Identifying Alternatives and Deciding Factors for a Data Mesh Architecture

Clara Voß[1]

Abstract: The data mesh was introduced in 2019 as a new type of data architecture. It promises a more democratic and scalable way of data production and consumption, while also solving current data engineering problems. These problems consist of siloed and hyper-specialized data engineering knowledge, a growing number of dependencies within data pipelines, and the rigidness of centralized monoliths. This paper uses expert interviews to identify the most significant current alternatives to the data mesh and abstract factors, with which companies can evaluate whether a data mesh can advance their move towards a data-driven, democratized future. The results show that the company's motivation, culture, structure, as well as IT history and IT structure should be evaluated before implementing a data mesh. This paper is based on a bachelor thesis.

Keywords: data mesh, data architectures, data integration, expert interviews

1 Introduction

In a world where the importance of data and its analysis continuously grows, choosing a fitting data architecture is crucial to the success of a company's data strategy. While many data architectures have been proposed since the start of enterprise data solutions in the 1970s, a lot of companies still struggle to find suitable solutions to their specific problems and to overcome challenges like the need for a growing number of increasingly specialized data engineers and a growing burden of dependencies.

In an attempt to solve these common architecture failure modes, Zhamak Dehghani presented the data mesh [De19]. This is a data architecture that is not focused on a centralized monolithic data architecture but instead provides a more distributed, democratized approach. Supported by a self-serve tooling and data platform, the business is separated into independent, empowered units, which create autonomous data products that are shared in a company-wide data marketplace. While the data mesh has started a big conversation about distribution among the data engineering community, many companies are still hesitant to implement it, because little work has been done to identify advantages and disadvantages. The findings of this paper are based on expert advice and practitioners' experiences and offer use cases, alternatives, and tangible deciding factors to allow for a greater understanding of the data mesh concept.

[1] Hochschule für angewandte Wissenschaften München, Fakultät für Informatik und Mathematik, Lothstr. 34, 80335 München, clara.voss97@gmail.com

2 Theoretical Foundation

2.1 Data Mesh

To define the data mesh, its idea and core concepts as presented in [De19, De20] will be summarized in the following paragraph: Central to the definition of the data mesh are its four main concepts: Domain Ownership, Data as a Product, Self-Serve Infrastructure Platform, and Federated Governance. According to Domain Ownership, a company should be organized into inter-disciplinary teams, that are formed around business functionality, have domain knowledge and are responsible for the full life cycle of its functionality. This pushes responsibility from the central IT or data engineering department to these independent teams. Each team can build data products, the smallest independent measure of data assets. These include the data, the describing metadata, the code leading to its finished product, the documentation of the infrastructure it is hosted on, documentation, and data lineage. Full responsibility for the data product, including the fulfillment of the DATSIS characteristics (discoverable, addressable, trustworthy, self-describing, interoperable, secure), stays with the team that created it. Data products are hosted on a company-wide data marketplace from which other teams can use them for analysis, further development, and creation of new data products (Data as a Product). To enable all teams to build data products without needing specialized technological knowledge, a self-serve infrastructure platform is built. This platform abstracts the complexity of choosing and provisioning technologies and allows for teams to independently choose the technologies that best suit their challenges. Lastly, federated governance is the principle that sets global standards, monitors the mesh, and is responsible for defining the balance between centralized guidelines and distributed responsibility / domain team autonomy.

Aspects of existing trends can be found in the main concepts leading to the data mesh. This shows that the data mesh can be seen as a natural progression, consolidation and application of these trends: Domain Driven Design (organizing teams around functionality), agile development (self-organized and cross-functional teams), cloud computing (abstraction of physical infrastructure and easier infrastructure administration), DevOps (collaboration between technical and business experts and removing traditional separation e.g. by organizational stages), microservices (splitting a monolithic system into small, independent, loosely coupled parts), data democracy (empowering a large, not necessarily technically trained group of people to use and produce data).

