Databricks Unity Catalog

Exploring Data Management and Governance Strategies

MC ID: 7124506 | Parveen KR | Parveen.R@hotmail.com

Agenda

- Data Management and Governance with Unity Catalog
- Understanding Databricks Unity Catalog
- Planning and Implementing Data Governance
- Encryption and Data Security
- Security and Access Control
- Fine-Grained Access Control
- Databricks Marketplace and Data Lifecycle

Data Management and Governance with Unity Catalog

MC ID: 7124506 | Parveen KR | Parveen.R@hotmail.com

Why Data Governance is Critical

Ensures Data Quality & Trust

Reliable data is essential for providing accurate insights and informed decision-making across an organization.

Maintains Security & Compliance

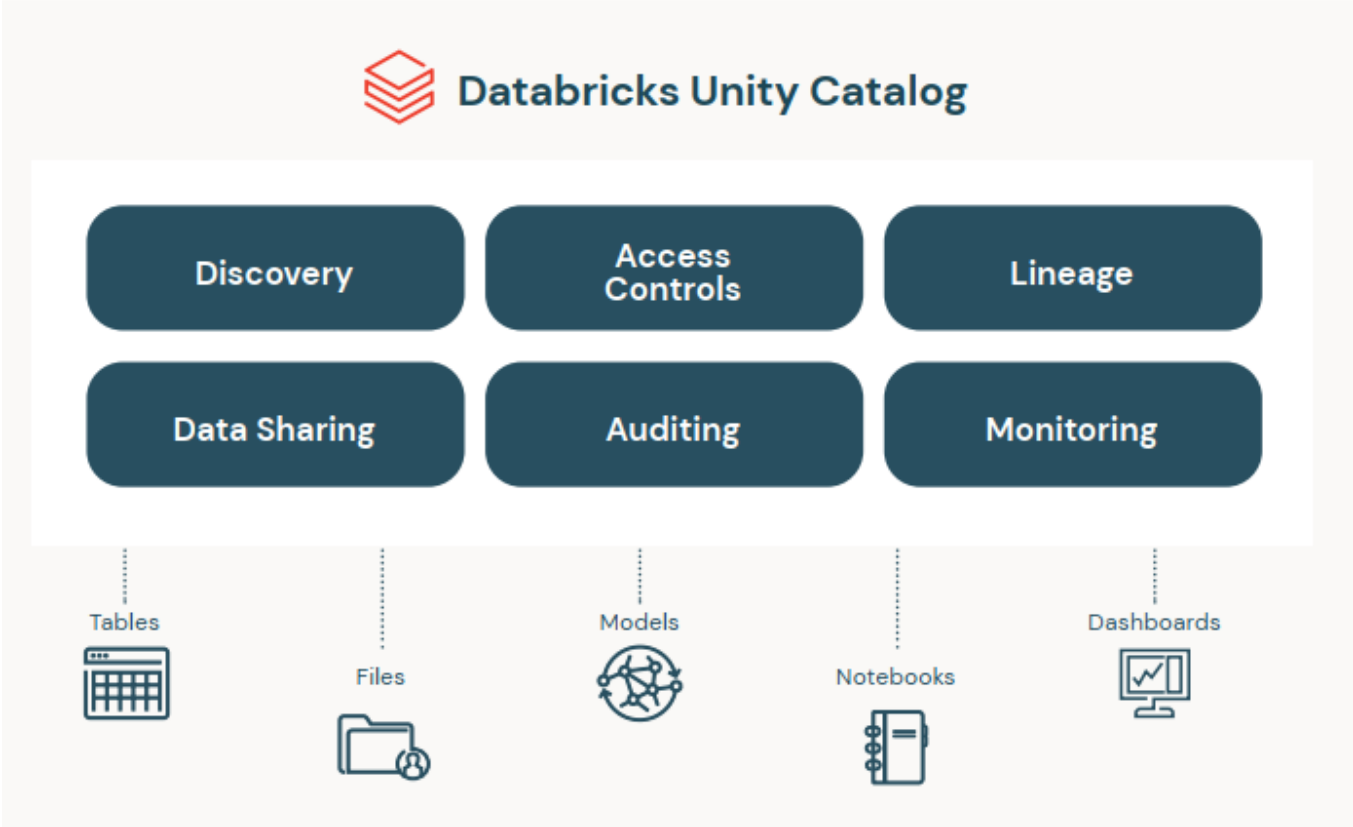
Effective data governance protects sensitive data and ensures compliance with regulations like GDPR and HIPAA.

Reduces Risk

A robust data governance strategy minimizes the risks of data breaches, unauthorized access, and misuse.

Improves Efficiency & Collaboration

Data governance promotes efficiency by breaking down silos and enabling self-service analytics for better teamwork.



Understanding Databricks Unity Catalog

MC ID: 7124506 | Parveen KR | Parveen.R@hotmail.com

What is Databricks Unity Catalog?

