# What is Elasticsearch?

## Data in: documents and Indices

Como se sabe Elasticsearch es una manera de almacenar documentos distribuidos el cual los guarda de manera estructurada de datos mas complejos como lo son los JSON, se crea un clúster cuando son varios nodos de Elasticsearch y se puede acceder a cualquiera de los documentos por cualquier nodo.

Cuando se guarda un documento este se indexa (adquiere un índice), Elasticsearch cuenta con un mecanismo de índices invertidos lo que permite la búsqueda de textos más complejos y de manera más rápida. La manera de ver los índices seria como una colección optimizada de documentos con una colección de campos (los datos). Además, que cuenta con la capacidad de no tener esquemas lo cual puede lograr que los documentos se pueden indexar sin especificar explícitamente y la utilización del mapeo dinámico.

A la hora de definir las propias asignaciones se pueden realizar las siguientes operaciones:

- *Distinguir entre campos de cadena de texto completo y campos de cadena de valor exacto.*
- *Realizar análisis de texto específico del idioma.*
- *Optimizar campos para coincidencias parciales.*
- *Usar formatos de fecha personalizados.*
- *Utilice tipos de datos como geo_point y geo_shape que no se puedan detectar automáticamente.*

## Information out: search and analyze

En lo que más residen las capacidades de Elasticsearch es en la posibilidad de realizar búsquedas integradas en la biblioteca del motor de búsqueda de Apache Lucene. Además, cuenta con API REST la cual es una forma simple y coherente para administrar clúster e indexar y buscar datos. Se pueden realizar consultas estructuradas, de texto completo o una combinación compleja de ambas; lo que ayuda a realizar consultas de términos individuales, frases, similitudes o prefijos. Asimismo, como Elasticsearch indexa los datos no textuales en estructuras de datos optimizadas, dando consultas de alto rendimiento y eficiencia.

Elasticsearch permite crear resúmenes complejos de los datos y obtener mejores resultado para métricas, patrones y tendencias, esto significa que de una simple frase podríamos obtener diferentes resultados, pero relacionadas a la frase inicial, dando mejores a la hora de realizar búsquedas mas avanzadas o mejor dicho adelantadas a la que el usuario amerite en el momento. Esta estructura de búsqueda es muy rápida y permite analizar y visualizar datos en tiempo real.

Agregando que se pueden realizar solicitudes de búsqueda con agregaciones y consigo la búsqueda de documentos, filtrar resultado y realizar análisis al mismo tiempo con esos datos o en una sola solicitud. Esta también cuenta con un mecanismo de aprendizaje automático para la creación de líneas de bases precisas a un comportamiento normal o bien, a la identificación de patrones anómalos que no sean del interés del programa en ejecución.

## *Scalability and resilience*

Elasticsearch está diseñado para estar siempre disponible y escalar a las necesidades de quién lo necesite, por lo que se pueden agregar varios servidores o en su defecto, nodos, a un clúster para aumentar la capacidad, lo que ejecuta automáticamente la distribución de carga de datos y consultas a todos los nodos que se encuentran disponibles. Por debajo, Elasticsearch es una agrupación lógica de uno o varios fragmentos físicos, que cada uno de estos es un índice autónomo. Esto ocasiona a su vez la redundancia de datos, por lo que se protege contra posibles fallas de hardware y claramente aumenta la capacidad de consultas.

Estos fragmentos se dividen en ***primaries and replicas***, cada documento le pertenece a un primario y las replicas son copias para proteger de fallas de hardware.

En cuando a gastos, entre más fragmentos más gastos generales y será mayor el tiempo en el que Elasticsearch le tome mover los fragmentos cuando necesite reequilibrar el clúster. Ahora bien, si hay menor cantidad de fragmentos quiere decir que el procesamiento por fragmento es más rápido, sin embargo, la cantidad de gastos generales es mayor. Para mejor eficiencia se recomienda trabajar en estos rangos: 20 GB a 40 GB para uso con datos basados en tiempo.

Cross-cluster replication (CCR): es una forma de sincronizar automáticamente los índice del clúster principal con un clúster remoto que sea secundario el cual puede servir como una copia de seguridad, entonces, si el principal falla el secundario tomaría el mando. El principal es el de índice activo que maneja todas las solicitudes de escritura, mientras que los índices replicados en clústeres secundarios son solo de lectura.