The 2024 State of

Data+AI Security.

By **Symmetry Systems**

Claude Mandy
Chief Evangelist
SYMMETRY

The Methodology

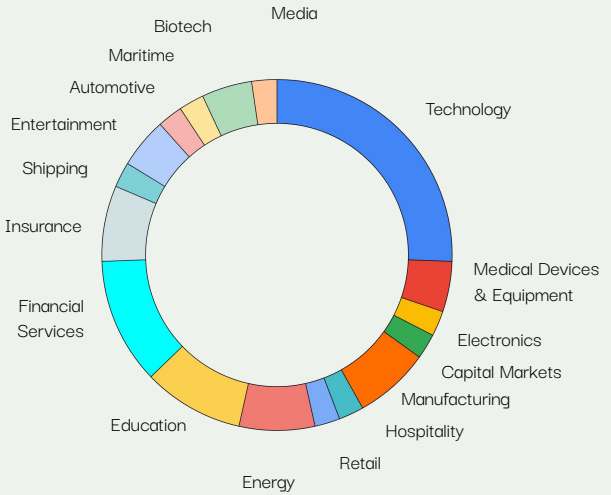
The Symmetry State of Data Security is based on insights gathered from over 50 organizational deployments of our platform. This includes ongoing customer deployments, as well point-in-time assessments up until Dec 2023. the scope and depth of data gathered varied depending on the specific deployment model and contractual arrangement.

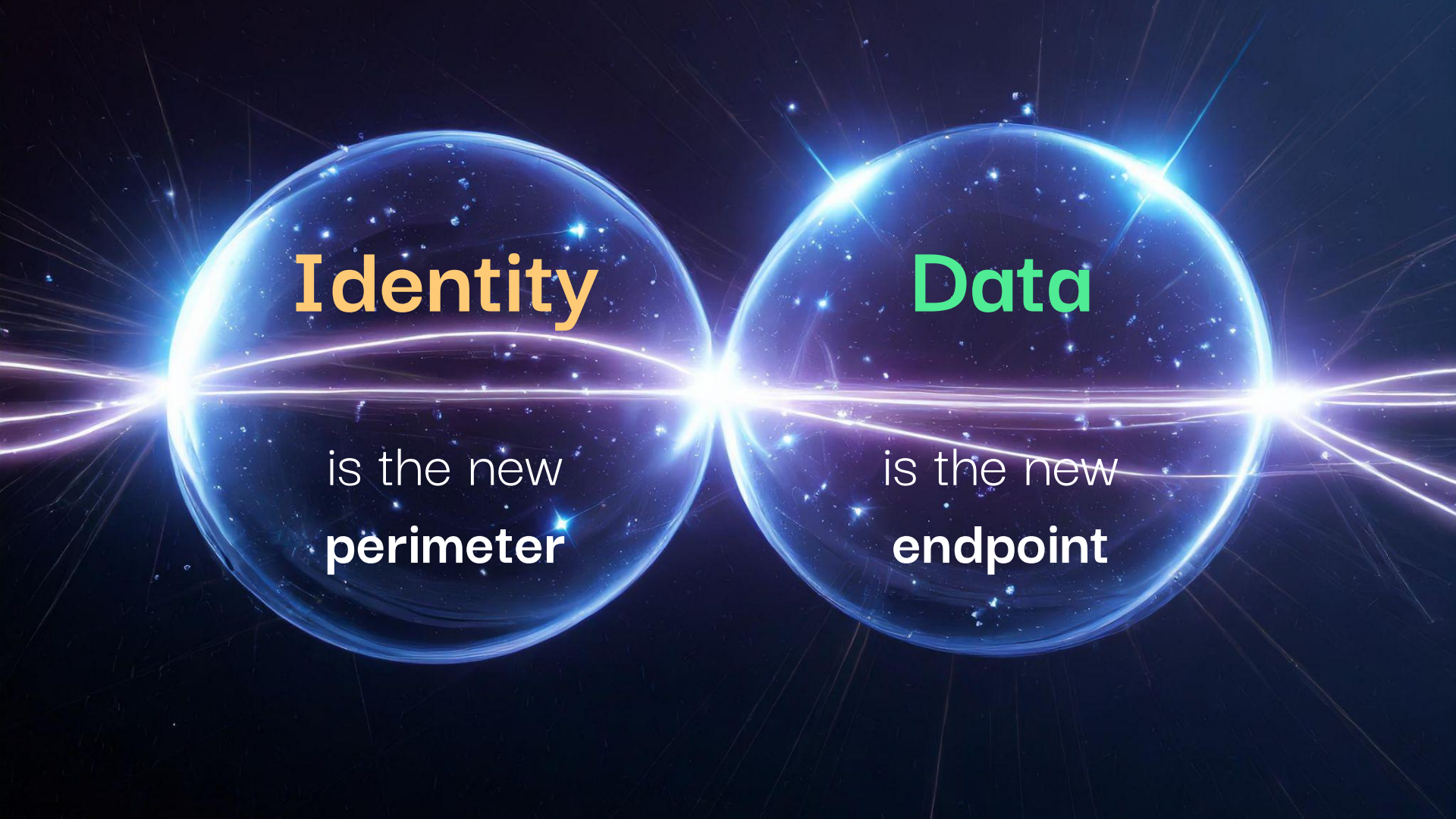
Only aggregated data from non air-gapped deployments was used in the creation of the study.

