

DATA QUALITY AND GOVERNANCE

Maximizing Data Value in the Age of AI

Data products accelerate outcomes while
enhancing governance

In today's fast-paced business landscape, data is a vital strategic asset — a linchpin for optimizing operations, driving innovation, and propelling growth. Beyond capturing mere numbers and statistics, data encapsulates insights, trends, and patterns that can illuminate pathways to success. Now more than ever, with the spotlight on generative AI and machine learning, businesses are grappling with a fundamental question:

How do they quickly unlock the full value of data to accelerate business outcomes — all while enhancing data trust and governance?

Speed is a key element in this equation. According to a recent IDC study, 75% of decision-makers say that data loses its value within days.¹

The challenge goes beyond managing overwhelming volumes of data. It involves refining information into high-quality, curated, and readily accessible assets suited to specific business domains. That, as anyone who works with data knows, is easier said than done, though. Data and analytical leaders are struggling with several challenges that hinder the full realization of their data initiatives:

- **Project-oriented data management lacks scalability.** Each project team operates within its own set of parameters, including scope, objectives, and stakeholders. This fragmented approach hinders cross-functional collaboration and limits the extraction of value from data. And the risks are manifold: stakeholder misalignment across projects, duplicated efforts, governance gaps, inefficient data utilization, increased costs, and ultimately, delays or missed business outcomes.

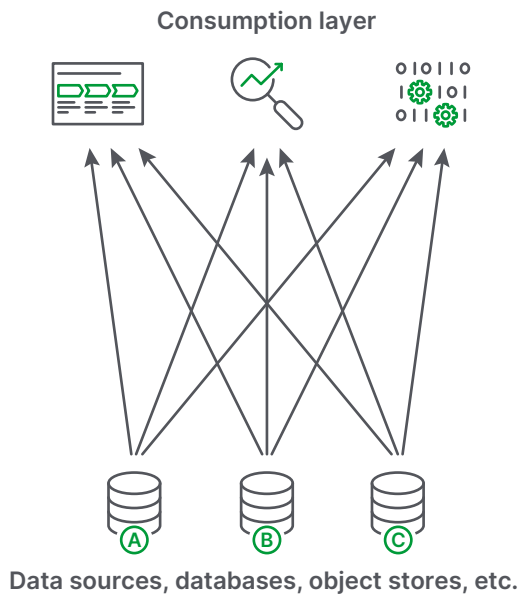
- Across the enterprise, **there is a lack of consistency in data quality across systems**, undermining decision making processes and eroding trust in data-driven insights.
- Across the data and value supply chain, **there is a lack of clear ownership and accountability** among data producers and consumers, thwarting the attempt to extract value at the speed of the business.
- **Relying on fragmented solutions and ad-hoc, do-it-yourself approaches** to data management prolongs time-to-value and is costly to scale and maintain without skilled staff.

These challenges widen the divide between data producers and consumers. On one end, data producers — who are IT-focused — prioritize data platforms over understanding data usage and consumer needs. On the other, data consumers — entrenched in their business domains — struggle to convey requirements and don't trust provided data. **To close this widening gap, organizations need a strategic, streamlined approach to data management.**

¹ IDC Blog, "Navigating the Planes of Enterprise Intelligence Architecture," June 2023.

Shifting data management approaches

Traditionally, organizations have employed two distinct approaches to extracting value from their data: the **fragmented** approach and the **centralized data** approach.



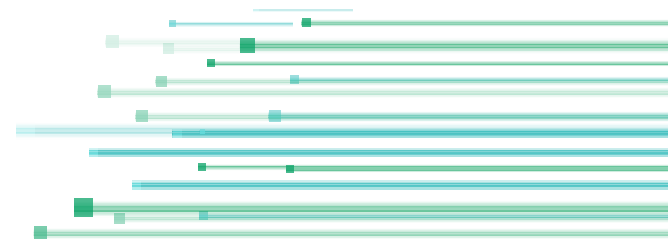
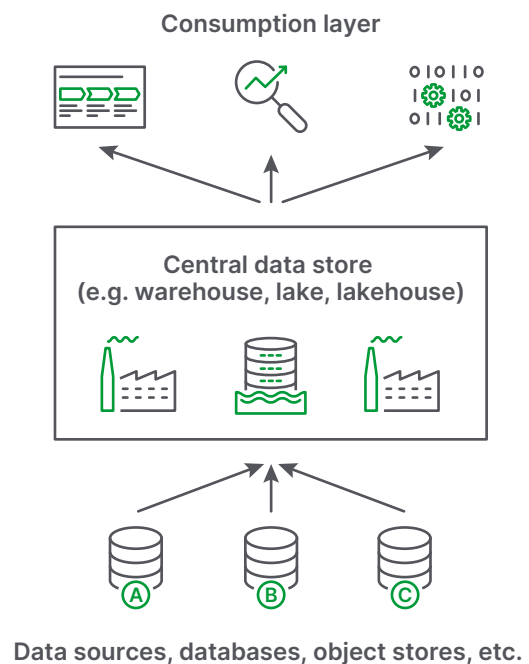
THE FRAGMENTED APPROACH

