Proyecto Final Análisis de datos

José Luis Colcha
jose.colcha@epn.edu.ec
Carlos Montesdeoca
Carlos.montesdeoca1@epn.edu.ec
Nathaly Bermeo
nathaly.bermeo@epn.edu.ec
Josselyn Stefania Troya Fuentes
Josselyn.troya@epn.edu.ec

I. Definición del caso de estudio

Se obtendrá dashboards de los siguientes casos de estudio:

A. Pulso político en 20 ciudades principales de Ecuador

En abril del año 2021 se realizarán las votaciones en el país por lo que es interesante realizar una búsqueda de datos para dar a conocer la opinión de la ciudadanía a cerca del pulso político, estos datos fueron recopilados por medio de Twitter ya que es una red social que utilizan a menudo los candidatos a la presidencia del Ecuador.

B. Pulso político por provincias en Ecuador

Ecuador no se caracteriza por ser un país que haga uso de la red Twitter, sin embargo, en cuestión de política utilizan diversas redes sociales como la mencionada anteriormente para expresar su opinión o apoyar al candidato de su preferencia, esto ha causado la curiosidad de saber que tan utilizada es esta red social en diversas provincias del país.
La recopilación de datos en cada una de las provincias acerca del uso de la red social Twitter y centrado en el ámbito político ha logrado dar resultados satisfactorios y despejar esa duda del porcentaje de provincias que tienen interés en la política

C. Juegos en línea por países.

Los juegos online son muy populares en el día de hoy y con la accesibilidad a internet cada país tiene su gran grupo de gamers.
Con ayuda de internet y diversas paginas se puede extraer datos de ratings de juegos, también crean estadísticas de jugadores por países y los ingresos que genera para la economía de cada país la industria de los videojuegos.

D. Tema definido por el estudiante

La recopilación de datos se realizó mediante un archivo csv exportado de la página data.world y hace referencia a los tiroteos por muertes fatales hechos por policías en EEUU, estos datos se realizan a menudo ya que es alto el índice de muertes en las diferentes ciudades del país.

E. Eventos o noticias mundiales

Se recopilaron los datos de kaggle relacionados a las personas vacunadas del Covid-19, para entender de mejor manera como va progresando mundialmente a un año de haberse declarado pandemia mundial.
También se recopilaron datos de una página web llamada statista, la cual proporciona la evolución del Covid-19 en el mundo desde el 3 de febrero de 2020 hasta el 12 de marzo de 2021.
Y finalmente se recopilaron datos de Facebook, de algunos portales de noticias más populares.

II. Objetivos Generales y específicos

A. Objetivos Generales

Recopilar datos a través de diferentes herramientas y técnicas para su análisis gráfico.

B. Objetivos Específicos

- Analizar los casos de estudio para el análisis
- Crear una arquitectura para recopilación de datos
- Desarrollar un cronograma de actividades.
- Dividir las actividades y herramientas a utilizar por integrante tomando como referencia la arquitectura.
- Recopilar datos de las fuentes establecidas en la arquitectura
- Realizar un análisis grafico usando las herramientas establecidas en la arquitectura.
- Documentar todo el análisis realizado por cada integrante del grupo.

III. Descripción del equipo de trabajo y actividades realizadas por cada uno

El equipo está conformado por 4 personas:

A. José Luis Colcha

- Recopilar datos en la red social Twitter en las 24 provincias del país.
- Guardar las bases de datos en el gestor CouchDB.
- Crear una réplica de todas las bases de datos.
- Migrar la base datos al gestor MongoDB Atlas.

- Exportar la base de datos al gestor de visualización PowerBi.
- Realizar las visualizaciones establecidas con los datos recolectados.

B. *Carlos Montesdeoca*

- Buscar datasets por internet
- Descargar dataset en formato csv
- Importar los datos a SQL server
- Exportar datos a Power BI
- Analizar grafico de los datos

C. *Nathaly Bermeo*

- Crear una base de datos en Mongo DB y Mongo DB Atlas con los datos recopilados de Twitter sobre el pulso político de 20 ciudades del Ecuador.
- Obtener visualizaciones en Power BI.
- Importar un archivo csv en MySQL phpMyadmin.
- Hacer conexión en tableu y obtener visualizaciones de los datos generados.

D. *Josselyn Troya*

- Recopilar datos relacionado al Covid-19 de la página web Kaggle y la página web statista.
- Guardar dicha información en la base de datos SQLite.
- Exportar los datos a ElasticSearch y poder realizar las visualizaciones con Kibana.
- Recopilar datos de *Facebook* de varios portales de noticias mundiales en la base de datos MongoDB.
- Con la librería de Python textBlog, se analizarán las noticias negativas y positivas.
- Exportar los datos para realizar las visualizaciones en PowerBI.

IV. CRONOGRAMA DE ACTIVIDADES

Tabla 1: Cronograma de actividades

Responsable	Actividades	2021			
		Semana 1	Semana 2	Semana 3	Total Horas
- José Colcha - Josselyn Troya - Carlos Montesdeoca - Nathaly Bermeo	**1. Planificación del Proyecto**				
	Selección de las herramientas a utilizarse.	3			3
	Distribución de tesis por miembro del equipo.	3			3
	Selección de bases SQL y no SQL.	2			2
	Establecimiento del tiempo de investigación.	2			2
- José Colcha - Josselyn Troya - Carlos Montesdeoca - Nathaly Bermeo	**2. Recolección de datos**				
	Búsqueda de datos en diferentes fuentes.	10	10		20
	Creación de scripts para recopilar datos.	2			2
	Ejecución de scripts para recopilar datos.	20	20	20	60
	Limpieza de datos			6	6
- Josselyn Troya - Nathaly Bermeo	**3. Concentración de datos**				
	Cargar los datos recopilados a sus respectivas bases de datos.		6	4	10
	Creación de índices en el servidor elasticsearch.		6	4	10
	Creación de scripts para migración a la herramienta elasticsearch.			5	5
- José Colcha - Josselyn Troya - Carlos Montesdeoca - Nathaly Bermeo	**4.Creación de dashboards.**				
	Importación de los datos recopilados a las herramientas de visualización.			4	4
	Limpieza de datos.			6	6
	Creación de visualizaciones.			8	8
- José Colcha - Josselyn Troya - Carlos Montesdeoca - Nathaly Bermeo	**5. Resultados**				
	Interpretación de cada visualización.			12	12
	Conclusiones y recomendaciones de los resultados obtenidos.			8	8
	Documentación de todo el proceso y resultados finales.			10	10
Total de horas/semana		39	42	87	171

V. RECURSO Y HERRAMIENTAS UTILIZADAS

A. *PYTHON*

Python es un lenguaje de programación con propósito general, por lo que puede desarrollarse en distintas áreas de la información. Es un lenguaje interactivo e interpretado con funciones incorporadas para el tratamiento de sus variables y cuenta con una sintaxis clara y fácil de entender

Para este proyecto en específico se está haciendo uso de las siguientes librerías:

- pymongo: utilizada para poder conectarnos a MongoDB.
- tweepy: librería fácil de usar para acceder a la API de Twitter.
- couchdb: utilizada para poder conectarnos a CouchDB.
- pymysql: utilizada para poder conectarnos a MySQL.
- pandas: es una biblioteca escrita como extensión de NumPy para manipulación y análisis de datos.

B. *APACHE COUCHDB*

El gestor de base de datos de código abierto Apache CouchDB o simplemente llamada CouchDB es una base NoSQL con el foco puesto para simular la web, utiliza Json para almacenar sus datos, Javascript como el lenguaje para realizar consultas sobre la base por medios de API's.

CouchDB no almacena los datos y sus relaciones en tablas. En cambio, cada base de datos es una colección de documentos independientes. Cada documento mantiene sus propios datos y su esquema auto-contenido.

El protocolo que emplea CouchDB en su replicación de datos se implementa en una variedad de proyectos y productos que abarcan todos los entornos informáticos desde clústeres de servidores distribuidos globalmente, pasando por teléfonos móviles hasta navegadores web.

Figura 1 CouchDB

C. *MONGODB*

MongoDB es una base de datos libre distribuida, basada en documentos y con aplicación general que ha sido diseñada para desarrolladores de aplicaciones modernas por su rango de escalabilidad y con tecnología acorde a la era de la nube.

El modelo para trabajar con los documentos en MongoDB resulta intuitivo y fácil de aprender ya que proporciona a los desarrolladores todas las funcionalidades que necesitan para

satisfacer los requisitos complejos a cualquier escala. Se proporcionan drivers para más de diez lenguajes, y además es sostenida por la comunidad desarrolladora.

Figura 2 MongoDB

D. XAMPP

XAMPP es una distribución de Apache gratuita y de fácil instalación que contiene MariaDB, PHP y Perl. El paquete de instalación de XAMPP ha sido diseñado para que resulte optimo y fácil de instalar y usar.

Una gran ventaja que nos ofrece es en cuanto a tiempo y recursos en instalar y ejecución, pues con XAMPP los componentes no actúan por separado, sino que sólo se requiere una pequeña fracción del tiempo necesario para descargar y ejecutar un archivo ZIP, tar, exe o fkl.

