WHITE PAPER

Responsible AI: How teams move faster and safer with Rubrik DSPM

Table of Contents

INTRODUCTION

> **“Data is the new oil,”**
>
> **– British mathematician Clive Humby, 2006**

We are in a new global AI race to unlock new, knowledge-based capabilities at a scale never before seen. Companies like OpenAI have proven that early movers reap outsized rewards, so speed is of the essence. The fuel for this new AI race? Data.

Whether a business’ chosen AI strategy is to use a vendor “co-pilot,” retrieval augmented generation (RAG), or fine-tuned LLMs; large quantities of a company’s proprietary information will be transiting the cloud, becoming broadly accessible, and potentially embedding into models that can’t be controlled.

In response, many CISOs will start by securing and governing the AI project team’s data and development infrastructure. This is a critical task given that development efforts and large data pipelines for Generative AI often span clouds, third-party transformers, LLMs, etc.

However, there is a far more business critical role for CISOs to play in accelerating and de-risking AI projects. In Rubrik’s conversations across thousands of IT teams, the leading obstacle for AI projects is controlling sensitive data and information in pipelines before it gets into AI. The goal is to prevent embarrassing or harmful data leakage from AI assistants.

CISOs who provide solutions that enable mass quantities of data to be deployed into AI without the risk of sensitive data or regulated data leakage, will be perceived as adding value vs. slowing projects down!

Lastly, CIOs and CISOs need to invest in future proof technologies for Responsible AI. These solutions will need to work quickly and cost-effectively across large data sets to identify emerging or possibly subjective data risks. These can include intellectual property, bias, culturally sensitive information, etc.

WHY DATA SECURITY FOR FASTER AI INNOVATION?

At Rubrik, our goal is to secure the world’s data. This takes the form of protecting you from data loss, cyber attacks, or even access by unintended parties and AI.

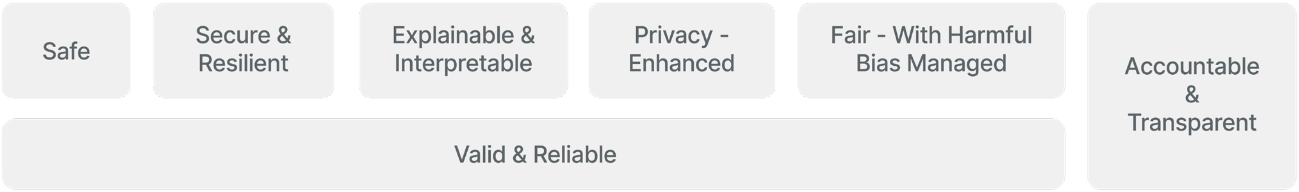
When it comes to enabling large language models (LLMs) and generative AI for the enterprise, we need to ensure that the data these systems use adheres to enterprise security standards. The data being processed and analyzed often contains personal data, including sensitive information like medical records, financial details, and location information.

Without robust data security measures, this data can fall into the wrong hands, leading to identity theft, financial fraud, and other privacy violations. Responsible AI practices emphasize protecting individual privacy by implementing strong security posture controls like encryption, granular access controls, and data anonymization techniques to minimize the risk of unauthorized access and misuse.

Using data security to implement data hygiene and implement strong policies around data usage is also critical for preventing biased and discriminatory outcomes in AI systems. When AI algorithms are trained on biased data, they can perpetuate existing societal inequities, leading to unfair decisions that disadvantage certain groups of people. By implementing data security measures, such as data provenance tracking, AI developers can ensure that the data used to train and operate AI systems is accurate, representative, and free from bias.

Responsible AI practices also emphasize transparency and accountability in data handling. Clear and accessible data governance policies, along with audit trails and reporting mechanisms, help ensure that AI systems are used in a transparent and accountable manner. Data security plays a key role in this process, providing tools for tracking data flows, identifying unauthorized access attempts, and maintaining a record of data usage.

The growing use of AI in various aspects of our lives has raised concerns about data privacy and security. When data breaches and misuse of personal information occur, it erodes public trust and hinders the acceptance of AI technologies. By prioritizing data security, AI developers and organizations can demonstrate commitment to responsible AI practices, building trust with users and fostering a more positive perception of AI.



Safe

Secure & Resilient

Explainable & Interpretable

Privacy - Enhanced

Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

Figure 1 – NIST Characteristics of trustworthy AI systems.

The trustworthiness characteristics of AI systems are a.o. tied to the datasets those AI systems rely on. By ensuring proper security controls and content validation of those datasets, we can increase the reliability and responsible nature of these systems and improve their enterprise readiness.

