ARTICLE

An Engineering Approach to Data Mesh

Stephen Brobst,
Chief Technology
Officer, Teradata

Ron Tolido,
EVP, CTO, and Chief Innovation Officer,
Insights & Data, Capgemini

A data-powered enterprise creates value by making data accessible across the enterprise. Yet a monolithic approach to building a data estate rarely succeeds. This is where the innovative concept of a Data Mesh comes in. It tackles monolithic data complexity by aligning to the notion of domains.

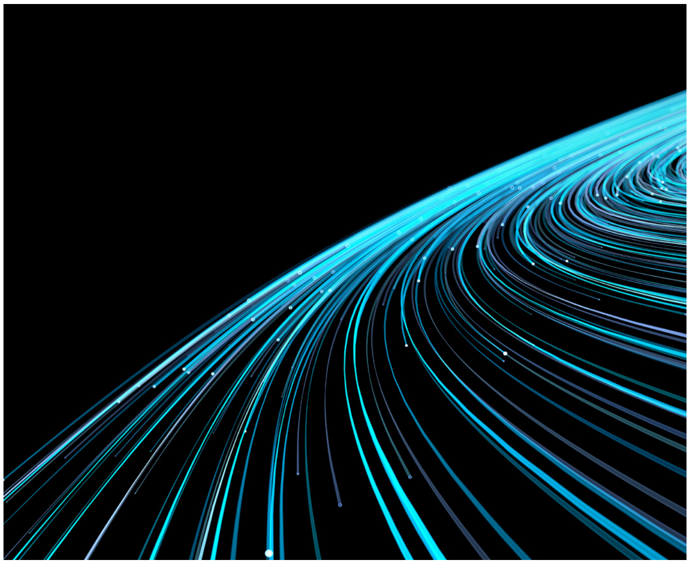
One of the primary advantages is the inherent agility associated with it, with loosely coupled teams focusing semi-independently on their specific areas of business (not technical) expertise. We recommend taking an engineering approach to implementing the Data Mesh concept, balancing different design patterns to actually make the vision come alive.

Data Products Aligned to Real-World Business

A domain may be aligned to a specific line of business within a bank–such as credit card, direct deposit, or mortgage. A domain may also align to a specific functional responsibility such as customer service or branch operations. Our recommended approach to implementation of the Data Mesh concept is to create separate schemas for each domain. Responsibility for data stewardship, data modeling, and population of the schema content is owned by experts with business knowledge about the specific domain under construction.

This approach removes many of the bottlenecks associated with attempting to implement a centralized consolidation of all enterprise data into a single schema. The domain-oriented schemas provide a collection of data products aligned to areas of business focus within the enterprise.

However, decomposition in alignment to business domains should not imply anarchy across the domains. It is critical to recognize the importance of global standards and interoperability when building a distributed data estate. Global standards include areas such as data typing, naming conventions, and quality metrics. Interoperability across the schemas requires consistency in primary and foreign key relationships across schemas, in addition to within a schema. Global access control and optimization of cross-domain query execution are also essential.



Balancing Enterprise Domains

For realization of enterprise data products, it will often be appropriate to create enterprise domains. Efficiency, consistency, and time-to-market considerations require that these enterprise domains will be constructed with the active support from cross-domain subject matter experts.

For example, in banking it would be useful to have an enterprise schema to capture supertype information about accounts across the credit card, mortgage, direct deposit, and other accounts. Subtype information that is specific to a particular business domain (e.g., credit card account attributes not relevant to other domains) would remain in the credit card business domain schema.

But supertype information that is common across all account types (such as open date, account status, balance amount) would be “promoted” into the enterprise domain to enable easy analysis across different account types. Similarly, customer information embedded in each business domain increases in value when promoted to an enterprise domain, facilitating a customer 360 view for purposes of enterprise marketing, risk, and other analytics.

In theory, it would be possible to implement a “union” operator across separate business domains to get an enterprise view of the data, but experience shows that the governance and integration of data associated with enterprise domains can have significant value for selected data products. There is additional work of a cross-functional nature to create enterprise domains, but the consistency and quality derived from selective use of enterprise domains is significant. Having robust governance and an integrated design for these domains creates huge value when performing analysis at an enterprise level.

