CONCENTRIC Ai

Data Risk Report

2H 2025

Table of Contents

Introduction

Data is growing exponentially, and not just in volume but also in velocity and variety, making it more difficult to track and protect. Even midsize organizations will easily accumulate tens or hundreds of millions of data records. And this rapid growth is seen in Concentric AI's customer base with scans of new customer data records for the first half of 2025 numbering in the hundreds of millions.

Data is more distributed than ever, flowing across cloud and on-premises environments, SaaS applications, mobile endpoints, and hybrid work environments. It's also growing more complex, often unstructured, embedded in everything from documents to emails to collaboration tools and beyond, and constantly in motion.

The use of generative AI (GenAI) is accelerating fast—and adding new data security and privacy concerns for organizations. Enterprises are also having to protect against "shadow AI," which is where employee use of unsanctioned AI applications can expose sensitive corporate data.

AI assistants like Copilot are accessing sensitive data. Enterprises are increasingly concerned not only about the risk of their sensitive data being shared with unauthorized users but also about the resulting legal, financial, and reputational consequences.

These combined risks create significant challenges for the security teams responsible for ensuring that sensitive data is both adequately protected and accessible for legitimate business operations. They are finding it increasingly difficult to locate the data, understand its context, and apply the right controls to prevent it from slipping through the cracks.

This edition of Concentric AI's semi-annual Data Risk Report reviews data analyzed and aggregated from a representative set of Concentric AI customers across multiple industry verticals for the first half of 2025 in order to identify broad and industry-specific trends.

Executive Summary

When users, applications, or systems have more access than they need to perform their function, data is more vulnerable. Oversharing and excessive permissions continue to drive increased risk for organizations' sensitive data.

Across all industry verticals in the sample, organizations averaged:

3M+
Sensitive data records shared with external domains

2M+
Sensitive data records shared without restrictions

400K+
Sensitive data records shared with personal accounts

Microsoft Copilot accessed on average almost three million sensitive data records per organization.

Organizations averaged more than 3,000 user interactions with Copilot.

More than one-third of data records with Anyone links contained sensitive information.

Close **to 60% of all data records** being shared organization-wide contained sensitive information.

At least half of all data shared with personal accounts contained sensitive information.

Organizations averaged 10 million duplicate data records and almost seven million data records older than 10 years.

- In the government & education sector, more than one-third of all data records were older than 10 years.

Organizations averaged more than four million orphaned data records and more than two million inactive user data records.

- Orphaned data accounted for 10% of retail organizations' total data.
- Inactive user data accounted for almost 10% of total data in the government & education sector.

Analysis and Key Findings

Generative AI (GenAI) is rapidly becoming one of the most transformative technologies of our time. Its use is growing across all industry verticals with aggregate revenue for the market projected to reach $85 billion by 2029.[1]

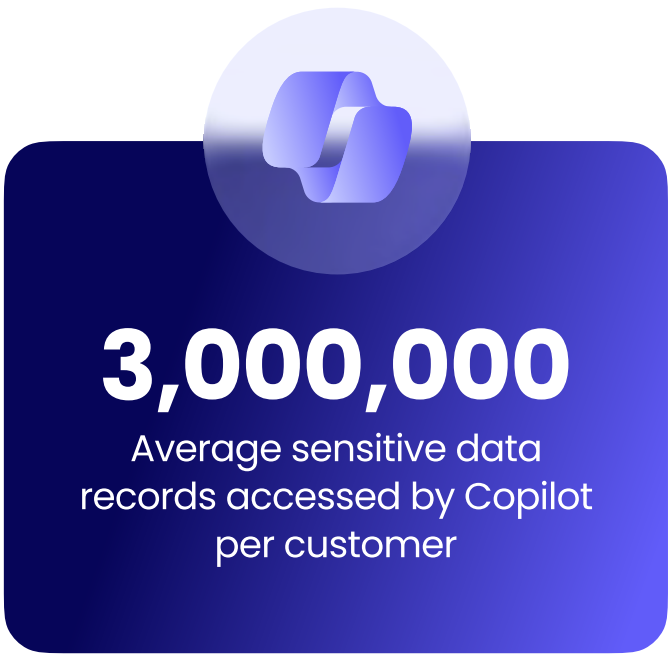
GenAI Is Here and Everyone's Using It

AI assistant tools like Microsoft Copilot are changing how we make decisions, solve challenges, create content, and engage with others. Sales teams have instant answers, customer service chatbots are tackling FAQs without breaking a sweat, and even security teams are using it to speed up incident response. But along with the tangible benefits there are significant data security risks. If data isn't properly classified and if the right protections are not in place, Microsoft Copilot can unintentionally expose sensitive company data—payroll records, intellectual property, or even that confidential layoff memo.

Copilot Activity

Across the data sample, the Semantic Intelligence platform observed Copilot accessing an average of almost three million sensitive data records per customer during the first half of 2025.

The data showed organizations averaging more than 3,000 user interactions with Copilot. During these interactions, Copilot could potentially modify or share data it was accessing with other users or systems.



[1] S&P Global

Public AI Tools

Public AI tools like ChatGPT and Claude AI offer material productivity and innovation advantages for organizations, but without the right governance they also introduce significant security and compliance risks for sensitive data.

