LOGDATEN

Logdaten und Datenschutz?

Kein Widerspruch!

Seit vielen Computergenerationen produzieren die meisten IT-Systeme Logdaten. Doch seit mindestens genauso langer Zeit werden Logdaten von vielen Administratoren leichtfertig ignoriert. Dabei enthalten Logdaten wichtige Informationen, die dabei helfen können, in großen Netzwerkumgebungen effizient Hard- und Softwareprobleme, sowie Ressourcenengpässe festzustellen und zu lokalisieren. Darüber hinaus ist ein immer deutlicherer Trend dahingehend erkennbar, Logdaten zur Erkennung von Sicherheitsproblemen und Angriffen heranzuziehen. Doch obwohl sich Logdaten zu diesem Zweck sehr gut eignen, zögern viele Unternehmen und Institutionen aus Gründen des Aufwandes und des Datenschutzes, eine Logdatenanalyse in ihr Sicherheitskonzept zu integrieren.

Jedes Unternehmen ist heutzutage von der kritischen Infrastruktur Internet und seinen Diensten existenziell abhängig. Doch mit der zunehmenden Bedeutung des Internets als Kommunikationsmedium steigt auch die potentielle Bedrohung, ein leichtes Ziel von Kriminellen zu werden, die von der Entwicklung dieses Mediums ebenfalls profitieren wollen. Intrusion Detection Systeme (IDS) gehören daher für jedes größere Unternehmen zur Grundausstattung. Bei IDS lassen sich verschiedene Realisierungsansätze unterscheiden. Networkbased IDSs analysieren die direkte Kommunikation auf der Netzwerkebene, indem die sensiblen Paketinhalte auf Schadmuster hin untersucht werden. Somit sind sie in der Lage, Angriffe noch „auf der Leitung" zu erkennen, bevor das Zielsystem erreicht wird. Ein bekanntes Tool dieser Gattung ist Snort. Ein weiterer Realisierungsansatz sind Hostbased IDSs, die unter anderem auf Logdaten operieren können. Im Gegensatz zu den Kommunikationsdaten auf der Leitung liefern Logdaten qualitativ viel hochwertigere Informationen, da sie komplexe Ereignisse beschreiben und somit auf einem höheren Level anzusiedeln sind als bloße Kommunikationsparameter. Da Logdaten erst im Laufe eines Angriffs, nämlich nach jedem einzelnen Angriffsschritt, entstehen, ermöglichen sie die Analyse eines Kommunikationsergebnisses durch die Bewertung von Fakten und realen Ereignissen. Logdatenanalyse ermöglicht also, Aussagen über die Reaktion eines Zielsystems auf einen Angriff zu treffen und diesen zu rekonstruieren. Ein Grund, der dazu führt, dass Logdatenanalyse trotz dieses Mehrwerts so selten eingesetzt wird, ist die Tatsache, dass die meisten Analysesysteme nur statistische Auswertungsmöglichkeiten bieten. Dazu müssen allerdings Logdaten über einen längeren Zeitraum persistent gemacht werden und verursachen somit datenschutzrechtliche Probleme. Das Institut für Internetsicherheit - if(is) - verfügt nun mit seinem Logdaten-Analyse-System (LAS) über ein System, das sowohl eine Echtzeitalarmierung, als auch eine statistische Langzeitanalyse bereitstellt und dabei einen optimalen Kompromiss zwischen Datenschutz und IT-Sicherheit umgesetzt hat [MRO08].

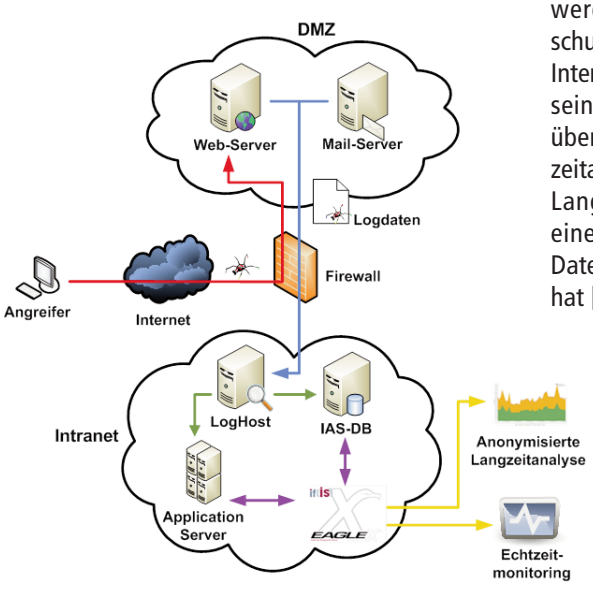

Abb. 1: Die Architektur des Logdaten-Analyse-Systems

Das Logdaten-Analyse-System

Die grundsätzliche Idee des LAS ist, dass jede Kommunikation mit einem Dienst über das Internet in einem Protokoll aufgezeichnet wird. Folglich muss auch die Kommunikation eines Angreifers mit dem jeweiligen Dienst erfasst werden. Da dessen Verhalten aber stark von dem eines normalen Nutzers differiert, müssen die von ihm verursachten Logdaten signifikante Muster aufweisen.

