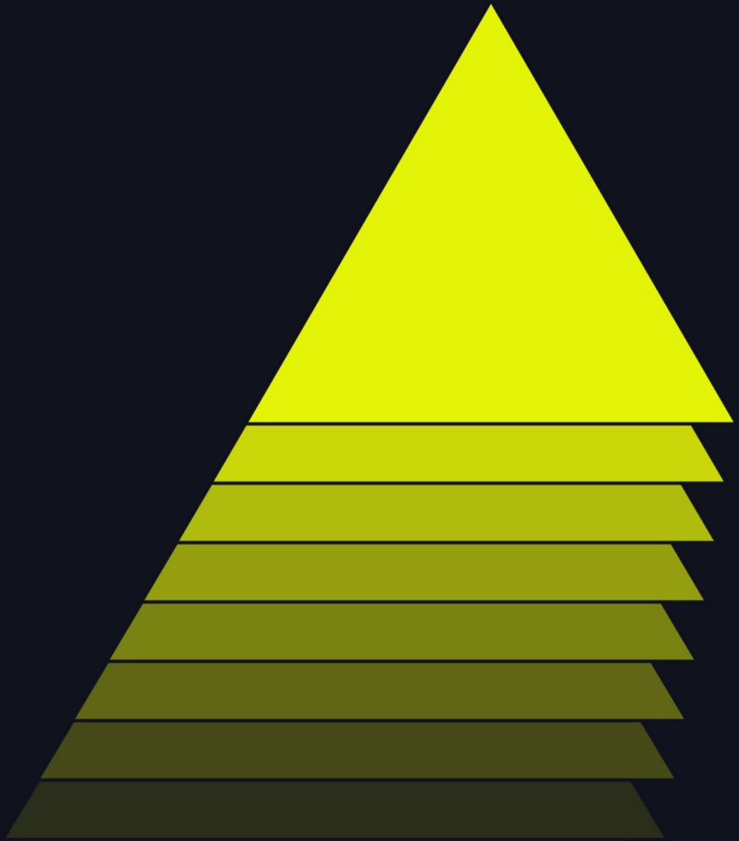
Product safe harbor statement

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

DATA+AI SUMMIT
BY databricks

ENTERPRISE COLLABORATION WITH DELTA SHARING



Tianyi Huang, Databricks | Jay Hugalavalli, Ontada | Javier Asensio, Kraken Technologies

June 2024

PRESENTERS

Tianyi Huang
Product Manager, Databricks

Javier Asensio
Head of Data Platform Engineering, Kraken Technologies

Jay Hugalavalli
Senior Director Data Management, Ontada

SHARING AND COLLABORATION IS A CRITICAL IMPERATIVE FOR ENTERPRISES

Life Sciences & Clinical Research

Advertising & Marketing

Commercialization

Financial Markets

Supply Chain & Operations

Regulatory & Reporting

HR

FLAVORS OF DATA COLLABORATION

- Data licensing from data vendors
- SaaS platform zero-copy bi-directional sharing
- Peer-to-peer sharing & collaboration

Today's focus

- Enterprise sharing across domains / business units

TOP CHALLENGE IN ENTERPRISE COLLABORATION: DATA SILOS

FORRESTER®

Study of data overload

- 60% of data leaders describe data silos as a top barrier (2nd highest-rated) to better capturing, analyzing, and acting on data

SOURCES OF DATA SILOS

Both organizational and technological reasons lead to data silos

- Organizational: centralized data teams and processes not keeping up with organizational growth
  - Geo expansion
  - Data localization
  - M&A and Conglomerates
- Technological: proprietary data platforms and formats causing a lack of interoperability
  - Multiple platforms

DATA NEEDS TO FLOW ACROSS THE ENTERPRISE

SOCIO-TECHNICAL APPROACHES

Various patterns emerged to tackle the organizational challenges

⬢ Data domain

⬢ Global hub

Publish and
discover data

- - - - - - - - - -

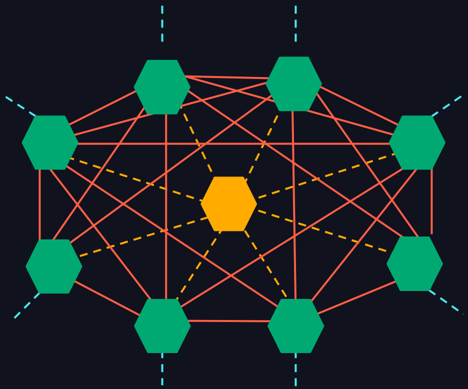
Consume data

———————

External sharing

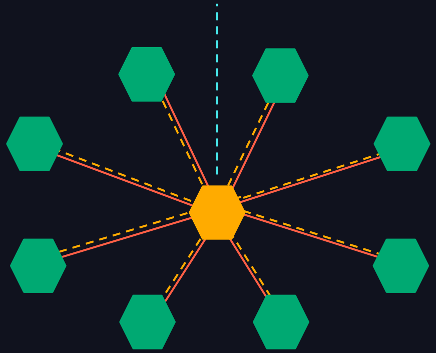
- - - - - - - - - -

Data mesh



Global hub as discovery catalog, each domain hosts and serves its own data

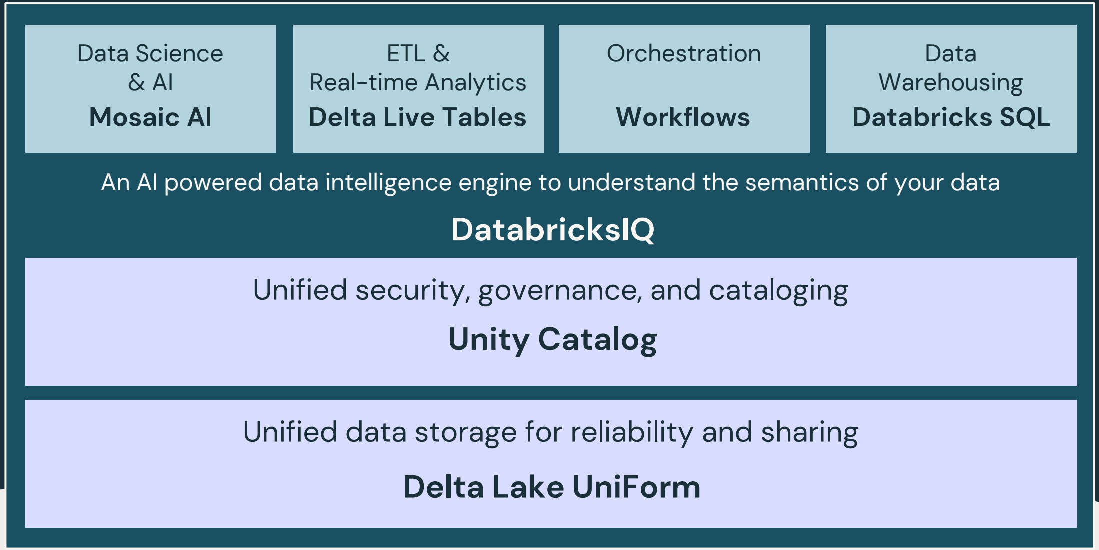
Hub-and-Spoke



Global hub for publishing, discovery and serving of enterprise data

DATABRICKS DATA INTELLIGENCE PLATFORM

Complementary to any enterprise mesh-like design pattern



Data Science & AI
Mosaic AI

ETL & Real-time Analytics
Delta Live Tables

Orchestration
Workflows

Data Warehousing
Databricks SQL

An AI powered data intelligence engine to understand the semantics of your data
DatabricksIQ