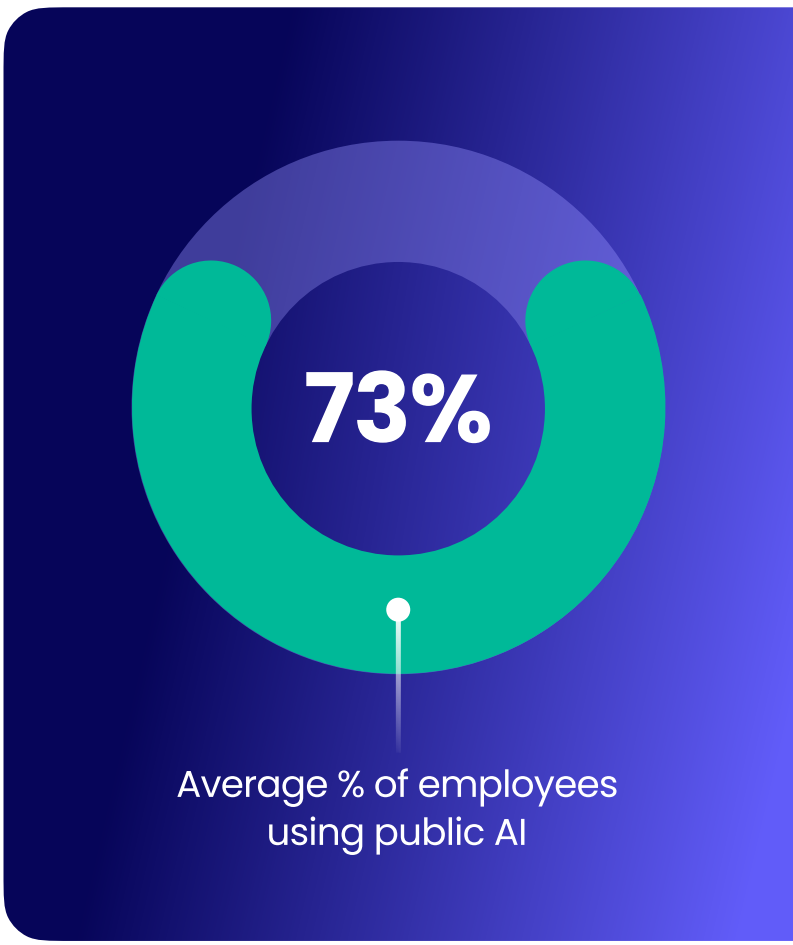
When employees share data with unsanctioned public AI applications, aka shadow AI, the organization no longer has visibility into—or control over—what happens to the data: who can access it, how it can be used, or where it can go. Confidential customer information, financial records, product roadmaps, intellectual property, source code, and more can quickly end up in the wrong hands.

Public AI Usage

Concentric AI sampled data from financial services customers over a 30-day period and found an average of 73% of employees used public AI applications.

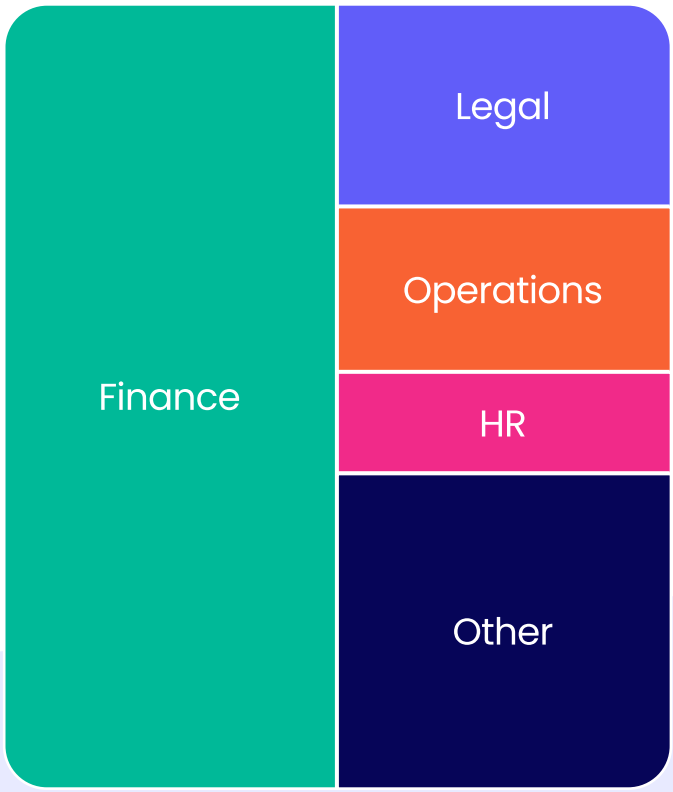
On peak usage days during this period, almost half of all employees (48%) were using public AI.

During the 30-day period, users submitted an average of 6.5 prompts per day, but on peak usage days, this number climbed to 20 prompts per user.



