eBook

A New Approach to Data Sharing

Open data sharing and collaboration for data, analytics and AI

Third Edition

databricks

Contents

Contents

Introduction
Data and AI Sharing in Today's Digital Economy

Today's economy revolves around data. Every day, more and more organizations must exchange data with their customers, suppliers and partners. Security is critical. And yet, efficiency and immediate accessibility are equally important. Where data sharing may have been considered optional, it's now required. More organizations are investing in streamlining internal and external data sharing across the value chain. But they still face major roadblocks — from human inhibition to legacy solutions to vendor lock-in. Gartner recently found that chief data officers who've successfully executed data sharing initiatives are 1.7x more effective in showing business value and return on investment from their data analytics strategy. To compete in the digital economy, organizations need an open — and secure — approach to data sharing.

In recent years, the emergence of AI has added new dimensions to data sharing. AI models thrive on large volumes of diverse data, making it essential for organizations to share not only structured datasets but also unstructured data (such as images, videos and text) and AI models themselves. The ability to share AI models and unstructured data efficiently is becoming a key differentiator for companies aiming to unlock advanced AI-driven use cases.

This eBook takes a deep dive into the modern era of data sharing and collaboration, from common use cases and key benefits to conventional approaches and their challenges. You'll get an overview of our open approach to data sharing and find out how Databricks allows you to share your data across platforms, to share all your data and AI, and to share all your data securely with unified governance in a privacy-safe way.

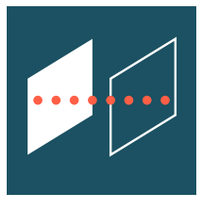
Chapter 1
What Is Data Sharing and Why Is It Important?

Data sharing is the ability to make the same data available to one or many stakeholders — both external and internal. Nowadays, the ever-growing amount of data has become a strategic asset for any company. Data sharing — within your organization or externally — is an enabling technology for data enrichment, enhanced analysis and/or monetization. Sharing data as well as consuming data from external sources allows companies to collaborate with partners, establish new partnerships and generate new revenue streams with data monetization. Data sharing can deliver benefits to business groups across the enterprise. For those business groups, data sharing can enable access to data needed to make critical decisions.

Common data and AI sharing use cases

Internal sharing across BUs

Within any company, different departments, lines of business and subsidiaries seek to share data so that everyone can make decisions based on a complete view of the current business reality. For example, finance and HR departments need to share data as they analyze the true costs of each employee. Marketing and sales teams need a common view of data as they seek to determine the effectiveness of recent marketing campaigns. And different subsidiaries of the same company need a unified view of the health of the business. Removing data silos — which are often established for the important purpose of preventing unauthorized access to data — is critical for digital transformation initiatives and maximizing the business value of data.

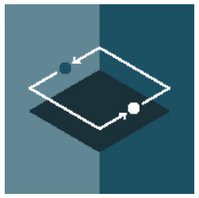


Peer-to-peer sharing

Many companies now strive to share data with partners and suppliers similarly to how they share it across their own organizations. For example, retailers and their suppliers continue to work more closely together as they seek to keep their products moving in an era of ever-changing consumer tastes. Retailers can keep suppliers posted by sharing sales data by SKU in real time, while suppliers can share real-time inventory data with retailers so they know what to expect. Scientific research organizations can make their data available to pharmaceutical companies engaged in drug discovery. Public safety agencies can provide real-time public data feeds of environmental data, such as climate change statistics or updates on potential volcanic eruptions.

Third-party data licensing

Across industries, companies are commercializing data, and this segment continues to grow. Large multinational organizations have formed exclusively to monetize data, while other organizations are looking for ways to monetize their data and generate additional revenue streams. Examples of these companies can range from a capital markets data provider such as S&P to a marketing data hygiene and enrichment company such as Epsilon to a telecommunications company with proprietary 5G data to retailers that have a unique ability to combine online and offline data. Data vendors are growing in importance as companies realize they need external data for better decision-making.



SaaS application sharing

Companies increasingly rely on various cloud-based services for different aspects of their operations. As a result, data becomes isolated within individual SaaS applications, making it difficult to gain a holistic view of business operations. SaaS application sharing addresses the growing need for businesses to integrate and analyze data from multiple SaaS platforms. This approach allows organizations to expand their data ecosystem by bringing information from various SaaS applications, enabling a more comprehensive and unified data strategy. For example, AVEVA has partnered with Databricks to allow customers to seamlessly and securely share reliable, high-quality industrial data across regions and platforms.

Key benefits of data and AI sharing

As you can see from the use cases described, there are many benefits of data sharing, including:

Greater collaboration with existing partners. In today's hyper-connected digital economy, no single organization can advance their business objectives without partnerships. Data sharing helps solidify existing partnerships and can help organizations establish new ones.

Ability to generate new revenue streams. With data sharing, organizations can generate new revenue streams by offering data products or data services to their end consumers.

Ease of producing new products, services or business models. Product teams can leverage both first-party data and third-party data to refine their products and services and expand their product/service catalog.

Greater efficiency of internal operations. Teams across the organization can meet their business goals far more quickly when they don't have to spend time figuring out how to free data from silos. When teams have access to live data, there's no lag time between the need for data and the connection with the appropriate data source.

Chapter 2
Conventional Methods of Data and AI Sharing and Their Challenges

Sharing data across different platforms, companies and clouds is no easy task. In the past, organizations have hesitated to share data more freely because of the perceived lack of secure technology, competitive concerns and the cost of implementing data sharing solutions.

Even for companies that have the budget to implement data sharing technology, many of the current approaches can't keep up with today's requirements for open format, multicloud, high-performance solutions. Most data sharing solutions are tied to a single vendor, which creates friction for data providers and data consumers who use noncompatible platforms. With the rise of generative AI, data is no longer limited to structured data. There's a proliferation of unstructured data (audio, video, images, PDF, etc.) and a need for AI models.

Over the past 30 years, data sharing solutions have come in three forms: legacy and homegrown solutions, cloud object storage and closed source commercial solutions. Each of these approaches comes with its pros and cons.

Legacy and homegrown solutions

Many companies have built homegrown data sharing solutions based on legacy technologies such as email, (S)FTP or APIs.

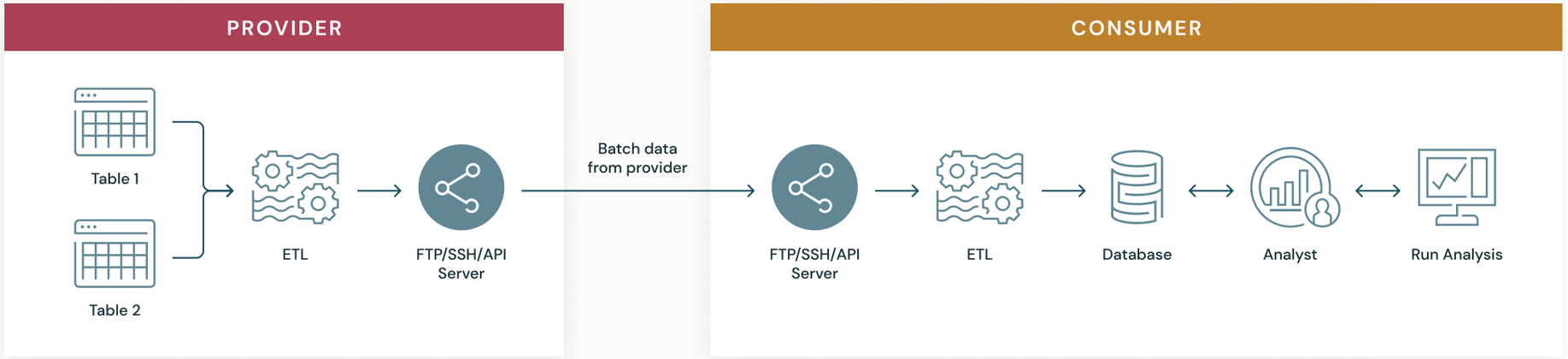


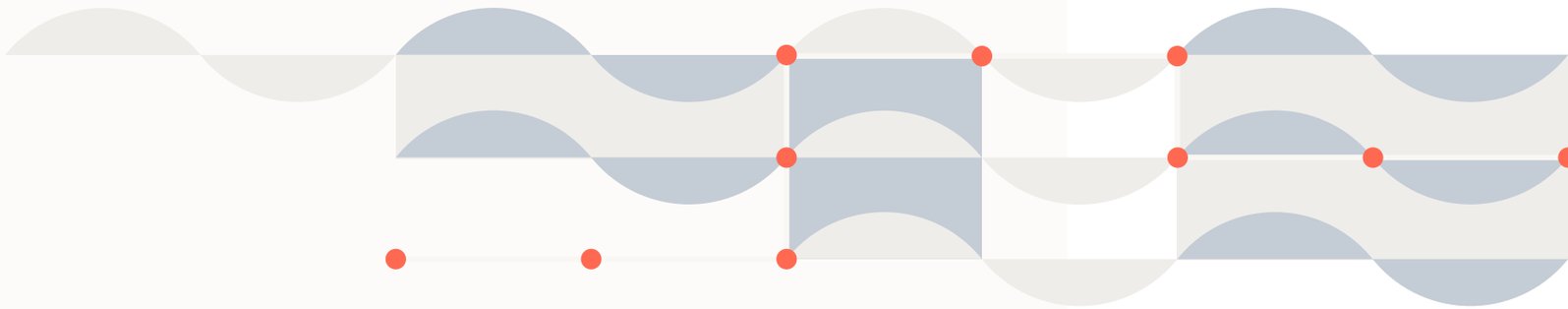
Figure 1: Legacy data sharing solutions

Pros

- **Vendor agnostic:** FTP, email and APIs are all well-documented protocols. Data consumers can leverage a suite of clients to access data provided to them.
- **Flexibility:** Many homegrown solutions are built on open source technologies and will work both on-premises and on clouds

Cons

- **Data movement:** It takes significant effort to extract data from cloud storage, transform it and host it on an FTP server for different recipients. Additionally, this approach results in data providers copying data to multiple platforms and dozens of regions manually. Data copying causes duplication and prevents organizations from instantly accessing live data.
- **Complexity of sharing data:** Homegrown solutions are typically built on complex architectures due to replication and provisioning. This can add considerable time to data sharing activities and result in out-of-date data for end consumers.
- **Operational overhead for data recipients:** Data recipients have to extract, transform and load (ETL) the shared data for their end use cases, which further delays the time to insights. For any new data updates from the providers, the consumers have to rerun ETL pipelines again and again.
- **Security and governance:** As modern data requirements become more stringent, homegrown and legacy technologies have become more difficult to secure and govern
- **Scalability:** Such solutions are costly to manage and maintain and don't scale to accommodate large datasets



Proprietary vendor solutions

Commercial data sharing solutions are a popular option among companies that don't want to devote the time and resources to building an in-house solution yet also want more control than what cloud object storage can offer.