Unified security, governance, and cataloging
Unity Catalog

Unified data storage for reliability and sharing
Delta Lake UniForm

Open Data Lake

All Raw Data
(Logs, Texts, Audio, Video, Images)

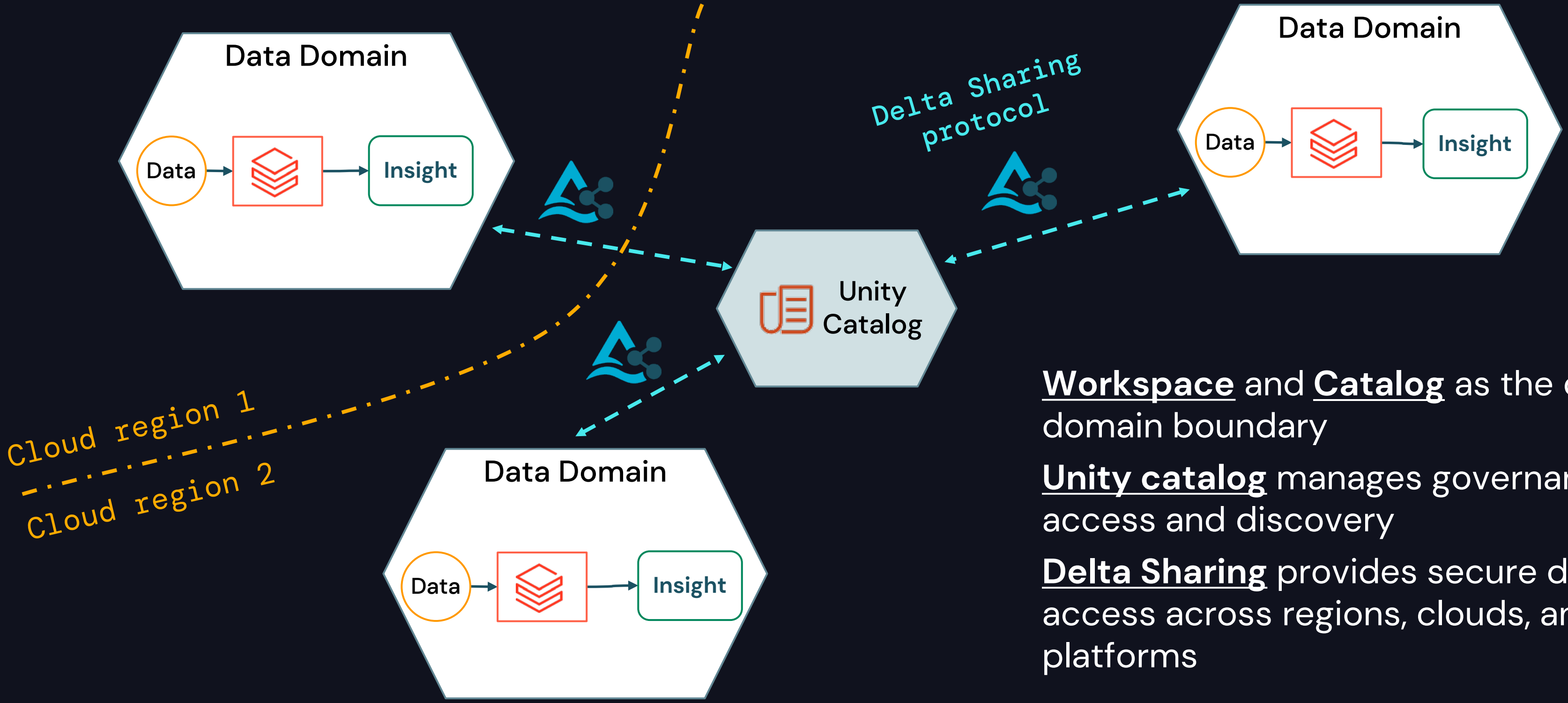
Interoperability : an open technology stack

Flexibility: a unified environment for all data personas

Governance and sharing: built-in data discovery, access, lineage and secure data sharing across domains

DATABRICKS DATA INTELLIGENCE PLATFORM

How to implement a hub-and-spoke model with Databricks



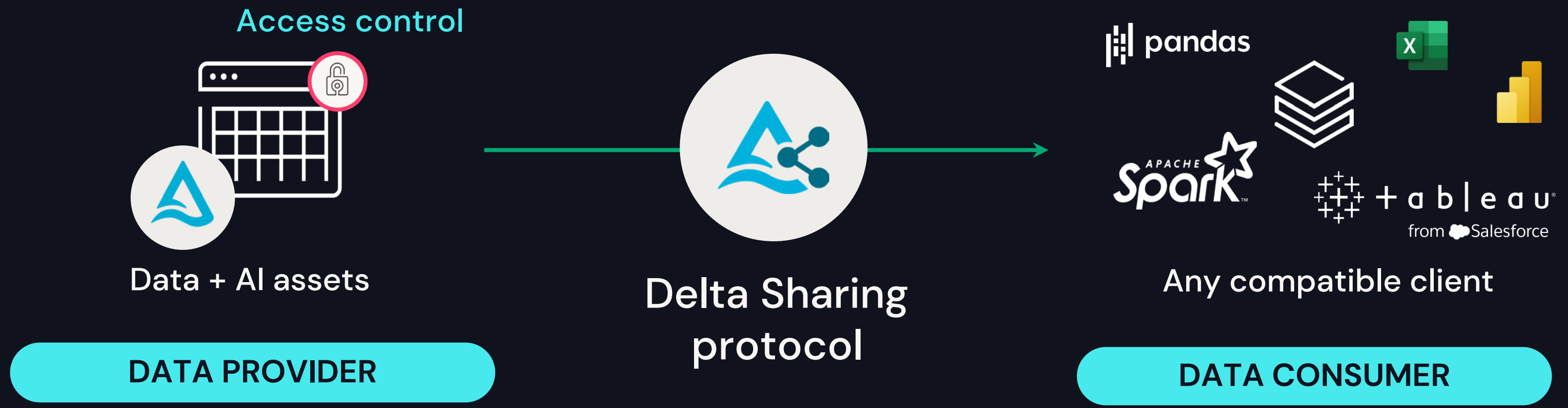
Workspace and **Catalog** as the data domain boundary

