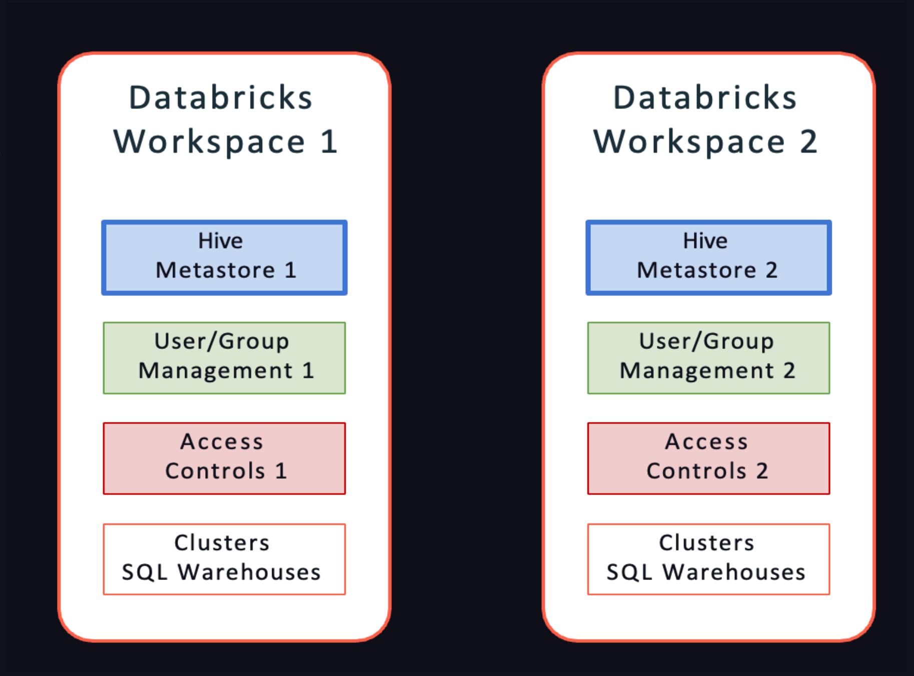
DATA+AI SUMMIT BY databricks

Empowering Centralized Data Governance @Comcast with Unity Catalog

Problem Statement

- Fragmented data governance & access control on 150+ Workspaces
- Inefficient data management and utilization
- Increased security risks and potential data breaches
- Difficulty in collaboration and data sharing
- Wasted storage costs due to data redundancy
- Increased administrative overhead for teams



Databricks Workspace 1

Hive Metastore 1

User/Group Management 1

Access Controls 1

Clusters SQL Warehouses

Databricks Workspace 2

Hive Metastore 2

User/Group Management 2

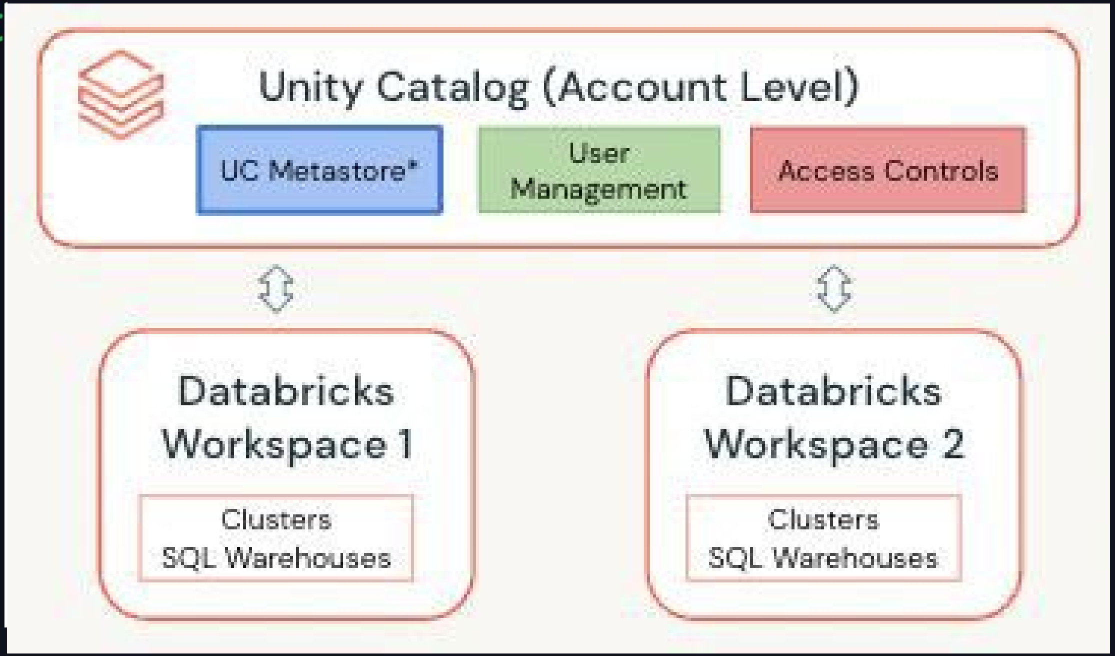
Access Controls 2

Clusters SQL Warehouses

Databricks Unity Catalog

Unified governance for data, analytics and AI

- Centralized Identity and access management
- Simplified administration
- Centralized governance for data and AI
- Enhanced auditing and data sharing
- Built-in data search and discovery
- Performance and scale
- Automated lineage for all workloads


Unity Catalog (Account Level)

UC Metastore*

User Management

Access Controls

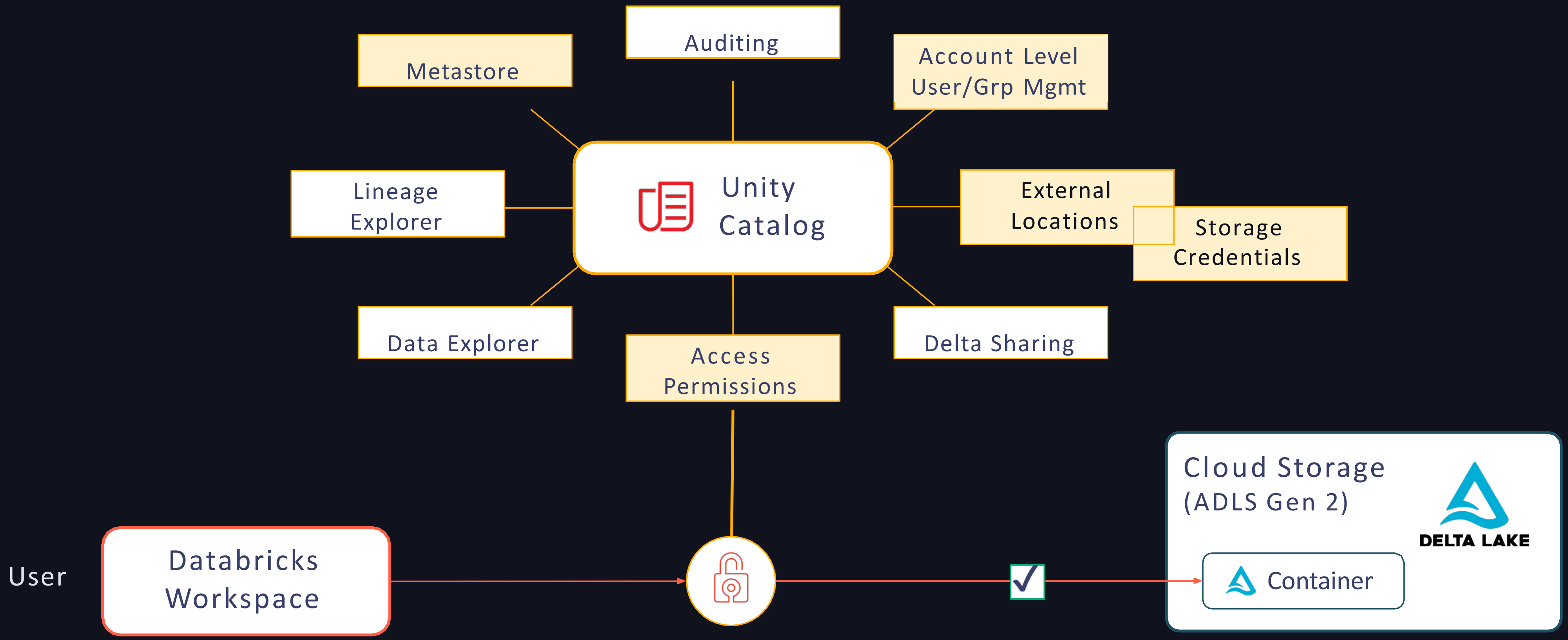
Databricks Workspace 1

Clusters
SQL Warehouses

Databricks Workspace 2

Clusters
SQL Warehouses

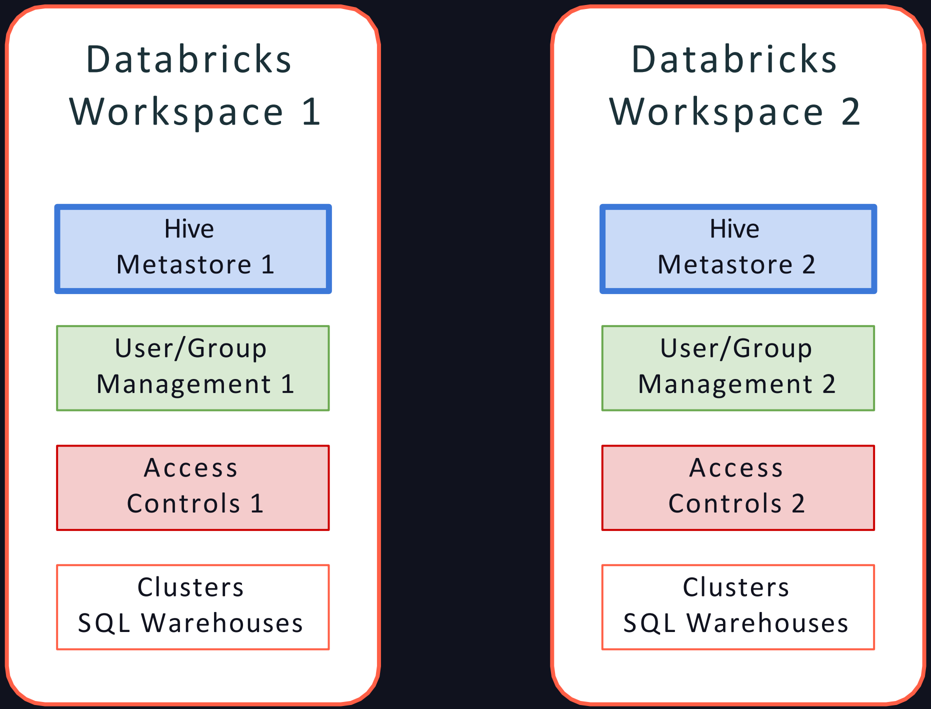
Unity Catalog - Architecture



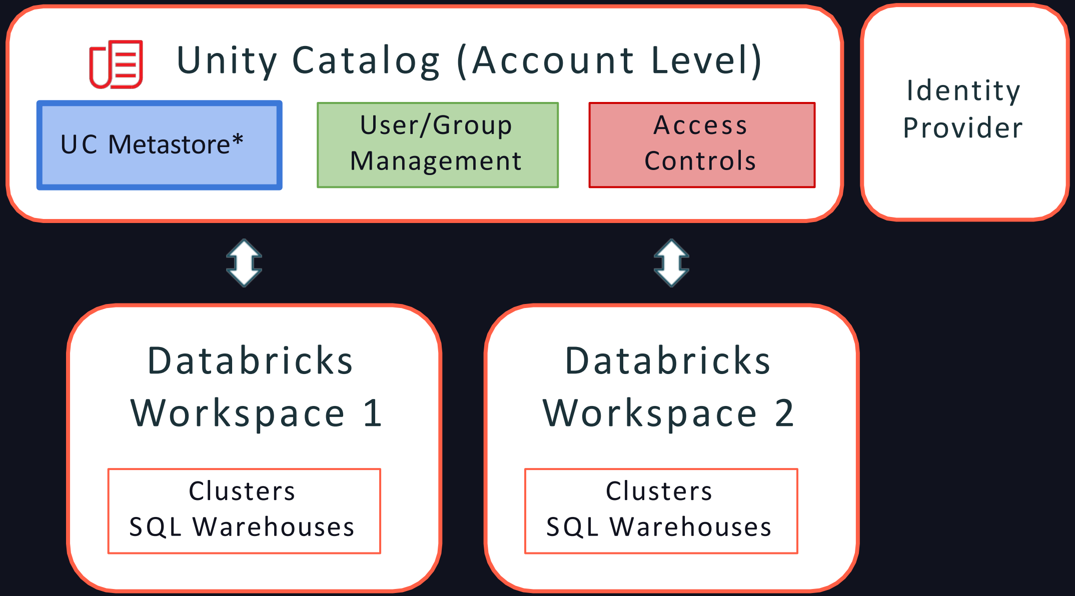
Access Management & Collaboration

Centralized Metadata, Identity, and Access Management

Without Unity Catalog



With Unity Catalog



*** Important: only one UC metastore per region!**

Identity Federation Enrollment before UC

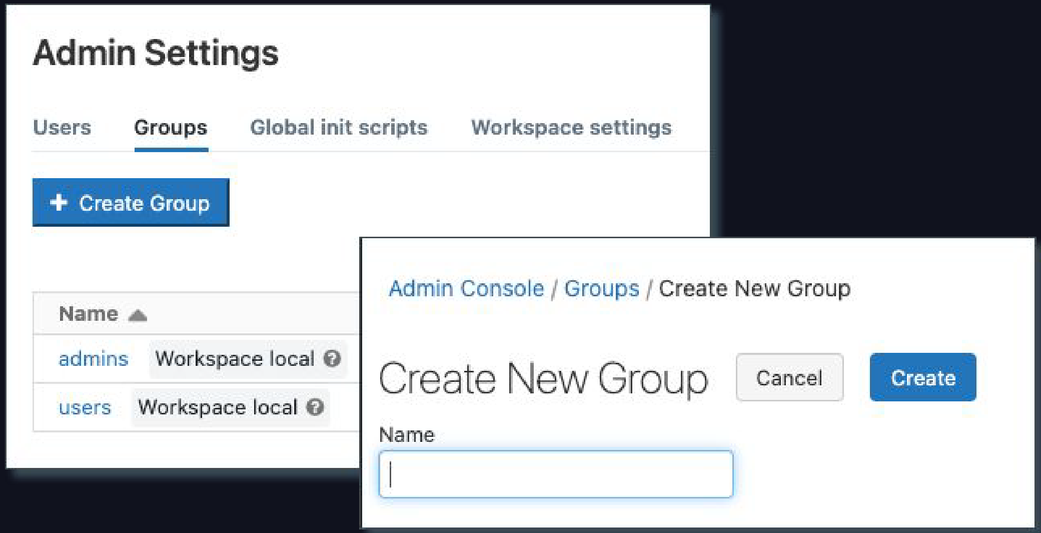
Group administration changes for Workspaces before UC

Before

- Users - created in the **Workspace**
- Groups - created in the **Workspace** with a **Create Group** button
  - These are now referred to as "Workspace local" groups

Challenge: duplicate administration work across workspaces

Before



Admin Settings

admins Workspace local
users Workspace local

Create New Group

Name

Identity Federation Enrollment with UC

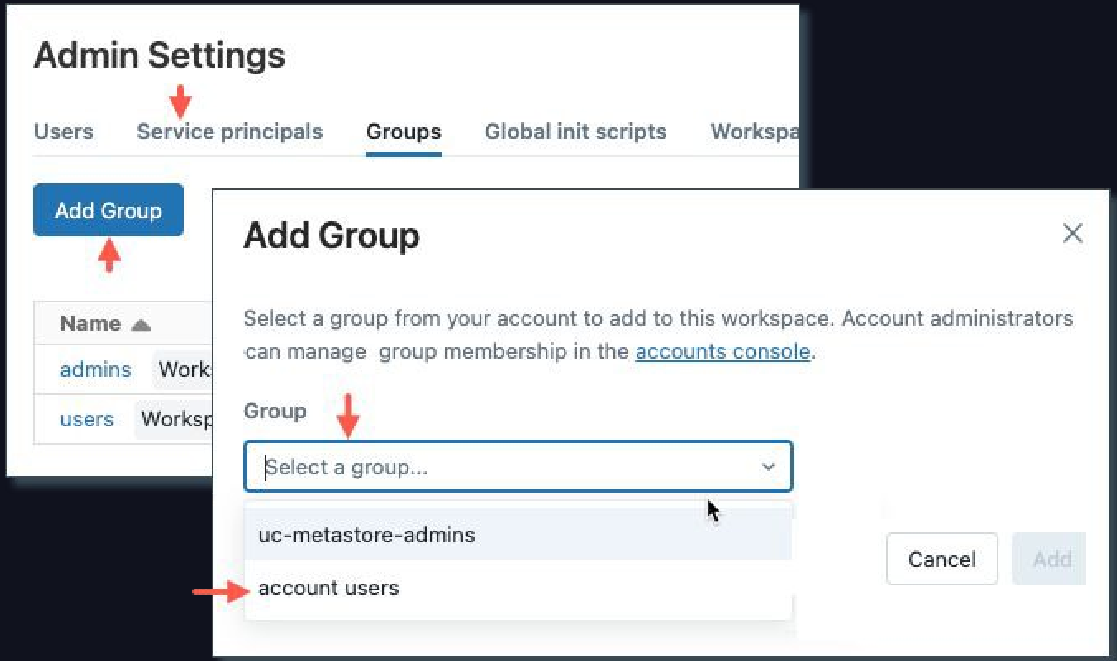
Group administration changes for Workspaces connected to UC

After

- Users - created in the Workspace **or at the Account level** - users created in the Workspace are synced to the account
- Groups - created **exclusively in the Account Console** and are *added* to the Workspace (**Add Group** button)
    - New **account users** group
- New Service principals menu

Simplified administration

After



Admin Settings

Users | Service principals | Groups | Global init scripts | Workspa

Add Group

Name	
admins	Work
users	Worksp

Add Group

Select a group from your account to add to this workspace. Account administrators can manage group membership in the accounts console.

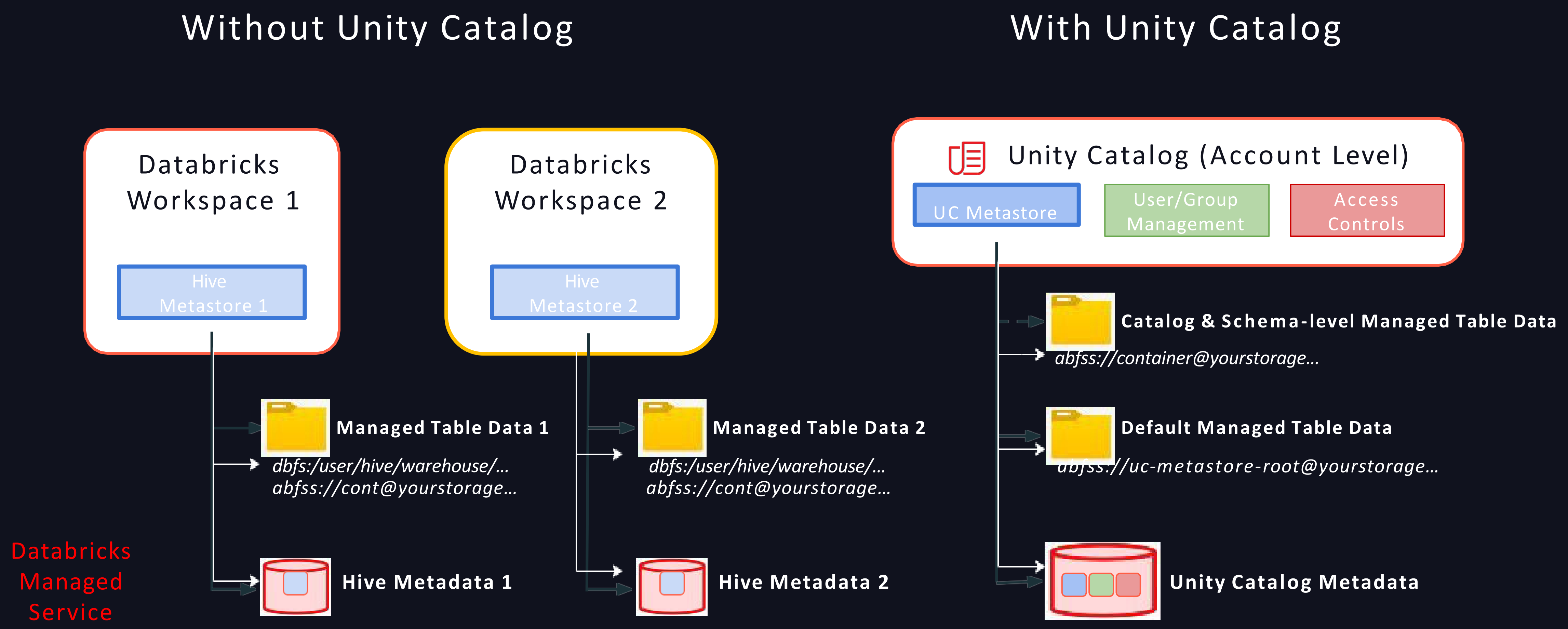
Group

Select a group...

- uc-metastore-admins
- account users

Cancel | Add

Hive vs UC metadata


Without Unity Catalog
With Unity Catalog
Databricks Workspace 1
Hive Metastore 1
Databricks Workspace 2
Hive Metastore 2
Managed Table Data 1
dbfs:/user/hive/warehouse/...
abfss://cont@yourstorage...
Managed Table Data 2
dbfs:/user/hive/warehouse/...
abfss://cont@yourstorage...
Hive Metadata 1
Hive Metadata 2
Databricks Managed Service
Unity Catalog (Account Level)
UC Metastore
User/Group Management
Access Controls
Catalog & Schema-level Managed Table Data
abfss://container@yourstorage...
Default Managed Table Data
abfss://uc-metastore-root@yourstorage...
Unity Catalog Metadata

Access Control Simplified

Centrally grant and manage access permissions across all workloads

Using ANSI Standard SQL

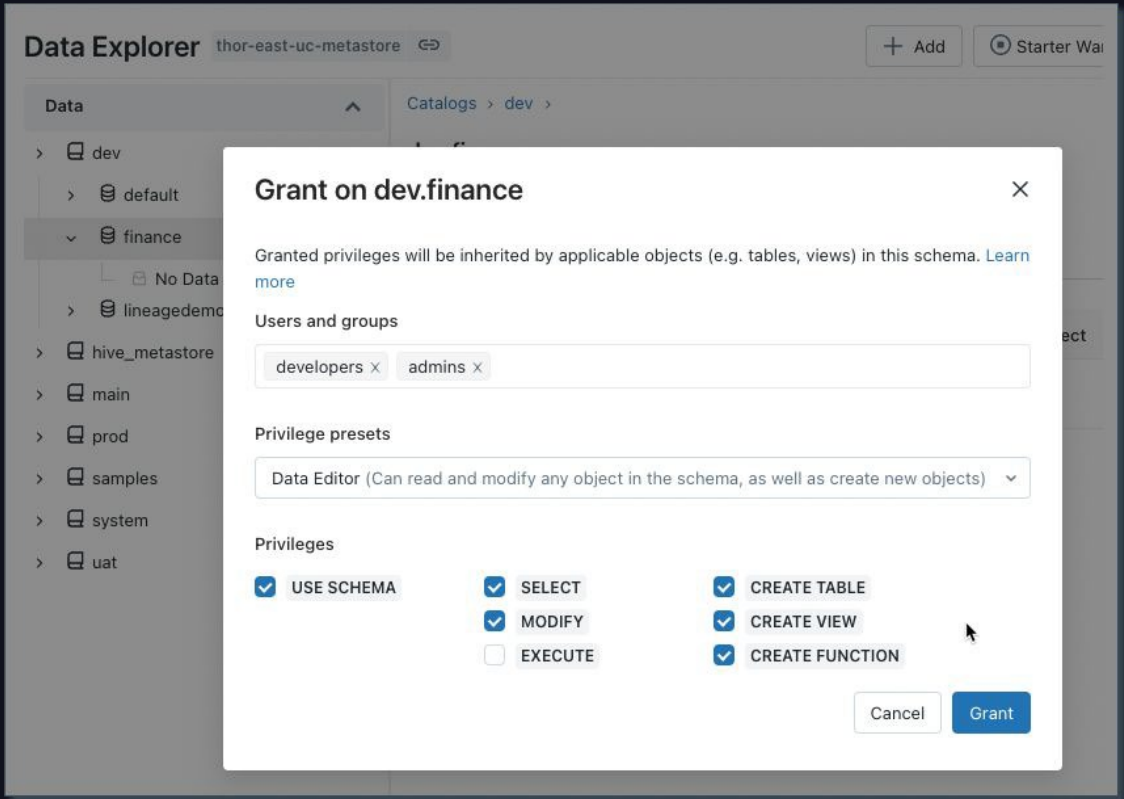
GRANT SELECT ON iot.events TO engineers

Choose permission level

Choose resource

Sync groups from identity provider

Using the UI



Grant on dev.finance

Granted privileges will be inherited by applicable objects (e.g. tables, views) in this schema. Learn more

Users and groups

developers × admins ×

Privilege presets

Data Editor (Can read and modify any object in the schema, as well as create new objects)

