Data Mesh : A Data Management Toolbox?

Master Thesis by Jan Nitschke at the Technical University Berlin. The following is an adaption of the introduction and the summary of the thesis. If you are interested to read the entire work, have questions or simply want to discuss the topic feel free to reach out via nitschke.jan@gmail.com or via LinkedIn.

May 2022

Contribution

1. This work provides evidence that Data Mesh can used as a data management toolbox. This entails that that individual parts of the Data Mesh solution proposal, namely the four Data Mesh principles, can be used individually to a certain extend.
2. The principle of Domain Ownership can be found to take on a central role in the use cases analyzed and there is evidence that it represents the central building block to distribute data ownership. Depending on the organizational context other principles become relevant to support that ideal.
3. The gap between data producers and data consumers is the most important challenge that organizations analyzed seek to overcome when implementing Data Mesh.

Summary

The relevance of data is undisputed in many of today's organizations [7, p.1]. Extensive research has been conducted to demonstrate the massive potential that data holds [6, p.214]. That potential lies especially in an organization's ability to use data to better understand its operations and customers, as a raw material to build and ship products and to drive innovation by identifying prospective business cases [2, p.1166] [9, p.191-192]. This surge in relevance of data for organizations is rooted in two developments: the rise in and increased accessibility of distributed compute platforms since the 2000's as well as the ability to accumulate ever larger amounts of data and resulting artificial intelligence capabilities in the 2010's [7, p.1].
Thus, being able to collect, analyze and operationally use data can be a competitive advantage, catalyze growth and ultimately lead to increased earnings in a variety of scenarios [9, p.192]. As a consequence, organizations of all kinds have committed to being "data-driven", postulating to use data to transform the way they are operating: towards putting data at the core of their existence [7, p.1]. This affects the fundamental role data plays within organizations: from a mere side product to a highly valuable asset at the center of their operations. To embrace that potential, fundamental organizational change, investments in and adaption of new technologies and methodologies as well as the incarnation of new professional roles is required [2, p.1174] [7, p.4-5]. The discipline that combines all of the above mentioned efforts is called data management [15, p.3].

Whereas the attention on data in organizations is vast, harnessing the true value that lies in data is often difficult even if major investments in data management are made [16, p.246]. The reasons are manifold and range from organizational to cultural and further to technological challenges [16, p.248] [11].
To better understand reasons of failure, it can be helpful to take a step back and to observe the scene from a distance. At the center of data management practices is a data platform [12, p.2]. It represents the technological system that drives data management efforts within the organization. That system is operated by a data team [1, p.12]. Hence, a data management system is a sociotechnical system and it lies in the nature of sociotechnical systems that its quality is demarcated by the ability to orchestrate the interaction between its technological and the social components. [19, p.17]. These interactions become increasingly fragile, the more parties need to interact with and within the system and the quicker and more drastic the context in which the system exists changes.

In the light of such challenges, Data Mesh proposes a novel way to orchestrate data management components. At its core, Data Mesh formulates a decentralized sociotechnical paradigm that aims to enable organizations to "share, access, and manage analytical data in complex and large-scale environments — within or across organizations" [3, p.3].
Initially proposed in 2019 by Zhamak Dehghani, the Data Mesh paradigm has been further refined and described throughout the years 2021 and 2022 [5, 4]. It has since sparked many, sometimes controversial, discussions among data practitioners about the exact outlines and relevance of the concept of Data Mesh [18]. Whether agreeing or not, one cannot deny that it is gaining more and more interest [20]. A possible evidence for this fact is Google Trends where the interest for the search term "Data Mesh" has steadily increased since 2021 [8] .

Research Goal

Given the growing interest in the topic albeit the controversial understanding about its core assumptions and goals, a thorough and sober analysis of the subject is paramount. However, little scientific research has been done on the subject at the time of writing this work ([10], [13], [14]). From the sparse research that exists, none is questioning or even analyzing the grounds on which Data Mesh alleges to stand on and build upon. However, such fundamental research is of utmost relevance in order to build a common understanding of the phenomenon. Thus, this work aims to help closing the named research gap by proposing an initial contribution to that field.
This is done by interviewing data management experts and qualitatively analyzing use cases that have started a Data Mesh implementation. The overarching goal of this work is to investigate if the Data Mesh paradigm can serve as a toolbox for data management. If so, then ideas proposed by Data Mesh could be used beyond the initial Data Mesh solution design. To achieve this overarching goal, it must be analyzed if individual parts of the Data Mesh paradigm can be applied individually and under which organizational circumstances. It is essential to understand that Data Mesh, for the remainder of this work, is specifically interpreted as the multitude of parts that form it and less as the single and indivisible abstract buzzword it is often reduced to. Only that divisibility will

allow for an in-depth analysis of the fundamental challenges Data Mesh addresses and solutions it formulates.

