

# Sharing in the Lakehouse

ORGANIZED BY  databricks



**Celia Kung**

Engineering Manager,  
Databricks



**Jay Bhankharia**

Sr. Director Data  
Partnerships, Databricks



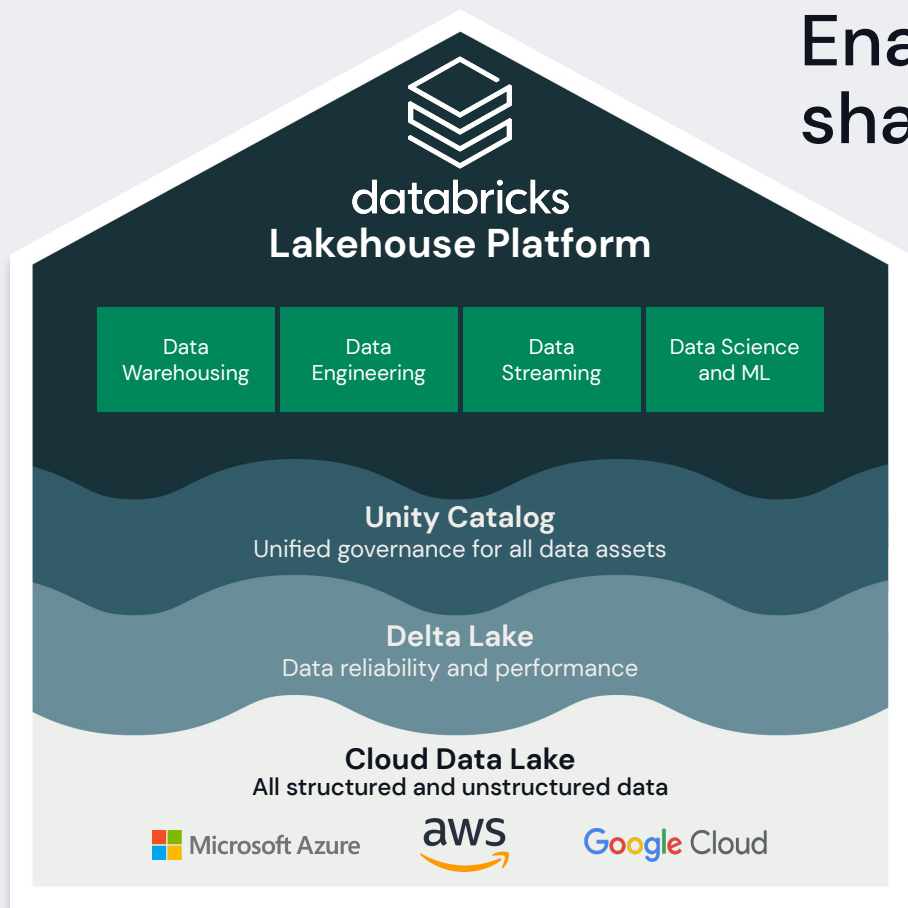
**Itai Weiss**

Lead Data Partner SA,  
Databricks

# Product Safe Harbor Statement

This information is provided to outline Databricks' general product direction and is for informational purposes only. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all.

