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

Management summary 1

1 Terms of reference and objectives 1
1.1 Challenges 1
1.2 Mission and vision 2
1.3 Motivations from the perspective of users and system vendors 3

2 Process models for data preparation 3
2.1 Definition and classification 3
2.2 Reference model of the DPDA project group 4

3 The role model of the DPDA project group 6

4 Outlook 7

5 Sources 8

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.

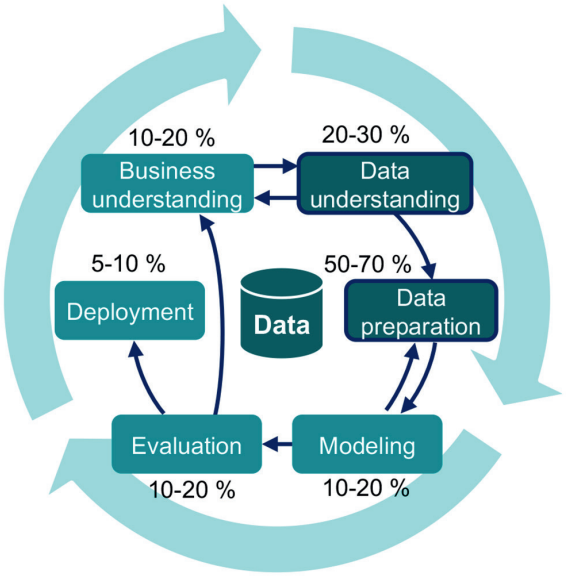

Figure 1: The Cross-Industry Standard Process for Data Mining (CRISP-DM) including the outlay for each phase of the process [2]

Following [3], the process of data preparation is defined as follows.

Definition of data preparation:
Data preparation is the process of transforming raw data into datasets that act as input to analysis functions or even machine learning algorithms in order to make predictions or gain insights. Data preparation includes the following activities:

- Obtaining and selecting data
- Cleansing data of errors and inconsistencies
- Integrating data from different sources
- Extracting features (feature engineering)
- Formating data for creating models

Furthermore, the associated activities such as collecting, organizing and cleansing datasets are described by data scientists as by far the most unpleasant part of their job [4]. Data preparation is therefore not only time-consuming and hence expensive, it is also not enjoyable given the time and expertise required to interpret it.

1.2 Mission and vision

The "Data Preparation for Data Analytics" (DPDA) project group was initiated by prostep ivip e.V. in 2018 with the remit of focusing on the low-outlay integration and preparation of data for modern analytics procedures and methods within an industrial context. Under the auspices of the Institute of Production Systems (IPS) of the TU Dortmund University and the Fraunhofer IPK, DPDA brings together stakeholders from business (IT vendors and industrial users) and the research community. The objective of DPDA is to develop concrete assistance for companies in respect of selecting and designing processes, methods, tools and information standards to allow them to master future data analytics projects and overcome the challenges associated with them. The focus is on two main requirements:

1. Assistive mechanisms for implementing data analytics projects are to be developed together with the companies involved. In particular, the roles to be filled, their functions and the procedure are to be specified and described.

2. It is necessary to ensure that data preparation is not done purely for its own sake. To achieve this, the phases of data preparation should be more closely aligned with those of data analysis.

All those involved share a common vision of developing a consistent, standardized and adaptable tool for process-driven

data preparation. To this end, the DPDA project group provides a platform where problems of data preparation that are relevant to industry can be collected, where approaches to solutions can be discussed and where collaborations across projects and across companies can be initiated. Consequently, strengthening collaboration between industry and the research community and initiating in-depth research projects represent another important component of the project group's mission.

1.3 Motivations from the perspective of users and system vendors

Users face the challenge that data preparation currently involves a lot of manual work, coupled with the fact that data scientists are also in short supply. There is considerable interest in an improved, more transparent process that requires less effort. Experiences with agile approaches in data preparation are also of interest to the DPDA group. Best practices and design options for procedures, analysis processes and architectures are to be presented and discussed between users in the project group. This accelerates the process of companies exchanging insights into the challenges they themselves face and obtaining a second opinion. In addition, the group investigates the potential for automation to simplify the exploratory phase of data preparation on the one hand and, on the other, to verify the health of data pipelines in the deployment phase. User companies are the primary target group of the DPDA initiative.

