

Build a modern, distributed Data Mesh with Google Cloud

Firat Tekiner, Thinh Ha, Johan Picard,
Victor Crowther, Susan Pierce

The building blocks of
Data Mesh

[Page 4](#)

Why Google Cloud is a
game changer

[Page 11](#)

Example Data Mesh
architecture in Google
Cloud

[Page 14](#)



Introduction

In an ideal world, every data user at a company has timely access to the data and insights they need, in the format they need, using tools that are familiar and comfortable to them. Teams spend little to no time managing data silos or building and maintaining pipelines to incorporate data. Instead, they can focus their efforts on deriving insights that impact product and business decisions.

A common first step along the pathway of data maturity is to remove silos by centralizing data storage. This reduces costs and is made relatively easy in the age of cloud computing and cloud-native data storage and analytics products. However, the organizational impact of this architectural change isn't always positive.

For example, as your data teams attempt to work in a centralized data model, they might rely on a central IT team to incorporate new data sources. This causes delays in deriving insights from data and results in stale data when pipelines are not kept up to date. These teams only have access to limited datasets which reduces innovation company-wide. As teams of data workers across the company grow and expand in scope, often the central IT team absorbs the additional overhead of ensuring that data is produced, processed, and incorporated in new ways. In the process, they also expand their own skill sets to match the different technologies being used across the company. You may have experienced this in your organization as it leads to delays in data processing and friction between the data owners (in this case, centralized IT) and the data users.

What is more concerning than delays and stale data is data without the proper context. When a team owns data that has been produced elsewhere in the company, they often are missing the business context for that data. They don't know the importance of various data sources, how the data is used downstream, and what combinations of data may yield rich insights. Having additional context can impact the technical decisions made about data processing, while the lack of that context can lead to additional processing overhead, higher total cost of ownership for systems, and additional technical debt in the form of patches and short-term fixes. When a centralized team acts as owner, data is treated as a liability to be managed rather than an asset that can be leveraged across the organization.

When a team owns data that has been produced elsewhere in the company, they often are missing the business context for that data.

Between the extremes of complete democratization and siloed ownership, we propose a middle ground: a data mesh

The other extreme is to fully decentralize data, letting teams handle not only their own processing and generation of insights but also the management of their own infrastructure and data policies. In this approach, teams frequently find that to get the right set of information and insights, they need to create multiple copies of existing data sets which leads to a proliferation of derived data in silos. This in turn contributes to data staleness, lack of quality insights, and an inability to follow regulations. Decentralized data also reinforces a culture of one-off development, where new pipelines are created for each project and never reused. Downstream consumers of the data lose confidence in the various data sources available and are unable to join across them effectively. The lack of confidence in data has a massive impact with [90% of employees](#) saying that their work is slowed by unreliable data sources. This also creates a huge risk of access to the wrong data, untrusted data, which has impact on outcomes.

What if there was a way to bridge decentralized data ownership with centralized data management and governance, ensuring discoverability of high-value datasets across an organization? It's possible, but it takes some work. Without a broad, organization-wide change to the way data is owned and treated by different teams, any attempt to modernize data analytics is going to cause unnecessary friction.

In our previous paper on building an [analytics platform](#), we briefly discussed two architectural constructs that can be used to address organizational conflict in data centralization and ownership: data warehouses and data lakes.

Data Lakehouse relies on a centralized IT team and infrastructure that can accommodate many data producing and consuming teams. Lakehouse brings the data warehouse and data lake together, allowing different types and higher volumes of data. This effectively leads to schema-on-read instead of schema-on-write, a feature of data lakes that was thought to close some performance gaps in enterprise data warehouses. You can read more about Lakehouse in our [recent open Data Lakehouse article](#).

Data Mesh, the focus of this paper, explores the idea of distributed data ownership, allowing teams to rely less on a centralized IT team and preserve context, while the organization's IT team is responsible for overall governance, efficiency, and infrastructure allocation. Data Mesh allows organizations to think more strategically about data as a product and distribute data ownership across teams that have the most business context. However, this approach requires buy-in from IT, data teams, and lines of business in order to be successful.

In the following sections, we will introduce the building blocks of Data Mesh, explore an implementation on Google Cloud, and discuss the solutions to potential challenges in adopting the approach. Before we get into the details of Data Mesh, it's important to understand the different users within an organization who play a part in the data lifecycle. Within each data domain, the roles include data product owners, data consumers, data stewards, and data domains. To see how we define each of these user types, check out the terminology section at the end of the whitepaper.

The building blocks of Data Mesh

Discovery

[From central data platform to federated data marketplace](#)

Valuable data for analytics and AI/ML across an organization are often locked away in database systems that are hard to access. Usually, this data is optimized for short-term uses by the team that produces it. As a result, it is usually not normalized for use by other teams or purposes within a company. Efforts to integrate data across the organization into a single analytics database often struggle to keep up with business demands for fresh, complete, and trustworthy data.

By solving the technological bottleneck in providing self-service data and analytics infrastructure, you have the opportunity to transform your processes, culture, and operating model to accelerate business value realization. The Data Mesh architecture [as articulated by Zhamak Dehghani at ThoughtWorks](#) offers a perspective on how this can be achieved.

Instead of relying on a central data team consisting of technology specialists, the Data Mesh architecture proposes the distribution of responsibility for creating high-quality data assets to federated domain teams. These teams are subject matter experts (SMEs) on data that originates from their domain and are therefore best suited to organize and generate value from this data. **Conceptually, each data domain maintains its own data warehouse, but these individual data warehouses can combine to form a Data Mesh.**

Instead of building a centrally-defined data model, the Data Mesh architecture proposes organizing data assets along a logical data domain structure as outlined in **Figure 1**. This effectively applies [Bounded Context](#) from [Domain-Driven Design](#) to the data and analytics space within an organization. Each domain owner is responsible for promoting the use of analytics and data-driven decision making in their area and thus across the business. Once published, data products can be easily discovered via a unified data access interface for consumption.

