eBook

A New Approach to Data Sharing

Open data sharing and collaboration for data, analytics, and AI

Second Edition

databricks

Contents

Contents

Introduction
Data Sharing in Today's Digital Economy

Today's economy revolves around data. Everyday, 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.

To be truly data-driven, organizations need a better way to share data. Gartner predicts that by 2024, organizations that promote data sharing will outperform their peers on most business value metrics. In addition, Gartner recently found that Chief Data Officers who have 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.

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 the challenges of those methods. 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.

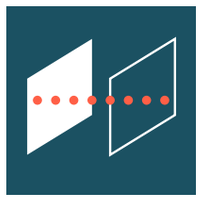
Chapter 1
What Is Data Sharing and Why Is It Important?

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 commercialization and enhanced analysis. 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. This includes but is not limited to roles such as the data analyst, data scientist and data engineer.

Common data sharing use cases

Data monetization

Companies across industries are commercializing data. 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 an agency with an identity graph to a telecommunication company with proprietary 5G data or 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.



Data sharing with partners or suppliers (B2B)

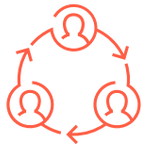
Many companies now strive to share data with partners and suppliers as similarly as they share it across their own organizations. For example, retailers and their suppliers continue to work more closely together as they seek to keep their products moving in an era of ever-changing consumer tastes. Retailers 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.

Internal lines of business (LOBs) sharing

Within any company, different departments, lines of business and subsidiaries seek to share data so 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 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.

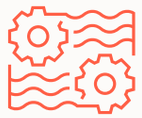
Key benefits of data sharing

As you can see from the use cases described above, 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 its 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 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, multi-cloud, high-performance solutions. Most data sharing solutions are tied to a single vendor, which creates friction for data providers and data consumers who use non-compatible platforms.

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.

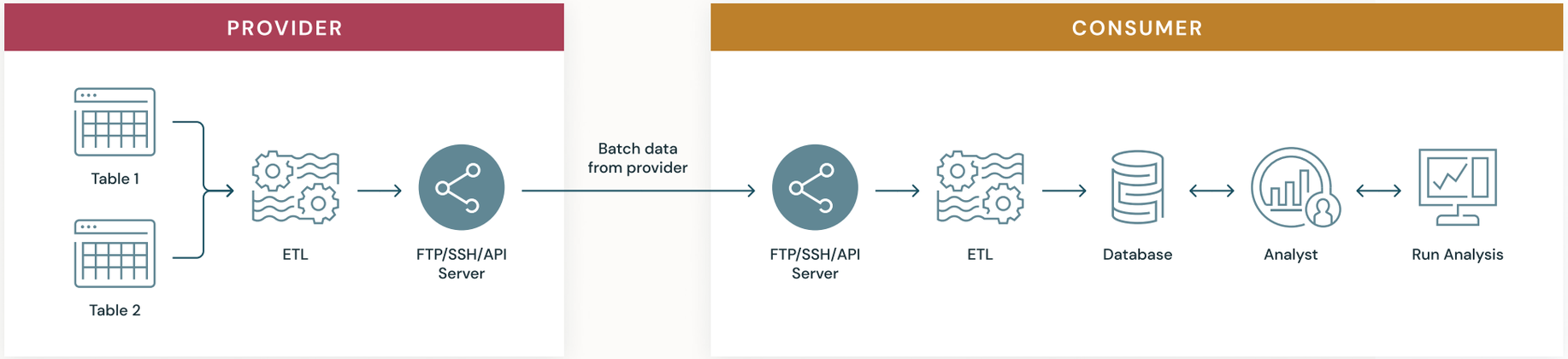


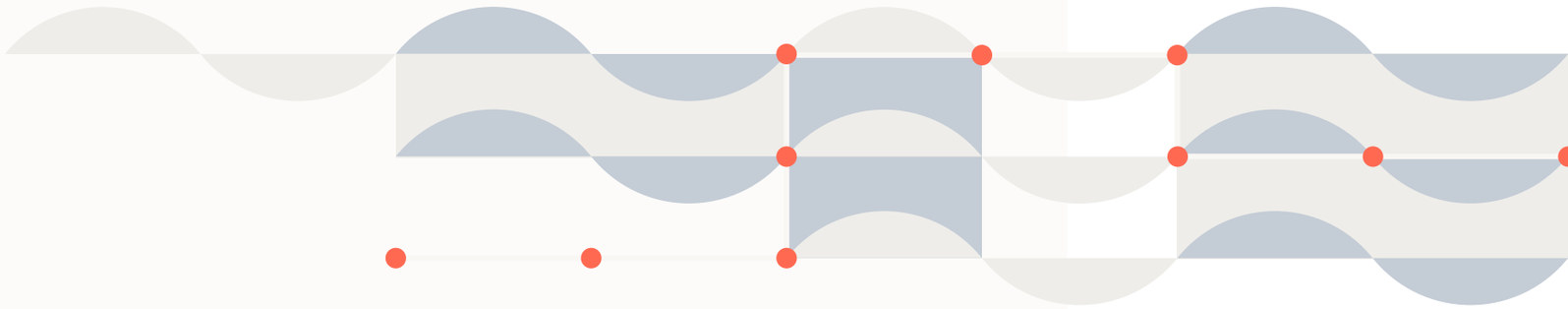
Figure 1: Legacy data sharing solutions

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-prem 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 creating copies of data sets. 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 data sets.



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.

