BASES DE DATOS MULTIDIMENSIONALES

2020

**ANÁLISIS DE INFORMACIÓN**

Calle Xavier, Gómez Samanta, Cristian Guamba, David Pisuña

Escuela Politécnica Nacional

Escuela de Formación de Tecnólogos

[xavier.calle@epn.edu.ec](mailto:xavier.calle@epn.edu.ec), [samanta.gomez@epn.edu.ec](mailto:samanta.gomez@epn.edu.ec), [cristian.guamba@epn.edu.ec](mailto:cristian.guamba@epn.edu.ec), [bryan.pisuna@epn.edu.ec](mailto:bryan.pisuna@epn.edu.ec)

1. **INTRODUCCIÓN**

En el presente documento se detalla el proceso de cosecha de datos, su transformación, mapeo, creación de índices, entre otros procesos para el análisis de datos de fuentes de Internet como facebook, twitter, webscrapping, tiktok, linkedin y además archivos estáticos de kaggle, INEC, etc.

1. **OBJETIVOS**
   1. **OBJETIVO GENERAL**

Obtener y analizar datos de diversas fuentes con el uso de varias herramientas usadas en clase.

* 1. **OBJETIVOS ESPECÍFICOS**

Diseñar una arquitectura de Data Lake en la cual tenga el insumo de datos.

Recopilar información con geolocalización o con filtro de palabras.

Realizar varias visualizaciones de datos para el análisis de los mismos.

Poner en practica todos los temas y herramientas aprendidas en clase durante el semestre en la materia de bases de datos multidimensionales.

1. **EQUIPO DE TRABAJO**

El equipo de trabajo está integrado por: Xavier Calle, Samanta Gómez, Cristian Guamba y David Pisuña.

Entre las actividades realizadas se organizó el equipo de la siguiente manera:

Xavier: Recolección de datos para el caso de estudio: juegos en línea.

Fuente de datos: Twitter.

Motor de base de datos: MySQL.

Samanta: Recolección de datos para el caso de estudio: Condiciones de vida-vivienda en el Ecuador

Fuente de datos: INEC.

Motor de base de datos: PostgreSQL.

Cristian: Recolección de datos para el caso de estudio: Pulso político en 20 ciudades principales de Ecuador, listas y candidatos, presidenciales y diputados. Pulso político por provincias en Ecuador, listas y candidatos, presidenciales y diputados.

Fuente de datos: Twitter.

Motor de base de datos: CouchDB.

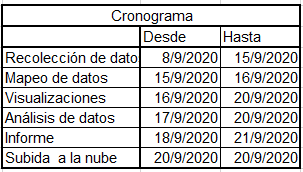
David: Recolección de datos para el caso de estudio: Tendencia actual – Red Social

Fuente de datos: Tik Tok.

Motor de base de datos: MongoDB.

1. **CRONOGRAMA**

Tabla 1 Cronograma por cada miembro del grupo



1. **RECURSOS Y HERRAMIENTAS**

**MYSQL**

MySQL es un sistema de gestión de bases de datos relacionales de código abierto (RDBMS, por sus siglas en inglés) con un modelo cliente-servidor. RDBMS es un software o servicio utilizado para crear y administrar bases de datos basadas en un modelo relacional. [1]

**POSTGRESQL**

PostgreSQL es una de las opciones más interesantes en bases de datos relacionales open-source. Michael Stonebraker inició el proyecto bajo el nombre Post Ingres a mediados de los 80’s con la idea de solucionar problemas existentes en las bases de datos en esa época. MySQL fue por mucho tiempo el motor más popular; pero hoy es propiedad de Oracle y esto limita su evolución. [2]

Es gratuito y libre, además de que hoy nos ofrece una gran cantidad de opciones avanzadas. De hecho, es considerado el motor de base de datos más avanzado en la actualidad. (Y Platzi tiene un Curso de PostgreSQL)

Una característica interesante de PostgreSQL es el control de concurrencias multiversión; o MVCC por sus siglas en inglés. Este método agrega una imagen del estado de la base de datos a cada transacción. Esto nos permite hacer transacciones eventualmente consistentes, ofreciéndonos grandes ventajas en el rendimiento. [2]

**COUCHDB**

Es una base de datos NoSQL capaz de replicarse en una amplia gama de entornos cliente y servidor. Por sus respectivas características, el Cloud Computing es el ambiente de funcionamiento más natural de CouchDB, otra de las aplicaciones Open Source que podemos instalar y desplegar sobre los Servidores Cloud de Arsys en sólo unos minutos. [3]

CouchDB es un sistema gestor de bases de datos no relacional, basado completamente en tecnologías sobre las que está implementada la web, como HTTP, JSON o Javascript. Es una base de datos distribuida, capaz de adaptarse tanto a ambientes de servidor como a clientes de diversos tipos. Se trata de un proyecto Open Source de la Fundación Apache pensado especialmente para usarse como infraestructura en aplicaciones web. Su principal característica es la definición del Couch Replication Protocol, un sistema que permite que una variedad de productos pueda comunicarse entre sí, compartiendo datos y realizando una sincronización de estos de manera automática. [3]

**MONGODB**

Es una base de datos orientada a documentos. Esto quiere decir que, en lugar de guardar los datos en registros, guarda los datos en documentos. Estos documentos son almacenados en BSON, que es una representación binaria de JSON. [4]

Una de las diferencias más importantes con respecto a las bases de datos relacionales, es que no es necesario seguir un esquema. Los documentos de una misma colección - concepto similar a una tabla de una base de datos relacional, pueden tener esquemas diferentes. [4]

**ELASTICSEARCH**

Se trata de un motor de búsqueda y análisis. Es distribuible y fácilmente escalable, enfocado sobre todo al mundo empresarial y científico. Es accesible a través de una extensa y elaborada API. Con esta herramienta podemos impulsar búsquedas extremadamente rápidas que respalden nuestras aplicaciones de descubrimientos de datos. [5]

Está escrito en Java, utilizando la filosofía de desarrollo de código abierto y utiliza una licencia Apache, similar a la de software libre.

Trabajo juntamente con un motor de recopilación de datos llamado Logtash, y una plataforma de análisis y visualización llamada Kibana. Los tres productos están diseñados para trabajar de manera conjunta, una solución llamada “Elastic Stack” (anteriormente llamada ELK) [5]

Sus características:

* Es orientado a documentos: Utiliza JSON
* No utiliza esquemas, aunque si son necesarios se pueden llegar a definir
* Distribuido: Realiza escalado de manera dinámica, implementa alta disponibilidad (HA)
* Utiliza una potente API: expone prácticamente todas sus funcionalidades utiliza una API REST
* Permite búsquedas tanto estructuras como no estructuradas [5]

**KIBANA**

Kibana es una herramienta open-source perteneciente a Elastic, que nos permite visualizar y explorar datos que se encuentran indexados en ElasticSearch, es decir, un plugin de ElasticSearch. Kibana también es conocido por el stack ELK:

* Elasticsearch
* Logstash
* Kibana [6]

**CEREBRO**

Cerebro es un entorno de trabajo colaborativo adecuado para proyectos de complejidad ilimitada. Esta es una herramienta para la planificación, distribución de tareas y seguimiento de la ejecución. Ofrece características únicas en términos de almacenamiento e intercambio de datos, así como anotaciones visuales para cualquier material de trabajo. [7]

Optimice el desarrollo de un extremo a otro, desde la creación de activos hasta los comentarios de los clientes. Revise el estado del proyecto, asigne trabajos, distribuya el trabajo, tan fácilmente como si estuviera en una sala de reuniones con su equipo. Cree nuevos proyectos con facilidad y calcule rápidamente tanto el plazo como la carga de trabajo por empleado. [7]

**LOAGSTASH**

Es una herramienta Open Source que nos permite centralizar la recogida de información, normalizarla y redistribuirla. [8]

Aunque es una herramienta que puede funcionar de forma independiente, como parte del Stack vemos que Logstash tiene total integración con éste. Su integración con Beats, Elasticsearch y Kibana es nativa, ofreciendo así gran valor para la transformación avanzada de nuestra información. [8]

**TWITTER**

Esta plataforma social, es un servicio de comunicación bidireccional con el que puedes compartir información de diverso tipo de una forma rápida, sencilla y gratuita, se trata de una de las redes de microblogging más populares que existen en la actualidad y su éxito reside en el envío de mensajes cortos llamados “tweets”. [9]

**INEC**

Instituto Nacional de Estadística y Censos fue creado mediante la Ley No. 7839 del Sistema Estadístico Nacional (SEN), como Institución Autónoma de derecho público, con personería jurídica y patrimonio propios, gozando de autonomía funcional y administrativa consagrada en el artículo No 188 de la Constitución Política. [10]

**TIK TOK**

Hablamos de una red social asiática basada en compartir vídeos musicales, y que está consiguiendo unos increíbles resultados en los últimos meses. De hecho, este pasado mes de octubre ha conseguido superar a Facebook, Instagram, YouTube y Snapchat en número de descargas confirmándose como una de las sorpresas del año. [11]

