Informe Proyecto final

Elvis Lenin Agila Paladines

[elvis.agila@epn.edu.ec](mailto:elvis.agila@epn.edu.ec)

Jonathan Efrain Alquinga Cajamarca

[jonathan.alquinga@epn.edu.ec](mailto:jonathan.alquinga@epn.edu.ec)

Yomara Isabel Diaz Oscullo

[yomara.diaz@epn.edu.ec](mailto:yomara.diaz@epn.edu.ec)

Guillermo Alexander Rivera Guerrero

[guillermo.rivera@epn.edu.ec](mailto:guillermo.rivera@epn.edu.ec)

1. **Definición de caso de estudio**

Los casos de estudio utilizados para el proyecto son:

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

Debido a la creciente actividad de las organizaciones políticas debido a la elección presidencial, es necesario recopilar información acerca de las ciudades del país y la opinión de los civiles con respecto al estado político, social y económico por el cual se está desarrollando la campaña electoral, de esta manera se permita implementar un análisis y presentar resultados con respecto a los parámetros por los cuales se realizan las filtraciones de los datos concluyendo en un proceso de conocimiento y aprendizaje sobre este tipo de temas de interés.

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

Un motivo importante es que el país va a ingresar a un proceso electoral, la segunda vuelta para los candidatos Andrés Arauz y Guillermo Lasso. El objetivo de este apartado es recopilar datos de los dos candidatos que se están postulando a presidente, con estos datos se procederá ah analizar 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 video juegos en línea para realizar un análisis sobre qué categoría, plataforma, editor del juego tuvo más ventas desde el año 1980 hasta 2017. A

su vez con una base de datos se puede observar el valor en dólares de la venta de video juegos durante algunos años en diferentes países.

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

Se ha tomado el tema del Covid-19 como caso de las noticias mundiales para la recopilación de datos. El propósito de la recopilación de datos es llevar un análisis de cómo va evolucionando la enfermedad alrededor del mundo, para ello hemos tomado datos mundiales y de dos países en específico en este caso son: Japón e India. Estos datasets fueron recopilados desde la aplicación de Kaggle y se realizó el análisis de los datos recopilados por medio de la aplicación de Power BI.

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

Receptar y analizar conjuntos de datos recopilados a través de técnicas y herramientas de extracción de data.

* 1. **Objetivos específicos**
* Definir las formas en las cuales se va a recolectar los datos, incluyendo los parámetros en aquellos en los que se tiene que extraer información de fuentes específicas.
* Definir las bases de datos en las cuales se almacenará la información.
* Establecer la arquitectura con la cual se desarrollará el proyecto, tomando en cuenta la vía más factible para cumplir con la problemática expuesta.
* Recopilar data con respecto a los temas expuestos en la problemática.
* Recopilada la información, establecer un punto en el cual se incorporará el conjunto de dicha información.
* Realizar las visualizaciones acordes a los resultados que presente la información receptada.
* Interpretar las visualizaciones y los resultados de tal manera que permita a todo tipo de usuario entender de qué trata el tema expuesto y las características que presente.

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

El equipo encargado del desarrollo del proyecto está compuesto por cuatro integrantes los cuales son: Elvis Agila, Jonathan Alquinga, Yomara Díaz, Guillermo Rivera; para cada integrante se definieron un conjunto de tareas que permitieron agilizar el proceso de recolección de datos para su posterior análisis y presentación por medio de exposiciones del procedimiento y visualizaciones de los procedimientos realizados.

|  |  |
| --- | --- |
| **Nombre** | **Tareas** |
| Elvis Agila | * Recopilar datos sobre los videojuegos por países utilizando las herramientas con bases de datos públicas como las páginas de Kaggle.com, Tableu Public Resources y UNdata. * Se exportará los datos a la base de datos SQL en este caso MySQL. * Las visualizaciones de los datos obtenidos se realizarán utilizando la herramienta de PowerBI ya que la misma cuenta con la función de extraer datos de MySQL. |
| Yomara Diaz | * 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. |
| Jonathan Alquinga | * Recopilar datos sobre el pulso político por provincias en Ecuador, listas y candidatos y presidenciales, 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 Kibana. * A partir del tema de preferencia elegido que es “Test de Personalidad”, se realizara las respectivas tareas mencionadas anteriormente. |
| Guillermo Rivera | * 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**

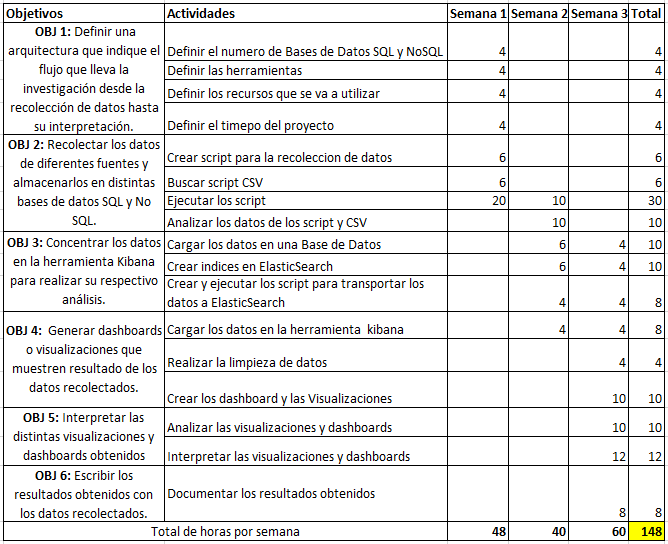


Figura 1. Cronograma de actividades

1. **Recursos y herramientas utilizadas**

En el proceso de recopilación de datos, se hizo uso de diferentes herramientas, recursos y técnicas de extracción y almacenaje como los son scripts desarrollados en Python en los cuales se incorporó librerías propias de las bases de datos a las cuales se enviaba la información, a parte de esas también se ocuparon credenciales de Twitter para poder filtrar la información por medio de parámetros de búsqueda que posteriormente se almacenaron en un bases de datos tipo NoSQL como lo son CouchDB y MongoDB.

En aquellos datasets que se buscaron y extrajeron están en formato CSV para la lectura y utilización de estos, para su extracción se usaron herramientas como Kaggle, INEC, Databank, para dichos archivos se implementó su almacenaje en MySQL que son bases de datos tipo SQL o bases de datos relacionales, aparte de que este tipo de archivos pueden ser editados directamente en herramientas de ofimática como Excel.

Si bien se hizo uso de varias herramientas de almacenaje y recopilación, es necesario detallar a aquellos recursos que facilitaron el proceso de visualización y análisis de los datos como lo es Kibana, para poder observar los datos en tiempo real y también Power BI que es otra herramienta que permite establecer dashboards en los cuales incorporar visualizaciones de forma estática, es importante destacar que dichas herramientas permiten el almacenaje en la nube.

Proyectos open source (ELK) permitieron concluir con el proyecto realizado, para ellos se utilizó Elasticsearch, Logstash y Kibana como se mencionó, debido a que Elasticsearch es un motor de búsqueda y análisis de data, por otro lado, Logstash es un pipeline de procesamiento, lo que significa que transforma un flujo de datos de manera continua en un proceso que es comprendido por varias fases secuenciales [1], posterior a eso Logstash envía la información a lugares “ocultos”, Kibana y Elasticsearch permiten la visualización de datos de manera dinámica para el usuario y finalmente pero no mes importante se hizo uso de la herramienta cerebro que permite la gestión de los índices que se incorporen en las herramientas anteriormente citadas.

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 datos, compuesta de dos partes. La primera parte consiste en recolectar datos y guardar en las bases locales de CouchDB, Mongo DB y MySQL estas bases de datos contendrán las principales fuentes de datos, estudiadas a lo largo de curso (Twitter, Facebook, Webscrapping, CSV, JSON, Kaggle, entre otras).

La segunda parte tiene como objetivo alojar los datos que tenemos individualmente, a un clúster o base concentradora que es Mongo DB Atlas para así poder llevar los datos a ElasticSearch así poder realizar su indexación, realizar dashboards y realizar las respectivas visualizaciones en Kibana.

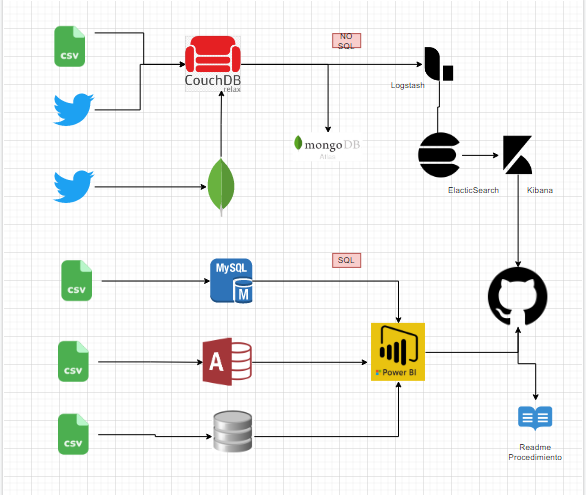
A continuación, la Figura 2 se muestra el diseño de nuestra arquitectura la cual como se detalló anteriormente nos permitirá luego de recopilar la información analizarla y plasmarla en visualizaciones.

