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.

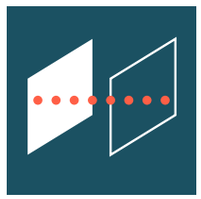
Chapter 1
What Is Data Sharing and Why Is It Important?

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.

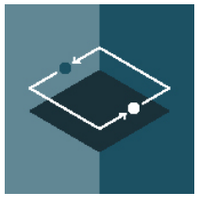


Peer-to-peer sharing

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.

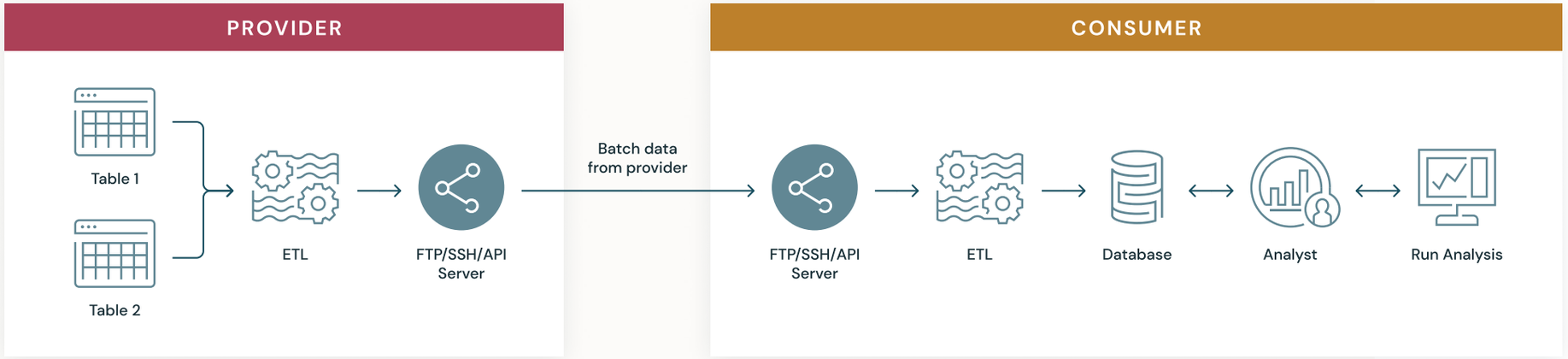


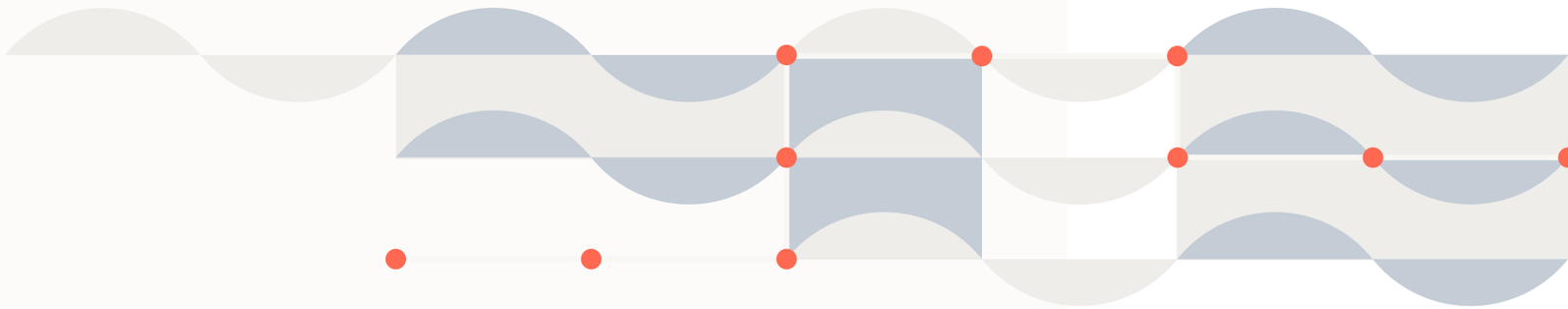
Figure 1: Legacy data sharing solutions

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.

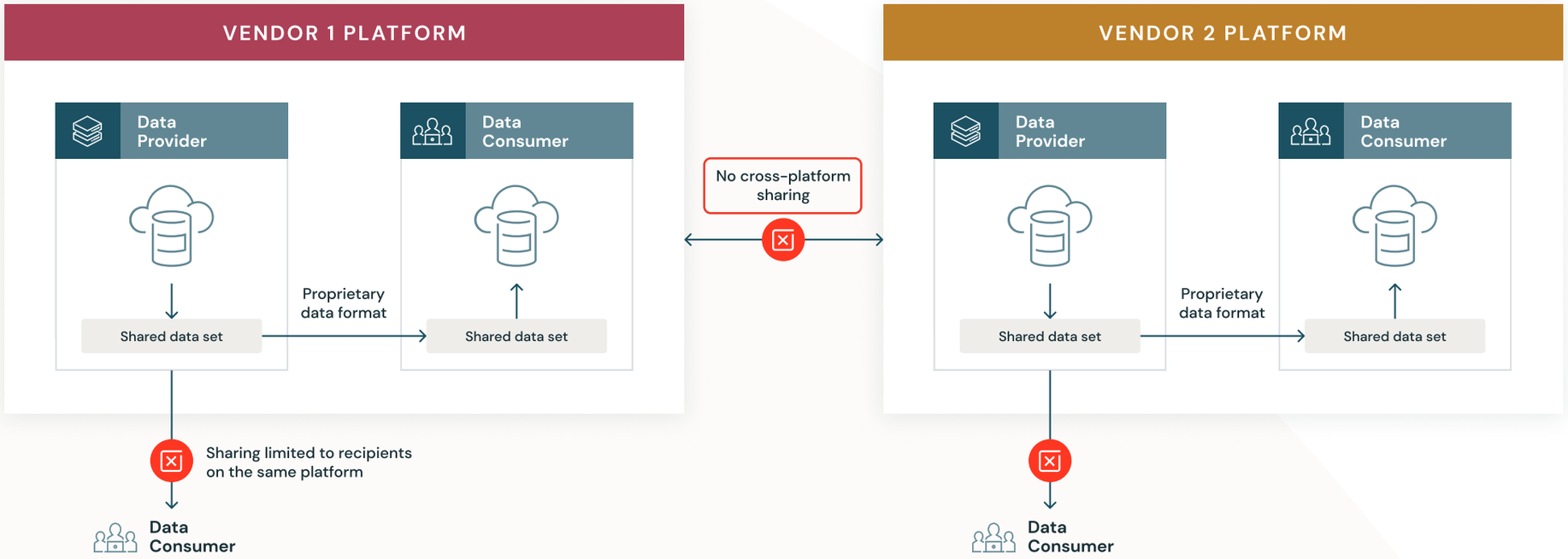


Figure 2: Proprietary vendor solutions

Pros

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