Documento de Visión y Alcance

# Proyecto: Sistema de Web Scraping para X (antes Twitter)

**RascaCazuelas S.A.**

**Administración de Proyectos I Profesora:** Alma Delia De Jesús Islao

**Ciclo escolar:** 2025-I

## Elaborado por:

Leonardo Valdes Palafox - 18-011-0249

Andrea Marlene Ortega Almendares - 17-003-1026

**Fecha de entrega:** 17 de febrero de 2025

**Versión:** 1.1

# Control de Versiones

|  |  |  |  |
| --- | --- | --- | --- |
| **Versión** | **Fecha** | **Descripción del cambio** | **Responsable** |
| 1.0 | 17/02/2025 | Versión inicial | Leonardo Valdes, Andrea Ortega |
| 1.1 | 30/04/2025 | Actualización y robustecimiento del documento | Leonardo Valdes, Andrea Ortega |
| C C | | | |

**Tabla de Contenidos**

1. Introducción
   1. Propósito
   2. Alcance
   3. Definiciones y Acrónimos
   4. Referencias
2. Posicionamiento
   1. Oportunidad de Negocio
   2. Declaración del Problema
   3. Declaración de Posición del Producto
3. Descripción de Stakeholders y Usuarios
   1. Resumen de Stakeholders
   2. Perfiles de Usuario
   3. Necesidades de Stakeholders y Usuarios
   4. Alternativas y Competencia
4. Visión General del Producto
   1. Perspectiva del Producto
   2. Necesidades y Características
   3. Objetivos de Negocio
   4. Criterios de Éxito
   5. Alcance y Limitaciones
      1. Alcance de las Entregas
      2. Criterios de Aceptación
      3. Diagrama de Contexto
      4. Entorno de Operación
      5. Fuera del Alcance
      6. Gestión de Riesgos
         1. Identificación de Riesgos
         2. Análisis Cualitativo
         3. Estrategias de Mitigación
      7. Recursos y Planificación
         1. Equipo del Proyecto
         2. Cronograma Preliminar
         3. Presupuesto Estimado
      8. Conclusiones

# Introducción

* 1. **Propósito**

Este documento describe la visión y el alcance del desarrollo de una herramienta de **web scraping** para la empresa RascaCazuelas S.A. La solución propuesta permitirá recopilar información de la red social **X (anteriormente Twitter)** con el fin de centralizar y analizar datos de interés para el negocio.

El propósito principal es establecer las bases para el desarrollo del sistema, definiendo claramente los objetivos, funcionalidades y límites del proyecto, así como las expectativas de los stakeholders y usuarios finales.

* 1. **Alcance**

El documento cubre la visión general del proyecto, los objetivos de negocio, las características principales del sistema, el alcance de las entregas y el entorno de operación del sistema de web scraping.

La herramienta permitirá a RascaCazuelas S.A.:

 Extraer datos públicos de X (antes Twitter) de manera automatizada  Configurar reglas específicas de extracción según necesidades

 Programar extracciones periódicas

 Validar y limpiar los datos obtenidos

 Almacenar datos estructurados en la base de datos corporativa

* 1. **Definiciones y Acrónimos**

 **Web Scraping:** Técnica utilizada para extraer información de sitios web de manera automatizada

 **API:** Application Programming Interface (Interfaz de Programación de Aplicaciones)

 **HTTPS:** Protocolo seguro de transferencia de hipertexto

 **BD:** Base de Datos

 **GUI:** Graphical User Interface (Interfaz Gráfica de Usuario)

 **X:** Anteriormente conocida como Twitter, plataforma de microblogging y red social

 **Tweet:** Publicación en la plataforma X, limitada a 280 caracteres

 **Hashtag:** Palabra o frase precedida por el símbolo # que categoriza el contenido

* 1. **Referencias**

 CasoEstudio\_S1 Documento de Visión y Alcance.pdf [Classroom] <https://drive.google.com/file/d/1LOQBe3-_B99oqc11rrOwCgB39yD9l_8r/view> [Último acceso: 15 de febrero 2025].

