USING UNITY CATALOG FOR DATA GOVERNANCE

Harmonizing Your Data: Establishing Definitive, Reliable Data Sources and Governance with Databricks Unity Catalog.

INFINITIVE X databricks

TABLE OF CONTENTS

INTRODUCTION

Data is one of the most valuable assets for any organization, but **effective data governance** is what helps unlock its full potential. It's a set of rules, tools, and practices that help manage data throughout its lifecycle—making sure it's secure, well-organized, and aligned with an organization's strategic business goals.

A strong data governance strategy helps:

- Improve visibility and control over data
- Track how data is accessed and used
- Protect against unauthorized access
- Meet data privacy and regulatory compliance requirements
- Build customer trust

Strong, effective data governance can actually create a competitive advantage for organizations, and as AI becomes more widely adopted, the need for data governance will become even more critical to an organization's data strategy and overall success.

That said, effective governance requires organizations to overcome many challenges. Chief among them:

Fragmented Data Landscape

- Data is often spread across multiple systems like data lakes, data warehouses, and cloud storage (e.g., AWS S3, Azure, Google Cloud). This creates **data silos**, making it hard to find, access, or analyze data effectively. Furthermore, unstructured data (like documents or images) makes up 80% of data in organizations, and moving it into structured formats creates even more silos. Adding to the complexity are tools like dashboards, ML models, and notebooks. The result? Delays in decision-making, higher costs, and reduced innovation.

Complex Data Access Management

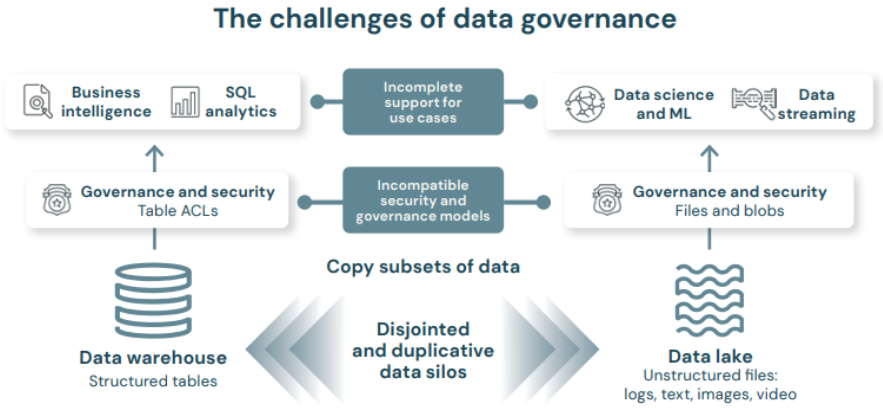
- Different platforms and tools use different ways to manage access controls, which creates considerable complexity, is time consuming, and increases the chance of missed controls that lead to unauthorized access to sensitive data. Lack of consistency and centralization affects collaboration, audits, and data sharing.

Inadequate monitoring and visibility

- Without full monitoring and visibility across the entire lifecycle of data and AI assets, it becomes difficult to perform effective audits, understand the impact of changes, or quickly identify and fix errors. When teams can't clearly trace where data came from, how it has changed, where it has moved, or how it's being used, it becomes much harder to ensure data quality. This lack of insight can lead to costly delays and increased effort when problems arise in data pipelines, dashboards, or reports, as it takes much longer to track down the source of the issue.

Limited cross-platform sharing and collaboration

- Without a standard way to securely share data and AI assets—like machine learning models, notebooks, and dashboards—across different platforms and cloud environments, collaboration becomes challenging. As a result, teams often end up copying the same data across multiple systems, clouds, or regions just to work together. This leads to unnecessary duplication, added storage costs, and inefficiencies.



