ESCUELA POLITÉCNICA NACIONAL

ESCUELA DE FORMACIÓN DE TECNÓLOGOS

TRABAJO FINAL

Paúl Guala
@epn.edu.ec

Nathaly Guallichico
nathaly.guallichico@epn.edu.ec

Ruben Maya
@epn.edu.ec

Ligia Elena Pérez Bautista
ligia.perez@epn.edu.ec

Quito, marzo 2021

INDICE

INDICE DE ILUSTRACIONES

INDICE DE TABLAS

1. INTRODUCCIÓN

El presente proyecto se centra en la extracción de datos de distintas temáticas siendo de fuentes como Facebook, Kaggle, Inec, TikTok entre otras para realizar una arquitectura en la cual utilizaremos bases de datos NoSQL y bases relacionales así recopilando todo o parcialmente los datos que fueron extraídos. Posterior realizaremos la indexación de los nodos con elasticsearch y visualizaremos en tiempo real.

1.1 Objetivo general

Desarrollar el proyecto final de Análisis de Datos

1.2. Objetivos específicos

- Almacenar la data de todas las temáticas en un clúster con su background.
- Generar un dasboard de cada temática establecida.
- Generar una arquitectura para las visualizaciones correspondientes.

2. METODOLOGÍA

En la actualidad el uso de metodologías ágiles ayuda a desarrollar proyectos con rapidez y flexibilidad, estas metodologías tienen la capacidad de responder al cambio durante el proceso de creación del producto o servicio.

2.1 Metodología de Desarrollo

Scrum permite definir claramente cada una de las etapas y el proceso adecuado para llevar a cabo con éxito la implementación en el desarrollo del proyecto. Una de las etapas más relevantes es la planificación, puesto que a través del levantamiento de la información se cubren todas las actividades del proyecto. Por otra parte, es el trabajo colaborativo ya que a través de reuniones y entregas continuas con los integrantes se ha logrado entregar progresivamente avances funcionales del proyecto.

2.1.1 Roles

En Scrum la participación de los roles es indispensable para la realización del proyecto, ya que deben estar comprometidos con el mismo y son responsables del éxito de cada Sprint. Es por ello por lo que, aplicando Scrum se han definido los siguientes roles para el proyecto integrador:

Propietario del Producto (Product Owner)

Este rol lo desempeña el Ingeniero, quien es el encargado de proporcionar toda la información sobre el proyecto final, permitiendo cumplir con la fase de planificación y determinar de esta manera: herramientas, arquitectura.

Scrum Master

Este rol lo desempeña el director del proyecto, quien guía al equipo de desarrollo a cumplir el objetivo planteado, organizando una serie de reuniones con el objetivo de que en cada Sprint sea finalizado de manera exitosa y asegurar de que se apliquen las buenas prácticas y reglas de manifiesto ágil.

Equipo de desarrollo (Developer Scrum)

Este rol lo cumplen los integrantes del proyecto, encargados de transformar los

requerimientos del ingeniero en pequeños avances funcionales al terminar cada uno de los Sprint.

El equipo Scrum se encuentra conformado por el siguiente grupo de trabajo, como se presenta en la **TABLA 1**.

Tabla 1 Roles asignados

NOMBRE	**ROL**
Ing. Juan Zaldumbide	Product Owner
Paúl Guala	Scrum Master
Nathaly Guallichico Ruben Maya	
Elena Pérez	Equipo de desarrollo

2.1.2 Cronograma de actividades

Tabla 2 Sripnt 1

Sprint 1		Semana 1	1-6 marzo	
id	Prioridad	Descripción (enfoque Front End)	Est. (horas)	Por
1	Alta	Extracción de datos de la temática 1 y 2	46	PG
2	Alta	Extracción de datos de la temática 5	26	NG
3	Alta	extracción de datos de la temática 4,3	26	RM
4	Alta	Extracción de datos de la temática 5	26	EP

