Deloitte. | databricks

# MIGRATING FROM HIVE METASTORE (HMS) TO DATABRICKS UNITY CATALOG

## OVERVIEW OF HIVE METASTORE

Hive Metastore (HMS) is a centralized metadata repository used by Hadoop-based systems to manage metadata about tables, partitions, and data schemas.

## OVERVIEW OF DATABRICKS UNITY CATALOG

Databricks Unity Catalog provides a unified and open governance solution for managing all data and AI assets. As the cornerstone of your data intelligence strategy, Unity Catalog combines the power of Lakehouse and AI to build a deep understanding of your data and deliver contextual, domain-specific insights that boost productivity for both technical and business users across any workload. With an open source foundation, Unity Catalog enables discovery, access, and sharing of trusted data and AI assets across any tool, compute engine, or cloud platform. This unified, open approach fosters collaboration, accelerates data and AI initiatives, and simplifies compliance in an ever-evolving data landscape.

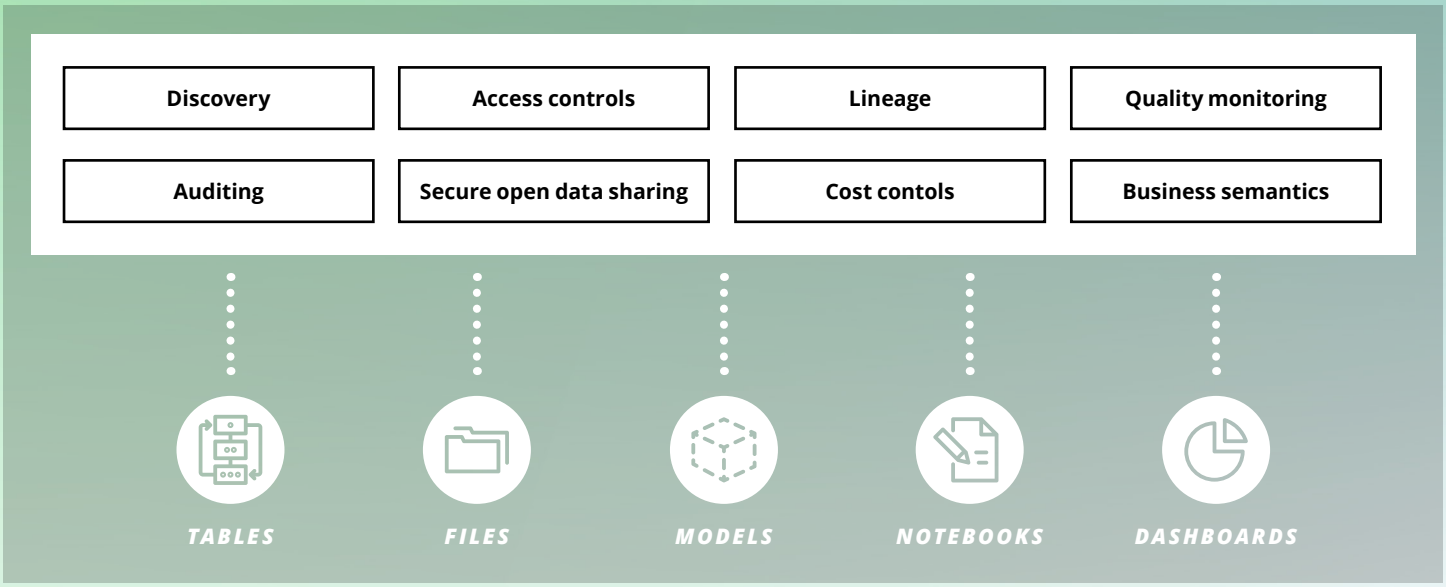
## KEY BENEFITS OF UNITY CATALOG INCLUDE:

**Unified governance for data and AI**
Build an enterprise catalog for all structured and unstructured data, ML models, AI tools, notebooks, metrics with unified governance. Leverage any open data format of your choice, including Delta, Iceberg, and Parquet. This unified approach brings together access management, cataloging, monitoring, and lineage tracking within a centralized platform, offering seamless visibility across all assets.

**Open connectivity to any tool, engine or platform**
Connect to any data source and create a single source of truth across data lakes, databases, data warehouses and catalogs. Access data and AI assets from any compute engine or tool of your choice with Open APIs. Share data and AI assets across platforms, regions, and clouds to drive collaboration at scale.

