UNITY CATALOG

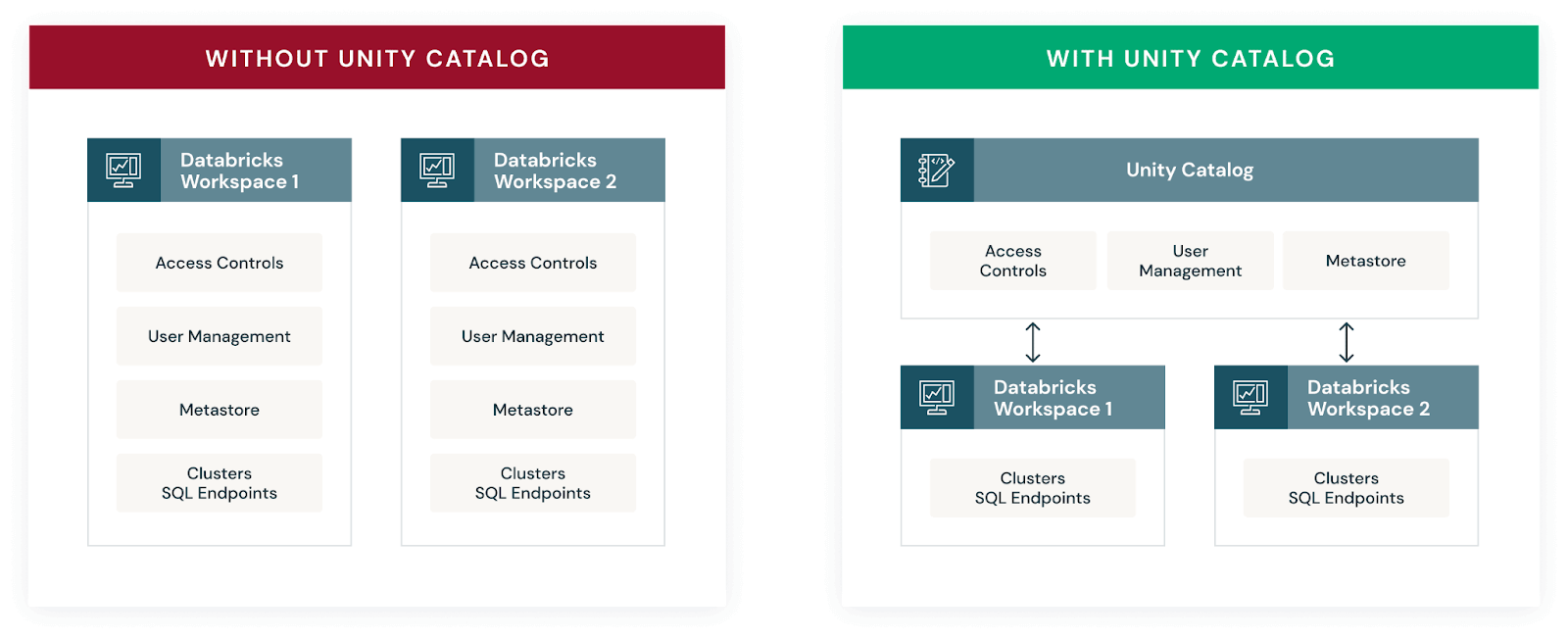
[Jobaoushadancse2021@jerusalemengg.ac.in](mailto:Jobaoushadancse2021@jerusalemengg.ac.in)

Job Aoushadan N

**What Is the Unity Catalog in Databricks?**

Databricks announced Unity Catalog at the Data and AI Summit in 2021 to address the complexities of data governance by providing fine-grained access control within the Databricks ecosystem. Before Databricks Unity Catalog, data governance in Databricks was typically handled by various third party and open source tools, which, while effective, often lacked seamless integration with the Databricks ecosystem. These tools often lacked the integration and granular security controls specifically tailored for data lakes and were sometimes limited to certain cloud platforms. This limitation highlighted the need for a data governance solution that could provide more fine-grained access control and work across different cloud platforms while integrating seamlessly with the Databricks ecosystem. Thus, Databricks Unity Catalog was born to exactly address these challenges.

Databricks Unity Catalog serves as a centralized governance layer within the Databricks Data Intelligence Platform, streamlining the management and security of various data and AI assets. It supports a wide range of assets including files, tables, machine learning models, notebooks, and dashboards. Unity Catalog uniquely identifies each asset type, simplifying access control and ensuring that only authorized users can interact with specific data elements.