1. **ARQUITECTURA DE LA SOLUCIÓN**

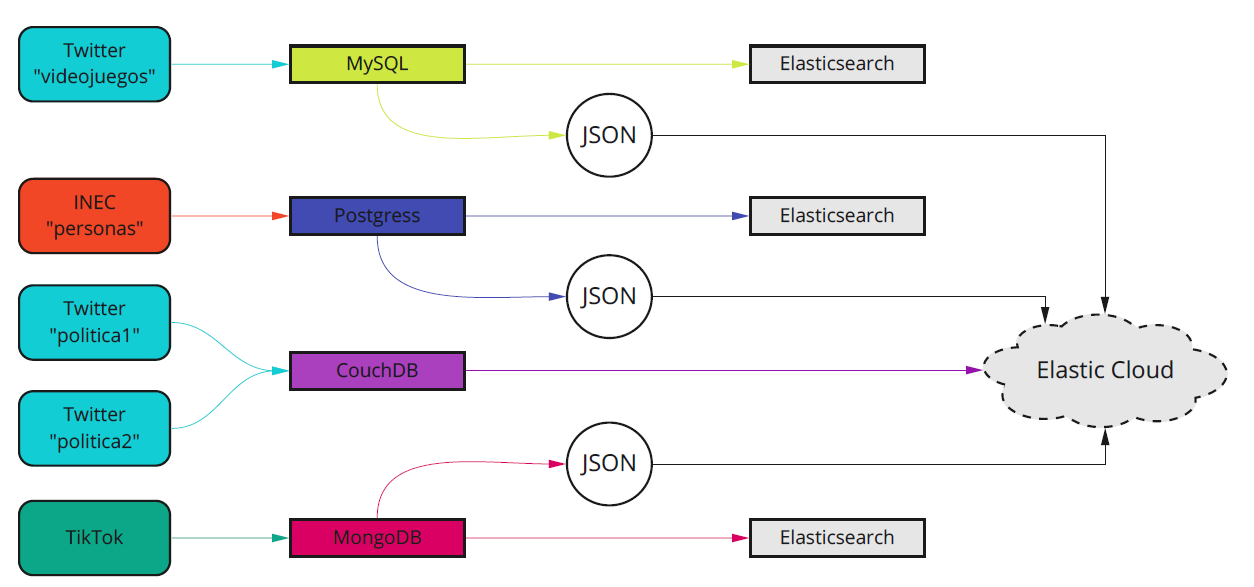


Figura 1: Arquitectura

1. **EXTRACCIÓN DE DATOS**

**Caso de estudio:** Pulso político en Ecuador

La extracción de datos se realizó desde Twitter mediante dos scripts de Python.  
El primero utilizó como criterio de búsqueda los nombres de los precandidatos a presidencia y vicepresidencia, y sus respectivos partidos políticos. Además, se añadieron los hashtags y nombres de usuario en Twitter.

El segundo script incluyó las coordenadas del territorio ecuatoriano extraídas del sitio web; <https://boundingbox.klokantech.com/>

Los datos fueron almacenados en CouchDB, y mediante Logstash, se enviaron en tiempo real a Elastic cloud.

Los datos enviados desde CouchDB hasta Elastic cloud, se unifican en mismo índice de nombre map\_twitter.

**Caso de estudio**: Juegos en linea

La extracción de datos se dio desde la fuente: Twitter

La extracción de datos se realizó desde Twitter utilizando un script de Python y un archivo de configuración.

Primeramente, se realizó la recolección de datos con el script de Python que se detalla a continuación:

Importación de librerías:

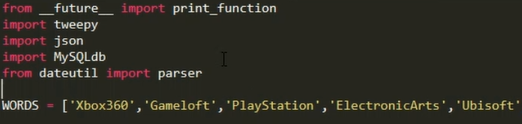


Figura 2:Importación de librerías y palabras clave

Registros de la base de datos:

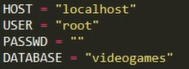


Figura 3: Base de datos

Transformación de datos

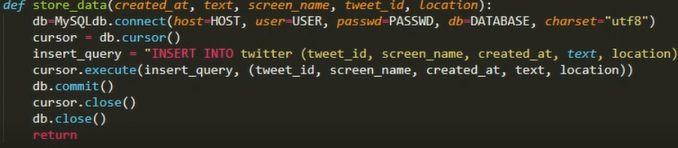


Figura 4: Transformación de datos

Siguiendo el diseño de la arquitectura se utilizó un archivo de configuración de Logstash para pasar los datos de MySQL a ElasticSearch. Cabe recalcar el uso del controlador jdbc para pasar los datos desde MySQL a ElasticSearch.

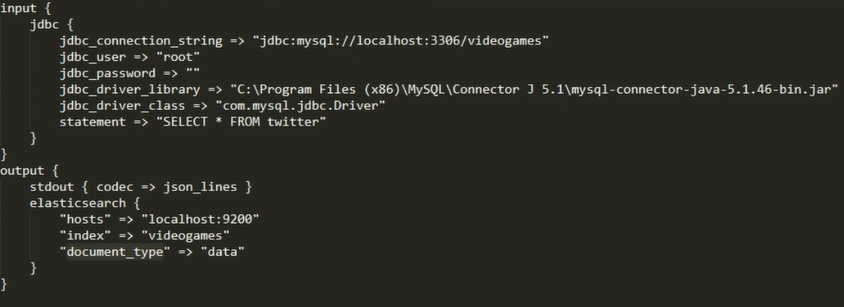


Figura 5: MySQL-ElasticSearch

**Caso de estudio:** Condiciones de vida-vivienda en el Ecuador

La extracción de datos se dio desde la fuente: INEC



Figura 6: INEC

Obteniendo un archivo .csv con la información que se analizará.

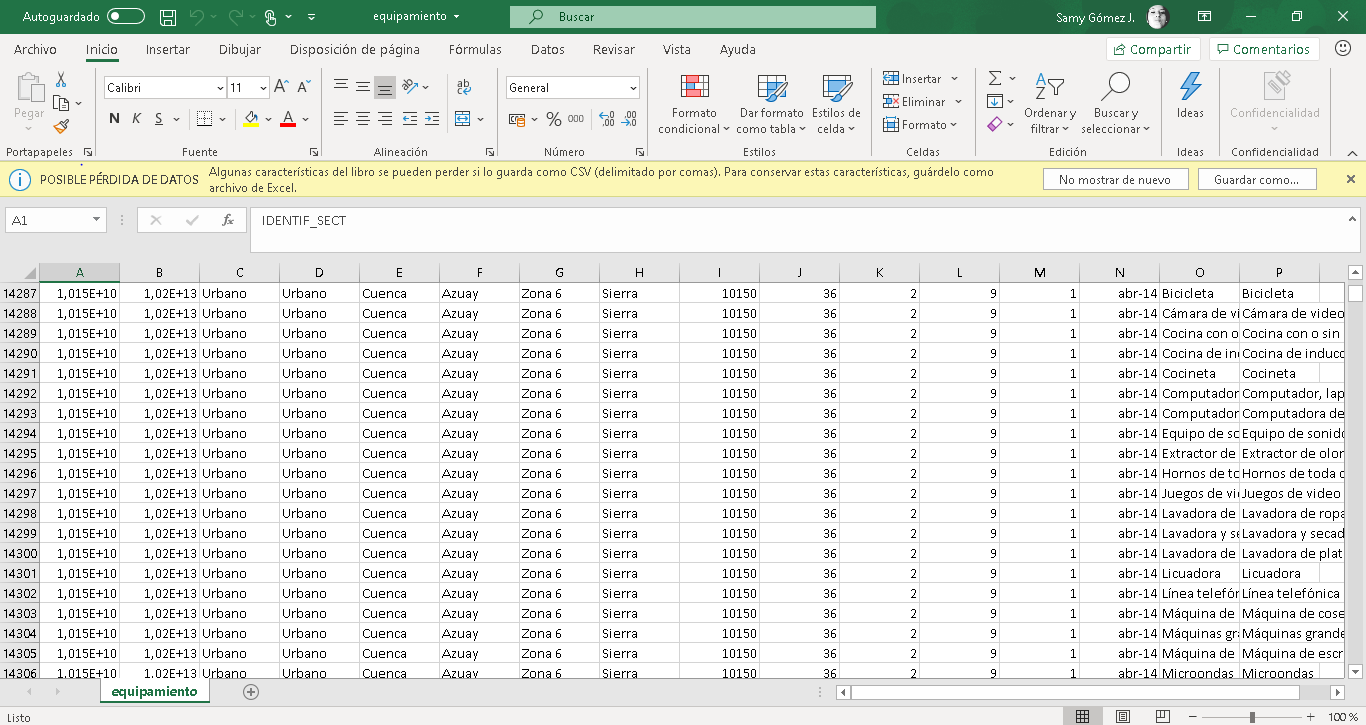


Figura 7:csv Cond\_vida\_vivienda

**Caso de estudio**: Tendencia actual – Red Social

La extracción de datos se dio desde la fuente: Tik Tok

Obteniendo un archivo .csv con la información que se analizará.

Desde el paso de obtención de datos, se estuvieron descargando los archivos desde Python y luego enviando las consultas desde CouchDB hasta Elastc cloud. Razón por la que los datos ya se encontraron en el punto de unificación de los resultados obtenidos, desde un principio.

