Enterprise Strategy Group by TechTarget

SHOWCASE

JULY 2024

Importance of Securing Workloads for Generative AI

Stephen Catanzano, Senior Analyst

Abstract: Generative AI (GenAI) has emerged as a transformative technology for businesses, streamlining operations, enhancing customer engagements, and building competitive advantages and innovations. In the world of IT, it is crucial to strengthen security by enhancing threat detection, access management, adversarial defense, and network security measures. To achieve its many benefits, securing the data and environments that GenAI solutions are built on is critical. Every GenAI solution created by an organization relies on storage and compute infrastructure, has a data foundation comprising specific internal data, and uses embeddings from a large language model (LLM) or foundation model (FM), along with other AI-related tools. Whether creating customer-facing GenAI solutions or using GenAI within your organization, clear security compliance is crucial to ensure that any data shared externally remains protected and private. This reduces the risk of model bias as well as the risk of model poisoning through malicious inputs. A combination of solutions from AWS Partners and tools from AWS—such as AWS Nitro, AWS Key Management Service, logging behavior, AWS PrivateLink, tenancy models, and more—can help solve these security challenges.

Security Ranks as a Top Challenge for AI

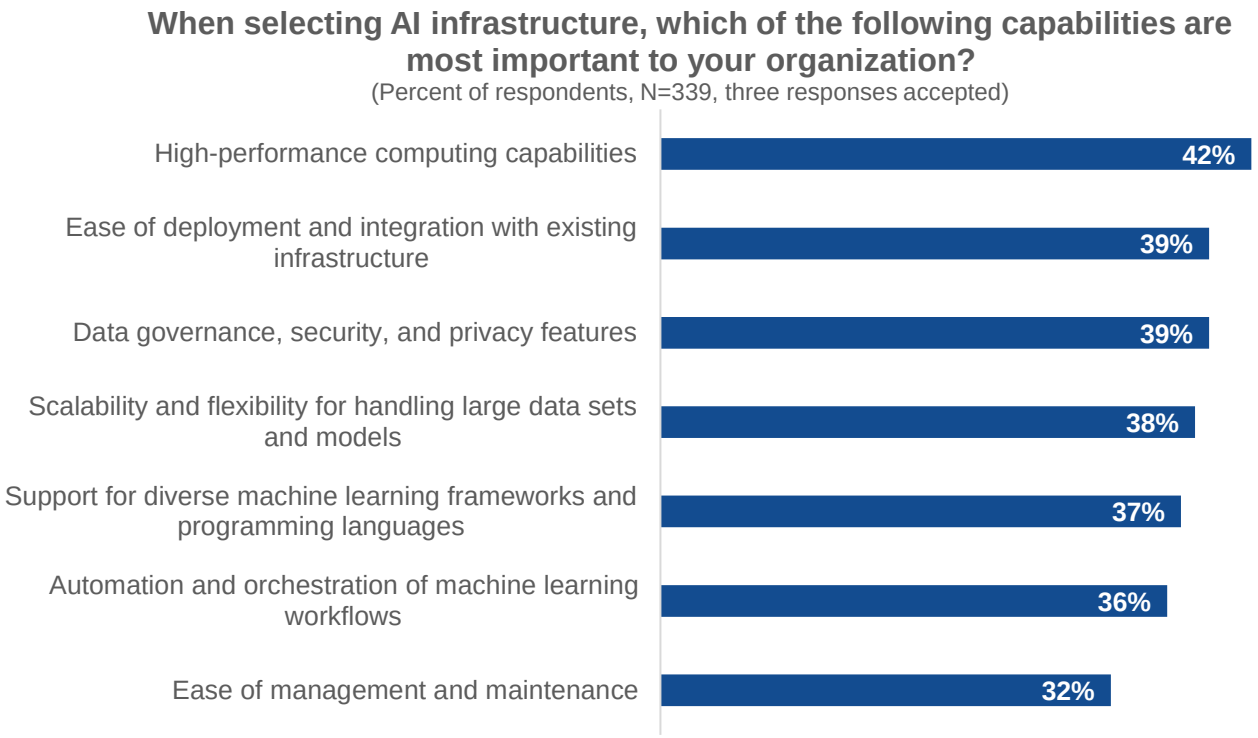
41% of organizations reported that they are challenged with the security risks and vulnerabilities associated with GenAI when integrating or associating GenAI with AI infrastructure.