Governance Solution

Databricks Unity Catalog provides a fine-grained governance solution for managing data and AI assets effectively.

Centralized Metadata Management

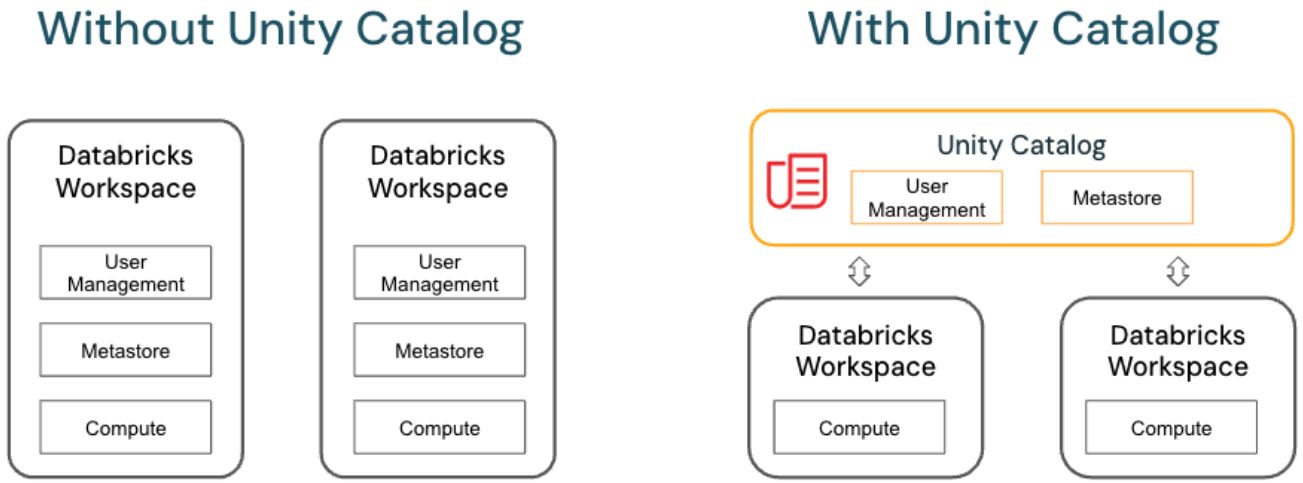
The catalog offers centralized metadata management, allowing for organized and efficient data handling across workspaces.

Access Control and Security

Access control and security are based on ANSI SQL, ensuring safe and compliant data access.

Automated Data Lineage

Automated data lineage tracking helps organizations understand data flow and transformations across systems.



Understanding the Unity Catalog Hierarchy

Metastore

The metastore is the top-level container for all metadata in the Unity Catalog, organizing all data assets effectively.

Catalog

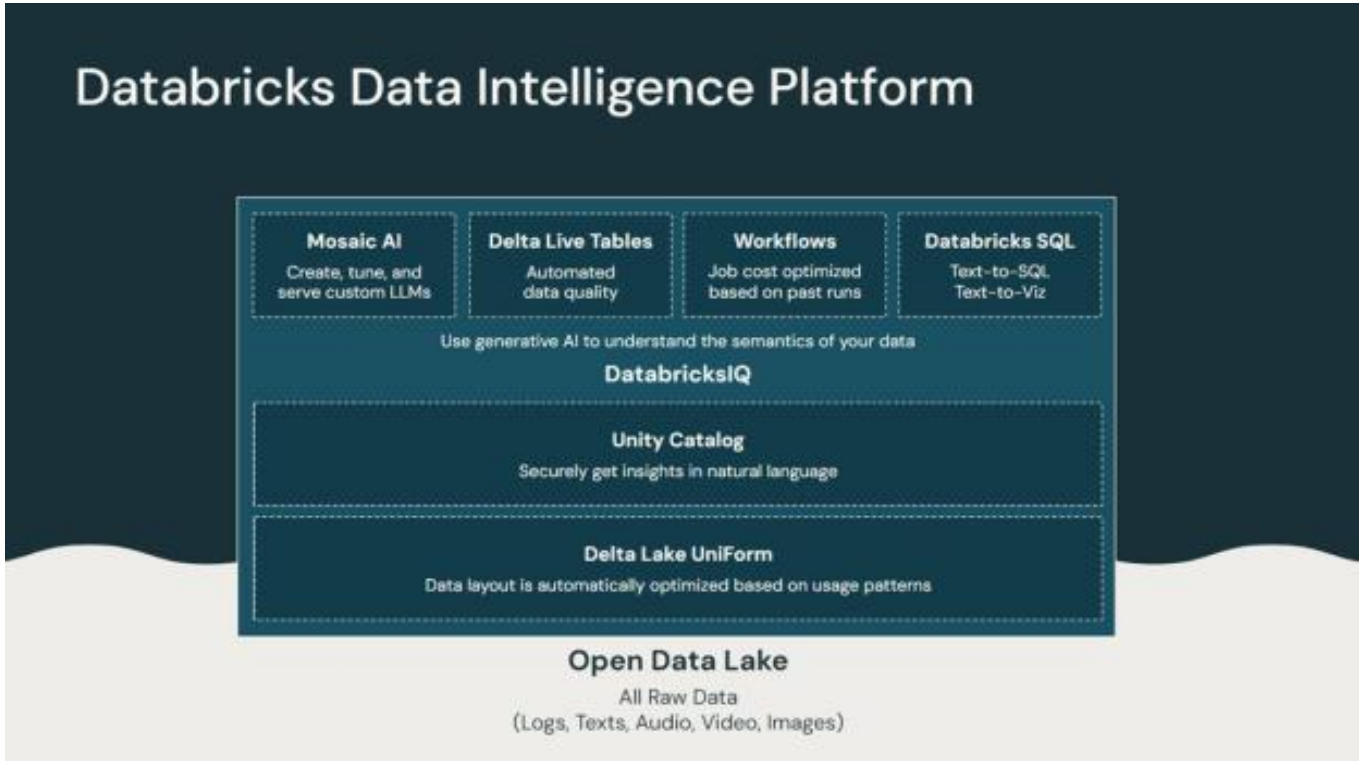
The catalog serves as the first layer of data organization within a metastore, grouping related schemas and assets.

