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.

Se propone y planifica el desarrollo de un sistema de notificación y alerta de tráfico de servidores 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. De esta manera se pretende analizar el volumen total de tráfico sin sacrificar velocidad ni precisión. Finalmente se implementa un sistema de alertas configurable en la que se notifique al cliente la actividad de su servidor, así como la categorización de actividad que asigne el sistema.

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 dicha información es valiosa en el apoyo a toma de decisiones.

Contenido

Introducción

A través del siguiente documento se propone un sistema que permita analizar y detectar anomalías en tiempo real en el uso de servidores web. Integrado con un sistema de alertas se espera que el sistema sea capaz de alertar a responsables de servicios en internet si alguno de sus servicios requiere medidas preventivas o correctivas de rubros como, infraestructura, mejorar seguridad, actividad sospechosa/anómala. Se considera que este tipo de herramientas es una gran oportunidad de integrar tecnologías de aprendizaje automático debido a que el volumen de información generado en la actualidad sobrepasa cualquier intento y capacidad de realizar un análisis manual pero integrarlo a un sistema inteligente y de alerta permite obtener conocimiento real de cómo los usuarios usan e interactúan con los servicios ofertados y al mismo tiempo permite estar mejor preparado para tomar tanto acciones correctivas como preventivas.

A continuación se presenta la solución desde la dimensión científica, financiera y de administración, áreas que se consideran esenciales para completar exitosamente el proyecto. En el apartado científico se ahonda en las características y alcances que se espera tengan el sistema, así como algunos de los requisitos que se deben tomar en cuenta para lograrlas. En la sección de financiera se evalúan los costos estimados para el desarrollo de cada uno de los componentes del sistema, así como una estimación salarial, y se exploran soluciones viables para obtener el financiamiento. Finalmente en la última sección enfocada en la gestión del proyecto, se exploran las capacidades esperadas del equipo, se detallan roles y funciones así como responsabilidades así como un cronograma inicial en el que se estiman los tiempos y el orden en que se espera se ejecuten todas las actividades.

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 ascendente demanda desde la producción en masa de dispositivos celulares y móviles, hasta la progresiva adopción del internet de las cosas [1], pero derivado de dicha disponibilidad, se genera una cantidad inmensa de tráfico que cualquier servidor web disponible desde internet debe dar seguimiento, procesar, contestar [2].

El origen de dicho tráfico puede ser generado por peticiones de usuarios reales, peticiones de bots, peticiones automatizadas y peticiones de usuarios con intenciones maliciosas cada uno de estos grupos de clientes se comportan de maneras muy específicas, por lo que se considera una área de oportunidad el analizar estos datos para reconocer y clasificar los usuarios que interactúan con los servicios ofrecidos [3].

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 y al que se le pueden aplicar distintas técnicas de aprendizaje automático así como análisis topológicos que permitan analizar la información desde múltiples dimensiones y perspectivas [4].

Como consecuencia del el acceso generalizado a servicios en internet, se ha notan tendencias de confiar en el proveedor del servicio y permitirle alojar datos personales y sensibles en sus 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

A continuación se presenta una propuesta de proyecto en la que mediante la integración de técnicas de ciencia de datos así como procesos de ingeniería y desarrollo de software, el objetivo del sistema es proporcionar una solución integral de monitoreo y detección de tráfico anómalo que sea útil en la toma de decisiones tanto para 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:*** Mediante este análisis se espera conocer las herramientas y procedimientos mediante los cuales se realizan análisis similares a los propuestos en el proyecto.
- ***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 sistema 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.
- ***Evaluar el sistema desarrollado implementándolo en un entorno controlado:*** Una vez que se ha completado el desarrollo es necesario realizar una evaluación y validación respecto a todos los subsistemas y procedimientos internos, con ello es posible determinar el grado de madurez del proyecto para buscar posibles clientes.

1.3 Preguntas de investigación

Para la efectiva realización del proyecto se considera pertinente plantearse y contestar las siguientes 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 el análisis de viabilidad, se detectaron las siguientes características que se deben cubrir para la correcta realización del proyecto;