Tabla 3 Sprint 2

Sprint 2		Semana 2	8-13 marzo	
id	Prioridad	Descripción (enfoque Front End)	Est. (horas)	Por
1	Alta	Visualización con couchdb usando logstash y kibana		PG
2	Alta	Visualización con SQLite usando logstash y kibana		NG
3	Alta	Visualización con SQLite usando PowerBI		RM
4	Alta	Visualización con MySQL usando logstash y kibana	26	EP

2.2 Diseño de la arquitectura

2.2.1 Herramientas utilizadas para el proyecto

Para guardar lo que son los datos extraídos de distintas fuentes se ha utiliza las bases de datos como CouchDB, MongoDB, MongoDB Atlas, MySQL, SQL Server y SQLite. Para realizar la indexación ocupamos elasticseacrh y cerebro, como también losgstash para integrar todos nuestros datos. Y por último para visualizaciones en tiempo real utilizamos lo que es Kibana.

2.2.2 Arquitectura de la solución

La **Ilustración 1** indica la estructura del proyecto como se va a realizar las extracciones de datos de distintas fuentes y en que bases van a ser guardadas, después realizaremos el Logstash y por último terminaremos con las visualizaciones en tiempo real utilizando Kibana, también se va a realizar visualizaciones estáticas con PowerBi.

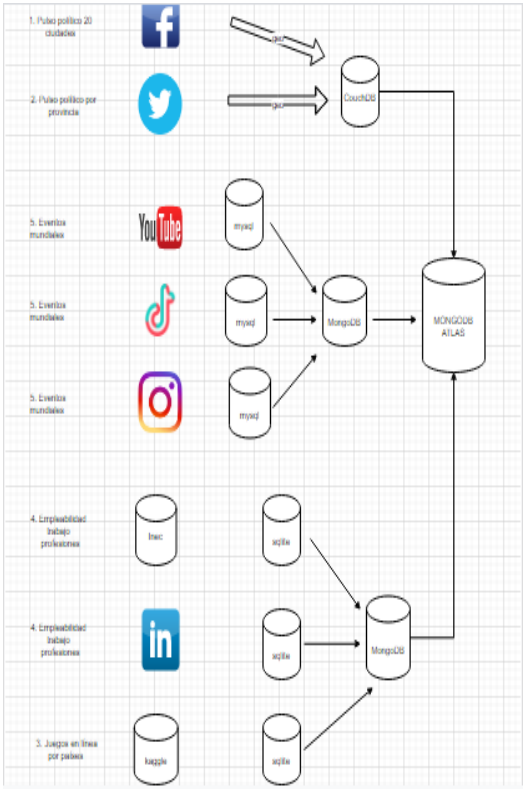


Ilustración 1 Estructura del proyecto

2.2.3 extracción de datos

2.2.3.1 Facebook y Twitter

Para la extracción de datos de la plataforma Twitter se utilizó la API oficial de la red social, tweepy, que permite la recolección de información de tweets basada en distintos parámetros de selección, como la ubicación geográfica, la selección por términos, el nombre de un usuario en específico, entre otras.

Las librerías Stream, OAuthHandler y StreamListerner permiten el acceso hacia la API de Twitter, y así, a la información almacenada en sus bases de datos.

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
```

Ilustración 2 Importación de los métodos alojados en la libreria tweepy

A continuación, se realiza la autenticación de las credenciales en las funciones anteriores, y se procede con la respectiva extracción de datos. Para su almacenamiento, se utilizó la base de datos local CouchDB.

```
auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
```

Ilustración 3 Configuración de las credenciales de acceso a la API de Twitter

El almacenamiento de información se realizó en la base de datos local CouchDB. Y a continuación, se utilizó la herramienta Logstash para la transferencia de datos hacia ElasticSearch, para su respectiva visualización en Kibana.