Steps were taken to ensure that data on each organization was anonymized and no sensitive information was indirectly or directly included in the analysis.

This process underscores Symmetry's dedication to to data privacy and security.

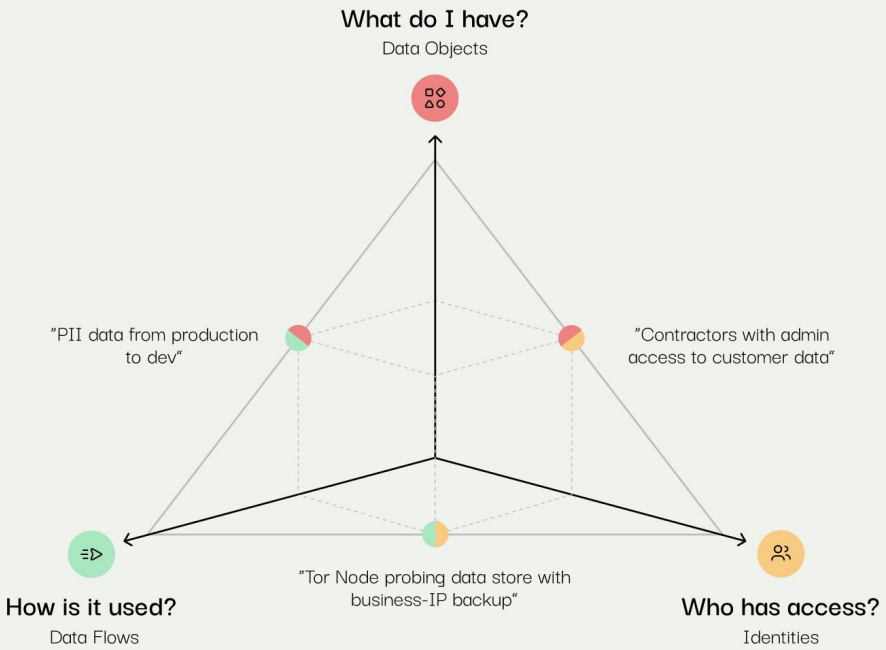
Industry Breakdown



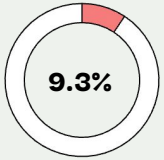

Identity
is the new
perimeter
Data
is the new
endpoint

Symmetry's Secret Sauce

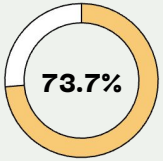
All questions about your data security posture fall into a combination of these three axes:



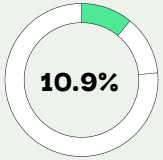
Key Findings



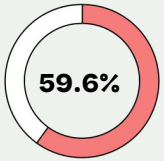
Data stores containing some form of **sensitive data.**



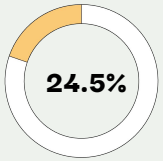
Organizations with at least 1 human account without MFA enabled & console access to their environment.



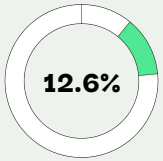
Cloud Accounts / Projects are unknown i.e. outside the organization's' control boundary & unknown to the organization.



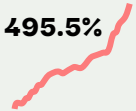
Data stores are considered **dormant** - no operations performed within last 90 days.



Dormant Identities with access to data - no operations performed within last 90 days.



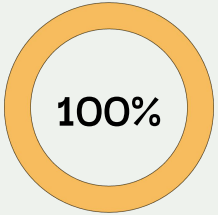
Cloud Accounts / Projects with access to data are known and external to the organization.



Avg. growth in number of dormant data stores **in the last 12 months.**

Organizations collecting & retaining personal information - with at least three types of personal information identifiers.

Organizations had at least one form of secret stored outside of a secrets manager within their environment



of organizations had **3+ types of PII** stored & used within their environment.