Privileges

☑ USE SCHEMA

☑ SELECT

☑ MODIFY

☐ EXECUTE

☑ CREATE TABLE

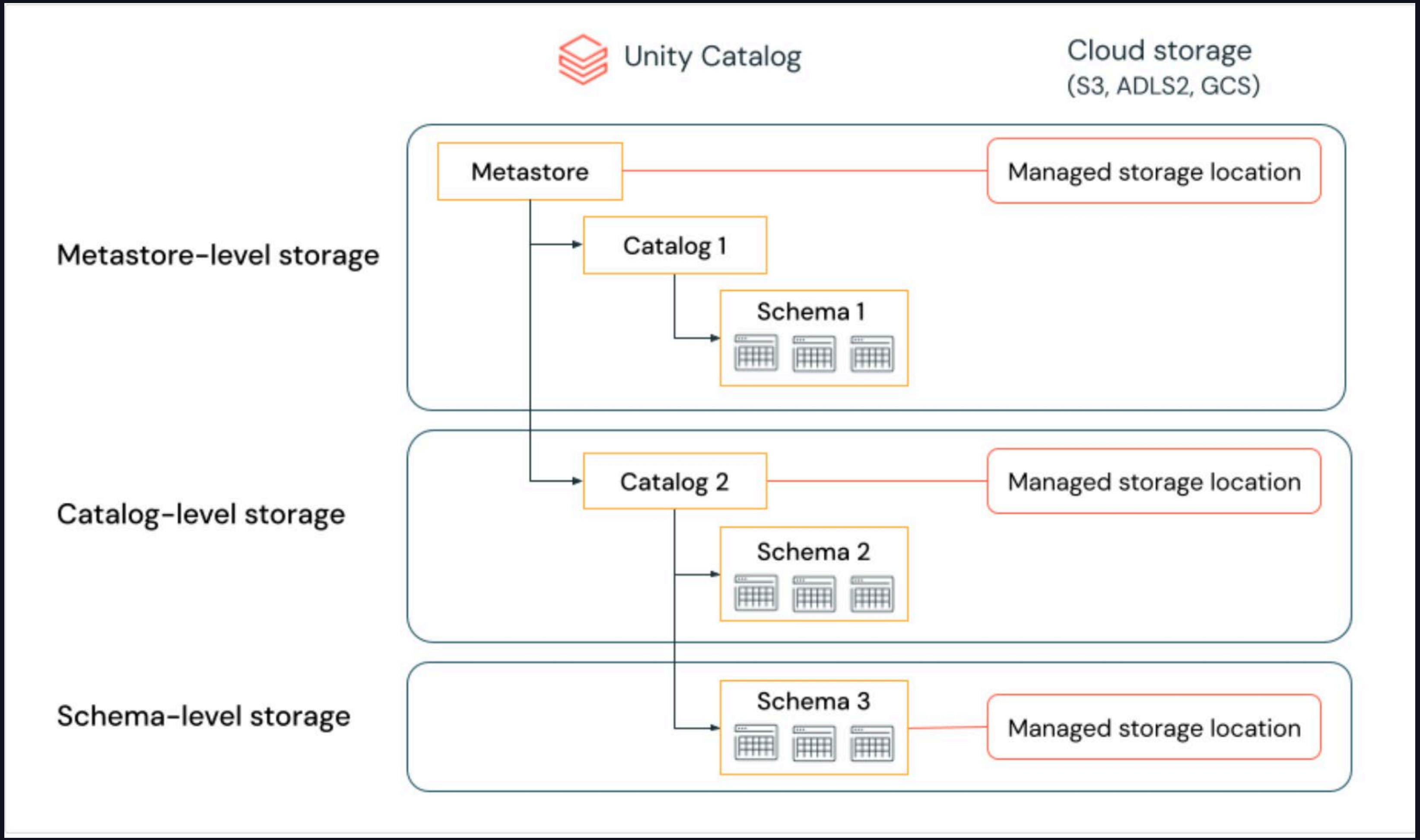
☑ CREATE VIEW

☑ CREATE FUNCTION

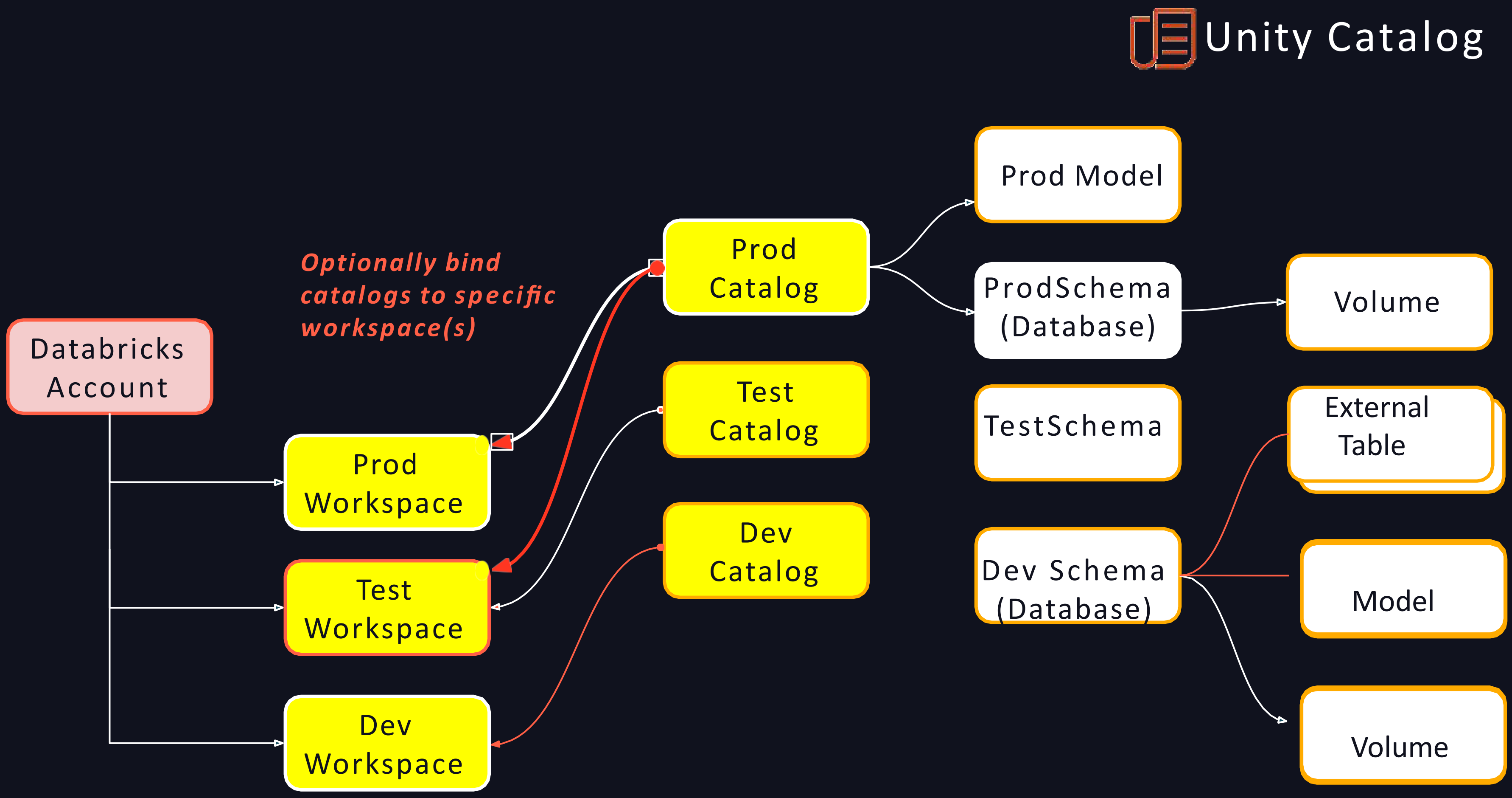
Cancel Grant

Data Isolation with UC at Comcast

Storage based isolation

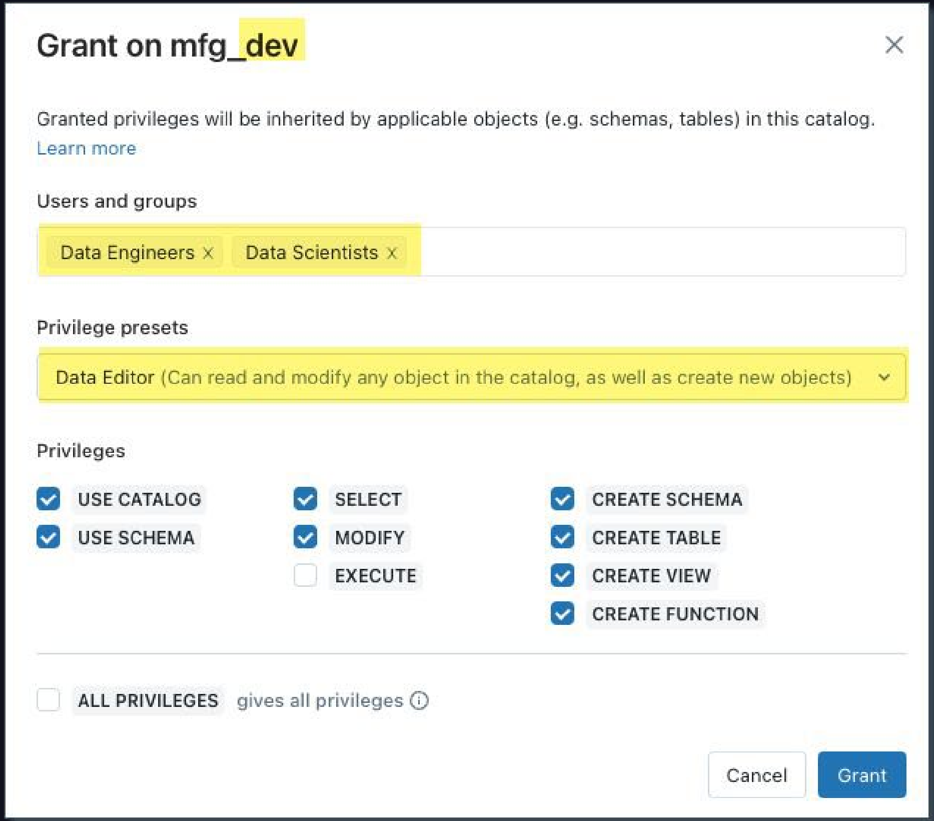


Workspace based isolation



User based isolation

Assign group, user, and service principal permissions as needed



Grant on mfg_dev

Granted privileges will be inherited by applicable objects (e.g. schemas, tables) in this catalog.
Learn more

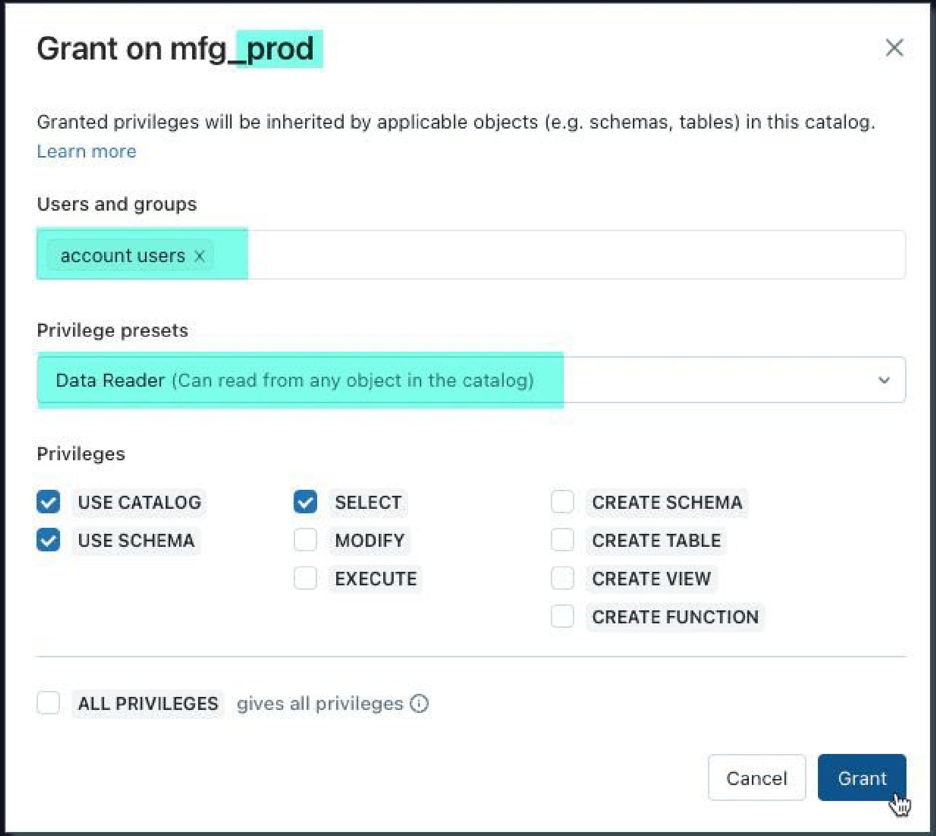
Users and groups: Data Engineers × Data Scientists ×

Privilege presets: Data Editor (Can read and modify any object in the catalog, as well as create new objects)

Privileges

- ☑ USE CATALOG
- ☑ USE SCHEMA
- ☑ SELECT
- ☑ MODIFY
- □ EXECUTE
- ☑ CREATE SCHEMA
- ☑ CREATE TABLE
- ☑ CREATE VIEW
- ☑ CREATE FUNCTION

□ ALL PRIVILEGES gives all privileges



Grant on mfg_prod

Granted privileges will be inherited by applicable objects (e.g. schemas, tables) in this catalog.
Learn more

Users and groups: account users ×

Privilege presets: Data Reader (Can read from any object in the catalog)

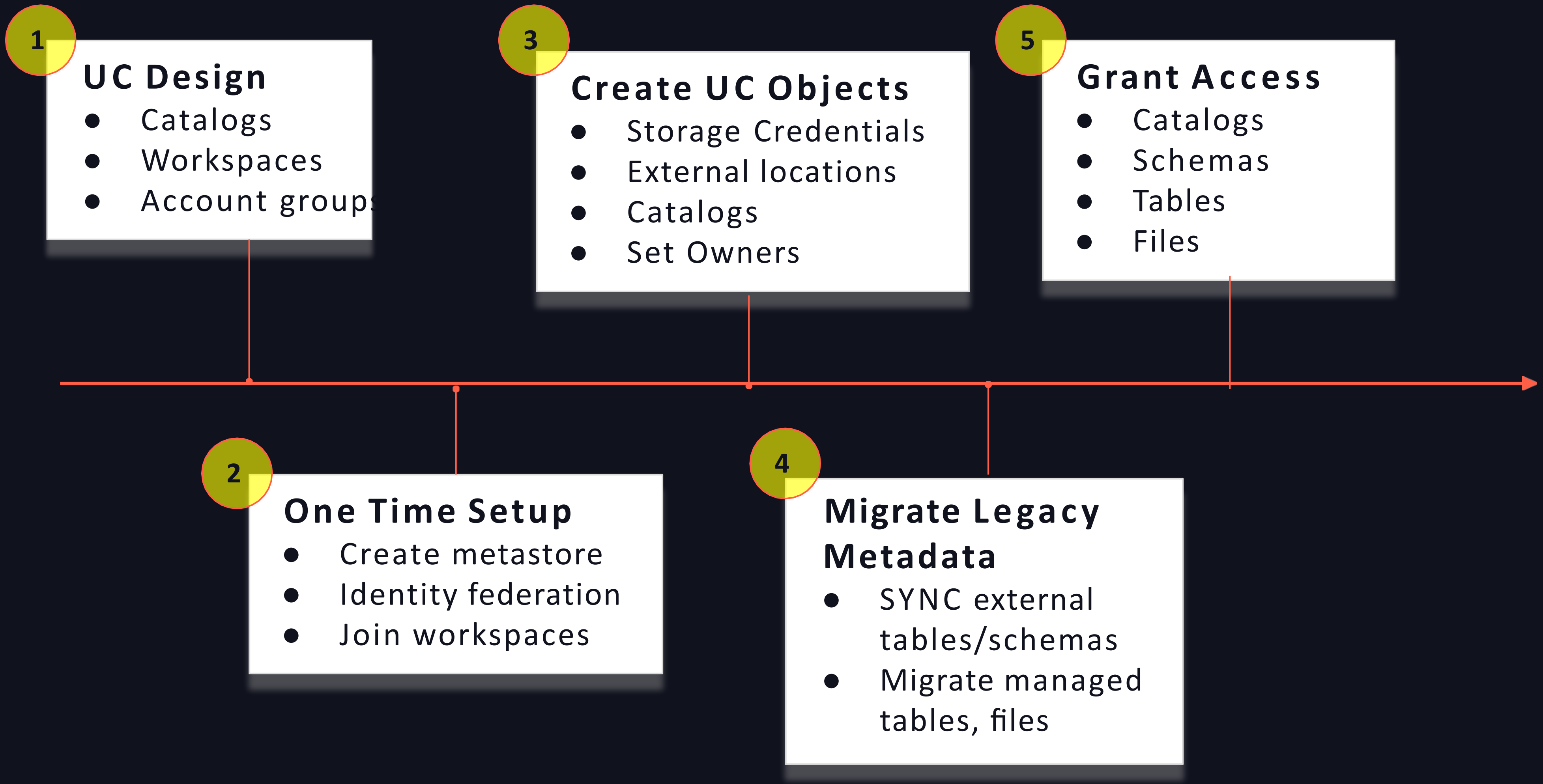
Privileges

- ☑ USE CATALOG
- ☑ USE SCHEMA
- ☑ SELECT
- □ MODIFY
- □ EXECUTE
- □ CREATE SCHEMA
- □ CREATE TABLE
- □ CREATE VIEW
- □ CREATE FUNCTION

□ ALL PRIVILEGES gives all privileges

High Level Roadmap to Unity Catalog

Steps to consider



1 **UC Design**
- Catalogs
- Workspaces
- Account groups

2 **One Time Setup**
- Create metastore
- Identity federation
- Join workspaces

3 **Create UC Objects**
- Storage Credentials
- External locations
- Catalogs
- Set Owners

4 **Migrate Legacy Metadata**
- SYNC external tables/schemas
- Migrate managed tables, files

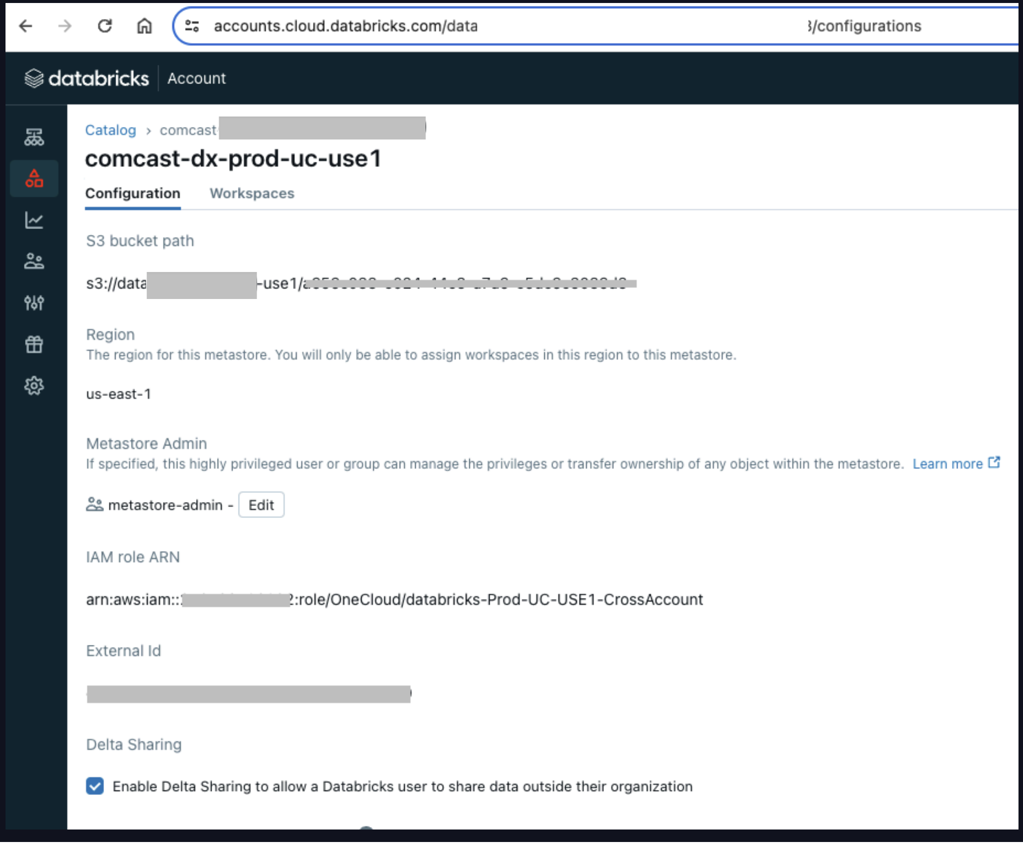
5 **Grant Access**
- Catalogs
- Schemas
- Tables
- Files

Create metastore and join workspaces

2

One Time Setup

- Create metastore
- Identity federation
- Join workspaces



accounts.cloud.databricks.com/data[redacted]3/configurations

Catalog › comcast[redacted]

comcast-dx-prod-uc-use1

S3 bucket path

s3://data[redacted]-use1/[redacted]

Region

The region for this metastore. You will only be able to assign workspaces in this region to this metastore.

us-east-1

Metastore Admin

If specified, this highly privileged user or group can manage the privileges or transfer ownership of any object within the metastore. Learn more

metastore-admin - Edit

IAM role ARN

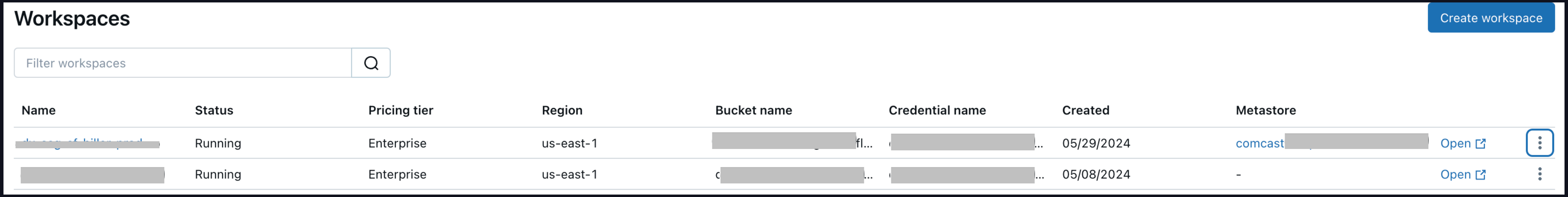
arn:aws:iam::[redacted]:role/OneCloud/databricks-Prod-UC-USE1-CrossAccount

External Id

[redacted]

Delta Sharing

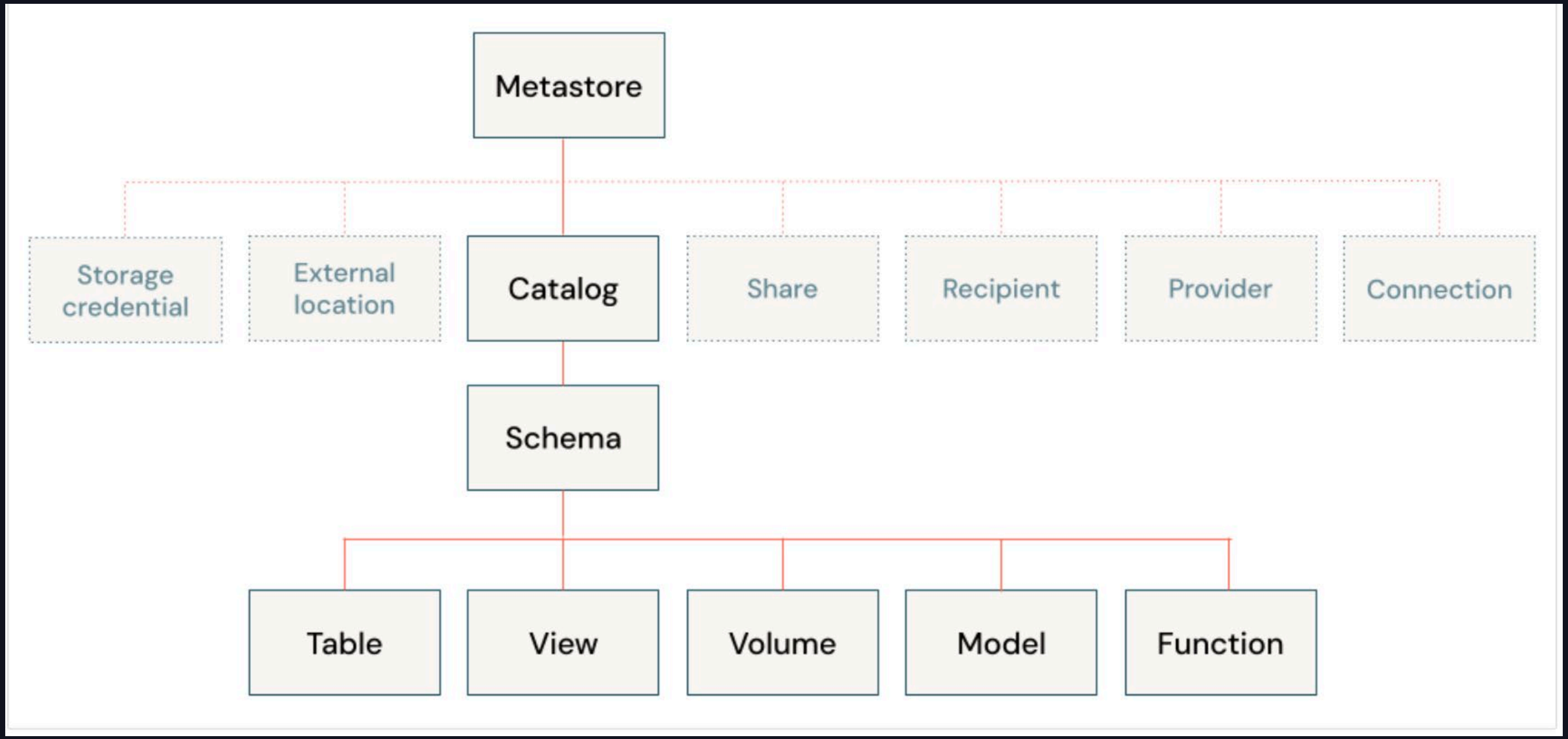
☑ Enable Delta Sharing to allow a Databricks user to share data outside their organization



Workspaces

Name	Status	Pricing tier	Region	Bucket name	Credential name	Created	Metastore		
[redacted]	Running	Enterprise	us-east-1	[redacted]fl...	[redacted]...	05/29/2024	comcast[redacted]	Open	⋮
[redacted]	Running	Enterprise	us-east-1	c[redacted]...	[redacted]...	05/08/2024	-	Open	⋮

UC Object Model

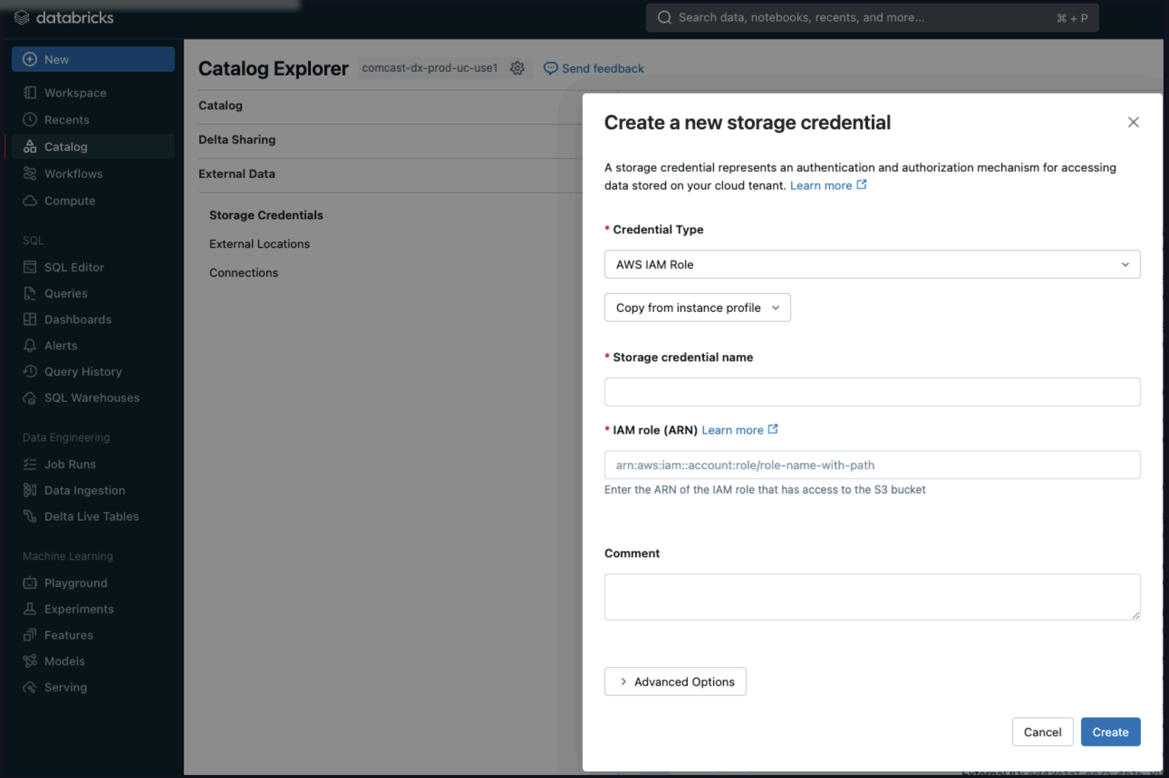


Create UC Objects (Storage Credential, External Locations)

3

Create UC Objects

- Storage Credentials
- External locations
- Catalogs
- Set Owners



Create a new storage credential

A storage credential represents an authentication and authorization mechanism for accessing data stored on your cloud tenant. Learn more