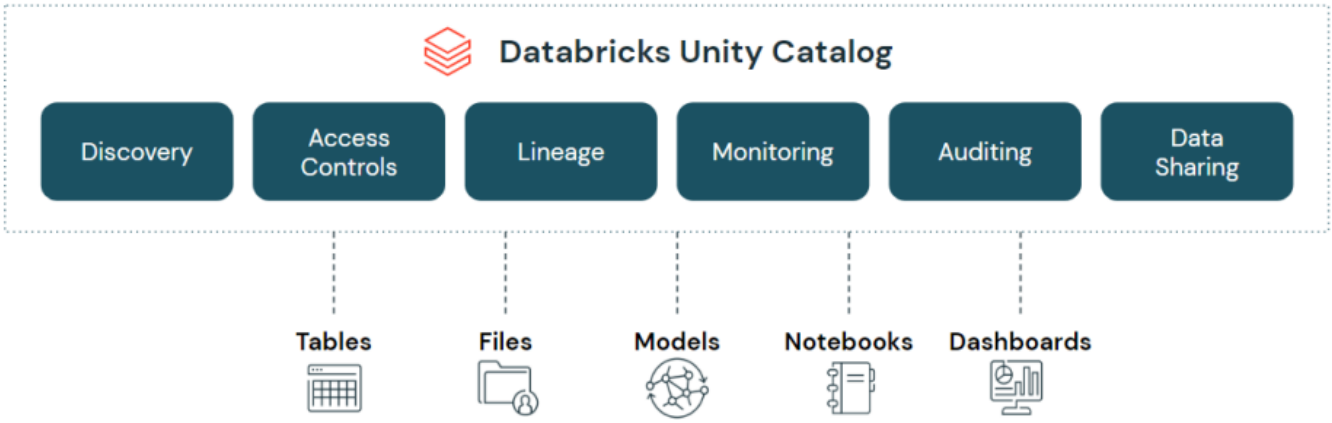
What organizations need, given these challenges, is a single data and analytics governance solution to streamline and automate governance efforts. This can be accomplished through a central system that gives you a big-picture view of your data—what you have, where it is, and how it's being used. Such a solution organizes data by areas like business units, software applications, or stages in development. To make sure only the right people have access to the right data, this system also needs to connect with your company's user and group information. That way, you can control who sees what.

Databricks Unity Catalog is a data governance solution that addresses many of these challenges by offering a central place to organize and share different kinds of data across your company, even if your teams are in different locations or using different cloud platforms.

WHAT IS UNITY CATALOG

Unity Catalog is a powerful tool for managing and securing data, providing centralized access control, auditing, lineage tracking, and data discovery capabilities. Imagine a large library with books, magazines, and digital content spread across multiple rooms and floors. Unity Catalog is like a smart cataloging system for this library that:

- Organizes all the content in a structured way
- Helps you find what you need
- Controls who can access which materials
- Keeps track of who accessed “what” and “when”
- Shows how different pieces of information are connected



Having this central system means your data security and compliance teams can manage access and keep records from one place. This reduces the risk of mistakes, like giving someone access to data they shouldn’t see.

It also lets you group and manage data in a structured way: Data is grouped into logical “catalogs” that can reflect business units or domains, making it easier for teams to find and use what they need—while ensuring sensitive information stays protected.

Key Features and Benefits

Define Once, Secure Everywhere

- Provides a unified platform for administering data access policies across Databricks workspaces and user personas.

Metadata Management

- Centralizes metadata management, enabling standardization of data definitions, structures, and formats. The ability to centralize and standardize field definitions plays a big role in managing data quality and in making sure that the same field, even if it comes from different data sources, has a common definition understood and used by the entire organization. The common standard definition and formatting informs the cleansing rules applied to the data.
- An example might be a standardized definition and format for customer-related fields. Customer data may come in from various sources, may have different data types, but the data must all be standardized and formatted consistently for downstream use by analytics applications, dashboards, reports, etc.

Data Discovery

- The process of finding and understanding what data exists in your organization—what it is, where it's stored, and how it can be used.
- In Unity Catalog, data discovery means that users can easily search, browse, and explore all the data assets your organization has—like tables, files, dashboards, or machine learning models—through a central, organized system. So, instead of guessing where to find the data they need (or not knowing it exists at all), teams can:
  - Quickly search for data by name, keyword, or category
  - See descriptions of what the data is and how to use it
  - Understand who owns the data and who can access it
  - Know when it was last updated and how it's being used
- This makes it much easier for analysts, data scientists, and business users to find trusted data and use it confidently—without needing to go through IT or dig through multiple systems, which can be extremely time consuming.

