

PROYECTO FINAL BASE DE DATOS MULTIMEDIA

ALTAMIRANO TACO JOSÉ LUIS
jose.altamirano@epn.edu.ec

TIPÁN VILLEGAS JENNY PATRICIA
jenny.tipan@epn.edu.ec

ESCUELA POLITÉCNICA NACIONAL
ESCUELA DE FORMACIÓN DE TECNÓLOGOS

I. INTRODUCCIÓN

El acceso cada vez más fácil a la información a través de fuentes electrónicas de almacenamiento, ya sean bases de datos, CD-ROM, o Internet, ha originado la constitución de bases de datos textuales de gran tamaño, formadas por artículos, patentes, informes, notas técnicas entre otros, información que resulta bastante importante para las organizaciones, empresas y la misma sociedad, ya que a través de un análisis de esta se puede tomar decisiones importantes ya sea para innovar algo, hacer cosas nuevas y estar informado de lo que ocurre a nuestro alrededor.

II. OBJETIVO GENERAL

- Analizar los datos recopilados sobre pulso político, top 10 twitters en cinco ciudades de Ecuador y un tema de interés personal para exponer los resultados del análisis.

III. OBJETIVOS ESPECÍFICOS

- Definir la arquitectura del proyecto.
- Recopilar datos sobre pulso político en cinco ciudades de Ecuador.
- Recopilar datos sobre top 10 twitters en cinco ciudades de Ecuador.
- Recopilar datos sobre un tema de interés personal.
- Visualizar los datos.

IV. RECURSOS Y HERRAMIENTAS PARA UTILIZAR

Para llevar a cabo el proyecto vamos a utilizar:

HARDWARE

Computador 1:

- Procesador: Intel® Core™ i3-2310M CPU @ 2.10 GHz
- Memoria instalada (RAM): 4,00 GB
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64

Ver información básica acerca del equipo

Edición de Windows

Windows 10 Education

© 2018 Microsoft Corporation. Todos los derechos reservados.

Sistema

Procesador: Intel(R) Core(TM) i3-2310M CPU @ 2.10GHz 2.10 GHz

Memoria instalada (RAM): 4,00 GB

Tipo de sistema: Sistema operativo de 64 bits, procesador x64

Lápiz y entrada táctil: La entrada táctil o manuscrita no está disponible para esta pantalla

Ilustración 1 Información básica del computador 1

Computador 2:

- Procesador: AMD E2-1800 APU with Radeon™ HD Graphics 1.70 GHZ
- Memoria instalada (RAM): 6,00 GB
- Tipo de sistema: Sistema operativo de 64 bits, procesador x64

Ver información básica acerca del equipo

Edición de Windows

Windows 10 Education

© 2018 Microsoft Corporation. Todos los derechos reservados.



Sistema

Procesador: AMD E2-1800 APU with Radeon(tm) HD Graphics 1.70 GHz

Memoria instalada (RAM): 6,00 GB (3,60 GB utilizable)

Tipo de sistema: Sistema operativo de 64 bits, procesador x64

Lápiz y entrada táctil: La entrada táctil o manuscrita no está disponible para esta pantalla

Ilustración 2 Información básica de Computador 2

SOFTWARE

- Python
- Base de datos CouchDB
- Java jdk

- Elasticsearch
- Logstash
- Cerebro
- Kibana

V. CASOS DE ESTUDIO

PULSO POLÍTICO

El próximo 24 de marzo se elegirán nuevos representantes para cumplir con las responsabilidades de la alcaldía, por lo cual vamos a recoger datos en cinco ciudades de Ecuador: Quito, Cuenca, Guayas, Ambato y Tulcán con el fin de analizar estos y conocer que candidatos son los más mencionados dentro de estas ciudades.

El propósito de analizar estos datos es llegar a conocer a breves rasgos quienes podrían llegar a ocupar el cargo de alcalde en estas ciudades.

ARQUITECTURA PARA LA SOLUCIÓN PULSO POLÍTICO

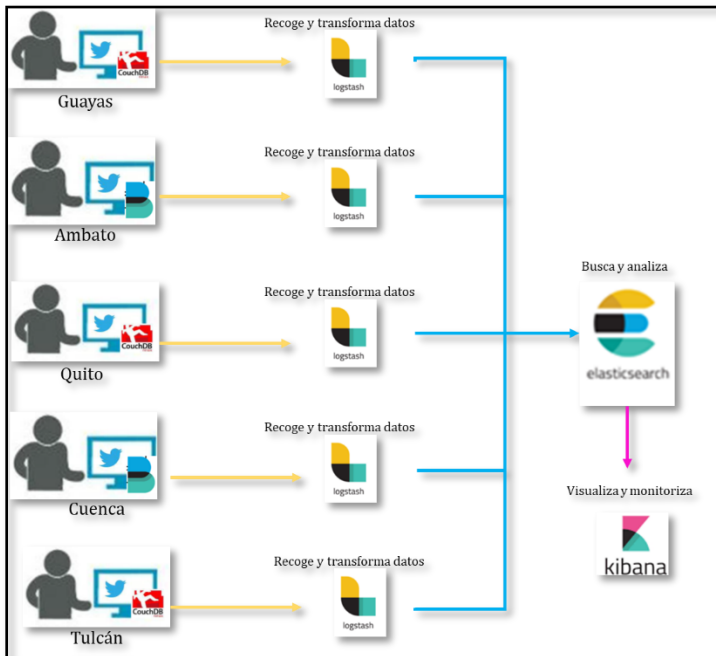


Ilustración 3 Arquitectura de la solución Pulso Político

EXTRACCIÓN DE DATOS

1. Vamos a usar un script en Python, para ello primero debemos instalar las librerías necesarias para su buen funcionamiento.
 - Buscamos donde se encuentra instalado Python e ingresamos a la carpeta Script y con esta ubicación ingresamos al **cmd** símbolo del sistema.
 - Aquí instalamos las librerías de CouchDB, Tweepy, json, de la siguiente manera:

```
C:\Python36\Scripts>pip install json
C:\Python36\Scripts>pip install tweepy
C:\Python36\Scripts>pip install couchdb
```

```
C:\Python27\Scripts>pip install couchdb
```

Ilustración 4 Instalación de librerías

2. Elaboramos el script en Python para que nos permita recoger datos de Twitter hacia nuestra base de datos CouchDB, cada script será nombrado de acuerdo con la ciudad de la cual se está minando datos. Así **política_Guayas.py**.

- Para el script primero importamos las librerías couchdb, tweepy y json.

```
import couchdb # Libreria de CouchDB (requiere ser
from tweepy import Stream # tweepy es la libreria d
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json # Libreria para manejar archivos JSON
```

Ilustración 5 Librerías de Python

- Usamos las credenciales de la cuenta de Twitter.

```
ckey = "b0hhIO0RfXbsgPjjBS1cAOEkB"
csecret = "3825kTPBIf0CkktKEZ1Q5aMjJe9HqMg8RD9P3sX0Tz1gf0p0dd"
atoken = "1268502774-9Glna0czUOXChiac8hs3S3lc976JHEXouRJGm5c"
asecret = "PAhtCrPxacsJ2AmTyIOCSKmGfPQzuF9j1107n2bT7ptgr"
```

Ilustración 6 Credenciales de Twitter

- Tenemos que especificar la dirección URL de nuestro servidor de CouchDB, poner las credenciales de usuario en caso de que las tengamos.

```
# Setear la URL del servidor de couchDB
server = couchdb.Server('http://admin:1234@localhost:5984/')
```

Ilustración 7 Dirección URL del servidor CouchDB

- Debemos especificar la base de datos en la que queremos recopilar los datos. Este script busca si la base de datos existe en el servidor y si no es así la crea. Como vamos a recopilar datos de cinco ciudades en cada script va a variar el nombre de la base de datos, que va a ser nombrada de acuerdo con cada ciudad de la que se va a recopilar datos.

```
# Si no existe la Base de datos la crea
db = server.create('guayas')
except:
# Caso contrario solo conectarse a la base existente
db = server['guayas']
```

Ilustración 8 Creación de la base de datos

- Vamos a utilizar palabras claves para la búsqueda de datos, en este caso vamos a usar las palabras **alcaldía**, **candidatos**, **guayas**, ya que queremos información acerca de los candidatos para alcalde de cada ciudad seleccionada. En cada script va a variar el nombre de la ciudad, ya que ahí deberá colocarse el nombre de la ciudad respectiva de la que se desea recoger los datos.

```
# Aquí se define el bounding box con los límites geograficos de
twitterStream.filter(track=["alcaldia","candidatos","guayas"])
```

Ilustración 9 Filtración de datos por palabras claves

3. Ejecutamos el script y si todo esta correcto los datos empezaran a guardarse en la base de datos de CouchDB.

```
File Edit Shell Debug Options Window Help
===== RESTART: C:/Users/jenti.
Guardado => 1092260295391748096
Guardado => 1092260299078623233
```

Ilustración 10 Almacenamiento de datos

4. Podemos verificar en CouchDB que los datos se están almacenando correctamente.

guayas	18.7 MB	6646
--------	---------	------

Ilustración 11 Base de datos en CouchDB

5. Una vez que ya hemos recogido los datos de las ciudades seleccionadas de Ecuador vamos a pasar los datos a Elasticsearch.

