UNIVERSIDAD DE LA CIUDAD DE AGUASCALIENTES

MAESTRÍA EN CIENCIA DE DATOS

GESTIÓN DE PROYECTOS DE CIENCIA DE DATOS

“Detección de anomalías de tráfico en servidores web”

-

Propuesta de Valor

Alumno:

E23S-18014: MITSIU ALEJANDRO CARREÑO SARABIA

Periodo Enero 2024 - Junio 2024, Aguascalientes, Ags

ÍNDICE

INTRODUCCIÓN

Cualquier empresa que ofrezca servicios en internet y especialmente aquellas que se encargan de recopilar, almacenar o procesar información sensible, confidencial, personal e identificable deben conocer y mitigar los riesgos que conlleva tener un servidor con conexión a internet. Además, está documentado (https://www.embroker.com/blog/cyber-attack-statistics/) que el impacto de un cyberincidente tanto en los ambitos económico, reputacional y técnico puede ascender a los millones de dolares de no ser detectado y manejado en tiempo y forma correspondientes. Por ello se propone un sistema de monitoreo y alerta cuando se detectan patrones de tráfico anómalos según el historial de la plataforma misma.

DESCRIPCIÓN DEL PROBLEMA

Con la expansión del acceso a servicios de internet, así como la creciente disponibilidad de dispositivos de distintas categorías para conectarse a la red, la demanda y tráfico de servicios web se encuentra en constante aumento. Mucho se ha desarrollado en términos de escalabilidad de infraestructura así como adopción de soluciones distribuidas para dar servicio a la creciente demanda.

Pero un aspecto muchas veces ignorado es la importancia de que las organizaciones evalúen cómo es realmente la interacción entre sus clientes y la infraestructura disponible, analizarlo puede ser útil para responder muchas preguntas cómo ¿se está obteniendo el máximo rendimiento de la infraestructura, o es necesario escalar? ¿Acceden desde un dispositivo móvil, una computadora, una pantalla inteligente? E incluso si el contenido que tiene el cliente es sospechoso, anómalo o malicioso. Es por ello que este trabajo propone una metodología y desarrollo de un sistema para procesar la enorme cantidad de conexiones que recibe un servidor, de manera automática, confiable y partiendo del tráfico habitual del servidor.

SOLUCIÓN PROPUESTA

Se propone una solución integral de monitoreo y alerta en la detección de tráfico anómalo mediante métodos basados en análisis topológicos y en modelos no supervisados de aprendizaje máquina para evaluar y detectar valores anómalos en las peticiones que recibe un servidor web. Mediante este método es posible evaluar nuevas peticiones basado en el tráfico histórico del servidor y obtener un índice de

similitud respecto a solicitudes pasadas, con ello es posible detectar anomalías o contenido malicioso y tomar acciones tanto preventivas como correctivas.

INNOVACIÓN Y DIFERENCIACIÓN COMPETITIVA

Un sistema que permita evaluar el tráfico en tiempo real, permite notificar a administradores de sistema y personas relevantes mucho más rápido, tanto para procesar posibles solicitudes maliciosas como para monitorear el performance del sistema así como de la infraestructura asociada, ampliando incluso a acciones preventivas como incrementar horizontal o verticalmente los recursos, agregar servicios adicionales de manejo de tráfico tales como balanceadores de carga, etc.

VIABILIDAD TÉCNICA

Actualmente para el desarrollo de este sistema se está partiendo de las configuraciones default de la tecnología NGINX reverse-proxy, el cuál actualmente es el líder del mercado dada su filosofía gratuita y open source. También se están usando las bitácoras de registros por default, las cuales tienen el formato “remote_addr - remote_user [local_time] request status body_bytes_sent http_referer http_user_agent gzip_ratio” (https://docs.nginx.com/nginx/admin-guide/monitoring/logging/) por lo que desacopla el sistema de monitoreo de una configuración única o específica, permitiendo incluso extenderlo a otros proveedores de reverse-proxy tal como apache (segundo en concentración de mercado).

Finalmente se considera la funcionalidad “Logging to syslog” (https://nginx.org/en/docs/syslog.html) para ofrecer la funcionalidad de monitoreo en tiempo real donde las solicitudes de recursos del servidor sean enviadas al sistema de monitoreo donde son alimentadas al algoritmo de aprendizaje automático y según sea catalogada la petición, se notifique a las personas adecuadas.

PLAN DE EJECUCIÓN

Para la ejecución del proyecto se propone el siguiente plan de acción:

- Realizar la recopilación cruda de las bitácoras de registro de distintos sistemas.

- Aplicar procesos de limpieza, desagregación de información y generar su contraparte estructurada.
- Realizar los análisis topológicos correspondientes para entender la naturaleza y relación de la información.
- Definir y desarrollar la arquitectura y algoritmo de aprendizaje automático y aplicar procedimientos de entrenamiento con los datos recabados.
- Desarrollar la implementación syslog y montar el algoritmo en un servidor para recibir flujos de información en tiempo real así como desarrollar la parte técnica de las notificaciones vía email o celular.

PRÓXIMOS PASOS

Actualmente se está realizando la exploración topológica y definiendo la técnica de aprendizaje automático más efectiva, el cuál es fundamental para la correcta aplicación del modelo y las fases subsecuentes del plan de ejecución. Cuándo el nivel de madurez del algoritmo sea el apropiado, así como los desarrollos tecnológicos esperados, se planea incrementar el número de fuentes y/o formatos aceptados por el algoritmo para permitir la integración con sistemas de bitácoras extra.