

ELASTIC STACK (ELK)

Maestría en Inteligencia de Datos Orientada a Big Data

Asignatura: Captura y Almacenamiento de Información

Profesor: Javier Bazzocco

Alumnos: María Paula Ponce, Gabriel Alejandro Migo

Año: 2021

1. **Introducción**

En el marco del aprendizaje de la materia “Captura y Almacenamiento de Información” hemos realizado este trabajo de investigación de herramientas que permiten capturar y alojar grandes volúmenes de datos.

Para nuestro caso, seleccionamos las herramientas vistas en clase del Stack Elk de Elastic Search. Una suite que contiene un motor de alojamiento NoSql , Elastic Search, un lenguaje de scripts para ETL (extracción, transformación y carga) Logstash y una herramienta de análisis - visualización Kibana.

1. **Objetivo**

A través de los procedimientos enunciados a continuación, se plantea la adquisición de conocimiento a partir del cumplimento de diferentes objetivos a saber:

- Elección de un dataset en formato CSV.

* Se instalarán las últimas versiones disponibles al momento del trabajo de investigación de los siguientes productos: Elasticsearch , Logstash y Kibana.
* Se realizará la configuración de un archivo “.conf” en el que se incluirá los datos de entrada
* Se le dará tratamiento a algunos campos de interés para adaptar la salida. Entre ellos, se aplicará filtros para que se reconozca la ubicación como tipo de dato geopoint.
* Se configurará un Mapping en Kibana del índice de Elastic Search que permita utilizar la información geo referencial.
* Se ejecutará un comando Logstash para efectivizar la ingesta del CSV en Elasticsearch.
* Se Establecerá la creación de un índice en Elastic Search para su análisis en Kibana
* Se generarán distintos tipos de visualizaciones con Kibana para su análisis.

1. **Elección del dataset**

Se seleccionó un dataset en formato CSV, que cuenta con información oficial de crímenes reportados en la ciudad de Nueva York desde el año 2006 al 2019. Se adjunta el archivo crimesNYdescripcionColumnas.xlsx el cuál detalla el significado de cada columna. Para el siguiente trabajo le daremos mayor relevancia a los datos “latitud” y “longitud”. El dataset puede ser accedido desde <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>

1. **Uso de Logstash**

**4.1- Confección del Archivo de configuración**

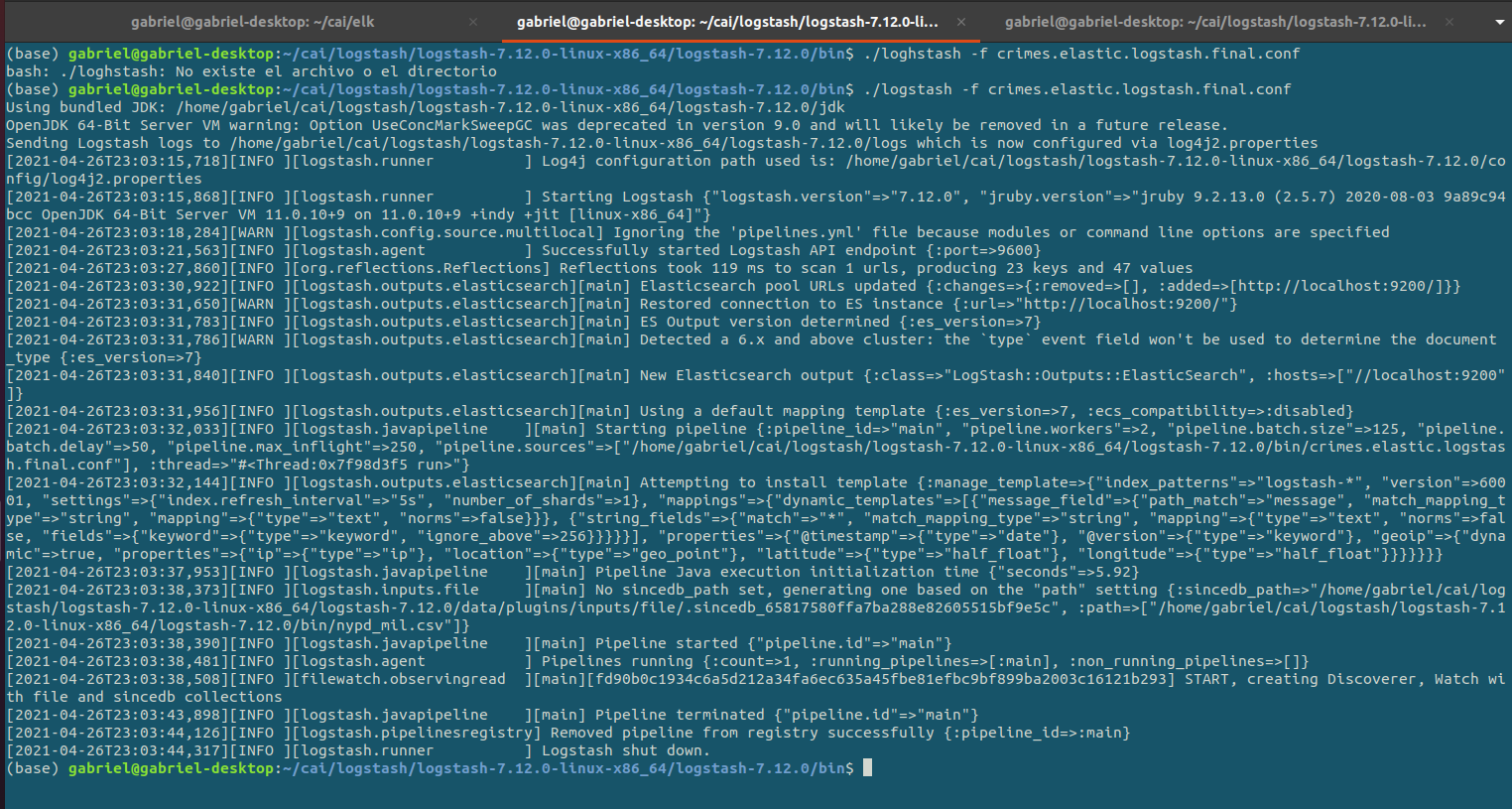
Se prepara el archivo de configuración a ser leído por Logstash en donde se incluye la definición de distintos tipos de datos pertenecientes al CSV NYPD\_Complaint\_Data\_Historic.csv (se adjuntan ambos archivos (**crimes.elastic.logstash.conf** y **NYPD\_Complaint\_Data\_Historic.csv**).

