Toward Standardization and Automation of Data Science Projects: MLOps and Cloud Computing as Facilitators

Christian Haertel[a], Christian Daase[b], Daniel Staegemann[c], Abdulrahman Nahhas[d], Matthias Pohl[e] and Klaus Turowski[f]
Magdeburg Research and Competence Cluster VLBA, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Keywords: Data Science, Project Management, Cloud Computing, MLOps, Automation.

Abstract: The significant increase in the amount of generated data provides potential for organizations to improve performance. Accordingly, Data Science (DS), which encompasses the methods to extract knowledge from data, has increased in popularity. Nevertheless, enterprises often fail to reap the benefits from data as they suffer from high failure rates in the conducted DS projects. Literature suggests that the main reason for the lack of success is shortcomings in the current pool of DS project management methodologies. Hence, new procedures for DS are required. Consequently, in this paper, the outline for a model for DS project standardization and automation is discussed. Following a summary of DS project challenges and success factors, the concept, which will incorporate MLOps and cloud technologies, and its individual components to address these issues are described on a high level. Therefore, the foundation for further research endeavors in this area is presented.

1 INTRODUCTION

The amount of data generated has risen significantly and is even predicted to massively increase in the future (Yin and Kaynak, 2015). Accordingly, due to the potentials for organizations to improve performance (Chen et al., 2012; Müller et al., 2018; Wamba et al., 2017), the importance of Data Science (DS), which is characterized as the application of (semi-)automated methods to extract knowledge from complex data (Schulz et al., 2020), has grown as well. Because of the limited applicability of project management approaches from traditional IT undertakings to DS (Das et al., 2015), various dedicated process models and methodologies (e.g., CRISP-DM) (Martinez et al., 2021b) emerged over the past decades to support the execution of DS projects. Nevertheless, the majority of these projects fail to achieve their goals and/or are cancelled (VentureBeat, 2019) which constitutes a significant issue (e.g., financial loss). In this regard, literature suggests several shortcomings in the current pool of DS methodologies (Martinez et al., 2021b). One of them is a lack of guidance for the execution of project management-related tasks that consider the specific characteristics of DS (Haertel et al., 2022c). Hence, new standardized approaches for DS project management are required (Saltz and Krasteva, 2022).

During project execution, a project team encounters challenges that can hinder the success of the undertaking. In DS projects, common issues are, amongst others, poor team coordination and collaboration, lack of analytics skills, a low level of process maturity for DS, setting adequate expectations due to uncertain business objectives, and the complexity of the IT infrastructure (Martinez et al., 2021b). Accordingly, an effective project management methodology for DS, which takes these problems into account, would be highly beneficial for project success (Martinez et al., 2021b; Saltz and

[a] https://orcid.org/0009-0001-4904-5643
[b] https://orcid.org/0000-0003-4662-7055
[c] https://orcid.org/0000-0001-9957-1003
[d] https://orcid.org/0000-0002-1019-3569
[e] https://orcid.org/0000-0002-6241-7675
[f] https://orcid.org/0000-0002-4388-8914

Haertel, C., Daase, C., Staegemann, D., Nahhas, A., Pohl, M. and Turowski, K.
Toward Standardization and Automation of Data Science Projects: MLOps and Cloud Computing as Facilitators.
DOI: 10.5220/0012235100003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 3: KMIS*, pages 294-302
ISBN: 978-989-758-671-2; ISSN: 2184-3228

Shamshurin, 2016). As a matter of fact, the incorporation of two emerging paradigms appear promising. First, Cloud Computing (CC) allows, for instance, the processing of Big Data at high speed (Cuquet et al., 2017). Moreover, the concept of MLOps, which constitutes a specialization of DevOps, can be utilized for the automation of analytical model development, deployment, and monitoring in DS projects (Kreuzberger et al., 2023; Makinen et al., 2021). Therefore, the following research question (RQ) shall be discussed in this paper:

RQ: *How can common challenges of Data Science project management be addressed, using MLOps and Cloud Computing?*

Answering the RQ will provide more clarity on how practitioners can potentially mitigate or solve typical impactful issues in the execution of DS projects. Additionally, this paper points out paths and goals for future research that is geared toward the implementation of the proposed solution concept(s). For this purpose, this article is structured as follows. After this introduction, the relevant terminology for this work are outlined in the theoretical background section. In particular, we focus on DS, project management, and CC. The third section initially contains the discussion of the DS project management state-of-the-art as well as challenges and success factors. Afterwards, the high-level concept of a model for DS project automation and standardization, using MLOps and CC, is introduced to address these challenges. Consequently, limitations of this paper and respective next steps are outlined. The work is concluded with a summary and an outlook to the future research endeavors.

