Elasticsearch

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

¿Qué es Elasticsearch?

Es el motor de búsqueda y análisis de ***Elastic Stack***. En Elasticsearch es donde se realiza la indexación, búsqueda y análisis de información. Puede almacenar e indexar gran variedad de tipos de datos.

Tiene elementos como **Logstach, Beats y Kibana**. Los primeros dos se encargan de manipular la información y en Kibana se visualiza e interactúa. Ofrece búsqueda y análisis casi en tiempo real, búsquedas rápidas y análisis de tendencias y patrones. Es funcional para sistemas de alto crecimiento de información.

Algunas de sus funciones son cajones de búsqueda, manejo eficiente de datos, machine learning, automatización, aplicación en geolocalización y bio informática.

Data in: Documentos e Indices

Elasticsearch no guarda la información en filas y columnas, utiliza estructuras complejas de datos que se serializan como JSON. Esto facilita la búsqueda y la conección de información de un nodo a otro en el cluster.

Inverted index: es una estructura de datos que lista todas las palabras únicas del documento e identifica cada aparición en el documento de cada una. Cuando el documento se indexa, se pueden hacer búsquedas en menos de 1 segundo.

Elasticsearch indexa toda la información en cada campo, y estos a su vez tienen estructuras de datos dedicadas dependiendo el tipo de dato. Elasticsearch detecta los tipos de datos (booleanos, punto flotante, enteros, ...) y los asocia a las estructuras específicas que facilitan la búsqueda. No es necesario configurar cómo se quiere que maneje los campos del documento. También está la otra opción, configurar todo manualmente con reglas.

Configurarlo permite análisis específico del idioma y usar formatos personalizados y datos que no se detectan automáticamente. Es posible indexar un campo de diferentes formas para diferentes propósitos. También cuando se consulta un campo de texto completo se puede aplicar la cadena de indexación que se utiliza en el texto completo.

Information out: Búsqueda y Análisis

Lo más importante de Elasticsearch es el acceso completo a Apache Lucene. REST APIs de Elasticsearch soporta consultas estructuradas, de texto completo y complejas. Se pueden hacer búsquedas de términos, frases y similitud. También se ofrece el autocompetado. Se puede usar el estilo JSON de Elasticsearch para hacer las consultas o utilizar consultas tipo SQL.

Se analiza la información haciendo un resumen de los valores clave, patrones y tendencias. De esta forma si se busca una frase, se pueden recomendar preguntas o información relacionada que puede ser útil para el usuario, son respuestas más completas que cumplen con el criterio de búsqueda.

Las agregaciones utilizan las misma estructuras que se utilizan en la búsqueda, lo que permite analizar la información en tiempo real. En una consulta documentos, filtros, ... en una misma consulta.

Adicionalmente se puede usa machine learning para:

- Detectar anomalías
- Comportamientos extraños

Escalabilidad y resilencia: clusters, nodes and shards

Elasticsearch es de alta disponibilidad, siempre va a estar disponible. Se pueden añadir servers (nodes) al cluster para aumentar la capacidad, Elasticsearch se encarga de manejar los multi nodos, entre más nodos mejor.

Se puede pensar que Elasticsearch agrupa lógicamente muchos fragmentos indexados, que distribuyéndolos por nodos se facilita la búsqueda y la redundancia, lo que ayuda a prevenir errores y pérdida de información. Existen dos tipos de fragmentos: primarios y réplicas. Entre más fragmentos existen la sobrecarga está en mantener los índices de estos.

Si se quiere hacer una búsqueda rápida, se pueden hacer varias consultas de pequeños fragmentos o pocos, pero grandes fragmentos. Cuál se usa depende de la situación. Se puede hacer un test para saber cuál es la mejor opción.

Para prevenir **desastres** existe la replicación Cross-cluster que provee sincronización automática de los índices del cluster principal a uno secundario que funciona como respaldo en caliente. Si el primero falla entonces trabaja el segundo.

Kibana es un centro de control para manejar el cluster de Elasticsearch. Es importante asegurar, manejar y monitorear los clusters.