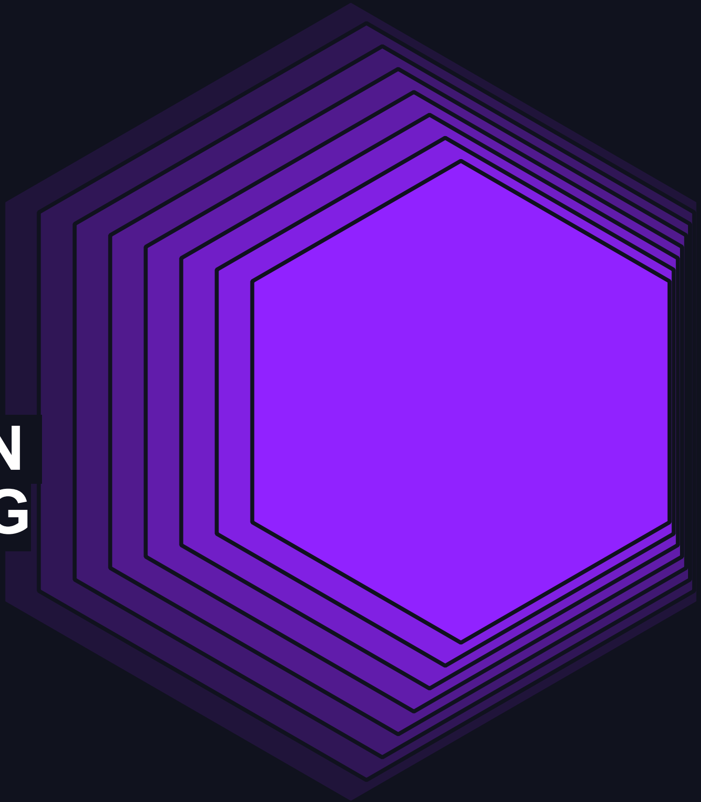
DATA+AI SUMMIT BY databricks

LAKEHOUSE FEDERATION:

DISCOVER, QUERY AND GOVERN ANY DATA WITH UNITY CATALOG



Can Efeoglu – Sr. Staff Product Manager
Todd Greenstein – Staff Product Manager
Andrew Li – Senior Software Engineer

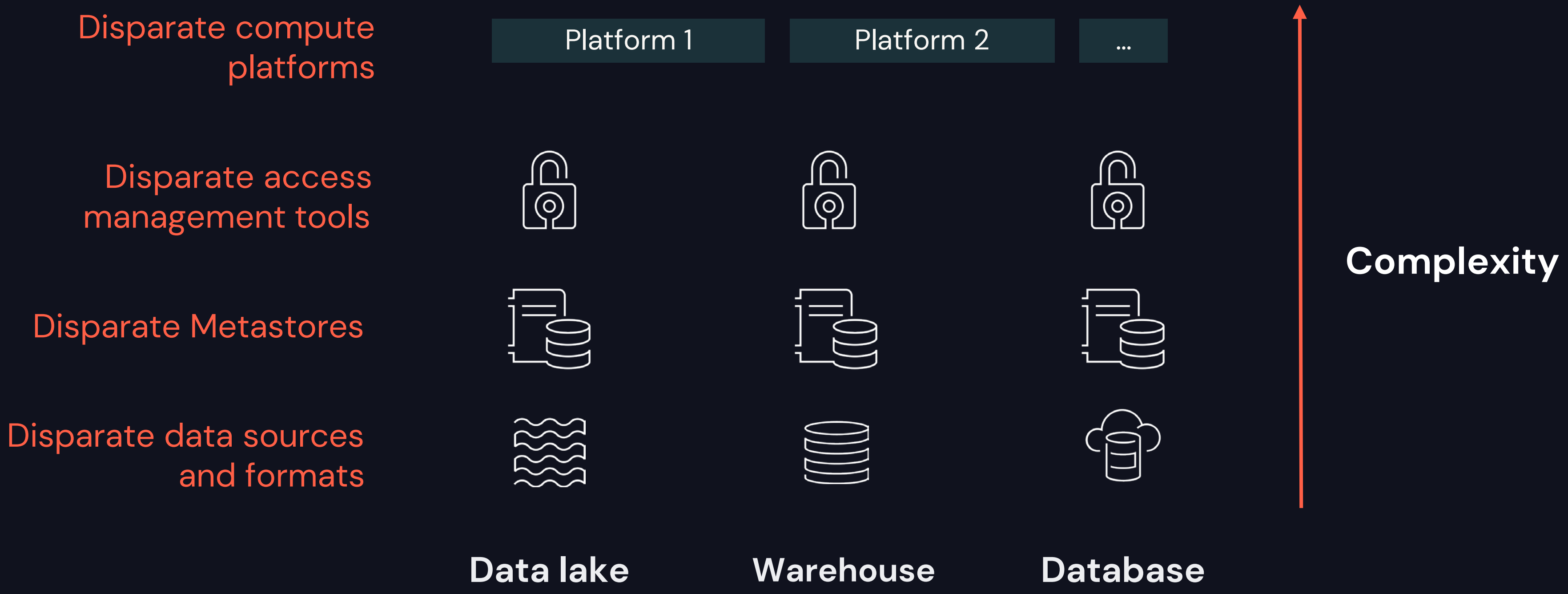
June 13, 2024

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 is scattered across many siloes

Silos cause complexity



Disparate compute platforms

Disparate access management tools

Disparate Metastores

Disparate data sources and formats

Platform 1

Platform 2

...

