Fraunhofer ISST

Fraunhofer Institute for Software and Systems Engineering ISST

DATAOPS FOR DATA SHARING

ISST-REPORT
Challenges and Requirements for interorganizational Data Sharing

Part of the Project

IEDS

This report is part of the research from the "IEDS – Incentives and Economics of Data Sharing" project funded by the German Federal Ministry of Education and Research (BMBF)

Federal Ministry of Education and Research

ISST-REPORT

DATAOPS FOR DATA SHARING

Challenges and Requirements for interorganizational Data Sharing

The digital transformation has changed the role of data from a by-product of business processes to a strategic resource for new business potential and innovation. At the same time, the value creation and innovation process has become increasingly complex. Today, companies are less and less able to offer digital services and products on their own. Companies must respond to strategic and operational challenges that require high-quality data. Therefore, new structures are forming in the data economy in the form of data ecosystems and data spaces.

However, participation in data ecosystems and in the data economy itself requires strategic and efficient data management so that companies are able to meet the needs of data consumers in an efficient, timely, and high-quality manner. However, it shows that many companies are not able to provide and share the data they need. In addition, the implementation of new data environments leads to enormous effort and data professionals are often busy fixing data quality errors instead of building new data pipelines to participate in data sharing. One solution to establish effective and efficient data management could be DataOps. DataOps is a data management practice that uses agile and lean structures to automate and standardize the building of data pipelines while ensuring high data quality. DataOps practices enable organizations to leverage new data sources and efficiently deliver across enterprise boundaries

This report serves as an overview of how DataOps can be established in an organization to optimize strategic data management and thus be able to participate efficiently in the data economy. It shows how agile and lean structures can be used to build new data pipelines and increase data quality through strategic data management.

ISST – Series of Reports:

Within the Series of "ISST Reports", the Fraunhofer Institute for Software and Systems Engineering ISST, publishes its white paper. Thematically, it examines trends and technologies in computer sciences and takes up innovative subjects from some of the Institutes research projects. They provide insights into the current state of research concerning "Innovations From Data", Fraunhofer ISST's main research topic.

AUTHOR

Inan Gür M.Sc.

ISST-REPORT

ISSN 0943-1624

Published: December 21, 2021

EDITOR

Prof. Dr.-Ing. Boris Otto
Prof. Dr. Jakob Rehof

CONTACT

Fraunhofer Institute
for Software and Systems Engineering ISST
Emil-Figge-Straße 91
Germany - 44227 Dortmund
info@isst.fraunhofer.de
+49 231 97677-0

IMAGE SOURCE

Cover: ©bingfengwu_iStock-1156830749

Table of Content

1. DataOps for Data Sharing

The crucial role of data in our society and economy

In our current, ever-shifting era of technological advancements and innovations, not only the velocity of change is accelerating, so is the global amount of data that we generate, capture, analyse, consume (figure 1). Whereas in 2020 the global amount of data reached 64.2 zettabytes, it is estimated to surpass 180 zettabytes by the year 2025 (European Commission 2020). This evokes changes not only in society, as new technologies and data sources alter the way we communicate and consume, but also in the way economic organizations pursue their business and seek for survival and success in the on-going market. The breakdown of global public cloud workloads show how enterprise applications, analytics and internet of things increasingly dominate the global data flow, showing how companies constantly face new challenges in dealing with environmental, technological and economical changes. The flood of data enabled entire industries through innovations, new digital services and ways to manage, analyse, govern, handle and use high quality data.

The new industries and businesses are often times shaped by a systemic worldview, which emphasizes the interdependence and interconnectedness of each market participant. This development shifts the focus of companies on innovation from their own, internal facilities towards outside their boundaries, searching and enabling collaboration. The valuation of data as a strategic resource pushes companies to utilize and look for data sources outside their organizational boundaries. Many smart services or products rely on high quality data that cannot be generated by an organization alone. These new data-driven innovations are increasingly difficult to develop by a single organization and in traditional value chains. Instead, the increasingly interconnected world is leading to the combination, enrichment, and sharing of different data sources by different actors in cross-sectoral, socio-technical networks-so called data ecosystems (Gelhaar et al. 2021; Oliveira and Lóscio 2018). Data ecosystems consist of complex networks of organizations and individuals that share and use data as a primary resource. Such ecosystems provide an opportunity and basis for creating, managing, and sustaining data sharing initiatives (Oliveira and Lóscio 2018). The value of digitally transforming a society and industries is unmistakable. However, the value creation process must be well thought out and deliberately accomplished (Mielli and Bulanda 2019). Ecosystems for data, innovation and collaboration of various actors are becoming increasingly more relevant. Data ecosystems offer participants the opportunity to extend their access to data sources as data will be exchanged and monetarized within these ecosystems. The collaborative use of data, also called "data sharing", enables organizations to pursue their business in new ways of digital services and products, as data ecosystems can offer a complete coverage of a data value chain. According to the current state of the literature, there is not yet a uniform and differentiated definition of the term data exchange. However, the definition by Thuermer et al. (2019) describes the central aspects as follows:

> *"The sharing of data between entities, typically for a specific purpose. This can happen between companies, or departments within an organisation. The data owner or provider provides a data user with access to some of its data. If the data is personal data, then the sharing will be regulated by GDPR."*

The value of data for industry and society is also emphasized by the German government in its data strategy[1] published by the Federal Chancellery. It states that data form the basis of the digital society. With its strategy, the German government therefore aims to increase the innovative and responsible provision and use of data, particularly in Germany and Europe. One of the strategy's goals is to make German and European data ecosystems more attractive to more participants by expanding

1 https://www.bundesregierung.de/breg-en/service/information-material-issued-by-the-federal-government/data-strategy-of-the-federal-german-government-1950612

data infrastructures in an interoperable, energy- and resource-saving, and decentralized manner. To this end, the cross-industry GAIA-X project is to be advanced to create open and transparent data ecosystems by making data and services available, merging them and sharing them in a trustworthy manner (Bundeskanzleramt 2021).

The European Commission has also focused on data ecosystems and data sharing in its data strategy. According to the European Data Strategy[2] and the European Commission's Communication to the European Parliament, the Council and the European Committees, the EU can become a model for a society that can make better decisions thanks to data in the public sector and in the economy, as digital technology has transformed the economy and society of the European people. One of the goals of the European data strategy is therefore to create a single European data space in the sense of a single market for data. In this data ecosystem, global data should be securely and easily available. Specifically, the intention is to drive forward the creation of EU-wide interoperable data spaces that will remove the legal and technical barriers associated with data sharing. In these data spaces, European rules, in particular privacy and data protection and competition law, should be fully respected and the rules for access to and use of data should be fair, practical and clear (European Commission 2020a). The basis for the interoperable data spaces is to be the federated data infrastructure of GAIA-X and the standard for sovereign data exchange of the International Data Spaces Association.

Considering the internal and external data assets of an organization as a valuable resource can be the crucial measure to turn the fate of a company (Gür et al. 2021). In order to strive for long term sustainable success and competitive advantage, companies need to excel in the way they manage and utilize their data in order to participate in data ecosystems and data sharing. However, organizations see themselves facing a variety of digital hardships. These challenges occur not only on a technological level when dealing with legacy technology and a colossal software landscape, but also on an organizational level. Often times data is stuck in data silos within the organization or the data team is not capable of utilizing new data sources in a quick manner. Businesses today require data teams that are agile in order to respond to the data needs of stakeholders as quickly as possible. At the same time the data must be trustworthy, so decision-makers can utilize it worry-free.

2 https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_de

How fast is your data growing per day?

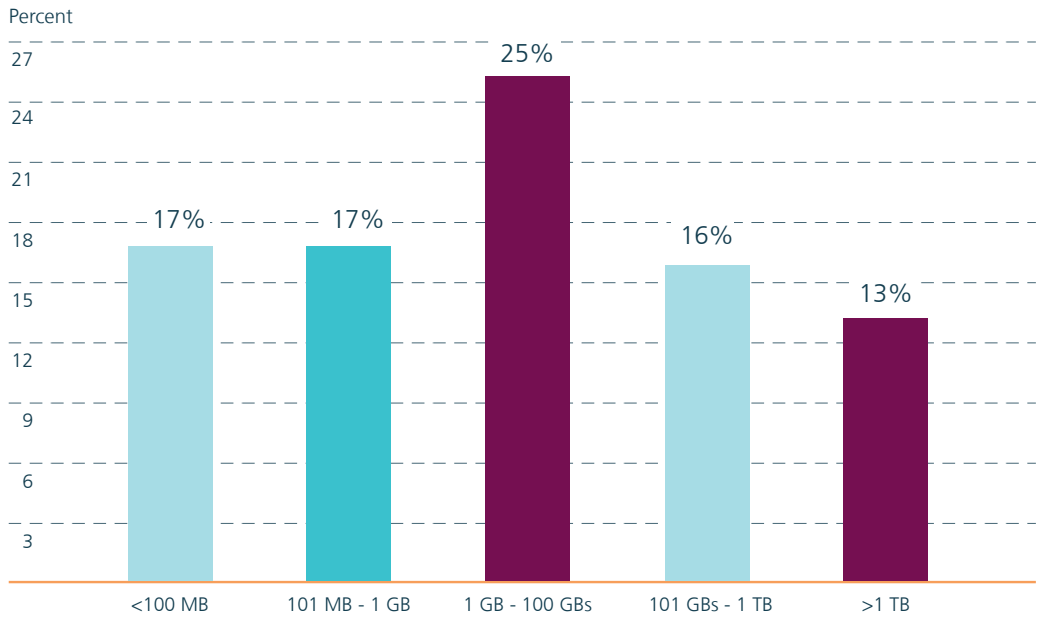

Figure 1: The growth of Data in organizations per day (Nexla Inc. 2018)

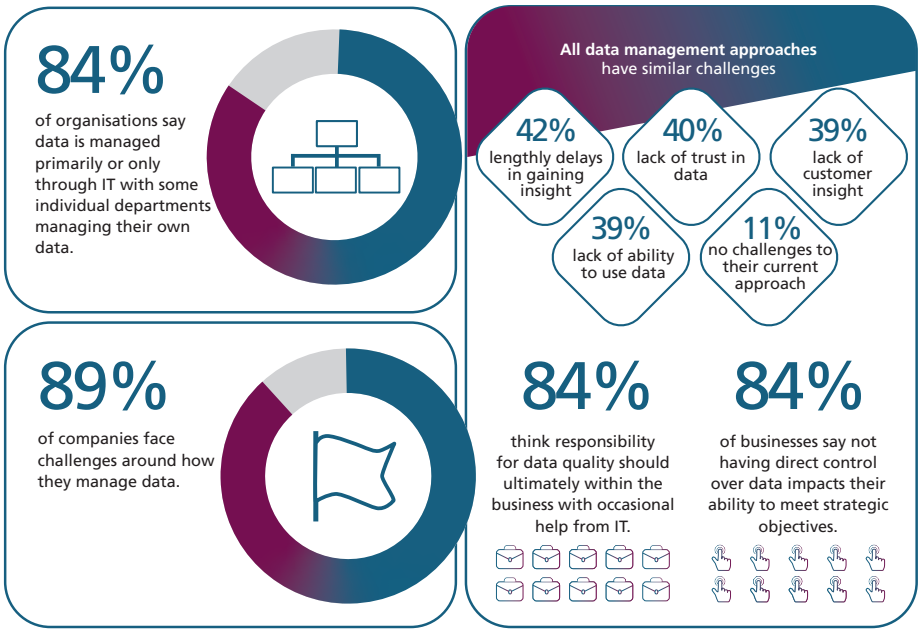

Figure 2: Challenges and Obstacles in Data Management (Experian ltd. 2019)

Studies show that organizations need to overcome a variety of barriers to maximize their strategic gain from data. According to a study conducted by Experian ltd. 89% of companies face challenges around how they manage data, due to lengthy delays in gaining insights, lack of trust in data or lack of ability to use data (see figure 2).

DataOps is a buzzword that has emerged in the recent past, but is now being used with increasing frequency. More and more companies are showing up that they have been able to realize and optimize their data analytics capabilities and data management through successful implementation of DataOps practices. The increasing use of this term is evident not only in

Figure 3: Global Google Trends results for DataOps in the last 5 years

the increasing amount of best practices being published, but also in an analysis of Google Trends over the past few years. Figure 3 shows the data on global DataOps Google search queries over the last 5 years. This shows that the interest in DataOps is continuously increasing.

At the same time, however, there is a lack of well-founded and scientific elaboration that underpins this concept and fills it with content. For example, a search using the Scopus database, which is one of the largest databases in the field of information systems research, yields only a small number of results. This underscores how great the interest in DataOps is for companies, but they do not have a solid base of knowledge about DataOps to fall back on. This white paper aims to counteract this development by providing a detailed analysis and summary of DataOps in science and business.Hence, an analysis and preparation of the body of knowledge regarding DataOps is necessary. For example, the study by the company (Seagate Technology LLC) shows that the status of DataOps in companies is still very weak, especially in Europe. More than half of the companies surveyed claim that DataOps is currently not being taken into account at all or is only being developed. Only five percent of the companies were able to report that DataOps capacities are fully implemented throughout the company (see figure 4).

DataOps status in the company

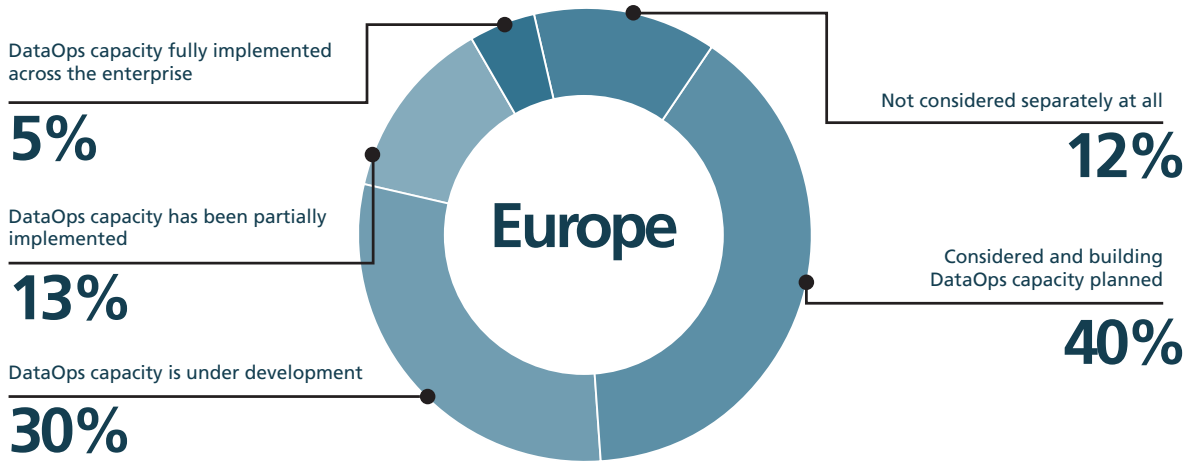

Figure 4: DataOps Status in organizations in europe (Seagate Technology LLC 2020)

This study reveals another interesting fact. The study compares the DataOps capacities of companies globally and compares them between the regions North America, China, Europe and Asia-Pacific (see figure 5). The global average among the surveyed companies that rate DataOps as extremely important is 23% and 42% that rate DataOps as very important. On average, only 35% of companies said they rated DataOps as relevant, relatively relevant or not very relevant. In Europe, however, the proportion of companies rating the relevance of DataOps in the lower spectrum is 42%. Only 15% of the companies surveyed in Europe said DataOps was extremely important. However, the study was conducted before the Corona pandemic began. Due to the need for digitization revealed by the pandemic, these numbers are expected to increase in all regions in the future. However, in Europe, DataOps is the least implemented area. In this regard, the study reveals that European companies responded rather conservatively in a number of areas. One potential explanation for this is that pan-European companies are less likely to share data across borders due to current GDPR rules and other regulations. According to the study, these rules, combined with legacy data management technologies, could be the barriers to full data use compared to other regions. Current rules require extra effort to leverage the value of available data, as GDPR regulations make data management more complicated and data protection easier to ensure in fixed data silos, inhibiting innovation (Seagate Technology LLC, 2020). How DataOps can counteract this will be demonstrated throughout this reports.

