DATA+AI SUMMIT 2022

Architecting the Future of Digital Fan Engagement

How Pumpjack Dataworks Became the Leading Fan Engagement Platform through Data Sharing

Corey Zwart
Head of Engineering, Pumpjack Dataworks

Itai Weiss
Data Partner SA, Databricks

Steve Touw
CTO, Immuta

ORGANIZED BY databricks

Agenda

- Data sharing: It’s complicated
- Immuta and access control
- Delta protocol for sharing data
- Pumpjack Dataworks Fan Engagement Platform
- Demo
- Questions

Challenges with data sharing

It's complicated!

Granularity

- Your data is your IP!
- Data use agreements
- Filtering data for privacy
- Filtering data for billing purposes
- Capturing audit for billing

"Live"

- Don't want copies floating around
- Consumers demand the latest version of your data

Multi-party

- Need to join source data from multiple contributors

To solve for these you need

Granularity

- Data use agreements
- Filtering data for privacy
- Filtering data for billing purposes
- Capturing audit for billing

Access control framework

“Live”

- Don’t want copies floating around
- Consumers demand the latest version of your data

Sharing Protocol

Multi-party

- Need to join source data from multiple contributors

Fine-grained access control and anonymization

Fine-grained access control

Get the right data into the right hands with Immuta

Data analyst

Purpose

Data anonymization

Protect the privacy of the data subjects

Direct Identifiers

Nulling

Hashing

Encryption

Examples

- Social security number
- Account number
- Phone number

Indirect Identifiers

K-anonymization

Examples

- Zip code
- Birthdate
- Race

Sensitive Data

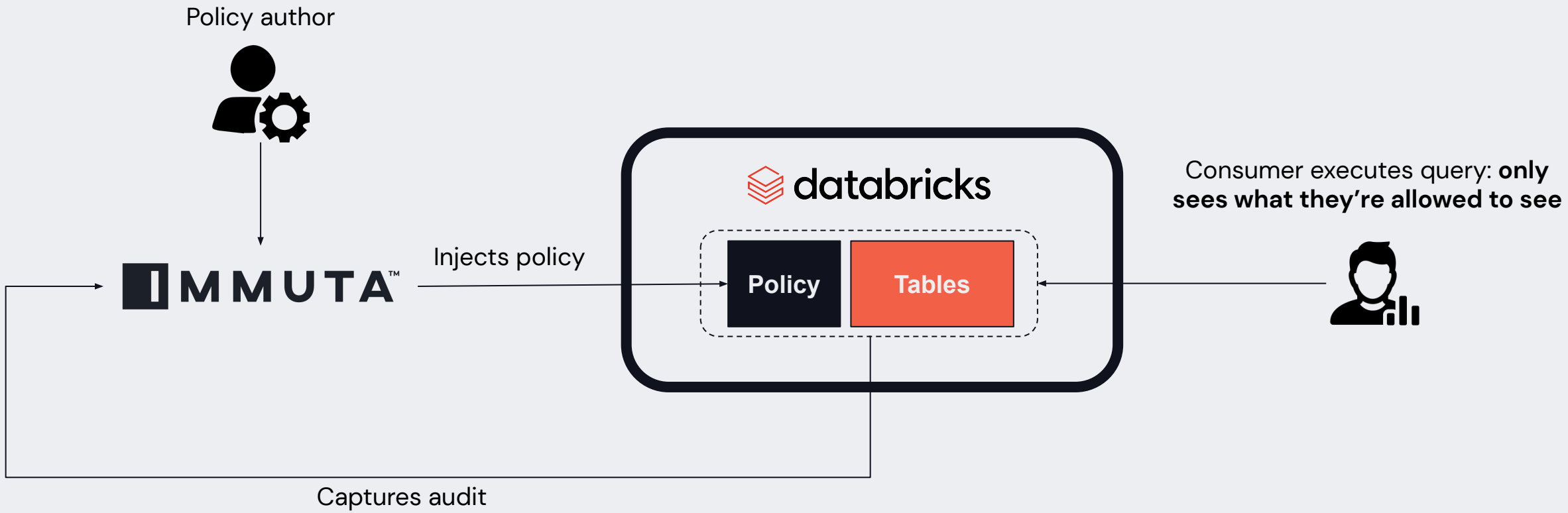
Randomized Response /
Local Differential Privacy

