Manejo eficiente de eventos de Seguridad de honeynets mediante BigData

Eduardo Casanovas, Mariano García Mattio, Francisco Coenda, Carlos Tapia, Fernando Boiero

Instituto Universitario Aeronáutico, Facultad de Ingeniería, Av. Fuerza Aérea 6500, Córdoba, Provincia de Córdoba, Argentina
ecasanovas@iua.edu.ar, magm3333@gmail.com, franciscocoenda@gmail.com, ctapia@iua.edu.ar, fboiero@gmail.com

Resumen. Es un hecho de que las organizaciones van incorporando cada vez más a las honeynets como herramientas de ciberseguridad flexibles y adaptables, que entregan resultados sumamente ricos respecto de ataques informáticos, pero ello plantea el problema de cómo gestionar la gran masa de datos obtenidos. Esta problemática obliga a avanzar hacia un siguiente paso en su madurez, implicando lograr una eficiente gestión de los registros de logs que se generen en dichas honeynets, a los efectos de poder aprovechar la información clave, y a partir de allí generar contramedidas específicas para los ataques. Es necesario continuar con el estudio de maneras eficientes de almacenar y utilizar grandes volúmenes de datos obtenidos en las honeynets, para así obtener el máximo provecho de la herramienta y aportar información de ataques informáticos para su correcta prevención.

Palabras clave. honeynet, honeypot, big data, Seguridad Informática, logs, eventos, estadísticas.

1 Introducción

Los términos honeypot y honeynet aún resultan desconocidos para muchas organizaciones, o sólo existen en el marco de pequeños proyectos de Seguridad Informática, fundamentalmente en el ámbito académico. Sin embargo, para aquellas organizaciones que efectivamente hayan implementado este tipo de herramienta como medida de Seguridad, más temprano que tarde se da el inconveniente del gran volumen de datos que acarrea el monitoreo de los sensores de la red ficticia.

Estas honeynets o redes ficticias cuyo único fin es ser atacadas para obtener datos de las intrusiones, son una herramienta invaluable a la hora de recabar información de primera mano y extremadamente actualizada sobre ataques

informáticos. Estos beneficios entregados tienen como contrapartida cierta capacidad de cómputo para que los sensores de las redes ficticias efectivamente puedan realizar su trabajo de "engañar" a los atacantes y así crean estar atacando recursos productivos de una organización, pero fundamentalmente, el insumo primario es el espacio en disco y la capacidad para procesar posteriormente grandes volúmenes de datos. No es inusual que al cabo de algunos días de funcionamiento expuesto a Internet, un sensor haya almacenado cientos de miles de registros, los cuales eventualmente deberán ser filtrados, ordenados o tratados para derivar en reportes útiles para la gestión de las honeynets y las estadísticas de ataques que se buscan obtener, siendo esta información tan valiosa para poder priorizar las contramedidas y dirigir más puntualmente cualquier inversión a realizar en Seguridad Informática.

2 Análisis de Datos de una Honeynet.

Un honeypot por sí solo produce información, pero no el suficiente volumen de datos para llevar a cabo un análisis respecto de los tipos de ataques que están sucediendo. No es posible mediante un único tipo de honeypot marcar patrones de ataques o tendencias. Es la combinación de distintos honeypots, que generan una honeynet, producen un gran volumen de datos útiles que pueden ser explotados para obtener información respecto a los ataques que se están llevando a cabo contra la infraestructura.

Una de las tecnologías utilizadas para analizar los datos de una honeynet es HoneyLog framework que utiliza PHP y JavaScript, siendo la librería D3.js la utilizada para visualizar la información. Este framework se compone de un modelo cliente-servidor, donde el webserver se compone principalmente de un módulo analítico [1].

Luego, teniendo como variable al tiempo, un enfoque para representar visualmente los datos es heatmap. También cabe destacar que se utiliza un servicio de Geo IP API, para poder trazar desde dónde se están realizando los ataques [1].

Otro enfoque es analizar los paquetes capturados por la honeynet y a través de procesos matemáticos representar la información en ejes de coordenadas, tanto 2D como 3D [2]. En estos análisis no se ha visualizado la utilización de herramientas particulares para procesar la información, solamente se ve el número de ocurrencias de las variables y de qué forma pueden ser representadas en ejes de coordenadas cartesianas [2] [5].

