Brought to you by:

Starburst

Data Mesh

for dummies®
A Wiley Brand

Elevate the value of data as a product

Streamline data flow through the organization

Enable autonomous data management domains

Colleen Tartow, Ph.D.
Andrew Mott, MBA
Adrian Estala

Starburst
Special Edition

About Starburst

Starburst is the analytics engine for all your data. We provide the fastest, most efficient analytics engine for your data warehouse, data lake, or data mesh. We unlock the value of distributed data by making it fast and easy to access, no matter where it lives. Starburst queries data across any database, making it instantly actionable for data-driven organizations. With Starburst, teams can lower the total cost of their infrastructure and analytics investments, prevent vendor lock-in, and use the existing tools that work for their business. Trusted by companies like Comcast, FINRA, and Condé Nast, Starburst helps companies make better decisions faster on all data.

starburst.io

Data Mesh

Starburst Special Edition

by Colleen Tartow, Ph.D., Andrew Mott, MBA, and Adrian Estala

for dummies®
A Wiley Brand

Data Mesh For Dummies®, Starburst Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-91088-6 (pbk); ISBN 978-1-119-91089-3 (ebk)

Publisher's Acknowledgments

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub.

Development Editor: Faithe Wempen
Project Manager: Dan Mersey
Acquisitions Editor: Ashley Coffey
Senior Managing Editor: Rev Mengle
Managing Editor: Camille Graves
Senior Client Account Manager: Matt Cox
Production Editor: Saikarthick Kumarasamy

IN THIS CHAPTER

- Introducing Data Mesh
- Learning how Data Mesh is relevant to today's business needs

Chapter 1
Discovering Data Mesh

New to Data Mesh? Start here! In this chapter you'll find out what it is and why it's recently been rising in popularity.

What is Data Mesh?

Data Mesh is a relatively new approach to data management. It combines several important trends in data management, including domain-driven design and data as a product, to decentralize the ownership of ingestion, processing, and serving of data. Zhamak Dehghani defined the term in 2019 as "a decentralized sociotechnical approach to share, access, and manage analytical data in complex and large-scale environments—within or across organizations." Data Mesh aims to overcome the limitations of traditional centralized data management approaches that use data warehousing and data lakes.

In a traditional data management approach, three functions are essential for providing data for end-users: data ingestion, data transformation, and data serving. Traditionally a separate centralized team is responsible for each of these functions, meaning that if you want to scale the overall process, you must scale the entire centralized team. In a Data Mesh, a single team (*domain*) is

responsible for all three functions. Each domain focuses solely on data from a specific business area. You can scale domains as needed to address new analytic requirements or to onboard new data.

Why Data Mesh Now?

REMEMBER

With today's increased data volume and expectations of nearly instantaneous access, centralized data management is no longer doing the job, because it can't provide the agility and performance organizations need. In the context of data and analytics, *agility* is an organization's ability to respond efficiently and swiftly to changes in external market conditions, fluctuating regulatory requirements, and internal organizational changes. A key factor for organizations to adopt a Data Mesh strategy is so that they can survive and thrive in the face of ever-changing conditions. Consider the following business drivers:

- **Market Volatility:** Market shifts will impact IT budgets. You'll need to optimize your analytics costs and prioritize budgets across domains.
- **New Business Models:** Emerging digital business models often start with the discovery of data sets and novel ways to analyze them. Data Mesh exposes your data to drive innovation, inventive products, and enhanced tools.
- **Competitive Pace:** Organizations that are leading the digital transformation journey are also leading their respective industries, and undergoing activities such as mergers and acquisitions. Data Mesh can help you realize strategic benefits sooner by quickly connecting these new decentralized sources without a full migration.

With Data Mesh, the activities performed by a centralized data team are replaced with empowered, distributed domains. Each domain operates in an agile, self-sufficient, and autonomous manner.

A decentralized data strategy is a thoughtful response to an ever-increasing need for data to support decision making. Data Mesh is the next evolutionary step in a modern data management strategy.

