Safeguarding Data at Scale in the AI Age

JANUARY 2026

CONTENTS

EXECUTIVE OVERVIEW 03

DATA IN THE AI AGE 04

THE SECURITY IMPERATIVE 08

SAFEGUARDING YESTERDAY’S WALLED PREMISES 10

SAFEGUARDING TODAY’S CLOUD DATA ASSETS 12

SAFEGUARDING DATA IN TOMORROW’S AI AGE 15

BRIDGING THE CULTURE GAP: ALIGNING PEOPLE AND SECURITY 21

CONCLUSION: THE JOURNEY AHEAD 22

EXECUTIVE OVERVIEW

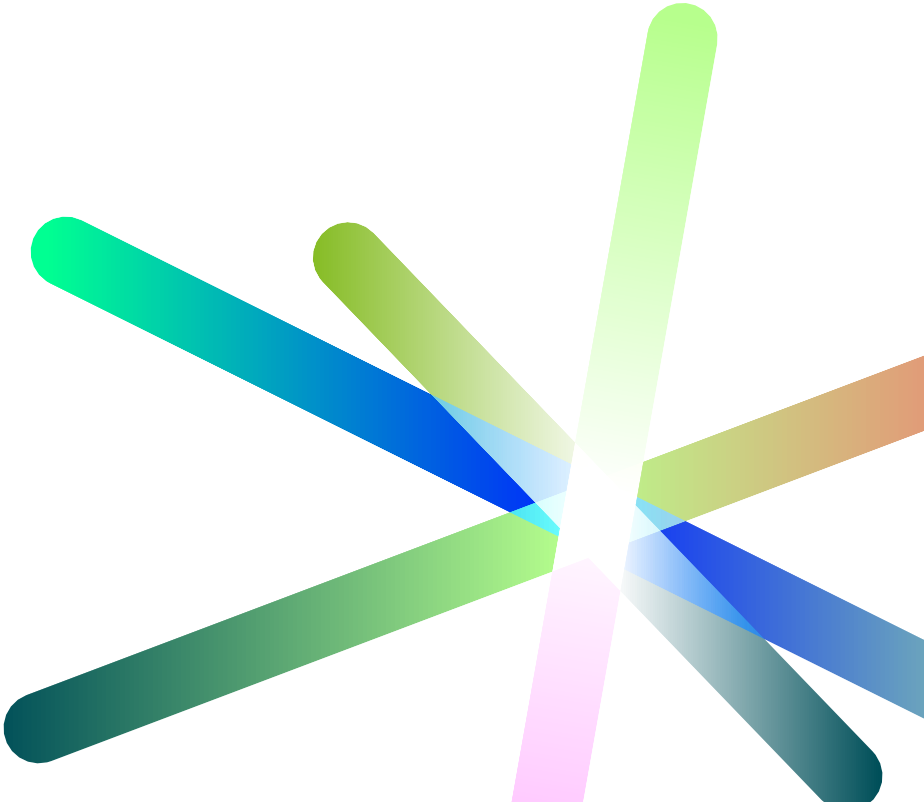
The Data Landscape: Over the last decade, organizations tackled an ever-increasing variety, volume, and velocity of data by leaning on a Big-Data landscape that enabled beneficial insights driving top-line revenue growth, marketplace positioning and brand loyalty. But achieving these benefits came at the cost of added complexity, operational disruptions, technical challenges, and significant investment in ongoing transformation of data platforms and storage infrastructure. Furthermore, this complex transformation has led organizations to be highly susceptible to data-related vulnerabilities and exploits, as demonstrated by a rise in adverse events impacting sensitive data (data loss, misuse etc.)

Security Imperative: Organizations face an imperative to safeguard data in response to rising adverse events from stakeholders (customers, employees, third-party partners, and regulators). Given non-compliance leads to loss of customer trust and punitive regulatory actions, executive decision-makers have a heightened awareness of the imperative and have continued to push security directives and programs. But, despite this push, organizations are still struggling to meet stakeholder needs as demonstrated by the continued rise and impact of adverse events.

Safeguarding Data at Scale: The challenges faced by many organizations today can generally be attributed to a combination of three factors. Firstly, organizations generally have issues defining data security programs with clear objectives, leading to underfunded teams and roadmaps with competing priorities. Next, inspite of innovative tools, there is an inherent technical complexity to secure a "tri-modal" landscape made up of legacy on-premises data stores, modern cloud data platforms, and the emerging class of Artificial Intelligence (AI) assets. This complexity can be a major hurdle for even well-funded data security teams to overcome, especially as they navigate business and user constraints. Finally, a misaligned organizational culture can contribute largely to undermining well-designed and well-funded data security capabilities, for instance, users circumventing security measures through "shadow IT" or haphazard storage practices.

Objective and Organization: This whitepaper is primarily intended for Data Security and Governance teams, Data Science and Product teams, and stakeholder groups tasked with safeguarding data at scale in the AI age.

The primary objective of this whitepaper is to create awareness of the data security challenges in the "tri-modal" landscape and outline emerging approaches to overcome these challenges. It describes the evolution of enterprise Data Security practices, along-with their application and limitations in the context of the "tri-modal landscape". This helps build a shared understanding amongst stakeholder groups and sets a common stage for a comprehensive approach centered around data posture management, access governance frameworks and emerging capabilities to tackle Data Security in the AI age. The whitepaper additionally covers a combination of concepts, illustrative architectures, and practical guidance for an Amazon Web Services (AWS) based environment. Finally, a theoretical journey is introduced to help organizations navigate their path to an effective program for safeguarding data at scale



DATA IN THE AI AGE

Organizations generate, collect, store, and manage vast amounts of raw and derived data. This includes structured transactional data, analytical workloads, unstructured assets such as documents, images, audio, and video, user-generated content like code files, credentials, secrets, and business intelligence reports. Initially these data assets were just utilized for operational and transactional processes, but over the past decade, teams have leveraged the massive volume of historical data assets to extract insights that transform the business and drive competitive advantages.

A Tri-Modal Landscape

Typically for organizations, a mixture of constraints ranging from system integration and compatibility requirements, application and data pipeline downtime limits, critical business process support, cost, and performance drivers end up hampering any modernization initiative. Thus, as data teams navigate these constraints over the course of their transformation efforts, they end up with a varied pace of adoption across the enterprise i.e., some teams may pivot faster than others to new modern platforms. This leads to organizations ending up with a complicated "tri-modal" landscape of legacy on-premises systems (past compatibility), modern cloud-hosted data stores (current state-of-art), and emerging platforms (early adoption) supporting the rise of Generative AI (GenAI) applications.

Data Platforms in the AI age: Organizations are gravitating towards a Tri-modal data landscape, dominated by versatile data architectures (such as lakehouses) enabled by scalable cloud-based infrastructures.

How we got here: Cloud-based Storage Infrastructure

A dominant shift with a clear path to business value is the transition of modern infrastructure to a cloud-based environment from legacy on-premises components. This shift has been underway for the better part of the last decade, with organizations gaining benefits like improved flexibility, scalable demand management, and effective cost management. From a purely data infrastructure standpoint, legacy on-premises components have been extremely fragmented with a variety of platforms, formats and services prevalent in a typical enterprise setup. Data engineering teams have leveraged the power of cloud environments to get a streamlined cohesive ecosystem across the entire pipeline from ingestion to analytics workloads. Teams can spin up a right-sized data environment rapidly and flexibly in line with changing resource demands and experimentation needs, avoiding upfront hardware investments and maintenance fees. The emergence of SaaS and vendor-managed platform further empowers data teams through pre-configured security, auto-scaling, and embedded governance features, minimizing the lift on technology teams.

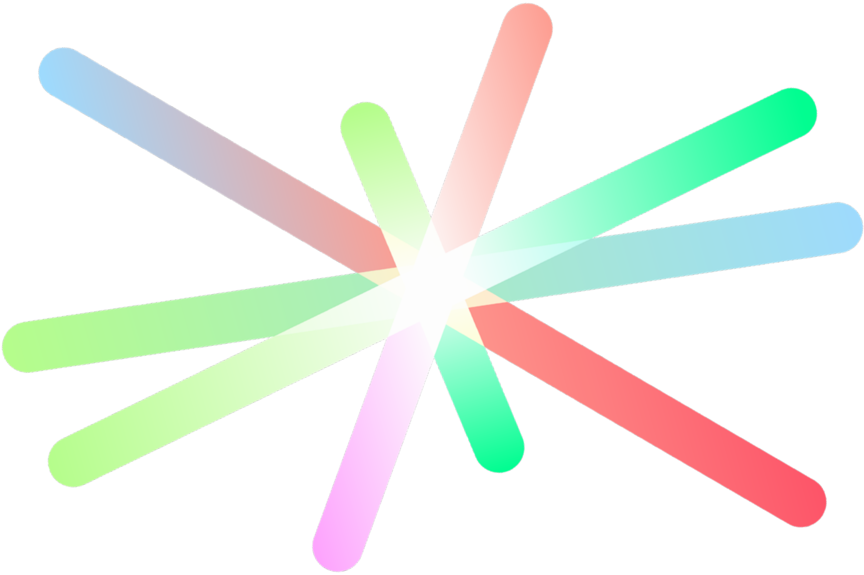
How we got here: Data Platforms

Irrespective of the storage infrastructure, enterprise teams depend on a variety of data platforms to manage their data products and assets, depending on the data format, consumption patterns, and state-of-art technology innovations.

