Name: Justin Grima

Student Number: 14248599

Course ID: MA5831 Advanced Data Processing and Analysis using SAS

Assessment: Data Processing Trends

Due Date: 11:59 PM AEST Sunday of Week 6

Word Count: 2,200 (2,000 +/- 10%) not including headers, sub-headers, and image and table description.

I. Introduction

In the current technological landscape, the rapid advancement and proliferation of products and services have led to unprecedented surges in big data. The International Data Corporation (IDC) predicts that data will more than double from 2022 to 2026, resulting in exponential growth in volume, velocity, and variety, posing significant challenges for organizations (Bode et al., 2023). Traditionally, organizations rely on monolithic data architectures like Data Warehouses, Data Lakes, and Data Lakehouses, with a central data management team handling the collection, processing, and management of organizational data for decision-making. However, the issues associated with these architectures call for a paradigm shift toward a truly data-oriented organization.

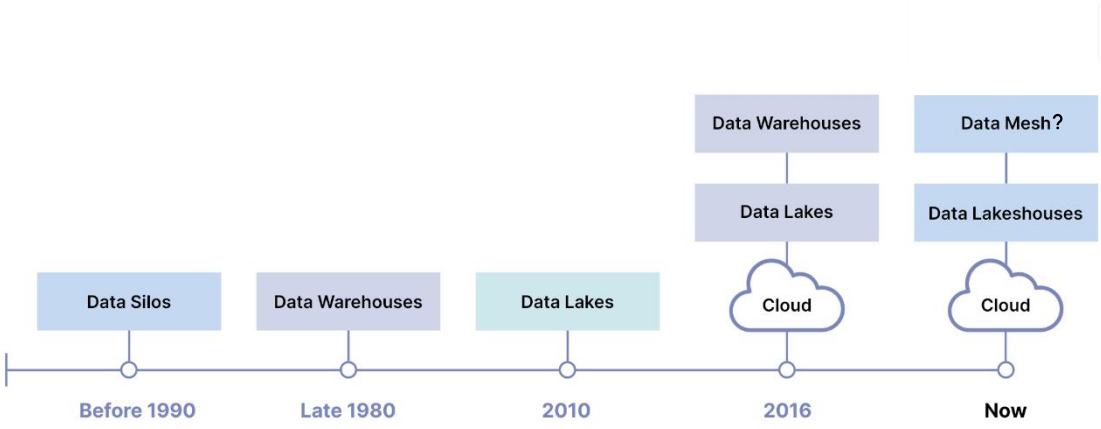
This literature review explores the emerging architectural paradigm of Data Mesh as a potential solution to challenges posed by the current monolithic architectures. Data Mesh adopts a decentralized, domain-oriented approach to data management, emphasizing self-service data handling and efficient collection, processing, storage, management, and governance. The transformative potential of Data Mesh lies in addressing the limitations of the monolithic approach, enabling companies to leverage their investments in Big Data and AI to generate measurable results, compete using data, drive business changes through analytics, and foster a data-driven culture (Info Q, 2019).

This paper investigates the concept of Data Mesh by reviewing peer-reviewed papers, discussion forums, recorded interviews and keynote speaker presentations, accredited websites, and professional papers. It provides an overview of the main data architectures and their associated challenges, a complete analysis of the Data Mesh architecture, its benefits, and concerns, and theorizes the role of a data scientist in a data mesh architecture. The outcome of this paper is to further contribute to the discussion of the Data Mesh and how it advances the current state of data systems.

II. Literature Review

Current Data Architectures: Analysis and comparison

Figure 1: Timeline of data architectures.



