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”

Alumno:

E23S-18014: MITSIU ALEJANDRO CARREÑO SARABIA

Periodo Enero 2024 - Junio 2024, Aguascalientes, Ags

Resumen.

[Resumen conciso del proyecto que incluya los objetivos principales, el alcance y la importancia del estudio.]

Contenido

Introducción

[Contextualización del problema o la pregunta de investigación que se aborda en el proyecto.]

1. Propuesta Científica

Se propone desarrollar una solución integral de monitoreo de tráfico en servidores web así como la detección automatizada de tráfico anómalo mediante la implementación de técnicas de análisis topológico así como aprendizaje automático las cuales en conjunto permitan por una parte el constante monitoreo de los usos y la toma de decisiones preventivas y correctivas.

Aplicando estas metodologías, 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 tanto acciones correctivas (protección ante uso anómalo) como preventivas (detectar picos o valles de tráfico y ajustar la infraestructura acorde).

1.1 Antecedentes

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 derivado de dicha disponibilidad, se genera una cantidad de tráfico que cualquier servidor web con acceso a internet inmensa la cuál debe dar seguimiento, procesar, contestar.

Dicho tráfico puede ser generado por peticiones de usuarios reales, peticiones de bots, peticiones automatizadas y peticiones de usuarios con intenciones maliciosas por lo que tan solo analizar la cantidad de información generada de manera manual es una tarea imposible.

El tráfico de servidores web tiene claras tendencias como recursos solicitados, región geográfica de donde se solicita, hora en que se solicitó, cantidad de bytes enviados, por lo que identificar las tendencias y detectar las anomalías es un trabajo que puede ser automatizado.

Con el acceso generalizado a servicios en internet, se ha generado una tendencia de permitir alojar datos personales y sensibles en servidores, lo que aumenta la relevancia de evaluar qué y cómo se están accediendo a los recursos solicitados, así como desarrollar herramientas que faciliten filtrar las anomalías para tomar acciones correctivas.

1.2 Objetivos del Proyecto

El objetivo del sistema es desarrollar una solución integral de monitoreo y detección de tráfico anómalo mediante la implementación de técnicas de aprendizaje automático para decidir las acciones preventivas y/o correctivas necesarias.

Para ello es necesario realizar múltiples acciones de soporte como:

- Analizar las técnicas y procesos tanto tradicionales como de aprendizaje automático mediante los cuales se analiza tráfico web actualmente

- Enumerar las características y casos de uso de sistemas de monitoreo y alerta efectivos
- Desarrollar un sistema detección de anomalías basado en aprendizaje automático
- Evaluar el sistema desarrollado implementandolo en un entorno controlado

1.3 Preguntas de investigación

¿De qué manera se analiza el tráfico web actualmente?

¿Qué elementos debe tener un sistema de alertas para ser útil (falsos negativos/falsos positivos, canales de comunicación, protocolos extras)?

¿Actualmente cómo se ha implementado el aprendizaje automático en análisis de tráfico web?