2 BACKGROUND

To enable a discussion for answering the RQ, it is initially necessary to develop an understanding of the topic at hand. Therefore, in the following, the relevant concepts for this work are briefly outlined.

2.1 Data Science

Due to the increased amount of generated data in today's society (Yin and Kaynak, 2015) and the consequent aspiration to use it in a way that is beneficial to the fulfilment of business objectives, several data-related disciplines have emerged. In this regard, DS, Data Mining (DM), Data Analytics (DA), Machine Learning (ML) and Big Data (BD) constitute common terms that are also linked, which is highlighted by approaches that find interdisciplinary application (e.g., process models, algorithms) across these domains.

DA is an originally statistics-dominated field with the aim to "definitively address a hypothesis" (Chang and Grady, 2019) by an experiment design that determined the exact required and sufficient input data. The DA lifecycle consists of the steps "data collection, preparation, analytics, visualization, and access" (Chang and Grady, 2019). There are also some conceptual similarities to ML, where computer systems are created with *"automatic statistical and data analytics methods"* (Haertel et al., 2022a) to perform autonomous decisions through the use of data (Helm et al., 2020) while having improvement potential through experience (learning) (Ho et al., 2007). Later, DM emerged as new analytics specialization that was able to widen the class of problems by taking into account domain knowledge next to math and statistics. In particular, DM denotes the use of the dedicated algorithms for "extracting patterns from data" (Chang and Grady, 2019).

As datasets grew to an increased extensiveness over the past decade(s), new scalable architectures for their efficient storage, manipulation, and analysis were required (Chang and Grady, 2019). This observation is embraced by the BD concept, typically characterized as *"large datasets that primarily exhibit the characteristics of volume, velocity, variety, and/or variability"* (Chang and Grady, 2019) and that have a significant impact on the design of the needed IT infrastructure. Being tightly coupled to BD, the National Institute of Standards and Technology (NIST) defines DS as the *„methodology for the synthesis of useful knowledge directly from data through a process of discovery or of hypothesis formulation and hypothesis testing"* (Chang and Grady, 2019). DS can be understood as the set of activities in an analytics pipeline to gain insights from data. Accordingly, this includes the use of ML techniques for this goal. Hence, approaches like MLOps, which is characterized as *"a paradigm, including aspects like best practices, sets of concepts, as well as a development culture when it comes to the end-to-end conceptualization, implementation, monitoring, deployment, and scalability of machine learning products"* (Kreuzberger et al., 2023), find application in DS (Chowdary et al., 2022; Testi et al., 2022). In general, DS is typically considered as a super-set of the previously mentioned concepts, while specifically including the analysis of BD. Because of the latter, the term was further developed into *"Big Data Science"* (Chang and Grady, 2019).

2.2 Project Management

According to the Project Management Institute (PMI), project management encompasses *"the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements"* (PMI, 2017). Consequently, adequate project management shall enable the project team to deliver the undertaking successfully (Pilorget and Schell, 2018). Therefore, generally speaking, project management needs to consider the three cornerstones scope, time, and budget.

Project management constitutes a widely researched domain that spawned several standards and methodologies such as the Project Management Body of Knowledge (PMBOK) (PMI, 2017). The PMBOK describes the ten knowledge areas of project management, including *Project Scope Management, Project Time Management, Project Cost Management*, and *Project Resource Management*. The latter is defined as the identification, acquisition, and management of the resources that are required for the "successful completion of the project" (PMI 2017). As a result, the management of the allocated resources (e.g., personnel, in-kind assets) is vital (Pilorget and Schell, 2018). Accordingly, the processes of Project Resource Management should help to ensure the availability of the relevant resources to the project manager and team "at the right time and place" (PMI, 2017). The resource management is closely connected with the project scheduling as it describes when the agreed deliverables from the project scope will be fulfilled. Using other words, the project schedule defines start and end dates for activities to determine when the respective resources are required (PMI, 2017). Thus, clearly established timelines for project milestones can be considered a success factor, especially for DS (Martinez et al., 2021a). Moreover, resource management is further impacted by the project requirements which determine scope and deliverables of the project.

