Informatica

eBook

Unlock the Benefits of AI-Powered Data Lineage to Boost Your Business Success

Where data & AI come to LIFE™

Contents

Introduction

As businesses increasingly shift towards complex, multi-cloud environments amid evolving regulatory landscapes, understanding and managing the lifecycle and integrity of data has never been more critical.

Leveraging AI technologies in **data lineage** tools can help you maintain high **data quality** and security standards, thereby reducing data redundancies, streamlining business operations and ensuring compliance with burgeoning global regulations involving the use of data and AI.

AI-powered data lineage can facilitate enhanced **data governance** and provide deeper insights into data transformations to drive informed business decisions.

Welcome to a new era of data lineage.

In this eBook, you'll learn how AI-powered data lineage helps you achieve your data-driven business objectives — and what to look for in a solution.

Part One

Data Lineage Is Crucial for Modern Business

Data lineage has become a critical component in business operations amid the advancements and complexities of Generative AI (GenAI), digital transformation and cloud modernization. Its role in tracking the origin and transformation of data is key to delivering trusted data and unlocking its value-creation potential. As organizations invest in innovative AI technologies, today's businesses require enterprise-wide visibility and AI-powered intelligent capabilities to deliver effective data governance.

From a recent Informatica CDO survey, top data strategy priorities for 2024 include delivering reliable and consistent data fit for generative AI (39%), improving data-driven culture and data literacy (39%) and improving governance over data and data processes (38%).[1]

[1] **CDO Insights 2024: Charting a Course to AI Readiness**

Data lineage can be a key element to achieve these goals by increasing data transparency.

Understanding data prior to its use is crucial for all enterprises. Data lineage is about discovering the origin of data and how it is transformed throughout its lifecycle. This information provides the visibility and traceability organizations need to help data professionals trust their data. By comprehending the origin of data (provenance), its user and their manipulation of it, a strong foundation is created to fuel any data-driven project that can help address critical strategic questions:

- How can data across the enterprise be identified for alignment with regulations, including the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPR) and others?

- What do more recent regulations such as the EU AI Act (and others) mean for data discovery and understanding risk exposure?

- What data should be prioritized during a cloud modernization journey, specifically for migration to a cloud data warehouse and who would be affected by this transformational shift?

- Where should data scientists gather reliable data for advanced analytics insights?

- Which data sources should serve as the basis for developing new customer experience initiatives?

Part One

Data Lineage Is Crucial for Modern Business (continued)

Data lineage requirements extend to both technical and business users, ranging from data architects, engineers, stewards, analysts and scientists to business managers, compliance professionals and tech experts. Data lineage empowers these diverse roles by giving them a deeper understanding and confidence in their data. It helps visualize the end-to-end path your data takes from its origin and facilitates the evaluation of data accuracy and quality, leading to data intelligence with new and improved insights.

The Next Evolution of Data Lineage Is Here

Data lineage gives you a graphical representation of how data flows from its origin to its destination, visualizing the different processes the data undergoes and its dependencies, as well as how data is transformed along the way. In the contemporary data landscape, amplified by the complexities of cloud modernization and the implications of technological advances, including GenAI, consolidating this information is a challenging task.

Data originates from various heterogeneous sources scattered across the organization, including on-premises systems, hybrid IT and multi-cloud environments. Enterprises might be dealing with hundreds or even thousands of such data sources, spiraling into tens of millions of data objects.

Informatica's CDO survey points to 41% (of respondents) already struggling with 1,000+ sources and 79% expect that number to increase in 2024.[2]

With the overwhelming volume and scale of data today, manually tracing data's route through system infrastructure is impossible.

Many tools and technologies offer limited lineage capabilities, providing a siloed view of data within an individual source — for example, lineage within Hadoop grants visibility only up to your Hadoop clusters. Relying on source-level lineage gives you a piecemeal perspective of your enterprise data.

Recognizing these challenges, enterprises have started pivoting towards the next evolution of modern data lineage tools to scale out their digital businesses. This transformation is powered by AI and machine learning (ML) capabilities and bolstered by expansive metadata connectivity to data sources. It enables granular drill-down views or high-level overviews for business analysts. More importantly, it automates the stitching together and inference of end-to-end lineage across both on-premises and cloud ecosystems, thereby strengthening a cloud modernization journey.