But system providers also participate in the DPDA group. On the one hand, the motivation is to derive clear requirements from real-life practice. This concerns not only problems faced by industry, but also interfaces and standards on which solutions can be built. For software providers in particular, the development of standard frameworks for data analytics and preparation as well as the required architectures in industrial practice are important areas of activity. In this way, the findings of the DPDA group can help to avoid custom solutions for a few customers and to increase the functional scope and quality of software solutions. Consulting and implementation services as well as matching customer requirements to the solution portfolios of system vendors may also be areas of interest.

2 Process models for data preparation

This section discusses the topic of data preparation and is intended to promote a common understanding.

2.1 Definition and classification

In earlier work of the project group, workshops and the findings of the latest research were used to bring together the most common industrial processes for data preparation. The methods analyzed comprise the Knowledge Discovery in Databases (KDD) process [5] and the Knowledge Discovery in Industrial Databases (KDID) process based on it [6], the Sample Explore Modify Model Assess (SEMMA) process [7], and the Cross-Industry Standard Process for Data Mining (CRISP-DM) [2]. The various processes sometimes divide steps into higher-level phases differently, although the contents usually show a strong correlation. All the processes described are iterative and are therefore not to be considered as a rigid sequence of steps. In many models, it is possible to go back one or more steps at any time in the event of problems or for validation purposes.

Before data preparation commences, an understanding of the business considerations must be built and data analytics objectives established. Problems associated with the company and production are collected and initial solutions are identified. In this phase, a use case is selected, taking into account the available data and the given data quality, applicable methods and the goals that have been set.

After the use case has been selected, the required data must be collected, viewed and checked. The data is collected from various sources and collated for common use. An initial exploratory assessment of the data takes place to determine its quality and identify issues. An initial data set is created from this data, and this forms the basis for initial exploratory analyses. It is often expedient to initially use a representative partial data set (subset), as large data sets quickly push the analysis environments to their limits and smaller datasets also reduce the effort involved in modeling. The dataset is also described using various statistical techniques in order to provide an overview.

Once the data has been selected, it must be pre-processed. Common challenges in industry include missing values and outliers, missing or discrepant timestamps, redundancies and inconsistencies, and discrepant IDs [8]. Furthermore, dimensions can be changed or reduced at this stage, e.g., using principal component analysis. For instance, colors can be translated to a different color space or the dimensions (number of variables) can be restricted to reduce complexity. The objective is to produce a dataset that can be used in the subsequent phase for training and evaluating machine learning models. It should be noted that the phases outlined above are usually not strictly sequential. Rather, the results of modeling can often lead to iterations in preliminary data preparation.

2.2 Reference model of the DPDA project group

This section presents the reference model for the data preparation process developed by the DPDA project group. The reference model outlined in Abbildung 2 is broken down into a sequential process that can be stepped through iteratively.