2.3 Cloud Computing

The term "Cloud Computing" has gained increased interest in the past years (Hasimi and Penzel, 2023). It is characterized as *"a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., network, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"* (Mell and Grance, 2011). CC is composed of five main characteristics, namely *"on-demand self-service"*, *"broad network access"*, *"resource pooling"*, *"rapid elasticity"*, and *"measured service"* (Mell and Grance, 2011). CC offers various benefits for organizations. For instance, CC can reduce expenses associated with materials such as costs for technical experts, physical space, and security (Alalawi and Al-Omary, 2020). Additionally, especially small companies can profit from the rapid provisioning of required computing equipment via easy interfaces to achieve the intended results (reduced time-to-market) (Alalawi and Al-Omary, 2020). This also offers the possibility to facilitate the execution of DS projects. Instead of traditionally conducting the time-intensive setup of the necessary infrastructure of the undertaking, a system can be quickly set up (Yan, 2017). Furthermore, the rapid elasticity characteristic allows the release of unused resources that were relevant for previous compute-intensive DS project tasks (e.g., model training). Accordingly, resource utilization is optimized in contrast to static scaling. Finally, the benefits of cloud for DS exceed infrastructure-related aspects. For instance, fully-managed AutoML tools like Google Cloud's Vertex AI offer support for different activities in a DS project (Google Cloud, 2023). Moreover, dedicated ML platforms (e.g., Kubeflow, Airflow) allow orchestrating entire MLOps workflows on cloud infrastructure (George and Saha, 2022). Therefore, numerous activities in a DS project can be supported, standardized, and automated.

3 DATA SCIENCE PROJECT MANAGEMENT: OVERVIEW, CHALLENGES, AND SOLUTIONS

This paper aims to propose how certain project management related challenges of DS projects could be addressed in the future. Therefore, it is necessary to initially discuss the current state-of-the-art of DS project management. Afterwards, we specifically outline the literature regarding identified problems, proposed solutions, and general success factors of DS projects.

3.1 Data Science Project Management

In the broadest sense, DS projects constitute IT undertakings as well. However, due to the data focus and the resulting more explorative nature (Das et al.,

2015), DS has unique characteristics in comparison to "traditional" IT projects. Consequently, the particularities of DS should be considered in the applied project management methodology. As a matter of fact, multiple DS process models emerged in the past (e.g., CRISP-DM) to support DS project execution. Nevertheless, a high failure rate of DS undertakings (VentureBeat, 2019) suggests methodological weaknesses in these process models. In the past, various studies already examined the established DS methodologies. In a previous literature analysis, existing DS process models and reviews of them were comprehensively investigated (Haertel et al., 2022b). Some of the involved articles and the key findings are outlined in the following. For example, Schulz et al. (2020) evaluated KDD, SEMMA, CRISP-DM, and the TDSP across certain requirements. Using the discovered weaknesses, the process model DASC-PM was introduced. Similarly, strengths, limitations, and weaknesses of eight different DS process models were reviewed in the publication of Oliveira and Brito (2022). A comparison of KDD, SEMMA, and CRISP-DM regarding their activities was conducted by Azevedo and Santos (2008). Kutzias et al. (2021) reviewed seven established DS methodologies to derive important features for future continuous DS process models. In an expansive study by Martinez et al. (2021b), a total of 19 DS process models were evaluated regarding their consideration of challenges in team, project, and data and information management. Finally, Haertel et al. (2022b) examined 28 DS methodologies, focussing on activities, team roles, and deliverables. As a consequence, a general Data Science lifecycle (DSLC) was derived. In general, the mentioned studies arrive at a comparable conclusion. Currently, DS appears to lack "integral methodologies" (Martinez et al., 2021b) which implies the need for new or revised approaches. This corresponds to the findings of similar research in the field (e.g., Saltz and Krasteva (2022)). Additionally, the newly generated process models from the above-mentioned articles were not yet used and evaluated in an organizational context.

The lacking success of DS projects is largely attributed to the process aspect instead of the technical side (Saltz and Krasteva, 2022), which coincides with the existing analyses of DS process models. The absence of "established and mature methodologies" (Saltz and Krasteva, 2022) for DS is also reflected by the discovery that only a minority of DS projects follow a dedicated methodology (Martinez et al., 2021a; Saltz et al., 2018). This phenomenon might be a result of missing guidelines on how to conduct specific project management-related activities with respect to DS in the process models, including, amongst others, requirements engineering, project scheduling, resource management, or technology selection (Haertel et al., 2023). Because of the specific DS characteristics, the portability of traditional approaches from software engineering is limited. For example, the relevant resources differ, and the explorative nature aggravates the time planning of activities (Saltz, 2015). Accordingly, further research is required in this area to bridge this gap, since the effectiveness of project management activities such as "planning, budgeting, solving conflicts, and controlling requirements" (Gökay et al., 2023) is influential on project success (Iriarte and Bayona, 2020).

3.2 Challenges and Success Factors of Data Science Projects

The low success rate of DS projects suggests the existence of several methodological issues that occur in the execution of such undertakings. Martinez et al. (2021b) reviewed the literature and identified a total of 21 "main challenges" in DS projects. A relevant subset for this work, including ways to address them, will be discussed in the following.

The above-outlined lack of standardized approaches for DS project management constitutes an indicator for the *low level of process maturity* for DS (Bhardwaj et al., 2014; Martinez et al., 2021b). Accordingly, it is recommended to employ a *Data Science lifecycle* that features the common high-level tasks, specific guidelines and a scheme for project management (Martinez et al., 2021b). This coincides with the findings of Saltz and Shamshurin (2016), who see a *well-defined organizational structure* and *project management process* as success factors.