1.4 Justificación

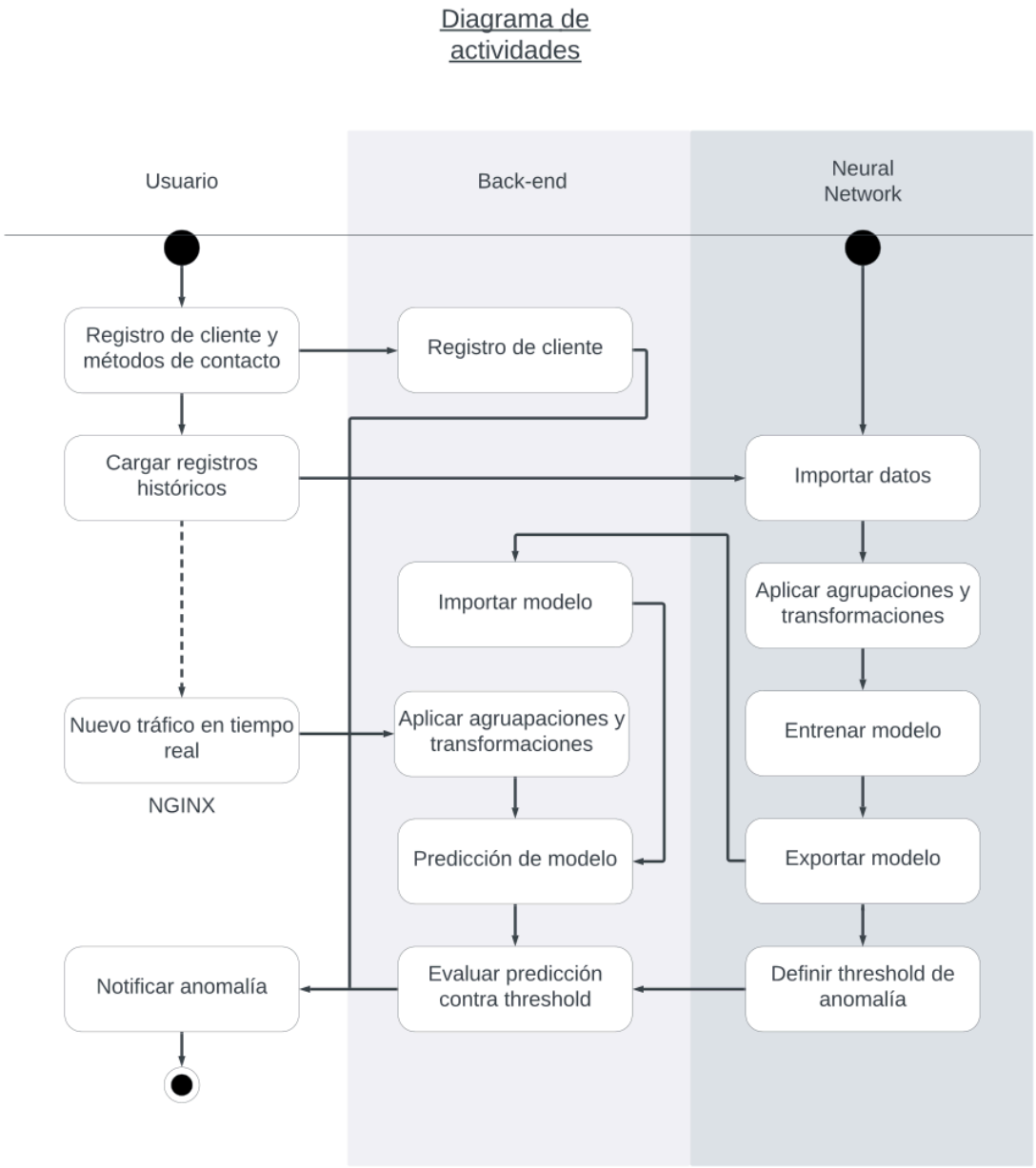
El tráfico a un servidor web provee datos confiables sobre la información y el contexto bajo el que se usan sus recursos, pero la cantidad de información generada es tan grande que un análisis manual no es viable. Entender los usos típicos y diferenciarlos de los atípicos es una herramienta poderosa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

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 e incluso calcular un parámetro de probabilidad de ser malintencionadas. 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.

1.5 Viabilidad

Para la realización del proyecto se cuenta con acceso a fuentes de información necesarias para realizar el análisis y desarrollo correspondiente con datos confiables y en cantidad suficiente para simular condiciones reales.

Respecto a la estructura e infraestructura se realizó el siguiente diagrama en los que se establecen las distintas entidades involucradas en el sistema así como su interrelación, dependencia y relevancia. Se considera que se cuenta con la suficiente experiencia técnica para construir el sistema de manera exitosa.



Por otra parte es necesario realizar una inversión inicial para rentar los equipos de cómputo adecuados para realizar el entrenamiento, transformación y manejo de datos adecuado, por lo que es necesario considerarlo tanto en las características de los equipos necesarios como en su financiamiento.

2. Propuesta Financiera

Para llevar a cabo de manera exitosa el proyecto es necesario categorizar y definir los recursos que permitirán el desarrollo del mismo:

Recursos Humanos:

- Equipo de ocho personas interdisciplinario integrando las áreas de (dirección, desarrollo,, ciencia de datos, finanzas y manejo de clientes)

Recursos Computacionales:

- Amazon Web Services - EC2 instance: Servicio de cómputo dedicado a ejecutar el proyecto en ambiente de producción
- Amazon Web Services - S3 storage: Servicio de almacenamiento empleado para almacenar los datos de entrada (tráfico en tiempo real), así como assets o versiones del modelo entrenado.
- Google Colab Pro: Servicio de computo enfocado en el desarrollo del código de desarrollo de la red neuronal
- Dominio de internet: Cadena de caracteres único que identifica un ámbito de autonomía, autoridad y control, con la finalidad de identificar servicios en internet.

Productos/Servicios personales:

- Coworking: Espacio de oficina compartido por distintas empresas, con diversos servicios como internet, muebles, bebidas, limpieza, etc

2.1 Presupuesto

En la hoja de cálculo adjunta , en la hoja “Costos” se detalla el procedimiento para las cantidades expuestas a continuación.

