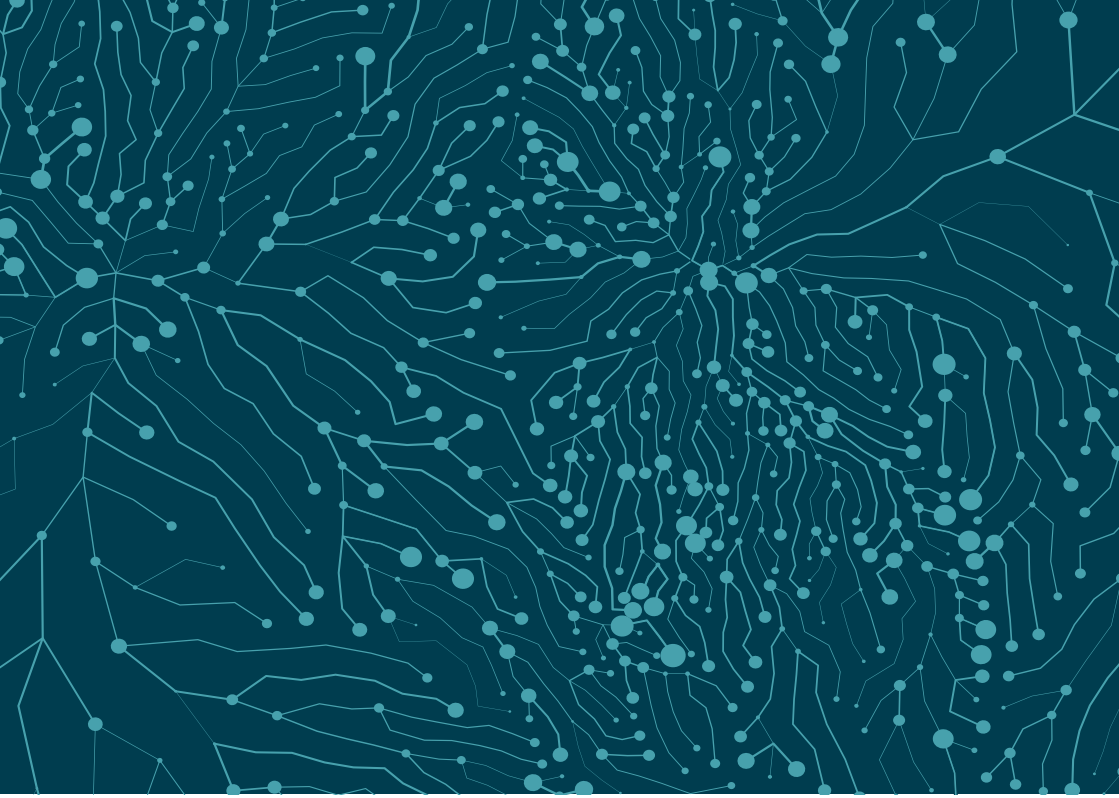

Bringing Data Mesh to life

Expert advice and reference architectures to guide your journey to a new data paradigm

thoughtworks | Google Cloud

In May 2019, Thoughtworks luminary Zhamak Dehghani **published an article** that would begin an entirely new data architecture paradigm. Zhamak introduced the world to Data Mesh — a decentralized socio-technical approach to managing data at scale designed to help organizations accelerate their ability to derive value from their data.

Since then, Thoughtworks' and Zhamak's thinking around **Data Mesh** has evolved significantly.

Alongside this thinking, Thoughtworks and Google Cloud have been helping many organizations implement the approach; building practical experience and expertise around the theory.

However, a key question remains for the thousands of organizations interested in harnessing the benefits of Data Mesh: What does it practically take to build and manage an enterprise Data Mesh? As the term socio-technical suggests, a successful Data Mesh journey requires teams to consider both the social and technical implications of adopting this new model.

In this paper, we'll show you how to bring a Data Mesh to life. We'll share practical advice from expert Thoughtworkers and Googlers based on their experiences helping customers start their Data Mesh journeys, and provide reference architectures to illustrate how you can build your own Data Mesh on Google Cloud.

The journey to Data Mesh adoption

Data Mesh isn't just a data architecture framework — it's an entirely new paradigm. It'll have a profound impact on how your teams work, how businesses act on the insights they uncover, how data is governed and controlled, and how it moves around your organization.

Because it represents such diverse and far-reaching change, it's essential that your Data Mesh journey begins with a clearly defined strategy. From there, you can create a detailed roadmap that considers every aspect of your transformation, and what it means for your teams and domains.

Thoughtworks breaks this planning process down into several manageable chunks, helping organizations work through logically and strategically. The process helps move them from an initial vision towards a clearly defined Data Mesh roadmap and prioritized list of use cases, and deliver measurable success as the organization progresses through its Data Mesh journey.

