

Domain 2: Secure and govern Unity Catalog objects in Azure Databricks

Secure Unity Catalog objects

1. Introduction

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/1-introduction>

Introduction

Completed



Securing data is fundamental to enterprise data platforms, yet traditional approaches often fragment governance across multiple systems and require constant maintenance. As organizations scale their analytics workloads on Azure Databricks, they need unified security that protects sensitive information while maintaining operational efficiency. **Unity Catalog** addresses this challenge by providing centralized governance for data assets, compute resources, and access controls.

When you work with Unity Catalog, security operates at multiple layers. You control who can access catalogs, schemas, and tables through **granular permissions**. You enforce **row-level and column-level restrictions** to ensure users see only the data they're authorized to view. You authenticate data access through **service principals** or **managed identities**, eliminating hardcoded credentials from your code. You retrieve sensitive configuration values from **Azure Key Vault** without exposing secrets in notebooks or job configurations.

Understanding how these security mechanisms work together enables you to build data solutions that balance accessibility with protection. You need to know when inherited permissions simplify administration and when explicit grants provide necessary control. You must choose between row and column security functions versus dynamic views based on your governance requirements. You should recognize when managed identities offer advantages over service principals for authenticating storage access.

This module guides you through Unity Catalog's security model, from understanding the **query lifecycle** to implementing **authentication strategies**. You explore access control patterns that scale across your organization, fine-grained filtering techniques that protect sensitive data, and credential management approaches that strengthen your security posture. By the end, you'll be equipped to design and implement comprehensive security for your Azure Databricks environment.

2. Understand query lifecycle

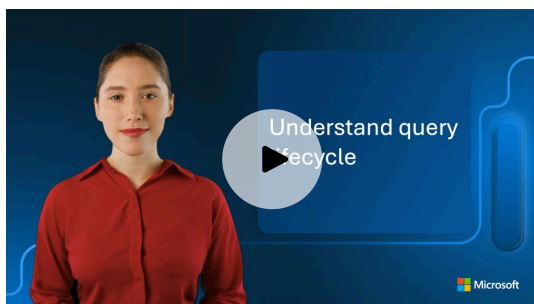
<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/2-understand-query-lifecycle>

Understand query lifecycle

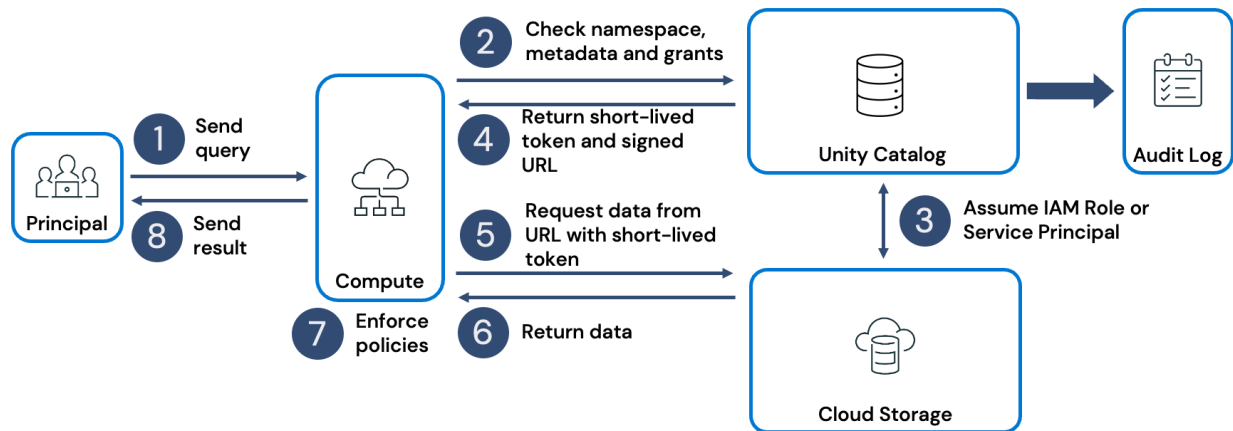
Completed

Let's examine how Unity Catalog manages query execution, with particular attention to security and administrative controls. We'll walk step by step through what happens when a query is submitted against a Unity Catalog table, covering the interactions between compute, the control plane, the data plane, and cloud storage.

The query lifecycle in Unity Catalog



The following diagram illustrates the overall flow of a query as it moves through different components in Unity Catalog.



Step 1: Query submission

The lifecycle begins when a principal (a user or service identity) issues a query. Queries can originate in different ways:

- A data scientist might use an **all-purpose cluster** for interactive Python or SQL workloads.
- A **service principal** might run a scheduled pipeline or job on a **job cluster**.
- A **data analyst** might send a query through **Databricks SQL**, using a SQL warehouse.
- A **business intelligence (BI) tool** connected to a SQL warehouse could also generate the request.

In each case, the compute resource—cluster or warehouse—receives the query and begins execution.

Step 2: Request validation in Unity Catalog

The compute resource forwards the request to **Unity Catalog**. Unity Catalog acts as the control plane for security and governance. It records the request in the **audit log** and checks whether the principal has the necessary permissions to access the objects referenced in the query. If permissions are denied, the query is blocked and the denial is logged. If access is granted, the request moves forward.

Step 3: Assuming cloud credentials

For every object referenced in the query, Unity Catalog assumes the appropriate **cloud credential** associated with that object. These credentials are configured by a cloud administrator.

- For **managed tables**, the credential typically points to the cloud storage tied to the Unity Catalog metastore.
- For **external tables or files**, the credential corresponds to an external location governed by a defined storage credential.

Note

In Azure, the preferred way to grant Unity Catalog access to Azure Data Lake Storage (ADLS Gen2) is via a **Managed Identity + Access Connector** (versus relying purely on service principals). They have the benefit of allowing Unity Catalog to access storage accounts protected by network rules, which isn't possible using service principals, and they remove the need to manage and rotate secrets.

Step 4: Issuing scoped tokens

Once credentials are validated, Unity Catalog generates a **scoped temporary access token** for each object. Along with the token, it provides a secure access URL. This allows the compute resource to read directly from storage without exposing long-lived credentials.

Step 5: Data access from storage

The compute resource (cluster or SQL warehouse) uses the token and URL provided by Unity Catalog to request data directly from the underlying ADLS Gen2 endpoints (via `abfss://` or `dfs.core.windows.net` URL)

Note

If your ADLS account has firewall or virtual network restrictions, you must explicitly allow the Azure Databricks access connector / managed identity to access the storage (in addition to allowing the compute nodes). If the storage account is locked down, even a valid token may be rejected if the identity is not allowed via firewall rules.

Step 6: Data transfer

Cloud storage returns the requested data to the compute resource. This process is repeated for every object referenced in the query.

Step 7: Fine-grained filtering

Unity Catalog enforces **row- and column-level filters** on the compute resource itself. This ensures that principals only see the exact subset of data they're entitled to access.

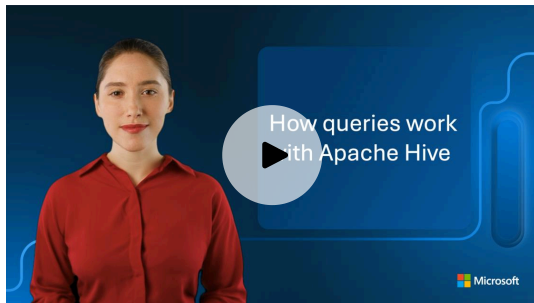
- **Row-level filtering** introduces conditions that are evaluated at query execution time inside the compute engine. Even though the underlying storage system may return a broader dataset, Unity Catalog ensures that only the rows matching the defined conditions are made visible to the requesting principal. This allows multiple users to query the same table but each receives a personalized, restricted view of the data. A row filter function is basically a predicate (a SQL expression or user-defined function) that is evaluated during query execution.
- **Column-level filtering** operates similarly but at the attribute level. Specific columns can be masked or transformed so that sensitive information is hidden or anonymized, depending on

the privileges of the principal. This makes it possible to share datasets broadly while ensuring that confidential fields are only fully visible to those with appropriate authorization.

Step 8: Returning the result

The filtered query result is returned from the compute resource to the calling user, job, or application.

How queries work with Apache Hive



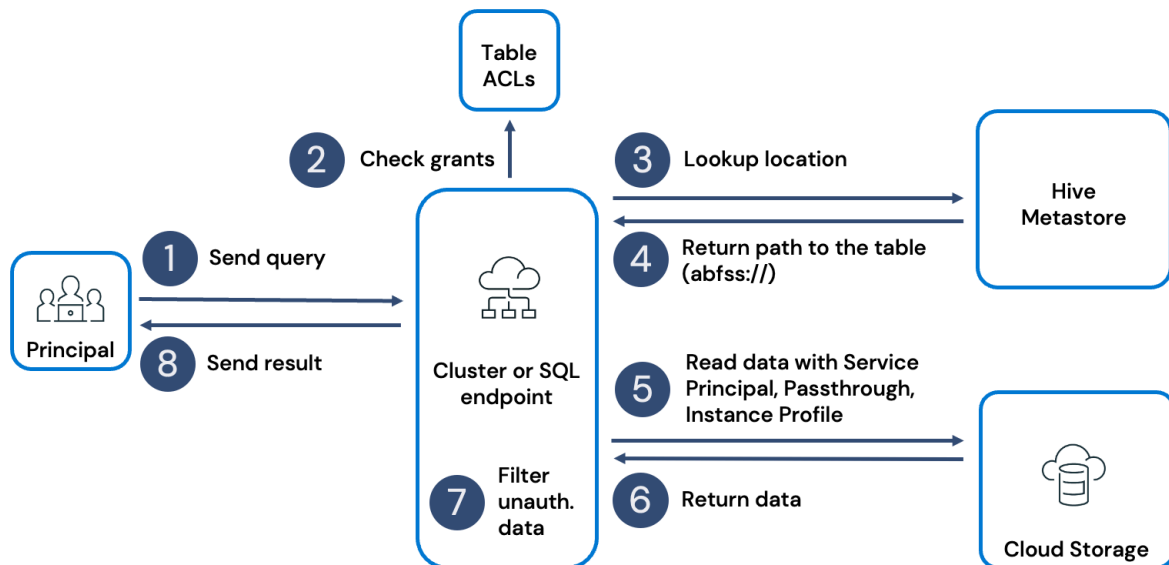
Unity Catalog was designed to streamline governance, but it also maintains compatibility with the legacy Hive metastore. When a workspace is assigned to a Unity Catalog metastore, Hive appears as a special catalog named **hive_metastore**. Tables stored there can still be queried by referencing this catalog in the namespace.

Note

Hive metastore table access control is a legacy data governance model.

The lifecycle of a query against a Hive table differs from Unity Catalog in several ways:

- Hive doesn't provide centralized, fine-grained governance. Instead, administrators often need to manage **service accounts, secrets, or instance profiles** for authentication and authorization.
- Access to data may involve setting up **mount points, passthrough authentication, or storage policies**, each requiring additional configuration.
- Audit logging and access controls are less integrated compared to Unity Catalog, which can lead to inconsistencies in security enforcement.



Step 1: Query submission

A principal (user or service) sends a query to a cluster or SQL endpoint. This is the starting point for all Hive queries.

Step 2: Access check

The cluster checks the query against **table access control lists (ACLs)**. These ACLs determine whether the requesting principal has the right to access the specified table.

Step 3: Location lookup

If access is granted, the cluster consults the **Hive metastore** to find the storage location of the table. The metastore contains metadata about tables, including their schemas and storage paths.

Step 4: Path returned

The Hive metastore returns the **storage path** to the cluster, usually as a URI (for example, `abfss://...`). This path points to the physical location of the table data in cloud storage.

Step 5: Data request

The cluster attempts to read data from the cloud storage. At this stage, it must authenticate using one of several mechanisms, such as a **service principal, passthrough authentication, or an instance profile**. Administrators often need to configure these credentials in advance, along with secrets and mount points.

Step 6: Data returned

Cloud storage sends the requested data back to the cluster.

Step 7: Enforcement of policies

The cluster enforces any last-mile filtering to remove data the principal shouldn't see. This may involve applying row- or column-level restrictions at query time.

Step 8: Result delivery

The final query result is sent back to the principal who issued the request.

In practice, this means querying Hive involves more manual administrative steps and ongoing maintenance. Unity Catalog reduces this overhead by automating credential handling, providing scoped tokens, and enforcing policies consistently across all objects.

3. Implement access control strategies

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/3-implement-access-control-strategies>

Implement access control strategies

Completed

Unity Catalog uses a privilege model to control access to securable objects such as catalogs, schemas, tables, and views. Grants define *who* (the principal, such as a group or user) can perform *what action* (a privilege like SELECT, MODIFY, or CREATE) on *which object* (such as a table). Privileges can be assigned explicitly at each level of containment or inherited from higher levels like schemas or catalogs. Understanding these patterns is key to designing secure and maintainable permission structures.

On Azure, a Unity Catalog **metastore** is a top-level container for all securable objects and is tightly integrated with Azure Databricks workspaces:

- **Region Bound:** A metastore is created in a specific Azure region and can only be assigned to workspaces in the same region.

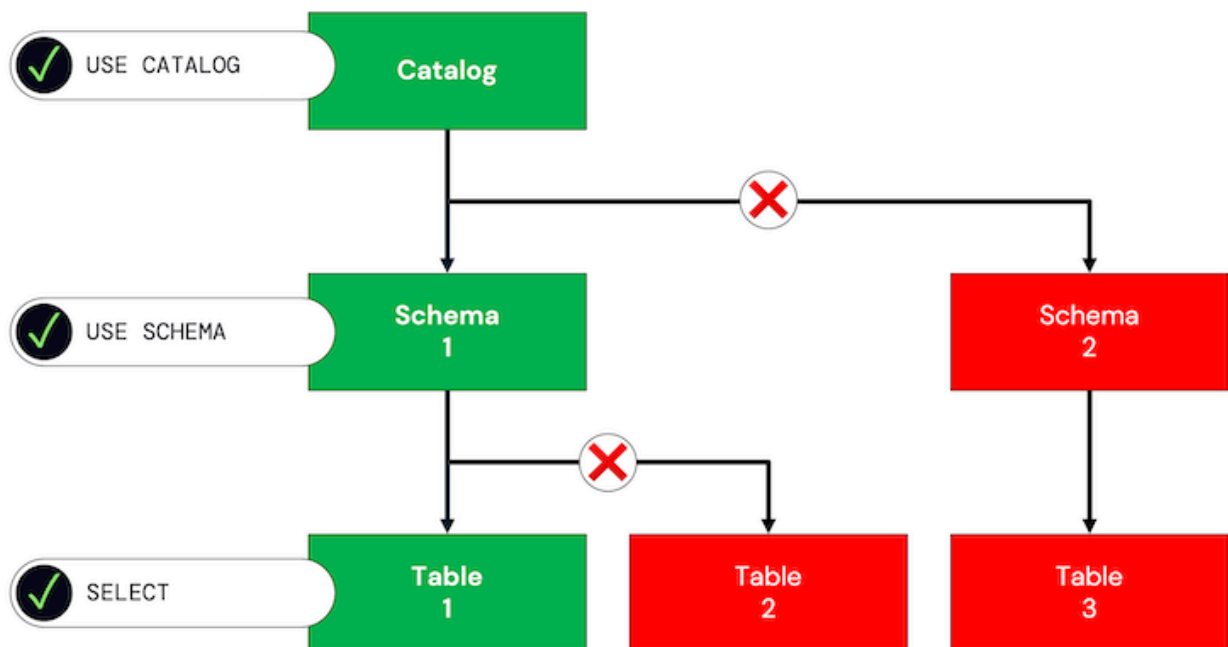
- **One per Workspace:** Each Azure Databricks workspace can be linked to only one metastore at a time.

Because of this region binding, organizations often create multiple metastores (one per region) if they operate across different Azure geographies. Data governance policies should account for this boundary.

Querying a table with explicit permissions at each level

When a user or group needs to query a table, it isn't enough to grant SELECT directly on the table. Unity Catalog requires permissions at all levels of containment. That means the grantee needs:

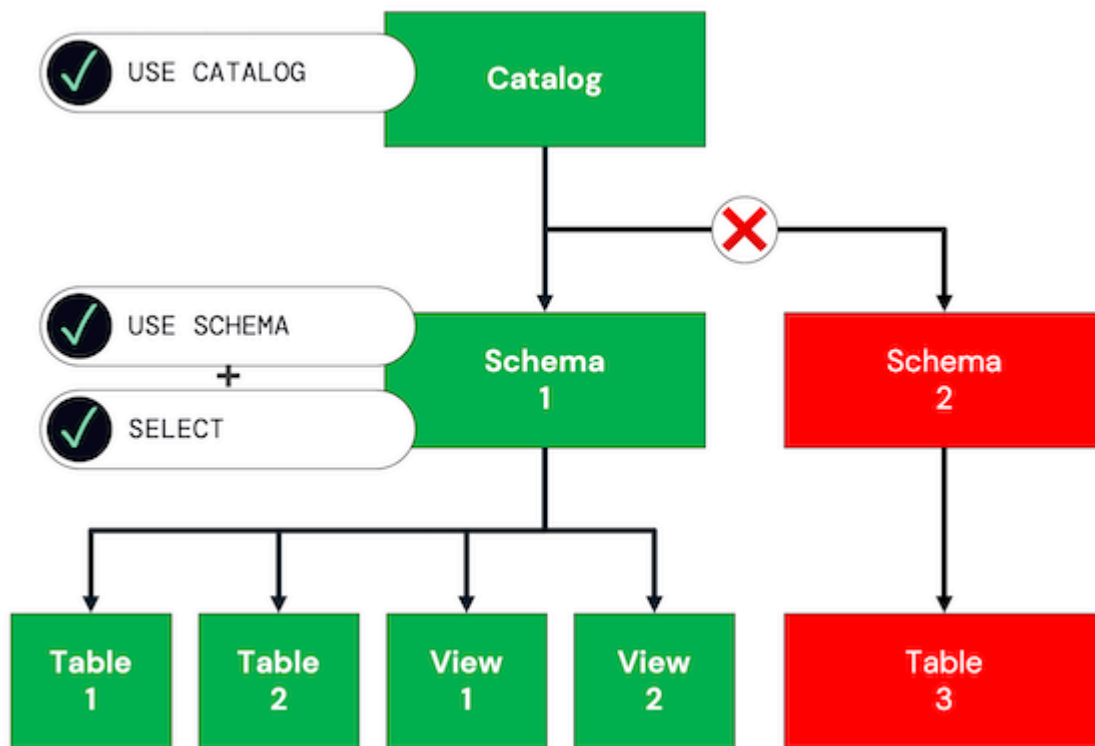
1. **USE CATALOG** on the catalog that contains the schema,
2. **USE SCHEMA** on the schema that contains the table, and
3. **SELECT** on the table itself.



If any of these privileges are missing, the user can't successfully query the table. This approach is called as the *explicit privilege model* because permissions are granted individually at each level. It's inherently secure because access to other objects is never granted unintentionally, but it requires more administrative effort. For example, if there are thousands of tables in a schema, each table must be explicitly granted, which is time-consuming to maintain.