AI TRUST, RISK, AND SECURITY MANAGEMENT (TRISM)

AI models and applications aren't innately reliable, trustworthy, fair and secure, AI TRiSM is a set of solutions and approaches to proactively identify and mitigate these risks. AI TRiSM is more broadly applicable than the data being used by the models and safeguarding the models itself, but they are core components of it.

AI TRiSM is concerned with:

- Explainability/Model Monitoring
- ModelOps
- AI Application Security
- Privacy

It is very hard to explain and clarify how a model functions, on a deeply technical level for example we don't really understand how neural networks, one of the key ingredients of modern AI, work and are generally considered to be black boxes. Therefore we need to control what we can, which is the data, and its inherent biases, that we train the models on.

From a TRiSM perspective we can focus on discovering the AI Models using DSPM capabilities, for example understanding if a model appears in a public cloud location and is attached to a cloud VM. We can also assess the AI model risk related to data and application input from IaaS, SaaS, and PaaS sources and potentially classify datasets according to sensitivity or regulatory relevance.

RUBRIK DSPM: DATA SECURITY FOR AI

ENTERPRISE DATA USAGE IN LLMS

Rather than solely relying on existing general purpose generative AI solutions like ChatGPT, enterprises are moving towards leveraging their own existing trove of data to feed large language models and build tailored solutions and custom applications.

Alternatively, enterprises can also explore fine tuning models with existing company data to improve their relevance for specific use cases. For example, creating a web enabled chat agent that has access to all your customer support data.

Other approaches use retrieval-augmented generation (RAG) to enhance the accuracy and reliability of generative AI models with facts fetched from external sources, in this case enterprise data sets.

SECURING ENTERPRISE DATA FOR AI USAGE

Enterprises should be enabled to deploy these AI systems in production. Rubrik DSPM can provide much-needed visibility and control over the data consumed by these systems, and help prevent inadvertent data exposure during model training or deployment.

Provide visibility into the datasets leveraged by LLM models.
Rubrik DSPM provides data discovery for structured, unstructured, and semi-structured data, and classification used for various frameworks like PCI, HIPAA, GDPR, etc. Additionally, it can map which users and roles have access to this type of data, allowing you to understand whether these models use this data.

AI Data Lineage

Rubrik can identify the source of data assets and copies of data that can ultimately become the input for these models. This way, you can trace the origin of the original information and provide more visibility into the model input.

Prevent model contamination with unintended data.

Once these models are trained, they often become black boxes, making it difficult to understand if sensitive or unintended data was used as input. By monitoring data being moved to a data asset known for input, we can prevent model contamination and help ensure responsible AI.

Data Access Governance for AI Data.

By mapping data access for internal and external identities, we can safeguard data against tampering by non-sanctioned employees or third party entities, ensuring only pre-approved personnel can access and work with the datasets that will ultimately end up as a baseline for these models.

Spotting unsanctioned models and systems.

By identifying shadow data and shadow models running on unmanaged cloud infrastructure, including databases running on unmanaged VMs, we can provide security teams with alerts on sensitive data stored or moved into these systems. We can also detect when a VM is used to deploy an AI model on the basis of these datasets.

