How To Realize A Context-Driven Data Mesh:
Engagement Among the Data Mesh Designer, Decision Maker, Data Analyst, and Information Consumer

Erica L. Dretzka
Chief Digital and AI Office (CDAO), CTO
Washington, DC
Erica.l.dretzka.civ@mail.mil

Abstract

The digital world is acknowledging its transition from centralized to distributed architecture. This evolution, however, requires deliberate action. Consider linking Intellectual Property (IP)-sensitive proprietary modules with context-tuned data management practices, compliant with regulations yet differentiated from peers. This challenge requires a sophisticated technical and strategic response with a well-formed implementation plan.

Evoking the data mesh as idealistically formalized in research, we will illustrate the viability and interoperability of four composable foundational building blocks for any organization. A precursor to at least ten additional mesh components, these establish a data-mesh-worthy technical backbone for any organization.

This paper builds on the 2023 presentation on [x]BOMs, knitting them into the next phase of the data mesh ecosystem, anticipating the need for well-defined DevSecOps environmental control components.

ABOUT THE AUTHOR

Erica Dretzka: Ms. Erica Dretzka is a seasoned data scientist with over 20 years of experience in various industries, including Insurance, Energy, and National Defense. She has established two data science teams inside the Department of Defense (DOD) and led the development of advanced Artificial Intelligence (AI) and Machine Learning (ML) models, which have been adopted by both DOD and Department of Homeland Security. A data scientist by trade she and her team are employing engineering-based methods to design the optimal reference architecture and bridge strategy to support AI and data-backed mission support at the scale and resilience required for DOD.

1 INTRODUCTION

The data mesh was popularized in May 2019 when Zhamak Dehghani authored the groundbreaking article “How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh”[1] In it, Dehghani asks the reader to momentarily suspend deep assumptions and biases established by current paradigms of traditional data platforms, noting that despite broad acknowledgement of the need to become intelligently empowered with data, organizations are hindered by cost, technical modernization, and organizational resistance.

Some of the largest organizations in the world, even those with legacy systems, are following the recommendation to suspend assumptions and are actively working to adopt the data mesh. Leadership is funding these initiatives because they see the promise of a high return on investment (ROI), particularly considering the exponential expansion of data creation. Consider that in 2018 the total amount of data created, captured, copied and consumed in the world was 33 zettabytes (ZB) – the equivalent of 33 trillion gigabytes. This grew to 59ZB in 2020 and is predicted to reach a mind-boggling 175ZB by 2025 [2].

Investments and continuing innovations in technology and data management approaches are making great strides. Yet, they leave something lacking as leaders and decision makers recognize the major role of organizational change. Culture shift is a key consideration as a new era of data driven intelligence will help to shape and optimize decisions and facts that inform the optimization of business process outcomes.

This paper is structured to help decision makers, leaders, information consumers, and professional practitioners understand each other's perspectives and challenges. Finally, it proposes four foundational pillars of a technical data mesh architecture. Ultimately, it aims to both overcome miscommunications on the data mesh concept and the associated terms and provide a tangible way ahead.

2 DATA MESH PRINCIPLES

A Data Mesh differs from a Data Fabric in that it incorporates an element of organizational change as well as comprehensive information interoperability across all data objects, both simple and complex. Thus, the four principles of a mesh include: domain-oriented decentralized data ownership and architecture, data as a product, self-serve infrastructure as a platform, and federated computational governance[3]. Figure 1 illustrates them and their interrelationships.

