Informatica

Solution Brief

Accelerate Analytics on Amazon Web Services with an AI-Powered Data Catalog

Key Benefits

- Empower 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 AWS and additional data sources
- Enable robust enterprise-wide data governance, privacy and regulatory compliance programs
- Accelerate your cloud journey with data intelligence

AI-Powered Rapid Data Discovery for Amazon Web Services (AWS)

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. 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, every enterprise needs to be able 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 extract 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.

But a recent survey of chief data officers, chief analytics officers, CIOs, CTOs and other senior technology leaders found that "just 13% of organizations excel at delivering on their data strategy."[1] How is this select group of high achievers delivering measurable business results across the enterprise? By building on a solid foundation of sound data management and architecture that can derive optimal value from the power of machine learning (ML).

In 2020, 64.2ZB of data was created or replicated. This growth is forecast to experience a compound annual growth rate (CAGR) of 23% over the 2020–2025 forecast period.[2] Although much of this data is being created in the cloud, many enterprises are running organizations across 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] MIT Technology Review Insights, "Building a high-performance data and AI organization," April 15, 2021

[2] 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)

One of the key challenges when operating in this complex environment 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 Amazon Web Services (AWS). 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 to AWS, this challenge is amplified when the metadata and data lineage is trapped and lacks transparency. This is often the case with metadata existing across various data sources, such as complex enterprise applications and systems, stored procedures for databases, data warehouses and multivendor ETL and BI tools. As a result, it is increasingly difficult to extract and understand business-critical data and analytics, posing operational and regulatory risks.

Building trust in AI and ML models and insights requires comprehensive visibility and understanding of source data. To accelerate actionable insights, data management professionals require a complete and unified data cataloging and governance foundation spanning data sources, across on-premises and multi-cloud environments. A foundation based on better, trusted data intelligence can enable enterprises to overcome counter-productive data and governance silos and accelerate valuable analytics and results with AI.

Informatica® Data Catalog enables enterprises to build a comprehensive inventory of metadata, inclusive of a variety of data sources such as Amazon S3, Redshift and AWS Glue. Powered by metadata-driven automation from the Informatica CLAIRE® AI engine, Informatica's intelligent data catalog enables rich, relevant context and 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 within and outside AWS at a granular level. Informatica Data Catalog is a foundational pillar for enabling a holistic and comprehensive data governance and data democratization strategy for all your data, regardless of where it resides.

Key Capabilities

Rapid Data Discovery Powered by Machine Learning

Informatica Data Catalog enables rapid discovery of data with powerful, semantic search capabilities — empowering data stewards, data scientists and analysts, 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 statistical and metadata-driven ML algorithms, Informatica Data Catalog tackles the inherent complexity in data. The solution discovers, tags, clusters, identifies similarities and patterns in data and captures system-wide data relationships, enabling enterprises to intelligently catalog all types of data at scale.

Informatica Data Catalog allows easy import of business glossary assets such as terms, policies and classifications from Informatica Data Governance, as well as third-party tools. You can 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
Informatica's AI-powered 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. These sources include structured and unstructured files. The solution automatically detects partitions in files in Amazon S3 and S3-compatible storage such as Scality RING. Additional data sources include tables and views from Redshift, RDS instances (Oracle, MSSQL, MySQL and PostgreSQL) and Athena; and databases, tables and metadata from AWS Glue.

Informatica Data Catalog has end-to-end data lineage and impact analysis capabilities, which allow you to easily visualize, trace and understand the flow of data within AWS. These capabilities also extend to linked data sources, such as Microsoft Azure and Google Cloud, enterprise applications and systems, databases and ETL and BI tools. You can perform detailed impact analysis of transformations within AWS, 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.

