Assessment 3: Data Processing Trends

MA5831 – Advanced Data Management and Analysis using SAS.

Name: Rijo Kuruvilla

Student ID: 14171011

I. Introduction

Big Data’s evolution revolutionized how organizations collect, store, and analyze data. Technological advancements in big data infrastructure, services, and analytics enable businesses to transform into data-driven enterprises (Lee, 2017, p. 1). The vast volumes of data from different sources created the need for efficient and scalable facilities that provide quick results for analytical purposes. Previous architectures like Data Warehouse and Data Lakes, although still popular, possess certain limitations that diminish their effectiveness when it comes to handling Big Data. Both data warehouses and data lakes follow a centralized, monolithic, and domain-agnostic structure differing in their approach to data storage and management (Dehghani, 2019).

Data warehouses (First generation architectures) were initially designed to centralize structured data from various operating systems providing a unified view for reporting and analysis (Dehghani, 2022, p.40), but often struggled in accommodating massive scale of data directly leading to bottlenecks and huge costs in storing diverse data types and schemas. Data lakes and data lake houses, which are less structured than Data warehouses, combated the issue by providing flexibility in data types and schemas, acting as a staging area for data warehouses, a platform for experimentation for data scientists and analysts, and a direct source for answering ad-hoc queries. However, data lakes posed issues in terms of data governance, discoverability, and quality control (Khine & Wang, 2018, as cited in Machado, 2021).

To tackle the limitations of Data warehouses and Data Lakes, Data Mesh architecture was introduced. Data Mesh architecture signals a paradigm shift in sourcing, managing, and accessing large amounts of data at an analytical scale (Dehghani, 2022, p.1). Data mesh gives teams more power by favouring a decentralized domain-oriented approach to data management that empowers domain teams to take ownership of their data, promoting a self-service data infrastructure, and embracing a data-as-a-product mindset. This architecture also promotes cross-functional collaboration, autonomy, and accountability among domain experts, data engineers, and data scientists. Each domain team oversees all aspects of its domain which includes the generation, utilization, and quality control of the data.

Data mesh also disperses data ownership and governance, promoting a more adaptable and flexible data ecosystem across an organization's domains, allowing them to effectively scale their data infrastructure, adapt to dynamic requirements, and use their data more effectively for better data-driven solutions.

II. Principles defining the Data Mesh Architecture

The primary objective of data mesh architecture is to create a base for getting insights and values from the constant change of data landscapes, diverse transformation and processing of each undertaken use case, and the response speed to the changes applied (Dehghani, 2020). There are four primary principles that define the data mesh architecture:

- *Decentralized Domain Ownership*: Domain-driven design introduces the idea of breaking down monolithic systems into chunks that are designed around the domain. This method enables the development of software and infrastructure based on the underlying domain models. Combining operational and analytical data from these business domains, domain experts and developers create data products. Building and maintaining data products and carrying out domain-level governance tasks are among the main duties of domain teams (Goedegebuure et al., 2023). Possessing decentralized domain ownership enables domain teams to determine which domain data assets should be provisioned as data products.

- *Data as a product*: A data product is an independent logical component comprising all the required data, code, and interfaces to serve the domain team’s analytical data needs (INNOQ, n.d.). Data products are connected to a variety of sources such as operational systems or other data products on which various transformations are applied to produce the results expected by customers. Five elements form a data product: data, metadata (data defining owners, schemas, metrics, and policies), product code (responsible for ingesting and transforming data from source systems and delivering the transformed data to the customers or other data products), interfaces (which enables communication with other data sources, products, and applications) and infrastructure (which consists of resources and platforms to consume, store, execute data operations and make the data and its capabilities accessible via interfaces) (Goedegebuure et al., 2023).