Querying a table with inherited permissions from schema

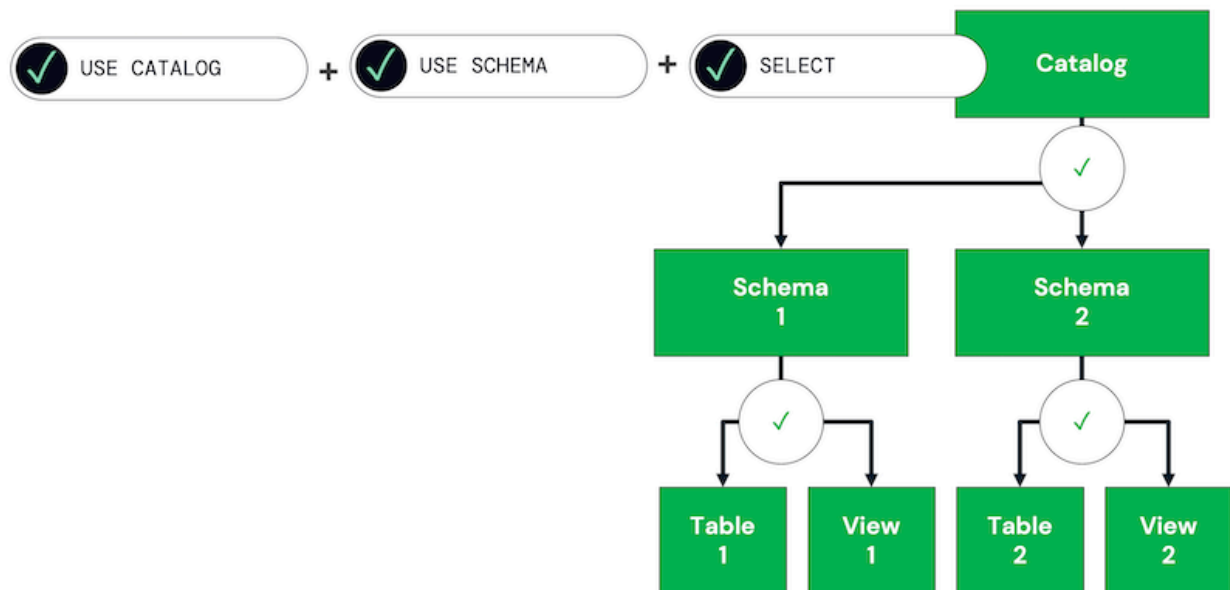
Privilege inheritance reduces administrative overhead by granting permissions at a higher level. Instead of assigning SELECT on each table individually, an administrator can grant **SELECT on the schema**. Combined with **USE SCHEMA** and **USE CATALOG** privileges, this automatically gives the grantee SELECT on all tables and views within that schema, including any created in the future.



This approach simplifies management, especially when large numbers of tables exist. However, it's less restrictive than the explicit model because all current and future objects within the schema become accessible. Access is still limited to that schema, though—other schemas remain inaccessible unless granted separately.

Querying a table with inherited permissions from catalog

Inheritance can also extend from the catalog level. By granting **USE CATALOG** and **ALL PRIVILEGES** (or **SELECT**) on the catalog, the grantee automatically inherits the necessary permissions on all schemas, tables, and views within that catalog. This includes any schemas or objects created in the future.

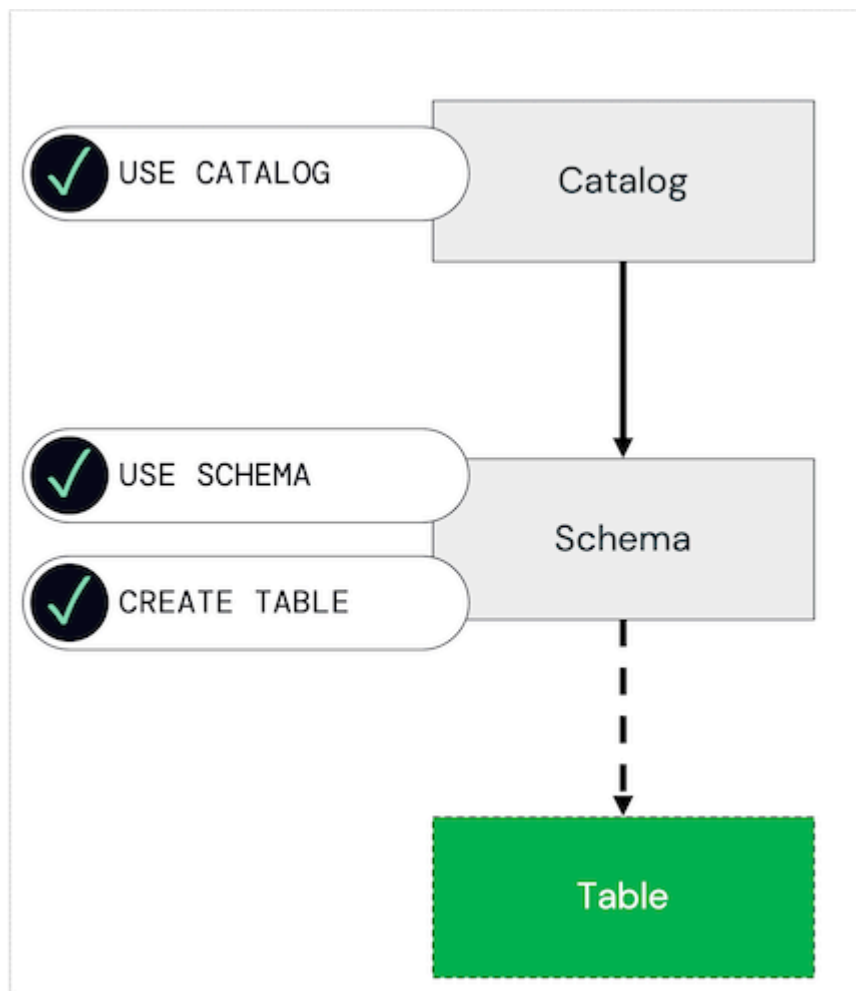


This model is the most permissive because it opens access across the entire catalog. It's convenient when broad access is desired, such as giving a group of analysts the ability to query everything within a specific catalog. However, administrators need to carefully assess whether all objects in the catalog should be exposed in this way, since there's no granularity at the schema or table level when inheriting from the catalog.

Creating a table: explicit at schema

Creating new tables requires more than SELECT permissions. At a minimum, the grantee must have:

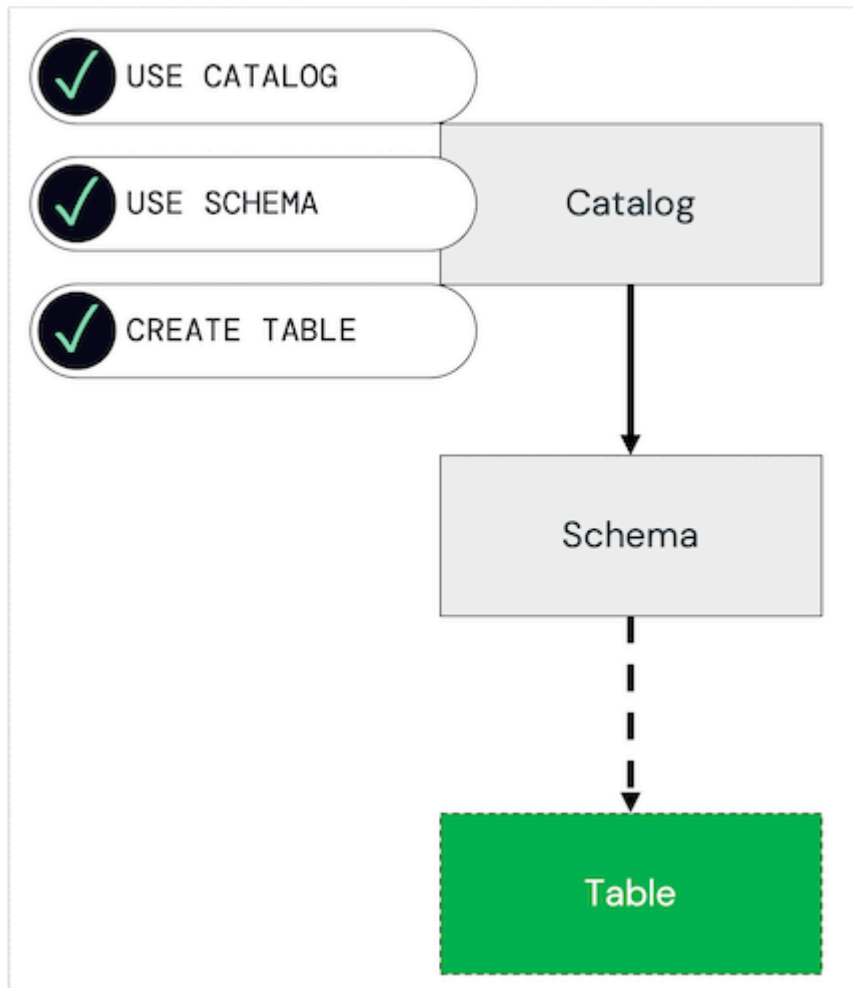
- **USE CATALOG** on the containing catalog,
- **USE SCHEMA** on the target schema, and
- **CREATE TABLE** on that schema.



With these permissions, the grantee can create new tables in that specific schema. This explicit setup restricts table creation to one schema and avoids granting unnecessary privileges elsewhere. It's a common choice when only certain schemas should allow new object creation.

Creating a table: inherited from catalog

In more permissive environments, table creation rights can be inherited from the catalog level. By granting **USE CATALOG**, **USE SCHEMA**, and **CREATE TABLE** at the catalog, the grantee automatically has the ability to create tables in any schema within that catalog, including schemas that may be added later.



This approach is easier to maintain but comes with greater risk, since it enables table creation everywhere within the catalog. It's best suited for catalogs designed for collaborative or exploratory work, where unrestricted creation is acceptable.

Granting permissions with ANSI SQL

Unity Catalog follows the ANSI SQL Data Control Language (DCL) standard for granting and revoking privileges. This means you can manage permissions with familiar statements like **GRANT** and **REVOKE**. For example, if you want to allow the group *analysts* to query a specific table, you can issue:

```
GRANT SELECT ON TABLE sales TO `analysts`;
```

If you want that same group to traverse the containing schema and catalog, you would also need:

```
GRANT USE SCHEMA ON SCHEMA finance TO `analysts`;  
GRANT USE CATALOG ON CATALOG corpdata TO `analysts`;
```

To permit table creation within a schema, you can grant:

```
GRANT CREATE TABLE ON SCHEMA finance TO `analysts`;
```

Permissions can be revoked in a similar way:

```
REVOKE SELECT ON TABLE sales FROM `analysts`;
```

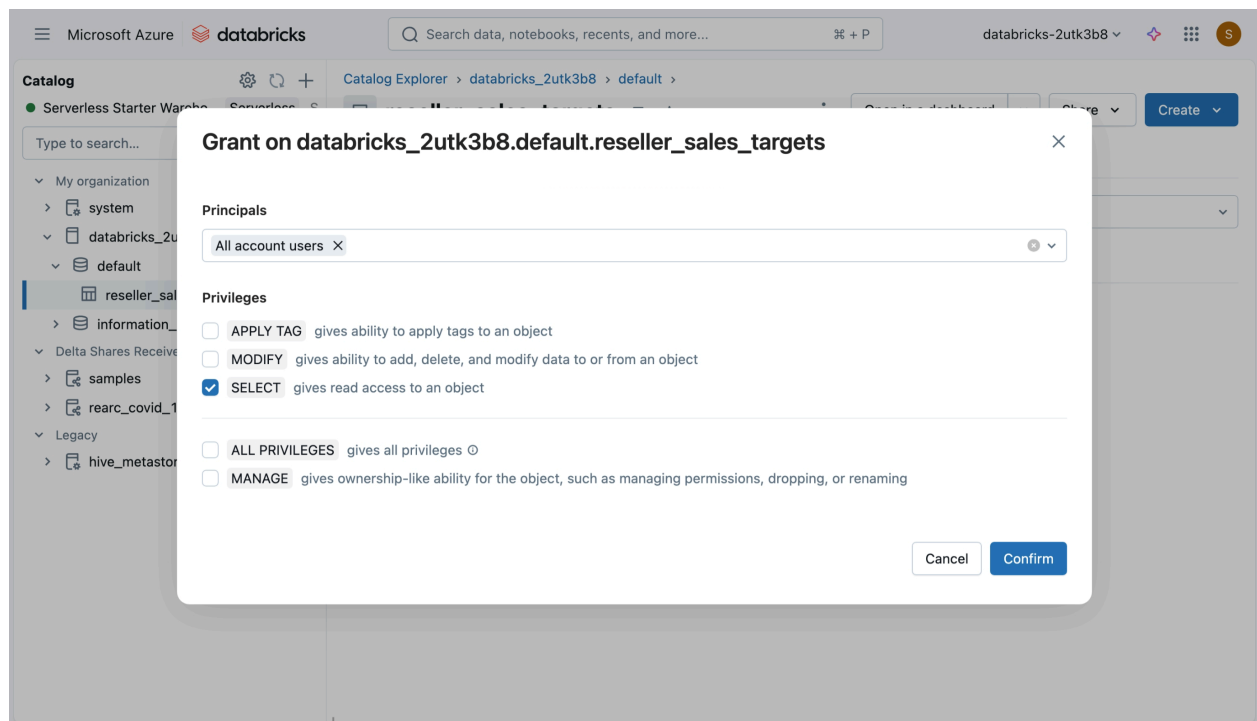
These SQL statements can be run from a Databricks notebook, the SQL editor in the workspace, or in Databricks SQL.

Granting permissions with the Azure Databricks user interface

For those people who prefer a graphical approach, Unity Catalog permissions can also be managed directly in the Azure Databricks UI. In **Catalog Explorer** (previously referred to as Data Explorer), you can navigate through catalogs, schemas, and tables. Each object has a **Permissions** tab where you can add users, groups, or service principals as principals, and assign privileges like SELECT, USE SCHEMA, or ALL PRIVILEGES.

For example, to grant SELECT on a table through the UI:

1. Open **Catalog Explorer**.
2. Navigate to the catalog and schema that contain the target table.
3. Select the table, then go to the **Permissions** tab.
4. Add the group (for example, *analysts*) and assign the **SELECT** privilege.



You can repeat the same process at the schema or catalog level to assign broader permissions. The UI also clearly shows inherited privileges, making it easier to audit access across levels.

Verify granted privileges

After granting privileges, you should verify that the correct permissions are in place. Unity Catalog provides the **SHOW GRANTS** command to display all privileges on an object or for a specific principal.

To see all grants on a table, use the SHOW GRANTS command with the object type and name:

```
-- Show all privileges granted on a table
SHOW GRANTS ON TABLE sales.reporting.monthly_revenue;
```

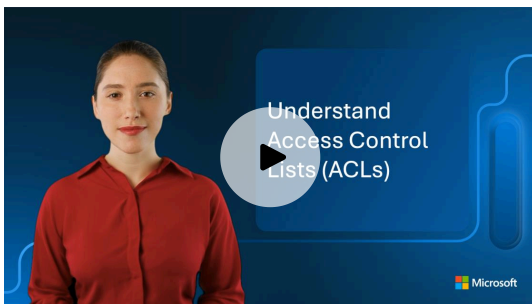
This command returns a list showing each principal, the privileges they hold, and the object those privileges apply to. You can also check grants on schemas, catalogs, or other securable objects using the same pattern.

If you want to see privileges for a specific principal, include the principal name in the command:

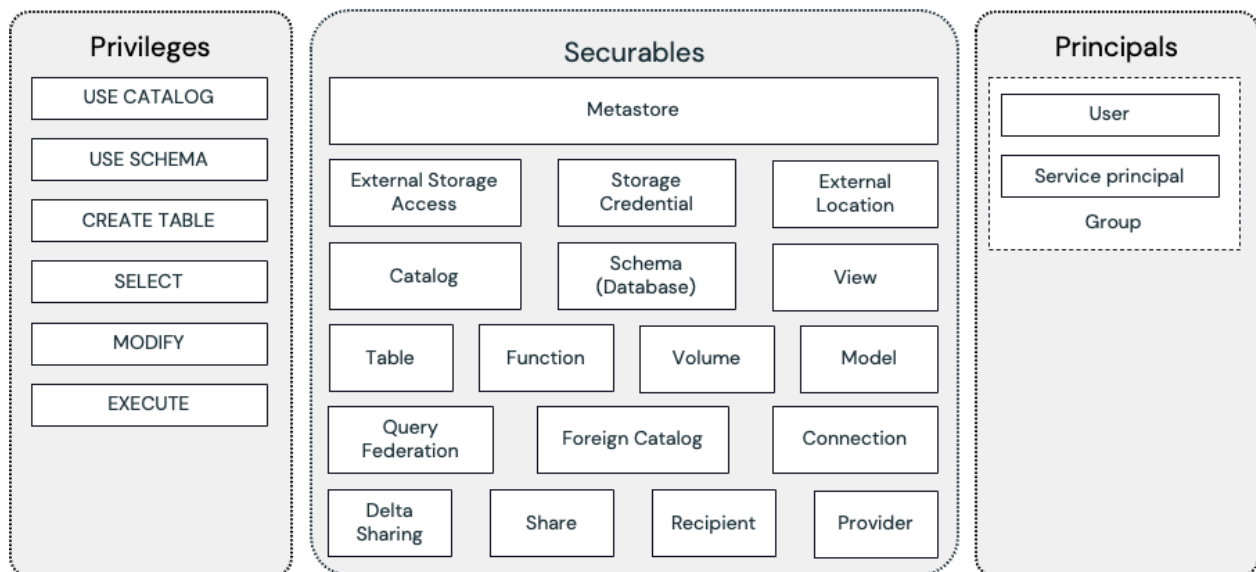
```
-- Show privileges for a specific group
SHOW GRANTS 'finance-team' ON TABLE sales.reporting.monthly_revenue;
```

When reviewing privileges, remember that **inherited grants** might not appear in the output for the specific table. If a user has SELECT granted at the schema or catalog level, that privilege applies to the table through inheritance, but SHOW GRANTS on the table itself might not display it. To get a complete picture, check grants at all levels of the hierarchy.

Understand Access Control Lists (ACLs)



Access Control Lists (ACLs) are the backbone of Unity Catalog's authorization model: they explicitly associate a *principal* with a *privilege* on a *securable object*. To understand how ACLs work, we can break down those three pieces—privileges, securables, principals—and then see how they interact in a typical grant.



Privileges

Privileges are the specific operations or actions that can be performed on securable objects. In Unity Catalog, different kinds of objects support different privilege sets. For example, for a **table** the allowed privileges include `SELECT`, `MODIFY`, `MANAGE`, `ALL PRIVILEGES` (which expands to all applicable privileges) among others. For a **schema**, privileges like `USE SCHEMA`, `CREATE TABLE`, `ALL PRIVILEGES`, and `MANAGE` apply. At the **catalog** level, you see privileges such as `USE CATALOG`, `CREATE SCHEMA`, `BROWSE`, `ALL PRIVILEGES`, and more.

Because not every privilege makes sense for every object (for example, `MODIFY` on a view is meaningless), Unity Catalog limits which privileges are valid on which securable types. The `ALL PRIVILEGES` shorthand is powerful in that it gives "every applicable privilege" on an object, and as new privileges are added over time, the `ALL PRIVILEGES` grant automatically covers those new privileges as well.