[2] **CDO Insights 2024: Charting a Course to AI Readiness**

Part Two

Four Core Data Lineage Use Cases

Data lineage is vital in managing a wide range of data-driven business requirements. Let's examine the key use cases in more detail.

(1) Ensuring Regulatory Compliance

Data lineage is critical for responding to regulators' increasingly frequent requests for granular transparency across a variety of financial regulations, such as Basel Committee on Banking Supervision's standard number 239 (BCBS239), the Markets in Financial Instruments Directive (MiFID II) and the Comprehensive Capital Analysis and Review (CCAR). Data lineage provides the required granular transparency to keep compliant with trusted business reporting for privacy and governance laws such as the GDPR, the California Consumer Privacy Act (CCPA) and the U.S. Foundations for Federal Evidence-Based Policymaking Act.

The recent European Union's Artificial Intelligence Act (EU AI Act)[3] creates new regulations to manage the opportunities and risks associated with AI implementation. The aim is to balance leveraging AI's transformative capabilities while safeguarding the public's fundamental rights and safety.

Data lineage plays a vital role in realizing this objective in several important ways:

1. **Transparency and Trust**: One of the EU AI Act requirements is transparency in AI operations. Data lineage aids this necessity significantly by tracking and visually representing data from its origin through its lifecycle, including any transformations it undergoes. This method fosters trust in the AI outputs as organizations understand how and where data is altered.

2. **Risk Management**: The EU AI Act categorizes AI applications based on their associated risks. Data lineage assists in this process by showcasing the data's journey, supporting comprehensive risk analysis.

3. **Data Accuracy**: For AI to generate meaningful results, it requires accurate and high-quality data. With data lineage, data stewards can track the data's origin and transformation, ensuring data accuracy and quality.

4. **Compliance and Auditing**: Adherence to regulations will be obligatory under the EU AI Act. Data lineage provides a transparent data map as it passes through AI processes, simplifying compliance audits. It enables immediate inspection and rectification of issues, ensuring compliance with GDPR and other regulations.

Part Two

Four Core Data Lineage Use Cases (continued)

5. **Documentation**: Firms will need to keep detailed documentation related to AI systems as per the Act. This stringent record-keeping includes data lineage to provide evidence of proper data management. By simplifying how comprehensive and complete trusted business reporting is derived, organizations can save time, effort and cost during audits and more easily demonstrate compliance.

Compliance with these regulations requires reporting that demonstrates a clear understanding of how key performance metrics are calculated. Data lineage empowers data stewards to authenticate the origins of data, monitor its movement through organizational systems and demonstrate its transformations in the process.

The benefits of data lineage extend beyond these crucial requirements. Data lineage exposes discrepancies in your data analysis, paving the way to proactively identify and fix gaps in required data. Furthermore, synchronizing data across various regulations simplifies the discovery of different datasets containing the same data, thereby minimizing data duplication. This drives operational efficiency and ensures smoother management of compliance processes.

2 **Improving Analytics**

AI-powered data lineage boosts understanding of data's relevance for analytic insights and helps assure the reliability of data-driven outcomes, which is essential for making confident and trustworthy business decisions. Data lineage provides both technical experts and business analysts' with greater confidence in the use of business-critical data by offering visibility into the organization's data, including its location and additional context for enabling effective use. Revealing the source and transformation of crucial data to enable transparency, data lineage supports identifying pivotal, often overlooked data for analytic insights that require complete information to be reliable. This capability is especially important for AI-related sources that suffer from bias and hallucinations when source data is incomplete. (AI hallucinations occur when an AI system generates responses or outputs that contain false or misleading information presented as fact.[4])

The dynamic field of advanced analytics is being transformed by the integration of analytics and AI-powered data lineage, which empowers data scientists to fully leverage their organizations' data assets. The integration of AI into data lineage tools is not just an enhancement; it's ground-breaking. It elevates the accuracy, efficiency and completeness

[4] https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

Part Two

Four Core Data Lineage Use Cases (continued)

of tracing data back to its origins to understand its transformation. By leveraging AI to automate, infer and refine data lineage processes, organizations can swiftly adapt to changing data strategy, identify and rectify issues through data observability and uncover meaningful insights that would otherwise remain hidden in the complexity of big data.

