

PROCESAMIENTO DE DATOS USANDO SCRAPING Y ETL PARA REALIZAR
RECOMENDACIONES PERSONALIZADAS SOBRE PRODUCTOS DE
TECNOLOGÍA

Juan Sebastián Méndez Mojica

11161824426

Giovanny Alexander Zamora Bustos

11161812822

Universidad Antonio Nariño

Programa Ingeniería de Sistemas y Computación

Facultad de Ingeniería de Sistemas

Bogotá, Colombia

2023

DEDICATORIA

Esta tesis está dedicada a mi tía Melissa Ospina y a mi abuelita Alicia Hernández (Giovanny), que, con dedicación y paciencia sobresalientes, posibilitaron la concreción de este proyecto de vida. Además, a mis padres Sandra Mojica y Edgar Mendez (Sebastián) quienes, con su apoyo y cariño incondicional, fueron pieza fundamental para seguir adelante ante cualquier adversidad presentada en el proceso.

AGRADECIMIENTOS

Infinitas gracias a todas aquellas personas que han hecho parte de este proceso, a nuestros colegas por su compromiso y su interés por enseñar, a la universidad por permitirnos ser parte de ellos, al profesor Jhonatan Rico por brindarnos su conocimiento y finalmente, a nuestros amigos más allegados.

De igual manera, agradecemos a nuestro directo de tesis Juan Camilo Ramírez por su tiempo, dedicación y compromiso durante todo el proceso del trabajo de grado.

CONTENIDO

Introducción.....	1
1. Planteamiento del problema.....	3
1.1. Descripción del problema.....	3
1.2. Formulación del problema.....	10
1.3. Objetivos.....	11
1.3.1. Objetivo general.....	11
1.3.2. Objetivos específicos.....	11
1.4. Justificación.....	12
1.5. Alcance y limitaciones del proyecto.....	13
1.5.1. Alcance.....	13
1.5.2. Limitaciones.....	14
2. Metodología.....	15
2.1. Scrum.....	15
2.2. Roles.....	15
2.3. Ritos.....	16
2.4. Artefactos.....	17
3. Marco de referencia.....	18
3.1. Marco Teórico.....	18
3.1.1. Scraping.....	18
3.1.2. Etl.....	18
3.1.3. Software libre.....	19
3.1.4. World wide web.....	19
3.2. Estado del arte.....	20
3.3. Marco legal.....	23
4. Desarrollo del proyecto.....	24
4.1. Levantamiento de información.....	24
4.1.1. Arquitectura del sistema.....	24
4.2. Análisis del sistema.....	25
4.2.1. Requerimientos funcionales.....	25
4.2.2. Requerimientos no funcionales.....	26

4.2.3. Casos de uso.....	26
4.3. Diseño del sistema.....	28
4.3.1. Mockups.....	28
4.3.2. Diagramas de secuencia.....	29
4.3.3 Historias de usuario.....	31
4.4. Base de Datos.....	33
4.4.1 Modelo entidad relación.....	33
4.5. Pruebas.....	34
4.5.1. Sprints.....	38
5. Resultados obtenidos.....	42
5.1. ScrapMaster.....	42
6. Conclusiones y recomendaciones.....	48
7. Bibliografía.....	49

LISTA DE FIGURAS

Figura 1: Primera pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma uru para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.....	4
Figura 2: Segunda pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma uru para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.....	5
Figura 3: Tercera pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma uru para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.....	5
Figura 4: Cuarta pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma uru para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.....	6
Figura 5: Quinta pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma uru para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.....	6
Figura 6: Sexta pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma uru para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.....	7
Figura 7: Séptima pregunta la cual es opcional.....	7
Figura 8: Resultado final después de haber respondido todas las preguntas.....	8
Figura 9: Página de celulares el cual contiene más de 350 productos.....	9
Figura 10: Arquitectura del software, patrón mvc modelo vista controlador.....	24
Figura 11: Caso de uso entre el administrador y el usuario final.....	27
Figura 12: Prototipo de la página de inicio para los usuarios.....	29
Figura 13: Diagrama de secuencia sobre responder preguntas en scrapmaster...	30
Figura 14: Diagrama de secuencia al ejecutar los scripts de scraping.....	30
Figura 15: Diagrama de secuencia al ejecutar los scripts de limpieza de datos....	31
Figura 16: Modelo entidad-relación simple con la tabla de celulares.....	34
Figura 17: Manejo de tareas en trello.....	38
Figura 18: Consideraciones importantes para los scripts de scraping en trello....	39
Figura 19: Consideraciones importantes para la limpieza de datos en trello.....	40
Figura 20: Consideraciones importantes para realizar scrapmaster en trello.....	41
Figura 21: Presentación página scrapmaster.....	42
Figura 22: Primera pregunta de scrapmaster.....	43
Figura 23: Segunda pregunta scrapmaster.....	44
Figura 24: Tercera pregunta de scrapmaster.....	44
Figura 25: Cuarta pregunta de scrapmaster.....	45
Figura 26: Resultado de responder las preguntas de scrapmaster.....	45
Figura 27: Resultado al no encontrar coincidencias en scrapmaster.....	46

Figura 28: Página oficial de uru.....46
Figura 29: Datos de contacto y qr a documento.....47

LISTA DE TABLAS

Tabla 1: Cronología de motores de búsqueda.....	21
Tabla 2: Herramientas de scraping y sus características.....	22
Tabla 3: Requerimientos funcionales.....	25
Tabla 4: Requerimientos no funcionales.....	26
Tabla 5: Historia de usuario scraping.....	32
Tabla 6: Historia de usuario limpieza de datos.....	32
Tabla 7: Historia de usuario aplicación web.....	33
Tabla 8: 1º Caso de prueba ejecución del scraping.....	35
Tabla 9: 2º Caso de prueba ejecución del scraping.....	35
Tabla 10: 3º Caso de prueba ejecución del scraping.....	36
Tabla 11: 1º Caso de prueba limpieza de información.....	36
Tabla 12: 2º Caso de prueba limpieza de información.....	37
Tabla 13: Caso de prueba obtener recomendaciones de scrapmaster.....	37

RESUMEN

Ádalo Tech, una empresa que comercializa productos de hardware y software, actualmente cuenta con una línea de negocio llamada URU que les ofrece a los clientes una búsqueda guiada del producto que desean a partir de un formulario en el cual deben contestar una serie de preguntas. No obstante lo anterior, las mediciones internas llevadas a cabo por la empresa evidencian que este proceso les puede tomar varias horas a los clientes, un problema que se ve acentuado con la variabilidad en el catálogo de productos y servicios. Para abordar esta problemática, este trabajo de grado consiste en el desarrollo de una plataforma de asistencia a los clientes durante el proceso de compra, en la cual se les guía con información recopilada utilizando técnicas de data scraping sobre un Marketplace para ofrecer a cada uno un informe personalizado, ofreciendo la mejor opción según sus requisitos. Las pruebas funcionales realizadas evidencian que la plataforma desarrollada permite ahorrar dinero y maximiza los beneficios de la compra del cliente.-

Introducción

La empresa Ádalo Tech está posicionada en el mercado colombiano para ofrecer productos tecnológicos por medio de una plataforma tecnológica donde los usuarios reciben recomendaciones sobre los productos que mejor atienden sus necesidades y donde pueden realizar hacer las compras de los mismos. A pesar de esta asesoría automatizada, los indicadores de la empresa señalan que a los clientes les puede tomar varias horas encontrar y comprar el producto que desean. Uno de los factores que contribuyen a este problema es la amplia variedad de productos, marcas y especificaciones, información con la que no todos los clientes están familiarizados. En algunos casos, las personas se dejan llevar más por la marca y el precio que por las características y sus funciones, haciendo que la inversión no sea la más adecuada.

Por lo anteriormente expuesto surge el proyecto URU de la compañía Ádalo Tech. URU es una línea de negocio que busca brindar las mejores opciones para sus clientes.

Se planea ayudar a Ádalo Tech en la solución de esta problemática. El tiempo de búsqueda de productos de tecnología es muy largo actualmente, ya que toma 1 semana aproximadamente el proceso de búsqueda. Se tomo como caso de pruebas un producto tecnológico el cual es el celular, mirando su precio, colores y demás características, esto tomando de ejemplo la página de Falabella, la cual cuenta con más de seis mil resultados.

Este trabajo de grado consiste en el desarrollo de un software en el cual se brindará la solución a Ádalo Tech, con el fin de optimizar los tiempos de búsqueda del proyecto de URU, beneficiando a sus clientes al momento de hacer su compra, brindándole al usuario la mejor opción a partir de una serie de preguntas. Las preguntas se realizan en relación de conocer un poco mejor al cliente y las necesidades que tiene, esto con el fin de que ahorre dinero. Siendo lo más eficiente y eficaz posible, la metodología de desarrollo de software que se empleó fue una metodología ágil (Scrum), accediendo a los datos del Marketplace de Falabella usando técnicas de scraping para automatizar los procesos de recolección de datos y de esta manera se cuenta con otra fuente de datos aparte de las que ya cuenta Ádalo Tech con otros Marketplace. Mejorando las comparaciones de los productos de tecnología que se encuentran en la actualidad siendo la búsqueda muy rápida. Se busca utilizar herramientas de open source. Por tal motivo se va a realizar scraping con el lenguaje de programación Python,

usando librerías de código fuente abierto, pasando por una serie de procesos y técnicas como lo son las ETL (Extract Transform Load).

Por lo anterior, se realizó una aplicación web donde se pueden ver los resultados obtenidos de las consultas realizadas por los usuarios de URU, integrando la información directamente con la interfaz de usuario, utilizando ETL para la base de datos correspondiente. Esto tuvo un impacto en los usuarios, ya que reciben su recomendación de inmediato en vez de hacer una compra normalmente la cual dura horas; debido a esto ahorran tiempo. La estructura de este documento se organiza de la siguiente manera para proporcionar una comprensión completa del proyecto. En el Capítulo 1, se aborda el Planteamiento del Problema, donde se identifican y analizan las cuestiones clave que motivaron este estudio. El Capítulo 2 se dedica al Marco Conceptual, donde se presenta una revisión exhaustiva de la literatura que respalda el marco teórico del proyecto. En el Capítulo 3, se describe detalladamente la Metodología utilizada para llevar a cabo la investigación, incluyendo los métodos de recopilación de datos y los procedimientos empleados. A continuación, en el Capítulo 4, se presenta el Desarrollo del Proyecto, destacando los pasos y las acciones ejecutadas durante su implementación. En el Capítulo 5, se revelan los Resultados Obtenidos, junto con sus análisis e interpretaciones. Las Conclusiones, que se presentan en el Capítulo 6, resumen los hallazgos clave y su relevancia. Finalmente, en el Capítulo 7, se proporciona una lista detallada de las Bibliografías que respaldaron la investigación y enriquecieron su contexto.