```
server = couchdb.Server('http://admin:admin8675423*@localhost:5984/')
try:
    db = server.create('arauz')
except:
    db = server['arauz']
```

Ilustración 4 Almacenamiento de los datos en la base CouchDB

Para la extracción de información de la red social Facebook se utilizó la librería facebook-scraper, que permite tener un acceso directo a la información sobre las publicaciones de una página pública, un perfil o un grupo.

A través del método get_posts, se puede recopilar información sobre las publicaciones, como el número de likes y reacciones, el número de comentarios, el texto, las imágenes o los vídeos incluidos en la publicación, etc.

```
from facebook_scraper import get_posts
```

Ilustración 5 Importación de los métodos alojados en la librería facebook_scraper

2.2.3.2 Youtube TikTok e Instagram

Para extraer datos de Youtube se ocupó Octoparse una aplicación que mediante el link de la página se puede extraer los datos.

La Ilustración muestra la aplicación donde ya está puesto el link para que se vaya extrayendo los datos que necesitamos.

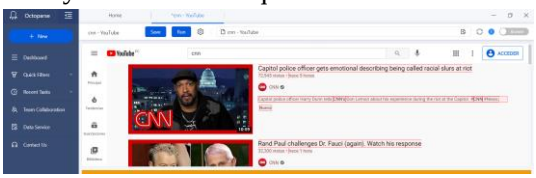


Ilustración 6 La aplicación octoparse

Para la extracción de datos de tik-tok utilizamos tiktok-scraper que consiste en utilizar 2 lineas de codigo en consola, este caso recopilamos datos del canal nbcnews en tiktok para determinar cuáles han sido sus tendencias en los últimos días. Para la inicialización de tik-tok scraper utilizamos la siguiente línea de código:

```
C:\Users\Dell\Desktop\DS\4toSemestre\AnalisisDatos\2do_Bimestre\proyectoFinal>npm i -g tiktok
-scraper
```

Ilustración 7 Línea de Código para la extracción

Posterior a esto vamos a generar nuestra segunda línea de código, en la cual definiremos el nombre de la cuenta de usuario de tik-tok, el número de datos que deseamos obtener y finalmente el formato del archivo.

```
C:\Users\Dell\Desktop\DS\4toSemestre\AnalisisDatos\2do_Bimestre\proyectoFinal>tiktok-scraper
user nbcnews -n 1000 -t csv
```

Ilustración 8 Definir el nombre de la cuenta de usuario de tik-tok

Este nos arroja la ruta donde se encuentra nuestro archivo con los datos recopilados

```
CSV path: C:\Users\Dell\Desktop\DS\4toSemestre\AnalisisDatos\2do_Bimestre\proyectoFinal/nbcne
ws_1615386619649.csv
```

Ilustración 9 Ruta donde se encuentra nuestro archivo

Para la extracción de datos de instagram utilizamos seobots.io que sirve para la recopilacion y extracción de datos de redes sociales como instagram, donde con las URL de las cuentas de donde queremos obtener información nos proporcionará un dataset en formato csv, excel y json.

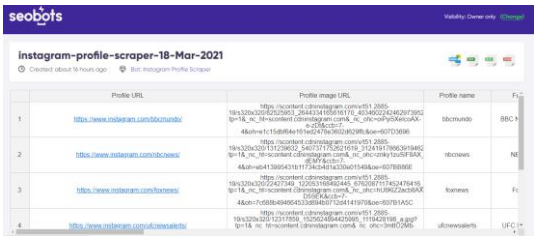


Ilustración 10 Seobots.io que sirve para la recopilación y extracción de datos

2.2.3.3 Kaggle LinkedIn INEC

Para la extracción de datos de la plataforma Linkedin se utilizó un script en python ejecutado el notebook de Jupyter para recolectar información de ofertas de trabajos en la plataforma, dentro de grupos, opcion empleos en el que seleccionamos los parámetros de empresa que oferta el empleo, el lugar, la descripción del trabajo a realizar y el nombre de la oferta; a qui hay que mencionar que se tomaron solo estos parámetro para que se pueda mezclar con otras fuentes que fueron consultadas también, como Computrabajo y indeed España y US. Con el objetivo de poder mostrar información contrastada y relevante para quien busca trabajo.

Se utilizó las librerías selenium de python, ya que esta nos permite navegar por la página raíz e interactuar con páginas dinámicas como en este el caso de Linkedin.

