databricks

DATA+AI WORLD TOUR

What's new in Unity Catalog

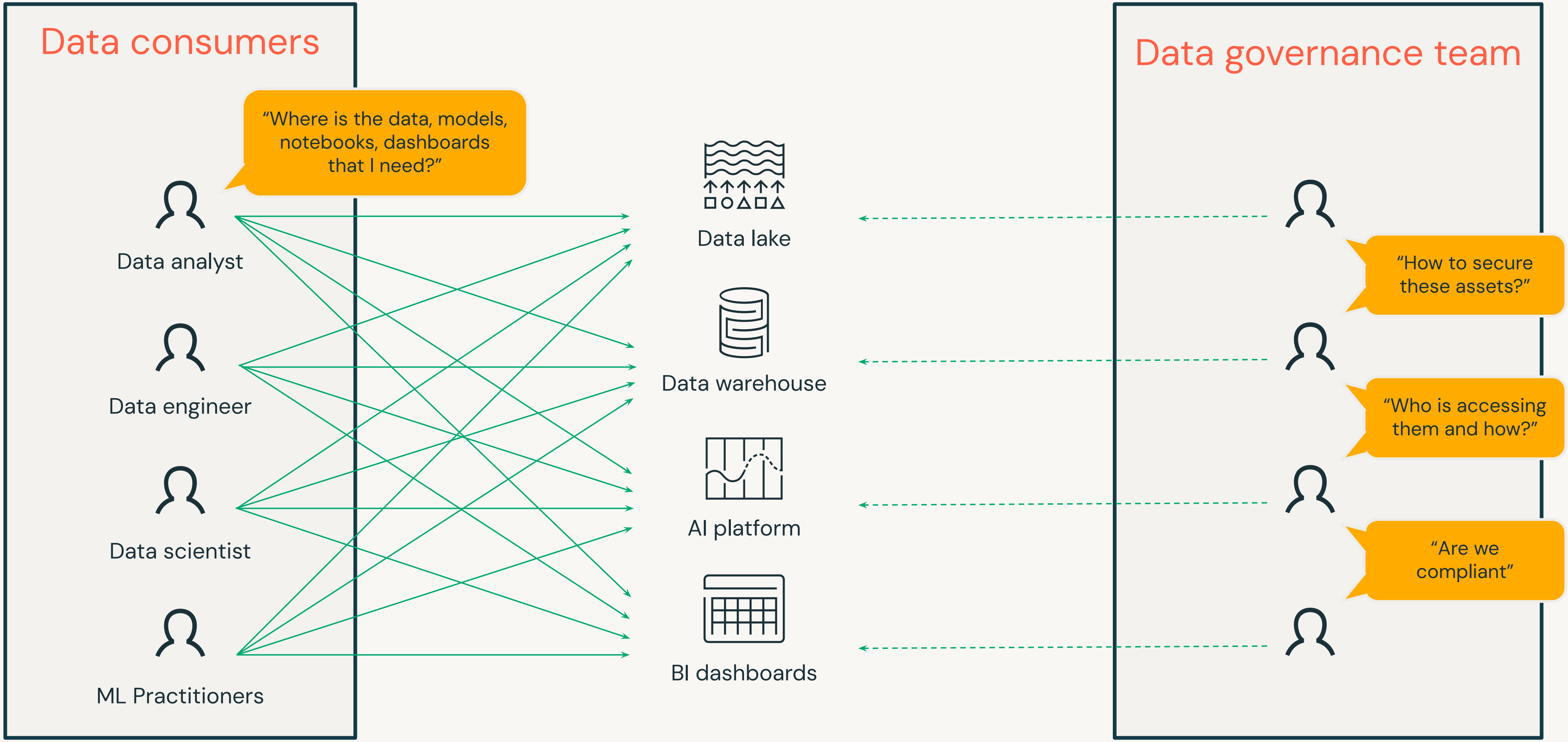
Dalya Altaha, Solution Architect @ Databricks
Zen Van Gaever, Databricks Data Engineer @ delaware

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.

Permissions on tables, [illegible]

Data and AI governance is complex



Data consumers

"Where is the data, models, notebooks, dashboards that I need?"

Data analyst

Data engineer

Data scientist

ML Practitioners

Data lake

Data warehouse

AI platform

BI dashboards

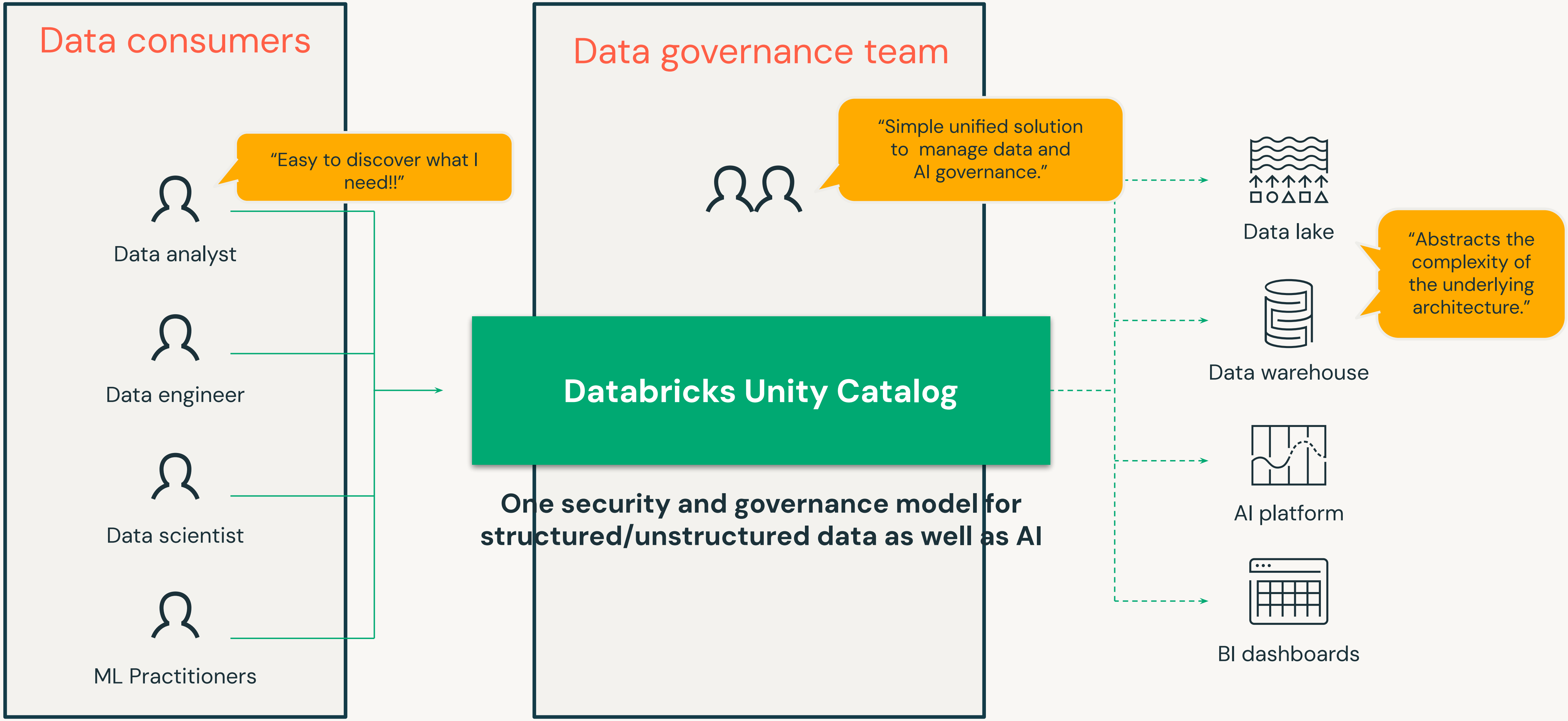
Data governance team

"How to secure these assets?"

"Who is accessing them and how?"

"Are we compliant"

Databricks unifies governance for data and AI



One security and governance model for structured/unstructured data as well as AI

Data and AI governance is growing in importance

> “Organizations are finally realizing the value of data as an asset that needs to be protected, managed and maintained to increase asset value.”
>
> —IDC

> “Organizations seeing the highest returns from AI have a framework for AI governance to cover every step of the model development process.”
>
> — The State of AI in 2022, McKinsey & Co.

> “AI is now an enterprise essential, and as such, AI governance will join cybersecurity and compliance as a board-level topic.”
>
> —Forrester, 2023 AI Predictions report

Key elements of governance

Secure and auditable

Discoverable

Usable and performant

Accurate and high quality

How do we get there?

Centralized access control for data & AI

Built-in data search and discovery

Performance & scalability

Lineage & data quality

How do we get there?