Databricks Unity Catalog diagram (Source:databricks.com)

Now, If you’re transitioning from Hive metastore to Databricks Unity Catalog, rest assured that the Databricks Unity Catalog enhances the Hive metastore experience. It not only supports the same functionalities but also introduces advanced governance features. Unity Catalog enables fine-grained access control, allowing you to designate column-level permissions to safeguard sensitive data, such as Personally Identifiable Information (PII). Also, it provides visibility into data lineage, tracing the flow and transformation of your data across different processes.

But that's not all! Databricks Unity Catalog also allows you to share selected tables across different platforms and clouds using Delta sharing.

Key features and benefits of Databricks Unity Catalog:

1. **Define once, secure everywhere**: Databricks Unity Catalog offers a single place to administer data access policies that apply across all workspaces and user personas, ensuring consistent governance.
2. **Standards-compliant security model**: Databricks Unity Catalog's security model is based on standard ANSI SQL, allowing administrators to grant permissions using familiar syntax at various levels (catalogs, schemas, tables, and views).
3. **Built-in auditing and lineage**: Databricks Unity Catalog automatically captures user-level audit logs and data lineage, tracking data access and usage patterns for compliance and troubleshooting purposes.
4. **Data discovery**: Databricks Unity Catalog enables tagging and documenting data assets, providing a search interface to help users easily find and understand the data they need.
5. **System tables (Public Preview)**: Databricks Unity Catalog provides access to operational data, including audit logs, billable usage, and lineage, through system tables.
6. **Managed storage**: Databricks Unity Catalog supports managed storage locations at the metastore, catalog, or schema levels, enabling organizations to isolate data physically in their cloud storage based on governance requirements.
7. **External data access**: Databricks Unity Catalog allows registering and governing access to external data sources, such as cloud object storage, databases, and data lakes, through external locations and Lakehouse Federation.

