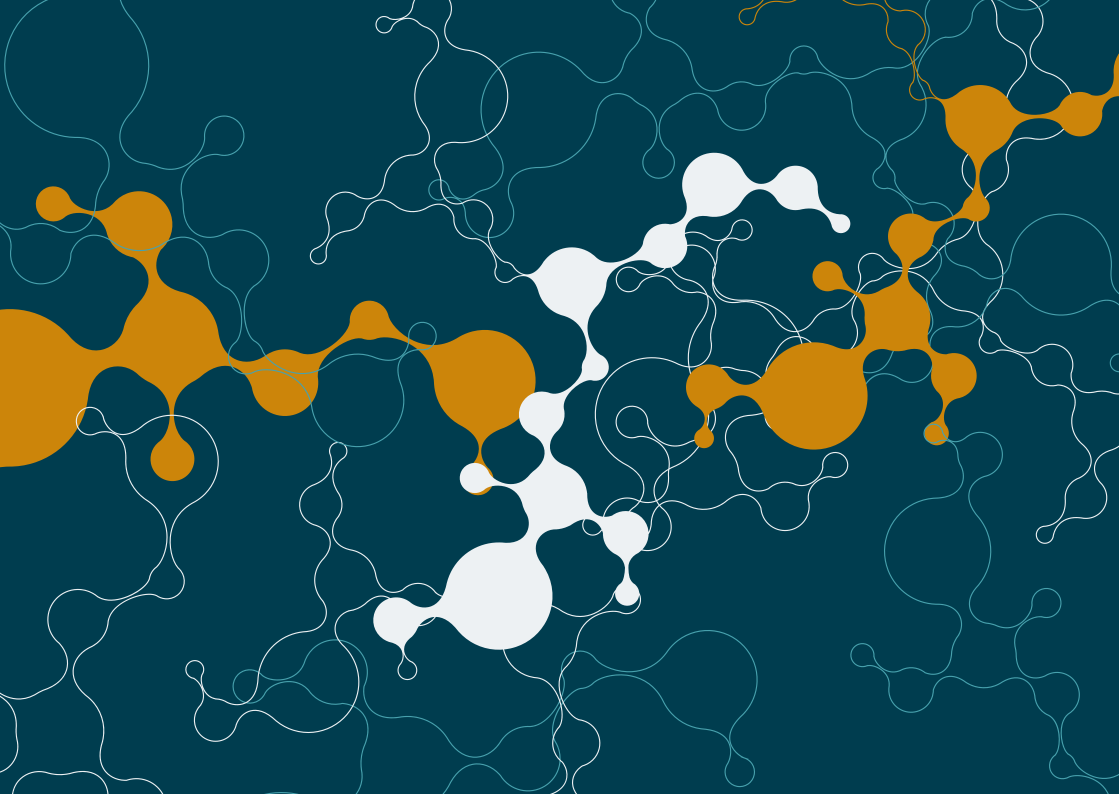


# The data mesh shift

## A new data architecture and paradigm to unlock business value

/thoughtworks

# Executive summary

Big data was supposed to transform the enterprise. Existing data architectures, however, have shortcomings that prevent them from delivering the promised business value. We are now experiencing a crisis of data confidence as a result. Conventional data architectures normally involve some form of central IT system.

Focusing instead on domains and data products allows us to avoid these types of silos and drive business value. Thoughtworks has developed just such a domain-driven approach rooted in organizational change called the data mesh. It represents a true paradigm shift — and an opportunity to successfully create a data-driven organization and unlock business value from data.

# Introduction

Not long ago, data mining was the hottest topic of the day. Companies were eager to get their hands on more data. Soon, the question would be what to do with it all. Many expected machine learning and AI to turn all those bits into valuable information to base decisions on.

What actually happened? Traditional approaches to generating big data insights have resulted in unhappy stakeholders. The reasons include data quality issues, problems with sourcing data engineering talent, overworked data platform teams, and endless and costly data infrastructure projects.

In this high-level decision-maker briefing, Thoughtworks describes the data mesh — a new paradigm built on proven software engineering principles. Implementing a new data architecture is, however, a considerable undertaking. How did we get here, and why can't we just optimize existing approaches?