TIP

By implementing a decentralized data architecture and organization, businesses can benefit by shortening the path between data sources and the value created by that data.

IN THIS CHAPTER

- Understanding the four principles of Data Mesh
- Creating and maintaining high-quality data products
- Learning how to optimize for maximum value

Chapter 2
Implementing and Optimizing Data Mesh

In this chapter, you'll find out how Data Mesh works, and learn out how it might work for your organization.

Identifying The Principles

Data Mesh is founded on four key principles: domain-driven ownership, data as a product, self-serve data platforms, and federated computational governance. Here's more about each of those.

Domain-driven ownership

A *domain* is a collection of people with relevant knowledge organized around a common business outcome or data subject area. A domain might include experts in a particular field, analysts, and data engineers who are focused around that specific data subject area. Domains can mirror organizational units such as finance, operations, and marketing, or they can represent subject areas such as customer or product.

A domain owns and is responsible for transforming and serving data to consumers via data products related to a specific area. Domains do not concern themselves with the minutiae of the data technology and infrastructure; their focus is on creating data products.

Data as a product

The *data product* is the heart of a modern data management strategy. A data product typically includes these components:

- **Data:** Structured information stored and organized in files, tables, views, and streams.
- **Metadata:** Data about the data, such as columns, definitions, numbers of rows, refresh patterns, security policies, and observability metrics.
- **Access patterns:** End-user instructions for how to query the data and what engine to use.
- **Code:** Code used to create the data within the data product.
- **Infrastructure:** Tools and platforms used to create data products.

REMEMBER

A data producer can create one or more data products, each of which is made up of one or more datasets. In this context a dataset might be a table, file, logical view, or a materialized view.

Producers should consider consumer needs as they create data products. Using a universal data language such as SQL enables producers to create and edit data products with ease, while data consumers can view, understand, and produce insights using their corresponding toolset. Relevant metadata such as usage metrics, sample queries, social commentary, versioning, and data lineage can streamline development and consumption as well.

Self-serve data platforms

Each domain needs a data platform to ingest, process, and serve its data, but the domain team members do not manage that underlying data platform themselves. A key pillar of Data Mesh is that a clear separation exists between data and infrastructure ownership.

Centralized data teams provide all the domains with a *self-serve data platform*. This platform enables the domains to focus on

building high-quality data products that ultimately garner business value in the form of data analytics.

The self-serve data platform should be domain-agnostic. A domain requiring additional technical capabilities can work with the central IT team to develop those capabilities, but the onus remains with the central IT team to provide the data platform.

Federated computational governance

Traditional centralized data governance can get in the way of using data to create value. In a Data Mesh, the domain teams are empowered to manage some elements of governance based on their unique value and risk position.

With Data Mesh, governance standards are set at the federated level (meaning at a level above the domains) and are implemented as policies computed as part of the data product. Policy enforcement is automated, making this a *federated computational governance* model.

Governance is composed of many business-critical capabilities, including:

- **Security:** Are the right people authenticated and authorized to use the data?
- **Compliance:** Does the data follow all required policies, such as the right to be forgotten in the General Data Protection Regulation?
- **Availability:** Can authorized users access the data?
- **Quality:** Is quantifiable information about the data's quality available to users?
- **Entity Standardization:** Do all domains have a common understanding of terminology?
- **Provenance:** Is it clear where the data came from and who is responsible for it?
- **Usage:** Who is using the data? How are they using it?

REMEMBER

Each data governance area can be defined on an inter-domain or intra-domain basis, depending on the organization's needs. That distinction is important, but varies between different organizations that adopt Data Mesh based on their specific

industry and business requirements. Governance should maximize value and manage risk in whatever way makes the most sense for each organization.

Diving Deeper into Data Products

The key to success for any company lies in deriving maximum business value from its data in a robust, scalable, and timely fashion.