**Built-in data intelligence for all users across any workload**
Enhance clarity and understanding with AI-powered comments and tags while boosting productivity through context-aware search and auto-generated data insights. Accelerate workloads with an intelligent assistant that delivers domain-specific intelligence for any user and any workload. Optimize performance and reduce total cost of ownership with AI-driven optimizations, ensuring efficiency across all workloads.

## DATABRICKS UNITY CATALOG KEY CAPABILITIES




**Discovery:** Unity Catalog provides a unified catalog for all files, tables, ML models, AI tools, notebooks, and business metrics. It supports open data formats such as Delta, Iceberg, Parquet, and more. AI-powered comments, tags, context-aware search, and auto-generated data insights help users quickly find and understand data.

**Access controls:** Unity Catalog centralizes access policy management for data and AI assets across all workspaces. It offers fine-grained controls at the row and column level, including attribute-based access controls.

**Lineage:** Captures real-time column-level lineage for data and AI workloads, providing visibility into data flow and dependencies.

**Auditing:** Automatically logs user-level access to data, ensuring comprehensive audit trails.

**Quality monitoring:** Leverages AI-powered quality monitoring for data and AI pipelines, with automated alerts to detect and address quality issues.

**Open access:** Provides open APIs, enabling data access from any tool, engine, or platform for seamless interoperability.

**Open connectivity:** Supports data federation, allowing users to discover, access, and govern data across external databases, data warehouses, and catalogs without duplicating data.

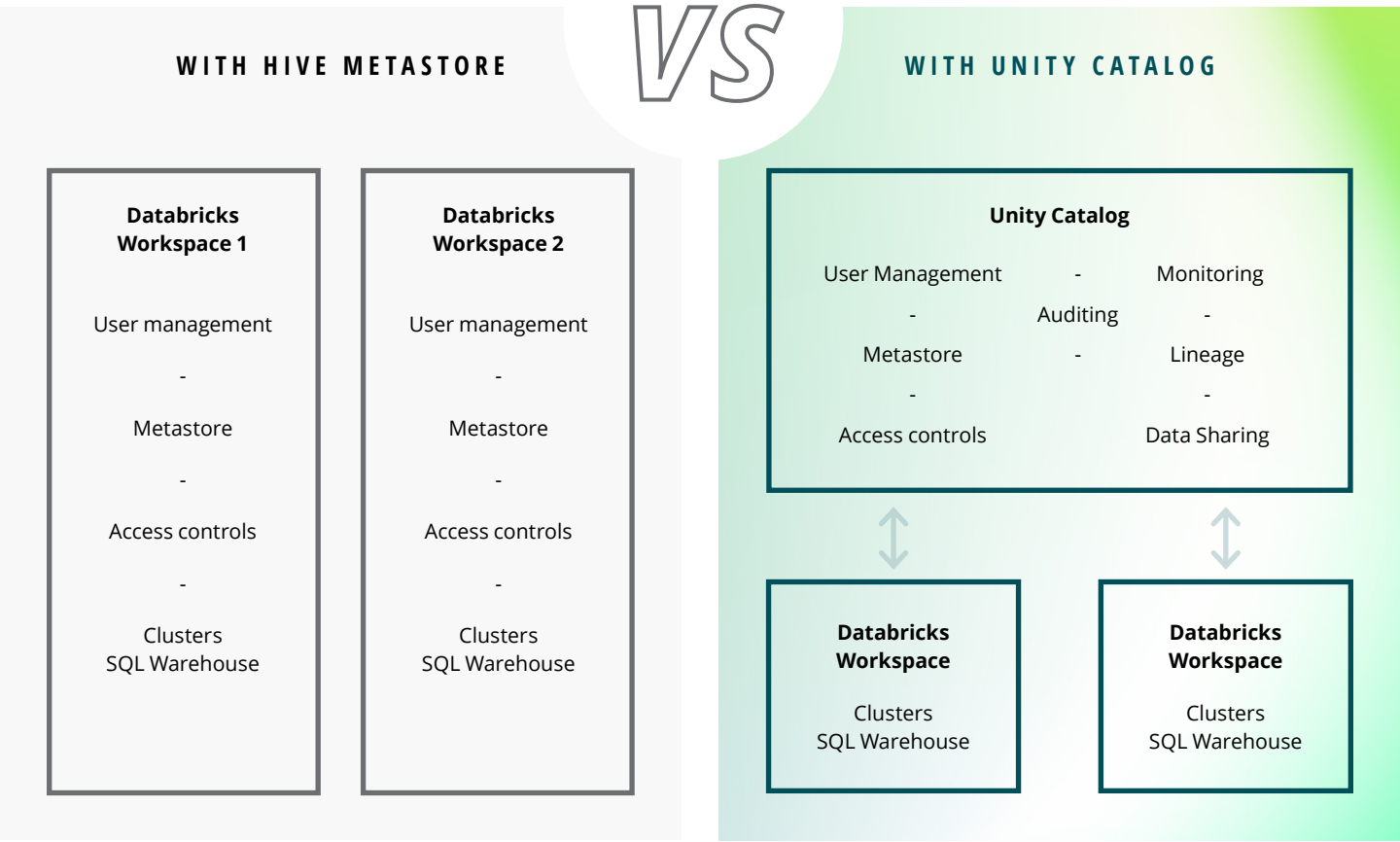
**Open sharing:** Enables secure, cross-cloud and cross-platform data sharing without ETL or data replication, leveraging open-source Delta Sharing.

