PROYECTO FINAL

Amagua Jhoel, Curipoma David, Soto Wendy
Escuela Politécnica Nacional Tecnología en Análisis de Sistemas Informáticos

***Resumen*— Saludos, en este proyecto se diseñara la arquitectura de un Data Lake la cual va a tener insumo de diferentes bases de datos SQL, NoSQL y de fuentes de Internet El concentrador de datos a utilizar será Elasticsearch, además este informe contendrá, objetivos, cronograma de actividades, recursos y herramientas utilizadas, la arquitectura que utilizamos para la realización de este proyecto, las visualizaciones de la información, los resultados obtenidos y al final del informe se incluirá el link de github en donde se subirá toda la información respectiva del proyecto.**

I. INTRODUCCIÓN

A. Definición

En cada caso de estudio se tendrá diferentes visualizaciones, cada una con su explicación y la respectiva indexación.

1. Eventos deportivos en los principales estadios de Ecuador. - Este tema nos arrojara información acerca de los partidos que se ha tenido en cada estadio del país, sea del campeonato nacional, copa Ecuador o partidos de copas internacionales jugados en estos estadios.

2. Tema de interés: Stack Overflow (Internet) Es una página donde se busca soluciones con respecto a problemas de programación y demás temas.

3. Pulso político en 5 ciudades de Ecuador. - En esta base de datos se muestran los diferentes partidos políticos, cuando se crearon y los principales participes de estos diferentes partidos, así como donde se encuentran ubicados y demás información.

4. Top 10 twitteros en 5 ciudades de Ecuador. - En esta base se tendrá a las personas que más han tenido actividad en Twitter en diferentes ciudades del Ecuador, así como su número de seguidores, la descripción de la publicación y demás información.

5. Conciertos y eventos públicos. - En esta base se tendrá eventos que se han realizado en el Ecuador, siendo estos conciertos, eventos públicos o demás. En la información se tendrán datos como fecha, lugar, y demás.

6. Eventos o noticias mundiales. - A diferencia de las bases anteriores, en esta base se tendrá registros de eventos y noticias mundiales, se obtendrá datos como fechas, lugares, personas, protagonistas y demás información importante de cada evento o noticia.

B. Objetivos Generales

Nuestro principal objetivo es la aplicación de los conocimientos aprendidos durante todo el semestre, conocer acerca de las diferentes bases de datos que existen y lograr poder unir estas en una sola para trabajar con datos mas completos.

C. Objetivos Específicos

- Unir diferentes bases de datos.
- Utilizar ElasticSearch
- Conocer acerca del funcionamiento de Logstash para poder la concentración de todos los datos.
- Usar Kibana para tener una mejor visualización de los datos que conseguimos.

D. Descripción del Equipo

En grupo de trabajo está compuesto por 3 personas, cada una con diferentes habilidades y cada una se encargará de una parte del proyecto para lograr el objetivo.

- **Amagua Jhoel. -** Recopilación de las bases de datos de: Eventos deportivos en los principales estadios de Ecuador y Top 10 twitteros en 5 ciudades de Ecuador.
- **Curipoma David. -** Recopilación de las bases de datos de: Eventos o noticias mundiales y Pulso Político en 5 ciudades de Ecuador.
- **Soto Wendy. -** Recopilación de las bases de datos de: Eventos o noticias mundiales y tema de interés elegido por el estudiante (Stack Overflow).

E. Cronograma de Actividades

Se describirá las actividades que se desarrollaran en las fechas previstas para cumplir con el objetivo del proyecto.

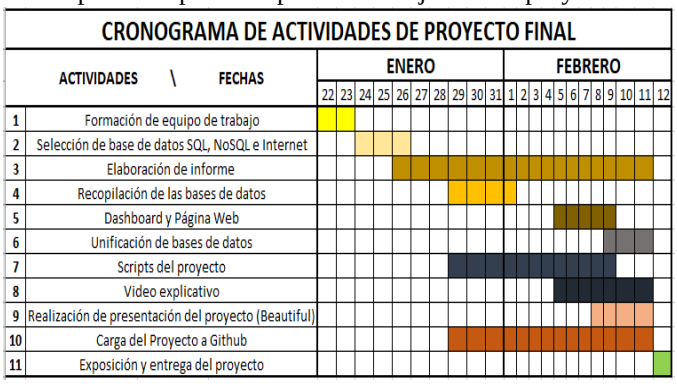


Ilustración 1.- Cronograma de actividades

F. Asignación de actividades a cada miembro del equipo

Cada miembro del equipo tendrá actividades a desarrollar, sin embargo, se necesita apoyo de una a otro para lograr el objetivo, así pues, varias actividades estarán realizadas en pares o entre todos los miembros del equipo:

- **Selección de base de datos SQL, NoSql e Internet:** Todos los miembros del equipo.
- **Elaboración de informe:** Todos los miembros del

equipo.

