IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124121

From Data Mess to Data Mesh: Solution for Futuristic Self-Serve Platforms

Satyajit Panigrahy[1], Bibhu Dash[2], Ramya Thatikonda[3]

Doctoral Student, National Institute of Technology, Rourkela, India[1]

Research Scholar and Data Architect, University of the Cumberlands, KY, USA[2]

Research Scholar, University of the Cumberlands, KY, USA[3]

Abstract: As technology advances, data volume and velocity increase in all domains. Data is everywhere, which creates a data mess situation in many organizations. Sometimes, it makes challenging to think how this data can be utilized to work for the betterment of the organization without storing it in many places in different formats or duplicating it again and again. There comes the 'Data Mesh' design, which is a relatively new concept focusing on data storage, data governance, and data management in an efficient way to encourage self-service data handling. Many organizations are now considering this new "Data Mesh" concept to address the major issues and barriers in data management and usage. They realize that by focusing on domain-specific data products enabled by common support functions, they can ensure flexible access to data with the significant benefit of reduced time-to-market and fast-to-product development. The data Mesh incorporates contemporary architectural concepts and focuses on data management rather than connectivity and orchestration.

Keywords: data mess, data mesh, SSP, SLAO, sustainability

I. INTRODUCTION

Every organization aspires to improve the perspective of business and life using data. But it is always a question of how to manage data for better scalability and governance. Data mess to Data mesh, which is founded on the four ideas of self-service data infrastructure as a platform, domain-oriented decentralized data ownership and architecture, data as a product, and federated computational governance, addresses these dimensions. Each tenet gives a fresh logical viewpoint on the organizational structure and technical architecture.

A data mesh is a decentralized data architecture that organizes data by business domain - for example, marketing, sales, customer support, and so on - giving producers of a given dataset more ownership. The producers' knowledge of the domain data enables them to establish data governance policies centered on documentation, quality, and access.

This, in turn, allows for self-service use throughout an organization. While this federated approach avoids many operational inefficiencies associated with centralized, monolithic systems, it does not exclude the usage of traditional storage solutions such as data lakes or data warehouses. It simply indicates that their application has transitioned away from a single, centralized data platform and towards several decentralized data repositories [1]. Zhamak Dehghani coined the word mesh in 2019, and it is built on four essential principles that combine the well-known below concepts [2].

(a) The domain ownership principle requires domain teams to take ownership of their data. This principle states that analytical data should be organized around domains, similar to how team boundaries coincide with the system's bounded context.

(b) The notion of data as a product applies a product-thinking mentality to analytical data. This principle implies that there are data users outside of the domain. The domain team is in charge of meeting the needs of other domains by supplying high-quality data. Domain data should be treated similarly to any other public API [2].

In order to advance the analytics paradigm, this study clarifies and summarises the architectural features of data mesh. In its subsequent sections, this paper talks about what and how a data mesh works, its benefits, its architecture, and its limitations, along with conclusions and future road maps.

DOI: 10.17148/IJARCCE.2023.124121

II. DATA CLASSIFICATION AND MODERN BUSINESS SIGNIFICANCE

Earlier, we always talk about operational data and transactional data. However, as technology advanced, that concept became increasingly diluted, and we began to refer to it as operational data (both operational and transactional) and analytical data [3]. As it lives in databases behind business capabilities provided by microservices, operational data has a transactional nature, preserves the current state, and satisfies the requirements of the business-running apps. Analytical data is a temporal and aggregated view of an organization's facts through time, typically modeled to provide retrospective or forward-looking insights; it either feeds analytical reports or is used to train machine learning models [4].

