cyberhaven

Data Lineage

Powering the *Next Generation* of Data Security

Strengthening security and compliance by understanding how data moves

Introduction

The evolution of data security

Data security is evolving faster than legacy tools:
Data is an enterprise's greatest strategic asset and existential risk. As technology advances, data is being created, shared, and transformed at speeds legacy security solutions can't keep up with.

Traditional methods are failing. Static scanning misses dynamic data flows, content-based classification struggles with complex data types, and investigations lack the context needed to understand what really happened.

Data lineage: Understanding data in motion

Data lineage tracks the complete history of data throughout its lifecycle—creating a comprehensive record of how information is created, moved, transformed, and used across your organization.

Unlike traditional approaches that focus on what data contains, data lineage reveals how data moves and transforms, providing security teams with crucial context about data usage. This shift in perspective fundamentally changes how organizations address their most critical security challenges.

Beyond traditional security

Data lineage represents a fundamental shift—from viewing data as static content to understanding it as dynamic flows across people, applications, and organizations.

This approach acknowledges how modern enterprises actually work: data constantly moves, transforms, and gets shared both within and beyond organizational boundaries. By capturing this movement pattern, security teams can make more informed decisions about risk and protection.

In this whitepaper, we'll explore the technical foundations of data lineage, examine how it works and how it's implemented in practice, and provide real-world examples of organizations that have transformed their security operations using this revolutionary approach.

Table of contents

Section 01

What is Data Lineage?

Origins of data lineage

Data lineage is the history of data from its origin through its entire lifecycle. It provides a detailed record of how data is created, moved, transformed, and used.

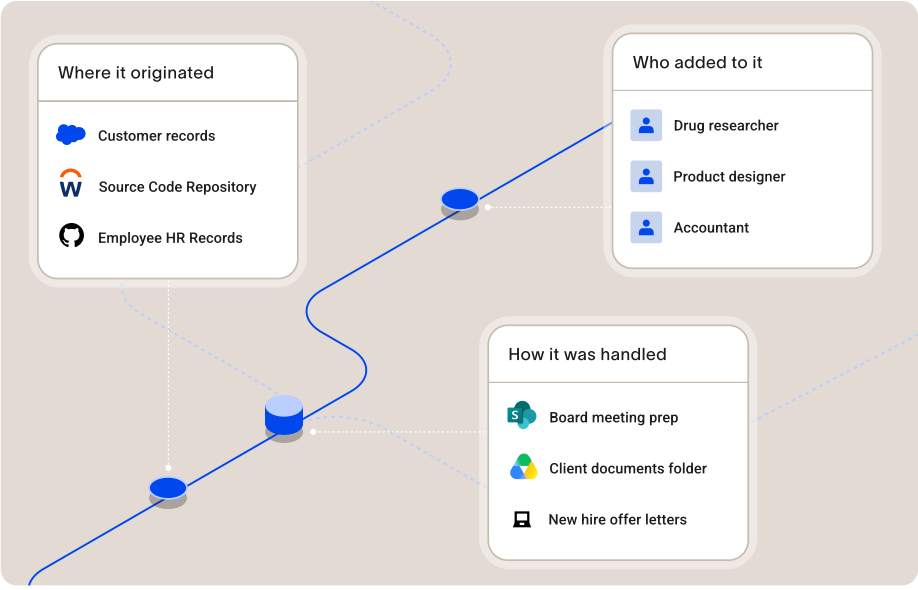
Armed with data lineage, one can answer fundamental questions about the origin, usage, and nature of that piece of data—insights that have broad applicability for anyone using data.

The earliest records of data lineage as a concept can be found among United States geologists and cartographers in the early 1980s. Resource surveyors cataloging and publishing data about mineral wealth in the United States were among the earliest adopters of electronic recordkeeping. Given the complexity of the data collection process and the billions of dollars in land management decisions that would be made based on the data, surveyors sought to document data lineage to enable quality assurance of the collection and processing system.

Data quality continued to be the primary driver behind the development of data lineage, especially as databases, electronic recordkeeping, and data analysis gained wider adoption. By 1999, SQL server guidebooks described how to leverage server logs to create data lineage and ensure data integrity. Likewise, data cataloging technologies were building lineage features to help database administrators ensure the quality of the data they were using.

> "Lineage records and auditing capabilities would be required to record changes, who made them, and when they were made."

