Leveraging Multicloud Data to Strengthen AI Strategies

Data is the magical elixir behind successful – or ineffectual – AI initiatives. Discover the steps needed to ensure and protect the integrity of critical data across multicloud environments

A RESEARCH REPORT FROM HMG STRATEGY AND NUTANIX

HMG STRATEGY

EXECUTIVE SUMMARY

If artificial intelligence (AI) is the engine driving the 21st-century enterprise, then data is the fuel for AI initiatives. AI relies heavily on real-time and historical data to build predictive models and to make intelligent decisions.

However, when there are issues with the management, integrity, quality, protection and security of data, including the accuracy and consistency of the data used in AI initiatives, this can negatively impact AI and machine learning outcomes. This includes the movement of data that's generated and stored in a multicloud environment, including private and hybrid clouds along with data generated on the edge. Data integrity – which includes the security, control and protection of data – is the foundation for successful and trustworthy AI and machine learning (ML) initiatives.

This is already causing problems with recent AI deployments. According to a recent Harvard Business Review study, almost half (47%) of newly created data records contain at least one critical error.

To help address these issues, the movement and scalability of data and models between cloud environments is critical since AI is compute-intensive for training and fine-tuning. Meanwhile, organizations need the ability to access and process data for AI initiatives quickly and efficiently while collecting data to update models periodically.

"The biggest challenge for customers (business tech executives) is the inablity to connect their private company data to an AI model," said **Luke Congdon**, Senior Director of Product Management at Nutanix. "AI models are based on massive amounts of public information, but usually that public information knows nothing except for what's on the external website about any particular company. So, safely getting their internal and proprietary data connected to a model is not straightforward."

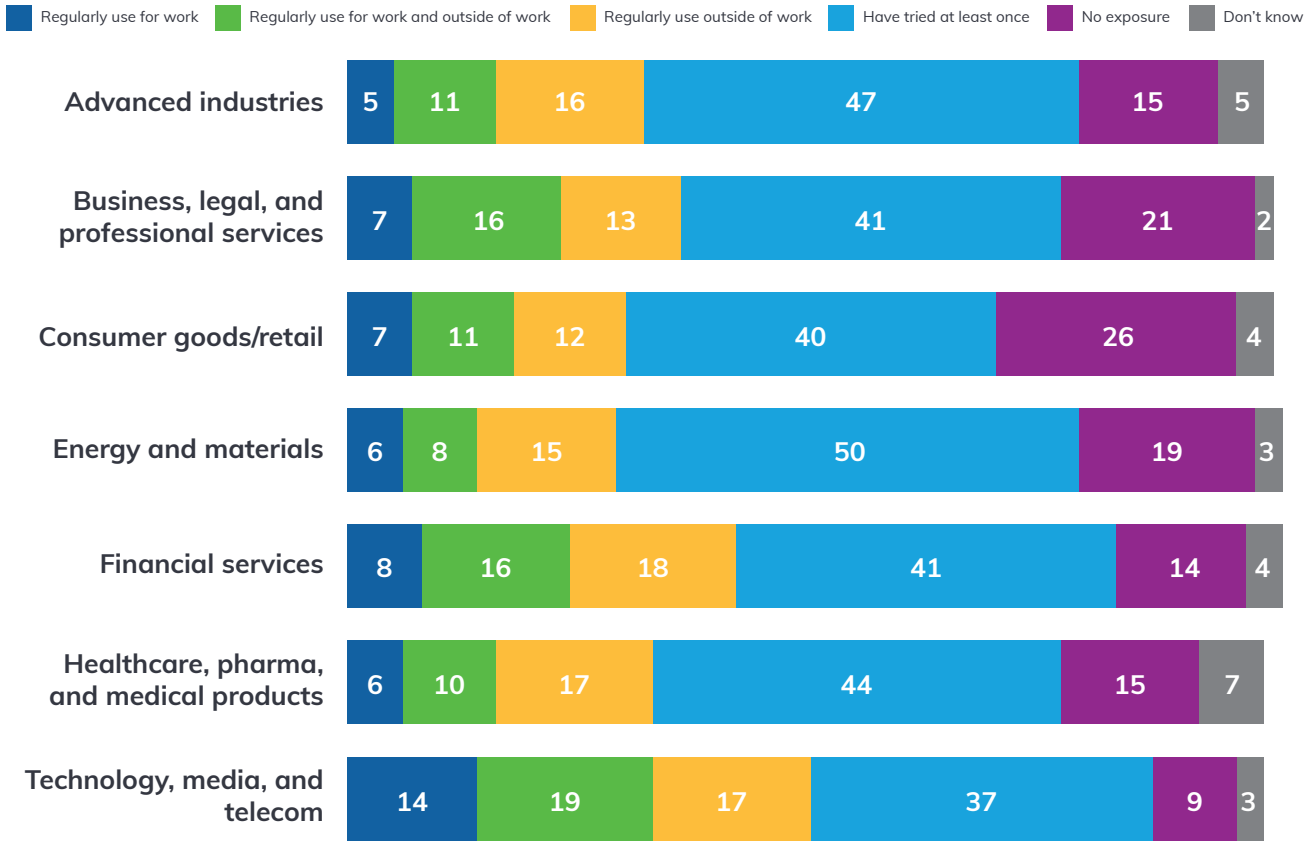
In this HMG Strategy research report, you'll discover:

- How data integrity and quality can impact the success – or downfall – of AI and ML initiatives
- Recommendations for ensuring the quality of data in a multicloud environment
- Advice for protecting data that resides in and is transferred to public, private and hybrid cloud environments to prevent the unprotected utilization of private and proprietary data
- The business and operational benefits of ensuring the integrity of data in hybrid multicloud-fueled AI initiatives

Figure 1: Generative AI Takes Off

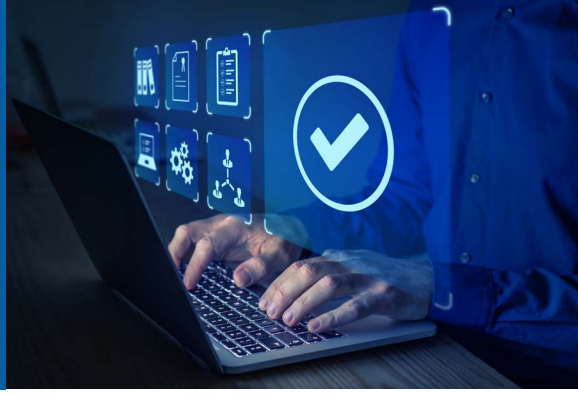
A McKinsey & Company survey of C-suite executives reveals that 79% of respondents have had some exposure to Generative AI – either at work or outside of work while 22% say they're regularly using GenAI at work. Moreover, 40% of respondents say their organizations plan to increase their investments in AI because of advances with GenAI.

Still, the study finds that fewer than half of the respondents are mitigating the risk they consider most relevant for AI initiatives: the inaccuracy of data and AI models.



Source: McKinsey Global Survey on AI; 1,684 respondents, April 2023

Data Quality: The Linchpin to AI Success – or Failure



It's been said that data is the new oil. But just like crude oil, raw data must be refined in order to be properly deployed and to deliver value.

Similarly, the quality of the data used for AI projects and applications – including structured and unstructured data – is also critical. If the data used to train an AI model is inaccurate, incomplete or inconsistent, then each model's results and decisions will also be flawed.

Poor quality data can have a variety of impacts on AI and ML projects. A survey of 600 business and technology executives conducted by Researchscape International found that 38% of respondents reported that their projects took longer due to poor data quality. Another 36% said their projects were more expensive and 33% said their projects didn't achieve the anticipated results due to data quality issues.

Data quality issues can take many forms. For instance, data that has invalid or missing values, duplicate data, or incomplete data.