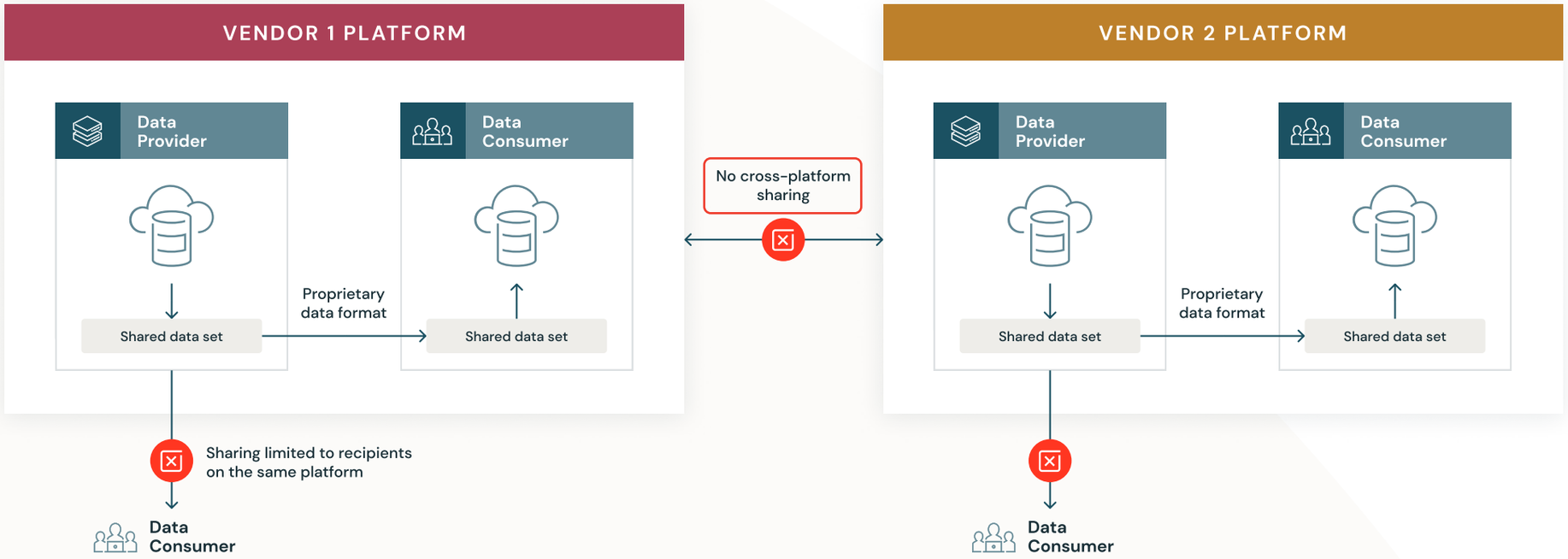


Figure 2: Proprietary vendor solutions

Pros

- **Simplicity:** Commercial solutions allow users to share data easily with anyone else who uses the same platform

Cons

- **Vendor lock-in:** Commercial solutions don't interoperate well with other platforms. While data sharing is easy among fellow customers, it's usually impossible with those who use competing solutions. This reduces the reach of data, resulting in vendor lock-in. Furthermore, platform differences between data providers and recipients introduce data sharing complexities.
- **Data movement:** Data must be loaded onto the platform, requiring additional ETL and data copies
- **Scalability:** Commercial data sharing comes with scaling limits from the vendors
- **Cost:** All of these challenges create additional cost for sharing data with potential consumers, as data providers have to replicate data for different recipients on different cloud platforms

Cloud object storage

Object storage is considered a good fit for the cloud because it's elastic and can more easily scale into multiple petabytes to support unlimited data growth. The big three cloud providers all offer object storage services (AWS S3, Azure Blob Storage, Google Cloud Storage) that are cheap, scalable and extremely reliable.

An interesting feature of cloud object storage is the ability to generate signed URLs, which grant time-limited permission to download objects. Anyone who receives the presigned URL can then access the specified objects, making this a convenient way to share data.

Pros

- **Sharing data in place:** Object storage can be shared in place, allowing consumers to access the latest available data
- **Scalability:** Cloud object storage profits from availability and durability guarantees that typically can't be achieved on-premises. Data consumers retrieve data directly from the cloud providers, saving bandwidth for the providers.

Cons

- **Limited to a single cloud provider:** Recipients have to be on the same cloud to access the objects
- **Cumbersome security and governance:** Assigning permissions and managing access is complex. Custom application logic is needed to generate signed URLs.
- **Complexity:** Personas managing data sharing (DBAs, analysts) find it difficult to understand identity and access management (IAM) policies and how data is mapped to underlying files. For companies with large volumes of data, sharing via cloud storage is time-consuming, cumbersome and nearly impossible to scale.
- **Operational overhead for data recipients:** The data recipients have to run extract, transform and load (ETL) pipelines on the raw files before consuming them for their end use cases

New challenges: AI model sharing and unstructured data sharing

As AI continues to evolve and shape the future of industries, organizations face additional challenges beyond traditional structured or tabular datasets. Today's enterprises must share not only structured datasets but also unstructured ones — such as images, videos, documents — and AI models themselves (e.g., machine learning models or notebooks).

1. Unstructured data sharing

- Sharing unstructured datasets (e.g., text documents or multimedia files) presents unique challenges because these formats are often larger in size or lack standardized schemas compared with structured datasets like databases or spreadsheets
- The complexity increases when unstructured volumes need real-time collaboration across different platforms or clouds while maintaining security standards

2. AI model sharing

- The inability to easily share AI models (e.g., trained machine learning models), notebooks or other AI artifacts across organizations limits innovation
- Without effective mechanisms for cross-platform AI model exchange — whether due to technical incompatibilities between frameworks or security concerns — organizations struggle to unlock the full potential of their shared datasets

Both unstructured dataset sharing *and* AI model sharing represent significant hurdles that prevent organizations from fully realizing advanced AI-driven use cases.

The lack of a comprehensive solution makes it challenging for data providers and consumers to easily share data and AI assets. Cumbersome and incomplete data sharing processes also constrain the development of business opportunities from shared data.

Chapter 3
Delta Sharing: An Open Standard for Secure Sharing of Data and AI Assets

We believe the future of data and AI sharing should be characterized by open technology. Data and AI sharing shouldn't be tied to a proprietary technology that introduces unnecessary limitations and financial burdens to the process. It should be readily available to anyone who wants to share data at scale. This philosophy inspired us to develop and release a new protocol for sharing data: Delta Sharing.

What is Delta Sharing?

Delta Sharing provides an open protocol to securely share live data from your lakehouse to any computing platform. Recipients don't have to be on the Databricks Platform or on the same cloud or on a cloud at all. Data providers can share live data without replicating it or moving it to another system. Recipients benefit from always having access to the latest version of data and can quickly query shared data using tools of their choice for BI, analytics and machine learning, reducing time to value.

Data providers can centrally manage, govern, audit and track usage of the shared data on one platform. Delta Sharing is natively integrated with Unity Catalog, enabling organizations to centrally manage and audit shared data across organizations and confidently share data assets while meeting security and compliance needs. Delta Sharing protocol also powers Databricks Marketplace, an open marketplace for exchanging data and AI products, and Databricks Clean Rooms, a secure and privacy-protecting environment where multiple parties can work together on sensitive enterprise data.

With Delta Sharing, organizations can easily share existing large-scale datasets based on the open source formats Apache Parquet, Apache Iceberg™ and Delta Lake without moving data. Teams gain the flexibility to query, visualize, transform, ingest or enrich shared data with their tools of choice.

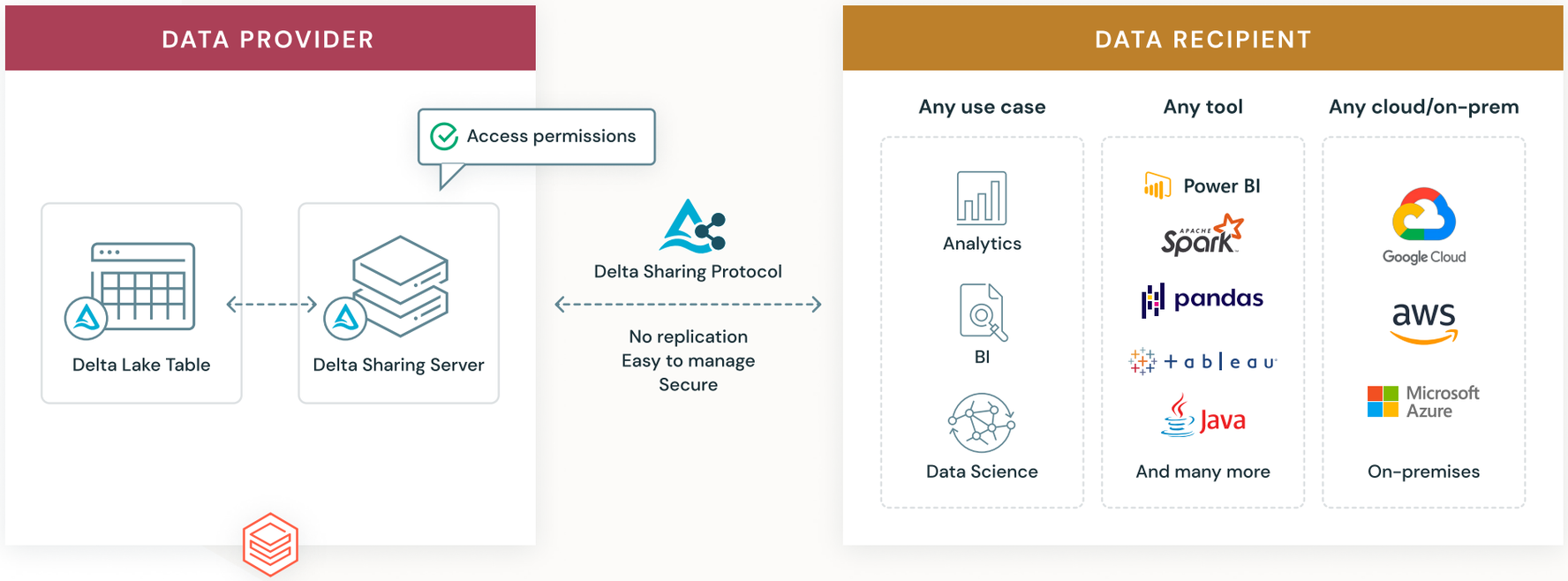


Figure 3:
Delta Sharing

Databricks designed Delta Sharing with five goals in mind:

- Provide an open cross-platform sharing solution
- Share live data without copying it to another system
- Support a wide range of clients such as Power BI, Tableau, Apache Spark™, pandas and Java, and provide flexibility to consume data using the tools of choice for BI, machine learning and AI use cases
- Provide strong security, auditing and governance
- Scale to massive structured datasets and also allow sharing of unstructured data, ML models, dashboards and notebooks, in addition to tabular data

Key benefits of Delta Sharing

By eliminating the obstacles and shortcomings associated with typical data sharing approaches, Delta Sharing delivers several key benefits.

Open cross-platform sharing. Delta Sharing establishes a new open standard for secure data and AI sharing and supports open source Delta and Apache Parquet formats. Delta Sharing supports cross-cloud and cross-platform sharing. Data recipients don't have to be on the Databricks Platform or on the same cloud, as Delta Sharing works across clouds and even from cloud to on-premises setups. To give customers even greater flexibility, Databricks has also released open source connectors for pandas, Apache Spark, Elixir and Python, and is working with partners on many more.

Securely share live data without replication. Most enterprise data today is stored in cloud data lakes. Any of these existing datasets on the provider's data lake can easily be shared without any data replication or physical movement of data. Data providers can update their datasets reliably in real time and provide a fresh and consistent view of their data to recipients.