An Application of Geographic Information System Technology to Sand and Gravel Resource Planning, Mary Kate Beard, University of Wisconsin-Madison (1984)



Data lineage allows one to answer critical questions surrounding the creation and history of data.

The history of data security and data lineage

Since the launch of the first data security solutions in 2000, commoditized approaches to data security have flooded the market for almost twenty years and failed to adapt to the evolving landscape of data usage faced by security teams.

01
Greater value and scrutiny

Data security started as checking straightforward regulatory boxes, but this has not remained the case. Data and intellectual property are viewed as an enterprise's greatest strategic assets. They empower employees and customers to create value for the company and represent an existential threat if leaks or data loss occur. Additionally, regulatory frameworks surrounding data have increased in complexity, such as GDPR or PCI DSS 4.0, demanding more visibility and control from security teams.

02
Breadth of data access

The rise of cloud software, cloud infrastructure, and mobile/personal device access has increased the number of locations where sensitive company data may be accessed, creating challenges around monitoring and controlling legacy tools that were never built to handle.

03
Speed of transformation and collaboration

With greater access to technology and greater integration of data with company value and processes, the speed at which sensitive data is created, transformed, and moved is far greater than legacy approaches to securing data were designed to handle.

In this new environment, the limitations of legacy approaches to data security were clear:

Discovery was focused on static scanning

The legacy approach to "data discovery" solutions was scanning known repositories of sensitive data. However, given the dynamism of data usage and the breadth of locations where sensitive data can exist, this approach gives a woefully incomplete picture of risk.

Dependent on content patterns to define sensitivity

Legacy technologies were developed to tackle regulatory challenges, like health data and credit card numbers. However, with the growing importance of more complex and text-free data types—unstructured documents, images, videos, blueprints, etc.—content analysis is a limited approach.

Limited context to triage and investigate

When data access patterns were simple, a brief understanding of what led to an alert could be enough to respond. However, with a greater breadth of data access and changing business needs surrounding data, security teams struggled to investigate and respond to incidents.

Cyberhaven brings data lineage to Cybersecurity in 2019

The application of data lineage to the challenge of securing data didn't occur until 2019, when Cyberhaven launched the industry's first Data Detection and Response (DDR) solution.

Cyberhaven's story begins with a team of Swiss PhDs and the 2016 DARPA Cyber Grand Challenge. Cyberhaven's founding team developed technology to rapidly analyze the flow of data between applications and applied this capability to machine-led hacking and vulnerability identification.

Building on the ability to rapidly analyze data flows, Cyberhaven developed its data lineage approach to help address the challenges of legacy data security. Even early versions of the technology were recognized by security and IT leaders for its potential.

Since its launch in 2019, Cyberhaven's data lineage approach to security has gained momentum amongst the most innovative security organizations. With the integration of data lineage with real-time protection and insider risk capabilities, industry analysts are recognizing the transformative potential of leveraging data lineage to better secure data.

Section 2 of this paper discusses how lineage works at a high level and the technological advances that were necessary to enable its usefulness for security teams. Section 3 breaks down the specifics of how data lineage helps address key challenges of data security.

Cyberhaven co-founder and CTO Volodymyr Kuznetsov accepting Most Innovative Solution at the Security Leadership Exchange.

"The DLP market has been saturated with traditional, content-heavy DLP solutions that may not fully cater to the dynamic data security requirements of modern organizations. Invest in a DLP solution that can understand the full lineage of the data."

Gartner 2023 Market Guide for Data Loss Prevention

Section 02

How to Build Data Lineage for Data Security

STEP 1

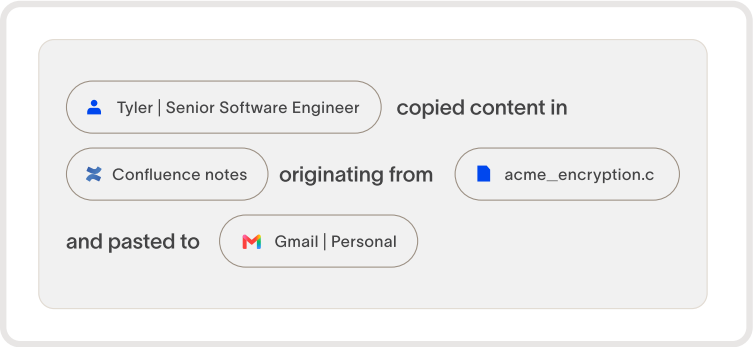
Signal collection

The first step in creating data lineage is the collection of relevant signals regarding data creation, usage, and movement. For the typical enterprise, critical entities to collect signals from include:

- Corporate endpoints
- SaaS applications
- Cloud infrastructure
- Corporate email
- Company data centers

Collecting signals from the audit logs of these systems and storing them for analysis is the first step in creating lineage.

