THE STATE OF
Cloud Data Security

2023

Dig Security

Table of Contents

Executive Summary

With more and more data being stored in the cloud, there is no better time than now to review how it is being handled and what we can learn from it. Until now, we could only estimate, via surveys, the answer to questions like where sensitive data resides and how much of it is currently at risk. Why? Because of the lack of tools enabling effective and timely visibility and assessment.

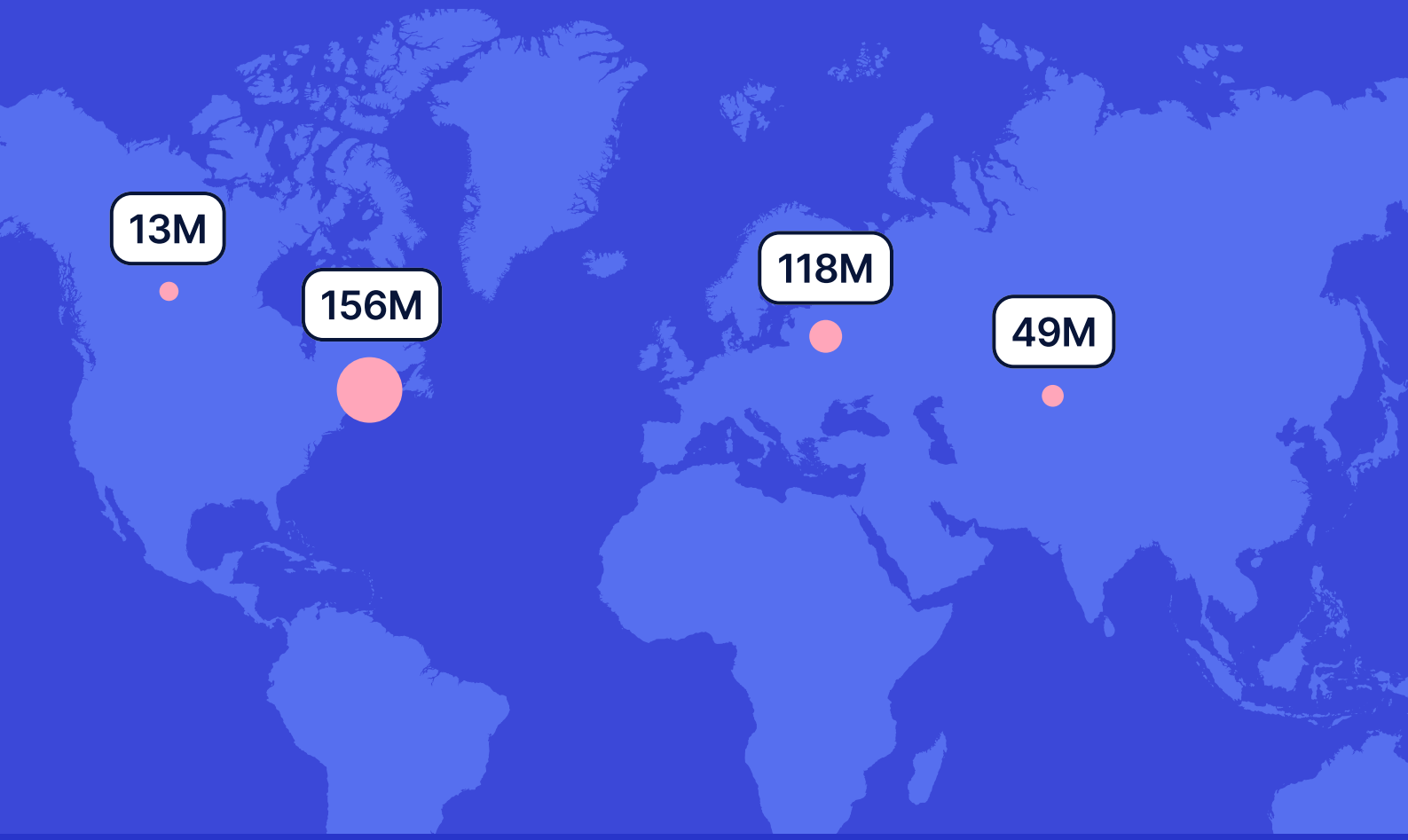
This paper follows up a study analyzing over 13 billion files and 8 petabytes of information stored in various public cloud environments – our “sample data set.” We hope the paper drives more awareness and responsibility over how we engage with sensitive data in today's working environments.