A recent research survey by TechTarget's Enterprise Strategy Group asked respondents about the challenges or concerns their organization encounters or anticipates when integrating or associating GenAI with infrastructure.[1] Organizations see a wide array of challenges when assessing their infrastructure needs to support GenAI initiatives, with the most commonly cited challenge being "security risks and vulnerabilities associated with generative AI" (41%). Other challenges include "computational resource requirements for generative AI techniques" (39%), "difficulty in validating and evaluating generated rules" (38%), "employee hesitancy to trust recommendations" (38%), "legal and regulatory implications of generated content" (37%), "ethical considerations and biases in generated content" (37%), and "integration complexity with existing infrastructure and tools" (36%).

In a separate Enterprise Strategy Group survey, shown in Figure 1, respondents were also asked about the capabilities most important to their organization when selecting AI infrastructure. The top three responses focused on infrastructure performance; ease of deployment with infrastructure; and security, privacy, and governance—all critical to building AI solutions.

[1] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023. All Enterprise Strategy Group research references and charts in this showcase are from this research report.

Figure 1. Most Important Infrastructure Considerations



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

The Infrastructure Needed to Build Secure GenAI Solutions

GenAI is a powerful and transformative tool. Organizations building GenAI solutions use cloud infrastructure, internal data foundations, and LLMs and FMs as core elements—all of which need to be secure.

- **Cloud infrastructure:** To build GenAI applications, an enterprise needs a robust and scalable infrastructure that encompasses several key components. First and foremost, a powerful computing environment is essential, typically leveraging high-performance cloud services like AWS Elastic Compute Cloud instances with GPUs or Tensor Processing Units for efficient model training and inference. Secure and scalable storage solutions, such as Amazon S3, are necessary for handling large data sets and model artifacts.
- **Internal data foundations:** Once a use case for a GenAI application is determined, a data foundation is developed with the trusted, accurate, and high-quality data needed for the project. For example, if this was GenAI as an internal knowledge base, this foundation may include the data from a knowledge base repository, support and ticketing data, technical configuration data, and even data from SAP and Salesforce systems.
- **LLMs and FMs:** LLMs are designed to be language experts, trained on massive amounts of text to understand and respond to questions and requests in an informative way. FMs are broader and can be trained on all sorts of data, such as text, images, and video, enabling them to potentially understand the world in a more comprehensive way. Using Amazon Bedrock, an organization has the flexibility to test and deploy all the top LLMs and FMs.

Organizations need cloud infrastructure, a data foundation, and an LLM or FM. The LLM or FM provides the language expertise, but it lacks context, which is where internal data comes in and becomes the real power for an organization looking to create unique and differentiated GenAI solutions. Then comes securing it all.

Why Security Matters for GenAI Workloads

Securing GenAI workloads is crucial for several reasons, as these workloads often involve sensitive data, advanced models, and critical business processes. It is important to secure GenAI workloads for the following key reasons.

Protection of Sensitive Data

- **Confidentiality:** GenAI models often require a data foundation that might contain sensitive information such as personal data, intellectual property, or proprietary business information. Securing this data ensures that it is not exposed to unauthorized access.
- **Compliance:** Many industries are subject to strict and rapidly evolving data protection regulations (e.g., GDPR, HIPAA, EU AI Act). Ensuring that GenAI workloads are secure helps organizations comply with these legal requirements and avoid penalties.
- **Preventing tampering:** Unauthorized access can lead to tampering, resulting in corrupted or biased outputs. Ensuring the integrity of models is crucial for maintaining their reliability and trustworthiness.

Confidentiality of AI Innovations

- **Intellectual property protection:** GenAI models represent significant intellectual property. Protecting them from theft or unauthorized access ensures the safeguarding of proprietary solutions and innovations.
- **Competitive advantage:** Securing GenAI workloads helps maintain a competitive edge by preventing competitors from gaining access to valuable data and models.

Trust and Reputation

- **Customer trust:** Securing GenAI workloads builds trust with customers, partners, and stakeholders who need assurance that the data and the AI services they rely on are protected.
- **Brand reputation:** A security event can severely damage an organization's reputation. Proactively securing AI workloads helps protect the brand and maintain a positive public image.

Preventing Malicious Use

- **Mitigating abuse:** Ensuring that only authorized users have access to GenAI models prevents their misuse, such as generating misleading information, deepfakes, or other harmful content.
- **Safeguarding against security events:** Robust security measures help protect against various cyberthreats, including data security events, ransomware, and denial-of-service attacks, which could compromise AI operations.