Schema (Database)

Schemas are the second layer of data organization, contained within a catalog, facilitating structured data management.

Data Assets

Data assets encompass tables, views, volumes, functions, and models, representing the various forms of data within the catalog.



Patterns for Organizing Your Data Assets

Data Organization by Environment

Organize data assets based on the environment, such as development, staging, and production catalogs.

Data Organization by Business Unit

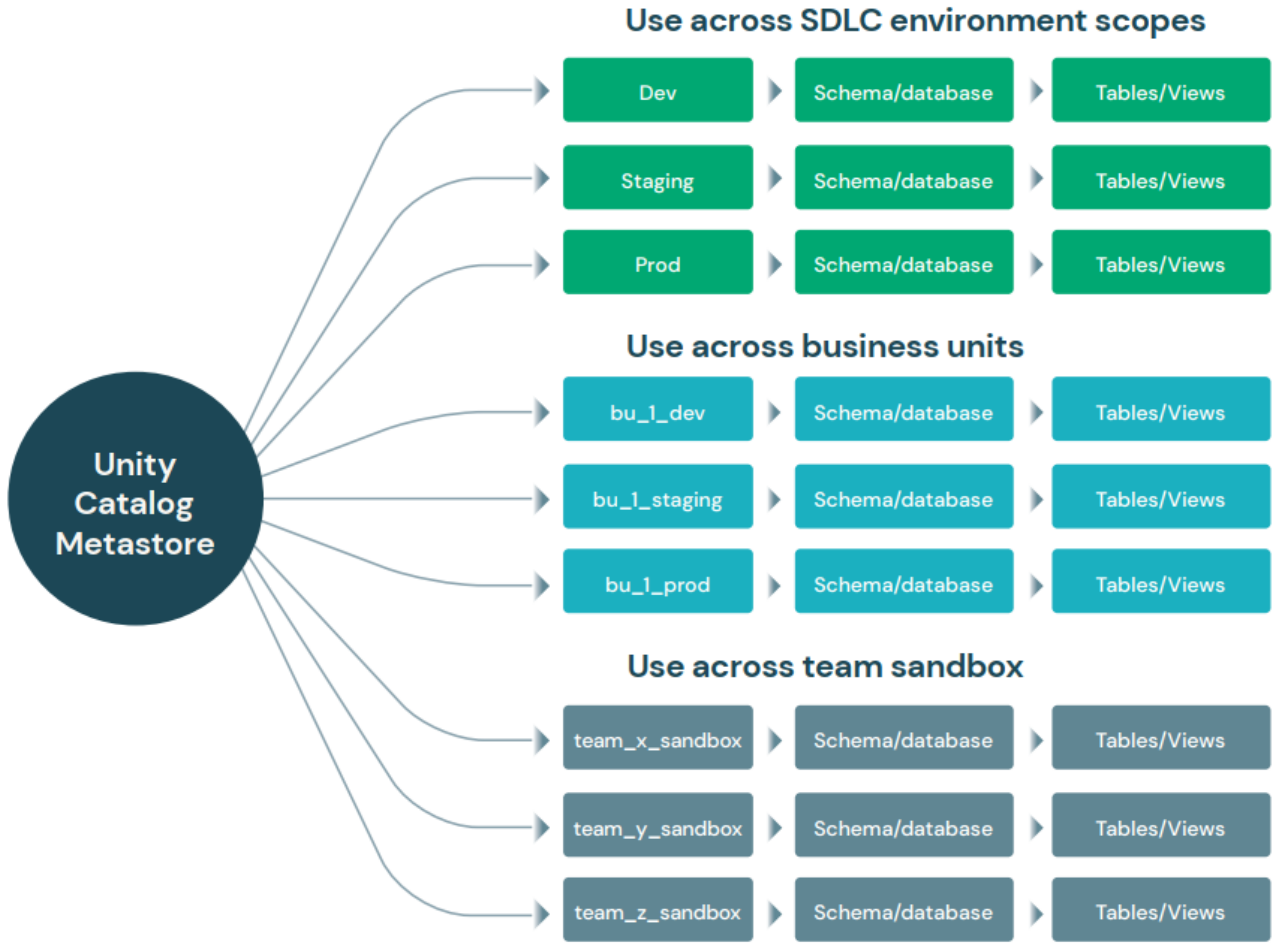
Classify data assets according to business units or departments like sales, marketing, and finance.