Dig Security, Data Security Research, 2023

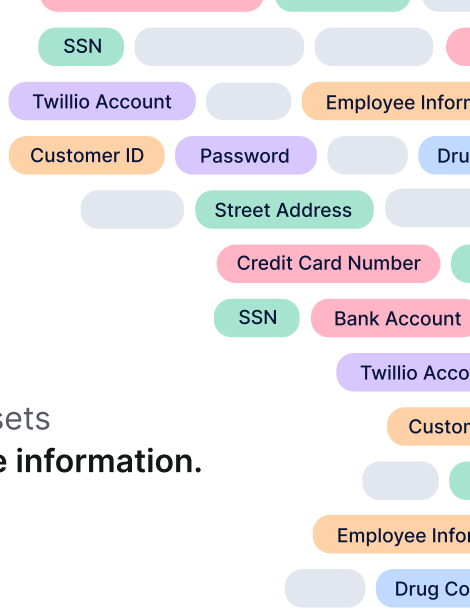
PART 1

Know Your Data - Where is Your Sensitive Data?

As much as the cloud empowers organizations through data democratization, it also increases data sprawl. Data is constantly being shared, copied, transformed, and forgotten. The ease of operating with data in the cloud causes sensitive data to proliferate across services, clouds, and geographies, often leading to security and compliance breaches.



Sensitive information is located everywhere.



More than

30%

of cloud data assets
contain sensitive information.

Types of sensitive information

Our research shows that the most common sensitive data type that organizations save is personal identifiable information (PII), which contains employee and customer data. A more surprising finding shows that companies also save employees' health information such as health insurance and COVID-19 vaccination status.

In a sample study of 1 billion records, 21% of all PII that we found contained email addresses, followed by phone number and street addresses.

PII type distribution *

21%	7%	6%
EMAIL ADDRESSES	PHONE NUMBERS	STREET ADDRESSES

More than

10M*

social security numbers were also found, making it the **sixth** most-common type of sensitive information.

About

3M*

credit card numbers were also found, making it the **seventh** most-common type.

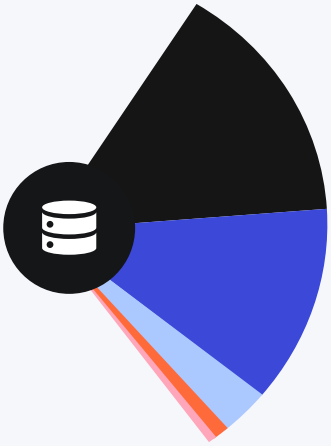
* out of 1 billion sample records

Data classification examples that are essential to many organizations:

1. **Personal identifiable information (PII):** Data that can be used to identify an individual such as full name, social security number, driver's license number, or passport number.
2. **Financial information:** Data related to financial transactions and accounts such as credit card numbers, bank account numbers, and investment information.
3. **Confidential business information:** Data that is proprietary to a company and gives it a competitive advantage such as trade secrets, business plans, and market research.
4. **Health information:** Data related to a person's health status and medical history such as diagnoses, treatment plans, and prescription information.
5. **Intellectual property:** Data related to patents, trademarks, copyrights, and trade secrets.
6. **Government information:** Data that is classified or restricted by government agencies such as national security information, law enforcement records, and classified military information.
7. **Employee information:** Data related to employees such as payroll information, job performance evaluations, and disciplinary records.

Over 30% of the data in structured and unstructured managed services contains sensitive information.

Below is a distribution graphic of different sensitive data types.



