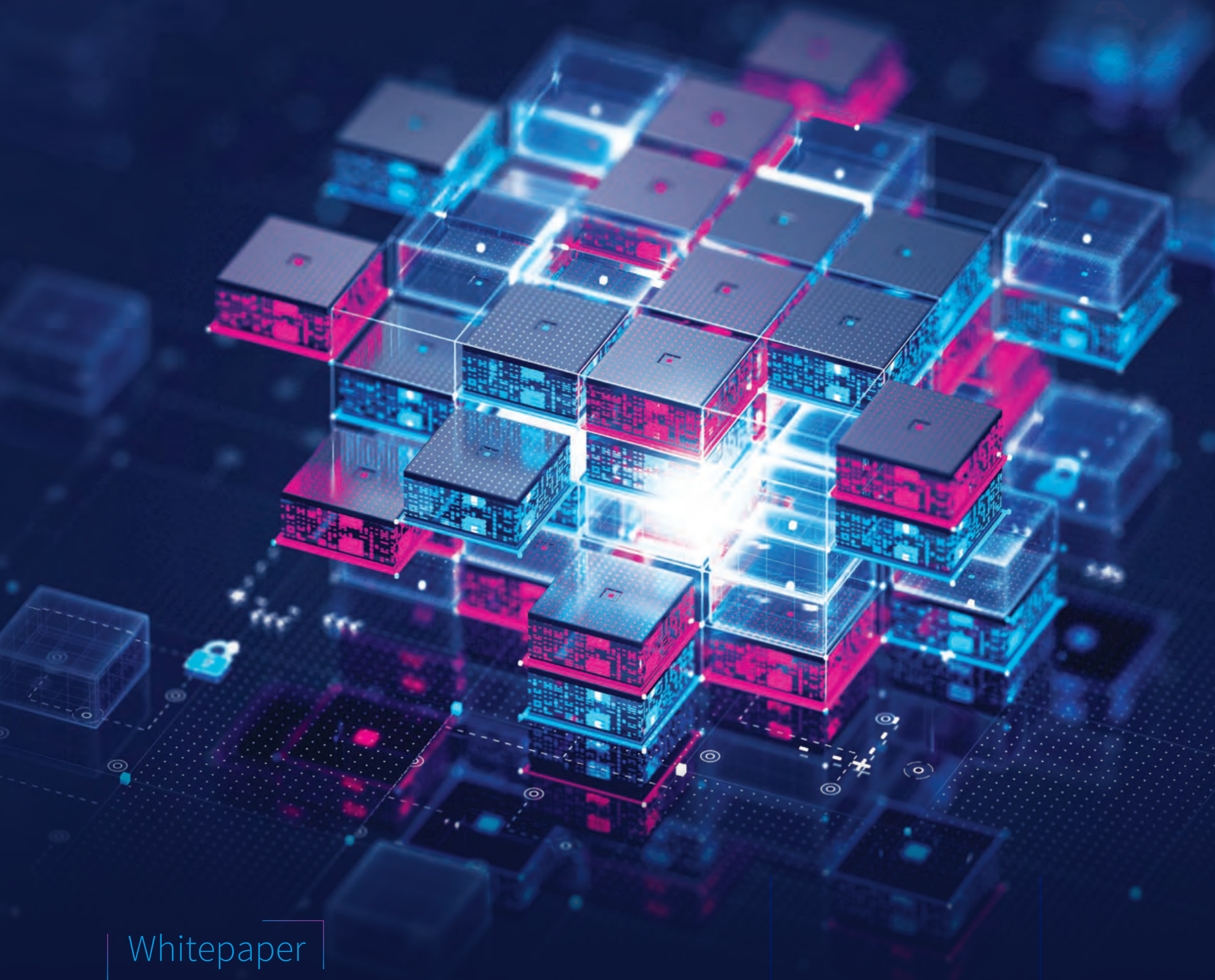
LTIMindtree



Whitepaper

Databricks Unity Catalog and Trust

Table of Contents

Executive Summary

Data is the lifeblood of modern organizations, driving critical decisions and fueling innovation. However, with data's ever-increasing volume and complexity, ensuring its security, accessibility, and trustworthiness becomes a significant challenge. Due to the lack of a unified data governance framework, security vulnerabilities, inconsistencies, and ultimately, a loss of trust in the data itself. Databricks Unity Catalog emerges as a game changer, offering a unified data governance solution that simplifies data management and builds confidence in our data ecosystem. By centralizing metadata management, fine-grained access control, data lineage, data privacy, data security, auditing, compliance, and data