LTIMindtree

Data Mesh & its Application in Data Analytics

Author:
Chaithanya M I Ramanath K P

LTIMindtree

Contents

Introduction 3

Principles 3

- Domain-driven Data Ownership and Architecture 4
- Data as a Product 4
- Self-serve Data Platform 4
- Federated Computational Governance 4

Challenges in Today's Data Ecosystem 5

Problems with Centralized Data Ownership 6

Benefits of Data Mesh in Data Management 7

Architecture of Data Mesh Approach 8

Implementing a Practical Data Mesh 9

Data Mesh Pattern: Enterprise Data Product Catalog 10

Conclusion 11

About the author 12

Introduction

Why Data Mesh

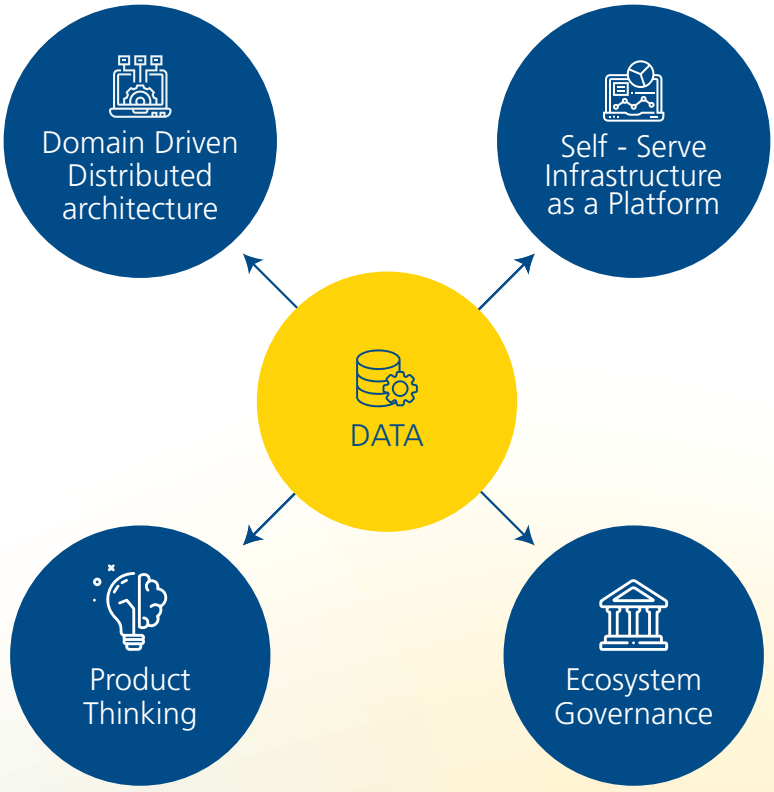
In this new paradigm, data is a strategic business asset. Almost every enterprise in the self-service era describes itself as data-first. However, not all companies equip their data architectures with necessary democratization and scalability.

The rampant lack of democratization and scalability of big data means current data architectures (typically comprised of data warehouses or data lakes) may not be able to keep up with the growing number of new data sources and support diverse use cases. To solve this, a new approach – data mesh is required to remain agile in large and complex environments by enabling quick and cost-effective value extraction from data.

At its core, a data mesh is about coordinating people, processes, and organizations, not just technology.

Principles

Data Mesh Principles



Domain-Driven Data Ownership and Architecture

A domain is a group of stakeholders organized to meet a shared business motive. Data mesh ensures the domain is accountable for managing data that is related to or is produced by the domain's commercial enterprise characteristic. These domains assimilate, transform, and provide data to end-users as data products whose lifecycles are entirely owned by that domain.

Data-as-a-Product

The mesh applies the concept of "product thinking" to the data, making it a top priority for organizations and eliminating data pipelining and storage concerns. Analytical data is seen as a product and consumers of this data are seen as customers, and their needs must be fulfilled.

Self-Serve Data Platform

A self-serve data infrastructure consists of several functionalities that domain members can leverage to create and manage their data products. Such a platform is supported by an engineering team that focuses on the management and operation of various technologies.