1. Planteamiento del problema

1.1. Descripción del problema

La empresa Ádalo Tech ofrece productos tecnológicos y se especializa en desarrollo de software. Es relativamente nueva, creada en enero del año pasado; sus fundadores cuentan con más de 6 años de experiencia en el mercado, realizando ejercicios de Arquitectura Empresarial y en Gerencia de Proyectos de desarrollo de software, participando en todas las etapas de desarrollo de productos y servicios de base tecnológica. Dentro de sus operaciones tiene una línea de negocio de innovación llamada URU. URU tiene la finalidad de facilitar las compras de tecnología al usuario final y apoyarlo en las decisiones de compra, basándose en una plataforma en línea donde los usuarios pueden realizar búsquedas y compras en línea. De acuerdo a una encuesta interna la compañía está interesada en mejorar el tiempo de los procesos al momento de realizar las recomendaciones de búsqueda. Los clientes de la compañía usan términos de búsqueda no tan conocidos o que pueden confundir al público en general y a la gran variedad de productos que se encuentran en el mercado. Esto hace necesario contar con una solución de software enfocada en ayudar al usuario común sin tantos conocimientos técnicos para comprar productos de tecnología, por medio de recomendaciones y búsquedas guiadas según sus necesidades particulares.

Actualmente, Ádalo Tech lleva a cabo un proceso de recopilación de datos desde fuentes en línea de manera manual y exhaustiva de acuerdo a ciertos productos. Este proceso manual consiste en que, de acuerdo al producto seleccionado por el usuario, se empiezan a aplicar ciertos filtros dentro de la página para que posteriormente se vayan reduciendo los resultados y de esta manera obtener el resultado óptimo. Este proceso implica la extracción de información específica de una variedad de páginas web seleccionadas como lo son Alkosto, Amazon, Linio, etc. La información recopilada es usada para realizar las comparaciones entre los distintos mercados y dar la mejor opción.

De la figura 1 a la 7 es el paso a paso de las preguntas que se pueden responder desde la página de URU¹, la figura 7 corresponde a una pregunta opcional; las demás son obligatorias, esto con el fin de recolectar la información y dar las mejores opciones por medio de un formulario de preguntas puntuales que realiza la función de filtro para seleccionar el producto que mejor se adapta dependiendo la necesidad de los clientes de Ádalo Tech. A continuación, se observarán más detalles:

Fuente: elaboración propia con datos de la plataforma de Ádalo Tech.

1 2 3 4 5 6 7

Usaré el celular principalmente para...
Puedes seleccionar varias opciones

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Entrar a redes y Whatsapp	Videollamadas	Transmitir en streaming	Escuchar música y ver videos	Tomar fotos y videos	Ver y editar documentos	Juegos Sencillos	Juegos de alto desempeño

Siguiente >

Figura 1: Primera pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma URU para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.

¹ Página de URU, <https://www.uru.com.co/>

Fuente: elaboración propia con datos de la plataforma de Ádalo Tech.

1 2 3 4 5 6 7

Durante el día uso el celular... *

En todo momento

De 4 a 6 horas

Máximo dos horas

Muy poco al día

< Anterior Siguiente >

Figura 2: Segunda pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma URU para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.

Fuente: elaboración propia con datos de la plataforma de Ádalo Tech.

1 2 3 4 5 6 7

Guardaré principalmente... *

Puedes seleccionar varias opciones

Imágenes y fotos de recuerdos

Mis videos favoritos

Mi música preferida

Juegos para relajarme y disfrutar

Mis documentos de trabajo y estudio

Mis e-Books

< Anterior Siguiente >

Figura 3: Tercera pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma URU para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.

Fuente: elaboración propia con datos de la plataforma de Ádalo Tech.

1 2 3 4 5 6 7

Prefiero los celulares...*

Puedes seleccionar varias opciones

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
				
iPhone	Android	Huawei	Xiaomi	Otros

< Anterior Siguiente >

Figura 4: Cuarta pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma URU para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.

Fuente: elaboración propia con datos de la plataforma de Ádalo Tech.

1 2 3 4 5 6 7

Lo más importante del celular para mí es...*

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
					
			Que tenga		

1 2 3 4 5 6 7

Tengo un presupuesto máximo de...*

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Menor de \$500.000	\$500.000 - \$999.999	\$1.000.000 - \$1.999.999	\$2.000.000 - \$2.999.999	Más de \$3.000.000

< Anterior Siguiente >

Figura 5: Quinta pregunta y opciones de respuesta que aparecen en el formulario utilizado actualmente por la plataforma URU para filtrar mejores resultados de productos al momento de realizar la compra de celulares según su uso.

Luego de responder cada una de las preguntas por el usuario, el sistema dará las me
me
COR

Fuente: elaboración propia

Estas son las mejores recomendaciones para ti

	Huawei P30 Pro	None V20	Huawei Nova 5T
descripción	El Huawei P30 Pro es un teléfono inteligente de gama alta de 2019. Tiene una pantalla de 6.47 pulgadas, una cámara cuádruple de 40 MP + 20 MP + 8 MP + TOF 3D, 8 GB de RAM y 256 GB de almacenamiento. Está equipado con una batería de 4200 mAh con carga rápida y compatible con carga inalámbrica.	El None V20 es un teléfono inteligente con Android 9 Pie de gama alta de 2019. Está equipado con una pantalla de 6.4 pulgadas, una cámara de triple lente de 48 MP + 8 MP + 5 MP y 8 GB de RAM. Está alimentado por una batería de 4000 mAh y tiene una ranura para tarjeta microSD.	El Huawei Nova 5T es un teléfono inteligente con Android 9.0 Pie de 2019. Está equipado con una pantalla de 6.26 pulgadas, una cámara cuádruple de 48 MP + 16 MP + 2 MP + 2 MP y 8 GB de RAM. Está alimentado por una batería de 3750 mAh y tiene una ranura para tarjeta microSD.
comentarios buenos	Gran diseño, pantalla vibrante, excelente cámara y lentes, buena batería, carga rápida y carga inalámbrica.	Gran pantalla, buena cámara, rendimiento excelente.	Gran pantalla, buena cámara, carga rápida y excelente rendimiento.
comentarios malos	El software de Huawei no es muy amigable para usuarios de Android, disponible solo en algunos países.	El diseño no es muy atractivo y la batería podría ser un poco mejor.	La batería no es muy duradera.
ventajas	Es un celular con una gran potencia, adecuado para ver videos, guardar documentos y tener una excelente experiencia de usuario.	Es un equipo con gran rendimiento, adecuado para ver videos, guardar documentos y disfrutar de una gran experiencia de usuario.	Es un equipo con gran rendimiento, adecuado para ver videos, guardar documentos y disfrutar de una gran experiencia de usuario.

Fuente: elaboración propia con datos de la plataforma de Ádalo Tech.

1 2 3 4 5 6 7

Además, me gustaría que tuviera...

Estas preguntas son opcionales, contéstalas solo si alguna es de tu preferencia

Tamaño

Pequeño, máximo de 5 pulgadas

Mediano, de 5 a 6 pulgadas

Grande, más de 6 pulgadas

Adicionales

Lector de huellas

Reconocimiento facial

Resistencia al agua

¿Algo más que quisieras que tuvieramos en cuenta?

Figura 7: Séptima pregunta la cual es opcional.

Figura 8: Resultado final después de haber respondido todas las preguntas.

La empresa ha tomado estadísticas internamente a partir de la plataforma en línea, se ha determinado que les tomaba a los clientes el proceso de selección del mejor producto 5 horas. Este proceso se debía mejorar, porque era lento al realizar el proceso manualmente y requería recursos de personas.

Se realizaron ejemplos prácticos tomando los celulares como producto de interés, consultando la información de la página web de K-Tronix, ya que de allí se recopilará la información. Posteriormente será tratada por URU y finalmente se dará el producto más óptimo. Este es uno de los Marketplace principales para comprar tecnología a nivel nacional; al momento de hacer la búsqueda hay más de 350 productos diferentes y aproximadamente para revisar cada una se demora alrededor de 5 minutos, validando todas sus características, tal cual lo muestra la figura 9. Esto tomaría alrededor de unas 5 horas el realizar una sola asesoría, haciendo que el proceso sea bastante demorado.

Fuente: elaboración propia con datos de la plataforma de Ktronix.

The screenshot shows the K-Tronix website interface. At the top, there is a navigation bar with the K-Tronix logo and the tagline 'pasión tecno'. Below the logo is a search bar with the placeholder text 'Buscar por producto, categoría, marca...'. To the right of the search bar are icons for 'Mi cuenta' and 'Mi carrito'. The main navigation menu includes categories like 'Computadores e Impresoras', 'TV', 'Audio', 'Electrodomésticos', 'Videojuegos', 'Cámaras', 'Casa Inteligente', 'Accesorios', 'Netflix y Pines', 'Deportes y Smartwatch', 'Ofertas', and 'Otros'. The featured product is a 'Celular HONOR X7A 6GB+128GB Azul'. The product listing includes a star rating of 4.5 (3 reviews), a list of specifications: 'Resolucion Camara Frontal 1 8 Mpx', 'Tipo de Camara Posterior Cuadruple', 'Tipo de Camara Frontal Sencilla', 'Memoria Interna 128 GB', and 'Memoria RAM 6 GB'. The price is shown as '\$669.900' with a crossed-out original price of '\$1.499.000' and a 'Hoy' price of '\$749.900'. A 'Ver más detalles' link is present below the product image. At the bottom of the page, there is a button that says 'Mostrar más productos' and a note that says 'Has visto 25 de 357 productos'.

Figura 9: Página de celulares el cual contiene más de 350 productos.

Es por lo anterior que en este trabajo de grado se planteó como solución la recopilación de datos de una manera más rápida y fácil, esto con el fin de realizar comparaciones con otros Marketplace lo cual es esencial para tomar decisiones informadas al evaluar productos, optimizar costos, acceder a una variedad más amplia de opciones y garantizar la confiabilidad. Esto es crucial para la empresa, ya que les permite encontrar las mejores ofertas, no solo de Ktronix el cual es un ejemplo práctico, sino de otras páginas importantes como lo son Alkosto, Éxito y Jumbo, encontrando productos únicos, oportunidades de expansión y mantenerse competitivos en el mercado. Las comparaciones son fundamentales para una experiencia de compra segura y satisfactoria, y también impulsan la mejora continua en el mundo del comercio electrónico. Realizando este proceso con la página de Falabella, la cual se obtendrá la información, para que posteriormente la compañía maneje estos datos y tenga una gama más amplia de mercados, siendo para Ádalo Tech más fácil el proceso de la recolección y comparación de datos, para que el desarrollo sea transparente para el usuario final.

Por lo tanto, se planteó un producto de software conformado por un conjunto de scripts utilizando el lenguaje de programación Python con el fin de que el proceso que se realiza manualmente sea más rápido. El software hará el proceso automáticamente arrojando resultados según las especificaciones dadas por el usuario, utilizando los datos obtenidos para realizar una comparación frente a la necesidad del proceso de la compra tecnológica. El sistema propuesto recolectó los datos de manera automática haciendo que el rendimiento sea más óptimo, con el fin de almacenar la información en la base de datos que se creará a partir de la información obtenida, permitiendo consultar y filtrar el reporte directamente desde la aplicación web, optimizando procesos y recursos.