- **Recopilación de las bases de datos:** 2 bases de datos cada miembro del equipo.
- **Dashboard y Pagina Web:** Jhoel Amagua
- **Unificación de bases de datos:** Todos los miembros del equipo, cada miembro unificara sus bases de datos
- **Scripts del proyecto:** Todos los Miembros del equipo
- **Video explicativo:** David Curipoma
- **Realización de presentación del proyecto (Beautiful)):** Jhoel Amagua
- **Carga del proyecto a Github:** Wendy Soto
- **Exposición y entrega del proyecto:** Todos los miembros del equipo.

II. DESARROLLO

A. *Recursos y herramientas utilizadas*

La herramienta a utilizar como concentrador de datos será ElasticSearch.
Las bases de datos SQL serán:

- Postgresql
- SQL Server

Bases de datos NoSQL serán:

- CouchDB
- Redis

Internet:

- Twitter
- Web Scraping

La herramienta para visualizar los datos será Kibana.
Adicionalmente los recursos usados son las computadoras de cada miembro del equipo y algunos editores de texto, asi como Python para correr el respectivo script para la obtención de datos.

B. *Arquitectura de la solución*

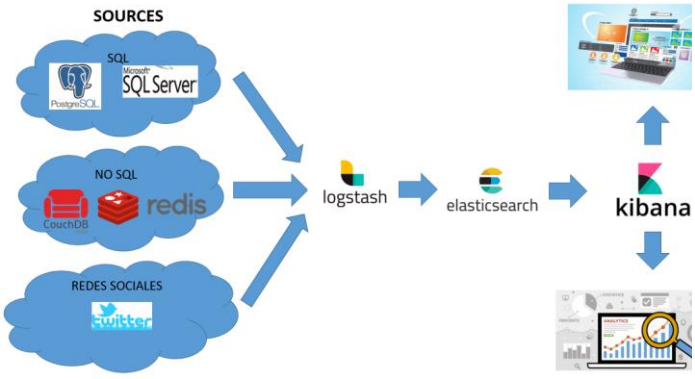


Ilustración 2.- Arquitectura de la solución

C. *Extracción de datos*

Los datos utilizados fueron extraídos de diferentes fuentes de internet:

- **Eventos deportivos en los principales estadios de Ecuador. -** Se extrajo los datos desde twitter mediante filtro de palabras.
- **Tema de interés:** Se escogió como tema de interés la página web Stack Overflow, pagina donde se busca soluciones con respecto a problemas de programación y demás temas.
- **Pulso político en 5 ciudades de Ecuador:** estos datos fueron extraídos de la base de datos del CNE (Consejo Nacional Electoral), link: http://cne.gob.ec/es/estadisticas/bases-de-datos/category/1585-organizaciones-politicas
- **Top 10 Twiteros en 5 ciudades de Ecuador:** Se extrajo desde Twitter mediante el archivo .py que se proporcionó en el semestre utilizando geolocalización para las ciudades.
- **Conciertos y eventos públicos:** Se extrajo los datos desde twitter mediante filtro de palabras.
- **Eventos o noticias mundiales:** Se extrajo los dartos de una subsidiaria de Google LLC , es una comunidad en línea de científicos de datos y profesionales del aprendizaje automático . Kaggle permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos basado en la web. Link: https://www.kaggle.com/datasets

D. *Análisis de la Información*

La información recopilada deberá coincidir con los temas que se buscó.

1. Eventos deportivos en los principales estadios de Ecuador.

```
            doc = db.save(dict_tweet)  # Aqui se guarda el tweet en la base de couchdb
            print ("Guardado " + "=> " + dict_tweet["_id"])
        except:
            print ("Documento ya existe")
            pass
        return True

    def on_error(self, status):
        print (status)


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

# Setear la URL del servidor de couchDB
server = couchdb.Server('http://localhost:5984/')
try:
    # Si no existe la Base de datos la crea
    db = server.create('eventos_deportivos')
except:
    # Caso contrario solo conectarse a la base existente
    db = server['eventos_deportivos']

# Aqui se define el bounding box con los limites geograficos donde recolectar los tweets.
twitterStream.filter(track=["fútbol","estadios","deportes","ecuador","eventos"])
#twitterStream.filter(locations=[-78.819158,-1.463512,-78.180577,-1.147738])  # GEOLOCALIZACION DE CUENCA
```