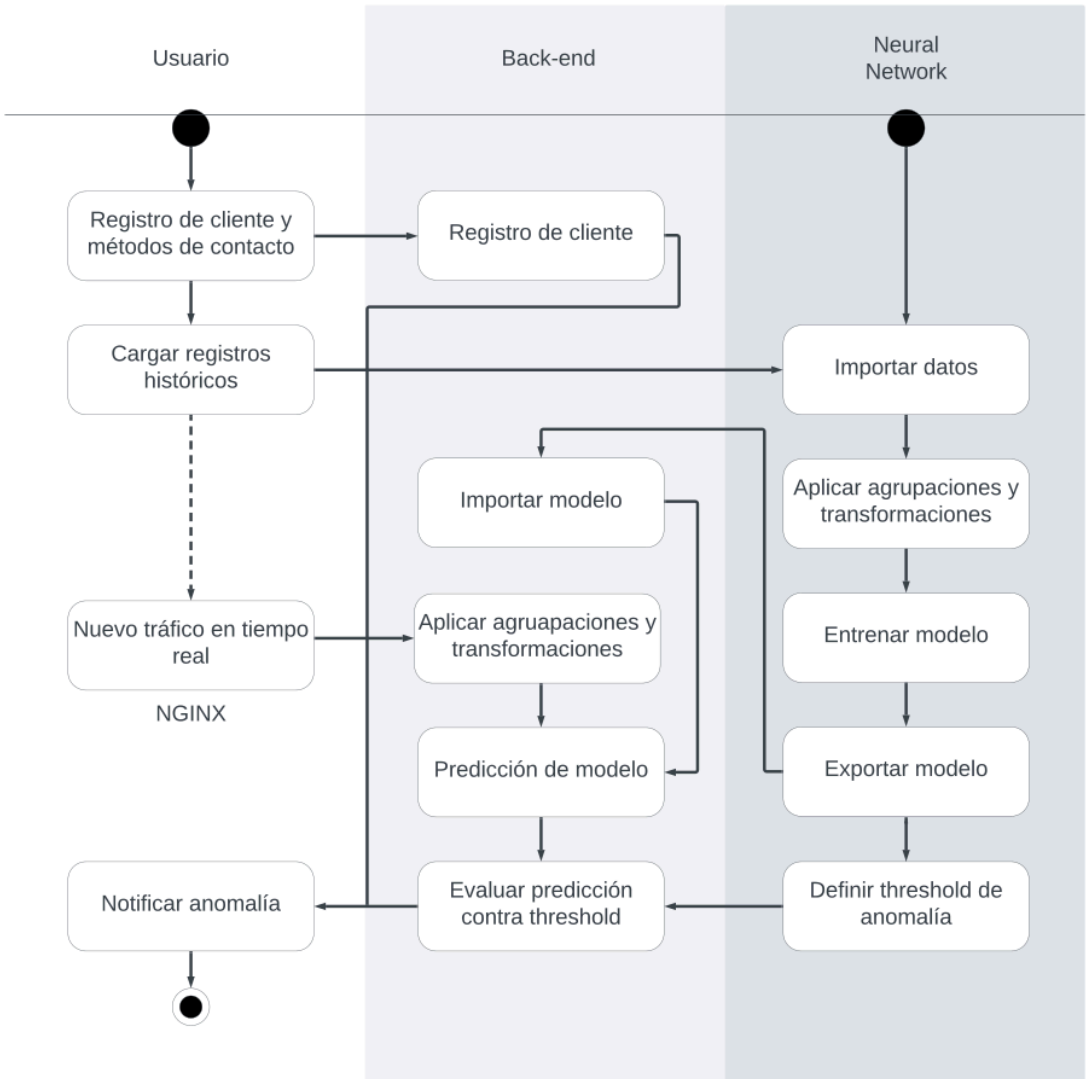
- Fuentes de datos fiables:
- Entendimiento general del proyecto que se quiere realizar
- Entendimiento tecnológico de las plataformas y herramientas necesarias para llevarlo a cabo.

Para la realización del proyecto se cuenta con acceso a fuentes de información necesarias para realizar el análisis y desarrollo correspondiente con datos reales, confiables, obtenidos bajo convenio de privacidad, con consentimiento y en cantidad suficiente para simular condiciones reales se realizó un análisis de la calidad de los datos el cuál se encuentra en la sección de anexos (4).

Respecto a la estructura e infraestructura basado en el siguiente diagrama en los que se establecen las distintas entidades involucradas en el sistema así como su interrelación, dependencia y relevancia.

Siendo los componentes principales:

- Arquitectura de red neuronal: Ingesta de datos de entrenamiento y validación mediante integración a sistema de almacenamiento aws S3 o similar. Entorno de ejecución Python 3 o superior, empleando una instancia de procesamiento aws EC2 (c7a.medium) o rentar una instancia dedicada según se ajuste mejor al presupuesto.
- Backend: Punto de conexión para los clientes y el público en general es responsable del manejo de flujos en tiempo real, la autenticación y administración de usuarios, y realiza predicciones sobre el modelo entrenado, también se encarga de procesar el envío de notificaciones una vez que se detecte actividad anómala.



De manera generalizada se considera que se cuenta con la suficiente experiencia técnica y de liderazgo para construir el sistema de manera exitosa.

Finalmente es necesario realizar un proceso de reclutamiento y selección de colaboradores basado en perfiles bien definidos los cuales se explican en la sección 3.1 de este documento. Se considera crítico elegir a integrantes que tengan experiencia teórica y práctica desarrollando sistemas.

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 el 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

A continuación se detalla el procedimiento mediante el cual se obtuvieron las cantidades expuestas.

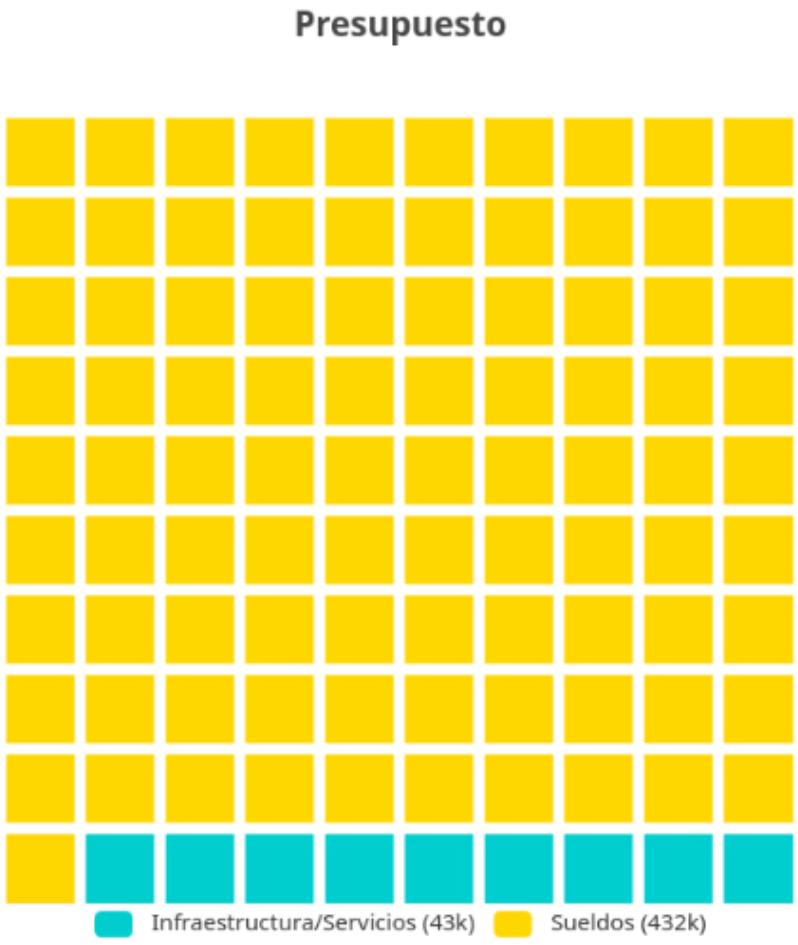
- ***Recursos humanos:*** Se requiere un equipo con 9 integrantes, con sueldos entre $15,000 y $25,000 pesos mensuales, total = $389,000.00
- ***Licencia google colab pro:*** Permitirá realizar el desarrollo y entrenamiento de la red neuronal usando hardware dedicado (unidades de procesamiento gráfico GPU e incluso unidades de procesamiento tensorial TPU) de esta manera se renta el hardware durante el periodo necesario en lugar de invertir en equipo que obsolesce, y requiere condiciones especiales, $3,500 al mes, 2 meses, 2 licencias, total = $14,000
- ***EC2 instance (c7a.medium):*** EC2 es el servicio que ofrece AWS para rentar maquinas virtuales para realizar procesamiento, de esta manera se ahorran costos de adecuación de espacio para un servidor físico así como no se exponen redes domésticas a internet. Costo por hora bajo demanda = $0.852, 2352 horas (14 semanas), total = $2003.90
- ***S3 storage (S3 Standard):*** Es la tecnología de almacenamiento que nos permitirá almacenar los datos completos de entrenamiento de la red, sino también el flujo de datos de clientes reales, ofreciendo un servicio 24/7, tolerante a fallos, autoincrementable. Costo por Gb = $0.34, Costo por Gb en transferencia = $0.34, 500 Gb, total = $340

