VIEW POINT

HUMAN CENTRIC DESIGN OF SECURITY DATA LAKES FOR BETTER ADOPTION OF AI FOR CYBER DEFENSE

Abstract

Cybersecurity is evolving into a strategic differentiator. Businesses today are investing in AI powered cyber defense capabilities to fight off enterprise scale Gen AI powered threat actors, and improve organizations' risk resilience. This POV provides a reference architecture and human centric implementation approach on how enterprises can build an AI-first AWS security data and accelerate AI adoption.

Infosys®
Navigate your next

How can enterprises use AI for Cyber Defense?

Cybersecurity is a part of C-suite discussions centred around business cases tied to strategic initiatives. It is no longer about protecting enterprise assets, infrastructure or managing risk. Enterprises would be spending 101.5 Bn USD by 20251 in reinforcing their enterprise defenses. It is a strategic business enabler that shapes product/ service delivery capability, organizational effectiveness, and relationships with customers, partners, and employees. In today's threat landscape, adversaries equipped with Gen AI are not limited to individuals but enterprise scale organizations.

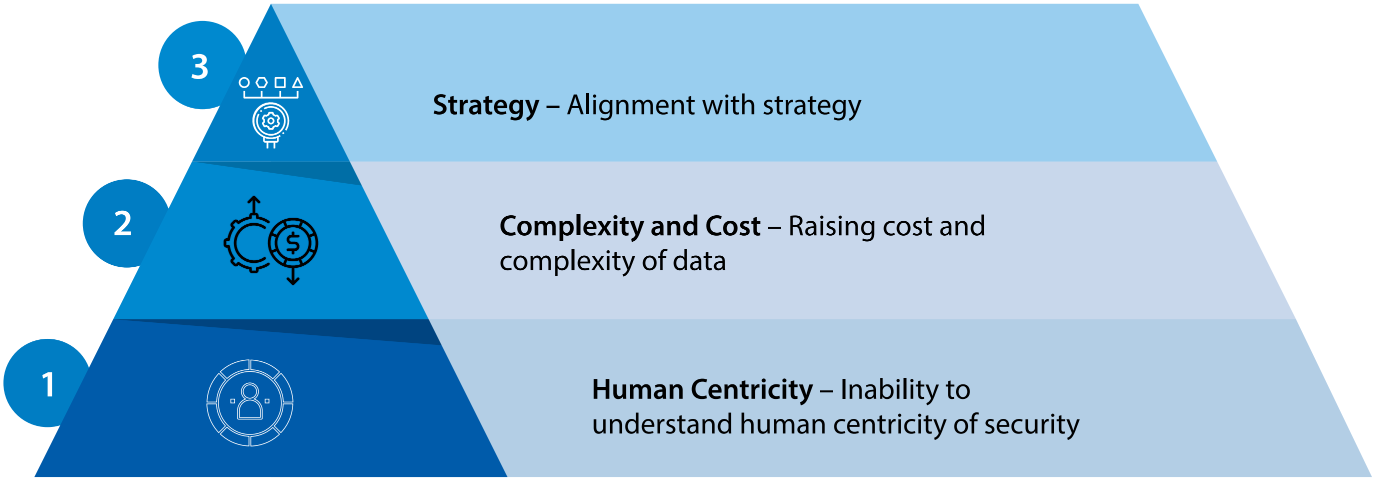
As more businesses move to cloud, and adopt AI, they are facing a Gen AI fuelled threat landscape. The variety, volume, and velocity of the data processed by enterprises has increased exponentially. Enterprises need to use AI for cyber defense by combining security data with contextual business data. Traditional security management systems have proven inadequate for handling this business demand. A security data lake allows enterprises to collect large volumes of data and apply AI and analytics to derive insights without spending too much effort on data processing. AI applications powered by a security data lake aims to achieve high search investigation, higher accuracy, and improved compliance standards.

Today, an enterprise centric view dominates the cybersecurity procedures including how enterprises build and use security data lakes. Deriving from the theory of protection of critical national infrastructure and economics, the enterprise centric cybersecurity focusses on protection against threats which impact- the enterprise assets and damage critical infrastructure. This paper explores an alternate method of human centric approach in building security data lakes which can enable enterprises to have 360-degree view of their enterprise risk. It provides a reference architecture and an implementation approach to protect customers, partners or employees who have the largest risk of attack and improve enterprise AI adoption for cyber defense.

