UNIVERSIDAD DE LA CIUDAD DE AGUASCALIENTES

MAESTRÍA EN CIENCIA DE DATOS

GENERACIÓN DE NEGOCIOS BASADOS EN DATOS

“Detección de anomalías mediante aprendizaje automático en tráfico de servidores web”

Alumno:

E23S-18014: MITSIU ALEJANDRO CARREÑO SARABIA

Periodo Agosto 2024 - Diciembre 2024, Aguascalientes, Ags

Problema que se desea resolver o atender

El tráfico de un servidor web provee datos confiables sobre accesos, solicitudes y procesamiento de peticiones y el contexto bajo el que se usan sus recursos, pero el volumen de información generada es tan grande que un análisis manual no es viable. Entender los usos típicos y diferenciarlos de los atípicos es una herramienta poderosa que aplicada en tiempo real permitirá mejorar la calidad, y resguardo de la información contenida.

“El viejo sistema de ‘almacenar y procesar’ obstacularizará en gran medida el sistema de cómputo [...] porque le toma tiempo procesar, por ello es necesario mirar a tecnologías que hagan todo el procesamiento al mismo tiempo” (W Sardjono et al, 2021)
Analizar los registros de tráfico web permite no solo entender la manera en que se consume la información que contiene un servidor, sino también detectar si el uso generalizado se transforma, o si existen anomalías. Dado el volumen de información que se genera, y la creciente sensibilidad de los datos alojados, aplicar herramientas de aprendizaje automático permitirá agilizar y perfeccionar cualquier proceso manual.

Dado el contexto anterior se propone desarrollar un producto que permita a las empresas proveedoras de servicios en internet monitorear el uso y abuso de sus productos implementando tecnologías de desarrollo de software tradicionales e incluyendo un motor de detección de anomalías basado en técnicas de aprendizaje automático, análisis topológicos y big data. Con este sistema se pretende apoyar a las empresas a poder monitorear, prevenir y corregir cualquier servicio que ofrezca en internet, por lo que pueden conocer mejor cómo usan e interactúan realmente sus usuarios, detectar actividad anómala e incluso detectar cambios y patrones de comportamiento esta información es valiosa en el apoyo a toma de decisiones importantes.

Propuesta de valor o factor diferenciador

El principal factor diferenciador es el desarrollo del producto basado totalmente en aprendizaje automático, este enfoque basado en datos, involucra el desarrollo de técnicas de aprendizaje automático “state of the art”, aplicando novedosas técnicas y métodos.

Este proyecto pretende implementar técnicas aplicadas en software de protección grado empresarial a un software nicho cuyos clientes tienen requisitos más laxos, ofreciendo monitoreo y protección a bajo costo comparado con soluciones multipropósito.

Objetivos

Objetivo general

El objetivo del producto es ofrecer monitoreo inteligente de tráfico a nivel aplicación, el cuál permita la toma de decisiones en un ámplio espectro de contextos por ejemplo, mitigación de ciberataques, insights de uso de aplicaciones y servicios, detección de cambios en patrones de uso, escalabilidad y aprovisionamiento de recursos en la infraestructura.

Objetivos específicos

Analizar las técnicas y procesos tanto tradicionales como de aprendizaje automático mediante los cuales se analiza tráfico web actualmente: Mediante este análisis se espera conocer las herramientas y procedimientos mediante los cuales se realizan análisis similares a los propuestos en la solución.

Enumerar las características y casos de uso de sistemas de monitoreo y alerta efectivos: A pesar de proponer técnicas de análisis innovadoras, el sistema sigue siendo en su núcleo, un sistema de alertas, por lo que es necesario evaluar cuales son las características de los sistemas de alertas útiles.

Desarrollar un modelo de detección de anomalías basado en aprendizaje automático: Esté objetivo comprende el punto innovador de la solución propuesta, ya que integra técnicas de vanguardia que ayudan a obtener mejores resultados de manera más rápida, cualidades que se esperan del proyecto.

Análisis de mercado o clientes potenciales

Desde inicios del milenio, ha existido un claro incremento en la cantidad de usuarios y servicios ofertados en internet, este incremento se debe principalmente a dos factores, el número de usuarios aumenta debido a mejoras en la accesibilidad a internet desde un amplio rango de dispositivos, y a la rápida adopción de servicios alojados en internet, ofreciendo una mejora disponibilidad, agilidad, seguimiento personalizado entre otras características. Ambos factores fomentan a más empresas a migrar sus servicios a internet, a pesar del grado de capacidad y experiencia tecnológica y de innovación que posean.