Ilustración 3.- Información de eventos deportivos

2. Tema de interés

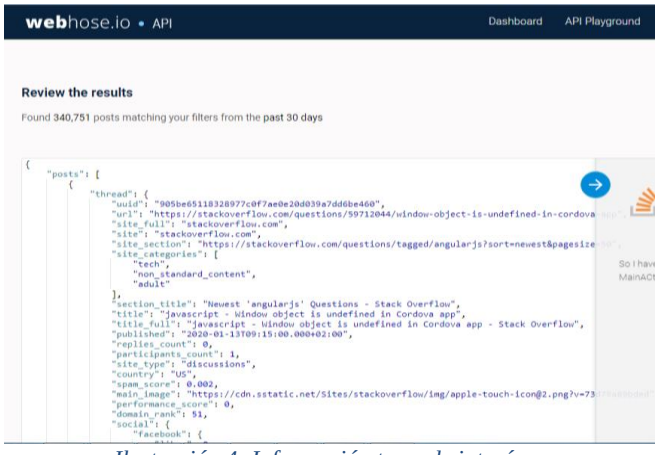


Ilustración 4- Información tema de interés.

3. Pulso político en 5 ciudades de Ecuador

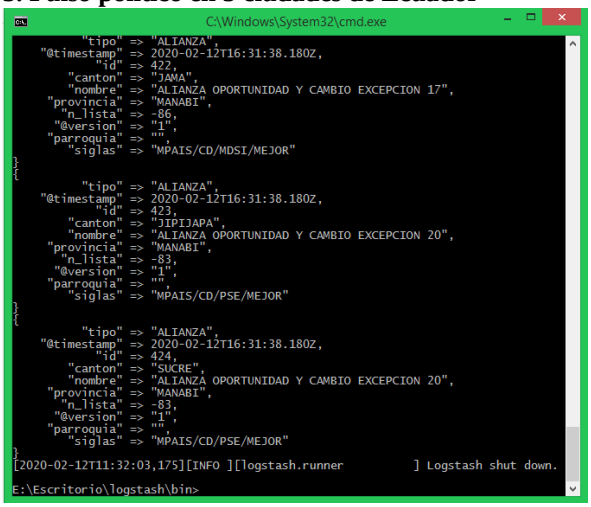


Ilustración 5.- Información de pulso político

4. Top 10 twitteros en 5 ciudades de Ecuador

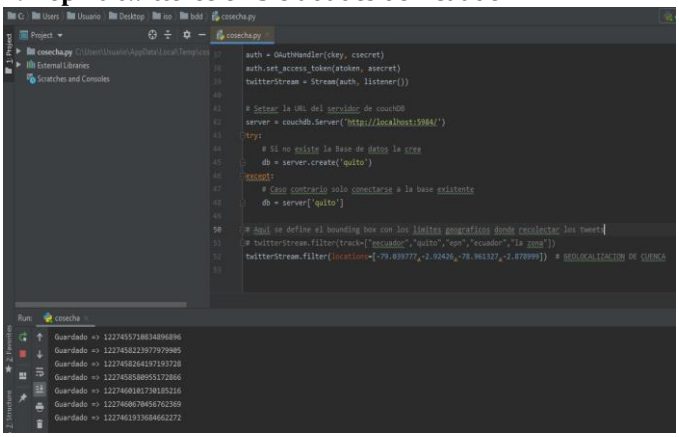


Ilustración 6.- Informacion de twitteros

5. Conciertos y eventos públicos

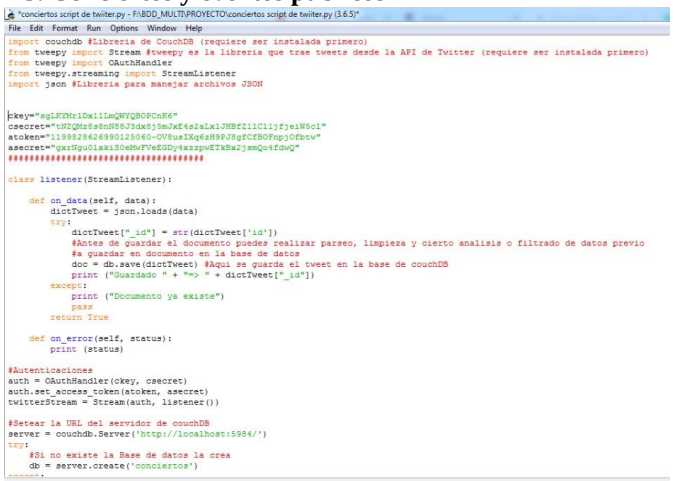


Ilustración 7.- Código para eventos o noticias mundiales

6. Eventos o noticias mundiales

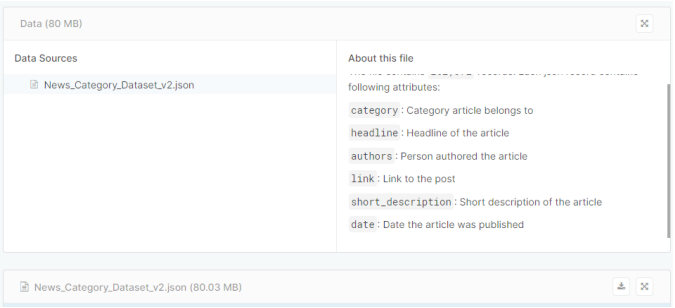
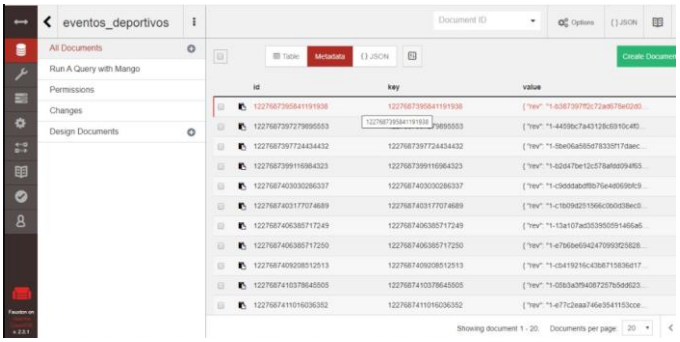


