Informe Proyecto final

Esteban Alejandro Rios Pérez

[esteban.rios@epn.edu.ec](mailto:esteban.rios@epn.edu.ec)

Freddy Geovanny Valverde Gallardo

[freddy.valverde@epn.edu.ec](mailto:freddy.valverde@epn.edu.ec)

Marvin Aníbal Zambrano Giler

[marvin.anibal@epn.edu.ec](mailto:marvin.anibal@epn.edu.ec)

Miguel Ángel Jurado Cedeño

[miguel.jurado@epn.edu.ec](mailto:miguel.jurado@epn.edu.ec)

1. **Definición de caso de estudio**

Los casos de estudio utilizados para el proyecto son los siguientes:

* 1. **Pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados.**

Como el país está ingresando en los procesos electorales y como en todos los años siempre hay distintos postulantes al puesto de mandatario y pues el objetivo es recopilar datos para ser analizados y determinar cómo es el movimiento sobre el pulso político en las 20 ciudades principales de Ecuador.

* 1. **Pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados.**

Por el mismo motivo que el país está ingresando en los procesos electorales y como en todos los años siempre hay distintos postulantes al puesto de mandatario. Entonces el objetivo es recopilar datos para ser analizados y determinar cómo es el movimiento sobre el pulso político o elecciones en las provincias de Ecuador.

* 1. **Juegos en línea por países.**

Se recopilaron datos relacionados con el tema de Juegos en línea para realizar un análisis sobre qué categoría, plataforma, editor del juego tuvo más ventas desde el año 1980 hasta 2017. ​

* 1. **Eventos o noticias mundiales.**

Para el caso de las noticias mundiales su decidió recopilar datos relacionados con el tema del Covid-19 alrededor del mundo. El propósito de esta recopilación es tener una representación más clara de cómo ha avanzado esta problemática a nivel mundial. Para poder realizar el análisis se utilizará los datasets que Kaggle pueda proporcionar sobre el tema y de ahí se realizará un análisis con graficas o visualizaciones que permitan explicar todo de mejor manera.

1. **Objetivos**
   1. **Objetivo General**

Analizar los datos recopilados de diferentes fuentes, utilizando diferentes herramientas para la recolección, ilustración, e interpretación de estos.

* 1. **Objetivos específicos**

Definir una arquitectura que indique el flujo que lleva la investigación desde la recolección de datos hasta su interpretación.

Recolectar los datos de diferentes fuentes y almacenarlos en distintas bases de datos SQL y No SQL usando distintos scripts.

Concentrar los datos en una herramienta de colección de datos masivos.

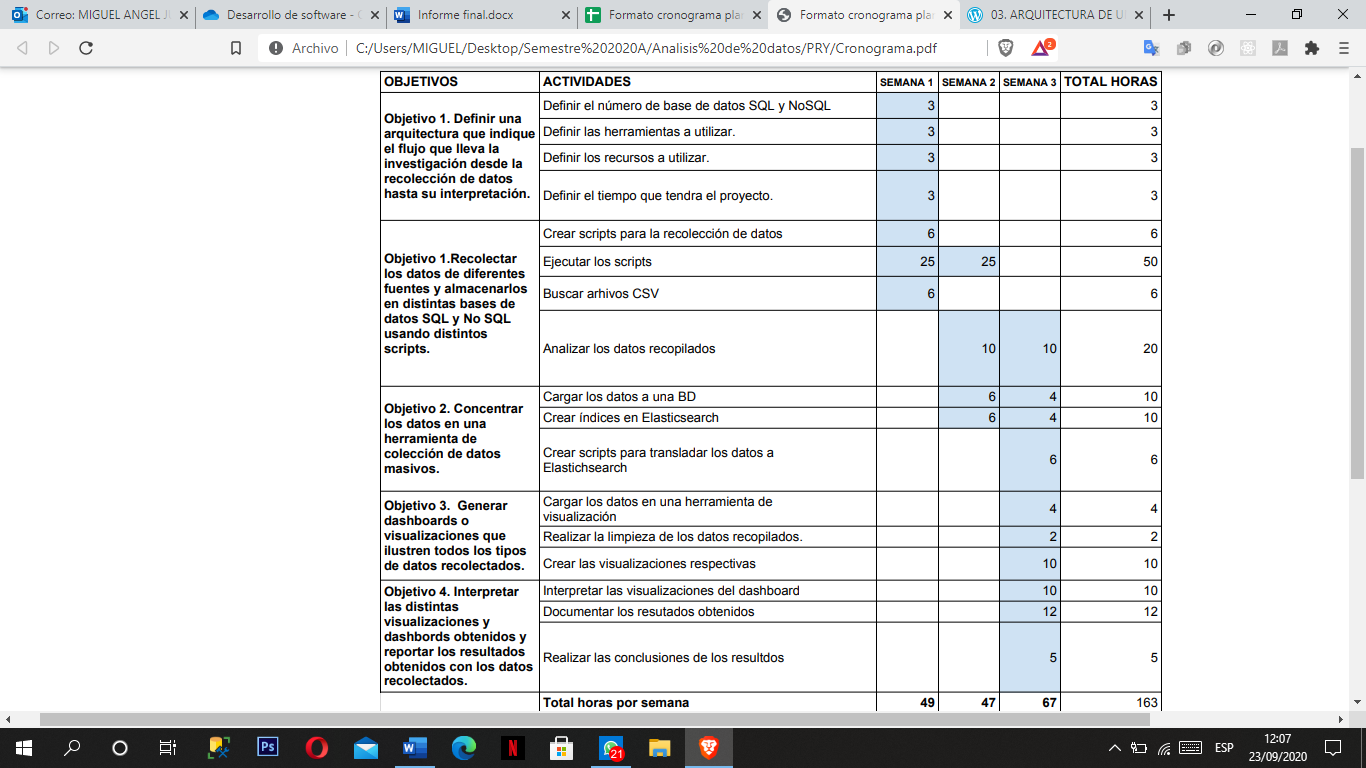
Generar dashboards o visualizaciones que ilustren todos los tipos de datos recolectados.

Interpretar las distintas visualizaciones y dashboards obtenidos y reportar los resultados obtenidos con los datos recolectados.

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

El equipo de trabajo está conformado de 4 personas: Miguel Jurado, Esteban Rios, Freddy Valverde y Marvin Zambrano; estudiantes de 4to semestre de la carrera de Tecnología Superior en Desarrollo de Software de la Escuela Politécnica Nacional. Las actividades realizadas por cada miembro son:

|  |  |
| --- | --- |
| **Nombre** | **Tareas** |
| Miguel Jurado | * Recopilar datos sobre videojuegos utilizando Kaggle para obtener la información. * Guardar dicha información en una Base de Datos SQL, en este caso MySQL. * Realizar las visualizaciones con la información obtenida utilizando PowerBI. |
| Esteban Rios | * Recopilar datos relacionados con el Covid-19 utilizando Kaggle para obtener la información. * Guardar dicha información en una Base de Datos SQL, en este caso MySQL. * Realizar las visualizaciones con la información obtenida utilizando PowerBI. |
| Freddy Valverde | * Recopilar datos sobre el pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados, a través de Twitter utilizando un script de Python. * Guardar la información obtenida en un Base de Datos NoSQL, en este caso CouchDB. * Exportar esta información a ElasticSearch para crear un nuevo índice y poder realizar las visualizaciones en tiempo real utilizando Kibana. * A partir del tema de preferencia elegido que es “Test de Personalidad”, se realizara las respectivas tareas mencionadas anteriormente. |
| Marvin Zambrano | * Recopilar datos sobre el pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados, a través de Twitter utilizando un script de Python. * Guardar la información obtenida en un Base de Datos NoSQL, en este caso MongoDB. * Exportar esta información a ElasticSearch para crear un nuevo índice y poder realizar las visualizaciones en tiempo real utilizando Kibana. |

1. **Cronograma de actividades**

**Figura 1: Cronograma de actividades**

1. **Recursos y herramientas utilizadas**

Para el desarrollo del proyecto se utilizaron recursos como scripts de Python, los cuales tienen como funcionalidad obtener datos desde Twitter o Facebook para llevarlos a una base de datos No SQL (CouchDB y MongoDB). Además, se utilizaron otros recursos como CSV’s los cuales fueron obtenidos desde Kaggle. Kaggle es plataforma gratuita que posee a disposición de los usuarios una serie de problemas para solucionar con temáticas como la ciencia de datos, el análisis predictivo y machine learning [1].

