Informatica

Extract Value from Your Data with AI-Powered Data Discovery

Use metadata-driven intelligence and automation to find and understand data at scale

Contents

Tip: click to jump straight to any section.

Too Much Data, Not Enough Value

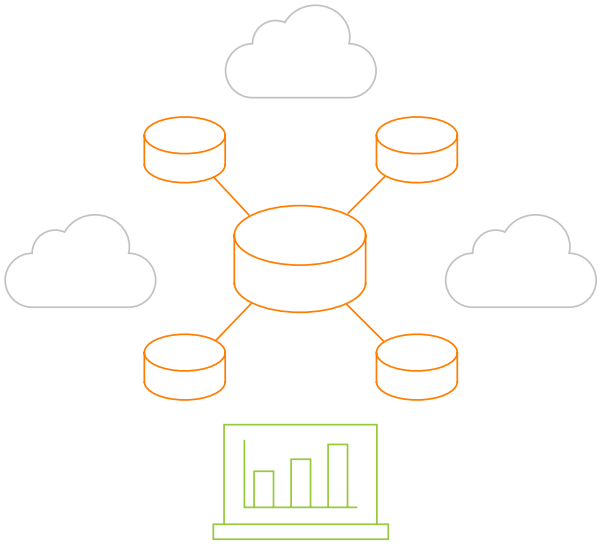
Data is the foundation of digital transformation. Yet for far too many businesses, finding and validating data that can drive value is like looking for the proverbial needle in a haystack.

Enterprises today operate in an increasingly complex multi-cloud environment where data is dispersed and siloed across hundreds or even thousands of data sources. Most organizations have diverse data distributed across innumerable complex on-premises and cloud-based enterprise applications, data lakes, data warehouses, databases, mainframe systems, to name a few. And enterprises are drowning in this complex data landscape with no end in sight as data continues to proliferate at an exponential rate. According to IDC[1], worldwide data will surpass 175 zettabytes by 2025.

In such conditions, data is difficult to find, understand, trust, and extract value from. In fact, Forrester Research reports that between 60% to 73% of all data that enterprises possess goes untapped.[2]

The monetary and competitive costs of not having a handle on what data you have, where it resides, who owns it, and the relationships between it, are staggering. According to research from IDC, "Data workers spend 70% of their time searching for and preparing data, 20% of their time governing data, and only 10% of their time performing analytics."[3]

Given the vast amounts of enterprise data today, data discovery at scale is possible only with the help of AI and automation. In this eBook, you'll learn how intelligent, automated data discovery that's powered by advanced machine learning algorithms can help you find and validate enterprise data accurately and efficiently across a widespread global IT environment, so you can maximize its value.