- Primero vamos a levantar cada uno de los servicios:
 - Elasticsearch
 - Cerebro

```
C:\Windows\System32\cmd.exe - elasticsearch
Microsoft Windows [Versión 10.0.17134.523]
(c) 2018 Microsoft Corporation. Todos los derechos reservados.

C:\elk\elasticsearch-6.5.4\bin>elasticsearch
[2019-01-14T21:00:30,284][INFO ][o.e.e.NodeEnvironment ] [_GDGCNY]
using [1] data paths, mounts [[(C:)], net usable_space [369.1gb], n
et total_space [464.4gb], types [NTFS]
[2019-01-14T21:00:30,308][INFO ][o.e.e.NodeEnvironment ] [_GDGCNY]
heap size [990.7mb], compressed ordinary object pointers [true]
[2019-01-14T21:00:32,354][INFO ][o.e.n.Node ] [_GDGCNY]
node name derived from node ID [_GDGCNYVS_6lMhePj6ehfQ]; set [node.n
```

Ilustración 12 Servicio Elasticsearch

- Una vez que ha iniciado Elasticsearch y cerebro vamos a ingresar al navegador y escribimos **localhost:9000** y podemos inicializar en cerebro.

6. Luego creamos los índices de mapeo.

- Escribimos el tipo de documento.

```
{
  "mappings": {
    "doc": {
```

Ilustración 13 Tipo de documento del índice de mapeo

- Estructuramos el mapeo, tomamos en cuenta los datos tipo **date**.

```
"created_at":{
  "type":"date",|
  "format":"EE MMM d HH:mm:ss Z yyyy||dd/MM/yyyy||dd-MM-yyyy||date_optional_time"
},
```

Ilustración 14 Estructura de mapeo datos tipo date

- Estructuramos el mapeo, tomamos en cuenta los datos de geolocalización o tipo **geo_point**.

```
"coordinates":{
  "properties":{
    "coordinates":{
      "type":"geo_point"
    }
  }
},
```

Ilustración 15 Estructura de mapeo datos tipo geo_point

7. Creamos los índices en los cuales se van a ir guardando los datos de cada base de datos de CouchDB, se usará el mapeo ya indicado para cada índice. Cada índice tendrá como nombre **ciudad_politica**, ciudad se reemplazará por el nombre de cada ciudad de la cual se recopiló los datos.

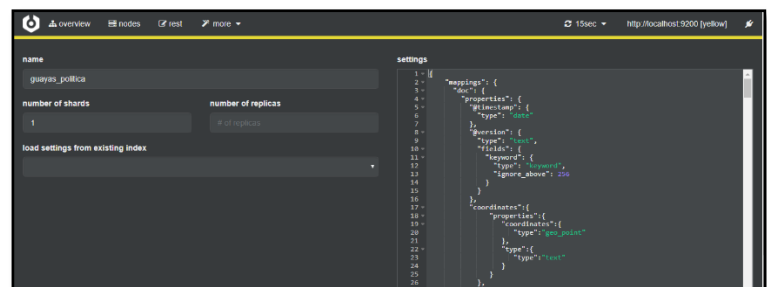


Ilustración 16 Creación de índices

ambato_politica shards: 5 * 2 docs: 0 size: 1.12KB [0] [1] [2] [3] [4]	cuenca_politica shards: 5 * 2 docs: 0 size: 1.12KB [0] [1] [2] [3] [4]
---	---

Ilustración 17 Índices caso de estudio Pulso Político

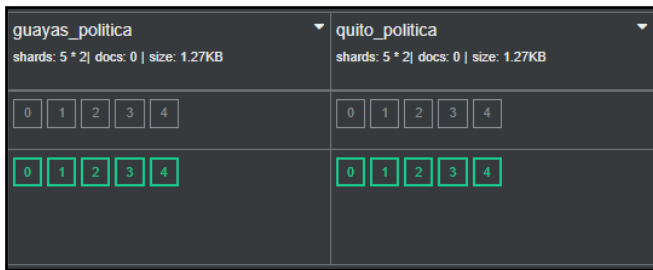


Ilustración 18 Índices caso de estudio Pulso Político

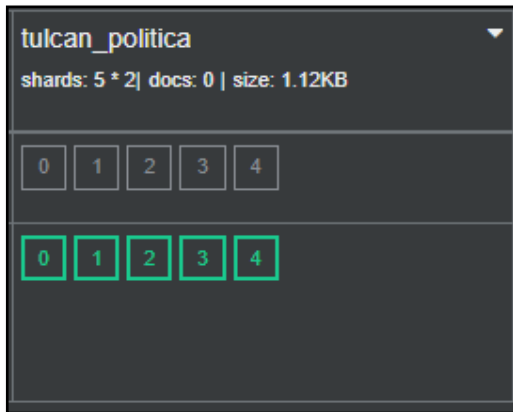


Ilustración 19 Índices caso de estudio Pulso Político

8. Ahora usamos Logstash.

- Primero instalamos el plugin **couchdb_changes**. Así `>logstash-plugin install logstash-input-couchdb_changes`

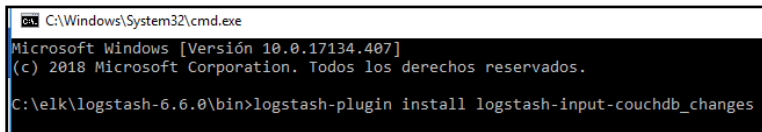


Ilustración 20 Instalar plugin couchdb_changes

- Luego vamos a usar el archivo de configuración **couchciudad.conf**, ciudad se ira reemplazando por cada ciudad seleccionada así **couchguayas.conf**. En este archivo vamos a nombrar la base de datos que vamos a utilizar de CouchDB y el nombre del índice que vamos a utilizar.

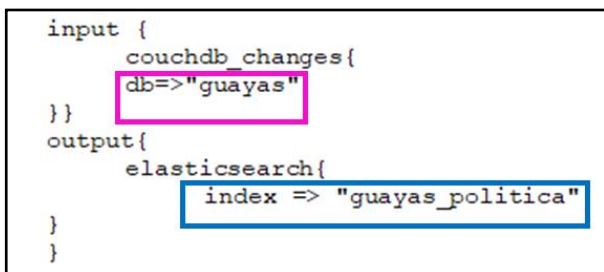


Ilustración 21 Archivo de configuración

- Este archivo debe estar guardado en la carpeta **bin** de Logstash. Ejecutamos el archivo desde Logstash.
`C:\elk\logstash-6.6.0\bin>logstash -f couchguayas.conf`

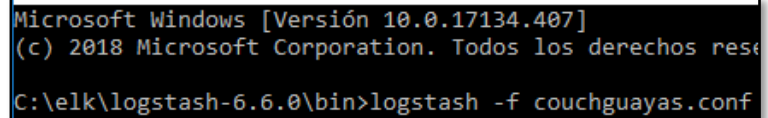


Ilustración 22 Ejecución del archivo de configuración en Logstash

- Podemos verificar que los documentos se están guardando correctamente desde cerebro, revisamos el índice que estamos usando y observamos el registro de los documentos.

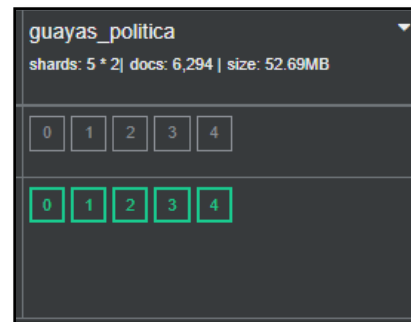


Ilustración 23 Carga de datos

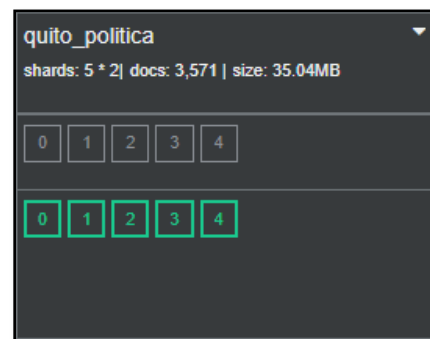


Ilustración 24 Carga de datos

- También vamos a recopilar datos directamente con Logstash con el fin de tener más datos para el respectivo análisis.

