Informatica®
CLOUD FIRST. DATA ALWAYS.™

6 Steps to Building Intelligent Cloud Data Warehouses and Data Lakes

Delivering value at enterprise-scale with the Informatica Intelligent Data Management Cloud™

Contents

Tip: click to jump straight to any section.

↑
⌂
↓

Introduction:

Next-Gen Insights Start with Efficient and Intelligent Data Management

The volume and complexity of data continue to explode. The reasons are many—cloud-driven data democratization initiatives; ongoing growth of unstructured and semi-structured data from sources like mobile and the Internet of Things (IoT); the rise of streaming data; the increasing complexity of heterogeneous hybrid and multi-cloud infrastructure; and mounting pressure to support next-gen analytics and machine-learning (ML) use cases—but the bottom-line impact is straightforward.

Data has become its own currency. The ability to tap data and to derive the maximum value from it—once a nice-to-have—is now critical for growth and survival in almost any industry. Organizations are racing to transform their legacy data platforms and infrastructure to better support exploding volumes of heterogeneously structured data and support modern analytics and machine learning (ML) workloads.

But successfully implementing cloud data warehouses and data lakes—and overcoming the challenges of data volume, complexity, scalability, security, trust, and velocity—is easier said than done. At the end of the day, modern cloud data platforms require modern data management solutions.

In this eBook, we'll explore how to successfully build a cloud data warehouse, and how to leverage comprehensive and automated intelligent data management to ensure it delivers value. Topics include:

- The benefits and challenges of successfully implementing cloud data warehouses and data lakes
- How to build your cloud data warehouse in six steps
- The role of intelligent, automated data management solution in maximizing the ROI of data transformation projects
- Examples of successful cloud data warehouse and data lake modernization in action

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization
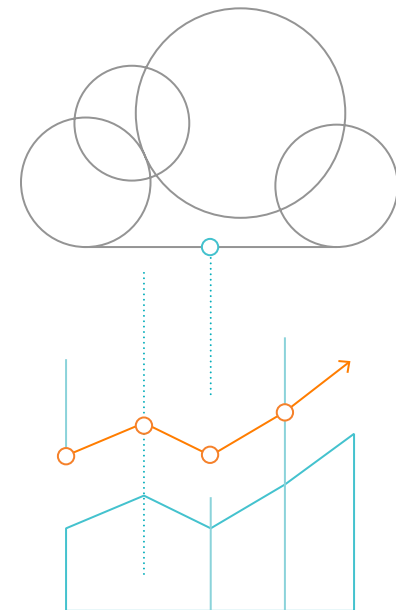
Cloud Data Warehouses and Data Lakes: Challenges and Opportunities

Data management has evolved significantly over the years—from the pre-internet days of countless disparate relational databases to centralized on-premises data warehouses, to Hadoop-based data lakes for higher volumes and less-structured data, and ultimately to Spark-based data architectures designed to leverage cloud efficiencies and simplify operational management.

Today, organizations are either building new cloud-native data warehouses and data lakes or modernizing their on-premises data warehouses in the cloud. According to a recent study from Ventana Research, 86% of organizations surveyed expect most of their data and analytics to reside in the cloud.[1]

But cloud data warehouse and data lake modernization are easier said than done. You need deep technical and platform-specific expertise; a comprehensive understanding of both your existing data ecosystem and future data opportunities; and a thorough, systematic implementation strategy—from data migration and integration, to pipeline optimization, governance and monitoring, and quality control.

As organizations embrace cloud data warehouses and lakes, cloud can also bring on cost overruns, resource constraints, and technology and implementation complexity. As a result, some may struggle to capture benefits like accelerated time to value and ROI. Let's look at the three key problems in modern data integration initiatives.

[1] Ventana Research, Avoiding Cloud Data Cost Overruns, 2021

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

The Cost, Constraints, and Complexity of Data Integration

Today CIOs and CDOs looking to optimize costs, increase productivity, and democratize data across the enterprise face three common challenges when it comes to data integration for analytics modernization.