Centralized governance. With Databricks Delta Sharing, data providers can grant, track, audit and revoke access to shared datasets from a single point of enforcement to meet compliance and other regulatory requirements. Databricks Delta Sharing users get:

- Implementation of Delta Sharing as part of Unity Catalog, the governance offering for the Databricks Data Intelligence Platform
- Simple, more secure setup and management of shares
- The ability to create and manage recipients and data shares
- Audit logging captured automatically as part of Unity Catalog
- Direct integration with the rest of the Databricks ecosystem
- No separate compute for providing and managing shares
- Sharing for Lakehouse Federation, which allows users to share data from existing data warehouses or databases without expensive ETL and without the need to copy it to Databricks

Share data products, including AI models, unstructured data, dashboards and notebooks, with greater flexibility. Data providers can choose between sharing an entire table or sharing only a version or specific partitions of a table. However, sharing just tabular data isn't enough to meet today's consumer demands. Databricks Delta Sharing also supports sharing of nontabular data and data derivatives such as data streams, AI models, SQL views, volumes and arbitrary files, enabling increased collaboration and innovation. Volume sharing allows providers to securely share large-scale unstructured data such as images, videos and logs stored in cloud volumes without replication. Data providers can build, package and distribute data products, including datasets, volumes, AI models and notebooks, allowing data recipients to get insights faster. Furthermore, this approach promotes and empowers the exchange of knowledge — not just data — between different organizations. With Databricks Delta Sharing, we're able to achieve both a truly open marketplace and a truly open ecosystem. In contrast, commercial products are mostly limited to sharing raw tabular data and can't be used to share these higher-valued data derivatives.

Share data at a lower cost. Delta Sharing lowers the cost of managing and consuming shares for both data providers and recipients. Providers can share data from their cloud object store without replicating, thereby reducing the cost of storage. Additionally, by integrating with Cloudflare's Bandwidth Alliance and R2 zero-egress fees object storage, Delta Sharing further minimizes costs by eliminating or significantly reducing egress fees — charges incurred when transferring data out of a cloud provider's network.

In contrast, existing data sharing platforms require data providers to first move their data into their platform or store data in proprietary formats in their managed storage, which often costs more and results in data duplication. With Delta Sharing, data providers don't need to set up separate computing environments to share data. Consumers can access shared data directly using their tools of choice without setting up specific consumption ecosystems, thereby reducing costs.

Reduced time to value. Delta Sharing eliminates the need to set up a new ingestion process to consume data. Data recipients can directly access the fresh data and query it using tools of their choice. Recipients can also enrich data with datasets from popular data providers. The Delta Sharing ecosystem of open source and commercial partners is growing every day.

Maximizing the value of data and AI with Delta Sharing

Delta Sharing is already transforming data and AI sharing activities for companies in a wide range of industries. Given the sheer variety of data available and the technologies that are emerging, it's hard to anticipate all the possible use cases Delta Sharing can address. The Delta Sharing approach is to share any data anytime with anyone easily and securely. In this section we'll explore the building blocks of such an approach and the use cases emerging from these.

"We're excited to continue working with Databricks to enhance our data distribution capabilities through their Delta Sharing platform, which provides an open ecosystem that makes a number of S&P Global content sets more seamlessly accessible and available to our clients. This expands our collaboration with Databricks, which began with S&P Global Capital IQ Workbench, leveraging their technology to create a collaborative analytics notebook environment for our users."

— **David Coluccio**, Head of Distribution Solutions, S&P Global Market Intelligence

"Most data platforms offer closed sharing solutions that restrict our ability to reach all of our customers. We prefer to invest in open solutions which enable us to share data with all of our customers and partners, not only across clouds, but also across platforms."

— **Derek Slager**, CTO and Co-founder, Amperity

"AI21 Labs is pleased that Jamba 1.5 Mini is now on Databricks Marketplace. With Delta Sharing, enterprises can access our Mamba-Transformer architecture, featuring a 256K context window, ensuring exceptional speed and quality for transformative AI solutions."

— **Pankaj Dugar,** SVP and GM, AI21 Labs

"We use reinforcement learning (RL) models in some of our products. Compared to supervised learning models, RL models have longer training times and many sources of randomness in the training process. These RL models need to be deployed in three workspaces in separate AWS regions. With model sharing we can have one RL model available in multiple workspaces without having to retrain it again or without any cumbersome manual steps to move the model."

— **Mihir Mavalankar,** Machine Learning Engineer, Ripple

"Delta Sharing makes it easy to securely share data with business units and subsidiaries without copying or replication. It enables us to share data without the recipient having an identity in our workspace."

— **Robert Hamlet**, Lead Data Engineer, Cox Automotive

"When we want to launch and grow a product with a partner, such as a point-of-sale consumer loan, the owner of the data needs to send massive datasets on tens of thousands of customers. Before, in the traditional data warehouse approach, this would typically take one to two months to ingest new data sources, as the schema of the sent data would need to be changed in order for our systems to read it. But now we point Databricks at it and it's just two days to value. We used Delta Sharing and we had tables of data showing up in our Databricks workspace in under 10 minutes."

— **Barb MacLean**, SVP and Head of Technology Operations and Implementation, Coastal Community Bank

Internal sharing across business units with Delta Sharing

Internal data sharing is becoming an increasingly important consideration for any modern organization, particularly where data describing the same concepts have been produced in different ways and in different data silos across the organization. So it's important to design systems and platforms that allow governed and intentional federation of data and processes, and at the same time allow easy and seamless integration of said data and processes.

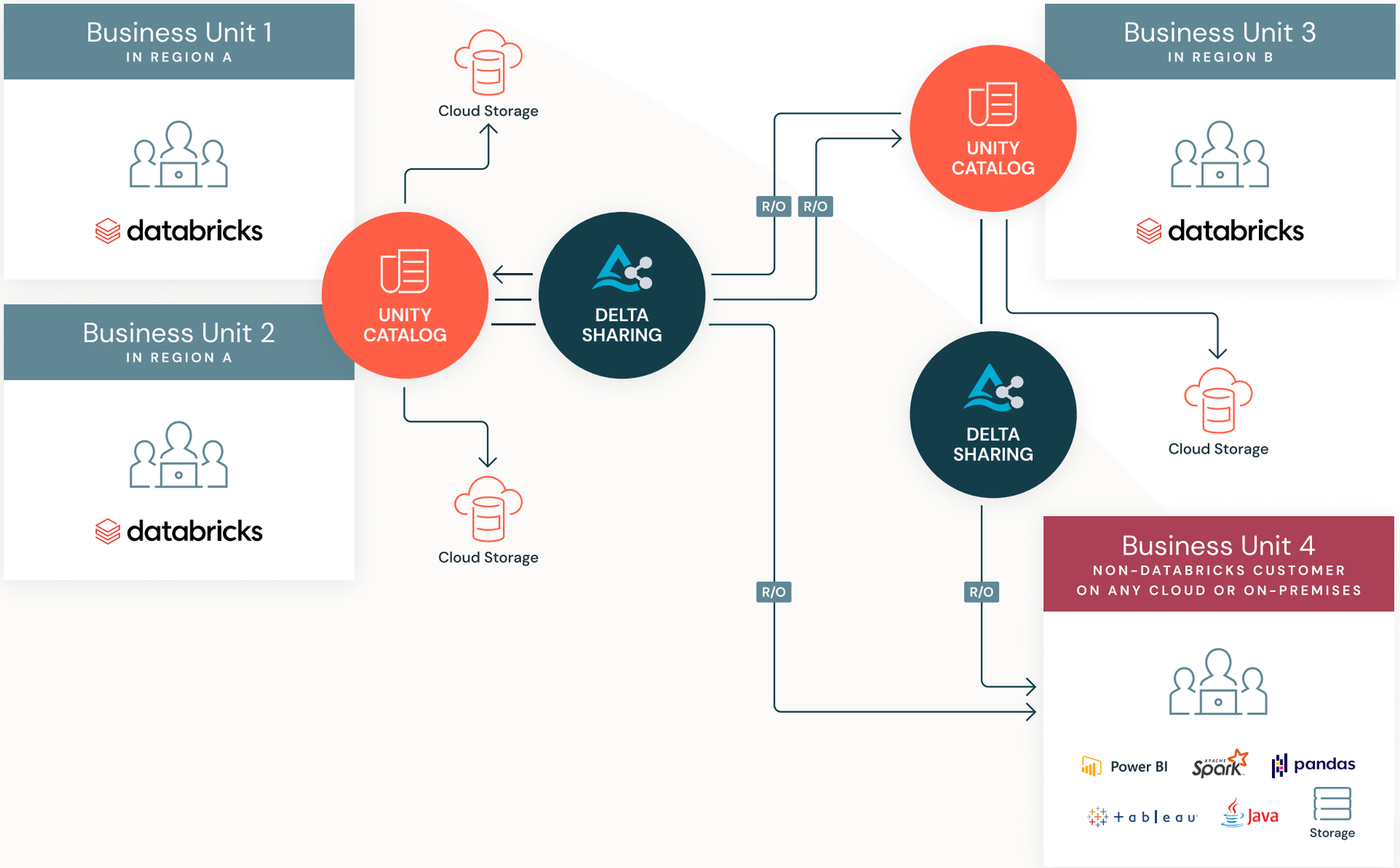


Figure 4: Internal sharing across BUs

To make matters even more complicated, organizations can grow through mergers and acquisitions. In such cases we can't assume that the organizations being acquired have followed the same set of rules and standards to define their platforms and to produce their data. Furthermore, we can't even assume that they've used the same cloud providers, nor can we assume the complexity of their data models.

Delta Sharing can simplify and accelerate the efforts of unification and assimilation of newly acquired organizations and their data and processes. Only selected data sources can be exchanged between the different clouds, platforms and regions. This enables teams to move freely between the organizations that are merging without losing their data — if anything, they're empowered to drive insights of higher quality by combining the data of both.

Databricks Marketplace further enhances internal sharing by offering a private exchange capability that enables organizations to securely share data and AI products with specific business units or partners. Unlike public marketplace listings, private exchanges allow providers to control who can discover and access their data and AI products. This capability is built on top of Delta Sharing, ensuring that shared data remains secure without requiring replication.

Unity Catalog plays a pivotal role in streamlining and governing internal data sharing across business units (BUs) within an organization. Unity Catalog provides centralized governance, auditing and access control for all data assets, ensuring that sharing is secure and compliant with organizational policies. It integrates seamlessly with Delta Sharing, which is the open protocol developed by Databricks for secure data sharing across different platforms.

This combination of Unity Catalog with Delta Sharing and Databricks Marketplace private exchanges provides a powerful framework for internal collaboration across BUs while maintaining robust security controls. Teams can move freely between different parts of the organization or newly acquired entities without losing access to critical data — enabling faster insights and better decision-making.