Figura 3 XAMPP

E. LOGSTASH

Este componente es el paso a ElasticSearch, y corresponde a la parte del procesamiento para ingresar nuestros datos que pueden ser en una multitud de fuentes, transformándolos y enviándolo a su destino mediante clases input y output respectivamente.

Figura 4 logstash

F. ELASTICSEARCH

Motor de búsqueda y analítica, que funciona como base de datos distribuida tanto a nivel de procesamiento como de información, otorgando consultas con mejores rendimientos.

Figura 5 ElasticSearch

G. KIBANA

Es el componente más visual de la ELK, en la que se puede manipular los datos a fin de conseguir visualizaciones en distintos tipos, comprender como fluyen las solicitudes en servicios y generación dashboards.

Figura 6 Kibana

H. TABLEAU

Es un software de visualización de datos gratuito que a partir de fuentes que puedes ser entre archivos de Excel como CSV u otros orígenes de datos, permiten generar visualizaciones de alto impacto gráfico e interactivo.

Además, una gran ventaja es que con la instalación de la herramienta se podrá crear libremente un perfil personal en la nube con 10 GB de espacio en donde podrás publicar tus visualizaciones y compartirlas de manera más eficiente y profesional.

Figura 7 Tableau

I. POWER BI

Power BI es una herramienta software de Microsoft, para el análisis de datos en el ámbito empresarial, que ofrece un alto rendimiento en visualizaciones de resultados, por sus graficas dinámicas e interfaz fácil de usar. Power BI incorpora capacidades de inteligencia empresarial, con el que facilita el entendimiento de gráficos y visualizaciones, para crear informes y paneles. Power BI ofrece también protección de

extremo a extremo en la transmisión de datos, conexión a servicios de nube y otros servicios de Microsoft.

Figura 8 Power BI

J. SQLITE

SQLite es una herramienta de software libre, que permite almacenar información en dispositivos empotrados de una forma sencilla, eficaz, potente, rápida y en equipos con pocas capacidades de hardware.
SQLite soporta desde las consultas más básicas hasta las más complejas del lenguaje SQL, y lo más importante es que se puede usar tanto en dispositivos móviles como en sistemas de escritorio, sin necesidad de realizar procesos complejos de importación y exportación de datos, ya que existe compatibilidad al 100% entre las diversas plataformas disponibles, haciendo que la portabilidad entre dispositivos y plataformas sea transparente.

Figura 9 SQLite

K. MICROSOFT SQL SERVER

SQL Server es un sistema de gestión de bases de datos relacionales (RDBMS) de Microsoft que está diseñado para el entorno empresarial. SQL Server se ejecuta en T-SQL (Transact -SQL), un conjunto de extensiones de programación de Sybase y Microsoft que añaden varias características a SQL estándar, incluyendo control de transacciones, excepción y manejo de errores, procesamiento fila, así como variables declaradas.

Figura 10 SQL Sever

L. MONGODB ATLAS

Una base de datos mundial basada en la nube y completamente administrada de MongoDB que combina modelos de datos similares a JSON, indexación y búsqueda avanzadas, y escalabilidad elástica, a la vez que automatiza las tareas administrativas que llevan mucho tiempo.

Figura 11 MongoBD Atlas

M. TEXTBLOB

Es una librería Python (2 y 3) para procesar datos de texto. Proporciona una API consistente para sumergirse en tareas comunes de procesamiento de lenguaje natural (PNL) tales como etiquetado de parte del habla, extracción de frases sustantivas, análisis de sentimientos, y más.

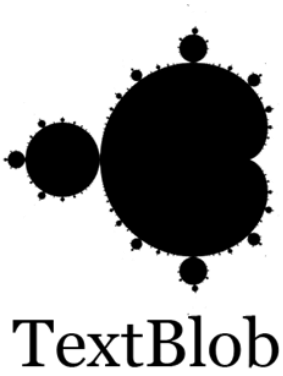

Figura 12 TextBlob

VI. ARQUITECTURA DE LA SOLUCIÓN

En la arquitectura se presenta una recopilación de datos a través de diferentes fuentes. La arquitectura está compuesta de dos partes. La primera parte consta de la recopilación de la red social *Twitter* mediante el uso de API's, para posteriormente almacenarlos en la base de datos CouchDB y MongoDB. También se recolecto de la red social *Facebook,* los datos se almacenaron en *MongoDB* y se realizó un análisis de sentimientos con la librería textBlob. Por último, se realizó la obtención de csv's en fuentes oficiales relacionadas con los temas a tratar y se los almaceno en SQL Server.
Tras la cosecha de la primera parte de la arquitectura, una parte de los datos se trasladaron a MongoDBAtlas. A final se realizó

los dashboards correspondientes en PowerBI.

La segunda parte de la arquitectura consta de la recopilación de csv's en fuentes oficiales relacionadas con los temas. Los datos se trasladaron a las bases de datos *SQLite* y *MySQL*, para finalmente utilizar *ELK* para generar las visualizaciones.

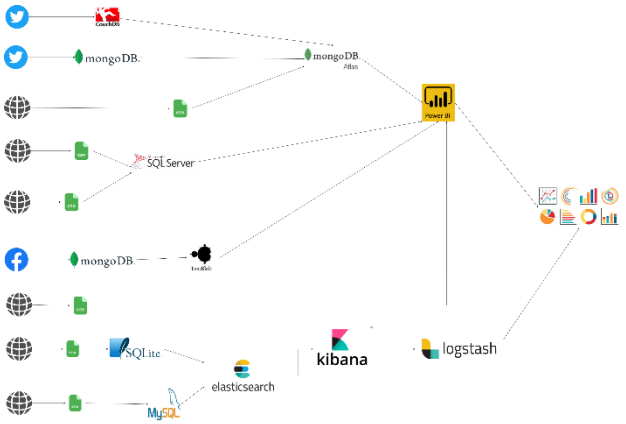

Figura 13 Arquitectura

VII. Extracción de Datos

A. *Pulso político en 20 ciudades principales de Ecuador*

La recopilación de datos de diferentes 20 ciudades del Ecuador se realizó a través de Twitter con el uso API's para la recolección de datos de dicha plataforma. Se recaudó datos acerca del pulso político con palabras claves como "Arauz" y "Lasso" ya que son los únicos candidatos para las siguientes elecciones en el país. Los datos son guardados en MongoDB para posteriormente vincular con MongoDB Atlas, para realizar las visualizaciones se utilizó la herramienta Power BI, en esta base de datos se recaudó aproximadamente más de 32000 documentos.

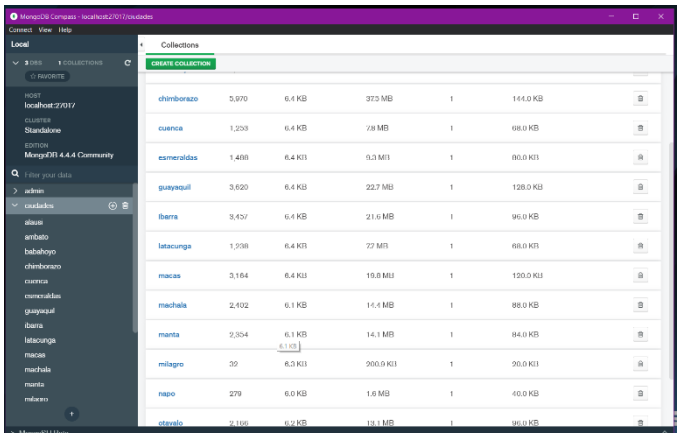

Figura 14 Colecciones en la base de datos de MongoDB

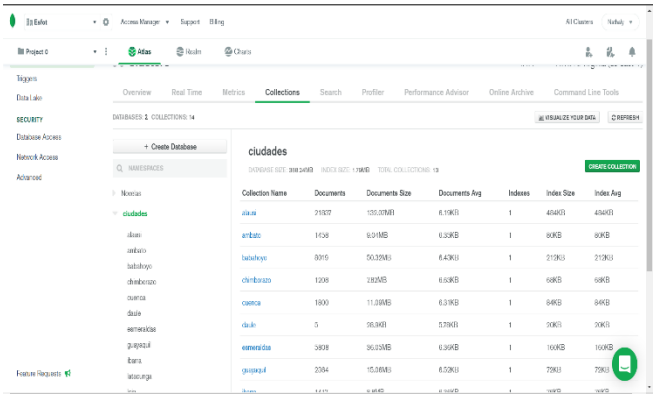

Figura 15 Colecciones en la base de datos de MongoDB Atlas.

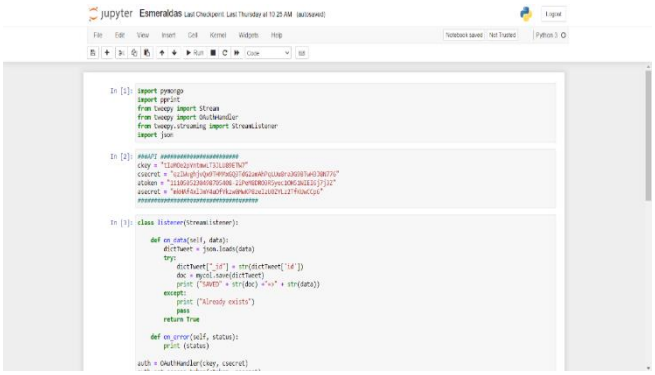

Figura 16 Script de la recopilación de Twitter a MongoDB.