insight, Unity Catalog empowers organizations to build a secure and trustworthy data foundation. This white paper explores Databricks Unity Catalog, a solution designed to enhance trust in the Lakehouse environment. We delve into the functionalities of the Unity Catalog that contribute to building trust. Additionally, we explore the limitations of the Unity Catalog and discuss potential future advancements.

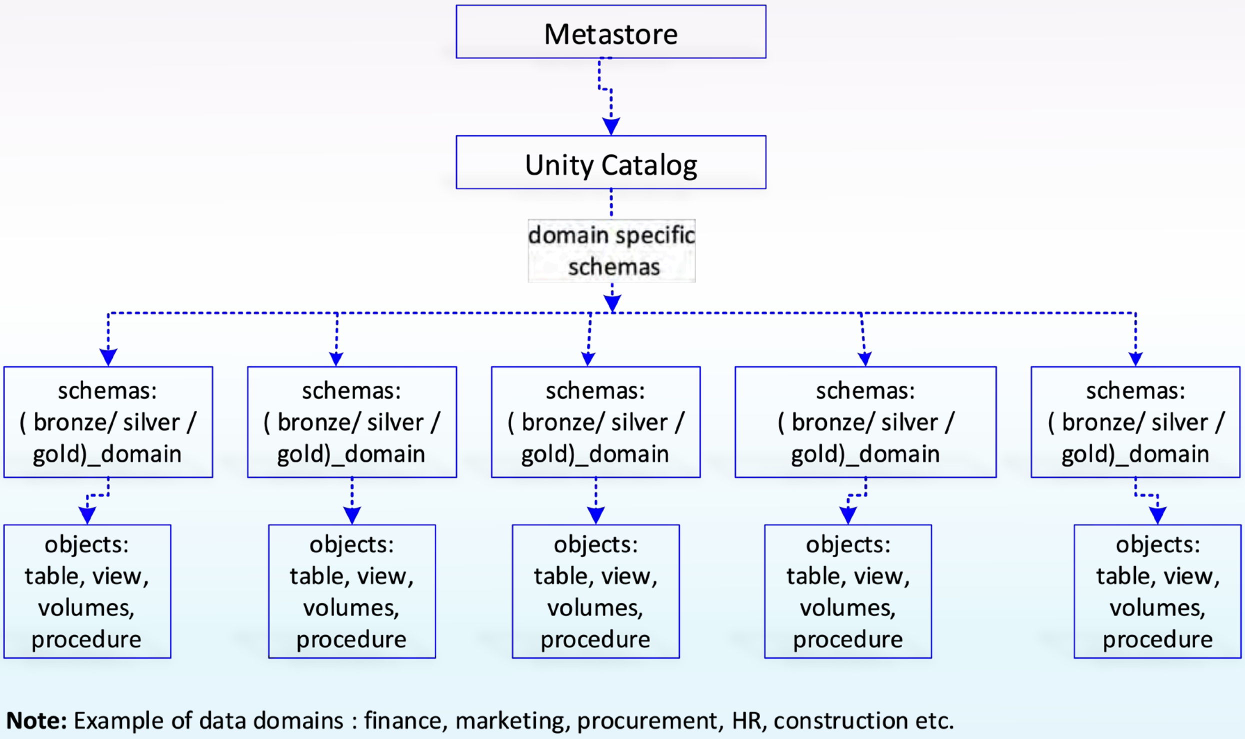
Introduction

The Lakehouse (Delta Lake + Unity Catalog), a hybrid data architecture combining data warehouse and data lake capabilities, has gained significant traction due to its flexibility and scalability. However, ensuring trust in the data within Lakehouse presents a challenge. Databricks Unity Catalog emerges as a solution, aiming to establish Lakehouse as a scalable, reliable, and secure data platform.