Based on concrete data management challenges, Data Mesh proposes a paradigm that consists of four principles that build the foundation of its implementation proposal(s) [17, p.10] [14, p.265-266]. It further states that these are "collectively necessary and sufficient" [3, p.9]. In the pursuit of the goal whether Data Mesh can serve as a toolbox for data management, an analysis is conducted in two parts.
First, the problem space that the Data Mesh paradigm targets is investigated. To do so, challenges that organizations were facing before starting their change process towards Data Mesh are collected (RQ 1). Second, parts of Data Mesh solution proposal that directly relate to those challenges are analyzed in depth. Two research questions are targeted to that second part of the analysis. RQ 2.1 aims at understanding whether the Data Mesh principles as introduced by Dehghani are universally and equally relevant in the use cases analyzed. RQ 2.2 analyzes if these principles play the role that Data Mesh envisions for and attributes to them.

The research questions are formulated as follows:

RQ 1 What are the data management challenges that organizations face prior to a Data Mesh implementation?

RQ 2.1 Are the four principles of Data Mesh as introduced by Dehghani universally and equally relevant in a Data Mesh implementation?

RQ 2.2 Do these four principles take on the role that they were foreseen by the Data Mesh paradigm as introduced by Dehghani?

Methodology

To examine these research questions, a research methodology of three major steps is synthesized and applied as shown in figure 0.1. In step I, primary use case data is collected through preparing and conducting data management expert interviews. The gained insights form the foundation for answering the three outlined research questions. In step II, to answer RQ 1, interviews are systematically analyzed to extract and categorize challenges that motivate organizations to adapt their data management strategies following the Data Mesh paradigm. Step III builds upon challenges identified in step II and adds the concrete Data Mesh implementations from the organizations interviewed to the analysis. The overall aim of step III is to investigate if and how the Data Mesh paradigm can be used as a toolbox for data management. Step III is split into two parts. Step III.I maps identified challenge categories onto the Data Mesh principles. In addition, use cases are analyzed with regard to the Data Mesh principles that they implement. This enables to understand the relevance of individual principles and to answer RQ 2.1. In order to answer the final research question, RQ 2.2, step III.II analyzes how individual principles are implemented in the use cases at hand. This reveals if the postulated role of the Data Mesh principles can be verified. The goal is to identify the contexts under which individual principles can be successfully implemented.

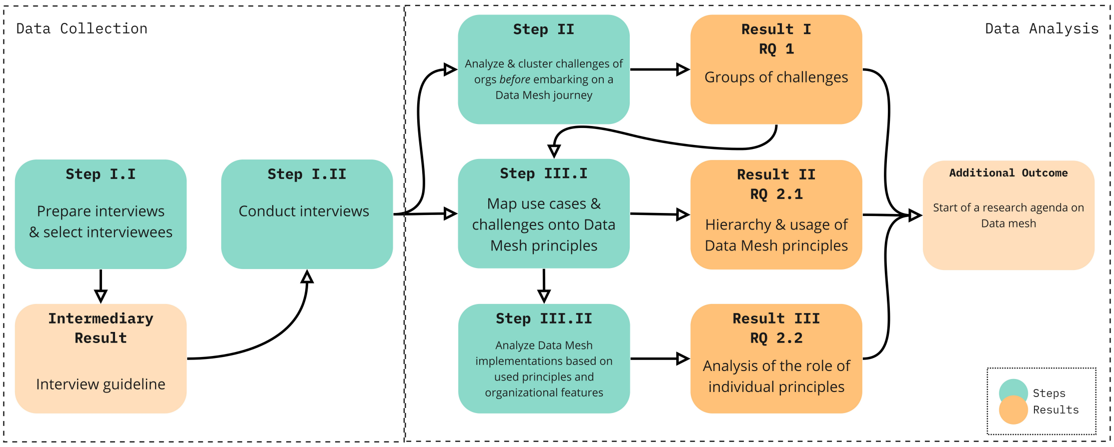
Data Collection
Data Analysis
Step I.I
Prepare interviews & select interviewees
Step I.II
Conduct interviews
Intermediary Result
Interview guideline
Step II
Analyze & cluster challenges of orgs *before* embarking on a Data Mesh journey
Result I
RQ 1
Groups of challenges
Step III.I
Map use cases & challenges onto Data Mesh principles
Result II
RQ 2.1
Hierarchy & usage of Data Mesh principles
Step III.II
Analyze Data Mesh implementations based on used principles and organizational features
Result III
RQ 2.2
Analysis of the role of individual principles
Additional Outcome
Start of a research agenda on Data mesh
Steps
Results