Start by defining your data strategy and set a clear vision of success

If you're thinking about building a Data Mesh, the very first question you need to answer is 'why?' What problem are you trying to solve? And what is the value to be delivered?

While there is a lot of hype and discussion around the approach today, it's important to recognize that a Data Mesh isn't a magic wand. It's designed to solve specific data management, sharing, governance and operationalization challenges. If you're not facing any of those challenges, it might not be the right thing for you yet.

In our experience, successful Data Mesh journeys are predicated on a well-articulated data strategy. As outlined in DORA's DevOps transformation guidance[1], all transformation needs a 'true north' around which every stakeholder and team can align themselves, and prioritize their efforts towards. In a Data Mesh journey, your data strategy provides that 'true north'.

The process of building that strategy starts with a detailed examination of the challenges an organization is facing, its current data architecture and processes, and the business outcomes it is working towards.

1 https://cloud.google.com/architecture/devops/devops-culture-transform

When evaluating your own data strategy, or creating an entirely new one, start by answering the following questions:

- What data assets do you have available, and what sort of condition are they in?
- How will you measure the success of your data strategy and your Data Mesh?
- What key decisions does your business need to make, and how could they be optimized? Who is making those decisions?
- What do your ideal ways of working look like, and how might your organizational and operational structures need to evolve towards that?

By answering those questions, you can work towards defining perhaps the most important element of your Data Mesh journey — your vision.

Your vision should clearly lay out how you want your organization to work with data and get value from it. From there, you can determine what your architecture and operations should look like after your Data Mesh has been created. Your target vision should show all the relevant domains and the data product. And as a primary purpose of Data Mesh is to evolve with the business, if the business changes your target vision will change.

A new data strategy demands operational evolution

Data Mesh will have a significant impact on how people engage with data across your organization. Your operating model, or the ways that teams will execute on the strategy to deliver value to their customers, will need to evolve to help them obtain the insights from Data Mesh.

Your operating model needs to align domain teams to your strategic vision and empower them to independently determine how best to innovate, adapt and respond to change. That's why across many of our customer engagements, Thoughtworks has advocated for operating models based on **the EDGE framework**.

There are many commonalities between EDGE and Data Mesh, which include:

- Emphasizing autonomy across domain teams
- Empowering teams to achieve their goals, their way, without prescriptive delivery requirements
- Advocating for developing multiple use case 'bets', sometimes referred to as 'value hypotheses', simultaneously, so teams can easily pivot between them if one doesn't work out
- Challenging traditional centralized structures, proposing new approaches to governance and the development and execution of strategy

That similarity makes the EDGE model a good fit for many organizations that are adopting Data Mesh for the first time. By taking the EDGE operating model as your starting point, your journey will be closely aligned with the principles of Data Mesh from the beginning. The model also ensures that every decision-maker has a consistent definition of what constitutes 'value' for the Data Mesh, helping leaders identify which domains are best suited to be onboarded into the Data Mesh first, and which should follow.

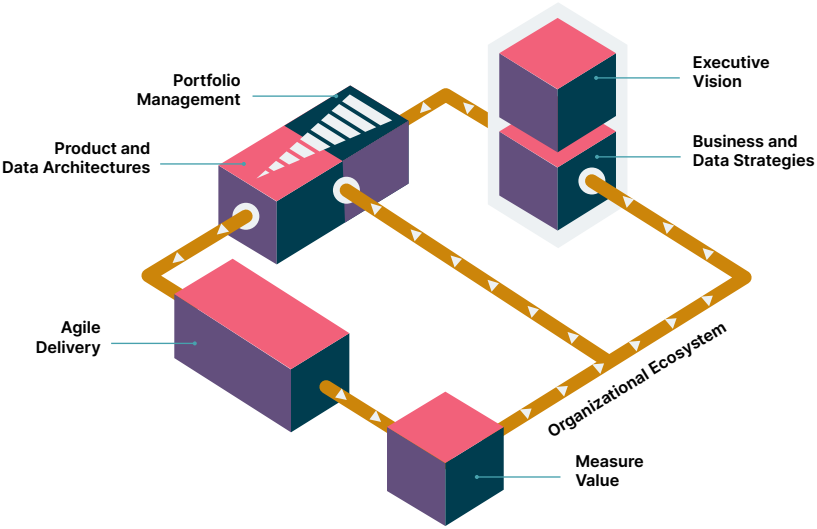

Figure 1 - The EDGE approach for portfolio management

Figure one shows how an EDGE-based organizational approach can help build agile, responsive organizations. The overarching executive vision and business strategy, also translated into your data strategy, are used to drive intelligent portfolio management, which in turn helps accelerate delivery.