 "¿Qué es el web scraping y para qué sirve?" Award-winning news, views, and insight from the ESET security community. [En línea]. Disponible: [https://www.welivesecurity.com/la-es/2023/07/05/que-es-](https://www.welivesecurity.com/la-es/2023/07/05/que-es-web-scraping-para-que-sirve/) [web-scraping-para-que-sirve/](https://www.welivesecurity.com/la-es/2023/07/05/que-es-web-scraping-para-que-sirve/) [Último acceso: 17 de febrero de 2025.]

 Laínez Fuentes, J. R. (2024). Gestión de proyectos de software: Metodologías, herramientas y casos prácticos (2ª ed.). RA-MA Editorial.

 Documentación de la API de X (Twitter): <https://developer.twitter.com/en/docs> [Último acceso: 28 de abril de 2025]

 IEEE Std 830-1998, IEEE Recommended Practice for Software Requirements Specifications

# Posicionamiento

* 1. **Oportunidad de Negocio**

La automatización de la recopilación de datos mediante web scraping permitirá a RascaCazuelas S.A. mejorar significativamente su eficiencia operativa, reducir costos y obtener información más precisa y actualizada para la toma de decisiones.

Los beneficios específicos incluyen:

 Reducción del tiempo necesario para la recopilación de datos de 40 horas/semana a aproximadamente 5 horas/semana

 Incremento en el volumen y diversidad de datos analizables  Capacidad para detectar tendencias en tiempo casi real

 Mejora en la calidad de los informes para clientes

 Ventaja competitiva por el acceso rápido a datos de la plataforma X

* 1. **Declaración del Problema**

|  |  |
| --- | --- |
| **Aspecto Descripción** | |
| Problema | La recolección manual de datos resulta en un proceso costoso, propenso a errores y excesivamente lento |
| Afecta a | Analistas de datos y toma de decisiones en RascaCazuelas S.A. |
| Impacto | Retrasos en la obtención de insights, costos elevados y posible pérdida de oportunidades de negocio |
| Solución | Una herramienta automatizada de web scraping que extraiga datos de manera eficiente y confiable |
| C C | |

## Consecuencias actuales:

 Los analistas dedican el 60% de su tiempo a la recopilación manual de datos  La detección de tendencias ocurre con 3-4 días de retraso

 Alto riesgo de errores humanos en la clasificación y estructuración de datos

 Limitación en el volumen de datos que pueden procesarse (actualmente 100-150 tweets diarios)

* 1. **Declaración de Posición del Producto**

|  |  |
| --- | --- |
| **Para RascaCazuelas S.A.** | |
| Quienes | Necesitan datos actualizados y confiables de redes sociales |
| El producto | Es una herramienta de web scraping automatizada |
| Que | Extrae, procesa y almacena datos de X (antes Twitter) |
| A diferencia de | El proceso manual actual |
| Nuestro producto | Ofrece automatización, mayor precisión y reducción de costos |
| C C | |

## Ventajas diferenciales:

 Capacidad para procesar hasta 10,000 tweets diarios  Actualización de datos cada 15 minutos

 Reducción del 95% en errores de clasificación

 Interfaz intuitiva que no requiere conocimientos técnicos avanzados  Integración directa con herramientas de análisis existentes

# Descripción de Stakeholders y Usuarios

* 1. **Resumen de Stakeholders**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Rol** | **Responsabilidades** | **Participación** |
| Leonardo Valdes | Gerente del  Proyecto | Coordinar el desarrollo del proyecto y  asegurar el cumplimiento de los objetivos | Supervisión general,  comunicación con directivos |
| Andrea Ortega | Desarrollador | Implementar la herramienta de web  scraping e integrar con la Base de Datos | Desarrollo técnico, pruebas y  mantenimiento |
| Alma Delia De  Jesús Islao | Profesora/Cliente | Evaluar el proyecto y proporcionar  retroalimentación | Validación de entregables,  definición de requisitos |
| Dirección General | Patrocinador | Aprobar recursos y presupuesto | Toma de decisiones  estratégicas |
| Departamento de  TI | Soporte técnico | Proporcionar infraestructura y  mantenimiento | Soporte técnico y acceso a  sistemas |
| Analistas de Datos | Usuarios finales | Utilizar los datos extraídos para análisis | Retroalimentación sobre  funcionalidad |
| C C | | | |