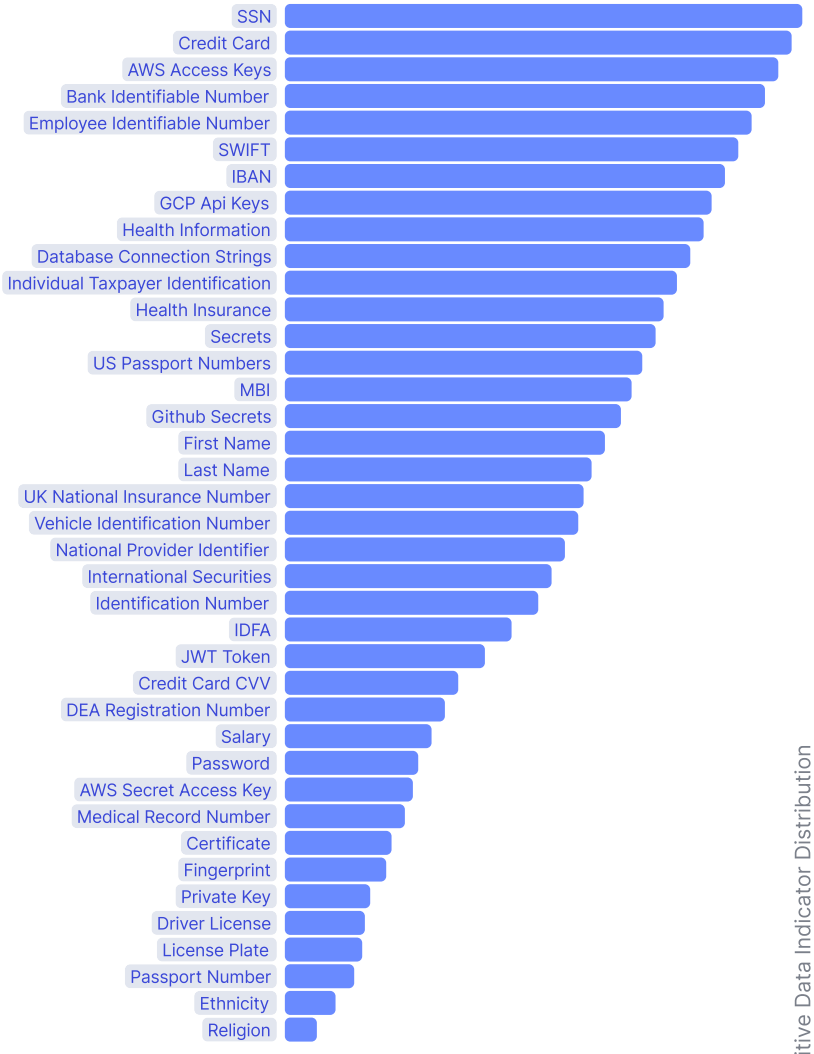
Sensitive information type distribution

PII	44%
PCI	7%
Developer secrets	1%
PHI	0.2%
Other sensitive entities	47.8%

We found no differences between the distribution of sensitive data types in structured and unstructured managed services.

Sensitive data comes in various forms.

The following graph is a sample of indicators and their distribution that we found in customer environments.



Production vs. Development

Non-production environments contain large volumes of sensitive information.
More than 40% of developer secrets, 30% of PII, and 20% of financial information are located in non-production environments.

SENSITIVE INFORMATION IN NON-PRODUCTION ENVIRONMENTS

More than

40%

Developer secrets

30%

Personal identifiable information (PII)

20%

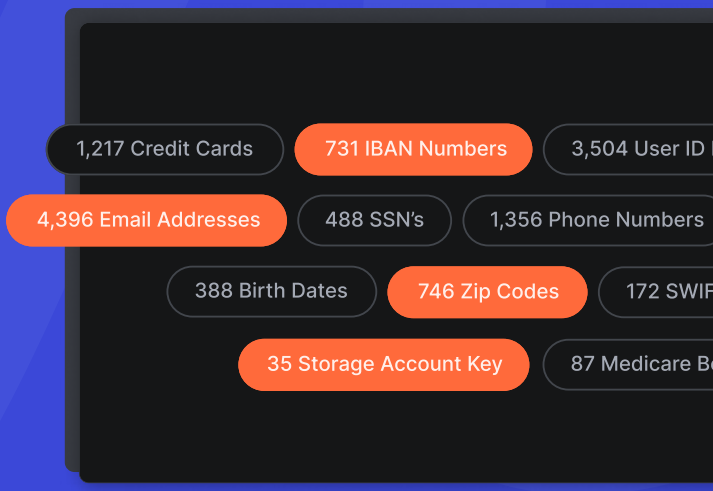
Financial information

Why is it important?