The importance of DataOps

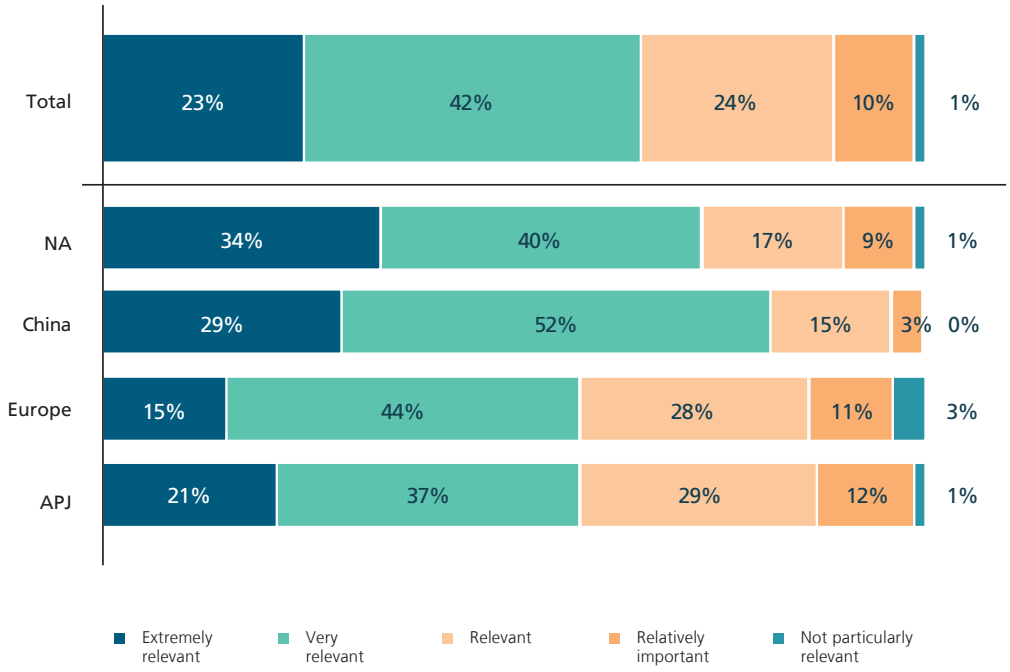

Figure 5: DataOps Status in organizations wolrdwide (Seagate Technology LLC 2020)

2. DataOps – agile and lean Data Management

We conducted a "Multi-vocal Literature Review" and "Structured Literature Review" to explain the concept of DataOps, give a thorough definition and elaborate core tenets. "In the following chapter, the analysis will be presented". Therefore, section 2.1 explains the method procedure, section 2.2 elaborates the definition of DataOps, section 2.3 gives an overview on DataOps frameworks and section 2.3 explains the core tenets of DataOps.

2.1. Market Analysis and Scientific Review

In the first chapter we elaborated the urgent necessity of a digital transformation of organizations to yield maximum value from their data and data management. Furthermore, we showed how efficient and automated data management still forms a significant challenge for organizations and that the rising trend provides a promising approach to overcome this challenge.

In this chapter, we want to provide a thorough definition of DataOps and elaborate the core tenets of this concept. Therefore, we conducted a "Multi-vocal Literature Review" (MVLR) for publications and insights of economic organizations and other market participants as well as a "Structured Literature Review" (SLR) to evaluate scientific publications of the Information Systems research community on DataOps (see figure 6). Our goal in this endeavour is to work towards a consensus in the field DataOps and agile Data Management, as first research results show how diverse the definitions and approaches of DataOps are. As it has been shown in the first chapter, the field of DataOps is a new and "young" research and operational field. Thus, in order to cover the entire field and yield sufficient resources, we opted for a market analysis with the MVLR and a scientific analysis with the SLR. In order to maintain rigor and depth in our analysis, we followed the procedure of (Vom Brocke et al. 2009; 2015; Webster and Watson 2002). As agile Data Management and DataOps can be mapped to the research field of Information Systems, we utilized the most common databases, namely Scopus, AISeL, ACM DL and IEEE Xplore, to reveal as many relevant publications as possible. Using the keyword "DataOps" our search yielded 32 publications, from which, after applying exclusion critera like "The study is not accessible on the web", "The study does not present any type of findings or discussion about DataOps" or "The study is not written in English", 12 relevant publications could be extracted. We conducted a forward and backward search on these 12 publications and by that extended the scientific literature basis to 15.

In order to conduct a market analysis, we searched for analysed economic whitepapers, reports and firm insights. We used the google search engine to find relevant publications from firms like IBM or McKinsey and the search yielded 13 relevant publications. We analysed these publications in terms of DataOps definition, DataOps tools, DataOps execution and DataOps frameworks. The analysis provided valuable insights, covering a variety of DataOps approaches, recommendations and indications on challenges and obstacles. In the following-section, we evaluate different definitions on DataOps give an overview on DataOps frameworks and present the core tenets of this concept..

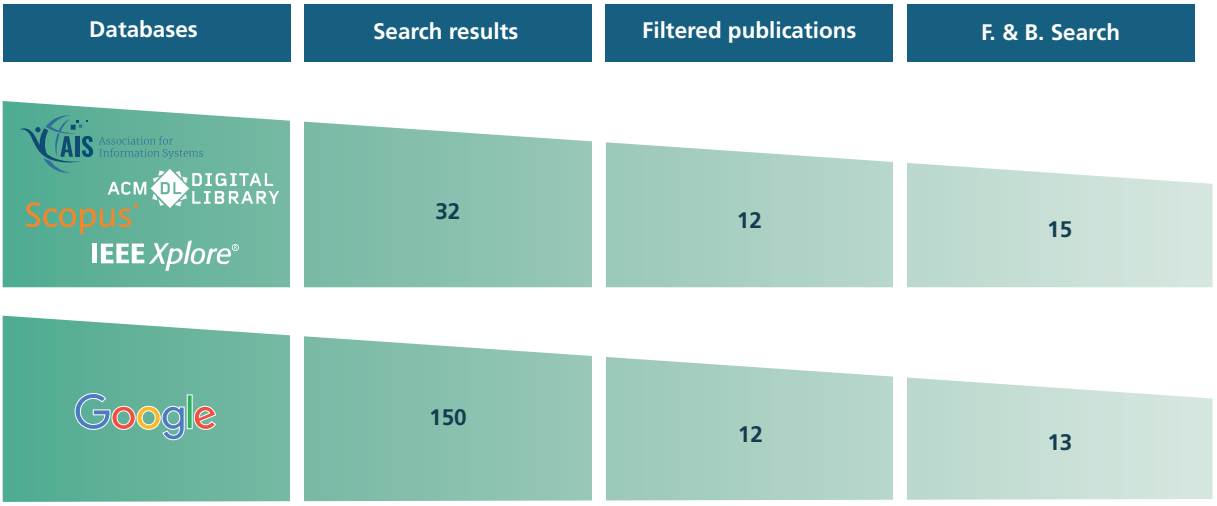

Figure 6: Scientific Analysis of DataOps in Research and Economy

2.2. Definition of DataOps

As shown in the previous section, DataOps is a new approach to efficient and effective data management and its attention is rising. In these circumstances, it is essential to have a consensus in the definition of the approach, in order to not only do research in the scientific field but also to build benchmarks throughout different organizations and enable collaborative innovations. Therefore, during our analysis we took a closer look on prevalent definitions in different companies and organizations as well as in scientific publications on DataOps.

The analysis revealed several definitions with some differences but also ubiquitous constructs and elements. In the following we want to point out the most commonly used and relevant definitions as well as emphasize core constructs of DataOps. Building on these definitions, the subsequent section presents frameworks and core tenets of DataOps and how they are grasped in this new approach.

We start with the most dominant and widely used definition for DataOps throughout economic publications given by Gartner.
In their online glossary[3] in the section for information technology, they describe DataOps.

A similar approach on DataOps is presented by IBM. In their whitepaper "Deliver business ready data fast with DataOps

> "DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization. The goal of DataOps is to deliver value faster by creating predictable delivery and change management of data, data models and related artifacts. DataOps uses technology to automate the design, deployment and management of data delivery with appropriate levels of governance, and it uses metadata to improve the usability and value of data in a dynamic environment."

3 **https://www.gartner.com/en/information-technology/glossary/dataops**

- An introduction to the IBM DataOps methodology and practice" (Madera and Aguilera 2020) they give a definition for DataOps that emphasizes the data governance aspect through the orchestration of people, process and technology to provide high-quality data to data consumers. By focussing on automation and collaboration across an organization, agility and speed in setting up new data initiatives is increased and inefficiencies in the data pipeline are decreased:

> "Data operations (DataOps) is the orchestration of people, process and technology to deliver trusted, high-quality data to data citizens fast. The practice is focused on enabling collaboration across an organization to drive agility, speed and new data initiatives at scale. Using the power of automation, DataOps is designed to solve challenges associated with inefficiencies in accessing, preparing, integrating and making data available."

Another often used definition for DataOps is given by 451 Research. 451 Research is an information technology industry analyst firm based in New York and formerly belonged to the 451 Group. The organization focusses on 9 research channels that align with prevailing issues in IT innovation. The definition of DataOps from Matt Aslett (2019) published by 451 research focusses on agile and automated data management provided by the alignment of people, process and technology.

> "DataOps is the alignment of people, process and technology to enable more agile and automated approaches to data management. It aims to provide easier access to data to meet the demands of various stakeholders who are part of the data supply chain (developers, data scientists, business analysts, DevOps professionals, etc.) in support of a broad range of use cases."

During our analysis we performed a structured literature review to include scientific publications in order to reveal results and conclusions on DataOps from different research streams. Our analysis yielded publications from different scientific conferences and journals within the research field of Information Systems.

According to Sahoo and Premchand (2019) article published in the International Journal of Applied Information Systems, DataOps focusses on automation and methods from agile software engineering:

> "DataOps is DevOps based and process-oriented methodology used by data and analytics teams to improve the quality of data analytics and reduce cycle time of the same, generally, to aid better business decision making, improving profits, furthering business and monetizing data available with an organization"

Following the colleagues of Capizzi et al. (2020) in their article published in the "Software Engineering Aspects of Continuous Development and New Paradigms of Software Production and Deployment" journal, DataOps can be defined as follows:

> "DataOps is a new approach that aims to improve quality and responsiveness of data analytics life-cycle. This approach is based on DevOps rules, in particular DataOps aims to bring DevOps benefits to data analytics, adopting Agile rules and Lean concepts"

There are several other definitions, which we have summarized in the following table. The table again reflects the high level of interest from both industry and academia. It also shows that while there is no single definition of DataOps that all publications refer to, there is consensus on core features of DataOps and its intent. DataOps is a methodology for aligning and orchestrating people, processes, and technologies to improve, accelerate, and improve the quality of the data integration process and the building of data pipelines. In doing so, the methodology draws on concepts from lean management, DevOps, quality management, and agile methodologies to standardize and automate data processes, as well as create collaboration and a culture of continuous improvement. DataOps thus represents a relevant part of an organization's data strategy that should lay the foundation for digital transformation and participation in data sharing. Through successfully implemented DataOps practices, organizations are able to meet the needs of data consumers efficiently and with high quality data. DataOps not only draws on technology, but provides a method to align, standardize, automate and structure the associated processes and stakeholders. In doing so, it accelerates the development and production process of data environments, new features and models, and increases the responsiveness of data teams. Considering data as a strategic resource, high-quality data delivery should be a standard and top priority for data-driven organizations and those that want to become one. In the following section, we would like to go into more detail about DataOps frameworks.

Table 1: DataOps Definitions

Author/ Organization	Year	Type	Definition
Antonio Capizzi, Salvatore Distefano, Manuel Mazzara	(2020)	Scientific Publication	"DataOps is a new approach that aims to improve quality and responsiveness of data analytics life-cycle. This approach is based on DevOps rules, in particular DataOps aims to bring DevOps benefits to data analytics, adopting Agile rules and Lean concepts."
Prabin Ranjan Sahoo, Anshu Premchand	(2019)	Scientific Publication	"DataOps is DevOps based and process-oriented methodology used by data and analytics teams to improve the quality of data analytics and reduce cycle time of the same, generally, to aid better business decision making, improving profits, furthering business and monetizing data available with an organization."
Aiswarya Raj Munappy, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, Anas Dakkak	(2020)	Scientific Publication	"DataOps adopts the best practices, processes, tools and technologies from Agile software engineering and DevOps for governing analytics development, optimizing code verification, building and delivering new analytics thereby promoting the culture of collaboration and continuous improvement."
Julian Ereth	(2018)	Scientific Publication	DataOps is a set of practices, processes and technologies that combines an integrated and process-oriented perspective on data with automation and methods from agile software engineering to improve quality, speed, and collaboration and promote a culture of continuous improvement.
Manuel Rodriguez, Luiz Jonata Pires de Araujo, Manuel Mazzara	(2020)	Scientific Publication	DataOps can be understood as the function within an organization that controls the data journey from source to value. The main focus of DataOps is to improve practices for data management and processes to increase the speed and the accuracy of analytics.
Damian A. Tamburri, Willem-Jan Van Den Heuvel, Martin Garriga	(2020)	Scientific Publication	Like DevOps, DataOps aims to combine the production, operation and delivery (of data) into a single, agile practice that directly supports specific business functions to improve quality, speed, and collaboration and promote a culture of continuous improvement.

Author/ Organization	Year	Type	Definition
Humza Naseer, Sean B. Maynard, Jia Xu	(2020)	Scientific Publication	DataOps combines ideas from information systems research areas including agile development methods, Development and Operations (DevOps), lean thinking and total quality management (TQM) and applies them to data analytics.
Matt Aslett - 451 Research	(2019)	Scientific	DataOps is the alignment of people, process and technology to enable more agile and automated approaches to data management. It aims to provide easier access to data to meet the demands of various stakeholders who are part of the data supply chain in support of a broad range of use cases.
Wayne Eckerson – Eckerson Group	(2019)	Economic Report	DataOps is a set of practices, processes, and technologies for building analytic solutions, including reports, dashboards, self-service analytics, and machine learning models. It applies the rigor of software engineering to the development and execution of data pipelines, which govern the flow of data from source to consumption. The purpose is to accelerate the delivery of data and analytics while simultaneously improving quality and lowering costs. By delivering data "faster, better, cheaper," data teams increase the business value of data and customer satisfaction.
J. Sparapani - MIT Technology Review	(2019)	Economic Report	Data operations, or DataOps, a confluence of advanced data governance and analytics delivery practices that encompasses the entire data life cycle, from data retrieval and preparation to analysis and reporting. Like DevOps, which aims to speed up software development, DataOps incorporates agile and continuous-delivery development methods supported by on-demand IT resources.
Castro et al. -McKinsey Technology	(2020)	Economic Whitepaper	DataOps (enhanced DevOps for data), which can help to accelerate the design, development, and deployment of new components into the data architecture so teams can rapidly implement and frequently update solutions based on feedback.
K. Madera and D.P. Aguilera - IBM	(2020)	Economic Whitepaper	Data operations (DataOps) is the orchestration of people, process and technology to deliver trusted, high-quality data to data citizens fast. The practice is focused on enabling collaboration across an organization to drive agility, speed and new data initiatives at scale. Using the power of automation, DataOps is designed to solve challenges associated with inefficiencies in accessing, preparing, integrating and making data available.