These above-discussed data kinds are separate but related. The disparity has led to an unstable architecture. Many people who attempt to connect these two planes—transferring data from the operational data plane to the analytical plane and back to the operational data plane—are aware of the frequently failing ETL (Extract, Transform, Load) processes and the steadily rising complexity of a web of data pipelines [4-5]. The analytical data planes are bound with two main architectural technology stacks: data lake and data warehouse. The modern data lake design supports data science and Machine Learning, whereas the data warehouse design supports business reporting and analytical dashboarding [5]. Data mesh is primarily an add-up to discuss existing contemporary analytical data plane architecture and interaction. A domain-driven data mesh design is shown below in Figure 1.

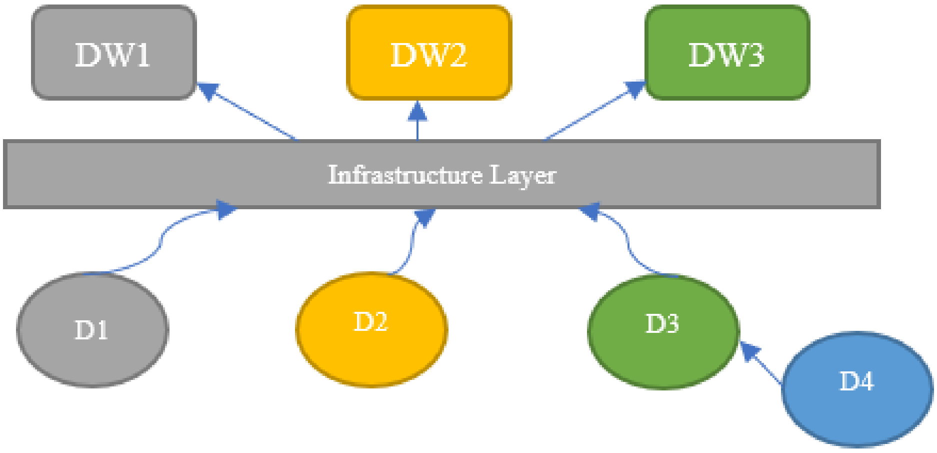

Figure 1. Data mesh design with ‘Domain-Driven’ Architecture (*D for Domain)

III. HOW DOES A DATA MESH WORK?

A data mesh demands a change in how firms view their data on a cultural level. Data no longer serves as a process by-product; instead, it serves as a product, with data creators working on behalf of data product owners [6]. Traditionally, a centralized infrastructure team would manage data ownership across domains; however, the product thinking focus of a data mesh model distributes this ownership to the producers, who are the subject matter experts (SMEs). Their expertise in the domain's core data consumers and how they exploit operational and analytical data enables them to create APIs with their best interests in mind.

While this domain-driven approach makes data producers accountable for articulating semantic definitions, cataloging metadata, and establishing policies for rights and usage, a centralized data governance team is still in place to enforce these standards and procedures. Furthermore, while domain teams are accountable for their ETL or ELT data pipelines in a data mesh design, a centralized data engineering team is still required. Their responsibilities, however, shift to finding the appropriate data infrastructure solutions for the data products being stored [1]. A contemporary data warehouse design is shown in Figure 2 below, together with specifications for operational and analytical data planes of domain data.

