Informatica

Solution Brief

# Boost Analytics and AI on Databricks with AI-Powered Data Governance

## Trusted Data: The Key to Analytics and AI Initiatives on Databricks

Analytics and artificial intelligence (AI) are transforming how digital businesses operate. However, many firms struggle to fully harness AI's potential and trust its outcomes due to shortcomings in data management capabilities. More than 70 percent of organizations report that managing data is already one of the top obstacles preventing them from scaling AI use cases.[1]

Organizations are adopting cloud data platforms such as the **Databricks Data Intelligence Platform** to enable enterprise-grade analytics and AI solutions at scale. However, the reliability of analytics and AI models built on the Databricks platform depends heavily on the quality and accuracy of your underlying data. Maintaining the integrity and accuracy of large volumes of data across complex data landscapes to train AI models can remain a challenge for your organization.

Building trust in your AI models and insights requires comprehensive visibility, understanding and accessibility to reliable data across your enterprise. To accelerate insights with Databricks, you need a **modern data governance** foundation that spans on-premises and multi-cloud environments and provides in-depth data intelligence. This framework helps you improve your understanding of what data you have and its fitness for use, including where it resides, who owns or produces it, its quality, and its compliance with policies.

### Key Benefits

- Achieve faster time-to-insights with accelerated data discovery
- Visualize, trace and understand end-to-end data lineage across your data sources
- Mitigate privacy and regulatory compliance risks
- Improve the accuracy and reliability of your analytics and AI insights
- Enable safe, secure and responsible data access and sharing

[1] https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20data%20dividend%20fueling%20generative%20ai/the-data-dividend-fueling-generative-ai.pdf?shouldIndex=false

## Scale Analytics and AI with Cloud Data Governance and Catalog for Databricks

Informatica® **Cloud Data Governance and Catalog** (CDGC), a service of Informatica Intelligent Data Management Cloud™ (IDMC), integrates with Databricks Unity Catalog to provide holistic data governance within Databricks SQL and across the wider data landscape. CDGC acts as a catalog of catalogs to create a comprehensive metadata inventory across Databricks and similar data sources with integrated scanners that enable companies to uncover where data resides and build a comprehensive picture.

The Informatica **CLAIRE**® AI engine helps you enrich the metadata with relevant context and enables advanced capabilities designed for rapid data identification, curation and collaboration at scale. End-to-end data lineage and impact analysis allow Databricks users to track and understand data flows at a granular level within and outside the Databricks environment and help ensure compliance with your organizational governance policies through trusted business reporting.

Integrated **data quality and observability** capabilities assist in identifying and managing data quality issues within Databricks and enable you to meet data quality standards as data gets utilized across the platform. CDGC can be combined with the Informatica **Cloud Data Marketplace** (CDMP) service to help ensure reliable and verified data products are accessible and discoverable to your technical and non-technical members for driving analytics and AI within Databricks. Additionally, capabilities in **Cloud Data Access Management** (CDAM) enable universal, automated controls to enforce security and privacy policies that help safeguard your data as it is shared across the organization.

With **CLAIRE GPT**, a GenAI-powered data management assistant, you can interact with and manage your data through a natural, language-based, business-friendly interface, driving productivity gains and accelerated access to capabilities by improving ease of use.

## Key Benefits

### Achieve Faster Time to Insights with Accelerated Data Discovery

Accelerate time to insights and improve productivity by reducing time and effort spent finding relevant data assets. Quickly locate reliable datasets to include in data pipelines from within and outside of your Databricks platform. This can provide a consolidated view of data and helps save your data analysts and data scientists time on data preparation. By automating data governance workflows and providing secure self-service access to data, your organization can reduce bottlenecks and use data confidently for decision-making.

### Visualize, Trace and Understand End-to-End Data Lineage Across Your Data Sources

Organizations should thoroughly understand their data to get the most value from it. Comprehensive data lineage provides you with the transparency to support self-service analytics and AI initiatives. End-to-end **data lineage graphs** with detailed information enable users to understand the context of the data flow, including its source, transformations and use. This empowers your technical and non-technical business users working across multiple systems to understand the data in depth and assess the impact that changes in data may have on related data assets within or outside the Databricks environment.

