ANALYST BRIEF
Sponsored by: NetApp

Data governance is fast becoming a foundational element of any company's IT strategy. They must consider a unified data architecture when designing AI workflows, which must be built with highly performant, resilient, flexible, and scalable storage systems that can seamlessly implement complex AI data pipelines.

Policy-Driven Data Governance and Security for Deploying AI Responsibly at Scale

February 2024

Written by: Dave Pearson, Vice President, Worldwide Infrastructure

Introduction

For organizations undergoing a digital transformation, data is the new oil. Using AI, companies are embarking on ways to extract new insights from data to bolster existing revenue streams, add new sources of revenue, improve profitability, increase customer satisfaction, accelerate the development of products and services, and increase employee productivity.

Using predictive and generative AI to extract deep insights from data requires companies to stitch together diverse internal and external data sets, potentially with lots of sensitive and personally identifiable information, most of which cannot be leaked or exposed in the public domain under any circumstances. For most companies, compliance with data sovereignty underpins many digital initiatives, including AI. Thus, deploying AI responsibly is not just a requirement but an existential mandate.

AT A GLANCE

KEY TAKEAWAYS

- Digital infrastructure deployment choices for AI are heavily influenced by cybersecurity and regulatory compliance requirements.
- Improving data governance remains a top priority for businesses embarking on AI.
- IDC encourages companies to take a holistic approach to AI with a unified data architecture.
- Deploying AI responsibly at scale means focusing on flexibility, governance, efficiency, and productivity

Data governance is fast becoming a foundational element of any company's IT strategy. Any digital transformation initiatives — specifically AI initiatives — must be bound by specific policies and procedures with the objective of deploying that initiative responsibly at scale. Such governance standards seek to ensure protection from cybertheft, avert the accidental leakage of sensitive data, and ensure compliance with prevailing laws and regulations. Unfortunately, legacy data architectures — often placed within existing infrastructure silos — are not well suited for this task. AI workflows are complex from the get-go and will only become more complex as adoption increases.

To succeed, companies must consider a unified data architecture when designing AI workflows. This data architecture must be built with highly performant, resilient, flexible, and scalable storage systems, thus enabling organizations to seamlessly implement complex AI data pipelines. Unified, hybrid, and multicloud storage systems provide observability across structured, semi-structured, and unstructured data residing across the cloud and on premises, automate data compliance, and crucially protect against data exposure and risk. A unified data architecture ensures the organization

remains steadfast in meeting corporate governance requirements, wherever they may be in their journey of AI transformation.

Trends — The Current State of Governance in AI

IDC's research finds that intellectual property risks are the most concerning issue for companies worldwide when it comes to AI due to worries about violations of security policies. Potential code manipulations are also a major concern. As companies increase their reliance on AI, they are increasing the diversity of data sets in use, including greater use of unstructured and semi-structured data sets. The data includes:

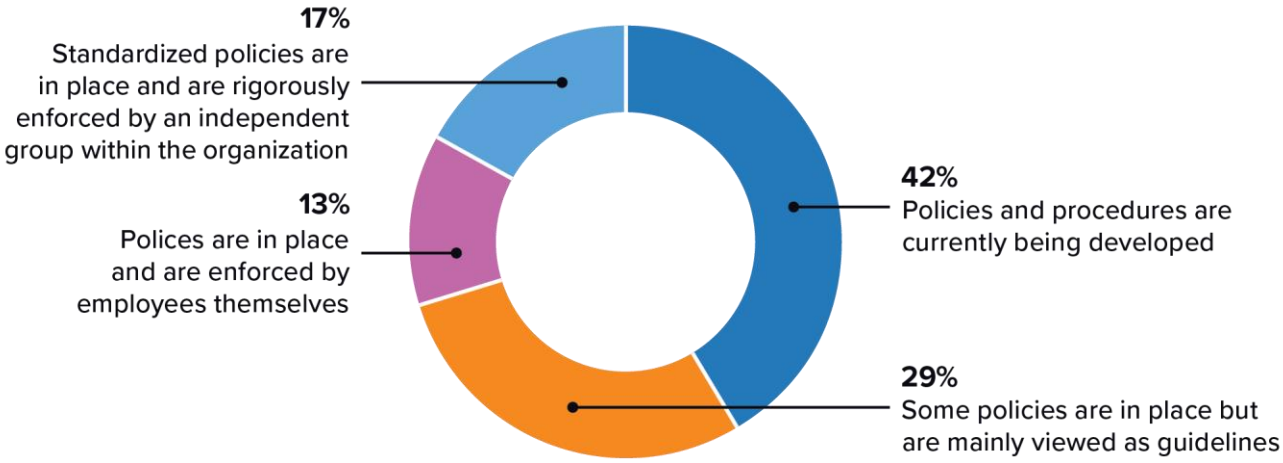
- Transactional data, including purchase orders, invoices, payments, timecards, and sales data
- Master data, including sensitive customer, partner, and employee information
- Social media data, including data that ties the company's business to its own end users and employees
- Enterprise blockchain data, including Hyperledger transactional data
- Streaming data, including stock, event, click streams, and data from IOT sensors
- Business-to-business data, including electronic interactions, catalogs, and other interchange data
- Human-generated data, including spreadsheets, documents, and presentation materials
- Multimedia data, including audio and video files, drawings, and pictures
- Spatial data, including geo-mapping, spatial, and remote-sensing data
- Synthetic and digital twin data

Many of these data types include personally identifiable information; sensitive personal information, such as credit card data and email addresses; and an organization's own intellectual property. The speed at which companies need to gain insights does not leave much room for them to sanitize source data sets prior to analysis. In many cases, personalized insights require them to operate on live data. As a result, improving data security, compliance, and governance must be a high priority.

As a part of a December 2023 study on data requirements for AI, IDC interviewed IT decision makers (ITDMs), practitioners, and developers on their organizations' views on governance and security in the context of AI initiatives. In other words, they were asked to elaborate on what deploying AI responsibly at scale meant to their organization. Figure 1 shows the maturity levels of interviewed companies in the context of governance for AI.

Figure 1

Readiness of Organizations in Deploying AI Responsibly at Scale



Source: IDC, 2024

Do companies have effective AI-specific data governance policies in place today?

A significant number of companies are in the early stages of developing AI-specific data governance policies, indicating a growing awareness of the need for such measures. Some already have AI-specific data governance policies in place, with varying degrees of effectiveness. They continuously review and modify these policies as their specific AI use cases and outcomes evolve to meet the needs of the business.

"AI-specific data governance policies are focused on responsible AI with ethics/bias as drivers" — VP of IT Ops and Data Science, Life Sciences, U.S.A.

In general, IDC finds that there is keen awareness and recognition of the need for continuous improvement and adaptation in AI data governance, particularly in areas such as real-time data monitoring, complex data management, and AI education at all organizational levels.

Finally, despite the existence of some policies, enforcement remains a challenge, with some companies relying on general data governance policies and others starting from basics to build AI-specific policies. "Build now, X later" development methodologies (where X can equal govern, manage, protect, or secure) create risks for organizations. A unified data architecture can aid in the implementation of governance policies by creating visibility across workstreams throughout the entire data life cycle, automating compliance, ensuring observation of the chain of custody of all types of data, and avoiding blind spots within which out-of-system activities can occur.

Who is responsible for developing and implementing data governance policies?

Most companies have dedicated data governance teams, often led by a chief data officer, responsible for setting data governance policies, including those related to AI and generative AI. These teams typically include members from various departments, such as IT, cybersecurity, legal, and AI research, and are expected to evolve over time as AI technologies and their associated risks develop.

The need for more specialized AI experts, including AI ethicists and technologists, is anticipated to address the unique and rapidly changing aspects of AI governance. There is a consensus that the governance structure may need to change in the future, with new roles and professions being created to provide greater insight and value in the face of advancing AI technology.

"Over time, as AI technologies evolve, we expect to see a shift toward the involvement of more specialized AI ethicists and technologists in these teams to address the unique and rapidly changing aspects of AI governance" — VP of IT Ops, Manufacturing, U.S.A.

