UAB
Universitat Autònoma de Barcelona

Automatización del Análisis de Seguridad Corporativa mediante Técnicas de Escaneo Combinado y Agentes de IA RAG

Trabajo de Fin de Grado
Grado en Empresa y Tecnología
Curso 2024-25

Autor: Alessandro Benigni
Tutor: Miguel Ángel de Cara Ruiz
Fecha de entrega: 28/05/2025

Resumen

La inteligencia artificial ha tenido un enorme impacto sobre nuestras vidas durante los últimos años, pasando de ser un concepto ciencia ficción hasta tener un rol central en muchos sectores de la sociedad, entre los cuales el de la ciberseguridad. Si, por un lado, nacen herramienta de detección y antivirus más sofisticados y potenciados por la IA, como es el caso de antivirus y EDR, por otro lado, se desarrollan nuevos vectores de ataques generados y automatizados por la inteligencia artificial, dificultando la adaptación y la protección de las empresas frente a esta realidad en cuanto se amplía la superficie de ataque. Además, según el Informe de Amenazas a los Datos de Thales 2025, casi el 70 % de las organizaciones identifican al rápido ecosistema de la IA como el principal riesgo de seguridad que está relacionado con la GenAI. Dicho en otras palabras, las empresas eligen adoptar soluciones IA lo más rápido posible, dejando de lado la segurización completa de sus activos. (Thales, 2025)[15] La adopción de estos sistemas implica desafíos de segurización de las informaciones, ya que sobre todo las IA generativa y los sistemas agénticos necesitan trabajar con datos de calidad. Por tanto, no son conscientes de toda la información expuesta en internet, ya que, también disminuyen sus inversiones en pruebas de seguridad para poder invertir en la automatización de sus procesos.

Por estas razones, en este trabajo, se presenta una plataforma integrada de análisis de seguridad que combina OSINT, escaneo web y de red, potenciada por inteligencia artificial mediante arquitectura RAG. El sistema automatiza la detección de vulnerabilidades usando herramientas automatizadas tanto propias como de terceros, y mejora la interpretación de resultados mediante recuperación de conocimiento contextualizado. La implementación de la arquitectura Rag, representa el punto principal del proyecto, ya que mejora significativamente el análisis e interpretación de los resultados obtenidos por las distintas herramientas. Este enfoque reduce considerablemente las “alucinaciones” o imprecisiones, típicas de los modelos de lenguaje LLM, al anclar el análisis en datos verificables y contextualizados. El sistema recupera información relevante de una base de datos y genera análisis precisos contextualizados.

La herramienta proporciona, por tanto, a las empresas una amplia visión de su situación de seguridad y mediante los informes sugiere un plan de acción basado en los hallazgos encontrados.

Palabras clave

OSINT, Ciberseguridad, Inteligencia Artificial, RAG, Análisis Vulnerabilidades, Automatización Seguridad, Generación Reportes.

Abstract

Artificial intelligence has had an enormous impact on our lives during recent years, going from being a science fiction concept to having a central role in many sectors of society, among which cybersecurity. If, on one hand, more sophisticated and AI-powered detection tools and antivirus are born, as is the case of antivirus and EDR, on the other hand, new attack vectors generated and automated by artificial intelligence are developed, making it difficult for companies to adapt and protect themselves against this reality as the attack surface expands. Furthermore, according to the Thales 2025 Data Threat Report, almost 70 % of organizations identify the rapid AI ecosystem as the main security risk that is related to GenAI. In other words, companies choose to adopt AI solutions as quickly as possible, leaving aside the complete securitization of their assets. (Thales, 2025)[15] The adoption of these systems implies information security challenges, since especially generative AI and agentic systems need to work with quality data. Therefore, they are not aware of all the information exposed on the internet, since they also decrease their investments in security testing to be able to invest in the automation of their processes.

For these reasons, in this work, an integrated security analysis platform is presented that combines OSINT, web and network scanning, powered by artificial intelligence through RAG architecture. The system automates vulnerability detection using both proprietary and third-party automated tools, and improves result interpretation through contextualized knowledge retrieval. The implementation of the RAG architecture represents the main point of the project, since it significantly improves the analysis and interpretation of the results obtained by the different tools. This approach considerably reduces the “hallucinations” or inaccuracies, typical of LLM language models, by anchoring the analysis in verifiable and contextualized data. The system retrieves relevant information from a database and generates precise contextualized analyses.

The tool therefore provides companies with a broad vision of their security situation and through reports suggests an action plan based on the findings found.

Keywords

OSINT, Cybersecurity, Artificial Intelligence, RAG, Vulnerability Analysis, Security Automation, Report Generation.

Agradecimientos

En primer lugar, quiero expresar mi agradecimiento a mi tutor, Miguel Angel de Cara Ruiz, por su orientación y apoyo durante todo el desarrollo de este trabajo. Sus ideas y conocimientos han sido claves para llevar a cabo el trabajo.

También quiero agradecer a mi familia, por su apoyo en los momentos más difíciles y a todos los que, de una u otra manera, han contribuido a mi crecimiento personal y académico durante estos años, tanto amigos como profesores.

Índice

1. Introducción **9**
1.1. Contexto y motivación . . . 9
1.2. Objetivos del proyecto . . . 11

2. Metodología **13**

3. Contribuciones principales **15**

4. Marco teórico **16**
4.1. Open-Source Intelligence (OSINT) . . . 16
4.2. Análisis de Red . . . 17
4.3. Análisis Web . . . 17
4.4. Inteligencia Artificial en Ciberseguridad . . . 17
4.5. Agentes IA . . . 19
4.6. Componentes de un agente IA . . . 20
4.7. Retrieval Augmented Generation (RAG) . . . 21

5. Arquitectura del proyecto **23**
5.1. Visión General de la Arquitectura . . . 23
5.1.1. Componentes Principales . . . 24
5.2. Módulos de Análisis OSINT . . . 24
5.2.1. Dorking . . . 24
5.2.2. Descubrimiento de Empleados . . . 24
5.2.3. Análisis de S3 Buckets . . . 24
5.2.4. Detección de Leaks . . . 25
5.2.5. Integración con IntelX y HunterHow . . . 25
5.3. Módulos de Análisis Web . . . 25
5.3.1. Nuclei Scan . . . 25
5.3.2. WhatWeb . . . 25
5.3.3. Sublist3r . . . 25
5.3.4. Detección de Command Injection . . . 26
5.3.5. Descubrimiento de Subpáginas . . . 26
5.4. Módulos de Análisis de Red . . . 26
5.4.1. Nmap Scan . . . 26
5.4.2. Análisis de Vulnerabilidades SMB . . . 26
5.4.3. Análisis de Vulnerabilidades SNMP . . . 27
5.5. Arquitectura RAG para Análisis Inteligente . . . 28
5.5.1. Fundamentos de la Arquitectura RAG implementada . . . 28

5.5.2. Componentes del Sistema RAG 28
5.5.3. Flujo de Procesamiento RAG 30
5.5.4. Ventajas del Enfoque RAG 30
5.6. Informes generados 30
5.6.1. Estructura de los Informes 31
5.7. Flujo de Ejecución 32

6. Análisis de resultados **33**
6.1. Objetivos del análisis 33
6.2. Métricas empleadas 33
6.3. Umbrales definidos para el análisis de resultados 35
6.4. Resultados e interpretación 35
6.4.1. Escaneo 1 36
6.4.2. Resultados informe escaneo 1 36
6.4.3. Escaneo 2 41
6.4.4. Resultados informe Escaneo 2 41
6.5. Análisis Comparativo de los escaneos 45
6.6. Evaluación de Calidad de Informes 45
6.7. Conclusiones sobre la Eficacia del Sistema 46

7. Conclusiones **47**
7.1. Principales Hallazgos 47
7.2. Implicaciones Prácticas 47
7.2.1. Para Profesionales de Seguridad 48
7.2.2. Para Organizaciones 48
7.2.3. Para el Ecosistema de Seguridad 48
7.3. Limitaciones del Trabajo 48
7.3.1. Limitaciones Técnicas 48
7.3.2. Limitaciones Metodológicas 49
7.3.3. Limitaciones de Alcance 49
7.4. Líneas Futuras de Investigación 50
7.5. Reflexión Final 50

8. Anexos **52**
8.1. Anexo I 52
8.1.1. Whatweb 52
8.1.2. HunterHow 54
8.1.3. IntelX 55
8.1.4. Sublist3r 56
8.1.5. Google Dorking 56

8.1.6. Nmap 57
8.1.7. Nuclei 57
8.2. Anexo II 58
8.2.1. Motor de procesamiento de datos 58
8.2.2. Configuración y fases del proceso 58
8.3. Agentes AI con arquitectura RAG 60
8.3.1. Interfaz gráfica 61
8.3.2. Resultado de la interfaz 61

Glosario de Términos

Antivirus Software diseñado para detectar, prevenir y eliminar malware de sistemas informáticos. Utiliza firmas de virus conocidas y técnicas heurísticas para identificar amenazas.

API (Application Programming Interface) Conjunto de protocolos y herramientas que permiten la comunicación entre diferentes aplicaciones de software.

CVE (Common Vulnerabilities and Exposures) Sistema de identificación estándar para vulnerabilidades de seguridad conocidas públicamente, gestionado por MITRE Corporation.

CVSS (Common Vulnerability Scoring System) Sistema estándar para evaluar la gravedad de las vulnerabilidades de seguridad informática mediante una puntuación numérica.

EDR (Endpoint Detection and Response) Solución de ciberseguridad que monitoriza continuamente los endpoints para detectar y responder a amenazas avanzadas en tiempo real.

Firewall Sistema de seguridad de red que controla el tráfico entrante y saliente basándose en reglas de seguridad predeterminadas.

IA (Inteligencia Artificial) Tecnología que permite a las máquinas simular procesos de inteligencia humana, incluyendo aprendizaje, razonamiento y autocorrección.

LLM (Large Language Model) Modelo de inteligencia artificial entrenado con grandes cantidades de datos textuales para generar y comprender lenguaje natural.

Machine Learning Subset de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente.

Malware Software malicioso diseñado para dañar, interrumpir o obtener acceso no autorizado a sistemas informáticos.

OSINT (Open Source Intelligence) Disciplina de recolección y análisis de información disponible públicamente para propósitos de inteligencia.

Penetration Testing Práctica de seguridad que simula un ataque cibernético contra un sistema para encontrar vulnerabilidades explotables.

RAG (Retrieval-Augmented Generation) Arquitectura de IA que combina la recuperación de información con la generación de texto para producir respuestas más precisas y contextualizadas.

Superficie de Ataque Conjunto total de puntos donde un atacante no autorizado puede intentar ingresar o extraer datos de un entorno.

Threat Intelligence Información basada en evidencias sobre amenazas de seguridad existentes o emergentes que ayuda a informar las decisiones de seguridad.

TTP (Tactics, Techniques, and Procedures) Patrones de comportamiento de los atacantes que describen cómo realizan sus ataques.

Vulnerabilidad Debilidad en un sistema informático que puede ser explotada por amenazas para obtener acceso no autorizado o causar daño.

WAF (Web Application Firewall) Sistema de seguridad que filtra, monitoriza y bloquea el tráfico HTTP entre una aplicación web e internet, protegiendo contra ataques específicos de aplicaciones web.

1. Introducción

1.1. Contexto y motivación

En el panorama actual de la ciberseguridad, las organizaciones se enfrentan a un entorno de amenazas en creciente evolución y complejidad, debido también a la expansión que experimenta la superficie de ataque gracias a la adopción de nuevas tecnologías, tipos de arquitecturas y servicios ofrecidos como por ejemplo en la nube. Durante los últimos años, la IA ha revolucionado el sector tanto desde el punto de vista de las empresas, como para los atacantes. Si bien, por un lado, la IA automatiza los procesos corporativos y nacen herramientas de detección y análisis potenciadas por ella, para los atacantes se traduce en una mayor capacidad de ataque. Los atacantes pueden usar la IA para generar payloads automatizados y también pueden aprovechan la IA de las empresas para exfiltrar información. Es por este motivo que las empresas, en esta realidad, necesitan la implementación de herramientas siempre más sofisticadas para prevenir ataques, constituir capas de protección frente a ataques externos e internos. ¿Pero qué tal la información expuesta en internet de las compañías?

Según el Informe de Amenazas a los Datos de Thales 2025, casi el 70 % de las organizaciones identifican al rápido ecosistema de la IA como el principal riesgo de seguridad que está relacionado con la GenAI. Dicho en otras palabras, las empresas eligen adoptar soluciones IA lo más rápido posible, dejando de lado la segurización completa de sus activos.(Thales, 2025)[15].

La adopción de estos sistemas implica desafíos de segurización de las informaciones, ya que sobre todo las IA generativa y los sistemas agénticos necesitan trabajar con datos de calidad y, por tanto, pueden ser causa de la exfiltración de informaciones. Las empresas, por tanto, no son conscientes de toda la información expuesta en internet y también priorizan en sus inversiones en la automatización de sus procesos mediante la IA en vez de la segurización de dichos procesos y sistemas, olvidando que la información de la empresa expuesta en internet puede también facilitar el bypass de las protecciones corporativas y que estas constituyen un real peligro para la empresa. Es el caso, por ejemplo, de leaks de credenciales tanto en el internet indexado que todos conocemos, como en la deep web, publicaciones de empleados que violan normativas empresariales, curriculums de empleados que indican tecnologías específicas internas como protocolos y redes reconducibles a la empresa, facturas y documentos expuestos en internet que pueden contener informaciones reservadas tanto de empresas como de los empleados y s3 buckets en caso de que se implemente AWS de una forma incorrecta, etc. (IT Digital Media Group, 2023)[9]

Es para este motivo que nacieron las investigaciones OSINT llevadas a cabo por profesionales investigadores. Estas investigaciones son importantes tanto en el ámbito de la ciberseguridad, contribuyendo al descubrimiento de la información pública de la empresa para

prevenir por ejemplo ataques de phishing, como a la hora de realizar fusiones entre empresas o a la hora de planear una expansión empresarial mediante una adquisición. Es aquí, donde la mayoría de las veces se contrata un investigador OSINT para verificar e investigar la situación de una empresa al fin de verificar su situación real.

Las investigaciones OSINT, se constituyen por un mix de escaneos realizados a través de herramientas automáticas e investigaciones manuales por parte de expertos. Al final, se generan resultados que han de ser analizados. Estas investigaciones presentan limitaciones debidas a:

- **Sobrecarga de información**: los escáneres de seguridad generan enormes volúmenes de datos, haciendo que el análisis manual sea lento y propenso a errores.
- **Falta de contextualización**: las herramientas individuales carecen de la capacidad para correlacionar hallazgos entre diferentes dominios, perdiendo así información crítica sobre posibles vectores de ataque complejos.
- **Dependencia de expertos**: La interpretación efectiva de los resultados requiere profesionales altamente cualificados y con experiencia en múltiples dominios.