Examples

- Injury
- Salary

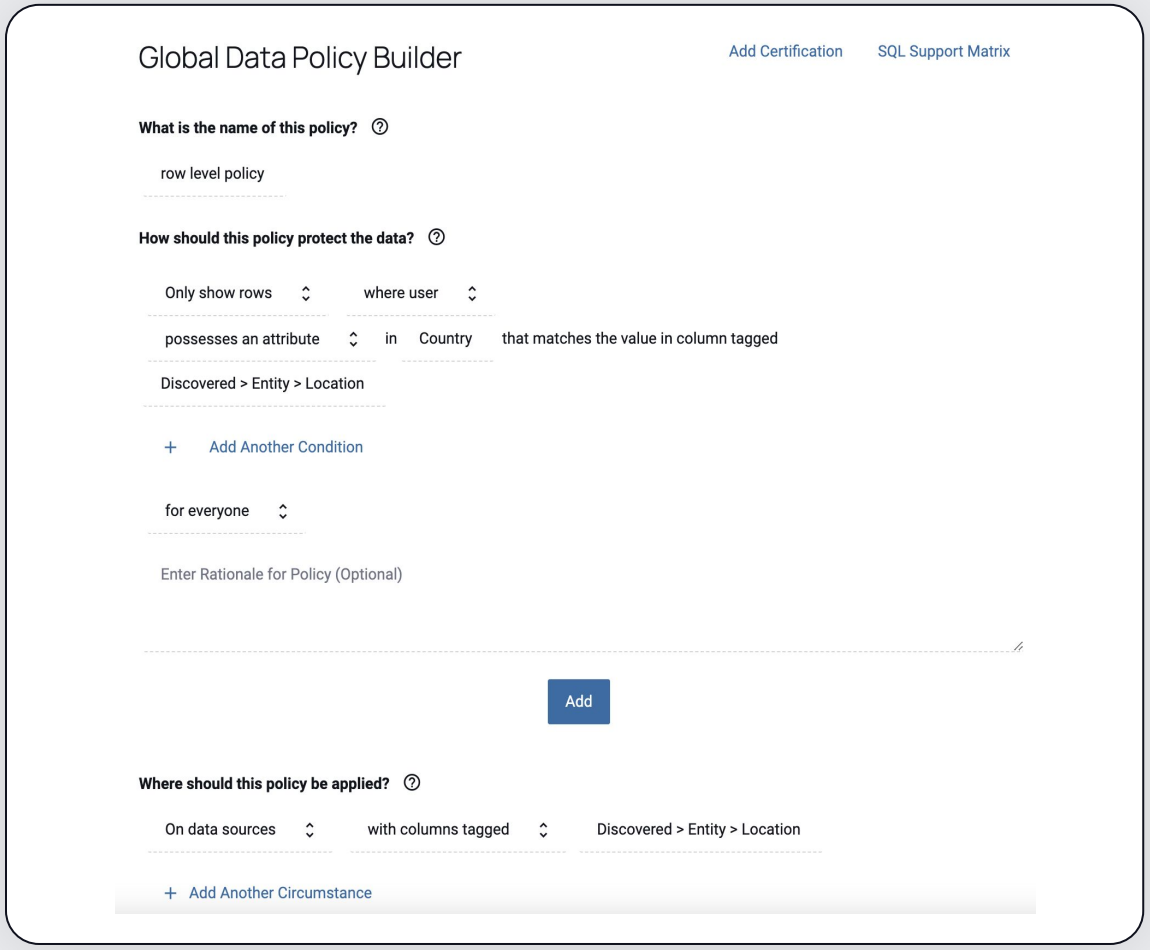
How to manage those controls?

Immuta



Policy authoring

Build policies simply and consistently across consumers

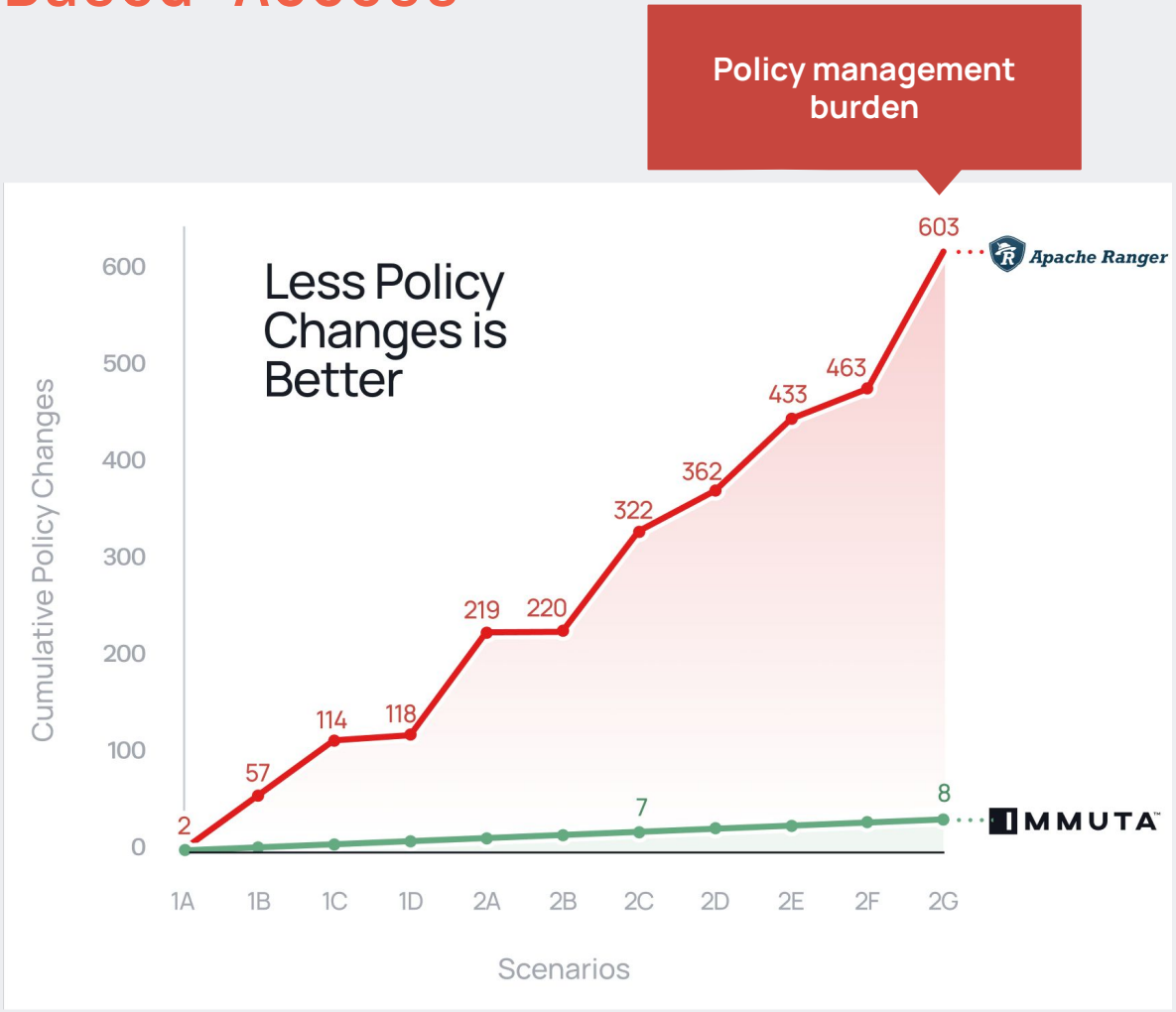


- Easy to author
- Easy to understand
- Easy to modify
- Automatically attaches to newly discovered metadata (proactive)
- Also available “as code” (and can easily verify policy applied “as code”)