By implementing comprehensive security measures, organizations can ensure that their GenAI workloads are protected against a wide range of risks, thereby supporting the safe, reliable, and ethical use of AI technologies.

How to Secure GenAI Workloads

Securing GenAI workloads on AWS involves a comprehensive approach that integrates various AWS products and features to provide robust compliance and management capabilities. Here are some of the AWS solutions to consider using together to create a comprehensive security posture for GenAI solutions.

AWS Nitro System

- **Security and isolation:** The AWS Nitro System enhances security by offloading hypervisor functions to dedicated hardware, minimizing the attack surface. This helps to ensure that GenAI workloads are isolated and secure.

AWS Key Management Service and Customer Managed Key

- **Data encryption and access controls:** AWS Key Management Service provides a secure way to create and manage cryptographic keys used for encrypting data at rest and in transit. This is vital for protecting sensitive AI

data and model parameters. Key Management Service also enables fine-grained control over who can access and manage encryption keys, ensuring that only authorized users and services can access sensitive data.

- **Cloud AI model protection:** In the context of AI workloads on AWS, AWS KMS Customer Managed Key is a Key Management Service key that you create, manage, and own. This allows you full control over your keys, including managing policies, grants, tags, and aliases.

AWS Logging Behavior

- **Monitoring and logging:** Using Amazon CloudWatch, AWS CloudTrail, and AWS Config, organizations can monitor the activities of GenAI workloads, log all API calls, and track changes. This allows visibility into the environment and helps to identify and respond to security incidents.
- **Auditing:** CloudTrail's detailed logs enable auditing of all actions taken within the AWS environment, helping ensure compliance with security policies and regulations.

AWS PrivateLink

- **Secure connectivity:** AWS PrivateLink enables organizations to access AWS services without exposing the data to the public internet. This helps ensure that communication between the customer Virtual Private Cloud and components in the GenAI workload, including Amazon Bedrock and S3 (e.g., between Elastic Compute Cloud instances and S3 storage), remains secure and private.
- **Reduced attack surface:** By keeping traffic within the AWS network, PrivateLink reduces the potential attack surface, making it harder for attackers to intercept or tamper with data.

Security Tools Working Together

The combination of AWS Nitro System enhanced security and performance, AWS Key Management Service encryption, and AWS PrivateLink secure connectivity ensures that GenAI workloads are running in a highly secure environment. Logging and monitoring tools like CloudWatch, CloudTrail, and AWS Config provide continuous visibility into the environment, enabling monitoring for suspicious activity and audit actions and maintaining compliance. Key Management Service ensures that all sensitive data, whether at rest or in transit, is encrypted and protected with robust key management practices. The AWS tenancy model enables the isolation of workloads for cost-efficiency or dedicated resources for maximum security and compliance.

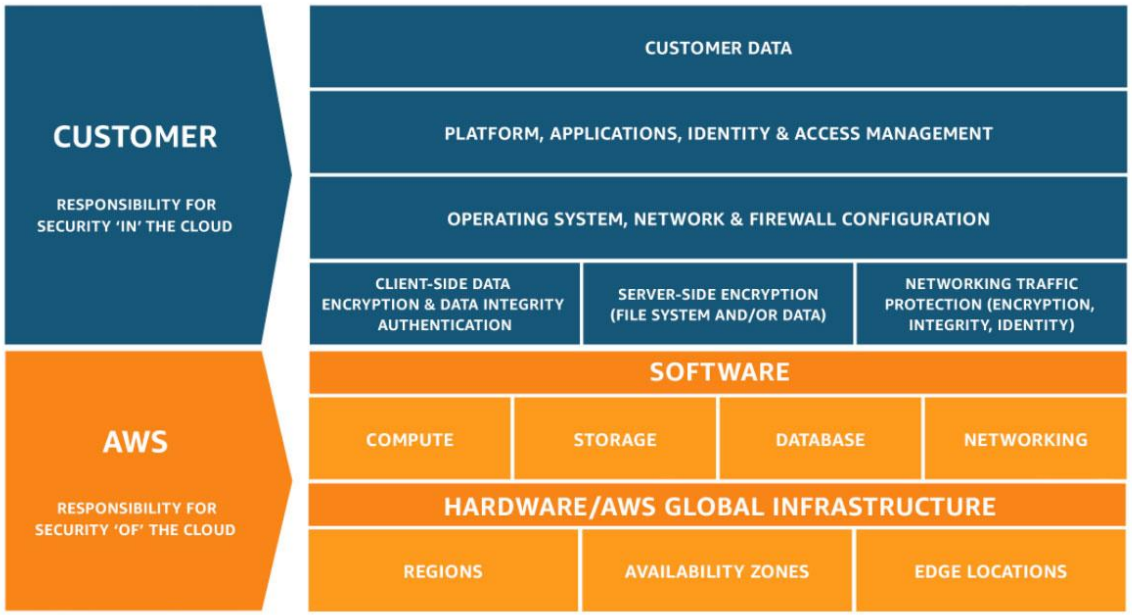
By integrating these AWS products, organizations can build a more secure, high-performance, and more compliant environment for running GenAI workloads, helping to ensure that both data and models are protected throughout their lifecycle.

