Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN
Computer Science and
Information Technology

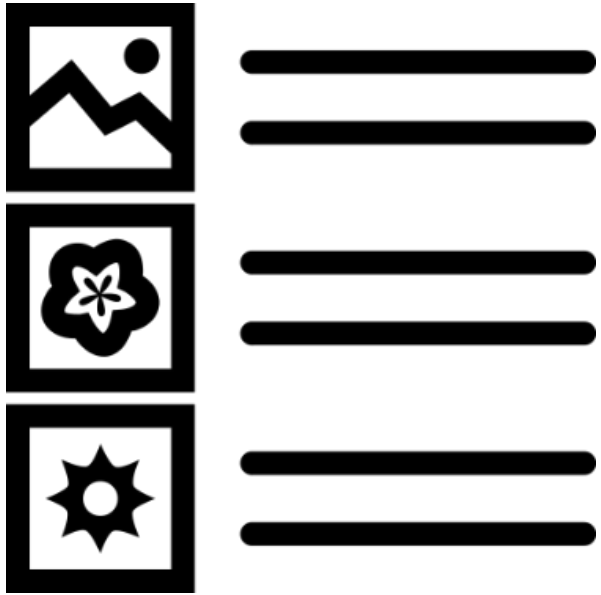
Logging in IT Umgebungen

Einführung

ITIA.H2401
Peter Infanger

Inhalt (Logging)

- Grundsätzliche Gedanken
- Um was geht es da eigentlich?
- Die Dimensionen des Logging:
    - Ziele
    - Daten- und Informationsquellen
    - Sammelart, Speicherung und Formate
    - Abnehmer
    - Filtern/Normalisieren
    - Korrelieren/Analytics
- Kriterien für gutes Logging: die 5 W's



Grundsätzliches - Logging

Fragen die ich mir bei der Vorbereitung gestellt habe:

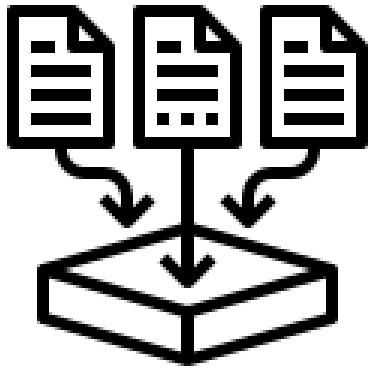
- Für was brauchen wir das eigentlich?
- Nutzen wir das überhaupt sinnvoll, effektiv?
- Könnte man noch mehr daraus machen?

Im Kurs möchte ich Euch auch die Erfahrungen mitgeben, welche ich im Verlauf der Zeit gesammelt habe.

Um was geht es beim Logging?

Sammeln und festhalten von Daten und Informationen welche:

- Aktivitäten
- Ereignisse
- Parameter
- Betriebszustände/Status
- …

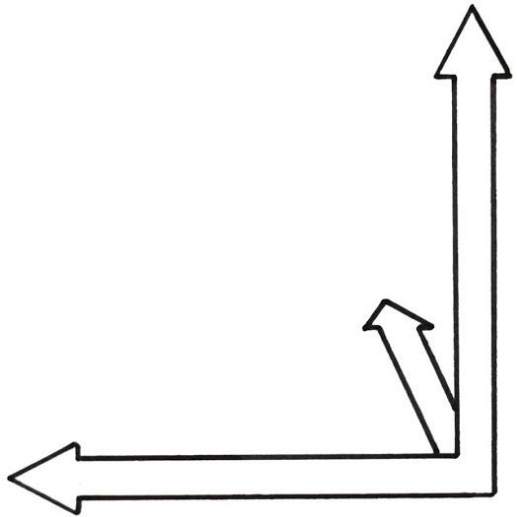


in IT Umgebungen beschreiben

Dimensionen des Loggings

Logging steht nicht im "luftleeren Raum" sondern ist immer eingebettet in einen Kontext:

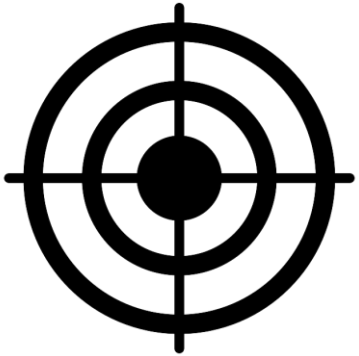
- Ziele (weshalb man überhaupt logged)
- Daten- bzw. Informationsquellen
- Sammelart, Formate und Speicherung
- Filtern und Normalisieren
- Korrelation und Analytics
- Abnehmer



Aber mit welchem Ziel?