* Credential Type: AWS IAM Role

Copy from instance profile

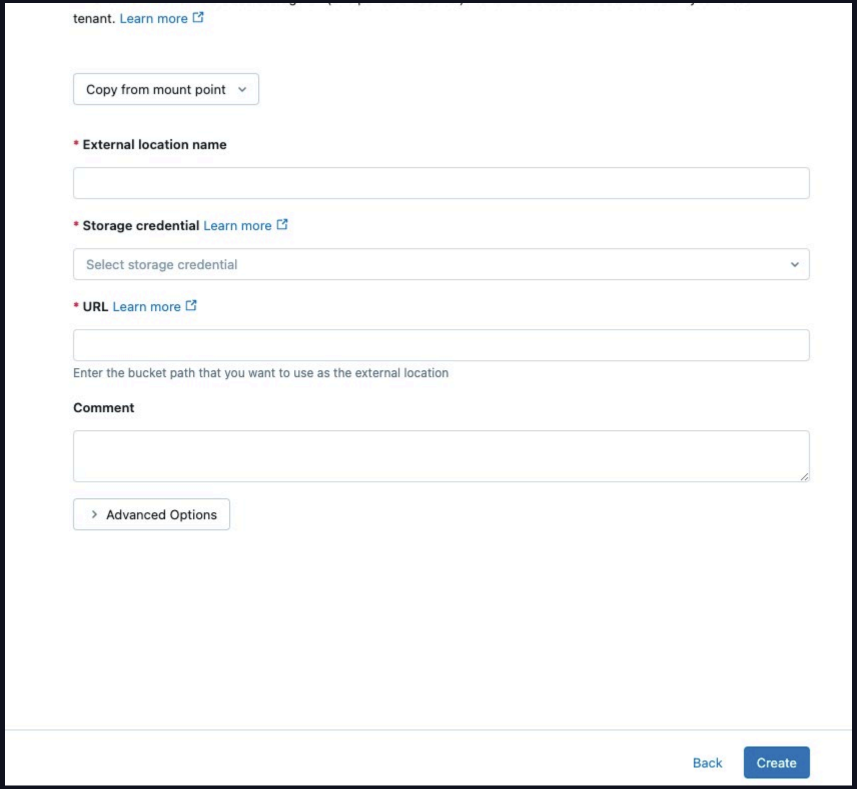
* Storage credential name

* IAM role (ARN) Learn more

arn:aws:iam::account:role/role-name-with-path

Enter the ARN of the IAM role that has access to the S3 bucket

Comment



tenant. Learn more

* External location name

* Storage credential Learn more

Select storage credential

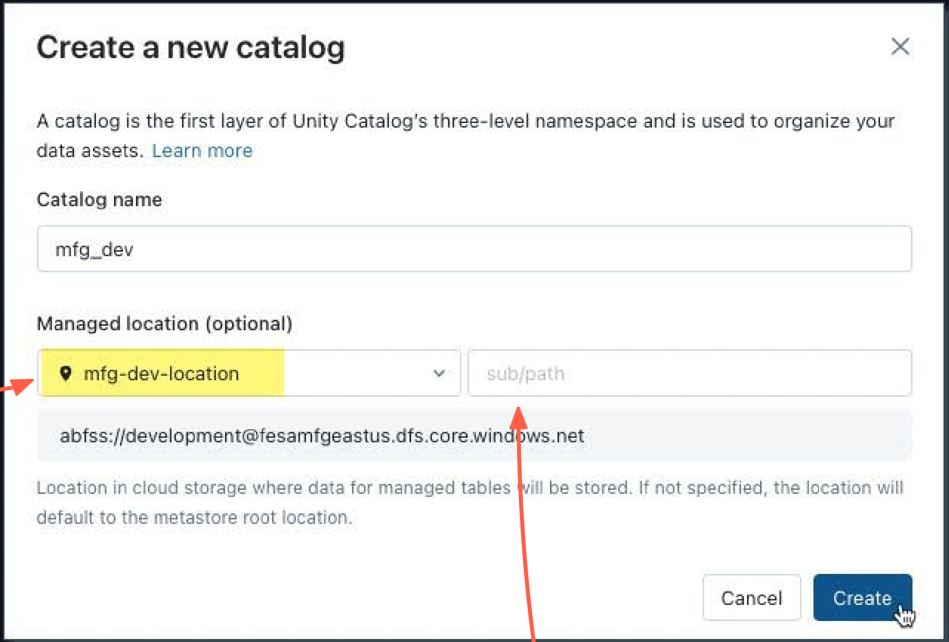
* URL Learn more

Enter the bucket path that you want to use as the external location

Comment

DATA+AI SUMMIT

Managed Location for Catalogs- Storage based data isolation



Create a new catalog

A catalog is the first layer of Unity Catalog's three-level namespace and is used to organize your data assets. Learn more

Catalog name

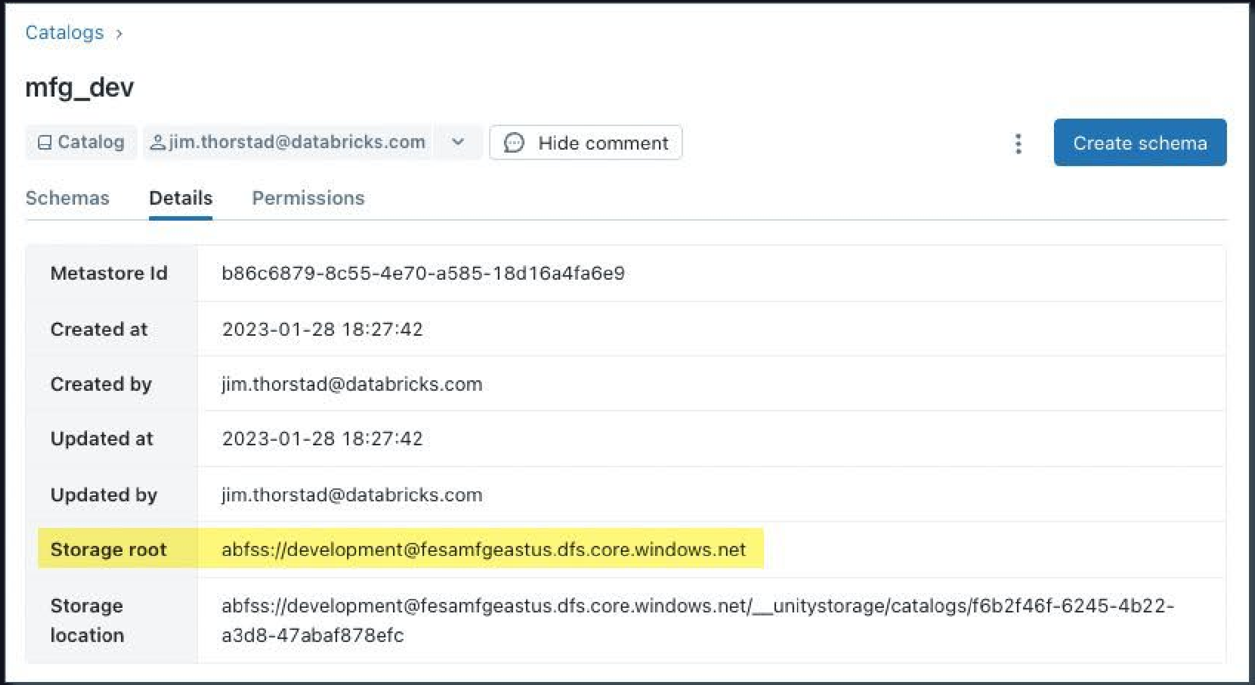
mfg_dev

Managed location (optional)

mfg-dev-location | sub/path

abfss://development@fesamfgeastus.dfs.core.windows.net

Location in cloud storage where data for managed tables will be stored. If not specified, the location will default to the metastore root location.



Catalogs ›

mfg_dev

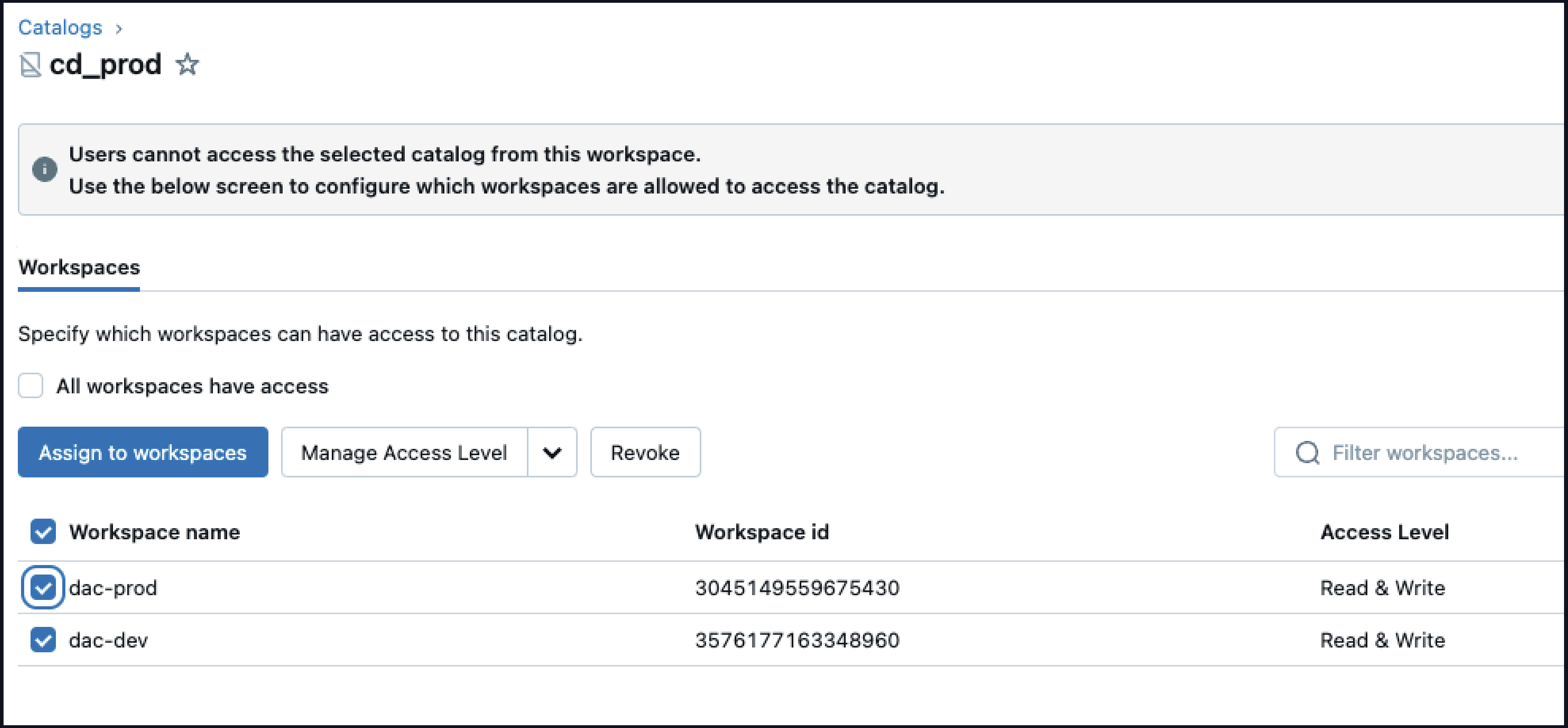
Metastore Id	b86c6879-8c55-4e70-a585-18d16a4fa6e9
Created at	2023-01-28 18:27:42
Created by	jim.thorstad@databricks.com
Updated at	2023-01-28 18:27:42
Updated by	jim.thorstad@databricks.com
Storage root	abfss://development@fesamfgeastus.dfs.core.windows.net
Storage location	abfss://development@fesamfgeastus.dfs.core.windows.net/__unitystorage/catalogs/f6b2f46f-6245-4b22-a3d8-47abaf878efc

Must choose from previously configured External Locations

Optionally include sub folder/path

DATA+AI SUMMIT

Catalog-Workspace binding – env based data isolation



cd_prod ☆

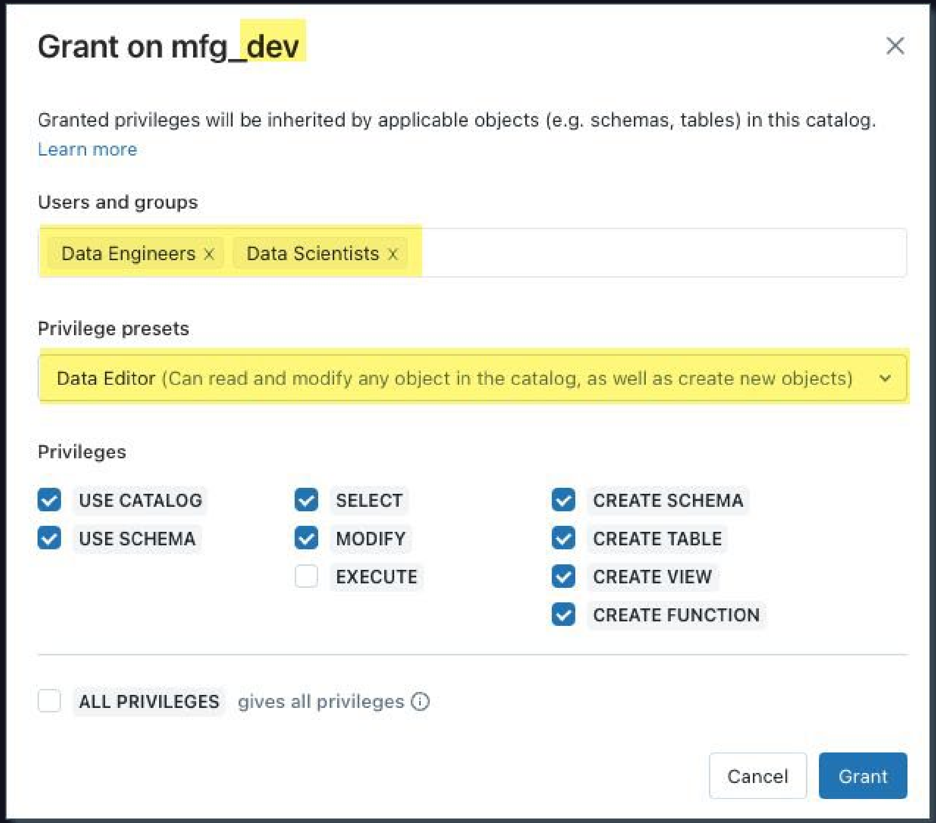
Users cannot access the selected catalog from this workspace.
Use the below screen to configure which workspaces are allowed to access the catalog.

Specify which workspaces can have access to this catalog.

□ All workspaces have access

☑ Workspace name	Workspace id	Access Level
☑ dac-prod	3045149559675430	Read & Write
☑ dac-dev	3576177163348960	Read & Write

Assign group, user & service principal permissions - User based isolation



Grant on mfg_dev

Granted privileges will be inherited by applicable objects (e.g. schemas, tables) in this catalog.
Learn more

Users and groups

Data Engineers x Data Scientists x

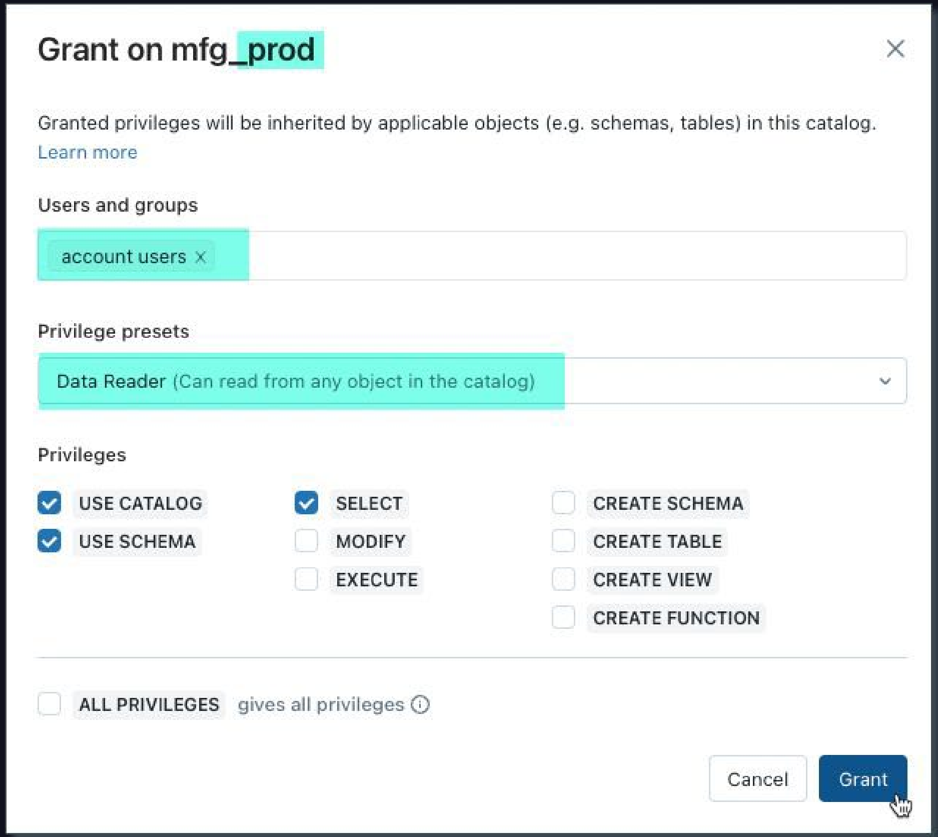
Privilege presets

Data Editor (Can read and modify any object in the catalog, as well as create new objects)

Privileges

☑ USE CATALOG
☑ USE SCHEMA
☑ SELECT
☑ MODIFY
□ EXECUTE
☑ CREATE SCHEMA
☑ CREATE TABLE
☑ CREATE VIEW
☑ CREATE FUNCTION

□ ALL PRIVILEGES gives all privileges



Grant on mfg_prod

Granted privileges will be inherited by applicable objects (e.g. schemas, tables) in this catalog.
Learn more

Users and groups

account users x

Privilege presets

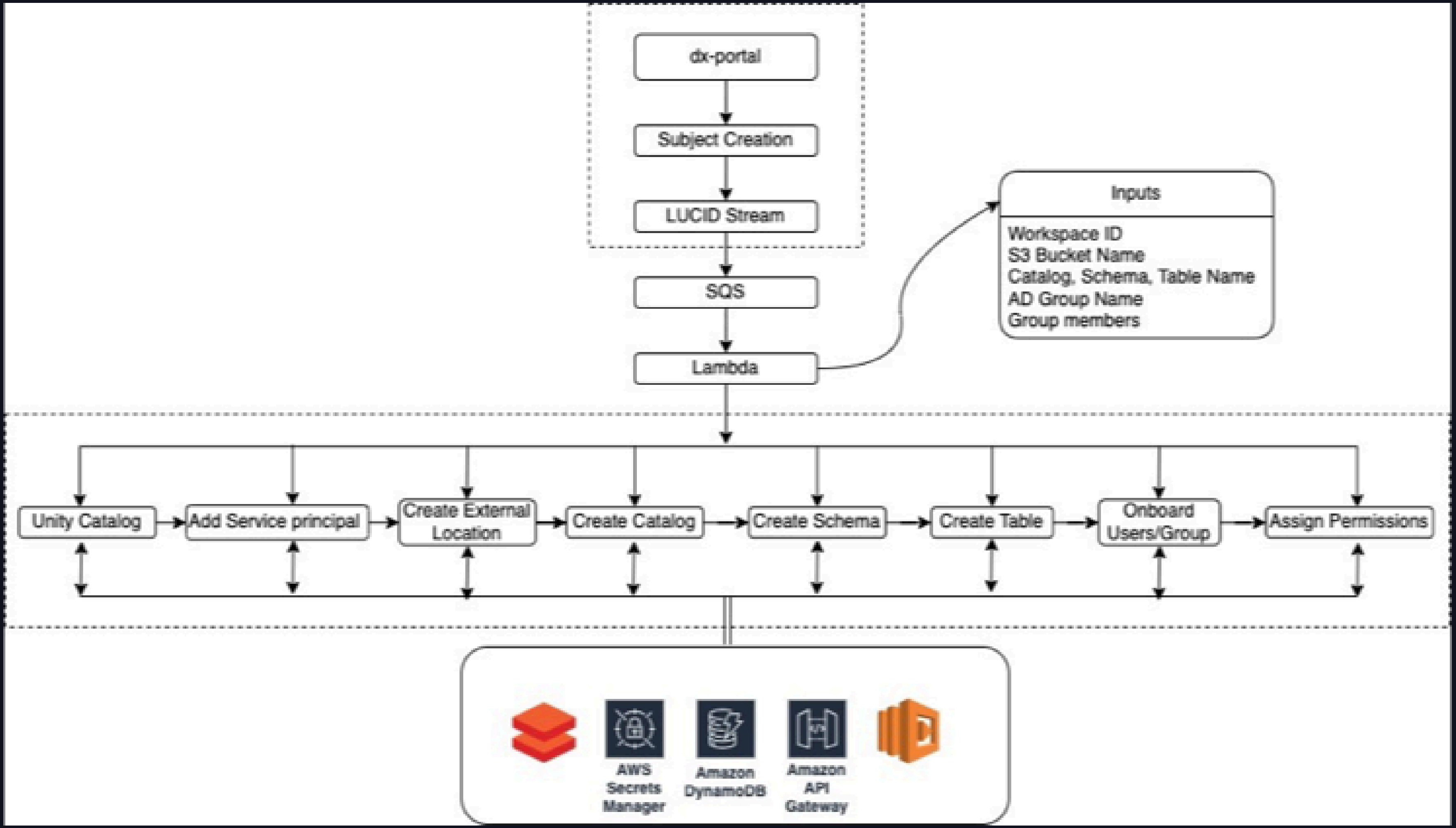
Data Reader (Can read from any object in the catalog)

Privileges

☑ USE CATALOG
☑ USE SCHEMA
☑ SELECT
□ MODIFY
□ EXECUTE
□ CREATE SCHEMA
□ CREATE TABLE
□ CREATE VIEW
□ CREATE FUNCTION

□ ALL PRIVILEGES gives all privileges

Unity Catalog Onboarding Automation Flow



dx-portal

Subject Creation

LUCID Stream

SQS

Lambda

Inputs

Workspace ID
S3 Bucket Name
Catalog, Schema, Table Name
AD Group Name
Group members