Types of Data Being Shared

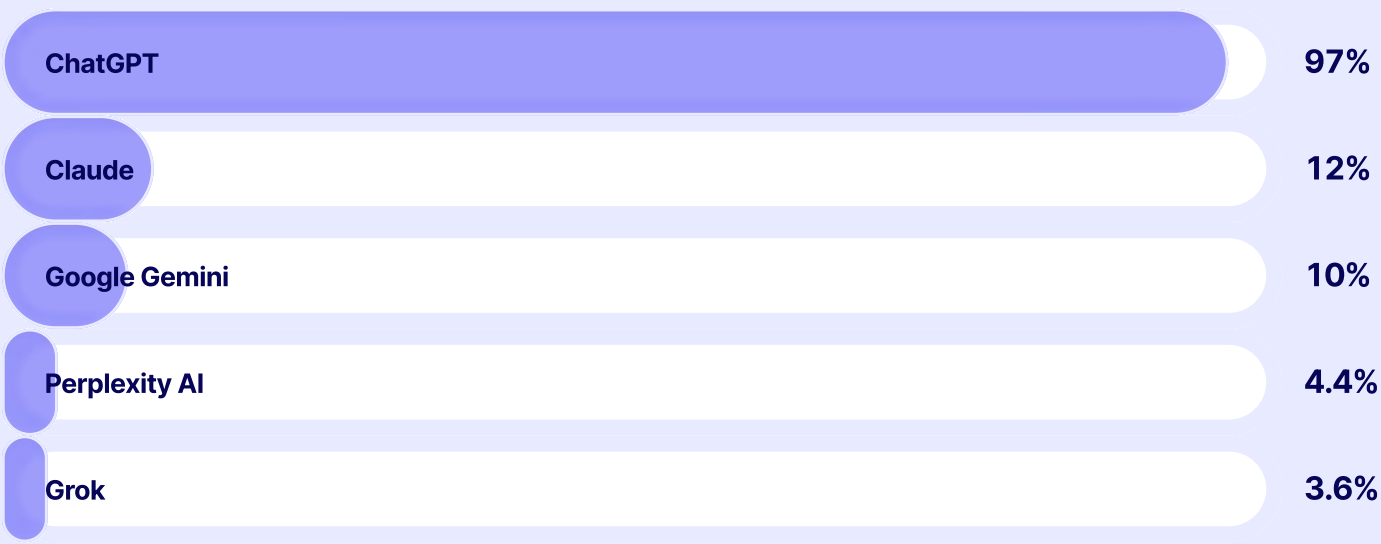
Analysis of the types of data being fed into public AI applications by the financial services organizations in the data sample revealed data from across multiple categories.
Not surprisingly, the most common category was financial (48.6%). This was followed by legal (13.1%), operations (10.7%), and HR data (6.9%). The remaining 20.7% of data being shared comprised a mix of sales, insurance, product, and other miscellaneous categories.



Most Used Public AI Applications

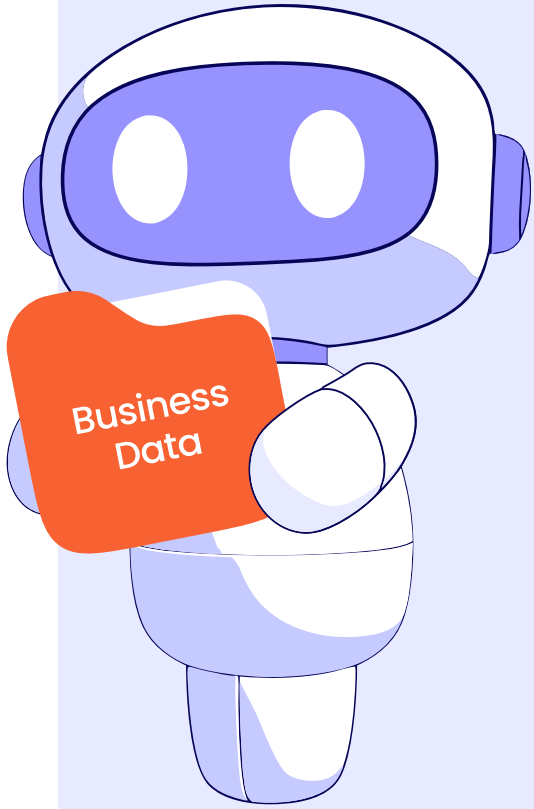
Organizations need clear visibility into all the public AI tools their users are accessing, not just the most popular tools but also the newer, more niche tools. Over the 30-day period, the sampled financial services organizations used an average of 19 public AI applications.

By far the most commonly used application was ChatGPT (97%). It was significantly trailed by Claude (12%) and Google Gemini (10%), which in turn were trailed by Perplexity AI (4.4%) and Grok (3.6%).



ANALYSIS AND KEY FINDINGS

Excessive Permissions Invite Risk



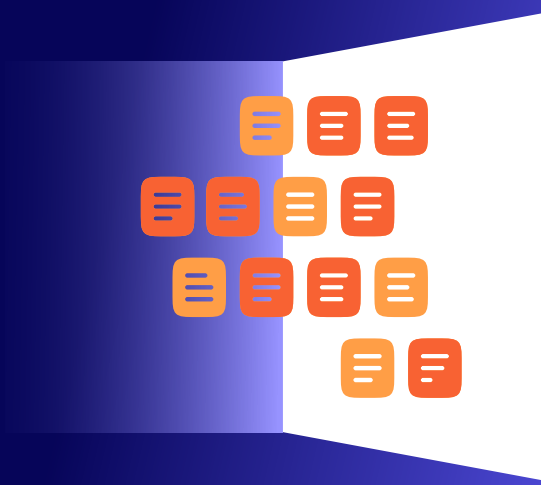
Data sharing has never been easier, and it brings clear benefits: better collaboration, faster decision-making, operational efficiencies, transparency and accountability; but uncontrolled sharing also increases risk.

Internal and external users as well as third-party systems and applications are accessing data connecting from many different devices and locations, but every access point is a vulnerability, and the more data is accessed, the greater the risk of it falling into the wrong hands.

Sharing With External Domains

Across all organizations in the sample dataset, in the first half of 2025, on average more than three million sensitive data records were shared outside the organization. This equates to 55% of the total number of files being shared externally by organizations.

By far the highest total number of files being shared with external domains (18%) was within the manufacturing vertical. The highest number of data records being shared externally and containing sensitive information (73%) was within the financial services vertical, followed by government and education (61%) and then technology (53%).



3,000,000+

Average number of sensitive data records being shared externally

Anyone Links

Anyone links allow any user, whether inside or outside the organization, to access Microsoft 365 content without needing to sign in or authenticate, and this puts sensitive data at risk of being unintentionally shared.