Define Unity Catalog

Databricks Unity Catalog is a centralized data governance tool for managing data security, access control, lineage, auditing, compliance, and discovery across Databricks workspaces. It offers a single control point for defining and enforcing data access policies, ensuring consistent rules are applied throughout your data lake environment. Unity Catalog also simplifies data lineage tracking by automatically capturing how data flows through various processing stages, helping you understand the origin and impact of data used in analytics and AI projects. Furthermore, it integrates with existing data storage systems and governance solutions, protecting your investments and simplifying the creation of a future-proof data governance model.

Unity Catalog organizes data assets in a hierarchical structure. At the top sits the Metastore, the central container for all metadata. Within the Metastore, Catalogs are the first layer, functioning like folders to group your data. Schemas, or databases, reside within catalogs and hold tables, views, and volumes. Tables store your structured data, views offer virtual representations of tables, and volumes manage non-tabular data. Here's a simplified figure to illustrate the object model of the Unity Catalog structure:


Metastore
Unity Catalog
domain specific schemas
schemas: (bronze/ silver / gold)_domain
objects: table, view, volumes, procedure
schemas: (bronze/ silver / gold)_domain
objects: table, view, volumes, procedure
schemas: (bronze/ silver / gold)_domain
objects: table, view, volumes, procedure
schemas: (bronze/ silver / gold)_domain
objects: table, view, volumes, procedure
schemas: (bronze/ silver / gold)_domain
objects: table, view, volumes, procedure

Note: Example of data domains : finance, marketing, procurement, HR, construction etc.

Fig.1 Object model of Unity Catalog

The Challenges

Ensuring data trustworthiness, data-driven decision-making, and AI development rely heavily on the quality and trustworthiness of underlying data. However, several challenges hinder data trust, such as:

- **Data silos and lineage gaps:** Fragmented data across various systems creates blind spots, making it difficult to understand the origin and transformations applied to data. This lack of lineage can lead to biases in AI models and hinder regulatory compliance.
- **Inconsistent data access control:** Inadequate access control practices expose data to unauthorized access or manipulation. This raises security concerns and can compromise the integrity of the data used for critical decision-making.
- **Complex data governance processes:** Managing data access policies across diverse data sources can be cumbersome and error-prone. Inconsistent governance practices further erode trust in data.

Building Trust with Databricks Unity Catalog

Building trust in your enterprise data is paramount, and Databricks Unity Catalog tackles this challenge from multiple angles. Its automated data lineage tracks how data flows throughout your system, providing transparency into its origin and transformations. Granular access control ensures that only authorized users can interact with specific data objects. Furthermore, Unity Catalog streamlines data governance by offering a central location for managing permissions and leveraging familiar SQL syntax. Finally, it integrates with existing identity providers and enforces strong access controls through features like Role-based access control (RBAC) , with plans for even more granular control with RBAC in the future. Unity Catalog fosters trust by providing clear visibility, secure access, simplified governance, and robust identity management. Let us analyze how the following key Unity Catalog features foster trust in managing enterprise data:

1. Unity Catalog for Centralized Governance

Databricks Unity Catalog shines in its ability to centralize data governance across your Databricks workspaces. Imagine managing data access policies from a single source, eliminating the need for complex, per-workspace configurations. This defines a once-secure-everywhere approach that ensures consistent data security standards. Administrators can manage access control for catalogs, schemas, tables, rows, and columns with familiar ANSI SQL syntax for defining permissions. This simplifies governance tasks and empowers you to enforce data security policies efficiently. By centralizing control, Unity Catalog streamlines data governance, saving time and resources while ensuring robust data security throughout your Databricks environment.