Structured data initially took center-stage for most organizations, with Online Transaction Processing (OLTP) systems i.e., relational databases driving operational transactions. As the need for analytical insights grew beyond transactional workloads, Online Analytical Processing (OLAP) systems built on data warehouses became a powerful tool for engineering teams

- Data Warehouse: Data warehouses rely heavily on predefined schemas (i.e., the data model and its logical organization of tables, attributes in columns, and elements stored in rows) consolidating multiple sources and enforcing consistent reporting
  - This structured approach enabled standardized analytics and faster insights but also imposed rigid constraints on data formatting and organization. As businesses increasingly need to handle semi-structured and unstructured data at scale, these limitations spurred the rise of flexible data-lake architectures

- **Data Lake:** Data Lakes support storage for large volumes of unstructured diverse data elements sourced from IoT sensor feeds, transactional logs, and other semi-structured or unstructured sources—using low-cost object repositories
  - These lakes are typically driven through a storage layer managing raw data in open file formats (e.g., Apache Parquet, Apache Avro, Optimized Row Columnar (ORC) formats), hosted in cloud-based or on-premises object storage
  - Data processing/query frameworks and engines (e.g., Hadoop, Apache Spark, Apache Flink) support the data pipeline based on the stored objects into a query-ready format for analytics tools (Tableau, Power BI) or self-service methods (SQL, Python and others)
  - However, data lakes devolved into "data swamps" as they grew unchecked—lacking catalogs, governance, and proper monitoring
  - Organizations often found that complex data pipelines are essential to extracting value from raw data, but their associated query processing (e.g., inserting or deleting records) was extremely slow, costly, and cumbersome over time. These challenges have prompted a shift toward "data lakehouse,' blending the flexibility of data lakes with structured advantages of data warehouses
- **Data Lakehouse:** Data Lakehouses add a structured 'table' layer on data lake architectures to enable effective management of data assets through cataloging and retrievability of data. Typically, some critical layers are typically incrementally included:
  - Table Layer: A format storing metadata attributes for files in the storage layer, just like tables in a database. These attributes consist of partitions, schema, file-level metadata that get ingested by query engines to provide easier access to data. This helps with adding organizational features like supporting schema evolution and "time-travel' rollbacks, ACID transactions, and query optimization
  - Catalog Layer: Adds business and data context to the underlying "tables' such as
  - Meta-store/Technical Catalog: Technical information like file path, schema organization, partition etc., that enables the computing layer to work with data (provided by Apache Iceberg as a REST enabled catalog or other managed options provided by lakehouse vendors like Snowflake Polaris)
  - Governance catalog: This is the secondary sub-layer for business context (data ownership, metadata, user access authorization and permissions, data sharing) through catalogs (generally enterprise vendors that span non-lakehouse assets or through a sub-layer offered by commercial vendors for more tighter integration)
- **Data Mesh:** A "data mesh' approach is gaining traction to simplify the business organization of data. Data meshes are not truly a technical evolution and can be treated more as an organizational shift making data assets (such as tables, warehouses, lakehouses) available to the organization in a structured fashion
  - Data meshes revolve around critical principles including domain-oriented data, self-service data infrastructure, federated governance, and most importantly treating "data-as-a-product'. A data-as-a-product philosophy is achieved by treating data as a product managed by individual teams with the enterprise being the "customer'. It covers building out the datasets, high-quality business and technical metadata, documented landing pages facilitating data exploration, pipeline / transformation logic, and usage statistics driving new "product features'



On the Horizon: A new paradigm

Now, with the emergence of Generative AI (GenAI), organizations are on the forefront of another paradigm shift. GenAI enables this shift by offering advanced large language models (LLMs) capable of inferring and generating content like human users. If the Big-Data era was characterized by the exponential volume, variety, and velocity of data, the GenAI era will be characterized by the data value, represented as an exponential growth in users, formats, and use-cases. From a people standpoint, data will expand beyond end-users and a limited set of service accounts (powering integrations and data pipelines) to cover machine identities, and more importantly agentic AI.

From a data standpoint, GenAI helps organizations to unlock a wide range of assets and move beyond structured and semi-structured data for business insights to encompass unstructured information such as textual documents, image files, and system logs. From a use-case standpoint, there is a rise in demand, especially from agentic AI seen both at scale (i.e., affecting multiple data elements across the enterprise), and at speed (i.e., a very heavy volume of diverse data requests). Organizations should therefore prepare to manage this new paradigm and incorporate considerations such as synthetic data generation, storage like vector databases and data formats like embeddings.

Bringing it together – an illustrative AWS-based enterprise data platform

AWS offers a host of data-oriented services that support engineering teams to spin up new assets within any architecture pattern in an easy, flexible, and scalable fashion. Product teams are typically responsible for their own technology stack ensuring interoperability with enterprise systems. Therefore, using native AWS services can help teams follow a repeatable blueprint with a consistent technical foundation that is well integrated, supporting core features at scale at higher performance.

Enterprise data platform - Component View

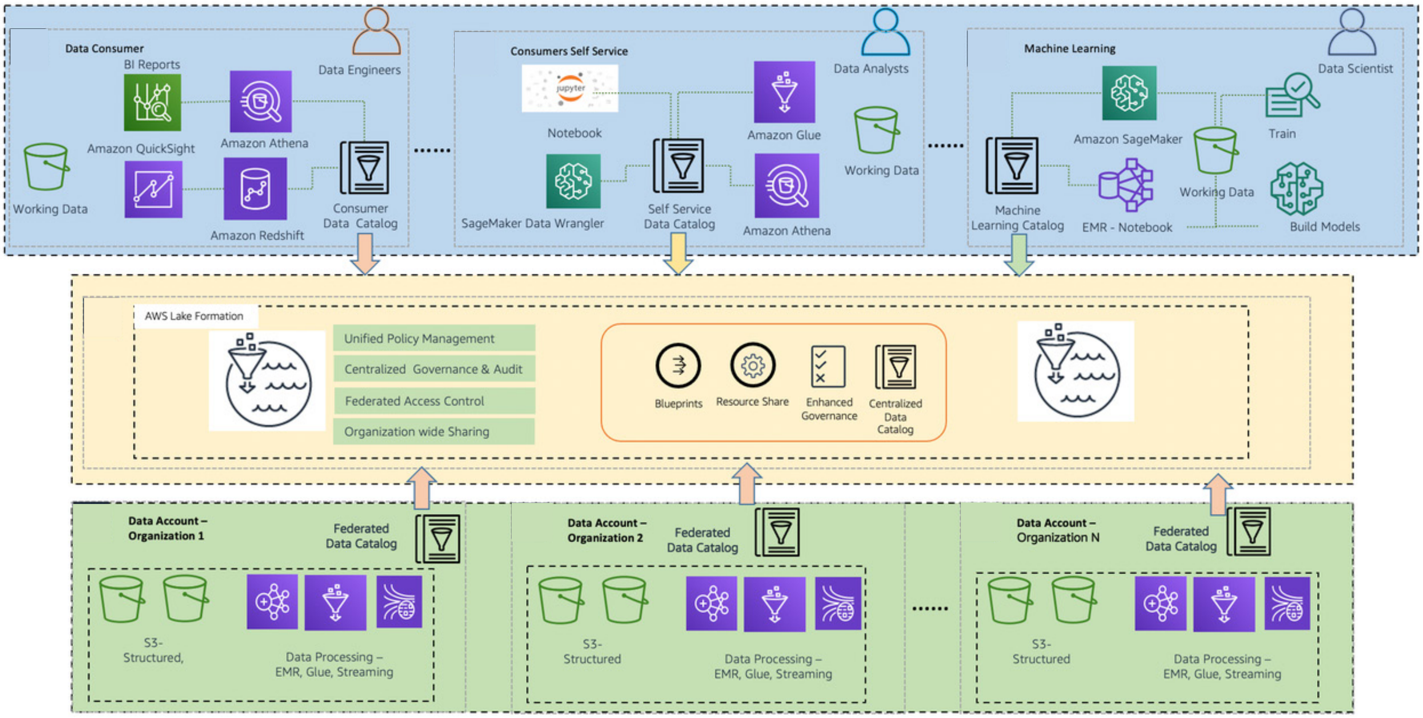
Enterprise Rights to Data Platform

Inputs		
Enterprise data management and governance • Data policy • Data classification • Data privacy • Data retention • Data security • Data quality • Data ownership • Data accessibility + Use cases + Regulatory compliance + Strategic alignment	**Governance and risk**	Access & control Data access & sharing Data governance Risk monitoring
	Operation	Feedback management Evaluation Cost tracking Talent onboarding & management* Performance metrics Bias detection (for ML use cases) Monitoring
	Data	Data Ingestion, preparation and labeling Data catalog management Data input and usage analysis
	Security	Encryption Sensitive data handling Data access control Threat management (deepfakes, attacks) Identity & access management Production (infrastructure, data, app) Threat detection & incident response

* Multi-tenant architecture

A common approach leveraging native services consists of the following structure

- Ingestion Layer: Data Extract-Transform-Load (ETL) through AWS Glue, AWS Elastic Map Reduce (EMR), Amazon Athena services
- Storage Layer: S3 buckets leveraging Apache Parquet; AWS S3 Tables; AWS Redshift
- Catalog Layer: AWS Glue Data Catalog
- Consumption Engines: Amazon SageMaker, Amazon Athena, AWS Redshift
- Governance Layer: SageMaker Unified Studio, Amazon Data Zone powered by AWS LakeFormation policy engine
- Security services: Identity and Access Management (IAM), Key Management Service (KMS)



Source: https://aws.amazon.com/blogs/big-data/design-a-data-mesh-architecture-using-aws-lake-formation-and-aws-glue/

THE SECURITY IMPERATIVE

Data Risks and their impact

Organizations are dependent on their data assets, to drive competitive advantages through data-driven insights. However, as organizations manage the data landscape, they should be aware and prepared to tackle a wide variety of data security risks like confidentiality (data breach by external actors and insider threats), availability (through system downtime, inadequate backups, ransomware events), and integrity (data tampering, misuse deviating from original purpose).