1.2. Formulación del problema

Siendo URU una de las líneas de negocio de Ádalo Tech, se debe ocupar personal para llevar a cabo el proceso de asesoría. En el momento, URU ofrece este servicio por su página web contestando una serie de preguntas y posteriormente cobrando una tarifa para que el usuario reciba un reporte a su correo electrónico con las mejores opciones de productos tecnológicos según las preguntas respondidas. Actualmente el tiempo que toma al realizarse el proceso de la recolección de datos para una asesoría toma alrededor de unas 5 horas para un solo producto de tecnología de cierto Marketplace, haciendo que al día solo se puedan atender 2 solicitudes. Es por esto que se desea mejorar los tiempos para que URU sea escalable en el futuro.

Debido a lo anterior se plantea solucionar la pregunta:

¿Cómo reducir el tiempo del proceso de recolección y procesamiento de datos para Ádalo Tech en cada asesoría gestionada por medio de URU, mejorando así la capacidad operativa actual de la empresa?

1.3. Objetivos

1.3.1. Objetivo general

- Desarrollar una plataforma en línea llamada ScrapMaster, para mostrar los productos tecnológicos (Celulares) optimizando las búsquedas de los clientes de Ádalo Tech, en el Marketplace de Falabella, usando técnicas de scraping y ETL con datos públicos, mejorando así el tiempo en el cual se le brindan las recomendaciones que realiza URU sobre productos de tecnología teniendo en cuenta las necesidades del usuario.

1.3.2. Objetivos específicos

- Diseñar la base de datos donde se almacenarán los datos obtenidos por medio de scraping usando el modelo entidad relación.
- Codificar los scrapers que permitan recopilar información de un Marketplace que vende productos de tecnología usando librerías del lenguaje Python.
- Recopilar los datos del Marketplace almacenándolos en la base de datos diseñada usando técnicas de ETL y asegurando la calidad de estos.
- Desarrollar el módulo web que permita que las personas realicen consultas sobre el conjunto de datos almacenados en la base de datos.

1.4. Justificación

Este proyecto ayudará a Ádalo Tech a que tenga una mejor capacidad operativa, debido a que actualmente el proceso de encontrar el mejor producto tecnológico tomaba mucho tiempo, lo cual hacía que realizar una asesoría tomará más de 5 horas estando expuesto a variaciones por cantidad de información, el método de recolectar la información era “manual”.

En la actualidad este proyecto beneficiará a las personas que no tienen los suficientes conocimientos en las áreas tecnológicas, por lo cual terminan comprando productos sin sacarles el mayor provecho a todas las utilidades, características, funciones compradas o incorporadas en el equipo. Al realizar uso de la herramienta URU de Ádalo Tech se reduce el tiempo del proceso de búsqueda.

A corto plazo, en el ámbito económico, Ádalo Tech se beneficiará con la línea de negocio de URU, ya que este proyecto no representará costo para la empresa, mejorando sus procesos de recolección de datos siendo automatizados. A mediano plazo Ádalo Tech reducirá costos de recursos como personas, tomará de base el desarrollo que se hará y lo usará en otros Marketplace para obtener los datos de otros productos; a largo plazo Ádalo Tech logrará expandirse en la capacidad operativa mejorando el rendimiento de asesorías para los clientes de la compañía.

En el ámbito tecnológico las herramientas empleadas son el uso de lenguaje de programación Python un lenguaje de análisis de datos muy popular en la actualidad, de librerías open source. Al hacer esta implementación usando el lenguaje Python, hará que la herramienta sea sostenible en el tiempo, debido a su portabilidad y al ser un lenguaje de alto nivel. Se usarán bibliotecas para hacer más fácil su implementación.

Este proyecto fue un gran reto para seguir en proceso de aprendizaje, teniendo en cuenta que el impacto es más positivo, adquiriremos más experiencia y podemos desempeñarnos en la rama de datos, y ciencia de datos, aplicando los nuevos conceptos y conocimientos de la práctica.

1.5. Alcance y limitaciones del proyecto

1.5.1. Alcance

La aplicación cuenta con las siguientes funcionalidades:

- **Funcionalidad 1 (Scripts de Scraping):** Extraerá la información usando Python como herramienta principal.
 - Recopilación de datos de un Marketplace, en este caso Falabella, con un solo producto tecnológico.
 - Almacenar la información en una base de datos.
- **Funcionalidad 2 (Scripts de transformación de datos):** Se integrarán los datos que no están optimizados, es decir que no cuentan con reglas de calidad, es decir información que contenga campos en blanco o caracteres incorrectos, para que puedan ser tratados correctamente.
 - Extraer los Datos en Bruto (raw).
 - Aplicar técnicas de ETL para asegurar la calidad de los datos.
- **Funcionalidad 3 (Interfaz de usuario para la consulta y visualización de productos):** Mostrará la diferente información obtenida por medio del scraping para que las personas realicen consultas.
 - Publicar el reporte en la web.

Tecnológicamente el proyecto usará:

- **Front-end:** HTML, CSS y Javascript
- **Backend:** Scraping: – Lenguaje Python, usando librerías
- **Bases de datos:** MySQL

1.5.2. Limitaciones

- Solo se incluirá un módulo web en la aplicación con el reporte de la información detallada a partir del desarrollo del scraping.
- El desarrollo para dicho objetivo partirá de solo un producto tecnológico y un solo Marketplace.
- La empresa nos dispondrá de las diferentes herramientas para desarrollar el proyecto de software, como son las bases de datos, el WordPress y las librerías de Python.
- Los gastos del hosting y del dominio serán cubiertos por la empresa.

2. Metodología

2.1. Scrum

Es una metodología de desarrollo ágil que inició en los años 80 donde se resalta el trabajo en equipo. Se busca la satisfacción del cliente, levantando requerimientos funcionales Scrum involucra al cliente en el proceso de desarrollo: el cliente define qué se hace y cuando se hace, el equipo es el que se compromete en las entregas de Sprint. Scrum cuenta con eventos los cuales se deben cumplir para la inspección. Estos eventos son de seguimiento para cumplir con los objetivos, realizando reuniones para hacer seguimiento al proceso de cada Sprint. Esta metodología está expuesta a la adaptabilidad porque se realizan mejoras continuas buscando todo el tiempo la satisfacción del cliente, siendo incremental la mejora (Rodríguez & Dorado, 2015).

2.2. Roles

El Scrum Master o también conocido como facilitador de proyectos, es el encargado de que los grupos de trabajo alcancen sus objetivos hasta llegar a la fase final del sprint, siguiendo las prácticas y objetivos según la metodología scrum (Canal, 2022).

El product owner es la persona encargada de darle valor al producto desarrollado por la compañía, haciendo las mejoras necesarias para su producto y que éste siempre este a un alto nivel (Scrum: roles y responsabilidades).

Los desarrolladores suelen estar formados de 3 a 9 personas que se encargan de desarrollar el producto, gestionando sus tareas para entregar un incremento en el software al final de cada sprint; todo esto a partir del product Backlog seleccionado durante todo el Sprint Planning (Scrum: roles y responsabilidades).

Esto aplicado en el proyecto se ve de la siguiente manera:

- SCRUM MASTER: Representante del trabajo de grado asignado por el CNTG (Comité Nacional de Trabajos de Grado).
- PRODUCT OWNER: Raissa Daniela Quintero, CEO de Ádalo Tech.
- DESARROLLADORES: Sebastián Méndez y Giovanni Zamora

2.3 Ritos

El Sprint es en donde se realizan todas las acciones y se testea si las acciones realizadas funcionan correctamente (Canal, 2022).

El Sprint review es una reunión para validar los avances y el progreso después de cada Sprint, esto con el fin de mirar si se cumplen los objetivos establecidos en un futuro (Garcia, 2020).

El Sprint Retrospective es una reunión periódica que se realiza al final de cada Sprint para analizar los puntos positivos y negativos de este ciclo y ver mejoras para el próximo Sprint (Adobe Communications Team, 2022).

Esto aplicado en el proyecto se ve de la siguiente manera:

- Daily: En este caso no se tuvo una sesión diaria, sino una reunión cada semana para validar el estado de los procesos.
- Sprint Planning: Se tuvo una sesión al inicio de cada Sprint para definir el Product Backlog y lo que se va a entregar en cada Sprint.
- Sprint Review: Al final de cada Sprint se revisó lo que se logró en cada Sprint, mostrando los resultados del trabajo por parte de los desarrolladores.
- Retrospective: Se trata de una reunión en la que se evalúan los resultados del Sprint, mirando posibles mejoras para las próximas entregas, teniendo claro que tuvimos 3 Sprints.

2.4. Artefactos

El Product Backlog es el archivo que recoge las tareas y funciones a desarrollar a lo largo del proyecto (Canal, 2022).

El Burn Down es el análisis y control de las tareas ejecutadas y todo lo que queda pendiente (Canal, 2022).

El Sprint Backlog es el documento que muestra la división de tareas entre los miembros del equipo (Canal, 2022).

Esto aplicado en el proyecto se ve de la siguiente manera:

Se manejó un Trello donde se llevó las tareas que se deben desarrollar durante todo el proceso de desarrollo de software, además de un documento con los avances que cada uno de los desarrolladores ha realizado.

3. Marco de referencia

3.1. Marco Teórico

3.1.1. Scraping

Scraping es un proceso de técnicas automatizadas para recolectar datos públicos de páginas web, (López, 2018) que será aplicadas las técnicas de Scraping para la empresa Ádalo tech enfocada en la línea de negocio URU. Se debe realizar un proceso de estructuración de datos para ejecutar un algoritmo de extracción y procesamiento de datos. Se analizan los datos para extraer la información necesaria. Las técnicas por web scraping representaría al proceso de recolectar datos manualmente siendo automatizados todos los procesos, como por ejemplo seleccionar el texto, copiarlo y luego pegarlo de manera automática (López, 2018).

3.1.2. Etl

ETL significa Extraer Transformar y Cargar datos. Hoy en día es utilizado para extraer información de algún sitio en la web, transformar esos datos para luego ser cargada a otro aplicativo o software. La manera en la que almacena la información es un data mart siendo una base de datos extensa, resguardando información de interés para el área. Se cuentan con varios pasos en la etapa de transformación de datos como lo son: aplicación de reglas comerciales, limpieza, filtrado, división, unión, transposición y aplicar la validación de datos (Pott, 2018). ETL (Extract, Transform and Load) es el proceso que permite mover datos de múltiples fuentes, reformatearlos y limpiarlos, cargarlos a otras bases de datos. Es un conjunto de técnicas cuyos procesos automatizan los procesos manuales, siendo más eficiente en el tiempo al utilizar algoritmos para la extracción el análisis, la transformación y la carga de datos relevantes en la base de datos (Loaiza, 2020).

En la primera etapa se extraen los datos relevantes, en la mayoría de los casos sin formato. La transformación requiere limpiar los datos, haciendo controles de calidad, para cumplir con el estándar dónde será almacenada la información. En este paso se realiza la eliminación de duplicados, se filtra según se defina por el interesado, se clasifica y agrupa la información. La última etapa de carga implica llevar los datos a la base de datos MySQL para ser consumida por el cliente.

3.1.3. Software libre

Desde 2013, uno de los debates más significativos en el campo del scraping se ha centrado en el uso del software libre, que ofrece acceso al código fuente y permite su uso, ejecución, modificación y distribución. A partir de 1991 se realiza el lanzamiento del primer núcleo Linux por Linus Torvalds. En el mismo año Guido van Rossum libera la primera versión del lenguaje de programación Python. Los lenguajes de programación son la herramienta indispensable de construcción de los programas de software. Python se destaca y gana por su sencillez de aprendizaje y por la posibilidad de concentrarse en problemas actuales. Con este lenguaje se facilita el uso educacional y científico y permite el uso de librerías de software libre (Challenger et al., 2014).