Peer-to-peer sharing with Delta Sharing

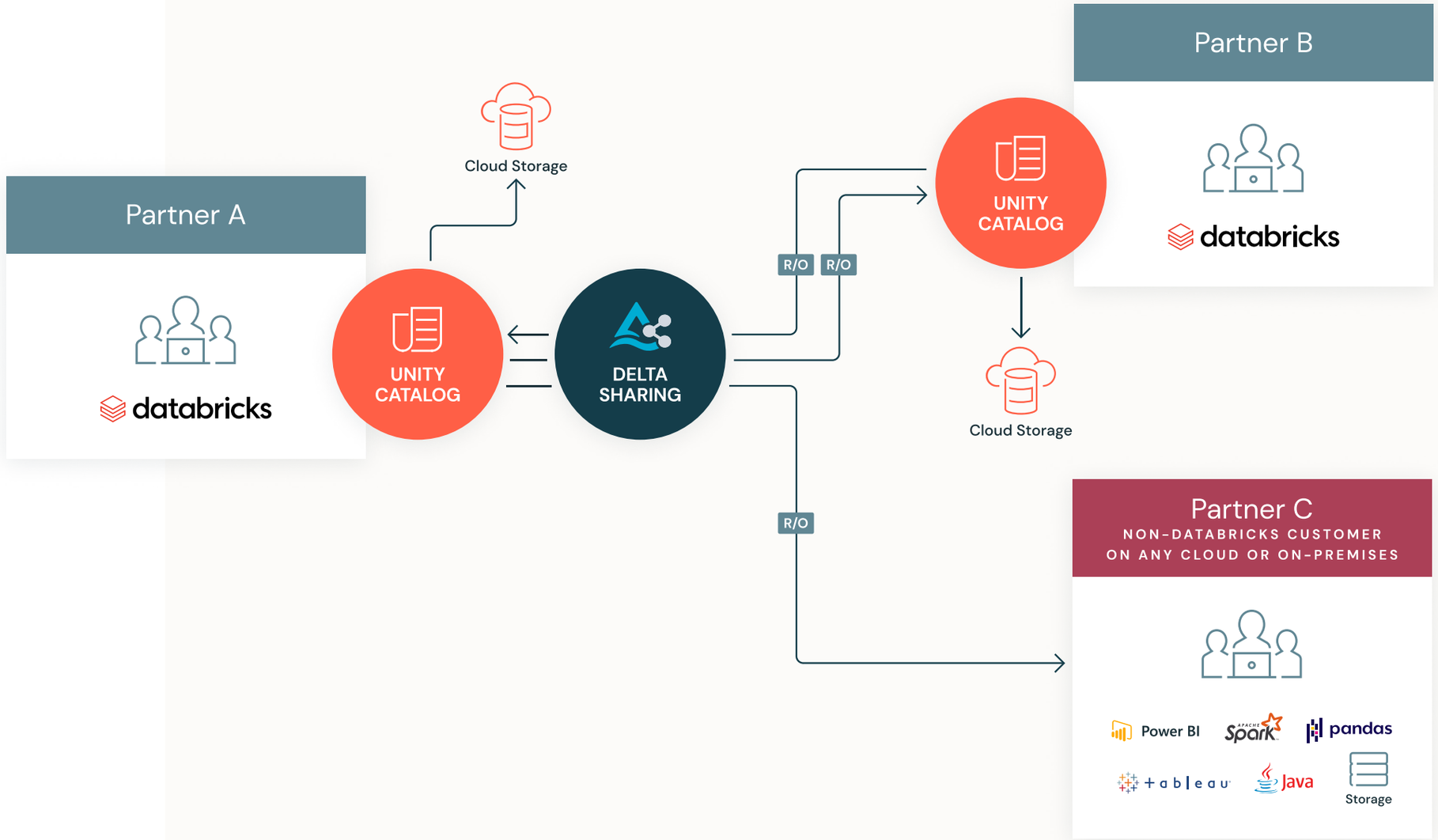


Figure 5:
B2B sharing with Delta Sharing

Delta Sharing has become a robust solution for bidirectional data exchange, enabling companies to seamlessly integrate partners, customers and suppliers into their workflows.

Traditionally peer-to-peer sharing isn't an easy task. An organization has no control over how their partners are implementing their own data platforms. Different partners may also use various formats, protocols and methods (APIs, CSV, JSON, FTP, HTTP). The complexity further increases when we consider that the partners and suppliers can reside in public cloud, private cloud or on-premises deployed data platforms. The choices of platform and architecture aren't imposed on your partners and suppliers. Delta Sharing addresses these challenges with its open protocol, allowing data to be shared across clouds, platforms and regions without imposing specific architecture choices on partners and suppliers. This flexibility is supported by a wide array of connectors that enable data to land wherever it's needed.

Beyond just data, Delta Sharing also allows sharing of AI models with external parties to add innovative ways to collaborate. You can train your models in one place and deploy them anywhere. The shared models work with Databricks AI capabilities out of the box. Shared models appear in Unity Catalog and customers gain access to AI and governance features to productionize any model. This includes end-to-end model development capabilities, from model serving to fine-tuning, along with Unity Catalog's security and management features.

This means that you can form much more agile data exchange patterns with your partners and suppliers and attain value out of your combined data much quicker than ever before.

Third-party data licensing with Delta Sharing

Delta Sharing enables companies to monetize their data product simply and with necessary governance.

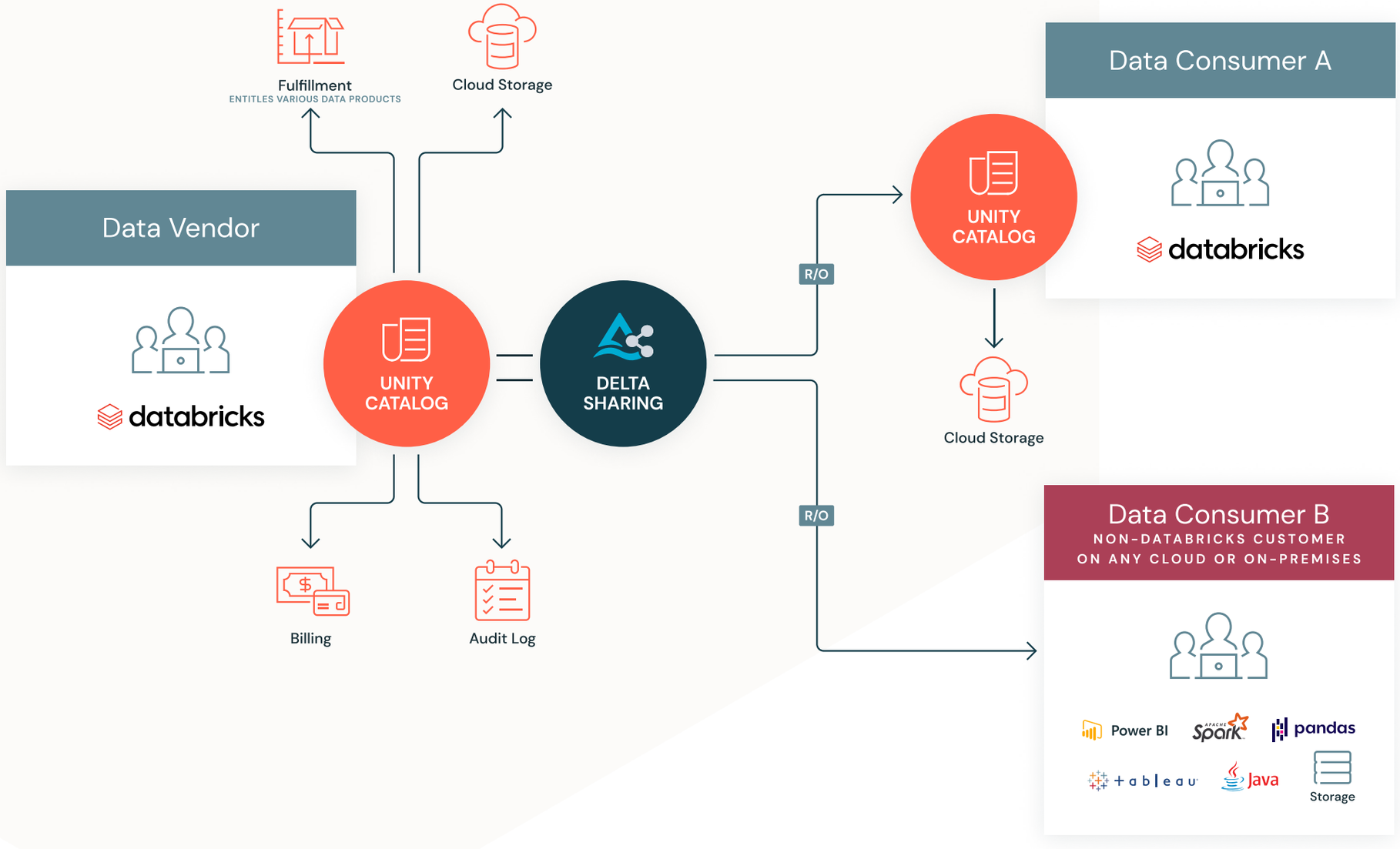
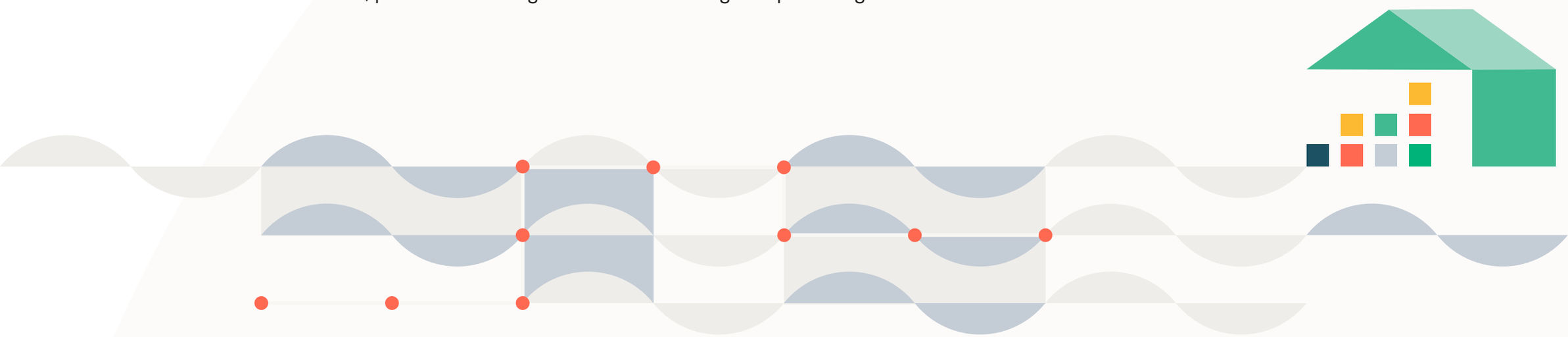


Figure 6:
Data monetization with Delta Sharing

Delta Sharing has significantly advanced the capabilities of data providers to license and monetize third-party data and AI models. It allows providers to seamlessly share large datasets without the scalability issues traditionally associated with SFTP servers. Unlike API services that require a dedicated service for each data product, Delta Sharing simplifies the process by enabling providers to grant and manage access to data recipients without replicating the data. There's no need to make multiple copies of data while sharing. Because there's no data movement while sharing, data providers avoid storage duplication and ensure that consumers get timely access to fresh, up-to-date data without delays. Any data exiting ELT/ETL pipelines can become a potential data product. Previously, data providers had to build complex integrations with a variety of platforms to reach all of their partners and customers. With cross-cloud and cross-platform sharing, data providers can expand their market reach to consumers across clouds, platforms and regions without needing complex integrations.