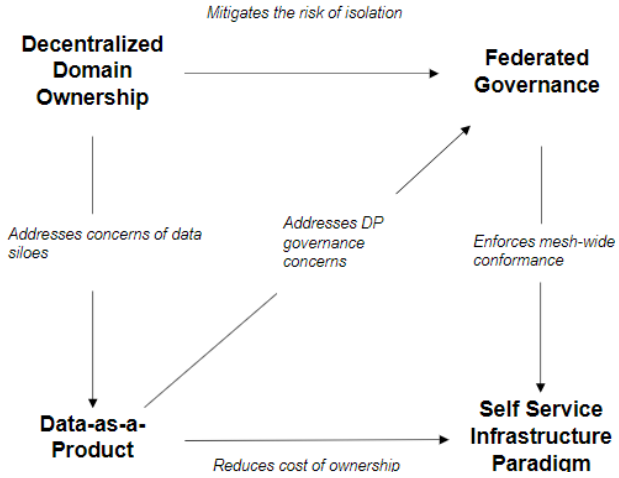

Figure 1 Interrelationships of the Four Mesh Principles

Decentralization of data ownership ensures decentralization and distribution of responsibility to people who are closest to the data[3]. Put simply, the data producer, wherever they are, knows the data best, and while oversight management functions have an essential role to regulating the enterprise behaviors of compliance, it is never possible to please everyone for all things, and compromises must be made. The mesh strategically supports this by digitizing data lineage and provenance, thus enabling traceability to the source.

Domain-oriented decentralized data teams calls for an architecture that arranges the analytical data by domains in which the domain's interface to the rest of the

organization not only includes the operational capabilities but also access to the analytical data that the domain serves[3]. They inform things such as why the data was generated, its definition and characteristics, and caveats to using it, as well as all subsequence changes, whether it enhances, corrects, extends or supplements the original source content. The mesh ideally supports organizational and program management insight subsequent to the data's release to the mesh as it can track endpoint ingests and transformations, and each contributing actor along the data journey.

Data as a product mindset is what happens to the data when it is ingested into the data mesh. The decentralized data teams provide raw data and its definitions to data product teams (DPTs), which perform data engineering activities to provide a more user consumable data object, or perhaps more advanced analytical techniques to massage it into an information set that is focused on a use case. This is important because the same data elements may be repurposed to answer different questions, causing its treatment, such as interpolation of missing data, merging with associated with other data sets, to be repositioned to meet the need.

Self-serve data infrastructure promotes the notion of a logically federated data environment, creating an interoperability landscape, by making data and data mediation and management tools available to those with appropriate privileges to orchestrate information across the entire mesh.

3 FOUNDATIONAL MESH COMPONENTS

Foundational Mesh Components

As illustrated in Figure 2, organizations are challenged to make legacy IT systems work with data of varied formats. The technical side of the mesh is a series of component building blocks that make interoperability possible.

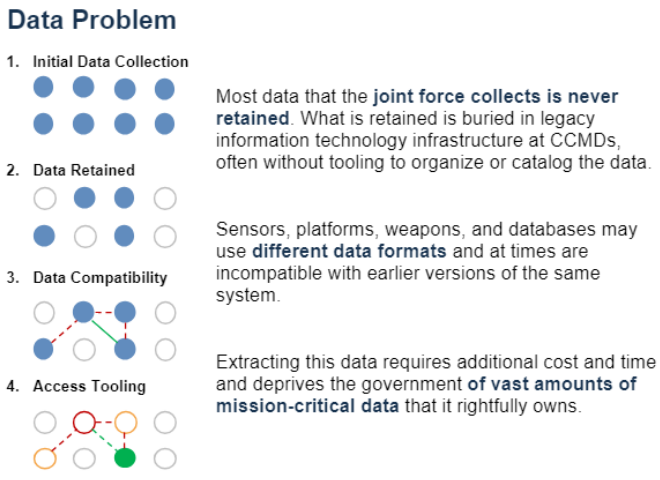

Figure 2 Access Tooling to Address the Data Problem

This paper proposes a combination of four foundation capabilities for any data mesh instantiation: a) Unique Identifiers (UID), b) Canonical Controlled Vocabulary (CCV), c) federated metadata catalog (FMdC), and d) Bills of Materials (xBOMs, eXtensibleBOMs). Operationally, each of these capabilities is benefitted by using graph database technology to manage the scale demands of larger organizations, while still natively supporting all size organizations. However, alternate technology implementation approaches are envisioned as viable if the mesh is envisioned to be bounded to a finite pre-computed size that will not scale beyond practical technology constraints of compute, memory, query mechanisms.