Table 1: DataOps Definitions

2.3. DataOps Frameworks

Despite the novelty of the DataOps methodology, there are some organizations that offer tool systems or enterprise-wide platform that incorporate the core elements of DataOps. For example, the SourceForge portal lists some DataOps tools that incorporate the elements of DevOps, agile development and Big Data analytics and are useful in implementing DataOps in the organization. For example, Nexla offers a platform to integrate, transform, and monitor data at scale, a single platform for all ETL, ELT, Data API, API Integration, or Data as a Service workflows. Or DataKitchen, which offers control of data pipelines to deliver value, without errors. The DataKitchen™ DataOps platform automates and coordinates all the people, tools, and environments in the data analytics organization –from orchestration, testing, and monitoring to development and deployment (https://sourceforge.net/software/dataops/). Due to the possibilities of using different technologies, software and services (which we will discuss in following chapter), it is necessary to stick to holistic processes and ideas

for decision making when establishing DataOps processes in the organization. In this regard, frameworks can help to set a holistic approach in architecture building when establishing new processes and tools. In the following, we would like to introduce some DataOps approaches and concepts.

One concept is presented by the Snowflake organization. Founded in 2012, Snowflake is a cloud-based software-as-a-service organization that provides services in data management, data warehouses, cloud platforms and the like. Snowflake's vision of DataOps illustrates how DevOps principles are applicable to DataOps in the cloud. In doing so, the DataOps framework consists of seven basic pillars that represent feature sets necessary for DataOps success. Figure 7 shows these pillars.

Component design and maintenance aim to build data processes similar to software. This means avoiding large monolithic systems and instead putting small, easy-to-understand, easy-to-maintain-and-test pieces that can then be assembled in needed advanced steps. Environment management describes the setup of production, development and test instances, but also the management of master and feature branch databases to enable CI/CD. Agility and CI/CD here provide DataOps processes that have a repeatable, orchestrated pipeline for all data and schemas to achieve continuous development and continuous integration goals. Automated testing receives a highlighted importance. Current approaches often consist of making infrequent changes to the production landscape and then manually running tests on them, or making frequent changes and not running automated or regular tests. Neither approach is recommended on the premise that data is a strategic asset. Automated testing and an automatically scaling platform can be used to run a large number of tests and ensure the quality of the data processes. Another pillar besides the ETL process, is collaboration and self-service in DataOps environments. Implemented DataOps practices should enable the entire organization to access controlled and high-quality data through structured anonymization.

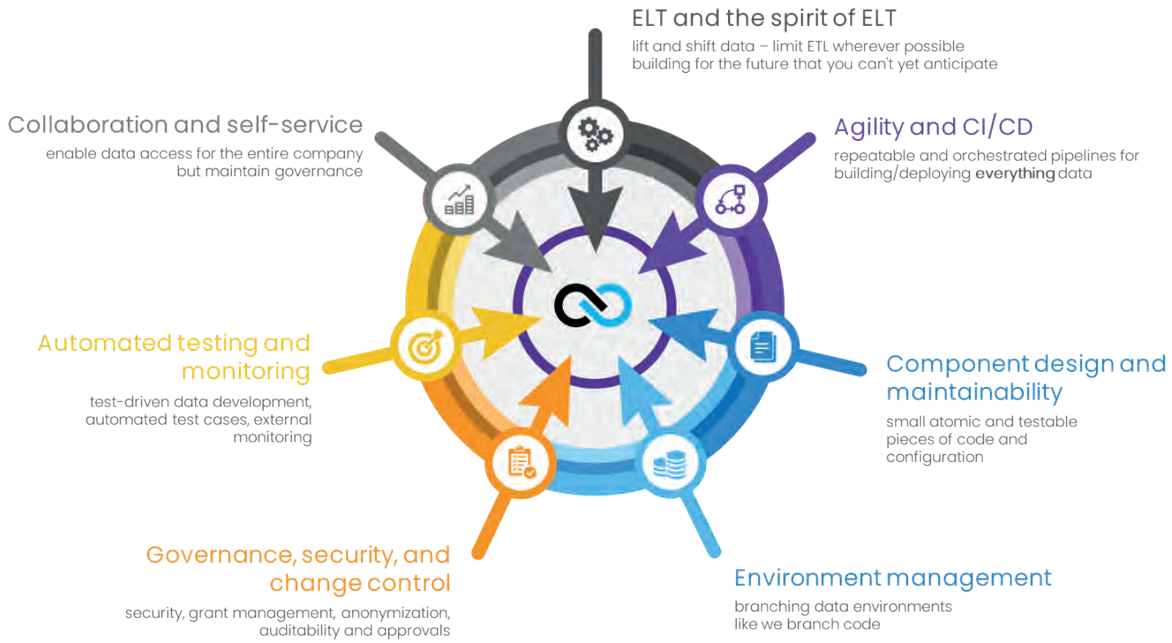

Figure 7: Snowflake's 7 pillars of DataOps[4]

Another DataOps framework is described by the Eckerson Group. In their publications, they use this framework to explain the DataOps process in more detail (Eckerson, 2019a, 2019b). Figure 8 shows the framework. The Eckerson DataOps framework can be divided into three areas. The first area is indicated by the arrow in the middle and describes a typical data pipeline that moves data from one source through three phases, namely data ingestion, data development, and data analysis. Data pipelines thus represent the data supply chain for making data from different systems available to different users and consumers. Data ingestion is usually done by IT/data engineers, data development by data engineers and analysts, and data analysis by data analysts and business users.

The lower part of the framework describes the technologies that are used to capture and analyze data. The technologies are divided into the categories of data integration, data management, data preparation and data analysis. In the data integration technologies for batch and streaming jobs, sql queries, file transfer and replication are located. For data

4 https://www.snowflake.com/blog/the-rise-of-dataops-governance-and-agility-with-truedataops/?lang=de

DataOps Framework

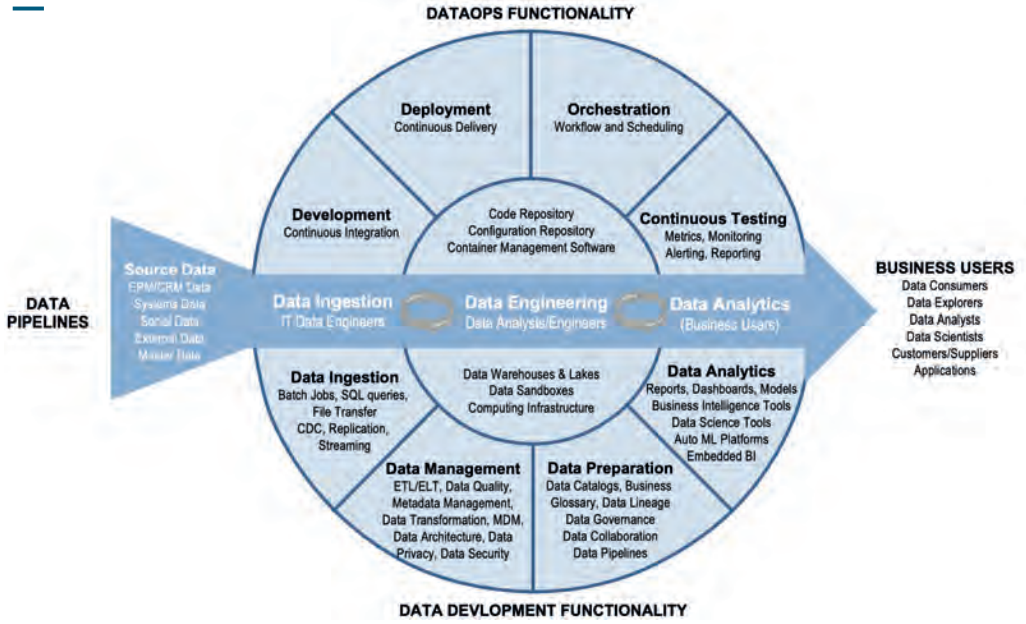

Figure 8: Eckerson DataOps Framework[5] (Eckerson 2019b)

management, according to the framework, technologies are located for the ETL process, for data quality assurance, for metadata management, for data architecture and for data security. For data preparation, technologies such as data catalogs, business glossaries, data lineage tools, data governance tools and data collaboration tools are used. Across categories, technologies are enumerated for computing infrastructure, which is increasingly cloud-based, virtualized, and elastic, as well as technologies for data warehouses and data sandboxes.

The top half of the framework describes the data processes necessary to establish effective DataOps structure in the organization. In addition to the technologies, it requires clearly structured and defined processes and methods for creating, modifying, testing, deploying, executing, and tracking new functionality because of the need to manage all artifacts such as code, data, scripts that are generated in the processes. It therefore requires event triggers, job scheduling, error handling and performance management to achieve a desired service level agreement. The first two phases of the framework are development and deployment. These phases are informed by agile and DevOps methodologies and are well defined. The goals of these phases are to develop new features with self-organized, business-oriented teams that produce fully tested and functional code in short sprints. Development stores and synchronizes codes in a central repository that also provides version control. In addition, tools are needed to seamlessly merge code and move it into production in accordance with continuous integration and continuous deployment. The next category is orchestration. Complex workflows are needed to orchestrate the multitude of tasks with diverse dependencies when creating and traversing a new data pipeline. Tools such as Apache Airflow coordinate the code, data, technology and infrastructure components. In DevOps environments, orchestration tools provide automated development, testing, staging, and production environments by using container management software such as Kubernetes to coordinate containers that support the processes. The final category is continuous testing and monitoring of processes. DataOps teams create tests before developing features and code, which are executed by orchestration tools before and after each phase of the data pipeline. Automatic detection of errors and problems in the spirit of TQM saves time and money and keeps quality standards high. A related capability for optimal uptime and performance is continuous monitoring of tools, applications, and infrastructure. (Eckerson, 2019a, 2019b)

Another framework is offered by the company Zaloni. Zaloni Inc. is a data management software company founded in 2007 by Ben Sharma and Bijoy Bora. Zaloni provides DataOps software for Big Data scale-out architectures such as Amazon AWS, Microsoft Azure and GCP. Among other things, the company focuses on streamlining DataOps for enterprises with its flagship product, the Arena data management platform, which provides end-to-end data ingestion service, metadata catalog, self-service provisioning and data governance (https://www.zaloni.com/). Zaloni's framework for DataOps consists of a feedback loop from which cost savings, accelerated analytics, faster product value, and advanced AI/Ml support are expected to result (see figure 9). At the center of the loop is a collaborative data catalog, which is an inventory of an organization's data assets that provides context by organizing and describing the data assets so that data consumers can discover, understand, and use their needed data sets. The

5 https://www.eckerson.com/articles/diving-into-dataops-the-underbelly-of-modern-data-pipelines

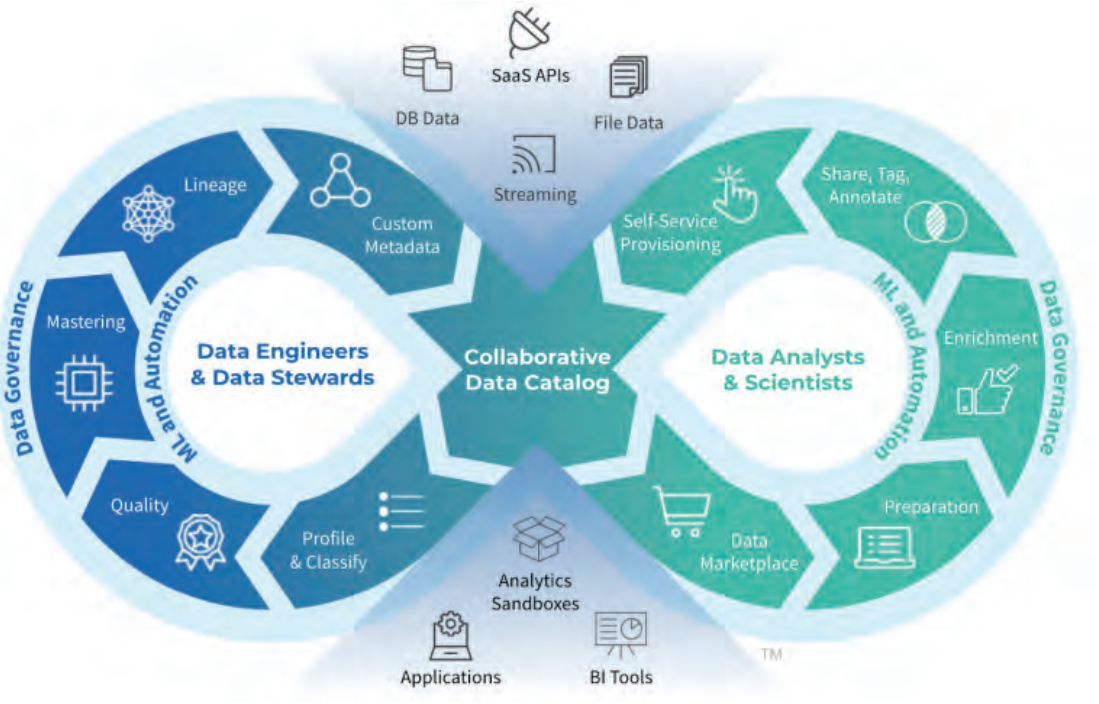

Figure 9: DataOps Cycle according to Zaloni[6]

DataOps feedback loop is designed to align stakeholders such as data stewards, engineers, analysts, and scientists and cover all data governance in the process. The left half of the loop covers the killing of data engineers and data stewards and focuses on activities related to metadata management, data lineage and data quality. The other half of the loop focuses on data analysts and scientists and their data needs, making self-service provisioning, data enrichment, data preparation and data marketplaces central. Zaloni's DataOps methodology considers tools, processes and people to realize collaboration, automation and self-service capabilities. Zaloni's approach is to develop a unified view and control of data pipelines within an organization to streamline DataOps so that it is governed, secure and of high quality.

In the publication »Good practices for the adoption of DataOps in the software industry«, the authors (Rodriguez et al.) provide an insight into the methodology for software development with DataOps and name relevant components. According to the authors, the development of a DataOps system follows an iterative and incremental approach to prioritize different functions and solicit user feedback during development. In doing so, DataOps teams work in iterations that typically last no longer than four weeks, which serves to ensure the project's continuous progress and solicit customer feedback.

Each iteration of the development process should produce functionality that has been tested and monitored at each stage of development. The testing and debugging in this process serves to maintain quality. The introduction of such agile methodologies require that all routine processes are automated. Automated testing thereby shifts the effort from routine tasks to the development of the quality of the functions. The lifecycle of a DataOps process can be seen in (figure 10). As described in the core elements, DataOps follows the same principles as DevOps, but focuses on streamlining data management processes.