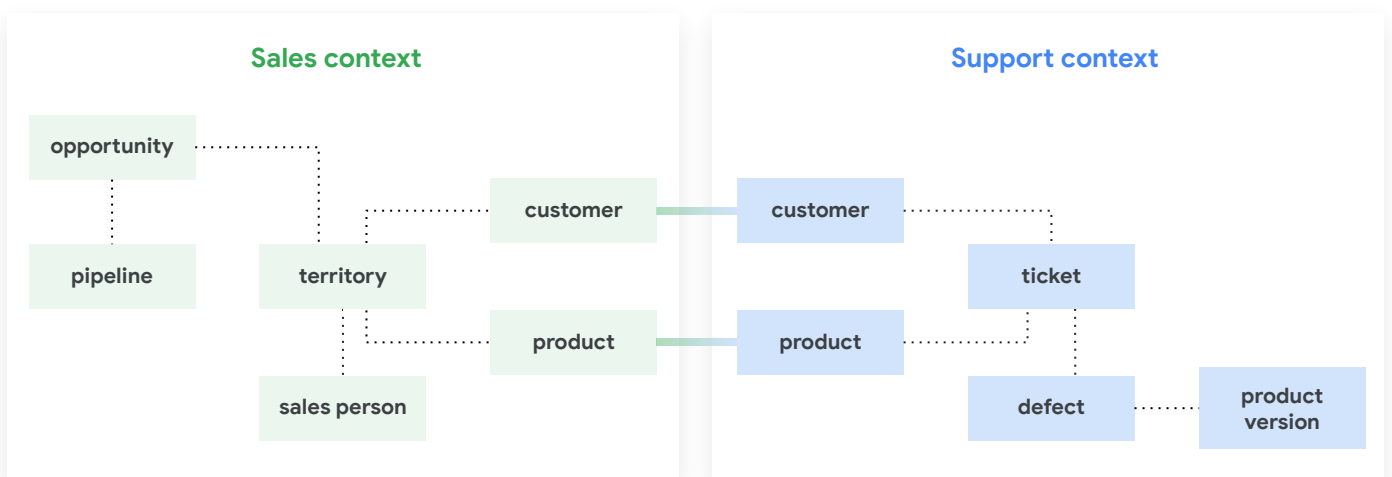


Figure 1: Organizing data assets logically

The decentralized operating model described by Data Mesh is not necessarily novel. Google is a distributed organization with high autonomy within each business unit, each with its own data and analytics functions. In many ways, we have already been operating for over a decade according to several core Data Mesh principles. At Google, distributed teams can publish data products into a unified Data Marketplace where data is organized, managed, and made readily accessible to analytics users within Google. In this

Data Marketplace, data producers can publish datasets for others to consume. Data consumers can easily discover and explore available datasets, request data access, and integrate the Data Marketplace into their products. We believe this to be a key Data Mesh principle to make accessibility of high-quality and trustworthy data products a primary concern, regardless of the underlying tooling or technology. **Figure 2** provides an example user interface of what a Data Marketplace can look like.

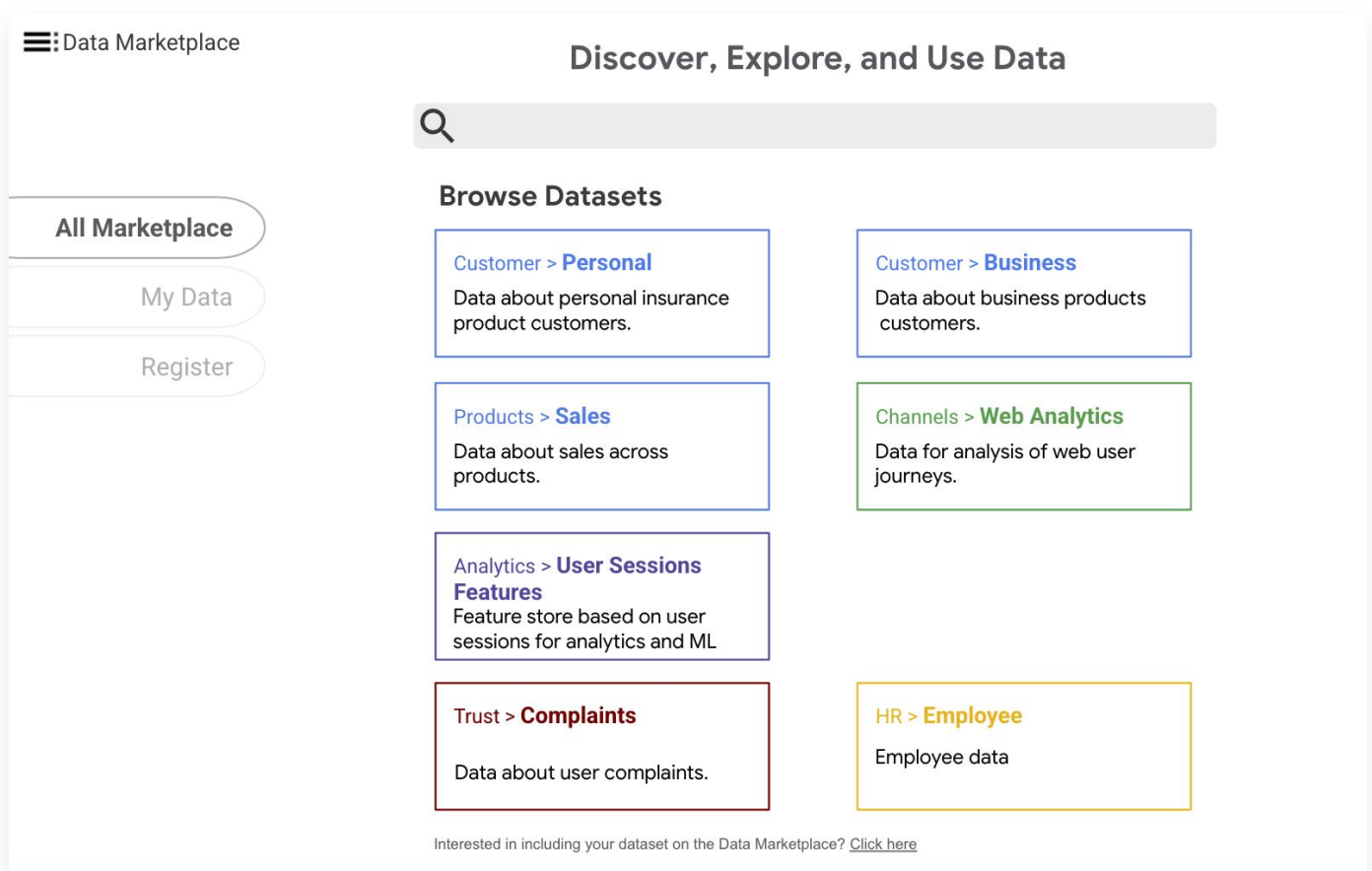


Figure 2: Example Data Marketplace for searching data

Accessibility

Trusted data products as a primary concern

Once a data consumer discovers a dataset, they need to understand its content, whether it can satisfy their use cases, and whether the quality and trustworthiness of the data are sufficient to base business decisions on. In many organizations, data assets are made available with little metadata or documentation to help consumers make informed decisions about their usefulness or purpose. Without clear ownership over datasets being produced, finding a SME to describe how a table was created and the meaning of its content can be a challenge. This creates a bottleneck in data understanding that can inhibit scaled adoption of data-driven decision making across the business.

In a Data Mesh, the producer or publisher of the data is responsible for providing useful information about the data. Every published data asset is accompanied by a [Data Card \(example Data Card\)](#) containing standardized metadata fields that producers can update to include information about its data model and dictionary, how the data was collected, the data security and privacy classifications, and how to access the data. This form is in effect a Data Sharing Agreement that can be used to drive data management controls across the organization. **Figure 3** provides an example of how this metadata can be used in practice. For example, we could check the Data Sharing Agreement to verify whether a financial report is using high-quality data, or whether the access policy for a particular dataset complies with the organization's privacy policy.

To assure downstream data consumers that they can base their critical business functions on a particular data product, data product owners can also publish

Discover, Explore, and Use Data



App > **User Sessions** : Data Card

About this data

.....

.....

Data model and Dictionary

.....

.....

How was this data collected and recommended uses

.....

.....

Data SLO

.....

.....

Who can access this data

.....

.....

[Access data](#)

Interested in including your dataset on the Data Marketplace? [Click here](#)

Figure 3: Example Data Card

guarantees regarding the [service level objectives \(SLOs\)](#). This then allows them to offer metrics such as data freshness and data quality. The SLOs express the promises that the data owners make to their data consumers. This also forms the basis for objective, data-driven conversations around service level expectations across engineering and business stakeholders alike. Achieving 100% SLO is realistic in some cases and infeasible in others. Focusing engineering attention on certain SLOs may require deprioritizing other SLOs. Choosing the relevant SLOs should be a collaborative effort between the data producer and data consumers. This allows the organization to prioritize engineering efforts towards making measurable improvements on the usability of data products on dimensions that matter to the business' performance goals.