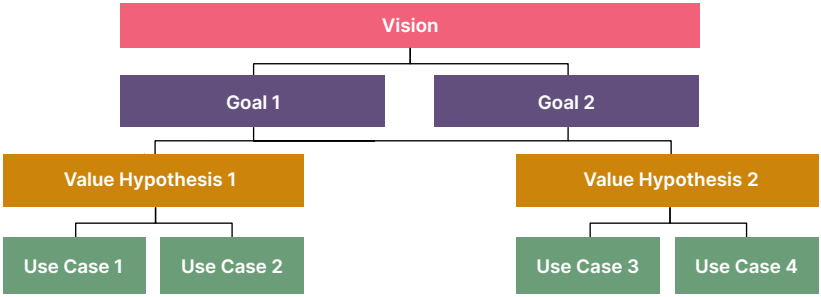

Figure 2 - An example of a lean value tree

The EDGE model enables intelligent portfolio management through the creation of Lean Value Trees (LVTs) as shown above.

As the example LVT in figure two shows, everything begins with a value-oriented vision. Then, the vision is broken down into multiple goals. From there, domains can develop multiple value hypotheses for how they might be able to achieve those goals with the right data products. Finally, the data products needed to bring those value hypotheses to life are mapped out into use cases.

By keeping everything focused on value, the LVT approach doesn't just ensure that domains end up creating the right data products, it also provides a solid foundation for prioritization.

From looking at hypothesized LVTs, leaders can understand which domains present the strongest and clearest cases for early onboarding into the mesh, and prioritize them accordingly. Looking at multiple LVTs side by side can also help identify cross-domain use cases, and enable Domain-Driven Design from the earliest stages of a Data Mesh implementation.

Turning your organizational and data vision into a roadmap

Once your vision is defined and you've mapped out the first domains and data products to onboard to your Data Mesh, it's time to create a clear roadmap that shows how you'll make your vision a reality. Every organization's roadmap will look slightly different depending on its 'current' state and its unique vision. But there are some important processes and common elements that should be in your roadmap:

- Workshops with domains and central teams to help you identify and prioritize use cases
- Domain onboarding, in line with your identified priorities list
- The building of your core Data Mesh platform team
- Revisions and changes to your operating model in line with the Data Mesh
- The creation or restructuring of teams to operate as cross-functional data product teams that deliver the prioritized use cases

Start bringing your domains onboard

With your overarching vision established, and a clear roadmap of how you'll get there, it's time to shift your focus down a level to the individual domains that will ultimately come together to be part of your Data Mesh.

Just as your organization needs a robust strategy and vision for what Data Mesh should help it achieve, so do your individual domains. Thoughtworks follows a 'double diamond' approach to domain onboarding (as shown below) to help establish what each domain wants from the Data Mesh, how it will measure success and which data products could deliver the most value for them.

Data Mesh vision to use case

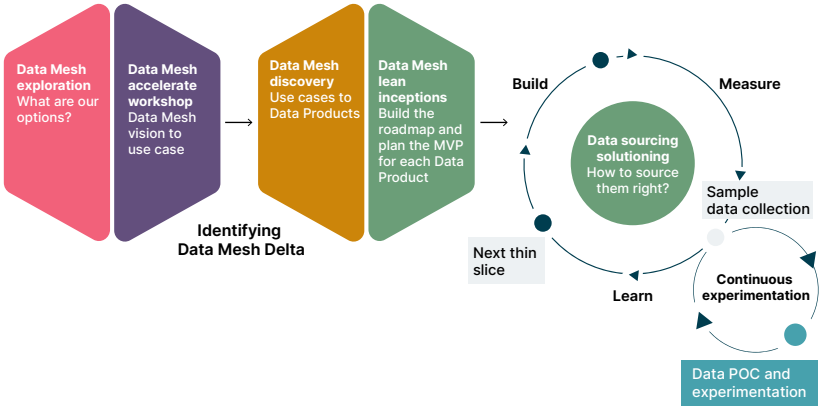

Figure 3 - The double diamond process with the four stages to onboard a new domain into the Data Mesh

Figure three shows how the double diamond approach breaks the domain discovery and onboarding process down into four stages:

- **Stage One: Domain exploration** - The process begins with a deep-dive exploration of the domain's needs, helping them understand how joining the Data Mesh could deliver value for them as well as the implications and changes required.
- **Stage Two: Workshops** - Accelerate Workshops bring the domain teams together and help them turn the vision into valuable use cases. At this stage the domain's goals are defined, along with their measures of success.
- **Stage Three: Data Mesh Discovery** - Here the team identifies potential data products that they could create to help them deliver their use cases and organize the teams and governance processes to support them.
- **Stage Four: Lean inceptions** - Finally, we bring everything together into a clear delivery roadmap for their data products. At this stage, the initial Minimum Viable Products (MVPs) are decided and the team can create its initial technical backlog.