Databricks Marketplace further enhances these capabilities. Databricks Marketplace acts as an open forum for exchanging data and AI products, leveraging Delta Sharing to provide secure sharing and easy access for data consumers. Providers can list datasets, AI models, notebooks and Solution Accelerators on Databricks Marketplace, making them accessible to a broader audience. This platform supports both public listings and private exchanges, where listings are shared only with approved users.

By integrating with Databricks Marketplace, Delta Sharing enables providers to reach new buyers and accelerate sales cycles by reducing time to insight for consumers. The robust Databricks Marketplace infrastructure allows providers to showcase their offerings effectively while ensuring compliance with security and governance standards through Delta Sharing's open protocol.

To mitigate cost concerns, Delta Sharing maintains an audit log that tracks any permitted access to the data. Data providers can use this information to determine the costs associated with any of the data products and evaluate if such products are commercially viable and sensible. Additionally, Delta Sharing minimizes costs by eliminating the need for data replication. The protocol also supports integration with Cloudflare R2, a storage solution that incurs no egress fees, further lowering costs for cross-region or cross-cloud data sharing.

Chapter 4
How Delta Sharing Works

Delta Sharing is designed to be simple, scalable, nonproprietary and cost-effective for organizations that are serious about getting more from their data. Delta Sharing is natively integrated with Unity Catalog, which enables customers to add fine-grained governance and security controls, making it easy and safe to share data internally or externally.

Delta Sharing is a simple REST protocol that securely grants temporary access to part of a cloud dataset. It leverages modern cloud storage systems — such as AWS S3, Azure ADLS or Google GCS — to reliably grant read-only access to large datasets. Here's how it works for data providers and data recipients.

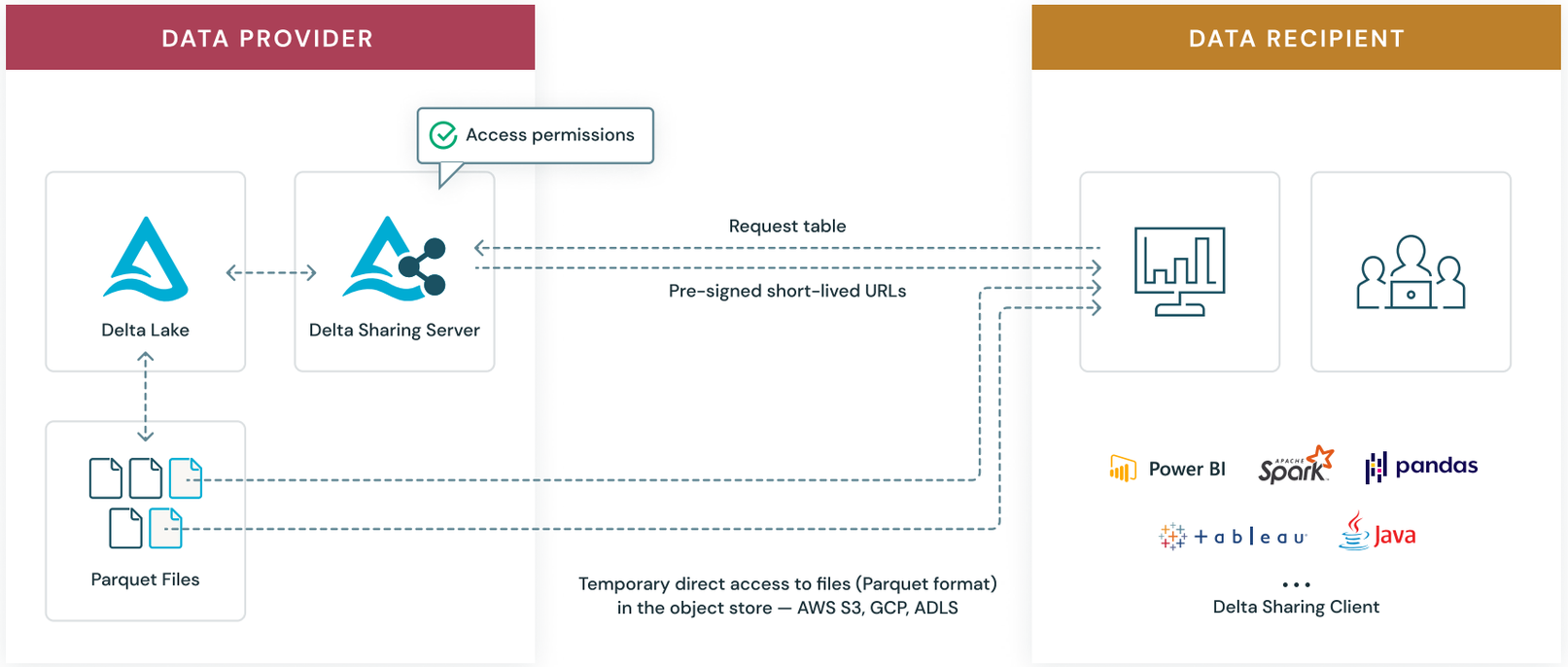


Figure 7: How Delta Sharing works connecting data providers and data recipients

Data providers

The data provider shares existing tables or parts thereof (such as specific table versions or partitions) stored on the cloud data lake in Delta Lake format. The provider decides what data they want to share and runs a sharing server in front of it that implements the Delta Sharing protocol and manages recipient access. To manage shares and recipients, you can use SQL commands, the Unity Catalog CLI or the intuitive user interface.

Data recipients

The data recipient only needs one of the many Delta Sharing clients that support the protocol. Databricks has released open source connectors for pandas, Apache Spark, Java and Python, and is working with partners on many more.

The data exchange

The Delta Sharing data exchange follows three efficient steps:

1. The recipient's client authenticates to the sharing server and asks to query a specific table. The client can also provide filters on the data (for example, "country=US") as a hint to read just a subset of the data.
2. The server verifies whether the client is allowed to access the data, logs the request and then determines which data to send back. This will be a subset of the data objects in cloud storage systems that make up the table.
3. To allow temporary access to the data, the server generates short-lived presigned URLs that allow the client to read Parquet files directly from the cloud provider. This allows read-only access in parallel at massive bandwidth without streaming through the sharing server.

Chapter 5
Introducing Databricks Marketplace

Enterprises need open collaboration for data and AI. Data sharing — within an organization or externally — allows companies to collaborate with partners, establish new partnerships and generate new revenue streams with data monetization.

The demand for generative AI is driving disruption across industries, increasing the urgency for technical teams to build generative AI models and large language models (LLMs) on top of their own data to differentiate their offerings.

Traditional data marketplaces are restricted and offer only data or simple applications, which limits their value to data consumers. They also don't offer tools to evaluate the data assets beyond basic descriptions or examples. Finally, data delivery is limited, often requiring ETL or a proprietary delivery mechanism.

Enterprises need a better way to share data and AI that is flexible, secure and unlocks business value. An ecosystem makes data sharing and collaboration powerful.

Data marketplaces present many challenges, and collaboration can be complex for both data consumers and data providers.

Data Consumers	Data Providers
Focus on data only or simple applications	Limited opportunities to monetize new types of assets
Lengthy discovery and evaluation	Difficulty reaching more users
Delayed time to insights with vendor lock-in	Lack of secure technology and unified governance

Challenges in today's data marketplaces

Data Consumers	Data Providers
Focus on data only or simple applications: Accessing only datasets means organizations looking to take advantage of AI and machine learning need to look elsewhere or start from scratch, causing delays in driving business insights.	**Limited opportunities to monetize new types of assets:** A data-only approach means organizations are limited to monetizing anything beyond a dataset and will face more friction to create new revenue opportunities with noncompatible platforms.
Lengthy discovery and evaluation: The tools most marketplaces provide for data consumers to evaluate data are simply descriptions and example SQL statements. Minimal evaluation tools means it takes more time to figure out if a data product is right for you, which might include more time in back-and-forth messages with a provider or searching for a new provider altogether.	**Difficulty reaching more users:** Data providers must choose between forgoing potential business or incurring the expense of replicating data.
Delayed time to insights with vendor lock-in: Delivery through proprietary sharing technologies or FTP means either vendor lock-in or lengthy ETL processes to get the data where you need it to be to work with it.	**Lack of secure technology and unified governance:** Without open standards for sharing data securely across platforms and clouds, data providers must use multiple tools to secure access to scattered data, leading to inconsistent governance.

What is Databricks Marketplace?

Databricks Marketplace is an open marketplace for all your data, analytics and AI, powered by Delta Sharing.

Since Databricks Marketplace is powered by Delta Sharing, you can benefit from open source flexibility and no vendor lock-in, enabling you to collaborate across all platforms, clouds and regions. This open approach allows you to put your data to work more quickly in every cloud with your tools of choice.

Databricks Marketplace brings together a vast ecosystem of data consumers and data providers to collaborate across a wide array of datasets without platform dependencies, complicated ETL, expensive replication and vendor lock-in.

Databricks Marketplace also supports AI model sharing to provide users access to both OSS and proprietary AI (first- and third-party) models. This enables data consumers and providers to discover and monetize AI models and integrate AI into their data solutions.

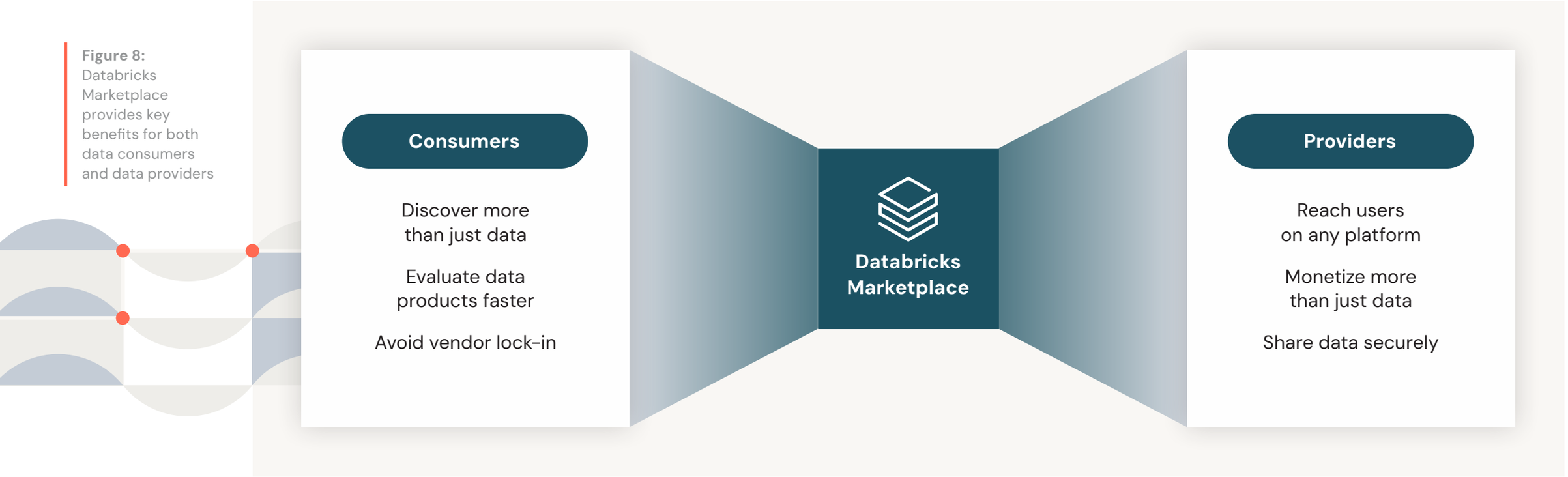
Databricks Marketplace offers a wide variety of AI assets, including pretrained generative AI models, large language models (LLMs) and industry-specific models from external providers like AI21 Labs, John Snow Labs, OLA Krutrim, Bitext and more. These models can be used as is or fine-tuned with custom data for specific use cases.

