



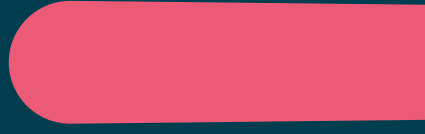
# Aligning data architecture

with organizational structure  
in the financial services sector

/thoughtworks



<b>Introduction</b>	<b>3</b>
<b>Enabling AWS services</b>	<b>8</b>
<b>Aligning data architecture with organizational structures</b>	<b>12</b>
<b>Scenario 1:</b> A centralized organization or business unit	<b>13</b>
<b>Scenario 2:</b> Federated business units within an organization	<b>18</b>
<b>Scenario 3:</b> Fully autonomous business units within an organization	<b>24</b>
<b>Scenario 4:</b> A group of commercially distinct organizations	<b>30</b>
<b>Conclusion</b>	<b>34</b>
<b>Defining your data strategy</b>	<b>36</b>



# Introduction



## **Data architectures and organizational structures have evolved significantly; but not in tandem**

For the past 20 years, driven by the decoupling of storage from computers and the advent of storage area networking, data architectures have favored centralized approaches for defining team, systems, data, and ownership boundaries. However, as the technologies used to manage data have evolved, the organizational structures and governance models that surrounded them largely stayed the same.

The conversations around data architecture evolution have been widely dominated by discussion of technology trends. The shift from data warehouses (schema on write) to data lakes (schema on read) was partially motivated by the growth of big data and the upfront effort required to design data models before migrating data from operational systems. Prior to that, the enterprise data warehouse tried to create a single model for the entire organization, which proved to be hard, if not impossible. This led to the creation of smaller, purpose-built data marts, but there was still high reliance on a central team to create, populate and operate them.

While data lakes allowed upfront modeling effort to be delayed by ingesting raw data at larger volumes, the ownership and operational model didn't change. Central data platform teams remained responsible for large-scale data ingestion, cleansing, cataloging and governance activities.

As technology and services evolved further, there was a movement to migrate data lake workloads into the cloud, decoupling the infrastructure for storage and processing while allowing more flexibility in usage and pricing models. However, the teams responsible for this migration faced similar challenges to scale. Once again, a major shift happened within data architectures, but the impact on how organizations were structured was very limited.

Today, Data Mesh represents the next major evolution in that journey. Data Mesh is a modern data architecture approach that favors a decentralized and federated model to structure teams and data.

The four main principles of a Data Mesh architecture are:

- **Domain oriented decentralization:** Ownership and responsibility for data are spread across the organization, with domains developing their own data products and taking responsibility for the data they expose.
- **Data as a product:** Putting data consumers' needs at the forefront and serving them with the data they need.
- **Self-serve data platform:** To enable data product teams to operate efficiently and independently, the data platform offers self-serve capabilities that increase consistency without requiring central coordination.
- **Federated computational governance:** Data products are interoperable and don't become silos, and global policies are embedded and computationally enforced or monitored.

It's an incredibly powerful proposition for many organizations — creating opportunities to bring data closer to the people who need it, enable self-service and build bespoke data products that can be controlled and optimized by end users. This approach gives organizations greater opportunity than ever before to do the one thing they've consistently failed to do over the last 20 years of data architecture evolution, align architecture with organizational structure.

An organization's data strategies should be shaped by its unique ownership, change management, security and regulatory needs. Data Mesh is making it easier to create architectures that align with all of those considerations. But, even then, it's still not always going to be the right fit for every kind of organizational structure.

With more architecture choices available, and more flexible

technology accessible in the cloud, now is the right time to address the fundamental disconnect between architecture and organizational structure that's persisted for decades. To get the most from any architecture, it needs to support how an organization and the people within it work.

In this paper, we'll explore the considerations for choosing a data architecture to ensure it aligns with organizational structure and supports the right outcomes for all stakeholders. We'll walk through common scenarios, exploring which architecture types are best suited to different organizational structures and show how Amazon Web Services (AWS) infrastructure and tools can bring those architectures to life.



# Enabling AWS services



## Enabling AWS services

The scenarios discussed in this paper explore data strategies for organizations and communities with different levels of centralization and federation. To provide actionable guidance for those organizations, we've created sample architectures based on the capabilities available through AWS. AWS provides a set of modular services which organizations can use in combination to create data architectures, which can be grouped into five broad categories:

- 1. Core:** [Amazon Identity and Access Management](#) (IAM) and [Amazon S3](#) provide the security and storage foundations for any data solution.
- 2. Data transformation:** [AWS Glue](#) is a serverless data integration service that makes it easy to discover, prepare and combine data for analytics, machine learning and application development. AWS Glue provides both visual and code-based interfaces, enabling users to easily find and access data using the [AWS Glue Data Catalog](#).
- 3. Data Lake Formation:** [AWS Lake Formation](#) enables the simple set up of a secure data lake. Lake Formation defines data sources, data access and security policies to apply. It then helps collect and catalog data from databases and object storage, move the data into an [Amazon S3](#) data lake, clean and classify the data using machine learning algorithms and secure access to sensitive data. Users can access a [centralized data catalog](#) which describes available datasets and their appropriate usage. Lake Formation builds on the capabilities available in [AWS Glue](#). Another option is to use [Amazon EMR](#) for running big data workloads on top of an Amazon S3 data lake.