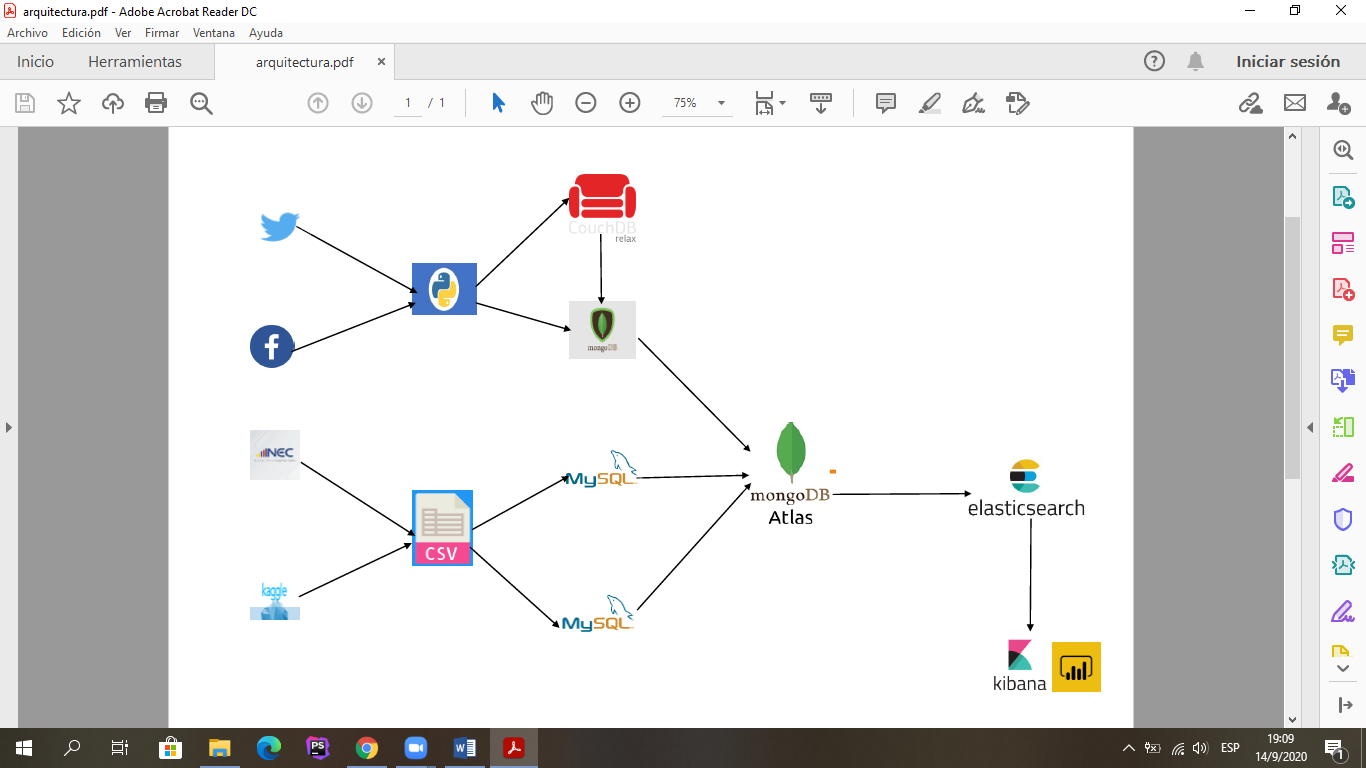
Por otro lado, para la facilidad del manejo de los datos se utilizaron herramientas tales como Excel para editar algunos de los CSV’s, MySQL para almacenar los datos estructurados, CouchDB y MongoDB para almacenar datos no estructurados, Mongo Atlas para unificar las bases de datos, Kibana para hacer visualizaciones en tiempo real y PowerBI para hacer visualizaciones estáticas.

Finalmente, se utilizó ELK que son las siglas de tres proyectos *open source:* Elasticsearch, Logstash y Kibana. Elasticsearch es un motor de búsqueda y analítica. Logstash es un *pipeline* de procesamiento de datos del lado del servidor que ingesta datos de una multitud de fuentes simultáneamente, los transforma y luego los envía a un “escondite”, como Elasticsearch. Kibana permite a los usuarios visualizar los datos en cuadros y gráficos con Elasticsearch. Adicional a esto, también se utilizó cerebro para la administración de índices en elasticsearch.

1. **Arquitectura de la solución**

Para el proyecto se definió una arquitectura para recopilar los datos, compuesta de dos partes. La primera parte consta en la recolección de datos hacia una base local (MongoDB, CouchDB y MySQL), estas bases contendrán datos de las principales fuentes estudiadas a lo largo del curso (Twitter, CSV, JSON, entre otras). La segunda parte consiste en alojar los datos que tenemos individualmente por los casos de estudios presentados a un clúster o base concentradora (MongoDB Atlas) para así llevar los datos a Elastic Search para su indexación y posteriormente realizar las visualizaciones estáticas y en tiempo real con Kibana y PowerBI.

A continuación, la figura muestra el diseño de nuestra arquitectura la cual como se detalló anteriormente nos permitirá luego de recopilar la información analizarla y plasmarla en visualizaciones.

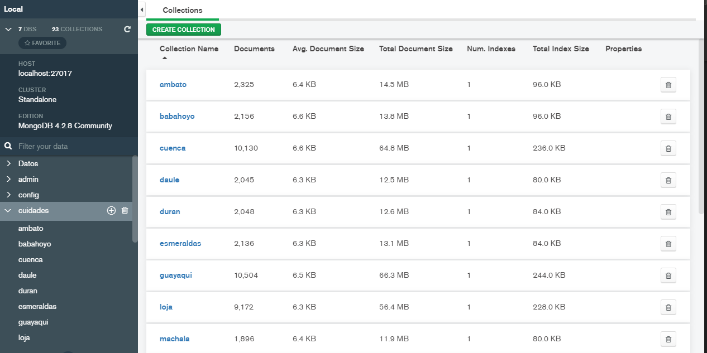


**Figura 2** Arquitectura de Solución

1. **Extracción de datos**
   1. **Pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados.**

Los datos se recopilaron por el script de “Tweets Cuidades.py”, los que se almacenarían en la base de datos NoSQL MongoDB. En el script constan las credenciales del API de Twitter y en que dataset se van a almacenar, adicional se definen las 20 colecciones que tendrán los datos, el nombre de cada colección es acorde a la cuidad. Además, se aplica un filtro por palabras con las palabras claves de: *Ecuador, Nombre\_Cuidad, Pulso Politico, Candidatos y Elecciones.*

Las principales ciudades del país tomadas en cuenta son: Quito, Guayaquil, Manta, Portoviejo, Babahoyo, Cuenca, Loja, Samborondón, Esmeraldas, Durán, Daule, Machala, Ambato, Tena, Salinas, Santo Domingo, Quevedo, Villamil, Otavalo y Tulcán como se muestra en la Figura 2.



**Figura 3.** Datos recolectados en MongoDB

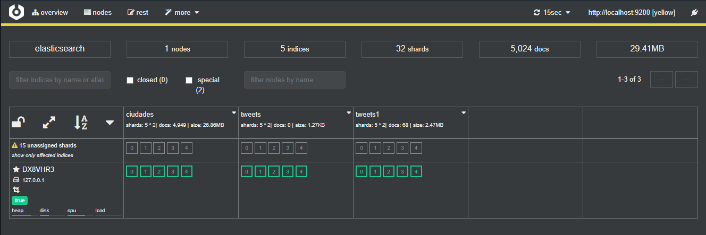
En Mongo DB constaran los datos recolectados que son un total aproximado de 100mil datos, los cuales se concentraran en un Cluster de MongoDB Atlas con una base de datos común y un dataset por cada tema que se haya analizado, desde aquí pasarán a CouchDB por el problema que se presenta con Logstash y MongoDB, como se muestra en la Figura 4.



**Figura 4:** Datos en CouchDB

Desde aquí se pasará a ElasticSearch mediante Logstash con el archivo de configuración *couchdb.conf.* El índice *ciudades* creado estará mapeado según características específicas de cada tweet entre los principales están la fecha y la geolocalización como se muestra en la Figura 4.

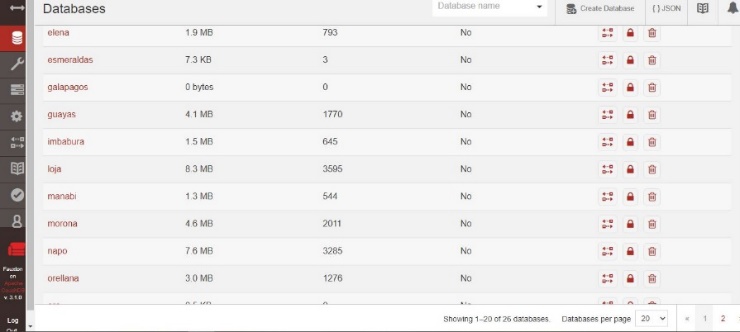
Los datos recopilados serán analizados mediante Kibana.