Avoids policy management explosion

Scalability through Attribute-Based Access Control (ABAC)

> It took **96 policies** to filter 12 tables using **Ranger**, versus **1 policy** in **Immuta** to filter hundreds to thousands of tables.
> This increased time to data from 4 weeks to **instant access saving $289,966** per rule.
>
> - MULTINATIONAL MANUFACTURER



SOURCE: Independent and Fully Reproducible Study by GigaOm Research, 2021
*Based on Conservative Policy Burden Cost Estimate by GigaOm

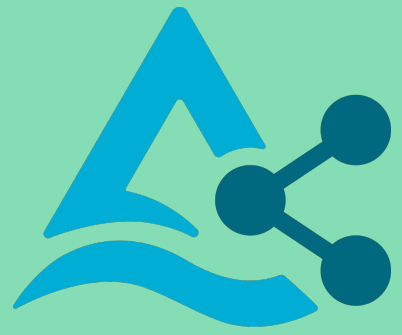
Audit all actions

See what data is most valuable, prove compliance

- The user
- The table
- Backing storage
- Time of query
- Purpose of query (if applicable)
- The spark plan
- The query
- The policy
- Who created the policy
  - When


```
“profileID”: 4,
“userID”: “mvogt@immuta.com”,
“dataSourceID”: 8,
“dataSourceName”: “Immuta Pov Immuta Fake Credit Card Transactions”,
“count”: 1,
“recordType”: “spark”,
“success”: true,
“component”: “dataSource”,
“accessType”: “query”,
“query”: “Project [id#3085, customer_last_name#3086, ImmutaHashMaskingExpression
“extra”: {
  “databricksWorkspaceID”: “0”,
  “maskedColumns”: {
  “transaction_time”: “Nullify”,
  “credit_card_number”: “Hashing”
  }
  “metastoreTables”: [
     “immuta_pov.immuta_fake_credit_card_transactions”
  ],
  “clusterName”: “steve-jonathan-demo”,
  “pathUris”: [

“dbfs:/user/hive/warehouse/immuta_pov.db/immuta_fake_credit_card_transactions”
  ],
  “queryText: “select * from immuta.pov.immuta_fake_credit_card_transactions:”,
  “queryLanguage”: “sql”,
  “clusterID”: “0504-135906-wp39eh6p”
},
```