DOI: 10.17148/IJARCCE.2023.124121

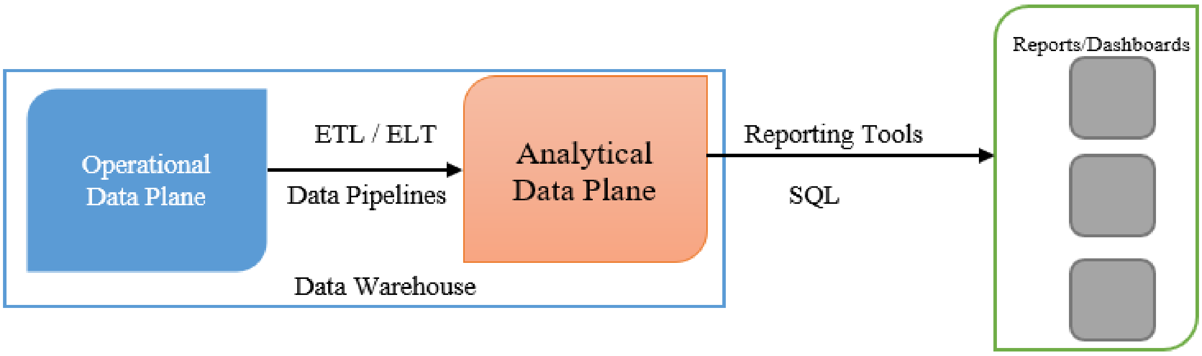

Figure 2. A divide of modern data-warehouse

IV. DATA MESH BENEFITS

Building a foundation for value extraction from analytical data and historical facts at scale is what data mesh aims to achieve. By scale, we mean the diversity of transformation and processing that use cases call for, the speed at which they react to change, and the constant change of the data landscape. Table 1 demonstrates some of the key differences between traditional data warehouse and data mesh architectural design. There are direct benefits for an organization adopting this data mesh design concept, and those are highlighted below.

A. Agility and scalability

Because of the decentralized data infrastructure as a service, there is a considerable improvement in time-to-market, scalability, and overall business domain agility, as well as a reduction in the IT backlog. This is also due to the agile project teams' ability to function autonomously, focusing on key data products [2, 7].

B. Strong central governance to control end-to-end compliance.

Traditional architectural setups with centralized data lakes fail to reconcile the semantics and volume of ingested data due to the rapidly rising number of data sources and their diverse data formats. Decentralizing data operations to a domain and following global data governance principles enhances data quality and accessibility [8]. There will be no more data dumps into data lakes in bulk. The data will be assessed basis of domain needs before categorization and storing in data lakes. It gives it an advantage in managing and setting rules for ETL or ELT pipelines and business intelligence details.

C. Transparency in cross-functional domain teams

In contrast to traditional data architecture approaches that encourage the isolation of skill teams, which frequently have large backlogs, Data Mesh presents a solution in which domain experts and owners are in charge. This is carried out through enhanced domain expertise, tighter collaboration between business and IT departments, and agile virtual teams [8-9]. It also gives businesses or SMEs to have data usage without having many dependencies on IT teams.

D. Faster data delivery

Storing data in specified traditional formats and creating data infrastructure is frequently a hindrance to data management. Data Mesh provides a self-service, governable, and centralized infrastructure with the underlying complexity hidden away for speedier data delivery with independent data storage formatting [3, 9].

E. High-quality self-serve data

Data mesh provides autonomous teamwork for each data domain by decentralizing data processes. As a result, operational costs and adaptability are enhanced. Data mesh offers high-quality data outputs because of its self-service data design (bring only what you need) and autonomous governance principles (cross-functional teams master it) [9-11].

DOI: 10.17148/IJARCCE.2023.124121

Table 1. Traditional data warehouse (DW) vs. Data Mesh Architecture

Traditional DW Architecture	Data Mesh Architecture
Techno-Functional Approach	Socio-Technical Approach
Centralized Architectural Design	Decentralized Architectural Design
Low scalability and complexity increase when more domains are added.	High scalability and domain-specific design
Data Warehouse(DW) designs are not interconnected	Interconnected design
Teams need IT team support to read and filter their data needs.	This design supports self-service platforms to filter and manage data by functional teams.
Data quality and governance is a choice per Business Intelligence (BI) needs.	Supports strong data quality and governance through cross-domain exchange.

V. PRINCIPLES OF DATA MESH ARCHITECTURE

A. Domain-driven data ownership and architecture

One of the core notions of a decentralized data model is this. A domain-driven strategy is replacing a centralized data hub. A domain is a business unit that generates data (for example, Marketing, Sales, Finance, and so on). They are the functional components of a data mesh.

In this model, data products are created by teams who understand the data the best and adhere to an organization-wide set of data governance principles [10-11]. People who were previously only stakeholders in specific data sources in a centralized system are now solely responsible for developing, disseminating, and managing that data (see Figure 3).