**Cost controls:** Offers system tables and built-in dashboards to monitor billing, track usage, and maintain cost efficiency.

**Business semantics:** Allows users to define, govern, and access business metrics across data and AI workloads, enhancing the accuracy of data insights.

# BUSINESS REASONS TO MIGRATE TO DATABRICKS UNITY CATALOG

Migrating from Hive Metastore to Databricks Unity Catalog can offer several advantages, particularly for organizations that need to improve their data governance, security, and scalability. Here are some key reasons why organizations might consider this migration:




## UNIFIED GOVERNANCE ACROSS DATA SOURCES:

**HMS**
Primarily designed for managing metadata for data stored in Hadoop environments. It's limited in its ability to provide governance across multiple cloud platforms and storage systems.

**Unity Catalog**
Offers centralized governance across a variety of data sources, including cloud storage systems (e.g., S3, ADLS), databases, and streaming data. This unified governance simplifies managing data across different environments and platforms.

## FINE-GRAINED ACCESS CONTROL:

**HMS**
Provides access control primarily at the database and table levels. More granular access control (e.g., column-level, row-level) can be complex to implement and manage.

**Unity Catalog**
Enables fine-grained access controls at the table, column, and row levels, allowing more precise control over who can access specific data. This feature is essential for maintaining compliance with data privacy regulations.

### *IMPROVED SECURITY AND COMPLIANCE:*

**HMS**
While it supports basic security features, it may require custom implementations to meet advanced security and compliance needs, particularly in multi-cloud environments.

**Unity Catalog**
Provides enhanced security features, including built-in auditing, automated policy enforcement, and integration with identity providers (e.g., Azure AD, Okta) for more secure and compliant data access.

### *DATA LINEAGE AND AUDITABILITY:*

**HMS**
Tracking data lineage and performing audits can be difficult and require custom solutions, which may not scale well as data environments grow.

**Unity Catalog**
Automatically tracks data lineage, showing how data flows through various processes. This built-in feature is crucial for auditing, debugging, and ensuring data quality.

### *SCALABILITY AND PERFORMANCE:*

**HMS**
While HMS works well for Hadoop-based environments, it may struggle with scaling in modern, cloud-based data architectures, especially when dealing with large, distributed datasets.

**Unity Catalog**
Designed to scale with modern cloud architectures. It supports large-scale, multi-cloud deployments, and provides better performance and manageability in distributed environments.

### *COLLABORATION AND PRODUCTIVITY:*

**HMS**
Collaboration in HMS is often manual, with teams needing to coordinate access and governance across multiple tools and environments.

**Unity Catalog**
Integrates seamlessly with Databricks notebooks and other data tools, promoting collaboration and cross-platform sharing of data and AI assets among data engineers, scientists, and analysts. It simplifies governance and improves productivity by maintaining consistent policies across different teams and projects.

### *FUTURE-PROOFING:*

**HMS**
Although still in use, HMS is increasingly seen as a legacy system for managing metadata, especially as organizations move towards more modern data architectures and multi-cloud environments.

**Unity Catalog**
Represents a modern solution designed to address the needs of contemporary data environments. Migrating to Unity Catalog can help future-proof data governance and ensure that organizations are ready for emerging technologies and compliance requirements.

### *SUPPORT FOR ADVANCED USE CASES:*

**HMS**
Limited in supporting advanced use cases, such as real-time data processing, machine learning, and streaming data governance.

**Unity Catalog**
Supports advanced use cases, including real-time data governance, streaming data management, and integration with machine learning workflows on Databricks.