Unity catalog manages governance, access and discovery

Delta Sharing provides secure data access across regions, clouds, and platforms

FOUNDATION: DELTA SHARING

Open and secure sharing across domains



Access control

Data + AI assets

DATA PROVIDER

Delta Sharing protocol

Any compatible client

DATA CONSUMER

Share all data + AI assets

Live sharing
across regions and clouds

Open collaboration
across platforms

DATABRICKS DELTA SHARING ECOSYSTEM

Fast growing ecosystem among customers and partners

16K+

data recipients using Delta Sharing

+300% YoY

growth in active Delta Shares

40%

active Delta Shares use open connectors

License data from vendors

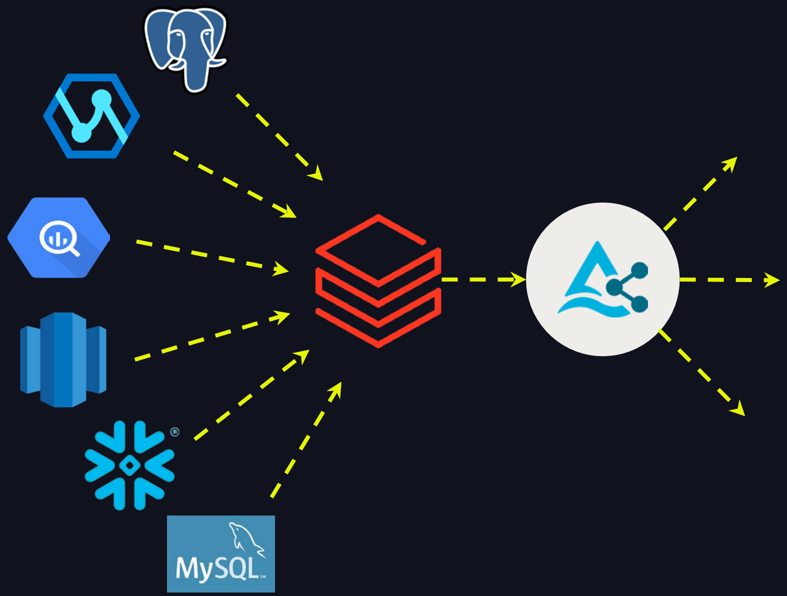
Exchange data with SaaS platforms

Peer-to-peer collaboration

WHAT'S NEW

More innovations to empower enterprise collaboration

Sharing for Lakehouse Federation



Private Exchange

- A storefront interface for discovering and accessing data products
- Control data product discoverability to specified consumers

Kraken Technologies

KRAKEN TECHNOLOGIES

Who are you? 🧐

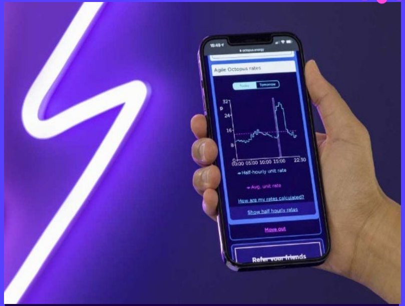
We use **technology** to drive the global green energy revolution – making it **cheaper and faster** for citizens and the planet.

KRAKEN TECHNOLOGIES

Who are you? 🧐

Octopus Energy

Retail Energy



- **9m+** customers
- **8** countries
- **#1** for customer service
- **World-leading** consumer flex

Services

- **EV** leasing
- **Heat pump** tech & installations
- **Solar and meter** installations
- **Electroverse EV charging** network

Generation

- **$7bn** generation assets managed
- **14** countries
- Fan Club community generation

Kraken

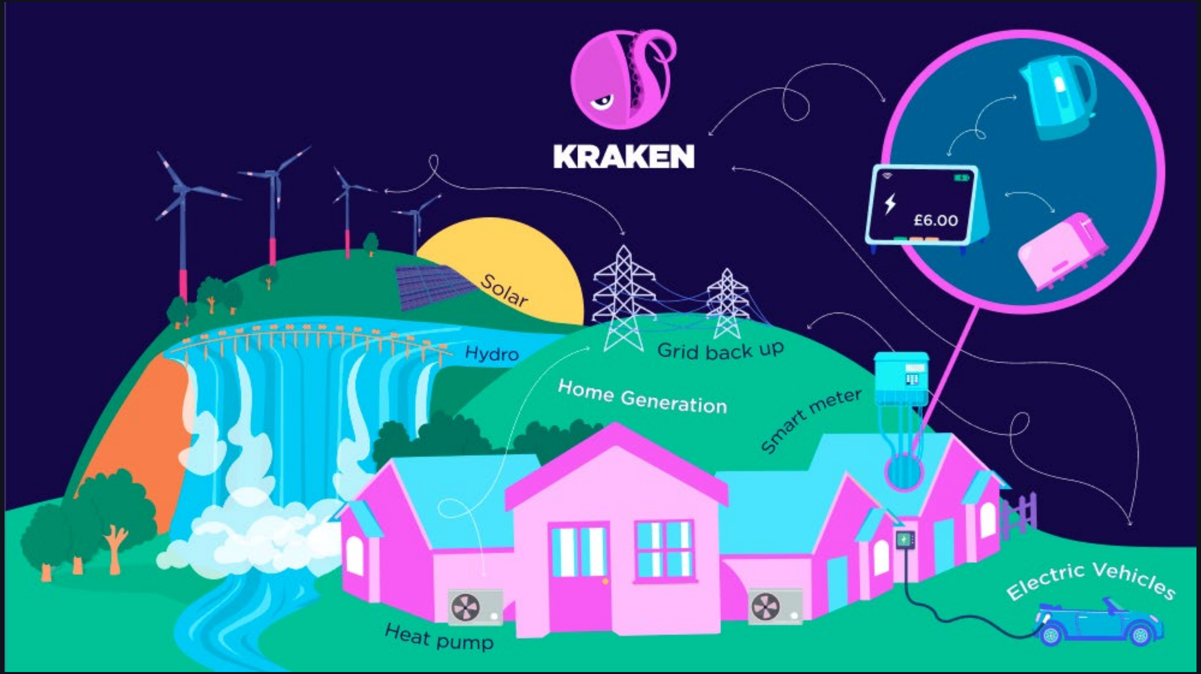
Tech

- **20 countries**
- **54m+** contracted accounts
- **30+ migrations** from 17 platforms
- KrakenFlex: **5GW** contracted

KRAKEN TECHNOLOGIES

An advanced *operating system* for utilities

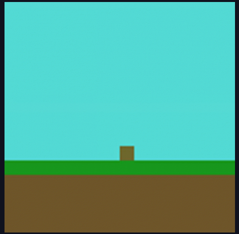
Kraken powers the integrated solutions our planet needs



THE KRAKEN DATA PLATFORM JOURNEY

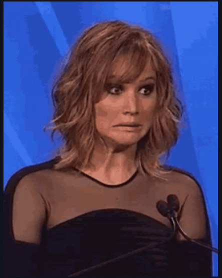
From smol to humongous (and beyond)

<2018	2020	2022	2024
1 Kraken Client	5 Kraken Clients	15 Kraken Clients	30 Kraken Clients
200K customers	3M customers	18M customers	54M customers



CHALLENGES

Oooops! too many clients



- We run 1 Databricks environment or more per client (we have around 30 accounts with more than 40 workspaces)

- Sharing data internally
  - Data centralisation can be challenging
  - Some datasets need to be everywhere or partially everywhere

- Sharing data externally
  - Sharing data with our clients is difficult for the most part
  - Onboarding clients can be **VERY** challenging

DELTASHARING TO THE RESCUE

Easy peasy

- Dead easy to share data across our databricks accounts
- We can share internally and with clients
- Everything can be implemented as infrastructure as code
- Before we were using just raw parquet files and S3 to share

SHOW AND TELL

Sharing data from all our Databricks accounts into a centralised platform

Many isolated clients

- ~30 dbx accounts
- Each client is completely isolated
- The architecture is **very** consistent
- The pipelines are **very** consistent

Needs for centralisation

- Centralised reports
- For business use cases
- For helping devs with metrics
- This data is **anonymised**

S&T: PROBLEM STATEMENT: MONEY?!