The explorative nature exacerbates *setting adequate expectations* for DS projects (Das et al., 2015; Martinez et al., 2021b). This increases the emphasis that should be put on the business understanding phase (Martinez et al., 2021b) to achieve *clarity of project deliverables* and *explore and communicate difficulties* of the project (Saltz and Shamshurin, 2016). Accordingly, conducting a *feasibility study* is important in DS (Saltz and Shamshurin, 2016).

Another consequence of the more explorative character of DS projects is the *difficulty to establish realistic project timelines* (Martinez et al., 2021b; Saltz et al., 2017). Therefore, *processes to control the duration of specific steps* are required (Martinez et al., 2021b) to enable increased concreteness in planning

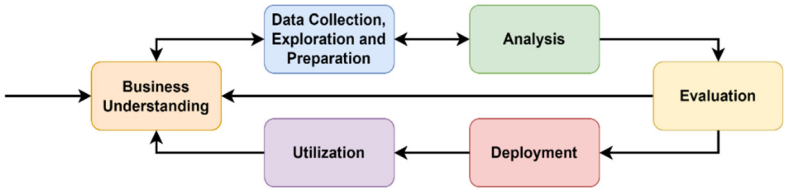

Figure 1: DSLC, adapted from Haertel et al. (2022b).

future undertakings. Nevertheless, a DS project (schedule) should still feature *flexibility and agility* to give freedom for experimentation (Saltz and Shamshurin, 2016).

A challenge directly related to the low success rate of DS undertakings is that the *project results* are *not used* by the business (Martinez et al., 2021b). Thus, *methods for evaluation of the results* and the *integration* of these *in the client environment* are required and should be accompanied by the *appropriate training* of the end users (Martinez et al., 2021b). In this regard, the application of MLOps procedures presents an interesting concept (Kreuzberger et al., 2023). Additionally, *clear strategic objectives* and *clear business requirements* (Gökay et al., 2023) are needed to avoid this pitfall. To further promote the benefit(s) of the project, methods like the value net could be considered to identify the business value of a DS application for the enterprise (Pohl et al., 2023).

Literature suggests that organizations executing DS projects suffer from a *lack of people with analytics skills* (Martinez et al., 2021b; Saltz et al., 2017). This observation complies with the widely spread shortage of skilled workers (Duell and Vettier, 2020). Consequently, DS projects often require time-consuming training to development the necessary (analytical) capabilities (Medeiros et al., 2020; Saltz and Shamshurin, 2016). As a result, technology providers have increased their offerings toward no-code and low-code solutions (Hintsch et al., 2021) that can support in DS projects (e.g., VertexAI AutoML) to also cater to non-technical users.

On the side of DS team management, *collaboration issues* are outlined by multiple publications (Martinez et al., 2021b; Saltz et al., 2017). Accordingly, a *clear role definition* would "aid coordination between team members and across stakeholders" (Martinez et al., 2021b). Other measures to mitigate this challenge are *continuous documentation*, employment of a *git workflow and coding agreements* (Martinez et al., 2021b) as well as an established *communication about the data and initiatives* (Saltz and Shamshurin, 2016).

Similarly, *inefficient governance models* for DS are highlighted (Becker, 2017; Martinez et al., 2021b). An integral DS methodology should describe *how to coordinate with the IT* department and feature *approaches for the deployment* stage (Martinez et al., 2021b). Gökay et al. (2023) propose the implementation of *IT Governance Frameworks* (e.g., ITIL), while Medeiros et al. (2020) argue for *mechanisms and data governance policies*. Other related recommended principles refer to *clear role definition* and *versioning code, data, and models* (Martinez et al., 2021b).

Furthermore, a *lack of reproducibility* and *knowledge retention and accumulation* are identified as challenges (Martinez et al., 2021b; Saltz et al., 2018). To tackle these obstacles, organizations should implement a *setup for assuring reproducibility and traceability* by ensuring *continuous documentation* and *creating a knowledge repository* (Martinez et al., 2021b). This includes findings about *data, models, experiments, project insights, best practices, and pitfalls* (Martinez et al., 2021b) that could be of use for future endeavors, too.

Due to the significance of the data for DS project success, the perceived *lack of quality assurance checks* from the literature can be detrimental for the outcome (Domino Data Lab, 2017; Martinez et al., 2021b). Therefore, DS approaches should prescribe *tests to check data limitations in quality and potential use* (Martinez et al., 2021b). A general *data governance* policy regarding *standardization* and documentation of *data preparation steps* (Gökay et al., 2023; Medeiros et al., 2020) as well as *data (quality) management and ownership* should be in place (Saltz and Shamshurin, 2016).