Most large enterprises use multiple cloud service providers and operate across multiple geographies, so a connected data warehouse is fundamental to Data Mesh implementation.

Deployment Design Patterns

Separate schemas aligned to different domains do not necessarily imply a distinct database for each domain. There are different design patterns for deploying schemas within a Data Mesh:

The co-located approach places domains aligned to different schemes under the management of a single database instance. This contributes to better performance when data across multiple domains is combined. Co-location allows for more efficient execution due to improved query optimization and lower overhead than is required for assembling data across multiple database instances (or even multiple clouds). However, there are cases where the co-located approach is undesirable.

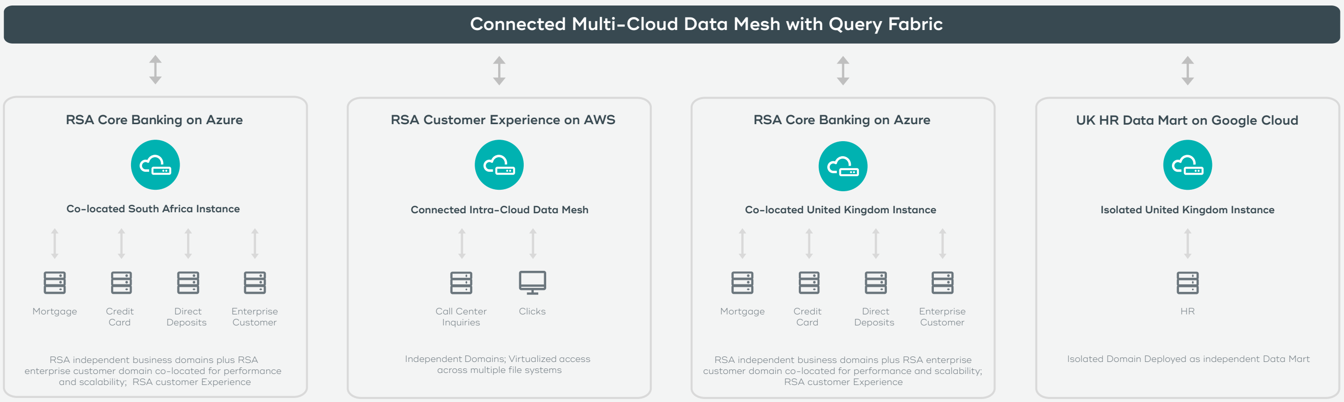

Connected Multi-Cloud Data Mesh with Query Fabric
RSA Core Banking on Azure
Co-located South Africa Instance
Mortgage
Credit Card
Direct Deposits
Enterprise Customer
RSA independent business domains plus RSA enterprise customer domain co-located for performance and scalability; RSA customer Experience
RSA Customer Experience on AWS
Connected Intra-Cloud Data Mesh
Call Center Inquiries
Clicks
Independent Domains; Virtualized access across multiple file systems
RSA Core Banking on Azure
Co-located United Kingdom Instance
Mortgage
Credit Card
Direct Deposits
Enterprise Customer
RSA independent business domains plus RSA enterprise customer domain co-located for performance and scalability; RSA customer Experience
UK HR Data Mart on Google Cloud
Isolated United Kingdom Instance
HR
Isolated Domain Deployed as independent Data Mart

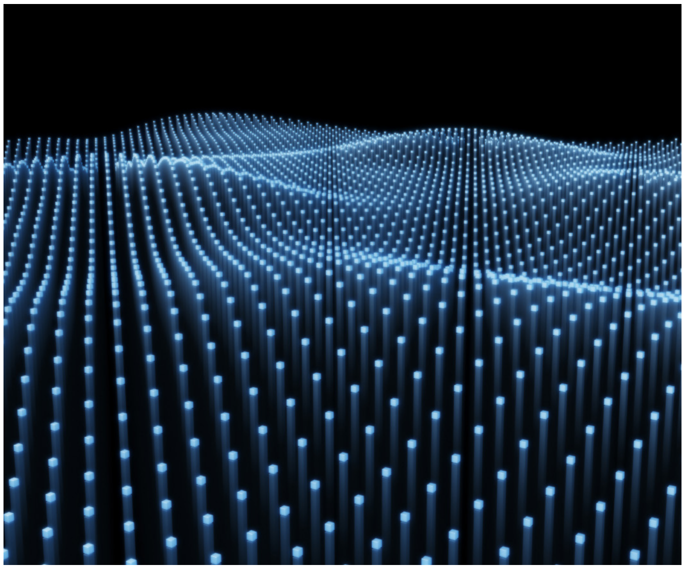
Figure 1. Connected Multi-Cloud Data Mesh with Query Fabric