Auditing and Logging

- Automatically captures detailed audit logs to track data access and usage. These logs support internal security reviews and play a critical role in meeting regulatory compliance requirements such as GDPR and HIPAA by providing traceability and accountability for data-related operations.

Data Lineage

- Automatically captures end-to-end data lineage across tables, views, columns, and other key data assets—tracking how data moves and transforms from initial ingestion through every stage of processing to its final consumption. This lineage is collected in real time and includes connections to notebooks, jobs, dashboards, and even machine learning models, providing comprehensive visibility into data flows throughout your organization
- Data lineage is like a detailed map of your data's journey—from where it starts, through all the changes it goes through, to where it ends up being used. It shows how data moves, who touches it, and what happens to it along the way. Think of it like tracking a package—you know where it came from, who handled it, what route it took, and when it was delivered. Being able to trace data lineage offers tremendous benefits, particularly with respect to data quality. Poor quality data leads to inaccurate analytics, poor decision-making and cost overhead.
- Lastly, data lineage also plays a role in compliance. For example, it helps prove where data in reports came from—important for meeting rules like GDPR or HIPAA. It saves time by avoiding manual tracking during audits.

The powerful capabilities Unity Catalog offers make it clear its value as a data governance solution, reducing complexities in managing your data landscape. Infinitive can help realize this potential for your organization by charting the course and performing your migrating to Unity Catalog.

OUR MIGRATION APPROACH

We tailor our Unity Catalog migration approach to clients' unique needs, adapting it according to the organization's technical environment and data governance requirements.

Phase 1: Assess & Plan	Phase 2: Setup & Design	Phase 3: Develop & Migrate	Phase 4: Test & Validate	Phase 5: Go Live & Handoff
Conduct discovery to gather technical and functional requirements. Run pre-migration assessment. Review catalogs, workspaces and access control patterns. Tailor the migration plan.	Setup and configure the Unity Catalog at the Account level. Document the current state, objects to be migrated, and design the Unity Catalog architecture. Create metastore.	Create catalog and schemas. Sync legacy metadata, and upgrade tables and views, then create access controls. Start with one workspace and validate, then repeat for other workspaces.	Test thoroughly to validate successful object migration. Test access controls. Create go live plan, checklist, and rollback options.	Perform the migration in production. Finalize and review documentation. Do knowledge transfer to ensure a smooth transition and handoff.

Assess & Plan

Getting Started

- To begin, we identify and meet with your key stakeholders, as well as key technical personnel (Databricks administrators, cloud administrators) who will be essential and instrumental throughout the migration process. This core group of people will not only provide the technical, functional, and governance requirements needed, but they play a crucial role in reviewing and signing off on phases and tasks as we go through the entire migration process.

Discovery & Requirements Gathering

- Next, we gather details about your technical environment, including which cloud provider you're using and other considerations relevant to the migration process. In addition, we gather functional and governance requirements related to access controls, regulatory compliance, privacy considerations, and more.

Pre-migration Assessment

- Once we have gathered requirements, we begin the detailed planning of Unity Catalog migration by using automated tools to conduct a comprehensive assessment to understand the current state of your data environment and identify any potential challenges. This step helps inform the migration strategy and ensures a smoother transition.