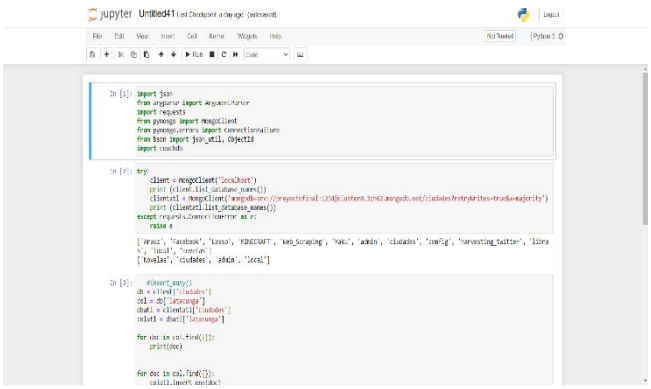

Figura 17 Script para migrar datos de MongoDB a MongoDB Atlas.

B. *Pulso político por provincias en Ecuador*

La recopilación de datos de cada una de las provincias del ecuador en la red social Twitter se llevó a cabo a través de un script, el cual contenía credenciales de desarrollador las mismas que permitieron acceder a la API de dicha red social, además se filtró por diversas palabras para poder obtener un mejor resultado se utilizó palabras como: candidatos 2021, votaciones 2021, pulso político, etc. La geolocalización también fue

utilizada para recolectar datos más precisos en cada una de las 24 provincias y almacenar las bases de datos en el gestor CouchDB con su respectivo nombre, se planifico a recopilar 5 horas al día datos y se logró recopilar una cantidad aproximada de 30000 documentos.

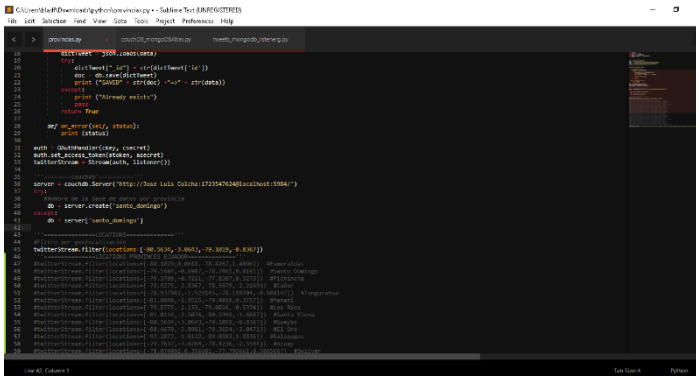

Figura 14 Script para recopilar datos a CouchDB

Luego de almacenar los datos de cada provincia en couchDB se procedió a crear un replica con todas las provincias incluidas y mediante la creación de un script se procede a migrar de base de datos la a mongoDBAtlas.

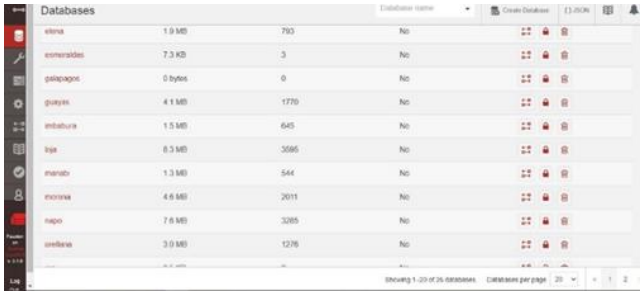

Figura 15 Bases de datos en CouchDB

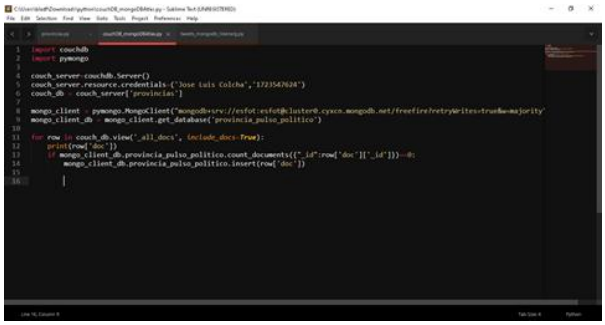

Figura 16 Script para migrar de CouchDB a MongoDB Atlas

C. Juegos en línea por países.

Para la extracción de datos se buscó páginas con contenido real tanto de ingresos económicos generados a través de los jugadores online por países, pobladores y gamers en las zonas analizadas y por último un último data set que presenta el número de jugadores por millones en el juego de WarZone.
Una vez obtenida la información se procede a importar los datos a SQL Server para lo cual se debe crear una base de datos llamada examen.

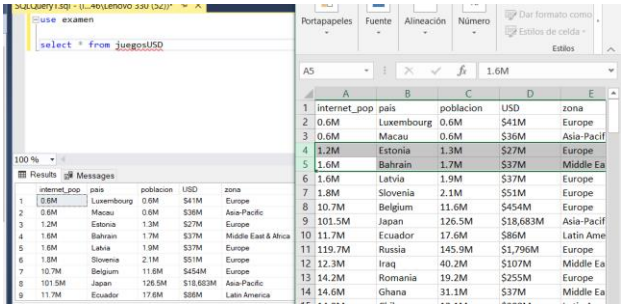

Figura 17 Bases de datos importada desde .csv a SQL Server

Luego de importar los dataset a SQL server se procedió a exportar las mismas bases de datos a la aplicación de Power BI para poder crear las visualizaciones.
La aplicación Power BI tiene dentro de sus funciones la de poder conectar directamente a SQL server por lo que el proceso rápido y se puede acceder directamente a todas las tablas creadas al momento de la importación.

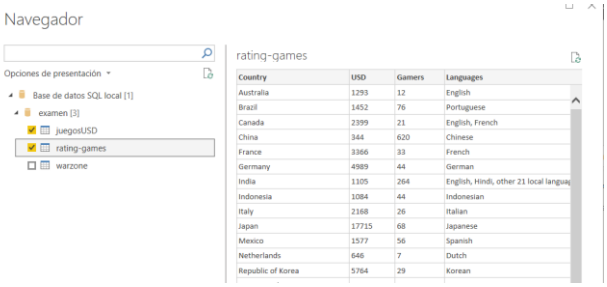

Figura 18 Bases de datos en exportados en Power BI.

D. Tema definido por el estudiante

Una vez descargada el archivo .csv de la página web data. world se procede a importar en phpMyadmin para tener la visualización de la tabla como tal, aproximadamente se encuentra 2.91 datos.

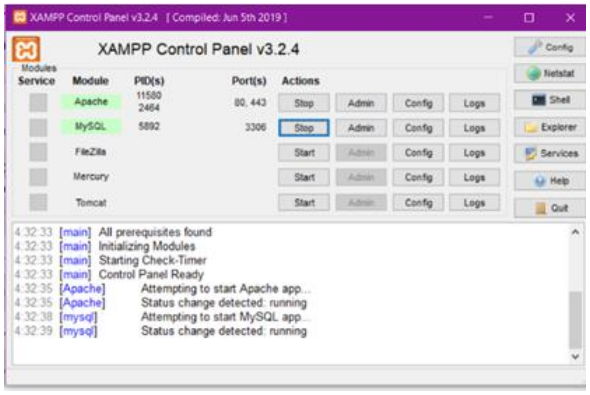

Figura 19 Herramienta XAMPP

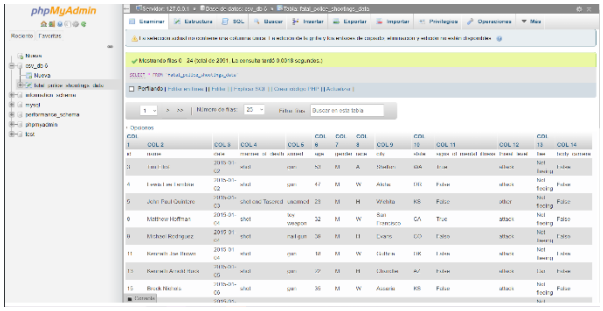

Figura 20 Data set importada en MySQL phpMyadmin.

