Resumen #1: "What is Elasticsearch?"

Max Richard Lee Chung - 2019185076

Data in: documents and indices

Elasticsearch es un almacenamiento distribuido de documentos que, en vez de guardar los datos en fila, las almacena en estrcutruas complejas en JSON. Si existen varios nodos en Elastic en un clúster, los documentos son distribuidos a todos los nodos de forma inmediata. Los documentos son almacenados por medio de índices que pueden ser buscados en tiempo real (menos de 1 seg). Elastic utiliza un sistema de índices invertidos que soporta rápidas búsquedas de texto completo (full text). El sistema de índices invertidos facilita la búsqueda de palabras en todos los documentos para mostrarlos al usuario.

Un índice se podría considerar como una colección de datos optimizada, en donde cada documento es una colección de campos (contienen la llave y el valor de la información) De forma predeterminada, Elasticsearch indexa todos los datos en cada campo y cada campo indexado tiene una estructura de datos dedicada y optimizada. Además, tiene la opción de ser schema less (sin estructura), el cual permite almacenar los índices sin tener que definir cómo manejar los diferentes campos. Cuando el mapeo dinámico está habilitado, se detecta y agregan los nuevos campos de forma automática, el cual permite mayor facilidad para indexar y explorar la información.

Se pueden definir reglas para el control de mapeo dinámico, el cual permite:

- Distinguir entre un string de texto completo y el valor exacto de un string.
- Realizar un análisis específico entre idiomas.
- Optimizar campos para encontrar informaicón parcial.
- Usar formatos de fechas personalizadas.
- Usar tipos de datos que no pueden ser detectados.

A menudo es útil indexar el mismo campo en diferentes formas para otros usos. La cadena de análisis que se aplica a un campo de texto completo durante la indexación también se usa en el momento de la búsqueda. Cuando consulta un campo de texto completo, el texto de consulta se somete al mismo análisis antes de buscar los términos en el índice.

Information out: search and analyze

Elastic provee un REST API coherente para administrar el clúster y administrar y buscar la información. Soporta consultas estructuradas (consultas SQL), consultas de texto completo (búsquedas de documentos) y consultas complejas que combinan las dos formas. En términos individuales, se pueden realizar búsquedas de frases, búsquedas de similitudes, búsquedas de prefijos y sugerencias. Elasticsearch indexa datos no contextuales en estructuras de datos optimizados que soportan altos rendimientos de consultas numéricas y geográficas. Además, se pueden construir consultas de estilo SQL para buscar y agregar datos dentro de Elastic.

Las agregaciones de Elasticsearch permiten crear resúmenes complejos de sus datos y obtener información sobre métricas, patrones y tendencias clave. Son muy rápidas las búsquedas, debido a que las agregaciones aprovechan las mismas estructuras de datos utilizadas, el cual permite al usuario analizar y visualizar sus datos en tiempo real. Sus informes y paneles se actualizan a medida que cambian sus datos para que pueda

tomar medidas en función de la información más reciente. Las agregaciones operan junto con las solicitudes de búsqueda, el cual permite filtrar resultados y realizar análisis al mismo tiempo, en los mismos datos, en una sola solicitud entre todos los documentos.

Se pueden utilizar características de machine learning para crear líneas bases del comportamiento normal de los datos de forma precisa e identificar anomalías en la información.

Scalability and resilience: clusters, nodes, and shards

Elasticsearch está diseñado para estar siempre disponible y escalar dependiendo de las necesidades. Se pueden agregar servidores al cluster para aumentar la capacidad y Elasticsearch distribuye o balancear la información de forma automática entre todos los nodos. Esto se logra porque los índices son agrupamientos de uno o más fragmentos físicos (primarias, el cual guarda el índcie del documento o réplicas, el cual son copias de la primaria), donde cada un fragmento contiene un índice para conseguir redundancia. De esta manera, se puede proteger la información al tener un error de hardware e incrementa la capacidad de consultas. El número de índices primarios están prediseñados, mientras que las réplicas se pueden cambiar en cualquier momento sin interrumpir las operaciones.

Sin embargo, entre más fragmentos se creen, más carga general se aplicará al servidor a la hora de realizar mantenimientos de los índices. Creando consultas con fragmentos pequeños, aumenta la velocidad de procesamiento de cada uno pero requiere más carga general. Es recomendado tener un promedio de fragmentos en el servidor entre la capacidad de almacenamento y la cantidad es proporcional al espacio disponible.

Tener una buena conexión entre los nodos es muy común para tener un buen rendimiento, en donde los nodos se almacenan en un mismo lugar o en data centers cercanos. Sin embargo, si se presenta una falla de un establecimiento, los nodos vecinos tienen que poder soportar la carga del nodo no disponible, por lo que se presenta el uso del cross cluster replication (CCR). Este método provee métodos para sincronizar de forma automática los índices primarios del cluster a un cluster secundario remoto que sirve de respaldo y los índices almacenados tienen la única función de solo lectura.

Referencia

- Elasticsearch (2013). "What is Elasticsearch?". Recuperado de Elasticsearch