The impact of these risks to organizations is primarily a set of direct and indirect losses stemming from financial damage of a data loss event, regulatory actions including punitive fines, operational disruption, and brand damage especially loss of trust. Thus, it is imperative for organizations to manage the security of their data assets across the landscape.

Data Risks

Data Misuse
Threats caused due to violations of data use policy, agreements, standards, privacy laws and regulation

External Threats
Threats that originate outside the network from malicious actors that can target vulnerabilities to gain access to data.

Insider Threats
Threats caused by associates having access to systems and sensitive data.

Third Party
Threats that originate through Third party and/or vendors that handle restricted, confidential and/or internal data.

A Challenging Path Forward

Challenge: Empowering the Right Team

Within a typical defense-in-depth cybersecurity model, organizations plan and design capabilities around users, applications, infrastructure, and networks. These capabilities contribute heavily to mitigating data risks and achieving objectives like data confidentiality. So, organizations may elect the path of managing Data Security capabilities across multiple teams aligned to the security organization.

Common examples, for instance, include business team responsibility for encrypting data assets in a federated model, identity and access management teams taking on data access responsibilities, data governance teams performing data discovery and classification, security operations center (SOC) teams driving database activity monitoring triage and alerting, and many others. This leads to typically underfunded, fragmented, and often deprioritized capabilities.

Effective programs are typically observed to have a well-defined "Data Security" function with appropriate oversight and leaders, right-sized funding and defined charter. Typically, the capabilities in this delineated 'Data Security' function cover three core areas: understanding and contextualizing data (Data Discovery and Classification), protecting data based on its context (Data Protection), and monitoring/ enforcing usage (Monitoring and Enforcement).

Data Protection

Discover

Identify sensitive data stored across environments so that controls may be applied where they are most effective

- Data Discovery Scanning
- Cloud Access Security Broker (CASB)

Monitor/Enforce

Monitor the movement and usage of sensitive data to detect and/or prevent unauthorized or risky activity

- Data Loss Prevention (DLP)
- Cloud Access Security Broker (CASB)
- Database Activity Monitoring (DAM)

Classify

Classify files, emails, and structured data sets to allow the application of downstream capabilities based on sensitivity

- Data Classification and Labeling
- Data inventory

Protect

Govern access to sensitive data by making it readable to only authorized users, prioritized by data sensitivity

- Encryption, Tokenization, and Masking
- Key and Certificate Management (KCLM)
- Data Access Governance (DAG)

While enforcement and adoption of these capabilities might be shared across business teams, data governance teams, and other relevent cross-functional groups; having a dedicated Data Security function driving capability maturity in alignment with leading industry standards and efficient operating models will be beneficial for organizations.

Challenge: Technology complexities of the "Tri-Modal Landscape"
Traditional data security controls are not scalable across the tri-modal landscape. Organizations should therefore plan to accommodate data security for the tri-modal landscape, focusing on a combination of legacy approaches, current modern cloud-focused controls, and emerging guardrails for AI data security.

Challenge: The Security Mindset
Security teams could face a variety of organizational challenges and user culture impediments.

While the prevailing concern is around users circumventing security controls (through shared accounts, shadow copies, and workarounds), there are other factors that impede teams from establishing effective programs. Data security is often perceived as a Compliance oriented barrier instead of true enablers, leading to a 'Check-the-Box' afterthought. Teams also often have limited visibility and understanding of data amplified by mismanaged data sprawl. Finally, data security is not integrated with governance, privacy, and risks teams leading to siloed and ineffective controls.

For the emerging AI-driven data paradigm, an effective Data Security program safeguards data across the tri-modal landscape at scale, i.e., "the right principals accessing right data and using it in the right manner".

SAFEGUARDING YESTERDAY'S WALLED PREMISES

Historically, organizations established data platforms within their enterprise datacenters including raw data sources, databases and their associated infrastructure servers, engineering components like the Extract-Transform-Load (ETL) platforms, and consuming Business Intelligence (BI) applications (e.g., data visualization and analytics tools).

The approach:

A walled premises approach was developed for such data platforms focused around securing the data platform in isolation, depending on manual processes, enabled through dominantly native capabilities or a limited set of specialized tools. System administrators established a limited set of data egress channels via access to networked interfaces and resources.

The data security goal of "the right principals accessing right data and using it in the right manner" is achieved through the following measures:

- Right Principals: Platform-specific users validated through enterprise Identity and Access mechanisms such as Single-sign-on (SSO), multi-factor authentication (MFA), and privileged access for database administrators and other privileged roles
- Right Data: Data Discovery and Classification is performed within each component, using manual approaches (or enterprise tools deployed by mature teams). Data destruction processes are also in place at the component level, though they are rarely exercised
- Right Access: Data authorization is established for each component within the pipeline through role-based access controls (RBAC), such as database roles
- Right Manner: Activity monitoring and data loss prevention is primarily implemented at external egress boundaries, such as web, email, USB, and print channels, and not for individual components

Analytics and AI workloads

On-premises analytical and AI workloads are typically limited, especially due to the technological complexities and cost of hosting these workloads through static resources. However, in certain nascent scenarios, organizations may still need to support these workloads.

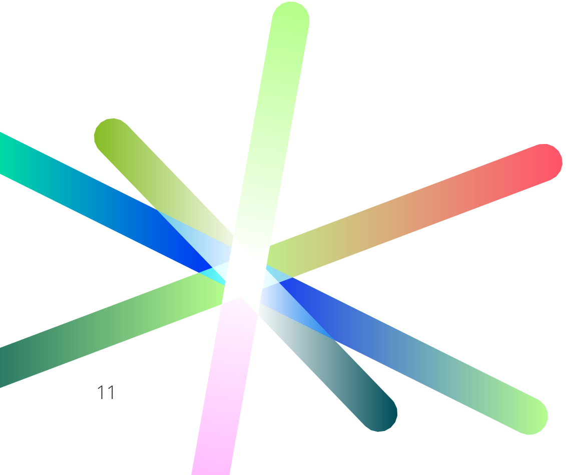
- Data Discovery and Classification: Manual processes with offline catalog tables, lineage tracked via project documentation and institutional knowledge; outsized risk of shadow copies hampering a centralized directory of sensitive data elements
- Data Protection: Limited encryption (mainly device and file-level encryption), role-based access controls to the platform with a lack of fine-grained authorization, limited data loss prevention that doesn't mitigate the risk of researchers and data scientists sharing large datasets through shadow IT channels
- Data Activity Monitoring: Some database activity monitoring via tools in mature organizations and activity logs within platforms (such as database audit tables), but a lack of centralized integrated views to flag suspicious activity

Understanding the limitations – "The Cracks in the Wall"

- Perimeter-less Data: With the rise of cloud-based storage and SaaS solutions powered by APIs, data is never static in a simple perimeter and flows through multiple egress points into a variety of connected infrastructure elements. In an isolated approach, teams miss out on applying guardrails uniformly (for example, ensuring encryption is applied at all physical storage drives). Additionally, integrated interfaces and data flows through downstream components often get overlooked unless a combined threat model is considered by each platform owner

- Overly Permissive: Many teams lean towards facilitating data science and business teams, thus enabling broad permissions for users within the component perimeter (i.e., an authenticated user can typically access broad sections of the data warehouse leading to risks in case of account compromise or insider threats). Additionally, guardrails established at individual component may not cover threats materializing over the pipeline. A good example is data access permissions. The BI tool has "full-trusted' access via service accounts to the underlying database. Fine-grained data permissions (column-level and row-level) are implemented within each individual component and granted to users as access roles by the administrator. If not done correctly at any platform, the entire dataset is vulnerable to data exploits and breaches

- Misaligned roles: Typically, engineering or technology teams manage guardrail implementation based on security requirements (or in certain cases, provision services supported by cybersecurity teams). Therefore, operating processes become an additional overhead for already stretched resources leading to inefficiencies. For example, the slow pace of access approvals (taking roughly weeks from request submission to provisioning) is prompting organizations to seek alternative approaches



The walled premises approach may still be needed, prominently critical on-premises systems such as mainframes or legacy enterprise systems that can't be migrated to the cloud. However, organizations should understand the limitations and plan appropriately.