Figura 2. Arquitectura de la solución.

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

Para la recopilación de datos desde Twitter por medio del script de Python se estableció la base de datos “Ciudades” MongoDB con 20 colecciones, una por cada ciudad que se estableció en la problemática, en el script definido como “Tweets Ciudades.py” se implementaron los parámetros de búsqueda como se muestra en la Figura 3, siendo estos: *El país, la ciudad, “Pulso político”, “Elecciones”, “Candidatos”, “Presidencia”, “Diputados”*.



Figura 3. Parámetros de búsqueda en el script

Las ciudades de las que se recopiló información por medio de la red social Twitter fueron: Tulcán, Otavalo, Salinas, Villamil, Santo Domingo, Salinas, Tena, Ibarra, Machala, Ambato, Esmeralda, Loja, Cuenca, Babahoyo, Manta, Quito, Guayaquil, Los Ríos, Portoviejo, Chone. Los datos recopilados como se muestran en la Figura 4 se procedieron a descargar posterior a su almacenaje, lo cual permitió aplicar el análisis y visualizaciones.

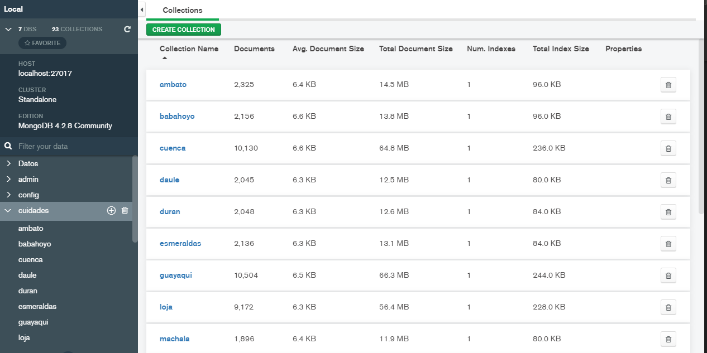


Figura 4. Datos recolectados en MongoDB

Se utilizaron 3 diferentes bases de datos No SQL para el almacenaje de datos, para lo cual se hizo uso de scripts en Python para realizar la conexión entre estas, primero se guardaron los datos en MongoDB y posterior a eso se mandaron a CouchDB debido a que no hay un input desarrollado para logstash que permita mandar los índices de una base en MongoDB hacia Elasticsearch, también se almacenó en MongoDB Atlas debido a que permite manejar en la nube y de esta forma no se consumen tantos recursos.

En total se recolectó un aproximado de 123 mil 227 datos con respecto a los parámetros de búsqueda, lo que permitió tener una amplia gama de opciones al momento de querer analizarlos.

Para realizar dicha conexión se definieron índices en Elasticsearch para cada ciudad permitiendo separar los datos recolectados por ese mismo tema, los datos almacenados en Couch como se muestra en la Figura 5, denotan también el nombre de los índices definidos en Elasticsearch como se muestra en la Figura 6, a parte de definir los Clusters previo al envío de los datos.

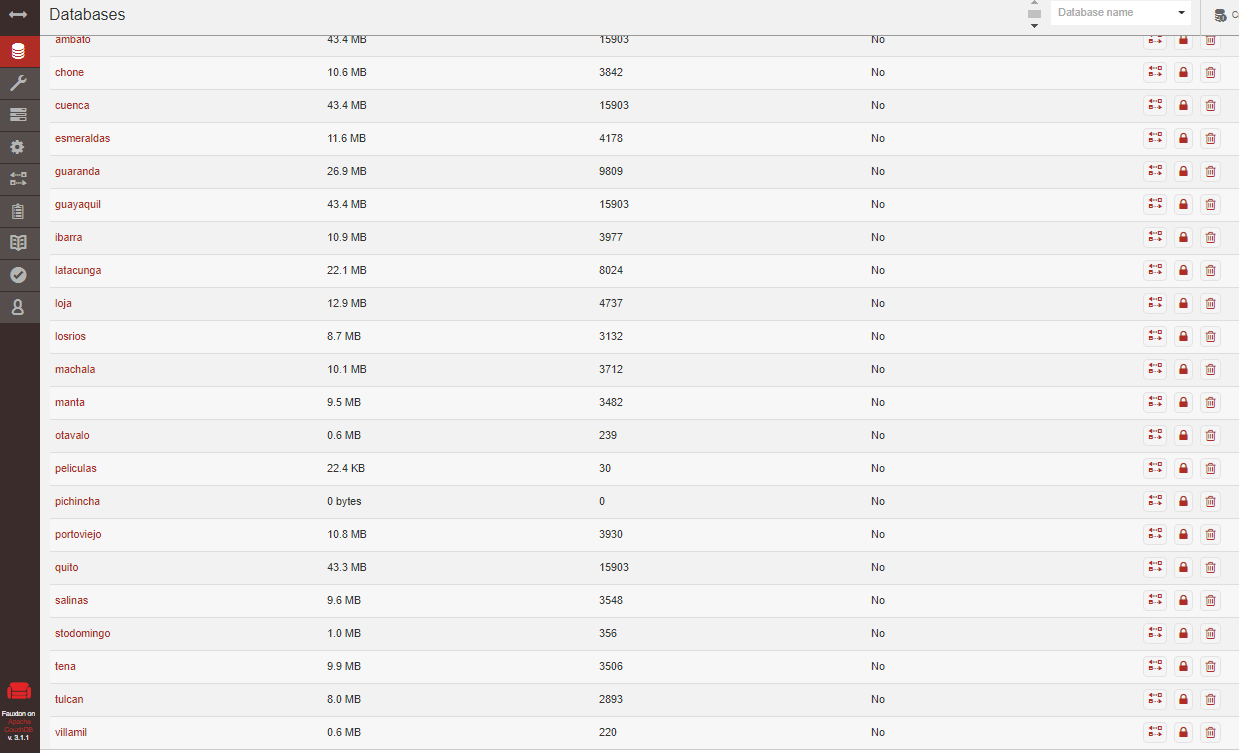


Figura 5. Datos almacenados en CouchDB.

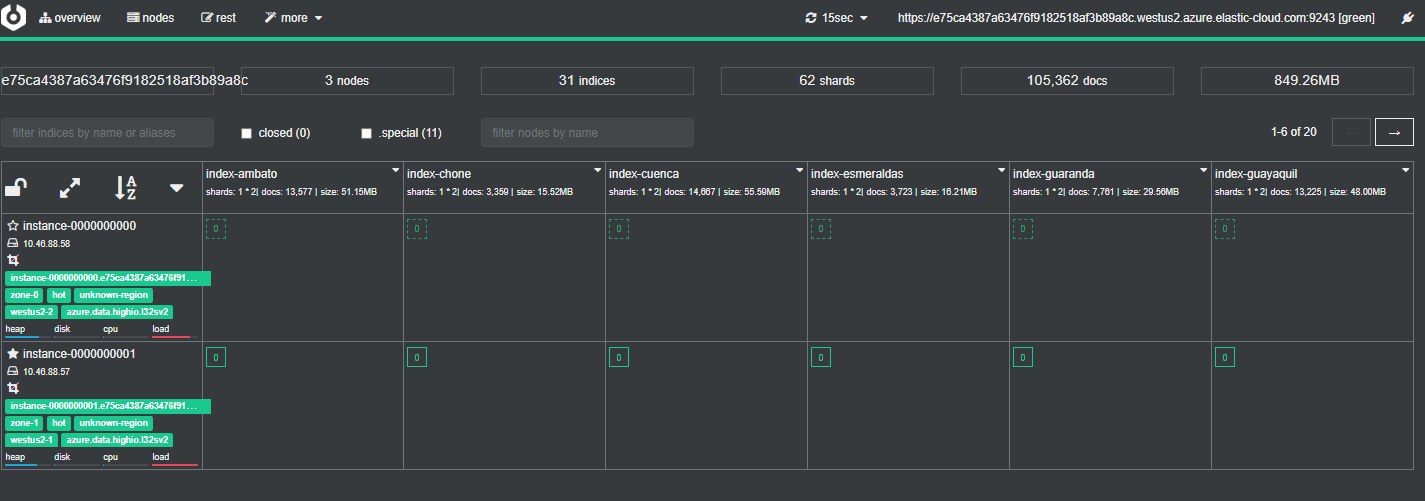


Figura 6. Índices y cluster en Elasticsearch.

