

E-BOOK

Data Security for Data Mesh Architectures

Benefits, Challenges, and Steps for a Secure Implementation

Table of Contents

Foreword	3
Understanding the Data Mesh	5
The Four Pillars of the Data Mesh	
Benefits of Data Mesh Architectures	8
The Complexity of Data Mesh Security	10
Common Data Mesh Security Challenges	
Regulatory Compliance Considerations	
3 Steps for Implementing a Secure Data Mesh	14
Step 1: Maintain Consistent Metadata	
Step 2: Employ Global & Local Policy Management	16
Step 3: Foster Organizational Change	19
Roche's Secure Data Mesh Implementation	21
Conclusion	22

Foreword

Despite computers and electronically-stored data being a commodity for nearly half a decade, the inability to efficiently realize value is a challenge that still plagues many data initiatives. Organizations invest substantial resources into collecting, storing, and managing vast troves of data, with the ultimate goal of turning it into actionable and valuable insights. Even so, a lack of efficient practices, platforms, and tools often results in a significant portion of data's potential remaining untapped.

When I first started working as a consultant in the data and analytics space seven years ago, I was faced with a grim reality: data was still largely a siloed resource within organizations. It was confined to the realms of data warehouses, accessible only to a select few who possessed the keys to unlock its potential. The analytics world often felt like a labyrinth, with data engineers seemingly being the only gatekeepers to retrieve any insights from data. Like many others, I was constantly looking to find ways that this could be done better–allowing more people to interact with data whilst retaining control over key principles such as quality, security and compliance.

I first came across the concept of data mesh when reading Zhamak Dehghani's articles where she outlined the complexities of the world of data, which I knew all too well from my own experience. Her idea of breaking down data silos and empowering domain experts to take ownership in the form of data products was a game-changer. Data mesh felt like a breath of fresh air in the stale corridors of traditional data management. The idea that data could be treated as a product, owned and managed by cross-functional teams, resonated deeply with me. It was a vision of a future where data democratization was not just a buzzword, but a reality.

Shortly after I first got introduced to the concept of data mesh, I had the fantastic opportunity to help Roche Diagnostics-a large pharmaceutical company headquartered in Switzerland-embark on their very own data mesh journey. My main responsibility was to develop a platform that would embed all of the data mesh principles into it. Given the decentralized nature of Roche Diagnostics' business model, the redesigned self-serve data infrastructure helped to alleviate the bottlenecks that had long hampered data engineering teams. Product-led thinking inspired a user-centric approach to data, emphasizing the delivery of data products tailored to the specific needs of users.

Moving from a centralized to a decentralized delivery model to empower self-service data management and ownership, however, posed many challenges-most notably upholding data privacy, security, and compliance without stifling innovation.

We at Immuta understand these challenges, as well as how they can be addressed. This eBook serves as a guide for those who, like me, are inspired by the possibilities of data mesh, and are eager to embark on their own journey of data transformation.



Claude Zwicker Senior Product Manager, Immuta

Understanding the Data Mesh

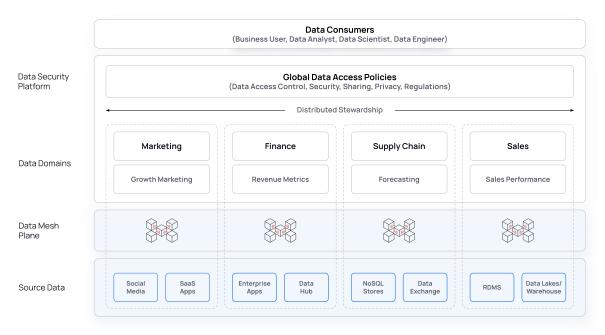
First defined by technology consultant Zhamak Dehghani in 2019, data mesh was developed as an alternative to the monolithic architectures of many legacy data ecosystems. These systems were built with centralized data warehouses, keeping data in as few storage platforms as possible. All data access and use was coordinated through these platforms and the teams-often engineers and developers-who are responsible for them.

As a concept, data mesh is based on the decentralization of data ownership and the enablement of domain experts to independently develop and manage their own data products. Rather than relying on a centralized data warehouse, a **data mesh architecture** separates data into various self-contained domains. Each of these data domains is aligned to a specific business purpose, use case, or project. This model gives ownership of data to the teams most closely aligned with it, enabling them to create context-aware data products, govern their data, increase self-service, and improve collaboration.

