TRUSTLOGIX

Navigating the *AI Data Security Landscape*

Introduction

The rapid adoption of artificial intelligence (AI) has ushered in a new era of innovation and efficiency for businesses across industries. However, this integration also brings significant security risks that organizations must proactively address to protect their sensitive data and AI assets. As AI becomes integral to business operations, robust AI data security has become a critical priority for technology leaders and security professionals.

Understanding the Increasing *Relevance of AI*

AI has rapidly evolved from simple algorithms to advanced deep learning models with near-human precision. Today, technologies like Large Language Models (LLMs) lead innovation, with Retrieval-Augmented Generation (RAG) enhancing LLMs by incorporating external data for better accuracy and flexibility. As a result, AI is now integrated into nearly every industry.

> *"For the past 2 years, I've been leading a generative AI research and development group at Zeta Global. We built a **large language model-based agent** that has access to various layers of data that can execute complex analytics queries and create marketing campaigns."*
>
> **Danny Portman,**
> Head of AI at Zeta Global,
> highlighted this trend.

Exploring the *Security Risks* Posed by AI

While organizations are attracted to AI because of the benefits it offers, the methods used to expose, share, train, and deploy AI models also make them appealing targets for malicious actors. Some of the critical security risks organizations face include:

Data Privacy and Confidentiality

As enterprises depend more on AI/ML for insights and decisions, safeguarding sensitive data becomes crucial. AI systems often require access to large amounts of business data, increasing the risk of unauthorized access or data breaches. Ganesh Kirti, CEO of TrustLogix, emphasized this point.

Model Protection

As per OWASP Machine Learning Security Top 10 project, Models can be subjected to several attacks:

MODEL POISONING

An attacker manipulates the model's parameters to cause it to behave in an undesirable way.

MODEL INVERSION

An attacker reverse-engineers the model to extract information from it.

MODEL SKEWING

Attacks occur when an attacker manipulates the distribution of the training data to cause the model to behave in an undesirable way.

Malcolm Harkins, CSO of Hidden Layer, shared a relevant example:

*"Six months after Blackberry bought Cylance, an adversarial research team posted a blog, saying, 'Cylance, I kill you.' **They were able to**, because of the interactions they were able to do with the machine learning capability, be able to **figure out the weights of the model**."*

HIDDENLAYER

Regulations & Standards

Several key regulations aimed at AI such as the EU AI Act and New York City Local Law 144, have been introduced to control the use of AI. Additionally, generic regulations such as GDPR add complexity due to the vast amount of data involved in AI projects.

Malcolm Harkins, CSO of Hidden Layer, noted the rapid pace of regulatory developments :

"In my 35 years in tech, I've never seen such rapid legal and regulatory developments as with AI. The EU AI Act, presidential executive orders, and NIST regulations are setting fast-paced expectations for the technology."

HIDDENLAYER

Key Recommendations for *AI/ML Security Measures*

Based on insights from the panel **discussion** and broader industry trends, here are some key recommendations for organizations navigating the AI data security landscape:

Secure the Entire AI Development Pipeline
Don't just focus on model security. Protect data throughout its lifecycle, from collection to storage in AI systems.

Adopt a Risk-based Approach
Identify where AI introduces substantial risk to the business. Focus security efforts on protecting the most critical data and AI assets.

Leverage Data Security Capabilities
Utilize advanced data access control methods, such as ABAC and PBAC, and integrate them into your organization's existing cloud infrastructure to provide comprehensive protection across platforms. Implement automated tools for data discovery, classification, and protection.

Foster Collaboration
Break down silos between data science, IT, and security teams. Establish governance frameworks that bring these groups together.

Raise Awareness
Help business leaders understand AI security risks and necessary investments. Work towards greater explainability in AI systems to aid in security auditing, building trust with stakeholders and Ongoing employee training on cloud data security

Audit & Assess Risks
Regularly conducting security assessments and penetration testing of cloud-based AI systems.