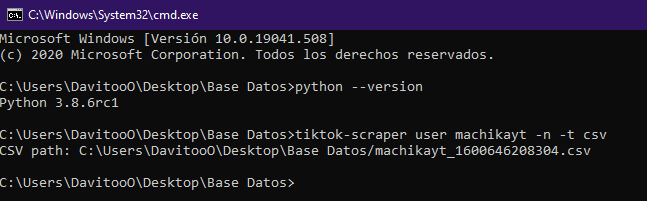


Figura 8: Comandos y verificación de Python para la extracción de datos

1. **ANALISIS DE DATOS Y DESARROLLO**

**Caso de estudio:** Pulso político en Ecuador

Para el análisis de datos se usó la herramienta Elasticsearch en la nube.  
Para un correcto formato de los datos, se realizó el mapeo de los atributos de tipo fecha y geolocalización en las propiedades del índice para los datos de Twitter.

De la totalidad de datos que se consiguieron solo el 44,21% tuvo la geolocalización activa.

Para filtrar los datos se crearon varias consultas para clasificar los datos por localidades y por partidos políticos.

Las subconsultas se realizaron en la herramienta Kibana de Elastic clud.

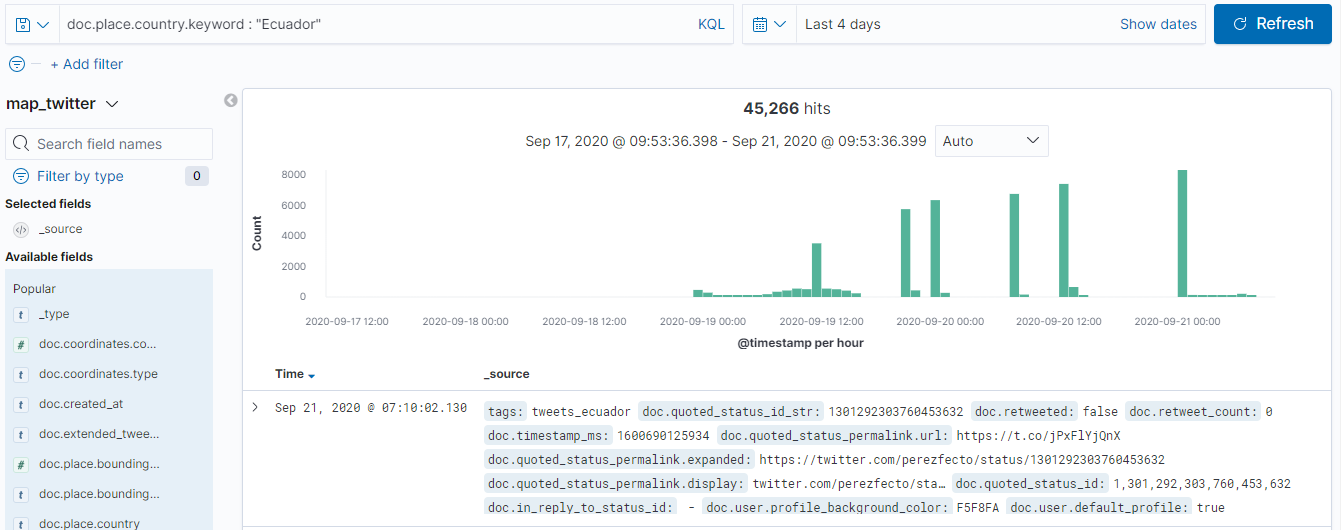


Figura 9: Filtrado de resultados en Kibana

Para el patrón de indexación se utilizó el atributo @timestamp, que indica la fecha y hora a la que se publicó el tweet, la misma a la que se subió el tweet a elastic search.

**Caso de estudio**: Juegos en línea

Los datos recolectados fueron: identificador de twitter, nombre de usuario, texto del tweet y localización como se indica en la figura 4 donde definimos que campos se van a guardar en la base de datos y el tipo de datos.

Se utilizó la herramienta cerebro para visualizar los datos recolectados en elastic search.

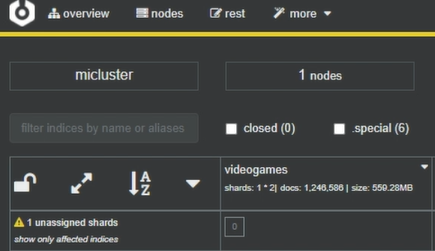


Figura 10: Visualización de resultados en cerebro

**Caso de estudio:** Condiciones de vida-vivienda en el Ecuador

El instituto nacional de estadísticas y censos realizó una encuesta nacional de condiciones de vida en todas las provincias del país dando un aproximado de 1040000 millones de datos.

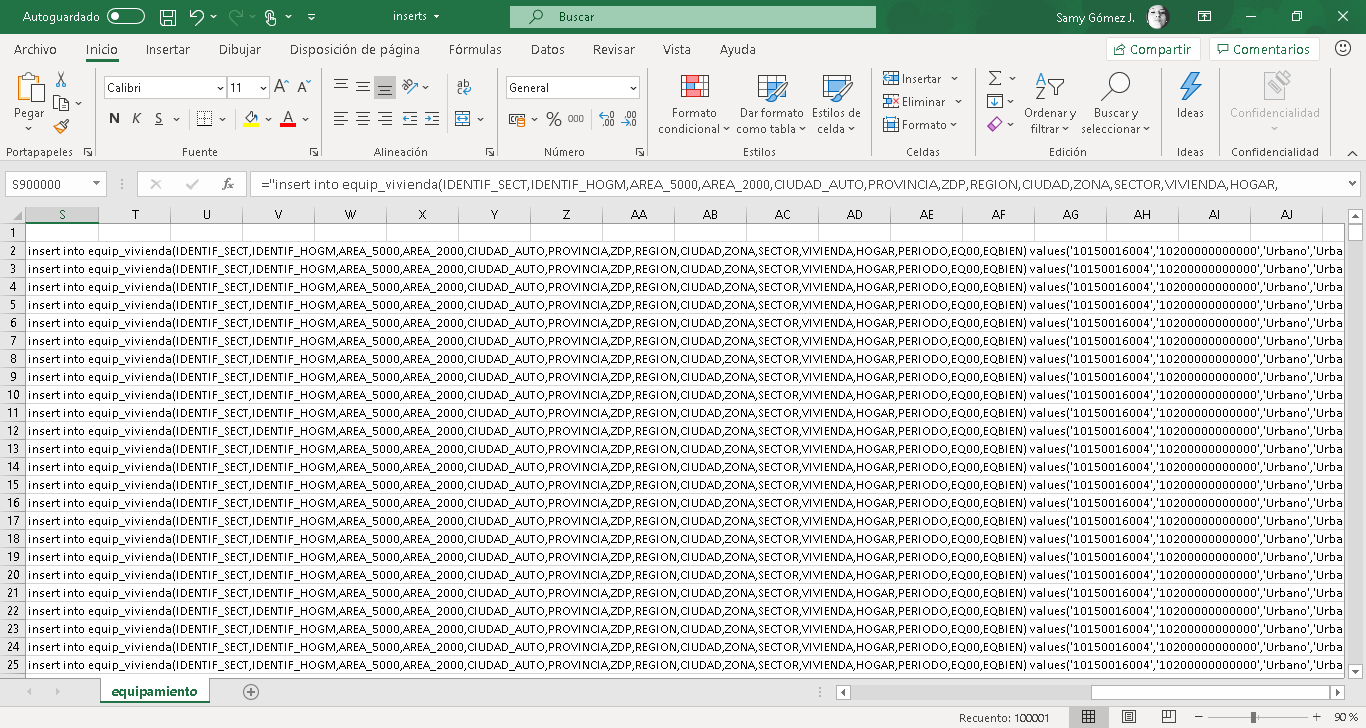


Figura 11: Insert\_cond\_vida

Se realiza la población de la base de datos, ésta puede realizarse de dos maneras ya sea por el método de importación de un archivo csv como se se muestra en la siguiente figura:

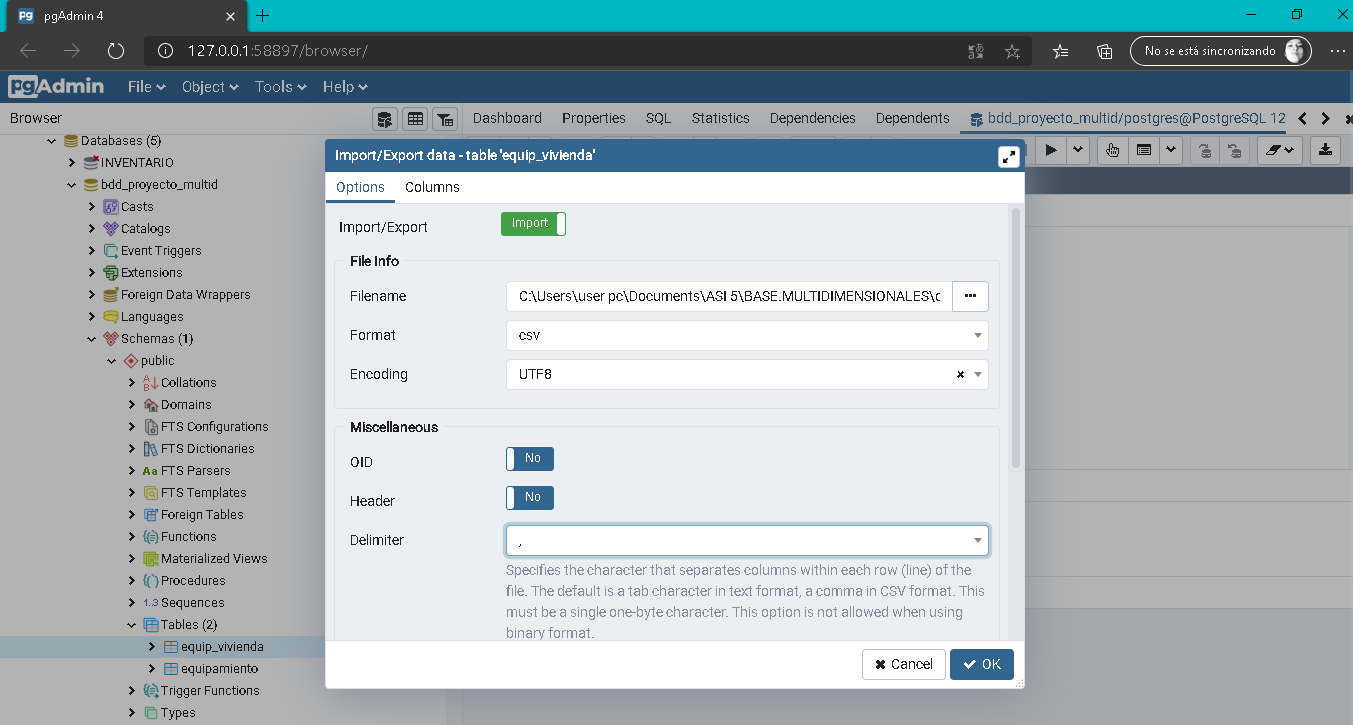


Figura12: Importación de archivo .csv

O a su vez por medio de fórmula se puede realizar las líneas de insertar en el documento .csv para luego copiar en un script y al correr ingresan los datos.

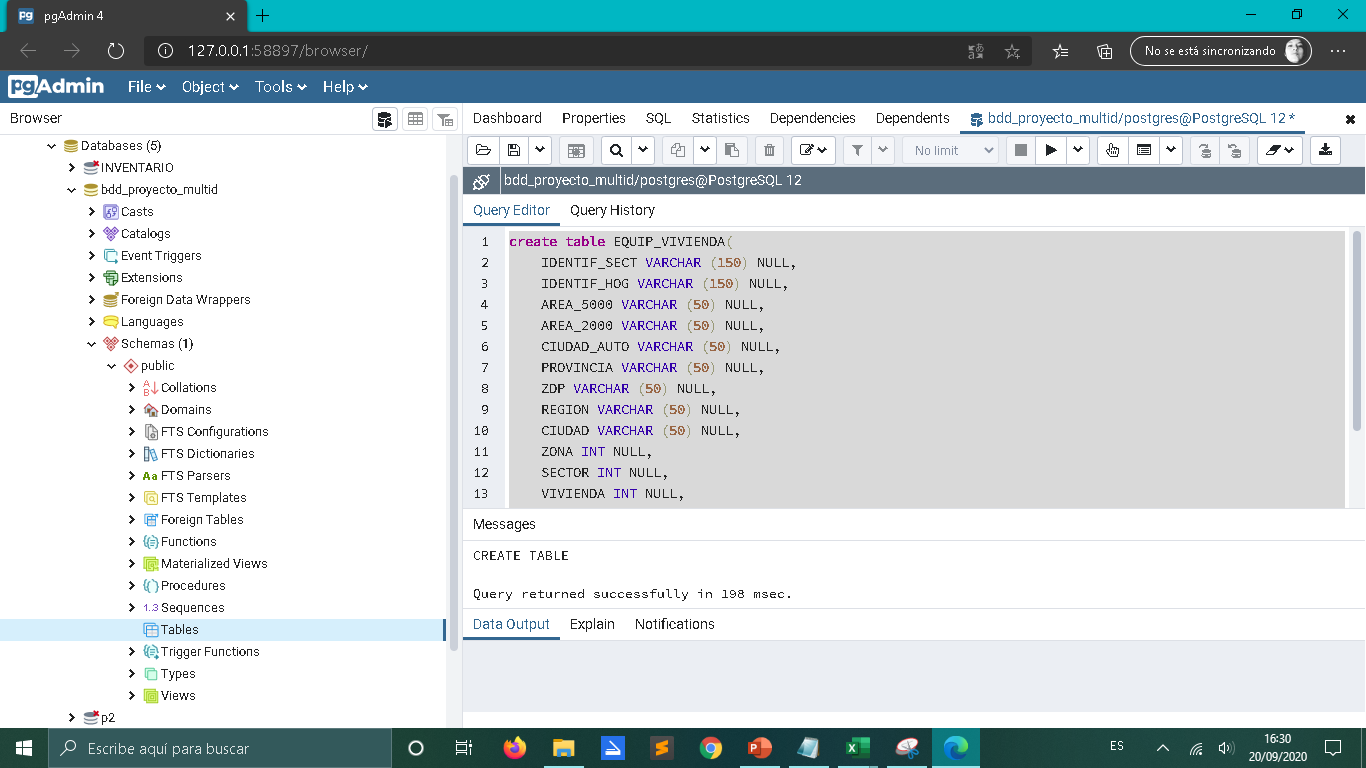


Figura 13 : Creación de BDD y tabla equip\_vivienda

Después de haber creado la base de datos se procede a crear una tabla la misma que va a hacer llenada de inmediato.

Al finalizar la población de la base de datos se puede verificar por medio de una consulta como se muestra en la siguiente figura.

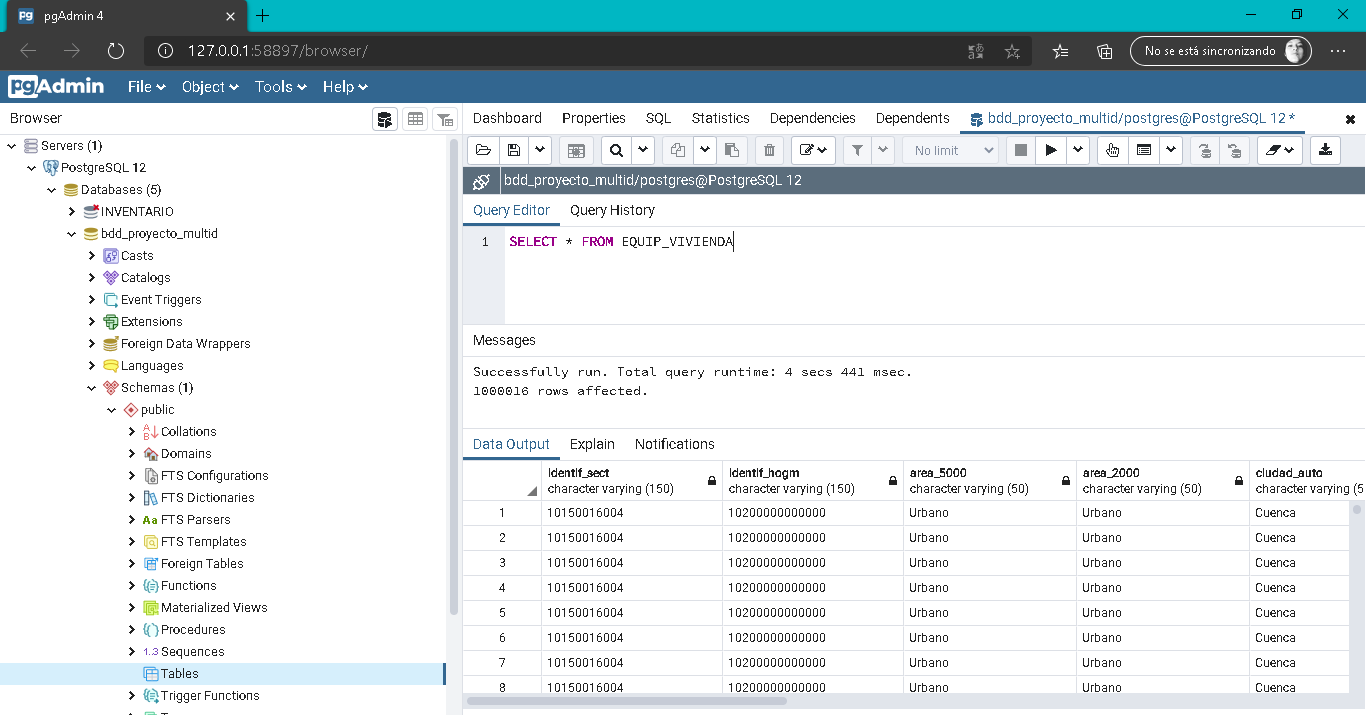


Figura 14: Población BDD en Postgresql

**Caso de estudio**: Tendencia actual – Red Social Tik-Tok

Para el análisis de datos se usó la herramienta MongoDb en la nube.  
Para un correcto formato de los datos, se realizó una cosecha de datos respecto a las personas más seguidas en Tik-Tok