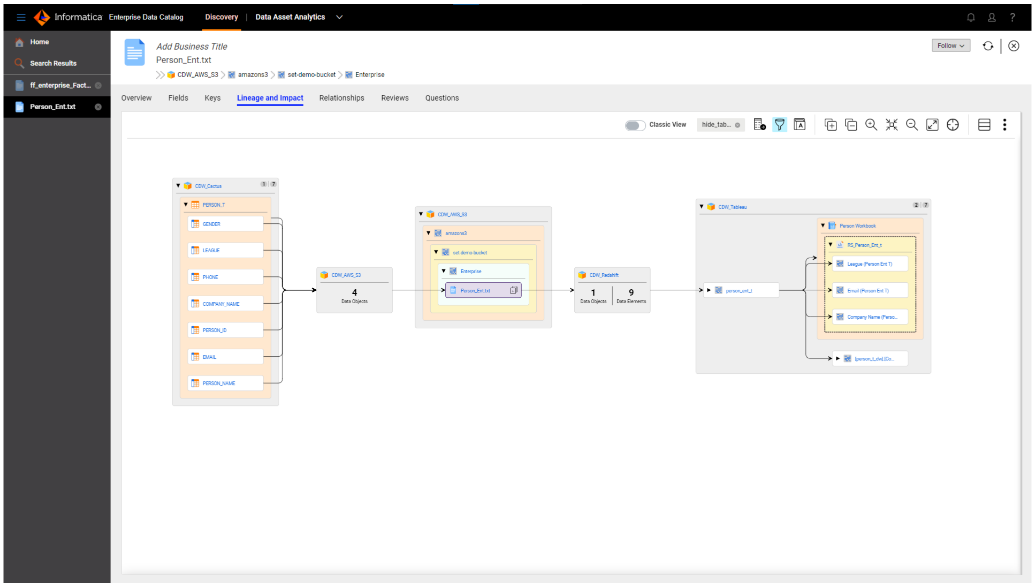

Figure 1: Informatica Data Catalog scanners enable detailed lineage for data impact and insights.

Purpose-Built Advanced Scanner for Amazon Redshift

Informatica Data Catalog Advanced Scanners are purpose-built for enabling deep extraction of metadata and deriving detailed data lineage for data intelligence. The Advanced Scanner connects seamlessly with Amazon Redshift to scan and extract metadata from SQL scripts (files), functions, stored procedures, materialized views and external tables for context. Support for Redshift enables teams to discover data asset dependencies and perform impact analysis. Data scientists, engineers and others can easily track data movement to identify related tables, views and domains to accelerate data-driven insights and analytics decision making.

Data Collaboration and Social Curation With Intelligent Crowdsourcing and Annotations

Informatica Data Catalog empowers data stewards, data scientists and data governance and analytics leads to easily find the most relevant and trusted data for analysis by harnessing the combined power of ML, human expertise and collaboration. Data owners and subject matter experts can certify datasets and provide ratings and reviews, enabling the social curation of data. A Q&A platform enables subject matter experts to answer common questions from users. The solution for AWS accelerates data discovery to help determine the best datasets for analytics use cases.

Integrated Data Quality

View data profiling statistics, data quality rules, scorecards and metric groups alongside technical metadata to understand the quality of data assets within AWS 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, enabling data leaders to share business adoption and value metrics with stakeholders. Automated Data Value Calculator, a first-of-its-kind capability, allows an enterprise to measure and optimize the value of its data assets based on key factors that impact data value. For instance, you can obtain information on what percentage of your data inventory exists in key data sources, as well as 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 AWS.

Key Benefits

Achieve Faster Time to Value With Trusted Insights

By parsing SQL scripts or other sources with the Informatica Advanced Scanner, data scientists, analysts, engineers and other data users can discover and apply relevant data, better arming them with the end-to-end lineage and in-depth data insights they need. The combined solution from Informatica and AWS allows enterprises to 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 use rapidly and cost-effectively with trust assurance, encompassing all your data — whether it resides on AWS or on alternative platform data sources.

Capture and Enforce Privacy Policies

As part of Informatica's complete Data Governance solution, data analytics users can leverage comprehensive, user-defined privacy policies to align stakeholders and workflows. Users can rapidly discover data that is subject to privacy regulation compliance that may reside in AWS as well as other data sources. 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 more).

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 in which data source systems it originates. End-to-end data lineage provides the intelligence needed to trace and understand the movement of sensitive data at a granular level. From there, users gain the intelligence needed to make informed decisions on data exposure, enabling the organization to get more value from its data.

Accelerate Migration to the Cloud

Migrating analytics and AI to the cloud offers improved economies and more agility for applications but requires the capability 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 AWS to help simplify your journey.

Cloud-Native or On-Premises Deployment

Advance your cloud-first, cloud-native data governance strategy with Cloud Data Governance and Catalog. Cloud Data Governance and Catalog is Informatica's software-as-a-service solution that brings Informatica Data Catalog natively to the cloud. Alternatively, you can deploy Informatica Enterprise Data Catalog as an on-premises data catalog solution.

Accelerate time to value and simplify consumption with Informatica Data Catalog support for Amazon AWS with extensibility to Microsoft Azure, Google Cloud Storage and BigQuery, Snowflake, Databricks, SAP and more. Data governance, data cataloging and data quality are enabled with scanners that support Amazon S3, Redshift, Athena and a full range of ecosystem partners, regardless of deployment method.

Next Steps

To learn more, visit our web pages for Cloud Data Governance and Catalog, Informatica Enterprise Data Catalog, Enterprise Data Catalog Advanced Scanners and Informatica on AWS.

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.™

Informatica®

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