En base a estas limitaciones, se podría argumentar de que los investigadores necesitan ayuda para analizar este gran volumen de datos no contextualizados para poder proporcionar sucesivamente a la empresa, un plan de contingencia y pasos a seguir para limitar la superficie de ataque. En este escenario, la IA puede dar un contributo significativo.

En los últimos 4 años, se ha experimentado un cambio radical de la sociedad debido a la introducción al público de la inteligencia artificial. La inteligencia artificial, y particularmente los modelos de lenguaje de gran escala (LLM), han demostrado un potencial significativo para transformar diversos campos, incluida la ciberseguridad. Sin embargo, estos modelos presentan limitaciones cuando se aplican directamente a tareas de análisis de seguridad, siendo la más notable su tendencia a generar información inexacta o .ªlucinaciones", especialmente cuando carecen de acceso a datos actualizados y específicos. Problema, que puede ser parcialmente resuelto si al LLM se le combina una arquitectura RAG (Retrieval-Augmented Generation), la cual emerge como una solución a estas limitaciones, combinando la potencia de los LLM con la precisión del acceso a información específica. Este trabajo propone un sistema integral que aprovecha el potencial de RAG para mejorar el análisis de seguridad, proporcionando evaluaciones de vulnerabilidades más precisas y contextualizadas a partir de una investigación OSINT previa, realizada mediante herramientas finalizadas a recolectar información de fuentes distintas. A partir del análisis, se genera automáticamente un informe que viene impreso en la interfaz visual del proyecto, indicando puntos críticos y acciones correctivas y preventivas que puede tomar la empresa para poder mejorar su situación.

1.2. Objetivos del proyecto

El presente trabajo tiene como objetivo principal el diseño, implementación y evaluación de un sistema de análisis de seguridad basado en arquitectura RAG que integre y mejore los resultados de múltiples herramientas de escaneo en los dominios OSINT, web y network. Esta herramienta automatizada, no solo recopila información expuesta de la empresa, sino que también proporciona un análisis inteligente para identificar posibles riesgos de seguridad, lo que es la base de la diferencia entre OSINT y OSINF. El OSINT, diferentemente del OSINF, no comprende simplemente la recopilación de información expuesta en internet, sino que consiste en "sacar la inteligencia", es saber extraer un resultado de la información. En investigaciones OSINT, se busca información en fuentes abiertas, ya que sean foros, papers de universidades y fuentes cerradas, más complejas de acceder y para las cuales se necesita un rol diferente del simple usuario, como por ejemplo una base de datos de la policía, para la cual se necesitan credenciales y un rol válido para poderla visualizar.

Específicamente, se persiguen los siguientes objetivos:

1. **Diseñar e implementar una arquitectura RAG** que permita reducir significativamente las alucinaciones en los análisis generados por IA.

2. **Implementación de una Base de Datos vectorial** alimentada por un sistema de tratamiento de los datos generados por la investigación automática.

3. **Integrar herramientas de escaneo de los dominios elegidos** en un sistema unificado controlado por un orquestador:

    - **OSINT:** dorking, descubrimiento de empleados, análisis de S3 buckets, detección de filtraciones tanto en internet indexado, como en la deepweb, descubrimiento de Ips y dominios asociados a la empresa.
    - **Web:** escaneos de vulnerabilidades, descubrimiento de informaciones sobre tecnologías implementadas en el dominio principal, descubrimiento subdominios, detección de parámetros inyectables en la url del dominio, y descubrimiento de subpáginas.
    - **Network:** escaneos de puertos abiertos, análisis de vulnerabilidades en servicios SMB y SNMP.

4. **Implementar mecanismos de parseo de la información** para poder almacenar la información de forma más estructurada en la base de datos vectorial.

5. **Reducir la tasa de falsos positivos y falsos negativos** mediante la aplicación de análisis contextual basado en IA con conocimiento específico proporcionado por la información encontrada por la investigación.

6. **Desarrollar un sistema de generación de informes** que proporcione análisis detallados y recomendaciones accionables adaptadas al contexto específico de la organización evaluada.

2. Metodología

Para el desarrollo de este proyecto se ha adoptado una metodología basada en la investigación teórica, desarrollo de código y validación de los resultados. Este desarrollo del sistema de análisis de seguridad se estructura en torno a tres componentes principales: la automatización de técnicas de reconocimiento y análisis de vulnerabilidades, la integración de inteligencia artificial mediante arquitectura RAG para el análisis contextualizado de resultados, y la generación automatizada de reportes automáticos, basados en la información obtenida. Se adopta una perspectiva centrada en los usuarios finales, cuáles profesionales de ciberseguridad y empresas. Por tanto, se ha priorizado en el desarrollo de la plataforma, la usabilidad, precisión y eficiencia del sistema. Asimismo, la implementación de la inteligencia artificial está enfocada en el análisis y reducción de las alucinaciones, comunes en los LLM, para evaluar de forma eficiente la situación de seguridad de la empresa. El desarrollo de la plataforma ha sido incremental. Se han conectado los componentes al cuerpo del programa, solo después de haber comprobado la funcionalidad del mismo.

A continuación se describen las fases del proyecto:

1. **Fase de investigación**: se ha llevado a cabo un estudio exhaustivo del funcionamiento tanto de los LLM y sus limitaciones, como de las herramientas de seguridad disponibles en internet, la arquitectura RAG y las técnicas de investigación OSINT.
    - La selección de las herramientas se ha llevado a cabo evaluando para cada una, el objetivo, la calidad de los datos generados, la integración mediante API, el formato de generación de los datos (preferiblemente JSON).
    - El modelo LLM para la generación del informe ha sido seleccionado según estándares de calidad y estabilidad de la implementación. La implementación del LLM, debe basarse en un código estable, que permita una ágil intercambiabilidad de modelos.
    - El lenguaje de programación implementado es Python, unos de los lenguajes más populares y de alto nivel.
2. **Fase de diseño**: se ha definido la arquitectura del sistema a implementar, además de la selección de componentes, diseño la interfaz y definición del flujo de información tanto entre las herramientas de escaneo, como entre base de datos vectorial y el LLM.
3. **Fase de implementación**: se ha llevado a cabo un desarrollo progresivo de los componentes del sistema:
    - Algoritmos separados de las varias herramientas de escaneo.
        - División de los algoritmos en carpetas, según el dominio correspondiente.

        - Interconexión básica entre los algoritmos que comparten datos y variables.
    - Integración de herramientas de escaneo.
        - Implementación de un orquestador, el cual gestiona la ejecución de los escaneos mediante llamadas asíncronas.
    - Implementación de la arquitectura RAG.
        - Creación del sistema de tratamiento de datos.
        - Creación de la Base de datos vectorial.
        - Definición de procedimiento de recuperación de la información a partir de la base de datos y la query del usuario.
    - Desarrollo de análisis contextual.
    - Interconexión integral del sistema de generación de informes.
4. **Fase de evaluación**:
    - Los algoritmos de escaneos han sido testeados previamente sobre máquinas disponibles en plataformas de ciberseguridad como HacktheBox o TryHackme.
    - Se ha efectuado una validación integral del sistema mediante:
        - Una primera investigación manual de la empresa elegida (con consentimiento).
        - Comprobación de los resultados de la herramienta
        - Análisis comparativo de los resultados.
    - Realización de pruebas sobre empresas inscritas a plataformas de Bug Bounty como HackerOne deshabilitando algoritmos de escaneos no permitidos en la plataforma correspondiente.
5. **Evaluación de los resultados**: cada ejecución del programa se validaba el informe generado, cuya calidad influenciaba los parámetros del LLM y el modelo a elegir, el prompt interno que define la estructura del informe y el algoritmo de tratamiento y almacenamiento de los datos. Los resultados deben cumplir con altos estándares definidos en el apartado correspondiente del presente trabajo.
6. **Documentación técnica**: comprende la elaboración de la memoria técnica, manual de uso.

3. Contribuciones principales

El trabajo se centra en aportar un sistema de investigación OSINT y búsqueda de vulnerabilidades automatizado, caracterizado la integración con una arquitectura RAG conectada a un LLM enfocada a la generación de informes de seguridad, lo que contribuye a la reducción de alucinaciones de la inteligencia artificial y permite un análisis más rápido respecto al manual. El algoritmo de la plataforma implementa mecanismos de contextualización que permiten priorizar hallazgos según su relevancia específica en el entorno analizado. El objetivo último del presente trabajo es la generación automática de informes de seguridad basados en el análisis automatizado y contextualizado de la información recopilada mediante la investigación.

4. Marco teórico

4.1. Open-Source Intelligence (OSINT)

OSINT es un conjunto de herramientas y técnicas utilizadas para recopilar información pública, analizar datos y relacionarlos para convertirlos en conocimiento útil. Además, son las siglas en inglés de Open Source Intelligence, que en español se traduce como inteligencia de fuentes abiertas y que hace referencia a este planteamiento. Esta metodología se utiliza en diversos sectores como los ámbitos militar, policial, tecnológico, financiero y de marketing, entre otros, y permite acceder a todos los datos disponibles en cualquier fuente pública, a cualquier cosa que queramos investigar, entre las que están las personas físicas o las empresas. Toda esta información proviene no solo de documentos públicos e imágenes satelitares, sino también de sitios web, foros y redes sociales, es decir, cualquier fuente que sea accesible y no presente restricciones legales (Ferreira, 2023)[3].

Entre las técnicas principales destacan:

- **Google Dorking:** uso de operadores avanzados de búsqueda para encontrar información sensible indexada por los motores de búsqueda como Google, Bing y Yahoo. Esta técnica permite descubrir archivos confidenciales, paneles de administración expuestos, y credenciales filtradas de diferentes formatos.
- **Descubrimiento y análisis de perfiles de empleados:** se realiza un mapeo de la estructura organizacional de una empresa mediante análisis de redes sociales profesionales como Linkedin combinado con técnicas de dorking en caso no sea permitido el scraping. Esta información facilita ataques de ingeniería social o spear phishing.
- **S3 Buckets:** se trata del proceso de identificación de repositorios de almacenamiento en la nube de AWS, mal configurados y que pueden contener información sensible o permitir modificaciones no autorizadas. Existen tanto herramientas automatizadas, como es posible la búsqueda manual mediante dorks con correspondiente análisis por parte de los investigadores.
- **Análisis de la infraestructura web de la empresa:** mediante herramientas de escaneo se pueden detectar dominios y tecnologías que implementa la empresa, cuya desactualización puede constituir un potencial punto de explotación.
- **Integración con servicios de inteligencia:** plataformas como IntelX permiten ampliar el alcance de la investigación mediante acceso a bases de datos especializadas de vulnerabilidades y exposiciones.

4.2. Análisis de Red

En esta fase se realiza mediante herramientas como nmap un descubrimiento del sistema operativo, un escaneo de puertos abiertos y servicios que se ejecutan por cada uno de ellos. Es muy importante, sobre todo en entornos reales, qué escaneos como el que realiza nmap, se ejecuten, sea en modalidad stealth o silenciosa, para no afectar a los servicios de la IP objetiva y evitar detección por parte de eventuales firewalls.

4.3. Análisis Web

Una página web es la "carta de presentación de una empresa", es fundamental para dar a conocer los que son sus valores y los servicios proporcionados a los clientes. La página web es un conjunto de recursos (páginas HTML, scripts, APIs, bases de datos) accesibles por un navegador o cualquier cliente HTTP.

Todavía, el hecho de estar naturalmente siempre expuesta a internet para obvios motivos, la puede convertir en un objetivo atractivo por parte de atacantes, ya que estas suelen manejar datos sensibles como credenciales, información personal de clientes y tarjetas de pago. Además, las aplicaciones web tienen múltiples puntos de entrada, cuáles formularios, parámetros en URL y APIs públicas, lo que amplía consideradamente su superficie de ataque.

Los atacantes suelen perseguir cuatro grandes objetivos:

1. **Robo de datos:** exfiltrar información confidencial.
2. **Control del servidor:** ejecutar código en el backend (RCE) o desplegar puertas traseras.
3. **Interrupción de servicio:** provocar fallos o denegación de servicio (DoS).
4. **Escalada de privilegios:** aprovechar fallos menores para llegar a niveles más profundos de la infraestructura.

El análisis web se puede llevar a cabo de forma manual a través de operaciones ejecutadas por expertos del sector. Todavía, aunque este análisis pueda ser del todo manual, la implementación de herramientas de escaneos de vulnerabilidades web y API, como Zap o OpenVAS, puede representar un consistente ahorro de tiempo y mayor precisión.

4.4. Inteligencia Artificial en Ciberseguridad

Un modelo de lenguaje de gran tamaño (LLM) es un tipo de modelo de inteligencia artificial que emplea técnicas de machine learning (aprendizaje automático) para comprender y generar lenguaje humano. Estos modelos pueden resultar muy valiosos para las empresas y las entidades que buscan automatizar y mejorar diversos aspectos de la comunicación y

del procesamiento de datos. Los LLM utilizan modelos basados en redes neuronales y técnicas de procesamiento del lenguaje natural (NLP) para procesar y calcular sus resultados. El NLP es un campo de la inteligencia artificial que se centra en lograr que las computadoras comprendan, interpreten y generen texto. Esto, a su vez, permite que los LLM realicen diversas tareas: analizar texto y sentimientos, traducir distintos idiomas y reconocer voces. (Los Modelos de Lenguaje de Gran Tamaño O LLM: ¿Qué Son Y Cómo Funcionan?, n.d.)[14])

La aplicación de inteligencia artificial en el ámbito de la ciberseguridad representa un avance significativo en la capacidad de análisis, detección y respuesta a amenazas. Desde el punto de vista defensivo, como se menciona en la página oficial de Fortinet: “La IA permite a los sistemas de ciberseguridad analizar grandes cantidades de datos, identificar patrones y tomar decisiones informadas, a velocidades y escalas más allá de las capacidades humanas.” Por tanto, la IA permite quitar a los humanos todas tareas repetitivas, acelera la ejecución de largas tareas de análisis y permite reconocer amenazas en tiempo real asegurando una respuesta y una mitigación más rápida.

Por otro lado, si se considera el uso de IA por atacantes, hay que tener en cuenta las siguientes consecuencias mencionadas en la página de malwarebytes (Malwarebytes, 2024)[10]:

1. Optimización de ataques cibernéticos: mediante grandes modelos (LLM) es posible optimizar técnicas de phishing y ataques de ransomware.