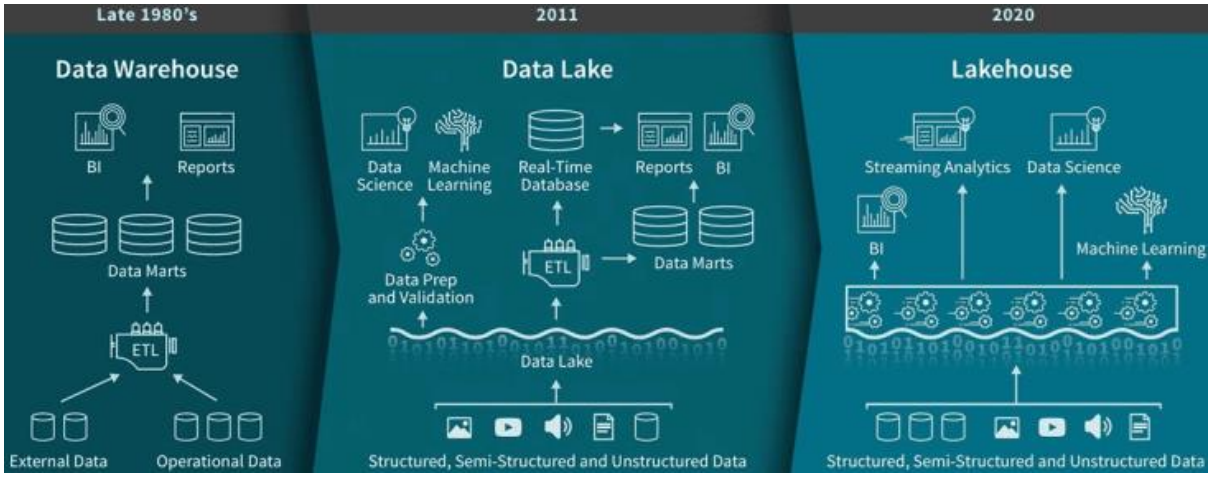
Note. Sourced from Agrawal, S. (2022)

In the late eighties, the Data Warehouse (Figure 2) emerged to address the proliferation of data, and organizations' need to store and access their information (SAS Institute Inc, 2017). Its centralized repository served for structured data from various sources, supporting historical analysis and reporting for decision-making. However, high costs, inability to handle semi-structured and unstructured data, slow data ingestion, and scalability challenges are its downfall. In 2011 the Data Lake (Figure 2) concept emerged as an improvement. Like the Data Warehouse, its' architecture serves as a centralized repository but stores structured, semi-structured, and unstructured data, at any scale, and offers significantly higher storage capacity at a lower cost while maintaining reasonable data access speeds (Amazon Web Services, Inc. 2023).

Soon after, cloud service was implemented for Data Warehouses and Data Lakes, providing cost-effective storage, scalable computing, and managed service delivery (Bassman, 2023). The emergence of the Data Lakehouse (Figure 2) followed, which combines the key benefits of Data Lakes and Data Warehouses: from the former, comes low-cost storage to handle structured, semi-structured, and unstructured data, in an open format accessible by a variety of systems, and powerful management and optimization features from the latter (Armbrust, M et al. 2021). Using technologies, like cloud storage and scalable query engines, the Data Lakehouse promotes an endless transition, from exploring raw data, to obtaining dependable, controlled access for analytics and reporting purposes.

However, the current architecture presents issues in keeping up with the scale and scope of data and analytics use cases (Bode et al., n.d.). Issues like data complexity, silos, ownership, and bottlenecks in data management teams further hinder efficient data management and integration which leads to processing inefficiencies, delayed decision-making, and limited agility in meeting business requirements. So, what's the solution? One concept gaining momentum is a paradigm shift to a Data Mesh.

Figure 2: Comparison – Data Warehouse, Data Lake, and Data Lakhouse principles and architecture.

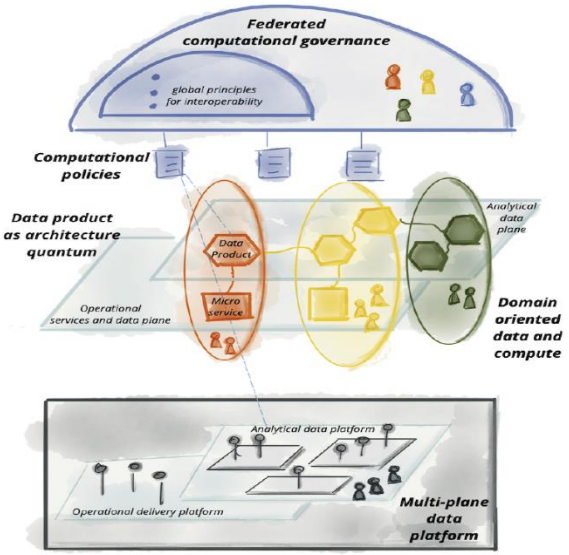


Note. Sourced from Sugumaran, H. (2021, p. 26).