Seeing the increased application of DS in the BD context, the necessary *investment in the IT infrastructure* poses a considerable factor (Martinez et al., 2021b; Saltz et al., 2017), which might require *preallocation of* (monetary) *resources* for these expenses. As a matter of fact, *sufficient investment and management of the IT infrastructure* for the relevant technology and tools is commonly described as a success factor (Gökay et al., 2023; Medeiros et al., 2020; Saltz and Shamshurin, 2016). However, there is a lack of holistic technology selection approaches for DS (Haertel et al., 2023), as most methods focus on specific project steps or technology

types (e.g., BD reference architectures like in Volk et al. (2019)). Nevertheless, the use of CC technologies can allow for reducing expenses associated with management of infrastructure (Alalawi and Al-Omary, 2020) and increase organizational agility (rapid (de-)provisioning of computing resources) (Yan, 2017), which benefits DS project execution.

3.3 Toward a Model for Data Science Project Standardization and Automation

The previous subsection outlined that various challenges can hamper successfully carrying out a DS undertaking. Regardless, the literature offers multiple abstract proposals and success factors on how to address the process-related issues in DS. In the following, we will refer to these aspects and argue how these obstacles can be overcome through application of a model for a standardized and (partly) automated DS workflow, leveraging both, MLOps and cloud tools. Addressing these problems will contribute to achieving better DS project success rates, which would support DS practitioners significantly.

In general, the outlined challenges in context of DS can be mainly categorized as project management and especially resource-related. Additionally, these are amplified through today's widely present personnel shortage of qualified workers, especially for IT personnel and data scientists (iMove, 2022). Accordingly, a new model for DS project planning and execution should take these factors into account and alleviate the burden that is created by these obstacles (Martinez et al., 2021b). Therefore, we propose the generation of standardized templates for the DS project activities, which would simplify DS project planning, resource management, and in turn, execution. These should include guidelines, logic and parameters relevant for these tasks across the DSLC (e.g., infrastructure management, analytical model development). Having a standardized procedure allows the automation of certain DS project tasks and therefore can increase process maturity and address issues in team coordination. In this regard, MLOps as specialization of DevOps constitutes a possibility for DS (Sweenor et al., 2020), since a significant portion of such projects is concerned with the application of analytical models. Here, MLOps aims to automate and monitor *"at all steps of ML system development and deployment, including integration, testing, […] and infrastructure management"* (Makinen et al., 2021). The combination with cloud tools seems especially promising, as time savings, agility, and bridging the analytic skill gap can be facilitated (Alalawi and Al-Omary, 2020; Hentschel and Leyh, 2018). In the following, we will further detail this concept and outline the necessary components of this artifact from a research standpoint.

Self-evidently, standardized approaches for project planning and resource management can only be constructed when a standard DS workflow with the commonly relevant activities and milestones is used as a foundation. In a recent work, Haertel et al. (2022b) derived a general DSLC (see Figure 1), which encompasses the six broad stages 1) *Business Understanding*, 2) *Data Collection*, *Exploration and Preparation*, 3) *Analysis*, 4) *Evaluation*, 5) *Deployment*, and 6) *Utilization*. In a supplementary BPMN process map, the respective detailed sub-activities of each phase, the functional roles, and relevant documentation deliverables are stated (Haertel et al., 2022b). The DSLC will be utilized as the basis for the model. As a next step, the commonly involved resources within the DSLC need to be identified for enabling adequate project resource management. Since team management constitutes an integral part of resource management, the enlisted team role profiles, namely *Business/Domain Users, Project Manager, Data Scientists, Data Engineers*, and *IT Infrastructure Team* (Haertel et al., 2022b), need to be defined in further detail.

An issue in the prospect of generating standardized project templates is that, naturally, not all DS projects are identical or similar. These undertakings have different characteristics, which impact the encountered challenges in the project (Saltz et al., 2017). In turn, this should also influence the project management approach, including planning and resources. Thus, a categorization of DS projects will be helpful to describe the undertaking and draw inferences for project templates. In particular, such model can facilitate an improved understanding of the project life cycle, timelines, and obstacles (Saltz et al., 2017). The rough project requirements serve as an important input, because they directly determine the scope. Moreover, other factors play a role. In this respect, a categorization approach of Saltz et al. (2017) raises only two major dimensions for DS classification: the *level of discovery* (explorative character of the project) and the *infrastructure* (level of computing requirements). Other criteria such as team aspects, type of analytics goal (Nalchigar and Yu, 2018), or the respective domain are not considered. From the categorization criteria, Saltz et al. (2017) identify four types of DS projects, namely *"smaller data"*, *"well-defined"*, *"exploratory"*, and *"hard to justify"*. Additionally, the derived concrete implications for DS project management are limited.