Why do we struggle to build or adopt AI for Security data lakes?

Despite making security investments and having unified data through a security data lake, enterprises are struggling to capture the real value of security data for AI and analytics. Further, security teams face restrictive compliance limits on which data they can collect from sources and how long they can keep/store it in their repositories for analysis. The following are the key challenges faced by enterprises building security data lake.

Challenges in building security data lakes



3 **Strategy –** Alignment with strategy

2 **Complexity and Cost** – Raising cost and complexity of data

1 **Human Centricity** – Inability to understand human centricity of security

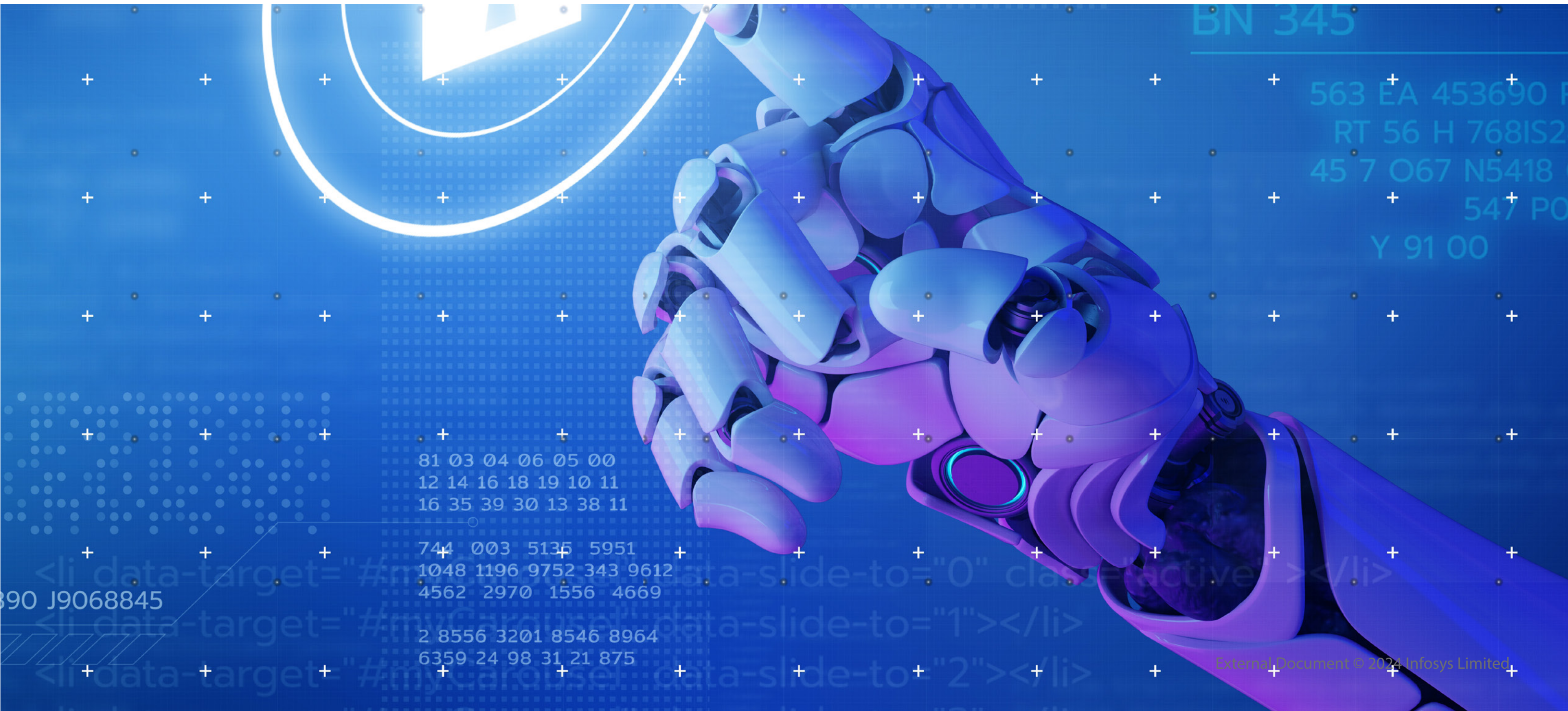
1. **Failing to Understand the Human Centricity of Data** - Security data lakes hold a treasure trove of sensitive data on end users like their logins, application access, downloads, and even email activity. Building Gen AI applications on the aggregated data without proper understanding of the user behind the data, can create ethical, privacy and security issues. For instance, projects like CYC 2 started in 1984, failed to create an impact as it was a knowledge representation of common sense rather than the human centric view. While technology is a key enabler, security ultimately relies on people. Enterprises fail to build a human-centric approach that considers user behavior and integrates security awareness training - essential to maximize the effectiveness of a security data lake.
2. **Complexity and Cost** - Security data lake tends to be used as a tactical solution to address the limitations of legacy SIEM. By moving away from SIEM systems to security data lakes, organizations have separated security data management from security operations. This separation can lead to inefficiencies, as businesses may find themselves navigating between disparate systems while trying to correlate data and insights effectively. Security data lakes can become data swamps if there's a lack of integration with existing security tools and SIEM systems. Without proper data governance and quality checks, the cost and complexity of the data increases without yielding a positive ROI of security data lakes.
3. **Alignment with Strategy** - Building business outcomes driven by AI from the security data lake has been a challenge for many enterprises. Just throwing data into a lake isn't enough. A larger retailer did not get any results for 3 years of their security data lake journey. Without clearly defining security objectives and use cases, the data becomes overwhelming and difficult to analyze. For instance, enterprises will have more success by working on clearly defined metrics and use cases, like "detect phishing attempts" or "identify risky user behavior," than vague goals like improve overall security. Data collection should be aligned to the right security metric chosen by enterprises for effective data capture, calibration, and control.

