LU08a - LogFile Basics

Lernziele

- Die wichtigsten Log-Dateien eines Computers nennen können.
- Zielsetzung der verschiedenen Log-Filles erklären können.
- Aufbau eines Logfiles darlegen können.

Einleitung

Logfiles sind von entscheidender Bedeutung für Computer, besonders für den Betrieb eines Webservers. Mit ihrer Hilfe können sowohl generelle Webserver-Probleme wie auch Angriffe auf den Server erkannt werden.

LogFiles protokollieren alle Benutzertransaktionen (Zugriffe) sowie Zustände, Aktionen und Fehler des WebServers. Aus ihnen lassen sich neben Informationen zum Benutzungs- und Navigationsverhalten auch Informationen und Rückschlüsse über Zugangswege und Verlinkungen von Webinhalten auswerten.

Definition Logdatei bzw. LogFile

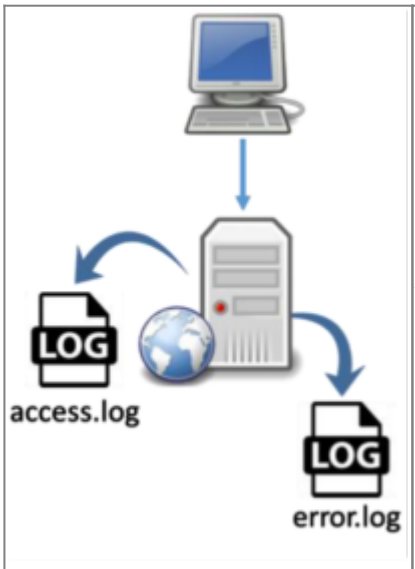
Eine Logdatei (auch Ereignisprotokolldatei, Englisch „log file“) enthält das automatisch geführte Protokoll aller oder bestimmter Aktionen von Prozessen auf einem Computersystem.

LogFiles eines WebServers sind strukturiert aufgebaute Textdateien. Sie bestehen aus ASCII-Zeichen, in denen die Zugriffe auf den WebServer gespeichert werden. Der WebServer schreibt dazu für jede http-Transkation (e.g. Aufruf einer Seite, Ausführen eines Scripts) eine Log Zeile mit Informationen, die diese Transkation betreffen. Eine Zeile im LogFile stellt somit eine Anfrage an den WebServer dar.

escuela_access_log

```
200.90.177.197 - - [07/Nov/2004:04:14:17 -0300] "GET /novedades/rss/ HTTP/1.1" 200 3284
200.73.40.132 - - [07/Nov/2004:04:14:19 -0300] "GET / HTTP/1.1" 200 7550
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/escmovil1.jpg HTTP/1.1" 200 7748
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/novedades1.jpg HTTP/1.1" 200 7952
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/mapa1.jpg HTTP/1.1" 200 8035
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/lado1.jpg HTTP/1.1" 200 12964
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/lado2.jpg HTTP/1.1" 200 8340
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/escuela1.jpg HTTP/1.1" 200 7858
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/servicios1.jpg HTTP/1.1" 200 6835
200.73.40.132 - - [07/Nov/2004:04:14:20 -0300] "GET /img_index/departamentos1.jpg HTTP/1.1" 200 7900
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/instructivos1.jpg HTTP/1.1" 200 8597
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/calendarios1.jpg HTTP/1.1" 200 7217
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/organizaciones1.jpg HTTP/1.1" 200 7543
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /imagenes/fmellado.jpg HTTP/1.1" 200 2675
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/cabierta.png HTTP/1.1" 200 5464
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/wap_ing_uchile_cl.jpg HTTP/1.1" 200 5419
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /imagenes/logo_ucursos.jpg HTTP/1.1" 200 36799
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /img_index/novedades2.jpg HTTP/1.1" 200 8089
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /novedades.htm HTTP/1.1" 200 655
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /barraizquierda2.htm HTTP/1.1" 200 3258
200.73.40.132 - - [07/Nov/2004:04:14:21 -0300] "GET /main_novedades.htm HTTP/1.1" 200 514
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/lupa.gif HTTP/1.1" 200 1566
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /head_principal.htm HTTP/1.1" 200 3361
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/organizaciones1.jpg HTTP/1.1" 200 7543
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/novedades2.jpg HTTP/1.1" 200 8089
200.73.40.132 - - [07/Nov/2004:04:14:22 -0300] "GET /img_index/escmovil2.jpg HTTP/1.1" 200 8077
```

Arten von LogFiles eines Webservers



Die meisten WebServer führen im Standard die zwei LogFiles *Acceess.log* und *Error.log*. Es lassen sich aber noch zusätzliche LogFiles für spezielle Anwendungen oder Auswertungen konfigurieren, sogenannte Custom_logs.

Access.log - Das Zugriffsprotokoll

Im Zugriffsprotokoll werden sämtliche grundlegenden Informationen über jede HTTP-Transaktion gespeichert. Es protokolliert auf welche Dokumente zu welchem Zeitpunkt zugegriffen wurde. Mit diesem Protokoll kann ein einen Überblick über die Funktion und Auslastung des WebServers verschafft werden.

Error.log - das Fehlerprotokoll

Im Fehlerprotokoll werden alle Fehler notiert. Gerade in der Aufbauphase eines WebServers, bei

Konfigurationsänderungen oder der Installation von Programmen, kann ein ausführliches Fehlerprotokoll schnell Hinweise auf mögliche Fehlerursachen geben.