- 4. Data productization:** [AWS Data Exchange](#) makes it easy for data producers to create self-describing data products, and for data consumers to find, subscribe to and use these data products. Once subscribed to a data product, the data consumer can use the AWS Data Exchange API to load data directly into [Amazon S3](#) and then analyze it with a wide variety of AWS [analytics](#) and [machine learning](#) services. Hence, AWS Data Exchange provides a mechanism to enforce strong data modularity with well defined boundaries.
- 5. Data consumption:** Services that enable analytics and business intelligence such as [Amazon Athena](#), [Amazon Redshift](#), [Amazon QuickSight](#) as well as more advanced analytics use cases using machine learning with [Amazon Sagemaker](#), or AI services such as [Amazon Polly](#), [Amazon Rekognition](#), [Amazon Comprehend](#), among others.

Core (1), transformation (2), and formation (3) and data consumption (5) tooling may be used as elements in all of the scenarios discussed within this paper. Data productization tooling (4) becomes essential in scenarios where strong data boundaries need to be enforced between business units in the same organization, or different organizations within an ecosystem.



**Aligning data  
architecture with  
organizational  
structure**

## **Aligning data architecture with organizational structure**

Selecting a data architecture shouldn't be a purely technology-based decision. Above all, it's essential that organizations select a data architecture that's well aligned with their current or intended organizational structure.

In this section, we'll take a detailed look at four common organizational structures, exploring which data architectures may align best with them to help every user and stakeholder get maximum value from their data.



**Scenario 1:  
A centralized  
organization  
or business unit**

## Scenario 1: A centralized organization or business unit

A small, centrally managed organization can realize value from data by using a centralized data architecture managed by a central team. This concentrates expertise and expedites data-driven insights. From a technology perspective, either a centralized data warehouse or data lake can deliver those insights effectively, and the choice between them will depend on the volume, velocity and variety of data managed by the organization.

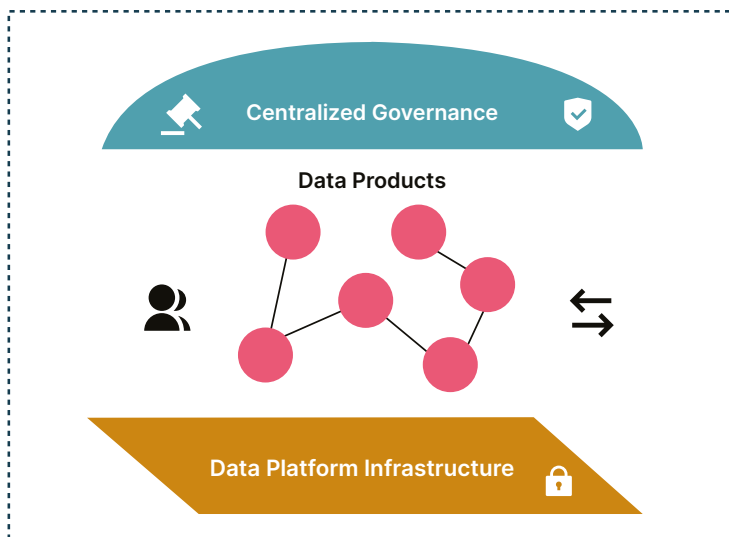


Figure 1: For a centralized organization, a central team can effectively manage data architecture, infrastructure and governance.

Figure 1 shows an example of an organization adopting a centralized data lake architecture. All elements are contained within the data lake boundary (dotted line). A single commercial/legal/governance entity owns all data products. Security and compliance are also the responsibility of this central authority and are enforced and applied to the shared data platform infrastructure, which is managed by a central operational team.

Within this environment, data product visibility and access is also managed centrally. Environmental change is centrally coordinated, with changes propagated across all the participants and data products in lock-step, as data coupling between producers and consumers is high.

Within this structure, both data consumers and data producers lie within the same operational and governance boundaries as the hosted data products. So, the organizational (commercial/legal/governance), operational and functional boundaries are all appropriately aligned.

It's important to note that this is a very common pre-growth scenario. Smaller organizations, or those implementing a data lake for the first time, may find it well aligned with their organizational structure at the point of implementation, but once change and scaling becomes a factor, it may prove restrictive.

This environment has high cohesion and tight coupling. However, the challenge of scaling data and insights becomes difficult for larger organizations with multiple teams, business units and data products. With scale, a centralized approach that requires coordinated change across all parties creates bottlenecks that delay the flow of value.

## **Architecture in action in finance: Control at the cost of scalability**

The centralized data lake approach maps extremely well to organizations that have a primary business objective. For example, the Financial Industry Regulatory Authority ([FINRA](#)) created a flexible data platform that can adapt to changing market dynamics while providing its analysts with the tools to interactively query multi-petabyte data sets. To respond to rapidly changing market dynamics, FINRA moved about 90% of its data volumes to AWS, using AWS to capture, analyze and store a daily influx of 37 billion records. However a centralized data lake approach also maps well to smaller financial institutions that have centralized IT teams. In both cases, the management and governance are the responsibility of a single team, it offers a high level of control and helps keep things like maintaining regulatory compliance simple.

However, it's a model that quickly loses its value as organizations become more complex with more decoupled business units with their own product offering. The control it offers can quickly become restrictive. As soon as an organization wants to collaborate or share data with another, the architecture stops being fit for purpose. For that reason, it can quickly become a barrier to innovation and growth as organizations and their data goals expand.