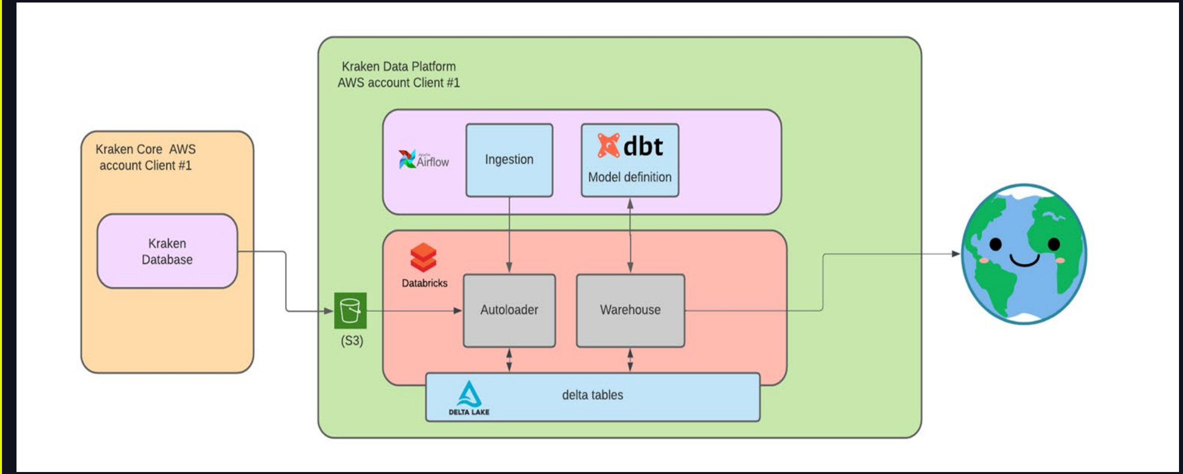
How much we we spend in Databricks across all our clients?

“David Sykes” - Head of Data Octopus Energy Group

**Dramatisation*

S&T: ARCHITECTURAL OVERVIEW

Single client

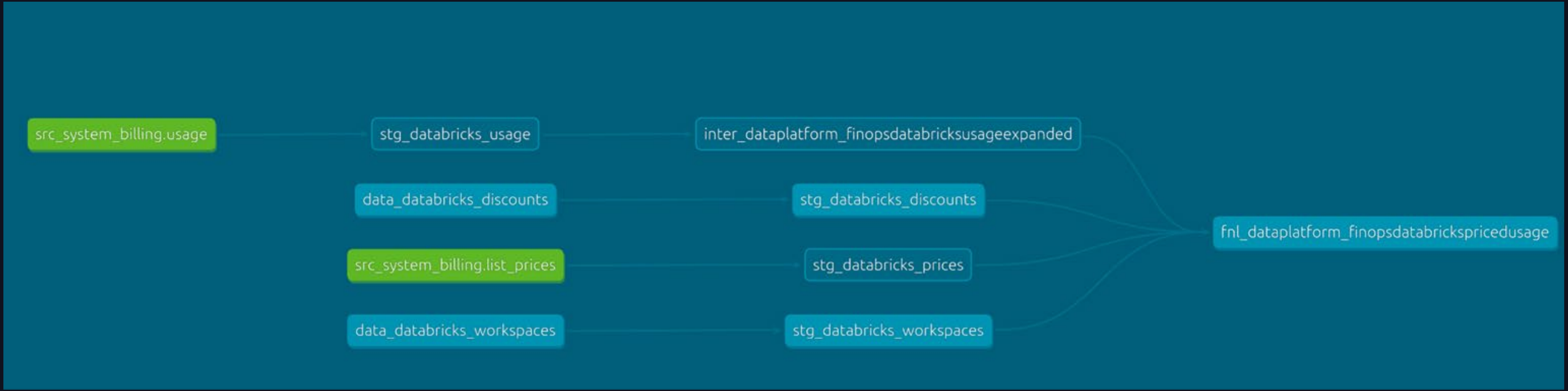


- One AWS Account of our CRM
- One AWS account for DBX
- Airflow for orchestration
- dbt for transforming data
- Medallion architecture in the inside