While that process is ongoing, the data platform team must go through a similar journey, exploring and discovering the required capabilities for the core platform. In our experience, this works best when the data platform team works closely alongside domain teams to create the first MVP.

Through this process, you will engage with the domain leadership to lay out goals and objectives, and create an implementation roadmap that covers product, technology, and operating model. You can read more details about these tracks in our **Roche Data Mesh implementation article series**.

Building your Data Mesh

With your vision set, your strategy defined, a clearly prioritized list of domains to be onboarded and data products for those domains to create, you can start to shift your focus towards the technical delivery side of your Data Mesh project.

Your current architecture and unique requirements will ultimately shape the core data platform you build. Regardless, there are several elements that are common across different Data Mesh implementations where pre-built capabilities and functionality can significantly accelerate the building process.

When it comes to building out those elements, you could choose to do so in any major cloud environment. For the purposes of this paper, we've chosen to create the reference architecture for Google Cloud. Key reasons for choosing Google Cloud for your Data Mesh include:

- **Simple and comprehensive data ecosystem** enabling a broad range of operational and analytical use-cases across batch and streaming, structured and unstructured data, data lake and data warehouse, all with integrated machine learning. This allows data product teams to quickly generate value from data and democratize access to data for consumers.

- **Serverless and auto-scaling allowing consumers fast access to data**, provides limitless scaling for multiple concurrent users, reducing operational overhead and total cost of ownership by virtue of being fully-managed, alleviating bottlenecks on the platform team and allowing data product teams to focus on creating value from data.
- **Integrated data product observability and governance** empowering data product teams to choose the storage and compute architecture that best fit their use-cases; reducing data duplication, enabling fine-grained cost attribution. Whilst ensuring peace of mind by providing out-of-the-box data observability, governance and security designed to support distributed and fast-paced data teams.

Now we'll look at three crucial elements of an enterprise Data Mesh in detail:

- Element #1: Self-service data platform and team
- Element #2: Building and managing data products
- Element #3: Enabling computational federated governance

For each, we'll look at example implementations and reference architectures that show what each element can look like in Google Cloud. These architectures show just one example of how these elements can be structured in Google Cloud, in practice, your own architecture could be structured differently, in line with your needs and goals.

Element #1: Self-serve data platform and team

To reduce the complexity of building, deploying and maintaining interoperable and trustworthy data products, the data platform team provides a self-service solution with all the tools and building blocks domain teams need to create and consume data products. By facilitating both creation and consumption of data products across domains, the self-serve data platform makes it easy for teams to turn domain-specific data products into innovative cross-functional use-cases.

To be successful, the data platform ultimately needs to make it easy for domain teams to consume data products. If it doesn't, they'll create their own alternatives, undermining the Data Mesh model. A good self-service data platform should make production and access workflows as intuitive as possible, and simplify technical and architectural decision-making by providing proven choices that get the job done without introducing excessive complexity.

By leveraging reusable frameworks and managed cloud services through automated pipelines, the data platform can reduce operational overhead and complexity across the Data Mesh. That allows domain teams to focus on value-adding activities, and ensures that Data Mesh delivers on its promise of making data more manageable and accessible for everyone.

The reference architecture in figure four below shows what a typical self-service data platform looks like when built in Google Cloud.

