cyberhaven

THE BROKEN PERIMETER:

Why stopping insider risk requires data lineage and AI

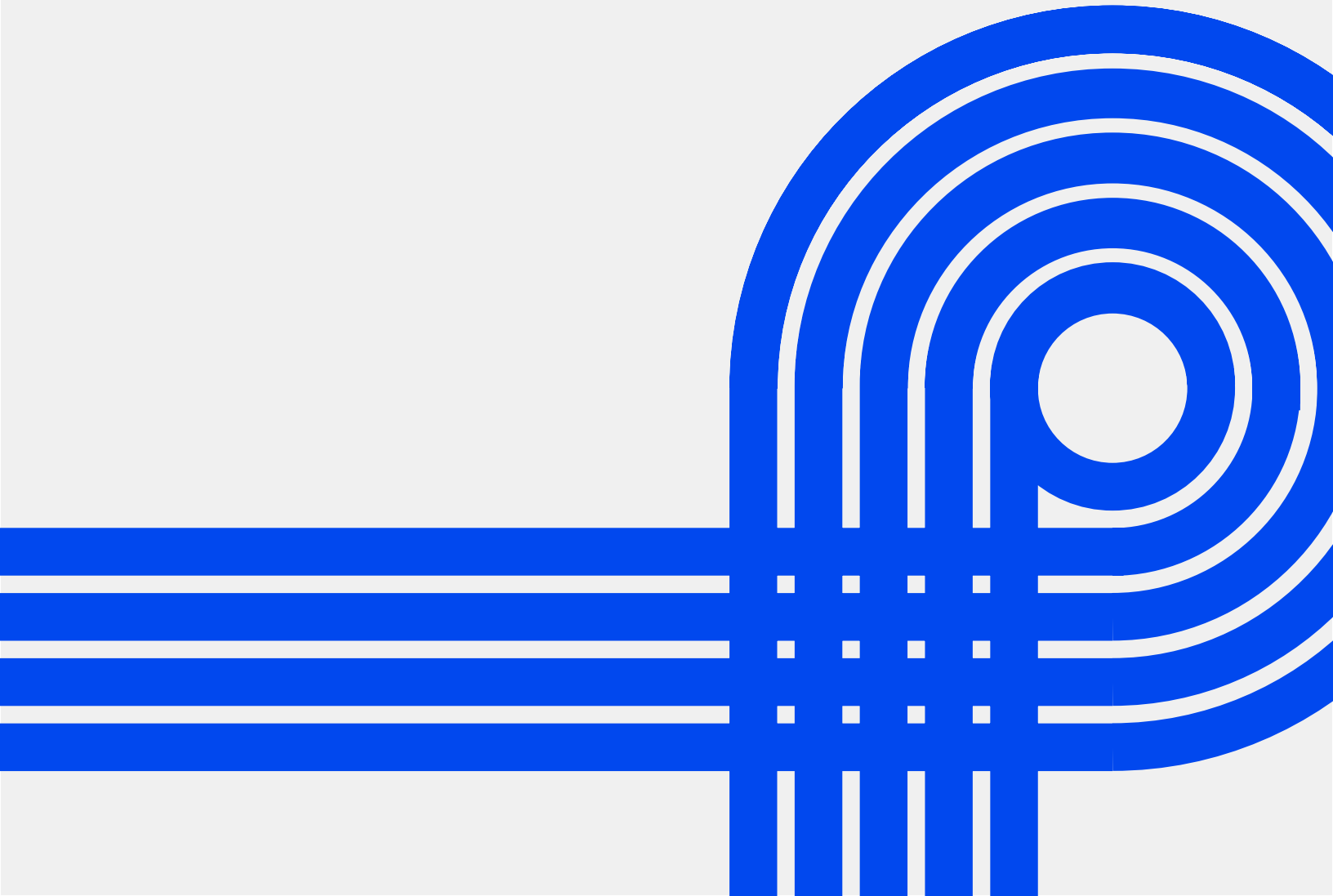


Table of Contents

Executive Summary 3

The fragmentation crisis: data has escaped the file 4

Why legacy tools come up short 6

The lineage advantage: seeing the whole story 7

AI: making lineage actionable at scale 8

The new reality: data without borders 9

A new model for data security 10

Conclusion 11

Executive Summary

The file perimeter is broken. Data no longer lives neatly inside databases, file shares, or SaaS applications. It fragments, multiplies, and moves through emails, screenshots, AI prompts, spreadsheets, and chat messages. Each fragment contains pieces of your most valuable intellectual property, but traditional security tools are blind to it.

Today over

80% of critical data exfiltrated by employees consists of fragmented or derivative data,

according to an analysis by Cyberhaven Labs. Legacy insider risk and DLP solutions weren't built for this world. They look for patterns in content and rely on static policies, blind to how data flows and transforms. In contrast, data lineage and AI enable security teams to see the full story — tracing where data came from, how it's changed, and why it's moving.