```
#Selenium imports here
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.webdriver.support.wait import WebDriverWait
import time
```

Ilustración 11 Importación de los métodos alojados en la librería Selenium.

A continuación, se realiza la autenticación de las credenciales, como correo electrónico, y contraseña.

```
#open the webpage
driver.get("https://www.linkedin.com/login/es")
#target username
username = WebDriverWait(driver, 10).until(EC.element_to_be_clickable((By.CSS_SELECTOR, "input[id='username']")))
password = WebDriverWait(driver, 10).until(EC.element_to_be_clickable((By.CSS_SELECTOR, "input[id='password']")))

#enter username and password
username.clear()
username.send_keys("rubenmaya753@yahoo.com")
password.clear()
password.send_keys("[redacted]")
```

Ilustración 12 Autentificación en Linkedin.

A continuación, se realiza un ciclo for para crear un ciclo de paginación y extracción de datos de la página que necesitamos. a través

de la selección de los elementos que queremos extraer mediante expresiones Xpath:

Ilustración 13 Ciclo de paginación y extracción de datos.

2.2.4 Análisis de la información

2.2.4.1 Análisis de la información de Facebook y Twitter

El análisis de la información se basó en los diferentes gráficos extraídos a través de la recolección de datos, alojados en la base local de couchd.

2.2.4.2 análisis de la información de YouTube, TikTok e Instagram

Para hacer un análisis de los datos que obtuvimos de youtube se utilizó elasticsearch.

En la Ilustración se muestra el análisis que se realizó y cuáles fueron los resultados.

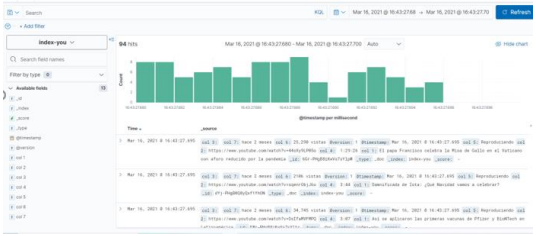


Ilustración 14 Analisis de datos de Youtube

Para realizar el análisis de tik-tok e instagram utilizamos elasticsearch, logstash y kibana (ELK) en donde instalamos logstash, mysql connector.jar ya que la base relacional que ocupamos es Mysql en donde estarán nuestros datos.

Generamos un código de conexión para establecer la conexión entre elasticsearch y nuestra base de datos relacional Mysql, los códigos que utilizamos para nuestras conexiones son las siguientes:

La Ilustración 15 nos muestra la conexión de nuestra base mysql nbc (datos extraídos de tik-tok) con nuestro elasticsearch nbc.