Además se borran campos creados automáticamente por Logstash como "message", "@version","host" , "path" y otros de nuestro CSV que encontramos irrelevantes. Como así también se co

nfigura un nuevo campo Location para la ubicación geo referencial. Se define que la salida sea a ElasticSearch y el nombre del indice.

**4.2- Ejecución del comando logstash –f crimes.elastic.logstash.conf**

Se ejecuta el comando logstash –f crimes.elastic.logstash.conf para realizar la ingesta del CSV en Elasticsearch (se adjunta log de consola **consola-logstash.log**)



**4.2.1- Inconvenientes encontrados**

Surgió el problema de limitaciones de capacidad de procesamiento de las PC hogareñas. Tomamos dos caminos: uno instaló cada producto por separado, y otro utilizando docker. Este proceso nos enriqueció en el dominio de los distintos productos y su infraestructura. Ya sea en la elaboración de un docker-compose (se adjunta el archivo **docker-compose.yml**) como así en el uso de los mapeos y configuraciones de logstash.

PC1:

Hardware y SO anfitrión: Intel Core i5-2500 3.30 Ghz, 8GB RAM - Windows 10 PRO

Hardware y SO huesped: Ubuntu Server 14.04.4 LTS 64 bits, memoria base 5281 MB, 4Procesadores, memoria Video 128 MB, aceleración 3D -- VirtualBox 6.1

Sobre Ubuntu, se instaló logstash-7-12.0 y el stack elk mediante docker compose, al no levantar elasticSearch, se realizó la misma prueba sobre Windows instalando las siguientes versiones:

Elasticsearch-7.12.0-windows-x86\_64

Logstash-7.12.0-windows-x86\_64

Kibana-7.12.0-windows-x86\_64

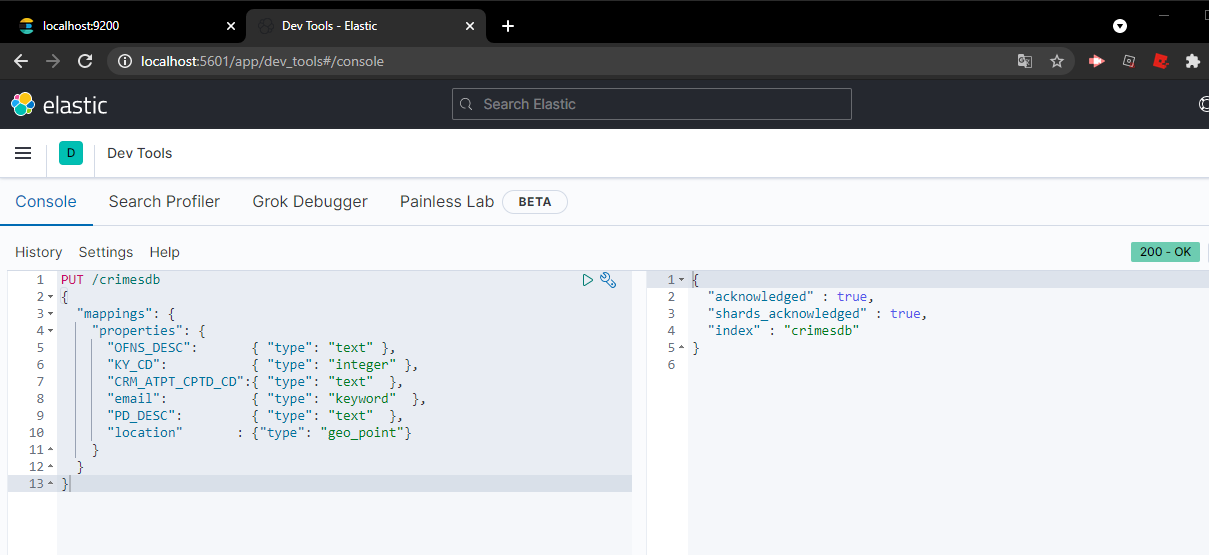
PC2:

Hardware y SO anfitrión: Intel(R) Celeron(R) CPU G3950 @ 3.00GHz, 8GB RAM - Ubuntu 20.04

Sobre Ubuntu, se instaló docker y el stack elk (Logstah, Elastic Search y Kibana todo en un mismo contenedor). Se creó un docker-compose para su arrancar y parar el contenedor.