Simple signal collection isn't enough for data lineage. The ability to create lineage and make it useful comes from collecting metadata associated with these signals. Detailed time, user, source, and destination information are all required to correlate events and harness the insights within lineage.



Event metadata enables the creation and provides the value of data lineage.

Endpoint Data Collection

Endpoint interaction with data represents some of the most risky access, sharing, and transformation of data and some of the hardest signals surrounding data to collect. Legacy data security technologies utilized invasive techniques with the kernel to gain visibility, making it difficult to collect granular signals and creating performance and stability issues. Both Windows and macOS have introduced significant developments that were not available to first-generation data security technologies, which have made it possible to collect detailed information regarding data usage reliably.

Event Tracing for Windows (ETW)

Event Tracing for Windows, known as ETW, allows programs to collect detailed information surrounding the operation of other applications and users. ETW was first introduced with Windows 2000 but underwent a significant upgrade with the release of Windows 7 in 2008, which added 20 times more granularity and significantly improved the developer experience.

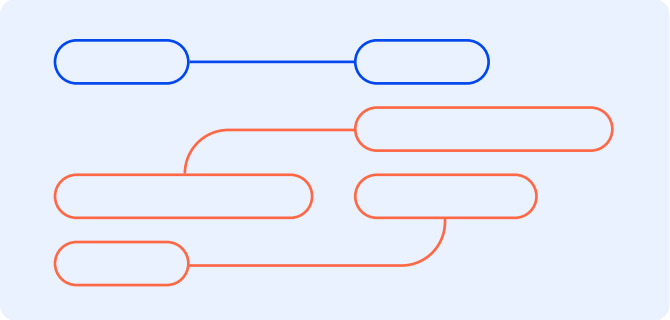
macOS Endpoint Security Framework (ESF)

ESF was introduced in 2019 to enable security technologies to work with Macs as a part of Apple's efforts to broaden enterprise adoption of macOS. Modeled after ETW, Apple provides developers with the ability to access granular telemetry from other applications for security purposes.

STEP 2

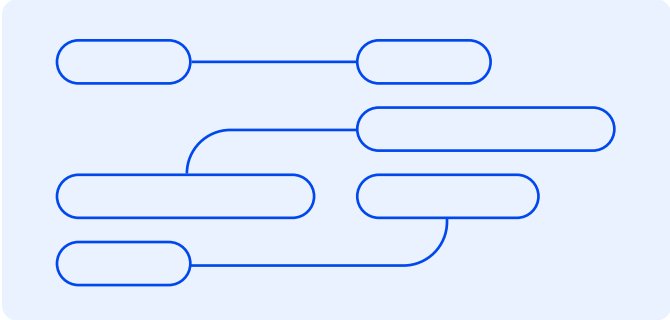
Processing and correlation

Once signals have been collected, metadata must be processed to create data lineage. The accuracy of this processing is dependent on the detail of the metadata collected and the rigor of data processing.



False link

Relying on signals like the operation, file hash, and file size is a good start. However, correlating details like destination and source are necessary to track the flow of data accurately. In this example, the file is identified as the same content via its hash and size, but the destination of the upload does not match the source of the download. Signals must be further analyzed to complete the picture of data flow.



True link

This processing is continued across all events to create a complete view of data lineage.