Ilustración 15 conexión de nuestra base mysql nbc (datos extraídos de tik-tok) con nuestro elasticsearch nbc

La Ilustración 16 nos muestra la conexión de nuestra base mysql nminsta (datos extraídos de instagram) con nuestro elasticsearch nm.

Ilustración 16 Conexión de nuestra base mysql nminsta (datos extraídos de instagram) con nuestro elasticsearch nm

2.2.4.3 Análisis de la información de INEC, LikedIn y Kaggle.

Para el análisis de la información recolectada tanto en la plataforma de Linkedin, Keggle y del instituto de estadísticas y censos del Ecuador (INEC)

se utilizó la Herramienta de análisis de datos de Microsoft PowerBI.

A continuación, se presenta los dashboard obtenidos

3. RESULTADOS Y DISCUSION

A continuación, se presenta los resultados obtenidos en la ejecución de las tareas de cada Sprint.

3.1. Sprint 0. Extracción de datos

El Sprint 0, define la extracción de datos sobre que fuentes se va a utilizar.

3.1.1 Extracción de datos en Facebook y Twitter

La extracción de los datos se realizó al acceder a las bases de datos de las redes sociales mencionadas: Facebook y Twitter. Una vez almacenados los datos, se trasladó la información hacia ElasticSearch en la nube, mediante la herramienta de indexación Logstash. Este proceso consiste en la creación de los diferentes índices donde se almacenará la información.

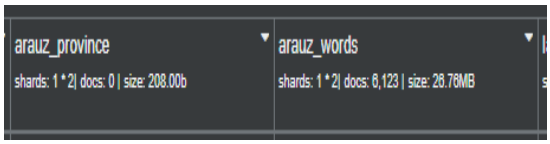


Ilustración 17 Creación de los índices en ElasticCloud

Finalmente, con los datos alojados en la nube, se crearon las diferentes gráficas utilizando la herramienta Kivana, para sus respectivos análisis.

3.1.2 Extracción de datos en TikTok, YouTube e Instagram

Mediante Octoparse se obtuvo los datos con los parámetros que se necesitaba guardandolos así en un archivo CSV para luego transferir a MySQL.
La Ilustración 18 muestra los datos extraídos.

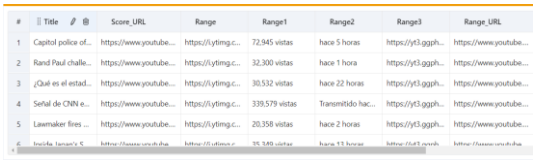


Ilustración 18 Datos de Youtube

La Ilustración 19 muestra el archivo ya en MySQL

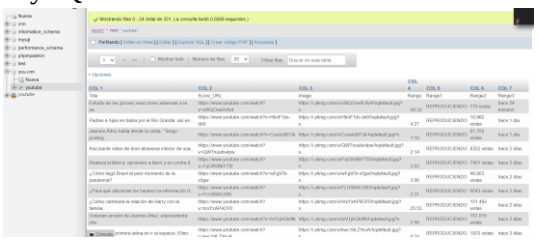


Ilustración 19 La base en MySQL

A través de nuestras líneas de comando previamente explicadas para la obtención de datos de tik-tok pudimos obtener nuestro dataset que se almaceno en la ruta previamente arrojada por el sistema.
La Ilustración 20 nos muestra nuestros datos de tik-tok ya almacenados.