**4.3- Elaboración de mapping en los índices de Elastic Search**

Se genera un mapeo de los atributos del índice a crear para que pueden ser ingestados por kibana en índices de patrones. Esto resulta necesario para poder crear capas de información sobre los mapas con la información del índice. Se ejecuta a través de la consola DevTools que ofrece Kibana desde la opción de configuración. Se adjunta el mapping **crimes\_map\_template.json**



**4.3.1- Inconvenientes encontrados**

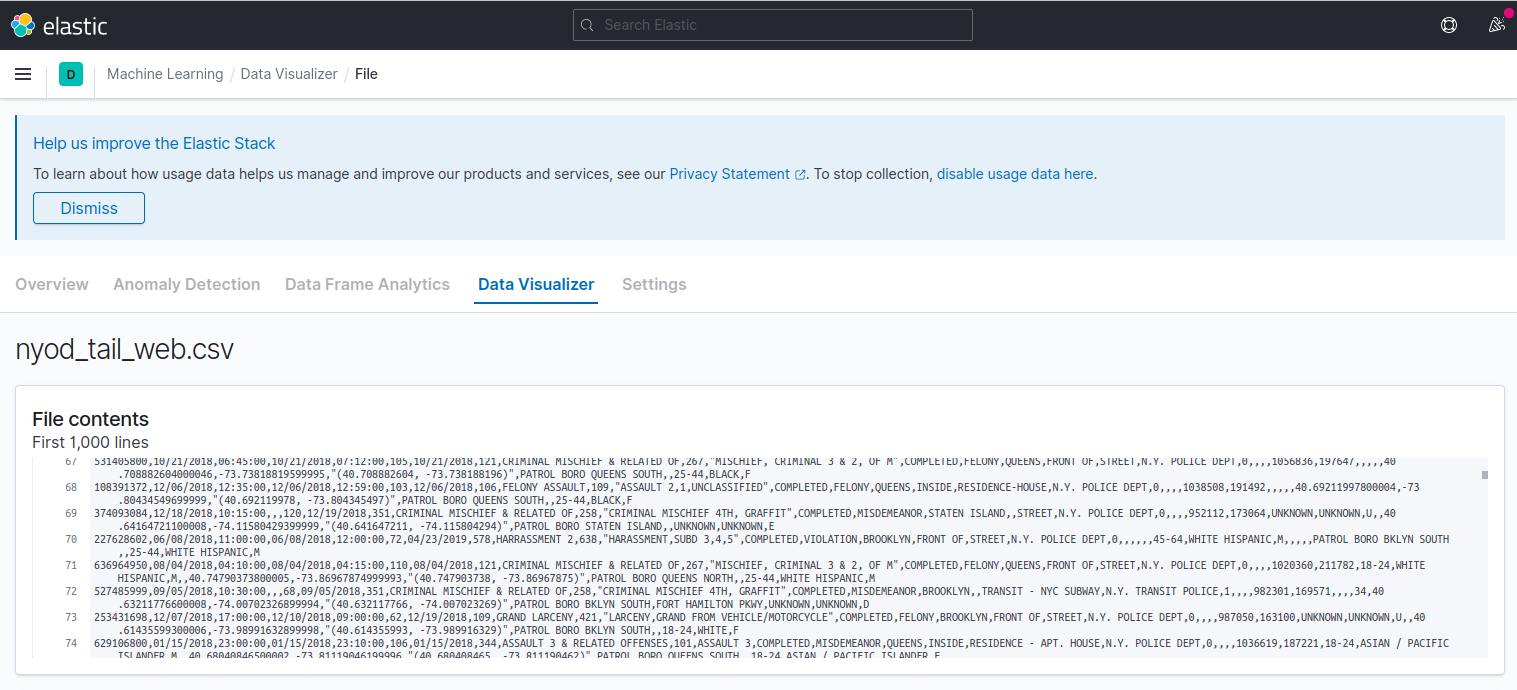
Se necesitaba utilizar la información Geo referencial. Se creó el campo “location” donde se unían latitud y longitud en el archivo de configuración del Logstash, pero esto no se reflejaba en el índice de la manera esperada. Luego de analizar el problema, nos dimos cuenta que la unión de los campos en el archivo .conf, no era suficiente para que el índice tome el tipo de datos adecuado. La solución resultó en realizar primero un mapeo específico para la creación del índice y luego asignarle los datos.