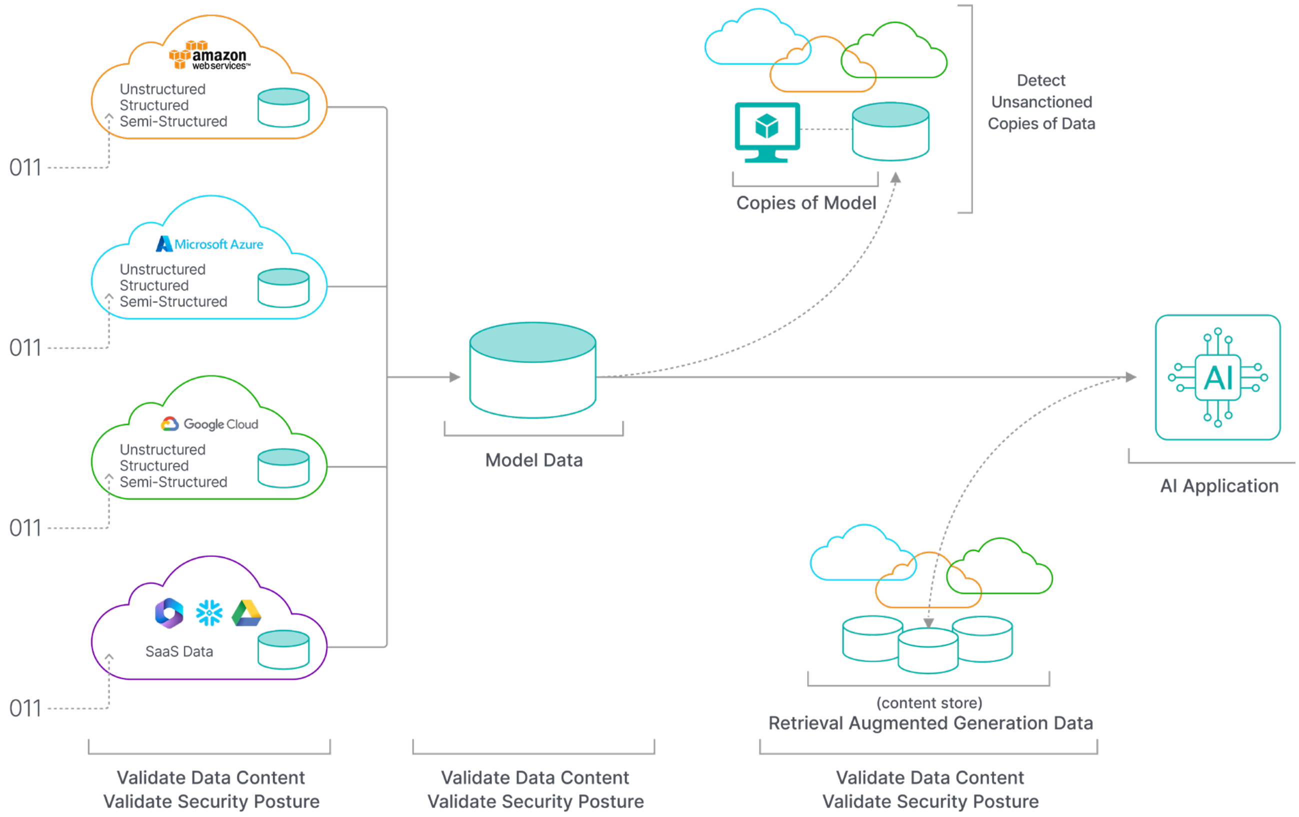


Figure 2 – Rubrik DSPM for AI Datasets

OVERVIEW OF INDUSTRY FRAMEWORKS FOR DATA SECURITY + AI

Several early frameworks exist that attempt to cover the potential security risks of using LLMs and AI more broadly, advocating for security best practices to avoid common pitfalls. Looking at the data security recommendations, some initial guidance can be extracted.

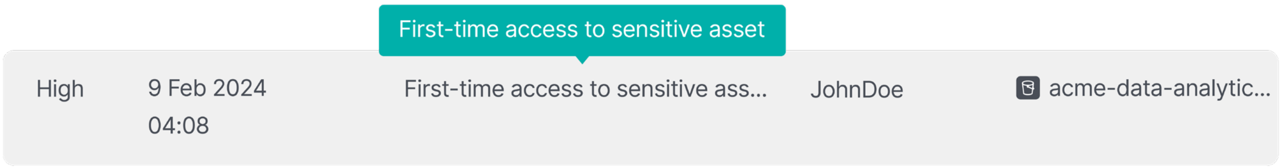
OWASP Top 10 for LLM Applications

The OWASP Top 10 for LLM Applications[1] is a list of the top ten security and safety issues developers and security teams must consider when building applications leveraging Large Language Models (LLMs). It was created by a team of over 370 security experts, AI researchers, developers, and industry leaders.

LLM03: TRAINING DATA POISONING

> **“Training data poisoning refers to manipulation of pre-training data or data involved within the fine-tuning or embedding processes to introduce vulnerabilities (which all have unique and sometimes shared attack vectors), backdoors or biases that could compromise the model’s security, effectiveness or ethical behavior. Poisoned information may be surfaced to users or create other risks like performance degradation, downstream software exploitation and reputational damage. Even if users distrust the problematic AI output, the risks remain, including impaired model capabilities and potential harm to brand reputation.”**

Recommendation: By understanding data access patterns to your training data assets, and mapping which identities have access to your highly sensitive data locations you can ensure the input data has not been compromised and ensure its integrity.