* 1. **Perfiles de Usuario**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tipo de Usuario** | **Descripción** | **Responsabilidades** | **Conocimientos Técnicos** |
| Administrador del Sistema | Personal de TI que  configura y monitorea el sistema | Instalación, configuración inicial, gestión de usuarios, mantenimiento | Altos (programación, bases de datos, redes) |
| Analista de Datos | Personal que define reglas  de extracción y utiliza los datos | Configuración de reglas,  programación de extracciones, análisis de datos | Medios (SQL básico, herramientas de análisis) |
| Usuario Ejecutivo | Directivos que consultan reportes basados en los  datos | Toma de decisiones basada en los insights generados | Bajos (ofimática básica) |
| C C | | | |

* 1. **Necesidades de Stakeholders y Usuarios**

|  |  |  |  |
| --- | --- | --- | --- |
| **Stakeholder/Usuario** | **Necesidad** | **Prioridad** | **Solución propuesta** |
| Analistas de Datos | Acceso rápido a datos actualizados de X | Alta | Extracción automática cada 15  minutos |
| Analistas de Datos | Filtrado eficiente por temáticas y  términos clave | Alta | Sistema de reglas personalizables |
| Dirección General | Reportes de tendencias en tiempo casi  real | Media | Dashboard integrado con datos  extraídos |
| Departamento de TI | Bajo consumo de recursos de sistema | Media | Optimización del proceso de  scraping |
| Analistas de Datos | Validación automática de datos | Alta | Algoritmos de limpieza y validación |
| Usuarios Ejecutivos | Visualización clara de insights | Baja | Informes automatizados y visuales |
| C C | | | |

* 1. **Alternativas y Competencia**

|  |  |  |
| --- | --- | --- |
| **Alternativa** | **Ventajas** | **Desventajas** |
| API oficial de X | Datos oficiales, estructurados y  confiables | Costos elevados, limitaciones en volumen,  restricciones de acceso |
| Servicios SaaS de análisis de  redes sociales | Solución lista para usar, soporte  profesional | Alto costo mensual, personalización limitada,  dependencia de terceros |
| Mantener proceso manual | Sin inversión inicial, control  humano del proceso | Lento, costoso a largo plazo, propenso a  errores |
| Subcontratación del servicio | Sin necesidad de desarrollo  interno | Costos recurrentes, posibles problemas de  confidencialidad |
| C C | | |

# Visión General del Producto

* 1. **Perspectiva del Producto**

La herramienta de web scraping será un componente integrado en la infraestructura de análisis de datos existente en RascaCazuelas S.A. Se encargará de la extracción automatizada de datos desde X (antes Twitter), su procesamiento, validación y almacenamiento en la base de datos corporativa.

El sistema funcionará como un servicio independiente que se comunicará con:

 La plataforma X (antes Twitter) para extraer datos públicos

 La base de datos corporativa para almacenar información procesada  El sistema de análisis existente para proporcionar datos estructurados

 El sistema de notificaciones para alertar sobre problemas o resultados relevantes

* 1. **Necesidades y Características**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Característica** | **Descripción** | **Prioridad** | **Justificación** | **Criterio de Aceptación** |
| CAR- 01 | Configuración de reglas de scraping | La herramienta debe permitir configurar reglas de scraping para X (antes  Twitter) | Alta | Necesario para extraer datos relevantes | El usuario puede definir al menos 5 tipos diferentes de reglas de  extracción |
| CAR- 02 | Extracción periódica | Debe realizar la extracción de datos de manera periódica y  programada | Alta | Automatización del proceso | El sistema ejecuta extracciones sin intervención humana  según programación |
| CAR- 03 | Validación de datos | Debe validar y limpiar los datos recolectados para  garantizar su precisión | Alta | Calidad de la información | Tasa de error en datos inferior al 5% |
| CAR- 04 | Integración con BD | Debe integrarse con la Base de Datos corporativa | Media | Centralización de datos | Los datos extraídos son accesibles desde las herramientas de análisis  existentes |
| CAR- 05 | Interfaz de usuario | Proporcionar una GUI intuitiva para configurar y monitorear el sistema | Media | Facilidad de uso | Usuarios pueden configurar una extracción completa en  menos de 10 minutos |
| CAR- 06 | Sistema de alertas | Notificar eventos importantes o anomalías  en los datos | Baja | Monitoreo proactivo | El sistema envía alertas por email ante  situaciones predefinidas |
| CAR- 07 | Reportes de operación | Generar informes sobre el  funcionamiento del sistema | Baja | Transparencia en operaciones | Reportes diarios  automatizados de actividad |
| C C | | | | | |