- Usamos el archivo de configuración **tweetguayas.conf**, del mismo modo el nombre va a variar de acuerdo con el nombre de cada ciudad seleccionada. En este archivo se establecerá:

- Las credenciales de Twitter
- Las palabras claves de búsqueda
- El nombre del índice que se va a usar
- El tipo de documento

```

input {
  twitter {
    consumer_key => "m4Fq2Pr4yHn1YLLg6nmPYxXYz"
    consumer_secret => "0WC0Z1D9sT4aMi8Y5xDrRjFIQqT3KbU8oSaNEsFEkKPHZCAsE4"
    oauth_token => "999027411613356032-NvGF9YveYjVjQq4sf61x5IbFDe0KBj"
    oauth_token_secret => "ZHTEn2rx8oKLIKfa57Ksm3Hs4jYtwimhgYcsq8TtAXVXv"

    keywords => ["alcaldia, candidatos, guayas"]

    full_tweet => true
  }
}
filter{
}
output {
  elasticsearch {
    index => guayas_politica
    document_type=> doc
  }
}

```

Ilustración 25 Archivo de configuración tweetguayas.conf

- Guardamos el archivo en la carpeta **bin** de Logstash. Ejecutamos el archivo desde Logstash.
C:\elk\logstash-6.6.0\bin>logstash -f tweetguayas.conf

```

Microsoft Windows [Versión 10.0.17134.407]
(c) 2018 Microsoft Corporation. Todos los derechos reservados
C:\elk\logstash-6.6.0\bin>logstash -f tweetguayas.conf

```

Ilustración 26 Ejecución del archivo de configuración en Logstash

- En cerebro podemos verificar que los datos se están guardando correctamente en el índice que estamos usando.



Ilustración 27 Carga de datos

VISUALIZACIÓN Y ANÁLISIS DE INFORMACIÓN

- Para visualizar la información que tenemos almacenada en Elasticsearch usamos la herramienta Kibana.

- Levantamos el servicio de Kibana.

```

C:\Windows\System32\cmd.exe - kibana
log [02:04:19.115] [info][kibana-monitoring][monitoring-ui] Starting m
onitoring stats collection
log [02:04:19.133] [info][status][plugin:security@6.5.4] Status change
d from yellow to green - Ready
log [02:04:19.342] [info][license][xpack] Imported license information
from Elasticsearch for the [monitoring] cluster: mode: basic | status: ac
tive
log [02:04:32.325] [info][listening] Server running at http://localhos
t:5601
log [02:04:32.343] [info][status][plugin:spaces@6.5.4] Status changed
from yellow to green - Ready

```

Ilustración 28 Servicio Kibana

- Ingresamos al navegador y escribimos **localhost:5601**, y podemos ver el servidor Kibana en ejecución.

- Aquí creamos los **index pattern** de cada una de las ciudades seleccionadas, si todo está bien podremos visualizar los datos registrados.

```

★ ambato_politica*
cuenca_politica*
guayas_politica*

```

Ilustración 29 Índice tweet en Kibana.

- Realizamos las visualizaciones correspondientes a cada ciudad y un análisis de cada visualización.

Ciudad de Guayas:

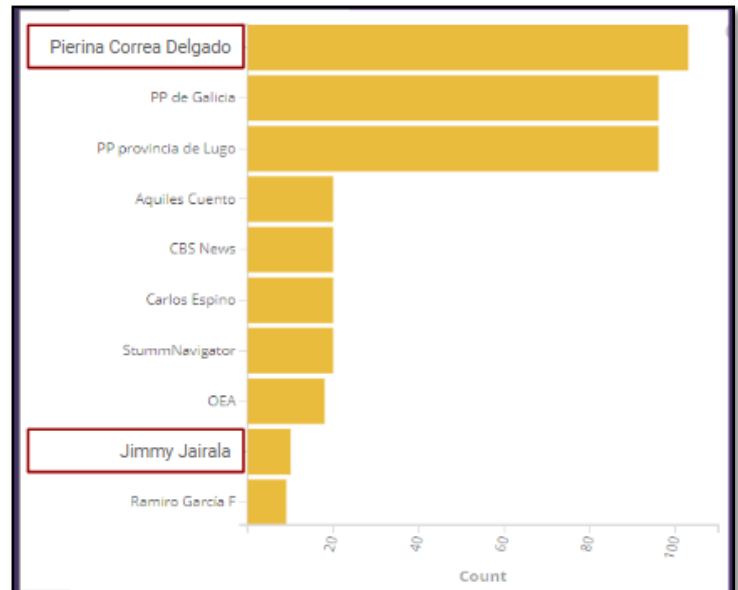


Ilustración 30 Candidatos de la ciudad de Guayas.

- En esta gráfica se puede observar que la candidata que ha sido mayormente mencionada para las próximas elecciones de alcaldía y prefecturas es Pierina Correa, pese

a ello hay que considerar que las menciones pueden o no ser positivas, pero la gente conoce sobre ella.

Otro de los demás candidatos mencionado es Jymmy Jairala candidato para la alcaldía de Guayaquil, no tiene gran número de menciones sin embargo la gente habla de él.

Ciudad de Ambato:

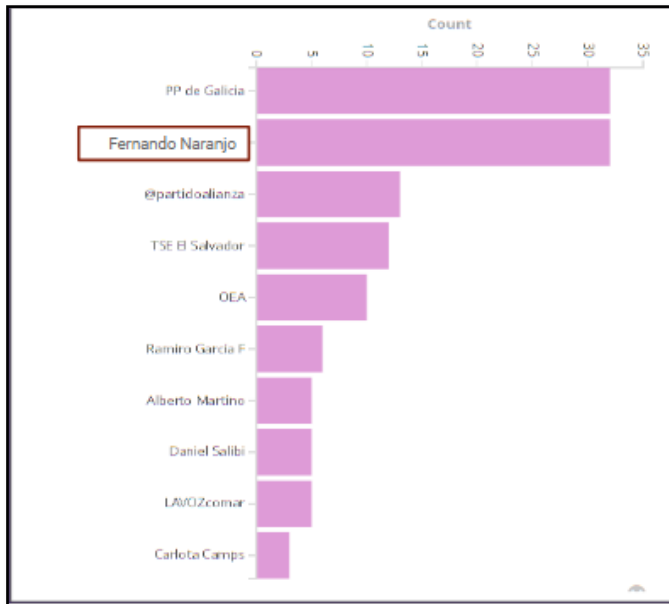


Ilustración 31 Candidatos de la ciudad de Ambato.

- En esta gráfica se puede observar que solo uno de los candidatos postulado para la alcaldía de la ciudad de Ambato ha sido mencionado en los tweets, pero tiene gran número de menciones. Lo que se puede concluir que el resto de los candidatos a un no están definidos o no ha hecho nada por darse a conocer. El único candidato mencionado es Fernando Naranjo.

Ciudad de Cuenca:



Ilustración 32 Candidatos de la ciudad de Cuenca.

- En esta gráfica se puede observar que el candidato con mayores menciones para la alcaldía de Cuenca es JEFF que es Jefferson Pérez, seguido de Gustavo Naranjo. No se ha podido observar el nombre de más de los candidatos por lo que suponemos que no tienen mayor relevancia en la población de esta ciudad.