Cost Overruns: CIOs and other IT leaders find it difficult to predict cloud costs, let alone manage them. They often lack visibility into costs and can't see who is using what cloud services and to what degree. Moving data into and out of cloud data warehouses and lakes incurs significant data transfer charges. Too often, organizations simply throw people and compute hours at these data processing problems instead of finding a more cost-effective solution.

Resource Constraints: CIOs and CDOs are challenged to find enough resources to do all the work that data-hungry organizations demand. Often, this work requires highly skilled and specialized resources. While IT struggles to meet demand, the business wants to move to the cloud even faster. Lack of resources and self-service capabilities make IT a bottleneck.

Complexity: As cloud adoption accelerates, data and IT leaders can find it difficult to connect cloud and multi-cloud with on-premises environments. Disparate tools don't work well together. And moving projects from development into production often takes too long because of system integration, data quality, and DataOps issues.

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

6 Steps to Build Your Cloud Data Warehouse and Data Lake: The ability to support large-scale, high-performance data processing, analytics, and ML initiatives starts with successful implementation. Here's a simple, six-step process to build your cloud data warehouse and data lake.
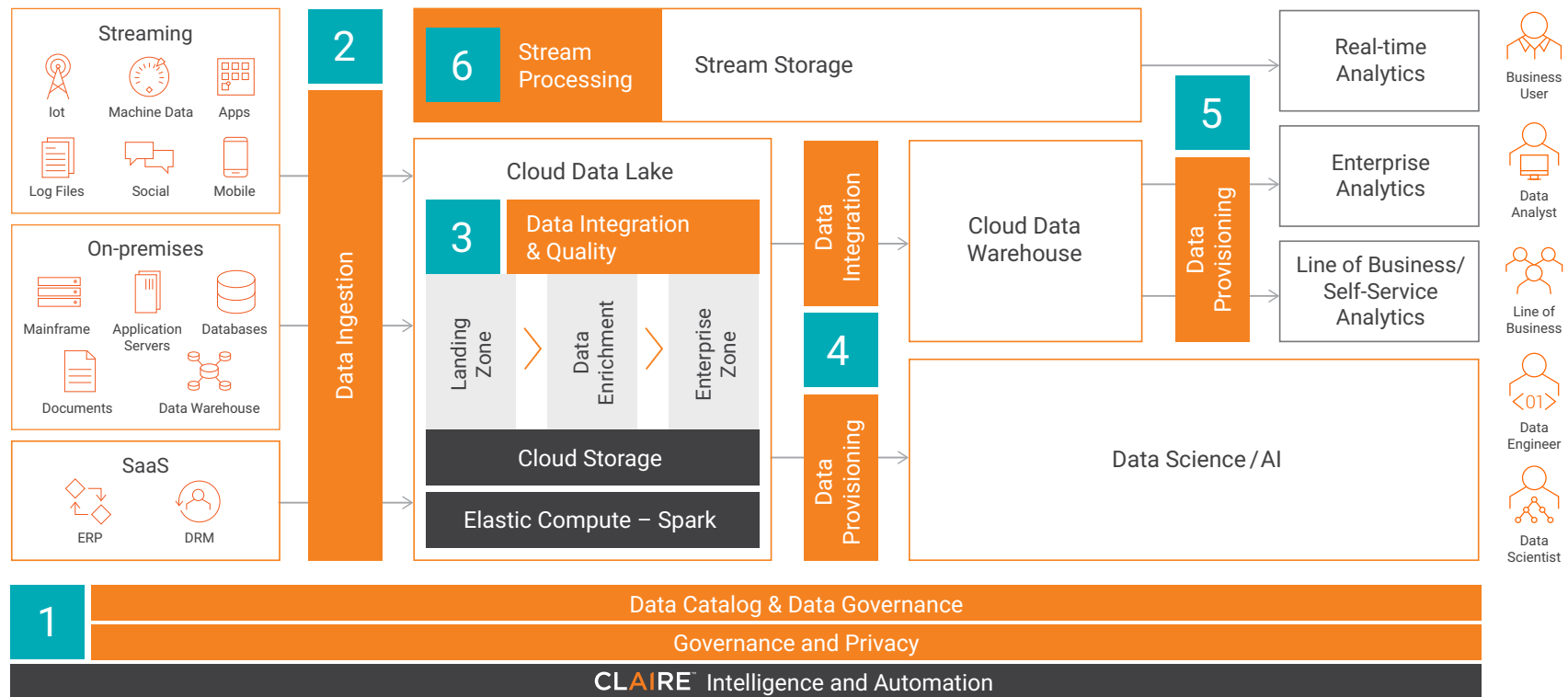

Figure 1: The above illustrates our six-step blueprint for building a cloud data warehouse and data lake.

