Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2024, 11(9):51-56

Research Article ISSN: 2394 - 658X

Implementing Data Mesh: Challenges and Strategies for Decentralized Data Governance

Puneet Matai

Data & AI Governance Lead,
Rio Tinto Commercial Pte Ltd, Singapore
puneet.matai@gmail.com

ABSTRACT

This whitepaper explains Data Mesh which is a decentralized approach to data architecture designed to overcome the limitations of traditional centralized systems. The focus is on the challenges and strategies for implementing decentralized data governance. It finds that while Data Mesh offers flexibility and improved data ownership, it requires careful planning, cultural shifts, and robust tooling. This paper is for data architects, engineers, and business leaders seeking to modernize their data infrastructure. The final recommendation emphasizes starting small, clearly defining ownership, investing in tooling, implementing a data-driven culture, and maintaining agility in execution.

Keywords: Data Mesh, Decentralized Data Governance, Domain-Oriented Decentralization, Data as a Product, Federated Computational Governance, Data Interoperability

---

INTRODUCTION

Background on Data Mesh

Data Mesh is a decentralized data architecture that emphasizes domain-oriented ownership. It treats data as a product and improves data accessibility. In 2018, Zhamak Dehghani introduced the concept of Data Mesh [1].

She defines Data Mesh as a "*decentralised socio-technical approach to managing and accessing data. This is designed to address the challenges of scaling data systems".*

Data Mesh shifts from a centralised data lake or warehouse to a more federated architecture where data is owned by cross-functional teams who treat it as a product.

Traditional Data Architecture vs. Data Mesh **[2]**

- **Centralization vs Decentralization:** Traditional data architecture centralizes data management in a single system, while Data Mesh decentralises it across domains.
- **ETL vs ELT[1]:** Traditional systems rely on ETL for data integration whereas, Data Mesh prefers, agile, real-time ELT processes.
- **Small vs Large Organizations:** Traditional architecture suits smaller, static environments while Data Mesh excels in large, dynamic organizations.
- **Central Control vs Domain Autonomy:** Traditional approaches centralize control while Data Mesh has domain teams with autonomy-reducing dependencies.

Purpose and Scope of the Article

The purpose of this article is to examine the challenges of implementing decentralized data governance in Data Mesh architecture and provide strategic solutions for effective governance. Read on to understand the concept of decentralized data governance and its benefits such as flexibility and increased ownership.

---

[1] ETL: Extract, Transform, Load; ELT: Extract, Load, Transform

UNDERSTANDING DATA MESH

Understanding Data Mesh starts with recognizing it as a transformative shift in data architecture, moving from centralized to decentralized data management.

Core Principles of Data Mesh

The key principles include:

Figure 1: Principles of Data Mesh [1]

- **Domain-Oriented Decentralization:** It enhances agility and responsiveness by distributing data ownership and reducing central bottlenecks.
- **Data as a Product:** It focuses on delivering high-quality and reliable data with clear ownership, accountability, and user-centric design for effective data utilization.
- **Self-Serve Product Infrastructure:** It provides scalable tools for autonomous management. This infrastructure reduces dependence on central teams.
- **Federated Computational Governance:** It integrates global standards with domain-specific flexibility.

Key Differences from Traditional Data Architectures

We explored a few differences before in this article, here, we will explain the key differences in data mesh from traditional data architectures:

1. **Centralized vs Decentralized Data Management**
    - Traditional data architectures rely on centralized management, where all data is stored and processed in a single repository. It often leads to bottlenecks and scalability issues.
    - In contrast, Data Mesh adopts a decentralized approach which distributes data ownership and management across domain-specific teams. This shift enhances scalability and reduces central bottlenecks.
2. **Data Quality and Access Control**
    - In traditional systems, data quality and access control are managed centrally. It can lead to inconsistencies and slower adaptation to domain-specific needs.
    - Data Mesh, on the other hand, treats data as a ‘product’ with each domain team responsible for its data quality and access control.

CHALLENGES IN IMPLEMENTING DATA MESH

Cultural and Organizational Challenges

Shifting to a Product Mindset

Transitioning to a product-oriented approach requires a cultural shift with teams. A data product mindset is a core principle of the Data Mesh paradigm as it treats data as a product.