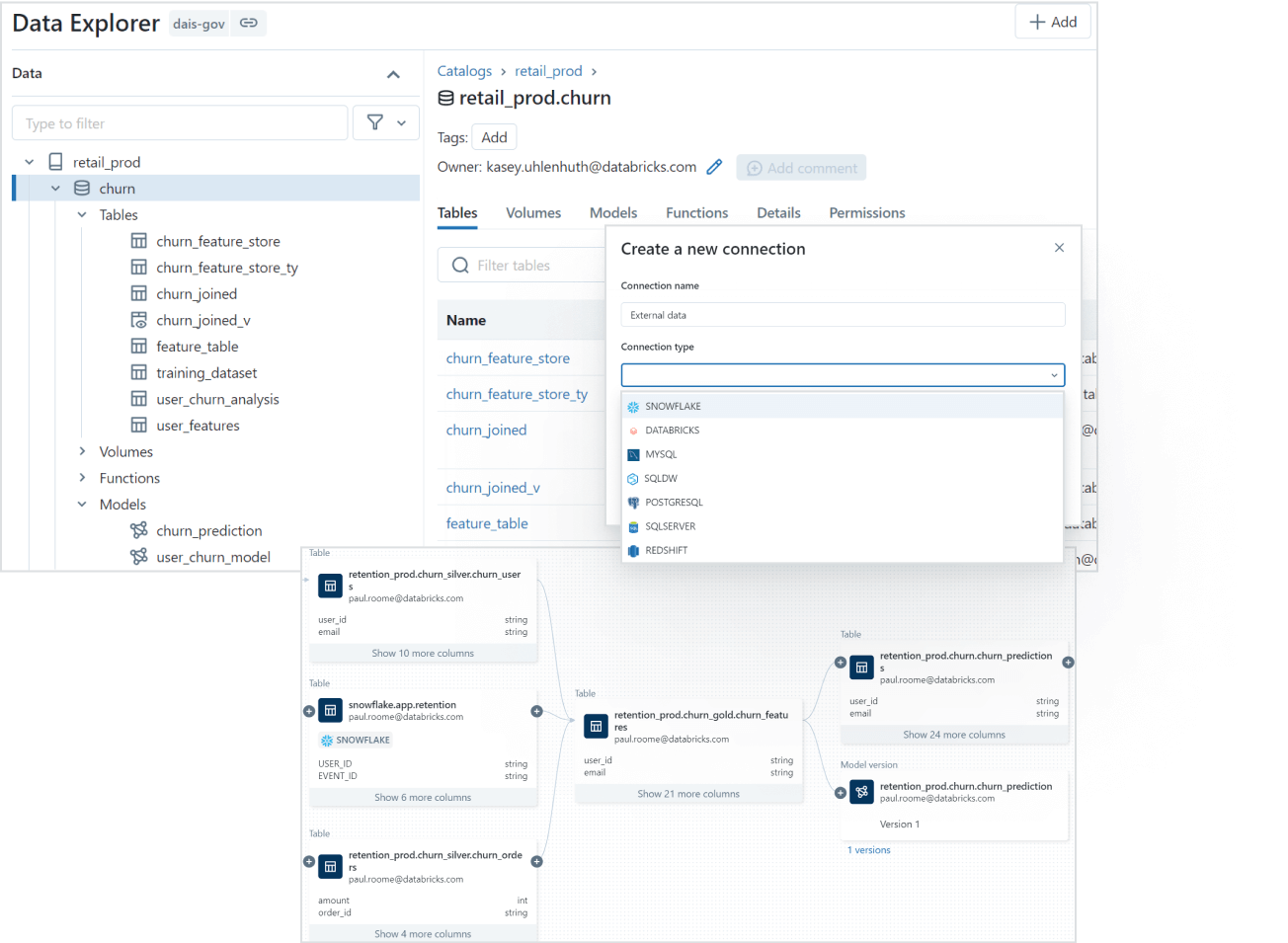
Source: Databricks

**Databricks Unity Catalog Architecture Breakdown**

Here’s an architecture breakdown of Databricks Unity Catalog:

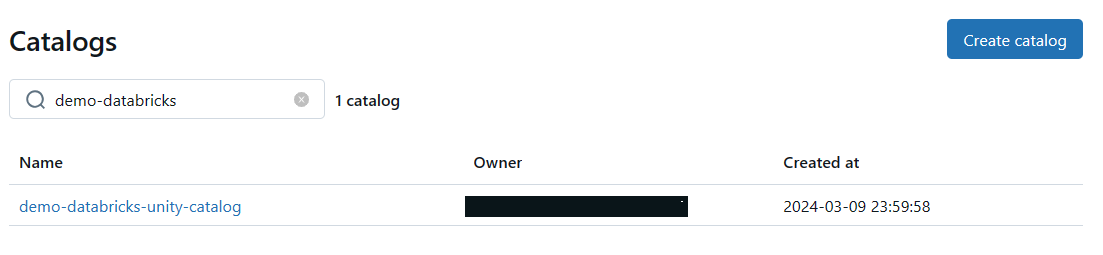
**1) Unified Governance Layer**

Databricks Unity Catalog offers a unified governance layer for both structured and unstructured data, tables, machine learning models, notebooks, dashboards and files  across any cloud or platform. This layer enables organizations to govern their data and AI assets seamlessly, ensuring regulatory compliance and accelerating data initiatives.

Unified Governance Layer of Databricks Unity Catalog (Source:databricks.com)

