Informatica

Solution Brief

Accelerate Analytics and AI Initiatives on Databricks With an AI-Powered Data Catalog

Key Benefits

- Empower non-technical and technical users with rapid data discovery and collaboration at scale
- Easily visualize, trace and understand your data from source to target with end-to-end data lineage
- Extract deep metadata and data lineage from Databricks and additional data sources
- Enable robust enterprise-wide data governance, privacy and regulatory compliance programs
- Accelerate your cloud journey with data intelligence

Rapid Data Discovery and Data Context for Multi-Cloud Environments

Data, analytics and artificial intelligence (AI) are transforming business. AI has arguably the most potential to drive new value from data; however, many organizations run into difficulties during implementation, due largely to a lack of modern data management capabilities.[1] The most challenging aspect of AI is managing the data that fuels the AI models. Whether they are using cloud data warehouses, data lakes, lakehouses or legacy storage, modern enterprises need the ability to easily and cost-effectively store, access, integrate and analyze massive volumes and varieties of data in real time. At the heart of this shift is the need to derive value from data to fuel innovation, improve customer experience and increase operational agility and speed. Enabling trusted and democratized data across the enterprise unleashes the promise of self-service analytics and AI workloads at scale.

A recent survey of CDOs, chief analytics officers, CIOs, CTOs and other senior technology leaders found that "just 13% of organizations excel at delivering on their data strategy."[2] How does this select group of "high achievers" deliver measurable business results across the enterprise? Their success depends upon a foundation of sound data management and architecture to democratize data and derive value from machine learning (ML).

In 2020, 64.2ZB of data was created or replicated. This growth is forecast to experience a 23% compound annual growth rate (CAGR) over the 2020–2025 forecast period.[3]

Although much of this data is being created in the cloud, many enterprises are running their organizations across a variety of on-premises, dedicated private cloud and multiple public cloud environments, with data located in modern and traditional sources. This further increases the operational complexity of the data landscape when governing vastly disparate data sources.

[1] Why Is It So Hard to Become a Data-Driven Company? Randy Bean, Harvard Business Review (February 5, 2021)
[2] IDC InfoBrief, sponsored by Informatica, Driving Business Value from Data in the Face of Fragmentation and Complexity, doc #US48293521, November 2021
[3] IDC, "Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All?" (Doc #US46410421, March 2021)

A key challenge for enterprises operating in a complex data landscape is the lack of end-to-end visibility and understanding of data. In today's enterprise, petabytes of data are dispersed across data platforms, including the Databricks Lakehouse Platform. Enterprises lack the necessary in-depth data intelligence to understand what data they have, where it resides, who owns it, its quality standards, and its compliance with governance and privacy policies. Just as important is the need to know what transformations each dataset has undergone throughout its lifecycle, and what data dependencies exist across the entire data ecosystem.

For enterprises modernizing their data, AI and analytics platforms with Databricks, this challenge is amplified when the metadata and data lineage are trapped and lack transparency. This is often the case with metadata that exists across various data sources, such as complex enterprise applications and systems, stored procedures for databases, data warehouses and multivendor ETL and BI tools. This lack of transparency makes it difficult to extract and understand data and creates additional operational and regulatory risks.

For example, analytics and AI workloads depend heavily on the quality and accuracy of underlying data. Building trust in AI and ML models and insights requires comprehensive visibility and understanding of source data. To accelerate data intelligence insights, data scientists require a complete and unified data cataloging and governance foundation that spans on-premises and multi-cloud environments.

Informatica® Enterprise Data Catalog allows you to build a comprehensive inventory of metadata, regardless of where it resides, inclusive of the Databricks Lakehouse Platform and other data sources. Powered by intelligence and automation from the Informatica CLAIRE® AI engine, Enterprise Data Catalog enriches this metadata with relevant context and delivers advanced capabilities designed for rapid data discovery, curation and collaboration at scale.

With end-to-end data lineage and impact analysis, you can easily visualize, trace and understand the flow of data at a granular level, within and outside Databricks. Enterprise Data Catalog is a foundational pillar for enabling a holistic and comprehensive data governance and data democratization strategy for all your data.