Figure 0.1: Steps of the research methodology

It is important to state that while this work is proposing an initial research approach to Data Mesh, it can only provide a coup d'oeil into what needs to be researched. The Data Mesh paradigm, and data management as a whole, has a far-reaching impact on organizations [2, p.1168-1169] [3, p.195-196]. A high variety of interpretations and implementations is thus to be expected. Although the contained use cases of this work are carefully selected to represent a high diversity of Data Mesh applications, the scope of this work only allows for a limited sample.

Results & Discussion

Understanding the challenges addressed as well as the solution proposed by the Data Mesh paradigm is paramount to be able to identify the context in which the Data Mesh paradigm can provide answers to data management challenges of organizations. Being able to delimit that context allows to analyze to what extend Data Mesh can provide tools that can be of general validity for data management. In the pursuit of this goal, several key findings result and interpretations are drawn.

With respect to research question RQ 1, results of this work identify the organizational challenges to be most important for organizations prior to a Data Mesh implementation. In addition, two other groups of challenges are identified: semantic and technical challenges. An intraorganizational gap between data producers and data consumers is the challenge that is raised the most from interviewees. Challenges identified, their classification as well as how they relate to the Data Mesh principles are shown in figure 0.2.

The challenges identified underline the sociotechnical implications of data management. Upon further analysis, it is indicated that the challenge of the gap between data producers and data consumers is a starting point for many other challenges identified. Further, challenges identified are not uniformly distributed across use cases which indicates that data management challenges are different depending on the organizational context. That distribution of challenges reveals that a solution proposal for data management challenges must be able to adapt to different organizational contexts.