Federated Computational Governance

Traditional data governance mode can impede the creation of value from data. To remediate this issue, a data mesh approach incorporates governance issues into the domain's workflow. There are several facets of data governance that must be measured and reported in a data mesh. This includes tracking how much data is used and how it is used to understand the value and determine the success of individual data products.

Challenges in today's data ecosystem

In the last 30 years, the evolution of data, and its management, can be broadly classified into two categories: operational data and analytical data. However, an unfortunate problem with "core data management " is that it is seen as a technical problem and ignores the domain organization of the data. Domains reappear in reporting, analysis, and data science, while data moves from source to analysis. Through the many stages of transformation and preservation, it creates cracks in lineage management, resilience to change, and other issues.

Problems with Centralized Data Ownership

- In a centralized ownership model, data needs to be imported and transferred to a central data lake – a time-consuming and expensive process. On the other hand, in a data mesh's distributed data, data is viewed as a product with separate domain ownership for each business entity. As a result, this decentralized model accelerates time to value and empowers teams with discoverable data.
- As data grows, queries become more complex, requiring changes across data pipelines that don't scale, reducing response time and agility across teams. By delegating ownership of records to domains (individual teams or business users) a data mesh architecture promotes agility and scalability, facilitating real-time decision-making in enterprises.

Data Mesh

Data mesh may be taken into consideration in the implementation of an architecture where data is deliberately distributed among numerous mesh nodes. This ensures there is no chaos or siloed records, considering there are centralized governance techniques and guarantees the sharing of middle concepts at some stage in the mesh nodes.

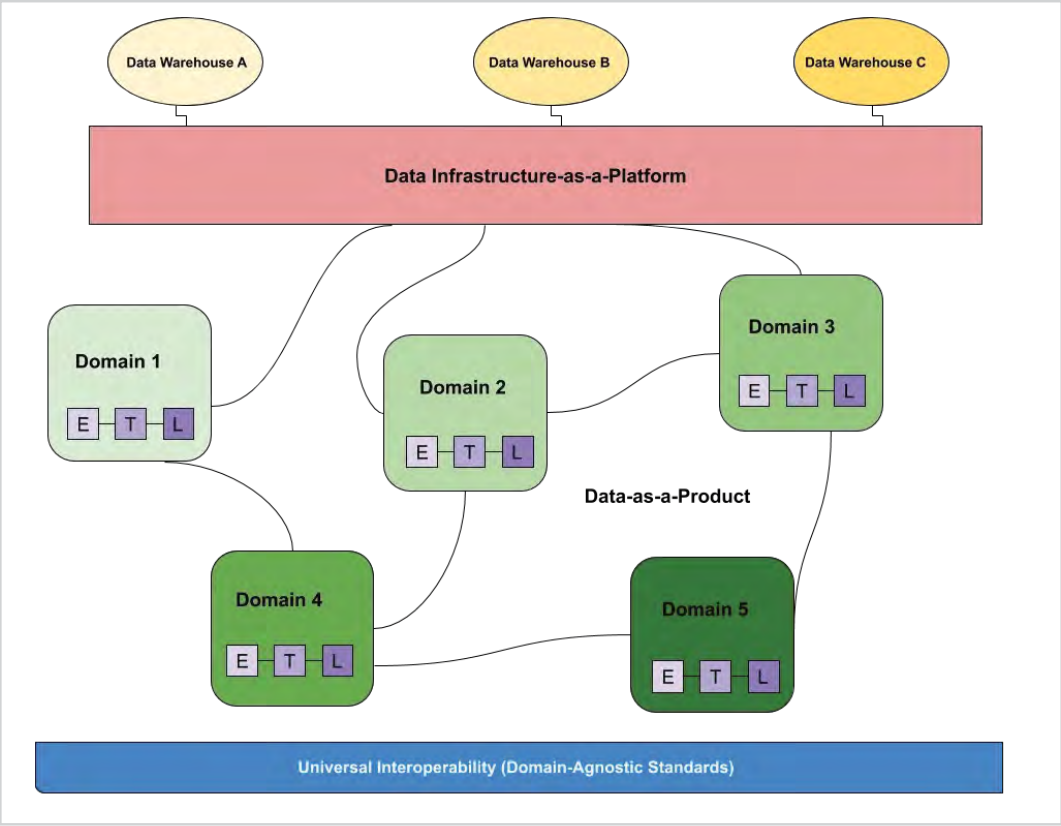
Data mesh emerges as an important paradigm shift that allows organizations to become information-oriented, enforcing a structure that brings the other of the modern models for efficient facts product cooperation. From a more structural perspective, records are prepared into domains and information teams manipulate themselves and perform their work in an agile and product-oriented way. This paradigm shift does not arise at simply a structural level, but also at an organizational level, since it enables organizations to focus on a specific domain and, therefore, become decentralized.