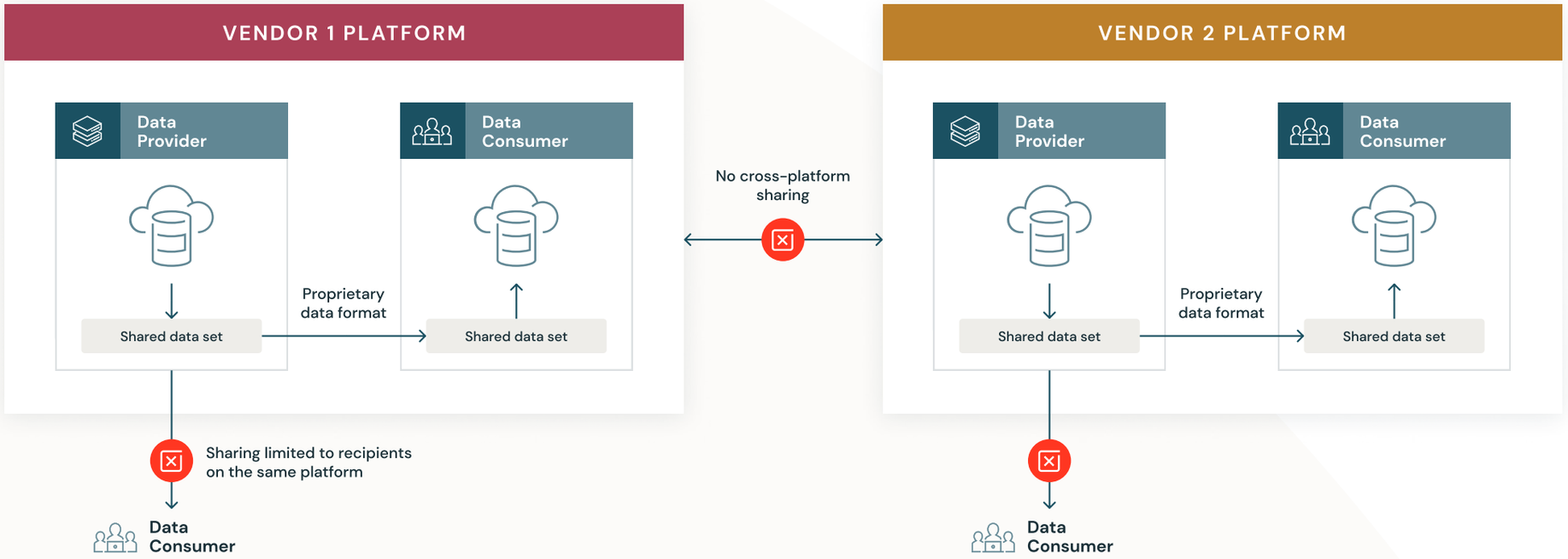


Figure 2: Proprietary vendor solutions

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 interop with other platforms well. 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 the above 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 is 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, 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 cannot 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 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.

The lack of a comprehensive solution makes it challenging for data providers and consumers to easily share data. 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 Assets

We believe the future of data sharing should be characterized by open technology. Data 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 solution 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 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.

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

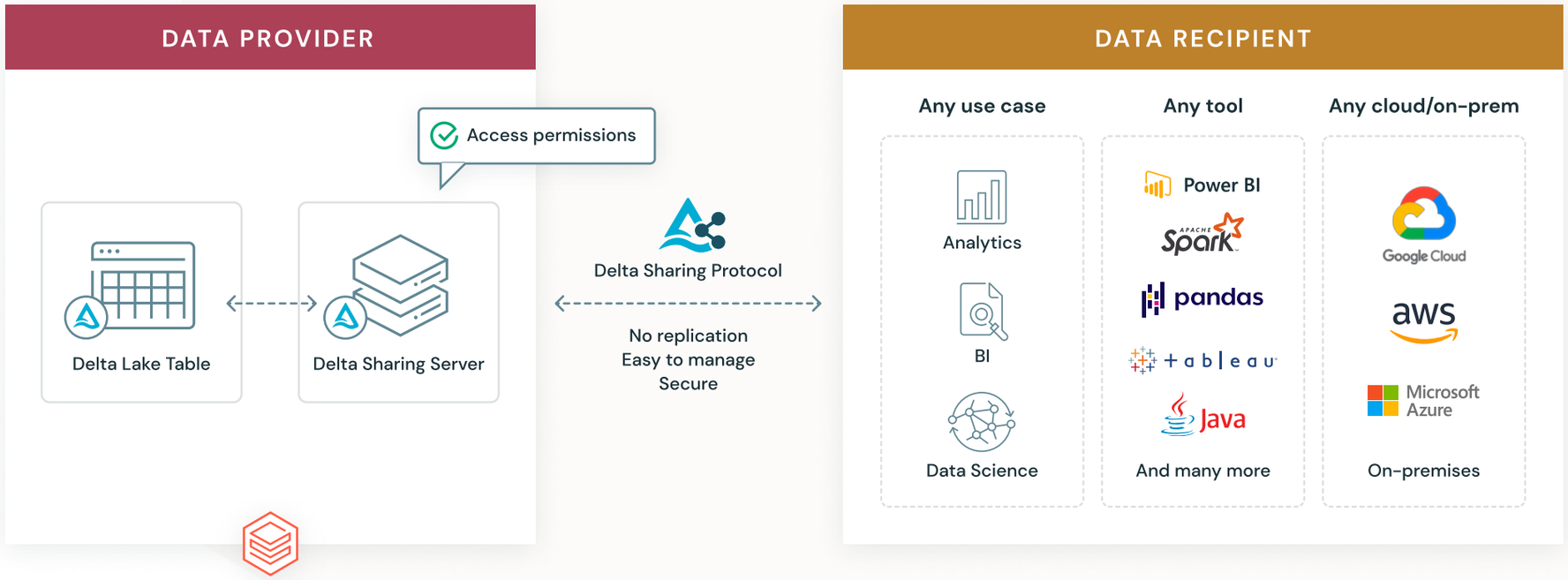


Figure 3:
Delta Sharing

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 data sets and also allow sharing of unstructured data and future data derivatives such as 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, including:

Open cross-platform sharing. Delta Sharing establishes a new open standard for secure data sharing and supports open source Delta and Apache Parquet formats. 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 data sets on the provider's data lake can easily be shared without any data replication or physical movement of data. Data providers can update their data sets 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 even revoke access to shared data sets 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 Databricks Lakehouse
- 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

Share data products, including AI models, dashboards and notebooks, with greater flexibility. Data providers can choose between sharing anentire table or sharing only a version or specific partitions of a table. However, sharing just tabular data is not enough to meet today's consumer demands. Delta Sharing also supports sharing of non-tabular data and data derivatives such as data streams, AI models, SQL views and arbitrary files, enablingincreased collaboration and innovation. Data providers can build, package and distribute data products including data sets, AI and notebooks, allowingdata recipients to get insights faster. Furthermore, this approach promotes and empowers the exchange of knowledge — not just data — between different organizations. With Delta Sharing we are able to achieve a truly open marketplace and truly open ecosystem. In contrast, commercial products are mostly limited to sharing raw tabular data and cannot 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. Incontrast, 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 data sets from popular data providers. The Delta Sharing ecosystem of open source and commercial partners is growing every day.

Maximizing value of data with Delta Sharing

Delta Sharing is already transforming data sharing activities for companies in a wide range of industries. Given the sheer variety of data available and the technologies that are emerging, it is 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 will explore the building blocks of such an approach and the use cases emerging from these.

