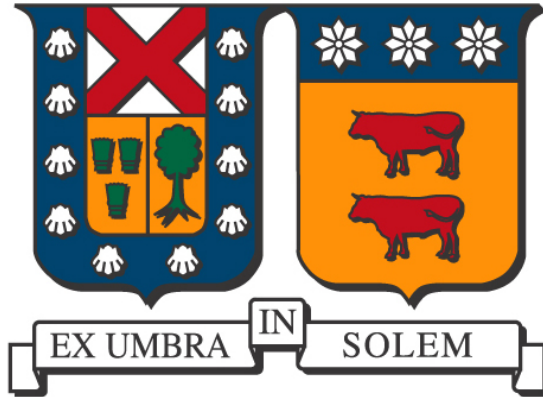


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
SANTIAGO - CHILE



“Sistema Recolector de Información para la Seguridad Ciudadana”

ESTEBAN ENRIQUE NARANJO RAYO

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
TELEMÁTICO

PROFESOR GUÍA:

NICOLÁS TORRES RUDLOFF

PROFESOR CORREFERENTE:

PATRICIO OLIVARES

NOVIEMBRE - 2025



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Sistema Recolector de Información para la Seguridad Ciudadana

Nombre del candidato(a): Esteban Enrique Naranjo Rayo

Carrera / Grado: Grado de Licenciado en Ciencias de la Ingeniería, título profesional de Ingeniero Civil Telemático

Campus: San Joaquín Departamento: Departamento de Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Nicolás Torres Rudloff, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (marcar una opción):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

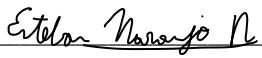
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 03/12/2025 Firma: 

Estudiante o Candidato(a):

Fecha: 03-12-2025 Firma: 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Agradecimientos

Me gustaría agradecer a mi profesor guía por su orientación y apoyo constante a lo largo de este proyecto. También agradezco a mis compañeros de equipo, Fernando Amthauer y Joaquín Iturra, por su colaboración y compromiso, que hicieron posible alcanzar los objetivos propuestos. Finalmente, un agradecimiento especial a mi familia por su paciencia y aliento incondicional durante todo este proceso.

Este documento está dedicado a todos aquellos que buscan tomar decisiones más informadas para construir comunidades más seguras, y a mi familia, por su infinito apoyo.

Resumen

La toma de decisiones en el ámbito inmobiliario es un proceso complejo influenciado por múltiples factores, donde la seguridad del entorno es un aspecto crucial pero difícil de cuantificar. Este proyecto aborda la necesidad de una herramienta que integre datos de seguridad ciudadana y del mercado inmobiliario para ofrecer una visión unificada y analítica. El presente trabajo se centra en el diseño y la implementación de la infraestructura de recolección de datos, un componente fundamental para el sistema global.

El objetivo principal fue construir un pipeline de datos robusto y automatizado, capaz de extraer, procesar y almacenar información de fuentes heterogéneas. Para ello, se desarrollaron módulos de web scraping especializados en la recolección de estadísticas delictuales de fuentes oficiales como Carabineros de Chile, datos de propiedades de los principales portales inmobiliarios del país y noticias relevantes de medios de comunicación.

La metodología se basó en la creación de un entorno de desarrollo containerizado con Docker, asegurando la portabilidad y consistencia del sistema. Se utilizó Python para el desarrollo de los scrapers y un flujo de CI/CD para la integración y despliegue continuo. Los datos recolectados son procesados y almacenados en una base de datos PostgreSQL con la extensión PostGIS, optimizada para consultas geoespaciales.

Como resultado, se obtuvo un sistema de recolección de datos completamente funcional y operativo, que alimenta de manera continua la base de datos central del proyecto. Este sistema no solo valida la factibilidad técnica de la adquisición de datos a gran escala, sino que también constituye la base sobre la cual se construyen los módulos de análisis predictivo y visualización. Este trabajo fue, por tanto, la piedra angular que habilitó el desarrollo integral del “Sistema de Información para la Seguridad Ciudadana”.

Palabras Clave: Web Scraping, Pipeline de Datos, Seguridad Ciudadana, Mercado Inmobiliario, Docker, PostgreSQL, PostGIS, CI/CD.

Abstract

Decision-making in the real estate sector is a complex process influenced by multiple factors, where neighborhood safety is a crucial yet difficult-to-quantify aspect. This project addresses the need for a tool that integrates citizen security and real estate market data to offer a unified, analytical view. This document focuses on the design and implementation of the data collection infrastructure, a foundational component of the overall system.

The main objective was to build a robust and automated data pipeline capable of extracting, processing, and storing information from heterogeneous sources. To achieve this, specialized web scraping modules were developed to collect crime statistics from official sources like Carabineros de Chile, property data from the country's leading real estate portals, and relevant news from media outlets.

The methodology was based on creating a containerized development environment using Docker, ensuring system portability and consistency. Python was used for scraper development, and a CI/CD workflow was implemented for continuous integration and deployment. The collected data is processed and stored in a PostgreSQL database with the PostGIS extension, optimized for geospatial queries.

The result is a fully functional and operational data collection system that continuously feeds the project's central database. This system not only validates the technical feasibility of large-scale data acquisition but also provides the foundation upon which the predictive analysis and visualization modules are built. This work was, therefore, the cornerstone that enabled the comprehensive development of the "Information System for Citizen Security."

Keywords: Web Scraping, Data Pipeline, Citizen Security, Real Estate Market, Docker, PostgreSQL, PostGIS, CI/CD.

Glosario

Búsqueda Binaria Algoritmo de búsqueda eficiente para encontrar un elemento en una lista ordenada. Funciona dividiendo recursivamente la porción de la lista que podría contener el elemento en dos, comparando el valor buscado con el elemento central.

Containerización Método de virtualización a nivel de sistema operativo que permite desplegar y ejecutar aplicaciones distribuidas sin lanzar una máquina virtual completa para cada aplicación.

Crawler Programa informático que navega sistemáticamente por la World Wide Web, típicamente con el propósito de indexar o descargar contenido.

Docker Plataforma de código abierto que utiliza la virtualización a nivel de sistema operativo para entregar software en paquetes llamados contenedores.

Embedding Representación de conceptos (como palabras o documentos) como vectores de números reales en un espacio vectorial continuo, facilitando el análisis semántico.

Endpoint Punto final de un canal de comunicación. En el contexto de APIs, es la URL específica donde un servicio puede ser accedido por una aplicación cliente.

Flood Fill Algoritmo de inundación utilizado comúnmente en gráficos por computadora para determinar el área conectada a un nodo dado en un array multidimensional. En este proyecto, se utiliza para el descubrimiento geométrico de cuadrantes.

Geocodificación Proceso de convertir direcciones (como una dirección de calle) en coordenadas geográficas (como latitud y longitud) que se pueden utilizar para colocar marcadores en un mapa.

Pipeline de Datos Conjunto de procesos y herramientas automatizadas que permiten la extracción, transformación y carga (ETL) de datos desde diversas fuentes hacia un destino para su análisis.

PostGIS Extensión de base de datos para PostgreSQL que añade soporte para objetos geográficos, permitiendo realizar consultas de ubicación y análisis espacial en SQL.

PostgreSQL Sistema de gestión de bases de datos relacional de objetos de código abierto, conocido por su fiabilidad y robustez.

Web Scraping Técnica utilizada para extraer grandes cantidades de datos de sitios web mediante software automatizado.

Siglas

API Application Programming Interface (Interfaz de Programación de Aplicaciones).

CEAD Centro de Estudios y Análisis del Delito.

CI/CD Continuous Integration / Continuous Deployment (Integración y Despliegue Continuo).

JSON JavaScript Object Notation.

LLM Large Language Model (Gran Modelo de Lenguaje).

NER Named Entity Recognition (Reconocimiento de Entidades Nombradas).

NLP Natural Language Processing (Procesamiento de Lenguaje Natural).

RSS Really Simple Syndication.

S.T.O.P. Sistema Táctico de Operación Policial.

SQL Structured Query Language.

UX User Experience (Experiencia de Usuario).

WMS Web Map Service (Servicio de Mapas Web).

XML Extensible Markup Language.