- Weshalb macht man das?
- Was könnte der Nutzen dieser Datensammlung sein?
- Bzw. soll man das überhaupt machen? ("man legt ja doch nur Datenfriedhöfe an")
- Aber auch: gegenteiliges Ziel → explizit nicht machen! (man denke an Protonmail)

Ziele beim Logging (1)



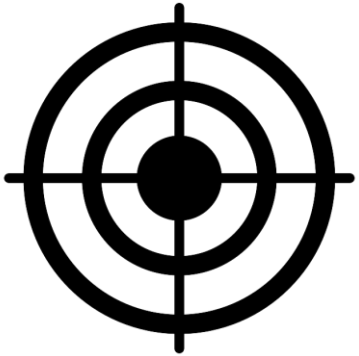
Mögliche Antworten:

- keine spezifischen, einfache weil sich das gehört (default-Einstellung).

Man macht also mit den gesammelten Informationen mal nichts; aber es könnte ja mal was kommen
→ siehe Windows Ereignisanzeige

- Gezielt, selektiv sammeln und verwenden, je nach Quelle und Abnehmer z.B. Webserverzugriffe um Herkunft, Sprache, ... der Webseitenbesucher rauszufinden

Ziele beim Logging (2)

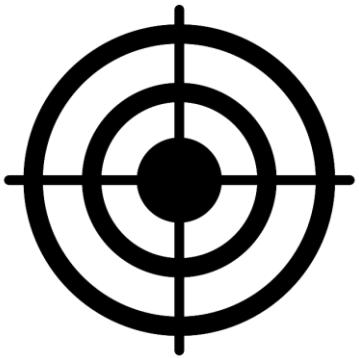


- Bewusstes Sammeln von Informationen mit fortlaufender Analyse zur Erkennung bestimmter Situationen wie:
  - reguläre/irreguläre System-Verhaltensweisen
    → Programmabsturz
  - abnormale/unzulässige Zustände, Situationen
    → CC wird gleichzeitig an verschiedenen Orten benutzt
  - Reguläre Zustandsänderung aufgrund gelaufener Automatisierungs-Tasks

Ziele beim Logging (3)

- Kontrollieren ob eine Automatisierungsaufgabe auch den gewünschten Erfolg hatte → z.B. rapportierter Statuswechsel

- Weil es Teil des Aufgabengebietes ist, z.B. in einem SOC



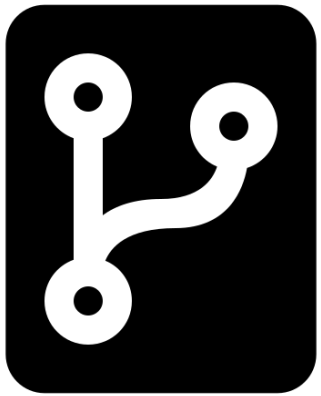
- aufdecken von Sicherheitsvorfällen
  → Fehler beim Login

- Weil ein Regulatorium das fordert (z.B. PCI DSS: https://cybersecurity.att.com/blogs/security-essentials/pci-dss-logging-requirements-explained)

Daten- und Informationsquellen (1)

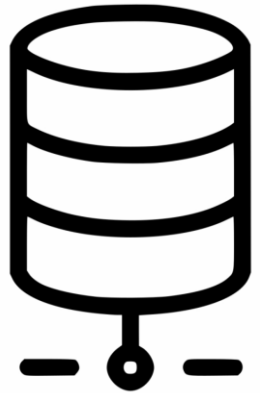
Alle möglichen **Programme & *Prozesse*** auf IT Equipment wie:

- Server, PC
- Router, Switch,
- Firewall, Proxy, ALG's
- DNS, Mail-, Web-Server
- …



→ mehrere Log-Quellen pro Gerät möglich und auch wahrscheinlich!