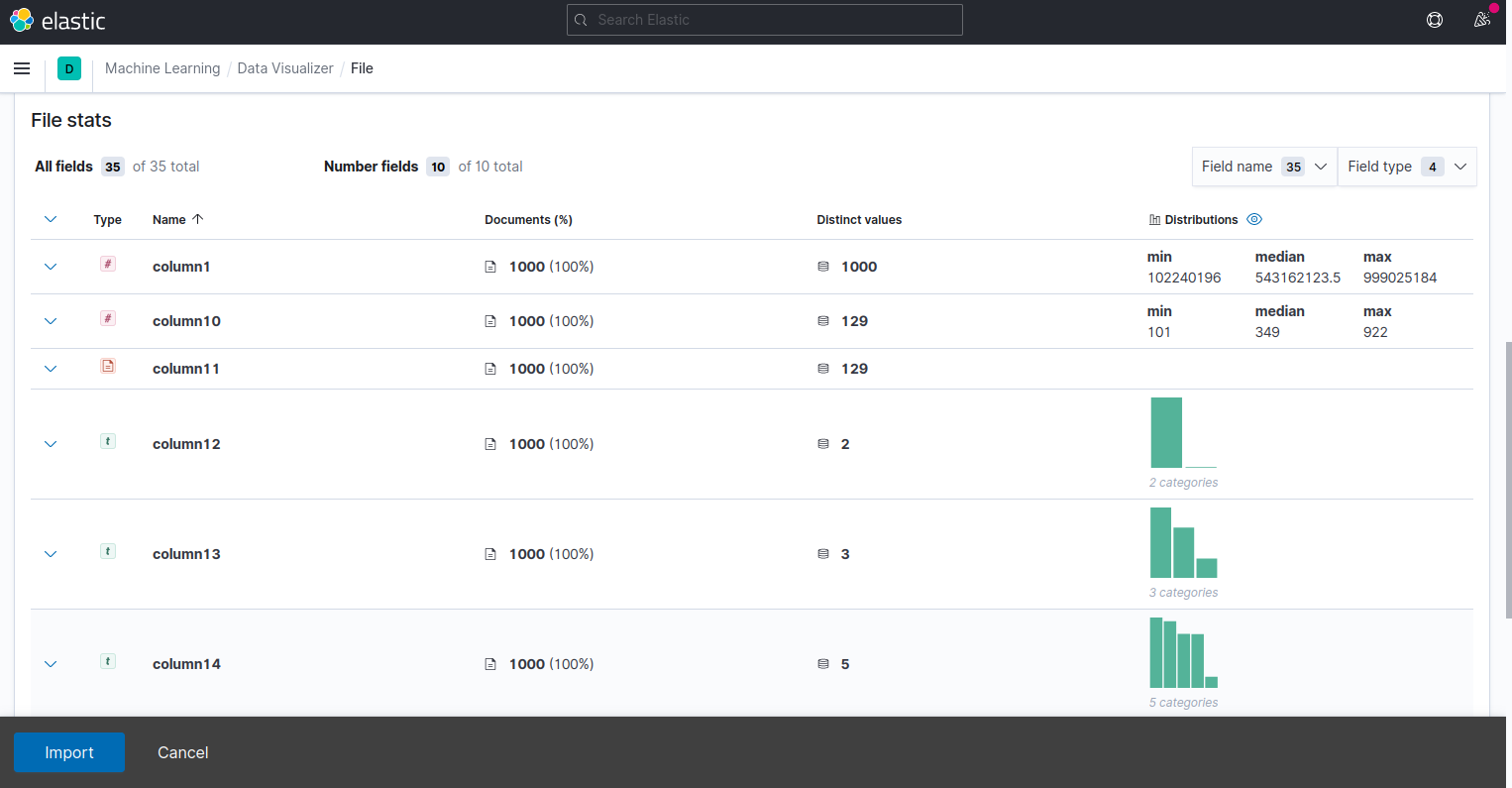
Por otro lado, para visualizar correctamente las coordenadas de los geo-point fue necesario en el archivo de configuración de Logstash primero especificar la longitud, seguido de la latitud, ya que de la otra manera mostraba las coordenadas invertidas.

Por último en el archivo de configuración de Logstash se trabajó para que se cree el nuevo campo location solamente para aquellos registros que tengan informada la latitud debido a que si no se realizaba lo anterior Kibana mostraba puntos con coordenadas (0,0) quedando errónea la visualización. Vale aclarar que los datos correspondientes a los registros sin latitud informada si se almacenan, sólo no se genera el geo-point.

**4.4- Carga de csv vía web**

La herramienta Kibana posee una interfaz de carga para distintos tipos de fuentes. Se procedió a cargar una parte del csv por este medio para la ingesta del Elastic Search.