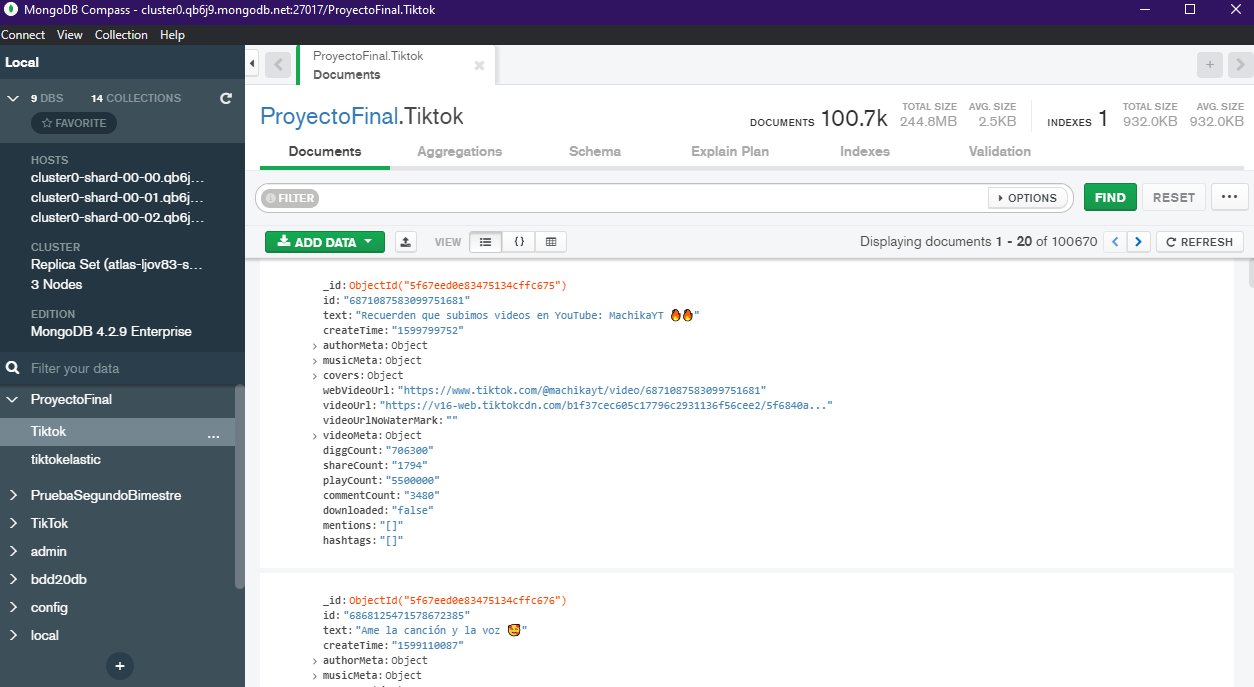


Figura 15: Archivos subidos en MongoDb

1. **VISUALIZACION DE INFORMACIÓN**

**Caso de estudio:** Pulso político en Ecuador

Para la visualización de información y la creación de consultas se usó kibana.  
Mediante la opción Discover, se analizarón resultados previos basados en grupos pequeños de datos. Estas consultas trabajan con datos limitados (500) para mostrar resultados de tendencias.

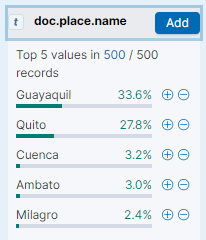


Figura 16: Previsualización en Kibana

Estos resultados se usaron para definir los atributos críticos de los tweets recabados, para poder generar las visualizaciones.

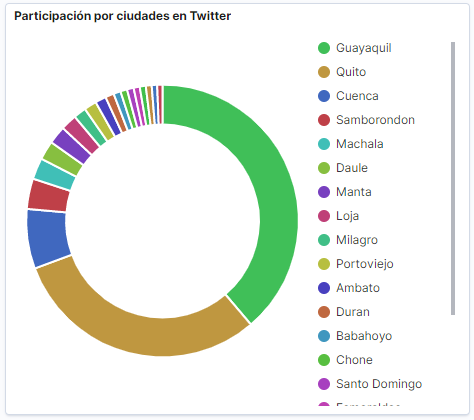


Figura 17: Visualización ciudades en Twitter

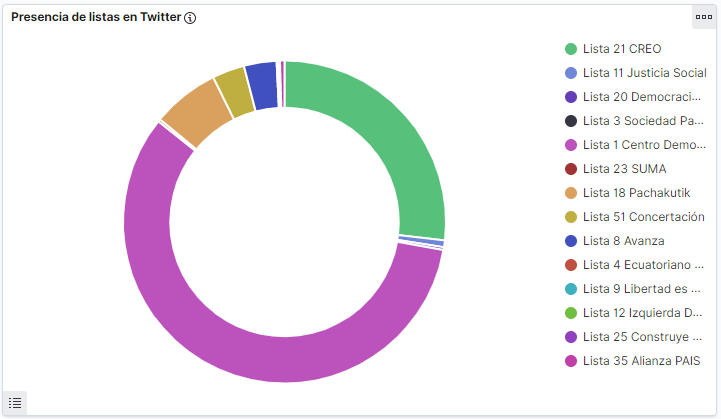


Figura 18: Presencia de listas en Twitter

**Caso de estudio**: Juegos en línea

Se utilizó Kibana para la visualización de resultados.

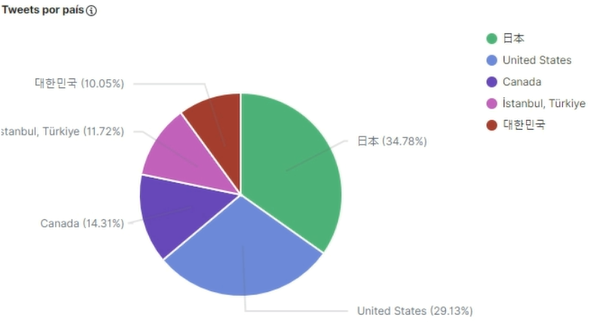


Figura 19: Tweets sobre videojuegos por país

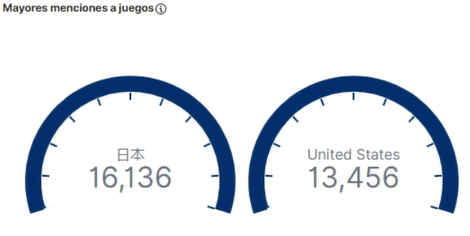


Figura 20: Países top en menciones de videojuegos

**Caso de estudio:** Condiciones de vida-vivienda en el Ecuador

Se requiere configurar un archivo para ejecutar con logstash, previamente ya iniciado cerebro, elasticsearch y kibana para ingresar el archivo json y poder ver algunas tablas como visualizaciones.

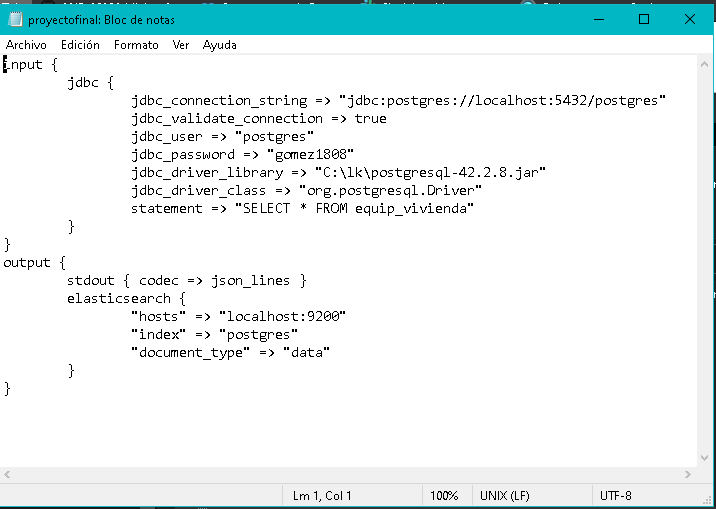


Figura 21: Configuración Postgresql-Elasticsearch

Inicia el proceso de carga del archivo. Json.

