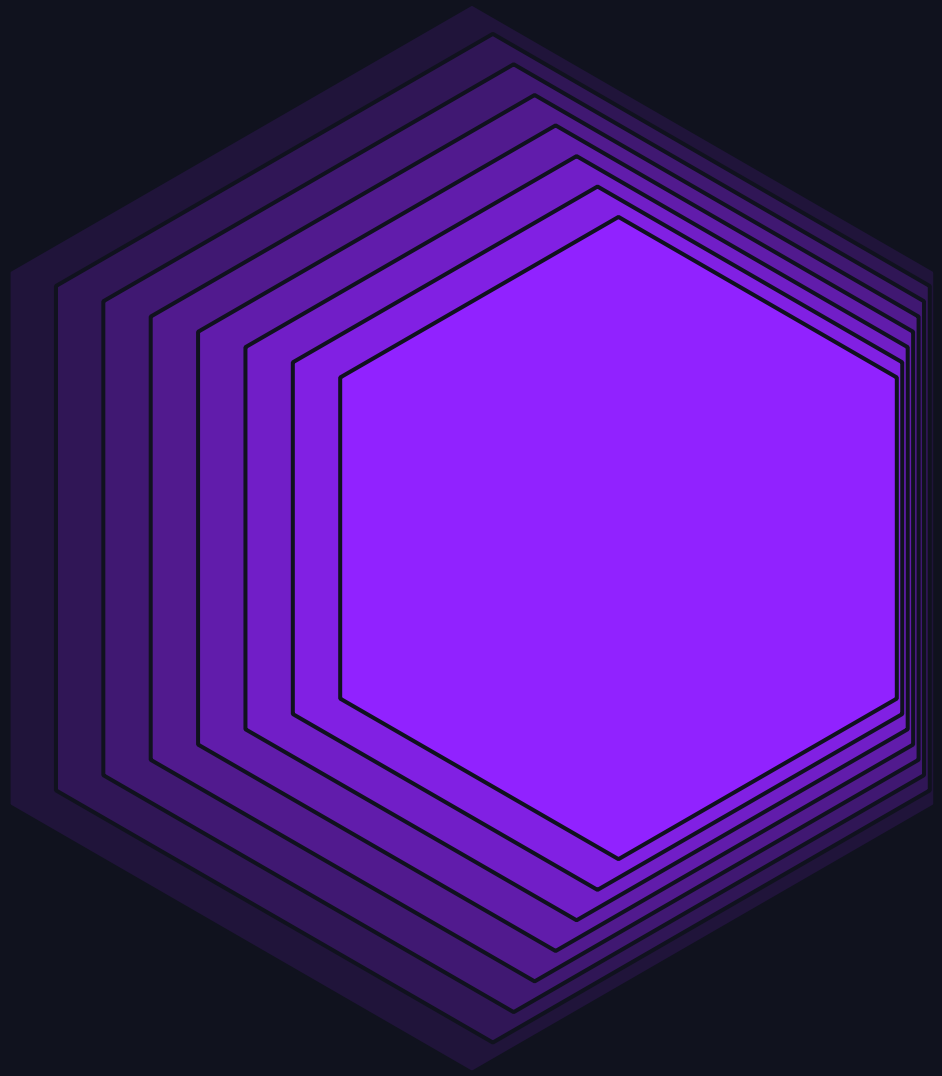
DATA+AI SUMMIT BY databricks

Open Data Lakehouse with Unity Catalog Open Source



Ramesh Chandra and Michelle Leon
June, 2024

Who Are We?

Michelle Leon

- Staff Product Manager
  - Previously Webflow, Airbnb
- Based in San Francisco
- Talk to me about
  - Delta Lake
  - Unity Catalog interoperability
  - Best burritos in the Mission neighborhood 🌯

Who Are We?

Ramesh Chandra

- Principal Software Engineer
  - Previously Google, Nutanix
- Based in Mountain View
- Talk to me about
  - Unity Catalog
  - Governance
  - Sharing

Agenda

What is a catalog?