- *Self-serve data infrastructure platform*: Data Mesh promotes building infrastructure and platform services around domain needs to empower teams to build and manage their product cycles, to share the mesh's emergent knowledge details with other domains, and to streamline data consumers' experiences in discovering, accessing, and using data products. A centralized data platform team of highly specialized infrastructure and platform service developers builds and maintains the platform (Goedegebuure et al., 2023). Some of the key services offered by the self-serve platform are offering computing resources, networking capabilities, polyglot storage (storage for diverse types of data), metadata repository, data monitoring, security and privacy, and policy reinforcement (Goedegebuure et al., 2023).

- *Federated Governance*: The federal governance group comprises representatives from all teams taking part in the data mesh. They create rules on how domain teams should build their data products and concur global policies for the mesh architecture (Li et al., 2022). Interoperability policies function as the starting point, which allows domain teams to use a product in a consistent way. The governance team also decides on the document type to discover and understand the data products available and to access the data products in a secure way. Additionally, they oversee developing incentive-based models to encourage domain teams, whose typical tasks might not include providing data products (to provide data as products), and promptly regulate the provided products.

III. Comparison between Data Mesh and other architectures

Table 1: Difference between Data Warehouse, Data Lakes, Data Lake House, and Data Mesh

Parameters	Data Warehouse	Data Lakes	Data Lakehouse	Data Mesh
Type of Data	Structured	Raw, Semi-Structured, Unstructured	Combination of both Data warehouse and Data lakes	Structured, Raw, Semi-structured and Unstructured
Governance	Centralized	Centralized or decentralized depending on the organization	Centralized	Decentralized-Domain Oriented
Architecture	Monolithic	Monolithic	Monolithic	Distributed
Technicality	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 one unit
Operationality	Top-Down	Top-Down	Top-Down	Federated computational governance
Transaction Cost	High	Low	Moderate	Varies
Processing	Batch	Real-Time and Batch	Real-Time and Batch	Real-Time
Scalability	Limited Scalability	Highly Scalable	Scalable	Scalable
Socio-technical perspective	A specialized team manages data governance and ownership through centralized architecture.	A centralized repository for data exploration and storage, but ownership and control are organization specific.	Balances the decentralized nature of the data mesh with the centralized qualities of the data warehouse	Focuses on a decentralized approach with domain teams being responsible for data ownership, encouraging autonomy and collaboration.

Note: The table above displays the comparison of Data Warehouses, Data Lakes, Data Lake houses, and Data Mesh. Adapted from "*Data Mesh: Data-driven value at scale*. O'Reilly Media, by Dehghani, Z., 2022 https://biconsult.ru/files/Data_warehouse/Data%20Mesh%20Delivering%20Data-Driven%20Value%20at%20Scale.pdf

IV. Benefits and Challenges of Implementing a Data Warehouse

Benefits:

- *Improved Data Accessibility*: Data warehouses provide a centralized and unified view of data from multiple domains, enhancing data accessibility for data science and analytics teams without going through complex data integration processes.
- *Data Consistency and Quality*: Better integration options including data governance and data wrangling directly improve data consistency and quality. By enforcing standardized data practices and conducting regular data checks during data integration, organizations can ensure the data's reliability and accuracy.
- *Improved Data Governance*: By judiciously distributing IT teams across domains with independent controls over all data-related activities, the data mesh architecture addresses Data Governance bottlenecks without sacrificing scale (Ghosh, 2023).
- *Improved Scalability and Flexibility*: Data mesh's distributed nature helps to easily scale up or down depending on demand without compromising performance or reliability (Ramos, 2022).
- *Self-manage or Self-serve*: The individual domain in a data mesh develops the capacity to self-manage and serve on all facets of its data processing and data science projects (Ghosh, 2023).