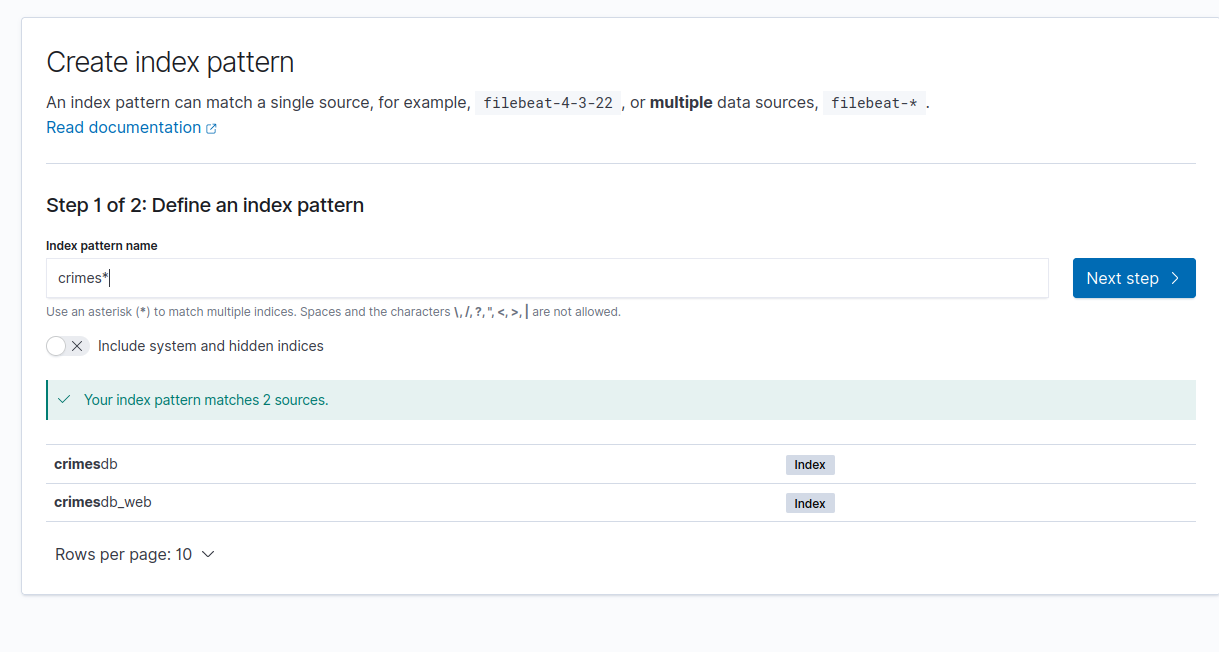


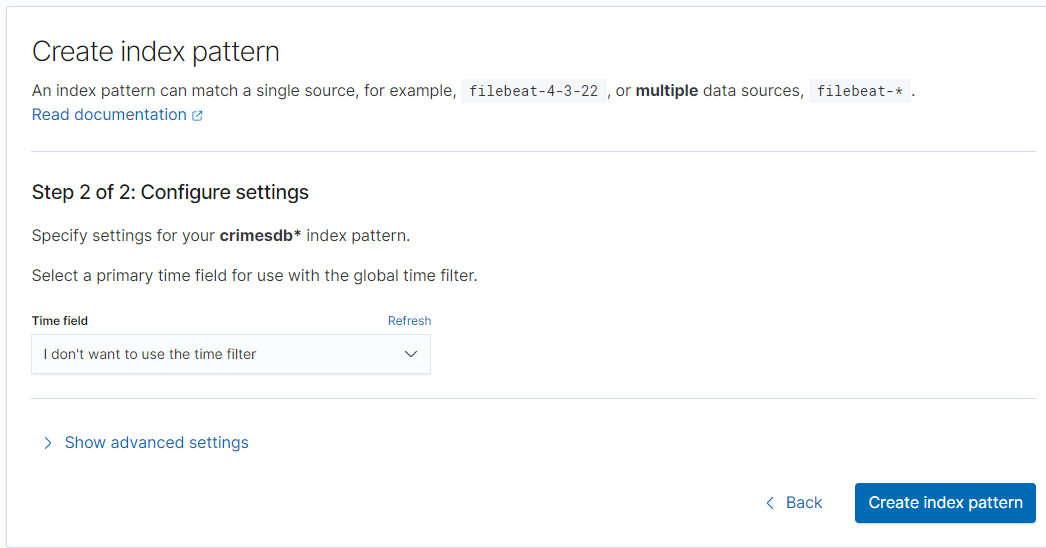
En este caso la interfaz ya realiza la creación del índice y el mapeo



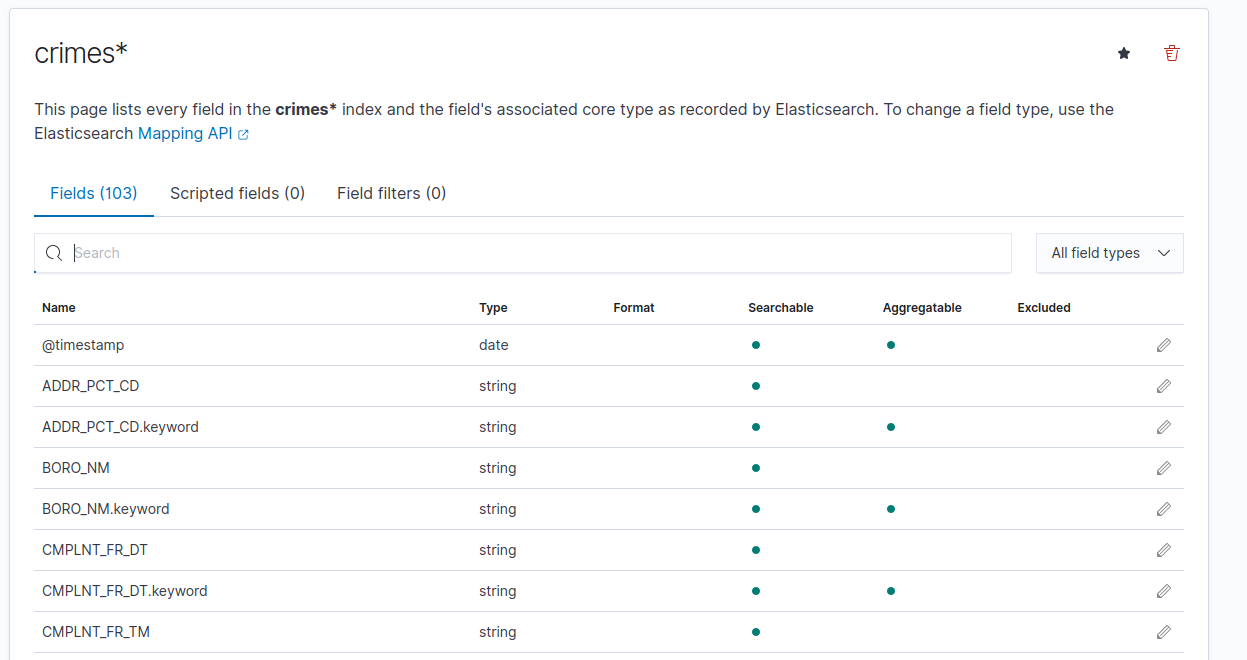
1. **Creación de índices de patrones en Kibana**

Se accede a través de la interfaz de kibana, ubicada en “Index paterns”. Para nuestro caso, juntamos el índice creado en Logstash con el índice creado y cargado por la interfaz web en kibana.