2. Malware automatizado: a programadores, los LLM permiten crear código malicioso automatizado.

3. Seguridad física: siendo la IA presente en muchas infraestructuras, vehículos autónomos, equipos de construcción, aumentan los riesgos de seguridad física al ser atacable.

4. Riesgo de privacidad: al integrarse con aplicaciones de uso cotidiano, muchas veces se usan para recopilar informaciones de usuarios de forma no autorizada. Es por ejemplo el caso de Chatgpt o Whatsapp, el cual implementa Meta AI.

5. Robo de modelos de IA: riesgo posible gracias a técnicas de ingeniería social y/o ataques de red.

6. Manipulación y envenenamiento de modelos: la IA es un elemento vulnerable a envenenamiento de datos, inyectado informaciones incorrectas que se usarán para el entrenamiento del modelo o a inyecciones de comandos, técnica que puede llevar a la revelación de informaciones confidenciales.

7. Suplantación de identidad: mediante la simulación de voz a través de modelos específicos o creación de deepfake.

8. Ataques sofisticados: automatización de scripts para sparephishing, aumento de la complejidad de scripting para la creación de malware.

9. Daño reputacional: si los sistemas de IA fallan o vienen vulnerados y resulta haberse dado una filtración de informaciones, la empresa puede experimentar una pérdida de reputación.

Los sistemas de IA no son conscientes de sí mismos por lo que son sistemas vulnerables e imprecisos. En los sistemas en los que se implementa IA, alucinaciones y obsolescencia de datos minan precisión de las respuestas.

4.5. Agentes IA

Un agente de inteligencia artificial es un programa que puede interactuar con el entorno de forma independiente. Un agente tiene unos objetivos definidos por el humano y para lograrlos es capaz de recolectar datos de un entorno específico y realizar tareas secuenciales de forma independiente. A diferencia de los LLM o Large Lenguage model como Chat GPT de OpenAi, los cuales responden a entradas de texto (prompts) y generan una respuesta, actuando, por lo tanto, no de forma independiente, un agente AI interactúa con un entorno físico o digital. También es posible incorporar una memoria y definir un algoritmo de aprendizaje continuo, permitiéndole adaptarse a nuevas tareas y mejorar la calidad de los resultados generados. Un claro ejemplo de agente IA es un asistente virtual de escritorio que puede organizar eventos de calendario, agendar citas y todas las tareas para las cuales ha sido definido.

4.6. Componentes de un agente IA

De acuerdo con Amazon Web Services, los agentes pueden tener objetivos diferentes, pero todos están compuestos por los siguientes componentes:

- **Arquitectura**: base del agente que puede ser física, software o ambas. Un agente robótico, por ejemplo, puede usar sensores y motores, mientras que un agente basado en una arquitectura software implementa APIS u otros componentes software.
- **Función del agente:** define como los datos se transforman en acciones para cumplir objetivos. Se puede implementar el uso de mecanismos de retroalimentación mediante bases de datos vectoriales, modelos de Inteligencia Artificial, como en caso de tener que generar textos comprensibles para el usuario y otras funciones o herramientas en relación con el objetivo a conseguir.
- **Programa del agente:** Es la implementación práctica de la función del agente, abarcando su desarrollo, entrenamiento y despliegue en la arquitectura correspondiente.

4.7. Retrieval Augmented Generation (RAG)

El RAG o Retrieval Augmented Generation, como confirma la página de IBM, es un tipo de arquitectura para potenciar el rendimiento de una inteligencia artificial conectándola a una base de datos externa. Al combinar los datos con las capacidades de los LLM, se genera un texto en output más preciso, ya que el proceso de generación del texto es actualizado con información relevante contenida en la base de datos.

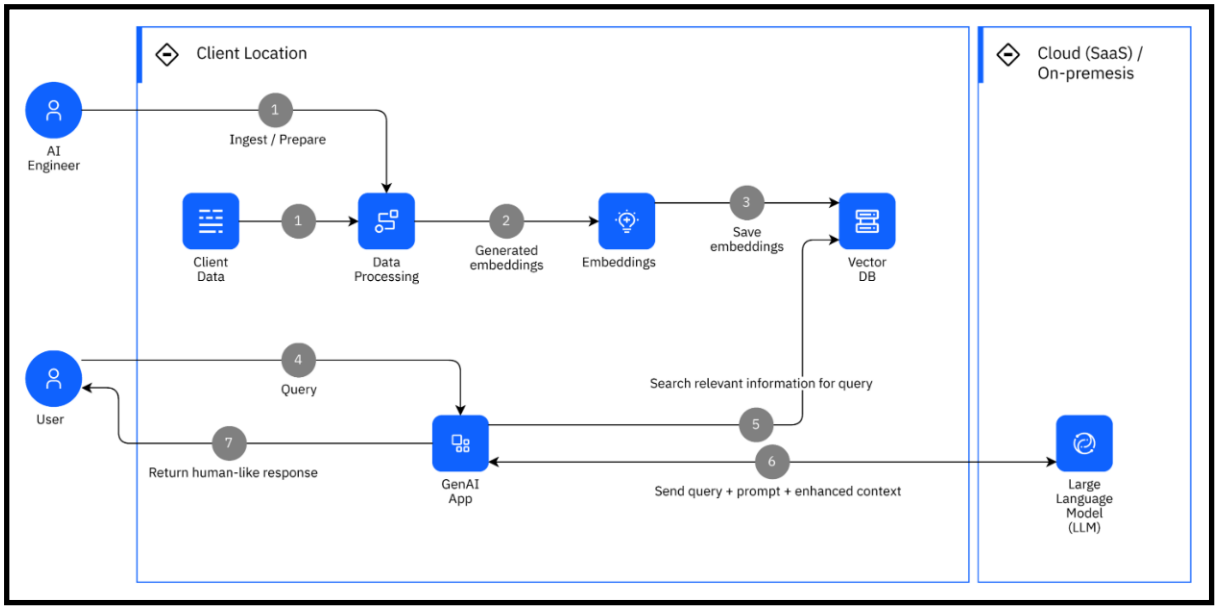
Esta arquitectura se compone de dos técnicas clave:

1. **Recuperación de información**: se consulta una base de datos de conocimiento o una base de datos vectorial para recuperar fragmentos de texto relevantes.

2. **Generación de lenguaje natural**: al implementar en el proceso un LLM preentrenado y del cual se mejora el contexto para que este tenga una mayor comprensión del tema, se genera una respuesta basada en los datos recuperados y de hecho más precisa.

Figura 1

Diagrama de arquitectura RAG de IBM.



Nota. Fuente: (AI RAG, n.d.) [2]

Cabe destacar como una de las ventajas de utilizar una arquitectura RAG, es que este ayuda a mitigar las que son definidas como “alucinaciones de la IA”.

Estas “alucinaciones” son errores que se producen debido a algunos factores como datos incorrectos, incoherentes o hasta la falta de estos. Estas alucinaciones de IA, en acorde a la página de Google, pueden presentarse en formas distintas:

- Predicciones incorrectas: consiste en el predecir un evento considerado poco probable o imposible.
- Falsos positivos: el modelo detecta por ejemplo una amenaza cuando en realidad no lo es.
- Falsos negativos: no se identifica correctamente una amenaza.

Además, mejora la interpretabilidad de los resultados al poder rastrear las fuentes y permite actualizar la base de conocimiento sin necesidad de reentrenar el modelo completo.

5. Arquitectura del proyecto

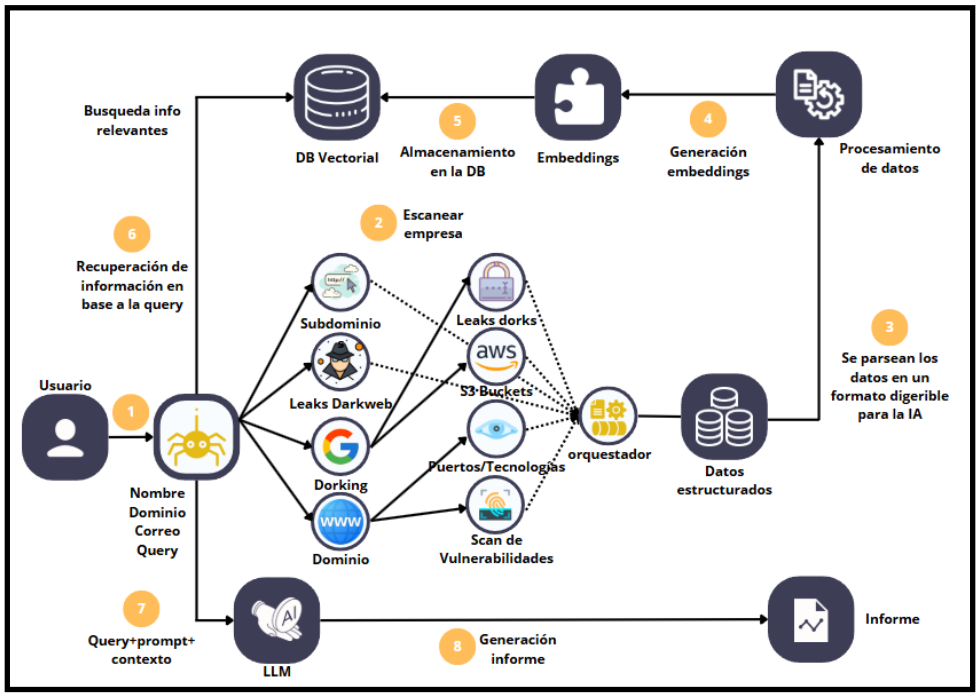
La arquitectura del sistema desarrollado está diseñada para proporcionar una solución integral de análisis de seguridad y se compone, para el módulo de análisis, de tres dominios principales: OSINT (Open Source Intelligence), seguridad web y seguridad de red. El sistema se compone también de otros módulos que llevan a cabo el tratamiento, procesamiento, análisis y recuperación de la información. Al final de los varios procesos secuenciales, se genera un reporte mediante un proceso de recuperación de la información y generación de texto a través de la conexión con un LLM.

5.1. Visión General de la Arquitectura

La arquitectura del sistema sigue un enfoque modular que permite la ejecución coordinada de diferentes herramientas de análisis de seguridad. Los algoritmos encargados de la recopilación de la empresa y de la ejecución de escaneos de seguridad. Los módulos están especializados y agrupados en dominios comunes y comparten una infraestructura común. El sistema implementa una arquitectura RAG (Retrieval-Augmented Generation) para el análisis inteligente de los resultados obtenidos por las distintas herramientas.

Figura 2

Esquema de la Arquitectura implementada.



Nota. Elaboración propia

5.1.1. Componentes Principales

La arquitectura se compone de los siguientes componentes principales:

- **Frontend:** interfaz de usuario para la configuración de escaneos y visualización de resultados.
- **Backend:** un algoritmo central denominado orquestador se encarga de la coordinación de las llamadas asíncronas a los varios módulos de escaneo.
- **Módulos de Escaneo:** algoritmos encargados de la recopilación de la información basados sobre tres dominios principales: OSINT, análisis web y escaneo de red.
- **Sistema RAG:** módulo de inteligencia artificial para el análisis y contextualización de resultados, constituido por una base de datos vectorial, un algoritmo de recuperación de la información basada en contexto y un LLM para la generación del texto del informe

5.2. Módulos de Análisis OSINT

El módulo OSINT se encarga de recopilar información de fuentes públicamente accesibles para identificar posibles vectores de ataque o filtraciones de información sensible.

5.2.1. Dorking

Implementación de búsquedas avanzadas en motores cuáles Google y Bing, para identificar información sensible o vulnerabilidades expuestas. El sistema automatiza la construcción y ejecución de consultas basadas en operadores de búsqueda avanzados llamados dorks.

5.2.2. Descubrimiento de Empleados

Algoritmo de dorking para identificar empleados asociados a una organización a través de dorks específicos que combinan en la query el nombre de la empresa con la plataforma profesional Linkedin. Al final del proceso se genera un archivo que contiene los nombres de los empleados encontrados y su respectivo link de Linkedin. Esta información se utiliza para mapear la estructura organizativa y potenciales objetivos de ingeniería social.

5.2.3. Análisis de S3 Buckets

Sistema de descubrimiento y análisis de buckets S3 públicos o mal configurados asociados al objetivo, capaz de:

- Identificar buckets basados en patrones de nomenclatura comunes.
- Verificar permisos y configuraciones de acceso.

5.2.4. Detección de Leaks

Componente especializado en la identificación de filtraciones de datos relacionadas con el objetivo, mediante la creación automatizada de dorks. Se buscan filtraciones según páginas comunes de filtraciones y por extensión de documentos comunes como pdf y json. Se implementa un algoritmo de validación de links guardados para reducir el número de falsos positivos.

5.2.5. Integración con IntelX y HunterHow

El sistema incorpora APIs de inteligencia de amenazas de terceros:

- IntelX: para búsqueda avanzada de filtraciones relevantes en la superficie, deep y dark web.
- HunterHow: para obtener información sobre tecnologías, dominios, ips y exposición de activos.

5.3. Módulos de Análisis Web

El componente de análisis web está diseñado para identificar vulnerabilidades en aplicaciones y servicios web expuestos.

5.3.1. Nuclei Scan

Implementación de escaneo basado en plantillas utilizando Nuclei, que permite tanto la detección automatizada de vulnerabilidades conocidas, como por ejemplo las que se encuentran clasificadas en el top 10 OWASP y la identificación de malas configuraciones.

5.3.2. WhatWeb

Integración de WhatWeb para el fingerprinting de tecnologías web que ha permitido:

- Identificación de CMS y frameworks.
- Detección de versiones potencialmente vulnerables.
- Reconocimiento parcial de componentes de infraestructura

5.3.3. Sublist3r

Se ha llevado a cabo una integración de la herramienta Sublist3r en el código del programa para cumplir un descubrimiento de subdominios mediante algoritmos automatizados de fuerza bruta. Siendo una enumeración basada en diccionarios largos, se ha mantenido

un bajo perfil para evitar detecciones y bloqueos de IP, alargando el período de ejecución gracias a un menor número de peticiones por segundo y a un menor número de hilos de ejecución en paralelo.

5.3.4. Detección de Command Injection

Módulo especializado en la identificación de vulnerabilidades de inyección de comandos en la URL de la página oficial. El algoritmo lleva a cabo diferentes tareas, como el análisis de parámetros en las peticiones GET, mediante el uso de diccionarios y un proceso de verificación de la respuesta. Una vez se hayan descubierto parámetros válidos, se realizan pruebas de inyección con payloads personalizados para comprobar si dicha vulnerabilidad existe. Las inyecciones también vienen validadas para minimizar falsos positivos.