Daten- und Informationsquellen (2)



Es ist üblich, dass Programme entweder regulär/automatisch oder aufgrund gewisser Situationen Informationen festhalten, sei es in lokalen Speichern (Files, DBs) oder an eine zentrale Stelle zur Speicherung und ev. Weiterverarbeitung weiterschicken.

Dies Funktion wurde wissentlich eingebaut, entweder war es eine Anforderung oder der Programmierer hat das für sich, z.B. zum Debuggen, eingebaut

Daten- und Informationsquellen (3)

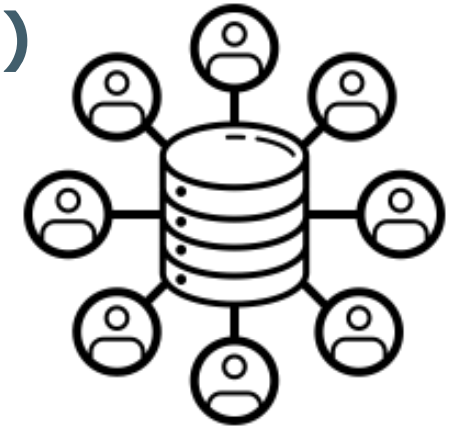
- Achtung: nicht selten ist das Sammeln von Daten mehreren, sich widersprechenden Zielen, unterworfen, siehe z.B. Windows Telemetrie oder Tesla Telemetrie*.

- Der Benutzer muss also selber abwägen, ob er dieser Datensammlung zustimmt/haben will oder nicht.

Detail am Rande:
IoT Devices sind relativ "schweigsam" was das Logging angeht (um Ressourcen zu sparen → billiger zu produzieren)

* https://www.tesla.com/support/privacy

Art der Datensammlung (1)



Das Sammeln kann:
- umfassend, unspezifisch
- gezielt, selektiv
- oder als ein Mix dazwischen

erfolgen

In der Realität wird der Mix am häufigsten anzutreffen sein.

- Nicht alle möglichen Daten werden für das angestrebte Ziel benötigt
- Alle Daten sammeln würde zu umfangreich
- Je nach Abnehmer werden die Daten eventuell noch aufbereitet.

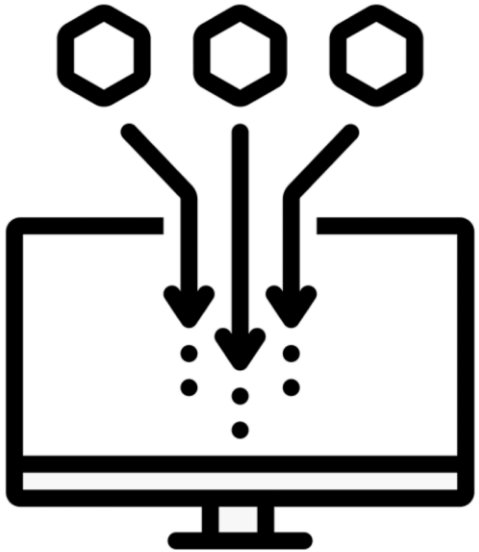
Art der Datensammlung (2)

Im Prinzip realisierbar:

- lokal
- remote

Lokal:

- einfachste Variante, wenig Abhängigkeiten (falls genug Speicherplatz)



Art der Datensammlung (3)

Remote ev. Cloud:
eine oder mehrere dedizierte Sammelstellen

Vorteile:

→ effizient z.B. in einem DBMS
→ entlastet lokale Ressourcen
→ definierte Anlaufstelle für Analytics
→ Cloud ideal für verteilte Systeme oder wenn Quellen ebenfalls in der Cloud sind

