Informatica

eBook

How to Build Trust in AI: The Data Leaders' Playbook

Ensure Reliable AI Performance by Prioritizing Data Quality

Where data & AI come to LIFE™

Contents

Drive Confident, AI-Enabled Decision-Making with Trusted Data

To fulfill the promise of generative AI and maximize business outcomes, chief data officers (CDOs) and chief data and analytics officers (CDAOs) must prioritize data quality. Only by using trusted, consistent, reliable and timely data can data leaders ensure reliable AI performance as they scale their AI initiatives.

Organizations increasingly recognize the value of generative AI. In a recent survey, nearly half (45%) of data leaders say their companies have already integrated generative AI into their business processes, and 54% expect to implement this technology[1] in the year ahead.

Yet data quality is a high-priority concern. Among data leaders implementing or planning to implement generative AI and large language models (LLMs), 42% say that the quality of data is their top data-related obstacle.[2] As a result, the need to improve data quality is a growing priority for data leaders who want to successfully adopt AI.

"Through 2025, at least 30% of GenAI projects will be abandoned after proof of concept due to poor data quality, inadequate risk controls, escalating costs or unclear business value."

Arun Chandrasekaran,
Distinguished VP Analyst at Gartner[3]

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

[2] Ibid.

[3] Source: Gartner® Article, Highlights From Gartner Data and Analytics Summit" March 13, 2024. Trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

Drive Confident, AI-Enabled Decision-Making with Trusted Data (continued)

This eBook strives to provide data leaders like you with a comprehensive understanding of the urgent need to deliver high-quality data to your business. It also reviews key strategies you can use to develop trusted data that helps support the success of your generative AI initiatives.

"Data quality has always been an important issue for CDOs. But the scale and scope of data that generative AI models rely on has made the 'garbage in/garbage out' truism much more consequential and expensive, as training a single LLM can cost millions of dollars."

McKinsey, "The data dividend: Fueling generative AI," 2023[4]

[4] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-data-dividend-fueling-generative-ai

Understand the Role of Data Quality in AI

Your need to develop a data strategy that helps you govern, organize, analyze and deploy your organization's data – effectively and efficiently – is well-understood. But to compete in today's world, your data strategy also must advance your company's strategic business goals.

It must supercharge business performance across the organization and create competitive advantage. Your data strategy also should help you build trust and a shared vision among key stakeholders and deliver practical, incremental and demonstrable value back to the business.

That's a tall order, one that requires targeted investment. In a recent publication,[5] Srikant Kanthadai, senior director of data management at Deloitte, argues that data leaders need to apply finance-like rigor to data management and custody.

"Treat data in a repository like money in the bank," he advises. "Keep it locked up and secure. Know who all the counterparties are in any exchange. Be able to say with confidence that everyone with access to a vault is following company and regulatory rules to the letter and, very importantly, the data meets the required quality and governance standards."

However, there's more to the story. When you manage data this way, you can begin unleashing more value from it – by combining responsible use with innovative new technology.

The Connection Between Generative AI and Data Quality

Modernizing data management and ensuring trust in the organization's internal and external data is critical to your success with generative AI. According to Ventana Research, "AI models must be based on accurate data to produce accurate results. And with many AI models being used in automated processes it is important invest in ensuring the trust and quality levels in the data that feeds AI."[6]

[5] https://www.informatica.com/content/dam/informatica-com/en/collateral/white-paper/how-cdos-can-enable-the-intelligent-enterprise_white-paper_4635en.pdf

[6] https://www.ventanaresearch.com/white_paper/big_data/data_and_ai_governance_for_trusted_insights

Understand the Role of Data Quality in AI (continued)

High-data quality correlates directly with the performance of today's AI models. To produce worthwhile results, your AI models must be based on accurate data. In contrast, using low-quality data that includes errors, inconsistencies or inaccuracies can result in skewed results and unreliable model predictions. In short, issues with data quality can lead to flawed business decisions, such as:

- AI using outdated sales data could predict incorrect stock needs, causing financial losses.
- An AI model providing financial advice may give misleading strategies if it uses inconsistent data, risking significant user losses.
- An AI system calculating credit scores could deny deserving candidates credit if it relies on inaccurate financial data.
- Chatbots trained on wrong or biased data could offer improper customer assistance, damaging a company's reputation. A recent story linked a company's error to a chatbot using unsuitable language with a customer.[7]

"As AI becomes part and parcel of decision-making of core aspects of life, the sanctity and quality of data powering these models takes on high importance."

Google Research, "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI," 2021[8]

[7] https://www.bbc.com/news/technology-68025677

[8] https://storage.googleapis.com/pub-tools-public-publication-data/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf

Prioritize Data Quality

Data quality refers to the overall utility of a dataset and its ability to be easily processed and analyzed for other uses. High-quality data is both fit for its intended purpose and clean enough to generate the desired results.