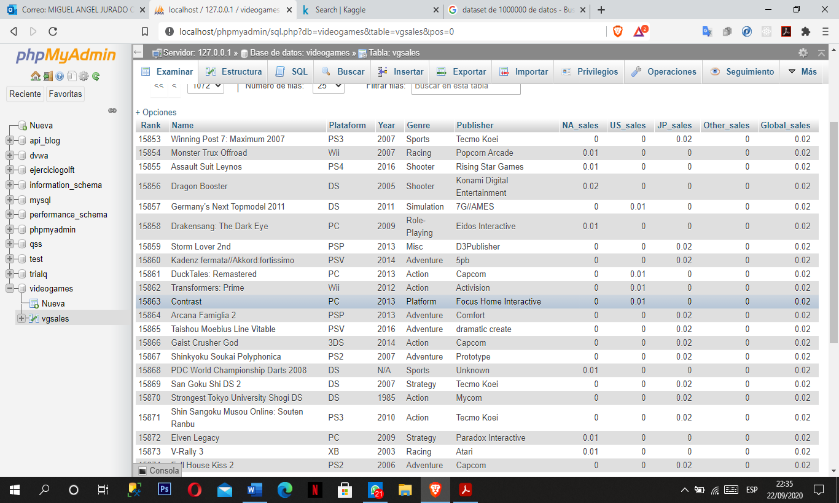
**Figura 5.** Índice Mapeado y con Datos

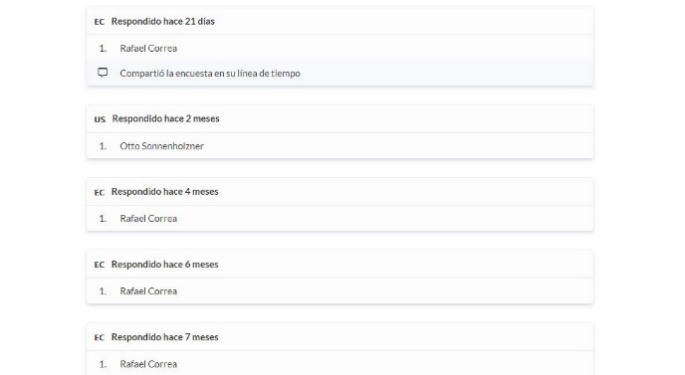
* 1. **Pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados.**

La extracción de datos de este tema tiene dos orígenes. El primero es Twitter donde hicimos uso del script “Ecuador” el cual exportará estos datos a una base de datos en couchDB dicha base de datos tuvo el nombre de cada provincia. El script igualmente para cada extracción de datos tuvo modificaciones en los filtros por coordenadas y los filtros por palabras. En los filtros de palabras fue común buscar “Elecciones Ecuador 2021, Nombre\_Candidato, provincia, Nombre partido político”. Existieron provincias donde no se obtuvieron resultados favorables. Se recopilaron alrededor de 60000 documentos.



**Figura 6:** Bases de datos en CouchDB

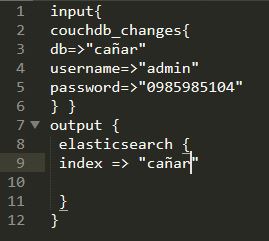
La segunda fuente fue una encuesta realizada por el diario El Mercurio de la ciudad de Cuenca, la cual fue lanzada a comienzos del año y en la cual pregunta al público lector ¿Si las elecciones fueran hoy por quién votaría usted para presidente de la república? Además, el dataset proporcionado contaba con datos como el país de origen de encuestado, la región o estado. Esta encuesta fue realizada por aproximadamente 4000 personas.



**Figura 7:** Dataset de videojuegos en MySQL

**Figura 8:** Encuesta de diario El Mercurio

Todos estos datos fueron llevados a elasticSearch para su posterior análisis, esto se pudo realizar mediante logstash, utilizando el archivo “csv.conf” y “couchdb.conf” los cuales realizan el respectivo traslado de los datos para los diferentes archivos y bases obtenidas.



**Figura 9:** Estructura del archivo couchdb.conf

Finalmente, todos los índices obtenidos desde couchDB fueron pasados a un índice general llamado Ecuador el cual contiene todos los datos de Twitter, donde de los 60000 después de mapearlos quedaron alrededor de 30000. Ya con los índices creados se procedió a realizar las visualizaciones en Kibana.

* 1. **Juegos en línea por países.**

Para la extracción de datos de este caso de uso se utilizó un csv obtenido de la plataforma Kaggle. El dataset contiene alrededor de 16000 tuplas y 11 columnas. Luego de obtener el *dataset* de la plataforma Kaggle, y siguiendo la arquitectura planteada en este proyecto, se cargaron los datos en la base de datos MySQL, como se muestra en la Figura 6, para un posterior análisis de los datos en la herramienta PowerBI.

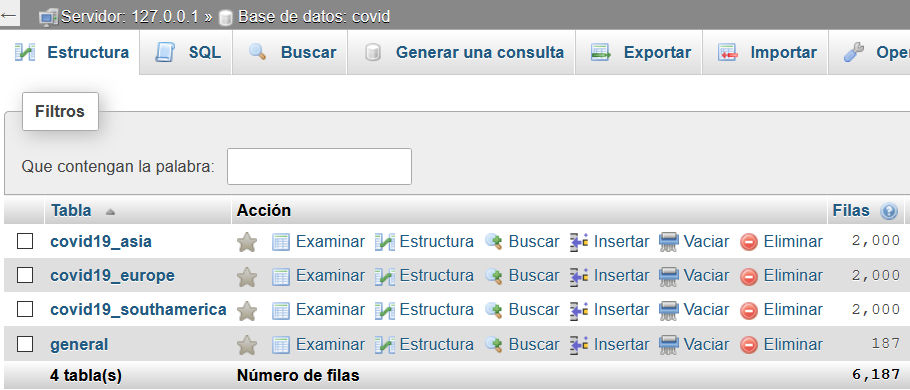
Cabe mencionar que, como se está manejando muchos datos, se debe configurar el tiempo ejecución para que se puedan cargar todos los datos.

Después de tener cargados los datos en MySQL, se procedió a llevar los datos a Mongo Atlas para tener todas las bases de datos unificadas tal y como se plantea en la arquitectura del proyecto.

* 1. **Eventos o noticias mundiales.**

Como ya se mencionó en un principio la fuente de obtención de datos es la página de Kaggle. Para el desarrollo del proyecto se buscó información relacionada con el tema del Covid-19 (para más información visitar el link: <https://www.kaggle.com/search?q=covid-19+date%3A90>), ya que es un tema relevante y es importante conocer como se ha venido evolucionando.

Los datos obtenidos de los datasets se los almacenó en MySQL en una base de datos con el nombre de “covid” como se muestra en la **Figura 1**.



**Figura 10:** Base de Datos covid

En las tablas *covid19\_asia, covid19\_europe y covid19\_southamerica* se encuentra información respecto a cada continente mientras que la tabla *general* muestra información del covid a nivel mundial.

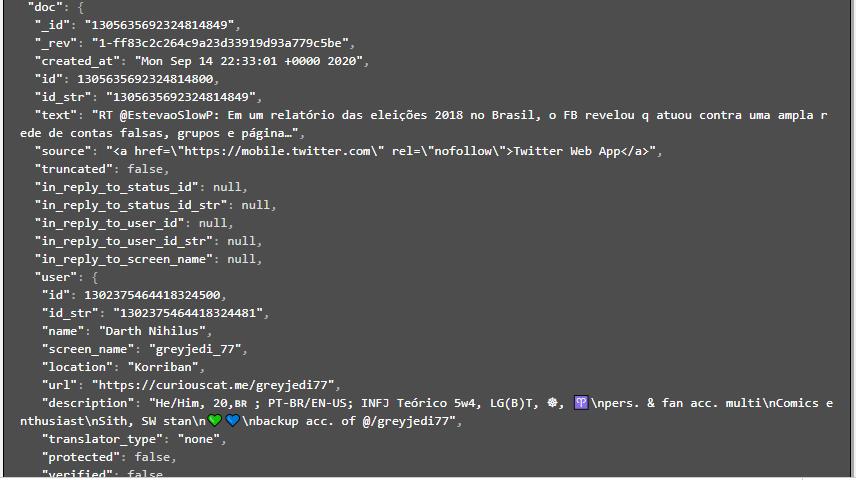
1. **Análisis de información**
   1. **Pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados.**

Se crea una index pattern en kibana para la visualización de los datos correspondientes, usando como filtro las fechas que contiene cada uno de los datos que fueron mapeados en elasticsearch.

Los campos en los que consta cada tweet son:

* Un identificador único (\_id).
* Fecha de creación (created\_at).
* Texto del tweet (text).
* Referencia externa (source)
* Los datos del usuario que lo publicó (user) dentro del cual se ubican:
  + Identificador único (id)
  + Nombre del usuario (name)
  + Nickname que tiene en la aplicación (screen\_name).
  + Ubicación (location).
  + Enlaces del usuario (url).
  + Descripción sobre el usuario (descripción).
  + Número de seguidores (followers\_count).
  + Número de amigos (friends\_count).
  + Cantidad de favoritos (favourites\_count).
  + Estados posteados (statuses\_count).
  + Fecha de creación de la cuenta (created\_at).
  + Zona horaria (time\_zone)
  + Ubicación habilitada (geo\_enabled)
  + Datos relacionados con la foto de perfil (profile\_\*)
* Ubicación del tweet (geo),
* Lugar (place) con un desglose de datos
  + Identificador (id).
  + Direccion web (url).
  + Tipo de lugar (place\_type).
  + Nombre (name) y nombre completo (full\_name).
  + Codigo de país (country\_code).
  + Pais (country),
  + Cuadro delimitador de palabras (boundig\_box), que consta de coordenadas (coordinates) y el tipo (type)
* Colaboradores (colaborators),
* El estado de los retweets (retweeted\_status).
  + Los datos del usuario que hizo el retweets.
* Los hashtags que tiene (hashtags).
* Contadores de retweets (retweet\_count)
* Contadores de favoritos (favorite\_count)
* Contadores de respuesta (reply\_count)
* Citas de recuento (quote\_count)