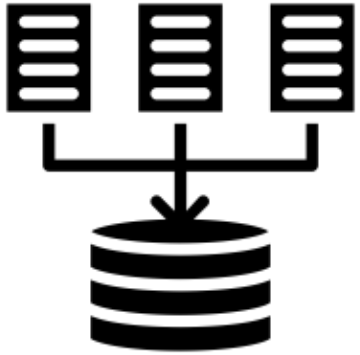
Nachteile:

→ setzt Connectivity voraus (was passiert aber bei Unterbruch?)
→ belastet Netzwerk

Art der Speicherung (1)

Files:
- einfachste Form
- direkter Zugriff möglich
  → aber Zugriffsschutz wenn heikle Informationen drin sind?

Datenbanken:
- Alle Vorteile eines DBM-Systems
- zusätzliche Ressourcen benötigt (ev. auch Lizenzen)



Hochschule Luzern
Computer Science and
Information Technology

Art der Speicherung (2)

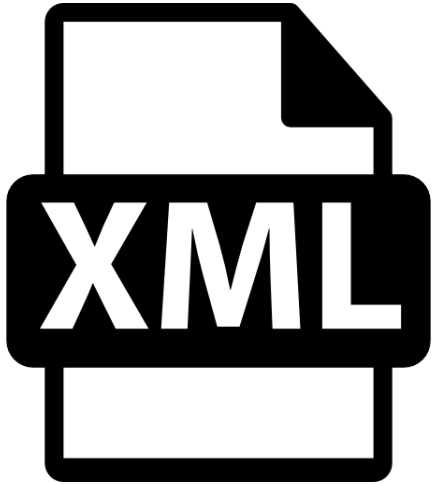
Punkte die zu beachten sind:

- Welche Datenmenge?
  (und schlussendlich wie weit zurück in der Vergangenheit sollen die Daten reichen? – Retention Time)
- Wird "direkt access" benötigt oder kann der Zugriff auch langsamer erfolgen (z.B. bei Speicherung auf Tapes)
- Wie sieht es mit dem Manipulationsschutz der gespeicherten Daten aus?
  → Revisionssicherheit!

Allgemein wird unter einer revisionssicheren Archivierung verstanden, digitale Daten aufzubewahren und zwar so, dass die rechtlichen Anforderungen in Bezug auf Ordnungsmässigkeit, Vollständigkeit, Sicherheit, Verfügbarkeit, Nachvollziehbarkeit, Unveränderlichkeit und Zugriffsschutz erfüllt sind.

(https://www.diamant-software.de/blog/revisionssicherheit/)

und die Formate … (1)



Logging erfolgt in verschiedenen Formaten wie:

(einfach) lesbare Form wie:
- ASCII
- JSON
- XML

nach bestimmten Regeln formatiert, z.B. Windows: https://docs.microsoft.com/en-us/windows/win32/eventlog/event-log-file-format)

oder in binärer Form (z.B. EBCDIC*)
- üblicherweise kompakter als lesbare Form

*Extended Binary Coded Decimal Interchange Code

und die Formate … (2)

Folgen aufgrund der Formate:

- Speicherbedarf
  → Textdaten lassen sich besser komprimieren!

- ev. nicht direkt zu verarbeiten, man muss zuerst umformen (decodieren, entzippen, …)

- Wie sieht es mit der Langzeitspeicherung aus? (können in Zukunft die Daten noch verarbeitet werden?)

