Why we need a data ingestion process

Cleaning, parsing, assembling and gut-checking data is among the most time-consuming tasks that a data scientist has to perform. In fact, the problem is not new as statisticians have been dealing with the same dilemma for many decades. The time spent on data cleaning can start at 60% and increase depending on data quality and the project requirements. One could debate the proportion and surely it depends on the team, the data and a number of other factors, but one statement that is difficult to argue against is

Very significant portions of time are often devoted to data ingestion pipelines.

For many enterprises data is the most important asset and when this is true maintaining the quality of those data is paramount. Poor data quality can result in project delays, budget projection shortfalls, or other avoidable challenges. The quality of data refers to both the observations themselves and the maturity of the data itself. Companies may consider improving their data ingestion infrastructure and methods for the benefits it could return.