



UNIVERSIDAD CÉSAR VALLEJO

ESCUELA DE POSGRADO

**PROGRAMA ACADÉMICO DE MAESTRÍA EN INGENIERÍA DE
SISTEMAS CON MENCIÓN EN TECNOLOGÍAS DE LA
INFORMACIÓN**

Automatización de web Scraping de los diarios de noticias para la empresa
Isuri, San Martín de Porres

TESIS PARA OBTENER EL GRADO ACADÉMICO DE:

Maestro en Ingeniería de Sistemas con Mención en Tecnologías de la
Información

AUTOR:

Br. Antonio Federico Martinez Nuñez (ORCID: 0000-0003-4364-2866)

ASESOR:

Dr. Edwin Alberto Martínez López (ORCID: 0000-0002-1769-1181)

LÍNEA DE INVESTIGACIÓN:

Sistema de información y comunicaciones

LIMA – PERÚ

2020

Dedicatoria:

A mis padres por haber formado la persona que soy actualmente, ellos me han acompañado en muchos de mis logros. Siempre me han motivado incondicionalmente a alcanzar mis metas.

Agradecimiento:

Agradezco a Dios ante todo por permitirme lograr mis metas y proyecto en compañía de mis padres, de la misma forma agradecer a las personas que me han apoyado en el desarrollo y culminación del proyecto.



DICTAMEN DE LA SUSTENTACIÓN DE TESIS

El / La Bachiller: **Martínez Núñez, Antonio Federico**

Para obtener el Grado Académico de **Maestro en Ingeniería de Sistemas con Mención en Tecnologías de la Información**, ha sustentado la tesis titulada:

Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres

Fecha: 16 de agosto de 2020

Hora: 8:00 a.m.

JURADOS:

PRESIDENTE:
Dr. Luis Alejandro Esquivel Castillo



SECRETARIO:
Dr. Alejandro Ramirez Rios



VOCAL:
Dr. Martínez Lopez Edwin Alberto



El Jurado evaluador emitió el dictamen de:

- Aprobar por excelencia

Habiendo encontrado las siguientes observaciones en la defensa de la tesis:

-
-
-

Recomendaciones sobre el documento de la tesis:

-
-
-

Nota: El tesista tiene un plazo máximo de seis meses, contabilizados desde el día siguiente a la sustentación, para presentar la tesis habiendo incorporado las recomendaciones formuladas por el jurado evaluador.

Declaratoria de autenticidad

Yo, Antonio Federico Martínez Núñez, estudiante de la Escuela de Posgrado, del programa de la Maestría en ingeniería de sistemas con mención en tecnologías de la información, de la Universidad César Vallejo, Sede Lima Norte; presento mi trabajo académico titulado: “Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres”, en 94 folios para la obtención del grado académico de Maestro en ingeniería de sistemas con mención en tecnologías de la información, es de mi autoría.

Por tanto, declaro lo siguiente:

- He mencionado todas las fuentes empleadas en el presente trabajo de investigación, identificando correctamente toda cita textual o de paráfrasis proveniente de otras fuentes, de acuerdo con lo establecido por las normas de elaboración de trabajos académicos.
- No he utilizado ninguna otra fuente distinta de aquellas expresamente señaladas en este trabajo.
- Este trabajo de investigación no ha sido previamente presentado completa ni parcialmente para la obtención de otro grado académico o título profesional.
- Soy consciente de que mi trabajo puede ser revisado electrónicamente en búsqueda de plagios.
- De encontrar uso de material intelectual ajeno sin el debido reconocimiento de su fuente o autor, me someto a las sanciones que determinen el procedimiento disciplinario.

Lima, 01 de agosto de 2020



Antonio F. Martínez Núñez

Índice

Dedicatoria	ii
Agradecimiento	iii
Página del jurado	iv
Declaratoria de autenticidad	v
Índice	vi
Índice de figuras	vii
RESUMEN	viii
ABSTRACT	ix
I. INTRODUCCIÓN	1
II. MÉTODO	16
2.1. Tipo y diseño de investigación	16
2.2. Escenario de estudio	18
2.3. Participantes	18
2.4. Técnicas e instrumentos de recolección de datos	19
2.5. Procedimiento	19
2.6. Método de análisis de información	20
2.7. Aspectos éticos	21
III. RESULTADOS	22
IV. DISCUSIÓN	28
V. CONCLUSIONES	36
VI. RECOMENDACIONES	38
REFERENCIAS	39
Anexo 1: Matriz de categorización	45
Anexo 2: Preguntas de la entrevista semi estructurada	46
Anexo 3: Matriz de desgravación de la entrevista	47
Anexo 4: Matriz de codificación de la entrevista	50
Anexo 5: Matriz de entrevistados y conclusiones	53

Anexo 6: Guía de Observación	56
Anexo 7: Ficha de Análisis documental	58
Anexo 8: Otras evidencias	59
Anexo 9: Propuesta de automatización de web Scraping	67

Índice de figuras

Figura 1. Triangulación de la observación de la unidad de estudio.	22
Figura 2. Triangulación del análisis documental.	23
Figura 3. Triangulación de las entrevistas semi estructurada.	24
Figura 4. Triangulación de las técnicas de investigación utilizadas.	25
Figura 5. Triangulación de los antecedentes, marco teórico y los resultados	27
Figura 6. Web Scraping de los diarios de noticias.	59
Figura 7. Proceso de extracción de noticias.	60
Figura 8. Proceso General de extracción de noticias.	61
Figura 9. Característica de los servidores usados – SO, Recursos y ubicación. Fuente: IBM.	62
Figura 10. Característica de los servidores usados - RED. Fuente: IBM.	62
Figura 11. Valor por estimado por hora de cada servidor. Fuente: IBM.	63
Figura 12. Costo similar al proveedor de Servicio de noticias Stor vtex.	64
Figura 13. Promedio de noticas ingresadas manualmente.	64

RESUMEN

La presente investigación titulada: Automatización de web Scraping de los diarios de noticias para la empresa Isuri, tuvo como objetivo el desarrollo la automatización de web Scraping de los diarios de noticias para la empresa Isuri, dedicada a la monitorización de noticias de web en el distrito de San Martín de Porres, la investigación fue de enfoque cualitativo, el método de investigación se basó en el paradigma interpretativo, tipo de investigación aplicada tecnológica y se utilizó el diseño de investigación acción. Se empleó como técnicas de recolección de datos, la entrevista a profundidad semiestructurada realizada a expertos, la observación a la unidad de estudio la cual fue la oficina el departamento de tecnología de la información de la empresa Isuri y el análisis documental. Además, se utilizó el método inductivo para el análisis de la información.

Se Concluye que la empresa tuvo un gran consumo de recursos de presupuesto y humano, los cuales son usados en tres procesos, un aplicativo interno que está desplegado en dos servidores en la nube, un proveedor especializado de donde se extraer cierta cantidad de noticias mensuales, y por último el personal dedicado al ingreso manual de las noticias al sistema comercial. A través de la automatización de este proceso mediante el uso de nuevas tecnologías, modelado de extracción, las reglas del negocio y los valores generados para la toma de decisiones basadas en datos, se vio reflejado la reducción de manera considerable en los recursos que son usados para el proceso de web Scraping. Para la reducción principal del uso de los recursos computacionales se usó Serverless, para el control y balance del flujo del proceso se usó Nifi con Kafka de apache.

Palabras claves: web Scraping, ELT, NoSql, Serverless, Nifi, bots.

ABSTRACT

The present investigation titled: Automation of Web Scraping of news newspapers for the Isuri company, had the objective of developing the automation of web Scraping of news newspapers for the Isuri company, dedicated to monitoring web news in the district From San Martin de Porres, the research was qualitative in approach, the research method was based on the interpretive paradigm, type of applied technological research, and the action research design was used. As data collection techniques, the semi-structured in-depth interview carried out with experts, the observation to the study unit which was the office of the information technology department of the Isuri company and the documentary analysis were used. In addition, the inductive method was used for the analysis of the information.

It is concluded that the company had a large consumption of budget and human resources, which are used in three processes, an internal application that is deployed on two servers in the cloud, a specialized provider from which to extract a certain amount of monthly news, and finally the staff dedicated to manual entry of news into the commercial system. Through the automation of this process through the use of new technologies, extraction modeling, business rules and the values generated for decision-making based on data, the reduction in the resources used was reflected considerably. for the web scraping process. For the main reduction of the use of computational resources, Serverless was used, for the control and balance of the process flow, Nifi with Kafka of apache was used.

Keywords: web Scraping, ELT, NoSql, Serverless, Nifi, bots.

I. INTRODUCCIÓN

En Brasil, Knewin (2019) indican que realizar el monitoreo de noticias en tiempo real, es una demanda muy importante para el panorama de la información digital, donde permite a sus clientes identificar menciones sobre sus propias marcas, competencia y movimientos del mercado. Sin embargo, LiveOn de Brasil empresa que monitorea las noticias puede tener un costo alto de USD 10,000.00 hasta USD 60,000.00, en la cual implementa su solución y está dirigido para el sector B2B, ellos indican que su producto es excelente para distribuidores, proveedores y empresas de reventa. Para el sector B2C la implementación y diseño de un Layout tiene un valor desde de los USD 200 hasta los USD 10,000.00 (Live On Solutions 2020).

A nivel mundial el 20 % del tráfico en internet eran provenientes de "bots" (en oposición a los bots benévolos como las arañas de los motores de búsqueda); Una cifra que ha bajado un 6,4 por ciento en 2017. La compañía citó a Amazon como la fuente de la mayoría del tráfico global de bots defectuosos con un 18 por ciento; un aumento del 10/6 por ciento en 2017. La gran mayoría del tráfico de bots malos vino de los EE. UU., con los Países Bajos en segundo lugar; cifras respaldadas por hallazgos recientes de honeypot (CrbOnline, 2019). Un juez en EEUU había ordenado que Microsoft elimine 'as soon as possible' toda tecnología destinada a impedir que hiQ Labs obtenga datos públicos de LinkedIn a través de web Scraping. LinkedIn es una de las redes sociales en que el web Scraping se aplica de manera más frecuente. (Reuters 2017).

Así también existe una empresa llamada IP noticias, que se encarga del monitoreamiento de medios y analítica de información, es un servicio de inteligencia informativa que busca conocer y entender el comportamiento multimedia de marcas. a través de la digitalización y gestión de contenido, la analítica, la creación de modelos prescriptivos y soluciones tecnológicas que mejoran la calidad de vida. (ipnoticias-latam, 2019). Las empresas buscan obtener información que les permita no solo posicionarse de mejor manera en el mercado donde comercializan, sino también buscan monitorear a la competencia, comparar sus precios y cuál es la tendencia social en el dónde se mueven sus productos o servicios, con lo cual poder reaccionar de manera inmediata a alguna crisis que pueda ocurrir donde se vean involucrados sus negocios.

Así mismo existe una técnica llamada Scraping dedicada a la mineración de información específica mediante bots que pueden lograr simular el comportamiento humano. Aunque web Scraping no es un término nuevo, en años pasados la práctica ha sido más comúnmente conocido como screen Scraping, data mining, web harvesting o variaciones similares. Grandes empresas reconocidas a nivel mundial hacen el uso de Scraping, la empresa que vienen viene realizando este proceso por bastante tiempo es Google, quien usa estas técnicas para poder indexar páginas web a su buscador. Ryan Mitchell (2015). En el Perú el Scraping ha sido implementado en 2016 bajo un proyecto denominado “Manolo” con la finalidad de poder exponer la corrupción en nuestro país. Para ayudar a los periodistas a rastrear a los cabilderos, comencé el proyecto Manolo, es una herramienta de búsqueda simple que contiene los registros de visitantes de varias instituciones gubernamentales en Perú. (Carlos Peña 2016).

En otros países la automatización de este proceso se realiza mediante bots que extraen información de páginas web mediante el uso de técnicas de Scraping y Crawler. El uso de Scraping más común en nuestro país es para la recuperación de cierto tipo de información, como el tipo de cambio de Sunat o SBC, lo mismo para obtener una lista de precios o productos de ciertas páginas Web de E-commerce, y otras categorías de menor impacto. En la presente investigación titulada “Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres”.

La empresa Isuri se dedica al procesamiento de datos de noticias Web, con la finalidad de obtener una las ultimas noticias del día, realiza web Scraping a algunos sitios de noticias posteriormente se depurada con diferentes parámetros y algoritmos básicos. Cada noticia de un sitio web debe ser raspado y almacenado en una base de datos centralizada. De la misma forma su procesamiento personalizado permite la elaboración de un producto final, con el cual ayude a sus clientes de diversos rubros comerciales puedan tomar una decisión basada en datos e identificar potencialmente como se encuentran posicionados en el mercado y como se encuentra su competencia, este tipo de información le aporta un panorama amplio de toma de decisiones, aumenta la posibilidad de éxito y permite tener un control adecuado de donde esta direccionando sus resultados. Para la empresa implementar toda una infraestructura con esta técnica realmente tiene un costo elevado en recursos computacionales.

Con respecto a los antecedentes de carácter internacional, Cárdenas y Chau (2018), exponen en su investigación que la automatización de web Scraping les permitió recolectar información relevante de varias páginas web durante 15 días sobre los precios, ubicación y otras características de inmuebles que se encuentran alquiler en las principales ciudades de Colombia, una vez descarga esta información fue depurada y proceso en un base de datos unificada donde se clasifica para cada una de las ciudades. Generando una lista de inmueble consistente, precios más acertados. Correa (2018), expone en su investigación que el uso de una arquitectura escalable para minería de datos en tiempo real. El uso de web Scraping requiere tener una capacidad alta de CPU y memoria RAM, muchas veces superando la arquitectura interna, el uso de micro servicios permitirá optimizar y monitorear el uso de los recursos, así mismo alertar de posibles fallas, el no controlar el aplicativo adecuadamente podría ocasionar la pérdida excesiva de datos y la degradación del servicio.

Stefano, Neves, Santana, Valadão y Hammes (2020), exponen en su investigación que el uso de web Scraping acompañado de otras tecnologías para el análisis de evolución de la ingeniería de transporte en los últimos 50 años, la extracción de estos datos de PDF que son más de 12,000 documentos hacerlo de forma manual requiere el uso de muchos recursos, la implementación de bots fue muy eficiente para este tipo de tareas. Fue posible realizar un estudio, dentro del espectro temporal, e identificar nuevas tendencias, basadas en el análisis de las series históricas de las más frecuentes y mayor relevancia.

Haralson (2016) expone en su investigación que con el uso de web Crawling facilito la selección de potenciales clientes, donde con un algoritmo con diversos parámetros evaluó un valor para este, permitiendo a personal de la compañía tener unas métricas de ayuda para la cotización de la venta. Así también, indican que el aplicativo en PHP es muy general, donde funciono en 90% de los casos. Verdes (2016) expone en su trabajo que la empresa E-Force ICT desea utilizar las funciones de Scraping para dar una ventaja competitiva a sus clientes. Esto permite a sus clientes de E-Force ICT poder monitorizar la competencia, sabiendo en tiempo real los precios de los productos de sus páginas web. la aplicación web almacena los enlaces y los XPath en la base de datos y los extrae de forma correcta.

En el contexto nacional, Córdova y Quispe (2019), manifiestan en su investigación que el uso de técnicas de Scraping y Api's, fue la mejor manera de abordar el problema dentro de la institución, fue necesario el uso de algoritmo personalizados para web Scraping, ya que cada sitio web tiene una estructura interna distinta. El uso técnicas de Scraping facilitaron raspar y recolectar información de distintos sitios web de entidades o empresas en tiempo real Los recursos usados para la ejecución de este proceso requieren basten tiempo, esto se debe a la comunicación entre los distintos servidores de páginas de web, en algunos casos por el bloqueo o también baneo de IP publica del servidor donde se ejecuta el Scraping. El resultado obtenido facilita la búsqueda de información de manera rápida y eficiente de los diferentes servicios básicos consultados.

Huamán (2019), manifiesto en su investigación se hizo uso de bots como asistentes virtuales y aprovecho las técnicas de web Scraping para la extracción de datos. Los chats bots se encargaron de ofrecen un servicio de mejor calidad y personalizado para cada cliente, la información extraída permite el entrenamiento de forma masiva y automatizada, este proceso les permitió obtener información actualizada y en tiempo real, mejorando la clasificación, precios y características de la información, el proyecto tuvo un resultado muy exitoso, esto se logró con la integración de web Scraping y mensajería de Facebook Messenger como interfaz de consulta. Midiendo el nivel de satisfacción, se obtuvo un resultado positivo.

Capuñay (2018), manifiesto en su trabajo que el uso de Web Scraping permitió la recolección masiva de datos, el análisis permitió identificar los requisitos que necesitan cada uno de los profesionales de la carrera de ingeniería de sistemas en las instituciones públicas peruanas. Los datos a partir de la aplicación de técnicas de Scraping de datos, entre marzo de 2014 y marzo de 2017 en un número de 109,825 convocatorias de las cuales 4777 aplican para el análisis de la presente investigación. Así mismo, Castañeda (2016), manifiesta en su investigación que logro identificar que el uso de estructuras repetitivas, facilita que el aplicativo desarrollado para que pueda conectarse y extraer la información de sitios Web específicos, Este proceso tiene un considerable uso de recursos computacionales por la conexión recurrente y constante para obtener información en tiempo real.

En referente al marco teórico, La automatización de web Scraping facilita la extracción masiva de noticias de los diarios de noticias, cada sitio web tiene un modelo de estructura distintas al otro, motivo por lo cual el algoritmo tiene que ser dinámico, los bots facilitarían que este proceso se automatizado y escalable. La infraestructura tecnológica que soportara este aplicativo constara de usos de servicios en la nube, gestos de mensajería para aplicaciones y almacenado en una base de datos NoSQL. Modelado de extracción, es el modelado de las páginas web que pueden contener información adicional como los metadatos, anotaciones semánticas, que pueden ser usadas para identificar fragmentos de datos específicos. La mayoría de anotaciones se pueden encontrar incrustadas en las páginas web, alojadas en el DOM (Document Object Model) de cualquiera de los navegadores. (Ghosh, Banerjee y Sengupta 2018).

WWW (World Wide Web, Traducción “Red Global”) este servicio que popularmente es conocido como web, la web fue creada en 1989 por Tim Berners, el concepto consistía en poder organizar la información usando medios físicos de comunicación que sería la red de internet y protocolos HTTP (Hyper Trasference Protocol) que es la transferencia de hipertexto que los navegadores usan para realizar peticiones y recibir las respuestas de los servidores web. Ramos, A. & Ramos, R. (2014). La evolución de la web ha pasado por varias etapas: la Web 1.0, su característica principal es por el contenido estático, la Web 2.0 su característica es por el contenido dinámico o interactivo, la Web 3.0 se caracteriza por el contenido colaborativo, la web 4.0 donde el sistema operativo está establecido por la web. (Onyanha, Plekhanova y Nelson 2017).

Web index, en su investigación indica que los motores de búsqueda más reconocido y usados, son los que se encargan constantemente de hacer el rastreo de los sitios web con la finalidad de garantizar un resultado más relevante a la búsqueda, normalmente es una combinación de matemática lingüística y psicología, encontrando resultados más exactos. GoogleBot es un robot Crawling de páginas web, se encarga de encontrar, recuperar e indexar. Es un Spider que va recorriendo todos los hilos de internet, dentro de estas indexaciones hay reglas de negocio donde google ignora (no indexa) palabras comunes como (como, es, en o de, cómo, por qué, así como ciertos dígitos y letras individuales). (Krunal 2014).

Técnicas de Web Scraping, las diversas técnicas de Scraping, 1) El tradicional copiar y pegar, 2) Grapado de texto y expresión regular comando UNIX (Sistema operativo portable). 3) Programación del protocolo de transferencia de hipertexto (HTTP). 4) Análisis del lenguaje de marcado de hipertexto, se usa lenguaje de consulta semiestructurado, como XQuery y consulta por hipertexto(HTQL). 5) Análisis del modelo de objetos de documento (DOM). 6) Software de Web Scraping, existen muchas herramientas disponibles para usar, 7) Plataformas de agregación vertical, existen diversas compañías con plataformas de cosechas específicas, las cuales crean y supervisan una multitud de bots, 8) Analizadores de páginas web de visión por computadora. (Sirisuriya 2015).

La extracción de información se utiliza para motores de búsqueda, bibliotecas de noticias, manuales, texto específico de dominio o diccionarios. La mineración de texto es una tarea de recuperación de información destinada a descubrir nuevos datos previamente desconocida. (Eloisa 2013). Así mismo Loreto (2019) web Scraping, se la conoce también como Web Crawling, data mining, creen Scraping, Web extracción entre otros. A esta técnica, como se ha mencionado anteriormente, se le puede dar varios usos entre los cuales destacan. La Web Scraping, para Borrego (2016) es una técnica que permite extraer información de una página web y de manera automatizada. Los datos se extraen mediante softwares de programación y, a estos softwares, se les denomina: bots, spider, Crawlers, etc. Grosso modo, lo que hacen es simular la navegación de un humano y substraer información, transformando un contenido no estructurado (normalmente HTML) a datos estructurados.

La Escuela de Datos (2016) para poder extraer y estructurar los datos de las páginas Web se utiliza la técnica llamada como Web Scraping. Scraping significa “raspado” se refiere a la extracción, limpieza y filtro de los datos Algoritmos para Scraping, para Heydt (2018) muchos procesamientos de lenguaje natural requieren dividir gran cantidad de texto en oraciones, aun para mucho de nosotros suele ser una tarea simple, para los ordenadores es un problema mayor. Es necesario utilizar otros algoritmos para poder clasificar el texto de mejor manera. Para Krunal (2014) que existe diferentes tipos de algoritmos para web Scraping que se aproximan a la mayoría de estructuras web, sin embargo, esto se puede mejorar con métodos que le permitan delimitar de mejor manera el resultado deseado.

Regla de negocio, la búsqueda de automatizar este proceso de extracción de información de sitios web ha generado una creciente demanda en la generación de Bots, por lo resultados obtenidos la oportunidad de incorporar la capacidad e integración con otros datos relevantes para las corporaciones. Big data, para el autor indica que el concepto esta medido con la “V” de Big data, donde considera que existen 3 V’s: está determinado por grandes de volúmenes de datos que se puedan generar grandes velocidades y con una variedad de fuentes de información nunca antes vistas hasta ahora. Que comprende 5 fases: a) Fuentes de información Big Data, b) Integración de Big Data, c) Sistema y repositorio Big data d) Procesamiento Big data y e) Interfaces y visualización Big Data, donde todas estas fases se encargaran de recopilar diversas fuentes de datos, bien por ser parte intrínseca y necesaria para un proyecto. Con el objetivo principal de enriquecer los datos para los cados de uso y reglas del negocio para mejorar la calidad. (Miranda 2015). Considerando que el uso de Big data brinda grandes oportunidades para la inferencia estadística, quizás desafíos aún mayores, especialmente cuando se compara con el análisis de conjuntos de datos cuidadosamente recopilados, y se identifican otros factores que pueden afectar las reglas de negocio. (Franke, Roscher y Annie 2016).

Bots, la principal actividad que realizan los bots es el descargar los paginas web. Esta es una labor realiza por los bots que son conocidos como arañas (Spiders), Crawlers, Web Crawlers o Web Walkers. Las arañas, descargan páginas web según los parámetros marcados en la aplicación, el uso de bots para fines comerciales, son muy útiles como recopiladores de información que van realizan la descargar página web buscara los enlaces contenidos dentro de ella y da seguimiento a otros enlaces dentro de ella para enlazar otras páginas, como esto podría ser una labor infinita. (Vukovic y Dujlovic 2016)

ETL, la tendencia en Big data ha convertido las ETL (Extract, transform y load) en ELT’s (Extract, load and transform), es quiere decir, que, tras lograr extraer los datos de la fuente de origen no se hace ninguna transformación, Dentro de estos procesos de transformación se suman la limpieza de datos (Data Cleansing) mejorando la calidad y el enriquecimiento de datos (Data Enrichment) con otras fuentes de información. (Miranda 2015). Se prioriza el almacenado, luego se organiza, prepara y procesa. Esto se logra a través de una red de alta velocidad utilizando ETL / ELT o herramientas de

procesamiento de datos grande. (Shekhar y Pandey 2019). El avance actual del uso de la nube, este proceso puede ayudar de forma enorme en la capacidad y una figura versátil, con el propósito de mantener toda la información cruda extraída, Cuando se apilan, los cambios y las razones comerciales se conectan utilizando controladores SQL locales. (Dipti 2019).

Serverless, la computación sin servidor es un modelo de ejecución que se hace cada vez más popular en la nube. Con servicios como AWS Lambda, (Google Cloud Functions y Azure Functions. Klimovic, Wang y Kozyrakis 2018). Los usuarios escriben aplicaciones como colecciones de funciones sin estado que implementan directamente en un marco sin servidor. Los 4 principales modelos son: 1) IaaS (Infrastructure as a Service, traducción Infraestructura como servicio). 2) PaaS (Platform as a Service, traducción Plataforma como servicio). 3) CaaS (Container as a Service, traducción Contenedor como servicio), 4) FaaS (Function as a Service, Traducción “Funciona como un servicio”). (Mohamed 2018).

La arquitectura sin servidor en AWS, para Malawski, Gajek, Zima, Balis y Figiela (2017), que están diseñadas principalmente para procesar tareas en segundo plano de la Web e Internet de las cosas aplicaciones o procesamiento de flujo controlado por eventos. Para Mohamed (2018) en la nube fue iniciada por AWS Lambda en el 2014, este modelo en la nube tiene diversos beneficios 1) NoOps: donde la responsabilidad por aprovisionar, mantener y parchar los servidores se transfiere de cliente a proveedores. Y los desarrollares se enfocan en optimizar, innovar y mejorar sus funciones, solo se paga el tiempo que la función demore en ejecutar. 2) Autoescalado y alta disponibilidad: proveedor de servicios para decidir cómo usar su infraestructura de manera efectiva para atender las solicitudes de los clientes y escalar horizontalmente las funciones en función de la carga. 3) Optimización de costes: sólo se paga por el calcule el tiempo y los recursos (RAM, CPU, red o tiempo de invocación) que consume. No paga por los recursos inactivos. 4) Polygot: se puede usar diversos lenguajes de programación una parte de la aplicación se puede escribir en Java, otra en Go, otra en Python; en realidad no importa mientras haga el trabajo. Para Debois y Doland (2017) un enfoque sin servidor no resuelve todos los problemas, ni elimina las complejidades subyacentes del sistema. Pero cuando se implementa correctamente, puede proporcionar oportunidades para reducir, organizar y gestionar la complejidad.

Go Serverless, Boucher, Kalia, y Andersen (2018) con la finalidad de monopolizar los CPU's, se usará un enfoque que funcione a escalas de microsegundos, aprovechando las multitareas que nos proporciona Go Lang a través de las Gorutinas (hilos ligeros). Según Mohamed (2018), el anuncio de AWS su apoyo al lenguaje de programación GO a partir del inicio del 2018, donde existe algunas framework para integrar con Lambda, una de las razones de uno usar servidores es poder usar Poligot, independientemente del lenguaje. Es donde Go entra en el juego, teniendo como ventajas lo siguiente: 1) Orientado a la nube: Lenguaje diseñado por Google, considerando la escalabilidad y reducción en el tiempo de compilación. 2) Rápido: Go tiene una sintaxis limpia y especificaciones de lenguaje claras. Esto ofrece un lenguaje fácil para que los desarrolladores aprendan y muestra buenos resultados rápidamente mientras produce código de fácil mantenimiento. 3) Escalable: o tiene una concurrencia integrada con goroutines en lugar de hilos. 4) Eficiente: la velocidad de compilación más rápida permite una retroalimentación igual de rápida; esto sin duda es la ventaja más importante para alguien con un presupuesto ajustado. Además, Go está respaldado por Google, tiene un ecosistema grande y un gran soporte IDE (IntelliJ, VSCode, Atom, GoGland) y depuración. Para Biradar, Shekhar y Reddy (2018) Golang se usa para aprovisionar para admitir la programación en contenedores son livianos y portátiles, ya que se trata de una encapsulación de un entorno en el que se ejecutará la aplicación

Sistemas de mensajería, indica que este tipo de aplicaciones desempeñan un papel bastante importante en proyectos de big data. Dentro de un proyecto de gran escala, tiene diferentes capas como: 1) capa de ingestión: los datos de entrada para ser ingeridos en una base de datos, los orígenes pueden ser una o varias fuentes de datos. 2) Capa de procesamiento: comprende la lógica empresarial la cual dicta como se va transformar los datos para convertirlo en información útil. 3) capa de consumo: donde la información útil se encuentra en un único punto, donde puede ser distribuido a diferentes puntos a través de consumidores. Según Guo y Ding (2018) está comprobado que el grupo Kafka puede garantizar 1) la fiabilidad del mensaje, 2) reducir significativamente la sobrecarga del recurso, 3) mejorar el rendimiento del clúster, 4) garantizar la fiabilidad del mensaje y 5) disponibilidad del sistema y el alto rendimiento. (Manish y Chanchal 2017).