The two major approaches to data management we see today rely either on data warehouse or data lake technologies. Data warehouses usually store structured data in queryable formats. A data lake, in its simplest form, stores raw data from various sources.

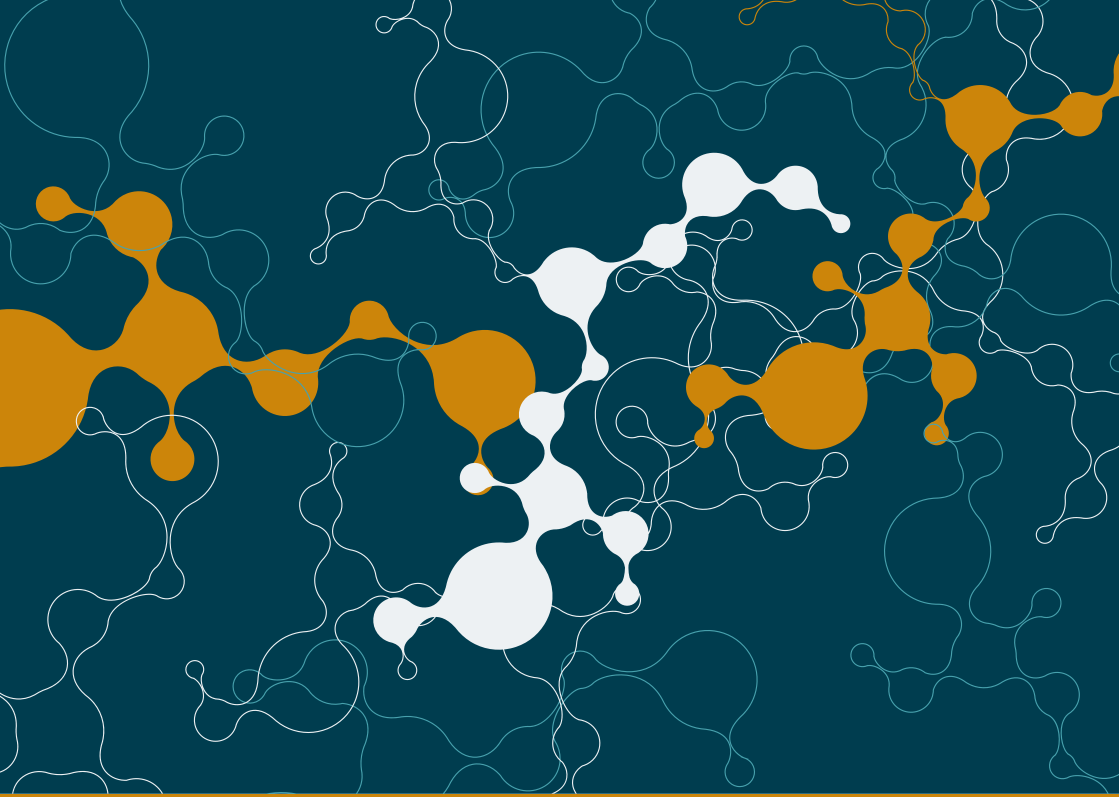
With this approach, the data can retain whatever schema the source system dictates until it is time to conduct an analysis.

A data mesh, in contrast, is a decentralized approach to data architecture. It applies the following principles borrowed from production IT architecture, but applied to data:

1. domain-driven distributed architecture
2. product thinking
3. self-service infrastructure platforms, and
4. federated governance

With it, businesses are creating so-called data products — decentralized expert offerings that focus on one domain and align data ownership and consumption. Working together and on top of each other, these data products exhibit a network effect which allows for an ongoing cycle of data, analysis, and action, resulting in a continuous flow of business value.

But existing data architectures are capable of collecting and analyzing data, too. So, why are they failing to live up to data's true potential?



# The data mesh paradigm

# The symptoms of big data dysfunction

Organizations experiencing data quality and value issues exhibit a common set of failure modes.