Universal Identifiers (UID) satisfy the mesh's first major need to track every asset uniquely and use a single "key" for all circumstances of use and context. The suboptimization of not having a UID would require extensive compute and query resources for each and every request. For example, without a UID, one can imagine an American citizen who has attended college, is employed, has a passport, a driver's license, and a monthly utility bill. Each of these has generated a distinct identification number for this person. If that person asks for their credit score they typically enter their social security number (SSN). A mesh would then ping multiple databases to develop a profile of this person's citizenship, education, employment, travel record, and driver's record, yet it only has the SSN.

Canonical Controlled Vocabulary (CCV) builds on the UID component to develop semantic congruence across all contexts providing a consistent understanding and meaning for every term used. Organizations are riddled with multiple ontologies which are tailored to a specific use case or domain. Reconciling those varied domains is non-trivial and often requires finding an ontologist to manage creating a consistent mapping or defining alternate differentiating terms to maintain correctness within and across all contexts. This is not tenable when performing analytics at scale. Historically Entity Relationship (ER) Diagrams or more recently ER Models, which identify entities to be represented in the database and representation of how those entities are related[13], have been used to perform a semblance of this task. Although ER models are used primarily for database design, they often do not store domain knowledge[14]. Reviewing that a key tenet of the mesh is its relatability to different contexts, this is a critical gap in strictly using ER models for this purpose. More complex ontological approaches are extending the traditional ERD/M, however, the multi-context challenge remains due to the independent nature of each context, or coercive techniques can be applied to "force" contexts to shift natural behaviors to be compliant with governed ontology models

The Federated Metadata Catalog (FMdC) is not necessarily a problem in recent years, as the industry has developed excellent solutions as long as the enterprise is able to use a single FMdC. The new challenge, especially with the distributed nature of data production and management, as envisioned in typical data mesh behavior, is how to contextually relate all of the catalogs to each other, which identifies the need for CCV and UID, or some other non-mediated semantically consistent approach.

xBOMs and eXtensibleBOMs orchestrate the decades-old concept of Bill of Materials (BOMs) to describe the components of information transiting across the mesh. The resulting BOM fabric utilizes UIDs and CCVs to formulate or populate BOM data which is referenced in FMdCs, developing a digital mesh traversable at scale.

The three types of BOMs are explained in Figure 3 using the process of brewing coffee. In this, coffee beans and water are the two discrete elements, each having its own xBOM type 1, which includes all of the necessary information about the beans and water that a user would need to understand. We are then given the understanding that a cup of steaming coffee is composed of coffee beans and water in xBOM type 2, where all of the necessary elements are identified, yet we don't know how they are "composed" into the final coffee product. It isn't until xBOM type 3 that we know how those two elements are treated and joined to develop a steaming cup of coffee.

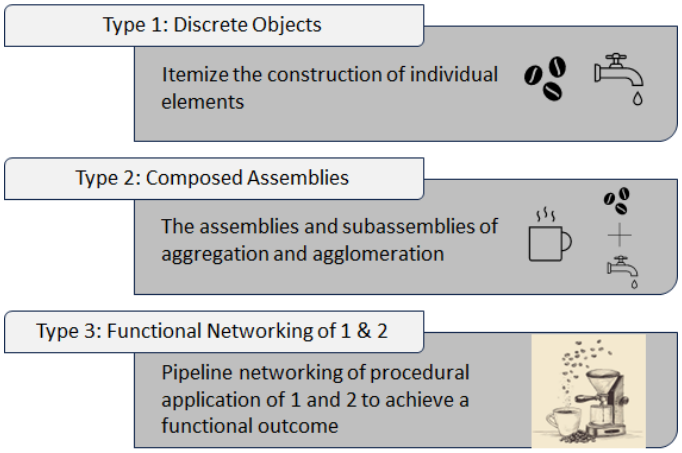

Figure 3 Three Types of xBOMs

These first four components do not portend to be the mesh in its entirety but assert themselves as the cornerstone for the data-first approach to its construction. As described above, the full mesh requires observability, protection, authentication, authorization, and many other components. These additional capability concepts and interoperability models are under construction as of the writing of this paper.