3.1.4. World wide web

La web se ha convertido en un medio de comunicación indispensable convirtiéndose en la principal fuente de información para obtener conocimiento. Cuenta con más de 6 mil millones de páginas sólo en su parte web pública indexada. La (AI) arquitectura de información presenta un problema principal la cual es la organización. Todos los días se tendrán visitas públicas de muchas personas accediendo por primera vez al sitio web de Ádalo Tech. El diseño debe ser intuitivo. Por esta razón se ha utilizado la arquitectura tradicional en la creación de la página web para la empresa, utilizando la herramienta WordPress y PHP. La web (HTTP) es un servicio muy reclamado para publicar información. Es uno de los protocolos más usados en el mundo (Baeza et al., 2004). WordPress es uno de los gestores de contenidos (CMS) más conocido. Se enfoca en la creación de blogs, páginas corporativas, tiendas virtuales, y mucho más. Es de fácil manejo. El panel de administración permite crear webs modernas y de gran utilidad. Su desarrollo se basa en PHP y MySQL. Una de las mayores ventajas es su código modificable y la inserción de mejoras extras al sitio web (Sánchez, 2021).

Por lo anterior se crea la necesidad de llegar a lugares donde no es fácil llegar el comercio. A partir de los 70's se empezaron a implementar métodos de comercio electrónico en línea: B2B (transacciones entre compañías) y B2C (ventas a clientes). Para 1995 nace AMAZON, lo que sería el primer concepto de Marketplace, seguida de EBAY. En ese punto empieza a ganar fuerza y nace una nueva forma de comercio. Un Marketplace es una plataforma online donde compradores,

vendedores y distribuidores se encuentran para intercambiar información y llevar a cabo operaciones comerciales. Se centra en transacciones online de empresa a empresa. Los tipos de Marketplace son: productos, servicios y trabajos (Rodríguez, 2021).

3.2. Estado del arte

A medida que ganó popularidad el uso de web scraping, se generaron controversias con sobre la ética y la legalidad. Algunos sitios prohíben el scraping en sus términos de servicio, mientras que otros lo permiten bajo ciertas condiciones (Qayoom, s.f.).

Se ha desarrollado técnicas más avanzadas para el scraping web, como el uso de proxies y VPNs para evitar bloqueos IP, el uso de encabezados de solicitud para emular un comportamiento humano y evitar detección, y el uso de algoritmos de aprendizaje automático para analizar y estructurar datos extraídos (ChatGPT). Una de las herramientas de pago que realiza scraping es Elasticsearch, Google Cloud, extensiones que incorporan la funcionalidad de Scraping como Selenium, Jira, Scrapyng web.

Para obtener la información y los datos se utilizan motores de búsqueda “arañas” que leen el contenido para la indexación y registran enlaces (Wall, 2017). Las herramientas para realizar web Scraping en la actualidad son el uso de lenguajes de programación: el más usado en la actualidad es Python por su versatilidad en el manejo de datos y por su facilidad de aprendizaje al utilizar este lenguaje que nos brinda librerías open source (Wall, 2017).

En la actualidad el web scraping es la práctica de extraer información de manera automatizada por medio de scripts que hacen uso de herramientas como bibliotecas open source siendo las más populares: Beautiful Soup, Scrapy, las dos con el lenguaje Python, Jira y Selenium para la automatización y exploración de navegadores, Puppeteer (Node.js), entre otras. Existen herramientas de pago como: Elasticsearch.

Fuente: Adaptado de (Krijnen, 2014) y (Wall, 2017)

Motores de Búsqueda	Años	Método
WebCrawler	1994	Spiders
Lycos	1994	Spiders
Infoseek	1994	Búsqueda Jerárquica
Inktomi	1996	Spiders
Ask.com	1997	Multimedia
Overture	1998	Spiders Rastreadores Web
AllTheWeb	1999	Spiders Rastreadores Web
Salesforce.com API-Ebay vs bidder Edge	2000	Spiders Indexado
Altavista	2003	Búsqueda-Meta
Anuncios de Google	2003	Búsqueda-Meta
Beautiful soup	2004	Scraping-Búsqueda Indexada
Búsqueda vertical	2005	Spiders
Google Base	2005-2010-	Scraping-Búsqueda

	2016	horizontal
Facebook vs power.com	2009	Scraping-Búsqueda-Meta
Yahoo	2009	Búsqueda-Meta-Scraping
Kimono Labs	2013	Scraping
Búsqueda y la web móvil	2014	Scraping-Búsqueda-Meta
Google	2015	Scraping-Búsqueda-Meta

Tabla 1: Cronología de motores de búsqueda.

Ya existen ciertas herramientas en el mercado las cuales hacen el scraping respecto a la página web que se necesite de manera automática, algunos de los ejemplos son:

Octoparse: Esta herramienta logra extraer datos de una web de manera efectiva para usuarios que no tienen competencias como programadores avanzados (Cómo hacer scraping: 10 herramientas web).

Cyotek WebCopy: Esta herramienta descarga el contenido, como sus enlaces, recursos y hojas de estilo, además de ser un software libre (Cómo hacer scraping: 10 herramientas web).

Dataminer.io: Esta herramienta es una extensión de Google y de Edge, donde la información se baja en un Excel (Cómo Hacer Scraping: 10 Herramientas Web).

Fuente: Adaptado de (Cómo Hacer Scraping: 10 Herramientas Web, n.d.)

Herramienta	¿Gratis o?	Acceso al código fuente de los scrapers	Integración automática con un sitio web	Trata con sitios como Amazon	Calidad de datos sobre los datos recopilados	Publicación de reporte web

Octoparse				X	X	
Cyotek WebCopy	X					
Dataminer.io	X			X		
ScrapMaster	X	X	X	X	X	X

Tabla 2: Herramientas de Scraping y sus características.

Lo que se concluye con las otras herramientas es que algunas no se adaptan a la página web que se quiere hacerle scraping o cuentan con alguna limitación. En el caso de Octoparse que es una herramienta que cuenta con un gran nivel de manejo tiene costo para su uso a partir de una prueba gratis, la tarifa cambia y comienza a ser de pago. En nuestro caso se adapta según las especificaciones requeridas, contando con el código fuente, aplicando reglas de calidad y sobre todo mostrando la información por medio de un reporte web para hacer más fácil la visualización para el usuario final.

3.3 Marco legal

- **La ley 1581 de 2012:** Habla sobre la protección de datos personales, esta tiene como objetivo desarrollar el derecho que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos, a través de la web, encuestas, llamadas, formularios, etc (Ley 1581 de 2012 - Gestor Normativo).
- Además de los derechos, libertades y garantías constitucionales a que se refiere el artículo 15 de la Constitución Política; (República de Colombia, 2011).
Legalidad del scraping:

- Por otro lado, el scraping es legal, siempre y cuando los datos recolectados estén disponibles libremente para terceros en la web, en este caso los Marketplace tienen esta información libremente por internet.

Es por esto que este proyecto tuvo en cuenta la ley 1581 de 2012 ya que se recolectarán los datos de aquellas personas que se encuentren interesadas en hacer una consultoría sobre la compra de un producto de tecnología, ya que se almacenará dicha información en una base de datos.

4. Desarrollo del proyecto

4.1. Levantamiento de información

En esta sección, se relata el progreso del software en desarrollo, siguiendo las etapas del método Scrum. Aquí se expone minuciosamente cómo se aplicó la metodología y se describe el camino seguido para crear la aplicación web.

4.1.1. Arquitectura del sistema

Se optó por la estructura MVC (Modelo Vista Controlador) para construir la aplicación web, con el propósito de distinguir claramente entre los datos de la aplicación, la interfaz de usuario y la lógica de control. Bajo este enfoque, el modelo se encarga de gestionar los datos, el controlador recibe y procesa las instrucciones, obteniendo los datos del modelo y pasándolos a la vista. La vista, por su parte, se responsabiliza de mostrar las respuestas al usuario, tal como se ilustra en la Figura 8.

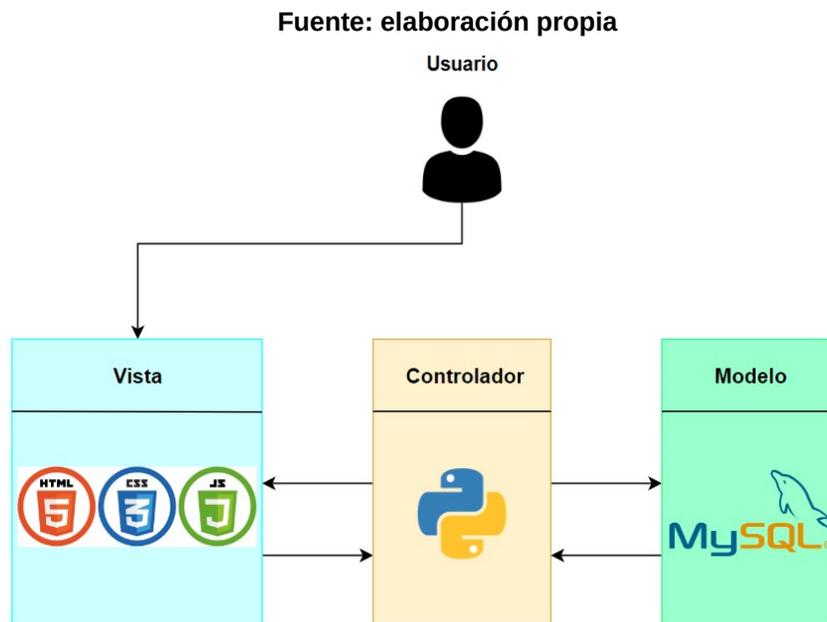


Figura 10: Arquitectura del Software, patrón MVC Modelo Vista Controlador

- En el Modelo, la atención se centra en los datos, por lo tanto, alberga los sistemas para obtener y modificar el estado de la información. La gestión de estos datos se llevó a cabo a través de MySQL.
- La Vista engloba el código necesario para mostrar las interfaces de usuario o representar los estados de la aplicación en formato HTML. Para lograr esto, se empleó CSS y JS para diseñar y mostrar las interfaces de usuario de la aplicación.
- En el Controlador, se encuentra el código destinado a manejar las acciones requeridas por la aplicación, como la visualización de elementos o búsquedas, entre otras. Para desarrollar esta funcionalidad, se utilizó Python.
-

4.2. Análisis del sistema

4.2.1. Requerimientos funcionales

Fuente: elaboración propia

ID	Requerimiento	Descripción dentro del alcance	Descripción fuera del alcance
01	Scripts de Scraping	<p>- El sistema debe ser capaz de extraer información sobre los dispositivos móviles, cumpliendo con las leyes y regulaciones de privacidad.</p> <p>- El software debe ser capaz de almacenar los datos extraídos de los dispositivos en una base de datos centralizada.</p>	- No se tendrá en cuenta el cambio de etiquetas y estilos en los HTML de Falabella, ya que esto no está a nuestro alcance
02	Scripts de transformación de Datos	- Los scripts deben limpiar y preparar los datos extraídos antes de almacenarlos en la base de datos.	