In theory, all four principles of Data Mesh are equally important in this effort and should be given equal weight from the outset. However, there are significant cultural, organizational, and architectural considerations that arise alongside the first two principles: domain-driven data management and treating data as a product. To enable success in implementing these two principles, the third and fourth principles (self-serve data infrastructure and federated computational governance) become a necessary component of an organization's data strategy.

A huge part of a successful data and business strategy is treating data as a product, where data is managed, valuable, and usable.

Data products are the heart of the Data Mesh. By initially focusing on data products, organizations will be able to create value from data quickly while simultaneously working through the organizational and architectural changes required to support Data Mesh.

By creating useful and effective data products early in the Data Mesh journey, organizations can benefit from the principles of Data Mesh while simultaneously navigating the more complex and lengthy changes required to implement the other three pillars.

Identifying the key qualities of an effective data product

REMEMBER

Data products that are useful and serve downstream consumers well must be:

- **Discoverable:** End-users and other domains must be able to discover and access a given data product.
- **Addressable:** The data should have a straightforward and documented way of being programmatically accessed.

- **Trustworthy:** End-users should be able to understand the level of data quality so they can be confident in the results of its analysis.
- **Self-describing:** Any end-user outside the domain that produces the data product should have all the information they require to use the data.
- **Interoperable:** Governance should ensure that the data complies to any inter- or intra-domain standards or regulations, so the end-user can confidently use the data without concern.
- **Secure:** Data products should be designed so that access control and security are part of the definition of the data product, and are enforced by the self-serve platform.

TIP

Choose technologies that support these qualities so that data product creators can focus on business logic and curating datasets that provide the greatest value to the downstream consumers. It should be simple for data product consumers to use the data products, so they can focus on using the data to create business insights.

TECHNICAL STUFF

An effective data product applies business logic to transform operational data into analytical data. As with anything in modern development, the fewer steps involved in this transformation, the better. In legacy data ecosystems, data is always copied and curated to an analytics-focused data storage format. Modern data architectures allow operational and analytical data to be curated and exposed to downstream users without necessarily creating multiple copies. Ideally, a modern data architecture allows users to build data products based directly on operational data, analytical data, or a combination of the two. This flexibility results in faster and simpler development to create high-value data products for a better overall downstream consumer experience.

Creating a data product

Modern development paradigms such as Agile can be quite useful in creating data products. The principle of "start small and iterate quickly" applies well. An initial data product should be created based on expected consumption patterns and then evolved with empathy for the consumer over time.

Your consumer should be part of the data product design activity so that you understand what they need and how they intend to use it. It is essential to provide data products that are easily accessed and understood by downstream consumers, without translation or external assistance.


TIP

A data product doesn't need to be based on a curated, gold-plated, perfect dataset — particularly not in the beginning of your Data Mesh journey. A simple representation of the data from a source system may be relevant and valuable for a specific set of consumers.

When you are starting your data product journey, focus on a few small products and one or two representative consumers. Over time, you may have many data products that are organized and discoverable from a catalog.

Ideally use a single platform where data products are published and consumed. However, if multiple platforms are required, then interoperability between those platforms should be seamless. For example, the platforms should be able to share a common metadata repository.

Reuse and interoperability are tremendously important to accelerate the speed of your analytics projects and to increase your ROI. The same data product can be used to drive a BI tool, a predictive algorithm, or an automation engine. That might be too ambitious for your initial data management efforts, but this is how you should envision products being used across the organization. It's equally important to ensure your consumer has the necessary data skills to work in a self-service manner.


TECHNICAL STUFF

There are multiple ways to access data, but the most common and nearly universal way these days is with SQL. SQL is the modern *lingua franca* of data, most modern query engines support it, and nearly every data worker will be fluent in it beginning early in their career.