Challenges today

Unity Catalog OSS

Demos

Roadmap ahead

What is a catalog?

A great catalog enables you to

1. Manage data and AI assets in one place
2. Govern assets through a single source of truth
3. Leverage best-of-breed tools with your data

Catalogs are the source of truth for critical properties about your assets

Catalogs are the source of truth for critical properties about your assets

Table metadata for lakehouse formats

- Table name
- Schema
- Storage location
- Storage format
- Created at
- Type

Catalogs are the source of truth for critical properties about your assets

Table metadata for lakehouse formats

- Table name
- Schema
- Storage location
- Storage format
- Created at
- Type

Governance and access control

- Roles and permissions
- Users / groups
- Data masking policies
- Access logs
- Compliance tags
- Credential management

Catalogs are the source of truth for critical properties about your assets

Table metadata for lakehouse formats

- Table name
- Schema
- Storage location
- Storage format
- Created at
- Type

Governance and access control

- Roles and permissions
- Users / groups
- Data masking policies
- Access logs
- Compliance tags
- Credential management

AI/ML object metadata

- Model name
- Model version
- Inference endpoints
- GenAI tool registry
- AI Gateway
- Index name
- Vector dimensionality

Catalogs are the source of truth for critical properties about your assets

Table metadata for lakehouse formats

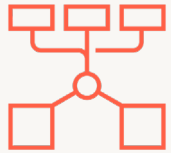
- Table name
- Schema
- Storage location
- Storage format
- Created at
- Type

Governance and access control

- Roles and permissions
- Users / groups
- Data masking policies
- Access logs
- Compliance tags
- Credential management

AI/ML object metadata

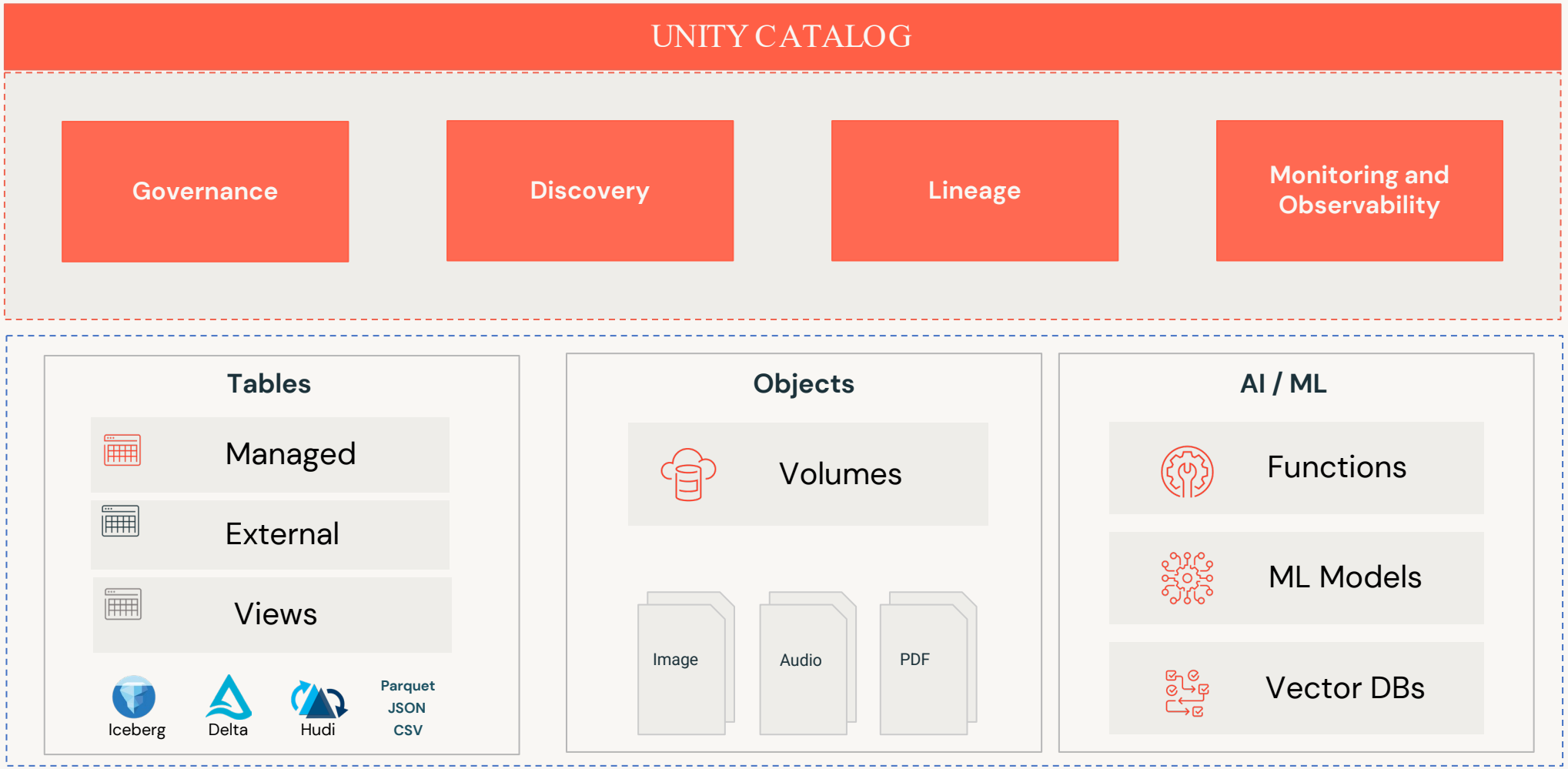
- Model name
- Model version
- Inference endpoints
- GenAI tool registry
- AI Gateway
- Index name
- Vector dimensionality