E. Eventos o noticias mundiales

1) Facebook

Se generan scripts .py para recopilar los datos de *Facebook*, del portal de noticias BBCnewsMundo, CNN en Español, El Clarin, El mundo y RT media.

```python
from facebook_scraper import get_posts
import pymongo
import json
import time

myclient = pymongo.MongoClient("mongodb://localhost:27017")

try:
    mydb=myclient['noticias_mundo']
    mycol=mydb['bbc_news_mundo']
except:
    mydb=myclient['noticias_mundo']
    mycol=mydb['bbc_news_mundo']

i=1
for post in get_posts('BBCnewsMundo', pages=100000, extra_info=True):
    print(i)
    i=i+1
    time.sleep(5)

    id=post['post_id']
    doc={}

    doc['id']=id

    mydate=post['time']

    try:
        doc['texto']=post['text']
        doc['date']=mydate.timestamp()
        doc['likes']=post['likes']
        doc['comments']=post['comments']
        doc['shares']=post['shares']
        try:
            doc['reactions']=post['reactions']
        except:
            doc['reactions']={}


        doc['post_url']=post['post_url']
        mycol.save(doc)


        print("guardado exitosamente")

    except Exception as e:
        print("no se pudo grabar:" + str(e))
```

Figura 21 Script para recopilar datos de BBCnewsMundo a MongoDB

```python
from facebook_scraper import get_posts
import pymongo
import json
import time

myclient = pymongo.MongoClient("mongodb://localhost:27017")

try:
    mydb=myclient['noticias_mundo']
    mycol=mydb['cnn_en_español']
except:
    mydb=myclient['noticias_mundo']
    mycol=mydb['cnn_en_español']

i=1
for post in get_posts('CNNee', pages=100000, extra_info=True):
    print(i)
    i=i+1
    time.sleep(5)

    id=post['post_id']
    doc={}

    doc['id']=id

    mydate=post['time']

    try:
        doc['texto']=post['text']
        doc['date']=mydate.timestamp()
        doc['likes']=post['likes']
        doc['comments']=post['comments']
        doc['shares']=post['shares']
        try:
            doc['reactions']=post['reactions']
        except:
            doc['reactions']={}


        doc['post_url']=post['post_url']
        mycol.save(doc)


        print("guardado exitosamente")

    except Exception as e:
        print("no se pudo grabar:" + str(e))
```

Figura 22 Script para recopilar datos de CNN en español a MongoDB

```python
from facebook_scraper import get_posts
import pymongo
import json
import time

myclient = pymongo.MongoClient("mongodb://localhost:27017")

try:
    mydb=myclient['noticias_mundo']
    mycol=mydb['clarincom']
except:
    mydb=myclient['noticias_mundo']
    mycol=mydb['clarincom']

i=1
for post in get_posts('clarincom', pages=100000, extra_info=True):
    print(i)
    i=i+1
    time.sleep(5)

    id=post['post_id']
    doc={}

    doc['id']=id

    mydate=post['time']

    try:
        doc['texto']=post['text']


        doc['post_url']=post['post_url']
        mycol.save(doc)


        print("guardado exitosamente")

    except Exception as e:
        print("no se pudo grabar:" + str(e))
```

Figura 23 Script para recopilar datos del Clarín a MongoDB

```python
from facebook_scraper import get_posts
import pymongo
import json
import time

myclient = pymongo.MongoClient("mongodb://localhost:27017")

try:
    mydb=myclient['noticias_mundo']
    mycol=mydb['elmundo']
except:
    mydb=myclient['noticias_mundo']
    mycol=mydb['elmundo']

i=1
for post in get_posts('elmundo', pages=100000, extra_info=True):
    print(i)
    i=i+1
    time.sleep(5)

    id=post['post_id']
    doc={}

    doc['id']=id

    mydate=post['time']

    try:
        doc['texto']=post['text']


        doc['post_url']=post['post_url']
        mycol.save(doc)


        print("guardado exitosamente")

    except Exception as e:
        print("no se pudo grabar:" + str(e))
```

Figura 24 Script para recopilar datos de El Mundo a MongoDB

```python
from facebook_scraper import get_posts
import pymongo
import json
import time

myclient = pymongo.MongoClient("mongodb://localhost:27017")

try:
    mydb=myclient['noticias_mundo']
    mycol=mydb['esRTmedia']
except:
    mydb=myclient['noticias_mundo']
    mycol=mydb['esRTmedia']

i=1
for post in get_posts('esRTmedia', pages=100000, extra_info=True):
    print(i)
    i=i+1
    time.sleep(5)

    id=post['post_id']
    doc={}

    doc['id']=id

    mydate=post['time']

    try:
        doc['texto']=post['text']


        doc['post_url']=post['post_url']
        mycol.save(doc)


        print("guardado exitosamente")

    except Exception as e:
        print("no se pudo grabar:" + str(e))
```

Figura 25 Script para recopilar datos de esRTmedia a MongoDB

Se recolectaron un total de 23.470 datos de los cuales se los almaceno en MongoDB.

Collection Name	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size	Properties
bbc_news_mundo	10,207	465.8 B	4.5 MB	1	228.0 KB	
clarincom	1,273	320.4 B	398.3 KB	1	60.0 KB	
cnn_en_español	9,336	474.1 B	4.2 MB	1	204.0 KB	
elmundo	1,474	358.3 B	515.8 KB	1	68.0 KB	
esRTmedia	1,180	365.2 B	420.8 KB	1	60.0 KB	

Figura 26 Datos en MongoDB

2) Kaggle

Se obtuvo un csv del progreso mundial de vacunación contra Covid-19. Vacunación diaria y total contra COVID-19 en el mundo, actualizado hasta el 14 de marzo.
En total se obtienen 6.187 datos.

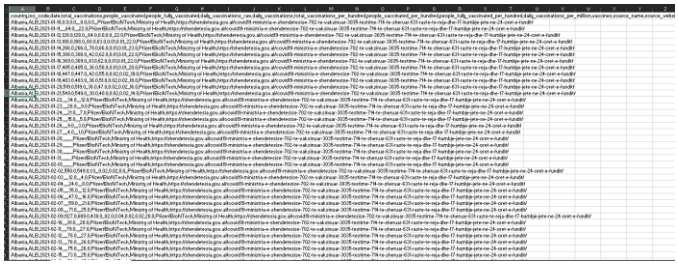

Figura 27 csv de personas vacunadas

3) Statista

Se obtuvo varios archivos .xls donde se registra el número de personas contagiadas, fallecidas y recuperadas de Covid-19.

En total se obtienen 475 datos.

	A	B	C
1	países afecta	total personas	
2	EE.UU	29.925.902	
3	India	11.308.846	
4	Brasil	11.284.269	
5	Rusia	4.370.617	
6	Reino Unido	4.241.677	
7	Francia	3.990.331	
8	España	3.178.356	
9	Italia	3.149.017	
10	Turquía	2.835.989	
11	Alemania	2.546.510	
12	Colombia	2.290.539	
13	Argentina	2.177.898	
14	México	2.151.028	
15	Polonia	1.868.297	
16	Irán	1.723.470	
17	Sudáfrica	1.525.648	
18	Ucrania	1.438.468	
19	Indonesia	1.410.134	
20	Perú	1.394.571	
21	Rep. Checa	1.376.998	
22	Países Bajos	1.138.796	
23	Canadá	899.757	
24	Chile	873.512	
25	Rumanía	845.352	
26	Israel	814.250	
27	Portugal	812.575	

Figura 28 Casos confirmados

	A	B	C	D
1	fecha	curados		
2	3-feb-20	643		
3	10-feb-20	4.043		
4	17-feb-20	12.712		
5	24-feb-20	29.467		
6	2-mar-20	48.108		
7	9-mar-20	64.058		
8	16-mar-20	79.658		
9	23-mar-20	102.168		
10	30-mar-20	169.773		
11	6-abr-20	307.554		
12	13-abr-20	525.013		
13	20-abr-20	807.327		
14	27-abr-20	1.147.774		
15	4-may-20	1.450.918		
16	11-may-20	1.796.412		
17	18-may-20	2.195.657		
18	25-may-20	2.710.384		
19	1-jun-20	3.316.022		
20	8-jun-20	3.996.711		
21	15-jun-20	4.717.314		
22	22-jun-20	5.506.787		
23	29-jun-20	6.331.694		
24	6-jul-20	7.286.012		
25	13-jul-20	8.358.992		
26	20-jul-20	9.571.001		
27	27-jul-20	10.912.089		

Figura 29 Casos recuperados

	A	B	C
1	Paises	Número de muertes	
2	EE.UU	543.721	
3	Brasil	273.124	
4	México	193.142	
5	India	158.326	
6	Reino Unido	125.168	
7	Italia	101.184	
8	Rusia	91.220	
9	Francia	89.830	
10	Alemania	73.560	
11	España	72.085	
12	Irán	61.016	
13	Colombia	60.858	
14	Argentina	53.493	
15	Sudáfrica	51.110	
16	Perú	48.484	
17	Polonia	46.724	
18	Indonesia	38.229	
19	Turquía	29.290	
20	Ucrania	27.915	
21	Rep. Checa	22.865	
22	Canadá	22.371	
23	Bélgica	22.370	
24	Chile	21.362	
25	Rumanía	21.252	
26	Portugal	16.635	
27	Hungría	16.627	

Figura 30 Número de fallecidos

VIII. ANÁLISIS DE LA INFORMACIÓN

A. Pulso político en 20 ciudades principales de Ecuador

Los datos recopilados en Twitter mostraron los reetweets que se hizo en cada ciudad, número de tweets, fecha, tweets, etc. Según el pulso político o candidato puesto en el script, se analiza el data set por medio de Power BI.