Development, or dev, environments are considered "sandbox" environments. By nature, they tend to be less secure and less monitored, and therefore, more vulnerable than others.

That is also why dev environments are more attractive for attackers – many known breaches like the Uber breach and Lastpass started from there.

Dev environments can attract the initial breach, which often leads to higher privileges or to environments with more sensitive information.



Is your sensitive data at risk?

The first step in establishing data security is to identify and locate all sensitive data throughout the organization. As soon as the data is discovered and classified, it is crucial to assess its posture and ensure that appropriate controls are in place to protect it.

Risk types associated with data assets mentioned in this paper include:

- **Lack of encryption**
  Enables unauthorized actors to access sensitive data
- **Lack of comprehensive logging**
  Prevents keeping track of user interactions and access attempts, which are essential to proactive threat detection, incident response, and forensic investigations
- **Open-to-the-world (public) sensitive data**
  Exposes data beyond the need-to-know principle and violates data privacy regulations

◆ Publicly exposed data types ◆

20% of **financial data** found is publicly exposed

15% of **developer secrets** found is publicly exposed

4% of **PCI and PII** found is publicly exposed

PUBLICLY EXPOSED

Storage assets at risk

More than 7% of storage services containing sensitive data are public. More than 60% of storage services are not encrypted at rest, and almost 70% lack comprehensive logging.

Database assets at risk

91% of database services with sensitive data are not encrypted at rest, 20% lack comprehensive logging, and 1.6% are open to the public.

Assets with sensitive data at risk

RISK	STORAGE	DATABASE
Open to the world	7%	1.6%
Lack encryption	60%	91%
Lack comprehensive logging	70%	20%

Sensitive data under compliance

The PCI regulation clearly requires that sensitive information be audited and encrypted at rest.

About

60%

of credit card information **lacks comprehensive logging**

More than

60%

of credit card information is **not encrypted at rest**

PART 1

Takeaways

1. Sensitive data is everywhere, but most companies don't know where their data resides and what types of sensitive information it contains, leaving them exposed.

2. Security in dev environments is important, just like in production, and attackers will use the weakest link in the chain to get in.

3. Despite expert risk assessments, it's likely some sensitive data is still vulnerable to threats, including publicly accessible, unlogged, or non-encrypted data.

4. By knowing where sensitive data is stored, risk management can be simplified and data security can be improved.

PART 2

Who Has Access to Your Sensitive Data?

This section deals with who has access to sensitive information by examining dangerous roles that can lead to its exposure. We also look at the risks of sharing sensitive information between cloud accounts, and explore access to sensitive information in both storage assets and managed databases.

Admin vs. Consumer

To better understand how sensitive data is exposed, we looked at the permissions a principal has on sensitive assets and divided them into two categories:

1 **Consumer** permissions → **Allows a principal to read/write data.**

2 **Admin** permissions → **Allows a principal to change the asset configuration.** For example, make the asset public or add direct access permissions to a foreign account.

---

* Only a principal with consumer privileges maintains actual access to sensitive information.

Admin permissions can lead to the exposure of sensitive information.

For example, a principal with admin privileges may grant direct access permission to remote accounts or expose the data to the public.

Our research shows that...

About **10%** have **admin access** to at least one sensitive asset

About **20%** have **consumer access** to at least one sensitive asset

More than **35%** have some **privileges** to sensitive data assets

This is troubling information, since it means that sensitive data can be easily accessed.

About

10%

of principals **have consumer access to PCI data**

More than

5%

of principals **have admin access to PCI data**

Separation of duties

Separation of duties (SoD) is an important best practice in security.
Take, for example, a DevOps engineer who requires access to manage data assets. That person is not a consumer of such data and should not have access to the data itself.

To determine if this best practice is enforced in a cloud landscape, we asked how many principals with admin permissions also have consumer permissions.