"Delta Sharing helped us streamline our data delivery process for large data sets. This enables our clients to bring their own compute environment to read fresh curated data with little-to-no integration work, and enables us to continue expanding our catalog of unique, high-quality data products."

— **William Dague**, Head of Alternative Data, Nasdaq

"Leveraging the powerful capabilities of Delta Sharing from Databricks enables Pumpjack Dataworks to have a faster onboarding experience, removing the need for exporting, importing and remodeling of data, which brings immediate value to our clients. Faster results yield greater commercial opportunity for our clients and their partners."

— **Corey Zwart**, Head of Engineering, Pumpjack Dataworks

"We recognize that openness of data will play a key role in achieving Shell's Carbon Net Zero ambitions. Delta Sharing provides Shell with a standard, controlled and secure protocol for sharing vast amounts of data easily with our partners to work toward these goals without requiring our partners be on the same data sharing platform."

— **Bryce Bartmann**, Chief Digital Technology Advisor, Shell

"Data accessibility is a massive consideration for us. We believe that Delta Sharing will simplify data pipelines by enabling us to query fresh data from the place where it lives, and we are not locked into any platform or data format."

— **Rayne Gaisford**, Global Head of Data Strategy, Jefferies

"As a data company, giving our customers access to our data sets is critical. The Databricks Lakehouse Platform with Delta Sharing really streamlines that process, allowing us to securely reach a much broader user base regardless of cloud or platform."

— **Felix Cheung**, VP of Engineering, SafeGraph

Data monetization with Delta Sharing

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

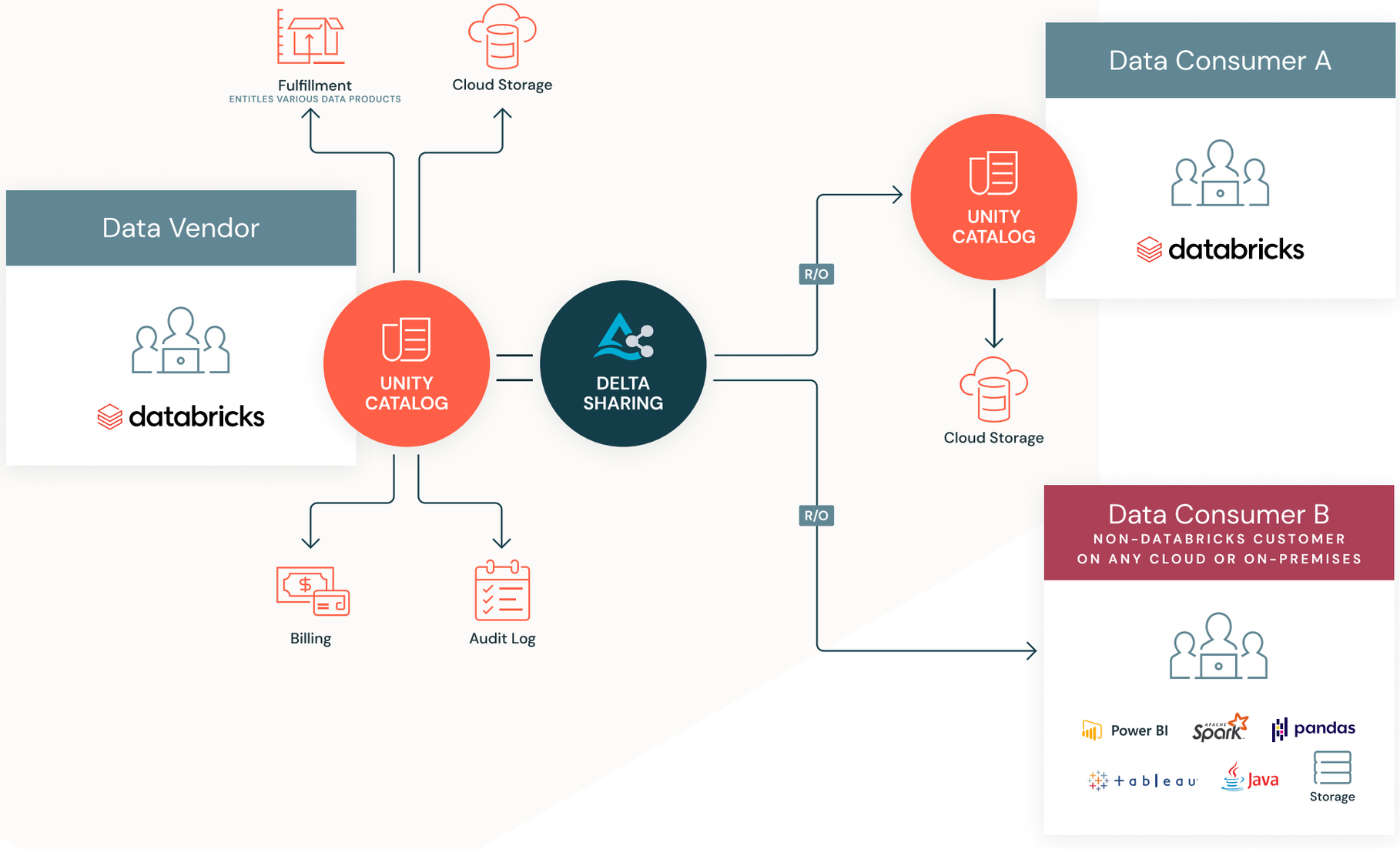
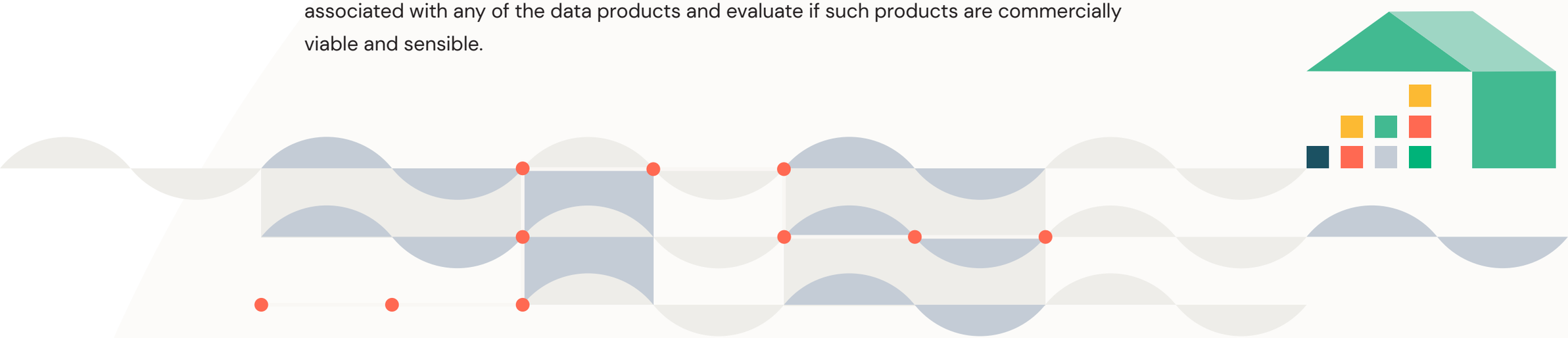