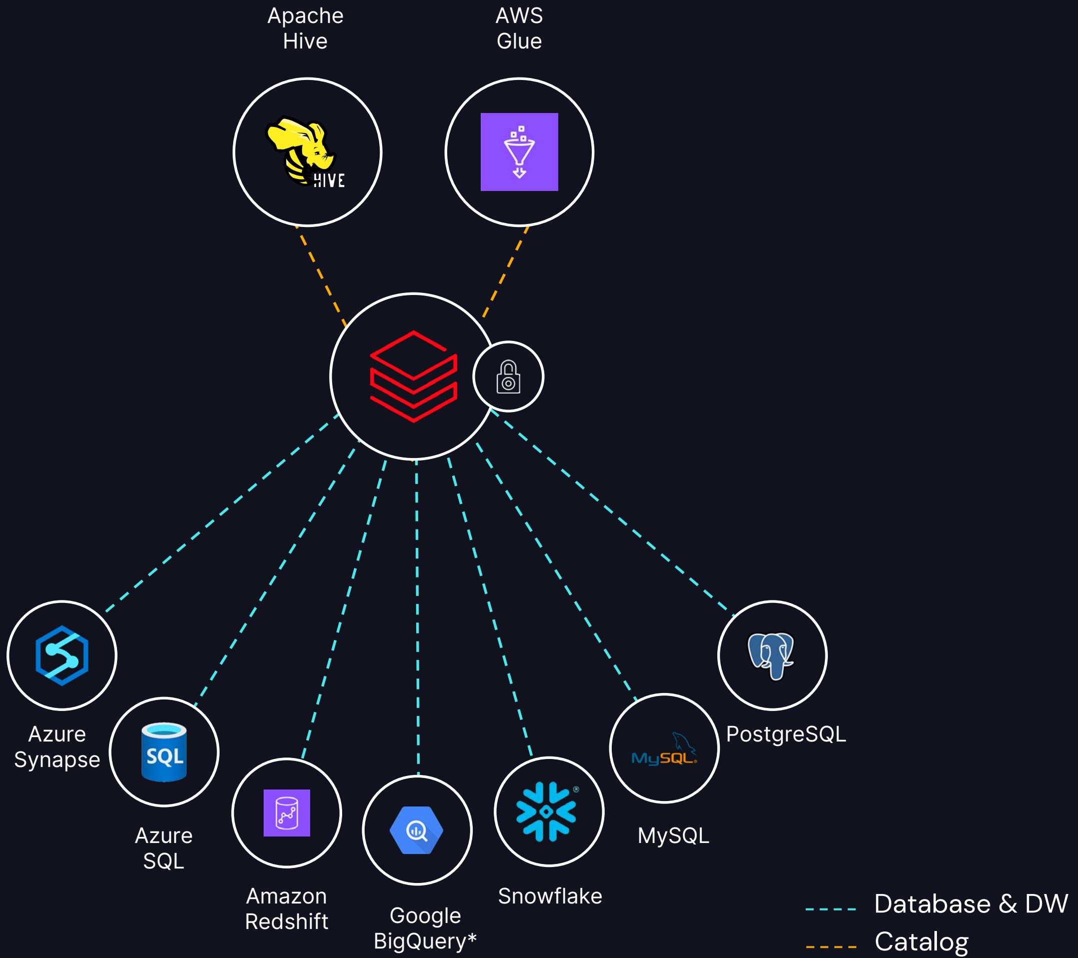
Data lake

Warehouse

Database

Complexity

You need a single point to access and manage all your data



Key benefits

1 **Unified view of all your data**

All users have a common approach to securely discover and explore all data, no matter where it lives.

2 **Unified engine for all data and use cases**

Accelerate ad-hoc analysis and prototyping by querying external data sources for all data, analytics, and AI use cases with a single engine - no ingestion required.

3 **Unified governance across all data sources**

One permission model for the entire data estate provides unified data governance with built-in data lineage and auditability.

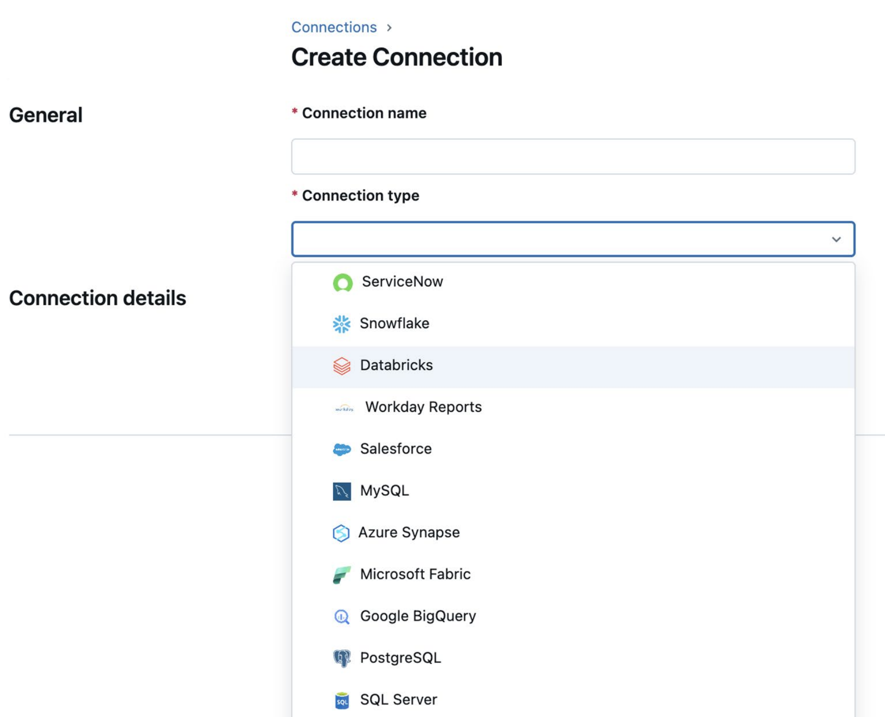
Database & DW Federation

Quick Recap

Central Credential Management

Use shared credentials to set up data source connections in UC

```
CREATE CONNECTION <foreign connection name>
    TYPE <connection type e.g. postgres>
    OPTIONS (
         host <host>,
         port <port>,
         user <username>,
         password <password>
    );
```