It is significant to bear in mind that a complete data mesh will achieve full interoperability of any data for any user in any context, where it makes sense and is appropriate, which includes the capabilities (potentially a service mesh) to allow meaningful use of the information.

4 DATA MESH ACTORS

Figure 4 illustrates the interaction of the policies inside the data mesh. It layers key concepts from three documents into a stack of high-level layers tuned for the data mesh. The first is the DoD DevSecOps Reference Architecture[15]; the second is the assembly of the seminal components to a skeleton data mesh; the third is the DoD Mission Engineering Guide [16]. They are depicted along a vertical continuum that builds from physical functionality (network and transport) through a layer of data processing (data product) up to mission effectiveness (mission engineering) and operational relevance (commander's intent).

While execution and purpose differ for each, their intents necessarily overlap.

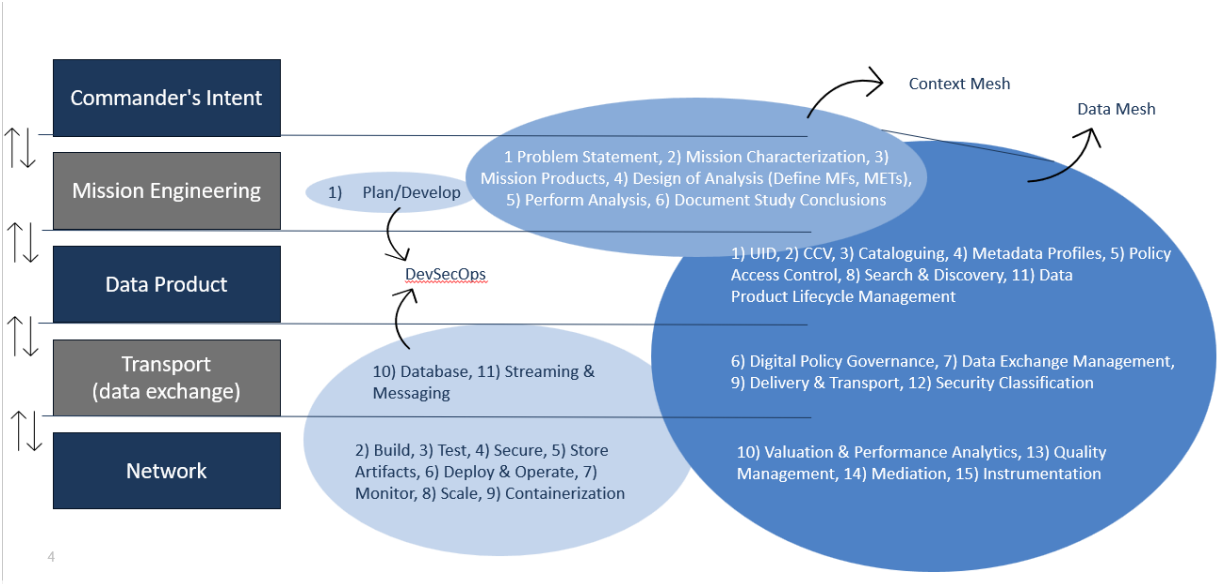

Figure 4 Policies Associated with the Data Mesh

The Decision Maker

Strategic vision and organizational alignment are the starting point for any project and generally correspond to the Mission Engineering layer. Notably, the data mesh is frequently referred to as a "socio-technical paradigm shift". The "socio" portion of that description requires that decision makers grow the organization towards the mesh principles (culture shift).