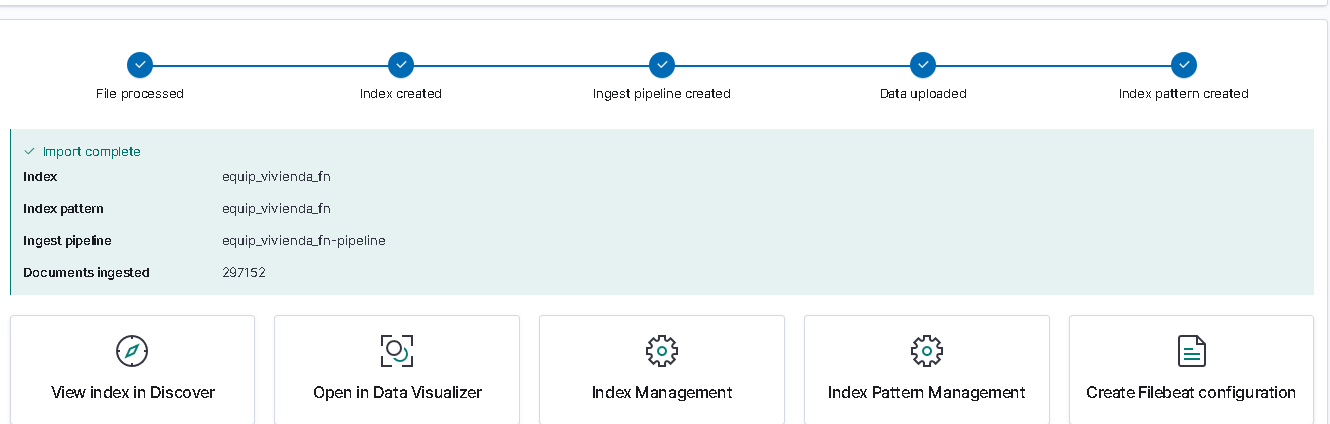


Figura 22: Carga de datos

Observamos varias vistas con diferente orden de tatos para interpretar de una forma mas práctica y facil la información.

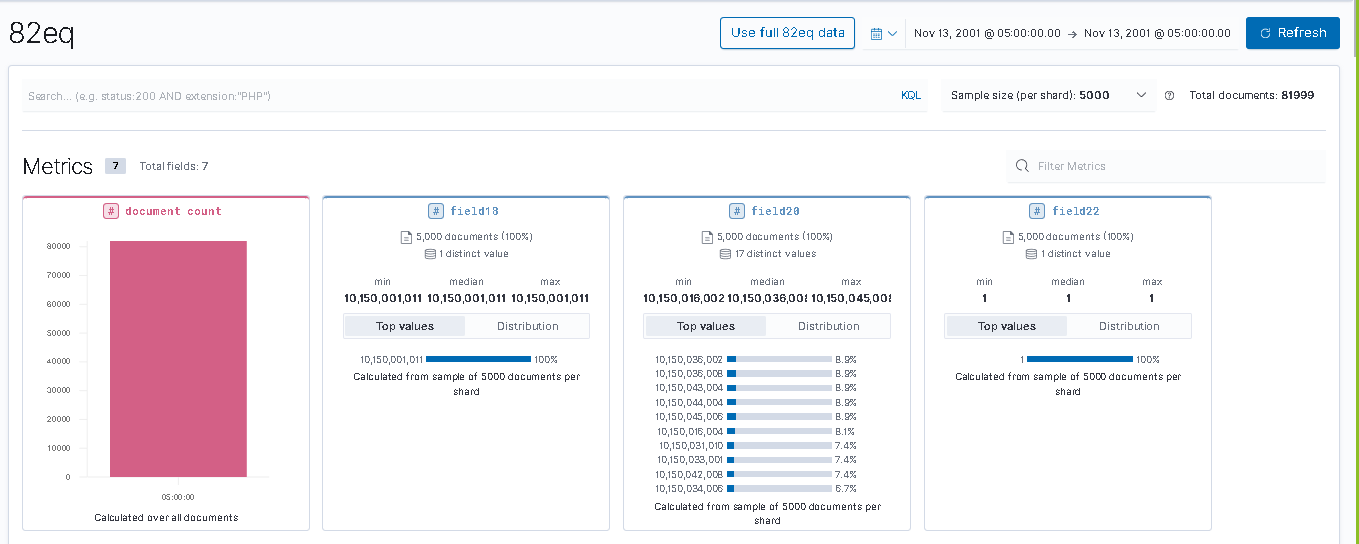


Figura 23: Vista 1 - Cond vivienda

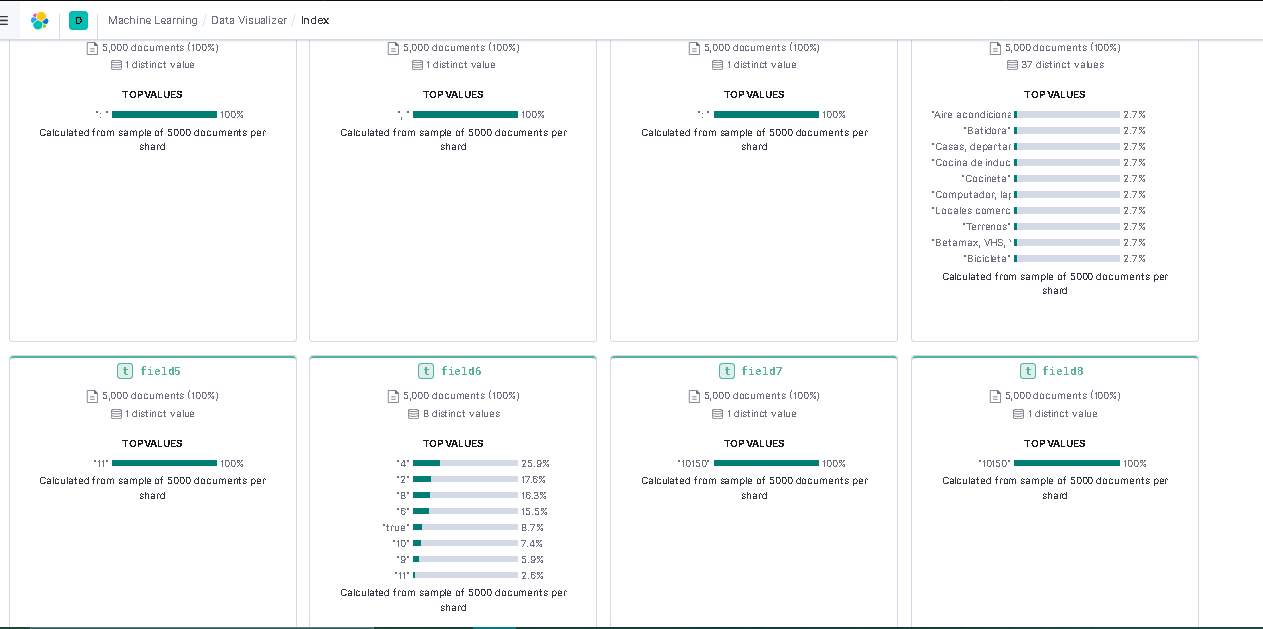


Figura24: Vista 2 - Cond vivienda

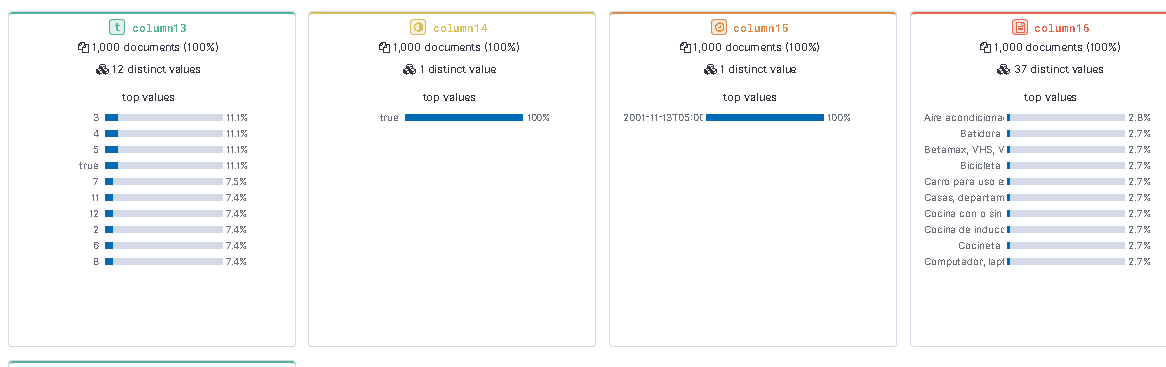


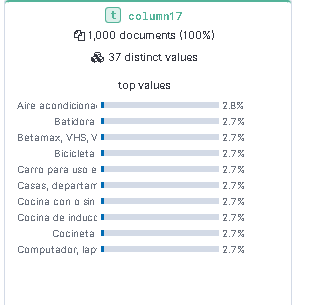
Figura25: Vista 3 - Cond vivienda 

Figura26: Vista 4 - Cond vivienda

Como se puede observar en las figuras tambien se puede notar por porcentajes los bienes que poseen las viviendas en el Ecuador.