With information about data assets readily accessible, data consumers can easily search for relevant datasets to their use case. In return, they can quickly get up to speed on what the dataset contains and make informed decisions about its consumption.

We can measure the value or success of a data product based on how often it is consumed or the number of derived data products created from it. By aligning metrics that express the value of a data product and by measuring the ease at which it can be consumed, the Data Marketplace recognizes the value created by data assets beyond the lifecycle of a single project or a team. It introduces metrics that can inform organizational incentives for data product owners to continue curating and distributing useful data assets for their organization.

Ownership

Domain-oriented federation and responsibility

In a Data Mesh, the responsibility for generating value from data is federated to the people who understand it best. In other words, the people who created the data or brought it into the organization are also responsible for creating data products from the data they produced. Establishing a useful data product necessitates having understanding and influence over how data is produced in the source systems. Therefore, the role of a data product owner is formally introduced to take on the responsibilities of onboarding data into the organization, building useful data products from this data, and promoting its consumption. Data Mesh building blocks from data product point of view is outlined as in **Figure 4**.

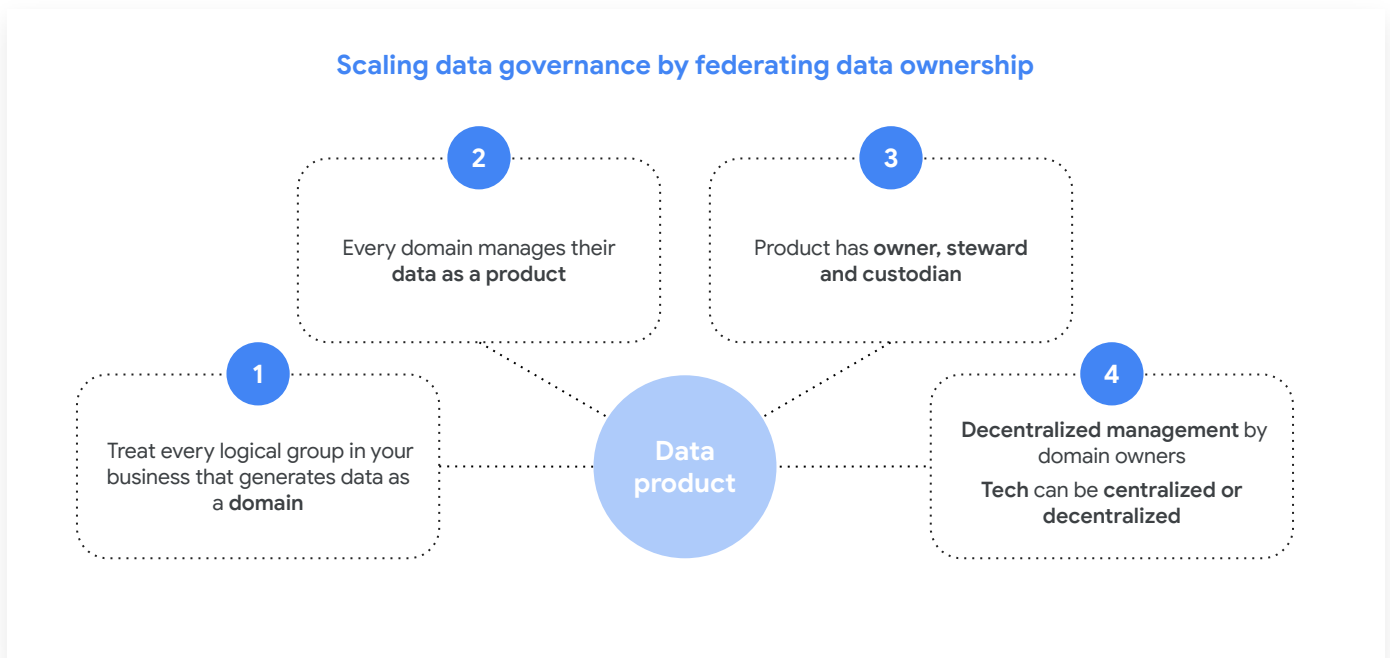


Figure 4: Data Mesh building blocks from data product point of view

In many organizations, establishing a single source of truth or authoritative data source is difficult due to the repeated extraction and transformation of data across the organization. This is because there isn't clear ownership of or responsibility for the newly created data. With Data Mesh, the authoritative data source is the data product published by the source domain, with a clearly assigned data product owner and data stewards who are responsible for that data.

If a data consumer were to take this data and transform it into another format for consumption by other teams, they would become the data producer of the transformed data and therefore its data owner. Data without owners or data that has been orphaned must go through a review process to either assign a new owner or be scheduled for deletion if they no longer serve a purpose in the business.

Aligning roles and responsibilities as in **Figure 5** for creating high-quality data products to the teams who produced the data addresses two problems:

1. Encourages data quality issues to be resolved where it is produced in the data origination domain, rather than relying on downstream consumers to patch this data
2. Reduces external or cross-team dependencies on data SMEs who are most informed about how the data was produced and what the data means.

As data SMEs are often aligned to the data-producing domains, data quality issues can be resolved locally within the same team. This can be achieved before it is packaged into a more readily-consumable data product for external consumption.