The Data Mesh: Principles, logical architecture, and sociotechnical perspective

In 2019, Zhamak Dehghani introduced the Data Mesh architecture to address the current, common architectural issues (Voß, C. 2022). While Dehghani's work establishes a foundation for the Data Mesh, it is still in its early stages, with limited research and scientific contributions available. The Data Mesh (Figure 3) core premise is decentralizing data to specific domains, empowering those closest to the data to take responsibility for continuous change and scalability (Dehghani, 2020). To drive this transformation, Dehghani outlines four key principles: "1) domain-oriented decentralized data ownership and architecture, 2) data as a product, 3) self-serve data infrastructure as a platform, and 4) federated computational governance" (Dehghani, Z., 2020). These principles enable flexible scaling, provide high-quality and reliable data, optimize operating costs, and prevent data silos. As a result, the organization's data assets are treated as products created by domains, promoting efficient discovery and utilization by consumers.

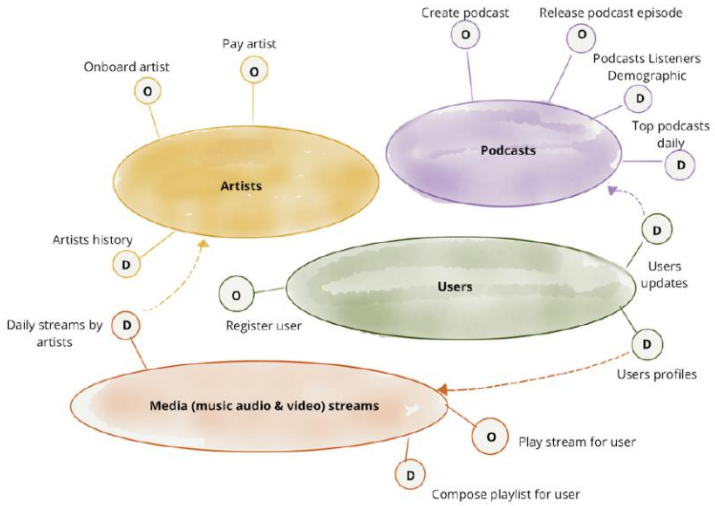
Figure 3: Logical architecture of the data mesh concept.



Note. Sourced from Dehghani, Z. (2020, p. 26).

The first principle, *domain-oriented decentralized data ownership* (Figure 4), distributes data to specific domains within the organization, granting responsibility and data ownership as close as possible to the people who know the systems, the terminology, and the data. This allows for efficient utilization aligned with the domains' objectives. The domain is viewed not only for its operational capabilities but for providing its domain's datasets/ data products in an easily consumable way (Dehghani, Z. 2019).

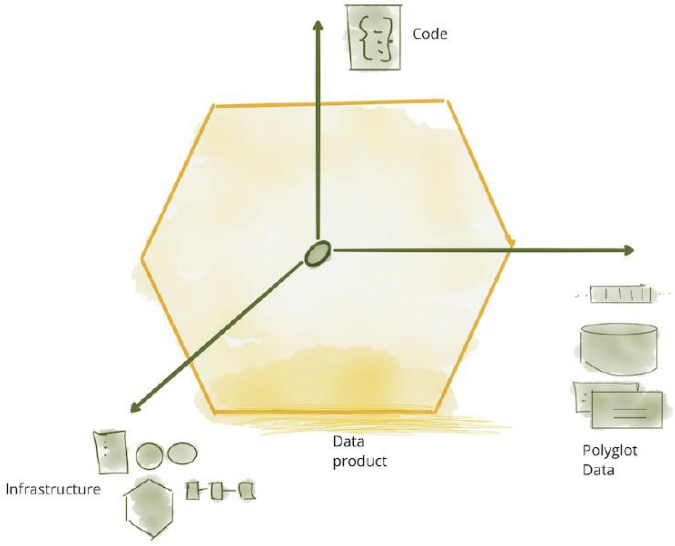
Figure 4: Example of domain-oriented ownership of analytical data and data products in addition to operational capabilities



Note. Sourced from Dehghani, Z. (2020, p. 10).