Organizations that use this architecture need to understand its limitations and recognize when it may stop being the right choice for them. Those that attempt to maintain this architecture as they scale can expect to see limited value created from data, and reduced ability to operationalize and contextualize data.



## Building the architecture in AWS

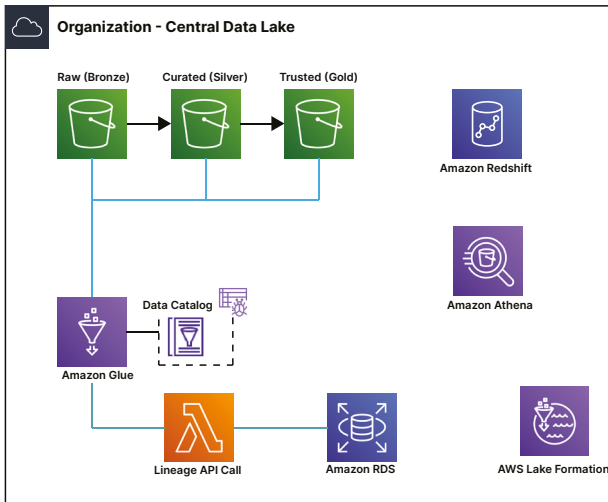


Figure 2: Central data lake on AWS Architecture.

The figure above gives a basic overview of how this kind of architecture looks when built and deployed in the AWS cloud:

- **Amazon Glue** helps discover, prepare and combine data from various raw, curated and trusted buckets; creating a single data catalog.
- **Amazon RDS** provides a scalable relational database solution for data in the lake, enabling the organization to operationalize the data lake however they choose to.
- **Amazon Redshift** uses SQL to analyze data across the lake at speed, opening up a variety of business intelligence and machine learning use cases.
- **AWS Lake Formation** helps automate and accelerate much of the work involved with establishing a data lake, enabling organizations to create these kinds of architectures in just days.



**Scenario 2:  
Federated business  
units within an  
organization**

## Scenario 2: Federated business units within an organization

As an organization grows, a single, central data lake may become too restrictive. To maintain agility, some degree of autonomy may need to be delegated down to each business unit. This may lead to a model where multiple business unit-aligned data lakes (regions bounded by dotted lines) are used instead of one central lake, as shown in the figure below:

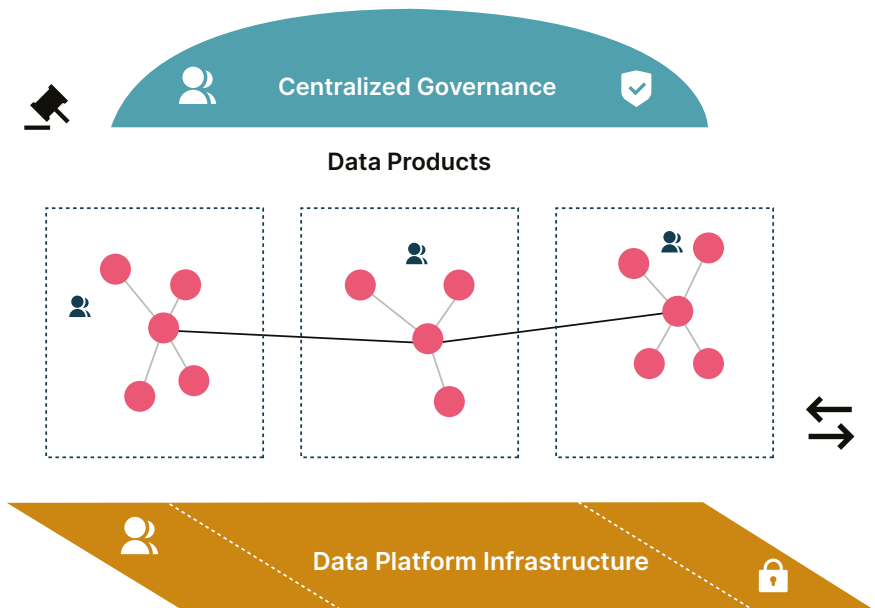


Figure 3: A semi-centralized architecture across multiple business units, with separate data lakes sharing data in a hub-and-spoke model, but central governance and platform infrastructure.

Data is now owned, managed and maintained by each business unit. Data producers and consumers within each business unit interact with data in the same manner as in scenario 1. Within each organization, data sharing is possible through localized data catalogs. However, across the business units, data still needs to be accessible, but might require different access controls. To achieve this, curated subsets or summaries of the data catalogs in each business unit data lake are published to an enterprise-wide catalog, represented by the bold lines between the business units.

While the data usage pattern is peer to peer (i.e. the consumer outside the business unit accesses the data from the business unit's data lake), the metadata interaction is still a centralized hub-and-spoke model. Each business unit publishes metadata to the central catalog, then each consumer searches the central catalog. However, the level of coordination required across all participants is flexible within a business unit, but remains high for data products that are exposed across business units.

While this scenario does not implement a full Data Mesh architecture, it does incorporate a few Data Mesh principles. One example is the creation of a central data platform team operating the shared infrastructure. This helps remove unnecessary operational complexity and reduce operational costs, as each business unit data lake can still use a common underlying technology substrate and a shared control plane. Security at the platform level becomes the responsibility of the central operational team.