Organizations have difficulties with the transition toward federated governance associated with the data mesh concept, the shift of responsibility for the development, provision, and maintenance of data products, and the comprehension of the overall concept[4]. Executive sponsorship and championing of this new strategic approach is a vital element of any successful data mesh plan.

The Data Analyst

Data practitioners operate primarily in the Data Product layer. They are moving as fast as possible to adopt the mesh without disrupting product delivery. This is no small feat, particularly as their tools and instruments are rapidly changing. For example, data storage has traditionally been a local large disk drive space, a data warehouse (DW), or a purpose-built relational database running on specialized hardware either on the premises or in the cloud[5] containing data in its processed or filtered form. In contrast, Data Lakes (DL) exist almost entirely as repositories storing raw data in their original, or raw, formats and providing a common access interface, thus they do not inherently have the analytics features associated with DWs[6]. Lakehouses (LH) are an emerging data management architecture trend as a unique data storage solution for all data—unstructured, semi-structured, and structured—while providing the data quality and data governance standards of a data warehouse. A data practitioner relies on both stable connections to each data storage location and data dictionaries and metadata explaining the contents of each item in the dataset. The data practitioner must change code and technology when an endpoint changes, a data element is added, or a new dataset is created, or new analytic tools are available.

Consider the data mesh (DM) a platform unifying the above. It makes all data held in any of the data storage models, mentioned above, accessible via pipelines and mediation tools with flexible security and consumers' knowledge of operations, contexts, and usefulness. This ultimately enables scalable Data Driven Decision Making (DDDM), even when data has a high degree of the popular 5 V's (Volume, Velocity, Variety, Veracity, and Variability), making it difficult to extract value from the data[8].

The Information Consumer

One or more degrees removed, the information consumer uses analytic products produced by the data analyst. It is typically using the Commander's Intent layer. All actors in the data mesh should develop use cases to ensure the data is prepared for mission relevance and will be available as needed. The information consumer may be a machine, human, or a combination. The mission space may be real-time operations, boardroom decision support, vulnerability exploration, or pure data science. The universe of uses for data is a continually evolving practicum and having the ability to identify if existing data mesh capabilities are sufficient is critical to continued success and strategic advantage.

Roles of the Data Mesh Designer

Although Zhamak Dehghani has proposed some core concepts about Data Mesh principles and architecture, there are limited, if any, consolidated contributions in the scientific community about the practical implementation of a Data Mesh and its contribution in the development of data-driven information systems[9]. This is the Systems Engineers' (SE) and Systems Architects' (SA) job. Their work concentrates in the Transport and Network layers.

SEs face in three directions: the system users' and stakeholders' needs and concerns, the enterprise and project managers' financial constraints, and the capabilities and ambitions of the specialists who must develop, built, test, and deploy the elements of the system[10]. In short, they must translate needs defined by the organization's strategists and practitioners into a mesh domain model that is organizationally sensitive, responsive to current state and desired future capability requirements, and financially responsible.

One such domain model composed of three components that support its operation has been proposed: the Mesh Catalog (where all the nodes of the Mesh are listed and detailed), Change Keeper (component where the changes that occur are registered), and a Mesh Communication Channel (which allows the various teams that work on the Mesh to communicate quickly and efficiently)[9]. This excellent characterization focuses on the operations of the mesh. In order to make the mesh relevant to strategists and practitioners, the following proposal iterates on this architecture to develop a version which focuses on four main tactical elements of how data flows through the system versus the system itself.

5 METADATA LAYER

Any use case for a Computer-Aided Engineering (CAE) task, including simulators, Artificial Intelligence, and High Performance Computing (HPC), assumes intelligible data. That is, it requires metadata that is easily searchable, consumable, and comprehensible in the format prescribed by that use case, yet despite this demand, metadata management is a core and oft-

neglected enabler of data analyses. Academic studies propose a “Scientific Data Bag”, essentially a container of metadata and an object-oriented central database or DL in a predefined schema[8].