Secondly, *data as a product* adopts a product-thinking mindset for analytical data. Like operational domains offering APIs to enhance value and functionality, datasets from these domains should be treated as products within the data mesh. These products consist of three structural components (Figure 5): Code, Data and Metadata, and infrastructure (Dehghani, Z. 2020). By implementing a "data product catalog," other data and ML engineers, and data scientists, can access and utilize these products within the organization.

Figure 5: Data product components

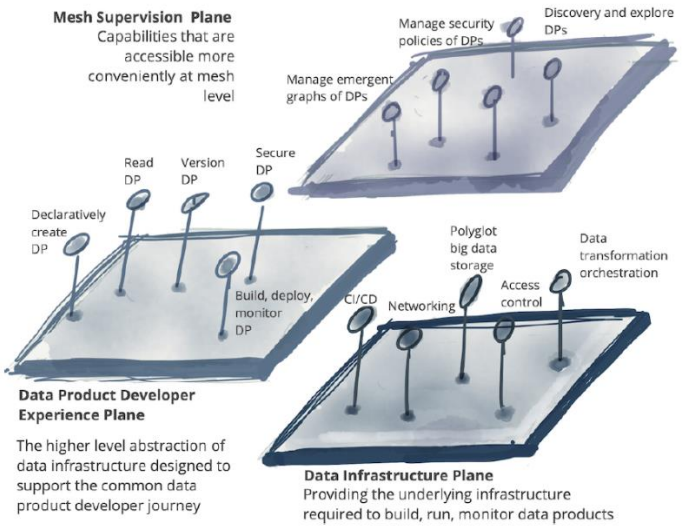


Note. Sourced from Dehghani, Z. (2020, p. 13).

The third principle, *self-serve data infrastructure as a platform*, enables domain developers to access platforms that provide tools for creating, maintaining, and running data products without requiring advanced specialized knowledge. This platform offers services like scalable polyglot data storage, data product schema, and data pipeline orchestration (Dehghani, Z. 2020). By adopting this principle, organizations can reduce costs and minimize

the need for specialized expertise in building data products. This involves different data platforms (Figure 6) catering to different needs for infrastructure provisioning, development experience, and data mesh supervision (Dehghani, Z. 2020).

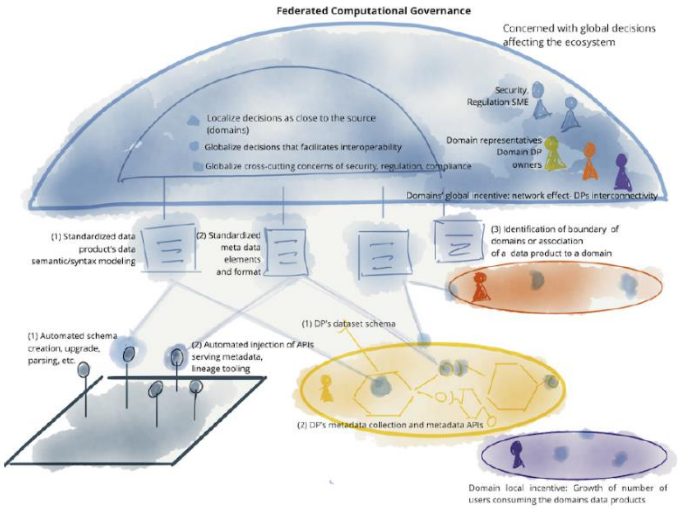
Figure 6: Multiple planes of a self-serve data platform.



Note. Sourced from Dehghani, Z. (2020, p. 18).