It's analogous to the use of enterprise service bus (ESB) and service-oriented architecture (SOA). An ESB architecture decouples business services from each other, but at the cost

of indirect coupling through a central managed infrastructure service. Changes in messaging schema are centrally defined and must be applied across the population of participants. For SOA environments, centralized service discovery and API management are required.

Note that where in scenario 1 all the boundaries were aligned (data, operational, security, governance and commercial), this is no longer true in scenario 2. Governance and legal remain at the organizational level, as does the technology implementation and operational management of the data platform infrastructure. Security is hybrid with local controls within each business unit, but also shared access policies in the central catalog. Data ownership is now aligned to the business units.

A structural hierarchy has also been introduced. Data consumers and producers interact within each business unit as in scenario 1. However, at enterprise scale, the nature of the interaction changes for data consumers and producers across different business units.

## **Architecture in action in finance: Creating organization-wide value from data products**

Data is highly valuable to the team that produces or collects it, but it has a lot of value beyond that context too. That's especially true in financial services, where data gathered on markets, customer needs, trends, and even fraud can be hugely valuable across a wide range of business units.

If you're an institution like JP Morgan Chase that has multiple lines of business and corporate functions, a federated data lake approach can help everyone in the organization benefit from valuable data, curated and controlled by the people closest to it. While this is not a full data mesh implementation, it was exactly why [the firm recently enabled a data architecture using AWS](#) that's aligned with its organizational structure

Control over data sits with the teams closest to it, who can curate and manage it in ways that make it easier for other teams to get value from it. Everyone is empowered to help themselves to datasets and products, within the terms defined by each of the owning teams. It's a decentralized approach that helps amplify the total value created from data, where a central data catalogue allows consumers to find the data they need.

## Building the architecture in AWS

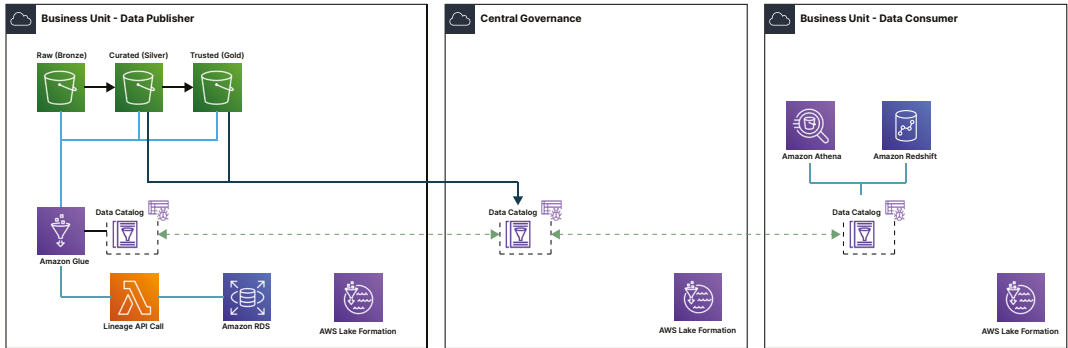
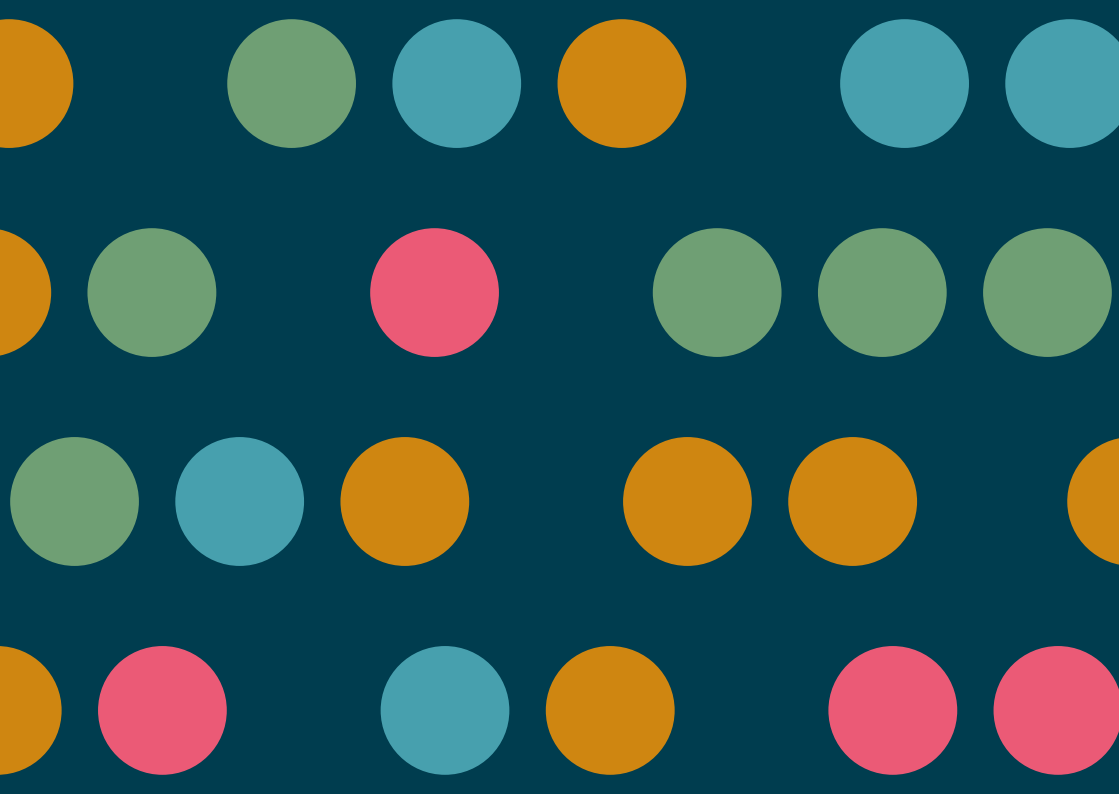