Figure 4:
Data monetization with Delta Sharing

With Delta Sharing, a data provider can seamlessly share large data sets and overcome the scalability issues associated with SFTP servers. Data providers can easily expand their data product lines since Delta Sharing doesn't require you to build a dedicated service for each of your data products like API services would. The company simply grants and manages access to the data recipients instead of replicating the data — thereby reducing complexity and latency. Any data that exits your ELT/ETL pipelines becomes a candidate for a data product. Any data that exists on your platform can be securely shared with your consumers. This grants a wider addressable market — your products have appeal to a broader range of consumers, from those who say "we need access to your raw data only" to those who say "we want only small subsets of your Gold layer data."

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.



B2B sharing with Delta Sharing

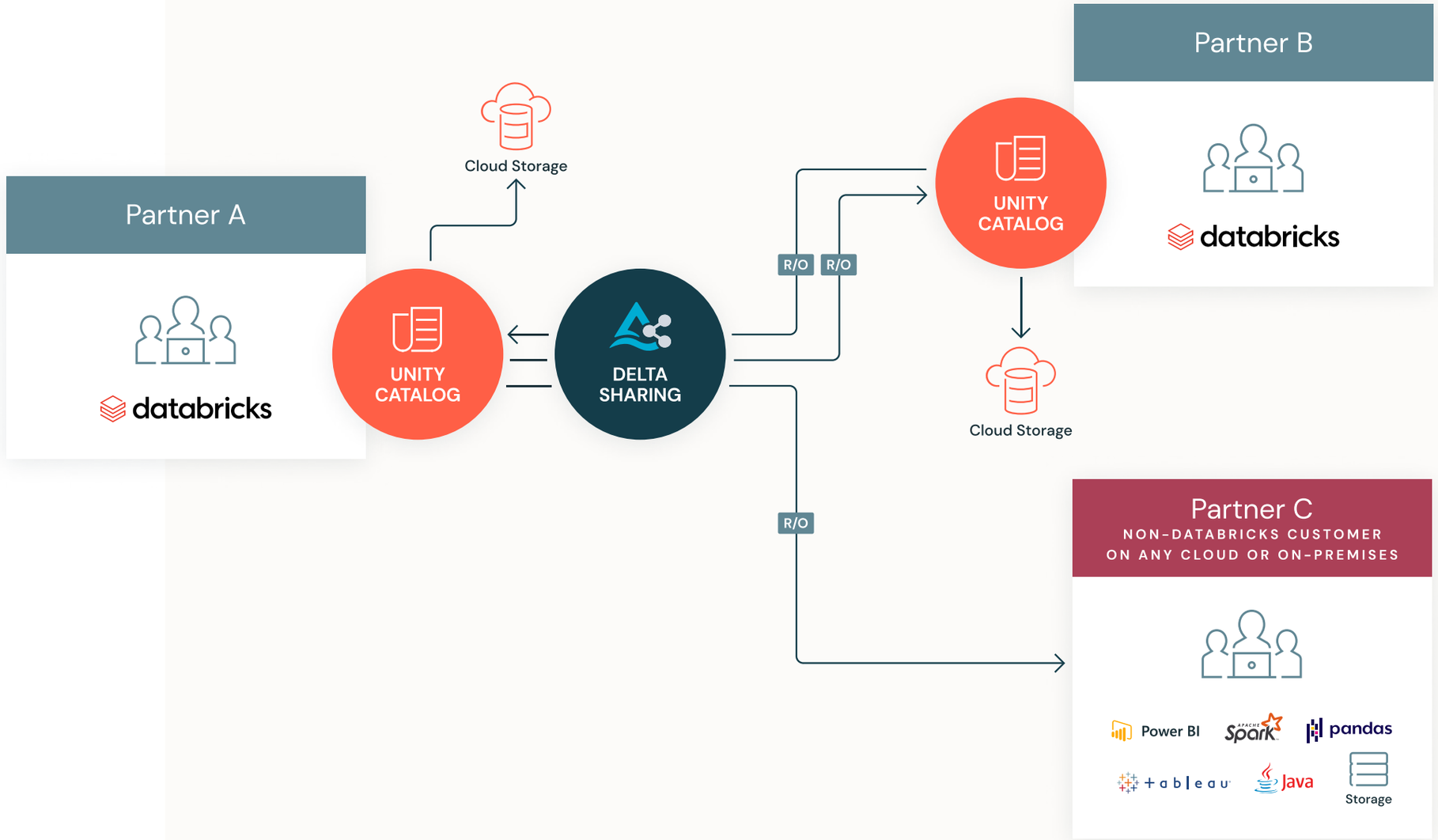
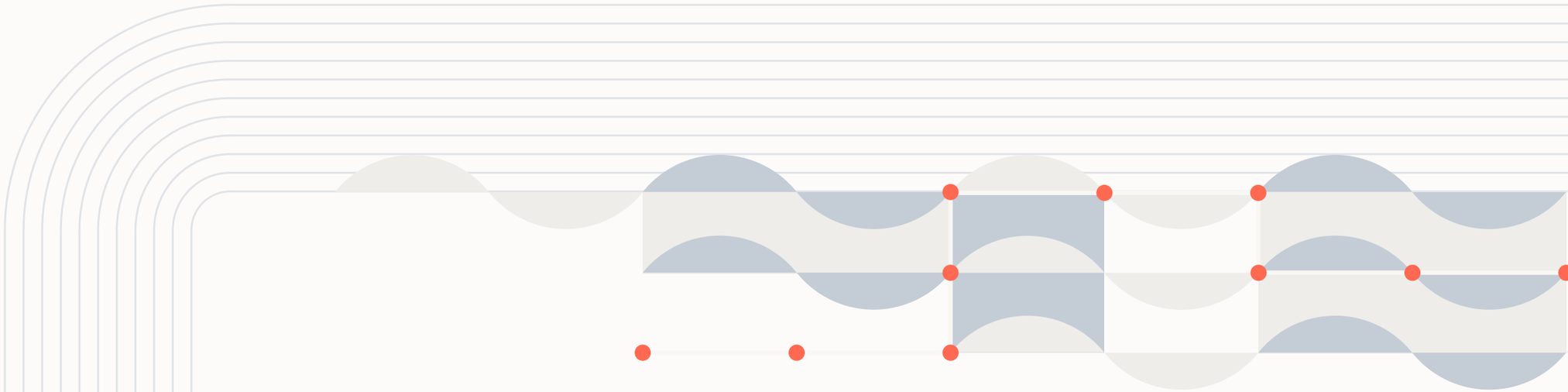