Connections ›

Create Connection

General

* Connection name

* Connection type

Connection details

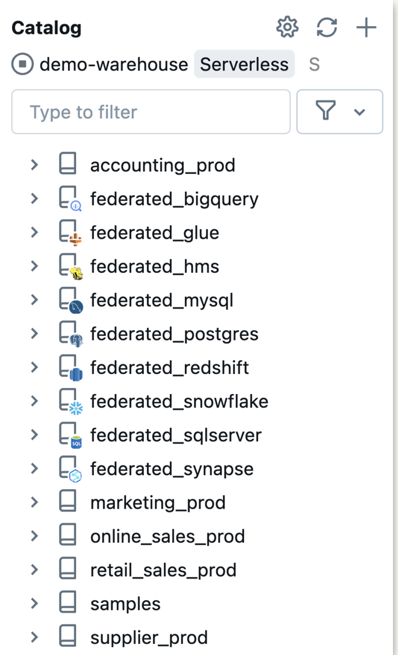
Automatic Metadata Mirroring

Create a foreign catalog in UC pointing to a connection

Data sources

```
CREATE FOREIGN CATALOG <catalog_name>
USING CONNECTION <connection name>
OPTIONS (
  database <db>,
  ...
);

SELECT *
FROM <foreign_catalog>.<schema>.<t>
```



Catalog

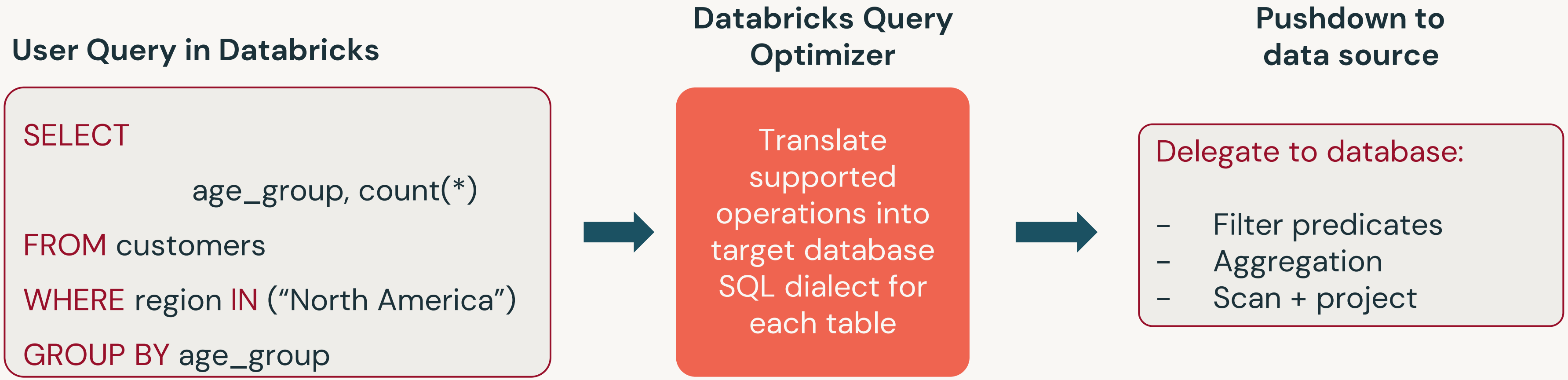
demo-warehouse Serverless S

Type to filter

- accounting_prod
- federated_bigquery
- federated_glue
- federated_hms
- federated_mysql
- federated_postgres
- federated_redshift
- federated_snowflake
- federated_sqlserver
- federated_synapse
- marketing_prod
- online_sales_prod
- retail_sales_prod
- samples
- supplier_prod

Smart Pushdowns

A simple example



User Query in Databricks

```
SELECT
              age_group, count(*)
FROM customers
WHERE region IN ("North America")
GROUP BY age_group
```

Databricks Query Optimizer

Translate supported operations into target database SQL dialect for each table

Pushdown to data source

Delegate to database:

- Filter predicates
- Aggregation
- Scan + project

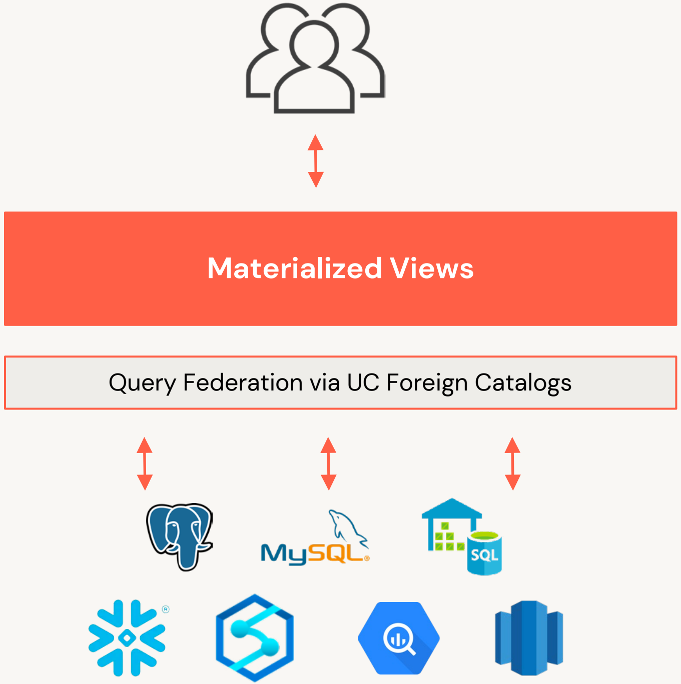
For each foreign table, maximize pushdowns for optimal query performance.