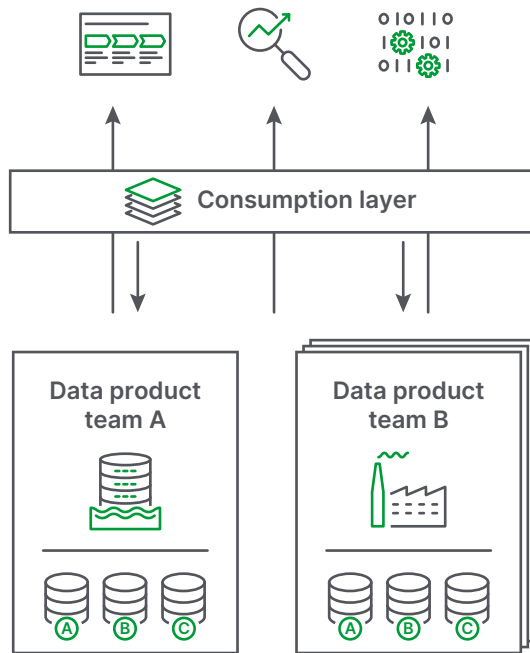
In the fragmented approach, individual “data-project” teams gather data from various source systems to address specific business use cases. Each team consumes data independently, using their own tools and technologies to extract, integrate, and analyze data for their specific needs. While this method generates immediate value in the short term, it’s not scalable for large enterprises in the long term, resulting in more data silos, duplicated efforts, increased costs, and delayed business outcomes.

THE CENTRALIZED APPROACH

Alternatively, a centralized approach makes a centralized data team responsible for extracting, cleansing, and aggregating data on a large scale. The central team’s objective is to address the analytical requirements of the organization comprehensively. They establish intricate data pipelines to retrieve data from various systems, cleanse it, and consolidate it within cloud-based data lakes, warehouses, or lakehouses (such as Databricks Delta Lake, Snowflake, Big Query, S3, etc.).

Centralization may sound like the solution to fragmentation — and to some extent, it is — but it has a fatal flaw: Because it requires extensive coordination across departments to make changes or updates, it impedes organizational agility. Additionally, by focusing on common data requirements across the organization, the central data team often overlooks the specific needs and preferences of business stakeholders and consumers, resulting in a suboptimal solution that fails to align with business requirements.





THE FEDERATED APPROACH

Both traditional approaches — fragmented and centralized — have limitations, leaving gaps in meeting the data consumption needs of businesses. However, in response to these challenges, a paradigm shift is underway. The increasingly popular **federated approach** strategically distributes data delivery by harnessing data products while centrally managing platform engineering functions.

Multiple data product teams can work in parallel, with each team focusing on their domain-specific data products. This approach accelerates business outcomes by leveraging cross-functional collaboration, including both business and technical stakeholders.

The federated approach offers several key advantages:



Agility and flexibility

By streamlining end-to-end data management, a federated approach enables organizations to **swiftly unlock business value** from data. It promotes agility by embracing a product-driven mindset where delivery and value creation are distributed across cross-functional product teams, while platform engineering functions are carried out centrally. This allows for **quicker responses to evolving business needs** — all while ensuring data integrity and accountability throughout the process.



Cost-effectiveness

The federated model embraces **reusability**, empowering organizations to maximize their investments and lower operational costs. Moreover, central sharing of foundational components among different product teams **minimizes redundant work and enhances long-term maintainability**.



Accelerated use case delivery

Facilitated collaboration among cross-functional teams yields **rapid, domain-specific outcomes** customized to meet evolving business requirements and new data-driven initiatives. The federated approach can achieve **scalability and deliver domain-specific use case customizations** through parallel development streams across multiple data product teams.



Enhanced trust

The federated model **fosters trust with enhanced governance, transparency, and compliance**, instilling confidence in data-driven decision making. Clear data domain ownership fosters alignment and accountability across data and analytical stakeholders, further strengthening the organization's data governance framework.