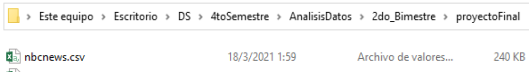


Ilustración 20 Datos de tik-tok

La ilustración 21 nos muestra nuestros datos de instagram ya almacenados.

Ilustración 21 Datos de instagram

La Ilustración 22 muestra los datos extraídos de tik-tok ya almacenados en nuestra base de datos Mysql.

Ilustración 22 los Datos extraídos de tik-tok ya almacenados en nuestra base de datos MySQL

La ilustración muestra los datos extraídos de instagram ya almacenados en nuestra base de datos Mysql.

Ilustración 23 Datos extraídos de instagram ya almacenados en nuestra base de datos MySQL

La Ilustración 24 muestra los datos recopilados de tik-tok que estaban almacenados en nuestra base de datos MySQL ya pasados a nuestra base elasticsearch.

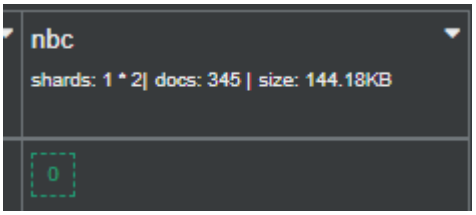


Ilustración 24 Base elasticsearch

La Ilustración 25 muestra los datos recopilados de instagram que estaban almacenados en nuestra base de datos MySQL ya pasados a nuestra base elasticsearch.

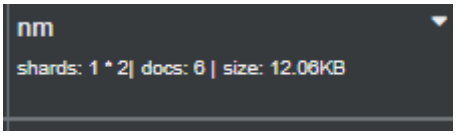
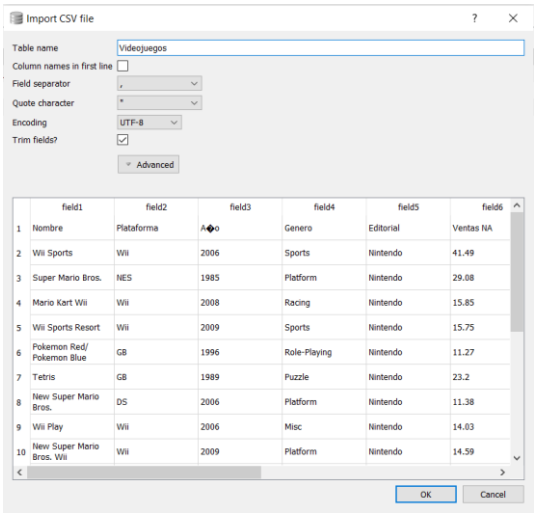


Ilustración 25 Base elasticsearch

3.1.3 Extracción de datos en INEC, Kaggle y LinkedIn

Mediante el dataset obtenido de kaggle se obtuvo los datos un archivo CVS
que luego del análisis con power BI serán migrados a la base de datos
Q Lite para luego ser enviado a MongoDB Compass y posteriormente a MongoDB Atlas a través de un archivo CSV.
A continuación, se indica cómo se realizó la migración de los datos



	field1	field2	field3	field4	field5	field6
1	Nombre	Plataforma	A◆o	Genero	Editorial	Ventas NA
2	Wii Sports	Wii	2006	Sports	Nintendo	41.49
3	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08
4	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85
5	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75
6	Pokemon Red/ Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27
7	Tetris	GB	1989	Puzzle	Nintendo	23.2
8	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38
9	Wii Play	Wii	2006	Misc	Nintendo	14.03
10	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59

Ilustración 26 Migración de los datos

Para el caso del INEC se descargaron mediante el sistema integrado de consultas (REDATAM) y los dataset disponibles en la página oficial de ecuador en cifras 2021 https://www.ecuadorencifras.gob.ec/.