In a data mesh architecture, facts from specific locations can now be linked inside the mesh. As a result, it supports complex control, promotes entry, and assists components through the connectivity layer it implements.

Dehghani's data mesh concept proposes a divide-and-conquer method for the delivery of data and analytic merchandise by aligning implementation with domain-driven design principles and patterns. Such an approach involves three strategies to successfully deliver data and analytic products faster in large and complex organizations:

- Intelligent use of decomposition for parallel creation of large and complex products by multiple, loosely coupled development groups.
- The use of agile, business-led development methods and teams to eliminate unnecessary tasks.
- The automation of the necessary development and testing work.

Benefits of Data Mesh in Data Management



Agility and Scalability

The data mesh supports distributed data operations to improve time-to-market, scalability, and business domain agility.

Flexibility and Independence

Companies that adopt a data mesh architecture avoid being locked into a data platform or product.

Fast Access to Critical Data

Data mesh provides easy access to centralized infrastructure with a self-service model that enables faster data access and SQL queries.

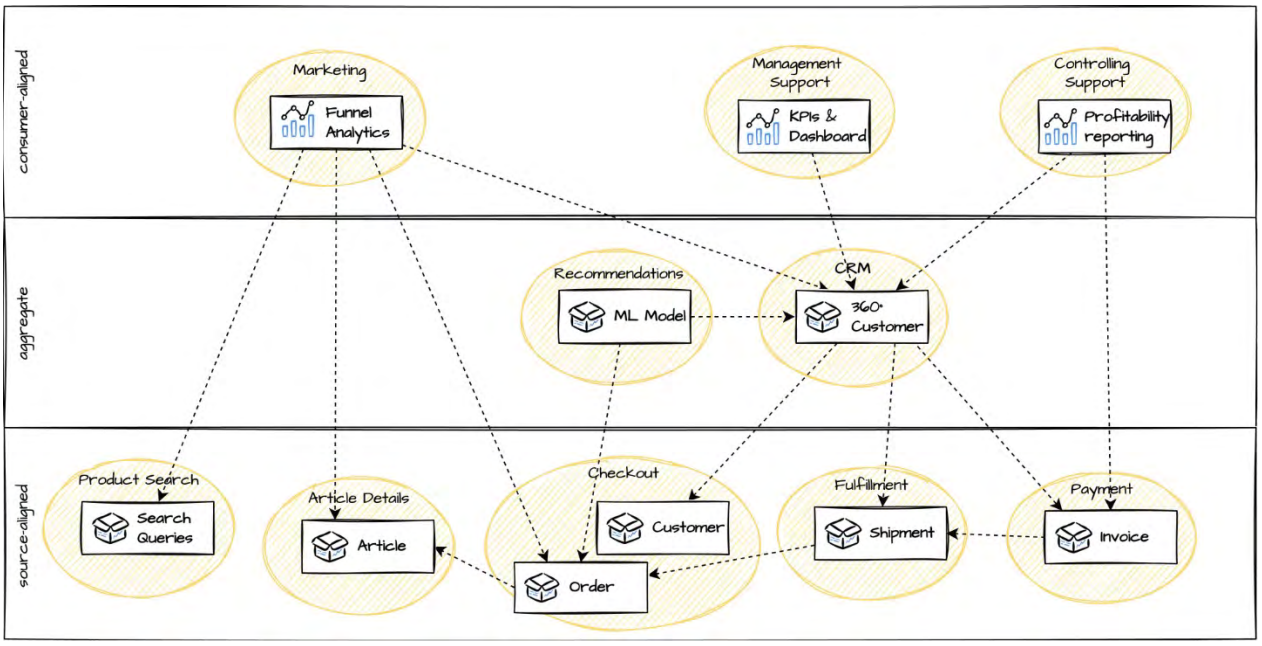
Visibility for Cross-Functional Teams

Centralizing data ownership in traditional data platforms leaves expert data teams isolated, highly dependent, and resulting in a loss of visibility. Data mesh decentralizes data ownership and distributes it to cross-functional domain teams.

Architecture of a Data Mesh Approach

A data mesh structure is a decentralized technology that allows domain teams to own domain records evaluations where each domain is responsible for its operational and analytical statistics. The domain teams ingest operational information and build analytical statistic trends to carry out their own evaluations. It uses analytical information to construct statistical products based on different domain requirements.