Figure 4: Decentralised Data Mesh model on AWS Architecture.

As the figure above shows, in this AWS architecture:

- **AWS Lake Formation** is used to create three distinct data lakes: one for business unit data publishers, one dedicated to centralized data governance and one to be utilized by business unit data consumers.
- **Amazon Redshift and Amazon Athena** are utilized by data consumers within distinct business units to operationalize the data within their own data catalog.

Within this architecture, data can be moved between each lake's data catalog, enabling a central governance team to ensure quality and compliance before it's put into the hands of consumers.



**Scenario 3:  
Fully autonomous  
business units  
within an  
organization**



## **Scenario 3: Fully autonomous business units within an organization**

Organizations that have grown organically or through acquisition will have a much more complex data and operational landscape. They might demonstrate even higher levels of federation.

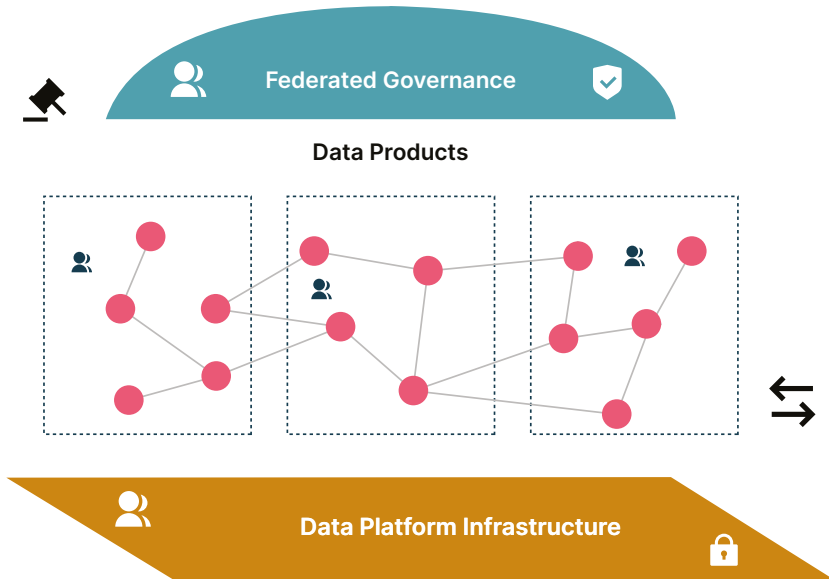
Within these structures, business groups operate completely autonomously, each with their own technology strategy and operational concerns. Only commercial, legal and governance policies remain defined by the organization.

The data strategy for a highly federated organization should decouple data interactions between participants in different business units in the same organization. In this scenario, Data Mesh principles can enable the scaling of data sharing and insights across the organization.

As shown in the figure below, data is curated into multiple data products — each owned and managed by an autonomous data product team. Those products are then advertised and made available across business units through a common mechanism offered by a self-serve data platform, enabling other teams to benefit from them and use them in new ways alongside their own data products.

To support this higher level of autonomy, while still being able to apply and enforce common governance policies, there is a need to create a strong self-serve data platform capability, alongside a set of shared services to facilitate interoperability and data access. For example, an organization-wide IAM framework and an agreed mechanism to register and discover data products. Each data product producer locally controls which third parties may consume the products they advertise. The data products

themselves are self-describing, meaning a third-party data consumer can infer the data structure or schema from the metadata embedded in the data product.



*Figure 5: A strong intra-organizational Data Mesh, with strong domain ownership and interoperability between data products.*

Within this architecture, operational and data boundaries are aligned with the organization's business unit boundaries and domains. The data platform is shared, but offered as self-service infrastructure to enable interoperability, sharing and computational federated governance. Figure 6 shows a more exhaustive list of the platform capabilities required to build and operate a Data Mesh architecture. Each team has autonomous control over its own data products, but those products are still accessible and available to be utilized across domains.