Formate: pros/cons

	XML logging	Syslog text logging	Text file logging	Proprietary logging
Consumption mode	Mostly machine reading	Mostly manual reading	Only manual reading	Only machine reading
Common use case	Security logging	Operational logging, debugging logging	Debugging logging (enabled temporarily)	High-performance logging
Example	Cisco IPS security appliance	Most routers and switches	Most application debugging	Checkpoint firewall logging, packet capture
Recommendation	Use when a rich set of structure information need to be transferred from producing to consumer and then analyzed	Add structure such as name=value to simplify automated analysis; use for most operational uses	Add structure to enable automated analysis, if the logs are to be left enabled during operations	Use for super high performance uses only
Disadvantages	Relatively low performance, large log message sizes	Lack of log message structure makes automated analysis complicated and expensive	Typically the logs can only be understood by the application developers	Not human readable without a dedicated application that can convert binary into text

Quelle: Logging and Log Management; ISBN13 9781597496353

Abnehmer der gesammelten Daten und Informationen

- Operating
- Security
- Software-Entwicklung
- Compliance
- Marketing
- …

→ Sichtweisen auf die Daten können also sehr unterschiedlich sein! (unterschiedliche Interessen)

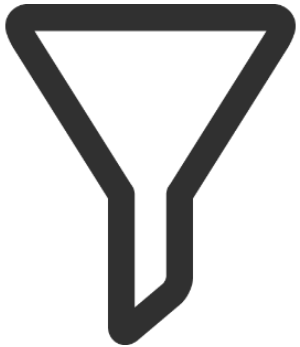
→→ hat Folgen fürs Sammeln und ggf. für die Aufbereiten der Daten

Filtern / Normalisieren (1)

Je nach Abnehmer der Logs lohnt es sich die Daten zu Filtern und/oder zu Normalisieren

- Filtern:
  - Nur das was interessiert wird überhaupt erfasst bzw. gespeichert
    → Problem: weiss man jeweils im voraus, was interessant sein könnte?

  - Empfehlung: zurückhaltend filtern (um nichts wichtiges zu verpassen) in der Analytics lässt sich dann das Interessante schon rausholen!



Hochschule Luzern
Computer Science and
Information Technology

Filtern / Normalisieren (2)

Normalisierung vereinfacht die Weiterverarbeitung, kann aber recht komplex sein.

Was ist einfacher, eine IP Adresse immer gleich zu formatieren oder folgendes Regex zur Suche/Überprüfung einer IP*?

```
^([01]?\d\d?|2[0-4]\d|25[0-5])\.([01]?\d\d?|2[0-4]\d|25[0-5])\. ([01]?\
   d\d?|2[0-4]\d|25[0-5])\.([01]?\d\d?|2[0-4]\d|25[0-5])$
```

* Wie könnte das wohl bei IPv6 aussehen? ☺

Filtern / Normalisieren (2)

- Beispiele für sinnvolles Normalisieren:
  - gleiche Datums- bzw. Zeitformate

10/15/21 11:18:03 PM → 15.10.2021 23:18:03

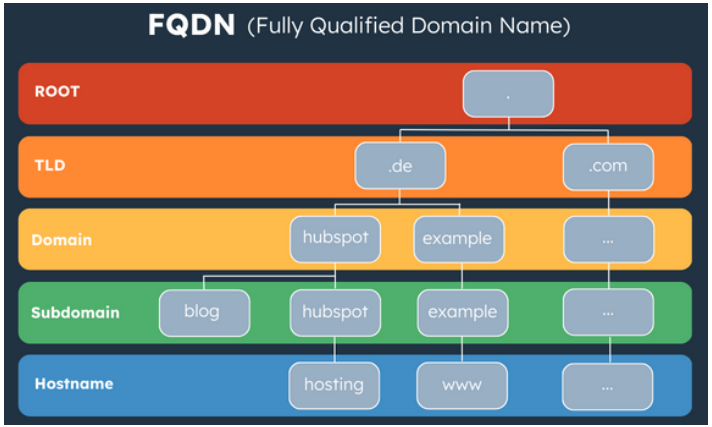
  - IP Adressen

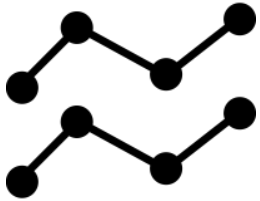
1.2.3.4 → 001.002.003.004

  - Hostnamen FQDN

sugus.cookies.sweets.com

host / SD? SLD TLD





Hochschule Luzern
Computer Science and
Information Technology

Korrelation (1)

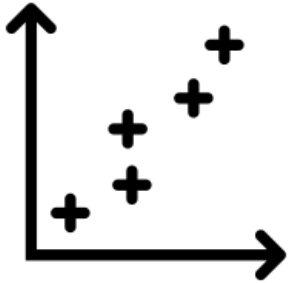
Führt man von verschiedenen Systemen die Logs zusammen so können sie korreliert werden um z.B. Gemeinsamkeiten zu entdecken.