- ***Dominio de internet (loggart.com):*** Permite a los clientes contactarnos de manera rápida en internet, ofreciendo una identidad del servicio, además usualmente incluyen un registro de email lo que facilita la comunicación dentro y fuera del equipo de trabajo. Proveedor godaddy, total = $1,199.97 + impuestos
- ***Coworking (alda - private desk):*** Corresponde a la renta de un espacio físico de trabajo, amueblado, con servicios de comida, y limpieza, ofreciendo un ambiente colaborativo entre los integrantes del equipo, así como promover la comunicación, socialización y camaradería, factores que se consideran esenciales para el éxito del proyecto. Costo$3,200 mensual, 4 meses, 4 personas total = $25,600.00

Total de infraestructura/servicios= $43,143.87

Inversión total de arranque de proyecto = $432,143.87

A continuación se comparte un gráfico con la distribución presupuestaria:



En la sección de anexos (1) se adjunta una tabla descriptiva tanto de los costos de operación así como los estimados de sueldos y salarios según el puesto y cantidad de empleados.

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.

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 cual genera un ahorro en la renta del espacio de oficina.

Finalmente una vez que el proyecto se encuentre en fase de implementación se estima cobrar un costo mensual de $80,000 por lo que si bien el retorno de inversión está condicionado de la cantidad de usuarios simultáneos, se espera que al menos se recupere en un plazo no mayor a 6 meses después de que el proyecto comenzar a ser operativo.

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

Para la correcta realización del proyecto se deberá contar con un equipo multidisciplinario con principal énfasis en las áreas de ingeniería de software y ciencia de datos, además se consideran algunos roles de coordinación como son líder de desarrollo y responsable de financiamiento. Dichos roles ejecutarán la toma de decisiones en sus áreas correspondientes y ayudarán a los distintos equipos a solucionar problemas brindando guía y experiencia. Finalmente se cuenta con el responsable del proyecto que es quién tiene la visión general del proyecto.

Además se consideran roles auxiliares para tareas de soporte como ingeniero devops o un responsable de manejo de clientes que apoyen en tareas asociadas a sus cargos.

Con un equipo de dichas características se estima que se poseerán las capacidades técnicas, de operación de negocio y de finanzas necesarias para resolver los retos y problemas que se presenten derivado del desarrollo mismo del proyecto.

3.1 Equipo de Trabajo y estructura organizativa

Para la correcta realización del proyecto se considera pertinente integrar un equipo de trabajo que cubra los siguientes roles y necesidades:

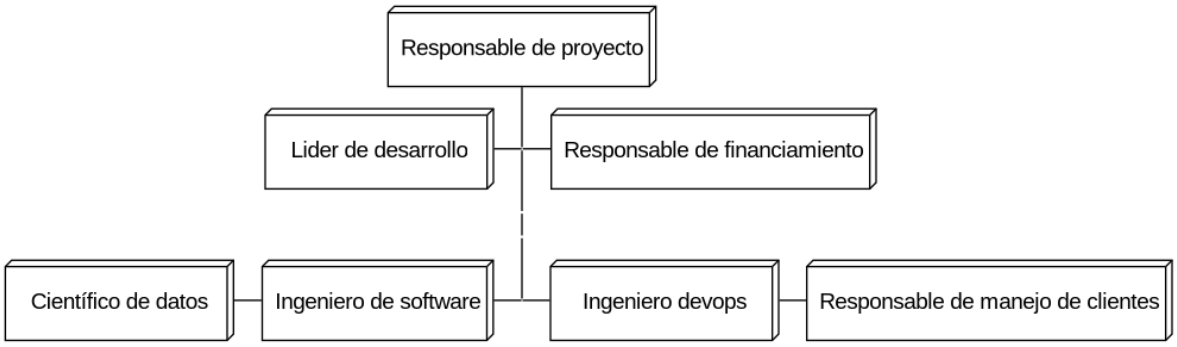
- Responsable de proyecto: Es la figura central que entiende todas las dimensiones del proyecto (técnica, financiera, de negocio, de gestión, etc) y es capaz de tomar decisiones y delegar responsabilidades así como definir y priorizar tareas.
- Responsable de financiamiento: Es la persona que entiende y está inmerso en áreas de manejo financiero, tiene entendimiento legal y de administración de recursos por lo que es capaz de solventar las cuestiones financieras del proyecto.
- Lider de desarrollo: Es el rol a cargo de gestionar el desarrollo general del proyecto por lo que debe tener una base fuerte en desarrollo de tecnología y ciencia de datos en sus diferentes áreas (tecnologías, integraciones, servicios, plataformas de desarrollo y despliegue etc). Es la figura responsable de que el proyecto desde el enfoque tecnológico se lleve a cabo de manera exitosa.
- Científico de datos: Es el cargo responsable de concretar las tareas y actividades concernientes a la ciencia de datos con acompañamiento y dirección del líder de desarrollo, se espera que tenga conocimiento fundamental sobre redes neuronales así como conocimiento de buenas prácticas de desarrollo y capacidad de trabajo en equipo.
- Ingeniero de software: Es el cargo responsable de generar el código para las distintas integraciones necesarias (alertas, procesamiento de información en tiempo real, manejar almacenamiento de información, etc), por lo que debe tener fuertes bases de conocimiento en sistemas y flujos de información, así como poder aplicar e integrar tecnologías existentes.
- Ingeniero devops: Rol con la capacidad de integrar los distintos sistemas generados (redes neuronales, despliegues automatizados, versionado automático, bitácoras de registros para los distintos sistemas, replicabilidad de ambientes) así como ser capaz de generar y