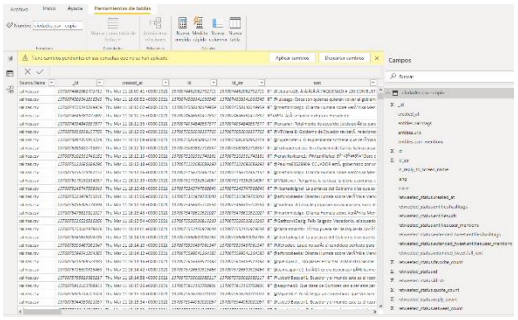

Figura 31 Datos de Twitter en Power BI.

B. Pulso político por provincias en Ecuador

La limpieza de datos fue lo principal antes del respectivo análisis, se la hizo mediante Excel se procedió a borrar espacios vacíos, suprimir columnas innecesarias, escribir cada provincia en mayúscula, de esta manera solo se conservó los datos necesarios como los son:

- Text Filter: Filtra los nombres de los 5 candidatos más mencionados.
- Country: Filtra el país de donde preceden los tweets en este caso Ecuador.
- Place: Filtra la provincia de donde proceden los tweets.
- Date: Filtra la fecha de la creación del tweet.
- Retweet count: Filtra el total de retweets de cada tweet.
- Reply count: Filtra el total de respuesta de cada tweet.
- Favorites: Filtra el total de favoritos en los que se ha colocado un tweet.
- TimeStamp: Filtra si en un tweet se ejecutó timestamp o no.

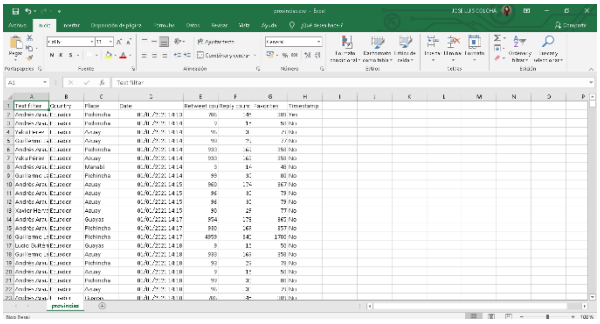

Figura 32 Bases de datos en Excel

Luego de haber realizado la debida limpieza de datos se exporto al gestor de visualización PowerBI para proceder a analizar los datos recopilados en cada una de las provincias y saber que tan frecuente se twittea acerca de algún tema político o ya sea brindar apoyo algún candidato a la presidencia.

	Candidates	Timestamp	Country	Province
1	Andrés Arauz	Yes	Ecuador	Pichincha
2	Andrés Arauz	No	Ecuador	Pichincha
3	Yaku Pérez	No	Ecuador	Azuay
4	Guillermo Lasso	No	Ecuador	Azuay
5	Andrés Arauz	No	Ecuador	Pichincha
6	Yaku Pérez	No	Ecuador	Azuay
7	Andrés Arauz	No	Ecuador	Manabí
8	Guillermo Lasso	No	Ecuador	Pichincha
9	Andrés Arauz	No	Ecuador	Azuay
10	Andrés Arauz	No	Ecuador	Azuay
11	Andrés Arauz	No	Ecuador	Azuay
12	Xavier Hervas	No	Ecuador	Azuay
13	Andrés Arauz	No	Ecuador	Guayas
14	Andrés Arauz	No	Ecuador	Pichincha

Figura 33 Bases de datos en PowerBI

C. Juegos en línea por países.

Con el primer dataset se analiza el número de jugadores por millones que existen en cada país y que juegan Call of Duty WarZone.
Con el segundo dataset se puede analizar:

- Country: país de origen
- Population: población por millones en el país en el campo Country.
- Lenguages: el idioma que hablan los pobladores del país.
- Gamers: la población por millones de jugadores en línea.
- USD: dinero generado por la población jugadora.

Country	USD	Gamers	Languages	Rank	Population
Australia	1293	12	English	14	25
Brazil	1452	76	Portuguese	13	211
Canada	2399	21	English, French	8	37
China	344	620	Chinese	1	1415
France	3366	33	French	7	65
Germany	4989	44	German	5	82
India	1105	264	English, Hindi, other 21 local languages	16	1354
Indonesia	1084	44	Indonesian	17	267
Italy	2168	26	Italian	10	59
Japan	17715	68	Japanese	3	127
Mexico	1577	56	Spanish	12	131
Netherlands	646	7	Dutch	20	17
Republic of Korea	5764	29	Korean	4	51
Russian Federation	1657	65	Russian	11	144
Spain	2202	25	Spanish	9	46
Taiwan	1231	15	Chinese	15	24

Figura 34 Bases de datos rating-games.csv

D. Tema definido por el estudiante

El data set escogido está limpio por lo que no hubo la necesidad de limpiar los datos. Para una mejor comprensión se transfirió el data set a phpMyadmin donde podemos observar cada uno de los campos con mayor determinación y para realizar las visualizaciones se utilizó la herramienta de Tableu.

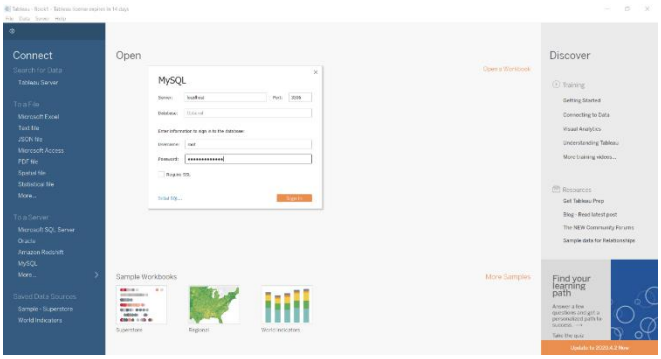

Figura 35 Conexión de MySQL phpMyadmin.a Tableu.

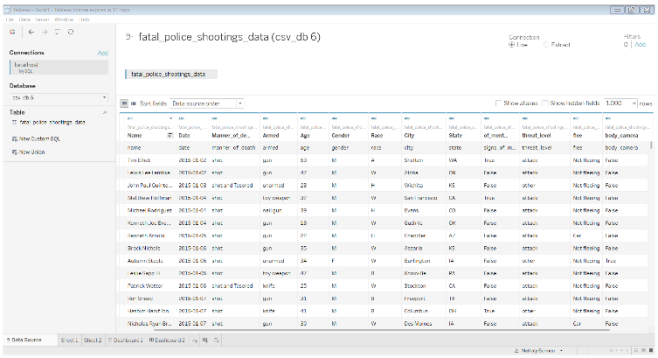

Figura 36 Tabla recopilada de la base de datos de phpMyadmin.

E. Eventos o noticias mundiales

1) Facebook

Los datos obtenidos por la red social Facebook, de los cuales solo se analizó la columna *texto* con la librería de Python textBlob, para obtener las noticias positivas, negativas.

```python
import pandas
from textblob import TextBlob
import seaborn as sns
#%

df = pandas.read_csv('bbc_news_mundo.csv')

a = list(df['texto'])

b = ' '.join(str(e) for e in a)


#%
texto=b.split('\n\n')

list_palabras=[]

for index in range(len(texto)):
    if index == 0:
        list_palabras.append(texto[index])
    else:
        try:
            list_palabras.append(texto[index].split('\n')[1])
        except:
            list_palabras.append(texto[index])

#%
df = pandas.DataFrame(columns=['texto', 'polaridad'])
df['texto'] = list_palabras
df['polaridad'] = ['' for _ in range(len(df))]


#%%
for i in range(len(df)):
    if TextBlob(df.at[i, 'texto']).sentiment.polarity < 0:
        df.at[i, 'polaridad'] = "Negativo"
    elif TextBlob(df.at[i, 'texto']).sentiment.polarity > 0:
        df.at[i, 'polaridad'] = "Positivo"
    else:
        df.at[i, 'polaridad'] = "Neutral"


#%%
df.to_csv('bbc_news_mundo_polaridad.csv')
```

Figura 37 Script para analizar sentimientos

2) Kaggle

Lo datos obtenidos de la página Kaggle, relacionados con las vacunas contra el Covid-19 se los almaceno en SQLite en una base de datos llamada “países_vacunados” y una tabla llamada “vacunados”.

```
Símbolo del sistema - sqlite3

C:\Users\tefo_\Documents\__EPN\__datos>sqlite3
SQLite version 3.34.1 2021-01-20 14:10:07
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .mode csv
sqlite> .import country_vaccinations.csv vacunados
sqlite> .shema vacunados
Error: unknown command or invalid arguments:  "shema". Enter ".help" for help
sqlite> .schema vacunados
CREATE TABLE IF NOT EXISTS "vacunados"(
  "country" TEXT,
  "iso_code" TEXT,
  "date" TEXT,
  "total_vaccinations" TEXT,
  "people_vaccinated" TEXT,
  "people_fully_vaccinated" TEXT,
  "daily_vaccinations_raw" TEXT,
  "daily_vaccinations" TEXT,
  "total_vaccinations_per_hundred" TEXT,
  "people_vaccinated_per_hundred" TEXT,
  "people_fully_vaccinated_per_hundred" TEXT,
  "daily_vaccinations_per_million" TEXT,
  "vaccines" TEXT,
  "source_name" TEXT,
  "source_website" TEXT
);
sqlite> .save paises_vacunados.db
sqlite>
```

Figura 38 Convertir de csv a db con SQLite

Posteriormente se utilizó *ELK* para generar las visualizaciones finales.

```
paises_vacunados.conf: Bloc de notas
Archivo  Edición  Formato  Ver  Ayuda
input {
        jdbc {
                jdbc_connection_string => "jdbc:sqlite:C:\Users\tefo_
\Documents\__EPN\__datos\paises_vacunados.db"
                jdbc_user => "root"
                jdbc_password => ""
                jdbc_driver_library => "C:\Users\tefo_\Documents\sqlite
conector\sqlite-jdbc-3.34.0.jar"
                jdbc_driver_class => "org.sqlite.JDBC"
                statement => "SELECT * FROM vacunados"
        }
}
output {
        stdout { codec => json_lines }
        elasticsearch {
                "hosts" => "localhost:9200"
                "index" => "paises_vacunados"
                "document_type" => "data"
        }
}
```

Figura 19 jdbc logstash

3) Statista

Los datos obtenidos de la página *Statista* se los analizo manualmente, ya que había celdas que interferían al momento de realizar las gráficas.