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

1   **Enable access to trusted and compliant data** at scale, across the enterprise by cataloging and governing the data. An enterprise data catalog provides a common enterprise metadata foundation for intelligent, automated, end-to-end visibility and lineage across your environment. It enables you to bridge on-premises data stores with your modern cloud ecosystem to achieve maximum value without disrupting your data consumers.

2   **Efficiently ingest the data** from various sources, such as on-premises databases or data warehouses, change data capture (CDC) sources, SaaS applications, IoT sources, and streaming applications, into a cloud data lake, cloud data warehouse, or messaging systems to make it immediately available for real-time processing, database replication, and application synchronization. A unified ingestion solution should provide a simple and intuitive wizard-driven experience for ingestion, increasing productivity and lowering costs relative to hand-coding-based approaches. Without having to code data pipelines, you can ingest any data—such as real-time streaming and IoT data or transactional CDC and bulk data—from on-premises systems such as file systems, mainframes, data warehouses like Teradata, and relational databases like Oracle, SQL Server, and MySQL.

3   **Integrate the data** by cleansing, enriching, and transforming it via the creation of zones such as a landing zone, enrichment zone, and enterprise zone. Next-gen cloud-native data integration now makes it possible to build enterprise-scale integration workloads in minutes—connecting hundreds of applications and data sources across on-premises and cloud infrastructure. And to further simplify and speed up complex integrations, you can employ a modern, codeless UI to build new integrations using drag-and-drop interfaces; rely on pre-built templates to get teams up and running more quickly, even for very involved scenarios; and leverage purpose-built wizards to complete the last mile of integrations.

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

4 **Apply data quality rules** to cleanse and manage data while making it available across the organization to support DataOps. Identify and resolve data quality problems within your cloud data warehouse, so that all your data is clean and ready to use in your most critical applications and workflows. It's important to empower both IT and business users to collaborate to discover data, define and apply business rules, and monitor data quality within your cloud data warehouse. Your ability to drive intelligent decisions, reduce costs, generate growth, and quickly and reliably extract value from data is all directly contingent on your ability to protect your cloud data warehouse and data lake investments.

5 **Prepare the data** to ensure refined and cleansed data moves to a cloud data warehouse to enable self-service analytics and data science use cases. AI-powered enterprise data preparation enables raw data to be systematically discovered, blended, cleansed, and transformed so that the DataOps teams, including data engineers, data scientists, and data analysts, are empowered to turn datasets into trusted and governed information for AI/ML use cases. Enterprise data preparation is critical to bringing standardization and efficiency to the data preparation processes necessary to increase data accuracy, quality, validity, completeness, and conformance at scale.

Organizations also need modern enterprise solutions to cleanse data, so users can uncover errors, inconsistencies, and anomalies. Modern solutions offer easy-to-use graphical interfaces to hide complexity from users, and use ML features to accelerate data discovery, cleansing, and other preparation tasks. They also offer AI-driven recommendations that help users shape data preparation specifically to their needs.

6 **Use stream processing** to derive insights from real-time streaming data from sources such as Kafka, then move said data to a cloud data warehouse for further analytics consumption. This allows you to prepare and process streams of data and uncover insights while acting in real time to suit business needs. Streaming solutions can scale out horizontally and vertically to handle petabytes of data while honoring service level agreements (SLAs). They also offer data transformation and data enrichment capabilities to process the streaming data and make it available for operationalization and downstream analytics.