Unity Catalog

Add Service principal

Create External Location

Create Catalog

Create Schema

Create Table

Onboard Users/Group

Assign Permissions

AWS Secrets Manager

Amazon DynamoDB

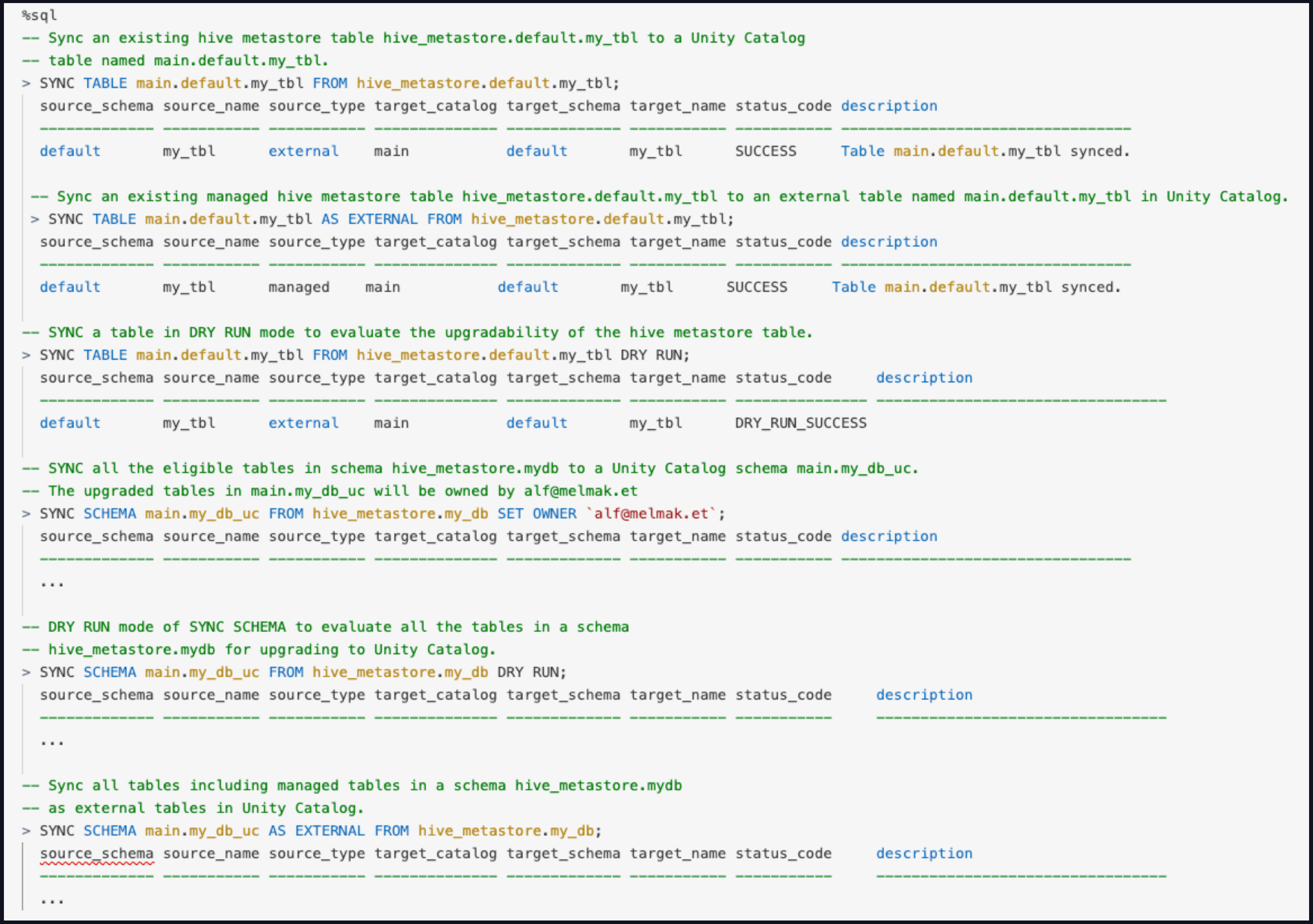
Amazon API Gateway

Migrate Legacy Metadata

4

Migrate Legacy Metadata

- SYNC external tables/schemas
- Migrate managed tables, files

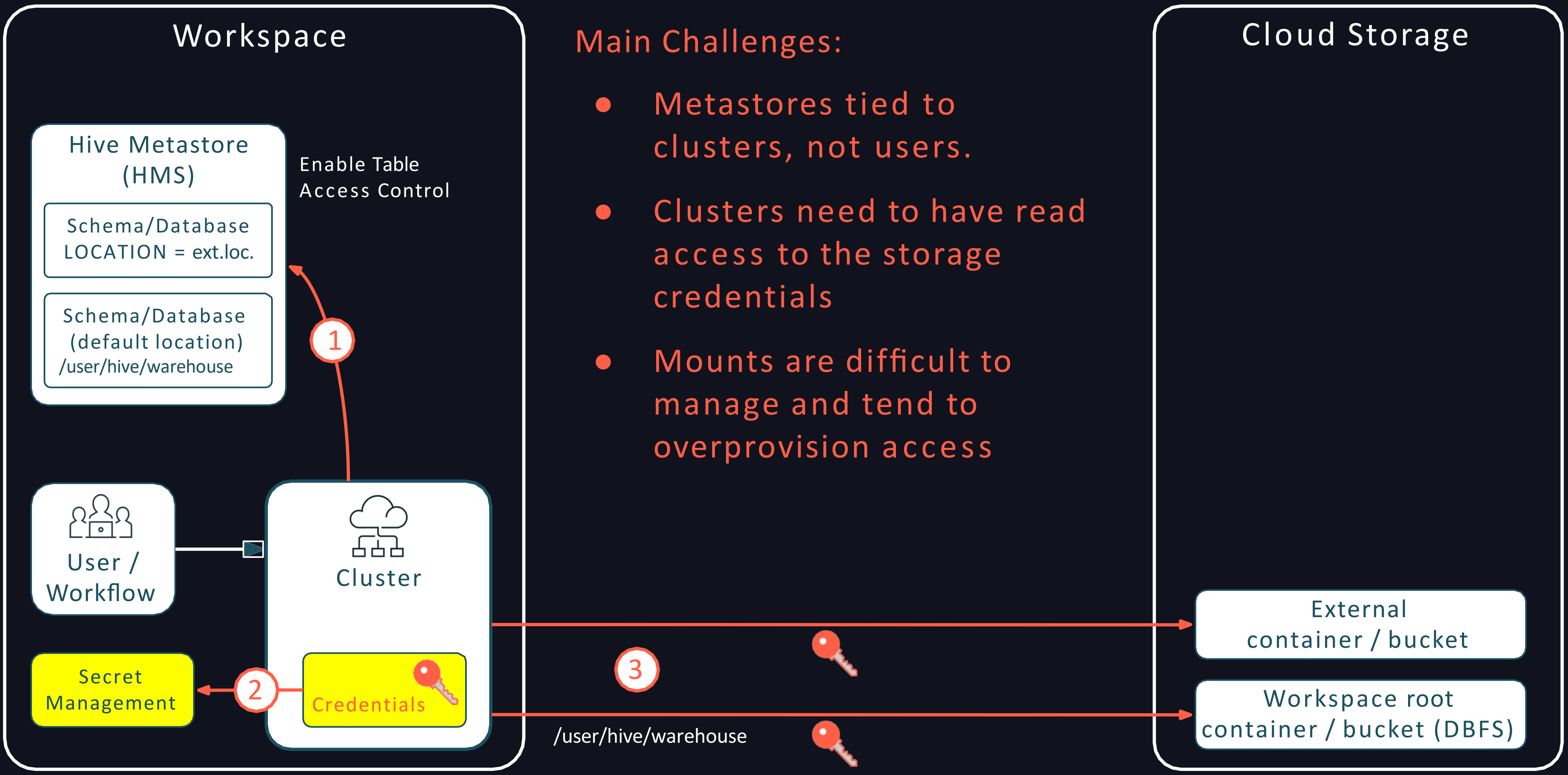


```
%sql
-- Sync an existing hive metastore table hive_metastore.default.my_tbl to a Unity Catalog
-- table named main.default.my_tbl.
> SYNC TABLE main.default.my_tbl FROM hive_metastore.default.my_tbl;
  source_schema source_name source_type target_catalog target_schema target_name status_code description
  ------------- ----------- ----------- -------------- ------------- ----------- ----------- ---------------------------------
  default       my_tbl      external    main           default       my_tbl      SUCCESS     Table main.default.my_tbl synced.

 -- Sync an existing managed hive metastore table hive_metastore.default.my_tbl to an external table named main.default.my_tbl in Unity Catalog.
 > SYNC TABLE main.default.my_tbl AS EXTERNAL FROM hive_metastore.default.my_tbl;
  source_schema source_name source_type target_catalog target_schema target_name status_code description
  ------------- ----------- ----------- -------------- ------------- ----------- ----------- ---------------------------------
  default       my_tbl      managed    main           default       my_tbl      SUCCESS     Table main.default.my_tbl synced.

-- SYNC a table in DRY RUN mode to evaluate the upgradability of the hive metastore table.
> SYNC TABLE main.default.my_tbl FROM hive_metastore.default.my_tbl DRY RUN;
  source_schema source_name source_type target_catalog target_schema target_name status_code     description
  ------------- ----------- ----------- -------------- ------------- ----------- --------------- ---------------------------------
  default       my_tbl      external    main           default       my_tbl      DRY_RUN_SUCCESS

-- SYNC all the eligible tables in schema hive_metastore.mydb to a Unity Catalog schema main.my_db_uc.
-- The upgraded tables in main.my_db_uc will be owned by alf@melmak.et
> SYNC SCHEMA main.my_db_uc FROM hive_metastore.my_db SET OWNER `alf@melmak.et`;
  source_schema source_name source_type target_catalog target_schema target_name status_code description
  ------------- ----------- ----------- -------------- ------------- ----------- ----------- ---------------------------------
  ...

-- DRY RUN mode of SYNC SCHEMA to evaluate all the tables in a schema
-- hive_metastore.mydb for upgrading to Unity Catalog.
> SYNC SCHEMA main.my_db_uc FROM hive_metastore.my_db DRY RUN;
  source_schema source_name source_type target_catalog target_schema target_name status_code     description
  ------------- ----------- ----------- -------------- ------------- ----------- -----------     ---------------------------------
  ...

-- Sync all tables including managed tables in a schema hive_metastore.mydb
-- as external tables in Unity Catalog.
> SYNC SCHEMA main.my_db_uc AS EXTERNAL FROM hive_metastore.my_db;
  source_schema source_name source_type target_catalog target_schema target_name status_code     description
  ------------- ----------- ----------- -------------- ------------- ----------- -----------     ---------------------------------
  ...
```

DATA+AI SUMMIT
BY databricks

Data Governance & Security

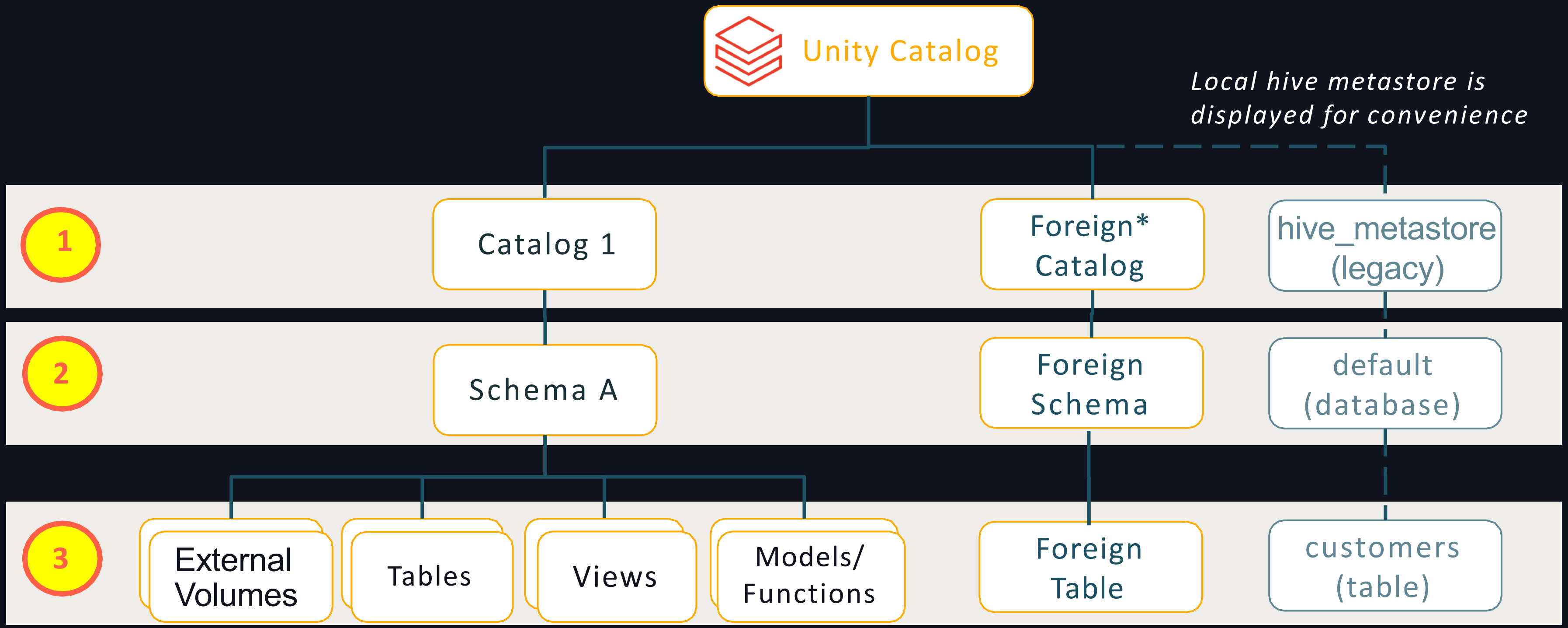
Hive metastore table access control (legacy)


Workspace
Hive Metastore (HMS)
Schema/Database LOCATION = ext.loc.
Schema/Database (default location) /user/hive/warehouse
Enable Table Access Control
1
User / Workflow
Cluster
Secret Management
2
Credentials
3
/user/hive/warehouse
Cloud Storage
External container / bucket
Workspace root container / bucket (DBFS)

Main Challenges:

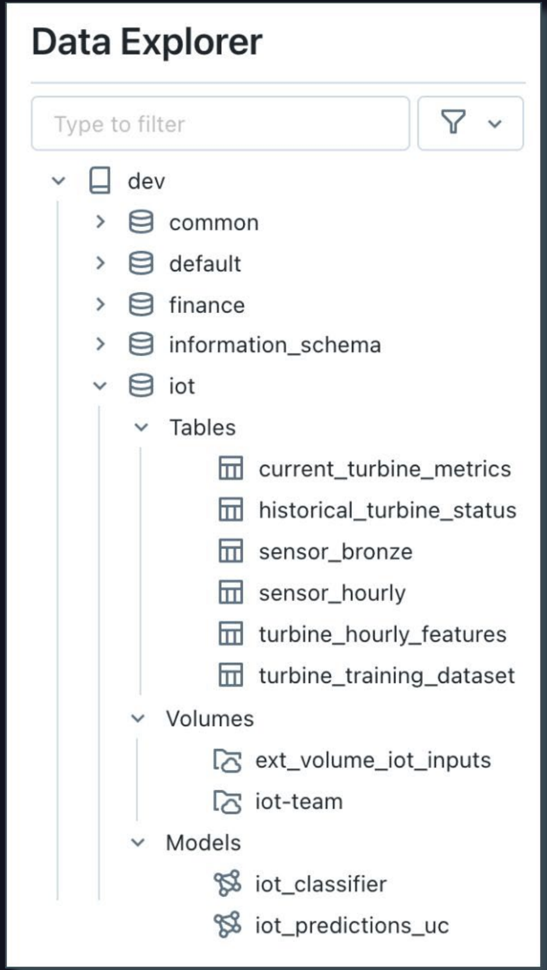
- Metastores tied to clusters, not users.
- Clusters need to have read access to the storage credentials
- Mounts are difficult to manage and tend to overprovision access

Legacy data access VS Three level namespace



Local hive metastore is displayed for convenience

```
With UC: SELECT * FROM dev.iot.sensor_bronze; <catalog>.<database>.<table>
Without UC: SELECT * FROM hive_metastore.default.customers;
```



Data Explorer

Type to filter

- dev
  - common
  - default
  - finance
  - information_schema
  - iot
    - Tables
      - current_turbine_metrics
      - historical_turbine_status
      - sensor_bronze
      - sensor_hourly
      - turbine_hourly_features
      - turbine_training_dataset
    - Volumes
      - ext_volume_iot_inputs
      - iot-team
    - Models
      - iot_classifier
      - iot_predictions_uc

IAM role credential passthrough (Legacy)

Difficult to administer; overprovisioned access

1. Create & manage instance profiles
2. Connection to S3 is not short-lived.
3. Every cluster needs setup with configs and instance profile
4. Create mounts for cloud locations
5. Risk of over provisioning

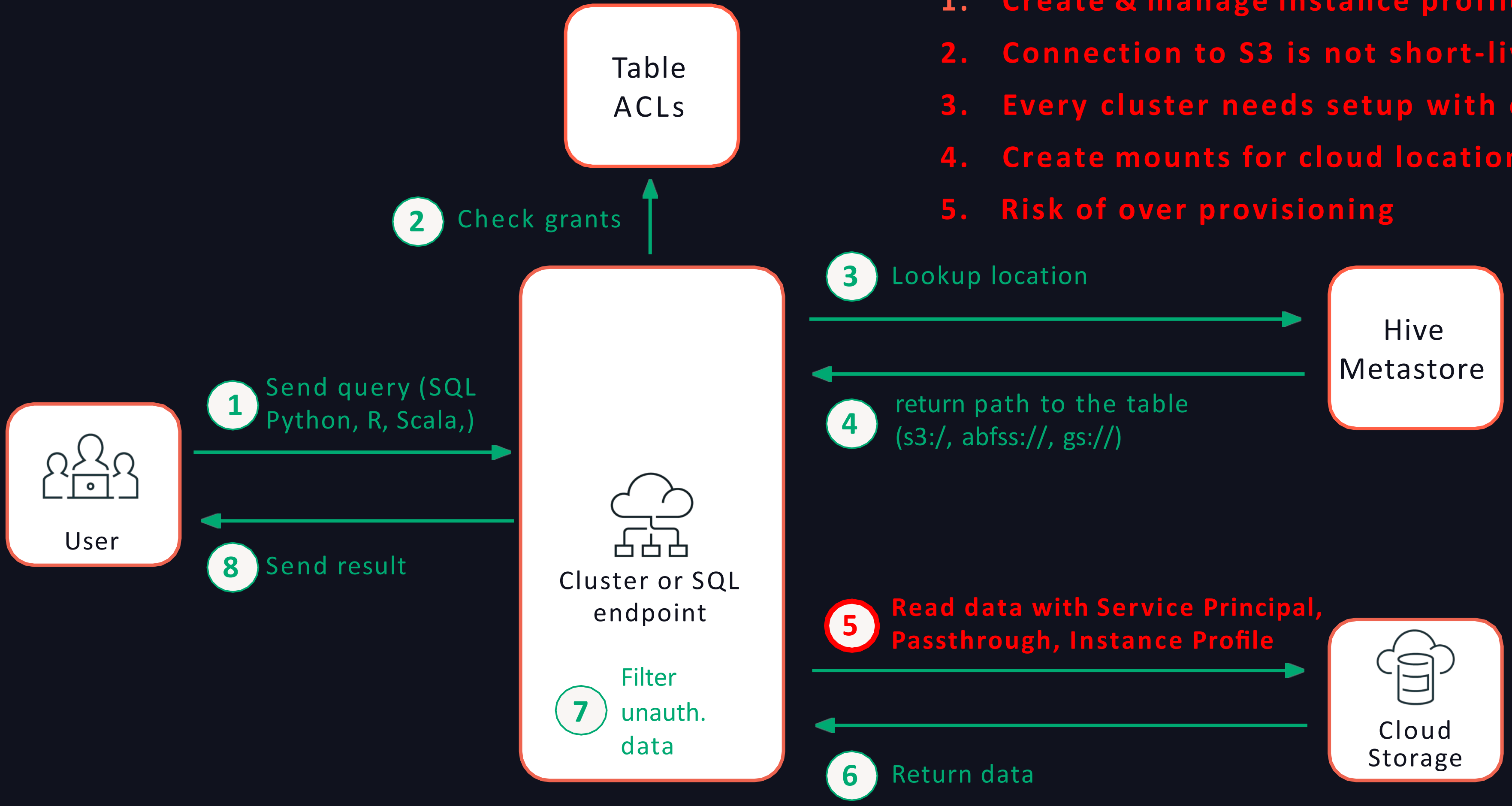


Table ACLs
2 Check grants
User
1 Send query (SQL Python, R, Scala,)
8 Send result
Cluster or SQL endpoint
7 Filter unauth. data
3 Lookup location
Hive Metastore
4 return path to the table (s3:/, abfss://, gs://)
5 Read data with Service Principal, Passthrough, Instance Profile
Cloud Storage
6 Return data

Need to configure spark.conf properties/instance profiles on clusters, configure table ACLs on each workspace, mounts are inflexible…

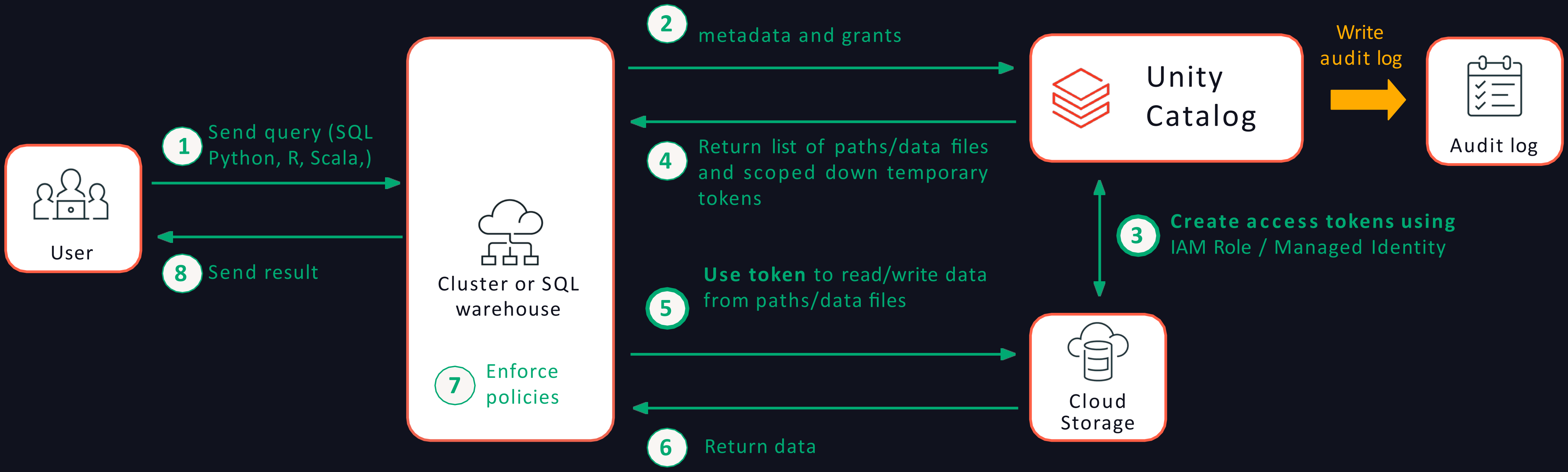
Life of a query with Unity Catalog

1. **No longer any need for Instance Profiles, Mounts, Secret Scopes, Spark.conf settings**

2. **No cluster level configs required.**

3. **Storage cred, External locations configured only once and can be used from any workspace**

4. **Data access for authorized users/service principals is via down scoped, short-lived, OAuth tokens**

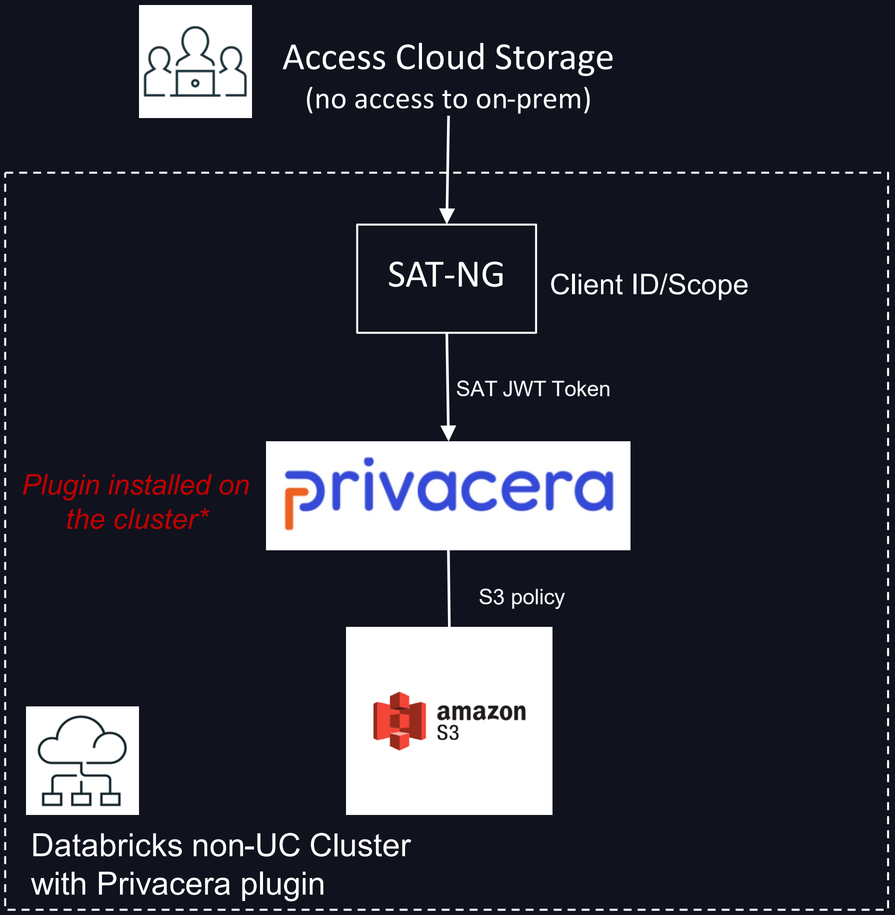

User
1 Send query (SQL Python, R, Scala,)
8 Send result
Cluster or SQL warehouse
7 Enforce policies
2 metadata and grants
4 Return list of paths/data files and scoped down temporary tokens
Unity Catalog
Write audit log
Audit log
3 Create access tokens using IAM Role / Managed Identity
5 Use token to read/write data from paths/data files
6 Return data
Cloud Storage

DATA+AI SUMMIT BY databricks

Secure access to Data Lake & On-prem Data

Pre-UC:MULTIPLE ACCESS PATTERNS

Privacera



Access Cloud Storage
(no access to on-prem)

SAT-NG

Client ID/Scope

SAT JWT Token

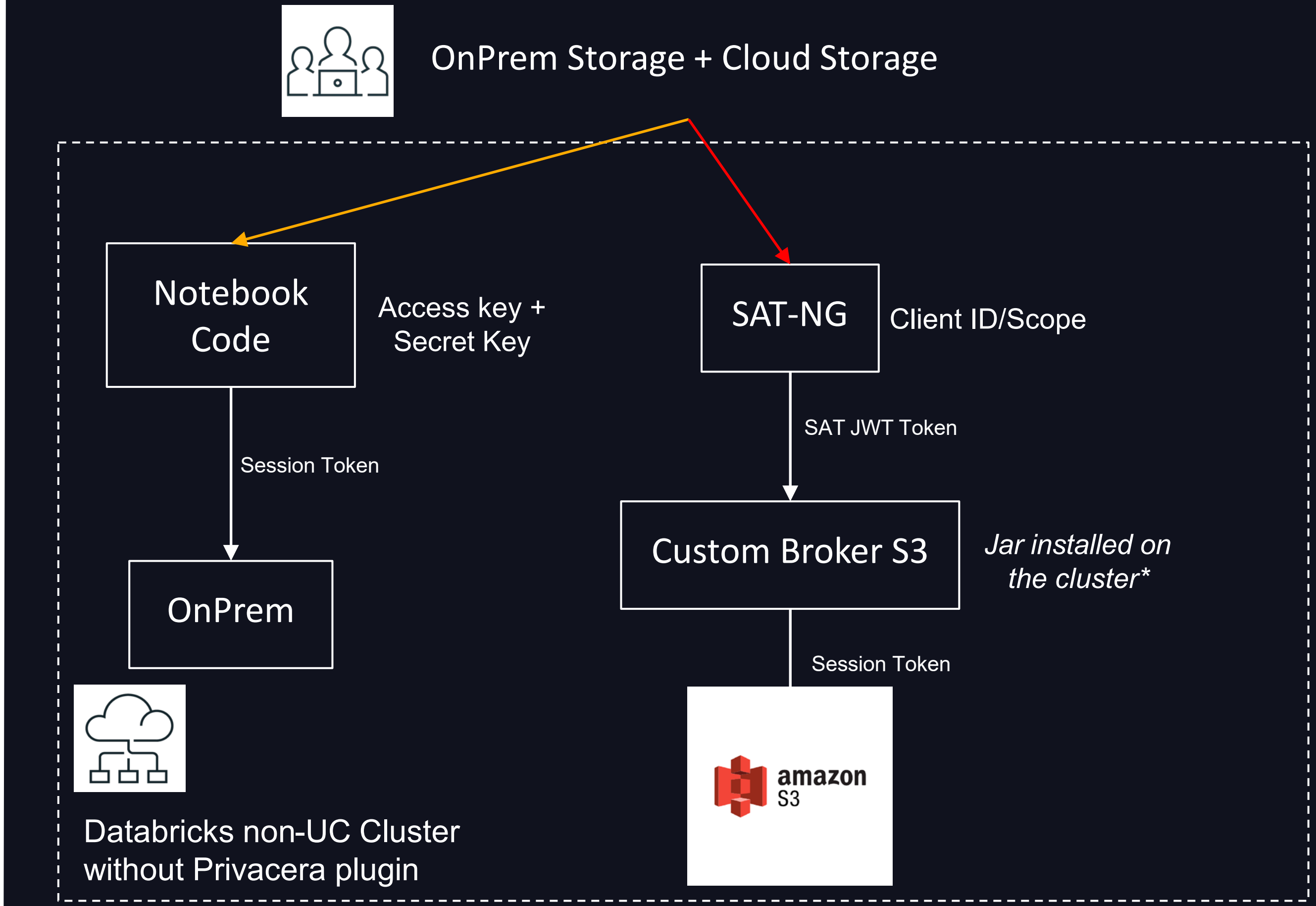
*Plugin installed on the cluster**

S3 policy

Databricks non-UC Cluster with Privacera plugin

Involves SAT portal that generates the JWT token
Authentication via JWT Token. No user level Identity
Any user with the token can access the data
Privacera needs to be installed on the cluster
User can't access minIO and non-DX S3 buckets

Custom Broker



OnPrem Storage + Cloud Storage

Notebook Code

Access key + Secret Key

Session Token

OnPrem

SAT-NG

Client ID/Scope

SAT JWT Token

Custom Broker S3

*Jar installed on the cluster**

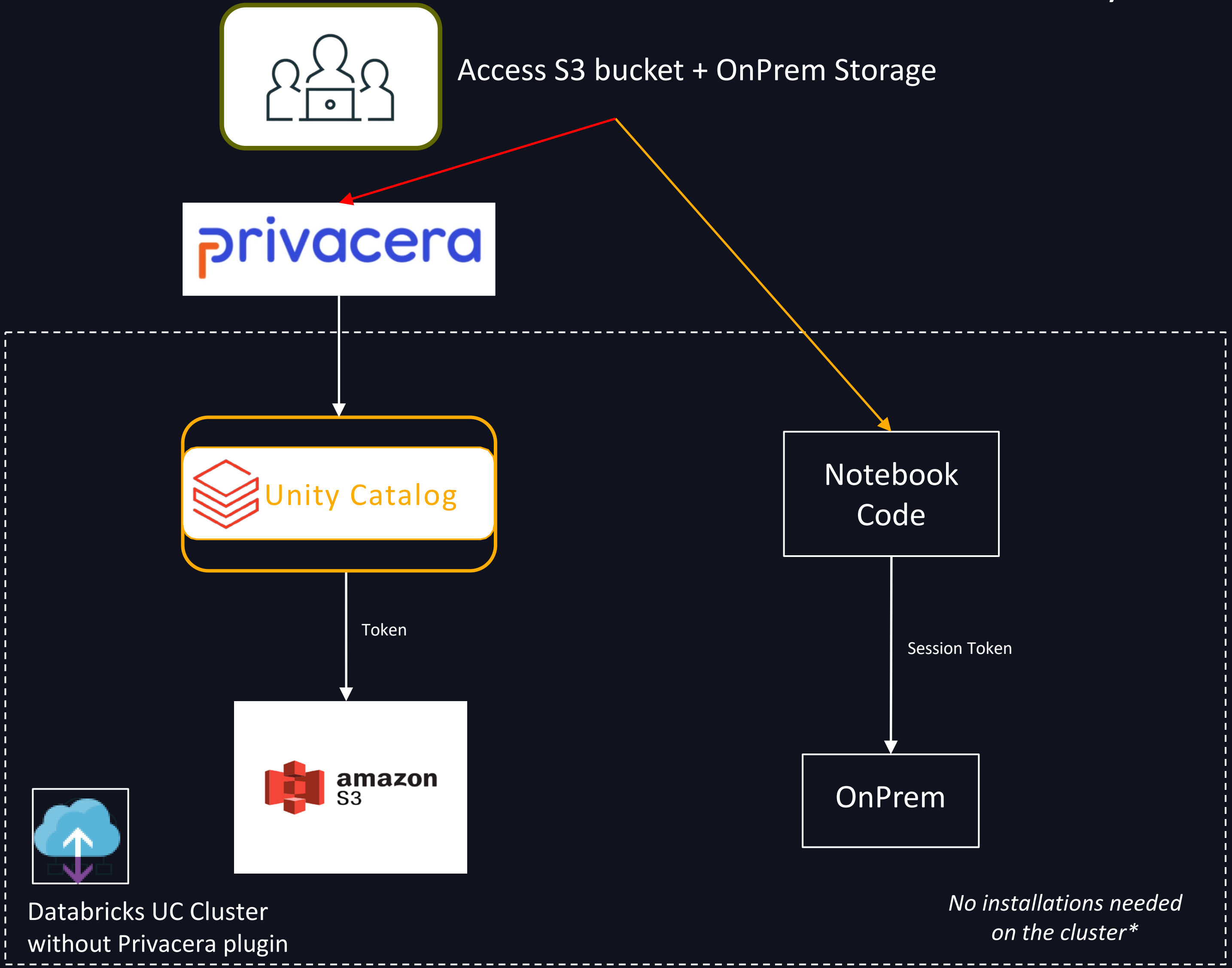
Session Token

Databricks non-UC Cluster without Privacera plugin

Doesn't integrate with Privacera
Involves SAT portal that generates the JWT token
Authentication via JWT Token. No user level Identity
Any user with the token can access the data
Custom Broker Jar needs to be installed on the cluster
No fine-grained access

Post UC - SINGLE ACCESS PATTERN

Privacera + Unity Catalog



Highlights:

User's own identity will be used for data access.