Ciudad de Quito:

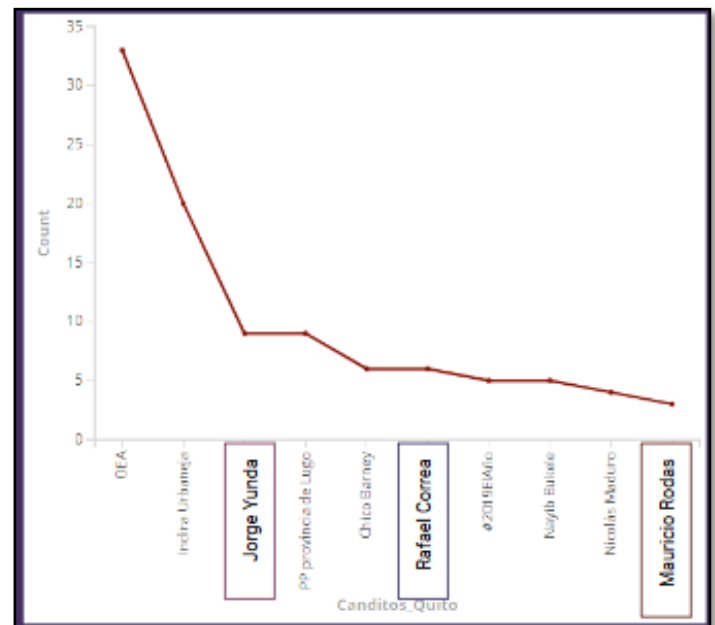


Ilustración 33 Candidatos de la ciudad de Quito.

- En esta gráfica se puede observar que los candidatos mencionados para la alcaldía de Quito son Jorge Yunda, Rafael Correa y Mauricio Rodas. Tiene mayor número de tweets Jorge Yunda, con respecto al Rafael Correa suponemos que el es mencionado a

favor de la lista 5. Mauricio Rodas también es mencionado, pero con menor incidencia.

Ciudad de Tulcán:

Candidatos_Tulcan	Count
Nayib Bukele	122
A Primera Hora	12
ARENA	12
OEA	11
@partidoalianza	10
PP de Galicia	10
PP provincia de Lugo	10
FMLN Oficial	8
Carlos Calleja	7
CNE Ecuador	6

Ilustración 34 Candidatos de la ciudad de Tulcán

- En esta gráfica podemos observar que ninguno de los candidatos postulados a la alcaldía ha sido nombrado, creemos que esto se debe a que a un no se han definido las personas que van a postularse para este cargo. O que no se han dado a conocer aún.

RESULTADOS OBTENIDOS

Lo que hemos podido observar es que las personas muestran desconocimiento de los candidatos a la alcaldía de sus respectivas ciudades por medio de las menciones en sus tweets.

TOP 10 TWITTEROS

Para este caso de estudio vamos a recopilar información sobre el Turismo en cinco ciudades de Ecuador: Santo Domingo, Manta, Machala, Ibarra y Portoviejo.

El propósito de analizar estos datos es exponer un Top 10 Twitteros sobre el turismo que se realiza en estas ciudades y conocer que ciudad les parece más atractiva y turística a estos twitteros.

ARQUITECTURA PARA LA SOLUCIÓN TOP 10 TWITTEROS

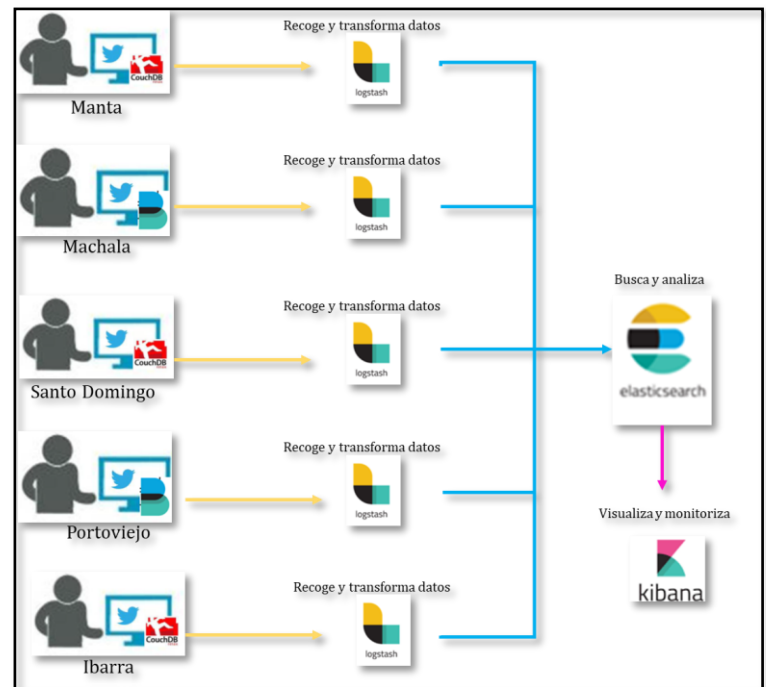


Ilustración 35 Arquitectura de la solución Pulso Político

EXTRACCIÓN DE DATOS

- Vamos a usar un script en Python, así que elaboramos el script en Python para que nos permita recoger datos de Twitter hacia nuestra base de datos CouchDB, cada script será nombrado de acuerdo con la ciudad de la cual se está minando datos. Así **turismo_Machala.py**.

- Para el script primero importamos las librerías couchdb, tweepy y json.

```
import couchdb # Libreria de CouchDB (requiere ser
from tweepy import Stream # tweepy es la libreria d
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json # Libreria para manejar archivos JSON
```

Ilustración 36 Librerías de Python

- Usamos las credenciales de la cuenta de Twitter.

```
ckey = "b0hhI00RfXbSgPjjBS1cAOEkB"
csecret = "3825kTPBIfoCkktKEZ1Q5aMjJe9HqMq8RD9P3sX0Tz1gf0p0dd"
atoken = "1268502774-9G1na0czUOXChiac8hs3S31c976JHExouRJGm5c"
asecret = "PAhtCrPxacs2AmTyIOCSKMGfPQzuF9j1l07n2bT7ptgr"
```

Ilustración 37 Credenciales de Twitter

- Tenemos que especificar la dirección URL de nuestro servidor de CouchDB, poner las credenciales de usuario en caso de que las tengamos.

```
# Setear la URL del servidor de couchDB
server = couchdb.Server('http://admin:1234@localhost:5984/')
```

Ilustración 38 Dirección URL del servidor CouchDB

- Debemos especificar la base de datos en la que queremos recopilar los datos. Este script busca si la base de datos existe en el servidor y si no es así la crea. Como vamos a recopilar datos de cinco ciudades en cada script va a variar el nombre de la base de datos, que va a ser nombrada de acuerdo con cada ciudad de la que se va a recopilar datos.

```
# Si no existe la Base de datos la crea
db = server.create('machala')
except:
# Caso contrario solo conectarse a la base existente
db = server['machala']
```

Ilustración 39 Creación de la base de datos

- Vamos a utilizar palabras claves para la búsqueda de datos, en este caso vamos a usar las palabras **Machala**, **turismo** ya que queremos información acerca de los twitters que se refieren al turismo en dichas ciudades. En cada script va a variar el nombre de la ciudad, ya que ahí deberá colocarse el nombre de la ciudad respectiva de la que se desea recoger los datos.

```
# Aqui se define el bounding box con los limites
twitterStream.filter(track=["Machala", "turismo"])
```

Ilustración 40 Filtración de datos por palabras claves

10. Ejecutamos el script y si todo esta correcto los datos empezaran a guardarse en la base de datos de CouchDB.

```
File Edit Shell Debug Options Window Help
===== RESTART: C:/Users/jenti
Guardado => 1092260295391748096
Guardado => 1092260299078623233
```

Ilustración 41 Almacenamiento de datos

11. Podemos verificar en CouchDB que los datos se están almacenando correctamente.

machala	449.6 KB	162
---------	----------	-----

Ilustración 42 Base de datos en CouchDB

12. Una vez que ya hemos recogido los datos de las ciudades seleccionadas de Ecuador vamos a pasar los datos a Elasticsearch.

- Primero vamos a levantar cada uno de los servicios:
 - Elasticsearch
 - Cerebro

```
C:\Windows\System32\cmd.exe - elasticsearch
Microsoft Windows [Versión 10.0.17134.523]
(c) 2018 Microsoft Corporation. Todos los derechos reservados.

C:\elk\elasticsearch-6.5.4\bin>elasticsearch
[2019-01-14T21:00:30,284][INFO ][o.e.e.NodeEnvironment ] [_GDGcNY]
using [1] data paths, mounts [[(C:)]], net usable_space [369.1gb], n
et total_space [464.4gb], types [NTFS]
[2019-01-14T21:00:30,308][INFO ][o.e.e.NodeEnvironment ] [_GDGcNY]
heap size [990.7mb], compressed ordinary object pointers [true]
[2019-01-14T21:00:32,354][INFO ][o.e.n.Node ] [_GDGcNY]
node name derived from node ID [_GDGcNYVS_6lMhePj6ehfQ]; set [node.n
```

Ilustración 43 Servicio Elasticsearch

- Una vez que ha iniciado Elasticsearch y cerebro vamos a ingresar al navegador y escribimos **localhost:9000** y podemos inicializar en cerebro.