Complex transaction management

- Table version
- Latest commit
- Multi-statement txns
- Multi-table txns

Catalogs becomes the source of truth for metadata, governance, and more for data and AI assets



How do various tools and platforms interoperate with my catalog?

Challenges today

Challenges today

Most cloud data platforms lack open access

Data and AI assets are arbitrarily siloed

Governance across Data + AI is inconsistent and hard

Most Cloud Platforms Lack Open Access

Many cloud DWs have “native tables” not in open format

Proprietary catalogs further limit access to metadata

Need always-on compute for external engine access

Data and AI assets are arbitrarily siloed

Tabular data

product_info/

<xyz>.parquet

customer_reviews/

<xyz>.parquet

Unstructured data

product_images/

img_001.jpg

ugc_review_images/

img_01234.jpg

Data and AI assets are arbitrarily siloed

Tabular data

product_info/

customer_reviews/

Unstructured data

product_images/

ugc_review_images/

Models

- text_sentiment_model_v1
- text_sentiment_model_v2
- image_quality_assessment_model_v1
- customer_segmentation_model_v1

Data and AI assets are arbitrarily siloed

Tabular data

product_info/

<xyz>.parquet

customer_reviews/

<xyz>.parquet

Unstructured data

product_images/

img_001.jpg

ugc_review_images/

img_01234.jpg

Models

- text_sentiment_model_v1
- text_sentiment_model_v2
- image_quality_assessment_model_v1
- customer_segmentation_model_v1

Functions

```
generate_sentiment_score(
  review_text: str
) -> float
```

```
extract_product_info(
  product_id: int
) -> dict
```

Governance is inconsistent and hard

Existing open catalogs lack governance support

- Encourages pattern of giving external engines direct storage access, which bypasses governance