Kafka, mencionan que este servicio nació a partir de una necesidad que tuvo una red social profesional altamente conocida como LinkedIn, durante su crecimiento el equipo técnico inicio con un sistema de recopilación de métricas del software utilizando componentes internos personalizados con soporte en herramientas de código abierto. Sin embargo, esta solución no duro mucho tiempo por varios problemas al momento de querer usar los mensajes en diversas aplicaciones. (Dobbelaere y Sheykh 2017). Todos los mensajes se almacenan como registro en sistemas de archivos persistentes. Tiene un sistema de registro anticipado que permite escribir todos los mensajes publicados antes de ponerlo a disposición de aplicaciones de consumo. Cada mensaje en Kafka es una colección de bytes, esta colección se representa como un matriz, estos se van almacenando mensajes en la secuencia que van llegando. (Manish y Chanchal 2017).

El número de particiones se configura al momento de la creación de “Topic “, físicamente, cada topic se extiende sobre diferentes corredores de Kafka, que albergan uno o más particiones. Un típico Kafka clúster consta de múltiples corredores, los cuales ayuda en las lecturas y escrituras de mensajes de equilibrio de carga en el clúster controlado por el líder. Wu, Shang y Wolter (2019). Zookeeper, Según Brenner, Wulf, Goltzsche, Weichbrodt y Lorenz (2016) es un servicio de coordinación que permite aplicaciones distribuidas la fácil implementación y coordinación. Tales mensajes se pueden nombrar, gestionar la configuración, elegir líderes, pertenecer a grupos, barreras y bloqueos distribuidos. Para Manish y Chanchal (2017), para mantener sus estados usan Zookeeper, cada topic tiene un líder acompañado de cero o más tuberías como seguidores. Los líderes gestionan cualquier solicitud de lectura o escritura para sus respectivas particiones, dicho seguidores replican al líder en segundo plano sin interferir activamente, según EL-Sanosi y Ezhilchelvan (2018) los servidores son réplicas entre sí y cada uno mantiene una copia del estado de la solicitud. Los clientes de Zookeeper pueden enviar sus solicitudes a cualquiera de los N servidores. Las solicitudes pueden ser ampliamente categorizado como leer o escribir.

NoSQL, para Malki, Hamadou, Chevalier, Péninou y Teste (2013), la interpretación de NoSQL (No solo SQL) son varias clases de sistema de gestión de base de datos en un enfoque “sin esquema” identificados por su no adherencia lo que permite una amplia variedad de representaciones y esta flexibilidad conduce a un gran volumen de datos heterogéneos.

Las bases de datos Nosql no se crean principalmente en tablas. Este tipo de movimiento NoSQL comenzó en los primeros años del siglo XXI, cuando el mundo comenzó a centrarse en creación de base de datos a escalas web, donde este tipo de escala se refiere a atender a cientos millones de usuarios y ahora crece a miles de millones de dispositivos conectados, incluidos, entre otros, teléfonos móviles inteligentes, tv por internet y muchos más. El NoSql hacer referencia a cualquier almacén de datos que no sigue el modelo de RDBMS (relational database management system, traducido “sistema de gestión de bases de datos relacionales”) tradicional, específicamente. (Gaurav 2013).

Los enfoques no son relacionales y no utilizan SQL como lenguaje de consulta. Este tipo de base de datos intentan resolver los problemas de escalabilidad y disponibilidad frente a los de atomicidad o consistencia. NoSQL no es una base de datos, ni siquiera es un tipo de base de datos, solo un término para identificar un conjunto de base de datos fuera del ecosistema. Namdeo y Suman (2020). Sistema de gestión de bases de datos relacionales, Para Gaurav (2013), Un RDBMS tradicional tiene un conjunto de características como: 1) Atomicidad: todo en una transacción tiene éxito para que no se revierta, 2) Consistencia: una transacción no puede dejar la base de datos en un estado inconsistente, 3) Aislamiento: una transacción no puede interferir con otra y 4) Durabilidad: una transacción completa persiste, incluso después de reiniciar las aplicaciones. Por muy indispensables que puedan parecer estas cualidades, son bastante incompatibles con la disponibilidad y el rendimiento en aplicaciones de escala web.

Los buscadores, una de las piezas importantes de todo aplicativo en crecimiento de Big data es el buscador. Lucene y SoIR son 2 proyectos apache íntimamente relacionados. Lucene constituye un buscador y SoIR es una aplicación que recubre Lucene con otras aplicaciones. Mishra (2019) las características de Elasticsearch ayudan a ofrecer una mejor oferta de búsqueda e idoneidad para lograr los objetivos de búsqueda actuales y futuros para el sitio web extraídos. Andre, Behrens, Branson, Brummer, Chaze y otros (2018) los documentos son inyectados en formato JSON dentro del clúster Elasticsearch utilizando protocolo HTTP, creando de una plantilla con un índice para el mapeo y búsqueda configurada para algunos campos posteriormente

Elasticsearch, es un motor analítico y de búsqueda escrita en Java, la cual tiene una base de SOIR, su primer lanzamiento en el 2010. ha sido ampliamente adoptado por la NASA, Wikipedia y GitHub, para diferentes casos de uso. Elasticsearch proporciona una API HTTP/JSON, que tiene estas principales características: 1) Distribuido: Puede comenzar con un clúster Elasticsearch de un solo nodo y puede escalar ese clúster a cientos o miles de nodos 2) Alta disponibilidad: replicación de datos significa tener múltiples copias de datos en su clúster, 3) Basado en REST: Elasticsearch está basado en la arquitectura REST y proporciona puntos finales API para no solo realizar operaciones CRUD a través de llamadas API HTTP, 4) Potente DSL de consulta: El DSL de consulta (lenguaje específico del dominio) es una interfaz JSON proporcionada por Elasticsearch para exponer el poder de Lucene para escribir y leer consultas de una manera muy fácil y 5) Sin esquema: Ser sin esquema significa que no tiene que crear un esquema con nombres de campo y tipos de datos antes de indexar los datos en Elasticsearch. (Bharvi 2016).

Comprensión de REST y JSON, Según Serrano, Stroulia, Lau (2017) Rest, es una API web, los datos y los servicios basados en REST están expuestos como recursos como URL. proporcionando una sintaxis simple y directa para acceder a recursos de datos enriquecidos. Cada recurso tiene un identificador de recurso, que se llama como URI. Según Lucas, Favaram, Felipe, Cris, Todt (2019) JavaScript Object Notation (JSON) es un formato ligero de intercambio de datos y, en el mundo NoSQL, se ha convertido en un formato estándar de serialización de datos. La razón principal para usarlo como formato estándar es la independencia del lenguaje y la compleja estructura de datos anidados que admite. JSON tiene el siguiente soporte de tipo de datos: Matriz, booleano, nulo, número, objeto y cadena. Para Powell, Nason, Elliott, Mayhew, Davies y Otros (2017), consideran que existe una relación entre la información raspada en la web y los índices basados en datos convencionales porque puede cuantificarse, documentarse y analizarse

Toma de decisiones basadas en datos, el uso de la analítica empresarial ha evolucionado a diferentes velocidades en diversas áreas. En diversas áreas como finanzas, contabilidad y operaciones era más sencillo obtener datos, pero otras áreas como recursos humanos han quedado rezagadas porque la necesidad no era ampliamente conocida. Tracey (2012). para poder ejecutar de manera correcta las tomas de decisiones

basadas en datos, se necesita una estrategia sólida acompañada de una serie de herramientas tecnológicas que le permitan trabajar de manera más inteligente, sin embargo, ninguna de estas aplicaciones logra mágicamente mejorar la toma de decisiones. Kannan y Li (2016). Así también, destacan los puntos de contacto en el proceso de marketing acompañados de procesos de estrategia como de tecnologías digitales están teniendo y tendrán un impacto significativo en las empresas. (Alvares, Muñiz, Morán 2019)

El posicionamiento de marca, Para Morales (2010) citado por Carpio, Hanco, y Cutipa (2019) en redes sociales y otros canales tiene una mayor relevancia a la web 2.0. El lograr posicionar una marca en internet permite a los usuarios ubicar a la empresa mediante búsquedas en la web mediante motores de búsquedas, interacción en redes sociales, blogs, sitios web u otros servicios. Para Koch y Gyrd-Jones (2019) como un proceso recurrente de varios niveles, es algo más que una actividad de marketing a nivel corporativo. Para analizar a la competencia, un camino práctico para conocer a nuestros competidores dentro del mercado. Donde se monitorea los activos digitales de la competencia, por ejemplo: sus sitios web y sus estructuras, redes sociales, aplicaciones móviles y en que canales en los que se encuentran.

Los reportes o informes, siempre fueron una instantánea del pasado, se generan procesamiento masivo de la información en muchas empresas por las noches, sin embargo, la barrera no es la capacidad de procesamiento, si no el acceso a los datos en tiempo real. (Tracey 2012). Las Métricas, no solo nos permite medir el impacto que tiene una publicación en cualquiera de los canales digitales que existen, también es necesario calcular el retorno de inversión respecto a otros medios, lo cual ayuda a determinar que parte del presupuesto es en aquel que tiene mayor rentabilidad. (Alvares, Muñiz, Morán 2019).

La analítica predictiva de un conjunto de datos permite identificar mediante métricas y objetivos para descubrir diversos fenómenos u tendencias, a diferencia de otras técnicas el Scraping es la base para la mayoría de estos procesos, centrándonos en los datos que son de interés de las diversas plataformas web, mediante un lenguaje de programación y herramientas podemos manipular, Ajax, JavaScript u otro elemento dentro de sitio en línea. Sandulescu (2018). Los datos alineados a la estrategia llevan el

análisis predictivo un paso más allá. En este paso, la estrategia comercial impulsa la priorización del análisis de los datos y las predicciones futuras ya no solo se basan en tendencias pasadas, enfocándose en el análisis del entorno competitivo externo generando la necesidad de elaborar escenarios a futuro. (Tracey 2012).

La empresa necesitaba que se pueda automatizar la extracción de noticias, cada página web tiene un modelo distinto en su estructura, usando diversos algoritmos con web Scraping, la empresa tiene reglas de negocios que se consideran al momento de la automatización con Bots personalizados. Así mismo, se necesita una estructura tecnológica de bajo costo llamada Serverless (Sin servidor) usando Lambda de Amazon Web Services (AWS) y servidores locales donde se diseñó una herramienta de control de flujos o ETL. Así como un gestor de colas llamado Kafka que permitirá el control de mensajes que finalmente fueron integrando la información en un base de datos NoSQL de nombre Elasticsearch, también se tuvo en consideración las reglas de negocio de la empresa que permitió tener un aplicativo escalable y robusto. Los datos recuperados pueden ser usados por los clientes para mejorar su posicionamiento de marca, monitorizar a la competencia y realizar métricas personalizadas.

La justificación teórica de la presente tesis radica en que sea un complemento a los fundamentos de esta investigación la cual permitirá identificar información acerca de la problemática relacionada con la automatización de web Scraping de los diarios de noticia, agrupando diversas técnicas de web Scraping, algoritmos, uso de bots permitiendo integrar diversas tecnologías que en su conjunto permitan tener un resultado de alto impacto con pocos recursos computacionales y de presupuesto, igualmente, sus resultados serán un aporte a las teorías existentes que indican que el uso de bots para web Scraping facilita la selección de datos de las distintas páginas web, considerando una estructura adecuada para cada una de ellas. Desde su intencionalidad práctica, la automatización de web Scraping tiene un alto impacto en la reducción de uso de recursos tecnológicos como de presupuesto esto mediante el uso de técnicas, herramientas y el como gran aliado a la tecnología. Lo cual permitirá centralizar la información en un solo repositorio y ser más accesible para diversos clientes o usuarios.

Los recursos para la automatizar este proceso de extracción de noticias, se ven soportadas por aplicaciones de bajo costo Serverless, Productos de Apache y base de datos NoSql. Así mismo, el producto de este desarrollo investigativo desde lo metodológico, se conocerá cuáles son los aspectos que intervienen en la automatización de web Scraping para lo cual se diseñará una entrevista a ser aplicada especialistas en ele tema, con ello poder obtener la información necesaria. También, basado en experiencias observables se aprecia que, la extracción de noticias manifiesta el uso muchos recursos, lo que motiva esta propuesta a manera de intervención sustentadas en técnicas y herramientas para un modelado de extracción de noticias, considerando las reglas del negocio, la tecnología y la explotación de los datos y DDDM lo mencionado anteriormente es un aporte de gran impacto para esta investigación.

A partir de lo descrito se formula el siguiente problema de investigación: ¿Cómo se automatiza el web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?, Asimismo, los problemas secundarios se describen de la siguiente manera: ¿Cómo es el modelado de extracción para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?, ¿Cómo son las reglas de negocio para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?, ¿Cuáles son las tecnologías para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?, ¿Cómo se interpreta la toma de decisiones basado en datos para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?. Así mismo se plantea el siguiente objetivo general de investigación: Desarrollar la automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres. Y como objetivos específicos: Describir el modelado de extracción para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres, Definir las reglas de negocio para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres, Identificar las tecnologías para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres, Evaluar la toma de decisiones basado en datos para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres.

II. MÉTODO

El presente trabajo de investigación busca automatizar el web Scraping en la empresa Isuri, la empresa realizaba la recolección de información de noticias con un aplicativo desarrollado por ellos, así también, tenían una alternativa donde pueden consultar a un proveedor especializado, este servicio tiene un costo alto en el presupuesto. En este estudio busca describir la automatización de web Scraping como una opción o complemento de bajo recursos para la extracción de noticias de los diarios en sus sitios web. En este sentido, se eligió el enfoque cualitativo debido a su idoneidad para examinar las interacciones del usuario en su ambiente natural. El enfoque cualitativo pretende que el investigador comience un proceso de examinar los hechos en sí y apoyándose en estudios previos con la finalidad de generar una teoría consisten de lo que está observando. Hernández y Mendoza (2018). Es una investigación de paradigma interpretativo, ya que busca interpretar las experiencias recogidas de distintos ángulos para crear un conocimiento valido y significativo, permitiendo comprender mejor la situación de la empresa y obtener una adecuada solución. (Villegas 2011).

2.1. Tipo y diseño de investigación

Tipo de investigación

El presente trabajo según su propósito de investigación es de tipo aplicada tecnológica, según la finalidad es descriptiva donde se realizó un cambio en la problemática, de tener un proceso costoso en recursos por costo del servicio y humano, a un conjunto de aplicaciones automatizadas a bajo costo, haciendo uso de diversas técnicas, y un conjunto variado de tecnología de gran impacto con pocos recursos. Puede decirse que es la puesta en práctica, mediante diseños adecuados para que los resultados sean mejoras del proceso convencional, mediante el paso intermedio en mucho de los casos a escala piloto. Cegarra (2016). Así mismo, la investigación tecnológica parte de la observación-reflexión-praxis, de la necesidad de análisis-síntesis del objeto de investigación que puede ser un sistema, una norma técnica, maquinas, herramientas, dependiendo del tipo de tecnológica. Es la reflexión de sobre la máquina, es el repensar la máquina, es el repensar la tecnología que se está aplicando y sobre la cual se está trabajando. (Bello (2008) cita por Ñaupas 2014).

Otro aspecto importante en la investigación tecnológica es un proyecto de transformación, no de las teorías sino de las tecnologías existentes para optimizar su eficiencia o eficacia; por ende, el proyecto se presenta como un conjunto metódico de mecanismos, pasos y técnicas de carácter reflexivo, con procesos evaluativos a ser aplicados mediante la observación, como base del método para captación del hecho tecnológico, una vez esto se sistematiza, a través de la experiencia reflexiva, ella en si se convertirá en un producto, en un método de investigación tecnológica. (Bello (2008) citado por Ñaupas 2014).

Diseño de investigación

En la presente investigación se ha utilizado el diseño investigación acción por qué busca la resolución del problema y el uso de mejores prácticas concretas dentro de la investigación, el proceso cotidiano se realizaba de forma manual por cada diario web, en el cual se extraía la información a una base de datos NoSql, se observó que luego esta se depuraba para ser insertada atreves de otro proceso operado de forma manual al repositorio de las noticias trabajadas, el análisis de esta operación resulta en que este tipo de trabajo toma un tiempo y recursos, para resolver este problema se optó por usar varias tecnológicas open source, acompañado de varios servicios y técnicas, creando un plan de mejorar permanente automatizado. En relación, el principal objetivo de la investigación acción es transformar la realidad, en definitiva, se centra en el cambio y transformación. Para el efecto este método se vincula a la resolución de problemas mediante un proceso cíclico que va desde la “actividad reflexiva a la actividad transformadora”. (Cabezas, Andrade y Johana 2018).

Se centran en aportar información que guíe la toma de decisiones para programas, procesos y reformas estructurales. Podemos encontrar dos diseños fundamentales de la investigación-acción: práctico y participativo. El diseño participativo implica que las personas interesadas en resolver la problemática ayudan a desarrollar todo el proceso de la investigación: de la idea a la presentación de resultados. Las etapas o ciclos para efectuar una investigación-acción son: detectar el problema de investigación, formular un plan o programa para resolver la problemática o introducir el cambio e implementar el plan, además de generar realimentación, la cual conduce a un nuevo diagnóstico y a una nueva espiral de reflexión y acción. (Hernández y Mendoza 2018).

2.2. Escenario de estudio

El escenario de estudio es en el departamento de tecnología de la información de la empresa Isuri, donde se encargan en la recolección de datos de diferentes medios de internet, radio, televisión y otros, lugar donde se llenó la guía de observación con la información obtenida de los colaboradores estudiados que se encargan de este proceso, se tomó en consideración este escenario porque el ambiente físico permite entender las necesidades para poder automatizar el web Scraping es donde se origina el problema principal de esta investigación que es el proceso que se realizaba para la obtención de noticias, las evidencias muestran una realidad de la sobre carga en los recursos que puede con llevar hacer este proceso. En esta investigación, el ambiente de observación el departamento de tecnología de la información, ello permitirá comprender como es el proceso en dicho espacio.

2.3. Participantes

Para este estudio de investigación se seleccionó representantes por conveniencia, donde está caracterizado por un esfuerzo voluntario y deliberado para encontrar los participantes “representativos” a través de la inclusión en la muestra de grupos aparentemente típicos. Se selecciona directa e intencionalmente a los colaboradores de la empresa que conocen respecto al tema de interés. Los métodos de muestreo no probabilísticos más usada es el muestro por convivencia, este procedimiento consiste en seleccionar las unidades de muestrales más convenientes para el estudio o en permitir que la participación de la muestra sea totalmente voluntaria. Por tanto, no existe control de la composición de la muestra representativa y la representatividad de los resultados es cuestionable. (Fernández 2004).

La selección de la muestra no es aleatoria, razón por la que se desconoce la probabilidad de selección de cada unidad o elemento del universo del universo, este tipo de muestreo con frecuencia tiene el sesgo de selección debido a la influencia que tiene el sujeto que determina la inclusión en la muestra. Borda (2009). Los participantes de este estudio fueron los colaboradores del área de sistema de la empresa Isuri, los cuales fueron observados en un ambiente de trabajo. Se examinó el contexto laboral durante las horas asignadas a este proceso, donde se pudo observar y registrar las interacciones, de igual forma poder vaciar en el instrumento correspondientes.

2.4. Técnicas e instrumentos de recolección de datos

Durante la recolección de datos se utilizaron técnica de entrevista, observación y análisis documental; los instrumentos fueron una guía de entrevista semi estructurada y la guía de observación estructurada; Al respecto, Hernández y Mendoza (2018) la recolección de datos resulta fundamental para la investigación cualitativa. Lo que se busca en un estudio cualitativo es obtener datos (que se convertirán en información) de personas, otros seres vivos, comunidades, situaciones o procesos en profundidad; en las propias "formas de expresión" de cada unidad de muestra. Al tratarse de seres humanos, los datos que interesan son conceptos, percepciones, imágenes mentales, creencias, emociones, interacciones, pensamientos, prácticas, experiencias, vivencias y roles manifestados en el lenguaje de los participantes, ya sea de manera individual, grupal o colectiva. Se recolectan con la finalidad de analizarlos y comprenderlos, y así responder a las preguntas de investigación y generar conocimiento.

2.5. Procedimiento

El procedimiento para realizar el trabajo investigativo consistió en aplicar los instrumentos de investigación siguientes: Inicialmente, se redactó un texto donde se expuso al encargado del departamento de tecnología de la información de la empresa Isuri, donde se detallan los objetivos del trabajo de investigación. Luego de solicitar los permisos requeridos para este trabajo, es aceptado. Lo cual permitió realizar la observación es realizada a los participantes, durante sus labores en el horario habitual de sus labores diarias. Así, mismo para las entrevistas a profundidad se realizó de manera personal y directa a cada especialista en el momento de efectuarlo, se aclaró las dudas existentes hasta asegurarse que ellos hayan comprendido la razón de la entrevista.

Se les comunico que no hay correctas ni incorrectas, sino que deben contestar, con la mayor sinceridad. Cabe señalar que, la entrevista no se hizo conjuntamente con el vaciado de información en la guía para la observación; Por un lado, se realizó la entrevista y en otro momento se vació la información lograda con los participantes en la guía de la observación. Para finalizar, la ejecución del análisis de los resultados se procedió a realizar la conclusión sobre la automatización de web Scraping. Se pretende desarrollar una manera más óptima de recolectar las noticias de los sitios web, haciendo un énfasis en la tecnología que se utilizada para lograr mejores resultados. A. Categoría 1: Modelado de extracción, Sub categoría A1: Web, Sub categoría A2: Web índice, Sub

categoría A3: Técnicas de Scraping, Sub categoría A4: Algoritmos, B. Categoría 2: Regla del negocio. Sub categoría B1: Big data, Sub Categoría B1: Bots, Sub categoría B1: ETL/ELT. C. Categoría 3: Tecnología. Sub categoría C1: Serverless, Sub categoría C2: Go Serverless, Sub categoría C3: Sistema de mensajería, Sub categoría C4: NoSql. D. Categoría 4: Marketing Digital. Sub categoría D1: Posicionamiento de marca, Sub categoría D2: Monitorizar a la competencia, Sub categoría D3: Métricas.

2.6. Método de análisis de información

La triangulación metodológica implica la triangulación dentro del mismo método, en el mismo proceso se puede utilizar diferentes técnicas e instrumentos provenientes de un método particular referidas del mismo objeto; o también se puede utilizar una combinación de métodos (la observación, la entrevista, el análisis documental, etc.). Yuni y Ariel (2006). Se elaboró un instrumento en base a las categorías y subcategorías para el uso de las entrevistas a profundidad, con esto por desarrollar 3 tipo de matrices: 1) preguntas y respuestas de cada entrevistados, 2) la codificación de las entrevistas de cada entrevistado para cada una de las preguntas y 3) consolidar las preguntas de los 3 entrevistados, determinando las semejanzas y diferencia por cada uno de los entrevistados respecto a la pregunta analizando mediante el análisis de contenido por frases para poder llegar a un conclusión por cada una de las preguntas.

El análisis de contenido en un sentido amplio, y como se entiende en esta investigación, es una técnica de interpretación de textos, bien sean escritos, grabados donde exista toda clase de padrón de datos, para la transcripción de las entrevistas. Bardin (1986). Este método será usado para el análisis de la información obtenida en las entrevistas aplicadas a los colaboradores del departamento de tecnología de la información en la empresa Isuri, en el distrito de San Martin de Porres.

El análisis temático se define como un procedimiento para el tratamiento de la información en estudios cualitativos, que ayuda determinar, organizar, examinar detalladamente y reportar ofrecer modelos o temas desde una minuciosa lectura y re-lectura de la información acumulada, para deducir resultados que favorezcan la conveniente comprensión/interpretación del fenómeno estudiado. Braun y Clarke (2006) citados por Mielles et al. (2012). Cuando el observador a recolectado suficiente información para realizar su análisis, mediante la codificación de las informaciones recogidas va reduciendo la información y haciéndola manejable para la búsqueda de

categorías de significado, necesarias para su interpretación. Martínez y Gonzales (2014). Cuanto mayor sea la diversidad de las metodologías, datos e investigadores empleados en el análisis de un problema específico, mayor será la confiabilidad del producto.

2.7. Aspectos éticos

El presente trabajo de investigación se desarrolló teniendo en cuenta el código de ética donde se contempla tener responsabilidad en las decisiones, colaboración profesional, honestidad, respetando los derechos de propiedad intelectual, así como las ideas de los autores dentro del contenido de esta investigación, manteniendo alto niveles de competencia profesional, evitando el daño, manteniendo el anonimato de los participantes y respetando la normativa vigente. Estos códigos de conducta, convergen a la misma orientación, elevar los estándares de competencia profesional, la de salvaguardar el bienestar de los participantes y de la investigación.

En esta investigación se tomó en consideración la autorización del encargado del departamento de tecnología de la información y la gerencia de la empresa Isuri en el distrito de San Martín de Porres para tener acceso a los colaboradores de la empresa, de igual manera se facilitó una copia de la guía de observación para que se pueda mantener al tanto de los aspectos de la observación. Este trabajo de investigación se desarrolló tomando en cuenta el cumplimiento de las disposiciones vigentes del Reglamento de Grados y Títulos de la Universidad César Vallejo contenida en la Resolución Rectoral N° 089 – 2019 – UCV, así también, se ha validado utilizando el modelo Normas APA UCV vigente y uso de la herramienta turnitin.

III. RESULTADOS

En cuanto al presente trabajo de investigación los resultados de esta investigación se han efectuado con técnicas de recolección de datos como la observación, entrevistas a profundidad a especialistas y análisis documentario, cada técnica se aplicó con su instrumento, y los instrumentos está en función de lograr los objetivos planteados. A continuación, se muestra las diversas conclusiones, la cual se llegó a través de las triangulaciones.

Se utiliza 03 maneras de hacerlo, se usa un aplicativo interno que se encarga de buscar materias en unos cuantos sitio, en algunos casos se hace manual la inserción de estas al sistema comercial, y adicionalmente se puede solicitar a un proveedor especializado este tipo de información, aproximadamente 4 horas por el grupo de sitios de noticias, así mismo, el seleccionar información de un servicio de tercero, demanda un tiempo adicional por cada diario de noticias entre 25 – 30 minutos adicionales, estas se tienen que ingresar manualmente.

P1 Encargado del departamento de Tecnología de información

Para la extracción se utiliza 03 maneras de hacerlo, un aplicativo interno que se toma para este proceso se demanda 4 horas por el grupo de sitios de noticias, en algunos casos se hace manual la inserción de estas al sistema comercial, y como una alternativa adicional se puede solicitar a un proveedor especializado este tipo de información, esto puede demorar entre 10 – 15 por cada cliente, en otros casos se tienen que ingresar manualmente algunas noticias relevantes para el informe que se le envía al cliente. Con esta investigación se reduciría a una solo forma de extracción y automatizada, reduciendo considerablemente los recursos, tiempos y errores.