Challenges:

- *Cultural change and organizational pushback*: The adoption of data mesh enforces many changes within the organization resulting in resistance from users, especially in smaller organizations which may strain staffing and training resources (PwC, n.d.). Additionally, organizations have been hesitant to adopt the self-service analytics aspect of Data mesh as they already have dedicated teams set up for data analysis and data science.
- *Data security and privacy concerns*: Integrating data from various domains may expose organizations to various security and privacy threats. Proper security measures must be set up to protect sensitive data.
- *Setting up a holistic data governance system and standardization*: Decentralised data engineering, federated governance, and platform components are key principles of Data Mesh. Enforcing decentralized data governance is difficult because data products coexist independently, increasing the possibility of governance inconsistencies between domains (PwC, n.d.).
- *High Cost of implementation*: Depending on the size of your organization's infrastructure, installing a data mesh system can be expensive, as it typically calls for sizable investments in hardware and software resources, which may not be feasible for all budgets (PwC, n.d.).

V. Business Scenario

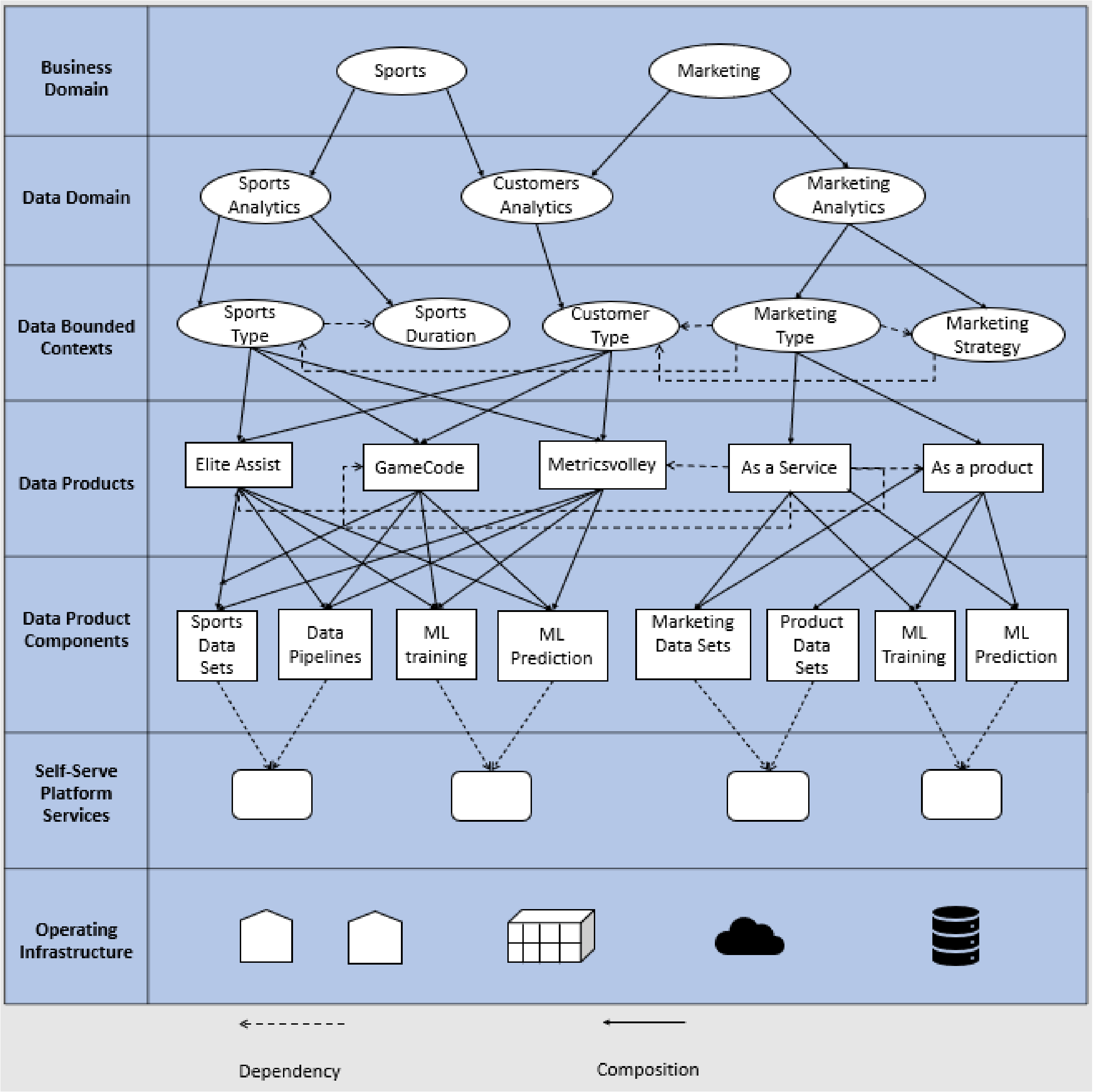
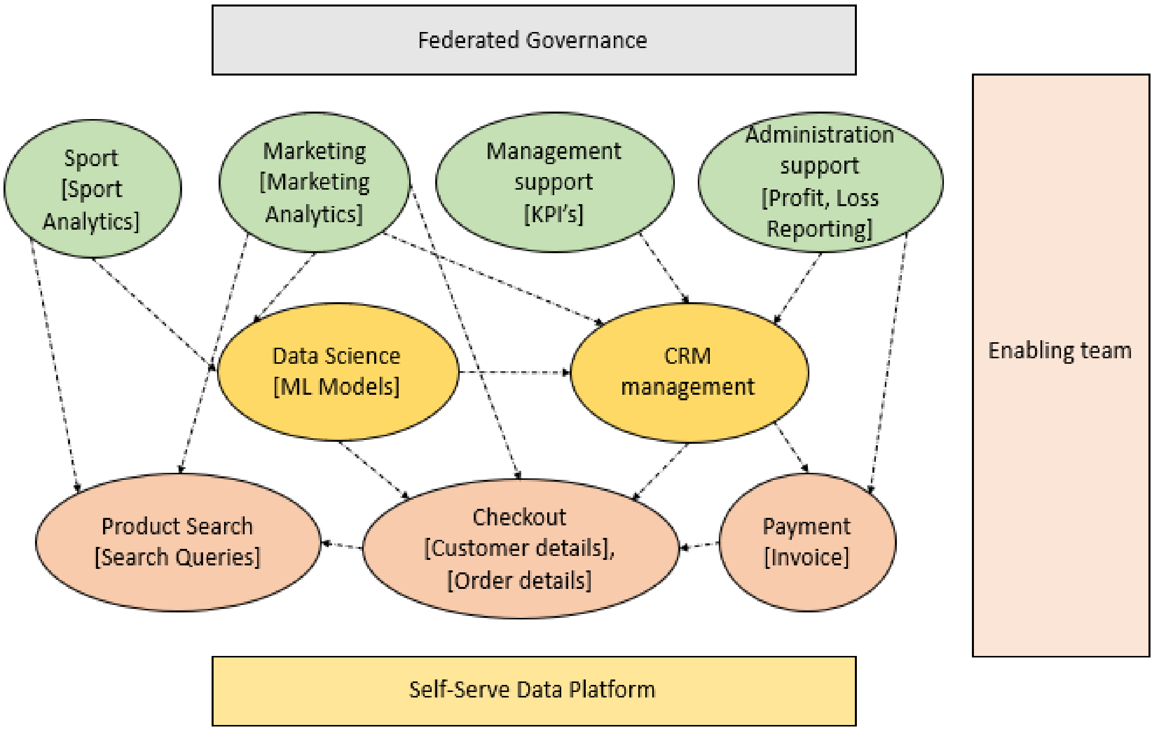


Figure 1: A layered model for data mesh development for a Sports Analytics organization