Governance is inconsistent and hard

jane@acme.co

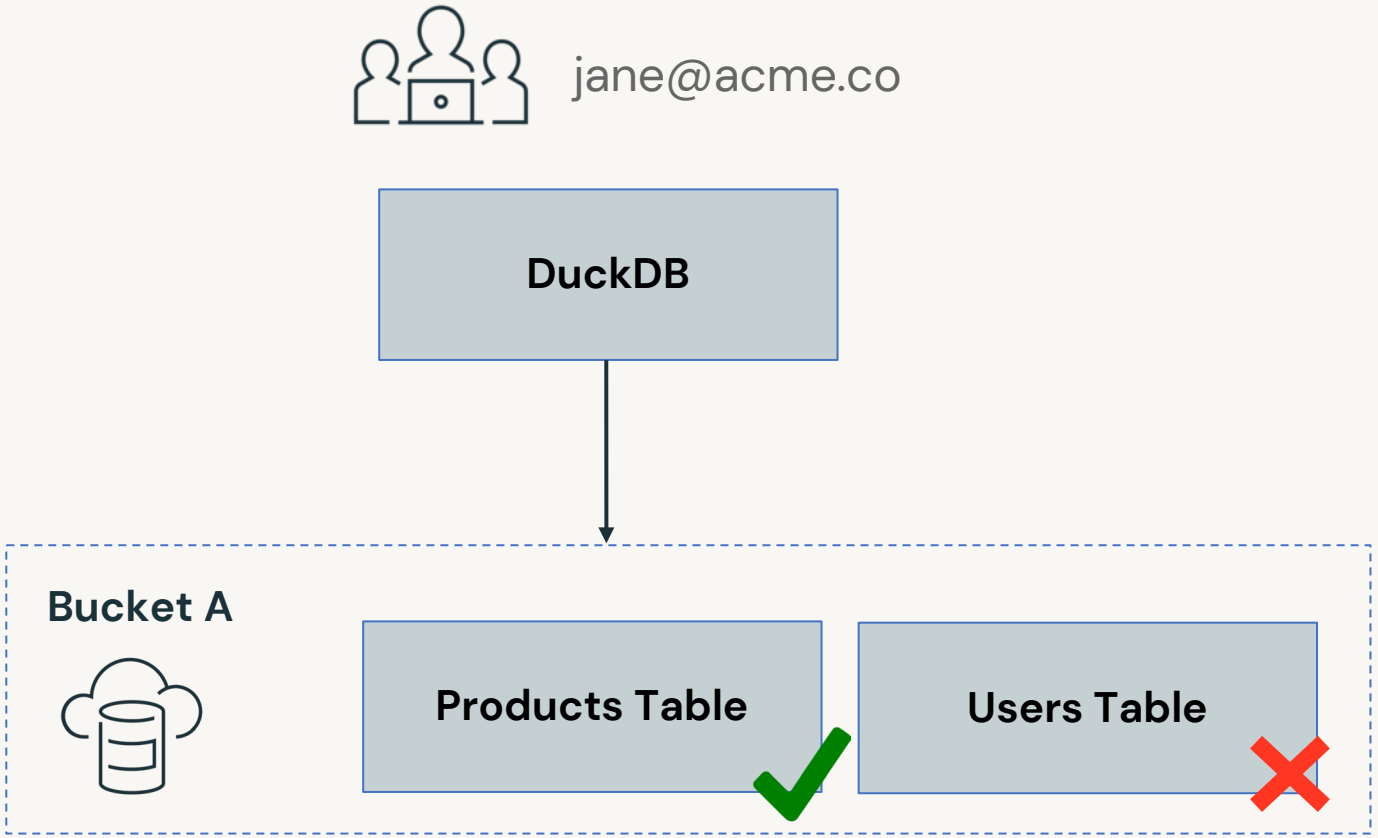
DuckDB

Bucket A

Products Table

Users Table

Governance is inconsistent and hard



Jane needs SELECT access to only the Products Table

Granting storage access also gives access to Users Table

Governance is inconsistent and hard

Data and AI silos also lead to fragmented governance

- Duplicate access policies increase burden and cause security bugs
- Auditing is hard – admins need to stitch together audit logs
- Silos can cause lineage gaps