For example, sovereignty laws may require that data created by a business unit within a specific country must remain in that country. For a multi-national company there needs to be multiple schemas deployed across different geographies–even if the database technology used is the same. Other reasons may be a counter-indication as well, such as data gravity created by applications producing data in different clouds or use of fit-for-purpose database technologies that are distinct for unique analytic use cases or data characteristics.

With the isolated approach, the data product is completely self-contained within a single domain. The schemas used with the isolated technique are usually narrow in scope and service operational reporting requirements rather than enterprise analytics. Isolated domains typically have more autonomy in their deployment–both in data modeling and technology selection.

Sometimes isolated domains are chosen on the need for strong security. More often though, the "real" reason has to do with politics or the desire for organizational independence. The connected approach involves query execution across multiple, disparate locations of data. There will typically be collections of co-located schemas that exist across multiple clouds and/or database technologies.

It is up to a powerful data fabric software infrastructure (including a global orchestration engine) to decide how the query execution takes place across the multiple collections of co-located schemas, optimizing data movement and efficiency of query processing.



Best of Two Worlds

Most large enterprises use multiple cloud service providers and operate across multiple geographies, so a connected data warehouse is fundamental to Data Mesh implementation. Within a cloud service provider and within a geography, co-location of multiple schemas aligned to specific business domains within a single, scalable database instance gives the best of two worlds: agility in implementation and high-performance in execution.

Applying engineering discipline is critical in high-performance deployment of Data Mesh using a combination of the co-located, isolated, and connected design patterns.

We recommend taking an engineering approach to implementing the Data Mesh concept, balancing different design patterns to actually make the vision come alive.

Innovation Takeaways

Business is a mesh

The concepts of a Data Mesh do justice to the distributed, federated reality of most businesses–and accordingly also their data estates.

Data is the product

In a Data Mesh, ownership and stewardship of data are assigned to the actual business domains that hold it; data thus becomes a key product, managed by a domain.

A mixed bag of design patterns

There are multiple ways of deploying schemas within a Data Mesh: co-located, isolated, and connected; which pattern is chosen depends on various enterprise considerations.

Engineering approach

A well-balanced combination of different deployment patterns is a matter of engineering a Data Mesh that does justice to both business and technology considerations.

About the Authors

Stephen Brobst is the Chief Technology Officer for Teradata Corporation and is widely regarded as a leading expert in data warehousing. Stephen performed his graduate work in Computer Science at the Massachusetts Institute of Technology where his Masters and PhD research focused on high-performance parallel processing. He also completed an MBA with joint course and thesis work at the Harvard Business School and the MIT Sloan School of Management. Stephen is a TDWI Fellow and has been on the faculty of The Data Warehousing Institute since 1996. He has been ranked the number four CTO in the USA and has authored numerous journal and conference papers.

Ron Tolido, EVP, CTO, and Chief Innovation Officer, Insights & Data, Capgemini, absorbs technology and business trends, then filters out what matters for impact. He guides global clients in understanding and embracing breakthrough technologies such as AI and autonomous systems to radically innovate and transform their business.

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of 270,000 team members in nearly 50 countries. With its strong 50 year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2020 global revenues of €16 billion. Learn more at **Capgemini.com**.

About Teradata

Teradata is the connected multi-cloud data platform company. Our enterprise analytics solve business challenges from start to scale. Only Teradata gives you the flexibility to handle the massive and mixed data workloads of the future, today. The Teradata Vantage architecture is cloud native, delivered as-a-service, and built on an open ecosystem. These design features make Vantage the ideal platform to optimize price performance in a multi-cloud environment. Learn more at **Teradata.com**.

17095 Via Del Campo, San Diego, CA 92127   Teradata.com

This article first appeared in "Data-powered Innovation Review | Wave 2," Capgemini, 2021.