13. Luego creamos los índices de mapeo vamos a usar el archivo de mapeo mencionado en el caso de estudio **Pulso Político**.

14. Creamos los índices en los cuales se van a ir guardando los datos de cada base de datos de CouchDB, se usará el mapeo ya indicado para cada índice. Cada índice tendrá como nombre **ciudad_turismo**, ciudad se reemplazará por el nombre de cada ciudad de la cual se recopiló los datos.

machala_turismo	manta_turismo
shards: 5 * 2 docs: 0 size: 1.12KB	shards: 5 * 2 docs: 0 size: 1.12KB
0 1 2 3 4	0 1 2 3 4
0 1 2 3 4	0 1 2 3 4

Ilustración 44 Creación de índices caso de estudio Top 10 twitters

portoviejo_turismo	ibarra_turismo
shards: 5 * 2 docs: 0 size: 1.12KB	shards: 5 * 2 docs: 0 size: 1.12KB
0 1 2 3 4	0 1 2 3 4
0 1 2 3 4	0 1 2 3 4

Ilustración 45 Índices caso de estudio Top 10 twitters



Ilustración 46 Creación de índices caso de estudio Top 10 twitteros



Ilustración 49 Carga de datos

15. Ahora usamos Logstash.

- Vamos a usar el archivo de configuración **couchciudad.conf**, ciudad se ira reemplazando por cada ciudad seleccionada así **couchmachala.conf**. En este archivo vamos a nombrar la base de datos que vamos a utilizar de CouchDB y el nombre del índice que vamos a utilizar.

```
input {
  couchdb changes{
    db=>"machala"
  }
}
output{
  elasticsearch{
    index => "machala_turismo"
  }
}
```

Ilustración 47 Archivo de configuración

- Este archivo debe estar guardado en la carpeta **bin** de Logstash. Ejecutamos el archivo desde Logstash.
C:\elk\logstash-6.6.0\bin>logstash -f couchmachala.conf

```
Microsoft Windows [Versión 10.0.17134.407]
(c) 2018 Microsoft Corporation. Todos los derechos reservados
C:\elk\logstash-6.6.0\bin>logstash -f couchmachala.conf
```

Ilustración 48 Ejecución del archivo de configuración en Logstash

- Podemos verificar que los documentos se están guardando correctamente desde cerebro, revisamos el índice que estamos usando y observamos el registro de los documentos.

16. También vamos a recopilar datos directamente con Logstash con el fin de tener más datos para el respectivo análisis.

- Usamos el archivo de configuración **tweetmachala.conf**, del mismo modo el nombre va a variar de acuerdo con el nombre de cada ciudad seleccionada. En este archivo se establecerá:

- Las credenciales de Twitter
- Las palabras claves de búsqueda
- El nombre del índice que se va a usar
- El tipo de documento

```
input {
  twitter {
    consumer_key => "m4Fq2Pr4yHn1YLLg6nmPYxXYz"
    consumer_secret => "0WC0Z1D9sT4aMiBY5xDrRjFIQqT3KbU8oSaNeSFEkKPHZCAsE4"
    oauth_token => "999027411613356032-NvGF9YveYjVjQq4sf61x5IbFDe0KBej"
    oauth_token_secret => "ZHTEn2rxBoKLikFa57Ksm3Hs4jYtwimhgYcsq8TtAXVXv"

    keywords => ["Machala, turismo"]

    full_tweet => true
  }
}
filter{
}
output {
  elasticsearch {
    index => machala_turismo
    document_type=> doc
  }
}
```

Ilustración 50 Archivo de configuración tweetguayas.conf

- Guardamos el archivo en la carpeta **bin** de Logstash. Ejecutamos el archivo desde Logstash.
C:\elk\logstash-6.6.0\bin>logstash -f tweetmachala.conf

```
Microsoft Windows [Versión 10.0.17134.407]
(c) 2018 Microsoft Corporation. Todos los derechos reservados.

C:\elk\logstash-6.6.0\bin>logstash -f tweetmachala.conf
```

Ilustración 51 Ejecución del archivo de configuración en Logstash

- En cerebro podemos verificar que los datos se están guardando correctamente en el índice que estemos usando.



Ilustración 52 Carga de datos

VISUALIZACIÓN Y ANÁLISIS DE INFORMACIÓN

- Para visualizar la información que tenemos almacenada en Elasticsearch usamos la herramienta Kibana.

- Levantamos el servicio de Kibana.

```
C:\Windows\System32\cmd.exe - kibana
log [02:04:19.115] [info][kibana-monitoring][monitoring-ui] Starting monitoring stats collection
log [02:04:19.133] [info][status][plugin:security@6.5.4] Status changed from yellow to green - Ready
log [02:04:19.342] [info][license][xpack] Imported license information from Elasticsearch for the [monitoring] cluster: mode: basic | status: active
log [02:04:32.325] [info][listening] Server running at http://localhost:5601
log [02:04:32.343] [info][status][plugin:spaces@6.5.4] Status changed from yellow to green - Ready
```

Ilustración 53 Servicio Kibana

- Ingresamos al navegador y escribimos **localhost:5601**, y podemos ver el servidor Kibana en ejecución.

- Aquí creamos los **index pattern** de cada una de las ciudades seleccionadas, si todo está bien podremos visualizar los datos registrados.

```
ibarra_t*
machala_t*
manta_t*
portoviejo_t*
```

Ilustración 54 Índice tweet en Kibana

- Realizamos las visualizaciones correspondientes a cada ciudad.

Ciudad de Manta:

Twitteros con más publicaciones sobre Turismo en Manta	Count
The Depths Below	112
Jorge Hurtado	83
Sin Censura	69
Amarildo, o possessor!	47
Daniel Mendoza	43
LaSole 🌞🌧️	34
Entropía.	30
Héctor Astudillo	26
Politburó de Tepetongo ru	19
LaHistoria	18

Ilustración 55 Top 10 de lugares turísticos de la ciudad de Manta.

- En la Ilustración 54 se puede observar que la persona con más números de tweets acerca de los lugares turísticos en la ciudad de Manta es la persona con el nickname The Depths Below

Ciudad de Ibarra:

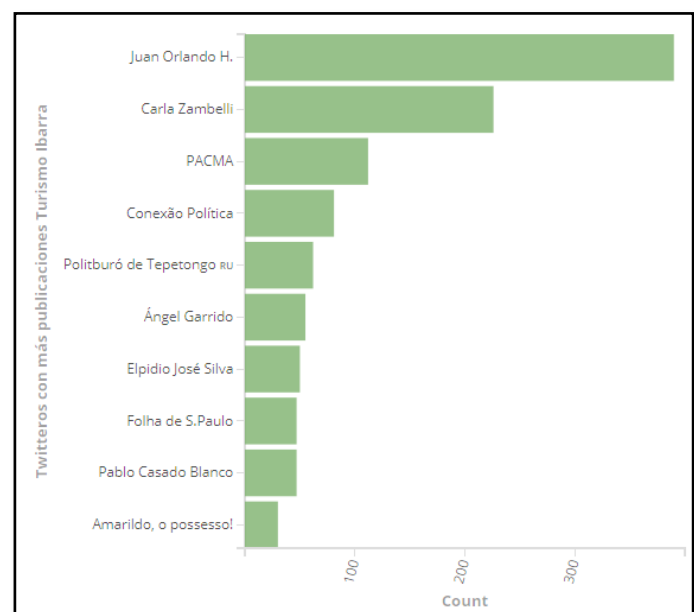


Ilustración 56 Top 10 de lugares turísticos de la ciudad de Ibarra.

- En la Ilustración 55 se puede apreciar que la persona con más menciones acerca de lugares turísticos de la ciudad de Ibarra es Juan Orlando H.