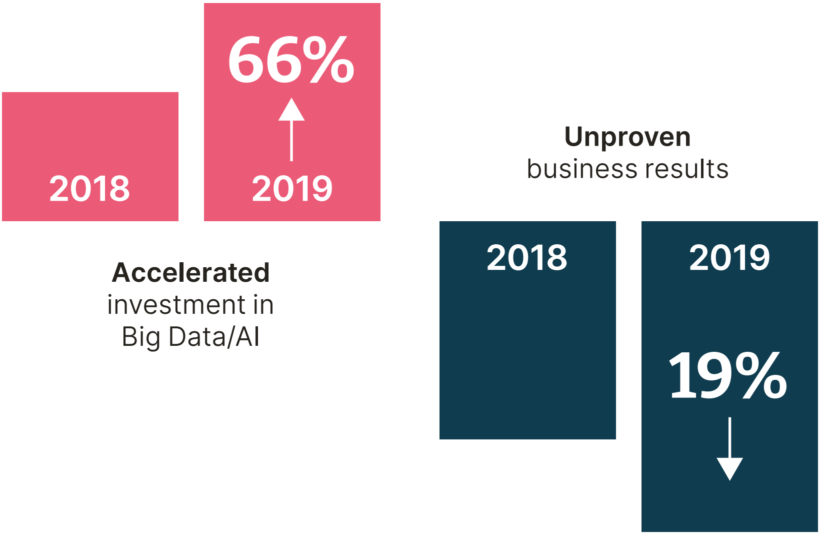
**Failure to bootstrap**
The envisioned use cases for data never get off the ground

**Failure to scale consumers**
The organization cannot keep pace with the needs of an increasing number of data consumers

**Failure to scale sources**
As more data becomes available within and outside the enterprise, sources cannot be integrated as quickly as they multiply

**Failure to materialize data-driven value**
Without alignment between data producers and data consumers, it becomes difficult or impossible to generate value

At Thoughtworks, we see the effects in our daily consulting practice. As investments in big data and artificial intelligence continue to grow, confidence in the business value of these investments is actually declining.



According to a NewVantage Partners study, big data continues to be a struggle for most enterprises. The survey reports that only 24% of firms claim to have succeeded in creating a data-driven organization. Only a similarly small minority of companies claim to have successfully built a data culture within their organizations. Yet the technology is not why they are failing. Over 90% cite people and processes as what stands between them and transforming their organizations through data.

## Why existing architectures aren’t enabling transformation

The difficulties organizations experience implementing data warehouses and data lakes have similar origins but occur at different points.

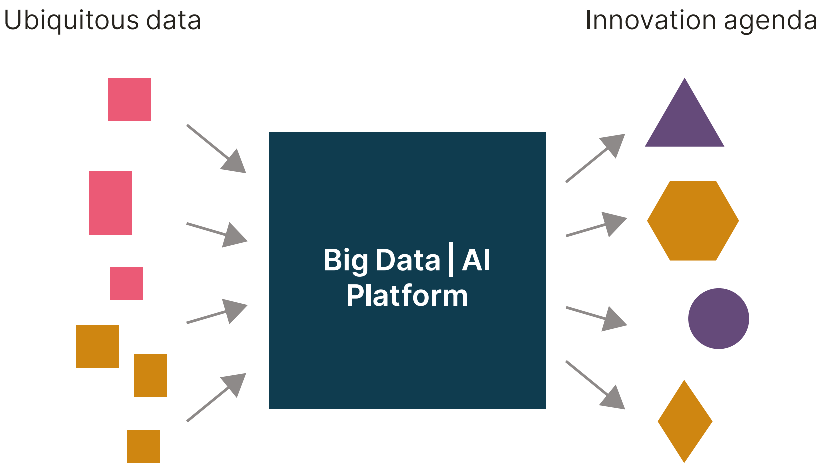
The goal of a data warehouse is to provide a structured data store with baked-in compute power to serve all the organization's needs. Yet the larger the enterprise, the less realistic this becomes. All but the simplest of domains will demand multiple bounded contexts and the corresponding data models. The overhead of data warehouse technologies makes adapting to these contexts more difficult. Even in the best-case scenario (using a single data model is feasible), data quality can still suffer.

