Base de Datos II (IC4302) – Semestre 1, 2023

Resumen #1 – What is Elasticsearch?

David Suárez Acosta – 2020038304

Data in: documents and índices

Elasticsearch es un almacén de documentos distribuidos en formato JSON. Cuando se tienen nodos de Elasticsearch en un clúster, se puede acceder a un documento en menos de un segundo ya que Elasticsearch usa una estructura de datos llamada “inverted index” que permite búsquedas rápidas de texto.

Un índice es una colección optimizada de documentos donde cada documento es una colección de “fields” que funcionan como llaves para la información almacenada, cada información almacenada tiene su propia estructura de datos optimizada.

Elasticsearch no necesita esquemas ya que los documentos se pueden indexar sin necesidad de especificar como se va a tratar cada campo del documento. Cuando se habilita la opción “dynamic mapping” Elasticsearch automáticamente añade nuevos campos al index.

Se pueden definir reglas para manejar el mapeo y tener todo el control sobre como los campos se almacenan e indexan.

Information out: search and analyze

El REST API de Elasticsearch hace posible la creación de “structured queries” (son similares a las queries de SQL), “full text queries” (se buscan todos los documentos que contienen el texto de la consulta realizada) y “complex queries” que combina las dos.

Las aggregations de Elasticsearch permiten la construcción de resúmenes muy completos que permiten obtener más información con relación a los datos y a los parámetros de llaves, patrones y trends.

Como las aggregations usan las mismas estructuras de datos que las búsquedas, estas son bastante rápidas, permitiendo un análisis a tiempo real. Estas aggregations también pueden funcionar simultáneamente junto con solicitudes de búsqueda, haciendo búsquedas filtradas.

Scalability and resilience

Elasticsearch está hecho para estar siempre disponible y para adaptarse a las necesidades del usuario. Se pueden agregar servers o nodos a un clúster para aumentar la capacidad, Elasticsearch automáticamente distribuye la información y las consultas por los nodos disponibles de forma balanceada.

Un índice de Elasticsearch es un grupo lógico de uno o más partes físicas, donde cada fragmento es un índice autónomo. Al distribuir los documentos en un índice por fragmentos y distribuyendo esos fragmentos por los diferentes nodos, Elasticsearch aplica redundancia. Conforme crece un clúster, Elasticsearch migra automáticamente los fragmentos para balancear el clúster.

Existen dos tipos de fragmentos, primarios y replicas. Cada documento en un índice pertenece a un fragmento primario mientras que un fragmento réplica es una copia de uno primario, esto se hace para

aplicar la redundancia. El número de fragmentos primarios en un índice es permanente desde que este se crea, pero el número de réplicas sí puede cambiar.

Entre más fragmentos hay o si aumenta mucho sus tamaños, puede afectar el rendimiento a la hora de balancear el clúster. Los nodos de un clúster necesitan estar en buen estado y tener conexiones confiables entre sí, esto se logra colocando los nodos en data centers cercanos, pero para mantener la High Avaliability, se hacen replicaciones del clúster.