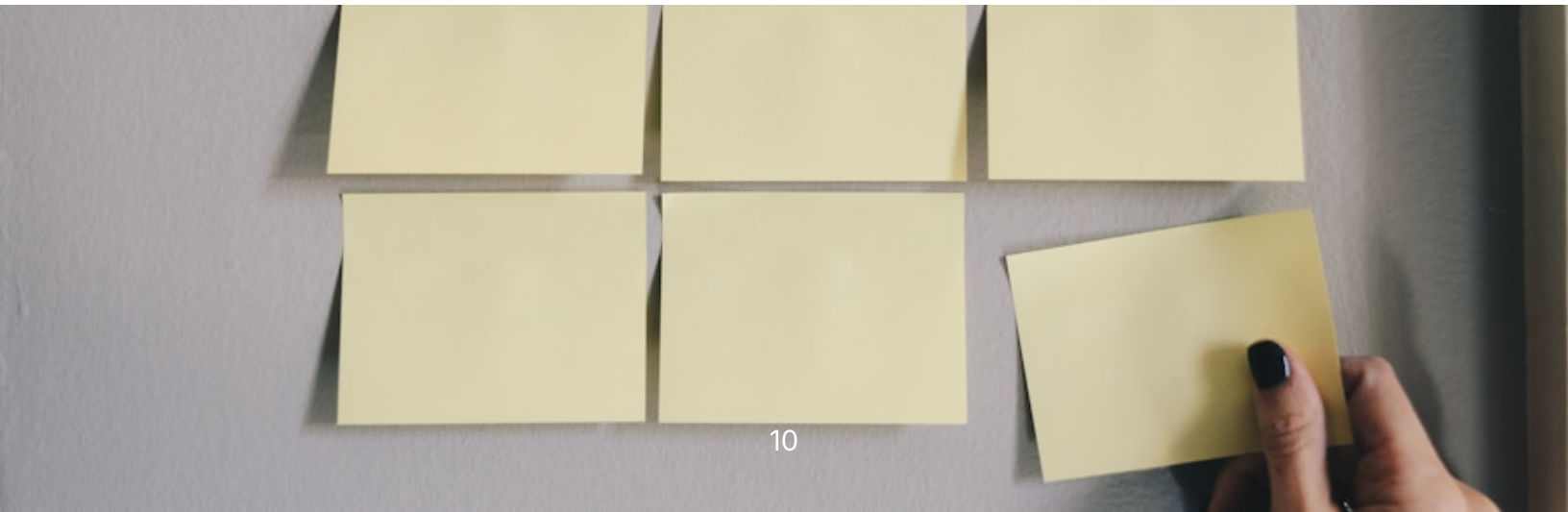
Based on the pre-migration assessment, we go through the following review process:

- **Review Existing Catalogs & Workspaces:** Evaluate all current workspaces, catalogs, and schemas (databases) across your Databricks environment. Document the current structure, usage patterns, data volume, and critical dependencies. Identify any overlapping or redundant assets and assess how they are currently organized and accessed. This will help in mapping existing resources to the new Unity Catalog structure and determining opportunities for consolidation or cleanup. Furthermore, it becomes part of the baseline we will use for comparison during the test and validation phase to ensure completeness.

- **Review Access Control Patterns:** Analyze current access control configurations, including workspace-level permissions, table- or schema-level ACLs, cluster policies, and any role-based access control (RBAC) implementations. Pay close attention to how users, groups, and service principals are granted access to different assets. This review helps identify which permissions need to be replicated or restructured in Unity Catalog, which uses a centralized, fine-grained access control model based on metastore-level governance.

Tailor Migration Plan

- Lastly, based on the assessment results, technical environment details, and requirements we then tailor the migration plan accordingly. We review with stakeholders to ensure alignment and understanding of the scope, and we define the success criteria.



Setup & Design

Account-Level Setup for Unity Catalog Migration

- Before migrating any data or workloads, a proper account-level setup is essential to enable Unity Catalog. This foundational step establishes the governance and security framework that Unity Catalog uses to manage data assets centrally across workspaces. This account-level setup lays the groundwork for a secure, scalable, and compliant Unity Catalog environment across your data platform.

Architecture Design

- Designing a Unity Catalog architecture before migration is a critical planning phase that sets the foundation for a successful rollout. It involves defining how data, users, and access controls will be structured and managed under Unity Catalog's centralized governance model.

Create Metastore

- Creating a metastore in Unity Catalog is like establishing a central hub for all your data and AI assets in Databricks. This central repository organizes your data resources—such as tables, views, volumes, and machine learning models—across multiple workspaces within a cloud region236. Once set up, the metastore enables you to centrally secure and manage these assets, enforce consistent data access policies, implement fine-grained access controls, streamline collaboration and data sharing, and strengthen the overall security posture of your data platform

Develop & Migrate

In this phase, the rubber really hits the road.

Creating catalogs and schemas

- Think of *catalogs* and *schemas* like folders and subfolders that help organize your data. In this step, you're setting up a clean, organized structure in Unity Catalog to hold all your data (structured and unstructured), ML models and experiments, and user-defined functions.

