Universidad Distrital Francisco José de Caldas
Facultad de Ingeniería

Herramienta de Diagnóstico Unificado para el Análisis de Seguridad y Geolocalización de Direcciones IP

Proyecto de Investigación

Investigadores:

Juan Manuel Serrano Rodríguez
Código: 20211020091
jumserranor@udistrital.edu.co

Nicolás Guevara Herrán
Código: 20211180026
nguevarah@udistrital.edu.co

Área de Investigación:
Ciberseguridad y Análisis de Redes

Grupo de Investigación:
Modelamiento en Ingeniería de Sistemas

Línea de Investigación:
Inteligencia de Amenazas Cibernéticas

September 21, 2025

1 Introducción

La evaluación de la legitimidad y seguridad de direcciones IP se ha convertido en una actividad fundamental para profesionales de ciberseguridad, administradores de sistemas y usuarios técnicos en el contexto actual de amenazas cibernéticas sofisticadas [1]. La creciente complejidad de los ataques cibernéticos y la fragmentación de información de seguridad en múltiples plataformas han impulsado el desarrollo de herramientas unificadas de diagnóstico que integren múltiples fuentes de datos para proporcionar evaluaciones comprehensivas de seguridad [2].

Este proyecto consolida, en un único flujo de consulta y presentación, información pasiva sobre direcciones IP usando exclusivamente fuentes gratuitas y de acceso público: (i) conjuntos de datos de Censys disponibles en Google BigQuery y (ii) la base GeoLite2 de MaxMind. No se usarán APIs con planes limitados ni servicios propietarios de pago.

1.1 Propósito

Aprovechar estas dos fuentes abiertas para producir una vista integrada que incluya, como mínimo: presencia observada en escaneos de Censys (puertos/servicios y certificados asociados) y geolocalización aproximada (país/ciudad si aplica) desde GeoLite2, presentadas de forma clara y reproducible.

1.2 Alcance

El trabajo comprende: (a) consulta y filtrado de datos en BigQuery (tablas públicas de Censys), (b) resolución local con GeoLite2, y (c) una plantilla de reporte que estandariza la lectura de resultados. No se crean nuevas fuentes ni métricas propietarias; se organiza y documenta lo existente.

1.3 Restricciones

- Solo se emplean fuentes gratuitas (BigQuery con datasets públicos de Censys y GeoLite2) - Geolocalización es aproximada y sujeta a las limitaciones de la base - No se ejecutan escaneos activos; se usa reconocimiento pasivo

1.4 Entregables

- Consultas de ejemplo en BigQuery sobre datasets de Censys - Procedimiento para resolución/localización con GeoLite2 - Estructura de reporte para presentar

hallazgos de manera consistente

2 Estado del Arte

2.1 Reconocimiento Pasivo en Ciberseguridad

2.1.1 Definición y Metodologías

El reconocimiento pasivo, definido como el proceso de recolección de información sin interactuar directamente con los sistemas objetivo, representa una metodología fundamental en el arsenal de herramientas de ciberseguridad [3]. A diferencia del reconocimiento activo, que implica el envío de consultas directas y puede ser detectado por sistemas de seguridad, el reconocimiento pasivo permite obtener información valiosa manteniendo un perfil bajo y minimizando el riesgo de detección [4].

2.1.2 Técnicas y Herramientas Modernas

Las técnicas modernas de reconocimiento pasivo han evolucionado significativamente con la integración de inteligencia artificial y aprendizaje automático. Los atacantes utilizan herramientas impulsadas por IA como Maltego, SpiderFoot y Recon-ng para automatizar la recolección de OSINT (Open-Source Intelligence), el análisis de actividades en redes sociales y la extracción de datos de bases de datos filtradas [17]. Esta evolución ha llevado a que el reconocimiento pasivo sea más eficiente y difícil de detectar, aumentando la necesidad de contramedidas defensivas más sofisticadas.