La forma en la que se enviaron los datos desde Couch hasta Elasticsearch fue desarrollado por medio de un input como se mencionó con anterioridad, el cual contiene el servidor desde donde se va a sacar los datos y hacia donde se los quiere enviar, en este caso se utilizó almacenamiento en la nube de Elasticsearch, el script con el Input como se muestra en la Figura 7, contiene también el “Id” que nos devuelve el servidor de Elasticsearch.

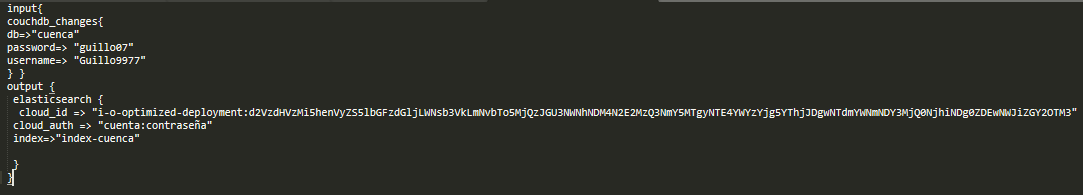


Figura 7. Script con input.

Cabe recalcar que los datos que se almacenen en Elasticsearch serán analizados en forma de visualización con Kibana, también de forma remota.

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

La extracción de datos de este tema tiene dos partes. El primero es Twitter donde hicimos el uso de un script en Python para poder obtener datos de diferentes provincias, todos estos datos se exportarán a la Base de Datos CouchDB. El script igualmente para la 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”. Por último, se creó una Base de Datos general llamada Ecuador la cual tiene como objetivo tener todos los datos de las provincias que se recolectaron individualmente como se muestra en la Figura 8, esta base de datos es la que se va a exportar a ElasticSearch donde se podrá realizar dashboards y visualizaciones respecto al tema planteado.

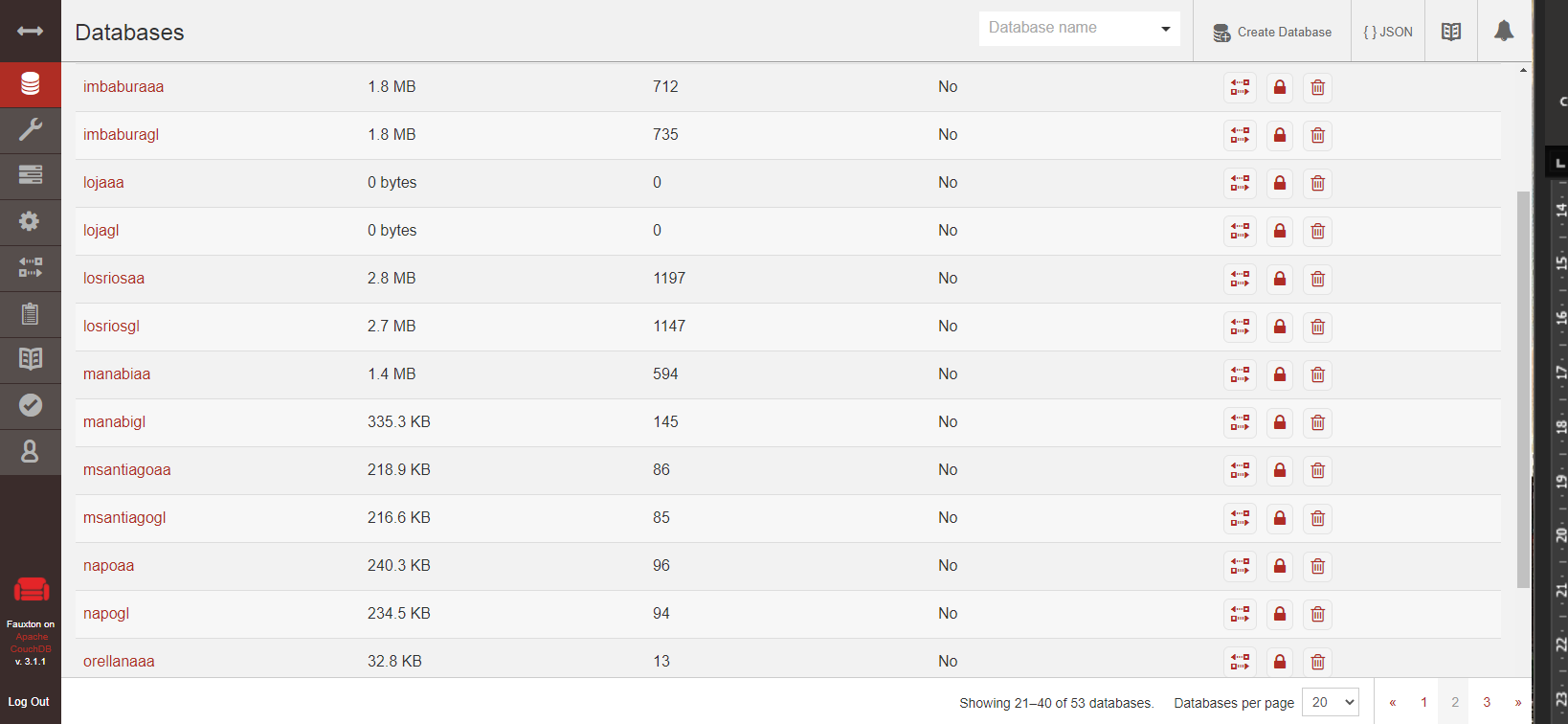


Figura 8. Datos en CouchDB.

La segunda fuente fue una encuesta realizada por el diario El Mercurio de la ciudad de Cuenca, la cual fue lanzada en el año 2020 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 que se obtuvo contaba con datos aleatorios de posibles presidentes que podrían llegar a ganar, cabe recalcar que esto solo es una encuesta que se realizó aproximadamente a 3000 personas como se muestra en la Figura 9.

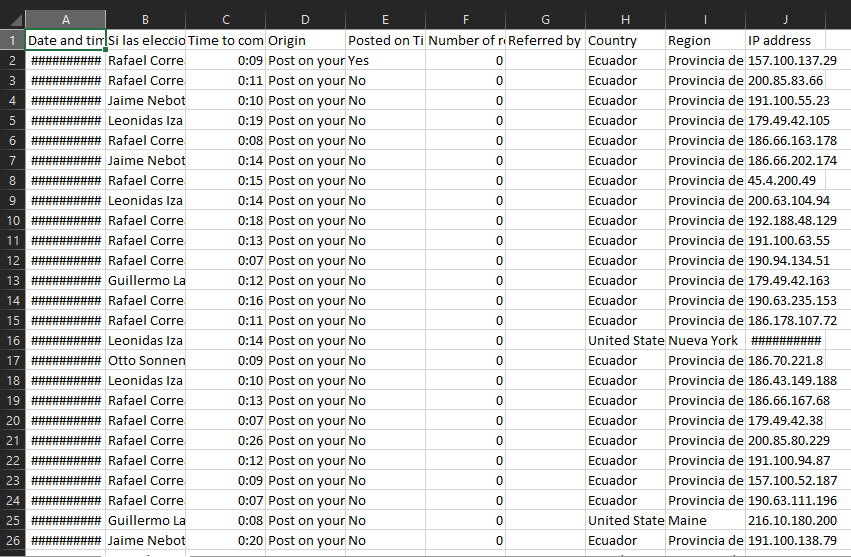


Figura 9. Encuesta de diario Mercurio.

Todos estos datos recolectados fueron llevados a ElasticSearch para su posterior análisis, esto se pudo realizar mediante Logstash, utilizando el archivo “couchdb.conf” como se muestra en la Figura 10, los cuales realizan el respectivo traslado de los datos para los diferentes archivos y bases obtenidas.

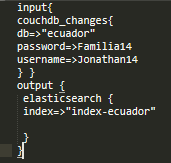


Figura 10. Estructura del archivo couchdb.conf.

Finalmente, El índice principal que es índex-ecuador que contiene todos los datos que se recolecto gracias al script de Python, en esta parte vamos a comparar los índex de los datos recolectados por Twitter y por la encuesta y sacar una conclusión que la redactaremos al final del documento.

* 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 11.