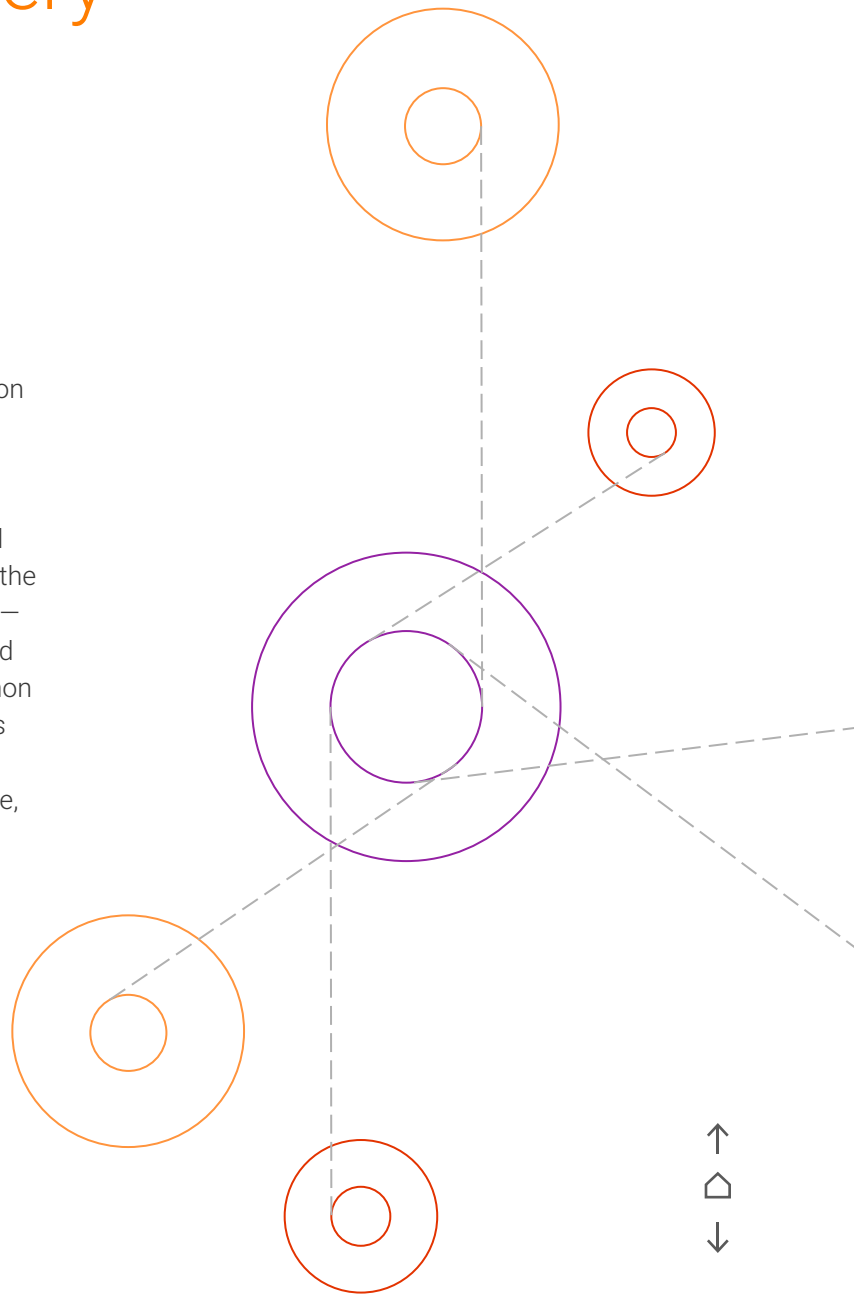
1 Network World, "IDC: Expect 175 zettabytes of data worldwide by 2025"
2 CEO Magazine, "Underuse of Analytics could be costing organisations millions"
3 IDC, "Automating Intelligent Data Governance"

Defining AI-Powered Data Discovery

Before you can extract value from your data, you first must discover and understand it.

But in a complex, petabyte-scale data landscape where data resides across a multi-cloud environment, manual data discovery is not only time and cost prohibitive, it is unsustainable. What's more, data must be validated and governed before it can be democratized for enterprise value creation programs. The solution lies in using artificial intelligence to scan and catalog enterprise metadata at scale.

An intelligent, automated solution for data discovery can provide context and organization for enterprise data. Such a solution is able to scan the metadata for tens of millions of enterprise data objects in minutes or hours (versus months), gathering metadata from all your enterprise systems regardless of where the data may reside—on-premises or in the cloud—including technical, business, operational, and usage-based metadata. The result is a common metadata foundation for the enterprise that is then used for discovery techniques such as intelligent search, domain discovery, and more, for a comprehensive understanding of your enterprise data.



Defining AI-Powered Data Discovery (continued)

Here are four defining characteristics of automated data discovery powered by metadata-driven AI:

1 **Intelligent data classifications:** An AI-powered data discovery solution can classify data fields by applying semantic tags (that is, data domains) to columns. An intelligent solution then automatically applies the same tag to similar columns in different tables, for instance by identifying domains like credit card number or account number and tagging them with that label. Once data domains are discovered, an intelligent and AI-powered data discovery solution can assemble individual fields into higher-level business entities. For instance, an entity like "Customer" is made up of multiple data domains, including name, account number, address, email, phone number, and so on.

2 **Intelligent, semantic search:** Automated data discovery powered by advanced machine learning algorithms and natural language processing (NLP) enables Google-like semantic search to dramatically speed the process of finding data across diverse, distributed enterprise systems.

3 **Data profiling:** Data profiling enables you to identify and understand data value frequencies, value distribution, anomalies and missing values, and so on. An AI-powered data discovery solution enables profiling so you can identify data quality issues and increase data accuracy across the organization. Profiling statistics can also be used to identify similar datasets and data relationships, which further enhances the ability to discover and understand data in a holistic manner.