2.1.3 Metodologías de Detección

La identificación de actividades de reconocimiento presenta desafíos significativos, pero varios indicadores pueden sugerir que alguien está recopilando información para planificar un ciberataque. Estos incluyen tráfico de red inusual, patrones de acceso anómalos, indicadores de ingeniería social, uso de herramientas de reconocimiento y alertas de inteligencia de amenazas externas [5]. Los sistemas modernos de detección emplean análisis de comportamiento y aprendizaje automático para identificar estas actividades de manera más efectiva.

2.2 Geolocalización de Direcciones IP

2.2.1 Técnicas Tradicionales y Limitaciones

La geolocalización de direcciones IP, aunque inherentemente aproximada debido a las limitaciones técnicas de los métodos disponibles, proporciona información contextual valiosa para el análisis de seguridad [6]. Las técnicas tradicionales se basan principalmente en mediciones de latencia y análisis topológico, pero estas presentan limitaciones significativas en términos de precisión y cobertura [18].

2.2.2 Enfoques de Aprendizaje Automático

Los avances recientes en técnicas de geolocalización han demostrado mejoras significativas en la precisión, particularmente cuando se combinan múltiples fuentes de datos y se implementan algoritmos de machine learning para procesar la información recolectada [7]. Un estudio de 2024 propuso un marco novedoso que combina redes convolucionales de grafos (GCNs) y optimización de función de energía con muestreo de Monte Carlo para abordar los desafíos de incertidumbre en la geolocalización IP [8].

2.2.3 Limitaciones de Precisión en Bases de Datos Comerciales

La investigación sobre la precisión de bases de datos comerciales de geolocalización ha revelado discrepancias significativas entre las precisiones reportadas y las medidas empíricamente. Un estudio de 2021 sobre la base de datos GeoLite2 City de MaxMind encontró que las precisiones reportadas por MaxMind caían fuera de los intervalos de confianza del 99% en el 72% de los casos evaluados, con diferencias de precisión promedio de 4.2% para un radio de 10 km, 11.7% para 50 km y 11.1% para 250 km [9]. Estas limitaciones subrayan la importancia de evaluar críticamente la confiabilidad de las fuentes de datos de geolocalización.

2.3 Inteligencia de Amenazas Cibernéticas

2.3.1 Evolución de los Marcos de Trabajo

La inteligencia de amenazas cibernéticas ha evolucionado considerablemente en los últimos años, transitando desde análisis manuales hacia sistemas automatizados sofisticados impulsados por inteligencia artificial [10]. Esta evolución ha sido impulsada por la necesidad de procesar volúmenes masivos de datos en tiempo real y la complejidad creciente de los patrones de ataque [11].

2.3.2 Marcos de Trabajo Contemporáneos

Los marcos de trabajo modernos incluyen NIST Cybersecurity Framework 2.0, que en 2024 expandió su alcance más allá de la infraestructura crítica para incluir organizaciones de diversos tamaños y sectores [20]. MITRE ATT&CK continúa siendo fundamental para el mapeo de tácticas, técnicas y procedimientos adversarios, mientras que marcos complementarios como D3FEND y ENGAGE proporcionan enfoques defensivos y de compromiso respectivamente.

2.3.3 Integración Multi-fuente

Las plataformas modernas de inteligencia de amenazas integran múltiples fuentes de datos, utilizan algoritmos avanzados de análisis y proporcionan capacidades predictivas que permiten a las organizaciones adoptar una postura proactiva frente a las amenazas emergentes. Un estudio reciente propuso un marco mejorado de inteligencia de amenazas que integra flujos de datos diversos, incluyendo logs de red corporativa, inteligencia de código abierto y monitoreo de la web oscura [2].