replicar ambientes de desarrollo, pruebas y producción así como soluciones de tolerancia a fallos

- Responsable de manejo de clientes: Es el rol responsable de manejar y mantener el contacto con los clientes, permitiendo la transparentización de los procesos tanto internos como externos y dando seguimiento al avance del equipo de trabajo.

Además se considera esencial que todos los perfiles cuenten con aptitudes sociales, comunicativas y de trabajo en equipo para que las los problemas y dudas se puedan escalar y clarificar de manera adecuada, además se espera que los distintos responsables y líderes sean capaces de instruir a sus equipos para que logren desarrollar sus habilidades más allá de sus tareas rutinarias.

A continuación se muestra el organigrama propuesto para el equipo de trabajo:



3.2 Plan de trabajo

Para el desarrollo del proyecto se dividirán las actividades en dos etapas, desarrollo e implementación, donde durante el desarrollo se realizan pruebas de concepto de las tecnologías así como de las ideas, mientras que en la fase de implementación se considera emplear las tecnologías y soluciones creadas y evaluadas durante la fase de implementación aplicado a clientes con datos reales, a continuación se listan las actividades que se consideran necesarias realizar en cada una de estas etapas.

3.2.1 Desarrollo:

- ***Definir requisitos y alcances del desarrollo (Responsable de proyecto [2 semanas]):*** Se espera definir los alcances del proyecto en general y su nivel de atracción y usabilidad en el mercado.
- ***Realizar cotizaciones para fase de entrenamiento (Responsable de financiamiento [3 semanas]):*** A la par de definir los requisitos y alcances, se recomienda tener el acompañamiento de responsable de financiamiento para definir objetivos alcanzables así como tener un idea general del nivel de inversión necesaria para el desarrollo del proyecto.
- ***Fase de preprocesamiento (Líder desarrollo [3 semanas]):*** Una vez concluida la planeación, es posible comenzar con la adaptación de los datos de prueba, los cuales se comienzan a reunir, limpiar, filtrar y extender, recordando que los datos son puramente textuales, por lo que es necesario clasificarlos y transformarlos para su posterior uso [5].

- ***Análisis topológicos (Científico de datos [2 semanas]):*** Hacia el final del preprocesamiento es posible comenzar con los diversos análisis topológicos mediante los cuales se obtiene un mejor entendimiento de los datos así como de su comportamiento.
- ***Desarrollo de red neuronal (Científico de datos / Líder desarr. [5 semanas]):*** En esta actividad se espera que basado en las perspectivas obtenidas de pre-procesar y analizar topológicamente los datos se pueda determinar la mejor tecnología para detectar anomalías en los datos, entre las que puede ser redes neuronales, autoencoders, o técnicas de aprendizaje profundo, etc.
- ***Validación de predicciones y gráficos (Científico de datos [3 semanas]):*** Hacia el final del desarrollo de la red neuronal es posible comenzar con las evaluaciones de sus predicciones, y comenzar a ingresarle datos más extensos y desconocidos para evaluar su comportamiento. A la vez es posible evaluar los gráficos e instrumentos derivados del análisis topológico y evaluarlos.
- ***Validar entrada, tratamiento y predicciones (Líder desarrollo [1 semana]):*** FInalmente con la red completada y valorada, es posible comenzar a usar el sistema en ambiente controlados con datos específicos pero desconocidos para la red, estandarizar cuales son los estándares de entrada de datos y salidas producidas.
- ***Desarrollar conexión con stream en tiempo real (Ingeniero de software [4 semanas]):*** Esta actividad tiene poca relación con las mencionadas anteriormente por lo que se puede llevar a cabo en paralelo al desarrollo de la red neuronal, en ella se evalúan y prueban distintas tecnologías que permitan conexiones en tiempo real que provean los datos a la red neuronal. Al finalizar esta actividad se debe tener una serie de tecnologías y protocolos definidos que ingresen la información de tráfico al sistema.
- ***Evaluar tecnologías para sistema de alertas (Líder desarrollo [2 semanas]):*** En esta actividad se deben analizar los beneficios, limitaciones y costos de distintas tecnologías que permitan enviar notificaciones (push, email, in app, sms) a dispositivos móviles.
- ***Desarrollar integración de sistema de alertas (Ingeniero de software [4 semanas]):*** Derivado del análisis de la actividad anterior se espera que se realice el código, configuraciones y adquisiciones necesarias para tener un sistema de alertas robusto, configurable y que se adapte a las necesidades específicas del proyecto.
- ***Despliegue de prueba de concepto (Ingeniero devops / Líder desarr. [2 semanas]):*** Para esta actividad se espera que la red neuronal, el análisis topológico y los sistemas de ingesta en tiempo real y de alertas estén finalizados y se integren en un solo sistema
- ***Validación general del sistema (Lider de desarrollo [1 semana]):*** Una vez completados todas las actividades anteriores, es posible realizar pruebas generales del sistema y realizar distintas evaluaciones sobre el (performance, responsividad, puntos de mejora, puntos de venta, etc) pero sobre todo se espera contar con una opinión completa del sistema con todas sus partes.