P2 Analista programador

Dentro de este proceso las noticias se almacenan en el sistema comercial tienes que ser seleccionadas para uno o varios clientes, este proceso puede demorar entre 10 – 15 por cada cliente, así también hay algunos retrasos en la extracción por la cantidad de pasos que se realizan, en otros casos se tienen que ingresar manualmente algunas noticias relevantes para el informe que se le envía al cliente.

P3 Analista de datos

El proceso de diversos diarios locales es muy demandante, se tiene que estar monitoreando cada uno de ellos constantemente, el aplicativo de extracción no realiza de forma escala cada sitio web de noticias, para esto busca solo los links que contienen noticias, sin embargo, al ser cada sitio web totalmente diferente respecto a la estructura se tiene que monitorear constantemente lo que está extrayendo. El ingreso manual de noticias tiene un tiempo de 20 min por cada diario.

Figura 1. Triangulación de la observación de la unidad de estudio.

Basado en la triangulación de la observación de la unidad de estudio, para la obtención de noticias mediante de web Scraping internamente se usa un aplicativo desarrollado internamente para la extracción este proceso demanda 4 horas por el grupo de sitios de noticias, así también, la extracción manual para esto se tiene que ir al sitio web del diario, buscar noticia y seleccionar los datos de interés este proceso demanda unos 20 minutos por cada diario a buscar. y como una alternativa adicional se puede solicitar a un proveedor especializado este tipo de información, así mismo, este proceso puede demorar entre 10 – 15 por cada cliente, el mantenimiento a la lógica no es muy constante, pero sin embargo solo se extrae ciertos diarios locales, este proceso dura aproximadamente 2 horas esto se integra directamente a una base de datos para su post proceso, donde seleccionan las noticias para clientes específicos. (Figura 1)

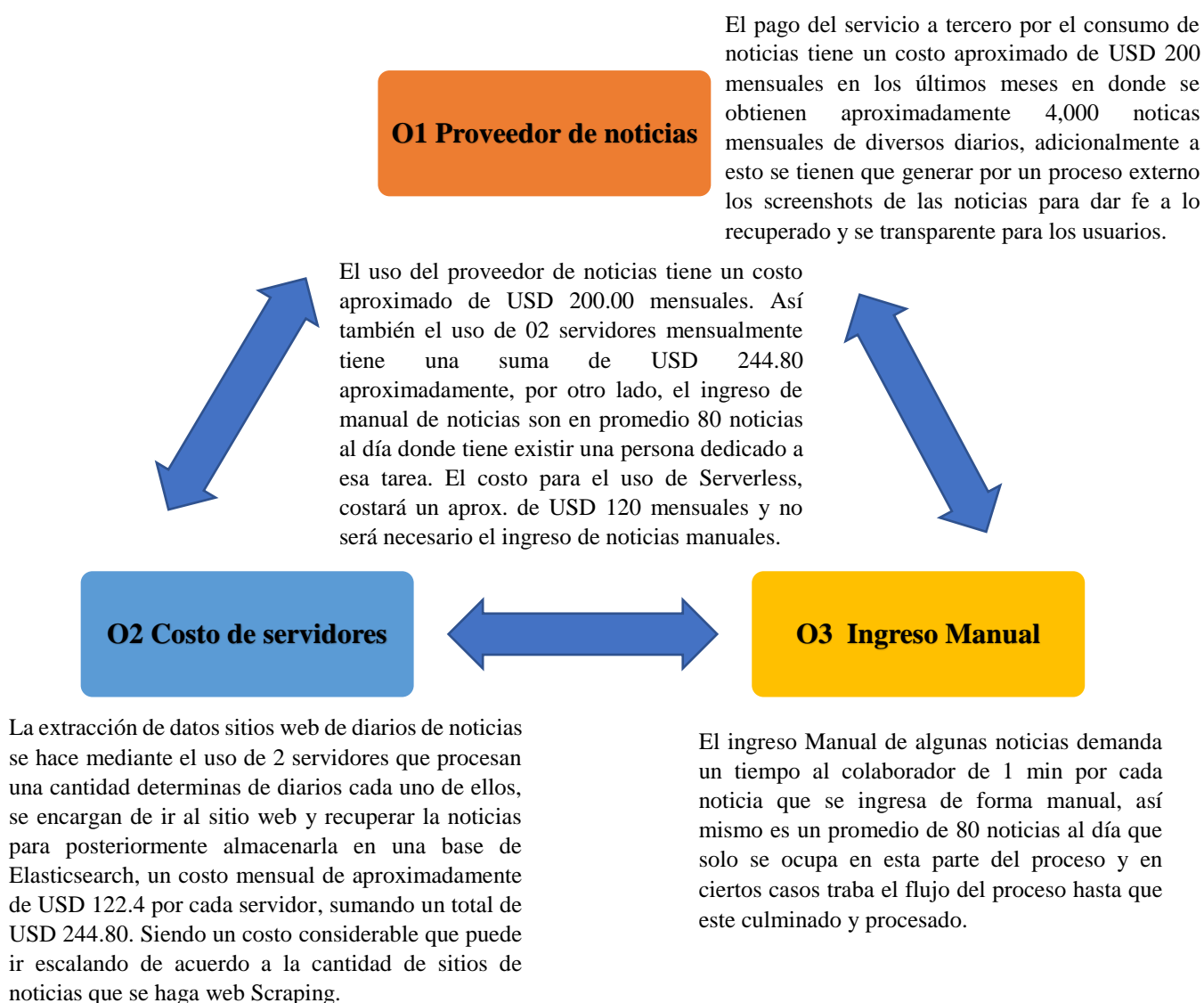


Figura 2. Triangulación del análisis documental.

Basados en la triangulación del análisis documental, el uso del proveedor de noticias tiene un costo aproximado de USD 200.00 mensuales, este costo es por el aproximado de 4,000 noticias seleccionadas, el usuario busca y selecciona no noticias de interés para poder integrarse al proceso. Así también se cuenta con 02 servidores donde correo en cada uno de ellos un grupo de diarios para separar la carga, el costo mensualmente tiene una suma de USD 244.80 aproximadamente en IBM, por otro lado, el ingreso de manual de noticias donde en promedio son 80 noticias al día y hay un personal dedicado para esta tarea. (Figura 2).

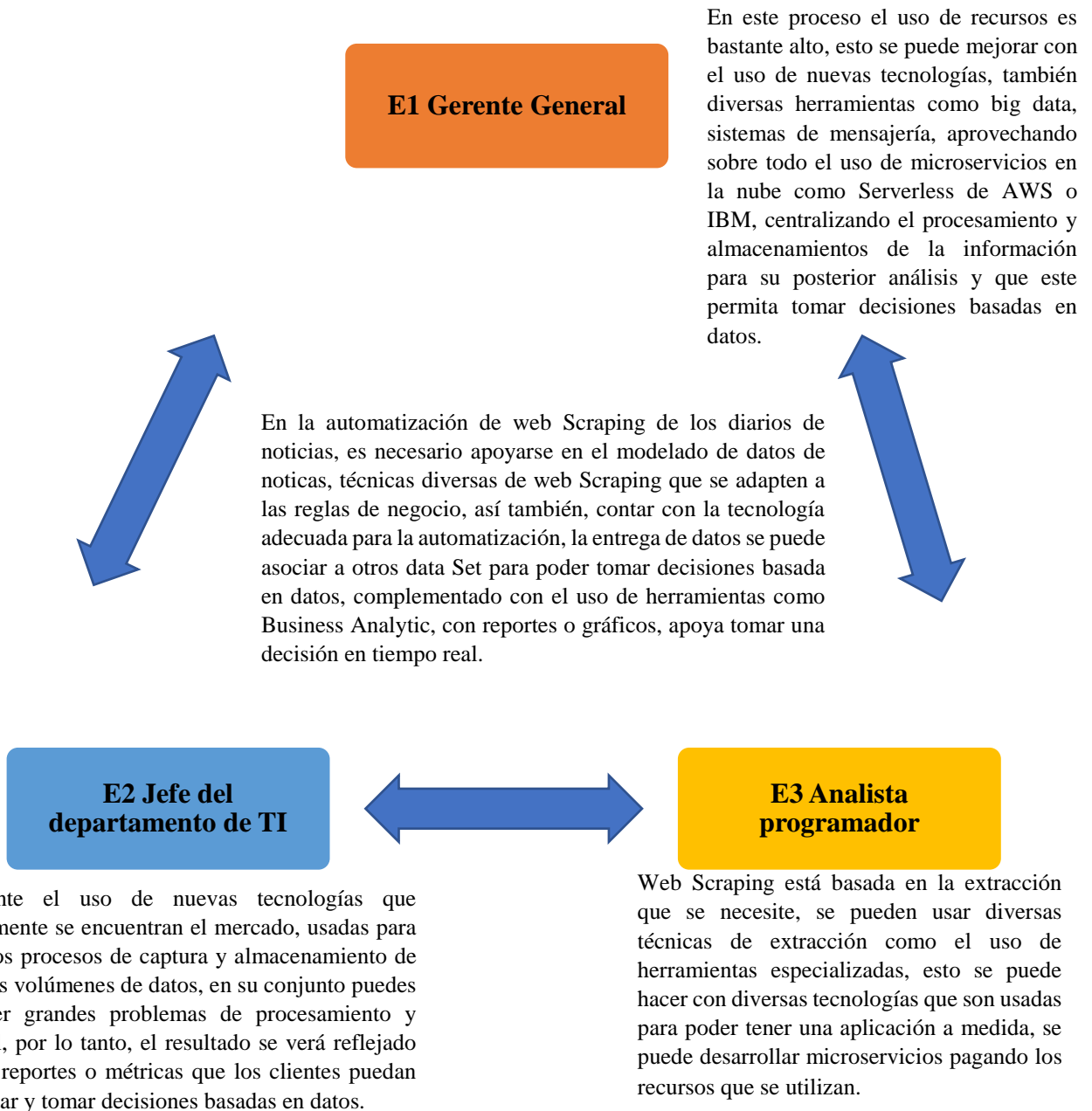


Figura 3. Triangulación de las entrevistas semi estructurada.

Basados en la triangulación de las entrevistas semi estructurada, la automatización de web Scraping de los diarios de noticias, es necesario apoyarse en el modelado de datos de noticas, técnicas diversas de web Scraping que se adapten a las reglas de negocios del negocio, contar con la tecnología adecuada para la automatización con el uso de ETL o ELT durante todo el flujo el web Scraping de noticias, los componentes tecnológicos están compuesto por Serverless por el costo que se genera solo de lo que se usa, Golang por el uso de concurrencia, hilos y fue desarrollado con el propósito de usar varios procesadores al mismo tiempo, un sistema de colas o mensajería que equilibre la carga de datos y sobre una base de datos NoSql. A través de la entrega de datos se puede asociar a otros data Set para poder tomar decisiones basada en datos, complementando con el uso de herramientas como Business Analytic, con reportes o gráficos, apoya tomar una decisión en tiempo real. (Figura 3).

I1 Entrevista semi estructurada

En la automatización de web Scraping de los diarios de noticias, es necesario apoyarse en el modelado de datos de noticas, técnicas diversas de web Scraping que se adapten a las reglas de negocio, así también, contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo, A través de la entrega de datos se puede asociar a otros data Set para poder tomar decisiones basada en datos.

En la automatización de web Scraping de los diarios de noticias, es necesario el modelado de datos de noticas, técnicas diversas de web Scraping, contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo, el uso de 02 servidores mensualmente tiene una suma de USD 244.80 aproximadamente, por otro lado, el ingreso de manual de noticias son en promedio 80 noticias al día donde tiene existir una persona dedicado a esa tarea y el uso del proveedor de noticias tiene un costo aproximado de USD 200.00 mensuales. Con esta investigación se reduciría a una solo forma de extracción y automatizada, reduciendo considerablemente los recursos, tiempos y errores. Así también, el costo para el uso de Serverless, costará un aprox. de USD 120 mensuales y no será necesario el ingreso de noticias manuales.

I2 Observación

Se utiliza 03 maneras de hacerlo, se usa un aplicativo interno que se encarga de buscar materias en unos cuantos sitios, en algunos casos se hace manual la inserción de estas al sistema comercial, y adicionalmente se puede solicitar a un proveedor especializado este tipo de información, aproximadamente 4 horas por el grupo de sitios de noticias, así mismo.

I3 Análisis documental

El uso del proveedor de noticias tiene un costo aproximado de USD 200.00 mensuales. Así también el uso de 02 servidores mensualmente tiene una suma de USD 244.80 aproximadamente, por otro lado, el ingreso de manual de noticias son en promedio 80 noticias al día donde tiene existir una persona dedicado a esa tarea.

Figura 4. Triangulación de las técnicas de investigación utilizadas.

En la automatización de web Scraping de los diarios de noticias, es necesario el modelado de datos de noticias, técnicas diversas de web Scraping, así también, contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo, apoyándose a través de algoritmos personalizados que optimizan el tiempo y el proceso de extracción. se puede complementar con el uso de herramientas como Business Analytic, con reportes o gráficos, apoya tomar una decisión en tiempo real. Actualmente se usa un aplicativo interno, en algunos casos se hace manual la inserción y adicionalmente se puede solicitar a un proveedor especializado este tipo de información, el proceso interno dura aproximadamente 4 horas, así mismo, el seleccionar información de un servicio de tercero, demanda un tiempo entre 25 – 30 minutos adicionales. Así también el uso de 02 servidores mensualmente tiene una suma de USD 244.80 aproximadamente, por otro lado, el ingreso de manual de noticias son en promedio 80 noticias al día donde tiene existir una persona dedicado a esa tarea y el uso del proveedor de noticias tiene un costo aproximado de USD 200.00 mensuales. (Figura 4).

Para la automatización de web Scraping requiere tener una capacidad alta computacional muchas veces superando la arquitectura interna, por lo tanto, es necesario el modelado de extracción de cada página web de noticias, usando diversos algoritmos combinadas con técnicas de web Scraping. Así mismo, se necesita una estructura tecnológica de bajo costo llamada Serverless usando Lambda de Amazon Web Services y servidores locales donde se diseñó una herramienta de control de flujos o ETL. Así como un gestor de colas llamado Kafka que permitirá el control de mensajes que finalmente fueron integrando la información en un base de datos NoSQL de nombre Elasticsearch, Los datos recuperados pueden ser usados por los clientes para mejorar tomar decisiones basada en datos e identificar su posicionamiento de marca, monitorizar a la competencia y realizar métricas personalizadas. así también, el contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo. Así también el uso de 02 servidores y el proveedor externo mensualmente asciende a una suma de USD 689.60 aproximadamente, por otro lado, el ingreso de manual de noticias son en promedio 80 noticias al día. (Figura 5).

I1 Técnicas

En la automatización de web Scraping de los diarios de noticias, es necesario el modelado de datos de noticias, técnicas diversas de web Scraping, así también, contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo. Actualmente se usa un aplicativo interno, en algunos casos se hace manual la inserción y adicionalmente se puede solicitar a un proveedor especializado este tipo de información, el proceso interno dura aproximadamente 4 horas.

Para la automatización de web Scraping requiere tener una capacidad alta computacional muchas veces superando la arquitectura interna, por lo tanto, es necesario el modelado de extracción de cada página web de noticias, usando diversos algoritmos combinadas con técnicas de web Scraping. Así mismo, se necesita una estructura tecnológica de bajo costo llamada Serverless usando Lambda de Amazon Web Services y servidores locales donde se diseñó una herramienta de control de flujos o ETL. Los datos recuperados pueden ser usados por los clientes para mejorar tomar decisiones basada en datos, así también, contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo. Así también el uso de 02 servidores y el proveedor externo mensualmente asciende a una suma de USD 486.60 aproximadamente, por otro lado, el ingreso de manual de noticias son en promedio 80 noticias al día. Con esta investigación se reduciría a una solo forma de extracción y automatizada, reduciendo considerablemente los recursos, tiempos y errores. Así también, el costo para el uso de Serverless, costará un aprox. de USD 120 mensuales y no será necesario el ingreso de noticias manuales.

I2 Antecedentes

La automatización de web Scraping requiere tener una capacidad alta computacional muchas veces superando la arquitectura interna, el uso de micro servicios permitirá optimizar y monitorear el uso de los recursos, para esto fue necesario el uso de algoritmo personalizados, ya que cada sitio web tiene una estructura interna distinta. Identificar las estructuras repetitivas, facilita que el aplicativo desarrollado para que pueda conectarse y extraer la información de sitios Web similares. Esto permite poder monitorizar la competencia con los datos extraídos.

I3 Marco teórico

Para la automatización de web Scraping es necesario el modelado de extracción de cada página web de noticias, usando diversos algoritmos combinadas con técnicas de web Scraping. Así mismo, se necesita una estructura tecnológica de bajo costo llamada Serverless (Sin servidor) usando Lambda de Amazon Web Services (AWS) y servidores locales donde se diseñó una herramienta de control de flujos o ETL. Así como un gestor de colas llamado Kafka que permitirá el control de mensajes que finalmente fueron integrando la información en un base de datos NoSQL de nombre Elasticsearch, también se tuvo en consideración las reglas de negocio de la empresa que permitió tener un aplicativo escalable y robusto. Los datos recuperados pueden ser usados por los clientes para mejorar tomar decisiones basada en datos e identificar su posicionamiento de marca, monitorizar a la competencia y realizar métricas personalizadas.

Figura 5. Triangulación de los antecedentes, marco teórico y los resultados

IV. DISCUSIÓN

En el desarrollo de la presente investigación se realizó la comparación de todos los resultados obtenidos, cada uno de estos llegaron a ser contrastados con la documentación incluida en la tesis, realidad problemática, trabajos previos, artículos indexados, marco teórico, todo ello relacionado con cada uno de los objetivos identificados dentro de la investigación. Por tal motivo el objetivo principal de la tesis fue desarrollar la automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres., para ellos se usó tres técnicas de recolección de datos la guía de entrevista semiestructurada, Guía de observación y ficha de análisis documental. Así mismo, el tipo de investigación es de aplicada tecnológica y finalidad es descriptiva.

Durante el análisis documental de los últimos meses el uso de un proveedor de noticias externo de donde se obtienen las noticias con un costo en promedio de USD 200.00 mensuales (Anexo 8), el este costo es por el aproximado de 4,000 noticias seleccionadas, así mismo, el usuario busca y selecciona no noticias de interés para poder integrarse al proceso. Este proceso se usa debido a la carencia de poder hacer web Scraping a algunos sitios web de noticias, y los sitios donde se extrae la información se necesita el uso de dos servidores donde correo en cada uno de ellos un grupo de diarios para separar la carga, el costo mensualmente tiene una suma de USD 244.80 aproximadamente en IBM (Anexo 8), cada uno de ellos hace Web Scraping, extraen las noticias con ciertos algoritmos internos, se ejecutan cada seis horas para traer un nuevo grupo de noticias que recién se han insertado o publicado.

Luego almacenarlo en un base de datos Nosql denominada Elasticsearch, para complementar el proceso hay un grupo de noticias que no se ingresa manualmente al aplicativo comercial, estas noticias no se ingresan por dos motivos, el primero es porque las colas en base de datos se pierden y la otra es porque el proveedor no logra recuperar todas las noticias de los sitios web, la cantidad de noticias que se ingresa de forma manual es en promedio 80, a esta tarea se tiene que asignar una persona dedicada a este proceso. El uso de estos recursos se reducirá a un valor mensual de USD 120.00 con la automatización de web Scraping de los diarios de noticias, Según Cárdenas y Chaux (2018) exponen en su investigación que la automatización de web Scraping les permitió recolectar información relevante de varias páginas web, ubicación y otras características que se encuentran en

alquiler en las principales ciudades, una vez descarga esta información fue depurada y se procesa en un base de datos unificada donde se clasifica el texto.

De acuerdo a lo observado en la unidad de estudio para completar la extracción de noticias se utiliza tres formas de hacerlo, la primera es el aplicativo interno que su proceso demora en promedio 4 Horas al día, se ejecuta en tres horarios distintos, la inserción manual de noticias se toma 80 min cada día aproximadamente, y como una alternativa adicional se puede solicitar a un proveedor especializado este tipo de información, esto puede demorar entre 10 – 15 min por cada cliente. El recurso usado para la extracción en tiempo suma diariamente cerca de cinco horas y medias por día. Mediante la automatización de web Scraping el uso de estos recursos se verán recudidos al proceso principal de extracción que durara una hora y media. Esto esta reforzado por Stefano, Neves, Santana, Valadão y Hammes (2020), realizar la extracción de información de forma manual requiere el uso de muchos recursos, computacionales, costos y humano, con la implementación de bots este proceso se vuelve muy eficiente para este tipo de tareas.

Así que la automatización de web Scraping para los diarios de noticias, ayudaría a reducir la problemática actual de la empresa, según las entrevistas realizadas a los expertos es importante el modelado de extracción para la recuperación uniforme de solo el texto de noticias, las reglas de negocio permiten identificar como extraer, almacenar y procesar las noticas recuperadas, lo más importante para este proceso es el uso de la tecnología donde está reflejado el mayor consumo de recursos, y aplicando tecnología más reciente podemos usar menos recursos con mejores resultados, donde se considera el lenguaje de programación, sistemas de colas y base de datos NoSql. Esta información recuperada se puede usar para complementar otros informes, reportes o métricas para la toma de decisiones basadas en datos, permitiendo conocer el posicionamiento de la marca y monitorizar a la competencia (Anexo 8).

Dentro de los objetivos específicos esta describir el modelado de extracción para web Scraping de los diarios de noticias para la empresa Isuri, por lo concluido según los tres expertos entrevistados para lograr la automatización de web Scraping de los diarios de noticias, es necesario contar con diversos componentes de diversa importancia dentro del flujo del proceso dentro de ellos se encuentra el modelado de extracción de datos de noticias, usando diversas técnicas de web Scraping dentro del código de programación, que permiten identificar la estructura de una noticia dentro de un HTML y mapear el Xpath donde se

encuentra el texto donde se desarrolla la noticia, de esta manera evitar que pueda extraer otros valores que se encuentran alrededor del texto. Esto es sustentado por Verdes (2016).

Para el modelado de extracción los especialistas mencionan el uso de técnicas de web Scraping más recurrentes a usar son la programación del protocolo de transferencia de hipertexto (HTTP), así como análisis del lenguaje de marcado de hipertexto, se usa lenguaje de consulta semiestructurado, como XQuery y consulta por hipertexto (HTQL) y Análisis del modelo de objetos de documento (DOM). Aprovechando estas técnicas podemos solo extraer el texto de las noticias publicada y aislando otros valores dentro de ellas, ya que normalmente existe publicidad u otras publicaciones dentro de la noticia. Esto esta reforzado en la investigación por (Sirisuriya 2015).

Dentro del modelado de extracción de acuerdo a los especialistas es necesarios limpiarlos datos recuperados de los sitios web que se realizan en el web Scraping, para esto es necesario poder contar con algoritmos personalizados que se centren en extraer la información lo más limpia posible, sin contenido adicional. Permitiendo transformar y clasificar la información obtenida, así mismo enriquecer esa información con datos adicionales, estas aseveraciones se sustentan en lo sostenido Krunal (2014) y Córdova y Quispe (2019), en la guía de observación se identifica que cada diario local tiene una estructura distinta, por lo tanto, el algoritmo debe considerar cada posible estructura y adaptarse a ella para conseguir el resultado deseado, de no ser así se tendría que hacer manualmente la limpieza de las noticias, esto se lograr mediante el uso de técnicas de web Scraping y algoritmos personalizados, esto esta reforzado por Capuñay (2018), que indica que el uso de Web Scraping permitió la recolección masiva de datos.

Que en su conjunto permitirán modelar la extracción de las noticias del sitio web solicitado, permitiendo tener el texto, fecha, autor de las noticias publicadas recientemente, esto forma parte de la primera fase de la automatización de web Scraping. Esto también se encuentra sustentado en la investigación por Haralson (2016), donde se hace mención de algoritmos personalizados con diversos parámetros para el uso de web Scraping. Otras técnicas que se pueden usar para el web Scraping son software especializados como la compra de esta información de noticias a diversos proveedores, pero sin embargo hay información adicional que las noticias necesita para tener una valorización por la llegada que esta pueda tener a un determinado sector, grupo de personas, así también, el enriquecimiento

de la noticia se puede agregar otros servicios adicionales como la clasificación de texto donde se obtiene las palabras principales o claves de las noticias y otros valores

Otro de los objetivos específicos de la investigación consiste en definir las reglas de negocio para web Scraping de los diarios de noticias para la empresa Isuri, sobre esta premisa los especialistas sostienen que se debe tener en cuenta las reglas del negocio que permiten a la empresa Isuri controlar el flujo de las noticias dentro su sistema, las noticias se almacenan para luego ser procesadas, este almacenamiento debe estar de un ecosistema de base de datos dinámico, robusto y escalable. Parte del proceso preparación posterior de estos datos que se agregan información adicional para complementarla, con el objetivo principal de enriquecer los datos para los casos de uso y reglas del negocio para mejorar la calidad. Coincide con lo sustentado por Miranda (2015), donde afirma que el uso de big data está determinado por grandes volúmenes de datos que se puedan generar grandes velocidades y con una variedad de fuentes de información nunca antes vistas hasta ahora.

Dentro de las reglas de negocio para la automatización de web Scraping es relevante el uso de Bots dentro del flujo del proceso, esta parte apoya en poder realizar las diversas tareas que demandarían gran cantidad de tiempo y personas para hacer lo de forma manual, por el volumen de información que debería procesar cada colaborador, así mismo, la cantidad de errores es mínima, y se lograra realizar todas las actividades planificadas para la extracción de noticias de los sitios web. Estos bots tendrá una estructura interna recurrente que le permitirá manejar hilos (Ejecutar varios eventos en paralelo), así mismo el control de fallos que permita tomar una acción ante cada uno de los errores que se puedan presentar, siempre notificando mediante correo electrónico las tareas que se van ejecutando.

En la Guía de observación se identificó que hay cierta parte del proceso que se realiza manualmente, con esta investigación se lograría reducir a lo mínimo cualquier ingreso manual de noticias a la base de datos. Los bots controlaran la extracción de la información de manera automática apoyado con el modelado de la extracción donde se encuentra estructurado el código para extraer solo las noticias y los algoritmos para limpiar los valores que no sean parte de ella. Esto es sostenido por Huamán (2019), donde en su trabajo de investigación los bots y las técnicas de web Scraping permitieron extraer la de forma masiva, automatizada, en tiempo real, mejorando la clasificación y características de la información. Para el control del proceso de la automatización de web Scraping, los especialistas sostienen que el uso de un ETL, sería lo más adecuado para controlar cada etapa por donde pasa el

flujo de información desde la recuperación de las noticias hasta el enriquecimiento con datos adicionales con otros datos y terminando por el almacenamiento NoSql. Según las reglas de negocio es importante identificar la carga que va tener este proceso, es normal extraer, transformar y cargar, el orden indicado cuando se trata de datos finales, básicamente para una estructura comercial donde normalmente esta base de datos SQL, sin embargo, al extraer datos de los sitios web no estructurados, que son cambiantes es necesario dedicar un recurso especial para limpiar y enriquecer los datos durante el proceso. En esta investigación Miranda (2015) y Shekhar y Pandey (2019), sustentan que para un proyecto de esta tamaño que tiene características de Big data es necesario cambiar el orden del proceso, en lugar de usar un proceso ETL (Extract, transform y load) de preferencia sería ideal usar un ELT (Extract, load and transform), donde tras lograr extraer las noticias de la fuente de origen no se realice ninguna limpieza ni transformación, si no se priorice el almacenado de los datos, luego ser organizar, se prepara y se procesa.