Otra aproximación cada vez más adoptada, es la de data mining, aplicado a la gestión de logs de honeynets. Como ejemplo de ello, es posible implementar un software desarrollado en C-Sharp, que recibe la información obtenida de la honeynet, la cual es volcada primero en archivos de texto plano para luego ser

pasada a través de este programa y así obtener un output con información del número de ocurrencias de acuerdo a una serie de criterios que se establecen a priori, antes de dar comienzo con el análisis de datos [3] [5].

Otros esquemas se basan en implementar soluciones y tratar de mejorarlas o realizarles algún tuning a fin de agregar mayor información o mejorar la información que entrega o su procesamiento. Se observaron avances en este aspecto, conformando el almacén de información vía un RDBMS y no un NoSQL storage [4].

3 Montado de la Honeynet.

En la primera fase de investigación se montó una honeynet con los sensores Snare y Kippo. Snare es un honeypot que finge ser un servidor web con un sitio que imita un website real, pero que tiene como único destino ser atacado y registrar los intentos de intrusión. Por su parte, Kippo es un honeypot de que registra los ataques de fuerza bruta y la interacción que tiene el atacante con la shell del sistema.

Debido al gran volumen que se obtiene de datos de esta red señuelo, se tomó la decisión de aplicar Big Data sobre el conjunto con el fin de poder verificar y determinar la factibilidad de buscar patrones en común entre los distintos ataques que se registran e incluso poder determinar si existen tendencias que nos puedan indicar o predecir los tipos de ataques. El foco se ha puesto en establecer un manejo del gran volumen de datos que nos permita su posterior explotación a través de herramientas específicas.

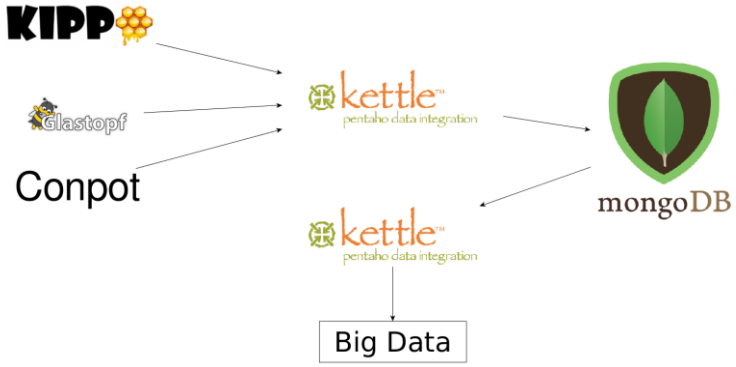

Imagen 1: esquema honeynet-Big Data

4 Centralización de los Datos.

Una vez que la honeynet fue expuesta a internet y empezó a generar datos, se estudió y generaron los procesos necesarios para almacenar la información en MongoDB. Se decidió trabajar con mongo DB ya que es una BD orientada a

documentos y puede manejar JSON. Esto dota de flexibilidad al sistema de análisis a la hora de tomar los datos de los distintos sensores y almacenarlos en MongoDB.
Para manipular los datos y almacenarlos en MongoDB se trabajó con Pentaho Data Integration, El proceso consiste en tomar los datos que se encuentran en los logs de los sensores, procesarlos e insertarlos en la base de datos. Este proceso se conoce como ETL. En esta primera fase, se ha establecido que las ETLs sean ejecutadas manera automática en periodos regulares.

4 Conclusión

En esta primera fase de investigación se ha logrado montar exitosamente una honeynet funcional y la infraestructura necesaria para almacenar los datos que produce ésta.
Se continuará estudiando otros sensores para ampliar la cantidad de datos que provee la honeynet y se realizarán pruebas de performance sobre la infraestructura a fin de mejorar los tiempos y procesos. También se buscará agregar medios que permitan visualizar la información almacenada en la base de datos, ya sea a través de dashboards o distintos reportings.

5 Referencias bibliográficas:

[1] Identifying Network Traffic Features Suitable for Honeynet Data Analysis. Pavol Sokol, Lenka Kleinová, Martin Husák. IEEE.

[2] Identifying Network Traffic Features Suitable for Honeynet Data Analysis.Mohammed H. Sqalli, Syed Naeem Firdous , Khaled Salah , and Marwan Abu-Amara. IEEE.

[3] Identifying Scanning Activities in Honeynet Data using Data Mining. Mohammed H. Sqalli, Shoieb Arshad, Mohammad Khalaf, Khaled Salah. IEEE.

[4] Improving Honeynet Data Analysis. Camilo Viecco. IEEE.

[5] Intrusion discovery with data mining on honeynet. Jian Yin.