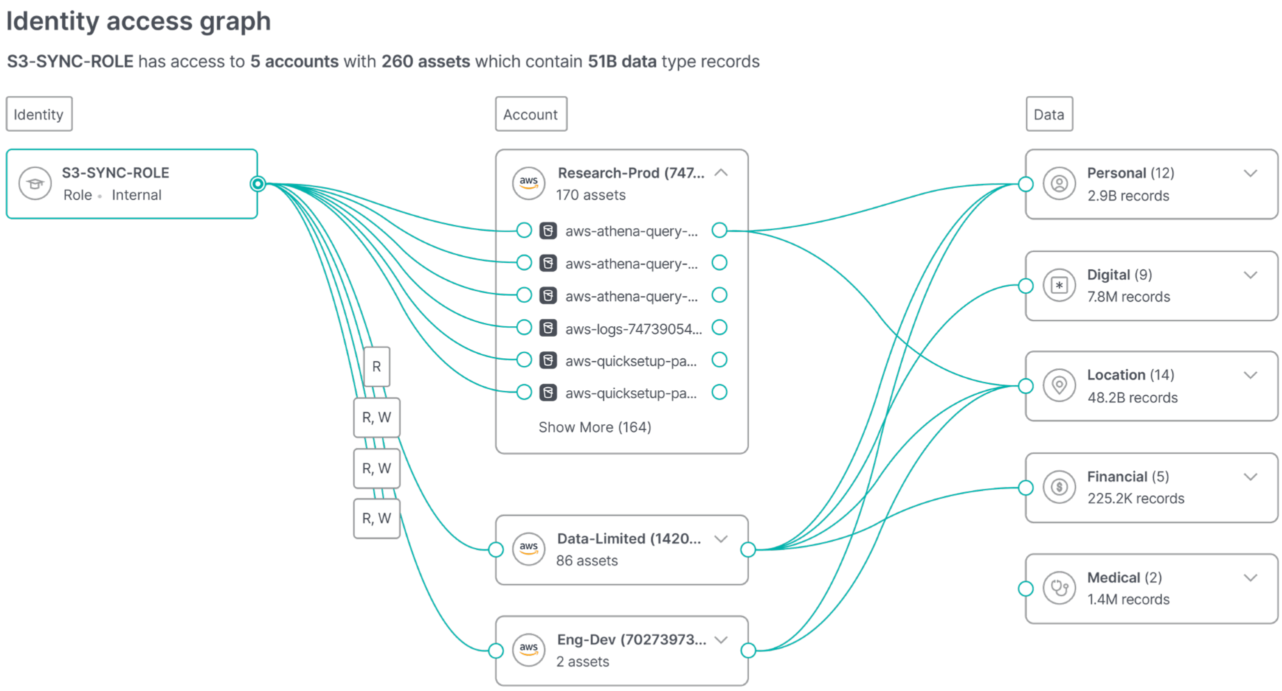


First-time access to sensitive asset

High	9 Feb 2024 04:08	First-time access to sensitive ass...	JohnDoe	acme-data-analytic...

Understanding which identity/role has access to these data assets allows you to understand potential risk to the model, by mapping the permissions you can differentiate between risk of data exposure and risk of data tampering.

[1] https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf



Identity access graph

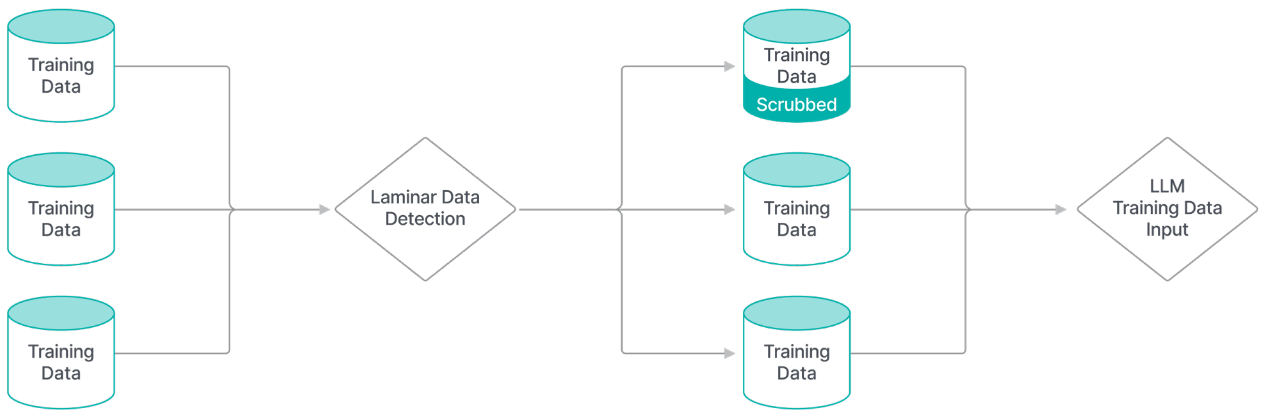
S3-SYNC-ROLE has access to **5 accounts** with **260 assets** which contain **51B data** type records