Introducing data products

A federated approach is the most appropriate starting point for data products, because for many organizations, it signifies an evolutionary shift — from control to collaboration, from project to product, and from rigidity to agility, all while delivering value at scale. **Data mesh** and **data fabric**, distinct in their approaches to data management, can together form a powerful synergy to unlock the full potential of data products. In the data mesh paradigm, data ownership is decentralized, owned, and managed by the teams that use it. Data product is the most tangible and impactful aspect of the data mesh paradigm, and the data fabric becomes the foundational data management architecture that enables the optimal delivery of data products to domain teams.²



According to the Gartner CDAO Agenda Survey for 2024, 50% of participating organizations have already deployed data products, and another 29% are committed to piloting or considering deployment within the next year.³

So, what is a data product? A data product delivers high-quality, curated, ready-to-use, and AI-ready datasets that people can easily access and apply to different business use cases across an organization. A data product can be thought of as a microservice for data. Just as microservices expose application capabilities via smaller, more contextual, and composable units, a data product provides access to domain-oriented and relevant data via service interfaces. A data product adheres to several fundamental principles:



Domain orientation and ownership

Every data product must be aligned to a specific domain and have a data owner (person or team) responsible for its success — i.e., delivering value and growing adoption.



Product thinking

Data products are created by applying product thinking to data, including ownership, versioning, and customers (in this case, data consumers). Products are enhanced by understanding usage patterns and obtaining feedback from customers. They have a lifecycle, just like real products.



Standardized and interoperable

Data products adhere to standardized formats, schemas, and metadata conventions. They facilitate interoperability and ease of integration with other components of the data ecosystem.



Findable and discoverable

Data products are registered in a data product catalog and can be easily found and discovered by data consumers. A data product describes itself — what data it represents, how it can be accessed, its business semantics, and its lineage and quality.



Reusable

Data products are built once and used multiple times in various use cases. Other data products are built using existing data products.

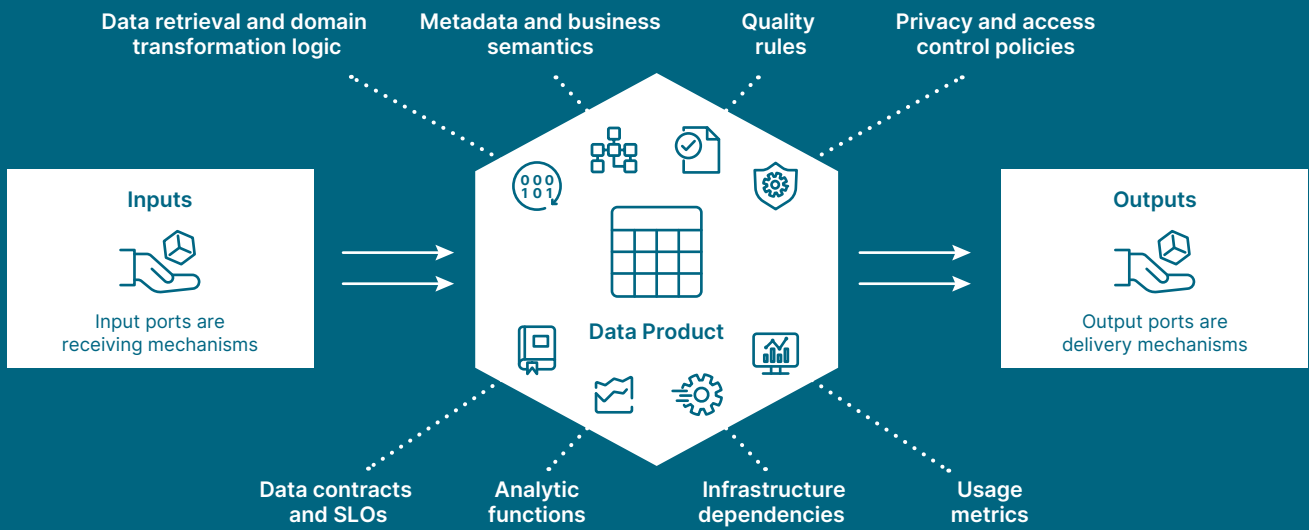


Accessible

Data products are accessed through a standard set of interfaces such as SQL or a REST API that are defined and exposed when the data products are created. The interfaces hide the complexity within the data product.