Key Capabilities

Rapid Data Discovery Powered by Advanced ML

Enterprise Data Catalog enables rapid discovery of data with powerful, semantic search — empowering data stewards, data scientists, and analytics, data governance and data architect teams to easily find the data they need. Users can quickly discover and profile data, identify its location and obtain key attributes about datasets at scale. Semantic search is also applied to inferred data domains, including synonyms and concept matching, so that no data asset is left undiscovered.

Using advanced statistical and metadata-driven ML algorithms, Enterprise Data Catalog tackles the inherent complexity in data to discover, tag, cluster and identify similarities and patterns in data and capture systemwide data relationships. This enables intelligent cataloging for all types of data at scale.

With Enterprise Data Catalog, you can easily import business glossary assets such as terms, policies and classifications from Informatica Axon™ Data Governance, as well as third-party tools. You can then add rich business context to data by automatically associating business terms with the right technical metadata.

Broad and Deep Metadata Connectivity With End-to-End Data Lineage and Impact Analysis

Enterprise Data Catalog is the "catalog of catalogs," with both broad and deep metadata connectivity. It offers the most comprehensive set of scanners that are purpose-built to extract deep metadata and data lineage from widely adopted data sources across on-premises, hybrid and multi-cloud environments.

Enterprise Data Catalog tracks data lineage of pipelines with Databricks' Unified Analytics Platform and makes Databricks tables available as part of the data catalog. End-to-end data lineage and impact analysis capabilities allow you to easily visualize, trace and understand the flow of data within Databricks. You can also perform those tasks within linked data sources such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud, enterprise applications and systems, on-premises databases, and ETL and BI tools.

You can perform detailed impact analysis of transformations within Databricks, as well as on third-party upstream and downstream data assets and linked systems. You can interactively trace data origin through lineage views at any level — from business-friendly, system-level views that highlight the endpoints to granular views that include all the complex details in between. Additionally, a drill-down lineage view expands any lineage path to show granular column- and metric-level lineage.

Purpose-Built Advanced Scanner for Databricks Notebook

Enterprise Data Catalog Advanced Scanners are purpose-built to enable deep metadata extraction, the derivation of detailed data lineage for data intelligence. The Advanced Scanner connects seamlessly with Databricks Lakehouse Platform to scan and index metadata such as Folder, Notebook and Command objects for context. This allows teams to discover data asset dependencies through detailed data lineage and impact analysis. Data professionals can easily track data movement, including column- and metric-level lineage, to identify related tables, views and domains. These capabilities accelerate data-driven insights and analytics decision making.

Data lineage with Enterprise Data Catalog is enabled through Python and Spark SQL methods to parse code to extract detailed data lineage and includes capabilities such as nested Notebook execution and embedded SQL parsing. In addition, the Advanced Scanner supports temporary tables and reference objects.

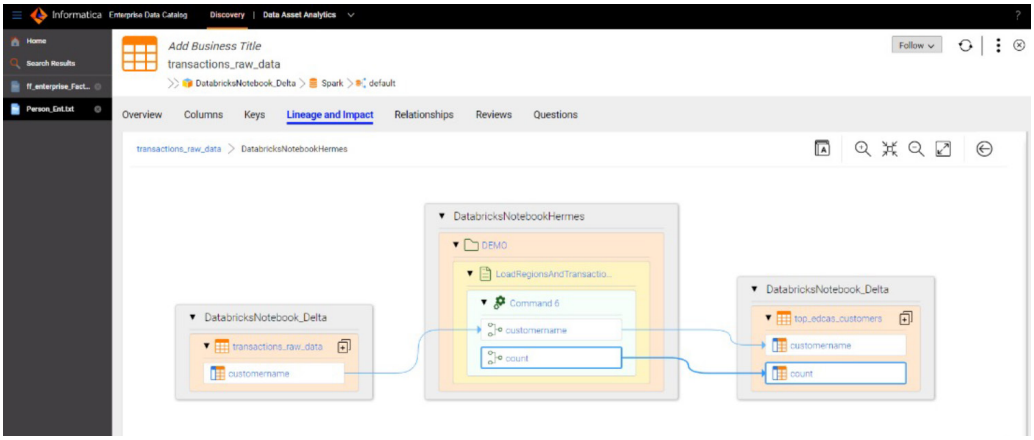

Figure 1: The Advanced Scanner for Databricks enables detailed lineage for data impact and insights.