LLM06: SENSITIVE INFORMATION DISCLOSURE

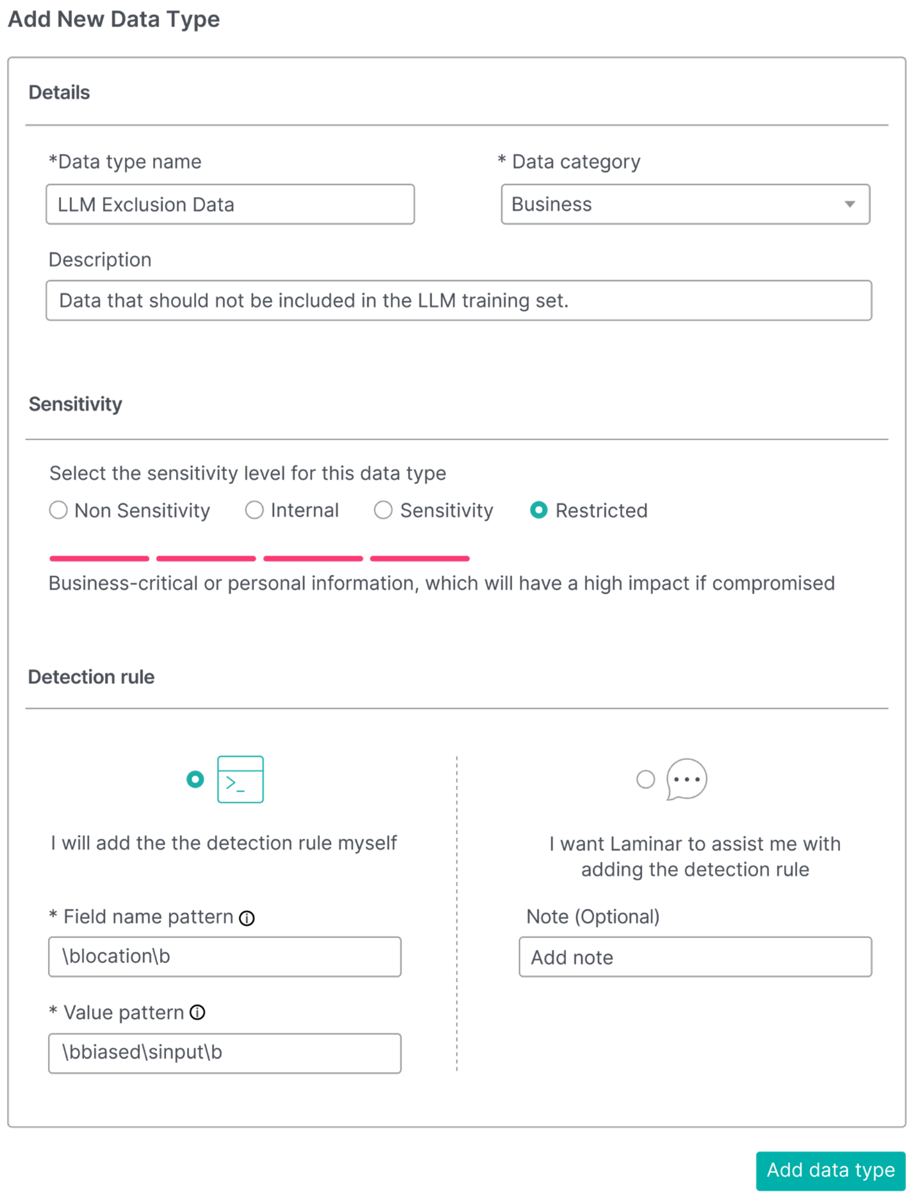
> **"LLM applications have the potential to reveal sensitive information, proprietary algorithms, or other confidential details through their output. This can result in unauthorized access to sensitive data, intellectual property, privacy violations, and other security breaches. It is important for consumers of LLM applications to be aware of how to safely interact with LLMs and identify the risks associated with unintentionally inputting sensitive data that may be subsequently returned by the LLM in output elsewhere.**
>
> **To mitigate this risk, LLM applications should perform adequate data sanitization to prevent user data from entering the training model data. LLM application owners should also have appropriate Terms of Use policies available to make consumers aware of how their data is processed and the ability to opt out of having their data included in the training model."**

Recommendation: By validating the training data set beforehand, and scanning for sensitive data or other specific terms and keywords, you can ensure the expected dataset is used as the basis of training.

Rubrik's DSPM's classification engine continuously scans your environment for (sensitive) data patterns. Scanning each location for all specific active data types using our proprietary algorithms, pattern matching, and metadata. Built-in and custom date types define each location's sensitivity level and the security policies that may apply.