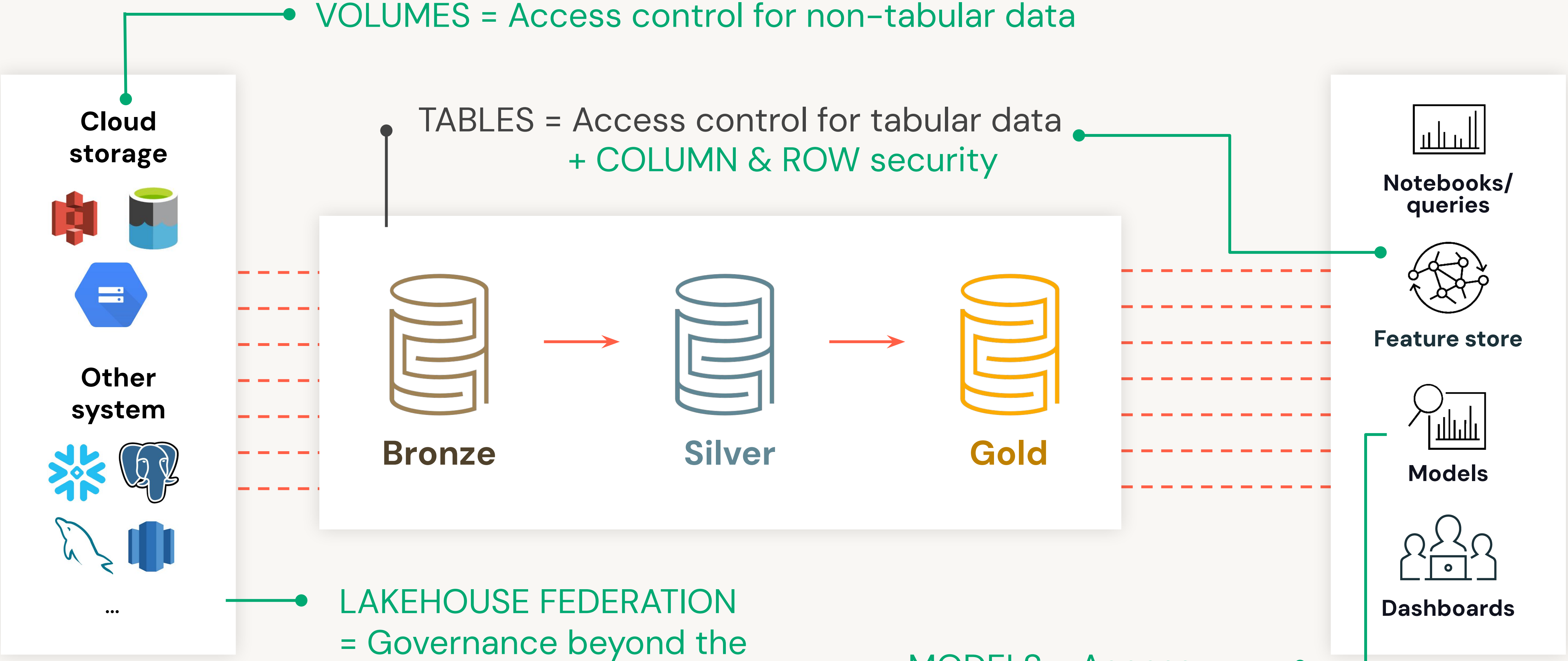
Centralized access control for data & AI

Built-in data search and discovery

Performance & scalability

Lineage & data quality

Access control for data + AI assets



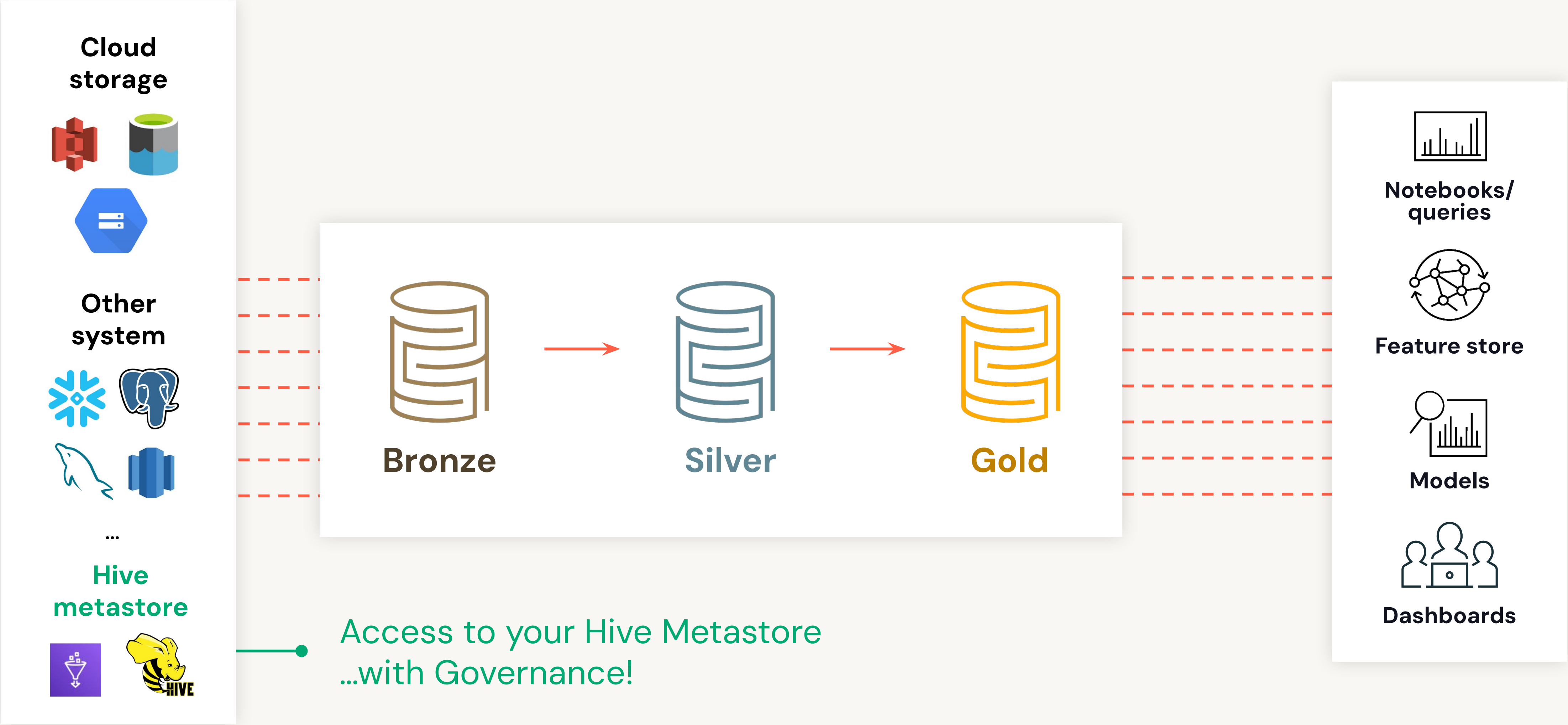
VOLUMES = Access control for non-tabular data

TABLES = Access control for tabular data
+ COLUMN & ROW security

LAKEHOUSE FEDERATION
= Governance beyond the lakehouse!

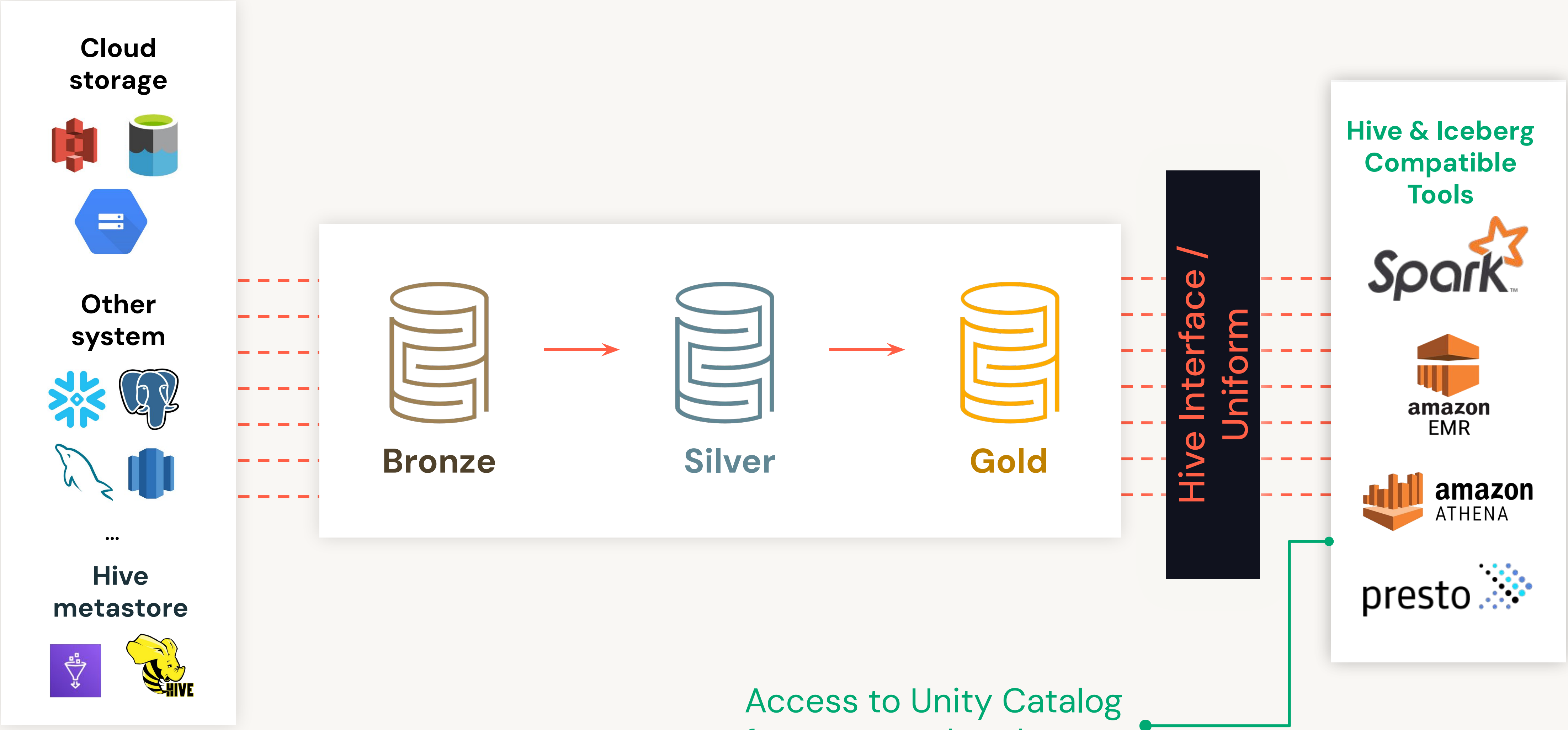
MODELS = Access control for AI

Hive metastore federation Coming Soon!