Securables

Securables are the objects to which you can grant or revoke privileges. In Unity Catalog, securables are hierarchical: the metastore (or catalog) is at the top, followed by catalogs, then schemas, then tables/views, and other object types such as functions, volumes, external locations, etc.

Because of this hierarchy, privileges can be inherited (if permitted) from higher levels down to lower levels.

Principals

Principals are the "who" in an ACL—the identity to which privileges are granted. In Unity Catalog / Azure Databricks, a principal can be a **user**, a **service principal** (for programmatic identities), or a **group**. Groups are often preferred for manageability: you assign privileges to groups, and then users inherit those privileges by membership, rather than granting to each individual.

The syntax for naming principals in SQL uses backticks when special characters exist (for example, an "@" in a user name or dashes in a service principal ID). For instance:

```
GRANT SELECT ON TABLE t TO `alice@company.com`;
```

Or

```
GRANT SELECT ON TABLE t TO `aaaaaaaa-bbbb-cccc-1111-222222222222`; -- service principal
```

Also worth noting: `account users` is a built-in principal representing all users of the account; you can grant privileges to `account users` (though not to the `users` workspace-local group) in Unity Catalog.

When you put these pieces together, the ACL is simply a statement like: *grant this privilege to that principal on this securable object*. At runtime, when a user tries to perform an operation (say select from a table), the system checks the ACLs along the hierarchy (catalog, schema, table) and evaluates whether the user (via group membership, explicit grants, or inherited grants) has the required combination of permissions to succeed.

Microsoft Entra ID integration

On Azure Databricks, Unity Catalog identities are managed through **Microsoft Entra ID**. This means that all users, service principals, and groups referenced in grants must exist in Microsoft Entra ID. Key points include:

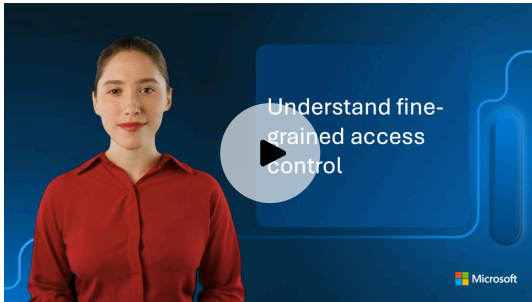
- **Users and Groups:** Unity Catalog does not support workspace-local users or groups. You must provision and manage them in Microsoft Entra ID, and Unity Catalog syncs them into the Azure Databricks account.
- **Service Principals:** Programmatic identities are registered as Microsoft Entra applications and can be granted privileges in Unity Catalog. Credentials are typically stored in Azure Key Vault or accessed through managed identities.
- **Account Users:** The special built-in principal `account users` represents all Microsoft Entra ID users in the Azure Databricks account, and can be used when broad access is acceptable.

Because Unity Catalog relies on Microsoft Entra ID, administrators should coordinate closely with their Microsoft Entra ID administrators to ensure that group memberships and role assignments remain accurate.

4. Understand fine-grained access control

Understand fine-grained access control

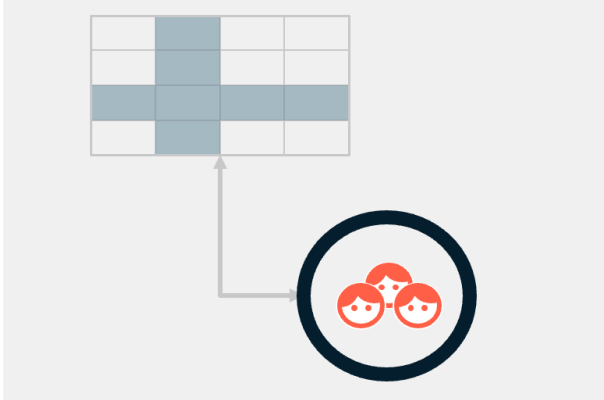
Completed



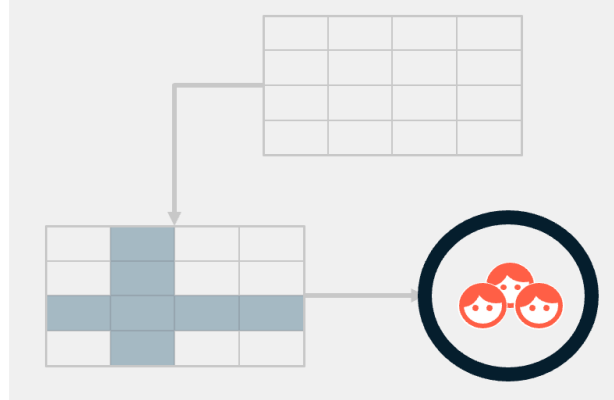
In the previous module, you looked at granting or denying access to entire tables. That works when every authorized user can see every row and column, but in real scenarios you often need to expose only part of a dataset. You might need to hide rows that a user isn't entitled to view, redact or mask specific columns that contain sensitive attributes, or partially transform values so that they remain useful without leaking full detail. Rather than duplicating tables or maintaining multiple extract copies, you can apply fine-grained access control. In Databricks Unity Catalog, you implement fine-grained control using one of two approaches:

- **Row and Column Security:** the preferred, modern method, composed of column masking and row filtering.
- **Dynamic Views:** the earlier pattern that still has valid use cases.

Row and Column Security



Dynamic View



In this unit, you'll learn what each approach is, when to use it, and why one is generally favored today. You'll explore the specific SQL syntax in the next unit, so here we stay at the conceptual level.

Why fine-grained control matters

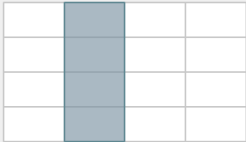
You frequently have a single canonical table that must serve different audiences: application users, data analysts, compliance reviewers, or support staff. Creating separate physical tables for each audience leads to data duplication, stale copies, extra orchestration, and security drift. Views can help, but managing many views—each encoding slightly different masking or filtering—becomes difficult over time. Fine-grained access control lets you keep one authoritative table while tailoring what each principal actually sees at query time.

Typical objectives you'll address:

- **Column masking:** Show a column's presence but hide (or transform) its sensitive values for most users.
- **Row filtering:** Suppress entire records that a user or group shouldn't see.
- **Partial transformation:** Preserve analytical utility (for example, last two digits, domain of an email) without exposing full identifiers.

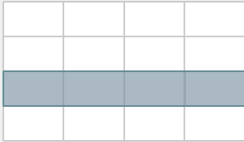
Limit Access to Columns

Omit column values from output



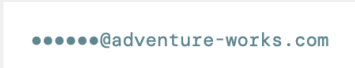
Limit Access to Rows

Omit rows from output



Data Masking

Obscure data



Row and column security

Row and Column Security is implemented through two core concepts:

- **Column masking:** You associate a masking function with a specific column. When a user queries the table, Unity Catalog invokes that function to decide whether to return the real value, a redacted placeholder, or a transformed version. You typically define one function per masked column.
- **Row filtering:** You associate a filter function with the table. That function returns a boolean determining whether each row is visible to the querying principal. One function can govern all row-level rules for that table.

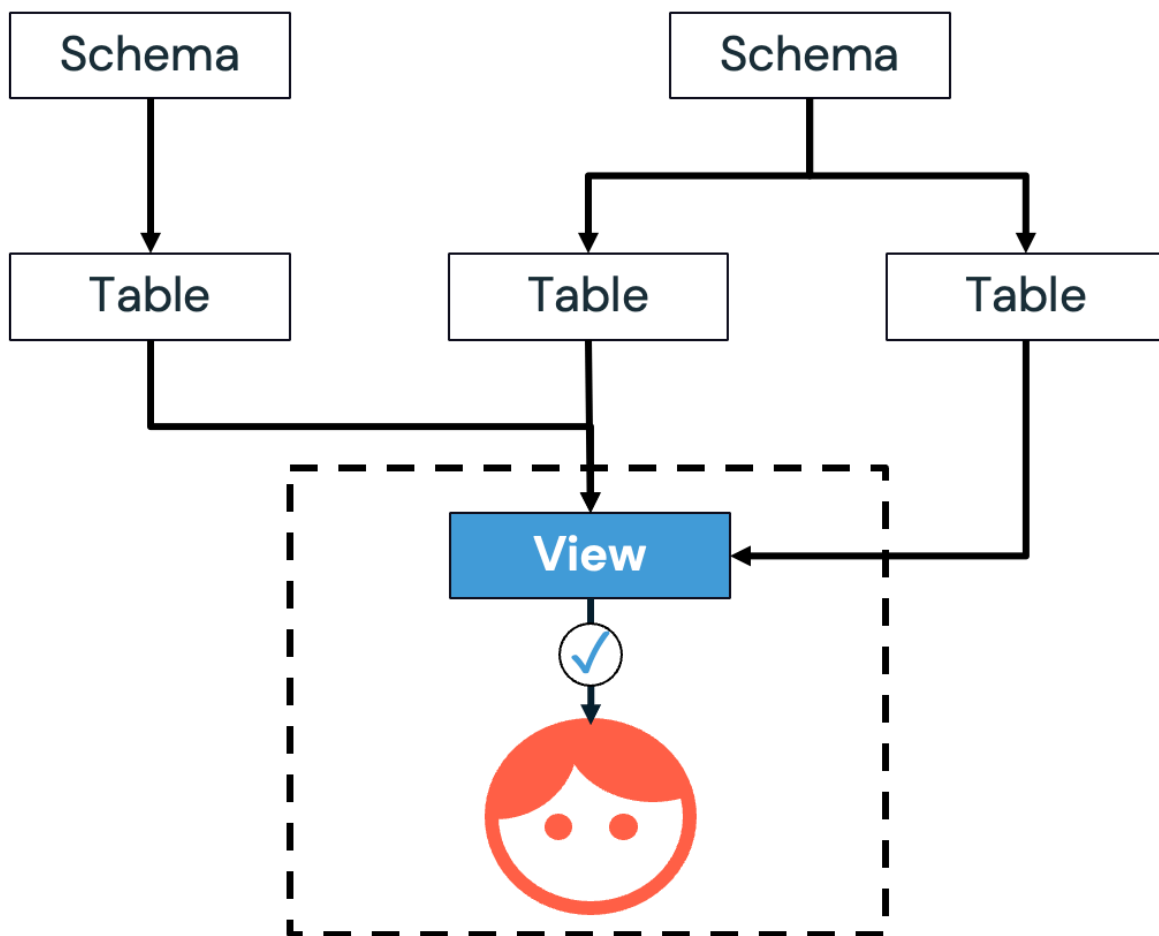
You create user-defined SQL functions that contain the decision logic (for example, checking the querying user's group membership). Because these functions are securable objects just like tables, you also decide where they live (schema/catalog) and who can alter or execute them. You then bind them directly to the table's metadata. At that point, anyone querying the table benefits from a single authoritative definition of masking and filtering—no extra layer required.

Key characteristics you should internalize:

- **Centralized logic:** Rules live alongside the table definition, reducing drift between data and policy.
- **Combinatorial flexibility:** A single set of masking and filtering functions handles all user combinations; you don't create separate artifacts per audience.
- **Performance alignment:** Operating at the table layer avoids the overhead of stacking multiple view abstractions.
- **Operational simplicity:** Fewer named objects to permission, audit, and lifecycle.

Dynamic views

Dynamic Views predate Row and Column Security. These have been with Azure Databricks for some time, and though they're no longer the preferred method for controlling access to rows and columns, they still have their place. You define a view that selects from one or more underlying tables and encodes conditional logic: CASE expressions to mask sensitive columns, predicates to exclude rows, or transformations that partially obfuscate values. Any user with permission on the view can query it without having direct access to the underlying tables, assuming the view owner has that access. This separation lets you shield source objects while presenting a curated projection.



You might create multiple dynamic views over the same table to serve different user cohorts (for example, a support view with partially masked identifiers and an analyst view with additional fields). Each view becomes its own securable object with its own ownership and grants.

Characteristics to note:

- **Level of separation:** Users can be isolated from source tables entirely.
- **Multi-table composition:** You can assemble data from several tables and apply masking/filtering in one place.
- **Read-only exposure:** Views can't themselves be updated; any data modification still requires access to the base tables.
- **Object proliferation risk:** Each audience-specific variant typically requires a new view.

Why row and column security is preferred

You should default to Row and Column Security because:

- **Reduced object sprawl:** One table plus a small set of functions replaces a potentially expanding family of views.
- **Easier maintenance:** Policy changes occur where the data resides, lowering chances of unsynchronized logic across parallel views.
- **Performance advantages:** Enforcement happens at the source, avoiding extra resolution layers and enabling better optimization.
- **Clearer governance:** Auditing and lineage focus on the primary table rather than a web of derivative objects.
- **Flexible combinations:** You avoid creating a new physical or logical artifact for every unique mix of masking and filtering rules.

When dynamic views still make sense

You should consider dynamic views instead (or in addition) when:

- Source tables are intentionally treated as **read-only interfaces**; you want an explicit contract boundary between producers and consumers.
- You need to present a logical dataset assembled from **multiple tables** with consistent masking and filtering logic applied across joins.
- **Organizational policy** requires strict separation such that consumers shouldn't even know the exact structure or existence of underlying tables.
- **Existing legacy patterns** already use dynamic views and a phased migration strategy is necessary.

Comparing row and column security with dynamic views

Here's your comparison table with **Row and Column Security** vs **Dynamic Views** side by side, so you can see the trade-offs and decide which approach best fits your needs.

Dimension	Row and Column Security	Dynamic Views
Maintainability	Minimizes moving parts by enforcing security at the table level.	More objects to manage if multiple views are created.
Flexibility	Limited to what's stored in the table.	Can combine multiple tables directly for flexible compositions.
Object governance	Fewer objects overall; simplifies grants and audits.	More objects to govern; grants may need to be managed at both view and table level.
Read/write patterns	Supports seamless querying and updating with masking or filtering applied at the table.	Typically read-only; write patterns require direct table access.
Abstraction boundary	Exposes storage design more directly.	Provides a presentation layer decoupled from underlying storage.

Typical workflow

Follow these steps to design the right approach and keep your model both secure and maintainable:

1. Identify sensitive columns and row-level segmentation rules (for example, region-based, role-based).
2. Decide if you truly need multi-table assembly or strict abstraction. If not, prefer Row and Column Security.
3. For Row and Column Security: design minimal user-defined functions—one per masked column and optionally one row-filter function—that evaluate the querying principal.
4. Attach these functions to the table so that queries automatically enforce the logic.
5. For Dynamic Views (when justified): craft a single view that applies CASE expressions and WHERE predicates; create additional views only when audiences require materially different exposures.
6. Periodically review whether evolving requirements still justify each dynamic view or whether consolidation into Row and Column Security is now feasible.

Common pitfalls

A few common pitfalls can undermine the effectiveness of a security model. One of the most frequent mistakes is **duplicating tables** for each audience. This approach quickly creates synchronization challenges and complicates governance, as any schema or data change must be applied consistently across all copies. Another recurring issue is the creation of **multiple, nearly identical views** instead of implementing generalizable masking logic. This not only increases maintenance overhead but also makes the system harder to audit and reason about. It's also important to be deliberate about where business logic resides.

Embedding complex transformations inside masking functions can blur responsibilities; these rules typically belong in upstream transformation pipelines, not in security enforcement. Equally critical is

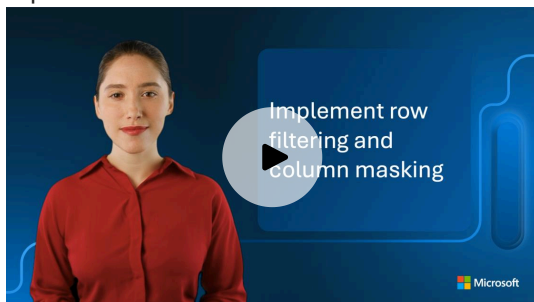
ensuring that the **masking and filtering functions themselves are properly secured**. If they're altered or bypassed, the entire protection model can be weakened. Finally, organizations should be cautious of **over-masking**, which removes more information than necessary and can inadvertently block legitimate analytical tasks that depend on those fields.

5. Implement row filtering and column masking

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/5-implement-row-filtering-column-masking>

Implement row filtering and column masking

Completed



When working with sensitive information in a data lakehouse, you need more than just role-based permissions. Unity Catalog allows you to enforce **fine-grained access control** so that different users can query the same tables but see only the data they're allowed to. Two approaches are commonly used, as explained in the previous unit:

- **Row and Column Security** – restricts which rows and columns a user can see at the table level.
- **Dynamic Views** – encapsulates filtering and masking logic in a view definition that adapts automatically to user context.

Both strategies depend on group membership checks, such as `is_account_group_member()`, to determine whether a user should see protected values.

Row and column security

Row and column security lets you define rules directly on a table. Two operations are available: **column masking** and **row filtering**.

Column masking

Column masking ensures that sensitive values are hidden unless a user belongs to an authorized group. For example, the `c_phone` column in a **customers** table can be masked with a function:

```
CREATE OR REPLACE FUNCTION phone_mask(c_phone STRING)
  RETURN CASE WHEN is_account_group_member('metastore_admins')
    THEN c_phone
    ELSE 'REDACTED PHONE NUMBER'
  END;
```

The masking function is then applied to the column:

```
ALTER TABLE customers
  ALTER COLUMN c_phone
  SET MASK phone_mask;
```

From this point forward, queries against the table return actual phone numbers only to `metastore_admins`. Other users always see `REDACTED PHONE NUMBER`.

Row filtering

Row filtering removes rows from query results for users who aren't authorized. You define a function that returns `true` only when a row should be visible. For example:

```
CREATE OR REPLACE FUNCTION nation_filter(c_nationkey INT)
  RETURN IF(is_account_group_member('admin'), true, c_nationkey = 21);
```

When applied as a row filter:

```
ALTER TABLE customers
  SET ROW FILTER nation_filter ON (c_nationkey);
```

This ensures that non-admins only see rows where `c_nationkey = 21`. Admin users continue to see all rows.

Together, column masking and row filtering provide a way to enforce restrictions directly on the base table.

Dynamic views

An alternative to modifying a table is to create a **dynamic view**. Views allow you to express row and column restrictions in SQL, and every user querying the view automatically sees the appropriate subset of the data.

Consider a **customers_new** table with sensitive information. A dynamic view can redact phone numbers and limit rows in one definition:

```
CREATE OR REPLACE VIEW vw_customers AS
SELECT
  c_custkey,
  c_name,
  c_address,
  c_nationkey,
  CASE
    WHEN is_account_group_member('admins') THEN c_phone
    ELSE 'REDACTED PHONE NUMBER'
  END as c_phone,
  c_acctbal,
  c_mktsegment,
  c_comment
FROM customers_new
WHERE
  CASE WHEN
    is_account_group_member('admins') THEN TRUE
    ELSE c_nationkey = 21
  END;
```

When users query this view, the rules apply automatically:

- Phone numbers are shown only for members of the `admins` group.
- Non-admins see only rows with `c_nationkey = 21`.