For organizations wanting to get an edge on the competition, there are many benefits of incorporating AI into data lineage tools. It marks the transition from manual, time-consuming and error-prone processes to a streamlined, accurate and agile approach. AI-powered data lineage provides a clear, interactive visualization of the data landscape, making it easier to manage data, understand its impact and implement governance across its lifecycle. This level of clarity and control empowers data stewards to confidently guide their organizations through the complexities of digital transformation, ensuring that data remains an accurate and trusted asset. This is not just an operational upgrade; it's a strategic advantage for any organization that aims to thrive in the data-driven age.

These important benefits include:

1. **Increased Transparency**: Data lineage highlights the data's journey from origin to endpoint, providing unparalleled transparency. This allows analysts to trust the data they work with and understand its context and pedigree, which is essential for accurate analysis.

2. **Improved Data Quality and Accuracy**: By tracing data back to its origins, organizations can quickly identify and correct errors or inconsistencies through data observability to help ensure robust data quality throughout the entire data lifecycle. This, in turn, allows data analysts to trust the data and make decisions based on accurate, timely and reliable information.

3. **Enhanced Regulatory Compliance**: As data privacy and compliance regulations become increasingly stringent, data lineage provides a detailed audit trail of data movement and transformation. This makes it easier to demonstrate compliance with laws such as the EU AI Act, GDPR and the CCPA.

Part Two
Four Core Data Lineage Use Cases (continued)

4. **Efficient Troubleshooting and Problem-Solving**: When issues arise, data lineage enables teams to swiftly pinpoint the source of the problem by tracing the data's movement. This capacity for rapid response not only saves time but also mitigates the potential impact of data-related issues in decision-making processes.

5. **Strategic Decision Making**: With a comprehensive understanding of data's origin, context and lifecycle, as illustrated in Figure 1, organizations can make more informed, confident, strategic decisions. Data lineage ensures that these decisions are based on solid, trustworthy data, thereby enhancing the overall quality and effectiveness of data analytics initiatives.
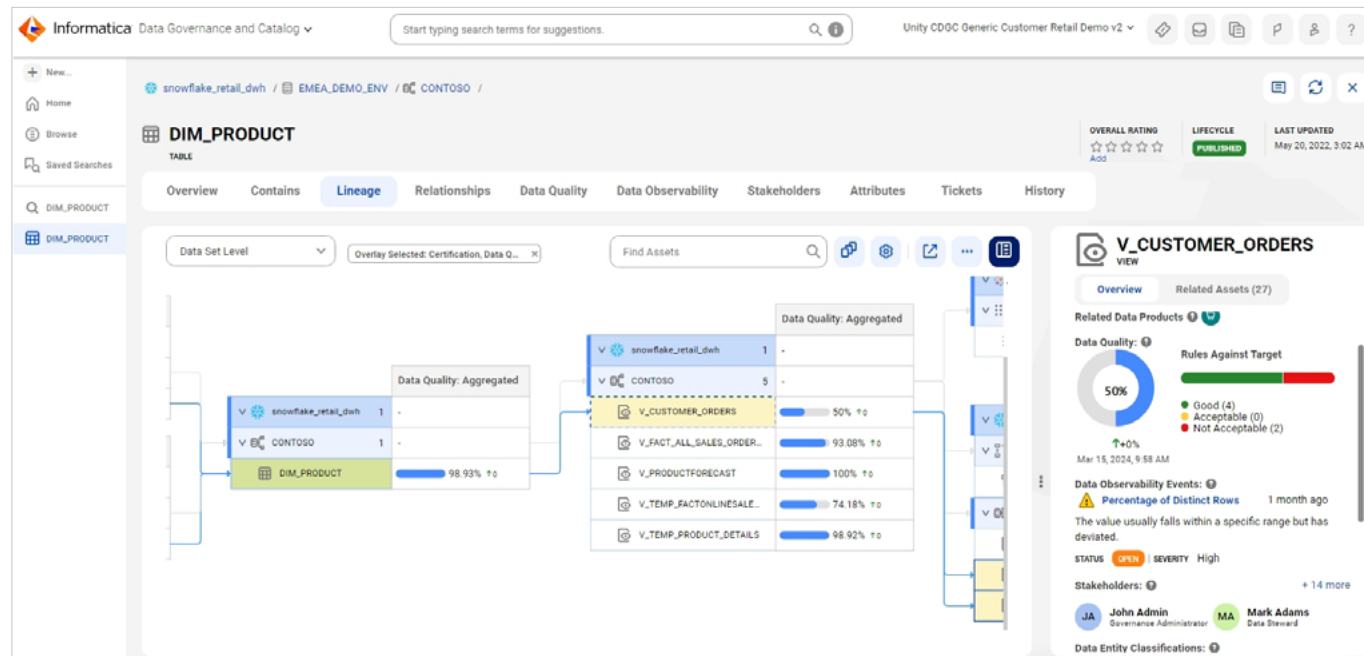