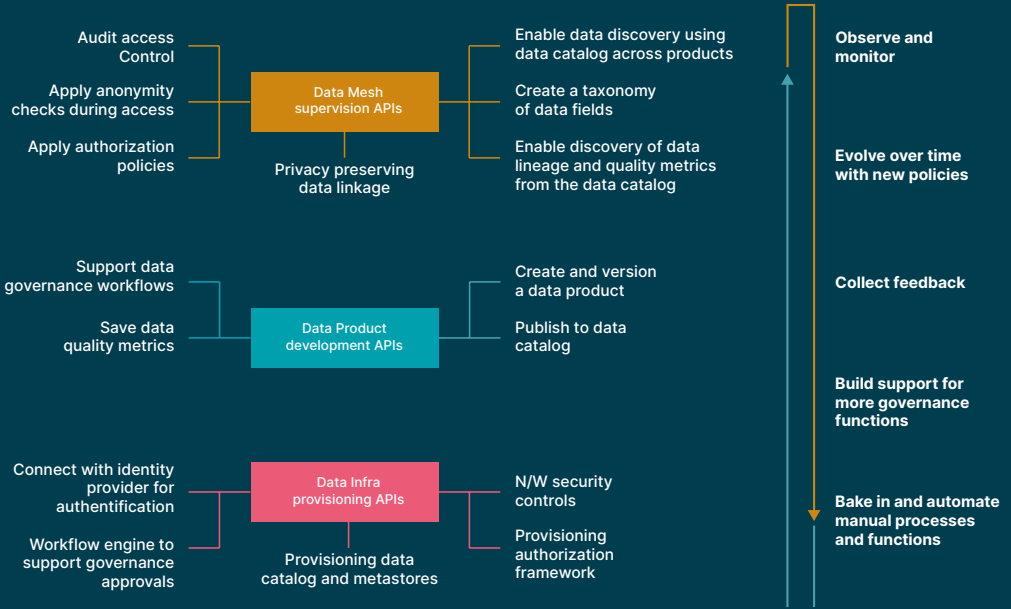


Figure 6: Self-serve and federated governance capabilities that must be supported in a Data Mesh architecture.

## Architecture in action in finance: Supporting governance and data quality efforts at Saxo Bank

In any financial organization, the flow of accurate and reliable data is critical to effective decision making, understanding of risk and maintaining regulatory compliance. According to [IDC research](#), knowledge workers spend approximately 30% of their working hours searching for and gathering data, and the [Bank of England](#) has highlighted that 57% of regulatory reporting resources are associated with process flow, due to the highly manual processes in banks.

Saxo Bank has been partnering with Thoughtworks on a journey to build a self-service data catalog and quality platform, Data Workbench, to enable domain teams to explore and consume data products. This Data Mesh approach eliminates avoidable dependencies and incorporates a new business glossary of unified definitions for business terms across the organization. This enables the easy search of data assets and their origins, giving users clarity, building trust and improving governance.

## Building the architecture in AWS

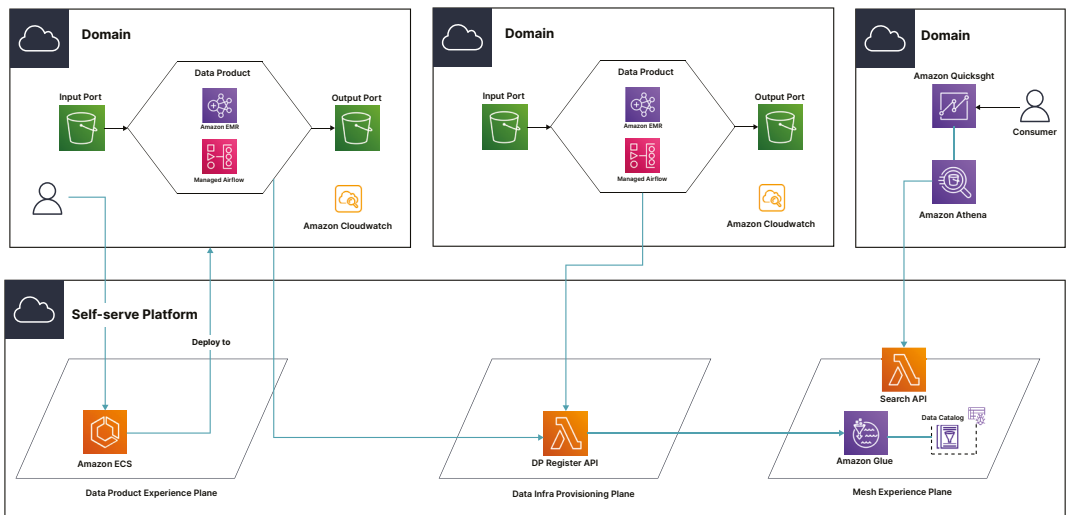



Figure 7: Example of domain-aligned teams on AWS Architecture.

As the above figure shows, in this AWS architecture:

- **Domain-aligned teams and data products:** Each domain team has its own AWS account and data ownership, but deployment of data products are enabled by a shared Elastic Container Service (ECS) built by the data platform team, that can read a common data product specification format and deploy the data product infrastructure on the target domain team's account.
- **Self-serve data platform:** The shared platform contains capabilities and shared services that enable interoperability, the organization-wide data catalog, data product registration and other capabilities listed in figure 6.
- **Federated computational governance:** The self-serve platform enables the ability to bake in the defined governance procedures and policies, and enforces them at various stages of a domain data product since its inception.

This architecture provides a high degree of autonomy to the different domain teams across the organization, but also requires a high maturity from the data platform team to build truly self-serve capabilities that support a federated architecture.



**Scenario 4:  
A group of  
commercially  
distinct  
organizations**

## Scenario 4: A group of commercially distinct organizations

In this scenario, we're looking at a structure that incorporates multiple autonomous companies operating in an ecosystem. Each company manages its own internal commercial, legal and governance concerns, but there may be a less formal community code of conduct in place to guide their operations.