Different systems within the IT organization often hold the same data. It may be difficult for central data engineers to know which one is the best to extract from. What's more, analytics scenarios can vary in the level of data quality they require. In both a data warehouse and a data lake model, that data quality is determined by actions that occur long before the affected data consumers come into play. And perhaps more crucially, individuals tasked with producing data belong to different organizational units than the consumers of that same data.

### A big silo and an insurmountable bottleneck

Both data warehouses and data lakes create what amounts to a gigantic silo that holds petabytes of data. Of course, these architectures are designed to enable access to all the enterprise's data. But unlike other silos, the problem here is not that data is locked away due to technical constraints. The issue is an organizational one. Namely, the teams that operate these monolithic data repositories must be populated with hyperspecialized data engineers.

### Centralized | Monolithic



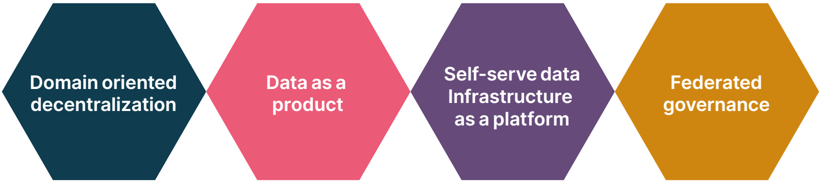
Talented data engineers are hard to find and expensive to hire. Recruiting the right people to create and run a data lake thus presents an enormous hurdle. But even if a company is able to recruit sufficient talent to build and operate its data lake, the problems don't end there. Because those engineers will be tasked with getting data from people and teams across the organization that have little incentive to ensure they are sharing only correct, trustworthy, and meaningful data.

Once that data has been acquired, the data engineering team now has the unenviable task of making it useful to the rest of the organization. Without any domain expertise to guide them. Or input from a continually rising number of data consumers.

# A paradigm shift that goes beyond technology

To remedy this situation, we must move on from the current system-oriented paradigm. If we do not, the inherent disconnect between data producers and data consumers will remain, as will the giant silo, recruiting bottleneck, and resulting failure modes.

Pillars of the **data mesh** paradigm shift


Domain oriented decentralization
Data as a product
Self-serve data Infrastructure as a platform
Federated governance

Instead of continuing to think of data as a by-product of other business functions, it is time to recognize that data has long been a product in its own right. With this new perspective, we can move on from monolithic systems and linear pipelines. In their place, an understanding of data as destined for specific consumers informs how we structure not only the supporting architecture, but also the organization itself.

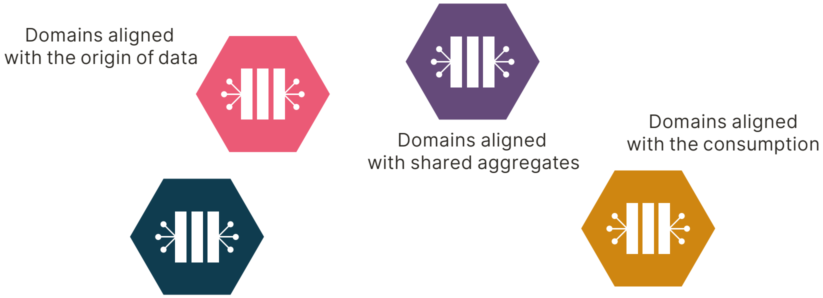
Most data scientists spend 80% or more of their time doing data discovery and extraction. But what if business activities were designed with data in mind from the start? What if those same data scientists were fully involved throughout the entire lifecycle of the data?

## Transitioning to domain-driven data products

Choosing the domain, instead of a monolith, as the organizing principle for big data allows us to combine domain expertise with the technological capabilities necessary to generate business value. Viewed through the lens of the domain, data can become a portfolio of discrete products. Any successful product must delight its consumers, in this case the wider organization: data analysts and anyone else who needs to work with it.

How do you create products that delight? By drawing on the wealth of knowledge product thinking provides. Furthermore, when organizations establish distributed data teams with baked-in domain expertise, they eliminate most of the friction surrounding data extraction, cleansing, and analysis.

