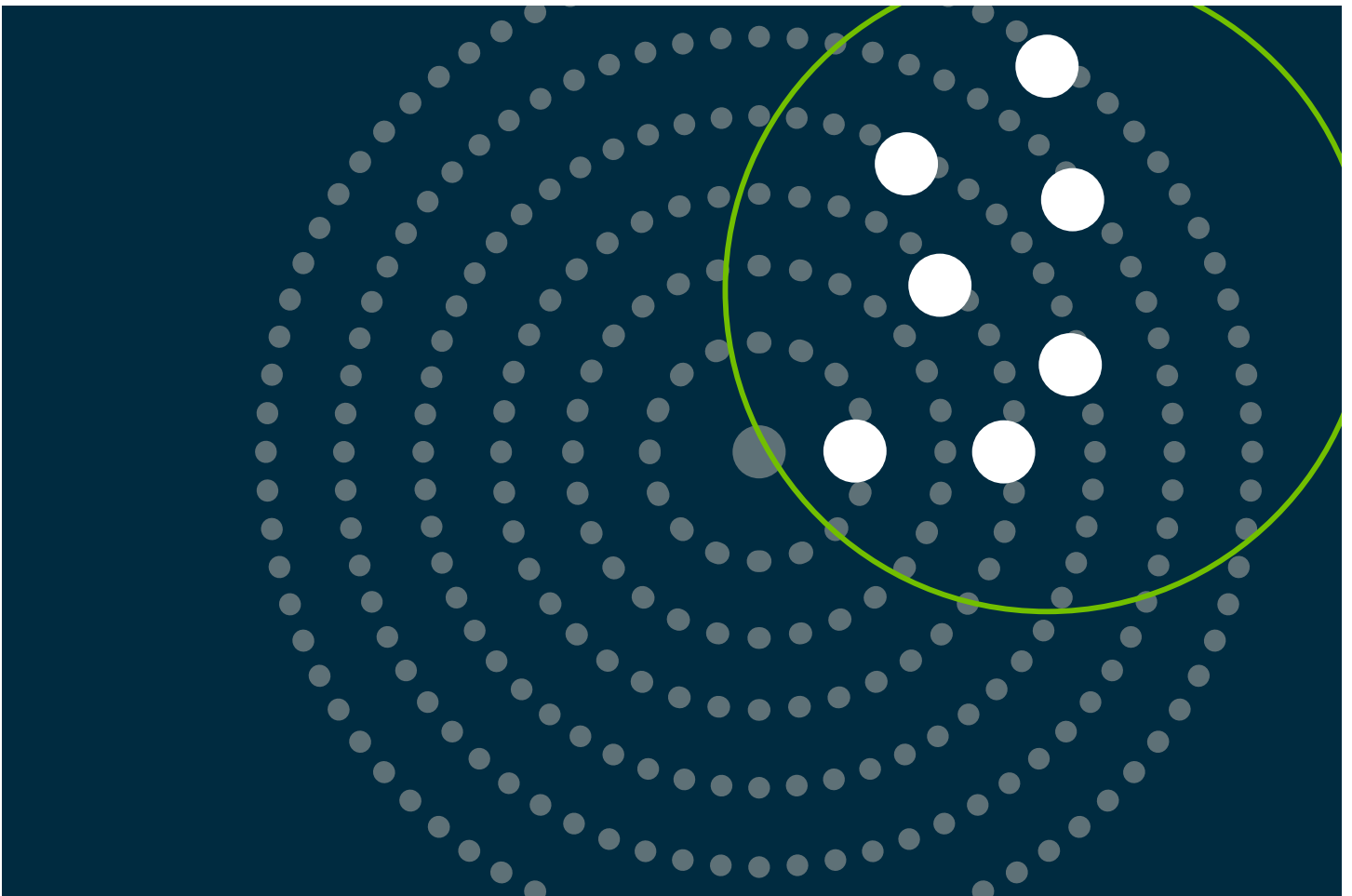
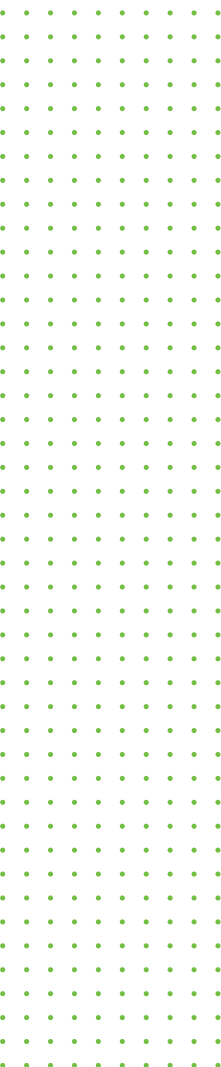