S&T: Pipeline

FinOps pipeline

- We gather the data for each client from the system tables and other info



src_system_billing.usage

stg_databricks_usage

inter_dataplatform_finopsdatabricksusageexpanded

data_databricks_discounts

stg_databricks_discounts

fnl_dataplatform_finopsdatabrickspricedusage

src_system_billing.list_prices

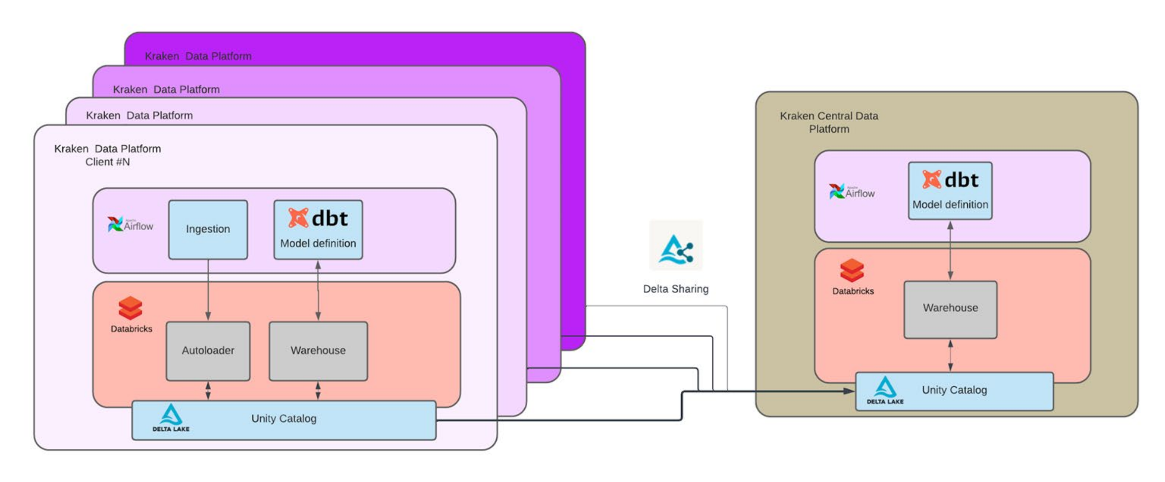
stg_databricks_prices

data_databricks_workspaces

stg_databricks_workspaces

S&T: ARCHITECTURAL OVERVIEW

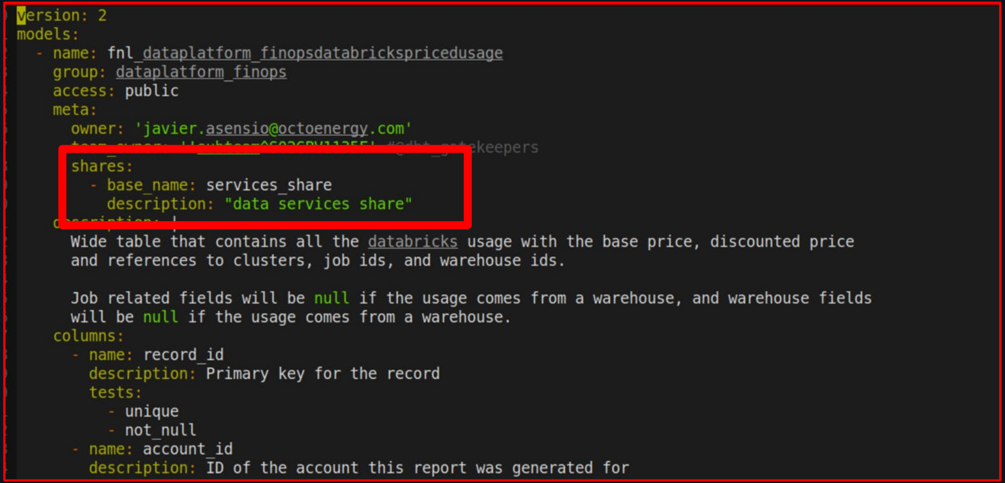
Many clients



- All the clients share with a central account
- Everything is terraformed
- Common upstream schemas
- We use dbt to share the tables and union them

S&T: Pipeline

FinOps pipeline



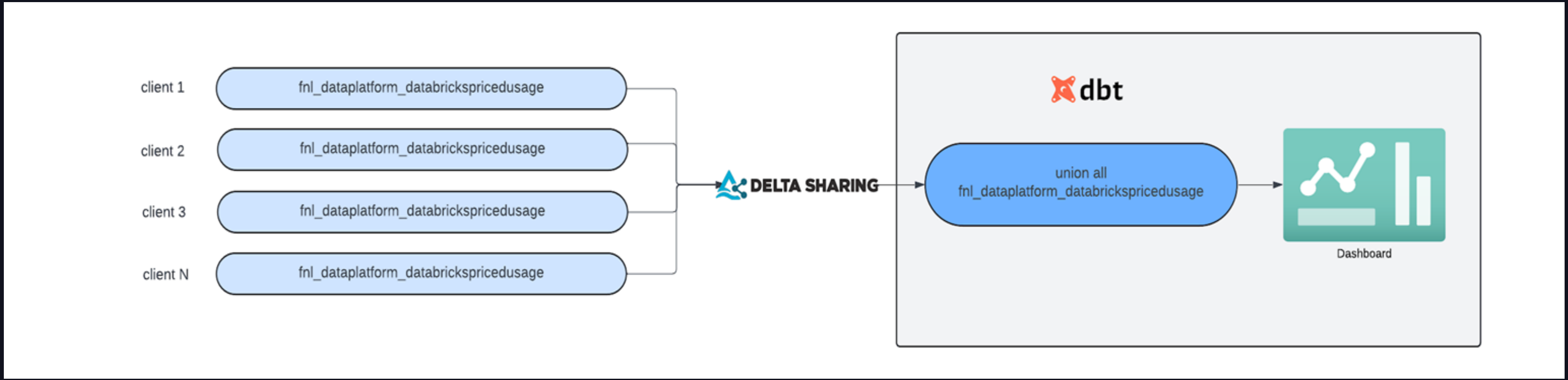
```yaml
version: 2
models:
  - name: fnl_dataplatform_finopsdatabrickspricedusage
    group: dataplatform_finops
    access: public
    meta:
      owner: 'javier.asensio@octoenergy.com'
      team_owner: '!subteam^S02CPV1135F' #@dbt_gatekeepers
      shares:
        - base_name: services_share
          description: "data services share"
    description: |
      Wide table that contains all the databricks usage with the base price, discounted price
      and references to clusters, job ids, and warehouse ids.

      Job related fields will be null if the usage comes from a warehouse, and warehouse fields
      will be null if the usage comes from a warehouse.
    columns:
      - name: record_id
        description: Primary key for the record
        tests:
          - unique
          - not_null
      - name: account_id
        description: ID of the account this report was generated for
```