Federation & Materialized Views

Accelerating federated workloads

Federation 💛 Materialized Views:

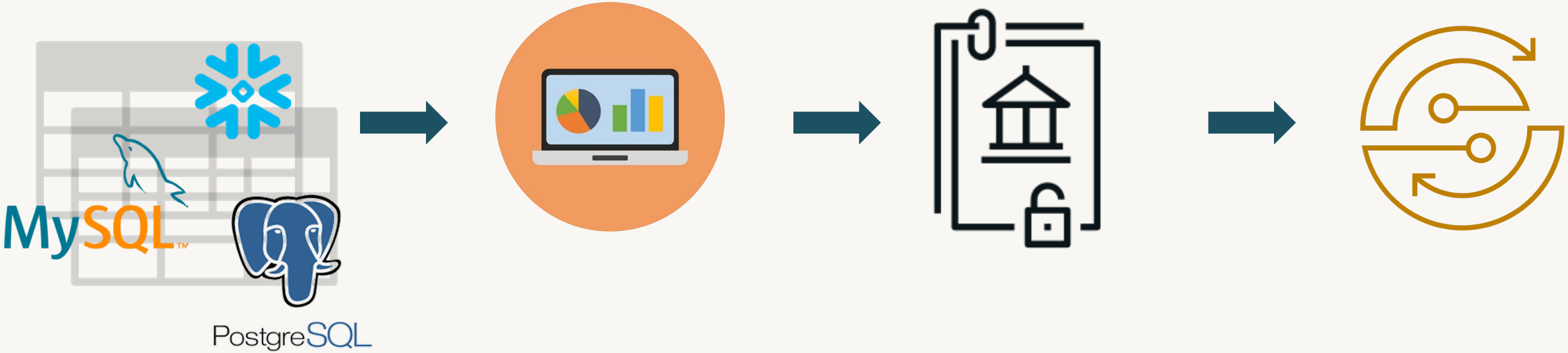
- **Consistent latency & concurrency** for data outside of the lakehouse
- **Accelerate cross-source joins and complicated transformation** logic
- **Offload access to underlying databases** via materialized views to avoid high/concurrent loads on operational databases.



Materialized Views

Query Federation via UC Foreign Catalogs

What are customers doing?



DISCOVER

Register data from **external data sources** in Unity Catalog

QUERY

Get **one single view** and securely explore **the most complete data** for all use cases – no ingestion needed

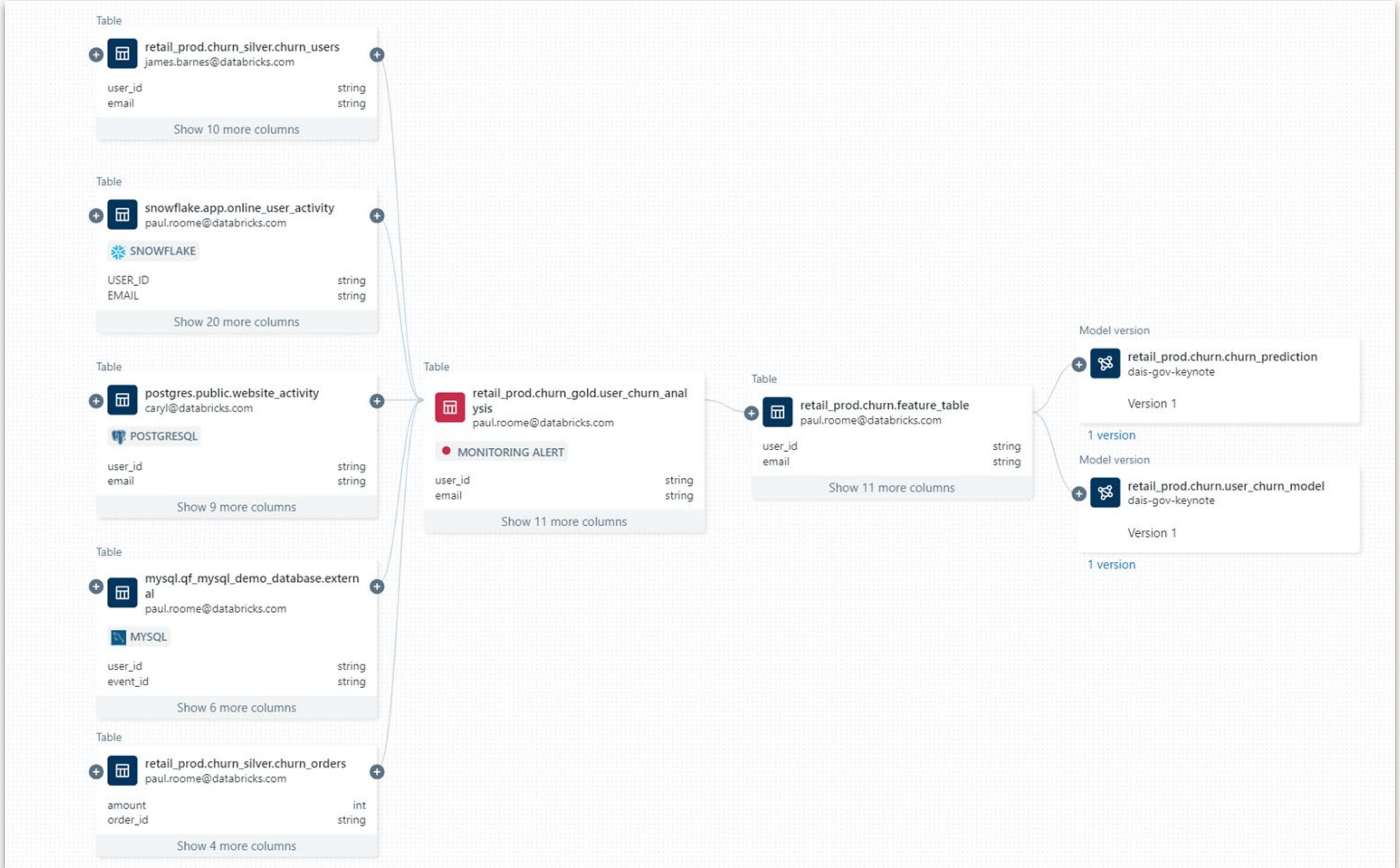
GOVERN

Get **data lineage, insights, auditing, permissions,** all from Unity Catalog

PRODUCTIONIZE

Improve speed & performance by using **Materialized Views** or building **pipelines** as needed