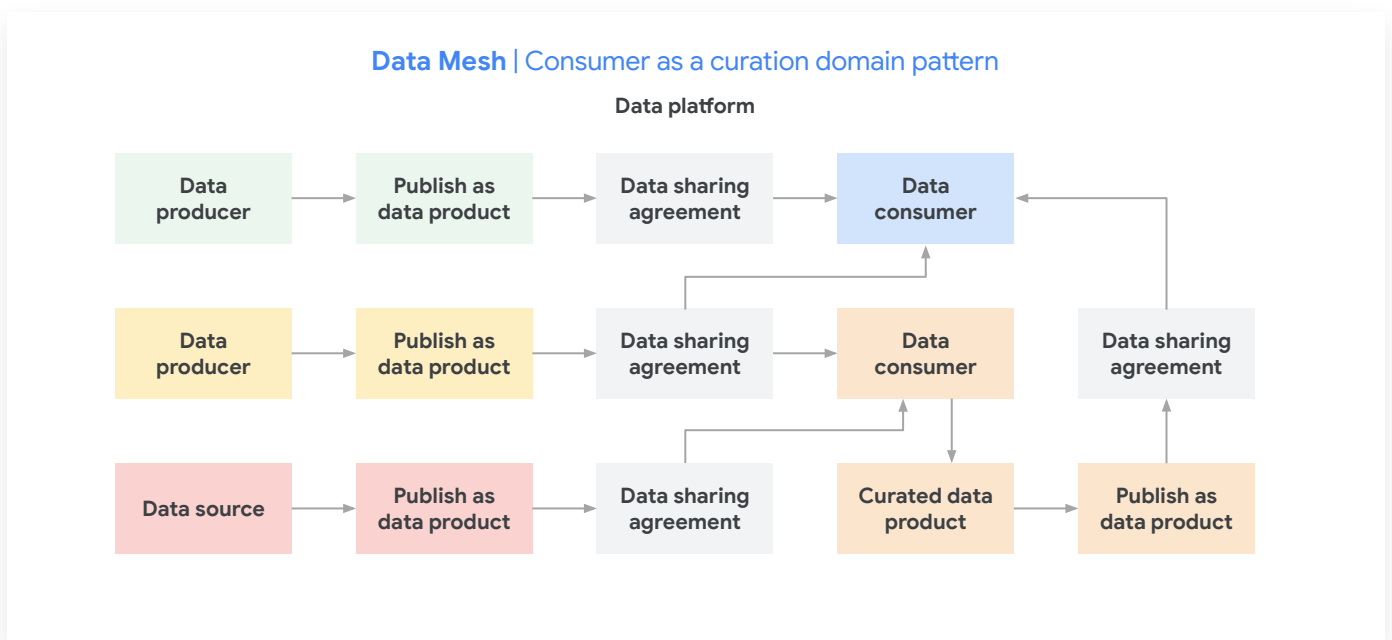


Figure 5: Consumer as a curation domain pattern

Governance

Embedded data controls in development processes and data exchange interfaces

By establishing a consistent data publishing and consumption interface via the Data Marketplace, we gain the ability to develop a set of automated controls targeting the same API interfaces.

Data access logs or usage data for products in the Data Marketplace can be easily analyzed for audit and compliance purposes as they follow a standardized format as seen in **Figure 6**. It is possible to understand how a data product is being consumed, which products are derived from each other, whether data consumption is following acceptable use guidelines, or how data products can be improved to improve query efficiency for its consumers.

We can index information provided in Data Sharing Agreements in a search engine to make data products and their metadata easily discoverable and understandable by consumers.

Data Sharing Agreements are designed to promote data consumption towards high quality and governed data products. To encourage data products to reach higher levels of data management maturity, we can directly measure the maturity of each data product by the completeness of its Data Sharing Agreements. Once data products reach a certain level of maturity, they can become certified via an official process to become an authoritative data source, and therefore more likely to be recommended for consumption by the Data Marketplace.

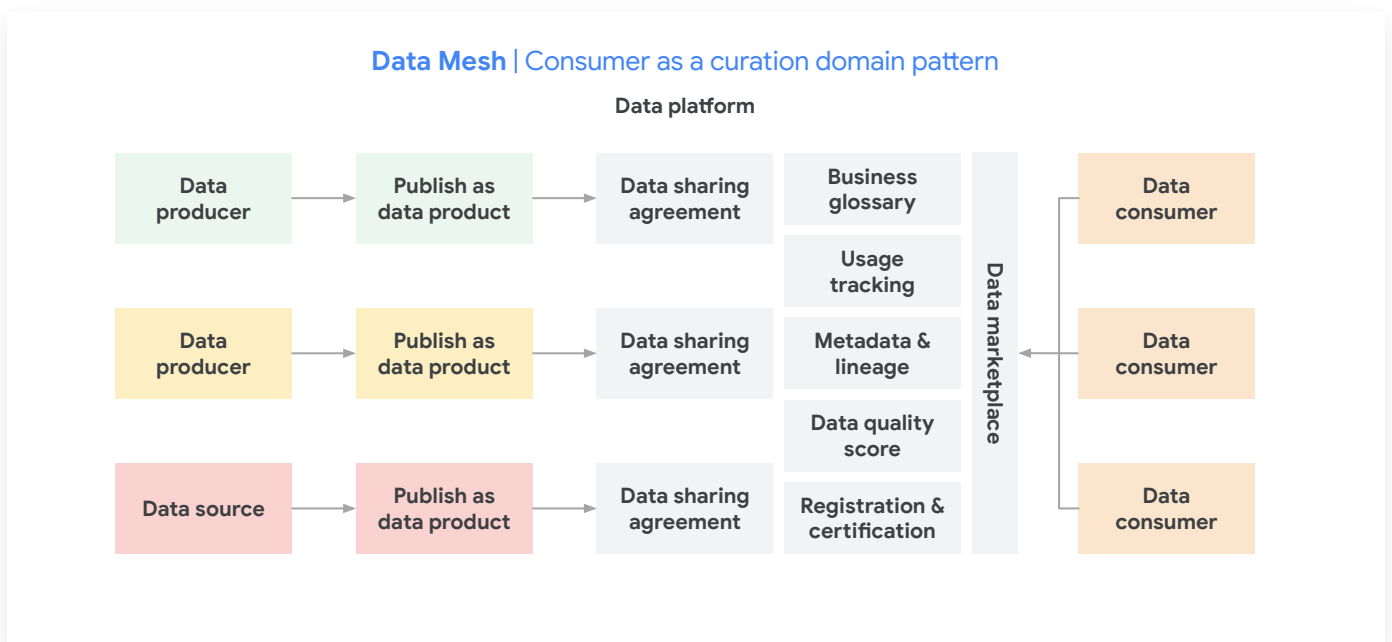


Figure 6: Federated data controls at consumption point

Data sets with better metadata completeness and data quality can be promoted in a curated catalog of data products to all data consumers. Certain data analytics journeys or compliance reports may be required to prove that they are consuming data from authoritative sources with sufficient data quality.

The annual [Google Accelerate State of Devops Report](#) shows that the highest performing organizations:

- Deploy code more frequently
- Have faster release cycles into production
- Recover from failures faster
- Have a lower failure rate than their competitors

All these points contribute to accelerated delivery of business value. This means that "[shifting left](#)" compliance and audit controls into the normal developer feedback cycle is essential for distributed organizations to maintain compliance with regulatory standards. The Data Mesh operating model applies this principle to data governance, mandating compliance controls to be implemented as computational policies that are embedded into standard engineering processes, and that can be continuously monitored and enforced in real-time. For these reasons, an effective Data Mesh implementation requires deep partnership with the security, risk, and compliance organizations as key stakeholders, as well as executive sponsorship to align objectives of the transformation journey to a common business goal.

Google's data cloud to the rescue

Google is uniquely positioned to help businesses transform, with data as the essential ingredient. Google products are used and loved by billions of users across the globe, bringing together the complex web of disconnected, disparate and rapidly changing data that makes up the internet. When you get an answer in milliseconds from google.com via a simple search bar, you know we have this down to a science. Google brings data and software together for businesses of all sizes looking to build a data cloud.

A data cloud offers a comprehensive and proven approach to cloud and embraces the full data lifecycle, from the systems that run your business, where data is born, to analytics that support decision making, to AI and ML that predict and automate the future. A data cloud allows you to securely unify data across your entire organization, so you can break down silos, increase agility, innovate faster, get value from your data, and support business transformation so you can be competitive. This is the heart of the data cloud as seen in **Figure 7**.