The fourth principle, *federated computational governance* (Figure 7), involves a federation of domain and platform owners who have decision-making power and follow a set of global rules. It requires the group to maintain a balance between global and domain-specific decisions. These rules ensure the interoperability of data products across domains, leading to efficient discovery, composition, and generating a network effect.

Figure 7: Notation – a federated computational governance model.



Note. Sourced from Dehghani, Z. (2020, p. 22).

	Data Warehouse	Data Lake	Data Lakehouse	Data Mesh
Architecture	Monolithic	Monolithic	Monolithic	Distributed
Data Type	Structured	Structured, semi-structured, and unstructured (raw)	Structured, semi-structured, and unstructured (raw)	Structured, semi-structured, and unstructured (raw)
Governance	Centralized ownership	Centralized ownership	Centralized ownership	Decentralized Domain ownership
Scalability	Limited	High	Moderate	Moderate
Processing	Batch	Batch and Real-Time	Batch and Real-Time	Real-Time
Technologically	Data is a by-product of code.	Data is a by-product of code.	Data is a by-product of code.	Data and code are a unit.
Order of operation	Top – Down	Top – Down	Top – Down	Federated computational governance
Principally	Data as an asset to collect	Data as an asset to collect	Data as an asset to collect	Data as a product to share

Table 1: Comparison of existing and upcoming architectures (Dehghani, Z. 2022, p. 9).

Data Mesh benefits

The following are some of the main benefits that an organization can have when adopting the data mesh architecture:

- Data ownership and leveraged domain knowledge: In current organizations, it can be "unclear whether certain data are owned by information processing units, data management units, or non-IS business units" (Winter, R., & Mornar, V. 2001). The data mesh architecture emphasizes data ownership, giving domains independence and responsibility for managing their own data. Data is also closer to domain experts, which improves quality, innovation, and efficiency while enabling trust in the data, data products, and decision-making.

- Data silo breakdowns and reduced data redundancies: Domains independently facilitate, maintain, and provide their data products to the organization, where consumers use a data catalog to find access, use, and/or modify existing products. This prevents duplication, and silos, and minimizes redundancy while encouraging data reuse.

- Better governance: Aside from the organization's global standards and policies, better governance also comes from a decentralized approach. By distributing governance responsibilities to domains, localized governance practices, and decision-making can occur with teams close to the data source. This domain-specific expertise ensures effective data governance within domains and across the

organization.

- Scalability and agility: With data no longer decentralized data in a data mesh, there are fewer dependencies on a central management platform (Goedegebuure et al., 2023). This enables domain teams to scale and make independent changes which allows faster experimentation, innovation, and agile responses to business needs, therefore reducing bottlenecks.

Challenges of a Data Mesh

Next, we discuss the challenges and concerns expressed regarding the implementation of the data mesh:

- Shift in responsibility: Implementing a data mesh shifts responsibility for data product accessibility and quality to the domains, which increases their workload without possible additional compensation. This may lead to deprioritizing these additional responsibilities over their current ones.

- Limitation of resources: Transitioning to the data mesh involves technical and structural changes, increasing costs and requiring additional skilled employees, software, and hardware. Limited finances could hinder acquiring additional required employees and necessary software and/ or hardware.

- Data products vs business products: Depending on the business, employees may prioritize revenue-generating products over creating data products. They may see data product creation as extra work that hampers performance, resulting in possible missed bonuses or pay increases and therefore deprioritizing data product creation.

- Comprehension: Lack of technical understanding of the data mesh concept can lead to confusion in terminologies, processes, and management which complicates the implementation and progression of a data mesh. Additionally, there are concerns regarding domain comprehension of protected and regulated data for federated data governance: lacking unawareness of compliance with protection regulations can have serious consequences (Goedegebuure et al., 2023).

- Resistance to change: Transitioning to a data mesh in established organizations brings significant technical and structural changes, adding additional responsibilities and disrupting current day-to-day tasks. This can potentially cause employee resistance due to unhappiness with the changes.

Role of a Data Scientist in a Data Mesh Architecture.

