

# What data does Monalog collect from you?

---

- Monalog collects what you type on the command line as well as some other contextual information.
- Some examples of the type of data collected include:
  - ◆ resolved aliases (e.g., `ll` may translate to `ls -al` ).
  - ◆ the path of binaries executed (e.g., `ls` may translate into `/usr/local/bin/ls`).
  - ◆ timestamps
  - ◆ working directory
  - ◆ user and process ID

# Why should we sanitize?

---

- Some of the data Monalog collects from you may be of a sensitive or private nature.
- The Sanitizer gives you the ability to not share such data with us.
- Data is precious, so we don't want to waste it.
- This is why we give you granular control over sanitizing the data.
- If you sanitize smaller pieces of data (single lines vs. entire sessions), the amount and quality of data available to us improves, without forgoing your privacy.

# What is a “session”? What are logs?

---

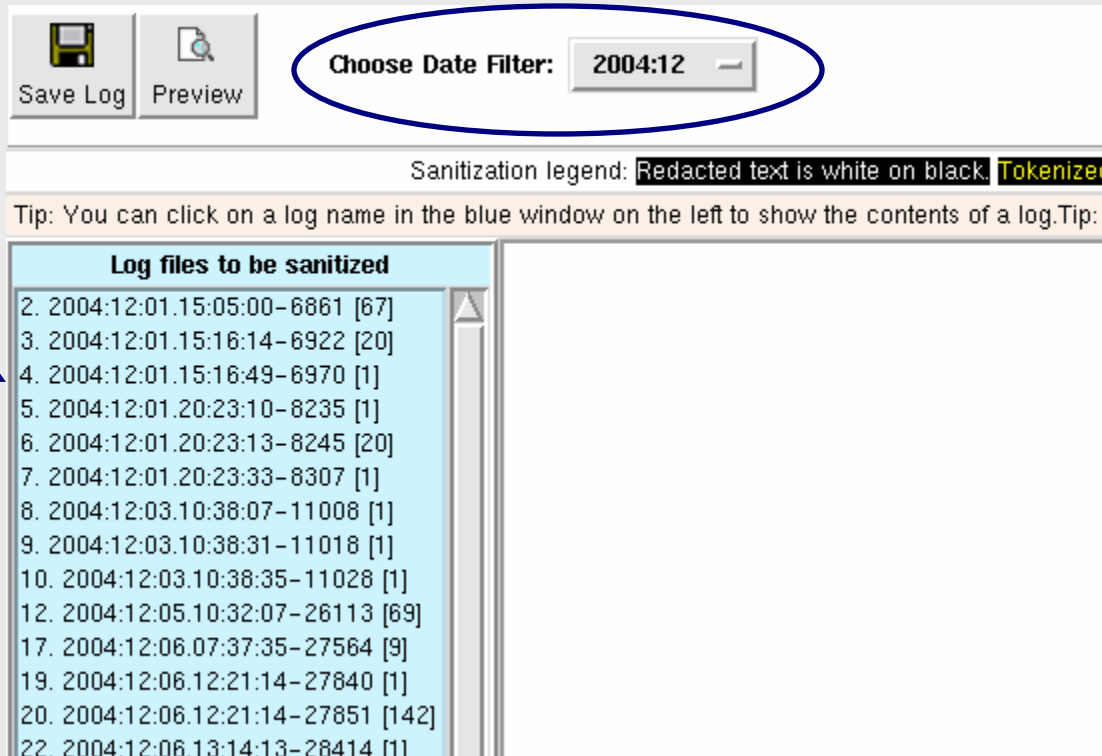
- An example of a session is a telnet instance or an xterm window.
- Each of your sessions is recorded in a single log file.
- Hence your data consists of multiple logs, each containing the commands typed and associated data (timestamps, etc.).
- The name of each log file contains a date & timestamp of when the logged session was initiated. For example:

`2004:12:03.10:38:07-11008.log`

was started on Dec 3 2004, at 10:38 am.

# How do I get started?

1. Choose a date filter or select "All Logs" from the drop-down menu.
2. Your logs will load into the blue panel on the left hand side.



The screenshot shows the Monalog Sanitizer interface. At the top, there are two buttons: "Save Log" and "Preview". To the right of these buttons is a "Choose Date Filter:" label followed by a drop-down menu currently set to "2004:12". A blue oval highlights this date filter. Below the buttons is a "Sanitization legend:" section with two entries: "Redacted text is white on black." and "Tokenized". Below the legend is a tip: "Tip: You can click on a log name in the blue window on the left to show the contents of a log. Tip: .". The main area is divided into two panels. The left panel, titled "Log files to be sanitized", contains a list of log files with their IDs and counts in brackets. The right panel is currently empty.