To maintain maximal decoupling, commercial licensing and usage conditions must be embedded in the data products published by each organization. All parties within the community agree on a marketplace through which data products can be advertised and consumed.

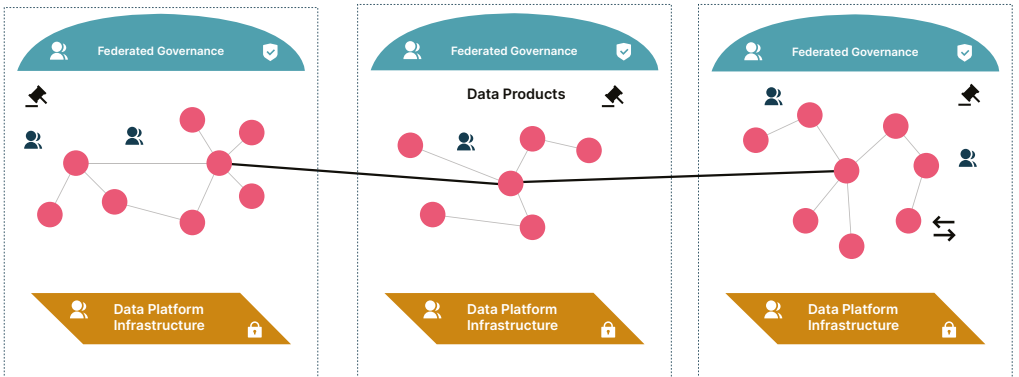


Figure 8: A strong inter-organizational Data Mesh.

Just as with the autonomous business units in figure 5, commercially distinct organizations are empowered by the architecture in figure 8 to create and manage their own data products, and define how those products are shared and used by the other organizations in their ecosystem.

Power over governance and control of data rests with the individual organizations, but the structure makes secure collaboration possible across those organizations. The organizations can learn from each other's data products, without having to relinquish control or ownership of their products.

## **Architecture in action in finance: Working hand-in-hand with regulators**

Perhaps the most obvious application of a structure like the one shown in figure 6 is cross-organizational commercial collaboration. But the model can also transform how financial institutions work with external organizations like regulators.

Organizations can create data products specifically for regulators. So, when a regulator wants information about the organization's activity, they can help themselves to it immediately, instead of lodging a request and waiting days or weeks to have it fulfilled.

It's a shift that can completely redefine the relationship between financial institutions and financial regulators. With shared access to specially curated data products, the relationship becomes seamless and collaborative. When new regulations are introduced, all the business needs to do is evolve its data products to align with the regulatory bodies, so they can get everything they need in an instant.



## Building the architecture in AWS:

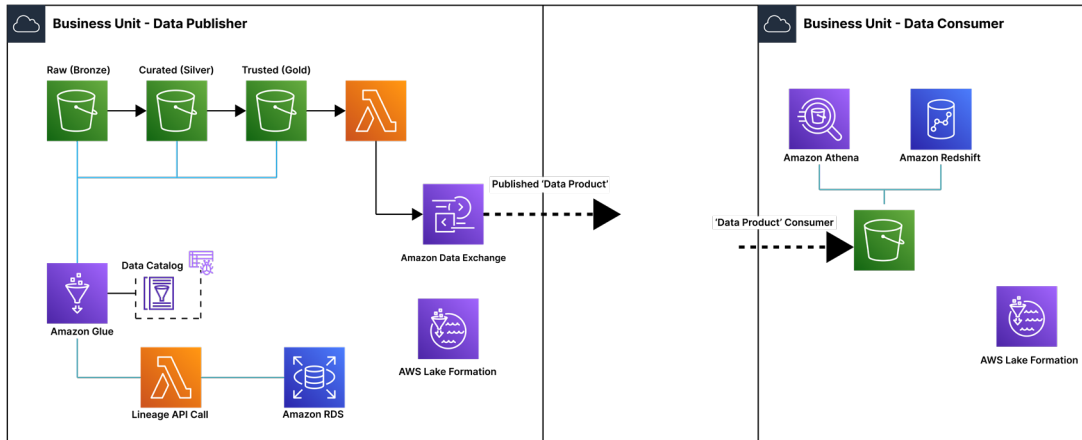


Figure 9: Multiple business units on AWS Architecture.

As the above figure shows, in this AWS architecture:

- **Amazon Data Exchange** is used to combine data from across the publisher's different buckets into an accessible data product. Using this product, consumers can utilize and benefit from that data, without needing to rely on direct transfers or APIs.
- **AWS Lake Formation and Amazon Glue** are used in the same way they would be when creating a single centralized data lake; creating distinct lakes that aren't directly connected to one another.

The two units function entirely independently, as entirely distinct lakes. But the publishing of data products enables consumers to pick up and securely use compiled, clean and operationalized data from the publishing team.



# Conclusion

## Conclusion:

The value of aligning organizational and data structures isn't a new revelation. Far from it in fact. In 1968 Melvyn Conway observed that "organizations which design IT systems are constrained to produce designs which are copies of the communication structures of these organizations," now known to us as Conway's Law. Now, architecture and organizational decision-makers must take Conway's Law to heart, and use it to guide their choices as they build future-ready data architectures, a technique that Thoughtworks calls the [Inverse Conway Maneuver](#).