IX. VISUALIZACIÓN DE INFORMACIÓN

A. Pulso político en 20 ciudades principales de Ecuador

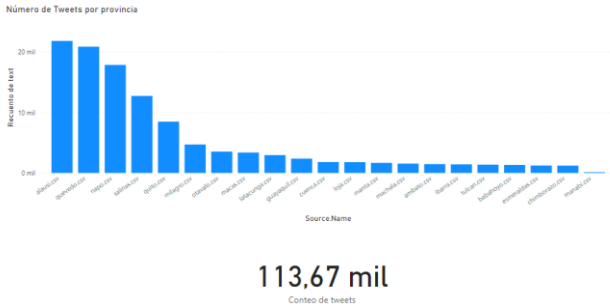

Figura 40 Número de tweets por provincia.

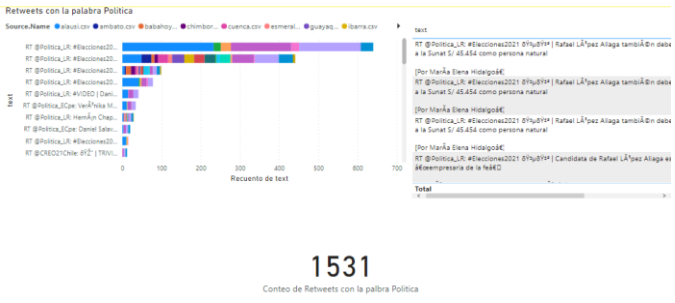

Figura 41 Retweeets con la palabra *Política*.

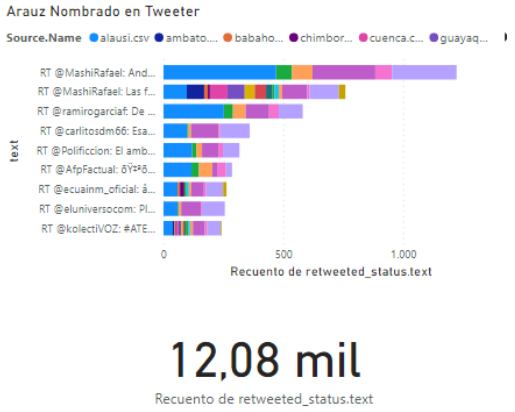

Figura 42 Conteo de Arauz nombrado en tweeter.

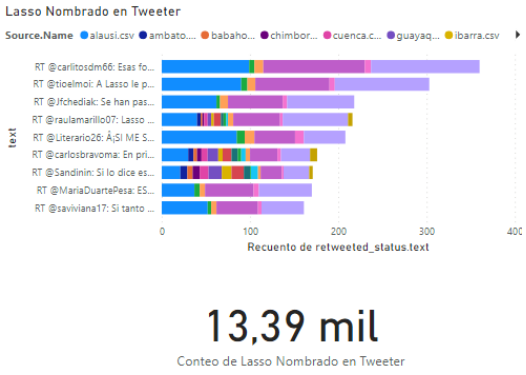

Figura 43 Conteo de Lasso nombrado en tweeter.

B. Pulso político por provincias en Ecuador

.

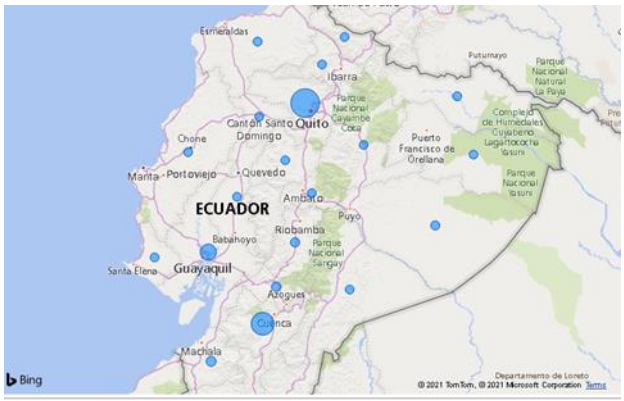

Figura 44 Total de Retweets por provincial

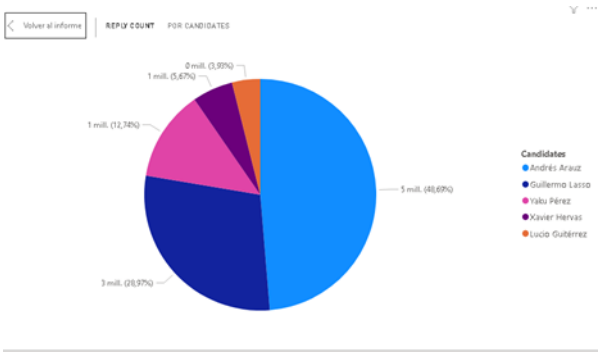

Figura 45 Total de tweets favoritos en todas las provincias por candidato

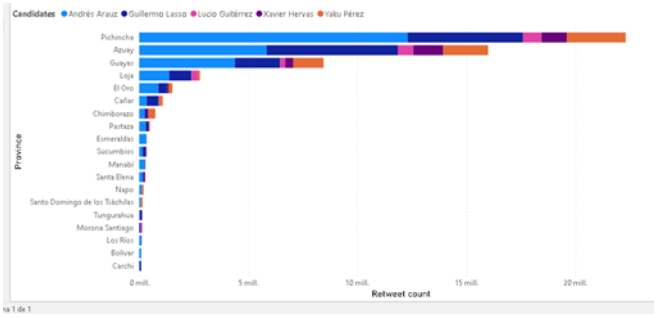

Figura 46 Total de tweets por provincia y candidato presidencial

C. Juegos en línea por países.

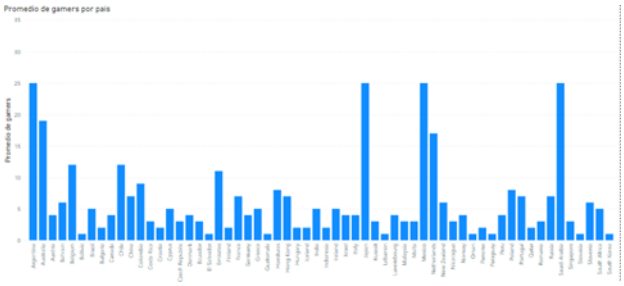

Figura 47 warzone.csv grafica de jugadores por país.

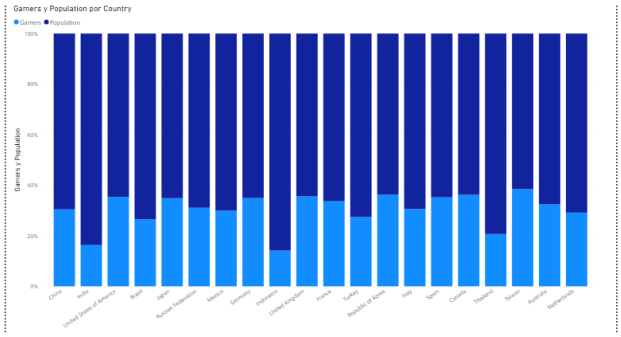

Figura 48 rating-gamers.csv población vs gamer.

Figura 49 rating-gamers.csv concentración poblacional gamer.

D. Tema definido por el estudiante

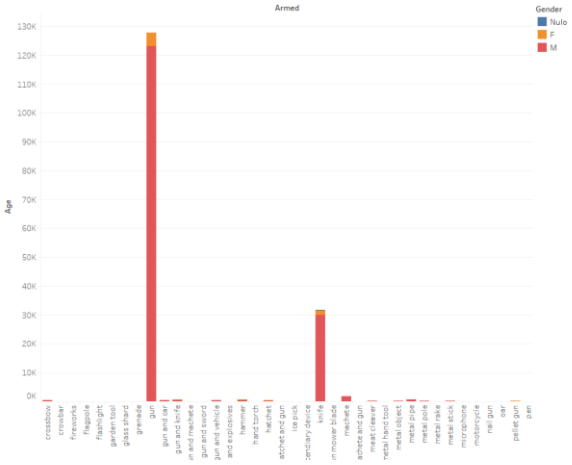

Figura 50 Armas que más utilizan en EE. UU. para tener enfrentamientos con la policía.