2.4 Plataformas de Datos y Fuentes de Información

2.4.1 Censys y Google BigQuery

Censys publica datasets que capturan observaciones de servicios y certificados en Internet, y su disponibilidad en BigQuery permite consultas SQL reproducibles sobre grandes volúmenes de datos sin necesidad de claves de API ni dependencias de planes comerciales [12]. La arquitectura CQRS (Command Query Responsibility Segregation) de Censys permite el procesamiento independiente de operaciones de lectura y escritura, manejando aproximadamente 5 mil millones de eventos por día en 2024.

2.4.2 MaxMind GeoLite2 y Limitaciones

GeoLite2 ofrece bases gratuitas para geolocalización aproximada de direcciones IP y es una referencia común en proyectos abiertos y académicos cuando no se requiere la precisión y soporte de las versiones comerciales. Sin embargo, estudios independientes han cuestionado las precisiones reportadas, encontrando diferencias significativas entre las métricas publicitadas y las mediciones empíricas [9].

2.4.3 Datos WHOIS en Ciberseguridad

Las bases de datos WHOIS proporcionan información crítica sobre el registro y propiedad de dominios, siendo esenciales para la investigación de amenazas y la aplicación de la ley. El uso de WHOIS reverso permite a los expertos en ciberseguridad rastrear campañas maliciosas, identificar infraestructuras de atacantes y detectar patrones de registro sospechosos. La integración de datos WHOIS con plataformas de inteligencia de amenazas mejora significativamente las capacidades de hunting de amenazas y respuesta a incidentes.

2.5 Evaluación de Riesgo y Puntuación Algorítmica

2.5.1 Modelos de Puntuación de Riesgo

Los sistemas modernos de evaluación de riesgo implementan modelos algorítmicos que procesan múltiples vectores de datos para generar puntuaciones comprehensivas. Estos modelos consideran no solo la presencia de anomalías, sino también la velocidad y contexto en que se encuentran, permitiendo clasificaciones más matizadas de riesgo.

2.5.2 Integración con AbuseIPDB

AbuseIPDB proporciona una plataforma colaborativa donde la comunidad de ciberseguridad reporta direcciones IP involucradas en actividades maliciosas. La integración de AbuseIPDB con sistemas de evaluación de riesgo permite obtener una segunda opinión basada en reportes de la comunidad, con puntuaciones de abuso calculadas utilizando sistemas de pesos ponderados que previenen la manipulación por usuarios individuales [13].

2.6 Herramientas de Escaneo y Análisis Comparativo

2.6.1 Análisis Comparativo de Herramientas

Un análisis comparativo reciente de herramientas de enumeración de red reveló que cada herramienta posee características y funcionalidades únicas. OpenVAS es altamente valorado por sus capacidades comprehensivas de escaneo de vulnerabilidades, mientras que Zenmap y Nmap sobresalen en mapeo de red y escaneo de puertos [14]. Nessus se destaca por su extensa base de datos de vulnerabilidades y robustas capacidades de reporte.

2.6.2 Eficacia de Herramientas de Escaneo de Puertos

Un estudio experimental comparativo de 2023 que evaluó Nmap, Zmap y masscan no encontró diferencias en el rendimiento general, pero reveló diferencias estadísticamente significativas en eficiencia [15]. Estos resultados pueden guiar la selección de herramientas de escaneo de puertos basada en necesidades específicas y requisitos de rendimiento.

2.7 Aprendizaje Automático en Ciberseguridad

2.7.1 Aplicaciones en Detección de Amenazas

El aprendizaje automático transforma la ciberseguridad al permitir detección de amenazas más inteligente, respuesta a incidentes mejorada y asignación más eficiente de recursos. Los algoritmos de ML pueden analizar patrones históricos y adaptarse a nuevos patrones, mejorando tanto las medidas de seguridad preventivas como reactivas.

2.7.2 Tipos de Aprendizaje y Sus Aplicaciones

El aprendizaje supervisado utiliza conjuntos de datos etiquetados para entrenar modelos que detecten tipos específicos de amenazas, como correos de phishing o variantes conocidas de malware. El aprendizaje no supervisado identifica amenazas desconocidas analizando patrones en datos no etiquetados, siendo particularmente efectivo para detectar anomalías en actividad de red que podrían indicar ataques de día cero. El aprendizaje por refuerzo permite que los modelos aprendan y mejoren optimizando mecanismos de defensa basados en el éxito de acciones previas.

