Instituto Tecnológico de Costa Rica
IC4302 - Bases de Datos II
Resumen 1
Profesor: Nereo Campos Araya
Estudiante: Melany Salas Fernández - 2021121147

What is Elasticsearch?

Elasticsearch es un **motor de analisis y busqueda distribuido**, lo que permite la disponibilidad, facilita agregar y coleccionar datos e incluso crecer de acuerdo a las necesidades. Tambien permite la visualización de datos en un tiempo cercano al real.

Data in- Documents and indices

Elasticsearch **guarda estructuras de datos complejas**, serializadas en documentos JSON, además, cuando se tienen múltiples nodos en un clúster, los documentos guardados se distribuyen y pueden ser accedidos desde cualquier nodo.
Cuando se guardan los documentos, son indexados. La **indexación** se hace mediante índice invertido (inverted index) que permite búsqueda en textos de forma rápida, listando cada palabra que aparece en el documento e identificándola. Este índice puede verse como una colección de documentos, donde estos son una colección de campos llave valor, cada una de estas tiene una estructura optimizada.
Elasticsearch también permite que los documentos se indexen sin que se especifique cómo manejar los campos, mediante el uso mapping dinámico, que hace una detección automática, facilitando la indexación y la exploración de datos. De igual forma, se pueden definir reglas para controlar el mapping dinámico y decidir cómo se guardan e indexan los campos, esto permite personalizar formatos, implementar análisis de texto en lenguajes específicos y otros.

Information out: search and analyze

Elasticsearch brinda una Rest Api para el manejo del clúster, la indexación y la búsqueda de datos, también permite subir solicitudes desde la línea de comandos o desde la Developer Console en Kibana.

Búsqueda en los datos

Las Rest Api soportan distintas queries:

- **Queries estructuradas**: Similares a SQL
- **Full text queries**: Encuentran documentos que hacen match con el querie string y los devuelven ordenados de acuerdo con que tan buen match hacen.

  También hay queries que combinen ambos tipos, llamadas **queries complejas**.

Por otro lado, para búsquedas de términos se puede usar **phrase searches**. Para datos geoespaciales y numéricos, Elasticsearch tiene **non-textual indexes** y **estructuras optimizadas**.

El acceso a las capacidades de búsqueda de Elasticsearch se permite mediante el **JSON-style query language** (Query DSL). También se puede construir SQL-style query para buscar y agregar datos nativamente.

Análisis de datos

Las **agregaciones** permiten construir resúmenes de datos para tener patrones, estas también implementan las estructuras optimizadas de Elasticsearch. Además, las agregaciones trabajan en conjunto con las solicitudes de búsqueda, por lo que se puede filtrar resultados y analizar datos a la vez.
Además, permite el **uso de machine learning**, este sirve para la detección de anomalías en el comportamiento de datos, identificar comportamientos...

Scalability and resilience

Gracias a la facilidad para que los datos crezcan, se pueden agregar nodos a un clúster para aumentar la capacidad. Elasticsearch se encargará de distribuir la información en los nodos disponibles.

Elasticsearch usa la **redundancia** para proteger contra fallos de hardware, además, cuando un clúster crece, el fragmento de la base de datos migra de forma automática para balancear el clúster.
Tipos de fragmentos:

- **Primarios.**
- **Replicas**: Copia el fragmento primario, usando la redundancia.
  Las réplicas también permiten el aumento de la capacidad de lectura para las consultas.

Algunas consideraciones

El uso de fragmentos implica:

- Entre más fragmentos, más difícil o costoso será el mantenimiento.
- Entre más grandes sean los fragmentos, más difícil balancear el clúster.
- Consultar muchos fragmentos pequeños podría hacer el procesamiento más rápido, pero es más complicado, por lo que, dependiendo, consultar pocos fragmentos más grandes podría ser más rápido.

Debido a esto, se sabe que lo mejor depende, lo ideal es hacer una revisión y análisis de los datos para escoger la solución óptima.

Cross-cluster replication

Forma automática de sincronizar los índices del nodo primario al secundario, esto para mantener la disponibilidad de la base de datos en caso de fallas, además estos índices pueden usarse para la lectura de datos.

Kibana

Es un centro de control para manejar los clúste, además, permite la interacción con los datos.

Resumen de What is Elasticsearch?