Man benötigt aber "Erkennungsmerkmale" für diese Gemeinsamkeiten → woher soll man sonst wissen, dass ein Eintrag im Log der DB mit einem Web-Request zusammenhängt?

Man kann beispielsweise Events oder Felder aus einem Log zur Korrelation heranziehen

Eine Korrelation ("Wechselbeziehung") beschreibt eine Beziehung zwischen zwei oder mehreren Merkmalen, Zuständen oder Funktionen. Die Beziehung **muss** keine kausale Beziehung sein.

Korrelation (2)



Hochschule Luzern
Computer Science and
Information Technology

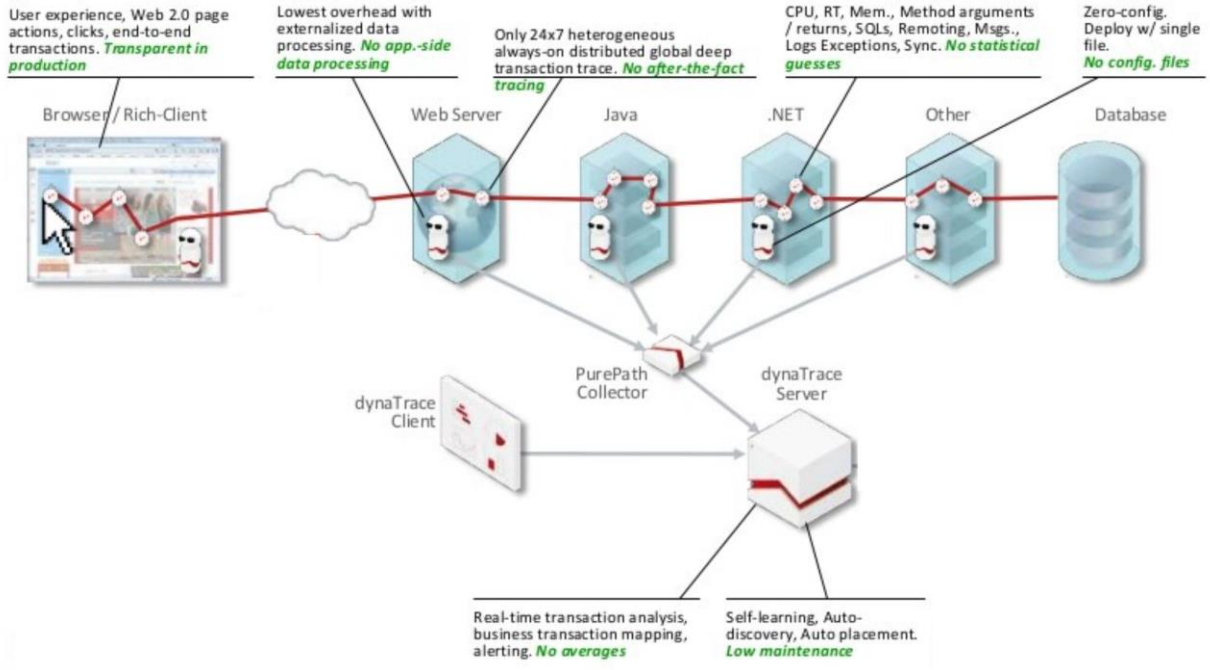
Eine Möglichkeit ist der Zeitstempel: alles was gleichzeitig passiert könnte zusammengehören, aber:

1. Das setzt aber eine gemeinsame Zeitbasis voraus!!!
2. Was wenn die Verarbeitung über mehrere System längere Zeit braucht?