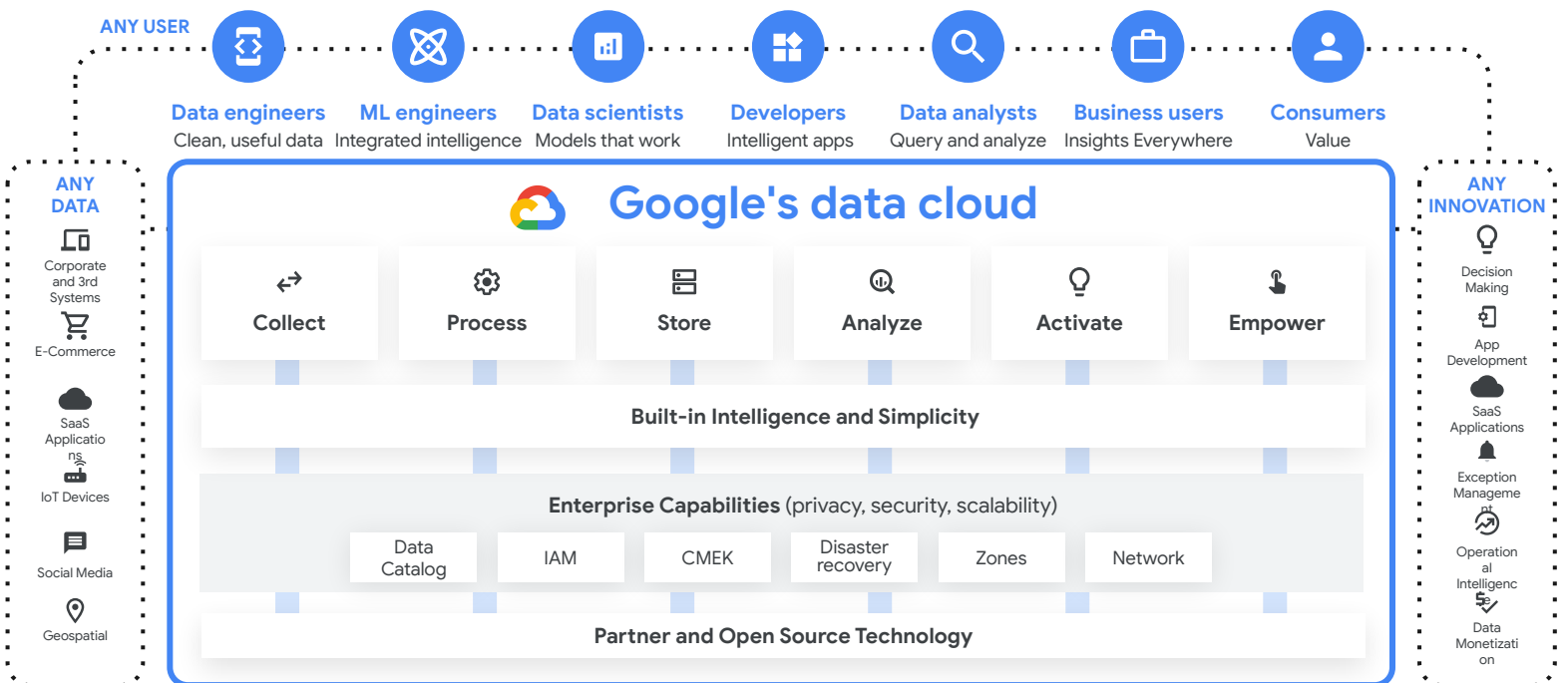


Figure 7: Data Cloud that handles industry's sophisticated needs

How Google Cloud is Designed for a Data Mesh Architecture?

In the previous sections, we covered the main Data Mesh concepts and how some of these concepts have been used internally in Google for over a decade. The good news is that most of these tools are also available as part of the Google's Data Cloud. Dataplex allows you to realize a Data Mesh by managing assets (whether in [BigQuery](#) Datasets, the [GCS Object store](#), or other storage systems) at scale and defining data domains within lakes and zones virtually layered over the actual physical location of the asset. With BigQuery, you no longer need to centralize engineering resources to provision and manage technical infrastructure for consumption by the business. With little to no tuning, BigQuery can process terabyte- to petabyte-scale datasets in seconds. In addition, BigQuery is at the heart of the data processing ecosystem, it allows not only federating access to databases such as [Bigtable](#), [Spanner](#) but provides direct access to data residing in GCS Buckets.

Google Cloud enables business users to create insights on the fly with self-service analytics using familiar tools such as SQL or BI. The flexible, on-demand compute power of BigQuery means that ad hoc analysis doesn't compete with scheduled reporting. Without being constrained by limitations of technical infrastructure, you can focus valuable engineering bandwidth on developing data into strategic assets for your organization, in turn allowing your business users to focus on generating business value from this data. One example of how Google Cloud can enable a serverless and integrated analytics architecture can be seen in **Figure 8**.

To get maximum value from data, Google Cloud allows organizations to use different execution engines, optimized for different workloads and personas, to run on top of the same data tiers. This is made possible because of the complete separation of compute and storage on Google Cloud.

Meeting users at their level of data access including SQL, Python, or more GUI-based methods means that technological skills do not limit their ability to use data for any job. Google Cloud's [Data Catalog](#), along with [Dataproc Metastore](#) and the BigQuery Metastore, allow for accessing data wherever it is located with any query engine. The process is simple: onboard the data asset in a [Dataplex](#) lake, which triggers discovery and publishes the data in BigQuery and Dataproc Metastore. Tools such as [Vertex AI](#) notebooks, Spark running on Dataproc, BigQuery, or Serverless Spark can then be easily integrated into the workflow. With BigQuery and Serverless Spark, data engineers can spend all their time on the code and logic. They do not need to manage clusters or tune infrastructure. They submit SQL or PySpark jobs from their interface of choice and processing is auto-scaled to match the needs of the job.

While BigQuery and Google Cloud Storage can serve as the infrastructure to store and process data, Dataplex then provides a single pane of glass to organize, govern, and index data in the Data Mesh.

Dataplex provides a managed service that enables enterprises to group data containers from across their Google Cloud deployment and organize them in a hierarchy — effectively overlaying a Data Mesh virtually on top of the Google Cloud project hierarchy — with zones representing data domains in the Data Mesh. You can also curate, catalog, secure, integrate, and explore any type of data at any scale with an integrated experience. To support building a Data Mesh, Dataplex enables quickly built lakes without needing to acquire and worry about different resources. It expands automatic data discovery and schema inference across different systems so that once you add a resource to Dataplex, it can be represented in the Data Mesh.

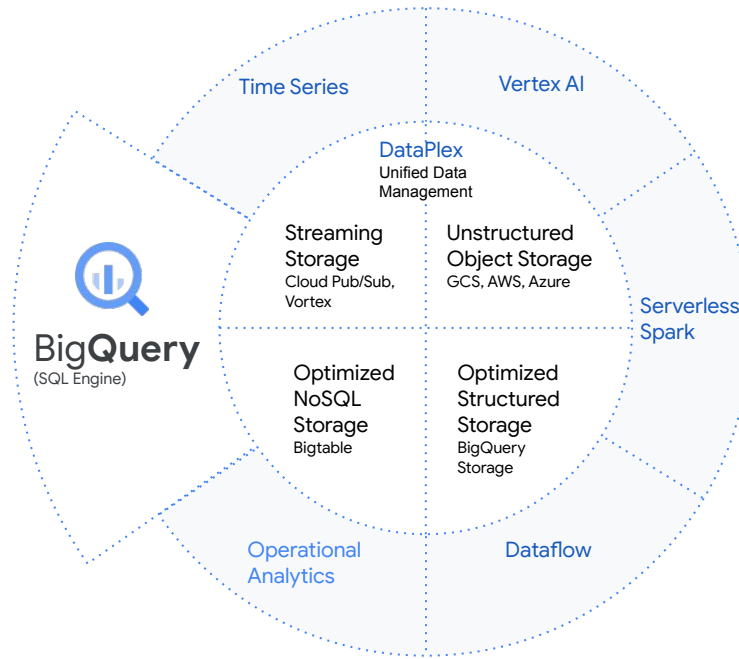


Figure 8: Serverless and integrated Data Cloud

Dataplex complements this by automatically registering metadata as tables and filesets as metastores and Data Catalog. Furthermore, with our integrated [Cloud Data Loss Prevention API](#) (DLP) and built-in data quality checks, the tagging of sensitive data is tightly integrated.