How are organizational policies evolving over the next 18 months to ensure AI is used responsibly and safely?

"We need policies that allow innovation and experimentation while keeping data safe" — VP of IT Ops, Education, U.K.

Companies are increasingly focusing on the safe and responsible use of AI, with governance committees gaining more influence to manage AI-related risks. Specifically, IDC finds that dedicated data governance teams focus on fostering innovation and experimentation while ensuring data safety, with a particular focus on AI applications that directly impact customer engagement and return on investment.

Over the next 18 months, companies anticipate their policies and procedures will become more comprehensive and robust, with updates to address the rapidly changing AI landscape and a greater emphasis on continuous employee education and training. Finally, companies are leaning toward seeking external consultation for the development of AI-related policies and procedures that comply with industry regulations. They focus on ensuring AI is used safely without exposing personal or company data.

How are companies addressing issues with bias or data sovereignty?

Many companies are still in the early stages of developing policies to address bias and data sovereignty, instead focusing on general academic rigor with initiatives to improve education, knowledge, and skills. For these organizations, there is a need for continuous data review and quality control in data interpretation to address bias, with AI policies clearly articulating how these issues are addressed.

"Bias is harder to address, and it will take time naturally as it relates to the data input" — CIO, IT Ops, Financial Services, U.K.

Organizations are addressing data sovereignty issues through data localization and residency policies, with data being stored in specific regional data centers and subject to audit. There is a strong emphasis on hiring data scientists with good ethical reasoning habits to tackle bias in AI with the use of tools and models that increase transparency and auditability around AI decisions.

.

How are companies addressing issues with commercial data sensitivity, security, and privacy?

"Anticipate, but don't assume! This proactive approach is vital for safeguarding sensitive commercial and operational data in an increasingly digital and interconnected world" — VP of Data Science, Education, U.K.

Companies are focusing on developing stringent data governance policies, with emphasis on data sensitivity, security, and privacy. This includes role-based access control, data classification, encryption, and regular security audits. They are continuously reviewing and updating their policies to align with prevailing regulations and address emerging AI developments. For example, many companies treat GDPR and CCPA compliance on a priority basis.

External audits and internal training are key strategies being employed to ensure data security and privacy. Companies are investing in continuous education and reinforcement to address data security issues upfront. Future strategies include the development of automated controls for data privacy and residency, the use of industry-specific clouds, and the creation of internal AI tools with added layers of processing to address security and privacy concerns.

Benefits — Implementing a Unified Data Architecture

A unified data architecture as a foundation eliminates data silos in any organization and enables the company to take a holistic approach to data security. By implementing a unified data architecture as a preamble to embarking on any AI initiatives, companies can:

- **Protect against data exposure and risk:** Regardless of where and how data resides, IT can remove sensitive or biased information from ingestion before it is used to train or infer from AI models.
- **Automate compliance:** IT can create automatic, auditable copies of AI models for traceability and error detection. It requires an AI-centric approach to copy data management.
- **Protect against cyberthreats in real time:** Using AI and ML-driven anomaly detection, IT can protect organizations' most important, confidential, and commercially critical data sets from bad actors.
- **Rely on validated security frameworks**: IT can implement an AI-ready data stack validated by the Commercial Solutions for Classified Program, led by the U.S. National Security Agency. Other validations include HIPAA, SOC 2, the proposed American Data Privacy and Protection Act in the U.S., GDPR, and the proposed EU AI Act in the European Union.