Log files to be sanitized	
2.	2004:12:01.15:05:00-6861 [67]
3.	2004:12:01.15:16:14-6922 [20]
4.	2004:12:01.15:16:49-6970 [1]
5.	2004:12:01.20:23:10-8235 [1]
6.	2004:12:01.20:23:13-8245 [20]
7.	2004:12:01.20:23:33-8307 [1]
8.	2004:12:03.10:38:07-11008 [1]
9.	2004:12:03.10:38:31-11018 [1]
10.	2004:12:03.10:38:35-11028 [1]
12.	2004:12:05.10:32:07-26113 [69]
17.	2004:12:06.07:37:35-27564 [9]
19.	2004:12:06.12:21:14-27840 [1]
20.	2004:12:06.12:21:14-27851 [142]
22.	2004:12:06.13:14:13-28414 [1]

3. Selecting a log displays it on the right hand side. In this window, you can examine and sanitize your data.

# Types of sanitization: Tokenization & Redaction

---

- **Tokenization** replaces the string by a numbered token. All identical strings in the data map to the same token.
- Hence we know that all instances of the token are the same string, but we don't know what the content of the string is.
- We prefer that you **tokenize** wherever you can, and redact as rarely as possible.
- **Redaction** replaces characters by 'X's. This is the physical equivalent of blacking out text with a black pen, only better, because even the length of the redacted text is hidden!

# How do I tokenize or redact?

- You can select or highlight any red-colored text and either tokenize or redact it by using the buttons on the top right.

The screenshot displays the Monalog Sanitizer interface. At the top right, there are four buttons: **Tokenize** (yellow text on black background), **Redact** (white text on black background), **Untokenize** (white text on grey background), and **Unredact** (white text on grey background). A blue arrow points from the **Tokenize** button to a yellow-highlighted log entry, and another blue arrow points from the **Redact** button to a black-highlighted log entry. The log entry shows a process execution with the following details:

```
[ Wed Dec 1 15:16:38 2004 - Wed Dec 1 15:17:58 2004 ]
Entid: 14
Init: /usr0/fahd/CVS/Sanitizer/IThreat-Pkg/IThreat
Init Time: Wed Dec 1 15:16:38 2004
Init Pid: 6922
User: ruid=9496 euid=9496 suid=9496
Group: rgid=100 egid=100 sgid=100
Line[ 1]: (pid:6922) (uid:9496) [Wed Dec 1 15:16:49 2004]: less Report.pm
Exec[ 1]: (pid:6969) (uid:9496) [Wed Dec 1 15:16:49 2004]: (/usr/bin/less) less Report.pm
Save Time: Wed Dec 1 15:17:58 2004
Save Pid: 6922
Exit: 0
```

On the left side, there is a list of log entries under the heading "s to be sanitized". The main log area has a tip: "Tip: When viewing a log file, you can select editable data and sanitize (redact or tokenize) it."