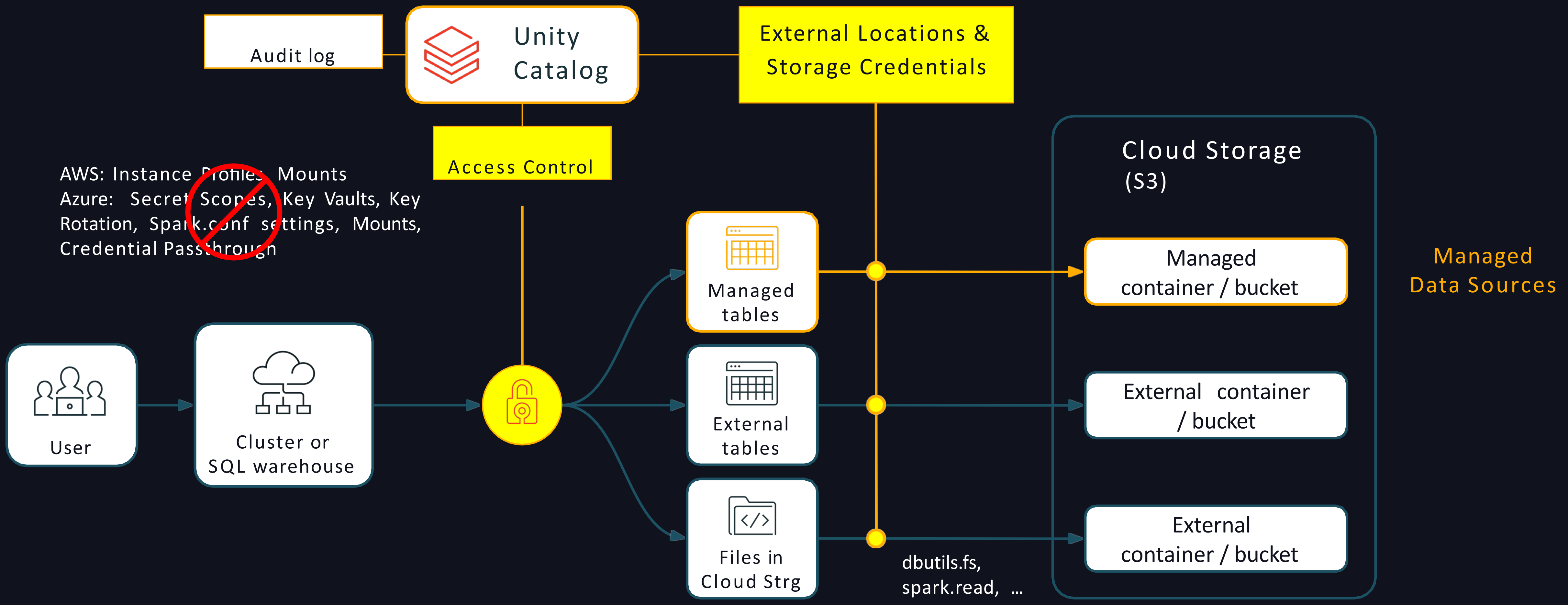
Hence, Governance can be managed without JWT tokens and clientID

Integrates with Privacera and at the same time lets users access On-prem and non-DX S3 buckets

No installations needed on the cluster

Managed Data Sources & External Locations

Simplify data access management across clouds



Audit log
Unity Catalog
External Locations & Storage Credentials
Access Control
AWS: Instance Profiles, Mounts
Azure: Secret Scopes, Key Vaults, Key Rotation, Spark.conf settings, Mounts, Credential Passthrough
User
Cluster or SQL warehouse
Managed tables
External tables
Files in Cloud Strg
dbutils.fs, spark.read, …
Cloud Storage (S3)
Managed container / bucket
External container / bucket
External container / bucket
Managed Data Sources

Comparison of Managed and External UC Tables

Consider the benefits of Managed tables

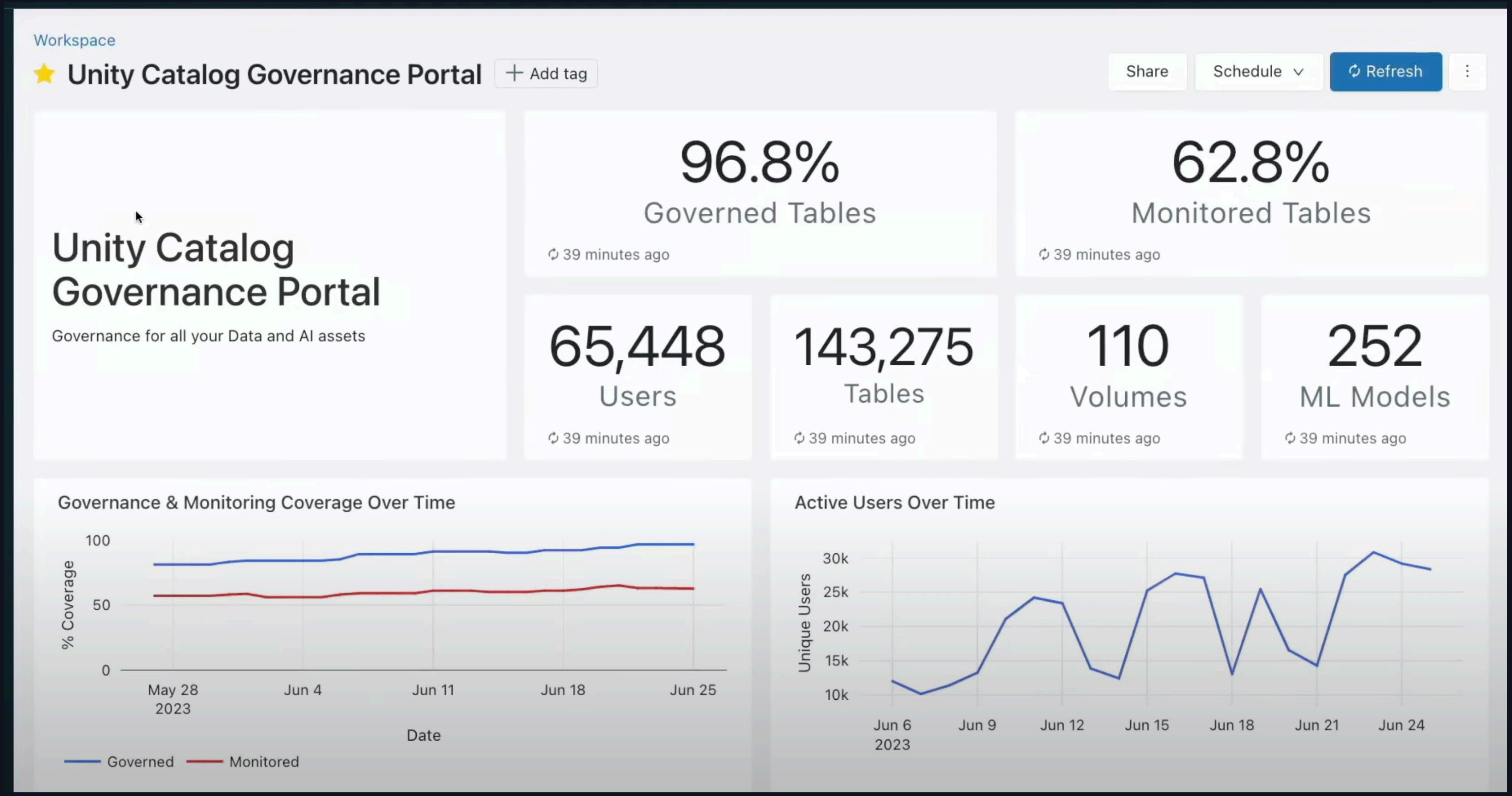
Characteristic	**Managed**	**Unmanaged** (a.k.a. “External”)
Table’s **Type** Property Value	“MANAGED”	“EXTERNAL”
DROP Table Behavior	**Deletes** the associated data - Generally what business analysts (SQL users) expect	Only deletes metadata. Does not delete the underlying S3 data. - May be helpful in certain use cases

Create Table Syntax	`CREATE TABLE [<catalog>.][<schema>.]<table> (`
	`<column_specification>;`
		`CREATE TABLE [<catalog>.][<schema>.]<table> (`
		`<column_specification>`
		`LOCATION 'abfss:/cont@stacct.dfs.core.windows.net' ;`

Data File Location	Whichever is found first: - Location specified for the database (if specified) - Location specified for the catalog (if specified) - Metastore default managed storage location	The path specified by the LOCATION keyword in your create table statement
Performance Optimizations	**Auto Tune** (In Preview)	Manually managed by the customer
Data Format Support	DELTA	DELTA, CSV, JSON, AVRO, PARQUET, ORC, TEXT

Data Management & Observability

Unity Catalog Governance Portal



Unity Catalog Governance Portal

Governance for all your Data and AI assets

96.8% Governed Tables — 39 minutes ago

62.8% Monitored Tables — 39 minutes ago

65,448 Users — 39 minutes ago

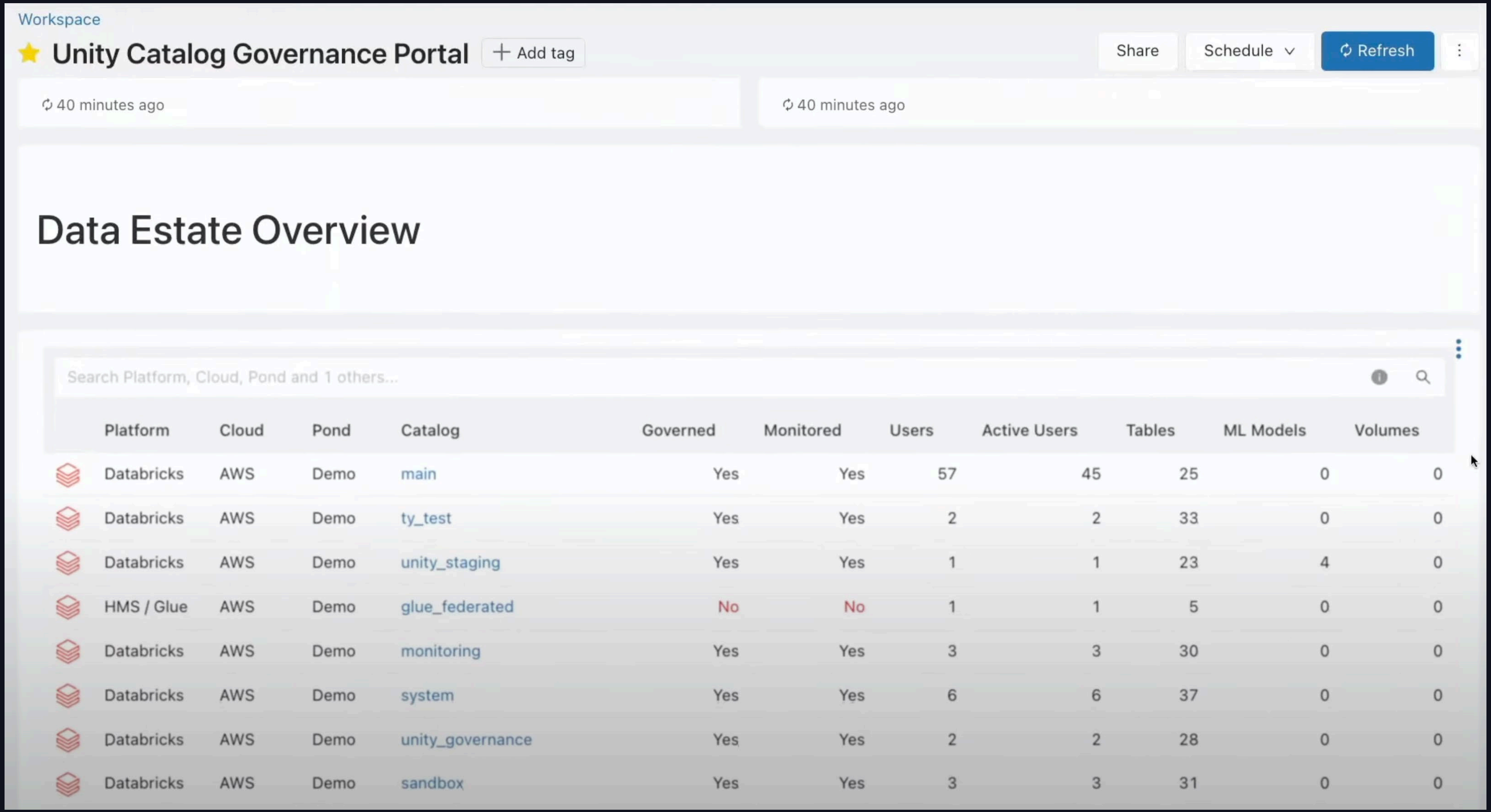
143,275 Tables — 39 minutes ago

110 Volumes — 39 minutes ago

252 ML Models — 39 minutes ago

DATA+AI SUMMIT

Unity Catalog Governance Portal



Workspace
★ Unity Catalog Governance Portal
+ Add tag
Share
Schedule
Refresh
40 minutes ago
40 minutes ago

Data Estate Overview

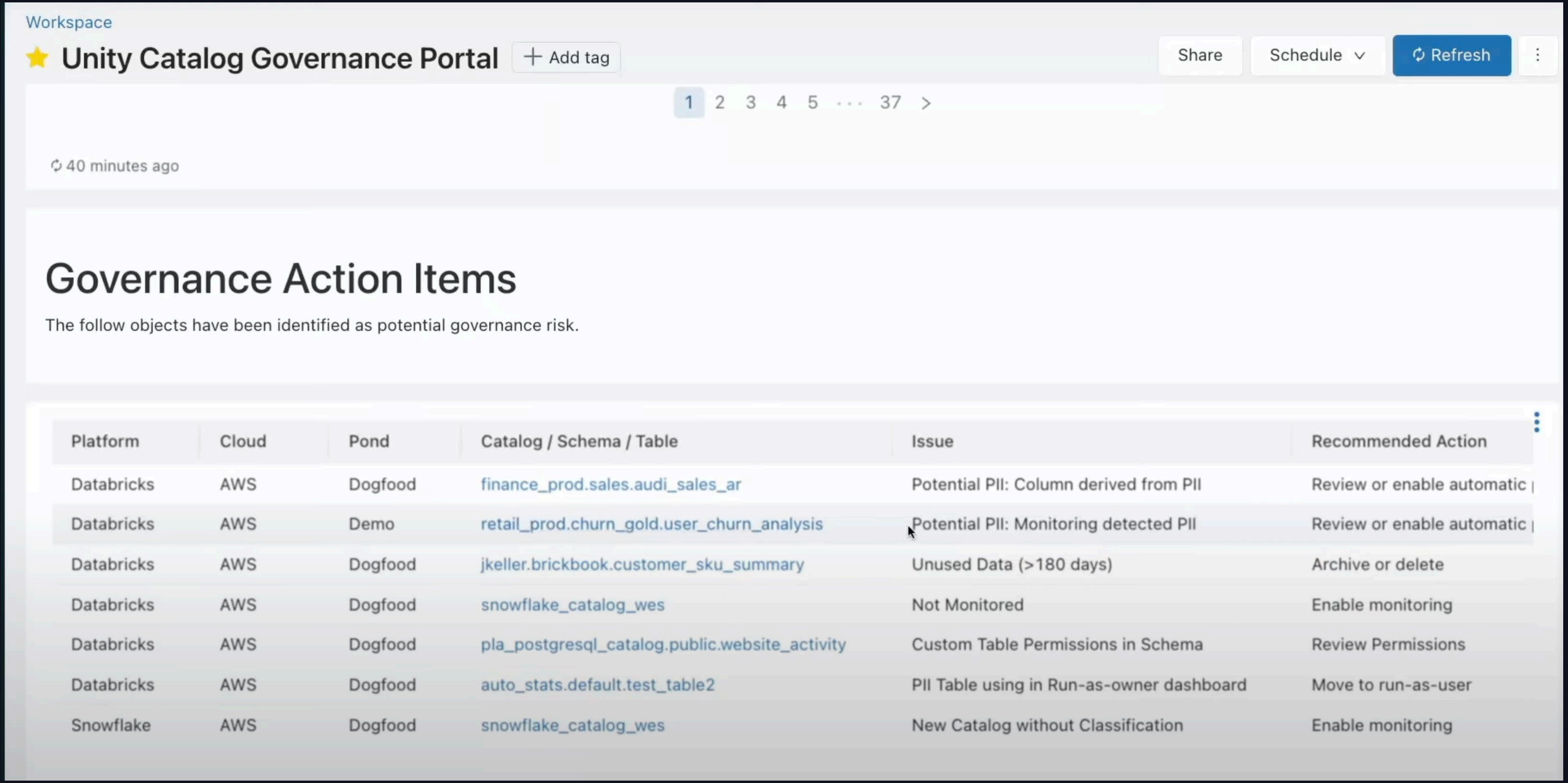
Search Platform, Cloud, Pond and 1 others...

	Platform	Cloud	Pond	Catalog	Governed	Monitored	Users	Active Users	Tables	ML Models	Volumes
	Databricks	AWS	Demo	main	Yes	Yes	57	45	25	0	0
	Databricks	AWS	Demo	ty_test	Yes	Yes	2	2	33	0	0
	Databricks	AWS	Demo	unity_staging	Yes	Yes	1	1	23	4	0
	HMS / Glue	AWS	Demo	glue_federated	No	No	1	1	5	0	0
	Databricks	AWS	Demo	monitoring	Yes	Yes	3	3	30	0	0
	Databricks	AWS	Demo	system	Yes	Yes	6	6	37	0	0
	Databricks	AWS	Demo	unity_governance	Yes	Yes	2	2	28	0	0
	Databricks	AWS	Demo	sandbox	Yes	Yes	3	3	31	0	0

DATA+AI SUMMIT BY databricks

AI for Governance

Unity Catalog Governance Portal



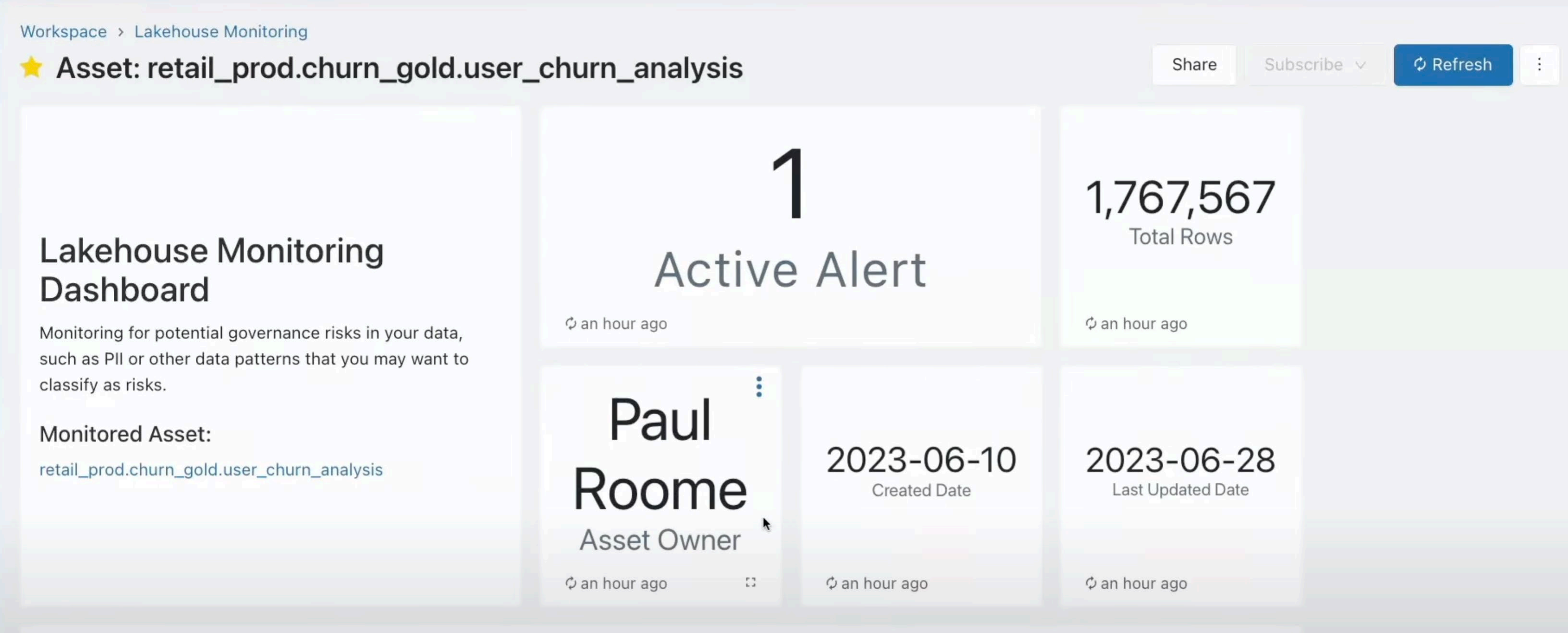
Unity Catalog Governance Portal

Governance Action Items

The follow objects have been identified as potential governance risk.

Platform	Cloud	Pond	Catalog / Schema / Table	Issue	Recommended Action
Databricks	AWS	Dogfood	finance_prod.sales.audi_sales_ar	Potential PII: Column derived from PII	Review or enable automatic
Databricks	AWS	Demo	retail_prod.churn_gold.user_churn_analysis	Potential PII: Monitoring detected PII	Review or enable automatic
Databricks	AWS	Dogfood	jkeller.brickbook.customer_sku_summary	Unused Data (>180 days)	Archive or delete
Databricks	AWS	Dogfood	snowflake_catalog_wes	Not Monitored	Enable monitoring
Databricks	AWS	Dogfood	pla_postgresql_catalog.public.website_activity	Custom Table Permissions in Schema	Review Permissions
Databricks	AWS	Dogfood	auto_stats.default.test_table2	PII Table using in Run-as-owner dashboard	Move to run-as-user
Snowflake	AWS	Dogfood	snowflake_catalog_wes	New Catalog without Classification	Enable monitoring

Lakehouse Monitoring Dashboard



★ Asset: retail_prod.churn_gold.user_churn_analysis

Lakehouse Monitoring Dashboard

Monitoring for potential governance risks in your data, such as PII or other data patterns that you may want to classify as risks.