5.3.5. Descubrimiento de Subpáginas

El descubrimiento de las subpáginas del dominio, viene llevado a cabo mediante un algoritmo de fuerza bruta. Dicho algoritmo, implementa un diccionario de posibles palabras comunes en cuanto a páginas web. Se realiza el Crawling en respeto a políticas robots.txt para la identificación exhaustiva de rutas y recursos dentro de la página web. Las subpáginas son validadas mediante la comprobación del código de estado de la petición web.

5.4. Módulos de Análisis de Red

El módulo de análisis de red se enfoca en la identificación de servicios expuestos y vulnerabilidades a nivel de infraestructura a partir de la IP descubierta mediante Whatweb.

5.4.1. Nmap Scan

Implementación avanzada de escaneo de puertos y servicios mediante Nmap:

- Escaneo lento y sigiloso.
- Identificación de servicios y versiones.
- Detección de sistemas operativos basado en ttl recibido de la traza ICMP.
- Detección de vulnerabilidades basadas en puertos y servicios encontrados.

5.4.2. Análisis de Vulnerabilidades SMB

Algoritmo que busca detectar configuraciones inseguras y vulnerabilidades en servicios SMB. Primero, se enumeran posibles recursos compartidos en caso de que los puertos en los que corre el servicio estén abiertos. Si existen dichos recursos, se busca identificar posibles

malas configuraciones de permisos, por ejemplo, intentando efectuar pruebas de autenticación nula y acceso anónimo.

5.4.3. Análisis de Vulnerabilidades SNMP

Componente para la evaluación de seguridad en servicios SNMP. Se efectúan, en caso de que esté activo el servicio, pruebas de community strings por defecto y extracción de información de configuración.

Para saber más sobre los escaneos implementados, consultar Anexo I.

5.5. Arquitectura RAG para Análisis Inteligente

Es el componente clave de este proyecto. La implementación de la arquitectura Rag conectada a un modelo de lenguaje o LLM permite un análisis inteligente e inmediato de los resultados obtenidos por los escaneos y la investigación OSINT.

5.5.1. Fundamentos de la Arquitectura RAG implementada

La arquitectura RAG implementada en el proyecto se ha caracterizado por las siguientes acciones:

- **Recuperación (Retrieval):** Obtención de información relevante de la base de conocimiento constituida por los resultados de la investigación previa
- **Aumento (Augmentation):** Enriquecimiento del contexto con datos almacenados en la base de datos vectorial, la query del usuario y un prompt interno que indicaba el rol al agente y como debe ser estructurado el reporte
- **Generación (Generation):** Creación de análisis y recomendaciones mediante LLM (Large Language Models). El LLM implementado para la generación del texto es Gemini en su versión gemini-2.5-pro-preview-05-06, pero es intercambiable con otros modelos.

Esta arquitectura mitiga significativamente el problema de las alucinaciones (generación de información incorrecta) al anclar las respuestas del modelo en datos verificables. Además de que, sin el escaneo previo, un LLM no tendría acceso a la información almacenada en la base de datos, ya que "privada".

5.5.2. Componentes del Sistema RAG

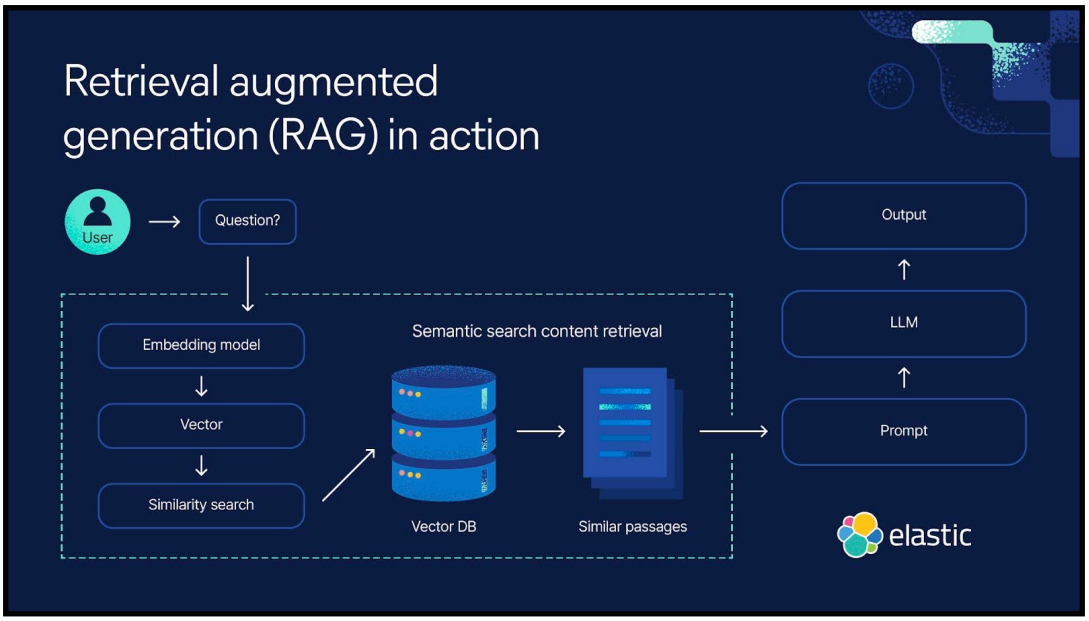
El sistema RAG implementado consta de los siguientes componentes:

1. **Base de Conocimiento Vectorial:** componente encargado del almacenamiento de información de seguridad indexada mediante embeddings.
2. **Sistema de Recuperación:** se basa en un sistema de búsqueda semántica realizada en función de la similitud vectorial entre las palabras. El sistema intenta priorizar la información crítica y recuperar todo lo que sea inherente a la query del usuario y otros parámetros presentes en el formulario de la interfaz de la plataforma.

3. **Modelo de Lenguaje:** se ha optado por la implementación de un modelo pre-entrenado por Gemini, ya que entrenar un modelo específico propio no era posible debido a limitaciones en cuanto a hardware necesario, tiempos y material disponible. La capacidad de interpretación técnica del modelo ha sido posible gracias al proceso de recuperación de la información contenida en la base de datos vectorial.

Figura 3

Componentes RAG y esquema



Nota. Fuente: (¿Qué Es la Generación Aumentada de Recuperación (RAG)? | una Guía Completa de RAG, s.f.) [13]

5.5.3. Flujo de Procesamiento RAG

El flujo de datos en la arquitectura RAG implementada sigues las fases aquí mencionadas:

1. **Procesamiento:** se procesan los datos obtenidos por la investigación previa y contenidos en archivos con extensión txt o json.
2. **Indexación:** cada documento se convierte en fragmentos o çhunks"más pequeños para facilitar la búsqueda semántica.
3. **Vectorización:** cada fragmento se vectoriza usando embeddings mediante la implementación de GoogleGenerativeAIEmbeddings. Los vectores generados se almacenan en una base de datos vectorial Chroma DB, para permitir búsquedas por similitud.
4. **Retrieval:** Cuando se recibe la consulta del usuario mediante la interfaz, esta se vectoriza y se buscan los fragmentos más relevantes, llamados "nearest neighbors", en ChromaDB usando la similitud vectorial de embeddings.
5. **Formación del contexto:** se ensamblan los fragmentos recuperados junto con la consulta del usuario y el nombre de la empresa a analizar para formar el contexto. El contexto viene enviado sucesivamente al LLM para la generación de la respuesta.
6. **Generación inteligente del informe:** el modelo LLM utiliza el contexto para generar una respuesta precisa y contextualizada.

5.5.4. Ventajas del Enfoque RAG

La implementación de esta arquitectura en el proyecto ha proporcionado beneficios significativos en cuanto a reducción de alucinaciones respecto a un normal modelo de LLM, ya que genera informes contextualizados según la información vectorial contenida en la base de datos, adaptación a los que son los objetivos del trabajo y escalabilidad, en cuanto se permite agregar cada vez nuevo conocimiento sin necesidad de ciclos de entrenamiento y ajuste de parámetros. Datos específicos sobre la precisión del sistema se encuentran descritas en el apartado de “Análisis de resultados”.

Para saber más sobre la arquitectura, consultar Anexo II

5.6. Informes generados

El informe generado es el resultado de todos los procesos de investigación y análisis automatizado e inteligente previos. Este tiene el compito de enseñar de forma clara las vulnerabilidades de la empresa basada en un análisis de riesgos contextualizado, además de los

pasos a seguir para mejorar la situación de la empresa en cuanto a exposición de informaciones en internet, filtraciones y vulnerabilidades estructurales.

5.6.1. Estructura de los Informes

Los informes generados siguen una estructura comprensiva que incluye:

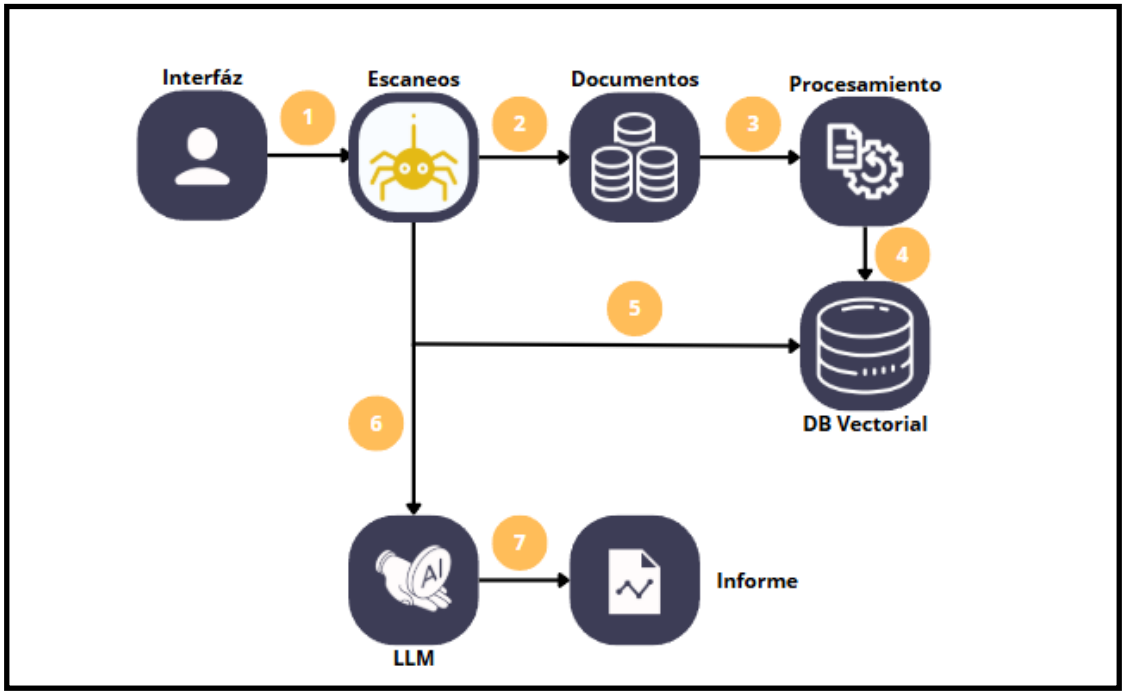
1. **Resumen ejecutivo:** contiene el resumen de los principales hallazgos y riesgos identificados
2. **Análisis de riesgos**
3. **Hallazgos técnicos detallados:** se indica criticidad, descripción, impacto, vector de ataque, recomendación y referencias para cada uno de los hallazgos
4. **Plan de mitigación:** contiene la lista de pasos a seguir por la empresa para remediar o limitar los principales problemas encontrados
5. **Recursos adicionales:** herramientas, guías, referencias externas

5.7. Flujo de Ejecución

Entrando a la plataforma, el usuario encontrará un formulario, con 4 parámetros a completar para poder iniciar el programa (nombre de la empresa, dominio de correo de la empresa, página web corporativa y query para la generación del informe). Una vez completado el formulario, el usuario puede dar inicio al programa. Primero, el backend coordina la ejecución secuencial o paralela de los distintos módulos de escaneo interconectados, los cuales han sido implementados para recopilar información basada en los datos del formulario. Una vez completados los escaneos, la información se almacena en documentos que tendrán el formato txt o json. Los documentos contenidos en la carpeta de la empresa analizada, serán divididos en chunks, vectorizados y almacenados en una base de datos semántica (ChromaDB). El sistema RAG coordina sucesivamente el análisis y recuperación semántica basada en contexto y completada por la pregunta del usuario. En fin, el modelo LLM integrado, produce un informe o respuesta estructurada que integra todos los hallazgos relevantes de la empresa (resumen ejecutivo, análisis de riesgos, detalles técnicos y plan de mitigación). Mediante la interfaz principal, el usuario puede consultar cualquier hallazgo presente en el informe impreso en pantalla y descargar el documento en formato markdown.

Figura 4

Esquema Flujo de Ejecución.



Nota. Elaboración propia

6. Análisis de resultados

En la siguiente sección se muestra el análisis de la eficacia de la arquitectura RAG implementada en la plataforma para la generación automática de reportes automáticos de seguridad. El análisis se ha realizado sobre una organización denominada .Empresa X"por motivos de privacidad y confidencialidad.

6.1. Objetivos del análisis

El análisis tiene como objetivo principal evaluar la eficacia del modelo RAG en la generación de reportes de seguridad precisos y útiles. Para lograr este objetivo, se han tomado dos casos de estudio sobre la misma empresa para evaluar el comportamiento del sistema bajo diferentes condiciones de carga de datos que permiten evaluar el comportamiento del sistema bajo diferentes condiciones de carga de datos y especificidad de la consulta del usuario.

En el primer escaneo se evalúa el rendimiento del sistema con una consulta general que requiere el procesamiento de todos los datos disponibles encontrados durante la fase de investigación, mientras que el segundo caso evalúa la capacidad del sistema para procesar consultas específicas con un subconjunto reducido de esa información.

Mediante el análisis de estos dos casos, se busca comprender no solo la precisión del sistema en cuanto a confiabilidad general y detección de reales vulnerabilidades, sino que también se evalúan los resultados en función del volumen de datos a recuperar y la especificidad de la consulta.

6.2. Métricas empleadas

Para la evaluación del sistema se han definido las siguientes métricas de evaluación:

1. **Cobertura de hallazgos**
    - **Total de vulnerabilidades detectadas:** se cuantifica el total de vulnerabilidades detectadas por cada herramienta de escaneo integrada en el sistema (Nuclei, Intel X, reconocimiento de dominios), así como el número de hallazgos únicos tras el procesamiento y filtrado realizado por el componente RAG. Cabe destacar que durante la ejecución de las pruebas, el escáner Nmap fue deshabilitado por restricciones de privacidad y permisos de la empresa objetivo, por lo que no forma parte del análisis de cobertura.
    - **Número de hallazgos únicos** tras el filtrado y generación RAG.
2. **Precisión y Recall (Precisión ponderada / Tasa de recuperación ponderada)**
    - **FP (Falsos Positivos):** hallazgos erróneos que no representan vulnerabilidades reales.