Consequently, we see the necessity to develop a revised DS categorization model that allows drawing inferences for the project management practices within the DSLC, depending on the respective category. Therefore, we will extensively analyze the literature regarding DS project case studies and other potentially existing artifacts for categorization.

Upon completion of this essential building block, a DS project categorization model should enable the creation of standardized templates for the identified DS project types (Saltz et al., 2017), which are designed to support dealing with project management-related challenges such as establishing timelines and deliverables (Martinez et al., 2021b). Furthermore, this allows recommending the respectively suitable technical setups for the DS project (Volk et al., 2022), using cloud technologies. With this knowledge, tools and techniques of MLOps can be used for automation of infrastructure management as well as model development, and deployment (Makinen et al., 2021). To the best of our knowledge, there is no such model for project templates for automation in DS that is based on project categories derived from the respective objectives and requirements. Designing such an approach will have multiple advantages for the management and execution of DS projects. Currently, the Business Understanding stage of the DSLC prescribes a so-called "Situation assessment" (Haertel et al., 2022b) that is, amongst others, concerned with identifying and setting up the relevant technological infrastructure and performing the necessary training for the team members. These time-intensive activities, especially the setup of architectures for the projects, can be simplified and automated with the application of cloud technologies and orchestration tools (Sousa et al., 2015) that offer increased scalability and agility (Hasimi and Penzel, 2023). On a similar note, the use of such standardized DS project templates in conjunction with CC and automation through MLOps (Chowdary et al., 2022) also offers the potential to bridge the gap caused by shortage of skilled (analytics) personnel to some extent (Hasimi and Penzel, 2023; Olsowski et al., 2022). In turn, active human involvement can largely be focused on the most important DS activities and project management challenges that require critical thinking.

3.4 Limitations and next Steps

This paper presents an overview of current challenges in DS project management and a concept for a model, involving MLOps and CC, to address them. However, further detailing of this research goal is required in future studies. From a methodological perspective, the Design Science Research (DSR) methodology will be leveraged for the construction of this artifact (Hevner et al., 2004). As the problem relevance constitutes an important guideline in DSR, this article puts strong emphasis on outlining the currently faced issues in DS and potential benefits when they are addressed with a dedicated approach. Accordingly, a comprehensive and structured review of the literature on DS project problems, challenges, and success factors would be beneficial to adequately highlight the research gap and motivate the need for revised DS methodologies. Additionally, since DSR prescribes the rigorous demonstration of the efficacy of an artifact (Hevner et al., 2004), these derived issues and best practices from the scientific body of knowledge can be utilized as parameters for showcasing how the envisioned model is superior to state-of-the-art approaches.

Furthermore, it has to be noted that cloud technologies and MLOps tools already find use in DS projects to automate certain activities (Sweenor et al., 2020). However, based on the definition of Kreuzberger et al. (2023), MLOps seems to predominantly focus on the engineering perspective (i.e., development and deployment of ML models) instead of employing a holistic view on the project that is aligned with the business goals (i.e., extract insights from data for a specific objective) and requirements. Hence, our envisioned model of standardized templates for DS undertakings aims to bridge this gap between the technical and project management perspective. For further clarification, a structured literature review of the current application of MLOps in DS is necessary to underscore this matter. Moreover, this will provide the foundation for a conceptual representation of the intended artifact.

4 CONCLUSION

DS has become a widely utilized discipline to extract actionable insights from data. Therefore, for organizations, the successful completion of the corresponding projects is required to reap the benefits. However, there are multiple challenges associated with process and resource aspects that hinder DS undertakings. Hence, new methodologies are needed (Martinez et al., 2021b) and the publication at hand proposes the idea for a model for DS project standardization and automation, using MLOps and cloud technologies, to tackle the aforementioned challenges (e.g., lack of (analytics) personnel). The

relevant components of the model are outlined on a high-level and the objectives of subsequent research are introduced to further detail the envisioned artifact. Therefore, for example, an abstract and conceptual representation of this model needs to be constructed. On the basis of the DSLC of Haertel et al. (2022b), the concretely impacted DS tasks shall be highlighted. This will also include the automation of the documentation artifacts which constitute the outputs of the individual steps within the DSLC (Haertel et al., 2022b).

REFERENCES

Alalawi, A., and Al-Omary, A. (2020). "Cloud Computing Resources: Survey of Advantage, Disadvantages and Pricing," in *ICDABI 2020*.

Azevedo, A., and Santos, M. F. (2008). "KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW,"

Becker, D. K. (2017). "Predicting Outcomes for Big Data Projects: Big Data Project Dynamics (BDPD)," *2017 IEEE International Conference on Big Data*.

Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A. J., Madden, S., and Parameswaran, A. G. (2014). "DataHub: Collaborative Data Science & Dataset Version Management at Scale,"

Chang, W. L., and Grady, N. (2019). "NIST Big Data Interoperability Framework: Volume 1, Definitions,"

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.

Chowdary, M. N., Sankeerth, B., Chowdary, C. K., and Gupta, M. (2022). "Accelerating the Machine Learning Model Deployment using MLOps," *Journal of Physics: Conference Series* (2327:1), p. 12027.

Cuquet, M., Vega-Gorgojo, G., Lammerant, H., Finn, R., and Hassan, U. u. (2017). "Societal impacts of big data challenges and opportunities in Europe," *Societal impacts of big data: challenges and opportunities in Europe*.

Das, M., Cui, R., Campbell, D. R., Agrawal, G., and Ramnath, R. (2015). "Towards Methods for Systematic Research on Big Data," *2015 IEEE International Conference on Big Data*.

Domino Data Lab. (2017). "Managing Data Science Projects," available at https://www.dominodatalab.com/resources/field-guide/managing-data-science-projects/, accessed on Jul 20 2023.

Duell, N., and Vettier, T. (2020). "The employment and social situation in Germany," *Policy Department for Economic, Scientific and Quality of Life Policies*.

George, J., and Saha, A. (2022). "End-to-end Machine Learning using Kubeflow," *5th Joint International Conference on Data Science & Management of Data*.

Gökay, G. T., Nazlıel, K., Şener, U., Gökalp, E., Gökalp, M. O., Gençal, N., Dağdaş, G., and Eren, P. E. (2023). "What Drives Success in Data Science Projects: A Taxonomy of Antecedents," *Computational Intelligence, Data Analytics and Applications* (643), pp. 448-462.

Google Cloud. (2023). "Vertex AI," available at https://cloud.google.com/vertex-ai?hl=en, accessed on Jul 11 2023.

Haertel, C., Nahhas, A., Daase, C., Volk, M., and Turowski, K. (2022a). "A Holistic View of Adaptive Supply Chain in Retailing Industry," *AMCIS 2022 Proceedings* (7).

Haertel, C., Pohl, M., Nahhas, A., Staegemann, D., and Turowski, K. (2022b). "Toward A Lifecycle for Data Science: A Literature Review of Data Science Process Models," *PACIS 2022 Proceedings*.

Haertel, C., Pohl, M., Nahhas, A., Staegemann, D., and Turowski, K. (2023). "A Survey of Technology Selection Approaches for Data Science Projects," *AMCIS 2023 (In Press)*.

Haertel, C., Pohl, M., Staegemann, D., and Turowski, K. (2022c). "Project Artifacts for the Data Science Lifecycle: A Comprehensive Overview," in *2022 IEEE International Conference on Big Data*.

Hasimi, L., and Penzel, D. (2023). "A Case Study on Cloud Computing: Challenges, Opportunities, and Potentials," *Developments in Information and Knowledge Management Systems for Business Applications* (466), pp. 1-25.

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., and Ramkumar, P. N. (2020). "Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions," *Current Reviews in Musculoskeletal Medicine 13*, pp. 69-76.

Hentschel, R., and Leyh, C. (2018). "Cloud Computing: Status quo, aktuelle Entwicklungen und Herausforderungen," in *Cloud Computing*, pp. 3-20.

Hevner, A. R., March, S. T., and Park, J. (2004). "Design Science in Information Systems Research," *MIS Quarterly*.

Hintsch, J., Staegemann, D., Volk, M., and Turowski, K. (2021). "Low-code Development Platform Usage: Towards Bringing Citizen Development and Enterprise IT into Harmony," *ACIS 2021* (11).

Ho, T. B., Kawasaki, S., and Granat, J. (2007). "Knowledge Acquisition by Machine Learning and Data Mining," in *Creative Environments*: *Issues of Creativity Support for the Knowledge Civilization Age*, pp. 69-91.

iMove. (2022). "Shortage of skilled workers in STEM areas reaches new all-time high – acute shortage in eastern Germany," available at https://www.imove-germany.de/en/news/Shortage-of-skilled-workers-in-STEM-areas-reaches-new-all-time-high-acute-shortage-in-eastern-Germany.htm, accessed on Jul 18 2023.

Iriarte, C., and Bayona, S. (2020). "IT projects success factors: a literature review," *International Journal of Information Systems and Project Management* (8:2).

Kreuzberger, D., Kühl, N., and Hirschl, S. (2023). "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," *IEEE Access* (11).

Kutzias, D., Dukino, C., and Kett, H. (2021). “Towards a Continuous Process Model for Data Science Projects,” *Proceedings of the AHFE 2021 Virtual Conference on The Human Side of Service Engineering* (266), pp. 204-210.