# Enable organizations to safely share and collaborate on data



## Delta Sharing

Secure sharing from your Lakehouse with no replication



## Marketplace

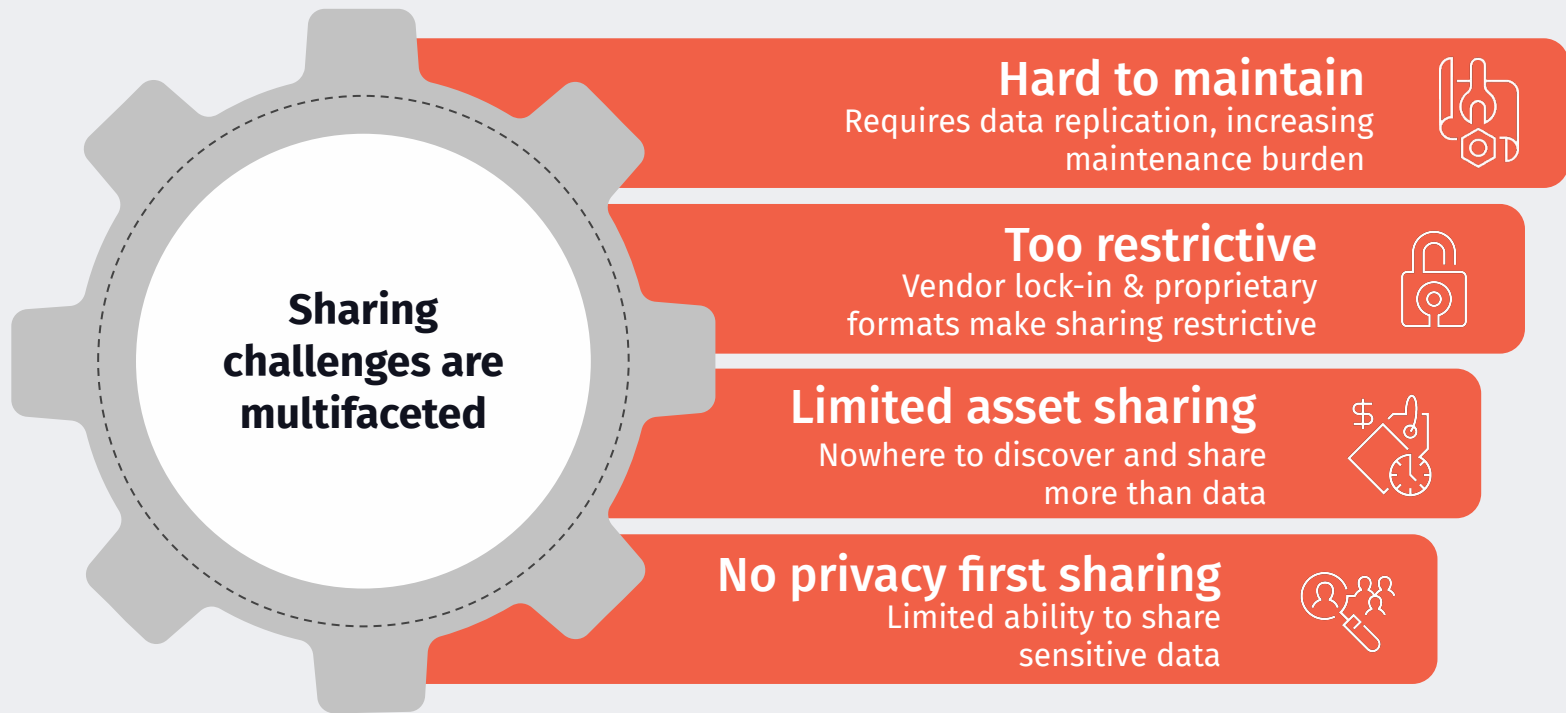
Discovery and access data products



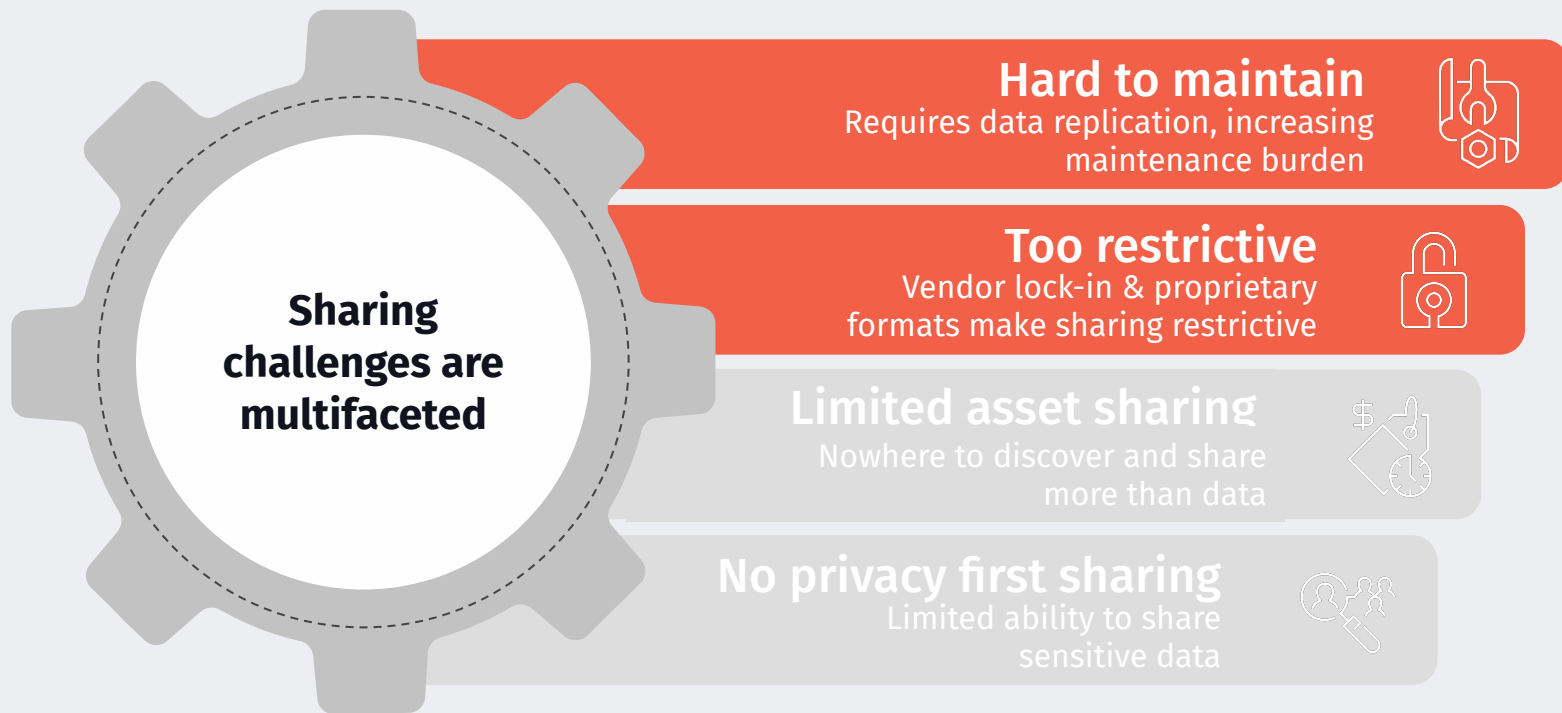
## Cleanroom

Privacy safe compute with multi-party code approval

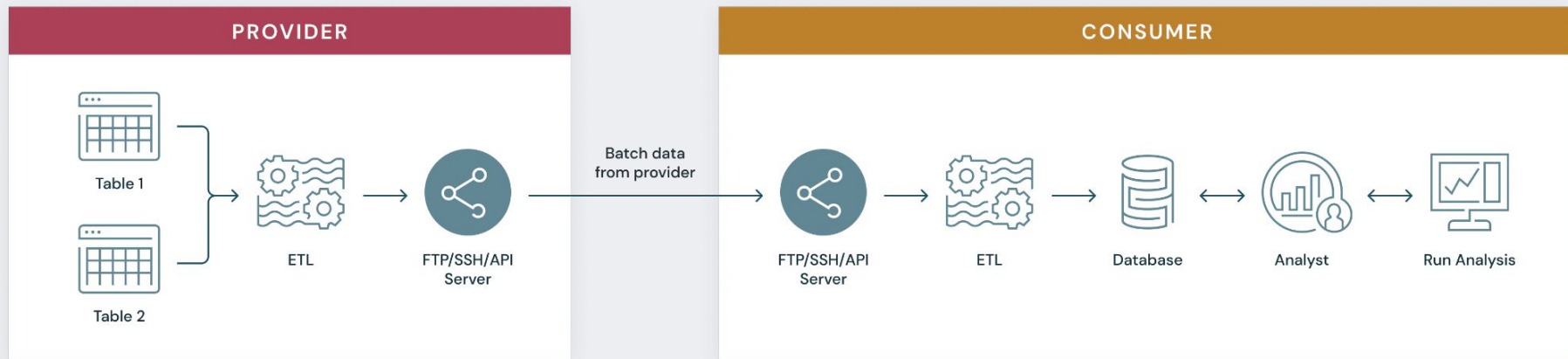
# Most organizations struggle with data sharing



