DATA+AI
SUMMIT 2022

How to Build a Complete Security and Governance Solution Using Unity Catalog

Don Bosco Durai
PMC - Apache Ranger
Co founder & CTO, Privacera

Zeashan Pappa
Senior Product Specialist - Data Governance

ORGANIZED BY databricks

Abstract

Unity Catalog unifies governance and security for Databricks in one place. It can store data classifications and privileges and enforce them.

This talk will go into the details of Unity Catalog and explains the core building blocks in Unity Catalog for Security and Governance. I will also explain how Privacera translates Apache Ranger policies into native policies of Unity Catalogs, audits are collected from Unity Catalog and imported into the centralized Audit Store of Apache Ranger, and Privacera can extend Unity Catalog.

Agenda

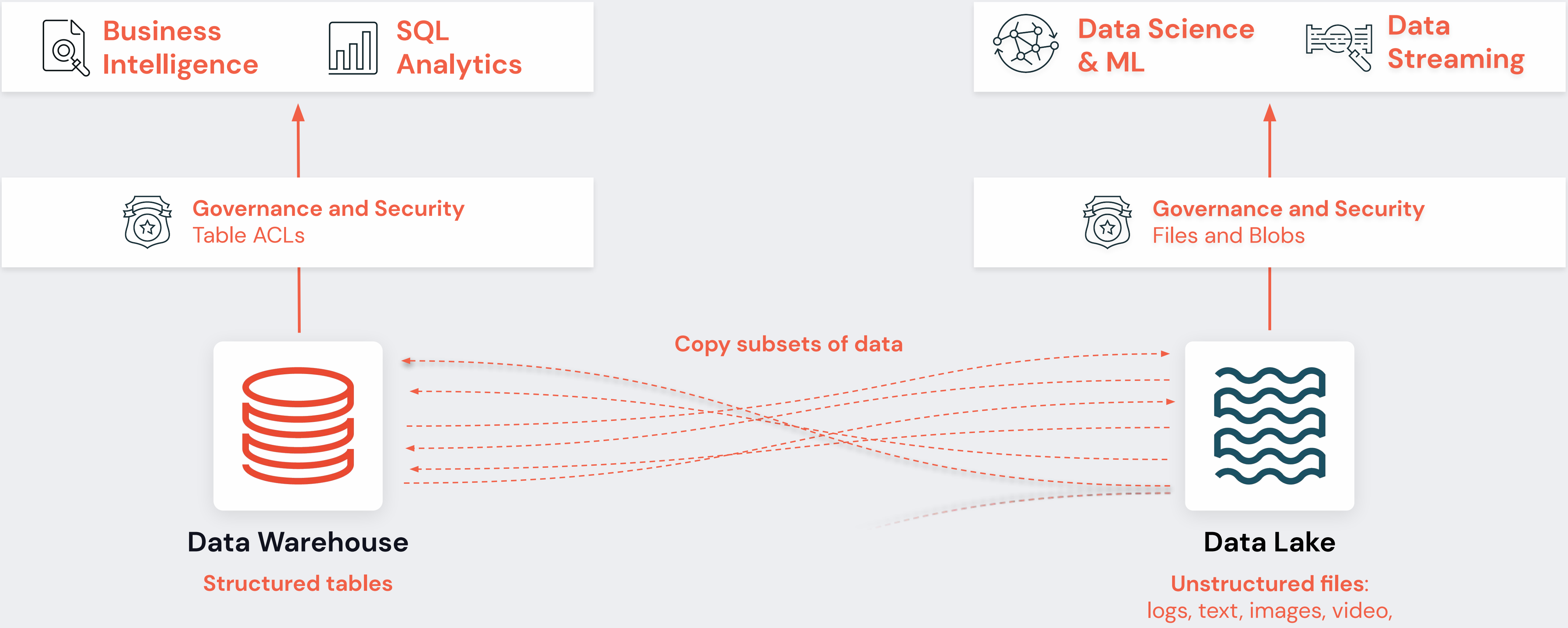
- Unity Catalog Overview
    - Foundational capabilities for Enterprise-wide Governance
- Security and Governance for Enterprise
- Best Practices for implementing Access Governance
- Demo

Product Safe Harbor Statement

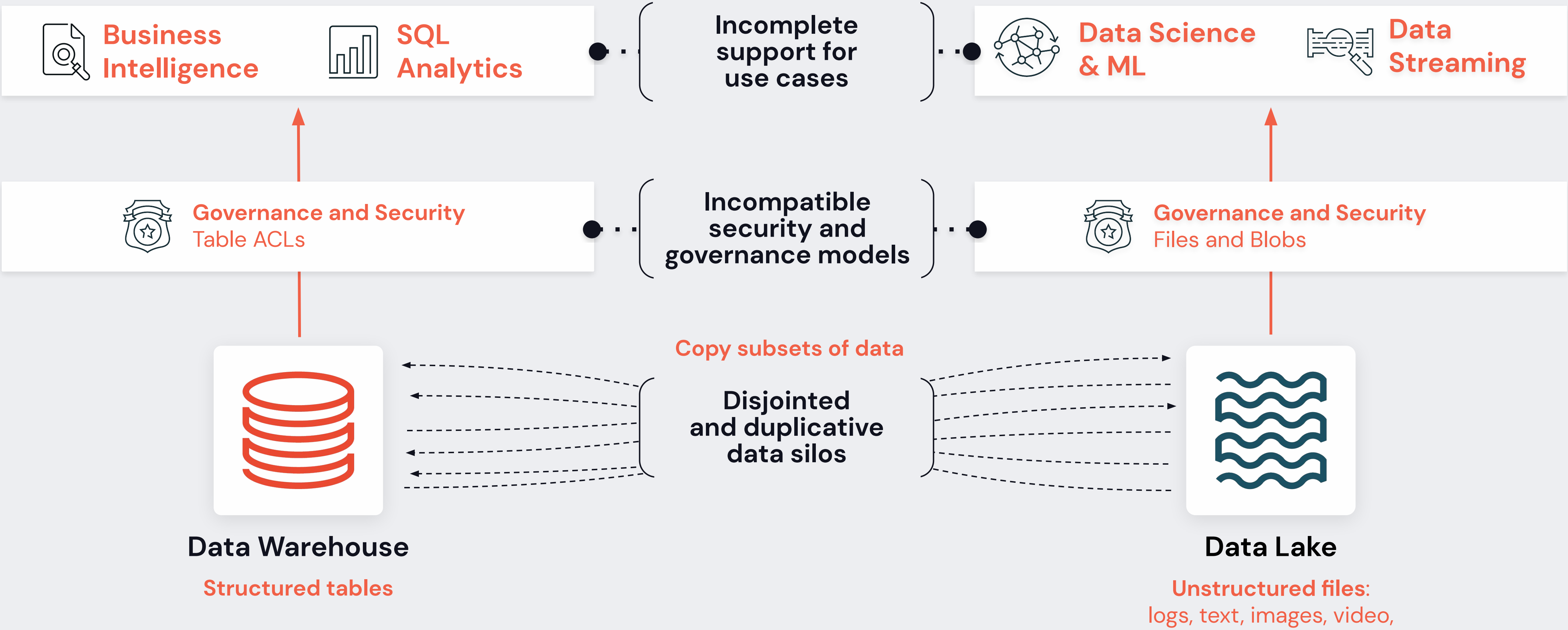
This information is provided to outline Databricks' and Privacera's general product direction and is for informational purposes only. Customers who purchase Databricks or Privacera services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks and Privacera's discretion and may not be delivered as planned or at all.