3.2.2 Implementación:

- ***Solicitar insumos a cliente (Responsable de cliente [4 semanas]):*** Para comenzar la implementación del sistema con clientes y datos reales se deben solicitar accesos para

integrar el stream en tiempo real, así como solicitar medios de contacto preferidos para las alertas y cualquier acceso necesario de su infraestructura para integrar el sistema.

- ***Preparar sistema de almacenamiento (Ingeniero devops / Líder desarr. [3 semanas]):*** Durante esta actividad se deben configurar distintos sistemas de almacenamiento que permitan al sistema realizar análisis agregados sobre periodos de tiempo definidos, (tráfico de las últimas 24 hrs) así como ofrecer robustez al sistema y tolerancia a fallos y finalmente alimentar nuevos modelos de aprendizaje para detectar con mejor precisión el tráfico anómalo.
- ***Realizar entrenamiento específico (Científico de datos [3 semanas]):*** Con datos reales del cliente se espera realizar un entrenamiento de la red específico y personalizado, lo que permite a la red aprender, cual es el comportamiento real del cliente.
- ***Desplegar implementación de sistema de alerta (Ingeniero devops [3 semanas]):*** Finalmente se realizan todos los chequeos de los sistemas internos así como la comunicación tanto de entrada (stream en tiempo real) como de salida (sistema de alertas)

En la sección de anexos (2) se comparte el cronograma completo de manera gráfica por lo que es más sencillo identificar las tareas paralelas así como el orden de las mismas

3.3 Riesgos y Mitigación

Todos los proyectos tienen ciertos riesgos y este no es la excepción, tras un análisis del concepto del sistema a realizar se detectaron los siguientes riesgos los cuales se categorizaron por su probabilidad, gravedad y tipo de riesgo.

3.3.1 Riesgo estratégico:

- ***El cliente no quiere compartir información / accesos [Probabilidad.: No es probable / Gravedad: Catastrófico]***: La información sobre la que se espera que opere el sistema propuesto es sensible, ya que puede exponer tanto especificidades del comportamiento del cliente, como detalles técnicos de la infraestructura y software que usa el cliente, por lo que es entendible si existe resistencia a compartir información, sin embargo se espera que el sistema ofrezca más beneficios al poner en uso dicha información, que de manera realista, pocas veces se analiza.
- ***El cliente usa tecnologías no compatibles [Probabilidad: Posible / Gravedad: Catastrófico]:*** Si bien la propuesta intenta mitigar este riesgo al es ser compatible con las tecnologías más comunes y basado principalmente en simples cadenas de texto, en el aspecto la tecnología constantemente están surgiendo nuevas soluciones por lo que no se descarta que un cliente tenga tecnologías/infraestructura que el sistema propuesto no pueda procesar.
- ***Exponer información del cliente de manera accidental [Probabilidad: Posible / Gravedad: Catastrófico]:*** Al concentrar información sensible de clientes se entiende que la infraestructura y código del proyecto sea un posible objetivo para actores maliciosos, por lo que es inevitable considerar el riesgo de ser víctimas. Para ello es necesario contar con altos estándares de calidad en las prácticas de programación y en el manejo de información que por naturaleza es sensible.

3.3.2 Riesgo operativo:

- ***El desarrollo del sistema rebasa las estimaciones de tiempo [Probabilidad: Probable / Gravedad: Importante]:*** En la industria una de las tareas más complejas es la correcta estimación de tiempo para las tareas, esto se debe a los "desconocidos desconocidos", no conocemos las aptitudes de los integrantes del equipo, ni las tecnologías que mejor se adapten a las necesidades del proyecto, por lo que siempre se intenta dejar una holgura para cubrir estas fases de descubrimiento. Sin embargo la mejor manera de mitigar estos desfases es ser muy cuidadoso a la hora de planear, intentar encontrar los posibles puntos ciegos y asegurar que los responsables de los distintos equipos prevean los errores, o inconsistencias para tomar acciones correctivas.
- ***El almacenamiento es más costoso de lo estimado [Probabilidad: Posible / Gravedad: Importante]:*** Dado el contexto del proyecto, se estima que el principal costo a cubrir del proyecto es el costo de almacenamiento en la nube, el cuál se cobra basado en la cantidad de escrituras y lecturas. Sin embargo con el auge del Software as a Service y Platform as a service, diversas soluciones están disponibles y operan bajo principios similares por lo que una migración de proveedor de servicio es posible.
- ***El sistema en tiempo real tenga latencia [Probabilidad: Probable / Gravedad: Moderada]:*** En sistemas en tiempo real es importante considerar la latencia, más aún en un sistema de monitoreo, sin embargo, un buen modelo de detección puede ser capaz de detectar y alertar de manera previa, por lo que si bien es un riesgo, se puede atender mejorando la infraestructura, los modelos y el código.

3.3.3 Riesgo financiero:

- ***Los costos de operación rebasan el financiamiento [Probabilidad: Posible / Gravedad: Moderada]:*** Los costos de operación son difíciles de estimar, debido a que existen partes del sistema que aun no estan concretas por lo que se desconoce su consumo computacional, de ancho de banda, de almacenamiento, pero como se mencionó en puntos anteriores, siempre existen alternativas, por lo que se puede buscar un nuevo proveedor de servicio, una tecnología que reduzca cierta característica o optimizar procesos dentro del código generado, por lo que es un riesgo manejable y común en el ramo tecnológico.