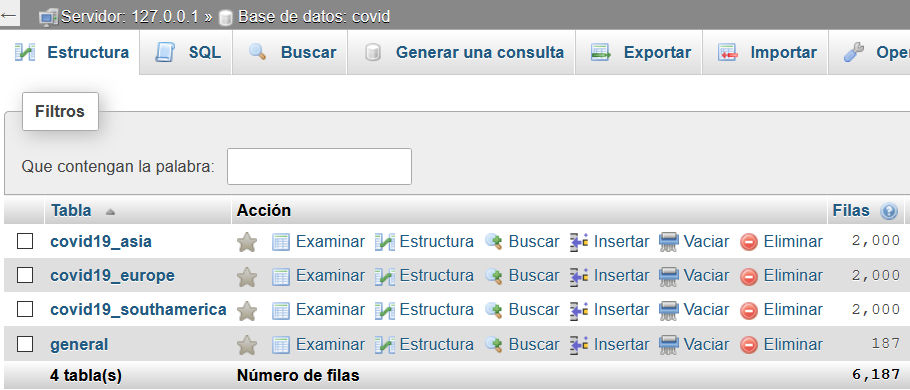


Figura 11. Base de Datos covid.

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.**

En la nube de Kibana se emplea la creación de un index pattern correspondiente a cada índice que creó en Elasticsearch para cada ciudad, de esta manera también se crea un index pattern especial que aglomera a todos los índices antes mencionados, para así, poder realizar la visualización de los datos correspondientes, y como filtro especial se usó el doc.date debido a que enlos índices se debe especificar un campo tipo date para la diferenciación

En cada tweet recopilado constan los mismos datos como muestra a continucación:

* 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)

Antes que nada, es necesario especificar que para poder tener esos parámetros o más bien que sean validados se tuvo que implementar un mapping como se muestra en la Figura 12 en la creación de cada índice en Elasticsearch, esto con la finalidad de que los datos recopilados cumplan con un formato general y así evitar incongruencia en los mismos.



Figura 12. Mapping para index en Elasticsearch.

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

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 como se muestra en la Figura 13:

* 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)

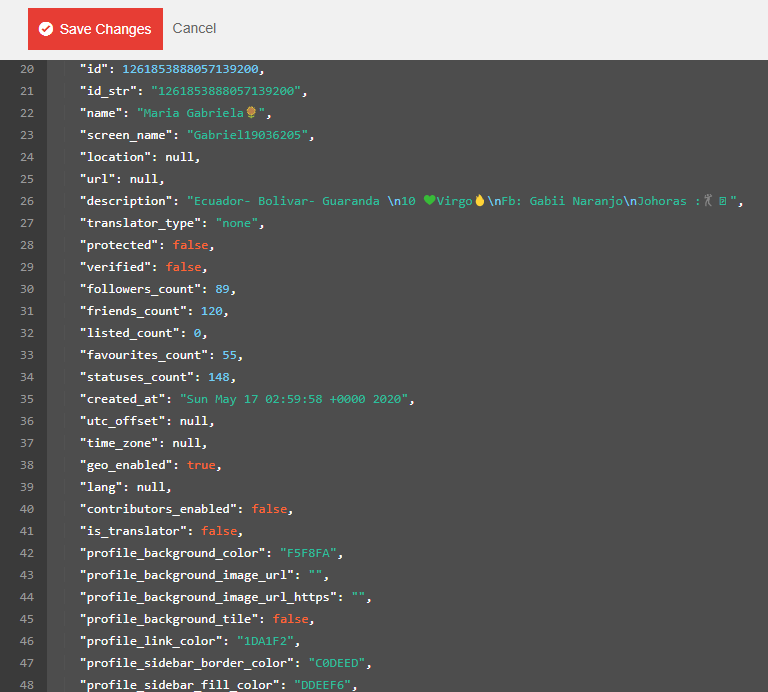


Figura 13. Resultado en CouchDb

Por otro lado, la estructura del dataset descargado de la encuesta está estructurado de las siguientes columnas, también se la puede observar en la Figura 14:

* 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.

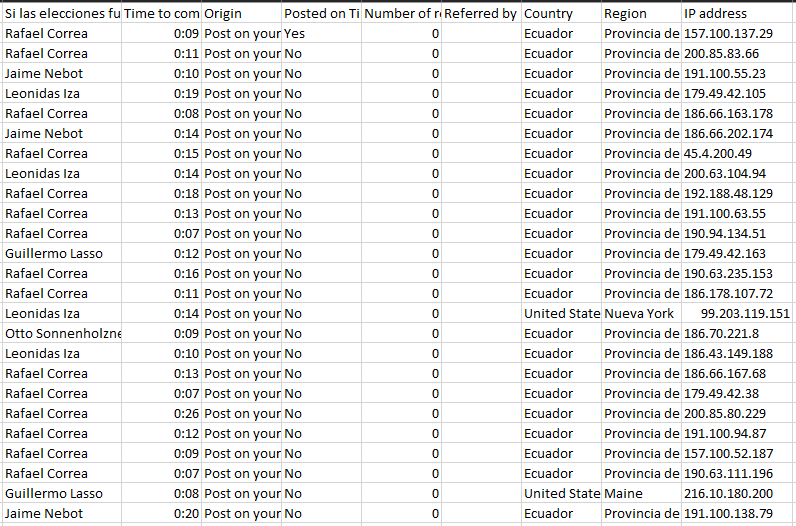


Figura 14. Datos de la encuesta.

* 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 12.

Como se muestra en la Figura 12, 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 la siguiente Figura 15, se muestra datos del covid-19 de todo el mundo, la tabla cuenta con los siguientes atributos:

* **Country**. - representa el nombre de los países del mundo.
* **Confirmed**. - número de casos confirmados.
* **Deaths.** - número de fallecidos.
* **Recovered**. - número de casos recuperados.
* **Active.** - número de casos con el virus activo.
* **New cases**. - número de casos nuevos.
* **New deaths**. - nuevos casos de fallecidos.
* **New recovered**. - nuevos casos recuperados
* **Deaths /100 Cases**. - numero de muertos de cada 100.
* **Recovered /100 Recovered**. - número de muertos de cada 100.
* **Confirmed last week.** – casos confirmados la semana pasada.
* **1 week change**. - cambios de casos entre semanas
* **1 week % increase.** - porcentaje de aumento en una semana.
* **WHO Region.** – qué región del país se realiza el análisis.

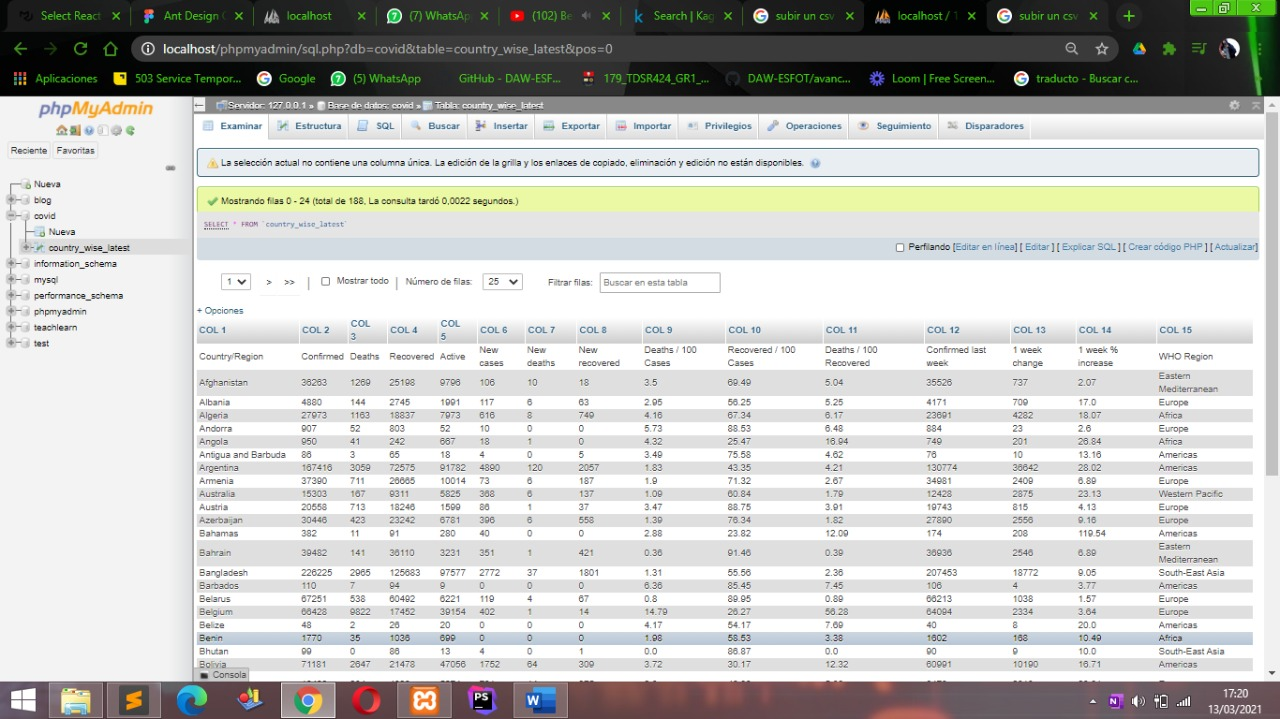


Figura 15. Tabla de estado del Covid en el mundo.