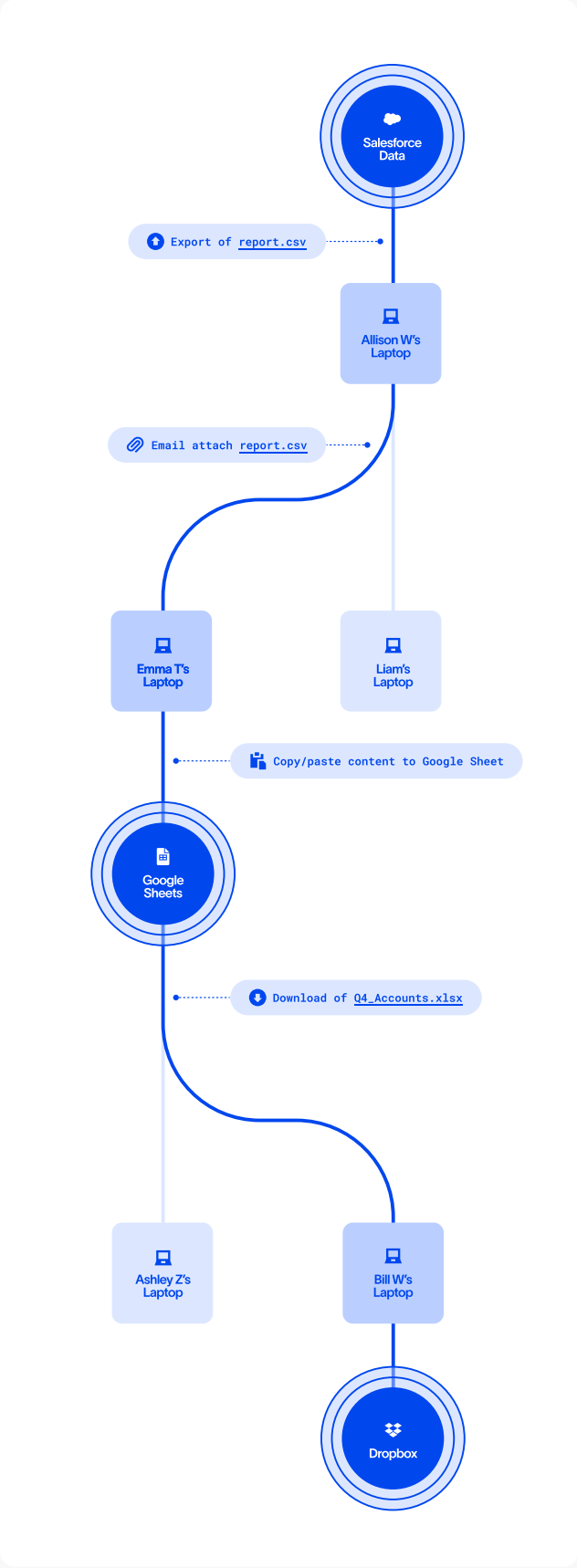
Global vs. local lineage

As data lineage has gained momentum in the industry, it's important to examine the scope of various approaches to data lineage.

Global Data Lineage traces the entire history of data across an entire organization. This reflects the natural and practical realities of how data is created, modified, and evolved in organizations: moving across endpoints, cloud, and between individuals over time.

Local Data Lineage is limited to a specific person, cloud, or endpoint machine and considers only how that person or machine has interacted with the data. That means it's blind to what might have happened before the data arrived and to what happens after it's sent to another part of the organization.

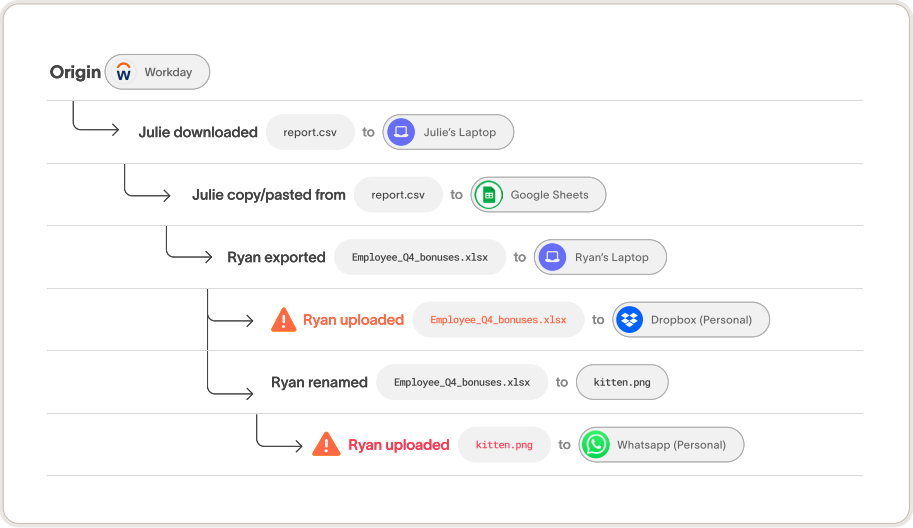
Limiting the scope of lineage can make signal processing significantly easier, but it also reduces the level of insight and security improvements lineage can provide a security organization. These limitations will be addressed in the third section.