- This bit shares this final table with the central account

S&T: Pipeline

FinOps pipeline



client 1

fnl_dataplatform_databrickspricedusage

client 2

fnl_dataplatform_databrickspricedusage

client 3

fnl_dataplatform_databrickspricedusage

client N

fnl_dataplatform_databrickspricedusage

DELTA SHARING

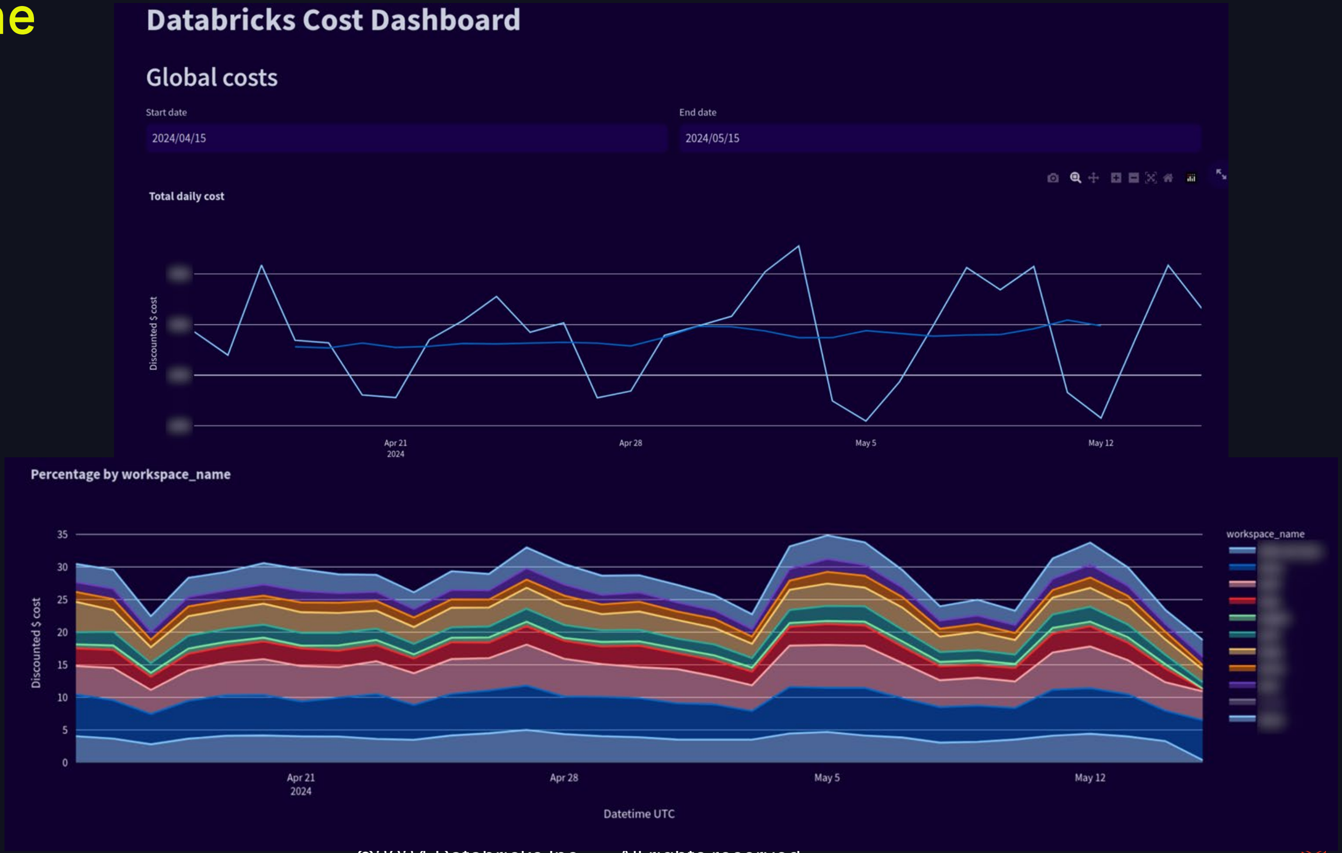
dbt

union all
fnl_dataplatform_databrickspricedusage

Dashboard

S&T: Pipeline

FinOps pipeline



Databricks Cost Dashboard

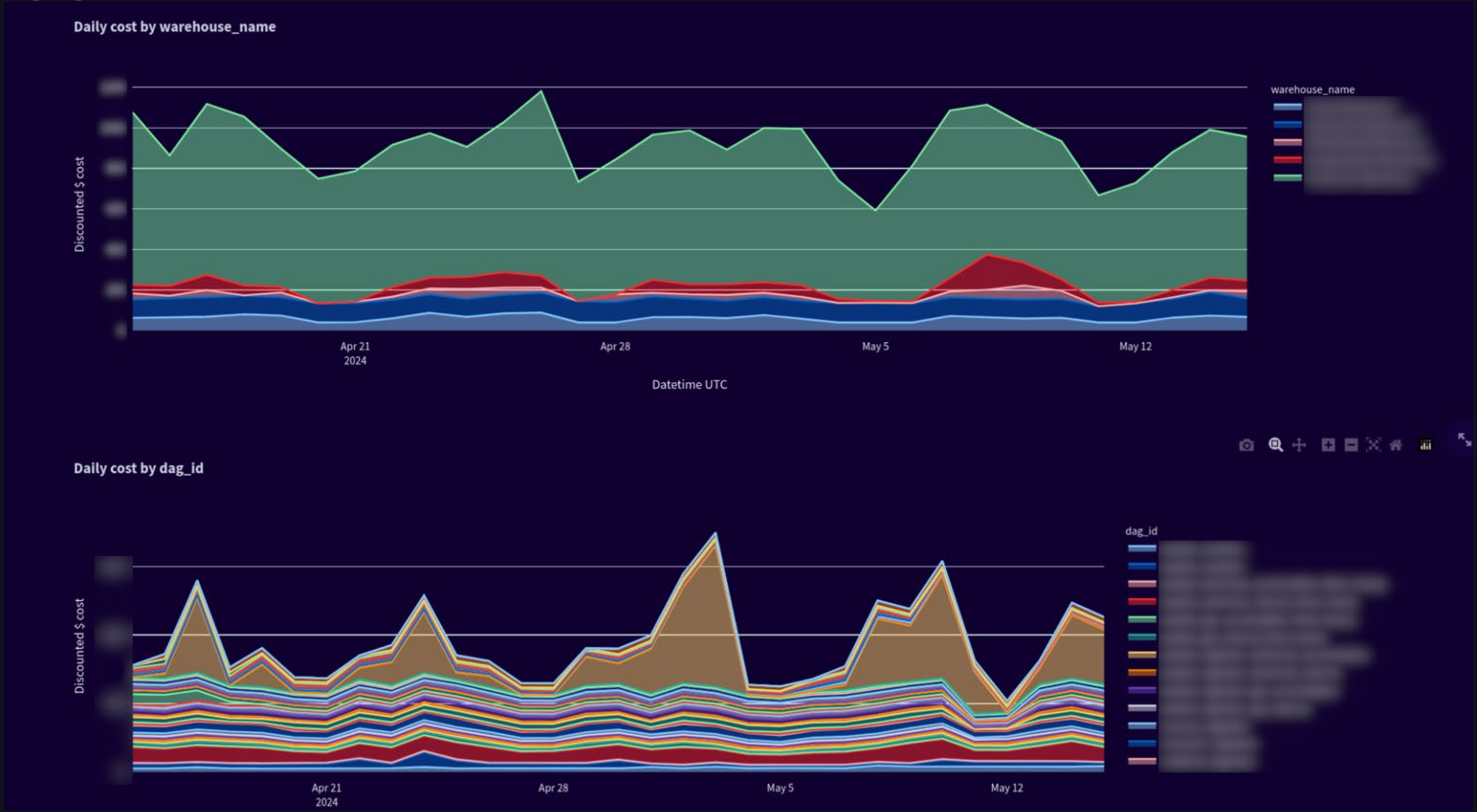
Global costs

Start date: 2024/04/15

End date: 2024/05/15

S&T: Pipeline

FinOps pipeline



OTHER USE CASES

There’s more!

- From the central account to other accounts
  - Electricity market data
  - Weather data
- For our clients*
  - Easy sharing of raw and processed data

CONCLUSIONS

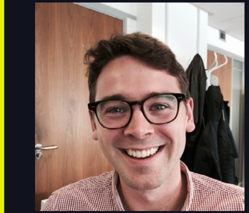
The takeaways

- No more siloed data
  - With deltasharing we can easily access data across different databricks accounts
  - Very easy and transparent configuration, analysts get all the power they need.
- Great for internal and external usage alike
  - Sharing data with out own accounts
  - Simplifies how we share data with out clients too

CONCLUSIONS

The takeaways

- y access da ts
- figuration, a
- usage ali
- ounts
- vith out clier



“Oh, wow!”

David Sykes

Head of data
Octopus Energy

"Delta Sharing creates simplicity for our clients so they can make greater impact quicker”

Mike Yorwerth

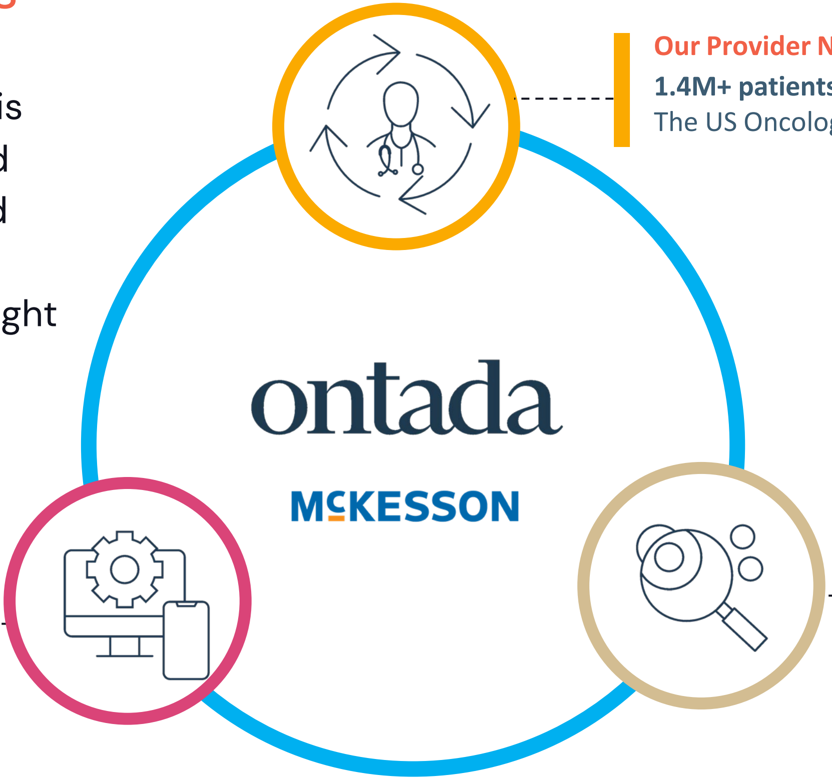
COO Kraken
Technologies

Ontada

Ontada – Introduction

Transforming the fight against cancer

Ontada, a McKesson business, is an oncology real-world data and evidence, clinical education, and provider technology business dedicated to transforming the fight against cancer.