**Covid\_india**

A continuación, se muestra la Figura 16, donde se encuentran los datos del covid-19 en India, la tabla cuenta con los siguientes atributos:

* **Son.** - código único.
* **Date**. - fecha.
* **Tiempo**. - hora en la cual está hecho el análisis.
* **State/Union Territory**. - territorio de la región.
* **ConfirmedIndiaNational**. - casos confirmados de personas de nacionalidad India.
* **ConfirmedForeignNational.** - casos confirmados de personas de nacionalidad extranjera.
* **Cured**. - casos curados.
* **Deaths.** - casos de muerte
* **Confirmed**. - casos confirmados.

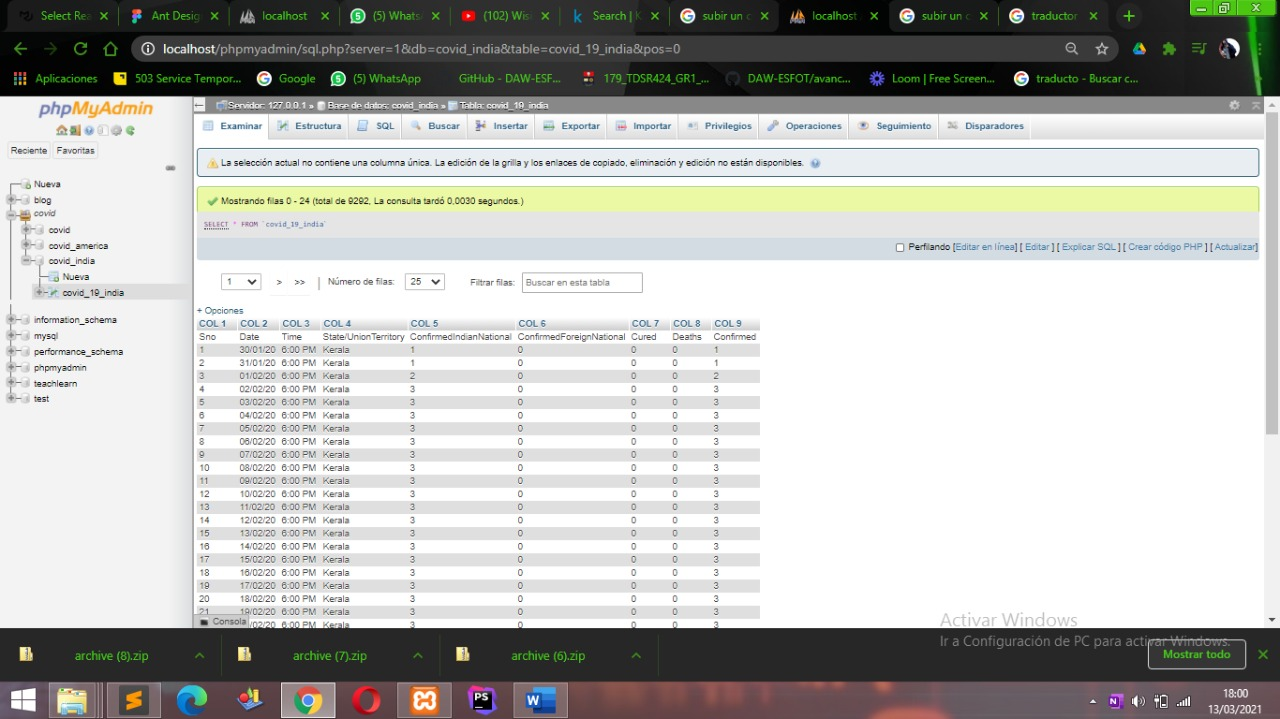


Figura 16. Tabla de Covid en India.

**Covid\_japon**

Como se muestra la Figura 17, donde se encuentran los datos del covid-19 en Japón, la tabla cuenta con los siguientes atributos:

* **Date. –** fecha
* **Location. –** ubicación
* **Positive. –** casos positivos
* **Tested. –** casos comprobados de covid.
* **Symptomatic. –** casos con síntomas de covid.
* **Asymptomatic. –** casos sin síntomas de covid.
* **Hosp\_mild. –** casos de hospitalización no severos.
* **Hosp\_severe. –** casos de hospitalización severos.
* **Hosp\_unknown. –** casos de hospitalización desconocidos.

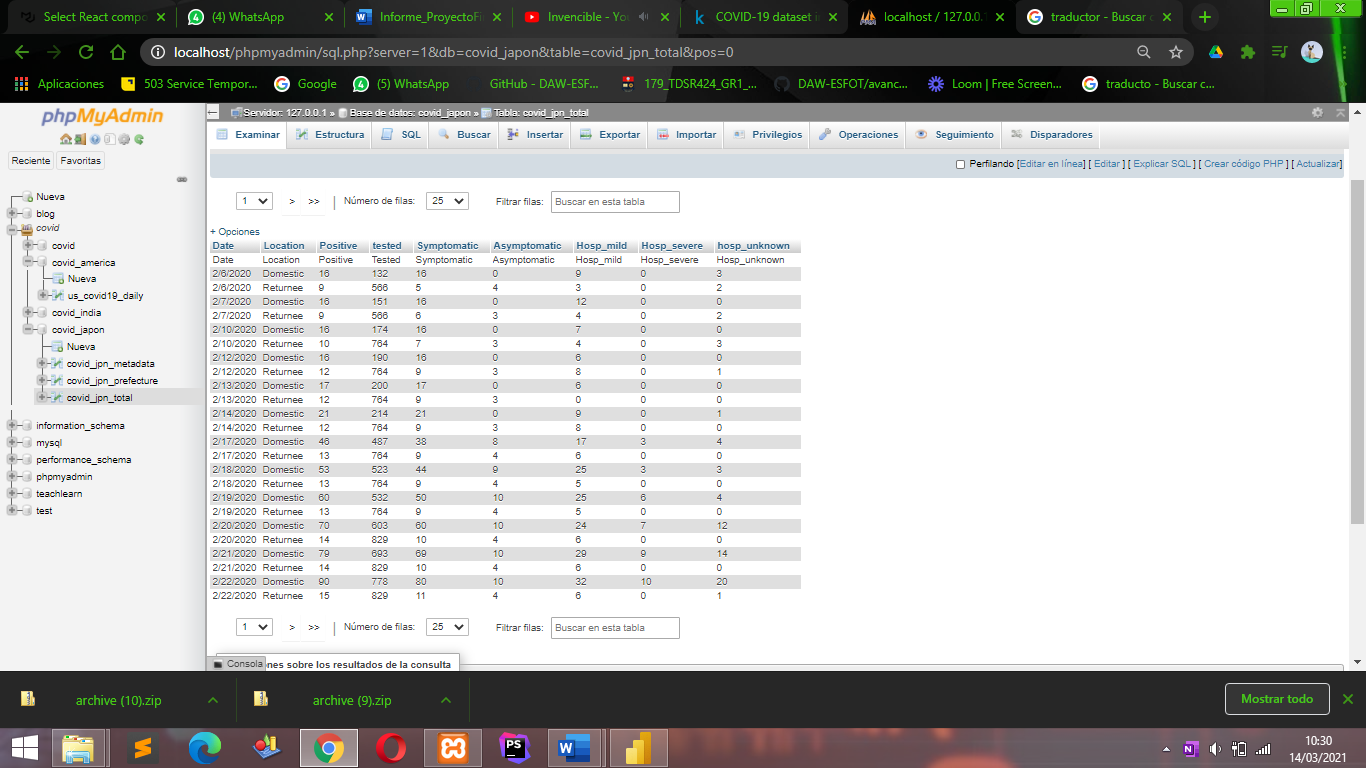


Figura 17. Tabla de Covid en Japón.

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

A continuación, se muestran las visualizaciones realizadas por medio de Kibana con respecto a la información que se recopiló:



Figura 18. Top personas que más twittearon acerca del tema.