Data Organization by Source System

Structure data assets based on their source systems, such as CRM, ERP, or IoT platforms.

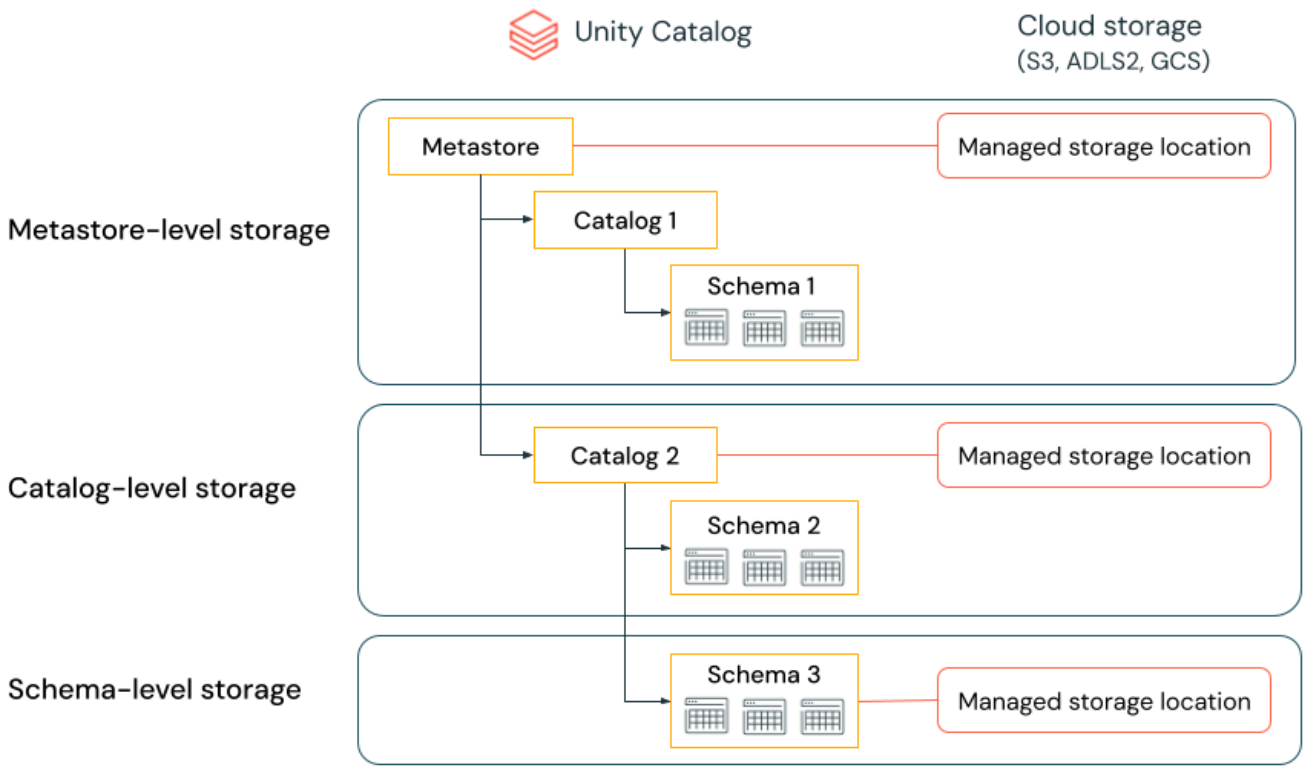
Hybrid Data Organization Approach

Combine elements from different patterns to create a hybrid approach for data organization.



Planning and Implementing Data Governance

Planning Your Data Governance Strategy



Define Roles & Responsibilities

Clearly defining roles such as Metastore Admins and Data Stewards ensures accountability in data management practices.

Establish Naming Conventions

Standardized naming conventions for catalogs, schemas, and tables facilitate better data organization and discovery.

Develop a Security Model

A robust security model must define who has access to specific data, ensuring data protection and compliance.

Plan for Data Quality

Establish processes to monitor and ensure data quality, which is crucial for reliable data governance.

Benefits of Moving to Unity Catalog

Centralized Governance

Unity Catalog provides a single platform for managing all data and AI assets, enhancing governance.

Fine-Grained Security

It ensures consistent access controls, providing fine-grained security across all data assets.

