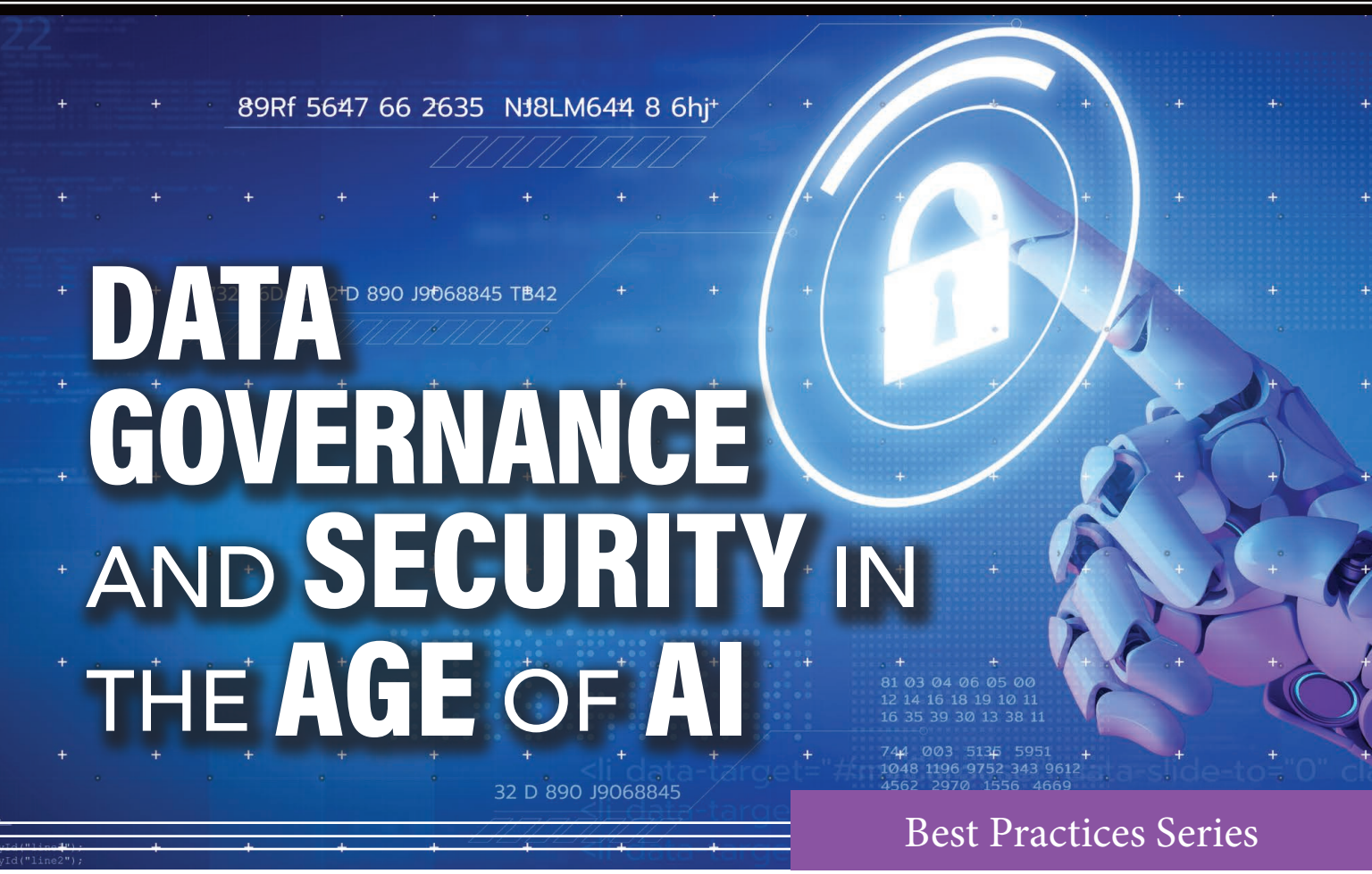
database
TRENDS AND APPLICATIONS

DATA GOVERNANCE AND SECURITY IN THE AGE OF AI

database
TRENDS AND APPLICATIONS

Best Practices Series



DATA GOVERNANCE AND SECURITY IN THE AGE OF AI

Best Practices Series

The more we advance into the age of AI, the more challenging it becomes to data managers and others charged with delivering solid data streams and access to their businesses. Some of the issues that have vexed data operations for decades are getting amplified as decision makers seek real-time insights, and greater hands-off automation drives critical business processes.

Blame the relentless pace of AI advancements, for starters. "Another day, another new AI large language model that's supposedly better than all previous ones," the *Atlantic*'s James Surowiecki recently wrote. "When I began writing this story, Elon Musk's xAI had just released Grok 3, which the company says performs better than its competitors against a wide range of benchmarks. As I was revising the article, Anthropic released Claude 3.7 Sonnet, which it says outperforms Grok 3. And by the time you read this, who knows? Maybe an entirely new LLM will have appeared."