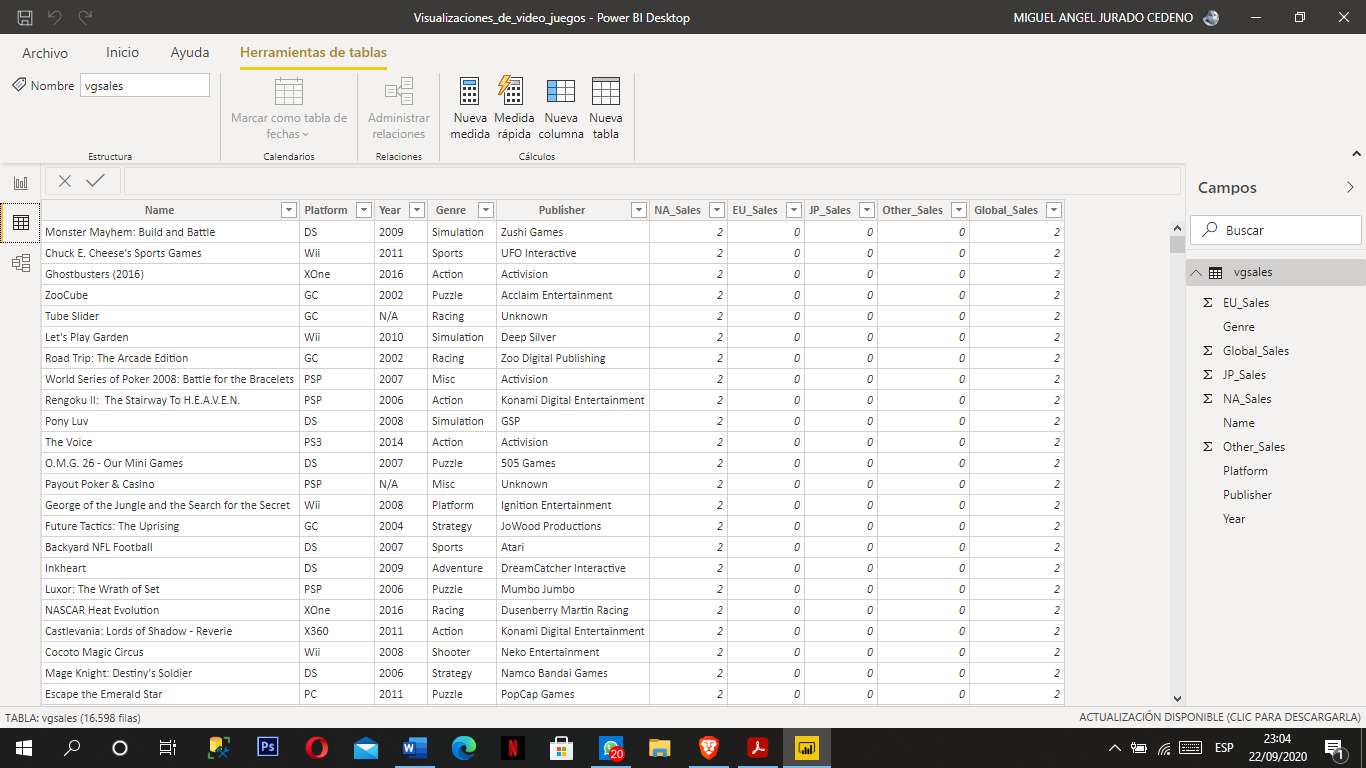
Cabe mencionar que no todos los campos no están llenos. Pero los campos mencionados usan los campos principales y algunos de ellos se muestran en la Figura 7, los cuales servirán para hacer las visualizaciones dentro de kibana.



**Figura 11.** Campos de los Tweets

* 1. **Pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados.**

La estructura principal de los documentos creados en base a los tweets es muy amplia y detallada de la cual se puede tener los siguientes campos:

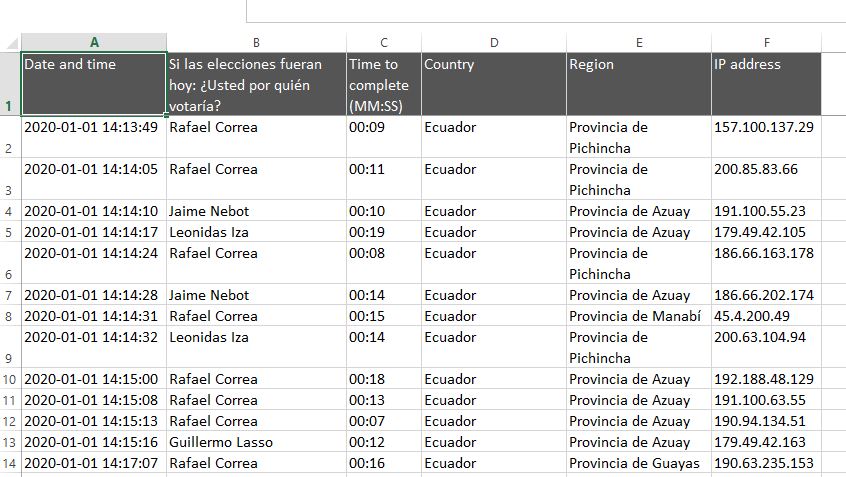
* Un identificador único (\_id).
* Fecha de creación (created\_at).
* Texto del tweet (text).
* Referencia externa (source)
* Los datos del usuario que lo publicó (user) dentro del cual se ubican:
  + Identificador único (id)
  + Nombre del usuario (name)
  + Nickname que tiene en la aplicación (screen\_name).
  + Ubicación (location).
  + Enlaces del usuario (url).
  + Descripción sobre el usuario (descripción).
  + Número de seguidores (followers\_count).
  + Número de amigos (friends\_count).
  + Cantidad de favoritos (favourites\_count).
  + Estados posteados (statuses\_count).
  + Fecha de creación de la cuenta (created\_at).
  + Zona horaria (time\_zone)
  + Ubicación habilitada (geo\_enabled)
  + Datos relacionados con la foto de perfil (profile\_\*)
* Ubicación del tweet (geo),
* Lugar (place) con un desglose de datos
  + Identificador (id).
  + Direccion web (url).
  + Tipo de lugar (place\_type).
  + Nombre (name) y nombre completo (full\_name).
  + Codigo de país (country\_code).
  + Pais (country),
  + Cuadro delimitador de palabras (boundig\_box), que consta de coordenadas (coordinates) y el tipo (type)
* Colaboradores (colaborators),
* El estado de los retweets (retweeted\_status).

**Figura 12:** Dataset de videojuegos en PowerBI

* + Los datos del usuario que hizo el retweets.
* Los hashtags que tiene (hashtags).
* Contadores de retweets (retweet\_count)
* Contadores de favoritos (favorite\_count)
* Contadores de respuesta (reply\_count)
* Citas de recuento (quote\_count)

Por otro lado, la estructura del dataset descargado de la encuesta está estructurado de las siguientes columnas:

* Date and Time: Fecha en la que el usuario realizo la encuesta.
* Candidato: Nombre del candidato.
* Time: El tiempo que le tomo al usuario llenar la encuesta.
* Country: País desde donde el usuario está realizando la encuesta.
* Región: Provincia o estado desde donde el usuario realizo la encuesta.
* IP: Dirección IP del usuario.



**Figura 13:** Datos de la encuesta

* 1. **Juegos en línea por países.**

Una vez que el dataset ha sido cargado en la base de datos MySQL se procede a realizar el análisis en la herramienta PowerBI. Gracias a que la herramienta PowerBI tiene la capacidad de obtener datos que se encuentren en MySQL, se importaron los datos de MySQL directamente a PowerBI para realizar las visualizaciones del dashboard. El dataset obtenido se lo puede observar en Figura 11.

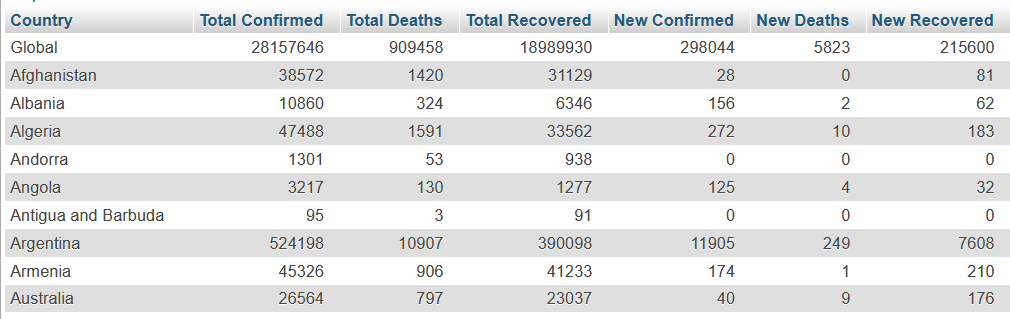
Como se muestra en la Figura 11, el *dataset* contiene las siguientes columnas:

* **Name:** Nombre del video juego.
* **Plataform:** Plataforma del video juego.
* **Year:** Año de publicación del video juego.
* **Genre:** Género o tipo del video juego.
* **Publisher:** Editor del video juego.
* **NA\_Sales:** Ventas del video juego en Norte América (en millones).
* **EU\_Sales:** Ventas del video juego en Estados Unidos (en millones).
* **JP\_Sales:** Ventas del video juego en Japón (en millones).
* **Other\_Sales:** Ventas en el resto del mundo (en millones).
* **Global\_Sales:** Total de ventas en el mundo.
  1. **Eventos o noticias mundiales.**

**General**

En esta tabla se almacena datos del covid-19 a nivel mundial como se muestra en la ***Figura 2***, en esta los datos se dividen en:

* **Country:** que representa el nombre de los países del mundo.
* **Total Confirmed:** el total de casos confirmados hasta la fecha.
* **Total Deaths:** el total de muertes por el covid-19 hasta la fecha.
* **Total Recovered:** el total de recuperados del covid-19 hasta la fecha.
* **New Confirmed:** los casos confirmados en los últimos días
* **New Deaths:** las muertes ocurridas en los últimos días a causa del covid-19.
* **New Recovered:** el total de personas recuperadas en los últimos días.