Most enterprises today struggle with data and AI Governance

Two disparate, incompatible data platforms



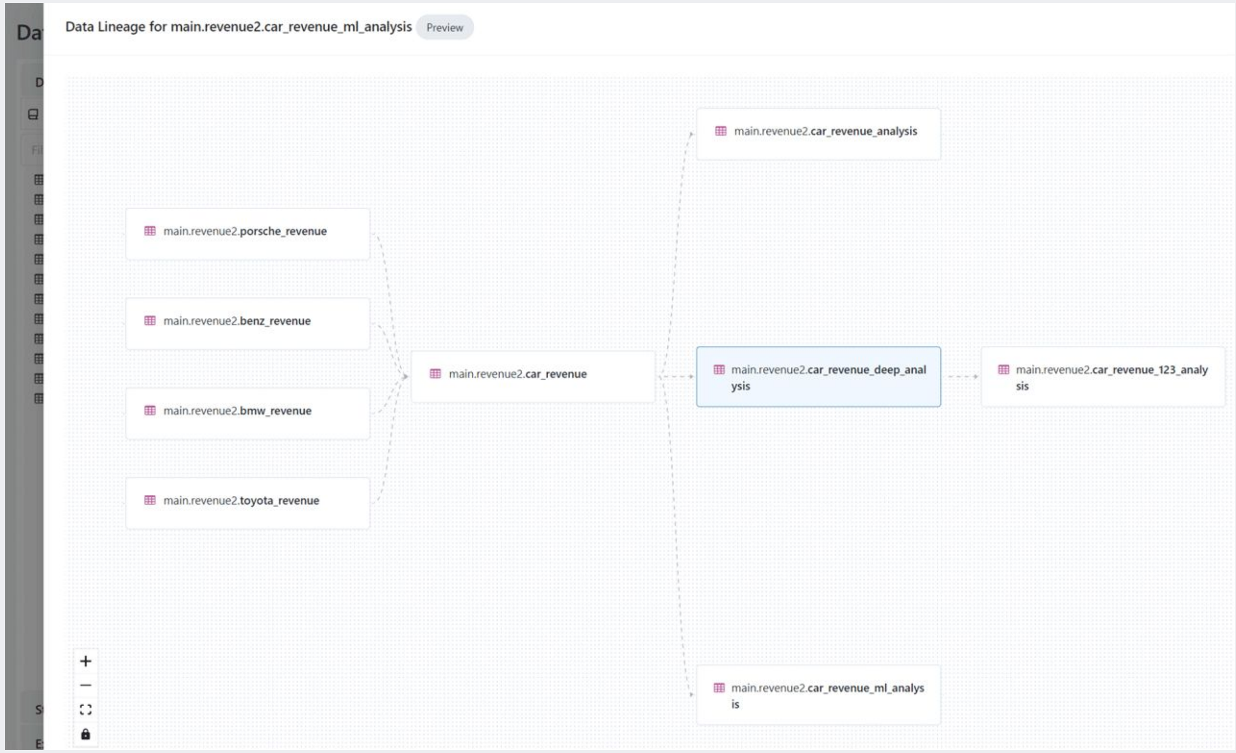
Two disparate, incompatible data platforms



Built-in Data Lineage

End-to-end visibility into how data flows and consumed in your organization

- Auto-capture runtime data lineage on a Databricks cluster or SQL endpoint
- Track lineage down to the table and column level
- Leverage common permission model from Unity Catalog



Data Lineage for main.revenue2.car_revenue_ml_analysis

Preview
main.revenue2.car_revenue_analysis
main.revenue2.porsche_revenue
main.revenue2.benz_revenue
main.revenue2.car_revenue
main.revenue2.car_revenue_deep_analysis
main.revenue2.car_revenue_123_analysis
main.revenue2.bmw_revenue
main.revenue2.toyota_revenue
main.revenue2.car_revenue_ml_analysis
+
−

Governance is also hard to enforce on data lakes

- Lack of fine-grained access controls
- Time consuming data discovery
- Hard to audit
- Non-standard cloud-specific governance model
- No governance support for other asset types – ML models, dashboards

Data Lake

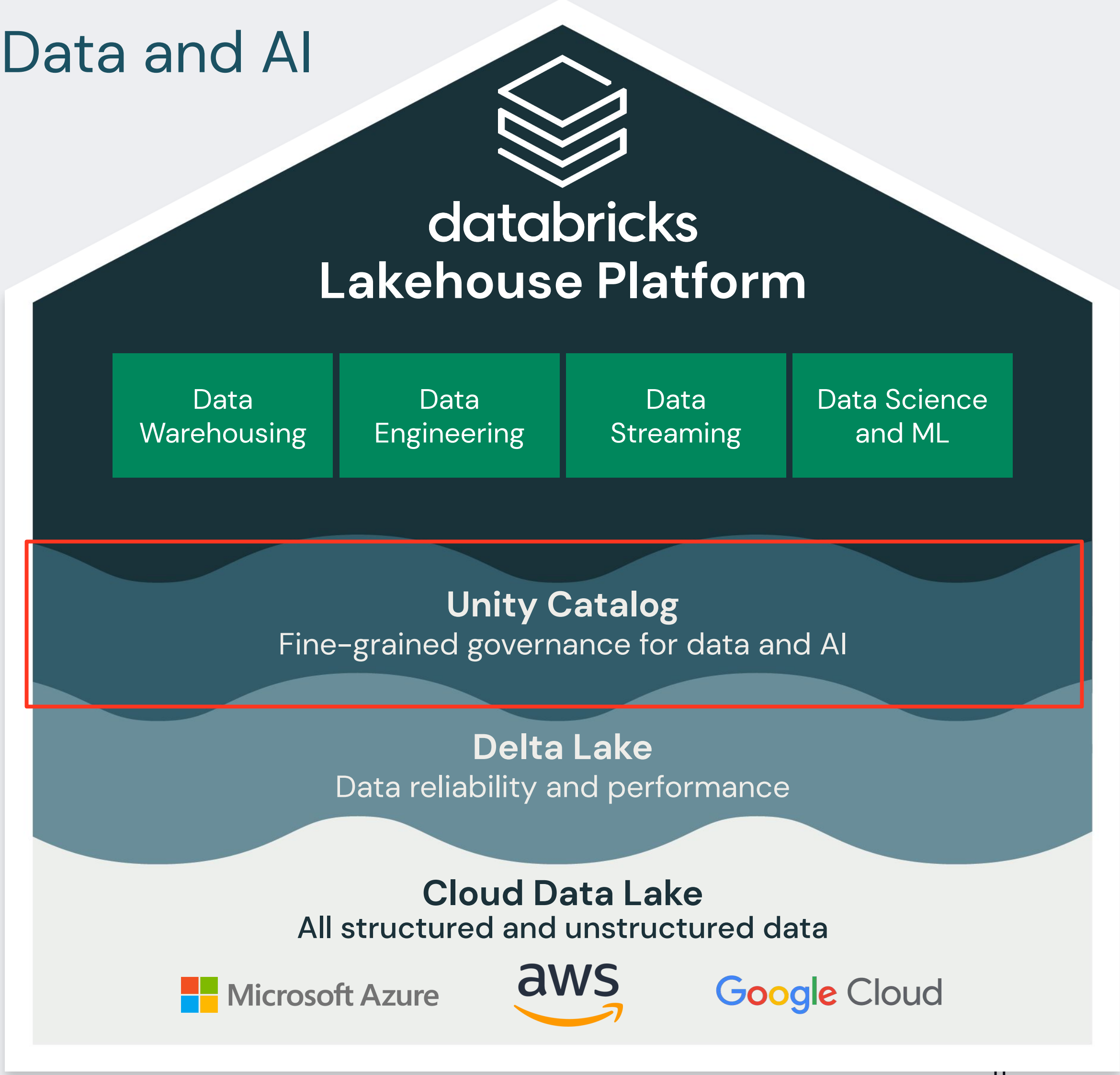
Unstructured files:
logs, text, images, video,

How Databricks helps

Unity Catalog

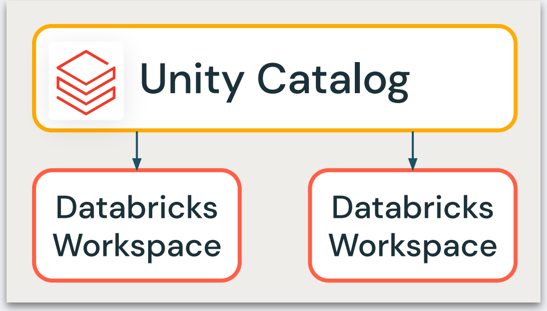
Provide the Building Blocks for Governance of Data and AI Assets on the Lakehouse Architecture

- Provides foundational, unified metadata and access control for data, analytics and AI across clouds within the Databricks Lakehouse
- Help to meet compliance requirements with Databricks by providing audit functionality
- Provides performant enforcement of access control across all languages and compute types to increase productivity.
- Seamlessly integrates across partner and customer ecosystems.