Figure 1: Informatica's Data Catalog asset overview in lineage

Part Two

Four Core Data Lineage Use Cases (continued)

(3) **Boosting Impact with Cloud Modernization**

Migrating your on-premises data warehouses and lakes to the cloud not only offers greater agility, flexibility and scalability, but also enhances data accessibility and comprehensive control. To ensure data migration is strategic and its impact managed efficiently, AI-guided data lineage becomes indispensable for inferring data movement and identifying gaps using automation.

By understanding the data journey, lineage can equip the IT team to review source data warehouse schema structures and identify necessary modifications. This approach can provide users with a blend of transparency and control, enabling them to efficiently manage, use and trust their information assets. Impact analysis helps you understand data usage, dependencies and how the migration will influence data assets, resources, users and the return on investment (ROI) of moving certain data assets and workloads to the cloud. Post-migration, comprehensive views into data lineage enable you to comprehend the data before and after it's moved, ensuring structural changes are executed accurately, maintaining data integrity and building trust in AI systems.

How is cloud modernization supported by data lineage?

1. **Trust and Compliance**: Data lineage provides full visibility into lifecycles from data origin to consumption. This enhances trust and regulatory compliance by tracing back to validate data when migrating from on-premises to the cloud. Compliance with regulations such as GDPR, CCPA and HIPAA requires visibility of the origins and destinations of sensitive data, including tracking cross-border data transfers. Data lineage improves quality by tracking data elements across the organization, ensuring system inputs and outputs are compatible and that business rules are accurate.

   By revealing critical data and its flow, data lineage enables the implementation of solutions to maintain data availability and enhance security. This transparency is a crucial part of **data governance**, encompassing the overall management of data ownership, quality, availability and security across the enterprise. By offering clarity on data location, usage, purposes and transformation, data lineage enables the assignment of ownership to relevant stakeholders who ensure compliance with internal and external regulations.

Part Two

Four Core Data Lineage Use Cases (continued)

2. **Data Quality**: Data lineage facilitates the identification of the root cause behind data quality problems. By understanding the journey of data throughout its lifecycle, it's possible to pinpoint where inaccuracies originate, enabling cleaner data and more effective data correction. Moreover, it aids in establishing standards for data migration, a crucial aspect when prioritizing cloud modernization for enhanced data visibility.

3. **Impact Analysis**: Data lineage helps to conduct impact analysis, as referenced earlier. If there's a proposed change to data, you can trace it back to its origins to understand what systems could be affected prior to migration. This can reduce post-migration risks by providing more clarity through business reporting, drawing data from numerous sources. Changes may include semantic alterations, modifications to business logic, data model restructuring, process adjustments and more. Data lineage is crucial for efficient change management in business intelligence and application development scenarios. Data Lineage can perform a "What If" analysis on potential changes at granular levels with business logic, facilitating an understanding of how the change will impact users, business processes and reports, providing trusted data.