Example:



Add New Data Type

Details

*Data type name: LLM Exclusion Data

* Data category: Business

Description: Data that should not be included in the LLM training set.

Sensitivity

Select the sensitivity level for this data type

○ Non Sensitivity ○ Internal ○ Sensitivity ● Restricted

Business-critical or personal information, which will have a high impact if compromised

Detection rule

● I will add the the detection rule myself

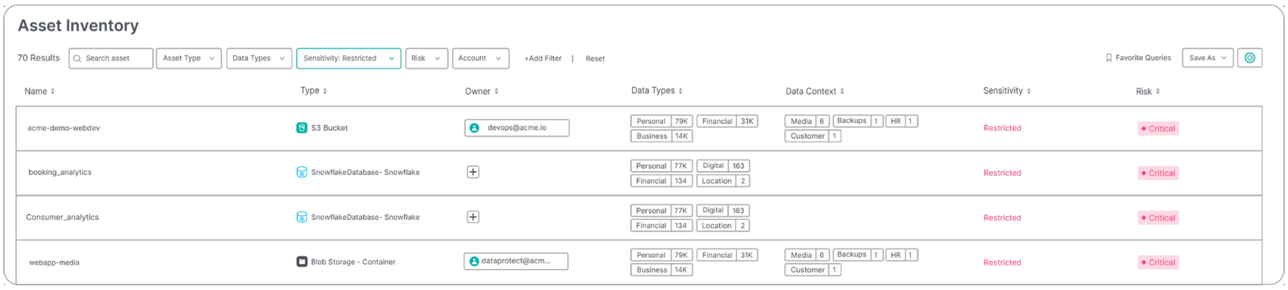
* Field name pattern: \blocation\b

* Value pattern: \bbiased\sinput\b

○ I want Laminar to assist me with adding the detection rule

Note (Optional): Add note

The Restricted data will then be identified across the data assets marked for LLM input and can be scrubbed or excluded for processing.



Asset Inventory

70 Results | Search asset | Asset Type | Data Types | Sensitivity: Restricted | Risk | Account | +Add Filter | Reset | Favorite Queries | Save As

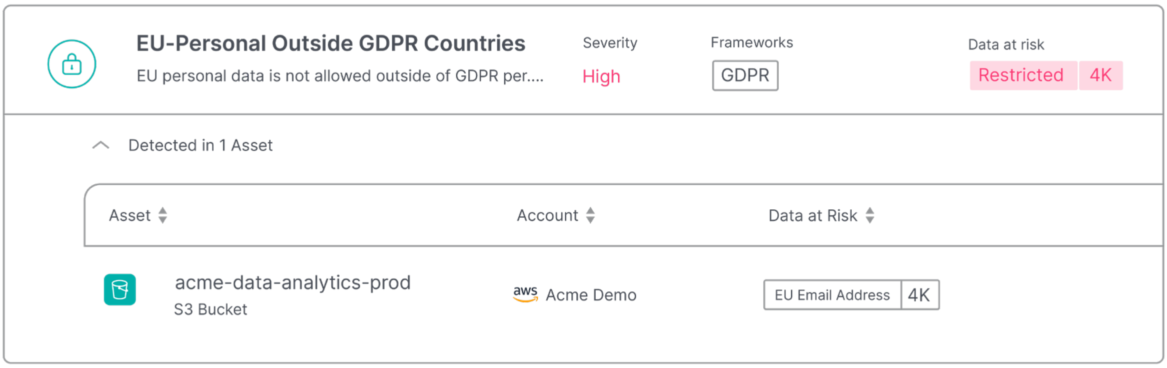
Name	Type	Owner	Data Types	Data Context	Sensitivity	Risk
acme-demo-webdev	S3 Bucket	devops@acme.io	Personal 79K Financial 31K Business 14K	Media 6 Backups 1 HR 1 Customer 1	Restricted	◆ Critical
booking_analytics	SnowflakeDatabase- Snowflake	+	Personal 77K Digital 163 Financial 134 Location 2		Restricted	◆ Critical
Consumer_analytics	SnowflakeDatabase- Snowflake	+	Personal 77K Digital 163 Financial 134 Location 2		Restricted	◆ Critical
webapp-media	Blob Storage - Container	dataprotect@acm...	Personal 79K Financial 31K Business 14K	Media 6 Backups 1 HR 1 Customer 1	Restricted	◆ Critical

The result will be an LLM trained on exactly the type of data you expect, excluding any training data poisoning influencing the ultimate outcome.