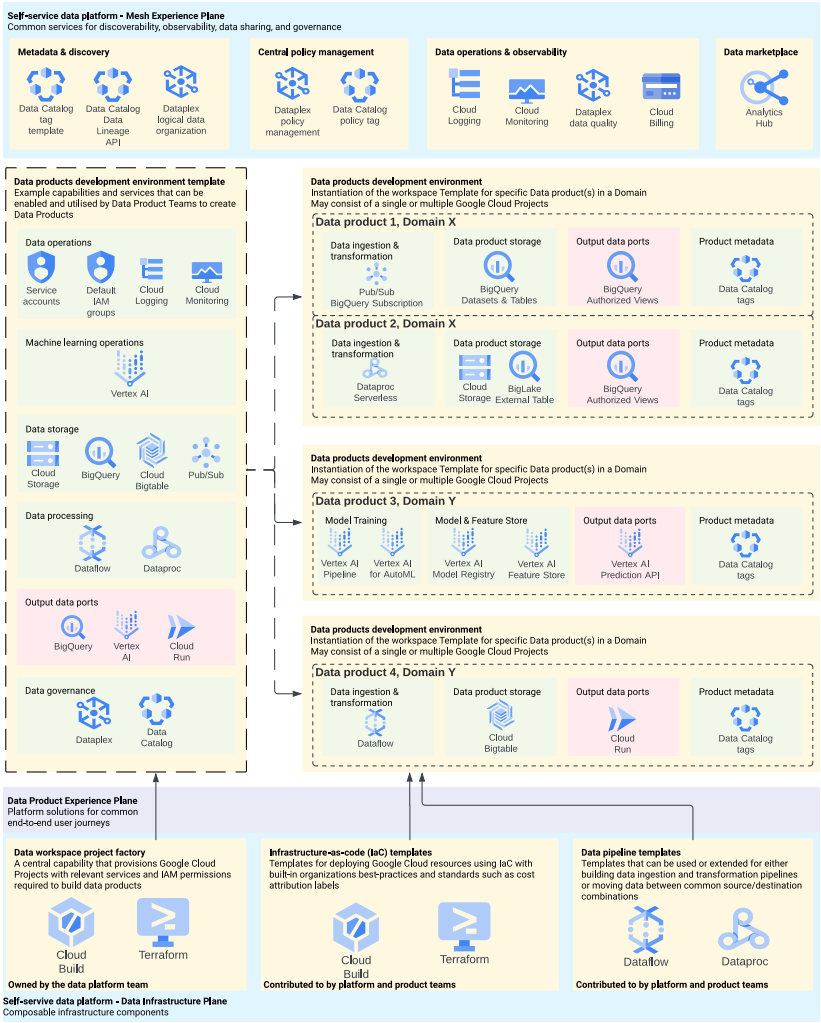

Figure 4 - Common services that can be used as part of the self-service data platform, as well as templates for different kinds of data products, using Google Cloud services

For the **domain teams** that will use this self-service data platform, the Google Cloud architecture in figure four can:

- **Enable rapid onboarding and time to value:** the data product development environment template and data pipeline templates can accelerate the onboarding and creation of new data products. So, when a domain team wants to start building data products within the Data Mesh they can leverage existing solutions and building blocks instead of reinventing the wheel for typical data pipeline architectures.
- **Simplify implementation of organizational standards and policies:** Organizational standards can be packaged into the Infrastructure-as-code and data pipeline templates to embed security and governance best-practices into development workflows and simplify security integration and collaboration through automation.
- **Promote best practices and foster a community of practice:** The pre-built templates and common services made available through the platform help teams follow Data Mesh best practice from day one, without having to go through extensive training or upskilling processes. As new domains are onboarded they too can contribute their own artifacts and knowledge to these templates which allows for natural growth as new data products are created.

Element #2: Building and managing data products

As soon as the core of the self-service data platform is built, domains can start getting into what Data Mesh is all about; creating and consuming purpose-built data products across a wide range of use cases.

The reference architecture in figure five below shows a simple example of a Data Mesh in Google Cloud, and how different types of data products (source-aligned, aggregate and consumer-aligned) can be used with one another across the mesh.