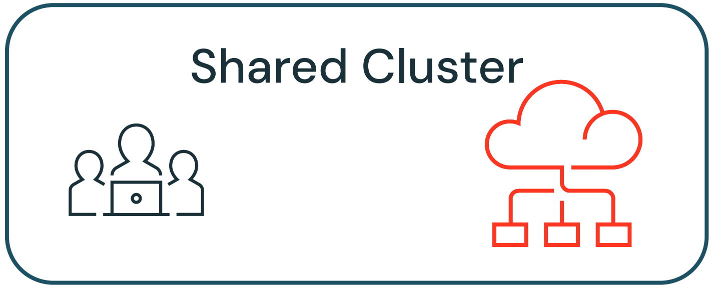
Access to your Hive Metastore
...with Governance!

Hive metastore interface & Uniform Private Preview

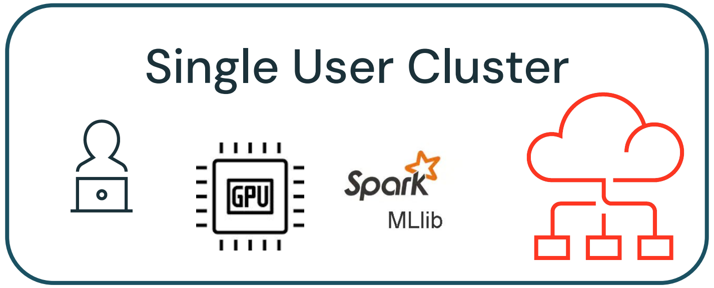


Access to Unity Catalog from external tools

Clusters



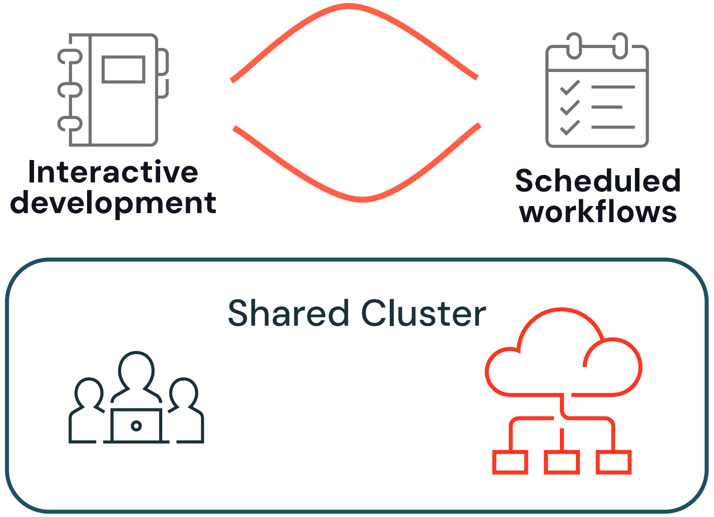
Support concurrent users/workloads with full isolation guarantees, fine grained security



Support advanced use cases such as GPUs, distributed ML, RDDs

UC Shared clusters

Develop and deploy your workloads on shared clusters!



Support concurrent users/workloads with full isolation guarantees

Recently launched (DBR 13+)

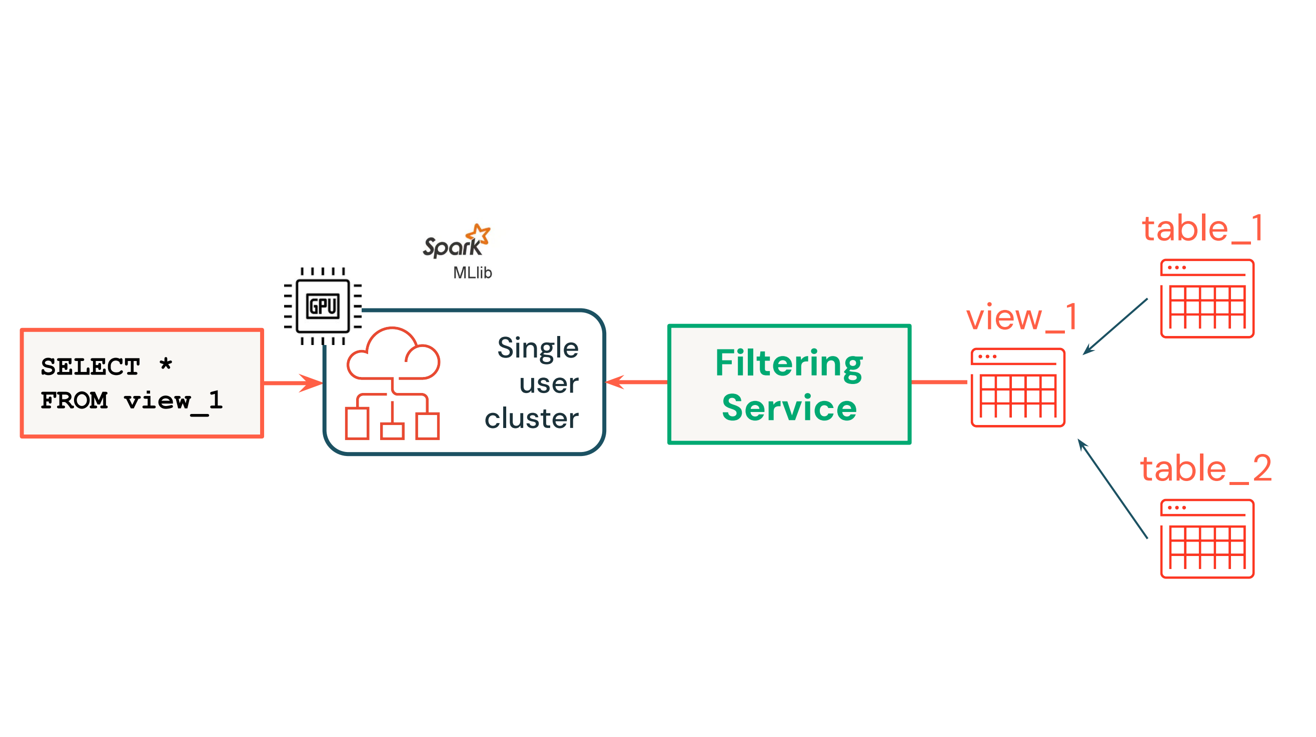
- Python cluster libraries (13.1)
- Scalar Python/Pandas UDFs (13.2)
- Single node ML & MLflow
- Structured Streaming
- Network connectivity from Python

Private previews:

- Scala
- Init scripts / Jars
- MLR (light, single node ML)

Fine-grained access control for single-user clusters

Coming Soon!



Query views and masked tables securely from single-user clusters!

View and masked table access:

- Data filtered via secure, serverless filtering service,
- Filtered results are streamed back to the single user cluster

Demo: Access Control for the Lakehouse

How do we get there?