Figure 5:
B2B sharing with Delta Sharing

Delta Sharing applies in the case of bidirectional exchange of data. Companies use Delta Sharing to incorporate partners and suppliers seamlessly into their workflows. Traditionally, this is not an easy task. An organization typically has no control over how their partners are implementing their own data platforms. The complexity increases when we consider that the partners and suppliers can reside in a public cloud, private cloud or an on-premises deployed data platform. The choices of platform and architecture are not imposed on your partners and suppliers. Due to its open protocol, Delta Sharing addresses this requirement foundationally. Through a wide array of existing connectors (and many more being implemented), your data can land anywhere your partners and suppliers need to consume it.

In addition to the location of data consumer residency, the complexity of data arises as a consideration. The traditional approach to sharing data using APIs is inflexible and imposes additional development cycles on both ends of the exchange in order to implement both the provider pipelines and consumer pipelines. With Delta Sharing, this problem can be abstracted. Data can be shared as soon as it lands in the Delta table and when the shares and grants are defined. There are no implementation costs on the provider side. On the consumer side, data simply needs to be ingested and transformed into an expected schema for the downstream processes.

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



Internal data sharing 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. In this situation it is 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.

Architectural design patterns such as Data Mesh have emerged to address these specific challenges and considerations. Data Mesh architecture assumes a federated design and dissemination of ownership and responsibility to business units or divisions. This, in fact, has several advantages, chief among them that data is owned by the parts of the organization closest to the source of the data. Data residence is naturally enforced since data sits within the geo-locality where it has been generated. Finally, data volumes and data variety are kept in control due to the localization within a data domain (or data node). On the other hand, the architecture promotes exchange of data between different data domains when that data is needed to deliver outcomes and better insights.

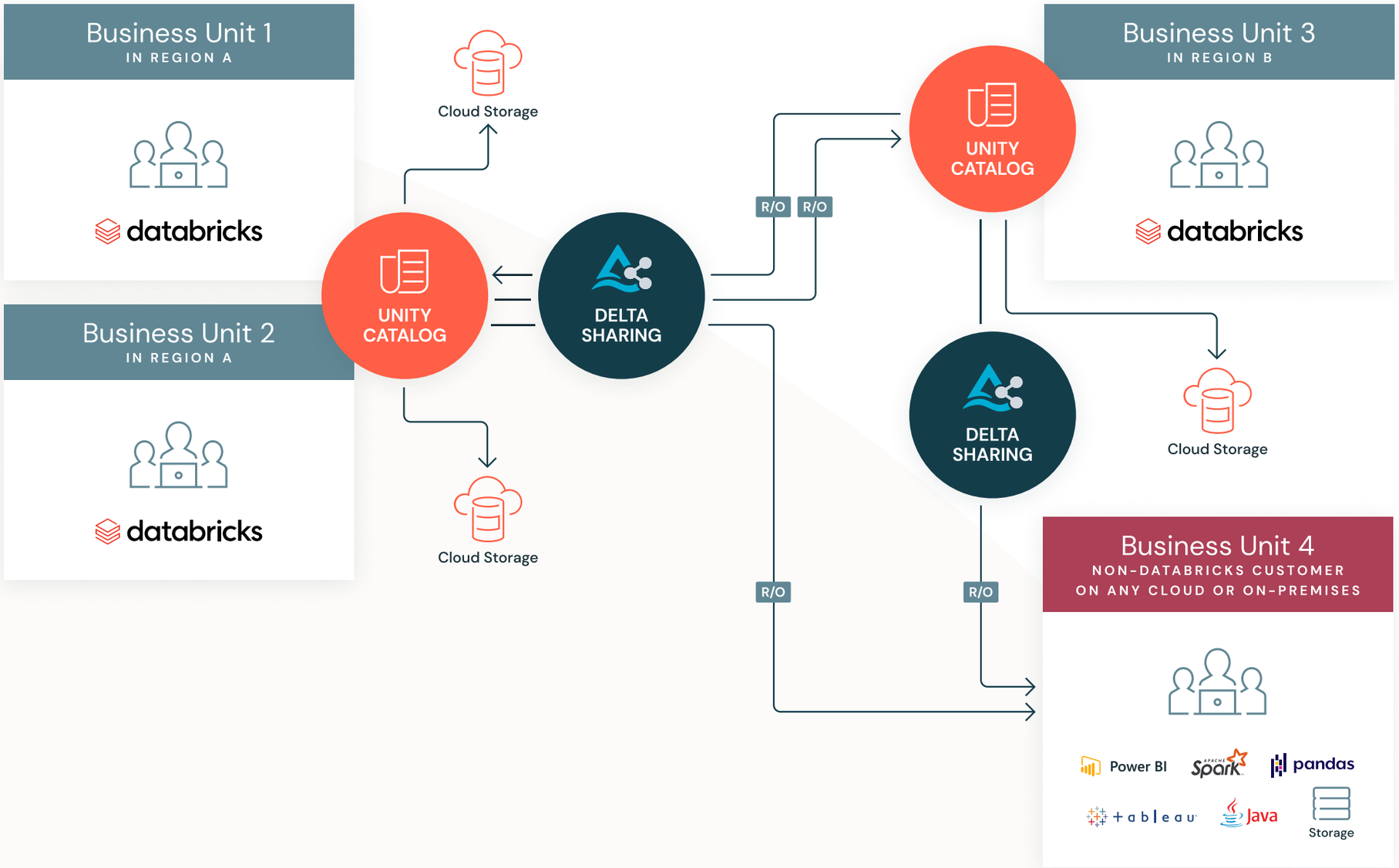


Figure 6:
Building a Data Mesh with Delta Sharing

Unity Catalog enables consolidated data access control across different data domains within an organization using the Lakehouse on Databricks. In addition, Unity Catalog adds a set of simple and easy-to-use declarative APIsto govern and control data exchange patterns between the data domains in the Data Mesh.