"Data mesh is a concept that really forces us to start thinking about data as a product, which is a great concept because when you have a product, you kind of need to assume that you're going to have a client, and if you have a client, you build a product for the consumer. It changes the way you engineer the product. It changes the way you're trying to make it available and changes the way you make it discoverable and accessible."



Yarik Chinkskiy Chief Architect, JPMorgan Chase & Co



Data Mesh Architecture

A typical data mesh architecture resembles the graphic above. The data mesh enables various teams-such as marketing, finance, and sales-to operate in their own self-managed domains. It separates contexts while still remaining connected, so as to avoid siloing of resources or responsibilities. Ultimately, this distributed system permits manageable data storage and control, while keeping it accessible for data consumers across an organization.

What is a Data Mesh?

Learn more about the importance of data mesh and when to consider its application.

READ IT HERE

The Four Pillars of the Data Mesh

Dehghani developed four core principles upon which to build a data mesh architecture:

1. Domain-Centric Ownership & Architecture

The responsibility of domain ownership and management are shifted to teams that work most closely with and possess the most knowledge about each respective data domain. Legacy architectures place the full range of domains under the care of a select few technical stakeholders that lack familiarity or business context, which can complicate approval workflows. The domain-centric approach ensures that the business team owning each specific domain works closely with data consumers to facilitate timely and reliable data access.

2. Data-as-a-Product

Teams should create data products that are well-documented and designed to meet the specific needs of different domain and consumer teams. By making use of data resources through a product lens, teams can better adopt practices centered around ease of use. This includes data discoverability and identification, documentation, and data quality measures that are all carried out by each domain-specific team. Ensuring data is good quality, documented, and discoverable means data users can access, share and leverage it more efficiently and consistently.

3. Self-Service Data Platform

Once quality data is documented and made available to consumers, it requires a framework to serve these users with streamlined access. By implementing consistent, domain-agnostic access and security measures that are low maintenance and easy-to-understand, teams create a structure that is both clearly defined and repeatable across efforts/domains.

4. Federated Computational Governance

With data readily available to so many users, maintaining consistent **data governance** and security measures across domains is an integral part of data mesh security. This requires teams to balance the delegation of domain-based policy management with the enforcement of centralized, consistent security standards across the ecosystem. The combination of centralized and decentralized governance capabilities ensures informed data protection and compliance across domains.

Benefits of Data Mesh Architectures

While the data mesh is still a relatively nascent architectural paradigm, there are a handful of consistent benefits that modern organizations have experienced following their decision to implement it. Some of the most common data mesh benefits include:

1. Enhanced Data Democratization

Data democratization is the process of enabling easy, scalable data access to a larger number of data users, regardless of their role. It is a concept designed to eliminate complicated data frameworks and bottlenecks by enabling a wider range of stakeholders with timely and secure access.

By operationalizing a decentralized architecture, domain-based data mesh models allow users to access the data they need, when they need it. As long as global controls are also in place, data access and use can reliably operate on a domain level. This ease of access allows cross-functional teams to become more data-driven, getting resources into the hands of users who can benefit from their insights. This not only helps reach business goals and objectives, but it also improves the dayto-day functions of both data users and their teams.

2. Optimized Stakeholder Alignment

By shifting ownership of specific domains to specialized teams, the data mesh ensures better alignment between business operations and data resources. It is a model that places the data resources that teams need to access under their control, rather than requiring them to move through complex, time-consuming, and bottlenecked access request and approval processes.

Domain-specific ownership simplifies the access request process by making clear to all data users who is responsible for access decisions on particular data sets. This also helps to pivot away from the monolithic and bogged-down access requests of centralized models, helping facilitate peer-topeer data sharing without sacrificing security.

3. Flexibility at Scale

In a centralized data ecosystem, data access and use can be controlled in one location, and any necessary changes can be made from the ground up. Whether adding new data sources, updating policies, or providing access to additional user groups, each update must be made to the centralized platform(s).

This kind of centralized model can work for a smaller organization with no forecasted growth, as it's likely that updates won't be frequent. It's comparable to having a one-terminal airport for a town or small province, one that can service the needs of its users without requiring massive scale or complexity.

For dynamic, growth-oriented organizations, however, this type of environment is simply not feasible. At a certain point, the rate of change in a rapidly scaling ecosystem will outpace its capabilities, potentially resulting in access breakdowns and increased risk of leak or breach. A one-terminal airport cannot service a bustling, up-and-coming city that is gaining new businesses, residents, and visitors at a steady pace.

