Administración de Sistemas - Curso 2023 / 2024

Laboratorio 10.- Elasticsearch

Contenido:

Objetivos: Profundizar en las técnicas de búsqueda de Elasticsearch.

Administración de Sistemas - Curso 2023 / 2024

1 Preparar el entorno

Para realizar este laboratorio es necesario lo siguiente:

- Un despliegue funcional de Elasticsearch. Se recomienda utilizar Docker Compose. Las instrucciones se encuentran en las diapositivas de teoría.
- El dataset de pruebas accounts.json distribuido por Elastic. El fichero contiene datos ficticios de 1000 personas, incluyendo su nombre, apellido, edad o ciudad de residencia.
    - Descargar el fichero comprimido que contiene el dataset: https://download.elastic.co/demos/kibana/gettingstarted/accounts.zip
    - Crear un índice "bank" en Elasticsearch e indexar los datos utilizando la carga en bruto. No es necesario definir un esquema concreto.
    - Antes de comenzar con los ejercicios, se recomienda explorar los datos para familiarizarse con el formato y los campos de los documentos. El esquema es el siguiente:

```
"account_number": INT,
"balance": INT,
"firstname": "String",
"lastname": "String",
"age": INT,
"gender": "M or F",
"address": "String",
"employer": "String",
"email": "String",
"city": "String",
"state": "String"
```

2 Búsquedas

En esta parte del laboratorio se trabaja con diferentes formas de hacer búsquedas en datos almacenados por Elasticsearch. Antes de comenzar, se recomienda completar los ejercicios de las diapositivas de teoría y repasar el funcionamiento básico de Elasticsearch.

2.1 Paginación

Por defecto, Elasticsearch devuelve los 10 primeros resultados que coinciden para una búsqueda. Sin embargo, es posible configurar una búsqueda para que devuelva tantos resultados como queramos y, si lo necesitamos, que se devuelvan en grupos. Obtener los resultados agrupados se llama "paginación" y se utiliza, p.e., para recuperar los resultados de búsqueda en bloques y después mostrarlos en una página Web.

La documentación oficial sobre cómo configurar la paginación en Elasticsearch está aquí:
https://www.elastic.co/guide/en/elasticsearch/reference/current/paginate-search-results.html

Revisar la documentación sobre paginación y realizar las siguientes tareas:

- Recuperar los primeros 20 documentos del índice.
- Recuperar los segundos 20 documentos de índice.
- Buscar los datos de las personas que residen en Texas (código de estado "TX") y devolver los primeros 15 resultados.

2.2 Ordenación

Los resultados de una búsqueda en Elasticsearch se pueden mostrar ordenados por algún criterio que definamos, p.e. en orden ascendente en base al campo "Año" del documento.

Sin embargo, aunque es posible ordenar resultados de búsqueda por valores numéricos, no es posible hacerlo por campos de texto que hayan sido marcados para su búsqueda como texto (p.e, como "text"). Esto es debido a que los campos de texto se dividen para ser analizados en forma de índice invertido. Como resultado, sólo se pueden utilizar los campos de texto marcados como "keyword" para ordenar.

La documentación oficial sobre cómo configurar búsquedas con resultados ordenados esta aquí: https://www.elastic.co/guide/en/elasticsearch/reference/current/sort-search-results.html

Revisar la documentación sobre ordenación de resultados y realizar las siguientes tareas:

- Recuperar los datos de los residentes en el estado de Los Ángeles (código de estado "LA") y mostrar los resultados ordenados por edad de forma ascendente.
- Recuperar los datos de los residentes en el estado de New Jersey (código de estado "NJ") y mostrar los resultados ordenados por su balance de forma ascendente.

2.3 Filtros

Elasticsearch dispone de filtros para eliminar documentos de los resultados de una búsqueda. Esta característica está descrita en las diapositivas de teoría y aquí se proponen algunas tareas para profundizar más en ella.

La documentación oficial sobre filtros de búsqueda está aquí:
https://www.elastic.co/guide/en/elasticsearch/reference/current/filter-search-results.html

Revisar la documentación sobre filtros de resultados y realizar las siguientes tareas:

- Recuperar los datos de usuarios cuyo estado de residencia sea Los Ángeles (código de estado "LA") pero que su ciudad de residencia no sea Loretto, y que además su edad sea superior a los 33 años.
- Recuperar los datos de usuarios cuyo estado de residencia sea Ohio (código de estado "OH") y su edad sea superior a 39 años.

2.4 Búsquedas difusas

Elasticsearch es capaz de búsquedas difusas, es decir, búsquedas tolerantes a fallos gramaticales y errores de escritura. Esta característica se basa en aplicar la distancia de Levenshtein a los términos de búsqueda para establecer un umbral con el que filtrar resultados.

La distancia de Levenshtein se aplica a las siguientes operaciones (en los siguientes ejemplos el valor de distancia es 1):

- Sustitución de caracteres: p.e., detectar "circuler" como "circular".
- Inserción de caracteres: p.e., detectar "cirrcular" como "circular".
- Borrado de caracteres: p.e., detectar "cirular" como "circular".
- Intercambiar 2 caracteres adyacentes: p.e., detectar "ciruclar" como "circular".

La documentación oficial sobre consultas difusas está aquí:
https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-fuzzy-query.html

Revisar la documentación sobre consultas difusas y realizar las siguientes tareas:

- Utilizar una búsqueda difusa para recuperar los datos de la persona residente en Wyoming, pero utilizando "Woyming" como término de búsqueda
- Modificar la búsqueda anterior para mostrar los datos de la misma persona, pero utilizando "Wyomin" como término de búsqueda.

2.5 Prefijos de búsqueda y comodines

En Elasticsearch es posible realizar la búsqueda especificando sólo una parte del comienzo de los términos a buscar, p.e. buscar en el campo "anyo" con el prefijo "201" para encontrar los valores 2011, 2012, ...

También se pueden utilizar comodines (*wildcards*) para expresar patrones de búsqueda, al igual que se hace en algunos comandos Linux, p.e. utilizar el comodín * como término de búsqueda "co*e" para encontrar "coche", "coge", "come", ...

La documentación oficial para la búsqueda con prefijos está aquí:
https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-prefix-query.html

La documentación oficial sobre búsquedas con comodines está aquí:
https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-wildcard-query.html

Revisar ambas documentaciones y realizar las siguientes tareas:

- Recuperar los datos de las personas cuyos apellidos comiencen por "Mc".
- Recuperar los datos de las personas cuya ciudad de residencia comience por la letra "G" y acabe con "field".

2.6 Expresiones regulares

Como un paso más allá de utilizar comodines en las búsquedas, Elasticsearch permite utilizar expresiones regulares para encontrar patrones concretos. Se puede encontrar un listado de los caracteres y operadores aceptados por Elasticsearch en: https://www.elastic.co/guide/en/elasticsearch/reference/current/regexp-syntax.html

La documentación oficial sobre los operadores de búsquedas utilizando expresiones regulares está aquí:
https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-regexp-query.html

Revisar la documentación oficial y realizar las siguientes tareas:

- Utilizando expresiones regulares, recuperar los datos de las personas cuyo nombre de empleador comience por la letra "A" y esté compuesto por 4 o 5 letras, p.e. los empleadores "Avit" o "Amtap".