Looking to establish a Data Fabric and a Data Mesh?

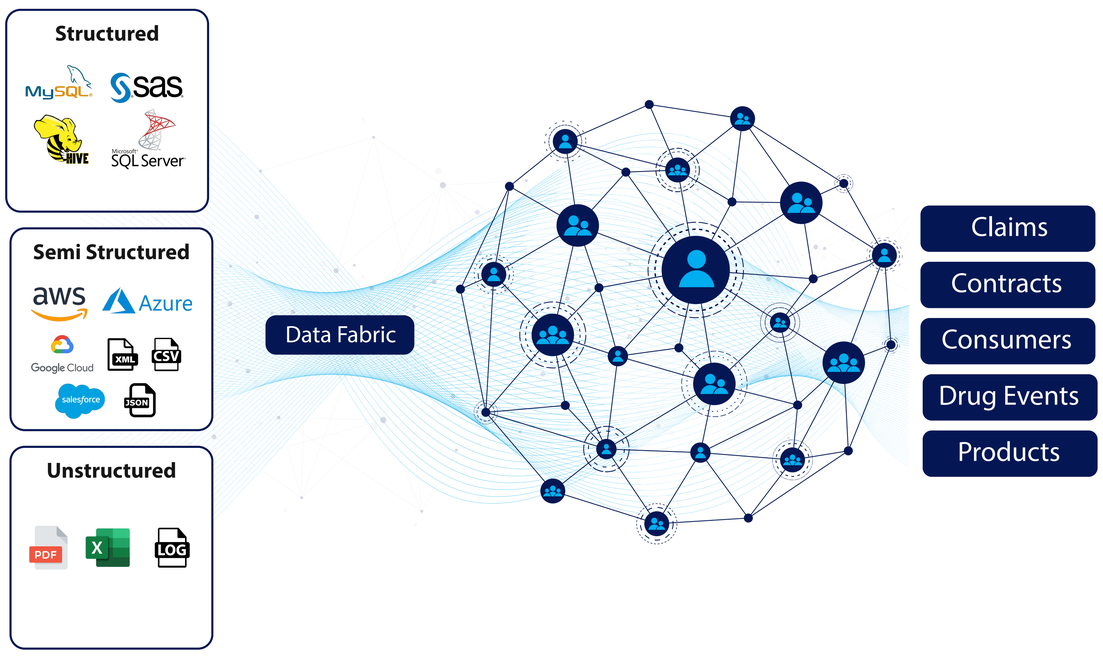
The terms Data Fabric and Data Mesh are now routinely used in the data management and engineering circles. Given the hype and marketing, reaching an agreement on their definitions and usage patterns is proving difficult. The purpose of this short paper is to provide clarity from an adoption perspective.

Context

Data is dispersed throughout an enterprise in a variety of structures and formats, spanning numerous applications, databases, data warehouses, and data lakes. The migration of on-premise data repositories to the cloud extends the data landscape even more, and with Data Scientists requiring external data sets to continuously feed self-learning models, the complexity of managing data is increasing exponentially. The need to think about new data management designs and practices is now front and center in the industry.

What is a Data Fabric?

A Data Fabric needs to be seen from a data management design viewpoint, not from an implementation perspective. No single solution can provide a comprehensive one-stop-shop to enable a Data Fabric. Instead, multiple providers and consumers of data need to be brought together focused on three core tenets for a Data Fabric: agility, integration, and automation. These are supported by using an active metadata repository to capture the source technical and business metadata and visualized through semantic knowledge graphs. A Data Fabric provides data engineers and subject matter experts with the foundations to curate and deliver data domain products.



The main objective of a Data Fabric is to provide a “net” that is cast to stitch together multiple heterogeneous data sources and types, through automated data pipelines that proliferate an active metadata repository.

This allows for logical groupings (without moving the data) to create virtual data domains where augmentation techniques to apply tags (for example classify PHI data) or ML algorithms can be applied to automate the data quality and cataloging of data sets.

As such, a data fabric design is a collection of data services that deliver agile and consistent data integration capabilities across a variety of endpoints throughout hybrid and multi-cloud environments. Further, a Data Fabric adds a layer of abstraction, data can remain distributed with no movement of the physical data aside from crawling and profiling to create a logical map of the data landscape. This removes the need to replicate data for no outcome-driven reason.

Many organizations know that point-to-point integration patterns scale very poorly when faced with too many integrations. What starts out as one to few integrations, quickly morph to become a spaghetti of integration points. A good data fabric design aims to liberate this nightmare scenario with an active metadata catalog to repurpose existing data pipelines. Furthermore, enhancing the productivity of scarce Data Engineers by shifting away from manual, time-consuming, and error-prone ETL tools and toward low-code, UI-driven data pipeline creation saves time and money.