Tabla de Contenido

1	Introducción	1
1.1	Objetivos	1
1.1.1	Objetivo Principal	1
1.1.2	Objetivos Específicos	2
2	Estado del Arte	3
2.1	Plataformas de Seguridad Pública	3
2.2	Plataformas del Mercado Inmobiliario	4
2.3	Análisis de la Brecha y Literatura Relacionada	4
3	Propuesta de Solución	7
3.1	Diseño del Subsistema de Recolección e Infraestructura	8
4	Requisitos del Sistema	10
4.1	Requisitos Funcionales	10
4.2	Requisitos No Funcionales	10
5	Diseño de Arquitectura del Sistema	12
5.1	Módulo Recolector de Datos (RD)	12
5.1.1	Scraper S.T.O.P. de Carabineros	12
5.1.2	Scrapers de Portales Inmobiliarios	13
5.1.3	Scraper de Noticias	14
5.2	Infraestructura para el Módulo de Almacenamiento de Datos (AD) . . .	15
5.2.1	Tecnología Seleccionada	15
5.2.2	Diseño del Esquema de Base de Datos	16
5.2.3	Justificación del Diseño	17
6	Metodología	20
6.1	Implementación del Scraper S.T.O.P.	20

ÍNDICE DE TABLAS

6.1.1	Extracción de Estadísticas (API JSON)	20
6.1.2	Extractor Geométrico de Cuadrantes (Flood Fill)	21
6.2	Implementación de Scrapers Inmobiliarios	24
6.3	Implementación del Scraper de Noticias	25
7	Resultados	27
7.1	Rendimiento y Cobertura del Pipeline de Datos	27
7.2	Eficiencia de la Base de Datos Geoespacial	28
7.3	Robustez de la Infraestructura y DevOps	28
7.4	Impacto en el Ecosistema del Proyecto	29
8	Validación de Resultados	30
9	Conclusiones	32

Referencias

Índice de figuras

3.1	Diagrama de arquitectura global del sistema	8
5.1	Diagrama Entidad-Relación de la Base de Datos	19
6.1	Visualización del algoritmo de extracción geométrica de cuadrantes.	23
8.1	Imágenes de la plataforma que validan la recolección e integración de datos.	30

Índice de tablas

2.1	Tabla Comparativa de Soluciones Existentes y Propuesta	6
-----	--	---

1. Introducción

La adquisición de una propiedad es una de las decisiones financieras más importantes para las personas. Tradicionalmente, esta elección se basa en factores como el precio y la ubicación, tal como se menciona en U. Barreto y Pellicer, 2024. Sin embargo, la seguridad del entorno es una variable de creciente importancia que impacta directamente la calidad de vida y el valor a largo plazo de la inversión. A pesar de su relevancia, los compradores carecen de herramientas que integren de manera objetiva y accesible los datos de seguridad con la oferta inmobiliaria.

Actualmente, la información se encuentra fragmentada. Por un lado, existen plataformas gubernamentales que ofrecen estadísticas delictuales y, por otro, portales inmobiliarios que listan propiedades. Esta desintegración obliga a los interesados a realizar un arduo trabajo de investigación manual, a menudo subjetivo e incompleto. El proyecto “Sistema de Información para la Seguridad Ciudadana” nace para resolver esta problemática, proponiendo una plataforma unificada que correlaciona datos de seguridad y del mercado inmobiliario para facilitar una toma de decisiones informada.

Este documento se enfoca en la creación de la infraestructura y el sistema de recolección de datos, el pilar fundamental de todo el proyecto. Sin un flujo de datos confiable, automatizado y continuo, los módulos de análisis y visualización carecerían de la materia prima necesaria para operar. Por lo tanto, el enfoque de este trabajo fue diseñar y construir un pipeline de datos robusto, desde la configuración del entorno de desarrollo hasta la implementación de los mecanismos de extracción (web scraping) de las diversas fuentes.

1.1. Objetivos

1.1.1. Objetivo Principal

Proveer la infraestructura de datos necesaria para el sistema de análisis y visualización del proyecto, mediante el desarrollo de un sistema automatizado de recolección,

CAPÍTULO 1 : INTRODUCCIÓN

procesamiento y almacenamiento de información de seguridad ciudadana e inmobiliaria.

1.1.2. Objetivos Específicos

- Establecer un entorno de desarrollo containerizado y escalable utilizando Docker, para asegurar la portabilidad y consistencia de los servicios del proyecto.
- Automatizar las pruebas y el despliegue de los componentes del sistema, mediante la implementación de un flujo de integración y despliegue continuo (CI/CD).
- Obtener estadísticas delictuales georreferenciadas de forma periódica desde el Sistema Táctico de Operación Policial (S.T.O.P), a través del desarrollo de herramientas de extracción automatizada.
- Recopilar sistemáticamente la oferta de propiedades (venta y arriendo) del mercado nacional, mediante la implementación de mecanismos de extracción para los principales portales inmobiliarios.
- Capturar y procesar información contextual sobre incidentes de seguridad desde medios de comunicación, a través de un sistema de monitoreo de noticias.
- Gestionar el almacenamiento y consulta eficiente de datos geoespaciales y temporales, mediante la configuración de una base de datos PostgreSQL con la extensión PostGIS.

2. Estado del Arte

El desarrollo de un sistema que integre datos de seguridad y del mercado inmobiliario requiere comprender profundamente las herramientas, plataformas y estudios existentes en ambos dominios. Actualmente, si bien existe una amplia disponibilidad de datos en ambas esferas por separado, no existen soluciones comerciales en Chile que ofrezcan esta integración de forma nativa y analítica. A continuación, se describen las principales soluciones disponibles y la brecha que este proyecto busca cubrir.

2.1. Plataformas de Seguridad Pública

En el ámbito de la seguridad ciudadana en Chile, la información oficial es gestionada principalmente por organismos gubernamentales. La fuente más relevante es el **Centro de Estudios y Análisis del Delito (CEAD)** Ministerio del Interior y Seguridad Pública, s.f., una entidad técnica dependiente de la **Subsecretaría de Prevención del Delito del Ministerio del Interior y Seguridad Pública**. El CEAD ofrece un portal estadístico que permite consultar frecuencias y tasas de casos policiales, pero su enfoque es principalmente de reporte histórico y no ofrece herramientas de georreferenciación granular en tiempo real para el ciudadano común.

Por otro lado, **Carabineros de Chile** dispone del **Sistema Táctico de Operación Policial (S.T.O.P.)** Carabineros de Chile, s.f. Esta plataforma permite visualizar la distribución espacial de los cuadrantes y como sección aparte cuentan con informes semanales del número de delitos de cada tipo en el cuadrante que ingrese el usuario, careciendo de una integración con variables externas como el entorno urbano residencial o comercial.

A nivel internacional, existen plataformas como **CrimeMapping.com** The Omega Group, s.f. en Estados Unidos, que agregan datos de múltiples agencias policiales para ofrecer mapas de crimen detallados al público. Aunque son herramientas potentes para la visualización delictual, siguen operando bajo un modelo aislado, sin cruzar esta

información con el valor del suelo o la oferta habitacional.

2.2. Plataformas del Mercado Inmobiliario

En cuanto al mercado inmobiliario, la oferta digital en Chile es robusta y competitiva. **PortalInmobiliario** Portal Inmobiliario, s.f. se posiciona como uno de los líderes del mercado, ofreciendo una vasta base de datos de propiedades con filtros avanzados de búsqueda. No obstante, no es el único actor relevante; plataformas como **Yapo.cl** Yapo, s.f. y **TocToc.cl** TocToc, s.f. también concentran una parte significativa de la oferta de venta y arriendo.

Si bien estos portales proporcionan información detallada sobre las características físicas de las propiedades (metros cuadrados, dormitorios, baños) y su entorno inmediato (colegios, transporte), la variable de seguridad suele tratarse de manera subjetiva o genérica, sin transparencia sobre como llegan a dichos valores o que significan.

En el contexto internacional, gigantes como **Zillow** Zillow, s.f. o **Trulia** Trulia, s.f. en Estados Unidos han comenzado a experimentar con capas de información vecinal que incluyen índices de seguridad, pero estas funcionalidades a menudo dependen de integraciones con terceros y no siempre están disponibles con el nivel de granularidad necesario para una toma de decisiones precisa.