* 1. **Objetivos de Negocio**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Descripción del Objetivo** | **Métricas de Éxito** | **Tiempo Estimado** |
| ON- 1 | Automatizar la recolección de datos para  reducir costos hasta en un 40% en los próximos 6 meses | Reducción verificable del 40% en  horas-persona dedicadas a recolección | 6 meses |
| ON-  2 | Incrementar la precisión y fiabilidad de los  datos hasta en un 30% | Tasa de error en datos por debajo  del 5% (vs 25-35% actual) | 3 meses post-  implementación |
| ON- 3 | Integrar la herramienta de web scraping con la plataforma de análisis de datos existente en la  empresa | 100% de compatibilidad con sistemas actuales | 1 mes post- desarrollo |
| ON-  4 | Desarrollar la herramienta en un periodo no  mayor a 4 meses | Entrega puntual del producto | 4 meses |
| ON-  5 | Ampliar la cobertura de datos analizados en  un 500% | Capacidad para procesar al menos  10,000 tweets diarios | 2 meses post-  implementación |
| C C | | | |

* 1. **Criterios de Éxito**

Para considerar exitoso el proyecto, se deben cumplir los siguientes criterios:

## Técnicos:

 El sistema extrae datos de X (antes Twitter) con una disponibilidad del 99%

 La tasa de errores en extracción de datos es inferior al 5%

 El consumo de recursos no supera el 30% de capacidad del servidor asignado

## De negocio:

 Reducción del 40% en costos asociados a la recolección manual de datos

 Aumento del 300% en el volumen de datos analizables

 ROI positivo dentro de los primeros 8 meses

## De usuario:

 90% de los analistas consideran la herramienta "fácil o muy fácil" de usar

 Tiempo de configuración de nuevas extracciones inferior a 10 minutos

 Satisfacción general superior a 4/5 en encuestas de usuario

# Alcance y Limitaciones

* 1. **Alcance de las Entregas**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Número**  **de entrega** | **Fecha**  **estimada** | **Tema Principal** | **Características**  **incluidas** | **Hitos clave** |
| 1.0 | Final Semana 8 | Configuración y reglas básicas de  scraping | CAR-01, CAR-02, CAR-03 | - Diseño de arquitectura<br>- Interfaz básica de configuración<br>- Módulo de  extracción funcional |
| 1.5 | Final Semana 12 | Mejoras de validación y UI | CAR-03  (ampliación), CAR- 05 | - Algoritmos avanzados de  validación<br>- Interfaz completa de usuario |
| 2.0 | Final Semana 16 | Integración de la Base de Datos y  finalización | CAR-04, CAR-06, CAR-07 | - Integración con BD completada<br>- Sistema de alertas implementado<br>-  Documentación finalizada |
| C C | | | | |

* 1. **Criterios de Aceptación**

Para cada entrega, se utilizarán los siguientes criterios de aceptación:

## Entrega 1.0:

 El sistema puede conectarse a X y extraer datos básicos

 Las extracciones pueden programarse a intervalos definidos

 Los datos extraídos pasan por validación básica de formato

## Entrega 1.5:

 La validación incluye detección de duplicados y datos incompletos

 La interfaz permite todas las operaciones de configuración

 Se puede monitorizar el estado de las extracciones en tiempo real

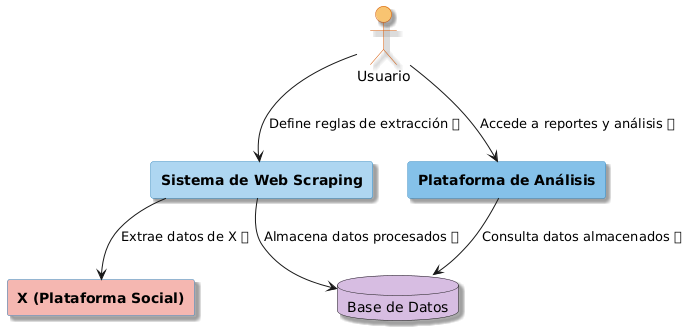
## Entrega 2.0:

 Los datos se almacenan correctamente en la BD corporativa  El sistema genera alertas ante eventos predefinidos

 Los informes de operación proporcionan información relevante

 La documentación completa está disponible para usuarios y administradores

* 1. **Diagrama de Contexto**



Flujo de información:

 Configuración de reglas → Sistema Web Scraping

 Plataforma X → Sistema Web Scraping (extracción de datos)  Sistema Web Scraping → Base de Datos (almacenamiento)  Base de Datos → Sistema de Análisis (procesamiento)

 Sistema Web Scraping → Sistema de Alertas (notificaciones)

 Sistema de Análisis → Administrador (reportes y visualizaciones)

* 1. **Entorno de Operación**

El sistema será ejecutado en un equipo doméstico y deberá soportar los siguientes requisitos:

## Navegadores web compatibles:

 Chrome (versión 120 o superior)  Firefox (versión 115 o superior)  Edge (versión 110 o superior)

 Brave (versión 1.50 o superior)

## Requisitos técnicos:

 Procesador: Intel Core i5 o equivalente (2.5 GHz o superior)  Memoria RAM: 8 GB mínimo, 16 GB recomendado

 Almacenamiento: 100 GB libres en disco  Conexiones seguras HTTPS

 Capacidad para acceder a la red social X (antes Twitter)

 Integración con la base de datos existente (MySQL 8.0 o PostgreSQL 14+)

## Capacidades del sistema:

 Ejecución en un equipo doméstico estándar  Acceso a navegadores web modernos

 Soporte para conexiones seguras HTTPS

 Capacidad para realizar extracciones periódicas y programadas de datos  Validación y limpieza de datos recolectados

 Integración con la base de datos de la empresa

 Consumo máximo de ancho de banda: 5 Mbps durante extracciones  Operación 24/7 con mantenimiento programado mensual

* 1. **Fuera del Alcance**

Los siguientes elementos están específicamente excluidos del alcance del proyecto:

 Extracción de datos de otras redes sociales diferentes a X

 Acceso a tweets o contenido privado o protegido

 Análisis semántico o de sentimiento de los datos extraídos

 Desarrollo de modelos predictivos basados en los datos

 Generación automática de informes de análisis de mercado

 Integración con sistemas CRM externos

 Capacidades de procesamiento de lenguaje natural avanzado

 Desarrollo de aplicaciones móviles para acceso al sistema

 Traducción automática de contenido extraído

# Gestión de Riesgos

* 1. **Identificación de Riesgos**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Descripción del Riesgo** | **Categoría** | **Impacto** | **Probabilidad** |
| R-01 | Cambios en la API o estructura de X que afecten la extracción | Técnico | Alto | Media |
| R-02 | Restricciones o limitaciones impuestas por X para web scraping | Legal | Alto | Alta |
| R-03 | Problemas de integración con la BD existente | Técnico | Medio | Baja |
| R-04 | Volumen de datos mayor al estimado que afecte el rendimiento | Técnico | Medio | Media |
| R-05 | Resistencia al cambio por parte de los analistas | Organizacional | Bajo | Media |
| R-06 | Retrasos en el desarrollo por complejidad técnica | Proyecto | Medio | Media |
| R-07 | Requisitos insuficientemente definidos | Proyecto | Alto | Baja |
| C C | | | | |

* 1. **Análisis Cualitativo**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Impacto en Objetivos** | **Impacto en**  **Cronograma** | **Impacto en**  **Presupuesto** | **Prioridad** |
| R-  01 | Podría impedir la extracción de datos o reducir  su calidad | 2-4 semanas de  retraso | 15-20% incremento | 1 |
| R-  02 | Podría requerir rediseño o cancelación del  proyecto | 4+ semanas de  retraso | 30-50% incremento | 1 |
| R-  03 | Funcionalidad limitada o duplicación de datos | 1-2 semanas de  retraso | 5-10% incremento | 3 |
| R-  04 | Degradación del rendimiento o necesidad de  hardware adicional | 1 semana de retraso | 10-15% incremento | 2 |
| R-  05 | Adopción limitada y ROI reducido | Sin impacto directo | Sin impacto directo | 4 |
| R-  06 | Retraso en entregas o reducción de  funcionalidades | 2-3 semanas de  retraso | 10-20% incremento | 2 |
| R-  07 | Producto final no alineado con necesidades | 3-4 semanas de  retraso | 20-30% incremento | 3 |
| C C | | | | |

