Informatica

Solution Brief

Unlock the Full Potential of Your Data With AI-Powered Inferred Data Lineage

Manage Data More Efficiently in Complex Data Ecosystems

Every day, modern businesses generate a staggering amount of data, fueled by big data growth, IoT advancements and digital transformation. Enterprises might have hundreds or even thousands of such data sources, spiraling into tens of millions of data objects. According to Informatica's 2024 CDO Insights Survey, 38% of respondents grapple with an increasing volume and variety of data. Currently, 41% already struggle with 1,000+ sources and 79% expect that number to increase beyond 2024.

With the advent of AI-powered, automated data scanning, organizations can now track data movement and sharing with unparalleled precision across a sea of data sources. This technological leap allows for the extraction of metadata, offering a clearer picture of the intricate relationships within your data ecosystem. The **Informatica Cloud Data Governance & Catalog** (CDGC) solution uses advanced techniques to automatically scan and extract metadata from various data sources, including cloud platforms, BI tools, databases, multi-vendor ETL and data science tools, enterprise applications, file formats, SQL dialects and stored procedures. This process helps to provide comprehensive visibility of data during its journey by deriving end-to-end data lineage.

Key Benefits

- Enable Intelligent Lineage Discovery and Visualization
- Enhance End-to-End Data Visibility
- Manage Regulatory Compliance
- Improve Data Quality and Trust
- Increase Operational Efficiency
- Deliver Scalability and Flexibility

Enhancing Data Governance Through Inferred Data Lineage

Due to technological limitations or security constraints, complete lineage may not be visible after metadata extraction in many enterprises. However, this gap can be bridged by employing inferred data lineage, which means that the data flow and relationships have been analyzed to make educated deductions about how data moves through processes, transformations and storage locations; for example, if a data pipeline extracts data from a source database, performs some transformations and then loads it into a target data warehouse, the inferred data lineage would show the flow from source to target, even if there isn't explicit documentation for each step.

How It Works

Source and target catalog sources can be linked to create lineage. As illustrated in Figure 1, users can select specific source and target schemas to restrict lineage inference to specific subsets of data objects within the data sources. CDGC uses **trained AI** models to automatically detect and build lineage between user-provided sources and targets using CLAIRE-powered linking, as shown in Figure 2. The linked assets and generated lineage links are auto-accepted by default and appear on the **Catalog Source Links** page in CDGC. However, stakeholders, such as data stewards of the source and target catalog sources, can reject these auto-accepted lineage links from the **Action** menu. If stakeholders initially reject the generated lineage links and later accept them, they are marked as accepted in CDGC. Stakeholders can also view the generated lineage on the **Lineage** tab of the asset.

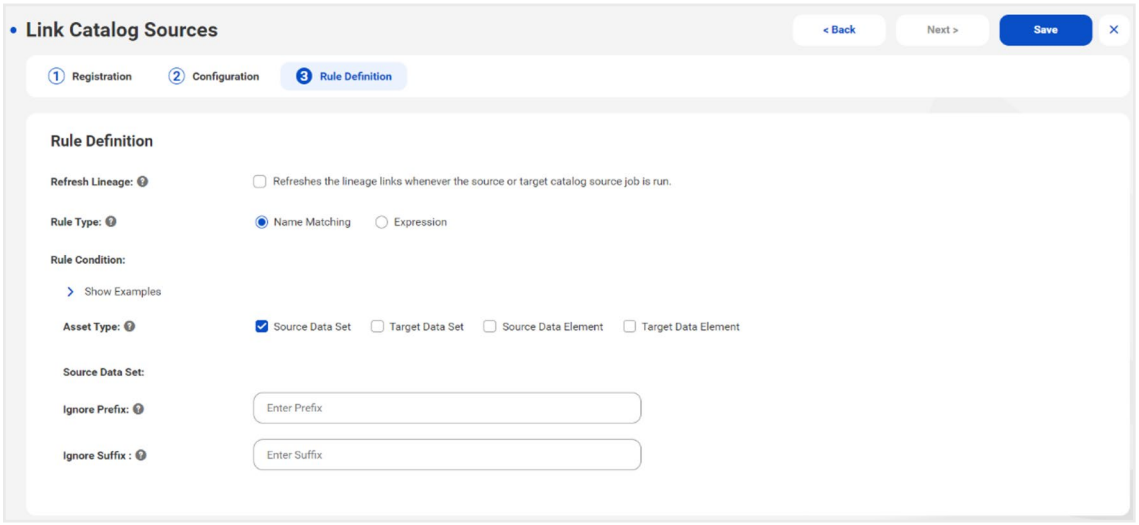

Figure 1: Showing the **Rule Definition** tab with **Name Matching** as the rule type

Lineage can then be viewed before or after curating the catalog source links, either directly in the **Lineage** tab of the asset or by accessing it from the **View Lineage** option on the **Action** menu of the linked assets on the **Catalog Source Links** page.