SAFEGUARDING TODAY'S CLOUD DATA ASSETS

Modern data systems (cloud-hosted, vendor managed, or even on-premises modern data architectures) have matured with a set of data security capabilities to help organizations meet risk objectives. Organizations are moving towards cloud-hosted data platforms, building on the shared responsibility model to secure these platforms, enabled by a comprehensive set of services and capabilities such as policy-as-code guardrails to help drive a self-service, embedded approach to security.

The approach

Data security can still be considered as a "walled premises" approach for the modern cloud-hosted data platforms and pipelines. However, the critical difference is the establishment of a "virtual perimeter" that can be scaled to cover the underlying infrastructure and integrations (APIs, cross-platform tooling, third-party data flows), providing an integrated view of data along-its journey across enterprise systems, cloud provider infrastructure, and even vendor-hosted applications. This perimeter helps organizations tackle the challenges faced in legacy systems, for instance, tackling data flows outside enterprise boundaries, managing misaligned roles and overly permissive access through risk-driven guardrails, and rapid speed of deployments through a secure-by-design approach (such as Infrastructure-as-Code – IaC driven by services such as AWS CloudFormation and Terraform).

A good example is key management for encrypted data assets. Whereas traditionally, teams depended on native solutions (e.g., hardware encryption keys, key stores for "transparent data encryption' offered by database engines), security teams can enforce key management through AWS Key Management Service (KMS) abstracting the complex technicalities of key management. Encryption can be performed at service build with a key provisioned through code snippets within the IaC pipeline.

The data security goal of "the right people accessing right data in the right manner" is achieved through the following measures:

- Right People: Data principals (roles, users) managed and authenticated through cloud services, and operating with secure tokens, layered with enterprise Identity and Access mechanisms such as Single-sign-on (SSO), multi-factor authentication (MFA), and privileged access mechanisms
- Right Data: Data Discovery and Classification natively performed within the broader cloud service (e.g., AWS Macie, Amazon DataZone) and leveraging asset tags; data retention and destruction driven through storage policies and using tiered storage options
- Right Access: Data authorization is still established for each virtualized component within the pipeline through role-based access controls (RBAC), such as database roles
- Right Manner: Activity monitoring and data loss prevention controls are implemented at component boundaries to restrict improper usage, primarily external enterprise boundaries, such as web, email, USB, and print channels. Cloud access brokers (CASB) platforms also help enforce data exfiltration protection measures by monitoring traffic between networks

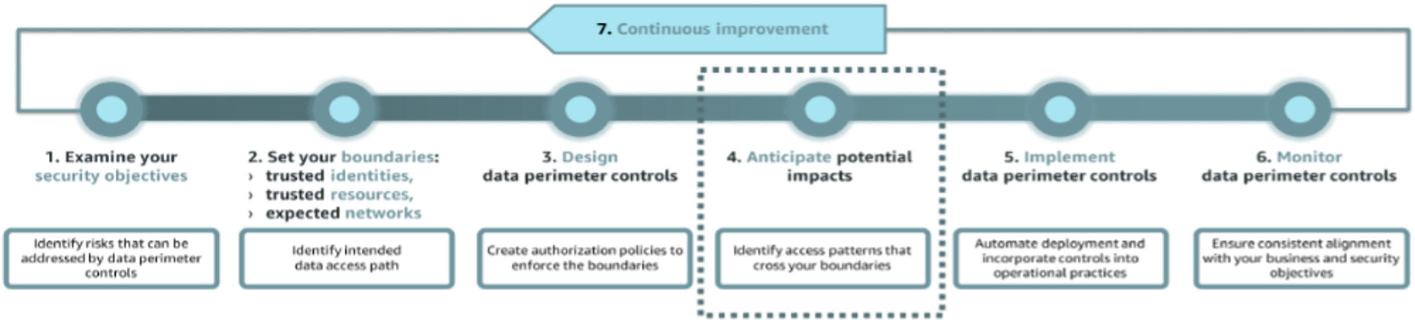
Analytics and AI workloads

Modern analytical and AI workloads are typically cloud-hosted leveraging object stores (such as Amazon Simple Storage Service, S3 buckets). Adaptive guardrails can be applied in a similar fashion across layers (such as ingestion, storage, consumption), leveraging cloud-native services

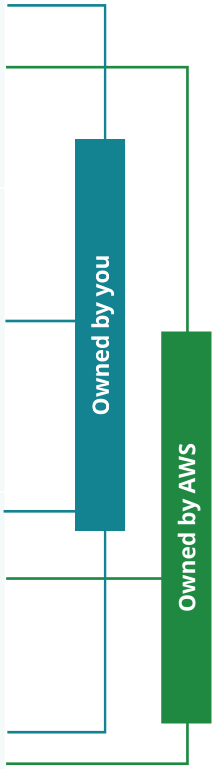
- Data Discovery and Classification: "Crawlers" i.e., scheduled scanners generally look at cloud-hosted storage (and in certain cases, leveraging specialized tools for SaaS or on-premises stores) and contextualize business data through pattern-based matching and apply resource tags
- Data Protection: Tag-based' resource policies generally drive automated protection capabilities such as access permissions depending on roles, sensitive data tokenization leveraging programmatic function. Data Loss prevention is focused on the cloud perimeter (accounts, resource groups, network policies) and management of egress data
- Data Activity Monitoring: Metadata events captured within cloud logs form the primary source of activity monitoring supported by specialized tools monitoring other capabilities like APIs in mature organizations

Bringing it together within the AWS environment

Organizations can set up “virtual’ data perimeters within the AWS environment and back it up with native cloud services (for Discovery, Classification, Encryption, Access Control, Data Loss Prevention) to protect cloud data assets. In this model, infrastructure elements are generally handled by AWS with “virtualized’ data layers handled by the organization.


7. Continuous improvement
1. Examine your security objectives
Identify risks that can be addressed by data perimeter controls
2. Set your boundaries:
› trusted identities,
› trusted resources,
› expected networks
Identify intended data access path
3. Design data perimeter controls
Create authorization policies to enforce the boundaries
4. Anticipate potential impacts
Identify access patterns that cross your boundaries
5. Implement data perimeter controls
Automate deployment and incorporate controls into operational practices
6. Monitor data perimeter controls
Ensure consistent alignment with your business and security objectives

Perimeter	Control Objective	Using	Primary IAM feature
Identity	Only trusted identities can access my resources	RCP	aws:PrincipalOrgID awe PrincipalOrgPaths aws: PrincipalAccount aws: PrincipalIsAWSService aws: SourceOrgID aws: SourceOrgPaths aws: SourceAccount
	Only trusted identities are allowed from my network	VPC endpoint policy	
Resource	My identities can access only trusted resources	SCP	aws.ResourceOrgID aws:ResourceOrgPaths aws:ResourceAccount
	Only trusted resources can be accessed from my network	VPC endpoint policy	
Network	My identities can access resources only from expected networks	SCP	aws:SourceIp aws:SourceVpc/aws:SourceVpce aws:ViaAWSService
	My resources can only be accessed from expected networks	RCP	aws:SourceIp aws:SourceVpc/aws:SourceVpce aws:ViaAWSService aws:PrincipalisAWSService



- Service control policies (SCPs): Enforce a permission boundary for data principals (i.e., roles or users set up using AWS IAM service) to access only trusted resources from expected networks. Leverage condition keys (aws:SourceIp, aws:SourceVpc, aws:SourceVpce) to establish these boundaries
- Resource control policies (RCPs): Enforce a permission boundary for resources (e.g., data assets, cloud workloads) to be accessed only by trusted data principals and from expected networks. Leverage condition keys (aws:PrincipalOrgID, aws:PrincipalOrgPaths, aws:PrincipalAccount)
- VPC endpoint policies: Enforce a permission boundary for the organization network (i.e., allowed endpoints and network infrastructure) to be accessed only by trusted identities and for trusted resources. Leverage condition keys (aws:SourceOrgID, aws:SourceOrgPaths, aws:SourceAccount)

Essentially, through SCPs, organizations can limit users from accessing resources outside the network (potentially malicious and non-trusted data), and via RCPs and VPC policies, external users can be restricted from accessing internal resources (potentially untrusted and malicious connections). In addition, to the "virtual' perimeters established through these policies, organizations can leverage the services like Right People (AWS IAM), Right data (Amazon Macie), Right Manner (Amazon Guard Duty, AWS CloudTrail, AWS Config)

The limitations – "Adaptive but not Optimized"

- Engineering Complexity: Given that business teams increasingly leverage cloud resources in an on-demand and ephemeral manner, security controls must be designed not as one-off implementations but as part of a continuous, proactive monitoring strategy that evolves alongside business needs. Traditional security practices must also be augmented with expertise in modern technologies such as Infrastructure as Code, cloud architecture, and DevSecOps practices. Many organizations lack the right skillset and training to handle cloud security challenges effectively
- Ineffective policy governance: Organizations don't establish robust policies and guardrails even with available services due to resource constraints, self-service workarounds, and blind spots. For instance, policy mismanagement and granularity is a major concern especially in cases of AWS Identity and Access Management (IAM) Roles and other risk-based rules like Data Loss Prevention, and activity alerts
- Shared Fate: Many organizations often rely too heavily on the cloud service provider and don't understand that shared responsibilities also lead to a shared fate for security. Organizations don't understand their roles in the defense-in-depth model and take appropriate actions. For instance, organizations operating under strict data movement regulations should ensure data storage and processing adhere to designated geographical boundaries and comply with regulatory mandates

Adaptive guardrails are robust security measures covering current security threats and vulnerabilities within cloud environments; "if done right"; otherwise, it often leads to largely symbolic security measures adding complexity and overhead.