We built Unity Catalog to address these problems

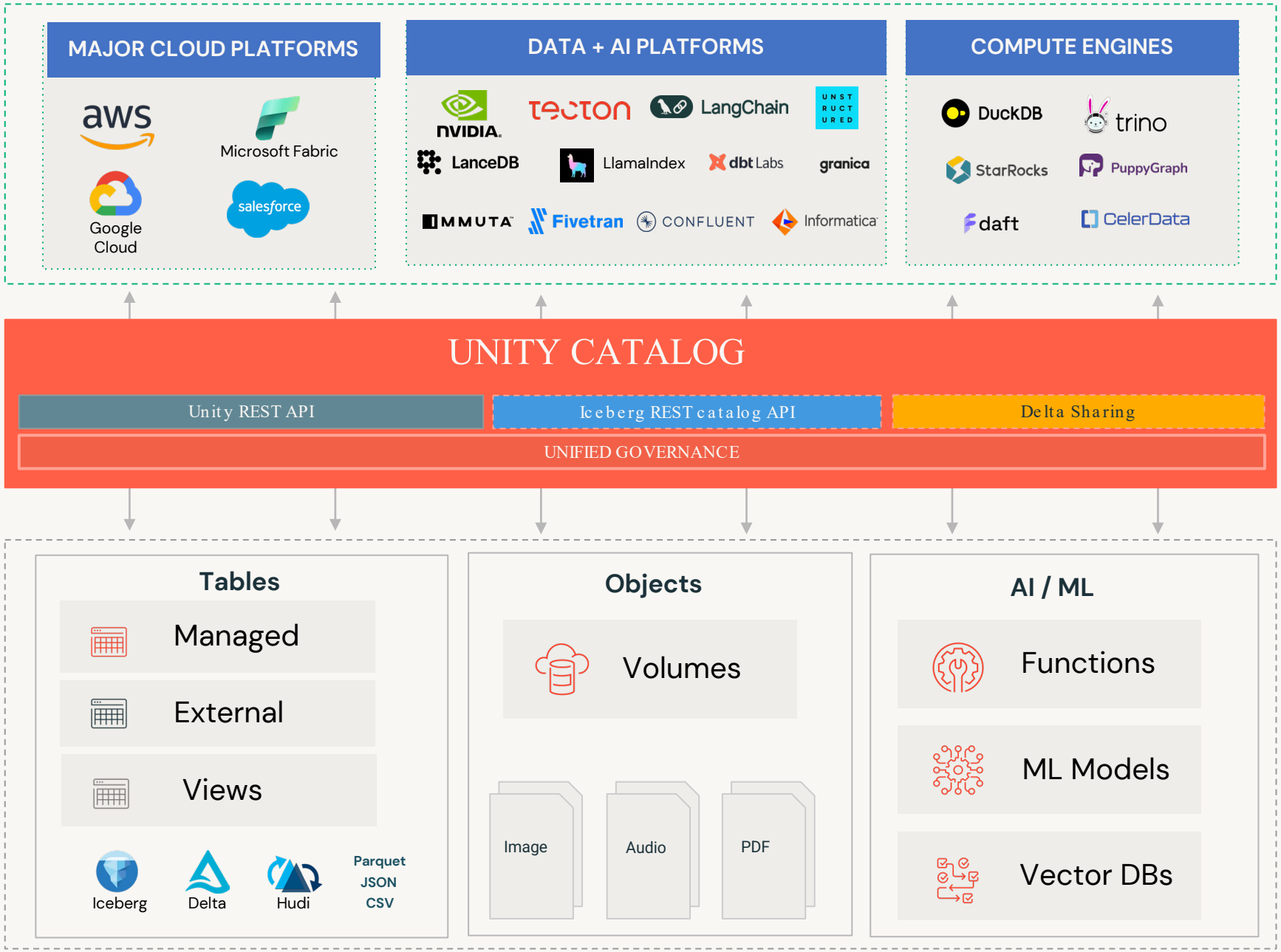
Unity Catalog: The industry's only universal catalog for Data and AI

Any engine
Client ecosystem

Any client
Universal standard

Any asset
Data + AI assets

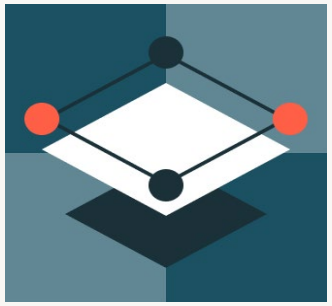
Any format
UniForm



And now Unity Catalog is open source!

