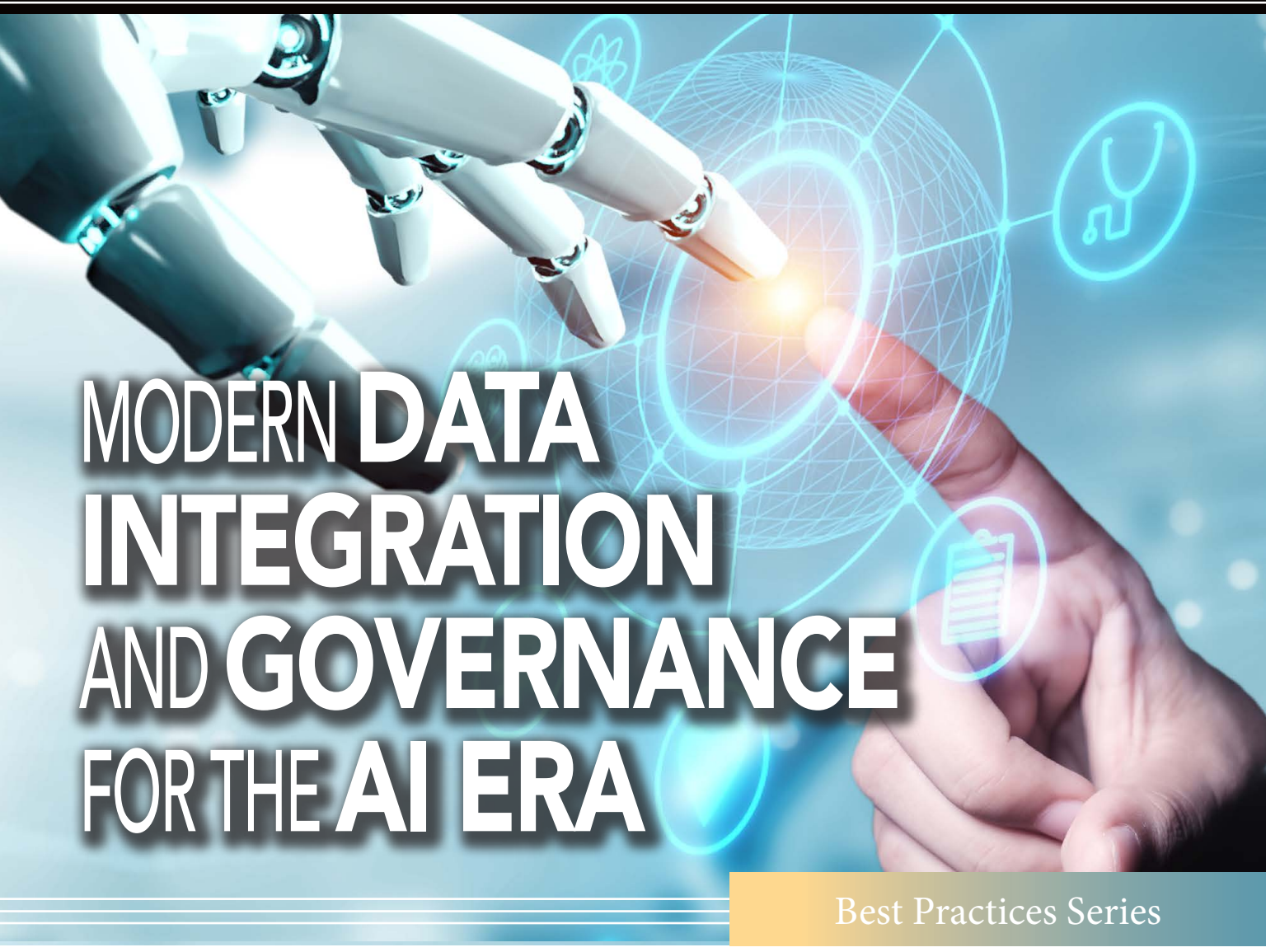
database
TRENDS AND APPLICATIONS

MODERN DATA INTEGRATION AND GOVERNANCE FOR THE AI ERA

Best Practices Series

database
TRENDS AND APPLICATIONS



MODERN DATA INTEGRATION AND GOVERNANCE FOR THE AI ERA

Best Practices Series

The meteoric rise of AI—with its insatiable appetite for timely, quality data—is exposing the inefficiencies and underfunding that have long dogged enterprise data management.

Data managers have been able to improvise, adapt, and overcome many deficiencies when it comes to keeping the data pipelines flowing. However, with businesses leaning heavily on their data, people, and assets to support AI at many levels, there's much more at stake than before.

First, without the right data collected, vetted, cleansed, and transformed, AI is just an empty and potentially very expensive vessel, delivering no value to the enterprise.

Second, with all the excitement and hype around AI, there's been a tendency for people across all organizations to jump into it without the full engagement of the enterprise, while ignoring alignment with its goals and values. As a result, there are documented cases of AI delivering erroneous and even harmful decisions that damage businesses financially and harm brand reputation. Witness the recent instance of an airline chatbot providing steep discounts to passengers—which the airline later attempted, unsuccessfully, to withdraw.