Unity Catalog – Foundational Capabilities

- Centralized metadata and user management Preview
- Centralized data access controls Preview
- Data lineage Private Preview
- Data access auditing Preview
- Data search and discovery Private Preview
- Secure data sharing with Delta Sharing Preview



Unity Catalog

Databricks Workspace

Databricks Workspace

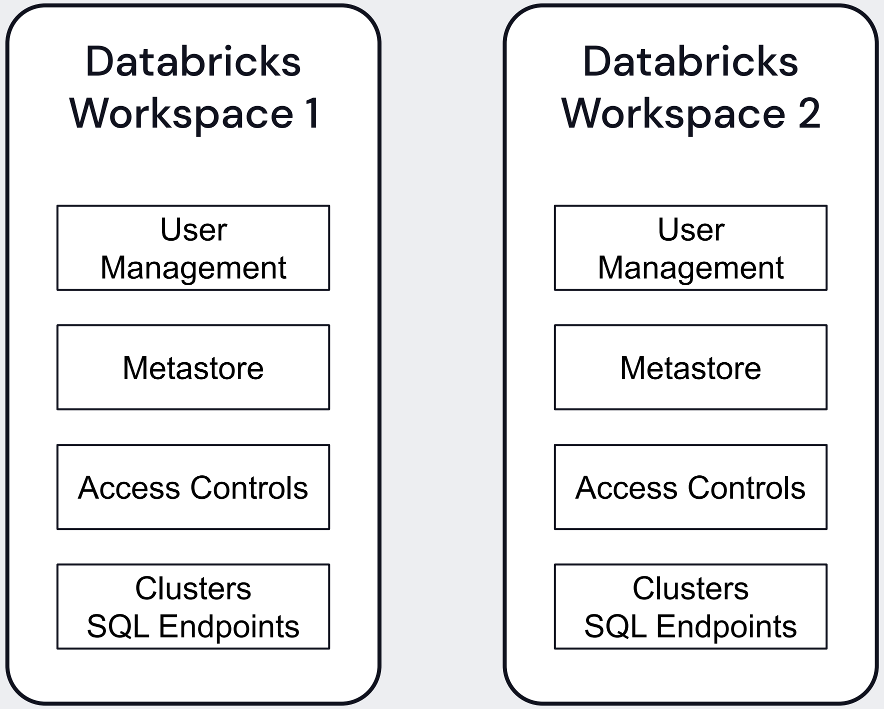
```
 GRANT ... ON ... TO ...
REVOKE ... ON ... FROM ...
```

Catalogs, Databases (schemas), Tables, Views, Storage credentials, External locations

Centralized Metadata and User Management

Create a unified view of your data estate

Without Unity Catalog



Databricks Workspace 1

User Management

Metastore

Access Controls

Clusters
SQL Endpoints

Databricks Workspace 2

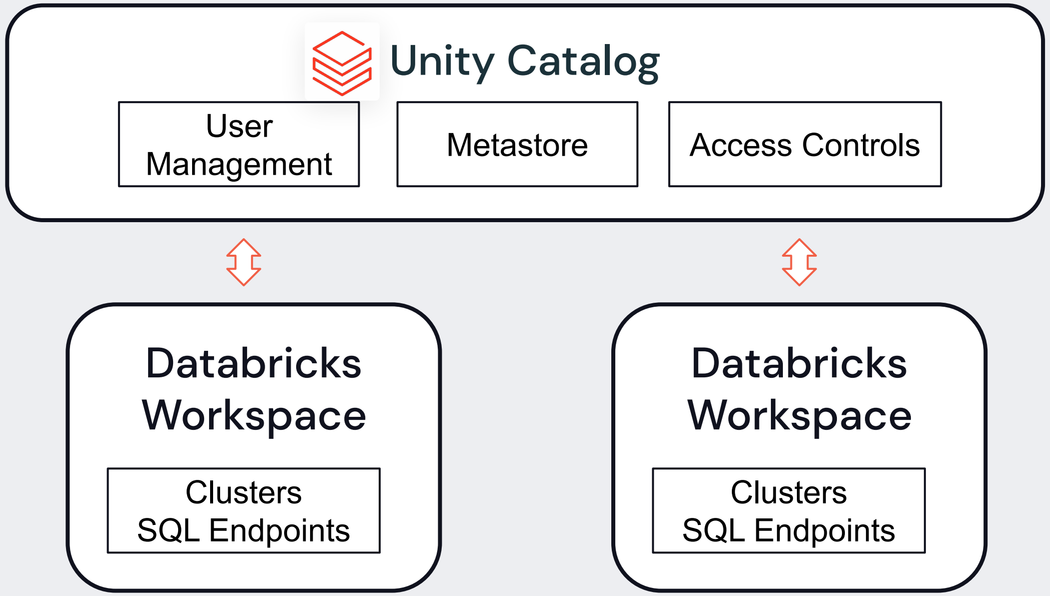
User Management

Metastore

Access Controls

Clusters
SQL Endpoints

With Unity Catalog



Unity Catalog

User Management

Metastore

Access Controls

Databricks Workspace

Clusters
SQL Endpoints

Databricks Workspace

Clusters
SQL Endpoints

Centralized Access Controls

Centrally grant and manage access permissions across workloads

Using ANSI SQL DCL

```
GRANT <privilege> ON <securable_type>
<securable_name> TO `<principal>`
```

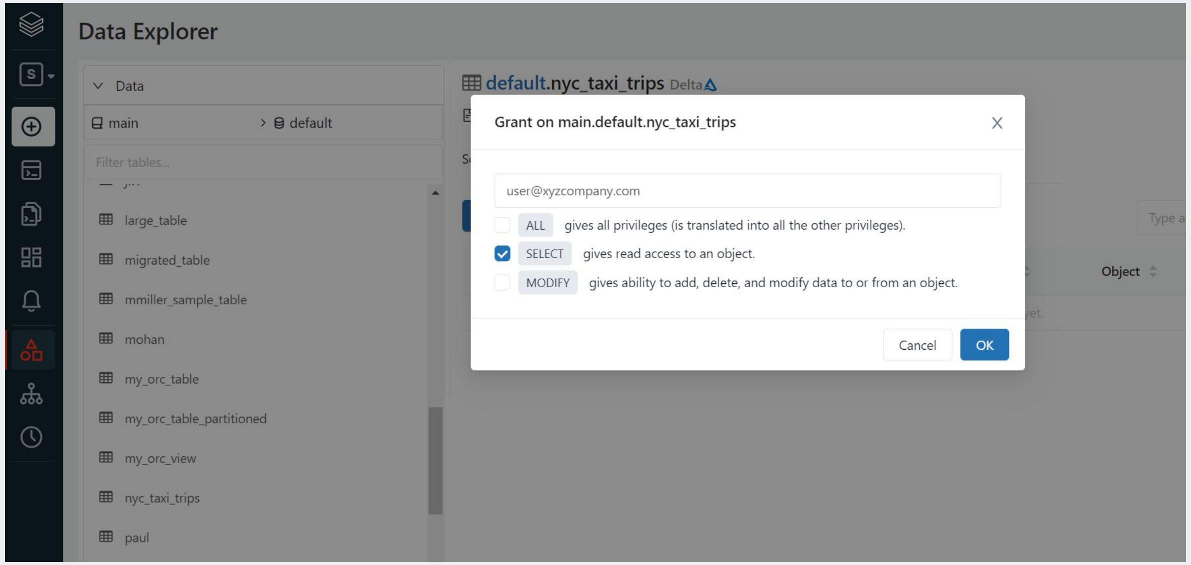
```
GRANT SELECT ON iot.events TO engineers
```