² Gartner: 6 Lessons Data Leaders Can Learn from Adopters of Data Mesh by Robert Thanaraj, et al. 4 March 2024. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

³ Gartner: Gartner Chief Data and Analytics Officer Agenda Survey for 2024. 12 March 2024. Results of this study do not represent global findings or the market as a whole but reflect sentiment of the respondents and companies surveyed. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

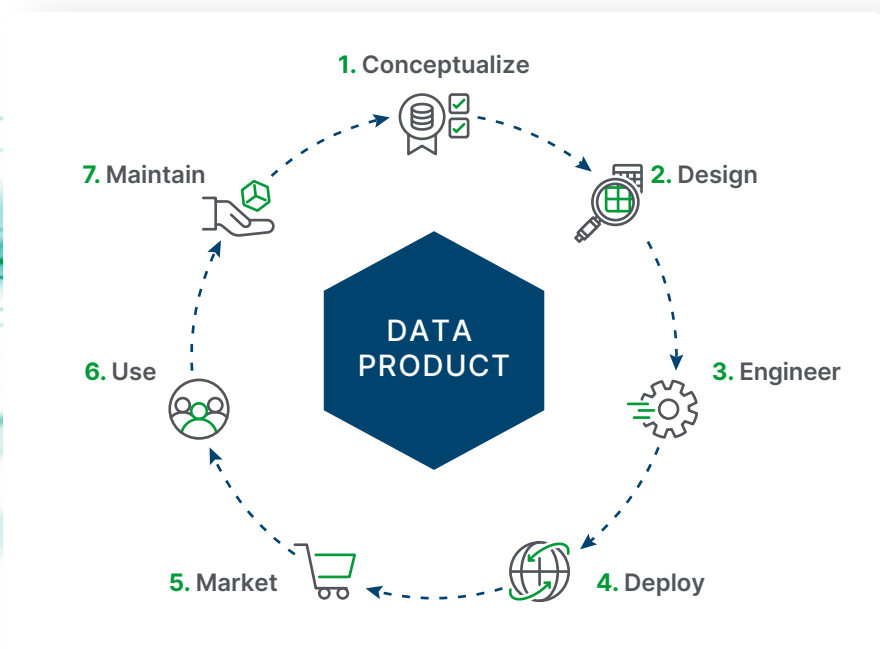


A data product is a self-contained quantum that encapsulates data, transformations, code, governance policies, SLAs, access patterns, and infrastructure dependencies as a unit. The diagram above shows an anatomy of the data product.

- 1 Dataset:** Exists at the center of the product. It could be a set of tables, files, graphs, or vector representations. There could be multiple tiers of data within a data product — a raw set that represents data as ingested from a source system, a curated set that represents cleansed and integrated data, and a consumption-ready dataset modeled for ease of use.
- 2 Inputs:** Represents the input ports or incoming source of data. This could be tables in an SAP application that need to be ingested, tables in Snowflake that need to be used as a source, another data product from which a new one is being derived, or a message queue to receive real-time data feeds.
- 3 Outputs:** Represents the consumption-facing data-delivery mechanisms. A data product can expose a variety of standardized interfaces like SQL, REST API, vector search, message, and file. Output ports can also vary based on the type of service the data product provides. One output may provide access to historical data, while the other may provide access to the latest snapshot of data.
- 4 Data retrieval and domain transformation logic:** Represents the data pipeline code and logic to acquire the data e.g., moving the data from a source database to a cloud data platform or reading a set of tables), and then applying transformation logic to shape the data from the source format to the target dataset format, which can be in the form of tables, views, or vector embeddings.
- 5 Metadata and business semantics:** Contain the technical metadata (table names, file names, column definitions) and business metadata (domain, product name, type of product, purpose, ownership, tag, etc.) to allow for a better understanding of the data product.
- 6 Data quality rules:** Define the various data quality checks that need to be executed on the data product. Each rule has configuration information around when to execute, the scope of the rule, whether to execute on all data or incremental, etc. The data quality rules will be executed at the prescribed schedule to ensure data accuracy and build trust.
- 7 Privacy and access control policies:** Establish the policies that control access to the dataset within the data product. For example, if the dataset contains personally identifiable information (PII), a privacy policy can mask it for all users, or only certain user groups may be granted access.
- 8 Data contracts and SLOs:** Define the target values for the service objective that the data product is expected to fulfill. For example, an SLO could require a data quality accuracy metric to be greater than 95% or the freshness of the data to be every five minutes.
- 9 Analytic functions:** Encapsulate additional custom business logic to make the function reusable. A function on a sales order data product could be delinquent orders, which will encapsulate the complex domain logic to determine which orders are delinquent. The analytic functions can be exposed via an output port like an API.
- 10 Infrastructure dependencies:** Information about the infrastructure requirements for the datasets associated with the data product. For example, the datasets are stored on a data platform (like Snowflake or Databricks), with the associated code (e.g., transformation logic) requiring a compute environment (e.g., SQL) to execute.

Data product lifecycle

Data products are not one-time creations. Like material products, they require ongoing management and continuous monitoring. Each data product is owned and managed by a domain-aligned data team responsible for success in delivering value, growing adoption, and maintaining its life cycle.