Syncing legacy metadata

- This means bringing over information about your existing data—like names, descriptions, and where things are stored—from your old system into Unity Catalog. It’s like copying the labels and tags from your old folders so they stay consistent in the new system.

Upgrade tables and views

- Your actual data (tables) and the ways you look at it (views) need to be updated so they work smoothly in Unity Catalog. This step is about converting or reformatting them to fit the new setup. Imagine switching your documents from an old file format to a newer, more compatible one.

Grant access control permissions

- Now that your data is moved and organized, you decide who can see or use what. You’re setting rules so only the right people or teams can access specific data. Like giving keys to the right people for the right rooms in a new office building.

The above steps are common to virtually all Unity Catalog migrations. Depending on the scope of the migration, other steps may be needed. However, these would be already identified and part of the tailored project plan.

Test & Validate

Thorough testing and validation are crucial to confirming a migration is successful. Our test and validation process includes the following steps, each of which includes a review with business stakeholders (and/or direct involvement as needed) to get buy-in.

Validate a successful object migration to Unity Catalog

- This is crucial to ensure data integrity, access control consistency, and functionality post-migration. To do this, we compare the inventory of Unity Catalog objects to the baseline.

Validate data

- Ensure that the data itself has been preserved accurately during migration. This is accomplished by doing row count validations, sample data checks, checks for unexpected changes in data completeness. If it applies, partitioning structures are also validated.

Validate managed and external tables

- Ensure that both managed and external tables are correctly linked to their designated cloud storage locations. Verify that storage paths have the appropriate permissions assigned to Unity Catalog's service principles and confirm that external locations and storage credentials are properly configured and accessible.

Validate access controls

- Verify that users and groups retain appropriate access to catalogs, schemas, and objects. Test permissions for common roles (e.g., data analyst, data engineer): Can they read the data they need? Are unauthorized users blocked?
- If row-level or column-level security has been applied, such as tag-based access policies, then confirm these controls function as expected.

Verify governance policies are enforced correctly

- Confirm any additional specifics related to governance policies, per the functional requirements gathered in the Assess & Plan phase, are completely enforced.

Audit and logging validation

- Confirm audit logging is enabled for Unity Catalog actions, and validate that important events (e.g., access attempts, object changes) are logged correctly. Review logs for unexpected errors or denied access.

Do additional validation steps, if needed

- Depending on the scope and details of your migration, there may be additional steps. They may include verifying query and pipeline functionality, ensuring they run and function correctly with Unity Catalog object references. Key projection queries may need to be run to validate performance and results. Any streaming pipelines, Delta Live Tables, and ML flows should be tested if they rely on migrated data.

Prepare Go Live Plan

- Finally, thorough preparations prior to deploying in a production environment are essential. We create a comprehensive plan to go live, ensuring a smooth rollout to a production environment. This includes a checklist of steps to follow, and part of the intent is to anticipate anything that could go wrong and minimize potential risks. However, in the event of an unforeseen issue, we also develop rollback options to minimize or eliminate any risk of disruption to the production environment.

Go Live & Handoff

Production Migration

- At this final stage, we go through the go live checklist and then migrate to the production environment, verifying all is working as expected.
- What a successful migration looks like:
  - All critical data assets are present and accurate
  - Access control and governance policies are enforced correctly
  - No broken jobs, pipelines, or dashboards
  - Performance is consistent or improved
  - Stakeholders confirm usability and trust in the data

Documentation

- Upon successful migration, we finalize all documentation and compile all artifacts produced along the way into one package, which is then reviewed with all stakeholders and those who will be maintaining and working with Unity Catalog going forward.

Handoff

- Lastly comes the handoff, confirming with all stakeholders that we have delivered a successful Unity Catalog migration, and they are prepared with the resources they need going forward to use, manage, and maintain the system.

CONCLUSION

As we've detailed in this e-book, the greatest challenges to data governance across your organization's data landscape can be streamlined and simplified with Unity Catalog. Its potent capabilities – from access controls and data discovery to lineage, auditing, and more -- can help you get the most out of your data assets now and in the future.

We at Infinitive are truly excited to offer you our expertise and comprehensive migration methodology to bring the power of Unity Catalog to your organization's data landscape.