1

The fragmentation crisis: data has escaped the file

In the past, securing data meant protecting files and systems. Today, the data itself has become fluid. Cyberhaven research shows that over half of exfiltration incidents involve unstructured intellectual property moving through multiple systems and channels before leaving the organization, including personal cloud accounts, removable drives, and generative AI tools.

Worse yet, research shows that

80% of critical data exfiltrated by employees consists of fragmented or derivative data,

according to an analysis by Cyberhaven Labs.

Data fragments are partial copies of the original data. Data fragments can occur when a user executes the copy/paste function from a Word document to a Slack channel or when a service syncs some data, but not all, from its source to the destination.

Data derivatives are copies of the original data presented in a format different from the original. Derivatives can be partial or complete copies of the original data. Derivatives can also include fragments of data.

Fewer than 20% of organizations can trace the full path of sensitive data. These figures highlight how fragmentation undermines traditional controls and reinforce why legacy DLP and insider risk tools can't keep up with the modern data landscape.

On average, **sensitive data is copied six times before it is exfiltrated,** making traditional policy enforcement nearly impossible. **Over 60% of incidents involve data fragments that no legacy DLP would flag**, leaving organizations unaware of critical data exposure until it's too late.

<20% of organizations can trace the full path of sensitive data.

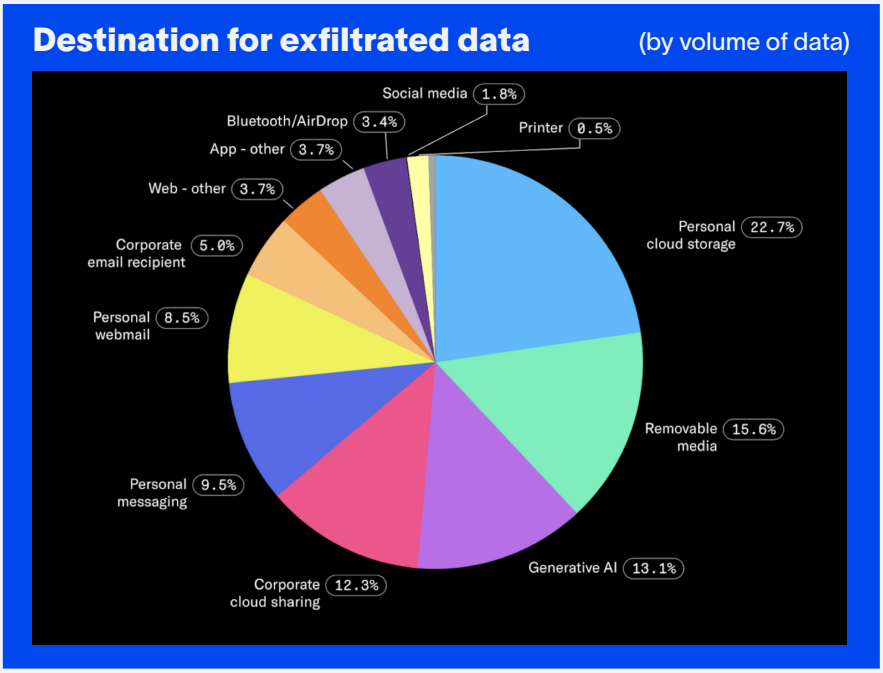
Sensitive data **is copied 6x** before it is exfiltrated.

>60% of incidents involve data fragments that legacy DLP would miss.

Each of these fragments—copied into a slide deck, pasted into an AI prompt, or exported into a CSV—carries business-critical context invisible to conventional tools. Once data crosses even a single transformation step, its original classification is lost.

The result: your sensitive data no longer exists in one place, one file, or one format.

In the past, securing data meant protecting files and systems. Today, the data itself has become fluid. As highlighted in Cyberhaven's Q1 2024 Insider Risk Report, **"75% of exfiltrated data is unstructured intellectual property, not regulated PII or PCI."** The same report found that **employees exfiltrate data through more than a dozen different vectors—most commonly personal cloud accounts (22.7%), removable drives (15.6%), and generative AI tools (13.1%).**



Each of these fragments—copied into a slide deck, pasted into an AI prompt, exported into a CSV—carries business-critical context that's invisible to conventional tools. Once data crosses a single transformation step, its original classification is lost.

The result: your sensitive data no longer exists in one place, one file, or one format.

2

Why legacy tools come up short

The limitations of legacy tools make clear the need for a new approach. As traditional DLP and insider risk systems struggle to keep pace with how data moves today, organizations require visibility that follows data across its entire journey—that's where data lineage becomes essential.

Legacy DLP tools were built for a world where sensitive data lived in databases and documents—static, structured, and easily labeled. Those assumptions no longer hold.

Lack of holistic visibility

Most traditional insider threat tools look at activities, build a statistical model, and then look for anomalies. While this might find some risky behavior, these tools lack visibility into the content, resulting in long tuning periods and high false positive rates.

Data loss prevention tools at least look at content, yet they still miss most of the picture. Content scanning looks for keywords, regexes, or file types. It can't tell whether a line of code came from an internal repository or the open web, or whether an employee is uploading source code or a blog draft. As data takes on different forms, from design files to screenshots to AI outputs, content alone tells security teams almost nothing.

