

INTRODUCCIÓN A ELASTICSEARCH



ELASTICSEARCH

Es un motor de búsqueda y análisis distribuido, capaz de resolver gran cantidad de casos de uso y el lugar donde se almacenan los datos para poder explotarlos posteriormente. Es el corazón de la pila ELK.





1.

INTRODUCCIÓN

¿CÓMO SE USARÁ ESTA **BASE DE DATOS**?

Es una colección de información organizada para un posterior aprovechamiento de la misma.

Esta información debe:

- Ser accesible
- Se pueda gestionar
- Se pueda actualizar

CARACTERÍSTICAS DE ELASTICSEARCH

- **Distribuido y escalable.** Permite crecer conforme lo hagan las necesidades **de forma horizontal**.
- **Datos en tiempo real.** Los datos están disponibles para su análisis segundos después de haber sido indexados (*real time*).
- **Alta disponibilidad.** Datos replicados a lo largo de distintos nodos permite el fallo de alguno de estos dentro del clúster sin que se vea afectado el funcionamiento.

CARACTERÍSTICAS DE ELASTICSEARCH

- **Multi-tenancy.** Los índices donde se almacenan la información pueden ser consultados de manera independiente.
- **Búsquedas full-text.** Usa Lucene aprovechando sus capacidades de búsqueda de texto, soportando geolocalización, autocompletado, expresiones regulares...
- **API Restful.** Proporciona una API sobre JSON para realizar consultas e interactuar con Elasticsearch.



2.

ELASTICSEARCH

ALMACENAMIENTO DE LA INFORMACIÓN

- **Índice**: Colección de documentos con características similares (catálogo de productos, clientes, pedidos...).

Identificado por un nombre (minúsculas) usado para realizar operaciones como *indexado*, *búsqueda*, *actualización* y *borrado* de los documentos que contiene.

ALMACENAMIENTO DE LA INFORMACIÓN

- **Documento:** Unidad básica de información que puede ser indexada (documento por producto, cliente o pedido) en formato JSON.

https://www.w3schools.com/js/js_json_intro.asp

En un índice se pueden almacenar todos los documentos que se requiera.

ALMACENAMIENTO DE LA INFORMACIÓN

- **Shards**: Cuando se almacena gran cantidad de información puede exceder los límites hardware de un nodo. Elasticsearch proporciona la posibilidad de subdividir un índice en distintas partes denominadas *shards*.
 - Permite subdividir/escalar la información almacenada.
 - Permite distribuir y paralelizar operaciones.

ALMACENAMIENTO DE LA INFORMACIÓN

- **Réplica:** Elasticsearch permite realizar una o más copias de los shards. Su importancia recae en:
 - Proporciona alta disponibilidad en el caso de que un shard/nodo falle. Importante que las réplicas no se encuentren en el mismo nodo.
 - Permite el escalado del volumen de búsquedas ya que estas pueden ejecutarse en paralelo en las réplicas.

ALMACENAMIENTO DE LA INFORMACIÓN

Una vez replicado, cada índice tendrá un **shard primario** y las **réplicas** (copias exactas del primario).

Los *shards* y *réplicas* se pueden definir por índice a la hora de creación de los mismos. Las réplicas podrán cambiarse en caliente al contrario que los shards.