Data Mesh

Don't drown in your data lake



Introduction

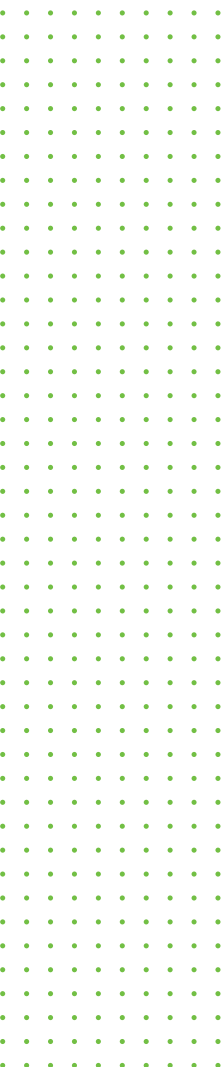


Almost every organization today understands the inherent value of data and the potential impact it can have on the business. With the right data-driven strategy, banks can better predict and avoid fraud, insurance companies can minimize losses, and retail organizations can better understand their customers and increase profitability. Over the years, businesses across all industries have invested in a variety of tools, solutions and platforms in the hope of mining, extracting and making smarter decisions based on the rich intelligence locked within their data.

Many have adopted traditional data warehouses and business intelligence platforms, and may have also incorporated a data lake, along with other cloud and machine learning capabilities. The common denominator across all these tools and solutions is that they are implemented as a centralized architecture. It's this monolithic, centralized approach that has inhibited or created significant friction in the ability for organizations to discover, understand and leverage their data assets to their full potential.

In today's modern landscape, data is coming from a myriad of sources, and the volume and variety of this data continues to accelerate. Often this deluge of data from multiple data sources and business domains is consolidated into a siloed repository to be put under the control and watch of an IT or data platform team. As this web of ever-growing data becomes more complex and intermingled, the process of finding the requested needles in the haystack, within a reasonable SLA, grows more complex as well.

In addition, these teams are typically disconnected from both the business domains that create the data, and the consumers that need to access it. Without any insight or knowledge about the data — what's in it, where it came from, or what it will be used for — how can they be expected to provide accurate, valuable data? By simply dumping massive amounts of data into the central repository, the true owners of the data lose their ability to continually curate it and ensure its quality. The goal should instead be to embrace a data management process that supports the collaboration of data owners, data engineers, business analysts and consumers of the data.



This becomes an issue for both the business user and the company as a whole. Take a leader of a bank's fraud division, who relies on understanding the massive amount of data about their customers' habits and usage, in order to better detect fraud. Without knowledge of the underlying data, what sources it came from or the quality of the data, he exposes his customers and his organization to significant risk.

In an ideal world, data should be available to everyone across the organization without the need to rely on technical teams who oversee the data. A key barrier to enabling self-service access is a lack of connective tissue across all data sources and systems. Without this connectivity, users are reliant on technical resources to get visibility of where data came from, understand what's in the data and confirm the quality of the data. This isn't a scalable model as data insights are delayed, productivity is impacted and resources are overburdened.

The idea with deploying data warehouses was to address the lack of integration and house data in a central storage so that the business could have a unified view across all their data. Creating a common view requires ongoing cleansing and transformation. Given the volume, variety and pervasive poor data quality, many data initiatives failed to live up to expectations. Data lakes helped avoid the requirement for conformance, but many have also fallen short due to a lack of governance together with a glut of outdated and irrelevant data.