4 **Automated, end-to-end lineage:** End-to-end data lineage capabilities make it easy to understand where data originates and how it changes over the entire data lifecycle, increasing trust and confidence in the data. Comprehensive data lineage gives you visibility into data dependencies, so you know where data is used and who's using it. An AI-powered data discovery solution with advanced relationship discovery can identify relationships, such as joins and primary keys, by inferring joins across datasets at scale. As a result, you can understand how a change will potentially impact users, business processes, and reports.

Customer Story

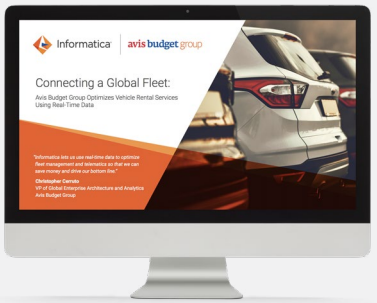
avis budget group

Avis Budget Group Spurs Innovation Using In-Depth Data About its Vehicle Fleet

Avis Budget Group offers some of the most recognized brands in the vehicle rental industry, including Avis, Budget, and Zipcar. As emerging competitors challenged the company to innovate and offer customers new experiences and services, Avis wanted to leverage terabytes of real-time data from and about its diverse fleet of 650,000 vehicles. This data came from GPS and navigation systems, Internet of Things (IoT)-enabled sensors, and ever advancing technology from vehicle original equipment manufacturers (OEMs).

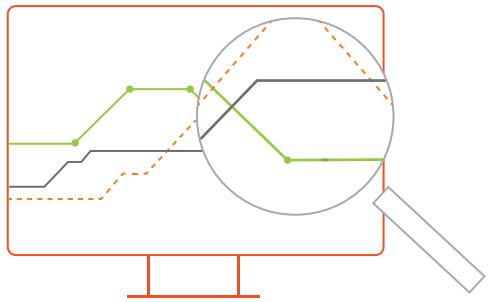
Avis Budget Group needed to profile and govern telematics data to uncover any data quality issues that might introduce business risk. It was also looking to document vehicle attributes and telematics data in an enterprise data catalog that could provide business context and capture knowledge from subject matter experts. In addition, it wanted to speed time to market for advanced analytics by giving end users simple tools to prepare data from the catalog for self-service analytics.

The car rental company now uses intelligent data discovery to provide visibility into data location, lineage, and business context. As a result, Avis Budget Group is able to organize fleet and telematics data so fleet managers have faster access to the data they need to optimize fleet management, save money, and drive the bottom line.



READ MORE

Key Data Discovery Use Cases



AI-powered data discovery is an essential first step to today's most critical digital transformation initiatives.

Advanced Analytics

Businesses today are racing to leverage AI and machine learning (ML) for advanced analytics and data science initiatives that can help them seize competitive advantage and deliver industry-changing innovation. But AI and ML projects require vast amounts of data. Intelligent, automated data discovery enables data scientists and DataOps teams to rapidly find the data they need. In iterative ML projects that require additional data assets, AI-powered data discovery can help DataOps teams find trusted datasets to make available in data pipelines and streamline data preparation.

Customer 360

Delivering the experience customers expect requires your organization to have a 360-degree view of your customers and an end-to-end view of each customer's journey and experience with the company each step of the way. Intelligent data discovery enables you to build the data foundation for successful customer engagement by discovering and cataloging customer data wherever it resides, identifying and classifying customer data domains, profiling data quality, and providing visibility into data lineage across customer touchpoints. Such a foundation enables you to understand customers' history, preferences, consents, relationships, and products owned in addition to finding important customer attributes across data sources and integrating them with the master customer model.

Modernizing Data Warehouses and Data Lakes in the Cloud

Organizations today are modernizing their traditional on-premises data warehouses and data lakes in the cloud to take advantage of cloud agility, flexibility, and cost savings. Intelligent, automated data discovery delivers end-to-end visibility and lineage across your environment so you can perform detailed impact analysis across data assets, resources, and users. Such analysis can help you demonstrate the cost benefits of moving certain data assets and workloads to the cloud, prioritize which ones to move, and develop a plan to minimize disruption, while providing ongoing visibility into the data for users.