Centralized access control for data & AI

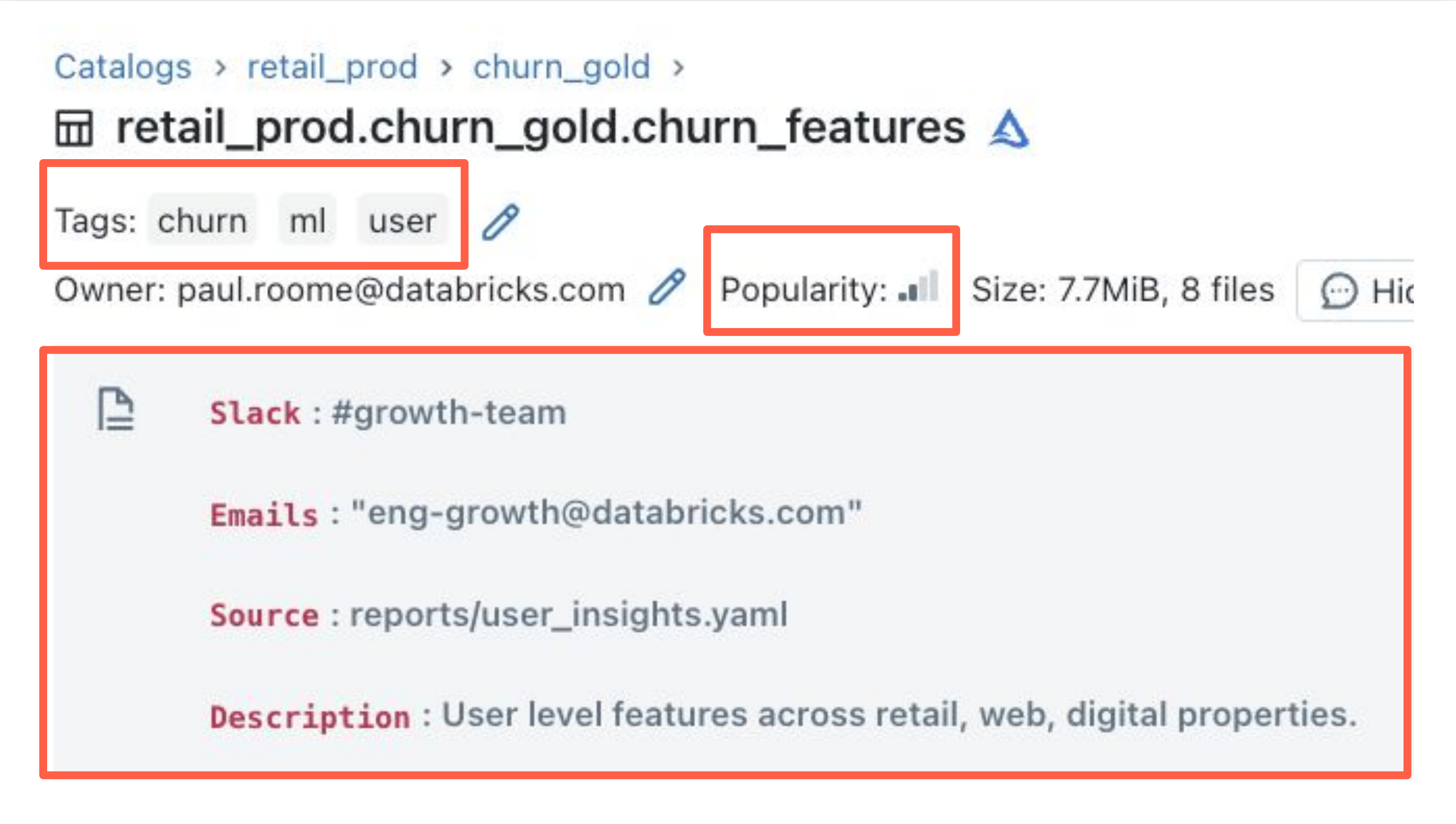
Built-in data search and discovery

Performance & scalability

Lineage & data quality

Describe and tag data

Public Preview



Catalogs › retail_prod › churn_gold ›

retail_prod.churn_gold.churn_features

Tags: churn ml user

Owner: paul.roome@databricks.com Popularity: Size: 7.7MiB, 8 files

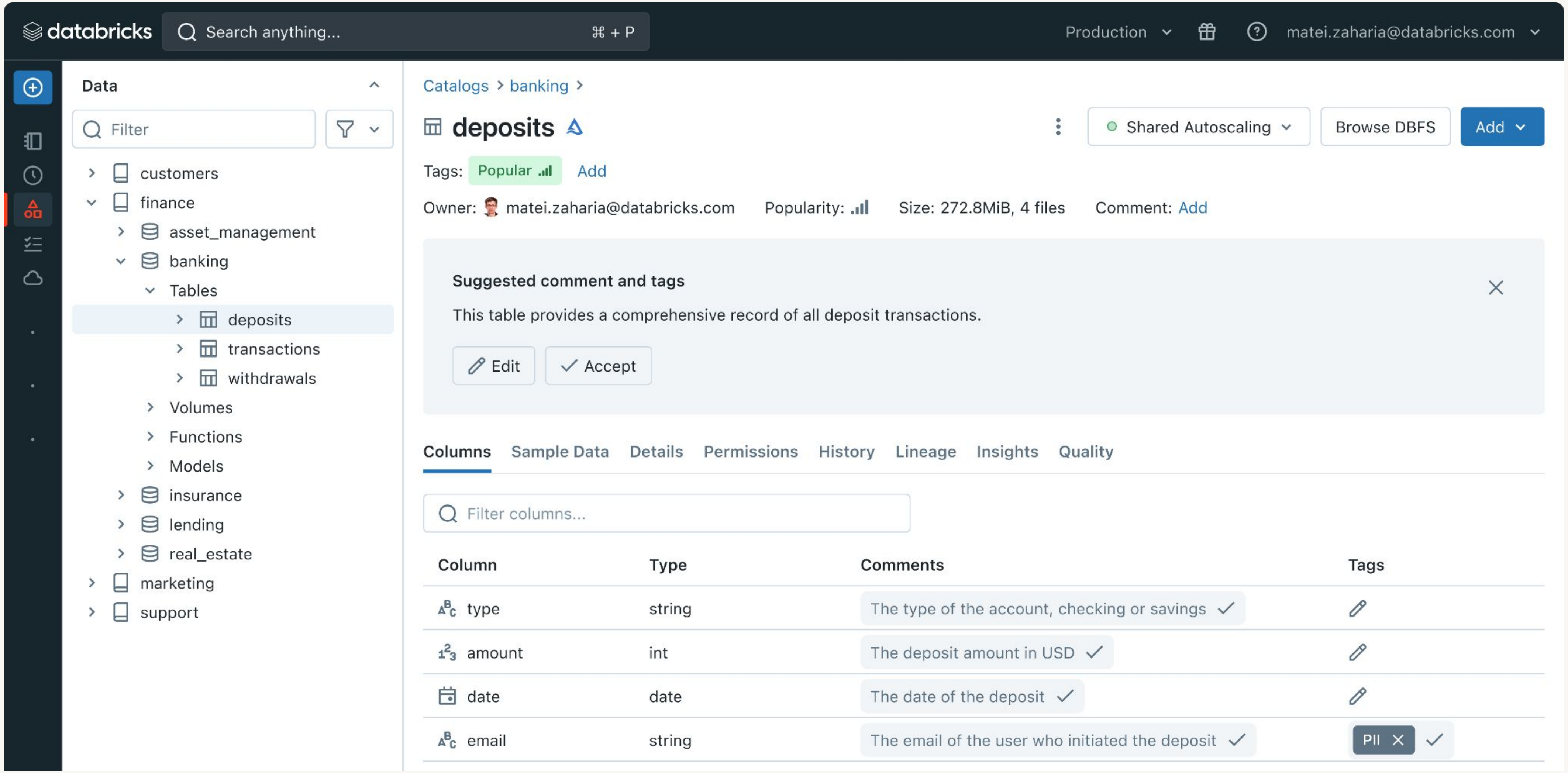
Slack : #growth-team

Emails : "eng-growth@databricks.com"

Source : reports/user_insights.yaml

Description : User level features across retail, web, digital properties.

Next Step : LakehouseIQ



Catalogs > banking >

deposits

Tags: Popular Add

Owner: matei.zaharia@databricks.com Popularity: Size: 272.8MiB, 4 files Comment: Add

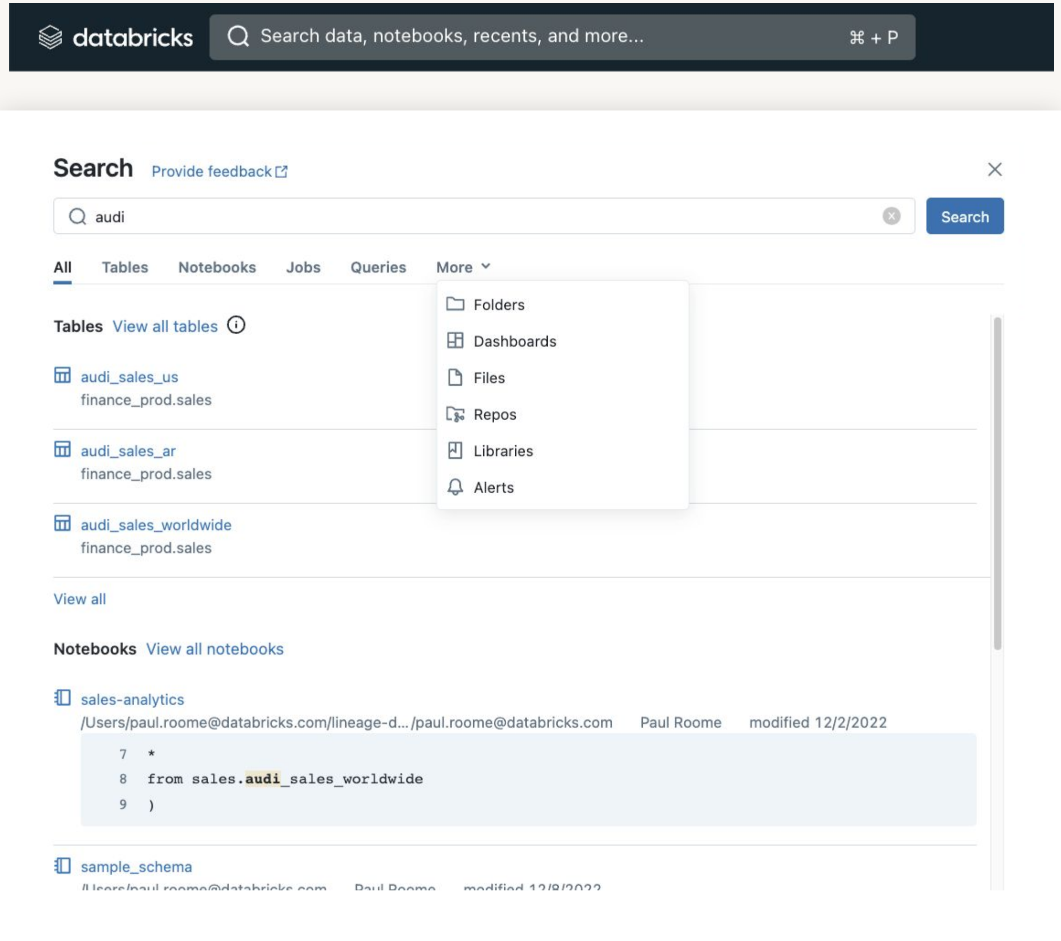
Suggested comment and tags

This table provides a comprehensive record of all deposit transactions.