SAFEGUARDING DATA IN TOMORROW'S AI AGE

What is changing?

Typically, organizations must plan for three major areas of change and its associated impact compared to their traditional application / data landscape, due to the rise of GenAI.

Changing Data Consumption: Due to GenAI, the value, criticality, and protection approach to data is changing e.g., rise of new classes like synthetic data, inferred data, vector embeddings; changing value to attackers e.g., unstructured documents are more easily parsed by attackers now with GenAI and becoming attractive targets. Compared to current platforms, GenAI (especially agentic AI workflows) will also lead to a change in consumption patterns including speed, scale, complexity, and ephemerality e.g., compared to today's human users and limited non-human identities, enterprises might see a consumption pattern of thousands of new virtual "user-like agents" spinning up on demand in workflows, and getting decommissioned post use within a few minutes.

Changing attack surface: Attackers are identifying new vulnerabilities in the AI landscape, e.g., prompt injection attacks, data poisoning, supply-chain attacks embedded via tool functionalities powering agents. Mitigation is further complicated by the autonomous behavior of GenAI based agents and their probabilistic outputs. A critical example is the Confused Deputy issue where an autonomous agent / data principal may get elevated access even if not explicitly allowed by interacting with a more privileged entity integrated within the pipeline. This issue is particularly scaled within the modern paradigm, enabled by agentic AI operating with complex data lake house environments (difficult to trace sensitive data assets and associated access permissions).

The enterprise security response to AI changes

Enterprise security teams have an inertia roadblock, finding it hard to react nimbly to the changing landscape. Data security teams don't have the appropriate understanding of the emerging controls in the AI age and lean towards treating them like the present landscape (i.e., mandating controls as a cloud data store and passing the burden on data teams that are ill-equipped to manage risk). Teams with more awareness or being more risk-averse, however electing the other end of the spectrum, restricting capabilities until security guardrails are available (observed in GenAI pipelines such as limiting data ingested into a landing zone), hampering business enablement.

Executive decision-makers have questions about identifying the specific impact unique to AI adoption as well as their accountabilities. Organizations have been tackling data risks in modern cloud platforms and SaaS applications, focusing on the shared responsibilities and associated security guardrails. Therefore, in today's landscape where enterprises are still largely dependent on other providers for AI models and associated components, some executives might consider AI applications equivalent to vendor-managed systems from a risk standpoint and consider themselves absolved from critical risks (such as training data-set leakage risks, model bias and hallucination risks). This can also naturally lead to questions about the perceived incremental risk of AI applications to the enterprise, furthering an inertial block to AI security and risk management.

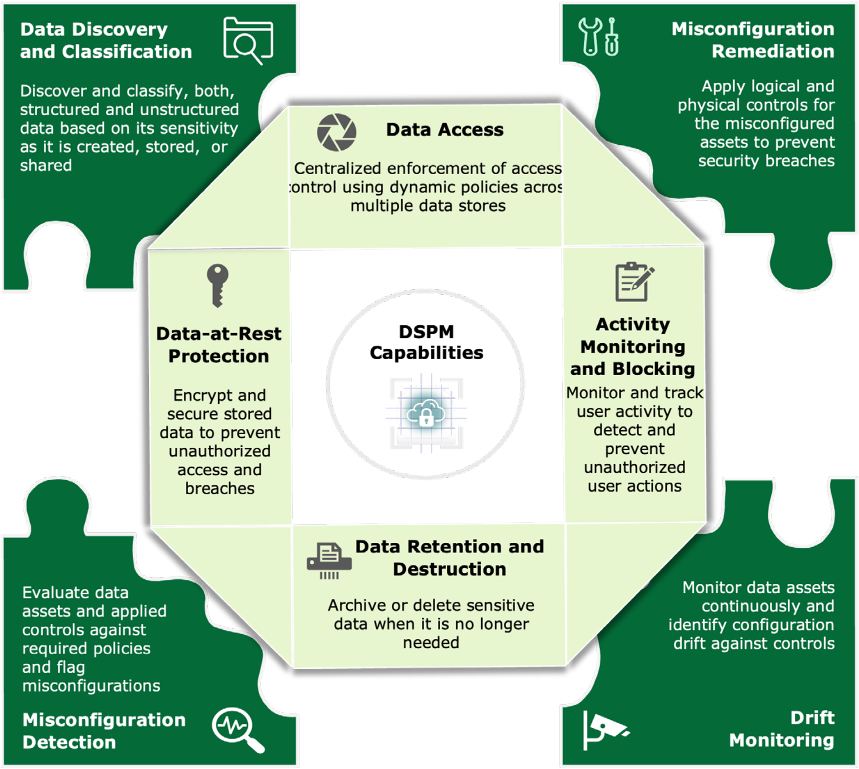
Organizations need to understand, just like the shared responsibility model, popularized within SaaS environments, they share fate with providers i.e., accountability in safeguarding data for their customers and stakeholders. In addition, while data risks and associated impact can be like those faced by current cloud / SaaS landscape; there is a need to adapt to the changes in the AI threat landscape and develop new capabilities needed for effective security.

The approach

Conceptually, the current "zero-trust" approach used for Cloud data security remains relevent to tackle AI data risks through continuous monitoring and validation, segmented micro-perimeters, and a least privilege approach for user access permissions. Organizations, however, need to develop new capabilities to tackle the changes in data consumption and the attack surface in the AI age. This is effectively powered at the data plane by a combination of two paradigms – "Data Security Posture Management (DSPM) coupled with a robust Data Authorization framework."

Consider the security goal for "the right people accessing right data in the right manner".

- The right data in the right manner: A "DSPM" platform is a modern solution comprising of multiple capabilities that can help discover, classify, contextualize, label / tag critical data elements across structured and unstructured files. Through a posture management approach, DSPM solutions can identify issues (e.g., sensitive data in unapproved storage, or production data in test environments), apply remedial actions (for instance, remove overly permissive access, or even delete data), and track issue remediation through streamlined workflows (integrating with corporate data catalogs and ticketing systems). DSPM platforms can help organizations implement proactive continuous monitoring aimed at establishing a set of foundational security measures
- The Right People and the Right Access: User validation and data access can be implemented with a robust data authorization approach. This approach leans on establishing access policies and enforcing coarse-to-fine-grained entitlements via platform-specific access roles and permissions. This helps organizations tackle the challenges of provisioning just-in-time access to enable AI applications and emerging analytics paradigm, but at the same time, confirm adequate security measures are in place

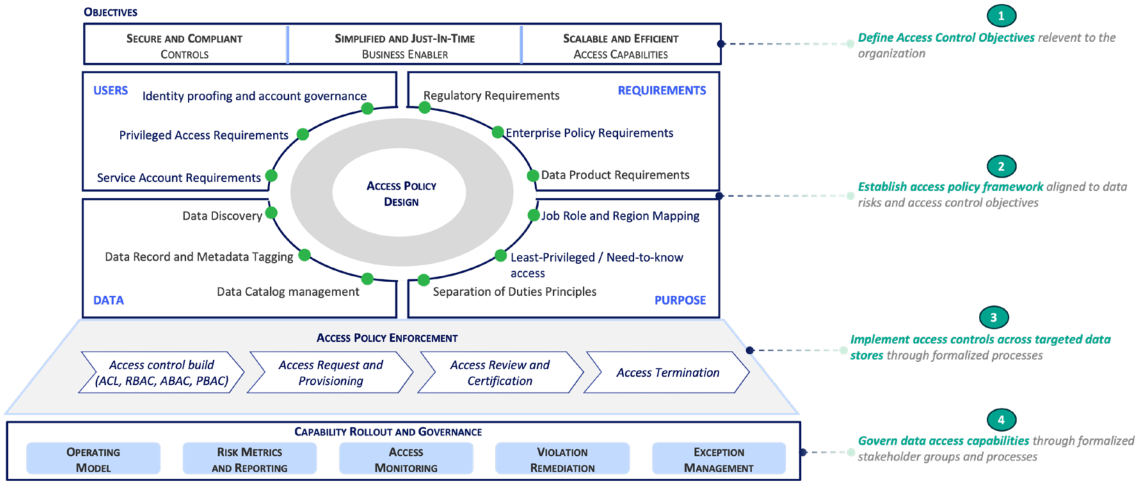

Data Discovery and Classification
Discover and classify, both, structured and unstructured data based on its sensitivity as it is created, stored, or shared
Misconfiguration Remediation
Apply logical and physical controls for the misconfigured assets to prevent security breaches
Data Access
Centralized enforcement of access :ontrol using dynamic policies acros multiple data stores
Data-at-Rest Protection
Encrypt and secure stored data to prevent unauthorized access and breaches
DSPM Capabilities
Activity Monitoring and Blocking
Monitor and track user activity to detect and prevent unauthorized user actions
Data Retention and Destruction
Archive or delete sensitive data when it is no longer needed
Evaluate data assets and applied controls against required policies and flag misconfigurations
Misconfiguration Detection
Monitor data assets continuously and identify configuration drift against controls
Drift Monitoring

Data authorization can help tackle challenges through establishing a scalable guardrail implemented at the data consumption layer, thus handling complex capabilities aligned to the requirements for data lake houses (technical implementations available to enforce data access at the catalog layer), data-as-a-product' approaches driven by data product owners (easier mechanisms to design data access policies), and GenAI application nuances (enforcement layer for new components like agentic AI with easier "just-in-time' policy-as-code approaches, visibility over complex access hierarchies solving confused deputy problem).