Already a core part of Databricks pipelines



Table

retail_prod.churn_silver.churn_users
james.barnes@databricks.com

user_id string
email string

Show 10 more columns

Table

snowflake.app.online_user_activity
paul.roome@databricks.com

SNOWFLAKE

USER_ID string
EMAIL string

Show 20 more columns

Table

postgres.public.website_activity
caryl@databricks.com

POSTGRESQL

user_id string
email string

Show 9 more columns

Table

mysql.qf_mysql_demo_database.external
paul.roome@databricks.com

MYSQL

user_id string
event_id string

Show 6 more columns

Table

retail_prod.churn_silver.churn_orders
paul.roome@databricks.com

amount int
order_id string

Show 4 more columns

Table

retail_prod.churn_gold.user_churn_analysis
paul.roome@databricks.com

● MONITORING ALERT

user_id string
email string

Show 11 more columns

Table

retail_prod.churn.feature_table
paul.roome@databricks.com

user_id string
email string

Show 11 more columns

Model version

retail_prod.churn.churn_prediction
dais-gov-keynote

Version 1

1 version

Model version

retail_prod.churn.user_churn_model
dais-gov-keynote

Version 1

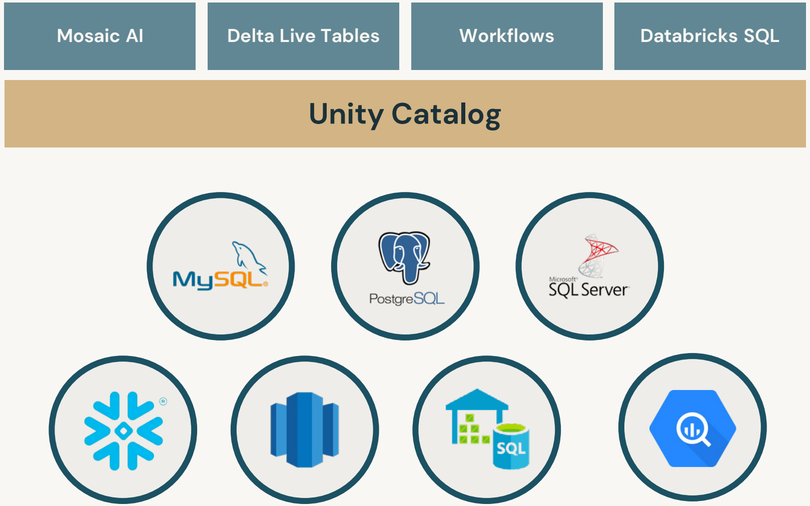
1 version

Announcing General Availability

General Availability – All Clouds

Database & DW Federation GA:

- Improved performance for connectors
- Enhanced security for Snowflake & Azure ecosystem connectors
- More data sources to connect to (Preview): Google BigQuery and Salesforce Data Cloud



Mosaic AI

Delta Live Tables

Workflows

Databricks SQL

Unity Catalog

*BigQuery GA Q3

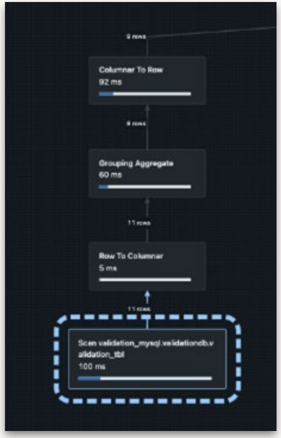
Performance

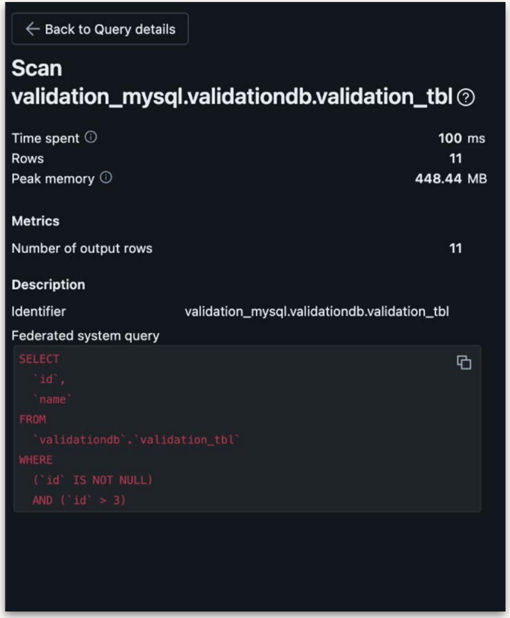
Improved pushdown coverage & performance

Expanded pushdown reliability for SQL Server, Postgres, MySQL, Snowflake, Redshift & Synapse.

Pushdown Query Profiles

As part of Query Profiles, you can now view the every query pushed down to source systems, as well as query execution metrics.





Scan validation_mysql.validationdb.validation_tbl

Time spent	100 ms
Rows	11
Peak memory	448.44 MB

Metrics

Number of output rows	11

Description

Identifier	validation_mysql.validationdb.validation_tbl

Federated system query

```
SELECT
  `id`,
  `name`
FROM
  `validationdb`.`validation_tbl`
WHERE
  (`id` IS NOT NULL)
  AND (`id` > 3)
```

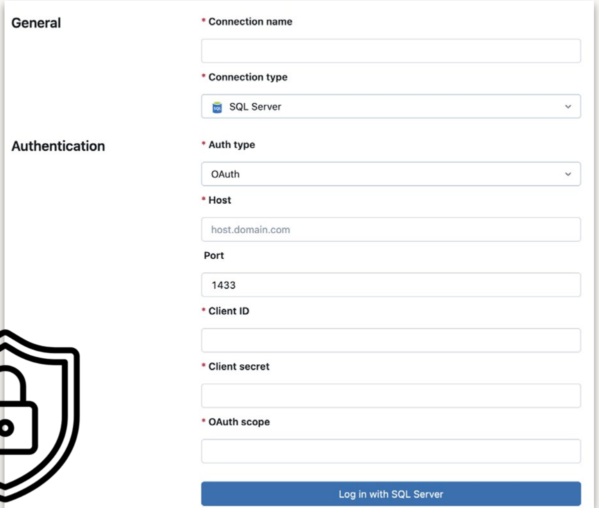
Security

Azure AD support

Azure ecosystem connections:
Synapse & SQL Server/Azure SQL

Snowflake OAuth support

Securely connect to your Snowflake instance using OAuth. Authentication experience built directly into Unity Catalog UI.