2.2 Alternatives

The most important alternative data architectures are the data warehouse, data lake, data lakehouse, data hub, and data fabric, which can be sorted into two main categories:

The data warehouse and data lake specify the way data is stored, the data scheme used and are closely connected to technologies like TeraData or Hadoop. Specialized to work with structured data, the data warehouse stores its data in relational databases and strictly enforces consistency with its scheme-on-write [Wh21]. The growing amount of semi- or unstructured data challenged the data warehouses' focus on structured data and its rigid consistency enforcement. This led to the development of the data lake, which stores data in its raw format, without transformations or enforced consistency [KW18].

In comparison, the data lakehouse, data hub, data mesh and data fabric do not predominately make suggestions about the way data is stored but focus on connecting distinct data sources to offer users a single point of access. The data hub is based on a hub-and-spoke layout and is optimized for sharing data between multiple companies, while conforming to data privacy regulations [Bh15]. Connecting the database management system functionalities and benefits of the data warehouse and the data lake has been the focus of the data lakehouse, which was proposed by Databricks [Ar21].

The data fabric is the closest alternative to the data mesh. Both are not technologies, but emerging data architecture concepts, which identify the same problem of managing data at scale. While the data mesh focusses on distributing the company into independent data domain teams with a federated governance, the data fabric still enforces central data management and control. Additionally, the concept of the data fabric relies on the evolution of artificial intelligence (AI) and knowledge graphs to solve data engineering problems and automatically create inference between data using AI [GB19].

3 Method: Expert Interviews

This paper uses semi-structured, guideline-based expert interviews to extract their experiences with alternatives and transfer the experts' knowledge onto the new topic of the data mesh. The introductory questions centered around the experts' own definition of the data mesh and its alternatives as well as data mesh implementations in past projects. The focus of the interviews was using the ISO 25010 criteria to consider different aspects of the data mesh concept and how they compare to the alternatives. Lastly, the experts were invited to contemplate the future of data architectures and the data mesh concept's place in it.

Planning and execution of the interviews followed the steps described by Mieg and Näf's [MN05], while the interviews' transcripts were analyzed using Mayring's method of qualitative content analysis [Ma15]. The resulting category system built the basis of the analysis and discussion.

The interviewed experts can be separated into three categories based on their background (managerial, technical, and academic), to allow for a broad discussion of the data mesh concept and include different experiences in implementing it. The interviews were conducted in English and German via video calls from 01.11.2021 to 30.11.2021.

Name	Viewpoint	Role	Experience with data integration	Interview length
Roy Kronester	Management	Director Technical Advisory	30 years	1:02:23 min
Dael Williamson	Management	European CTO for Data & AI	25 years	1:05:57 min
-- (anonymous)	Technical, academic	Professor & Hub Lead	11 years	1:01:21 min
-- (anonymous)	Technical	Data Engineering Senior Analyst	4 years	0:29:40 min
Inês Machado	Academic	Author of the first paper on the data mesh	5 years	0:47:48 min
Yvonne Niedling	Technical	Data Engineering Manager	10 years	0:36:30 min

Tab. 1: Overview of the experts

4 Findings

The category system, that resulted from the qualitative content analysis, consolidates the content of the transcribed interviews, and is summarized in the following paragraphs.

4.1 Data architectures in comparison

The data hub, data lakehouse and data fabric are more contemporary data architectures and solve the same use case of connecting distinct data sources. None the less, especially the older data warehouse is mentioned as a suitable alternative to the data mesh. The consolidating nature of the data warehouse allows for representation of all business units within a company and the connections between them. The development of a “data swamp”, a non-functioning data lake that lacks governance and standards, can also be mirrored in a data mesh leading to a “data mess”. Modern architectures like the data lakehouse and data fabric have the added challenge, that much of their vision is based on future advancement of technologies like AI. The level of AI necessary to implement a data fabric has not been reached yet, limiting the evaluation of these data architectures for practical use cases and the implementation itself.

4.2 Factors in the decision for a data mesh

The main factors a company should consider before implementing a data mesh can be grouped into “motivation”, “IT history”, “company culture”, “IT structure”, and “company structure”.