Data Collaboration and Social Curation with Intelligent Crowdsourcing and Annotations

Enterprise Data Catalog harnesses the combined power of sophisticated ML algorithms, human expertise and collaboration, and makes it easy for data stewards, data scientists, and data governance and analytics leads to find the most relevant and trusted data for analysis. Data owners and subject matter experts can certify datasets and provide ratings and reviews, enabling social curation of data. A Q&A platform lets subject matter experts answer common questions from users. The solution for Databricks accelerates data discovery to help determine the best datasets for analytics use cases.

Integrated Data Quality

View data profiling statistics (such as value distributions, patterns and data type, and data domain inference), data quality rules, scorecards and metric groups alongside technical metadata to understand the quality of data assets within Databricks before using data for analysis. Profiling statistics include value distributions, patterns and data type and data domain inference.

Advanced Data Asset Analytics

Data Asset Analytics provides prepackaged reports and dashboards on data asset inventory, usage, enrichment, level of collaboration and more. Reports are extensible and can be exported, allowing data leaders to share business adoption and value metrics with stakeholders. Automated Data Value Calculator, a first-of-its-kind capability, lets you measure and optimize the value of enterprise data assets based on key factors that impact data value. For instance, you can obtain information about the percentage of your data inventory that resides in key data sources, along with the types of data your users are accessing. This will help you proactively prioritize, manage and optimize the value of your data assets when migrating to Databricks Lakehouse Platform.

Key Benefits

Achieve Faster Time to Value with Trusted Insights for Data Science

By parsing code in Databricks Notebooks with the Informatica Advanced Scanner, data scientists, engineers and others are better armed with the end-to-end lineage and in-depth data insights they need by discovering and applying relevant data. The combined solution from Informatica and Databricks lets enterprises build more accurate AI and analytics models from trusted data and pipelines that enable self-service analytics with confidence.

Enable Comprehensive Data Governance at Scale

The Informatica Data Governance solution brings together advanced capabilities using a consistent metadata-driven platform to share data intelligence. The intelligent, integrated and modular solution allows you to democratize data rapidly and cost-effectively with trust assurance, encompassing all your data.

Capture and Enforce Privacy Policies

As part of Informatica's complete Data Governance solution, in addition to rapidly discovering data residing in Databricks (or other data sources) that is subject to privacy regulations, data professionals can leverage comprehensive, user-defined privacy policies. For example, using Boolean match conditions and acceptance thresholds, users can search any of the multiple data elements controlled by privacy policies (e.g., CCPA, GDPR, BCBS 239, HIPAA and others).

Empower Non-technical and Technical Users

Data stewards, data scientists, data governance teams, data architects and other stakeholders can rapidly discover, certify and collaborate on data at scale. Users can easily identify which datasets may contain personally identifiable information (PII) and the source systems where it originates. End-to-end lineage lets users trace and understand the movement of sensitive data at a granular level, so they have the intelligence they need to make informed decisions about data exposure and value to the organization.

About Informatica

At Informatica (NYSE: INFA), we believe data is the soul of business transformation. That's why we help you transform it from simply binary information to extraordinary innovation with our Informatica Intelligent Data Management Cloud™. Powered by AI, it's the only cloud dedicated to managing data of any type, pattern, complexity, or workload across any location —all on a single platform. Whether you're driving next-gen analytics, delivering perfectly timed customer experiences, or ensuring governance and privacy, you can always know your data is accurate, your insights are actionable, and your possibilities are limitless. Informatica. Cloud First. Data Always.™

Accelerate Migration to the Cloud

Migrating analytics and AI to the cloud offers improved economies and more agility for applications but requires improved data intelligence to identify and prioritize key datasets while reducing risk exposure. You can build a comprehensive and unified view of your critical data that resides within and outside the Databricks Lakehouse Platform to help simplify your journey.

Next Steps

To learn more, please visit Informatica's Databricks solution page and the product pages for Informatica Enterprise Data Catalog and Enterprise Data Catalog Advanced Scanners.

Informatica

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

IN17_0422_04239