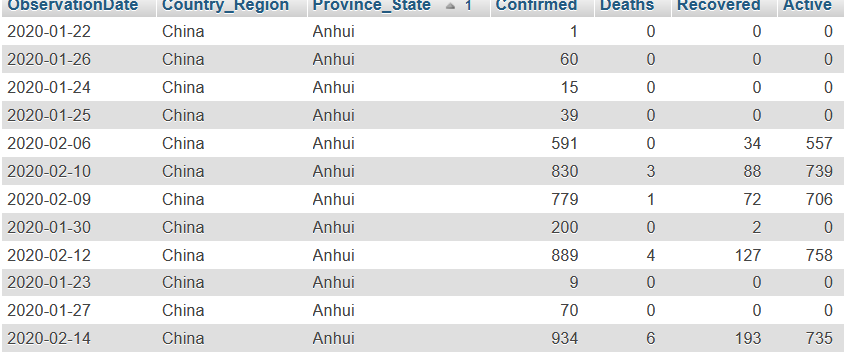


**Figura 14:** Tabla general del covid

**Covid19\_asia**

En esta tabla se almacena datos del covid-19 en el continente asiático como se muestra en la **Figura *3***, en esta los datos se dividen en:

* **Observation/date:** En este campo se registra las fechas que van desde el 22 de enero hasta el 10 de septiembre.
* **Country\_Region:** Se registra el nombre de cada uno de los países asiáticos que se han visto afectados en las fechas registradas.
* **Province\_state:** Este campo registra un lugar más específico de cada uno de los países (no tomado en cuenta porque en su mayoría está representado por null)
* **Confirmed:** El total de casos confirmados en cada país en la fecha indicada.
* **Deaths:** El total de muertes por covid-19 en cada país en la fecha indicada.
* **Recovered:** El total de casos recuperados en cada país en la fecha indicada.
* **Active:** El total de casos activos en cada país en la fecha indicada.

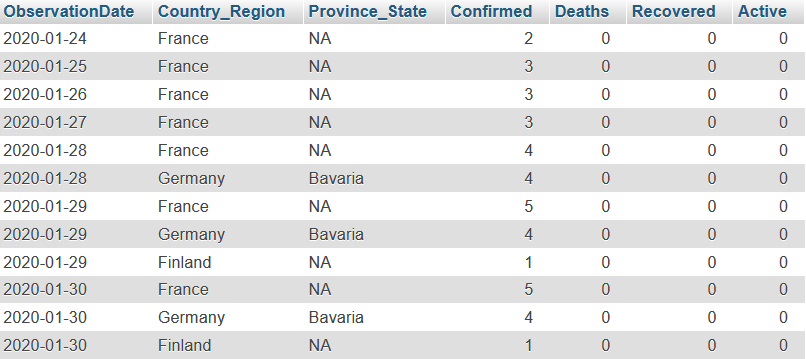


**Figura 15:** Tabla covid19\_asia

**Covid19\_europe**

En esta tabla se almacena datos del covid-19 en el continente europeo como se muestra en la **Figura 16**, en esta los datos se dividen en:

* **Observation/date:** En este campo se registra las fechas que van desde el 22 de enero hasta el 10 de septiembre.
* **Country\_Region:** Se registra el nombre de cada uno de los países europeos que se han visto afectados en las fechas registradas.
* **Province\_state:** Este campo registra un lugar más específico de cada uno de los países (no tomado en cuenta porque en su mayoría está representado por null)
* **Confirmed:** El total de casos confirmados en cada país en la fecha indicada.
* **Deaths:** El total de muertes por covid-19 en cada país en la fecha indicada.
* **Recovered:** El total de casos recuperados en cada país en la fecha indicada.
* **Active:** El total de casos activos en cada país en la fecha indicada.



**Figura 16:** Tabla covid19\_europe

**Covid19\_southamerica**

En esta tabla se almacena datos del covid-19 en el continente americano especialmente de los países latinoamericanos como se muestra en la **Figura 5**, en esta los datos se dividen en:

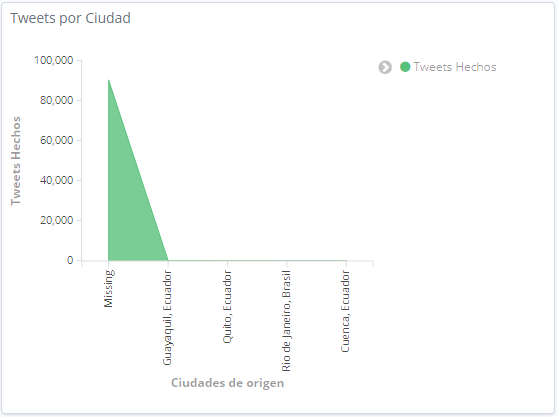
* **Observation/date:** En este campo se registra las fechas que van desde el 22 de enero hasta el 10 de septiembre.
* **Country\_Region:** Se registra el nombre de cada uno de los países latinoamericanos que se han visto afectados en las fechas registradas.
* **Province\_state:** Este campo registra un lugar más específico de cada uno de los países (no tomado en cuenta porque en su mayoría está representado por null)
* **Confirmed:** El total de casos confirmados en cada país en la fecha indicada.
* **Deaths:** El total de muertes por covid-19 en cada país en la fecha indicada.
* **Recovered:** El total de casos recuperados en cada país en la fecha indicada.
* **Active:** El total de casos activos en cada país en la fecha indicada.



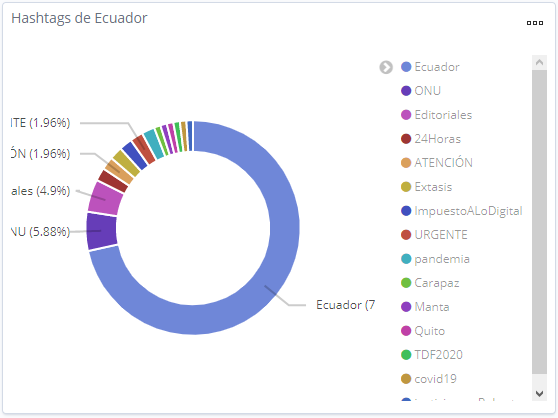
**Figura 17:** Tabla covid19\_southamerica

1. **Visualización de información**
   1. **Pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados.**

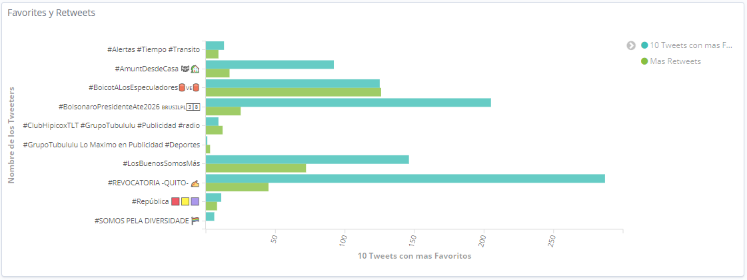
A partir de los datos obtenidos se obtienen las siguientes visualizaciones.



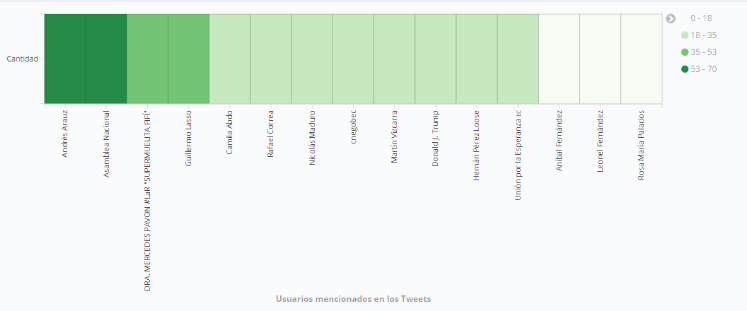
**Figura 18.** Visualización según la cantidad de tweets por cuidad



**Figura 19** Visualización según los hashtags de cada Tweet



**Figura 20** Visualización de ranking de los 10 Tweets con más Favoritos y Retweets.

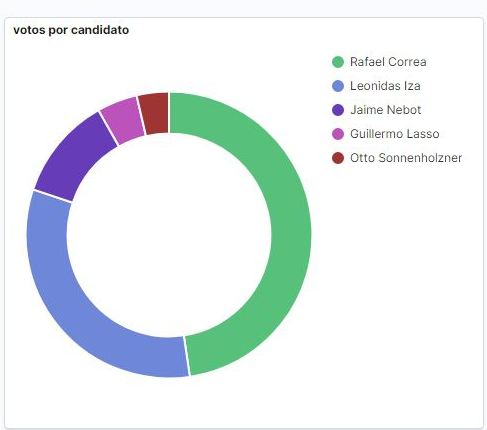


**Figura 21** Visualización de los 10 usuarios de Twitter más mencionados

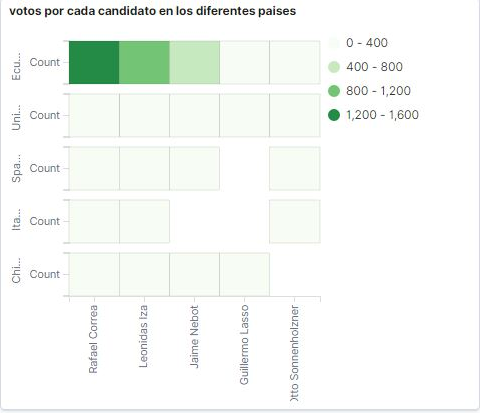


**Figura 22** Visualización de los Hashtags mas usados dentro de los Tweets.

* 1. **Pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados.**



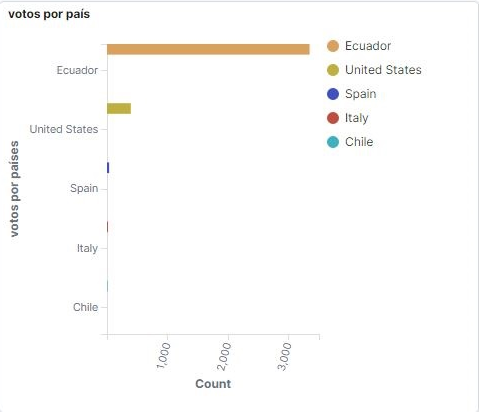
**Figura 23:** Total de votos por candidato