4. **Data Discovery and Understanding**: Data lineage makes it easier to discover and understand critical data during cloud migration, prioritize it for migration and safeguard its journey if sensitive. It helps users to accurately locate the data they need, understand it and use it effectively once modernized in the cloud.

5. **Operational Efficiency**: Data lineage facilitates performance optimization, reducing costs and resources in the cloud by helping identify redundant processes and eliminating duplicate data across hosted **data management** services and pipelines. Data lineage can help simplify efforts to optimize your data footprint, reducing costs. Data lineage tools can help identify and remove duplicate data, detect and address data silos, locate, remove unused data and systems. It simplifies **data modernization** and migration programs, offering businesses quicker access to reliable data.

Part Two
Four Core Data Lineage Use Cases (continued)

6. **Security**: By tracking data from its origin to its current state, data lineage helps to enhance security by mapping data flow, which helps organizations identify data security and privacy vulnerabilities. Data lineage is combined with other cloud-native services to offer end-to-end visibility and reduce risk exposure, whereby data security failures can result in severe consequences. For instance, GDPR sets high penalties for data breaches, including fines of up to 10 million euros or up to 2% of your global turnover for the fiscal year preceding a data breach, whichever is higher.[5] Organizations must thus safeguard data to meet regulatory requirements and reduce risk. Data lineage renders visibility into sensitive data flow within your organization, ensuring appropriate controls at each step.
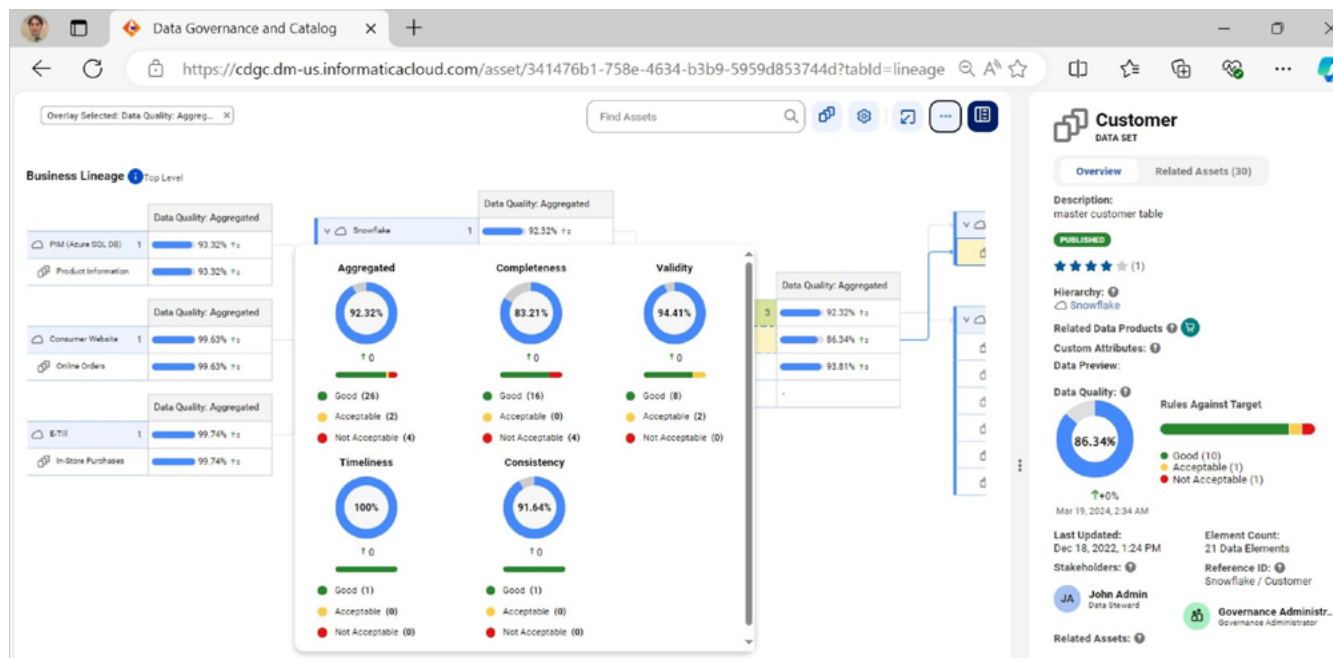