### Mitigate Privacy and Regulatory Compliance Risks

With CLAIRE, your organization can automatically identify, classify and protect sensitive data in Databricks and other data sources to help reduce the risk of non-compliance. AI Governance capabilities provide insights into AI models and the data used to train them. Improved accountability, explainability and transparency across the data landscape help ensure you are compliant with rules for sensitive information use, reducing the risks of improper intellectual property exposure and data misuse.

### Improve the Accuracy and Reliability of Your Analytics and AI Insights

Databricks users can more quickly identify and understand data anomalies and issues by leveraging the integrated data quality and observability capabilities in CDGC. Additionally, CDGC can improve your organization's operational efficiency by monitoring diverse data pipelines, reducing the time invested in manual efforts. This approach helps ensure the data that fuels analytics and AI initiatives remains reliable and trustworthy. Further, automated **data quality** rules can be instituted for maintaining data consistency across cloud platforms, including Databricks, reducing the possibility of data anomalies. Enabling prompt data delivery and maintaining a high level of data quality and trustworthiness helps your organization make accurate decisions and ultimately achieve reliable business results.

### Enable Safe, Secure and Responsible Data Access and Sharing

Facilitate data sharing and empower your data consumers with trusted data. With CDGC for Databricks, users can find, understand, govern and trust the data they need within its context and terms of usage. This allows them to utilize it for decision-making with fewer dependencies on IT teams. It provides visibility into data assets across different parts of your organization, including their origin and relationships. Informatica **Cloud Data Marketplace** (CDMP) capabilities complement CDGC, allowing you to create a centralized platform for self-service access to verified data products. Additionally, Informatica **Cloud Data Access Management** (CDAM) capabilities combine with CDGC to enable safer accessibility to trustworthy data for data-driven decision-making across your organization.

## Key Capabilities

### Broad and Deep Metadata Connectivity for Unified Metadata Management

CDGC builds a centralized view of the data landscape by offering broad and deep metadata connectivity with hybrid, multi-cloud and on-premises environments. Pre-built scanners for various cloud platforms and tactical catalogs allow CDGC to act as the primary catalog, enabling users to explore the wider data ecosystem for relevant datasets. A unified, AI-powered metadata intelligence foundation combined with a business glossary to define and standardize business terms helps ensure Databricks-related metadata and business terms are consistent with the broader enterprise business glossary and fosters a common understanding of data definitions across the organization.

**Contact Informatica** for the most current list of supported data sources.

## AI-Powered Rapid Data Discovery and Classification

CLAIRE GPT and powerful semantic search capabilities allow your users to perform queries in natural language to locate critical data across cloud and on-premises platforms. CDGC provides filtering and preview capabilities to quickly review and identify desired assets. Multiple personas can explore data assets using browsable hierarchical views for context, relating technical data sources to business-curated datasets. CDGC empowers your users to add business context to data by automatically associating business terms with the correct technical metadata to help enhance the value of your data for analytics and AI.

CLAIRE-powered automation generates data classifications without the need for human input. CDGC recognizes the meaning of data in the catalog sources and classifies the columns automatically. Your data stewards can curate, review and accept from more than 215 out-of-the-box automated data classification associations recommended by CLAIRE. Users can further modify, rename and extend these classifications or add new ones as needed. (See Figure 1.)