Choose permission level

'Table'= collection of files in S3/ADLS

Sync groups from your identity provider

Using UI



Grant on main.default.nyc_taxi_trips

user@xyzcompany.com

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

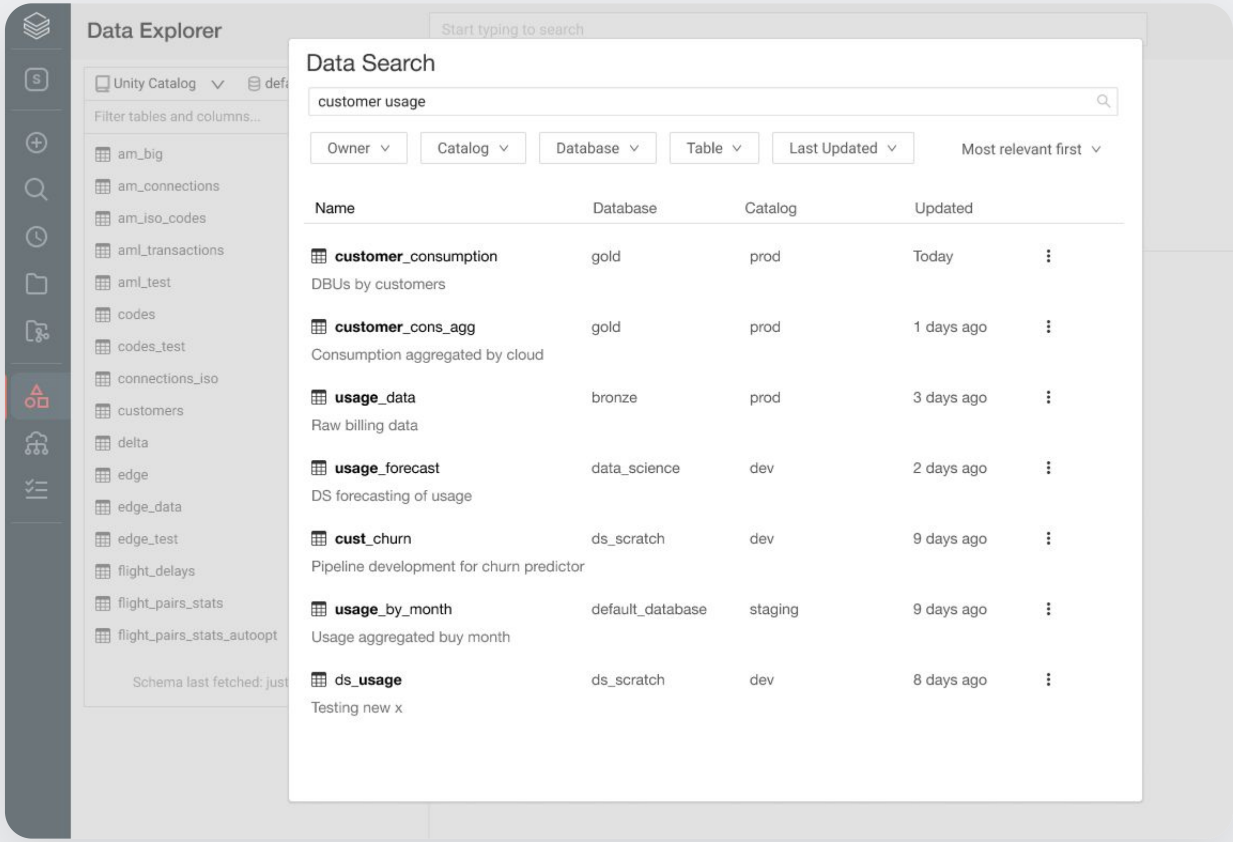
☑ SELECT gives read access to an object.

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

Data Discovery

Accelerate time to value with low latency data discovery

- UI to search for data assets stored in Unity Catalog
- Unified UI across DSML + DBSQL
- Leverage common permission model from Unity Catalog



am_big
am_connections
am_iso_codes
aml_transactions
aml_test
codes
codes_test
connections_iso
customers
delta
edge
edge_data
edge_test
flight_delays
flight_pairs_stats
flight_pairs_stats_autoopt

Schema last fetched: just

customer usage

Name	Database	Catalog	Updated
customer_consumption DBUs by customers	gold	prod	Today
customer_cons_agg Consumption aggregated by cloud	gold	prod	1 days ago
usage_data Raw billing data	bronze	prod	3 days ago
usage_forecast DS forecasting of usage	data_science	dev	2 days ago
cust_churn Pipeline development for churn predictor	ds_scratch	dev	9 days ago
usage_by_month Usage aggregated buy month	default_database	staging	9 days ago
ds_**usage** Testing new x	ds_scratch	dev	8 days ago

What’s next for Unity Catalog?

Attribute-based access control

Apply tags to any data asset

Create policies on tags

Simplify access control at scale

```
CREATE TAG finance_data
APPLY TAG finance_data TO TABLE global_sales
APPLY TAG finance_data TO SCHEMA cost_of_goods

GRANT SELECT ON TAG finance_data TO finance_group
```

Row filtering and column masking

Use SQL functions to create row filters and column masks

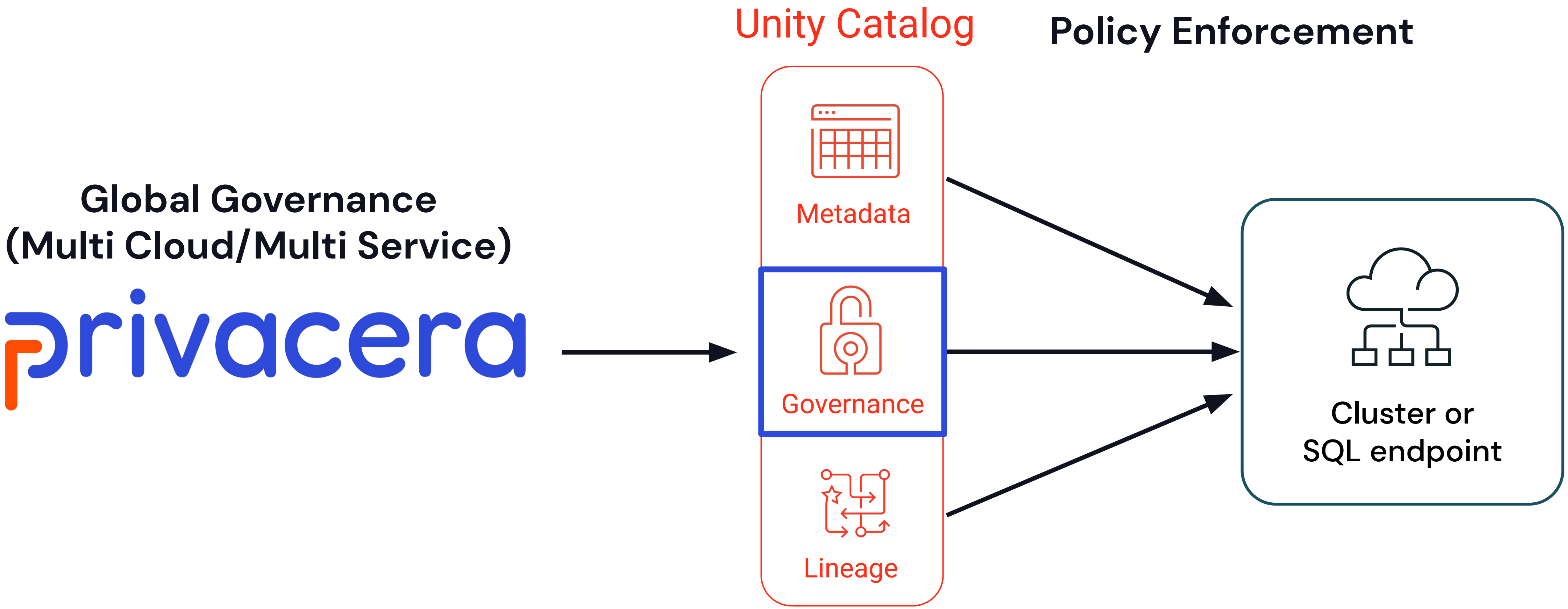
Apply to existing tables

```
CREATE FUNCTION us(region STRING) RETURNS BOOLEAN
  RETURN if(is_member(‘admin’), true, region=’US’)

ALTER TABLE sales SET ROW FILTER us ON (region)
```

```
CREATE FUNCTION mask(ssn STRING) RETURNS STRING
  RETURN if(is_member(‘admin’), ssn, ‘****’)

ALTER TABLE users ALTER COLUMN ssn SET MASK mask
```