Figure 2: Informatica's Data Catalog showing Business lineage detailed data quality score

[5] https://www.itgovernance.eu/en-ie/dpa-and-gdpr-penalties-ie#:~:text=What is the maximum GDPR,global turnover - whichever is higher

Part Two

Four Core Data Lineage Use Cases (continued)

4

Enhancing Customer Experience

In today's consumer-centric world, the experience a business provides often outweighs product attributes or cost. For organizations to provide exceptional experiences, a holistic view of the customer and an understanding of their journey is paramount. Master data management (MDM) is critical to providing reliable consolidated customer data in a single source of truth. An AI-powered data lineage tool can help pinpoint information for your master data. It also provides insights into data quality and usage, which can deliver a more seamless, transparent and effective customer experience.

Leveraging an AI-powered data lineage tool can be instrumental in further enhancing customer experiences. Providing data visibility across its journey is of paramount importance because it allows your organization to:

1. **Locate specific points of information within master data**: This ease of traceability supports quicker decision-making, as business stakeholders can base decisions on data that's been tracked and documented, thus improving its reliability; for example, data lineage could be used to trace customer data history and build segmented profiles that can be used in targeted, personalized marketing campaigns.

2. **Gain insights into data quality**: By tracking how data is transformed across its journey, companies can understand where errors or inconsistencies are introduced and fix them at source. This ensures high-quality data that leads to high-quality insights, directly impacting the customer experience. For example, by tracking the lineage of customer call data and logging errors in a timely manner, you can improve customer support experiences.

3. **Understand data usage**: Seeing where and how data is used can indicate what information is most valuable for enhancing the customer experience. It can shed light on what data is used — by whom, when and why — painting a clearer picture of customer needs and interactions. For example, lineage can help you understand where users spend most of their time within an application. This could lead to user interface improvements or new feature development that makes the most commonly used functions more accessible.

4. **Facilitate greater transparency**: In an era where data privacy and security are of prime importance, being able to report on clear data lineage to customers can enhance their trust in a business and offer a more seamless, transparent customer experience. For example, a hospital that maintains clear data lineage can demonstrate to patients exactly how, where and why their data has been shared, thus building trust through greater transparency of their data records.

Part Three
Case Study | Data Lineage in Action

Lowell

Lowell Helps Millions Improve Financial Health with Data-Driven Services

Since 2004, the **Leeds-based Lowell Group** has been improving credit for European consumers and businesses.

Goal

As a leader in responsible credit management, Lowell Group, UK, was looking to make trusted data more readily available across its organization so that more teams can help consumers manage their debts in a personalized, affordable way.

Solution

Lowell partnered with Informatica to accelerate and automate consumer insight and help eliminate human error in data management. By leveraging Informatica's cloud-native platform, Lowell's various teams could finally collaborate around a common data foundation. **Informatica Data Governance and Catalog** solutions provided a structured business glossary, bringing much-needed governance to data across the Lowell group of brands. These tools helped facilitate discovery and classification and end-to-end lineage tracking of data across every department and public or private cloud. With **Informatica Data Quality** and AI-powered automation, Lowell cleansed and profiled assets further for greater accuracy.

Results

This initiative empowered 1,500 employees by providing them with a common data foundation that includes a user-friendly catalog, a comprehensive business glossary and an end-to-end lineage that spans both on-premises and cloud data assets.

"Our old approach made it difficult to dive deeper into our data, much less integrate it into our analytics. Informatica has been a game-changer. For the first time in Lowell's history, anyone in the company can search for and understand the language of the business and its associated data and lineage."

Robin Miller
Group Data Manager

Part Four

Eight Essential Features for Comprehensive AI-Powered Data Lineage

In summary, the complexity of today's data landscape demands data lineage solutions with advanced functionalities that leverage automation and AI. Here are seven capabilities your solution should possess:

1. **Automatic Data Lineage Stitching from Multiple Sources**
   The ideal solution should 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. Along the way, it gets 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.