Recursos humanos: Se requiere un equipo con 8 integrantes, con sueldos entre $15,000 y $25,000 pesos mensuales, total = $353,000.00

Licencia google colab pro: $3,500 al mes, 2 meses, 2 licencias, total = $14,000

EC2 instance (c7a.medium): Costo por hora bajo demanda = $0.852, 2352 horas (14 semanas), total = $2003.90

S3 storage (S3 Standard): Costo por Gb = $0.34, Costo por Gb en transferencia = $0.34, 500 Gb, total = $340

Dominio de internet (loggart.com): Proveedor godaddy, total = $1,199.97 + impuestos

Coworking (alda - private desk): $3,200 mensual, 4 meses, 4 personas total = $25,600.00

Total de infraestructura/servicios= $43,143.87

2.2 Justificación Económica

Para el desarrollo del proyecto se consideran dos aspectos, el tecnológico necesario para desarrollar, desplegar y almacenar toda la información necesaria y el aspecto humano que abarca remuneraciones y espacios de trabajo.

Profundizando en el aspecto tecnológico, durante el desarrollo se requiere poder de cómputo, tanto durante la fase de desarrollo como durante el despliegue de la aplicación, para ello se consideran dos licencias de google colab pro, las cuáles ofrecen acceso a equipos de cómputo potentes así como tarjetas gráficas que permiten acelerar el proceso de desarrollo de la red neuronal, así como los análisis topológicos correspondientes. Por otra parte el resto del desarrollo de código (sistema de alertas) se considera apropiado usar tecnologías de uso libre (github, github workflows, infisical, podman, etc, postgres).

Una vez que el código y la red estén en un avance considerable, se planea usar infraestructura de Amazon Web Services para el manejo de almacenamiento en tiempo real, y cómputo para calcular las predicciones según la red neuronal desarrollada, así como las métricas derivadas de implementar el análisis topológico.

Finalmente respecto al aspecto humano, se considera que el rentar un cowork es una opción viable, ya que se espera que la colaboración entre los integrantes del equipo mejore y agilice tareas de resolución de problemas, brainstorming y organización de tareas a la vez que impacte positivamente en las dinámicas interpersonales, elementos clave para concluir el proyecto en el tiempo estimado, aunque cabe la posibilidad de contratar colaboradores vía remota lo cuál genera un ahorro en la renta del espacio de oficina.

2.3 Fuentes de Financiamiento

Se considera que existen diversas opciones capaces de generar el financiamiento necesario para cubrir los costos de desarrollo y operación del proyecto.

Entrar a concursos es una gran oportunidad para dar visibilidad al proyecto, además de ser una puerta para realizar tareas de networking con gente inmersa en el desarrollo tecnológico o en el desarrollo de proyectos en general, además existe este potencial de ganar y obtener reconocimiento y apoyos económicos al ganar.

Apoyos del gobierno a MIPyMES puede ser una opción viable, en los que además se obtienen asesorías sobre el manejo de negocio además de contactos con gente que conoce la regulación y administración de negocios que si bien no ofrecen una fuente de financiamiento directo, pueden tener un impacto significativo en la toma de decisiones del negocio y son gratuitas.

Una opción que requiere una inversión inicial son las incubadoras, que ofrecen lo mejor de los concursos y de los apoyos del gobierno, ya que permiten realizar mucho networking, se reciben asesorías especializadas y se conoce gente inmersa en el desarrollo de tecnologías emergentes y de riesgo.

Finalmente es posible considerar una inversión inicial propia o de conocidos cercanos que permita tener la liquidez inicial necesaria con una baja tasa de interés.

Independientemente de la fuente de financiamiento, cabe resaltar que el objetivo principal es generar un producto útil y deseable que capture el interés del mercado objetivo, que ofrezca un beneficio real y cuantificable y que sea superior a la competencia, para de esta manera poder generar ingresos.

3. Propuesta de Gestión del Proyecto

[La propuesta de gestión del proyecto establece un marco organizativo y operativo para garantizar la ejecución exitosa y el cumplimiento de los objetivos establecidos.]

3.1 Equipo de Trabajo y estructura organizativa

[Descripción de los miembros del equipo, sus roles y responsabilidades. Detalle de la estructura de gestión del proyecto, incluyendo roles, comunicación y toma de decisiones]

3.2 Plan de trabajo

[Establece de manera detallada las actividades, tareas, recursos y plazos necesarios para alcanzar los objetivos del proyecto de manera eficiente y efectiva.]

3.3 Riesgos y Mitigación

[Identificación de los posibles riesgos del proyecto y estrategias para mitigarlos.]

3.4 Plan de Comunicación

[Recomendaciones para la comunicación y cooperación entre los tres equipos de trabajo del proyecto, incluyendo reuniones regulares, informes de progreso, etc.]

3.5 Ética y Cumplimiento

[Consideraciones éticas y legales relevantes para el proyecto, incluyendo la protección de datos y la conformidad con regulaciones aplicables.]

4. Siguientes Pasos

[Detalle de las actividades que permitirán comenzar la implementación del proyecto, una vez que este ha sido autorizado]

5. Conclusiones

6. Referencias