Implementation approach - Infosys 4D for building security data lakes on AWS

Enterprises can build a foundation and then accelerate their AI powered security data lake journey based on executive commitment, engineering maturity and security investments. Infosys is committed to the 4D approach of Diagnose-Design-Deliver-Defend[5] for building a holistic implementation of security data lake and accelerating the AI journey for cyber defense.

- Diagnose enables assessment and analysis of the security data collected into the security data lake.
- Design phase focusses on design of scalable and future ready capabilities to build AI applications along with guardrails for the implementation of AI.
- Deliver is for automation, transformation, and orchestration of AI capabilities on AWS security data lake.
- Defend is critical to govern, support operations, audit and respond to attacks based on data.

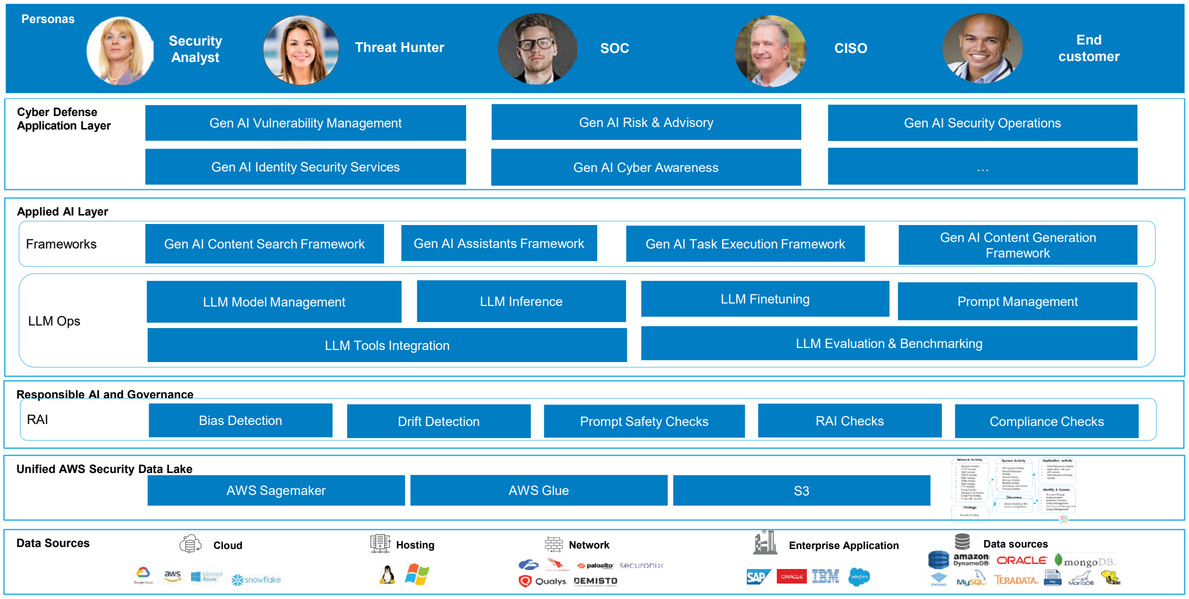
4D process framework for AWS security data implementation has enabled enterprises to build holistic resilience, embed security by design and build a collaborative platform between customers, partners, third parties, and regulators.