2. **Detailed Drill-Down Capability**
   The solution should have a reliable drill-down lineage feature that allows both business and technical users to track transformation logic at an appropriate level of detail for each audience. This feature should start from a summary level and enable users to drill down to the smallest details, including the transformations taking place at each step.

3. **Impact Analysis with Shareable Exports**
   Whenever changes are made to a data source, for example, if business analysts are requesting changes or you're migrating a data warehouse to the cloud, the solution should divulge how the dataset is utilized based on end-to-end lineage, the impact of these alterations and who might be affected. Providing an exportable and shareable form of this information can simplify reporting.

4. **Automatic Lineage Derivation**
   Understanding how data transforms at each step is integral to comprehensive lineage visibility. A solution that can automatically derive lineage from code used to restructure, transform or merge data, including SQL scripts, stored procedures, BI reports, ETL jobs and mappings, eliminates the need for manual documentation of lineage, accelerating efficiency.

Part Four

Eight Essential Features for Comprehensive AI-Powered Data Lineage (continued)

5. **Change Notifications**
   The solution should automatically alert database administrators of any modifications made to predefined resources, datasets or the entire data source. This allows them to proactively manage reports using data lineage to understand what has been impacted by these changes.

6. **Data Similarity Discovery**
   AI-powered data similarity discovery can "infer" data lineage by identifying similar datasets across various sources. 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.

7. **Data Relationship Discovery**
   During impact analysis, understanding both data flow and "control" relationships is vital. For instance, a column deletion that's used in a join (the operation of combining data from two or more tables) can impact a report relying on that join. An AI-powered solution that deduces joins contributes to a more thorough impact analysis, even when relationships aren't documented.

Part Four

Eight Essential Features for Comprehensive AI-Powered Data Lineage (continued)

8. **Deep and Broad Metadata Connectivity**

   A cloud data governance and catalog solution should have broad and deep metadata connectivity that spans multi-cloud and on-premises environments to automate the extraction of metadata that is deeply buried in your most complex data sources. 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. Scan both static and dynamic code, as well as perform language parsing for automated data lineage.
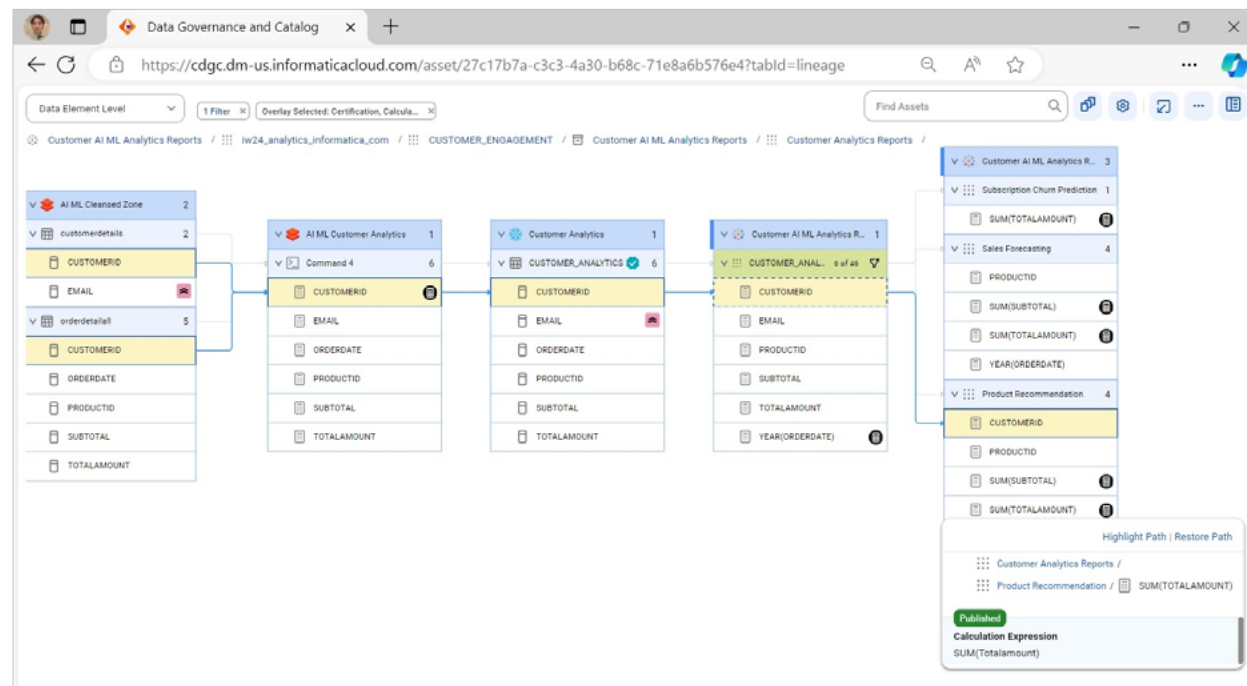