Data is the refined material powering the decisions and directions coming out of AI systems. Enterprises are hungry for all the data available to boost their decision making and automated applications, hopefully as close to real time as possible. The challenge is that while there may be plentiful data out there, much of it is either siloed, out of reach, or even unknown to business leaders and data managers. Whether the data is compliant with various laws and mandates is another open question.

Quality and consistency—vital to a well-functioning AI system—are growing concerns. If anything, confidence among data managers in data quality is going the wrong way, according to a survey published in October 2023 by Unisphere Research, a division of Information Today, Inc. Only 23% express full confidence in their organization's data—down 7 percentage points from a similar survey conducted 2 years ago.

That's why a comprehensive effort to step up data integration and governance is so critical in today's economy. (Note: Throughout this article, we combine integration and governance into a unified category, as both have the same objectives.) Successful programs

have a direct impact on the widescale performance of analytics and AI. The goal is to bring together data from multiple sources both within and outside the organization, arriving in various formats. This data is then made available to the business in a secure, well-vetted fashion.

There are a range of technologies, tools, and platforms to help deliver trustworthy, AI-ready data. Data management and analytic software has grown increasingly sophisticated, providing the ability to make predictions, take action, and present insights via graphic interfaces. With many tools and platforms, creating and managing metadata—data about data—are also vital pieces of the integration and governance equation.

INTEGRATION AND GOVERNANCE NOW TOP PRIORITIES

Data governance is one of three top priorities for the year ahead, another Unisphere Research study showed. With AI pressing in on all sides, data managers are ramping up their commitment to modernize their data architecture and data management environments, the recent survey, conducted by Unisphere Research in conjunction with John O'Brien, principal advisor to Radiant Advisors, finds.

Forty-six percent of data managers view data quality as their top priority for the year ahead, followed closely by data governance (44%), upgrading their data management platform (43%), and moving forward with their data integration platform (40%). The strong focus on data quality and governance underscores their critical role in ensuring data integrity and regulatory compliance, reflecting ongoing concerns around data accuracy and usage standards.

The survey also shows data fabric gaining more acceptance as a novel approach, with data virtualization and metadata activation recognized as key components. Data catalogs and data observability alike continue to garner more attention. The demand for real-time and cloud-based data integration tools is growing steadily. And the ongoing adoption of DataOps is helping organizations build more flexible, agile processes.

There are still organizations that have yet to embrace data governance. This is especially a vexing problem with the complexities that come with today's enterprises, with multiple databases of varying types supporting varying functions. Many organizations may not have a full picture of the data that is passing through their ranks and systems. What is needed is an all-encompassing approach to data governance and integration that meets the needs of organizations going into the later 2020s and beyond.

INTEGRATION AND GOVERNANCE BUILDING BLOCKS

There are many components that go into a data integration and governance initiative, starting with relational databases, data warehouses, data lakes, and lakehouses and the analytical layers that sit on top of them. Along with widely available data management technologies, a new generation of tools and platforms is emerging, providing new ways for organizations to mitigate these long-standing obstacles.

Importantly, data integration and governance initiatives need to be built on human involvement and engagement, with an emphasis on developing processes, policies, and guidelines to ensure that the right data is being used for the right purposes in a secure and reliable way.

A data integration and governance initiative needs to put the business's goals and values front and center, incorporating rules and identifying key performance indicators linked to a business's data-driven initiatives. At the database level, this means putting security protocols and access controls in place, as well as ensuring availability and meeting service level agreements. Again—and this can't be emphasized enough—the entire organization needs to take part in developing, enforcing, and continuously improving data governance.

NOT JUST TECH

It's difficult to put an effective data integration and governance initiative in place, however. The technology may be mismatched or unable to scale to the requirements of such efforts. Even more of a challenge is getting everyone on board and on the same page with it. Strengthening and improving such initiatives may mean changing processes or putting up additional guardrails that employees may resist.

Often, there's too much reliance on technology to manage data integration and governance, versus enabling human oversight. AI—which is supposed to be enhanced and improved through such efforts—may, rightly or wrongly, also be seen as a panacea for accelerating such efforts.

A successful enterprise-scale data integration and governance program is not just a technology project. It requires cultural adaptation as well. Organizations rely on data to achieve customer satisfaction, as well as to provide insights to guide decision making. Data collection and stewardship concern everyone in one way or another, and effective integration and governance will help establish what is—and what is not—possible with these assets.

People at all levels within the organization—not just data managers, scientists, and C-suite executives—need to be engaged with the processes to surround data with the right tools and technologies, assure its security, and guide the business processes that it enables. This may be expressed through a cross-enterprise integration and governance committee or board, as well as through individuals tasked with its oversight and delivery, such as a chief data officer. There needs to be clarity and well-defined roles overseeing data management as well.