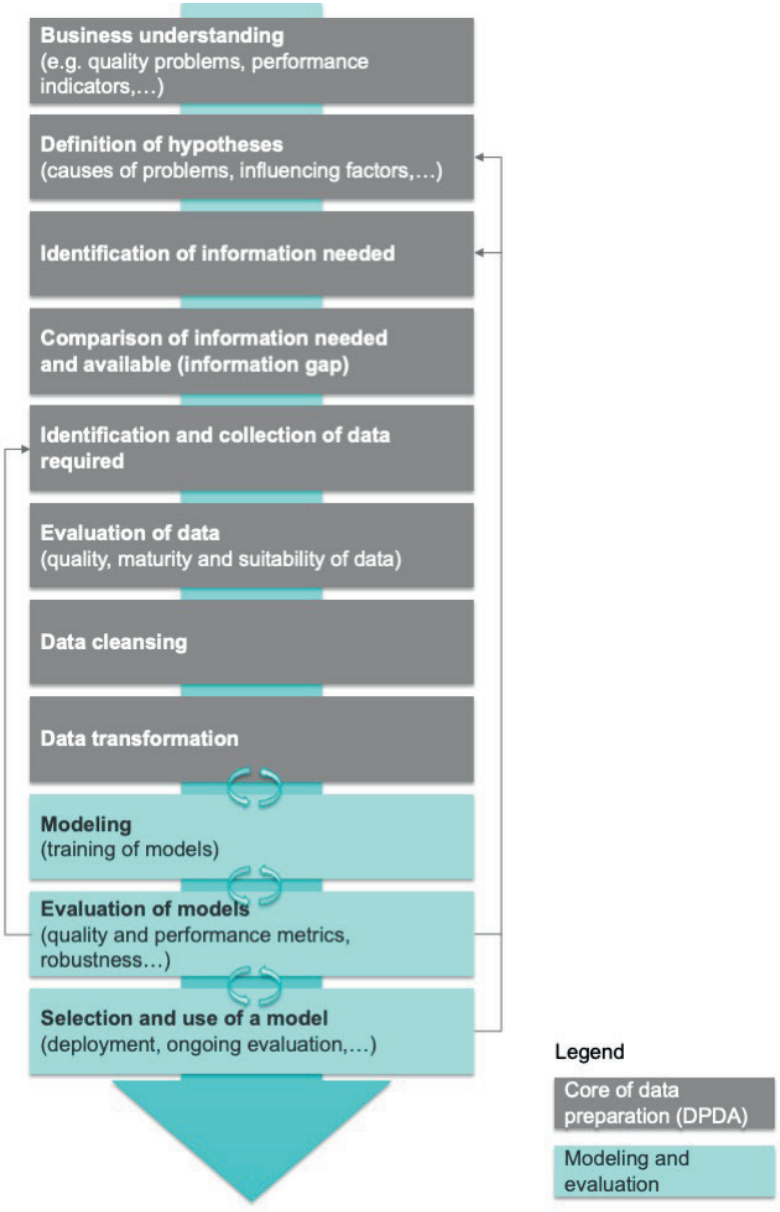

Figure 2 Reference model for data preparation

First, in the business understanding phase, a basic understanding of the business considerations must be established along the lines of CRISP-DM and goals must be derived. This includes a precise analysis of pain points and problems as well as opportunities that are to be addressed in an analytics project. Fields of application from industry can include such things as quality analysis and forecasting in order to reduce wastage or time expended, analysis and forecasting of plant malfunctions, but also higher-level aspects such as buffer and bottleneck predictions or the analysis of customer and supplier data. It is important to have measurable goals in order to assess the success of the project, but also to be able to derive target values to allow the training of models.

In the second phase, hypotheses are defined as to which influencing factors are important and how the problem defined initially can be solved. From these, information that is required to formulate a learning task is derived. Greater focus can be achieved using such things as sensitivity analyses or on the basis of conventional methods of structured problem solving such as Ishikawa diagrams in the context of lean manufacturing or Six Sigma initiatives. This forms the basis for the next step, in which the needs for and availability of information are compared in order to identify possible gaps and select potential data sources in a targeted manner.

The required data is identified and collected on the basis of the information needs. In the context of the defined influencing and target variables and the hypotheses, it is necessary to assess what data is available to describe the problem and possible causes. As a rule, this involves various data sources that must be merged in order to describe an issue. There are challenges associated with the use of different IT systems and access to heterogeneous underlying data sources. Moreover, errors may have been recorded or malfunctions documented manually in some cases, with the result that information is not necessarily available in machine-readable form. If a gap is still identified in the previous step, it may be necessary to initiate data collection again, to reject initial hypotheses, or to adjust the objective to reflect the actual situation in respect of the data. In a first step, it may be helpful to extract limited excerpts from larger data sources in order to be able to evaluate the basic data structures as well as to conduct initial exploratory analyses.

Data quality and availability are analyzed during subsequent assessment. Possible criteria include completeness of feature coverage, sample sizes, methods of data storage, data formats and structures, measurement scales and aggregation levels of feature instances, missing values, statistical distributions, or traceability through uniform IDs or time stamps [8].

The next step is to cleanse the data that has been selected and deemed suitable. For example, unnecessary time periods, duplicates in data sets and redundant or uninformative attributes can be removed. The removal of (obvious) outliers and measurement errors is also part of this phase. In addition, pre-processing steps for handling missing values must be included.

And closely related to this is the subsequent transformation of the data. The objective is to convert the data formats, which have already been cleansed but are still heterogeneous, into a form that can be processed by machine learning algorithms. Essentially, this step involves merging the various data sources based on uniform product IDs or time stamps. An important decision has to be made regarding how to handle different measurement scales or measurement intervals, e.g. by aggregation or discretization of nominal values. Feature engineering plays a crucial role in this context. This involves extracting highly informative properties with distinguishing power (known as “features”) from datasets and selecting them for the training of models. In this context, it should also be noted that many algorithms are based on statistical offset and position metrics or weighting factors, and that normalization of features may be expedient if different units and value levels are placed in relation to each other. For example, in the real world, it makes no difference whether a frequency is expressed as 1000 Hz or 1 kHz. But the difference is significant for algorithms that only calculate using the value 1000 or 1 and place this in relation to other features.