**Figura 24:** Votos a cada candidato por cada país

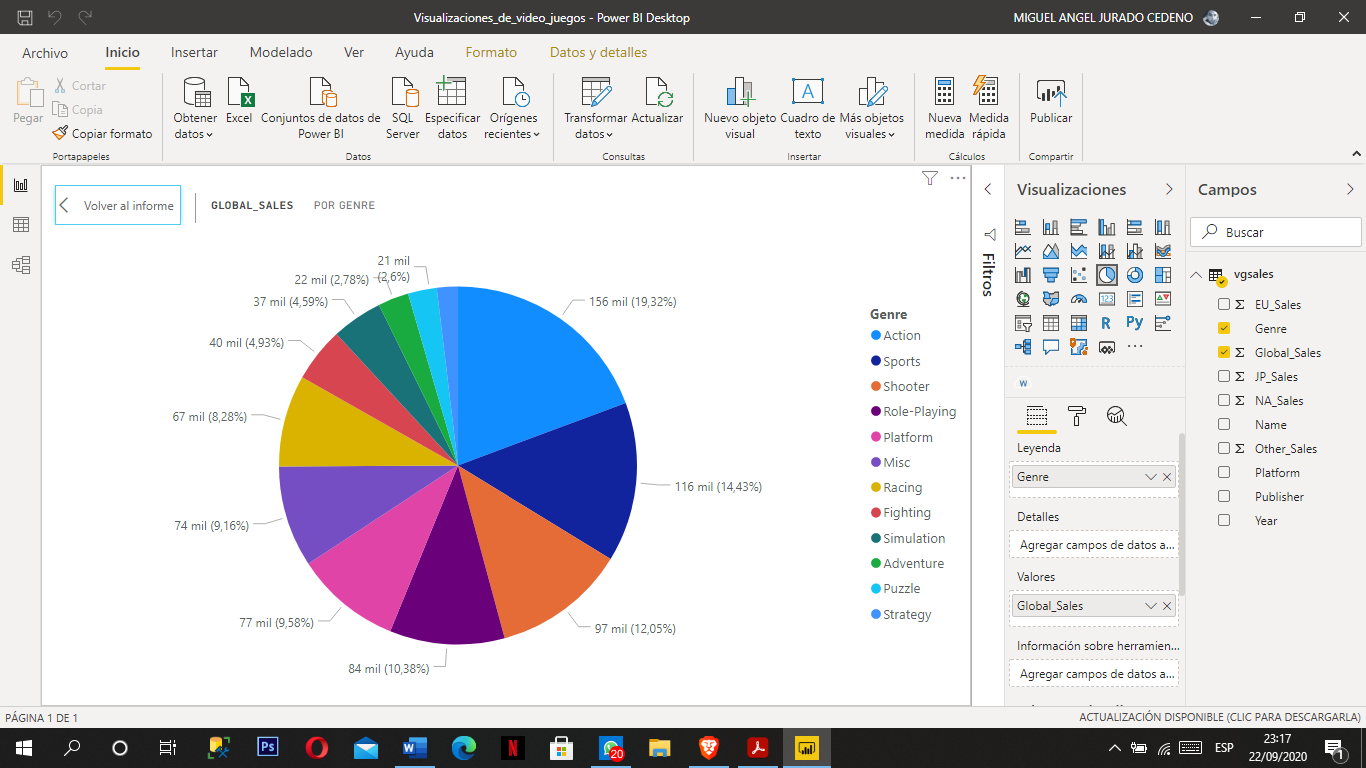


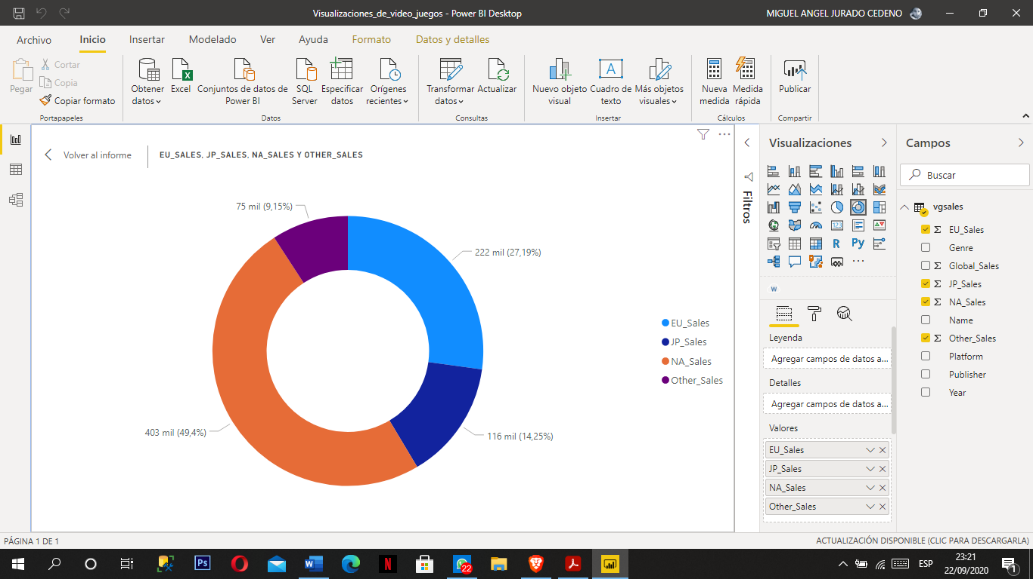
**Figura 25:** 10 Regiones y provincias mencionadas.