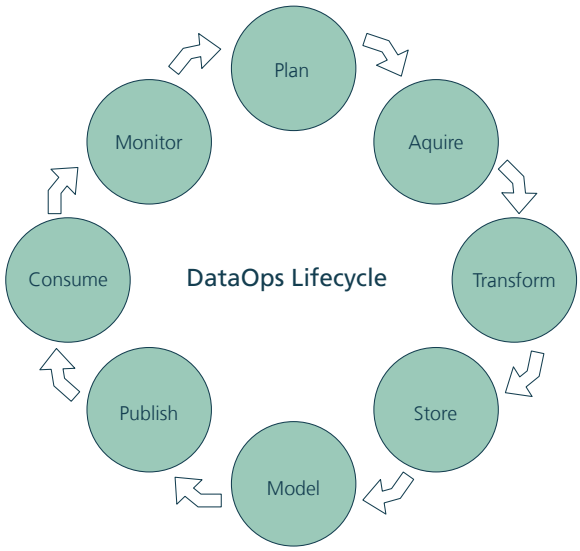

Figure 10: DataOps lifecycle ((Rodriguez et al. 2020)

6 https://www.zaloni.com/resources/what-is-dataops-and-what-is-it-not/

The DataOps components the authors enumerate consist of analytics tools in conjunction with toolchain components. The subcomponents deal with source control management, process management and effective communication between teams. The four main components here are data pipeline orchestration, quality assurance automation, continuous deployment and the use of data science models. Data pipeline orchestration is a software entity responsible for managing processes, handling exceptions, and controlling steps. In this context, DataOps requires a directed and graph-based workflow that fulfills all aspects of data analysis. This includes the process steps of data analysis production such as data collection, access, integration, modeling and visualization. Quality assurance automation lies on continuous monitoring of production quality of both data and artifacts through automated testing. Any changes made to the code must be automatically tested during the deployment process. As described earlier, the continuous deployment process deals with the seamless transfer and adaptation of code in a development environment.
A good practice here is to create sandboxes for a secure development and testing environment. The use of data science models is to gain insight from the data and must be continuously deployed. Continuous changes to the models are an effective way to gain new insights from existing data and thus add value to data consumers. To achieve this, there must be continuous interaction between the data analysis teams, the development team, and the data consumers in the DataOps process. This interaction is described in the following chapter. (Rodriguez et al., 2020)

2.4. Core tenets of DataOps

The collaborative data management practice DataOps aims to improve the communication, integration and automation of data flows across an organization in order to move data from its source to its target destination. Its purpose is to accelerate the delivery of high quality data from various sources. However, this new approach originates from rather known practices. For example, DataOps can be seen as an extension of the DevOps movement (Dhiraj 2019; Sahoo and Premchand 2019) "that uses code repositories, testing frameworks, and collaborative development tools to scale software development, increase code reuse, and automate deployments" (Eckerson 2019b). But next to DevOps, DataOps also applies principles from other mindsets, procedures and production methods like lean manufacturing, total quality management and agile management (Naseer et al. 2020) (see figure 11). In the following we want to give a brief dive into the core tenets of DataOps.

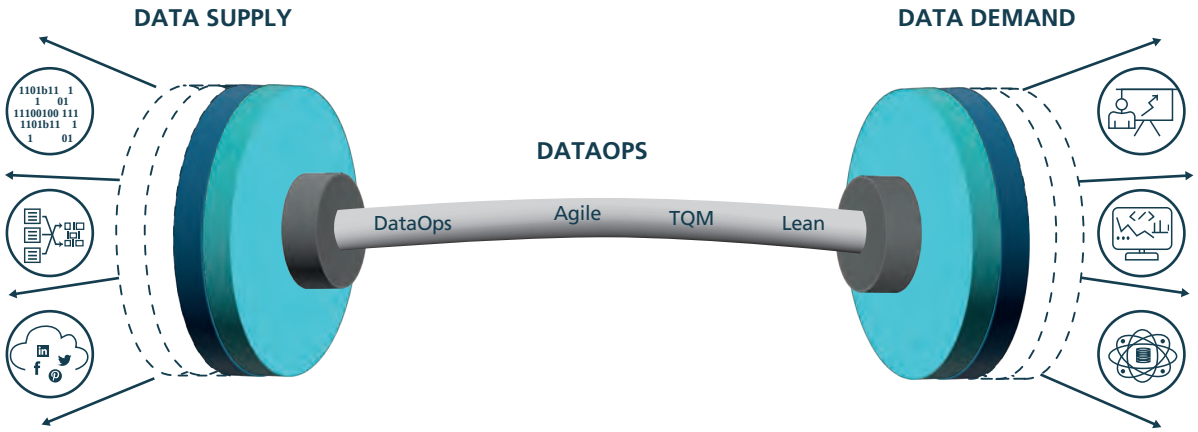

Figure 11: Core Tenets of DataOps according to and adopted from Nexla4[7]

From DevOps to DataOps

The most influencial concept for DataOps and in certain manners its origin and inspiration is DevOps. DevOps is methodology for agile software development (Sahoo and Premchand 2019) and can be seen as a "a set of practices that aim to decrease the time between changing a system and transferring that change to the production environment" (Balalaie et al. 2016). DevOps practices use collaborative development tools like code repositories and testing frameworks to increase code reuse, scale development, automate deployments and accelerate delivery (Eckerson 2019a; Sparapani 2019). The implementation of DevOps is carried out through a set of software tools and collaboration techniques to enable the management of environments and frameworks in which software is built, tested and released in a fast, frequent and

7 https://www.gartner.com/en/information-technology/glossary/dataops

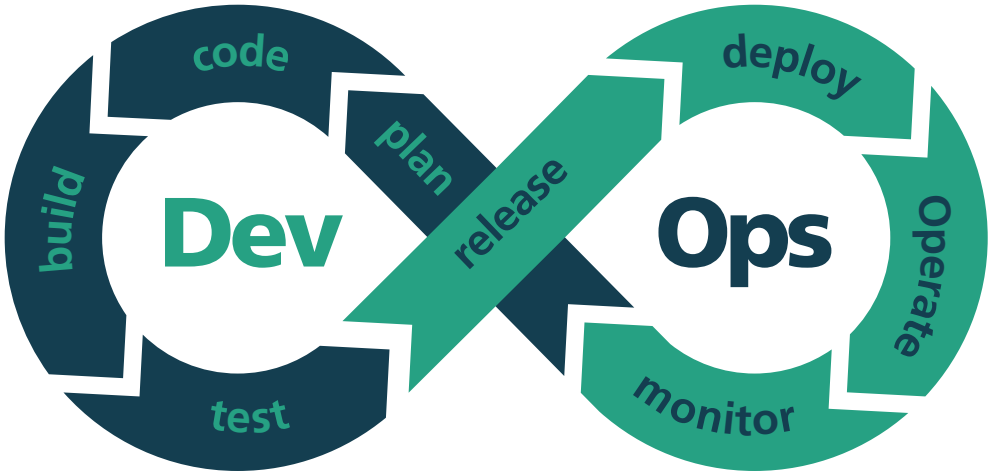

Figure 12: The DevOps process (https://geekflare.com/de/config-management-tools/)

reliable manner (Capizzi et al. 2020). Successful DevOps practices (see figure 12 establish cross-functional collaboration within and between teams to accelerate delivery of changes and increase quality of software and code (Ereth 2018). In terms of data analysis DevOps is used to maintain and update scripts as well as manage tools using continuous integration server and a central code repository (Mainali et al. 2021). Continuous integration "aims at automating the software/ product integration process of codes, modules and parts, thus a CI/CD pipeline" (Capizzi et al. 2020, p. 2). Further important features of DevOps are the continuous delivery process to deliver new features fast and incrementally by implementing a flow of changes of code into a given environment (Capizzi et al. 2020), and continuous deployment, which focusses on movement of code as well as configuration into production from development (Rodriguez et al. 2020; Sahoo and Premchand 2019). Similar to DevOps's automation of deployment and closing of the gap between operational and development team, DataOps brings data pro's (like data architects, data engineers, data, application developers) and data consumers (like business analysts or data analysts) together. It aims at the optimization of code verification, building and delivery of analytical solutions (Naseer et al. 2020). DataOps aims to streamline data management processes to maximize value of massive data stores (Sparapani 2019). According to Andy Palmer, DevOps emerged as traditional systems management failed to meet the needs of modern, web-based application development and deployment and represents the combination of software engineering, quality assurance and technology operations. DataOps, however, acknowledges the interconnected nature of data engineering, data integration, data quality and data privacy and supports an organization's endeavour to rapidly deliver data to accelerate and enable previously impossible analytics (Palmer 2015).

Agile Methods in DataOps

Another core tenet of DataOps are agile collaboration methods. DataOps uses agile development so data analytic teams, data consumers and other stakeholders work together more effectively and efficiently. In agile development self-organizing teams perform short development sprints to deliver fully tested code and regular process reviews (Eckerson 2019a). Innovations in agile environments and teams happen in regular intervals and teams publish new or updated analytics and code in short frequency into the value pipeline (Munappy et al. 2020). Erickson et al. define agility as "means to strip away as much of the heaviness, commonly associated with the traditional software-development methodologies, as possible to promote quick response to changing environments, changes in user requirements, accelerated project deadlines and the like" (Erickson et al. 2005, p. 89).The practice of regular retrospectives not only guarantees that the short sprints deliver high quality and working code but also ensure the business engagement in all initiatives and activities. The primary priority from agile methodologies in DataOps is to grant customer and consumer satisfaction by ensuring that valuable business insights are delivered quickly and continuously. Self-organizing teams are crucial in DataOps and mature agile organizations (Sahoo and Premchand 2019), since they represent process-oriented methodologies to improve cross-functional collaboration, integration and automation of work flow between various stakeholders (Sparapani 2019). There are several agile method approaches consisting of a set of practices that can be applied to software development. They serve as a reaction to traditional and plan-based methods and emphasize a high rate of change rather than rejecting it. The crystal methodologies is a collection of practices for co-located teams of varying size and relevance, identified by different colors. The most agile method, crystal clear, focuses on communication in small teams developing non-critical software. Dynamic software development method (DSDM) divides projects into three

phases: pre-project, project life-cycle and post project. 9 core principles are intended to enable undo changes, scope definition before project start, testing throughout the life-cycle, and efficient and effective communication. Feature-driven development combines model-driven and agile development with a focus on the initial object model, the division of work into features, and the iterative design of each feature. Lean software development is an adaption of principles from lean production and the toyota production system to software development. Scrum focuses on project management for projects that are difficult to plan with mechanisms for process control, where short sprints and feedback loops are the core elements. Software is developed in teams in the sprints that start with planning and end with review. Extreme programming (XP) consists of 13 primary practices divided into team, planning, programming and integration practices. The practices consist of planning game, small realeases, metaphor, simple design, testing, refracting, pair programming, collective responsibility, continuous integration, 40-hour work week, customer on-site, and coding standards. (Atwal 2020; Dybå and Dingsøyr 2008)

Lean Management in DataOps

Many problems or inefficiencies in data science are a result of waste. For example, a study by Nexla shows that much of a data engineer's work is cleaning up data or troubleshooting problems (Nexla Inc., 2018). Such hurdles are obvious cases of waste, both cleaning data unnecessarily and waiting for data or provisioning systems and resources. Waste also exists in other industries, supply chains, or products. Therefore, "Lean Thinking" has been successfully established in many areas such as manufacturing processes (Atwal, 2020). Originated in the 20th century in Japan at Toyota, Lean Thinking is defined by thinking principles, procedures and methods to make value chains efficient and waste-free. It aims to avoid superfluous activities and to optimally coordinate activities in order to remove waiting times and to continuously improve systems from the customer's and the company's point of view. The way for Lean Thinking into software development takes its origin from agile methods in coding. In recent ahren, many companies have adopted agile development. However, its applicability stagnated at the enterprise level. Agile practices were often limited to teams and projects and focused on short-term goals, such as reducing documentation and unnecessary features, potentially negatively impacting total lifecycle costs. Lean attempts to bridge this gap (Ebert et al., 2012). Mary and Tom Poppendick published a book on lean software development several years ago that was closely related to agile software development (Poppendieck and Poppendieck, 2010). Since then, the principles of "eliminate waste", "build quality in", "create knowledge", "defer commitment", "deliver fast", "respect people", and "optimize the whole" have taken a major role in software development (Atwal, 2020). These principles and thought patterns can also be found in DataOps. DataOps practices focus on efficiency. The use of version control systems and code repositories are intended to avoid unnecessary steps, repetitive activities, and erroneous procedures. In addition, DataOps focuses on the customer or data consumer by involving the consumer in the development process in short feedback loops and by establishing self-service infrastructures. By developing simple, standardized and automated processes, care is taken to reduce waste, redundancy and costs (Naseer et al., 2020).

TQM for DataOps

A crucial factor for successful data pipelines and DataOps initiatives are high quality standards. DataOps uses concepts from Total-Quality-Management (TQM) practices for end to end orchestration of the entire process. TQM can be viewed as a process-oriented approach to improve customer satisfaction by offering goods and services of high-quality (Sila 2007). The concept represents a fundamental change in the definition and treatment of quality in product development. It originated in the early 20th century when Walter Shewart introduced statistical controls to reduce variability in testing and experimentation at Bell Labs (Li et al. 2000). In their publication, (Li et al.) apply the "deming's 14 points for management" (Neave 1987) to the managing of the software development process to implement TQM. The guidelines include constancy of purpose, constant improvement, dependence on mass inspection to achieve quality and the breakdown of barriers between departments or staff areas (Li et al. 2000). In order to implement these guidelines in to data management and data ingestion processes, DataOps requires continuous focus in monitoring and improving quality and performance throughout the entire value generation process in a data pipeline (Sahoo and Premchand 2019). It "espouses continuous testing, monitoring and benchmarking to detect issues before they turn into major problems" (Eckerson 2019a, p. 6). The performance of tests and orchestration of data pipeline will be supported by continuous integration tools and deployment is done through continuous development tools. The automation of deployment reduces reconfiguration on the pipeline. Deployment task automation reduces the workload of reconfiguration and reworks on the pipeline in another environment. The implementation of continuous monitoring processes follow the pipeline input, performance, and output and cross-validate the monitoring outcomes with test results from the development environment and business requirements (Mainali et al. 2021). Continuous monitoring and alert management solutions are included in order to automatically test code changes during the process of deployment. The focus of the assurance automation in DataOps lays in automatic testing and monitoring of production quality of data and artefacts in the data analytics production process (see Figure 13).