Our Provider Network

1.4M+ patients seen, including The US Oncology Network

Market-Leading Provider Technology

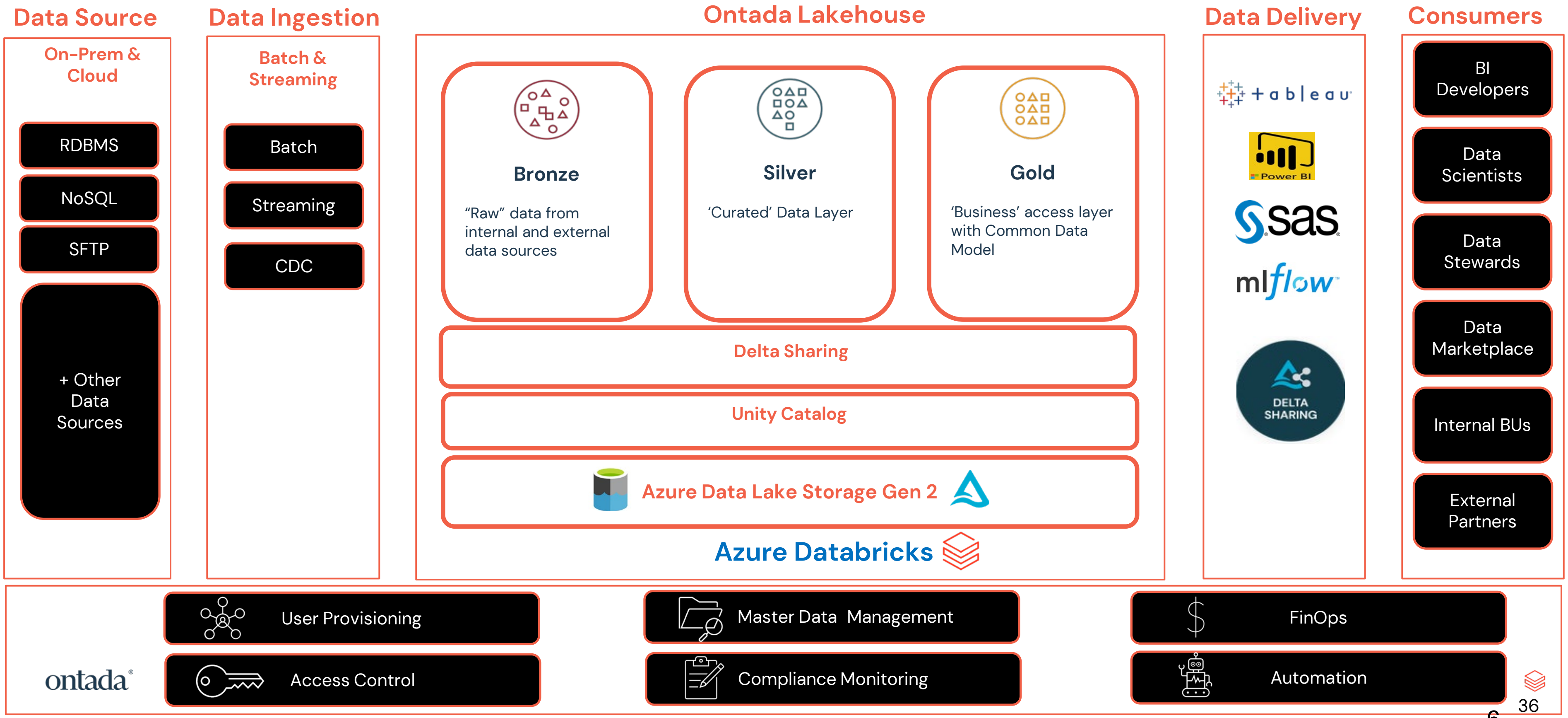
2.6K+ providers use iKnowMed^{SM}

Real-World Oncology Data & Insights

2.4M+ patient records available for research

Reference: https://www.ontada.com

Ontada Lakehouse



Data Source

On-Prem & Cloud

- RDBMS
- NoSQL
- SFTP
- + Other Data Sources

Data Ingestion

Batch & Streaming

- Batch
- Streaming
- CDC

Ontada Lakehouse

Bronze

“Raw” data from internal and external data sources

Silver

‘Curated’ Data Layer

Gold

‘Business’ access layer with Common Data Model

Delta Sharing

Unity Catalog

Azure Data Lake Storage Gen 2

Azure Databricks

Data Delivery

Consumers

- BI Developers
- Data Scientists
- Data Stewards
- Data Marketplace
- Internal BUs
- External Partners

- User Provisioning
- Access Control
- Master Data Management
- Compliance Monitoring
- FinOps
- Automation