This city would need an airport with a number of interconnected terminals that can serve the needs of any and all travelers. Similarly, by decentralizing data into more manageable domains, data mesh lays out a model that can sustain consistent growth. New data, users, and use cases can be added into additional domains that are controlled by relevant teams, and governed by access and security policies.

The Complexity of Data Mesh Security

While the benefits of data mesh are enticing, achieving secure implementation is not without its challenges. The data mesh is still an evolving concept, being tested and iterated with each new attempt at configuration.

The fledgling nature of the data mesh, combined with its distributed composition, makes securing and protecting it a complex process. Data security must be applied in a way that simultaneously protects individual domains and the entire ecosystem, without hindering data accessibility. This is true for any modern data ecosystem, but is especially applicable for one built around ease of access and silo breakdowns. Any security measures that bottleneck access or increase time-to-data negate the desired benefits of the data mesh and leave teams back at square one.

Common Data Mesh Security Challenges

As data mesh implementations become more regular, organizations are regularly being challenged in implementing and securing a specific aspects of the model:

1. Decentralized Ownership & Access Control

While decentralizing data ownership offers a range of benefits, it can be difficult to set up and maintain in practice. In a centralized ecosystem, data ownership and access management responsibilities are traditionally within the realm of either IT or engineering teams. These teams, usually overseen by a Chief Technology Officer (CTO) or Chief Data Officer (CDO), are tasked with facilitating all of the organization's data collection, processing, sharing, and use.

Adding separate domains to this model effectively distributes ownership. While this is done with the intent of streamlining data access and use, it can become difficult to know just who owns what. Systems need to be built to enable cross-domain data discovery, access, and sharing, which can create both a larger attack surface and more potential points of failure. There also must be a guarantee that rules are consistent across data products. Inconsistent or even contradictory rules compromise data security.

2. Data Governance in a Distributed Ecosystem

When it comes to data governance and compliance with regulations, centralized ecosystems are inherently easier to protect. All data is contained within one location, and compliant controls can be set around that platform to determine who can access what, when, and for which purposes. These controls should not only protect data within the ecosystem, but also lessen the risk of breach resulting from any external attack.

When data is separated out into distributed domains, its security requirements become exponentially more complex. Now data lives in a variety of locations, each owned and controlled by a different team. This necessitates additional governance policies for each domain, as well as a way in which to oversee the security and compliance of the entire domain-based framework. Access and governance requirements are federated, and therefore harder to protect in a consistent and comprehensive manner.

3. Privacy in a Self-Service Environment

Organizations today are increasingly collecting sensitive personal data in order to provide consumers with more personalized products, experiences, and services. This can include data such as personally identifiable information (PII), protected health information (PHI), financial information, and more-all of which could easily harm the data subject if it were to be exposed or accessed by an unauthorized party. Because of this, modern data rules and regulations require strict privacy protections be enforced on data ecosystems.

As with data governance, privacy is, by nature, harder to maintain in a distributed data mesh. Controls need to be applied within each respective domain, and kept up-to-date with evolving regulations. There is also less oversight into access in a data mesh, since there is no longer a central team that oversees and determines all access requests. This lack of oversight increases the likelihood of inadequate privacy controls or data misuse across domains.

Regulatory Compliance Considerations

Data mesh architectures are not excluded from the purview of modern compliance rules and regulations. If anything, they are more heavily scrutinized due to their possible security and privacy risks. As this architectural paradigm continues to evolve, so too do the regulatory standards it must meet.

A Guide to Data Compliance Laws and Regulations

Understand which data compliance laws might apply to your organization.

READ IT HERE

Whether it be foundational compliance laws like the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), or evolving state-level regulations like the California Consumer Privacy Act (CCPA), data protection regulations provide necessary standards for any data ecosystem-data mesh or otherwise.

Some important regulatory concepts to consider when implementing a data mesh are:

- Data Sharing Regulations/Agreements: Data sharing is a key enabler of many modern data-driven business objectives. Sharing data from one domain to another requires an internal data sharing protocol, but organizations also need a standard in place for sharing with external stakeholders, like contractors. Whether it be internal or external, sharing data is subject to many protective regulations or individual data sharing agreements. Domain-based sharing practices should be created and carried out with compliance in mind, to ensure that exchanging data between users is not done at the risk of privacy or security.
- Data Localization and Sovereignty Laws: By helping teams to connect distributed domains under one architecture, data mesh is a feasible model for organizations who are operating on an increasingly global scale. Even so, the current growth of data localization and sovereignty laws has an important impact on this cross-border potential. Data localization refers to a mandatory legal or administrative requirement directly or indirectly stipulating that data be stored or processed, exclusively or non-exclusively, within a specified jurisdiction. Data sovereignty refers to the rights of countries to govern and control data that has been generated within their borders, regardless of where it is being stored or accessed. These laws should be considered when creating any kind of data mesh

architecture that connects international domains. Just because data can be shared across borders does not mean that it is always legally permissible.

