Informatica | databricks

Solution Brief

Combining Informatica and Databricks to Expand Data and AI Governance Across Your Enterprise

How Informatica Cloud Data Governance and Catalog (CDGC) and Unity Catalog Strengthen Data Trust.

As AI models, analytics tools, and data-driven organizations are dependent on reliable stores of trusted data, the importance of effective data and AI governance is no longer up for debate. What's less settled is how best to build them.

Enterprise IT estates are highly complex, with data held in different systems, platforms, applications, and physical locations – on-premises and cloud. These organizational silos hinder a cohesive view of data, resulting in inefficient discovery, compliance risks and incomplete insights from analytics.

Leading data platforms have purpose-built governance solutions that are great at what they do, but they often operate in large enterprise environments where multiple data sources, systems, and architectures are the norm. Having control over and visibility of large volumes of data across such a complex landscape is difficult. To operationalize and comply with governance mandates throughout the enterprise, a centralized platform is needed to govern data.

That's why more CDOs and CISOs are opting to support the capabilities of **Databricks Unity Catalog (UC)** with Informatica's Cloud Data Governance and Catalog (CDGC). Used in tandem, they blend file-level data discovery with a 'catalog of catalogs' overview that captures all assets across the enterprise, enabling the extension of data and AI governance policies throughout the entire data estate.

The result is a symbiosis that delivers enhanced, end-to-end governance providing improved efficiency, accelerated insights, and reduced risk exposure as data gets consumed by generative AI models.

Leverage the Complementary Capabilities of Databricks and Informatica

While it's natural to assume all data catalogs do much of the same thing, the truth is more nuanced. Not all catalogs are created equal and some are designed for different tasks.

Take cataloging raw data for example. This is where Unity Catalog excels: diving deep into files to discover every data point and interact at the row and column level. Unity Catalog is unmatched at:

- Creating comprehensive data repositories
- Cataloging the contents, and
- Managing the data points within, particularly inside a Databricks environment

Data rarely resides in a single environment. It lives across silos, geographies, platforms and applications. Global organizations need a comprehensive solution that can catalog all enterprise sources in a single centralized catalog. This consolidation adds complementary data intelligence to platforms like Databricks, bringing extended visibility and accessibility.

Complementing UC's ability to locate, transform, and integrate data for quality and consistency is Informatica CDGC. It provides a unified view across the wider enterprise data estate, capturing metadata from the Databricks environment and everywhere else.

Using these combined and integrated capabilities together provides customers a complete, enterprise-grade data governance and security solution, delivering granular discovery and cross-enterprise visibility.

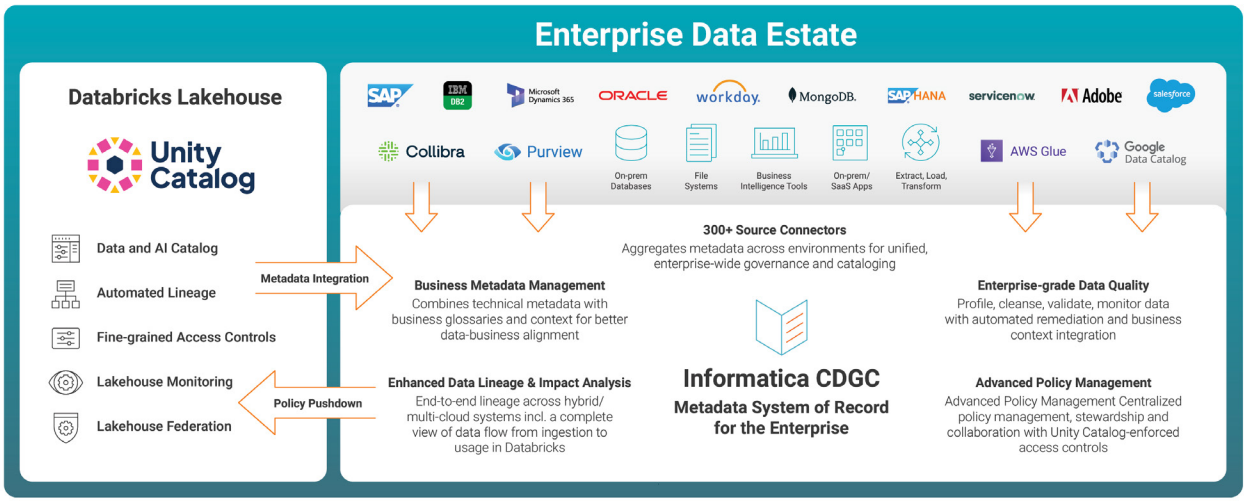
Consider the benefits for improving control of data lineage. Adding Informatica CDGC allows data teams to connect data flows from upstream and downstream systems, providing an end-to-end picture of where data is being created and how it's used by end users across multiple systems.