ontada

Data Sharing – the Old Way without Delta Sharing

Data Silos

Complex ETL

Legacy Data Transfer Mechanisms

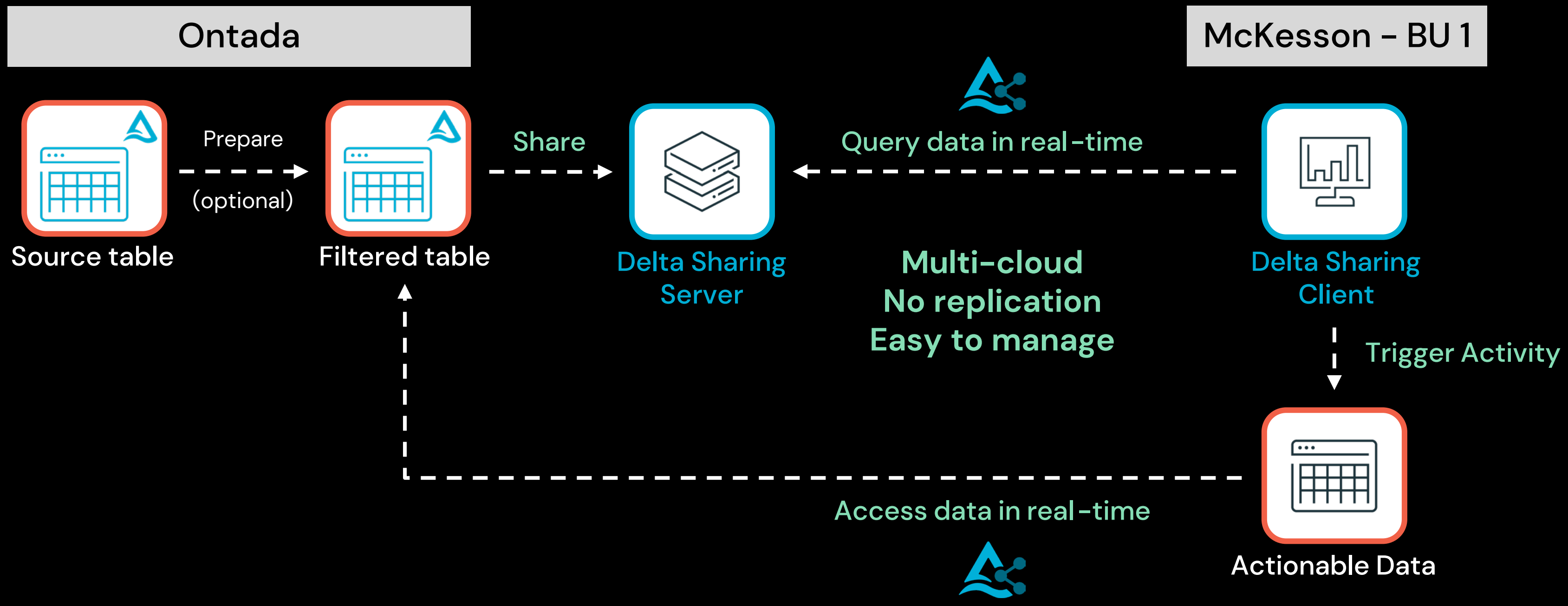
Latency Challenges

Data Rights Management

Data Disharmony

Streamlined Sharing with Delta Sharing

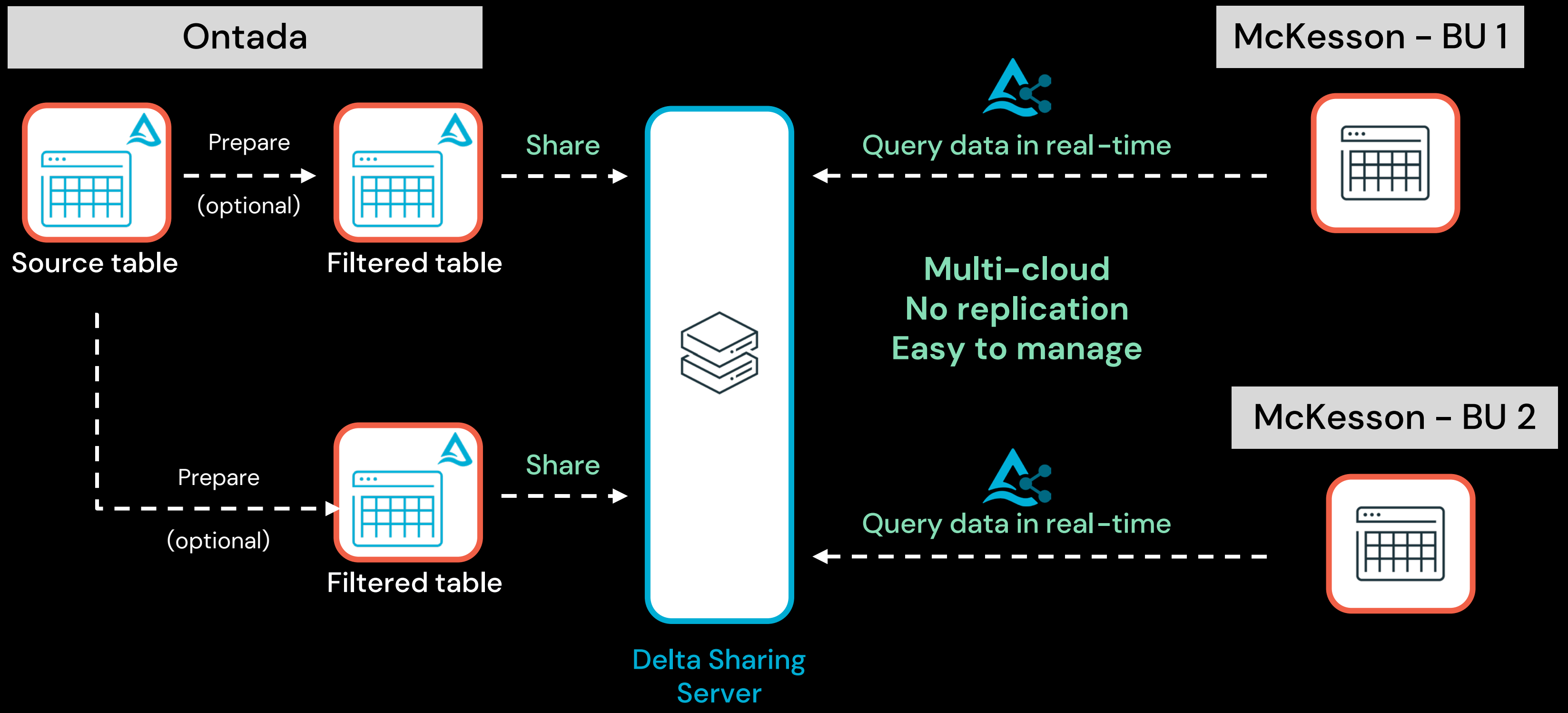
Delta Sharing cuts collaboration time with partners from days to real-time



Multi-cloud
No replication
Easy to manage

Streamlined Sharing with Delta Sharing

Delta Sharing cuts collaboration time with partners from days to real-time



Modernized Data Sharing with Delta Sharing

- Simplified ETL
- Reliable
- Timeliness of Data Delivery

- Centralized Auditing
- Data Rights Management
- Cost efficiency

Future State – Delta Sharing

Clean Rooms (Preview)

External Collaboration

Integration with Partners – Beyond Databricks

Data Marketplace

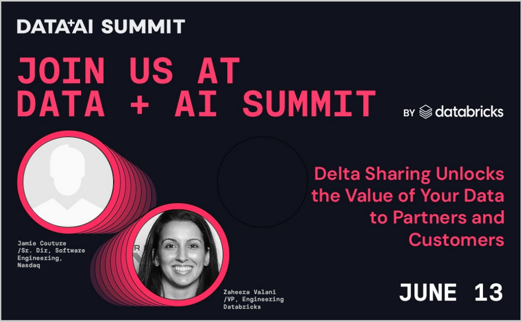
Accelerate Collaboration and Innovation

Q&A

Other sessions on enterprise collaboration

Breakout

- **Atlassian:** Data Mesh and Compliance in a Multi-Regional Data Lake at Atlassian
- **T-Mobile:** Delta Sharing and Unity Catalog — Lessons Learned at T-Mobile
- **Shell:** AI and the Lakehouse: Shell’s Journey Towards Effective Data Governance (Thursday, Jun 13 | 12:30 PM - 1:10 PM PDT | West, Level 2, Rm 2004)
- **Nasdaq**: Delta Sharing Unlocks the Value of Your Data to Partners and Customers (Thursday, Jun 13 | 2:50 PM - 3:30 PM PDT | South, Level 3, Rm 302)



Deep dive

- Best Practices for Architecting Data Collaboration At-Scale Across Clouds, Regions, Platforms (Thursday, Jun 13 | 4:00 PM - 5:30 PM PDT | South, Level 3, Rm 302)