Here's a breakdown of the key stages in the data product life cycle:

- 1 Conceptualize:** The process begins with identifying domain-specific business needs, including understanding the users, their use cases, and how they typically consume data.
- 2 Design:** The next step is to shape the solution, which includes identifying the needed datasets and their sources as well as defining the characteristics of the data product by functionality, interactivity, presentation approach, etc.
- 3 Engineer:** After design, it's time to build the technical infrastructure and functionalities of the data product. This includes gathering and integrating the raw datasets from their sources, ensuring data quality, and transforming the data.
- 4 Deploy:** This is the "making-it-accessible" stage, which involves completing product documentation, assigning ownership, integrating with existing business systems or consumption workflows for seamless access, and activating the data product, or making it ready for consumption.
- 5 Market:** After activation, it's important to create awareness and promote usage. This involves training consumers and setting up communication channels to capture their feedback.
- 6 Use:** Given a successful launch, consumers will begin interacting with the data product to search, find, understand, and gain access. This stage focuses on monitoring usage and gathering feedback. Typically, data is collected on how users interact with the data product, how it's performing, and whether it's meeting the domain-centric needs of the business.
- 7 Maintain:** Data products are not static entities; they need to be improved to ensure continued value. Based on the user feedback collected and the evolving needs of the business, new data dimensions, features, and functionalities are proposed to be added into the next version of the data product.

By following these life cycle stages, organizations can create data products that deliver meaningful insights, drive data-driven decision making, and achieve their business objectives.

What a data product is NOT

The world of data is cluttered with terms, and the usage of those terms isn't always consistent. Case in point: The phrase "data product" is often conflated with terms like "dataset" or "data catalog," which aren't the same thing. This section highlights some fundamental differences among these concepts.



While a dataset is a fundamental component of many data products, it's crucial to understand that it's only one component. A data product extends far beyond just the raw dataset.

DATASETS: THE RAW MATERIAL

A dataset is usually a bunch of organized data points, either structured or unstructured. This can include raw information like customer details, financial records, or readings from sensors. Think of datasets as raw materials — for example, piles of timber — that have potential to become valuable but need refining first. One of the most important things that differentiates a dataset from a data product is the lack of domain-specific business context.

DATA PRODUCTS: THE FINISHED GOODS

Data products take meticulously curated and processed datasets, often combined with additional artifacts discussed in the previous section to deliver value. To extend the timber metaphor from above: Think of data products as beautifully constructed furniture from the timber (i.e., the datasets).

Here is what differentiates data products from datasets:

- **Context and domain centricity:** Data products are tailored to address specific business needs or user demands. They go beyond raw data to include contextual information like definitions, lineage, and quality checks.
- **Usability and accessibility:** Data products are designed to boost value and ease consumption. They are curated and packaged to be readily consumable by domain-specific data consumers with limited technical expertise.
- **Actionable insights:** Data products do not just present data; they offer operational, analytical, predictive, and prescriptive insights. These insights can be delivered through visualization dashboards, pre-built analytical models, or integration hooks with other tools through SQL or APIs.
- **Ownership and accountability:** Data products have clear lines of ownership, ensuring accountability for various aspects of their development, consumption, and feedback within the organization. Ownership of data products can depend on the business context, with the owner being the producer who generates the data product, the consumer within the domain team responsible for the relevant use cases, or a newly emerging role known as the Data Product Manager.



Data products transform datasets into valuable, consumable assets that drive informed decision making and business growth. Harvard Business Review estimates that "companies that treat data like a product can reduce the time it takes to implement it in new use cases by as much as 90%, decrease their total ownership (technology, development, and maintenance) costs by up to 30%, and reduce their risk and data governance burden."⁴

⁴ "A Better Way to Put Your Data to Work," Harvard Business Review, July-August 2022.

DATA PRODUCT CATALOG: THE HUB OF FINISHED GOODS

A data product catalog serves as a comprehensive repository that houses data products and associated metadata — descriptions, tags, version, ownership labels, etc. — to aid in data product discovery and comprehension. It caters to both data producers and consumers, offering tailored solutions to meet their needs. For Data Product Managers, the catalog provides tools to configure, activate, and manage the entire life cycle of a data product, from creation to activation to end-of-life. For consumers, the catalog serves as a marketplace, which enables a shopping-like experience that drives utilization.