2.7.3 Beneficios y Desafíos

Los beneficios principales incluyen precisión mejorada en detección de amenazas, capacidades de análisis y respuesta en tiempo real, reducción de falsos positivos y negativos, y escalabilidad para manejar grandes volúmenes de datos. Sin embargo, los desafíos incluyen la necesidad de datos de entrenamiento de alta calidad, la interpretabilidad de modelos complejos y la adaptación a amenazas en evolución constante [16].

2.8 Marcos de Ciberseguridad Contemporáneos

2.8.1 NIST Cybersecurity Framework 2.0

El NIST Cybersecurity Framework 2.0, lanzado en 2024, representa la actualización más significativa desde la versión 1.1 de 2018. La nueva versión extiende su alcance más allá de la infraestructura crítica, dirigiéndose a una gama más amplia de organizaciones independientemente de su expertise en ciberseguridad. Una adición notable es el énfasis en governance de ciberseguridad, reconociendo la ciberseguridad como un componente clave de la gestión de riesgos empresariales [20].

2.8.2 Otros Marcos Relevantes

Los marcos MITRE (ATT&CK, D3FEND, ENGAGE) han ganado popularidad como herramientas comprehensivas para reconocer y combatir amenazas cibernéticas. Estos marcos proporcionan bases de conocimiento que cubren tácticas, técnicas y procedimientos cibernéticos, con actualizaciones consistentes para reflejar cambios en el panorama de amenazas [17].

2.9 Brechas y Oportunidades de Investigación

2.9.1 Limitaciones Actuales

A pesar de los avances significativos, persisten brechas importantes en el campo. La fragmentación de información entre múltiples servicios y modelos de acceso continúa siendo un desafío, particularmente para organizaciones con recursos limitados. La precisión de datos de geolocalización sigue siendo problemática, y la dependencia de fuentes comerciales introduce vulnerabilidades en términos de disponibilidad y costo [18].

2.9.2 Oportunidades Emergentes

Las oportunidades de investigación incluyen el desarrollo de algoritmos más sofisticados de fusión de datos, la mejora de técnicas de geolocalización mediante integración de múltiples metodologías, y el desarrollo de marcos de evaluación más robustos que consideren la incertidumbre inherente en las fuentes de datos. La integración de técnicas de inteligencia artificial explicable (XAI) podría mejorar la interpretabilidad de sistemas de evaluación de riesgo complejos [19].
Otra área promisoria es el aprendizaje federado y colaborativo. En lugar de depender de bases de datos centralizadas (que pueden ser costosas o introducir riesgos de privacidad), el aprendizaje federado permite a múltiples organizaciones entrenar

modelos conjuntos sin compartir datos sensibles entre sí. Aplicado a análisis de IP e inteligencia de amenazas, esto significa que varias entidades podrían contribuir con sus eventos de red para mejorar un modelo global de detección, sin revelar los datos brutos de sus usuarios. Este enfoque descentralizado preserva la privacidad y al mismo tiempo enriquece la diversidad de datos, produciendo modelos más robustos y menos sesgados frente a diferentes entornos [21]. Por ejemplo, un banco en Europa y otro en Latinoamérica podrían colaborar mediante aprendizaje federado para detectar fraudes vinculados a ciertas IP, beneficiándose mutuamente de las huellas de ataque vistas en cada región, pero sin intercambiar información de clientes.