3.3.4 Riesgo técnico:

- ***El sistema es lento [Probabilidad: No es probable / Gravedad: Importante]:*** Como se mencionó anteriormente, el sistema debe operar dentro de parámetros aceptables de responsividad, un sistema de alertas lento va contra el principio al que intenta servir, para mitigarlo además de los puntos de mejora existentes (infraestructura, modelos, código) también se cuentan con espacios en el cronograma para evaluar y validar que el sistema responde y actúa en los parámetros esperados, permitiendo realizar acciones correctivas tan pronto como sean detectadas.
- ***El equipo no tiene los conocimientos/expertise [Probabilidad: Posible / Gravedad: Moderada]:*** Si bien todos tenemos la capacidad de aprender y adaptarnos, los tiempos de

desarrollo del proyecto parten de ciertas capacidades esperadas en el personal, por lo que la mitigación está en el proceso de selección y entre mejor se tenga conceptualizado el proyecto (primera actividad del cronograma) más concretos serán los requisitos esperados de los integrantes del equipo.

3.3.4 Riesgo externo

- ***El cliente no sabe cómo reaccionar a alertas del sistema [Probabilidad: Muy probable / Gravedad: Menor]:*** Es posible e incluso probable que el cliente carezca del personal y/o experiencia para responder a las alertas generadas por el sistema, para ello es posible realizar cursos de capacitación con el departamento de sistemas/devops en las que se aborden situaciones y resoluciones comunes.
- ***Las alertas del sistema son ignoradas por los clientes [Probabilidad: Probable / Gravedad: Menor]:*** Es necesario trabajar desde el ámbito comercial y mercadológico para explicar desde un inicio, cuál es la relevancia y usos del sistema, así como realizar capacitaciones mencionadas en el punto anterior.

De manera generalizada, los riesgos se encuentran situados en un equilibrio entre riesgo y probabilidad, lo cuál nos demuestra que el proyecto es lo suficientemente innovador como para atender una problemática real y poco explorada en la actualidad, y al mismo tiempo es realizable con los recursos y tiempo adecuado. Además, integrando una evaluación periódica de riesgos cada vez que se alcance un hito, se puede mantener actualizadas las estratégias de mitigación en caso de requerirse.

En la sección de anexos (3) se muestra la representación gráfica de la matriz de riesgos.

3.4 Plan de Comunicación

La comunicación se considera un factor principal para lograr los objetivos y para mejorar la capacidad de cada uno de los integrantes del equipo por lo que se consideran una serie de acciones, tecnologías y protocolos para mejorarla. En primer lugar además de los medios tradicionales de comunicación moderna (email, aplicaciones de mensajería, videollamadas, etc), se considera en el presupuesto un espacio de coworking que permita a los integrantes reunirse, conocerse y convivir como estímulo para la comunicación, si bien no se rechazan estratégias de trabajo remoto, el postulante debe ser sobresaliente en habilidades comunicativas, para considerarlo una opción viable.

Además se considera apropiado implementar **Jira** o **Github issues** para dar seguimiento a las tareas, pendientes, nuevas funcionalidades etc ya que ambas nos permiten:

- Visibilidad del progreso y pendientes
- Asignar responsables
- Dar seguimiento de la tarea a nivel código
- Dar feedback o solicitar cambio de manera persistente, abierta y directa
- Realizar evaluaciones de código de manera jerárquica en el área

Además se cree oportuno generar reuniones re arranque de día muy rápidas (15-20 min) y segmentadas por área en las que se aborden los problemas persistentes que se topa el equipo, mediante esta junta, se puede escalar los problemas con los responsables de área y abre el espacio a un análisis detallado del problema y una solución concreta.

Se considera desarrollar una junta con todo el equipo cada vez que se cumpla un hito (completar el preprocesamiento, entrenar la red, completar el sistema de alertas, etc) para mantener una visión clara del progreso, así como realizar ajustes al cronograma.

Finalmente durante las fases de implementación se considera realizar una reunión semanal entre los integrantes del equipo involucrados y el cliente para ofrecer actualizaciones sobre el avance, coordinar esfuerzos y solicitar insumos.

3.5 Ética y Cumplimiento

Para garantizar el correcto uso de la información así como buenas prácticas de consentimiento y transparencia es necesario realizar varias acciones comenzando por los integrantes del equipo mismo, se debe elaborar un acuerdo de confidencialidad y no divulgación, acompañado de un taller de capacitación sobre el correcto uso de la información recibida. Además es necesario evaluar los términos de servicio con los proveedores de servicio (procesamiento, almacenamiento, etc) asegurando que los datos que manejan y alojen no sean divulgados ni empleados en otras tareas.

Respecto al cliente, es necesario elaborar un acuerdo en el que el cliente exprese su consentimiento para compartir información del tráfico de sus servidores y establecer como su responsabilidad notificar a sus usuarios que la información que generen en sus plataformas será compartida para su evaluación. Además el acuerdo con el cliente debe expresar puntualmente qué información y cuánto tiempo se guardará en nuestro sistema, si se podrá usar o no para futuros re-entrenamientos o posibles usos de otra naturaleza.

Por otra parte se debe establecer si los usuarios que generan la información (que acceden a los servidores monitorizados) pueden o no tener derecho a rechazar este tratamiento y queda en responsabilidad del cliente únicamente compartir la información de clientes que hayan aceptado en caso de ser opcional.

Sin embargo se debe ahondar en las regulaciones mexicanas respecto al uso y tratamiento de datos personales para seguir la normativa vigente y establecer los mecanismos que la ley mande.

Finalmente es necesario establecer políticas y protocolos de seguridad internos que aseguren la integridad, confidencialidad y disponibilidad de los datos al mismo tiempo que se protege ante accesos no autorizados o uso indebido de la información recopilada y almacenada.

