Bases II GR 1 - Resumen 1

Gerald Núñez Chavarría - 2021023226

1. ¿Qué es Elasticsearch?

Elasticsearch es un motor de búsqueda distribuido que ha sido diseñado para manejar grandes cantidades de datos casi en tiempo real, esta basado en la biblioteca Lucene y es una base de datos orientada a documentos (almacena datos en documentos JSON).

Es altamente escalable y brinda funciones de búsqueda y análisis, como la búsqueda de texto completo, la agregación y la búsqueda geoespacial. Presenta una gran variedad de APIs (Java, REST y JavaScript). Se puede utilizar para casos de búsqueda, análisis, registro y monitoreo.

1.1 Datos en: documentos e índices.

En lugar de almacenar la información como filas en columnas, Elasticsearch la almacena en estructuras de datos complejas que se han serializado como documentos JSON. Al agregar un documento, este se indexa y ya se puede buscar en 1 segundo aproximadamente. Un índice es una colección optimizada de documentos y cada documento es una colección de campos, luego indexa todos los datos en cada campo y cada campo indexado tiene una estructura de datos optimizada y dedicada.

Tiene la capacidad de no tener esquemas (no es necesario especificar como manejar cada uno de los campos que aparecen en un documento), mediante el mapeo dinámico, Elasticsearch permite indexar datos en los cuales no viene especificado el tipo de dato, pero este los detecta y le asignará valores booleanos, de punto flotante y enteros, fechas y cadenas a los tipos de datos de Elasticsearch apropiados.

El mapeo dinámico puede ser configurado para controlar como se almacenan e indexan los datos, esto le brinda beneficios cómo: Distinguir entre campos de cadena de texto completos y campos de cadena de valor exacto, análisis de texto especifico en un idioma, usar formatos de fechas personalizados...

1.2 Salida de información: Buscar y analizar.

Para buscar los datos en Elasticsearch, la API REST permite consultas estructuradas (similares a las de SQL), de texto complejo (buscan en todo un texto y devuelven coincidencias) o combinar ambas. También se admiten consultas de datos geográficos y numéricos de alto rendimiento. La mejor manera de acceder a estas habilidades de búsqueda es utilizando el lenguaje de consulta estilo JSON de Elasticsearch.

Se analizan los datos almacenados al punto de poder realizar resúmenes, obtener métricas, patrones y tendencias claves. Las agregaciones de análisis aprovechan las estructuras de datos de la busqueda, esto quiere decir que también se analizan los datos de manera muy rápida, casi en tiempo real. Puede buscar documentos, filtrar resultados y realizar análisis al mismo tiempo, en los mismos datos, en una sola solicitud.

Además, las herramientas de análisis utilizan funciones de aprendizaje automático, para crear métricas que identifican si el comportamiento de los datos almacenados es normal. Esto permite analizar anomalías con desviaciones temporales en valores, conteos o frecuencias, rarezas estadísticas o comportamientos inusuales para un miembro de la población.

1.3 Escalabilidad y resiliencia.

Elasticsearch está siempre disponible gracias a qué es distribuido por naturaleza. Se pueden agregar nodos a un clúster para aumentar la capacidad y Elasticsearch automáticamente distribuirá sus datos. Entre más nodos, mejor.

Un índice es la agrupación de uno o más fragmentos, y cada fragmento es un índice autónomo. Se distribuye los documentos en varios fragmentos y estos en varios nodos, generando redundancia, que protege contra fallo de software y aumenta la velocidad de consulta. Hay dos tipos de fragmentos, los primarios (cada documento en un índice) y los réplicas (copia de los principales).

Cuando un cluster crece o se reduce, Elasticsearch migra fragmentos para equilibrarse. No obstante, aunque "entre más nodos mejor" hay que ser cuidadosos con la cantidad de nodos primarios, si es muy grande genera costos en mantenimiento y de equilibrio. Lo recomendado es que la cantidad de fragmentos por GB de espacio sea menor a 20.

Los cluster son muy confiables dada la cercanía en los centros de datos y la confiable conexión que tienen los nodos entre sí. En el caso de falla en una localidad, los servidores de otra localidad deben estar preparados para asumir la responsabilidad, se llama replicación entre clústeres (CCR cross-cluster replication).

La replicación entre clústeres permite una manera automática de crear clústeres réplicas y conectarlas al índice del clúster primario, las réplica pueden sustituir al primario si este falla, haciendo un backup. Los clústeres secundarios son activo-pasivos, ya que solo hacen seguimiento de lectura del clúster primario, todas las solicitudes de escritura las maneja el clúster primario.

Como cualquier sistema empresarial, es recomendable tener herramientas para manejar, proteger y administrar su clúster.