First, we will propose a distilled version of an architecture such as the Open Systems Interconnection (OSI) tailored for the modern concept of a data mesh. Second, we will explain the foundational four mesh components and how they interoperate or notionally assemble to achieve the basic core capabilities of a data mesh, that ultimately allow other capabilities to be added to complete full interoperability.

Tailored Version of the Data Mesh Stack
Below is an explanation of the five layers in Figure 4 above.

> *Transport and Network layer* contains data storage and exchange services such as Application Programming Interfaces (APIs), Radio Frequency (RF), and even manual handoff of data for highly classified and air-gapped systems.
>
> *The Communication and Translation layer* contains services which automate semantic collisions and deviations in different data sets and negotiate disjoint keys and identifiers built for specific purposes. This layer contains most of the mediation tools in the mesh. Both this and the Transport Layer account for the first set of authorization, authentication, and continuous monitoring of user access. It accounts for considerations such as classification level, access controls, and encryption.
>
> *The Data Product Layer* is where the cataloging, metadata profiling, policy access control (authentication, etc. for data), and other data-specific functions occur.
>
> *The Mission Engineering and Commander’s Intent layer* architects how the mesh and the outputs of the mesh connect to the organization’s mission, both at a high-level mission setting function and an on-the-ground manager’s function. This frequently overlooked step is critical to ensuring the structure and operations of the mesh are relevant to the organization.

Services such as performance optimization and monitoring provide constant Communication across all these layers.

6 USE CASE DEVELOPMENT AND ILLUSTRATION

Ultimately, the data mesh design must materially support the full range of mission use cases imaginable by the organization’s strategic leaders. While some consider this a far-fetched dream, recent technological advances have lowered the bar, making this feasible. Examples of problem sets that will be addressed by any well-constructed data mesh include:

> *Boardroom*, such as data and analytics (data science, Artificial Intelligence (AI), and Machine Learning (ML) development) and operational use cases (Supply Chain Risk Management (SCRM) and Cyber-SCRM (C-SCRM)).

Operations, such as real time edge operations and predictive maintenance.

Modeling and Simulation (M&S), such as digital twins[14] and Model Based Systems Engineering (MBSE)

Use Case Illustration

Figure 5 illustrates a walkthrough of the mesh. Imagine a military exercise with friendly and unfriendly aircraft accompanied by a friendly ship and a civilian sailboat in the open ocean. They must first authorize each other's access to intel and do so through authentication and authorization. Next, aircraft declare themselves to be "flying blue forces" while the ship refers to the open water civilian boat as "blue water". This is confusing without context, so they must consult a semantic converter which relies on mediation hub services. Next, considering that there is an unfriendly ahead, they must consult the federated data catalog to know what they have available to them should a fight ensue, requiring consultation of the data product search. Without delay the friendly ship and aircraft must intuit the nature of the unfriendly forces ensuring accuracy via data quality checkers, privileges via policy administration, and currency in life cycle management. All of this is communicated via data exchange management (e.g., APIs, radio frequency) and informs the decision maker which both determines the best course of action and notifies logistics of any requirement for items to be restocked upon return to home base.