Unity Catalog and Governance Partners

Better together

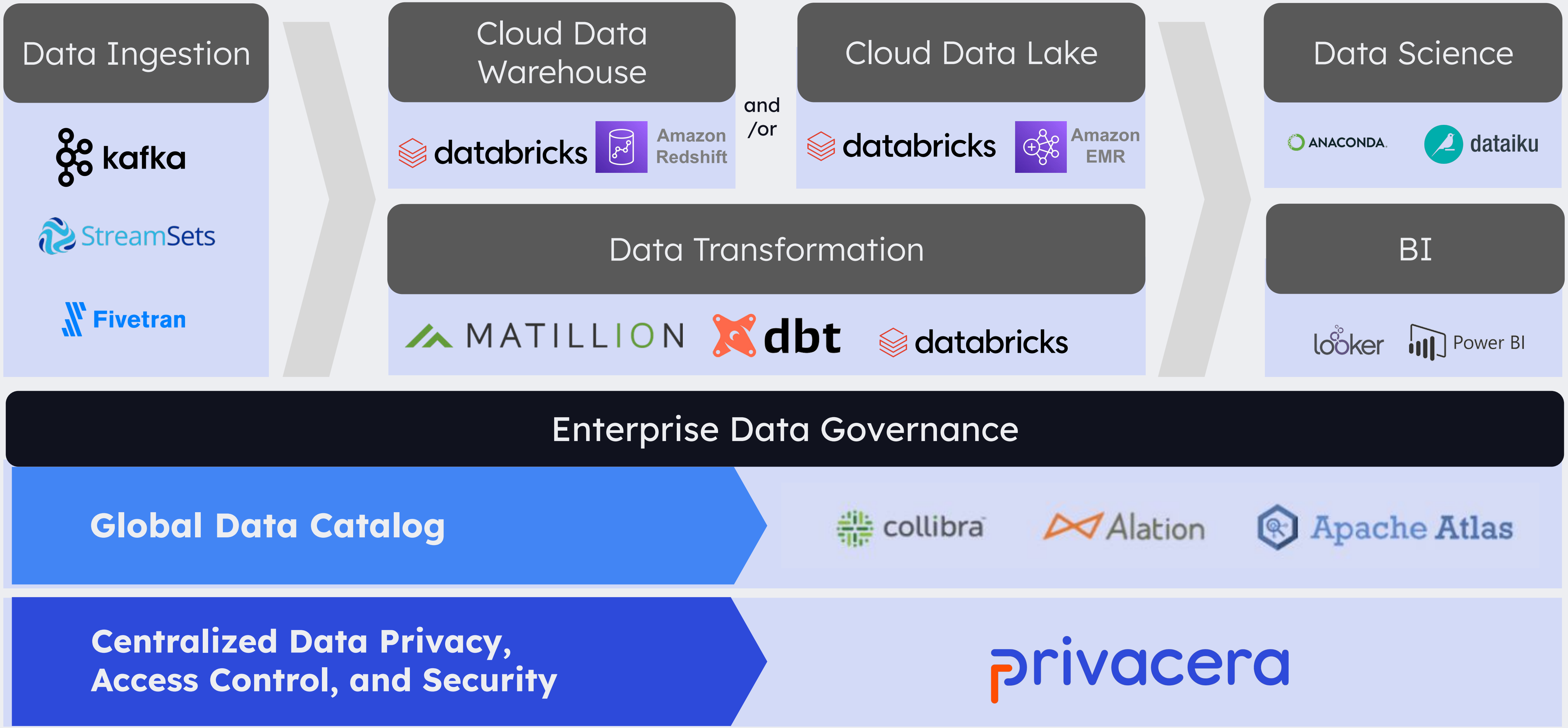


Extending Unity Catalog

Why extend Unity Catalog?

To address compliance and privacy requirements, **Data Governance** should be consistent across **all** of your **data services, tools** and **access patterns**

Multiple Data Services, Tools and Access Patterns



Data Ingestion

kafka

StreamSets

Fivetran

Cloud Data Warehouse

databricks

Amazon Redshift

and /or

Cloud Data Lake

databricks

Amazon EMR

Data Transformation

MATILLION

dbt

databricks

Data Science

ANACONDA.