Se corrobora que la estructura sea la definida en el archivo de configuración de Logstash y la generada por la interfaz web

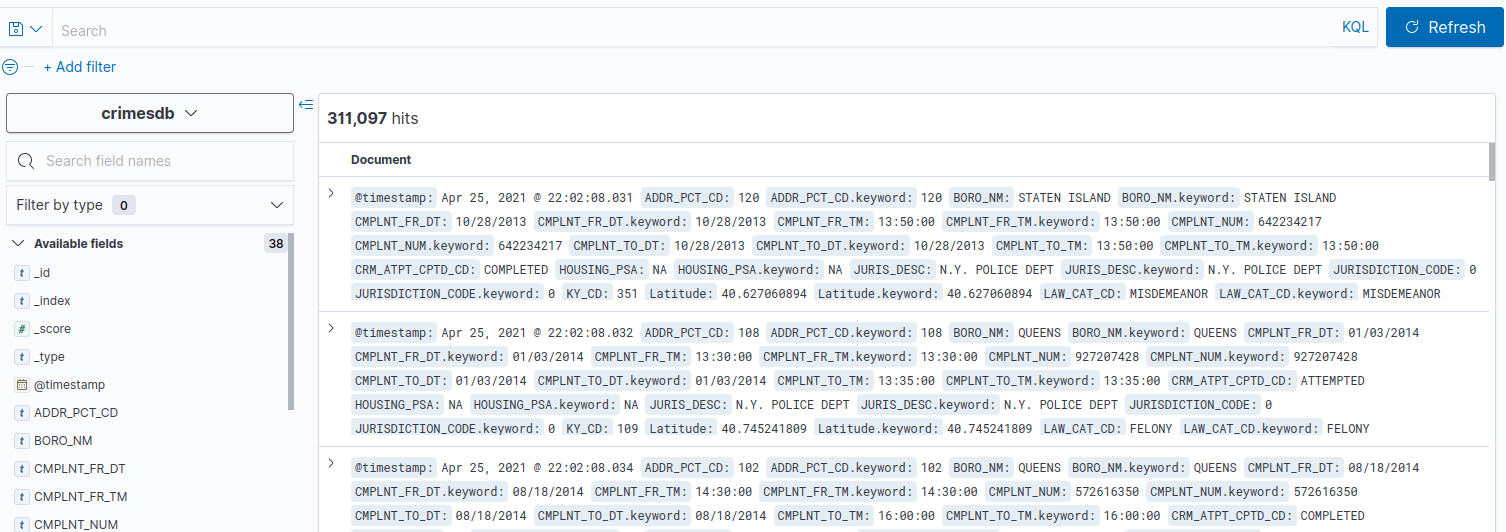


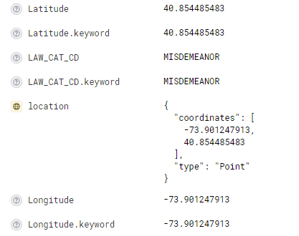
Lo más importante, el campo “Location” aparece como geo-point.



1. **Ejecución de consultas de Discover**

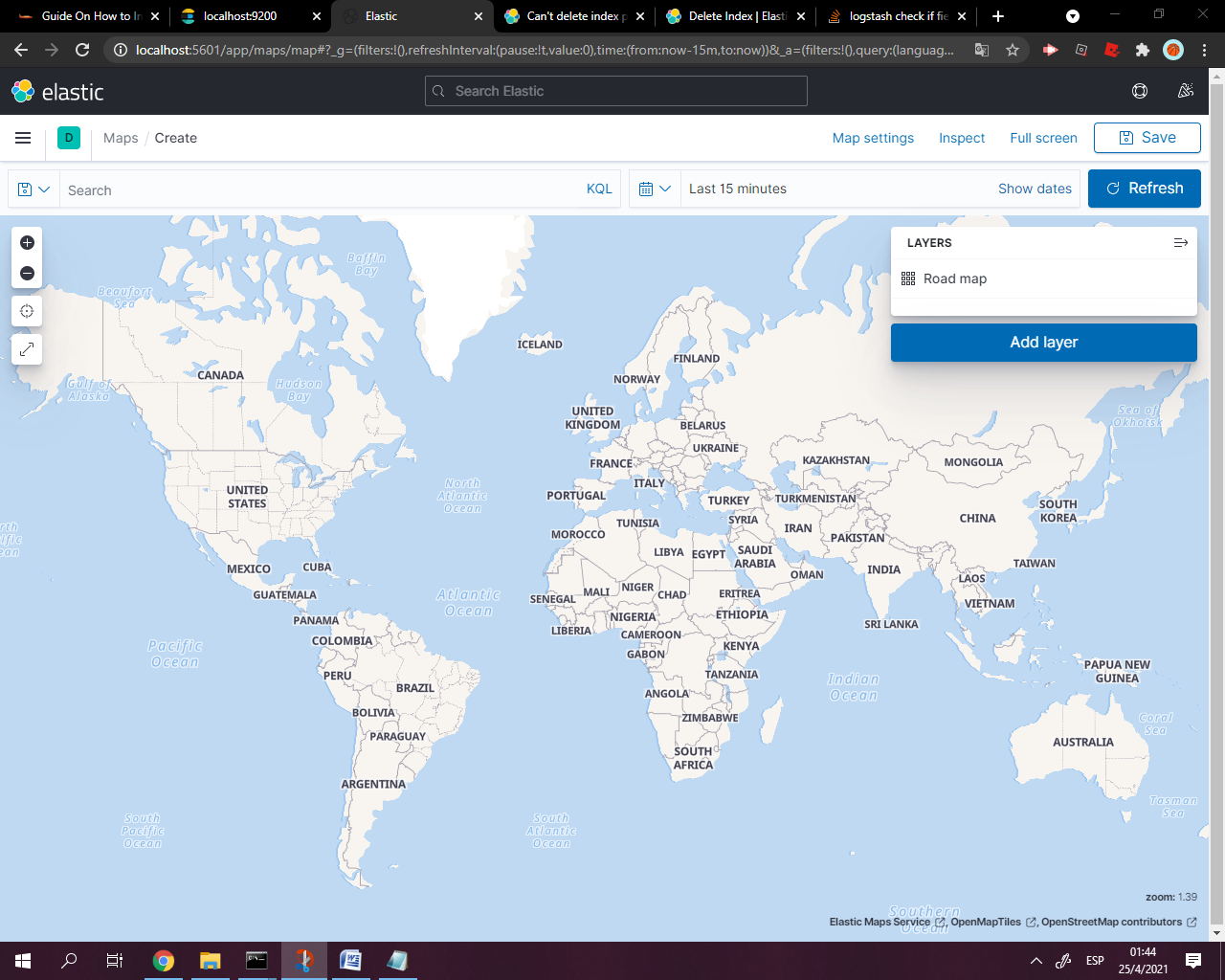
Se realiza una consulta de Discover sobre el dataset generado para la exploración de los datos cargados a través de Logstash. Se analiza la creación del campo “location” con sus coordenadas siempre que esté informada la latitud.