# Traditional data sharing challenges



# Existing sharing solutions are hard to maintain



Data  
Replication



Complex to  
maintain and scale

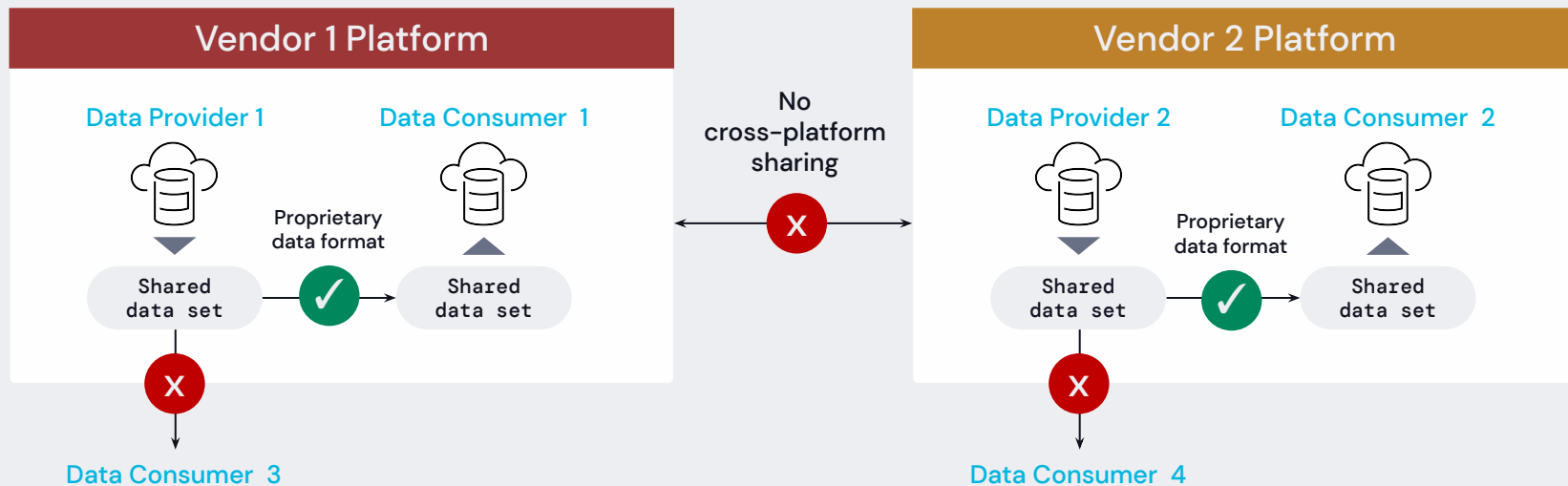


Days to process  
data



Out of sync  
Data

# Existing sharing solutions are too restrictive



Vendor lock-in

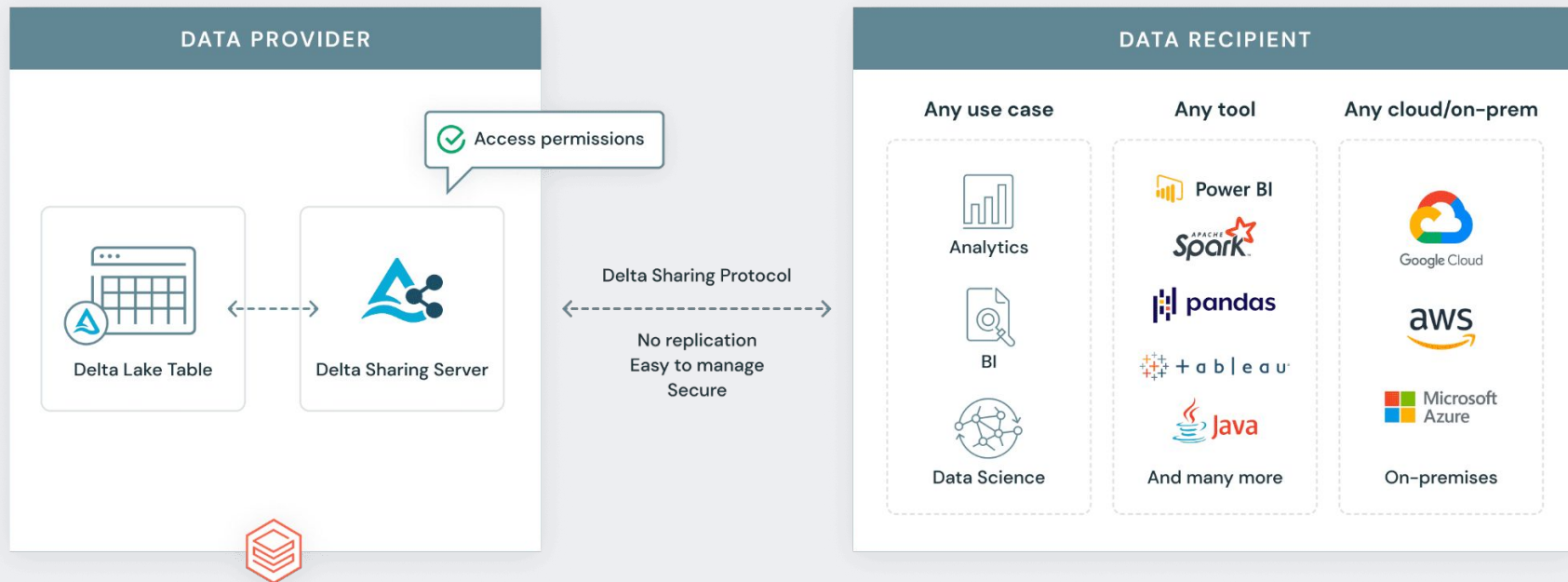


Expensive



Limited data sharing

# Delta Sharing: An open standard for secure sharing of data assets



Share live data with no replication

Centralized governance

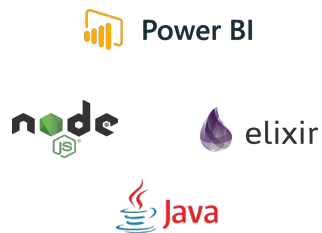
Open cross-platform sharing

# Delta Sharing connectors

## Original Launch



## Added This Year

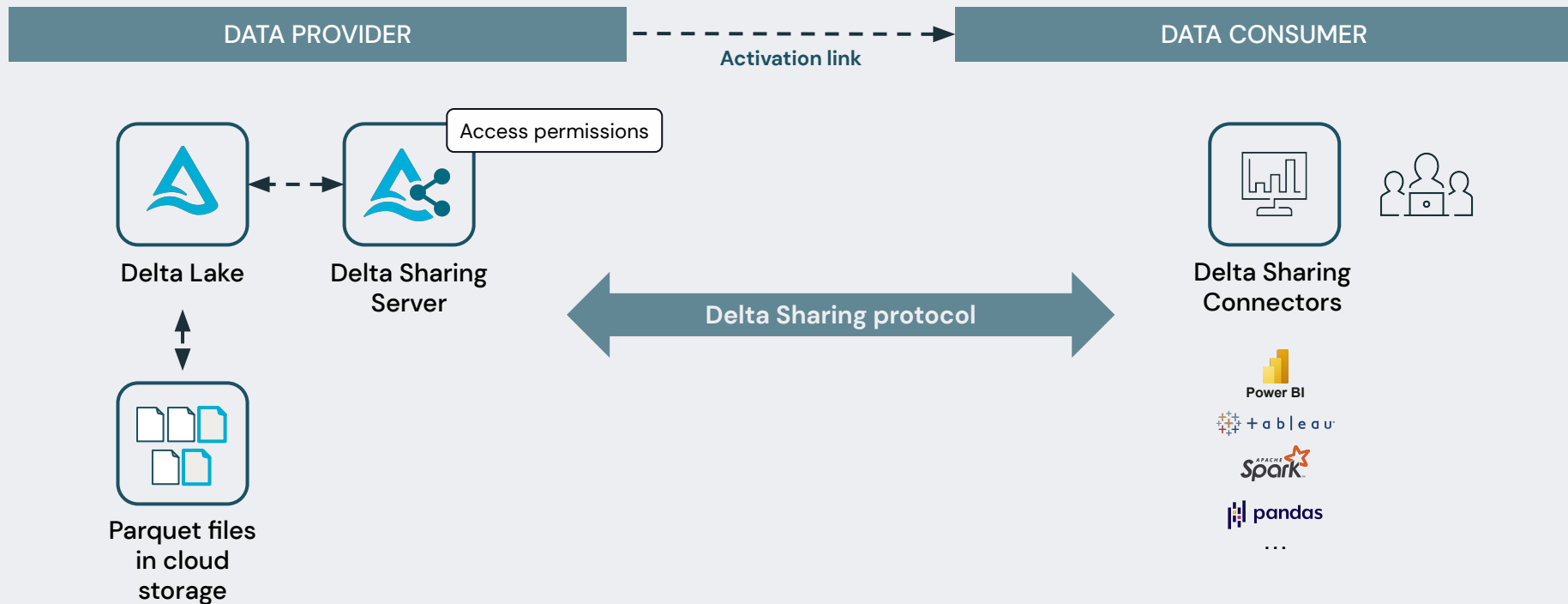


## Coming Soon

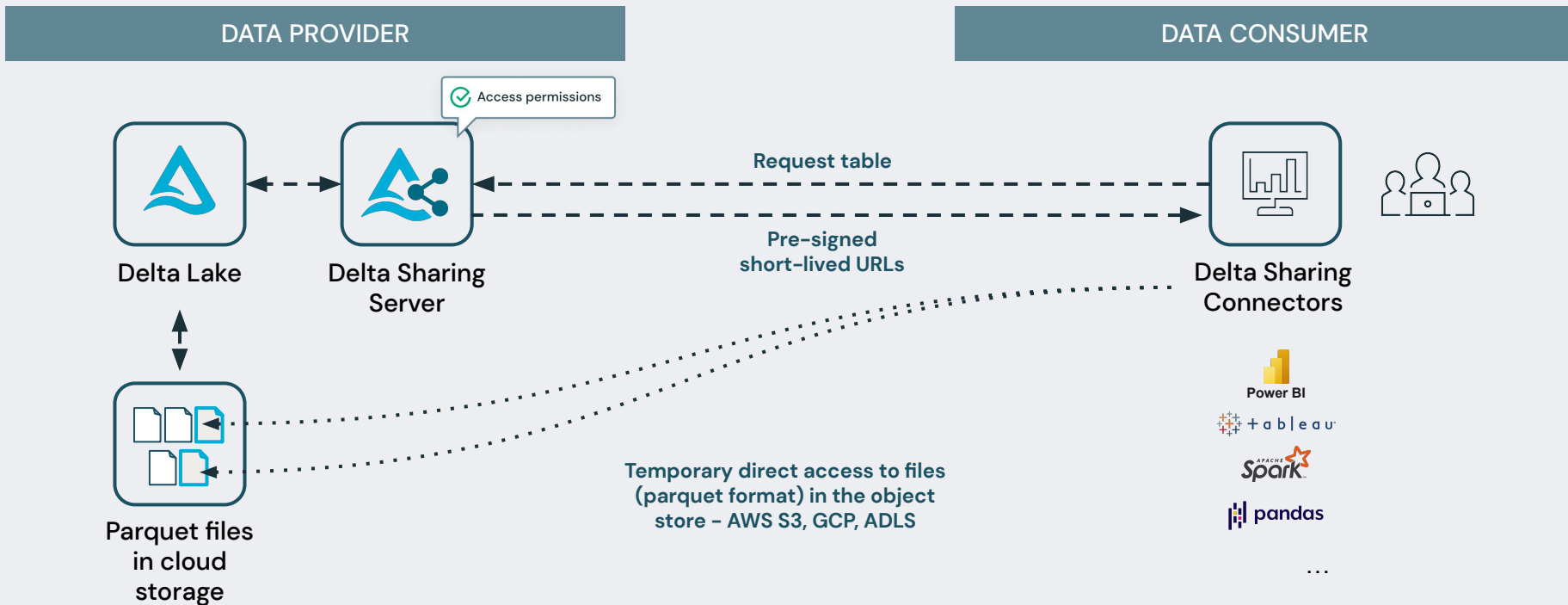


And others...

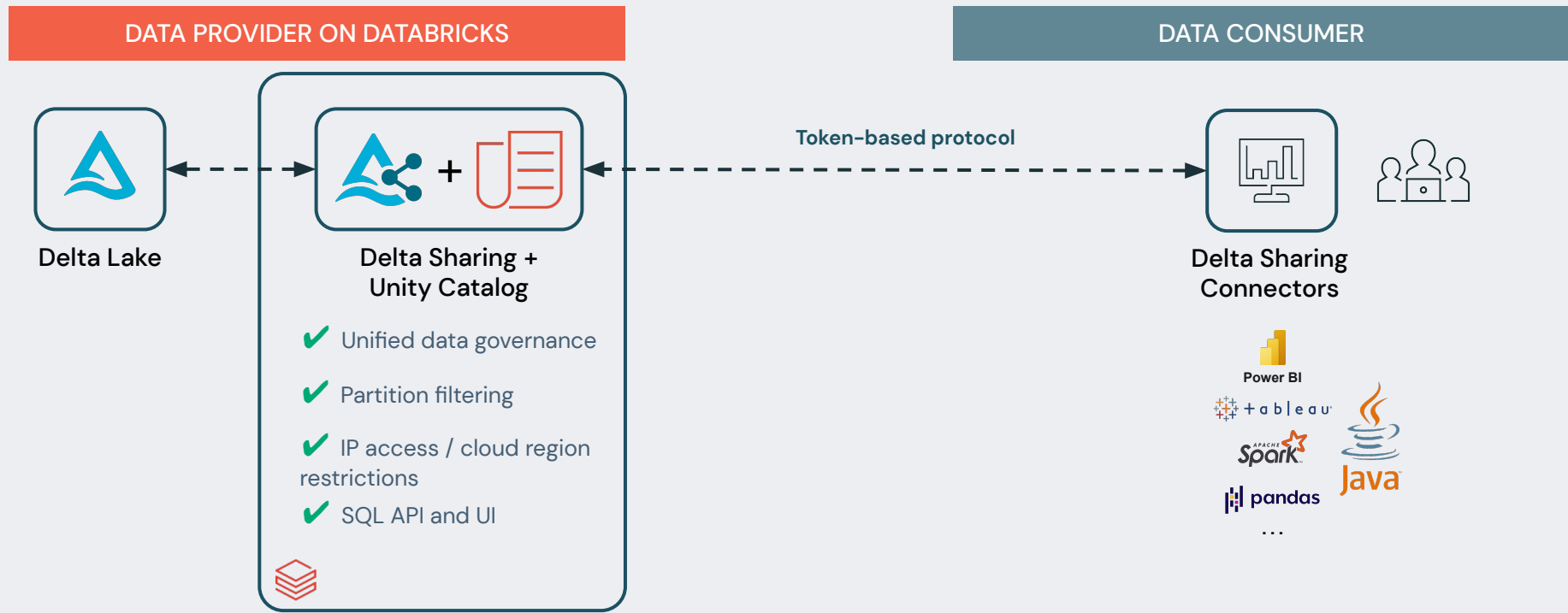
# How does it work?