LLM10: MODEL THEFT

> **“The theft of LLMs represents a significant security concern as language models become increasingly powerful and prevalent. Organizations and researchers must prioritize robust security measures to protect their LLM models, ensuring the confidentiality and integrity of their intellectual property. Employing a comprehensive security framework that includes access controls, encryption, and continuous monitoring is crucial in mitigating the risks associated with LLM model theft and safeguarding the interests of both individuals and organizations relying on LLM.”**

Recommendation: By leveraging Rubrik's DSPM capabilities, you can control access to sensitive data sets like AI models and prevent them from being exfiltrated from your environment. Rubrik's DSPM’s Data Detection and Response (DDR) capability can detect external and internal threats to your sensitive data by identifying anomalous data access and suspicious behavior—alerting you on data exfiltration, unusual third-party access, insider threats, accidental data leaks, data misuse, and other threats. Furthermore with Data Access Governance (DAG) you can map out who has access to your most sensitive data—whether they are privileged users or not. For high-risk identities, like your AI developers you can right size their permissions and add extra protections to limit the potential scope and damage of security incidents.



EU-Personal Outside GDPR Countries

EU personal data is not allowed outside of GDPR per....

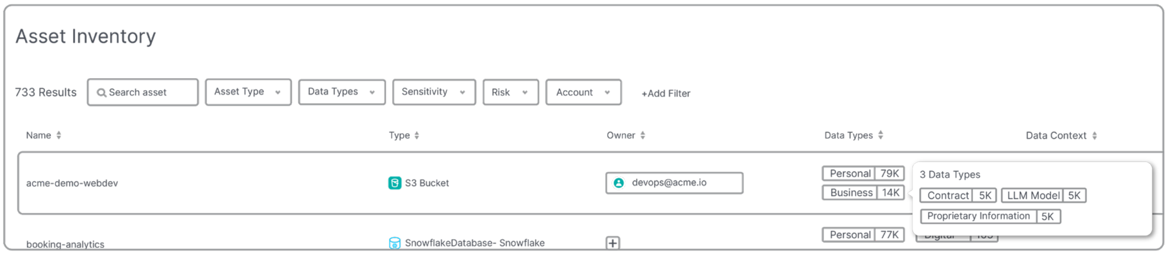
Severity: High

Frameworks: GDPR

Data at risk: Restricted 4K

Detected in 1 Asset

Asset	Account	Data at Risk
acme-data-analytics-prod S3 Bucket	aws Acme Demo	EU Email Address 4K



Asset Inventory

733 Results | Search asset | Asset Type | Data Types | Sensitivity | Risk | Account | +Add Filter

Name	Type	Owner	Data Types	Data Context
acme-demo-webdev	S3 Bucket	devops@acme.io	Personal 79K Business 14K	
booking-analytics	SnowflakeDatabase- Snowflake	+	Personal 77K	

3 Data Types

Contract 5K | LLM Model 5K | Proprietary Information 5K

NIST AI Risk Management Framework

The AI Risk Management Framework[2] (AI RMF) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

The framework mentions several areas where a DSPM approach could be relevant.

RISK PRIORITIZATION

> **"Attempting to eliminate negative risk entirely can be counterproductive in practice because not all incidents and failures can be eliminated. Unrealistic expectations about risk may lead organizations to allocate resources in a manner that makes risk triage inefficient or impractical or wastes scarce resources. A risk management culture can help organizations recognize that not all AI risks are the same, and resources can be allocated purposefully. Actionable risk management efforts lay out clear guidelines for assessing trustworthiness of each AI system an organization develops or deploys.**
>
> **Higher initial prioritization may be called for in settings where the AI system is trained on large datasets comprised of sensitive or protected data such as personally identifiable information, or where the outputs of the AI systems have direct or indirect impact on humans."**

Recommendation: By understanding which datasets marked for use in AI contain sensitive data, and what type of sensitive data, you can alleviate the risk associated with inadvertently overexposing this type of information.

ORGANIZATIONAL INTEGRATION AND MANAGEMENT OF RISK

> **"The AI RMF may be utilized along with related guidance and frameworks for managing AI system risks or broader enterprise risks. Some risks related to AI systems are common across other types of software development and deployment. Examples of overlapping risks include: privacy concerns related to the use of underlying data to train AI systems."**

Recommendation: By understanding the privacy implications of datasets used in LLMs and AI systems, you can more confidently and quickly deploy these systems, while limiting the risk of data overexposure.

[2] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

SECURE AND RESILIENT

> **"Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure."**

Recommendation: By safeguarding datasets, models, and other intellectual property using Data Security Posture Management capabilities, you can more effectively and reliably build enterprise AI systems and safely operate them over their entire life-cycle.