Makinen, S., Skogstrom, H., Laaksonen, E., and Mikkonen, T. (2021). “Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help?” *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI*.

Martinez, I., Viles, E., and Olaizola, I. G. (2021a). “A survey study of success factors in data science projects,” in *2021 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 2313-2318.

Martinez, I., Viles, E., and Olaizola, I. G. (2021b). “Data Science Methodologies: Current Challenges and Future Approaches,” *Big Data Research 24*.

Medeiros, M. M. de, Hoppen, N., and Maçada, A. C. G. (2020). “Data science for business: benefits, challenges and opportunities,” *The Bottom Line*.

Mell, P. M., and Grance, T. (2011). “The NIST definition of cloud computing,” *NIST Special Publication 800-145*.

Müller, O., Fay, M., and vom Brocke, J. (2018). “The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics,” *Journal of Management Information Systems* (35:2), pp. 488-509.

Nalchigar, S., and Yu, E. (2018). “Business-driven data analytics: A conceptual modeling framework,” *Data & Knowledge Engineering* (117), pp. 359-372.

Oliveira, D. F., and Brito, M. A. (2022). “Development of Deep Learning Systems: A Data Science Project Approach,” in *Information Systems and Technologies*: *WorldCIST 2022, Volume 2*, pp. 325-332.

Olsowski, S., Schlögl, S., Richter, E., and Bernsteiner, R. (2022). “Investigating the Potential of AutoML as an Instrument for Fostering AI Adoption in SMEs,” in *Knowledge Management in Organisations*.

Pilorget, L., and Schell, T. (2018). *IT Management*: *The art of managing IT based on solid framework leveraging the company's political ecosystem*.

PMI. (2017). *A guide to the project management body of knowledge (PMBOK guide)*.

Pohl, M., Haertel, C., and Turowski, K. (2023). “Value Creation from Data Science Applications - A Literature Review,” *22nd International Conference on Perspectives in Business Informatics Research (in-press)*.

Saltz, J., Hotz, N., Wild, D., and Stirling, K. (2018). “Exploring Project Management Methodologies Used Within Data Science Teams,” *Americas Conference on Information Systems 2018: Digital Disruption*.

Saltz, J., Shamshurin, I., and Connors, C. (2017). “Predicting data science sociotechnical execution challenges by categorizing data science projects,” *Journal of the Association for Information Science and Technology* (68:12), pp. 2720-2728.

Saltz, J. S. (2015). “The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness,” *IEEE International Conference on Big Data 2015*.

Saltz, J. S., and Krasteva, I. (2022). “Current approaches for executing big data science projects - a systematic literature review,” *PeerJ Computer Science* (8:e862).

Saltz, J. S., and Shamshurin, I. (2016). “Big data team process methodologies: A literature review and the identification of key factors for a project's success,” in *2016 IEEE International Conference on Big Data.*

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kuehnel, S., Badewitz, W., Dann, D., Kloker, S., Alekozai, E. M., and Lanquillon, C. (2020). “Introducing DASC-PM: A Data Science Process Model,” *ACIS 2020*.

Sousa, T. B., Correia, F. F., and Ferreira, H. S. (2015). “Patterns for Software Orchestration on the Cloud,” *HILLSIDE Proc. of Conf. on Pattern Lang. of Prog.*

Sweenor, D., Hillion, S., Rope, D., Kannabiran, D., Hill, T., and O'Connell, M. (2020). *ML Ops: Operationalizing Data Science*, O'Reilly Media Inc.

Testi, M., Ballabio, M., Frontoni, E., Iannello, G., Moccia, S., Soda, P., and Vessio, G. (2022). “MLOps: A Taxonomy and a Methodology,” *IEEE Access* (10).

VentureBeat. (2019). “Why do 87% of data science projects never make it into production?” available at https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/, accessed on Jul 21 2023.

Volk, M., Bosse, S., Bischoff, D., and Turowski, K. (2019). “Decision-Support for Selecting Big Data Reference Architectures,” in *Business Information Systems*, pp. 3-17.

Volk, M., Staegemann, D., Saxena, A., Hintsch, J., Jamous, N., and Turowski, K. (2022). “Lowering Big Data Project Barriers: Identifying System Architecture Templates for Standard Use Cases in Big Data,” in *Proceedings of the 19th ICSBT*.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., and Childe, S. J. (2017). “Big data analytics and firm performance: Effects of dynamic capabilities,” *Journal of Business Research* (70).

Yan, G. (2017). “Application of Cloud Computing in Banking: Advantages and Challenges,” *ICPEL 2017*.

Yin, S., and Kaynak, O. (2015). “Big data for modern industry: challenges and trends [point of view],” *Proceedings of the IEEE* (103:2), pp. 143-146.