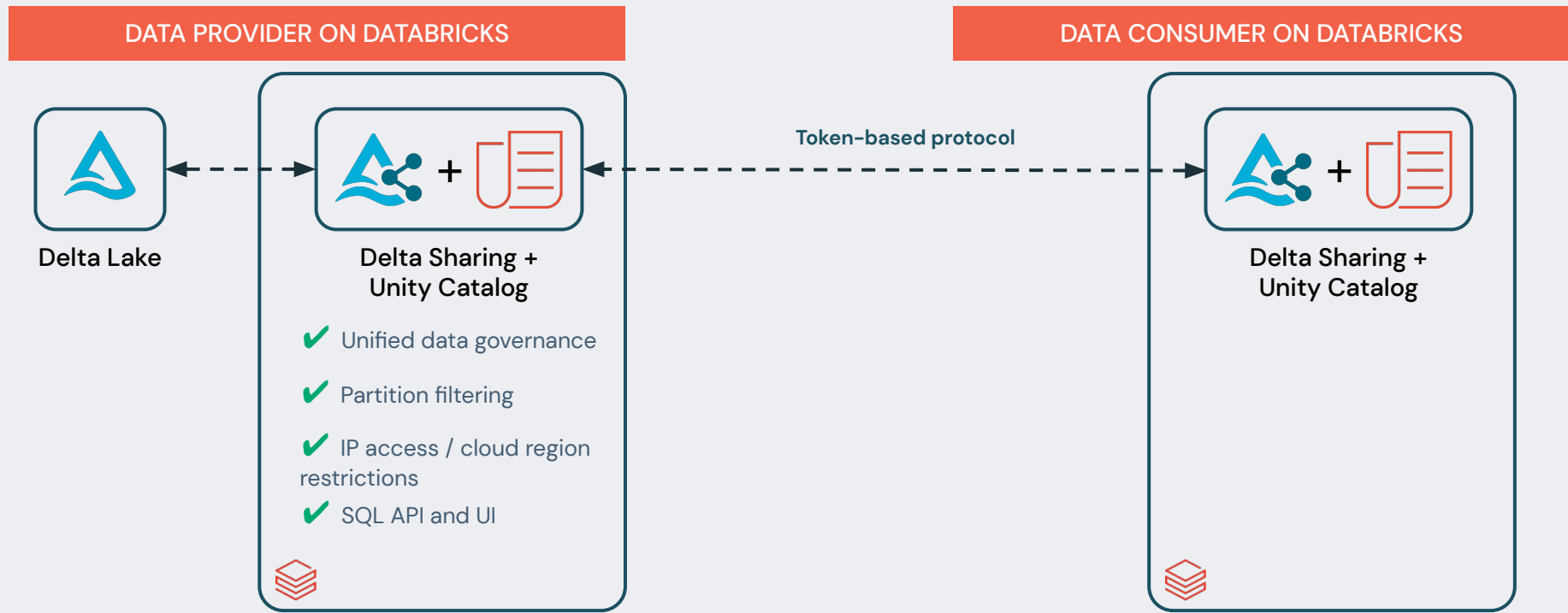
# Under the hood



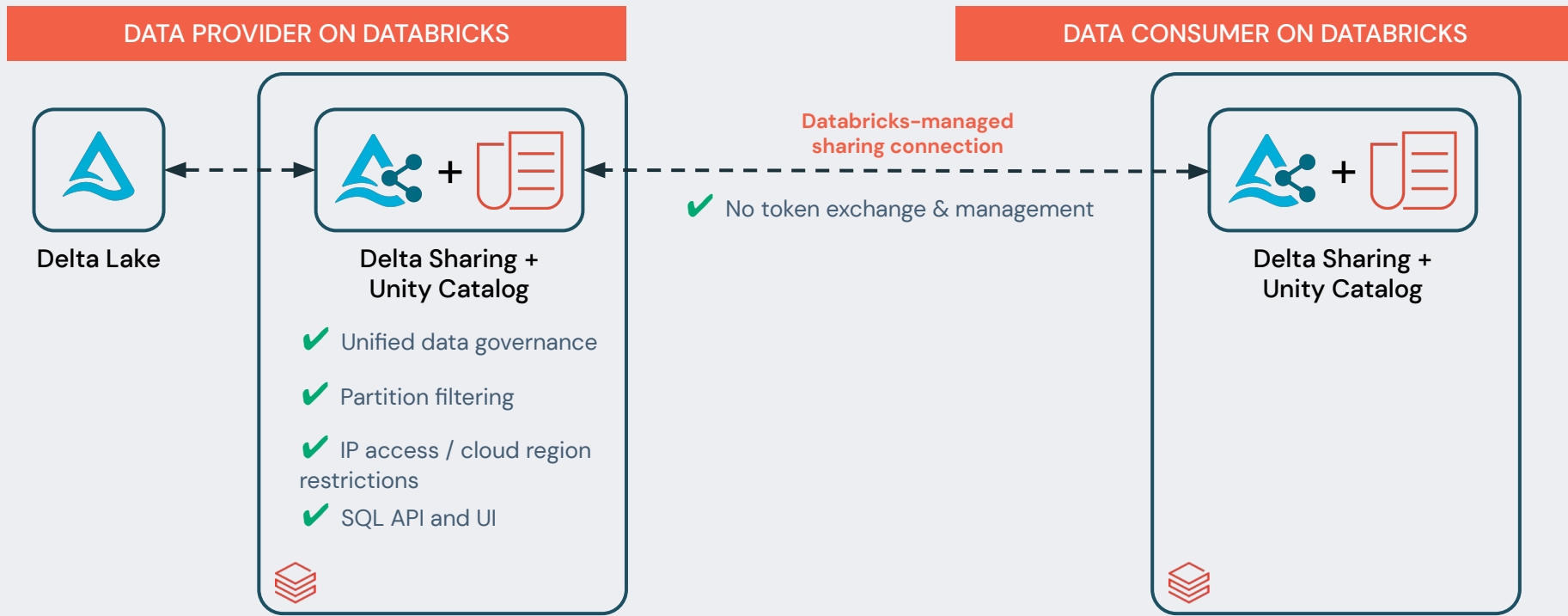
# Delta Sharing on Databricks



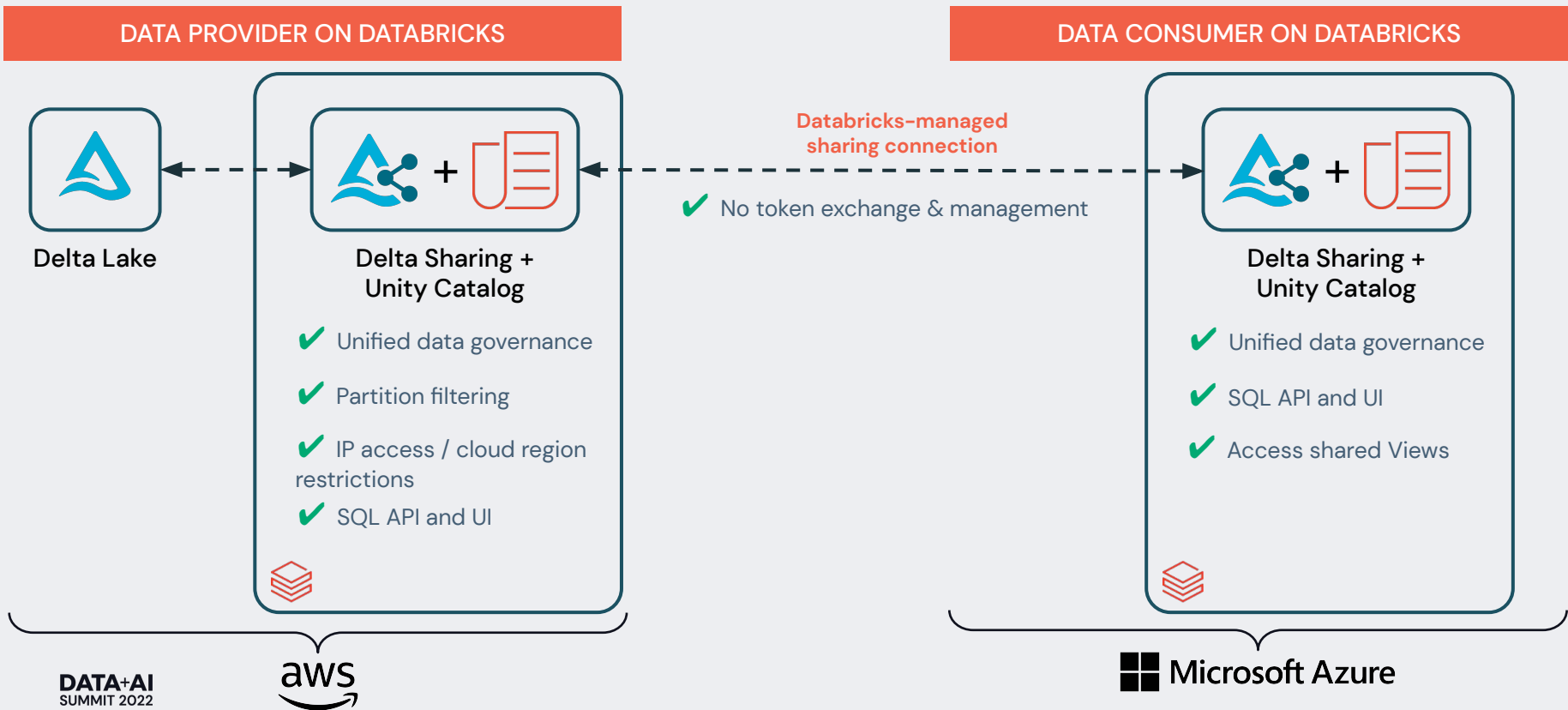
# Delta Sharing on Databricks



# Delta Sharing on Databricks



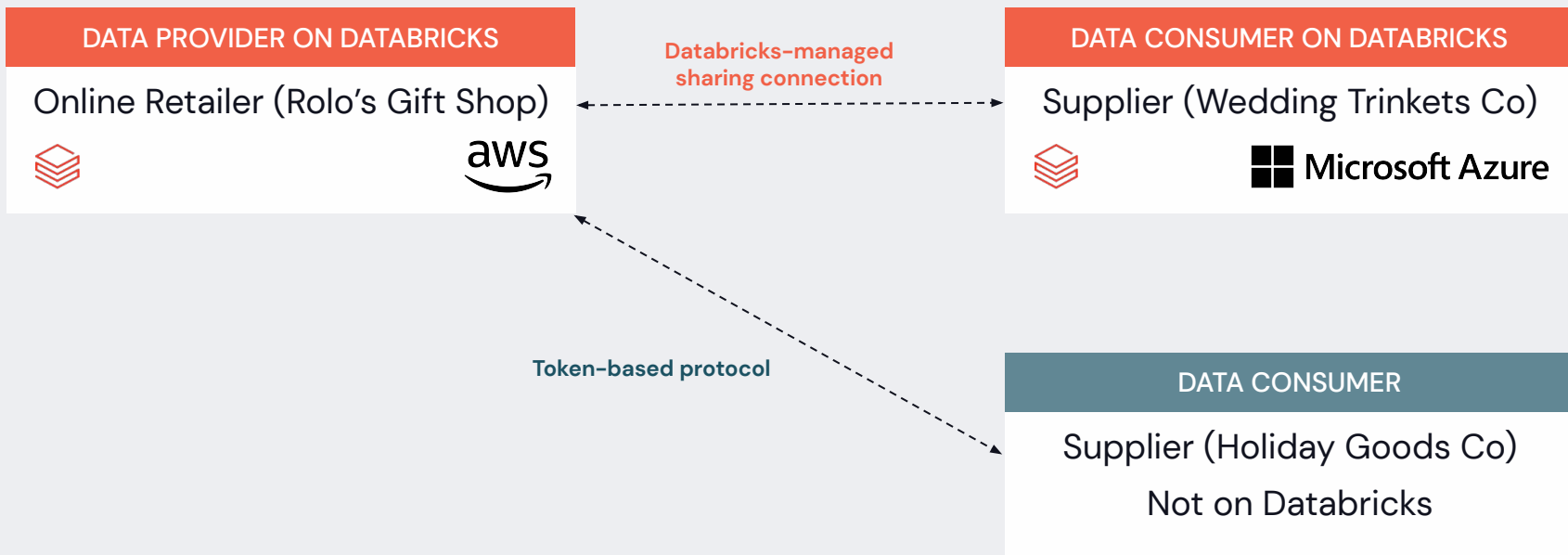
# Delta Sharing on Databricks



# Delta Sharing demo

# Demo Scenario

Retailer sharing its online sales data with its suppliers



\*Company names and use cases are purely fictional



▼

itai.weiss@databricks.com

# GA in the next few weeks!

Databricks-to-Databricks sharing

Data Explorer UI

Sharing incremental changes (Change Data Feed)

```
ALTER SHARE my_share ADD TABLE my_table WITH CHANGE DATA FEED
```

# Post-GA

## View sharing

Share views with fine-grained data filters

