Section 1 of 11

Log Anonymization

Welcome to Our Survey on Software Logs Anonymization

Software logs play a crucial role in various domains, providing valuable insights and supporting diverse applications. However, many organizations face challenges in managing, sharing, or transferring logs due to privacy concerns.

In our research, we aim to explore what constitutes "sensitive" information in software logs. In this survey, we aim to understand the perspective of industries and organizations on the subject of log privacy, ultimately facilitating safer and more effective log management practices.

This survey will only take 5-10 minutes of your time. Your input will be invaluable in shaping our findings and recommendations. Thank you for participating!

Are you ready to begin the survey? *

- ○ Yes, I'm ready!

After section 1 Continue to next section

Section 2 of 11

Section title (optional)

Description (optional)

Do you have any experience in analyzing, processing, or managing software logs? *

- ○ Yes
- ○ No

After section 2 Continue to next section

Section 3 of 11

Section title (optional)

Description (optional)

Are you familiar with the following data protection regulations and standards? (Select all that apply)

- [ ] GDPR (General Data Protection Regulation)
- [ ] CCPA (California Consumer Privacy Act)
- [ ] HIPAA (Health Insurance Portability and Accountability Act)
- [ ] PIPEDA (Personal Information Protection and Electronic Documents Act)
- [ ] NONE
- [ ] Other...

After section 3 Continue to next section

Section 4 of 11

Section title (optional)

Software logs come in various formats and contain a diverse range of attributes, which can differ significantly in structure and detail. This variability in log shapes and contents can impact how sensitive information is identified and handled. Below we provide some examples.

---

1. Dec 10 07:07:38 LabSZ sshd[24206]: pam_unix(sshd:auth): authentication failure; logname= uid=0 euid=0 tty=ssh ruser= rhost=ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn

This log entry indicates a failed authentication attempt via SSH.
Attributes: Date and Time, Host, Service, Process ID, Component, Message, IDs, Remote user and host

information

---

2. 76723 node-55 node temperature 1077205904 ambient=33

This log entry reports the temperature of a node.
Attributes: Log ID, Node ID, Component, Time, Environmental data

---

3. 2015-10-18 18:01:53,713 INFO [main] org.apache.hadoop.mapreduce.v2.app.rm.RMContainerAllocator: maxContainerCapability: <memory:8192, vCores:32>

This log entry shows the maximum resource capabilities for containers in Hadoop MapReduce.
Attributes: Date and Time, Log level, Thread, Component, Configuration details

---

4. 132.201.126.207 - - [28/Aug/1995:17:23:48 -0400] "GET /pub/atomicbk/images/spicy2.gif HTTP/1.0" 200 29018

This log entry is an HTTP access log entry showing a request made to a web server.
Attributes: IP address, Date and Time, Request method, File path, Request protocol, Status code, Response size

---

Based on the above examples, which of the following attributes do you consider sensitive in software logs? Each row includes an example after the colon (:) (Select all that apply) *

- [ ] IP address: 192.168.1.1
- [ ] Mac address: 5c:50:15:4c:18:13
- [ ] Date and Time: 2024-08-15 - 12:11:37
- [ ] File path: /user/root/rand/_temporary/part-00742
- [ ] URL: http://cs-www.bu.edu/lib/pics/bu-logo.gif
- [ ] Log level: INFO
- [ ] Component (the component that generated the log): org.apache.hadoop.mapreduce.v2.app.MRAppMast...
- [ ] IDs: Process ID, Thread ID, Job ID, Node ID, Application ID, Device ID
- [ ] Host name: ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn
- [ ] Port number: 8080

- [ ] Request protocol: HTTP/1.0
- [ ] Request method: GET
- [ ] Request status code: 200
- [ ] Request response size
- [ ] Request response time
- [ ] Environmental data: temperature ambient=33
- [ ] Configuration details: vCores:32
- [ ] Username: cheng
- [ ] NONE
- [ ] Other...

After section 4 Continue to next section

Section 5 of 11

Section title (optional)

Description (optional)

Have you ever anonymized log data? *

- ( ) Yes
- ( ) No

After section 5 Continue to next section

Section 6 of 11

Section title (optional)

Description (optional)

When you want to share software logs, which attributes do you anonymize? Each row includes an example after the colon (:) (Select all that apply) *

- [ ] IP address: 192.168.1.1
- [ ] Mac address: 5c:50:15:4c:18:13
- [ ] Date and Time: 2024-08-15 - 12:11:37
- [ ] File path: /user/root/rand/_temporary/part-00742
- [ ] URL: http://cs-www.bu.edu/lib/pics/bu-logo.gif
- [ ] Log level: INFO
- [ ] Component (the component that generated the log): org.apache.hadoop.mapreduce.v2.app.MRAppMast...
- [ ] IDs: Process ID, Thread ID, Job ID, Node ID, Application ID, Device ID
- [ ] Host name: ec2-52-80-34-196.cn-north-1.compute.amazonaws.com.cn
- [ ] Port number: 8080
- [ ] Request protocol: HTTP/1.0
- [ ] Request method: GET
- [ ] Request status code: 200
- [ ] Request response size
- [ ] Request response time
- [ ] Environmental data: temperature ambient=33

- [ ] Configuration details: vCores:32
- [ ] Username: cheng
- [ ] NONE
- [ ] Other...