To make matters even more complicated, organizations can grow through mergers and acquisitions. In such cases we cannot assume that organizations being acquired have followed the same set of rules and standards to define their platforms and produce their data. Furthermore, we cannot even assume that they have used the same cloud providers, nor can we assume the complexity of their data models. Delta Sharing can simplify and accelerate the unification and assimilation of newly acquired organizations and their data and processes.. Individual organizations can be treated as new data domains in the overarching mesh. Only selected data sources can be exchanged between the different platforms. This enables teams to move freely between the organizations that are merging without losing their data — if anything, they are empowered to drive insights of higher quality by combining the data of both.

With Unity Catalog and Delta Sharing, the Lakehouse architecture seamlessly combines with the Data Mesh architecture to deliver more power than ever before, pushing the boundaries of what's possible and simplifying activities that were deemed daunting not so long ago.

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 data set. It leverages modern cloud storage systems — such as AWS S3, Azure ADLS or Google's GCS — to reliably grant read-only access to large data sets. Here's how it works for data providers and data recipients.

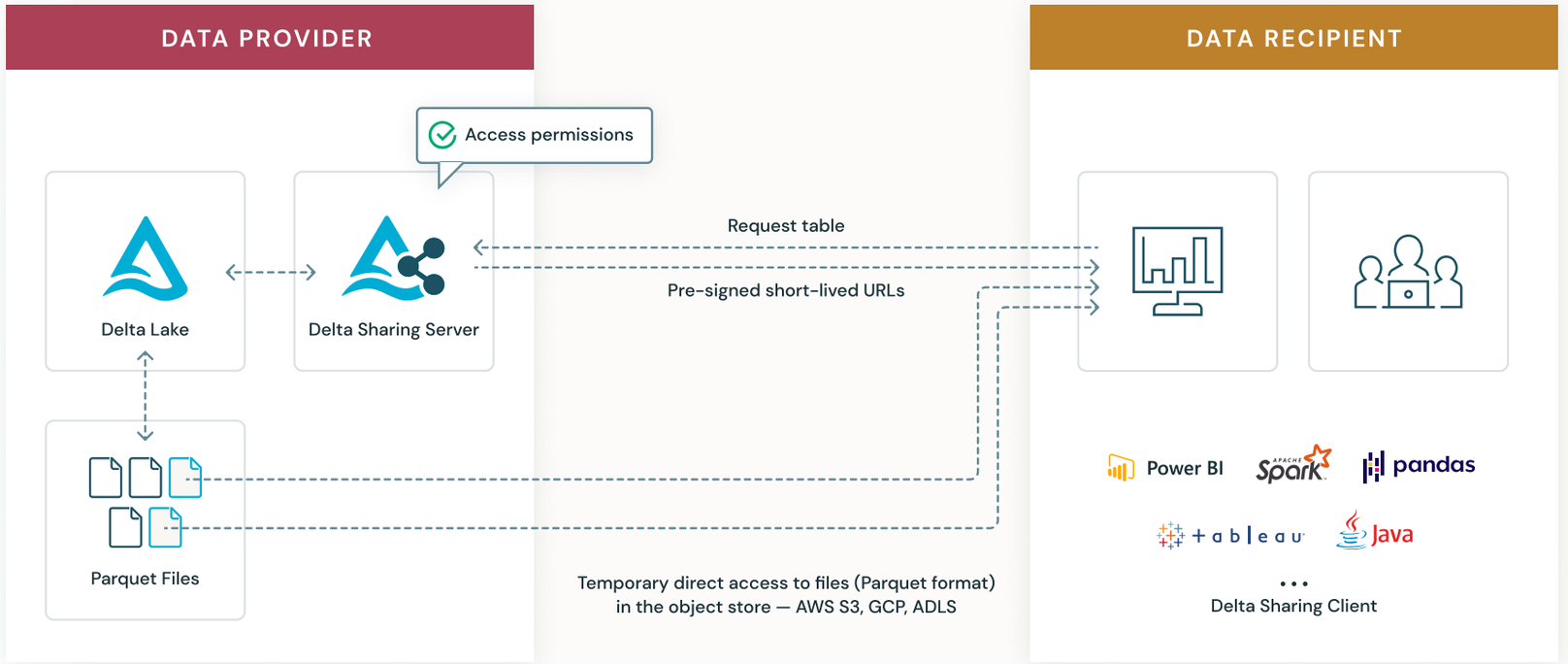


Figure 7: How Delta Sharing works connecting data providers and data recipients

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 so that the read-only access can happen 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, therefore limiting 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.

Today, 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 data sets means organizations looking to take advantage of AI/ML 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 data set and will face more friction to create new revenue opportunities with non-compatible 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 mean 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 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 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.

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

Key Benefits of Databricks Marketplace

Databricks Marketplace provides key benefits for both data consumers and data providers.

Databricks Marketplace drives innovation and expands revenue opportunities

Data Consumers

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

Discover more than just data: Access more than just data sets, 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 and right where you work.

Data Providers

For data providers, the Databricks Marketplace enables 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 data sets, notebooks, AI models to reach more data consumers.

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

Enable collaboration and accelerate innovation

Powered by a fast, growing ecosystem

Enterprises need open collaboration for data and AI. In the past few months, we've continued to increase partners across industries, including Retail, Communications and Media & Entertainment, Financial Services, with 520+ listings you can explore in our open Marketplace from 80+ 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 41.

New upcoming feature: AI model sharing

Nowadays, it may seem like every organization wants to become an AI organization. However, most organizations are new to AI. Databricks has heard from customers that they want to discover out-of-the-box AI models on Marketplace to help them kickstart their AI innovation journey.

To meet this demand, Databricks will be adding AI model sharing capabilities on Marketplace to provide users access to both OSS and proprietary AI (both first-and third-party) models. This will enable data consumers and providers to discover and monetize AI models and integrate AI into their data solutions.

Using this feature, data consumers can evaluate AI models with rich previews, including visualizations and pre-built notebooks with sample data. With Databricks Marketplace, there are no difficult data delivery mechanisms — you can get the AI models instantly with the click of a button. All of this works out-of-the-box with the AI capabilities of the Databricks Lakehouse Platform for both real-time and batch inference. For real-time inference, you can use model serving endpoints. For batch inference, you can invoke the models as functions directly from DBSQL or notebooks.

With AI model sharing, Databricks customers will have access to best-in-class models from leading providers, as well as OSS models published by Databricks which can be quickly and securely applied on top of their data. Databricks will curate and publish its own open source models across common use cases, such as instruction-following and text summarization, and optimize tuning or deployment of these models.