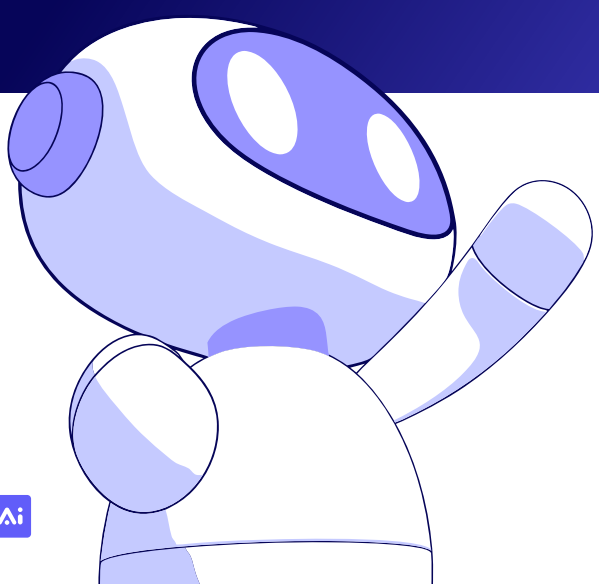
Looking across all organizations in the dataset, we see that, on average, more than one-third of all organizations' data records with Anyone links contained sensitive information.

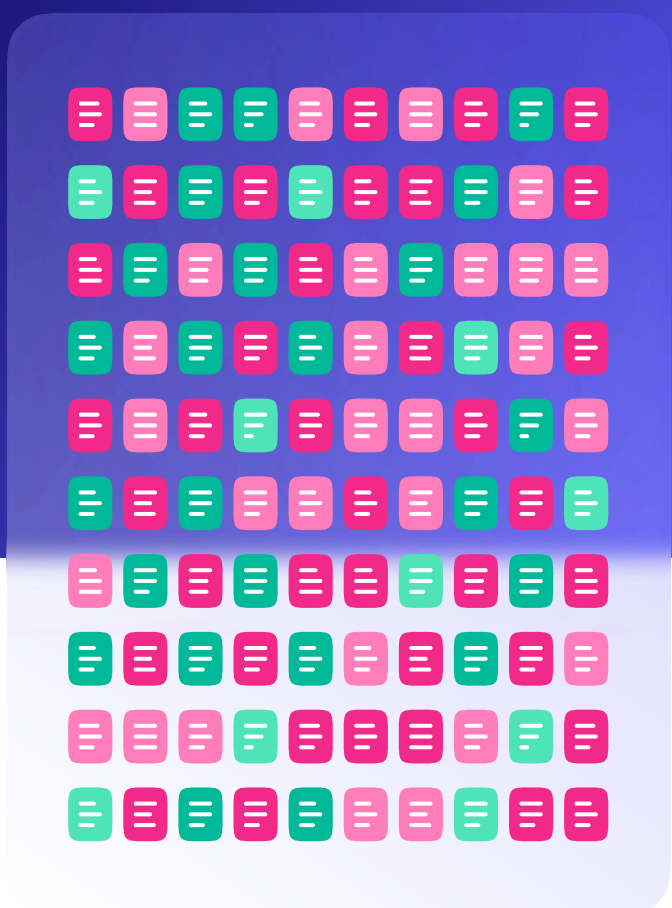
Organizations in the healthcare vertical shared the highest percentage of data records containing sensitive information using Anyone links (66%).

At least half of all data records being shared with Anyone links by organizations in the technology, retail, and financial services verticals contained sensitive information.