- **FN (Falsos Negativos):** vulnerabilidades reales que el sistema no detectó.
- **Precisión ponderada:** es la proporción de hallazgos encontrados por el sistema que efectivamente corresponden a vulnerabilidades reales.

  Dado que no todos los hallazgos tienen el mismo impacto en la seguridad organizacional, se implementa una métrica ponderada que asigna diferentes pesos según la categoría de riesgo establecida:

  - Riesgo **ALTO**: peso 3.
  - Riesgo **MEDIO**: peso 2.
  - Riesgo **BAJO**: peso 1.

  La métrica se calcula como:

$$\text{Precisión\_ponderada} = \frac{\sum(TP_i \times peso_i)}{\sum((TP_i + FP_i) \times peso_i)}$$

  proporcionando una evaluación más realista del impacto de la precisión del sistema.

  La evaluación de la métrica tiene como objetivo definir el grado de confiabilidad de los reportes generados.

- **Tasa de recuperación ponderado o Recall ponderado:** es la proporción de vulnerabilidades reales presentes que el sistema logra detectar.

  Similar a la precisión ponderada, la tasa de recuperación se pondera según la criticidad de las vulnerabilidades no detectadas y se calcula de la siguiente forma:

$$\text{Recall\_ponderado} = \frac{\sum(TP_i \times peso_i)}{\sum((TP_i + FN_i) \times peso_i)}$$

  Su objetivo es cuantificar el grado de exhaustividad del sistema.

3. **Evaluación cualitativa de la structura del informe**

   - Se mide de forma cualitativa la estructura de los reportes generados, incluyendo la presencia de todos los apartados indicados en el algoritmo del programa.

6.3. Umbrales definidos para el análisis de resultados

Para la interpretación de los resultados de las métricas calculadas, se han establecido umbrales de elaboración propia. Entre 0.90 y 1.00 se considera alta calidad, indicando que el sistema rara vez comete errores y la mayoría de los hallazgos son confiables para uso directo. El rango de 0.80 a 0.89 representa buena calidad, con algunos falsos positivos o negativos presentes, pero resultados útiles para toma de decisiones. Un rango entre 0.70 y 0.79 se categoriza como moderada, requiriendo revisión manual de algunos hallazgos, mientras que valores inferiores a 0.70 indican un bajo ratio de las métricas, con alta presencia de hallazgos erróneos.

Tabla 1

Tabla de clasificación por Umbrales

RANGOS DE CALIDAD	INTERPRETACIÓN
0.90 – 1.00	**Alta calidad** – El sistema rara vez comete errores. La mayoría de los hallazgos son confiables.
0.80 – 0.89	**Buena calidad** – Algunos falsos positivos o negativos, pero los resultados son útiles para tomar decisiones.
0.70 – 0.79	**Calidad moderada** – Aumentan los errores. Se recomienda revisión manual de algunos hallazgos.
<0.70	**Baja calidad** – Muchos hallazgos erróneos.

Nota. Elaboración propia.

6.4. Resultados e interpretación

Para el análisis de resultados, se presentan 2 ejemplos de reportes generados a partir de la misma cantidad datos sobre la misma empresa. El objetivo es cuantificar la precisión de la arquitectura implementada en función de la generación del reporte de seguridad, al variar de la cantidad de datos a recuperar. Por este motivo, el primer reporte se basa sobre una query del usuario sobre toda la información almacenada para la respectiva empresa, mientras que el segundo se basa sobre la recuperación de informaciones encontradas mediante un escáner específico, lo que requiere una menor cantidad de datos a recuperar.

6.4.1. Escaneo 1

En el primer escaneo se ha llevado a cabo un análisis integral de la empresa. Se listan los siguientes parámetros del escaneo:

1. Nombre Empresa: Empresa X
2. Correo Empresa: @empresa.es
3. Dominio Empresa: https://empresa.es
4. Query: Genera un reporte de seguridad en base a todas las informaciones encontradas sobre la empresa.

6.4.2. Resultados informe escaneo 1

La imagen del resumen ejecutivo muestra el estado general de seguridad de la organización, incluyendo el panorama de riesgo y recomendaciones prioritarias.

Figura 5

Resumen ejecutivo Escaneo 1.

Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

Durante este escaneo, el sistema RAG recuperó y analizó datos procedentes de todos los escáneres y ha individuado los siguientes hallazgos de seguridad clasificados según las categorías de riesgo ALTA, MEDIA y BAJA, como se puede notar en la imagen siguiente.

Para la categoría ALTA se han individuado dos vulnerabilidades, por un lado, se ha detectado un panel básico de login en uno de los subdominios encontrados y por el otro se hace referencia a las filtraciones detectadas gracias a la integración de la plataforma con

Intelx, las cuales pertenecen tanto a darkweb como al internet indexado. Estas filtraciones contienen archivos de diferentes extensiones, como es el caso de SQL, lo que se puede interpretar como una posible filtración de usuarios y contraseñas. También para la categoría MEDIA, han sido identificadas 2 vulnerabilidades principales. Primero se ha rescontrado la presencia de 2 puertos abiertos para el dominio principal de la empresa que pueden ser vulnerables a enumeración de usuarios. La segunda, hace referencia a desactualizaciones de plugins Wordpress, que constituyen posible objetivo de entrada por parte de atacantes. En cuanto a la categoría BAJA, aunque no es una vulnerabilidad, se hace referencia a la detección de un WAF. Si bien es un aspecto positivo tener un WAF, su efectividad depende de sus configuraciones y actualizaciones. Una mala configuración o desactualización puede representar un problema de protección frente a ataques.

Figura 6

Análisis de Riesgos Escaneo 1.

2. ANÁLISIS DE RIESGOS

Nivel de Riesgo	Hallazgo	Impacto	Probabilidad	Prioridad
ALTA	Subdominio docs[redacted].es con autenticación básica (Basic Auth) expuesta y panel de login delectado.	Acceso no autorizado a documentación sensible, posible escalada de privilegios y compromiso de la infraestructura.	Alta	1
ALTA	Detección de múltiples filtraciones de datos en la web oscura (Intelx), incluyendo referencias al dominio www.[redacted].es y archivos SQL (odigital.cc.sql).	Exposición de credenciales de usuarios, información confidencial de la empresa y datos personales, lo que podría resultar en robo de identidad, ataques dirigidos y daños a la reputación.	Alta	2
MEDIA	Presencia de puertos 2082, 2083 y 2096 expuestos para el dominio noel.es.	Estos puertos son comúnmente utilizados por cPanel/WHM, lo que podría permitir la enumeración de usuarios y la explotación de vulnerabilidades conocidas en estas plataformas.	Media	3
MEDIA	Utilización de versiones desactualizadas de plugins en el sitio web www.[redacted].es , incluyendo All-In-One Security (AIOS), Contact Form 7 Database Addon, Contact Form 7, Custom CSS and JS y Custom Post Type UI.	Exposición a vulnerabilidades conocidas en las versiones antiguas de los plugins, lo que podría permitir la ejecución remota de código, la inyección de scripts maliciosos y el acceso no autorizado a la base de datos.	Media	4
BAJA	Detección de un Web Application Firewall (WAF) en www.[redacted].es .	Si bien la presencia de un WAF es una medida de seguridad positiva, su efectividad depende de su configuración y mantenimiento. Una configuración incorrecta o desactualizada podría permitir la elusión del WAF y la explotación de vulnerabilidades.	Baja	5

3. HALLAZGOS TÉCNICOS DETALLADOS

3.1 EXPOSICIÓN DE CREDENCIALES Y DATOS SENSIBLES

- **CRITICIDAD**: ALTA
- **DESCRIPCIÓN**: Se han detectado múltiples filtraciones de datos relacionadas con la empresa [redacted] en la web oscura, incluyendo referencias al dominio www.[redacted].es y archivos SQL (odigital.cc.sql). Estos archivos podrían contener credenciales de usuarios, información confidencial de la empresa y datos personales.
- **IMPACTO**: Robo de identidad, ataques dirigidos, acceso no autorizado a sistemas internos, daños a la reputación y posibles consecuencias legales por incumplimiento de normativas de protección de datos.
- **VECTOR DE ATAQUE**: Explotación de vulnerabilidades en aplicaciones web, ataques de phishing, ingeniería social y reutilización de contraseñas.
- **RECOMENDACIÓN**:
  i. Realizar una investigación exhaustiva para determinar el alcance de la filtración y los datos comprometidos.
  ii. Invalidar las contraseñas de todos los usuarios y forzar el restablecimiento de contraseñas.
  iii. Implementar autenticación multifactor (MFA) en todos los sistemas críticos.
  iv. Monitorear la web oscura en busca de nuevas filtraciones de datos.
  v. Notificar a las autoridades competentes y a los usuarios afectados, según lo requerido por la ley.
- **REFERENCIAS**: OWASP Top 10, GDPR, CCPA
- **FUENTE**: leaks.json

Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

Para cada hallazgo de seguridad, en el informe se expone el nivel de criticidad, una breve descripción de la vulnerabilidad, una estimación del impacto que pueda tener una explotación de la misma y las recomendaciones aconsejadas para reducir el nivel de riesgo.

Figura 7

Hallazgos Escaneo 1.

Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

En fin, como se puede notar en la siguiente imagen, en el mismo informe se adjunta un plan de mitigación, basado en la prioridad y criticidad de los hallazgos, para llevar el porcentaje de riesgo a un nivel más bajo.

Figura 8

Plan de Mitigación Escaneo 1.

Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

Valoración precisión del modelo

Los resultados cuantitativos del primer escenario muestran que el escáner Nuclei identificó 2 hallazgos técnicos, correspondientes a la detección del WAF y la identificación de plugins WordPress desactualizados. La integración con IntelX proporcionó 114 filtraciones detectadas, las cuales fueron correctamente reportadas en el informe. El reconocimiento de dominios identificó 1 dominio crítico entre todos los detectados. No se registraron hallazgos adicionales a través de técnicas de Google dorking. A continuación se enseña una tabla resumen en la cual se ha aplicado el sistema de pesos a los hallazgos reportados según la criticidad asignada. Los puntos serán utilizados sucesivamente para el cálculo de las métricas ponderadas.

Tabla 2

Tabla Cálculo de puntos de Hallazgos Escaneo 1.

CATEGORÍA DE RIESGO	TIPO DE HALLAZGO	CANTIDAD	PESO	CLASIFICACIÓN	PUNTOS PONDERADOS
ALTO	Filtraciones Intel X	114	3	TP (Verdadero Positivo)	342
ALTO	Panel Login Básico	1	3	TP (Verdadero Positivo)	3
MEDIO	Puertos Abiertos	1	2	TP (Verdadero Positivo)	2
MEDIO	Plugins WordPress	1	2	TP (Verdadero Positivo)	2
BAJO	WAF Detectado	1	1	FP (Falso Positivo)	1

Nota. Elaboración propia.

Métricas Ponderadas por Criticidad:

A continuación se muestran los cálculos de ponderación realizados por categoría de riesgo (ALTO=3, MEDIO=2, BAJO=1):

- Riesgo ALTO: 114 filtraciones (TP=114, peso=3) + 1 panel login (TP=1, peso=3) = 345 puntos positivos.
- Riesgo MEDIO: 2 puertos abiertos (TP=1, peso=2) + plugins WordPress (TP=1, peso=2) = 4 puntos positivos.
- Riesgo BAJO: WAF detectado (FP=1, peso=1) = 1 punto negativo.
- En total, se han obtenido 349 puntos + un punto por el falso positivo.

$$\text{Precisión ponderada} = \frac{349}{349+1} = 99{,}7\,\%$$

$$\text{Tasa de recuperación ponderada} = \frac{349}{349} = 100\,\%$$

De la evaluación ponderada, considerando los elementos encontrados versus los reportados en el informe, se ha calculado una precisión general del 99.7 %, con la identificación de un único falso positivo correspondiente a la detección del WAF. La tasa de recuperación es del 100 %, indicando que el sistema detectó todas las vulnerabilidades reales presentes detectadas por los escáneres. Según los umbrales establecidos anteriormente, ambas métricas se caracterizan por una calidad de categoría alta.

Tabla 3

Tabla Resumen Métricas ponderadas Escaneo 1.

ESCENARIO 1: Análisis Integral					
ESCENARIO	∑(TP × PESO)	∑(FP × PESO)	∑(FN × PESO)	PRECISIÓN PONDERADA (%)	RECALL PONDERADO (%)
Escenario 1	349	1	0	99.7%	100.0%

Nota. Elaboración propia.

6.4.3. Escaneo 2

En el segundo escaneo, se ha evaluado la capacidad del sistema para procesar consultas específicas para poder cuantificar la precisión del filtrado de información y la capacidad del sistema para mantener coherencia cuando se trabaja con subconjuntos específicos de datos. A constinuación se listan los siguientes parámetros del escaneo:

1. Nombre Empresa: Empresa X
2. Correo Empresa: @empresa.es
3. Dominio Empresa: https://empresa.es
4. Query: Qué información ha sido reportada por Nuclei?

6.4.4. Resultados informe Escaneo 2

El sistema generó un reporte focalizado exclusivamente en los hallazgos provenientes del escáner Nuclei, como se puede ver también por el resumen ejecutivo de la imagen, cumpliendo con la consulta específica del usuario.

Figura 9

Resumen ejecutivo Escaneo 2.

Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

El reporte identificó correctamente las desactualizaciones de plugins WordPress, las cuales, al hacer referencia a la misma vulnerabilidad, son indicados en la tabla como un único hallazgo y son clasificándas con nivel de riesgo de categoría MEDIA. La presencia del WAF, también ha sido categorizado con riesgo de categoría BAJA. La presencia del WAF es un falso negativo que se repite en ambos informes y no representa de por sí una vulnerabilidad, sino en caso de malas configuraciones, puede representar un elemento vulnerable. La estructura del informe mantuvo la calidad observada en el primer escenario, con recomendaciones específicas y un plan de mitigación dirigido únicamente a los hallazgos de Nuclei.

Figura 10

Análisis de Riesgos y Hallazgos Escaneo 2.