**Decompose data around domains**
Distribute the ownership



## How to define data success

For this decentralized, domain-oriented approach to succeed, several prerequisites must be met. The data products need to be:

- **Shared / discoverable**
- **Addressable**
- **Interoperable**
- **Self-describing**
- **Trustworthy**
- **Secure**

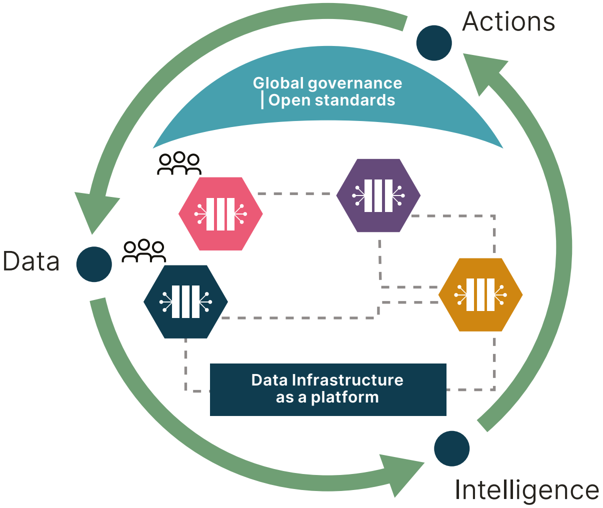
Once these qualities have been achieved, the model will be able to scale.

The true measure of success of any data product is always how happy data consumers are. If that is the case, it can also make sense to define the product's success in clear and measurable terms. To this end, we can provide outstanding documentation and publish quality metrics. One obvious performance indicator to track, for instance, would be the lead time necessary for a data scientist to find the relevant data and use it.

## Connecting interoperable data to create a cycle of intelligence

One of the core features of the data mesh is its federated governance model that achieves interoperability through standardization. Only with interoperable data can analyses involving multiple data products lead to valuable insights and action. These, in turn, influence the next cycle of data, establishing a connected cycle of intelligence.

### Execute through iterations of connected intelligence



## A leading internet retailer starts building its own data mesh

A leading German e-commerce site recently engaged Thoughtworks to help implement thin slices of the data mesh for specific domains such as web tracking and customer lifetime value. The project resulted in a number of key insights.

One central realization was that data producers must feel they are responsible for their own data. Conscious decisions also had to be made about what data should be stored. Because extracting and transforming data that will never be used only generates unnecessary costs. After deciding which data to offer stakeholders, the new domain-driven teams now have the responsibility to maintain and serve those users. In this way, data quality has become a contract between data consumers and producers.

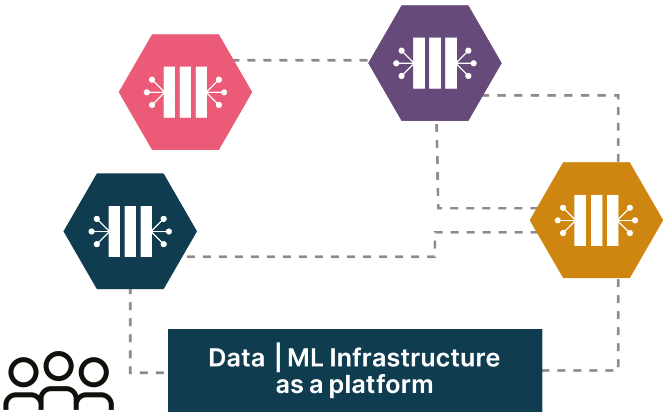
In moving from centralized ownership of data to a decentralized model, the company also saw a solution to its data engineering bottleneck. But the solution was essentially an organizational one. Before, no one had complete responsibility for or ownership of domain data. The new data products are source-aligned. Ownership now remains within the domain for the entire life cycle of the data product.

## What about existing infrastructures?

In introducing the concept of the data mesh, one concern we have encountered is that it could make recent investments such as a data lake obsolete. Another is that, in a distributed system, each data product would require its own separate infrastructure.

The data mesh resolves these issues by offering data infrastructure as a platform. Instead of requiring each domain team to engineer its own data platform, the necessary infrastructure is provisioned from a self-service platform. This gives the teams a high degree of autonomy while also allowing the integration of central assets such as an existing data catalog.