66%

of healthcare records with Anyone links contained sensitive information



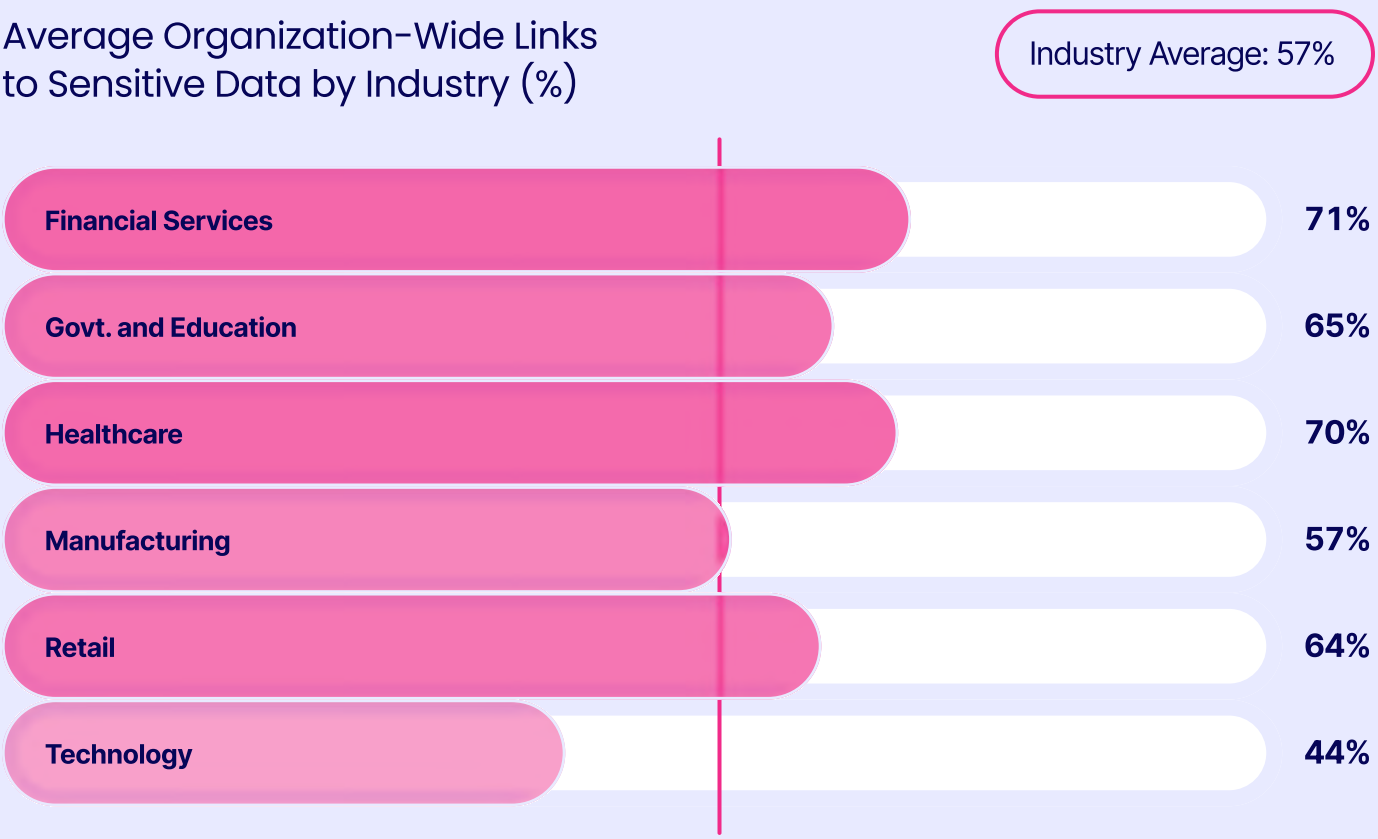


Organization-Wide Sharing

Organization-wide links allow anyone in an organization to access the data. However, while granting access to everyone, regardless of role, department, or business need, is convenient for collaboration, it also introduces significant risk if data is not properly protected.

Across the dataset, an average of 57% of all data records shared organization-wide contained sensitive information, while organizations in the financial services and healthcare sectors were even higher—averaging around 70%.

Organizations in the government and education, retail, and manufacturing verticals averaged sharing more than half of their sensitive data records organization-wide.

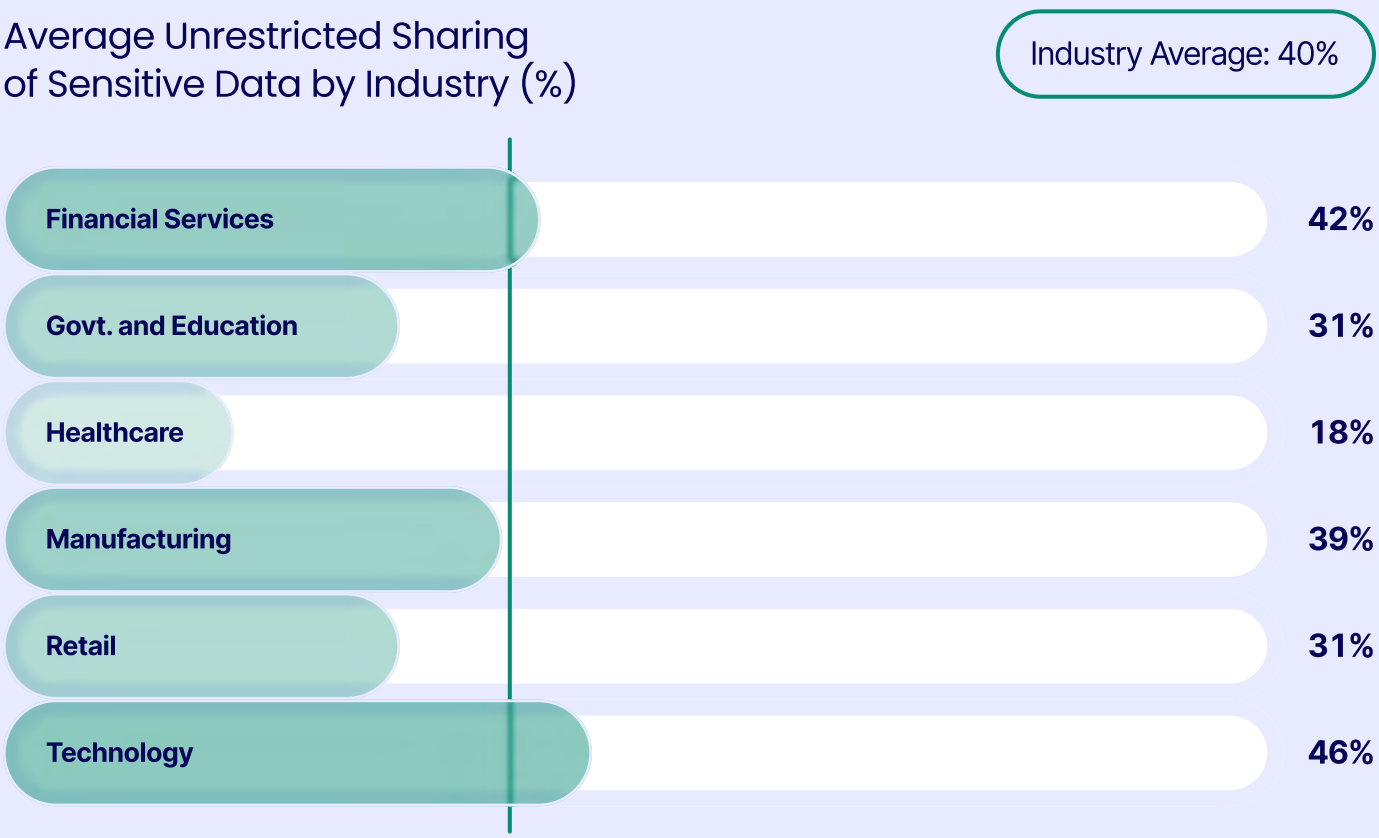


Unrestricted Sharing

When everyone in an organization has permission to access data, this increases the risk of the organization losing control over how or where the data is used.