How can we build a security data lake which will accelerate AI adoption for Cyber defense?

Infosys recommends a simplified reference architecture for building AWS based security data lake at enterprise scale. This architecture consists of five layers which enables better AI adoption and embraces human-centric security as the core principle.

Accelerating Gen AI adoption with Infosys Applied AI Platform



Personas: Security Analyst, Threat Hunter, SOC, CISO, End customer

Cyber Defense Application Layer: Gen AI Vulnerability Management, Gen AI Risk & Advisory, Gen AI Security Operations, Gen AI Identity Security Services, Gen AI Cyber Awareness, …

Applied AI Layer

Frameworks: Gen AI Content Search Framework, Gen AI Assistants Framework, Gen AI Task Execution Framework, Gen AI Content Generation Framework

LLM Ops: LLM Model Management, LLM Inference, LLM Finetuning, Prompt Management, LLM Tools Integration, LLM Evaluation & Benchmarking

Responsible AI and Governance

RAI: Bias Detection, Drift Detection, Prompt Safety Checks, RAI Checks, Compliance Checks

Unified AWS Security Data Lake: AWS Sagemaker, AWS Glue, S3

Data Sources: Cloud, Hosting, Network, Enterprise Application, Data sources

1. **Persona based UX layer** - Human centric security needs understanding of the intent of each persona in the enterprise security team. This persona-based approach helps us understand how security team members will think and behave. For instance, for a security analyst, it is critical to have the right visibility across the business that allows them to understand unusual user behavior, capture it, calibrate it against expected behavior and control it. These business metrics must be mapped to each personal before the designing of the security data lake so that AI application are adopted and used correctly.

