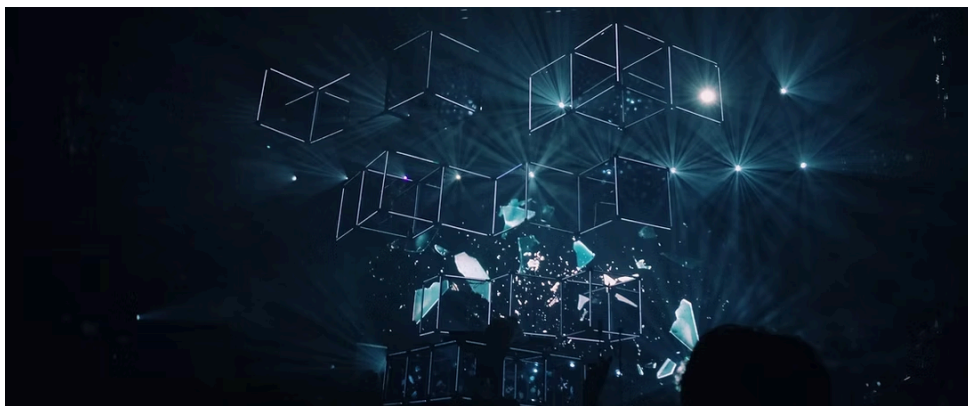Sep 13, 2022 · 6 min read

Data Mesh: A Paradigm Shift to Make Your Data Great Again

Updated: Jan 12, 2023

Imagine: you are an insurance agent dealing with insurance claims. You often have to pull tons of data from other departments in order to review these. The company has built a giant, all-encompassing **data warehouse** to collect, process and share everything, but as your business and clients grows, the waiting time for getting up-to-date data is also getting longer.

The company tried to introduce machine learning models to saving your precious decision time, for example, detecting possible frauds. The models regularly require latest training data to keep up with the client pattern, however, and the time to wait for cleaned-up datasets is just as long. Worse, the data team who is responsible for the data cleaning understands very little about the insurance domain. You wasted more time to communicate with them and improves almost nothing.

Your company stops growing, lose clients even, because people thought your company is either trying to stall the payment or simply does not care. Both the bottleneck and bad reputation becomes a vicious cycle that is hard to get rid off.

Data mesh, a high-level logical architecture first proposed by Zhamak Dehghani at Thoughtworks in 2019, is in fact trying to answer a pretty simple question: how do we solve bottlenecks in modern centralised data warehouses?

According to the KPMG 2020 data strategy survey, 43% of the respondents' organisation does not even have a data strategy. For those who have, 39% are using centralised data repository (the percentage rise to 60% for larger companies). Data warehouse has became the key for business intelligence (BI) for the past three decades, but as an organisation grows to a certain scale, the traditional **ETL** (*Extract, Transform, Load*) processes simply cannot keep up anymore.

> The ETL bottleneck is just one bottleneck holding back the full exploitation of data, but it is the most important one…the result of ETL and other bottlenecks is that the amount of data being stored is rising but the percentage of data being put to use is falling.
>
> – Forbes: How The ETL Bottleneck Will Kill Your Business (2016)

The data lose insight and value, and the users are unsatisfied from waiting. Unscalable data means unscalabe customers. Even **data lakes** - dump unprocessed unstructured data first, which is quicker and cheaper, then output processed format via ETLs to data science and machine learning usage - can easily become *swamp*s that trap everything inside.

Why, we are already splitting monolithic software applications to smaller, containerised **microservices**. There is even a trend called ***micro-frontends***, which is to have smaller teams each managing one domain component of the website. Can't we do the same to how we produce and consume data?