STEP 3

Querying, alerting, and real-time protection

With data lineage created, the next step is making it useful for security use cases by enabling querying, alerting, and real-time data protection. To query data lineage, the user defines the metadata fields they are interested in and matching events are returned.



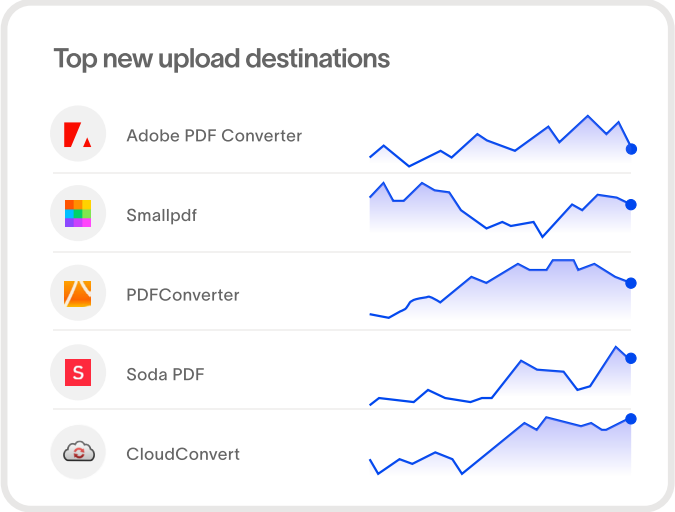
Example visualization of data lineage

Real-time response and graph database advances

The immense volume of signals generated by enterprise data usage and latency requirements make alerting and real-time response based on data lineage a challenging technical problem. When an event occurs, the system must first process its metadata to determine its lineage. Then, it must check if any policies have been violated and respond appropriately if necessary.

Utilizing data lineage at the scale of modern enterprise data and within the latency requirements of real-time policy enforcement would not be possible without recent advancements in the performance and scalability of graph database technology. Graph databases are a natural choice for representing data lineage because they natively support the concept of connections between various data points and are optimized to create and query these connections. This enables easy development and lower latency when creating and utilizing data lineage.

In the 2010s, graph databases grew in popularity with the rise of social networking software. Foundational advances made by the engineering teams at companies such as Facebook have been instrumental in making data lineage practically useful as a data security technology.



Example visualization of results from lineage analysis, aggregated by upload destination

Section 03

How Data Lineage Is Transforming Data Security

A Risk discovery

Establishing a robust data security program begins with identifying risks within the organization. Understanding risk is crucial not only for enforcing compliance but also for aligning security policies with business needs. In recent years, tools like Data Discovery, Data-at-Rest Scanning, and Data Security Posture Management (DSPM) have aided in this risk identification process. While effective to some degree, these traditional tools reveal three key limitations:

1. Reactive: As businesses adopt new technologies and processes, traditional scanning tools often lag, failing to keep pace with emerging risks. Organizations need to know about a SaaS app or new infrastructure to initiate scanning, leading to Shadow IT challenges where unmonitored applications pose hidden risks.

2. Latent risk: Traditional tools primarily focus on static, stored data, which leaves critical risk blind spots. Data at rest is only part of the picture; sensitive data is often most vulnerable in motion. Data lineage reveals these active risks by showing when data moves to unsanctioned storage, or if it's shared inappropriately.

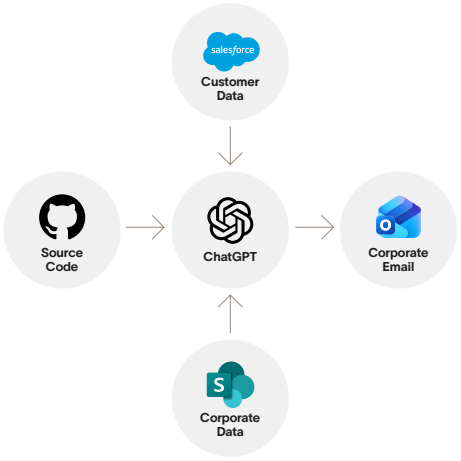
3. Lack of context: A static scan only shows that sensitive data exists, but lacks insight into how the business uses it. Without context, security teams miss out on understanding the team or business process that depends on the data, making it difficult to assess the impact of exposure or loss.

In contrast, data lineage changes the game by providing continuous, proactive monitoring that's independent of data sources.