More than 20 Types of Personal Information Including:

- ☑ Full Name
- ☑ Email Address
- ☑ Phone Number
- ☑ Mailing Address
- ☑ Date of Birth
- ☑ Date of Death
- ☑ Place of Birth
- ☑ Social Security
- ☑ Driver's License
- ☑ Passport Number
- ☑ Financial Information
- ☑ Medical Conditions
- ☑ Illnesses
- ☑ Gender
- ☑ Racial or Ethnic Origin
- ☑ Genomic Data
- ☑ Sexual Orientation
- ☑ IP Address
- ☑ GeoLocation Data
- ☑ ITIN

Most common identifiers are:
name, email address, and phone number.

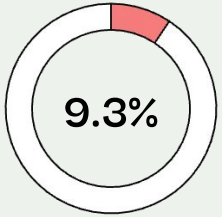
Useful intel for social engineering campaigns.

Personal Information can get Personal

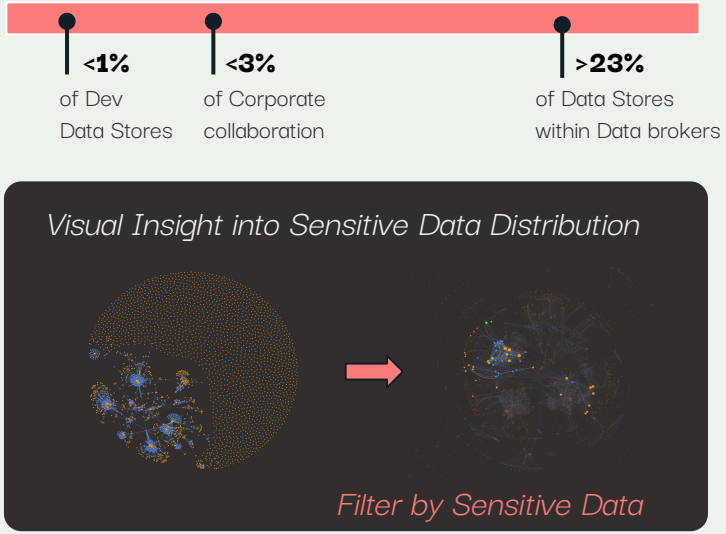
There is no doubt that organizations need personal information of their customers and employees to stay in business. Our findings reveal that 100% of organizations collect at least three pieces of personal information, **typically name, contact email, and phone number.**

However, some organizations go further in offering personalization of their products and services, amassing over **20 different types of personal data identifiers,** including other sensitive information such as birthplace, date of birth, race, eye color, and details about relatives.

The collection and retention of these broader organizations to implement robust data governance practices, including data minimization, purpose limitation, and strict access controls.



Data stores contain some form of sensitive data.



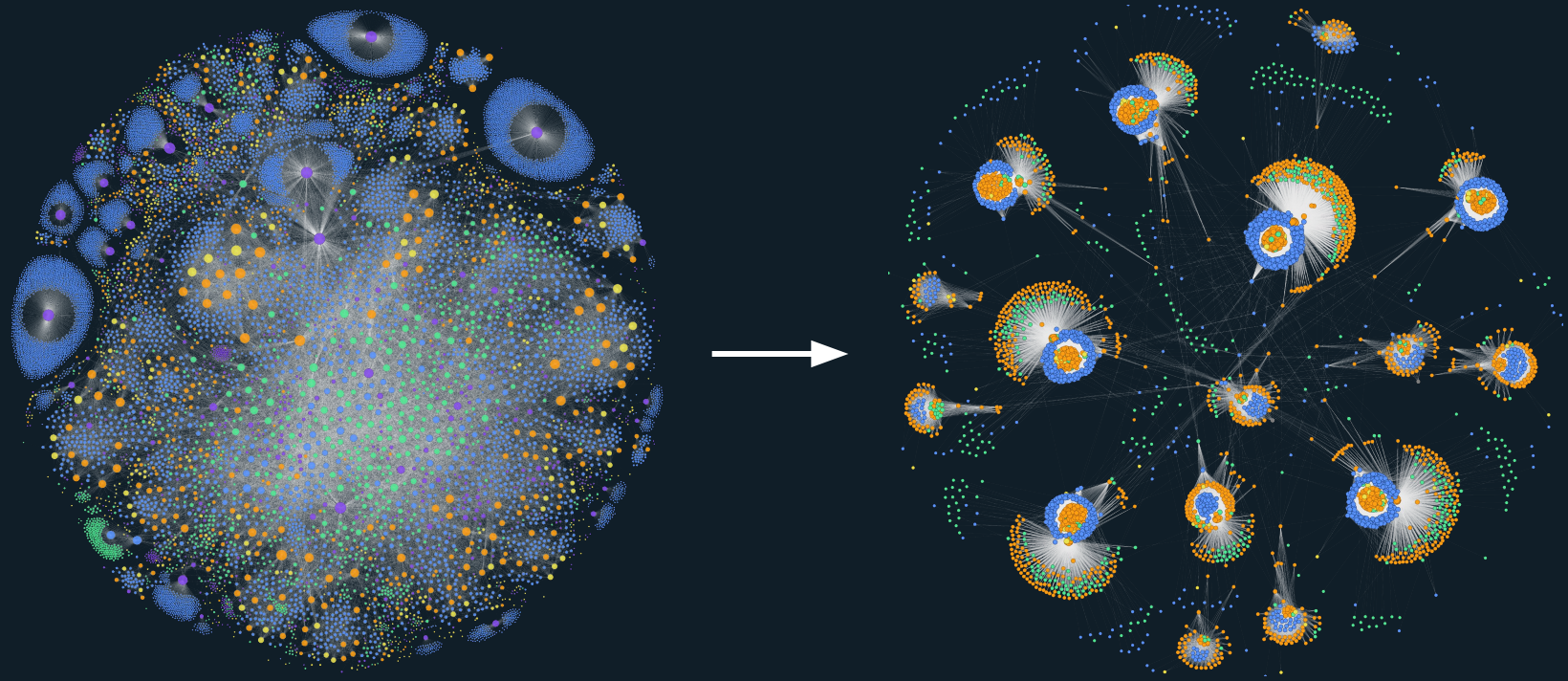
Industry & Env. Play Big Role in Sensitive Data Distribution