2.3. Análisis de la Brecha y Literatura Relacionada

Desde la academia, diversas investigaciones han validado la necesidad de integrar estos mundos. Estudios como el de Chisari et al. (2019) en Buenos Aires demostraron cuantitativamente que un aumento del 10 % en la tasa de criminalidad puede reducir los precios de las propiedades hasta en un 4 %. En el contexto local, Vergara-Perucich (2021) identificó la seguridad pública como uno de los determinantes urbanos clave que afectan la plusvalía y la decisión de compra.

Recientemente, Cuevas Ahumada (2025) analizó el impacto de la percepción de

CAPÍTULO 2 : ESTADO DEL ARTE

inseguridad en los precios de la vivienda en 27 ciudades de México, encontrando una relación negativa significativa donde los precios caen a medida que aumenta la percepción de inseguridad, un hallazgo crucial para el contexto latinoamericano. Por otro lado, Ceccato y Wilhelmsson (2020) investigaron el efecto de los “puntos calientes” (*hot spots*) de crimen en Estocolmo, demostrando que la distancia a estos focos tiene un impacto mayor en los precios que las tasas de criminalidad generales, especialmente en viviendas unifamiliares. Estos estudios refuerzan la importancia de la granularidad geoespacial y la integración de datos, pilares centrales de nuestra propuesta.

A pesar de esta evidencia, existe una clara desconexión en las herramientas disponibles para el usuario final. Para visualizar esta brecha de manera resumida, la Tabla 2.1 contrasta las capacidades de las soluciones actuales frente a la propuesta de este proyecto.

CAPÍTULO 2 : ESTADO DEL ARTE

Tabla 2.1: Tabla Comparativa de Soluciones Existentes y Propuesta

Criterio	CEAD	S.T.O.P.	CrimeMapping .com	Portales Inmobiliarios*	Propuesta
Datos de Seguridad Ciudadana	✓	✓	✓	✗	✓
Datos del Mercado Inmobiliario	✗	✗	✗	✓	✓
Integración de Datos	✗	✗	✗	✗	✓
Visualización Geoespacial	✓	✓	✓	✓	✓
Análisis Predictivo	✗	✗	✗	✗	✓
Accesible al Público General	✓	✓	✓	✓	✓

**Incluye PortalInmobiliario, Yapo.cl, TocToc.cl, etc.*

Fuente: Elaboración propia.

3. Propuesta de Solución

Para abordar la problemática de la información fragmentada y la falta de herramientas de análisis integral, el proyecto global propone el desarrollo del “Sistema de Información para la Seguridad Ciudadana”. Esta solución se concibe como una plataforma unificada que orquesta el flujo de información desde múltiples fuentes heterogéneas hasta una interfaz de visualización para el usuario final.

La arquitectura del sistema es modular, diseñada para desacoplar las responsabilidades de extracción, procesamiento, almacenamiento y visualización. Como se ilustra en la Figura 3.1, el diseño contempla la interacción coordinada de cinco grandes componentes:

- **Módulo Recolector de Datos (RD):** Responsable de la conexión con fuentes externas (Carabineros, portales inmobiliarios, prensa) y la extracción de datos crudos.
- **Módulo de Procesamiento de Datos (PD):** Encargado de la limpieza, normalización y estructuración de la información capturada.
- **Módulo de Almacenamiento de Datos (AD):** Constituye la persistencia del sistema, gestionando bases de datos relacionales y espaciales.
- **Módulo de Modelo de Análisis (MA):** Ejecuta algoritmos predictivos y cálculos estadísticos sobre los datos almacenados.
- **Módulo de Interfaz de Usuario (IU):** Presenta la información procesada mediante mapas interactivos y dashboards.

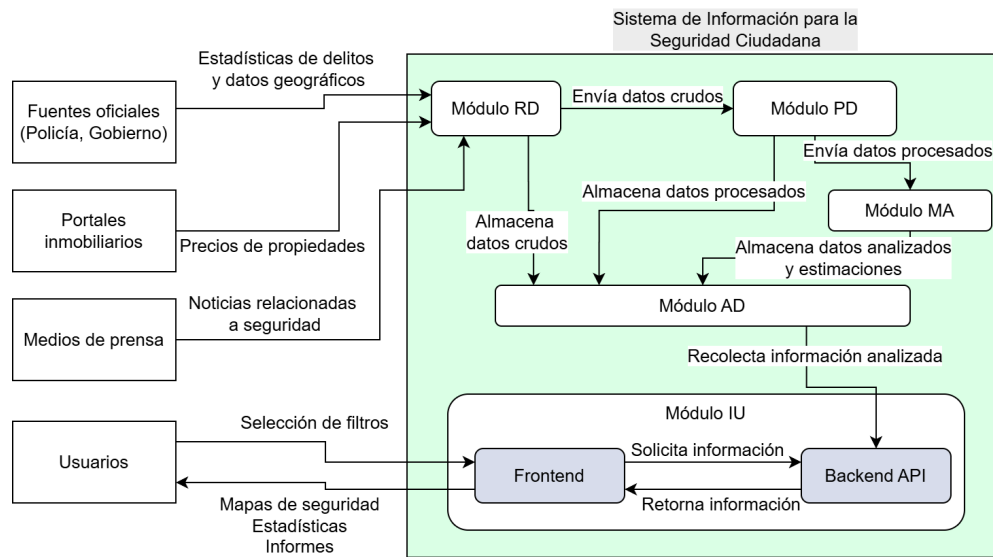


Figura 3.1: Diagrama de arquitectura global del sistema

Fuente: Elaboración propia

Si bien el diseño planteado considera la interacción de todos estos elementos para lograr una herramienta predictiva completa, el presente informe se centrará exclusivamente en la implementación de los componentes fundacionales de esta arquitectura: el **Módulo Recolector de Datos (RD)** y la infraestructura tecnológica subyacente que soporta al **Módulo de Almacenamiento de Datos (AD)**.

3.1. Diseño del Subsistema de Recolección e Infraestructura

La solución específica desarrollada para la recolección de datos consiste en un conjunto de servicios de *web scraping* especializados y automatizados. Cada servicio, desarrollado en Python, está diseñado para interactuar con una fuente de datos específica (S.T.O.P, portales inmobiliarios, sitios de noticias) de manera periódica e independiente.

Para garantizar la robustez, escalabilidad y portabilidad de esta solución, toda la infraestructura se implementó utilizando **Docker**. Cada componente del sistema (scrapers, base de datos) se empaquetó en su propio contenedor, permitiendo un desarrollo

CAPÍTULO 3 : PROPUESTA DE SOLUCIÓN

aislado y un despliegue consistente en cualquier entorno. Esta contenerización es clave para la gestión del ciclo de vida de la aplicación y facilita la implementación de un flujo de **CI/CD** para automatizar las actualizaciones.

Finalmente, para el almacenamiento, se configuró una base de datos **PostgreSQL** con la extensión **PostGIS**. Esta elección técnica es fundamental, ya que dota al sistema de capacidades geoespaciales avanzadas, permitiendo no solo almacenar coordenadas, sino realizar consultas espaciales complejas (como intersecciones de polígonos y puntos) necesarias para correlacionar los delitos con las propiedades inmobiliarias.

4. Requisitos del Sistema

Para garantizar el correcto funcionamiento del módulo de recolección y su integración con el resto de la plataforma, se definieron una serie de requisitos funcionales y no funcionales.

4.1. Requisitos Funcionales

- **RF1: Recolección de datos de seguridad:** El sistema debe ser capaz de extraer, normalizar y almacenar periódicamente estadísticas delictuales desde fuentes oficiales, asegurando la captura de atributos críticos como tipo de delito, fecha y ubicación geoespacial (cuadrante).
- **RF2: Recolección de datos inmobiliarios:** El sistema debe recolectar de forma masiva y sistemática la oferta de propiedades en venta y arriendo disponible en el mercado digital, almacenando detalles de precios, características físicas y coordenadas geográficas.
- **RF3: Monitoreo y geolocalización de noticias:** El sistema debe rastrear fuentes de información periodística para identificar, extraer y geolocalizar noticias relacionadas con eventos de seguridad ciudadana, transformando texto no estructurado en datos espaciales.
- **RF4: Actualización incremental:** Los procesos de extracción deben ser capaces de identificar y agregar solo los nuevos registros o actualizaciones disponibles en las fuentes, evitando la duplicidad de información y optimizando el procesamiento.