Dataplex enables a Data Mesh with simple security controls, as outlined in **Figure 9**. Consistent security policy and enforcement across Cloud Storage and BigQuery is enabled out of the box and allows central governance teams to audit the environment. Furthermore, managed Data Lake Storage with fine-grained access control, ACID transactions on files, and a BigQuery is managed through a single interface.

As mentioned in the earlier section, data ownership and governance is key to the success of Data Mesh and requires federated data sources with integrated governance. Dataplex offers a data management and governance layer so that you can organize your data

across BigQuery and Cloud storage with centralized governance rules. This means that data administrators can set up and manage workspaces together with appropriate environment profiles, including compute parameters, libraries, etc. At the same time, they can control user access and manage costs through one seamless interface.

With Dataplex, data scientists have one-click access to notebooks. Further, they can discover notebooks by using a notebook repository with links to associated data while being able to save and share notebooks as if they were sharing another asset within the organization. Data analysts can use SQL Workspace for ad hoc analysis without being dependent on any data processing environment. Effectively, through a single pane of glass they will be able to use Presto, Hive, or BigQuery without needing to access various environments.

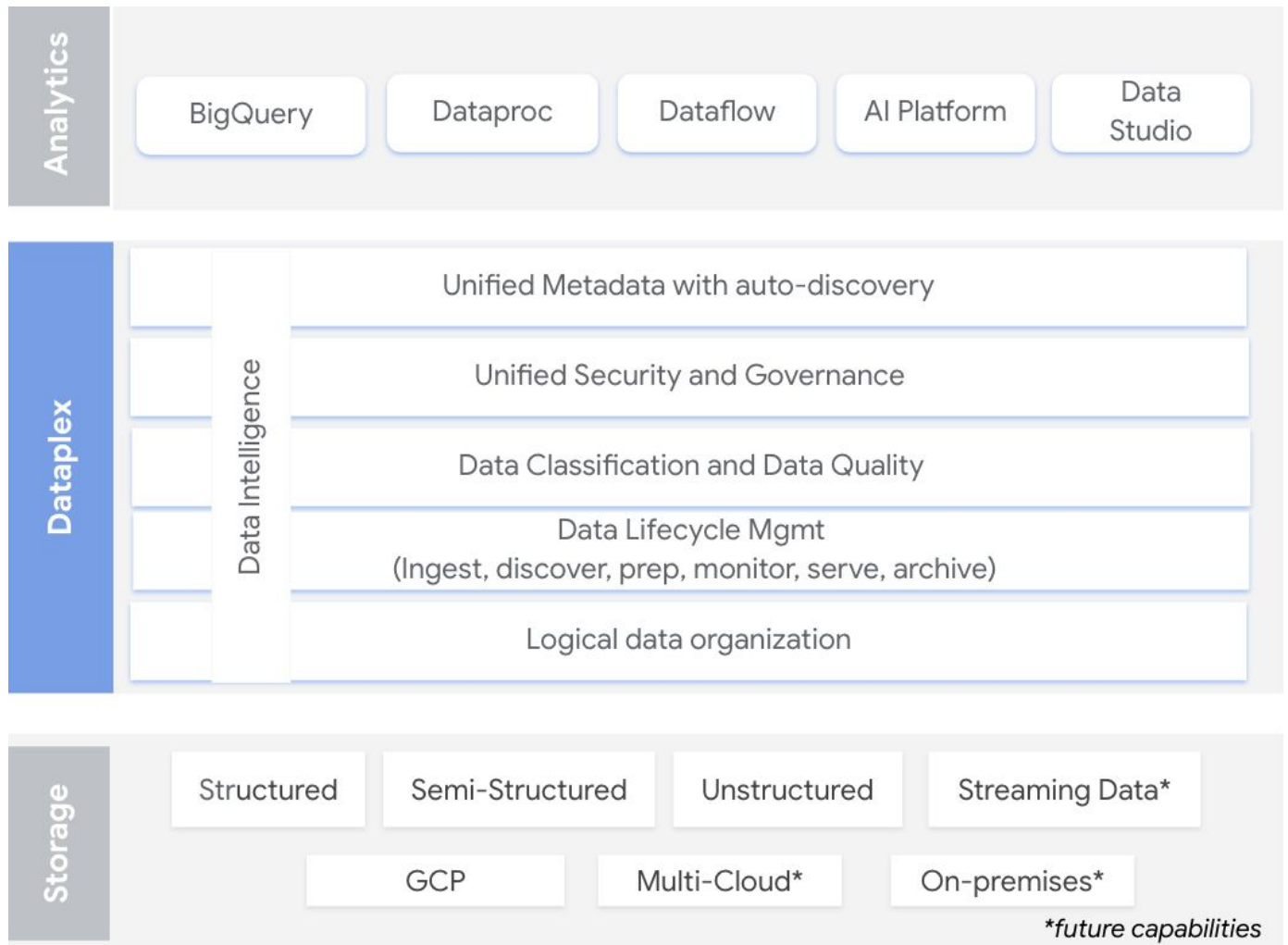


Figure 9: Single pane of glass - Dataplex

Last, but not least, data access is made simple and straightforward. An integrated experience across all Google Cloud data analytics services provides virtual Lakehouse experiences. This is complemented with an integrated serverless notebook experience with

serverless Spark for data science. All Cloud Storage data is automatically made queryable through OSS tools and BigQuery while enabling search and discovery across the board by using Data Catalog.

Example Data Mesh architecture in Google Cloud

Let us take a simple scenario of a Data Mesh comprising the following three data domains:

- Sales
- Supply chain
- Products

Each data domain can host several products, each of which is managed, owned, and maintained by their respective teams through their lifecycle and publication. These data product teams will leverage a Google Cloud based, self-service data platform and will be supported by a central team defining IAM, modeling, security, compliance, and architecture best practices. An example of what a data product looks like in Google Cloud context can be found in **Figure 10**.

Once the data processing and storage requirements have been dealt with, it is essential to provide standard interfaces for accessing the data. For

example, such an interface can be as easy as providing access to files stored in Cloud Storage, tables stored in BigQuery, or any other standard data interface specified by the governance standards. This can then be accessed directly via another Google Cloud project (Data Lake with Dataplex) if the team is given access rights. Since storage and computation are separated on Google Cloud, the data producer can choose to only expose the data itself without the need to think about providing the computational means to process it for the potential consumers. This alleviates an enormous operational burden for the product teams who do not have to think about managing the concurrent access to their data products or any potential performance issues that may arise due to lack of computational resources. In this way, users pay for the storage while the data subscriber will be charged for usage. This is a critical enabler of a Data Mesh approach that enabling billing and monitoring while and preventing data products from becoming silos in the future.