Data Governance and Privacy

Modern data governance leaders, along with data stewards throughout the organization, need a solution that automatically discovers and identifies key data elements that need to be protected in order to safely enable use with revenue-generating applications. Intelligent, automated data discovery helps you discover and classify data—including sensitive data such as personal information—so it can be trusted, understood, and secured. This helps ensure compliance with data privacy mandates, while reducing risk of abuse that could violate consumer rights policies or result in a data security breach. And by automatically relating data to business terms and definitions, an intelligent, automated data discovery solution can provide business context to data for enterprise data governance and fuel business value with trusted data.

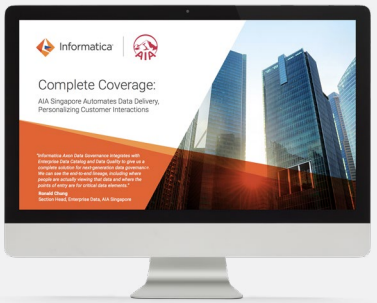
Customer Story

AIA Singapore Discovers Data Relationships Across Terabytes of Data

AIA Singapore is a leading insurance company that has served generations of Singaporeans for close to 90 years. In response to growing demographic challenges and technological advances, AIA Singapore was looking for deeper insights that would allow it to connect with customers in a more personalized manner.

To deliver the personalized services its customers expect, AIA Singapore wanted to increase enterprise-wide understanding of its business data, data standards, and policy holders. It was also looking to discover and gain a deeper understanding of data in context based on lineage and intelligent metadata.

By using AI-powered data discovery, AIA Singapore was able to rapidly discover, tag, and understand data relationships across terabytes of datasets. The company gained a better understanding of its business terms, data lineage, and the quality of data at the source. Ultimately, AIA Singapore used intelligent, automated data discovery to help develop a complete next-generation data governance framework that enables the company to win more business and retain existing customers.



READ MORE

Five Key Techniques for AI-Powered Data Discovery

An AI-powered data discovery solution should apply a broad range of AI techniques and machine learning algorithms to enterprise metadata across technical, business, operational, and usage categories. Here are five key techniques to look for in a solution:

1 **Intelligent Data Classifications with Data Domains**

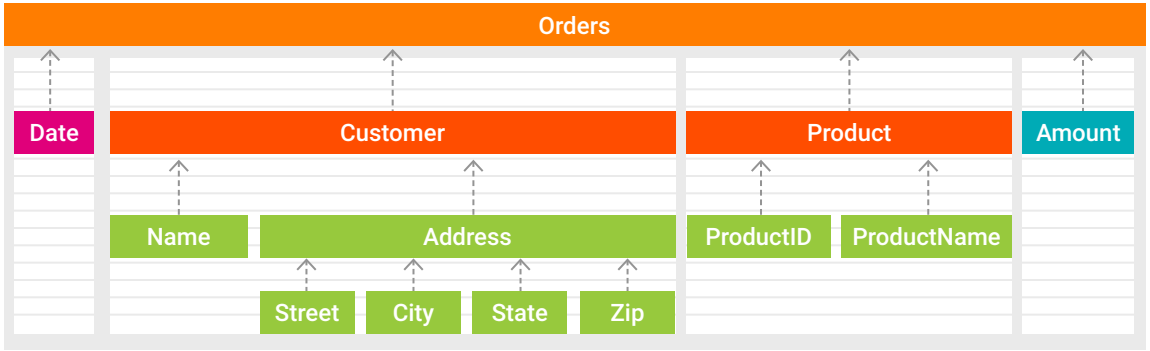
Advanced machine learning-enabled data discovery classifies data fields by applying semantic labels, called data domains, to each column.

Traditional techniques of applying semantic labels involve evaluating rules based on regular expressions (RegEx), reference tables, or other complex hand coded logic. While these techniques have a place in data discovery, defining and maintaining thousands of these rules involves significant manual effort. ML-based approaches to discovering and labeling data domains can dramatically simplify the process. For columns not already classified, the user provides a simple tag that labels the column content. The system learns by association and automatically propagates the tag to all similar columns. For comprehensive data discovery, you need a combination of RegEx- and ML-based approaches.