Here are the key features of centralized governance in Databricks Unity Catalog:

- **Single point of control:** Define and enforce data access policies once for all workspaces, eliminating the need for individual configuration and reducing inconsistencies.
- **Simplified administration:** Manage user permissions, data ownership, and security settings from a central location, saving time and effort compared to managing individual workspaces.
- **Consistent rules:** Ensure data access and security policies are applied consistently across all workspaces, promoting data integrity and compliance.
- **Improved visibility:** Gain a centralized overview of data access and usage across your entire data Lakehouse environment.

2. Unity Catalog for Fine-Grained Access Control

In Databricks Unity Catalog, fine-grained access control refers to the ability to define and enforce access permissions for users and groups at a granular level. This means you can control access not just to entire tables or datasets but also to specific columns or rows within those datasets. This granular control allows for a more secure and flexible approach to data governance, ensuring that users can only access the data they need to perform their jobs. Here are some critical aspects of fine-grained access control in Unity Catalog:

- **Row-level filtering:** Define conditions based on specific data values to determine which rows a user can see. This is useful when you want to hide sensitive information or restrict access to data based on specific criteria.
- **Column masking:** This method masks specific columns in a table, allowing users to see the data's existence but not the actual values. It can help protect sensitive information, such as personally identifiable information (PII) or confidential business data.
- **Dynamic views:** Create logical views of underlying tables that filter and mask data based on user permissions or other defined conditions. This allows you to grant different users access to different views of the same data, ensuring they only see what they're authorized to see.
- **Role-based access control (RBAC):** Define pre-configured roles with specific permissions for user groups. This simplifies access management and ensures users only have the level of access needed for their designated role.

Implementing fine-grained access control in Databricks Unity   Catalog offers several benefits, including:

- **Enhanced data security:** Limiting access to specific data can minimize the risk of unauthorized access and data breaches.
- **Improved compliance:** Granular control helps ensure adherence to data privacy regulations.
- **Increased efficiency:** Users only have access to the data they need, which helps them be more efficient and productive.
- **Reduced risk of errors:** Limiting access reduces the risk of users accidentally accessing or modifying data they should not.

3. Unity Catalog for Data Lineage

In Databricks Unity Catalog, data lineage automatically captures the origin and transformations of data as it flows through your data Lakehouse environment. This provides a comprehensive record of how data is created, modified, and used, which offers numerous benefits for building trust and ensuring data integrity. Here's how data lineage works in Unity Catalog:

- **Automatic capture:** Unity Catalog automatically tracks lineage across various data processing activities, including SQL queries, Python scripts, R scripts, and Pyspark notebooks. It captures information like:
    - **Source data:** The original data tables or files used in the processing.
    - **Transformations:** Any modifications applied to the data, such as filtering, aggregation, or joining.
    - **Destination data:** The resulting tables or files created by the processing.
- **Detailed lineage view:** Lineage graphs easily visualize the data lineage. These graphs show the data flow from source to destination, highlighting all intermediate steps and transformations.
- **Down to the column level:** Unity Catalog captures lineage down to the column level, providing granular detail about how specific data elements are modified throughout the processing pipeline.

Benefits of Data Lineage in Unity Catalog:

- **Improved data quality:** Understanding data lineage can identify potential issues or discrepancies in data transformations, improving data quality and accuracy.
- **Enhanced debugging:** When errors occur, lineage tracking helps you quickly trace the problem back to its source, facilitating faster troubleshooting and issue resolution.
- **Regulatory compliance:** Data lineage helps demonstrate compliance with data privacy regulations like GDPR, CCPA, and HIPPA, which require organizations to understand the flow and processing of personal data.
- **Impact analysis:** Understanding data lineage allows you to assess the potential impact of changes made to upstream data sources on downstream data consumers, ensuring informed decision-making.
- **Increased trust:** Data lineage tracking fosters trust in the data and the overall data Lakehouse environment by providing transparency into how data is processed.

4. Unity Catalog for Lakehouse Federation

Incorporating Lakehouse Federation with Unity Catalog empowers organizations to seamlessly discover, query, and govern data wherever it lives. Databricks Unity Catalog acts as the control center for Lakehouse Federation, a feature that tackles data silos.