03	Aplicación Web (Interfaz de usuario)	- Se requiere una interfaz de usuario web que permita a los usuarios acceder y visualizar los resultados de la información extraída y procesada.	
----	---	--	--

Tabla 3: Requerimientos Funcionales

4.2.2. Requerimientos no funcionales

Fuente: elaboración propia

ID	Descripción dentro del alcance	Descripción fuera del alcance
01	El software podrá operar en cualquier navegador web, como Chrome, Firefox, Brave y otros similares, siempre que esté conectado a internet.	
02	Se deben emitir mensajes claros en caso de no encontrar resultados	

Tabla 4: Requerimientos No Funcionales

4.2.3. Casos de uso

Para lograr una comprensión más profunda y una visión más clara del sistema que se ha desarrollado, es esencial crear un diagrama de casos de uso. Este diagrama, que se presenta de manera visual, proporciona una descripción de las tareas hechas por los diferentes roles. Además, desempeña un papel fundamental en la programación y en la ejecución de la estructura planificada. En la Figura 11, se puede apreciar el diagrama de casos de uso del aplicativo.

Fuente: elaboración propia

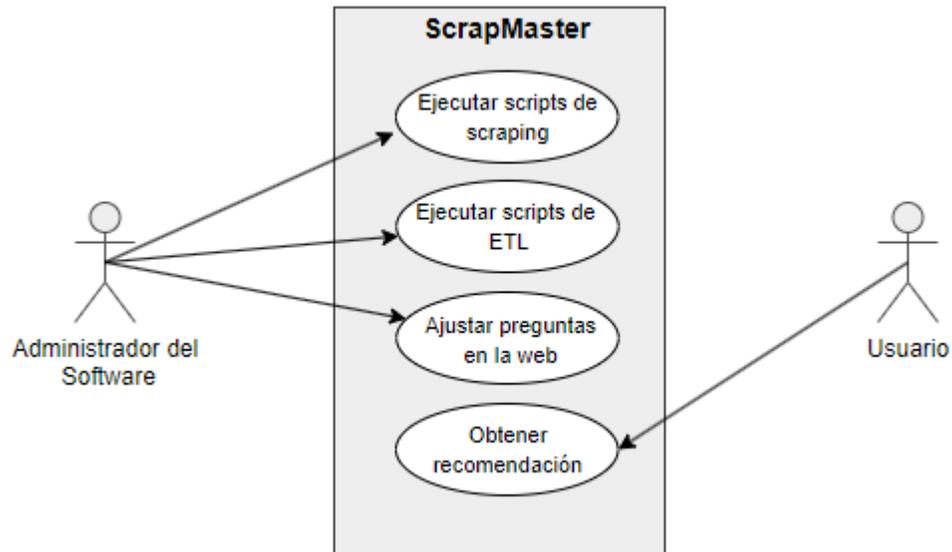


Figura 11: Caso de uso entre el administrador y el usuario final.

4.3. Diseño del sistema

4.3.1. Mockups

En la Figura 12 se nos brinda un adelanto inicial del prototipo de la aplicación en cuestión. Este prototipo se caracteriza por presentar un conjunto de tres elementos interactivos, específicamente botones, cada uno de los cuales desempeña una función particular.

El primer botón, denominado "Inicio", constituye el punto de partida en la interfaz de la aplicación, este para acceder a la funcionalidad principal o la pantalla inicial del programa. Es decir, su acción se orienta hacia el comienzo de la experiencia del usuario en la aplicación.

El segundo botón, identificado como "Preguntas", se dirige a una sección de la aplicación diseñada para mostrar las preguntas. Esta sección emplea un algoritmo para ofrecer las mejores opciones basadas en las respuestas proporcionadas por los usuarios a las preguntas planteadas.

Finalmente, el tercer botón, etiquetado como "URU", es un enlace o acceso directo a la página principal relacionada con "URU". Se entiende que, al hacer clic en este botón, el usuario será redirigido o conducido a la página principal de URU.

Fuente: elaboración propia

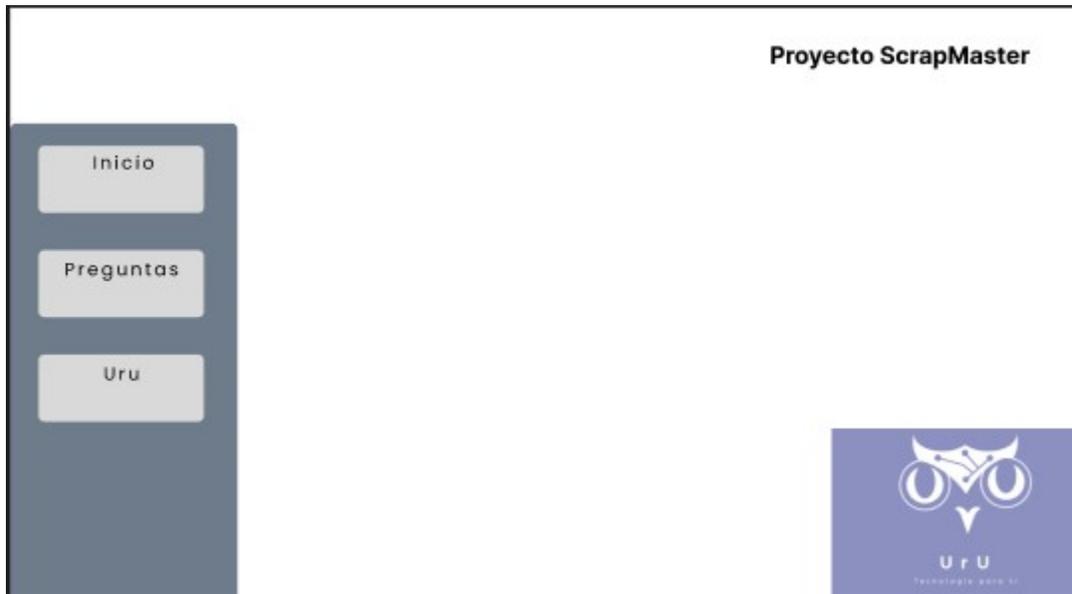


Figura 12: Prototipo de la página de inicio para los usuarios.

4.3.2. Diagramas de secuencia

De la figura 13 a la 15 se muestran los diagramas de secuencia, una herramienta esencial en la ingeniería de software y el diseño de sistemas para representar la interacción entre distintos objetos o componentes en un proceso o escenario particular. Estos diagramas se utilizan comúnmente en el desarrollo de software para visualizar la secuencia de eventos y comunicación entre objetos a lo largo del tiempo, lo que permite a los desarrolladores comprender y planificar con precisión el comportamiento de un sistema, en este caso no se tuvo en cuenta el diagrama de secuencias respecto a 'Ajustar preguntas en la web' ya que el administrador debe tener en cuenta múltiples factores como el cambio en el diseño de la página de Falabella.

Los diagramas de secuencia son especialmente útiles para modelar sistemas complejos, identificar posibles problemas de diseño y garantizar que un sistema cumpla con los requisitos de funcionamiento esperados. En general, son una herramienta valiosa para la comunicación efectiva entre equipos de desarrollo y stakeholders en proyectos de ingeniería de software.

Fuente: elaboración propia

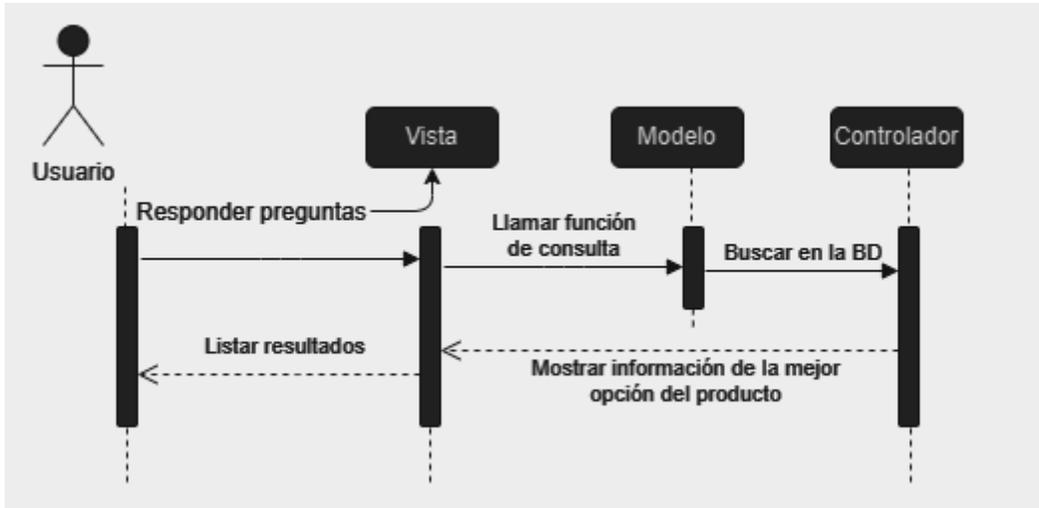


Figura 13: Diagrama de secuencia sobre responder preguntas en ScrapMaster.

Fuente: elaboración propia

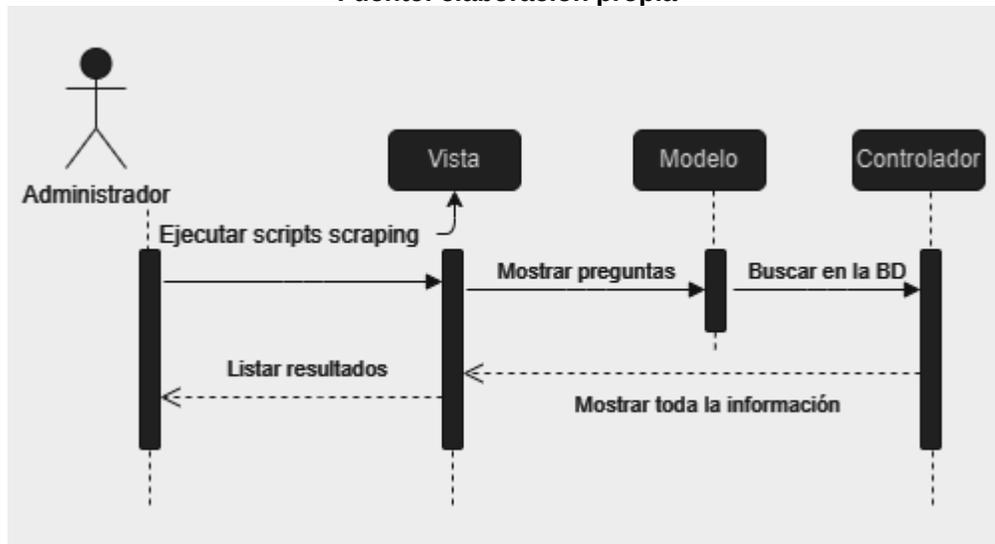


Figura 14: Diagrama de secuencia al ejecutar los scripts de scraping.