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

The Need for Cloud-native Intelligent, Automated Data Integration and Management

The steps outlined above provide a simple, achievable framework for building and deploying cloud data warehouse and data lake platforms. However, working with enterprise-scale data volumes demands taking an extremely efficient and scalable approach to each of the various data cataloging, integration, ingestion, and transformation processes required.

Many organizations initially rely heavily on manual methods, such as hand-coded data integration. But sooner or later, they discover that manual integrations and data pipeline construction are costly, inflexible, difficult to maintain, and unable to scale out data projects and support data self-service.

Getting the most out of your cloud data warehouse and data lake means finding and implementing the right tools and automation to handle data management and integration at-scale. Modern data integration solutions should provide best-of-breed ETL, ELT, and elastic Spark-based data processing for any cloud data integration need, allowing you to easily ingest, process, and enrich much larger volumes of data faster and more cost-effectively.

To truly maximize both time-to-value and the ROI of cloud data warehouse and data lake initiatives, you need a unified, cloud-native data management solution.

What Does Comprehensive, Cloud-native, Intelligent Cloud Data Management Look Like?

A unified cloud data management solution provides an end-to-end platform to catalog, ingest, process, prepare, transform, and enrich structured, unstructured, and semi-structured data in a governed manner. In other words: every cloud-native service you might need for data management. Organizations can improve data transparency, tap and integrate siloed data, and manage increasingly complex multi-cloud environments.

This approach enables people across your company—from business analysts to data scientists and data engineers—to access high-quality data quickly, easily, and securely for their analytics initiatives, driving innovation and providing organizations with a competitive edge.

A comprehensive cloud data management solution is built on three pillars: metadata management, data integration, and data quality.

↑
⌂
↓

Part One:

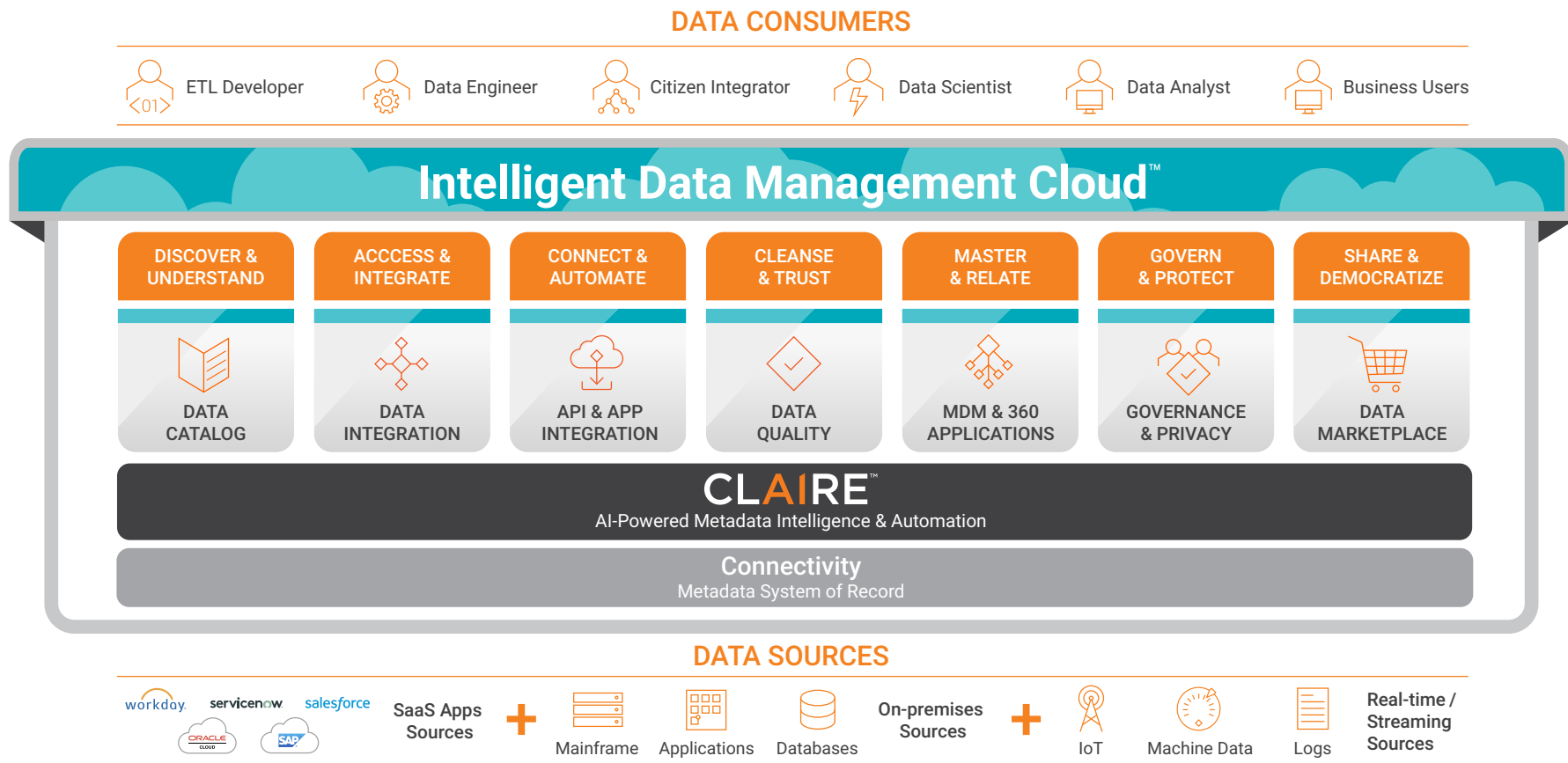# How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