Using AI models from Databricks Marketplace can help your organization summarize complex information quickly and easily to help accelerate the pace of innovation.

Chapter 6
Share securely 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-safe environment where participants can bring their sensitive data, which might include personally identifiable information (PII), 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 any sensitive data.

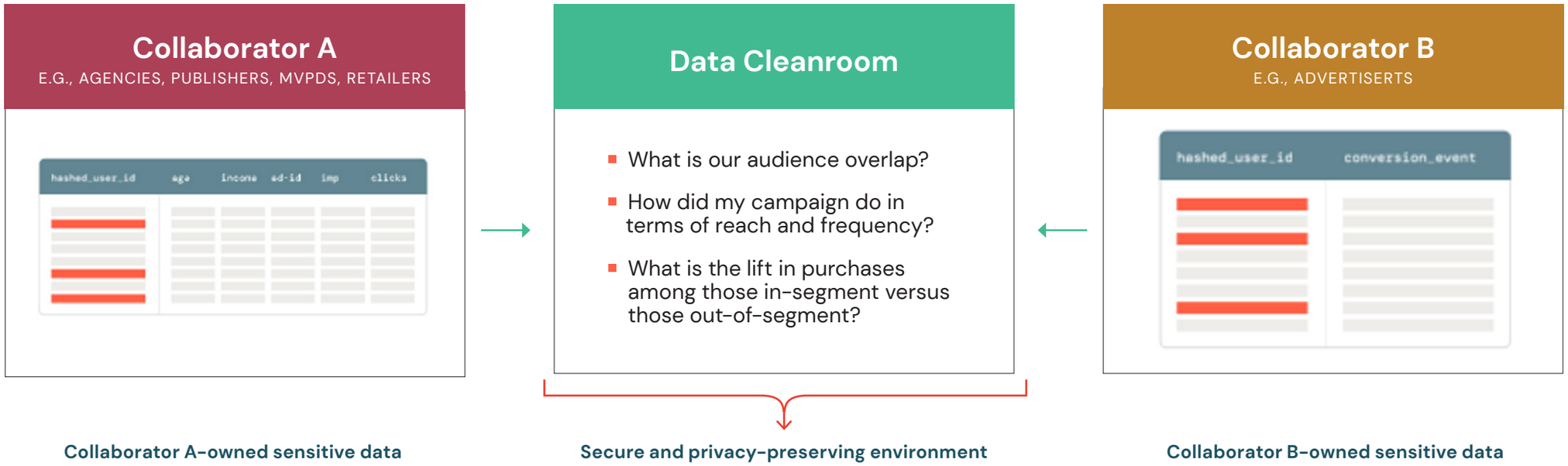


Figure 8: Data clean room diagram example for audience overlap analysis in advertising

A data clean room is not a new concept. Google introduced the idea in 2017 when it announced Ads Data Hub, which allows advertisers to gain impression-level insights about cross-device media campaigns in a more secure, privacy-safe environment. In the last few years, the demand for clean rooms has accelerated. IDC predicts that by 2024, 65% of G2000 enterprises will form data sharing partnerships with external stakeholders via data clean rooms to increase interdependence while safeguarding data privacy. There are various compelling needs driving this demand:

Privacy-first world. Stringent data privacy regulations such as GDPR and CCPA, along with sweeping changes in third-party measurement, have transformed how organizations collect, use and share data. For example, Apple's App Tracking Transparency Framework (ATT) provides users of Apple devices the freedom and flexibility to easily opt out of app tracking. Google also plans to phase out support for third-party cookies in Chrome by late 2024. As these privacy laws and practices evolve, the demand for data cleanrooms is likely to rise as the industry moves to new identifiers that are PII based, such as UID 2.0, and organizations try to find new ways to share and join data with customers and partners in a privacy-centric way.

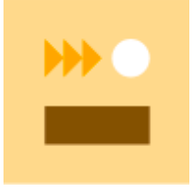
Collaboration in a fragmented ecosystem. Today, consumers have more options than ever before when it comes to where, when and how they engage with content. As a result, the digital footprint of consumers is fragmented across different platforms, necessitating that companies collaborate with their partners to create a unified view of their customers' needs and requirements. To facilitate collaboration across organizations, cleanrooms provide a secure and private way to combine their data with other data to unlock new insights or capabilities.

New ways to monetize data. Most organizations are looking to monetize their data in one form or another. With today's privacy laws, companies will try to find any possible advantages to monetize their data without the risk of breaking privacy rules. This creates an opportunity for data vendors or publishers to join data for big data analytics without having direct access to the data.

Common data clean room uses cases

Category management 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.



Audience overlap exploration for media and entertainment

By creating a clean room environment, media companies can securely share their audience data with advertisers or other media partners. This allows them to perform in-depth analysis and identify shared audience segments without directly accessing or exposing individual user information.

Real-world evidence (RWE) for healthcare

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

Know Your Customer (KYC) in banking

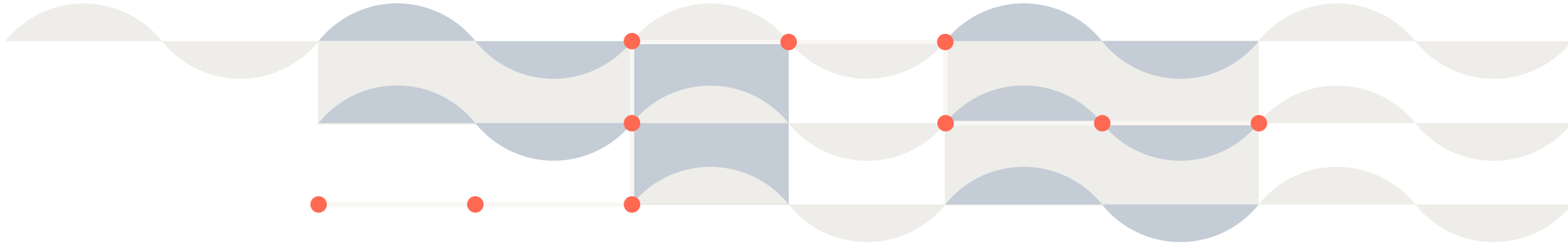
KYC standards are designed to combat financial fraud, money laundering and terrorism financing. Clean rooms can be used within a given jurisdiction to allow financial services companies to collaborate and run shared analytics to build a holistic view of a transaction for investigations.

Personalization with expanded interests for retailers

Retailers want to target consumers based on past purchases, as well as other purchases with different retailers. Clean rooms enable retailers to augment their knowledge of consumers to suggest new products and services that are relevant to the individual but have not yet been purchased.