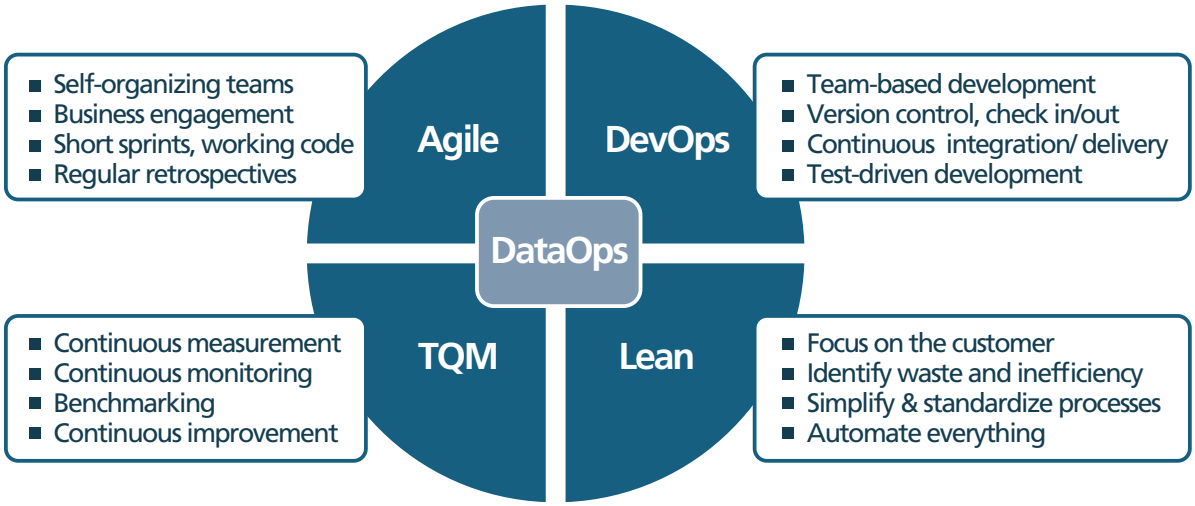

Figure 13: DataOps core tenets (adopted from (Eckerson 2019b)

3. Business Benefits and Challenges of DataOps

Even though DataOps is a fairly new approach, there has been studies on organizations that implemented these practices. These studies show how organizations that have ingrained DataOps in their company culture experience various benefits but also reveal different challenges and obstacles in the implementation of DataOps in the organization.

3.1. Benefits of DataOps

While DataOps is closely associated with operational efficiency, quality and agility, studies show how a mature DataOps approach in the organizational culture can create crucial business benefits for long-term and sustainable competitive advantage. The DataOps practice leverages automation and standardisation to drive significant impact on data curation services, metadata management, data governance (Madera and Aguilera 2020) as well as other business functions. Organizations employ their data to increase efficiency and drive advanced functions like AI-driven and dataintensive applications like IoT, advanced R&D efforts and complex financial analysis (Sparapani 2019).

Some examples of substantial DataOps benefits are shown in the study of 451 Research (Aslett 2019). In this study, the authors conducted a survey with 150 representative organizations with more than 1000 employees. Several industries were included as well a range of different positions and roles covering development, testing, database management, IT operations and business intelligence. The survey respondents were mostly mature in their DataOps adoptions and implementation (see figure 14).

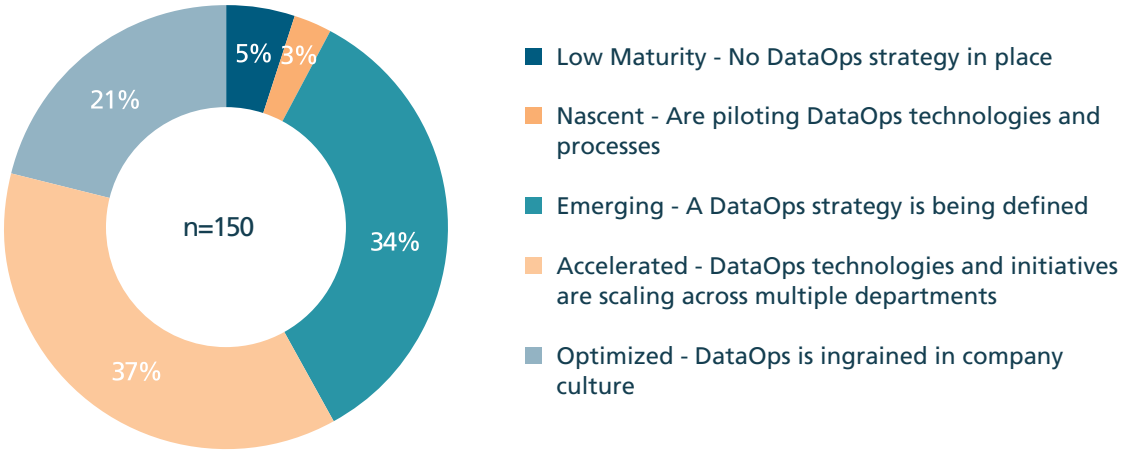

Figure 14: Level of DataOps Maturity of the survey participants (Aslett 2019)

Within the survey, the most mentioned benefit resulting from the implementation of DataOps practices are the alleviation of security and compliance as cross-functional concerns related to data management. Considering the core tenets of DataOps, an obvious advantage was the increased business agility and faster time-to-market (see figure 15).

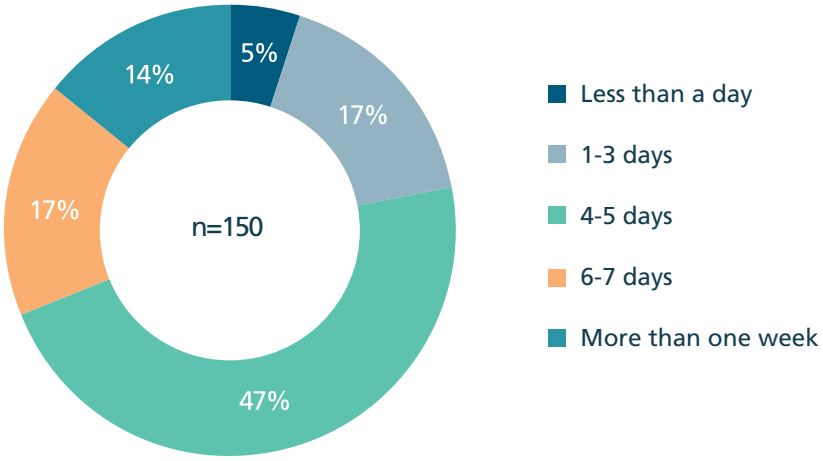

Figure 15: Time to provision a New Data Environment (Aslett 2019)

The responses illustrated that organizations with the greatest adoption of DataOps are successful in involving people from a variety of roles in using data management products and services. DataOps reduced the potential for friction between different group of users involved in data management, like data operators, business executives or data consumers (Aslett 2019) (see figure 16).

Another valuable study has been conducted by Eckerson Group (Eckerson 2019a, 2019b). The researchers coordinated a survey with 175 respondents, mostly from BI directors and managers. The organizations ranged from companies with more than 10000 employees to companies with less than 500 employees. The survey asked the respondents among other things about their conception of DataOps components and their perceived benefits of these practiced methods.

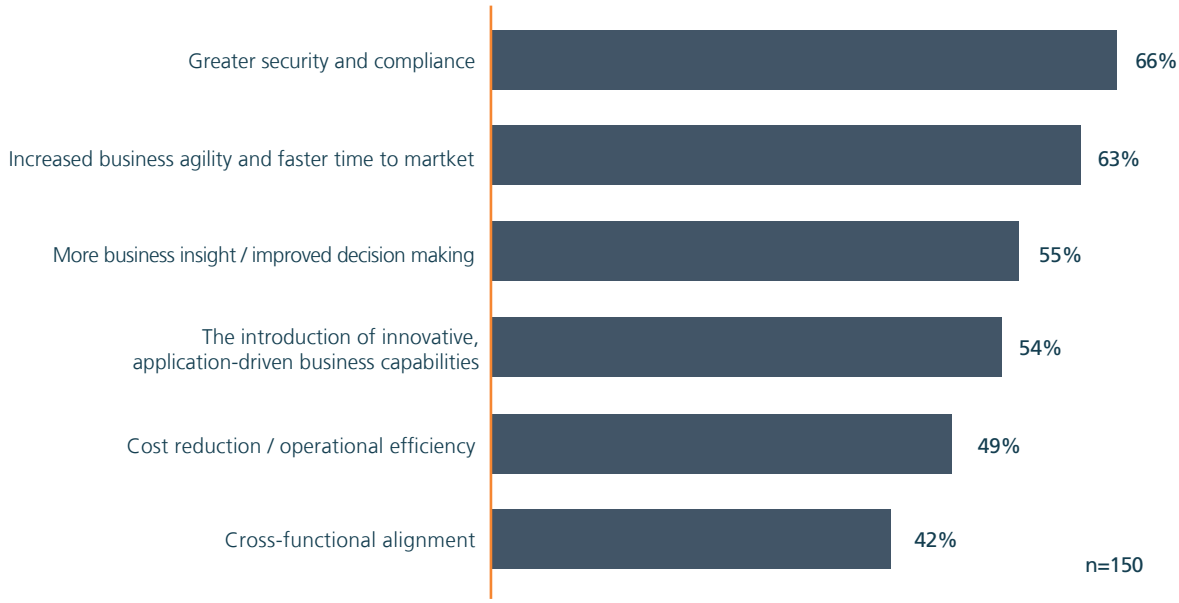

Figure 16: Business Benefits of DataOps (Aslett 2019)

The highest mentioned component was the agile development. Agile, in data warehouse teams, has become a standard practice in those organizations. Therefore, it is appropriate that the biggest benefits of the respondents regarding their DataOps practices are faster cycle times (see figure 17). According to the survey answers, most of the respondents see results in quicker application delivery, rapid ingestion of new data sources and faster change requests due to DataOps investments and initiatives. Other perceived benefits ranged from "fewer defects and errors", which results from the total-quality-management approach of DataOps, and "increased development capacity" as well as "improved data governance", which results from the lean management and agile approach of DataOps (Eckerson 2019b, pp. 13-14).

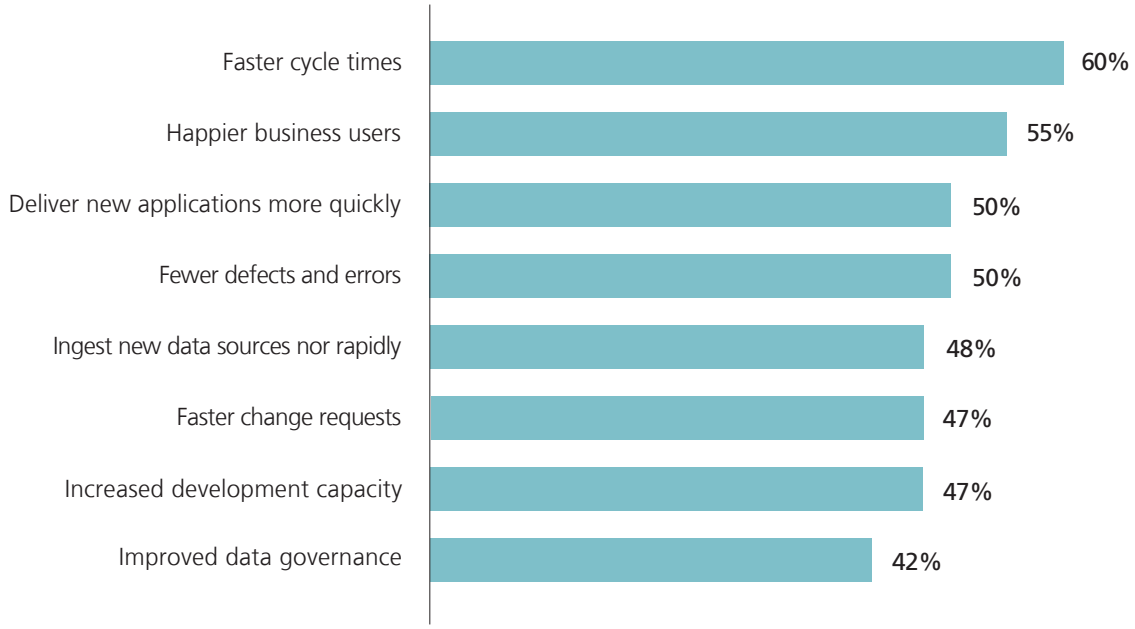

Figure 17: Benefits of DataOps according to (Eckerson 2019b) survey

(Eckerson 2019b)In a report by Sparapani (2019) published in MIT Technology Review Insights, the author provides insights on what benefits successful data analytics and use of data brings (see figure 18).

Figure 18: Business gains from the use of data analytics

According to the paper, the core priorities of DataOps are federated data management, developing pipelines and APIs for real-time operational services, establishing labs to support insights from advanced data analytics, and addressing the regulatory requirements of data. Given the data from the Seagate study shown in the introduction regarding potential barriers to companies not participating in data sharing due to regulatory rules, DataOps offers an opportunity in this regard to expand opportunities for value creation. According to the publication, successful use of data analytics techniques can thus result in improved operational efficiency, enhanced customer experience, achievement of competitive advantage, faster response to opportunities, increase in revenue, reduction in costs, reduction in business errors, and reduction in risk exposure (Sparapani, 2019).

It can be said that successful implementation of DataOps not only brings direct positive impact on business processes and improvement of various business metrics, but also solves barriers related to data sharing. Thus, the implementation of DataOps practices is elemental in the strategic direction of enterprises in the context of digital transformation. However, restructuring current processes and corporate cultural change is a lengthy process. Therefore, in the following chapter, we would like to discuss the hurdles and challenges in the implementation of DataOps.

3.2. Challenges of Data Management and Establishing DataOps

In addition to the benefits that DataOps generates, however, the hurdles and challenges that data management and the creation of value with data entail must still be considered. Since DataOps is not a technology innovation but rather a process innovation, the implementation of DataOps practices is associated with effort and must be completed piece by piece over a period of time. The motto here is evolution instead of revolution. DataOps reflects a true evolution of process, culture and philosophy (Sparapani 2019). Several different studies have already dealt with the hurdles and challenges of data management and DataOps, the results of which we would like to present in the following.
An extensive study on DataOps and the use of data has been conducted by Seagate Technology LCC. The report from Seagate Technology is based on the results of a global study sponsored by Seagate-sponsored global study, conducted between December 2019 and January 2020 by independent market research firm IDC was conducted. The quantitative online study 1,500 people were surveyed worldwide worldwide (375 in North America, 475 in in Europe, 500 in the Asia-Pacific (APJ) region and 150 in China). Participants in the study include CIOs, CTOs, IT VPs, Board members, executives, COOs/LOBs, storage architects and solution architects (Seagate Technology LLC 2020). Some of the results from this study can be seen in figure 19. In this figure are various challenges in realizing the full potential of data.