Cloud Data Governance and Catalog also brings the power of generative AI into play. Informatica's CLAIRE AI engine drives greater automation in essential governance processes including data discovery, data lineage, classification, and profiling.

Informatica | databricks

Informatica Data Governance & Catalog (CDGC)

Extending Unity Catalog for Enterprise-wide Data Governance and Security



About Informatica CDGC

Cloud Data Governance and Catalog is a multi-tenant governance solution designed for diverse on-premises and cloud data sources. It excels at collecting metadata – the schematic representation of data within a repository, its structural patterns, and the types of data within to enhance the understanding of data.

Combining data discovery, data lineage, profiling, business glossary creation and stakeholder and policy management, CDGC enhances data governance for Databricks, and other cloud data lakes or warehouses deployed by the organization.

Part of Informatica's IDMC platform, CDGC looks across the data landscape and helps establish a foundation of secure, trustworthy and high-quality data – a fundamental requirement for successful AI model training and advanced analytics.

CDGC's Capabilities Include:

1. AI-powered metadata extraction from Databricks and other heterogeneous sources
2. AI-assisted data discovery, data classification and data lineage
3. Intelligent data quality and observability
4. Data democratization through Informatica's enterprise data marketplace
5. Data privacy and protection with fine-grained controls to deliver realistic de-identified data to allow organizations to gain reliable insights
6. Natural language search for easily locating critical assets across business and technical domains
7. Data lineage traceability through business-friendly and granular views that include all more intricate details
8. AI model governance for visibility into underlying algorithms and training data used

About Databricks Unity Catalog

Native to the Databricks Data Intelligence Platform, UC is a unified and open governance solution that eliminates silos, simplifies governance and accelerates insights at scale.

It enforces consistent discovery, access, quality monitoring and compliance controls across structured and unstructured data, ML models and business metrics — in any cloud. With Unity Catalog, you can reduce risk, simplify audits and accelerate data access without compromising control.

Unity Catalog also leverages generative AI to simplify documentation curation, and discovery of data and AI assets by automating the addition of descriptions and comments for tables and columns.

Unity Catalog's Key Capabilities Include:

Automatic detection of personally identifiable information (PII) and model drift to proactively resolve issues in data and AI pipelines

Enhanced security with fine-grained control at the rows and columns level

Streamlined debugging, root cause analysis, and impact assessment with automated data lineage

Open APIs and standard interfaces to securely access data and AI assets from any compute engine

CDGC + UC: Better Together

Integrating CDGC with Databricks UC provides data leaders with a holistic data governance solution which spans Databricks SQL, Databricks Notebooks, and the wider enterprise data environment.

Databricks manages a vast landscape of data assets and objects: tables, views, histories, procedures, functions, models, agents, dashboards, notebooks, jobs, workflows, and more. As a result, it needs a dedicated Databricks-native solution to unify governance over such a diverse environment. This provides an enterprise-wide picture incorporating the rich metadata intelligence from Databricks and other sources.

Unblocking Analytics and AI Projects. Reducing Time and Effort.

Gartner predicts that through 2026, organizations will abandon 60% of AI projects due to lack of AI-ready data.[1] Boston Consulting Group says the proportion of companies demonstrating data management maturity has declined since 2021 due to the demands and complexities of generative AI.[2]

When combined, UC and CDGC's complementary capabilities can unblock advanced analytics, ML and AI projects by operationalizing both governance and access policies at every level of the organization. Oversight and control over the full enterprise data estate means governance, compliance, and quality rules can be applied consistently across data sets.

Using CDGC's end-to-end data lineage and impact analysis, Databricks users can track and understand their data flows at a granular level, locating reliable datasets both inside and outside of the Databricks environment.