Securing Emerging Analytics and AI workloads

Organizations are shifting as part of their data evolution into cloud-hosted data lakehouses managing structured, unstructured, and synthetic data, and empowering AI applications and agents with broad autonomy.

- Data Discovery and Classification:
  - The catalog layer is the critical element of the data lake house supporting data management as compared to a data lake. Modern lake houses (e.g., powered by Apache Iceberg open table formats) offer data discovery, tagging, cataloging, and contextualization through native approaches. This can be augmented through services and tools (e.g., AWS Glue Data Catalog) to integrate controls such as data lineage, ownership and accountability within the business hierarchies (for e.g., data stewards associated with the product), and enrich discoverability for business users through a "marketplace'
  - Synthetic data classes need additional capabilities for data discovery, classification, and labelling in line with regulatory requirements (such as labeling AI-generated images)
- Data Protection:
  - Data encryption can be applied within lake houses through native capabilities typically or in conjunction with enterprise tools
  - Data Loss prevention is typically applied at lake house boundaries focusing on system integrations, downstream consuming applications, and in conjunction with enterprise tools
  - Data access is a critical security measure as outlined through the Data Authorization framework
  - Data destruction must be executed at the storage layer, especially since table formats are based on certain engineering considerations (including snapshots, time-travel)
- Data Activity Monitoring:
  - Lake houses offer native logging capabilities supported by specialized tools monitoring other capabilities like APIs in mature organizations


OBJECTIVES
SECURE AND COMPLIANT CONTROLS
SIMPLIFIED AND JUST-IN-TIME BUSINESS ENABLER
SCALABLE AND EFFICIENT ACCESS CAPABILITIES
USERS
Identity proofing and account governance
Privileged Access Requirements
Service Account Requirements
REQUIREMENTS
Regulatory Requirements
Enterprise Policy Requirements
Data Product Requirements
ACCESS POLICY DESIGN
Data Discovery
Data Record and Metadata Tagging
Data Catalog management
DATA
Job Role and Region Mapping
Least-Privileged / Need-to-know access
Separation of Duties Principles
PURPOSE
ACCESS POLICY ENFORCEMENT
Access control build (ACL, RBAC, ABAC, PBAC)
Access Request and Provisioning
Access Review and Certification
Access Termination
CAPABILITY ROLLOUT AND GOVERNANCE
OPERATING MODEL
RISK METRICS AND REPORTING
ACCESS MONITORING
VIOLATION REMEDIATION
EXCEPTION MANAGEMENT
1
Define Access Control Objectives relevent to the organization
2
Establish access policy framework aligned to data risks and access control objectives
3
Implement access controls across targeted data stores through formalized processes
4
Govern data access capabilities through formalized stakeholder groups and processes

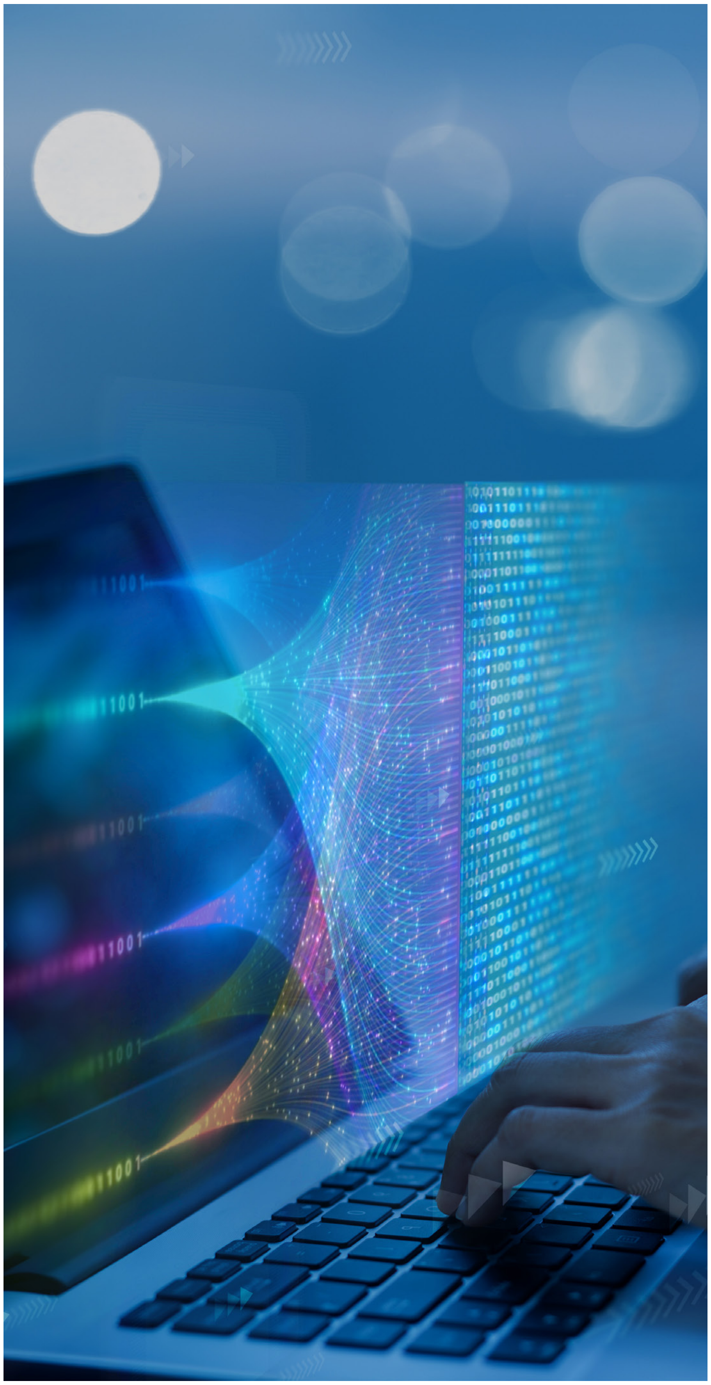
- Access Enforcement Model:
  - For the data lake house, there are two modes of data access – either through the catalog layer (e.g., Apache Iceberg) enforcing authorization on each request, or a query engine (for instance, data processing tools or consuming applications) getting a full-access service account and enforcing user-level authorization prior to final output
  - Typically, it is recommended to have a hybrid model of query engines for trusted connections, and controlled catalog-level access for untrusted or time-bound connections. This helps balance the best of both approaches by minimizing risk and enabling business users
- Policy enforcement
  - Access policies can be enforced with decision points (validate requests based on user attributes and permissions requested) and enforcement points (executes grants / deny based on policy decision)
  - Coarse authorization can be enabled (i.e., table / namespace / view and other securable objects) with a combined policy decision and enforcement point via the technical catalog (i.e., the metadata store such as the Apache Iceberg REST catalog with external permission management). The catalog layer helps establish a centralized decision point (one-set of permissions on Tables to be shared across many tools or on multiple storage systems)
  - Credentials can be provided for data storage elements via the lake house technical catalog, instead of applying it at storage layers (e.g., S3 buckets). Authorization shouldn't be applied to the storage layer as it will be challenging to replicate and maintain across flexible scalable stores such as S3 buckets, Amazon Redshift, and others)
- General Authorization Considerations
  - Enforce a "deny by default' approach to sensitive data categories (that may lead to data risks), and enable access limitations such as time-bound access or access recertifications for risky users (especially insider threats and those with broad access)
  - For the data-as-a-product approach, enable data stewards to establish access systems based on their specific architectures including "a marketplace' authorization layer to facilitate self-service
  - For data-as-a-product, empower data stewards to establish additional access policies, but not overrule enterprise-level restrictions without proper exceptions, in line with their data products and regulatory restrictions
  - Enrich user profiles with strong attributes and drive access approvals in a risk-based fashion e.g., on user attestation of secure handling of risky datasets or self-service access for low-risk data
- Coarse Authorization Considerations
  - Establish permission management tied to business hierarchies and attributes (through an integrated data governance catalog or alternative methods offered by the metadata store/Iceberg technical catalogs)

  - Leverage Role-based Access (RBAC) to commence by establishing a few key roles grouping different permissions and assigning these roles to users (via systems such as Active Directory groups), For instance, within AWS, ActiveDirectory can be integrated through the Identity Center or via Managed AD services, or via bringing in an identity provider of the enterprise's choice
  - Leverage Relationship-based Access (ReBAC) to scale up (extending Role-based access to natural relationships, for instance, the owner of a folder has access to each file in the folder). While ReBAC is still maturing, there are implementation methods using open-source formats like OpenFGA integrated in a few Iceberg catalogs (like Lakekeeper)
  - Leverage Attribute-based access (ABAC) to scale up further tackling complicated use-cases and user hierarchies including extremely segmented access (e.g., users in a certain division, role, and tenure can access sensitive data)
- Fine Grained Authorization
  - Catalogs in the Data Lakehouse generally align towards securable object authorization (e.g., tables/views)
  - Establishing row-level or column-level security is generally handled by the consumption layers (based on authorization controls and permission management such as ABAC)Bringing it together within the AWS environment

AWS offers a set of services to secure and manage data in the AI age (GenAI applications, agentic AI, and other applications).