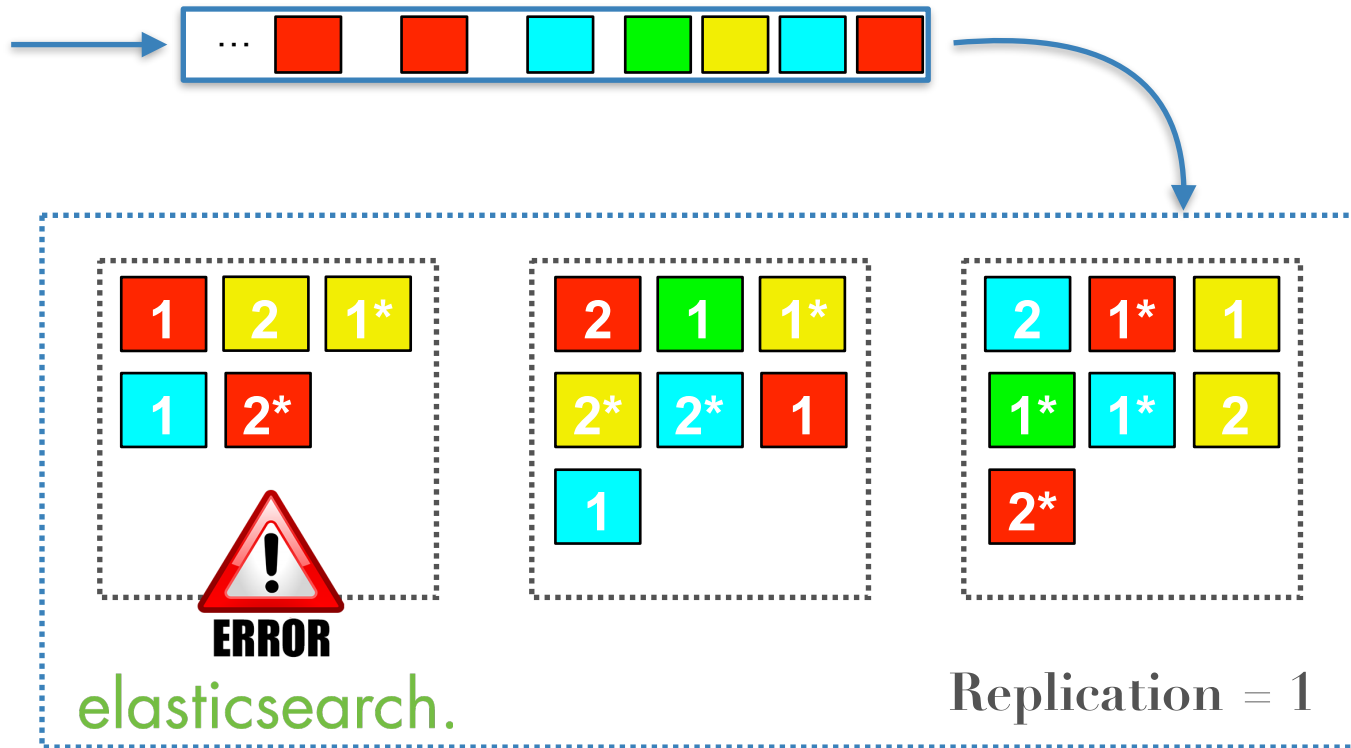


Existe un máximo de documentos que pueden estar en un único shard. Aproximadamente 2.147.000.000 documentos.

BASES DE DATOS DISTRIBUIDAS

- La información se almacena en diferentes sitios de la red (físicamente) pero lógicamente es una única base de datos.
- Transparente para un usuario final.
- Independencia respecto al sistema operativo.
- Procesamiento distribuido de consultas.
- Información fragmentada.
- Réplicas = Alta disponibilidad.

DISTRIBUCIÓN DE DATOS





3.

TIPOS DE NODOS

CONJUNTO DE NODOS -> CLÚSTER

- Instancia de Elasticsearch = **nodo**.
- Conjunto de nodos conectados forman un **clúster**.

Cada nodo se encarga de peticiones HTTP realizadas por los clientes REST API y también mantienen la comunicación en la capa de transporte entre los nodos del clúster.

Nodos conocen el estado del resto y pueden reenviar las peticiones de los clientes al nodo apropiado.

TIPOS DE NODOS: **MASTER NODE**

Responsable de acciones como:

- Crear y borrar índices.
- Controlar qué nodos forman partes del clúster.
- Decidir en qué nodo alojar cada *shard*.

Indexado y búsqueda de datos: CPU, RAM e I/O ↑↑↑

Buena práctica: ¡**Máster dedicado!**

TIPOS DE NODOS: DATA NODE

Almacenan los *shards* que contienen los documentos que se han indexado.

- Operaciones: CRUD, búsquedas y agregaciones.
- Importante monitorizar el uso de CPU, RAM e I/O.

- **Añadir más Data Nodes si se sobrecargan** para el balanceo de los procesamientos y reparto de datos.

TIPOS DE NODOS: **INGEST NODE**

Pueden ejecutar pre-procesamiento de la información entrante antes de su indexado.

- Si el procesamiento previo consume muchos recursos

Buena práctica: **¡Ingest dedicados!**

TIPOS DE NODOS: COORDINATING NODE

Pueden encaminar las peticiones, encargarse de las respuestas de las búsquedas y distribuir los indexados.



Balanceadores de carga inteligentes

- Ayudan a descargar la tarea de “*coordinación*” perteneciente a los Master Nodes.
- Usan el estado del clúster para encaminar búsquedas y peticiones al lugar apropiado.

CLÚSTER: ARQUITECTURA