Access to the view can then be granted with standard SQL permissions:

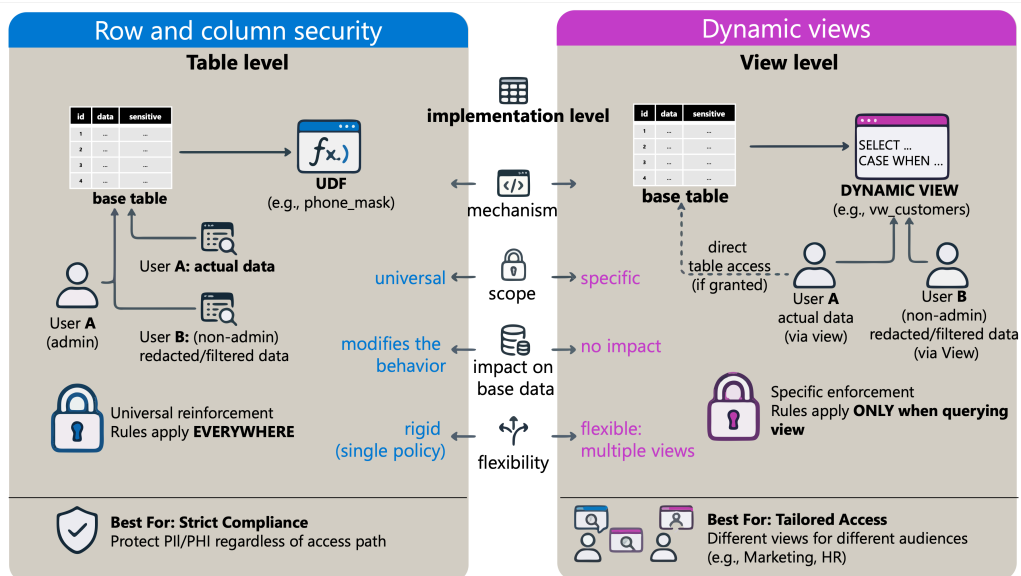
```
GRANT SELECT ON VIEW vw_customers TO `account users`;
```

The underlying table remains untouched, while the view enforces security logic.

Choosing between the two approaches

Both approaches achieve fine-grained security, but they differ in scope:

- **Row and Column Security** is embedded at the table level and guarantees that restrictions apply everywhere the table is used.
- **Dynamic Views** provide flexibility: you can create multiple views with different rules for different audiences, while keeping the base table unrestricted.



In practice, you use table-level controls when you need strict enforcement, and dynamic views when you want adaptable, shareable abstractions.

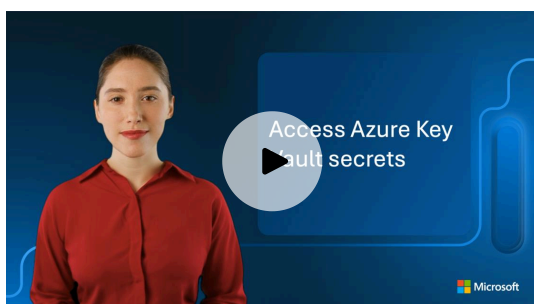
6. Access Azure Key Vault secrets

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/6-access-azure-key-vault>

Access Azure Key Vault secrets

Completed

When you work with Azure Databricks, you often need to connect to external data sources, APIs, or services that require authentication credentials. Storing passwords, connection strings, or API keys directly in notebooks creates security risks and makes credential management difficult. Azure Key Vault integration solves this challenge by providing a secure, centralized way to manage secrets in your Databricks workflows.

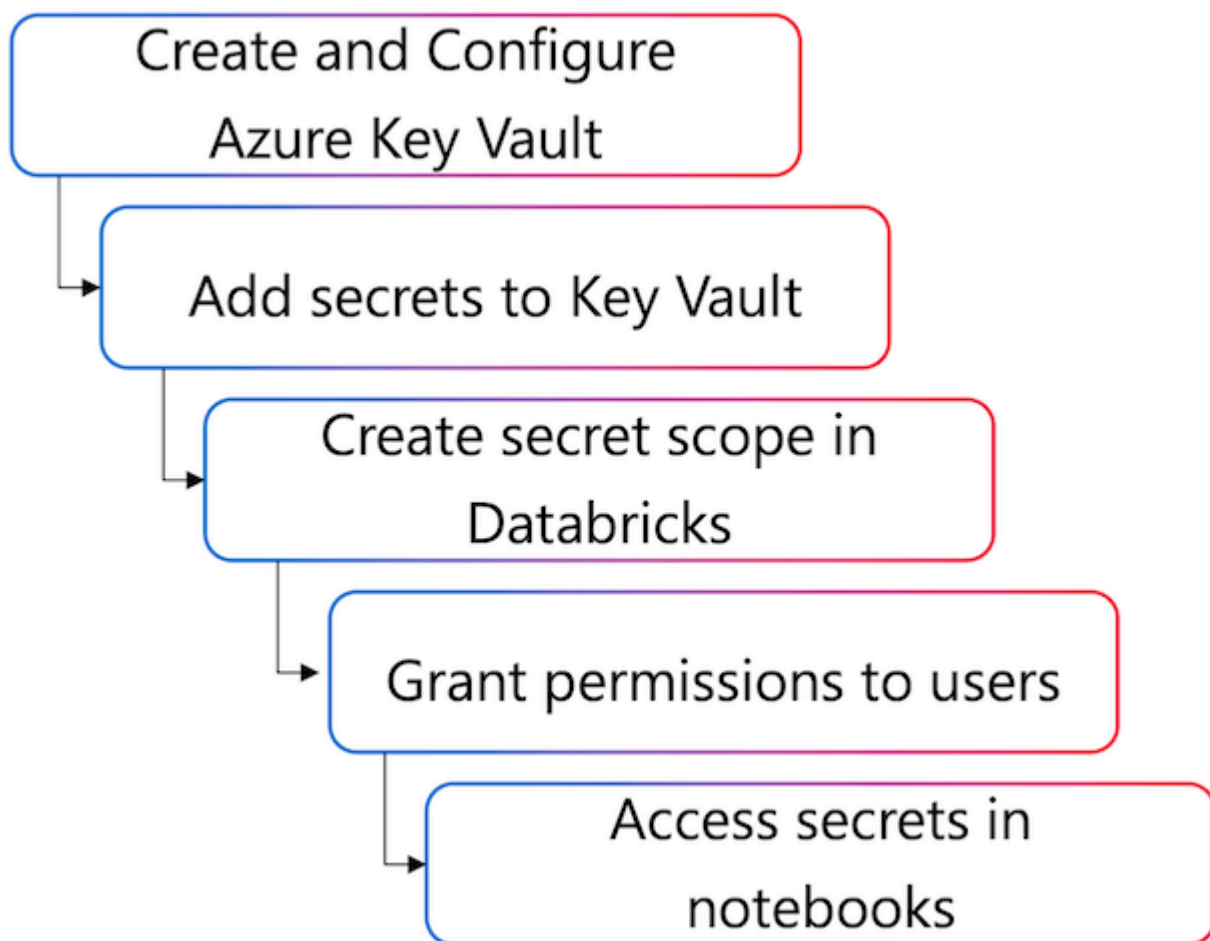


As a data engineer working in Azure, you typically use Azure Key Vault rather than Databricks-backed secret scopes. This approach aligns with enterprise security practices and provides better integration with other Azure services. You can reference secrets stored in Azure Key Vault from your notebooks, jobs, and cluster configurations without exposing sensitive values.

Understanding secret scopes in Azure Databricks

Azure Databricks uses secret scopes to organize and control access to secrets. A secret scope is a named collection of secrets that you reference in your code. Each scope can contain multiple secrets, and you assign permissions at the scope level.

The following diagram shows the overall workflow for using Azure Key Vault-backed secret scopes in Databricks:



Azure Databricks supports two types of secret scopes:

- **Databricks-backed scopes** store secrets in an encrypted database managed by Azure Databricks. You create and manage these secrets using the Databricks CLI or API.
- **Azure Key Vault-backed scopes** provide a read-only interface to secrets stored in an Azure Key Vault instance. You manage the actual secrets in Azure Key Vault, while the scope in Azure

Databricks provides controlled access.

For production workloads, Azure Key Vault-backed scopes offer several advantages. They integrate with your organization's existing Azure security infrastructure. They support network restrictions that Databricks-backed scopes cannot. They eliminate the need to synchronize secrets across multiple systems. When you update a secret in Key Vault, all Azure Databricks workspaces that reference that scope automatically use the updated value after you restart your clusters.

Important

Permissions on Azure Key Vault-backed scopes apply to all secrets in the linked Key Vault instance. If you need to restrict access to specific secrets, create separate Key Vault instances and link them to different secret scopes.

Creating an Azure Key Vault-backed secret scope

Before you create an Azure Key Vault-backed scope, you need an Azure Key Vault instance configured correctly. Your Key Vault must use the **Vault access policy** permission model rather than Azure role-based access control. Under the Key Vault's networking settings, you must allow trusted Microsoft services to bypass the firewall, even if you restrict access to specific virtual networks.

You create the secret scope using a special URL in your Azure Databricks workspace. Navigate to `https://<databricks-instance>#secrets/createScope`, replacing `<databricks-instance>` with your workspace URL. This URL is case-sensitive—the **S** in `createScope` must be uppercase.

When you create the scope, you provide a descriptive name and specify who can manage it. The **Manage Principal** setting determines whether only the creator or all workspace users have **MANAGE** permission on the scope. **MANAGE** permission allows reading secrets, writing new secrets, and controlling who else can access the scope. Restricting management to the creator requires the Premium plan.



HomePage / Create Secret Scope

Create Secret Scope

CancelCreate

A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)

Scope Name

Manage Principal

Creator



Azure Key Vault

DNS Name

https://xxx.vault.azure.net/

Resource ID

/subscriptions/xxxxxx/...

You also need two pieces of information from your Azure Key Vault: the **DNS Name** (for example, `https://databrickskv.vault.azure.net/`) and the **Resource ID** (for example, `/subscriptions/xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx/resourcegroups/databricks-rg/providers/Microsoft.KeyVault/vaults/databrickskv`). You can find both values in the **Properties** section of your Key Vault in the Azure portal.

When you create the scope, Databricks automatically grants the Azure Databricks service principal the **Get** and **List** permissions on your Key Vault. This allows the service to retrieve secret values on behalf of authorized users. Your account must have the Key Vault Contributor, Contributor, or Owner role to complete this configuration.

Note

After creating the scope, you manage the actual secrets in Azure Key Vault, not in Databricks. You use the Azure portal or Azure CLI to add, update, or delete secrets in the Key Vault instance.

Retrieving secrets in notebooks and jobs

Once you create a secret scope, you access secrets in your notebooks using the `dbutils.secrets.get()` function. This function takes two parameters: the scope name and the secret key name. The function returns the secret value, which Azure Databricks automatically redacts in notebook outputs.

```
username = dbutils.secrets.get(scope="jdbc", key="username")
password = dbutils.secrets.get(scope="jdbc", key="password")

df = (spark.read
      .format("jdbc")
      .option("url", "jdbc:sqlserver://server.database.windows.net:1433")
      .option("dbtable", "customers")
      .option("user", username)
      .option("password", password)
      .load()
    )
```

In this example, you retrieve database credentials from a secret scope named `jdbc`. The `username` and `password` variables contain the actual secret values, but if you print these variables or display them in notebook output, Azure Databricks replaces the values with `[REDACTED]`.

This redaction provides protection against accidental exposure, but it doesn't prevent intentional access by authorized users. Any user with permission to run code on the cluster can access secret values programmatically. You control access to secrets by managing permissions on the secret scope and the cluster.

You can also list the secrets available in a scope to discover what keys exist:

```
dbutils.secrets.list("jdbc")
```

This returns metadata about the secrets, including their key names, but not the actual values. This helps you verify that the secrets you need exist in the scope before you reference them.

Tip

Name your secrets descriptively so that their purpose is clear when you list them. For example, use `database-username` instead of `user1`, or `storage-account-key` instead of `key`.

Using the `secret()` function in SQL

In addition to `dbutils.secrets.get()`, you can retrieve secrets directly in SQL queries using the `secret()` function. This function is available in Databricks SQL and Spark SQL, making it convenient when working with SQL notebooks or queries.

```
SELECT secret('jdbc', 'username') AS db_username;
```

You can also use the `secret()` function in SQL statements that configure connections or perform operations:

```
CREATE TABLE customer_data
USING jdbc
OPTIONS (
  url "jdbc:sqlserver://server.database.windows.net:1433",
  dbtable "customers",
  user secret('jdbc', 'username'),
  password secret('jdbc', 'password')
);
```

The `secret()` function provides the same redaction behavior as `dbutils.secrets.get()` —secret values are replaced with `[REDACTED]` in query outputs. For cases where you want to return `NULL` instead of raising an error when a secret doesn't exist, use the `try_secret()` function.

Using secrets in Spark configurations

Sometimes you need to reference secrets in cluster configurations rather than in notebook code. This is useful when you want to set Spark properties or environment variables that multiple notebooks on the cluster can access.

You reference a secret in a **Spark configuration property** using double curly braces:

```
spark.storage.key {{secrets/storage/account-key}}
```

This syntax tells Databricks to retrieve the secret named `account-key` from the scope named `storage` and assign it to the `spark.storage.key` configuration property. You configure this in the **Advanced Options** section when you create or edit a cluster.

In your notebook, you retrieve the value using:

```
storage_key = spark.conf.get("spark.storage.key")
```

Reference secrets in environment variables

You can also reference secrets in environment variables. This is particularly useful for init scripts that run before notebooks execute:

```
STORAGE_KEY={{secrets/storage/account-key}}
```

In an init script, you access this environment variable like any other:

```
if [ -n "$STORAGE_KEY" ]; then
    echo "Configuring storage with key"
fi
```

Keep in mind that any user with CAN ATTACH TO permissions on the cluster can read environment variables and Spark configuration properties. This means they can access secrets even if they don't have direct permission to read from the secret scope. Use this approach only when the secrets need to be available to all cluster users.

When you update a secret in Azure Key Vault, you must restart your cluster for Azure Databricks to fetch the new value. The cluster caches secret values when it starts, and it doesn't automatically detect changes in Key Vault.

Managing secret scope permissions

Access to secrets is controlled through permissions on the secret scope. When you create a scope, you receive MANAGE permission by default. This allows you to read secrets, add new secrets (in Databricks-backed scopes), and control who else can access the scope.

You assign permissions using the Databricks CLI. Three permission levels exist: **READ** allows retrieving secret values, **WRITE** allows adding or updating secrets (only for Databricks-backed scopes, not Key Vault-backed scopes), and **MANAGE** allows reading secrets and controlling permissions.

Important

For Azure Key Vault-backed scopes, you can only manage who has access to the scope through Databricks ACLs. You cannot add or update secrets through Databricks—secrets must be managed directly in Azure Key Vault. The WRITE permission is not applicable to Key Vault-backed scopes.

To grant the READ permission to a group:

```
databricks secrets put-acl jdbc data-engineers READ
```

This command grants the `data-engineers` group permission to retrieve secrets from the `jdbc` scope. You can grant permissions to individual users by using their email addresses or to service principals using their application IDs.

You can view all permissions on a scope:

```
databricks secrets list-acls jdbc
```

This shows which principals have access and their permission levels. You can also check the permissions for a specific principal:

```
databricks secrets get-acl jdbc data-engineers
```

To remove a principal's access to a scope:

```
databricks secrets delete-acl jdbc data-engineers
```

Note

The ability to restrict MANAGE permission to only the creator (rather than all workspace users) requires the Premium plan. On the Standard plan, you must grant MANAGE permission to all workspace users when creating an Azure Key Vault-backed secret scope.

With these capabilities, you can securely access Azure Key Vault secrets in your Databricks workflows while maintaining control over who can use those secrets. This integration strengthens your security posture and aligns with enterprise governance requirements.

7. Authenticate data access with service principals

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/7-authenticate-data-access-service-principals>

Authenticate data access with service principals

Completed

As a data engineer working with Azure Databricks, you often need to access data stored in Azure Data Lake Storage or other cloud storage services. Service principals provide a secure, automated way to authenticate these data access requests without using personal credentials.

Understanding service principals for data access

A service principal is an identity created for applications and services to access Azure resources. Think of it as an automated user account specifically designed for applications rather than humans.

When your Databricks notebooks or jobs need to read data from Azure storage, they can authenticate using a service principal. This approach offers several advantages over using personal credentials. The authentication happens automatically without manual login, works consistently across different users and environments, and allows precise control over which storage locations the principal can access.

A service principal consists of three key components you'll work with:

- **Application (client) ID:** A unique identifier for the service principal
- **Directory (tenant) ID:** The Microsoft Entra ID tenant where the service principal exists
- **Client secret:** A password-like credential that proves the identity of the service principal

Note

While service principals remain a valid authentication method, Microsoft recommends using **managed identities** for Unity Catalog storage credentials when possible. Managed identities eliminate the need to manage and rotate secrets and can access storage accounts protected by network rules. However, service principals are still useful in scenarios where managed identities aren't available or when you need more explicit control over credentials.

We will cover managed identities in the next unit.

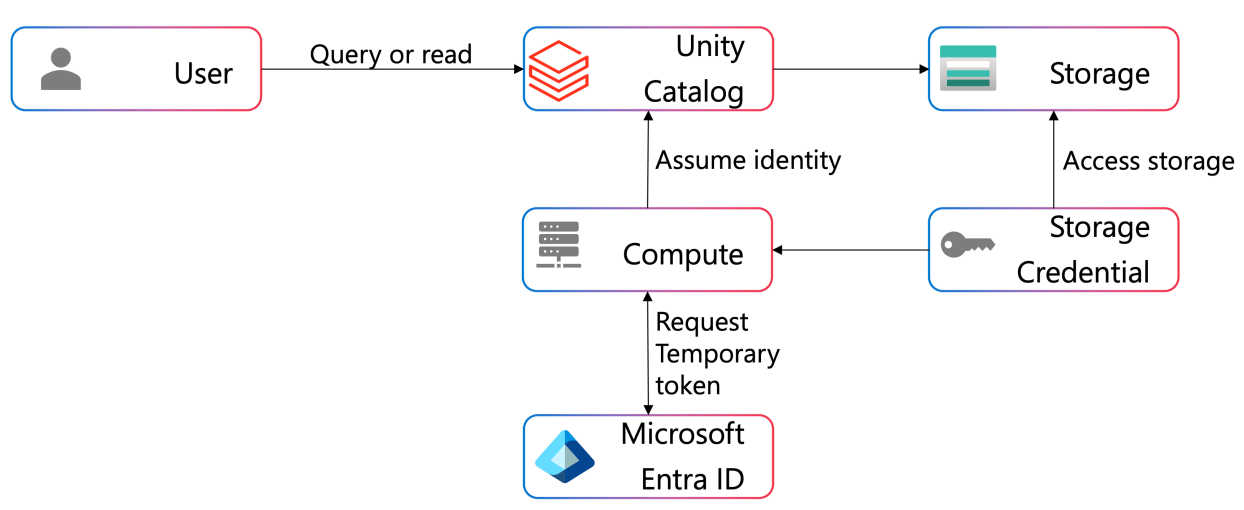
Using service principals with Unity Catalog

Unity Catalog provides a centralized way to manage data access using service principals through storage credentials and external locations.

A **storage credential** in Unity Catalog encapsulates a service principal's authentication details. Instead of embedding credentials directly in your code, you reference the storage credential by name. Unity Catalog handles the authentication behind the scenes, generating temporary tokens that allow your queries to access the underlying storage.