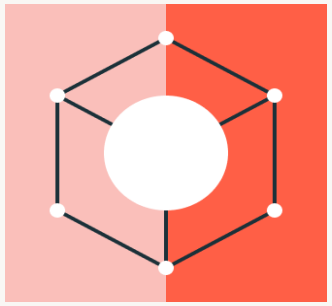
Introducing

Unity Catalog Open Source



Open

Open APIs and OSS server maximize flexibility and customer choice



Interoperable

Universal interface supports any format, engine, data and AI asset

Unified

Unified governance across tabular, non-tabular data and AI assets

Available today

Available today

OpenAPI spec

Managed tables APIs

External tables APIs

Volumes APIs

Functions APIs

Credential vending APIs

OSS server

OSS server

Available in new Unity Catalog Github repo

New developer resources

Unity Catalog OSS SDK

REST API docs

Updated Databricks SDKs (Java, Python, Go)

Secure credential vending

Unity Catalog secure credential vending

Available on Databricks in Private Preview

What this means for Databricks customers

Commitment to open source
Both Unity Catalog in Databricks and Unity Catalog OSS will co-evolve with richer functionality

What this means for Databricks customers

Commitment to open source
Both Unity Catalog in Databricks and Unity Catalog OSS will co-evolve with richer functionality

Full interoperability from Day 1
External client access to *all* tables and unstructured data in Unity Catalog
Read and write from any external engine or platform

True interoperability can only be unlocked with a rich ecosystem

The most open and interoperable catalog for data and AI

{Iceberg REST catalog APIs }

ICEBERG ECOSYSTEM

Amazon Web Services, Microsoft Azure, Google Cloud, Snowflake, Salesforce, Apache Impala, Clickhouse, Daft, Doris, Dremio, Flink, Hive, Presto, Apache Spark, StarRocks, Trino

Unity Catalog

DELTA ECOSYSTEM

Amazon Web Services, Azure, Google Cloud, Airbyte, Arrow, Athena, Apache Ballista, aws-pandas-sdk, beam, Daft, Dask, Apache DataFusion, Delta-RS, DLT (spark-r), DuckDB, EMR, Flink, Glue, Hive, Pandas, Polars, PrestoDB, Pulsar, PuppyGraph, Python, Ray, Rust, Apache Spark, StarRocks, Startree (pinot), Trino

UNITY CATALOG PARTNERS

Amazon Web Services, Azure, Google Cloud, Salesforce, NVIDIA, CONFLUENT, daft, databricks, dbt Labs, DuckDB, Fivetran, granica, IMMUTA, Informatica, LanceDB, LangChain, LlamaIndex, ONEHOUSE, PuppyGraph, StarRocks, tecton, UNSTRUCTURED, Apache XTable