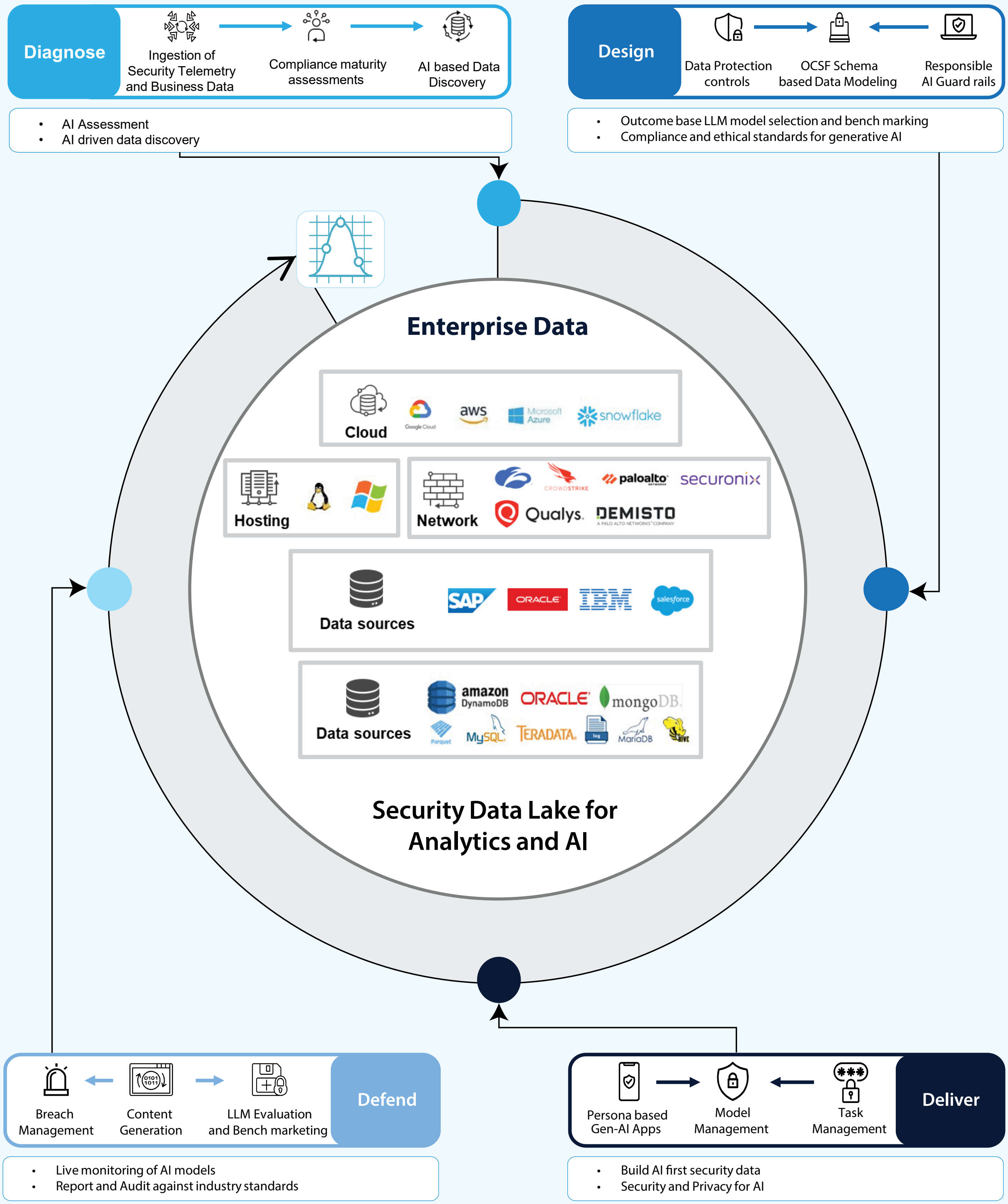
2. **Cyber Defense Application Layer** - Enterprises can build security application as per the domains which that are catered by th security data lake. The Gen AI applications are built on the insights generated from the AWS security data lake created by Sage maker across each domain of cybersecurity. For example, there is a rapid overlap between OT and IT environment in manufacturing set ups. The mapping of the interdependency of different business functions enables teams to build applications which gives a threat hunter a unified view to grasp the implication of a ransomware attacks across the organization.

3. **Applied AI layer** – Enterprises can adopt an applied AI layer. This is a collection of LLM Models, frameworks, adapters, and synthetic training data which accelerates the build of AI based pipelines for security data lake. For instance, enterprises might need semantic content search across different types of security documentation such as SOPs, advisories, and latest threat intel across multiple security domains. An applied AI layer acts as a library to help in re-use of semantic search capability across the stack. This accelerates development through automation of model training, tuning and deployment.

4. **Responsible AI and Governance Layer** – It is critical to build and deploy a balanced cybersecurity approach that doesn't invade users, partners or employee's privacy or security but still allows visibility over security operations metrics derived from the security data lake. Acts like EU-AI Act 3 are making regulation the top imperative for adoption of AI for security analysis. Responsible AI and governance capabilities such as bias detection, regulatory compliance and prompt safety checks must be designed before the data lake design can be taken up.