This means:

- Security teams gain real-time visibility into emerging risks as new technologies are adopted.
- They get a complete picture of data movement and sharing patterns, enabling them to identify active risks.
- They can better understand the "who," "what," and "where" of data interactions, offering valuable context on business processes and the value data contributes to them.

Discovery without data lineage	Discovery with data lineage
Reactive	Proactive
Latent risk	Active risk
No context	Insight into business usage



With data lineage, security teams can work alongside business units to understand technology usage and proactively identify data security risks.

Global vs. local lineage: Risk discovery

Global lineage provides a holistic view of data movement across the entire organization, making it possible to proactively discover and monitor risks as data flows between departments, applications, and locations. This approach enables security teams to see the entire data journey, identifying risk hotspots that may be missed if only local lineage is used. In contrast, local lineage, which focuses on data interactions within a single endpoint or individual's activities, offers a narrower, more isolated view. While useful for immediate tracking, local lineage can overlook broader patterns of risk that emerge as data moves across different users and systems.

Case Study

Understanding and securing ChatGPT at VillageMD

Background

VillageMD is a leading healthcare innovator, serving seven million patients across 26 states in the U.S. With such a large footprint, VillageMD constantly innovates while balancing regulatory obligations in healthcare.

Challenge

When ChatGPT was introduced, VillageMD recognized its potential as a productivity tool but also the risks it posed to data security, especially considering healthcare regulations. They didn't want to restrict access outright but needed a way to ensure ChatGPT's secure usage.

How data lineage helped

Data lineage allowed VillageMD to monitor how ChatGPT was being used across teams, enabling them to identify business cases for its usage, educate employees on potential risks, and ultimately implement an enterprise license for secure usage. This approach allowed VillageMD to retain the benefits of ChatGPT without compromising patient data.

"Are we using ChatGPT to write a marketing email or a patient letter? Two different things. One has PHI in it; one doesn't. What are we entering into it in order to get what we need? Is it simply a question, or are we entering source code or confidential data? Lineage helps us understand the use cases and the roles of those using these models, which is key."

— Dan Walsh, CISO, VillageMD

B Data classification

Effectively protecting sensitive data starts with accurate data classification. Historically, data security technologies have used two main approaches for classification, each powerful in its own way but with notable limitations:

1. Content-based classification

This method identifies sensitive data by scanning for specific keywords, regex patterns, or exact matches against predefined databases. This approach is effective in many cases, especially for structured data types like Social Security numbers or credit card details. However, it requires direct access to data content, which introduces a set of limitations:

- **Content access limitations:** Not all valuable data has identifiable content patterns. Many modern data types, like images, design files, and unstructured documents (e.g., presentations, blueprints), lack reliable content markers.
- **Encrypted and compressed files:** Sensitive data may be encrypted, compressed, or password-protected, making content-based scanning difficult or impossible in these cases. Without direct access to the content, classification methods that rely on specific text patterns are ineffective.

2. Label- and tag-based classification

Labels and tags can be added to data files to indicate sensitivity, either manually by users or automatically by systems. Both approaches have unique challenges:

- **User-based labels:** This relies on employees to correctly label data, which assumes a high level of accuracy and consistency that's often unrealistic. Mislabeling or omitting tags can expose sensitive information to unintended audiences.
- **Automated labels:** Automated tagging, while useful, often depends on content analysis, which means it inherits many of the same limitations as content-based classification.
- **Location of tags:** Tags can be stored either as part of a file's properties or in the file system itself. However, storing tags in file properties isn't always feasible, as not all file types support metadata properties. Similarly, file system-based tags can be lost or stripped away as data moves, is shared, or is modified, undermining the accuracy of the classification over time.

Moreover, sensitive data isn't limited to specific files; it exists across SaaS applications, sentences or paragraphs in emails, and small snippets of text that may be copy-pasted between documents.

Data lineage acts as a powerful supplement to these traditional classification approaches by addressing these limitations with a context-based strategy. Rather than relying solely on data content, lineage enables classification by adding a broader, contextual layer. Key benefits include:

- **Contextual classification:** Lineage can identify sensitive data based on how and where the data originated, who created or modified it, and what applications or locations it has interacted with. For instance:
  - Data created by certain users, like software engineers, drug researchers, or financial analysts, may indicate specific types of sensitive information.
  - Data sourced from particular applications, domains, or storage locations can imply sensitivity, such as customer records in Salesforce, confidential documents in certain SharePoint folders, or strategy documents in an internal network.
  - Data originating from external domains, like a customer's email, can signal a sensitive client relationship.