This section explores ideas, issues, and perceptions surrounding data scientists within an organization implementing a data mesh architecture. In transitioning to a data mesh, each domain now has its own team with possible data engineers, analysts, and data scientists

closely collaborating with individuals with comprehensive knowledge of the systems, terminology, and data within their domains. This requires that data scientists understand their respective domains and align their analytical and modeling efforts with the domain's priorities and goals. Additionally, data scientists are required to acquire knowledge in data engineering, data governance, and data operations to effectively cater to the end-to-end data requirements of their domains.

Data mesh architecture promotes domain data ownership and democratization. As a data scientist, I assist the team in leveraging their data for decision-making and innovation. This involves providing the necessary tools and procedures to effectively work with the data. Further establishment of data quality metrics within the domain, monitoring processes, and implementing feedback loops for continuous improvement resolves concerns regarding data quality and reliability organization-wide. Open communication with other domain teams is vital to identify quality issues, applying cleansing and validation techniques, and ensure reliable data and interoperability. In the data mesh, data acts as a product used by other domains, which requires high-quality and trustworthy data products.

In addition to mainly working within a specific domain, there is some cross-functional collaboration between other domains (example mentioned above). This includes data as a product. In conducting data exploration, cleaning, statistical analysis, ML, and data visualization, a data mesh allows for more efficient execution by accessing other domains' data products to use for similar tasks, therefore reducing redundancies. For example, instead of creating my own pipeline to process and transform raw data into a more usable format (a time-consuming process), utilizing another domain's data pipeline, found in the organization's catalog, creates more efficiency. If this data product doesn't exist, I would proceed to create my own to use on my domain's data which would then be available as a data product for other domains to use. When creating data products, finding a balance between domain-specific needs and organizational data standards is a priority. This may involve collaborating with other data scientists and data platform teams, ensuring seamless integration of these products within the broader data mesh, by incorporating common standards, governance frameworks, and best practices.

Furthermore, collaboration with data privacy and cybersecurity teams is crucial to implementing access controls, encryption methods, and data de-identification techniques for distributed data products across domains. Within the domain, I would also promote privacy best practices and ensure regulatory compliance. Data mesh architecture and the role of data scientists are evolving areas, and the suggested insights provided above are based on literature reviews, with real-world organizational implementations possibly varying.

III. Conclusion

From 2022 to 2026, data is expected to double, leading to exponential growth in volume, velocity, and variety. This report discusses challenges faced by organizations with monolithic architectures and centralized data management teams, such as data complexity, scalability, data silos, ownership, and bottlenecks. These challenges hinder efficient data management and integration, resulting in processing inefficiencies, delayed decision-making, and limited agility in meeting business requirements.

To address these issues, the Data Mesh architecture is introduced as a solution. It shifts from a centralized architecture to a decentralized approach, distributing data to respective domains. This brings data closer to experts, empowering them to take responsibility for continuous change and scalability. The four principles driving this transformation are domain-oriented decentralized data ownership and architecture, data as a product, self-serve data infrastructure as a platform, and federated computational governance of the data mesh.

Implementing a data mesh offers benefits such as data ownership, leveraged domain knowledge, breakdown of data silos, reduced redundancies, better governance, scalability, and agility. However, there are concerns about resistance to change, limited resources, shift in responsibility, integration of data products, and comprehension of the concept. From the acquired knowledge through literature reviews, we have theories ideas, issues, and perceptions surrounding data scientists working in an organization having implemented a data mesh architecture. This report aims to contribute to the discussion of the Data Mesh and its advancements in data systems, with hopes of global implementation in the future.

References

Agrawal, S. A. (2022, May 19). *Don't decide on a data architecture until you read this!* Infocepts Data & AI. Retrieved June 16, 2023, from https://www.infocepts.com/blog/dont-decide-on-a-data-architecture-until-you-read-this/

Amazon Web Services, Inc. (2023) *What's The Difference Between A Data Warehouse, Data Lake, And Data Mart?* (2023). Amazon Web Services, Inc. Retrieved June 14, 2023, from https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/#:~:text=It%20is%20a%20central%20repository,raw%20data%20and%20unstructured%20data.