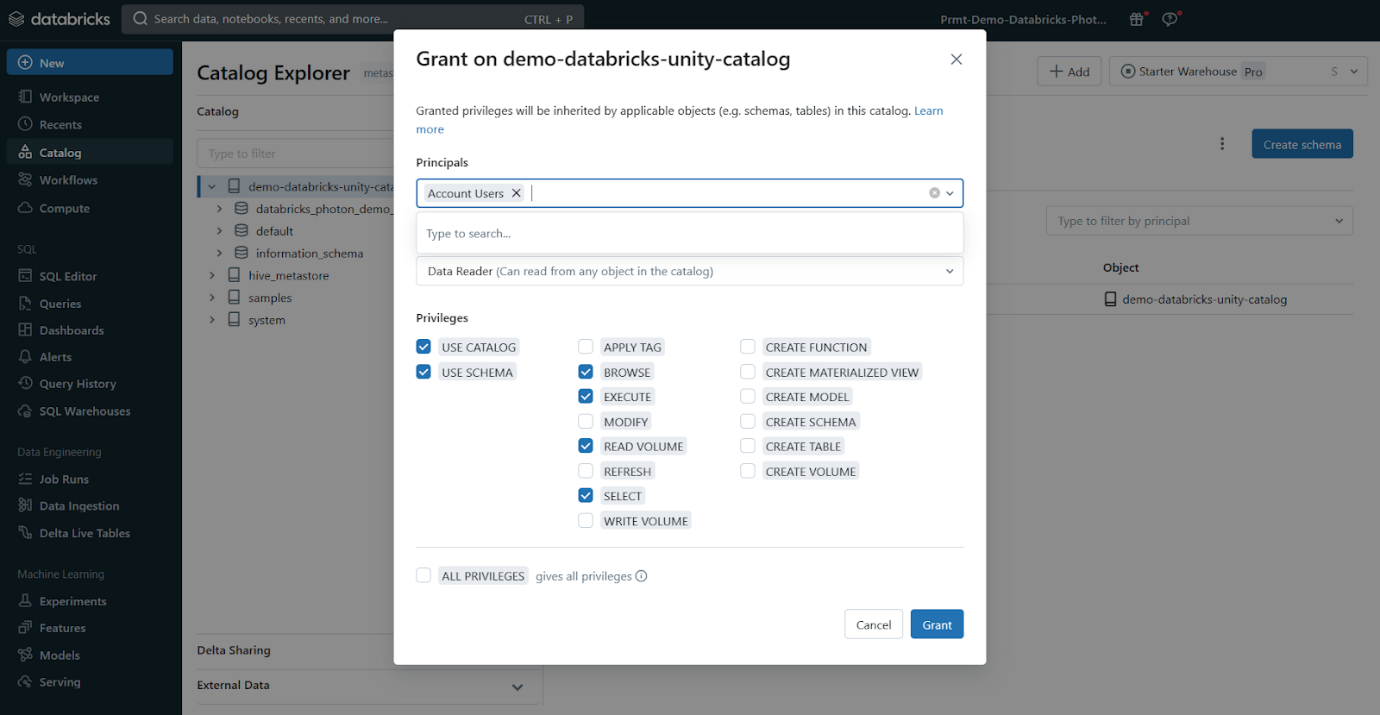
**2) Data Discovery**

Finding the right data for your needs is simple with Databricks Unity Catalog's discovery features. You can tag and document your data assets, and then use the search interface to locate the specific data you need based on keywords, tags, or other metadata.

Databricks Unity Catalog's data discovery features

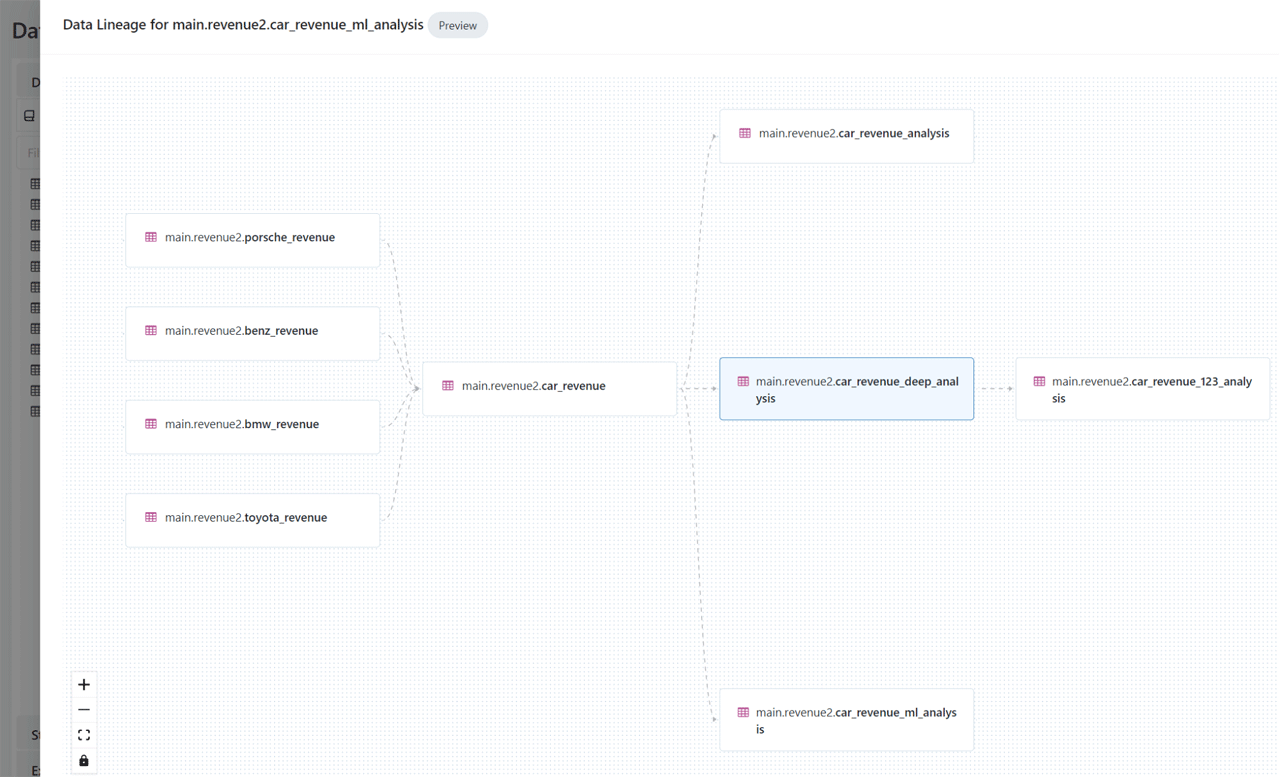
**3) Access Control and Security**

Databricks Unity Catalog simplifies access management by providing a single interface to define access policies on data and AI assets. It supports fine-grained control on rows and columns and manages access through low-code attribute-based access policies that scale seamlessly across different clouds and platforms.