Die Logdaten aller überwachten Dienste werden als Livedatenstrom an einen zentralen LogHost übertragen (Centralized Logging). Auf diesem LogHost läuft eine LogSonde, ein Dienst, der einerseits die Logdaten für die Langzeitanalyse anonymisiert und in einer Datenbank ablegt, und andererseits den Livedatenstrom auf Angriffe hin untersucht. Die Logdaten, anhand derer ein Angriff festgestellt wurde, werden dann an einen Application Server übertragen. Mit Hilfe des EagleX-Clients, Bestandteil des umfassenden Internet-Analyse-Systems (IAS) des if(is), lassen sich dann beide Datenquellen abrufen [PoPr06].

Anonymisierte Langzeitanalyse

Die Langzeitanalyse stützt sich auf das Prinzip der Zählung anonymisierter Deskriptoren. Ein Deskriptor ist die Definition eines Ereignisses, dessen Auftreten für feste Zeitintervalle gezählt und abgespeichert wird. Durch dieses Prinzip ist es möglich, die wesentlichen Informationen der Logdaten anonymisiert persistent zu machen, da keine sensiblen Daten wie Benutzernamen, E-Mail- oder IP-Adressen festgehalten werden. Die Deskriptoren bestimmter Ereignisse lassen sich dann im zeitlichen Häufigkeitsverlauf graphisch darstellen. Mit Hilfe einer statistischen Auswertung dieser Daten können dann Kommunikationsmuster und Profile beschrieben, Angriffssituationen und Anomalien erkannt und Prognosen zu Mustern und Angriffen abgegeben werden. Darüber hinaus ist es möglich, die anonymisierten Daten mehrerer Betreiber des Logdaten-Analyse-Systems in einer globalen Sicht zu aggregieren und somit einen Überblick über den aktuellen Zustand des Internets zu bieten.

Echtzeitanalyse

Die Echtzeitanalyse arbeitet zwar im Gegensatz zur Langzeitanalyse auf den

sensiblen Logdaten, gewährt dem Nutzer des Logdaten-Analyse-Systems aber lediglich Zugang zu den Daten, die potentiell in Zusammenhang mit einer Schadsituation stehen. Auf dem LogHost werden geeignete Algorithmen und Analysetechniken wie Signaturerkennung und Schwellenwertanalyse eingesetzt, um die Logdaten eines schadhaften Verhaltens in nahezu Echtzeit zu identifizieren.

Ziel dabei ist, nach der Erkennung eines erfolgreichen Angriffs einen Alarm auszulösen, so dass eine unmittelbare Reaktion, nämlich das Ergreifen von Schutzmaßnahmen gegen eine konkrete Bedrohung, möglich ist. Derartige Maßnahmen zur Schadensminimierung könnten das Anhalten von Diensten, Schließen von Ports oder Sperren von Benutzerkonten sein. Die Logdaten, die durch die Analysemodule einem Angriff zugeordnet werden konnten, werden dann im Schadszenario in der GUI dargestellt, allerdings nach 24 Stunden automatisch wieder entfernt, ohne dabei automatisiert persistent gemacht worden zu sein. Für den Fall, dass die Daten zu einer Angriffssituation auch zu einem späteren Zeitpunkt noch relevant sein könnten, besteht die Möglichkeit, diese bei Bedarf manuell zu speichern. Um den Datenschutz noch weiter voranzutreiben, wäre es denkbar, die Logdaten innerhalb der GUI in Abhängigkeit von Nutzerrechten pseudonymisiert anzuzeigen.

