Instituto Tecnológico de Costa Rica

Ingeniería en Computación.

Bases de Datos II.

1er semestre

Resumen #1

Resumen ElasticSearch

ESTUDIANTE:

Edgar André Araya Vargas

2020142856

PROFESOR:

Gerardo Nereo Campos Araya

Grupo: 1

Fecha de entrega:

Viernes 24 de febrero del 2023.

¿Qué es ElasticSearch?

Data in: documentos y índices.

Elasticsearch es un popular motor de análisis y búsqueda de código abierto que le permite buscar, analizar y almacenar grandes cantidades de datos rápidamente y en tiempo real, esto mediante un motor de búsqueda de texto completo distribuido y con capacidad para múltiples inquilinos con una interfaz web **HTTP** y documentos **JSON** sin esquema. Elasticsearch indexa los datos dividiéndolos en términos individuales, creando un índice invertido y luego almacenando el índice de manera distribuida en varios **nodos** en un **clúster**, así logrando la recuperación de datos en *tiempo real*.

Como se mencionó anteriormente, Elasticsearch posee la habilidad de trabajar **sin esquema** ya que puede mapear campos automáticamente en función de los datos que recibe. Permite una **indexación flexible** de los datos con un esquema variable aunque no esté explícitamente especificado como tratar cada diferente campo dentro del documento. Esto es llamado *"Dynamic Mapping"* y podemos utilizarlo para definir mappings personales que tomen control de los campos a ser guardados y indexados.

Information out: buscar y analizar.

Además de permitirnos recuperar documentos y su metadata Elasticsearch también nos provee un **REST API** para manejar nuestro cluster y el indexing.

Buscar tu data.

Los **REST API's** soportan consultas estructuradas, consulta de texto completo y consultas complejas.

- **Estructuradas:** son aquellas utilizadas en SQL.
- **Texto Completo**: se encargan de devolver resultados por *relevancia*, que tan bueno es el match para los términos de búsqueda.
- **Complejas:** combinación entre estructuradas y de texto completo.

Analizar tu data.

Elasticsearch cuenta con **agregaciones** que permiten analizar la informacion con mayor eficiencia y eficacia de lo que permitirian otras herramientas. Estas agregaciones funcionan agrupando datos según ciertos criterios, calculando métricas o estadísticas en esos grupos y permitiendo anidar para crear análisis más complejos y ya que funciona con la misma estructura de dato que la busqueda funcionan de manera muy veloz.

Ya que las agregaciones funcionan de manera paralela a las **consultas de búsqueda** podemos buscar los documentos, filtrar resultados e incluso realizar análisis al mismo tiempo en la misma data. Hasta podemos incorporar *features* de *machine learning*.

Escalabilidad y resiliencia: clusters, nodes y shards

Elasticsearch está diseñado para ser **altamente escalable y resistente** mediante el uso de una arquitectura distribuida. Elasticsearch logra escalabilidad y resiliencia mediante el uso de **clusters, nodes y shards**.

Se pueden añadir **nodos** a un **cluster** para aumentar la capacidad, de esta manera Elastisearch puede distrubuir automaticamente la data y el peso de las consultas hacia todos los nodos disponibles, es decir: *Elasticsearch sabe balancear clusters de múltiples nodos.*

Esto mediante una técnica llamada **sharding** para distribuir datos a través de **múltiples nodos** en un **cluster**. La **fragmentación** permite que Elasticsearch se *escale horizontalmente* al agregar más nodos a un clúster. Cada índice en Elasticsearch se divide en múltiples fragmentos, y cada fragmento se almacena en un nodo diferente en el clúster. Los shards pueden ser primarios o replicas redundantes de la data para protegerle en caso de fallo de hardware.

Hay que tener cierto cuidado con los **shards** es importante recordar y tomar en cuenta que entre mayor cantidad de shards mas supervisión existe en el índex, y entre más largo el tamaño del shard, más tiempo tomara para moverse al rebalanceo de un cluster.

Por último, Elasticsearch Cross-Cluster Replication es un *feature* que permite que los datos se repliquen en tiempo real entre múltiples clusters de Elasticsearch para recuperación ante desastres, migración de datos y rendimiento y disponibilidad mejorados. CCR admite opciones de filtrado, resolución automática de conflictos y herramientas de monitoreo para garantizar la confiabilidad e integridad del proceso de replicación.