Fuente: elaboración propia

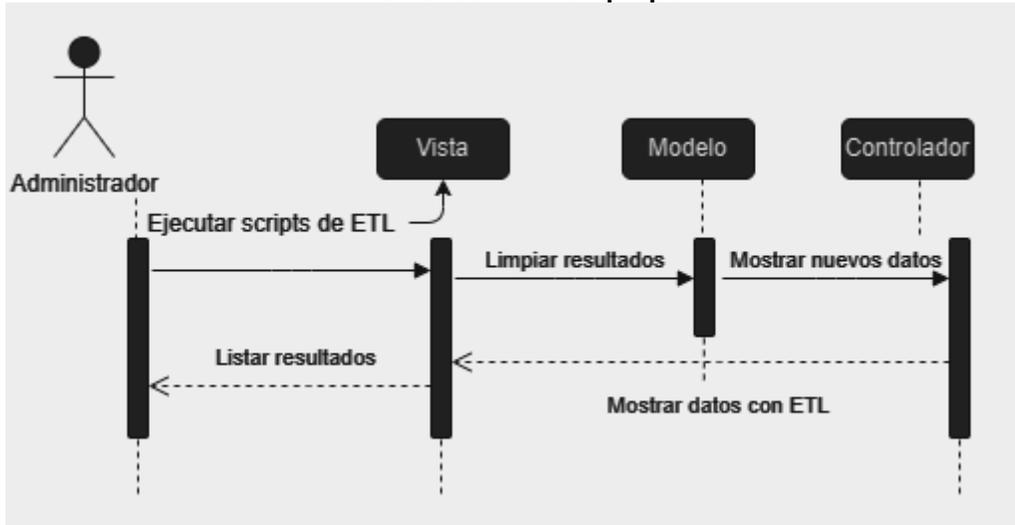


Figura 15: Diagrama de secuencia al ejecutar los scripts de limpieza de datos.

4.3.3 Historias de usuario

Las historias de usuario son una herramienta fundamental en la gestión y desarrollo del software. Explicando una breve descripción de las funcionalidades o características que un usuario puede realizar, en la Tabla 5, se presenta la descripción detallada de la historia de usuario que está relacionada con el requisito de scripts de Scraping.

Fuente: elaboración propia

Historia de usuario	
Número: H01	Usuario: Administrador del sistema
Nombre historia: Realizar Scraping	
Prioridad: Alta	
Descripción: Como analista de datos, deseo tener la capacidad de automatizar el proceso de scraping de información de teléfonos móviles desde el sitio de Falabella, recopilando detalles actualizados sobre modelos, especificaciones técnicas y precios.	
Criterios de aceptación: - Los datos extraídos deben incluir detalles como el nombre, especificaciones técnicas y	

precios.

- Se debe tener en cuenta de manera automática todas las páginas del sitio web

Tabla 5: Historia de usuario Scraping

En la Tabla 6, se muestra la historia de usuario correspondiente al requerimiento de Scripts de transformación de datos.

Fuente: elaboración propia

Historia de usuario	
Número: H02	Usuario: Administrador del sistema
Nombre historia: Ejecutar limpieza de datos	
Prioridad: Media	
Descripción: Requiero la habilidad de automatizar la obtención de los datos de los celulares, llevando a cabo un proceso de depuración de datos para asegurar que la información recopilada sea de alta calidad y esté uniformemente estructurada.	
Criterios de aceptación: - Se debe configurar el link desde donde se extrae la información sobre teléfonos móviles, en este caso Falabella. - Los datos limpios y procesados deben almacenarse en una base de datos para su posterior análisis y uso en la plataforma.	

Tabla 6: Historia de usuario Limpieza de Datos.

En la tabla 7, se muestra la historia de usuario correspondiente al requerimiento de aplicación web:

Fuente: elaboración propia

Historia de usuario	
Número: H03	Usuario: Consumidor de ScrapMaster

Nombre historia: Realizar búsquedas en ScrapMaster
Prioridad: Alta
Descripción: Como usuario de ScrapMaster, lo que necesito es que el sistema me ofrezca resultados precisos cuando respondo a diferentes preguntas, permitiendo definir palabras clave o categorías relevantes para mis necesidades.
Criterios de aceptación: - Se debe mostrar resultados a cada una de las diferentes opciones disponibles, ya sea que se encuentren resultados o que no haya celulares con esas características. - La página debe ser responsive, ya sea en computadores portátiles como en pantallas como monitores, ya que la aplicación funcionara en computadores.

Tabla 7: Historia de usuario aplicación web.

4.4. Base de Datos

4.4.1 Modelo entidad relación

El modelo entidad-relación desempeñó un papel importante en la creación de la base de datos, ya que facilitó la comprensión de los datos y sus características, tal como se ilustra en la figura 16.

Fuente: elaboración propia

celulares	
Marca	TEXT
Modelo	TEXT
Precios	INT
Memoria_interna	TEXT
Marca_y_modelo_del_procesador	TEXT
Memoria_RAM	TEXT
Sistema_operativo	TEXT
Conectividad	TEXT

Figura 16: Modelo entidad-relación simple con la tabla de celulares.

4.5. Pruebas

Para asegurarse de que las funcionalidades del sistema operaran correctamente y cumplieran con los requisitos establecidos, se crearon casos de prueba específicos para cada una de las historias de usuario.

En la Tabla 8, se presenta el caso de prueba que está vinculado con la acción de realizar un scraping en la página web de Falabella.

Fuente: elaboración propia

Caso de prueba			
Id caso de prueba	Nombre caso	Historia asociada	
CP01	Ejecución del scraper	H01	
Descripción caso de prueba: Verificar que al realizar la ejecución del scraping este guarde los resultados según los parámetros establecidos			
Requisitos			
#	Prerrequisitos		
01	Tener instalado Jupyter Notebook		
Resultados			
#	Detalles de los escenarios	Resultado esperado	Estado
1	Abrir el script de Python y ejecutar paso por paso según los comentarios	El script toma la url y hace la extracción de los datos, bajando la información según las etiquetas HTML	Falló

Tabla 8: 1º caso de prueba ejecución del scraping.

En la tabla 9, se ajustaron las etiquetas HTML en el scraper, ya que estas no permitían que se bajaron los datos y adicional se tuvo en cuenta todas las hojas o páginas de Falabella.

Fuente: elaboración propia

Caso de prueba		
Id caso de prueba	Nombre caso	Historia asociada
CP02	Ejecución del scraper	H01

Descripción caso de prueba: Verificar que al realizar la ejecución del scraping este guarde los resultados y adicional tenga en cuenta todas las páginas.			
Requisitos			
#	Prerrequisitos		
01	Tener instalado Jupyter Notebook		
Resultados			
#	Detalles de los escenarios	Resultado esperado	Estado
1	Abrir el script de Python y ejecutar paso por paso según los comentarios	El script toma la url y hace una iteración en todas las páginas del sitio, bajando la información según las etiquetas HTML	Falló

Tabla 9: 2º caso de prueba ejecución del scraping.

En la tabla 10 nuevamente se ajustaron los parámetros ya que la página de Falabella estuvo en constante remodelación, como el cambio de colores y etiquetas en el código, con lo cual es algo impredecible

Fuente: elaboración propia

Caso de prueba			
Id caso de prueba	Nombre caso	Historia asociada	
CP03	Ejecución del scraper	H01	
Descripción caso de prueba: Verificar que al realizar la ejecución del scraping este guarde los resultados y adicional tenga en cuenta todas las páginas.			
Requisitos			
#	Prerrequisitos		
01	Tener instalado Jupyter Notebook		
Resultados			
#	Detalles de los escenarios	Resultado esperado	Estado
1	Abrir el script de Python y ejecutar paso por paso según los comentarios	Se deben bajar los resultados de cada página, según las características de los celulares	Pasó

Tabla 10: 3º caso de prueba ejecución del scraping.

En la Tabla 11, se presenta el caso de prueba que está vinculado con la acción de limpiar los datos del scraping en la página web de Falabella.

Fuente: elaboración propia

Caso de prueba			
Id caso de prueba	Nombre caso	Historia asociada	
CP04	Optimizar datos	H02	
Descripción caso de prueba: Al realizar el proceso de depuración, los datos ya obtenidos en el caso de prueba anterior se deben separar según sus características			
Requisitos			
#	Prerrequisitos		
01	Tener instalado Jupyter Notebook		
Resultados			
#	Detalles de los escenarios	Resultado esperado	Estado
1	Continuar con el script de Python y ejecutar paso por paso según los comentarios	Se debe separar la información por las columnas correspondientes, Marcas, Modelo, etc.	Falló

Tabla 11: 1º caso de prueba limpieza de información.

En la Tabla 12, se ajustaron las expresiones regulares para obtener la información correspondiente, adicional se quitaron las etiquetas HTML, las cuales baja la extracción de Falabella.

Fuente: elaboración propia

Caso de prueba			
Id caso de prueba	Nombre caso	Historia asociada	
CP04	Optimizar datos	H02	
Descripción caso de prueba: Al realizar el proceso de depuración, los datos ya obtenidos en el caso de prueba anterior se deben separar según sus características			
Requisitos			
#	Prerrequisitos		
01	Tener instalado Jupyter Notebook		
Resultados			
#	Detalles de los escenarios	Resultado esperado	Estado
1	Continuar con el script de Python y ejecutar paso por paso según los comentarios	Se debe separar la información por las columnas correspondientes, Marcad, Modelo, etc.	Pasó

Tabla 12: 2º caso de prueba limpieza de información.

En la Tabla 13, se presenta el caso de prueba que está vinculado con la acción de realizar búsquedas en ScrapMaster.

Fuente: elaboración propia

Caso de prueba			
Id caso de prueba	Nombre caso	Historia asociada	
CP05	Obtener resultados	H03	
Descripción caso de prueba: Se debe obtener resultados de todas las opciones posibles al responder las preguntas			
Requisitos			
#	Prerrequisitos		
01	Tener la aplicación corriendo		
Resultados			
#	Detalles de los escenarios	Resultado esperado	Estado
1	Interactuar con la aplicación de ScrapMaster	Obtener un mensaje de los productos o en caso de no encontrar, mostrar que no hay resultados	Pasó

Tabla 13: Caso de prueba obtener recomendaciones de ScrapMaster.