LF AI & DATA

SANDBOX PROJECT

Unity Catalog

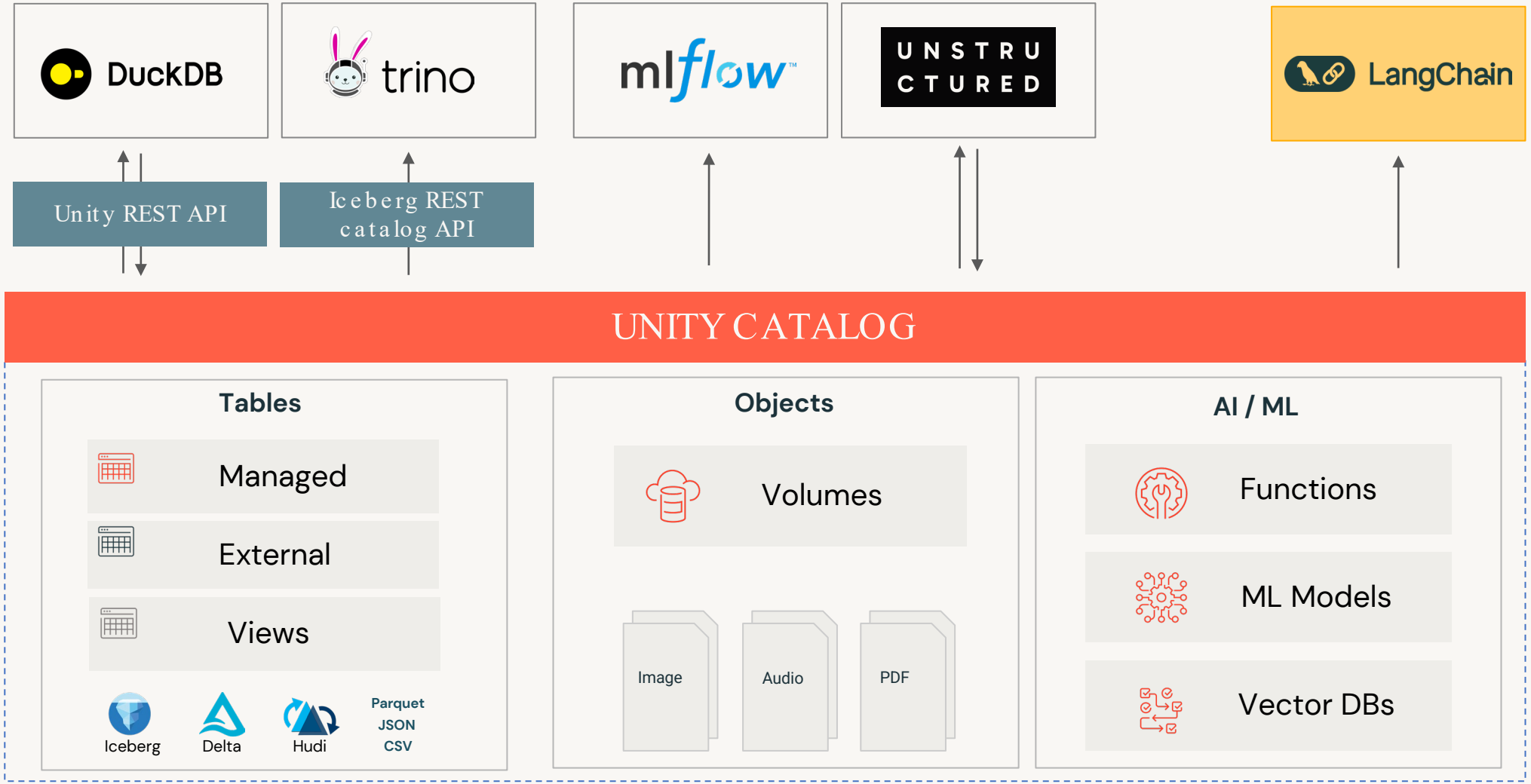
Demo Time

External access by vending temp credentials - enabling asset-level access control