General

* Connection name

* Connection type

SQL Server

Authentication

* Auth type

OAuth

* Host

host.domain.com

Port

1433

* Client ID

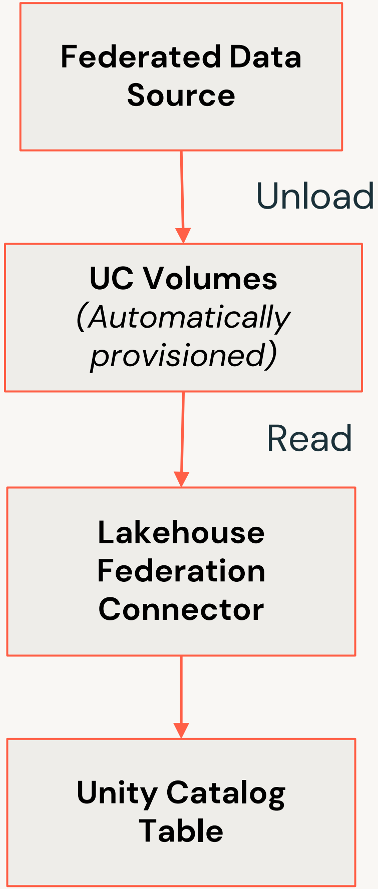
* Client secret

* OAuth scope

Looking Ahead

High Throughput Data Transfer

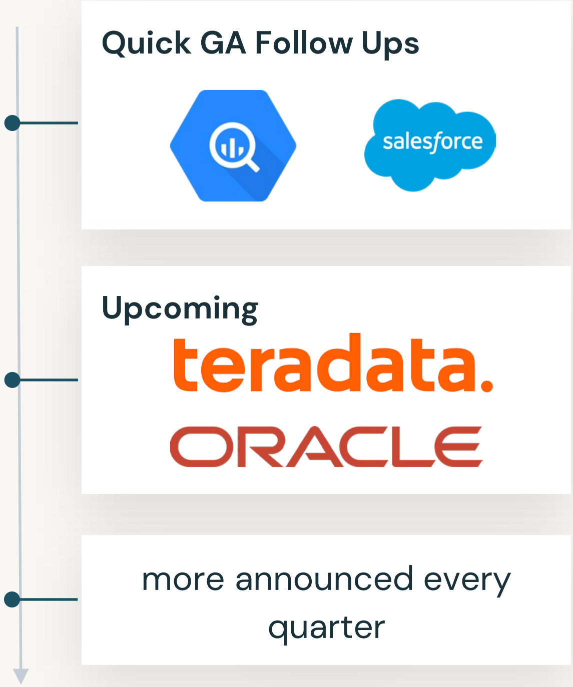
Simple access & ingest of large tables via SQL


Federated Data Source
Unload
UC Volumes *(Automatically provisioned)*
Read
Lakehouse Federation Connector
Unity Catalog Table

Automatic high-throughput data transfers:

- **Increased throughput transfer** by leveraging object storage unloading
- **Transparent to end-user**
- Lakehouse federation **automatically provisions and manages** UC volumes
- **Snowflake & Redshift** as first data sources to support.

More Connectors



Quick GA Follow Ups

salesforce

Upcoming

teradata.