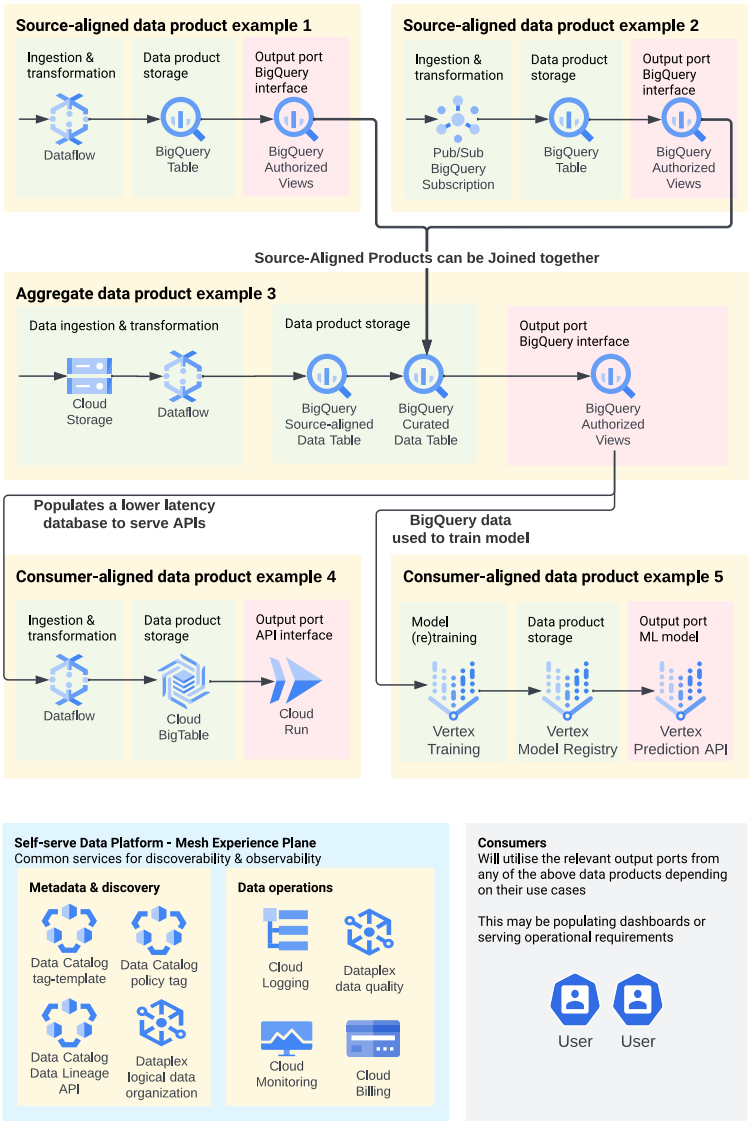

Figure 5 -Data Mesh architecture example with different types of data products (source-aligned, aggregate, and consumer-aligned). Highlighting different Google Cloud services that can fulfill analytics and Machine Learning needs

For the **Data Product Owners** who are responsible for creating and curating their own data products, the Google Cloud reference architecture above demonstrates multiple benefits:

- **Self-governed data product architecture:** Each data product can be built using the architecture pattern that best suits that use case then publishing the data to a suitable output port. Organizations can define multiple different types of architectural patterns that are in line with their standards and provide code templates to simplify and accelerate the data product build.
- **Easily combine data products:** Each data product can be accessed directly using its output port by any consumer if it suits their purpose. Data products can also consume data from other data products in the mesh. New aggregate data products or consumer-oriented data products can transform and aggregate information from other data products to facilitate common data querying patterns by consumers across the mesh.
- **Centralized capabilities co-existing with decentralized Data Mesh:** Whilst Data Mesh encourages decentralization for the creation and sharing of data products, there will continue to be some common functions that should be centralized for consistency. For example, defining common metadata patterns that users can follow to standardize discoverability makes it simpler for data product owners to publish their data products.

Element #3: Enabling computational federated governance

One of the biggest reasons why many large organizations look to build a Data Mesh is their ability to enable federated governance, and make it much easier to track and maintain high-quality, compliant data across diverse domains.

The reference architecture in figure six below shows how Google Cloud can enable metadata management in a Data Mesh, which is one aspect of federated computational governance.

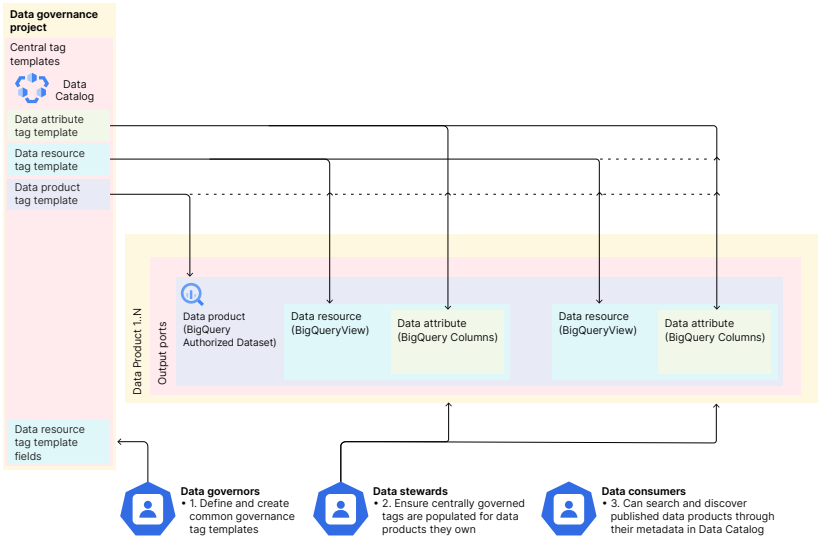

Figure 6 - How to tag templates managed centrally in Google Data Catalog, which can be used by federated data products to drive consistency and enable data governance activities

Examples:

Data product tag template fields include:	Data resource tag template fields include:	Data attribute tag template fields:
• Data domain • Data product description • Data product name • Data subdomain • Data confidentiality • Business criticality • Business owner • Technical owner • Documentation link • Access request link • Data product status • Last modified date	• Data sensitivity • Last updated time • Number of records • Number of fields • Data freshness • Global ID customer • Global ID account • Global ID location • Global ID product	• Sensitive data type • Fine-grained access control policies applied

Federated governance enabled by the architecture above can help **data governance teams**:

- **Consistently apply standards and policies across the Data Mesh:** Using Google Cloud Data Catalog, the central governance team can create detailed tagging templates for data products, resources (tables and views) and attributes (columns). By using common tag templates, governance teams can ensure that data product teams can easily and consistently tag their own data products using centrally-defined metadata standards.
- **Enable data discoverability and facilitate data understanding:** Google Cloud Data Catalog provides a single pane of glass allowing consumers to search across the Data Mesh to discover data products relevant to their needs. Metadata relevant to a product is readily available to help consumers understand how the data product can be used and how to request access.
- **Govern data through automation:** Aligning on the same metadata taxonomy allows the governance team to develop automation based on the metadata attributes on data products. Automation may include data management tasks that ensure data products are discoverable, addressable, trustworthy, self-describing, interoperable and secure.