- *Motivation:* It is crucial to evaluate who proposes the implementation of a data mesh and why. Because of the new distribution of responsibility to the business units, they must be fully aware of the effects a data mesh has on their workload. The work of the data engineering and IT departments becomes less stressful, and they can refocus on their intended assignments, creating a big incentive to implement a data mesh. That is why the motivation must be closely analyzed, if the idea of implementing a data mesh is brought forth by the IT department and not fully understood by the business units.
- *IT history:* A data mesh can be a very useful architecture, when the company has conducted many IT and data experiments leading to isolated systems, that otherwise cannot be connected to the main system or whose data cannot be accessed.
- *Company culture:* To support the successful implementation of a data mesh, it is important to make data a priority in the company and allow for flexible problem solving. Additionally, management must be willing to give up responsibilities to the self-governing independent teams and support them in their local decision-making. If this empowerment of the teams cannot be guaranteed by management, then a data mesh should not be implemented.
- *IT structure:* According to the experts’ opinions, a data mesh is especially useful in connecting isolated or legacy systems, clouds or data silos, which have a great amount of friction in data workflows and pipelines.
- *Company structure:* The data architecture should mimic the companies’ own structure: Geographically distributed companies or companies that are comprised of loosely coupled, independent business units can reap a greater benefit, because the domain boundaries can be easily derived. Consolidating these independent business units in a centralized solution would lead to more organizational overhead and problems. Additionally, the company size and amount of data should be considered, as small and mid-sized companies with a small amount of data might be overwhelmed by the organizational changes necessary to implement a data mesh.

5 Conclusion

5.1 Theoretical and Practical Implications

The compiled factors can be used to help companies in deciding whether a data mesh can solve the data engineering challenges and further the usage of data or whether the overhead of implementing a data mesh is too much in comparison to its benefits. If the data mesh is not feasible, the company might consider the mentioned alternatives to find a suitable data architecture solution.

5.2 Limitations and Outlook

Because the conversation about the data mesh has only started in 2019, not many implementation tools have been developed and distributed widely and the topic has not been fully academically discovered, many questions remain unanswered. Some companies report anecdotal successes and benefits due to the implementation of a data mesh, but an academic evaluation is still missing. To conclude, the conversation of moving away from monolithic data architectures might highlight the data usage within companies and bring on a new era of data integration.

6 Acknowledgments

I would like to thank the supervisor of my bachelor thesis, Prof. Dr. Johannes Ebke, for his continuous feedback and the encouragement to take part in this program. Additionally, I want to give thanks to the interview partners, who shared their experiences and invested time. Lastly, I would like to thank Avanade, the company that initially proposed the topic of the data mesh and supported my work along the way.

7 References

[Ar21] Armbrust, G. et al: Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In: 11th Annual Conference on Innovative Data Systems Research (CIDR ’21), 2021.

[Bh15] Bhardwaj, K. et al: Collaborative Data Analysis with DataHub. In (VLDB Endowment): Proceedings of the VLDB Endowment, Vol. 8, No. 12, S. 1916-1919, 2015.

[De19] Dehghani, Z.: How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh, https://martinfowler.com/articles/data-monolith-to-mesh.html, Stand: 09.05.2022.

[De20] Dehghani. Z.: Data Mesh Principles and Logical Architecture, https://martinfowler.com/articles/data-mesh-principles.html, Stand: 09.05.2022.

[GB19] Ghiran, A., Buchmann, R. A.: A Model-Driven Enterprise Data Fabric: A Proposal Based on Conceptual Modelling and Knowledge Graphs. In (Springer, Cham): Knowledge Science, Engineering and Management. KSEM 2019. Lecture Notes In Computer Science, vol. 11775, S. 572-583, 2019.

[KW18] Khine, P. P., Wang, Z. S.: Data Lake: A new ideology in big data era. In: ITM Web Conferences, vol. 17, 030525, 2018.

[Ma15] Mayring, P.: Qualitative Inhaltsanalyse. Grundlagen und Techniken, Beltz, Weinheim, 2015.

[MN05] Mieg, H. A, Näf, M.: Experteninterviews, 2. Aufl., Institut für Mensch-Umwelt-Systeme (HES), ETH Zürich, 2005.

[Wh21] Oracle (Hg.): What is a Data Warehouse?, https://www.oracle.com/database/what-is-a-data-warehouse/, Stand: 09.05.2022