Security Is a Shared Responsibility

It's important to remember that security is a shared responsibility. The key areas of shared responsibility are shown in Figure 2.[2] AWS takes responsibility for the "Security of the Cloud" and is responsible for protecting the infrastructure that runs all the services offered in the AWS Cloud. This infrastructure is composed of the hardware, software, networking, and facilities that run AWS Cloud services.

[2] Source: AWS, *Shared Responsibility Model*, 2024.

Figure 2. AWS Shared Responsibility Model



Source: Amazon Web Services

AWS customers are responsible for implementing "Security in the Cloud," even with AI workloads. This means the specific AWS services chosen for the AI task.

For instance, if an organization is building a GenAI application, such as a knowledge base that uses internal data with an LLM or FM, they may use Amazon Bedrock, Amazon Q, Amazon SageMaker, and Amazon Elastic Compute Cloud. Securing all these environments is crucial. Amazon Bedrock provides access to LLM and FM models for testing and deployments and gives organizations full control over the data used to customize the foundation models for GenAI applications. Data is encrypted in transit and at rest, and you can create, manage, and control encryption keys using AWS Key Management Service. Amazon Q, which is a GenAI-powered assistant that interfaces with the foundation data, can understand and respect existing governance identities, roles, and permissions while personalizing interactions accordingly. Amazon SageMaker handles some of the underlying infrastructure security, enabling organizations to focus more on the AI project itself. SageMaker provides a managed environment for building, training, and deploying machine learning models. With Amazon Elastic Compute, the organization is responsible for most security configurations and access controls. This includes managing security patches, user permissions, and encryption for AI data and models. Thus, AWS and its partners work together with customers to build a comprehensive security model for GenAI workloads regardless of where they may live in your environment.

Conclusion

GenAI unlocks a new era of business transformation, but its potential hinges on a secure foundation. To harness its power safely, organizations must prioritize securing the data and environments underpinning GenAI solutions. This includes securing the storage and compute infrastructure, the data used for training, access controls, and integration with the LLM. Robust security practices are essential to prevent data security events, mitigate model bias, and safeguard against malicious manipulation.

By prioritizing security and cultivating a trusted partnership with AWS, businesses can unlock the full potential of GenAI while minimizing risks, ensuring a smooth and secure journey in their emerging market endeavors. If your organization is looking to securely build GenAI solutions, Enterprise Strategy Group strongly recommends AWS and its partners.

AWS Partner Spotlight: Cyera

By leveraging AWS's infrastructure and Cyera's data security capabilities, customers can input secure, compliant, and accurate data into their AI models. Cyera enables this by effectively identifying and classifying data before it's used in AI models. This approach, known as AI data assurance, helps prevent the misuse of data, which can lead to unreliable and high-risk AI outputs. Equally important is customers' ability to understand their "AI blast radius." Should an issue arise, knowing what data was used in AI models and its sensitivity can accelerate incident investigation. Together, AWS's scalable cloud resources and Cyera's expertise in data security make it easier to manage large data sets and complex AI initiatives. Ultimately, the AWS and Cyera combination supports the creation of trustworthy AI applications that drive innovation and deliver valuable insights, while providing security, privacy, and compliance.

For more information about Cyera, click **HERE**.

Note: Content in the above Partner Spotlight section was provided by Cyera and edited for clarity by Enterprise Strategy Group. Enterprise Strategy Group has not necessarily been briefed by the featured partner and readers should perform their own research into the partner's offerings and capabilities.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com