Data ingestion and automation

Data engineers exist in many organizations to ease the burden of the data ingestion process. If the target data source is a database, then there are some useful tools and procedures under the umbrella term database testing. Data warehouse automation is the general term used to improve the overall process of data ingestion. Testing is an essential piece of data warehouse automation, because the quality of downstream models are tied to the quality of the available data.

IMPORTANT: The testing process is data-centric and it helps validate that data has been transformed and loaded into the target destination as expected. It is a critical part of data ingestion automation.

Testing can involve comparing large volumes of data which may contain millions of records. The size of the data can pose challenges, but in some cases a more significant challenge can be heterogeneous nature of data. You may find that you are using data of various types and sources: flat files, relational databases, open API feeds like twitter to XML web services and many others. Connecting all these heterogeneous sources in a standardized way can be a non-trivial task. With more sources of data comes an increased need for testing.

In reality, any form of data movement from source to target can be considered as data ingestion. In large enterprises like hospitals it is not uncommon to have dozens of independent systems saving data—oftentimes in a redundant way. A common database as a target is next to impossible due to logistical and privacy concerns, but a well-constructed gateway in the form of an Application Programming Interface (API) and API keys could be a viable solution towards automation.

Outside of more comprehensive solutions automation can be achieved with scripting. If the data ingestion code exists as a script (e.g. Bash or Python), then cron jobs are an incredibly powerful way to automate the process. The testing process can be automated with cron as well.