Here's how the workflow operates in practice:

When you query a table or read from an external location governed by Unity Catalog, the system checks which storage credential is associated with that location. Unity Catalog assumes the service principal identity, requests a temporary access token from Microsoft Entra ID, and provides that token to your compute resources. Your cluster or SQL warehouse then uses the token to access the storage directly.



This approach offers important benefits. Your sensitive credentials never appear in notebook code or job configurations. Unity Catalog audit logs record which principal accessed which data and when. You can grant and revoke access to storage locations through Unity Catalog permissions without changing any code.

Important

As a data engineer, you typically won't create the storage credential itself—that's an administrative task. You'll work with existing storage credentials that have been configured by your workspace administrators or metastore admins. Your role focuses on using these credentials through external locations and tables.

To access data through a Unity Catalog-managed external location, you simply query the table or reference the external location path. Unity Catalog handles the authentication automatically:

```
-- Query a table in an external location secured by a service principal
SELECT * FROM catalog_name.schema_name.external_table;
```

No explicit credential configuration is needed in your code. The table's metadata includes the external location reference, which points to a storage credential containing the service principal details.

Using service principals in notebooks

You might encounter scenarios where you need to configure service principal authentication directly in your notebook code. This pattern is common when working with storage that isn't registered in Unity Catalog or when using legacy configurations.

To authenticate to Azure Data Lake Storage using a service principal, you configure Spark session properties. The service principal's client secret should always come from a secure secret store, never hardcoded in your notebook.

Here's how you retrieve a secret and configure Spark to use service principal authentication:

```
# Retrieve the client secret from Databricks secrets
service_credential = dbutils.secrets.get(scope="storage-secrets", key="sp-client-secret")

# Configure Spark to authenticate using the service principal
spark.conf.set("fs.azure.account.auth.type.<storage-account>.dfs.core.windows.net",
               "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth.provider.type.<storage-account>.dfs.core.windows.net",
               "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.<storage-account>.dfs.core.windows.net",
               "<application-id>")
spark.conf.set("fs.azure.account.oauth2.client.secret.<storage-account>.dfs.core.windows.net",
               service_credential)
spark.conf.set("fs.azure.account.oauth2.client.endpoint.<storage-account>.dfs.core.windows.net",
               "https://login.microsoftonline.com/<directory-id>/oauth2/token")
```

With these configurations in place, you can read data using the standard ABFS (Azure Blob File System) protocol:

```
# Read data from ADLS Gen2 using the service principal for authentication
df = spark.read.parquet("abfss://<container>@<storage-account>.dfs.core.windows.net")
```

The authentication happens transparently. Spark uses the OAuth 2.0 client credentials flow to obtain an access token from Microsoft Entra ID, then presents that token when accessing the storage account.

Security considerations and best practices

When working with service principals for data access, several security practices help protect your credentials and data.

Never hardcode secrets in notebooks or code. Client secrets are sensitive credentials that grant access to your data. Always retrieve them from a secure secret store like Azure Key Vault or Databricks secrets. This practice prevents accidental exposure if you share notebooks or commit code to version control.

Understand that secrets require rotation. Unlike managed identities, service principal client secrets have expiration dates and must be rotated periodically. As a data engineer, you should be aware that if a secret expires, your data access will fail. Work with your administrators to understand the rotation schedule and know who to contact if authentication suddenly stops working.

Apply the principle of least privilege. Each service principal should have access only to the specific storage locations and data it needs. If a service principal is compromised, limited permissions reduce the potential impact. When requesting access, specify exactly which storage accounts and containers you need rather than requesting broad permissions.

As you move forward in your work with Unity Catalog, you'll discover that managed identities offer a more streamlined approach to authenticating data access. In the next unit, we'll explore how

managed identities eliminate many of the challenges associated with service principals while providing enhanced security and simplified operations.

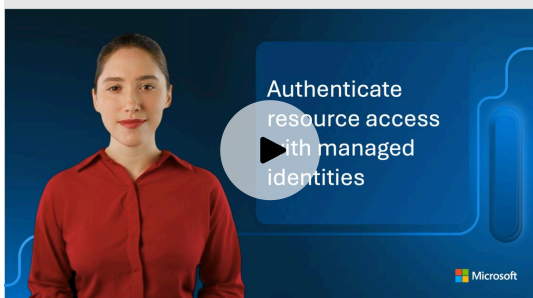
8. Authenticate resource access with managed identities

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/8-authenticate-resource-access-managed-identities>

Authenticate resource access with managed identities

Completed

Managed identities provide a secure, automated way to authenticate your Unity Catalog workspaces to storage resources without managing credentials in your code or configuration files.



As a data engineer, you use managed identities to create **storage credentials** in Unity Catalog. These credentials form the foundation for accessing external data through external tables or reading files directly from cloud storage. Unlike traditional authentication methods that require you to create, store, and rotate secrets, managed identities eliminate credential management overhead while providing enterprise-grade security.

Understanding managed identities for Unity Catalog

Managed identities are Azure resources that provide an identity for your applications when connecting to services that support Microsoft Entra ID authentication. Azure manages the lifecycle of these identities automatically, removing the burden of credential rotation and secret management from your team.

With Unity Catalog, you use managed identities through an **Azure Databricks access connector**. This first-party Azure resource acts as a bridge between your managed identity and your Azure Databricks account. The access connector can contain either a **system-assigned managed identity**

that Azure creates automatically, or one or more **user-assigned managed identities** that you create and manage separately.

Two primary use cases exist for managed identities in Unity Catalog. First, you use them to connect to the metastore's root storage account where Unity Catalog stores managed tables. Second, you use them to access external storage accounts for reading files or creating external tables. Both scenarios use the same configuration approach but serve different purposes in your data architecture.

Note

System-assigned managed identities are tied to the lifecycle of the access connector. If you delete the access connector, Azure automatically deletes the system-assigned managed identity. User-assigned managed identities persist independently and can be attached to multiple access connectors.

Creating storage credentials with managed identities

Storage credentials in Unity Catalog encapsulate the authentication information needed to access cloud storage. When you create a storage credential using a managed identity, you establish a secure connection pathway that Unity Catalog can use to access data on behalf of authorized users.

The process begins in the Azure portal where you create an access connector for Azure Databricks. You select the same Azure region as your storage account to minimize latency and avoid cross-region data transfer costs. During creation, you choose whether to enable a system-assigned managed identity or attach user-assigned managed identities. The resource receives a unique identifier in the format `/subscriptions/{subscription-id}/resourceGroups/{resource-group}/providers/Microsoft.Databricks/accessConnectors/{connector-name}`.

Microsoft Azure

[Home](#) > [Access Connector for Azure Databricks](#) >

Create an Access Connector for Azure Databricks

Basics

Tags

Managed Identity

Review + create

System assigned managed identity

Enable system assigned identity to grant the resource access to other existing resources.

Status ⓘ

☐ Off☒ On

User assigned managed identity

Add user assigned identities to grant the resource access to other existing resources.

+

Add

🗑

Remove

Name	↑↓	Resource group	↑↓	Subscription	↑↓
No user assigned managed identities assigned to this resource. Select 'Add' to add more.					

After creating the access connector, you grant the managed identity permissions on your Azure Data Lake Storage account. The **Storage Blob Data Contributor** role provides read and write access to blob data, which is typically what you need for Unity Catalog operations. You assign this role at either the storage account level for broad access or at the container level for more granular control.

For optimal performance, you should also grant the **Storage Queue Data Contributor** role to enable file event notifications. This configuration allows Azure Databricks to subscribe to change notifications from your storage account, making file processing operations more efficient. When files are added or modified in your storage containers, Unity Catalog receives immediate notifications rather than scanning for changes.

With the Azure configuration complete, you create the storage credential in Unity Catalog. In Catalog Explorer, you navigate to the Credentials tab and create a new storage credential. You select **Azure Managed Identity** as the credential type and provide the access connector's resource ID. If you're using a user-assigned managed identity, you also include its resource ID. The storage credential receives a unique name that you reference when creating external locations.

Create a new credential ✕

A storage credential represents an authentication and authorization mechanism for accessing data stored on your cloud tenant. [Learn more](#)

☒ Storage Credential ☐ Service Credential ⓘ

Credential Type*

Azure Managed Identity

Credential name*

my-credential

Access connector ID [Learn more](#) *

/subscriptions/aaaa0a0a-bb1b-cc2c-dd3d-eeeeee4e4e4e/resourceGroups/databricks-rg-da1

User assigned managed identity ID (optional)

Comment

> Advanced Options

Cancel

Create

Important

To create storage credentials, you need the **CREATE STORAGE CREDENTIAL** privilege on the Unity Catalog metastore. Account admins and metastore admins have this privilege by default. Without this privilege, you can't create new storage credentials but can still use existing ones that you have permissions to access.

Accessing external storage through storage credentials

Storage credentials work together with **external locations** to provide governed access to your cloud storage. An external location combines a storage credential with a specific cloud storage path, creating a managed access point that you can grant permissions on.

When you create an external location, you specify a path in your Azure Data Lake Storage account using the `abfss://` protocol. For example, `abfss://data@mystorageaccount.dfs.core.windows.net/raw/` points to the raw folder in the data container. You associate this path with your storage credential, which provides the authentication mechanism. Users who have privileges on the external location can read from or write to that path, depending on their specific permissions.

This architecture provides several benefits for data engineering teams. You can create multiple external locations that use the same storage credential but point to different paths, allowing you to organize access to different datasets or environments. You can grant different sets of users access to different external locations, even when they all use the same underlying storage account. You can mark storage credentials as read-only, which prevents creation of external locations that allow write access.

External locations also support **workspace binding**, which restricts access to specific workspaces in your metastore. This feature is valuable when you use workspaces to separate production and development environments. An external location pointing to production data can be bound to only the production workspace, ensuring that development workspaces can't accidentally access or modify production data. You can also bind storage credentials themselves to specific workspaces, controlling which workspaces can use those credentials to create new external locations.

```
# Create an external table using a managed identity-backed external location
CREATE EXTERNAL TABLE sales_data
USING DELTA
LOCATION 'abfss://data@mystorageaccount.dfs.core.windows.net/sales/';

# Query the external table
SELECT * FROM sales_data WHERE sale_date >= '2024-01-01';
```

When your Unity Catalog workspace accesses data through a storage credential, the managed identity authenticates to Azure Data Lake Storage. The workspace doesn't need to know the actual credentials—it simply presents the managed identity. Azure verifies the identity and checks whether the assigned roles allow the requested operation. This happens transparently in the background, requiring no additional code or configuration in your notebooks.

Tip

When you work with network-restricted storage accounts, ensure your workspace's virtual network can reach the storage account. Managed identities handle authentication, but network connectivity must already be established through private endpoints or service endpoints.

9. Exercise - Secure Unity Catalog Objects

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/9-exercise-secure-unity-catalog-objects>

Exercise - Secure Unity Catalog Objects

Completed

Now it's your chance to secure Unity Catalog objects. In this lab, you secure Unity Catalog objects in Azure Databricks by granting fine-grained access control to a Databricks group, applying row filters to restrict customer data by region, and masking PII email addresses using column mask functions. You also create an Azure Key Vault-backed secret scope and retrieve secrets securely inside a notebook, so sensitive credentials are never exposed in code.

Note

To complete this lab, you need an [Azure subscription](#) in which you have administrative access.

Launch the exercise and follow the instructions.

[Launch Exercise](#)

10. Module assessment

<https://learn.microsoft.com/en-us/training/modules/secure-unity-catalog-objects/10-knowledge-check>

Module assessment

Completed

11. Summary

Summary

Completed

Security in **Unity Catalog** extends beyond simple permission checks to encompass the entire lifecycle of data access. Throughout this module, you explored how Unity Catalog enforces security at multiple layers—from validating permissions and issuing **scoped tokens** during query execution, to controlling table and schema access through **inherited and explicit grants**, to protecting sensitive data through **fine-grained row and column filtering**.

You learned practical approaches for managing credentials and authentication. **Azure Key Vault** integration lets you retrieve secrets securely without exposing sensitive values in your code. **Service principals** provide programmatic access to storage resources with explicit credential control. **Managed identities** eliminate credential management overhead entirely while enabling access to network-protected storage accounts. Each authentication method serves specific scenarios, and choosing the right approach depends on your security requirements and operational constraints.

The access control strategies you encountered—from hierarchical permissions that simplify administration to fine-grained masking functions that protect individual columns—form a comprehensive security framework. **Row and Column Security** offers centralized enforcement with minimal object sprawl, while **dynamic views** provide flexibility when you need to compose data from multiple sources. Understanding when to apply each pattern helps you build solutions that balance governance with usability.

As you implement these security practices in your own environments, start with clear requirements for who should access what data. Design your permission model to scale with your organization, using groups and inherited permissions where appropriate. Test your row and column filtering logic thoroughly to ensure users see exactly what they should. Regularly review your storage credentials and secret scope permissions to maintain a strong security posture. Unity Catalog provides the tools—your role is to apply them thoughtfully to protect your data assets while enabling productive analytics workflows.

Govern Unity Catalog objects

1. Introduction

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/1-introduction>

Introduction

Completed



Effective data governance requires more than just controlling access—it demands visibility into your data assets, accountability for how they're used, and enforcement mechanisms that scale with your organization. As data platforms grow in complexity, managing table definitions, tracking data lineage, enforcing retention policies, and sharing data securely become increasingly challenging. **Unity Catalog** in Azure Databricks provides the foundation for comprehensive governance that addresses these challenges through centralized metadata management and policy enforcement.

When you govern Unity Catalog objects, you work across several interconnected capabilities. You document tables and columns with **comments and tags** that help data consumers discover and understand your assets. You implement **attribute-based access control (ABAC)** using governed tags and policies that automatically enforce fine-grained permissions. You configure **data retention policies** using Delta Lake's `VACUUM` and predictive optimization to manage storage costs and meet compliance requirements.

Beyond access control, governance extends to **data lineage tracking** that shows how data flows through your pipelines, enabling impact analysis and troubleshooting. **Audit logging** captures who did what and when, supporting security investigations and regulatory compliance. For external collaboration, **Delta Sharing** lets you share data with partners and customers while maintaining governance controls over what they can access.

This module guides you through implementing these governance capabilities in your Azure Databricks environment. You explore techniques for preserving metadata, enforcing policies at scale, managing data lifecycle, and sharing data securely. By mastering these concepts, you position yourself to build data platforms that balance accessibility with control, enabling your organization to derive value from data while meeting its governance obligations.

2. Create and preserve table definitions

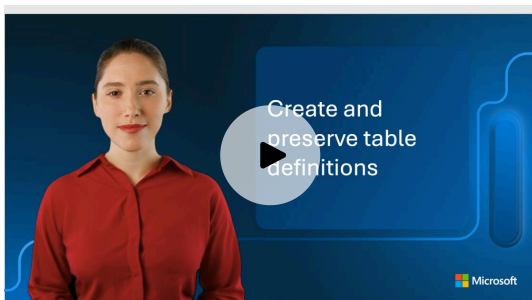
<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/2-preserve-table-column-definitions>

Create and preserve table definitions

Completed

Well-documented table and column definitions are the foundation of effective data governance. When you create and preserve comprehensive **metadata** for your data assets, you enable data consumers to discover, understand, and trust the data they need. This becomes especially important as AI-driven data discovery tools rely on high-quality metadata to surface relevant information.

In this unit, you learn how to create tables with descriptive comments and tags, and use discovery tools to explore your data assets in Unity Catalog.



Add comments to tables and columns

Comments provide human-readable descriptions that help data consumers understand the purpose and content of your data assets. You can add comments to any securable object in Unity Catalog, including catalogs, schemas, tables, views, and columns.

Consider the following example, where we add comments directly to columns as part of the `profiles` table definition:

```
CREATE TABLE sales.customers.profiles (
  customer_id  BIGINT COMMENT 'Unique identifier for each customer',
  email        STRING COMMENT 'Customer primary email address',
  created_date DATE   COMMENT 'Date when customer account was created',
  preferences  STRUCT<notifications: BOOLEAN, language: STRING>
                COMMENT 'Customer preference settings'
);
```

Add comments using SQL

Use the `COMMENT ON` statement to add or update comments on existing objects:

```
-- Add a comment to a table
COMMENT ON TABLE sales.customers.profiles
IS 'Contains customer profile information including contact details and preferences';

-- Add a comment to a column
ALTER TABLE sales.customers.profiles
ALTER COLUMN email COMMENT 'Primary email address used for account notifications';
```

You can also use markdown formatting in your comments to add structure:

```
COMMENT ON TABLE sales.orders.transactions IS
'## Order Transactions
Contains all customer orders with the following key fields:
- **order_id**: Unique order identifier
- **customer_id**: Foreign key to customers table
- **order_date**: Date order was placed';
```

Use AI-generated comments

Unity Catalog can automatically generate comments using AI-powered suggestions. This feature analyzes your table schema and column names to create initial descriptions. To generate AI comments in Catalog Explorer:

1. Open Catalog Explorer and select your table.
2. In the **About this table** panel, select **AI generate**.
3. Review the suggested comment and select **Accept** or **Edit** to modify it.

For column comments, select **AI generate** above the column list to generate suggestions for all columns.

Edit column comments



Column	Comment
customer_id int	A unique identifier assigned to each customer, allowing for easy tracking and analysis of individual customer data.
customer_name string	The full name of the customer, which can be used for personalized communication and marketing efforts.
region string	Indicates the geographical area where the customer is located, useful for regional analysis and targeted marketing strategies.
email_valid boolean	A flag indicating whether the customer's email address is valid, which is essential for effective communication and marketing campaigns.
total_orders bigint	Represents the total number of orders placed by the customer, providing insight into their purchasing behavior and loyalty.
total_spent double	The cumulative amount of money spent by the customer, which helps in assessing their overall value to the business.
avg_order_value double	Calculates the average value of orders placed by the customer, useful for understanding spending

8 comments unsaved

✓ Save all

Important

AI-generated comments are suggestions based on schema analysis. Always review these comments before saving, as AI models might generate inaccurate descriptions. Don't rely on AI comments for data classification tasks like detecting personally identifiable information (PII).

Required permissions for comments

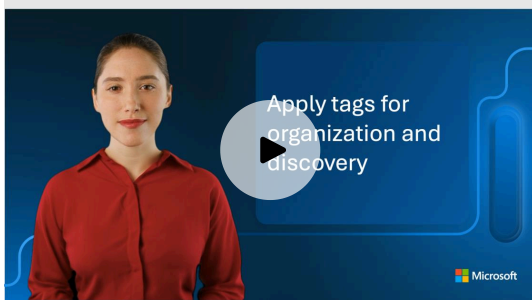
To add or edit comments, you need:

- Ownership of the object, or
- `MODIFY` and `SELECT` privileges on the table
- `USE CATALOG` and `USE SCHEMA` on the parent catalog and schema

After adding comments, any user with the `BROWSE` privilege can view them in Catalog Explorer.

Apply tags for organization and discovery

Tags are key-value pairs that help you organize and categorize data assets in Unity Catalog. Unlike comments, which provide descriptive text, tags enable structured classification and improve search functionality.