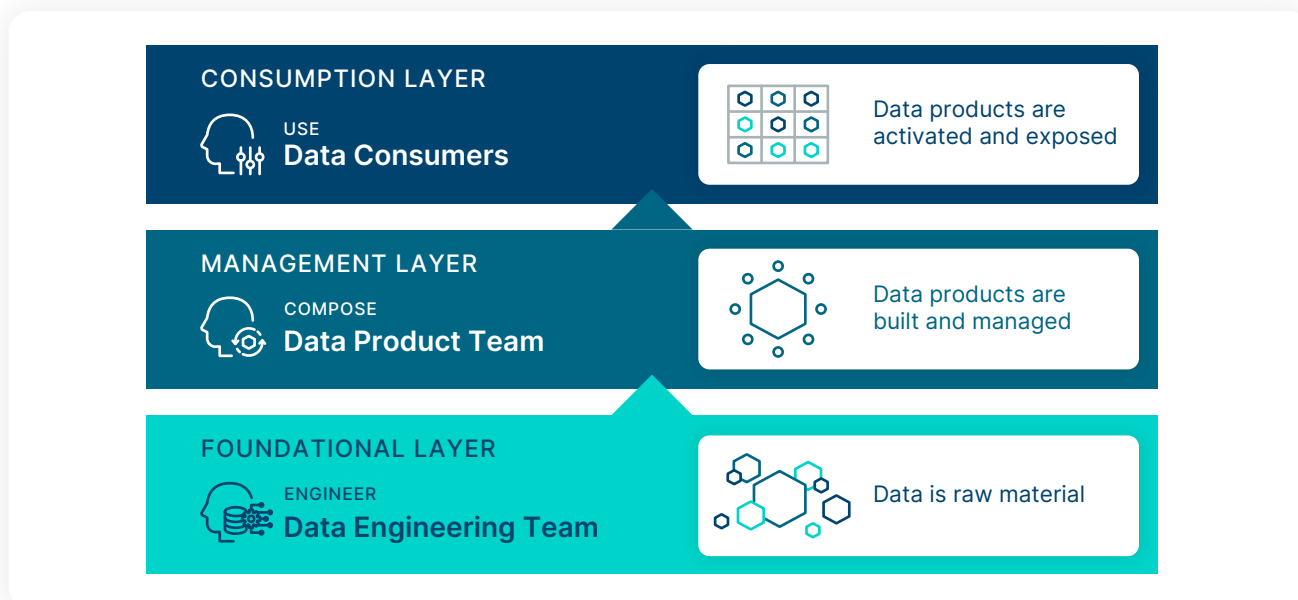
Here's what sets a data product catalog apart from a data catalog:

FEATURE	DATA CATALOG	DATA PRODUCT CATALOG
Mission	Strengthen data governance	Increase data value
Scope	Enterprise-focused	Domain-focused
Management model	Centralized	Federated
Purpose	Discoverability, compliance, transparency	Usability, trust, actionability
Content	Metadata of diverse kinds, including technical, operational, and business related to every data asset	Trusted data products, data product life cycle metadata, data product documentation, and ownership
Entities	Order of hundreds of thousands	Order of hundreds
Audience	Primarily data governance	Data producers and consumers

In the realm of data management, the challenge lies not merely in creating data platforms but in ensuring their effective adoption. Too often, organizations invest substantial resources — frequently millions of dollars — in building data infrastructures, only to encounter a serious problem: **data-sharing barriers between producers and consumers**. Why? Because the outcomes generated by many technology platforms lack reliability and fail to align with specific business objectives of consumers. **The bottom line is that you need both a data catalog and a data product catalog in your enterprise**. You can think of them as different aisles in your grocery store — one with raw materials like flour, butter, and sugar and the other with finished goods like cakes and pastries.

Data product catalog: layers and players

The data product catalog is the primary hub for organizing data products within an organization. It's structured into three layers, each tailored to a different set of stakeholders. Whether it's data engineers, stewards, analysts, or emerging roles such as Data Product Managers, each group interacts uniquely with these layers, driven by their objectives and requirements. This approach ensures separation of concerns based on user needs, with the right level of abstraction to ease adoption and consumption.



THE FOUNDATIONAL LAYER

The foundation of the data product catalog is a critical layer that establishes the technical infrastructure and delivers the capabilities supporting the subsequent layers.

The key capabilities of this layer are:

- **Data engineering:** Moving and processing data from various sources, including files, cloud data warehouses, data lakes, and real-time data streams.
- **Metadata management:** Collecting and organizing valuable information about the data, such as data lineage, data dictionaries, schema, etc.
- **Security and access control:** Enacting data security controls to protect data and ensure that only authorized users can access it.
- **Data quality and lineage tracking:** Monitoring data quality to ensure accuracy and data lineage tracking for reliability.