Ciudad de Santo Domingo:

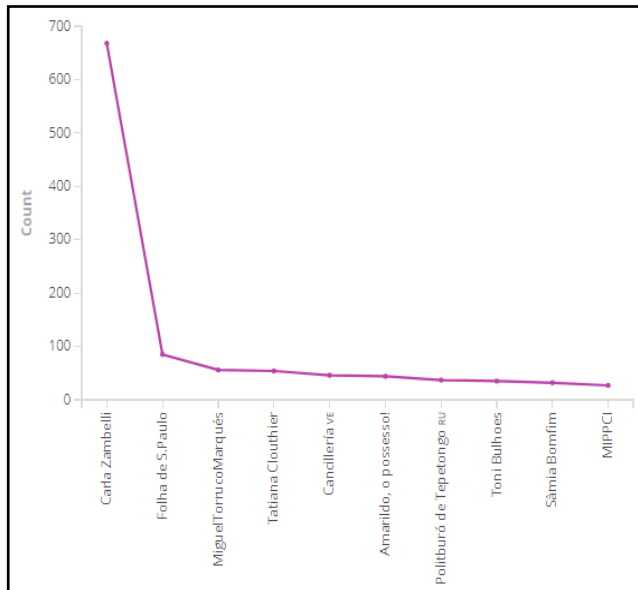


Ilustración 57 Top 10 de lugares turísticos de la ciudad de Santo Domingo.

- En la ilustración 56 podemos observar que la persona con más menciones acerca de los lugares turísticos de la ciudad de Santo Domingo de los Tsachilas es la persona con el nombre Carlita Zambelli.

Ciudad de Machala:

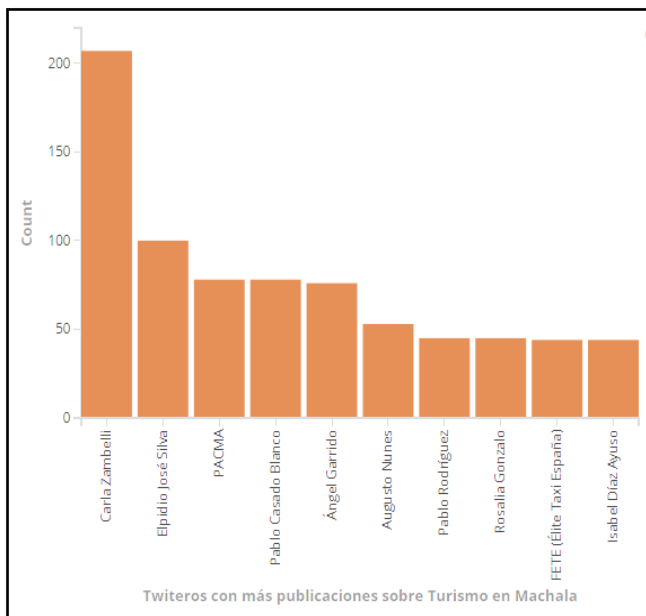


Ilustración 58 Top 10 de lugares turísticos de la ciudad de Machala.

- Como podemos observar en la ilustración 57 la persona con más menciones en sus tweets acerca de los sitios turísticos de la ciudad de Machala es Carla Zambelli.

Ciudad de Portoviejo:

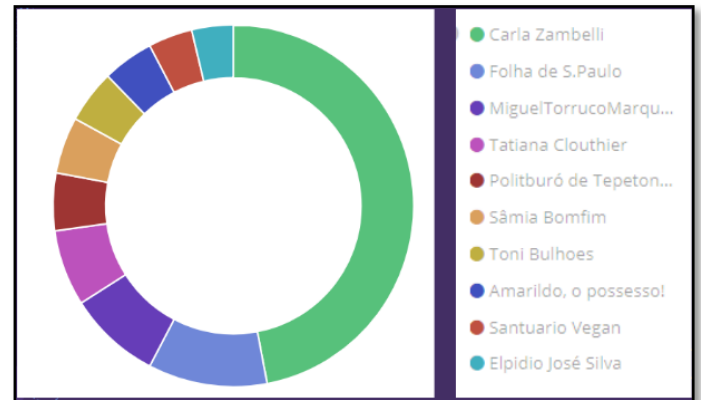


Ilustración 5960 Top 10 de lugares turísticos de la ciudad de Portoviejo.

- En esta gráfica se puede observar que Carla Zambelli es una de las personas que mayormente poste sobre turismo en la ciudad de Portoviejo.

RESULTADOS OBTENIDOS

Se puede observar que una de las ciudades con más número de tweets acerca de sus lugares turísticos es Santo Domingo. Y que la persona que realiza gran número de publicaciones sobre turismo es Carla Zambelli ya que su nombre se ha podido evidenciar en todos los tweets de las ciudades estudiadas.

VIOLENCIA EN ECUADOR

Para este caso de estudio vamos a recopilar información sobre la violencia que actualmente vivimos en el país, continuamente hemos venido escuchando que se han producido robos, asesinatos, violaciones, niños que son apartados de sus hogares, etc.

De acuerdo con esto queremos conocer que es lo que está comentando la sociedad de estos delitos y queremos conocer que ciudades y países postean sobre violencia.

ARQUITECTURA PARA LA SOLUCIÓN VIOLENCIA EN ECUADOR

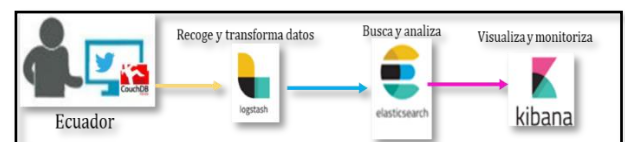


Ilustración 61 Arquitectura de la solución Violencia en Ecuador

EXTRACCIÓN DE DATOS

5. Vamos a usar un script en Python, así que elaboramos el script en Python para que nos permita recoger datos de Twitter hacia nuestra base de datos CouchDB. Este archivo se llamará **violencia_Ecuador.py**.

- Para el script primero importamos las librerías couchdb, tweepy y json.

```
import couchdb # Libreria de CouchDB (requiere ser
from tweepy import Stream # tweepy es la libreria de
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json # Libreria para manejar archivos JSON
```

Ilustración 62 Librerías de Python

- Usamos las credenciales de la cuenta de Twitter.

```
ckey = "b0hhIO0RfXb9gPjjBSlcAOEkB"
csecret = "3825kTPBIf0CkktKEZ1Q5aMjJe9HqMq8RD9P3sX0Tzlgf0p0dd"
atoken = "1268502774-9Glna0czUOXCbic8hs3S3lc976JHExouRJGm5c"
asecret = "PAhtCrPxacjS2AmTyIOCSKmGfPQzuF9j1107n2bT7ptgr"
```

Ilustración 63 Credenciales de Twitter

- Tenemos que especificar la dirección URL de nuestro servidor de CouchDB, poner las credenciales de usuario en caso de que las tengamos.

```
# Setear la URL del servidor de couchDB
server = couchdb.Server('http://admin:1234@localhost:5984/')
```

Ilustración 64 Dirección URL del servidor CouchDB

- Debemos especificar la base de datos en la que queremos recopilar los datos. Este script busca si la base de datos existe en el servidor y si no es así la crea.

```
try:
    # Si no existe la Base de datos la crea
    db = server.create('violencia')
except:
    # Caso contrario solo conectarse a la base existente
    db = server['violencia']
```

Ilustración 65 Creación de la base de datos

- Vamos a utilizar palabras claves para la búsqueda de datos, en este caso vamos a usar las palabras **violencia**, **Ecuador** ya que queremos información acerca de la violencia que actualmente estamos viviendo en el país.

```
# Aqui se define el bounding box con los limites geo
twitterStream.filter(track=["violencia", "Ecuador"])
```

Ilustración 66 Filtración de datos por palabras claves

17. Ejecutamos el script y si todo esta correcto los datos empezaran a guardarse en la base de datos de CouchDB.

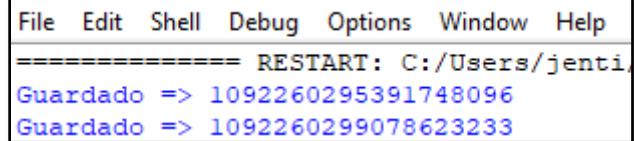


Ilustración 67 Almacenamiento de datos

18. Podemos verificar en CouchDB que los datos se están almacenando correctamente.

violencia	49.3 MB	17134
-----------	---------	-------

Ilustración 68 Base de datos en CouchDB

19. Una vez que ya hemos recogido los datos de violencia en Ecuador vamos a pasar los datos a Elasticsearch.

- Primero vamos a levantar cada uno de los servicios:
 - Elasticsearch
 - Cerebro

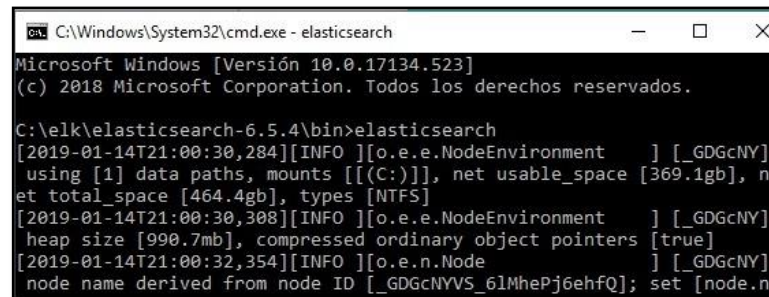


Ilustración 69 Servicio Elasticsearch

- Una vez que ha iniciado Elasticsearch y cerebro vamos a ingresar al navegador y escribimos **localhost:9000** y podemos inicializar en cerebro.