Key benefits of Databricks Marketplace

Since Databricks Marketplace is powered by Delta Sharing, you can benefit from open source flexibility and no vendor lock-in, enabling you to collaborate across all platforms, clouds and regions. This open marketplace brings together a vast ecosystem of data consumers and data providers to collaborate across a wide array of datasets without platform dependencies, complicated ETL, expensive replication and vendor lock-in.

Beyond data, providers can monetize AI models. Consumers can evaluate those AI models with rich previews, including visualizations and pre-built notebooks with sample data. With the click of a button, consumers can install the AI models. All of this works out of the box with the AI capabilities of the Databricks Data Intelligence Platform for both real-time and batch inference.

Figure 8: Databricks Marketplace provides key benefits for both data consumers and data providers



Here's what data providers and data consumers are saying about Databricks Marketplace.

Dun & Bradstreet delivers real-time access to global datasets with Databricks Marketplace

"Reliable, trusted and up-to-date data is the backbone of informed decision-making. The power of Dun & Bradstreet's datasets and analytical insights and the openness, scalability and security of Databricks Marketplace provide a strong foundation for organizations to put the power of data to work for them when and where needed to accelerate their business objectives."

— **Ginny Gomez,** President, Dun & Bradstreet, North America

HealthVerity offers de-identified healthcare data on Databricks Marketplace

"We're excited to deepen our partnership with Databricks and expand our presence through Databricks Marketplace. This collaboration empowers our clients — including pharmaceutical companies, government agencies and healthcare organizations — with unparalleled ease of access to comprehensive, de-identified healthcare data from the nation's largest real-world healthcare data ecosystem. This enables them to accelerate scientific discoveries and achieve transformative health outcomes. Together we're setting a new standard for healthcare data privacy, governance and interoperability."

— **Andrew Kress**, CEO, HealthVerity

The Trade Desk offers customers the ability to leverage first-party data through Databricks Marketplace for the first time

"Our partnership with Databricks revolutionizes how our customers unlock their data in digital media buying, enhancing their ability to harness real-time insights for more effective advertising campaigns. By integrating with the open ecosystem of Databricks Delta Sharing and the powerful predictive analytics capabilities of the Databricks Data Intelligence Platform, coupled with The Trade Desk's industry-leading advertising technology, we're empowering marketers to optimize their campaigns and achieve unprecedented levels of precision and efficiency."

— **Jay Goebel**, VP of Data Partnerships, The Trade Desk

Shutterstock's image datasets are now available on Databricks Marketplace

"Shutterstock is bringing their vast collection of nearly a billion creative content assets to Databricks Marketplace, a platform renowned for fostering open data and AI collaboration. This integration provides unparalleled access to our extensive library of ethically sourced visual content, propelling responsible AI and ML initiatives forward across various industries. We're excited to add Delta Sharing as a method to deliver data. Customers utilizing our rich dataset on Databricks can tap into new opportunities, catalyze product innovations and secure a competitive advantage."

— **Aimee Egan**, Chief Enterprise Officer, Shutterstock

Databricks Marketplace drives innovation and expands revenue opportunities.

Data Consumers

For data consumers, Databricks Marketplace dramatically expands the opportunity to deliver innovation and advance analytics and AI initiatives

Discover more than just data: Access more than just datasets, including AI models, notebooks, applications and solutions

Evaluate data products faster: Pre-built notebooks and sample data help you quickly evaluate and have much greater confidence that a data product is right for your AI or analytics initiatives. Obtain the fastest and simplest time to insight.

Avoid vendor lock-in: Substantially reduce the time to deliver insights and avoid lock-in with open and seamless sharing and collaboration across clouds, regions or platforms. Directly integrate with your tools of choice right where you work.

Data Providers

For data providers, Databricks Marketplace gives them the ability to reach new users and unlock new revenue opportunities

Reach users on any platform: Expand your reach across platforms and access a massive ecosystem beyond walled gardens. Streamline delivery of simple data sharing to any cloud or region, without replication.

Monetize more than just data: Monetize the broadest set of data assets, including datasets, notebooks and AI models, to reach more data consumers

Share data securely: Share all your datasets, notebooks, AI models, dashboards and more securely across clouds, regions and data platforms

Enable collaboration and accelerate innovation

Powered by a fast, growing ecosystem

Datarbricks Marketplace is the fastest-growing data marketplace. Since its launch, we've continued to increase partners across industries, including Retail; Communications, Media & Entertainment; and Financial Services, with 2500+ listings you can explore in our open marketplace from 250+ providers and counting.

Use cases for an open marketplace

Organizations across all industries have many use cases for consuming and sharing third-party data, from the simple (dataset joins) to the more advanced (AI notebooks, applications and dashboards).

Advertising and Retail

Incorporate shopper behavior analysis | Ads uplift/ performance | Demand forecasting | "Next best SKU" prediction | Inventory analysis | Live weather data

Finance

Incorporate data from stock exchange to predict economic impact | Market research | Public census and housing data to predict insurance sales

Healthcare and Life Sciences

Genomic target identification | Patient risk scoring Accelerating drug discovery | Commercial effectiveness | Clinical research

For more on Databricks Marketplace, go to marketplace.databricks.com, or refer to the Resources section on page XX.

Chapter 6
Privacy-Enhanced Sharing With Databricks Clean Rooms

While the demand for external data to make data-driven innovations is greater than ever, there is growing concern among organizations around data privacy. The need for organizations to share data and collaborate with their partners and customers in a secure, governed and privacy-centric way is driving the concept of "data clean rooms."

What is a data clean room?

A data clean room provides a secure, governed and privacy-enhanced environment where participants can bring their sensitive data and perform joint analysis on that private data. Participants have full control of the data and can decide which participants can perform what analysis without exposing the underlying sensitive data.

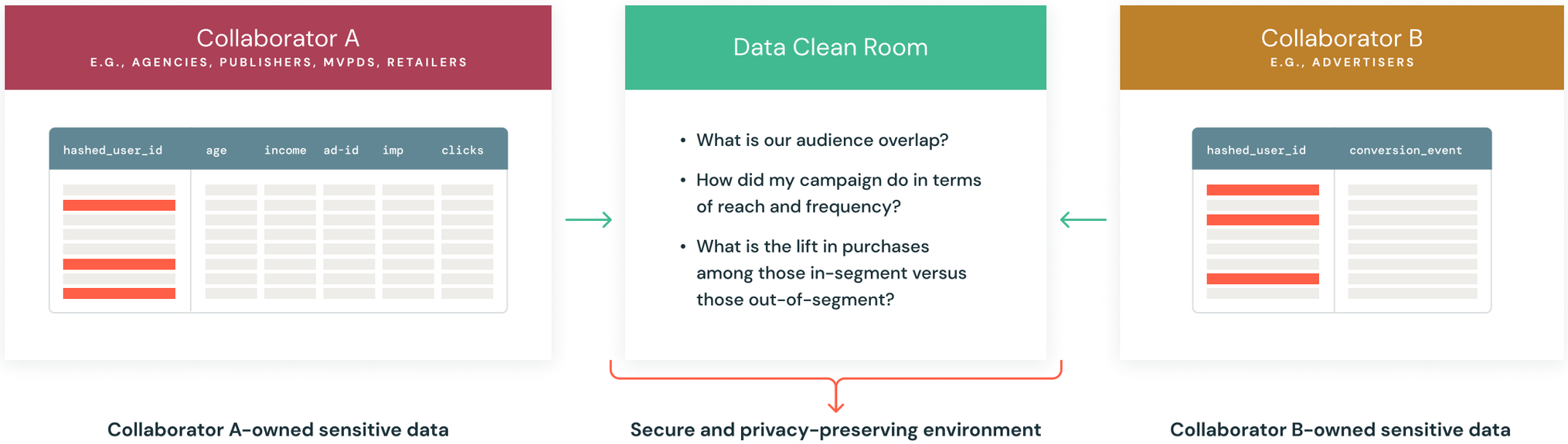


Figure 9: Data clean room diagram example for audience overlap analysis in advertising

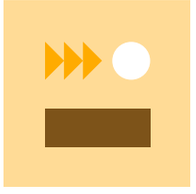
Data clean rooms have gained renewed attention as organizations seek privacy-compliant ways to collaborate on data. This trend is driven by stringent data privacy regulations like GDPR and CCPA, which have reshaped how data is collected, used and shared.

Despite the growing interest in data clean rooms, enterprises are still in the early stages of their data collaboration journey. According to the 2024 IDC External Data Sourcing and Collaboration Survey, only one-third of enterprises have started using data clean rooms, and those that have are only working with one or two data collaboration partners. Many organizations face significant challenges related to technology and data management when implementing data clean rooms. Nearly 56% of enterprises have concerns about privacy and/or consent regarding the treatment of data being shared.

Common data clean room uses cases

Campaign optimization for media and entertainment

By creating a clean room environment, media companies can securely unlock the value of their audience data by combining it with their advertiser's first-party data. This allows them to perform in-depth analysis and identify shared audience segments and post-campaign measurements without directly accessing or exposing individual user information.



Supply chain optimization for retail and consumer goods

Clean rooms enable real-time collaboration between retailers and suppliers, ensuring secure information exchange for demand forecasting, inventory planning and supply chain optimization. This improves product availability, reduces costs and streamlines operations for both parties.

Real-world evidence (RWE) for healthcare

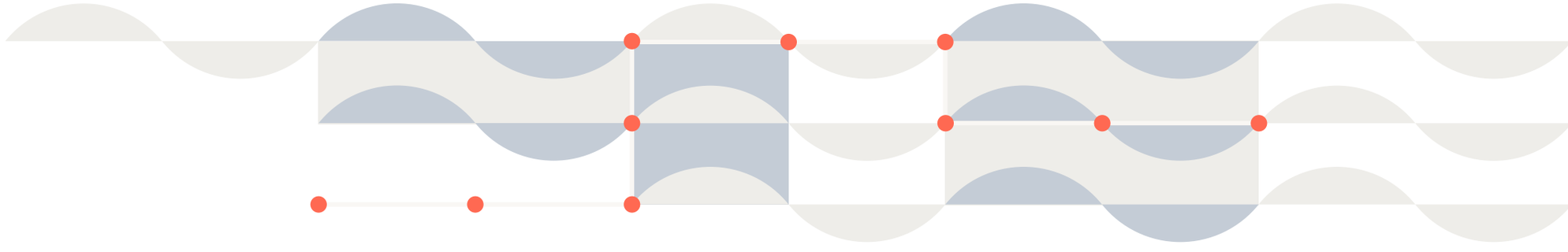
Clean rooms provide secure access to healthcare datasets, allowing collaborators to connect and query multiple sources of data without compromising data privacy. This supports RWE use cases such as regulatory decisions, safety, clinical trial design and observational research.