The product mindset strategizes on creating a customer-centric focus to prioritise the needs and preferences of the customers. It commits to ongoing improvement and iteration, similar to how products are refined over time based on user feedback.

It also concentrates on delivering tangible value through each business function or service. This aligns efforts with customer goals and expectations.

Aligning Cross-Domain Teams:

The coordination and alignment of objectives across multiple domain teams can be difficult. Therefore, effective collaboration and communication can ensure that the data products meet organizational standards and integrate across different domains seamlessly.

It can take place by integrating:

- **Clear Contracts and Interfaces**: To establish clear data contracts and API.
- **Centralized Data Catalog**: It can help in registering data products and their metadata.
- **Standardized Processes:** Framework implementation for data transformation and access controls.

Thus, by addressing these aspects, organizations can achieve effective alignment between cross-domain teams.

Technical Challenges

Ensuring Data Interoperability

Data interoperability is a fundamental challenge due to the decentralized nature of data management. Each domain manages its data products, which can lead to inconsistencies in data formats, schemas, and standards across different domains. The issues can arise from:

- **Diverse Data Formats and Schemas:** Different domains might use varying data formats which leads to compatibility issues.
- **Complex Data Integration:** Ensuring that data flows correctly across domains and maintains consistency can address schema evolution[2].

Integrating with Existing Systems

Integrating Data Mesh with existing systems presents significant challenges:

- **Legacy Systems:** It requires creating interfaces that bridge old and new technologies. The solutions include adapters and wrappers, incremental integration, and API development.
- **Data Pipeline Complexity:** It shows how data mesh must handle diverse data sources and formats. It can be done by modular pipelines, unified data contracts, and automation & orchestration.
- **Resource Allocation:** Allocating resources is crucial for integrating Data Mesh. The strategies may include cloud-based solutions, a dedicated team to handle integration tasks, and monitoring & optimization using tools.

Governance and Compliance Issues

Maintaining Data Quality and Consistency

Data quality and consistency are challenging as the lack of quality can lead to discrepancies and the creation of unreliable data for decision-making.

For example: A global retail chain struggling with inconsistent inventory data across regions due to lack of consistency and low data quality maintenance.

Addressing Data Security and Privacy Concerns

Securing sensitive data and adhering to privacy laws becomes more complex in a decentralized system. Each team must implement its security measures, increasing the risk of breaches.

For example: A financial firm might struggle with maintaining security across various departments while handling customer data.

STRATEGIES FOR EFFECTIVE DECENTRALIZED DATA GOVERNANCE

Developing a Data Mesh Strategy

Setting clear objectives and goals

A Data Mesh Strategy begins with forming clear objectives and goals:

1. **Identify Business Needs:** Define the key business objectives that the Data Mesh strategy aims to support, such as improved data accessibility, faster decision-making, or enhanced data quality. Align these objectives with broader organizational goals to ensure the data strategy drives overall business success.
2. **Establish Success Metrics:** Develop specific, measurable, achievable, relevant, and time-bound (SMART) metrics to evaluate the effectiveness of the Data Mesh implementation. There are metrics such as data product usage rates, response times etc.
3. **Prioritize Use Cases:** Identify and prioritize high-impact use cases that will benefit from a Data Mesh approach. Focus on areas where decentralization can address current bottlenecks or provide significant value.
4. **Self-Implementation Milestones:** Define clear milestones and timelines for the implementation of Data Mesh components. These should include domain team formation, data product development, platform deployment, and governance setup [6].

Defining domain boundaries and responsibilities

1. **Map Business Domains:** Each domain should align with business functions or operational units, such as sales, finance, or customer service.
2. **Assign Domain Ownership:** Each domain should have a dedicated team responsible for end-to-end management of their data assets. These teams are responsible for the data ingestion and transformation which allows them to independently manage data products.
3. **Define Data Interfaces:** Businesses should establish clear interfaces and contracts between domains. Specify how data will be shared, accessed, and integrated, including data formats, access controls etc.