Traditional DLP tools miss roughly 70% of incidents involving derivative or transformed data because they depend on static content matching rather than understanding how information moves and changes. This quantified failure rate highlights why a shift toward lineage-based visibility is essential.

They depend on fixed perimeters

Legacy DLP assumes data stays within managed systems. But according to Cyberhaven's Q1 2024 Insider Risk Report, **employees are 400% more likely to move data via Bluetooth or AirDrop when offsite and 254% more likely to use removable media.** In hybrid environments, the boundary between corporate and personal systems no longer exists.

They generate noise instead of insight

Without understanding data's history or intent, traditional tools flood security teams with false positives. One U.S. manufacturer reported spending thousands of analyst hours each quarter investigating incidents that were legitimate collaborations.

The takeaway:

DLP and insider risk tools see static snapshots. Data lineage sees motion.

3

The lineage advantage: seeing the whole story

Data lineage changes the security model by mapping every event in a data object's life—from creation, through transformations, to movement across people, apps, and systems.

This capability transforms how organizations detect and respond to insider risk:

- **Visibility into derivative data:** Track sensitive information even as it's copy-pasted, embedded in slides, or used to train AI models.
- **Context-aware classification:** Determine sensitivity based on origin, creator, and use, not just keywords or file type.
- **Automated investigations:** Trace an incident across the entire organization—who touched the data, how it changed, and where it went.

Unlike static scanning, lineage reveals not just what happened, but why it happened—the context behind every data movement.

4

AI: making lineage actionable at scale

AI transforms lineage data into action. It interprets the complex web of data movements, surfacing the risks that matter most. For example, AI can detect when sensitive source code copied from an internal repository later appears in an AI prompt or personal cloud account—flagging an incident that traditional tools would miss.

By applying context-aware models, AI highlights abnormal behavior, prioritizes the most severe events, and reduces alert fatigue. Together, lineage provides the map and AI provides the analysis, enabling security teams to respond faster and with clearer insight into intent and impact.

Tracing every data event across millions of users and billions of workflows is impossible for humans. AI makes lineage operational by:

- **Detecting anomalies:** Machine learning models can identify deviations in normal data movement patterns.
- **Automating triage:** Instead of 10,000 low-value alerts, AI can summarize incidents and prioritize those with clear risk indicators.
- **Understanding intent:** By analyzing behavior, AI distinguishes between legitimate collaboration and exfiltration.

Linea AI combines semantic risk detection with contextual investigation, automatically producing detailed reports that explain what happened, why, and what to do next.

5

The new reality: data without borders

Work is no longer confined to an office or network. According to Cyberhaven's Q1 2024 Insider Risk Report:

Office-based employees are **77% more likely** to exfiltrate sensitive data than remote workers.

When those office-based workers log in from offsite, they're **510% more likely** to exfiltrate data—the riskiest moment for corporate information.

Meanwhile, layoffs and job transitions amplify the threat: in the 24 hours before a layoff, data exfiltration spikes **by 720%.**

This fluid world—of hybrid work, generative AI, and cloud sprawl—has permanently erased the file perimeter. The only constant is movement.

6

A new model for data security

Modern insider threat programs require not just detection but a lifecycle approach: identify, protect, detect, respond, and recover. Data lineage and AI strengthen every phase by providing visibility into who accessed data, how it moved, and whether its use was legitimate.

Data lineage provides:

Legacy DLP was designed for a world that no longer exists. The only way to protect fragmented, fast-moving data is to understand its lineage—to know where it came from, how it's evolved, and what it means.

Data lineage provides:

- End-to-end visibility across every movement of data.
- Contextual understanding of user intent and data value.
- AI-driven detection and investigation at enterprise scale.

This is not just an incremental improvement. It's a redefinition of data security—from static control to dynamic understanding.

Conclusion

The broken perimeter demands action. Organizations can no longer rely on outdated tools to manage modern data risks. By adopting lineage-based security powered by AI, security teams gain continuous visibility and context to stop insider threats before data escapes.

Your data now lives everywhere—woven through emails, AI prompts, and shared documents. Traditional insider risk and DLP tools, built to guard walls that no longer exist, can't protect what they can't see. Data lineage and AI reveal how data truly moves and why it matters.

Organizations that embrace lineage-based data security will see everything. Those that don't will only see what's already gone.

Practical next steps

To operationalize lineage-based security, establish a cross-functional Insider Threat Program Office that includes HR, Legal, Security, and IT. Implement an incident response playbook aligned with the insider threat framework: triage alerts, investigate using lineage insights, contain risk in real time, and perform post-incident reviews to refine policies. Continuously measure response speed and detection accuracy to strengthen your posture as data and AI evolve.

cyberhaven

About Cyberhaven

Cyberhaven is reimagining data security. Until now, data security products have been limited to scanning data content or looking for specific user actions. Our AI-enabled data lineage technology analyzes billions of workflows to understand every piece of data within an organization, identify when it's at risk, and take action to protect it.

To learn more, visit **cyberhaven.com**