Personal Information Tagging for Logs


Sandeep Rao, Shivan Sahib, Ryan Guest

IETF 106, PEARG

Nov 18 2019
Motivations

Log Data

- Information recorded by a system

Uses

- Developers - Debugging, Troubleshooting
- Operations - Performance, Maintenance
- Business - Analysis and Marketing
- Security - Profiling, Monitoring, Incident Response
Life of a Log

Network Application Service

Syslogs
Mobile/App/Web logs
Infrastructure/Security/Service logs
API Gateways

Log Servers

Anonymizer Reformer

Detect or Reform sensitive Data
\{Name, IP, NID, Location\}

Privacy Data Description
schema

Compliance Checker

Cold Storage

Security / Business Analytics

3rd Party/Partner
Log Evolution

- Anonymization to Personalization
- Owned to Shared
- Silos to Collaborated
- Insignificant to Vulnerable
- Monitoring to Monetization
logs and privacy

- privacy
  - which, how, who
  - unregulated to regulated

- challenges
  - subjective - many types of log data
  - dictionary, data-set based training model
  - vendor specific schema / privacy policy
  - no standard - (what) data to be protected and (how) redaction
Privacy Framework for Logs

Source

- Log Generation
  - Privacy Framework for Logs
    - Personal Data Tagging
    - Privacy Policy Assertion

Parser

- Log Detection
  - Detect Personal Data
    - No hard-coding
    - Explicit detection methods

Enforce

- Log Redaction
  - Enforce Privacy Protection
    - Action based on source intents
    - Differential reduction actions
Insert “pii” metadata at Source (Field Level, Log Level)
Current Vendor Approaches (Prior Art)

- dynatrace OneAgent configuration file
- Log4j Framework by Apache
- apigee - custom variables prefixed with "private" and masking configuration
- Avi Networks - Hiding PII in Logs using "sensitive_log_profile"
Abstract

Software applications typically generate a large amount of log data in the course of their operation in order to help with monitoring, troubleshooting, etc. However, like all data generated and operated upon by software systems, logs can contain information sensitive to users. Personal data identification and anonymization in logs is thus crucial to ensure that no personal data is being inadvertently logged and retained which would make the logging application run afoul of laws around storing private information. This document focuses on exploring mechanisms to specify personal or sensitive data in logs, to enable any server collecting, processing or analyzing logs to identify personal data and thereafter, potentially enforce any redaction.
Hackathon Results

Loggly Output (Tagging Method)

Loggly Output (Regex Method)
Next Steps

● Feedback?
● Request for RG adoption?