Monitored Asset:

retail_prod.churn_gold.user_churn_analysis

1

Active Alert

an hour ago

1,767,567

Total Rows

an hour ago

Paul Roome

Asset Owner

an hour ago

2023-06-10

Created Date

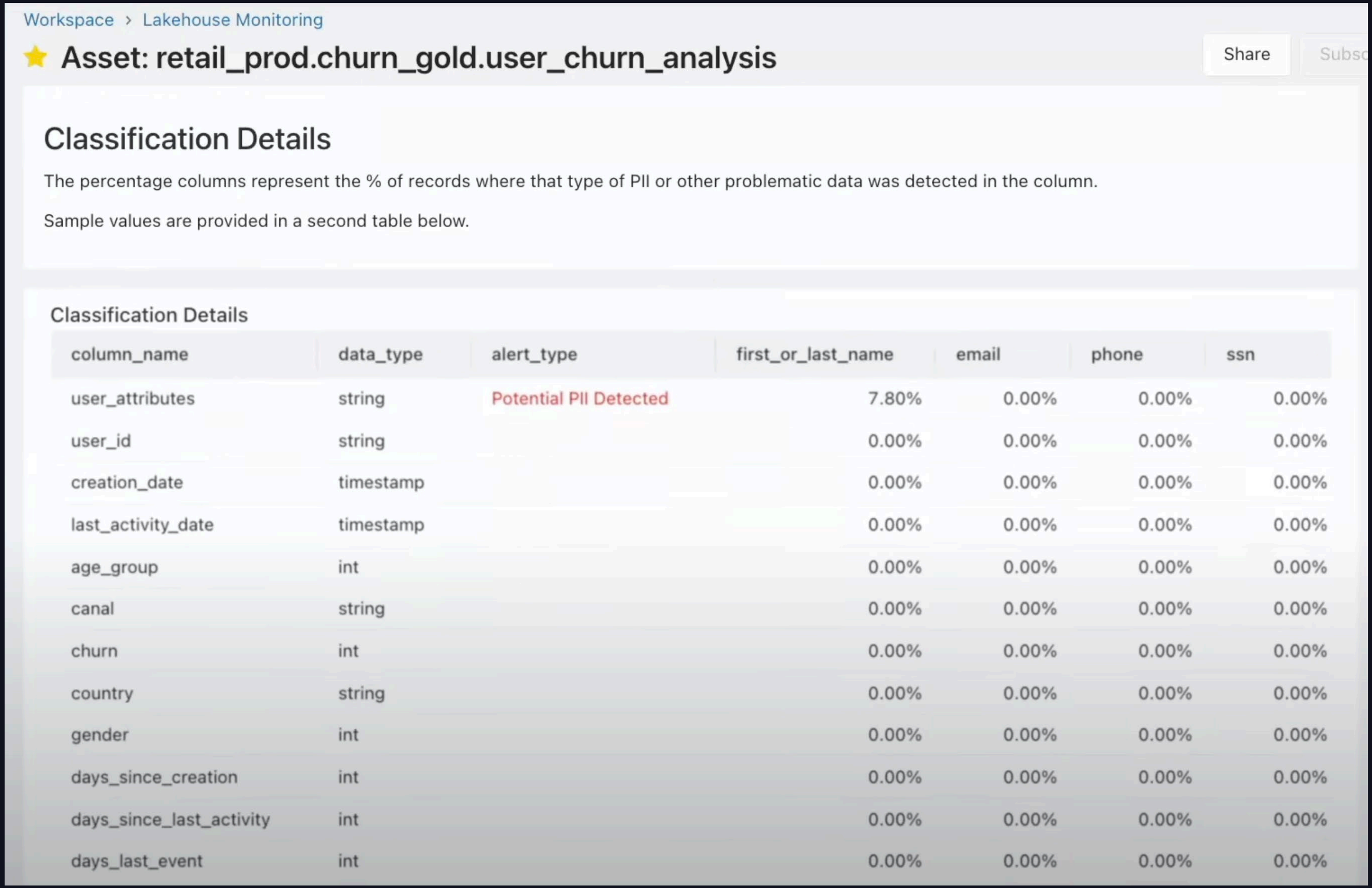
an hour ago

2023-06-28

Last Updated Date

an hour ago

Lakehouse Monitoring Dashboard



★ Asset: retail_prod.churn_gold.user_churn_analysis

Classification Details

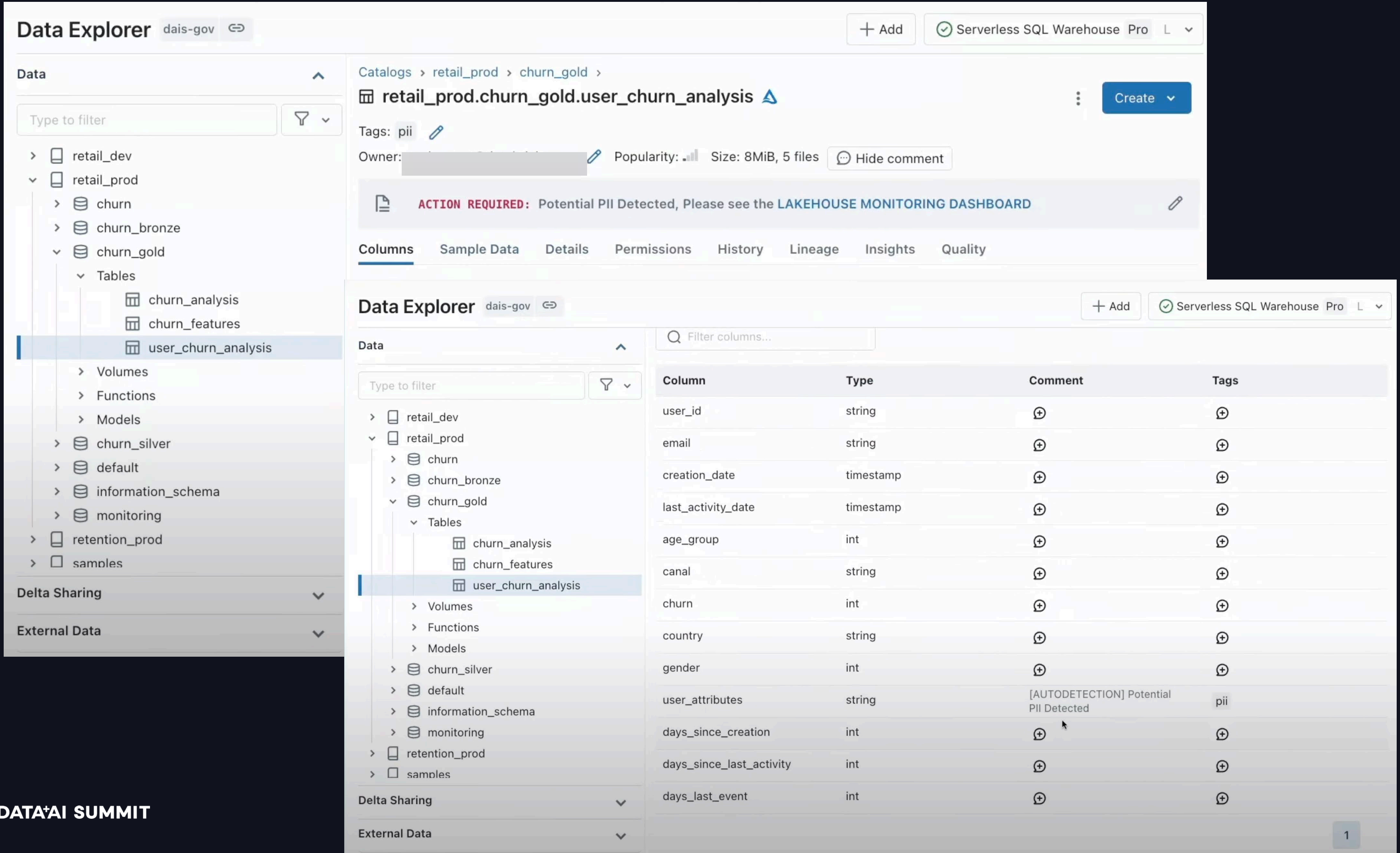
The percentage columns represent the % of records where that type of PII or other problematic data was detected in the column.

Sample values are provided in a second table below.

Classification Details

column_name	data_type	alert_type	first_or_last_name	email	phone	ssn
user_attributes	string	Potential PII Detected	7.80%	0.00%	0.00%	0.00%
user_id	string		0.00%	0.00%	0.00%	0.00%
creation_date	timestamp		0.00%	0.00%	0.00%	0.00%
last_activity_date	timestamp		0.00%	0.00%	0.00%	0.00%
age_group	int		0.00%	0.00%	0.00%	0.00%
canal	string		0.00%	0.00%	0.00%	0.00%
churn	int		0.00%	0.00%	0.00%	0.00%
country	string		0.00%	0.00%	0.00%	0.00%
gender	int		0.00%	0.00%	0.00%	0.00%
days_since_creation	int		0.00%	0.00%	0.00%	0.00%
days_since_last_activity	int		0.00%	0.00%	0.00%	0.00%
days_last_event	int		0.00%	0.00%	0.00%	0.00%

Data Explorer Dashboard



Data Explorer dais-gov

+ Add | Serverless SQL Warehouse Pro L

Data

Type to filter

- retail_dev
- retail_prod
  - churn
  - churn_bronze
  - churn_gold
    - Tables
      - churn_analysis
      - churn_features
      - user_churn_analysis
    - Volumes
    - Functions
    - Models
  - churn_silver
  - default
  - information_schema
  - monitoring
- retention_prod
- samples

Delta Sharing

External Data

Catalogs › retail_prod › churn_gold ›

retail_prod.churn_gold.user_churn_analysis

Create

Tags: pii

Owner: [redacted] Popularity: Size: 8MiB, 5 files Hide comment

ACTION REQUIRED: Potential PII Detected, Please see the LAKEHOUSE MONITORING DASHBOARD

Columns | Sample Data | Details | Permissions | History | Lineage | Insights | Quality

Data Explorer dais-gov

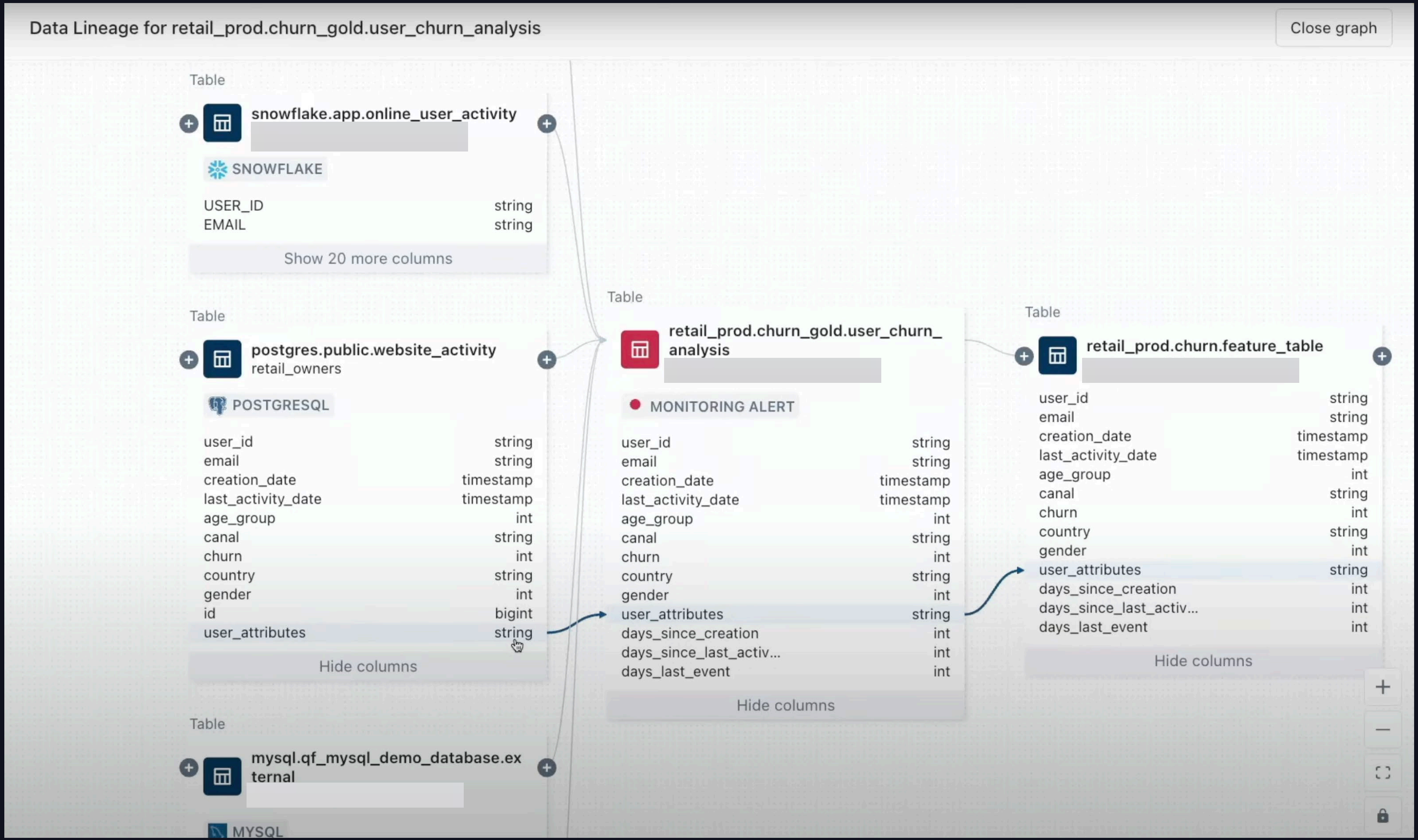
+ Add | Serverless SQL Warehouse Pro L

Filter columns...

Column	Type	Comment	Tags
user_id	string	⊕	⊕
email	string	⊕	⊕
creation_date	timestamp	⊕	⊕
last_activity_date	timestamp	⊕	⊕
age_group	int	⊕	⊕
canal	string	⊕	⊕
churn	int	⊕	⊕
country	string	⊕	⊕
gender	int	⊕	⊕
user_attributes	string	[AUTODETECTION] Potential PII Detected	pii
days_since_creation	int	⊕	⊕
days_since_last_activity	int	⊕	⊕
days_last_event	int	⊕	⊕

1

Data Lineage Dashboard



Data Lineage for retail_prod.churn_gold.user_churn_analysis

Close graph

Table

snowflake.app.online_user_activity

SNOWFLAKE

USER_ID	string
EMAIL	string

Show 20 more columns

Table

postgres.public.website_activity
retail_owners

POSTGRESQL

user_id	string
email	string
creation_date	timestamp
last_activity_date	timestamp
age_group	int
canal	string
churn	int
country	string
gender	int
id	bigint
user_attributes	string

Hide columns

Table

mysql.qf_mysql_demo_database.external

MYSQL

Table

retail_prod.churn_gold.user_churn_analysis

● MONITORING ALERT

user_id	string
email	string
creation_date	timestamp
last_activity_date	timestamp
age_group	int
canal	string
churn	int
country	string
gender	int
user_attributes	string
days_since_creation	int
days_since_last_activ...	int
days_last_event	int

Hide columns

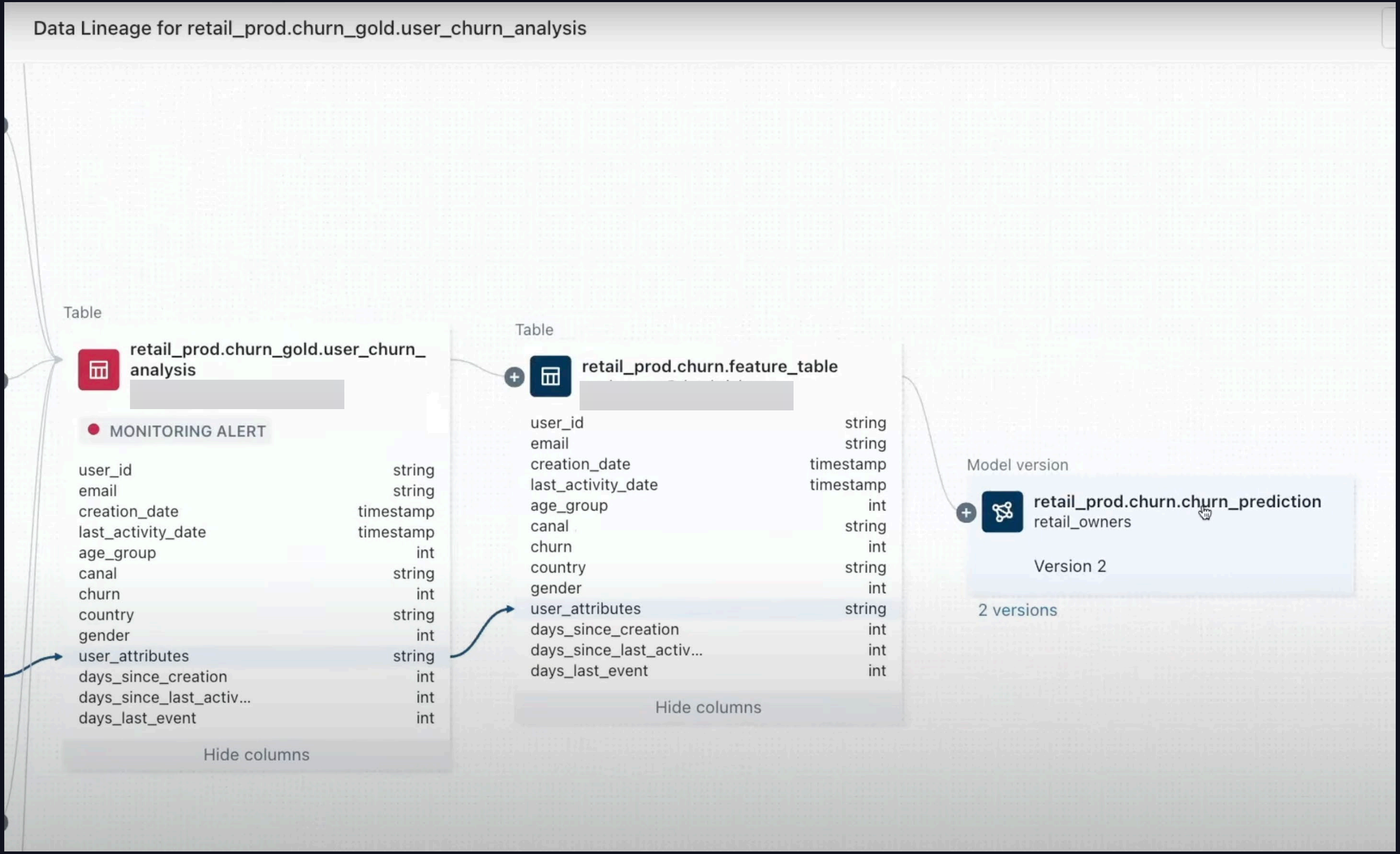
Table

retail_prod.churn.feature_table

user_id	string
email	string
creation_date	timestamp
last_activity_date	timestamp
age_group	int
canal	string
churn	int
country	string
gender	int
user_attributes	string
days_since_creation	int
days_since_last_activ...	int
days_last_event	int

Hide columns

Data Lineage Dashboard



Data Lineage for retail_prod.churn_gold.user_churn_analysis

Table

retail_prod.churn_gold.user_churn_analysis

[redacted]

● MONITORING ALERT

user_id	string
email	string
creation_date	timestamp
last_activity_date	timestamp
age_group	int
canal	string
churn	int
country	string
gender	int
user_attributes	string
days_since_creation	int
days_since_last_activ...	int
days_last_event	int

Hide columns

Table

retail_prod.churn.feature_table

[redacted]

user_id	string
email	string
creation_date	timestamp
last_activity_date	timestamp
age_group	int
canal	string
churn	int
country	string
gender	int
user_attributes	string
days_since_creation	int
days_since_last_activ...	int
days_last_event	int

Hide columns

Model version

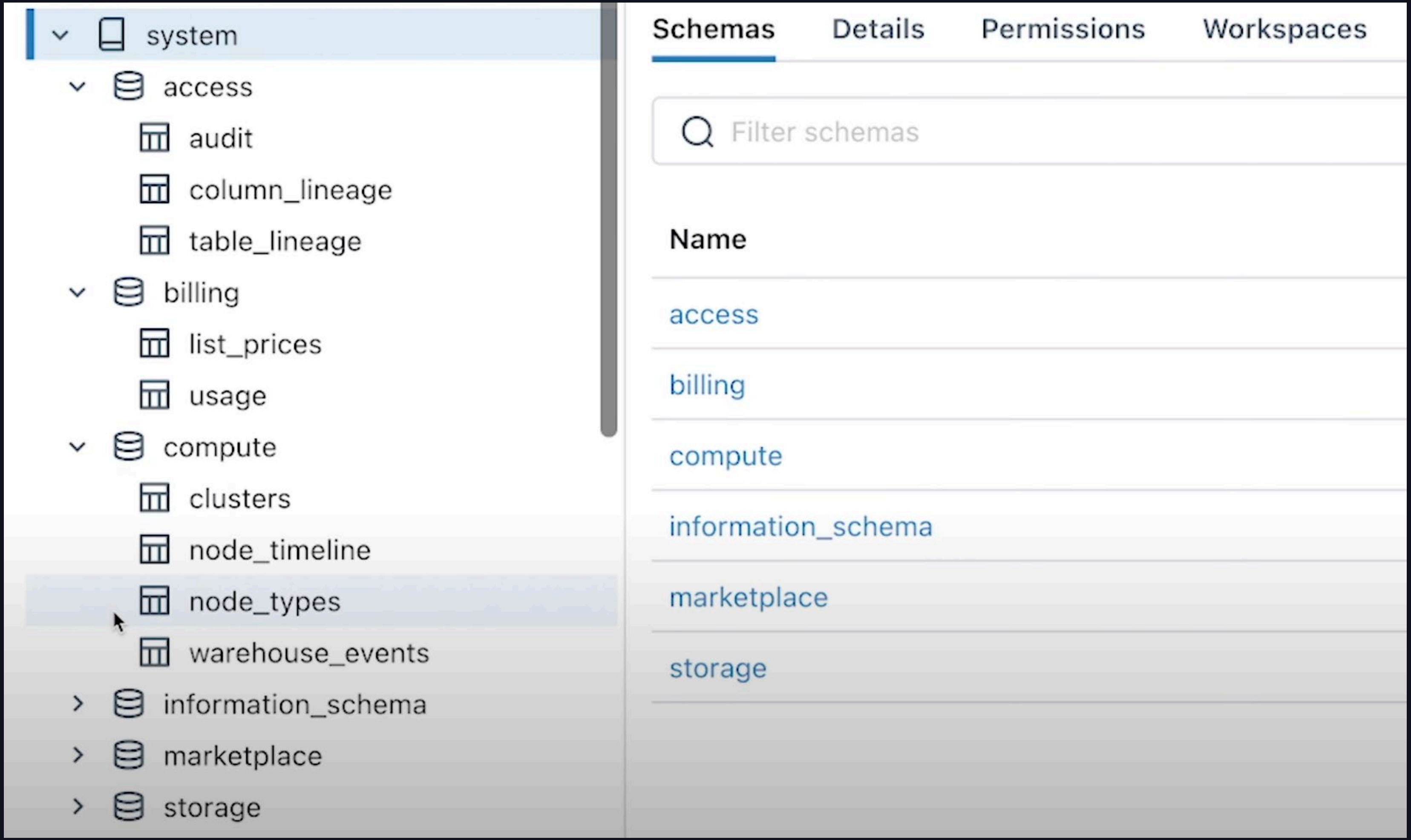
retail_prod.churn.churn_prediction
retail_owners

Version 2

2 versions

Data Audit : Systems Catalog

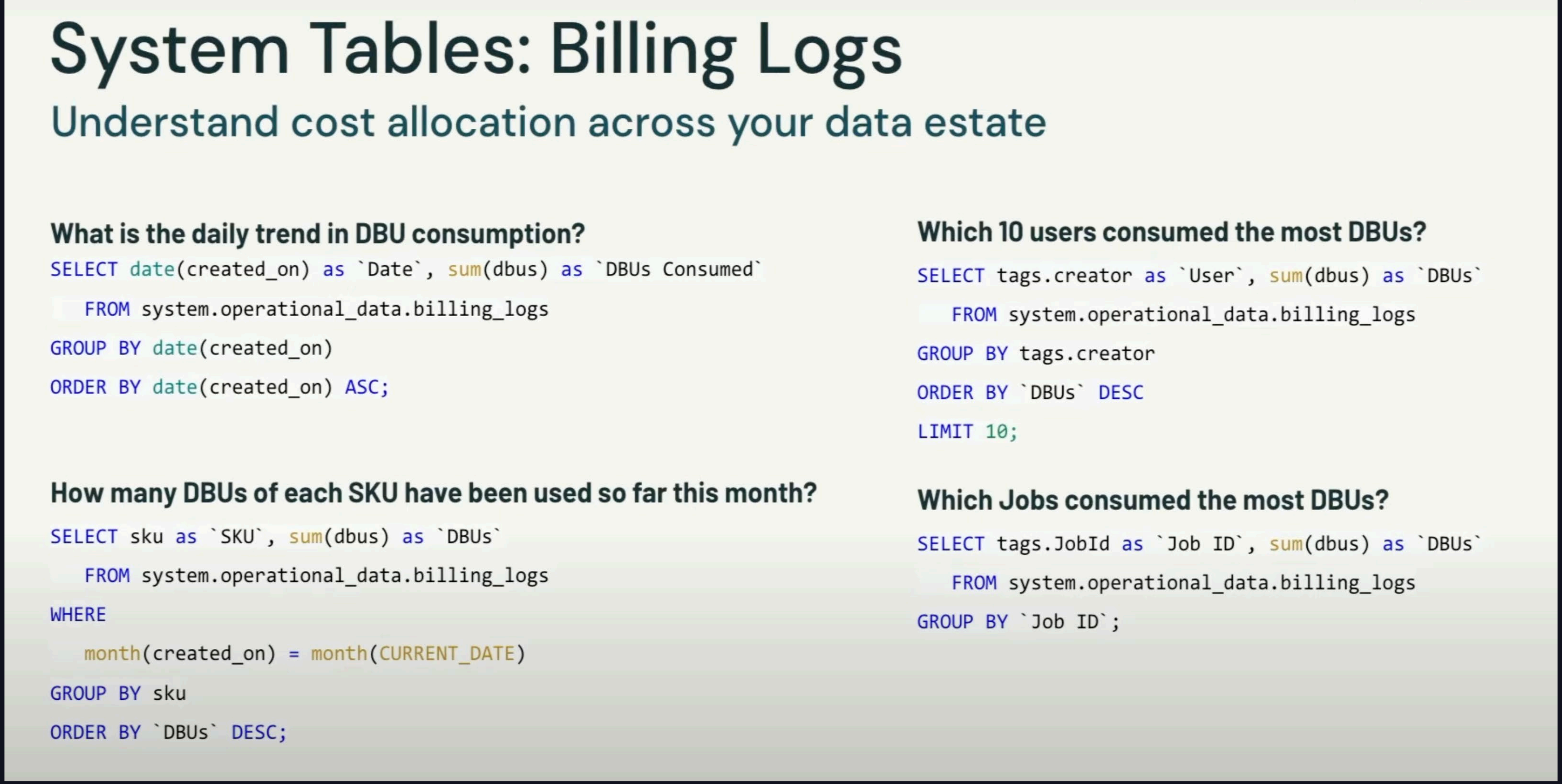
Audit and Billing Monitoring Using System Catalog



- system
  - access
    - audit
    - column_lineage
    - table_lineage
  - billing
    - list_prices
    - usage
  - compute
    - clusters
    - node_timeline
    - node_types
    - warehouse_events
  - information_schema
  - marketplace
  - storage

Name
access
billing
compute
information_schema
marketplace
storage

Billing Usage Sample View, Queries



System Tables: Billing Logs

Understand cost allocation across your data estate

What is the daily trend in DBU consumption?

```
SELECT date(created_on) as `Date`, sum(dbus) as `DBUs Consumed`
   FROM system.operational_data.billing_logs
GROUP BY date(created_on)
ORDER BY date(created_on) ASC;
```

How many DBUs of each SKU have been used so far this month?

```
SELECT sku as `SKU`, sum(dbus) as `DBUs`
   FROM system.operational_data.billing_logs
WHERE
   month(created_on) = month(CURRENT_DATE)
GROUP BY sku
ORDER BY `DBUs` DESC;
```

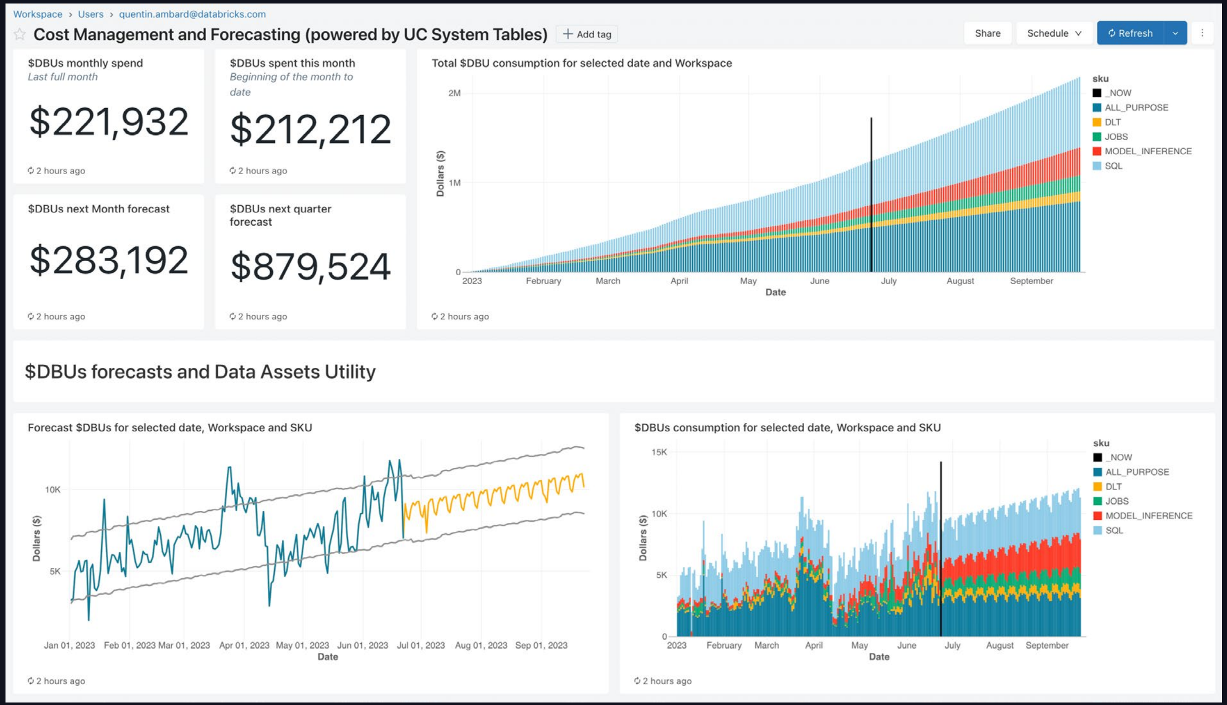
Which 10 users consumed the most DBUs?

```
SELECT tags.creator as `User`, sum(dbus) as `DBUs`
   FROM system.operational_data.billing_logs
GROUP BY tags.creator
ORDER BY `DBUs` DESC
LIMIT 10;
```