Ilustración 8.- Información de eventos o noticias mundiales

E. Visualización de la información

1. Eventos deportivos en los principales estadios de Ecuador



2. Tema de interés: Stack Overflow (red social)

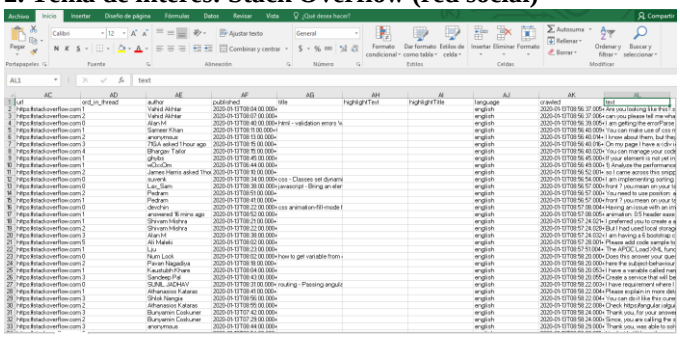


Ilustración 9.- Visualización de Stack Overflow

3. Pulso político en 5 ciudades de Ecuador

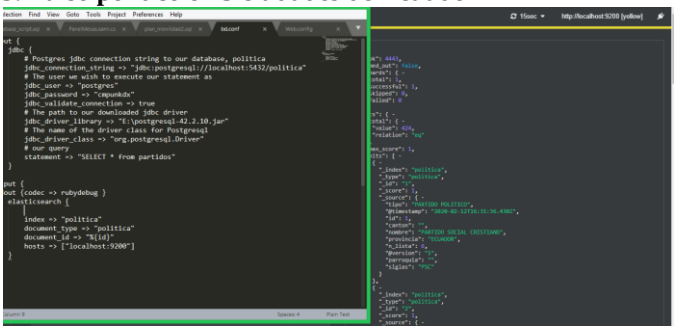


Ilustración 10.- Visualización de pulso político

4. Top 10 twitteros en 5 ciudades de Ecuador

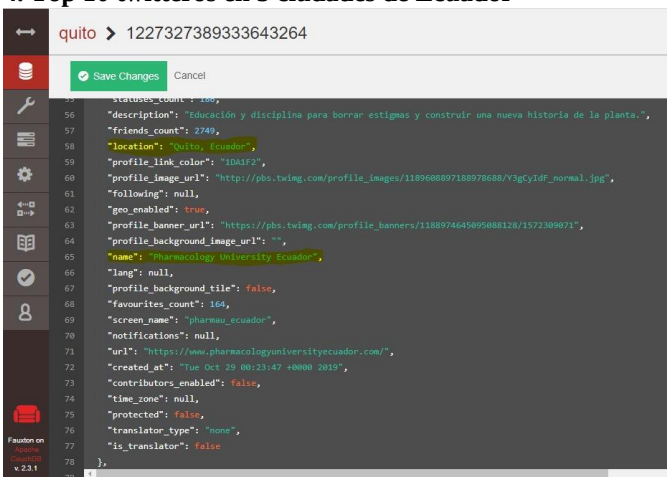


Ilustración 11.- Visualización twitteros 1

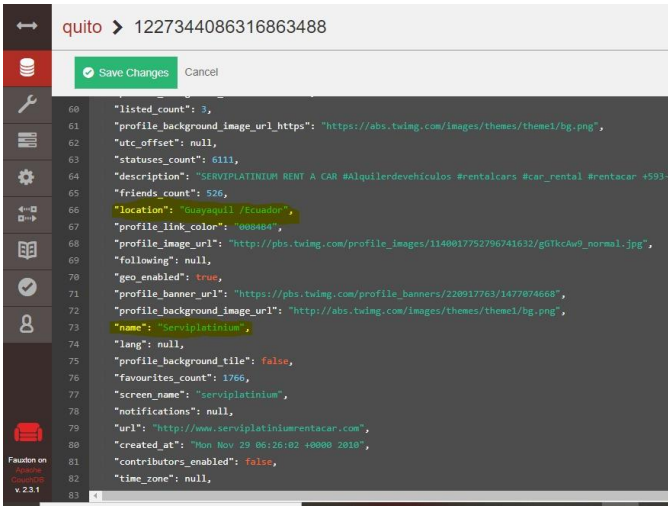


Ilustración 12.- Visualización twitteros 2

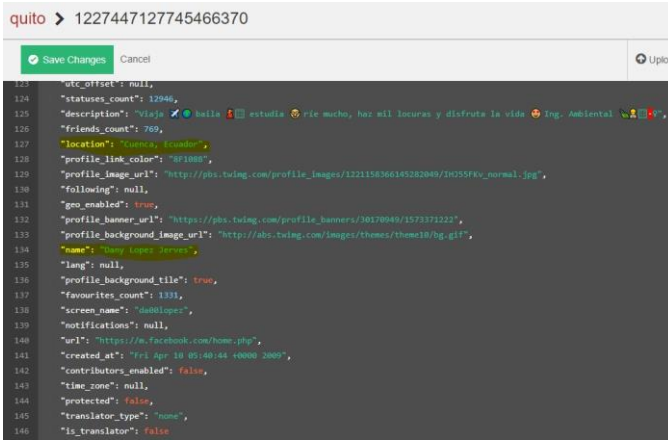


Ilustración 13.- Visualización twitteros 3

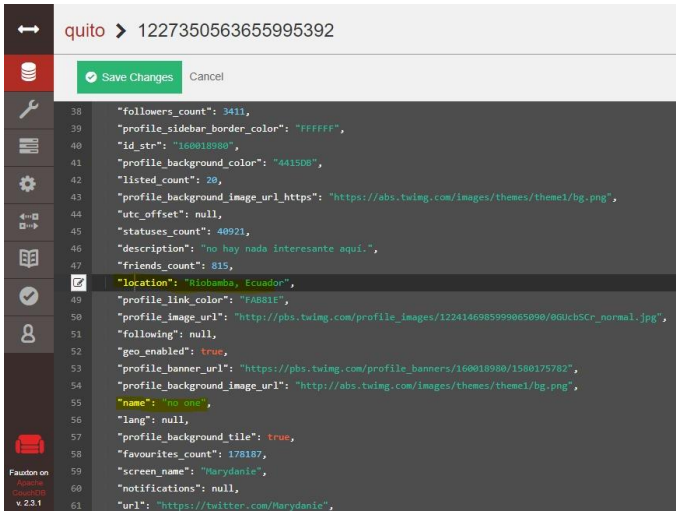


Ilustración 14.- Visualización twitteros 4