Having a consolidated view of data saves data scientists time on data preparation. CDGC knows which data has already been scanned, avoiding duplication of effort and added cost. Its automated data quality rules can be applied to Databricks and other cloud platforms, reducing the possibility of anomalies.

For non-technical users, clarity on data quality, compliance, and lineage supports greater data literacy and democratization. Data marketplaces can be leveraged to provide governed access to a range of trusted data assets, enabling self-service analytics and line of business (LoB) level AI initiatives.

Success Story: Accelerating Insights in Financial Services

In most large organizations, tasks related to preparing, discovering and applying governance consume the vast majority of time for data engineers. For example, our client, a global hedge fund, found that only 12% of their data team's time was spent on generating insights.

Traders and analysts, hungry for the latest reports and market measurements, wanted faster updates, but the firm struggled to accelerate data delivery with their existing data catalog.

By marrying the capabilities of Databricks UC and Informatica CDGC, the firm discovered it could accelerate data preparation dramatically. By letting Databricks handle initial scanning, CDGC could then scan metadata pre-gathered by Databricks. That allowed the hedge fund's data scientist to quickly sort through over 5,000 different data sets and schemas and identify the 100 or so that would have value for downstream users.

Combining Databricks UC with Informatica CDGC unlocked up to 75% faster data discovery and preparation, providing the hedge fund's data team with more time analyzing investment opportunities and mitigating risk.

[1] Gartner 2025: **Lack of AI-ready data puts AI projects at risk**

[2] Boston Consulting Group 2024: **Leaders in Data and AI Are Racing Away from the Pack**

About the Databricks-Informatica Partnership

Informatica and Databricks have partnered to help customers modernize their data estates by integrating Informatica's IDMC platform and the Databricks Data Intelligence Platform, providing a solid foundation for analytics and AI workloads. Within IDMC, Informatica's CDGC solution offers multiple Databricks capabilities, enhancing the ability of large enterprises to manage, prepare, transform and cleanse data at scale.

In 2024, Informatica was named **Databricks Data Integration Partner of the Year**, highlighting Informatica's commitment to industry-leading, AI-powered, cloud-native data integration and data management at scale for Databricks customers. CDGC and many other Informatica IDMC services are fully validated for Databricks UC.

Creating Trusted Data Together

Building trust in company data requires comprehensive visibility, understanding and accessibility across the enterprise. A complementary data governance foundation combining Databricks UC and Informatica CDGC can provide the in-depth data intelligence and agility data leaders need:

- Consolidate enterprise data into the Databricks Data Intelligence Platform
- Unlock generative AI innovations with trusted data
- Drive consistent, organization-wide governance

Used on Databricks repositories and beyond, this joint framework improves understanding of the data you have and assess its fitness for use.

You'll be in a better position to know where your data resides, where it came from, what it's quality is, and how to enforce robust compliance with governance and regulatory rules.

Check out CDGC in our demo center: **Democratize data with Cloud Data Governance and Catalog | Informatica**

Ready to take a governance deep dive? **Review our technical documentation for CDGC**

About Databricks

Databricks is the data and AI company. More than 10,000 organizations worldwide — including Block, Comcast, Condé Nast, Rivian, Shell and over 60% of the Fortune 500 — rely on the Databricks Data Intelligence Platform to take control of their data and put it to work with AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache SparkTM, Delta Lake and MLflow.

About Informatica

Informatica (NYSE: INFA), a leader in enterprise AI-powered cloud data management, brings data and AI to life by empowering businesses to realize the transformative power of their most critical assets. We have created a new category of software, the Informatica Intelligent Data Management Cloud™ (IDMC), which is the only platform that manages the entire lifecycle of data and interoperates with everything. It is an end-to-end data management platform, powered by CLAIRE® AI, that connects, manages and unifies data across virtually any multi-cloud, hybrid system, democratizing data and enabling enterprises to modernize their business strategies. Customers in approximately 100 countries and more than 80 of the Fortune 100 rely on Informatica to drive data-led digital transformation. **Informatica. Where data and AI come to life.™**

Worldwide Headquarters
2100 Seaport Blvd., Redwood City, CA 94063, USA Phone: 650.385.5000, Toll-free in the US: 1.800.653.3871

IN17-5201-1125