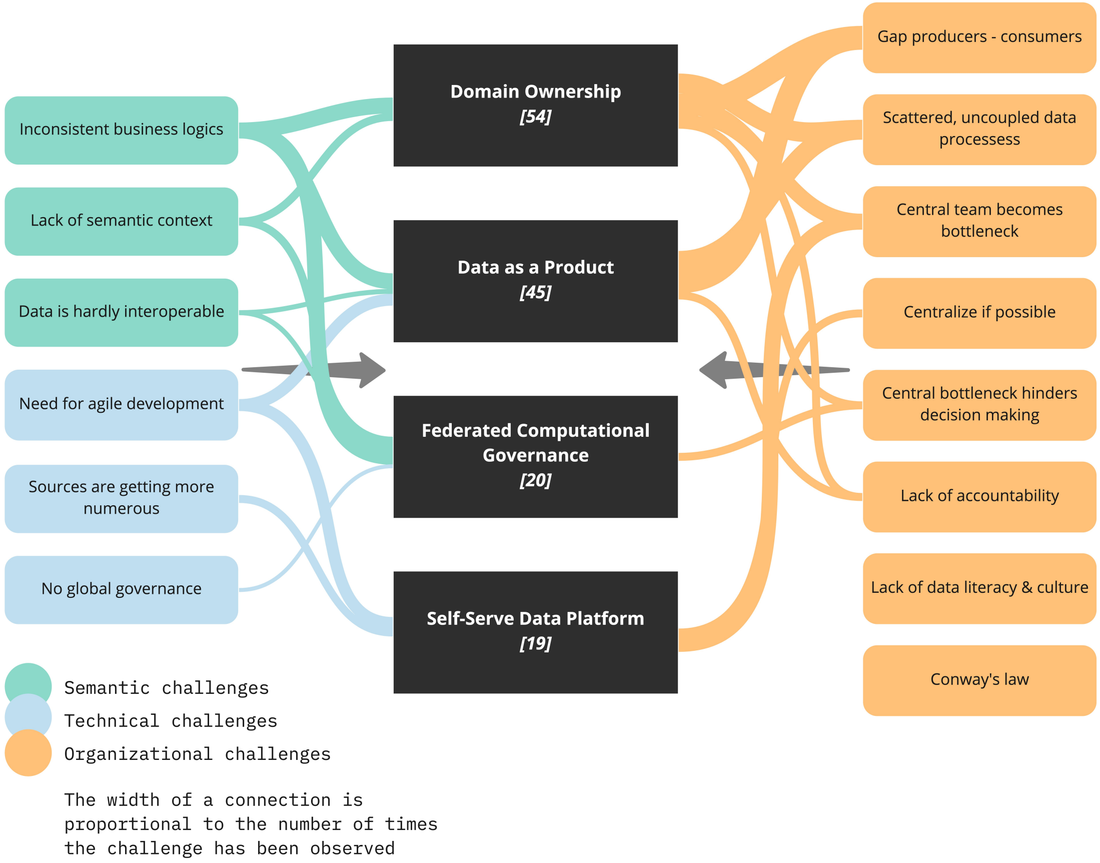
Inconsistent business logics
Lack of semantic context
Data is hardly interoperable
Need for agile development
Sources are getting more numerous
No global governance
Domain Ownership [54]
Data as a Product [45]
Federated Computational Governance [20]
Self-Serve Data Platform [19]
Gap producers - consumers
Scattered, uncoupled data processess
Central team becomes bottleneck
Centralize if possible
Central bottleneck hinders decision making
Lack of accountability
Lack of data literacy & culture
Conway's law
Semantic challenges
Technical challenges
Organizational challenges
The width of a connection is proportional to the number of times the challenge has been observed

Figure 0.2: Weighted mapping of challenge groups and Data Mesh principles

With regard to the second set of research questions and looking at the examined use cases it can be observed that the Data Mesh use cases implement different Data Mesh principles. The implementations are show in figure 0.3. However, the principle of Domain Ownership is implemented by all use cases. The analyses of the Data Mesh principles show that the different dimensions of scale of company size, number of data sources as well as the proportion of data and software personal can be used to explain the implementation of the principle of Data as a Product as well as Self-Serve Data Platform. However, these used scale dimensions were not able to describe the implementation of the principle of Federated Computational Governance.

From these results, it is concluded that the principle of Domain Ownership must be considered essential to contribute to the overall goal of the Data Mesh paradigm: distributing data ownership. The other three principles are implemented depending on the organizational context. Hence, considering research question RQ 2.1, it is deduced that not all principles are equally and universally relevant. However, the results from the use case analysis indicate that the Data Mesh paradigm can provide the required flexibility given the variety of challenges identified and implementations observed.
With respect to research question RQ 2.2, the results of this work suggest that the foreseen interplay and role of Data Mesh principles can be related to different types of organiza-