**Caso de estudio**: Tendencia actual – Red Social TikTok

Se utilizó Kibana para la visualización de resultados

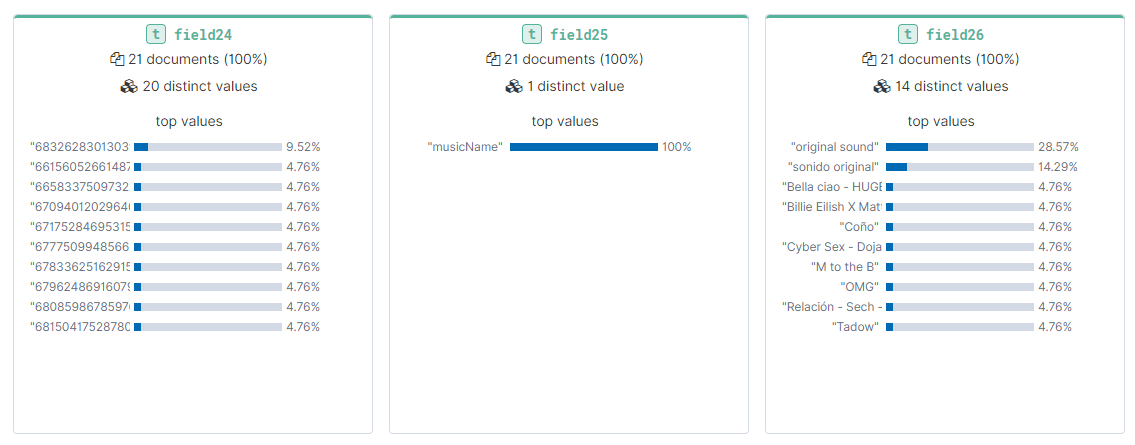


Figura 27: Vista de datos de TikTok

En la figura 27 observamos la visualización de los datos de una parte donde se observa la música original que utilizan de la aplicación.

1. **RESULTADOS OBTENIDOS UNIFICADOS**

**Caso de estudio:** Pulso político en Ecuador

Desde el paso de obtención de datos, se estuvieron enviando las consultas desde CouchDB hasta Elastc cloud. Razón por la que los datos ya se encontraron en el punto de unificación de los resultados obtenidos, desde un principio.

**Caso de estudio**: Juegos en linea

Se descargaron los datos de MySQL en un archivo con formato json.



Figura 25: Configuración para ECloud-condvivienda

Se subieron los datos del archivo generado con anterioridad a la nube con el siguiente archivo de configuración

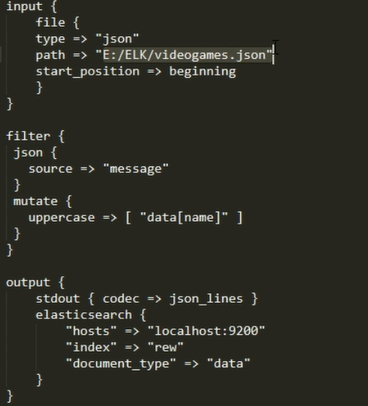


Figura 28: Json a ElasticSearch en la nube

**Caso de estudio:** Condiciones de vida-vivienda en el Ecuador

Se procede a configurar el archivo con el nombre y la ruta del archivo. json necesario para la inserción de datos en la nube.

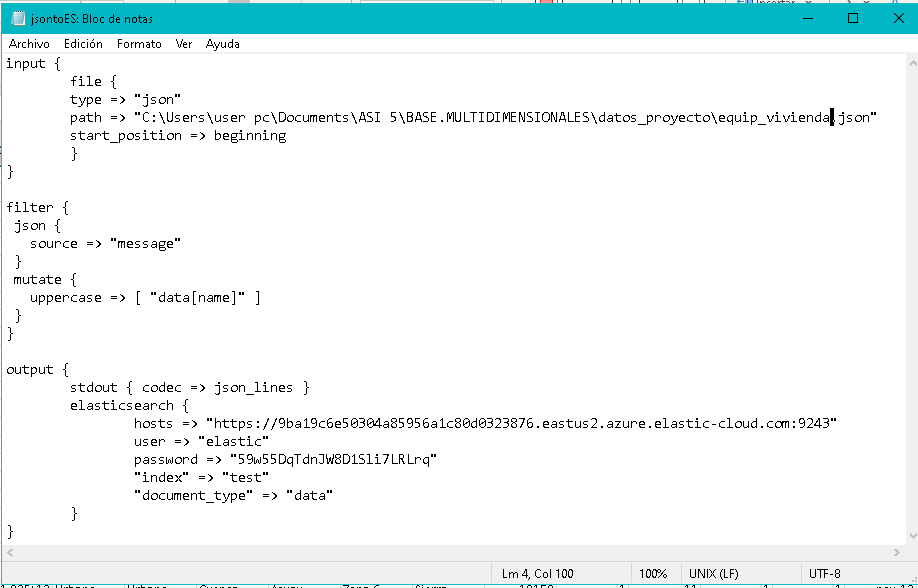


Figura 29:Configuración para ECloud-condvivienda

El comando para ejecutar el archivo es:



Figura 30: Ejecución de archivo de conf-condvivienda

Empieza a ingresar datos, se tiene que dejar que lea todo el archivo y no interrumpir para que pueda subir los datos sin percances.

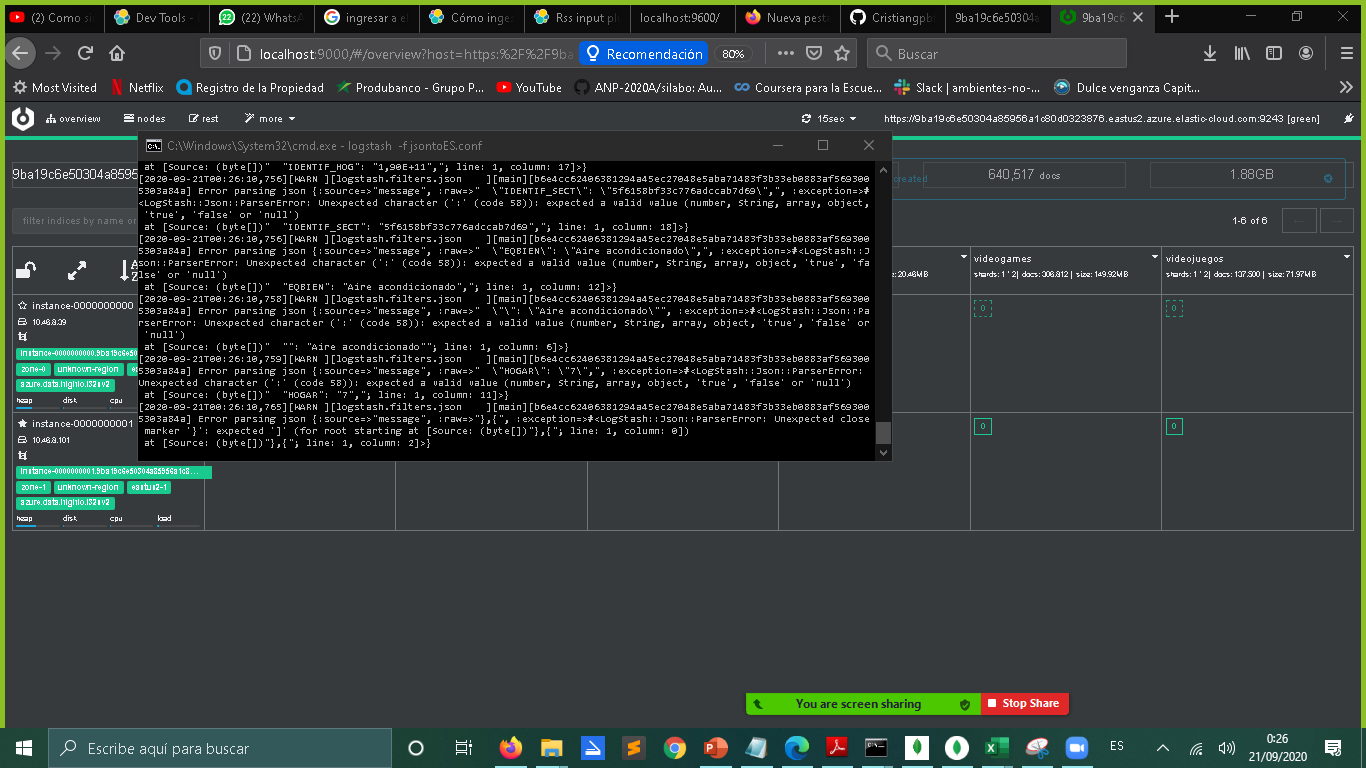


Figura 31:Ingreso datos cond-vida

Se puede llamar con el método get a los datos que se están ingresando y poder comprobar su correcto ingreso.