Imagine you have data scattered across various databases (Oracle, MySQL, PostgreSQL, Amazon Redshift, Snowflake, Azure SQL, Azure Synapse, Google BigQuery); Lakehouse Federation lets you write Structured Query Language (SQL) queries in Unity Catalog that access this data directly, without physically moving it. This central hub approach seamlessly combines data from different sources for real-time analytics and reports. Additionally, Unity Catalog's data governance ensures proper access control and audit trails for these federated queries, maintaining data security and trust within the Databricks environment.

Key Benefits

- **Unified data access:** Query data residing in diverse sources directly from Databricks.
- **Reduced data movement:** Eliminate the need for manual data transfers or copies, saving storage costs and streamlining data pipelines.
- **Enhanced collaboration:** Facilitate team collaboration by providing a unified view of data across different platforms.
- **Improved efficiency:** Gain faster insights from a broader data landscape without performance bottlenecks.

5. Unity Catalog for AI and ML

Databricks Unity Catalog is the backbone for AI and machine learning (ML) within the Databricks Lakehouse platform. Imagine a well-organized library; you can't effectively research without knowing where books are stored and how they are categorized. Unity Catalog acts like that library system for your data. It establishes a single source of truth for all data and AI assets, ensuring consistent access control and security policies. This secure foundation is crucial for AI and ML models, as it guarantees the trustworthiness and quality of the data they rely on for training and analysis. Furthermore, Unity Catalog automatically tracks data lineage, allowing AI and ML practitioners to understand the flow and origin of data used in models. This transparency is essential for debugging issues, ensuring data quality, and building reliable AI and ML applications.

It goes beyond traditional data catalogs by offering a unified view of structured data and machine-learning models, notebooks, dashboards, and unstructured files. A new feature, AI-generated documentation, allows Unity Catalog to automatically create descriptions for your data and models, saving you time and effort. It also ensures data quality and reproducibility for trustworthy AI development.

Unity Catalog provides a central location to discover, access, monitor, and collaborate on various AI assets, including:

- **ML models:** Track, manage, and govern access to different versions of your machine learning models.
- **Notebooks:** Organize and collaborate on notebooks containing your AI code and experiments.
- **Dashboards:** Centralize and manage dashboards that visualize the outputs and insights from your AI models.
- **AI-powered documentation:** Leverage AI to automatically generate descriptions and comments for your data and AI assets, improving your team's discoverability and understanding.

6. Unity Catalog for LakehouseIQ

LakehouseIQ is a knowledge engine developed by Databricks that is specifically designed to understand your business and its data. It uses generative AI to analyze various aspects of your organization, including industry jargon, data usage patterns, and even your company structure. This allows LakehouseIQ to answer your questions in a contextual way tailored to your specific business needs. Unity Catalog acts as a semantic layer and the foundation for LakehouseIQ, providing the data organization, security, and contextual understanding necessary for the AI engine to function effectively. This allows LakehouseIQ to democratize data access within your organization by facilitating secure and insightful exploration for all users. Here are some of the key features of LakehouseIQ:

- **Natural language interface:** With LakehouseIQ, you can ask questions about your data in plain English, eliminating the need to learn complex query languages.
- **Democratized data access:** LakehouseIQ empowers all employees to access and understand data relevant to their job function, regardless of their technical background.
- **Generative AI:** By leveraging generative AI, LakehouseIQ can continuously learn and improve its understanding of your specific business and data environment.

Unity Catalog plays a critical role in enabling LakehouseIQ, Databricks' AI-powered knowledge engine for data exploration, in a few key ways:

- **Metadata and lineage:** LakehouseIQ relies on metadata, essentially data about your data, to understand the structure and relationships between different datasets in your organization. Unity Catalog serves as the central repository for this metadata, providing LakehouseIQ with a comprehensive view of your data landscape. This allows LakehouseIQ to interpret your specific data terminology and context.
- **Security and governance:** When LakehouseIQ surfaces data based on user queries, compliance with your organization's data access and security policies must be ensured. Unity Catalog acts as the enforcement layer, guaranteeing that LakehouseIQ only delivers data users can see.
- **Understanding business context:** While LakehouseIQ leverages AI to understand your data, Unity Catalog provides additional clues by capturing information like data usage patterns and schemas. This combined knowledge empowers LakehouseIQ to deliver accurate and relevant results for your business needs.