Figure 2: Informatica Intelligent Data Management Cloud™ (IDMC) is the industry's first cloud-native and AI-powered data management platform.

Part One:

How to Maximize Value from Cloud Data Warehouse and Data Lake Modernization (continued)

All three should be entirely cloud-native, built on top of microservices and API-based architectures, with AI-driven automation baked in at every level:

- **Cloud-native/AI-powered:** Scale to support all enterprise workloads with elastic and serverless processing. Get faster insights when you apply AI and ML to your data and metadata.

- **Multi-cloud/multi-hybrid:** Most organizations run on multi-cloud and multi-hybrid environments. A modern data management solution should help you run, interoperate, and support any combination of cloud and hybrid infrastructures to support your data workloads.

- **Low code/no code:** A low-code/no-code experience empowers data practitioners to focus on what they do best: extract value from your data, and help ensure the highest levels of security, data quality, data governance, and privacy.

Maximizing ROI from Your Cloud Data Warehouse and Data Lakes with the Informatica Intelligent Data Management Cloud

Informatica's Intelligent Data Management Cloud offers the industry's first and most comprehensive AI-powered data management platform. Now you can manage and innovate with your data on any platform, any cloud, for any user in multi-cloud and multi-hybrid environments.

- **Boost productivity with an integrated approach to cloud data management:** Informatica provides enterprise-scale cloud data integration, data quality, and metadata management in an integrated, end-to-end solution. So instead of cobbling together tools and wasting time hand-coding integrations and data pipelines, you can speed data warehouse and data lake ROI.

- **Future proof your cloud data warehouse and data lake:** Migrating cloud data warehouses and data lakes to the cloud means supporting multi-cloud environments. An independent, best-of-breed cloud data management solution with built-in automation capabilities helps you evolve your cloud analytics alongside ever-shifting cloud platforms and infrastructure.

- **Improve agility with automation and codeless integration:** Using manual efforts to solve data quality and data management challenges reduces your ability to adapt and innovate. Hand-coding is complicated, expensive, and prevents you from reusing code as your infrastructure evolves and changes. Advanced metadata-driven AI capabilities give you flexibility and speed: You can quickly build data pipelines to fuel your cloud data warehouse and data lake—and reuse your work across cloud platforms and processing engines.

↑
⌂
↓

Part Two: Case Studies

Cloud Data Management in Action

BANCO ABC BRASIL

Banco ABC Brasil: Improving Service by Accelerating the Credit Application Process

Banco ABC Brasil has a wealth of financial and customer data, but most of that information was difficult to access. Not only was extracting data from source systems a manual process, but by the time the data was available to analysts, it was often out-of-date. To have the right data on hand to evaluate the credit worthiness of applicants, the bank needed to make data more accessible, actionable, and auditable. It also wanted to improve data governance and data quality to help streamline the credit application process and reduce risk.