On average, **9.3% of data stores contain some form of sensitive data,** highlighting the need for robust data protection measures. However, this percentage can significantly increase for organizations operating as data brokers, where over 23% of their data stores may house sensitive information.

In contrast, data stores within corporate collaboration platforms like OneDrive, SharePoint, and Google Drive incorporate far less sensitive data per data store or bucket, with less than 3% of their data stores containing such information.

Pleasingly, the distribution of sensitive data drops to less than 1% in development environments for most organizations, indicating a better adherence to data segregation practices as a result of deploying DSPM.

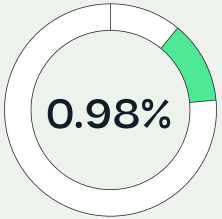
Cross Account Permissions Dilute Environment Segmentation



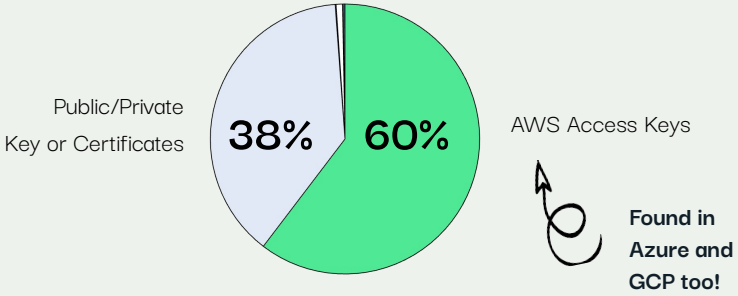
SYMMETRY

Users Services Roles Groups Data Stores

of organizations had 1+ form of secret stored outside of a secrets manager within their envs.



of sensitive data objects that Symmetry found within each environment are deemed “Secrets”

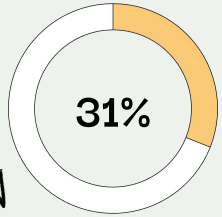


The Secrets Are Out There

1% of sensitive data objects or identifiers that Symmetry identified were classified as secrets, this illustrates simply the vast amount of data that is stored by data that is considered it was concerning that 100% of organizations had at least one form of secret stored outside of a dedicated secrets manager.

Of the secrets found, 38% were public/private keys or certificates, and 60% were AWS access keys. Other sensitive data like passwords, API keys, and credential files were also discovered scattered across environments. Startling we found AWS access keys in Azure, and GCP blob storages.

The uncontrolled proliferation of secrets outside secure storage exposes organizations to potential breaches, unauthorized access, and compliance issues.



of data identifiers that Symmetry found within each env. are IP addresses

Up to 80% of these are public IP addresses

GDPR

Although IP addresses are included as an online identifier of natural persons, they must still be associated with a natural person, this requires determining if there is any other information that can be combined with the IP and used to profile the individual using the IP address

CCPA

Similarly the CCPA considers IP addresses as an online identifier, but indicates that if an organization does not link the IP address to any particular consumer or household, and could not reasonably link the IP address with a particular consumer or household, then the IP address would not be 'personal information."