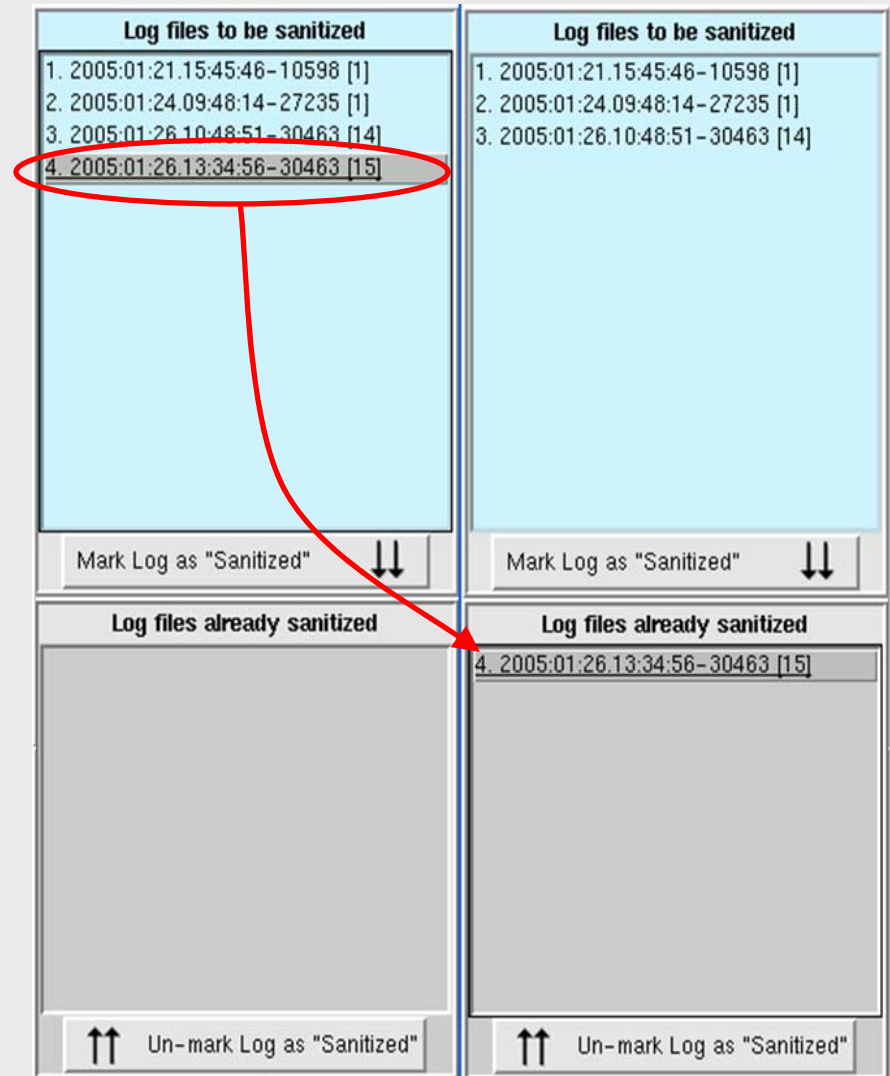
# What happens to my data when I sanitize?

---

- When you sanitize parts of your data, the actual data Monalog collected is not changed.
- Instead, your sanitizations are stored as rules or masks.
- When you finish sanitizing a log, you can mark it “Sanitized” and save your work.
- Then, when you have sanitized a number of log files, you can “Export” the logs.
- Exporting applies the masks to your actual data, and creates a sanitized copy of the data that you can share with us.
- **The original data stays on your machine, and we never have access to it.**

# How do I save my work?

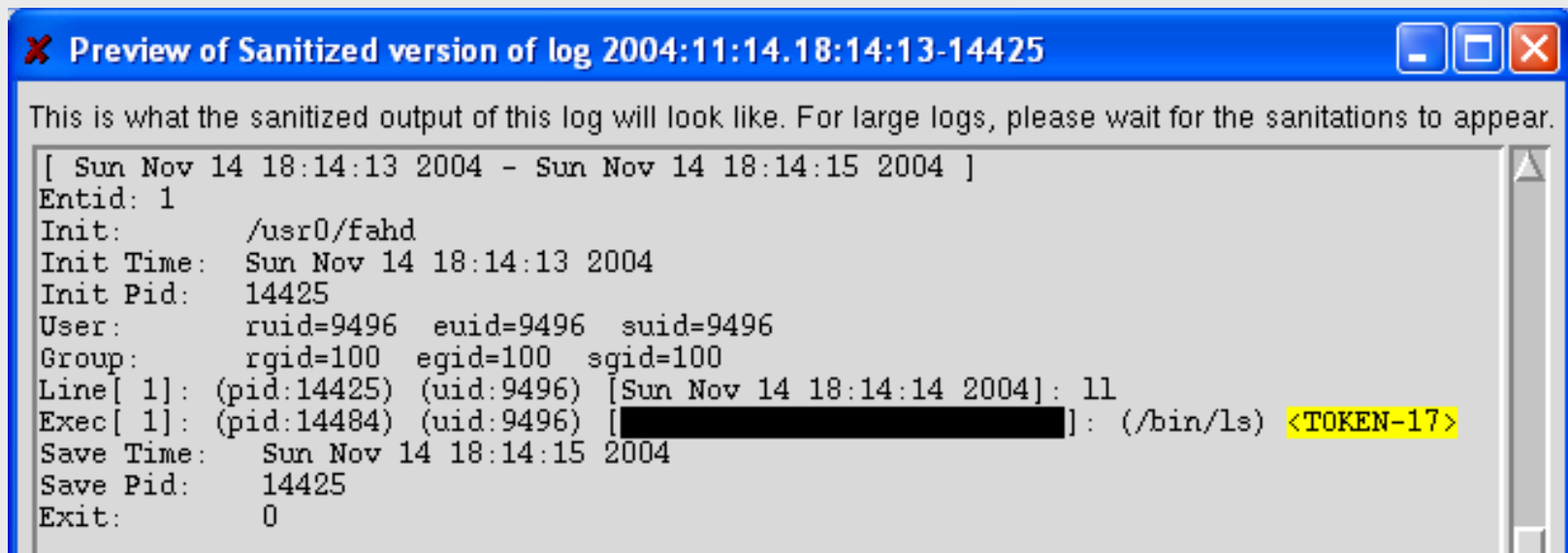
- After sanitizing a log, choose it in the blue window, where the unsanitized log files are listed. Then press the [Mark Log as "Sanitized"] button. The log will move into the list of sanitized logs on the lower left.
- The [Save] button stores both your sanitizations for the log you are working on and which logs you have marked as sanitized.