dataiku

BI

looker

Power BI

Enterprise Data Governance

Global Data Catalog

collibra

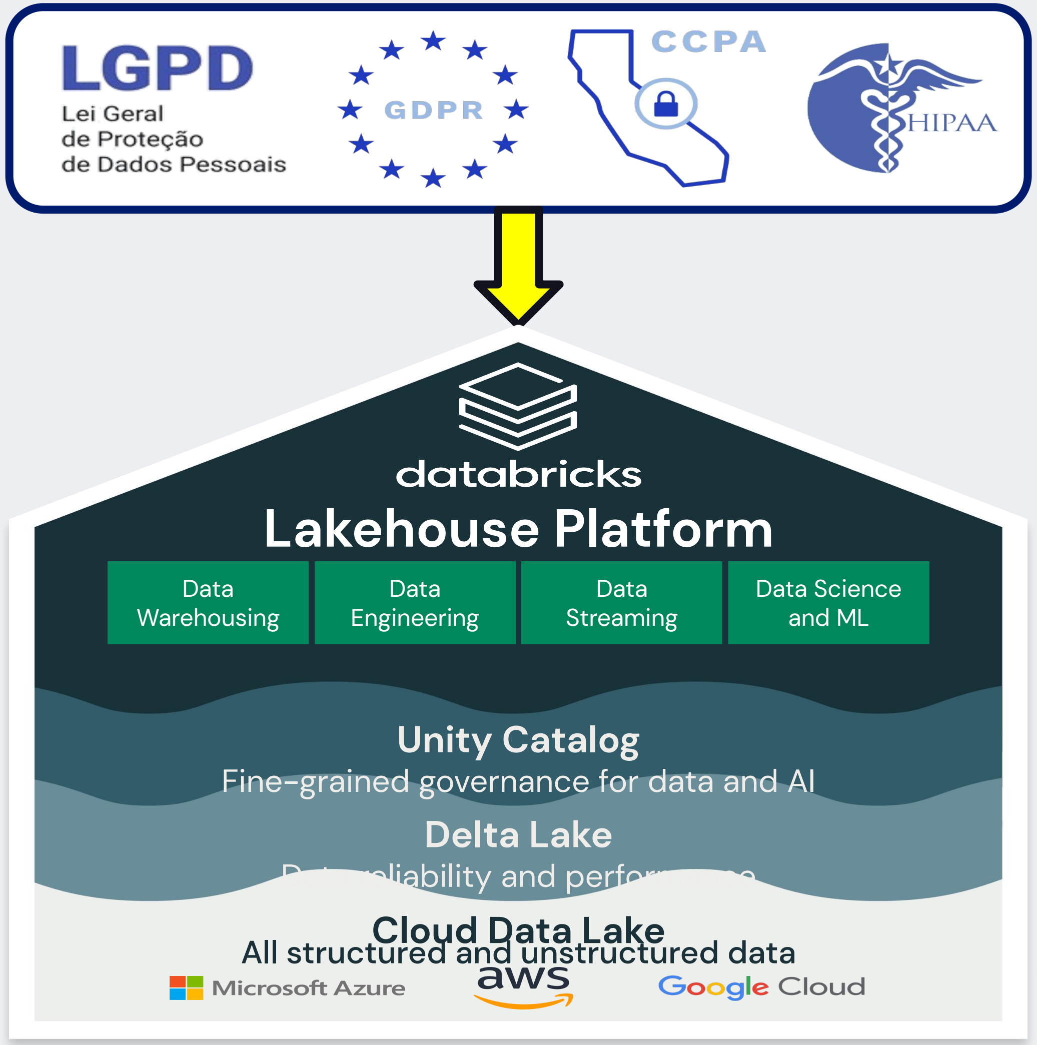
Alation

Apache Atlas

Centralized Data Privacy, Access Control, and Security

privacera

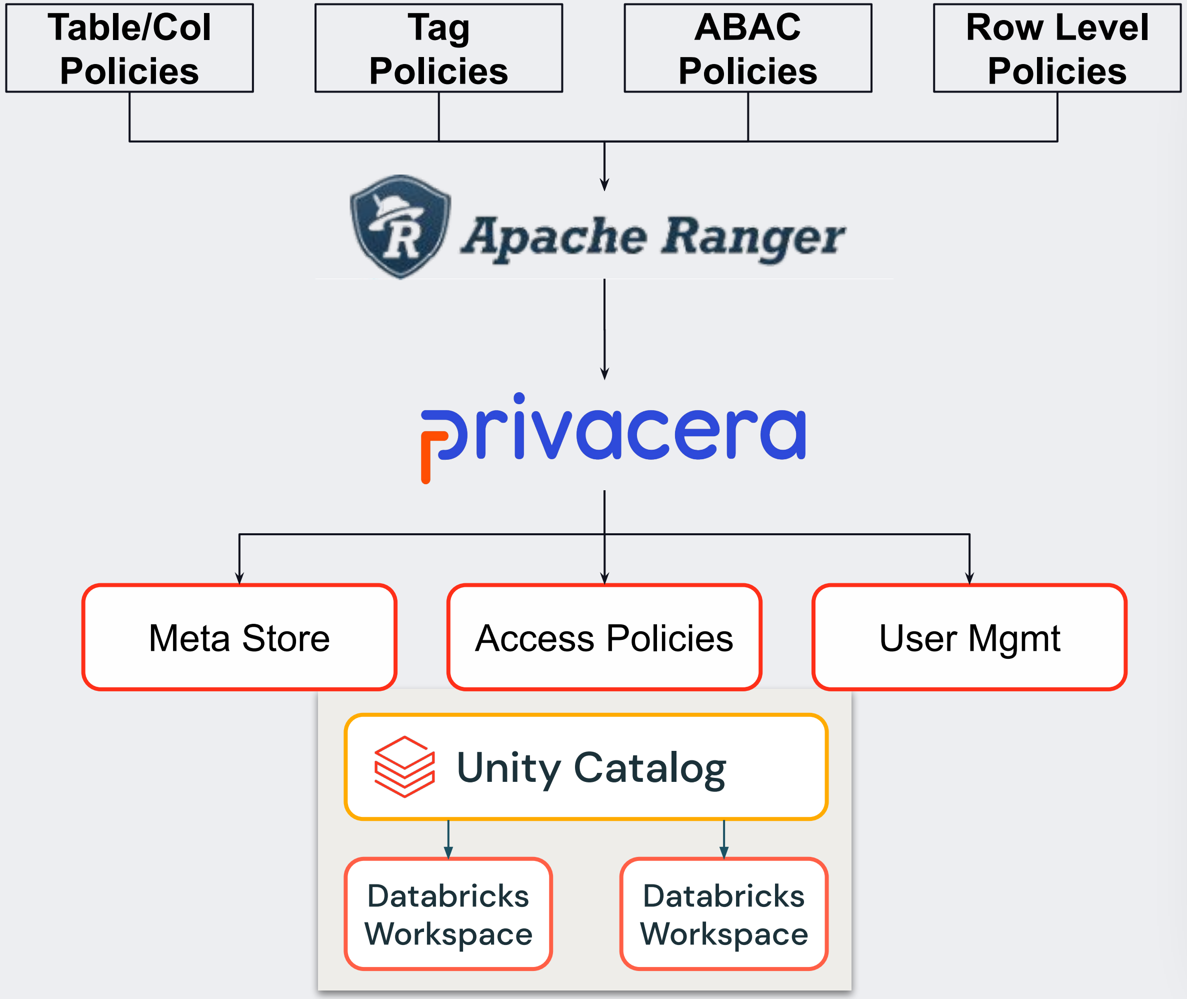
Compliance and Governance Requirements



Extend Unity Catalog to …

- Holistically apply compliance and privacy policies
- Integrate with central access management tools
- Integrate with external consent management systems
- Integrate with home grown entitlement systems

How does Unity Catalog Help?



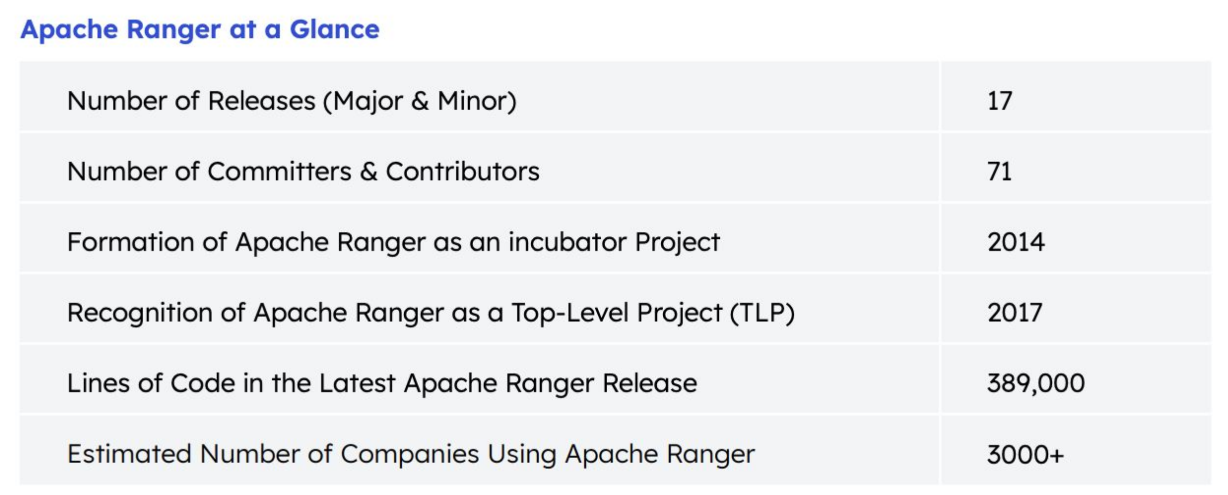
Unity Catalog – Foundational Capabilities

- REST APIs v/s GRANT/REVOKE
- Centralized metadata and user management for Databricks workspaces
- Centralized data access controls for Databricks workspaces
- Central access auditing for all workspaces

Apache Ranger

Centralized Access Management and Auditing Framework

- Used by 3000+ enterprises around the world
- Supporting plugins
  - AWS EMR, GCP Dataproc, Azure HDInsight
  - Dremio, Starburst, Trino, PrestoDB
  - Apache Submarine
  - Privacera –> Most Cloud Data Sources
- Centralized Auditing



Apache Ranger at a Glance

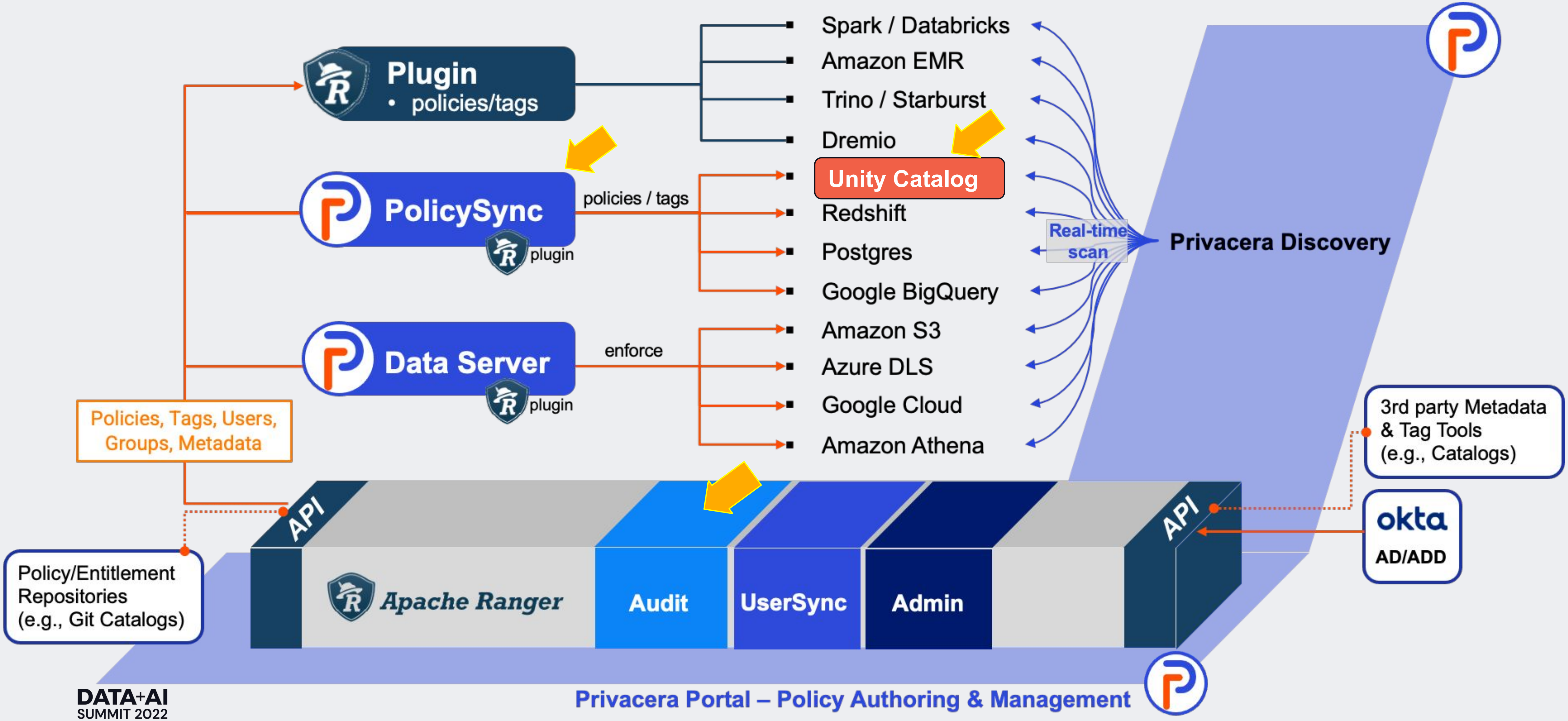
Number of Releases (Major & Minor)	17
Number of Committers & Contributors	71
Formation of Apache Ranger as an incubator Project	2014
Recognition of Apache Ranger as a Top-Level Project (TLP)	2017
Lines of Code in the Latest Apache Ranger Release	389,000
Estimated Number of Companies Using Apache Ranger	3000+