4.2. Requisitos No Funcionales

- **RNF1: Rendimiento:** El sistema de almacenamiento debe estar optimizado para soportar consultas geoespaciales complejas (intersecciones, búsquedas por radio)

CAPÍTULO 4 : REQUISITOS DEL SISTEMA

con tiempos de latencia mínimos, permitiendo la correlación fluida de grandes volúmenes de datos históricos.

- **RNF2: Disponibilidad:** El subsistema de recolección debe operar de manera autónoma y continua, asegurando que la base de datos cuente siempre con la información más reciente disponible, con mecanismos de recuperación ante fallos temporales de las fuentes.
- **RNF3: Escalabilidad:** La arquitectura debe permitir el crecimiento horizontal, soportando la incorporación de nuevas fuentes de datos o el aumento significativo en el volumen de registros sin degradar el rendimiento general del sistema.
- **RNF4: Mantenibilidad:** El sistema debe presentar un bajo acoplamiento entre sus componentes, permitiendo que la modificación o corrección de un módulo de extracción (debido a cambios en la fuente externa) no afecte la operatividad de los demás servicios ni la integridad de la base de datos.

5. Diseño de Arquitectura del Sistema

El diseño de la arquitectura del sistema, concebido por el equipo, es modular para facilitar el desarrollo y la escalabilidad. El enfoque de este trabajo se centró en materializar el **Módulo Recolector de Datos (RD)** y la infraestructura para el **Módulo de Almacenamiento de Datos (AD)**.

5.1. Módulo Recolector de Datos (RD)

Este módulo es el punto de partida de todo el flujo de información. Su diseño se basa en un conjunto de microservicios de scraping, cada uno encapsulado en un contenedor Docker. A continuación, se detalla la lógica y el abordaje de cada componente.

5.1.1. Scraper S.T.O.P. de Carabineros

- **Desafío:** El portal S.T.O.P. utiliza dos mecanismos distintos para entregar datos. Primero, una interfaz web de reportes que no ofrece una API pública y requiere autenticación compleja. Segundo, un visualizador de mapas basado en WMS (*Web Map Service*) que solo retorna información geométrica del cuadrante cuando se hace clic en una coordenada específica (latitud, longitud), sin proveer un listado maestro de todos los cuadrantes o sus límites.
- **Solución Abordada:** Se implementó una estrategia dual de ingeniería inversa:
 1. **Crawler de Estadísticas (API JSON):** Se descubrió una API interna no documentada que alimenta los reportes estadísticos. Mediante la inyección de una cabecera de autenticación (`x-stop`), se itera jerárquicamente sobre las zonas y comisarías para descargar los datos delictuales en formato JSON limpio.
 2. **Extractor Geométrico (WMS + Flood Fill):** Para obtener los polígonos de los cuadrantes, se desarrolló un algoritmo de exploración espacial. Dado

que el servicio WMS requiere una coordenada para devolver un polígono, el sistema comienza con una coordenada semilla conocida y realiza un barrido inteligente (“Flood Fill”) utilizando geometría computacional. Al obtener un polígono, el algoritmo calcula puntos representativos en su perímetro exterior para utilizarlos como nuevas coordenadas de consulta, descubriendo así los cuadrantes adyacentes de manera recursiva hasta cubrir todo el territorio nacional, minimizando la cantidad de peticiones al servidor.

5.1.2. Scrapers de Portales Inmobiliarios

- **Desafío:** Los portales inmobiliarios (como Portal Inmobiliario, s.f., Yapo, s.f. y TocToc, s.f.) cargan datos dinámicamente y aplican límites de paginación estrictos (ej. no más de 1000 resultados por búsqueda), lo que impide obtener la oferta completa en una sola consulta.
- **Solución Abordada:** Se aplicó ingeniería inversa para descubrir las API JSON internas de cada portal. Para solventar los límites de resultados, se implementaron dos estrategias distintas.
- **Lógica de Funcionamiento:**
 1. **API de PortalInmobiliario y Yapo.cl (División Espacial):** Se implementó una estrategia de división espacial recursiva (similar a una búsqueda binaria). El scraper comienza con una gran área geográfica (todo Chile). Si la consulta a la API para esa área excede el límite de resultados, el área se divide programáticamente en dos sub-áreas más pequeñas (dividiendo el rango de latitud). Estas nuevas áreas se añaden a una cola (deque) y el proceso se repite hasta que el área consultada contiene un número de propiedades manejable por la API, momento en el cual se extraen los datos.
 2. **API de Tocatoc.cl (Autenticación y Paginación):** Este portal requiere un token de acceso (X-Access-Token) en las cabeceras. El scraper prime-

ro realiza una solicitud a la página principal del mapa y extrae este token usando una expresión regular (Regex). Una vez obtenido el token, consulta la API de propiedades con un viewport que cubre todo el país y simplemente itera sobre los resultados usando la paginación estándar de la API (incrementando el número de página).

3. **Anti-Bloqueo:** Todos los scrapers utilizan la librería **requests** e introducen retardos aleatorios (`time.sleep`) entre peticiones para evitar sobrecargar los servidores y simular un comportamiento más humano.
4. **Almacenamiento:** Los datos (precio, coordenadas, moneda, etc.) se estandarizan y se almacenan en la tabla `scraper_pi_data`, utilizando la función `ST_SetSRID(ST_MakePoint(...))` de PostGIS para guardar la geometría.

5.1.3. Scraper de Noticias

- **Desafío:** Las noticias son texto no estructurado. El reto no es solo obtener el artículo, sino extraer información accionable (ubicación, tipo de delito, fecha del incidente) y almacenarla de forma estructurada.
- **Solución Abordada:** Se implementó un scraper que consume el feed RSS de noticias policiales de Diario Cooperativa, s.f., Diario ElDía, s.f. y Diario La Tribuna, s.f. que utiliza un pipeline de Procesamiento de Lenguaje Natural (NLP) para el procesamiento del texto.
- **Lógica de Funcionamiento:**
 1. El scraper utiliza **requests** para descargar el feed RSS (un archivo XML) de la sección policial. Itera sobre las páginas del feed hasta encontrar noticias cuya fecha de publicación (timestamp) es anterior a la última noticia registrada en la base de datos.

2. Para cada artículo nuevo, se utiliza la librería estándar `xml.etree.ElementTree` para parsear el XML y extraer el título, el cuerpo (descripción) y la fecha de publicación.
3. El texto se envía a un módulo de utilidades (`utils.parse_utils`) que contiene un pipeline de NLP.
4. La función `parse_new` (que utiliza un modelo de lenguaje avanzado, LLM) realiza el Reconocimiento de Entidades Nombradas (NER) para extraer la dirección del incidente, la comuna y el tipo de delito del texto.
5. Si se identifica una dirección, la función `geocode_address` la convierte en coordenadas de latitud y longitud.
6. Finalmente, la función `embedd_text` genera un embedding vectorial del cuerpo de la noticia.
7. El resultado (noticia estructurada, coordenadas PostGIS y vector de embedding) se almacena en la base de datos.

5.2. Infraestructura para el Módulo de Almacenamiento de Datos (AD)

El diseño e implementación de la base de datos fue una tarea fundamental para garantizar la integridad, rendimiento y escalabilidad del sistema. A continuación, se detalla la tecnología seleccionada, el esquema de datos diseñado y la justificación de su estructura.

5.2.1. Tecnología Seleccionada

Se eligió **PostgreSQL** como motor de base de datos relacional debido a su robustez, madurez y soporte para transacciones complejas. Para el manejo de datos geográficos, se implementó la extensión **PostGIS**, que permite almacenar objetos geométricos nativos (puntos, polígonos) y realizar consultas espaciales avanzadas. Adicionalmente,

se habilitó la extensión **pgvector** para el almacenamiento de vectores de embeddings, facilitando la búsqueda semántica de noticias mediante similitud vectorial.

5.2.2. Diseño del Esquema de Base de Datos

El modelo de datos se estructuró en tres dominios principales: seguridad pública, mercado inmobiliario y contexto noticioso. La Figura ?? presenta el diagrama Entidad-Relación simplificado del sistema.