En la investigación se identificó que para que esta automatización de web Scraping se pueda desplegar en un proyecto de este tamaño, es necesario tener como gran aliado y principal recurso la tecnología, lo cual otro objetivo de la investigación es identificar las tecnologías para web Scraping de los diarios de noticias, los especialistas sustentaron que es necesario apoyarse en servicios tecnológicos de bajo costo, robusto y escalable, que permita ser el soporte y base de todo el flujo del proceso de la automatización, desde que se extraer y almacena. Para realizar web Scraping se necesita gran cantidad de recursos dedicado para esta tarea, ya que tiene que interactuar con varios sitios web de noticias, donde cada uno de ellos tienen estructuras internas diferentes y a su vez tienen secciones donde se publican gran cantidad de noticias al día, Según la ficha de análisis documental el costo para extraer información con dos servidores y el uso del servicio del proveedor externo asciende a USD 689.60 (Dólares americanos) mensualmente, esta investigación busca reducir el costo a no más de USD 120.00 (Dólares americanos) mensualmente con la misma cantidad de diarios que se procesan actualmente. Es necesario apoyarse en nuevas tecnologías que garanticen su disponibilidad, esto es sustentado por Correa (2018), que expone en su investigación que el uso de una arquitectura escalable para minería de datos en tiempo real y el uso de web Scraping requiere tener una capacidad alta de CPU y memoria RAM, muchas veces superando la arquitectura interna, sin embargo el uso de micro servicios permitirá optimizar y monitorear el uso de los recursos, así mismo alertar de posibles fallas, el no controlar el

aplicativo adecuadamente podría ocasionar la pérdida excesiva de datos y la degradación del servicio.

En el contexto tecnológico los especialistas concuerdan que la captura de las noticias tiene un alto costo computacional, por eso recomiendan usar Serverless donde los costos de los recursos se miden solo del consumo que con lleva ejecutar una función específica, aquí es donde se ejecutara el modelado de extracción y los algoritmos de limpieza de información, cada uno de estos en funciones distintas por cada sitio web de noticias, durante la investigación la mejor arquitectura para Serverless en la nube es AWS, Según Mohamed (2018), que el uso de este tipo de arquitectura tiene diversos beneficios, el costo de mantenimiento es cero, al ser una arquitectura con kubernetes que permite el autoescalado y la alta disponibilidad, así también, es polygot, que se puede usar diversos lenguajes de programación para una sola aplicación.

Otro de los componentes de gran importancia es el lenguaje de programación que se usara para aprovechar esta arquitectura Serverless, junto a todo ello unos de los lenguajes de programación que se ha desarrollado para aprovechar cada uno de los procesadores de forma individual es Golang, un lenguaje desarrollado por Google, con una comunidad creciente, que dentro de sus principales características destacan el uso de concurrencia o hilos que permitirán ejecutar un mismo evento de forma paralela infinitamente, es un lenguaje de programación orientado a la nube, por la reducción en el tiempo de compilación y escalabilidad, su sintaxis es limpia y clara al ser fuerte mente tipado no acepta código que se declare y no se use, así como violaciones los tipos de datos de una variable en concreto, esto es sustentado por Boucher, Kalia, y Andersen (2018), que mencionan que con la finalidad de monopolizar los CPU's, se usa un enfoque que funcione a escalas de microsegundos, aprovechando las multitareas que nos proporciona Go Lang a través de las Gorutinas (hilos ligeros).

En la guía de observación se logró identificar que por la cantidad de información que se recuperaba y se enviaba al servidor de base de datos Elasticsearch, alguna de las noticias no se lograba almacenar correctamente, motivo por el cual ciertas publicaciones se ingresaban manualmente, en la investigación se pretenden solucionar esta situación con un gestor de colas o mensajería, que permita recepcionar e ir almacenando mensaje por mensaje sin perder algunos de ellos en el proceso de extracción, esto es mencionado por los especialistas que sugieren el uso de Kafka como balanceador de carga dentro del proceso

ELT, garantizando la fiabilidad de los mensajes y el alto rendimiento, esto es sustentado por la publicación de (Manish y Chanchal 2017).

De acuerdo a lo sostenido por los especialistas se debe usar base de datos NoSql, para almacenar la información extraída de los sitios web de noticias mediante Scraping, ellos concuerdan que este tipo de base de datos permite ser dinámica a la estructura interna que puede tener una colección o un índice internamente, además de almacenar de forma rápida la información al no depender de un esquema fijo. Según sostiene Namdeo y Suman (2020), Este tipo de base de datos intentan resolver los problemas de escalabilidad y disponibilidad frente a los de atomicidad o consistencia, así mismo, los especialistas indican que el uso de Elasticsearch como base de datos NoSql, ayudara a la segmentación, enriquecimiento y clasificación del texto, para Mishra (2019) las características de Elasticsearch ayudan a ofrecer una mejor oferta de búsqueda e idoneidad para lograr los objetivos de búsqueda actuales y futuros para el sitio web extraídos. se suman Andre, Behrens, Branson, Brummer, Chaze y otros (2018) donde hacen mención que los documentos son inyectados en formato JSON dentro del clúster Elasticsearch utilizando protocolo HTTP, creando de una plantilla con un índice para el mapeo y búsqueda configurada para algunos campos posteriormente.

Durante la investigación otro objetivo específico que se identificó era el de evaluar que es la toma de decisiones basado en datos para web Scraping de los diarios de noticias para la empresa Isuri, según los especialistas esta es la parte final del proceso donde la noticia ya ha sido procesada, depurada y enriquecida, esta parte es donde el cliente consume según ciertos parámetros las noticias que necesite, el cliente usa esta información como complemento de otros datos internos que ellos tienen para conocer el posicionamiento de su marca, monitorizar a la competencia, con el conjunto de información se pueden elaborar diversos reportes, métricas que le permitan cuantificar su valor en el mercado. Lo cual se sostiene por Morales (2010). Es necesario tener en cuenta que, para toma de decisiones basadas en datos, según Kannan y Li (2016), se necesita una estrategia sólida acompañada de una serie de herramientas tecnológicas que le permitan trabajar de manera más inteligente, sin embargo, ninguna de estas aplicaciones logra mágicamente mejorar la toma de decisiones.

En la presente investigación se usaron tres técnicas para la recolección de datos: tales como las entrevistas semi estructuradas, la observación y el análisis documental, se logró realizar el análisis completo a la unidad de estudio, con la finalidad de lograr la

automatización de web Scraping de los diarios de noticias. Mediante la entrevista a los tres especialistas externos de la institución con mayor experiencia en este campo se logró reducir el uso de los recursos computacionales, humano, presupuesto y errores, mediante lo aportado recomendaron el uso de técnicas de web Scraping, algoritmos, tecnología robusta, escalable, un óptimo lenguaje de programación, el balanceo de la carga para las colas de mensajes, el procesamiento del texto, enriquecimiento del texto, finalmente para que los clientes de la empresa puedan tomar decisiones basadas en los datos, entre otros, lo mencionado se contrasta con búsqueda bibliográfica consultada en el marco teórico. El cual ha mostrado métodos adicionales de control y mejorar la automatización que demanda mayor inversión y tiempo para poder aplicar (Anexo 8).

V. CONCLUSIONES

Primera:

El presente trabajo concluye que el factor más importante para la automatización de web Scraping es el uso fundamental de la tecnología, que usando los componentes adecuados como Serverless, Golang, Sistema de mensajería y base de datos Nosql, enfocado esto en un proceso ELT, para poder controlar el flujo de la captura de noticias; por lo tanto, se logró la reducción de recursos computacionales, presupuesto y errores.

Segunda:

Respecto al modelado de extracción de noticias, de acuerdo a lo sustentado por los expertos y la revisión bibliográfica es necesario el uso de técnicas de web Scraping para la extracción de solo las noticias, y apoyarse en los algoritmos ya existentes o personalizados que permitan limpiar y enriquecer las noticias, considerando que cada sitio web noticioso tiene una propia estructura HTML personalizada.

Tercera:

Las reglas de negocio de la empresa Isuri, permitió identificar como funciona el flujo de información desde la captura de las noticias hasta el almacenamiento, lo que permitió considerar las bases necesarias para la automatización de web Scraping, los datos se alojaron en un Big data, la extracción se realizó mediante bots y el control del proceso se realizó mediante un proceso ELT (Extraction, Load, Transformation).

Cuarta:

La tecnología dentro de esta investigación tuvo un papel de alta importancia para la automatización de web Scraping, esto permitió ser más eficiente y reducir los recursos usados para la captura de noticias, el estudio concluye que el usar Serverless en lugar de un servidor la reducción de los costos es muy evidente, aprovechar los diversos procesadores que tienen esta arquitectura para usar Hilos en la ejecución de funciones mediante Golang, usar un sistema de mensajería para balancear la carga de los paquetes y almacenar los datos de forma dinámica a través de bases de datos NoSql.

Quinta:

La finalidad de extraer las noticias y enriquecerlas, es que se pueda usar esta información para que las empresas puedan tomar decisiones basadas en datos, mediante reportes, métricas

les permite conocer cómo se encuentra su marca y monitorizar a su competencia, así mismo estos datos sirven de complementos para cruzar con otros tipos de resultados, dentro del Business Analytic, permitiendo el análisis predictivo y análisis de los datos alineados a la estrategia de la empresa.

VI. RECOMENDACIONES

Primera:

Se recomienda al responsable del departamento de tecnología de la información, que el proceso de automatización de web Scraping se encuentre alojado en la nube sobre todo el mineración de las noticias, por tener una alta disponibilidad para el proceso de extracción, el conjunto de datos de la recolección se debe guardar en una base de datos NoSql, y posterior de su procesamiento, depuración y enriquecimiento.

Segunda:

Se recomienda al responsable del departamento de tecnología de la información, evaluar la estructura de cada sitio web de noticia para poder determinar cuál será el procedimiento o técnica a usar para ese tipo de diario, y la posibilidad de usar algoritmos que se asemejen más al comportamiento humano, donde se carga la página web totalmente antes de realizar la extracción.

Tercera:

Se recomienda al responsable del departamento de tecnología de la información, revisar periódicamente el mantenimiento de los bots, y mejorar su autonomía respecto a la ejecución de eventos respecto a algunos inconvenientes que puedan identificarse luego de la automatización, como por ejemplo la falta de conectividad al servidor de las noticias o base de datos, la respuesta tardía de las funciones en Serverless, u otras situaciones.

Cuarta:

Se recomienda al responsable del departamento de tecnología de la información, implementar este tipo de soluciones en la nube, para montar los bots como servicios se debe usar Docker, así también, para el control del flujo de proceso de automatización de web Scraping es ideal usar Nifi apache en conjunto con Kafka.

Quinta:

Se recomienda al responsable del departamento de tecnología de la información, que la información enriquecida cuente con los datos necesarios para la generación de reportes o métricas de interés de las empresas, tener en consideración que siempre hay algún valor nuevo que se puede recuperar con los avances tecnológicos.

REFERENCIAS

- Alvares C., Muñiz L., Morán J., Merchán L., Conforme G, Nevárez E., (2019). Romero R. Las ideas de negocios, el emprendimiento y el marketing digital. Alacant, España: 3ciencias. ISBN: 978-84-120756-7-0
- Andre, J., Behrens, U., Branson, J., Brummer, P., Chaze, O. y otros (2018). A scalable online monitoring system based on elasticsearch for Distributed Data Acquisition in CMS. *23 rd International Conference on Computing in High Energy and Nuclear Physics*. DOI: 10.1051/epjconf/201921401048. Recuperado en <https://bit.ly/2LN3mEY>.*
- Bardin., L (1996). El análisis del contenido (Searez César, Trad.). Madrid: Ediciones AKAL. (Obra original publicada en 1977).
- Bharvi D. (2016). Elasticsearch Essentials. Birmingham, Reino Unido: Packt Publishing. ISBN: 978-1-78439-101-0. Recuperado en <https://bit.ly/2zqxDqa>.
- Biradar, S., Shekhar, R. y Reddy, A. (2018). Build Minimal Docker Container Using Golang. Madurai, India: Second International Conference on Intelligent Computing and Control Systems. DOI: 10.1109/ICCONS.2018.8663172. Recuperado en <https://bit.ly/2WRnj3V>.*
- Brenner, S., Wulf, C., Goltzsche, D., Weichbrodt, N., Lorenz, M. y otros (2016). New York, United States: *Association for Computing Machinery*, DOI: 10.1145/2988336.2988350. Obtenido en <https://bit.ly/2LIubdc>.*
- Borda, M. (2009). Metodos Cuantativos Herramientas para la Investigacion en Salud. Barranquilla, Colombia: Universidad del Norte, ISBN: 978-958-741-010-5. Recuperado en <https://bit.ly/3e9gX5w>.
- Borrego, F. (2016). Alternativas para realizar Web Scraping. Feliciano Borrego Recuperado en <https://bit.ly/35wYP2g>.
- Boucher, S., Kalia, A., y Andersen, D. (2018). Putting the “Micro” Back in Microservice. Boston, USA: *2018 USENIX Annual Technical Conference (USENIX ATC '18)*. ISBN 978-1-939133-02-1. Recuperado en <https://bit.ly/2ZkQKwE>.*
- Capuñay, C. E. (2018). Análisis del perfil profesional de los profesionales de ingeniería de sistemas en instituciones del sector publico peruano a partir de la aplicación de técnicas de inteligencia de negocios y Scraping. Universidad Pedro Ruiz Gallo, Lambayeque - Perú.
- Cárdenas, J. y Chau F. (2018). Una base de datos de precios y características de vivienda en Colombia con información de Internet. Colombia: *Revista de Economía del Rosario*, 22(1), 75-100. DOI: 10.12804/revistas.urosario.edu.co/economia/a.7768. Recuperado en <https://bit.ly/2M3LXbn>.*
- Carpio M., A., Hanco G., M., Cutipa L., A., & Flores M., E. (2019). Estrategias del marketing viral y el posicionamiento de marca en los restaurantes turísticos de la Región de Puno. Puno, Perú: *Comunicación vol.10 no.1 Puno ene./jun. 2019*. DOI: 10.33595/2226-1478.10.1.331. Recuperado en <https://bit.ly/2WVnjjq> *

- Castañeda, E. B. (2016). *Propuesta de patrón de diseño de software orientado a prevenir la extracción automatizada de contenido web*. Pontificia Universidad Católica Del Perú, Lima - Perú.
- Cabezas, Andrade y Johana (2018). *Introducción a la metodología de la investigación científica*. Sangolquí, Ecuador: Universidad de las Fuerzas Armadas ESPE. ISBN: 978-9942-765-44-4.
- Cbr (2019). Un tercio del tráfico web en 2018 fue Bots, Computer business review <https://bit.ly/3feNg4q>
- Cegarra J. (2016). *Metodología de la investigación científica y tecnológica*. Albasanz, Madrid: Ediciones Díaz de Santos. ISBN: 978-84-9969-027-8. Recuperado en <https://bit.ly/2ZywO9M>.
- Córdova, C. A. (2019). *Metodología basada en Minería de Datos para la detección de usuarios Influencers en Twitter*. Universidad Nacional de Trujillo. Trujillo, Perú.
- Correa, D. A. (2018). *Propuesta de una arquitectura distribuida altamente escalable para minería de datos en tiempo real en redes sociales*. Universidad Central Del Ecuador. Quito – Ecuador.
- Debois, P. & Ferguson D. (2017). *Serverless Architectures on AWS*. Manning Publication Co. EEUU. ISBN 9781617293825.
- Del Alcázar Ponce, J. P. (2015). *Cifras, estadísticas y estado del e-commerce en Ecuador*. Formación Gerencial: <https://bit.ly/3fhwcLd>.
- Dipti M. (2019). Comparative Study of ETL and E-LT in Data Warehousing. *International Research Journal of Engineering and Technology*, e-ISSN: 2395-0056. Recuperado en <https://bit.ly/3dvc3zs>.*
- Dobbelaere, P. y Sheykh, K. (2017). Industry Paper: Kafka versus RabbitMQ. Barcelona, Spain: *DEBS '17: Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, DOI: 10.1145/3093742.3093908. Recuperado en <https://bit.ly/2ZlxvTY>.*
- Escuela de Datos. (2016). *Introducción a la extracción de datos de sitios Web: Scraping*. Schoolofdata <https://bit.ly/35uSuV5>.
- Eloisa V, y Mirko U. (2013). Exploiting web scraping in a collaborative filteringbased approach to web advertising. Barcelona, España. *Digital Technology Center*. DOI: 10.5430/air.v2n1p44, Recuperado en <https://bit.ly/2WKO2Oi>.*
- EL-Sanosi, I. y Ezhilchelvan, P. (2018). Improving ZooKeeper Atomic Broadcast Performance When a Server Quorum Never Crashes. Newcastle Upon Tyne, UK: *EAI Endorsed Transactions on Energy Web and Information Technology*. DOI: 10.4108/eai.10-4-2018.154455. Recuperado en <https://bit.ly/3ebVt84>.*
- eMarketer (2014). World wide ecommerce sales to increase nearly 20% in 2014. <https://bit.ly/3dC4BT0>.
- Fernández, A. (2004). *Investigación y técnicas de mercado*. Madrid, España: Esic Editorial. ISBN: 84-7356-392-1. Recuperado en <https://bit.ly/2WWRir5>.

- Franke, B., Roscher, R., Annie, E. (2016). Statistical Inference, Learning and Models in Big Data. Oxford, Usa: *International Statistical Review / Volume 84*. <https://doi.org/10.1111/insr.12176>. Recuperado en <https://bit.ly/2WJLG36>.*
- Ghosh, B., Banerjee, D. y Sengupta, S. (2016). An Intelligent Survey of Personalized Information Retrieval using Web Scraper. *I.J. Education and Management Engineering, 2016*. DOI: 10.5815/ijeme.2016.05.03. Recuperado en <https://bit.ly/36bQGRc>.*
- Guarav V. (2016). Getting Started with NoSQL. Birmingham, Reino Unido: Packt Publishing. ISBN: 9781849694988. Recuperado en <https://bit.ly/2ztDYkC>.
- Guo, Z. y Ding, S. (2018). Adaptive replica consistency policy for Kafka. Chongqing, China: *International Conference on Smart Materials, Intelligent Manufacturing and Automation (SMIMA 2018)*. DOI: 10.1051/mateconf/201817301019. Recuperado en <https://bit.ly/2LSgmsV>.*
- Haralson, D. (2016). Automating Website Crawling Using Web Scraping Techniques Provided by PHP. Helsinki Metropolia University of Applied Sciences. Helsinki – Finlandia.
- Hernández, R. y Mendoza, C. (2018). Metodología de la Investigación. México, México: McGraw Hill. ISBN: 978-1-4562-6096-5.
- Heydt, M. (2018) Python Web Scraping Cookbook. UK. Published by Packt Publishing Ltd. ISBN 978-1-78728-521-7. Recuperado en <https://bit.ly/2Wj2VXC>.
- Ipnoticias - latam (2019), Monitoreo de medios <https://bit.ly/2zV5ZSJ>.
- IBM (2020), Instancia de servidor virtual [Figura]. Recuperado en <https://ibm.co/30B9EOW>.
- Kannan, P. y Li, H. (2016). Digital marketing: A framework, review and research agenda. United States: *International Journal of Research in Marketing*. DOI: 10.1016/j.ijresmar.2016.11.006. Recuperado en <https://bit.ly/36oHh97>.*
- Klimovic, A., Wang, Y. y Kozyrakis., C. (2018). Understanding Ephemeral Storage for Serverless Analytics. Boston, USA: *The 2018 USENIX Annual Technical Conference (USENIX ATC '18)*. ISBN: 978-1-939133-02-1, Recuperado en <https://bit.ly/2ZiETPR>.*
- Knewin (2020), Monitoreo de noticias en tiempo real: ¿por qué es importante?, <https://bit.ly/2Ywol6s>.
- Koch, C. y Gyrd-Jones, R. (2019). Corporate brand positioning in complex industrial firms: Introducing a dynamic, process approach to positioning. Lund, Suecia: *Industrial Marketing Management*. DOI: 10.1016/j.indmarman.2019.03.011. Recuperado en <https://bit.ly/2Zv5edq>.*
- Krunal, V. (2014), Content Evocation Using Web Scraping and Semantic Illustration, Rajkot, India: *R.K. University*, ISSN: 2278-0661. Obtenido en <https://bit.ly/35tjBjm>.*
- Live On Soluções. Desenvolvimento de Aplicativo iOS Android e Sistemas para internet com parceria Vtex. Obtenido en <https://bit.ly/2Bo1znU>.
- Loreto V. (2019). Web Scraping: Intro para “desorientados” <https://bit.ly/2VZZc2r>.

- Lucas, A., Favarim, F., Felipe, J., Cris, R. y Todt E. (2019). Sensor Monitoring and Supervision Using Web Applications and Rest API. New York, United States: *Proceedings of the Intelligent Embedded Systems Architectures and Applications Workshop 2019*. DOI: 10.1145/3372394.3372398. Recuperado en <https://bit.ly/2TtCxd0>.*
- Malawski, M., Gajek, A., Zima, A., Balis, B. y Figiela, K. (2017), Serverless execution of scientific workflows: Experiments with HyperFlow, Krakow, Poland: *AWS Lambda and Google Cloud Functions*. Krakow, Poland: AGH University of Science and Technology, Department of Computer Science. DOI: 10.1016/j.future.2017.10.029. Recuperado en <https://bit.ly/36eunud>.*
- Malki, M., Hamadou, H., Chevalier, M., Péninou, A. y Teste, O. (2018) Querying Heterogeneous Data in Graph-Oriented NoSQL Systems. Toulouse, France: International Conference on Big Data Analytics and Knowledge Discovery. ISBN: 978-3-319-98539-8. Recuperado en <https://bit.ly/3g7YNT0>.*
- Manish K. & Chanchal S. (2017). Building Data Streaming Applications with Apache Kafka. Birmingham, Reino Unido: Packt Publishing. ISBN: 978-1-78728-398-5. Recuperado en <https://bit.ly/2YWUiVJ>.
- Martínez, C. y Gonzales, G. (2014). Técnicas e instrumentos de recogida y análisis de datos. Madrid: Editorial UNED. ISBN: 978-84-362-6822-5. Recuperado en <https://bit.ly/3c1IMwq>.
- Mieles, M., Tonon, G. y Alvarado, S. (2012). Investigación cualitativa: el análisis temático para el tratamiento de la información desde el enfoque de la fenomenología social. Universitas humanística. Bogotá, Colombia. Recuperado en <https://bit.ly/2AYtA5f>.
- Miranda, A. (2015). Big Intelligence: Nuevas capacidades Big data para los sistemas de vigilancia estratégica e inteligencia competitiva. Madrid – España: Fundación EOI. ISBN: 978-84-15061-8. Recuperado en <https://bit.ly/2SIe3wn>.
- Mohamed, L. (2018). Hands-On Serverless Applications with Go: Build real-world, production-ready applications with AWS Lambda. Birmingham, Reino Unido: Packt Publishing. ISBN: 978-1-78913-461-2. Recuperado en <https://bit.ly/2WBi86I>.
- Mishra, A. (2019). Moving search on science direct to elasticsearch. *Proceedings of the 3 rd Annual RELX Search Summit*. Recuperado en <https://bit.ly/2ZiqV0m>.
- Namdeo B., Suman U. (2020) Análisis de rendimiento de los enfoques de diseño de esquemas para la migración de bases de datos RDBMS a NoSQL. Springer, Singapur: *Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 94*. DOI: 10.1007/978-981-15-0694-9_39. Recuperado en <https://bit.ly/3cPBiwC>.*
- Ñaupas, H. (2014). Metodología de la investigación: cuantitativa-cualitativa y redacción de la tesis. Bogotá, Colombia: Ediciones de la U. ISBN: 978-958-762-188-4. Recuperado en <https://bit.ly/3gIVVmm>.
- Onyancha, J., Plekhanova, V. y Nelson, D. (2017). Noise Web Data Learning from a Web User Profile: Position Paper. London, U.K: *Proceedings of the World Congress on Engineering 2017 Vol II*. ISSN: 2078-0966. Recuperado en <https://bit.ly/3cM5SHI>.*

- Powell, B., Nason, G., Elliott, D., Mayhew, M., Davies J. y Otros (2017). Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting. UK: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. DOI: 10.1111/rssa.12314. Recuperado en <https://bit.ly/2LO4I7N>.*
- Quispe, P. (2019). Desarrollo de una aplicación móvil para el acceso a información de los servicios básicos de los usuarios en la ciudad de Piura. Universidad Nacional De Piura, Piura – Perú.
- Ramos, A. y Ramos, R. (2014). Aplicaciones Web. Madrid, España: Ediciones Paraninfo. ISBN: 978-84-283-9875-6. Recuperado en <https://bit.ly/3aYVLxj>.
- Sandulescu, A., (2018). Fundamentos de métrica digital en Ciencias de la Comunicación. Barcelona, España: Editorial UOC. ISBN: 978-84-9116-916-1.
- Serrano, D., Stroulia, E., Lau, D. (2017). API REST vinculadas: un middleware para la integración semántica de API REST. Honolulu, United States: *Conferencia Internacional IEEE 2017 sobre servicios web*. DOI: 10.1109 / ICWS.2017.26. Recuperado en <https://bit.ly/3bKYMBG>.*
- Shekhar, C. y Pandey, P. (2019), A review of big data environment, tools and challenges. India: *Research Scholar A.P.S University Rewa*. ISSN: 2349-5162. Recuperado en <https://bit.ly/3cKif1U>. *
- Sirisuriya, S. (2015). A Comparative Study on Web Scraping. Rathmalana, *Sri Lanka: International Research Conference*. Obtenido en <https://bit.ly/2Sxdl5a>. *
- Stefano, E., Neves, F., Santana, F., Valadão, G. y Hammes, G. (2017), Análise da evolução da pesquisa em engenharia de transportes. Brasil: *Brazilian journals v. 6, n. 2*, p. 6424-6439. DOI: 10.34117/bjdv6n2-081. Recuperado en: <https://bit.ly/2ZuXcRJ>.*
- Store.Vtex (2020), Tipos de implementación [Figura]. Recuperado en <https://bit.ly/30Bk2Gq>.
- Tracey, S. (2012). Data Driven Decision Making for Small Businesses: Unleashing the Power of Information to Drive Business Growth. Numerical Insights LLC. ISBN: 978-1470187453. Obtenido en <https://bit.ly/3e11btG>.
- Verdes, D. (2016). E-Force Scraper. Universidad Juame I, Castelló, España.
- Villegas, L. (2011). Investigación y práctica en la educación de personas adultas. Valencia, España: Nau L Libres. ISBN: 978-84-7642-813-9. Recuperado en <https://bit.ly/2WYzR9F>.
- Vukovic, D. y Dujlovic, I. (2016), Facebook messenger bots and their application for business, Belgrade, Serbia: *2016 24th Telecommunications Forum (TELFOR)*, DOI: 10.1109/TELFOR.2016.7818926. Recuperado en <https://bit.ly/2WU2i7p>.*
- Wu, Shang y Wolter (2019). Performance Prediction for the Apache Kafka Messaging System. International Conference on High Performance Computing and Communications. Zhangjiajie, China: *EEE 5th International Conference on Data Science and Systems*. DOI: 10.1109/HPCC/SmartCity/DSS.2019.00036. Recuperado en <https://bit.ly/2ZC6bRv>.*

Yuni, J. y Ariel, C. (2006). Técnicas Para Investigar 2. Argentina: Editorial Brujas 2da Edición. ISBN: 987-591-020-1. Recuperado en <https://bit.ly/2WYBXq>.