**Enable autonomy**
Abstract technical complexity in
self-serve data infrastructure



## Implementation and organizational change

In the data mesh, data producers and data consumers should be working together as closely as possible. From an organizational perspective, the ideal situation is when the same team is both producing and consuming the same data, uniting data responsibility with capability. Often, however, the many duties of data producing teams require splitting the roles across two teams, producers and consumers, that remain in direct communication.

So it's not just the technology stack that must change. Responsibilities and structures must also shift to implement the data mesh. This change process requires buy-in from the highest levels in the organization. This transformational change can be achieved incrementally by moving toward implementation in thin slices. In the interim phase, before an infrastructure-as-a-service platform has been completed, teams form around domains, using a data warehouse or lake as an intermediate source if necessary.

And yes, creating the infrastructure-as-a-service platform does require the exact same data engineering skills that frequently become a bottleneck in a data warehouse or data lake architecture. Once the platform has been established, however, it divorces domain knowledge from infrastructure. Data engineers no longer must dive into the domain knowledge to do their jobs, alleviating the pressure monolithic systems create.

# Conclusion

Implementing traditional approaches to data management often feels like running around putting out fires: trying to solve quality issues with more quality control, trying to solve data platform bottlenecks with more data engineers, trying to support a growth of data sources with more powerful infrastructure. The beauty of the data mesh is that it teaches us to look at the problem from an entirely different perspective. Set up in the correct way, a data mesh gets better the more data sources there are and the more data consumers there are. Instead of creating more and more problems, a working data mesh unleashes business insights from data, the more collaboration around data is happening.

Implementing a new data architecture can be a turning point in the growth of a business or a personal career. Approaching big data as a problem that can be solved with technology alone has limited the value existing strategies can generate. Success depends instead on adapting organizational structures to align the incentives of data producers with those of data consumers. Product thinking gives us the tools to make excellent data products that can unlock the true potential of big data. Implementing the data mesh does not require a big-bang migration. You don’t have to abandon your existing data lake. Instead, adoption can proceed in thin slices with each one contributing insights to the next.

Perhaps your company has experienced difficulty scaling data sources after migrating from a data warehouse to a data lake. You may simply be looking for a faster route to data ROI or a competitive advantage. Whatever scenario you are facing, a proven way to get results faster is bringing in experts who have experience with the specific data architecture you seek to implement.

Thoughtworks has worked with a number of enterprises on their data mesh implementations. One observation that has proven universal is: the sooner an organization gets started with data mesh, the faster it can start generating more value from its data. Although most IT departments already possess a great deal of the expertise required to implement a data mesh, external partners can accelerate and guide the process with best practices.

# References

Dehghani, Zhamak. 2019. “How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh.” ***martinFowler.com*** **http://martinfowler.com/articles/data-monolith-to-mesh.html**

Dehghani, Zhamak. 2020. “Data Mesh Principals and Logical Architecture.” ***martinFowler.com*** **https://martinfowler.com/articles/data-mesh-principles.html**

Fowler, Martin. 2015. “DataLake.” ***martinFowler.com*** **https://martinfowler.com/bliki/DataLake.html**

New Vantage Partners, 2021. “Big Data and AI Executive Survey 2021.” ***newvantage.com*** **http://c6abb8db-514c-4f5b-b5a1-fc710f1e464e.filesusr.com/ugd/e5361a_76709448ddc6490981f0cbea42d51508.pdf**

Wider, Arif. Nov. 16, 2020. “Data mesh: it’s not just about tech, it’s about ownership and communication” ***thoughtworks.com*** **http://www.thoughtworks.com/insights/blog/data-mesh-its-not-about-tech-its-about-ownership-and-communication**

Pallozzi, Daniel. Dec. 13, 2018 “The end of data gluttony: Principles to rejuvenate your data strategy” ***thoughtworks.com*** **https://www.thoughtworks.com/perspectives/edition2-data-article**

## About Thoughtworks

Thoughtworks is a global technology consultancy and community of passionate purpose-led individuals, 9,000+ people strong across 48 offices in 17 countries. Over 27+ years, we have helped our clients solve complex business problems by integrating strategy, design and engineering to drive digital innovation.

For more information visit: thoughtworks.com

/thoughtworks