Add tags using SQL

Use the `ALTER` command with `SET TAGS` to apply tags:

```
-- Add tags to a table
ALTER TABLE sales.customers.profiles
SET TAGS ('domain' = 'sales', 'data_classification' = 'internal');

-- Add tags to a specific column
ALTER TABLE sales.customers.profiles
ALTER COLUMN email
SET TAGS ('pii' = 'email', 'sensitivity' = 'high');
```

To remove tags, use `UNSET TAGS`:

```
ALTER TABLE sales.customers.profiles
UNSET TAGS ('domain');
```

Use governed tags

Governed tags provide centralized control over tag definitions. With governed tags, administrators define allowed keys and values at the account level, ensuring consistent classification across the organization. Governed tags appear with a lock icon in Catalog Explorer.

System tags are a special type of governed tag predefined by Azure Databricks for common use cases like data classification, ownership tracking, and lifecycle management.

Tag constraints

Keep these constraints in mind when working with tags:

- Maximum 50 tags per securable object
- Maximum 1,000 column tags per table
- Tag keys have a 255-character limit
- Tag values have a 1,000-character limit
- Tag keys are case-sensitive

Explore and discover database objects

After you create and document your data assets, you need ways to find and explore them. Unity Catalog provides several tools for data discovery.

Use Catalog Explorer

Catalog Explorer provides a visual interface for navigating your data assets. When you select a table, you can:

- View the **Columns** tab to see column names, data types, and comments
- Select the **Sample Data** tab to preview table contents
- Check the **Details** tab for table properties and metadata
- Review the **History** tab for Delta table version history

The **Insights** tab shows frequent queries and users who accessed the table in the past 30 days, helping you understand how data is being used.

Explore objects using SQL

Use SQL commands to programmatically explore your catalog:

```
-- List all catalogs available to you
SHOW CATALOGS;

-- List schemas in a catalog
SHOW SCHEMAS IN sales;

-- List tables in a schema
SHOW TABLES IN sales.customers;

-- View table details including comments
DESCRIBE TABLE EXTENDED sales.customers.profiles;

-- View table columns
SHOW COLUMNS IN sales.customers.profiles;

-- View table properties
SHOW TBLPROPERTIES sales.customers.profiles;
```

You can filter results using pattern matching:

```
-- Find tables matching a pattern
SHOW TABLES IN sales.customers LIKE 'profiles_*';
```

Search using tags

Use the workspace search bar to find tables and views by their tags. Enter tag keys or values to locate matching objects. Only objects you have at least the **BROWSE** privilege on appear in search results.

View table relationships

For tables with defined foreign keys, select **View relationships** in Catalog Explorer to display an Entity Relationship Diagram (ERD). This diagram visualizes how tables connect through primary and foreign key relationships.

Integrate with AI/BI tools

AI-powered discovery tools like **AI/BI Genie** use your table comments and tags to provide natural language search and question answering. High-quality metadata directly improves the accuracy of AI-assisted data exploration.

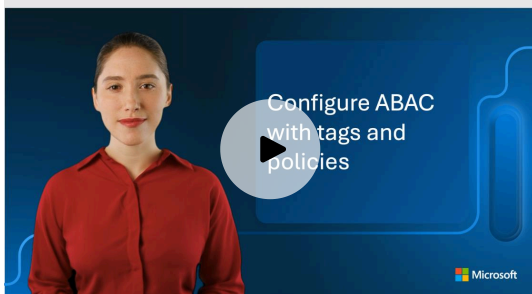
Maintaining comprehensive table and column definitions is an ongoing responsibility. As your data evolves, update your comments and tags to keep them accurate. This investment in metadata quality pays dividends through improved discoverability, governance compliance, and AI-readiness.

3. Configure ABAC with tags and policies

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/3-configure-attribute-based-access-control-tags-policies>

Configure ABAC with tags and policies

Completed



Attribute-based access control (ABAC) provides a flexible, scalable approach to data governance in Unity Catalog. Instead of managing individual permissions for every table and user combination, you define policies based on attributes. When you tag data assets with meaningful classifications, ABAC policies automatically enforce the right access controls.

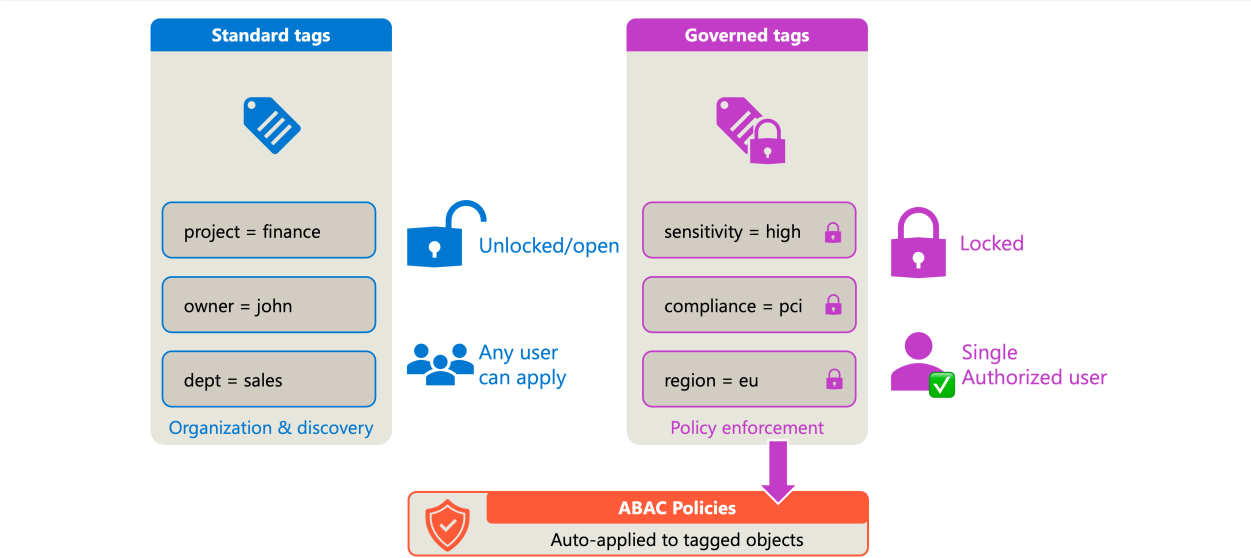
Important

ABAC in Unity Catalog is currently in **Public Preview**. Features may change before general availability.

In this unit, you learn how to use governed tags and ABAC policies to implement dynamic, centralized access control across your Unity Catalog environment.

Understand governed tags

Governed tags differ from the standard tags you've already encountered. While standard tags help with organization and discovery, governed tags add enforcement and consistency at the account level. Administrators define governed tags with specific allowed values, and only authorized users can assign them.



Consider the following differences between tag types:

Aspect	Standard tags	Governed tags
Scope	Any key-value pair	Predefined keys and allowed values
Control	Any user with <code>APPLY TAG</code> privilege	Only users with <code>ASSIGN</code> permission
Purpose	Organization and discovery	Policy enforcement and compliance
Visual indicator	None	Lock icon in Catalog Explorer

Governed tags serve as the foundation for ABAC policies. When you tag a table with `sensitivity=high`, you can then create a policy that masks certain columns for all tables with that tag. This approach scales efficiently because adding the tag to new tables automatically applies the policy.

Create governed tags

To create a governed tag, select **Catalog > Governance > Governed Tags**, then select **Create governed tag**. Enter a tag key, an optional description, and define the allowed values. Only these predefined values can be assigned to objects with this tag key.

For example, you might create a governed tag with key `pii` and allowed values `ssn`, `email`, and `address`. This ensures consistent classification of personally identifiable information across your organization.

Create governed tag



Creating a governed tag will define a tag policy with a set of allowed values and permissions for who can manage and apply the tag

* Tag key

Description

Allowed values

CancelCreate

Note

Tag data is stored as plain text. Don't use tag names or values that contain sensitive information.

Assign governed tags to objects

After creating a governed tag, apply it to columns, tables, schemas, or catalogs using the `SET TAGS` command:

```
-- Tag a column as containing SSN data
ALTER TABLE customers.profiles
ALTER COLUMN SSN
```

```
SET TAGS ('pii' = 'ssn');

-- Tag an entire table as containing sensitive data
ALTER TABLE customers.profiles
SET TAGS ('sensitivity' = 'high');
```

Governed tags applied to catalogs or schemas automatically inherit to all objects within them. This inheritance simplifies governance across large data estates.

Note

Tag inheritance applies only at the catalog and schema levels. Tags on individual table columns aren't inherited.

Create row filter policies

Row filter policies control which rows users can see based on governed tags. You define a user-defined function (UDF) that returns `TRUE` for rows the user should see and `FALSE` for rows to hide.

Before creating a policy, you need a UDF that implements your filtering logic. The following example creates a function that filters out European addresses:

```
CREATE OR REPLACE FUNCTION filter_non_eu(address STRING)
RETURNS BOOLEAN
RETURN NOT (
    LOWER(address) LIKE '%eu%' OR
    LOWER(address) LIKE '%europe%'
);
```

Now create a row filter policy that uses this function:

```
CREATE POLICY hide_eu_customers
ON CATALOG sales
ROW FILTER filter_non_eu
TO 'analysts'
FOR TABLES
MATCH COLUMNS
    hasTagValue('pii', 'address') AS addr
USING COLUMNS (addr);
```

This policy applies the `filter_non_eu` function to any table in the `sales` catalog that has a column tagged with `pii=address`. The `analysts` group only sees rows where the function returns `TRUE`.

You can also create policies using Catalog Explorer. Select the catalog or schema, choose the **Policies** tab, and select **New policy**. The visual interface guides you through selecting principals, scope, and conditions.

New policy

General

Configure which users and tables are governed by this policy.

Excluded users will remain unaffected and will still be able to use operations that requires full table access, such as cloning, Delta Sharing, and time travel.

Name*

hide_eu_customers

Description

Enter a description for this policy

Applied to...*

All account users X

Except for...

Search principals

Scope*

ms_learn X

All schemas

ms_learn (All schemas)

☒ Apply to tables that have specific tags

☒ Tags

pii : address X

Condition: hasTagValue('pii','address')

☐ Custom condition

Purpose

What should this policy do?

☐ Mask column data

Columns must have all selected tags to be masked.

☐ Hide table rows

Restrict access to individual rows in a table based on their content.

Create column mask policies

Column mask policies control what values users see in specific columns. Like row filters, these policies rely on UDFs to implement the masking logic.

Create a masking function that returns a redacted value:

```
CREATE OR REPLACE FUNCTION mask_ssn(ssn STRING)
RETURNS STRING
DETERMINISTIC
RETURN '***-**-****';
```

Then create a column mask policy:

```
CREATE POLICY mask_sensitive_ssn
ON SCHEMA customers
COLUMN MASK mask_ssn
```

```
TO `all_users`  
EXCEPT `compliance_team`  
FOR TABLES  
MATCH COLUMNS  
  hasTagValue('pii', 'ssn') AS ssn_col  
ON COLUMN ssn_col;
```

This policy masks SSN values for all users except the compliance team. Any column in the `customers` schema tagged with `pii=ssn` displays `***-**-****` instead of the actual value.

Follow UDF best practices

The functions that power your ABAC policies run on every query against protected tables. Poorly designed UDFs can significantly impact query performance.

Follow these guidelines when writing UDFs for ABAC:

- **Keep logic simple:** Use basic `CASE` statements and boolean expressions
- **Stay deterministic:** The same input should always produce the same output
- **Avoid external calls:** Don't make API calls or lookups to other databases
- **Reference only table columns:** This enables query optimization and predicate pushdown
- **Test at scale:** Validate performance with at least 1 million rows

Avoid common performance pitfalls:

- Calling `is_account_group_member()` or `is_member()` inside UDFs
- Using complex subqueries or joins within the function
- Performing heavy regex operations on large text fields
- Including non-deterministic logic that prevents caching

Keep access control decisions in the policy definition, not the UDF. The policy specifies who the rules apply to; the UDF specifies how to transform the data.

Understand policy inheritance and scope

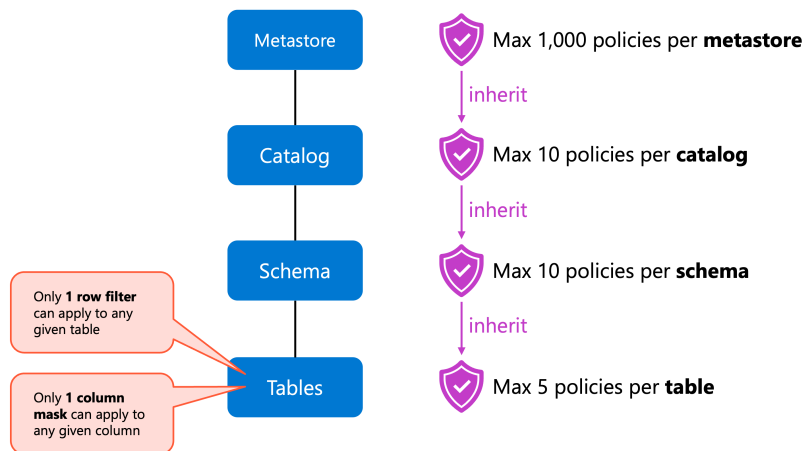
ABAC policies follow an inheritance model that simplifies governance at scale. A policy defined at the catalog level automatically applies to all schemas and tables within that catalog. Similarly, schema-level policies apply to all tables in the schema.

Policy quotas limit how many policies you can create:

- 10 policies per catalog
- 10 policies per schema
- 5 policies per table

Only one row filter can apply to any given table, and only one column mask can apply to any given column. If multiple policies would result in multiple filters or masks, Azure Databricks blocks access

and throws an error.



Important

You must use Databricks Runtime 16.4 or above, or serverless compute, to access tables secured by ABAC policies. Users not subject to the policy can use any runtime.

When you configure ABAC effectively, you create a governance model that scales with your organization. Adding the appropriate governed tag to a new table immediately brings it under the protection of existing policies, without requiring additional configuration.

4. Apply data retention policies

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/4-apply-data-retention-policies>

Apply data retention policies

Completed

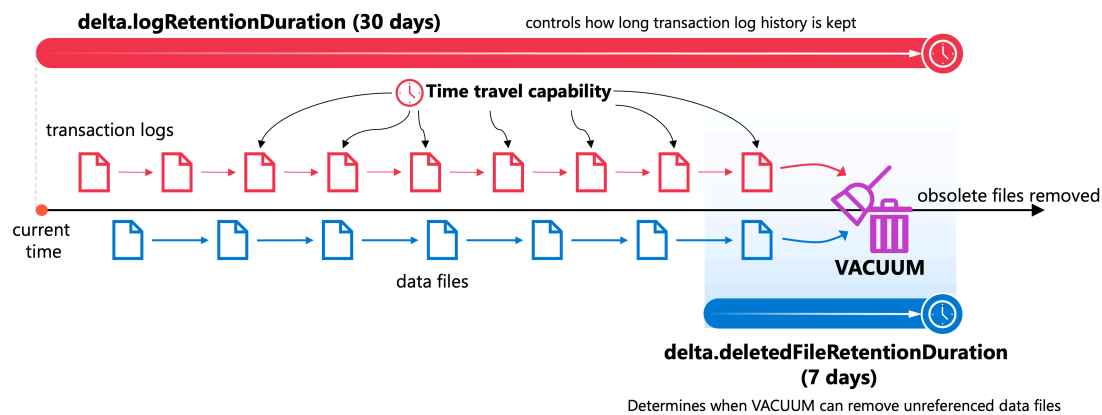
Every organization faces the challenge of balancing data availability with compliance requirements and storage costs. When you implement **data retention policies** in Azure Databricks, you ensure that data is kept only as long as necessary while remaining accessible for legitimate business needs. This becomes especially critical when regulations require you to delete personal data upon request.

In this unit, you learn how to configure retention settings, use **VACUUM** to remove obsolete data, handle deletion requests for compliance, and automate maintenance with **predictive optimization**.

Configure Delta Lake retention settings

Delta Lake maintains table history through **transaction logs** and **data files**. Two key properties control how long this historical data remains available:

Property	Default	Purpose
<code>delta.logRetentionDuration</code>	30 days	Controls how long transaction log history is kept
<code>delta.deletedFileRetentionDuration</code>	7 days	Determines when VACUUM can remove unreferenced data files



These settings work together to define your **time travel** capabilities. For example, if you need 30 days of historical data access, you must configure both properties accordingly:

```
ALTER TABLE sales.customers.transactions
SET TBLPROPERTIES (
  'delta.logRetentionDuration' = 'interval 30 days',
  'delta.deletedFileRetentionDuration' = 'interval 30 days'
);
```

Important

Increasing retention duration increases storage costs because more data files are preserved. Before adjusting these settings, evaluate your compliance requirements against storage budget constraints.

You must set both of these properties to ensure table history is retained for longer duration for tables with frequent `VACUUM` operations

To view current retention settings for a table, use the following command:

```
SHOW TBLPROPERTIES sales.customers.transactions;
```

Remove obsolete data with `VACUUM`

The `VACUUM` command removes data files that are no longer referenced by your Delta table and are older than the retention threshold. This operation permanently deletes historical data, so run it carefully.

```
-- Remove files older than the default retention period (7 days)
VACUUM sales.customers.transactions;

-- Remove files older than a specific retention period
VACUUM sales.customers.transactions RETAIN 168 HOURS;
```

When you run `VACUUM`, Delta Lake identifies files associated with versions older than the retention threshold and removes them from storage. After this operation, time travel queries to those older versions fail because the underlying data no longer exists.

For tables with **deletion vectors** enabled, you must also run `REORG TABLE ... APPLY (PURGE)` after deleting records to permanently remove the underlying data. `VACUUM` by itself does not remove data that deletion vectors have marked for deletion.

Note

Deletion vectors are a storage optimization feature you can enable on Delta Lake tables. By default, when a single row in a data file is deleted, the entire Parquet file containing the record must be rewritten. With deletion vectors enabled, it marks rows as deleted without rewriting the entire file.

Handle compliance deletion requests

Privacy regulations grant individuals the "**right to be forgotten**," requiring you to delete their personal data upon request. Delta Lake's **ACID transactions** make these deletions straightforward.

Delete individual records

Use the `DELETE` statement to remove specific records:

```
DELETE FROM bronze.users
WHERE user_id = 12345;
```

Process bulk deletion requests

For multiple deletion requests, use a `MERGE` statement with a control table that tracks pending requests:

```
MERGE INTO bronze.users AS target
USING (
  SELECT user_id
  FROM compliance.deletion_requests
  WHERE status = 'pending'
) AS source
ON target.user_id = source.user_id
WHEN MATCHED THEN DELETE;
```

After executing deletions, update your control table to mark requests as completed and run maintenance operations to permanently remove the data.

Propagate deletions through your data layers

When you delete data in your bronze layer, you must also remove it from silver and gold layers:

- **Materialized views** automatically reflect deletions after a refresh.
- **Streaming tables** require special handling because they expect append-only data. Add the `skipChangeCommits` option to ignore upstream deletions:

```
spark.readStream.option('skipChangeCommits', 'true').table("bronze.users")
```

After running deletion operations, execute `VACUUM` to permanently remove the data files from storage.

Automate retention with predictive optimization

Manually scheduling `VACUUM` and `OPTIMIZE` operations across many tables is time-consuming and error-prone. **Predictive optimization** in Unity Catalog automates these maintenance tasks for managed tables.