We found that all the principals with admin permissions also have consumer privileges, thereby violating the separation of duties best practice.

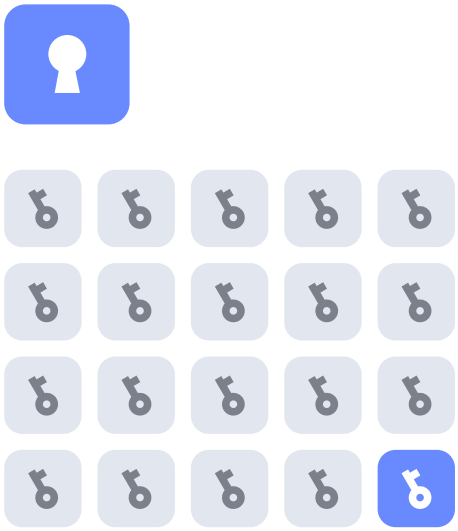
100% of the principals **with admin permissions also have consumer privileges**

Excessive permissions to sensitive assets

Access to an asset in a cloud platform can be granted specifically to a principal, to all principals, or to all assets. For example, in AWS, you can use the symbol "*" for all resources in the cloud or "**S3:***" for all S3 buckets.

Less than 5% of roles are granted with an explicit account type (consumer or admin), while the remaining 95% are granted through excessive privileges.

This is a warning sign that sensitive data should be protected and guarded more carefully.



95%

of principals with management or consumer permissions **are granted them through excessive privilege**

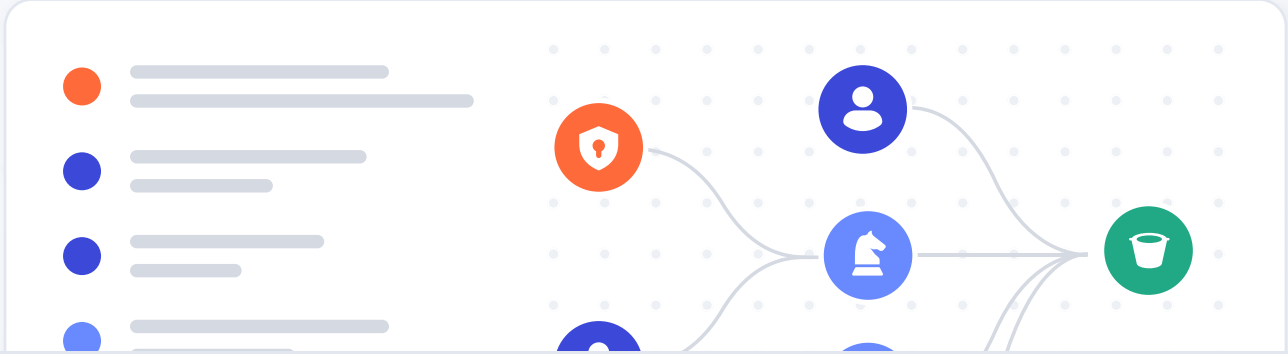
Access to **managed databases**

Managed databases provide two layers of role-based protection:

1 Role-based protection **of the cloud platform (IAM)**

2 Role-based protection **of the content inside the database**

The first layer represents management permissions on the asset, while **the second** is consumer permissions that require credentials to connect to the database and to use the data.



Given certain platform management permissions, a cloud principal could access the data within the database without receiving implicit access to the data.

This could be done through certain actions, for example, by resetting the admin password, making the database public, or exporting the database. These are powerful permissions and should be granted on a need-to-know basis.

We collected all possible actions to access the data inside a managed database and checked:

Who actually had permissions to carry out dangerous actions?

The results are even bleaker when looking at managed databases.

"Excessive" is the name of the game for them.