When those elements come together, things get easier for everyone

Each Data Mesh architecture element has the potential to transform ways of working for the people closest to them. When you bring them together, the result is an architecture that makes everyone's data lives easier.

Domain teams are empowered to self-serve and build their own data products at speed, without having to retrain as data and architecture experts. Data product owners can easily create, iterate on and curate data products from the data sets they're closest to. Data governance teams can keep everything controlled, consistent and compliant, without limiting anybody's ability to derive value from their data.

That's the real value of Data Mesh in action.

Are you ready to begin your Data Mesh journey?

Data Mesh has moved past being a theoretical architecture approach. Today, there's a wealth of evolving best practice available to guide your strategic planning and domain onboarding journey. And all the elements needed to create an open, intuitive Data Mesh architecture are readily available in the cloud.

So, all that's left for organizations to do is determine whether a Data Mesh is right for them — and if they're ready to get started.

Who is Data Mesh for?

Figure seven below[2] shows a point in an organization's data journey where adopting Data Mesh could deliver significant value.

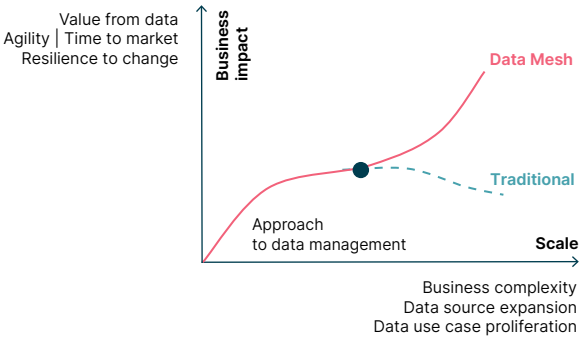

Figure 7 - Shows the inflection point in the approach to analytical data management

The organization in question is starting to see the value it's getting from its data, agility and ability to change plateau as it scales. If that trajectory continues, all three are set to fall below current levels as the company scales further.

To reverse the trend, the organization changes its approach to data management and builds a Data Mesh — setting itself on a new path towards exponential growth in the value it sees from data, its agility and enterprise adaptability.

2 Data Mesh: Delivering Data-Driven Value at Scale

Those goals will resonate with a lot of organizations. If you're looking for ways to turn data into value and make it more discoverable and accessible at scale, there's a good chance that a Data Mesh could be the right architectural approach for you.

But assessing how ready for it you are takes a little more reflective analysis. To help, we've created an eight-point self-assessment survey.

Data Mesh readiness self-assessment survey

This survey is split across eight areas to help determine how ready your organization is for Data Mesh. For each question, choose where your organization sits on the spectrum of responses — high, low, or medium if you're somewhere in between the two.

Chart your responses on the spider diagram below. The points closest to the middle represent the areas you need to work on before beginning your Data Mesh journey. But, if the majority of your responses are far from the middle, you should be in a good position to get started.

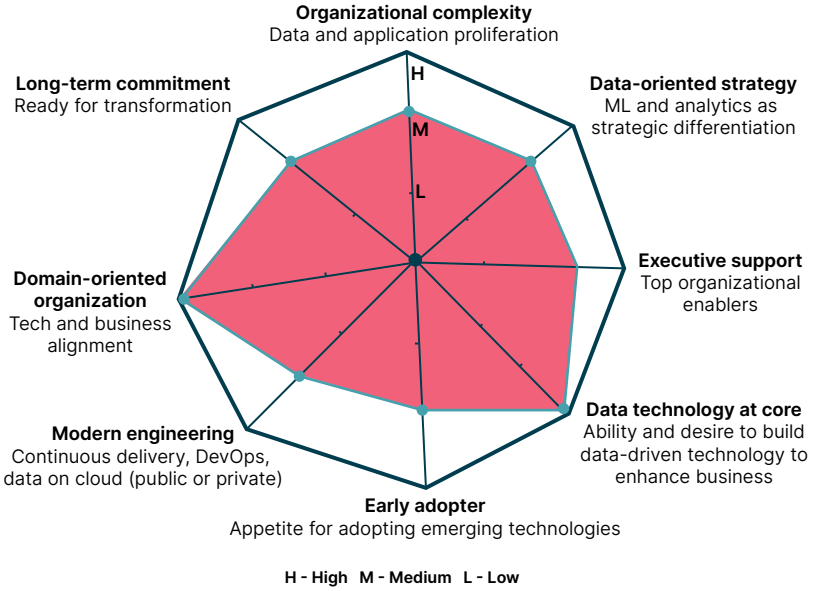

Figure 7 - Example of a Data Mesh readiness spider chart

Criteria #1: Organizational complexity

- **High:** My organization has a highly complex data and application landscape, where the two proliferate across every area of the business
- **Low:** My organization has a simple data and application landscape. Data rarely moves outside of the functions it originates in