After this, the models are trained in the modeling phase. This can be done on a limited number of datasets with the objective of exploration, or on the largest possible set of data resources with the objective of optimizing quality and

performance metrics. The training of models is closely and iteratively connected with the transformation of the data, especially the feature engineering aspect, but also with the evaluation of the models. Different quality and performance metrics can be applied depending on the process. In addition to algorithm-related criteria, it is also necessary to check what inferences can be drawn in respect of the real-life business processes by checking hypotheses, but also by collecting new data or adjusting the objectives. Finally, one or more models must be selected for deployment and put into use. It should be noted that the same data pipeline, complete with the resulting features, that was used to train a model must also be set up during deployment, put into operation, and monitored and maintained on an ongoing basis. This means that the job of the DPDA is not finished after one-off data preparation prior to modeling. Instead, it is a constant task in the company and forms the basis for the reliable application of models in real-life practice.

3 The role model of the DPDA project group

Because it was a common element of industrial use cases for data preparation, the challenge of assembling appropriate teams was prioritized by the domain experts. In particular, discussion revolved around the mixture of domain expertise and IT competencies as well as objective-driven coordination. This allowed a role model to be developed (see Figure 3). This section describes in detail the individual roles and the associated tasks in the data preparation process.

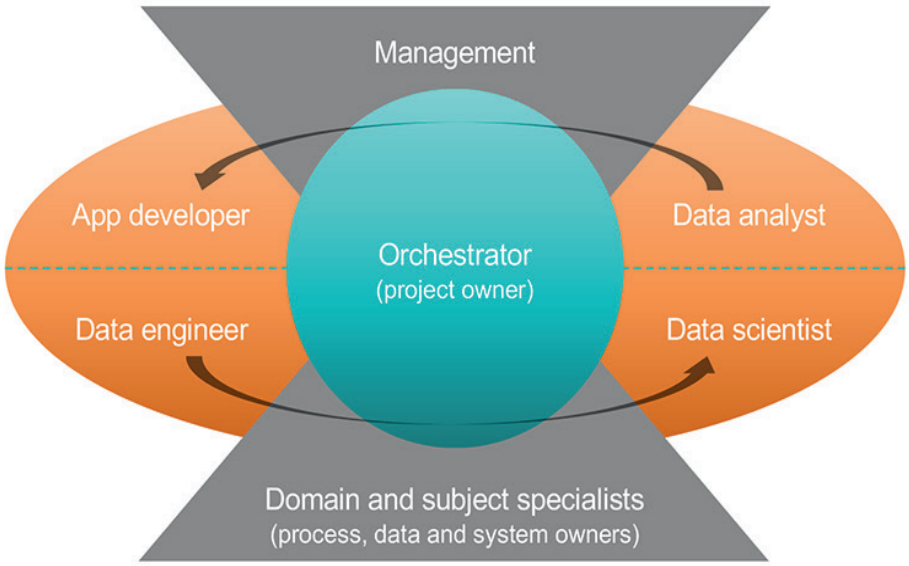

Figure 3: Role model for data preparation

Experts in relevant domains and subjects play a leading role in the process, as they often drive the initiation of the process, for instance by raising problems that cannot be solved by conventional improvement approaches. They are responsible for identifying problems, for instance regarding the quality of the product, and derive the required hypotheses from this. Finally, domain experts are responsible for identifying information and data sources that lead to the acceptance or rejection of the hypothesis.

In particular in the initiation or business understanding phase, domain experts are supported by management, who assist them in defining objectives and prioritizing possible application scenarios. Likewise, management can initiate projects out of strategic considerations and assemble suitable teams.