4.5.1. Sprints

Se realizo desde un principio la lectura del Libro 'Python para todos', además de ver unos videotutoriales para involucrarse con el lenguaje de Python y de los términos del scraping como se ve en la Figura 17:

Fuente: elaboración propia

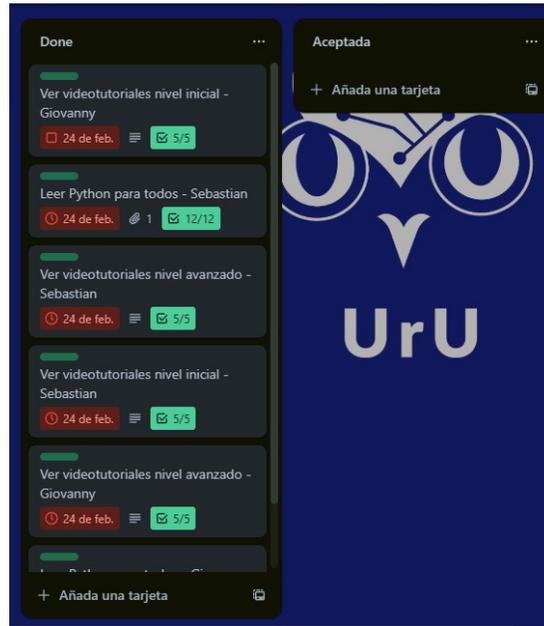


Figura 17: Manejo de tareas en Trello

A continuación, se detallan las actividades ejecutadas de los 3 requerimientos funcionales en 3 tarjetas diferentes con la herramienta de Trello:

En la Figura 18 están los puntos más importantes que se tuvieron en cuenta para el primer requerimiento funcional:

Fuente: elaboración propia

The screenshot shows a Trello card titled "RF 1. Scripts de Scraping" with the following content:

- Miembros:** GB +
- Notificaciones:** Seguir
- Descripción:**
 - Elegir una herramienta o lenguaje de programación:** Decidir qué herramientas o lenguajes de programación se utilizarán para realizar el web scraping. Python es una opción popular con bibliotecas como BeautifulSoup y Scrapy.
 - Configurar un entorno de desarrollo:** Se debe preparar un entorno de desarrollo con las herramientas y bibliotecas necesarias.
 - Identificar la estructura de la página web:** Examinar la estructura del sitio web (Falabella) para determinar cómo se organizan los datos que desea extraer.
 - Seleccionar las rutas y elementos:** Utilizar selectores CSS o XPath para identificar las rutas y elementos HTML que contienen los datos que quiere raspar.
 - Manejar la paginación:** Si los datos se encuentran en múltiples páginas, se debe diseñar un mecanismo para navegar a través de ellas.
 - Establecer un ritmo de solicitud adecuado:** Evitar sobrecargar el servidor del sitio web con demasiadas solicitudes en un corto período de tiempo.
 - Almacenar los datos:** Decidir dónde y cómo almacenar los datos extraídos, ya sea en una base de datos, un archivo CSV o en otro formato.
- Actividad:** Escriba un comentario...

On the right side of the card, there is a sidebar with various options:

- Sugerencias:** Unirse
- Añadir a la tarjeta:** Miembros, Etiquetas, Checklist, Fechas, Adjunto, Portada, Campos personaliz...
- Power-Ups:** Añadir Power-Ups
- Automatización:** Añadir botón
- Acciones:** Mover, Copiar, Crear plantilla, Archivar

Figura 18: Consideraciones importantes para los scripts de scraping en Trello.

La Figura 19 presenta los aspectos más relevantes que se consideraron al abordar el segundo requisito funcional:

Fuente: elaboración propia

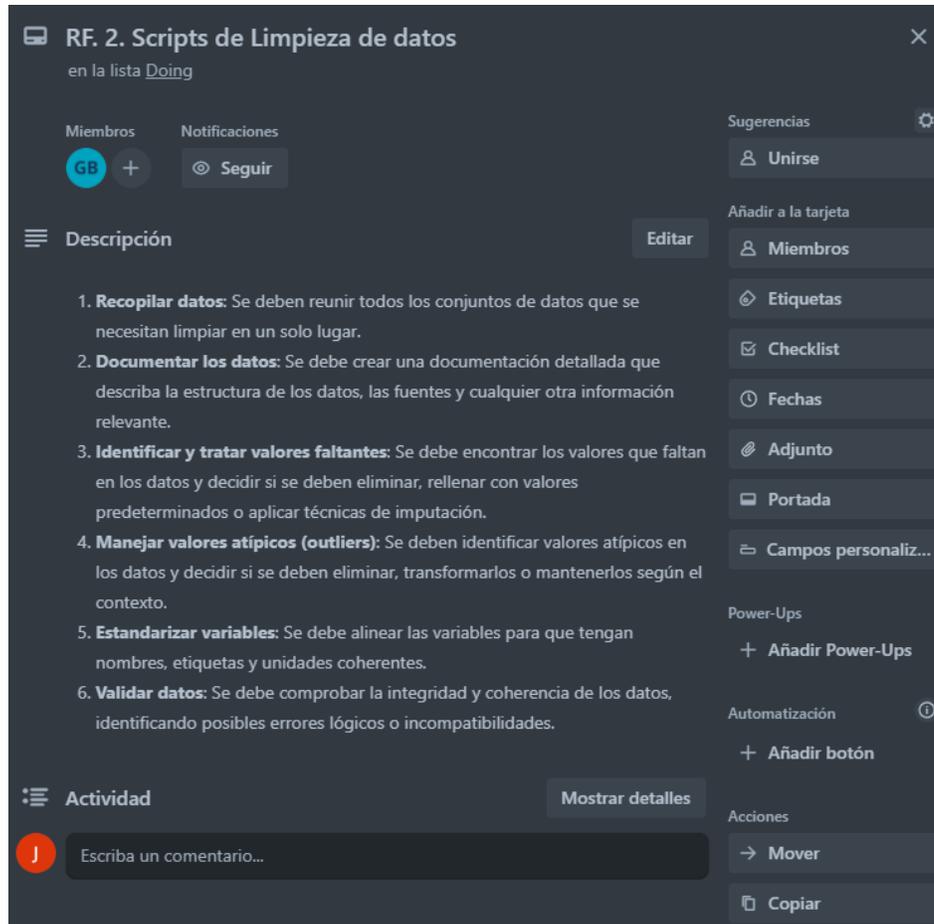


Figura 19: Consideraciones importantes para la limpieza de datos en Trello.

La Figura 20 presenta los aspectos más relevantes que se consideraron al abordar el tercer requerimiento funcional:

Fuente: elaboración propia

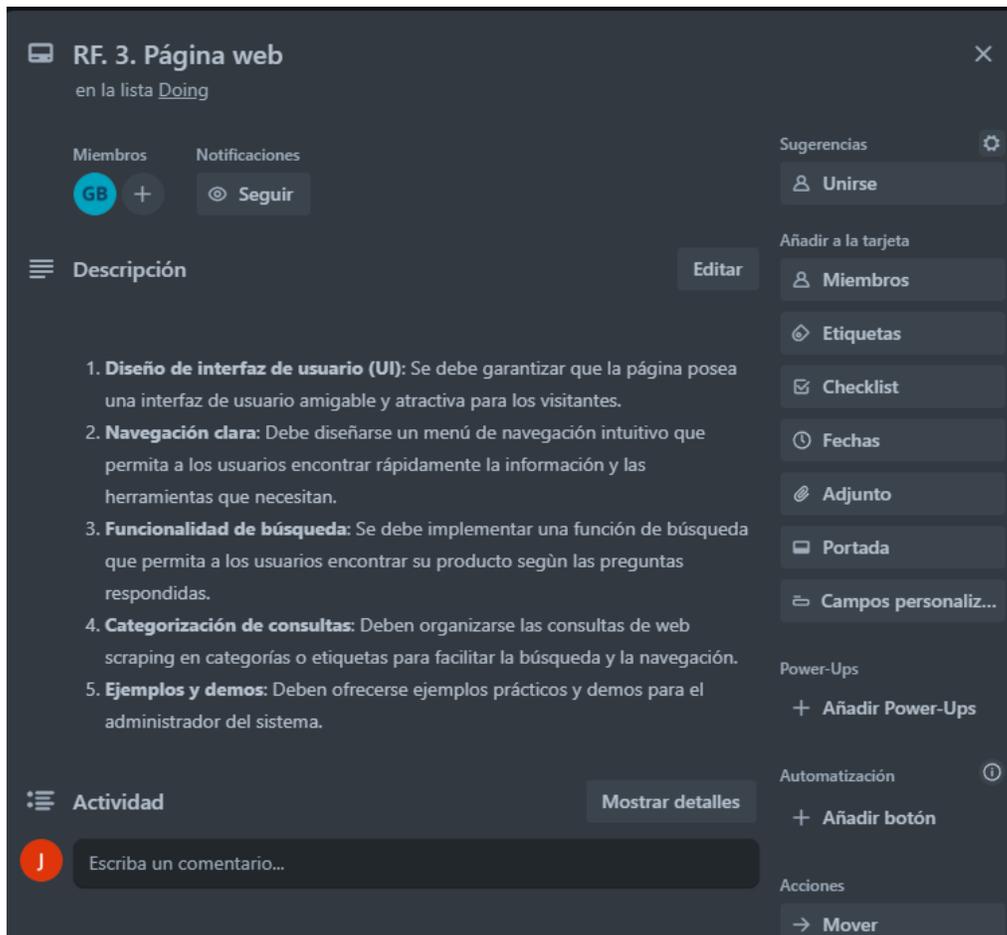


Figura 20: Consideraciones importantes para realizar ScrapMaster en Trello.

5. Resultados obtenidos

En esta fase final del trabajo de grado, se representan los resultados obtenidos de la ejecución e implementación realizadas en los tres Sprint previamente mencionados.

5.1. ScrapMaster

La aplicación ha sido diseñada pensando en la accesibilidad, lo que significa que está disponible y es adecuada para su uso por todas las personas. Su enfoque inclusivo garantiza que independientemente de la ubicación, la aplicación está destinada a ser utilizada de manera efectiva por una amplia audiencia global.

En el inicio de esta, presenta de manera corta el fin que esta tiene, como se ve en la Figura 21.

Fuente: elaboración propia



Figura 21: Presentación página ScrapMaster.

Seguidamente muestra las preguntas, las cuales se clasifican en marca, almacenamiento, precio y memoria RAM: la primera pregunta es la marca, entre las opciones esta “Samsung”, “Apple”, “Motorola” y “Xiaomi” tal como se ve en la Figura 22.

Fuente: elaboración propia



Figura 22: Primera pregunta de ScrapMaster.

Estas preguntas van cambiando según se le dé en el botón de ‘SIGUIENTE’. La segunda pregunta es el almacenamiento, entre las opciones esta 32GB, 64GB, 128 GB y 512GB, tal como se ve en la Figura 23.

Fuente: elaboración propia



Figura 23: Segunda pregunta ScrapMaster.

Seguidamente van los precios, entre los cuales se ubican menor de \$500.000, entre \$500.000 - \$999.000, de \$1.000.000 a \$1.999.999 y más de \$2.000.000, como se ve en la Figura 24.

Fuente: elaboración propia



Pregunta 3: ¿Cuál es tu presupuesto?

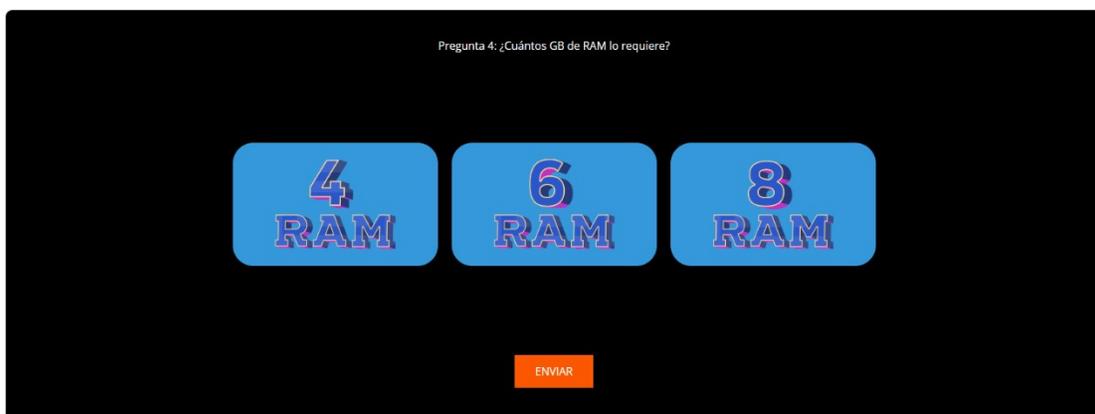
Opción	Rango de Precio
1	MENOR DE \$500.000
2	DE \$500.000 A \$999.999
3	DE \$1.000.000 A \$1.999.999
4	MÁS DE \$2.000.000

SIGUIENTE

Figura 24: Tercera pregunta de ScrapMaster.