• Inter- or Intra-Organization Standards: Legislative regulations are not the only standards with which data mesh users need to comply. Many organizations maintain a diverse set of inter- and intra-business rules, contracts, and best practices that have a regulatory impact on data use. These can include service-level agreements (SLAs), binding corporate rules (BCRs), industry standards, and other types of legal agreements. Any organization that uses data should maintain an understanding of which of these standards apply to them, which provides a baseline when building out a data mesh framework.

3 Steps for Implementing a Secure Data Mesh

Achieving a secure data mesh architecture is not necessarily a quick or easy task. Rather, it is a gradual process that, if carried out with intention, can result in an efficient, secure distributed data ecosystem. To help teams facilitate a successful implementation, we've assessed various approaches and distilled three key steps for securing any modern data mesh application.

Data Mesh Security Best Practices

Learn additional best practices for implementing a data mesh architecture.

READ IT HERE

Step 1: Maintain Consistent Metadata

When securing any kind of data ecosystem, it is essential that users and administrators understand the resources at their disposal. This is especially true for a distributed ecosystem like the data mesh, a structure not just composed of various platforms and domains, but also one that contains vastly different types of data and users.

All of these ecosystem parts-from data sets to data users-can be identified and understood using metadata. Metadata refers to the "what," providing contextual information about resources or users that is integral to the system's operation. It can take the form of tags or descriptions, and often includes information such as file size, author, date created, date modified, and more. Metadata is crucial to efficient access management, analytics, monitoring, and compliance-all of which are important parts of the data mesh.

In line with he data mesh's decentralized ideology, it follows that domain teams will have the most intimate, timely understanding of their data's metadata-who is creating, editing, owning, and modifying data in their domain. Even so, metadata is ineffective unless it can be consistently attributed and understood across domains.

This is why the first step in securing any data mesh architecture should be creating and applying a consistent metadata identification and tagging schema. Using **sensitive data discovery (SDD)**, teams can assess their data and ensure that it is discovered, tagged, and classified appropriately. This can be implemented in a transformative manner to data that already exists within the data mesh, as well as during the ingestion of new data sources.

Sensitive data discovery enables:

- **Cross-Domain Tagging:** By setting and managing a collection of specific tags to help delineate data, teams can classify data in a standard manner across their domains. This is critical to fostering an organizational awareness of data assets, as well as consistently applying data access and security policies.
- Data Classification: Data classification is the identification of the types, levels of sensitivity, and criticality of an organization's data. By discovering and tagging data based on its attributes, teams can create a standard taxonomy that reinforces data quality and security throughout the ecosystem.
- **Consistent Policy Application:** Once data is tagged and classified according to discovered metadata, an organization has the foundations upon which to build robust data access and security policies.

With data discovered and tagged according to a consistent schema, data teams gain a holistic view of the resources across their distributed data mesh. Regardless of domain, the data can be classified and tagged in a manner that enhances cross-functional comprehension and interdomain collaboration.

"So we're beginning to see a lot of change that these leaders are bringing into play. And they are leveraging all these different kinds of technologies, whether it's access control technologies, whether it's governance, security layer, or even all the capabilities in the cloud. They're bringing it all together to create that data mesh architecture."



Archana Venkatraman

Research Director, Cloud Data Management & CloudOps, IDC Europe

Step 2: Employ Global & Local Policy Management

In a centralized database, **data access control** and governance are straightforward: everything is stored in and accessed from a single, controlled location. It's like going to a library; all books are stored in the building, and any access needs to be routed through the librarians.