More than

8% of the principals **have modification permissions**

About

8% of the principals **have connection permissions**

7% of the principals **have export permissions**

Sensitive managed instances

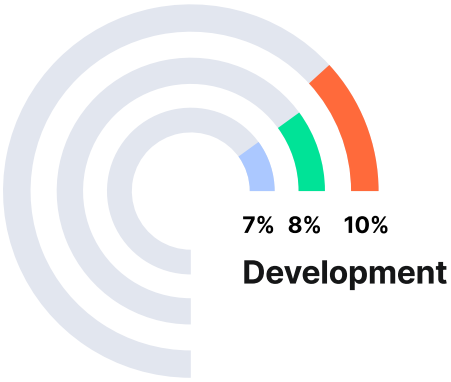
Excessive permissions in production and development

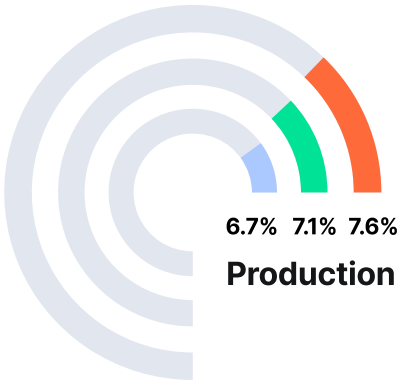
As mentioned in the previous section, sensitive information exists not only in production, or prod, environments, but also in development environments.

We examined access permissions between prod and dev environments for sensitive assets, and found that **the permission exposure ratio between the two is similar**.

Moreover, **the exposure of principals to sensitive assets in production is lower than in development environments.**

Distribution of excessive permissions in dev vs. prod managed databases





Managed databases

- Principals with modification permissions
- Principals with export permissions
- Principals with connect permissions
- All other permissions

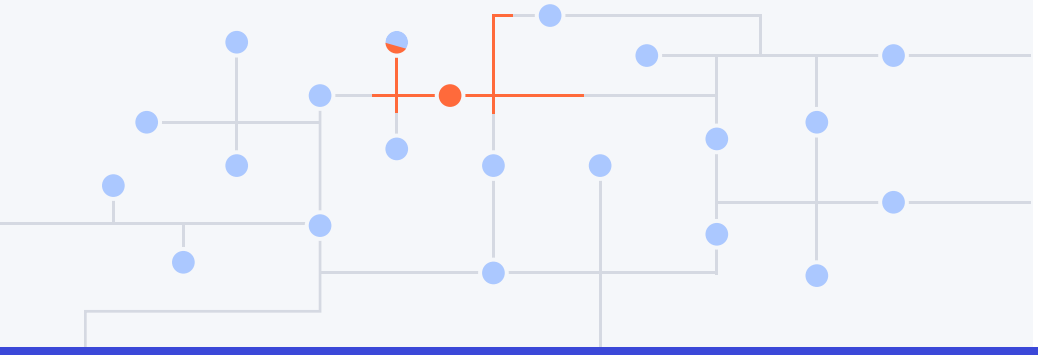
Storage services	Development	Production
Principals with management permissions	8.86%	10.21%
Principals with consumer permissions	17.56%	20.31%
Principals with other permissions	35.1%	37.79%

Sensitive assets with access risk

Sensitive assets shared between accounts or projects increase the potential risk of data exposure and reduce the security posture of the asset itself, since the privileges are managed in a remote account.

We found that almost 8% of sensitive data assets are shared with other accounts or projects.

> Overall, more than 2% of sensitive data assets are at risk due to direct access from a remote account.



8%

of sensitive data is shared with other accounts

- **55%** **of the shared assets** are shared with other accounts **in the same organization**
- **36%** **of the shared assets** are shared with other accounts **outside the organization**
- **9%** **of the shared assets** are shared with **third-party vendors**

PART 2

Takeaways

1. We showed how the separation of duties concept is neglected and not enforced in the cloud. It is recommended to remove consumer access from administrative roles.
2. While the majority of access is granted through excessive permissions, it is recommended to grant explicit permissions to each asset.
3. Sensitive data shared between accounts weakens control and increases the risk of data exposure. Reduce the exposure of sensitive data to multiple accounts.
4. Permissions and role-based access control (RBAC) are insufficient protections in the cloud. Another security layer is needed to manage the sensitive information and all paths leading to it. See the Summary for more.

PART 3

Where Does Your Sensitive Data Flow?

When a security incident happens, the questions we ask include:

What was taken?

When was it taken?

Who took it?

Where was it taken from?

How was it taken?