Anexo 1

Matriz de categorización

Título: Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres

Autor: Antonio Federico Martínez Núñez

Problema General	Objetivo General	Categorías	Subcategorías	Técnicas	Instrumentos
¿Cómo se automatiza el web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?	Desarrollar la automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres	Modelado de Extracción	<ul style="list-style-type: none"> ▪ Web index ▪ Técnicas de Web Scraping ▪ Algoritmos 	Entrevista	Guía de Entrevista
Problemas Específicos	Objetivos Específicos:	Regla de negocio	<ul style="list-style-type: none"> ▪ Big data ▪ Bot ▪ ETL / ELT 		
¿Cómo es el modelado de extracción para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?	Describir el modelado de extracción para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres	Tecnología	<ul style="list-style-type: none"> ▪ Serverless ▪ Go Serverless ▪ Sistemas de mensajería ▪ NoSql 	Observación	Guía de observación
¿Cómo son las reglas de negocio para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?	Definir las reglas de negocio para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres	Toma de decisiones basado en datos	<ul style="list-style-type: none"> ▪ Posicionamiento de marca y Monitorizar a la competencia ▪ Reporte y Métricas ▪ Análisis predictivo y análisis de datos alineados a la estrategia 	Análisis documental	Ficha de análisis documental
¿Cuáles son las tecnologías para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?	Identificar las tecnologías para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres				
¿Cómo se interpreta la toma de decisiones basado en datos para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres?	Evaluar la toma de decisiones basado en datos para web Scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres				

Fuente: Miranda (2015)

Anexo 2:

Preguntas para la entrevista semi estructurada

“Automatización de web Scraping de los diarios de noticias para la empresa Isuri”

1. ¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?
2. ¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?
 - a. ¿En qué consiste la indexación de las páginas web por los grandes buscadores?
 - b. ¿Qué técnicas de web Scraping que usted conoce permitirán ser más eficiente el modelado de extracción?
 - c. ¿Cuáles son las funciones de los algoritmos personalizados para el web Scraping?
3. ¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?
 - a. ¿Considera que la automatización de web Scraping tiene las características de un proyecto de Big Data? ¿Por qué?
 - b. ¿Cuál es la importancia del uso de bots en la automatización del proceso?
 - c. ¿Según su experiencia cual es mejor forma de procesar la información extraída por ETL o ELT?
4. ¿Cuál es la importancia de la tecnología en la automatización de web Scraping?
 - a. ¿Cuál es el impacto del uso de Serverless en la automatización del web Scraping?
 - b. ¿De qué manera el uso del Lenguaje de programación Go impactara en automatización del web Scraping?
 - c. ¿Cuál es la función del uso del servicio de sistemas de mensajería(colas)?
 - d. ¿Qué tan importante es el uso de NoSql en el almacenado de la información?
5. ¿Qué componentes de la tecnología son más usados en la automatización del web Scraping?
 - a. ¿Cuál es el impacto del uso de Serverless en la automatización del web Scraping?
 - b. ¿De qué manera el uso del Lenguaje de programación Go impactara en automatización del web Scraping?
 - c. ¿Cuál es la función del uso del servicio de sistemas de mensajería(colas)?
 - d. ¿Qué tan importante es el uso de NoSql en el almacenado de la información?
6. ¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?
 - a. ¿En qué consiste el posicionamiento de una marca y monitorizar a la competencia con respecto a la toma de decisiones basada en datos?
 - b. ¿En qué consiste los reporte y métricas respecto a la toma de decisiones basada en datos?
 - c. ¿En qué consiste el análisis predictivo y análisis de datos alineados a la estrategia respecto a la toma de decisiones basada en datos?

Anexo 3:

Matriz de desgravación de la entrevista

N.º	Preguntas	Entrevistado 1 – Gerente General
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	En el contexto actual la empresa tiene un trabajo manual para este proceso, el uso de los recursos es bastante alto, esto se puede mejorar con el uso de nuevas tecnologías, apoyado en lenguajes de programación de gran impacto que usen concurrencias como Golang, también herramientas para big data, sistemas de mensajería, aprovechando sobre todo el uso de microservicios en la nube como Serverless de AWS o IBM, ya que es un servicio optimizado e ideal para este proceso de automatización, centralizando el procesamiento y almacenamientos de la información para su posterior análisis que permita tomar decisiones, apoyados con gráficos o métricas que el cliente puede generar a su necesidad para el monitoreamiento de las noticias de sus interés.
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	Se tiene que seguir modelo que se pueda estar usando en la actualidad, como los grandes buscadores que indexan su información a través de bots, con diversas técnicas de web Scraping o framework que están alienados a algunas formas de trabajo. Para poder ser más eficientes, en el proceso de extracción se puede lograr con el uso de algoritmos que estén dirigidos a ciertos grupos de sitios de noticias similares, aprovechando toda esta revolución tecnológica. Los algoritmos personalizados permiten interactuar con diversas web con cierta similitud, a través de ciertas librerías o mecanismos personalizados se puede generar un código por cada sitio web, ya que cada página web tiene una forma de mostrar su información.
3	¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	La regla de negocio para la captura de la información se podría automatizar a través de la nube, por los gran de volúmenes de información entraría como un proyecto de Big data, el uso de bots en este proceso aceleraría y automatizaría la mineración de información, reduciendo considerablemente las diversas cantidades de errores que se encuentran en este proceso. Así mismo el uso de ETL a lo largo de este proceso permitirá tener un mayor control del procesamiento de los datos en los diversos servicios locales o en la nube, siendo más óptimo subir lo datos ya procesados al final de este proceso para reducir el uso de más recursos.
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	El uso adecuado de la tecnología permitirá poder reducir el nivel de error que, en gran parte del proceso, ya que controlando los datos por un flujo automatizado y en cada etapa de mineración y enriquecimiento de la información. Así mismo el poder usar micro servicios como parte de la solución, permitiría que este se más robusto y escalable, en el mercado existen diversos tipos de tecnología que permiten centralizar los datos, a través de colas o balanceadores como el Kafka o Rabbit.
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	Hay varias tecnologías que han evolucionado con el tiempo permitiendo madurar todo este ambiente, existe un servicio dedicado a la ejecución de funciones llamado Serverless, permite trabajar con concurrencias y en bloques, acompañado del lenguaje de programación Golang que optimiza los recursos donde se ejecuta este proceso, ya que está orientado a este tipo de proceso. El uso de sistemas de mensajería como Kafka para el control de colas por la cantidad de información a procesar, también el uso de bases de datos NoSql, que permite ser dinámico al momento de almacenar la información
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	Toda esta información extraída con el web Scraping es información en duro, se tendría que pasar por un proceso de análisis, generando reportes gráficos o métricas para ser interpretados de manera más rápida para la toma de decisiones, las empresas con estos datos puedan tener una visión rápida de lo que pueda estar pasando con su marca, si se está moviendo en forma positiva o negativa dentro del mercador, así mismo analizar a su competencia. Con la información obtenida se puede realizar diversas estrategias que permitan anticipar alguna situación desfavorable o oportuna para el negocio.

N.º	Preguntas	Entrevistado 2 – Jefe del departamento de TI
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	Se podría automatizar el web Scraping mediante el uso de nuevas tecnologías que actualmente se encuentran en el mercado, estas son usadas para diversos procesos de captura y almacenamiento de grandes volúmenes de datos, cada una de ellas por separado cumplen una función específica, pero en su conjunto pueden resolver grandes problemas de procesamiento y control, como Nifi, Kafak y Elasticsearch. Con la aplicación de estas herramientas la empresa va a lograr tener mejor información, por lo tanto el resultado se verá reflejado en los reportes o métricas que los clientes puedan necesitar y tomar decisiones basadas en datos.
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	El uso de indexación de datos como el Crawling y el Scraping, así como vienen siendo usadas por grandes buscadores web como, por ejemplo, Google, Yandex. Otro punto importante es el uso de técnicas de Scraping que podrían facilitar en obtener la información en mejor tiempo y actualizado, el uso de un software especializado o librerías que permitan interactuar con cada DOM dentro de HTML, en algunos casos los sitios web son estáticos y en otros dinámicos, por este motivo, se puede usar algún algoritmo que ya existe o desarrollar un nuevo algoritmo para obtener los datos a medida, según se requiera dentro de la extracción de noticias.
3	¿En qué consisten las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	Dentro de una empresa consiste en tener diversos procesos automatizados, que permitan construir una arquitectura de Big data o Data lake que se encargaran de soportar el almacenamiento de masivo de datos en bruto, ya que su arquitectura es plana e ideal para este proceso. El uso de bot apoya en la resolución de diversos problemas de tiempo, hacer esto manualmente por el volumen de datos que existe es prácticamente imposible, por eso motivo Google tiene la cantidad de bots que tiene rodando en la internet para poder obtener la información actualizada que es la espina dorsal de su negocio. Acompañado de un ETL o ELT durante todo el proceso, esto facilitara que se pueda ir interactuando en cada punto por donde la información va pasando para ser depurada y enriquecida.
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	El papel que juega la tecnología dentro de este proceso es demasiado importante, ya que gracias a ello el trabajo manual se verá reducido, sería una carga humana prácticamente insostenible, como décadas atrás donde el costo sería hiper elevado, lo cual con llevaría a no poder desarrollar negocios como los que actualmente existen, estas tecnologías están compuestas por diversas herramientas que facilitaran el poder recolectar, almacenar y procesar todo tipo de información.
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	El uso constante de la base de datos, así como de los servidores, CPUs, balanceadores y otros, para lo cual se puede usar una alternativa llamada Serverless, que soporta diversos lenguajes de programación como Golang por las concurrencias que maneja, El NoSql es una de las alternativas que se puede usar para este tipo de proceso debido a los grandes volúmenes de información que va depender mucho de la necesidad del proceso, se puede usar mongo, Elasticsearch, o spark. Los sistemas de mensajería como Kafka permiten tener un control e interacción con los datos que van llegando para almacenarse de forma masiva, permitiendo controlar la sobre carga en las validaciones que existen antes de insertar o actualizar la base de datos.
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	El poder tomar decisiones analizando resultados en tiempo real desde varios orígenes de datos, buscando posicionar la marca, son muchas empresas que utilizan este tipo de solución para conocer la salud de su marca comercial y a su vez monitorear a sus competidores. Existen diversos tipos de métricas o reportes que le va permitir a la empresa tomar decisiones a futuro basado en la mineración de datos. El poder aplicar el análisis predictivo en una organización viene acompañado de complementar diversos orígenes de datos, para obtener resultado que permitan identificar tendencias no solo locales, si no globales que se pueden acompañar de inteligencia artificial para poder simular diversos escenarios a futuro y alinear las decisiones a las estrategias del negocio.

N.º	Preguntas	Entrevistado 3 – Analista programador
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	Todo lo que es web Scraping está basada en la extracción que se necesite de los diarios noticias, se pueden usar diversas técnicas de extracción como el uso de herramientas especializadas o usando bots de extracción del DOM, va depender mucho de lo que se quiera lograr, esto se puede hacer con diversas tecnologías que son usadas para poder tener una aplicación a medida, permitiendo procesar los datos de manera general a bajo costo, se puede desarrollar aplicativos o microservicios para la nube donde solo se paga los recursos que se utilizan.
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	Basado en los ejemplos de grandes buscadores como google, ellos indexan las páginas web para obtener mejor resultado a la hora de la búsqueda de información, el uso de bots internamente ayuda a automatizar este proceso, de esta manera se podría crear un bot para ciertas tareas específicas o determinados diarios que cumplan con un patrón adecuado. Eso permitirá tener bien en claro los algoritmos que se va usar, si se va usar para realizar print o para extraer datos mediante técnicas de web Scraping podemos definir diversos algoritmos. Todo ellos basado los diversos modelos de estructura que puede tener un sitio web, por la experiencia se ha usado una librería liberada por Chrome para extraer el DOM puro, obtenido mejores resultados, pero también se puede usar Xquery, o Regex.
3	¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	El tema de mineración es un punto clave para este tipo de procesos, ya que este tipo de negocios maneja grandes volúmenes de datos, acompañando el proyecto de Big data con diferentes orígenes de datos. Finalmente es un proyecto de Big data por el manejo que se tiene que tener la información, luego de su captura. Se tiene que manipular las estructuras de páginas web para la extracción de datos, el uso de bots ayudaría con la automatización de este proceso controlando el web Scraping y reduciendo el uso de diversos recursos, tanto como tiempo y computacionales. Poder aplicar un ETL para todo el proceso completo desde la captura hasta el almacenamiento permitirá tener el monitoreamiento y control por cada etapa que se esté procesando la información,
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	Como programador normalmente se realiza un borrador de lo que se va ser, pero cuando estamos dentro de un proceso de una empresa que ya está en funcionamiento y lograr ser productivo, se tiene que utilizar herramientas tecnológicas que son de usadas por grandes empresas, apoyando en base de datos NoSql y otros servicios, la demanda creciente de datos, es necesario implementar servicios tecnológicos más recientes y escalables. El papel de la tecnología seria clave, porque permite a la empresa seguir evolucionando y mejorando sus diversos proceso internos.
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	El uso de Serverless dentro de web Scraping reducirá el impacto en los recursos que se usan actualmente, ya que este servicio está diseñado solo cobrar por lo que se consume, hay framework y librerías para este tipo de proyectos de microservicios proveedores como AWS y IBM, Golang como otros lenguajes de programación son ideales para este tipo de servicio. El control de inserciones, el uso de nifi como arquitectura de control, acompañado del Kafka, permite el manejo de datos a grandes escalas por el tipo de negocio que es, el manejo de grandes volúmenes de datos al estar soportado por base de datos NoSql, permite que la inserción de datos pueda ser dinámica, usando mongo y Elasticsearch.
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	La información que se recupera viene sucia de alguna manera, por lo cual se tiene que procesar los datos y enriquecerlos para poder usarlos como reportes o métricas personalizados, con esta información las empresas pueden tomar decisiones más acertadas permitiéndole saber si su marca comercial está bien posicionada, como le va en el mercado que se desarrolla, de la misma manera poder observar a su competencia y quizás tomar una decisión respecto a ello. Así mismo permite a la empresa poder predecir diversas situaciones del mercado actual y futuro aplicando Business Analytics.

Anexo 4:

Matriz de codificación de la entrevista

N.º	Preguntas	Entrevistado 1 – Gerente General	Entrevista 1 Codificada
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	En el contexto actual la empresa tiene un trabajo manual para este proceso, el uso de los recursos es bastante alto, esto se puede mejorar con el uso de nuevas tecnologías, apoyado en lenguajes de programación de gran impacto que usen concurrencias como Golang, también herramientas para big data, sistemas de mensajería, aprovechando sobre todo el uso de microservicios en la nube como Serverless de AWS o IBM, ya que es un servicio optimizado e ideal para este proceso de automatización, centralizando el procesamiento y almacenamientos de la información para su posterior análisis que permita tomar decisiones, apoyados con gráficos o métricas que el cliente puede generar a su necesidad para el monitoreamiento de las noticias de sus interés.	En este proceso el uso de recursos es bastante alto, esto se puede mejorar con el uso de nuevas tecnologías, también diversas herramientas como big data, sistemas de mensajería, aprovechando sobre todo el uso de microservicios en la nube como Serverless de AWS o IBM, centralizando el procesamiento y almacenamientos de la información para su posterior análisis y que estos permita tomar decisiones basadas en datos.
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	Se tiene que seguir modelo que se pueda estar usando en la actualidad, como los grandes buscadores que indexan su información a través de bots, con diversas técnicas de web Scraping o framework que están alienados a algunas formas de trabajo. Para poder ser más eficientes, en el proceso de extracción se puede lograr con el uso de algoritmos que estén dirigidos a ciertos grupos de sitios de noticias similares, aprovechando toda esta revolución tecnológica. Los algoritmos personalizados permiten interactuar con diversas web con cierta similitud, a través de ciertas librerías o mecanismos personalizados se puede generar un código por cada sitio web, ya que cada página web tiene una forma de mostrar su información.	Los Grandes buscadores indexan su información a través de bots, ellos usan diversas técnicas de web Scraping o framework personalizados, el proceso de extracción se puede lograr con el uso de algoritmos que estén dirigidos a ciertos grupos de sitios web de noticias con estructura similares. Los algoritmos personalizados permiten interactuar con diversas web con cierta similitud, a través de librerías, mecanismos personalizados o software especializado.
3	¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	La regla de negocio para la captura de la información se podría automatizar a través de la nube, por los gran de volúmenes de información entraría como un proyecto de Big data, el uso de bots en este proceso aceleraría y automatizaría la mineración de información, reduciendo considerablemente las diversas cantidades de errores que se encuentran en este proceso. Así mismo el uso de ETL a lo largo de este proceso permitirá tener un mayor control del procesamiento de los datos en los diversos servicios locales o en la nube, siendo más óptimo subir lo datos ya procesados al final de este proceso para reducir el uso de más recursos.	Por los gran de volúmenes de información es un proyecto de Big data, con el uso de bots en este proceso se aceleraría y automatizaría la mineración de información, reduciendo considerablemente las diversas cantidades de errores. El uso de ETL a lo largo de este proceso permitirá tener un mayor control del procesamiento de los datos en los diversos servicios locales o en la nube, finalmente este proceso reducirá el uso de más recursos.
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	El uso adecuado de la tecnología permitirá poder reducir el nivel de error en gran parte del proceso, ya que controlando los datos por un flujo automatizado y en cada etapa de mineración y enriquecimiento de la información. Así mismo el poder usar micro servicios como parte de la solución, permitiría que este se más robusto y escalable, en el mercado existen diversos tipos de tecnología que permiten centralizar los datos, a través de colas o balanceadores como el Kafka o Rabbit.	La tecnología permitirá poder reducir el nivel de error en gran parte del proceso, ya que controlando el flujo de la automatización en cada etapa de mineración. usar micro servicios como parte de la solución, permitiría que este se más robusto y escalable, permitiendo centralizar los datos, a través de colas o balanceadores como el Kafka o Rabbit.
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	Hay varias tecnologías que han evolucionado con el tiempo permitiendo madurar todo este ambiente, existe un servicio dedicado a la ejecución de funciones llamado Serverless, permite trabajar con concurrencias y en bloques, acompañado del lenguaje de programación Golang que optimiza los recursos donde se ejecuta este proceso, ya que está orientado a este tipo de proceso. El uso de sistemas de mensajería como Kafka para el control de colas por la cantidad de información a procesar también el uso de bases de datos NoSql, que permite ser dinámico al momento de almacenar la información	existe un servicio dedicado a la ejecución de funciones llamado Serverless, permite trabajar con concurrencias y en bloques, Golang optimiza los recursos donde se ejecuta este proceso, ya que está orientado a este tipo de proceso y servicios. El uso de Kafka para el control de colas por la cantidad de información a procesar, el uso de bases de datos NoSql, que permite ser dinámico al momento de almacenar la información
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	Toda esta información extraída con el web Scraping es información en duro, se tendría que pasar por un proceso de análisis, generando reportes gráficos o métricas para ser interpretados de manera más rápida para la toma de decisiones, las empresas con estos datos puedan tener una visión rápida de lo que pueda estar pasando con su marca, si se está moviendo en forma positiva o negativa dentro del mercado, así mismo analizar a su competencia. Con la información obtenida se puede realizar diversas estrategias que permitan anticipar alguna situación oportuna o desfavorable para el negocio.	Se tendría que pasar por un proceso de análisis, generando reportes gráficos o métricas para ser interpretados de manera más rápida para la toma de decisiones sobre su marca, si se está moviendo en forma positiva o negativa dentro del mercado, así mismo analizar a su competencia. permitan anticipar alguna situación oportuna o desfavorable para el negocio.

N ^o	Preguntas	Entrevistado 2 – Jefe del departamento de TI	Entrevista 2 Codificada
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	Se podría automatizar el web Scraping mediante el uso de nuevas tecnologías que actualmente se encuentran en el mercado, estas son usadas para diversos procesos de captura y almacenamiento de grandes volúmenes de datos, cada una de ellas por separado cumplen una función específica, pero en su conjunto puedes resolver grandes problemas de procesamiento y control, como Nifi, Kafak y Elasticsearch. Con la aplicación de estas herramientas la empresa va a lograr tener mejor información, por lo tanto el resultado se verá reflejado en los reportes o métricas que los clientes pueda necesitar y tomar decisiones basadas en datos.	Mediante el uso de nuevas tecnologías que actualmente se encuentran en el mercado, usadas para diversos procesos de captura y almacenamiento de grandes volúmenes de datos, en su conjunto puedes resolver grandes problemas de procesamiento y control, por lo tanto el resultado se verá reflejado en los reportes o métricas que los clientes pueda necesitar y tomar decisiones basadas en datos..
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	El uso de indexación de datos como el Crawling y el Scraping, así como vienen siendo usadas por grandes buscadores web como, por ejemplo, Google, Yandex. Otro punto importante es el uso de técnicas de Scraping que podrían facilitar en obtener la información en mejor tiempo y actualizado, el uso de un software especializado o librerías que permitan interactuar con cada Dom dentro de HTML, en algunos casos los sitios web son estáticos y en otros dinámicos, por este motivo, se puede usar algún algoritmo que ya existe o desarrollar un nuevo algoritmo para obtener los datos a medida, según se requiera dentro de la extracción de noticias.	La indexación de datos como el Crawling y el Scraping, siendo usadas por grandes buscadores web, El uso de técnicas de Scraping facilitan el obtener la información, interactuando con cada Dom dentro de HTML, los sitios web son estáticos y en otros dinámicos, se puede usar algún algoritmo que ya existe o desarrollar un nuevo algoritmo para obtener los datos a medida.
3	¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	Dentro de una empresa consiste en tener diversos procesos automatizados, que permitan construir una arquitectura de Big data o Data lake que se encargaran de soportar el almacenamiento de masivo de datos en bruto, ya que su arquitectura es plana e ideal para este proceso. El uso de bot apoya en la resolución de diversos problemas de tiempo, hacer esto manualmente por el volumen de datos que existe es prácticamente imposible, por eso motivo Google tiene la cantidad de bots que tiene rodando en la internet para poder obtener la información actualizada que es la espina dorsal de su negocio, Acompañado de un ETL o ELT durante todo el proceso, esto facilitara que se pueda ir interactuando en cada punto por donde la información va pasando para ser depurada y enriquecida.	Construir una arquitectura de Big data o Data lake que se encargaran de soportar el almacenamiento de masivo de datos en bruto. El uso de bot apoya en la resolución de diversos problemas de tiempo. hacer esto manualmente por el volumen de datos es prácticamente imposible, Acompañado de un ETL o ELT durante todo el proceso, esto facilitara que se pueda ir interactuando en cada punto por donde la información va pasando para ser depurada y enriquecida.
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	El papel que juega la tecnología dentro de este proceso es demasiado importante, ya que gracias a ello el trabajo manual se verá reducido, sería una carga humanada prácticamente insostenible, como décadas atrás donde el costo sería hiper elevado, lo cual con llevaría a no poder desarrollar negocios como los que actualmente existen, estas tecnologías están compuestas por diversas herramientas que facilitaran el poder recolectar, almacenar y procesar todo tipo de información. El uso de actores como modelo de programación para sincronizar las funciones internas del programa.	La tecnología juega dentro de este proceso demasiada importancia, el trabajo manual se verá reducido, sería una carga humanada prácticamente insostenible, estas tecnologías están compuestas por diversas herramientas que facilitaran el poder recolectar, almacenar y procesar todo tipo de información.
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	El uso constante de la base de datos, así como de los servidores, CPUs, balanceadores y otros, para lo cual se puede usar una alternativa llamada Serverless, que soporta diversos lenguajes de programación como Golang por las concurrencias que maneja, El NoSql es una de las alternativas que se puede usar para este tipo de proceso debido a los grandes volúmenes de información que va depender mucho de la necesidad del proceso, se puede usar mongo, Elasticsearch, o spark. Los sistemas de mensajería como Kafka permiten tener un control e interacción con los datos que van llegando para almacenarse de forma masiva, permitiendo controlar la sobre carga en las validaciones que existen antes de insertar o actualizar la base datos.	Se puede usar una alternativa llamada Serverless, que soporta diversos lenguajes de programación como Golang por las concurrencias que maneja, NoSql es una de las alternativas que se puede usar para este tipo de proceso debido a los grandes volúmenes de información. Los sistemas de mensajería como Kafka permiten tener un control e interacción con los datos permitiendo controlar la sobre carga.
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	El poder tomar decisiones analizando resultados en tiempo real desde varios orígenes de datos, buscando posicionar la marca, son muchas empresas que utilizan este tipo de solución para conocer la salud de su marca comercial y a su vez monitorear a sus competidores. Existen diversos tipos de métricas o reportes que le va permitir a la empresa tomar decisiones a futuro basado en la mineración de datos. El poder aplicar el análisis predictivo en un organización viene acompañado de complementar diversos orígenes de datos, para obtener resultado que permitan identificar tendencias no solo locales, si no globales que se pueden acompañar de inteligencia artificial para poder simular diversos escenarios a futuro y alinear las decisiones a las estrategias del negocio.	El poder tomar decisiones analizando resultados en tiempo real desde varios orígenes de datos, buscando posicionar la marca, a su vez monitorear a sus competidores. Existen diversos tipos de métricas o reportes que le va permitir a la empresa tomar decisiones a futuro basado en la mineración de datos. El análisis predictivo en un organización permite identificar tendencias no solo locales, si no globales para poder simular diversos escenarios a futuro y alinear las decisiones a las estrategias del negocio.

Nº	Preguntas	Entrevistado 3 - Analista programador	Entrevista 3 Codificada
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	Todo lo que es web Scraping está basada en la extracción que se necesite de los diarios noticias, se pueden usar diversas técnicas de extracción como el uso de herramientas especializadas o usando bots de extracción del DOM, va depender mucho de lo que se quiera lograr, esto se puede hacer con diversas tecnologías que son usadas para poder tener una aplicación a medida, permitiendo procesar los datos de manera general a bajo costo, se puede desarrollar aplicativos o microservicios para la nube donde solo se paga los recursos que se utilizan.	Web Scraping está basada en la extracción que se necesite, se pueden usar diversas técnicas de extracción como el uso de herramientas especializadas, esto se puede hacer con diversas tecnologías que son usadas para poder tener una aplicación a medida, se puede desarrollar microservicios pagando los recursos que se utilizan.
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	Basado en los ejemplos de grandes buscadores como google, ellos indexan las páginas web para obtener mejor resultado a la hora de la búsqueda de información, el uso de bots internamente ayuda a automatizar este proceso, de esta manera se podría crear un bot para ciertas tareas específicas o determinados diarios que cumplan con un patrón adecuado. Eso permitirá tener bien en claro los algoritmos que se va usar, si se va usar para realizar print o para extraer datos mediante técnicas de web Scraping podemos definir diversos algoritmos. Todo ellos basado los diversos modelos de estructura que puede tener un sitio web, por la experiencia se ha usado una librería liberada por Chrome para extraer el DOM puro, obtenido mejores resultados, pero también se puede usar Xquery, o Regex.	Grandes buscadores como google, ellos indexan las páginas web para obtener mejor resultado, el uso de bots internamente ayuda a automatizar este proceso en ciertas tareas específicas o determinados diarios que cumplan con un patrón adecuado. Los algoritmos que se va usar para extraer datos mediante técnicas de web Scraping Todo ellos basado los diversos modelos de estructura que puede tener un sitio web, para extraer el DOM puro, obtenido mejores resultados.
3	¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	El tema de mineración es un punto clave para este tipo de procesos, ya que este tipo de negocios maneja grandes volúmenes de datos, acompañando el proyecto de Big data con diferentes orígenes de datos. Finalmente es un proyecto de Big data por el manejo que se tiene que tener la información luego de su captura. Se tiene que manipular las estructuras de páginas web para la extracción de datos, el uso de bots ayudaría con la automatización de este proceso controlando el web Scraping y reduciendo el uso de diversos recursos, tanto como tiempo y computacionales. Poder aplicar un ETL para todo el proceso completo desde la captura hasta el almacenamiento permitirá tener el monitoreamiento y control por cada etapa que se esté procesando la información.	La mineración es un punto clave para este tipo de procesos, ya que este tipo de negocios maneja grandes volúmenes de datos, acompañando el proyecto de Big data, luego de su captura. Se tiene que manipular las estructuras de páginas web para la extracción de datos, el uso de bots ayudaría con la automatización de este proceso de web Scraping y reduciendo el uso de diversos recursos, Poder aplicar un ETL para todo el proceso completo desde la captura hasta el almacenamiento permitirá tener el monitoreamiento y control.
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	Como programador normalmente se realiza un borrador de lo que se va ser, pero cuando estamos dentro de un proceso de una empresa que ya está en funcionamiento y lograr ser productivo, se tiene que utilizar herramientas tecnológicas que son de usadas por grandes empresas, apoyando en base de datos NoSql y otros servicios, la demanda creciente de datos, es necesario implementar servicios tecnológicos más recientes y escalables. El papel de la tecnología sería clave, porque permite a la empresa seguir evolucionando y mejorando sus diversos proceso internos,	Lograr ser productivo, se tiene que utilizar herramientas tecnológicas que son de usadas por grandes empresas, la demanda creciente de datos, es necesario implementar servicios tecnológicos más recientes y escalables. El papel de la tecnología sería clave, porque permite a la empresa seguir evolucionando
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	El uso de Serverless dentro de web Scraping reducirá el impacto en los recursos que se usan actualmente, ya que este servicio está diseñado solo cobrar por lo que se consume, hay framework y librerías para este tipo de proyectos de microservicios proveedores como AWS y IBM, Golang como otros lenguajes de programación son ideales para este tipo de servicio. El control de inserciones, el uso de nifi como arquitectura de control, acompañado del Kafka, permite el manejo de datos a grandes escales por el tipo de negocio que es, el manejo de grandes volúmenes de datos al estar soportado por base de datos NoSql, permite que la inserción de datos pueda ser dinámica usando mongo y Elasticsearch. Montar estas solución a través de contenedores como Docker, permite la gestión de los recursos internamente.	Serverless dentro de web Scraping reducirá el impacto en los recursos que se usan actualmente, ya que este servicio está diseñado solo cobrar por lo que se consume. Golang como otros lenguajes de programación son ideales para este tipo de servicio. Acompañado del Kafka, permite el manejo de datos a grandes escalas. NoSql, permite que la inserción de datos pueda ser dinámica, usando mongo o Elasticsearch.
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	La información que se recupera viene sucia de alguna manera, por lo cual se tiene que procesar los datos y enriquecerlos para poder usarlos como reportes o métricas personalizadas, con esta información las empresas pueden tomar decisiones más acertadas permitiéndole saber si su marca comercial está bien posicionada, como le va en el mercado que se desarrolla, de la misma manera poder observar a su competencia y quizás tomar una decisión respecto a ello. Así mismo permite a la empresa poder predecir diversas situaciones del mercado actual y futuro aplicando Business Analytics.	La información que se recupera viene sucia, por lo cual se tiene que procesar los datos y enriquecerlos para poder usarlos como reportes o métricas personalizadas, permitiéndole saber si su marca comercial está bien posicionada, como le va en el mercado que se desarrolla. Así mismo permite a la empresa poder predecir diversas situaciones del mercado actual y futuro aplicando Business Analytics.