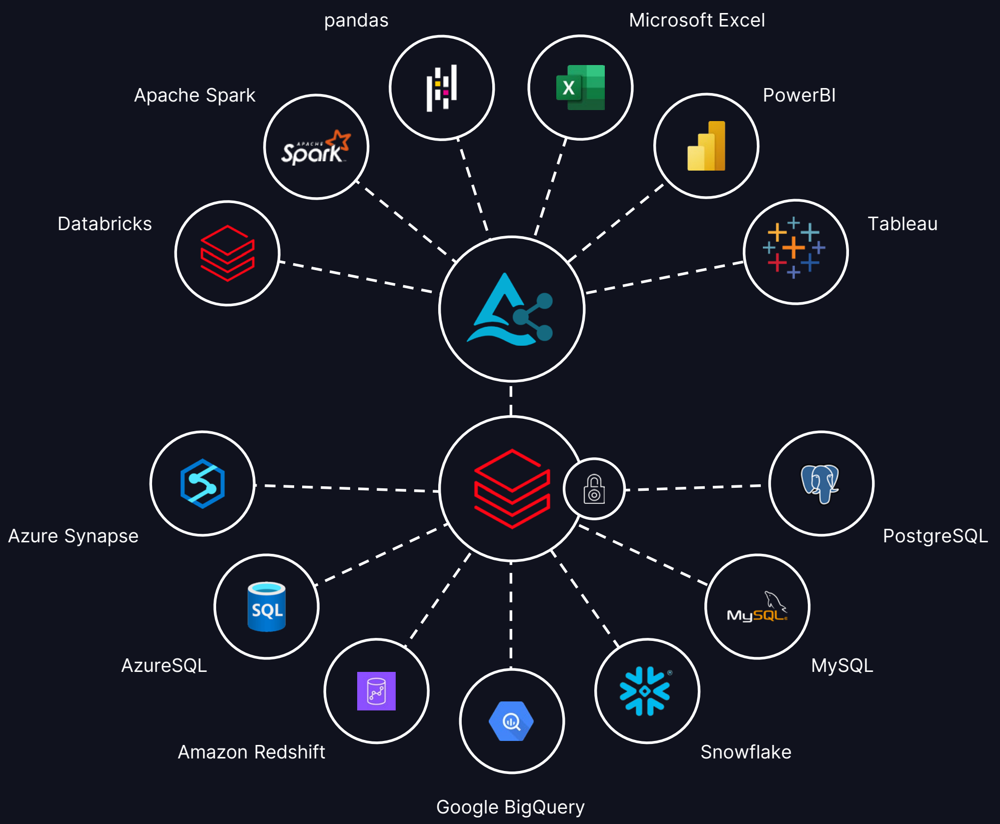
ORACLE

more announced every quarter

ANNOUNCING

Sharing for Lakehouse Federation

Share data from any database without ETL



Catalog Federation:
Hive Metastore & AWS Glue

HMS Federation

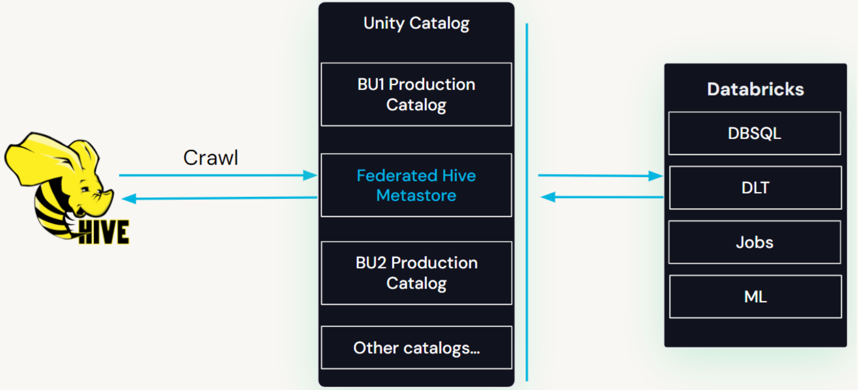
Simplify migration from Hive Metastore (HMS), transparent access.

Discover, govern and access data from internal/external HMS, Glue

Mount external Hive Metastore or Glue as foreign catalog in Unity Catalog

Simple and straightforward upgrade process to Unity Catalog

Continued transparent access to an external Metastore



In Private Preview

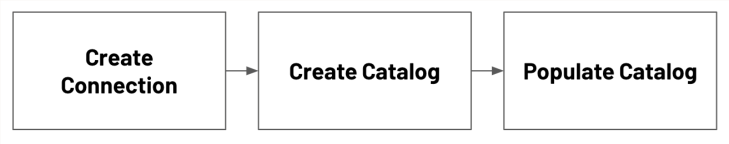
Why?

Lakehouse Federation gives us the perfect framework for giving you access to all your data

- **Reduce friction** for customers onboarding and/or migrating to UC.
- Give everyone using UC **access to all their existing data**.
- Transparent access to external Metastores.
- Foreign catalog data sources means:
  - **First class citizens in UC – *Just like any other catalog in UC***
  - **End to end Governance** including ABAC/FGAC
  - **Full audibility** for all workloads

How does it work?

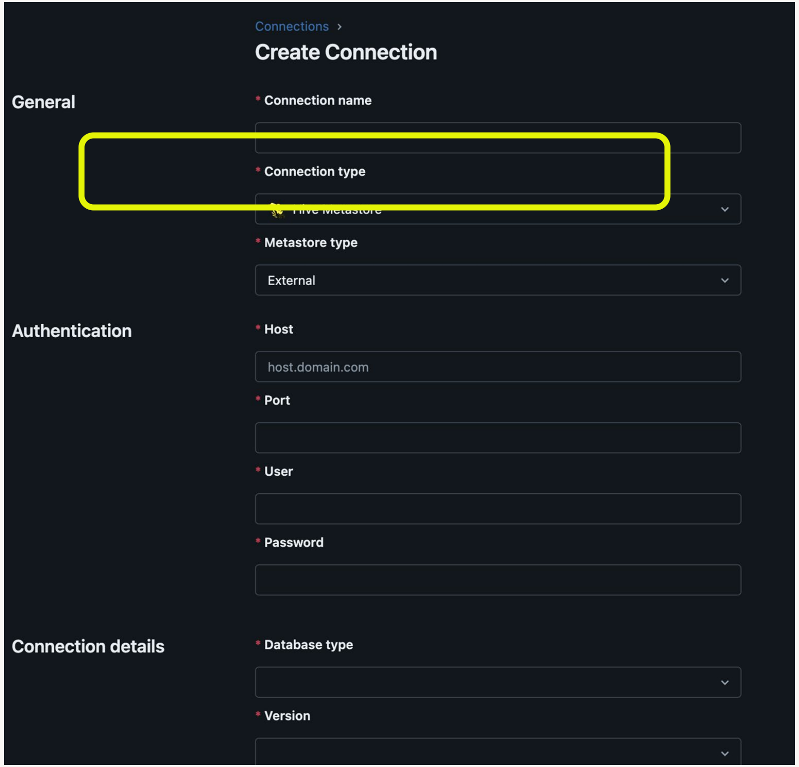
How?



Lakehouse Federation FTW!

- Very easy to plug new federation targets into Unity. *(setup a connection object and go!)*

- J-I-T Metadata means you are never querying stale metadata.