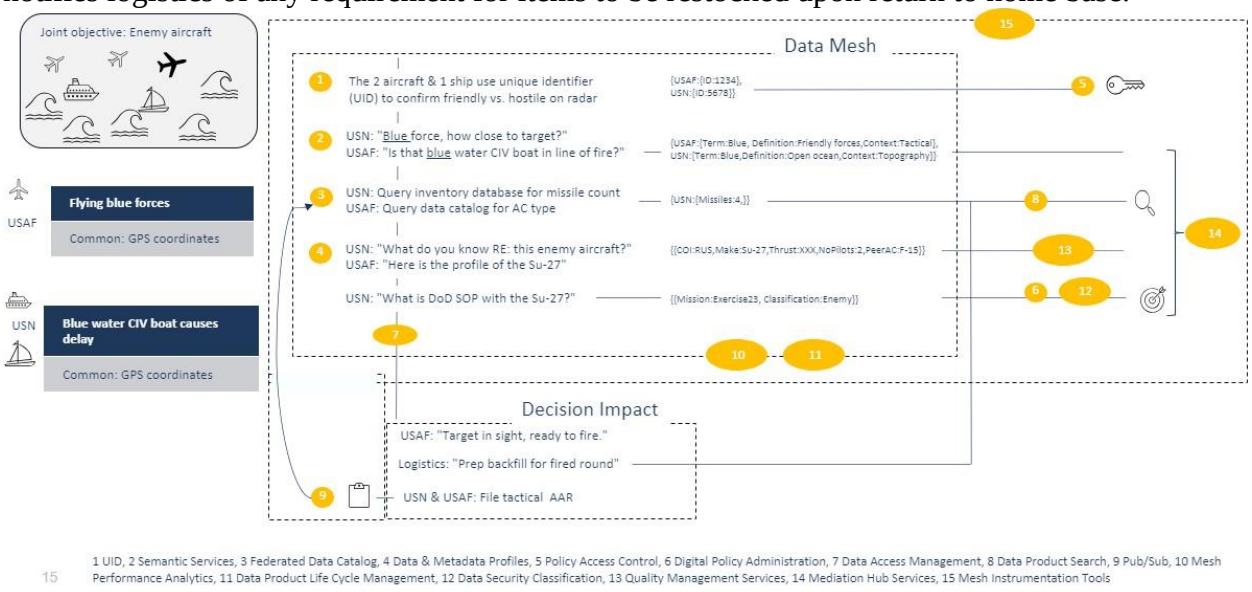

Figure 5 Sample Use Case Illustrating Data Mesh Component Interoperability

This is merely one scenario of how the data mesh serves the enterprise. Countless others are imaginable, ranging the gamut from boardroom to operational use cases.

7 CONCLUSION AND FUTURE WORK

The compelling argument for adopting a mesh is challenged as organizations often misinterpret the key principles of data mesh. They tend to label any effort to work with decentralized data as "data mesh." [18] This means that any successful mesh implementation must engage decision

makers, leaders, information consumers, and professional practitioners with full respect for each other's disciplines.

Gartner recommends a seven-step process to adoption: Build a Strategy, Engage with Stakeholders, Assemble the Team, Develop Best Practices, Implement the Concept, Collect Feedback, and Scale to Other Domains (cycling back to Engage with Stakeholders)[19]. Thoughtworks has proposed a similar adoption model. Learning best practices from those who have gone before them is particularly important for legacy organizations, which are ladened with technical debt and due to their stature in their industry cannot afford to slip in delivery.

The mesh is possible. Organizations such as the pharmaceutical company Gilead Sciences and JP Morgan Chase are said to have adopted it. That should encourage other organizations that once they adopt the crystal-clear understanding of data mesh as a fundamental organizational construct.

Future Work

The next steps of this research are to develop use cases for mesh application, perform technical experiments, and initiate organizational change. Use case development is a complex process with its own discipline, but which each organization can tailor. Where existing solutions either exist or do not fit with an organizations' needs, they must understand how to best prototype where necessary and make interoperable the proposed key foundational components detailed above. Finally, and perhaps the most difficult of them all is to initiate cultural change from the top down and from the bottom up.