The foundation layer primarily serves the data platform team, including hands-on technical roles such as data engineers, stewards, and architects. The goal is to build the capabilities in this layer as self-service so that non-technical users can easily invoke them. The capabilities required in the foundational layer can be found in various metadata management, data ingestion, data transformation, data quality, and data security tools.

THE MANAGEMENT LAYER

The data product management layer — a new concept — has the capabilities to fulfill the needs of the emerging Data Product Manager (or team) role. This layer offers capabilities to develop, manage, and maintain trusted, domain-oriented data products. It relies on the services in the foundational plane to inventory, move, transform, and secure data assets. It serves as a bridge between the raw metadata residing in the foundational layer and the actionable insights that consumers seek from their data.

The key capabilities within this layer are:

- **Data product composition:** Capabilities to engineer a data product, define the inputs and outputs, and configure all its various artifacts.
- **Data product deployment:** Version management and the workflow to activate a data product when it's ready for consumption by end users.
- **Data product maintenance:** Usage tracking, feedback gathering, changes and updates, and capturing all stages of the data product lifecycle described in the previous sections.

The data product management layer primarily serves the product management team, especially Data Product Managers, who steer the life cycle of data products to meet business objectives.

THE CONSUMPTION (A.K.A. MARKETPLACE) LAYER

The marketplace layer sits at the top of the data product catalog, representing the consumer-facing part of the system. It enables self-service data exploration and consumption, fostering a data-driven culture within the organization.

The key capabilities within this layer include:

- **Search and discovery:** Intuitive interfaces allow users to easily explore and discover the data products that best suit their needs.
- **Learn and trust:** Data consumers can understand the data product by viewing both metadata and sample data, learn how to use it by looking at sample code and data documentation, and gain confidence in it by reviewing the data quality and lineage.
- **Access and permissions:** This layer controls who can access specific data products, ensuring data security and compliance with data governance policies.
- **Data consumption interfaces:** APIs, visualization tools, or even direct database access points provide users with the means to consume the data products. Depending on the product, different interfaces might be available to cater to diverse user needs and skill sets.

The data consumption layer primarily serves business analysts, data scientists, line-of-business users, and other key business decision-makers.

BENEFITS OF DATA PRODUCTS TO DIFFERENT PLAYERS IN THE DATA ECOSYSTEM

Data is an organization's greatest asset, but its true potential can be realized only if it's readily accessible, well understood, and used effectively. Data products bridge this gap, transforming raw data into consumable, curated assets that empower stakeholders across the organization.

Let's explore how data products impact the key players in the data ecosystem:



For **data engineering teams**, data products are a common language — a well-defined agreement — that bring structure and clarify communication between data producers and consumers. This allows producers to develop reusable data products in an iterative fashion, reducing the total cost of ownership (TCO). In addition, the data engineering team doesn't need to build large centralized and complex data pipelines; they can focus on developing smaller, independent pipelines dedicated to each product's implementation.



For **data governance teams**, data products become a central audit point — and since the products are trusted and governed, it becomes simpler to minimize risk. The governance team centrally defines the policies and standards; different data product teams enforce the policies in a federated manner.



For **data product managers**, data products better align data producers and consumers, driving more collaboration and accountability across the data ecosystem. This brings ownership and accountability to data.



For **data consumers**, data products are domain-centric, curated data assets that deliver exactly what's needed to maximize business value. This reduces rework, leading to faster outcomes.

Together, these stakeholders can leverage a data product catalog to harness the full value of their data assets efficiently, drive data-driven decision making, and achieve their business objectives.

“ In the mortgage industry, data products go beyond mere numbers and algorithms — we view them as transformation agents. By defining curated self-service data products, we want to empower our loan officers to offer the right mortgage solution for the right borrower, while the finance teams can better evaluate costs and control risk. The approach of leveraging data products in our business is truly redefining the future of mortgages for us. ”