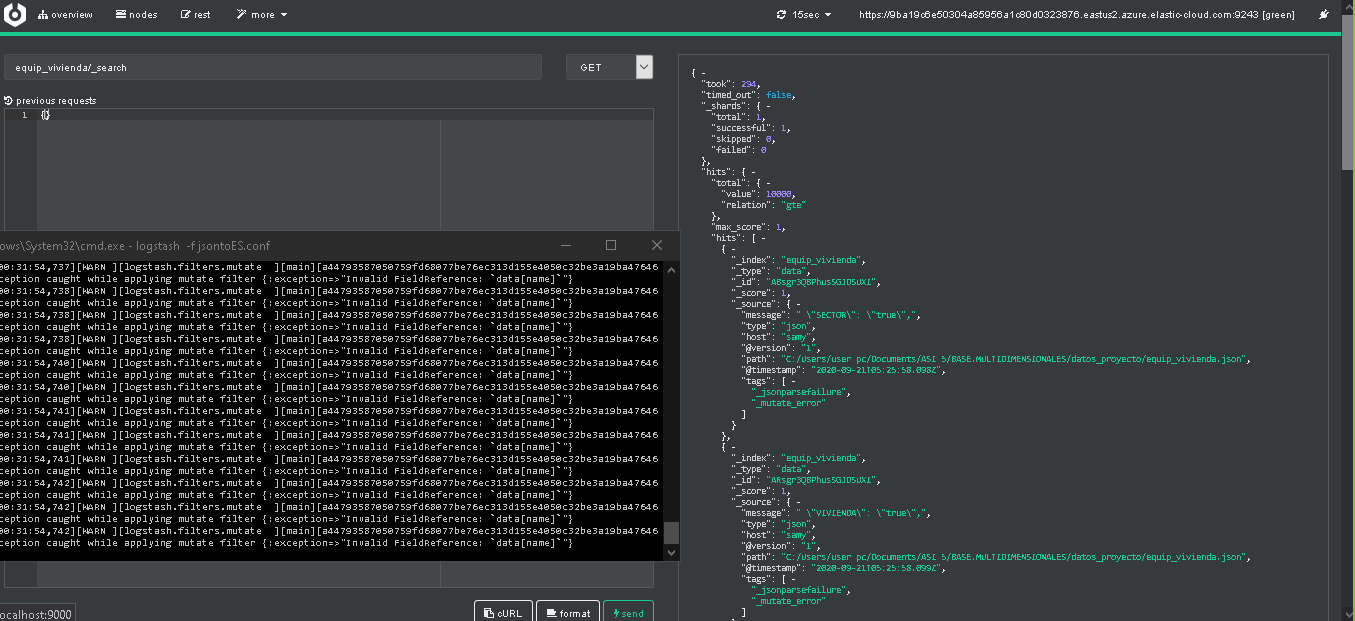


Figura 32:Comprobación cond-vida

**Caso de estudio**: Tendencia actual Red Social Tik-Tok, es una de las plataformas de redes sociales que más rápido ha crecido en el mundo y que presenta una versión alternativa de compartir e interactuar online, ya que permite a los usuarios crear vídeos cortos con música, filtros y otras características.

Se subieron los datos del archivo generado con anterioridad a la nube con el siguiente archivo de configuración



Figura 33: Json a ElasticSearch en la nube

Empieza a ingresar datos, se tiene que dejar que lea todo el archivo y no interrumpir para que pueda subir los datos sin percances.

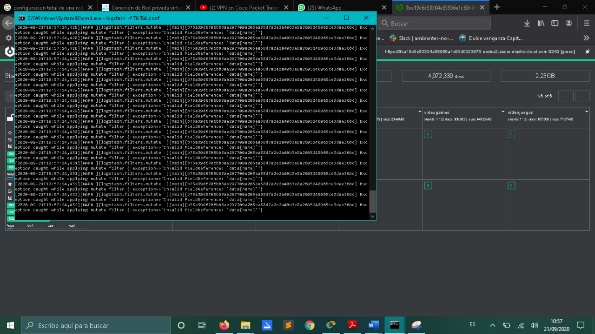


Figura 34: Comprobación TikTok

**Elastic Cloud:**

**Se observa la creación de 3 nodos, 16 índices para la carga de datos en conjunto de los distintos casos de estudio.**

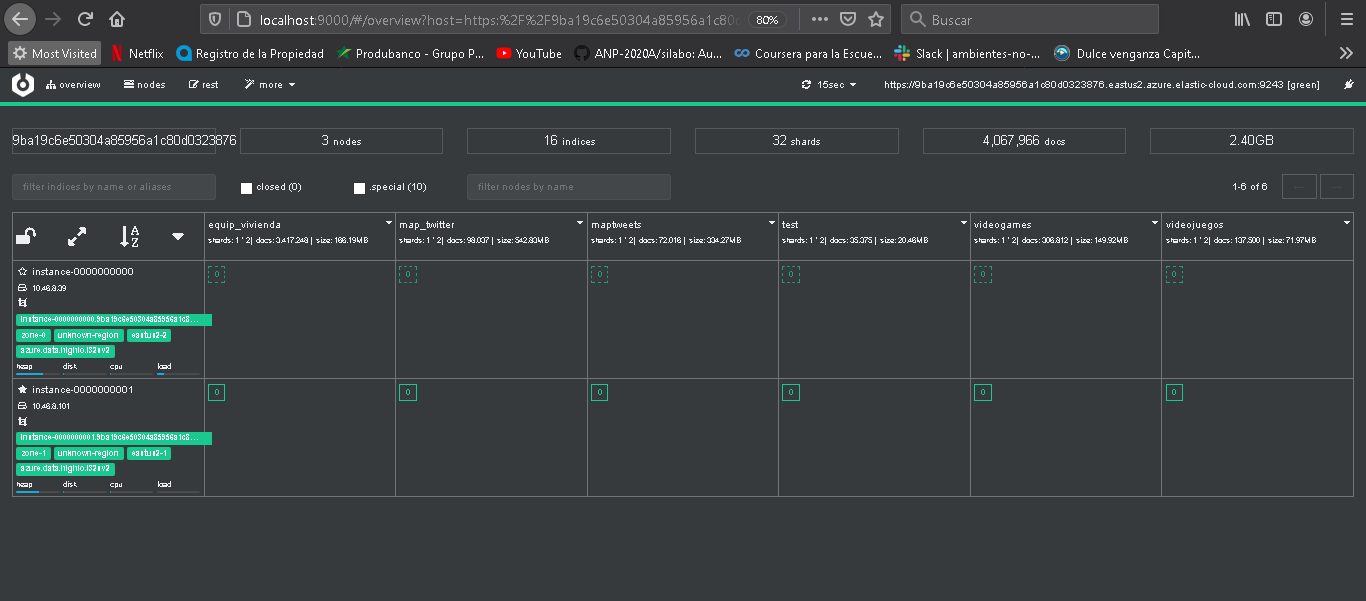


Figura 35: Elastic Cloud unificada

1. **CONCLUSIONES**

* El uso de varias fuentes permite relacionar la cosecha de datos sin importar los casos de estudio y a su vez darse cuenta de que es importante actualizar la información por lo que se concluye que es mejor trabajar con datos en tiempo real.
* El mapeo de los datos es primordial para el análisis de resultados, puesto que si no se cuentan con datos mapeados no se puede realizar un análisis correcto de los datos recopilados.

1. **DESAFÍOS Y PROBLEMAS ENCONTRADOS**

* La carga de información estática puede registrar un significativo aumento de tiempo empleado que los datos en tiempo real.
* Los datos no mapeados generan conflictos en el análisis de las gráficas.
* La carga de datos en elastic fue un problema por la cantidad de datos que poseemos en el archivo json.

.

1. **LINKS**

* <https://youtu.be/t84Wulo_by4>
* <https://github.com/Cristiangpbf/Proyecto_Final_BDDM.git>

1. **REFERENCIAS**

|  |  |
| --- | --- |
| [1] | Jorge, «Hostinger Tutoriales,» 09 Noviembre 2019. [En línea]. Available: https://www.hostinger.es/tutoriales/que-es-mysql/. [Último acceso: Septiembre 2020]. |
| [2] | César, «Platzi,» 2015. [En línea]. Available: https://platzi.com/blog/que-es-postgresql/. [Último acceso: Septiembre 2020]. |
| [3] | ARSYS, 27 09 2017. [En línea]. Available: https://www.arsys.es/blog/programacion/cloud-couchdb-bbdd/. [Último acceso: Septiembre 2020]. |
| [4] | Rubenfa, «GENBETA,» 3 Febrero 2014. [En línea]. Available: https://www.genbeta.com/desarrollo/mongodb-que-es-como-funciona-y-cuando-podemos-usarlo-o-no. [Último acceso: Septiembre 2020]. |
| [5] | David, «Ocho bits,» 14 Enero 2019. [En línea]. Available: https://www.ochobitshacenunbyte.com/2018/08/28/que-es-y-como-funciona-elasticsearch/. [Último acceso: Septiembre 2020]. |
| [6] | A. Chabir, «Adictos añ trabajo Tutoriales,» 27 Diciembre 2015. [En línea]. Available: https://www.adictosaltrabajo.com/2015/12/27/introduccion-a-kibana/. [Último acceso: Septiembre 2020]. |
| [7] | «Software Advice,» 2020. [En línea]. Available: https://www.softwareadvice.com/project-management/cerebro-profile/. [Último acceso: Septiembre 2020]. |
| [8] | davinci group, 12 Mayo 2020. [En línea]. Available: https://www.davincigroup.es/que-es-logstash-ejemplo-practico-de-uso/. [Último acceso: Septiembre 2020]. |
| [9] | WebEmpresa, 01 Marzo 2018. [En línea]. Available: https://www.webempresa.com/blog/que-es-twitter-como-funciona-2.html. [Último acceso: Septiembre 2020]. |
| [10] | INEC, «Instituto nacional de estadiística y censos,» 2020. [En línea]. Available: https://www.ecuadorencifras.gob.ec/la-institucion/. [Último acceso: Septiembre 2020]. |
| [11] | Y. FM, «xataka,» 26 Febrero 2020. [En línea]. Available: https://www.xataka.com/basics/que-tiktok-donde-viene-que-ofrece-red-social-videos. [Último acceso: Septiembre 2020]. |