Following are several data dimensions that you should consider when assessing data quality.

- **Uniqueness,** where no entity exists more than once in the dataset
- **Accuracy,** the degree to which data correctly represents the real-life objects it is intended to model
- **Consistency,** where data values in one data set are consistent with values in another data set
- **Completeness,** where certain attributes are assigned values in a data set and those values are met
- **Timeliness,** where the expectation for accessibility and availability of information is met when expected
- **Relevancy,** which considers whether the data is pertinent to the problem AI is trying to solve
- **Currency,** the degree to which information is current with the world that it models
- **Conformance,** whether data is stored, exchanged, or presented in a format consistent with expected values

How Automation Helps You Maintain Data Quality

Implementing and monitoring data quality requires constant vigilance. You need to adopt a consistent approach to deal with all data, especially as you start using multiple data sources or types.

Augmented data management solutions often use AI to maintain data quality by standardizing your data and ensuring that it is complete and consistent. Using machine learning algorithms, they can identify inconsistent or erroneous data, correct anomalies, and monitor changes in data quality over time.

These solutions automatically check for incomplete or invalid entries, resolve conflicts, and add missing information from third-party sources. They can rapidly detect and rectify inconsistencies in records, remediating data accuracy, completeness and duplication. They also can ensure your data remains standardized, clean and accurate – without incurring the delays created by manual data cleansing and data quality rule creation.

For example, AI models in the retail industry use automated data management solutions to provide accurate predictions about sales trends, customer behavior, and inventory-level management. High-quality data allows retail sites to use AI to recommend products based on customer behavior, purchase history, and trends. Advanced data quality solutions help ensure that the recommendations are accurate, supporting enhanced customer engagement, cross-selling, and revenue gains.

Enhance Data Observability

To maintain transparency and traceability, you must monitor and observe the complete data lifecycle – as data flows from collection to processing to usage. Enhanced data observability helps you understand real-time aspects of the data pipeline, including how data is consumed, protected and kept compliant with policies and regulations. You can see the health of your data, identify the impact of issues and take prompt preventative or remedial action.

Data can be observed through the following lenses:

- **The data itself,** with checks for anomalies or anything out of the norm
- **Data pipelines,** where you check for consistency in data volumes and size of records
- **Data consumption,** with an eye toward privacy issues and authorized data access

How Data Observability Affects AI Performance

Data observability helps you automate the identification and resolution of data problems that could impact downstream tasks such as data analysis and machine learning model training. A strong level of data observability helps organizations rely on and effectively use their data.

The accuracy and reliability of AI models requires high quality, consistency, and dependability of the data they are trained on. By observing the data, you can identify issues such as missing data, data corruption or changes in data distribution that can negatively impact the accuracy of the AI model. This process is also known as anomaly detection. By observing data pipelines and how data is consumed, data observability helps ensure the optimal performance of the model.

For example, e-commerce and telecommunications companies are using machine learning models to predict customer churn and personalize customer experiences. With the help of data observability, they can detect and fix data-related problems in advance, leading to improved accuracy in predictions and better customer experiences. Healthcare organizations use AI to predict patient risks and identify optimal treatments. Maintaining data observability helps ensure the reliability and timeliness of medical data.

Bringing together AI and data observability is an important step forward. Using AI-powered data observability tools can help organizations monitor and maintain data health at scale, automatically detecting anomalies and reducing the time-to-insight for data issues, effectively boosting AI performance.

Ensure Data Privacy

As AI systems process huge volumes of data, including sensitive data, ensuring data privacy and security becomes a critical challenge. You must balance the benefits of AI with the need to protect individual privacy rights. With regulations such as the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA) and the European Union's Artificial Intelligence Act, and increasing focus on environmental, social and governance (ESG) reporting, this challenge becomes more complex.

Many companies are using personal data to train AI models. To protect data privacy, teams must be able to leverage AI-powered insights to execute and enforce policies around data collection, quality, protection, access, use and retention. Without the right governance and stewardship, you could increase business risk exposure and limit opportunities to promote business value.

It's important to understand whether your company uses personal data in a way that complies with data privacy laws and regulations. For example, is your company properly anonymizing data and providing access to data that users are authorized to use for their intended purposes?

What Are the Best Practices to Ensure Data Privacy in AI?

As your AI initiatives develop, you must evaluate your data privacy practices:

- **Data collection and storage processes,** which should adhere to data protection regulations and carefully control who has access to data, how it's collected, where it's stored, and how securely it's stored.
- **Risk remediation,** which should offer anonymization and pseudonymization techniques that protect personally identifiable information (PII) and intellectual property (IP) yet preserve the quality and utility of data, along with technologies that prevent or limit re-identification to specific authorized conditions.
- **Consent and transparency,** which must fully inform data subjects and allow them to explicitly consent to the processing of their data by AI systems.
- **Data purpose limits,** which ensure that collected data is used only for its intended purpose, aligned to policies, and respect the rights of data subjects, preventing misuse.
- **Data retention policies,** which prevent data from being kept longer than necessary with data retention expiration and disposal policies in place.