5. Conciertos y eventos públicos

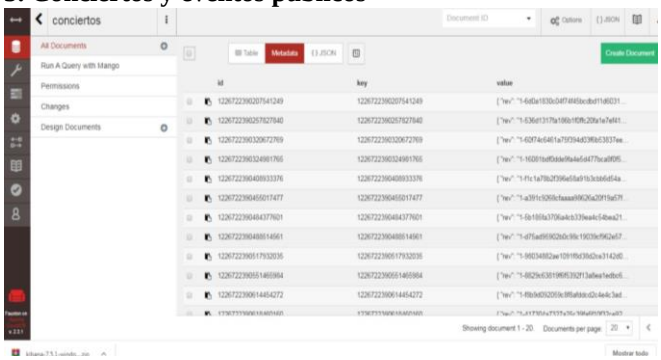


Ilustración 15.- Visualización de conciertos y eventos públicos

6. Eventos o noticias mundiales

Ilustración 16.- Visualización de eventos o noticias mundiales

F. Resultados Obtenidos

1. **Eventos deportivos en los principales estadios de Ecuador**

2. **Tema de interés: Stack Overflow (red social)**

- Número de preguntas hechas por los autores

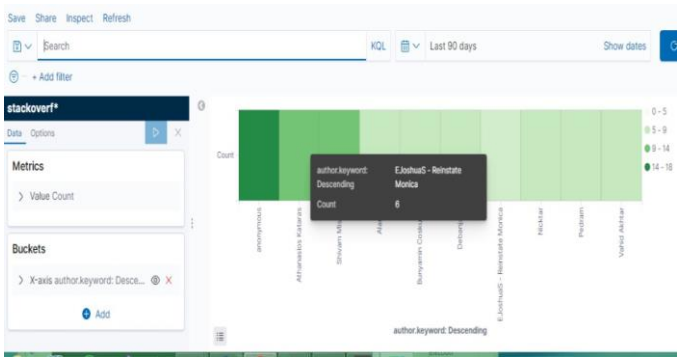


Ilustración 17.- Número de preguntas hechas por los autores

- Respuestas más utilizadas en stackoverflow

Ilustración 18.- Respuestas más utilizadas en stackoverflow

- Tema que más se busca solución

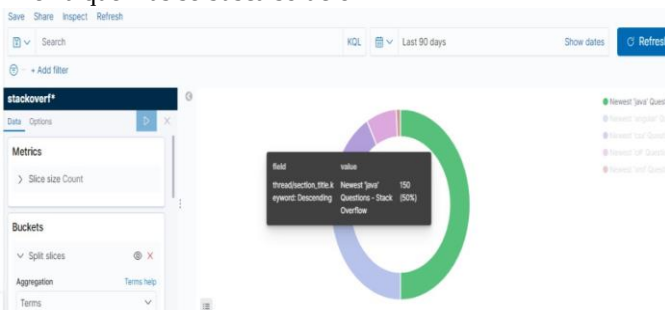


Ilustración 19.- Tema que más se busca: solución

Dashboard de Tema de interés

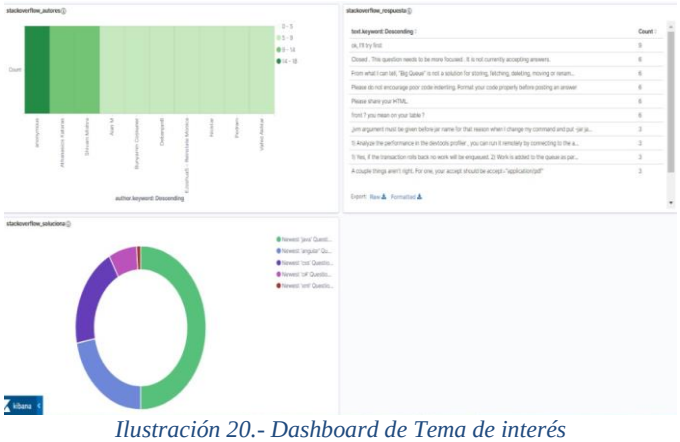