- **Handling encrypted and compressed data:** Traditional classification struggles when data is encrypted or compressed. With lineage, classification doesn't rely on content access and is unaffected by data state. Sensitive data can be accurately tracked even when compressed, encrypted, or password-protected.
- **Granular classification:** Data lineage allows classification to occur at a highly detailed level, capturing sensitivity as text is copied and pasted, even in cases where data moves between documents or formats. Classification becomes independent of file structure, allowing sensitive data to be monitored even as it's embedded into new contexts.

Global vs. local lineage: Data classification

With global lineage, classification becomes richer and more context-aware, as it considers not only data content but also its movement and interactions across the organization. This enables sensitivity tagging based on data's journey through various applications, departments, or even interactions with external sources, leading to more accurate and comprehensive classification. Local lineage, however, limits classification to the specific endpoint or individual's actions, potentially missing critical context about the data's origin or usage within the broader organization, and risking a less precise classification approach.

Case Study

Detecting and investigating leaks at Motorola Mobility

Background

Motorola Mobility, a prominent manufacturer of consumer electronics, faced challenges in protecting their intellectual property, especially their product designs and specifications, from competitors and media leaks. With over 5,000 global employees, Motorola needed an accurate and reliable method to secure this sensitive information.

Challenge

The Motorola security team initially relied on a legacy DLP tool, but high false positives and excessive alerts, especially with unstructured data types, created burdensome workloads and undermined the team's efficiency.

How data lineage helped

By implementing data lineage, Motorola's security team leveraged contextual indicators to classify and protect sensitive data accurately. Instead of relying solely on content, they could analyze factors like the teams involved with the data and the SharePoint locations where collaboration occurred. This provided a fuller picture and allowed for more effective, targeted protection of their valuable data.

"Staying ahead of the competition means guarding against insider threats. Cyberhaven gives us visibility into how data flows within our company and stops insider threats in real time."

— Richard Rushing, CISO, Motorola Mobility

C Accelerating investigations

In data security, rapid investigation and response are critical. When an alert is triggered, analysts must quickly assess the scope, cause, and potential impact of the incident to respond effectively. However, traditional approaches often make investigations slow and reactive, leading to delays in understanding what happened and taking corrective action. Data lineage fundamentally transforms this process by providing a clear, step-by-step view of data interactions, allowing analysts to efficiently investigate incidents with comprehensive context.

Key elements of this approach include:

- **Understanding data exfiltration details:** Data lineage provides visibility into exactly what data was accessed or exfiltrated, including details on where it went and how it was handled. If an employee renames or compresses a file to obscure it, lineage captures these actions, ensuring that security teams can identify these evasion attempts. This level of insight is especially useful for understanding unauthorized data transfers, intentional misuse, or insider threats.
- **Identifying contributors to the leak:** When data leakage incidents occur, it's essential to understand not only who accessed or moved the data, but also any other individuals who may have played a role, whether intentionally or unintentionally. Lineage reveals these connections, enabling analysts to see how data was shared, if permissions were misapplied, or if employees handled data improperly.
- **Responding rapidly with comprehensive context:** With data lineage, each alert is contextualized with a complete sequence of events leading up to it. This chronological view allows security teams to see the entire data journey, making it easier to understand the circumstances around a security event. This end-to-end visibility simplifies complex investigations, providing analysts with a clear path to trace data movement, access points, and actions taken with the data.

Beyond alert-driven investigations, data lineage also supports proactive investigation capabilities, allowing security teams to quickly review data interactions following specific events or changes in the organization. Examples include:

- **Layoffs or departures:** A sudden layoff or employee resignation, especially of an executive or knowledge worker, may prompt an investigation into what sensitive data the individual had access to or may have taken with them. Data lineage enables swift reviews, showing a detailed history of data access, modification, and sharing, allowing security teams to mitigate potential risks.
- **Media leak of sensitive data:** In the event of a data breach or leak to the media, data lineage provides a fast, structured approach to investigate and trace the source of the leak. Analysts can track the flow of sensitive data to identify any vulnerabilities in sharing protocols, permissions, or access control that may have led to exposure.

By automating the collection of detailed information on sensitive data usage, data lineage empowers security teams to respond quickly and accurately, even in the face of unexpected incidents or threats. Lineage data allows analysts to make rapid, informed decisions on containment, policy adjustments, and future preventative actions.