7. Unity Catalog for Lakehouse Observability and Monitoring

In data engineering, data science, ML, and AI, Lakehouse observability and monitoring encompass practices for continuously overseeing the health, performance, and quality of data within a Lakehouse architecture. It goes beyond essential monitoring by offering a deeper understanding of data behavior, lineage, and potential issues affecting downstream analytics and machine learning applications. The key aspects are:

- **Data health and quality:** Monitoring data validity, completeness, consistency, and adherence to defined rules ensures trust in the data for critical decision-making.
- **Pipeline performance:** Tracking data pipeline execution times, resource consumption, and potential bottlenecks. This helps identify areas for optimization and maintain efficient data flow.
- **Model drift:** Detecting changes in model predictions over time, potentially indicating a need for retraining or model adjustments.

Here, the Unity Catalog plays a crucial role in Lakehouse observability and monitoring by providing a unified metadata layer. It acts as a central repository for information about all your data assets, including:

- **Location of data in the Lakehouse**
- **Schema definitions**
- **Lineage information**

Lakehouse observability and monitoring tools leverage the Unity Catalog to understand the context and relationships between data elements. This facilitates:

- **Automated monitoring:** Unity Catalog helps automate the discovery and monitoring of data assets within the Lakehouse.
- **Lineage-based troubleshooting:** When an issue arises, Unity Catalog's lineage information enables you to trace the origin of the problem through data pipelines.
- **Standardized monitoring:** Unity Catalog ensures consistent monitoring practices across your entire Lakehouse environment.

The following diagram illustrates the end-to-end enterprise data processing architecture using a Lakehouse platform and Unity Catalog.

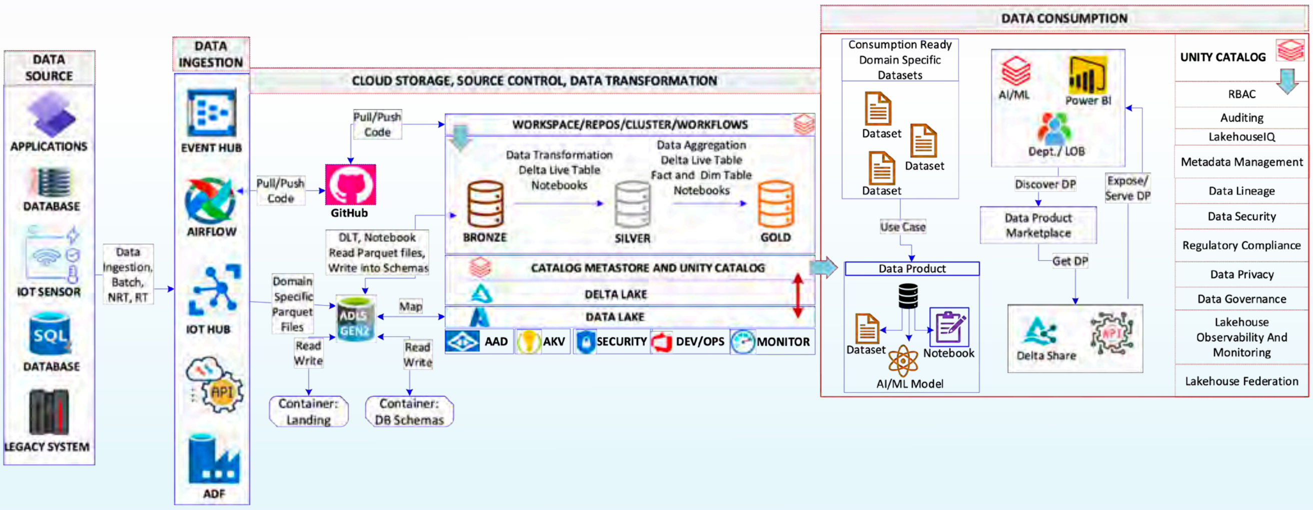


Fig.2 End-to-End next-generation data processing platform on Lakehouse (Delta Lake + Unity Catalog) medallion architecture aligned with data mesh pattern