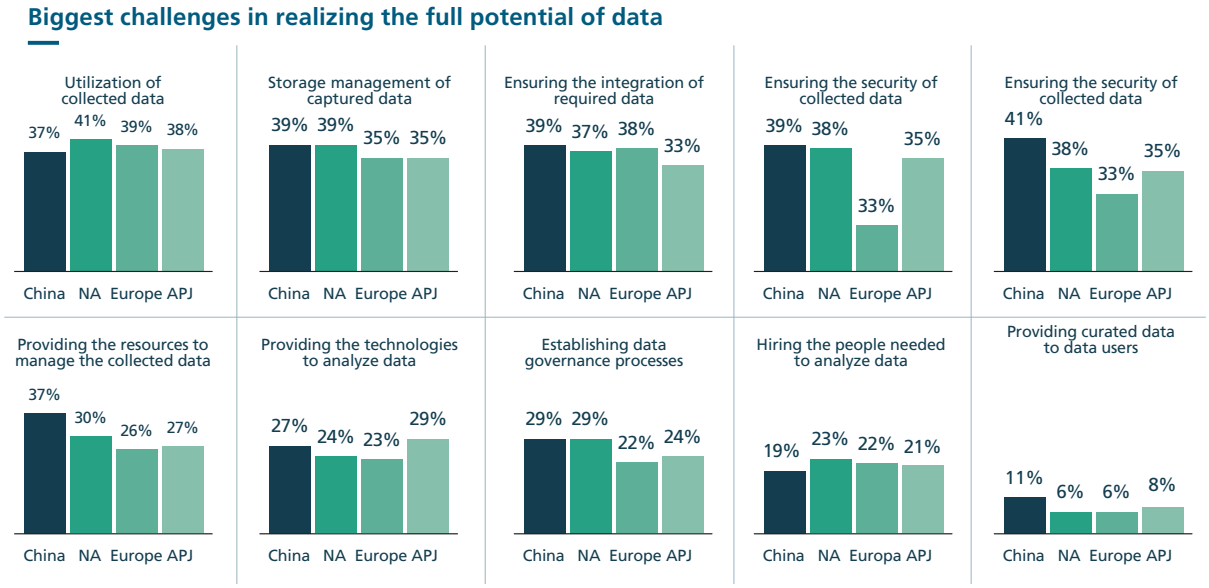

Figure 19: Biggest challenges in realizing the full potential of data according to (Seagate Technology LLC 2020)

The survey data collected is broken down into the regions of China, North America, Europe and Asia-Pacific. For the European region, the challenge in harnessing collected data and storage management of collected data play the most significant role. In addition, ensuring the security of captured data and providing resources to manage captured data played a major role. Along with this were technical challenges such as providing the technologies to analyze data, as well as

organizational challenges such as defining data management processes. In this context, the Seagate analysis also showed that Europe still lags behind in establishing and implementing DataOps practices by comparison. Analysis of the study data shows that Europe tends to react conservatively compared to the other regions, and therefore only 15% of the organizations surveyed rate the importance of DataOps as »extremely important.« This is also reflected in the DataOps status of the companies. The study shows that only 5% of the companies have implemented DataOps capacities completely and company-wide, and 13% have partially implemented DataOps capacities. In the European region, therefore, DataOps practices do not yet appear to be widely implemented to address data management challenges and data analytics challenges.

A 2019 publication in Harvard Business Review looked at data strategies in multi-cloud environments to analyze how organizations can get the maximum economic and strategic value from their data (see figure 20). In doing so, they also analyze the challenges of data management and analysis to gain the most value from their data. Lack of analytical expertise was cited as the biggest challenge, along with lack of internal resources for data analysis. Technical challenges were described in terms of outdated technologies and lack of interoperability between systems. This appeared to lack market solutions that supported the existing environment for the organization. Organizational challenges were reflected in the lack of clear direction from management and organizational silos within the organizations. These organizational silos subsequently led to the creation of data silos as well, or the creation of multiple copies of data, which is why the data quality of the organizations surveyed suffered (Harvard Business Review Analytic Services, 2019). To analyze the challenge of providing the right resources and staff to build and execute data management activities and data analytics, Nexla's survey looked at, among other things (Nexla Inc., 2018). For this purpose, 266 data professionals from 25 industries were asked whether they have enough backend resources in their organization for their activities and what the ratio is between frontend data users such as analysts or business users to backend data engineers. 50% of the respondents said that there are not enough backend resources to fully and consistently support the organization's potential data consumption. Within the survey, it was found that on average there is one backend engineer for every five frontend data professionals. However, there were organizations in which there is only one backend engineer for every 29 data consumers. The survey also asked respondents what activities they do most on average per month. Among the most frequently mentioned activities were »data clean-up/prep« (17%), »managerial activities (12%) and »troubleshooting« (12%). At the same time, however, the classic data operations activities such as »etl jobs« (5%), »data integration« (8%) and »building data pipelines« (5%) were mentioned almost least frequently. Therefore, the greatest challenges faced by the data professionals surveyed are also not surprising. Figure 21 shows the analyses of the responses to the question about the biggest challenges in handling and working with data. Matching the most common activity, data quality issues such as data format consistency are the most frequently cited problem. This is followed by challenges in data backend activities such as data integration processes and access to external data. (Nexla Inc., 2018)

Barriers to maximizing data for strategic gain

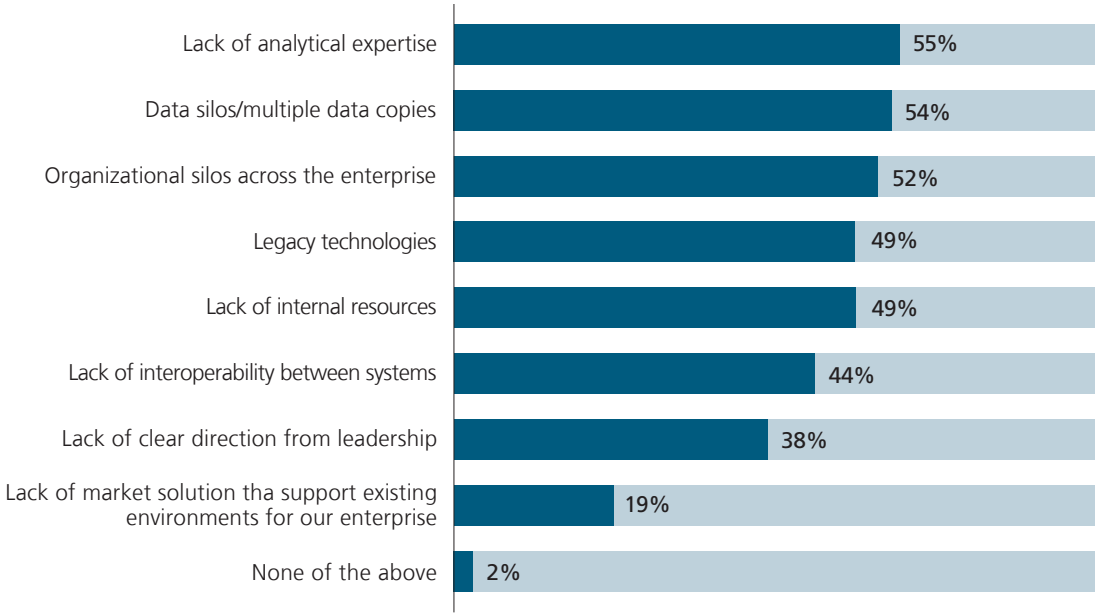

Figure 20: (Harvard Business Review Analytic Services 2019)

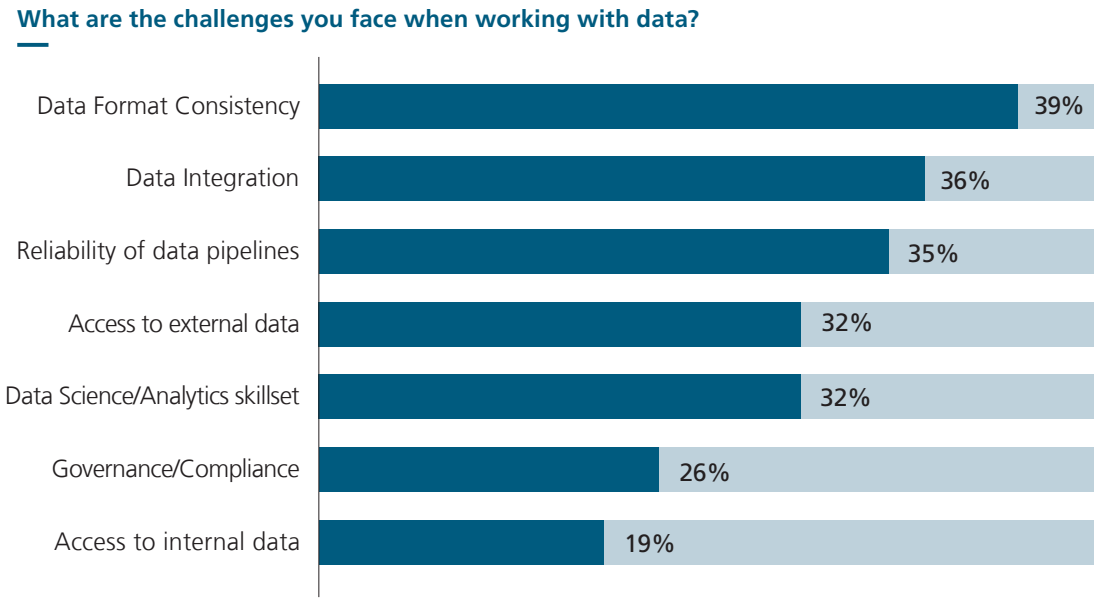

Figure 21: Challenges working with data (Nexla Inc. 2018)

In their survey and analysis, the researchers also asked Eckerson about the challenges of DataOps initiatives. The most frequently cited answer, at 55%, was »establishing formal processes« (see figure 22). The challenge describes that the acquisition of DataOps tools and software is simple, but establishing formal processes and using them effectively is difficult. Procedures and policies need to be defined and established, employees need to be trained in the new processes, and permissions need to be set for who can see what data and code. In this regard, some interviewees described how, when new DataOps initiatives were introduced, regular processes first slowed down as new tools and processes had to be learned and documented (Eckerson, 2019b).

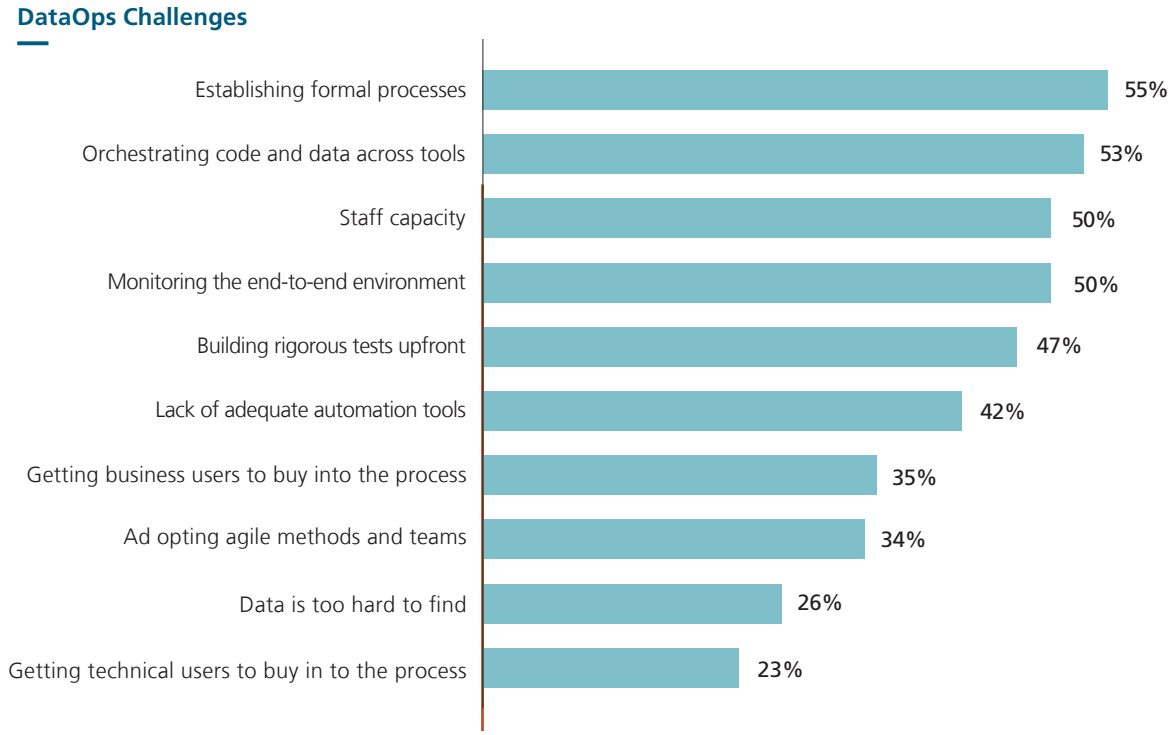

Figure 22: DataOps Challenges (Eckerson 2019b)

Along with this challenge, problems were highlighted that arose due to resistance from developers and backend engineers to new practices. The new DataOps practices were initially described as rigid and restrictive, especially by developers and employees who had previously worked independently and free of formal processes and controls. However, after the initial adjustment difficulties, developers and staff embraced the changes positively, as it provided a framework to better and more quickly deploy code without worrying about breaking anything in production. Furthermore, the new customizations left them more time to deal with activities related to predictive analytics or non-relational data. Another frequently cited challenge was the orchestration of code and data across different tools (53%). Most data pipelines operate in heterogeneous environments where data is retrieved from different source systems and moved through a series of ingestion, transformation, and deployment applications, most of which run on different platforms. Moving the data at an adequate scale and speed is a challenge. Another technical challenge cited was monitoring the entire end-to-end environment (50%), as DataOps is data middleware, between business applications and data infrastructure. In addition, »creating rigorous tests up front,« »lack of automation tools,« and »adopting agile methodologies and teams« were cited as challenges.

The hurdles and challenges described here must be taken into account when implementing DataOps practices in the company so that failures do not jeopardize the long-term digital transformation project. In order to meet with acceptance in this change process, it is therefore important to communicate the change not as a new technology trend, but as a new business and optimization opportunity. The idea of DataOps should therefore be communicated holistically and not by introducing new software. Since DataOps combines many different processes and technologies, as we will show in the further course, it is therefore helpful to make the change and communication process as simple as possible. Communicating the benefits described here serves the goal of involving all stakeholders in this process. Furthermore, the implementation of DataOps practices should be done in small steps so that ongoing processes are not hindered or disrupted by new structures. Nevertheless, it can be stated that the benefits of successfully implemented DataOps justify the effort. Participation in the digital data economy and in future national and international initiatives for data ecosystems and data spaces hold opportunities for progress and innovation. Therefore, effective data management to make participation in data sharing as efficient and successful as possible is relevant for the strategic direction of organizations that want to walk the path of digital transformation to a data-driven organization.

4. Technology and organizational Structure in DataOps

In the following, we would like to delve deeper into the topic of DataOps. Even though DataOps is a fairly new practice, there are already proven best practices and proven technologies. For this purpose, we will look at various DataOps frameworks and concepts that are currently offered by different companies. In addition, we give a brief overview of technologies and tools in DataOps as well as DataOps services that are currently offered by cloud providers. Finally, we give an insight into how DataOps can enable collaboration across functional boundaries using different roles and improve the data governance of an organization.

4.1. DataOps in the Cloud – Examples of AWS, IBM and Azure

The insight into DataOps frameworks shows that although DataOps is more of a process innovation than a technology innovation, it is essential to consider options regarding technologies and tools. In doing so, the frameworks show that many different types of tools and software are needed to realize and orchestrate data pipelines. However, the flexibility and scalability of software and services sometimes present challenges. Today's services provided by global cloud computing providers enable organizations to leverage a variety of technologies and trends. Therefore, to remain competitive, it is imperative for enterprises to incorporate cloud computing into their strategic digital transformation and operations. Cloud computing services provide the flexibility, scalability, security and agility needed to achieve strategic and long-term business goals. Various cloud providers already offer solutions for customers to establish DataOps processes in their organizations. In the following, we would like to present some such examples.