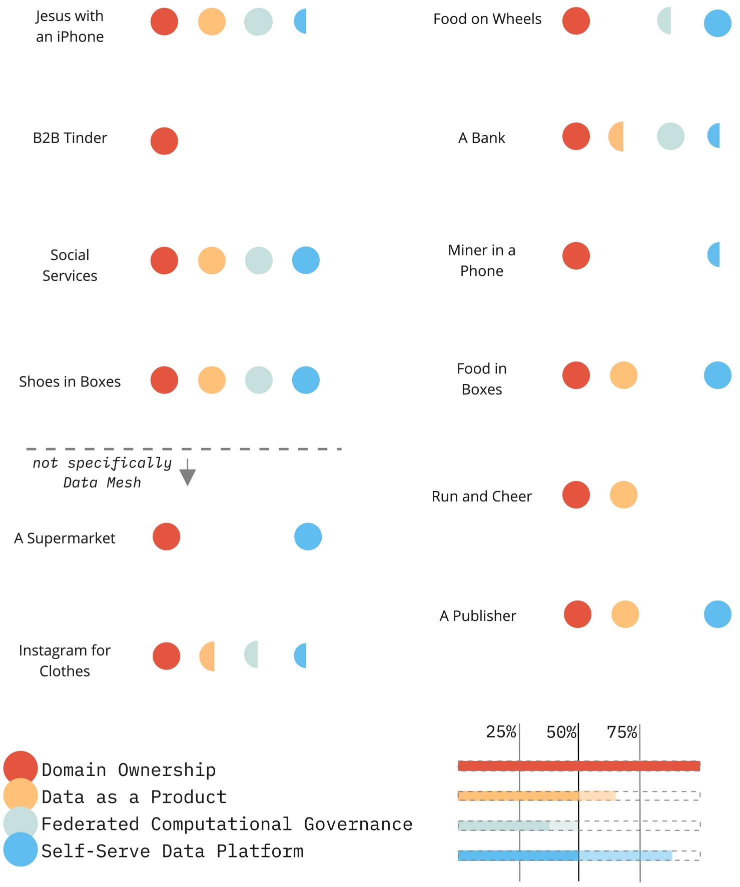

Figure 0.3: Attribution of Data Mesh principles to use cases

tional scale expressed as scale dimensions. The principle of Data as a Product is found to be relevant in scenarios with high data complexity, Self-Serve Data Platform applies in contexts where many data sources need to be integrated. For the principle of Federated Computational Governance, the foreseen role could not be validated. This might be a result of missing tools or wrong assumptions made by Data Mesh.

To conclude, this work provides evidence that the Data Mesh paradigm can be used as a toolbox to solve data management challenges that are mainly rooted in the sociotechnical nature of the discipline. Data Mesh is built around the idea of distributing data ownership across domain teams in order to overcome the fundamental challenge of bridging the gap between data producers and data consumers. At the center of this solution proposal lies the principle of Domain Ownership. Depending on the organizational context at hand, the principles of Data as a Product, Self-Serve Data Platform as well as Federated Computational Governance are used to further support that ideal.

Bibliography

[1] Jesse Anderson. "Data Teams". In: *Data Teams: A Unified Management Model for Successful Data-Focused Teams*. Ed. by Jesse Anderson. Berkeley, CA: Apress, 2020, pp. 3–17. ISBN: 978-1-4842-6228-3. DOI: 10.1007/978-1-4842-6228-3_1. URL: https://doi.org/10.1007/978-1-4842-6228-3_1 (visited on 12/15/2021) (cit. on p. 2).

[2] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4 (2012). Publisher: Management Information Systems Research Center, University of Minnesota, pp. 1165–1188. ISSN: 0276-7783. DOI: 10.2307/41703503. URL: https://www.jstor.org/stable/41703503 (visited on 03/24/2022) (cit. on pp. 1, 4).

[3] Zhamak Dehghani. *Data Mesh: Delivering Data-Driven Value at Scale*. S.l.: O'Reilly UK Ltd., May 17, 2022. 270 pp. ISBN: 978-1-4920-9239-1 (cit. on pp. 2 sqq.).

[4] Zhamak Dehghani. "How to Build a Foundation for Data Mesh: A Principled Approach with Zhamak Dehghani". Oct. 4, 2021. URL: https://www.youtube.com/watch?v=p1rCF-WWqhU (visited on 03/18/2022) (cit. on p. 2).

[5] Zhamak Dehghani. *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*. martinfowler.com. May 20, 2019. URL: https://martinfowler.com/articles/data-monolith-to-mesh.html (visited on 01/09/2022) (cit. on p. 2).

[6] Nada Elgendy and Ahmed Elragal. "Big Data Analytics: A Literature Review Paper". In: *Advances in Data Mining. Applications and Theoretical Aspects*. Ed. by Petra Perner. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 214–227. ISBN: 978-3-319-08976-8. DOI: 10.1007/978-3-319-08976-8_16 (cit. on p. 1).

[7] George Fletcher, Paul Groth, and Juan Sequeda. "Knowledge Scientists: Unlocking the data-driven organization". In: *arXiv:2004.07917 [cs]* (Apr. 16, 2020). arXiv: 2004.07917. URL: http://arxiv.org/abs/2004.07917 (visited on 03/24/2022) (cit. on p. 1).