Edit ✓ Accept

Column	Type	Comments	Tags
type	string	The type of the account, checking or savings ✓	
amount	int	The deposit amount in USD ✓	
date	date	The date of the deposit ✓	
email	string	The email of the user who initiated the deposit ✓	PII ✕ ✓

Find and discover assets



Search Provide feedback

audi

Tables View all tables

audi_sales_us
finance_prod.sales

audi_sales_ar
finance_prod.sales

audi_sales_worldwide
finance_prod.sales

View all

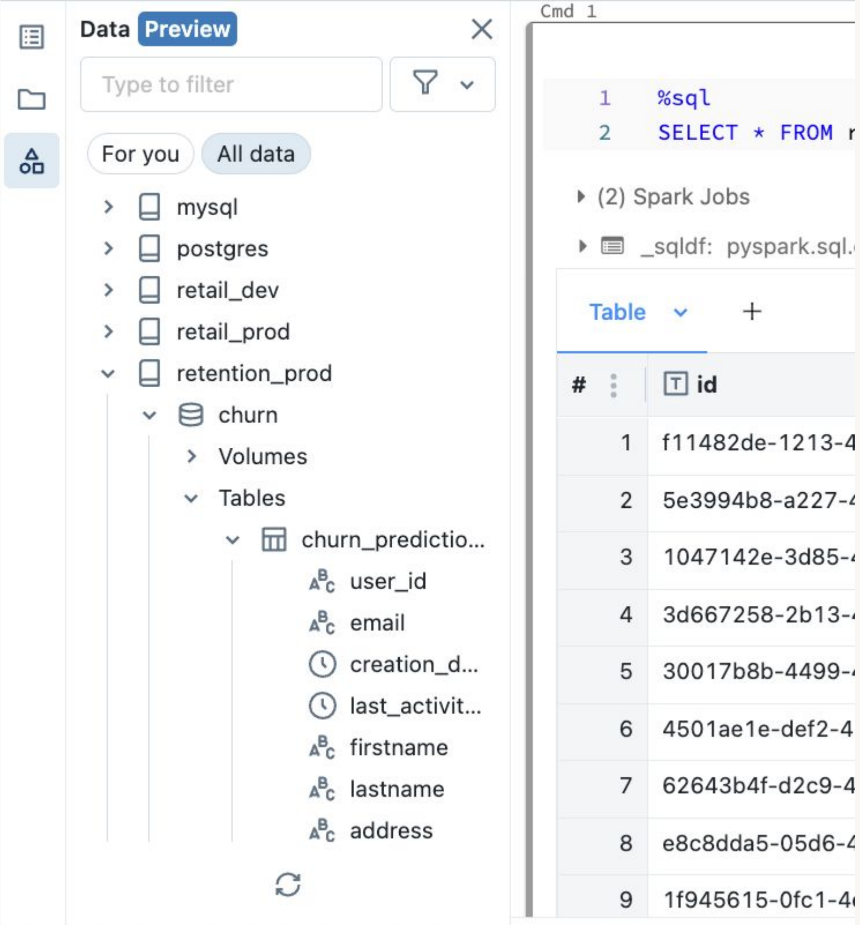
Notebooks View all notebooks

sales-analytics
/Users/paul.roome@databricks.com/lineage-d.../paul.roome@databricks.com Paul Roome modified 12/2/2022

```
*
from sales.audi_sales_worldwide
)
```

sample_schema
/Users/paul.roome@databricks.com Paul Roome modified 12/8/2022

Search for any lakehouse entity!



- mysql
- postgres
- retail_dev
- retail_prod
- retention_prod
  - churn
    - Volumes
    - Tables
      - churn_predictio...
        - user_id
        - email
        - creation_d...
        - last_activit...
        - firstname
        - lastname
        - address

Cmd 1

```
%sql
SELECT * FROM r
```

(2) Spark Jobs

_sqldf: pyspark.sql.

#	id
1	f11482de-1213-4
2	5e3994b8-a227-4
3	1047142e-3d85-4
4	3d667258-2b13-4
5	30017b8b-4499-4
6	4501ae1e-def2-4
7	62643b4f-d2c9-4
8	e8c8dda5-05d6-4
9	1f945615-0fc1-4

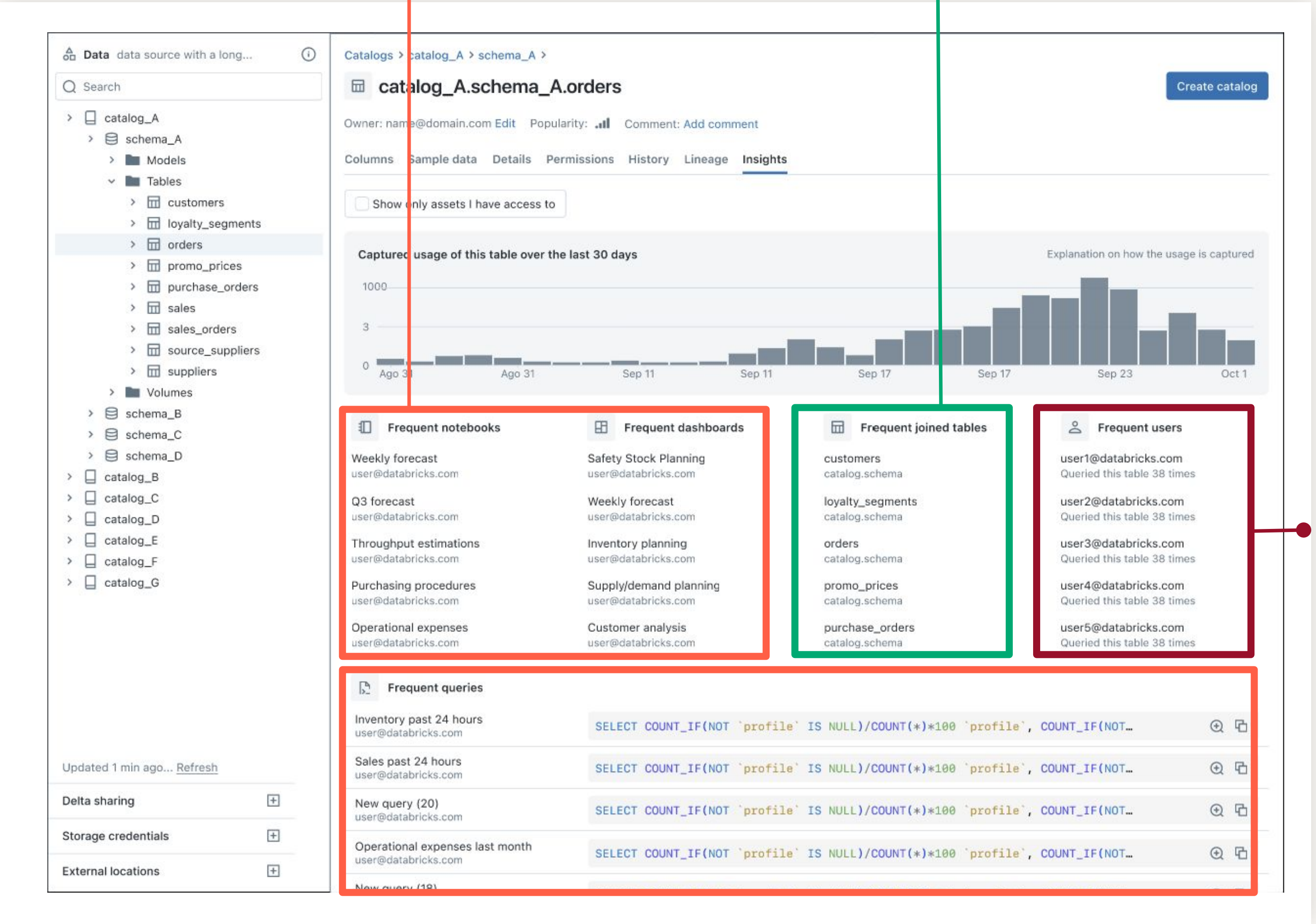
Data browser in the notebook!

Data insights

Frequent notebooks/queries/dashboards:
How do I get started?

Frequent Joins:
How can I enrich?

Frequent users:
Who are the experts?



Data data source with a long...