Which Jobs consumed the most DBUs?

```
SELECT tags.JobId as `Job ID`, sum(dbus) as `DBUs`
   FROM system.operational_data.billing_logs
GROUP BY `Job ID`;
```

Cost Governance Dashboard



Workspace › Users › quentin.ambard@databricks.com

Cost Management and Forecasting (powered by UC System Tables)

$DBUs monthly spend
Last full month
$221,932
2 hours ago

$DBUs spent this month
Beginning of the month to date
$212,212
2 hours ago

$DBUs next Month forecast
$283,192
2 hours ago

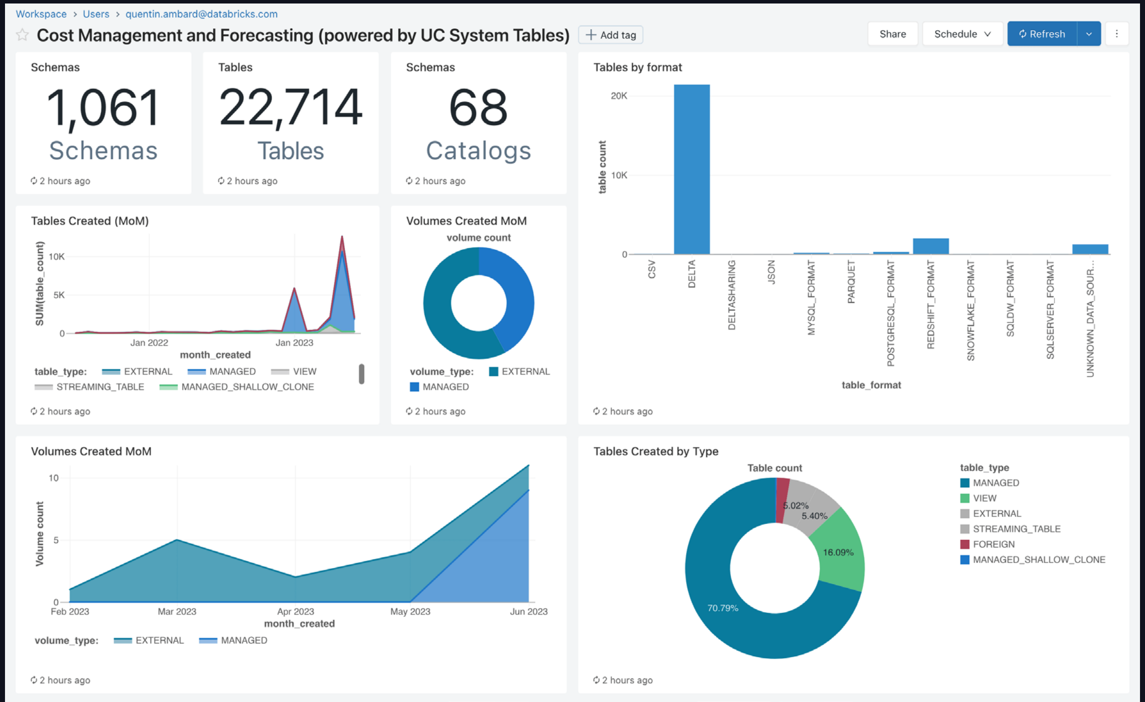
$DBUs next quarter forecast
$879,524
2 hours ago

Total $DBU consumption for selected date and Workspace
2 hours ago

$DBUs forecasts and Data Assets Utility

Forecast $DBUs for selected date, Workspace and SKU
2 hours ago

$DBUs consumption for selected date, Workspace and SKU
2 hours ago



Workspace › Users › quentin.ambard@databricks.com

Cost Management and Forecasting (powered by UC System Tables)

Schemas
1,061
Schemas
2 hours ago

Tables
22,714
Tables
2 hours ago

Schemas
68
Catalogs
2 hours ago

Tables by format
2 hours ago

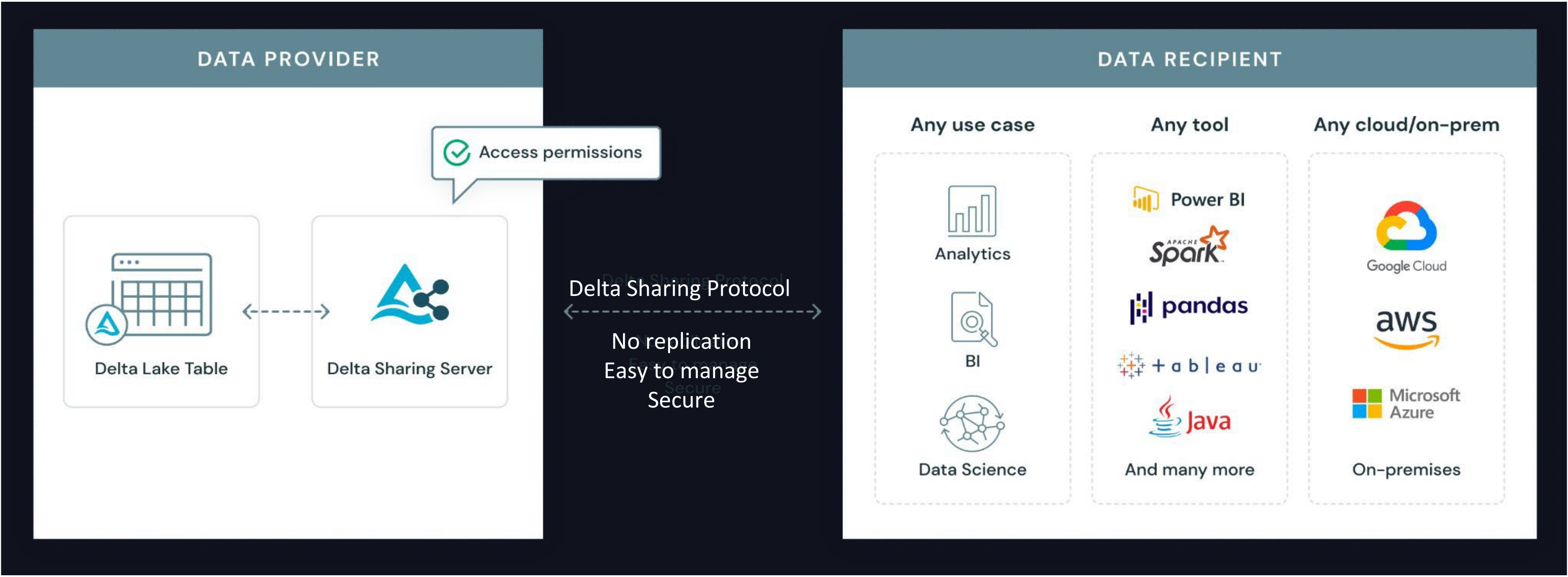
Tables Created (MoM)
2 hours ago

Volumes Created MoM
2 hours ago

Volumes Created MoM
2 hours ago

Tables Created by Type
2 hours ago

Open data sharing between teams and collaboration



DATA PROVIDER

Access permissions

Delta Lake Table

Delta Sharing Server

Delta Sharing Protocol

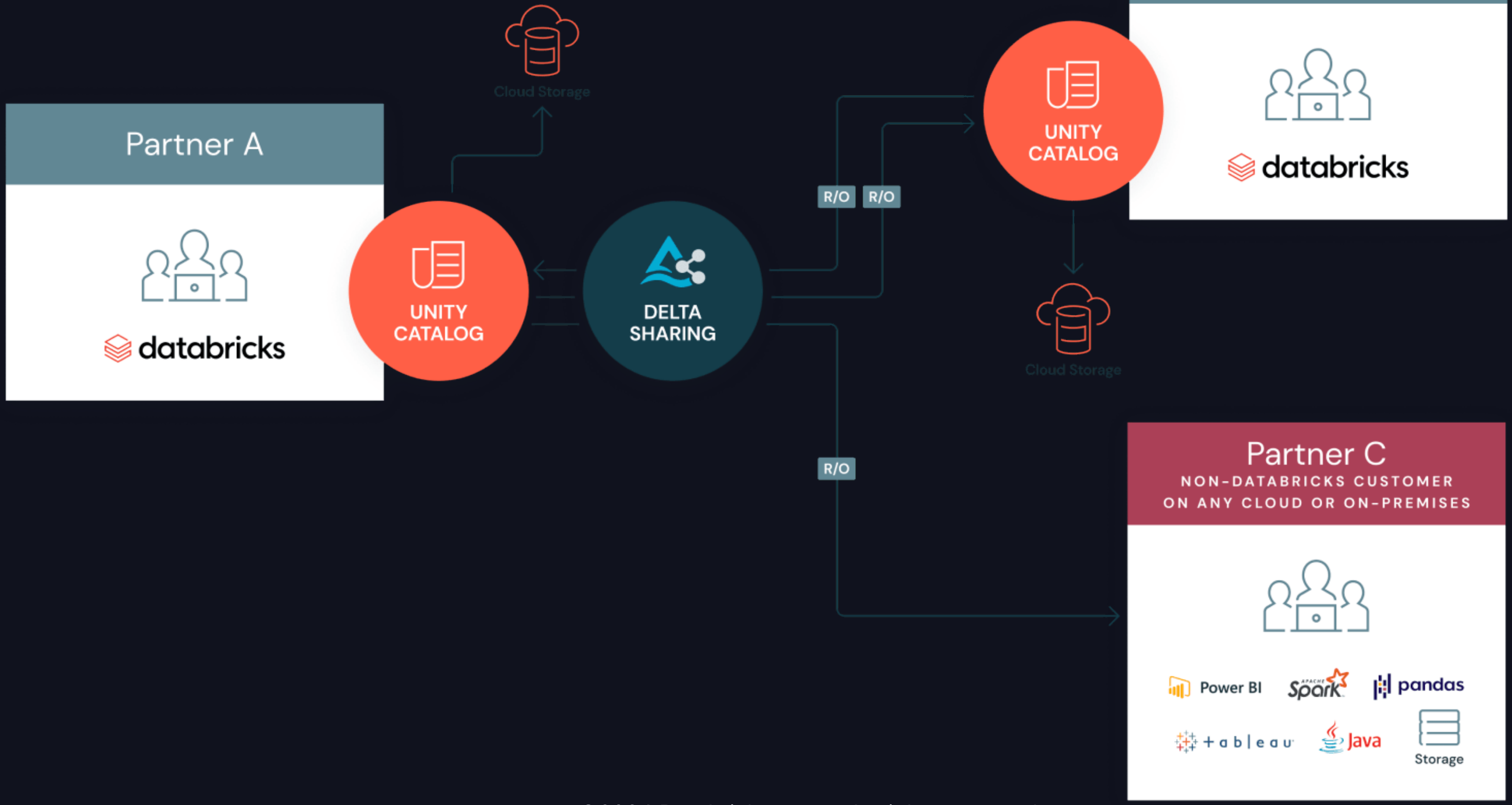
No replication
Easy to manage
Secure

DATA RECIPIENT

Any use case	Any tool	Any cloud/on-prem
Analytics BI Data Science	Power BI Spark pandas tableau Java And many more	Google Cloud aws Microsoft Azure On-premises

- Securely share live data across cloud regions, and data platforms and eliminate vendor lock-in
- Reduce operational overhead of maintaining multiple copies of data for different recipients
- Centrally manage, monitor usage and audit access to shared data with ease

Types of delta sharing



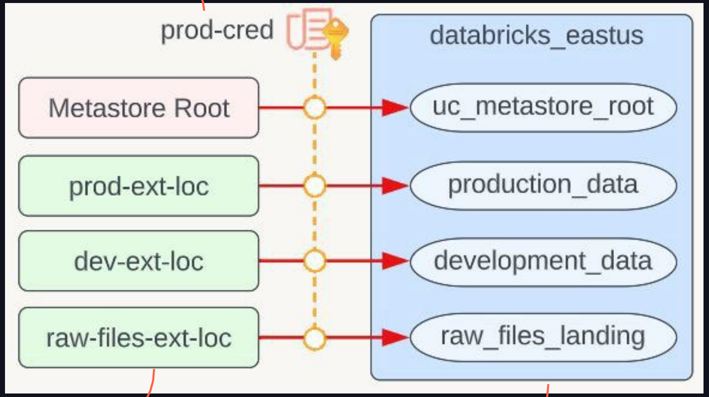
DATA+AI SUMMIT BY databricks

Questions

External Location Patterns

Comcast governance requirements drive the pattern

Simple

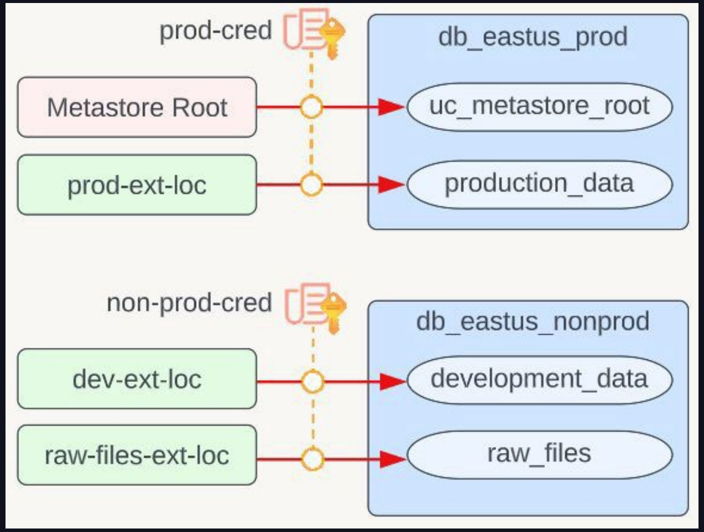


Storage Credential

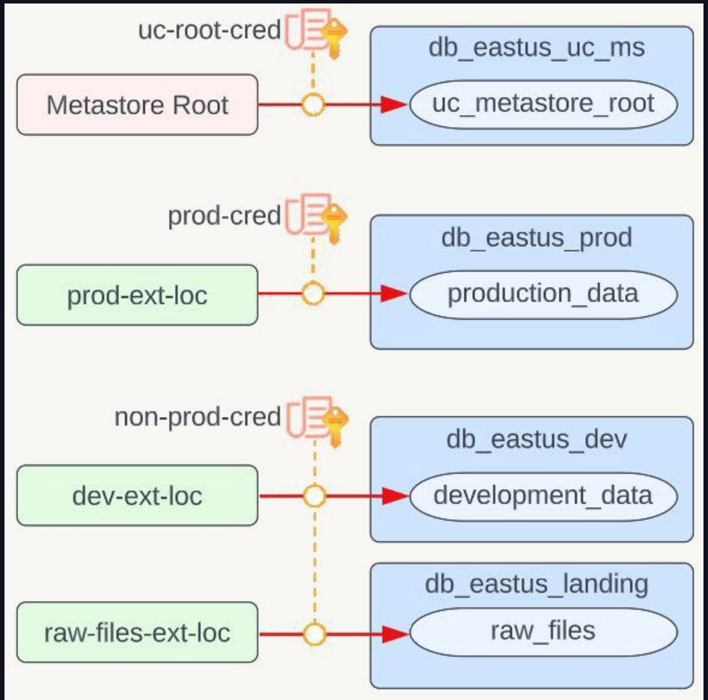
External Location

Storage Account

Prod/Non-Prod

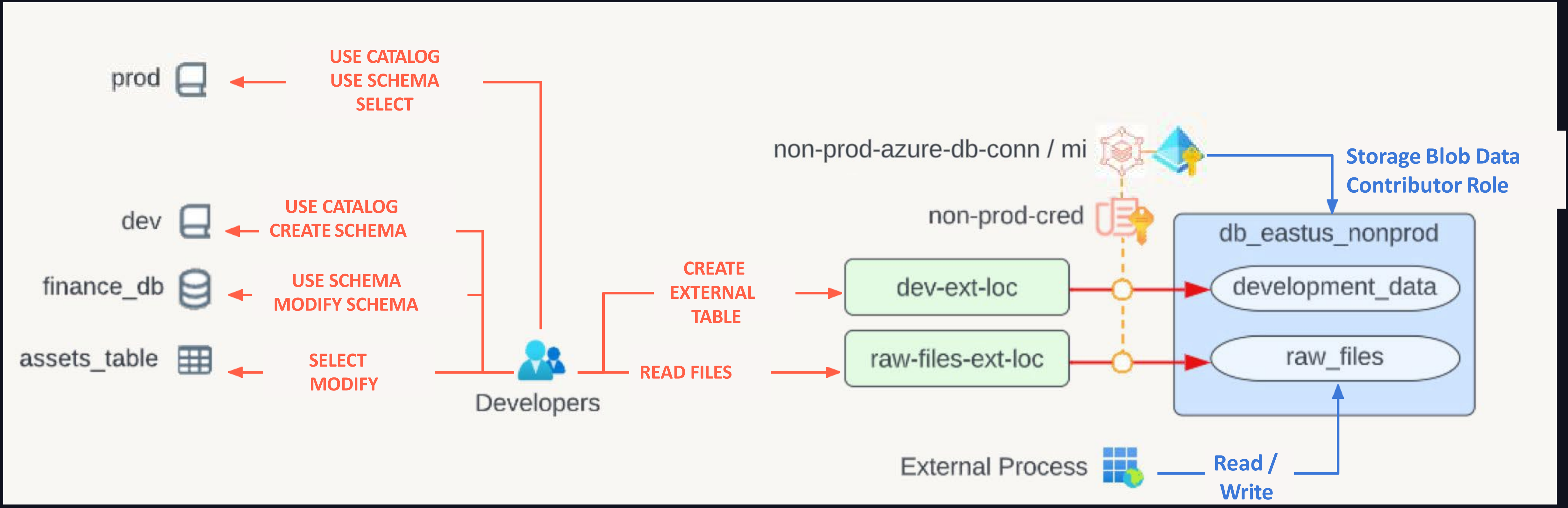


As Necessary

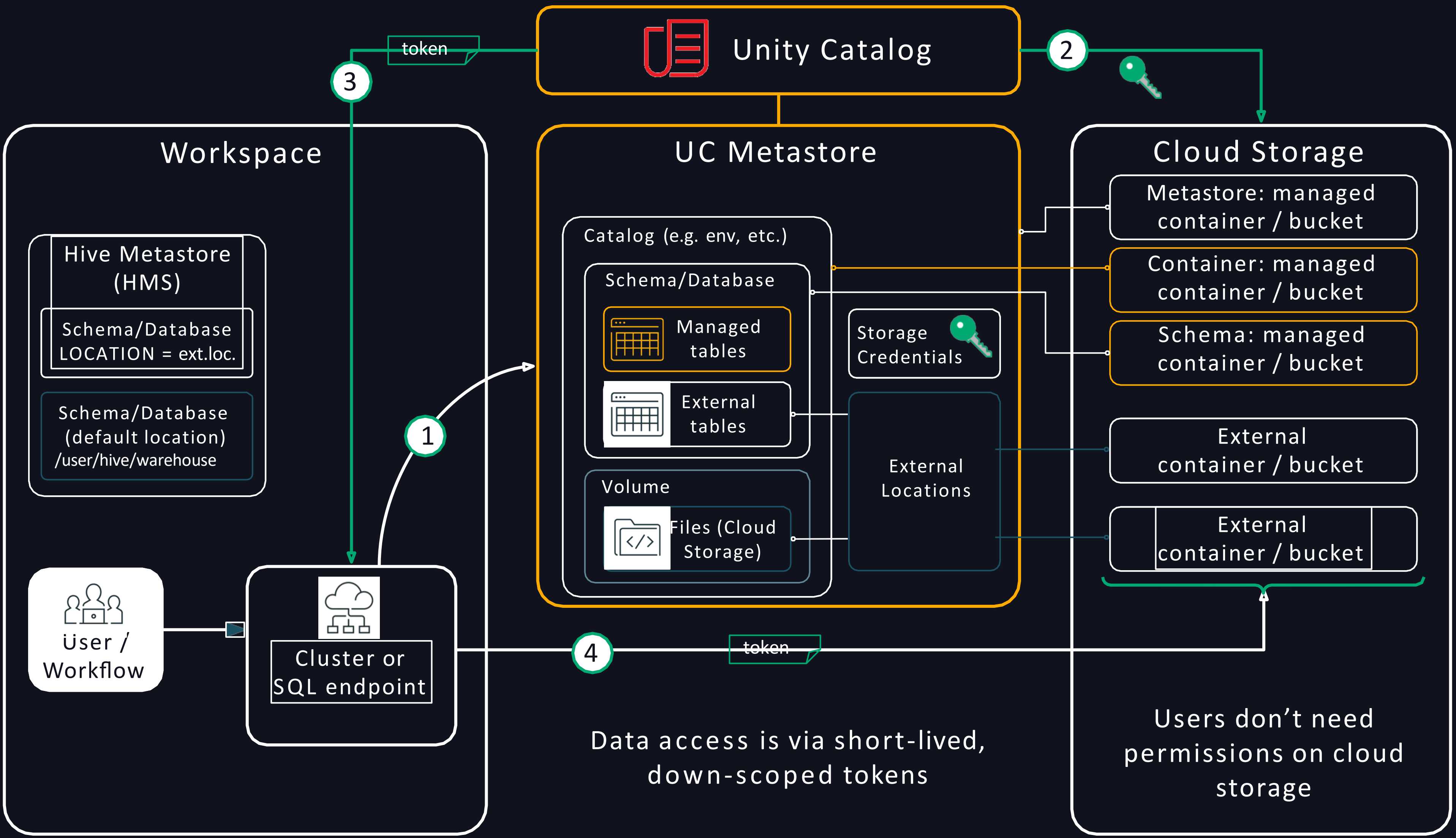


Permissions Example – Prod/Dev Catalogs

Developers need Unity Catalog permissions on External Locations to read non-UC files and create external tables



Data Access Model with Unity Catalog



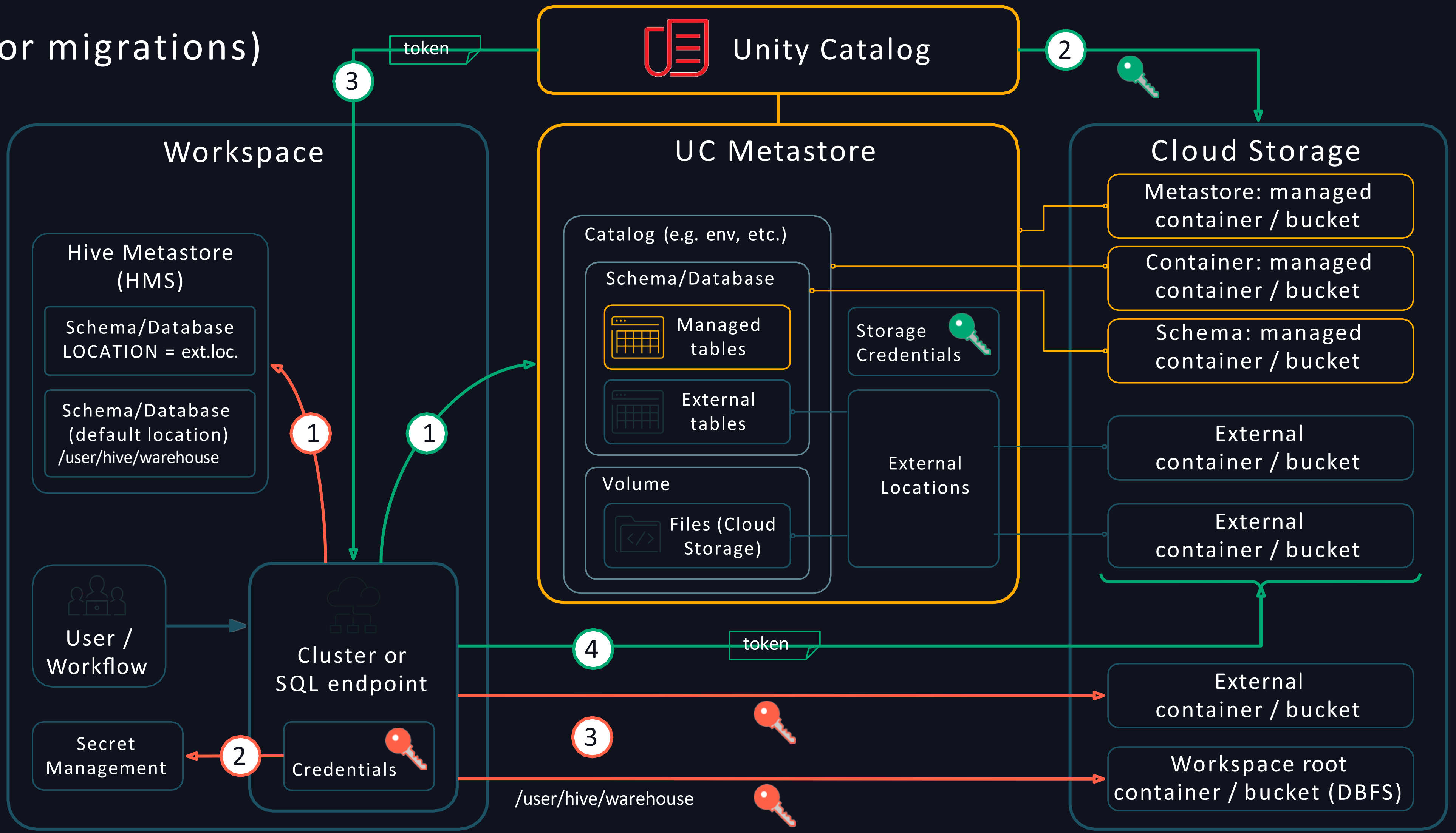
Unity Catalog
token
3
2
Workspace
UC Metastore
Cloud Storage
Hive Metastore (HMS)
Schema/Database LOCATION = ext.loc.
Schema/Database (default location) /user/hive/warehouse
Catalog (e.g. env, etc.)
Schema/Database
Managed tables
External tables
Volume
Files (Cloud Storage)
Storage Credentials
External Locations
Metastore: managed container / bucket
Container: managed container / bucket
Schema: managed container / bucket
External container / bucket
External container / bucket
1
User / Workflow
Cluster or SQL endpoint
4
token