Taking a legacy approach is impacting your business

IDC predicts that datasphere will increase from 33 ZB now to 175 ZB by 2025

32%
of companies reported being able to realize tangible and measurable data

Undoubtedly, data will continue on its explosive growth trajectory. The International Data Corporation (IDC) predicts that the datasphere will increase from 33 ZB now to 175 ZB by 2025¹. A majority of this data will be stored in the enterprise core and in public cloud environments². This growing data footprint will continue to tax a centralized architecture, which will be difficult to scale and slow to support new and expanding data sources.

This data growth also presents a challenge in extracting additional insight from data assets. Today only 32% of companies are able to realize tangible and measurable value from data³. Continuing with a current approach may not provide the agility to take advantage of the additional insights and value possible from new data.

77% of organizations are integrating up to 5 different types of data in pipelines⁴. The inability to homogenize different data types will further burden data teams and slow access to trusted data.

Data can only be valuable if it is trusted and accurate, yet only 3% of companies' data meets basic quality standards. A core factor contributing to this problem is the lack of data ownership with a centralized strategy.

65% of organizations are using at a minimum 10 different data engineering tools⁵. The digital ecosystem within organizations will continue to expand, and without broad integration across all tools the business will lack a unified view throughout the enterprise.

A new way forward: Data mesh

With today's enterprise landscape consisting of massive data volumes, a multitude of data sources, and a desire to drive smarter, data-driven business decisions, legacy data strategies have not evolved to meet the need and speed of business. Data mesh is not just another technology, but a completely new architecture to address the use cases of a complex data environment.

Data mesh takes a different approach to data by focusing on several core concepts, including domain ownership, data as a product, distributed data governance and self-service design. These concepts have been largely adopted across other parts of the tech industry with sustained success, and now provide for a transcendent methodology beyond the monolithic, centralized data warehouse/data lake architecture.

One of the primary pitfalls of the legacy architecture is the concept of pumping everything into a data lake and relinquishing the expertise and stewardship of the data from the actual owners of it. In an effort to decentralize the platform of yesteryear, the idea is to empower individual business domains to host and own their data rather than flowing it into a data lake. As the experts of the data within their business domain, domain owners are responsible for cleansing, enriching and making the data readily available to be served to consumers of the data throughout the organization. These domain owners must establish and maintain the quality of the data and provide the necessary trusted facts and documentation about the data. Data consumers should no longer have to take this on. Nor should they need to interface directly with the owners of the data source, as is often required in today's data warehouse and data lake solutions, which contributes to the friction of getting the right data to the right people.

A key part of the data mesh equation is delivering data as a product. This simply means that the owners are providing data to their consumers in a state that is ready-to-use, and does not require additional cleansing or rework. In order for consumers to derive value from the data, there are a core set of qualities required.

Firstly, the data must be easily **discoverable**. This can be done through a data catalog which provides a variety of metadata information, such as owners, origin and lineage. With each domain registering its data products, consumers can easily find the data for their purposes.

The data product needs to be **addressable** with standardized naming conventions and formats. By developing and agreeing on a common set of conventions, the business will be able to find and access the information they need in a ready-to-go format.