References

[1] Dehghani, Zhamak.(May 20, 2019). *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*. https://martinfowler.com/articles/data-monolith-to-mesh.html

[2] Vopson, Melvin. (May 7, 2021). *Just How Much Data The world's data explained: how much we're producing and where it's all stored.* https://www.weforum.org/agenda/2021/05/world-data-produced-stored-global-gb-tb-zb/

[3] Dehghani, Zhamak. (Dec 3, 2021). *Data Mesh Principles and Logical Architecture.* https://martinfowler.com/articles/data-mesh-principles.html

[4] Bode, J. (2023, February 3). *Towards Avoiding the Data Mess: Industry Insights from Data Mesh Implementations*. arXiv.org. https://arxiv.org/abs/2302.01713

[5] Rehman, K.U.U.; Ahmad, U.; Mahmood, S. A Comparative Analysis of Traditional and Cloud Data Warehouse. *VAWKUM Trans. Comput. Sci.* 2018, *6*, 34–40.

[6] Nambiar A, Mundra D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data and Cognitive Computing*. 2022; 6(4):132. https://doi.org/10.3390/bdcc6040132

[7] Armbrust, M.; Ghodsi, A.; Xin, R.; Zaharia, M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In Proceedings of the Conference on Innovative Data Systems Research, Virtual Event, 11–15 January 2021.

[8] Julian Ziegler, Peter Reimann, Florian Keller, Bernhard Mitschang, A Graph-based Approach to Manage CAE Data in a Data Lake, Procedia CIRP, Volume 93, 2020, Pages 496-501, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2020.04.155.

[9] Machado, Ines Araujo. (2021). *Data-Driven Information Systems: The Data Mesh Paradigm Shift.* https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1373&context=isd2014

[10] Kossiakoff, Alexander. Systems engineering Principles and practice. Wiley, 2020.

[11] Valle, M., Favre, J., Parkinson, E., Perrig, A., Fahrat, M.. Scientific Data Management for Visualization Implementation Experience. In: SimVis. 2005, p. 347–354.

[12] Kimball R, Ross M. The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling. Wiley. 2013.

[13] Geeks for Geeks. (Sept 20, 2023). Introduction of ER Model. https://www.geeksforgeeks.org/introduction-of-er-model/

[14] Singh S, Shehab E, Higgins N, et al. Data management for developing digital twin ontology model. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture. 2021;235(14):2323-2337. doi:10.1177/0954405420978117

[15] Department of Defense Chief Information Office. (Aug 12, 2019). *DoD Enterprise DevSecOps Reference Design.* https://dodcio.defense.gov/Portals/0/Documents/DoD%20Enterprise%20DevSecOps%20Reference%20Design%20v1.0_Public%20Release.pdf?ver=2019-09-26-115824-583

[16] Department of Defense Office of the Under Secretary of Defense for Research and Engineering. (Oct 1, 2023). *Department of Defense Mission Engineering Guide.* https://ac.cto.mil/wp-content/uploads/2023/11/MEG_2_Oct2023.pdf

[17] Majchrzak, Jacek. Data Mesh in Action. Manning Publications, Shelter Island, NY, 2023. https://books.google.com/books?hl=en&lr=&id=sPynEAAAQBAJ&oi=fnd&pg=PA1&dq=data+mesh+principles&ots=sObSO00uU5&sig=GAgRzVBBBitaCtknBnMX58OVshQ#v=onepage&q=data%20mesh%20principles&f=false

[18] Launi, M., Edjlali, R, Simone, M. (2024, January 3). *Data and Analytics Essentials: How D&A Leaders Can Accelerate Data Mesh Adoption*. https://www.gartner.com/document/5080631?ref=solrResearch&refval=395382297&
[19] Launi, M., Edjlali, R, Simone, M. (2024). *Data and Analytics Essentials: How D&A Leaders Can Accelerate Data Mesh Adoption* [PowerPoint slides]. Gartner. URL
[20] Joshi, D., Spens, J. (2022, October 25). Data Mesh: Real Examples and Lessons Learned. https://www.thoughtworks.com/insights/blog/data-engineering/data-mesh-real-examples-and-lessons-learned