ACCOUNTABLE AND TRANSPARENT

> **"Maintaining the provenance of training data and supporting attribution of the AI system's decisions to subsets of training data can assist with both transparency and accountability. Training data may also be subject to copyright and should follow applicable intellectual property rights laws."**

Recommendation: By validating that the training data sets do not contain the aforementioned data types, or if they do bring awareness and traceability to those contents, you can manage requirements of applicable rules and regulations, including specific intellectual property rights laws.

PRIVACY-ENHANCED

> **"Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals. Privacy-enhancing technologies ("PETs") for AI, as well as data minimizing methods such as de-identification and aggregation for certain model outputs, can support design for privacy-enhanced AI systems."**

Recommendation: By identifying potential sensitive data types in data sets, you can inform data tokenization systems about which data to anonymize and use more confidently in your AI applications.

FAIR – WITH HARMFUL BIAS MANAGED

> **"Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples."**

Recommendation: By building your own list of data types, and looking for those data types across your input data, you can investigate potential AI datasets and filter out unwanted samples, resulting in a more appropriate basis for your AI applications.

U.S. Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence[3]

The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

The executive order mentions data safety as an area where a DSPM approach could be relevant.

> **"Without safeguards, AI can put Americans' privacy further at risk. AI not only makes it easier to extract, identify, and exploit personal data, but it also heightens incentives to do so because companies use data to train AI systems."**

Recommendation: Using sensitive data discovery, you can identify which types of data are used as potential input sources for these AI applications, and action can be taken to remove PII or other specific data out of the data set.

CISA Roadmap for AI[4]

To promote responsible AI use, CISA will create their own NIST AI Risk Management Framework (RMF) profile to help develop and implement security and privacy controls for AI. The framework will include AI data requirements and uses.

The roadmap mentions adversarial machine learning, where a DSPM approach could be relevant.

ADVERSARIAL MACHINE LEARNING

> **"Malicious cyber actors target vulnerabilities throughout the AI supply chain to cause certain behavior, unintended by the system owner or operator, in machine learning (ML) systems—referred to as adversarial machine learning. For example, malicious actors may manipulate training data, affect the performance of the ML model's classification and regression, or exfiltrate sensitive ML model information."**

Recommendation: To avoid manipulation of training data, you can validate not only user access to data locations, but also the data content itself.

[3] https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[4] https://www.cisa.gov/sites/default/files/2023-11/2023-2024_CISA-Roadmap-for-AI_508c.pdf

EU Artificial Intelligence Act[5]

The EU Artificial Intelligence Act proposes comprehensive rules for trustworthy AI in the European Union. Other regions like ASEAN, or even specific countries like Japan, have set their own, sometimes competing, intentions on the use and regulation of these AI systems.

EU negotiators reached a provisional agreement on the Artificial Intelligence Act. This regulation aims to ensure that fundamental rights, democracy, rule of law and environmental sustainability are protected from high risk AI, while boosting innovation and making Europe a leader in the field. The rules establish obligations for AI based on its potential risks and impact.

Recommendation: The proposed Act mentions several areas where a DSPM approach could be relevant. They specifically call out certain banned applications using sensitive personal data to provide categorization.

By understanding the sensitive data being fed into these AI systems, we can prevent potential misuse of the applications working on the basis of these data sets.

CONCLUSION

To safely and reliably gain the benefits of Generative AI for the enterprise, we need to ensure secure use of the data used as its basis. Data Security Posture Management and Data Detection and Response capabilities like those found in Rubrik DSPM bring a much needed level of security to elevate LLMs and Generative AI systems to a level acceptable for enterprise usage. Focusing on the security of the most crucial parts of the AI equation, the data and models, we can keep the pace of innovation high, whilst not compromising on trust and reliability.

Author: Filip Verloy | Field CTO EMEA & APJ Rubrik X

[5] https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

rubrik

Global HQ
3495 Deer Creek Road
Palo Alto, CA 94304
United States

1-844-4RUBRIK
inquiries@rubrik.com
www.rubrik.com

Rubrik is on a mission to secure the world's data. With Zero Trust Data Security™, we help organizations achieve business resilience against cyberattacks, malicious insiders, and operational disruptions. Rubrik Security Cloud, powered by machine learning, secures data across enterprise, cloud, and SaaS applications. We help organizations uphold data integrity, deliver data availability that withstands adverse conditions, continuously monitor data risks and threats, and restore businesses with their data when infrastructure is attacked.

For more information please visit www.rubrik.com and follow @rubrikInc on X (formerly Twitter) and Rubrik on LinkedIn.

Rubrik is a registered trademark of Rubrik, Inc. All company names, product names, and other such names in this document are registered trademarks or trademarks of the relevant company.