Key Capabilities to *Protect Data and LLMs*

To mitigate these risks, organizations need to implement robust security measures across their AI development and deployment pipeline. Some key prevention strategies include:

Model Security

Tracking model versions through a model registry is essential to track the model throughout the lifecycle. Also the LLM's must be **protected** from unauthorized modifications and access.

Malcolm Harkins emphasized the importance of this:

"Ask your teams how they're doing a security development life cycle and validating that models have integrity and/or that they do not have embedded malicious code. Ask if they built the models themselves or downloaded public foundational models. These create exposure points for you, and then also ask, "How are you doing real-time, runtime protection of your AI projects?"

HIDDENLAYER

Data Discovery, Classification and Lineage

Implementing a strong data discovery and classification process helps identify and categorize sensitive information in AI training datasets, such as PII and PHI. Tracking the origin, movement, and transformation of this data ensures transparency and understanding of how it's used in AI/ML models.

Data Access Control

is a fundamental part of protecting AI/ML systems as it helps in:

1. **Securing Model Access** through strict access control and authentication measures and ensuring only authorized individuals can access and manipulate the models.
2. **Protect Data used in Training and Inference:** Controlling access to sensitive data used in prompts and inference is critical to prevent unauthorized disclosure or misuse.

Granular access controls such as role-based access control (RBAC) and attribute-based access control (ABAC) **policies** ensure that roles, user attributes, as well as the data and context of the request is used to determine access to data objects.

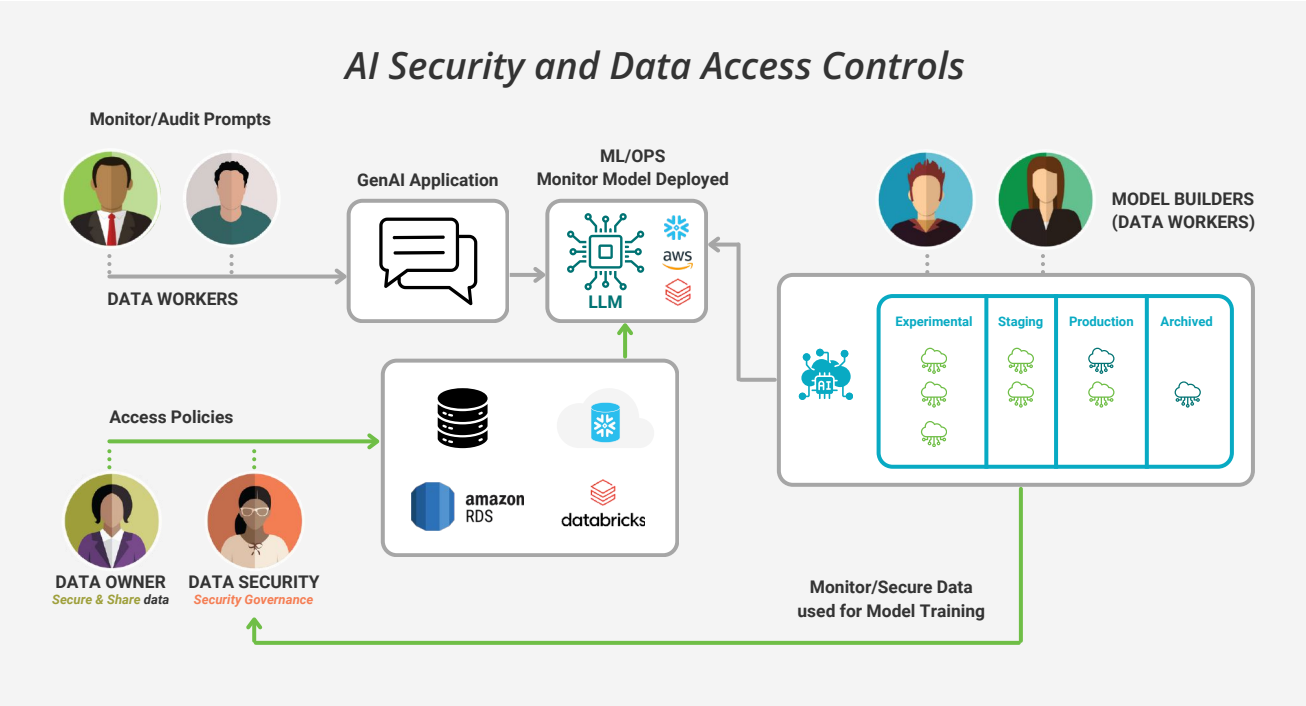
Right Sizing the privileges across each data platform is essential so that only the necessary objects are granted access. **Templatized** Approach ensures a consistent set of permissions across data platforms and a clear overview of who has access to what data and services.

