associated with this part of the AI workflow. Some examples of technologies that have forced the industry to re-think the boundaries of ETL are:

- source and target locations are not always SQL databases
- maintaining quality data can become more expensive than just storing everything
- streaming data as an alternative
- NoSQL technologies and enterprise solutions IBM’s enterprise data warehousing solution are common alternatives.

Another important limitation to the traditional ETL philosophy is that up front in the process decisions have to be made about which data are going to be important. Sometimes the specific data or form that the data must take is not evident. Sometimes, the best data for a model can change if a different model is selected. Flexibility in the early stages of the AI workflow is critical to avoid complications or errors later on.