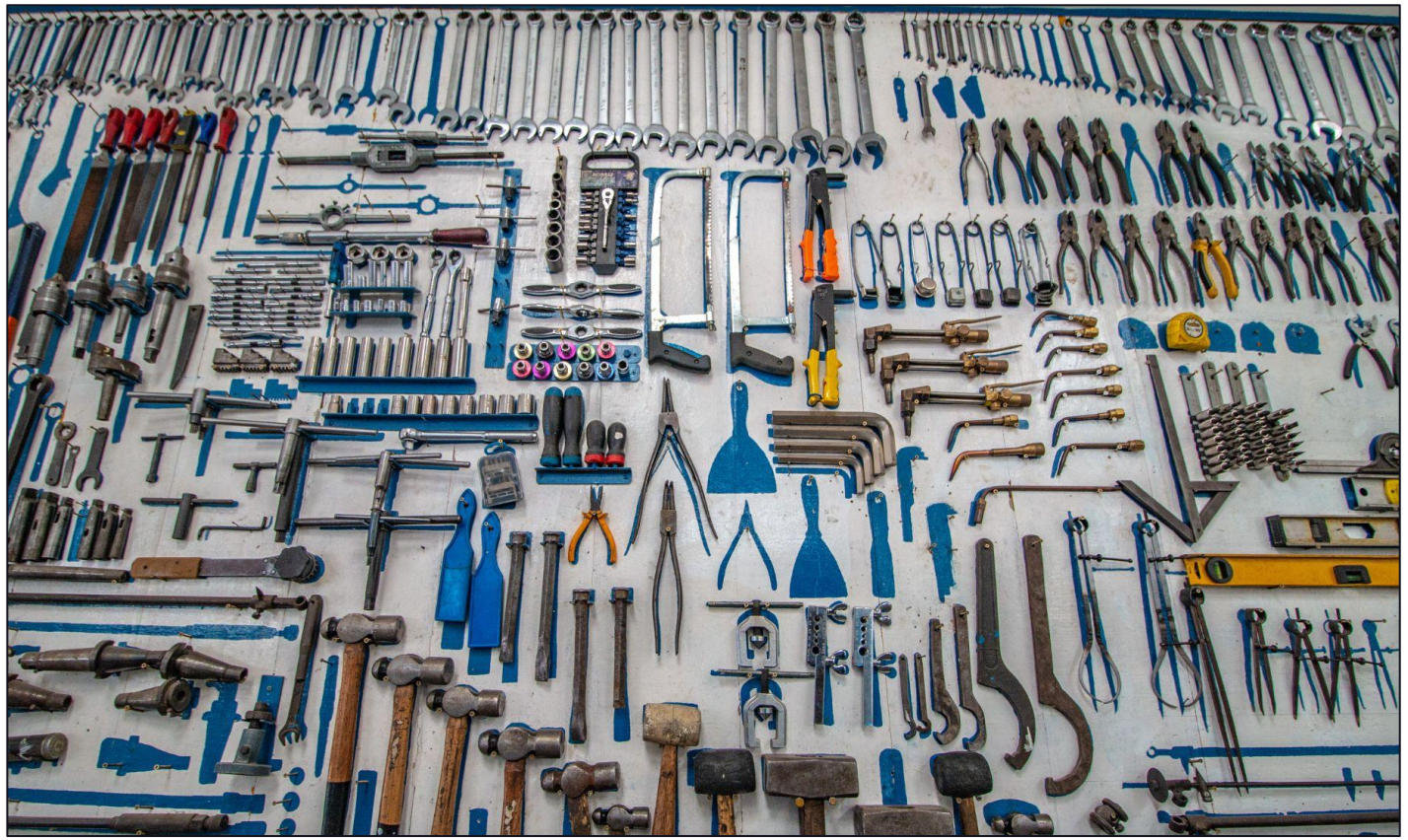


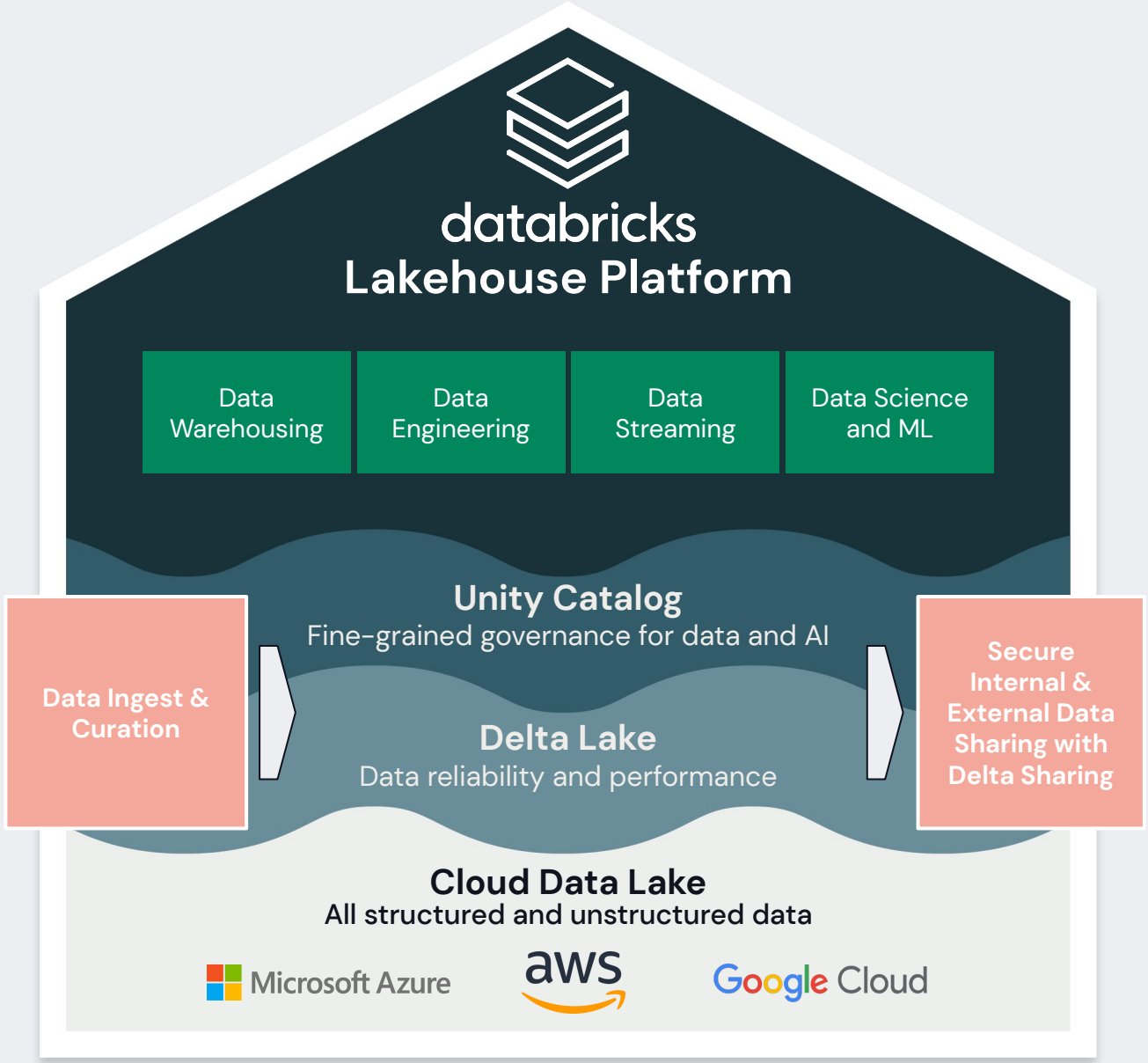
Delta protocol for data sharing

We have the policy in place, and now

There are still many challenges with private data sharing…

- Data replication
- Data access
  - So many tools are out there
  - So many formats
- Compute
  - Managed
  - Cloud
  - Bring Your Own
- DevOps
  - Getting this to run, and in scale





