Maping – định dạng một kiểu dữ liệu cho document

Analyse

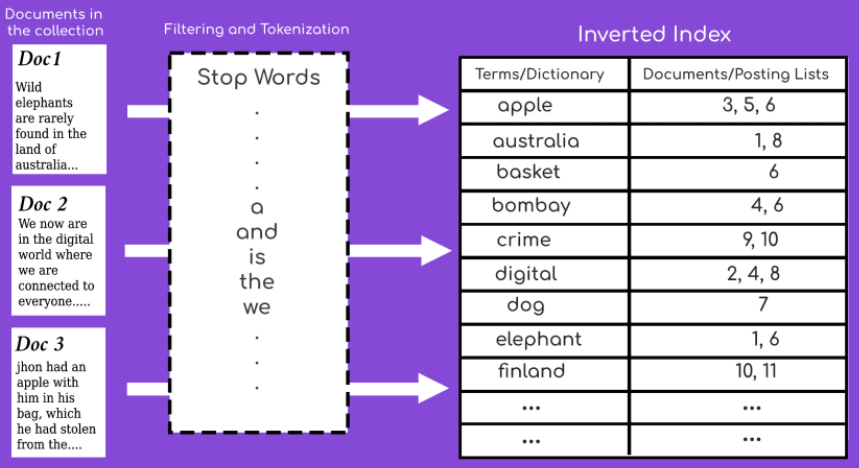
1. Inverted index

Trong Elasticsearch, inverted index là một phần quan trọng của cơ sở dữ liệu văn bản, được xử lý bới Apache lucence, giúp tăng tốc quá trình tìm kiếm và truy xuất thông tin từ các văn bản lớn.

Invertd index bao gồm tất cả các từ duy nhất xuất hiện trong bất kỳ document nào, và đối với mỗi từ sẽ có 1 danh sách document mà nó xuất hiện trong đó.

Một số khái niệm cơ bản để hiểu về inverted index trong Elasticsearch:

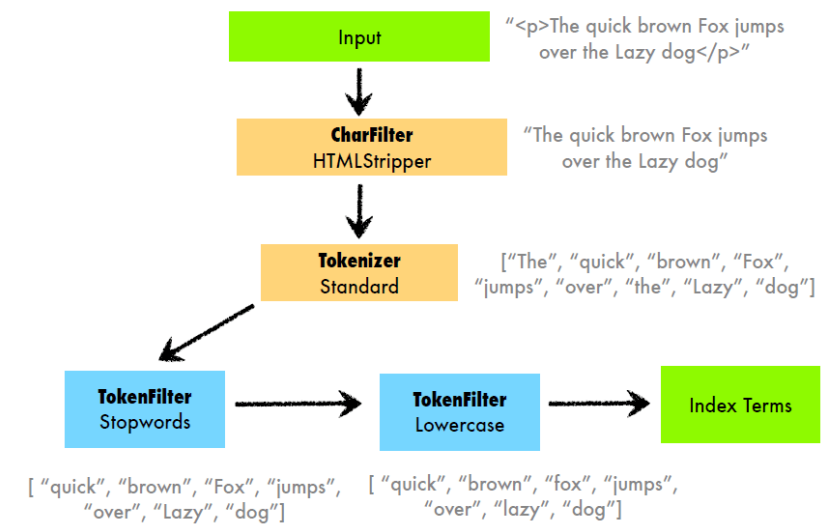
* Tokenization:
  + Trước khi xây dựng inverted index, văn bản thường được chia thành các “tokens” (các đơn vị nhỏ nhất như từ hoặc từ khóa)
  + Các bước này gọi là quá trình tokenization, và Elasticsearch có thể tùy chỉnh quá trình này để phù hợp với nhu cầu của ứng dụng.
* Stop Words và Stemming:
  + Elasticsearch thường loại bỏ từ ngữ không quan trọng như stop words (ví dụ: “and”, “the”) để giảm kích thước của inverted index.
  + Nó cũng thực hiện stemming, là quá trình chuyển đổi từ về dạng gốc của nó, để tăng khả năng tìm kiếm.



1. Analysis và Analyzer

Analysis là quá trình bao gồm 2 bước:

* Đầu tiền, thực hiện tokenizing văn bản trở thành các inverted index
* Sau đó chuẩn hóa các inverted index này theo một quy tắc để cải thiện khả năng tìm kiếm.



Analyzer là quá trình bao gồm 3 chức năng CharFilter, Tokenizer, Token filters được gom thành một gói duy nhất.

|  |  |
| --- | --- |
|  |  |

Có thể sử dụng các loại analyzer sẵn có hoặc cũng có thể tự tạo một analyzer riêng ([tham khảo](https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-analyzers.html))

1. Mapping

Trong Elasticsearch, Mapping là một khái niệm quan trọng liên quan đến cách dữ liệu được định dạng và lưu trữ trong chỉ mục. Mapping định rõ cấu trúc của các trường trong tài liệu và xác định kiểu dữ liệu của chúng.

Trong elasticsearch có 2 loại mapping:

* Dynamic mapping: tự động mapping dựa trên cấu trúc dữ liệu mới được thêm vào index
* Explicit mapping: định rõ mapping cho các trường trong index.

Trong Elasticsearch, **datatype** (kiểu dữ liệu) định rõ loại dữ liệu mà một trường cụ thể trong tài liệu có thể chứa. Dưới đây là một số kiểu dữ liệu phổ biến trong Elasticsearch:

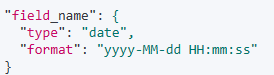
* **Text**: được sử dụng để lưu trữ chuỗi văn bản dài.



* **Keyword:** thường được sử dụng để lưu trữ giá trị không phân tích, ví dụ như các từ khóa hoặc giá trị không thay đổi.



* **Date:** sử dụng để lưu trữ thông tin về thời gian.



* **Interger:** sử dụng lưu trữ số nguyên.



* **Float và Double:**  sử dụng lưu trữ số thực với độ chính xác khác nhau.



* **Boolean:** sử dụng lưu trữ giá trị đúng sai.



* **Object:** là kiểu dữ liệu đặc biệt lưu trữ đối tượng.



* **Nested:** là một kiểu đặc biệ của object, được sử dụng khi bạn muốn lưu trữ một mảng các đối tượng.



|  |  |
| --- | --- |
|  |  |

[Snap Labs](https://www.snaplabs.io/insights/snort-and-elastic-stack)