## Stream sharing

Share data streams and enable consumers to process continuously

## Share other types of data assets

Share ML models, notebooks, etc.

# Marketplace challenges



**Sharing  
challenges are  
multifaceted**

## Hard to maintain

Requires data replication, increasing maintenance burden



## Too restrictive

Vendor lock-in & proprietary formats make sharing restrictive



## Limited asset sharing

Nowhere to discover and share more than data



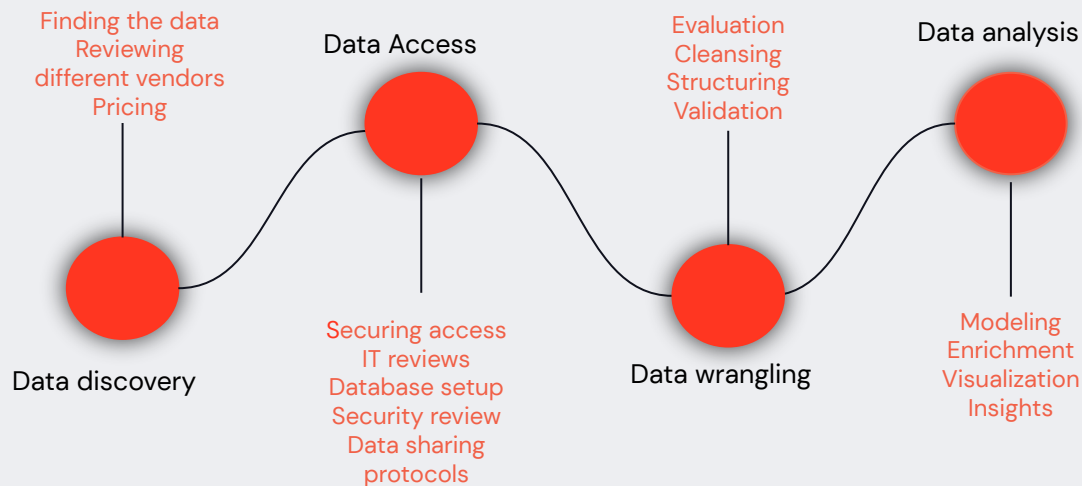
## No privacy first sharing

Limited ability to share sensitive data

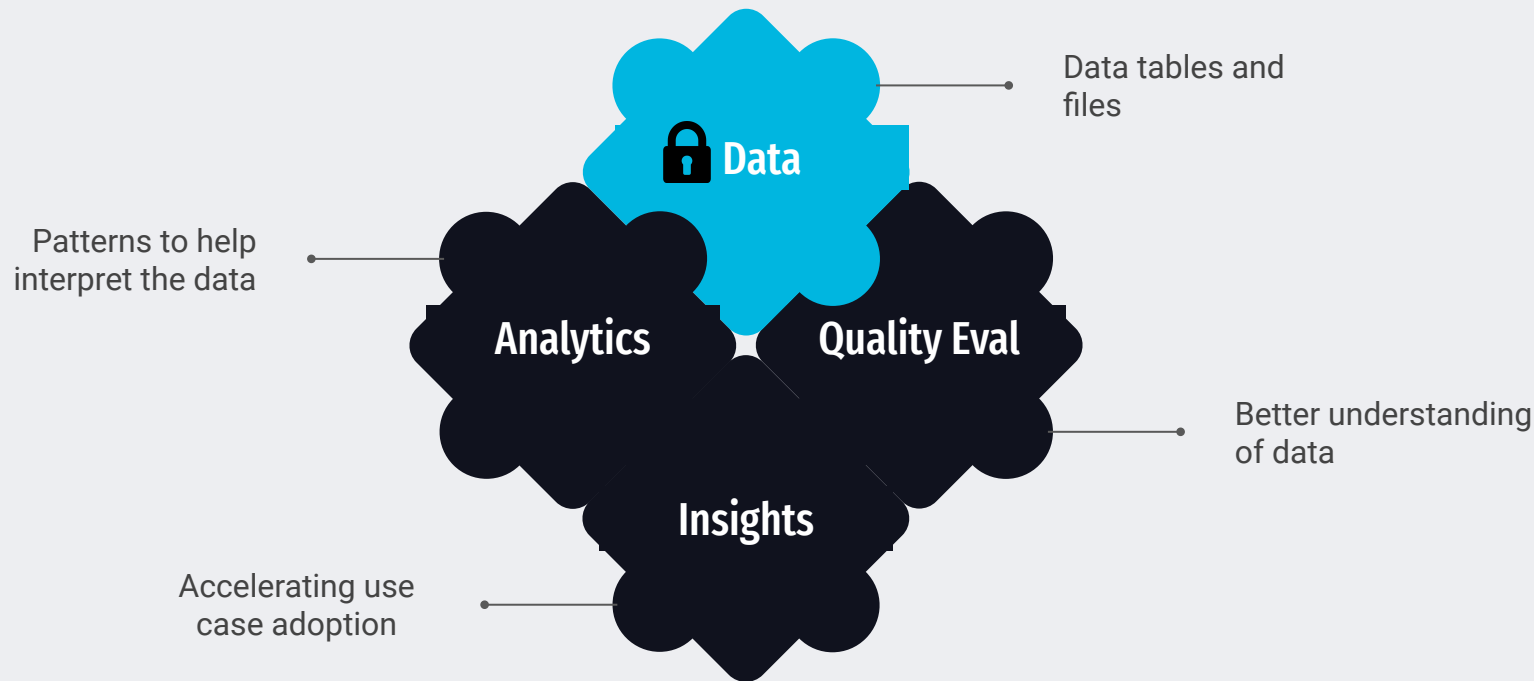


# Data consumers are frustrated

Time to value can be lengthy with 3rd party data



# “Data Marketplaces” are restrictive and not providing the full picture



# Introducing Databricks Marketplace

More than just data...