Connections ›

Create Connection

General

* Connection name

* Connection type

Hive Metastore

* Metastore type

External

Authentication

* Host

host.domain.com

* Port

* User

* Password

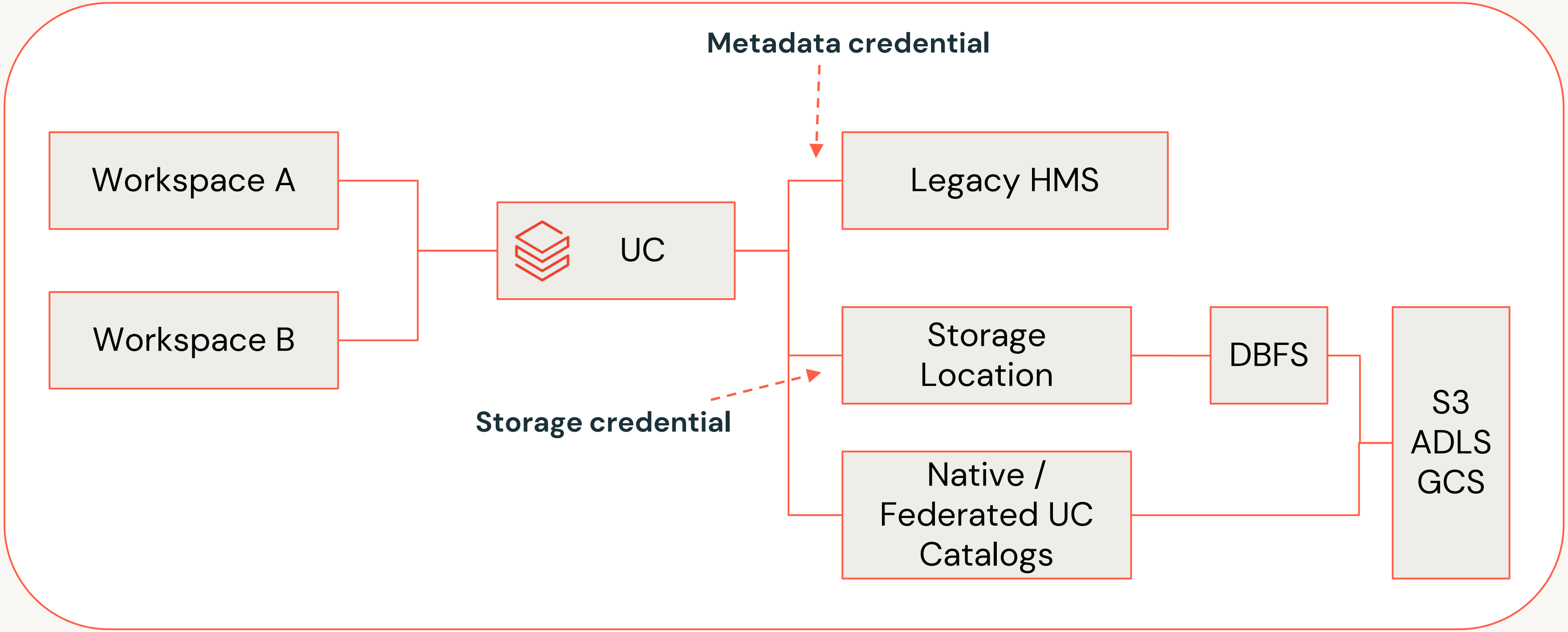
Connection details

* Database type

* Version

Catalog Federation

Use Case – Federate to Internal Workspace or external Hive Metastore/Glue



Demo - Campaign Analytics

Setup Experience

UC Native Experience

Consume: SQL, Dashboards & Genie

Administration

Catalog Federation

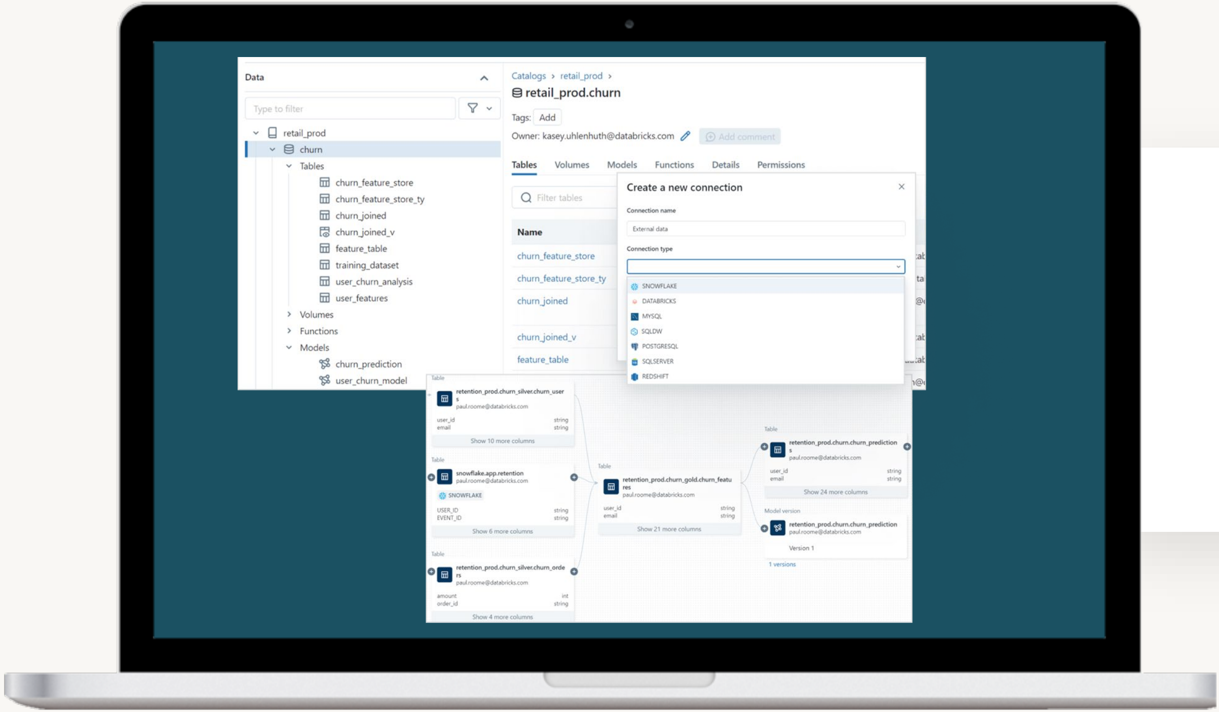
Wrap Up

Database & DW
General Availability

Catalogs
Coming soon

Explore Unity Catalog

Unified governance for all your data, analytics, and AI



Owner: kasey.uhlenhuth@databricks.com

Connection name

External data

Connection type

Explore Databricks Unity Catalog

databricks.com/unity

Demos

databricks.com/demos

Learn more at the summit!

Databricks
Events App

Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.

What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!

Get trained and certified

- Visit the Learning Hub Experience at Moscone West, 2nd Floor!
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! databricks.com/learn