When enabled, predictive optimization automatically:

- Runs `VACUUM` to remove unreferenced files
- Runs `OPTIMIZE` to improve file layout and query performance
- Runs `ANALYZE` to update table statistics

Enable predictive optimization at the schema level:

```
ALTER SCHEMA sales.customers ENABLE PREDICTIVE OPTIMIZATION;
```

Tip

Before enabling predictive optimization, set your desired `delta.deletedFileRetentionDuration` on tables that require longer retention periods. The default `VACUUM` retention is 7 days, which might be shorter than your compliance requirements.

To check whether predictive optimization is enabled for a table:

```
DESCRIBE TABLE EXTENDED sales.customers.transactions;
```

You can monitor predictive optimization operations through the system table `system.storage.predictive_optimization_operations_history`, which shows when maintenance ran, what operations completed, and associated costs.

Note

Predictive optimization is enabled by default if your account was created on or after November 11, 2024.

Integrate with Azure Storage lifecycle policies



For data stored in Azure Blob Storage, you can complement Delta Lake retention with **Azure Storage lifecycle management policies**. These policies automatically transition data to cooler storage tiers or delete blobs based on access patterns and age.

Consider using Azure lifecycle policies to:

- Move infrequently accessed historical data to **cool** or **archive** tiers
- **Delete** temporary files or staging data after a defined period
- Reduce storage costs for data that remains accessible but rarely queried

Lifecycle policies operate at the storage layer, while Delta Lake retention operates at the table layer. Use both together for a comprehensive data retention strategy.

Challenges with archiving Delta tables

When implementing storage lifecycle policies with Delta Lake, you must account for how Delta Lake manages files:

- **File creation time vs. modification time:** Azure Databricks typically relies on file creation time for lifecycle management. Operations like `UPDATE`, `MERGE`, `DELETE`, and `OPTIMIZE` rewrite Parquet files, which resets their creation time and can delay archival.
- **Querying archived data:**
 - **Offline tiers (Archive):** Querying files in offline tiers fails unless they're restored
 - **Online tiers (Cool, Cold):** Queries succeed but may incur higher access costs if data is frequently accessed.
- **Preventing frequent access:** It can be challenging to ensure users don't accidentally query archived data, which would trigger costly retrieval or failures.

Strategies for effective archiving

To overcome these challenges and implement a cost-effective archival strategy:

- **Partition data by date:** Partitioning tables by ingestion date or transaction date makes it easier to identify and manage older data files.
- **Define storage lifecycle policies:** Configure Azure Storage lifecycle management rules to move older files to cool or archive tiers based on their age (for example, move to Cool after 30 days, Archive after 365 days).
- **Set the `delta.timeUntilArchived` property:** Configure this table property to match your storage lifecycle policy. This informs Azure Databricks that files older than the specified period are archived, preventing the query optimizer from assuming they're immediately available.

```
ALTER TABLE sales_data
SET TBLPROPERTIES (delta.timeUntilArchived = '365 days');
```

- **Use views to restrict access:** Create views that filter out archived data to prevent users from accidentally querying it.

```
CREATE VIEW active_sales_data AS
SELECT * FROM sales_data
WHERE ingestion_date > CURRENT_DATE() - INTERVAL 365 DAYS;
```

- **Optimize and Vacuum with predicates:** When running maintenance commands, use predicates to avoid touching archived partitions, which would trigger retrieval.

```
OPTIMIZE sales_data
WHERE ingestion_date >= current_timestamp() - INTERVAL 30 days;
```

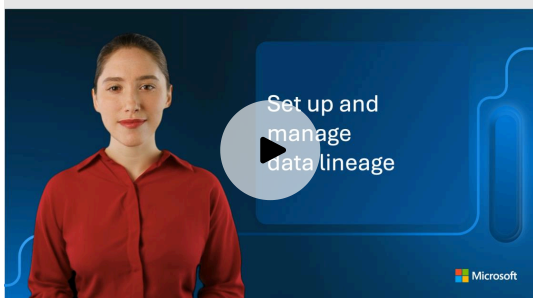
Implementing data retention policies requires ongoing attention as business requirements and regulations evolve. Establish clear processes for reviewing and updating your retention settings, and document your compliance procedures so your team can respond quickly to deletion requests.

5. Set up and manage data lineage

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/5-manage-data-lineage-tracking>

Set up and manage data lineage

Completed

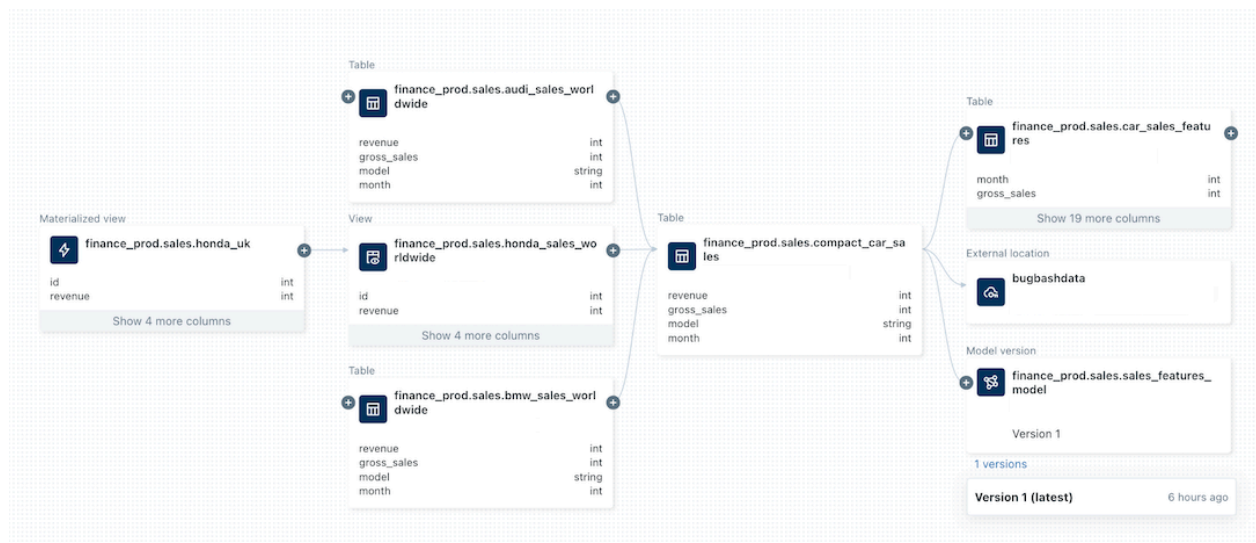


Understanding where your data comes from, how it transforms, and where it flows is essential for effective data governance. When you set up and manage **data lineage tracking** in Unity Catalog, you gain visibility into data dependencies, ownership, and history. This visibility helps you perform **impact analysis** before making changes, troubleshoot data quality issues, and meet compliance requirements.

In this unit, you learn how to use Catalog Explorer to view and manage data lineage, explore table history and ownership, and understand the connections between your data assets.

Understand data lineage in Unity Catalog

Unity Catalog automatically captures **runtime data lineage** across queries run on Azure Databricks. This lineage tracking works across all languages and captures relationships down to the **column level**. Lineage data includes the notebooks, jobs, and dashboards that interact with your tables.



Consider a scenario where a sales analytics dashboard suddenly shows incorrect revenue figures. With lineage tracking, you can trace back through the data flow to identify which upstream table or transformation caused the issue. This capability transforms troubleshooting from guesswork into a systematic investigation.

Configure compute to capture lineage

For Unity Catalog to capture lineage automatically, queries must run on **Unity Catalog-enabled compute**. This includes:

- **Compute clusters** configured with Unity Catalog access mode (shared, single user, or serverless)
- **SQL warehouses** that have Unity Catalog enabled
- **Jobs** running on Unity Catalog-enabled clusters

When you create a cluster with access to Unity Catalog, the compute automatically captures lineage for all queries that read from or write to Unity Catalog tables. This happens without requiring additional configuration or code changes, making lineage capture seamless and requiring minimal administrative effort.

Clusters that are not Unity Catalog-enabled, or queries that access tables by direct file path instead of catalog names, do not generate lineage data. To ensure comprehensive lineage coverage, configure all production workloads to use Unity Catalog-enabled compute and reference tables using their three-level namespace (`catalog.schema.table`).

Lineage is aggregated across all workspaces attached to a Unity Catalog metastore. When you capture lineage in one workspace, users in other workspaces that share the same metastore can view that lineage information. This **cross-workspace visibility** is particularly valuable for organizations with distributed data teams.

Note

Lineage data is retained for **one year**. You can filter lineage data by time frame within this window to focus on specific periods.

View data lineage using Catalog Explorer

Catalog Explorer provides a visual interface for exploring lineage relationships. To view lineage for a table:

1. In your Azure Databricks workspace, select **Catalog** in the sidebar.
2. Search or browse to locate your table.
3. Select the **Lineage** tab to display related tables in the lineage panel.
4. Select **See Lineage Graph** to open an interactive visualization of the data flow.

The lineage graph displays tables as nodes with arrows showing data flow direction. By default, one level of connections is visible. Select the **+** icon on any node to reveal more upstream or downstream connections.

When you select an arrow connecting two nodes, the **Lineage connection** panel opens. This panel shows details about the connection, including source and target tables, the notebooks that created the relationship, and the jobs that ran the transformations.

Explore column-level lineage

Column-level lineage shows how individual columns derive from source columns. This granularity helps you understand complex transformations and identify which source fields contribute to a calculated column.

To view column-level lineage, select a column in the lineage graph. The visualization highlights the upstream columns that contribute to your selected column and the downstream columns that depend on it.

View job and dashboard lineage

Beyond table relationships, Catalog Explorer shows which jobs and dashboards consume your data:

- To view **job lineage**, go to the table's **Lineage** tab, select **Jobs**, and choose **Downstream** to see jobs that consume the table.
- To view **dashboard lineage**, select **Dashboards** in the **Lineage** tab to see which dashboards query the table.

This information supports impact analysis. Before modifying a table schema, you can identify which downstream jobs and dashboards might be affected.

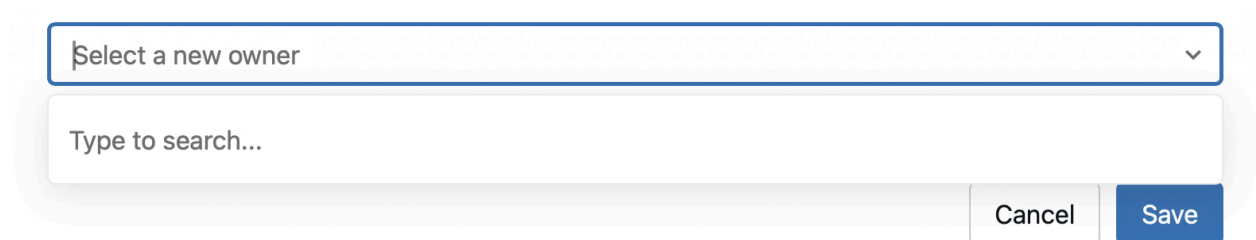
Manage table ownership

Every securable object in Unity Catalog has an **owner**. The owner has full privileges on the object and can grant permissions to other users. Understanding and managing ownership is essential for governance accountability.

To view or change ownership in Catalog Explorer:

1. Navigate to the table in Catalog Explorer.
2. On the **Overview** tab, locate the current owner.
3. Select the edit icon next to **Owner**.
4. Search for and select a user, group, or service principal.
5. Select **Save**.

Set owner for ms_learn.pipeline_lab.customers_bronze ×



Only the current owner or a **metastore admin** can transfer ownership. After transfer, the previous owner loses ownership privileges unless explicitly granted.

To view lineage, users need at least the **BROWSE** privilege on the parent catalog. Objects that a user doesn't have permission to access appear as **masked nodes** in the lineage graph. This security model ensures that lineage visibility respects your access control policies.

View table history and version changes

For **Delta tables**, Catalog Explorer displays a complete history of all operations. The **History** tab shows each version with timestamps, operation types, and the user who made the change.

You can also query table history using SQL:

```
-- View full table history
DESCRIBE HISTORY sales.customers.orders;

-- View only the most recent operation
DESCRIBE HISTORY sales.customers.orders LIMIT 1;
```

The history includes detailed information about each operation:

Column	Description
version	Table version number
timestamp	When the operation occurred
operation	Type of operation (WRITE, UPDATE, DELETE, MERGE)
userName	User who performed the operation
operationMetrics	Metrics like rows affected and files modified

This history supports auditing and compliance by providing a complete record of who changed what and when.

Query lineage data programmatically

For reporting and advanced analysis, you can query the **lineage system tables** directly. Unity Catalog provides two lineage tables:

- `system.access.table_lineage` records read and write events at the table level
- `system.access.column_lineage` tracks column-level dependencies

The following query finds the most frequently accessed tables in the past week:

```
SELECT
  source_table_full_name,
  COUNT(DISTINCT event_id) AS query_count
FROM system.access.table_lineage
WHERE event_date > CURRENT_DATE() - INTERVAL 7 DAYS
  AND source_table_full_name IS NOT NULL
GROUP BY source_table_full_name
ORDER BY query_count DESC
LIMIT 10;
```

You can also use *Databricks Assistant* to explore lineage interactively. From Catalog Explorer, select the Assistant icon and type:

- `/getTableLineages` to view upstream and downstream dependencies
- `/getTableInsights` to access metadata-driven insights like user activity patterns

These queries help answer questions such as "who queries this table most often" or "what are the downstream consumers of this data."

Understand lineage limitations

Lineage tracking has certain constraints you should be aware of:

- Lineage is **not preserved for renamed objects**. If you rename a table, the lineage connection to its previous name is lost.
- Tables accessed by path rather than name (for example, `delta.'abfss://...'`) show only the path in lineage, not the table name.
- Global temporary views and system tables under `system.information_schema` are not captured in lineage.
- Some user-defined functions can obscure column-level lineage by hiding the mapping between source and target columns.

Despite these limitations, lineage tracking provides comprehensive visibility for most data engineering workflows. By understanding these constraints, you can design your pipelines to maximize lineage capture.

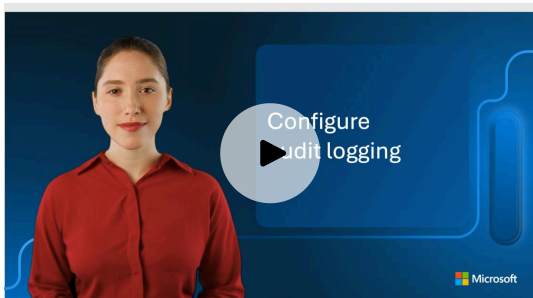
Effective data lineage management transforms how you govern and troubleshoot your data platform. With the visibility that Catalog Explorer and system tables provide, you can confidently manage dependencies, assess change impact, and maintain accountability across your data assets.

6. Configure audit logging

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/6-configure-audit-logging>

Configure audit logging

Completed



Tracking who did what, when, and to which data objects is essential for governance, security, and compliance. When you configure **audit logging** in Azure Databricks, you gain visibility into every significant action across your workspace and account. This visibility helps you meet regulatory requirements, investigate security incidents, and maintain accountability across your data platform.

Unity Catalog audit logging is **configured at the account level** and automatically applies to all workspaces within that account. This account-level configuration ensures consistent audit capture

across your entire Databricks environment without requiring per-workspace setup.

In this unit, you learn how to access and query audit logs, understand what events are captured, and use these logs for common governance scenarios.

Understand the audit log system table

Azure Databricks automatically captures audit events and stores them in the **audit log system table** located at `system.access.audit`. This table records actions across all workspaces attached to your Unity Catalog metastore, providing a **centralized view** of activity across your organization.

Consider a scenario where your security team needs to investigate unauthorized data access. Rather than searching through multiple log files, you query a single table that contains every tracked event. This centralized approach transforms compliance investigations from manual searches into straightforward SQL queries.

The audit log system table uses a consistent schema that includes the following key columns:

Column	Description
<code>event_time</code>	Timestamp when the action occurred
<code>event_date</code>	Calendar date of the action
<code>user_identity</code>	Identity of the user who initiated the action
<code>workspace_id</code>	The Azure Databricks workspace in which the audit event took place
<code>service_name</code>	Service category (for example, <code>unityCatalog</code> , <code>notebook</code> , <code>clusters</code>)
<code>action_name</code>	Specific action performed (for example, <code>getTable</code> , <code>createTable</code>)
<code>request_params</code>	Parameters included in the request
<code>response</code>	Response details including status codes and results
<code>source_ip_address</code>	IP address where the request originated

Account-level events record `workspace_id` as `0`, while workspace-level events include the actual workspace identifier. This distinction helps you filter events by scope when investigating issues.

Note

Azure Databricks retains audit logs for up to **one year** for security and fraud analysis purposes. Plan your long-term retention strategy if you need logs beyond this period.

Identify tracked events and services

The audit log captures events across numerous services within Azure Databricks. Understanding which actions are logged helps you design effective monitoring and compliance strategies.

Workspace-level services log events for actions taken within a specific workspace:

- **Unity Catalog:** Table creation, deletion, updates, and permission changes
- **Notebooks:** Commands executed, notebooks created, modified, or deleted
- **Clusters:** Cluster creation, resizing, and termination
- **Jobs:** Job creation, runs, and failures
- **Databricks SQL:** Query execution, warehouse management, and alert activity
- **Accounts:** User logins, token generation, and access control changes

Account-level services capture events that span **multiple workspaces**:

- **Account management:** User and group management across the account
- **Billable usage:** Access to usage reports and aggregated metrics
- **Delta Sharing:** Provider and recipient actions for shared data

For example, when a user queries a table in Unity Catalog, the `unityCatalog` service logs a `getTable` action. If the same user runs a command in a notebook, the `notebook` service logs a `runCommand` action. Each service provides specific action names that describe precisely what occurred.

Tip

For a comprehensive reference of audit log services and events, consult this article [Diagnostic log reference](#)

Query audit logs for common scenarios

You can write SQL queries against the audit log system table to answer common governance questions. The following examples demonstrate patterns you can adapt for your organization's needs.

Find who accessed a specific table

This query identifies users who accessed a particular table in the past seven days:

```
SELECT
  user_identity.email AS user_email,
  action_name AS access_type,
  event_time AS access_time
FROM system.access.audit
WHERE
  request_params.full_name_arg = 'catalog.schema.table_name'
  AND action_name IN ('getTable', 'createTable', 'deleteTable')
```

```
AND event_date > CURRENT_DATE() - INTERVAL 7 DAYS
ORDER BY event_time DESC;
```

Track permission changes across your metastore

This query shows all permission modifications, helping you audit access control changes:

```
SELECT
  event_time,
  user_identity.email AS changed_by,
  request_params.securable_type,
  request_params.securable_full_name,
  request_params.changes
FROM system.access.audit
WHERE
  service_name = 'unityCatalog'
  AND action_name = 'updatePermissions'
ORDER BY event_time DESC;
```

Identify failed login attempts

Security investigations often require finding authentication failures. This query surfaces denied access attempts:

```
SELECT
  event_time,
  user_identity.email AS user_email,
  source_ip_address,
  action_name,
  response.error_message
FROM system.access.audit
WHERE
  service_name = 'accounts'
  AND response.status_code != 200
  AND event_date > CURRENT_DATE() - INTERVAL 30 DAYS
ORDER BY event_time DESC;
```

These queries form a foundation for your monitoring strategy. You can extend them to create alerts, build dashboards, or export data to external security information and event management (SIEM) platforms.