El producto está enfocado a estas pequeñas y medianas empresas que ofrecen sus servicios en internet y requieren de herramientas que les permita monitorear y conocer el desempeño de sus productos en internet.

Por nombrar algunos potenciales clientes se listan:
Instituciones educativas en las que cada vez es más común habilitar plataformas y servicios que se ofertan en internet, pero a la vez deben resguardar la confidencialidad de los datos estudiantiles y del cuerpo docente y administrativo.

Organizaciones de la sociedad civil especialmente aquellas con causas de transparencia y justicia gubernamental y política, que pueden ser objetivo de ciberataques.

Periodismo y medios masivos de comunicación similar a las organizaciones de la sociedad civil, sitios de periodismo pueden ser objeto de ataque para reprimir la libertad de expresión.

Se considera que el modelo de negocio con mayor viabilidad es un esquema de suscripción mensual ofreciendo distintos niveles de servicio:

- **Básico (Costo estimado $3,000 mensuales por plataforma)**- Únicamente se realiza análisis básico estadístico y monitoreo de uso de plataforma (a que secciones acceden sus clientes, cuáles transacciones son más tardadas, etc)
- **Básico avanzado (Costo estimado $9,000 mensuales por plataforma)** - Se realiza un análisis general del sistema evaluando la actividad de 1 mes contra la actividad de hasta 3 meses anteriores.
- **Full protection (Costo estimado $30,000 mensuales por plataforma)** - Se da un seguimiento detallado de la plataforma, evaluando tres distintas ventanas de temporalidad histórica (1 minuto, 1 semana y 1 mes) además se integra un sistema de alertas.
- **Premium (Costo estimado $50,000 mensuales por plataforma)** - Se realiza un seguimiento detallado de la plataforma, evaluando distintas ventanas de temporalidad histórica para buscar patrones que se extiende hasta 1 año de registros históricos, además se integra un sistema de alertas.

Equipo técnico y presupuesto considerado

Actualmente en la fase inicial de desarrollo y prototipado se cuenta con acceso a un equipo de cómputo con las siguientes especificaciones:

- Procesador Intel® Xeon® Gold 5317 (12 núcleos)
- Memoria: 128 Gb RAM
- Almacenamiento: 1Tb Solid State Drive

Respecto al personal, se considera una persona cumpliendo el rol de arquitecto tecnológico cuyas responsabilidades serán, la planeación, diseño y desarrollo del cluster de procesamiento así como el desarrollo del algoritmo para el aprendizaje automático.

Una vez esté concluida la fase de prototipado y asegurado el primer cliente, se espera ampliar el equipo de la siguiente manera:

- Un analista de datos (Sueldo $16,000 mensuales)
- Un ingeniero devops (Sueldo $16,000 mensuales)
- Un ingeniero de software (Sueldo $18,000 mensuales)

Plan de trabajo y actividades (para desarrollar el primer prototipo o versión en 2 meses)

Se consideran las siguientes actividades para completar un primer prototipo funcional:

- **Revisión de literatura** (1 semana 28 oct - 1 nov): Se realiza una investigación exhaustiva de las técnicas y procesos estadísticos y de aprendizaje automático así como la naturaleza de los datos de entrada.
- **Creación de cluster de procesamiento** (2 semanas 28 oct - 8 nov): Se diseña y crea una solución tecnológica que permita extraer, procesar y analizar la información, con énfasis en la paralelización y concurrencia de tareas.
- **Desarrollo de algoritmo de clasificación** (4 semanas 4 nov - 29 nov): Se diseña y desarrolla la red neuronal, así como los procesamientos en lote para obtener conocimiento e intuición estadística de los datos así como su comportamiento
- **Documentación** (1 semana 2 dic - 6 dic): Se documenta la tecnologías empleadas así como la propuesta de procesamiento desarrollada para la detección de patrones automática.

Referencias

Sardjono et al, (2021), *The relationship between internet growth and implementation of the internet of things,* Journal of Physics: Conference Ser. 1836 012030 https://iopscience.iop.org/article/10.1088/1742-6596/1836/1/012030/pdf