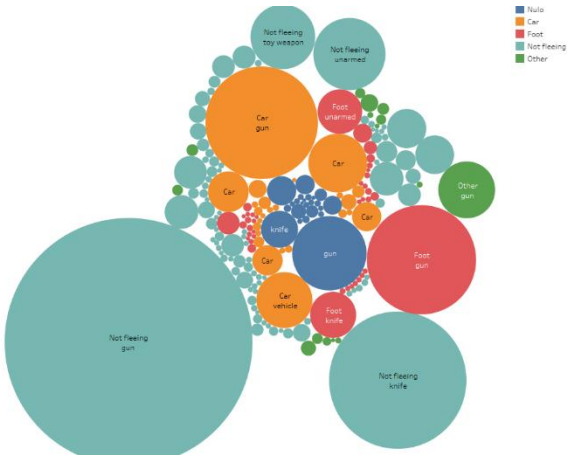

Figura 51 Acciones que toman los policías para defenderse según el arma.

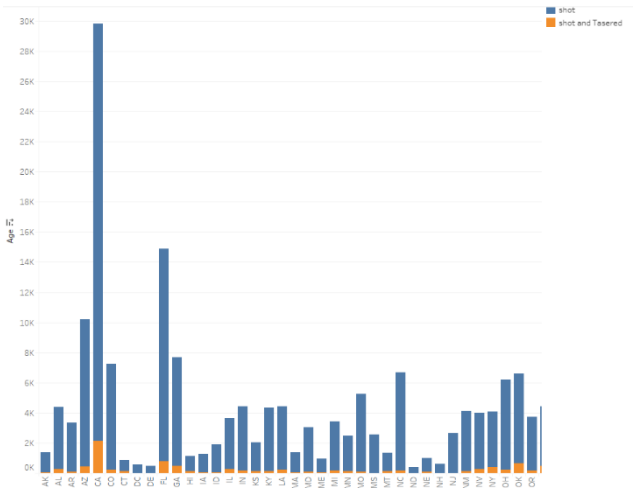

Figura 52 Armas que utilizan los delincuentes según el estado y la edad.