Anexo 5:

Matriz de entrevistados y conclusiones

Nº	Pregunta	E ₁ – Gerente General	E ₂ – Jefe del departamento de TI	E ₃ – Analista programador	Similitud	Diferencias	Conclusión
1	¿Cómo se podría automatizar el proceso de web Scraping de los diarios de noticias?	En este proceso el uso de recursos es bastante alto, esto se puede mejorar con el uso de nuevas tecnologías, también diversas herramientas como big data, sistemas de mensajería, aprovechando sobre todo el uso de microservicios en la nube como Serverless de AWS o IBM, centralizando el procesamiento y almacenamientos de la información para su posterior análisis y que estos permita tomar decisiones basadas en datos.	Mediante el uso de nuevas tecnologías que actualmente se encuentran el mercado, usadas para diversos procesos de captura y almacenamiento de grandes volúmenes de datos, en su conjunto puedes resolver grandes problemas de procesamiento y control, por lo tanto el resultado se verá reflejado en los reportes o métricas que los clientes pueda necesitar y tomar decisiones basadas en datos.	Web Scraping está basada en la extracción que se necesite, se pueden usar diversas técnicas de extracción como el uso de herramientas especializadas, esto se puede hacer con diversas tecnologías que son usadas para poder tener una aplicación a medida, se puede desarrollar microservicios pagando los recursos que se utilizan.	Los tres especialistas concuerdan en aseverar que el uso de tecnologías, servicios de en la nube son necesario para la automatización, dos de los especialistas enfatizaron que el uso de bases de datos para el almacenamiento de grandes volúmenes de dato apoyaran el automatización de web Scraping de los diarios de noticas.	El especialistas E2 enfatizo que el procesamiento final de esta mineración puede acompañar a la toma de decisiones basadas en datos.	Para lograr automatizar el web Scraping de los diarios de noticias es necesario el uso de las tecnologías, herramientas , servicios en la nube y base de datos robustas, que permitan tener resultados para la toma de decisiones.
2	¿Cuál es la estrategia más adecuada para modelar la extracción de datos de un sitio web de noticias?	Los Grandes buscadores indexan su información a través de bots, ellos usan diversas técnicas de web Scraping o framework personalizados, el proceso de extracción se puede lograr con el uso de algoritmos que estén dirigidos a ciertos grupos de sitios web de noticias con estructura similares. Los algoritmos personalizados permiten interactuar con diversas web con cierta similitud, a través de librería o software especializado.	La indexación de datos como el Crawling y el Scraping, siendo usadas por grandes buscadores web, El uso de técnicas de Scraping facilitan el obtener la información, interactuando con cada Dom dentro de HTML, los sitios web son estáticos y en otros dinámicos, se puede usar algún algoritmo que ya existe o desarrollar un nuevo algoritmo para obtener los datos a medida.	Grandes buscadores como google, ellos indexan las páginas web para obtener mejor resultado, el uso de bots internamente ayuda a automatizar este proceso en ciertas tareas específicas o determinados diarios que cumplan con un patrón adecuado. Los algoritmos que se va usar para extraer datos mediante técnicas de web Scraping Todo ellos basado los diversos modelos de estructura que puede tener un sitio web, para extraer el DOM puro, obtenido mejores resultados.	Los tres especiales expresan que el uso de técnicas de web Scraping y algoritmos personalizados, en su conjunto permiten modelar la extracción de información necesaria para este proceso, interactuando con el HTML y Dom en cada sitio web de noticias	El entrevistado E1 menciona la posibilidad de usar software especializados para la modelación de la extracción de noticias de un sitio web.	Para modelar la extracción de datos de un sitio de noticias, es necesario el uso de técnicas de web Scraping, generación algoritmos personalizados o a medida para poder recuperar la información solicitada.
3	¿En qué consiste las reglas del negocio para el proceso de web Scraping de los sitios web de noticias?	Por los gran de volúmenes de información es un proyecto de Big data, con el uso de bots en este proceso se aceleraría y automatizaría la mineración de información, reduciendo considerablemente las diversas cantidades de errores. El uso de ETL a lo largo de este proceso permitirá tener un mayor control	Construir una arquitectura de Big data o Data lake que se encargaran de soportar el almacenamiento de masivo de datos en bruto. El uso de bot apoya en la resolución de diversos problemas de tiempo. hacer esto manualmente por el volumen de datos es prácticamente imposible, Acompañado de un ETL o ELT durante todo el proceso, esto facilitara que se pueda ir interactuando en	La mineración es un punto clave para este tipo de procesos, ya que este tipo de negocios maneja grandes volúmenes de datos, acompañando el proyecto de Big dat, luego de su captura. Se tiene que manipular las estructuras de páginas web para la extracción de datos, el uso de bots ayudaría con la automatización de este proceso	Los tres entrevistados coinciden que las reglas de un negocio debe estar comprendido por el almacenamiento de los grandes volúmenes de datos atreves de big data, así también en la extracción de noticias mediante bots y	Los tres entrevistados no presentan en su respuestas.	Las reglas de negocio para el Scraping web, es necesario considerar el uso de ETL o ELT durante el web Scraping, acompañado de bots y por el volumen de

		del procesamiento de los datos en los diversos servicios locales o en la nube, finalmente este proceso reducirá el uso de más recursos.	cada punto por donde la información va pasando para ser depurada y enriquecida.	de web Scraping y reduciendo el uso de diversos recursos, Poder aplicar un ETL para todo el proceso completo desde la captura hasta el almacenamiento permitirá tener el monitoreamiento y control.	acompañado de un proceso ETL o LTE para el control y monitoreamiento.		almacenamiento debe usarse big data.
4	¿Cuál es la importancia de la tecnología en la automatización de web Scraping?	La tecnología permitirá poder reducir el nivel de error en gran parte del proceso, ya que controlando el flujo de la automatización en cada etapa de mineración. usar micro servicios como parte de la solución, permitiría que este se más robusto y escalable, permitiendo centralizar los datos, a través de colas o balanceadores como el Kafka o Rabbit.	La tecnología juega dentro de este proceso demasiada importancia, el trabajo manual se verá reducido, sería una carga humanada prácticamente insostenible, estas tecnologías están compuestas por diversas herramientas que facilitaran el poder recolectar, almacenar y procesar todo tipo de información.	Lograr ser productivo, se tiene que utilizar herramientas tecnológicas que son de usadas por grandes empresas, la demanda creciente de datos, es necesario implementar servicios tecnológicos más recientes y escalables. El papel de la tecnología sería clave, porque permite a la empresa seguir evolucionando	Para los tres especialistas están de acuerdo en que para minimizar los recursos, el error humano y la carga insostenible de este proceso, es de suma importancia la tecnología en el web Scraping, así también herramientas tecnológicas que soporten el crecimiento robusto.	Los tres especialistas no presentan diferencias en su respuestas.	La importancia de la tecnología en la automatización de web Scraping, permite un mejor uso de los recursos y facilita la carga manual pueda ser automatizada con diversas herramientas tecnológicas.
5	¿Qué componentes de tecnología son más usados en la automatización del web Scraping?	existe un servicio dedicado a la ejecución de funciones llamado Serverless, permite trabajar con concurrencias y en bloques, Golang optimiza los recursos donde se ejecuta este proceso, ya que está orientado a este tipo de proceso y servicios. El uso de Kafka para el control de colas por la cantidad de información a procesar, el uso de bases de datos NoSql, que permite ser dinámico al momento de almacenar la información	Se puede usar una alternativa llamada Serverless, que soporta diversos lenguajes de programación como Golang por las concurrencias que maneja, NoSql es una de las alternativas que se puede usar para este tipo de proceso debido a los grandes volúmenes de información. Los sistemas de mensajería como Kafka permiten tener un control e interacción con los datos permitiendo controlar la sobre carga.	Serverless dentro de web Scraping reducirá el impacto en los recursos que se usan actualmente, ya que este servicio está diseñado solo cobrar por lo que se consume, Golang como otros lenguajes de programación son ideales para este tipo de servicio. Acompañado del Kafka, permite el manejo de datos a grandes escalas. NoSql, permite que la inserción de datos pueda ser dinámica, usando mongo o Elasticsearch.	Los tres entrevistados coinciden en que usar Serverless, un lenguaje de programación que maneje concurrencia, un sistema de mensajería para las colas y bases de datos NoSQL, son los componentes más usados para la automatización del web Scraping.	El entrevistado E2 menciona que el NoSql es una de las alternativas, y va depender der la necesidad del proceso de mineración.	Los componentes de tecnología más usado para la automatización de web Scraping, está compuesto por Serverless, Golang por el uso de concurrencia, sistemas de colar o mensajería y bases de datos NoSql. Los cuales reducirían el uso de recursos y errores.
6	¿En qué consiste la toma de decisiones basada en datos dentro de web Scraping?	Se tendría que pasar por un proceso de análisis, generando reportes gráficos o métricas para ser interpretados de manera más rápida para la toma de decisiones sobre su marca, si se está moviendo en forma positiva o negativa dentro del mercador, así mismo analizar a su competencia. permitan anticipar alguna situación oportuna o desfavorable para el negocio.	El poder tomar decisiones analizando resultados en tiempo real desde varios orígenes de datos, buscando posicionar la marca, a su vez monitorear a sus competidores. Existen diversos tipos de métricas o reportes que le va permitir a la empresa tomar decisiones a futuro basado en la mineración de datos. El análisis predictivo en un organización permite identificar tendencias no solo locales, si no globales para poder simular diversos escenarios a futuro y alinear las decisiones a las estrategias del negocio.	La información que se recupera viene sucia, por lo cual se tiene que procesar los datos y enriquecerlos para poder usarlos como reportes o métricas personalizados, permitiéndole saber si su marca comercial está bien posicionada, como le va en el mercado que se desarrolla. Así mismo permite a la empresa poder predecir diversas situaciones del mercado actual y futuro aplicando Business Analytics.	Los tres especialistas coinciden en que es necesario conocer el posicionamiento de la marca, monitorear a la competencia, atreves de reportes o métricas, con ello obtener un análisis predictivo y usando los datos alienado a sus estrategias, para poder una decisión basada en datos.	El especialista E3 menciona que se podría usar estos datos en herramientas como Business Analytics.	La toma de decisiones basada en datos, se puede complementar con datos extraídos por los procesos anteriores y el uso de herramientas, reportes o gráficos, apoya tomar un decisión en tiempo real.

Conclusión de las entrevistas semi estructuradas

Es trabajo de investigación permitió llegar a la conclusión, que la automatización de web Scraping de los diarios de noticias, es necesario apoyarse en el modelado de datos de noticas, acompañando con procesos de indexación, técnicas diversas de web Scraping que se adapten a las reglas de negocios del negocio que facilita una visión panorámica de cómo se recolecta, almacena y procesa las noticias, así también, contar con la tecnología adecuada para la automatización reduciendo considerablemente los tiempo y errores dentro del flujo. El procesamiento de estos datos podrá sumarse a otros para tomar decisiones basadas en datos. El modelar la extracción de noticias de un sitio web, mediante el uso de técnicas de Scraping para lograr recuperar solo la noticia sin otros valores dentro de esta, apoyándose a través de algoritmos personalizados que optimizan el tiempo y el proceso de extracción. En las reglas de negocio para el Scraping web, es necesario considerar el uso de ETL o ELT durante todo el flujo el web Scraping de noticias, acompañado de bots especializados en estas tareas y por el volumen de almacenamiento se debe preparar una arquitectura con big data.

La tecnología juega un papel de mucha importancia en el web Scraping permite un mejor uso de los recursos y facilita que la carga manual pueda ser automatizada con diversas herramientas tecnológicas. Los componentes tecnológicos de mayor importancia para la automatización de web Scraping, están compuesto por Serverless por el costo que se genera solo de lo que se usa, Golang por el uso de concurrencia, hilos y fue desarrollado con el propósito de usar varios procesadores al mismo tiempo, un sistema de colas o mensajería que equilibre la carga de datos y sobre una base de datos NoSql, esto reducirían el uso de recursos y errores considerablemente. A través de la entrega de datos se puede asociar a otros data Set para poder tomar decisiones basada en datos, se puede complementar con el uso de herramientas como Business Analytic, con reportes o gráficos, apoya tomar un decisión en tiempo real.

Anexo 6:

Guía de Observación

Empresa :	Isuri SAC
Ubicación :	San Martin de Porres
Área :	Departamento de Tecnología de la información
Observador :	Antonio Federico Martínez Núñez
<p>Redacción de lo observado sobre las tres personas que trabajan dentro de la unidad de estudio, donde P1: Encargado del departamento de Tecnología de información, P2: Analista programador y P3: Analista de datos.</p> <p>P1: Para realizar la extracción de las noticias en un sitio web es necesario, se utiliza 03 maneras de hacerlo, se usa un aplicativo interno que se encarga de buscar materias unos cuantos sitio, en algunos casos se hace manual la inserción de estas al sistema, y como una alternativa adicional se puede solicitar a un proveedor especializado este tipo de información, sin embargo también es limitado por que no tiene todas los sitios web de noticias del ámbito local, este proceso demanda mucho tiempo para procesar las noticias aproximadamente 4 horas por el grupo de sitios de noticias, por ser un proceso manual, donde el colaborador tiene que ir revisando si la integración de algunos sitios de noticia se realizó, o se tiene que ir ingresando de forma manual al sistema comercial de la empresa, así mismo el seleccionar información de un servicio de tercero, demanda un tiempo adicional por cada diario de noticias entre 25 – 30 minutos adicionales, se observó que por estas interacciones del personal existen algunos errores de integración, que pueden terminar generando aun mayor pérdida de tiempo, ya que estas noticias se puede usar para diversos clientes. También están propenso a que el servicio pueda detenerse durante el web Scraping, retomar este proceso se debe realizar desde el inicio, lo cual es tiempo y recursos adicionales a todo el proceso. Por otro lado, cuando se recupera la información de cada diario se ha observado que hay ciertas noticas que no se logran almacenar durante el proceso, estas se tienen que ingresar manualmente, posteriormente el proceso de generación de reportes y gráficos para que los clientes puedan dar seguimiento a la información recuperada.</p> <p>P2: El proceso de diversos diarios locales es muy demandante, se tiene que estar monitoreando cada uno de ellos constantemente, el aplicativo de extracción no realiza de forma escala cada sitio web de noticias, para esto busca solo las links que contienen noticias, sin embargo al ser cada sitio web totalmente diferente respecto a la estructura se tiene que monitorear constantemente lo que está extrayendo, el mantenimiento a la lógica no es muy constante pero sin embargo solo se extrae</p>	

ciertos diarios locales, este proceso dura aproximadamente 4 horas esto se integra directamente a una base de datos para su post proceso, donde seleccionan las noticias para clientes específicos. El uso recursos es muy alto para este procesamiento, así también ser observado que el proceso de selección de noticias, el personal tiene que seleccionar noticia por noticias ciertas palabras claves para poder separar estas por clientes específicos, lo cual demanda por cada diario de noticias web un promedio de 10 – 15 por cada sitio, la arquitectura del aplicativo no permite seguir escalando más diarios de noticas sin que se pueden perder algunas de ellas.

P3: Dentro de este proceso las noticias se almacenan en el sistema comercial tienes que ser seleccionadas para uno o varios clientes, este proceso puede demorar entre 10 – 15 por cada cliente, así también hay algunos retrasos en la extracción por la cantidad de pasos que se realizan, en otros casos se tienen que ingresar manualmente algunas noticias relevantes para el informe que se le envía al cliente. Así mismo la cantidad de sitios web de diarios de noticias es limitada con el proceso que se tiene actualmente, se observó que en algunos casos las noticias recuperar se editan para mejorar el resultado, este proceso se realiza diariamente duran las primeras horas de la mañana, esta información procesada complementa una reporte y gráficos para apoyar al cliente y este puede tener una mejor referencia de su marca. En la extracción manual se tiene que ir al sitio web del diario, buscar noticia por noticia y seleccionar las noticias de interés este proceso demanda uno 20 minutos por cada diario a buscar. Otro punto a considerar que mientras se está insertando noticias en el primer proceso, el sistema demora más tiempo en cargar las materias nuevas para poder enriquecerlas y prepararlas.

Anexo 7:

Ficha de Análisis documental

Empresa :	Isuri SAC
Ubicación :	San Martin de Porres
Área :	Departamento de Tecnología de la información
Observador :	Antonio Federico Martínez Núñez
<p>(1) El pago del servicio a tercero por el consumo de noticias tiene un costo aproximado de USD 200 mensuales en los últimos meses en donde se obtienen aproximadamente 4,000 noticas mensuales de diversos diarios, también es importante reconocer que esto puede ser variables según la necesidad propia de cada cliente y los datos que necesite obtener para generar sus análisis. El proveedor solo entrega el texto de las noticias, adicionalmente a esto se tienen que generar por un proceso externo los screenshots de las noticias para dar fe a lo recuperado y se transparente para los usuarios. Así mismo dentro de esto se ha observado que en algunos casos hay ciertas noticias que no están consideradas dentro de las extracciones por tener una mayor dificultad al momento de hacer el web Scraping por parte del proveedor.</p> <p>(2) La extracción de datos sitios web de diarios de noticias se hace mediante el uso de 2 servidores que procesan una cantidad determinas de diarios cada uno de ellos, se encargan de ir al sitio web y recuperar la noticias para posteriormente almacenarlas en una base de Elasticsearch, esto servidores tiene un costo por hora de USD 0.17, sientu un costo diario a de USD 4.08 y un costo mensual de aproximadamente de USD 122.4 por cada servidor siendo un total de USD 244.80.</p> <p>(3) El ingreso Manual de algunas noticias demanda un tiempo al colaborador de 1 min por cada noticia que se ingresa de forma manual, así mismo es un promedio de 80 noticias al día que solo se ocupa en esta parte del proceso y en ciertos casos traba el flujo del proceso hasta que este culminado y procesado.</p>	

Anexo 8: Otras evidencias

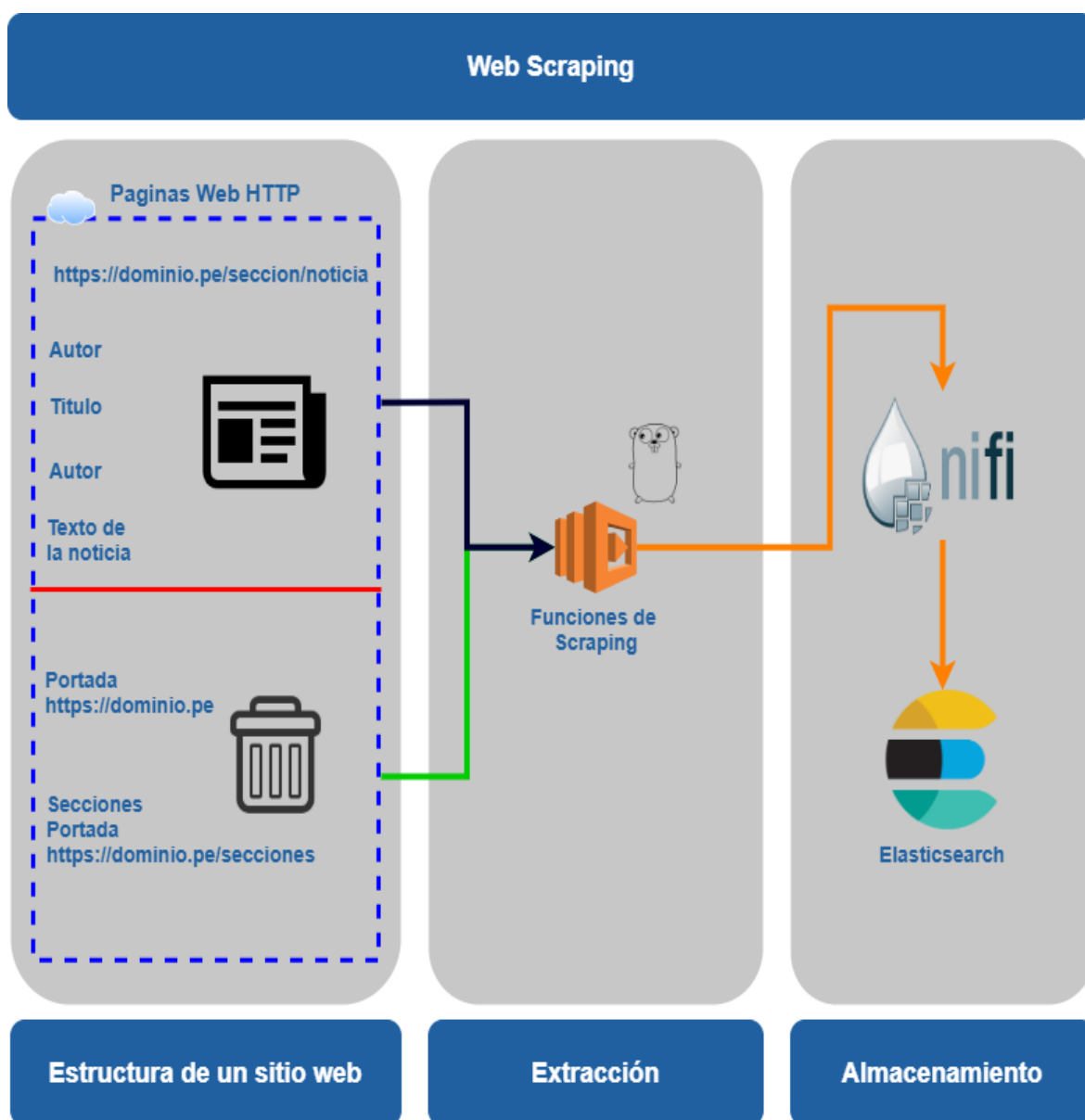


Figura 6. Web Scraping de los diarios de noticias.

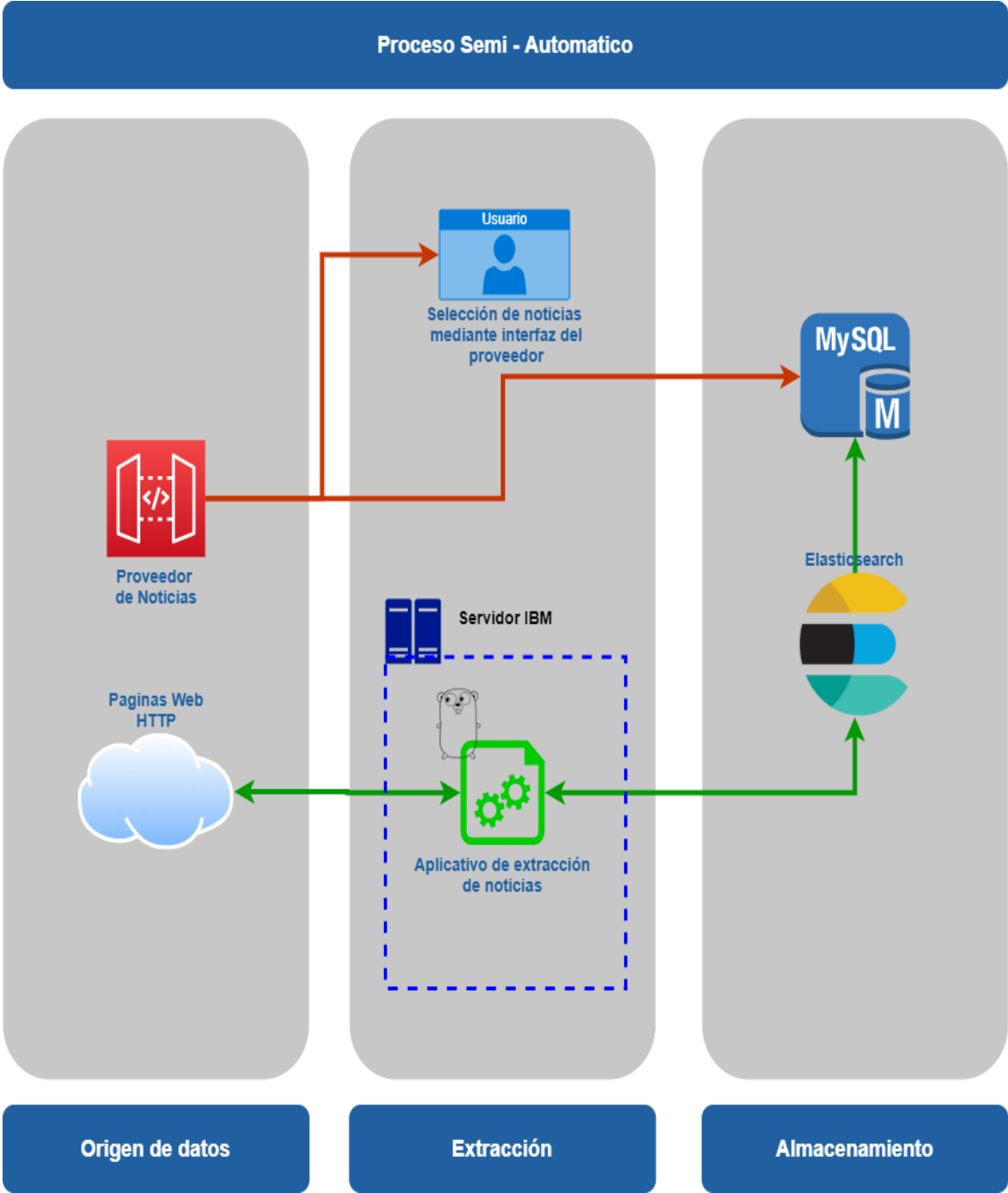


Figura 7. Proceso de extracción de noticias.