2. ANÁLISIS DE RIESGOS

Nivel de Riesgo	Hallazgo	Descripción
Medio	Outdated Plugin: All-In-One Security (AIOS) – Security and Firewall Detection	Se detectó la versión 5.3.3 (la última es la 5.4.0). Los plugins desactualizados pueden contener vulnerabilidades conocidas.
Medio	Outdated Plugin: Contact Form 7 Database Addon – CFDB7	Se detectó la versión 1.2.6.7 (la última es la 1.3.0). Los plugins desactualizados pueden contener vulnerabilidades conocidas.
Medio	Outdated Plugin: Contact Form 7	Se detectó la versión 5.1.9 (la última es la 6.0.6). Los plugins desactualizados pueden contener vulnerabilidades conocidas.
Medio	Outdated Plugin: Custom CSS and JS	Se detectó la versión 3.44 (la última es la 3.50). Los plugins desactualizados pueden contener vulnerabilidades conocidas.
Medio	Outdated Plugin: Custom Post Type UI	Se detectó la versión 1.5.8 (la última es la 1.17.3). Los plugins desactualizados pueden contener vulnerabilidades conocidas.
Bajo	WAF Detection	Se detecto un Web Application Firewall (WAF). Si bien esto es positivo, es importante mantener las reglas del WAF actualizadas para proteger contra las últimas amenazas.

3. HALLAZGOS TÉCNICOS DETALLADOS

3.1 Plugins Desactualizados

- **CRITICIDAD**: MEDIA
- **DESCRIPCIÓN**: Se identificaron varios plugins desactualizados en el sitio web www[redacted]es. Específicamente:
  - All-In-One Security (AIOS) – Security and Firewall Detection: Version 5.3.3 detected (latest: 5.4.0).
  - Contact Form 7 Database Addon – CFDB7: Version 1.2.6.7 detected (latest: 1.3.0).
  - Contact Form 7: Version 5.1.9 detected (latest: 6.0.6).
  - Custom CSS and JS: Version 3.44 detected (latest: 3.50).
  - Custom Post Type UI: Version 1.5.8 detected (latest: 1.17.3).
- **IMPACTO**: Los plugins desactualizados pueden contener vulnerabilidades conocidas que podrían ser explotadas por atacantes para comprometer el sitio web. Esto podría resultar en la pérdida de datos, la defacement del sitio o la inyección de malware.
- **VECTOR DE ATAQUE**: Un atacante podría aprovechar vulnerabilidades conocidas en las versiones antiguas de los plugins para ejecutar código malicioso en el servidor.
- **RECOMENDACIÓN**: Actualizar todos los plugins a sus versiones más recientes. Realizar pruebas en un entorno de staging antes de aplicar las actualizaciones en el entorno de producción.
- **REFERENCIAS**: OWASP Top 10, CWE-693: Protection Mechanism Failure
- **FUENTE**: nuclei_report.txt

3.2 Detección de WAF

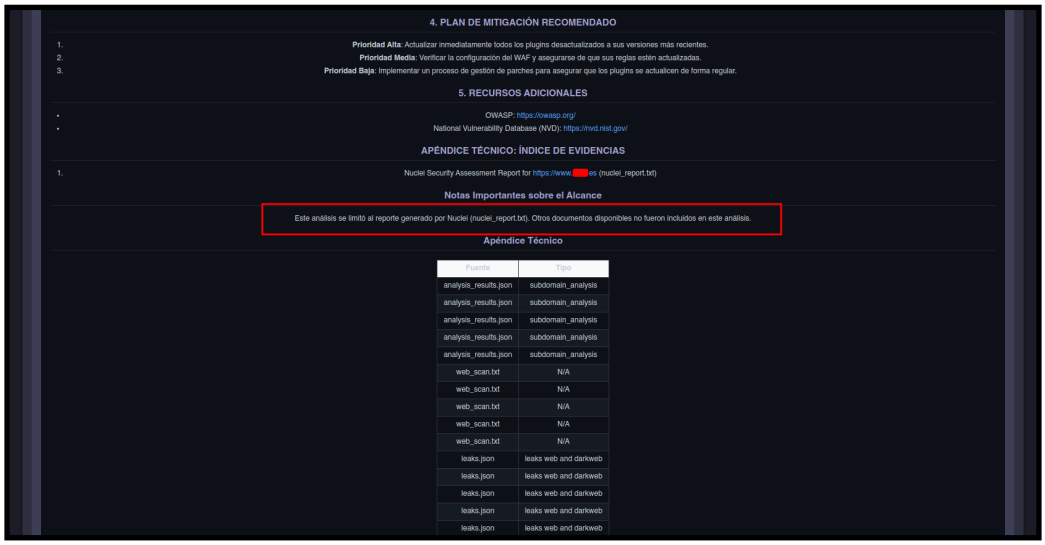
- **CRITICIDAD**: BAJA
- **DESCRIPCIÓN**: Se detectó un Web Application Firewall (WAF) protegiendo el sitio web.
- **IMPACTO**: La presencia de un WAF ayuda a mitigar ataques web comunes, como inyecciones SQL y cross-site scripting (XSS).
- **VECTOR DE ATAQUE**: N/A, el WAF es una medida de protección.
- **RECOMENDACIÓN**: Asegurarse de que el WAF esté configurado correctamente y que sus reglas estén actualizadas para proteger contra las últimas amenazas. Monitorear los logs del WAF para detectar y responder a posibles ataques.
- **REFERENCIAS**: OWASP Top 10
- **FUENTE**: nuclei_report.txt

Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

En el recuadro rojo de la imagen se indica el nombre del fichero utilizado para el análisis de vulnerabilidades (nuclei_scan.txt), lo que remarca la confiabilidad del informe generado en cuanto a precisión y tasa de recuperación ya que se cumple con los requisitos de la consulta. El plan de mitigación también se basa sobre los hallazgos definidos por Nuclei.

Figura 11

Plan de Mitigación Escaneo 2.



Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

Valoración precisión del modelo

Los resultados de este nuevo escaneo, muestran que de los 2 hallazgos encontrados por Nuclei, ambos fueron correctamente encontrados, pero uno de ellos (el WAF), como mencionado anteriormente, representa un falso positivo.

Tabla 4

Tabla Cálculo de puntos de Hallazgos Escaneo 2.

CATEGORÍA DE RIESGO	TIPO DE HALLAZGO	CANTIDAD	PESO	CLASIFICACIÓN	PUNTOS PONDERADOS
MEDIO	Plugins WordPress	1	2	TP (Verdadero Positivo)	2
BAJO	WAF Detectado	1	1	FP (Falso Positivo)	1

Nota. Elaboración propia.

Mediante la aplicación de los cálculos ponderados por categoría de riesgo a los hallazgos, como en el primer escaneo, se han obtenido los siguientes resultados:

- Riesgo MEDIO: Plugins WordPress desactualizados (TP = 1, peso = 2) = 2 puntos positivos.
- Riesgo BAJO: WAF detectado (FP = 1, peso = 1) = 1 punto negativo.
- En total, se han obtenido 2 puntos + 1 punto por el falso negativo.

$$\text{Precisión ponderada} = \frac{2}{2+1} = 66{,}7\,\%, \quad \text{Tasa de recuperación ponderada} = \frac{2}{2} = 100\,\%$$

Según los valores de las métricas calculadas (Precisión = 66.7 %, Tasa de recuperación = 100 %), se puede parecer que la consulta específica parece haber reducido la presión respecto a la consulta general. Todavía, hay que considerar que han habido menos hallazgos que recuperar y que el falso positivo generado se ha reportado en ambos casos. Dicho en otras palabras, ya que la pregunta, al ser específica, requiere al modelo una recuperación de datos menores respecto a la general, lo que disminuye la presión sobre los límites de tokens del modelo generativo, tanto en términos de input (número de chunks utilizados como contexto) como de output (extensión de la respuesta generada). Por este motivo, en caso de un mayor número de hallazgos, la precisión podría aumentar en vez de disminuir como en este caso.

En cuanto a calidad general del informe, como para el escenario anterior, se ha configurado un límite de 1000 palabras. Se han generado correctamente en ambos escaneos, y cada informe de seguridad se compone correctamente de los apartados correspondientes,

definidos en el programa. Además, se citan correctamente los documentos relevantes para la creación del informe.

Tabla 5

Tabla Resumen Métricas ponderadas Escaneo 2.

ESCENARIO 2: Consulta Específica					
ESCENARIO	∑(TP × PESO)	∑(FP × PESO)	∑(FN × PESO)	PRECISIÓN PONDERADA (%)	RECALL PONDERADO (%)
Escenario 2	2	1	0	66.7%	100.0%

Nota. Elaboración propia.

6.5. Análisis Comparativo de los escaneos

La comparación entre ambos escenarios revela diferencias significativas en cuanto a las métricas calculadas.

Análisis con Métricas Ponderadas:

La evaluación ponderada proporciona una perspectiva más realista del impacto:

- Escenario 1: Precisión ponderada 99.7 %
- Escenario 2: Precisión ponderada 66.7 %

La diferencia principal entre los valores de las métricas ponderadas calculadas, es la precisión. En el primer escenario, las 114 filtraciones críticas detectadas compensan el único falso positivo de bajo impacto. En el segundo escenario, el falso positivo tiene mayor impacto, aunque categorizado como riesgo de categoría BAJA, ya que la cantidad de hallazgos recuperados es significativamente menor respecto al primer caso.

La diferencia en precisión entre ambos escenarios no indica necesariamente una degradación del sistema, sino más bien una sensibilidad a la clasificación de elementos técnicos que no constituyen vulnerabilidades directas. En ambos casos, el elemento clasificado como falso positivo fue la detección del WAF, lo que indica una necesidad de refinamiento los criterios de clasificación para distinguir entre hallazgos informativos y vulnerabilidades actuales. Este problema se podría solucionar, mediante una mejora del tratamiento de los resultados de Nuclei, antes de ser guardados en el fichero correspondiente.

6.6. Evaluación de Calidad de Informes

En términos de calidad de informes, ambos escenarios produjeron documentos estructurados según indicado en el flujo del programa. Se generaron informes coherentes con los resultados de proporcionados durante la investigación, así como análisis técnicos y una clasificación apropiada.

6.7. Conclusiones sobre la Eficacia del Sistema

La implementación métricas ponderadas ha permitido dar una versión más real de los resultados. Según las métricas ponderadas calculadas, se puede decir que el sistema es efectivo en la detección de vulnerabilidades de alto impacto, que representan el mayor riesgo para las organizaciones.

Todavía, hay limitaciones inherentes a la clasificación de elementos informativos (como WAF) como vulnerabilidades, los cuales tienen bajo peso en el análisis ponderado y representan un impacto mínimo en la efectividad general del sistema.

En fin, la plataforma desarrollada parece ser efectiva para la detección de vulnerabilidades críticas y los informes mantienen la calidad y la precisión esperadas.

7. Conclusiones

El desarrollo de esta aplicación, CorpCrawler, la cual combina capacidades OSINT, análisis web y escaneo de red con procesamiento mediante inteligencia artificial y recuperación de la información almacenada, ha permitido obtener las siguientes conclusiones, tanto desde el punto de vista técnico como metodológico.

7.1. Principales Hallazgos

El presente trabajo ha demostrado la eficacia de integrar múltiples herramientas de escaneo de seguridad bajo una única arquitectura unificada. La implementación del sistema RAG (Retrieval-Augmented Generation) ha supuesto una mejora significativa en cuanto a interpretación de los resultados de la investigación en términos tanto de calidad, como de tiempo, gracias a su capacidad de generar un resumen contextualizado en unos pocos segundos y reduciendo considerablemente las imprecisiones de comunes LLM.

Los resultados obtenidos durante las dos pruebas realizadas evidencian los siguientes hallazgos:

1. La combinación de diferentes fuentes de información o dominios (OSINT, web y red), proporcionadas por escáneres unificados en una única arquitectura, proporciona una amplia visión del perfil de seguridad de la empresa bajo investigación.

2. El enfoque RAG para el análisis de resultados mejora significativamente la precisión y relevancia de los hallazgos en comparación con los métodos tradicionales o el uso directo de modelos de lenguaje sin aumentación.

3. Se requiere menos conocimiento técnico a la hora de analizar los resultados, ya que en el reporte se proporciona un análisis completo de todos los aspectos considerados relevantes.

4. La automatización del proceso completo de análisis de seguridad, desde la recopilación inicial de información hasta la generación de informes, reduce el tiempo necesario para obtener una evaluación comprehensiva y permite a los analistas de seguridad centrarse en tareas de mayor valor añadido.

7.2. Implicaciones Prácticas

Las implicaciones prácticas de este proyecto tienen relevancia tanto para expertos o juniors en el sector, como para empresas:

7.2.1. Para Profesionales de Seguridad

Se optimiza el tiempo, ya que la plataforma permite automatizar tareas de investigación, estandariza los procesos y las metodologías de la inspección, liberando tiempo para que los expertos encargados, puedan ocuparse de tareas más valiosas y que requieren un esfuerzo manual.

No se necesita conocimiento teórico específico para entender de cada elemento del informe, sino que es la misma herramienta que proporciona un análisis contextualizado mediante IA y permite aun teniendo menos experiencia, a novatos obtener más conocimiento rápidamente gracias a los insights valiosos y recomendaciones presentes en el reporte.

En fin, el enfoque basado en datos y aumentado con contexto reduce la influencia de sesgos personales en la interpretación de resultados.

7.2.2. Para Organizaciones

En cuanto a organizaciones, la automatización y eficiencia del sistema posibilitan la realización de análisis de seguridad con mayor regularidad, ya que la herramienta proporciona una amplia visión de la situación real en tema de seguridad de la empresa, facilitando la implementación de estrategias de seguridad continua.

La herramienta puede representar para la empresa una fuente de conocimiento propio, ya puede no ser consciente de la exposición de informaciones en internet.

7.2.3. Para el Ecosistema de Seguridad

Como evidenciado en los casos prácticos descritos en el presente trabajo, se ha demostrado la relevancia que puede tener la herramienta en entornos reales, gracias a la implementación de inteligencia artificial aplicada al campo de la ciberseguridad.

7.3. Limitaciones del Trabajo

A pesar de que los resultados obtenidos sean relevantes, es importante reconocer las limitaciones identificadas durante el desarrollo y evaluación del sistema que se indican en los apartados siguientes.

7.3.1. Limitaciones Técnicas

Se han rescontrado las siguientes limitaciones técnicas durante la definición de la plataforma y su programación:

- El rendimiento global del sistema está condicionado por las capacidades y limitaciones de las herramientas integradas, tanto externas como de elaboración propia, algunas de las cuales pueden presentar falsos positivos o negativos.

- Los escaneos han sido ejecutados en respeto de las actividades diarias de la empresa escaneada. Se ha intentado ralentizar los escaneos para no causar posibles interrupciones de los servicios y evitar bloqueos por parte de herramientas de seguridad corporativas como WAF. En algunos casos se ha omitido el uso de algunas herramientas.
- Cada herramienta, en fase prueba, ha sido testeada precedentemente en entornos controlados y no reales.
- El procesamiento simultáneo de múltiples herramientas y el análisis mediante modelos de inteligencia artificial requiere recursos significativos, lo que puede ha sido limitado por restricciones de hardware.