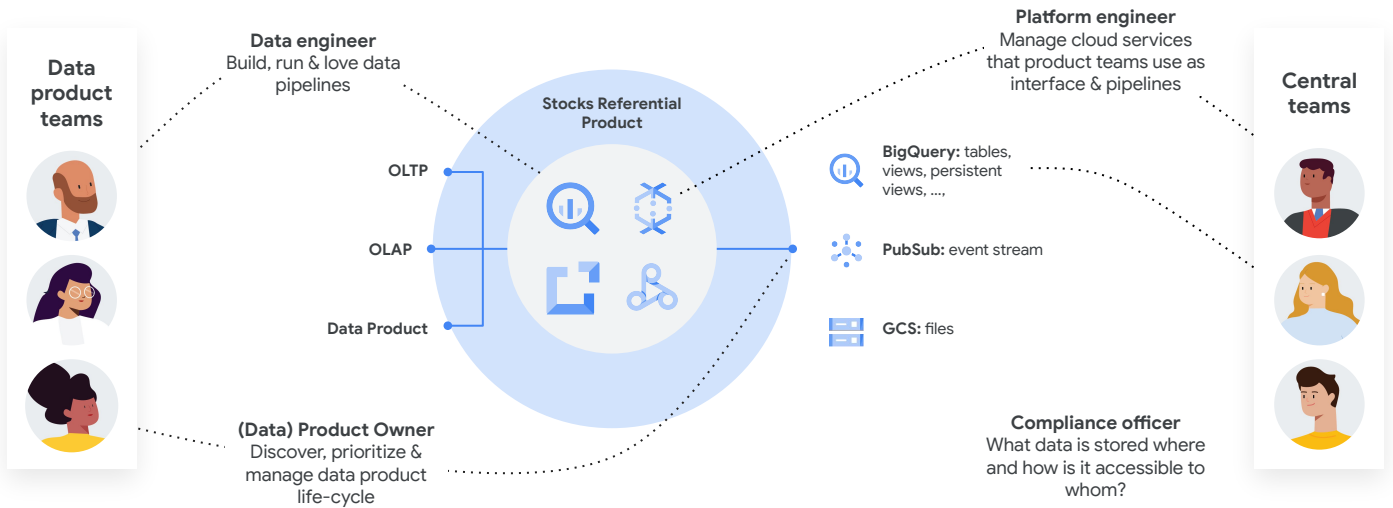


Figure 10: What a data product looks like

Figure 11 outlines a global picture of our sample scenario. Each data product is tied to a data domain and a sub-domain, where Data Lakes and Data Domains are governed by Dataplex.

As shown in **Figure 11**, a data product can be a set of files stored on Cloud Storage, a table exposed through BigQuery, a topic of events on [Pub/Sub](#), a dashboard on [Looker](#), a machine learning prediction endpoint from Vertex AI, or even a data application hosted on [Google Kubernetes Engine](#). Each data product is implemented in a Dataplex Lake with its

respective data pipelines, enabling the respective product teams to set up a very fine-grained permission control (including on the sub-lake, or zone, level) on each one independently. As defined by the company best practices, these products are to be versioned, documented, and published on the data catalog to ensure discoverability and usability. They can then be consumed and leveraged by other data products using the standard published interfaces such as APIs and datasets. In this scenario, the consumption and exchange of the various data products would look as in **Figure 12**.

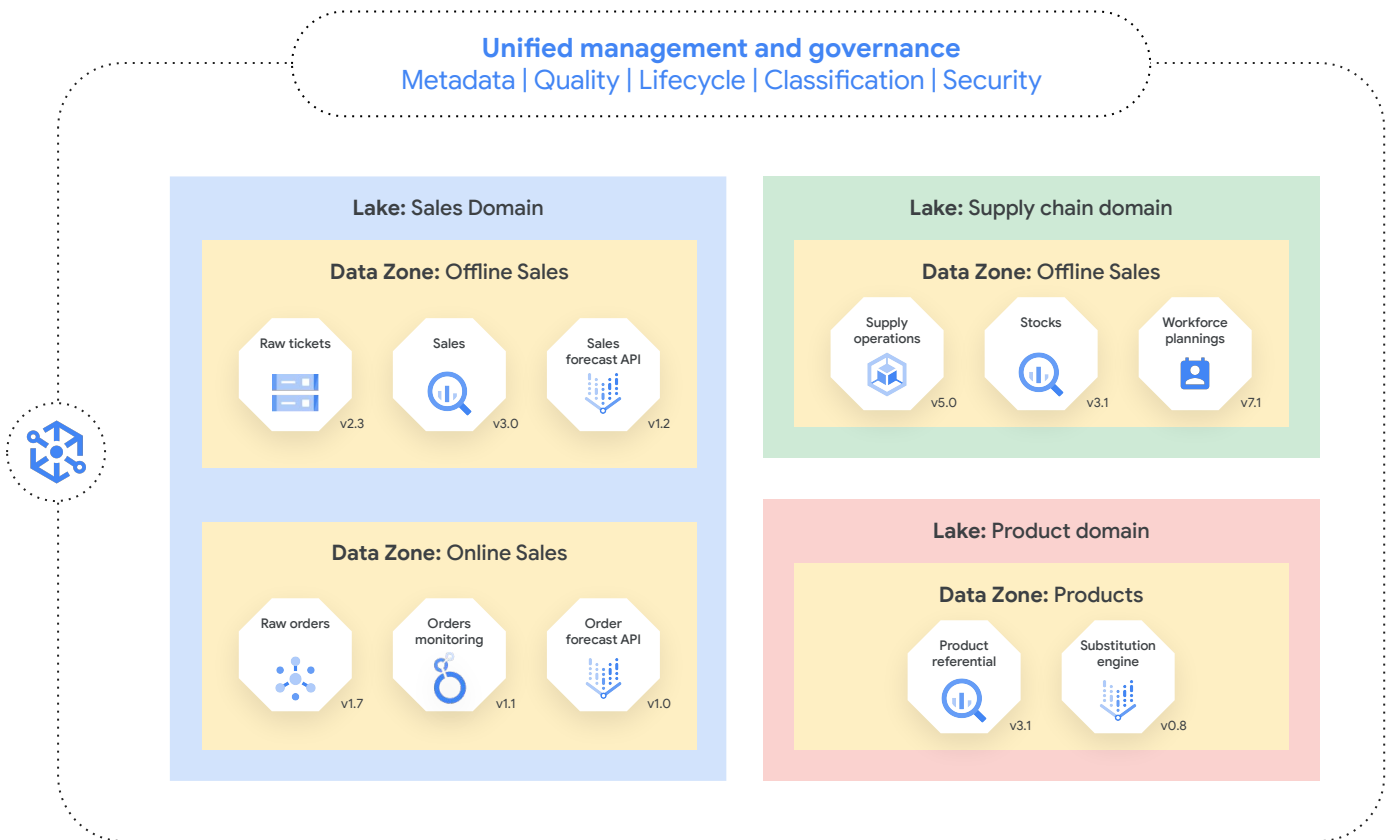


Figure 11: Data Mesh domains

Google Cloud

with a Data Mesh in Google Cloud, you can benefit from reduced friction and easier data consumption, under unified governance