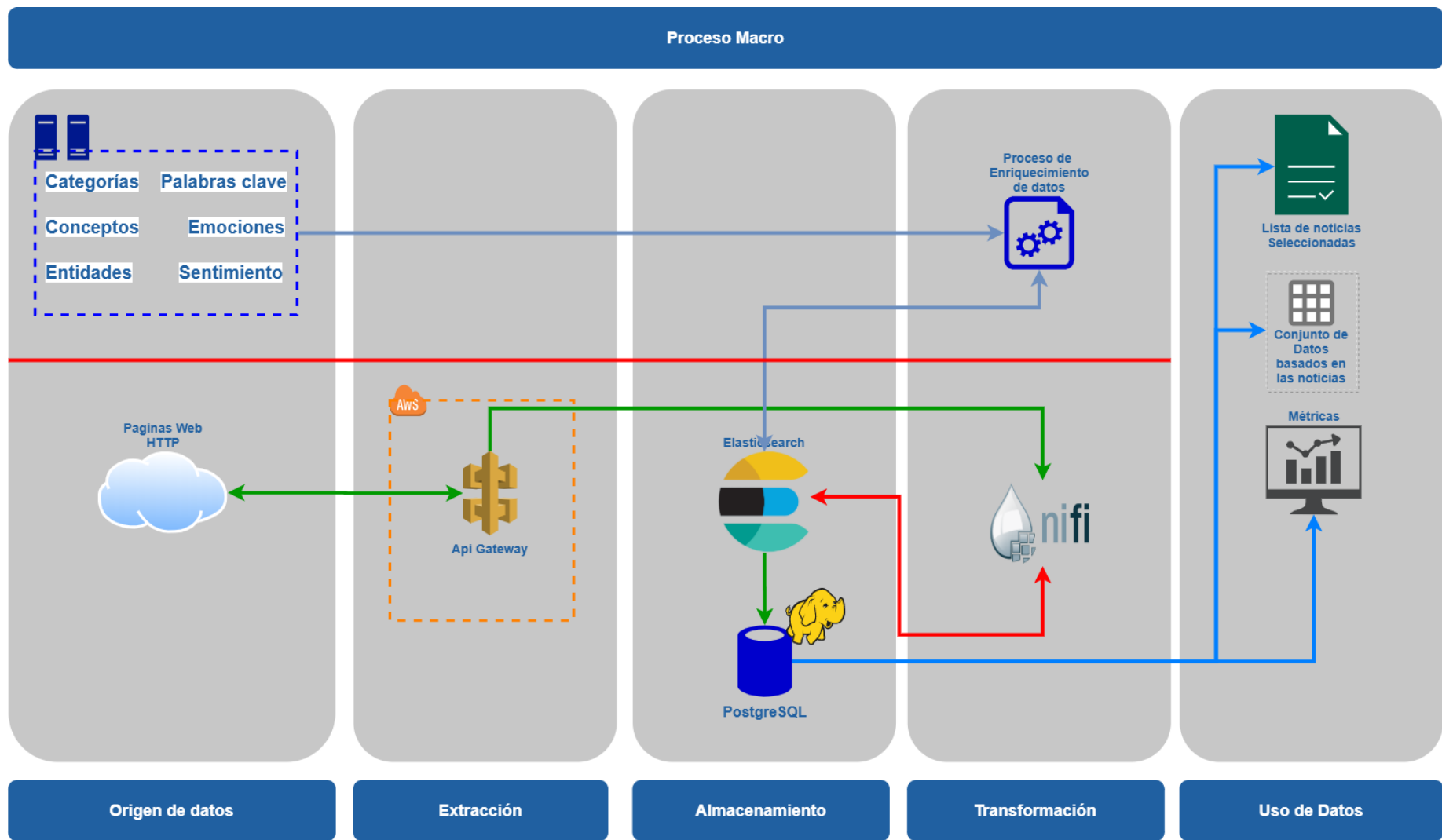


Figura 8. Proceso General de extracción de noticias.

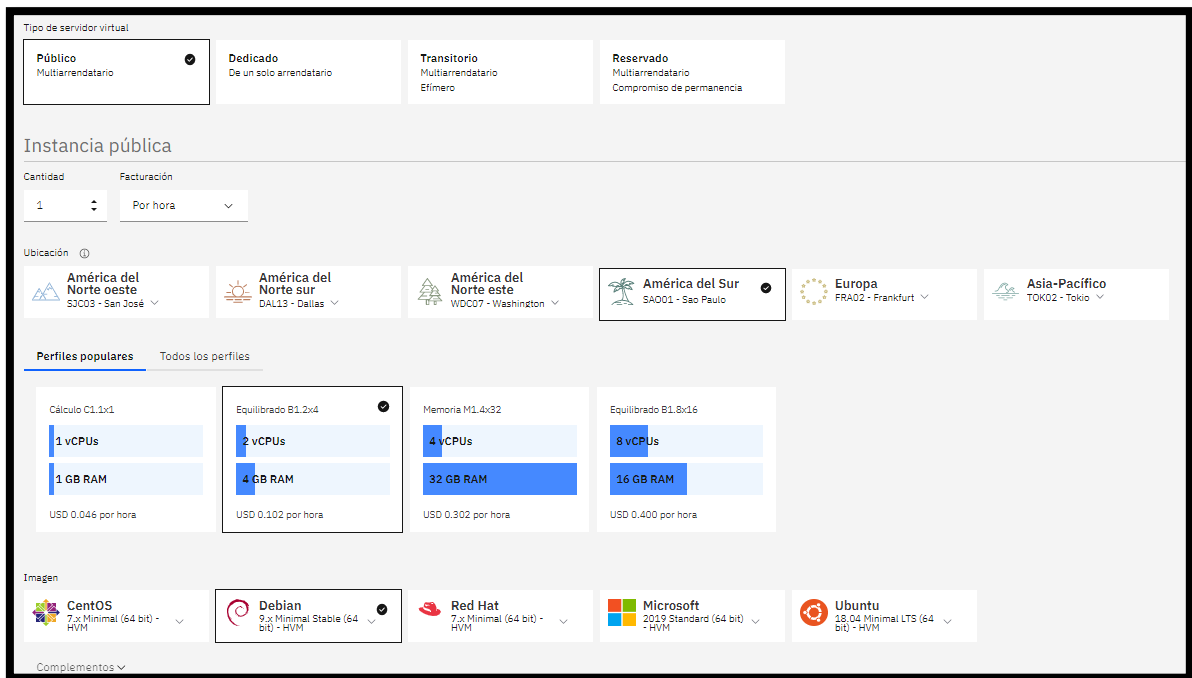


Figura 9. Característica de los servidores usados – SO, Recursos y ubicación. Fuente: IBM.

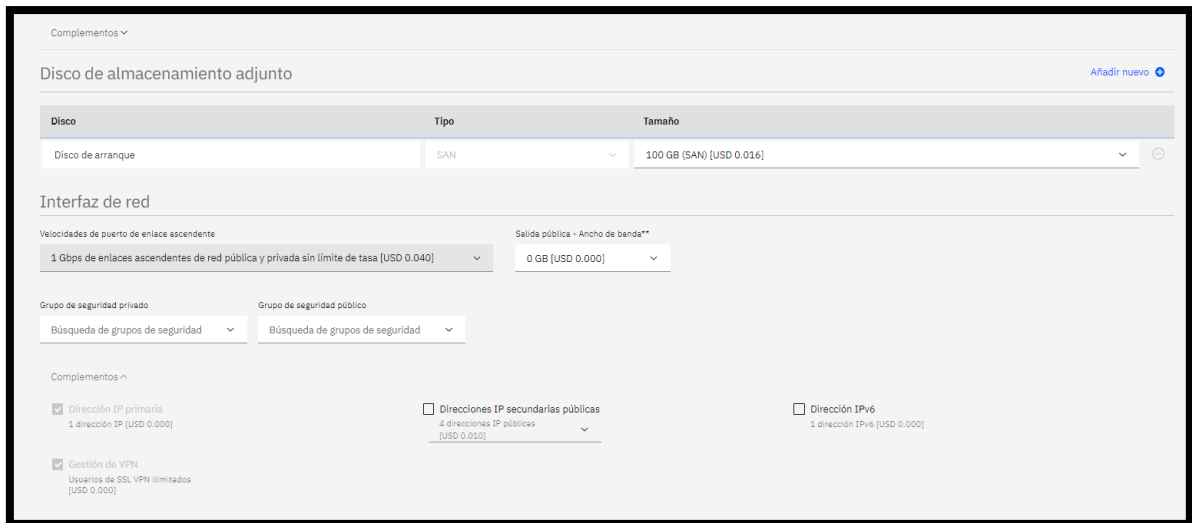


Figura 10. Característica de los servidores usados - RED. Fuente: IBM.

Resumen	
Estados Unidos	USD
1	Instancia de servidor virtual (Público) USD 0.102/hr Equilibrado B1.2x4 2 vCPU 4 GB RAM SAO01 - Sao Paulo Debian GNU/Linux 9.x Stretch/Stable - Minimal Install (64 bit) Complementos
	Disco de arranque - 100 GB USD 0.019
	Interfaz de red USD 0.048 1 Gbps de enlaces ascendentes de red pública y privada sin límite de tasa Complementos
Aplicar código promocional	
Vencimiento total por hora* USD 0.17 <i>estimado</i>	

Figura 11. Valor por estimado por hora de cada servidor. Fuente: IBM.

The screenshot displays a webpage for 'liveON SOLUTIONS'. On the left, there is a contact section with the following details:

- Empresa: Live On Solutions
- País: BR
- Estado: SP
- Dirección: Rua Treze de Maio, 675.
- liveonsolutions.com

 A red button labeled 'CONTACTO' is positioned below the contact information.

On the right, under the heading 'Tipos de implementación', there is a list of services with their respective price ranges in BRL:

- Implementación de layout**: BRL 1,000.00 - 50,000.00. Description: Creación e implementación de diseño, nuevas características. Optimización SEO.
- B2B**: BRL 50,000.00 - 300,000.00. Description: Desarrollo e implementación de tiendas B2B. Excelente para distribuidores, proveedores, empresas de reventa.
- Tienda internacional**: BRL 50,000.00 - 500,000.00. Description: Desarrollo de tiendas internacionales en colaboración con Vtex USA. Integraciones de mercado multicanal en venta con Amazon, Ebay, Walmart. B2C y B2B.
- Mercado**: (No price range shown)

Figura 12. Costo similar al proveedor de Servicio de noticias Stor vtex.

Etiquetas de fila	Promedio de count(*)
Abr	78.46666667
May	85.58064516
Jun	75.16666667
Total general	79.8021978

Figura 13. Promedio de noticias ingresadas manualmente.

CARTA DE SOLICITUD PARA USO DE INFORMACIÓN DE LA EMPRESA ISURI



San Martín de Porres, 03 de mayo de 2020

Señores

Universidad Cesar Vallejo - UCV

Presente. -

De nuestra consideración:

Por medio de la presente, tenemos el agrado de dirigirnos a ustedes, a fin de informarles sobre la solicitud para el uso de información de mi representada requerida por vuestro alumno de posgrado Br. Antonio Federico Martínez Núñez identificado con DNI: 45676320, para el desarrollo de su Tesis titulada “Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres”.

Al respecto, de manera expresa autorizamos que la información recogida en la presente investigación pase a ser de carácter pública dentro de los fines académicos que son propios de la naturaleza de este tipo de trabajos, entre los cuales está su publicación, una vez concluido el mismo, en el repositorio de la Universidad.

Sin otro particular, nos despedimos de Ustedes, expresándole las muestras de nuestra mayor consideración.

Atentamente,

Arturo A. Valenzuela Pizarro

Jr. Camaná 629
San Martín de Porres - Lima
Tel: (051) 9-949-40495



GUÍA PARA LA AUTOMATIZACIÓN DE WEB SCRAPING DE LOS DIARIOS DE NOTICIA

Preparado por:
Antonio F. Martínez Núñez
antonio0318@gmail.com

Anexo 9: Propuesta de automatización de web Scraping

I. Diseño de Automatización

1.1. Diseño Global

Está comprendida por el conjunto de técnicas, algoritmos, bots, tecnología, enriquecimiento de los datos, todo esto soportado dentro de dos ámbitos principales Nifi y Serverless AWS (Figura 1).

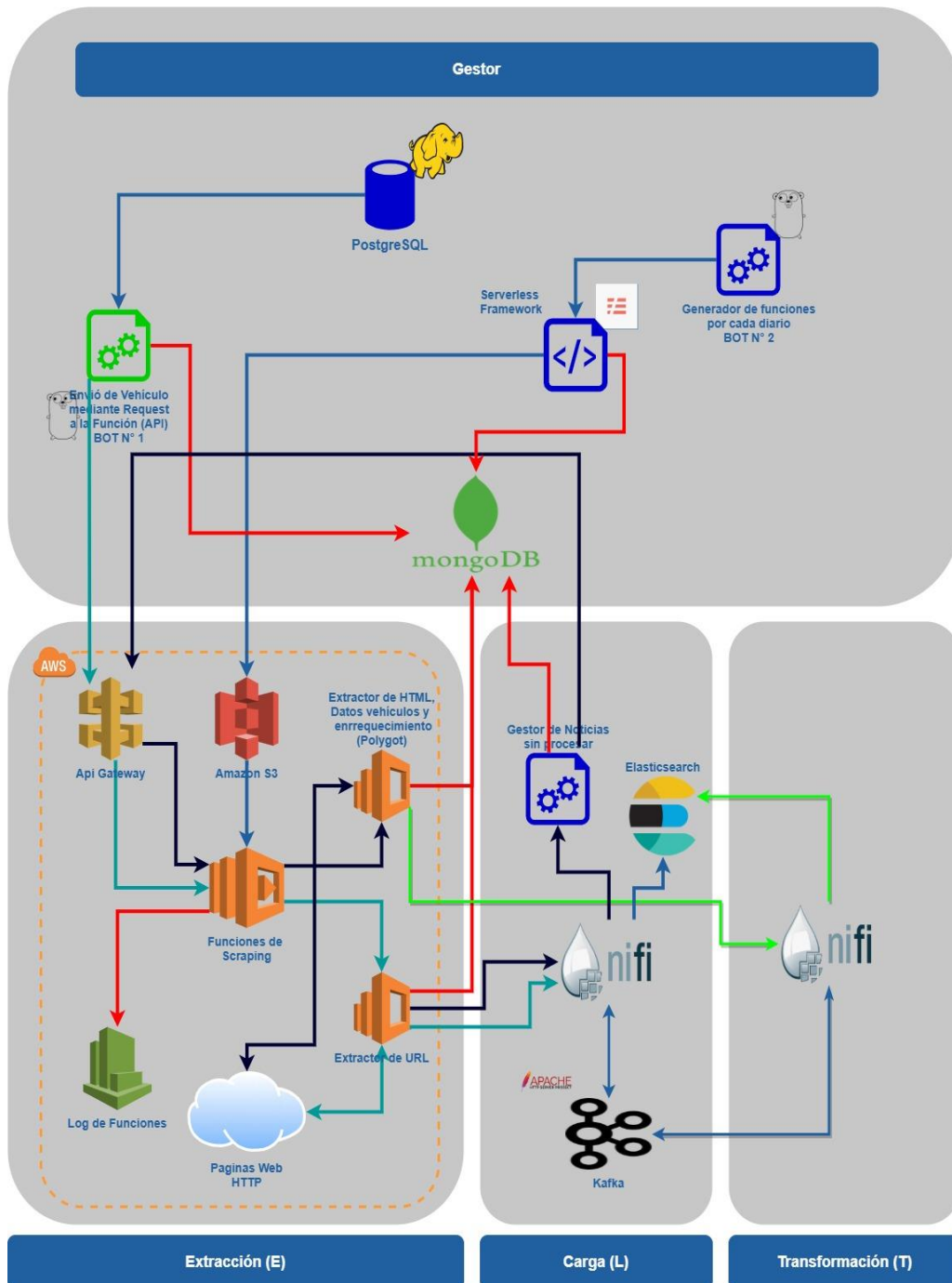


Figura 1. Modelo General de automatización.

1.2. Diseño Bot N° 1 - Web Scraping URL

Comprende el primer paso para la ejecución del web Scraping en un determinado horario pre establecido (Figura 2).

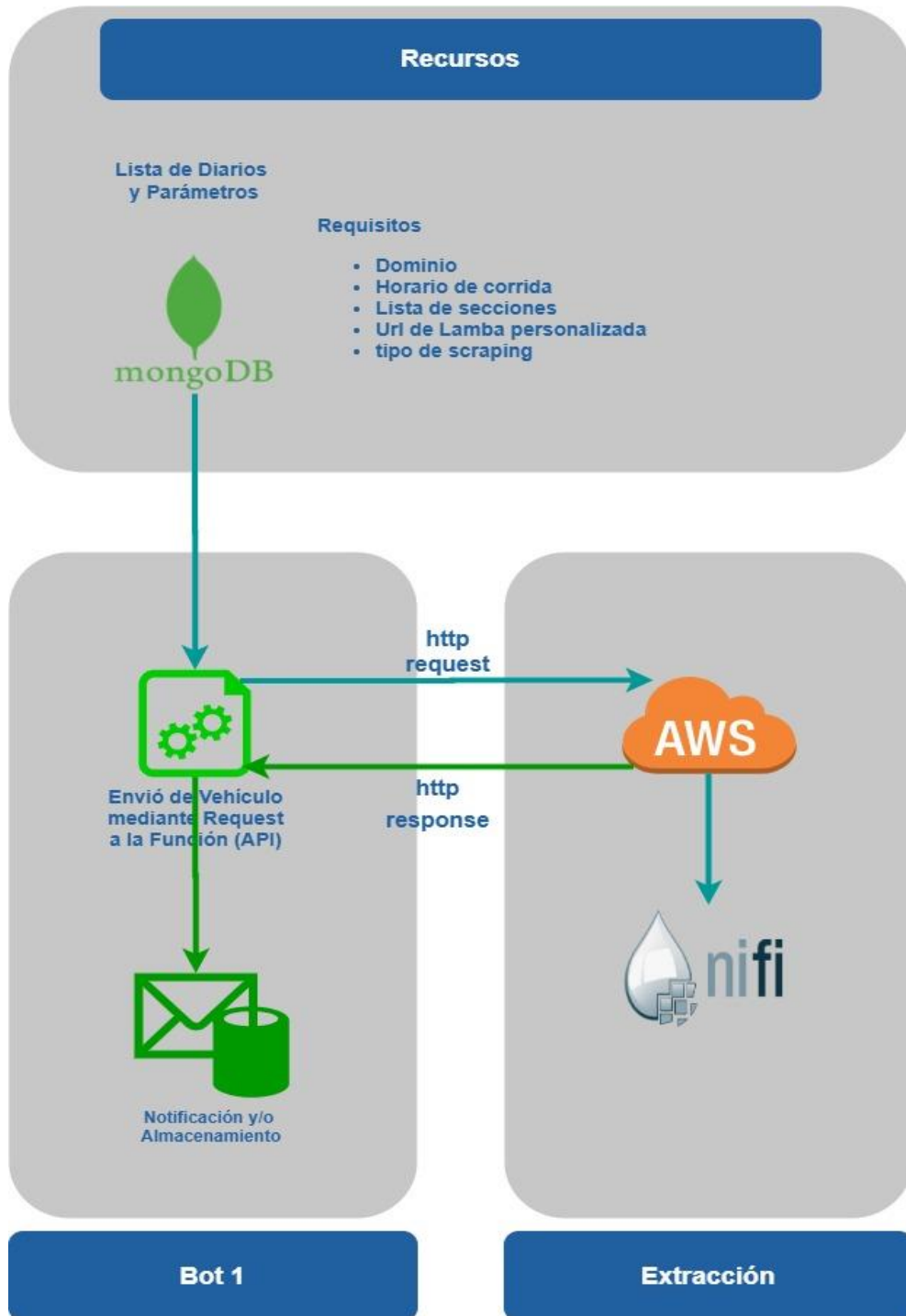


Figura 2. Modelo Bot N° 1 – Web Scraping URL.

1.3. Diseño Bot N° 2 - Creación de funciones

Dentro de este proceso permite generar automáticamente las nuevas funciones para los diarios que han sido agregados recientemente, este proceso revisa cada 20 min, si hay algún nuevo registro agregado (Figura 3).

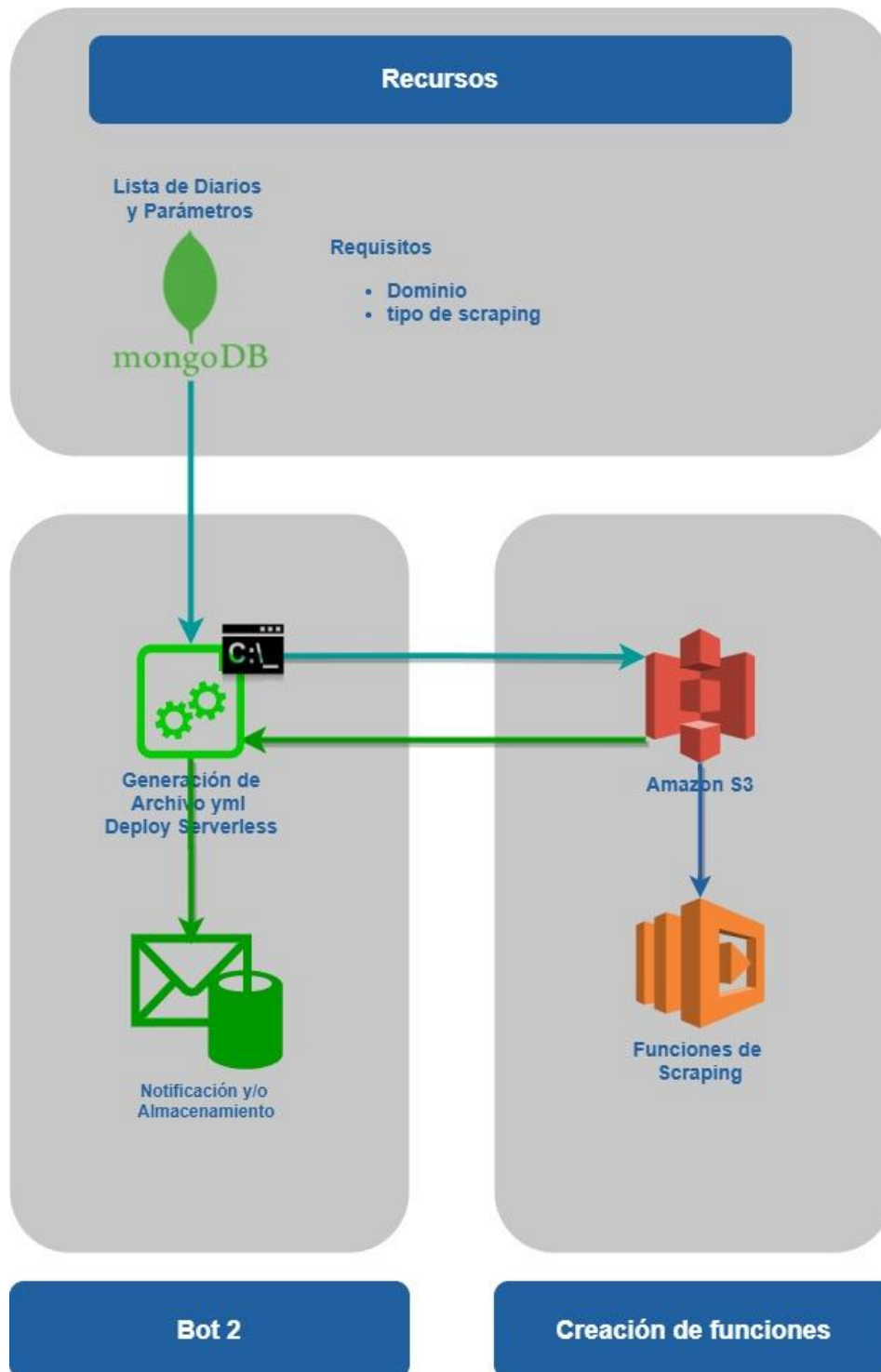


Figura 3. Modelo Bot N° 2 – Creación de funciones.

II. Aplicaciones

Para lograr la funcionalidad optima de la automatización de web Scraping es necesario instalar una serie aplicaciones, cada uno de ellas se debe configurar adecuadamente para que en su conjunto logren funcionar automáticamente.

Los aplicativos y software necesario son:

2.1. Nifi Apache

Instalación en Linux, post original de www.bidataguide.com (Figura 4, 5 y 6):

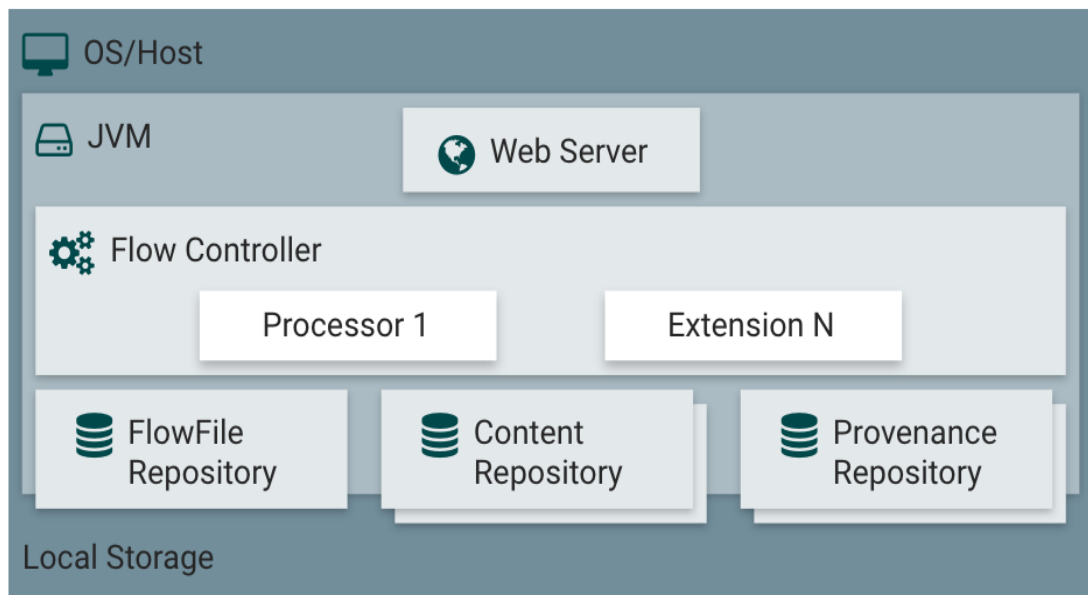


Figura 4. NiFi Architecture. Recuperado de <https://nifi.apache.org/docs.html>

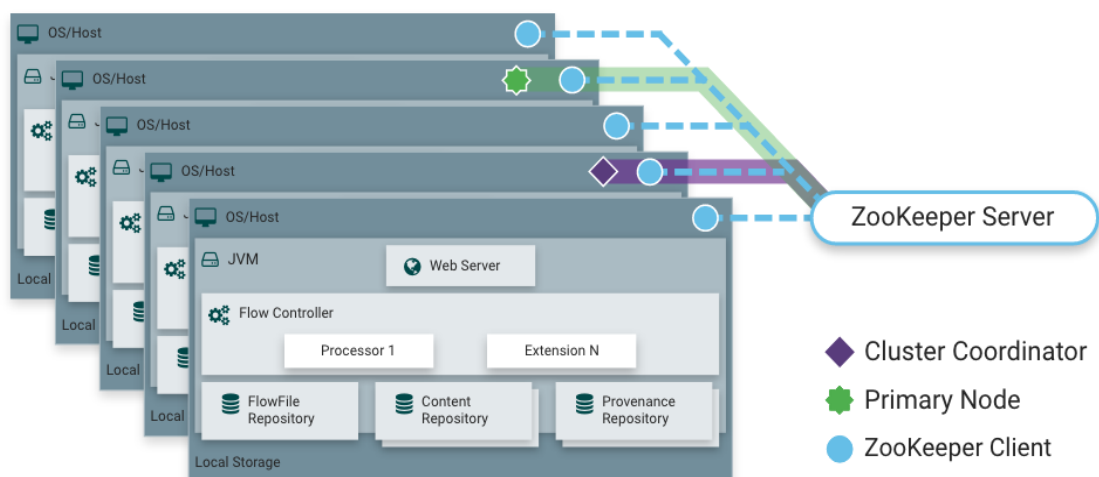


Figura 5. Zookeeper Server. Recuperado de <https://nifi.apache.org/docs.html>

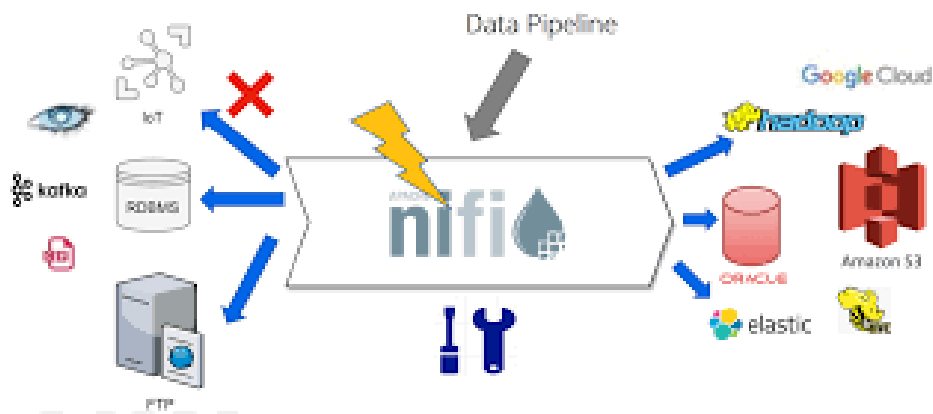


Figura 6. Nifi Performance. Recuperado de <https://www.informatik-aktuell.de>

Guías para implementación sugerida:

- <https://techexpert.tips/es/apache-nifi-es/apache-nifi-instalacion-en-ubuntu-linux/>
- <https://www.theninjacto.xyz/Pull-data-Twitter-push-to-Elasticsearch/>
- <https://www.nifi.rocks/using-the-executescript-processor/>

2.2. Kafka Apache

Este servicio nació a partir de una necesidad que tuvo una red social profesional altamente conocida como LinkedIn, durante su crecimiento el equipo técnico inició con un sistema de recopilación de métricas del software utilizando componentes internos personalizados con soporte en herramientas de código abierto. Sin embargo, esta solución no duro mucho tiempo por varios problemas al momento de querer usar los mensajes en diversas aplicaciones. Todos los mensajes se almacenan como registro en sistemas de archivos persistentes. Tiene un sistema de registro anticipado que permite escribir todos los mensajes publicados antes de ponerlo a disposición de aplicaciones de consumo. Cada mensaje en Kafka es una colección de bytes, esta colección se representa como un matriz, estos se van almacenando mensajes en la secuencia que van llegando(Figura 7).

