Interactive and Iterative Automated Data Wrangling in Big Data Applications

Abstract
Identifying data, extracting data and data cleansing is time consuming and may be technically challenging. Data engineers may nearly spend 80% of their time on this process rather than analyzing or using the data. Therefore user centered and more cost-effective workflows are needed. These workflows should allow automatic data profiling, extraction and cleansing processes which take user preferences into account and also give the user the freedom to explore data in an iterative manner. Also, the processes should improve in each iteration by using the feedback provided by the user.

In each iteration, users have the chance to review the processed data. For example, data can be visualized in data analysis tools (e.g. Tableau or Qlik). Additionally, an interactive interface (Wrangler) can be used to visualize distributions and inconsistencies and provide recommendations to improve the processes. The user then has the possibility to provide feedback, data context, user context and to make interactive modifications to the processes. The resulting processes can be exported into big data applications.