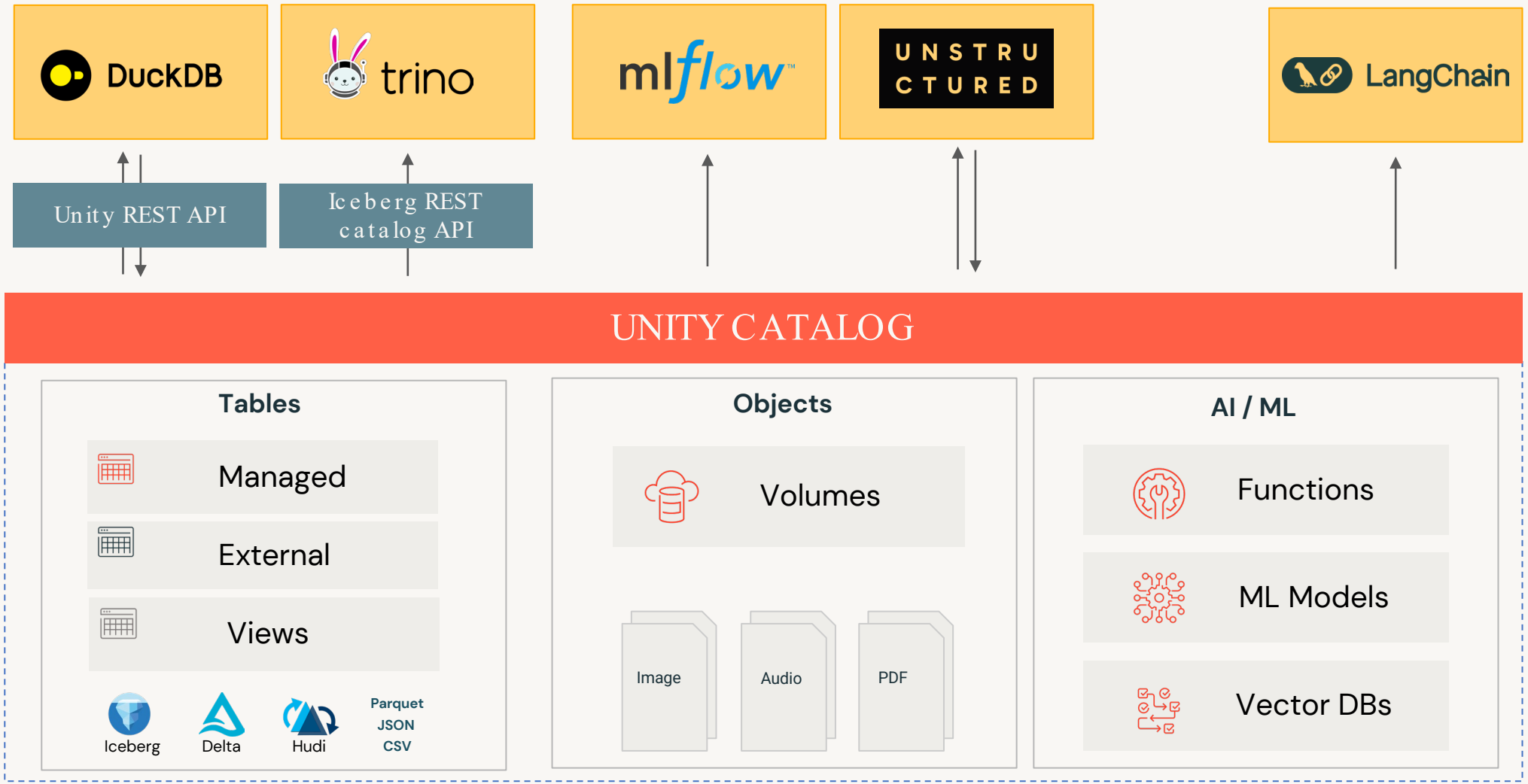
External access to Unity Catalog tables for Iceberg readers

Leverage multimodal data in Unity Catalog Volumes for AI/ML training and applications

Register functions in a central catalog to standardize AI tool use in your organization



Unity Catalog OSS enables broad interoperability



What’s next

Roadmap ahead

Enhanced support for data and AI assets in UC OSS v1.0

Format-agnostic table write APIs

Views (multi-dialect, beyond Spark)

Delta Sharing

Models

Remote Functions (OpenAPI endpoints)

Access Control APIs

**Coming soon with multi-cluster writes to UC managed tables*

Collaborate with us!

unitycatalog.io

github.com/unitycatalog/unitycatalog

Unity Catalog

go.unitycatalog.io/<>

go.unitycatalog.io/slack

go.unitycatalog.io/GitHub

go.unitycatalog.io/user

go.unitycatalog.io/dev

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