But, even once you accept that conclusion and resolve to align the two, creating the ideal data architecture or strategy isn't as simple as examining an organization's current structure and building something that mirrors it. At least, not in the long term. It's important to consider that organizational structures and data architectures are evolving entities. If you build the ideal architecture around today's structure, what happens when that structure shifts in line with new business requirements or a change in strategy? And if data itself holds the key to growing an organization, how can you create architectures that are aligned with today's structure and ready to support the structures of tomorrow?

While there's no single right answer to those questions, highly federated and flexible structures like data meshes can be a huge help. They enable organizations to take a highly modular approach to data architecture, where governance and agility can be balanced in line with the current state of the business and where the organization is heading.

By incorporating product thinking into data strategy, placing domain experts in control of each data product and operating a structure that can scale freely as needs change, organizations can create strong alignment between organizational structures and data architecture. And, more importantly, they can actively maintain it as both entities evolve.

## Defining your data strategy

As we've shown throughout this paper, choosing an appropriate data architecture requires a strong understanding of an organization's data strategy and structure. When Thoughtworks helps companies define and implement their data strategy, we take a value-driven approach as depicted in figure 10 below - looking holistically across multiple dimensions to answer these questions:

- **Value proposition:** How is the enterprise's mission supported, enabled and driven by data, analytics and AI?
- **Data assets:** What data is available within the organization, and how accessible is it to the people and processes that need it?
- **Organizational culture:** Does the organization have a data-driven culture? And if not, what will need to change to enable that evolution?
- **Operating model:** What are the key processes across the organization, and how should teams be organized to optimize operations?
- **Skills and capabilities:** Does the organization have the right people and skills to achieve its data goals?
- **Technology:** How is the organization's data strategy supported by architecture and technology?

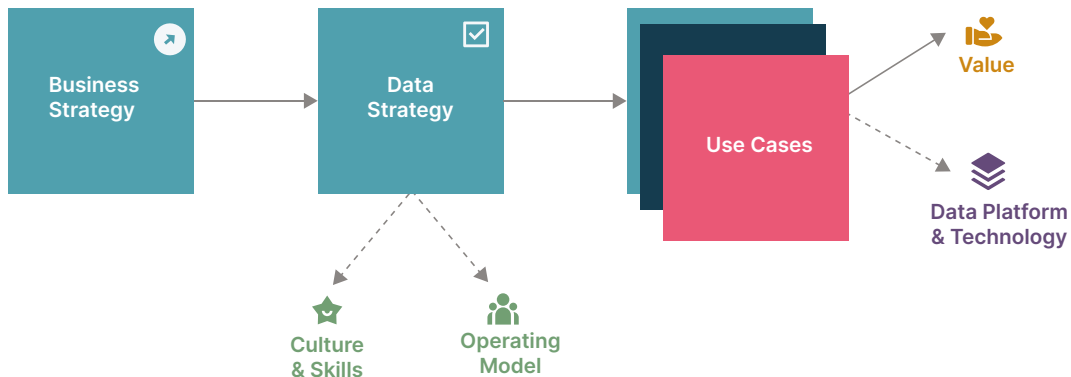


Figure 10: A value-driven approach to defining your data strategy and how it impacts your organizational structure and technology architecture.

Looking holistically at all those dimensions helps you make technology and architecture decisions that are aligned with your organization’s goals. As we’ve seen, the decision between opting for a federated, Data Mesh approach or a more traditional centralized data architecture will depend on the constraints, ambitions and characteristics of your organization.

Thoughtworker Zhamak Deghani originally outlined the fundamentals of the Data Mesh approach based on work Thoughtworks has done while helping clients navigate that process at scale. We have decades of experience helping clients across industries — from pharma and healthcare to financial services and banking — build data architectures that align with their organizational needs, structure and goals.

If you need help defining and implementing the right data strategy for your organization, then [speak](#) to a Thoughtworks expert today about our Data and AI services.



## About the author

**Richard Nicholson**

**Principal Solution Architect, Amazon Web Services**

Richard is a Principal Solution Architect in the AWS Financial Service EMEA business and market development team. Richard works on areas such as front office risk system architectures and back office core mainframe migration.

Prior to AWS, Richard spent 18 years in his own company focusing on the development and use of highly modular and runtime-adaptive Java/OSGi based software systems, for a range of industries including Financial Services and Industrial IoT.

An Astrophysicist by training, Richard entered the Financial Service industry in 1995, as an Infrastructure Systems Administrator for Salomon Brothers.



## About the author

**Danilo Sato**

**Head of Data & AI Services UK and Europe, Thoughtworks**

As the Data service line lead for Thoughtworks UK and Europe, Danilo is responsible for building high-performing teams to solve our client's most complex data problems. He leads technical projects in many areas of architecture and engineering, including software, data, infrastructure, and machine learning.

As an acknowledged thought-leader in the data space, Danilo has published books such as *Devops in Practice*, and has spoken at conferences around the world on data architecture and machine learning.

Danilo was recently nominated by DataIQ as one of the most influential people in data in 2022.

## Get in touch with us

[uk-partnerships@thoughtworks.com](mailto:uk-partnerships@thoughtworks.com)

Thoughtworks Ltd.  
First Floor, 76-78 Wardour Street  
London, W1F 0UR, UK  
+44 (0)20 3437 0990  
[thoughtworks.com](https://www.thoughtworks.com)