2.10 Aspectos Legales en el Análisis de Direcciones IP

2.10.1 A Nivel Mundial

El análisis de direcciones IP en ciberseguridad se encuentra regulado por un marco diverso de normativas que convergen en un principio común: las direcciones IP son tratadas como datos personales cuando permiten identificar, directa o indirectamente, a un individuo. En Europa, el Regalmento General de Protección de Datos (RGPD) establece que las direcciones IP deben ser protegidas de la misma forma que otros datos sensibles, lo que implica obligaciones de transparencia, limitación en la finalidad, consentimiento y medidas de seguridad en su tratamiento [22]. En Estados Unidos, aunque no existe una ley federal integral, legislaciones estatales como la Ley de Privacidad del Consumidor de California (CCPA/CPRA) incluyen las IP dentro de la definición de información personal, siempre que puedan vincularse a un consumidor o hogar [23]. Otras jurisdicciones han seguido caminos similares, como la Ley General de Protección de Datos (LGPD) en Brasil y la Ley de Protección de Información Personal (PIPL) en China, lo que refleja una tendencia global hacia mayores restricciones en el uso y recolección de este tipo de datos [24].

2.10.2 En Colombia

En el caso colombiano, la protección de datos personales está enmarcada principalmente en la Ley Estatutaria 1581 de 2012, que regula la recolección, almacenamiento, uso y circulación de información personal. Bajo esta normativa, una dirección IP puede ser considerada un dato personal en la medida en que permita identificar a una persona, lo que activa obligaciones específicas para quienes la recolectan o procesan. La ley establece principios rectores como legalidad, finalidad, veracidad, transparencia, seguridad y confidencialidad, y reconoce a los titualres derechos como conocer, actualizar y rectificar los datos que se poseen sobre ellos [25].

De la misma forma, el Registro Nacional de Bases de Datos, administrado por la Superintendencia de Industria y Comercio (SIC), funciona como mecanismo de control y vigilancia del cumplimiento normativo, mientras que la misma SIC está facultada para imponer sanciones en casos de incumplimiento [26]. Complementariamente, la Ley 1621 de 2013, conocida como la Ley de Inteligencia y Contrainteligencia, obliga a operadores de telecomunicaciones a conservar información de usuarios y comunicaciones técnicas durante cinco años, lo cual plantea un delicado equilibrio entre las necesidades de seguridad nacional y la protección del derecho a la intimidad [27]. Más recientemente, se han planteado propuestas para actualizar la legislación nacional, incorporando mayores controles sobre algoritmos, sanciones más estrictas y nuevas restricciones a las transferencias internacionales de datos hacia países con menores estándares de protección. Estas disposiciones y debates muestran cómo Colombia, al igual que otras jurisdicciones, enfrenta el desafío de armonizar la defensa cibernética con los principios de privacidad y derechos fundamentales, especialmente en lo referente al análisis y geolocalización de direcciones IP.

3 Problemática

3.1 Necesidad y contexto

La información relevante para contextualizar una dirección IP existe, pero suele estar fragmentada entre múltiples servicios y modelos de acceso. Parte importante del ecosistema impone límites de uso por API o requiere suscripciones, lo que afecta reproducibilidad y acceso.

3.2 Criterio de selección de fuentes

Para garantizar gratuidad, acceso sostenible y replicabilidad, este proyecto se restringe a:

- **Censys en Google BigQuery**: conjuntos de datos públicos que permiten consultas reproducibles sobre observaciones de puertos/servicios y certificados sin depender de claves de API
- **MaxMind GeoLite2**: base gratuita ampliamente utilizada para geolocalización aproximada de IP

3.3 Limitaciones asumidas

- La geolocalización es aproximada (propia de GeoLite2)

- Las observaciones de Censys reflejan cortes/ventanas de tiempo de los datasets públicos
- Sin escaneo activo ni enriquecimientos propietarios

3.4 Objetivo práctico

Reducir la fricción operativa mediante una presentación integrada y documentada de estas dos fuentes, priorizando claridad, trazabilidad y facilidad de replicación sin costos.

References

[1] M. Barni et al., "Information Forensics and Security: A quarter-century-long journey," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2856–2893, 2024.