Ilustración 27 Dataset disponibles en la página

3.2 Sprint 1. Visualización

El Sprint 1, define las visualizaciones que se van a realizar y que bases se ocuparan.

3.2.1. Visualización con Kibana

3.2.1.1 Vizualizacion de pulso político en 20 ciudades principales de Ecuador.

La primera gráfica muestra las ciudades del país con mayor presencia o menciones en la red social Twitter del candidato Andés Arauz. Estas las conforman Quito, con un aproximado del 17% de mención (504 menciones), seguida de Guayaquil, con un un aproximado de 15% (307 menciones), seguidas de Cuenca, Loja y Machala, con porcentajes menores al 5% (101 menciones).

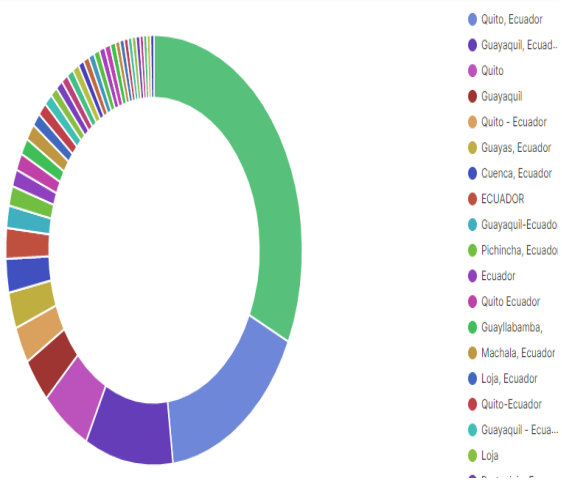


Ilustración 28 Ciudades del país con mayor presencia o menciones

La segunda gráfica muestra las ciudades del ecuador con mayor mención del candidato Guillermo Lasso, ubicándose en primer lugar la ciudad de Quito, con un aproximado de 140 menciones en una muestra de 2094. A continuación, se encuentra la ciudad de

Guayaquil de 100 menciones en la muestra, seguida de las ciudades de Cuenca, Esmeraldas y Sangolquí, con menciones de aproximadamente 40, 22 y 15 en la muestra, respectivamente.

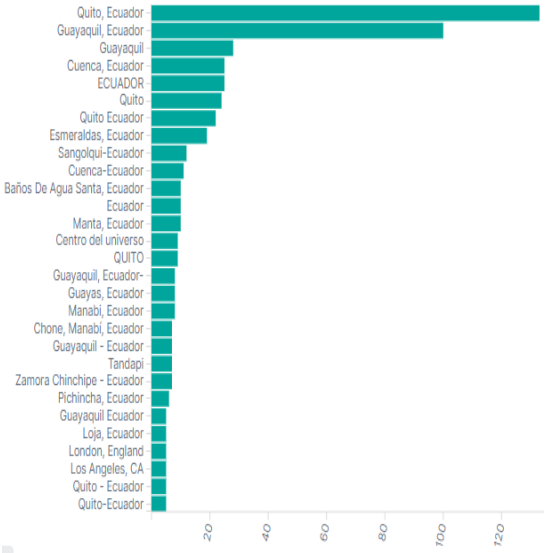


Ilustración 29 Ciudades del ecuador con mayor mención del candidato Guillermo Lasso

3.2.1.3 Vizualizacion de eventos o noticias mundiales.

La siguiente visualización muestra el top 10 de las noticias más reproducidas en la página nbcnews de tik-tok.

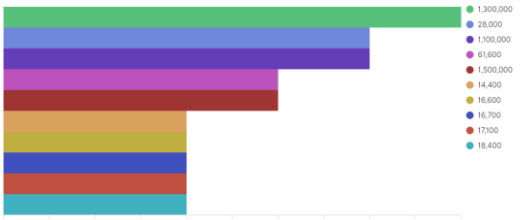


Ilustración 30 Top 10 de las noticias más reproducidas

Ilustración 31 Top 10 de las noticias más reproducidas

La siguiente visualización muestra el número de seguidores a nivel mundial de las noticias más reproducidas en las distintas cuentas de usuarios de instagram.

Ilustración 32 Número de seguidores a nivel mundial

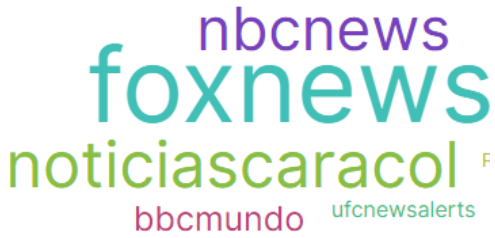


Ilustración 33 Número de seguidores a nivel mundial

Las siguientes visualizaciones se realizaron con elasticsearch son de los datos de Youtube.

La Ilustración 34 muestra el top 10 de las mejores noticias CNN.

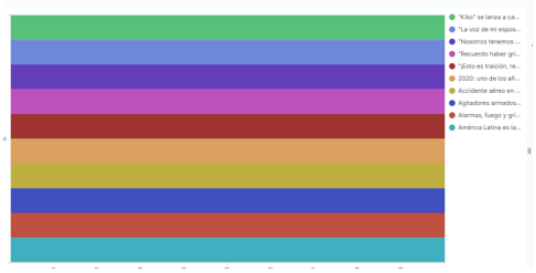


Ilustración 34 Top 10 de las mejores noticias CNN