5. **Unified AWS Security Data Lake** - Enterprises need to establish visibility into their enterprise technology assets, data sources and business systems. Security teams cannot protect to what they do not have visibility on. The journey begins with gaining and maintaining real-time visibility into the assets and correlate thatedthem with the business systems. OCSF4 based schema for the security data lake enables enterprises to get insights from the business data with their security telemetry. While some insider threats might be deliberate, others are accidental due to human error or lack of awareness. Analysis of surge in access attempts from an unusual location or attempts to access unauthorized files analyzed through the security data lake can enable better protection of assets and areas where security controls need to enhance.

Implementation approach



Diagnose

Ingestion of Security Telemetry and Business Data → Compliance maturity assessments → AI based Data Discovery

- AI Assessment
- AI driven data discovery

Design

Data Protection controls → OCSF Schema based Data Modeling ← Responsible AI Guard rails

- Outcome base LLM model selection and bench marking
- Compliance and ethical standards for generative AI

Enterprise Data

Cloud

Hosting

Network

Data sources

Data sources

Security Data Lake for Analytics and AI

Defend

Breach Management ← Content Generation → LLM Evaluation and Bench marketing

- Live monitoring of AI models
- Report and Audit against industry standards

Deliver

Persona based Gen-AI Apps → Model Management ← Task Management

- Build AI first security data
- Security and Privacy for AI

The AWS security lake implementation can be visualized by the pipeline moving through 4 key stages of implementation:

1. **Diagnose** – First step is to build capabilities to ingest, process and validate multiple log sources and security telemetry. It is the cornerstone or the first step in building high quality data in the security data lake. This stage also includes data mapping int the OCSF format to harmonize the data in the data lake. A data discovery exercise will also help identify the critical sensitive data getting ingested into the data lake.
2. **Design** - Intent driven persona-based experiences are designed for each AI application built on the security data lake. The right AI data model is picked based on the business use cases and the right guard rails of AI are designed and deployed in the AI pipeline.
3. **Deliver** - Scale and performance is the key in this stage when each layer of the solution works together to solve specific security business problems. Enterprises should focus onbuilding key use cases such as threat hunting, incident response management, Data Loss prevention, Analytics / forensic and next gen SIEM analytics for each persona of the security team.
4. **Defend** - Enterprises should focus on building proactive defense in this layer including capabilities to manage breaches on the security data. Enterprises can build automated compliance assessments and audits against industry standards. Further, there should be centralized communication and awareness of privacy among all the stakeholders which include end-users, partners, employees, and external vendors

Conclusion

Journey of AI for Cyber Defense using Security Data Lake

Studies show that high impact attacks on enterprises can paralyze a significant part of the enterprise. The largest attack in the oil and gas industry, Colonial pipeline attack[6], was a ransomware attack on the billing systems, however it could effectively shut down pipeline operation. This attack, despite targeting IT systems, created fuel shortages across the country. It highlights the interconnectedness of modern infrastructure and the need for a holistic approach to security. These attacks show that enterprises still struggle with fragmented technology and growing skill gap. A human-centric approach to building the security data lake and AI applications using that data empowers CISOs and business leaders to demonstrate the value and performance indicators of the cybersecurity program. By building foundational capabilities first, enterprises can embark on a journey to reduce risk, achieve holistic resilience, and embed security by design into next-generation security data lakes, all aligned with the security team's business priorities.

About the Author

Karthik Nagarajan,

Practice Manager and Senior Industry Principal

Karthik heads Infosys Data Protection and Privacy services. He has 17+ years of experience in product design and consulting services, with an expertise in AI, data privacy and customer experience strategy.

Vasudha Vasudev,

Senior Associate Consultant

Vasudha is a marketing enthusiast working for Infosys CyberSecurity marketing team. She has 2 years of experience in marketing, content generation, with a novice subject expertise in AI, data privacy and protection, CSR, Architecture and healthcare and life sciences.

References and Further reading

1. Cybersecurity trends: Looking over the horizon - McKinsey 2022
2. CYC - The Next Generation of Enterprise AI
3. The EU AI Act: An enterprise response strategy, IKI Infosys, Apr 2024
4. OCSF Schema – v1.2
5. Infosys 4D approach of Diagnose-Design-Deliver-Defend
6. U.S. Pipeline Cyberattack Forces Closure, WSJ, May 2021

For more information, contact askus@infosys.com

Infosys®
Navigate your next