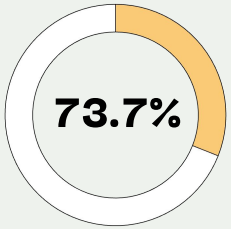
The Possible Sensitivity of IP Addresses in B2C

No surprises here - on average, **31% of the data identifiers** found across the environments were IP addresses. The widespread use of cloud services such as CloudTrail, which leverages S3 buckets and other blob storage for log data is the obvious culprit.

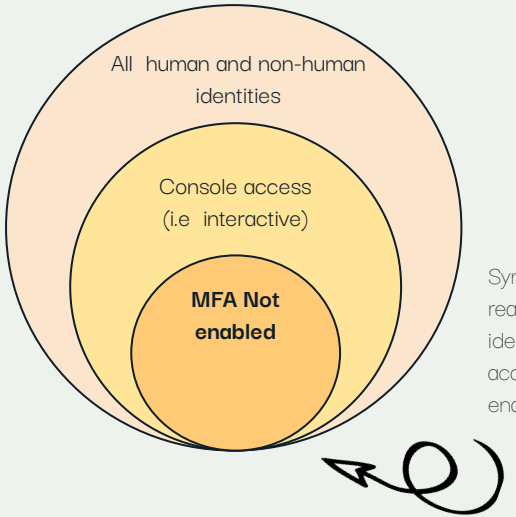
But given that up to **80% of these IP addresses are public IP addresses (outside the organization),** this raises potential compliance considerations in relation to privacy for B2C organizations. IP addresses can be considered personal information under various Privacy laws. Further analysis is necessary to determine the extent of regulation that needs to be applied. In such cases, organizations must ensure compliance with the legal basis for recording this data, likely for security purposes and prevent its use for unauthorized purposes, such as marketing activities.

Alas, poor MFA! A non-negotiable security measure once, of undeniable protection. Yet still not in place everywhere. Your security teams chop-fallen! For **73.7% of organizations** are an unguarded city, lacking thy multi-factor embrace. **Alas!**

SYMMETRY



of organizations have at least one account without MFA enabled & console access to their environment.

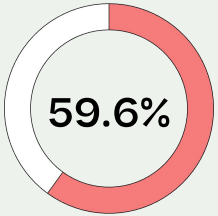


Symmetry focuses on the real risk from MFA, identities with console access without MFA enabled.

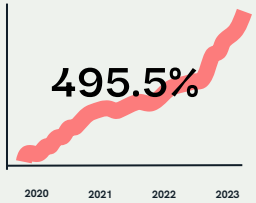
The Continued struggle to MFA

Alas, poor **Multi-factor Authentication!** Alas! Every organization knows that identities without multi-factor authentication (MFA) enabled poses a significant security risk that cannot be ignored, but a **concerning 73.7% of organizations have at least 1 user** account putting it at risk from credential stuffing, phishing and more.

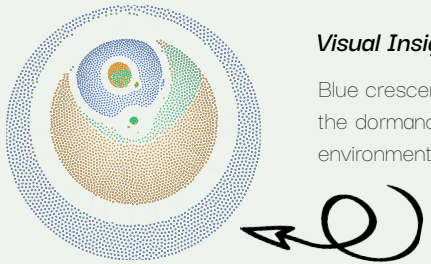
These user accounts without MFA enabled, particularly those with console access privileges, can provide entry points for malicious actors to gain unauthorized access to sensitive systems and data subject to compliance regulations. Leaving these accounts unprotected by MFA needlessly expands the attack surface and potential blast radius of a major security breach.



of Data stores are considered dormant - with no operations performed within last 90 days.



Average growth in the number of dormant data stores **in the last 12 months.**



Visual Insight into Data Dormancy

Blue crescent of data stores is indicative of the dormancy of the data stores in this environment.

Silent explosion of Dormant Data

The issue of **dormant data stores** poses a significant security risk that cannot be ignored. Dormant data stores are those that have not been accessed or contain data that has not been utilized in an extended period, typically defined as 90 days or more of inactivity.

Nearly **60% of data stores have not had any operations** performed on them in the last three months. Even more concerning, the amount of these inactive and neglected data stores has skyrocketed by a staggering 120% over the past 12 months. These dormant repositories, left unattended for 90 days or more, often contain sensitive information, intellectual property, or personal data subject to regulations, yet permissions to them remain unchanged over time remain, widening the attack surface and increasing the potential blast radius of a data breach.