Aufbau eines Logfiles

Das Format der Einträge in das Zugriffsprotokoll kann konfiguriert werden. Die übliche Variante ist das CLF (Common Log Format). Ein Eintrag in diesem Format sieht folgendermassen aus.

```
120.0.0.7 - - [06/Jan/2016:11:14:34 +0100] "GET /~lara/ HTTP/1.1" 200
13872 "http://www.google.de/search?q=lara&start=20&sa=N" "Mozilla/5.0
(Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/56.0.2924.87 Safari/537.36"
```

Feld	Beispiel	Beschreibung
Host [%h]	120.0.0.7	IP-Adresse oder vollständiger Domainname des zugreifenden Rechners.
Ident [%l]	-	Wenn die IdentityCheck-Anweisung in der Konfigurationsdatei aktiviert wurde und auf dem Clientrechner ein ident-Daemon installiert ist, wird hier der vom Client gelieferte Name des Benutzers auf dem Clientrechner eingetragen, sonst „-“.
Authuser [%u]	-	Bei Zugriffen auf zugangsgeschützte Dokumente oder Verzeichnisse wird hier der verwendete Benutzername eingetragen, sonst „-“.
Date [%t]	06/Jan/2016:11:14:33 +0100	Dieses Feld zeigt das Datum und die Uhrzeit der Anfrage, sowie Informationen zu der Zeitzone an.

Feld	Beispiel	Beschreibung
Status [%>s]	200	Dieses Feld zeigt den Statuscode der Antwort an und damit, ob die Anfrage erfolgreich war, oder ob ein Fehler aufgetreten ist. Die wichtigsten Codes: 200 = OK 206 = Partial Content 301 = Moved Permanently 302 = Found 304 = Not Modified 401 = Unauthorised (password required) 403 = Forbidden 404 = Not Found 500 = Internal Server Error
Bytes [%b]	13872	Hier wird die Größe der vom Server an den Client ausgelieferten Daten (ohne HTTP-Header) in Bytes angegeben.
Referer [%{Referer}i]	http://www.go.ch/search?q=Java&start=20&sa=N	Der Referer gibt an, welche Seite ein Besucher vor dem aktuellen Aufruf besucht hat und beschreibt somit die Navigation eines Benutzers zwischen verschiedenen Web-Seiten über definierte Linkbeziehungen.
UserAgent [%{User-Agent}i]	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/537.36	Der User Agent beschreibt der vom Benutzer verwendete Browser.

Nachteile der Logfile Analyse

Die Nachteile der Logfile Analyse:

- **Caching und Proxys:** da ein Logfile nur Daten aufzeichnen kann, die durch direkte Serverzugriffe entstehen, entfallen im Protokoll alle Zugriffe, die über den Cache-Speicher des Browsers sowie über Proxyserver erfolgen. Der Traffic einer Seite wird demnach mit der Logfile Analyse nur ungenau bestimmt.
- **Regelmässige Updates nötig:** damit Logfiles stets korrekte Zahlen liefern, muss die Software zur Datenerhebung immer wieder vom Webmaster aktualisiert werden. Hierdurch entsteht zusätzlicher Wartungsaufwand.
- **zusätzlicher Speicheraufwand:** da Logfiles automatisiert protokolliert werden, können die Datenmengen für die LogFiles bei hohen Besucheraufkommen schnell sehr gross werden, da jeder Serverzugriff registriert wird. Wer selbst Logfile Analysen von grossen Websites durchführt benötigt daher zusätzliche Speicherressourcen.
- **Aufwändige Datenaufbereitung bei grossen Datenmengen:** für eine Log Analyse müssen die einzelnen LogFiles zuerst in ein Programm zur Datenaufbereitung eingepflegt werden. Dies bedeutet besonders bei vielen Datensätzen zusätzlichen Arbeitsaufwand.

- **Kein Tracking von Widgets oder AJAX:** ein Logfile kann nur Daten speichern, die durch Serveranfragen entstehen. Werden z.B. Aktionen innerhalb einer Seite mit Hilfe von AJAX durchgeführt, finden sich diese nicht in dem Logfile wieder, da es sich hierbei nicht um echte Serverabfragen handelt.
- **Ungenaue Zuordnung von Visits:** wenn ein User eine dynamische IP-Vergabe beim Surfen verwendet und mehrfach auf eine Website zugreift, erscheinen im Logfile mehrere Zugriffe, obwohl es jeweils nur ein User war. Dadurch wird die Traffic-Auswertung ungenau. Gleiches gilt dann, wenn mehrere Nutzer mit der gleichen IP auf eine Website zugreifen. Diese werden dann nur als ein Besucher gezählt.
- **Weniger Daten:** Im Vergleich mit Webanalysetools bietet die Logfile Analyse weit weniger Daten. Sie kann zum Beispiel wichtige KPIs wie Absprungraten nicht anzeigen.

Quellennachweis

- https://www.teialehrbuch.de/Kostenlose-Kurse/Apache/15434-Protokolldateien.html
- https://www.feistyduck.com/library/apache-security/online/apachesc-CHP-8.html
- http://httpd.apache.org/docs/2.0/mod/mod_log_config.html
- http://www.linux-praxis.de/lpic1/manpages/logrotate.html
- http://www.webalizer.org/
- https://www.sans.org/reading-room/whitepapers/logging/detecting-attacks-web-applications-log-files-2074