La Ilustración 35 muestra el top 20 de las urls que más vistas tuvieron.

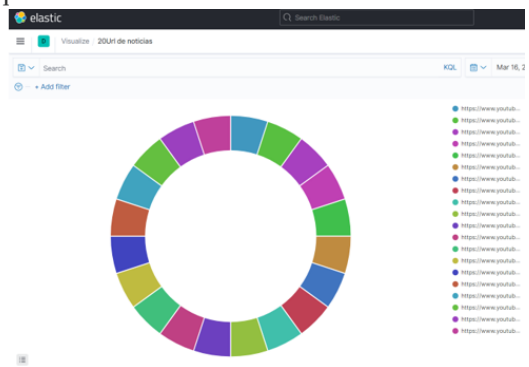


Ilustración 35 Top 20 de las urls que más vistas tuvieron

3.2.2. Visualización con PowerBI

3.2.2.1 Visualización de juegos en línea por países.

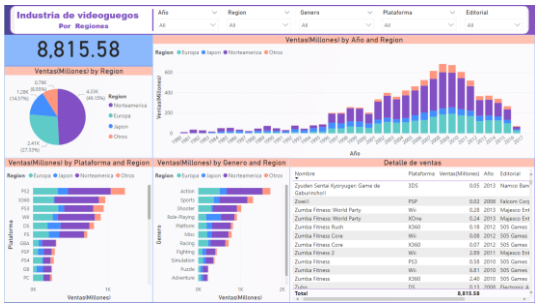


Ilustración 36 Juegos online

3.2.2.2 Visualización de ofertas de empleo.

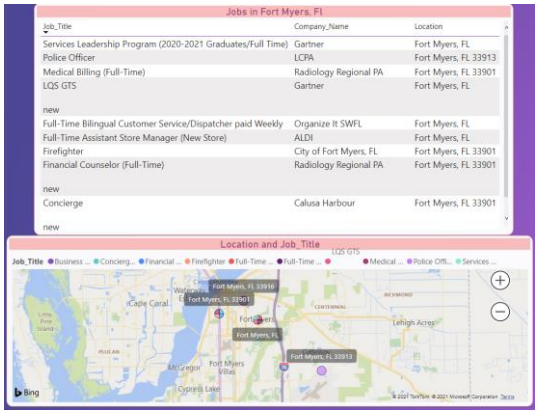


Ilustración 37 Ofertas de empleos

4. CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- ELK (elasticsearch, logstash, kibana) es un conjunto de herramientas de gran potencial de código abierto que se combinan para crear una herramienta de administración de registros permitiendo la monitorización, consolidación y análisis de logs generados en múltiples servidores.
- RapidMiner es una herramienta de Minería de Datos ampliamente usada para el pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de prueba de modelos, visualización de datos, etc.
- El análisis de los datos de redes sociales pone en manifiesto la poca interacción con las en algunas zonas del país, lo cual dificulta el llegar a conclusiones netamente objetivas o que comprendan en una gran mayoría el pensar ciudadano.
- Nos pudimos dar cuenta que Youtube no se lo utiliza solo para escuchar música sino también para ver noticias internacionales como lo es el canal CNN.

4.2 Recomendaciones

- El uso de una herramienta para realizar una limpieza correcta de los datos es necesaria para que nuestros datasets nos puedan dar buenas visualizaciones.
- Analizar el caso de estudio que se realizará, respecto a su región, sus sujetos de prueba, sus distintos casos, etc., para escoger la mejor estrategia que permita recolectar la información, procesarla y analizarla con la menor cantidad de recursos posibles.
- Definir correctamente cuáles serán las conexiones entre las distintas bases de datos y las fuentes de extracción de información, para llevar un flujo de trabajo y organización ordenados, evitando posibles futuras confusiones al momento de analizar la información y visualizarla.
- La etapa de cosecha para la recolección se le debe otorgar su

correspondiente tiempo y con anticipación para cubrir con los datos masivos sin retrasar las siguientes actividades.