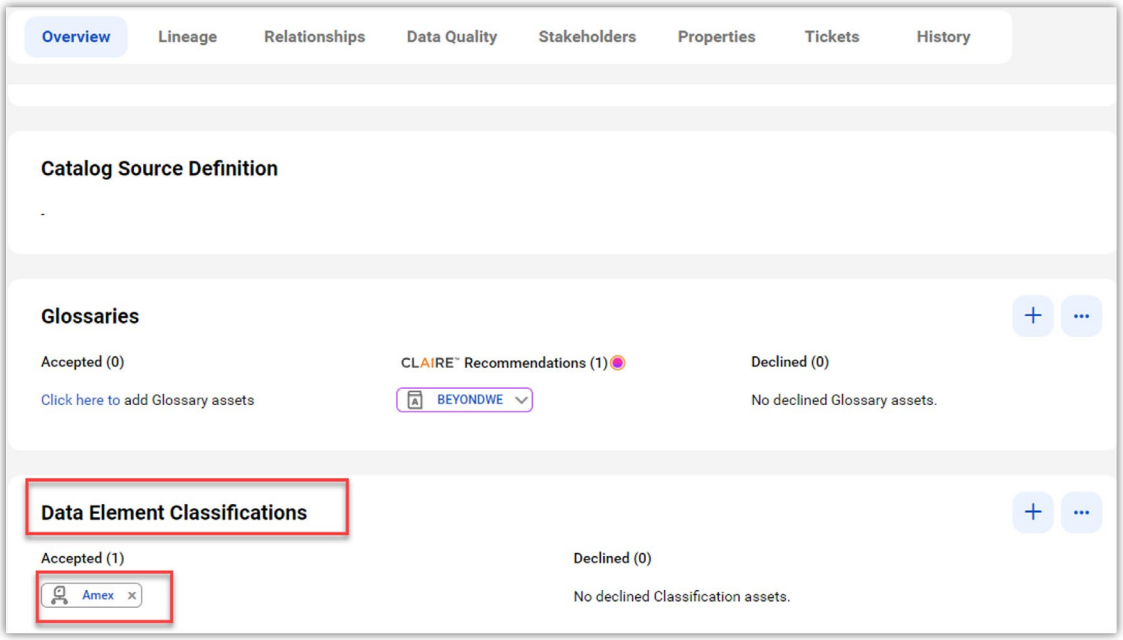


Figure 1: CLAIRE automatically infers and recommends data classifications for data assets based on rules.

## End-to-End Data Lineage and Impact Analysis

End-to-end data lineage and impact analysis capabilities help users track data flows within the Databricks platform and across other sources and applications. Inferred lineage capabilities automatically detect dataflows between sources and targets at the dataset and data element level, as shown in Figure 2. Users can also perform these tasks within your linked data sources such as hyper scalers, enterprise applications and systems, on-premises databases, and ETL and BI tools. CDGC interactively traces the origins of your data with lineage views at multiple levels, from business-friendly, system-level views, highlighting the endpoints to more granular views that include complex details in between.

You can leverage CDGC to perform detailed impact analysis of transformations within Databricks, as well as third-party upstream and downstream data assets and linked systems. You can conveniently visualize essential details associated with data, such as business glossary terms, domains, policies and processes directly within data lineage overlays. Additionally, data quality overlays allow your users to monitor data quality scores and how they change throughout the data flow across the data estate.

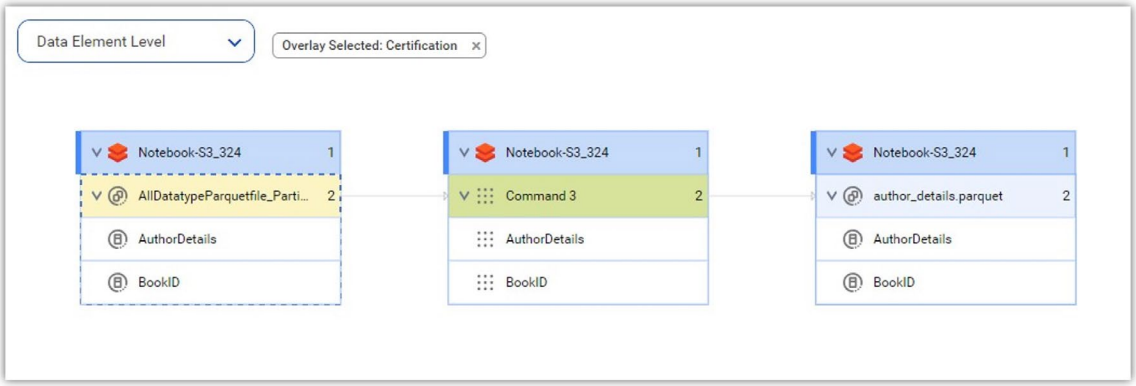


Figure 2: Explore data lineage at the data element level and view transformations on the data.