LogReduction

Der wichtigste Eingabeparameter für ein derartiges System ist die Menge an Logdaten, die ihm zugeführt wird. Logdaten treten nicht nur in schlecht konfigurierten Umgebungen in großen, unübersichtlichen Mengen auf. Daher ist es nötig, die relevanten Daten bereits im Voraus einzugrenzen (LogReduction). Grundsätzlich würde man annehmen, je mehr Logdaten zur Analyse herangezogen werden, desto mehr Informationen liegen vor und desto genauere Aussagen können getroffen werden. Das bedeutet für den Fall, dass zu wenig geloggt wird, dass das System dann nicht funktioniert. Wird aber ausnahmslos jede Information geloggt, funktioniert das System ebenso wenig, da einerseits wichtige Informationen „im Rauschen untergehen" und andererseits die großen Datenmengen die Performance des Systems stark beeinträchtigen. Ob ein Logdatum aber relevant ist, lässt sich nur schwer erkennen. Entweder ist die Information nur implizit enthalten, oder sie ergibt sich erst durch den Kontext zu anderen Logdaten. Ein fehlgeschlagener Loginversuch für sich allein beispielsweise stellt noch kein großes Sicherheitsrisiko dar. Folgen diesem aber viele weitere Loginversuche, liegt der Verdacht eines Einbruchversuchs nahe. Grundsätzlich kann davon ausgegangen werden, dass der Anteil sicherheitsrelevanter Informationen in Logdaten kleiner als 5% ist. Im derzeitigen Entwicklungszustand unterstützt das LAS iptables-Logdaten, anhand derer die LogReduction veranschaulicht werden soll. Die Informationen, die bei einer Firewall geloggt werden, sind hauptsächlich Paket-Header-Informationen. Um aber bei den vielen Paketen einen extremen Overhead zu vermeiden, wird nicht jedes Paket geloggt, sondern durch ein Regelwerk spezifiziert, welche Teilmenge erfasst werden soll. Das korrekte Verhalten des LAS ist von der richtigen Konfiguration des Logging-Regelwerks abhängig.

Es hat sich als sinnvoll erwiesen, immer das erste Paket des Verbindungsaufbaus, also ein Logdatum pro TCP-Verbindung, sowie alle von der Firewall zurückgewiesenen oder verworfenen Pakete zu erfassen. Alle Pakete, die über das erste Paket einer Verbindung hinaus die Firewall passieren, stimmen in vielen Kommunikationsparametern mit dem ersten überein und enthalten damit einen hohen Anteil redundanter Information. Abgelehnte Pakete hingegen sind unter Umständen sehr interessant, denn ein Verstoß gegen die Firewall Policy ist eine unerwünschte Kommunikation und daher Grund genug für weitere Untersuchungen. Insgesamt lässt sich die Menge aller Pakete somit auf eine kleine Teilmenge reduzieren, die geloggt wird und die nötigen Informationen enthält, um den Großteil aller Angriffe erkennen zu können.

Ergebnisse

Das LAS hat im Untersuchungszeitraum vom 1. bis zum 31. März insgesamt 555 Angriffe auf das Netzwerk der FH Gelsenkirchen und des if(is) festgestellt. Im Schnitt sind dies 18 Angriffe pro Tag. 56% dieser Angriffe waren allerdings nur Portscans. Insgesamt 45% machten allein die Scans des UDP-Ports 53 für DNS aus. Der Angriffstyp, der am meisten Verkehr erzeugt hat, war die SSH-Dictionary-Attacke, bei der versucht wird, durch Raten von Zugangsdaten Zugriff auf die Shell des Zielsystems zu erlangen. Angriffe dieser Art über 12 Stunden hinweg, parallel gegen bis zu 30 oder mehr Zielrechner gerichtet, waren keine Seltenheit. Neben diesen Angriffen und weiteren Portscanvarianten, traten noch Spam, meist von asiatischen IP-Adressen stammend, und Angriffe über HTTP, häufig aggressive Webcrawler, auf. DoS-Attacken konnten im gesamten Untersuchungszeitraum nicht festgestellt werden.

Fazit

Die Logdatenanalyse ist speziell in der Angriffserkennung ein sehr mächtiges Werkzeug, denn mit Hilfe von Logdaten lassen sich nicht nur Angriffe erkennen, sondern auch Aussagen über deren Verlauf sowie deren Ausgang treffen. Der Datenschutz ist bei der Logdatenanalyse ein ernst zu nehmendes Problem, das einen Einsatz in vielen Fällen verhindert. Allerdings gibt es Konzepte, um die bestehenden Bedenken zu zerstreuen, somit eine umfassende Logdatenanalyse praktikabel zu gestalten und dadurch bestehende Sicherheitskonzepte zu erweitern.

Literatur
[MRO08] Johannes Mrosek: „Entwicklung eines logdatenbasierten Analysesystems", Bachelor Thesis, Institut für Internet-Sicherheit if(is), FH Gelsenkirchen 2008

[PoPr06] N. Pohlmann, M. Proest: „Datenschutzkonforme Kommunikationsanalyse zum Schutz der IT-Infrastruktur", IT-Sicherheit – Management und Praxis, DATAKONTEXT-Fachverlag, 1/2006

Autoren

B. Sc. Johannes Mrosek

Prof. Dr. Norbert Pohlmann
Institut für Internet-Sicherheit – if(is)
Fachhochschule Gelsenkirchen