aws

Microsoft Azure

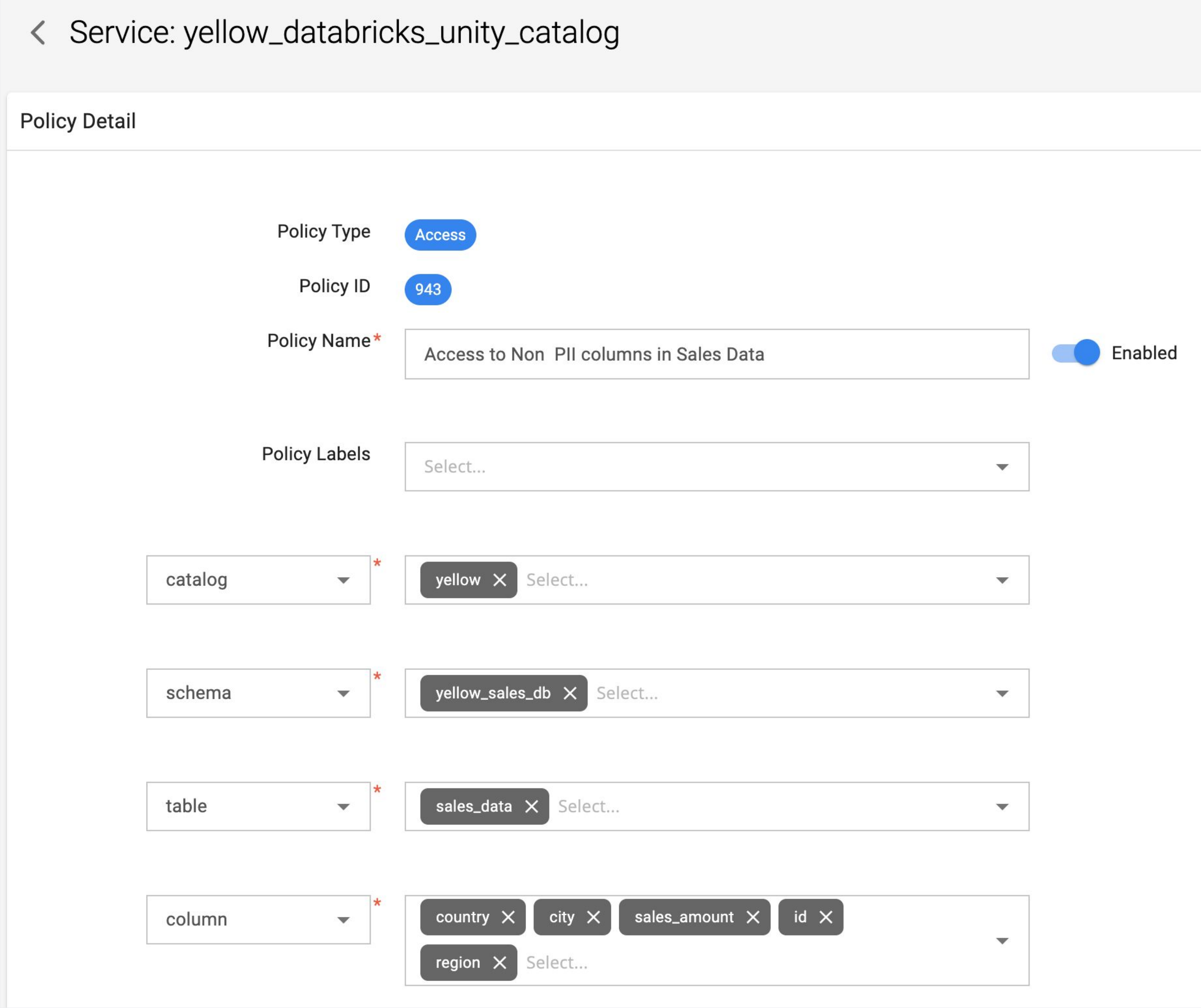
Google Cloud

Apache Ranger – Centralized Governance Framework



DATA+AI SUMMIT 2022

Apache Ranger Service Definition for Unity Catalog



Service: yellow_databricks_unity_catalog

Policy Detail

Policy Type: Access

Policy ID: 943

Policy Name*: Access to Non PII columns in Sales Data — Enabled

Policy Labels: Select...

catalog*: yellow

schema*: yellow_sales_db

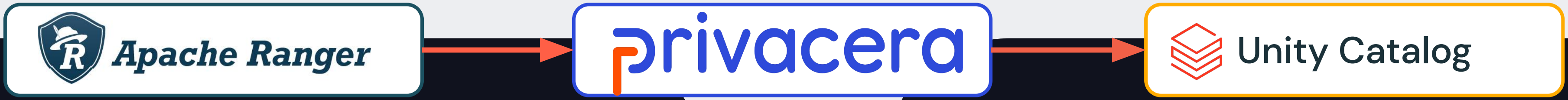
table*: sales_data

column*: country, city, sales_amount, id, region

DATA+AI SUMMIT 2022

Fine Grained Authorization

Translate Ranger Policy from YAML format to Unity Catalog JSON format



```
service: databricks_unity_catalog
resources:
  catalog:
    values:
      - sales_catalog
  schema:
    values:
      - sales_schema
  table:
    values:
      - sales_table
policyItems:
  - accesses:
      - type: Select
        isAllowed: true
    users:
      - emily.hope
```