* 1. **Estrategias de Mitigación**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Estrategia** | **Acción Preventiva** | **Plan de Contingencia** |
| R-  01 | Mitigar | Diseñar arquitectura modular que permita  adaptación rápida | Equipo de desarrollo asignado para  actualización inmediata |
| R-  02 | Evitar | Revisión legal de términos de servicio y consulta  con especialista | Plan alternativo utilizando API oficial con  alcance reducido |
| R-  03 | Mitigar | Pruebas de integración tempranas | Desarrollo de capa de adaptación específica |
| R-  04 | Aceptar | Diseño escalable desde el inicio | Optimización de consultas y almacenamiento |
| R-  05 | Mitigar | Involucrar a usuarios finales desde etapas  tempranas | Capacitación personalizada y soporte  dedicado |
| R-  06 | Mitigar | Planificación conservadora con margen | Priorización de funcionalidades críticas |
| R-  07 | Evitar | Validación iterativa de requisitos con  stakeholders | Ciclos de retroalimentación rápidos |
| C C | | | |

# Recursos y Planificación

* 1. **Equipo del Proyecto**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rol** | **Responsabilidades** | **Dedicación** | **Habilidades Requeridas** |
| Gerente de Proyecto  (Leonardo Valdes) | Coordinación general,  comunicación, seguimiento | 100% | Gestión de proyectos,  comunicación, planificación |
| Desarrollador Principal  (Andrea Ortega) | Diseño e implementación del  sistema | 100% | Programación, web scraping,  bases de datos |
| Asesor Técnico  (Departamento TI) | Soporte en infraestructura y  despliegue | 25% | Sistemas, redes, seguridad |
| Analista de Pruebas | Validación y testing | 50% | Testing, control de calidad |
| Representante de Usuario | Validación de funcionalidad y  usabilidad | 25% | Análisis de datos, conocimiento  del negocio |
| C C | | | |

* 1. **Cronograma Preliminar**

|  |  |  |  |
| --- | --- | --- | --- |
| **Fase** | **Duración** | **Actividades Principales** | **Hitos** |
| Iniciación | 2  semanas | Definición de requisitos, planificación inicial | Documento de visión y alcance  aprobado |
| Diseño | 3  semanas | Arquitectura, diseño de BD, interfaces | Diseño técnico aprobado |
| Desarrollo Fase  1 | 4  semanas | Implementación de núcleo y componentes  básicos | Prototipo funcional |
| Pruebas Fase 1 | 1 semana | Validación de funcionalidades básicas | Entrega 1.0 |
| Desarrollo Fase  2 | 3  semanas | Mejoras de UI y validación avanzada | Componentes integrados |
| Pruebas Fase 2 | 1 semana | Validación de mejoras | Entrega 1.5 |
| Desarrollo Fase  3 | 3  semanas | Integración BD y funcionalidades finales | Sistema completo |
| Pruebas Finales | 2  semanas | QA completo, pruebas de aceptación | Aprobación para despliegue |
| Despliegue | 1 semana | Instalación, configuración, documentación | Sistema en producción |
| C C | | | |

* 1. **Presupuesto Estimado**

|  |  |  |
| --- | --- | --- |
| **Categoría** | **Descripción** | **Monto Estimado (MXN)** |
| Personal | Salarios del equipo de desarrollo | 320,000 |
| Hardware | Servidor de desarrollo y producción | 45,000 |
| Software | Licencias y herramientas | 30,000 |
| Contingencia | Reserva para imprevistos (15%) | 59,250 |
| **Total** |  | **454,250** |
| C C | | |

# Conclusiones

RascaCazuelas S.A. necesita una herramienta de web scraping para automatizar la recolección de datos de X (antes Twitter) y mejorar su eficiencia operativa. Esta herramienta permitirá:

 Ahorrar costos significativos en el proceso de recolección de datos (reducción del 40%)

 Obtener datos más precisos y fiables (mejora del 30% en calidad)

 Aumentar el volumen de datos analizables (incremento del 500%)