Delta Sharing with Databricks Lakehouse

Simple

Unify your data warehousing and AI use cases on a single platform

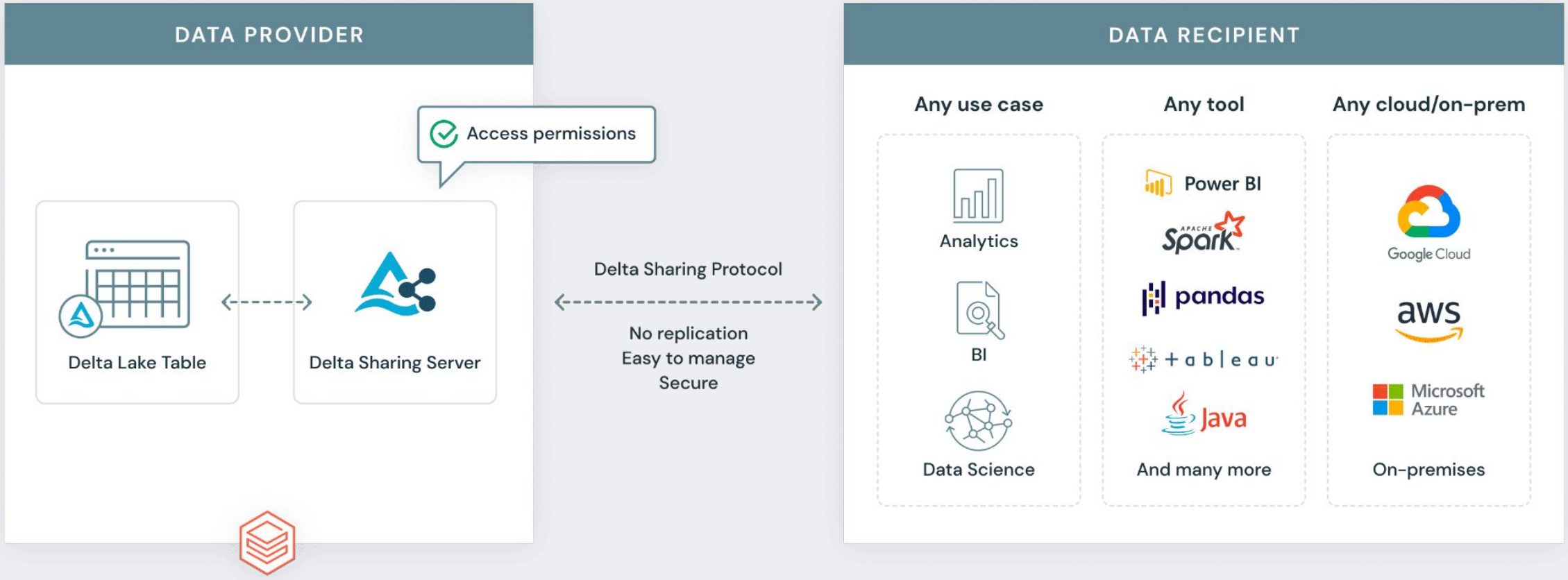
Open

Built on open source and open standards

Multi-cloud

One consistent data platform across clouds

Delta Sharing - The world's first open protocol for secure data sharing

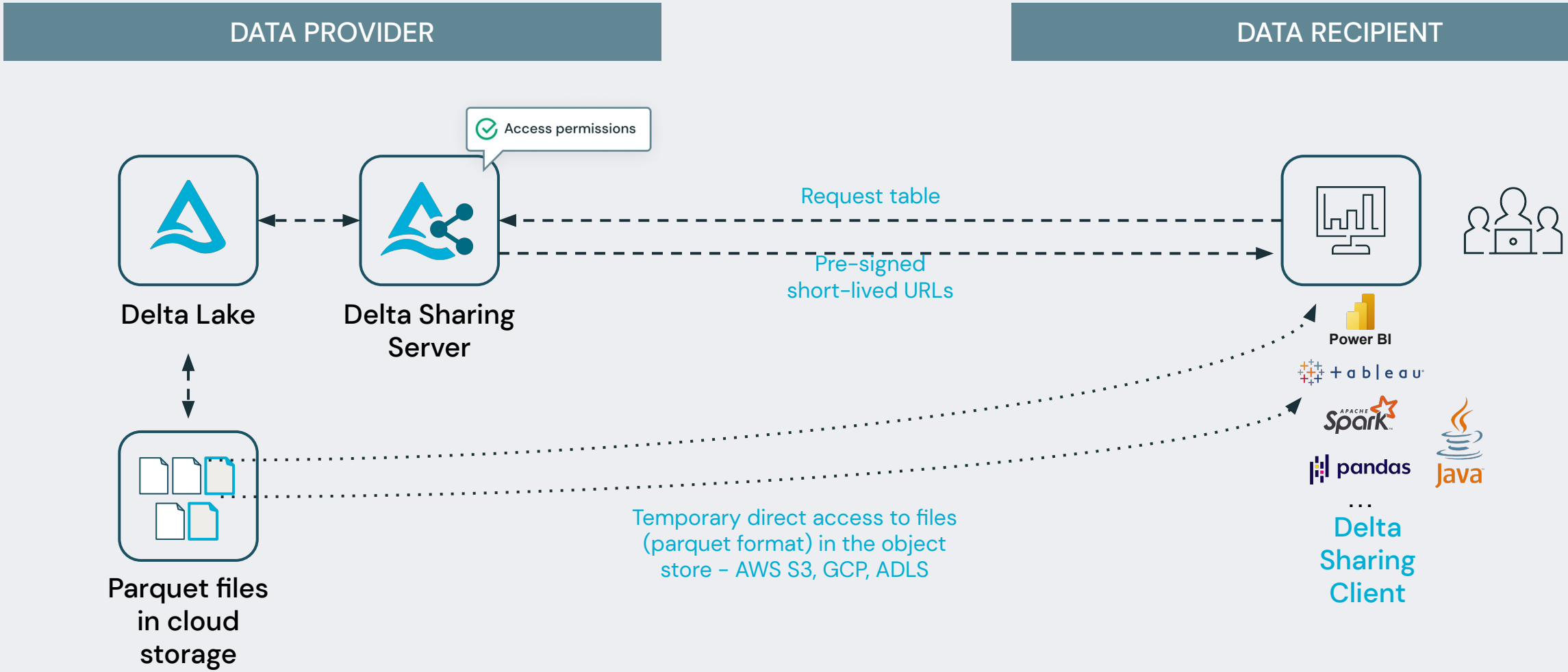


Open Cross-Platform Sharing

Share live data with no replication

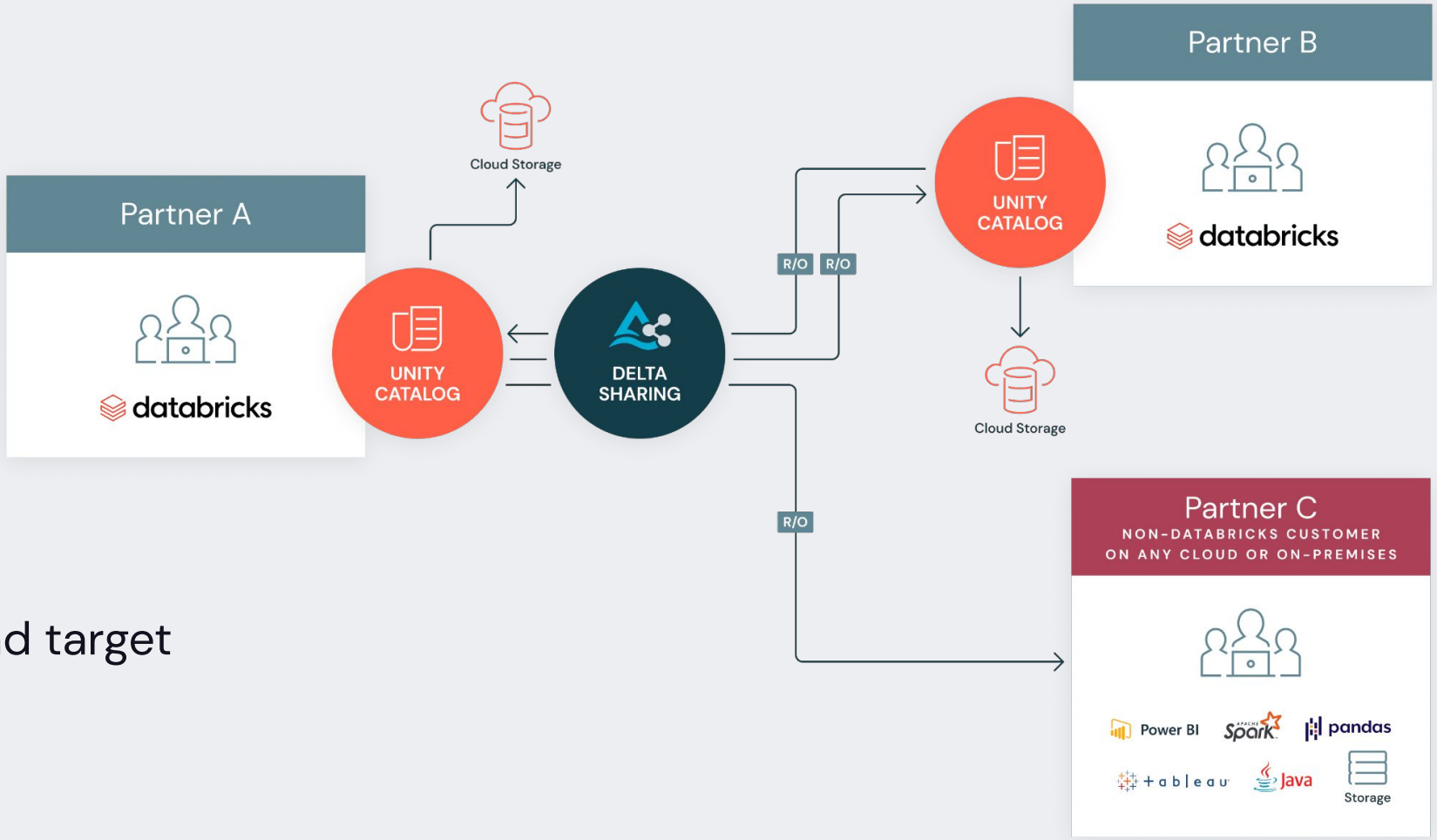
Ingestion & Dist.

How Delta Sharing works



Sharing is easy

- Read Only
- Always secured
- Multi-cloud or Multi-region
- Audited
- Managed
  - Tokenless
  - Known fixed source and target metastores



Secure data Share data –

Managed or standard sharing

Private data sharing – managed

- Tokenless
- Known, fixed source and target metastores
- Databricks trusted compute:
  - Governed or Not governed – as needed