Automated Lineage

Automated lineage helps users understand data dependencies, streamlining data management processes.

Enhanced Discovery

Unity Catalog enhances data discovery, making it easy to find and understand data assets.

Encryption and Data Security

MC ID: 7124506 | Parveen KR | Parveen.R@hotmail.com

Securing Your Data: Encryption Fundamentals

Importance of Encryption

Encryption serves as a critical layer to protect sensitive data from unauthorized access and breaches.

Data at Rest

Protecting stored data in cloud environments (e.g., S3, ADLS) ensures that information remains safe when not in use.

Data in Transit

Safeguarding data as it moves between systems prevents interception and unauthorized access during transmission.

Integration with Databricks

Databricks enhances data security by leveraging cloud provider capabilities for encryption, offering additional features.

Unity Catalog Governance

Unity Catalog works alongside encryption methods to manage access to protected data, ensuring secure data governance.

Protecting Stored Data: Encryption at Rest

Cloud Provider Encryption

Cloud providers offer server-side encryption (SSE) to automatically encrypt data before storage. This ensures data security without user intervention.

Customer-Managed Keys

Using customer-managed keys (CMK) allows for greater control over encryption for Databricks workspace storage. This enables tailored security measures.

Encryption for Managed Services

CMK can also encrypt data in managed services, including notebook files, secrets, and SQL queries, enhancing overall data security.

Advanced Encryption: Application-Level & Key Handling

Application-Level Encryption

Implement encryption directly within applications for highly sensitive data control, such as PII columns. Ensure performance and implementation complexity are considered.

Key Management Practices

Never hardcode encryption keys. Use secure storage solutions like Databricks Secrets, Azure Key Vault, and AWS Secrets Manager.

Key Rotation Strategy

Implement a strategy for key rotation, managing multiple key versions, and re-encrypting data to maintain security.

Security and Access Control

Unity Catalog's Approach to Security

Centralized Administration

Unity Catalog allows defining access policies once for seamless application across all workspaces, enhancing management efficiency.

Standards-Compliant Security

Utilizes ANSI SQL syntax with `GRANT` and `REVOKE` statements, making it user-friendly for database administrators.

Hierarchical Permissions

Supports granting privileges at various levels, including Metastore, Catalog, and Schema, providing flexible security control.

Ownership Model

Every securable object has an owner with full privileges, ensuring clear accountability and control over access rights.

Core Principles for Secure Data Access

Principle of Least Privilege

Grant users only the minimum permissions required for their tasks to enhance security.

Ownership Management

The creator of an object is its initial owner. Ownership can be transferred for flexibility.

Group Permissions Management

Granting permissions to groups simplifies management and enhances security across teams.

Types of Privileges You Can Grant

Catalog Privileges

Privileges for catalogs allow users to select and create schemas, enhancing data organization.

Schema Privileges

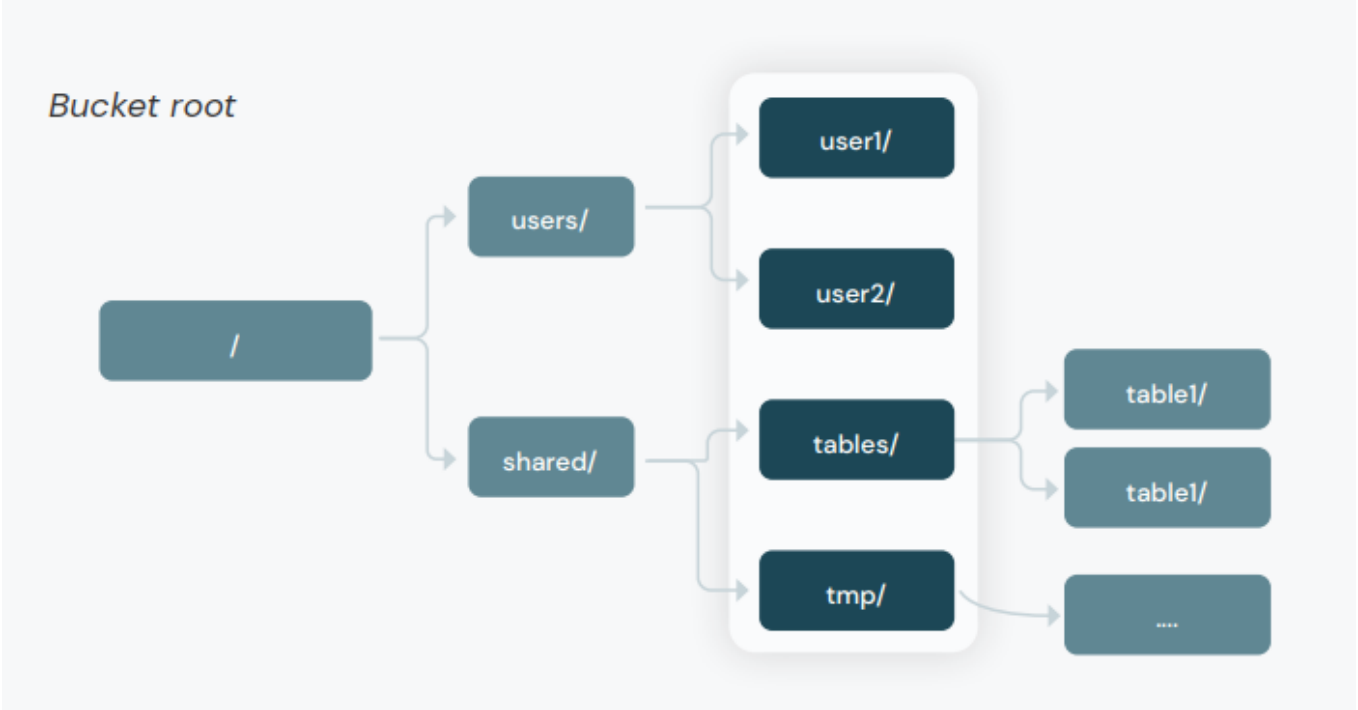
Schema privileges enable users to manage tables, views, and functions within the specified schema.