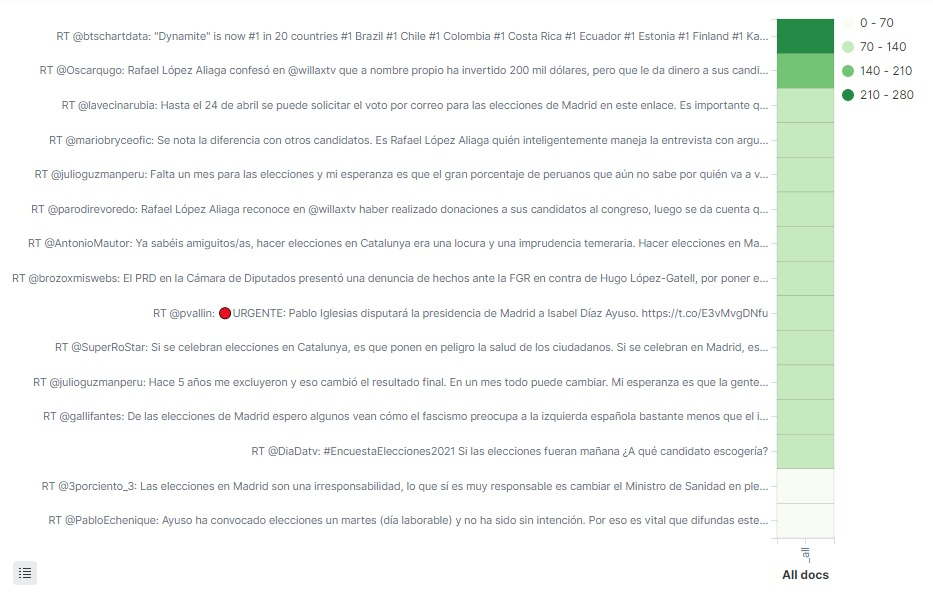


Figura 19. Tweets que contienen una descripción detallada sobre el tema buscado.

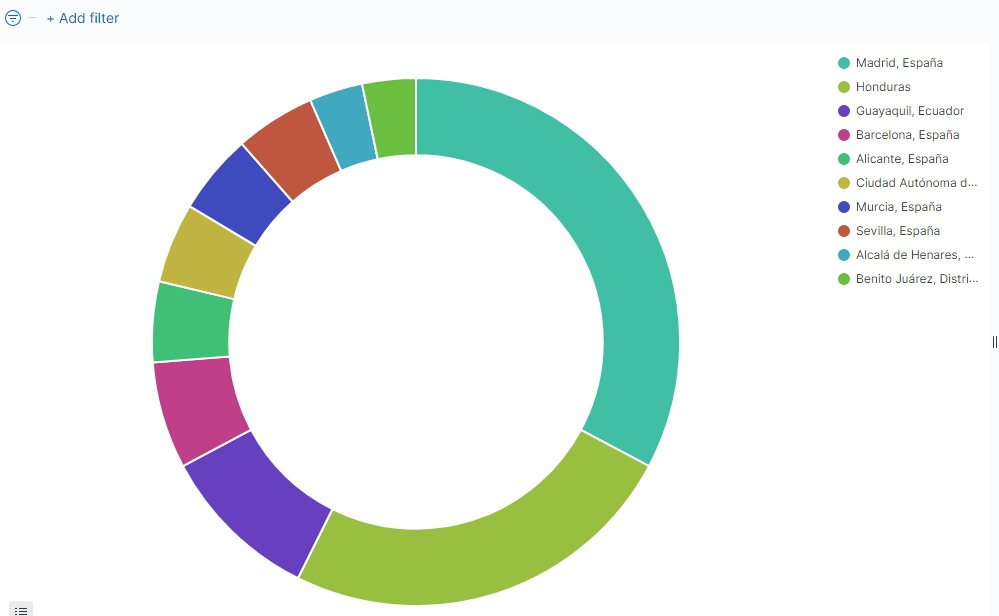


Figura 20. Muestra la ubicación desde dónde se hace el tweet.

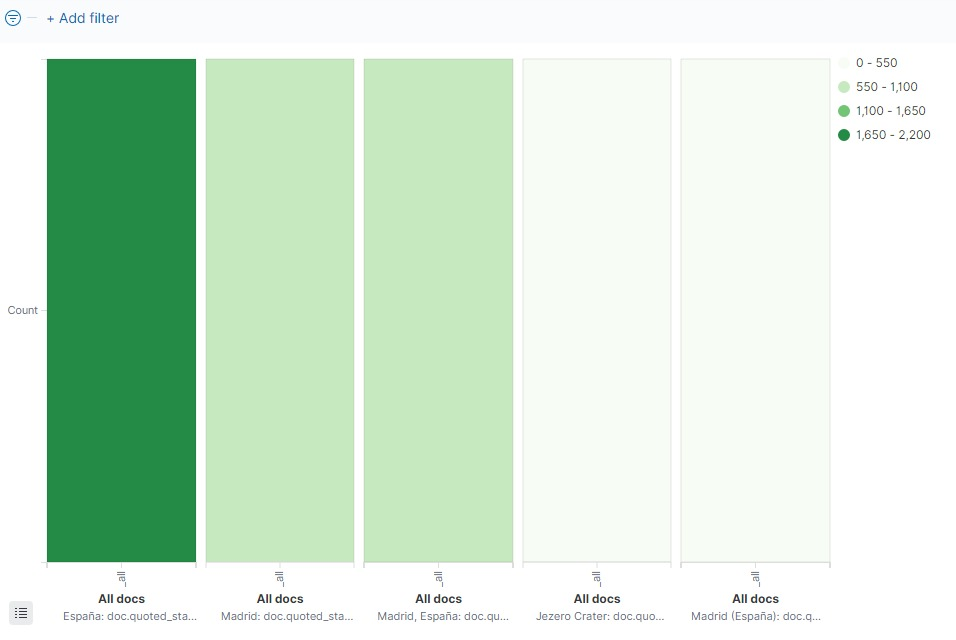


Figura 21. Cantidad de tweets por ubicación.



Figura 22. Revisión realizada por hora en Index-quito.

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

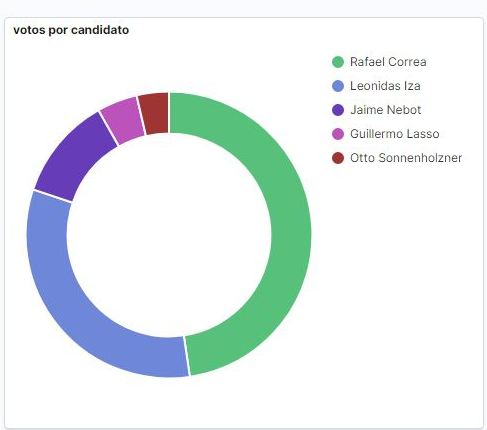


Figura 23. Total de votos por candidato.

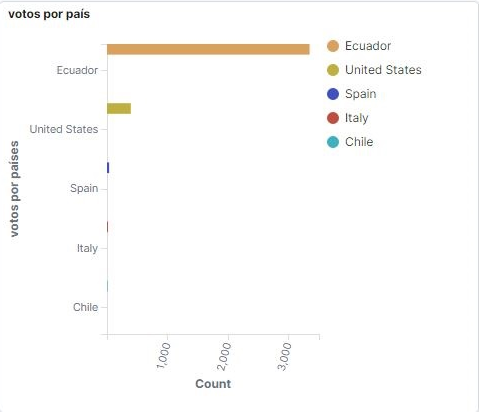


Figura 24. Votos generados por cada país.



Figura 25. Los hashtags más utilizados.

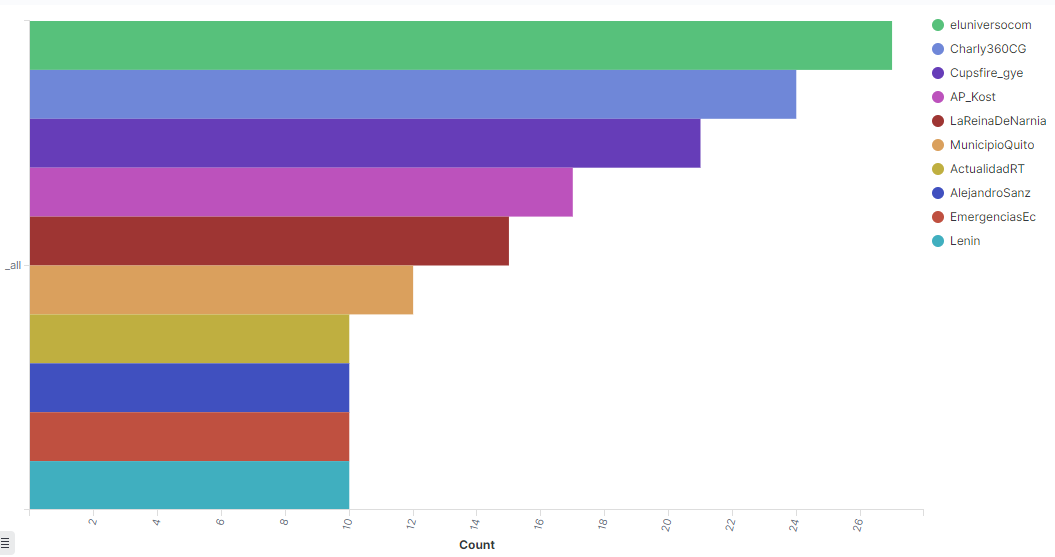


Figura 26. Top 10 usuarios que realizan más twitts.