For instance, inaccurate or incomplete customer data regarding a customer's financial and credit history could prompt a bank's AI model to reject a mortgage application. Not only would this negatively impact the loan applicant, it would also result in a missed business opportunity for the lender.

Another critical challenge that many practitioners face is blending proprietary data with public data in AI models. "Being able to consume your own data but keep it private inside your firewall is where the value exists," said Congdon. "But to leak or lose access to your data externally can be a death blow for a company."

Companies that adhere to data quality standards stand to achieve several benefits. These include improved business efficiencies, faster innovation cycles, more effective R&D and, ultimately, improved products and services.

Steps to Maximize Data Quality

Addressing data quality issues requires the use of technological solutions such as data management and cleaning software. However, achieving good data quality standards also entails organizational and cultural changes, including implementing prescriptive data governance policies and training staff on data handling best practices.

Here are 6 recommended steps that IT teams can take to improve the quality of their private and public cloud data – which, in turn, will benefit the business:

1. **Assess your organization's multicloud data.** Conduct a formal data assessment to determine the types of data that are collected, where it is stored (public, private, hybrid cloud environment), who has access to it, and whether the data is structured or unstructured.

2. **Define what is acceptable data quality.** Each organization should set a standard for what it considers to be acceptable for data quality. For instance, if certain types of data can't be 100% accurate, how close

to perfect is acceptable? For instance, our research found that most practitioners strive for data accuracy rates in the 97%+ range.

3. **Identify and correct data quality issues upfront.** Identifying and fixing multicloud data issues should be a part of any data quality program. To help make this process efficient, this can include the use of automated data cleansing tools.

4. **Establish a defined set of values for multicloud data.** Some data errors occur from employees and other users entering freeform data. For instance, California may be entered as 'CA', 'Calif.' or misspelled as 'Califronia.' Offer users a list of values or drop-down menus to ensure cleaner data.

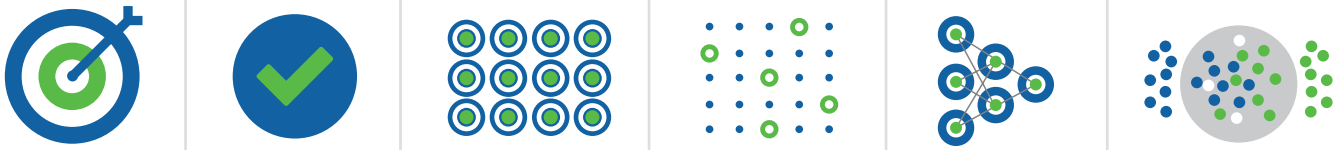
5. **Secure data.** Securing customer and other sensitive data stored in private and public clouds can help protect data against data breaches and other types of attacks. These protections can also ensure that unauthorized users don't tamper with or edit data and compromise its integrity.

6. **Conduct regular data quality reviews.** Conduct periodic reviews of your organization's data in public, private and hybrid cloud environments to determine the effectiveness of data quality initiatives and to determine whether additional steps are needed.

As we've discussed, clean and secure data is the lifeblood of any successful AI project. In this next section, we'll share advice for protecting data that resides in and is transferred to public, private and hybrid cloud environments to prevent the utilization of 'bad' data.

Figure 2: The Six Dimensions of Data Quality

Determining the quality of a data set takes numerous factors into account.



Accuracy
The degree to which the data correctly represents the entity or attribute being described.

Completeness
The percentage of missing data from a given data set.

Consistency
The absence of difference or contradiction in data irrespective of the data's source.

Validity
Invalid data affects the accuracy and completeness of a data set.

Integrity
The validity of relationships across various data entities.

Uniqueness
Ensures duplicate or overlapping data is identified and marked.

"Being able to consume your own data but keep it private inside your firewall is where the value exists. But to leak or lose access to your data externally can be a death blow for a company."

LUKE CONGDON
Senior Director of Product Management
Nutanix

Protecting Cloud and Transferrable Data

The volume and velocity of cyber-attacks are on the rise. Cybersecurity Ventures predicts that the cost of cyber-crime will jump from $3 trillion in 2015 to a projected $10.5 trillion in 2025.