E. Eventos o noticias mundiales

1) Facebook

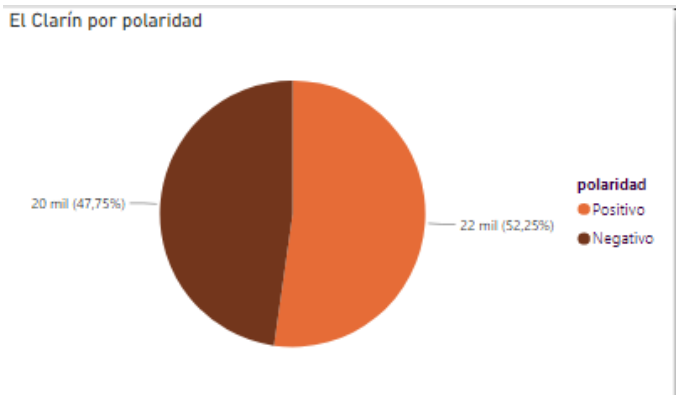

Figura 53 El Clarín

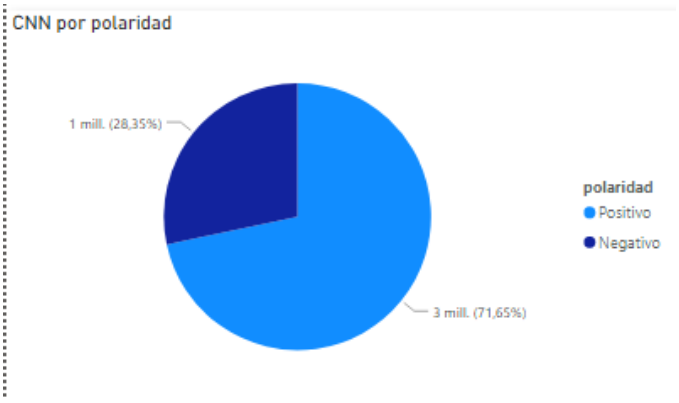

Figura 54 CNN en Español

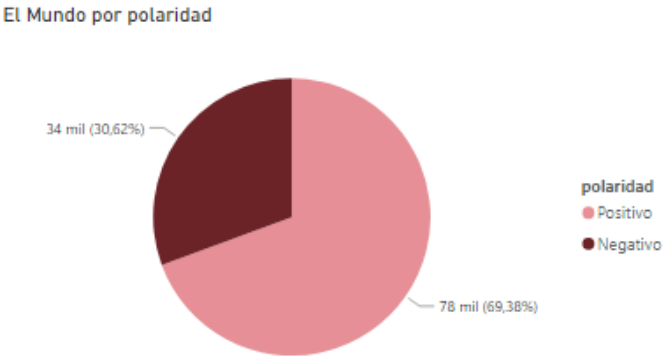

Figura 20 El Mundo

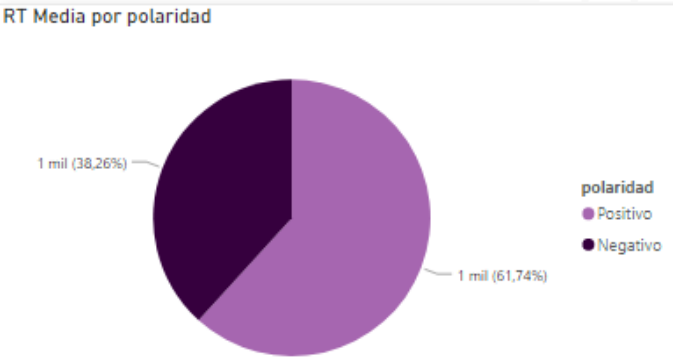

Figura 55 RT Media

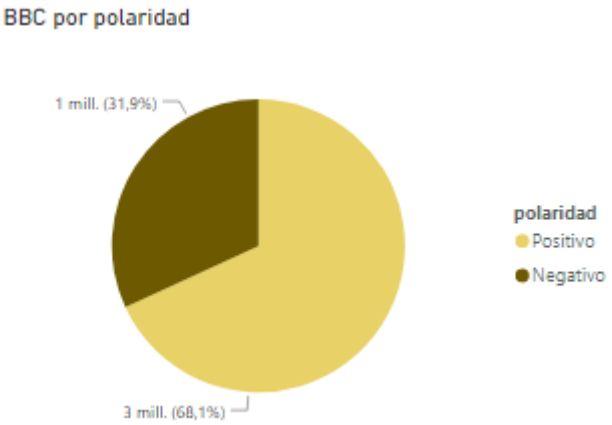

Figura 56 BBC news Mundo

2) Kaggle

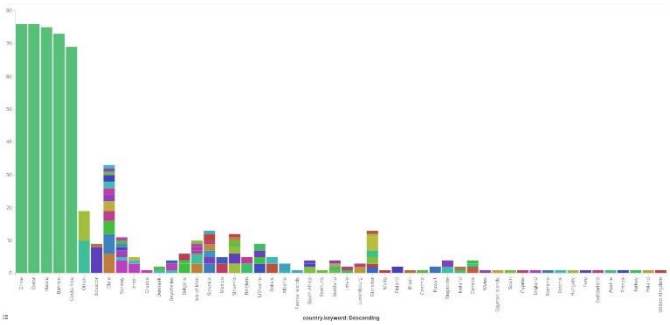

Figura 57 Personas completamente vacunadas por país

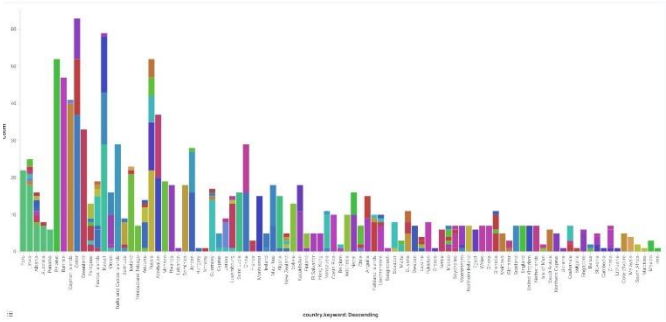

Figura 58 Personas vacunadas diariamente por país

3) Statista

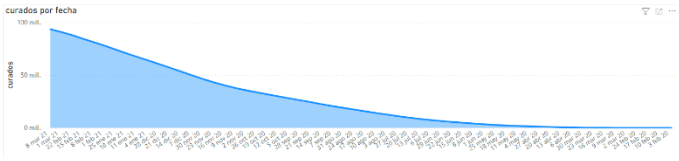

Figura 59 Número de personas recuperadas

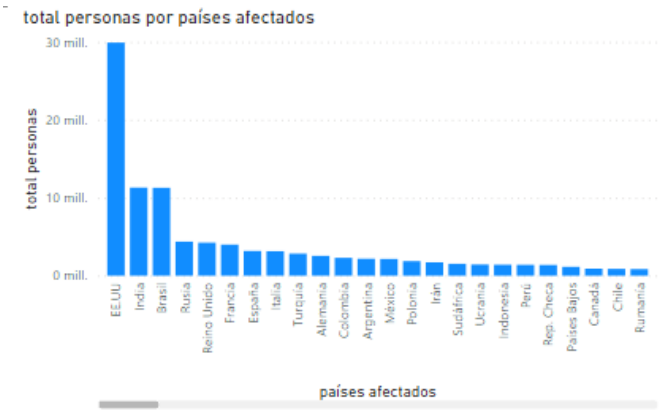

Figura 60 Número de personas confirmadas con Covid-19 por país

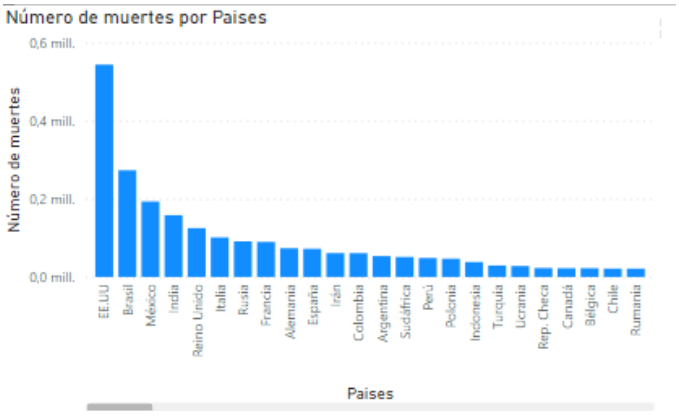

Figura 61 Número de personas fallecidas confirmadas con Covid-19 por país

X. Resultados obtenidos

A. Pulso político en 20 ciudades principales de Ecuador

1) Retweets por ciudad.

En Alausí se hizo un seguimiento y se observó que tienen más retweets en el ámbito político.

2) Conteo de tweets.

Se analiza que el candidato Lasso gana en el aspecto de conteo de tweets ya que tiene una cifra más alta que Arauz.

3) Conteo sobre pulso político.

Se puede observar que muchas personas hablan de política en sus cuentas ya que hacen muy seguido retweets sobre la situación del país.

B. Pulso político por provincias en Ecuador

1) Total, de retweets por provincias.

Los resultados obtenidos en cada una de las provincias acerca de que tan activos son los ecuatorianos en la red social Twitter dio como resultados que las primeras 3 provincias que más se habla de política son: Pichincha, Azuay y Guayas por otro lado las provincias en las cuales no se hablan acerca de política ni se twittea son: Carchi. Cotopaxi y Galápagos.

2) Total de tweets favoritos en todas las provincias por candidato.

Los resultados obtenidos dieron como resultados que en Andrés Arauz es el candidato presidencial más nombrado en la red social Twitter y agregado a la sección de favoritos, seguido por Guillermo Lasso y Yaku Pérez además Lucio Gutiérrez y Xavier Hervas también son nombrados, aunque en menor cantidad.

3) Total de tweets por provincia y candidato presidencial.

Con los distintos filtros de palabras que se hicieron uso se pudo apreciar que tan nombrados son los candidatos a la presidencia en la red social Twitter, dividiendo los resultados por provincias se comprobó que en la provincia de Pichincha el nombre del candidato Arauz es muy utilizado para twittear seguido por el candidato Lasso y Pérez, también se notó que Galápagos es la provincia con menos tweets.

C. Juegos en línea por países.

1) Warzone.csv grafica de jugadores por país

Se puede analizar la mayor concentración de los jugadores de warzone en cada país del dataset, además ver qué países tienen menor aceptación hacia el juego.

2) Rating-gamers.csv población vs gamer

Se puede analizar la diferencia que existe entre la población general y gamer que existen en los diferentes países, las franjas azules permiten ver la población total y las franjas turquesa son la población de gamers en ese país.

3) Rating-gamers.csv concentración poblacional gamer

Se pude analizar con uso de un mapa la concentración de población gamer y los ingresos económicos al país.

D. Tema definido por el estudiante

1) Armas que utilizan más seguido los delincuentes en EE. UU.

En la gráfica se observó que los delincuentes utilizan más pistolas y cuchillos para poder enfrentarse con los policías.

2) Género

Se puede observar que hay más delincuentes hombres en cada estado de EE. UU. y muy poco porcentaje son mujeres.

3) Muertes de policías.

El arma con el que más se enfrentan los policías según las visualizaciones hechas en Tableu es la pistola ya que es la principal causa de muerte.

E. Eventos o noticias mundiales

1) Portales de Noticias

Se realizaron varias gráficas, en las que se puede observar que existen mas noticias positivas que negativas en el mundo de los 5 portales escogidos de Facebook.

2) Personas completamente vacunadas por país

Se realizó la gráfica de las personas completamente vacunadas en el mundo por países. Dichos datos dan como resultado a China y Qatar como los países con más personas vacunadas completamente.

3) Personas vacunadas diariamente por país

Los resultados obtenidos de los países que son vacunados diariamente. Dando como primer lugar al país de Qatar, con más personas vacunadas diariamente.

4) Número de personas recuperadas

A fecha de 12 de marzo de 2021, se habían registrado en el mundo más de 93 millones de casos de pacientes curados de

COVID-19, causada por el coronavirus (SARS-CoV-2). El brote, originado en la ciudad China de Wuhan, se había extendido a países de los cinco continentes.

5) Número de personas confirmadas con Covid-19 por país

Los resultados muestran los países afectados por el coronavirus de Wuhan (SARS-CoV-2) en función del número de casos confirmados a 12 de marzo de 2021. China, país en el que se cree que se originó el brote, ha confirmado hasta el momento algo más de 89.800 casos de COVID-19. Sin embargo, la clasificación la encabeza Estados Unidos, con alrededor de 30 millones de positivos confirmados. En cuanto al Viejo Continente, los 47 países europeos han registrado hasta el momento infectados entre sus ciudadanos, destacando España , Rusia, el Reino Unido, Italia y Alemania.

6) Número de personas fallecidas confirmadas con Covid-19 por país

Los resultdos muestran el número de muertes causadas por el SARS-CoV-2, conocido popularmente como el coronavirus de Wuhan, a nivel mundial a fecha de 12 de marzo de 2021. Hasta ese día se habían contabilizado aproximadamente 2,64 millones de muertes debidas al virus, de las cuales 4.636 ocurrieron en China, lugar en el que se originó el virus. Sin embargo, el país asiático ya no es el territorio donde el nuevo coronavirus se ha cobrado más vidas. Estados Unidos encabeza la clasificación al superar los 543.700 decesos, seguido de Brasil con alrededor de 273.125. A 12 de marzo de 2021, había más de 119 millones de casos confirmados de COVID-19 en todo el mundo.

XI. Conclusiones y Recomendaciones

La recopilación de datos es un proceso de mucha utilidad cuando se requiere saber datos o información relevante acerca de algún tema en boga nacional o internacionalmente, de esta forma se hace uso de esta información para dar a conocer datos estadísticos.

Gracias al análisis de las gráficas se puede analizar que la población oriental tiende a tener menor índice de población gamer en plataformas online, y que los países que más generan ingresos económicos son los países latinoamericanos.

Las herramientas como Power BI y Tableu ayuda a tener una mejor visualización de los datos recolectados ya que facilita el entendimiento y el aprendizaje.

XII. Desafíos y problemas encontrados

Un problema a resaltar es la recopilación de datos por geolocalización en la red social Twitter, puesto que en ciertas provincias no se twittea o no cuentan con una conectividad a internet y no se pudo recolectar la cantidad de datos propuestos.

Otro problema al realizar una importación directamente a SQL Server es que de vez en cuando los dataset que se usan se crean dentro de las veces como caracteres tipo String lo que limita al momento de querer crear estadísticas en el análisis.

XIII. Enlace de github del proyecto

References

[1] Available: http://www.halcyon.com/pub/journals/21ps03-vidmar
[2] Apache. (2020). Logstash [Online]. Available: https://www.elastic.co/es/logstash
[3] A.Robledano.(2020). ¿Qué es MySQL? [Online]. Available: https://openwebinars.net/blog/que-es-mysql/
[4] Apache. (2020). Kibana [Online]. Available: https://www.elastic.co/es/what-is/kibana
[5] M.Alvarez. (2003). ¿Qué es Python? [Online]. Available: https://desarrolloweb.com/articulos/1325.php
[6] IONOS. (2020). CouchDB [Online]. Available: https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/presentacion-de-couchdb/
[7] Elastic. (2021). Elasticsearch [Online]. Available: https://www.elastic.co/es/what-is/elasticsearch
[8] Davinci. (2020). ¿Qué es logstash? [Online]. Available: https://www.davincigroup.es/que-es-logstash-ejemplo-practico-de-uso/
[9] Tableu. (2021). ¿Qué es Tableu? [Online]. Available: https://www.tableau.com/es-es/why-tableau/what-is-tableau
[10] M.Parada. (2019). ¿Qué es SQL server? [Online]. Available: https://openwebinars.net/blog/que-es-sql-server/