Table and View Privileges

Privileges related to tables and views allow users to read, modify, and tag data as needed.

Volume Privileges

Volume privileges enable users to read, write, and create external tables from files, enhancing data accessibility.



Fine-Grained Access Control

The Mechanics of Access Control: GRANT'' & REVOKE

Understanding GRANT Syntax

The `GRANT` command assigns specific permissions to users or groups for various objects in the database.

Understanding REVOKE Syntax

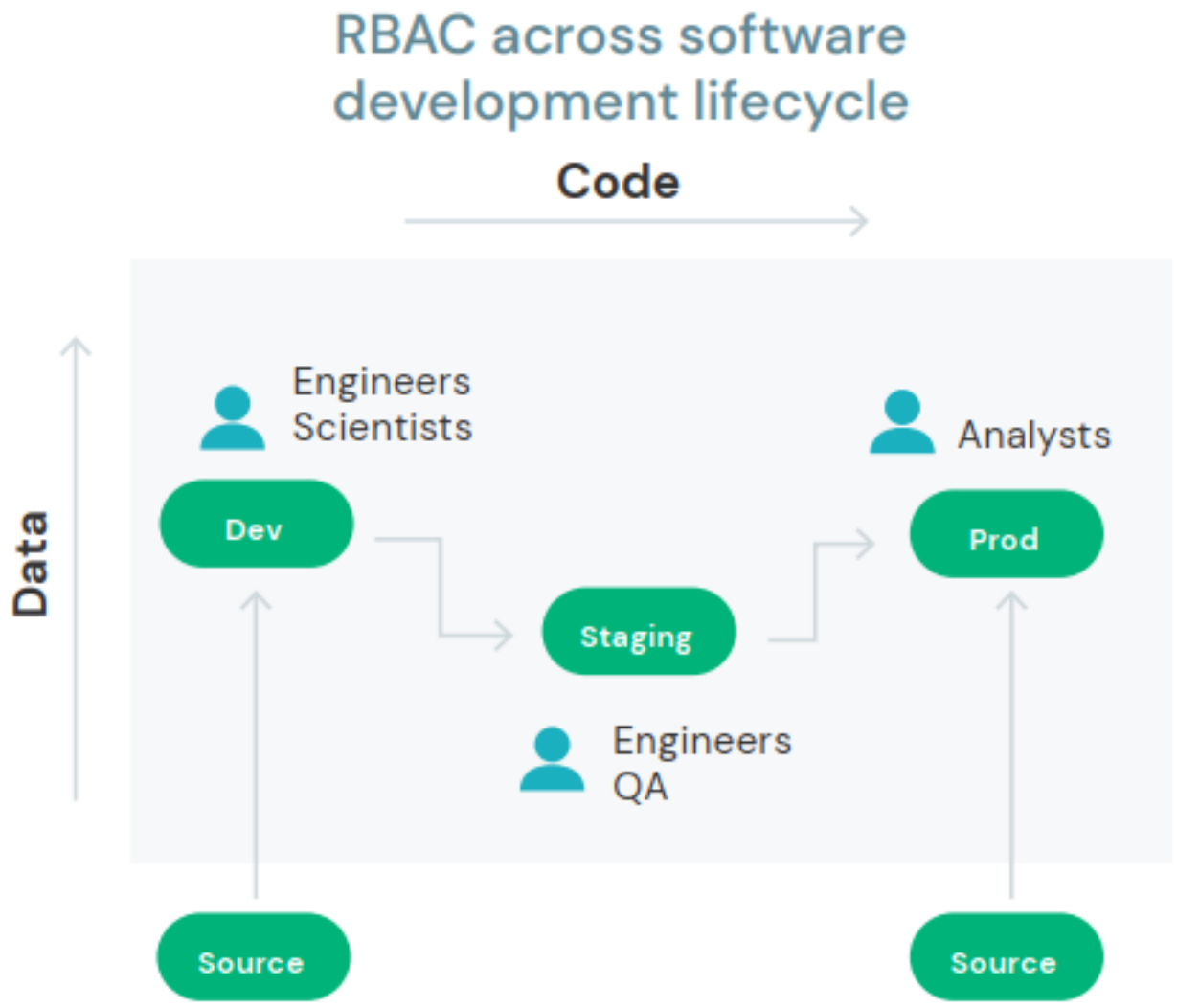
The `REVOKE` command removes specific permissions from users or groups, managing access control effectively.