  - Raw data protected – from any query
- Query anonymized data from a view
  - K-anonymization
  - Hide identifiers
  - Differential Privacy – Noise generation
- The recipient still ends up with a copy of the shared data

Anonymized data sharing – standard

- Client compute:
  - Not governed
  - Can't expose raw data
- Prepare or Serve anonymized data
  - K-anonymization – limited number of "dimensions"
  - Remove identifiers
  - Noise generation
  - Can power interactive and batch queries.
- The recipient still ends up with a copy of the shared data

PUMPJACK

DATAWORKS

Developing the industry leading fan engagement platform

Empower Organizations, Leagues and Teams

A Fan Data Platform tuned specifically for the Sports Industry - to help teams unlock the value of their fan data, and yield immediate results.

PUMPJACK

DATAWORKS

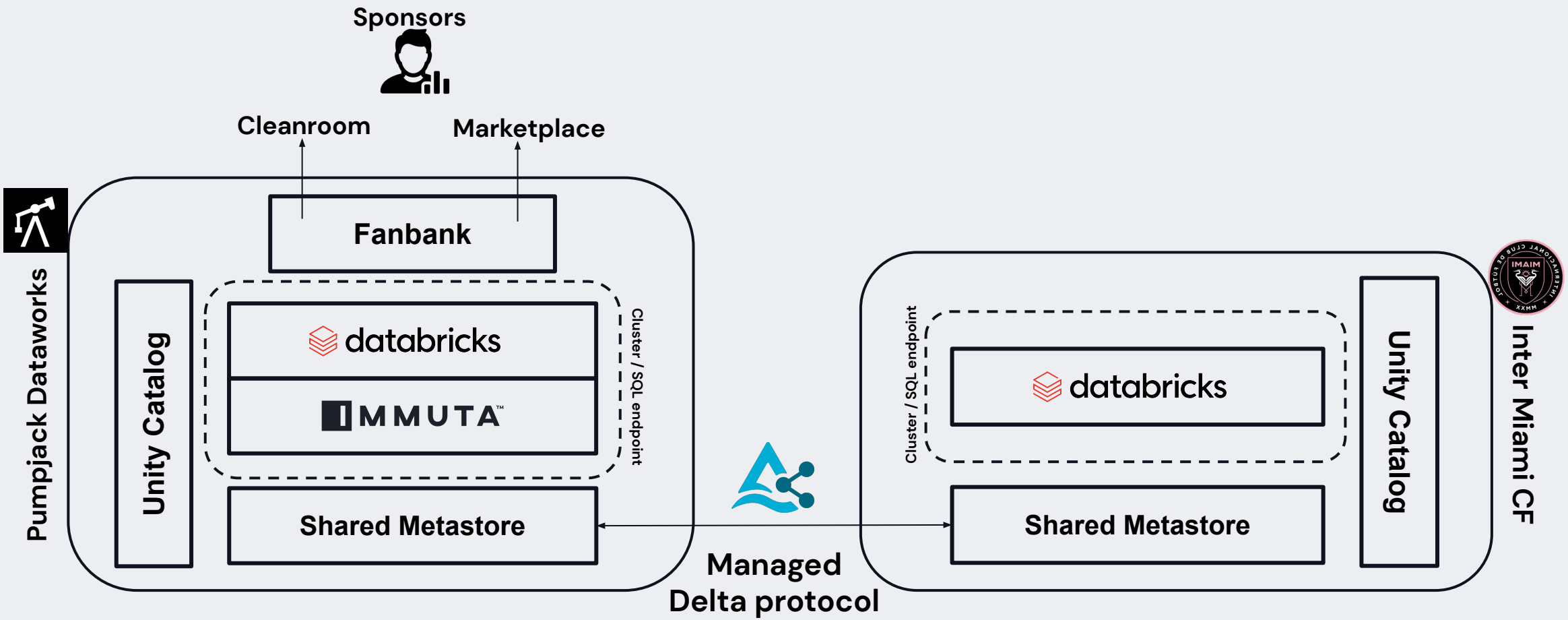
Modern data challenges require modern solutions