An open marketplace to exchange data products including datasets, notebooks, dashboards, and ML models all **powered by Delta Sharing**

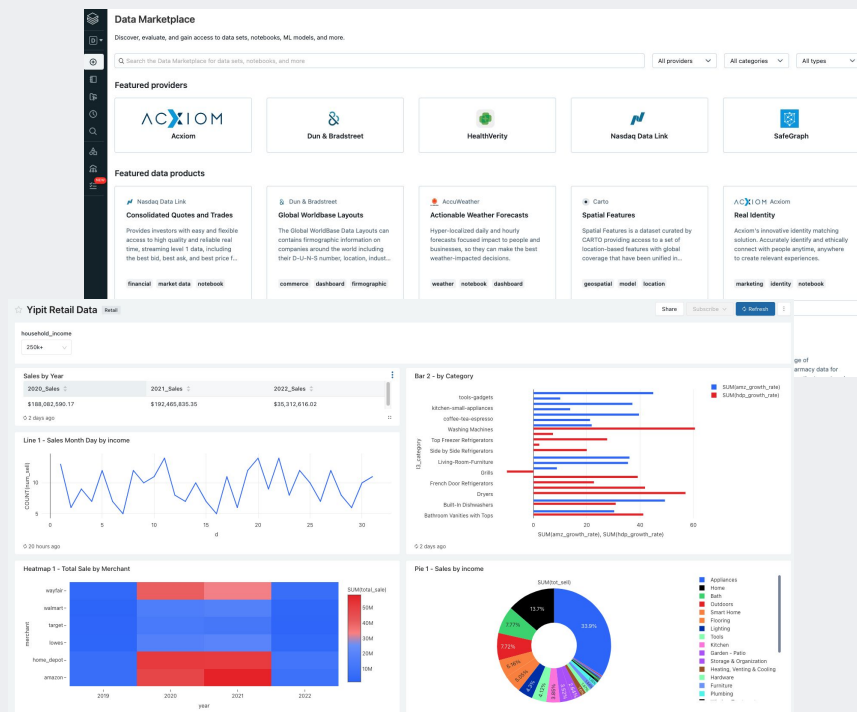
# Data Consumers Benefits

Turn data into insights quicker

Open and seamless access  
to data products

Frictionless evaluation

Accelerated insights



# Data Providers Benefits

Accelerate your growth

## Acquisition



Attract new customers

## Expansion



Reach new customers in current accounts

## Retention



Better experiences for current customers

Distribute and monetize **all your data assets**

**One share to anywhere.** No lock in to just one platform

**8,000+** and growing Databricks customer base

# What Data Providers are saying...

1

## Analytics on top of the data

"Customers need solutions, not only raw data. Being able to package raw data along with the code and analytics on top of it is how we see customers consuming raw data in the future"



SAFEGRAPH

2

## Expanded reach

"Databricks Marketplace is a compelling platform for us. We like the fact that it is open and provides us a way to reach existing and new types of personas for our data offerings. We see the platform as a key enabler to accelerate value with our data offerings to our customers"



LexisNexis

3

## Open sharing standards

"We are extremely excited to be part of the inception of the Databricks Marketplace. A marketplace built on their Delta Share protocol is a huge step forward in democratizing and simplifying data access."



Facteus

# Marketplace demo

ORGANIZED BY  databricks

June 27-30, 2022

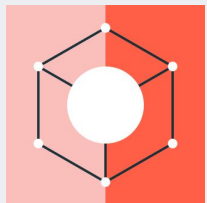
DATA + AI SUMMIT 2022

# Building the modern data stack with Lakehouse

Register now

# Join the Marketplace as a partner

Databricks helps our Data Provider Partners monetize data assets to a large, open ecosystem of data consumers all from a single platform.



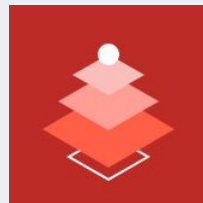
**Reach more  
consumers**



**Better  
customer  
experience**



**Industry  
solutions**



**GTM  
support**



**Technology  
for data  
products**

# Cleanroom challenges



**Sharing  
challenges are  
multifaceted**

## Hard to maintain

Requires data replication, increasing maintenance burden



## Too restrictive

Vendor lock-in & proprietary formats make sharing restrictive



## Limited asset sharing

Nowhere to discover and share more than data



## No privacy first sharing

Limited ability to share sensitive data



# Compelling events are driving this demand



Shift to privacy-first



Fragmented ecosystem



New ways to leverage  
data & IP

## Compelling events:

- **GDPR & CCPA**
- **Apple IDFA & App Tracking Transparency Framework (ATT)**
  - ~75% opt-out rate ("Ask App Not to Track")
- **Apple Mail Protection Privacy (MPP)**
- **Cookieless world**
  - Google will phase out support of 3P cookies in Chrome in 2023. Chrome owns ~42% of the market)

## Impact

- Reduced targeting capabilities
- Emergence of **new identifiers** (e.g. UID2) that backed by PII
- Increased focus on **harvesting PII** (publishers/retailers)
- New challenges with measuring **marketing attribution**

# The cleanroom approach

## Collaborator A

e.g. Agencies, Publishers, MVPDs, Retailers

hashed_user_id	age	income	ad_id	imp	clicks

**Collaborator A owned  
sensitive data**

## Data Cleanroom

- What is our audience overlap?
- How did my campaign do in terms of reach and frequency?
- What is the lift in purchases among those in-segment versus those out-of-segment?