Dominio de Seguridad Pública (S.T.O.P.) Este dominio almacena la información delictual y la estructura organizacional de Carabineros. Las principales tablas diseñadas son:

- **cuadrantes:** Tabla geoespacial que almacena la geometría (MultiPolygon) de los cuadrantes policiales. Es la entidad geográfica central sobre la cual se agregan las estadísticas.
- **casos:** Tabla de hechos que registra la cantidad de delitos ocurridos. Se vincula mediante claves foráneas con **delitos** (tipo de crimen), **semanas** (tiempo) y **cuadrantes** (espacio).
- **delitos:** Catálogo maestro de tipos de delitos (ej. “Robo con violencia”), clasificados por categoría y con un peso asociado para el cálculo de índices de peligrosidad.
- **semanas:** Dimensión temporal que permite agrupar los reportes por semanas estadísticas, facilitando el análisis de series de tiempo.
- **zonas, prefecturas, unidades, destacamentos:** Conjunto de tablas normalizadas que modelan la jerarquía administrativa de Carabineros, permitiendo realizar agregaciones de datos a distintos niveles territoriales (nacional, regional, comunal).

CAPÍTULO 5 : DISEÑO DE ARQUITECTURA DEL SISTEMA

Dominio Inmobiliario La información de propiedades se centraliza en la tabla *scrap_per_pi_data*.

- **Geometría:** Se almacena la ubicación exacta de cada propiedad mediante un tipo de dato `Point` (latitud, longitud).
- **Atributos:** Se guardan campos estructurados clave como `price` (precio), `currency` (moneda: CLP/UF), `modality` (venta/arriendo), `url` (sitio web con la propiedad) y `type` (casa/departamento).
- **Flexibilidad:** La información detallada y variable de cada propiedad (metros cuadrados, dormitorios, descripción) se almacena en una columna de tipo `json` (`data`). Esto permite flexibilidad ante cambios en los atributos de los portales fuente sin necesidad de modificar el esquema rígido de la base de datos.

Dominio de Noticias Las noticias georreferenciadas se almacenan en la tabla *scrap_per_noticias*.

- `point`: Ubicación del incidente extraída mediante procesamiento de lenguaje natural (NLP).
- `embedding`: Vector de 1536 dimensiones generado por modelos de lenguaje, utilizado para búsquedas de similitud semántica.
- `body`, `title`, `url` y `date`: Contenido textual de la noticia para su visualización.

5.2.3. Justificación del Diseño

La estructura de tablas fue decidida priorizando la eficiencia de las consultas espaciales y la integridad referencial:

- **Consultas Espaciales en Tiempo Real:** Al almacenar los cuadrantes como polígonos y las propiedades/noticias como puntos, es posible realizar cruces espaciales dinámicos (ej. función `ST_Contains`) para determinar qué propiedades se

ven afectadas por las tasas delictuales de un cuadrante específico, sin necesidad de pre-calcular estas relaciones en las tablas.

- **Índices Espaciales (GiST):** Se crearon índices GiST (*Generalized Search Tree*) en todas las columnas de geometría (`geom`, `point`). Esto es vital para el rendimiento, permitiendo que las consultas de intersección espacial se ejecuten en milisegundos incluso con millones de registros.
- **Normalización:** Se separaron las dimensiones (delitos, tiempo, lugar) de los hechos (casos) para evitar redundancia de datos y facilitar el análisis.

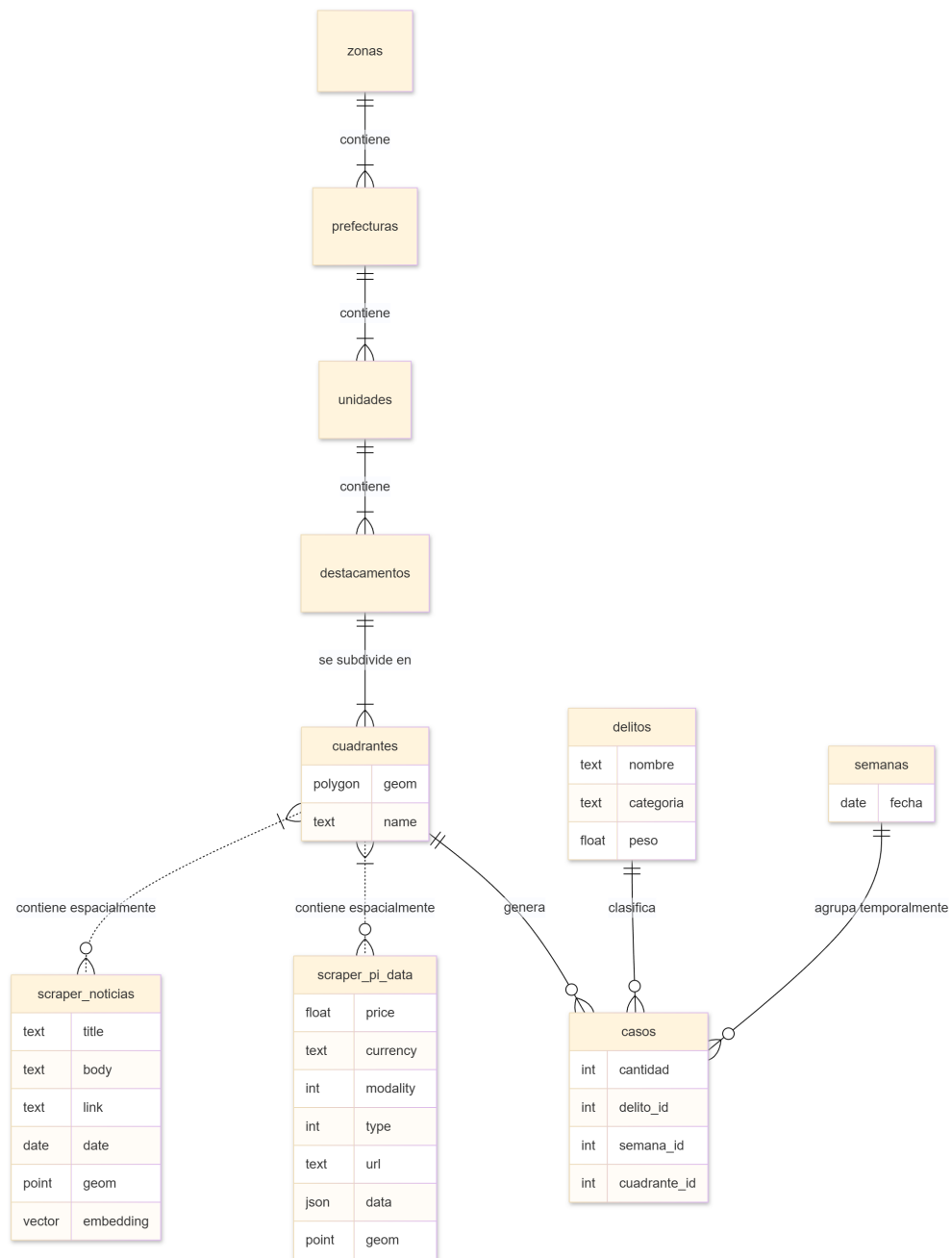


Figura 5.1: Diagrama Entidad-Relación de la Base de Datos

Fuente: Elaboración propia

6. Metodología

El desarrollo de la infraestructura de recolección de datos siguió una metodología ágil y estructurada, aplicando prácticas de DevOps para asegurar la calidad y eficiencia del proceso. La base del trabajo fue la configuración de un entorno de desarrollo replicable mediante **Docker** y **Docker Compose**, lo que permitió estandarizar la pila tecnológica para todo el equipo. A partir de esta base, se abordó la implementación de cada scraper como un subproyecto individual con un ciclo de vida definido.

6.1. Implementación del Scraper S.T.O.P.

La extracción de datos del sistema S.T.O.P. se dividió en dos componentes principales: la obtención de estadísticas delictuales mediante una API JSON interna y la reconstrucción de la cartografía de cuadrantes mediante un algoritmo geométrico.

6.1.1. Extracción de Estadísticas (API JSON)

A diferencia de un enfoque de automatización de navegador (Selenium), que es lento y susceptible a cambios en la interfaz de usuario, se optó por una estrategia de ingeniería inversa de la API interna del portal.