[2] M. Alazab et al., "Enhanced threat intelligence framework for advanced cybersecurity resilience," *Egyptian Informatics Journal*, vol. 27, art. 100521, 2024.

[3] A. S. Alqahtani, "Security threats and countermeasures in software defined network using efficient and secure trusted routing mechanism," *Computer Networks*, vol. 177, pp. 102–115, Mar. 2020.

[4] F. Alghamdi, "A Comprehensive Investigation of Reconnaissance Threats in Cybersecurity," *International Journal of Advanced Research in Management and Social Sciences*, vol. 13, no. 11, Nov. 2024.

[5] X. Qin et al., "A hybrid cyber defense framework for reconnaissance attacks in 5G networks," *Computers & Security*, vol. 136, art. 103569, 2024.

[6] Z. Dong et al., "Network measurement based modeling and optimization for IP geolocation," *Computer Networks*, vol. 56, no. 1, pp. 85–98, Jan. 2012.

[7] A. Hong et al., "A Cheap and Accurate Delay-Based IP Geolocation Method using Machine Learning and Looking Glass," in *Proc. IEEE INFOCOM Workshops*, 2023.

[8] X. Liu et al., "Robust IP geolocation through the lens of uncertainty quantification," *Computer Networks*, vol. 257, 2025.

[9] M. Schopman, "Validating the accuracy of the MaxMind GeoLite2 City database," Bachelor's thesis, Radboud University, 2021.

[10] M. A. Ferrag et al., "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Network and Computer Applications*, vol. 50, pp. 41–65, Feb. 2019.

[11] P.-C. Lin et al., "Correlation of cyber threat intelligence with sightings for intelligence assessment and augmentation," *Computer Networks*, vol. 229, art. 109751, Jun. 2023.

[12] Z. Durumeric et al., "Censys: A Map of Internet Hosts and Services," Technical Report, 2025.

[13] "A Multi-Functional Web Tool for Comprehensive Threat Intelligence and Risk Assessment," arXiv preprint arXiv:2412.03023, 2024.

[14] F. Alghamdi, "A Comparative Analysis of Network Enumeration Tools," *Journal of Computer Security and Cybercrime Management*, 2024.

[15] J. M. Pittman, "A Comparative Analysis of Port Scanning Tool Efficacy," arXiv preprint arXiv:2303.11282, 2023.

[16] "A comprehensive survey on concept drift and feature selection in machine learning for cybersecurity," *ScienceDirect*, 2024.

[17] W. Mazurczyk et al., "Cyber reconnaissance techniques," *Communications of the ACM*, vol. 64, no. 3, 2021.

[18] G. Ciavarrini et al., "Accuracy limits through Cramér–Rao lower bound for IP geolocation," *Computer Networks*, vol. 131, pp. 37–50, 2018.

[19] "A Machine Learning Approach for Detecting Cybersecurity Vulnerabilities," *Knowledge-Based Systems*, 2025.

[20] "NIST Cybersecurity Framework 2.0," NIST, 2024.

[21] Tripwire, "Federated Learning for Cybersecurity: Collaborative Intelligence for Threat Detection," Tripwire, 18 Mar. 2024.

[22] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the council (General Data Protection Regulation)," European Union, Apr. 2016.

[23] State of California, "California Consumer Privacy Act (CCPA), California Civil Code Title 1.81.5, as amended by California Privacy Rights Act (CPRA)," State of California, 2020.

[24] National People's Congress of China, "Personal Information Protection Law (PIPL)," China, Nov. 2021.

[25] Congreso de la República de Colombia, "Ley Estatutaria 1581 de 2012 - Protección de Datos Personales," Oct. 2012.

[26] Superintendencia de Industria y Comercio, "Registro Nacional de Bases de Datos - RNBD," 2025.

[27] Congreso de la República de Colombia, "Ley 1621 de 2013 - Ley de Inteligencia y Contrainteligencia," Apr. 2013.