## SUMMARY

Migrating from HMS to Databricks Unity Catalog can help organizations enhance their data governance, security, and scalability while improving collaboration and simplifying compliance. For organizations leveraging modern data architectures or operating in multi-cloud environments, Unity Catalog provides a more robust and future-proof solution compared to HMS.

## HOW UNITY CATALOG CAN UNLOCK VALUE

| Unified visibility into data and AI | Simple permission model for data and AI | AI-powered monitoring and observability | Open data sharing |
|---|---|---|---|
| Accelerate innovation | Reduce platform complexity and cost | Mitigate risk and ensure compliance | Collaborate and monetize the value of data |

## OPTIONS TO MIGRATE TO UNITY CATALOG

There are two primary methods for migrating your catalog: creating a parallel environment to gradually migrate your environment for less business interruption, or an in-place upgrade which requires downtime. Below are the considerations for each method.

| MIGRATION METHOD | CONSIDERATIONS |
|---|---|
| **Option 1: In place upgrade** | • Not yet in production or you are migrating a non-critical business environment<br>• Minimal business users<br>• Downtime is tolerated |
| **Option 2: Build parallel environment** | • Existing environment is business critical and must remain in production during build<br>• Many business users<br>• Downtime of 1+ day(s) will impact SLAs |

## HMS TO UC MIGRATION STEPS

There are multiple phases involved in HMS to UC migration.

### STEP 1 DISCOVERY AND ASSESSMENT

**Review the inventory of objects in HMS provided by UCX:** Analyze the existing metadata and data setup.

**Backup existing data:** Ensure backups are created for data and metadata.

Review the current HMS configuration, including table schemas and metadata. **Leverage Databricks provided UCX utility to get the inventory of the existing HMS setup.**

### STEP 2 PLANNING

**Capture requirement:** Leverage the Upgrade plan template from Databricks to capture the customer requirement.

**Gain approval:** Get the necessary approvals for the project.

Who's approving your upgrade plan?
Typical stakeholders include:

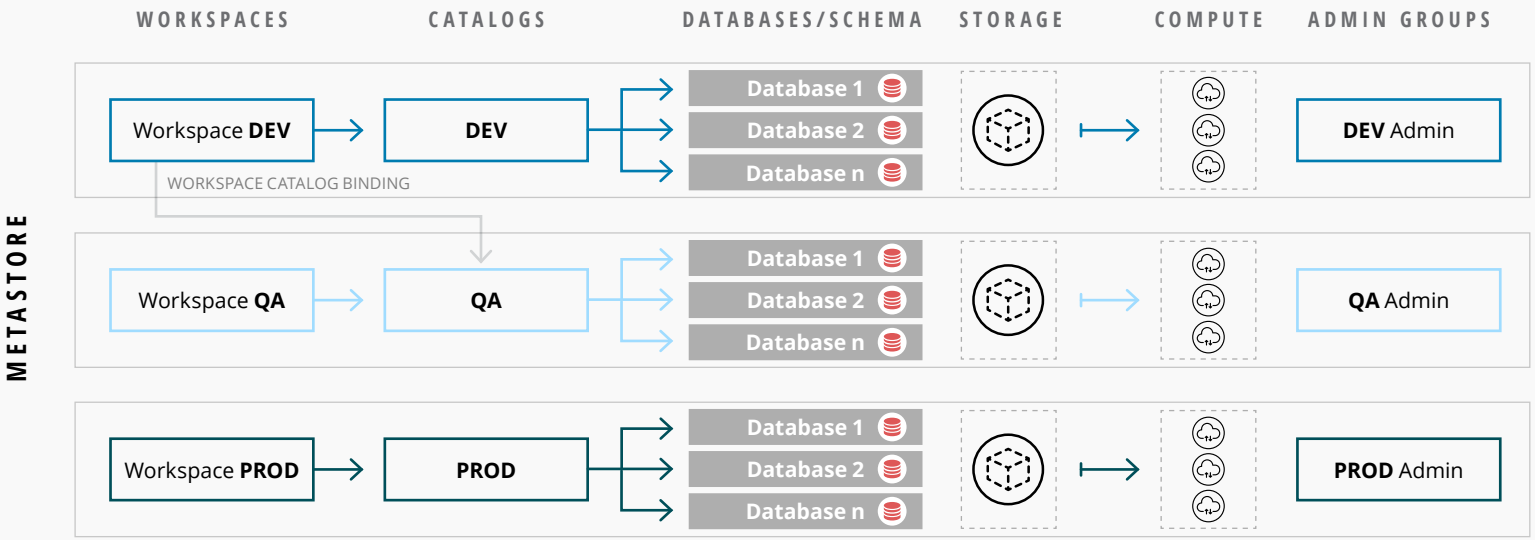
- Workspace owner
- Data Platform owner
- Business/Data Steward
- IT Owner
- CISO

Be prepared to speak to the implementation, risks, rollback options, and the benefits.