- AWS Glue Data Catalog: Glue Catalog is the metadata repository that keeps track of tables, schemas, and data locations within a data lake house framework, including those built on top of Apache Iceberg. By standardizing metadata, it centralizes knowledge of data assets to ensure consistent interpretation and governance. From a security standpoint, the Data Catalog works IAM policies to manage which users and services can read or update the metadata. By integrating with other AWS services and Lake Formation, Glue Data Catalog helps ensure that only authorized entities can modify crucial metadata or view sensitive schema details. This can help prevent tampering or unauthorized data access. Additionally, Glue Data Catalog supports resource-level permissions, enabling administrators to grant or revoke privileges on specific Glue Data Catalog objects (databases, tables, etc.), thus improving fine-grained access controls
- AWS Lake Formation: Lake Formation provides a centralized mechanism to set up data lakes securely and manage data access across complex organizational settings. For a data lake house based on Apache Iceberg, Lake Formation offers granular permissions aligned with tables, columns, and underlying storage infrastructure. This fine-grained authorization is facilitated by a unified policy framework that uses IAM-based roles and Lake Formation permissions, bridging the gap between raw storage (e.g., S3) and higher-level analytics services (e.g., Athena, EMR). Administrators can define explicit data governance rules—such as column-level restrictions or row-level filtering—to confirm that sensitive data is accessible only to authorized users or groups. Lake Formation's tight integration with Glue Data Catalog means that security policies are consistently enforced, offering a single point of control for classification, encryption, and auditing

- Amazon S3 Tables: Amazon S3 Tables is a fully managed, Iceberg-native storage layer that lets you land, organize, and query high-volume tabular data—such as purchase transactions, IoT telemetry, and ad impressions—at S3 prices but with data-warehouse speed: up to 3× faster queries and 10× more concurrent transactions than self-managed alternatives. Each purpose-built "table bucket" functions as an analytics warehouse inside your chosen AWS Region, automatically optimizing Parquet files for cost and performance while preserving S3's 11-nines durability. Within a bucket, namespaces group related to Iceberg tables, simplifying fine-grained IAM policies, and each table benefits from continuous, automated maintenance (compaction, snapshot pruning, orphan cleanup) plus its own S3 API endpoint for ACID-compliant reads and writes
- Amazon DataZone: Datazone focuses on helping organizations discover, govern, and share data on a scale across various teams. In the context of a data lake house built on top of Apache Iceberg, Datazone extends governance coverage by automatically cataloging data sets, applying metadata tagging, and ensuring consistent access policies across domains. DataZone's domain-based structure allows different business units or functional teams to enforce their own security and compliance needs within specific boundaries, while still enabling controlled data sharing across the enterprise. Through integration with IAM and Lake Formation, DataZone ensures that only entitled users can view, request, or utilize data assets. It also offers a controlled onboarding workflow, where data owners can review requests and apply context-specific policies, further refining who can access, manipulate, or publish data in the lake house environment
- Integrated tools: Apache Iceberg REST catalog / metastores supported by Glue Data Catalog, Business Catalogs (enterprise-level) with AWS integrations (for instance, tools supporting AWS consumption services such as Athena, Redshift, Amazon Bedrock, SageMaker)



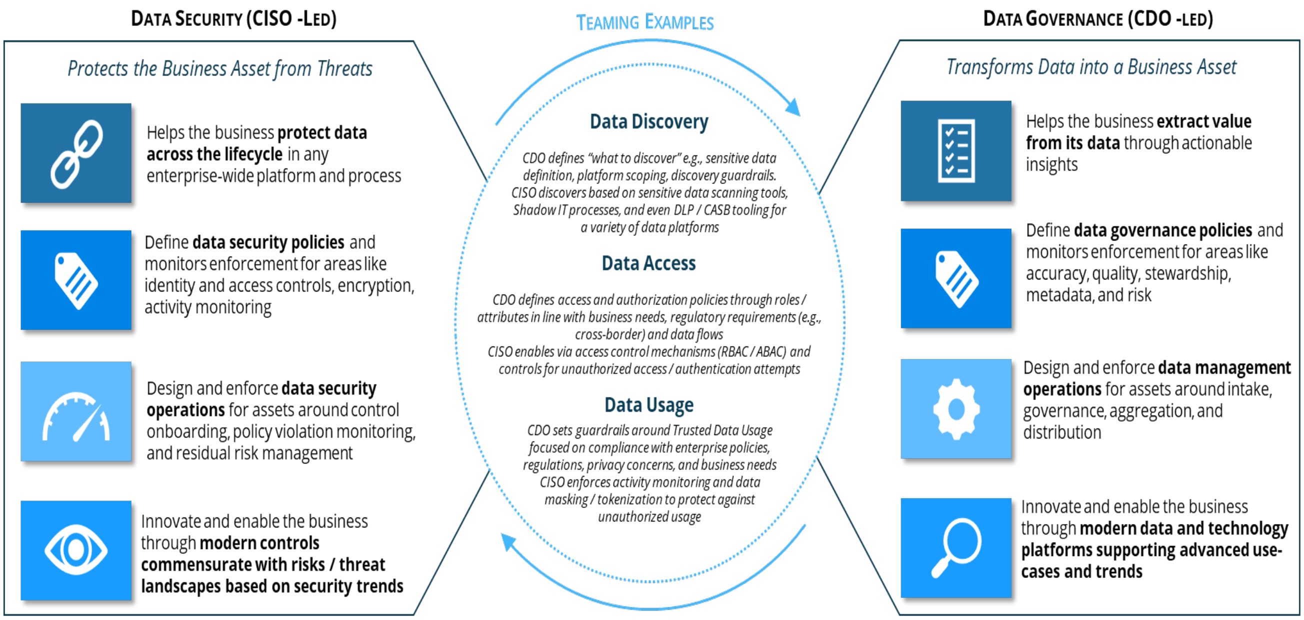
Safeguarding data assets in the AI age is evolving, with new capabilities emerging at a fast pace. Organizations must approach safeguards in a right-sized fashion focusing on no-regret capabilities well suited for the changing AI attack surfaces and threats.

BRIDGING THE CULTURE GAP: ALIGNING PEOPLE AND SECURITY

In addition to the technical challenges dealing with the tri-modal landscape, data security teams also must deal with a culture challenge. This can lead to business teams prioritizing data usage over security leading to lack of trust, transparency, and compliance. Users also lean towards non-secure workarounds such as creation of shadow data-stores outside security controls, overly permissive access policies to enable broader data usage, and limited engagement with security teams.

Building a security culture

- Change the vision
  - Treat Data Security as an enabler rather than a "Check-the-box' afterthought (e.g., simplifying developer experience, increasing production stability, and reducing risks in an embedded transparent manner)
  - Manage the shift from governance to enablement, driving a transition from policy enforcers to operators to enablers, delivering data services for strategic business initiatives
- Create a shared understanding
  - Coordinate with the business for integrated solutions that contribute to business initiatives and security goals by design
  - Integrate with cross-functional teams, such as Data governance, privacy, risk management, to ensure well-designed controls and reducing barriers
  - Enable end-user adoption and awareness through training and communications for effective change management
- Develop capabilities that enable
  - Select an approach that aligns with and enhances existing processes minimizing disruption during implementation
  - "Lighten the load" by deploying capabilities that help increase visibility, automation, and protection to meet requirements, moving away from restrictive solutions that are applied in a blanket fashion which increase security but hamper the business
  - Conduct pilots to identify opportunities to refine the enablement approach, configurations or awareness materials, gain support and momentum for enterprise deployment



DATA SECURITY (CISO -LED)

Protects the Business Asset from Threats

Helps the business **protect data across the lifecycle** in any enterprise-wide platform and process

Define **data security policies** and monitors enforcement for areas like identity and access controls, encryption, activity monitoring

Design and enforce **data security operations** for assets around control onboarding, policy violation monitoring, and residual risk management

Innovate and enable the business through **modern controls commensurate with risks / threat landscapes based on security trends**

TEAMING EXAMPLES

Data Discovery

CDO defines "what to discover" e.g., sensitive data definition, platform scoping, discovery guardrails.
CISO discovers based on sensitive data scanning tools, Shadow IT processes, and even DLP / CASB tooling for a variety of data platforms

Data Access

CDO defines access and authorization policies through roles / attributes in line with business needs, regulatory requirements (e.g., cross-border) and data flows
CISO enables via access control mechanisms (RBAC / ABAC) and controls for unauthorized access / authentication attempts

Data Usage

CDO sets guardrails around Trusted Data Usage focused on compliance with enterprise policies, regulations, privacy concerns, and business needs
CISO enforces activity monitoring and data masking / tokenization to protect against unauthorized usage

DATA GOVERNANCE (CDO -LED)

Transforms Data into a Business Asset

Helps the business **extract value from its data** through actionable insights

Define **data governance policies** and monitors enforcement for areas like accuracy, quality, stewardship, metadata, and risk

Design and enforce **data management operations** for assets around intake, governance, aggregation, and distribution

Innovate and enable the business through **modern data and technology platforms supporting advanced use-cases and trends**

CONCLUSION: THE JOURNEY AHEAD

In conclusion, the "Modern Data Paradigm" underscores the imperative of adopting a risk-centric framework that aligns governance principles with federated operating models and AI-driven consumption. Enterprises should consider implementing flexible controls that uphold trust without stifling innovation. Ultimately, this whitepaper is intended to help equip Chief Data Officers, governance and security teams, enterprise architects, and platform engineers with tools to confidently safeguard data and drive value in an ever-evolving digital landscape.