* 1. **Juegos en línea por países.**

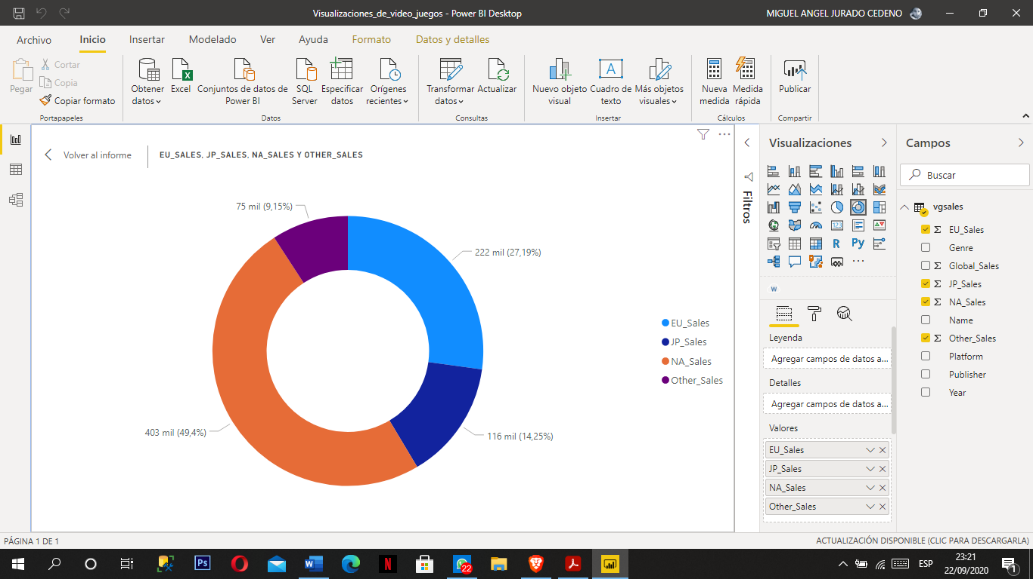


Figura 27: Ventas por país

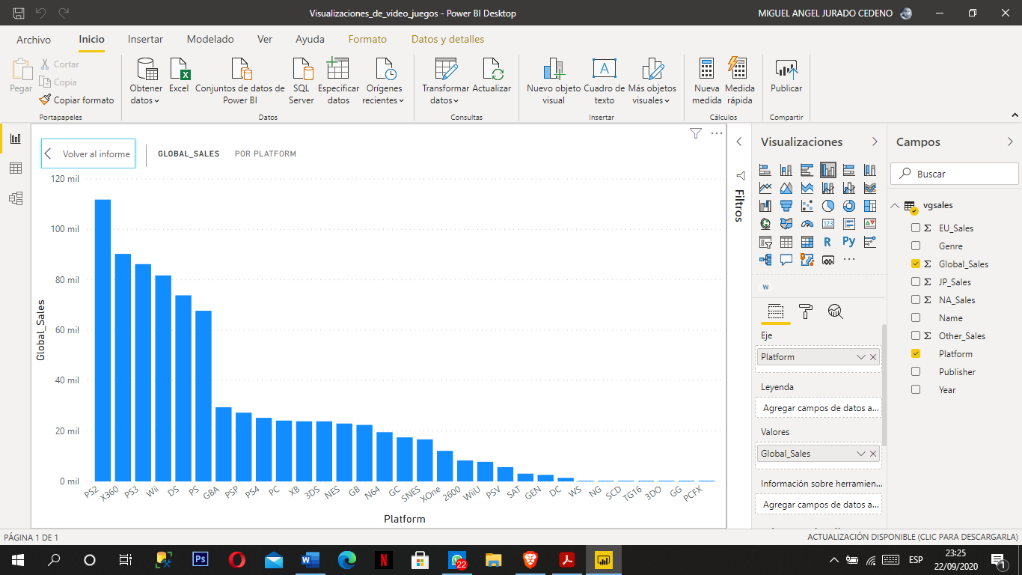
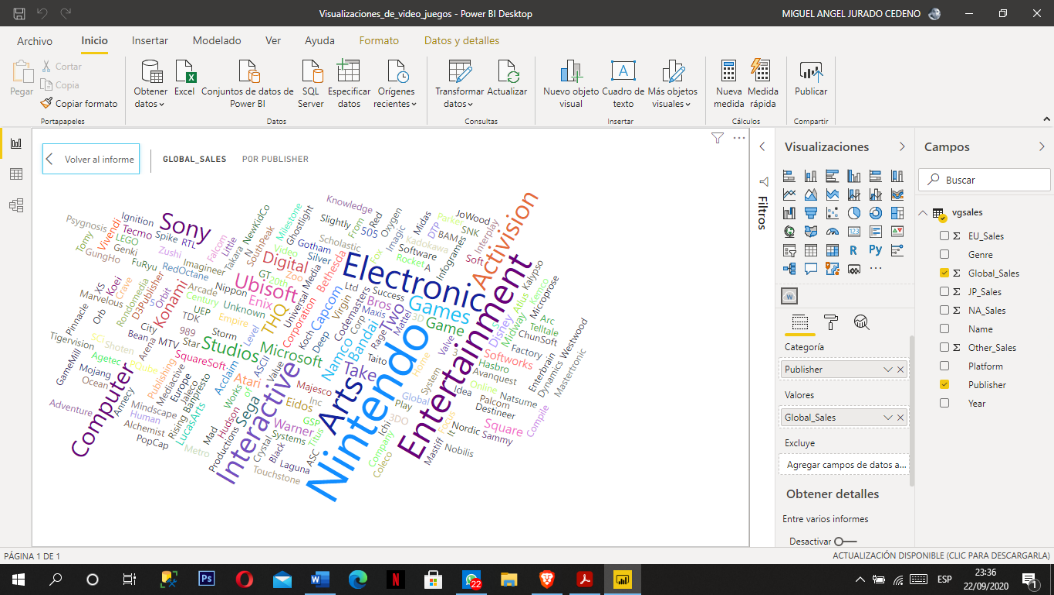