After section 6 Continue to next section

Section 7 of 11

Section title (optional)

Description (optional)

Why do your responses regarding which attributes are sensitive (the question on page 4) differ from those you actually anonymize (the question on page 6) in software logs? (Select all that apply) *

- [ ] Some attributes that I consider sensitive are not present in the logs I work with, so I cannot anonymize th...
- [ ] I anonymize only those attributes that are required by company policies or regulations.
- [ ] Anonymizing certain sensitive attributes could compromise the utility of the logs, so I avoid doing so.
- [ ] There are technical limitations or challenges in anonymizing certain sensitive attributes.
- [ ] The logs I handle are not always used for sharing externally, so I anonymize fewer attributes.
- [ ] I rely on automated tools that may not cover all sensitive attributes.
- [ ] My answers do not differ.
- [ ] Other...

After section 7 Continue to next section

Section 8 of 11

Section title (optional)

Description (optional)

What factors influence your decision to anonymize certain log attributes? (Select all that apply) *

- [ ] Legal compliance
- [ ] Risk of re-identification
- [ ] Impact on data utility
- [ ] Company policies
- [ ] Customer requirements
- [ ] Other...

Which anonymization techniques do you primarily use for log data? (Select all that apply) *

Suppression: Completely removing specific values from the data, often replacing them with a placeholder. For example, "192.168.1.1" can be replaced with "-".

Tokenization: Substituting sensitive data with a unique identifier or token that represents the original value. For instance, "192.168.1.1" could be replaced with "IP ADDRESS".

Truncation: Hiding part of the data while leaving a portion visible. For example, "192.168.1.1" might be represented as "192.168.x.x".

Generalization: Broadening the specificity of data by aggregating it into larger categories. For example, "192.168.1.1" can be generalized to "192.168.0.0/16".

Permutation: Replacing original values with new ones while preserving their order. For example, the list [3.6, 16.8, 21, 0.9, 27.9, 14.4] could be permuted to [3, 12, 15, 3, 18, 9], keeping the order but removing specific information.

Aggregation: Summarizing multiple values into a single metric. For example, instead of logging individual IP addresses, you might log the count of unique IPs per day.

Hashing: Converting data into a fixed-length string using a hash function. For example, "192.168.1.1" could be hashed to "c5eb5a4cc76a5cdb16e79864b9ccd26c3553f0c396d0a21bafb7be71c1efcd8c".

- [ ] Suppression
- [ ] Tokenization
- [ ] Truncation
- [ ] Generalization
- [ ] Permutation
- [ ] Aggregation
- [ ] Hashing
- [ ] Other...

How much do you believe anonymization impacts the usefulness of the data? *

- ○ Not at all
- ○ Slightly
- ○ Moderately
- ○ Significantly

How challenging do you find **balancing** anonymization with the need to preserve the utility of log data? *

	1	2	3	4	5	
Not challenging	○	○	○	○	○	Extremely challenging

How effective do you believe your current anonymization practices are in **protecting sensitive information**? *

	1	2	3	4	5	
Very ineffective	○	○	○	○	○	Very effective

How effective do you believe your current anonymization practices are in **preserving utility**? *

	1	2	3	4	5	
Very ineffective	○	○	○	○	○	Very effective

How efficient do you believe your current anonymization practices are in terms of **computation cost and time**? *

	1	2	3	4	5	
Very inefficient	○	○	○	○	○	Very efficient

After section 8 Continue to next section

Section 9 of 11

Section title (optional)

Description (optional)

Can you share any strategies or best practices you use to maintain data utility while anonymizing logs?

Long answer text

What challenges do you face when applying anonymization techniques to software logs?

Long answer text

After section 9 Continue to next section

Section 10 of 11

Section title (optional)

Description (optional)

Do you have any additional comments, suggestions, or insights you'd like to share regarding log anonymization practices or this survey?

Long answer text

After section 10 Continue to next section

Section 11 of 11

Section title (optional)

In this last section, we will ask you a few questions about your organization to better understand the context of your experiences with software logs.

Which best describes your role in the organization? *

- Data privacy officer
- Security Engineer

- Risk management specialist
- IT manager
- IT security engineer
- Network / system adminstration
- Network / System architect
- Data scientist
- Data engineer
- Software engineer
- Other...

How many years of experience do you have? *

- Less than 1 year
- 1-3 years
- 4-6 years
- 7-10 years
- More than 10 years

In which industry does your organization operate? *

- Technology
- Finance
- Healthcare
- Government

- [ ] Retail
- [ ] Other...

What is the approximate size of your organization? *

- [ ] 1-100 employees
- [ ] 101-500 employees
- [ ] More than 500 employees