Unity Catalog in Azure Databricks

1. What is Unity Catalog?

Unity Catalog is a **data governance and security feature** in Azure Databricks.
It provides a **central place** to manage access, permissions, and auditing for **all your data and AI assets** (tables, files, ML models, dashboards).

Think of it as the **"security + organization system"** for Databricks data.

---

2. Why Do We Need Unity Catalog?

- In big companies, data is spread across multiple databases, data lakes, and warehouses.
- Without governance, it's hard to control **who can see what**.
- Unity Catalog ensures:
    - **Centralized access control** (one place to set permissions).
    - **Consistent governance** across workspaces.
    - **Data lineage tracking** (who used what data, and how).
    - **Compliance** with regulations (GDPR, HIPAA, etc.).

---

3. Key Features

1. **Centralized Data Access Control**
    - Manage permissions for users and groups at **table, schema, or catalog level**.
2. **Data Lineage**
    - Track where data comes from, how it was transformed, and where it's used.

3. **Audit & Security**
    - Full logging of data access for compliance.
4. **Supports Multiple Storage Systems**
    - Works with data in **ADLS, Amazon S3, GCP Storage, Delta tables**.
5. **Fine-Grained Permissions**
    - Grant access at different levels:
        - Catalog → Database/Schema → Table/View → Column.

---

4. How Unity Catalog Organizes Data

Unity Catalog introduces a **3-level namespace**:

`catalog.schema.table`

- **Catalog** → Highest level (e.g., `sales_catalog`)
- **Schema** → Group of tables/views (like a database, e.g., `retail_schema`)
- **Table** → Actual dataset (e.g., `orders`)

Example: `sales_catalog.retail_schema.orders`

---

5. Benefits

- Single control point for permissions (no scattered policies).
- Works across **multiple Databricks workspaces**.
- Helps with **compliance and audits**.
- Makes collaboration easier without risking data leaks.

---

6. Simple Example

Imagine a company with sales data stored in ADLS:

- Without Unity Catalog → Each workspace manages access separately, causing duplication and risk.

- With Unity Catalog → Define access once (e.g., *"Only analysts can query customer tables"*), and it applies everywhere.

---

7. Summary

Unity Catalog = **One place to secure, organize, and audit all data in Databricks.**
It helps companies manage **who can access what data** across all projects, while keeping track of usage and ensuring compliance.

Short analogy:

- **Databricks** = the big library.

- **Unity Catalog** = the librarian who controls who can borrow which books, keeps a record of usage, and ensures no rules are broken.