For businesses utilizing hosted models like those from Amazon or Microsoft, it's essential to monitor which services are being used and what sensitive data is being transferred into these models. For instance, gaining visibility into the sensitive data entering the Open Source Llama model within the Snowpark Container Service—including details such as the dataset used, data classification, and recommendations for managing these concerns—is crucial.

With the growing popularity of Retrieval-augmented generation (RAG) in terms of return more specific, context-based answers – without the need to build, train, and deploy a proprietary LLM. TrustLogix RBAC and ABAC policies help audit around access rights and protect sensitive information from making its way into LLM responses, as well as monitor and audit the data being used in models or responses.

Data Quality Monitoring

Data Quality involves assessing the state and integrity of data, including its freshness and accuracy, and identifying null or blank fields. Snowflake offers built-in Data Metric Functions as well as allow users to define their own custom DMFs. These insights enhance your data governance posture and the security of the AI/ML pipeline and offer protection against model poisoning and obsolete models that could break the compliance.



Strategies to *Future-Proof AI Security*

To protect AI systems from emerging threats, implement flexible security models and invest in advanced research. The strategies outlined below are crucial for keeping up with the evolving threat landscape and ensuring lasting AI security.

Stay Updated on Threats and Regulations

Keep abreast of emerging attack vectors targeting AI systems. Stay updated on evolving AI regulations and design data architecture to support compliance requirements.

> **Malcolm Harkins** recommended:
>
> *"Go look at MITRE ATLAS and look at those attack factors, and then you'll be able to frame some additional questions you should be asking."*
>
> HIDDENLAYER

Leverage Advanced Tools

Use AI-powered tools for threat detection, anomaly detection, and automated incident response to enhance overall security posture. Explore emerging techniques like federated learning and differential privacy that enable AI training on distributed datasets while preserving privacy.

Plan for the Future

As **Danny Portman** emphasized:

"Don't just build for today's AI capabilities. Try to anticipate where the technology is heading and future-proof your security architecture. Ensure data infrastructure can scale to handle growing volumes of training data and model artifacts."

ZETA

Continuous Monitoring and Auditing

Implementing robust logging and monitoring solutions to detect unusual access patterns or potential security incidents such as lineage, Data Quality, Excessive Access, Shadow Client tool usage, unused privileges, etc. Adopt a proactive approach to data risk management, anticipating potential threats and implementing preventive measures.

Summary

As AI becomes increasingly central to business operations, taking a proactive, risk-based approach, implementing robust prevention strategies, and continuous assessment will help organizations navigate this complex landscape successfully. This requires ongoing collaboration between security, data science, and business teams. Data protection needs a methodical **process** and a step-by-step approach.

This not only protects sensitive data and AI assets but also builds trust with customers and stakeholders, enabling the organization to fully leverage the power of AI in cloud environments.

Visit the page trustlogix.io/safeguarding-your-ai-models-and-data to learn more on how to streamline Data Security and Privacy for your AI Models and Data. **Sign up** for a free 90-day protection service today by visiting : trustlogix.io/free_trial