**Figura 26:** Votos generados por cada país.

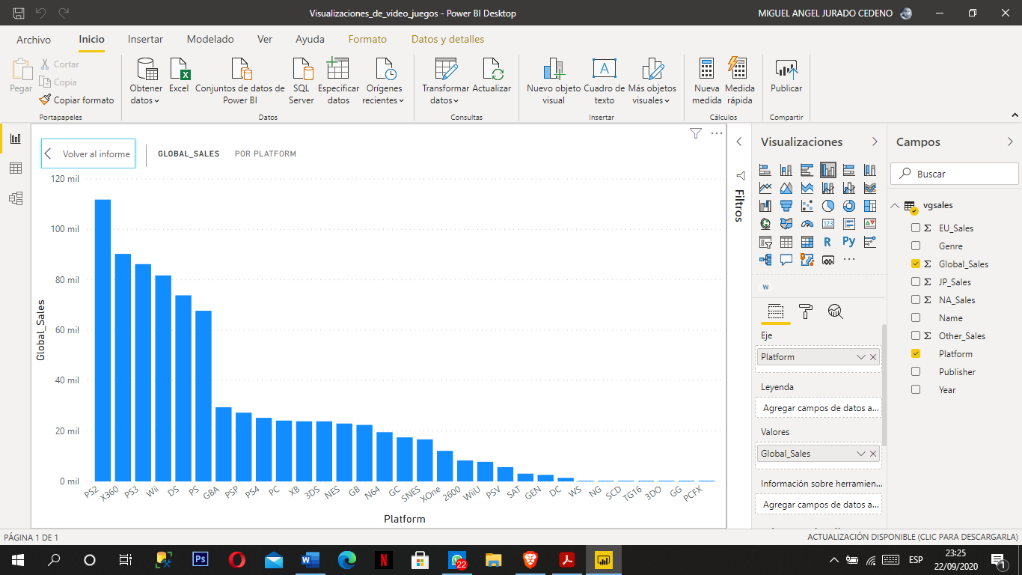
* 1. **Juegos en línea por países.**



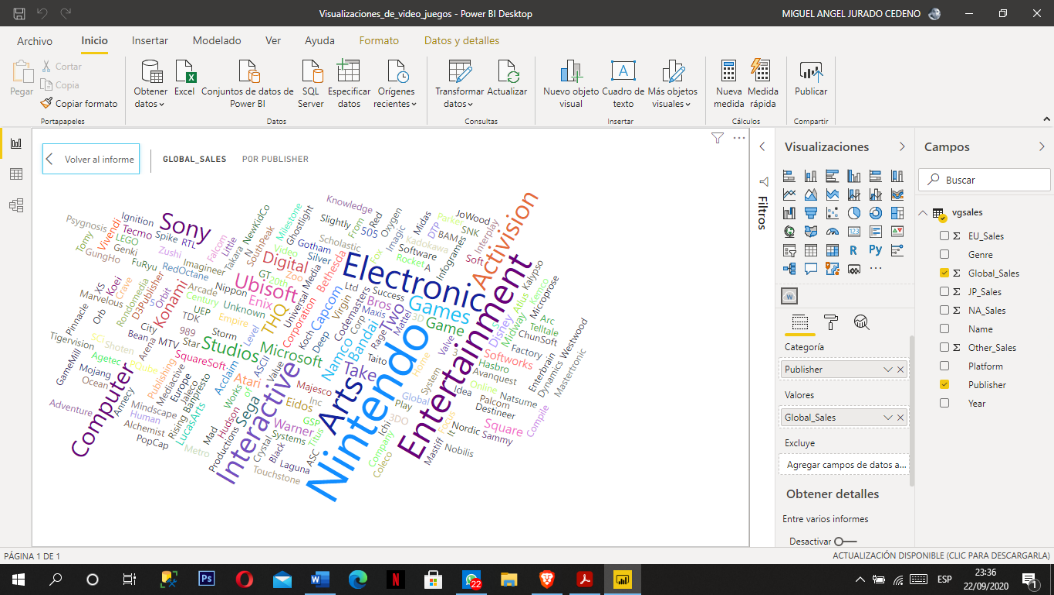


**Figura 27:** Ventas globales por género.

**Figura 28:** Ventas por país



**Figura 29:** Ventas globales por plataforma



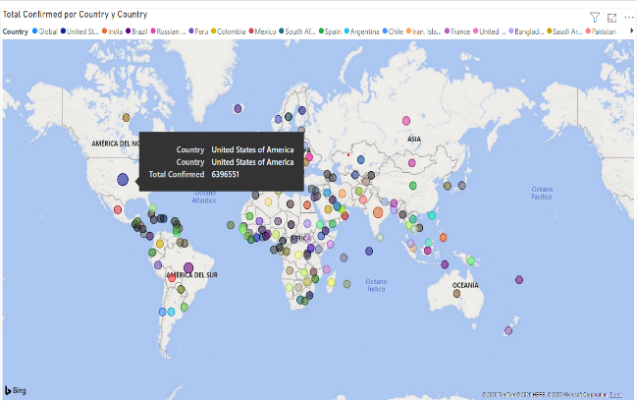
**Figura 30:** Ventas globales por editor

****

**Figura 31:** Filtro por año del dashboard

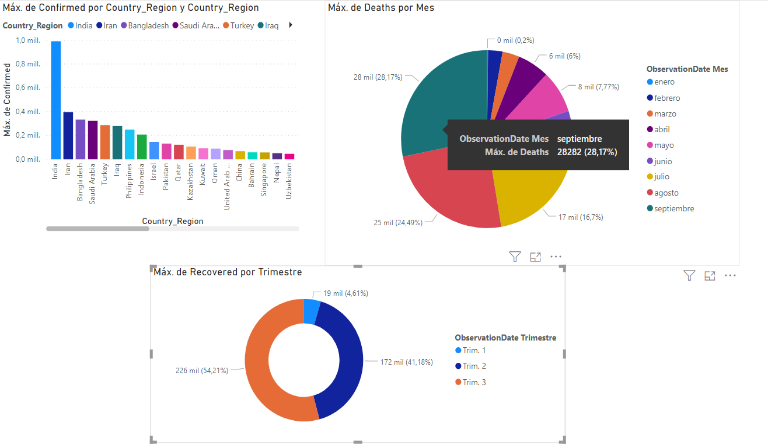
* 1. **Eventos o noticias mundiales**

**General**



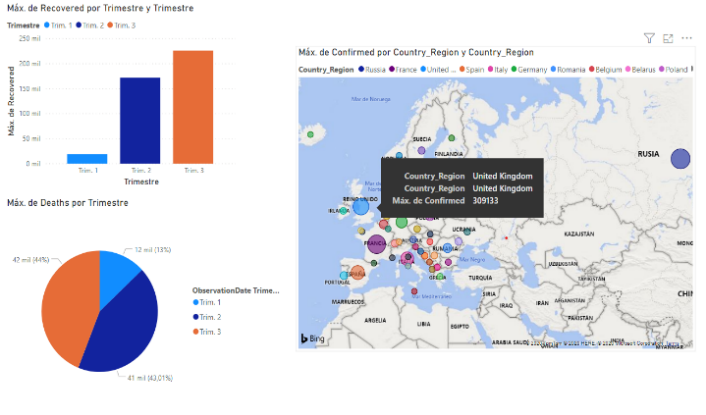
**Figura 32:** Total de casos confirmados por país

**Covid19\_asia**



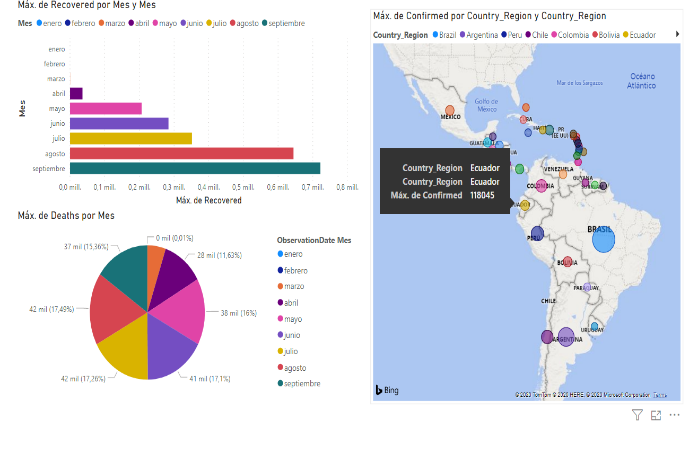
**Figura 33:** Dashboard obtenido de la tabla covid19\_asia

**Covid19\_europe**



**Figura 34:** Dashboard obtenido de la tabla covid19\_europe

**Covid19\_southamerica**



**Figura 35:** Dashboard obtenido de la tabla covid19\_southamerica

1. **Resultados obtenidos**
   1. **Pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados.**

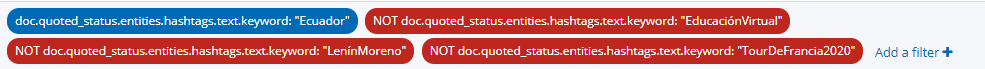
Los resultados obtenidos a partir de cada una de las visualizaciones que conforman el dashboard de Tweets por Cuidad del País.

La interpretación de cada una de las visualizaciones se presenta a continuación:

algún dato, para el resto de los puntos en la gráfica se observa que están los datos de la cuidad de Guayaquil, Quito, Rio de Janeiro y Cuenca, debido a la razón expuesta de los datos impuros se muestra Rio de Janeiro, es así que si a la gráfica se amplían los campos de visualización saldrán ciudades de otros países.

**Las veces de un Hashtags fue mencionado en los Tweets.**

La ***Figura 13*** muestra los hashtags que más fueron usados en todos los datos recolectados, pero por la diversidad de los datos estos fueron filtrados, es decir se quitaron ciertas palabras que no van de la mano con el tema principal, dándole prioridad a todos los que iban relacionados con Ecuador, de ahí los que no tenían nada que ver con la política, suprimiendo los hashtags de deportes entre otros, estos cambios aplicados se muestran en la siguientes figura.



**Figura 36.** Filtros aplicados a la Figura 14

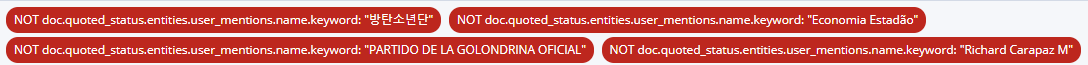
Una vez aplicados los cambios, en el gráfico de pastel, se definen que los hashtags más usados son los que llevan la palabra Ecuador, seguidos por temas no tan relacionados con la política, pero que generan noticas de sobre el tema.

**Ranking de los 10 tweets con más favoritos y retweets ordenados en orden ascendente.**

La ***Figura 14*** muestra los tweets con los más favoritos o en el caso de otra red social serian likes o me gusta, además de todos los retweets que se han producido en cada uno de estos, pero fueron seleccionados en el orden ascendente. También se muestra el nombre de los propietarios de cada uno de los tweets y los que también está ordenados alfabéticamente, cabe mencionar que los datos no son puros, por eso de los 10 datos que se muestran solo 2 están relacionados con el tema de la política del país.

**Usuarios de Twitter más mencionados en los Tweets**

La ***Figura 15*** muestra el top 10 de los usuarios que fueron mencionados en más tweets, donde se registran la mayoría de los temas de los que se está hablando, es decir, que todos están relacionados, pero para esto se usaron algunos filtros debido a que los últimos temas se han visto enmarcados con los deportes, los filtros se muestran en la siguiente figura.



**Figura 37** Filtros de Usuarios mencionados en Twitter

Con estos filtros se determina que los datos se enmarcan los pre-candidatos como el Sr Víctor Arauz con un total de 67 menciones, además del Sr. Guillermo Lazo con 37 entre otros representantes de los diferentes partidos. También destacan el poder legislativo del país en este caso la Asamblea Nacional que cuenta con 63.