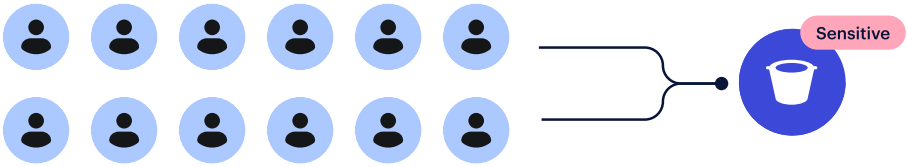
Answering the first four questions helps us answer the fifth:

Who is accessing the sensitive data?

In the previous chapter, we asked who has access to sensitive data.
Now we want to know who is actually using their permissions to access sensitive data.

Sensitive data assets, on average, are accessed by **14 different principals.**

Some 85% of the buckets containing sensitive data are accessed, on average, by **12 principals.**



What types of services access the data?

40% of the data **flows to data lakes**

30% of the activity involving sensitive information is replication between storage assets

16% of the activity represents applications operating on the data

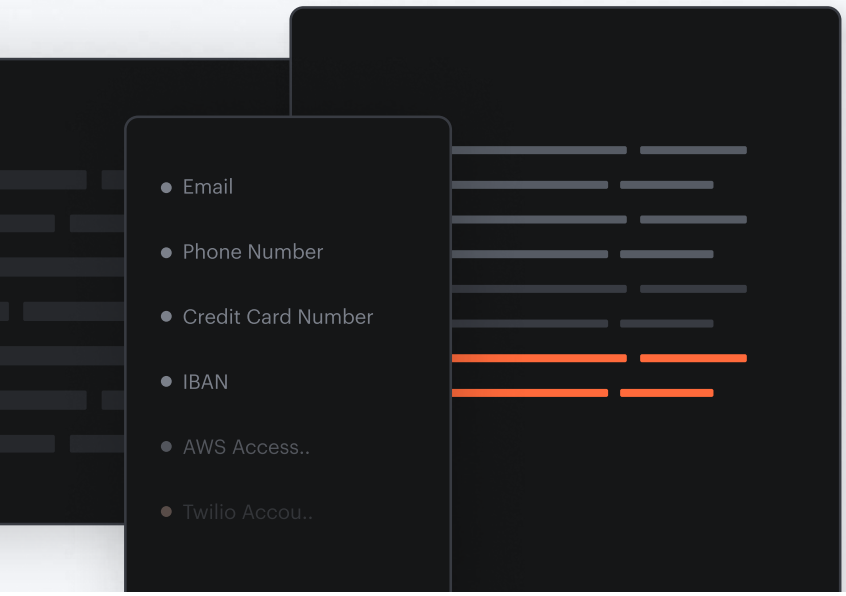
Hadoop

Hadoop exhibits major traffic of 37%, which puts sensitive data at risk, since it is being duplicated into an unmanaged environment.

Snowflake

An additional 2.6% of the activity is data ingested by Snowflake or other data lake solutions.

How many services, on average, access sensitive storage assets?



More than

50% of services are accessed by **5-10 applications**

About

20% of services are accessed by **10-20 applications**

From where are your sensitive data assets being accessed?

Sensitive information accessed from different geolocations is common.

Some regulations like GDPR clearly restrict sensitive information from leaving its geolocation.

More than one of every two assets, 56%, are accessed from multiple geographic locations.

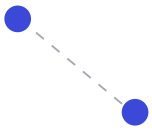
26% of sensitive data assets are accessed by five different geolocations.

Where does the data flow to?

As the data flows, the risk grows.