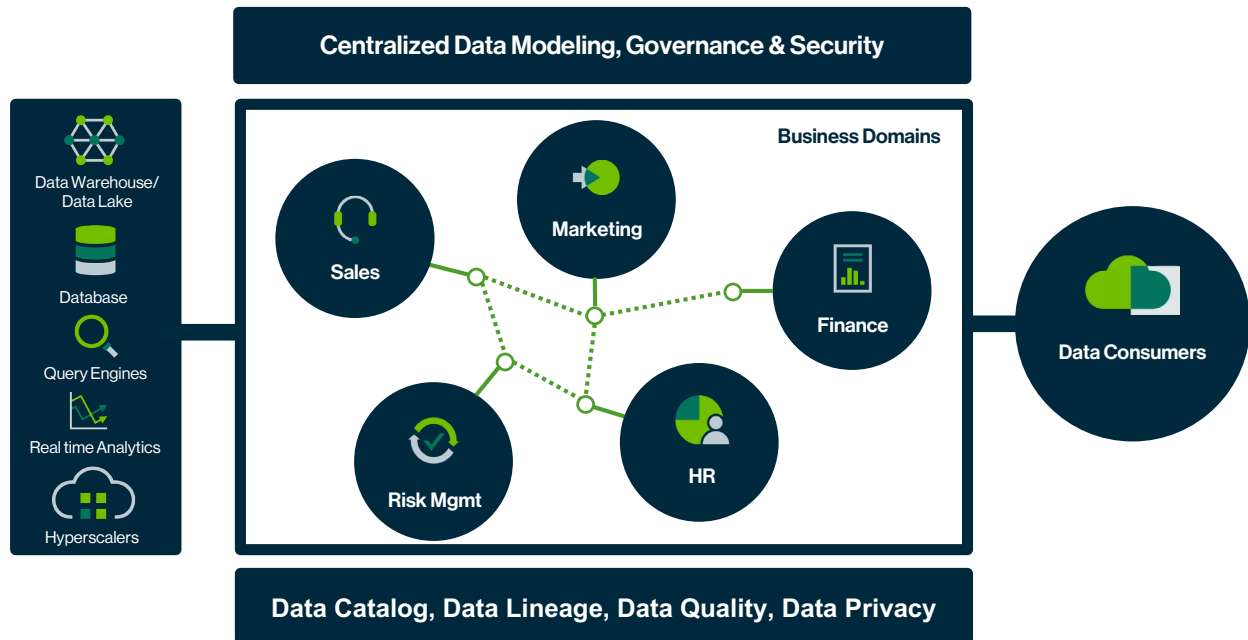
In order for the data to be of any value it must be **trustworthy**. Data should be cleansed and tested by the data owners prior to making it available. Consumers can confirm the trustworthiness of the data when data provenance and data lineage are also provided from the data domain.

To help facilitate the self-service nature of data mesh, the data should be **self-describing**. The consumer should not have to go back to the data owner to make sense of the data for their use. The semantics and syntax of the data should be clearly described so that it can immediately be consumed and avoid any bottleneck requesting assistance from the owner.

The nature of a distributed architecture solicits the need to support **interoperability**. Consumers will likely need to correlate and aggregate different data sets from different domains. To ensure data can be leveraged and compiled across multiple business domains, a set of standards and conventions need to be addressed and governed at a global level.

Finally, the data products must be **secured and governed** in accordance with security and compliance requirements such as GDPR. Access controls should be in place, both at the global and data domain level, for additional granularity.

One of the core tenets of data mesh is the ability to democratize data by enabling consumers to access the data they need in a self-service manner, without the need to engage IT or the data owner. With a traditional approach, users are often encumbered with discovering where to find data, what's in the data, where it came from, can it be trusted, as well as interfacing with data and IT teams to access, cleanup and understand the data. This gauntlet of obstacles has contributed to the failure of many data projects, and has limited the ability for users to get the intelligence from their data. Data mesh caters to the end user by hiding and automating all of the complexity, while providing all the necessary insights they need when they need it with self-service. The organization becomes a well-oiled machine that embraces the frictionless exchange of data between data owners and data consumers, but includes guardrails implemented with a system of distributed data governance to ensure security and compliance.



Benefits of a data mesh

As every organization's data landscape continues to advance and evolve, so too must the data intelligence and governance strategy. There are many benefits to incorporating a modern data mesh architecture in order to keep in lockstep with today and tomorrow's data environment.

A data mesh architecture is designed to decentralize much of the heavy lifting related to data operations that was previously put upon the shoulders of data platform and IT teams. By transferring the onus of these tasks to individual business domains, it's now possible to remove IT bottlenecks and empower these domains to produce and curate data at enterprise scale and speed.

With domain data ownership, organizations can now put the control of data in the hands of the true stewards who have deep expertise and knowledge of the data. Rather than boiling an ocean of data these business domain teams can focus on ensuring their data is cleansed, trustworthy, and always available to support business agility.