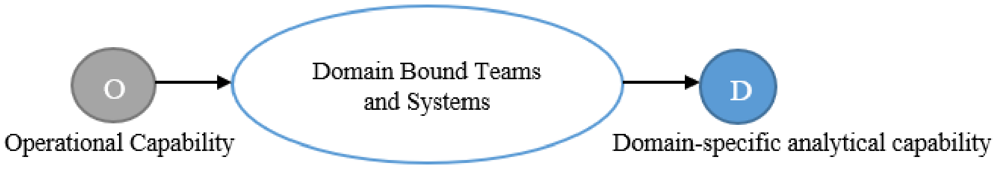
O
Operational Capability
Domain Bound Teams and Systems
D
Domain-specific analytical capability

Figure 3. A domain-driven architecture design

B. Self-Serve platform (SSP)

The goal of the self-service data platform is to provide a centralized way of consuming (discovering, exploring, and ultimately using) data from many domains. Microservices are to software development what data mesh is to data engineering. Consider our data products to be microservices and our self-service data platform to be an app that consolidates those microservices and teaches consumers how to use them. In the previous centralized paradigm, a company's "data team" was involved in everything, whereas in a data mesh approach, they assumed charge of the self-service platform for self-service analytics [8, 11]. Their primary purpose is to develop technologies that increase the efficiency of domain teams. It is important to note that the whole idea of the data mesh design is to decrease bottlenecks and shorten the time it takes to generate value; technology can be a bottleneck, and this is where platform teams can help.

C. Data as a first-class product

Treating data as a product has several instant advantages. We would not make changes to any product without extensive testing, quality assurance, and being confident that we are pleased for it to be consumed by the intended market [12]. The same might be said about our company's data. Customers can be confident that it will operate exactly as intended and

DOI: 10.17148/IJARCCE.2023.124121

documented, that it is easy to access, and that the data they get is of the greatest caliber. We, as the data product owners, are totally responsible for fixing any issues with the product, and we need to act quickly because the rest of the business depends on the results. In a design where the data product appears as the design quantum, data as a product aid in managing code, infrastructure, and polyglot data to act as the design's three main pillars (see Figure 4) [2, 12].

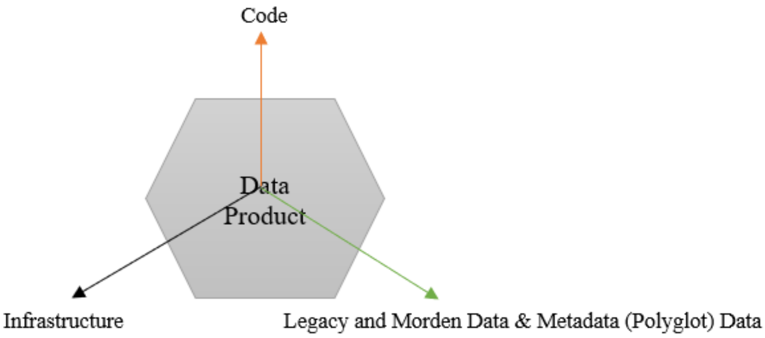

Figure 4. Architecture Quantum with data product components

D. Federated computational governance.

Traditional centralized governance (pre-data mesh governance structure) is fading out, and federated governance is getting more popular in data-driven supportive organizations. The goal of federated governance is to establish (quite clearly) data governance rules at the centralized level while giving domain teams the autonomy and responsibility to apply those standards most effectively for their scenario. Data governance at this level puts everyone at the team level on the same page and serves as the foundation for collaboration across distant teams. Data contracts are a useful technique to formalize all relevant information about a given product when data is viewed as a product [4, 12].