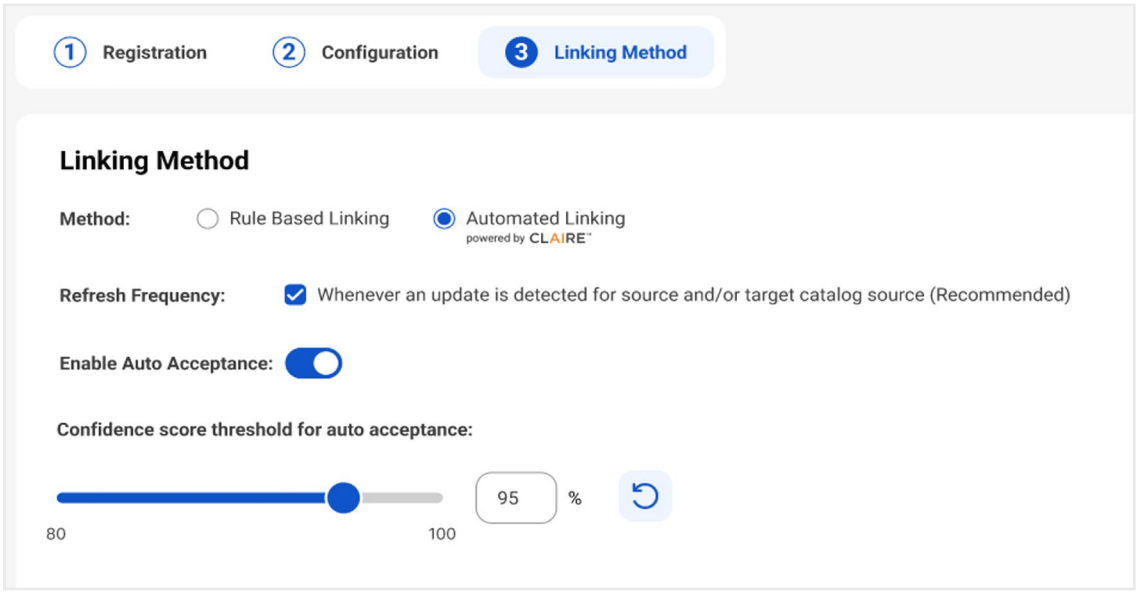

Figure 2: Automated linking powered by CLAIRE, the AI engine in IDMC

The system infers the lineage based on the observed patterns and dependencies, filling in gaps where documentation might be missing and providing a more complete picture of data movement across an organization. This is critical to ensure that your data is AI-ready.

Inferred data lineage, as part of Informatica CDGC, avoids time-consuming manual lineage processing, thereby reducing the risk of user errors. Recommended lineage is visible only to stakeholders, who can then make it available to other users. This capability can help organizations deliver the end-to-end visibility of data, which builds confidence in their analytics and AI models, improves customer experience programs, helps ensure regulatory compliance with industry policies, accelerates cloud modernization initiatives and much more.

Key Benefits

Enable Intelligent Lineage Discovery and Visualization

AI-powered lineage automation reduces manual effort by discovering and visualizing data flows across different systems. This improves data accuracy and saves time, allowing your team to focus on more strategic tasks that can drive better business value.

Enhance End-to-End Data Visibility

It provides comprehensive traceability into data's journey from origin to consumption. This transparency helps in understanding data dependencies and relationships, which is crucial for data governance and building trust in data usage.

Manage Regulatory Compliance

With extensive lineage tracking, it ensures compliance with global data regulations. This is particularly beneficial for industries with stringent regulatory requirements, such as finance and healthcare.

Improve Data Quality and Trust

By identifying and addressing data quality issues, it builds confidence in the data used for analytics and AI models. This trust is essential for making informed, data-driven decisions.

Increase Operational Efficiency

Automating data lineage helps identify redundancies and inefficiencies in data pipelines, streamline operations and accelerate cloud modernization initiatives.

Deliver Scalability and Flexibility

AI-powered data lineage can automatically scale up or down the resources needed to process the data based on load, ensuring efficient data lineage tracking without manual intervention. This makes it suitable for enterprises with extensive and diverse data environments.

Key Capabilities

The complexity of today's data landscape demands data lineage solutions with advanced functionalities that leverage automation and AI.

CDGC AI-powered data lineage provides the following key capabilities:

Automatic Data Lineage Stitching From Multiple Sources

Extract and decipher lineage from metadata collected from all your data systems across the cloud and on-premises for thorough end-to-end visibility. For instance, data may flow from transactional databases to a data warehouse and then be consumed in a BI tool like Tableau, Power BI or Qlik. Along the way, data is transformed using an ETL tool or stored procedure. Viewing the lineage of data in a Tableau report requires stitching that view back to the source data.

Detailed Drill-Down Capability

Allow both business and technical users to track transformation logic at an appropriate level of detail for each audience, leveraging a reliable drill-down lineage feature. This feature starts from a summary level and enables users to drill down to the smallest details, including the transformations taking place at each step.