Platform-based development and domain-driven design are the key focuses of data mesh architecture. Figure 1 displays a snapshot of the layered model architecture for data mesh development for a sports organization that primarily focuses on delivering analytics and event-based insights to its consumers. The primary consumers for this organization are schools, universities, and professional sports clubs. Sports and marketing are the primary business domains (First Layer) under discussion. The Second layer (Data Domain Layer) consists of the analytical data for various sports (Soccer, Football, Basketball, Volleyball, etc.), and clients that the organization serves can be derived utilizing these business domains. Analytical data for marketing is also extracted in this manner. The Third Layer presents some of the various contexts associated with the data domain, such as sports type, game duration, customer type, marketing type, and marketing strategy. The associated data products with each bounded context are identified in the fourth layer.

Data products are created utilizing tactical design patterns, which may be helpful to improve a domain's fundamental model of constrained context, from which architecture components for applications can be derived. Data products for this organization for the sports domain involve specific software developed for each kind of sport, whereas marketing involves using the products as either a service (sports analysis, player analysis) or as individual products (sports gears, motion sensors). The components responsible for the implementation of the functional and non-functional requirements of data products are displayed in the fifth layer. These include data sets from sporting and marketing events, data pipelines, and machine learning training and prediction services. The architecture used changes depending on the data product. The Self-serve data platform (sixth layer) empowers teams with autonomy to manage their domains (Machado et al., 2022) with available tools from the platform to create, monitor, discover, analyze, and maintain the data product's capabilities without possessing the advanced expertise usually required to operate such tools. The operating infrastructure (Seventh Layer) facilitates the use of data products and can be cloud-based, local, or a hybrid of the two.

Figure 2: Working of a Data Mesh Model in a Sports Analytics Organization


Federated Governance
Sport [Sport Analytics]
Marketing [Marketing Analytics]
Management support [KPI's]
Administration support [Profit, Loss Reporting]
Data Science [ML Models]
CRM management
Product Search [Search Queries]
Checkout [Customer details], [Order details]
Payment [Invoice]
Enabling team
Self-Serve Data Platform

Note: Adapted from Data Mesh Architecture, [Working of a data mesh], INNOQ. https://www.datamesh-architecture.com/

Figure 2 displays a simplified example of the working of a data mesh in a sports analytics organization. The federated governance team in this sports organization mainly consists of domain and platform owners who have decision-making power and follow a set of global rules. This includes, but is not limited to, coaches, and analysts that will help teams build data products, varying from sport to sport.

The mesh unfolds when teams use other domains' data products. Data products are either source-aligned, aggregated, or consumer-aligned. Source-aligned data products represent the data from operational systems which have undergone minimal transformations (Mott, 2022). In this case, the sports online store is split down into smaller parts to track the customer's journey from product search to a product purchase. Software teams using the self-serve data platform can use their own data analysis to confirm and improve the value of new features, as well as the data from other domains to streamline their queries and gain an understanding of the effects of independent domains.

Consumer-oriented data products are aimed at departments that make strategic decisions. Employees in these departments are mostly business experts who analyze reports and key performance indicators (KPIs) from various domains within an organization to determine their strengths and weaknesses. The data model for these domains is optimized for the needs of a specific department, and such consumer-aligned teams can feed from other domains' data products to gather domain knowledge, resulting in better analytical reports. The sport domain team can feed from data generated from the online store (here, the data product being the product store) by accessing product search queries for various software purchases made by different customers and use this data to improve their software or generate updates for existing software. Similarly, the marketing team can improve their marketing strategies based on the product popularity generated by each product or piece of software sold. The management team can create KPIs using visual analytics to track the overall performance of the various domain teams, whereas the administration team can track the company's financial performance.

Aggregated data products are usually generated for domain teams where significant mathematics, statistics, and technical expertise are required (Skelton & Pais, 2019). Data Science teams can use data generated from the online stores and the CRM (customer relationship management) domain team to generate machine learning and recommendation models. The CRM team can also utilize the data from the online store to generate detailed reports tracking customer behaviour. Crucial metrics include the number of products purchased, the number and types of products searched, the number of click events, etc, all of which can be aggregated and analyzed to solve several business problems.