Emphasize Efficient Data Management

Integrating and managing data from disparate sources is an ongoing challenge. You must reconcile data that uses various formats and standards into a unified view and effectively manage that data to ensure consistent analytics and reporting.

An automated, AI-powered data management platform can help you scale your generative AI efforts and deliver greater value to the business by supporting efficient data management. A unified, cooperative data platform facilitates shared learning from the experiences of different departments and business units. It offers an environment for collective wisdom, catalyzing innovative applications that propel the organization forward.

Data integration capabilities help you unify disparate data sources and heterogeneous applications at the speed your business demands – including batch, real-time, streaming or serverless. They should help you automate master data integration, using AI to significantly reduce the time required to onboard new data sources so you can adapt quickly to changing data landscapes and business needs.

Automation enabled by your data management platform is indispensable to your AI success. By bringing unmatched efficiency to data management and freeing teams from the intricate, time-consuming tasks of data stewardship, automation allows users to focus on turning insights into innovation.

Emphasize Efficient Data Management (continued)

How Can You Ensure Trusted Data Throughout Your Organization?

Data governance provides the foundation for the trusted data and models needed for critical AI initiatives. However, with the rapid growth of predictive AI and generative AI, taking a traditional approach to data governance is no longer enough.

A modern data and AI governance approach requires a focus on three pillars:

- **Risk and compliance,** helping you protect data privacy and reduce the potential for data misuse across complex data landscapes while allowing convenient access to reliable data.
- **Data sharing and democratization,** making sure data assets are easily discoverable, trustworthy and accessible to virtually all levels of data consumers and providing context that helps data consumers use AI solutions confidently.
- **Intelligent data observability,** allowing you to monitor data usage for data quality issues, improve data flow and help ensure data security, often by using automation and AI to detect data quality problems and anomalies at an early stage.

Using automation and AI can simplify the process of managing data in a responsible, ethical manner. Key technologies that can support modern data governance include metadata scanning and extraction, semantic search, data classification and curation automation, automated data quality and data observability, automated discovery of inferred relationships, and policy management for sensitive data and data protection. With these capabilities, you can manage, share and use data more effectively, resulting in better decision-making, increased operational efficiency and improved policy compliance.

How Informatica Helps Data Leaders Ensure Data Quality

Informatica offers automated, AI-powered data management solutions that ensure the data quality and data observability needed to responsibly enable your AI initiatives.

The **Informatica Intelligent Data Management Cloud™ (IDMC)** offers a comprehensive set of services to help establish proactive data quality and data observability. The platform provides a unified solution encompassing data profiling, data quality, data cataloging and lineage, data governance, data democratization and sharing, data protection and pipeline optimization. This common platform facilitates a flexible yet unified approach to data observability with tailored insights for data engineers, IT operations, chief data and analytics officers and data analysts.

The IDMC Cloud Data Quality service leverages **AI-powered CLAIRE®** to identify issues and anomalies in the data, and checks whether data meets predefined quality standards on an ongoing basis. Once issues and anomalies are identified, key stakeholders can be alerted. These alerts serve as early warnings, prompting timely remedial actions. The insights derived from continuous monitoring and remediation efforts contribute to enhancing data systems, improving data quality and making the organization more resilient against data-related challenges.

IDMC supports data governance and privacy needed to create trusted AI and analytics. The platform offers visibility into data sources and AI models, supporting explainable and responsible AI. Users can view profile statistics and monitor scorecards from a single pane to improve data quality and insights. A governed marketplace makes data available to business users with policy-based access for safe use. AI-powered intelligence and automation from CLAIRE uses active metadata to simplify data governance processes, increase efficiency and deliver trusted data faster.

[9] Gartner, Magic Quadrant for Augmented Data Quality Solutions, Melody Chien, et al., 6 March 2024.

Informatica Named a Leader in the Gartner® Magic Quadrant™ Report

For the 16th time, **Informatica has been named a Leader** in the Gartner Magic Quadrant for Augmented Data Quality Solutions. In its 2024 report, Gartner recognized Informatica for both "Ability to Execute" and "Completeness of Vision." With augmented data quality capabilities driven by AI, machine learning, convergence and integrations, our data management solutions enable game-changing automation.[9]

Conclusion

To scale your AI initiatives, you need a modern approach to data management that prioritizes data quality. Using automation can help you simplify the process of managing data effectively and realize maximum value from your AI projects.

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

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

IN19-4747-0524