Any effort needs to be designed to assure the highest possible levels and procedures for security and compliance. It helps to lock in security protocols and best practices, with constant review and improvement by all members of the enterprise. In the AI era, these aren't just luxuries for larger, big-budget organizations: Everyone needs to ensure that their AI output is as updated and accurate as possible. A comprehensive data integration and governance effort will assure success. ■

—Joe McKendrick

Integrating Data with Governance

IRI Voracity
An Insatiable Appetite for Data

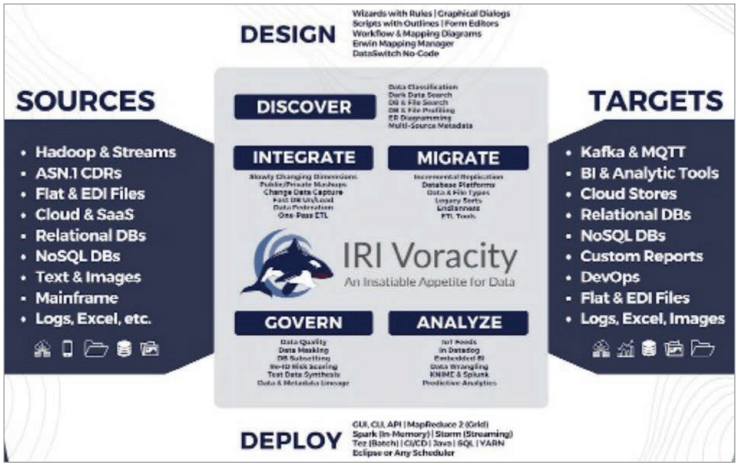
Data integration involves unifying data from different sources to support informational insights and drive innovation.

Data governance involves establishing policies, procedures, and standards to manage data assets effectively. It ensures data accuracy, consistency, and security, which are essential for making informed business decisions.

Performing data governance tasks in combination with data integration offers several benefits:

Enhanced Data Quality: Data cleansing, validation, and enrichment produces more accurate, consistent, and reliable information for decision masking. Filtering out bad or duplicate data *during* wrangling improves analytic results.

Improved Data Security: Data masking and RBAC policies protect sensitive information from unauthorized access. Adding key-managed encryption in the data transformation layer kills two birds with one stone.



IRI Voracity: Data Discovery, ***Integration****, Migration,* ***Governance,*** *and Analytics*

Privacy Law Compliance: In addition to pseudonymization, encryption or anonymization of data feeding analytics, you can score re-ID risk, support DSARs, anonymize PI, and log everything so your data practices align with regulatory requirements.

Streamlined Data Access: Data integration (and master data management) unify views of data for ease of access and analysis. When combined with RBACs, only authorized users will be able to access sensitive data.

WHY USE VORACITY?

Voracity is a proven data management platform that combines data discovery, integration, migration, governance, and analytics in a single, unified framework. Voracity users benefit from its:

Rapid Data Integration: Voracity exploits the power of IRI Fast Extract (FACT) and IRI CoSort to optimize and combine high-volume ETL – plus data cleansing, masking, multi-targeting, and reporting – tasks in a single job and I/O pass, making it highly versatile and efficient.

Robust Data Governance: Built-in data profiling, cleansing, and masking ensure data quality and security across a wide range of data sources. Voracity supports compliance with data privacy laws by encrypting, pseudonymizing, or redacting personally identifiable information (PII), and by facilitating common DSAR requests.

User-Friendly Interface: Built on Eclipse™, Voracity offers an ergonomic design environment for job creation, deployment, and management. Its 'Workbench' GUI also supports DBA jobs, Git asset management and BIRT.

Scalability and Performance: Voracity can process, protect, present, and prototype big data in a variety of formats and silos through multi-threading, memory and I/O optimization, and even distributed computing if needed.

PULLING IT ALL TOGETHER

To bring data integration and governance together in a seamless way, you can perform these tasks with Voracity:

PII Discovery and Reporting: Launch fit-for-purpose data profiling, ERD, and search wizards to discover and report on the statistical character and location of data in on-premise and cloud silos. These features identify data risk, opportunity and quality issues.

Data Classification and Masking: Classify data based on sensitivity levels and privacy law groups, and apply deterministic masking techniques to protect PII with referential and data integrity in structured, semi-structured, and unstructured sources. Apply re-ID risk scores, real-time incremental masking, RBACs, and audit data as needed.

Governed Data Integration: Combine data quality, masking and reporting functions with your filter/sort/join/aggregate/reformat/lookup/pivot transforms. Streamline data staging, security, and reliable analytics simultaneously.

Data Analytics: Leverage included math and stat functions to gain insights from integrated data. Create reports directly in the jobs above or wrangle data into subset handoffs to tools like Cubeware, Datadog, KNME and Splunk.

Test Data Management: Beyond masking, you can subset or intelligently synthesize data for test files, reports, NoSQL and relational databases, and provision data in multiple ways (e.g., MQs, DB clones, DevOps pipelines).

In conclusion, governing data while you're integrating it is possible and beneficial. IRI Voracity is a comprehensive platform for ensuring data quality, security, and compliance while enabling seamless data access and analysis. ■

IRI
www.iri.com