   → mögliche Lösung wäre ein Merkmal/ID/Tag, welche mit durch die Systeme "wandert" und man so die Zusammenhänge erkennen könnte.

Beispiel Dynatrace mit n-Tier Anwendung

- Eingefügte ID ermöglicht komplette Betrachtung des Abarbeitens eines Requests, Aufrufes oder sonstiger Bearbeitungssequenz



Quelle:
ITJobZone.biz

Analytics (1)

Korrelation bietet zwar die Möglichkeit, aufgrund dem Eintreten verschiedener Ereignisse oder Zustände z.B. Aktionen auszulösen (aka. Rule Based Correlation).

Sie ist aber nicht dazu geeignet, z.B. Abweichungen von einer Baseline oder überschreiten eines Schwellwertes zu erkennen .

Für diese Zwecke benötigt man typischerweise statistische Funktionen

Moderne Systeme verwenden dazu auch gerne Methoden aus dem Gebiet von AI & ML

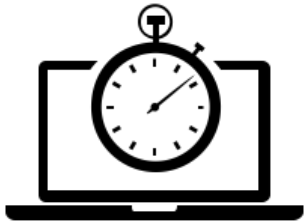
Analytics (2)

Klassische Themen die mit Analytics angesprochen werden können/müssen sind:

- Schwellwerte (Thresholds)
- Anomalie-Detektion
- Windowing (Beobachtungsfenster)

Schwellwerte (Thresholds)

- Schwellwerte werden verwendet, um festzustellen, wann etwas einen Basiswert überschreitet. Angenommen man weiss, dass bei einem Server durchschnittlich 10 fehlgeschlagene Anmeldungen pro Stunde registriert werden. Wenn diese Zahl plötzlich auf 23 ansteigt, dann sollte man das untersuchen.



Analytics (3)

Anomalie Detektion

- Bei der Erkennung von Anomalien geht es darum, Dinge zu entdecken, die noch nie zuvor gesehen wurden. Die Verwendung einer Baseline ist eine Form der Anomalie-Erkennung. Dies funktioniert weil man die Konfidenzintervalle beobachtet. Ein klassischer Anwendungsfall wäre z.B. ein DoS Angriff wo die Anzahl der Webzugriffe unaufhörlich steigt.

Windowing (Beobachtungsfenster)

- Diese legen üblicherweise das Zeitfenster fest, in welchem man eine Beobachtung macht. Ein Anwendungsfall wären z.B. die Anmeldezeiten der Benutzer (Normalfall zwischen 08:00 und 18:00). Sobald sich dieser verschiebt bzw. die Anmeldungen ausserhalb des Fensters geschehen wäre das eine Untersuchung wert.

Kriterien für gutes Logging

Was macht also ein "gutes" Logging aus?

Welche Informationen sind notwendig, damit eine Log-Meldung für den angedachten Verwendungszweck reicht?

Es gibt viele Arten von Logging und noch mehr Arten von Geräten, die Logs erstellen so dass es schwer ist, ein einziges Kriterium zu definieren

→Im Allgemeinen sollten Protokolle Auskunft geben über die **5 W**'s

Hochschule Luzern
Computer Science and
Information Technology

The 5 W's of Logging

- **What** happened (with appropriate detail; "Something happened" is not usually particularly useful).
- **When** did it happen (and when did it start and end, if relevant).
- **Where** did it happen (on what host, what file system, which network interface, etc.).
- **Who** was involved.
- **Where** he, she or it came from.

Quelle: Logging and Log Management ISBN13 9781597496353

Aaaaaber ...

Die 5 W's sind eigentlich nur das Minimum!

Weiter Informationen die man zudem gerne sehen würde sind:
- **Woher** kriege ich noch mehr Informationen?
- **Wie** zuverlässig ist die Information?
- **Was** alles ist betroffen?

und (wenn wir schon am träumen sind☺):
- **Was** könnte als nächstes passieren?
- **Was** ist sonst noch passiert, was mich interessieren sollte?
- **Was** sollte ich dagegen tun?