Search

- catalog_A
  - schema_A
    - Models
    - Tables
      - customers
      - loyalty_segments
      - orders
      - promo_prices
      - purchase_orders
      - sales
      - sales_orders
      - source_suppliers
      - suppliers
    - Volumes
  - schema_B
  - schema_C
  - schema_D
- catalog_B
- catalog_C
- catalog_D
- catalog_E
- catalog_F
- catalog_G

Updated 1 min ago... Refresh

Delta sharing ⊞

Storage credentials ⊞

External locations ⊞

Catalogs > catalog_A > schema_A >

catalog_A.schema_A.orders

Create catalog

Owner: name@domain.com Edit   Popularity: Comment: Add comment

Columns   Sample data   Details   Permissions   History   Lineage   **Insights**

□ Show only assets I have access to

Captured usage of this table over the last 30 days

Explanation on how the usage is captured

Frequent notebooks

- Weekly forecast
  user@databricks.com
- Q3 forecast
  user@databricks.com
- Throughput estimations
  user@databricks.com
- Purchasing procedures
  user@databricks.com
- Operational expenses
  user@databricks.com

Frequent dashboards

- Safety Stock Planning
  user@databricks.com
- Weekly forecast
  user@databricks.com
- Inventory planning
  user@databricks.com
- Supply/demand planning
  user@databricks.com
- Customer analysis
  user@databricks.com

Frequent joined tables

- customers
  catalog.schema
- loyalty_segments
  catalog.schema
- orders
  catalog.schema
- promo_prices
  catalog.schema
- purchase_orders
  catalog.schema

Frequent users

- user1@databricks.com
  Queried this table 38 times
- user2@databricks.com
  Queried this table 38 times
- user3@databricks.com
  Queried this table 38 times
- user4@databricks.com
  Queried this table 38 times
- user5@databricks.com
  Queried this table 38 times

Frequent queries

Inventory past 24 hours user@databricks.com	`SELECT COUNT_IF(NOT `profile` IS NULL)/COUNT(*)*100 `profile`, COUNT_IF(NOT…`
Sales past 24 hours user@databricks.com	`SELECT COUNT_IF(NOT `profile` IS NULL)/COUNT(*)*100 `profile`, COUNT_IF(NOT…`
New query (20) user@databricks.com	`SELECT COUNT_IF(NOT `profile` IS NULL)/COUNT(*)*100 `profile`, COUNT_IF(NOT…`
Operational expenses last month user@databricks.com	`SELECT COUNT_IF(NOT `profile` IS NULL)/COUNT(*)*100 `profile`, COUNT_IF(NOT…`
New query (18)	

How do we get there?

Centralized access control for data & AI

Built-in data search and discovery

Performance & scalability

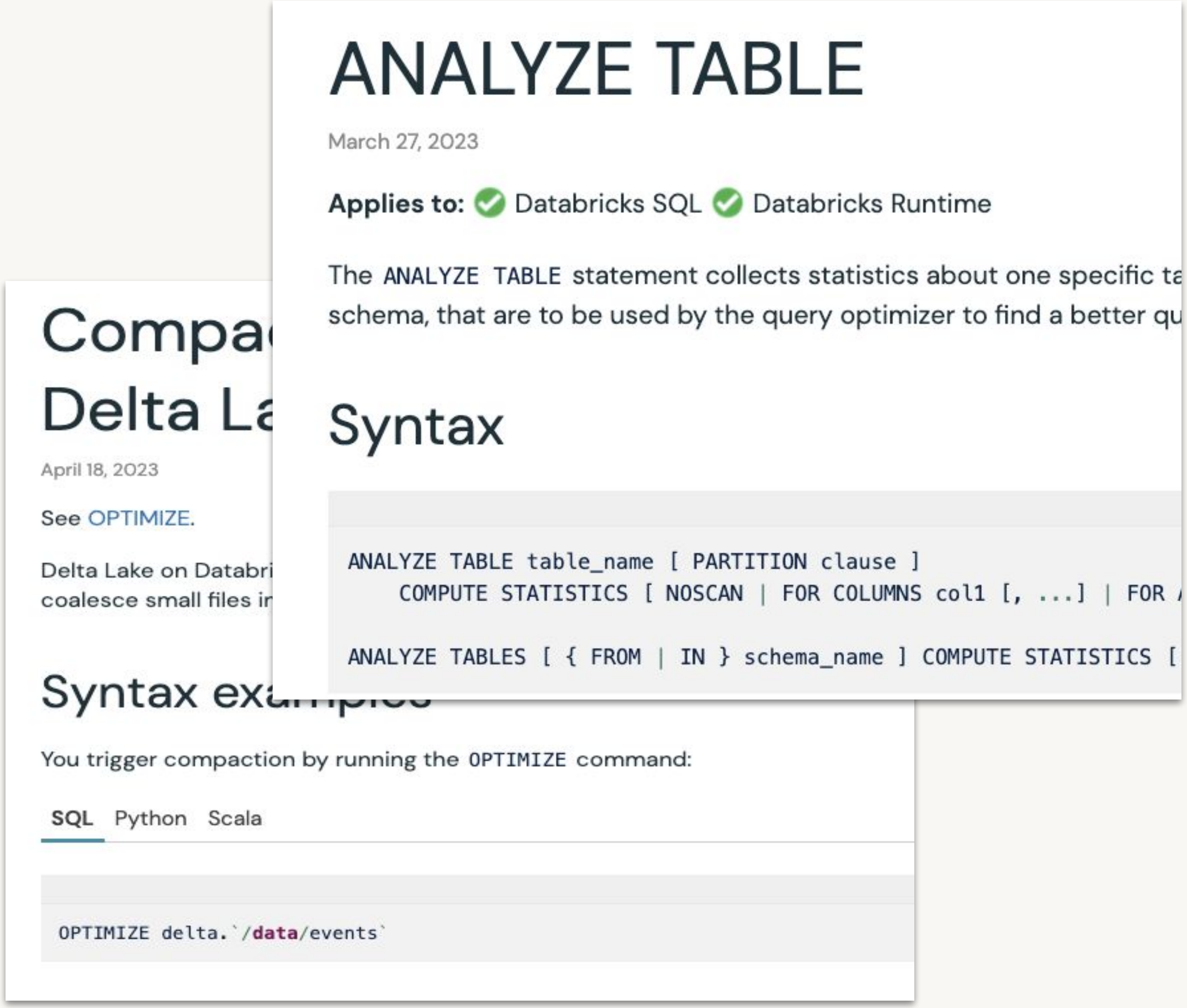
Lineage & data quality

Predictive optimization

Challenge

Tables can can be optimized for faster price-performance, but…

- Which tables?
- How often?
- What optimizations?



Compa
Delta La

April 18, 2023

See OPTIMIZE.

Delta Lake on Databri
coalesce small files ir

Syntax exa

You trigger compaction by running the `OPTIMIZE` command:

SQL Python Scala

```
OPTIMIZE delta.`/data/events`
```

ANALYZE TABLE

March 27, 2023

Applies to: ✓ Databricks SQL ✓ Databricks Runtime

The `ANALYZE TABLE` statement collects statistics about one specific ta
schema, that are to be used by the query optimizer to find a better qu

Syntax

```
ANALYZE TABLE table_name [ PARTITION clause ]
    COMPUTE STATISTICS [ NOSCAN | FOR COLUMNS col1 [, ...] | FOR A

ANALYZE TABLES [ { FROM | IN } schema_name ] COMPUTE STATISTICS [
```

Predictive Optimization

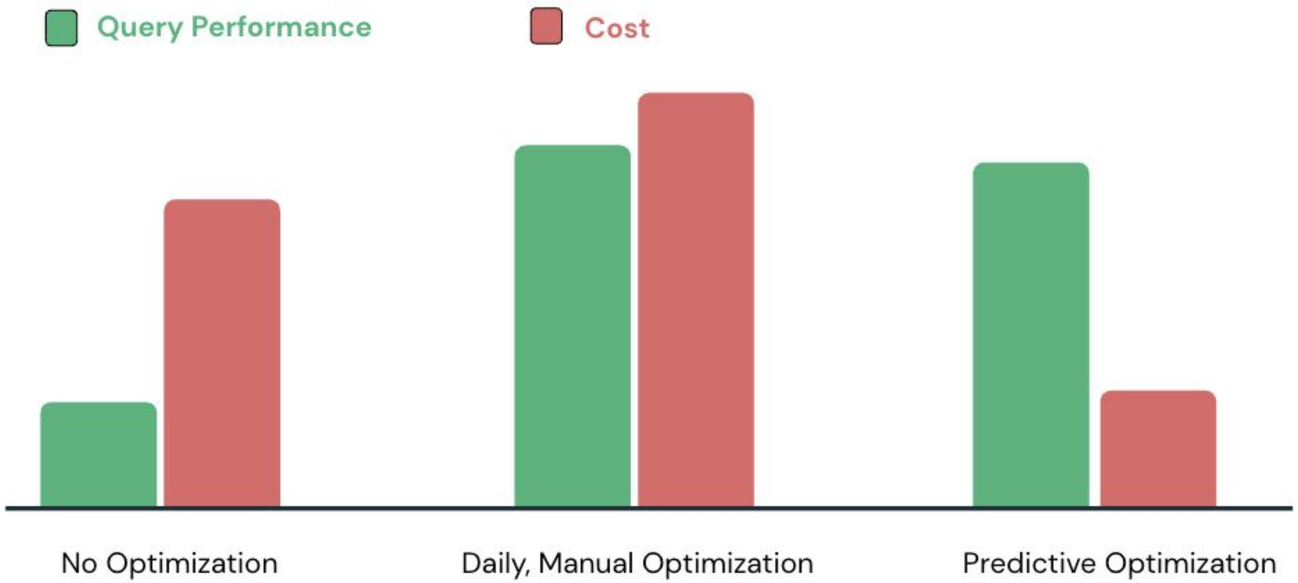
Public Preview

Solution

- AI model determines which tables to optimize
  - Bin-packing
  - Liquid clustering
  - Vacuuming
  - Table statistics
  - Intelligent storage tiering
- Databricks automatically performs maintenance
  - Using hyper-optimized serverless compute

Predictive optimization Public Preview

Results:



Fewer knobs

No need to:

- Determine which tables to optimize
- Run and manage the clusters

How do we get there?