VI. Conclusion

Data Mesh architecture represents a paradigm shift in how organizations approach data management and analytics by focusing on decentralized domain-oriented teams and invoking a data-as-a-product approach. This architecture's primary benefits include the ability to break down data silos, improve data-driven decision-making in various domains present within the organization by promoting a collaborative approach and provide a more scalable architecture. Sports organizations can take advantage of the benefits of data mesh architecture by integrating data and analysis from various sources, such as player and team performance, and marketing analytics, profit/loss analysis to make critical data-driven decisions, providing customers with a view of performance optimization and driving revenue growth. As a relatively new architecture, there are some limitations and concerns that come with implementing a data mesh in terms of federated data governance, data security, and privacy issues while integrating data from various domains, resisting cultural change, and the possibility of huge costs. However, if considerable precautions are taken, a data mesh can be a boon to organizations in terms of data management and can help boost their productivity and revenue.

Number of words: 2198

References

[1] Dehghani, Z. (2019, May 20). How to move beyond a monolithic data lake to a distributed data mesh. martinfowler.com. https://martinfowler.com/articles/data-monolith-to-mesh.html

[2] Dehghani Z. (2020, February 27). *Data mesh paradigm shift in data platform architecture* [Video]. InfoQ. https://www.infoq.com/presentations/data-mesh-paradigm/

[3] Dehghani, Z. (2022). *Data mesh: Delivering data-driven value at scale*. O'Reilly Media. https://biconsult.ru/files/Data_warehouse/Data%20Mesh%20Delivering%20Data-Driven%20Value%20at%20Scale.pdf

[4] Ghosh, P. (2023, January 26). *Data mesh architecture benefits and challenges*. DATAVERSITY. https://www.dataversity.net/data-mesh-architecture-benefits-and-challenges/

[5] Goedegebuure, A., Kumara, I., Drissen, S., Van Den Heuvel W., Monsieur G., Tamburri D., & Di Nucci D. (2023). *Data Mesh: a Systematic* Gray Literature Review [Master's thesis]. https://arxiv.org/pdf/2304.01062.pdf

[6] INNOQ. (n.d.). *Data mesh architecture*. Data Mesh Architecture. https://www.datamesh-architecture.com/

[7] Lee I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*(3), 293-303. https://www.sciencedirect.com/science/article/pii/S0007681317300046

[8] Li, J., Cai, S., Wang, L., Li, M., Li, J., & Tu, H. (2022). *A novel design for data processing framework of park-level power system with data mesh concept*. IEEE Xplore. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8048917

[9] Machado, i. A. (2021). *Proposal of an Approach for the Design and Implementation of a Data Mesh* [Master's thesis]. https://repositorium.sdum.uminho.pt/bitstream/1822/82290/1/Ines%20Araujo%20Machado.pdf

[10] Machado I., Costa C., & Santos, M. (2022). *Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures* [Master's thesis]. https://www.sciencedirect.com/science/article/pii/S1877050921022365

[11] Mott A. (2022, December 16). *What are the different types of data products*. Starburst. https://www.starburst.io/blog/what-are-the-different-types-of-data-products/

[12] PwC. (n.d.). *Pros and cons of data mesh*. https://www.pwc.ch/en/insights/data-analytics/data-mesh-challenges.html

[13] Ramos, E. (2022, December 26). *Exploring The Benefits Of A Data Mesh For Your Organization*.
Forbes. https://www.forbes.com/sites/forbestechcouncil/2022/12/26/exploring-the-benefits-of-a-data-mesh-for-your-organization/?sh=3daa03e3159f

[14] Skelton M., & Pais M. (2019, July 3). *Key concepts — Team Topologies*. Team Topologies. https://teamtopologies.com/key-concepts