”There are **known knowns**;
there are things we know we know.

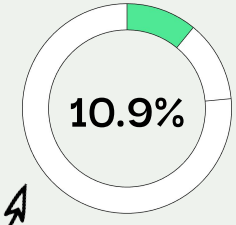
We also know there are **known unknowns**;
that is to say, we know there are some
things we do not know.

But there are also **unknown unknowns** –
the ones we don’t know we don’t know.”

———

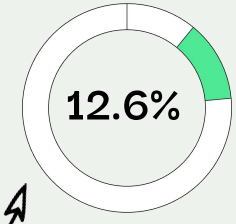
Donald Rumsfeld

Former United States Secretary of Defense



of Cloud Accounts / Projects are unknown - outside the organization's control boundary & unknown to the organization.

Almost entirely made up of AWS accounts, where the architecture model encourages connection based on Account number



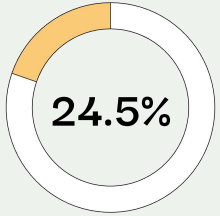
of Cloud Accounts / Projects with access to data are known and external to the organization*.

Identified by leveraging data provided by Symmetry's proprietary insight across customers, and open source insight from the Cloud Security Forum's known_aws_accounts repository.

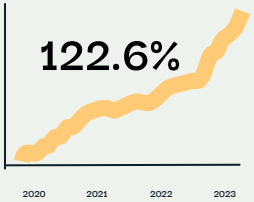
The Known Unknowns of AWS

Understanding who has access to your data, starts with understanding the Cloud Accounts and Projects that your Cloud Accounts and Projects are connected to. Alarmingly, **10.9% of accounts/projects are unknown** to the organizations they are connected to, although have access to their data. This is largely driven by the AWS's account number-based architecture model, which differs considerably in GCP and Azure.

In addition, of the known accounts/projects with access to data, **12.6%** can be mapped to **known external entities** underscores the critical need for robust access controls, continuous monitoring, and rigorous risk assessments to mitigate potential threats from these unseen entities within cloud environments.



of Identities with access to data are considered dormant - no operations in last 90 days.



Average growth in the number of dormant identities **in the last 12 months.** (normalizing for small populations)

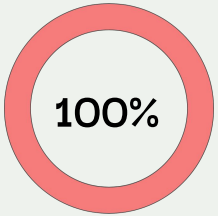
Dormant Identities Slumber Unhealthily On

Dormant identities pose a significant risk to data security that cannot be overlooked. **A staggering 24.5% of identities** that we identified with permissions giving them access to data are considered dormant, having performed no operations on any date in the last 90 days or more. For environments without automated remediation, the **number of dormant identities increase by 122.6% on average over 12 months.**

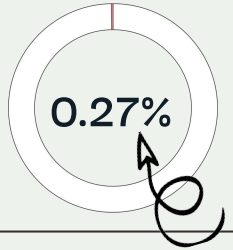
These dormant identities are often easy entry points to an organization's network and data stores with no-one to cry wolf when credential stuffing blocks access, or change a breached password.

Without regular monitoring and access reviews, these unused yet privileged identities can go unactioned for months, providing ample time for threat actors to exploit them.


A Deep Dive into
Microsoft Copilot

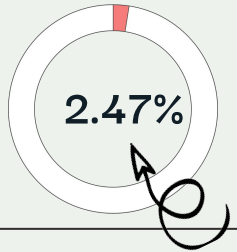


of organizations with Microsoft 365 have plans to enable CoPilot for Microsoft 365 in some form in next 12 months.



of files stored in Onedrive or Sharepoint allow anonymous access.

5.5% contain sensitive information.



of files stored in Onedrive or Sharepoint allow organization wide access.

6.5% contain sensitive information.