Por último, está la pregunta de la RAM, entre las opciones se encuentra 4RAM, 6RAM y 8RAM, como se ve en la Figura 25.

Fuente: elaboración propia



Pregunta 4: ¿Cuántos GB de RAM lo requiere?

Opción	RAM
1	4 RAM
2	6 RAM
3	8 RAM

ENVIAR

Figura 25: Cuarta pregunta de ScrapMaster.

Luego de responder todas las preguntas, mostrara las mejores opciones de acuerdo a las preguntas contestadas, como se ve en la Figura 26.

Fuente: elaboración propia

Marca: SAMSUNG
Modelo: Celular Samsung Galaxy A04 64gb Verde
Precio: 500.000
Marca: SAMSUNG
Modelo: Samsung Galaxy A04 Dual SIM 64 GB verde 4 GB RAM
Precio: 779.000
Marca: SAMSUNG
Modelo: Samsung Galaxy A04e 64gb
Precio: 570.000

Figura 26: Resultado de responder las preguntas de ScrapMaster.

En ocasiones no arrojará resultados ya que no se encontraron opciones según lo respondido, como se ve en la Figura 27.

Fuente: elaboración propia

No se encontraron resultados

Figura 27: Resultado al no encontrar coincidencias en ScrapMaster

Adicional a esto, se cuentan con dos páginas adicionales las cuales son Contacto y URU. Al darle clic en URU redirigirá a dicha página, como se ve en la Figura 28.

Fuente: elaboración propia con los datos de URU



Figura 28: Página oficial de URU.

Por último está la página de Contacto en la cual estarán los datos de los desarrolladores del proyecto y un código QR con la documentación y link a este archivo, como se ve en la Figura 29.

Fuente: elaboración propia

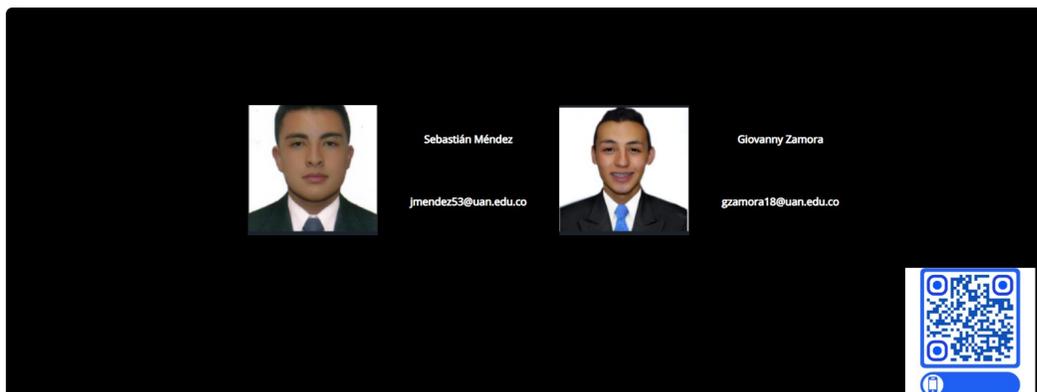


Figura 29: Datos de contacto y QR a documento.

6. Conclusiones y recomendaciones

La creación de la aplicación web se convirtió en un enfoque eficaz y sencillo para recomendar dispositivos móviles utilizando tecnología. Además, su desarrollo implicó la aplicación práctica de los conocimientos adquiridos a lo largo de la carrera, lo que se refleja en este trabajo de grado.

Durante el proceso de desarrollo e implementación del proyecto, surgió la aplicación "ScrapMaster", diseñada para agilizar el proceso de compra, convirtiéndose en una innovación en el mercado y siendo de utilidad para todas las personas, independientemente de su conocimiento en esta área.

Sin embargo, es importante destacar que uno de los desafíos más significativos durante el desarrollo fue la extracción de datos. Esta fase involucró múltiples intentos y pruebas de diversas técnicas y bibliotecas de extracción de datos, ya que la página de la cual se extraían los datos se encontraba en constante remodelación, lo que afectaba la estabilidad de los scrapers. Esto presentó un desafío constante, ya que las actualizaciones en el sitio web requerían ajustes continuos en el proceso de extracción.

Otro desafío significativo radicó en la presentación de la información según las opciones de las consultas. Esto implicaba que, una vez que el usuario terminaba de responder las preguntas, la aplicación debía mostrar la información de manera inmediata y personalizada en función de las respuestas proporcionadas. Este proceso requería una gestión eficiente de la base de datos y una programación cuidadosa para garantizar una experiencia de usuario fluida y relevante.

A pesar de estos desafíos, el proyecto se llevó a cabo con éxito mediante la metodología de trabajo Scrum, garantizando una entrega funcional, compatible y de fácil uso en el plazo de desarrollo establecido.

En cuanto a futuras mejoras, se tiene la intención de ampliar las funcionalidades de la aplicación, incluyendo otros productos tecnológicos además de los teléfonos móviles, así como variar las preguntas en la interfaz y realizar otros ajustes para hacerla más completa y potente. Además, se seguirá trabajando en la adaptación y mejora de las técnicas de extracción de datos y en la optimización de la presentación de la información para seguir proporcionando un servicio de calidad a los usuarios.

7. Bibliografía

Adobe Communications Team. (18 de March de 2022). *What is a Sprint Retrospective? Agenda & Questions | Adobe Workfront*. Recuperado el 18 de October de 2022, de Adobe Experience Cloud: <https://business.adobe.com/blog/basics/sprint-retrospective>

Baeza, R., Rivera, C., Velasco, J., & Baeza, R. (02 de 10 de 2004). *Artículos*. Recuperado el 12 de October de 2022, de Arquitectura de la información y usabilidad en la web: http://eprints.rclis.org/14480/1/arquitectura_informacion_y_usabilidad.pdf

Canal, P. (12 de May de 2022). *Definición y características del Scrum Master*. Recuperado el 15 de October de 2022, de IEBS: <https://www.iebschool.com/blog/definicion-y-caracteristicas-del-scrum-master-agile-scrum/>

Challenger, I., Díaz, Y., & Becerra, R. (14 de 04 de 2014). *The programming language Python*. Recuperado el 17 de October de 2022, de Redalyc: <https://www.redalyc.org/pdf/1815/181531232001.pdf>

Cómo hacer scraping: 10 herramientas web. (s.f.). Recuperado el 28 de October de 2022, de Baloriza Digital: <https://baloriza.com/herramientas-de-web-para-hacer-scraping/>

Garcia, M. (25 de May de 2020). *¿Qué es el 【 Sprint Review 】 en Scrum?* Recuperado el 1 de November de 2022, de ITtude: <https://ittude.com.ar/b/scrum/que-es-el-sprint-review/>

Gómez, L., & Gómez, L. A. (12 de 06 de 2012). *TEORÍAS DEL EMPRENDIMIENTO*. Recuperado el 12 de October de 2022, de TEORÍAS DEL EMPRENDIMIENTO Por: Luis Alberto Gómez Economista 1. Teoría de Andy Freire: Según la teoría del triángulo invert: <https://curso.ihmc.us/rid=1NCYQZM9N-1519FM6-201S/TEORIAS%20DEL%20EMPRENDIMIENTO.pdf>

Krijnen, D. (16 de 08 de 2014). *Automated Web Scraping APIs*. Recuperado el 3 de October de 2022, de web scraping: https://staas.home.xs4all.nl/t/swtr/documents/wt2014_web_scraping.pdf

Ley 1581 de 2012 - Gestor Normativo. (s.f.). Recuperado el 21 de September de 2022, de Función Pública: <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981>

Loaiza, J. A. (2020). *Autor Julián Andrés Loaiza López Universidad de Antioquia Facultad de Ingeniería, Departamento de Ingeniería de Sistemas Medellín,*

- Colombia 2020. Recuperado el 21 de September de 2022, de Repositorio Institucional Universidad de Antioquia: https://bibliotecadigital.udea.edu.co/bitstream/10495/17262/1/LoaizaJulian_2020_AutomatizacionProcesoExtraccion.pdf
- López, J. (5 de December de 2018). *Web scraping*. Recuperado el 21 de September de 2022, de Accelerating the world's research: <https://d1wqtxts1xzle7.cloudfront.net/55775125/web-scraping-with-cover-page-v2.pdf?Expires=1663798870&Signature=daabH8h-bmgkyM62dIlKZkgkDz9oqSw~Ta8yZ13JGWW53x3~OBQc2o8JXGI1-Qj~oQ3ZQIYsJ8YAn3IshUWmvgqZrH04CdaPu81KdONJWcD1ARD-3qZAFhun7oXA5ssZHzAymVbBJsLzB8hF>
- Pott, T. (04 de Junio de 2018). *Extract, transform, load?* Recuperado el 28 de September de 2022, de Extract, transform, load? More like extremely tough to load, amirite?: https://theregister.com/2018/06/04/data_integration_is_hard/
- qayoom, s. (s.f.). *LeadGen*. Obtenido de <https://leadgenapp.io/es/implicaciones-legales-y-%C3%A9ticas-del-web-scraping/>
- Rodriguez, C., & Dorado, R. (30 de 10 de 2015). *¿Por qué implementar Scrum?* Recuperado el 5 de October de 2022, de Vista de ¿Por qué implementar Scrum?: <https://journal.universidadean.edu.co/index.php/Revistao/article/view/1253/1218>
- Rodriguez, J. (12 de 02 de 2021). *MARKETPLACE Y SU ALCANCE A NIVEL SOCIAL*. Recuperado el 14 de October de 2022, de INFORMÁTICA Y DERECHO REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO: http://www.mpsp.mp.br/portal/page/portal/documentacao_e_divulgacao/doc_biblioteca/bibli_servicos_produtos/bibli_informativo/2021_Periodicos/Informatic-Derech_n.10.pdf#page=87
- Sánchez, J. (15 de 09 de 2021). *Wordpress*. Recuperado el 14 de October de 2022, de Empezando-con-wordpress: https://quantika14.com/wpresa/docu_download/Empezando-con-wordpress.pdf
- Scrum: roles y responsabilidades*. (s.f.). Recuperado el 15 de October de 2022, de Deloitte: <https://www2.deloitte.com/es/es/pages/technology/articles/roles-y-responsabilidades-scrum.html>
- Sierra, F. (15 de 10 de 2013). *Estudio y análisis de los framework en php basados en el modelo vista controlador para el desarrollo de software orientado a la web | Investigación y desarrollo en TIC*. Recuperado el 14 de October de

2022, de Revistas Científicas Universidad Simón Bolívar:
<https://revistas.unisimon.edu.co/index.php/identific/article/view/2480>

Wall, A. (20 de 11 de 2017). *History of Search Engines: From 1945 to Google Today*. Recuperado el 28 de September de 2022, de Search Engine History.com: <http://www.searchenginehistory.com/>