prostep IVIP

Management Summary
"Data Preparation for Data Analytics"

prostep ivip

Management Summary

"Data Preparation for Data Analytics"

Data preparation for industrial value creation

Table of Contents

Abstract

Whether they are multinational corporations or hidden champions, companies in the developing and manufacturing industry are facing enormous challenges due to advancing digitalization. Growing demands for flexibility, shorter development and product life cycles, and new business models are permanently changing the way data is handled as a factor in production. However, in order to use data to generate value across the product life cycles, suitable prerequisites in the form of processes, methods and technology need to be put in place to make the transformation of raw data into usable information manageable while at the same time incorporating domain know-how.

The DPDA (data preparation for data analytics) project group unites industry users, system vendors and the research community under a common vision of developing a universal, standardized and adaptable tool for process-driven data preparation. In joint workshops, participants discuss practical use cases and best practices that demonstrate, among other things, that the systematization of data preparation and anchoring it as an integral part of product development and production have the potential to facilitate control and optimization of complex processes and products. With the development of role and procedure models, the project group is making an initial contribution to putting the existing wealth of experience to real use in industry.

1 Terms of reference and objectives

1.1 Challenges

It is not just since the advent of Industry 4.0 that the issue of data analytics has become an important element in building new solutions and business models in development and manufacturing companies. However, the Internet of Things (IoT), the ubiquity of sensors, high-performance in-memory and non-relational NoSQL databases and innovative visualization tools have all made scalable and cost-effective building blocks available to establish and operate data analytics as an integral part of product development and production. Another driving force on the path to the data and information economy is machine learning. Its ability to reveal correlations in complex data sets and the possibility of designing automated responses based on them promise further gains in efficiency and effectiveness that are conducive to securing competitiveness.

Process models such as the Cross-Industry Standard for Data Mining (CRISP-DM) and Knowledge Discovery in Databases (KDD) have become established to allow the raw material data to be refined into information relevant to value creation.

A closer look at these models reveals two aspects. On the one hand, there is a lack of explicit consideration of industrial needs such as the acquisition of existing data sources, IT systems, and the domain or subject expertise required for modeling and interpretation [1]. On the other hand, experience shows that 50-70 % of project outlay is accounted for by the "data preparation" phase (Abbildung 1). Efficient data preparation is thus both a pitfall and a lever for successful data analytics in the context of heterogeneous industrial data landscapes.