Maintaining a data product

Once a data product is created and available for downstream users, the domain may iterate on it as feedback comes in, just as with any other product the domain produces. If a data product changes, they must notify any downstream users of those

changes in advance. Such notifications should inform data consumers about both breaking and non-breaking changes in a data product. Data products should have a well-defined lifecycle, and if a data product is shown to be unused via standard metrics, it should be retired. Data products can be used as an input to create other data products in the same or different domains; this use case will become increasingly common as the overall number of data products and domains grows.

REMEMBER

As the Data Mesh journey evolves and the organization restructures and architects itself around the decentralized data strategy, additional domains will begin producing data products. Those domains can benefit from the learnings of the domains that adopted Data Mesh earlier. For example, as the number of domains grows, the capabilities of the self-serve data infrastructure will become more well-developed. The federated computational governance approach will also evolve alongside the growth of adoption of the decentralized data strategy. Simply put, by focusing on a small set of domains and data products to start, the organic growth of the Data Mesh will lend itself naturally toward the development of the more complex principles around infrastructure and governance.

The self-serve data platform should provide telemetry and usage information on data products, to enable domains to understand and learn from consumers how their data products should evolve to better serve them. The interface should also provide usage information (for example, how to connect to a given data product from a BI tool), and give example queries which users find interesting or common. The same platform can provide data consumers with usage information to better understand data product popularity, ratings, and commentary. These are similar metrics to those available to consumers when purchasing products in an online store.

Data products in a mature Data Mesh

A mature Data Mesh provides a simple interface for both creating and consuming data products. Recall that consumers don't necessarily care which domain is responsible for a particular data product, provided they have all the information they need to use it without going back to the domain for more context.

The goal is a state in which multiple domains produce many data products. Once the Data Mesh journey is begun, creating, maintaining, and iterating upon data products becomes a more straightforward prospect. In this state:

- Published data products conform to clearly defined standards for shared aspects such as cross-domain entity rules, and enable data products to meet governance and regulatory standards.
- Each data product should be subject to access control, for example by a definition of roles and privileges granted to a given user.
- The self-serve data platform enables consumers access to data products in an easy-to-use environment that enables discoverability, accessibility, and ease of consumption.
- The platform does not assume that data product developers or consumers have specific technical knowledge, so lack of training does not prohibit anyone from performing the necessary data activities.
- There is shared ownership and motivation for producing and consuming high-quality data products, so the time to value is minimized and the data's ultimate value is maximized.

Optimizing for Maximum Value

To optimize a decentralized data strategy, you should focus on three main areas: culture, process, and technology. Here's a look at how each of those applies to Data Mesh.

Culture

The culture of a business is reflected in the behaviors and practices it encourages and rewards. To truly embrace Data Mesh, a company must adopt practices across the board that drive its employees to embrace the idea of a decentralized data philosophy:

- **Domain ownership:** Domain owners must be encouraged to — and even rewarded for — producing valuable data products that drive business strategy. For example, usage

metrics can be tied to a reward structure so data producers are lauded for creating the most valuable data products.

- **Data consumption:** Data consumers should be rewarded when they find new ways to use data products to discover valuable insights about the business, and when providing feedback to data producers about data products.
- **Embracing a data-driven mindset:** The entire organization should understand the value of accessible data and how it benefits individual careers, team charters, departmental accomplishments, and the business's overall growth and success.
- **Data enablement:** All employees across the organization must have the needed skills and knowledge to derive value from the data that they now have available to them.

Process

When considering implementing a Data Mesh, you will likely need to reconsider some of your organization's internal processes.


TECHNICAL STUFF

These changes will likely include adopting industry standard methodologies and approaches. Here are some suggestions:

- Use standard Agile methods such as Scrum to define the development lifecycle of data products. These practices are likely well-known to your development groups, and will simply need to be extended to data products.
- Apply iterative product management processes to data as well, such as clear product requirements definitions, minimum viable data products followed by usage testing and metrics-driven evolution.
- Make sure that domain-level roadmaps account for data products as well as business goals.
- Apply standard engineering practices such as testing, documentation, and retrospection to all data products.