Data access is via short-lived, down-scoped tokens

Users don't need permissions on cloud storage

Data Access Model combining both approaches

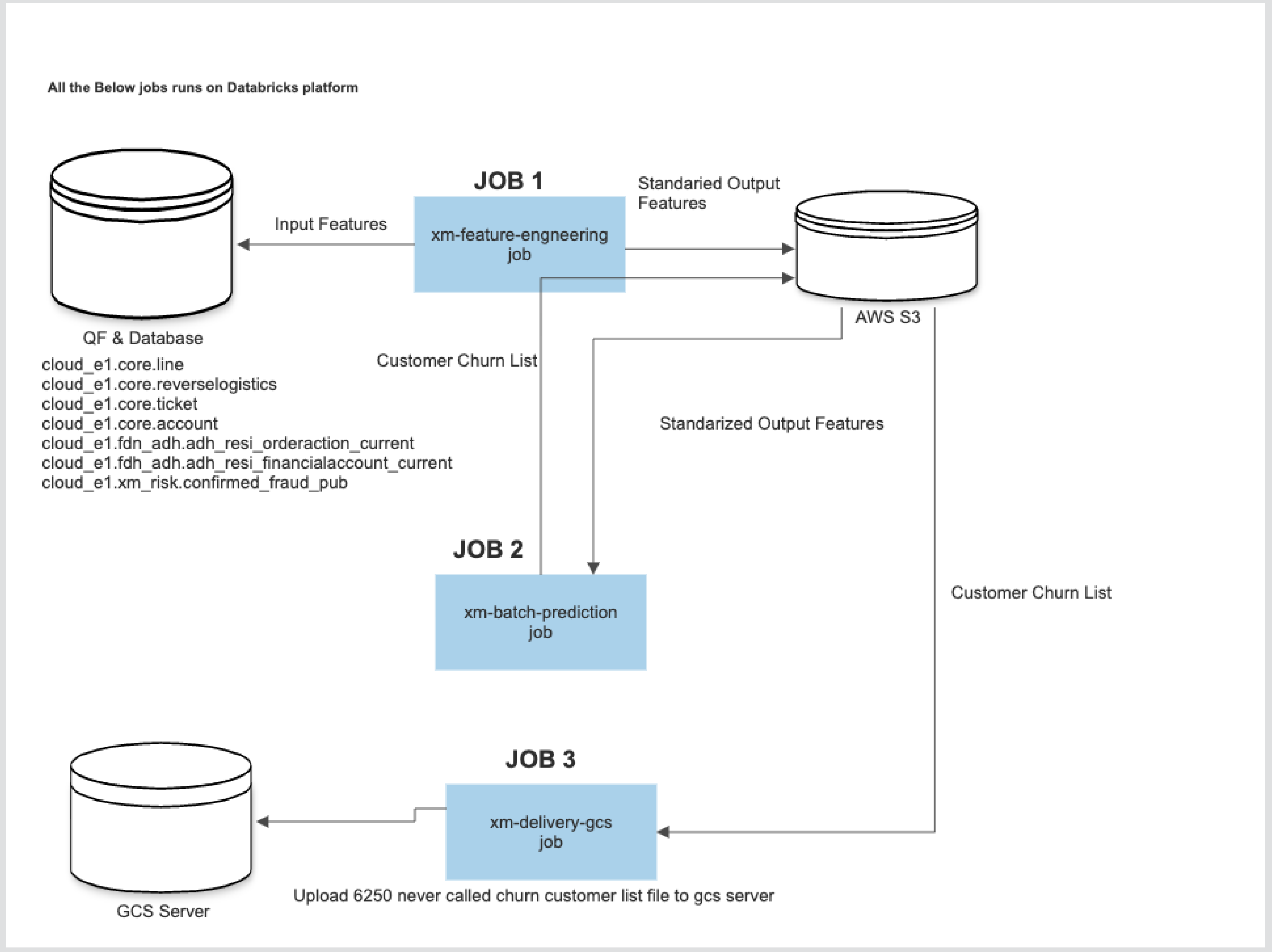
(e.g. for migrations)



token
3
Unity Catalog
2
Workspace
UC Metastore
Cloud Storage
Hive Metastore (HMS)
Schema/Database LOCATION = ext.loc.
Schema/Database (default location) /user/hive/warehouse
Catalog (e.g. env, etc.)
Schema/Database
Managed tables
External tables
Volume
Files (Cloud Storage)
Storage Credentials
External Locations
Metastore: managed container / bucket
Container: managed container / bucket
Schema: managed container / bucket
External container / bucket
External container / bucket
1
1
User / Workflow
Cluster or SQL endpoint
4
token
Secret Management
2
Credentials
3
External container / bucket
Workspace root container / bucket (DBFS)
/user/hive/warehouse

Advanced Use Cases

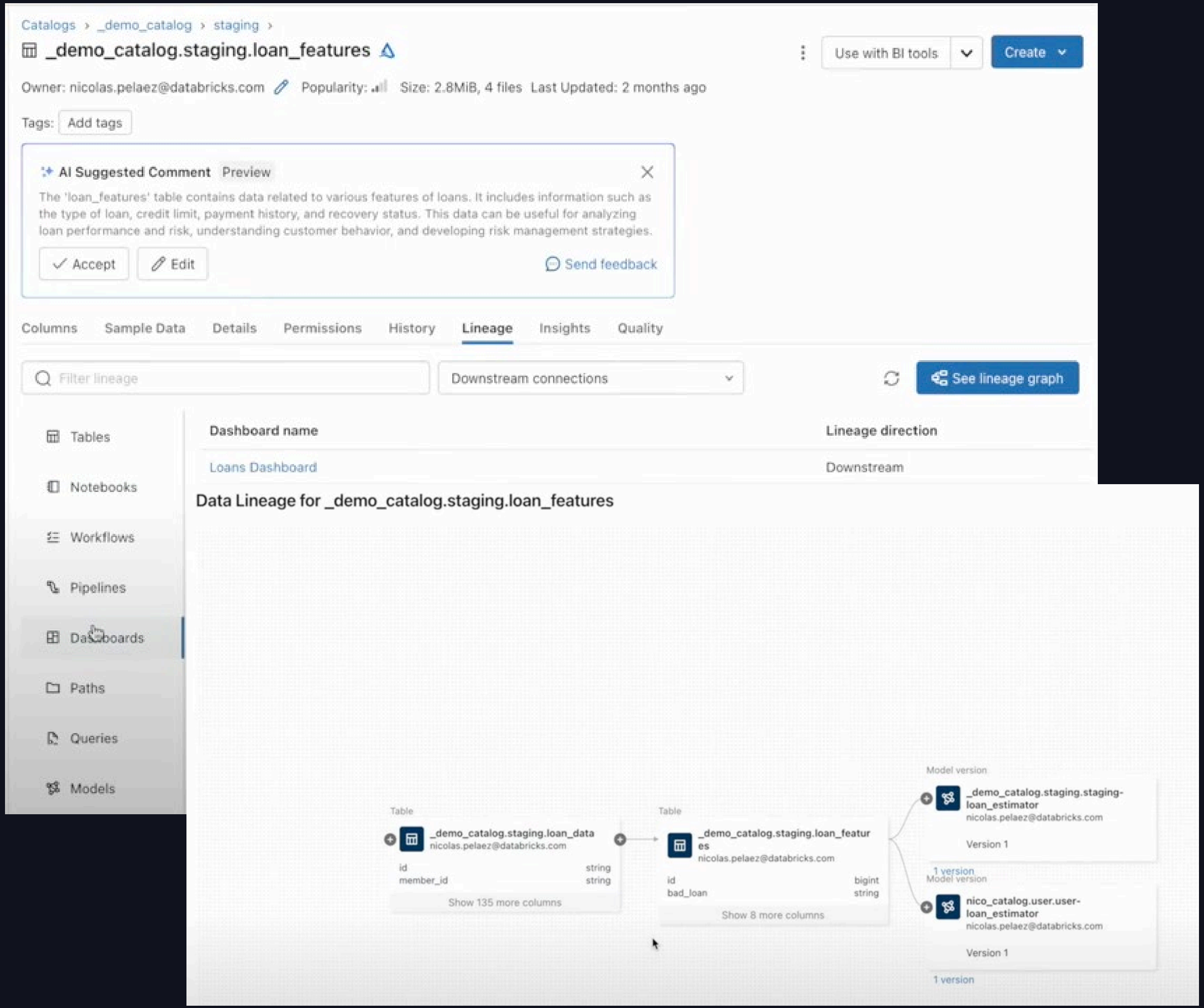
UC Machine Leaning Model for Customer Churn


All the Below jobs runs on Databricks platform
JOB 1
xm-feature-engneering job
Input Features
Standaried Output Features
QF & Database
cloud_e1.core.line
cloud_e1.core.reverselogistics
cloud_e1.core.ticket
cloud_e1.core.account
cloud_e1.fdn_adh.adh_resi_orderaction_current
cloud_e1.fdh_adh.adh_resi_financialaccount_current
cloud_e1.xm_risk.confirmed_fraud_pub
AWS S3
Customer Churn List
Standarized Output Features
JOB 2
xm-batch-prediction job
Customer Churn List
JOB 3
xm-delivery-gcs job
GCS Server
Upload 6250 never called churn customer list file to gcs server

DATA+AI SUMMIT

Data observability and quality with Lineage

- End-to-end visibility into how data flows in my org with automated data lineage down to column level
- View lineage across tables, columns, notebooks, workflows, dashboards
- Captured in real time across all workloads in SQL, Python, Scala, and R
- Integration with enterprise catalogs



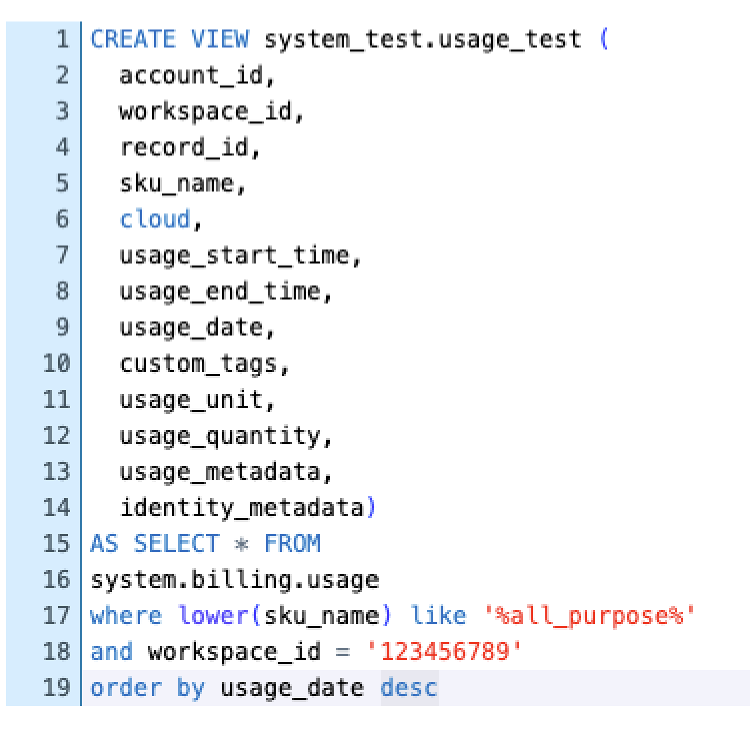
Owner: nicolas.pelaez@databricks.com Popularity: Size: 2.8MiB, 4 files Last Updated: 2 months ago

The 'loan_features' table contains data related to various features of loans. It includes information such as the type of loan, credit limit, payment history, and recovery status. This data can be useful for analyzing loan performance and risk, understanding customer behavior, and developing risk management strategies.

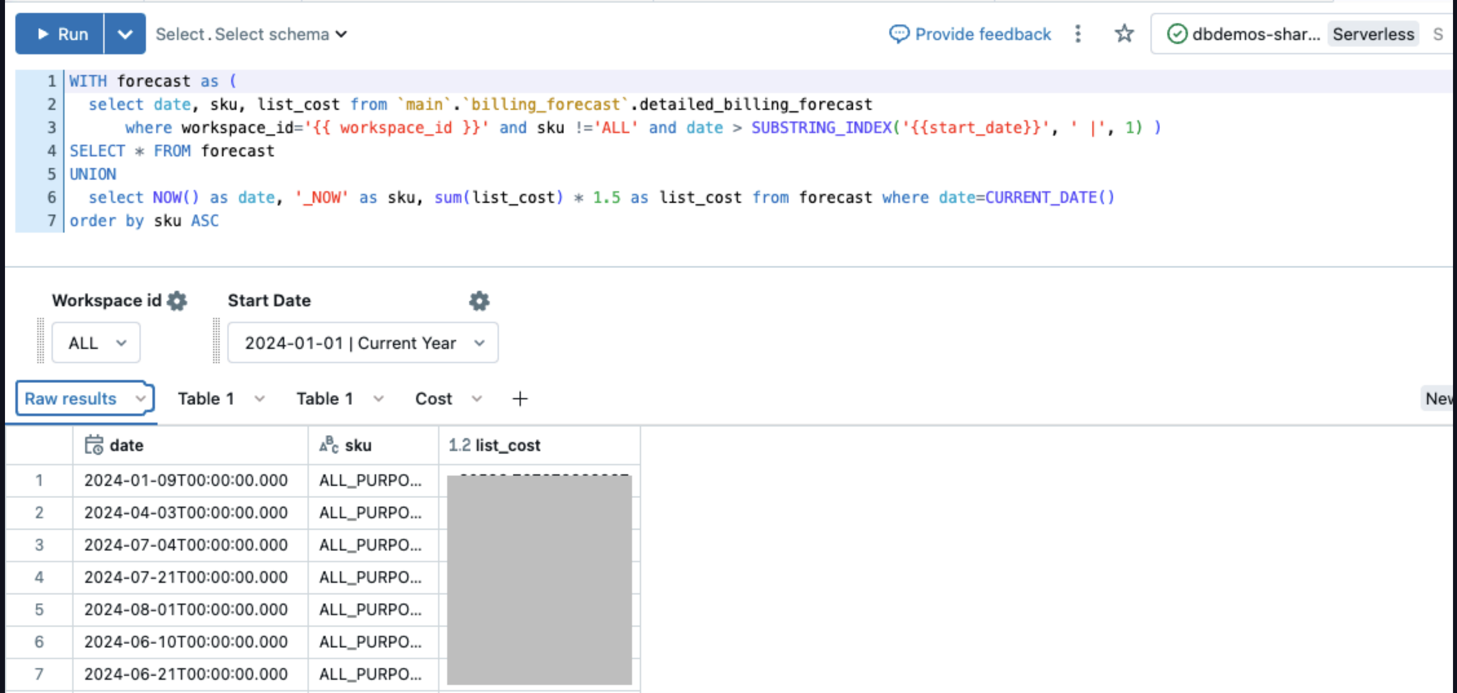
Dashboard name	Lineage direction
Loans Dashboard	Downstream

Data Lineage for _demo_catalog.staging.loan_features

Billing Usage Sample View, Queries



```sql
CREATE VIEW system_test.usage_test (
  account_id,
  workspace_id,
  record_id,
  sku_name,
  cloud,
  usage_start_time,
  usage_end_time,
  usage_date,
  custom_tags,
  usage_unit,
  usage_quantity,
  usage_metadata,
  identity_metadata)
AS SELECT * FROM
system.billing.usage
where lower(sku_name) like '%all_purpose%'
and workspace_id = '123456789'
order by usage_date desc
```



```sql
WITH forecast as (
  select date, sku, list_cost from `main`.`billing_forecast`.detailed_billing_forecast
      where workspace_id='{{ workspace_id }}' and sku !='ALL' and date > SUBSTRING_INDEX('{{start_date}}', ' |', 1) )
SELECT * FROM forecast
UNION
  select NOW() as date, '_NOW' as sku, sum(list_cost) * 1.5 as list_cost from forecast where date=CURRENT_DATE()
order by sku ASC
```

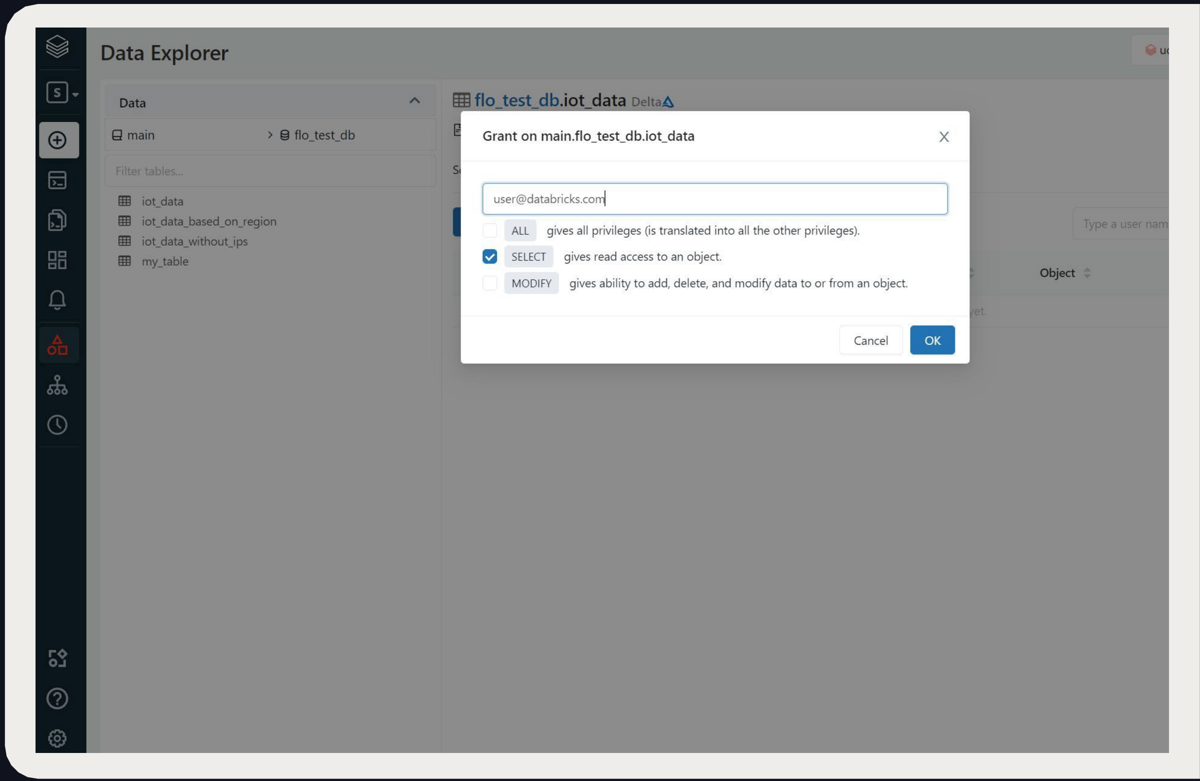
Workspace id: ALL

Start Date: 2024-01-01 | Current Year

	date	sku	1.2 list_cost
1	2024-01-09T00:00:00.000	ALL_PURPO...	[redacted]
2	2024-04-03T00:00:00.000	ALL_PURPO...	[redacted]
3	2024-07-04T00:00:00.000	ALL_PURPO...	[redacted]
4	2024-07-21T00:00:00.000	ALL_PURPO...	[redacted]
5	2024-08-01T00:00:00.000	ALL_PURPO...	[redacted]
6	2024-06-10T00:00:00.000	ALL_PURPO...	[redacted]
7	2024-06-21T00:00:00.000	ALL_PURPO...	[redacted]

Map, secure and audit data across clouds

- Catalog all data, analytics and AI assets and create a unified view of team's entire data estate
- Centrally manage access permissions and audit controls for files, tables across all workspaces and workloads using a familiar interface based on ANSI SQL



Grant on main.flo_test_db.iot_data

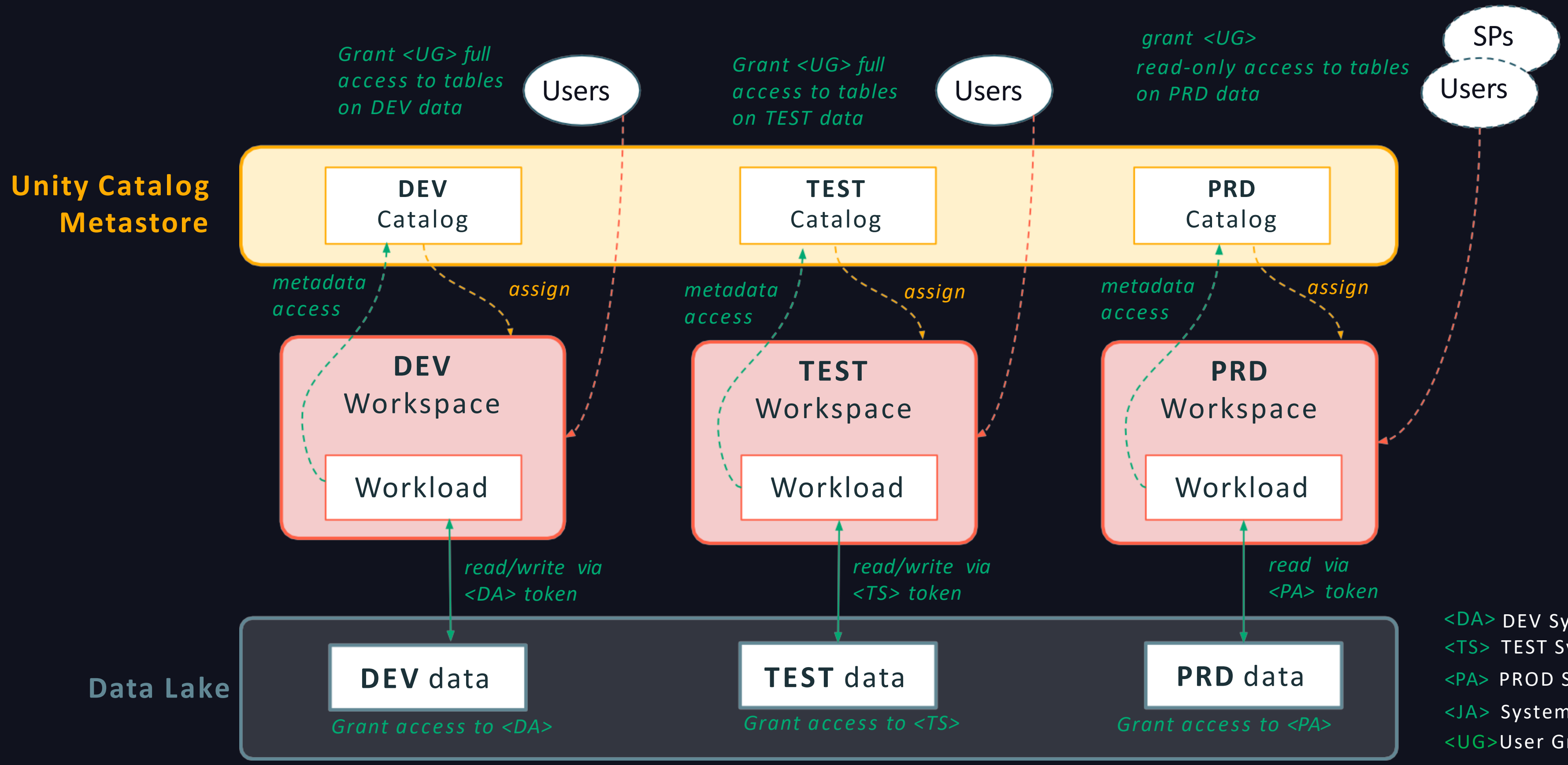
user@databricks.com

□ ALL gives all privileges (is translated into all the other privileges).

☑ SELECT gives read access to an object.

□ MODIFY gives ability to add, delete, and modify data to or from an object.

Software Development Lifecycle setup with UC



Grant <UG> full access to tables on DEV data

Grant <UG> full access to tables on TEST data

grant <UG> read-only access to tables on PRD data

Users

Users

SPs

Users

Unity Catalog Metastore

DEV Catalog

TEST Catalog

PRD Catalog

metadata access

assign

metadata access

assign

metadata access

assign

DEV Workspace

Workload

TEST Workspace

Workload

PRD Workspace

Workload

read/write via <DA> token

read/write via <TS> token

read via <PA> token

Data Lake

DEV data

Grant access to <DA>

TEST data

Grant access to <TS>

PRD data

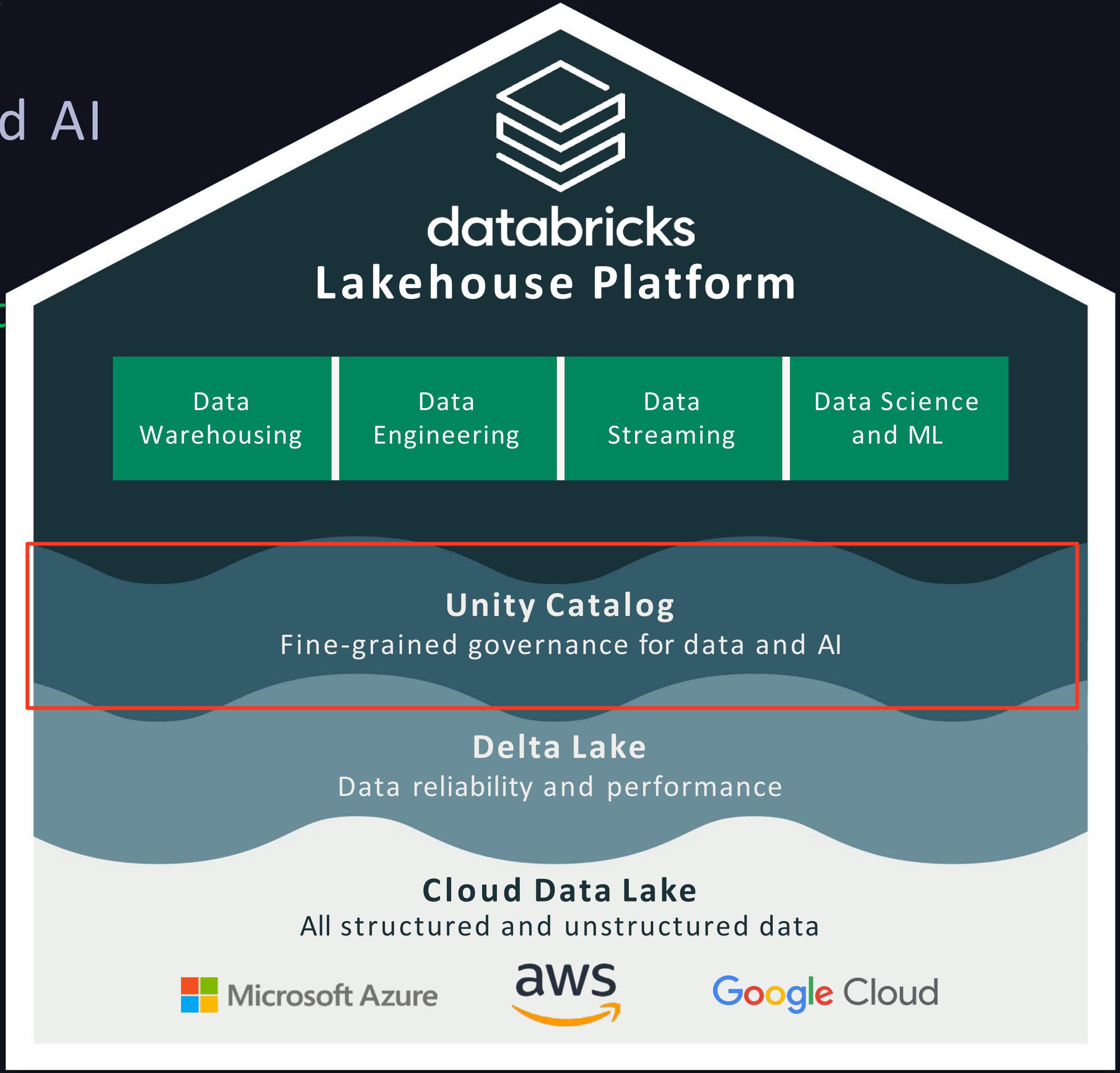
Grant access to <PA>

<DA>	DEV System Account for UC
<TS>	TEST System Account for UC
<PA>	PROD System Account for UC
<JA>	System Account to execute jobs in PRD
<UG>	User Group working in DEV, TEST and Prod

Databricks Unity Catalog

Unified governance for data, analytics and AI

- Centralized Identity and access management
- Simplified administration
- Centralized governance for data and AI
- Enhanced auditing and data sharing
- Built-in data search and discovery
- Performance and scale
- Automated lineage for all workloads



databricks

Lakehouse Platform

Data Warehousing | Data Engineering | Data Streaming | Data Science and ML

Unity Catalog
Fine-grained governance for data and AI

Delta Lake
Data reliability and performance

Cloud Data Lake
All structured and unstructured data