Since they are familiar with the intricate details of the domain processes that generate the data in the first place, domain data product owners are in the best position to decide how to monitor the data quality of their domain in this architecture. They must follow quality modeling and SLAOs (Service level Agreement and Objectives) definition based on a global standard, as stated by the global federated governance council and automated by the platform, notwithstanding their decentralized decision-making and autonomy [2, 11]. This design encourages automated processes to detect errors and recover through platforms' automated processing designs. The overall design of the data mesh architecture is shown in Figure 5.

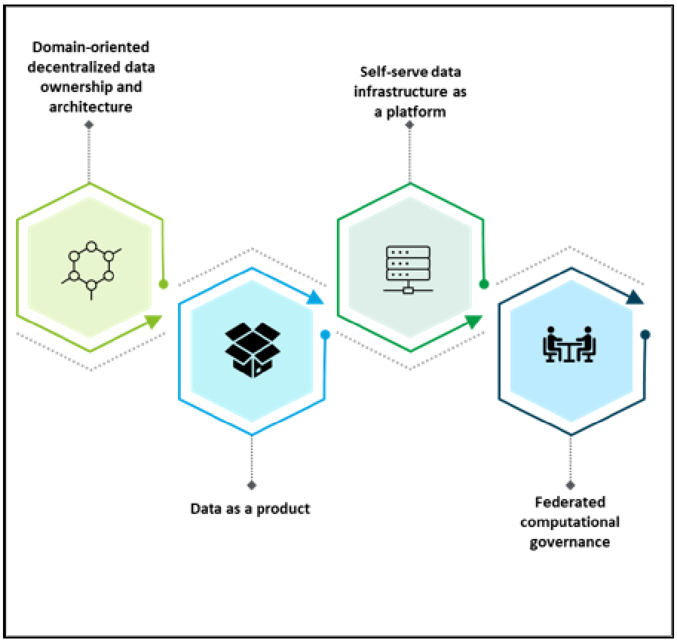
Domain-oriented decentralized data ownership and architecture
Self-serve data infrastructure as a platform
Data as a product
Federated computational governance

Figure 5. Pillars of Data Mesh Architecture [4]

DOI: 10.17148/IJARCCE.2023.124121

VI. LIMITATIONS OF DATA MESH

Though this architecture is getting a lot of attention today, irrespective of domain, this design has some potential limitations or disadvantages. These are discussed in detail below.

A. Team limitations

Setting it up and managing it might be challenging, particularly if a team has no prior knowledge of the architecture. It needs cross-functional SME support, so it will fail if team culture is not promoted to contribute independently [2].

B. Needs customization to fit into a particular organization.

There is no one-size-fits-all model. So, each organization needs to build it as per their use case, reporting, and analytical needs. Customization gives the design an independent organizational touch for better usability in the long run [2, 14].

C. Needs both strategic and tactical mindset.

These solutions are incredibly efficient, but to identify the strengths and weaknesses of your teams, you must think strategically and tactically. These solutions are a waste of time and money if any team is not performing at its highest level technically. To create long-lasting designs, functional and technological skills are required. It may create bottlenecks to both operating and maintaining such architectures.

D. Can create data silos.

Once implemented, these designs are very difficult to change frequently. It can cause difficulty to share as it creates domain-specific fragmented data. As the system and organization grow, it may create silos to change people’s use habits and may lead to duplication of data at each team level [4, 11].

VII. CONCLUSION

From our detailed analysis, we may now specify the design of the data mesh's components, such as the data product, the platform, and the requisite standardizations, using a consistent vocabulary and a coherent mental model. The ability of organizations to test the idea and learn from their failures will decide the viability of data mesh in light of shifting data conditions. Because they enable us to do things we never thought were conceivable, new technology developments are something we all appreciate. Nevertheless, it could be overwhelming, and staying up with the latest trends can be challenging. Will data mesh change the environment, and will it set a new standard? The evolving technical needs and data requirements will determine the long-term viability of this design, but we see a major influence for this architecture to change the data usage and standards in the near future. Currently, there are many software tools and techniques available that support designing customized enterprise data mesh data warehouses.