Technology

To implement Data Mesh, you need technology that enables domains to deliver at the speed and scale necessary. A key consideration should be reducing cognitive load on the creators and

downstream consumers of data products by not requiring them to learn complex new technologies.

Some of the key characteristics of valuable technology in a Data Mesh are:

- **Interoperability:** The chosen technologies should integrate well with each other. They should be able to share common aspects such as metadata, and to remove the need for data product developers to concern themselves with infrastructure.
- **Simplicity:** The technologies you use should focus on business value and make data products easy to create, use, and understand.
- **Performance:** Data product usage should be focused on time to value for data and a seamless experience for the downstream data consumer. Similarly, data product creation should be focused on business logic and data curation rather than complex development pipelines.
- **Optionality:** The technologies you select should avoid vendor lock-in and the creation of technical debt. For example, the self-serve data platform should enable you to adopt different data storage technologies without impacting data product developers. This enables organizations to easily adopt newer technologies as they become available.

IN THIS CHAPTER

- Reviewing the benefits of Data Mesh
- Discovering tips for a successful implementation

Chapter 3
Ten Key Takeaways

In a hurry? Here are some quick tips for a successful Data Mesh implementation.

Understand why Data Mesh is a win. Data Mesh shortens the path between the data produced and its business value by organizing and architecting a decentralized data strategy.

Focus on creating valuable data products. Understand how your consumers use data products, prioritize use cases based on business value, and establish a Data Mesh strategy that fits the organization's objectives.

Start small, and then refine and iterate. Successful implementations start with a small scope, strong sponsorship, and a well-defined business value. These fast wins create immediate benefit for the selected business function, and provide a tremendous learning experience for the Data Mesh sponsor and the organization itself.

Begin with domains and data products. Focus initially on a small set of domains and data products to start. As the Data Mesh grows, the more complex principles around infrastructure and governance will develop organically.

Deliver data products like a service. Apply standard release and change management principles to data products. Changes should

be planned, assessed, and clearly communicated to ensure value for downstream consumers.

Federate data governance. Give domains the autonomy to decide how to best apply risk and compliance standards, as they are closest to the data and best positioned to understand the risk.

Develop a self-serve infrastructure. In a decentralized Data Mesh design, domains are focused on creating and serving products to the consumer. Remove the burden of implementing and managing infrastructure so they can focus on providing business value in the form of data products.

Keep it simple. Use technologies and access methods that data product creators and consumers already understand well. This will ease development and usage, and encourage adoption of Data Mesh principles.

Be intentional. As Data Mesh capabilities grow, maintain a pace of maturity that supports the organization's data strategy and business outcomes.

Invest in a strong data culture. Design domains and data products with consumers in mind, and teach them how to use the data products effectively. Time to value will be vastly improved by their ability to find, access, and use data products for decisions and actions, and they will become proponents for the decentralized strategy.