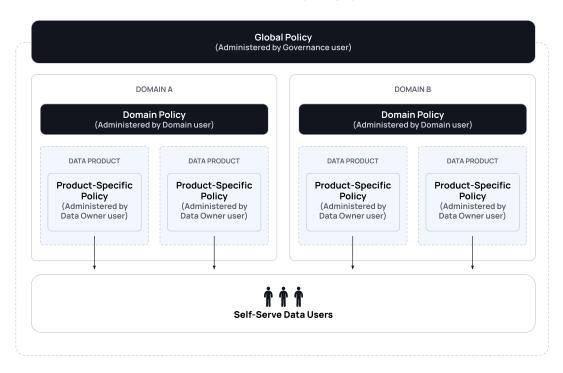
In the distributed domains of a data mesh architecture, data governance and access control are not as simple. Policies can be applied vertically (locally) within specific domains, but relying solely on domain-based policies limits consistency across the data ecosystem and requires intense manual effort to maintain. On the other hand, applying policies horizontally (globally) across domains in an attempt to foster consistency neglects the unique requirements of each domain's purpose(s), users, and specific data resources.

This is why employing a balance of both horizontal and vertical policies is crucial to data mesh security. Global policies should be authored and applied in line with the ecosystem's most generic and all-encompassing principles, regardless of the data's domain. Localized domain-level policies should be fine-grained and applicable to only domain-specific purposes or use cases.

Global	Local	
mask all PII data	redact all rows that contain data tagged 'credit_card_number'	
anonymize all PHI data	hash all columns marked 'home_address'	
encrypt all financial data sets	null all values tagged 'social_security_number'	

Examples of Global vs. Local Policies

Understanding the need to balance horizontal and vertical policy management is one thing. Authoring, applying, and maintaining these policies at scale, however, is an entirely new challenge. Policies-whether global or local-must be enforced when and where they're needed, in a manner that does not impede time-to-data for domain users.



Data Mesh Policy Application

A federated governance framework helps create, apply, and maintain policies at the global and domain/local level. The responsibility of domain-level policy management is delegated to the teams that own the data, while the responsibility of global policies remains with security and governance teams. All security and access control policies should be dynamic, using attribute-based access controls (ABAC) that are capable of scaling with added domains and users. These can take the form of subscription or data policies.

- 1. **Subscription Policies:** These determine who can request access to a data source or group of data sources based on whether the user is subscribed to the source. They can be helpful for local policy enforcement, as users can subscribe to the data within their specific domains.
- 2. **Data Policies:** These are policies that are applied directly to the data source, determining what data the user can see regardless of their subscriptions. They can be helpful for global policy enforcement, ensuring that certain data is protected/restricted no matter where it is accessed. Examples include masking, redaction, and differential privacy.

In order to keep track of policy enforcement, teams should continuously monitor activity across domains. Maintaining rigorous monitoring gives teams complete oversight of global policy application, as well as local policy enforcement within specific domains. Threat detection capabilities alert domain owners of anomalous data use, and provide ecosystem-wide surveillance for any data misuse or risky behavior. This gives teams the capacity to respond to and manage incidents in real-time, and effectively manage global and local data security.

"I came from software engineering and DevOps and microservices, there was this great 'solve all your problems and create new ones' kind of vibe. I think data mesh is exactly the same way. You need to be comfortable with automation. You need to have tools like Immuta in place. You need to have a culture of automation and understanding how all the bits hold together and how you're going to govern that."



Kurt Gardiner

Senior Engineering Manager, nib Group

Step 3: Foster Organizational Change

Thus far, we've focused on the technical requirements of a secure data mesh implementation. However, it would be a mistake to assume that data mesh implementation is a solely tech-based process.

Ultimately, the data mesh is just as much an organizational framework as it is a data architecture. In closing her original article introducing the data mesh, Zhamak Dehghani noted:

"The needs are real and tools are ready. It is up to the engineers and leaders in organizations to realize that the existing paradigm of big data and one true big data platform or data lake, is only going to repeat the failures of the past, just using new cloud based tools."

The technology is available, but actual implementation requires the collaboration, alignment, and commitment of all stakeholders in order to take root.

It's important to note that each organization will approach the data mesh from a different starting point. Factors like data maturity level, organizational structure, and team engagement can vary significantly across businesses and industries. By identifying champions to lead the charge and initial use cases to build around, teams can jumpstart their data mesh with a clear path to implementation.

With this in mind, we've delineated the following categories as a blueprint to help guide secure data mesh implementations, regardless of where your organization's starting point.