Access Control and Security - Databricks Unity Catalog

**4) Auditing and Lineage**

Databricks Unity Catalog automatically captures audit logs that record who accessed which data assets and when. It also tracks the lineage of your data, so you can see how assets were created and how they're being used across different languages and workflows. This lineage information is crucial for understanding data flows and dependencies.

Auditing and Lineage - Databricks Unity Catalog

**5) Open Data Sharing**

Databricks Unity Catalog integrates with open source Delta Sharing, which allows you to securely share data and AI assets across clouds, regions, and platforms without relying on proprietary formats or complex ETL processes.

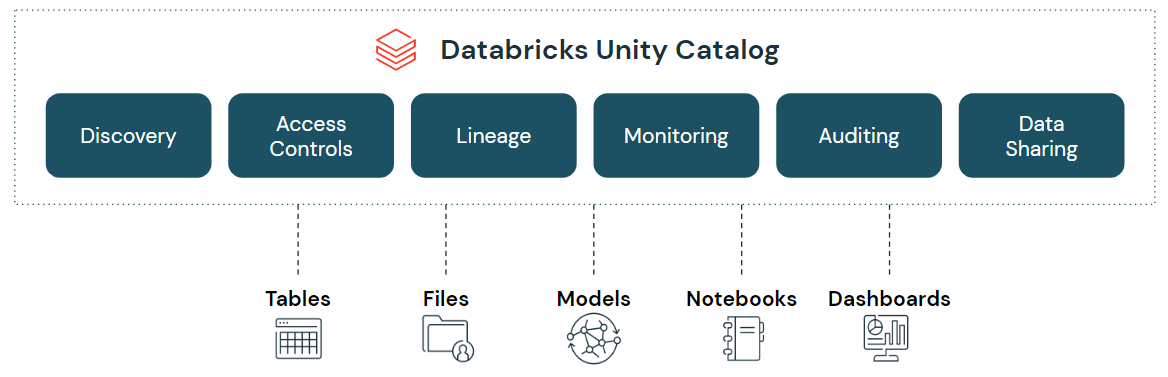
Databricks Delta Sharing - Databricks Unity Catalog

**6) Object Model**

Databricks Unity Catalog organizes your data and AI assets into a hierarchical structure: Metastore ► Catalog ► Schema ► Tables, Views, Volumes, and Models. At the top level, you have the metastore, which contains your schemas. Within each schema, you can have tables, views, or volumes (for unstructured data). To reference any asset, you use a three-part naming convention: <catalog>.<schema>.<asset>. We will dive deeper into the object model in Unity Catalog in next section

**7) Operational Intelligence**

Databricks Unity Catalog provides AI-powered monitoring and observability capabilities that give you deep insights into your data and AI assets. You can set up active alerts, track data lineage at the column level, and gain comprehensive visibility into how your assets are being used and managed.

Databricks Unity Catalog (Source:databricks.com)

Databricks Unity Catalog (Source: databricks.com)

**Databricks Unity Catalog Object Model**

Databricks Unity Catalog follows a hierarchical architecture with the following components:  Metastore ► Catalog ► Schema ► Tables, Views, Volumes, and Models.

* **Metastore**: The top-level container for metadata, exposing a three-level namespace (catalog.schema.table) to organize data assets.
* **Catalog**: The first layer of the object hierarchy, used to organize data assets logically, often aligned with organizational units or data domains.
* **Schema**: Also known as databases, schemas are the second layer of the object hierarchy, containing tables, views, and volumes.
* **Tables, Views, and Volumes**: The lowest level in the data object hierarchy, where:
  + **Table**: A structured data asset that represents a collection of rows with a defined schema.
  + **View**: A virtual table that is defined by a query.
  + **Volume**: A container for unstructured (non-tabular) data files.
* **Models**: Machine learning models registered in the MLflow Model Registry can also be managed within Databricks Unity Catalog.