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. *In Proceedings of CIDR* (p. 8). Retrieved June 13, 2023, from https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

Bassman, A. (2023, February 2). *5 misconceptions about cloud data warehouses*. IBM Blog. Retrieved June 15, 2023, from https://www.ibm.com/blog/5-misconceptions-about-cloud-data-warehouses/

Bode, J., Kühl, N., Kreuzberger, D., Hirschl, G., & Carsten Holtmann, G. (2023, April 5). *Data Mesh: Best Practices to Avoid the Data Mess*. Retrieved June 12, 2023, from https://arxiv.org/pdf/2302.01713

Dehghani, Z. (2019). *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*. (n.d.). Martinfowler.com. Retrieved June 12, 2023, from https://martinfowler.com/articles/data-monolith-to-mesh.html

Dehghani, Z. (2020). *Data Mesh Principles and Logical Architecture*. Martinfowler.com. Retrieved June 12, 2023, from https://martinfowler.com/articles/data-mesh-principles.html

Dehghani, Z. (2022). *Data Mesh: Delivering Data-driven Value at Scale*. Japan: O'Reilly Media. Retrieved June 9, 2023, from https://biconsult.ru/files/Data_warehouse/Data%20Mesh%20Delivering%20Data-Driven%20Value%20at%20Scale.pdf

Goedegebuure, A., Kumara, I., Driessen, S., Van, W.-J., Heuvel, D., Monsieur, G., Tamburri, D., Van Den Heuvel, W.-J., & Di, D. (2023). Data Mesh: A Systematic Gray Literature Review. *Data Mesh: A Systematic Gray Literature Review*, *1*(1), 29. Retrieved June 12, 2023, from https://arxiv.org/pdf/2304.01062

InfoQ (2019). *Data Mesh Paradigm Shift in Data Platform Architecture*. [Video]. YouTube. Retrieved May 27, 2023, from https://www.youtube.com/watch?v=52MCFe4v0UU

Sugumaran, H. (2021, January 9) *Data Lake House vs. Data Warehouse: Tips to pick the Right Solution for Your Stack*. Www.linkedin.com. Retrieved June 16, 2023, from https://www.linkedin.com/pulse/data-lake-house-vs-warehouse-tips-pick-right-solution-sugumaran/

SAS Institute Inc (2017). *Data Warehouse*. *What is it and why it matters.* Sas.com. Retrieved June 14, 2023 from https://www.sas.com/en_au/insights/data-management/data-warehouse.html

Thoughtworks. (2020). *Keynote - Data Mesh by Zhamak Dehghani.* [Video] YouTube. Retrieved May 27, 2023, from https://www.youtube.com/watch?v=L_-fHo0ZkAo

Voß, C., (2022, September 29-30). Identifying Alternatives and Deciding Factors for a Data Mesh Architecture. In:, . (Hrsg.), SKILL 2022. *Gesellschaft für Informatik, Bonn.* (S. 93-99). [Seminar] Retrieved June 13, 2023, from https://dl.gi.de/items/324d48dd-373f-48e9-a602-cb82ffdfe469/full

Winter, R., & Mornar, V. (2001). *Organization of Data Warehousing in Large Service Companies: A Matrix Approach Based on Data Ownership and Competence Centers.* [Conference session] AMCIS 2001 Proceedings, 65. Retrieved June 14, 2023, from https://aisel.aisnet.org/amcis2001/65/?utm_source=aisel.aisnet.org%2Famcis2001%2F65&utm_medium=PDF&utm_campaign=PDFCoverPages

Assessment Declaration

By submitting this piece of assessment electronically, I declare that:

- This assignment is my original work and no part has been copied/reproduced from any other person's work or from any other source, except where acknowledgement has been made.

- I have clearly declared any sections in the literature review where a generative AI program was used and the type used (e.g GPT3, Chat GPT).

• This work has not been submitted previously for assessment and received a grade, nor concurrently for assessment, either in whole or part, for this subject (unless part of integrated assessment design/approved by the Subject Coordinator), any other subject or any other course.

X 

Justin Grima