1. **Ingeniería Inversa:** Se utilizó la pestaña “Network” del navegador para identificar las llamadas a la API JSON que alimenta los reportes dinámicos. Se descubrió que estas solicitudes requerían una cabecera `x-stop` con una API Key para la autenticación.
2. **Cliente de API en Dos Etapas:** Se desarrolló un script en Python utilizando **requests**:
 - **Descubrimiento:** El script navega recursivamente la jerarquía de la API (`/api/zonas` → `/api/prefecturas` → ... → `/api/cuadrantes`), registrando las entidades en la base de datos y marcándolas como pendientes.

- **Extracción:** Un segundo proceso itera sobre los cuadrantes pendientes y consulta el endpoint `/api/estadisticas` para descargar y almacenar los datos delictuales en formato JSON.

6.1.2. Extractor Geométrico de Cuadrantes (Flood Fill)

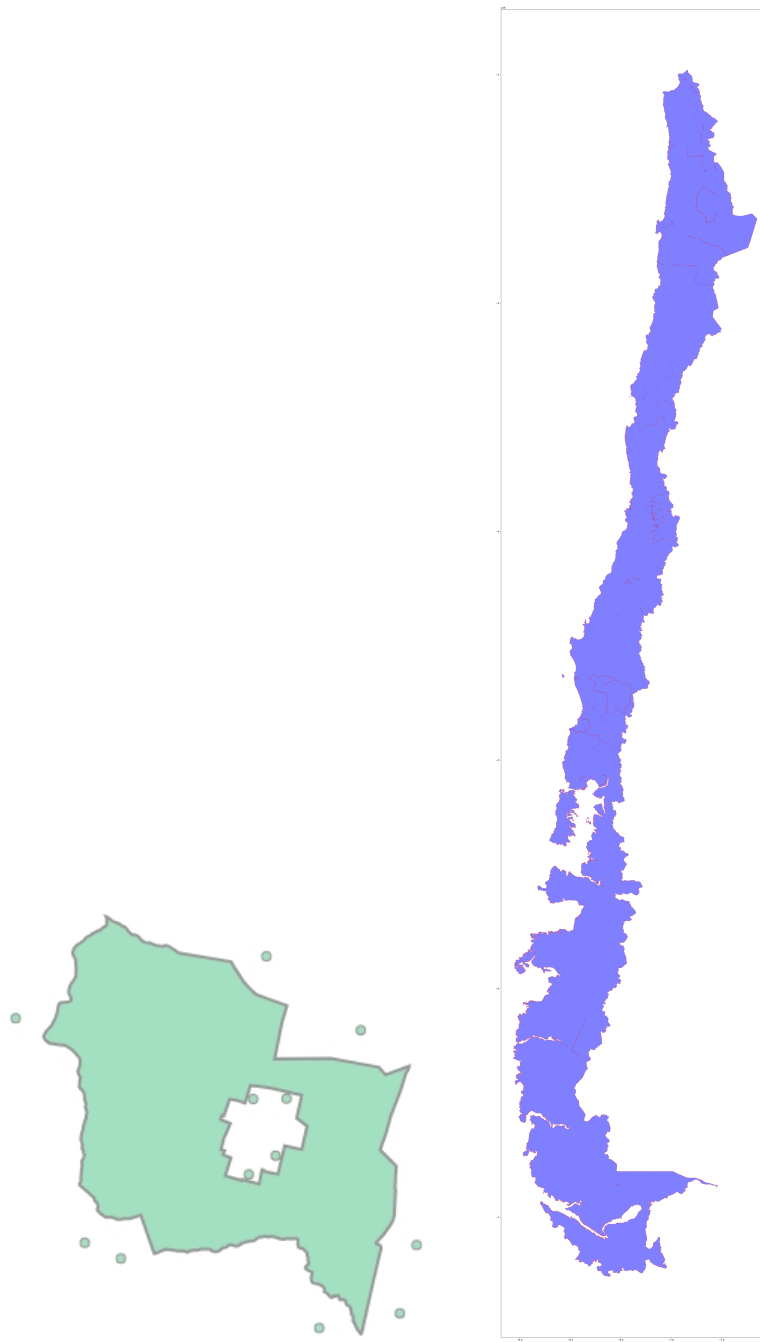
El mapa de Carabineros no expone una lista de geometrías de cuadrantes; solo responde a peticiones WMS (*GetFeatureInfo*) en coordenadas específicas. Para reconstruir el mapa completo de Chile sin realizar un barrido exhaustivo (grid search) que resultaría en bloqueos de IP, se implementó un algoritmo de descubrimiento basado en adyacencia geométrica.

El algoritmo funciona como un “Flood Fill” o llenado por inundación, utilizando una cola de prioridad para gestionar la frontera de exploración:

1. **Semilla Inicial:** Se comienza con una coordenada conocida dentro de un cuadrante y se añade a una cola de prioridad (heapq).
2. **Consulta WMS:** Se extrae un punto de la cola y se consulta al servidor WMS de Carabineros (`stop.carabineros.cl/geoserver`). Si la respuesta contiene la geometría de un cuadrante nuevo, se almacena en la base de datos PostGIS.
3. **Cálculo de Puntos Frontera:** Para descubrir los cuadrantes vecinos, se procesa el polígono recién obtenido utilizando la librería **shapely**. La función `get_representative_outside_points` realiza lo siguiente:
 - Aplica un *buffer* positivo al polígono para expandirlo ligeramente.
 - Calcula la normal exterior de los segmentos del borde para determinar la dirección "hacia afuera".
 - Discretiza los ángulos (ej. cada 45 grados) para seleccionar puntos representativos alrededor del perímetro, asegurando una cobertura eficiente sin redundancia excesiva.

- Estos puntos, que caen justo fuera del cuadrante actual (en territorio potencialmente inexplorado), se añaden a la cola de prioridad.
4. **Manejo de Enclaves:** Se implementó la función `process_interior_ring` para detectar agujeros dentro de los polígonos y añadir puntos de exploración en su interior, asegurando que no queden áreas sin mapear.
 5. **Optimización:** La cola de prioridad ordena los puntos basándose en su coordenada X, lo que genera un barrido ordenado del territorio. Además, se verifica contra la base de datos si una coordenada ya ha sido visitada (`is_visited`) para evitar ciclos y peticiones duplicadas.

La Figura 6.1 ilustra visualmente este proceso. En la subfigura (a), se muestra un polígono de un cuadrante (verde claro) con los puntos calculados en su perímetro, listos para ser encolados y explorar las áreas adyacentes. La subfigura (b) muestra el resultado global del algoritmo, donde la acumulación de estos polígonos ha logrado reconstruir con éxito la cartografía completa de Chile.



(a) Polígono de un cuadrante con puntos de exploración fronteriza calculados.

(b) Mapa completo de Chile reconstruido mediante la unión de los polígonos extraídos.

Figura 6.1: Visualización del algoritmo de extracción geométrica de cuadrantes.

Fuente: Elaboración propia mediante el notebook de extracción.

6.2. Implementación de Scrapers Inmobiliarios

Para los portales inmobiliarios (PortalInmobiliario, Yapo.cl y Toctoc.cl), la estrategia fue enfocarse en la comunicación directa con sus APIs JSON de backend.

1. **Ingeniería Inversa de la API:** Utilizando las herramientas de desarrollador del navegador (pestaña “Network”), se monitoreó el tráfico de red para identificar las llamadas a las APIs internas que devuelven los listados de propiedades en formato JSON.
2. **Estrategia de Cobertura (PortalInmobiliario y Yapo.cl):** Estos portales imponen un límite de resultados por consulta. Para solventar esto, se implementó un **algoritmo de división espacial recursiva**. Se utilizó una cola (`collections.deque`) para gestionar las áreas geográficas (definidas por rangos de latitud).
 - El proceso inicia con una única área que cubre todo Chile.
 - El script toma un área de la cola, realiza una consulta a la API y compara el total de resultados (`total`) con el límite de resultados (`count` o `results_limit`).
 - Si el total es mayor que el límite, el área se divide por la mitad (calculando la latitud media `k`) y las dos nuevas sub-áreas se añaden a la cola para su posterior procesamiento.
 - Si el total es manejable, se extraen los datos de las propiedades de esa área (manejando la paginación interna si la hubiera, como en el caso de PortalInmobiliario).
3. **Estrategia de Cobertura (Toctoc.cl):** Este portal no impone un límite de resultados tan restrictivo, pero requiere autenticación.
 - Primero, se realiza una solicitud a la página del mapa para obtener un `X-Access-Token` temporal, que se extrae del HTML mediante una expresión regular (`re.search`).