This means data managers and their colleagues in data security need to keep a step ahead of this fast-changing word, knowing that AI—both in its generative and machine learning/statistical forms—is essentially dead in the water without well-vetted, timely, secured, quality data.

AI, in all its forms, is a voracious consumer of data, requiring new levels of scalability and availability.

Data managers and their colleagues need to not only focus time, attention, and budget on the large-scale foundational AI models but also on the smaller, internal models. "It is much more likely that your corporate data will be used to train one of these functional models," according to an analysis out of Forrester Research.

Data readiness can either make or break AI initiatives, a survey of 565 data and analytics professionals, conducted through Drexel University, finds. Sixty-two percent of organizations lack the effective data governance that is needed to drive trustworthy AI into their organizations, the survey shows. The challenge is to manage data usage in their environments—including such factors as where it's stored, its lineage, who has access to it, and whether it includes personally identifiable information.

At least 42% of organizations in the Drexel survey identified data governance as a top challenge to data integ-

As AI has become a critical competitive tool, both data governance and security have emerged as critical areas for modernization.

rity, second only to data quality, and up from 37% in the previous year's survey. Organizations that have invested in data governance programs report benefiting from improved data quality (58%), improved quality of data analytics and insights (58%), increased collaboration (57%), increased regulatory compliance (50%), and faster access to relevant data (36%), the Drexel survey shows.

What are the essential elements of a robust data governance and security effort that can manage the AI boom? The following are issues that data managers and their colleagues need to ask about and address as they assess the degree of risk associated with generative and machine learning/statistical AI:

Does the AI model's training reflect the current real world? "AI models learn from experience, based on the training data they are provided with initially, and (if permitted) the examples they encounter in use," according to a Google analysis of AI governance approaches. "Problems can arise if the training data is incomplete and misses some key aspects, or even if relevant aspects of the world have changed since the training data was collected. Part of due diligence to ensure the safety of an AI model is thus to pay close attention to the provenance and quality of the training data set, and adjust to mitigate against any shortfalls."

What data catalysts and dependencies are associated with AI? Generative AI's dependencies include "1) massive amounts of data to train the models, 2) skilled practitioners to build the technology, and 3) substantial compute resources," the Forrester analysts report. "For most organizations, the adoption timeline for generative AI went from 'soon' to 'yesterday' due to… vendors announcing plans to bundle these capabilities into widely adopted and deployed enterprise tools."

What are the implications of third-party data management? "When you buy a product or service that includes generative AI, you depend on your suppliers to secure the solution," the Forrester analysts point out. "Microsoft and Google are taking that responsibility as they bundle and integrate generative AI into services like Copilot and Workspace, but other providers will source AI solutions from their own supplier ecosystem."

What safeguards prevent malicious or unauthorized use of AI models? This is looking for "appropriately documented controls for the most obvious misuse scenarios," according to Forrester.

Can the risk of "data poisoning" be mitigated? "AI systems that are continuously learning—rather than learning in lab conditions and then having the underlying model frozen before real-world use—are likely to be at greatest risk of having the data they learn from corrupted," the Google report stated. "As a general rule, developers should think carefully about the data poisoning risks associated with having their AI systems learn in real-time in a real-world environment."

What limits are in place for session lengths/rate? The purpose of raising this question is to ensure that "proper safeguards that prevent tools from behaving in unexpected or unintended ways" are in place.

What monitoring, detection alerting, and response capabilities protect against the risk of users making queries that are considered unsafe, unethical, or dangerous? "Controls are necessary but not sufficient," stated the Forrester analysts. "A solution must also have a proves from the provider—and your own internal teams—to identify, investigate, respond to, and remediate any issues discovered."

What controls prevent data leakage, and what monitoring exists for data leakage? This depends, the Forrester analysts point out, on "how the provider protects sensitive corporate data from being shared with unauthorized entities, and how are you notified in the event an employee attempt to share sensitive information?"

How much human intervention is required in data governance and security? "From the beginning stages of problem and goal articulation, through to data collection and curation, and model and product design, people are the engine for the system's creation," according to the Google analysis. "Even with advanced AI systems able to design learning architectures or generate new ideas, the choice of which to pursue should still be overseen by human collaborators, not least to ensure choices fall within an organization's legal and financial constraints. Similarly, people play a vital role in the upfront verification and monitoring of a system, such as choosing which tests to run, reviewing results, and deciding if the model satisfies the performance criteria so as to enter (or remain) in real-world use."