El número de particiones se configura al momento de la creación de "Topic ", físicamente, cada topic se extiende sobre diferentes corredores de Kafka, que albergan uno o más particiones. Un típico Kafka clúster consta de múltiples corredores, los cuales ayuda en las lecturas y escrituras de mensajes de equilibrio de carga en el clúster controlado por el líderes un servicio de coordinación que permite aplicaciones distribuidas la fácil implementación y coordinación. Tales mensajes se pueden nombrar, gestionar la configuración, elegir líderes,

pertenecer a grupos, barreras y bloqueos distribuidos. Para mantener sus estados usan Zookeeper, cada topic tiene un líder acompañado de cero o más tuberías como seguidores. Los líderes gestionan cualquier solicitud de lectura o escritura para sus respectivas particiones, dicho seguidores replican al líder en segundo plano sin interferir activamente, los servidores son réplicas entre sí y cada uno mantiene una copia del estado de la solicitud. Los clientes de Zookeeper pueden enviar sus solicitudes a cualquiera de los N servidores. Las solicitudes pueden ser ampliamente categorizado como leer o escribir (Figura 8 y 9).

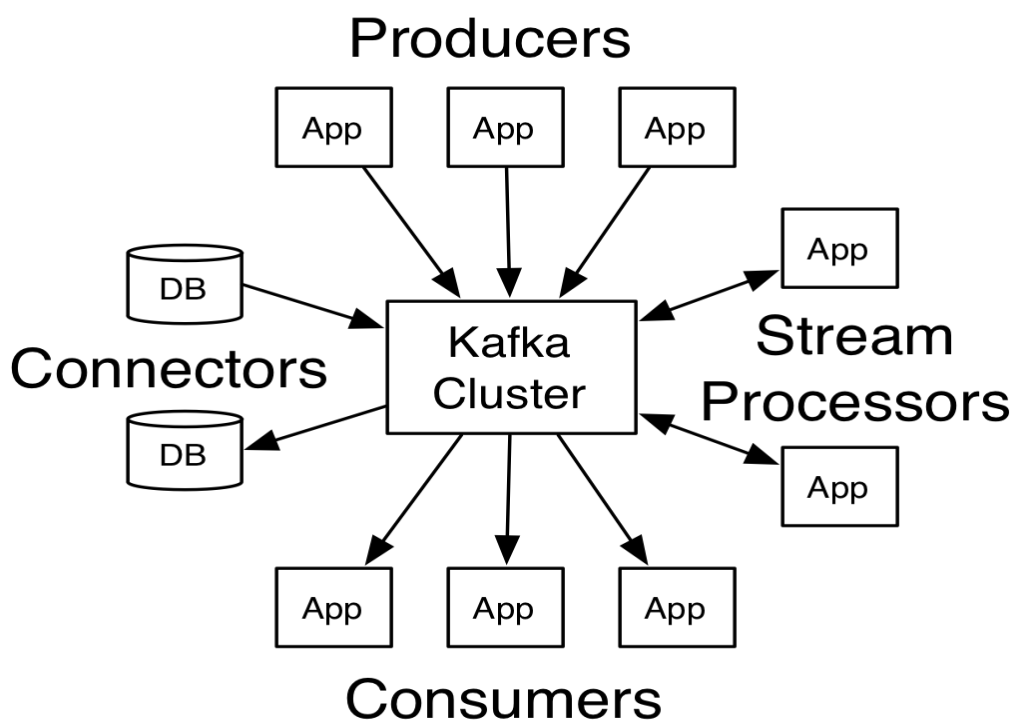


Figura 7. Plataforma de transmisión distribuida de <https://kafka.apache.org/>

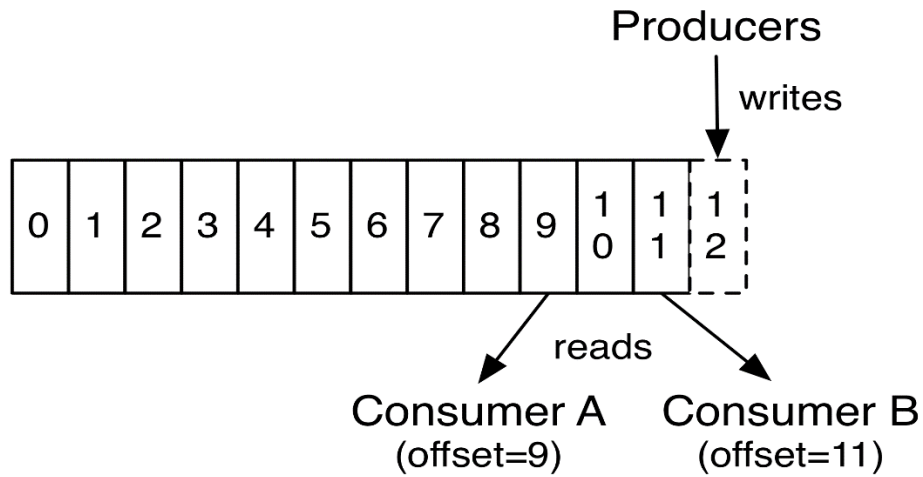


Figura 8. Kafka Productor, recuperado en <https://kafka.apache.org/>

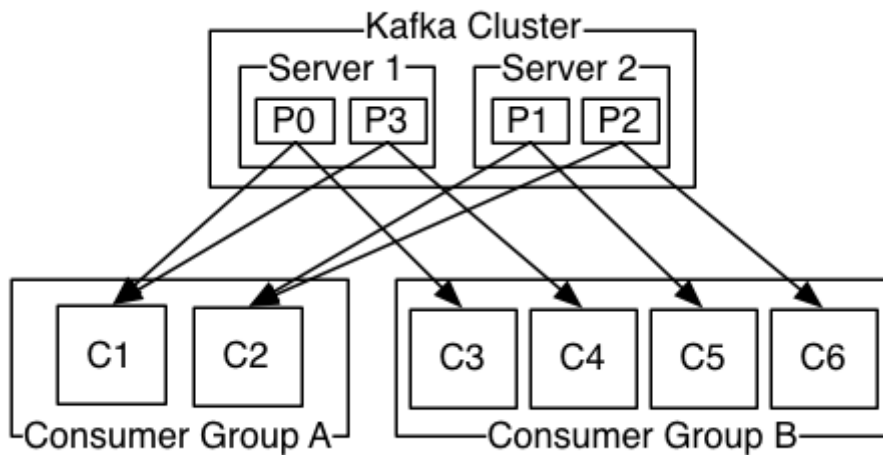


Figura 9. Kafka Consumidores, recuperado en <https://kafka.apache.org/>

Guías para implementación sugerida:

- <https://kafka.apache.org/quickstart>
- <https://www.digitalocean.com/community/tutorials/how-to-install-apache-kafka-on-debian-9>
- <https://stackoverflow.com/questions/34512287/how-to-automatically-start-kafka-upon-system-startup-in-ubuntu>

2.3. Elasticsearch

Es un motor analítico y de búsqueda escrita en Java, la cual tiene una base de SOIR, su primer lanzamiento en el 2010. ha sido ampliamente adoptado por la NASA, Wikipedia y GitHub, para diferentes casos de uso. Elasticsearch proporciona una API HTTP/JSON, que tiene estas principales características: 1) Distribuido: Puede comenzar con un clúster Elasticsearch de un solo nodo y puede escalar ese clúster a cientos o miles de nodos 2) Alta disponibilidad: replicación de datos significa tener múltiples copias de datos en su clúster, 3) Basado en REST: Elasticsearch está basado en la arquitectura REST y proporciona puntos finales API para no solo realizar operaciones CRUD a través de llamadas API HTTP, 4) Potente DSL de consulta: El DSL de consulta (lenguaje específico del dominio) es una interfaz JSON proporcionada por Elasticsearch para exponer el poder de Lucene para escribir y leer consultas de una manera muy fácil y 5) Sin esquema: Ser sin esquema significa que no tiene que crear un esquema con nombres de campo y tipos de datos antes de indexar los datos en Elasticsearch (Figura 10).

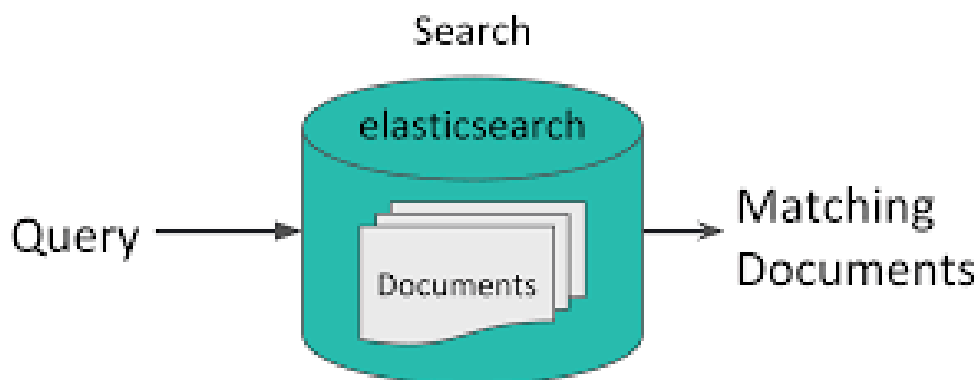


Figura 10. Estructura de Elasticsearch, recuperado en <https://blog.athento.com/>.

Guías para implementación sugerida:

- <https://www.rosehosting.com/blog/how-to-install-the-elk-stack-on-debian-9/>
- https://www.busindre.com/configuracion_de_rendimiento_de_elasticsearch

2.4. Mongo

El NoSql hacer referencia a cualquier almacén de datos que no sigue el modelo de RDBMS (relational database management system, traducido “sistema de gestión de bases de datos relacionales”) tradicional, específicamente. Los enfoques no son relacionales y no utilizan SQL como lenguaje de consulta. Este tipo de base de datos intentan resolver los problemas de escalabilidad y disponibilidad frente a los de atomicidad o consistencia. NoSQL no es una base de datos, ni siquiera es un tipo de base de datos, solo un término para identificar un conjunto de base de datos fuera del ecosistema. Namdeo y Suman (2020). Sistema de gestión de bases de datos relacionales. Un RDBMS tradicional tiene un conjunto de características como: 1) Atomicidad: todo en una transacción tiene éxito para que no se revierta, 2) Consistencia: una transacción no puede dejar la base de datos en un estado inconsistente, 3) Aislamiento: una transacción no puede interferir con otra y 4) Durabilidad: una transacción completa persiste, incluso después de reiniciar las aplicaciones. Por muy indispensables que puedan parecer estas cualidades, son bastante incompatibles con la disponibilidad y el rendimiento en aplicaciones de escala web (Figura 11).



Figura 11. Conexiones a MongoDB, recuperado en <https://bit.ly/2OFoMVL>.

Guías para implementación sugerida:

- <https://docs.mongodb.com/manual/tutorial/install-mongodb-on-windows/>
- <https://desarrolloactivo.com/articulos/mongodb-auth/>
- <https://www.howtoforge.com/tutorial/install-mongodb-on-ubuntu/>
- <https://chachocool.com/como-instalar-mongodb-en-debian-9-stretch>

2.5. Serverless Framework

Serverless, la computación sin servidor es un modelo de ejecución que se hace cada vez más popular en la nube. Con servicios como AWS Lambda, los usuarios escriben aplicaciones como colecciones de funciones sin estado que implementan directamente en un marco sin servidor. Los 4 principales modelos son: 1) IaaS (Infrastructure as a Service, traducción Infraestructura como servicio). 2) PaaS (Platform as a Service, traducción Plataforma como servicio). 3) CaaS (Container as a Service, traducción Contenedor como servicio), 4) FaaS (Function as a Service, Traducción “Funciona como un servicio”). La arquitectura sin servidor en AWS, están diseñadas principalmente para procesar tareas en segundo plano de la Web e Internet de las cosas aplicaciones o procesamiento de flujo controlado por eventos, la nube fue iniciada por AWS Lambda en el 2014, este modelo en la nube tiene diversos beneficios 1) NoOps: donde la responsabilidad por aprovisionar, mantener y parchar los servidores se transfiere de cliente a proveedores. Y los desarrollares se enfocan en optimizar, innovar y mejorar sus funciones, solo se paga el tiempo que la función demore en ejecutar. 2) Autoescalado y alta disponibilidad: proveedor de servicios para decidir cómo usar su infraestructura de manera efectiva para atender las solicitudes de los clientes y escalar horizontalmente las funciones en función de la carga. 3) Optimización de costes: sólo se paga por el calcule el tiempo y los recursos (RAM, CPU, red o tiempo de invocación) que consume. No paga por los recursos inactivos. 4) Polygot: se puede usar diversos lenguajes de programación una parte de la aplicación se puede escribir en Java, otra en Go, otra en Python; en realidad no importa mientras haga el trabajo (Figura 12).

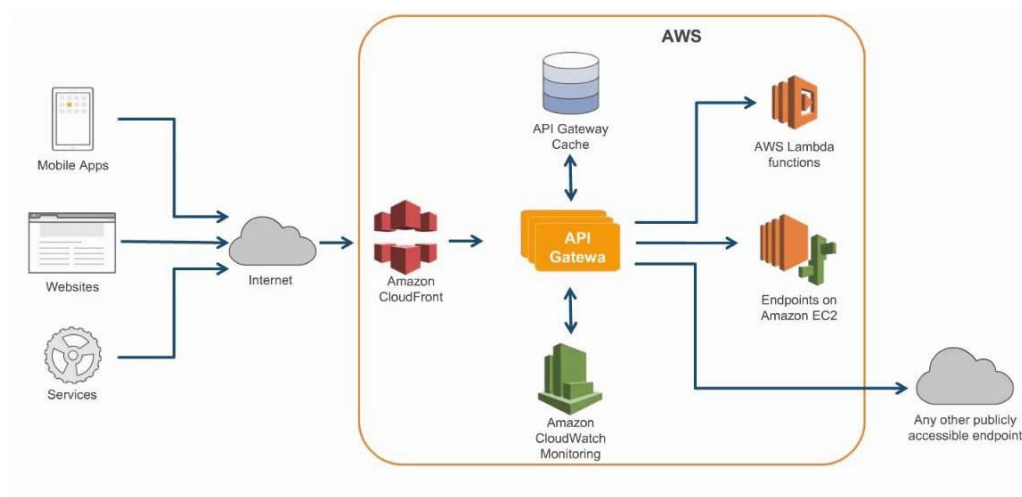


Figura 12. AWS Lambda, arquitectura Serverless para implementar APIs, recuperado de <https://bit.ly/2WwTMvw>.

Guías para implementación sugerida:

- <https://www.serverless.com/framework/docs/getting-started/>
- <https://enmilocalfunciona.io/aprendiendo-serverless-framework-parte-2-instalacion/>
- <https://www.paradigmadigital.com/dev/aws-lambda-arquitectura-serverless-implementar-apis/>
- <https://medium.com/@daniel.woods/deploying-a-golang-package-to-aws-lambda-in-5-minutes-cd11685f576>
- <https://www.serverless.com/blog/framework-example-golang-lambda-support>
- <https://dev.to/wingkwong/building-serverless-crud-services-in-go-with-dynamodb-part-1-2kec>
- <https://medium.com/@ranrib/how-to-make-lambda-faster-memory-performance-benchmark-be6ebc41f0fc>

III. Web Scraping

3.1. Técnicas de web Scraping

Las diversas técnicas de Scraping, 1) El tradicional copiar y pegar, 2) Grapado de texto y expresión regular comando UNIX (Sistema operativo portable). 3) Programación del protocolo de transferencia de hipertexto (HTTP). 4) Análisis del lenguaje de marcado de hipertexto, se usa lenguaje de consulta semiestructurado, como XQuery y consulta por hipertexto(HTQL). 5) Análisis del modelo de objetos de documento (DOM). 6) Software de Web Scraping, existen muchas herramientas disponibles para usar, 7) Plataformas de agregación vertical, existen diversas compañías con plataformas de cosechas específicas, las cuales crean y supervisan una multitud de bots, 8) Analizadores de páginas web de visión por computadora (Figura 13).



Figura 13. Web crawling, recuperado de <https://bit.ly/3fHdufJ>.

Guías para implementación sugerida:

- <http://go-colly.org/>

- <https://benjamincongdon.me/blog/2018/03/01/Scraping-the-Web-in-Golang-with-Colly-and-Goquery/>
- <https://github.com/gocolly/colly>

3.2. Algoritmos

Para poder extraer y estructurar los datos de las páginas Web se utiliza la técnica llamada como Web Scraping. Scraping significa “raspado” se refiere a la extracción, limpieza y filtro de los datos Algoritmos para Scraping, muchos procesamientos de lenguaje natural requieren dividir gran cantidad de texto en oraciones, aun para mucho de nosotros suele ser una tarea simple, para los ordenadores es un problema mayor. Es necesario utilizar otros algoritmos para poder clasificar el texto de mejor manera. Existe diferentes tipos de algoritmos para web Scraping que se aproximan a la mayoría de estructuras web, sin embargo, esto se puede mejorar con métodos que le permitan delimitar de mejor manera el resultado deseado.

3.3. Golang

Con la finalidad de monopolizar los CPU's, se usará un enfoque que funcione a escalas de microsegundos, aprovechando las multitareas que nos proporciona Go Lang a través de las Gorutinas (hilos ligeros). El anuncio de AWS su apoyo al lenguaje de programación GO a partir del inicio del 2018, donde existe algunas framework para integrar con Lambda, una de las razones de uno usar servidores es poder usar Poligot, independientemente del lenguaje. Es donde Go entra en el juego, teniendo como ventajas lo siguiente: 1) Orientado a la nube: Lenguaje diseñado por Google, considerando la escalabilidad y reducción en el tiempo de compilación. 2) Rápido: Go tiene una sintaxis limpia y especificaciones de lenguaje claras. Esto ofrece un lenguaje fácil para que los desarrolladores aprendan y muestra buenos resultados rápidamente mientras produce código de fácil mantenimiento. 3) Escalable: o tiene una concurrencia integrada con goroutines en lugar de hilos. 4) Eficiente: la velocidad de compilación más rápida permite una retroalimentación igual de rápida; esto sin duda es la ventaja más importante para alguien con un presupuesto ajustado. Además, Go está respaldado por Google, tiene un ecosistema grande y un gran soporte IDE (IntelliJ, VSCode, Atom, GoGland) y depuración.

Golang se usa para aprovisionar para admitir la programación del sistema en el contenedor. Los contenedores son livianos y portátiles, ya que se trata de una encapsulación de un entorno en el que se ejecutará la aplicación

Guías para implementación sugerida:

- <https://golang.org/>
- <https://code.tutsplus.com/es/tutorials/3-things-that-make-go-different--cms-28864>
- <https://nathanleclaire.com/blog/2014/12/29/shelled-out-commands-in-golang/>
- <https://medium.com/@nikolay.bystritskiy/how-i-tried-to-do-things-asynchronously-in-golang-40e0c1a06a66>
- <https://fabianlee.org/2017/05/21/golang-running-a-go-binary-as-a-systemd-service-on-ubuntu-16-04/>
- <https://www.simplified.guide/ubuntu/install-vim>

IV. Implementación

4.1. Desarrollo

4.1.1. Robots

El desarrollo de los robots y las funciones dentro de la infraestructura están basadas en Golang, dentro de la automatización existen 2 robots:

- Robot N°1: Este Robot se encarga de la extracción de las url de un diario en específico, el cual tiene como función principal recorrer todo el sitio web en busca de todas las secciones que estén dentro de ella, luego de recuperar cada una de las secciones envía el resultado a Kafka que se encargara de la carga de los mensajes.
- Robot N°2: Este Robot se encarga de generar todas las funciones necesarias para el proceso que se encuentren en Serverless de AWS, generando los archivos necesarios que serán cargados para que se complete el flujo del proceso, logrando recuperar en este proceso las URL de las funciones generadas al momento de subirlas y actualizando esta información dentro de los registros internos de control.

4.1.2. Funciones

Las funciones fueran diseñadas para lograr realizar Scraping web, con el uso de librerías como el Go Colly, que trae un sitio Web como texto y lo parsea por cada nodo HTML, ventaja de poder usar esta librería en Golang, es el uso de Hilos dentro de la arquitectura permitiendo procesar varias noticias de forma paralela, se utiliza las técnicas de programación del protocolo de transferencia de hipertexto y análisis del modelo de objetos de documentos, dentro de estas funciones se programa un proceso para poder limpiar los datos y extraer ciertos caracteres interno que no corresponden al texto.

4.2. Servicios en la nube

El servicio en la nube que se usa principalmente es el de AWS llamado Lambda, es donde van a ejecutar las funciones desarrolladas en el punto anterior, esta se encargara de almacenar cada una de ellas y poder ejecutar a demanda cada solicitud que se realice mediante un API desde el Robot N°01 (Figura N° 2), otras de las funciones dentro de este ecosistema se ejecutaran desde algunos de los procesadores que tiene Nifi Apache para realizar solicitudes tipo Get.

ACTA DE APROBACIÓN DE ORIGINALIDAD DE TRABAJO ACADÉMICO

Yo, Edwin A. Martinez Lopez, docente de la Escuela de Posgrado de la Universidad César Vallejo - filial Lima Norte.

La tesis titulada "Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres" en el curso de Diseño y Desarrollo del Trabajo de Investigación para la Maestría en Ingeniería de Sistemas con mención en Tecnologías de la Información, del estudiante **Antonio Federico Martínez Núñez**, constato que la investigación tiene un índice de similitud de 13% verificable en el reporte de originalidad del programa Turnitin.

El suscrito, analizó dicho reporte y concluyo que cada una de las coincidencias detectadas no constituye plagio. A mi leal saber y entender la tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad César Vallejo.

Lima, 01 de agosto del 2020



Dr. Edwin A. Martinez Lopez
Docente de la EPG – UCV



ESCUELA DE POSGRADO

**PROGRAMA ACADÉMICO DE MAESTRÍA EN INGENIERÍA DE SISTEMAS
CON MENCIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN**

**Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San
Martín de Porres**

TESIS PARA OBTENER EL GRADO ACADÉMICO DE:

Maestro en ingeniería de sistemas con mención en tecnologías de la información

AUTOR:

Br. Antonio Federico Martinez Nuñez (ORCID: 0000-0003-4364-2866)

ASESOR:

Dr. Edwin Alberto Martínez López (ORCID: 0000-0002-1769-1181)

Resumen de coincidencias X

13 %

1	repositorio.ucv.edu.pe Fuente de Internet	4 %	>
2	Entregado a Universida... Trabajo del estudiante	1 %	>
3	es.scribd.com Fuente de Internet	1 %	>
4	www.academia.edu Fuente de Internet	1 %	>
5	issuu.com Fuente de Internet	1 %	>
6	api.eoi.es Fuente de Internet	1 %	>
7	Entregado a Universida... Trabajo del estudiante	<1 %	>
8	Entregado a Universida... Trabajo del estudiante	<1 %	>
9	docplayer.es Fuente de Internet	<1 %	>
10	repositorio.espe.edu.ec Fuente de Internet	<1 %	>
11	aleph.uned.ac.cr Fuente de Internet	<1 %	>



UNIVERSIDAD CÉSAR VALLEJO

AUTORIZACIÓN DE LA VERSIÓN FINAL DEL TRABAJO DE INVESTIGACIÓN

CONSTE POR EL PRESENTE EL VISTO BUENO QUE OTORGA EL ENCARGADO DE INVESTIGACIÓN DE

ESCUELA DE POSGRADO

A LA VERSIÓN FINAL DEL TRABAJO DE INVESTIGACIÓN QUE PRESENTA:

Antonio Federico Martinez Nuñez

INFORME TITULADO:

Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San

Martin de Porres.


PARA OBTENER EL TÍTULO O GRADO DE:

Maestro en Ingeniería de Sistemas con Mención en Tecnologías de la Información

SUSTENTADO EN FECHA: 16 de agosto de 2020

NOTA O MENCIÓN: Aprobar por excelencia




DOCENTE DE LA ESCUELA DE POSGRADO
FILIAL LIMA NORTE



Centro de Recursos para el Aprendizaje y la Investigación (CRAI)
"César Acuña Peralta"

FORMULARIO DE AUTORIZACIÓN PARA LA PUBLICACIÓN ELECTRÓNICA DE LAS TESIS

1. DATOS PERSONALES

Apellidos y Nombres: (solo los datos del que autoriza)

Martinez Nuñez, Antonio Federico

D.N.I. : 45676320

Domicilio : Mz E15 Lte 09 – Mi Perú

Teléfono : Fijo : 5538365 Móvil : 987273044

E-mail : Anthonio0318@gmail.com

2. IDENTIFICACIÓN DE LA TESIS

Modalidad:

Tesis de Pregrado

Facultad :

Escuela :

Carrera :

Título :

Tesis de Posgrado

Maestría

Doctorado

Grado : Maestro en Ingeniería de Sistemas

Mención : Tecnología de la información

3. DATOS DE LA TESIS

Autor (es) Apellidos y Nombres:

Martinez Nuñez, Antonio Federico

Título de la tesis:

Automatización de web Scraping de los diarios de noticias para la empresa Isuri,
San Martin de Porres.

Año de publicación : 2020

4. AUTORIZACIÓN DE PUBLICACIÓN DE LA TESIS EN VERSIÓN ELECTRÓNICA:

A través del presente documento, autorizo a la Biblioteca UCV-Lima Norte, a publicar en texto completo mi tesis.

Firma :

Fecha : 23 de setiembre de 2020