ACKNOWLEDGMENT

We would like to thank our advisors**, Dr. Azad Ali** and **Dr. S. Swayamsiddha**, for their advice and support in completing this paper on time and to high standards.

REFERENCES

[1]. IBM. (2022). What is a data mesh? IBM. Retrieved April 26, 2023, from https://www.ibm.com/in-en/topics/data-mesh

[2]. Dehghani, Z. (2022). Data Mesh. Marcombo.

[3]. Pechovska, P. (2021). Data Mesh Architecture. Retrieved April 24, 2023, from https://www.datamesh-architecture.com/

[4]. Strengholt, P. (2023). Data Management at scale. " O'Reilly Media, Inc.".

[5]. Mutatiina, J., & Blaauw, E. (2022, February 13). From Data Mess to a data mesh. Deloitte Netherlands. Retrieved April 26, 2023, from https://www2.deloitte.com/nl/nl/pages/strategy-analytics-and-ma/articles/from-data-mess-to-a-data-mesh.html

[6]. Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: concepts and principles of a paradigm shift in data architectures. Procedia Computer Science, 196, 263-271.

[7]. Bode, J., Kühl, N., Kreuzberger, D., & Hirschl, S. (2023). Data Mesh: Motivational Factors, Challenges, and Best Practices. arXiv preprint arXiv:2302.01713.

DOI: 10.17148/IJARCCE.2023.124121

[8]. Dash, B., & Ansari, M. F. (2022). Self-service analytics for data-driven decision making during COVID-19 pandemic: An organization's best defense. Academia Letters, 2.

[9]. Drobac, D. (2023, March 7). Data Mesh: The future of Data Modeling and Scalable Analytics. Medium. Retrieved April 26, 2023, from https://medium.com/@danilo.drobac/data-mesh-the-future-of-data-modeling-and-scalable-analytics-57670a460a47

[10]. Sharma, P., Chetti, P., Dash, B., & Ansari, M. (2023). Data Modeling Best Practices Key to Data Mining and Data Standardization. Available at SSRN 4337595.

[11]. Li, J., Cai, S., Wang, L., Li, M., Li, J., & Tu, H. (2022, December). A novel design for Data Processing Framework of Park-level Power System with Data Mesh concept. In 2022 IEEE International Conference on Energy Internet (ICEI) (pp. 153-158). IEEE.

[12]. Yathiraju, N., & Dash, B. BIG DATA AND METAVERSE REVOLUTIONIZING THE FUTURISTIC FINTECH INDUSTRY.

[13]. Butte, V. K., & Butte, S. (2022, October). Enterprise Data Strategy: A Decentralized Data Mesh Approach. In 2022 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 62-66). IEEE.

[14]. Driessen, S., Monsieur, G., & van den Heuvel, W. J. (2023, March). Data Product Metadata Management: An Industrial Perspective. In Service-Oriented Computing–ICSOC 2022 Workshops: ASOCA, AI-PA, FMCIoT, WESOACS 2022, Sevilla, Spain, November 29–December 2, 2022 Proceedings (pp. 237-248). Cham: Springer Nature Switzerland.

BIOGRAPHY

Satyajit Panigrahy is a doctoral student at the National Institute of Technology, Rourkela, India. His discipline is in Electrical and Computer engineering with a main focus on AI, Bigdata, High voltage engineering, external insulation, and Condition monitoring.

Dr. Bibhu Dash is a research scholar and Lead Data architect in Wisconsin, USA. He is a tech. guide, author, and research fellow with a special interest in AI, Bigdata, IoT, Cloud computing, Data modeling and governance, and Cybersecurity.

Dr. Ramya Thatikonda is a researcher and software professional. She received her Ph.D. in IT from the University of Cumberlands, KY. Her research interests are in AI, BigData, Cloud, DevOps, and Cybersecurity.