Centralized access control for data & AI

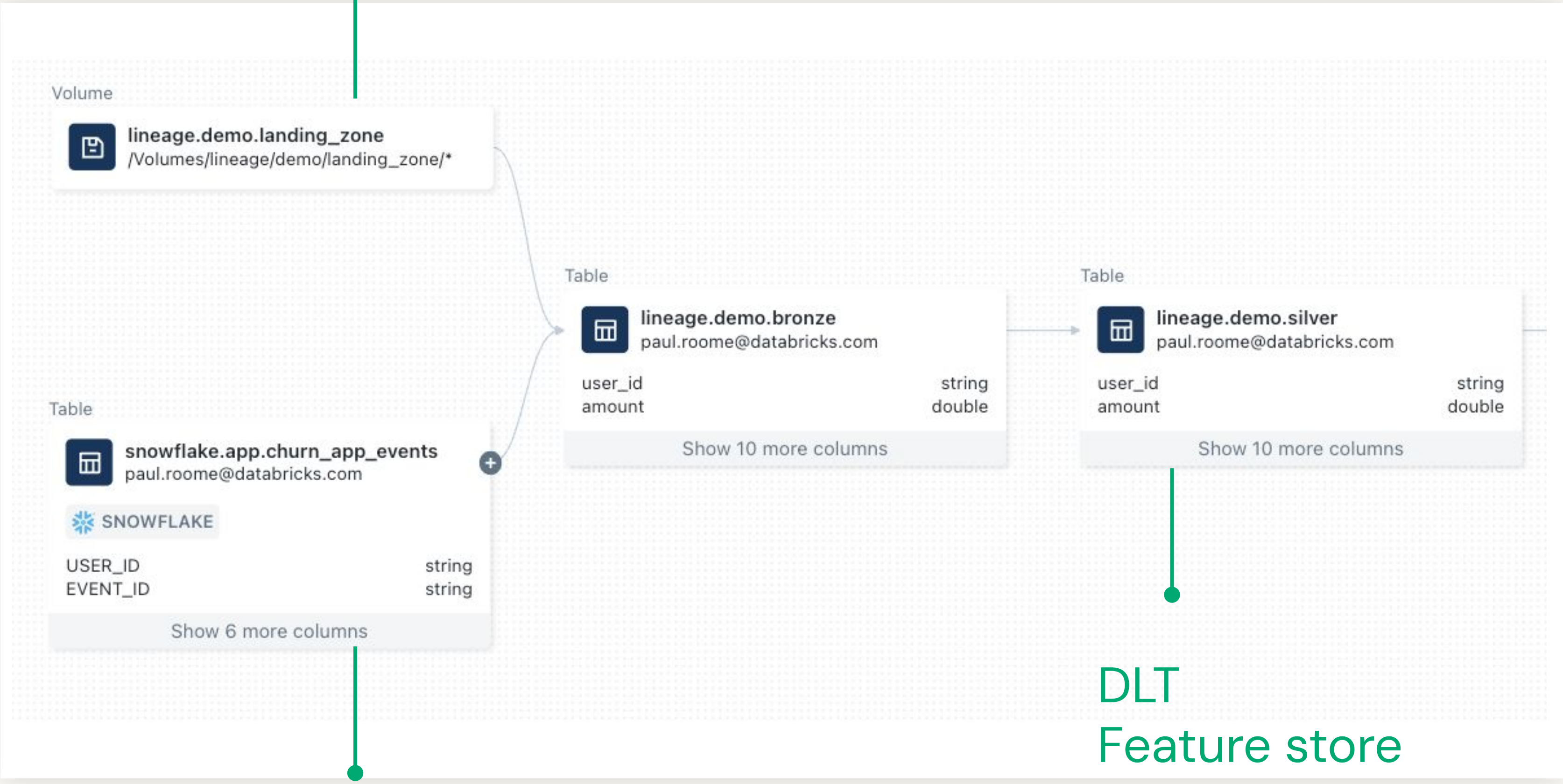
Built-in data search and discovery

Performance & scalability

Lineage & data quality

Lineage for the Lakehouse

Files



Volume

lineage.demo.landing_zone
/Volumes/lineage/demo/landing_zone/*

Table

lineage.demo.bronze
paul.roome@databricks.com

user_id	string
amount	double

Show 10 more columns

Table

lineage.demo.silver
paul.roome@databricks.com

user_id	string
amount	double

Show 10 more columns

Table

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

SNOWFLAKE

USER_ID	string
EVENT_ID	string

Show 6 more columns

Federated sources

DLT
Feature store
Models
Democratized access (BROWSE)
Improved retention logic

Automatic Data Documentation

Impact analysis

Root cause analysis

Lakehouse monitoring

Public Preview

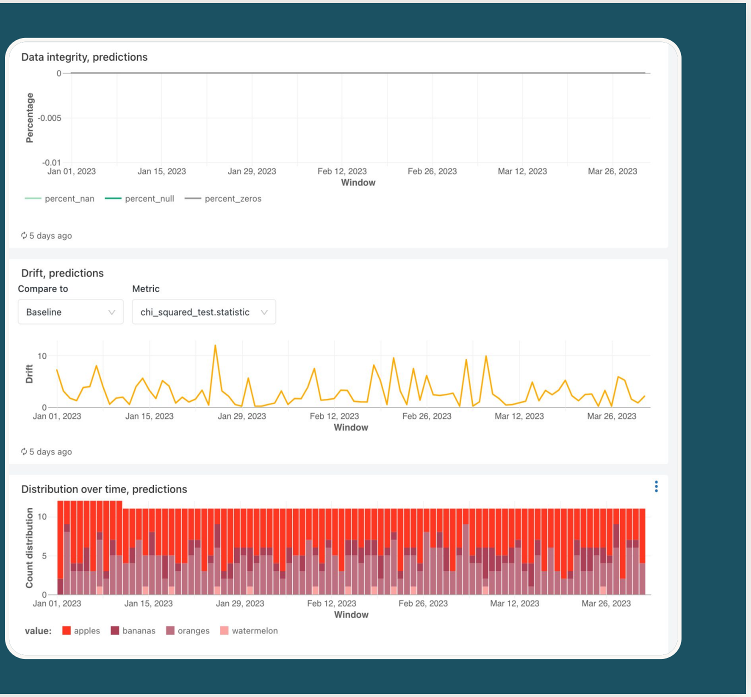
Data quality alerting

Auto classification of sensitive data

Model performance and drift monitoring

Lakehouse monitoring

The first AI-powered monitoring service for both Data and ML

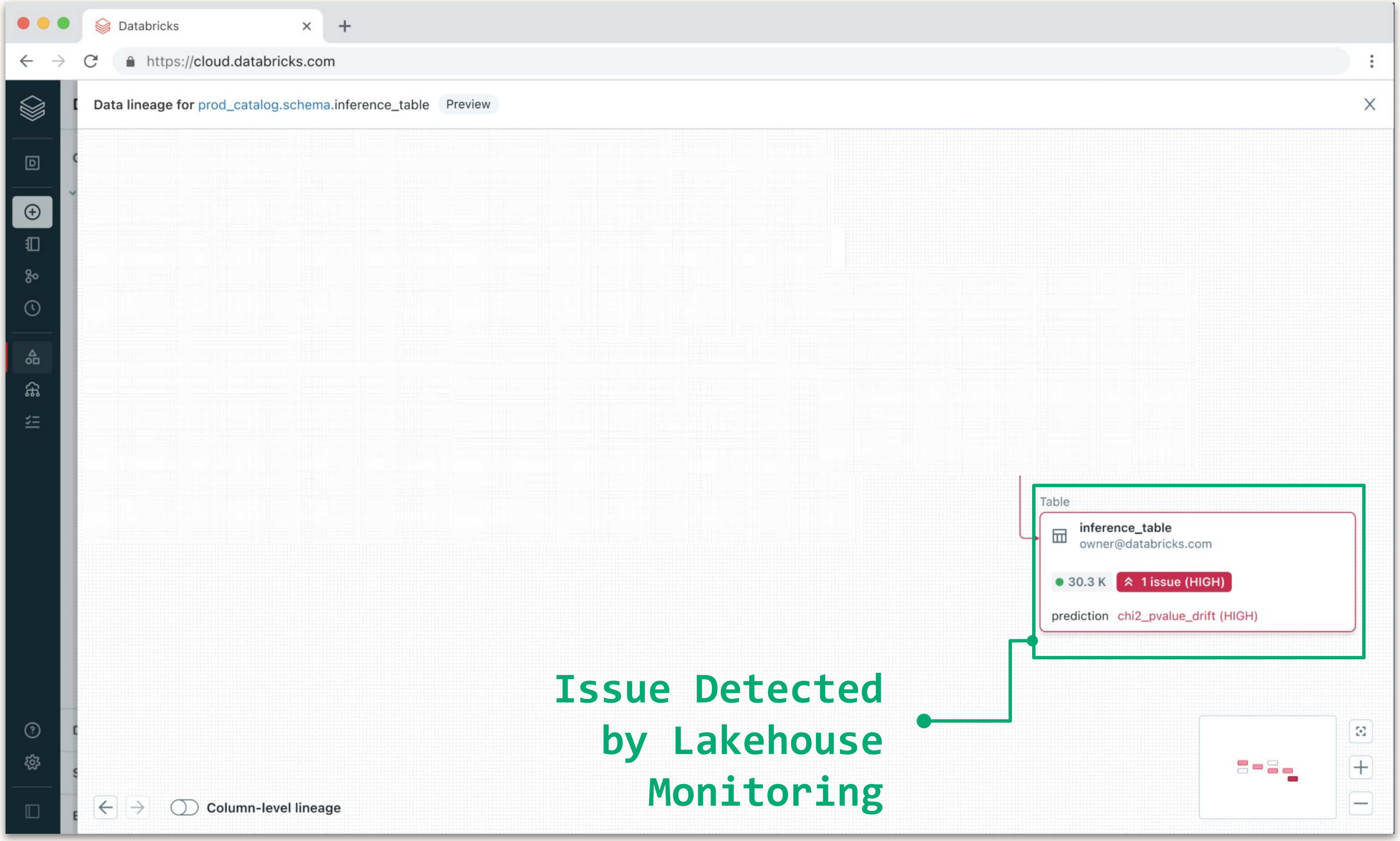


Compare to: Baseline

Metric: chi_squared_test.statistic

Lineage + Lakehouse Monitoring = 💖

Root cause analysis



Data lineage for prod_catalog.schema.inference_table Preview

Table

inference_table
owner@databricks.com

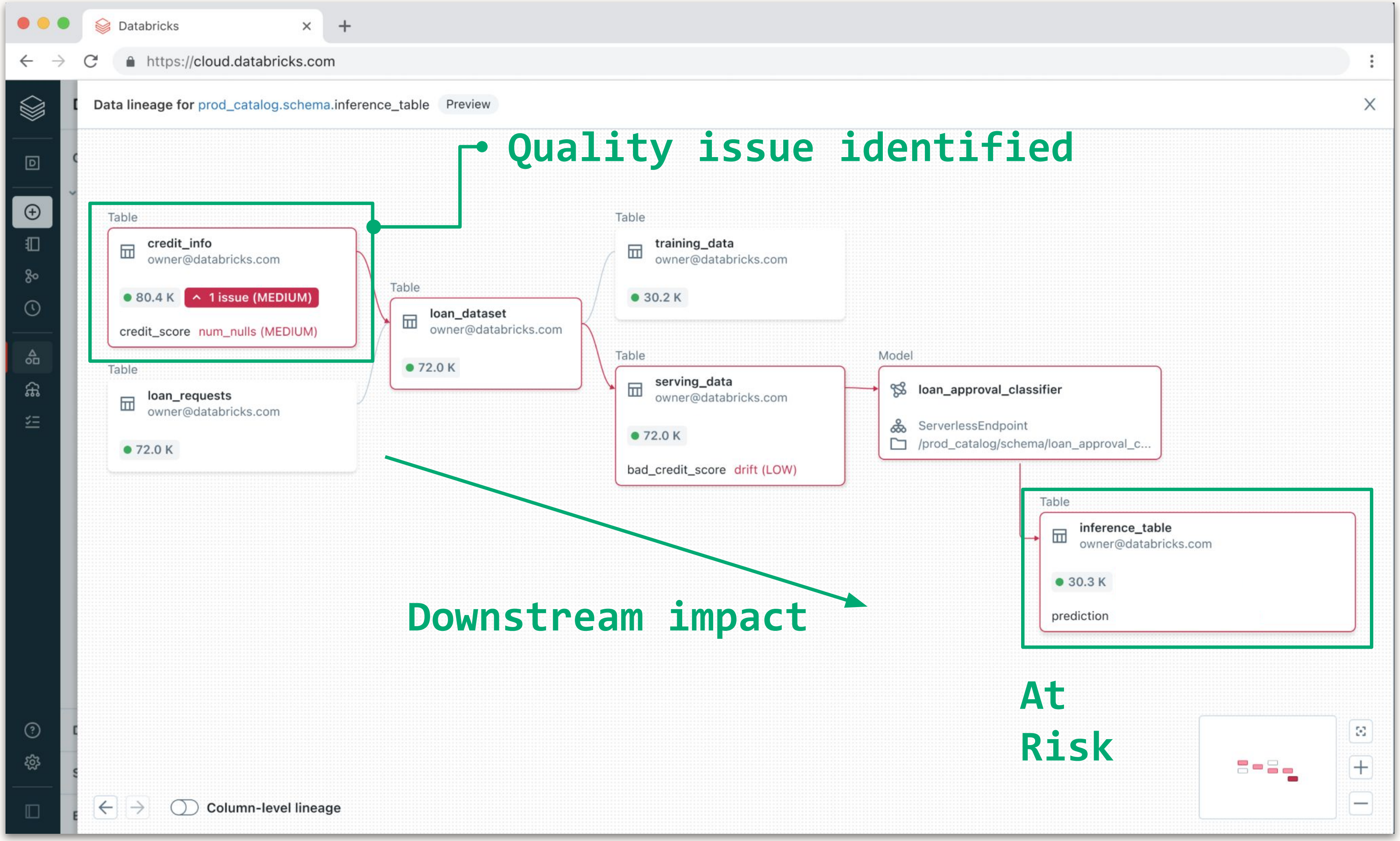
● 30.3 K 1 issue (HIGH)

prediction chi2_pvalue_drift (HIGH)