With the lion's share of enterprise data residing in cloud environments (public, private, hybrid), CISOs and security leaders need to ensure that resident and transferable data is being protected effectively.

CISOs and security leaders require a comprehensive approach to safeguarding their multicloud data. This can be achieved by using a single platform that can combine the simplicity of hyperconverged infrastructure (HCI) with native platform hardening, encryption, compliance and controls to block cyber-attacks using a zero-trust approach.

At the network level, microsegmentation is the key to zero trust by providing application-centric visibility and protection.

It's also critical to protect data at rest. This can be done through the use of multi-cloud-enabled encryption tools. Meanwhile, data access should also be limited to specific sets of users so that data is not changed or corrupted. This can be done through the use of identity and access management (IAM) tools.

Moreover, if organizations are using a cloud service such as Microsoft Azure, Google Cloud or AWS, " I would want to be especially sure that the connection points from on-premises to that cloud were secure," says Congdon. "Either I'm using encryption or a virtual private cloud, or I just have guarantees that my data, both at rest and in transit, are secure."

It's also recommended to work with an HCI provider who can protect data on the node from single drive failure and provide native back-up end-to-end.

Meanwhile, enterprise organizations are increasingly vulnerable to data loss and downtime during disasters, as they rely on virtualized and increasingly containerized applications that their legacy infrastructure can no longer adequately support. It's recommended instead to utilize an HCI platform that can provide native data protection and disaster recovery capabilities and disaster recovery orchestration features that are available for on-premises, managed and hosted service providers and with public cloud providers such as AWS.

The Business and Productivity Benefits of Utilizing Quality Multicloud Data

There are multiple business and operational benefits for protecting the integrity and quality of multicloud data used for AI implementations.

For starters, the cleaner the data is, the more accurate the output from the AI model will be. "If the data is good and applied to the right model, you can dramatically save time and augment human knowledge worker efforts," says Congdon.

Organizations that utilize clean data will also typically see better decision-making as the data in use will direct leaders to more accurate outcomes. It can help determine which products need increased production and which products should be dropped.

Meanwhile, customers will be happier and feel more valued if they are receiving appropriate offers based on their interests and behavior. This will also reduce marketing costs by avoiding a scattershot approach to marketing. Higher quality data run through algorithms can also help marketers and brand managers to better understand what customers like and the channels they prefer to use to view products and to receive support.

Plus, utilizing high quality data saves organizations money by reducing the costs associated with fixing bad data while preventing costly errors and disruptions. Good quality data improves the accuracy of AI models, leading to better business decisions that boost sales, streamlined operations and the ability to gain insights about customer or market conditions ahead of rivals that can provide a competitive edge.

Teams that use the right data tools are able to identify new business use cases 90 percent faster than their peers while lowering the total cost of ownership by up to 30%, according to McKinsey & Company.

Says Congdon, "The biggest opportunity is not just to pursue the excitement behind AI initiatives but to identify the business value that we can achieve."

About HMG Strategy

HMG Strategy is the world's leading digital platform for technology executives to reimagine the enterprise and reshape the business world. The HMG Strategy global network consists of more than 500,000 CIOs, CTOs, CISOs, CDOs, senior business technology executives, search industry executives, venture capitalists, industry experts and world-class thought leaders.

HMG Strategy's Global Advisory Services are a unique set of peer-driven research services that are designed to keep business technology executives up to speed on the latest leadership, business, technology and global macro-economic trends that are impacting businesses and industries.

About Nutanix

Nutanix is a global leader in cloud software, offering organizations a single platform for running apps and data across clouds. With Nutanix, companies can reduce complexity and simplify operations, freeing them to focus on their business outcomes. Building on its legacy as the pioneer of hyperconverged infrastructure, Nutanix is trusted by companies worldwide to power hybrid multicloud environments consistently, simply, and cost-effectively. Learn more at www.nutanix.com or follow us on social media @nutanix.