Automatic Lineage Derivation

Understand how data transforms at each step by automatically deriving lineage from code used to restructure, transform or merge data, including SQL scripts, stored procedures, BI reports, ETL jobs and mappings, eliminating the need for manual documentation of lineage and accelerating efficiency.

Data Similarity Discovery

Identify similar datasets across various sources using AI-powered data similarity discovery to "infer" data lineage. This capability significantly improves operational efficiency and reduces costs by detecting and eliminating duplicate assets. Impact analysis can help understand the potential impact of such changes.

Data Relationship Discovery

Understand both data flow and "control" relationships to conduct a thorough impact analysis. For instance, a column deletion used in a join (the operation of combining data from two or more tables) can impact a report relying on that join. The AI-powered solution deduces joins to enhance impact analysis, even when relationships aren't documented.

Deep and Broad Metadata Connectivity

Automate the extraction of metadata that is deeply buried in your most complex data sources with broad and deep metadata connectivity that spans multi-cloud and on-premises environments.

With these capabilities, you can gather metadata across:

- Cloud platforms
- BI tools
- Databases
- Multi-vendor ETL and data science tools
- Various enterprise applications and file formats
- SQL dialects
- Stored procedures

You can also obtain complete column-level data lineage, as shown in Figure 3, including a full inventory of all potential data lineage sources with rich details. For automated data lineage, scan both static and dynamic code and perform language parsing.

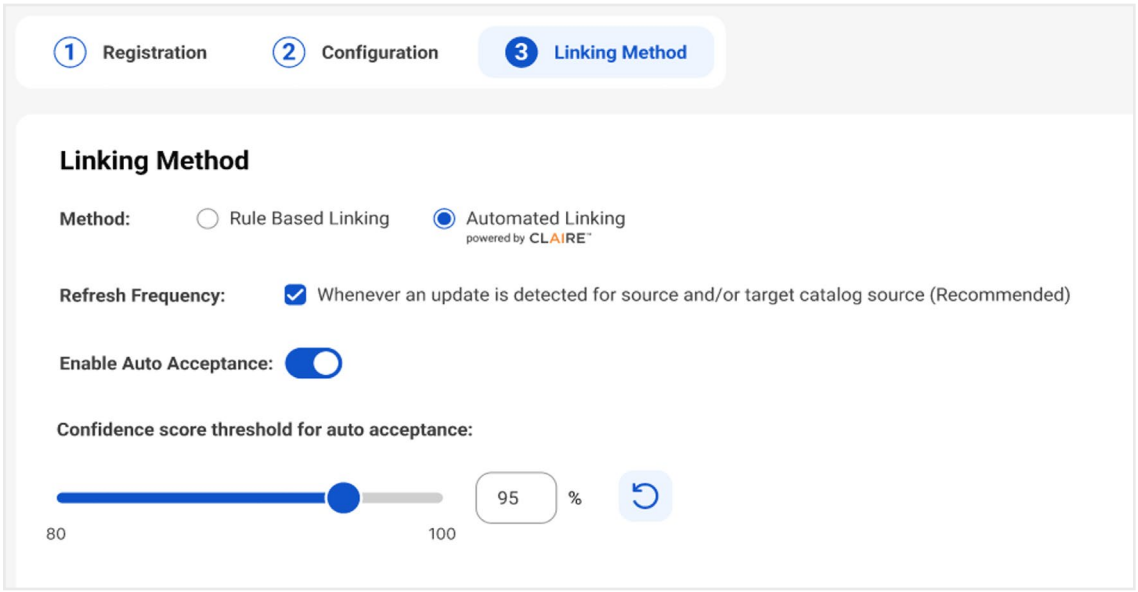

Figure 3: Informatica Data Catalog showing calculations in lineage

Deliver the Data That Builds Confidence in Your Analytics and AI

Invest in an intelligent, enterprise-scale data catalog designed for multi-cloud and on-premises environments that encompasses all eight capabilities and empower your organization to succeed in the new age of AI systems. This comprehensive catalog includes critical capabilities such as AI-powered data lineage, which aids in enrichment propagation and curation use cases by providing enrichment recommendations.

Automated, end-to-end data lineage is a key feature of Informatica® Cloud Data Governance and Catalog (CDGC), which combines the capabilities of data governance, data catalog and data quality into a singular tool for automating data intelligence insights. CDGC provides deep connectivity to a broad range of data sources across cloud, multi-cloud and on-premises data environments and applications. It allows users to track and view data lineage from origin to consumption across even the most fragmented and complex data landscapes.

Deliver the data that builds confidence in your analytics and AI models by utilizing inferred data lineage. Improve customer experience programs, help ensure regulatory compliance with industry policies, accelerate cloud modernization initiatives and much more. Your business users can enhance governance and privacy, deepen data analytics, transition to the cloud and augment the customer experience with greater ease and assurance. Concurrently, your IT teams and data analysts can refine change management, improve operational efficiency, reinforce data security and enhance responsible AI governance.

To learn more about how Informatica can bring your data to life, visit **www.informatica.com**.

Where data & AI come to

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.™**

IN17-5114-0125