	Purpose	Responsibilities	Stakeholders
Cross-Domain Governance Board	Establish and oversee data governance practices that span multiple domains.	Develop and communicate global governance policies, monitor adherence, and facilitate stakeholder enablement and collaboration.	Chief Data Officer (CDO)/ Chief Information Officer (CIO), Data Governance Manager, Domain Owners, Compliance Officer(s), Data Privacy Specialist(s), Data Stewardship Lead(s)
Domain Governance Team	Own data governance within a specific domain or business area.	Define and enforce data ownership roles and policies within the domain, monitor and ensure compliance with policies, and generally oversee domain-specific data quality, ownership, and usage.	Domain Owner, Domain Data Steward, Domain Data Architect, Data Product Owners, Domain Compliance Coordinator
Data Product Team	Design, develop, and maintain data products, which are self- contained units of data with well-defined use cases.	Identify domain-specific data needs, design data products, ensure products meet governance standards, implement pipelines and processing logic, and monitor data product usage and performance.	Data Product Manager, Data Engineer, Data Scientist/ Analyst/Bl Developer
Self-Serve Data Platform Team	Build and maintain the technical infrastructure and tools that enable domain teams to manage their data independently and efficiently.	Develop and maintain a user- friendly platform for domain teams to create, manage, and monitor data pipelines. Provide tools, interfaces, and enablement for data discovery, cataloging, and access control.	Platform Architect, Platform Engineer, Platform Operations Specialist, Platform Enablement Trainer

By aligning technical, business, security, and compliance stakeholders, and involving them in the implementation process, teams of any size and maturity can create an effective data mesh framework. Once this framework is built, dedicating resources to educating teams on ownership and accountability over their new domains will foster a data mesh that is accessible, secure, and effective across an organization.

Roche's Secure Data Mesh Implementation

To understand the necessity of stakeholder involvement, consider the example of Roche Diagnostics.

Roche's Federated Governance and Access Controls for Data Mesh

Hear directly from the Roche team about their secure data mesh implementation.

WATCH HERE

When reexamining how their data ecosystem was structured, the Roche team recognized the benefits the data mesh could provide for their objectives. There were a number of different teams involved in their original centralized data ecosystem, all of which were part of very manual data access and use processes that were unable to scale with business needs.

"For years, we tried to serve a decentralized business organization through a central IT," said Immuta Senior Product Manager Claude Zwicker (formerly of Accenture). "And we just realized it doesn't work."

Achieving a secure and scalable data mesh implementation needed to be a group effort, not just an IT change. With this mindset, Head of Data Management Platforms Paul Rankin noted "we started to speak to the business and ask them if they wanted to work as one partnership, one team to really move this forward at an organizational level. And this was key, really, to get that buy-in from the business organization."

By putting the business first and coordinating user enablement, Rankin and team were able to achieve stakeholder buy-in. This inter-team alignment quickly snowballed, becoming the driving force behind the success of their data mesh implementation.

Transitioning to an effective and secure data mesh required the alignment of technical teams that understood the structural aspects of the transformation, and business teams that were meant to actually use the domain-based framework. By involving these different stakeholders, and dedicating resources to educating them on self-service domain management and use, the Roche team achieved a secure data mesh that currently unites over 200 data products and counting.

Conclusion

To achieve the steps for a secure data mesh in a streamlined and unified manner, teams should consider implementing a dynamic data security platform like Immuta. With tools that Discover, Secure, and Detect sensitive data in a data ecosystem, Immuta enables streamlined, fine-grained data security and governance across a distributed data mesh.

Discover: Sensitive data discovery enables teams to automatically discover data and apply 60+ prebuilt classifiers, alongside domain-specific and custom classifiers. This helps to create and maintain a holistic classification and tagging schema, and keep metadata consistent across domains for an improved understanding of resources and application of policies.

Secure: This allows teams to build data policies in plain language or as-code, which are then enforced automatically in real time across teams, regions, and domains. ABAC policies deliver scalable data access without the challenges of role explosion, and enable the creation and automated enforcement of both global (horizontal) and local (vertical) data policies.

Detect: Continuous monitoring and detection capabilities help teams leverage timely insights into data access and user activity across domains, with anomaly indicators for faster analysis and risk remediation and proactive actions. Teams can receive notifications, alerts, and audit reports to prove compliance easily.

To learn more about the Immuta Data Security Platform's data mesh enabling capabilities, you can check our demo Powering Your Data Mesh with Snowflake and Immuta. If you'd like to connect to discuss the data mesh with an Immuta expert, request a demo today.

About Immuta

Immuta enables organizations to unlock value from their cloud data by protecting it and providing secure access. The Immuta Data Security Platform provides sensitive data discovery, security and access control, data activity monitoring, and has deep integrations with the leading cloud data platforms. Immuta is now trusted by Fortune 500 companies and government agencies around the world to secure their data. Founded in 2015, Immuta is headquartered in Boston, MA.

ΙΜΜυτα