Built on an architecture with self-service as a priority, data consumers can free themselves from complexities and focus on getting rapid access to the right data when they need it.

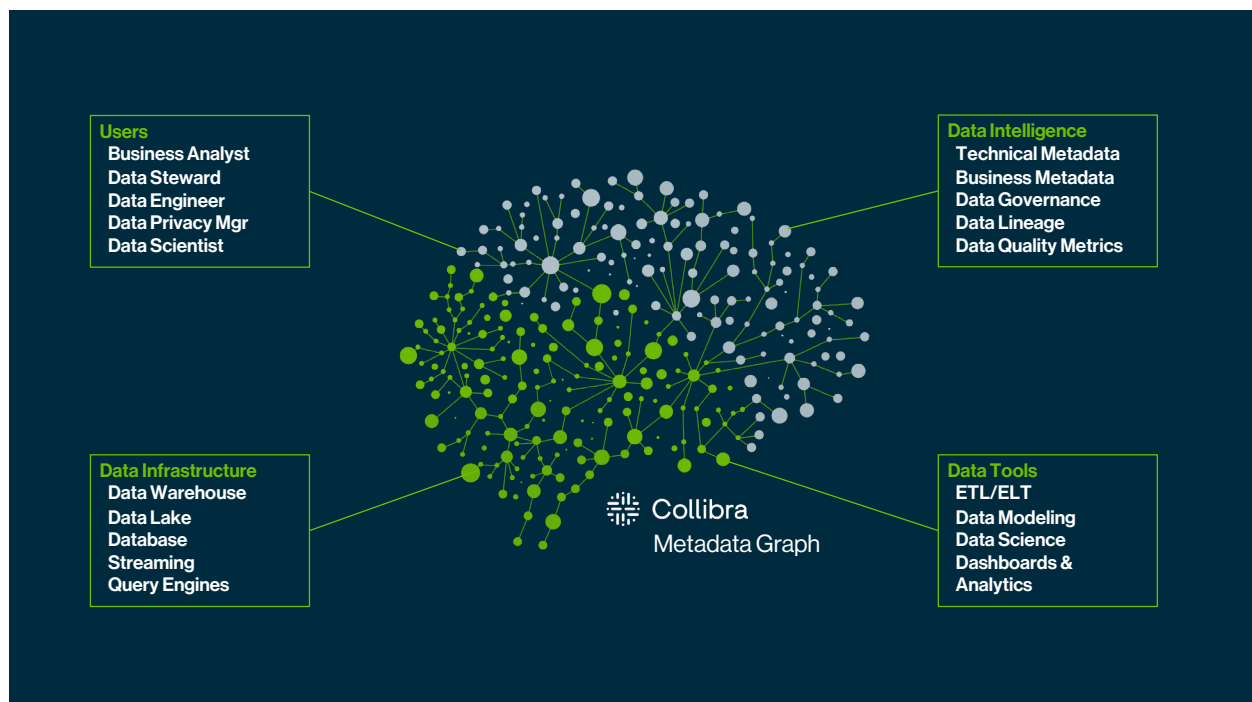
The digital ecosystem will continue to expand across a variety of data sources, formats and repositories. Relying on a centralized set of data governance, security and compliance policies, companies can enable users to easily correlate data, protect access to sensitive information and adhere to compliance requirements.

Supporting a data mesh with Collibra

To implement a data mesh architecture, a data fabric can be used as the connective tissue across the array of data sources and domains. A data fabric, supported with a composition of tools, technologies and processes, acts as a layer that integrates and connects all the various components of the mesh—from the data and source systems to the cataloging and centralized governance—and makes the data accessible where it's needed.

Although there are many business domains working independently, they still need to communicate with other business domains and also need to conform to the centralized set of data modeling, security and compliance policies. Rather than dealing directly with underlying sources and systems across the organization, centralized teams and business domains can instead share data through APIs and pipelines.

Once a global set of standards, definitions and policies are agreed upon these can be centralized across all domains. The data fabric brings together all data and domains for a unified view and enables the streamlined exchange of trusted data that is accessible via self-service.