The mesh emerges when teams use other domains' data products. The usage of statistics from upstream domains simplifies data references and lookups (besides getting an article's rate), at the same time as statistics from downstream domains enable reading consequences, e.g., A/B checks. Data from more than one other domain name can be aggregated to build complete reports and new data products. Let us consider a simplified e-commerce example:



Source-aligned

In this situation, an e-commerce operation is divided into domains alongside consumer journeys. In a data mesh, each domain presents its facts as data products, allowing others to access them. Engineers then analyze their own data to enhance operational structures and validate new features. Using their neighbor's records, domains can simplify their queries and gain insights into downstream impacts.

Aggregate

In a complex subsystem framework, an efficient approach is to have a dedicated team focusing solely on aggregating and distributing data products from other domains. Building sophisticated ML (machine learning) models that require improved, enhanced data science competencies is a typical example of a complicated subsystem. In such a scenario, data scientists can develop and train an advice model by using checkout records alongside a 360° view of the customer, while another group uses this model to provide the calculated guidelines within the online store or across promotional content.

Consumer-aligned

In a business enterprise, there are departments that need statistics from the complete value circulation to make realistic selections. This calls for detailed reports and KPIs gathered from across domains to detect strengths and weaknesses. The marketing department analyzes the funnel and net evaluation throughout the customer journey using tools, such as Google Analytics or Adobe Analytics. As such, the data model is optimized for a selected business unit's desires and aligned with the customer.

Implementing A Practical Data Mesh

Zalando's Data Mesh Implementation

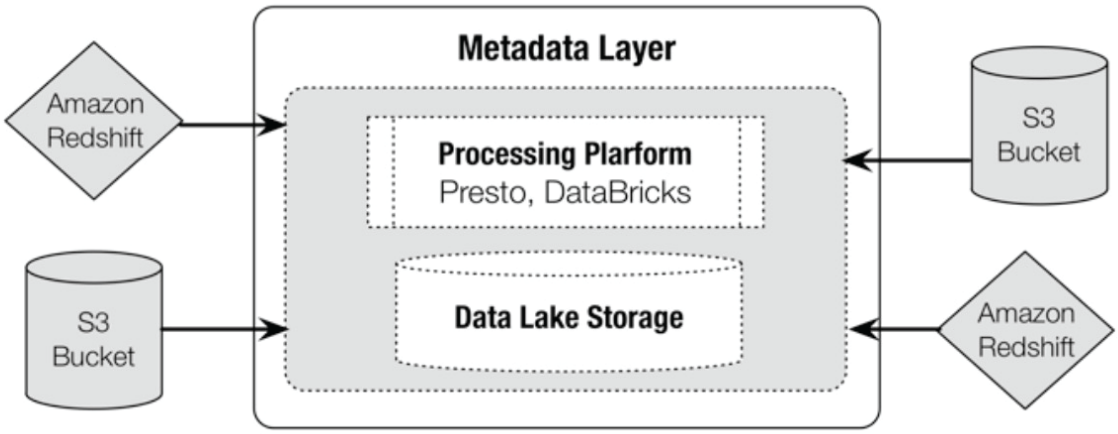
As Europe's leading fashion platform, Zalando needs to store, process, and consume large amounts of data on a daily basis.