Across all organizations in the sample dataset, organizations shared an average of two million sensitive data records with no restrictions, which was close to 50% of all data records being shared without restrictions.

Organizations in the technology vertical shared the highest percentage of sensitive data with no restrictions (close to half of all data shared with no restrictions), followed by the financial services and manufacturing verticals (around 40%).

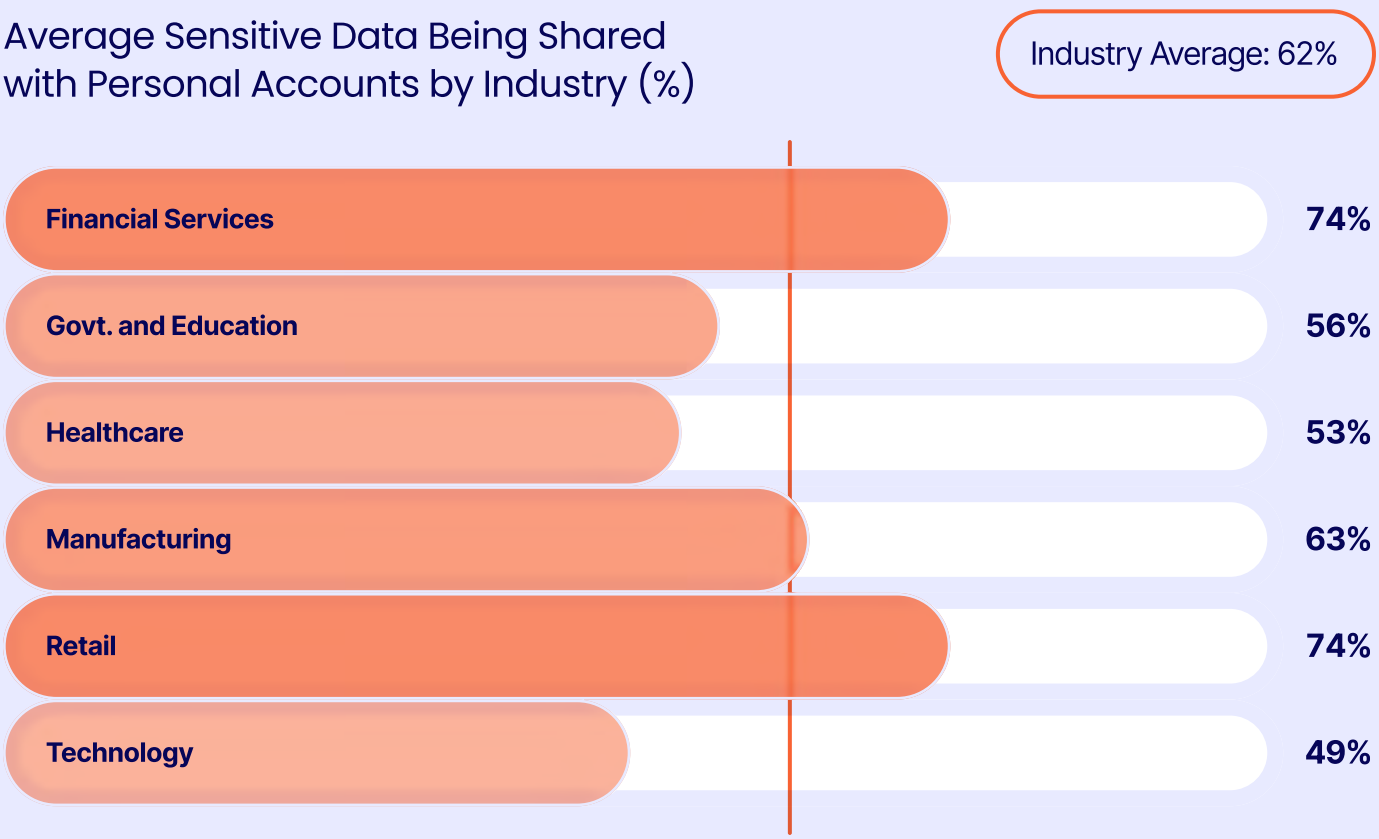


Sharing Data Records With Personal Accounts

Across the sample dataset, organizations shared an average of approx. 400K data records containing sensitive information with personal accounts, which was 62% of all data records shared with personal accounts.

Across every vertical, at least half of all data records being shared with personal accounts contained sensitive information.

For organizations in the financial services and retail verticals, three out of four data records shared to personal accounts contained sensitive information.