Before & after marking a log as sanitized

# How can I see the sanitized log?

- You can preview a log to see what the final, sanitized output will look like.
- In the preview mode, tokenized text is replaced by an enumerated token and highlighted in yellow for visual cuing.
- Redacted strings are replaced by the string '<XXXXXX>' and also blacked out for visual cuing.



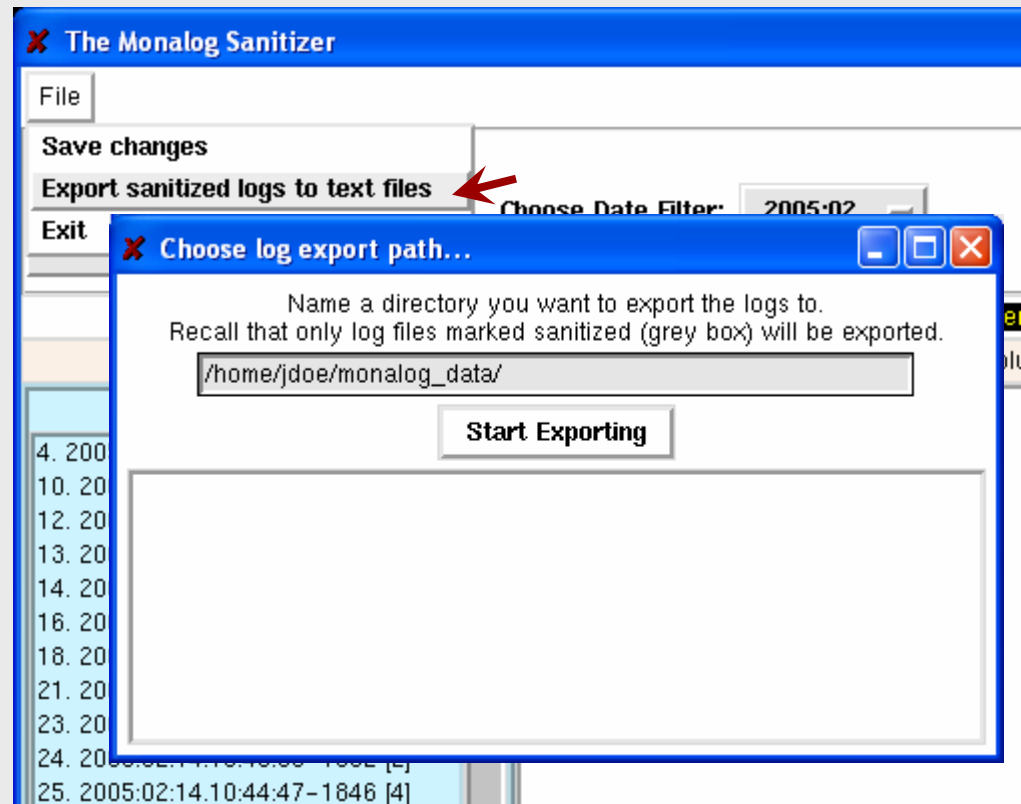
The screenshot shows a window titled "Preview of Sanitized version of log 2004:11:14.18:14:13-14425". The window contains the following text:

```
This is what the sanitized output of this log will look like. For large logs, please wait for the sanitations to appear.  
[ Sun Nov 14 18:14:13 2004 - Sun Nov 14 18:14:15 2004 ]  
Entid: 1  
Init:      /usr0/fahd  
Init Time: Sun Nov 14 18:14:13 2004  
Init Pid:  14425  
User:      ruid=9496  euid=9496  suid=9496  
Group:     rgid=100  egid=100  sgid=100  
Line[ 1]: (pid:14425) (uid:9496) [Sun Nov 14 18:14:14 2004]: ll  
Exec[ 1]: (pid:14484) (uid:9496) [REDACTED]: (/bin/ls) <TOKEN-17>  
Save Time: Sun Nov 14 18:14:15 2004  
Save Pid:  14425  
Exit:      0
```