Know Your Customer (KYC) in banking

KYC standards are designed to combat financial fraud, money laundering and terrorism financing. Clean rooms can be used within a given jurisdiction to allow financial services companies to collaborate and run shared analytics to build a holistic view of a transaction for investigations.

Personalization with expanded interests for retailers

Clean rooms enable retailers to augment their knowledge of consumers to suggest new products and services that are relevant to the individual but haven't yet been purchased.



Shortcomings of existing data clean rooms

Organizations exploring clean room options are finding some glaring shortcomings in the existing solutions that limit the full potential of the "clean rooms" concept.

First, many existing data clean room vendors require data to be on the same cloud, same region and/or same data platform. Participants then have to move data into proprietary platforms, which results in lock-in and additional data storage costs.

Second, existing clean rooms often lack the flexibility needed to support complex analyses, and they're pretty much restricted to SQL. While SQL is great and absolutely needed for clean rooms, it can be limiting and prevent organizations from fully leveraging their data for advanced insights and innovation. The inability to run diverse workloads or integrate with AI/ML frameworks limits the potential applications of data clean rooms in unlocking valuable insights.

Finally, setting up data clean rooms can be complex due to a lack of automation. This manual setup process leads to longer ramp-up times and increased total cost of ownership (TCO), making it challenging for organizations to deploy and scale these solutions quickly. The high costs associated with implementation and maintenance can be prohibitive, especially for smaller enterprises.

Privacy-safe collaboration with Databricks Clean Rooms

Databricks Clean Rooms is powered by Delta Sharing and allows businesses to easily collaborate with their customers and partners on any cloud without compromising privacy or sharing sensitive data. When collaborating in a clean room, your data stays in place and you're always in control of where and how the data is being used.

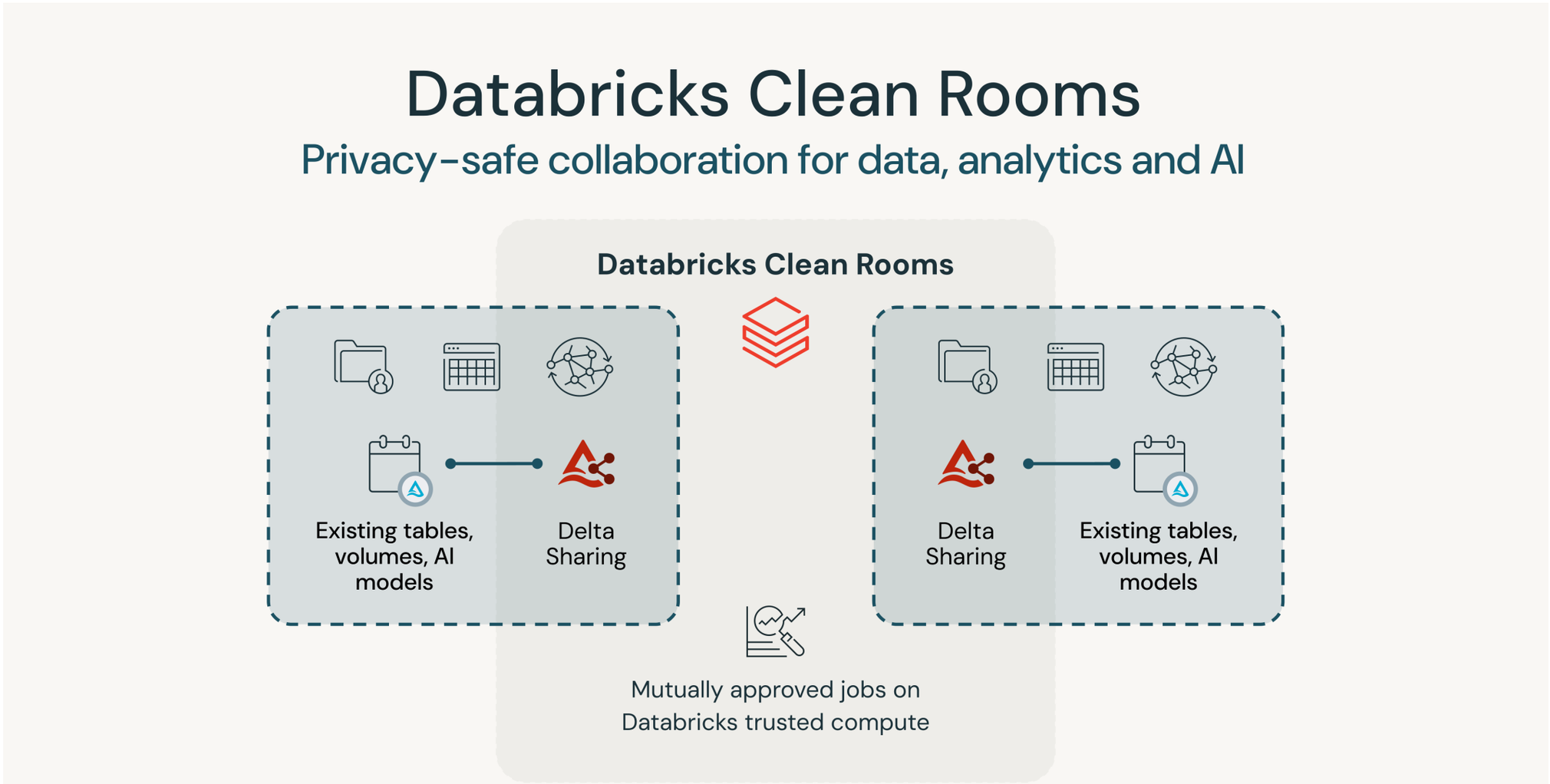


Figure 10: Databricks Clean Rooms

Any cloud, any platform

Databricks Clean Rooms is built for collaboration across clouds and across platforms. You can choose any cloud and region to start your clean rooms. Data stays where it is, and collaboration happens without the need to copy data. You collaborate on not only Databricks data, but also source data from outside of Databricks, thanks to Sharing for Lakehouse Federation. Databricks Clean Rooms truly supports you and meets your collaborators' needs wherever their data resides.

Any scale, any trust level

Databricks Clean Rooms is scalable and supports automating your privacy-safe workload with APIs, SQL commands and built-in workflow orchestration. You can also easily access your Databricks Clean Rooms outputs in notebooks or in your Unity Catalog so you can use them for other workloads. Databricks Clean Rooms also supports collaboration between multiple parties at different trust levels using different approval modes.

Any language, any workload

When collaborating using data clean rooms, the need for different programming languages becomes essential due to the diverse nature of tasks that participants from multiple organizations may need to perform. For instance, SQL is excellent for data querying and manipulation, while Python is preferred for machine learning and statistical analysis. Scala and Java are often used for building scalable data processing applications. Multilanguage support in Databricks Clean Rooms enables users to choose the best language for their specific workload, whether it's simple data joins or complex ML/AI computations. Databricks Clean Room supports collaboration on any format of data and AI models while protecting the privacy of the raw content. Leveraging the full power of Databricks Notebooks, you can run SQL or Python for complex compute and ML/AI workloads. And more language support is on the way.

"Mastercard is leading with new insights that solve our customers' needs and real problems. These insights are founded in data that can be a company's biggest asset" said Andrew Reiskind, Mastercard's chief data officer. "Accordingly, we're always looking at how we protect the confidentiality, privacy and security of that data when we use it. Databricks Clean Rooms is a solution that allows us to protect the information aligned to our Data and Tech Responsibility Principles, while giving visibility into trends. Databricks Clean Rooms offers new innovative opportunities for our customers to drive insights and value-added services. Partnering with Databricks, we've piloted both frameworks and developed a set of integrated PET capabilities to offer flexibility, scale and transparency."

— **Andrew Reiskind,** Chief Data Officer, Mastercard

How it all comes together

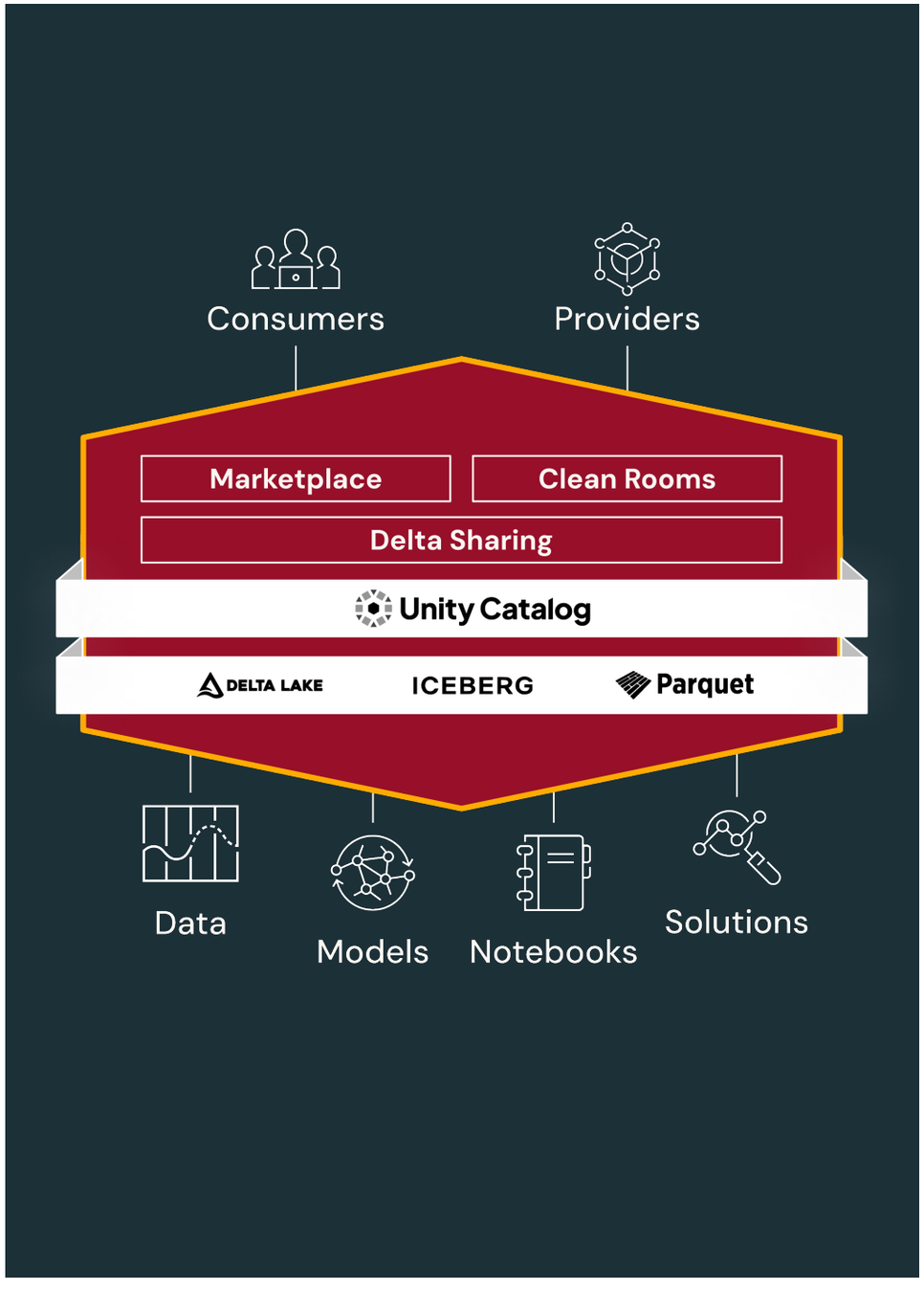
All of this is made possible with the Databricks Data Intelligence Platform, which is built for sharing and collaboration. Data sharing and collaboration in the Databricks Platform is built on top governance (Unity Catalog) and storage layer (Delta Lake, Iceberg, Parquet).