An illustrative prioritization helps enterprises follow a "crawl-walk-run' approach scaling controls in phases, thus dealing with change management and stakeholder buy-in through minimum viable proof-of-concept.

- Initialize the core foundations and stop immediate threats (Crawl)
  - Formalize data expectations aligned with the organization's risk appetite to established centralized guidelines, practices, and policies
  - Establish a well-defined operating model outlining roles and responsibilities across the enterprise for managing data risks for the modern data paradigm
  - Execute "no-regret" controls around the "trimodal data landscape"
    - Focus on critical on-premises systems and establish immediate guardrails
    - Identify a holistic approach for cloud-hosted systems (both cloud providers and SaaS), and establish immediate "quick-wins" through native services
    - Work with early adopters for emerging AI and analytics workloads, and implement coarse authorization through a technical catalog (right access) and data encryption for sensitive data
  - Establish shared capabilities / controls (across both enterprise data landscape and data lake house) on Identity providers (right people / users), service workflows (right time), activity monitoring (right manner)
  - Empower data stewards to tackle data-as-a-product approaches by building a foundational business catalog (centralized guardrails around data definitions, metadata and data quality guidelines, data ownership and accountability)
- Expand to foundational services across the data landscape (Walk)
  - Execute immediate controls around the enterprise data landscape (this helps organizations mitigate obvious threats at a storage layer)
    - Focus on right access (define access policies, establish foundational RBAC access enforcement, and encrypt sensitive data)
    - Focus on right data (Data Discovery and Contextualization using DSPM)
    - Focus on data loss prevention around enterprise boundaries (right manner)
  - Execute advanced controls for the enterprise data landscape and lakehouse (fine-grained authorization for row and column level security, business catalog driven permissions enforcement)
  - Empower data stewards with advanced business catalogs and orchestration tools to drive control rollouts consistently across data products
  - Establish assurance techniques and failure handling mechanisms for meeting stakeholder requirements (such as auditors, regulators, customers and partners)
    - Assurance is built through regular periodic assessments around controls and residual risks. Failure handling includes issue and exception management from assessment observations. A combination of these mechanisms helps stakeholders understand the overall security posture and awareness about any open identified issues and residual risks
- Scale and Mature (Run)
  - Integrate a cohesive single-pane-of-glass across data lakehouse and the enterprise data landscape (legacy) as governance tools mature from their current nascent stage
  - Prepare for Generative AI consumption use-cases especially as currently nascent tools and approaches mature to a production-ready manner

AUTHORS

LAXMAN TATHIREDDY

Deloitte & Touche LLP

Principal, Cyber Risk Services

Email: ltathireddy@deloitte.com

GREGORY, KOSZORUS

Deloitte & Touche LLP

Principal, Cyber Risk Services

Email: gkoszorus@deloitte.com

ADITYA KANITKAR

Deloitte & Touche LLP

Senior Manager, Cyber Risk Services

Email: akanitkar@deloitte.com

SPECIAL THANKS

KEITH HODO

Amazon Web Services

Partner Solutions Architect

Email: hodok@amazon.com

TANNEASHA GORDON

Deloitte & Touche LLP

Principal

Email: tagordon@deloitte.com

JUSTIN ROWE

Deloitte & Touche LLP

Partner Solutions Architect

Email: jurowe@deloitte.com

ANKIT SHRIVASTAVA

Deloitte & Touche LLP

Manager, Cyber

Email: ankshrivastava@deloitte.com

BARATHI KRISHNAMURTHY

Deloitte & Touche LLP

Manager, Cyber Risk Services

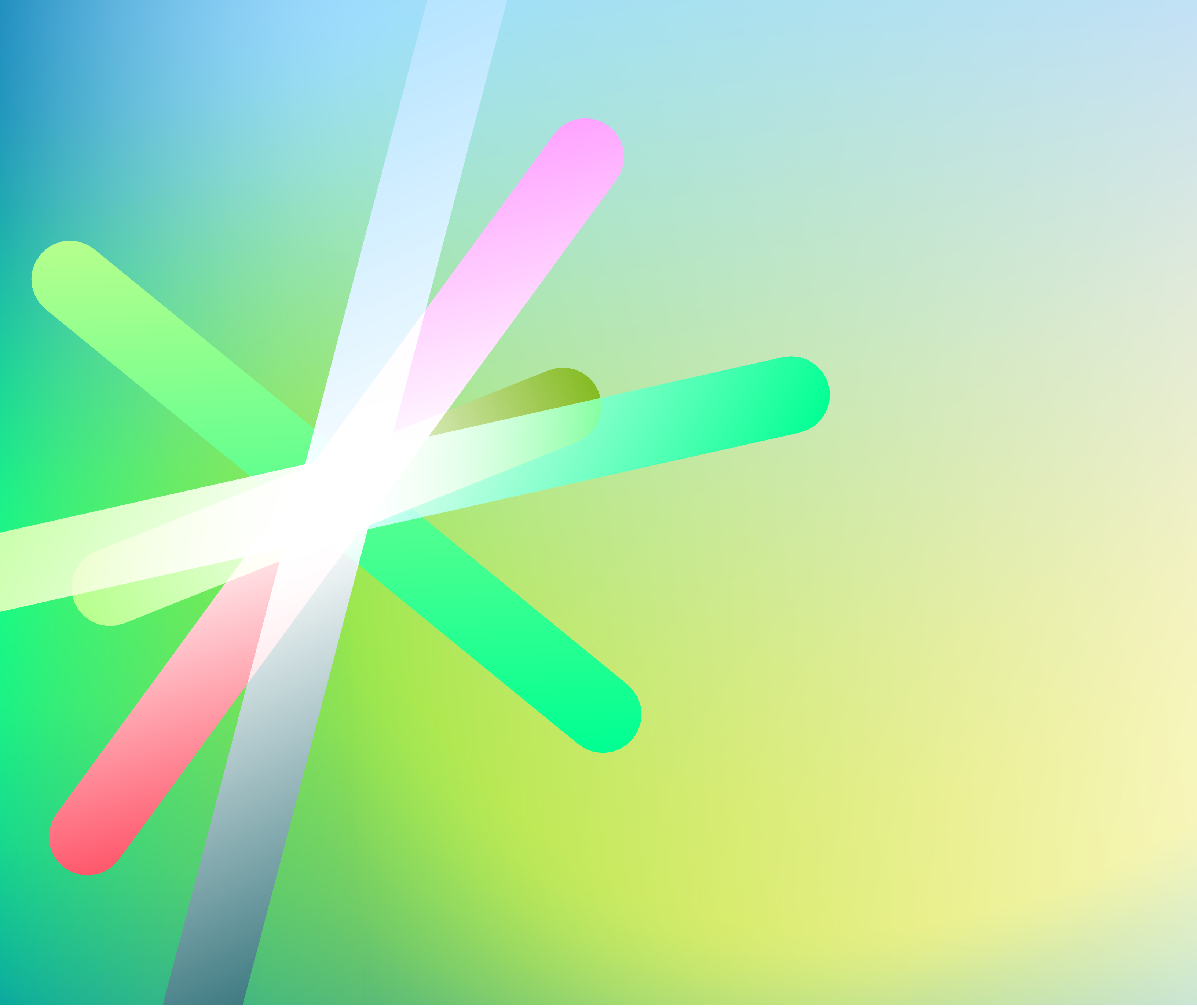
Email: bakrishnamurthy@deloitte.com

ANDREW BESLEY

Deloitte & Touche LLP

Consultant, Cyber

Email: abesley@deloitte.com



This document contains general information only and Deloitte and AWS are not, by means of this document, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This document is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte and AWS shall not be responsible for any loss sustained by any person who relies on this document.

All product names mentioned in this document are the trademarks or registered trademarks of their respective owners and are mentioned for identification purposes only. Deloitte is not responsible for the functionality or technology related to the vendor or other systems or technologies as defined in this document.

As used in this document, "Deloitte" means Deloitte & Touche LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

Designed by CoRe Creative Services. RITM2244833