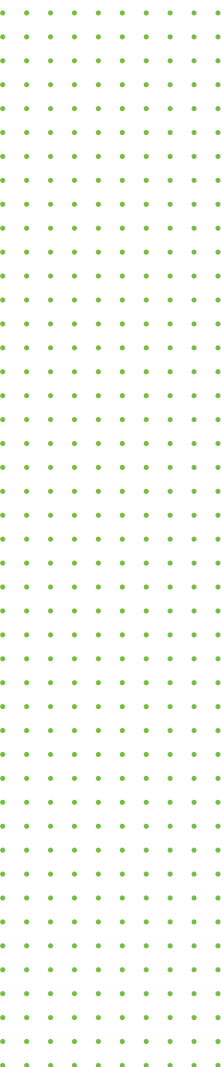


Collibra provides the needed foundation to help organizations start down their data mesh journey with a single source of truth for data across distributed environments.

The first step is enabling business domains to make their data discoverable. Collibra provides full visibility into data across the organization. Data is cataloged with relevant business context, including business definitions, ownership, policies and usage. Data consumers can easily find the data they're looking for across all data sources, business applications, data science and BI tools. Users can search and understand data using common business terms, and also get visibility to see how data transforms and flows from system to system.

Distributed business domains provide the framework that allows data owners to maintain their own data. Collibra provides the tools these business domains need to ensure their data is relevant and high-quality by understanding where data has been, who has used it, and know with confidence that it is reliable and accurate. Business domains can continuously monitor data for completeness, timeliness and accuracy, and quickly detect data quality issues before they become more serious. Collibra can help keep up with the scale of today's data environments by leveraging machine learning to generate data quality rules, and implementing quality checks along every step of the DataOps journey.

A decentralized data management model creates enormous opportunities to improve data quality and streamline data access, but seizing this opportunity goes beyond the data itself. Getting visibility to data assets isn't enough. Organizations must gain a complete understanding about the relationships and how the data is connected. Collibra's active metadata graph creates context-rich connections across all distributed business and data domains. The ability to repurpose data, collaborate and drive better decisions with data intelligence is the end result of connecting the right data, insights and algorithms to all users.



With Collibra, data producers and consumers can take advantage of rapid, self-service access to data. Users can shop for data, reports, models and other data assets through a business friendly, easy-to-use interface. Together with core data quality capabilities, users can quickly access the data they need with high confidence that the data is complete, accurate and reliable. Furthermore, organizations can ensure policies are adhered to and data access is granted in compliance with usage and privacy policies.

In support of distributed data governance, Collibra provides a flexible operating model that allows configuration of any type of entity (business term, data type, business rule, quality rule, process), any reference data, and any type of relationship. This means organizations can model their desired data strategy to fit their needs and put the data in the hands of the people who need it, using the right organization hierarchy structure. A role-based access and permissioning model provides the granularity necessary to support security and compliance requirements and deploy enterprise wide. Embedded governance capabilities, within the data catalog, ensure that access to trusted data can be granted enterprise-wide without risk.

Collibra helps companies easily integrate with all their data sources, BI and analytics tools, to ensure employees can be as productive as possible. Collibra further helps organizations eliminate disparate, point solutions and standardize on a single platform for data intelligence and management. The platform provides tools to reduce the effort related to manual rule writing, and can help save time and resources in identifying and migrating data to the cloud. Organizations can also avoid compliance fines with centralized security controls compliance policies.

Taking the next step

Data mesh is still a fairly nascent and evolving concept in the data management sphere, but has already been proven in other tech disciplines. Many companies see the promise and potential impact on business outcomes, with many already beginning to put this type of architecture in place. Those looking to start down the data mesh path need to think about leveraging the right tools to get the most out of this approach. Explore the Collibra Data Intelligence Cloud to learn more about how Collibra can help your organization embark on its data mesh journey.

¹ IDC 2018

² <https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html>

³ Accenture, Closing the Data-Value Gap, How to become data-driven and pivot to the new

⁴ IDC 2021

⁵ IDC 2021



To learn more about the Collibra Data Intelligence Cloud,
visit www.collibra.com/data-intelligence-cloud