As we discussed earlier in this book, Databricks Marketplace is the open marketplace for all your data, analytics and AI. Databricks Clean Rooms allows businesses to easily collaborate in a secure environment with their customers and partners on any cloud in a privacy-safe way. Both of these are powered by Delta Sharing. And all of this is secured and governed by Unity Catalog.

The integration of Delta Sharing with Unity Catalog ensures that data sharing is secure and governed, meeting compliance and privacy requirements. Unity Catalog provides centralized governance and security for data sharing. It allows organizations to manage, audit and track the usage of shared data, ensuring compliance with security and regulatory requirements.

Delta Lake, Apache Iceberg and Parquet are foundational technologies that enable efficient data sharing and collaboration across industries.

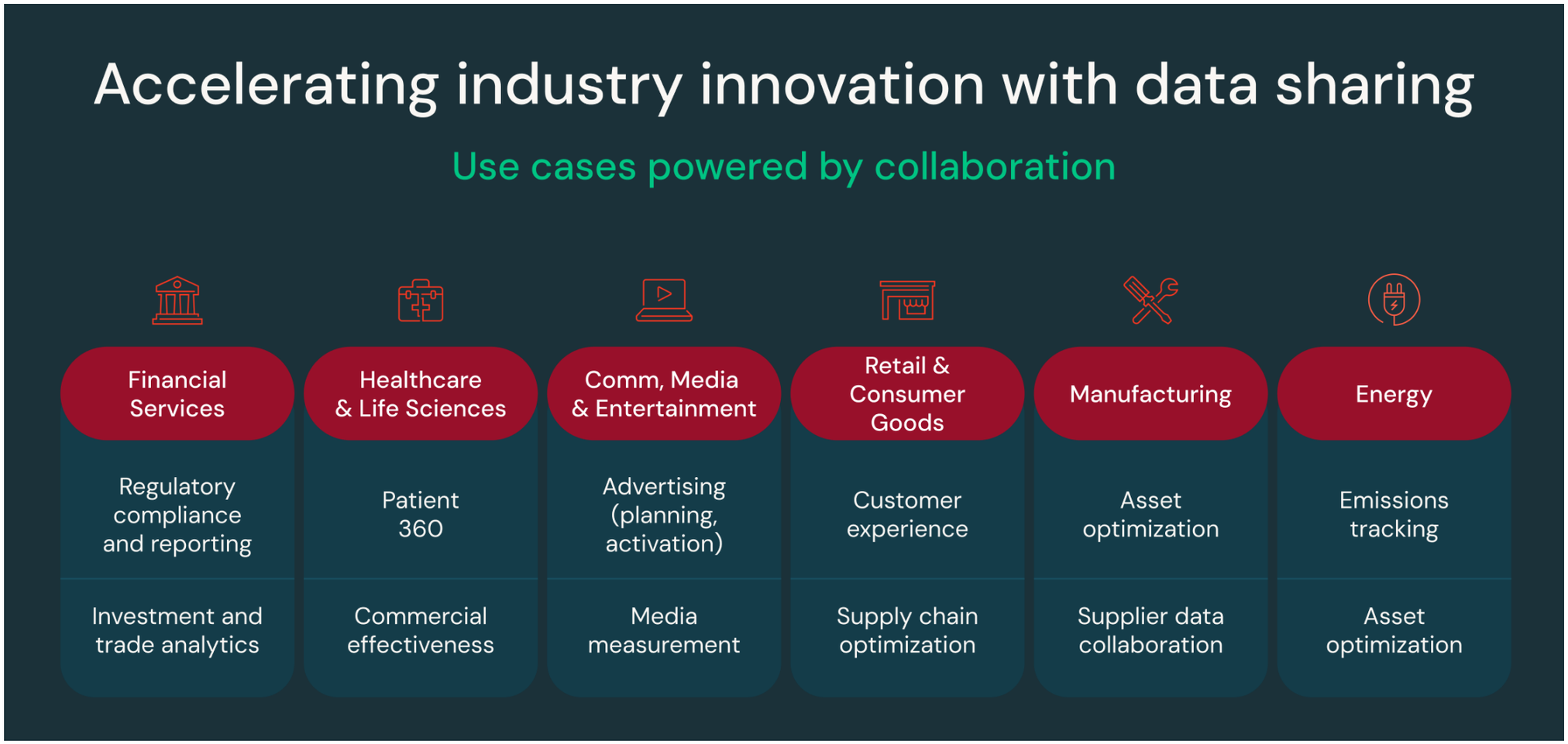
By integrating the capabilities required by data engineers, data scientists and business analysts into a single platform, Databricks eliminates the complexity and inefficiencies associated with using disparate tools. This unified approach supports a wide range of data analysis and AI tasks.



Data sharing across industries

Data sharing has become a critical enabler across multiple industries, driving innovation, efficiency and collaboration. Here are some key examples:

- **Retail:** Data sharing helps retailers create a unified customer view by integrating data from various sources like weather, events and pricing to offer personalized marketing and optimize supply chains. It also facilitates real-time collaboration with suppliers to improve inventory management and reduce response times.
- **Financial Services:** In an industry where regulatory compliance is paramount, data sharing ensures timely and accurate reporting for regulations such as AML (anti-money laundering) and KYC (Know Your Customer). Real-time access to data enhances transparency and efficiency in meeting compliance requirements.
- **Healthcare and Life Sciences:** Data sharing powers initiatives like Patient 360 by combining clinical data from electronic health records (EHRs), insurance claims and wearable devices. This holistic view improves patient outcomes and enables better collaboration across the healthcare ecosystem. Data sharing is highly relevant to real-world evidence (RWE) for life sciences companies to understand how medical treatments perform in everyday settings, beyond the controlled environment of clinical trials.
- **Manufacturing:** In industrial manufacturing, data sharing is essential for predictive maintenance and asset optimization. By securely sharing equipment performance data with suppliers, manufacturers can predict failures before they occur, driving efficiency across production lines.
- **Energy:** Data sharing supports emissions tracking and carbon offset verification by integrating diverse data sources. This helps energy companies optimize asset performance and collaborate on sustainability initiatives without compromising sensitive information.



Across these sectors, secure and privacy-compliant data sharing is transforming operations, enhancing decision-making and fostering deeper collaboration between stakeholders.

Resources
Getting Started With Data Sharing and Collaboration

Data sharing plays a key role in business processes across the enterprise, from product development and internal operations to customer experience and compliance. However, most businesses have been slow to move forward because of incompatibility between systems, complexity and security concerns.

Data-driven organizations need an open — and secure — approach to data sharing. Delta Sharing answers this need without imposing restrictions or additional costs. It's the first-ever open protocol, an open standard for sharing a dataset securely. With Delta Sharing, organizations can easily share existing large-scale datasets, based on open source formats like Apache Parquet, Iceberg and Delta Lake, without moving data. Databricks Marketplace expands on this by offering an open platform for exchanging not just data but also AI and analytics assets such as AI models, notebooks and Solution Accelerators. Databricks Clean Rooms provides a secure, privacy-safe environment for multiple parties to collaborate on sensitive data without exposing raw data.

- **Share across platforms:** You can share live datasets, AI models and notebooks across platforms, clouds and regions. This open approach is powered by Delta Sharing, the world's first open protocol for secure data sharing, which allows organizations to share data for any use case, any tool and on any cloud.
- **Share all your data and AI:** Databricks Marketplace is an open marketplace for all your data, analytics and AI, enabling both data consumers and data providers to deliver innovation and advance analytics and AI initiatives
- **Share securely:** Databricks Clean Rooms allows businesses to easily collaborate with customers and partners on any cloud in a privacy-safe way. With Delta Sharing, clean room participants can securely share data from their data lakes without any data replication across clouds or regions. Your data stays with you without vendor lock-in, and you can centrally audit and monitor its usage.

Get started with these products by exploring the resources below.

Delta Sharing

- Data Sharing on Databricks
- Learn About Databricks Unity Catalog
- Blog Post: What's New With Data Sharing and Collaboration on the Lakehouse
- Learn About Open Source Delta Sharing
- Video: What's New With Data Sharing and Collaboration on the Lakehouse
- AWS Documentation
- Azure Documentation

Databricks Marketplace

- Learn About Databricks Marketplace
- Explore Databricks Marketplace
- Video: Databricks Marketplace — Going Beyond Data and Applications
- Demo: Databricks Marketplace
- AWS Documentation: What Is Databricks Marketplace?
- Azure Documentation: What Is Databricks Marketplace?

Databricks Clean Rooms

- Learn About Databricks Clean Rooms
- Video: What's New With Data Sharing and Collaboration on the Lakehouse
- eBook: The Definitive Guide to Data Clean Rooms
- Webinar: Unlock the Power of Secure Data Collaboration with Clean Rooms
- Product Tour

About the Authors

Vuong Nguyen is a Solutions Architect at Databricks, focusing on making analytics and AI simple for customers by leveraging the power of the Databricks Data Intelligence Platform. You can reach Vuong on LinkedIn.

Somasekar Natarajan (Som) is a Solutions Architect at Databricks specializing in enterprise data management. Som has worked with Fortune organizations spanning three continents for close to two decades with one objective — helping customers to harness the power of data. You can reach Som on LinkedIn.

Harish Gaur is a Director of Product Marketing on the Data Sharing and Collaboration team. He drives product marketing for Databricks Marketplace, Databricks Delta Sharing and Databricks Clean Rooms, as well as the data partner marketing efforts. You can reach Harish on LinkedIn.

Jay Bhankharia is a Senior Director on the Databricks Data Partnerships team. His passion is helping customers gain insights from data so they can use the power of the Databricks Data Intelligence Platform for their analytics needs. You can reach Jay on LinkedIn.

Sachin Thakur is a Principal Product Marketing Manager on the Databricks Data Engineering and Analytics team. His area of focus is data governance with Unity Catalog, and he's passionate about helping organizations democratize data and AI with the Databricks Data Intelligence Platform. You can reach Sachin on LinkedIn.

Giselle Goicochea is a Senior Product Marketing Manager on the Databricks Data Engineering and Analytics team. Her area of focus is data ingestion with LakeFlow Connect, where she's dedicated to helping customers extract value from their data and accelerating innovation. You can reach Giselle on LinkedIn.

Kelly Albano is a Product Marketing Manager on the Databricks Data Engineering and Analytics team. Her area of focus is security, compliance and Databricks Clean Rooms. You can reach Kelly on LinkedIn.

About Databricks

Databricks is the data and AI company. More than 10,000 organizations worldwide — including Block, Comcast, Condé Nast, Rivian, Shell and over 60% of the Fortune 500 — rely on the Databricks Data Intelligence Platform to take control of their data and put it to work with AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake and MLflow. To learn more, follow Databricks on LinkedIn, X and Facebook.

Sign up for a free trial