20. Luego creamos el índice de mapeo vamos a usar el archivo de mapeo mencionado en el caso de estudio **Pulso Político**.

21. Creamos el índice en los cuales se van a ir guardando los datos de cada base de datos de CouchDB, se usará el mapeo ya indicado para este índice. El índice tendrá como nombre **violencia_ecuador**.



Ilustración 70 Creación de índice

22. Ahora usamos Logstash.

- Vamos a usar el archivo de configuración **couchviolencia.conf**. En este archivo vamos a nombrar la base de datos que vamos a utilizar de CouchDB y el nombre del índice que vamos a utilizar.

```
input {
  couchdb changes{
    db=>"violencia"
  }
}
output{
  elasticsearch{
    index => "violencia_ecuador"
  }
}
```

Ilustración 71 Archivo de configuración

- Este archivo debe estar guardado en la carpeta **bin** de Logstash. Ejecutamos el archivo desde Logstash.
C:\elk\logstash-6.6.0\bin>logstash -f couchviolencia.conf

```
Microsoft Windows [Versión 10.0.17134.407]
(c) 2018 Microsoft Corporation. Todos los derechos reservados
C:\elk\logstash-6.6.0\bin>logstash -f couchviolencia.conf
```

Ilustración 72 Ejecución del archivo de configuración en Logstash

- Podemos verificar que los documentos se están guardando correctamente desde cerebro, revisamos el índice que estamos usando y observamos el registro de los documentos.



Ilustración 73 Carga de datos

23. También vamos a recopilar datos directamente con Logstash con el fin de tener más datos para el respectivo análisis.

- Usamos el archivo de configuración **tweetviolencia.conf**. En este archivo se establecerá:
 - Las credenciales de Twitter
 - Las palabras claves de búsqueda
 - El nombre del índice que se va a usar

➤ El tipo de documento

```
input {
  twitter {
    consumer_key => "m4Fq2Pr4yHn1YLLg6nmPYxXYZ"
    consumer_secret => "0WC0Z1D9sT4aMi8Y5xDrRjFIQqT3KbU8oSaNEsFEkKPHZCAsE4"
    oauth_token => "999027411613356032-NvGF9YveYjVjQq4s61x5IbFDe0KBej"
    oauth_token_secret => "ZHTEn2rx8oKLIkFa57Ksm3Hs4jYtwimhgYcsq8TtAXVXv"

    keywords => ["violencia, Ecuador"]

    full_tweet => true
  }
}
filter{
}
output {
  elasticsearch {
    index => violencia_ecuador
    document_type=> doc
  }
}
```

Ilustración 74 Archivo de configuración tweetguayas.conf

- Guardamos el archivo en la carpeta **bin** de Logstash. Ejecutamos el archivo desde Logstash.
C:\elk\logstash-6.6.0\bin>logstash -f tweetviolencia.conf

```
Microsoft Windows [Versión 10.0.17134.407]
(c) 2018 Microsoft Corporation. Todos los derechos reservados
C:\elk\logstash-6.6.0\bin>logstash -f tweetviolencia.conf
```

Ilustración 75 Ejecución del archivo de configuración en Logstash

- En cerebro podemos verificar que los datos se están guardando correctamente en el índice que estamos usando.



Ilustración 76 Carga de datos

VISUALIZACIÓN Y ANÁLISIS DE INFORMACIÓN

- Para visualizar la información que tenemos almacenada en Elasticsearch usamos la herramienta Kibana.
 - Levantamos el servicio de Kibana.


```

C:\Windows\System32\cmd.exe - kibana
log [02:04:19.115] [info][kibana-monitoring][monitoring-ui] Starting m
onitoring stats collection
log [02:04:19.133] [info][status][plugin:security@6.5.4] Status change
d from yellow to green - Ready
log [02:04:19.342] [info][license][xpack] Imported license information
from Elasticsearch for the [monitoring] cluster: mode: basic | status: ac
tive
log [02:04:32.325] [info][listening] Server running at http://localhos
t:5601
log [02:04:32.343] [info][status][plugin:spaces@6.5.4] Status changed
from yellow to green - Ready

```

Ilustración 77 Servicio Kibana

7. Ingresamos al navegador y escribimos **localhost:5601**, y podemos ver el servidor Kibana en ejecución.
 - Aquí creamos el **index pattern** del índice creado en Logstash si todo está bien podremos visualizar los datos registrados.
8. Realizamos las visualizaciones correspondientes a los datos.

Violencia en Ecuador:

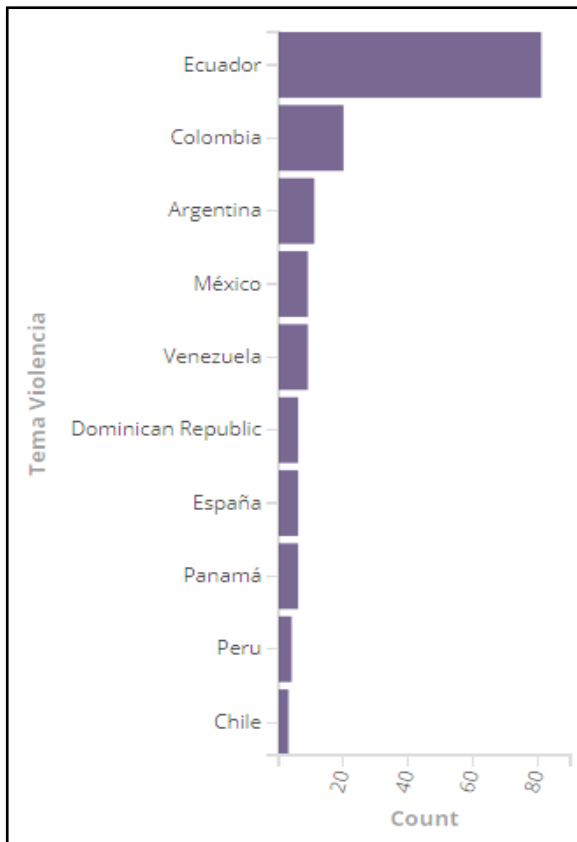


Ilustración 78 Países desde donde se hicieron tweets de temas de violencia

- En esta gráfica se puede observar los diez países que más postean sobre violencia, teniendo como resultado que Ecuador es uno de los países que más postea sobre violencia

lo que nos hace pensar que en la actualidad debemos cuidarnos mucho más para no ser víctimas de delitos.

El país que menos tweets sobre violencia ha posteado es Chile.

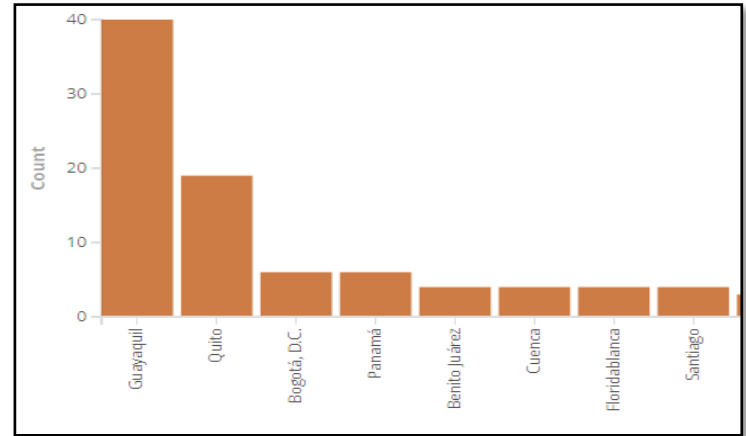


Ilustración 79 Ciudades desde donde se hicieron tweets de temas de violencia

- En esta gráfica podemos observar las diez ciudades que más postean sobre el tema violencia, siendo Guayaquil la ciudad que más índice de violencia tiene, seguido de Quito y Cuenca dentro del país.

RESULTADOS OBTENIDOS

La mayor cantidad de número de tweets acerca del tema se encuentra en la provincia de Guayaquil y podemos agregar que el país que más ha difundido tweets acerca del tema de violencia en el Ecuador es Colombia.