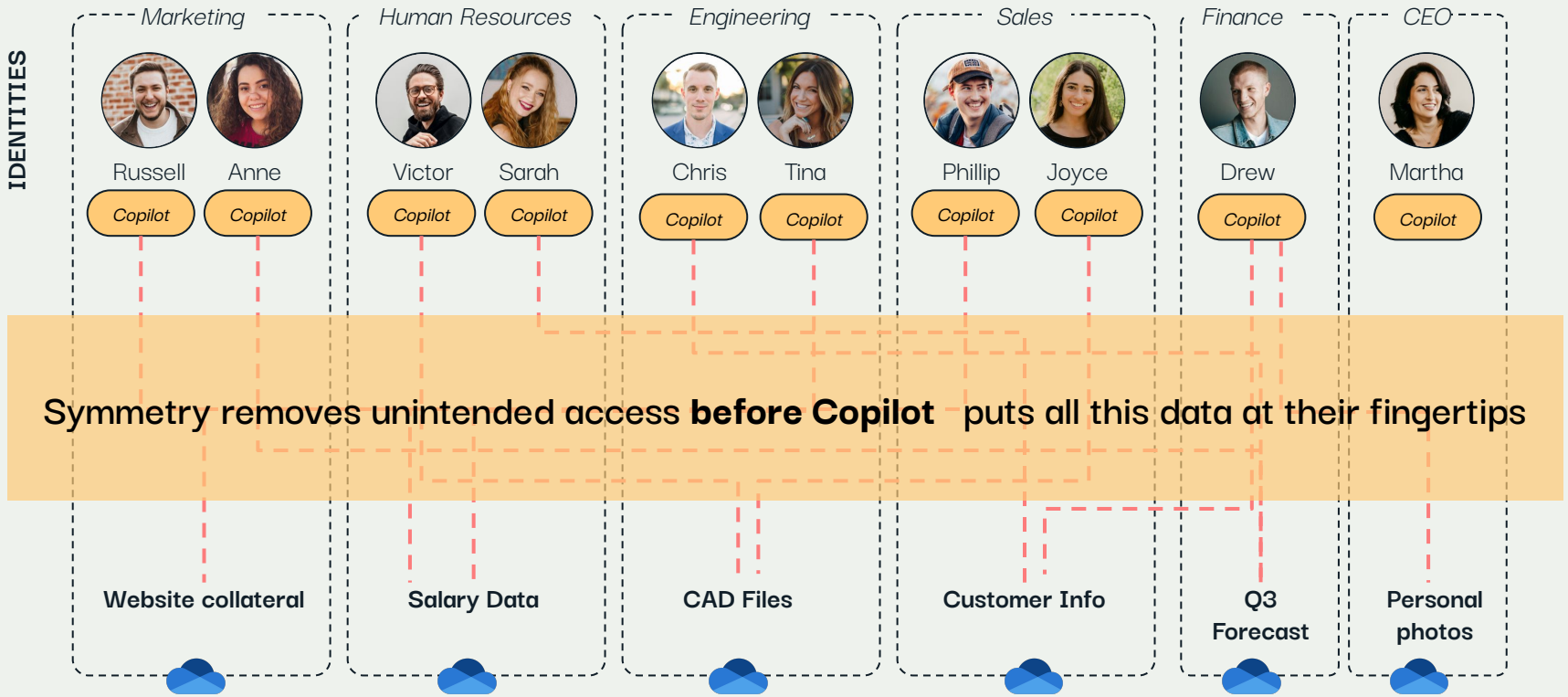
Copilot Adoption Set to Explode Access to Sensitive Data

A striking 100% of organizations utilizing Microsoft 365 have plans to enable Copilot for Microsoft 365 in some capacity within the next 12 months. This will continue to energize the data security landscape as it brings existing data access risks to the forefront.

Our findings show that **0.27% of files** stored in OneDrive or SharePoint allow **anonymous access**, with a concerning **5.56% of those files containing sensitive information**. Furthermore, **2.47% of files** stored in these cloud-based platforms permit **organization-wide access,** and an alarming **6.55% of those files contain sensitive data**.

At the scale of millions of files stored in OneDrive and SharePoint, this translates to **tens if not hundred of thousands of potential security alerts** that will demand investigation and mitigation.

Data+AI Security Outcomes that Matter



SYMMETRY

Unintended access to sensitive data.

"We predict that Microsoft 365 **Copilot will reach 6.9 million** US knowledge workers in 2024"

Forrester Research

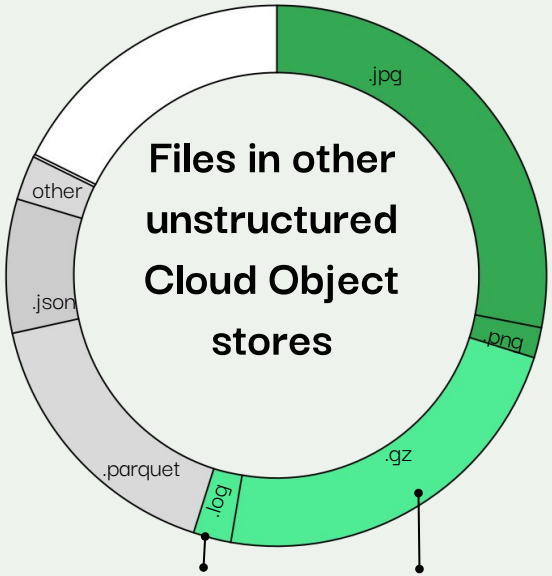
Mohit Tiwari • 1st
Co-Founder & CEO at Symmetry Systems
2mo •

Hot off the press -- Microsoft Copilot solves the shortage of pentesters and offensive security engineers. 🤗

It turns every identity in your org into an AI-powered pen-tester.