Considerations — Deploying AI Responsibly at Scale

IDC encourages ITDMs to take a holistic approach to data management when deploying AI. Figure 2 shows the four pillars of any responsible and scalable AI initiative. Companies that want to transform themselves with AI must start with a unified data architecture that offers:

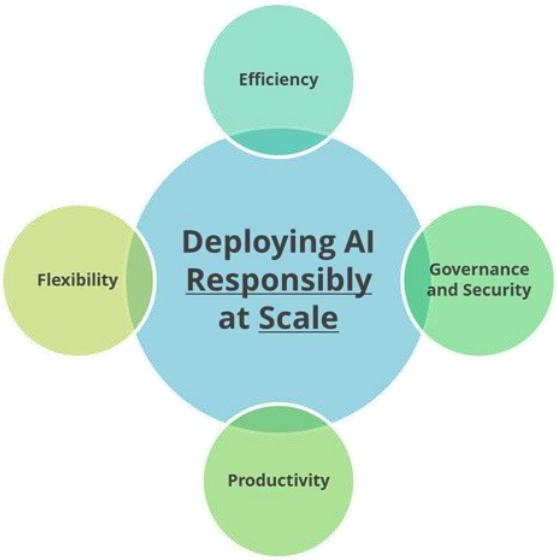
- **Flexibility:** This provides a single-stop access to corporate data for the myriad AI use cases being implemented across the entire organization.
- **Governance:** This enables organization-wide enforcement of data security and governance policies with reliable business outcomes.

- **Efficiency:** This enables the organization to achieve resource efficiency and sustainability objectives from their AI initiatives relative to their overall datacenter footprint.
- **Productivity:** This enables all teams involved in AI workflows — IT, data science, data engineers, and developers — to remain productive, thus ensuring that the desired business outcomes are achieved in a consistent, timely, and accurate fashion.

For companies to scale their AI initiatives responsibly and efficiently, they must ensure that one of the design considerations for the AI project is a fit-for-purpose data storage system that enables the implementation of an integrated data pipeline that can scale securely, efficiently, and sustainably.

Figure 2

Deploying AI Responsibly at Scale



Source: IDC, 2024

Conclusion

Changing the approach to managing data responsibly cannot happen overnight; organizations should expect resistance from groups that value agility and time to value over security and governance. Up until now, most digital transformation initiatives have been insular and inward facing. With AI and generative AI, that situation changes rapidly. An increase in data sources, repositories, and access means threat profiles are changing for organizations. Cyberevents continue to increase in scope, speed, and sophistication. As a result, companies must put governance and security at the forefront of attention and investment requirements when embarking on any AI initiative.

Companies must put governance and security at the forefront of attention and investment requirements when embarking on any AI initiative.

Appendix — Definitions

Data architecture

The collective set of models, policies, rules, and standards that govern the type of data collected, stored, and accessed by an organization. Data architecture enables the organization to collect, integrate, and analyze data from a variety of sources and place it into IT systems.

Unified data architecture

A unified data architecture should incorporate the tenets of unified, hybrid, and multicloud storage into a coordinated, holistic approach to data storage, utilization, management, protection, and governance and ensure the optimal utilization of data resources for all key workloads, no matter the data format, structure, access mechanism, location, or deployment model.

Unified storage

Storage infrastructure, software, and management tools with the ability to store and manage a wide variety of:

- Data formats (i.e., image, video, text, audio) in a single environment
- Data structures (i.e., structured, unstructured, or semi-structured)
- Data-access mechanisms (i.e., block, file, or object)

Hybrid storage

Storage infrastructure, software, and management tools that:

- Enable data to be placed (i.e., stored) across a variety of dissimilar IT systems, deployments, and environments
- Provide data mobility across a variety of shared (public cloud) and dedicated (private cloud or traditional non-cloud) environments
- Enable application-aware data placement

Multicloud storage

A data storage paradigm that:

- Enables more than one shared and/or private cloud environment for placement of data
- Stores and manages dissimilar data in terms of formats, structures, and access mechanisms
- Benefits from a common control plane, management interface, and data protection and security functionality where such tools exist

About the Analyst

Dave Pearson, *Research Vice President, Worldwide Infrastructure*

Dave Pearson is Research Vice President for Storage and Converged Systems practice within IDC's worldwide infrastructure research organization. He also oversees IDC Canada's Infrastructure Solutions research practice. He and his team are responsible for IDC's storage, integrated, hyperconverged, and composable systems and platforms. This includes storage for performance intensive use cases such as high-performance computing, artificial intelligence, and analytics

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.