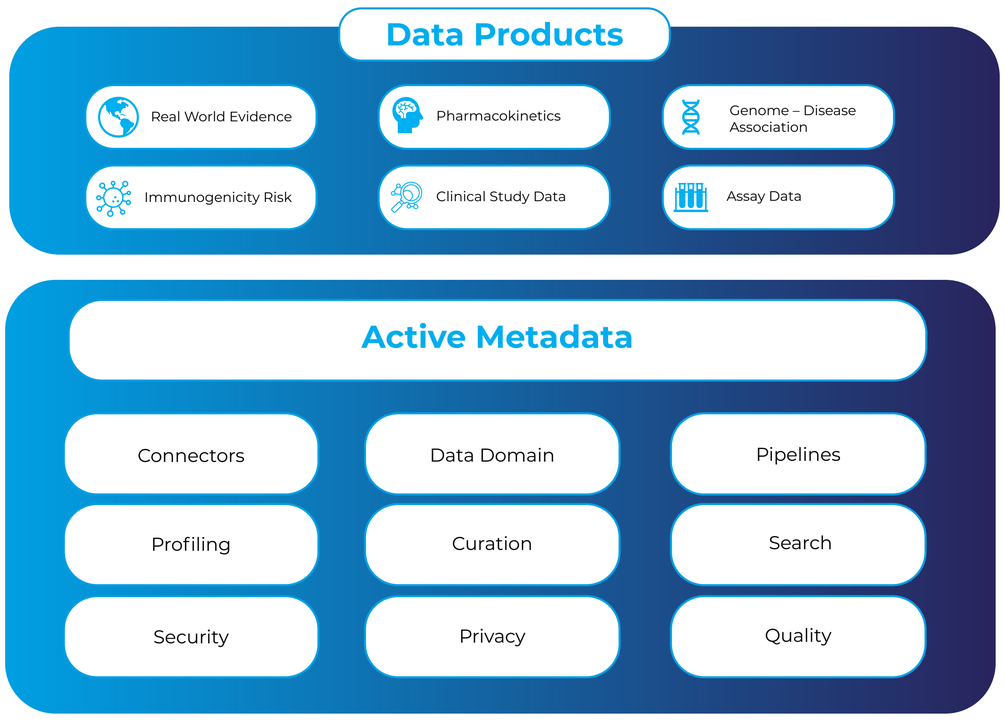
In summary, a Data Fabric can provide data architects and engineers with a design pattern where the focus is on the communication and collaboration with business users on the high-value use cases and less on the data infrastructure.

What is a Data Mesh?

The term Data Mesh was coined by Thoughtworks[1] to address moving from monolithic data platforms to distributed data management. Data mesh aims to connect the two planes of operational and analytical data sets and deliver business-owned data products with a lifecycle (just as software) and consumed through APIs.

Consequently, a Data Mesh can be thought of as a consulting-driven data implementation paradigm that requires customers to balance the decentralized vs. centralized data domain creation, orchestration, governance and management pendulum.

The development of domain-specific Data Products follows the principles that they are discoverable via a self-service data marketplace, trustworthy as the business has validated and interoperable with other data domains and data sets.



A data domain product can be regarded as "dossiers" of institutionalized business knowledge that have been collaboratively curated and made available to a wide range of users. They complement currently limited and focused data marts that give information on specialized or targeted use cases and are based on structured (relational) data sets. Domain-driven Data Products, on the other hand, will have a broader and richer mix of structured and unstructured data to suit a variety of use cases, including fueling AI model design and development to answer the questions of tomorrow.

About Modak

Modak is a solutions company that enables enterprises to manage and utilize their data landscape effectively.
We provide technology, cloud, and vendor-agnostic software and services to accelerate data migration initiatives.
Using machine learning (ML) techniques to transform how structured and unstructured data is prepared, consumed, and shared.

Modak's portfolio of Data Engineering and DataOps studios provides best-in-class delivery services, managed data operations, enterprise data lake, data mesh, augmented data preparation, data quality, and governed data lake solutions.

Modak Nabu™

Modak Nabu™ enables enterprises to automate data ingestion, curation, and consumption processes at a petabyte-scale. Modak Nabu™ empowers tomorrow's smart enterprises to create repeatable and scalable business data domain products that improve the efficiency and effectiveness of business users, data scientists, and BI analysts in finding the appropriate data, at the right time, and in the right context.

- **Find out more**
  https://modak.com/contact/
- **Follow us**
  https://www.linkedin.com/company/modak/
  https://www.facebook.com/ModakData/

[1] Get value from data at scale (https://www.thoughtworks.com/en-in/what-we-do/data-and-ai/data-mesh)

modak