Julia Fryk

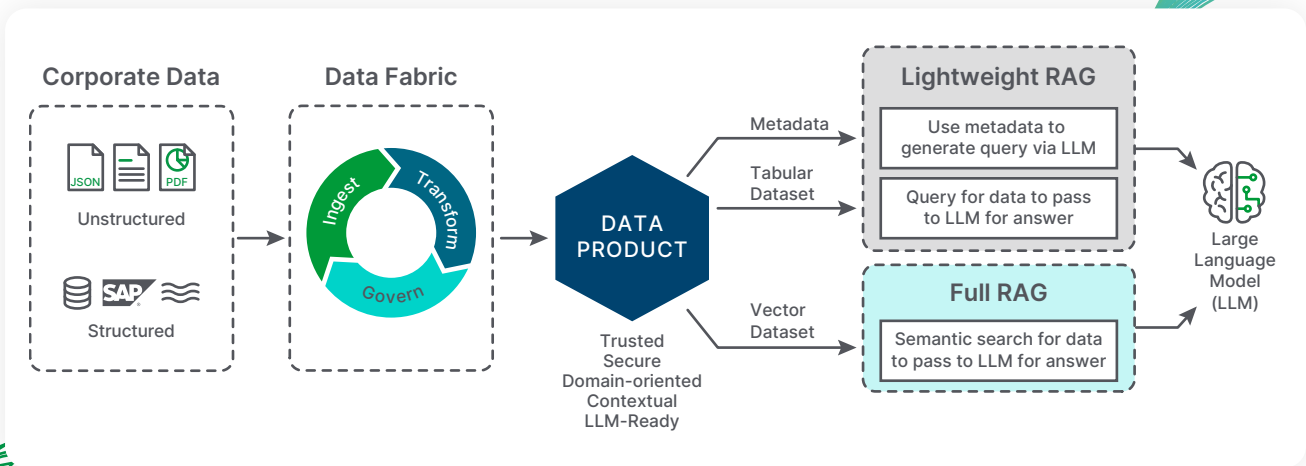
Principal Data Architect, Waterstone Mortgage

Delivering contextual data to LLMs with data products

Generative AI is creating tremendous opportunities for productivity enhancements in every aspect of the business. At the heart of GenAI are foundation models — such as LLMs like GPT-3, Anthropic, Llama, and BERT — that have been pre-trained on a massive corpus of text data (web pages, documents, etc.). However, foundational LLMs do not understand internal, company-specific data. That’s why they have to first be grounded with corporate data.

One of the most effective techniques for providing corporate data to LLMs is through a mechanism called RAG, or retrieval-augmented generation. As the name suggests, contextual data related to a user prompt (question) is **retrieved**, and then the user prompt is **augmented** with this data and passed to the LLM to generate an answer. There are 2 RAG methods, lightweight RAG and full RAG.

- In lightweight RAG (for structured data), the RAG application passes the metadata to the LLM to generate a SQL query. Then the query is executed on the tabular dataset, and the results are passed to the LLM to generate an answer. This method is used when an explicit answer is required from the data.
- In full RAG (for structured or unstructured data), the data is first vectorized. Then the RAG application performs a semantic search on this vectorized data to retrieve the data. That data is then passed to the LLM to generate an answer.



In both methods, it’s important to have high-quality, clean, and contextual data available for RAG.

Data products provide contextual and domain-oriented data in different ways for consumption.

As mentioned earlier, data products contain rich metadata (business, technical, and domain-oriented) and highly curated datasets (files, tables, or vectorized) for consumption. Hence, data products become the perfect vehicle for enabling RAG-based applications.

Building a RAG application starts with first searching the data product catalog/marketplace for the right data product to use based on the question being posed. For example, a question related to sales may be best served by an “Opportunity” data product sourced from the CRM, based on the detailed metadata associated with the data product. Or a question related to an order may be best served by a “Sales Order” data product sourced from SAP, based on the detailed metadata associated with that data product. Once the data product has been identified, the RAG application can directly query the data product dataset — tabular or vectorized — based on the use case.

Conclusion

As data continues to drive modern business, data products become increasingly critical to **unlock the full value of data** and accelerate business outcomes while **enhancing data trust and governance**. At Qlik, we are committed to assisting you on this journey as you construct your data foundation for self-service capabilities. This involves empowering your data teams to build and manage data products while granting data consumers the ability to swiftly search, discover, and utilize data products to drive successful business outcomes.

One final note: While data products are powerful tools, they're not a silver-bullet solution. The real power of data products lies in identifying business needs, having a well-defined federated data strategy, ensuring effective governance, and collaborating across teams.

Want to unlock
the full value of
your data?

Learn More

“ Organizations are increasingly seeking to create domain-centric data products that transcend data integration and quality. The focus of Qlik’s data quality and governance offerings to embody the concept of data products highlights a maturity that resonates with today’s customer needs. ”

Stewart Bond

VP, Data Intelligence and Integration Software, IDC



About Qlik

Qlik transforms complex data landscapes into actionable insights, driving strategic business outcomes. Serving over 40,000 global customers, our portfolio leverages advanced, enterprise-grade AI/ML and pervasive data quality. We excel in data integration and governance, offering comprehensive solutions that work with diverse data sources. Intuitive and real-time analytics from Qlik uncover hidden patterns, empowering teams to address complex challenges and seize new opportunities. Our AI/ML tools, both practical and scalable, lead to better decisions, faster. As strategic partners, our platform-agnostic technology and expertise make our customers more competitive.

[qlik.com](https://www.qlik.com)