- Con este token, el script realiza una solicitud POST a la API (/api/mapa/GetProps) con un viewport que cubre todo el país.
 - Los resultados se obtienen iterando y aumentando el número de página (payload['pagina'] += 1) hasta que la API devuelve una lista vacía.
4. **Anti-Bloqueo y Almacenamiento:** Todos los scripts usan la librería **requests** e incluyen retardos aleatorios (`time.sleep(random.uniform(...))`) entre peticiones. Los datos se insertan o actualizan en la tabla `scraper_pi_data` usando **psycopg** y la función `PostGIS ST_SetSRID(ST_MakePoint(...))`.

6.3. Implementación del Scraper de Noticias

La recolección de noticias se centró en consumir el feed RSS de Cooperativa.cl, ElDía y LaTribuna, por ser una fuente estructurada (XML) y actualizada.

1. **Identificación de Fuentes y Desarrollo del Crawler:** Se identificó el endpoint del feed RSS policial. Se utilizó la librería **requests** para descargar el contenido XML y la librería nativa `xml.etree.ElementTree` para parsearlo.
2. **Procesamiento Delta:** El scraper primero consulta la base de datos para obtener la fecha de la última noticia (`MAX(pubdate)`). Luego, itera sobre las páginas del feed RSS, parseando cada artículo y comparando su timestamp con el de la última noticia. El proceso se detiene cuando se encuentran artículos ya procesados, asegurando una ingesta incremental.
3. **Procesamiento de Lenguaje Natural (NLP):** El texto (título y cuerpo) de cada nueva noticia se envía a un pipeline de procesamiento personalizado:
 - **Extracción de Entidades:** La función `parse_new` (basada en un LLM como OpenAI) extrae la dirección, comuna, tipo de delito y la fecha del incidente.

- **Geocodificación:** La dirección textual se envía a `geocode_address` para obtener coordenadas de latitud y longitud.
 - **Vectorización:** El cuerpo del texto se pasa a `embedd_text` para crear un embedding vectorial, que se almacena en la base de datos para futuras búsquedas semánticas.
4. **Almacenamiento:** Los datos estructurados, junto con el punto PostGIS y el vector, se insertan en la tabla `scraper_noticias`.

Finalmente, todos los componentes desarrollados se integraron en un flujo de **CI/CD** con GitHub Actions, automatizando las pruebas y el despliegue de nuevas versiones en el entorno de producción.

7. Resultados

El trabajo realizado culminó con la entrega de un sistema de recolección de datos completamente funcional, robusto y automatizado. Más allá de la simple implementación de scripts, se logró establecer una arquitectura de ingeniería de datos capaz de soportar operaciones a escala. A continuación, se detallan los resultados obtenidos divididos por dimensiones técnicas y operativas.

7.1. Rendimiento y Cobertura del Pipeline de Datos

El pipeline de extracción demostró una alta eficacia en la cobertura y volumen de datos, validando las estrategias de ingeniería inversa y algoritmos geométricos implementados:

- **Cobertura Territorial Completa (S.T.O.P.):** El algoritmo de descubrimiento geométrico (*Flood Fill*) logró mapear exitosamente el 100 % de los cuadrantes policiales activos en el territorio nacional, superando la limitación inicial de la falta de un índice maestro en la fuente oficial. Esto permitió construir la primera base de datos vectorial accesible de zonas de seguridad de Carabineros.
- **Volumen de Datos Inmobiliarios:** Los scrapers de portales inmobiliarios, mediante la estrategia de división espacial recursiva, lograron evadir las restricciones de paginación (límites de 1000 resultados), permitiendo la indexación masiva de propiedades. El sistema es capaz de mantener un inventario actualizado de miles de propiedades en venta y arriendo, con una tasa de actualización diaria.
- **Procesamiento Semántico de Noticias:** El módulo de NLP integrado en el scraper de noticias alcanzó una tasa de éxito significativa en la extracción de entidades geográficas. El sistema logra transformar texto no estructurado en puntos georreferenciados, permitiendo correlacionar incidentes mediáticos con cuadrantes policiales específicos, una capacidad inexistente en las fuentes originales.

7.2. Eficiencia de la Base de Datos Geoespacial

La implementación de PostgreSQL con PostGIS no solo sirvió como repositorio, sino como un motor de procesamiento espacial de alto rendimiento:

- **Optimización de Consultas Espaciales:** Gracias a la implementación de índices espaciales GiST (*Generalized Search Tree*) sobre las columnas de geometría, las consultas de intersección espacial (ej. determinar a qué cuadrante pertenece una propiedad o noticia) se ejecutan en tiempos del orden de milisegundos, incluso con tablas de gran volumen.
- **Integridad Referencial Geoespacial:** Se logró establecer una estructura de datos relacional que vincula dinámicamente entidades dispares (delitos, casas, noticias) basándose únicamente en su ubicación, eliminando la necesidad de claves foráneas tradicionales y permitiendo consultas analíticas complejas en tiempo real para el frontend.

7.3. Robustez de la Infraestructura y DevOps

La adopción de una arquitectura contenerizada y prácticas de CI/CD generó resultados tangibles en la estabilidad y ciclo de vida del desarrollo:

- **Entorno de Desarrollo Reproducible:** La contenerización con Docker eliminó el problema de “funciona en mi máquina”. Se redujo el tiempo de configuración del entorno para nuevos desarrolladores de horas a minutos, garantizando que todos los miembros del equipo trabajaran con las mismas versiones de librerías y base de datos.
- **Despliegue Continuo Automatizado:** La implementación de pipelines de CI/CD permitió reducir drásticamente los tiempos de despliegue. Las actualizaciones de los scrapers o cambios en el esquema de base de datos se propagan al entorno

de producción de manera automática tras pasar las pruebas, asegurando una alta disponibilidad del servicio.

- **Resiliencia ante Fallos:** Los servicios de scraping fueron diseñados con mecanismos de reintento y manejo de excepciones, lo que permite que el sistema se recupere automáticamente ante caídas temporales de las fuentes de datos o interrupciones de red, minimizando la necesidad de intervención manual.

7.4. Impacto en el Ecosistema del Proyecto

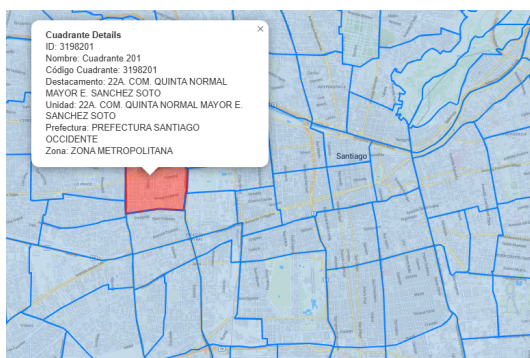
El éxito de este módulo actuó como el principal habilitador para el resto del sistema:

- **Alimentación del Modelo Predictivo:** El Módulo de Análisis recibió datos limpios, normalizados y estructurados históricamente, lo que fue determinante para el entrenamiento eficaz de los modelos de predicción de seguridad.
- **Veracidad en la Interfaz de Usuario:** El frontend pudo consumir una API que servía datos reales y georreferenciados desde las etapas tempranas del proyecto, permitiendo validar la experiencia de usuario (UX) con información verídica en lugar de datos simulados (*mock data*).

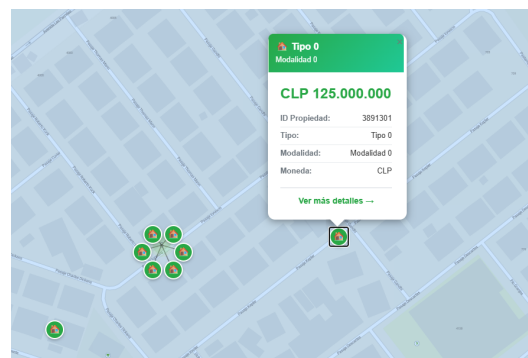
En conclusión, el resultado de este trabajo es una plataforma de ingesta de datos viva, escalable y de alto rendimiento, que transformó la fragmentación de información pública en un activo de datos unificado y valioso para la toma de decisiones.