One example to implement DataOps using AWS services is provided by S. D'Souza, B. Repeka and M. Fung (D'Souza et al. 2020). They implemented DataOps atop a data lake on AWS. The application of the DataOps practice was conducted through ten steps using AWS technology. To add data and perform logic tests, Git and AWS CodeBuild has been used. AWS CodeBuild is a fully managed continuous integration service to compile source code, run tests, and generate deployable software packages. For the version control system AWS offers the CodeCommit. AWS CodeCommit is a secure, highly scalable, managed source code management service that hosts private Git repositories. It makes it easy for teams to securely collaborate on code with encrypted contributions in transit and at rest. For the merging and brange process, AWS CodePipeline and Lambda were used. AWS Lambda is a serverless, event-driven computing service that lets you run code for virtually any type of application or backend service without provisioning or managing servers. In order to use multiple services within the architecture, AWS CloudFormation has been used. With AWS CloudFormation, you can model, provision, and manage AWS and third-party resources by treating infrastructure as code. Docker and Amazon EKS have been used to build container. Amazon Elastic Kubernetes Service (Amazon EKS) is a managed service that you can use to run Kubernetes on AWS without needing to install, operate, and maintain your own Kubernetes control plane or nodes. Kubernetes is an open-source system for automating the deployment, scaling, and management of containerized applications. In order to parametrize the processes, Amazon EMR, Amazon CloudWatch, AWS Step Functions, and Amazon Simple Notification Service (Amazon SNS) have been used. Amazon CloudWatch for example is a transparent monitoring and observation service for DevOps engineers, developers, site reliability engineers (SRE), and IT managers. CloudWatch provides data and meaningful insights to monitor applications, respond to system-wide performance changes, optimize resource usage, and get an overall view of operational health. The data storage has been set up using Amazon S3. All the processed data went into Amazon S3, as the service offers an object storage service that delivers scalability, data availability, security, and performance. So customers of all sizes and from all industries can use this service to store and back up any amount of data for a wide variety of use cases, including data lakes, websites, mobile apps, backup and recovery, archiving, enterprise applications, IoT devices, and Big Data analytics. For the data management processes, AWS Database Migration Service (AWS DMS), AWS Glue Data Catalog, and Amazon

Athena have been utilized. AWS DMS for example helps to migrate databases to AWS quickly and securely. The source database remains fully operational during the migration, minimizing downtime for applications that depend on the database (D'Souza et al. 2020). Figure 23 shows examplary an AWS DataOps Solution Architecture. The authors of this solution give an overview on the technology, prerequisites and processes within the solution. According to them, the implementation of DataOps process for data analysts requires business logic and tests in SQL, code submission to a Git repository, code review and automated tests as well as deployment in a running production data warehouse (Softic et al. 2021).

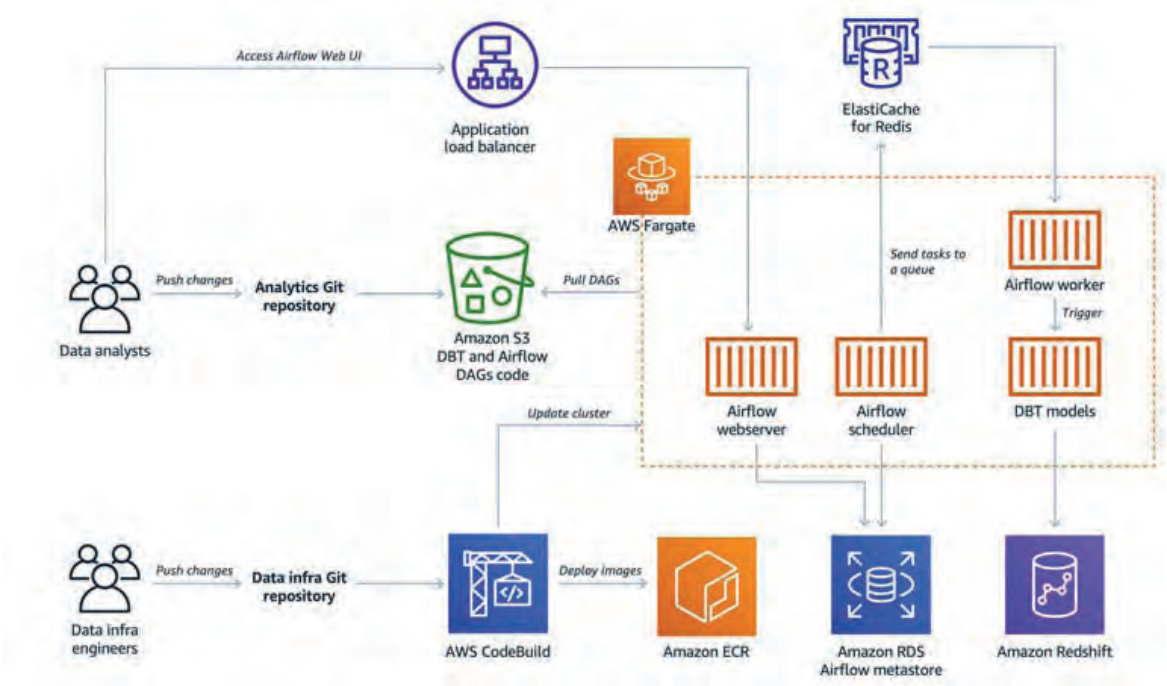

Figure 23: Amazon DataOps Solution Architecture (from (Softic et al. 2021))

Another way to flexibly and scalably establish DataOps technologies in the organization is offered by Microsoft's Azure cloud platform. In their article, the company explains how DataOps processes can be used with Azure Cloud Services to leverage a modern data warehouse for data aggregation (Azure 2020). Through the process described, structured, unstructured and partially structured data can be used to agilely provide analytics dashboards, operational reports and advanced analytics to the organization. Their article describes a use case scenario of how data from urban sensors can be transferred, cleansed, validated, and then used to apply data visualization tools and create reporting data. Their use case creates a DataOps Solution Architecture, shown in Figure 24. The Solution Architecture uses the following solutions. Azures Data Factory

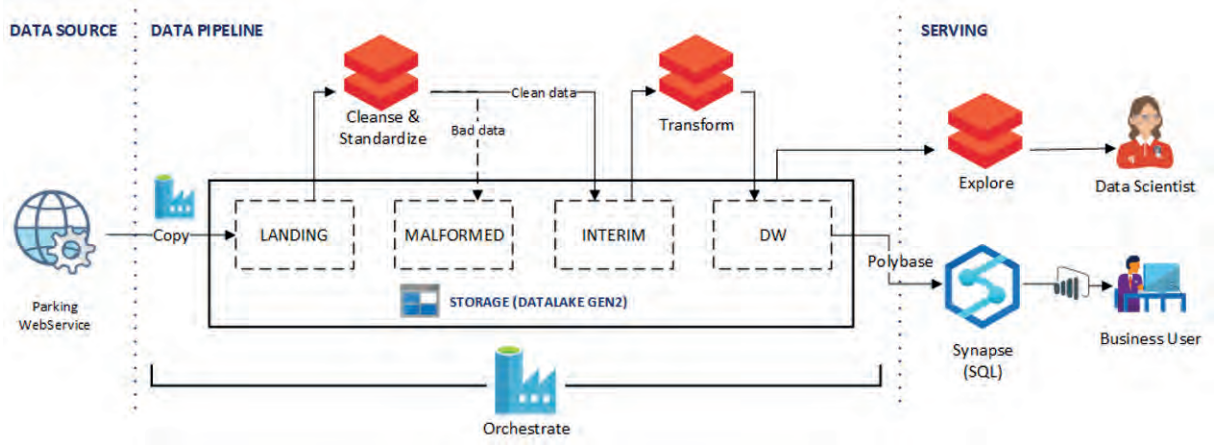

Figure 24: Microsoft Azure DataOps Solution Architecture (Azure 2020)

is used to integrate the data. The cloud service provides a fully legacy and serverless data integration service to orchestrate the data. Data cleansing and standardization takes place with the help of Azure Bricks. Azure Bricks can be used to establish rich data processing for batch and streaming workloads, data analytics, machine learning environments and experiment monitoring in the DataOps process. The basis for storing the data is provided by Data Lake Storage Gen2 for creating enterprise data lakes. In this Data Piepline, data is provided in two ways. On the one hand, Data Scientists get access to the data with the help of Databricks in order to be able to train models. Second, the data is moved from the Data Lake to Azure Synapse Analytics using Polybase, and Power BI can then access the data to make it available to a business user. Azure DevOps can be used to perform CI/CD and schedule and build development services for the AI models around team. Azure Key Vault secures cryptographic keys and other relevant objects used by cloud applications and services (Azure 2020).

Among others, IBM's cloud platform also offers solutions for DataOps processes (see figure 25). The organization began early to provide DataOps techniques and „IBM is a pioneer of the term DataOps and offers numbers of products for different data lifecycle stages (e.g., data collecting, analysis, storage, organization, and publishing) under the umbrella of their cloud service" (Mainali et al. 2021, p. 62). IBM defines DataOps as "the orchestration of people, process and technology to deliver trusted, high-quality data to data citizens fast. The practice is focused on enabling collaboration across an organization to drive agility, speed and new data initiatives at scale. Using the power of automation, DataOps is designed to solve challenges associated with inefficiencies in accessing, preparing, integrating and making data available" (Quoma 2019). According to (Quoma), for the consideration of using tools and technology for the DataOps practice of an organization, there needs to be automation in data curation services, metadata management, data governance, master data management and ser-service interaction in order to transform a data pipeline. For that IBM offers several cloud services, to realize DataOps practices in organizations. According to IBM, the data governance catalogs, protects and governs all data types, traces data lineage, and manages data lakes. For that, they offer the IBM Watson knowledge Catalog, which is a modern data catalog designed to help data scientists, data governance professionals and business analysts activate data for AI, business operations and analytics. For the data integration, replication and virtualization

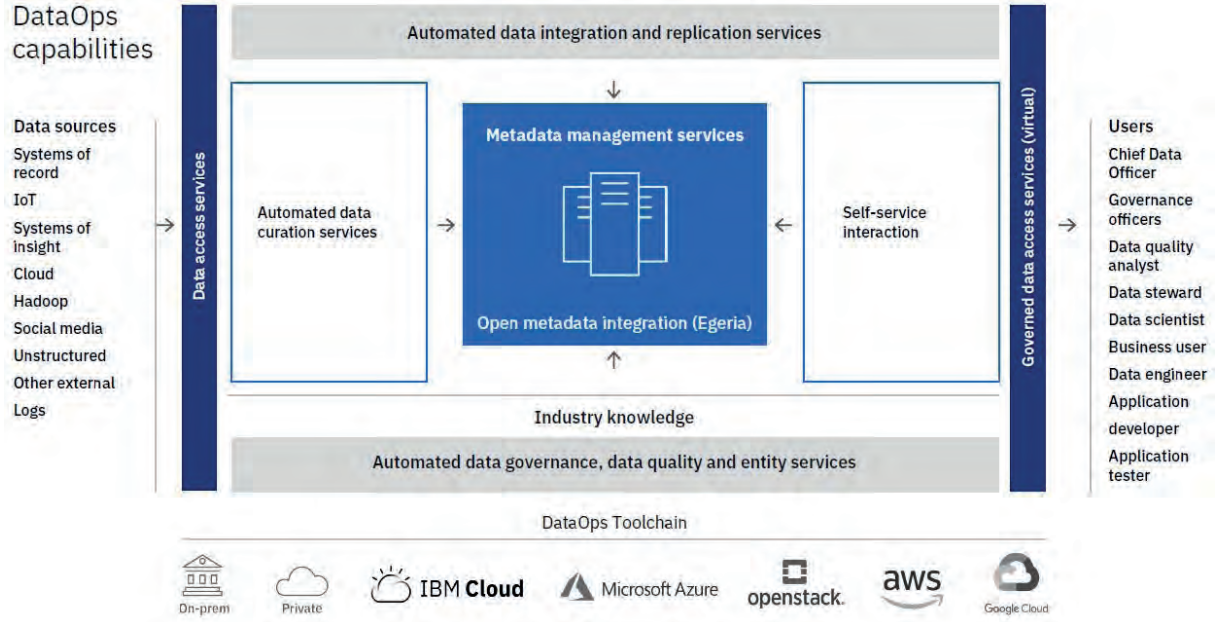

Figure 25: Examining an information architecture in support of DataOps (Quoma 2019)

in real time, IBM offers IBM DataStage, which is a scalable data integration tool for designing, developing and running jobs that move and transform data.

For the master data management, IBM offers the IBM Cloud PAK for data, which is a flexible multicloud data platform that integrates data, whether on premises or on any cloud (IBM 2021).

4.2. Improving your Data Governance with DataOps

As already mentioned in previous chapters, a strategy for implementing DataOps in the organization must not only consider the technological level, but first and foremost corporate cultural aspects and the behaviour of the respective stakeholders in a data value chain. Modern and data-driven organizations are often highly complex, as a structural flow of information requires breaking down internal silos to allow data to flow throughout the organization (Hupperz et al. 2021). However, many organizations are hindered in their value creation by this complexity. A study by McKinsey, for example, indicates that leaders in organizations often have too little understanding of the form of complexity in their structures and processes and whether that complexity creates or reduces value. While the study distinguishes between institutional and individual complexity, it finds that organizations with low and controlled complexity have higher returns and lower costs. It shows that reducing and controlling organizational complexity at both the employee and executive levels can serve as a competitive advantage and a basis for successful growth (Heywood et al., 2010). Complexity in data-driven organizations and data teams stems from the multitude of roles, functions, and business units that work with data together or in parallel. Dissonance of tools, departments, roles, and critical outcomes has made managing and leading the data organization challenging. As mentioned earlier, Nexla's study describes how data engineers often have to deal with challenges due to project failures, delayed schedules, and poor data quality (Nexla Inc., 2018). Implementing DataOps practices seeks to address this complexity and establish effective data governance. Among other things, DataOps aims for data governance to improve the ability to respond to regulatory requirements as well as to accelerate the definition of governance rule to accelerate value creation activities with data. The trend points to the fact that companies with mature DataOps also have mature data governance (Hyperight, 2020). In this regard, DataOps supports highly productive teams with automation technologies to efficiently and successfully manage the collaboration of roles and teams that rarely work together. In their publication, (Rodriguez et al.) show with which business units DataOps has a direct impact. Several different functional groups within an organization work collaboratively with the goal of establishing DataOps practices to create value in production and data value creation. Figure 26 highlights these different groups. Data Scientists apply techniques that include machine learning and deep learning models.

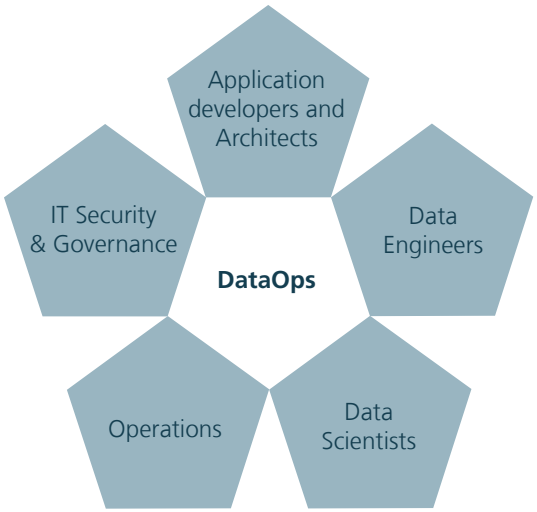

Figure 26: Cross-functional collaboration required for DataOps process (Rodriguez et al. 2020)

They build rigorous and efficient models using frameworks available, for example, in the form of software libraries in Python or R, as well as Big Data processing tools such as Spark or Tensorflow and others. Data Engineers deal with the aggregation of data and manage the data used to evaluate and train models. End-to-end applications are created by developers and architects, incorporating the models created by Data Scientists. In this process, data governance and IT define the access controls that allow data scientists to access historical and relevant data. Such controls can be accessed in a read-only state to optimize and reflect on models. The DevOps team is responsible for deploying applications to production environments to support service-level agreements (Rodriguez et al., 2020).