[2] Schema evolution: The process of updating or modifying the structure of a database schema over time.

4. **Implement Coordination Mechanisms:** It is important to set up mechanisms for inter-domain collaboration and conflict resolution.

Implementing Data Mesh Components

Data product management and ownership

Businesses should empower the domain teams to manage and own data products. They should be responsible for the lifecycle and quality of the data they provide.

The implementation steps include defining ownership > developing data products > and maintaining quality.

Domain teams should manage and publish their data products in a cleaned and managed form.

Source-aligned domains should include reference IDs and can add data from other domains if it's transformed and useful for decisions. Aggregate and consumer-aligned domains can use data from other domains to meet their users' needs.

Establishing a self-serve data platform

A self-serve data platform means creating an environment where end-users can access, understand, and use data independently. The goal is to minimize reliance on data engineers by providing users with the tools and information they need to manage and analyse data on their own. The implementation process is through:

- **Data Process Creation:** Gather user requirements to develop and validate datasets. Build foundational datasets, ensuring they are clean, structured, and traceable.
- **Access Mechanisms:** Provide static and dynamic reports, BI tools, SQL access, and APIs. This enables users to generate and customize reports based on their needs.
- **Reduce Dependencies:** Implement data catalogues and comprehensive documentation for easy data discovery and understanding.
- **Support and Training:** Offer training resources and ongoing support to help users effectively utilize the self-serve platform. It will minimize the direct reliance on data engineers.

Enabling federated governance frameworks

A federated governance framework allows each domain to manage its data independently, scaling its own pace. Federated governance strikes a balance between domain autonomy and inter-domain coordination.

Designing a Federated Data Governance Model blends centralized and decentralized elements, implementing governance priorities in organizational contexts. It involves centralizing (data standards, regulatory policies) and decentralizing (data curation and processing tasks) [3].

Best Practices for Data Mesh Adoption

Follow the best practices while implementing Data Mesh:

Incremental implementation and piloting

- Begin with a pilot project or a limited number of data domains.
- Implement data mesh incrementally, addressing one domain or capability at a time.
- Regularly assess the pilot's effectiveness, gather insights, and make necessary adjustments.

Building a culture of data stewardship

- Assign clear responsibilities to data domain owners and ensure they understand the value of their roles in maintaining data quality.
- Encourage collaboration among data engineers, domain experts, and data stewards t enhance accountability.
- Provide ongoing training and resources to empower data stewards and domain owners with the skills needed to manage their data domains effectively.

Continuous improvement and feedback loops

- Implement feedback mechanisms to refine processes, tools, and governance practices.
- Regularly track key performance indicators (KPIs) related to data quality, accessibility, and usage.
- Stay responsive to changes in the business environment, technology, and regulatory requirements.

CASE STUDIES AND REAL-WORLD EXAMPLES

Successful Implementations

Overview of companies that have adopted Data Mesh

- **Delhiivery's** deployment of a data mesh architecture is a strong example of successfully managing large-scale data operations. They ingest terabytes of data daily using Apache Hudi on Amazon EMR and Amazon RDS PostgreSQL [4]. It processes and enriches data inputs for their data lake.

- **Netflix's** Data Mesh is an evolution from their Keystone system. It serves as a robust data processing platform. It handles diverse data sources, supports various database connectors like MySQL and Cassandra, and accommodates complex processing patterns. The architecture comprises a Control Plane for pipeline management and a Data Plane for processing [5].

Key takeaways and lessons learned

1. **Scalability and Flexibility**
   - Delhiivery's deployment of Data Mesh has demonstrated the power of handling large-scale data operations, processing terabytes of data daily.
   - The evolution from Keystone to Data Mesh has allowed Netflix to handle diverse data sources and complex processing patterns.
2. **Data Ownership and Domain-Centric Design**
   Both Delhivery and Netflix have highlighted the importance of domain-oriented ownership of data. By allowing teams to manage their data pipelines and processes, they've been able to increase efficiency and reduce bottlenecks.
3. **Technology Stack and Tooling**
   The successful implementation of Data Mesh requires a strong technological foundation. Delhivery's use of Apache Hudi and Netflix's integration of various connectors and processors illustrate the need for robust tooling that supports diverse data processing needs while enabling seamless data movement across platforms.