**Palabras (Hashtags) más usados en los Tweets**

La **Figura 16** muestra las palabras más usadas en todos los datos o tweets recopilados, siendo la principal es “Ecuador” que está en el medio, y de donde se despliegan hashtag de partidos políticos importantes y las próximas elecciones de Ecuador con la etiqueta (hashtag) Elecciones2021Ec. En este caso se puede observar que en las descripciones de cada uno de loa tweets se encuentran datos sobre un tema común o la mayoría de estos los tiene.

Es así como se muestra que los datos por políticos por ciudades no son precisos los momentos de buscarlos porque solo se puede aplicar un filtro y si se los usa por palabras en específico las ciudades importantes del país que son tomadas en cuenta son Quito, Guayaquil y Cuenca. Por eso una gráfica de ubicación geográfica es difícil establecer ese tipo de visualizaciones.

* 1. **Pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados.**

**Total, de votos por candidato**

La **Figura 21** muestra la cantidad de votos obtenidos en la encuesta realizada por diario El Mercurio, donde se posiciona Rafael Correa como ganador. Se puede notar fácilmente que la encuesta como dice el artículo del diario fue realizada a inicio de año donde aún no se sabía a ciencia cierta el estado judicial y político de Correa. En segundo lugar, se posiciona Leónidas Iza, seguido de Jaime Nebot.

Para cerrar el listado encontramos a dos candidatos que también fueron mencionados en Twitter hablamos de Guillermo Lasso y Otto Sonnenholzner. Es aquí donde existe una diferencia de los datos ya que la extracción de datos de Twitter no permite obtener partes de texto claras para analizar nombres, todo lo realizado en Twitter es en base a hashtags.

**Votos a cada candidato por cada país**

Teniendo ya clara la idea de los posibles candidatos que el pueblo prefiere, en la se muestra cómo se distribuye estos votos, pero por país. Los indicadores de la gráfica **Figura 22** muestra cómo se distribuye la aceptación especialmente en el Ecuador. Como dato curioso podemos notar como Guillermo Lasso no tiene aceptación ni en España ni tampoco en Italia.

**10 regiones y provincias más mencionadas**

La **Figura 23** muestra 10 de las regiones, provincias o estados que más mencionaron a los candidatos. Esta muestra una gran participación en algunas provincias del país. Además, de mostrar un par de estados de los Estados Unidos. Esta visualización se la realizó uniendo los dos campos similares que tenían las dos fuentes.

Es notoria aquí la parte en donde la recolección de tweets nos refleja que a pesar de no ser provincias muy grandes las provincias de Loja, Azuay y Manabí están muy activas y pendientes del proceso electoral.

**Votos por países**

La encuesta realizada por diario El Mercurio nos da un gran análisis y con este se pudo definir como indica la **Figura 24**, la participación de las personas del Ecuador, Estados Unidos, España, Italia y sorpresivamente Chile. Al parecer esta encuesta llego a muchos lugares y tuvo una gran repercusión. Es bueno ver que principalmente en Ecuador la gente vaya definiendo una visión futura a las elecciones.

* 1. **Juegos en línea por países.**

**Ventas totales por género**

En la ***Figura 26*** se puede consultar con mayor detalle cual es la cantidad de ventas totales por género. Claramente se puede observar que el género que más ha tenido ventas es el Acción, sin embargo, esto puede variar dependiendo el año. Para esto se utilizó un filtro de años, el cual permitirá visualizar la información en el año que se solicite, y esto será reflejado en las demás visualizaciones que tiene el dashboard.

**Ventas por países**

Las ventas de los países que se reflejan en el dataset son solo Estados Unidos y Japón, sin embargo, existen columnas que unen el resto de los países, por ejemplo, NA\_sales, donde se encuentran todas las ventas de los países de allí, Other\_sales, donde se unen las ventas de otros países y Global\_sales que une las ventas totales. Teniendo en cuenta esto y observando la ***Figura 27***, se puede concluir que obviamente donde hay más ventas es en Norte América porque engloba muchos países, sin embargo, Estados Unidos no se queda atrás siendo solo la suma de un país.

**Ventas globales por plataforma**

Una vez que se tiene conocimiento de cuáles son los géneros de videojuegos más vendidos y en donde se venden más los videojuegos, es momento de saber cuál es la plataforma que más ha vendido videojuegos durante todos estos años. Para saber esto podemos consultar la ***Figura 28***, la cual nos muestra justamente esto y se puede concluir que la plataforma que más ha vendido durante estos años ha sido la PS2, esto debido a que cuando empezó su producción tuvo una gran demanda. Cabe mencionar que gracias al filtro de año que se tienen en el dashboard se puede consultar que plataforma vendió más videojuegos en un año específico.

**Ventas globales por editor**

También es bueno saber cuáles fueron los editores de videojuegos que tuvieron más ventas. Para saber esto se puede consultar la ***Figura 29***. Analizando la visualización de la ***Figura 29***, se puede concluir que entre los editores de videojuegos que más ventas tuvieron se encuentran Nintendo, Entertainment, Electronic, Activision, Arts, Interactive. Esto puede cambiar si se cambia el año en el filtro del dashboard.

* 1. **Eventos o noticias mundiales.**

**General**

Con base a la ***Figura 32*** la información obtenida nos deja como resultado que los países que tienen una mayor cantidad de datos confirmados son:

* Estados Unidos (6396551)
* India (4562414)
* Brasil (4238446)
* Rusia (1042836)

Esto se puede determinar porque estos países se representan con un punto más grande que los demás.

**Covid19\_asia**

Con base a la ***Figura 33*** se puede observar que se realizaron tres gráficos diferentes: el primero es un diagrama de barras donde se relaciona los casos confirmados por país. La información obtenida del grafico es la siguiente:

* India (una máxima de casos confirmados de 990795)
* Irán (una máxima de casos confirmados de 395448)
* Bangladesh (una máxima de casos confirmados de 332970)
* Arabia Saudita (una máxima de casos confirmados de 323720)

El segundo es un diagrama circular que relaciona el máximo de muertes por mes, donde se puede observar que los últimos meses ha sido donde más personas fallecieron siendo septiembre el mes en donde más muertes ocurrió con una máxima de 28282 fallecidos.

El tercer diagrama es un gráfico de anillos donde se muestra la máxima de recuperados por trimestre, dando como resultado que el último trimestre (julio, agosto, septiembre) ha sido donde más personas se han recuperado un valor aproximado de 226274.

**Covid19\_europe**

Con base a la **Figura 34** se puede observar que se realizaron tres gráficos diferentes: el primero es un diagrama de barras donde se relaciona los casos recuperados por trimestre, dando como resultado que de igual manera el último trimestre (julio, agosto, septiembre) ha sido donde más personas se han recuperado con valor aproximado de 226274.

El segundo es un gráfico circular donde se muestran en cambio las muertes por trimestre, donde nos da como resultado que el último trimestre (julio, agosto, septiembre) han fallecido aproximadamente 42072 personas. Si se realiza una comparación entre el primer y el segundo grafico se puede ver que existe una pequeña relación entre los casos recuperados con los fallecidos.

En el tercer y último grafico se presenta un mapa en donde se puede observar cuales son los países con más casos confirmados, dando los siguientes resultados:

* Rusia (una máxima de casos confirmados de 405843)
* Francia (una máxima de casos confirmados de 372501)
* Reino Unido (una máxima de casos confirmados de 309133)
* España (una máxima de casos confirmados de 236899)

**Covid19\_southamerica**

Con base a la ***Figura 35*** se puede observar que se realizaron tres gráficos diferentes: el primero es un diagrama de barras horizontales donde se relaciona los casos recuperados por mes, dando como resultado que septiembre es el mes donde más personas se han recuperado con un valor aproximado de 722025.

El segundo es un gráfico circular en donde se relacionan la cantidad de muertos por mes donde nos da como resultado que agosto fue el mes en donde más personas fallecieron con un valor aproximado de 42072.

En el tercer y último grafico se presenta un mapa en donde se puede observar cuales son los países con más casos confirmados, dando los siguientes resultados:

* Brasil (una máxima de casos confirmados de 874754)
* Argentina (una máxima de casos confirmados de 524198)
* Perú (una máxima de casos confirmados de 328903)
* Chile (una máxima de casos confirmados de 278713)

1. **Conclusiones y recomendaciones**

* La importación y exportación de datos es algo que se debe hacer para cumplir la arquitectura del proyecto, así que, hay que saber bien la estructura de los datos para no tener problemas al momento de exportar o importar estos.
* La limpieza de datos es un tema importante que se debe hacer para que al momento de realizar las visualizaciones estas puedan ser lo más exacta posibles y puedan cumplir con su propósito.
* La información que se puede obtener en base a los datos obtenidos es muy importante y variada, hemos visto como la información construye y nos da una idea clara de preferencias e ideas, de que se está hablando actualmente.

1. **Desafíos y problemas encontrados**

Uno de los problemas encontrados fue que para poder cargar todos los datos del dataset a la base de datos MySQL, se tuvo que cambiar el tiempo máximo de ejecución, ya que, de manera predeterminada está establecido en 30 segundos lo que no permitía que los datos fueran cargados totalmente.

Otro problema presentado fue que no existe el complemento input de MongoDB usando Logstash, una de las soluciones fue tomar los datos de MongoDB y moverlos a CouchDB desde el cual se integraran a Elasticsearch en un índice.  
Se descargó cada una de las colecciones y se las agregaron a una sola base de datos

.

**Link del repositorio de github**

<https://github.com/marvinzambrano310/ProyectoFinalAnalisis>