Executing Access Control Commands

Access control commands can be executed via SQL editor, notebooks, or Databricks CLI/API.

Viewing Existing Permissions

Use `SHOW GRANTS ON



Fine-Grained Access Control: Column & Row Filters

Column Masks

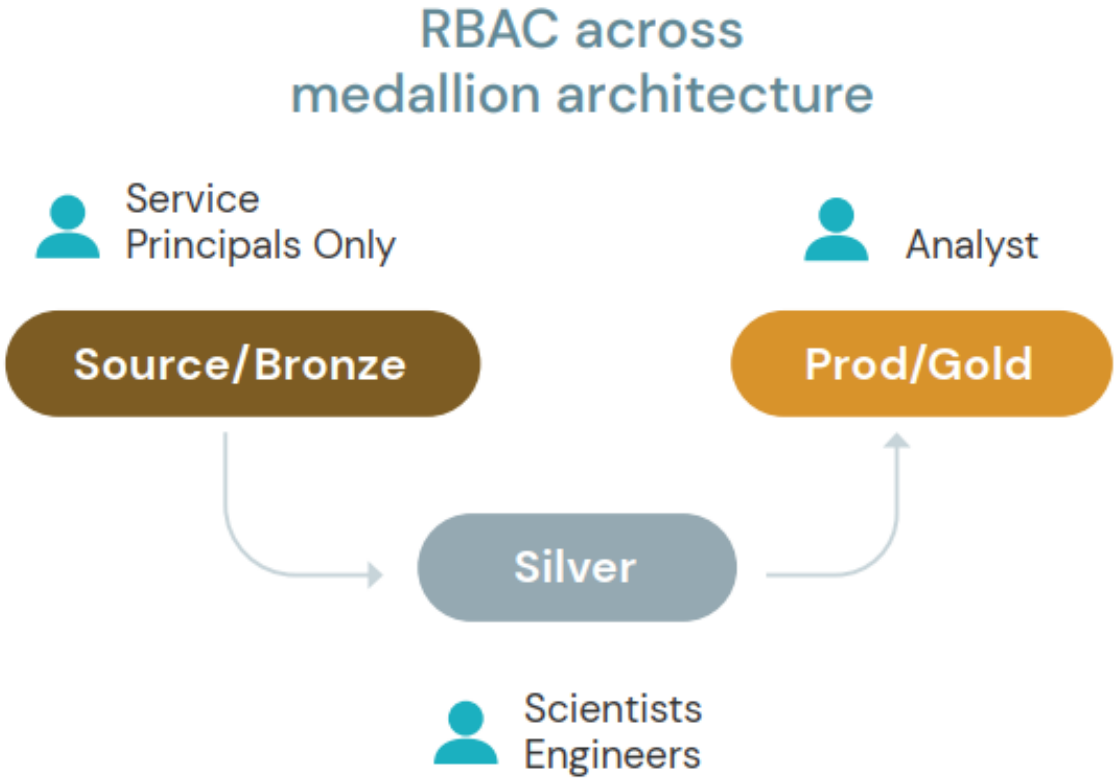
Column masks redact or tokenize sensitive data like PII and credit card numbers, ensuring privacy during data access.

Row Filters

Row filters limit user visibility to certain rows in a table, based on defined criteria and user permissions.

Data Privacy and Security

Fine-grained access control enables data privacy and security policies without creating multiple data versions, enhancing data management.



Administering Unity Catalog: Roles and Responsibilities

Metastore Admins

Metastore Admins possess the highest privileges, managing Catalogs and securing access across the Metastore.

Catalog Owners

Catalog Owners manage specific Catalogs and have the ability to create Schemas and grant associated permissions.

Schema Owners

Schema Owners focus on managing tables and other objects within their assigned Schemas, ensuring proper permissions.

Data Consumers

Data Consumers require access to read and analyze data, typically utilizing SELECT and USE privileges.