Common Pitfalls and How to Avoid Them

Lack of Clear Data Ownership

In many organizations, the transition to a Data Mesh model creates confusion around data ownership. The solution is to establish clear guidelines for domain-centric ownership.

For example, Delhivery avoided this pitfall by defining distinct domains and assigning dedicated teams to manage their respective data products.

Overwhelming Complexity

Data Mesh architecture can become overly complex, especially in large organizations with numerous data sources and processing requirements. The solution is to start with a pilot project and focus on a single domain.

For example, Netflix implemented their Data Mesh incrementally, first addressing Change Data Capture (CDC) use cases before expanding to other areas.

CONCLUSION

Summary of Key Insights

Data Mesh represents a transformative shift in data architecture, moving from centralized to decentralized management. This model promotes domain-oriented decentralization which treats data as a product and enables self-serve infrastructure with federated governance.

Future Outlook for Data Mesh

As data needs continue to evolve, Data Mesh is poised to adapt and grow, with advancements in technology and methodology. Trends like Artificial Intelligence (AI)-driven data governance, advanced real-time analytics, and enhanced automation in data pipelines are expected to further strengthen Data Mesh architectures. The rise of cloud-native tools and serverless architectures will also play a crucial role in the scalability and agility of future

Final Recommendations for Practitioners

For practitioners looking to adopt Data Mesh, the following recommendations are crucial:

- **Start Small:** Begin with a pilot project, focusing on a single domain to manage complexity and ensure smooth adoption.
- **Emphasize Ownership:** Clearly define domain boundaries and responsibilities to avoid confusion and enhance data quality.
- **Invest in Tooling:** Ensure your technology stack is robust and capable of supporting diverse data processing needs.
- **Implement a Data-Driven Culture:** Encourage a product mindset and build a culture of data stewardship to maintain high-quality and reliable data products.
- **Stay Agile:** Continuously assess and adapt your Data Mesh strategy to evolving business needs and technological advancements

REFERENCES

[1]. Z. Dehghani, "Data Mesh," *Google Books*, 2022. https://books.google.com/books?hl=en&lr=&id=ReWYEAAAQBAJ&oi=fnd&pg=PT6&dq=zhamak+dehghani+data+mesh&ots=cM_n90mYVv&sig=XRC23KmIJHiWHTJytgazhpEYKlU (accessed Aug. 26, 2024).

[2]. Sameer Paradkar, "Beyond Centralization: Data Mesh and the Reimagining of Data Architectures," *Medium*, Jan. 13, 2024. https://medium.com/oolooroo/beyond-centralization-data-mesh-and-the-reimagining-of-data-architectures-ae66bcbb4b2d#:~:text=Decentralized%3A%20Traditional%20data%20architectures%20like (accessed Aug. 26, 2024).

[3]. K. Burchardi, D. Nath, and J. Lan, "Federated Data Governance Model Add EyeEm Image," 2024. Available: https://media-publications.bcg.com/Federated-Data-Governance-Model.pdf (accessed Aug. 27, 2024).

[4]. Data, "Delhivery: Data mesh implementation ingesting terabytes of data," *Awscloud.com*, 2023. https://in-resources.awscloud.com/aws-summit-online-india-2021-big-data-and-analytics/delhivery-data-mesh-implementation (accessed Aug. 28, 2024).

[5]. Kumar Ashutosh, "Netflix's ability to keep viewers glued to their screens," *Linkedin.com*, May 19, 2024. https://www.linkedin.com/pulse/netflixs-data-mesh-look-next-gen-pipeline-solution-kumar-ashutosh-nrqyc/ (accessed Aug. 28, 2024).

[6]. I. A. Machado, C. Costa, and M. Y. Santos, "Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures," *Procedia Computer Science*, vol. 196, pp. 263–271, Jan. 2022, doi: https://doi.org/10.1016/j.procs.2021.12.013. (accessed Aug. 26, 2024).