Issue Detected by Lakehouse Monitoring

Lineage + Lakehouse Monitoring = 💖

Proactive alerting



Quality issue identified

Downstream impact

At Risk

Demo: Lineage for the Lakehouse

What we've seen…

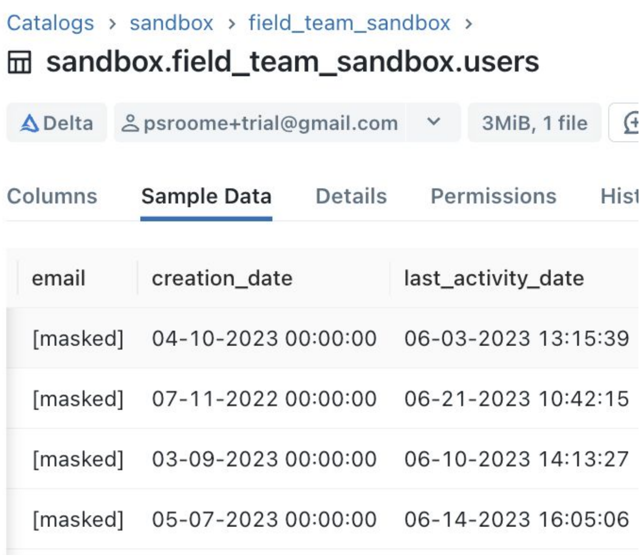
Secure and auditable

Discoverable

Usable and performant

Accurate and high quality

Even more on the horizon…

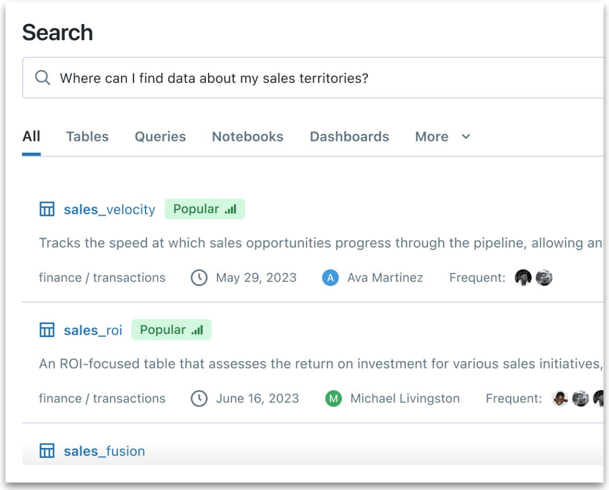


Catalogs › sandbox › field_team_sandbox ›

sandbox.field_team_sandbox.users

email	creation_date	last_activity_date
[masked]	04-10-2023 00:00:00	06-03-2023 13:15:39
[masked]	07-11-2022 00:00:00	06-21-2023 10:42:15
[masked]	03-09-2023 00:00:00	06-10-2023 14:13:27
[masked]	05-07-2023 00:00:00	06-14-2023 16:05:06

Auto Tag + Mask



Search

Where can I find data about my sales territories?

sales_velocity Popular

Tracks the speed at which sales opportunities progress through the pipeline, allowing an

finance / transactions May 29, 2023 Ava Martinez Frequent:

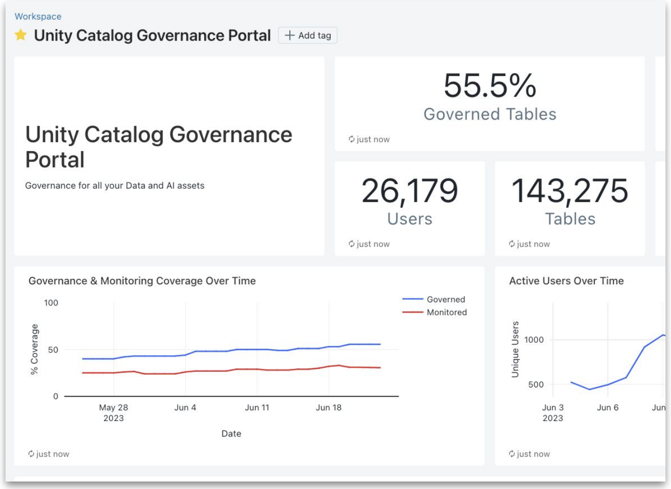
sales_roi Popular

An ROI-focused table that assesses the return on investment for various sales initiatives,

finance / transactions June 16, 2023 Michael Livingston Frequent:

sales_fusion

LakehouseIQ Augmented Search + Discovery



Workspace

Unity Catalog Governance Portal

Unity Catalog Governance Portal

Governance for all your Data and AI assets

55.5% Governed Tables

26,179 Users

143,275 Tables

Governance Portal

databricks