

Base de Datos I

Trabajo Práctico Obligatorio

**Parte A: Trabajo de Investigación - Elasticsearch**

**GRUPO 1**

Integrantes

Canclini, Alexis Angel LU:133215

Gerszonowicz, Matias LU:133341

Morcecian, Aram Christian LU:1072240

Docentes:

Cancela, Julio Eduardo

[Parma, Marcelo Alejandro](javascript:;)

**Introducción**



Elasticsearch es un motor de búsqueda de texto completo y de estadísticas de código abierto. Permite analizar, almacenar y buscar grandes volúmenes de datos de forma rápida y en casi tiempo real.

Este motor procesa un texto identificando las palabras claves, crea un índice por cada una de ellas y los almacena. Esto permite que la búsqueda sea mucho más rápida y efectiva.

**Características**

Elasticsearch es una plataforma de búsqueda en casi tiempo real, esto quiere decir que existe un tiempo de latencia entre la carga del documento, la indexación del mismo y la disponibilidad del mismo para su búsqueda.

**Componentes**

**Clúster**

Un clúster es un conjunto de nodos. Toda la información se encuentra almacenada en los nodos. El clúster debe ser capaz de buscar en todos los nodos para encontrar la información solicitada.

**Nodo**

Un nodo es un servidor, que es parte de un clúster, donde se almacenan los datos.

**Índice**

Un índice es una colección de documentos que tienen características similares.

**Tipo**

Dentro de un índice, se pueden seleccionar uno o más tipos. Un tipo es una categoría o partición lógica de su índice cuya semántica es configurable por el usuario. En general, un tipo se define para documentos con un conjunto de campos comunes. Por ejemplo, vamos a suponer que se quieren almacenar todos los datos de una plataforma de blogs en un solo índice. En este índice, es posible definir un tipo para datos de usuario, otro tipo para datos del blog, y otro tipo para datos de comentarios.

**Documento**  
Un documento es una unidad básica de información que puede ser indexado. Por ejemplo, se puede tener un documento para un solo cliente, otro para un solo producto, y otro para un solo pedido. En este documento se expresa en JSON (JavaScript Object notación), que es un formato de intercambio de datos de Internet.  
  
Dentro de un índice/tipo, se pueden almacenar tantos documentos como desee. Tenga en cuenta que si bien un documento reside físicamente en un índice, un documento en realidad debe ser indexado o asignado a un tipo en el interior de un índice.

**Fragmentos y Réplicas**  
  
Un índice potencialmente puede almacenar una gran cantidad de datos, lo cual puede exceder los límites de hardware de un solo nodo. Por ejemplo, un único índice de mil millones de documentos de 1 TB de espacio, puede no caber en el disco de un solo nodo o puede ser demasiado lento para servir a las solicitudes de búsqueda de un nodo solo.  
  
Para resolver este problema, Elasticsearch ofrece la posibilidad de subdividir el índice en varios trozos llamados fragmentos. Cuando se crea un índice, sólo tiene que definir el número de fragmentos que desee. Cada fragmento es en sí mismo un índice totalmente funcional e independiente.

Este método es importante por dos razones principales:  
  
- Le permite dividir horizontalmente y reducir el volumen de contenido  
- Permite distribuir y paralelizar las operaciones a través de fragmentos (potencialmente en varios nodos) aumentando así el rendimiento  
  
Los algoritmos de distribución de un fragmento y cómo los documentos se agregan de nuevo en las solicitudes de búsqueda está completamente administrado por Elasticsearch y es transparente.  
  
En un entorno de red donde puede haber fallas, es muy útil y muy recomendable contar con un mecanismo de conmutación por error en caso de que un nodo se desconecte. Con este fin, Elasticsearch le permite hacer una o más copias de fragmentos de su índice en lo que se denominan fragmentos de réplicas.  
  
La replicación es importante por dos razones principales:  
  
    Proporciona alta disponibilidad en caso de que falle un fragmento / nodo. Por esta razón, es importante tener en cuenta que un fragmento de réplica no se asigna en el mismo nodo que el fragmento original donde ha sido copiado.  
    Se le permite escalar el volumen de búsquedas, puesto que las búsquedas se pueden ejecutar en todas las réplicas en paralelo.  
  
En resumen, cada índice se puede dividir en múltiples fragmentos. Un índice también se puede replicar cero (es decir, sin réplicas) o más veces. Una vez replicado, cada índice tendrá fragmentos primarios (los fragmentos originales desde donde fueron replicadas) y fragmentos de réplica (las copias de los fragmentos primarios). El número de fragmentos y réplicas pueden ser definidos por el índice en el momento en el que se crea.   
Una vez creado el índice, es posible cambiar el número de réplicas dinámicamente en cualquier momento, pero no se puede cambiar los fragmentos.   
  
De forma predeterminada, cada índice en Elasticsearch asigna 5 fragmentos primarios y 1 réplica, lo que significa que si usted tiene por lo menos dos nodos en el clúster, el índice tendrá 5 fragmentos primarios y otros 5 fragmentos de réplica (1 réplica completa) para un total de 10 fragmentos por índice.

**Casos de uso más populares**

**Monitorización de aplicaciones en tiempo real**

Se recopilan logs de aplicaciones y sitios web a través de un módulo llamado Logstash en el cual se indexan los datos . Luego con el complemento Kibana se visualizan los datos de los cuales se pueden extraer rápidos análisis con una simple vista. Un ejemplo claro son los logueos en diferentes aplicaciones de forma geolocalizada.

**Análisis de Streaming**

A través de la función de percolación de Elasticsearch, se permite observar patrones predefinidos sencillos, como un conjunto de direcciones IP, y activar una alerta instantánea.

**Análisis de sentimientos en las redes sociales**

Es posible monitorizar los sentimientos y las tendencias más recientes que los personas expresan en las redes sociales en relación con una determinada marca, producto o evento social.