7.3.2. Limitaciones Metodológicas

Se han experimentado las siguientes limitaciones metodológicas:

- Aunque el sistema integra múltiples fuentes de información, no puede garantizar una cobertura exhaustiva de todas las posibles vulnerabilidades o vectores de ataque.
- Conocimiento limitado en cuanto a inteligencia artificial y RAG. La fase de estudio de estos dos componentes ha sido larga y llena de dificultades.
- Cambios continuos de las formas de implementación de LLM a nivel de código. Anteriormente se implementaba Huggingface para integrar modelos de inteligencia artificial. Pero, con el pasar de los meses se han rescontrados problemas en la ejecución del programa debido a cambios en permisos y formas de integración de los LLM, lo que ha llevado a optar por la integración con los modelos de Gemini, más estable y amplia disponibilidad de modelos que disponían de tokens gratuitos. Cada vez que se agota el número de tokens disponibles, es posible intercambiar fácilmente el modelo en uso con otro adapto.

7.3.3. Limitaciones de Alcance

La ejecución de las investigaciones de seguridad estaba sujeta a las siguientes limitaciones:

- Técnicas intrusivas de escaneo, en algunos casos, han sido deshabilitadas en cuanto podían haber causado alguna interrupción de los servicios corporativos además de consecuencias legales.
- Se ha buscado, para la realización de las pruebas, empresas que participaran a programas de Bug Bounty o en caso contrario, se ha optado para la deshabilitación de escaneos intrusivos y el análisis se ha basado sobre informaciones públicas.

7.4. Líneas Futuras de Investigación

Una de las posibles implementaciones a este proyecto, es la integración de un modelo de inteligencia artificial específico para ciberseguridad, con conocimiento especializado en vulnerabilidades, técnicas de ataque y estrategias de mitigación, el cual no ha sido posible por limitaciones de tiempo de entrenamiento y hardware. Al implementar este modelo, no se dependería tanto de grandes empresas como es el caso de Google y sus modelos Gemini u Openai con sus versiones de Chatgpt, lo que sería significativo en tema de privacidad. Hasta la fecha, la herramienta integra, en su arquitectura, los modelos de Gemini, lo que significa que toda información encontrada alimenta su entrenamiento. Aunque es discutible el hecho de que siendo información indexada en Google no haya sido ya utilizada para el entrenamiento, pero que por motivos de privacidad y éticos no pueda ser reflejada en las plataformas correspondientes.

La aplicación se podría componer en futuro por otras herramientas de seguridad ofensiva automatizada, como por ejemplo escáneres de Api o análisis de perfiles de redes sociales de los empleados. Estas nuevas herramientas ampliarían tanto la complejidad de la investigación, como la base de conocimiento de seguridad sobre la empresa, ayudándola a protegerse gracias al plan de acción recomendado.

Por último, se pueden mejorar los mecanismos para explicar las conclusiones del sistema y el mismo sistema de recuperación de la información. Esto está directamente conectado con el desarrollo de un modelo propio y específico, ya que los LLM están limitados por tokens tanto en input como en output, dificultando así la elección de los parámetros del modelo.

7.5. Reflexión Final

Este proyecto tenía como objetivos, por un lado, implementar una plataforma automatizada de análisis de seguridad, y por otro, averiguar el potencial de la inteligencia artificial aplicada a la ciberseguridad.

La plataforma permite a las empresas obtener una visión detallada de su exposición en internet y del estado de su seguridad, mediante una combinación de escaneo web, escaneo de red y recopilación de inteligencia de fuentes abiertas. La integración de la arquitectura RAG ha sido clave para contextualizar los resultados, reduciendo las “alucinaciones” típicas de los LLM, permitiendo la generación de informes más precisos. Gracias a la implementación de esta arquitectura se ha permitido generar valor añadido a la plataforma, ya que los informes generados pueden ayudar empresas y profesionales en reducir la superficie de ataque.

Los resultados obtenidos en los escenarios presentados se conforman con un alto nivel de calidad según los rangos definidos anteriormente, lo que garantiza una alta confianza en las recomendaciones proporcionadas. La validación empírica ha demostrado que la combinación de técnicas de escaneo automatizado con análisis inteligente mediante IA puede

identificar vulnerabilidades y riesgos de seguridad que podrían pasar desapercibidos en evaluaciones manuales tradicionales, sobre todo cuando el volumen de datos es mucho más grande.

Personalmente, puedo decir que este proyecto ha representado una oportunidad para explorar la intersección entre la inteligencia artificial y la ciberseguridad, dos campos en constante evolución y relacionados por el continuo nacimiento de herramientas y exploits. El desarrollo de la plataforma ha permitido extender mis conocimientos en ambos campos, y comprender como se puede aprovechar la intersección de los dos para intentar ayudar las empresas en el proceso de segurización. La automatización y la inteligencia artificial, han sido componentes claves del proyecto y han demostrado como pueden reforzar y agilizar el trabajo de los profesionales especialmente en un ámbito tan crítico como la ciberseguridad, donde las decisiones incorrectas pueden tener consecuencias significativas.

8. Anexos

8.1. Anexo I

8.1.1. Whatweb

WhatWeb identifica sitios web. Su objetivo es responder a la pregunta: "¿Qué es ese sitio web?". WhatWeb reconoce tecnologías web como sistemas de gestión de contenido (CMS), plataformas de blogs, paquetes de análisis estadísticos, bibliotecas JavaScript, servidores web y dispositivos integrados. WhatWeb tiene más de 1800 complementos, cada uno para reconocer algo diferente. WhatWeb también identifica números de versión, direcciones de correo electrónico, ID de cuentas, módulos de frameworks web, errores SQL y más.

WhatWeb puede ser sigiloso y rápido, o exhaustivo pero lento. WhatWeb soporta un nivel de agresividad que controla el equilibrio entre velocidad y fiabilidad. Cuando visitas un sitio web en tu navegador, la transacción incluye muchas pistas sobre qué tecnologías web están impulsando ese sitio web. A veces, una sola visita a una página es suficiente para identificar un sitio web, pero cuando no es así, WhatWeb puede interrogar más a fondo el sitio. El nivel de agresividad predeterminado, llamado 'sigiloso', es el más rápido y solo requiere una solicitud HTTP al sitio web. Esto es adecuado para escanear sitios web públicos. Se desarrollaron modos más agresivos para ser utilizados en pruebas de penetración.

La mayoría de los complementos de WhatWeb son exhaustivos y reconocen una gama de señales, desde las más sutiles hasta las más obvias. Por ejemplo, la mayoría de los sitios web de WordPress pueden ser identificados por la etiqueta meta en el HTML, e.g. ", pero una minoría de los sitios WordPress elimina esta etiqueta identificadora, lo que no impide que WhatWeb lo identifique. El complemento de WordPress de WhatWeb tiene más de 15 pruebas, que incluyen la comprobación del favicon, archivos de instalación predeterminados, páginas de inicio de sesión y la verificación de `/wp-content/` dentro de los enlaces relativos. (Horton, 2021)[5]

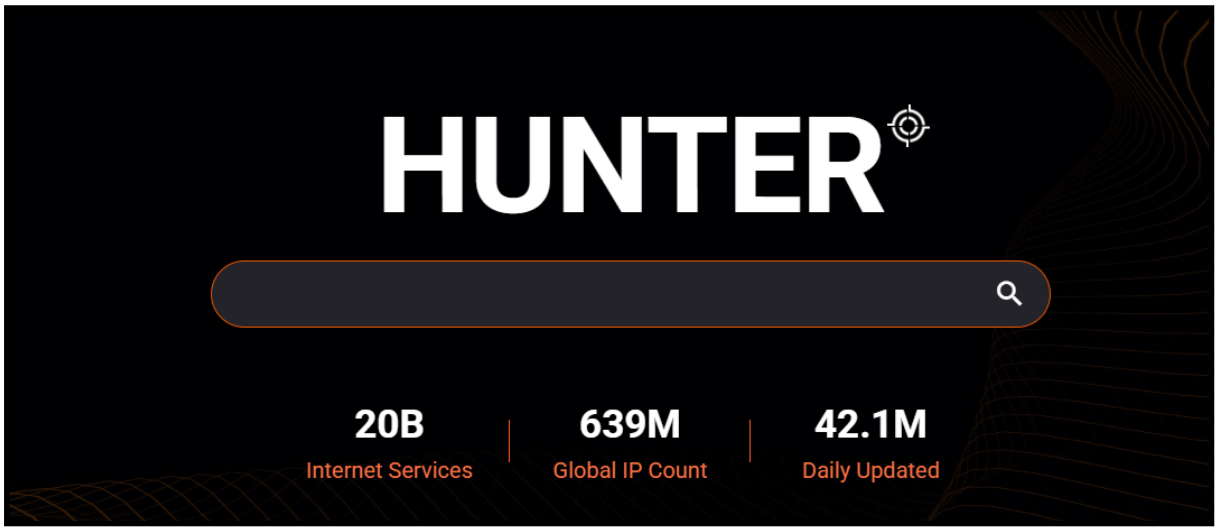
Características principales

- Más de 1800 complementos.
- Control sobre la relación entre velocidad/sigilo y fiabilidad.
- Ajuste del rendimiento: control sobre la cantidad de sitios web a escanear simultáneamente.
- Varios formatos de salida: Resumen (fácil de procesar), Verboso (legible por humanos), XML, JSON, MagicTree, RubyObject, MongoDB, ElasticSearch, SQL.
- Soporte de proxies, incluyendo TOR.
- Cabeceras HTTP personalizadas.
- Autenticación HTTP básica.
- Control sobre la redirección de páginas web.
- Rango de direcciones IP.
- Coincidencia difusa.
- Conciencia sobre la certeza de los resultados.
- Complementos personalizados definidos desde la línea de comandos.
- Soporte para Nombres de Dominio Internacionales (IDN).

8.1.2. HunterHow

Figura 12

HunterHow.



Nota. Fuente: (Hunter Search Engine, 2025)[6]

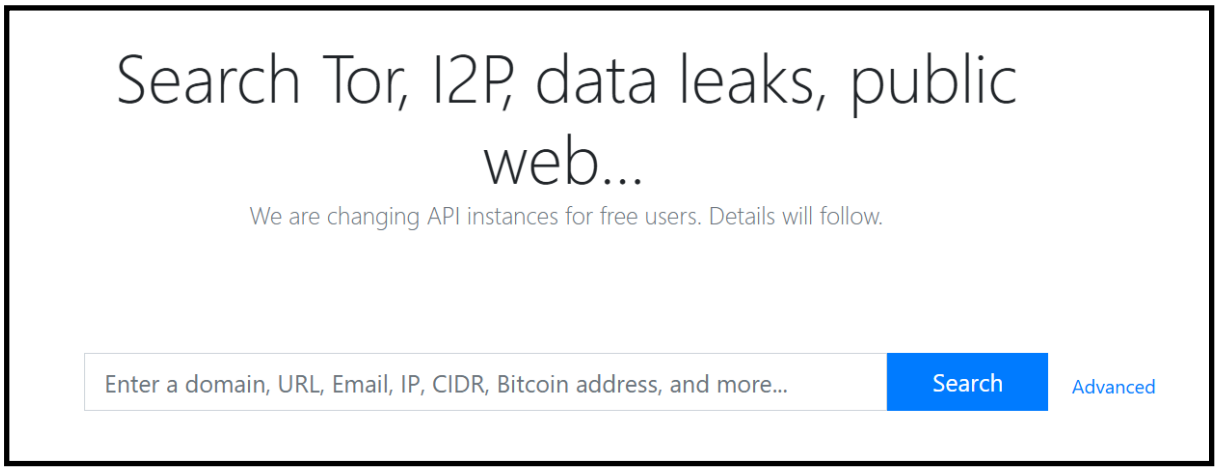
Hunter.how es un motor de búsqueda gratuito para investigadores de Internet. Ofrece una amplia cobertura de los servicios globales de Internet que abren servicios externos, con actualizaciones diarias de más de 40 millones de datos. Permite realizar búsquedas específicas con varios filtros (IP/Dominio/Puerto/Producto) y soporta el puerto 65535.

Este motor de búsqueda recién lanzado es capaz de realizar la recuperación de huellas dactilares de dispositivos y servicios conectados a Internet. Puede ayudar a encontrar diferentes cosas, como computadoras que ejecutan un determinado software (por ejemplo, Nginx), qué versión de Tomcat es la más popular, cuántos servidores FTP anónimos existen, o cuántos hosts se ven afectados por una nueva vulnerabilidad. (In-text citation: (Hunter Search Engine, 2025)[7]

8.1.3. IntelX

Figura 13

IntelX.



Nota. Fuente: (Intelligence X, n.d.)[8]

IntelX es una plataforma avanzada de búsqueda y análisis de información que permite a los usuarios acceder a fuentes de datos filtradas, incluyendo una variedad de fuentes de la deep web y la dark web. Esta herramienta se especializa en la recuperación de información relevante que normalmente no está disponible a través de búsquedas tradicionales en la web.

Funcionalidades principales

- **Búsqueda filtrada:** permite buscar información filtrada, lo que significa que los usuarios pueden acceder a contenido específico que no es fácilmente accesible a través de motores de búsqueda convencionales.
- **Exploración de la dark web:** se incluyen capacidades para explorar la dark web, proporcionando a los usuarios acceso a información que de otro modo permanecería oculta.
- **Algoritmos avanzados de búsqueda:** se implementan algoritmos para rastrear información relevante en diversas fuentes de datos y ofrecer un análisis detallado de los resultados obtenidos.
- **Análisis de resultados:** Una vez que se obtiene la información, IntelX ofrece herramientas de análisis para organizar y examinar los resultados, lo que facilita la extracción de datos relevantes para los usuarios.

8.1.4. Sublist3r

Es una herramienta desarrollada en python especifica para el descubrimiento de subdominios a partir de un determinado dominio. Los dominios son enumerados mediante requests hechas a través de diferentes motores de búsqueda como Google, Yahoo, Bing e implementado en sus procesos el uso de herramientas como Netcraft, Virustotal, ThreatCrowd, DNSdumpster y ReverseDNS para mejorar la búsqueda.(Aboul-Ela, 2021) [1]

8.1.5. Google Dorking

El algoritmo está diseñado para realizar búsquedas automatizadas en los motores de búsqueda Google y Bing y YAhoo, utilizando dorks específicos para localizar información relevante. A continuación se describe su funcionamiento:

- **URLs Base:** Se definen una URL base, como la siguiente, para cada motor de búsqueda a partir de las cuales se añadirán dorks:
  - `Google`: `https://www.google.com/search?q=`
- **Guardar Resultados:** Los resultados de las búsquedas se almacenan en un archivo JSON en el directorio con el nombre de la empresa.
- **Generar Consultas (Dorks):** Se ha definido una función que genera las consultas de búsqueda (dorks) basadas en un correo electrónico y el motor de búsqueda seleccionado. Estas consultas se componen de términos relacionados con contraseñas, credenciales y otro tipo de información sensible que puede ser recuperada a través de búsquedas avanzadas en los motores de búsqueda.
- **Realizar Búsqueda y Guardar Resultados:** Las búsquedas se llevan a cabo utilizando `Selenium` sin interfaz, el cual permite una navegación automática por las páginas de resultados, recogiendo la información mediante `BeautifulSoup` y guardándola en un archivo JSON.
- **Analizar Resultados de Búsqueda:** Además, se ha definido una función que encarga de analizar los resultados obtenidos en las búsquedas. Extrae información relevante de cada resultado, como:

- Título del resultado.
- Enlace o URL del resultado.