They selected Informatica's Intelligent Data Management Cloud for integration platform as a service (iPaaS) and the AI-powered Informatica Enterprise Data Catalog. These solutions positioned the bank to better understand its data using a full business glossary. Banco ABC Brasil also chose Informatica Cloud Data Quality

to run automated data quality checks to validate that the data going into the lake is ready for analysis, while Informatica Cloud Application Integration automates Banco ABC Brasil's credit analysis process.

By moving to a cloud data lake, Banco ABC Brasil has improved data self-service capabilities, scaled data analytics, reduced costs, and monetized its data assets. Using Informatica's ELT capabilities, the data integration process takes place on an intermediate server before it is loaded into the target. And, because the data is already on Google Cloud, the team doesn't need to worry about moving data once it is ingested. Analysts now have a consistent experience across data types, allowing them to build better predictive models of client credit worthiness. This means credit decisions are turned around up to 30% faster – and customer needs are met sooner.

"Informatica's deep understanding of what we want and need is outstanding. It's why we've formed a lifetime alliance for data innovation that helps us disrupt traditional financial services."

— **Rodrigo Andreos Cordeiro**
Chief Data Officer, Data Protection Officer, Chief Innovation Officer and Chief Information Security Officer, Banco ABC Brasil

Part Two: Case Studies

Cloud Data Management in Action (continued)

Intermountain® Healthcare

Intermountain Healthcare: Driving Positive Healthcare Change with Self-service Analytics

Informatica helped Intermountain Healthcare solve most of their data challenges by combining different sources of data and making it available for analytics in near real time. Informatica Intelligent Cloud Services (IICS), our industry-leading integration platform as a service (iPaaS), helped segregate the datasets and establish access controls and permissions for different users, strengthening data security and compliance. Intermountain began converting approximately 5,000 batch jobs to use Informatica PowerCenter, Informatica Cloud Data Integration, or both.

Data is fed into a homegrown, Oracle-based enterprise data warehouse that draws from approximately 600 different data sources, including Cerner EMR, Oracle PeopleSoft, and Strata cost accounting software, as well as laboratory systems. Affiliate providers and other partners often send data in CSV files via secure FTP, which IICS loads into a staging table before handing off to Informatica PowerCenter for the heavy logic.

"We chose Informatica because we felt it could solve most of our data challenges, either on-premises or in the cloud. We see Informatica Intelligent Cloud Services as being able to close the gap between the well-known capabilities of PowerCenter and newer technologies as they come out."

— **Siju Nedungadappallil**
Data Architect,
Intermountain Healthcare

Conclusion:

Delivering Value with a Cloud Warehouse and Data Lake

Cloud data warehouses and data lakes offer the potential for faster insight delivery at greater scale. It is essential to have a plan for data management that includes the cloud data warehouse and/or data lake, all its data sources, and its projected growth.

The key to unleashing the power of data and ensuring the cloud data warehouse or data lake delivers business value quickly is to select the right cloud data management solution. This solution must handle any data and any integration pattern, offer comprehensive data management capabilities, and enable you to start small and scale with the needs of your business. Visit our cloud modernization solution page to learn more about how a comprehensive cloud-native data management platform can speed time to value and boost ROI.
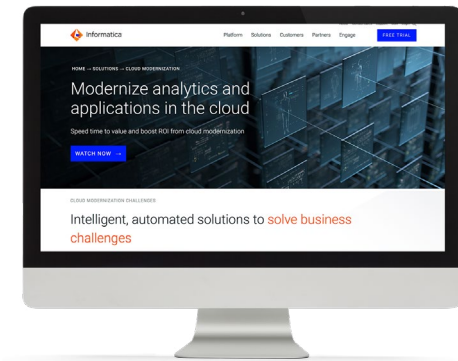
LEARN MORE

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™.

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

IN19-0122-3460

Informatica®
CLOUD FIRST. DATA ALWAYS.™