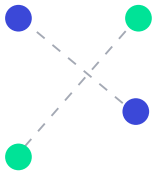
77%

of sensitive data assets **have 1 cross-service flow**



40%

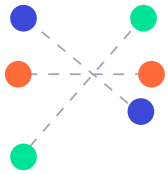
of sensitive data assets **have 2 cross-service flows**



More than

20%

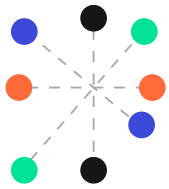
of sensitive data assets **have 3 cross-service flows**



More than

10%

of sensitive data assets **have 4 cross-service flows**

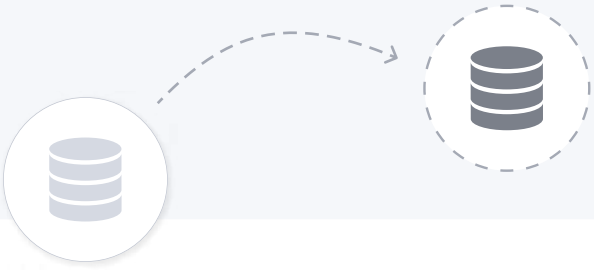


Data flow risks

30%

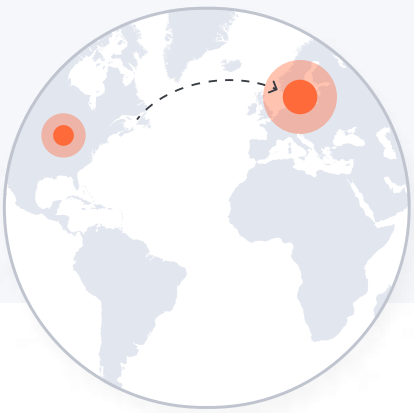
of the traffic in sensitive storage assets is from other storage assets, which means the data is replicated.

Duplication of data across different regions doubles the risks of exposure, and could lead to a compliance breach if the replication is carried out across different geolocations.



More than **1%**

of sensitive data assets **are replicated outside their geolocation.**



Sensitive assets at data flow risks

2%

are transferred to open-to-world assets

2%

are transferred out of production environments

4%

are transferred to a foreign project

Data flow risks by company

6% OF COMPANIES

have sensitive data **that has been transferred to publicly open assets**

10% OF COMPANIES

are at risk of sensitive information flowing from production to development environments

*

Since development environments are, by nature, less secure and less monitored, sensitive data is at risk.

PART 3

Takeaways

1. Sensitive data is accessed by many principals regularly. Minimize excessive permissions and continuously monitor principals' access to sensitive data, which will help reduce your sensitive data exposure.
2. Turn on logging for sensitive data assets to enable monitoring.
3. Data flows represent duplication that increases exposure risk. Reduce flows to the minimum required and make sure the destination is secured.
4. Ensure that your data flows do not violate your internal governance and external compliance mandates.

Summary and Closing Notes

We divided our journey into three parts.

In the first part, we saw that sensitive information has no specific location: it is located throughout the cloud ecosystem. We also saw how the borders between production and development environments are blurred when it comes to sensitive information, thereby increasing the risk of sensitive data exposure.

In the second part, we focused on permissions to assets containing sensitive information.
We emphasized the need for separation of duties to avoid excessive access. We reviewed the differences between a consumer of the information and a manager, as well as the ease with which a manager can become a consumer of sensitive information. We showed that although security is a common practice today, many permissions are granted through excessive roles that may expose your sensitive information.

We also discussed the danger of sharing sensitive information between cloud accounts, from production to development environments, and with foreign accounts and third-party vendors.

In the third and final part of our research, we delved deeper into where sensitive data flows. We saw that sensitive data flows outside of its original geolocation. As more sensitive data is accessed by principals, applications, and geolocations, the risk of its being exposed or exfiltrated is greater.

Cloud providers continue to implement controls for securing data assets. Nevertheless, the responsibility to turn on these controls falls on cloud data service consumers. We highlighted in our research the absence of such security controls when it comes to sensitive data.

It is clear that there is a need for a security layer to ensure that data is protected in cloud assets.

About Dig Security

Dig Security helps organizations discover, classify, protect, and govern their cloud data.
With organizations shifting to complex environments with dozens of database types across clouds, monitoring and detecting data exfiltration and policy violations has become a complex problem with limited fragmented solutions. Dig's cloud-native, completely agentless approach reinvents cloud DLP with data detection and response (DDR) capabilities to help organizations better cope with the cloud's data sprawl.
Dig is founded by 3 cyber security veterans from Microsoft and Google, and is backed by Team8, CrowdStrike, CyberArk, SignalFire, Okta Ventures, and others

Dig Security