### Integrated Data Quality and Observability

Informatica Cloud Data Quality (CDQ) enables your users to view data profiling statistics, rules, scorecards and metric groups alongside technical metadata. Refer to Figure 3 for an example. This helps them understand the data quality of assets, an integral part of any data governance program. Profiling statistics provided by CDQ, including value distributions, patterns, data type and domain inferences, automate data quality measurement. This helps to significantly reduce the stakeholder burden. Your users are also offered data quality previews to determine the accuracy and usability of their data. Additionally, CDGC can automatically notify stakeholders of data quality status changes via the user interface or email, allowing them to act swiftly and insightfully.

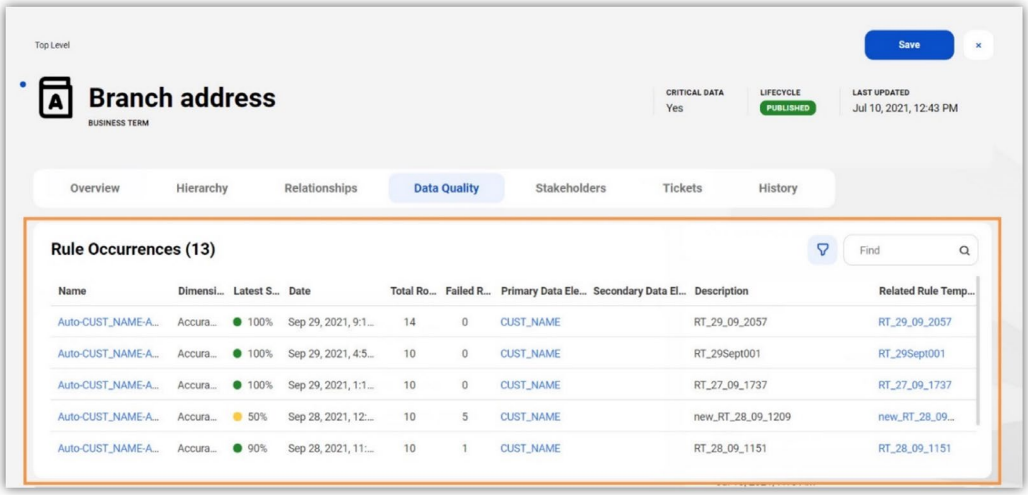


Figure 3: View rule occurrences in the data quality tab to determine the health of datasets.

### Automated Policy Enforcement and Compliance

CDGC connects with the Databricks Unity Catalog to enforce data governance policies consistently across the data landscape. CLAIRE can automatically link policies to data assets by using AI/ML-based data classification. Users can assign sensitivity levels to data to highlight its importance, evaluate risks and execute control measures. Your organization can deploy these policies using IDMC data protection, such as masking for obfuscation capabilities. Alternatively, you can use third-party or platforms-as-a-service offerings that host the data to manage policy-compliant data access. The automated workflows within CDGC enable your stakeholders to create and modify assets in compliance with **data governance** principles within Databricks and other data platforms.

## Safe, Secure and Protected Data Sharing and Access Management

By combining CDGC with other IDMC services such as CDMP and CDAM, organizations can establish a storefront environment for sharing data products safely and securely. With CDMP, data owners can package, publish and promote curated data products, AI/ML models and pipelines from a broad variety of sources. Data owners and operations teams can track requests for data and make data available in the format, location and access method most convenient for use. Data consumers can view policies and guidelines that are transparently enforced at the row, column and cell levels to determine terms of use. **CDAM** can deploy access controls that work universally across systems, including Databricks, and helps organizations to trust that their data is protected from its source to its use by applications and consumers. It uses conditional logic to compose rules, mapping data classifications and contextual attributes to controls. CDAM can enforce policies automatically across multiple workloads and data flows in diverse environments and logs policy decisions, data provisioning and data access for auditing and general observability.

## For More Information

To learn more about Informatica's intelligent data governance and privacy tools that can help you connect data consumers with trusted data, **visit our website**.

**Where data & AI come to LIFE™**

Informatica

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

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), powered by AI and an end-to-end data management platform 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.™**

IN17-5059-0924

© Copyright Informatica LLC 2024. Informatica and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and other countries. A current list of Informatica trademarks is available on the web at https://www.informatica.com/trademarks.html. Other company and product names may be trade names or trademarks of their respective owners. The information in this documentation is subject to change without notice and provided "AS IS" without warranty of any kind, express or implied.