The role of orchestrator lies at the heart of the model. This role is responsible for project implementation, coordination and networking of other stakeholders, and communication. Ideally, this role has cross-discipline knowledge from the relevant domain and information technology as well as a basic understanding of data science. In this context, the term "citizen data scientist" has become established. This is a person who often comes from the domain and brings practical skills from the areas of IT and data science. The orchestrating role acts as a contractor for both strategic, management-initiated projects and problem-related projects initiated from an operational context by domain or subject matter experts. From the perspective of organizational structure, the role can be understood as a central staff position, for example.
The data or business analyst first compares the existing information with the information requirements. Since the information required and information available do not necessarily match, this results in an information gap. This information gap must be closed by a data engineer collecting additional relevant data. Once the information gap has been closed, the data is evaluated, and then cleansed and transformed. These steps can additionally be carried out in close coordination with data governance and monitored by a data steward insofar as these roles exist within the company. Data access rights must also be observed and clarified. In the data cleansing and transformation phases, the data scientists coordinate closely with the domain experts to decide how to deal with any problems identified in the event of poor data quality.

The subsequent modeling step forms an interface. If this involves simple modeling, it is performed by a data analyst. However, if the modeling is complex and exceeds the competency of the data analyst, it is performed by the data scientist. The solution strategies used depend heavily on the problem and the competence of the users and data scientists.

The data scientist, in conjunction with the data analyst and the domain expert, is also responsible for evaluating the model and applying it to the initial hypothesis. Furthermore, the IT departments must also be heavily involved in the case of real-life implementations. The final results may give rise to the need to implement automation mechanisms in the form of software. The software developers (app developers) responsible for this are advised and supported by the other specialist team members. Furthermore, deployment-oriented architectures and data pipelines must be established in the company and the results from the use of the model must be integrated into the business processes.

4 Outlook

In the next phase of the project, the use cases will be specified in close coordination with the project partners. These use cases describe example problem scenarios that companies face when implementing DPDA. On the basis of the use cases that are developed, the role and process models will also be further refined and validated. In particular, greater emphasis will be placed on the user perspective and the focus will be broadened to include the initiation and evaluation phases (pre-project and post-project phases).

Users who are involved in carrying out analyses often lack in-depth knowledge of IT or databases. In many cases, the individual departments want to carry out the analyses independently without having to call on the resources of the IT department. However, because data preparation requires specialized IT and database expertise, efforts are underway to provide business departments with tools for self-service data preparation. These tools enable users themselves to carry out the cleansing and preparation of data to a limited extent with the help of user-friendly interfaces. A further objective of the DPDA project group is thus to identify proposals for data standardization and further automation potential, to develop generalized process models and to offer users design support.

If you have your own use cases and would like to discuss them with the group's experts, please feel free to contact the coordinators.

5 Sources

[1] Lieber, Daniel; Erohin, Olga; Deuse, Jochen (2013): Wissensentdeckung im industriellen Kontext. In: ZWF 108 (6), pp 388–393. DOI: 10.3139/104.110948.

[2] Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger (2000): CRISP-DM 1.0. Step-by-step data mining guide: CRISP-DM Consortium.

[3] Dresner Advisory Services (publisher) (2019): Wisdom of Crowds® Data Preparation market study. Available online at http://dresneradvisory.com/products/2016-end-user-data-preparation-market-study.

[4] CrowdFlower (publisher) (2016): Data Science Report. Available online at https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.

[5] Fayyad, Usama M.; Piatetsky-Shapio, Gregory; Smyth, Padhraic (1996): From Data Mining to Knowledge Discovery. An Overview. In: Usama M. Fayyad (publisher): Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press.

[6] Deuse, Jochen; Erohin, Olga; Lieber, Daniel (2014): Wissensentdeckung in vernetzten, industriellen Datenbeständen. In: Hermann Lödding (publisher): Industrie 4.0. Wie intelligente Vernetzung und kognitive Systeme unsere Arbeit verändern. 27. HAB-Forschungsseminar. Hamburg, 12.-13.09.2014. Berlin: GITO (Schriftenreihe der Hochschulgruppe für Arbeits- und Betriebsorganisation e.V. (HAB)), pp 373–395.

[7] Matignon, Randall (2007): Data mining using SAS Enterprise miner. Hoboken, NJ: Wiley-Interscience.

[8] Eickelmann, Michel; Wiegand, Mario; Deuse, Jochen; Bernerstätter, Robert (2019): Bewertungsmodell zur Analyse der Datenreife. In: ZWF 114 (1-2), pp 29–33. DOI: 10.3139/104.112037.

prostep IVIP

prostep ivip association

Dolivostraße 11
64293 Darmstadt
Germany

Phone +49-6151-9287336
Fax +49-6151-9287326

psev@prostep.com
www.prostep.org

ISBN 978-3-948988-16-6
2021-12