Adriel Araya Vargas 2019312845 Instituto Tecnológico de Costa Rica Bases de Datos 2 IC-4302

¿Qué es Elasticsearch?

Elasticsearch es el motor de búsqueda y analítica distribuido, parte esencial de Elastic Stack. Elastick Stack está conformado por Elasticsearch, Kibana, Beats, y Logstash. Elasticsearch provee búsquedas cerca de tiempo real y analíticas para todo tipo de datos. Este puede guardar e indexar de una forma eficiente que soporta búsquedas rápidas.

Datos dentro: documentos e índices.

Elasticsearch es un almacenamiento de documentos distribuidos. En vez de usar un formato de tablas, este guarda estructuras de datos complejas que han sido serializadas como documentos JSON. Cuando se corren múltiples nodos de Elasticsearch en un cluster los documentos de distribuyen a través del mismo y pueden ser accedido inmediatamente desde cualquier nodo. Cuando un documento esta guardado, es indexado y buscable en casi tiempo real.

El índice se puede pensar como colección optimizada de documentos y los documentos son una colección de campos. Elasticsearch indexa todos los datos de los campos y cada dato indexado tiene una estructura de datos optimizada y dedicada. Este además posee la habilidad de indexar sin tener la especificación de como manejar cada uno de los campos. Esto se le llama mapeo dinámico. Estos mapeos pueden ser incluso manipulados para añadir reglas y tener control total de como los campos son manejados. Es útil indexar el mismo campo de diferentes formas dependiendo del próposito.

Información fuera: búsqueda y análisis.

Elastic Search provee con un REST API simple para manejar el cluster e indexar y buscar la información. Con su respectivo cliente Elasticsearch se puede usar en el lenguaje de elección.

Buscando los datos

Este REST API soporta consultas estructuradas, consultas full-text y consultas complejas que combinan ambas. Las estructuradas son similares a las que se construyen en SQL. Las full-text encuentran todos los documentos que parean con el string de consulta y los devuelve en orden de relevancia(que tan certero es el pareo). Además se puede buscar por términos individuales, por frases, por prefijos y obtener recomendaciones autocompletadas. También se puede buscar información no textual en estructuras de datos optimizadas para datos numéricos y geoespaciales. Todas estas capacidades pueden ser accedidas desde el lenguaje de consultas de Elasticsearch: QueryDSL.

Analizando nuestros datos.

Las agregaciones de Elasticsearch permiten crear resúmenes complejos de la información y obtener conocimiento de las métricas de llave, patrones y tendencias. Como estas agregaciones permiten aprovechar las mismas estructuras de datos usados para la búsqueda, estas también son muy rápidas. Esto permite analizar y visualizar los datos en tiempo real. La información va a ser actualizada mientras cambian los datos. Las

agregaciones operan al mismo tiempo que las consultas de búsqueda, por lo cual se puede buscar documentos. filtrar resultados y hacer analíticas en los mismos datos, en una sola consulta.

Resilencia y escalabilidad: clusters, nodos y shards.

Elasticsearch esta construido para siempre estar disponible y escalar a base de las necesidades del usuario. Se pueden añadir servidores a un cluster para incrementar la capacidad y Elasticsearch automáticamente va a distribuir los datos y la consulta va a cargar a través de todos los nodos disponibles. Esto funciona debido a que el índice de Elasticsearch es una agrupación lógica de uno o más fragmentos físicos, donde cada fragmento es en realidad un índice auto-contenido. Al distribuir estos documentos en un índice a tráves de varios fragmentos y luego distribuir estos fragmentos en los múltiples nodos, Elasticsearch puede asegurarse de la redundancia que lo protege ante fallas de hardware e incrementa la capacidad de consultas así como se añaden los nodos.

Hay dos tipos de fragmentos: primarios y réplicas. Cada documento en un índice pertenece a un fragmento principal. Una réplica es la copia del fragmento principal. Las réplicas proveen copias redundantes de los datos para proteger de fallos e incrementar la capacidad para leer consultas.

Hay varias consideraciones de rendimiento con respecto al tamaño de los fragmentos y el número de fragmentos primarios configurados en el índice.

Algunas recomendaciones para empezar: -Intenta mantener el promedio de fragmentos entré unos cuantos GB y unas decenas GB. -Evite el problema de muchísimos fragmentos. El número de fragmentos que un nodo puede soportar es proporcional a la disponibilidad de espacio de almacenamiento dinámico.

En caso de desastre.

Los nodos de un cluster necesitan una buena y confiable conexión el uno al otro. Para habilitar buenas conexiones se necesitan localizar los nodos en data centers cercanos o en el mismo. Sin embargo en caso de falló por una avería de zona lo nodos deben ser relocalizados. Para esto se usa Replicación a tráves de clusters o Cross-Clusters replication (CCR). CCR provee una forma de sincronizar automáticamente los índices del cluster primario para que el segundo cluster secundario pueda servir como un respaldo. El CCR es activo-pasivo. El índice en el cluster primario es el líder activo que maneja todas las consultas. Los índices replicados a los clusters son los pasivos que solo leen.

Referencia:
https://www.elastic.co/guide/en/elasticsearch/reference/current/elasticsearch-intro.html