Databricks Marketplace and Data Lifecycle

MC ID: 7124506 | Parveen KR | Parveen.R@hotmail.com

Introducing the Databricks Marketplace

For Providers

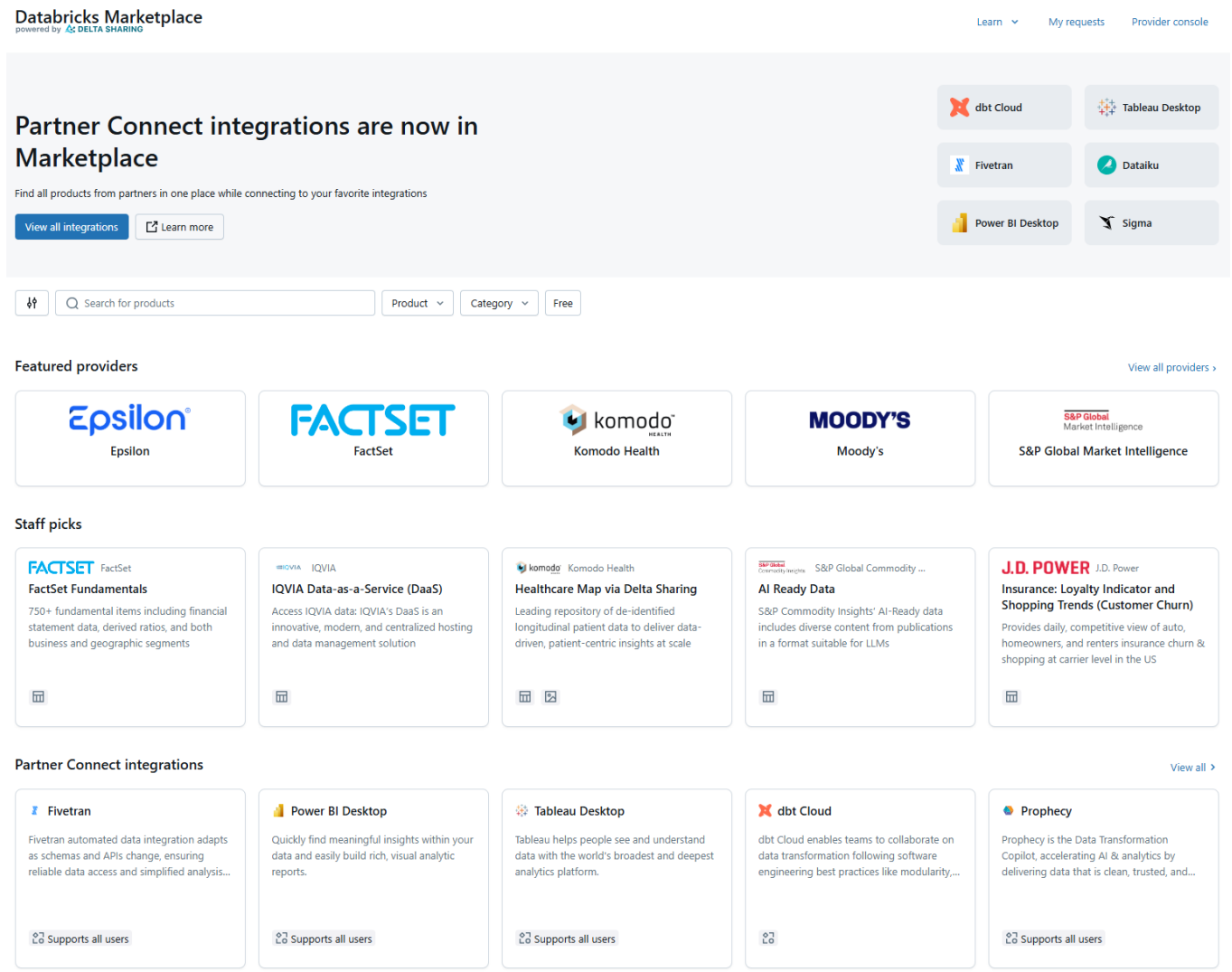
Providers can share datasets, notebooks, ML models, and solutions, enabling them to monetize their data products effectively.

For Consumers

Consumers can discover and access valuable external data and AI assets directly within their Databricks workspace, enhancing analytics capabilities.

Data Governance

Unity Catalog ensures secure access and governance for shared assets, providing a solid foundation for the marketplace.



How the Marketplace Works with Unity Catalog

Delta Sharing Technology

Delta Sharing is a technology that enables secure data sharing without the need to copy data, ensuring data integrity.

Creating Shares

Providers can create Shares within Unity Catalog, bundling data assets like tables and volumes for sharing with recipients.

Marketplace Integration

The Marketplace UI allows consumers to easily discover, access, and request shared data as read-only objects.

Governance in Unity Catalog

Unity Catalog governs access to shared data, ensuring that permissions are managed just like the user's own data.

Data Lifecycle Management: From Ingestion to Archiving

Data Ingestion

Ingestion involves bringing data into the system, such as the data lakehouse, for further processing.

Data Transformation

Transformation includes cleaning, enriching, and modeling data to prepare it for analysis and usage.

Data Storage

Deciding on storage formats and locations is essential for efficient data management and retrieval.

Retention and Archiving

Defining retention policies and archiving older data are crucial for cost management and compliance.

MC ID: 7124506 | Parveen KR | Parveen.R@hotmail.com

Hand on Lab