- Same old problems in a modern data landscape…
  - How do I connect all my data into one place to drive valuable insights?
  - I already have a data lake or warehouse but it’s painful to work with.
  - By the time I get an answer from my data, it's too late.
  - Sponsors are asking for information and I can’t provide it.

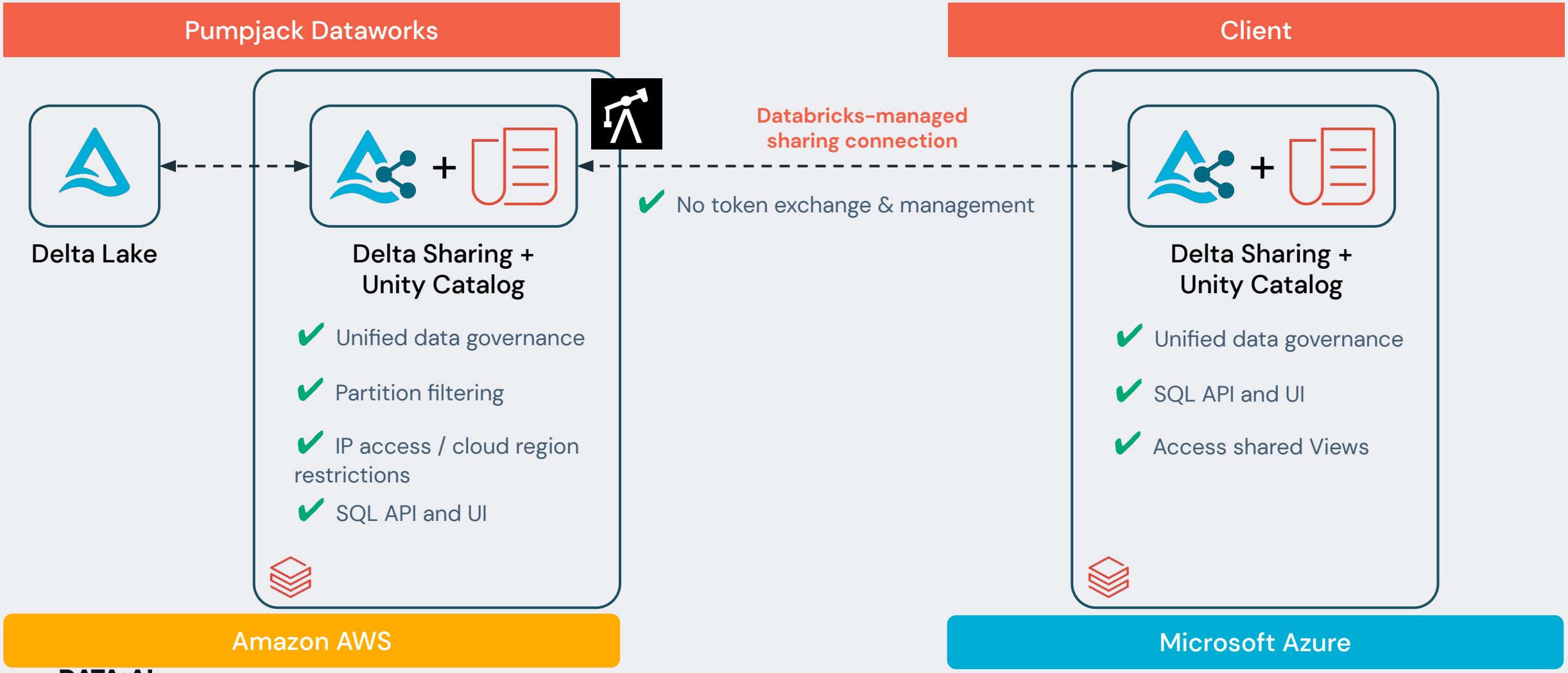
“Our data is a cost & doesn’t generate revenue!”