[8] *Google Trends Data Mesh*. Google Trends. Mar. 24, 2022. URL: https://trends.google.de/trends/explore?date=today%205-y&q=%22data%20mesh%22 (visited on 03/24/2022) (cit. on p. 2).

[9] Wendy Arianne Günther et al. "Debating big data: A literature review on realizing value from big data". In: *The Journal of Strategic Information Systems* 26.3 (Sept. 1, 2017), pp. 191–209. ISSN: 0963-8687. DOI: 10.1016/j.jsis.2017.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0963868717302615 (visited on 03/24/2022) (cit. on p. 1).

[10] Divya Joshi, Sheetal Pratik, and Madhu Podila. “Data Governance in Data Mesh Infrastructures: the Saxo Bank Case Study”. In: (Dec. 4, 2021). Number: 7150 Publisher: EasyChair. ISSN: 2516-2314. URL: https://easychair.org/publications/preprint/qZ3m (visited on 03/18/2022) (cit. on p. 2).

[11] Mayur P. Joshi et al. “Why So Many Data Science Projects Fail to Deliver”. In: *MIT Sloan Management Review* (Mar. 2, 2021). URL: https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/ (visited on 03/24/2022) (cit. on p. 2).

[12] Alexandros Labrinidis and H. V. Jagadish. “Challenges and opportunities with big data”. In: *Proceedings of the VLDB Endowment* 5.12 (Aug. 1, 2012), pp. 2032–2033. ISSN: 2150-8097. DOI: 10.14778/2367502.2367572. URL: https://doi.org/10.14778/2367502.2367572 (visited on 12/02/2021) (cit. on p. 2).

[13] Inês Machado, Carlos Costa, and Maribel Yasmina Santos. “Data-Driven Information Systems: The Data Mesh Paradigm Shift”. In: *International Conference on Information Systems Development (ISD)* (Aug. 10, 2021). URL: https://aisel.aisnet.org/isd2014/proceedings2021/currenttopics/9 (cit. on p. 2).

[14] Inês Araújo Machado, Carlos Costa, and Maribel Yasmina Santos. “Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures”. In: *Procedia Computer Science*. International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021 196 (Jan. 1, 2022), pp. 263–271. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.12.013. URL: https://www.sciencedirect.com/science/article/pii/S1877050921022365 (visited on 03/18/2022) (cit. on pp. 2 sq.).

[15] Rupa Mahanti. “Introduction to Data, Data Governance, and Data Management”. In: *Data Governance and Data Management: Contextualizing Data Governance Drivers, Technologies, and Tools*. Ed. by Rupa Mahanti. Singapore: Springer, 2021, pp. 1–3. ISBN: 9789811635830. DOI: 10.1007/978-981-16-3583-0_1. URL: https://doi.org/10.1007/978-981-16-3583-0_1 (visited on 04/05/2022) (cit. on p. 1).

[16] Gianna Reggio and Egidio Astesiano. “Big-Data/Analytics Projects Failure: A Literature Review”. In: *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). Aug. 2020, pp. 246–255. DOI: 10.1109/SEAA51224.2020.00050 (cit. on p. 2).

[17] Max Schultze and Arif Wider. *Report: Data Mesh in Practice - How to Set Up a Data-Driven Organization*. O’Reilly Media, 2022 (cit. on p. 3).

[18] Benn Stancil. *The Modern Data Experience*. benn.substack. Aug. 20, 2021. URL: https://benn.substack.com/p/the-modern-data-experience (visited on 03/13/2022) (cit. on p. 2).

[19] Ali Sunyaev. “Introduction to Internet Computing”. In: *Internet Computing: Principles of Distributed Systems and Emerging Internet-Based Technologies*. Ed. by Ali Sunyaev. Cham: Springer International Publishing, 2020, pp. 1–24. ISBN: 978-3-030-34957-8. DOI: 10.1007/978-3-030-34957-8_1. URL: https://doi.org/10.1007/978-3-030-34957-8_1 (visited on 04/04/2022) (cit. on p. 2).

[20] Matt Turck. *Red Hot: The 2021 Machine Learning, AI and Data (MAD) Landscape*. Matt Turck. Section: AI. Sept. 28, 2021. URL: https://mattturck.com/data2021/ (visited on 12/15/2021) (cit. on p. 2).