**Secure and privacy-preserving  
environment**

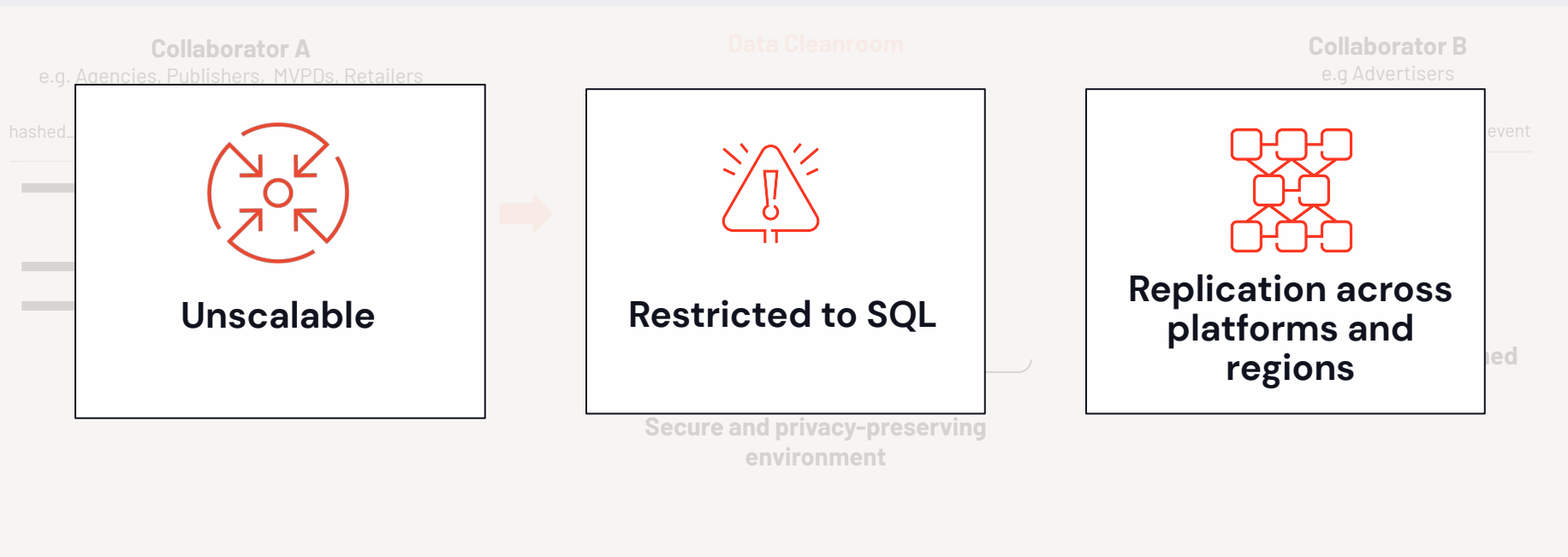
## Collaborator B

e.g Advertisers

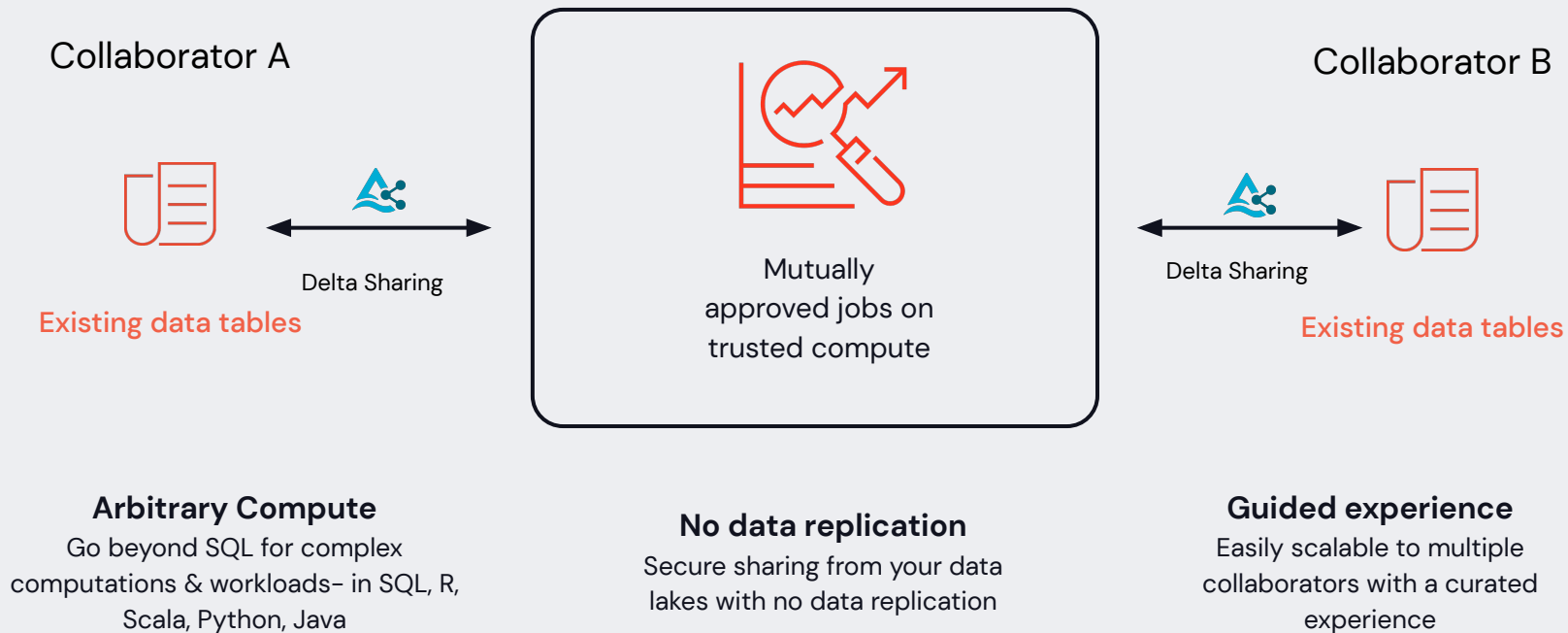
hashed_user_id	conversion_event

**Collaborator B owned  
sensitive data**

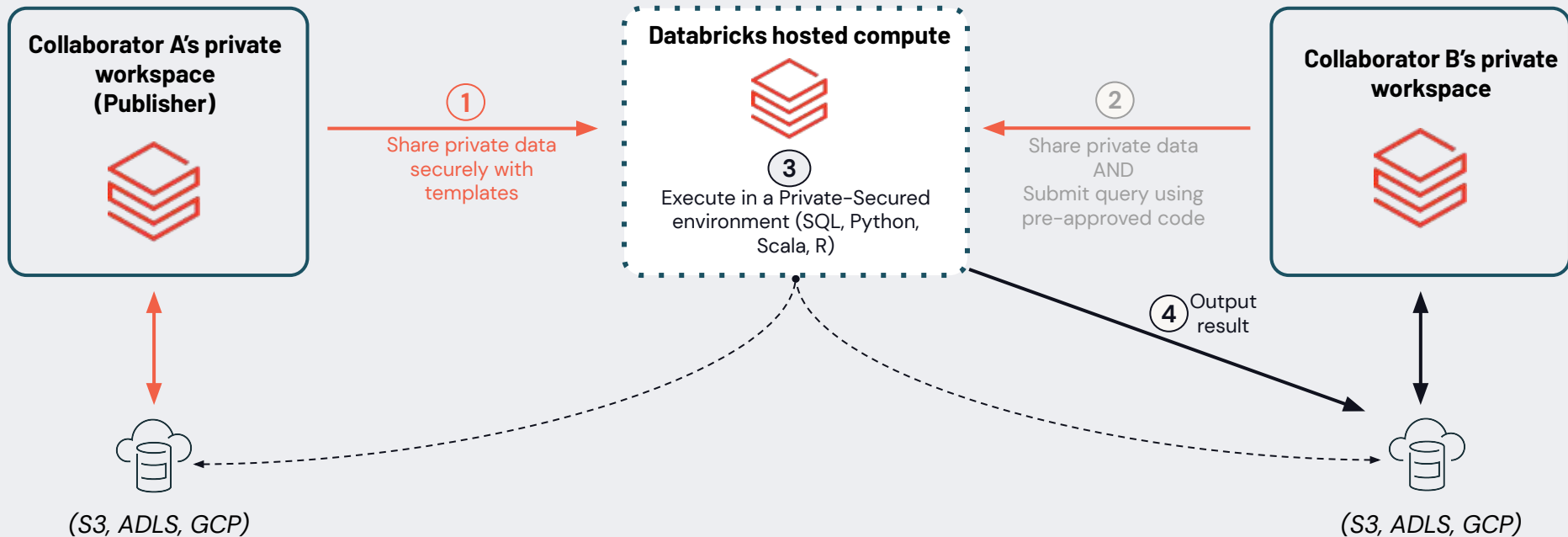
# Current solutions come with big drawbacks



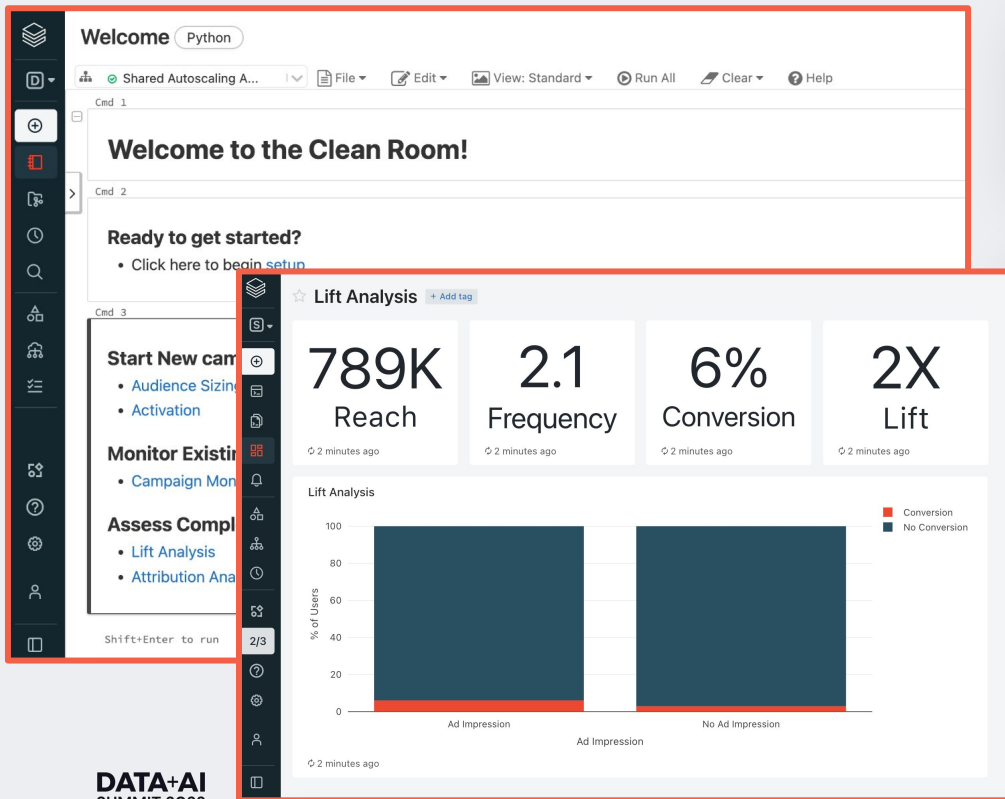
# Databricks cleanroom: any compute, at scale