## STEP 3

### *UC DESIGN*

Hierarchy of objects in UC.




## STEP 4

### *CREATE FOUNDATIONAL OBJECTS FOR UC*

**Create metastore** → **Catalog** → **Credentials** → **External location** → **Delta share** → **Connection**

## STEP 5

### *MIGRATION OF WORKSPACE OBJECTS*

1. **Upgrade tables & views**
2. **Upgrade groups**
3. **Job upgrade**
4. **Mount point upgrade**
5. **Upgrade compute sources**
6. **Delta Live Tables Pipeline upgrade**
7. **INIT Script upgrade**
8. **Databricks SQL upgrade**
9. **Migrate grants on groups**

# *MIGRATION STRATEGY*

Our automation-led approach mitigates risk and accelerates delivery, which can lead to significant cost reduction and faster realization of benefits while minimizing the risk of human error.

**Define migration scope**
Identify which metadata and data need to be migrated.

**Identify dependencies**
Recognize any dependencies or integrations that could impact migration.

| PHASE | JOURNEY MILESTONE | DIGITAL ASSETS |
|---|---|---|
| **Upgrade planning and strategy** | **Plan and analyze**<br>Perform Current State analysis and create data migration plan, address legacy challenges etc.<br>**Define upgrade approach and design**<br>Choose the right approach for the upgrade – in place vs close<br>**Infra. setup**<br>Setup cloud resources required for the upgrade such as Metastore and other resources (Infrastructure & tools) |  <br>DATABRICKS UCX TOOL |
| **Execution** | **Upgrade tables**<br>Leverage the Deloitte automated table upgrade accelerator to speed up the table upgrade process<br>**Upgrade notebooks**<br>Refactor the notebooks code to include 3 level namespace, external location and libraries<br>**Upgrade other objects**<br>• Upgrade the clusters<br>• Upgrade groups and migrate the permissions<br>• Upgrade the jobs, Init scripts and cluster policies |  <br>DELOITTE UPGRADE ACCELERATOR |
| **Validation** | **Parallel runs**<br>• Execute parallel runs between legacy hive Meta store and new Unity catalog<br>• Conduct reconciliation testing and certify migrated Data<br>• Leverage the validation accelerator<br>**Post go live support**<br>• Debug and fix issues after go-live<br>• Knowledge transfer and hand-over |  <br>DELOITTE VALIDATION ACCELERATOR |

## *CONCLUSION*

Migrating from Hive Metastore (HMS) to Databricks Unity Catalog presents a transformative opportunity for organizations seeking to enhance their data governance, security, and scalability. Unity Catalog offers a comprehensive solution that centralizes metadata management, enables fine-grained access control, and ensures robust compliance with regulatory requirements. Its support for advanced features such as data lineage tracking, automated policy enforcement, and multi-cloud integration makes it a future-proof choice for modern data environments.

The migration process, while complex, can be streamlined through careful planning, utilizing tools like the UCX utility for assessment and planning, and adopting an automation-led approach to mitigate risks provided by Deloitte. By following a structured migration strategy that includes evaluating the existing HMS setup, defining the migration scope, and identifying dependencies, organizations can seamlessly transition to Unity Catalog while maximizing the value of their data assets.

With Unity Catalog, businesses can future-proof their data governance strategies, improve collaboration, and unlock the full potential of their data in a secure, compliant, and scalable environment. This migration marks a significant step forward in transforming how organizations manage and secure their data, providing a solid foundation for innovation and growth in the data-driven era.

## *READY TO GET STARTED?* ***CONTACT US TODAY***


**Mani Kandasamy**
AI & Data Engineering
Technology Fellow
Deloitte Consulting LLP
mkandasamy@deloitte.com

**Vamsi Vangala**
Strategy & Analytics
Senior Manager
Deloitte Consulting LLP
vavangala@deloitte.com

**Syed Zaheerulla**
Strategy & Analytics
Delivery Manager
Deloitte Consulting LLP
szaheerulla@deloitte.com

**Emily Cole**
Databricks Alliance Manager
Deloitte Consulting LLP
emcole@deloitte.com

**Naveen K**
DC Specialist Senior
Deloitte Consulting LLP
naveenk2@deloitte.com