A critical part of DataOps is the development of a corporate culture based on collaboration. It must, as mentioned earlier, bring together groups and functions that otherwise do not interact with each other on a daily basis. A successful DataOps practice creates a smooth collaboration between developers, analysts, and consumers (Sparapani, 2019). In doing so, DataOps focuses the division of labor and access rights to data and code between multidisciplinary teams from business and technical business functions. The principle of collaboration thereby creates the agile and cross-functional teams consisting of different roles. In this way it is possible to better understand business requirements and deliver BI solutions. (Naseer et al. 2020). The way of digital transformation of an organization starts from understanding the goal and strategy to extract the maximum value from data (Gür and Spiekermann, 2020). To do this, strategic issues must be clarified about decision rights and services that affect the customer, what competitive advantage can be gained, and what strategic partnerships

can be fulfilled using data sharing. According to the strategic direction, executives should then define roles that all people in the data value creation process will take to be able to establish a DataOps culture in the organization. The unique needs and goals of the stakeholders must be considered in terms of the business goals in order to be able to ensure a smooth process within the company. With the help of defined roles and responsibilities using data governance committees, for example, the development and production culture that is aimed for can thus be established in the company and quality features such as data security and integrity can be ensured. Figure 27, for example, shows a sequence of roles that depicts a generalized workflow for data provision.

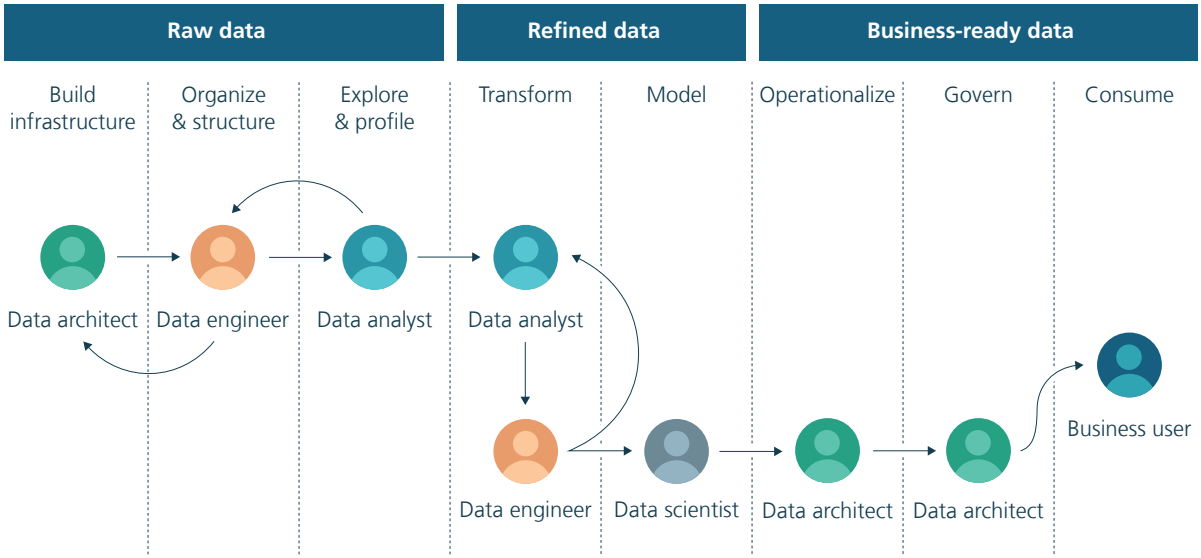

Figure 27: Example of a DataOps workflow by (Madera and Aguilera 2020)

In summary, the implementation of DataOps practices is a recommended methodology to standardize and automate the development process of new data environments and data pipelines, with the goal of accelerating processes, avoiding errors, minimizing waste of resources, and maintaining quality standards for both activities and data. DataOps helps inter- and intra-organizational collaboration, which can break down data silos in organizations and drive the empowerment of employees and companies to make data-driven decisions.

In doing so, DataOps offers the opportunity to counter the challenges of data management, enabling participation in the data economy and data ecosystems. As described in this report, organizations today still face challenges stemming from ignorance of required resources, outdated technologies, lack of strategic alignment, and obstructive organizational structures. As a result, the use of external data or the provision of high-quality data is not possible for many organizations, or the benefits did not justify the effort required. Thus, many organizations were denied participation in data ecosystems.

Using the DataOps methodology, existing resources can be used in a targeted and efficient manner and it provides clarity on what other resources, expertise and technologies are needed to make data management successful. By streamlining and standardizing the development process of new data environments, models and data pipelines, organizations are able to quickly leverage new data sources and thus participate in existing and emerging data ecosystems, fostering the growth of data spaces in Germany and Europe and solving barriers to entry.

A dissemination of DataOps practices in organizations can therefore promote the fulfilment of national and European data strategies. This report is intended to serve as an introduction to the methodology and to provide an overview with the aim of supporting and facilitating the digital transformation of organizations.

5. Contact

We would be pleased to give you an understanding of our offer with regard to the design of DataOps and to cooperate with you. Together we will take this important step in the digital transformation and make your company fit for the future.

Contact us without obligation for a first conversation!

Inan Gür
Data Business, Scientist
Tel. +49 231 97677-418
inan.guer@isst.fraunhofer.de

Fraunhofer Institute
for Software and Systems Engineering ISST
Emil-Figge-Straße 91
44227 Dortmund
www.isst.fraunhofer.de

Markus Spiekermann
Data Business, Head of Department
Tel. +49 231 97677-424
markus.spiekermann@isst.fraunhofer.de

Fraunhofer Institute
for Software and Systems Engineering ISST
Emil-Figge-Straße 91
44227 Dortmund
www.isst.fraunhofer.de

6. Acknowledgement

This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the IEDS0001 and implemented by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the content of this publication.

7. References

ASLETT, M., 2019. DataOps Lays the Foundation for Agility, Security and Transformational Change.

ATWAL, H., 2020. Practical DataOps. Berkeley, CA: Apress. ISBN 978-1-4842-5103-4.

AZURE, M., 2020. DataOps für das moderne Data Warehouse [online] [Zugriff am: 12. November 2021]. Verfügbar unter: https://docs.microsoft.com/de-de/azure/architecture/example-scenario/data-warehouse/dataops-mdw

BALALAIE, A., A. HEYDARNOORI und P. JAMSHIDI, 2016. Microservices Architecture Enables DevOps: Migration to a Cloud-Native Architecture [online]. IEEE Software, 33(3), 42-52. ISSN 0740-7459. Verfügbar unter: doi:10.1109/MS.2016.64

CAPIZZI, A., S. DISTEFANO und M. MAZZARA, 2020. From DevOps to DevDataOps: Data Management in DevOps processes [online]. In: Software Engineering Aspects of Continuous Development and New Paradigms of Software Production and Deployment. Verfügbar unter: http://arxiv.org/pdf/1910.03066v1

CASTRO, A., J. MACHADO, M. ROGGENDORF und H. SOLLER, 2020. How to build a data architecture to drive innovation—today and tomorrow. McKinsey Technology.

DHIRAJ, S., 2019. Nach DevOps kommt DataOps. Wirtschaftsinformatik & Management. Wirtschaftsinformatik & Management.

DIE BUNDESREGIERUNG – BUNDESKANZLERAMT, 2021. Datenstrategie der Bundesregierung.

D'SOUZA, S., B. REPAKA und M. FUNG, 2020. Modern data engineering in higher ed: Doing DataOps atop a data lake on AWS [online] [Zugriff am: 12. November 2021]. Verfügbar unter: https://aws.amazon.com/de/blogs/publicsector/modern-data-engineering-higher-ed-dataops-data-lake/

DYBÅ, T. und T. DINGSØYR, 2008. Empirical studies of agile software development: A systematic review [online]. Information and Software Technology, 50(9-10), 833-859. ISSN 09505849. Verfügbar unter: doi:10.1016/j.infsof.2008.01.006

ECKERSON, W., 2019a. Best Practices in DataOps How to Create Robust, Automated Data Pipelines.

ECKERSON, W., 2019b. Trends in DataOps Bringing Scale and Rigor to Data and Analytics.

ERETH, J., 2018. DataOps - Towards a Definition. In: Lernen, Wissen, Daten, Analysen.

ERICKSON, J., K. LYYTINEN und K. SIAU, 2005. Agile Modeling, Agile Software Development, and Extreme Programming [online]. Journal of Database Management, 16(4), 88-100. ISSN 1063-8016. Verfügbar unter: doi:10.4018/jdm.2005100105

EUROPEAN COMMISSION, 2020. The European Data Strategy.

EXPERIAN LTD., 2019. 2019 Global data management research - Taking control in the digital age.

GELHAAR, J., T. GUERPINAR, M. HENKE und B. OTTO, 2021. Towards a Taxonomy of Incentive Mechanisms for Data Sharing in Data Ecosystems. In: Twenty-fifth Pacific Asia Conference on Information Systems,

GÜR, I., M. SPIEKERMANN, M. ARBTER und B. OTTO, 2021. Data Strategy Development A Taxonomy for Data Strategy Tools and Methodologies in the Economy. In: Internationale Konferenz Wirtschaftsinformatik.

HARVARD BUSINESS REVIEW ANALYTIC SERVICES, 2019. Critical success factors to achieve a better enterprise data strategy in a multi-cloud environment. Harvard Business Review Analytic Services.

HUPPERZ, M., I. GÜR, F. MÖLLER und B. OTTO, 2021. What is a Data-Driven Organization? In: American Conference on Information Systems.

HURLEY, J., 2018. WHY YOUR DATA STRATEGY IS YOUR B2B GROWTH STRATEGY. Harvard Business Review. ISSN 00178012.

IBM, 2021. IBM DataOps Organize your data to be trusted and business-ready for your Journey to AI [online] [Zugriff am: 12. November 2021]. Verfügbar unter: https://www.ibm.com/analytics/dataops

LI, E.Y., H.-G. CHEN und W. CHEUNG, 2000. Total Quality Management in Software Development Process. The Journal of Quality Assurance Institute, (14). The Journal of Quality Assurance Institute.

MADERA, K. und D.P. AGUILERA, 2020. Deliver business-ready data fast with DataOps. An introduction to the IBM DataOps methodology and practice.

MAINALI, K., L. EHRLINGER, M. MATSKIN und J. HIMMELBAUER, 2021. Discovering DataOps: A Comprehensive Review of Definitions, Use Cases, and Tools. In: DATA ANALYTICS 2021 : The Tenth International Conference on Data Analytics.

MIELLI, F. und N. BULANDA, 2019. Digital Transformation: Why Projects Fail, Potential Best Practices and Successful Initiatives. In: 2019 IEEE-IAS/PCA Cement Industry Conference (IAS/PCA): IEEE, S. 1-6. ISBN 978-1-7281-1160-5.

MUNAPPY, R.A., D.I. MATTOS, J. BOSCH, H.H. OLSSON und A. DAKKAK, 2020. From Ad-Hoc Data Analytics to DataOps. In: International Conference on Software and Systems Process. Seoul.

NASEER, H., S.B. MAYNARD und J. XU, 2020. Modernizing business analytics capability with DataOps: A decision-making agility perspective. In: European Conference on Information Systems.

NEAVE, H.R., 1987. Deming's 14 points for management: framework for success. The Statistician. The Statistician.

NEXLA INC., 2018. The Definitive Data Operations Report.

OLIVEIRA, M.I.S. und B.F. LÓSCIO, 2018. What is a data ecosystem? In: M. JANSSEN, S.A. CHUN und V. WEERAKKODY, Hg. Proceedings of the 19th Annual International Conference on Digital Government Research Governance in the Data Age - dgo '18. New York, New York, USA: ACM Press, S. 1-9. ISBN 9781450365260.

PALMER, A., 2015. From DevOps to DataOps [online] [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://www.tamr.com/blog/from-devops-to-dataops-by-andy-palmer/

QUOMA, S., 2019. What is DataOps? [online]. Organize your data to be trusted and business-ready for your Journey to AI [Zugriff am: 12. November 2021]. Verfügbar unter: https://www.ibm.com/blogs/journey-to-ai/2019/12/what-is-dataops/

RODRIGUEZ, M., L.J.P. de ARAÚJO und M. MAZZARA, 2020. Good practices for the adoption of DataOps in the software industry [online]. Journal of Physics: Conference Series, 12032. ISSN 1742-6588. Verfügbar unter: doi:10.1088/1742-6596/1694/1/012032

SAHOO, P.R. und A. PREMCHAND, 2019. DataOps in Manufacturing and Utilities Industries. International Journal of Applied Information Systems, (12). International Journal of Applied Information Systems.

SEAGATE TECHNOLOGY LLC, 2020. Rethink Data. Bessere Nutzung von mehr Unternehmensdaten – vom Netzwerkrand bis hin zur Cloud.

SILA, I., 2007. Examining the effects of contextual factors on TQM and performance through the lens of organizational theories: An empirical study [online]. Journal of Operations Management, 25(1), 83-109. ISSN 02726963. Verfügbar unter: doi:10.1016/j.jom.2006.02.003

SOFTIC, D., T. BERGER, D. GREENSHTEIN und R. BASSETTO, 2021. Build a DataOps platform to break silos between engineers and analysts [online] [Zugriff am: 12. November 2021]. Verfügbar unter: https://aws.amazon.com/de/blogs/big-data/build-a-dataops-platform-to-break-silos-between-engineers-and-analysts/

SPARAPANI, J., 2019. DataOps and the future of data management.

TAMBURRI, D.A., W.-J. VAN DEN HEUVEL und M. GARRIGA, 2020. DataOps for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction and Matching [online]. In: IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), S. 391-394. Verfügbar unter: http://arxiv.org/pdf/2104.01966v1

THUERMER, G., J. WALKER und E. SIMPERL, 2019. Data Sharing Toolkit. Data Pitch.

VOM BROCKE, J., A. SIMONS, B. NIEHAVES, K. REIMER, R. PLATTFAUT und A. CLEVEN, 2009. Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. In: Proceedings of the 17th European Conference on Information Systems. Verona, Italy: AIS.

VOM BROCKE, J., A. SIMONS, K. RIEMER, B. NIEHAVES, R. PLATTFAUT und A. CLEVEN, 2015. Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research [online]. Communications of the Association for Information Systems, 37(1), 205-224. Communications of the Association for Information Systems. Verfügbar unter: doi:10.17705/1CAIS.03709

WEBSTER, J. und R.T. WATSON, 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review [online]. MIS Quarterly, 26(2), xiii-xxiii. MIS Quarterly. Verfügbar unter: http://dl.acm.org/citation.cfm?id=2017160.2017162

8. Figure Index

Imprint

Editor
Fraunhofer Institute for Software and Systems Engineering ISST
Emil-Figge-Str. 91
Germany - 44227 Dortmund

Authors
Fraunhofer Institute for Software and Systems Engineering ISST
Inan Gür M.Sc.

Typesetting and layout
Peter Michatz, Fraunhofer Institute for Software and Systems Engineering ISST