# Current Databricks cleanroom data flow



# Databricks cleanroom benefits



## See the entire picture

Multi party collaboration

Leverage 3rd party & open source libraries

## Improved tools for insight

Open standards

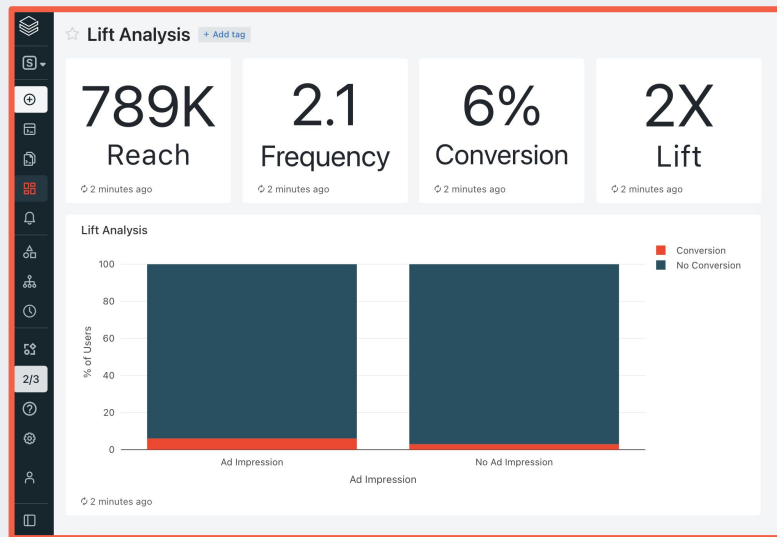
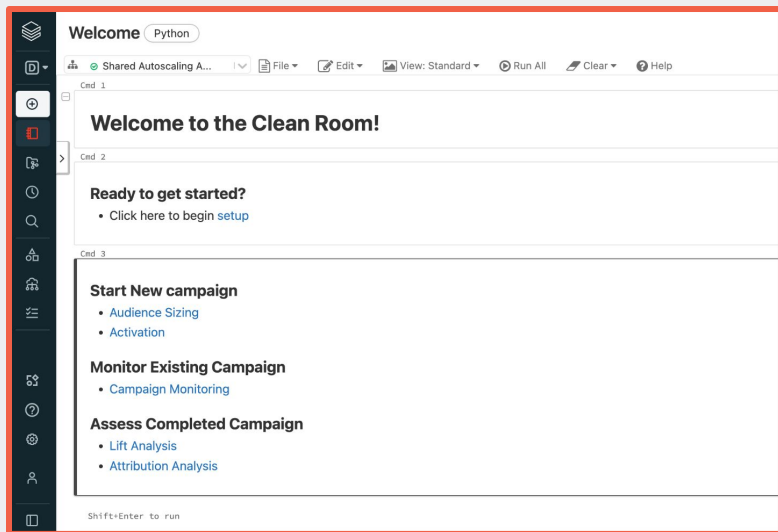
Multi languages + ML

## Faster Insights

No data movement

Enable near-instant time to value

# Databricks cleanroom benefits



## Faster Insights

No data movement

Enable near-instant time to value



## Improved tools for insight

Open standards  
Multi languages + ML



## See the entire picture

Multi party collaboration

Leverage 3rd party & open source libraries

# Databricks Cleanroom Brought to Life

Extending the Cleanroom for Identity Matching with Acxiom Data

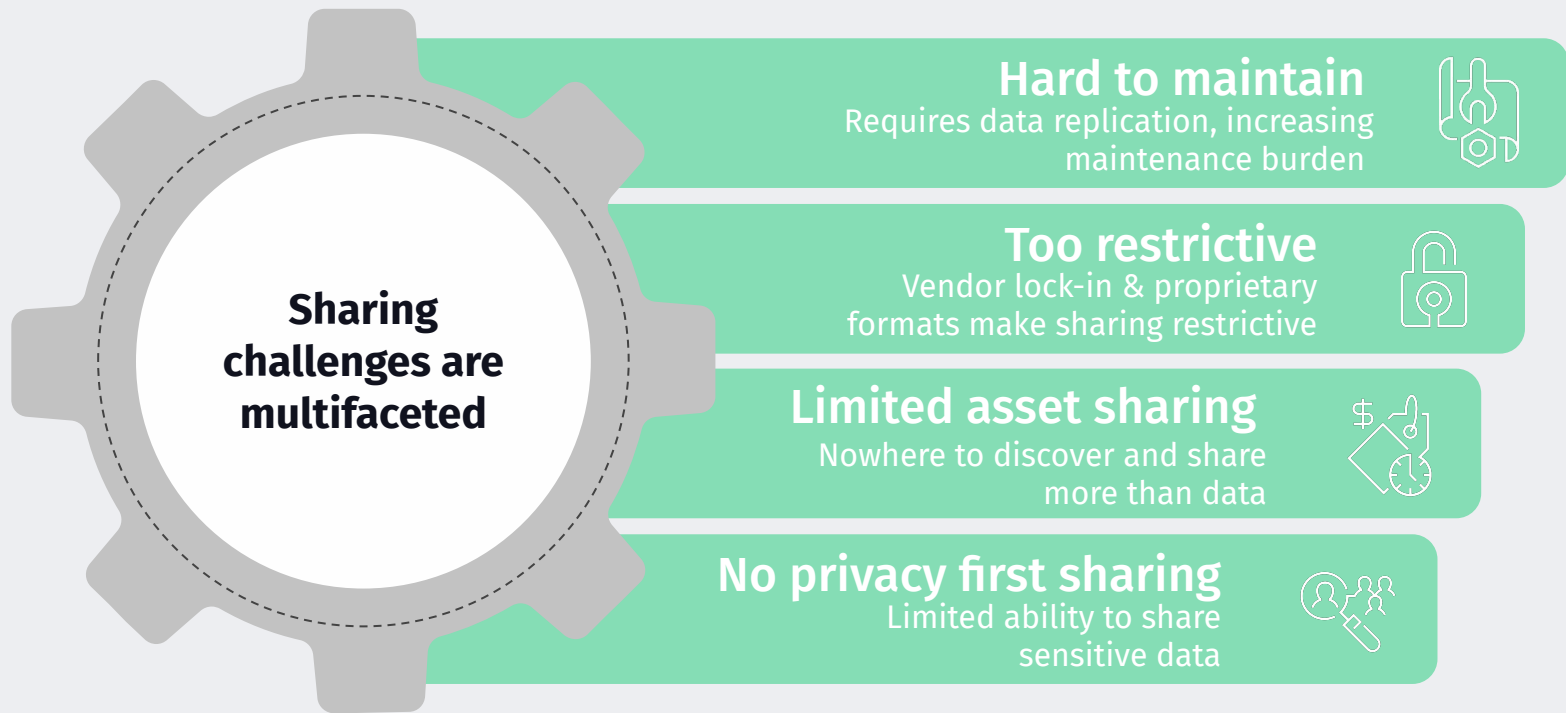


Acxiom enables a simple solution to **join different datasets based on identity resolution** within Databricks cleanroom to provide a comprehensive/complete picture of the household or segment

Precise   Persistent   Ethically sourced

Multi-sourced   Evidence-based

# Most organizations struggle with data sharing



# Roadmap Features

## Current

- Token-based Delta Sharing
- Token management
- Partition filtering
- Audit logs
- IP allow lists

## GA

- Databricks-managed Delta Sharing
- Cloud region restrictions
- Data Explorer UI
- Change Data Feed

## Post-GA

- Views
- Streaming
- Arbitrary File Sharing
- Support for Cleanrooms & Marketplace

# Roadmap

## Databricks Marketplace

- Open discovery
- Quickstart Assets
- Private Exchange

## Delta Sharing

- Views
- Streaming
- Arbitrary File Sharing

## Data Cleanrooms

- SQL, Python & ML
- Code approval
- Trusted compute

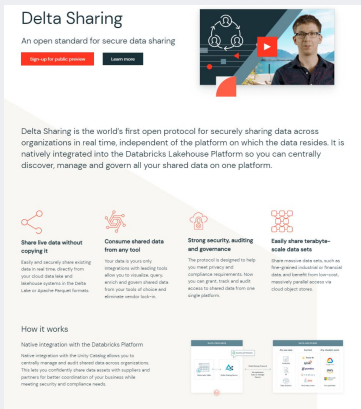
# What's next

**Contact your Databricks account executive!**

**Come to the Delta Sharing or CM&E Booth in the expo**

**Review past webinars for HLS and FSI**

**Look out for our eBook, coming up in the next few weeks**



**Delta Sharing**  
An open standard for secure data sharing

[Sign up for public preview](#) [Learn more](#)

Delta Sharing is the world's first open protocol for securely sharing data across organizations in real time, independent of the platform on which the data resides. It is natively integrated into the Databricks Lakehouse Platform so you can centrally discover, manage and govern all your shared data on one platform.

- Share live data without copying it**  
Easily and securely share existing data in real time directly from your cloud data lake and maintain existing data lakes. Lake or Apache Hadoop format.
- Consume shared data from any tool**  
Your data is yours only. Integrations with leading tools allow you to maintain, query, audit and govern shared data from your tools of choice and without vendor lock-in.
- Strong security, auditing and governance**  
The protocol is designed to help you meet privacy and compliance requirements. How you use data is tracked and audited across all shared data from one single platform.
- Easily share terabyte-scale data sets**  
Share massive data sets, such as large-scale scientific research data, and benefit from low-cost, managed parallel storage on cloud object stores.

**How it works**  
Native integration with the Databricks Platform  
Native integration with the Databricks Platform allows you to centrally manage and audit shared data across organizations. This lets you confidently share data across with suppliers and partners for better collaboration of your business while meeting security and compliance needs.

<https://databricks.com/product/delta-sharing>

# Thank you



**Celia Kung**

Engineering Manager,  
Databricks



**Jay Bhankharia**

Sr Director Data  
Partnerships, Databricks



**Itai Weiss**

Lead Data Partner SA,  
Databricks