Enable verbose audit logs

By default, Azure Databricks logs high-level events but doesn't capture the full text of commands executed in notebooks. When you need detailed visibility into exactly what code ran and when, enable **verbose audit logs**.

With verbose logging enabled, the audit log captures additional events:

Service	Action	Description
notebook	runCommand	Logs the full command text executed in a notebook cell
jobs	runCommand	Logs commands executed by job runs
databrickssql	commandSubmit	Logs SQL commands submitted to warehouses

To enable verbose audit logs:

1. Sign in to your Azure Databricks workspace as an administrator.
2. Navigate to **Admin Settings** by selecting your username and choosing **Admin Settings**.
3. Select the **Advanced** tab.
4. Locate **Verbose Audit Logs** and enable the feature.

Settings

 Workspace admin

Appearance

Identity and access

Security

Compute

Development

Notifications

Advanced

 User

Profile

Preferences

Developer

Linked accounts

Notifications

DBFS File Browser

Enable or disable DBFS File Browser

Off 

Databricks Autologging

Enable or disable Databricks Autologging for this workspace. When enabled, ML model training runs executed interactively on clusters with supported versions of the Databricks Runtime for Machine Learning will automatically be logged to MLflow.

On 

[Learn more](#) 

Legacy MLflow Model Serving

Enable or disable legacy MLflow model serving for this workspace. Disabling this option will not disable existing model serving.

On 

Verbose Audit Logs

Enable or disable verbose audit logs.

Off 

FileStore Endpoint

Enable or disable FileStore endpoint /files.

FileStore files are accessible at /files. When enabled, names of the files stored in FileStore have to be considered public.

On 

[Learn more](#) 

When you enable or disable verbose logging, Azure Databricks logs this configuration change as an auditable event. This creates accountability for who modified the logging configuration and when.

Important

Verbose logs capture command text, which might include sensitive information such as query parameters or hardcoded values. Ensure your security policies account for this additional data exposure.

Configure log delivery to external systems



While the audit log system table provides direct access within Databricks, many organizations route logs to external systems for centralized monitoring. Azure Databricks integrates with Azure diagnostic settings to deliver logs to multiple destinations, each optimized for different use cases:

- **Azure Log Analytics:** Provides near-real-time ingestion (typically within 15 minutes), centralized storage, and powerful querying using Kusto Query Language (KQL). This makes it ideal for fast analysis, interactive investigations, and building custom dashboards for auditing Unity Catalog access events.
- **Azure Storage accounts:** Offers low-cost, long-term archival and batch analysis. Best suited for compliance scenarios where logs are accessed infrequently but must be retained for extended periods.
- **Azure Event Hubs:** Enables real-time streaming to downstream consumers and processing systems. Ideal when you need to integrate audit logs with existing event-driven architectures or SIEM platforms.

Platform administrators configure log delivery at the account level through the Azure portal. When building monitoring solutions, account for typical ingestion latency of approximately **15 minutes** in your alerting thresholds.

Microsoft Azure

Search resources, services, and docs (G+)

[Home](#) >
[ms-learn](#) |
[Diagnostic settings](#) >

Diagnostic setting

Save

Discard

Delete

Feedback

A diagnostic setting specifies a list of categories of platform logs and/or metrics that you want to collect from a resource, and one or more destinations that you would stream them to. Normal usage charges for the destination will occur. [Learn more about the different log categories and contents of those logs](#)

Diagnostic setting name *

Logs

Category groups

☐ allLogs

Categories

☐ Databricks File System

☐ Databricks Clusters

☐ Databricks Accounts

☒ Databricks Jobs

☒ Databricks Notebook

☐ Databricks SSH

Destination details

☒ Send to Log Analytics workspace

Subscription

ms-learn

Log Analytics workspace

No workspaces in this subscription.

☐ Archive to a storage account

☐ Stream to an event hub

☐ Send to partner solution

The combination of the system table for interactive queries and external delivery for operational monitoring provides comprehensive coverage. You can use the system table for ad-hoc investigations while relying on your SIEM platform for continuous security monitoring.

Understanding audit logging capabilities positions you to support compliance requirements and respond quickly to security incidents. With the audit log system table, you have a powerful tool for tracking activity, investigating issues, and demonstrating governance controls to stakeholders.

7. Design secure Delta Sharing strategy

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/7-secure-delta-sharing-strategy>

Design secure Delta Sharing strategy

Completed

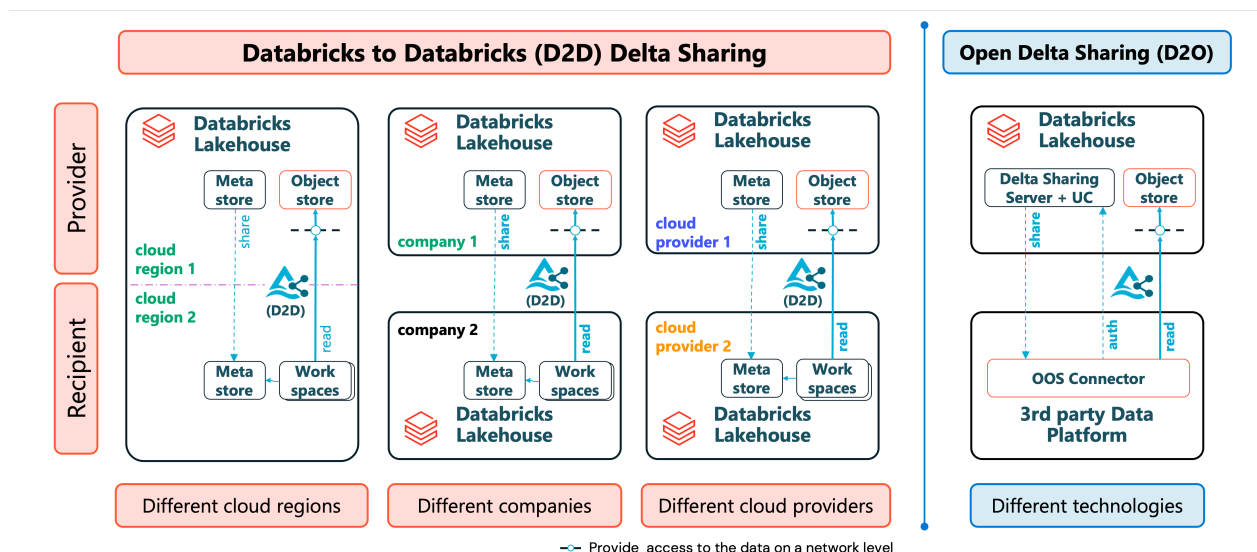


Sharing data securely with external partners, other teams, or across organizational boundaries requires a deliberate strategy. **Delta Sharing** in Azure Databricks lets you share live data without copying it, while maintaining full control over who accesses what. Designing a secure sharing strategy means choosing the right authentication method, applying appropriate access controls, and establishing monitoring practices that keep your data governance intact.

In this unit, you learn how to evaluate sharing protocols, implement security controls, and design fine-grained access patterns that align with your organization's compliance requirements.

Choose the right sharing protocol

Delta Sharing offers two primary protocols, each suited for different recipient scenarios. Your choice affects how authentication works and what assets you can share.



Databricks-to-Databricks sharing is the preferred option when recipients have access to a Unity Catalog-enabled Azure Databricks workspace. This protocol handles all security automatically through the platform. Recipients provide their **sharing identifier**—a string containing their metastore's cloud, region, and unique ID—and the connection establishes without token management. You can share tables, views, volumes, notebook files, and AI models through this approach.

Open sharing works when recipients use platforms other than Azure Databricks or lack access to Unity Catalog. You manage authentication by generating **bearer tokens** or configuring **OIDC**

(OpenID Connect) federation. While more flexible, this approach requires careful token lifecycle management and supports only tabular data.

Consider these factors when selecting your protocol:

Factor	Databricks-to-Databricks	Open sharing
Recipient platform	Databricks with Unity Catalog	Any platform
Authentication	Platform-managed	Token or OIDC
Shareable assets	Tables, views, volumes, notebook files, models	Tables only
Token management	Not required	Required
Performance	Optimized with history sharing	Standard

When recipients exist on both Databricks and non-Databricks platforms, you can create multiple recipient objects for the same logical organization—one for each access pattern.

Design authentication and access controls

A secure sharing strategy starts with proper authentication configuration and follows the principle of **least privilege**.

For **open sharing recipients**, configure **token expiration** to limit exposure if credentials are compromised. Set the default token lifetime at the metastore level:

1. Navigate to the account console and select your metastore.
2. Enable **Allow Delta Sharing with parties outside your organization**.
3. Set an expiration period rather than allowing tokens to live indefinitely.

For recipients who shouldn't rotate tokens frequently, consider **OIDC federation** instead. This approach grants short-lived Databricks OAuth tokens in exchange for JWT tokens from the recipient's identity provider, removing long-lived credentials entirely.

Restrict network access for sensitive shares by assigning **IP access lists** to open sharing recipients. You can limit access to specific IP addresses or CIDR ranges (Classless Inter-Domain Routing):

```
-- View recipient details including IP restrictions
DESCRIBE RECIPIENT finance_partner;
```

Each recipient supports up to 100 IP/CIDR values. When a request originates from an unlisted address, Delta Sharing denies access and logs the attempt.

Delegate sharing privileges carefully. Rather than granting metastore admin access broadly, use specific privileges:

- **CREATE SHARE** : Allows creating new shares
- **CREATE RECIPIENT** : Allows creating new recipients

- `USE SHARE` and `SET SHARE PERMISSION` : Together allow granting share access to recipients
- `USE RECIPIENT` : Allows viewing recipient details

Assign these privileges to groups representing your data stewardship team rather than individuals.

Configure shares and recipients

After designing your security approach, you need to create the actual share and recipient objects in Unity Catalog. This section walks through the configuration steps for both sharing protocols.

Create a share

A **share** is a securable object that contains a collection of tables, views, volumes, notebooks, and models you want to share. Create a share using SQL or Catalog Explorer:

```
-- Create a new share
CREATE SHARE IF NOT EXISTS partner_analytics
COMMENT 'Sales analytics data for partner integration';
```

Add assets to a share

Once created, add the data assets recipients should access. You can share individual tables, views, or entire schemas:

```
-- Add a table to the share
ALTER SHARE partner_analytics ADD TABLE catalog.schema.sales_summary;

-- Add a table with an alias (useful when internal naming differs from external)
ALTER SHARE partner_analytics ADD TABLE catalog.schema.internal_metrics
AS external_catalog.analytics.metrics;

-- Share a table with history for time travel queries
ALTER SHARE partner_analytics ADD TABLE catalog.schema.transactions
WITH HISTORY;

-- Add an entire schema (includes all current and future assets)
ALTER SHARE partner_analytics ADD SCHEMA catalog.analytics_schema;
```

For Databricks-to-Databricks sharing, you can also add volumes and views:

```
-- Add a volume
ALTER SHARE partner_analytics ADD VOLUME catalog.schema.documentation;

-- Add a view
ALTER SHARE partner_analytics ADD VIEW catalog.schema.aggregated_sales;
```

Note

To add notebook files to a share, use Catalog Explorer. Navigate to **Delta Sharing > Shared by me**, select your share, click **Manage assets > Add notebook file**, and browse for the notebook you want to share.

Create a recipient for Databricks-to-Databricks sharing

For recipients with Unity Catalog-enabled workspaces, request their **sharing identifier** first. This string contains their metastore's cloud, region, and UUID:

```
-- Create a recipient using their sharing identifier
CREATE RECIPIENT IF NOT EXISTS contoso_analytics
USING ID 'azure:westeurope:a1b2c3d4-e5f6-7890-abcd-ef1234567890'
COMMENT 'Contoso data analytics team';
```

The recipient can find their sharing identifier by running:

```
SELECT CURRENT_METASTORE();
```

Create a recipient for open sharing

For recipients without Databricks access, create a token-authenticated recipient:

```
-- Create an open sharing recipient
CREATE RECIPIENT IF NOT EXISTS external_partner
COMMENT 'External analytics platform integration';
```

After creation, retrieve the **activation link** to send to your recipient:

```
DESCRIBE RECIPIENT external_partner;
```

The activation link allows the recipient to download their credential file once. Share this link through a secure channel.

Grant recipients access to shares

After creating both the share and recipient, grant access:

```
-- Grant a recipient access to a share
GRANT SELECT ON SHARE partner_analytics TO RECIPIENT contoso_analytics;

-- View current grants on a share
SHOW GRANT ON SHARE partner_analytics;

-- View shares accessible to a recipient
SHOW GRANTS TO RECIPIENT contoso_analytics;
```

Manage shares and recipients

Use these commands to view and manage your sharing configuration:

```
-- List all shares
SHOW SHARES;

-- View share details including assets
DESCRIBE SHARE partner_analytics;
SHOW ALL IN SHARE partner_analytics;

-- List all recipients
SHOW RECIPIENTS;

-- Remove a table from a share
ALTER SHARE partner_analytics REMOVE TABLE catalog.schema.deprecated_table;

-- Revoke recipient access
REVOKE SELECT ON SHARE partner_analytics FROM RECIPIENT former_partner;

-- Delete a recipient
DROP RECIPIENT former_partner;
```

Implement fine-grained access with dynamic views

Standard table sharing grants recipients access to all rows and columns. When you need to restrict what specific recipients see, use **dynamic views** that filter data based on recipient properties.

Attach **custom properties** to recipients that correspond to your data segmentation:

```
-- Set country_region property for partition filtering
ALTER RECIPIENT regional_partner SET PROPERTIES ('country_region' = 'US');
```

Create a dynamic view that filters rows based on the recipient's property:

```
CREATE VIEW catalog.schema.regional_sales AS
SELECT * FROM catalog.schema.all_sales
WHERE region = CURRENT_RECIPIENT('country_region');
```

For column-level security, mask sensitive data for recipients who shouldn't see it:

```
CREATE VIEW catalog.schema.customer_data AS
SELECT
  customer_id,
  CASE
    WHEN CURRENT_RECIPIENT('access_level') = 'full' THEN email
    ELSE 'REDACTED'
  END AS email,
```

```
purchase_history
FROM catalog.schema.customers;
```

When you share this dynamic view, each recipient sees only the data their properties permit. The provider can't query views using `CURRENT_RECIPIENT()` directly—test by sharing with yourself as a recipient.

For simpler scenarios, **partition filtering** lets you share specific slices of a table without creating views:

```
ALTER SHARE partner_share ADD TABLE inventory
PARTITION (region = CURRENT_RECIPIENT('region'));
```

Monitor sharing activity through audit logs

Tracking who accesses shared data and when is essential for compliance and security investigations. Azure Databricks captures Delta Sharing events in the **audit log system table**.

Query the audit logs to understand sharing patterns:

```
-- Track share and recipient creation
SELECT
  event_time,
  user_identity.email AS performed_by,
  action_name,
  request_params
FROM system.access.audit
WHERE service_name = 'unityCatalog'
  AND action_name IN ('createShare', 'createRecipient', 'grantOnShare')
ORDER BY event_time DESC;
```

Monitor recipient access to shared tables:

```
-- View recipient data access events
SELECT
  event_time,
  request_params.recipient_name,
  request_params.share_name,
  action_name
FROM system.access.audit
WHERE service_name = 'unityCatalog'
  AND action_name IN ('deltaSharingQueriedTable', 'generateTemporaryTableCredential')
  AND event_date > CURRENT_DATE() - INTERVAL 7 DAYS;
```

The `deltaSharingQueriedTable` event appears for open sharing queries, while `generateTemporaryTableCredential` appears for Databricks-to-Databricks access with **history sharing**. Both events include details about what was accessed and by whom.

Establish regular review cycles for your sharing configuration:

- Audit active recipients monthly to confirm continued business need
- Review share contents to ensure only necessary data is exposed
- Verify token expiration settings haven't been modified
- Check for failed access attempts that might indicate credential compromise

Tip

Configure log delivery to **Azure Event Hubs** or **Log Analytics** if you need real-time alerting on suspicious sharing activity.

Apply security best practices

Building a robust Delta Sharing strategy requires combining technical controls with operational practices.

Document sharing agreements before configuring access. Establish what data is shared, for what purpose, and for how long. This documentation supports compliance audits and helps when reviewing whether shares should continue.

Share only what's necessary. When adding tables to a share, consider whether recipients need the entire table or just specific partitions. Use schema sharing with caution—it automatically includes future tables added to that schema.

Combine controls for defense in depth. A sensitive share might include:

- Databricks-to-Databricks authentication for platform-managed security
- Dynamic views for row-level filtering
- Recipient properties for parameterized access
- Regular audit log reviews

Plan for credential rotation. For open sharing recipients, implement processes to **rotate tokens** before expiration. Rotating a token invalidates the previous one after the old token's expiration, giving recipients time to update their configuration.

Review shares when personnel changes. When a recipient organization restructures, verify that the current sharing identifier still represents the appropriate metastore and that access should continue.

By designing your Delta Sharing strategy around these principles, you create a framework for secure collaboration that scales with your organization's needs while maintaining the governance controls your compliance requirements demand.

8. Exercise - Govern Unity Catalog Objects

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/8-exercise-govern-unity-catalog-objects>

Exercise - Govern Unity Catalog Objects

Completed

Now it's your chance to govern Unity Catalog objects. In this lab, you apply Unity Catalog governance controls to a connected vehicle data platform built on Azure Databricks. You use SQL to tag tables and columns for PII classification, configure Delta Lake retention policies and run VACUUM to purge deleted data, and enable predictive optimization. You then query system tables to trace data lineage programmatically and analyze the audit log to answer compliance questions about who accessed data and when.

Note

To complete this lab, you need an [Azure subscription](#) in which you have administrative access.

Launch the exercise and follow the instructions.

[Launch Exercise](#)

9. Module assessment

<https://learn.microsoft.com/en-us/training/modules/govern-unity-catalog-objects/9-knowledge-check>

Module assessment

Completed

10. Summary

Summary

Completed

Governing Unity Catalog objects requires a comprehensive approach that spans metadata management, access control, compliance, and collaboration. Throughout this module, you explored the tools and techniques that enable effective data governance in Azure Databricks—from documenting your data assets with **comments and tags** to enforcing access policies through **attribute-based access control (ABAC)**. You learned how **governed tags** enable policies that scale automatically as new tables are added to your catalog.

Data lifecycle management emerged as a critical governance capability. You configured **Delta Lake retention settings** and used **VACUUM** to remove obsolete data files. You handled **compliance deletion requests**, propagating deletions through bronze, silver, and gold data layers. **Predictive optimization** automates these maintenance tasks, reducing operational overhead while ensuring storage efficiency.

Understanding your data's journey proved essential for both troubleshooting and compliance. **Data lineage tracking** in Catalog Explorer reveals how data flows through your pipelines, enabling impact analysis before making changes. **Audit logging** through the system table provides visibility into who accessed what data and when, supporting security investigations and regulatory audits. For external collaboration, **Delta Sharing** offers secure data exchange with partners and customers, with options for both Databricks-to-Databricks sharing and open protocol sharing.

Apply these governance practices incrementally as you build your data platform. Start by documenting tables and columns with meaningful comments and tags. Implement ABAC policies for sensitive data classifications. Configure retention policies that balance compliance requirements with storage costs. Review lineage and audit logs regularly to maintain visibility into your data operations. These foundational practices create a governance framework that scales with your organization's data needs.