Figure 3: Informatica's Data Catalog showing calculations in lineage

Part Four

Eight Essential Features for Comprehensive AI-Powered Data Lineage (continued)

Investing in an intelligent, enterprise-scale data catalog designed for multi-cloud and on-premises environments that encompasses all eight capabilities can empower organizations to succeed in the new age of AI systems. AI-powered data lineage can also help in enrichment propagation and curation use cases by enrichment recommendations.

Making strategic decisions encompassing digital transformation while remaining compliant with an array of regulations, including new AI controls, requires trustworthy data.

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 data governance.

In today's data-driven corporate world, enterprises cannot depend exclusively on siloed data lineage tools. As businesses seek timely access to relevant, trustworthy data and the need for compliance intensifies, data lineage is gaining recognition as a fundamental requirement. Armed with AI-powered data lineage, you have the means to accomplish all your data-centric business goals.

Further Reading

Automated, end-to-end data lineage is a critical component of **Informatica® Cloud Data Governance and Catalog**, a service of **Informatica's Intelligent Data Management Cloud™ (IDMC)** that combines the capabilities of **data governance**, **data catalog** and **data quality** into a singular tool for automating data intelligence insights. Cloud Data Governance and Catalog 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 its origin to consumption across even the most fragmented and complex data landscapes.

The IDMC service enables organizations to leverage **predictive data intelligence**. This includes automated and recommendation-driven data classification, data curation, relationship building and sensitive **data discovery**, powered by Informatica's **CLAIRE® AI** and ML engine to promote next-best actions on data. This combined approach helps ensure that users can efficiently gather data intelligence insights and quickly drive business value from their data. Cloud Data Catalog enables you to scan and catalog assets across the enterprise to unleash the power of your data, combining multiple capabilities that enable you to discover your data, apply governance policies and ensure appropriate business strategy is enforced with the data for appropriate use. For more details, check out our **Informatica Cloud Data Governance & Catalog Data Sheet**.

Informatica's cloud data governance solution helps you govern your data responsibly and monitor its quality, ensuring trusted data and support for compliance regulations. Build a data governance and cataloging program strategy that improves business performance. Read our guide, **Cloud Data Goverance Adoption Guide: 8 Best Practices for Success**, for more information.

About Us

Informatica (NYSE: INFA) brings data and AI to life by empowering businesses to realize the transformative power of their most critical assets. When properly unlocked, data becomes a living and trusted resource that is democratized across your organization, turning chaos into clarity. Through the Informatica Intelligent Data Management Cloud™, companies are breathing life into their data to drive bigger ideas, create improved processes, and reduce costs. Powered by CLAIRE®, our AI engine, it's the only cloud dedicated to managing data of any type, pattern, complexity, or workload across any location — all on a single platform.

Where data & AI come to

LIFE™

Informatica

Worldwide Headquarters

2100 Seaport Blvd,

Redwood City, CA 94063, USA

Phone: 650.385.5000

Fax: 650.385.5500

Toll-free in the US: 1.800.653.3871

informatica.com
linkedin.com/company/informatica
twitter.com/Informatica

CONTACT US

IN19-3654-0724

informatica.com