As AI has become a critical competitive tool, both data governance and security have emerged as critical areas for modernization. AI, in all its forms, is a voracious consumer of data, requiring new levels of scalability and availability. An effective governance and security initiative isn't a luxury; it's a necessity for today's enterprises.■

—Joe McKendrick

Governing Data for Cloud & AI Environments

IRI Voracity
An Insatiable Appetite for Data

As stewards for the newer sources of data in the cloud and feeding AI models, today's GRC teams and developers must grapple with increasingly complex requirements for data security, compliance, and quality.

THE NEWER LANDSCAPE

Recent studies by Database Trends and Applications (DBTA) confirm that IT departments are dedicating more time and effort to data governance and security than ever before. Let's delve into the factors driving this trend:

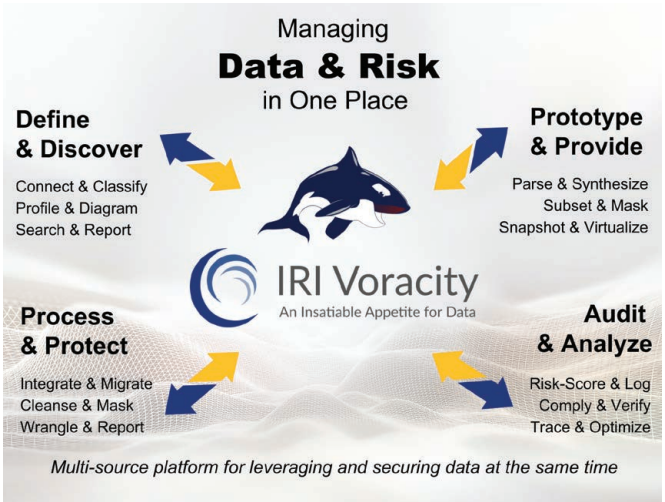
1. **Data Proliferation:** Organizations are facing an explosion of data sources and formats. From structured databases to unstructured logs, the sheer volume of information demands effective governance.
2. **Distribution Challenges:** Data transcends traditional boundaries. It flows rapidly between cloud, on-premise, edge and remote platforms. Ensuring consistent governance across this diverse landscape can be hard.
3. **Insatiable Demand:** Business users crave fast, easy access to data insights. Whether for reporting, analytics, or AI model training, the pressure to deliver timely information is relentless.
4. **Security Imperatives:** Evolving cyber threats pose risks to data integrity and confidentiality. Protecting sensitive information while enabling data access requires a delicate balance, in both production and test environments.
5. **Compliance and Quality:** Stricter regulations demand adherence to data privacy rules. Simultaneously, maintaining data quality standards is essential for accurate decision-making.

MODERNIZING AND IMPROVING DATA GOVERNANCE

IRI Voracity is a comprehensive data management platform designed to address these challenges head-on.

Voracity supports a range of data governance functions—especially for cleaning bad data and masking sensitive data—for a myriad of data sources on-premise and in the cloud. It uniquely offers:

1. **Unified Capabilities:** Voracity integrates data discovery, integration, migration, governance, and analytics within a single, intuitive interface. Built on the Eclipse™ framework, it streamlines the entire data lifecycle.
2. **Integrated Data Quality:** Standalone or during ETL, masking, wrangling, reporting, or migration jobs, Voracity users can validate, filter, scrub, de-duplicate, enrich, and standardize data in structured tables, files, and streams.
3. **Robust Data Masking:** Voracity excels in data masking and test data synthesis—critical data security activities. By classifying, discovering, and consistently obfuscating or synthesizing data, you can comply with data privacy regulations while enabling data utilization and preserving referential integrity.



REDEFINING TRADITION

Modern data governance and security practices must be scalable, agile, and adaptable to accommodate cloud and AI use cases. Voracity embodies this shift from traditional approaches, enabling organizations to harness the full value of their data while simultaneously governing it.

More specifically, Voracity and its included "IRI Data Protector" suite component products like DarkShield in particular, can search, report on, and mask a broad range of data classes pertaining to multiple data privacy laws across an even broader range of structured, semi-structured, or unstructured data sources.

Voracity also supports the ongoing cloud and AI movements through:

1. **Design Efficiency:** Voracity streamlines data processes, optimizing workflows from discovery to analytics. Its unified data management capabilities and consolidated data manipulation reduce project and design complexity, enabling teams to work faster and smarter.
2. **Runtime Speed:** Powered by CoSort and its built-in resource management features, Voracity saves time through smarter algorithms and job configurations.
3. **Team Collaboration:** A familiar GUI and common metadata foster a shared job design environment, bridging typical gaps between IT and business intelligence teams. As collaboration becomes seamless, governance and insight improve. ■

LEARN MORE ABOUT Voracity and the data-centric governance and security capabilities at https://www.iri.com/products/voracity/technical-details#capabilities.