Data Clutter Undermines Productivity and Efficiency

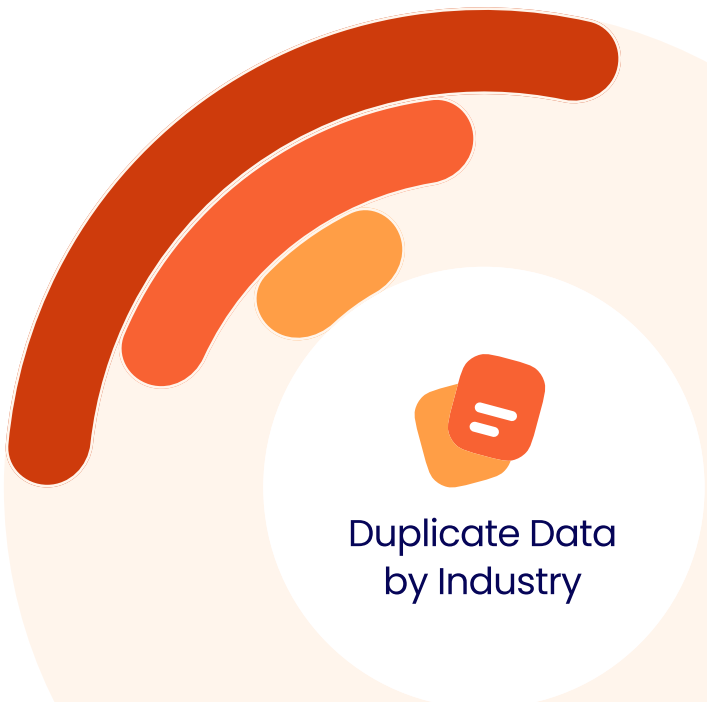
With sensitive data sprawled everywhere and growing exponentially, it's not uncommon for organizations to have multiple copies of the same file or to be hanging onto data way past its retention period.

But retaining stale data or duplicate data increases storage costs, hikes cyber insurance premiums, creates versioning chaos, and impacts productivity. Data hygiene is essential for enabling accurate, reliable decision-making and better operational efficiency. The previous Concentric AI Data Risk Report analyzed unstructured data records from companies in the technology, financial, energy, and healthcare industries and found one in three to be duplicates or near duplicates. This report analyzed data from companies across six verticals and discovered that on average one-fourth of their total data was duplicate.

Duplicate Data

Duplicate and near duplicate data is frequently improperly classified, has incorrect permissions, and is stored in unauthorized locations, which puts it at significant risk. How do you adhere to regulations such as the Privacy Act and protect data from unauthorized disclosure when you have no idea how many copies you have out there to begin with?

Across all verticals in the sample, organizations averaged a total of 10 million duplicate data records.



Greater than 10%: Financial Services

Greater than 20%: Manufacturing, Technology

Greater than 30%: Govt. & Education, Retail

Stale Data

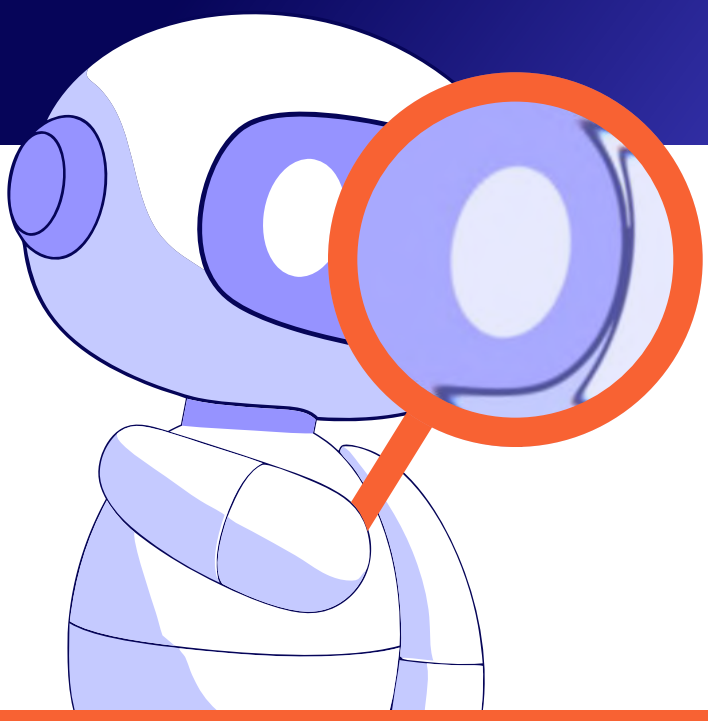
Even companies with clearly defined data retention policies end up holding onto thousands of data records that should've been deleted or archived years ago. Without visibility into how old data is, or when it was last modified, employees are reluctant to hit the "delete" button in case they accidentally get rid of data the company still needs.

In the first half of 2025, organizations across all verticals in our sample averaged almost seven million stale data records (i.e., data that is more than seven years old). On average, stale data made up approximately 12% of organization's total data records.

The manufacturing vertical had the highest percentage of stale data at approximately one-fourth of all data records.

≈7M

Average stale data records



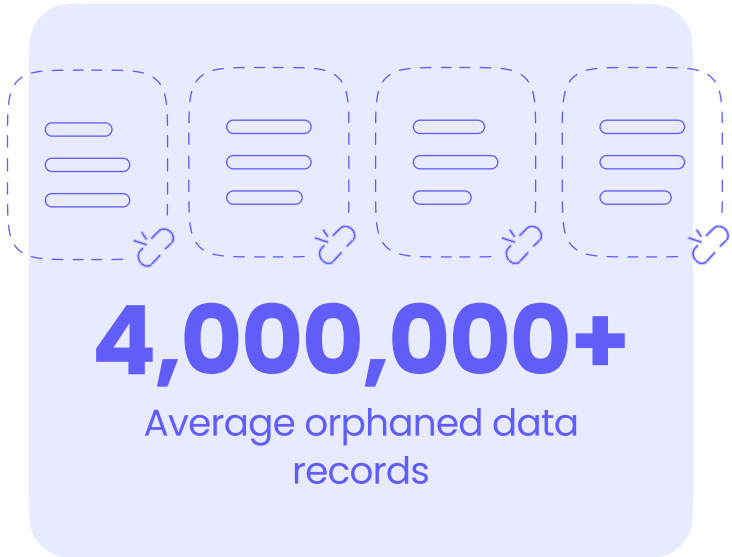
Unclaimed Data Is Unsafe Data

When data no longer has a clear system or owner responsible for its management, it is referred to as orphaned and inactive user data. Because this data is difficult to track, many organizations have little to no idea of how much they have.