TIP

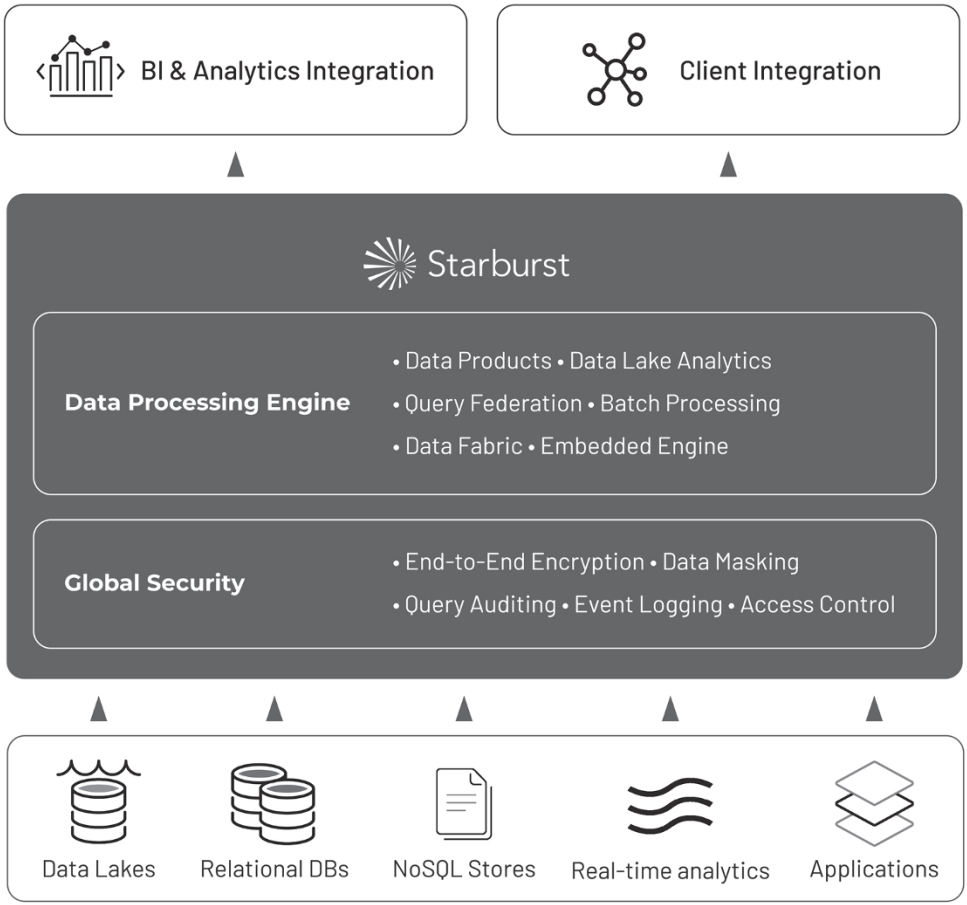
Find out more! Check out these useful resources:

- **Data Mesh, Delivering Data-Driven Value at Scale eBook:** `https://www.starburst.io/info/oreilly-data-mesh/`. The seminal text on Data Mesh by Zhamak Dehghani.
- **Data Mesh in Practice, How to Set Up a Data-Driven Organization analyst report:** `https://www.starburst.io/info/data-mesh-in-practice-ebook/`. The first of its kind, mapping Data Mesh Theory to reality by Max Schultze and Dr. Arif Wider.
- **Data Mesh Resource Center:** `https://www.starburst.io/info/distributed-data-mesh-resource-center/`. Enjoy exclusive access to Data Mesh content including the 90 Day Data Mesh Pathfinder series, on-demand talks, panel discussions, and more!
- **Data Mesh and Starburst blog series:** `https://blog.starburst.io/tag/data-mesh-and-starburst`. How Starburst supports the four key principles of Data Mesh.

Starburst **for Data Mesh**

Enable everyone in your organization to make better decisions with fast access to all your data, no matter where it lives.

- Single point of access to all your data
- Sharing and discovery with curated Data Products
- Access to 50+ data sources
- Advanced security and fine-grained access controls



Discover the benefits of Data Mesh

Data Mesh is a strategic approach to data management that decentralizes ownership of data ingestion, processing, and serving. This approach enables organizations to quickly leverage their data where it sits today, avoiding disruptive and expensive data migration projects. In this book, you will learn how Data Mesh works and gain valuable insight from experienced data leaders.

Inside...

- Understand the four principles of Data Mesh
- Learn how Data Mesh is relevant to today's business needs
- Build valuable, secure data products
- Optimize your data systems for maximum value
- Discover tips for a successful Data Mesh implementation

Colleen Tartow, Ph.D., is an engineering and data analytics leader, accelerating speed to insights. **Andrew Mott, MBA,** is a business analytics leader and speaker. **Adrian Estala** is a Fortune 25 CDO with a focus on leading digital and IT portfolio transformations.

ISBN: 978-1-119-91088-6
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.