The Four Principles of Data Mesh

In her groundbreaking articles, Zhamak Dehghani listed the four underlying principles of data mesh:

- *Domain-oriented decentralised data ownership and architecture*
- *Data as a product*
- *Self-serve data infrastructure as a platform*
- *Federated computational governance*
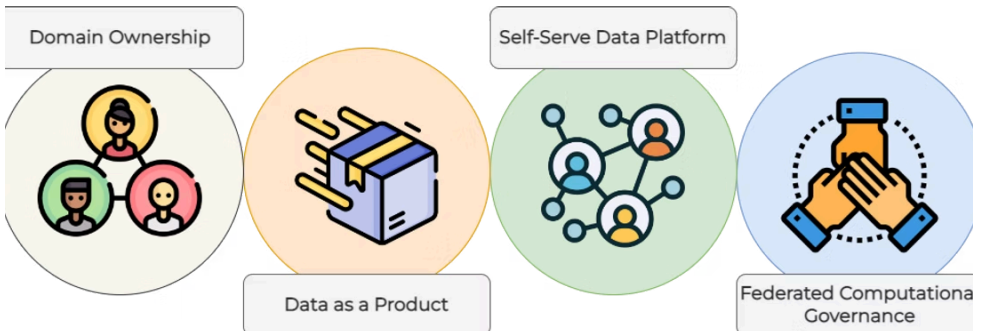
The four principles of data mesh

More importantly, these principles are built upon two key concepts:

> Data mesh, at core, is founded in *decentralisation* and *distribution of responsibility* to people who are closest to the data in order to support continuous change and scalability.
>
> – Zhamak Dehghani: Data Mesh Principles and Logical Architecture (2020)

By granting power and autonomy to people who create and maintain data, data can be created faster with better quality (because, you know, people *do* care if you trust them). The bottleneck is bypassed and the data processing loading is distributed and thus become scalable.

We will take a more detailed look at each of the principles in our following blog posts - but here we'll give you an quick overview:

Domain Ownership

The first step is to have **domain teams** control their own data and decide the way how data can be used by other teams - including providing code interfaces like API endpoints. This encourages more data to be created and used while staying trustworthy.

Data as a Product

The data has to be treated as a product - the domain team has to figure out who use their data, how to discover them, and the ways to improve the usefulness and quality of data. The **data products** in fact contains data pipelines that generates the data and metadata.

Self-Serve Data Platform

In order to enable data products being autonomously deployed, run and discovered by teams, a high-level abstraction platform on an infrastructure is necessary. The platform also has to be easy and safe enough to use without requiring expert software knowledge.

Federated Computational Governance

Who can use what data products, and which can be combined in what way? The decision can be either from the domain team or a federation of teams. The data product policies have to be able to run automatically on the platform itself.

Bringing Data Mesh to Reality

From the very beginning, FST Network's **Logic Operating Centre (LOC)** is deeply influenced by the four data mesh principles. LOC itself *does not* create data mesh, of course - but it is designed to help you make the transition to data mesh paradigm a lot *easier*.

We will talk more about these topics in the following articles, but here's some quick sneak peaks:

- LOC is deployed on **Kubernetes**, which is a distributed environment itself and can be run almost anywhere and on anything, from third party clouds or on premise (private) servers.
- Users can build and deploy data processes (data pipelines or data products in LOC) in a extremely easy way that is very similar to creating microservices on **FaaS (Function-as-a-Service)** platforms.
- LOC also provides ways to discover data processes as well as implementing access control (data policies).

Back to our insurance company example. If the company adopts LOC, every department will now be able to create and serve their own data, on a distributed yet logically unified platform (the data fabric). You can easily discover everything, integrate endpoints into your own data pipeline, and get the review job done more quickly. Even AI models can be enormously benefit from this - getting high-quality datasets directly from the source, and output predictions via another data product (discoverable by other LOC users.).

> Just 13% of organisations excel at delivering on their data strategy… They are succeeding thanks to their attention to the foundations of sound data management and architecture, which enable them to "democratize" data and derive value from machine learning.
>
> – MIT Tech Review, Building a high-performance data and AI (2021)

In the end, what data mesh is trying to achieve is pretty simple: to make your data agile, accessible and trustworthy (again), by applying a decentralised approach to break up bottlenecks formed by the data explosion, and to give people the ability and trust to create more value. Data has already be named as the most valuable commodity of businesses - thus how to get more value and insight from them should be your very first concern in the brave new data age.