Orphaned Data

If data is not properly migrated, archived, or deleted, it is considered orphaned.

Across all verticals, organizations averaged more than four million orphaned data records, with the amount in the retail sector—close to 10% of all data records.



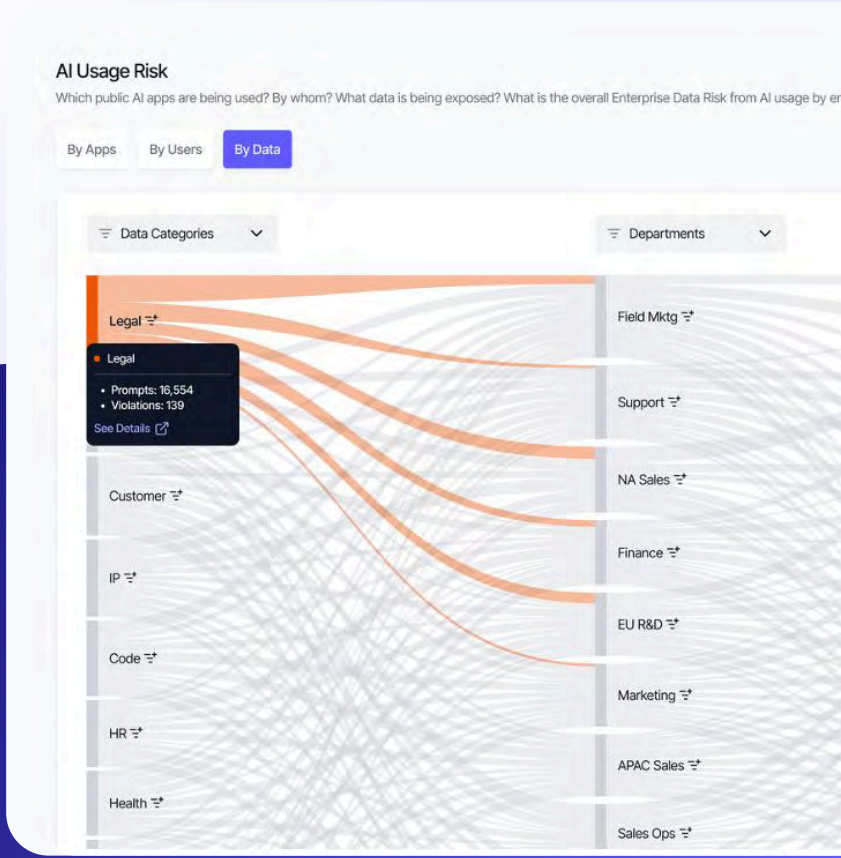
Inactive User Data

Inactive user data refers to data that can no longer be associated with an assigned owner; for example, because the owner left the company or because an account has become dormant.

Without clear ownership, the data is unlikely to be actively managed or monitored, which introduces significant security risks, particularly if the data contains sensitive information.

Across all verticals, organizations averaged more than two million inactive user data records. For organizations within the government & education vertical, this number accounted for almost 10% of their total data records.

Protecting Data Wherever It Lives and However It Travels



Concentric AI's Semantic Intelligence platform provides intelligent data security governance that aligns with organizational zero-trust policies and business goals and protects data whether it's at rest or in motion and across all the GenAI applications users interact with.

The platform is unique in its capability to discover and categorize data. Not only can it tell whether there is sensitive content in a data record, but it will also understand the type of record it is.

This is because its deep learning capabilities allow for a contextual understanding of data, which results in far more accurate classification and access policies.

The platform performs ongoing monitoring to identify data risks such as excessive permissions, inappropriate sharing, unclassified or misclassified data, stale and duplicate data, and more, and to help organizations maintain compliance across cloud and on-premises environments.

It can also take actions to remediate risks, including classifying data, managing permissions, relocating sensitive data, and deleting, archiving, blocking, or masking data.

This includes protecting data across multiple GenAI scenarios. From discovering and revealing shadow GenAI, to protecting sensitive data from being revealed to unauthorized users and tracking all user interactions with Copilot—the Semantic Intelligence platform can determine which users accessed which sensitive files and help organizations understand the risks. The platform's GenAI capabilities also include enabling organizations to curate the data for their proprietary GenAI tools, so the models are trained only on what they should have access to.

Most importantly, Concentric AI doesn't just hand our customers a platform and leave them high and dry until renewal time. Our security pros are right there to provide the guidance and support needed to ensure our customers have what they need to build a robust data security program.

Report Methodology

All content in this report is based on live data gathered from the Concentric AI Semantic Intelligence™ platform between January – June 2025 and across a representative set of customers in the financial services, government & education, healthcare, manufacturing, retail, and technology industries.

During calculations, outliers were removed, and values were rounded to the nearest whole number. Values below 1,000 were excluded under the assumption that these areas of risk had been remediated.

Concentric AI is intelligent data security made easy. Its Semantic Intelligence™ platform uses context-aware AI to discover sensitive data, monitor risks, automate remediation, simplify compliance, and accelerate investigations. It delivers smart, targeted protection by understanding how data is used, shared, and exposed.

Concentric AI also offers managed services to keep security programs lean, scalable, and effective. This end-to-end platform protects data at rest, data in motion, and all the GenAI tools users interact with—so organizations can stay compliant, reduce exposure, and safeguard critical information wherever it lives and however it travels.