VI. RÉPLICAS DE LAS BASES DE DATOS DE COUCHDB

1. Puesto que los datos fueron recopilados en diferentes máquinas tenemos que agruparlos en una sola base de datos para ello hemos realizado las respectivas réplicas de las bases de datos.

- Primero obtenemos la dirección IP de la máquina de la cual se va a extraer los datos.

```

Dirección IPv4. . . . . : 192.168.1.4
Máscara de subred . . . . . : 255.255.255.192
Puerta de enlace predeterminada . . . . . : fe80::1%17
192.168.1.1

```

Ilustración 80 Dirección IP de la máquina

- Desde la otra máquina realizamos la réplica, tenemos que escribir la dirección IP de la máquina de la cual se va a extraer los datos y las credenciales del usuario.

Source

Type:

Remote database

Database URL:

http://192.168.1.4:5984/guayas

Authentication:

Username and password

admin

....

Ilustración 81 Fuente origen de los datos

- En la fuente destino vamos a escribir el nombre de la base de datos en la que vamos a guardar los datos y las credenciales del usuario.

Target

Type:

Existing local database

Name:

guayaquil

Authentication:

Username and password

admin

....

Ilustración 82 Fuente destino de los datos

- La réplica será de tipo **One Time** ya que no queremos que se guarden cambios o actualización que se hagan en dicha base de datos.

Options

Replication type:

One time

Replication document:

Custom ID (optional)

Start Replication

Clear

Ilustración 83 Tipo de réplica

2. Una vez que se han realizado todas las réplicas de las bases de datos, ya tenemos nuestras bases de datos de cada ciudad unificadas.

Filter replications						New Replication
Source	Target	Start Time	Type	State	Actions	
http://192.168.1.4:5984/santo_domingo	http://localhost:5984/santo_domingo_1	Feb 4th, 4:58 pm	One time	Completed		
http://192.168.1.4:5984/portoviejo	http://localhost:5984/portoviejo_1	Feb 4th, 4:58 pm	One time	Completed		
http://192.168.1.4:5984/machala	http://localhost:5984/machala_1	Feb 4th, 4:57 pm	One time	Completed		
http://192.168.1.4:5984/manta	http://localhost:5984/manta_1	Feb 4th, 4:55 pm	One time	Completed		
http://192.168.1.4:5984/guayas	http://localhost:5984/guayaquil	Feb 4th, 4:51 pm	One time	Completed		

Ilustración 84 Réplicas realizadas exitosamente

VII. DESCRIPCIÓN DEL EQUIPO DE TRABAJO Y ACTIVIDADES REALIZADAS

Tabla 1 Descripción de actividades realizadas por cada miembro del equipo

PROYECTO FINAL MINERÍA Y ANÁLISIS DE INFORMACIÓN	
MIEMBROS DEL EQUIPO	ACTIVIDADES REALIZADAS
Luis Altamirano	Minería de datos caso de estudio Pulso Político, ciudades Quito, Cuenca, Ambato y Azuay.
	Minería de datos caso de estudio Top 10 Twitteros, ciudades Ibarra, Santo Domingo, Portoviejo, Manta y Machala.
	Minería de datos caso de estudio Violencia en Ecuador.
	Réplicas de las bases de datos de couchDB.
	Creación de los índices en Logstash.
	Visualizaciones de los datos en Kibana caso de estudio Pulso Político.
	Análisis de los datos de Pulso Político.
	Visualización de los datos en Kibana caso de estudio Violencia en Ecuador
Jenny Tipán	Análisis de los datos de Violencia en Ecuador.
	Minería de datos caso de estudio Pulso Político, ciudad de Guayas.
	Minería de datos caso de estudio Top 10 twitteros, ciudades Manta, Machala, Portoviejo, Santo Domingo e Ibarra.
	Visualización de datos caso de estudio Top 10 twitteros en Kibana.
	Análisis de los datos de caso de estudio Top 10 twitteros.
	Elaboración del informe del proyecto.
	Elaboración del documento de presentación.
	Almacenar los datos recopilados en un CD.
	Subir archivos al repositorio de GitHub.

VIII. CRONOGRAMA DE ACTIVIDADES

Tabla 2 Cronograma de actividades

CRONOGRAMA DE ACTIVIDADES						
No.	FECHAS ACTIVIDADES	Sem1	Sem2	Sem3	Sem4	Sem5
1	Minería de datos caso de estudio Pulso Político					
2	Minería de datos caso de estudio Top 10 twitteros					
3	Minería de datos caso de estudio Violencia en Ecuador					
4	Réplicas y unificación de bases de datos					
5	Creación de índices mapeados en Logstash					
6	Inserción de datos en Elasticsearch					
7	Crear visualizaciones en Kibana de los datos recopilados por caso de estudio					
8	Análisis de los datos					
9	Elaboración del informe del proyecto					
10	Almacenamiento de los datos en un CD					
11	Elaboración de la presentación					
12	Subir archivos en el repositorio de GitHub					

- Se recomienda hacer con tiempo cada una de las actividades que involucran la ejecución del proyecto de tal forma que si se presenta inconvenientes podamos realizar averiguaciones o consultas de los mismo, caso contrario nos estancaremos y no podremos llevar a cabo el proyecto de la manera que se nos ha pedido.

X. PROBLEMAS ENCONTRADOS

- El mayor problema que hemos encontrado en la ejecución del proyecto es realizar visualizaciones de los datos que reflejen información relevante de estos, puesto que los datos de Twitter tienen varios parámetros y resulta complicado encontrar parámetros importantes para su estudio.
- Otro problema fue el uso de las credenciales de cuenta privada de Twitter para la búsqueda de datos, puesto que unas funcionaban en un momento y luego ya no, por lo que fue complicado la recolección de la información.

IX. CONCLUSIONES Y/O RECOMENDACIONES

- De acuerdo con los resultados encontrados podemos decir que la minería de datos sobre temas particulares y de interés social nos permite conocer que está pasando a nuestro alrededor y en base a ello tomar decisiones.
- El estudio de los datos llega a ser importantes si se realiza un buen análisis de estos, ya que de nada sirve tener una gran cantidad de datos sin no sabemos cómo manejarlos o explotarlos.
- Es recomendable recoger con anticipación los datos que necesitamos para su posterior análisis, ya que resulta demoroso hacerlo de un día al otro y sobre todo no se puede llegar a tener resultados óptimos o aceptables.
- Antes de empezar a realizar proyectos de minería de datos es necesario revisar antes las herramientas que necesitamos puesto que una vez que llegamos a ejecutar el proyecto y una de las herramientas usadas no es compatible con las demás provocaría que el proyecto se detenga, que tengamos que volver al inicio, que tengamos que nuevamente realizar las búsquedas y descargas de las herramientas apropiadas, esto nos hace perder tiempo y productividad.

CONTENIDO

I.	INTRODUCCIÓN	1
II.	OBJETIVO GENERAL	1
III.	OBJETIVOS ESPECÍFICOS	1
IV.	RECURSOS Y HERRAMIENTAS PARA UTILIZAR...	1
	HARDWARE	1
	SOFTWARE	1
V.	CASOS DE ESTUDIO	2
	PULSO POLÍTICO	2
	ARQUITECTURA SOLUCIÓN PULSO POLÍTICO	2
	EXTRACCIÓN DE DATOS	2
	VISUALIZACIÓN Y ANÁLISIS DE INFORMACIÓN	5
	RESULTADOS OBTENIDOS	7
	TOP 10 TWITTEROS	7
	ARQUITECTURA SOLUCIÓN TOP 10 TWITTEROS	7
	EXTRACCIÓN DE DATOS	7
	VISUALIZACIÓN Y ANÁLISIS DE INFORMACIÓN	10
	RESULTADOS OBTENIDOS	11
	VIOLENCIA EN ECUADOR	11
	ARQUITECTURA VIOLENCIA EN ECUADOR	11
	EXTRACCIÓN DE DATOS	12
	VISUALIZACIÓN Y ANÁLISIS DE INFORMACIÓN	13
	RESULTADOS OBTENIDOS	14
VI.	RÉPLICAS BASES DE DATOS DE COUCHDB	14
VII.	DESCRIPCIÓN DEL EQUIPO DE TRABAJO Y ACTIVIDADES REALIZADAS	15
VIII.	CRONOGRAMA DE ACTIVIDADES	16
IX.	CONCLUSIONES Y/O RECOMENDACIONES	16
X.	PROBLEMAS ENCONTRADOS	16