Data Mesh is *not* the End - But a Necessary Shift

Interestingly, Gartner has listed data mesh as one of the "on the rise" technologies in mid-2022, but also pointed out that it may be obsolete before reaching the plateau of productivity.

Hype Cycle for Data and Analytics Programs and Practices, 2022
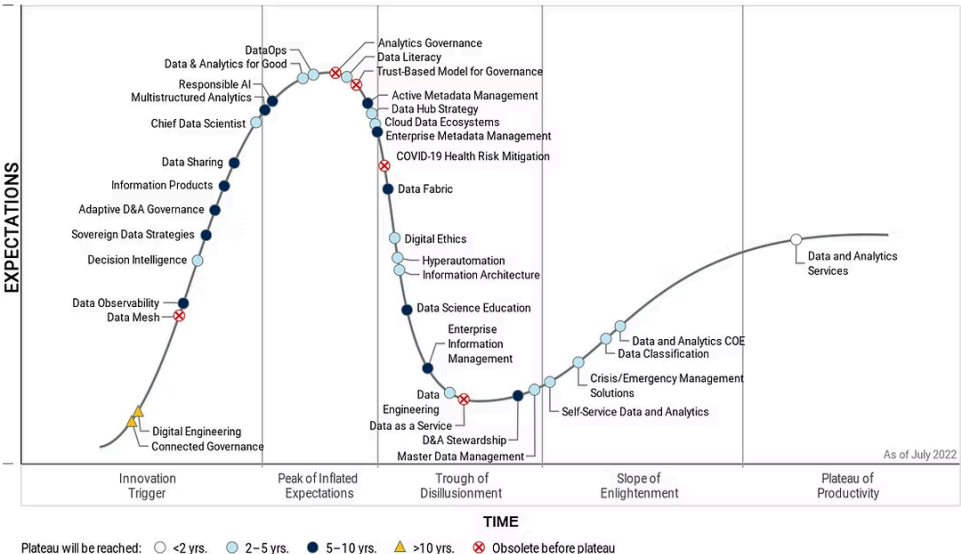
This doesn't mean *data mesh is already dead* - instead, data mesh may very well be transformed or combined with something else in the next few years. It is still one leading force of the great data decentralisation movement, and has already influenced some major changes.

For example, data mesh has successfully revived another old concept from 2010s, **data fabric** (a relative of **data virtualization**), which is to integrate data from separate silos into a continuous, seamless *fabric*. This is not that different from **data lakehouse** - to create a distributed access and governance layer upon data lakes, and to avoid the centralised ETL problem as well. Both are deeply focused on improving data analysis and machine learning, which are present in most of Gartner's rising data tech trends. The AI age is already here, and you'll need *lots* of data to use it.

But how exactly can you get there? In fact, data mesh is about this kind of paradigm shift to make data more agile and accessible, as well as being trustworthy. And this is what we are hoping to achieve for you, by creating the Logic Operating Centre.

LOC provides some features that point at the same directions:

- The executed results (data products) of LOC data processes are stored in a Amazon S3-compatible unstructured database, which is essentially the "lake" in LOC. The data can be read later if the data processes are invoked asynchronously.
- The clients can also deploy data processes to query results from other processes, along using various external data sources - relational databases, S3, FTP server or HTTP endpoints.
- Data processes can generate execution history logs, audit logs and *user-defined data events* - which is also referred as **data lineage** or **active metadata**, another popular data trend in recent years. Both data processes and data events are discoverable in LOC.
- Data processes can also have access controls and resource quota (computational governance).

Like we've mentioned before, we will go through these exciting features in our following articles. Stay tuned to FST Network's brand new blog series to find out more!

FST Network aims to revolutionise the data landscape by creating world's first cloud-native and serverless Data Mesh platform. Logic Operating Centre (LOC) is fully capable to productise and eventify your data, with easy-to-deploy data pipelines that seamlessly connects everything across your organisation, bringing in agile, accessible and trustworthy business insights. The data race is constantly changing and fiercer than ever - but we are there to give you a winning edge.

Icon credits: flaticon.com