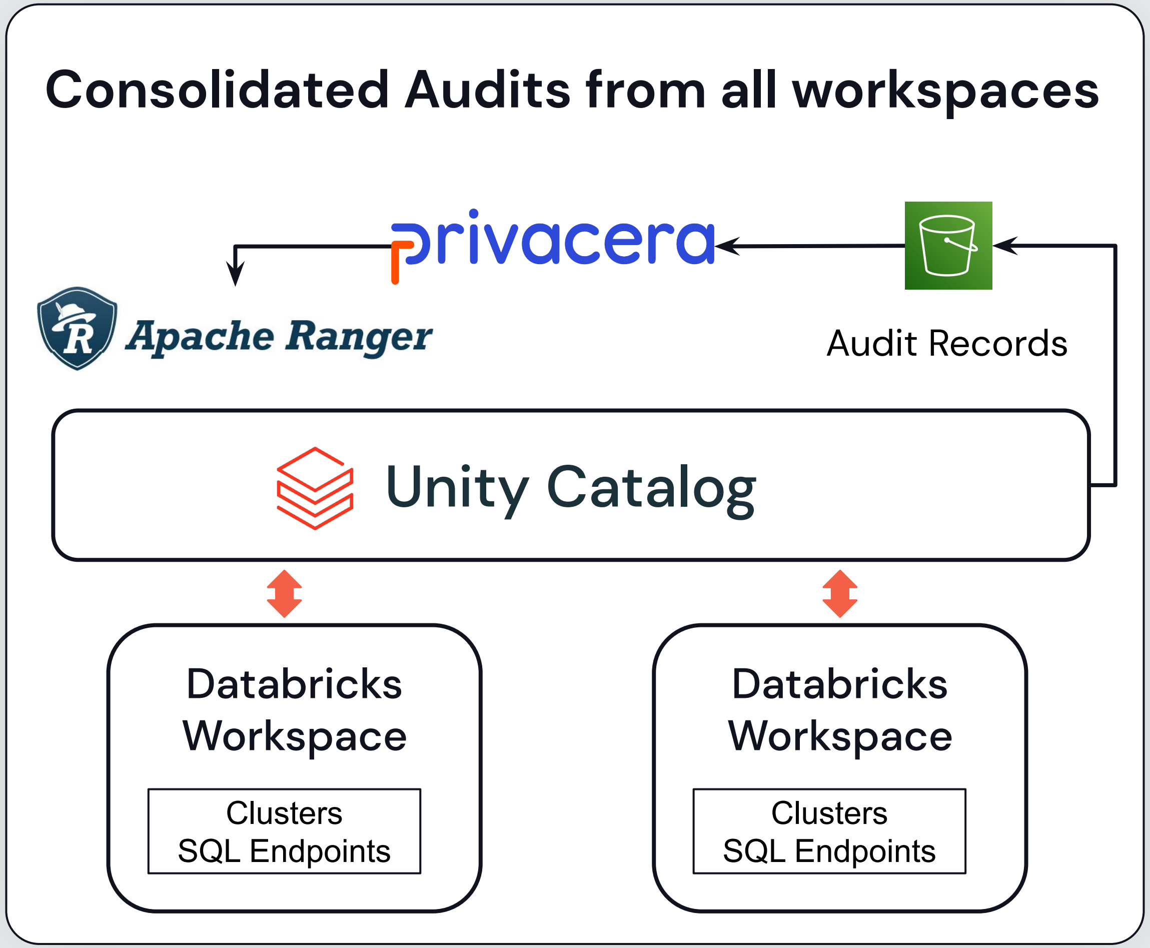
```
https://your-uc-workspace.cloud.databricks.com/api/2.0/unity-catalog//permissions/table/sales_catalog.sales_schema.sales_table

{
    "privilege_assignments": [
        {
            "principal": "emily.hope@acme.com",
            "privileges": [
                "SELECT"
            ]
        }
    ]
}
```

Centralized Auditing

Synchronize Audits from Unity Catalog to Apache Ranger

Steps

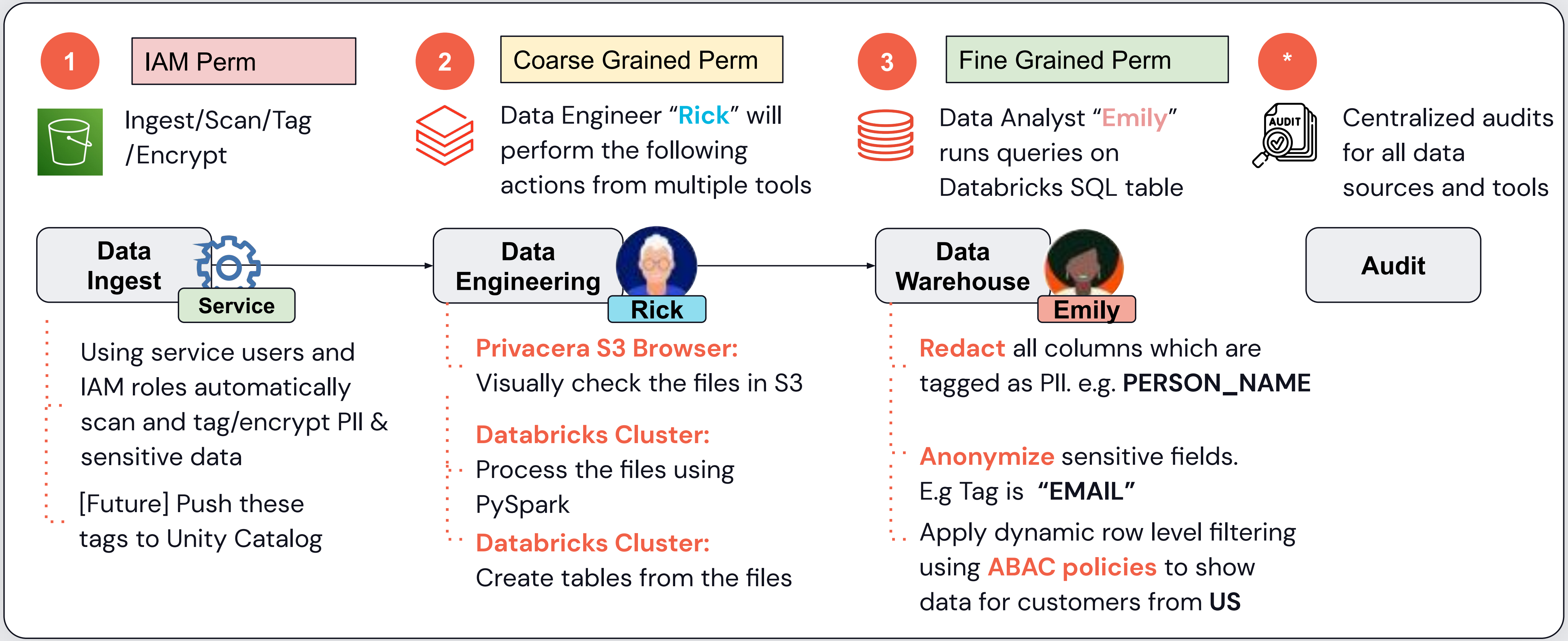
- Unity Catalog consolidates audits from workspace and stores in object store (S3/GCS/ADLS)
- Audits can be access using SQL table over the audit storage location
- Privacera pulls the audits by Unity Catalog and stores in Apache Ranger
- Consolidated audits from all services can be accessed from Apache Ranger



DATA+AI SUMMIT 2022

Simplifying Access Governance Implementation

3 Stages of Data Life Cylce and Personas



1 IAM Perm

Ingest/Scan/Tag /Encrypt

2 Coarse Grained Perm

Data Engineer “**Rick**” will perform the following actions from multiple tools

3 Fine Grained Perm

Data Analyst “**Emily**” runs queries on Databricks SQL table

* Centralized audits for all data sources and tools

Data Ingest → **Data Engineering** → **Data Warehouse**   **Audit**

Data Ingest (Service)

Using service users and IAM roles automatically scan and tag/encrypt PII & sensitive data

[Future] Push these tags to Unity Catalog

Data Engineering (Rick)

Privacera S3 Browser: Visually check the files in S3

Databricks Cluster: Process the files using PySpark

Databricks Cluster: Create tables from the files

Data Warehouse (Emily)

Redact all columns which are tagged as PII. e.g. **PERSON_NAME**

Anonymize sensitive fields. E.g Tag is **“EMAIL”**

Apply dynamic row level filtering using **ABAC policies** to show data for customers from **US**

Demo

Privacera and Unity Catalog together

Brings simpler governance across any data, any cloud

Privacera + Unity Catalog

- ✓ Data Governance across hybrid and multi-cloud
- ✓ Sensitive Data discovery, fine grained access management and encryption across any data
- ✓ Automated workflows to reduce data and user onboarding time
- ✓ Centralized auditing and canned reports for security and compliance

Unity Catalog

- ✓ Metadata and user management for lakehouse
- ✓ Access control and auditing for the lakehouse
- ✓ APIs to integrate with partner solutions

DATA+AI
SUMMIT 2022

Thank you

Don Bosco Durai

Co founder & CTO, Privacera

Zeashan Pappa

Senior Product Specialist - Data Governance