Given the large and complex nature of the organization that relied on a data lake, several challenges emerged including a lack of data ownership, poor quality of data after processing, and lack of organizational scalability (number of data sources and consumers). As a result, data teams became a bottleneck. To overcome these difficulties, Zalando decided to build its own data mesh that involved:

i) Evolving towards decentralized data ownership
ii) Prioritizing data domain at the expense of pipeline
iii) Envisioning data as a product, not a by-product
iv) Establishing multi-functional teams organized by sector
v) Relinquishing centralized data environment

These changes allowed Zalando to remove the bottleneck at the data team level, distributing this infrastructure responsibility to the data infrastructure as a platform, and interoperating from a monolithic data architecture (data lake) service environment (datanets).

The figure below shows the data mesh architecture implemented by Zalando. The initial core service (data lake storage) is retained, and a metadata layer and governance are implemented to hold information about it. Zalando then implemented the concept of "bring your own bucket" allowing users to integrate their S3 buckets (Simple Storage Service) with their data into a common infrastructure. Zalando uses technologies like Databricks and Presto to hold its central processing platform. Clusters (Spark clusters) are available on the processing platform and users can configure clusters and navigate their complexity without the team responsible for the infrastructure needing to know what they are doing.



Data Mesh Pattern: Enterprise Data Product Catalog

The Enterprise Data Product Catalog (EDPC) is a repository that aggregates metadata from all on-premises Data Product Catalogs (DPCs). Enterprise Data Catalog is used to store information and statistics (metadata) about all data managed by the enterprise data mesh, making it easier to find, view, use, and manage data. In such an approach:

- Business users use the EDPC to find the information they need to make business decisions.
- Developers use the EDPC to understand the data structures required for the application.
- Governance professionals use EDPC to understand and monitor data across the enterprise, enabling federated computational governance within the enterprise data mesh.

Conclusion

Over time, data architectures have evolved, from data warehouses to data lakes, which are currently the most widely used by enterprises. But despite all the advantages that come with data lakes (e.g., large amounts of data without worrying about rigid data schemas), they often fail to efficiently meet organizational needs. There are issues that need to be fixed.

Data meshing is not just about inserting new technology into existing architectures, adding new functionality, or reorganizing components. Data mesh brings the need to change the current paradigm as the infrastructure of the platform itself leads to a reorganization of data teams.

Data mesh ensures compliance with DATSIS principles, transforms data into products, organizes different teams and data products according to organizational domains (derived from them), and decentralizes the entire process. Data mesh also provides the concept of self-service infrastructure, allowing different teams to create and maintain data products and aggregated data products. This is a faster, lower-cost approach to creating these data products. It is also better because it reduces data duplication, which can lead to inconsistencies, complexity, and increased technical debt.

References

https://martinfowler.com/articles/data-mesh-principles.html

https://www.teradata.com/getattachment/eeb3142a-26d2-437c-bcd7-4d64002395ff/Data-Mesh-MD010585.pdf?lang=en-US&origin=fd

About the Authors

Chaithanya M

Data Engineer, LTIMindtree

A data engineer at DM-NXT COE in LTIMindtree. Technical skills include Power BI, Azure, and AWS QuickSight. Has experience developing and maintaining data pipelines and creating interactive data visualizations.Chaithanya is also passionate about using data to drive business decisions and is committed to producing high-quality work.

Ramanath K P

Data Engineer, LTIMindtree

Full stack developer at DM-NXT COE for Recast Product. His technical skills include Java, Spring boot, and Angular. Ramanath is passionate about solving problems and constantly strives to improve his skills.

LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by 81,000+ talented and entrepreneurial professionals across more than 30 countries, LTIMindtree — a Larsen & Toubro Group company — combines the industry-acclaimed strengths of erstwhile Larsen and Toubro Infotech and Mindtree in solving the most complex business challenges and delivering transformation at scale. For more information, please visit **www.ltimindtree.com.**