Limitations and Potential Future Improvements

While Databricks Unity Catalog offers a powerful solution for data governance, it's essential to acknowledge some limitations and areas for potential future improvements:

Limitations

Limited native data quality tools: While Unity Catalog tracks lineage and provides auditing, it currently lacks built-in tools for comprehensive data quality checks and cleansing. Integration with external data quality solutions might be necessary for some organizations.

Focus on structured data: Unity Catalog primarily manages structured data formats. Support for semi-structured and unstructured data, increasingly common in modern data lakes, could be further improved.

Limited customization: While offering granular access control, the current RBAC model might not accommodate the highly complex permission structures required by some organizations. More granular control over object-level permissions could be beneficial.

Vendor lock-in: Unity Catalog is tightly integrated with the Databricks Lakehouse platform. Though it offers some support for external data sources, organizations heavily invested in other cloud platforms might face challenges with full adoption.

Potential Future Improvements

Integrated data quality management: Embedding native data quality checks and cleansing functionalities within Unity Catalog could streamline data governance workflows and enhance integration with popular data quality tools.

Enhanced unstructured data support: As the volume of unstructured data grows, Unity Catalog could evolve to provide better management capabilities for text, images, and other non-tabular data formats. Integration with advanced data lake solutions could address this need.

Advanced RBAC and permission management: Future advancements could introduce more granular control over data access beyond the current RBAC model. User-defined roles and object-level permissions might be valuable additions to complex data governance scenarios.

Openness and multi-cloud support: While offering value within the Databricks ecosystem, Unity Catalog could evolve to support a broader range of cloud platforms and data storage solutions. This would address vendor lock-in concerns and cater to organizations with hybrid cloud deployments.

Conclusion

Databricks Unity Catalog is a cornerstone for building trust in your data Lakehouse. By offering centralized governance, familiar security models, fine-grained access control, robust auditing, data security, and regulatory compliance, Unity Catalog empowers organizations to manage data access and collaborate securely with confidence. As a result, enterprise data teams can:

Confidently leverage data for informed decision-making: With trust in the data's security and accuracy, organizations can confidently base critical decisions on insights derived from their data.

Unlock the power of AI: Secure and reliable data access is essential for training and deploying AI models. Unity Catalog facilitates this by ensuring high-quality data is readily available for AI initiatives.

Achieve greater collaboration and innovation: Secure data sharing and governance foster a collaborative environment where data teams can work together effectively, leading to faster innovation and improved business outcomes.

References

- *Databricks Unity Catalog: A Comprehensive Guide to Features, Capabilities, and Architecture, Atlan, September 28th, 2023:* *https://atlan.com/databricks-unity-catalog/*

- *Databricks documentation on Unity Catalog, April 24, 2024:* *https://docs.databricks.com/en/data-governance/unity-catalog/index.html*

- *YouTube video on data trust with Unity Catalog: https://www.youtube.com/playlist?list=PLTPXxbhUt-YWOfnmciX3rhByRgzUM282T*

- *YouTube video on data trust with Unity Catalog: https://www.youtube.com/playlist?list=PLTPXxbhUt-YVWi_cf2UUDc9VZFLoRgu0l*

Author Profile

Arttatran Parida,

Associate Principal Data Engineering

Arttatran Parida, with 24 years in IT and a master's in computer science, is a seasoned data strategist and architect. A published author, he excels in data strategy, aligning with business goals, and covering the entire data lifecycle. As a Lead Data Engineer, he specializes in MDM, Data Governance, and data modeling, bringing deep expertise in Data Analytics and BI.