Ilustración 20.- Dashboard de Tema de interés

3. Pulso político en 5 ciudades de Ecuador

- Número de partidos políticos por Provincia

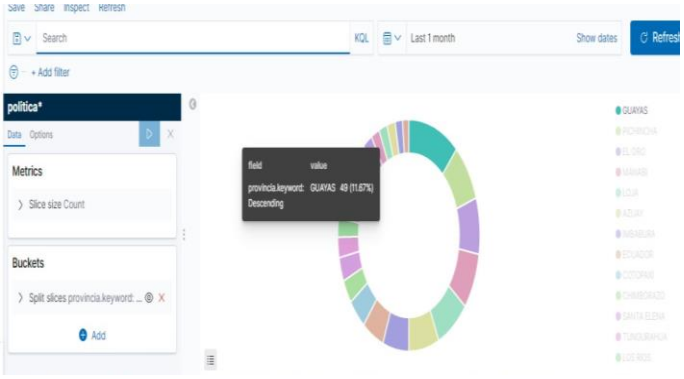


Ilustración 21.- Numero de partidos políticos por provincia

- Partidos políticos

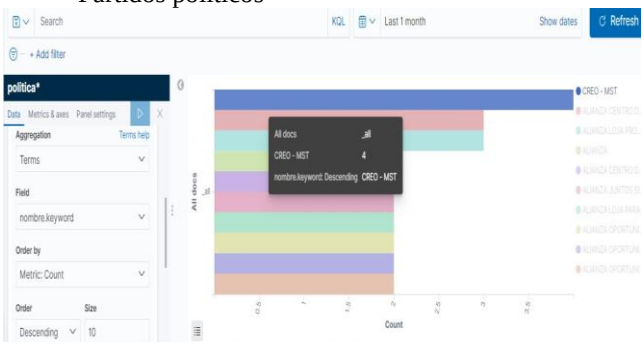


Ilustración 22.- Partidos políticos

- Tipo de partido

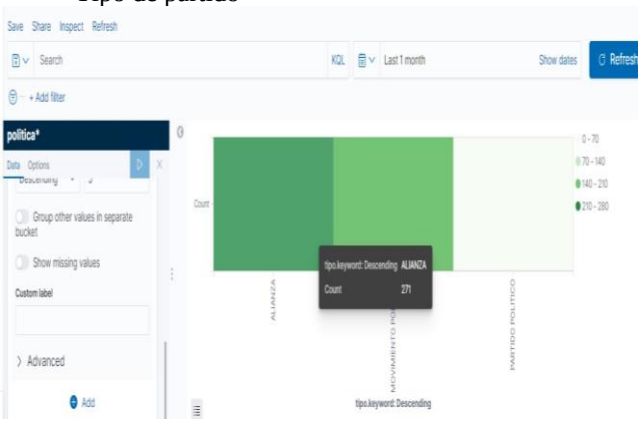


Ilustración 23.- Tipo de partido

Dashboard de pulso Político

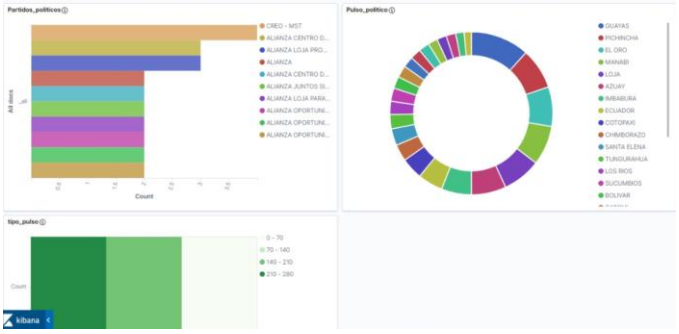


Ilustración 24.- Dashboard de pulso Político

4. Top 10 twitteros en 5 ciudades de Ecuador

5. Conciertos y eventos públicos

- Evento más twitteado

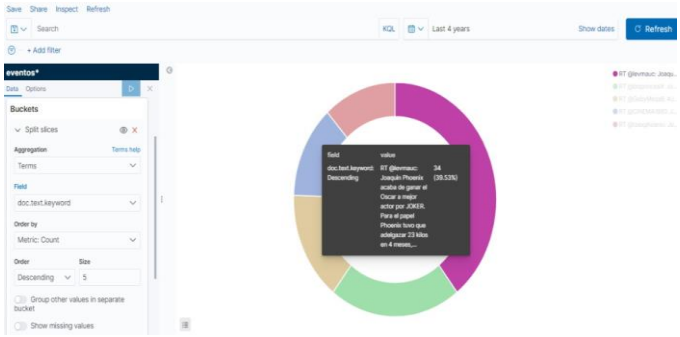


Ilustración 25.- Evento más twitteado

- 80% de los twitteros tenían activado la opción geolocalización