2 **Intelligent Entity Discovery**

Once domains for columns have been identified, data discovery powered by advanced machine learning algorithms can assemble individual fields into higher-level business entities. Entity discovery learns from the way users have assembled disparate data fields in their analytics or data-integration processes and applies this learning to derive entities across the enterprise data landscape. This classified data enables better search, filtering of search results, and business glossary recommendations.

The example below shows how an entity called Purchase Order is created by combining fields identified as Customer and as Product.



Combining data domains to detect entities from table and files.

Five Key Techniques for AI-Powered Data Discovery (continued)

3 **Advanced Relationship Discovery**
AI-powered data discovery significantly speeds the time it takes to use relationships between datasets to identify composite business entities (for instance, POs and invoices). Advanced machine-learning techniques automatically identify primary keys, unique keys, and joins across structured datasets, reducing months of documentation effort to minutes. These techniques work by combining profiling statistics like uniqueness, null counts, column metadata (e.g., column names containing "ID") and others to discover primary and unique keys. Machine-learning techniques such as column signature analysis then discover joins and join keys at scale across many potential datasets.

4 **Data Similarity and Clustering**
An Intelligent data similarity capability computes the extent to which data in two columns is the same. This capability helps users find the relevant and trusted data they need by identifying data, detecting duplicates, combining individual data fields into business entities, propagating tags across datasets, and recommending datasets to users. Using intelligent domain discovery techniques, an intelligent and automated data discovery solution can then propagate domain labels to similar datasets.

A brute-force approach to comparing all two-column pairs in an enterprise setting (which could have hundreds of millions of columns) would be computationally prohibitive. Instead, intelligent data similarity uses machine-learning techniques to cluster similar columns and identify likely matches. The process works by clustering columns with similar features and computing data overlap. For columns with enough data overlap to suggest they might be matches, similarity is computed using the Bray-Curtis and Jaccard coefficients.

Five Key Techniques for AI-Powered Data Discovery (continued)

5 **Sensitive Data Mapping**
Your organization needs to safeguard personally identifiable information (PII) and protected health information (PHI) data and comply with data privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA).

Intelligent data discovery leverages the power of advanced ML to identify sensitive data—it automatically evaluates and scores data that in combination can identify data subjects.

End-to-end data lineage and impact analysis capabilities then identify how sensitive data proliferates across repositories to support security and privacy compliance requirements. These capabilities can determine both upstream and downstream movement as well as related metadata, such as the specific type of data, process, protection status, and location of the data, to evaluate if violations have occurred.

For example, if personal data moves from a source to a target across geographic boundaries, you could be violating data sovereignty regulations. Or if data onboarded for billing processes is proliferated to other departments or locations for marketing processes, you could be violating privacy regulations. Intelligent data governance capabilities then notify policy or process stakeholders for remediation.

Customer Story

L.A. Care Health Plan Better Protects PII and PHI with Intelligent Data Catalog

L.A. Care Health Plan is one of the largest publicly operated health plans in the United States, offering Medi-Cal and Medicare plans to California's most vulnerable residents in Los Angeles County.

The plan has vast amounts of patient information to protect, govern, manage, and ultimately leverage for analytics efforts to help improve population health. But IT and data silos were preventing L.A. Care from easily identifying all of its sensitive data and enabling a comprehensive data governance strategy.