4. Siguientes Pasos

Una vez que las fases de desarrollo e implementación han sido completadas cabe resaltar que para seguir siendo un servicio atractivo, se deben mejorar las capacidades logradas, así como entender

las necesidades de los clientes para realizar nuevas propuestas de valor, mejorar la precisión de la detección de anomalías, detectar mejor distintos tipos de ataques cibernéticos, ofrecer alternativas en el sistema de alertas, realizar entrenamientos que permitan evaluar rangos mayores de tiempo, incluso refinar estratégias de marketing, manejo de clientes solo por nombrar algunas, además se pueden explorar maneras de abaratar los costos de operación, o mejorar el performance logrado, en fin al tratarse de un servicio ofertado, siempre se deben buscar maneras de mantenerse en el mercado.

5. Conclusiones

Los sistemas de aprendizaje por computadora han llegado a cambiar muchos aspectos y flujos de la vida moderna, hay tareas en que las computadoras en general son increíblemente eficientes, una de ellas es la capacidad de procesar información en grandes volúmenes de manera tan rápida que se puede realizar en flujos en tiempo real, esta capacidad es perfecta para integrarse en un sistema de monitoreo, a pesar de ello, poco se ha explorado en las áreas de aprendizaje por computadora y ciberseguridad. Es por ello que se considera que este proyecto es pionero en un campo tan fértil y proporciona una ventaja real sobre otros sistemas de monitoreo y alerta, ofreciendo análisis de tráfico personalizado, con la capacidad de estar en constante re-aprendizaje y que ofrezca alertas reales y útiles basadas en matemática, ingeniería y aprendizaje automático.

6. Referencias

1. Abid, A., Manzoor, M. F., Farooq, M. S., Farooq, U., & Hussain, M. (2020). Challenges and Issues of Resource Allocation Techniques in Cloud Computing. *KSII Transactions on Internet & Information Systems*, *14*(7). https://itiis.org/journals/tiis/digital-library/manuscript/file/23716/TIIS%20Vol%2014,%20No%207-5.pdf
2. Chung Yung, Chia-Ching Chen, Yu-Lan Yuan, Ching Li "A Systematic Model of Big Data Analytics for Clustering Browsing Records into Sessions Based on Web Log Data" in Journal of Computers doi: 10.17706/jcp.14.2.125-133 http://www.jcomputers.us/vol14/jcp1402-06.pdf
3. Naidu, K. B., Prasad, B. R., Hassen, S. M., Kaur, C., Al Ansari, M. S., Vinod, R., ... & Bala, B. K. (2022). Analysis of Hadoop log file in an environment for dynamic detection of threats using machine learning. *Measurement: Sensors*, *24*, 100545. https://www.sciencedirect.com/science/article/pii/S2665917422001799
4. Ashwini ladekar, Pooja Pawar, Dhanashree Raikar, Jayashree Chaudhari "Web Log based Analysis of User's Browsing Behavior" in International Journal of Computer Applications (0975 – 8887)Volume 115 – No. 11, April 2015 https://research.ijcaonline.org/volume115/number11/pxc3902430.pdf
5. Łukasz Korzeniowski; Krzysztof Goczy "Landscape of Automated Log Analysis: A Systematic Literature Review and Mapping Study," in IEEE Access, vol. 10, pp. 21892-21913, 2022, doi: 10.1109/ACCESS.2022.3152549. https://ieeexplore.ieee.org/document/9716129

7. Anexos

1. Desglose de gastos operativos y salarios:

Recursos Humanos					
Área	**Responsable**	**Sueldo mensual**	**Miembros de equipo**	**Meses requeridos**	**Sueldo total**
Dirección	Responsable de proyecto	25,000	1	4	100,000.00
Finanzas	Responsable de financiamiento	15,000	1	1	15,000.00
Desarrollo	Lider desarrollo	25,000	1	4	100,000.00
Ciencia de datos	Científico de datos	15,000	2	2	60,000.00
Desarrollo	Ingeniero de software	15,000	2	2	60,000.00
Desarrollo	Ingeniero devops	18,000	1	2	36,000.00
Manejo Clientes	Responsable de cliente	18,000	1	1	18,000.00
Total					389,000.00
Infraestructura y servicios					
Servicio	**Unidad**	**Costo por unidad**	**Unidades requeridos**	**Notas**	**Costo total**
Google colab pro	Meses	3,500	4	2 licencias, 2 meses	14,000.00
EC2 instance (c7a.medium)	Horas bajo demanda	0.852	2,352	14 semanas	2,003.90
S3 storage (S3 Standard)	Gb almacenamiento	0.340	500		170.00
S3 storage (S3 Standard)	Gb transferencia	0.340	500		170.00
Dominio de internet		1,199.970	1		1,199.97
Coworking	Meses	3,200.000	8	4 meses, 4 personas	25,600.00
Total					43,143.87

2. Cronograma de Gantt completo:

			Semanas															
Área	**Responsable**	**Actividad**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Desarrollo	Responsable de proyecto	Definir requisitos y alcances del desarrollo	X	X														
Desarrollo	Responsable de financiamiento	Realizar cotizaciones para fase de entrenamiento	X	X	X													
Desarrollo	Lider desarrollo	Fase de preprocesamiento			X	X	X											
Desarrollo	Científico de datos	Análisis topológicos					X	X										
Desarrollo	Científico de datos / Líder desarr.	Desarrollo de red neuronal					X	X	X	X	X							
Desarrollo	Científico de datos	Validación de predicciones y gráfos								X	X	X						
Desarrollo	Lider desarrollo	Validar entrada, tratamiento y predicciones										X						
Desarrollo	Ingeniero de software	Desarrollar conexión con stream en tiempo real			X	X	X	X										
Desarrollo	Lider desarrollo	Evaluar tecnologías para sistema de alertas			X	X												
Desarrollo	Ingeniero de software	Desarrollar integración de sistema de alertas					X	X	X	X								
Desarrollo	Ingeniero devops / Líder desarr.	Despliegue de prueba de concepto											X	X				
Desarrollo	Lider de desarrollo	Validación general del sistema												X				
Implementación	Responsable de cliente	Solicitar insumos a cliente												X	X	X	X	
Implementación	Ingeniero devops / Líder desarr.	Preparar sistema de almacenamiento												X	X	X		
Implementación	Científico de datos	Realizar entrenamiento específico														X	X	X
Implementación	Ingeniero devops	Desplegar implementación de sistema de alerta														X	X	X