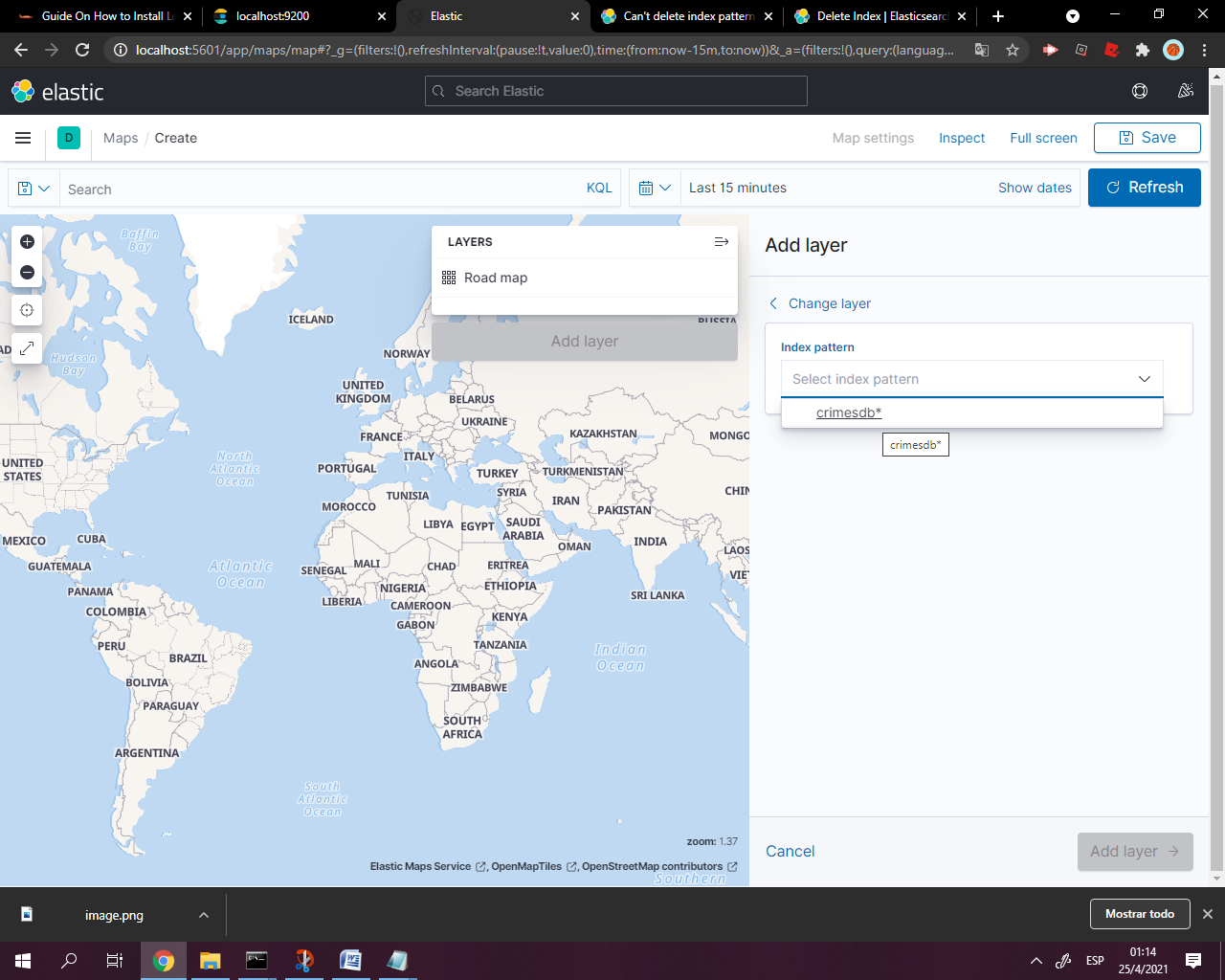


**7- Creación de mapas en Kibana**

Se ingresa a través de la página principal de Kibana a MAPS, se agrega una capa con la opción “Add Layer”