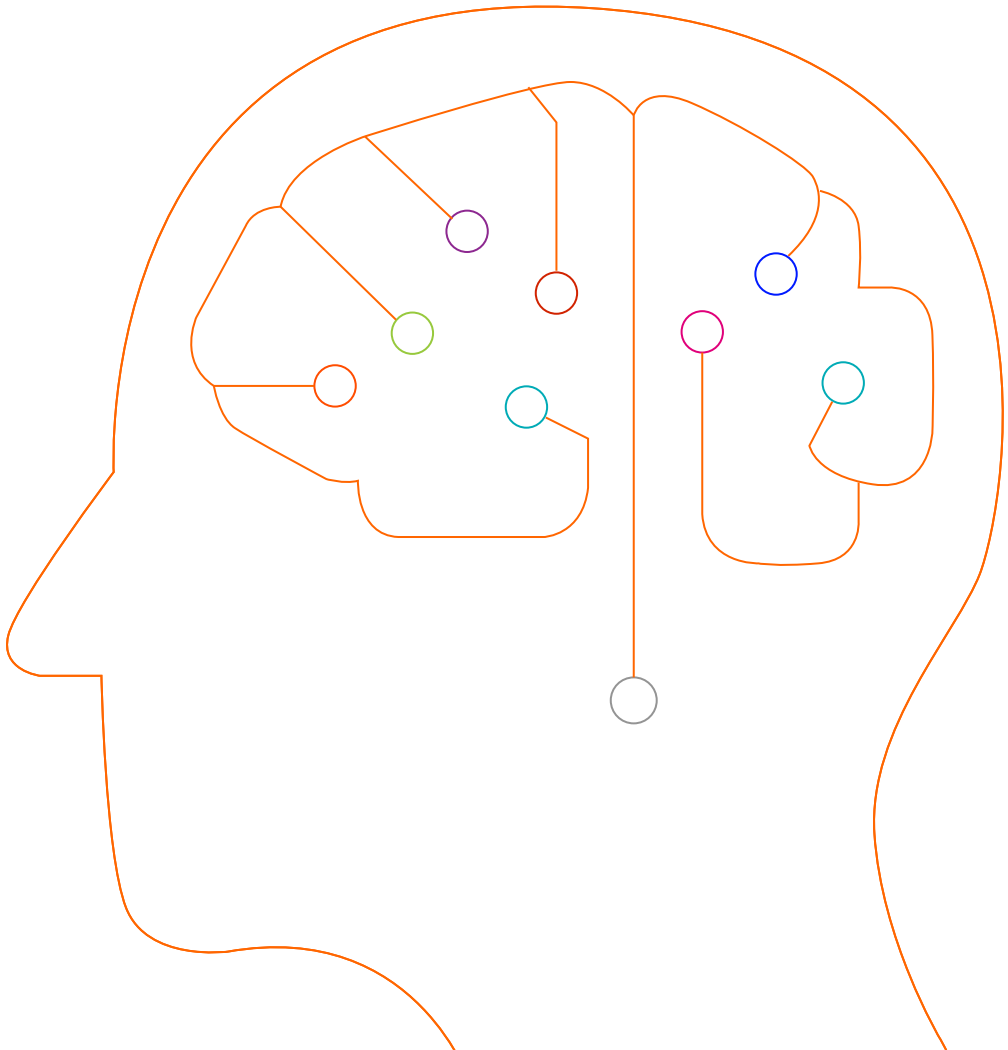
The solution was to use intelligent data discovery to locate where protected health information (PHI) and personally identifiable information (PII) exists and how it moves across the enterprise. Once the data is scanned, cataloged, and curated, the provider then links business metadata with technical and operational metadata. All the information provided through intelligent data discovery informs the architecture and implementation of security zones, entitlements, user roles, access management, and data asset handling.

READ MORE

Conclusion: Your First Step to Harnessing the Power of Your Data

Every digital transformation journey begins with data. Intelligent, automated data discovery is the foundational first step in extracting value from your data.

By leveraging the power of metadata-driven artificial intelligence, an AI-powered data discovery solution enables you to quickly find and understand all of your critical data assets and catalog them. As a result, you can see how this data impacts your digital transformation initiatives, understand the business context surrounding the data, and enable data democratization of trusted data at scale.



Further Reading

Informatica Enterprise Data Catalog Data Sheet

AI-powered data discovery capabilities are included as part of the Informatica Enterprise Data Catalog solution. Learn how the Enterprise Data Catalog enables you to find data assets across the enterprise and unleash the power of your data.

READ MORE

Drive Your Business Forward with a Catalog of Catalogs

Many tools and solutions come with their own data catalog to inventory the data they store. But specialty data catalogs can only give you visibility into data from one source. Most organizations rely on an ever-growing number of data sources. Informatica's AI-powered enterprise-class catalog of catalogs creates a centralized, standardized, end-to-end view of all your enterprise data.

READ MORE

About Informatica

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world's leader in enterprise cloud data management, we're prepared to help you intelligently lead—in any sector, category or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

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

IN19-0920-3956

Informatica®