3. Matriz de riesgos

<u>Categorías:</u>

a- Riesgo estratégico

b- Riesgo operativo

c- Riesgo financiero

d- Riesgo técnico

e- Riesgo externo

	Gravedad				
	1 - Insignificante	2 - Menor	3 - Moderada	4 - Importante	5 - Catastrófica
5 - Muy Probable	5	10	15	20	25
		e- El cliente no sabe cómo reaccionar a alertas del sistema			
4 - Probable	4	8	12	16	20
		e- Las alertas del sistema son ignoradas por los clientes	b- El sistema tiempo real tenga latencia	b- El desarrollo del sistema rebasa las estimaciones de tiempo	
3 - Posible	3	6	9	12	15
			c- Los costos de operación rebasan el financiamiento d- El equipo no tiene los conocimientos/ expertise	b- El almacenamiento es costoso	a- El cliente usa tecnologías no compatibles a- Exponer información de clientes de manera accidental
2 - No es probable	2	4	6	8	10
				d- El sistema es lento	a- El cliente no quiere compartir información/accesos
1 - Muy improbable	1	2	3	4	5

4. Evaluación de calidad de datos

Criterio	Nivel 1 Deficiente	Nivel 2 Aceptable	Nivel 3 Bueno	Nivel 4 Excelente	Evaluación	Comentarios
Relevancia	Los datos no son pertinentes para los objetivos del proyecto	Parte de los datos es relevante, pero no toda la información necesaria está presente	La mayoría de los datos es relevante y útil para los objetivos del proyecto	Todos los datos son completamente pertinentes y adecuados para los objetivos del proyecto	4 - Excelente	Los datos ofrecen datos únicos que expresan el contexto a partir del cual se generó la conexión y cómo la manejó el servidor
Exactitud	Los datos contienen errores significativos	Algunos datos son precisos, pero hay errores ocasionales	Los datos son mayoritariamente precisos, con pocos errores	Los datos son completamente precisos y reflejan la realidad sin errores.	4- Excelente	Los datos representan fielmente los datos que recibe el servidor, y se entienden las limitaciones al existir mecanismos y tecnologías de ofuscación
Completitud	Los datos están incompletos, faltan muchas piezas clave	Algunos datos faltan, pero no impiden significativamente el análisis	La mayoría de los datos está completa, con pocas piezas faltantes	Todos los datos están completos y abarcan todas las dimensiones necesarias	4- Excelente	Los datos al ser generados de manera automática como bitácora de un servidor web, ofrecen toda la información relevante.
Consistencia	Los datos presentan numerosas inconsistencias y contradicciones	Hay algunas inconsistencias, pero la mayoría de los datos son coherentes	Los datos son mayoritariamente consistentes, con mínimas discrepancias	Los datos son completamente coherentes y uniformes en todos los aspectos	4- Excelente	Los datos tienen una granularidad alta, lo que significa que un solo renglón representa una conexión completa, además de seguir el modelo RESTful

Estructura	La estructura que presentan los datos ha cambiado	La estructura de los datos ha cambiado parcialmente	La mayoría de la estructura de los datos se preserva y es pertinente para el estudio	La estructura de todos los datos es completamente relevante y sigue en uso	4- Excelente	La estructura de los datos a pesar de ser de hace un año sigue exactamente igual por lo que tiene valor para el estudio
Accesibilidad	Los datos no están disponibles o son muy difíciles de acceder	Los datos están disponibles, pero el acceso es complicado y lento	Los datos son accesibles con algunos obstáculos menores	Los datos son fácilmente accesibles y disponibles cuando se necesitan	3- Bueno	Los datos son difíciles de acceder por diseño, al formar parte de los datos sensibles de la empresa, además tienen una duración limitada de 15 días (configurable) para controlar la demanda de almacenamiento
Comprensibilidad	La documentación de los datos es confusa o inexistente	La documentación está presente pero es incompleta o difícil de entender	La mayoría de la documentación es clara y comprensible	Toda la documentación es clara, completa y fácil de entender	2- Aceptable	La documentación indica las posibles configuraciones a aplicar a los datos (duración, campos, formato, ubicación en disco) pero no ofrece una explicación directa del dato por lo que se necesita experiencia en el campo y sus tecnicismos
Fiabilidad	Las fuentes de datos no son confiables	Algunas fuentes de datos son confiables, pero otras no lo son	La mayoría de las fuentes de datos son confiables	Todas las fuentes de datos son altamente confiables y verificables	4- Excelente	Los datos provienen de una única fuente, que es el servidor web mismo, y de una única aplicación, el proxy que se

						encarga de manejar el tráfico, por lo que son altamente confiables
Seguridad	Los datos no están protegidos adecuadamente y pueden ser vulnerables	La protección de datos es consistente, con algunas áreas vulnerables	Los datos están mayoritariamente protegidos, con pocas vulnerabilidades	Los datos están completamente protegidos y cumplen con todas las normativas de seguridad y privacidad	4-Excelente	Los datos están únicamente alojados en el servidor en cuestión, protegidos con diversas medidas de seguridad entre las que destacan, conexión encriptada y acceso únicamente mediante contraseña
Rango temporal	Los datos son antiguos, de un periodo temporal corto y el comportamiento que describen pudo haber cambiado	Los datos son moderadamente recientes, de un periodo temporal corto y parte del comportamiento que describen aún se mantiene	Los datos son recientes o comprenden un rango temporal hasta el pasado reciente y generalmente describen comportamientos recientes	Los datos son completamente recientes y comprenden un rango de tiempo extenso que permite analizar tendencias nuevas y pasadas	2-Aceptable	Los datos comprenden un periodo de 15 días del año pasado, por lo que permiten realizar un análisis de comportamiento generalizado, pero es posible que las tendencias hayan cambiado moderadamente