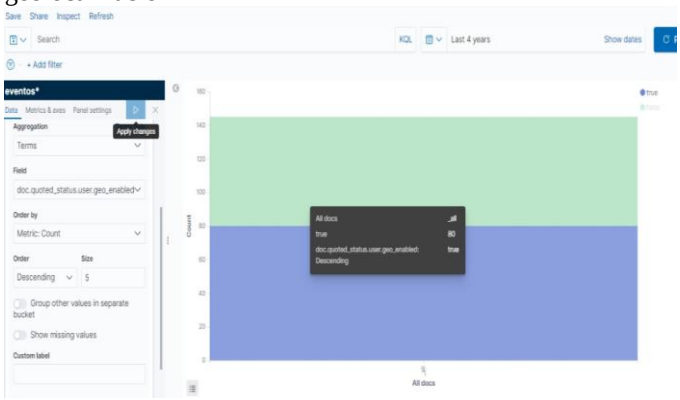


Ilustración 26.- 80% de los twitteros tenían activado la opción geolocalización

- 10 eventos más twitteados el fin de semana

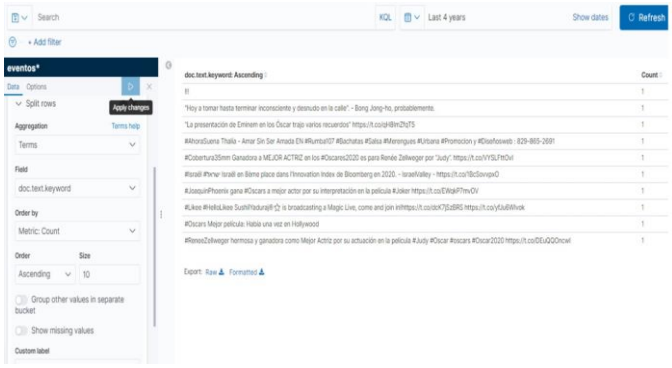


Ilustración 27.- - 10 eventos más twitteados el fin de semana

Dashboard de conciertos y eventos públicos

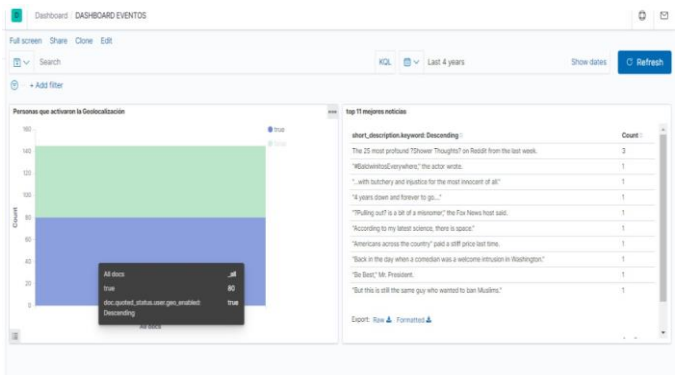


Ilustración 28.- Dashboard de conciertos y eventos públicos

6. Eventos o noticias mundiales

- Política como mayor noticia mundial

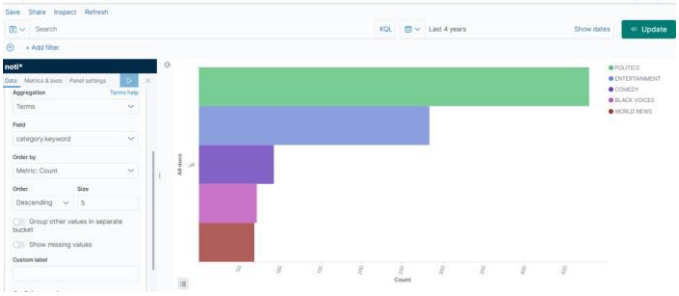


Ilustración 29.- Mayor noticia mundial

- Autores con más noticias publicadas

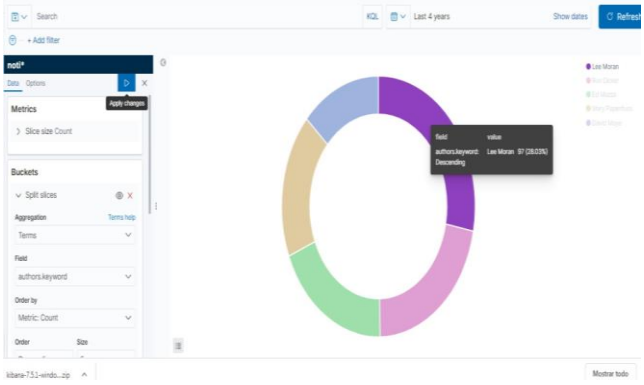


Ilustración 30.- - Autores con más noticias publicadas

- 11 Noticias más receptadas

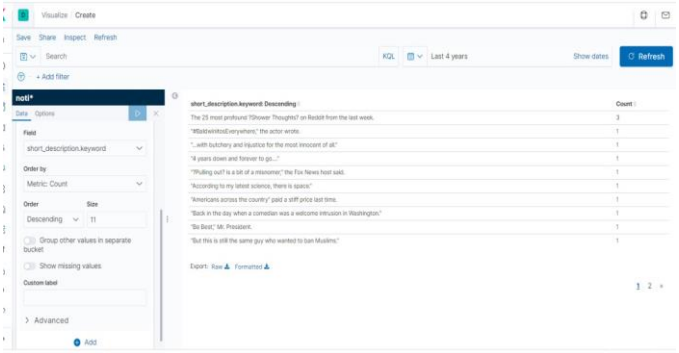
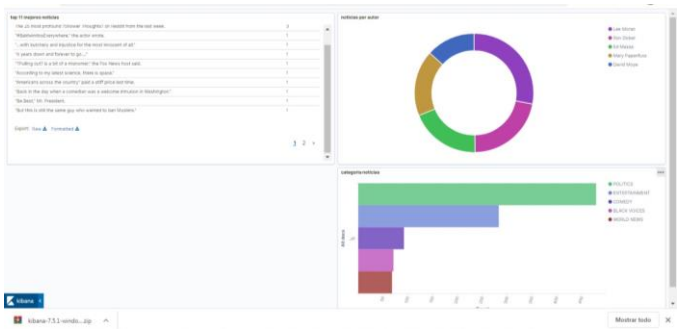


Ilustración 31.- 11 Noticias más receptadas

Dashboard de eventos o noticias mundiales



III. CONCLUSIONES Y RECOMENDACIONES

- Las bases de datos puedes ser uno de los mayores problemas, ya que muchas veces no se tiene suficiente información para la extracción, por lo que la recopilación se la debe hacer lo antes posible.
- El trabajo se desarrolló de manera correcta, aun cuando se tuvo varios inconvenientes con las bases de datos y la migración desde las diferentes fuentes a elasticsearch.

IV. DESAFÍOS Y PROBLEMAS ENCONTRADOS

- Muy poca información sobre los temas a recolectar datos.
- Información acerca de Logstash para la migración de información.
- Visualizaciones de la información recolectada no se mostraba en Kibana.
- Adaptar el código de Logstash para poder migrar desde las diferentes fuentes hacia elasticsearch