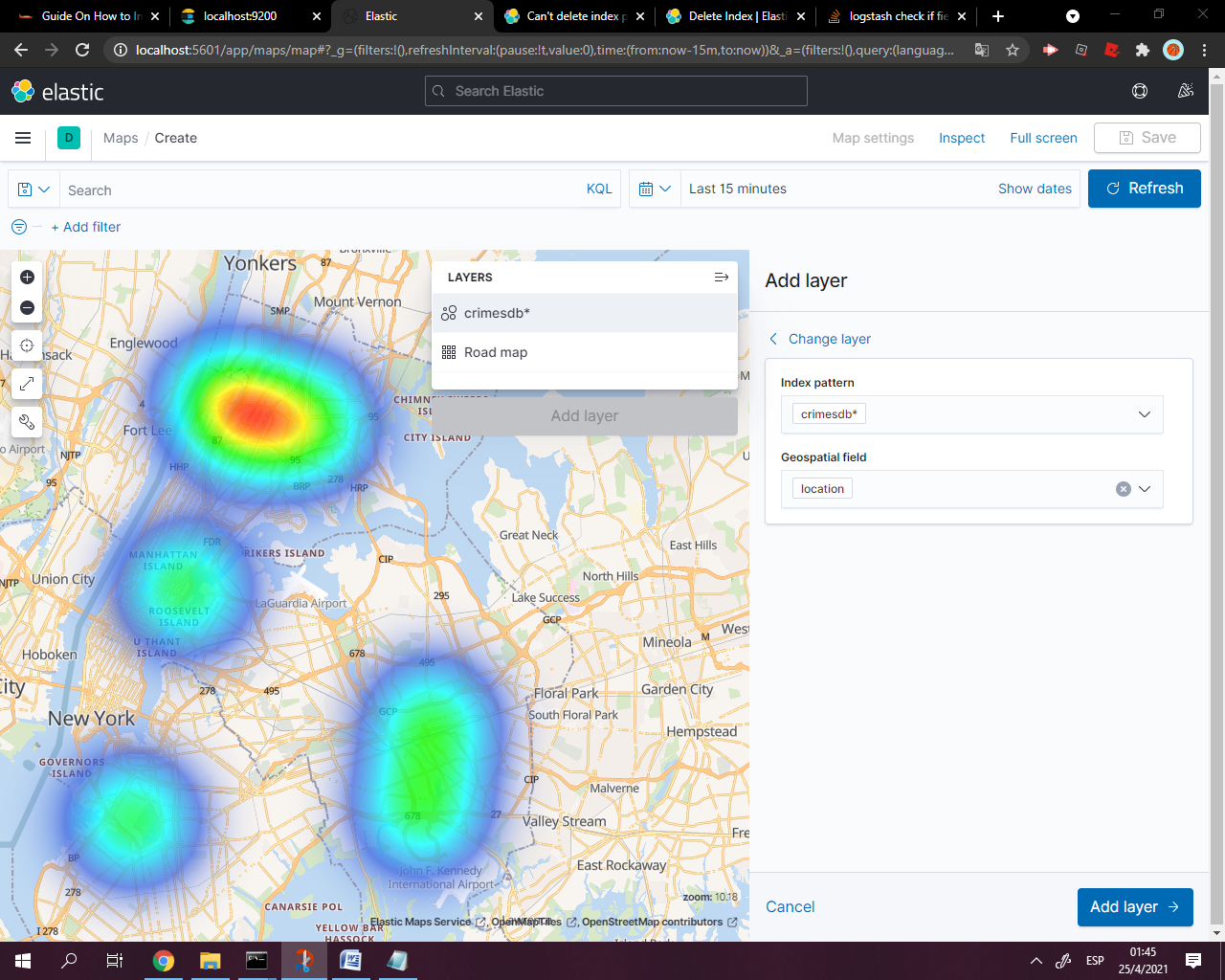


Se selecciona la opción “Heat Map” y luego se elige el índice recién creado



**7.1- Visualización del mapa**

Se agrega la capa creada y visualiza los indicadores elegidos. En nuestro caso



1. **Conclusión**

Se trabajó con un archivo CSV con datos de latitud y longitud en diferentes campos y se aplicó lógica con Logstash (herramienta que realiza el parseo y manipulación de datos) permitiéndonos de una manera práctica la carga, transformación e ingesta en un motor de base datos abstrayéndose del mismo, para nuestro caso el motor elegido es Elasticsearch a través de una configuración simple y accesible. Es importante aclarar que si se desea cambiar el motor solo debe modificarse el output en el archivo de configuración de Logstash (en nuestro caso comprobamos esto importando los datos de manera satisfactoria a MongoDB y visualizando los mismos con MongoDB Compass) .

Elasticsearch es un motor que demuestra su capacidad de almacenamiento, facilidad para alojar información, indexación y búsquedas rápidas de todo tipo de datos, incluyendo los geoespaciales, con buen tiempo de respuesta cuando se trata de grandes volúmenes de datos.

Por último, los mapas generados por Kibana, herramienta analítica, permitieron realizar visualizaciones de los datos almacenados en el motor de Elasticsearch que permiten identificar de una manera simple e intuitiva donde se ubican los mayores indicadores analizados.