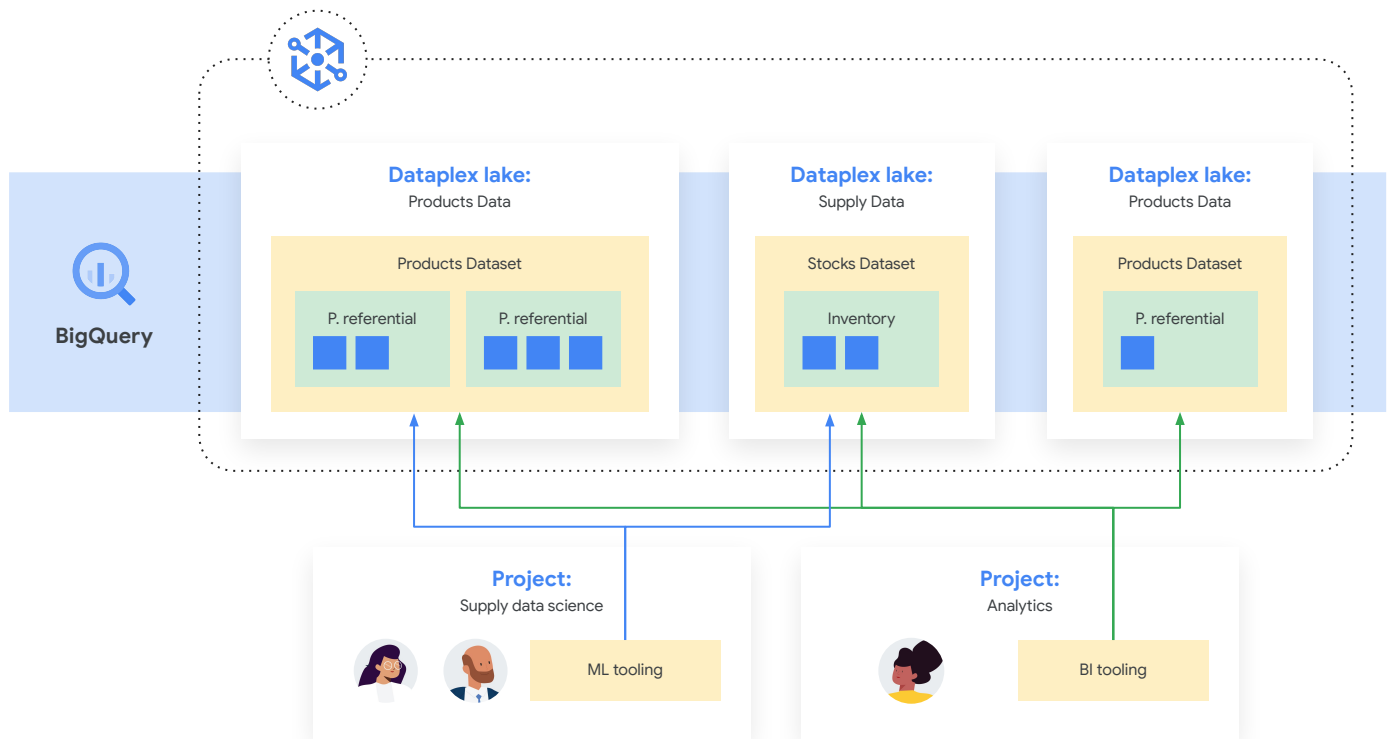


Figure 12: Data Mesh example architecture

Users will first search through the data product catalog for documentation regarding the desired domains and associated data products along with details about the data owners. They will then follow the specified access process, justify their demand, and get access to the relevant data products. They can now access the data directly from their own project with no need for duplication of the data. Since

a centrally and collegially defined data model has been defined for the basic exposition products, they can also join data coming from different domains, and do so with no data movement thanks to BigQuery's serverless engine. They will use their own resources to process the data, without access concurrency from other potential users, and be billed accordingly.

Conclusion and challenges

Data Mesh is appealing as a technical architecture paradigm because it promises solutions to break the organizational barriers that organizations face while enabling them to extract the most value from their data. To successfully adopt it, organizations must be willing to fully embrace the four principles of discoverability, accessibility, ownership & self service, and governance (federated), which require a coordinated effort across multiple teams and disciplines. This coordinated effort should start with the leadership buy-in as some concepts can be disruptive. In your organization, individual business units may need to own many more data curation and data management responsibilities than they currently do. This is in addition to developing data engineering and data stewardship talents internally to effectively leverage and scale Cloud technologies.

It is essential that the offices of the chief information security officer (CISO), Chief Data Officer (CDO), and chief information officer (CIO) are engaged as the key stakeholders as early as possible to enable business units to effectively manage the process. If your organization's key stakeholders are less involved in initial planning, this may result in inadequate resources being allocated and the project failing. **Fundamentally, Data Mesh is not just a technical architecture but rather an operating model shift towards distributed ownership of data and technology to enable business units to locally optimize for agility.**

Your organization may be predisposed for distributed ownership to naturally adopt similar operating models to Data Mesh. This is because you may already be using Google Cloud to provide a self-service technology and data platform that can enable individual business units to optimize for business value with fewer dependencies on any central technology and data teams. However, if you are coming from a brick-and-mortar type organization, you may not have been set up for distributed ownership over technology or data. This in itself is a fundamental operating model transformation shift that will require committed leadership buy-in to align the transformation journey to the overarching business strategy.

With solutions such as BigQuery, Cloud Storage, Dataplex, and Serverless Spark, Google Cloud is in a good position to reduce the friction and break the barriers of implementing Data Mesh. These solutions not only provide NoOps operating models but are designed to federate data production and consumption through different intents and personas. Furthermore, this is complemented by an integrated stack of security and privacy controls whereby data sharing is effortless but can be governed. This is further complemented by the existence of solutions such as Analytics Hub, which allows creating organizational and external data exchanges. As a result, we have been [working with customers](#) on their Data Mesh and Lakehouse journeys to create a unified analytics platform that allows them to get the most out of their data.

Looking to begin advancing your data journey? [Contact us](#) to get started.

Terminology

There are many users within an organization who have a part to play in the data lifecycle. Because a Data Mesh is decentralized in nature, different people within an organization may perform roles in the data lifecycle that are not captured by their job title. To avoid confusion and for completeness, we want to start out with a brief introduction to the different roles and key concepts involved in building and maintaining a Data Mesh.

Data product owner:

The people or teams that creates data and makes it available for others to consume. In addition to data creation and maintenance, they are responsible for providing useful information to preserve context and to make it more discoverable for other users. They also provide documentation and APIs to access the data. For example, the designated person from a CRM team who makes customer data available for downstream uses within the company while ensuring that Personal Identifiable Information (PII) is redacted or masked for most uses.

Data consumer:

The people or teams that access and use the data for their needs. In a modern organization, this group is effectively everyone — from data engineers, data analysts, and data scientists all the way to business users. An example here may be a product manager who performs an ad hoc query for fast insights, or a data scientist team who does deep discovery across datasets and builds models that cross data domains. This may also include an ML engineer who is responsible for the model to work without issues to end users in production systems, among others.

Data steward:

The people or teams responsible for defining and implementing policies and processes to collect, collate, and evaluate any issues that may arise with data. They are involved in data governance and make sure that data used is the right data, meets the data quality requirements and is accessed by the right personas. In our earlier example above, this person may define the organizational approach for data masking and redaction that is employed by the data product owner for obfuscating PII. In some organizations, data stewards report up to the legal department or would also be highly connected to internal risk and compliance teams. In others, they report to central IT or a chief data officer.

Data domain:

The subject area for a given set of data, usually the business context for a data product owner. For example, CRM systems and their data may fall within the sales data or CRM domain. **Within each domain, a dedicated data product owner and their team of data stewards and data engineers are incentivized to publish high-quality, useful, and trustworthy data assets for consumption by other teams.**

Build a modern, distributed Data Mesh with Google Cloud

February 2022

Interested in getting started? [Contact us](#) to learn more.