Global vs. local lineage: Accelerating investigations

In an investigation, global lineage allows security teams to retrace a data breach or misuse event across the full organizational landscape, providing visibility into how and where data is moved across systems and individuals. This end-to-end view is essential for uncovering hidden connections or identifying multiple contributors to an incident. Local lineage, while useful for analyzing specific endpoints or interactions, may lack the depth required to understand complex incidents involving multiple departments or touchpoints. As a result, investigations using local lineage alone may face gaps in understanding how the data was handled throughout its lifecycle.

Case Study

Securing data and enabling productivity at Cooley LLP

Background

Cooley LLP, a prominent Bay Area-based law firm, represents some of the world's most innovative software and biotech companies. With over 1,000 lawyers and paralegals, Cooley is entrusted with highly sensitive client information and must navigate complex data-sharing and movement requirements, often dictated by law enforcement agencies, judges, or clients.

Challenge

Given the nature of its work, Cooley's security team needed a flexible yet robust system to secure sensitive data while accommodating certain exceptions to rules when necessary. They required the ability to understand data movement context quickly to make informed decisions on policy exceptions without compromising client confidentiality or legal compliance.

How data lineage helped

Data lineage provided Cooley's security team with comprehensive, contextualized insights for every alert. This allowed them to assess each incident with full visibility into data movement and access patterns, enabling informed decision-making for policy exceptions. With lineage, Cooley's team could quickly determine whether data sharing was legitimate and aligned with client needs or if it represented a potential risk.

"Every piece of data we have can be sensitive, depending on timing. If all data is sensitive, it's crucial to understand how data moves, who handles it, and what they're doing with it."

— Mike Santos, Head of Security, Cooley LLP

Conclusion

01 Innovating with data lineage

Data lineage transforms how organizations secure sensitive data by providing complete visibility into how data moves and is used. Unlike legacy tools that rely on static rules and content scanning, lineage maps the entire lifecycle of data, helping security teams understand risks in real time. With this approach, organizations can proactively prevent data leaks, insider threats, and compliance violations before they escalate. As security teams struggle with increasing data complexity, lineage provides the missing link to securing modern data environments.

02 Discovering and monitoring risks

Legacy security tools rely on outdated assumptions about where sensitive data resides, often leaving security teams blind to emerging risks. Data lineage shifts risk discovery from a passive, reactive approach to an active, real-time model—tracking data across endpoints, cloud apps, and users. This approach enables security leaders to identify high-risk behaviors, such as shadow IT usage and unauthorized sharing, before they lead to breaches. By leveraging lineage-based risk discovery, organizations can stay ahead of threats without waiting for compliance audits to expose weaknesses.

03 Adding context to classification

Traditional data classification methods struggle to keep pace with modern data environments, particularly when dealing with unstructured, encrypted, or dynamically shared data. Data lineage enhances classification by analyzing the full context of data creation, movement, and usage patterns. This approach eliminates reliance on error-prone user-based labeling and instead leverages broader context to determine data sensitivity. With contextual classification, security teams can confidently enforce policies that align with real-world business processes, reducing false positives and compliance risks.

04 Accelerating investigations and reducing response time

Security teams often face delays in investigations due to fragmented data visibility and limited forensic insights. Data lineage streamlines investigations by reconstructing the complete history of data access, sharing, and modifications in a single view. This capability accelerates incident response, allowing teams to quickly identify the root cause of security events and take immediate corrective action. Whether responding to insider threats, unauthorized data exfiltration, or regulatory violations, lineage-driven security significantly reduces dwell time and operational burden.

05 Driving productivity and security

Security should not come at the cost of productivity. Unlike restrictive security models that block legitimate workflows, data lineage enables security teams to enforce policies, while factoring in business needs. This allows organizations to maintain regulatory compliance and data protection without disrupting daily operations. As organizations embrace AI, SaaS, and remote work, lineage-based security provides a scalable, intelligent approach that balances business agility with strong data protection.

About Cyberhaven

Cyberhaven is the AI-powered data security company revolutionizing how companies detect and stop the most critical insider threats to their most important data. Until now, data security products were limited to scanning data content and looking for specific user actions. Our AI technology analyzes billions of workflows to understand every piece of data within an organization, when it's at risk, and takes action to protect it. It's like nothing that's come before and protects data like nothing else. For more information, visit cyberhaven.com.