- Al final se genera también un archivo que contiene el top 5 de los dorks más relevantes.

8.1.6. Nmap

Figura 14

Nmap.

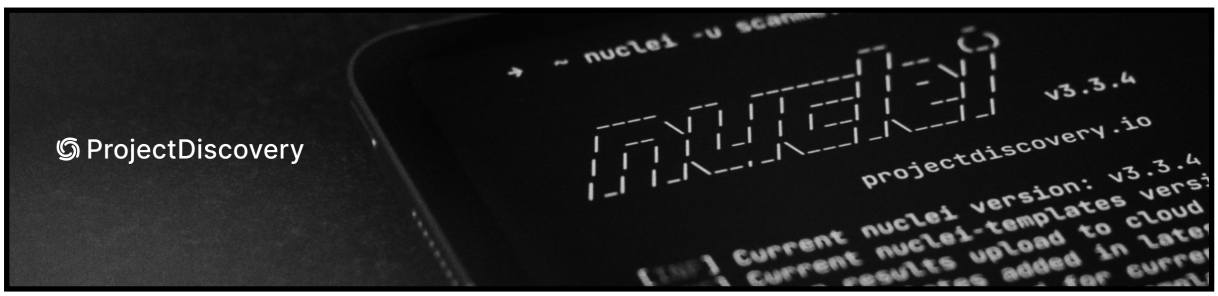
Nota. Fuente: (Instalación de NMAP En Ubuntu 22.04 | Digital_Educas, s.f.)[11]

Nmap, es una herramienta de seguridad que se implementa en auditorías de seguridad para poder efectuar un análisis de red. Permite descubrir tanto dispositivos conectados a la misma red, como enumerar puertos y servicios de una específica dirección IP.

8.1.7. Nuclei

Figura 15

Nuclei.



Nota. Fuente: (Projectdiscovery, s.f.-b)[12]

Nuclei es la herramienta open source implementada para el descubrimiento de vulnerabilidades de la página principal corporativa. La herramienta lleva a cabo procesos de descubrimiento de vulnerabilidades y aprovecha plantillas de detección basadas en el lenguaje de programación YAML, las cuales permiten clasificar y definir las vulnerabilidades encontradas. La herramienta permite efectuar miles de pruebas de vulnerabilidades y sus reportes son extremadamente precisos, lo que lo convierte en uno de los escáneres de seguridad open source de referencia.

8.2. Anexo II

8.2.1. Motor de procesamiento de datos

El motor de procesamiento de datos es la parte de algoritmo responsable de organizar, analizar y estructurar la información obtenida en el proceso de recolección de la investigación. Es el módulo que permite al agente IA utilizar los datos para generar reportes más precisos y de forma más eficaz.

El motor de procesamiento de datos se compone de estas etapas principales:

- Definición de funciones para el tratamiento de los datos
- Segmentación de los datos en chunks.
- Creación de embeddings.
- Almacenamiento de los datos en una base de datos vectorial (Chroma).

8.2.2. Configuración y fases del proceso

Se ha definido una clase `SecurityProcessor` que define toda la lógica principal del procesamiento de los ficheros generados por el motor de búsqueda del programa y la extracción de la información contenida en los archivos. Se ha definido, para cada tipo de extensión de archivo, un método diferente de extracción del contenido. Mediante los algoritmos de búsquedas, se generan solamente ficheros con extensión *Json* o *Txt*, por lo que en cada uno de los flujos de extracción de los datos, se definen controles basados en el nombre del archivo con el fin de mejorar y proporcionar un tratamiento y una extracción específicos basados en los tipos de datos definidos en cada uno de los ficheros.

El proceso de extracción y tratamiento se desarrolla en las siguientes fases:

- **Inicialización de Patrones de Seguridad**

  Al principio del fichero se inicializa un diccionario con expresiones regulares que permite identificar datos sensibles dentro de los archivos analizados, como por ejemplo direcciones IP, puertos, dominios, servidores, tecnologías y vulnerabilidades dentro de los documentos procesados.

- **Procesamiento de Archivos**

  En este punto del código, como mencionado anteriormente, se clasifican los archivos según su tipo y se redirige su procesamiento al método correspondiente.

- **Procesamiento de Archivos Json**

  En la siguiente función se analizan los archivos JSON en busca de información estructurada como listas de empleados, resultados de dorking y filtraciones de datos.

- **Ejemplo de procesamiento de fichero Json**

  Extracción de Información sobre los empleados encontrados mediante técnicas de dorking.

- **Procesamiento de Archivos txt**

- **Segmentación de Documentos** En esta fase, los datos contenidos en los documentos generados por las funciones de recolección de datos, se dividen en fragmentos más pequeños o *chunks* mediante el algoritmo `RecursiveCharacterTextSplitter` de LangChain. El tamaño de estos fragmentos ha sido elegido basándose en la largueza media del contenido de los archivos Para mejorar la indexación se almacena como índice el nombre de la carpeta y el nombre del texto, para que cada vez que se quiera analizar una empresa diferente, se inicie un proceso de recuperación o *retrieval* solo de las informaciones relativas a dicha empresa.

- **Generación de Embeddings**

  El proceso de generación de embeddings implementado, mediante un modelo específico Google Generative AI, ha permitido indexar los documentos en la base de datos vectorial. Este proceso permite convierte los chunks obtenidos anteriormente en representaciones numéricas multidimensionales que pueden ser comparadas de manera eficiente. El objetivo es permitir búsquedas semánticas sobre los documentos procesados, es decir, que no solo se puedan recuperar documentos mediante palabras clave exactas, sino también mediante términos relacionados o consultas con significado similar.

- **Almacenamiento de los embeddings en Croma DB**

  Una vez generados los embeddings, estos se almacenan en ChromaDB, base de datos diseñada específicamente para manejar datos vectoriales. Esta base de datos es específicamente estructurada para poder realizar consultas basadas en similitudes entre embeddings. Diferentemente de las bases de datos que almacenan datos estructurados y no estructurados (SQL y NOSQL), Chroma DB ChromaDB almacena representaciones numéricas de los documentos y permite hacer consultas en función de la dis-

tancia entre los vectores. Dicho en otras palabras, si dos vectores resultan ser muy similares en formato numérico, significa que contienen un contenido similar.

8.3. Agentes AI con arquitectura RAG

El código del agente es contenido en el fichero principal del programa, `app_d.py`. El agente ha sido programado para analizar los datos almacenados en la base de datos vectorial Chroma DB y proporcionar un análisis más preciso mediante el contexto que obtiene mediante la búsqueda en la base de datos y un prompt interno que indica como generar el informe. Además, es posible hacer consultas mediante el prompt de la interfaz gráfica para poder requerir un informe solo en base a un tipo de datos o integral.

El algoritmo del agente consta de las siguientes fases:

- **Definición de la plantilla del informe**
- **Extracción del contexto con ChromaDB**

  El contexto es clave para una generación del informe más precisa.

  Cada vez que un usuario realiza una consulta, el agente consulta la base de datos vectorial (ChromaDB), obteniendo resultados similares al texto de la consulta. Esta comparación es posible mediante la transformación del texto de la consulta en un vector y sucesivamente se recuperan los vectores similares.

- **Filtrado y clasificación del contenido**

  Para evitar incluir información irrelevante, se aplican criterios de filtrado basados en la media y la desviación estándar de las puntuaciones de similitud.

- **Generación del reporte**

  Para la generación del reporte se ha incluido un modelo de inteligencia artificial. La interacción con el modelo se verifica mediante la implementación de la api de Huggingface. El modelo en este caso recibe el prompt del usuario, el contexto y los datos para poder devolver un reporte en formato markdown.

- **Postprocesamiento del informe**

  Se realizan mejoras postproducción al contenido del informe para que resulte ser más claro y estructurado.

- **Ejemplo informe generado**

  Este es un ejemplo de informe generado por el agente basándose en el análisis realizado sobre la empresa analizada.

8.3.1. Interfaz gráfica

La interfaz gráfica consta de 3 componentes principales:

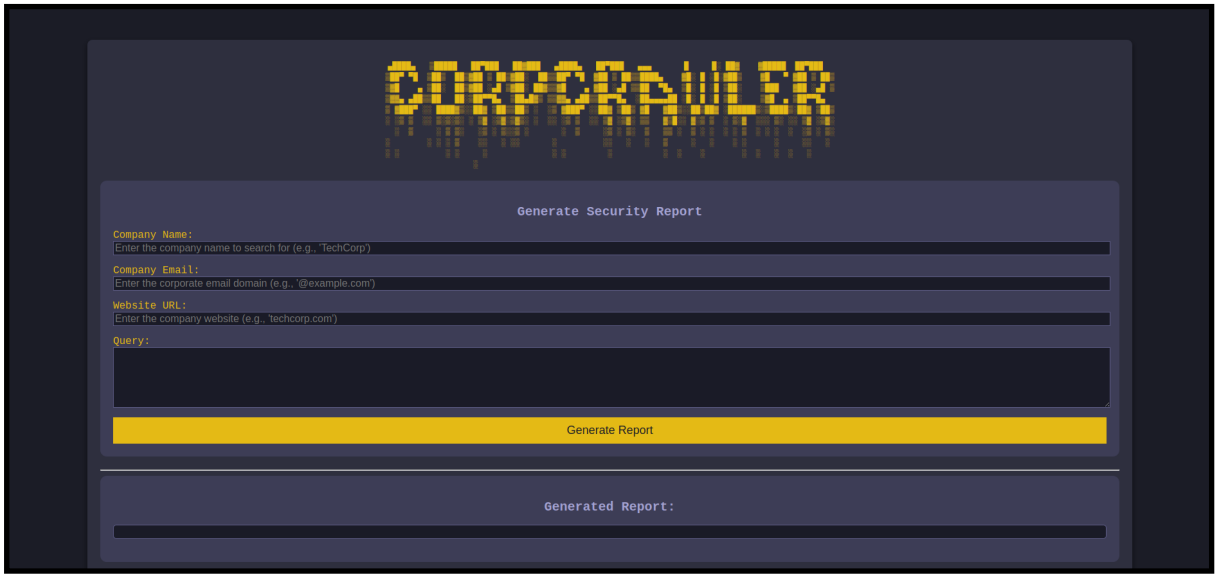
- Motor de la aplicación definido en el fichero `app_d.py`.
- Fichero que define el html de la interfaz (`templatesindex.html`).
- Fichero que define la estética de la interfaz (`staticstyle.css`).

8.3.2. Resultado de la interfaz

La web tiene un form principal con 3 valores de entradas para definir las variables de la empresa con un placeholder que define como tienen que ser introducidos los valores. Al envío de los datos se verifica que la empresa no haya sido ya analizada, para evitar cargar datos ya existentes en la base de datos. Si ha sido ya analizada se imprime el informe automáticamente, de lo contrario se inicializa todo el motor de búsqueda y se genera un reporte basado en los nuevos datos.

Figura 16

Interfaz de la plataforma.



Nota. Captura de pantalla de la interfaz de la plataforma desarrollada. Elaboración propia.

Referencias

[1] *Aboul-Ela, A. (2021, October 31). aboul3la/Sublist3r. GitHub.* https://github.com/aboul3la/Sublist3r

[2] AI RAG. (n.d.). *www.ibm.com*. https://www.ibm.com/architectures/hybrid/genai-rag

[3] Ferreira, A. C. (2023). *BDO España*. Inboundcycle.com. https://doi.org/1095489433/1741461324839

[4] Google. (2024). *What are AI hallucinations?* Google Cloud. https://cloud.google.com/discover/what-are-ai-hallucinations

[5] Horton, A. (2021, May 5). *urbanadventurer/WhatWeb*. https://github.com/urbanadventurer/WhatWeb

[6] Hunter Search Engine. (2025). *Hunter.how*. https://hunter.how/

[7] Hunter-How. (2022). *GitHub - Hunter-How/Support*. GitHub. https://github.com/Hunter-How/Support

[8] Intelligence X. (n.d.). *Intelx.io*. https://intelx.io/

[9] *IT Digital Media Group. (2023, 31 octubre). Las empresas invierten más de 220.000 millones de dólares en ciberseguridad. Actualidad | IT Digital Security.* https://www.itdigitalsecurity.es/actualidad/2023/10/las-empresas-invierten-mas-de-220000-millones-de-dolares-en-ciberseguridad/

[10] *Malwarebytes. (2024, 1 noviembre). Riesgos de IA y ciberseguridad | Riesgos de la inteligencia artificial.* https://www.malwarebytes.com/es/cybersecurity/basics/risks-of-ai-in-cyber-security

[11] *Instalación de NMAP en Ubuntu 22.04 | Digital_Educas. (s. f.).* https://digitaleducas.com/tutoriales/es/Instalacion_de_NMAP_en_Ubuntu_22.04

[12] *Projectdiscovery. (s. f.-b). GitHub - projectdiscovery/nuclei: Nuclei is a fast, customizable vulnerability scanner powered by the global security community and built on a simple YAML-based DSL, enabling collaboration to tackle trending vulnerabilities on the internet.*

It helps you find vulnerabilities in your applications, APIs, networks, DNS, and cloud configurations. GitHub.
https://github.com/projectdiscovery/nuclei

[13] ¿Qué es la generación aumentada de recuperación (RAG)? | Una guía completa de RAG. (s. f.). *Elastic*.
https://www.elastic.co/es/what-is/retrieval-augmented-generation

[14] *Los modelos de lenguaje de gran tamaño o LLM: qué son y cómo funcionan?* (n.d.). *www.redhat.com*.
https://www.redhat.com/es/topics/ai/what-are-large-language-models

[15] *Thales. (2025, 21 de mayo). El Informe de Amenazas a los Datos de Thales 2025 revela que casi el 70 % de las organizaciones identifican al rápido ecosistema de la IA como el principal riesgo de seguridad que está relacionado con la GenAI [Nota de prensa]. Business Wire. www.redhat.com*.
https://www.businesswire.com/news/home/20250520873994/es