

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

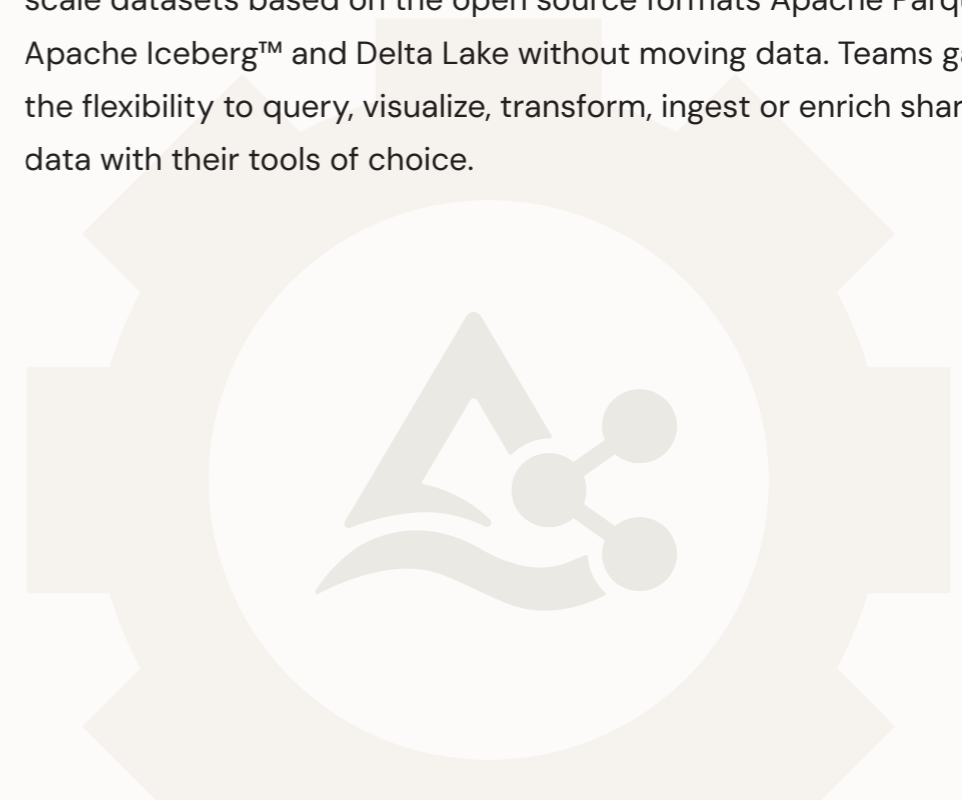
We believe the future of data and AI sharing should be characterized by open technology. Data and AI sharing shouldn't be tied to a proprietary technology that introduces unnecessary limitations and financial burdens to the process. It should be readily available to anyone who wants to share data at 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.



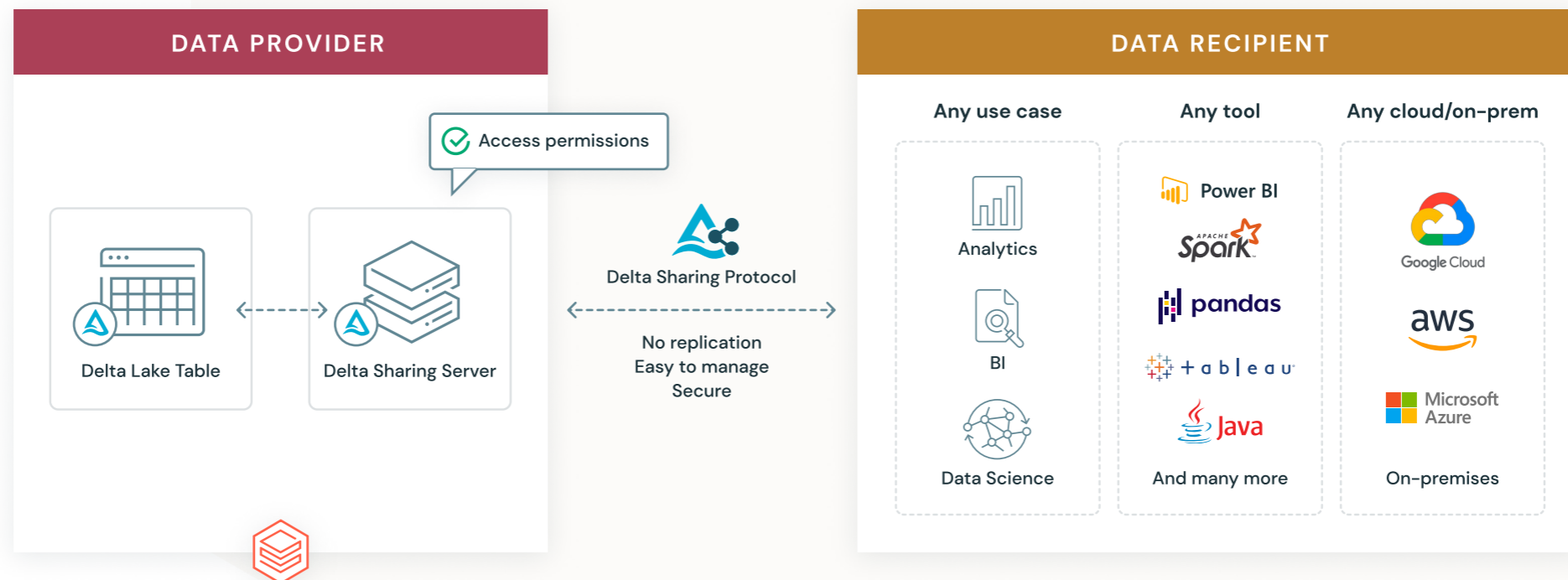


Figure 3:
Delta Sharing

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