5G data monetization for telecom

5G data monetization enables telecoms to capitalize on data from 5G networks. Clean rooms provide a secure environment for collaboration with trusted partners, ensuring privacy while maximizing data value for optimized services, personalized experiences and targeted advertising.



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, most existing solutions are not scalable to expand collaboration beyond a few collaborators at a time. For example, an advertiser might want to get a detailed view of their ad performance across different platforms, which requires analysis of the aggregated data from multiple data publishers. With collaboration limited to just a few participants, organizations get partial insights on one clean room platform and end up moving their data to another clean room vendor to aggregate the data, incurring the operational overhead of collating partial insights.

Finally, existing clean room solutions do not provide the flexibility to run arbitrary analysis and are mainly restricted to SQL, a subset of Python, and pre-defined templates. While SQL is absolutely needed for clean rooms, there are times when you require complex computations such as machine learning or integration with APIs where SQL doesn't satisfy the full depth of the technical requirements.

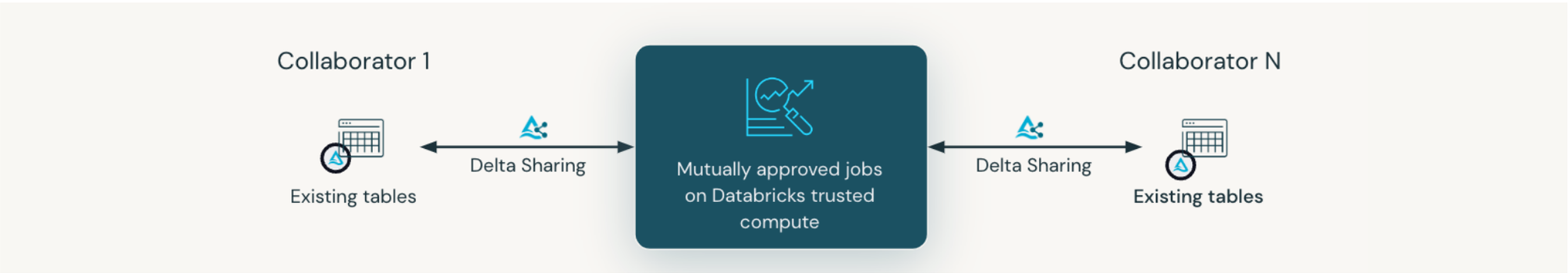
Key benefits of Databricks Clean Rooms

Databricks Clean Rooms allow businesses to easily collaborate with their customers and partners in a secure environment on any cloud in a privacy-safe way. Key benefits of Databricks Clean Rooms include:

Flexible - your language and workload of choice. Databricks Clean Rooms empower collaborators to share and join their existing data and run complex workloads in any language —Python, R, SQL, Java and Scala — on the data while maintaining data privacy. Beyond traditional SQL, users can run arbitrary workloads and languages, allowing them to train machine learning models, perform inference and utilize open-source or third-party privacy-enhancing technologies. This flexibility enables data scientists and analysts to achieve more comprehensive and advanced data analysis within the secure Clean Room environment.

Scalable, multi-party collaboration. With Databricks Clean Rooms, you can launch a clean room and work with multiple collaborators at a time. This capability enables real-time collaboration, fostering efficient and rapid results. Moreover, Databricks Clean Rooms seamlessly integrate with identity service providers, allowing users to leverage offerings from these providers during collaboration. The ability to collaborate with multiple parties and leverage identity services enhances the overall data collaboration experience within Databricks Clean Rooms.

Interoperable - any data source with no replication. Databricks Clean Rooms excel in interoperability, ensuring smooth collaboration across diverse environments. With Delta Sharing, collaborators can seamlessly work together across different cloud providers, regions and even data platforms without the need for extensive data movement. This eliminates data silos and enables organizations to leverage existing infrastructure and data ecosystems while maintaining the utmost security and compliance.

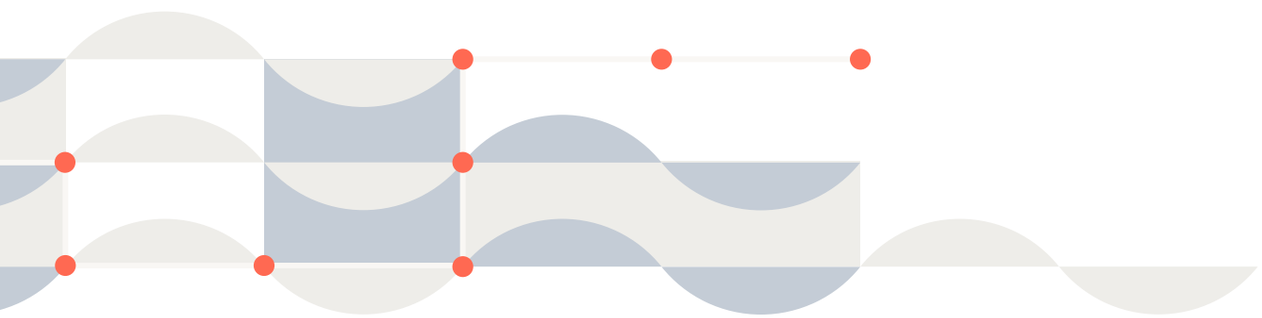


Resources

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.



Databricks offers an open approach to data sharing and collaboration with a variety of tools to:

- **Share across platforms:** You can share live data sets, as well as AI models, dashboards 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 with the ability 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 the usage of your data.

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

About the Authors

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

Milos Colic is a Senior Solution Architect at Databricks. His passion is to help customers with their data exchange and data monetization needs. Furthermore, he is passionate about geospatial data processing and ESG. You can reach Milos on LinkedIn.

Itai Weiss is a Lead Delta Sharing Specialist at Databricks and has over 20 years of experience in helping organizations of any size build data solutions. He focuses on data monetization and loves to help customers and businesses get more value from the data they have. You can reach Itai on LinkedIn.

Somasekar Natarajan (Som) is a Solution 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.

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 is passionate about helping organizations democratize data and AI with the Databricks Lakehouse Platform. You can reach Sachin on LinkedIn.

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

Giselle Goicochea is a Senior Product Marketing Manager on the Databricks Data Engineering and Analytics team. Her area of focus is data sharing and collaboration with Delta Sharing and Databricks Marketplace. 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 7,000 organizations worldwide — including Comcast, Condé Nast, H&M and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

Sign up for a free trial

databricks