Figura 28: Ventas globales por plataforma



Figura 29: Código QR para acceder a visualizaciones



**Figura 30: Ventas globales por editor**

* 1. **Juegos en línea por países.**

**General**

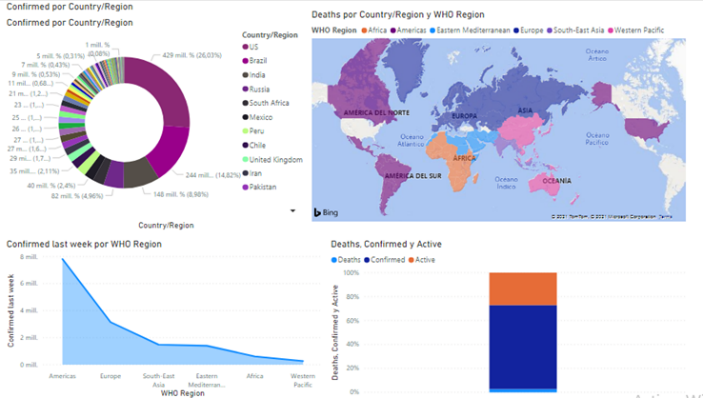


Figura 31 Dashbord del COVID en el mundo

**Covid\_india**

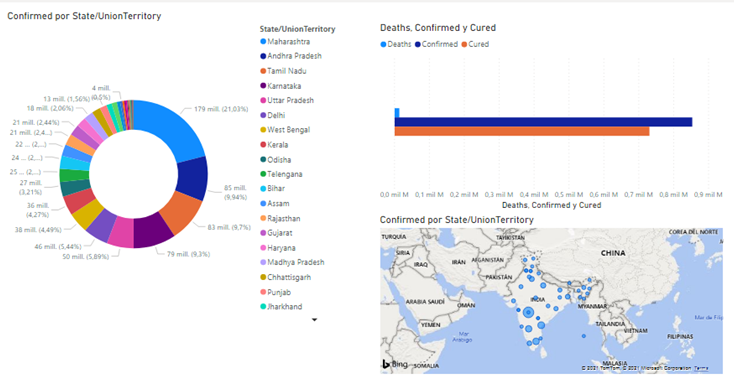


Figura 32 Dashbord del COVID en india

**Covid\_japan**

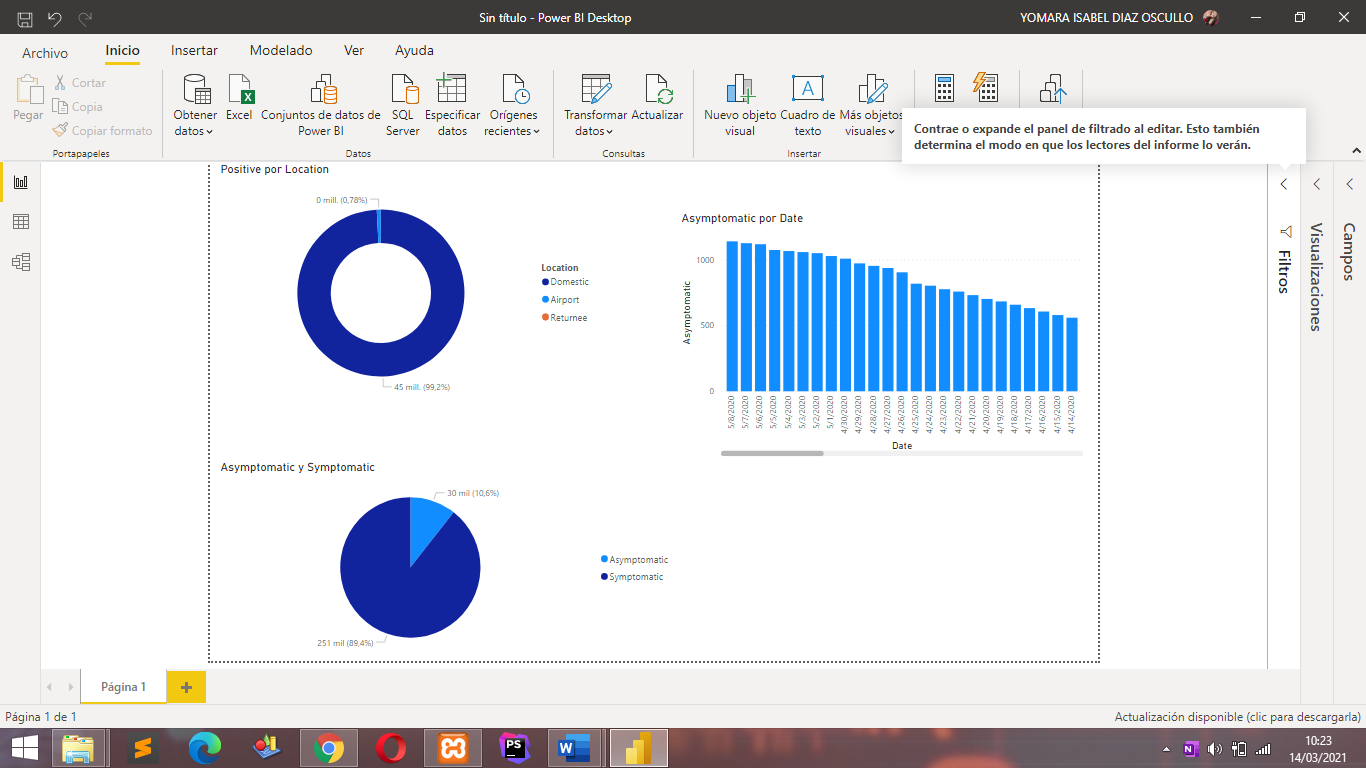


Figura 33 Dashbord del COVID en Japón

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

Con respecto a los resultados obtenidos no solamente por medio de las visualizaciones sino también con respecto a la revisión realizada se obtuvo que gran parte de ecuatorianos que se encuentran haciendo esidencia en el exterior demuestran un alto interés político por la situación del país y lo que conlleva este proceso de elección presidencial lo que permite desarrollar un análisis mucho más extenso, por lo cual en las gráficas obtenidas por medio del sondeo se demuestra la cantidad de tweets que se obtienen si se intenta tomar en cuenta un tema de interés o alguna problemática en sí por lo cual este tipo de análisis permiten entender la forma en la que se expresan los datos sea tomando en cuenta los diferentes tipos de acceso a la información, ya sea por medio de las redes sociales o por los medios de comunicación oficiales.

Sin lugar a duda se muestra la cantidad de interés mostrado por los usuarios en la plataforma tweeter.

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

**Total, de votos por candidato**

La **Figura 23** 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.

**Votos generados por cada país**

Los indicadores de la gráfica Figura 25 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.

**Los Hashtags más mencionados en Twitter**

En esta parte lo que hicimos es poner los 100 hashtag que las personas usan al momento de repostear un Twitter ganando la provincia de Guayas donde escribe cual es el posible candidato a presidente del Ecuador, en la **Figura 25** se muestra los resultados Obtenidos y el hash mas utilizado por los usuarios es Guayaquil, seguido de otros que son de diferentes provincias del Ecuador

**10 usuarios que realizan más twist**

La **Figura 26** muestra los 10 usuario que realizaron Retuitee que más mencionaron a los candidatos. Esta muestra una gran participación en algunas provincias del país. Por ejemplo, en Pichancha se recaudo 1500 datos en Guayaquil se recaudo 1750 datos y en Cuenca se recaudo 1300 los cuales nos ayudaron un gran análisis para saber de que provincia se realiza más Twists y hashtag postean en esa plataforma.

**Usuarios que realizaron más Hashtag.**

En este apartado realizamos que personas realizaban mas posteas en la plataforma de Twitter, aquí se procedió a poner un límite, ejemplo el top 10 de los usuarios que mas postearon a su candidato mas preferido y que sin duda piensa que va a ganar el. En la Figura 26 se muestra un grafico de barras donde mostramos los resultados obtenidos, cabe recalcar que los usuarios no son de una sola provincia sino de todo el Ecuador.

* 1. **Juegos en línea por países.**

**Ventas totales por género**

En la Figura 27 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**

En la Figura 31 se muestra el análisis del covid en el mundo, en la cual podemos identificar que Estados Unidos tiene el primer lugar en porcentaje de pacientes covid el cual es el 26.03%, como segundo país esta Brazil con el 14.82% y en tercer lugar India con el 8.98% de contagios. También se realizó el análisis de contagiados por continente el cual dio como resultado en primer lugar América (7815198 casos), seguida por Europa (3146816 casos) y en último lugar Western pacific (266359 casos). En cada uno de los continentes y por ende en cada país tenemos una cantidad de muertes, casos confirmados (activos y curados) y también casos activos de coronavirus.

**Covid\_japon**

En la Figura 32, se analizó los casos positivos de covid en tres localidades: Domestic (99.2%), airport (0.78%), return (0.02%), es decir la detección de casos mas grande se dio dentro del país. También se analizó los casos de covid con personas asintomáticas y con síntomas la cual lleva al análisis que existen más personas que presentan síntomas de covid con el 88.6% y el resto de las personas 9.25%, según los médicos es mejor presentar síntomas de covid para la detección de esta enfermedad y así poder combatirla. También se analizó en los meses de febrero, marzo, abril y mayo el aumento de personas asintomáticas, el aumento en la estadística realizada es totalmente notorio.

**Covid\_india**

En la Figura 33, nos muestra el porcentaje de casos contagiados de covid por cada región tomando el primer lugar Maharashtra (21.03%), segundo Andhra Pradesh (9.94%), tercer lugar Tamil Nadu (9,7%). Además, en cada región se cuenta con un porcentaje de muertos, contagiados confirmados y sin confirmar. Tenemos una gran cantidad de muerte en todo el territorio de India (13821481 casos)

1. **Conclusiones y recomendaciones**

* El acceso a la información como se hizo en el proyecto es netamente de análisis mas no se intenta evidenciar el interés del grupo por algún tipo de partido político, sin tomar en cuenta el sondeo se puede observar que las herramientas utilizadas tienen un alcance muy extenso en cuanto al análisis de datos se refiere.
* Es necesario realizar una limpieza en los datos debido a que por lo general los datasets que se encuentran en internet o en las diferentes fuentes de información, no contienen el formato correcto que se necesita para ser utilizados o exportados en sí.

1. **Desafíos y problemas encontrados**

Sin ligar a duda uno de los problemas más evidentes en la elaboración del proyecto sería el limitante de información que se encuentra en internet refiriéndose a herramientas que permitan hacer conexiones y análisis.

Otro punto a tomar en cuenta fue la creciente cantidad de datos que se fueron recolectando a lo largo del tiempo que conllevó el desarrollo del proyecto, puesto que utilizaron herramientas de almacenamiento masivo produjo una notoria limitante en los dispositivos, por lo tanto, hay que tomar en cuenta la capacidad de los equipos previo al análisis de datos, si es que se toma en cuenta grandes cantidades de información y herramientas utilizadas.

**REPO DE GITHUB CON EL PROYECTO:** https://github.com/Jonathan141999/proyecto-analisis-de-datos-JEGY