# How do I share the sanitized data with you?

1. Choose "Export sanitized logs to text files" from the "File" menu.
2. Choose a directory to store the sanitized files in, and click the "Start Exporting" button.

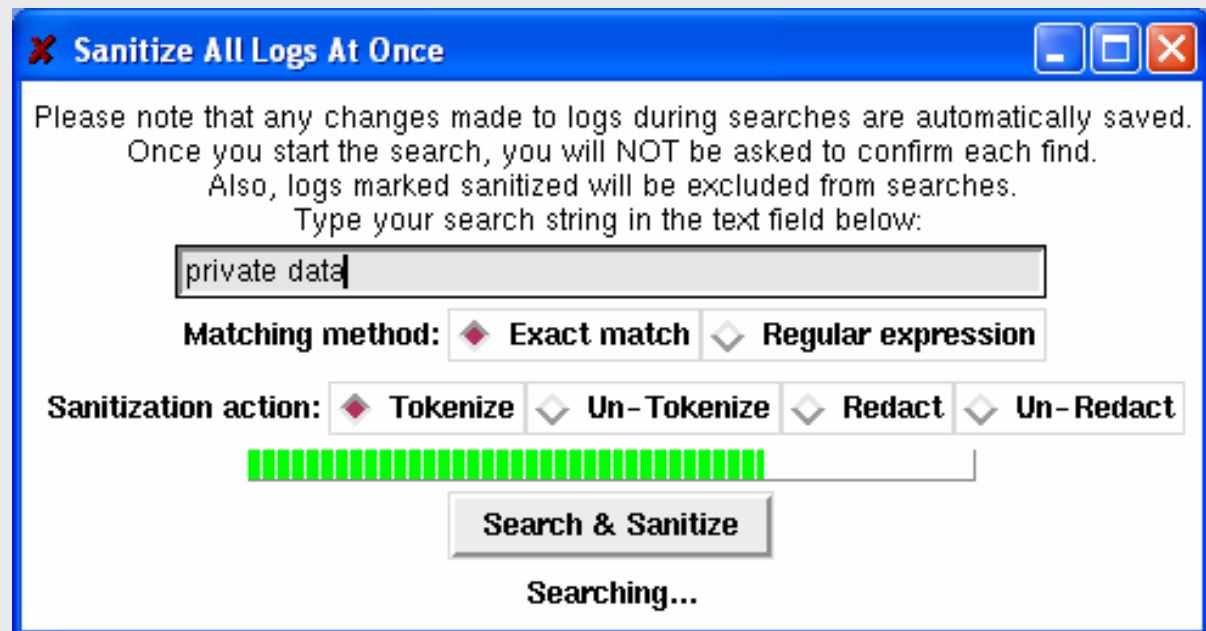
3. This directory will now contain a sanitized copy of your data. You can share this with us!



# How can I sanitize something in all logs at once?

- Pressing the [Redact/Tokenize All Logs] button will allow you to sanitize data across all logs, instead of one log at a time.
- You can choose a strict or regular expression search and either redact or tokenize all resulting matches.

- The search only works on logs you have not already marked as sanitized.



# How do regular expressions work?

- If there is a set of parenthesis in the regular expression string, the sanitization will be applied only to the match in the parenthesis. Otherwise, all the text that matches the regular expression will be matched.

- For example:

/home/\w+ /

will sanitize the whole string in red:

/home/myname /

but

/home/(\w+) /

will sanitize just the inner part (in red):

/home/myname /



# Thank you

---

You should now be ready to use the Sanitizer.

For help on command-line options, try the man pages.

For bugs, comments, or additional help, email  
[fahd+sanitizer@cs.cmu.edu](mailto:fahd+sanitizer@cs.cmu.edu).