Criteria #2: Data-oriented strategy

- **High:** Data-intensive capabilities such as machine learning and analytics are key to our strategy and are driving differentiation for our organization
- **Low:** Data is not critical to our strategy today, and we have no major plans to implement new data-intensive capabilities

Criteria #3: Executive support

- **High:** Our senior leadership team are passionate advocates for change, and they see Data and AI playing a key role in driving that transformation
- **Low:** It's very hard to secure senior buy-in for change and transformation projects in our organization or they don't believe Data and AI is a key driver for the digital transformation

Criteria #4: Data technology at core

- **High:** My organization has both the ability and desire to build new data-driven technology to enhance our business
- **Low:** Technology is only applied at the edge of our operations, and we have no plans to enhance our business with new core data-driven technology

Criteria #5: Early adoption

- **High:** My organization and the people within it have a high appetite for adopting new technologies before our competitors
- **Low:** My organization is risk averse, and we prefer to wait until technologies are widely proven before adopting them

Criteria #6: Modern engineering practices

- **High:** We have established and embedded modern engineering practices including Continuous Delivery, DevOps and Cloud, and they're now core parts of how our engineering teams work
- **Low:** Our teams still follow traditional engineering practices and ways of working

Criteria #7: Domain-oriented organization

- **High:** My company is made up of a diverse set of domains, each with their own digital priorities that our team works to align our technology with
- **Low:** My organization isn't aligned around many distinct domains or operate in a more centralized manner

Criteria #8: Long-term commitment

- **High:** We have a strong long-term mindset, and we can confidently commit to major transformation and change projects that span multiple years
- **Low:** We have trouble sticking with long-term transformation plans and seeing them through to their intended endpoint

Start your Data Mesh journey today

The Data Mesh era has begun. As more and more organizations embark on their own Data Mesh journeys, the full power and potential of the architectural paradigm are becoming clearer every day.

If you're ready to break down the bottlenecks created by centralized data architectures, empower domains to create and curate their own data products, improve data quality and accessibility, and streamline data governance, we can help.

By combining Thoughtworks' foundational work and unmatched practical Data Mesh experience with Google Cloud's powerful and intuitive toolset for building adaptive Data Mesh architectures, we can help accelerate and streamline your transformation journey.

To learn more about how you can build your own Data Mesh on Google Cloud and explore our reference architectures in more detail **here**. Or to start your own Data Mesh journey today, **book a free one hour consultation session with Thoughtworks today** (subject to Thoughtworks qualification).

About the authors

Danilo Sato, Head of Data & AI Services UK and Europe, Thoughtworks
Danilo is responsible for building high-performing teams to solve Thoughtworks client's most complex data problems. He leads technical projects in many areas of architecture and engineering, including software, data, infrastructure, and machine learning.

Thinh Ha, Strategic Cloud Engineer, Google Cloud Professional Services
Thinh specializes in Data and Analytics. He provides trusted advisory for Google's most strategic customers to help them make the best use of Google Cloud through deep product expertise, implementation experience and engineering capability. Thinh is a Global SME on Data Mesh and Data Governance.

Phill Radley, Principal Data Consultant, Thoughtworks
Phill specializes in shaping data strategies and guiding Data Mesh journeys. Before joining Thoughtworks he was Chief Data Architect at AstraZeneca (R&D) and BT; where he led their big data and AI initiatives.

Connor Lynch, Cloud Consultant, Google Cloud Professional Services
Connor specializes in Data and Analytics. He supports customers across EMEA with their Digital and Data Transformation journeys across all industries, recently working with a number of large organizations advising them on using Google Cloud to achieve their Data Mesh ambitions.

Elena Mata Yandiola, Cloud Consultant, Google Cloud Professional Services
Elena supports EMEA customers with multiple Data & Analytics and Machine Learning use cases. Her advise helps organizations to mobilize and adopt clearly defined next steps in the Data Mesh operating model, data products and technology streams through Google Cloud's technology and in-house developed frameworks.

Darren Young, Lead Data Consultant, Thoughtworks
Darren helps Thoughtworks' clients solve complex challenges and get the most value from their engineering and data environments. He has worked across a number of industries including pharma, media and mobility.

Thoughtworks is a global technology consultancy that integrates strategy, design and engineering to drive digital innovation. We are 12,000+ people strong across 50 offices in 17 countries. Over the last 29+ years, we've delivered extraordinary impact together with our clients by helping them solve complex business problems with technology as the differentiator.

/thoughtworks