Quelle: Logging and Log Management ISBN13 9781597496353

Konsequenz

Wenn man all die vielen W's betrachtet wird einem sofort klar, dass man Logging ohne ein bewusstes Ziel gar nicht richtig umsetzen kann.

Ist man sich im Klaren, was man mit Logging erreichen möchte, so kann man sich in all den angesprochenen Dimensionen aufstellen, die richtigen Quellen auswählen, die Daten entsprechend aufbereiten und adäquat speichern und mit optimalen, zielgerichteten Methoden analysieren.

Beispiel 1

SW Entwicklung benötigt Unterstützung bei der Konfiguration und Setup eines API-Gateways über den man externe Services bezieht bzw. selber anbietet. Das Gateway logt nur intern in eine eigene DB und Console (syslog) und bietet nur beschränkten Monitorzugang via spezielle Clientsoftware.

→Lösung: Verarbeitungslog wird mit zusätzlicher Information ergänzt und in ein externes Logging System zur besseren Analyse weitergereicht

Demo: API Gateway

Beispiel 2

SOC braucht für Monitoring des Internet-Zugriffs via Proxy detaillierte Informationen. Standardmässig wird aber nur geloggt, ob ein Request geblockt oder durchgelassen wurde.

→Lösung: anreichern des Logs mit zusätzlichen Informationen und weiterleiten ans SIEM des SOC

Demo: Proxy Log

Ranum’s Laws of Logging (von 2004)

Ranum's first law of Log Analysis:
"Never keep more than you can conceive of possibly looking at"
(Bewahre niemals mehr auf, als du dir überhaupt vorstellen kannst zu betrachten.)

Ranum's second law of Log Analysis:
"The number of times an uninteresting thing happens is an interesting thing"
(Die Anzahl, wie oft eine uninteressante Sache passiert, ist eine interessante Sache.)

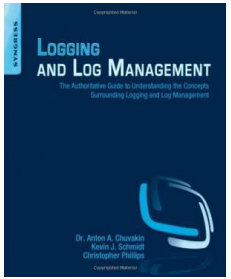
Ranum's third law of Log Analysis:
"Keep everything you possibly can except for where you come into conflict with the First Law"
(Behalte alles, was du kannst, ausser du kommst in Konflikt mit dem ersten Gesetz.

Quelle: https://honor.icsalabs.com/pipermail/firewall-wizards/2004-October/017383.html

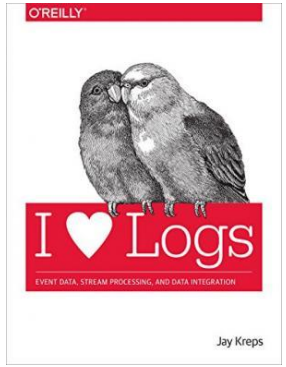
Best Practice

- Zeitstempel!
- Redundanzen reduzieren
- Normalisieren bei der Quelle lohnt sich
- Informationsgehalt der Logs prüfen, hat man alles mögliche rausgeholt → Ziele bekannt
- Selektiv loggen (aber aufgepasst beim Filtern!)

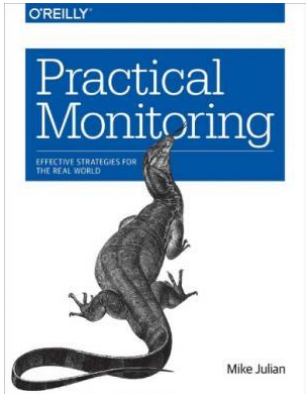
Quellen und Literatur



- Logging and log management: The authoritative guide to understanding the concepts surrounding logging and log management
  Anton A. Chuvakin, Kevin J. Schmidt
  Publisher: Syngress ISBN: 1597496359 ISBN13: 9781597496353



- I Heart Logs: Event Data, Stream Processing, and Data Integration
  Jay Kreps
  Publisher: O'Reilly Media ISBN: 978-1-491-90938-6



- Practical monitoring : effective strategies for the real world, Julian, Mike
  Publisher: O'Reilly Media ISBN: 1491957328