Pumpjack Dataworks Architecture

Fan engagement platform made real

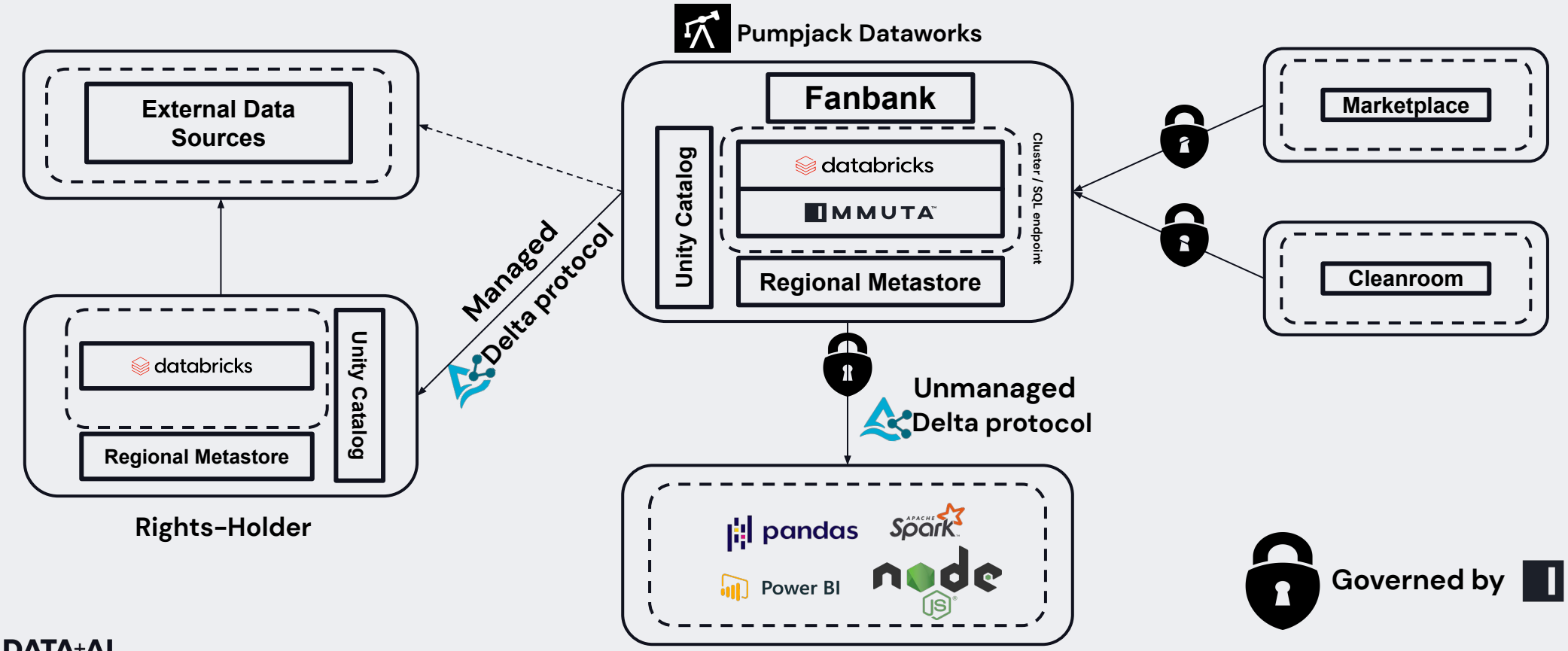


Delta Sharing on Databricks... Multi Cloud



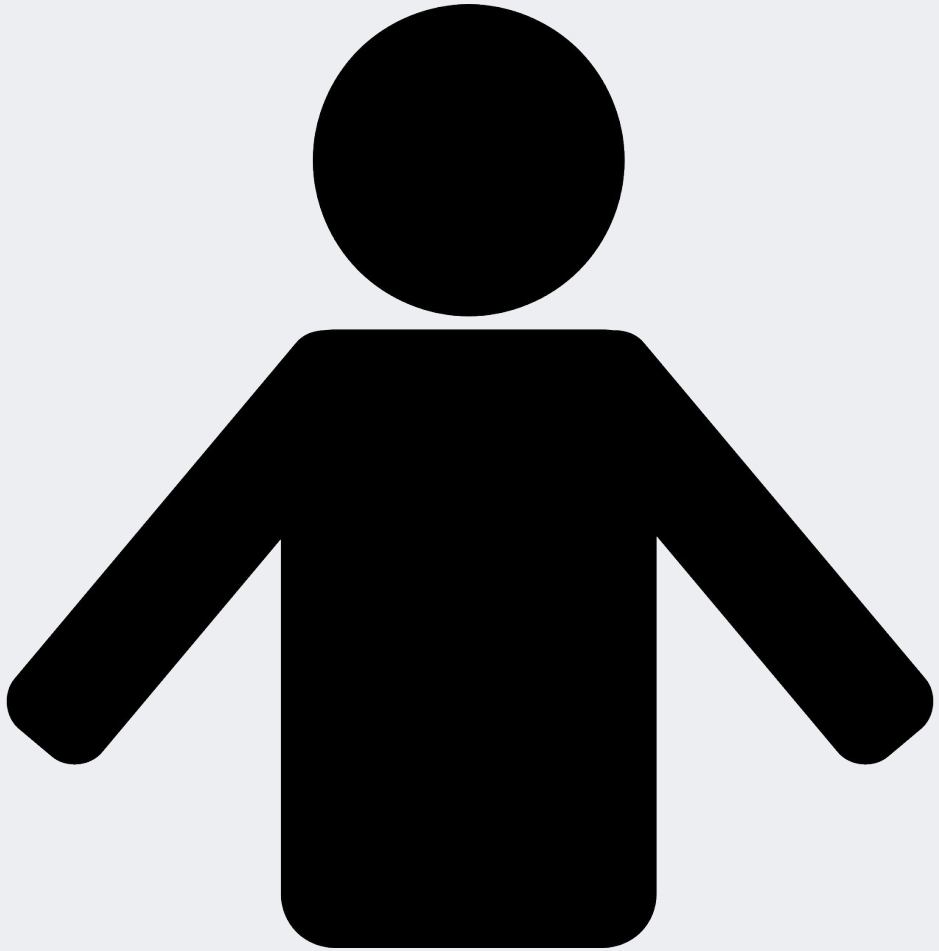
Source of Truth

Keeping control of the data in the rights-holders hands



People or boxes

ABAC… dont put me in a box



VS

Demo

Conclusion slide

What Immuta has done for us…

ABAC, Governance, Masking and Auditing… it has it all.
Every organization, large or small needs these in order to ensure their fans' data remains secure.

One easy answer to questions that our future clients always ask.

"We protect your data with Immuta."

Questions?

Open questions

Interested in additional information

- Come visit Immuta at booth 727
- Join us tomorrow to learn more about Unity + Immuta
  - *Complete Data Security and Governance Powered by Unity Catalog and Immuta*
    Time: 2:05 pm – 2:45 pm

DATA+AI
SUMMIT 2022

Thank you

Corey Zwart
Head of Engineering,
Pumpjack Dataworks

Itai Weiss
Data Partner SA,
Databricks

Steve Touw
CTO, Immuta