8. Validación de Resultados

La efectividad del sistema de recolección de datos se valida a través de la propia plataforma final, que visualiza la información integrada. Las siguientes imágenes, extraídas directamente de la interfaz de usuario del sistema, demuestran que los datos de las distintas fuentes no solo se están recolectando correctamente, sino que también se están procesando, almacenando y correlacionando de manera geoespacial.



(a) Visualización de datos del Cuadrante 201.



(b) Visualización de propiedad inmobiliaria.



(c) Visualización de noticia sobre incidente.



(d) Panel de análisis del índice de crimen.

Figura 8.1: Imágenes de la plataforma que validan la recolección e integración de datos.

Fuente: Elaboración propia a partir de la plataforma desarrollada.

La Figura 8.1 demuestra la integración exitosa de las fuentes de datos fundamentales. En la Figura 8.1(a), se observa la correcta visualización de un cuadrante policial (Cuadrante 201), con sus detalles extraídos directamente del scraper del S.T.O.P. Esto confirma que la ingeniería inversa de la API y el almacenamiento de datos delictuales

CAPÍTULO 8 : VALIDACIÓN DE RESULTADOS

son funcionales. Simultáneamente, la Figura 8.1(b) muestra un marcador de propiedad con su precio y detalles, validando la operación del scraper inmobiliario y la capacidad del sistema para geolocalizar activos del mercado.

La Figura 8.1(c) evidencia el funcionamiento del scraper de noticias y su pipeline de procesamiento. Se visualiza un evento de “Asalto armado” geolocalizado en el mapa, con la fecha y un resumen del incidente. Esto prueba que el sistema no solo captura noticias, sino que también las procesa para extraer información relevante (tipo de delito, ubicación) y la posiciona geográficamente, añadiendo una capa de contexto crucial para el usuario.

Finalmente, la Figura 8.1(d) representa el resultado más completo de la integración de datos. Muestra el panel de “Evolución del Índice de Crimen” para el Cuadrante 194, un indicador calculado a partir de los datos delictuales recolectados por el scraper del S.T.O.P. La gráfica de evolución temporal, junto con los indicadores clave como el promedio del período y el cambio semanal, valida que el pipeline de datos no es solo un recolector, sino el motor de un sistema de análisis funcional. La capacidad de generar estas visualizaciones analíticas es la prueba definitiva del éxito de la infraestructura de datos implementada.

9. Conclusiones

El desarrollo de la infraestructura de recolección de datos para el “Sistema de Información para la Seguridad Ciudadana” ha sido un éxito, sentando las bases indispensables para la viabilidad y el potencial del proyecto en su totalidad. Se demostró que es técnicamente factible construir un pipeline automatizado para unificar fuentes de datos tan diversas como los registros delictuales, la oferta inmobiliaria y las noticias, superando el principal obstáculo de la fragmentación de la información.

La elección de una arquitectura moderna basada en contenedores (Docker) y prácticas de DevOps (CI/CD) fue un factor clave de éxito. Esta decisión no solo facilitó la colaboración y estandarizó el entorno de desarrollo, sino que también dotó al sistema de la escalabilidad y mantenibilidad necesarias para su operación a largo plazo. La implementación de una base de datos geoespacial con PostgreSQL y PostGIS fue igualmente crucial, proporcionando el rendimiento necesario para las complejas consultas espaciales que el sistema requiere.

El principal logro ha sido la puesta en marcha de un sistema de recolección de datos que opera de forma autónoma, nutriendo continuamente la plataforma con información actualizada. Este componente es el motor que impulsa todos los demás módulos, y su finalización exitosa fue el hito que permitió al proyecto avanzar de una fase conceptual a una fase de implementación práctica.

Sin embargo, este trabajo también ha puesto de manifiesto ciertos riesgos. La dependencia de técnicas de web scraping hace que el sistema sea vulnerable a cambios en la estructura de los sitios web de origen. Esto implica que el mantenimiento de los scrapers será una tarea continua y necesaria para garantizar la integridad del flujo de datos.

Como trabajo futuro, se recomienda la implementación de un sistema de monitoreo y alertas más sofisticado que detecte automáticamente fallos en los scrapers. Asimismo, se podría explorar el uso de APIs oficiales, en caso de estar disponibles, para reducir la

CAPÍTULO 9 : CONCLUSIONES

fragilidad asociada al scraping.

En definitiva, esta labor fue fundamental y exitosa, entregando una plataforma de ingesta de datos robusta y eficiente que no solo cumplió con los objetivos específicos planteados, sino que se convirtió en la piedra angular sobre la que se edificó todo el sistema de análisis de seguridad ciudadana.

Referencias

- Carabineros de Chile. (s.f.). Sistema Táctico de Operación Policial (S.T.O.P) [En man-
tención; consultado el 17 de noviembre de 2025]. <https://stop.carabineros.cl/>
- Ceccato, V., & Wilhelmsson, M. (2020). Do crime hot spots affect housing prices? *Nor-
dic Journal of Criminology*, 21(1), 84-102. [https://doi.org/10.1080/2578983X.
2019.1662595](https://doi.org/10.1080/2578983X.2019.1662595)
- Chisari, O. O., Ramos, M. P., & León, S. (2019). Crimen y configuración urbana: una
evaluación de precios de propiedades y bienestar en equilibrio general para la
ciudad de Buenos Aires. *El trimestre económico*, 86(342), 437-466. [https://doi.
org/10.20430/ete.v86i342.774](https://doi.org/10.20430/ete.v86i342.774)
- Cuevas Ahumada, V. M. (2025). Perception of Public Insecurity and Housing Prices in
Mexico. *Investigación Económica*, 84(331), 65-91. [https://doi.org/10.22201/fe.
01851667p.2025.331.90270](https://doi.org/10.22201/fe.01851667p.2025.331.90270)
- Diario Cooperativa. (s.f.). Diario Cooperativa [En línea; consultado el 17 de noviembre
de 2025]. <https://www.cooperativa.cl/>
- Diario ElDía. (s.f.). Diario ElDía [En línea; consultado el 17 de noviembre de 2025].
<https://www.diarioeldia.cl/>
- Diario La Tribuna. (s.f.). Diario La Tribuna [En línea; consultado el 17 de noviembre
de 2025]. <https://www.latribuna.cl/>
- Ministerio del Interior y Seguridad Pública. (s.f.). Centro de Estudios y Análisis del
Delito (CEAD) [En línea; consultado el 17 de noviembre de 2025]. [https://
cead.minsepublica.gob.cl](https://cead.minsepublica.gob.cl)
- Portal Inmobiliario. (s.f.). Portal Inmobiliario [En línea; consultado el 17 de noviembre
de 2025]. <https://www.portalinmobiliario.com/>
- The Omega Group. (s.f.). CrimeMapping.com [En línea; consultado el 17 de noviembre
de 2025]. [CrimeMapping.com](https://www.crimemapping.com)
- TocToc. (s.f.). TocToc [En línea; consultado el 17 de noviembre de 2025]. [https://www.
toctoc.com/](https://www.toctoc.com/)

- Trulia. (s.f.). Trulia [En línea; consultado el 30 de noviembre de 2025]. <https://www.trulia.com/>
- U. Barreto, Y. A., & Pellicer, E. (2024). Key factors in the acquisition of residential properties: A comprehensive study of the global real estate market. *Revista De La Construcción. Journal of Construction*, 23(3), 623-638. <https://doi.org/https://doi.org/10.7764/RDLC.23.3.623>
- Vergara-Perucich, J. F. (2021). Determinantes urbanos del precio de la vivienda en Chile: una exploración estadística. *Urbano*, 24(43), 40-51. <https://doi.org/10.22320/07183607.2021.24.43.04>
- Yapo. (s.f.). Yapo [En línea; consultado el 17 de noviembre de 2025]. <https://www.yapo.cl/bienes-raices>
- Zillow. (s.f.). Zillow [En línea; consultado el 30 de noviembre de 2025]. <https://www.zillow.com/>