It indexes all data that each person has access to and then answers everyone who's curious to learn***:
* what is Bob's salary?
* what do CAD files with our core IP look like?
* how do our sales results and next year's plans look?

Congrats -- no more offensive skills shortage! ✊



2.29% 22.78%

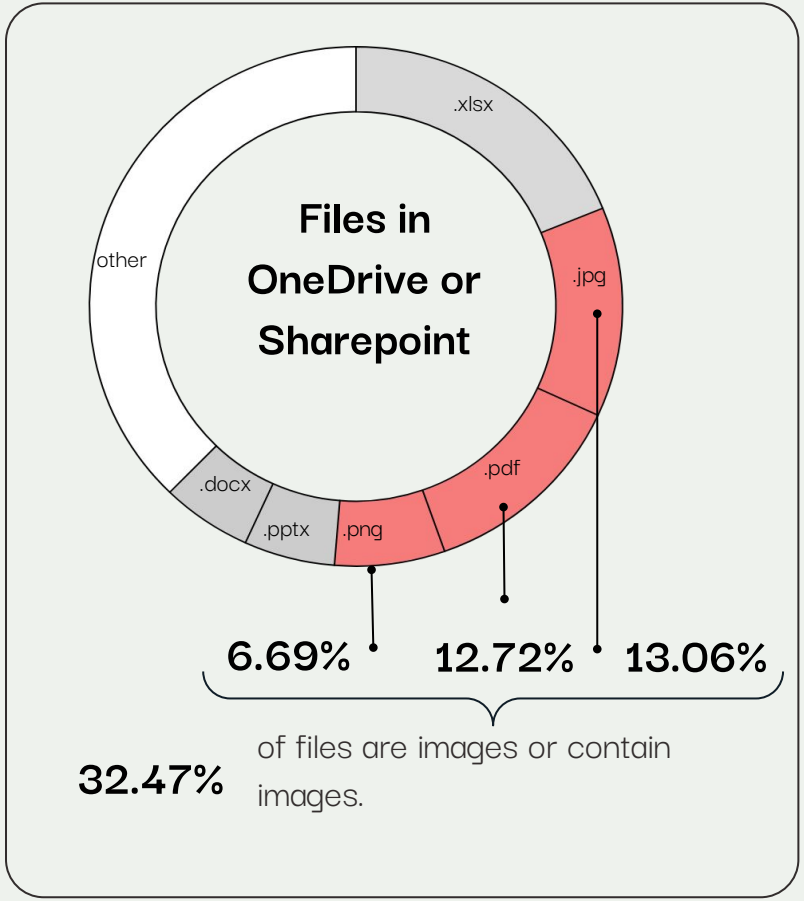
25.07% of files are logs or compressed log files.

When you can't see the Forest for all the Logs and Photos

Compressed files (gz: 22.78%) and uncompressed logs (2.29%) make up a significant portion of the stored data. While gz files can contain other compressed data, they are also commonly used to store compressed log files. While Logs generally contain limited sensitive information and have restricted internal access, sharing log data externally requires proper governance and redaction processes to protect any sensitive content.

Image data, including JPG (28.08%) and PNG (1.83%) files, is also prevalent. Identifying sensitive information within images requires OCR for text extraction and machine learning classifiers for content categorization. Overall, the diverse data formats underscore the need for robust data management, access controls, and a holistic governance strategy tailored to the organization's specific needs and data types.

SYMMETRY



Unstructured Data Types have their own AI security needs

A significant portion of data residing in organizations' OneDrive and SharePoint repositories (the primary scope of Microsoft CoPilot for 365) consists of image files or contain images, with **13.06% being JPG image files, 12.72% PDFs, and 6.69% PNGs.** Identifying sensitive information within this vast trove of unstructured data requires **both optical character recognition (OCR) for extracting text from images and machine learning-based classifiers for categorizing the images.**

However, attempting to process and classify every single file can be computationally intensive, cost prohibitive and impractical. In our opinion, the priority should be to reduce broad access to these repositories and implement targeted classification on a subset of files that are most likely to contain sensitive data based on factors such as file metadata, access patterns, and user context.

Thank You!