Elasticsearch 기초

Elasticsearch란

- Elasticsearch : **텍스트, 숫자, 위치 기반 정보, 정형 및 비정형 데이터 등 모든 유형의 데이터를 위한 분산형 오픈 소스 검색 분석 엔진**
- Elasticsearch의 특징 :
    - 간단한 REST API
    - 확장성 및 빠른 속도
    - 수많은 종류의 콘텐츠를 색인할 수 있는 능력

Elasticsearch의 사용 사례

- 애플리케이션 검색
- 웹사이트 검색
- 엔터프라이즈 검색
- 로깅과 로그 분석
- 인프라 매트릭과 컨네이너 모니터링
- 애플리케이션 성능 모니터링
- 위치 기반 정보 데이터 분석 및 시각화
- 보안 분석
- 비즈니스 분석

Elasticsearch 인덱스란

- **Elasticsearch index(인덱스, 한국어로 색인)** : 서로 관련되어 있는 문서들의 모음. 각 문서는 **일련의 키**(필드나 속성의 이름)와 **그에 해당하는 값**(문자열, 숫자, 부울, 날짜, 값의 배열, 지리적 위치 또는 기타 데이터 유형)을 서로 연결
- Elasticsearch는 **inversed index**를 사용. 이는 문서에 나타나는 모든 고유한 단어의 목록을 만들어, 각 단어가 발생하는 모든 문서를 식별.
- 색인 프로세스 : 문서를 저장하고 inversed index를 구축하여 거의 실시간으로 문서를 검색 가능한 데이터로 만듦. 인덱스 API를 사용해 색인이 시작되며, 이를 통해 사용자는 특정한 인덱스에서 JSON 문서를 추가하거나 업데이트 할 수 있음

Logstash의 사용 목적

- Logstash : Elastic Stack의 핵심 제품 중 하나로, **데이터를 집계하고 처리**하여 **Elasticsearch로 전송**하는데 사용됨. 즉, Logstash를 이용해 다양한 소스에서 동시에 데이터 수집이 가능하고, 이 데이터를 강화하고 변환한 다음, Elasticsearch에서 색인되도록 함.

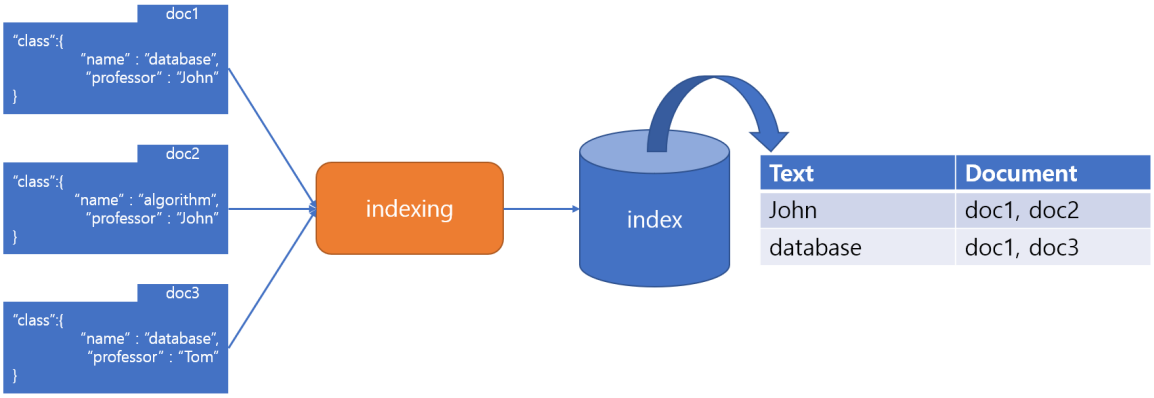
Kibana의 사용 목적

- Kibana : Elasticsearch를 위한 시각화 및 관리 도구로서, 실시간 히스토그램, 선 그래프, 파이 차트, 지도 등을 제공.

검색 과정

데이터 수집 → 색인 → 검색 요청 → 조회/평가 → 결과

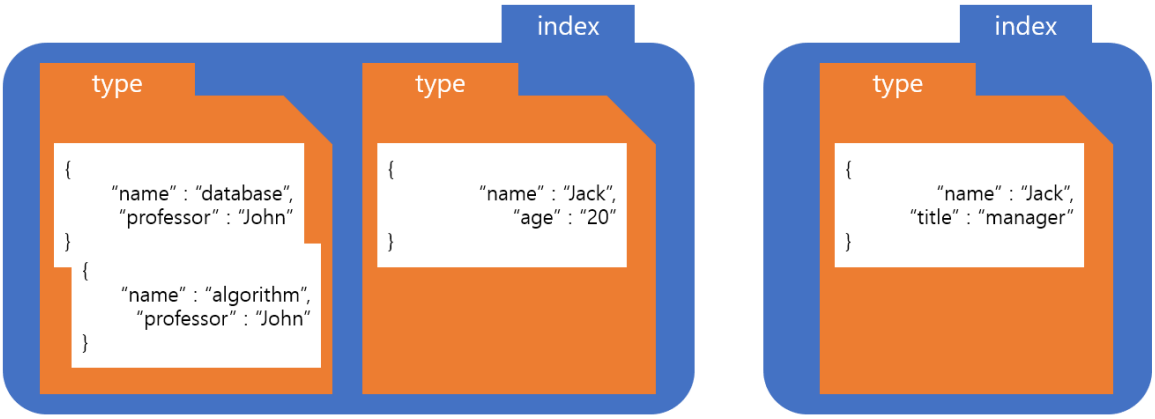
색인 할 때의 DATA FLOW



Elasticsearch VS. Relational DB

- Elasticsearch : **inversed index** 구조를 사용. keyword가 어떤 document에서 나왔는지만을 저장 → 더 빠름.
- Relational DB : 일반적인 index 구조를 사용 → 느림. query에 해당하는 단어가 있는지 모든 documet를 살펴봐야 하기 때문

Elasticsearch Data structure



- index > type > document

Elasticsearch VS. Relational DB

Elasticsearch	Relational DB
index	database
type	table
document	row
field	column
mapping	schema

Elasticsearch	Relational DB
GET (조회)	Select (조회)
PUT (수정)	Update (수정)
POST (삽입)	Insert (삽입)
DELETE (삭제)	Delete (삭제)

Elasticsearch의 기본 개념/용어

- **index/type/document** : elasticsearch의 데이터 계층. REST API에서 문서를 표현할 때는 `/news/article/10000`과 같이 표현. 이때, `news` : index, `article` : type, `10000` : document
- **field** : 문서의 property.
- **mapping** : index/type/document의 규칙을 정의한 것. 사용자가 직접 정의할 수 있음.
- **index(색인)** : V. elasticsearch가 문서를 검색할 수 있도록 색인 데이터를 만들어두는 과정.
- **index(색인)** : N. 색인 작업을 거쳐 만들어진 색인 데이터를 의미.
- **cluster/node** : 여러 대의 서버를 묶어서 구동하기 위해 사용되는 개념. 각 서버가 note, 서버의 묶음이 cluster.
- **shard/replica(샤드/복사본)** : elasticsearch는 index data를 여러개의 저장공간에 나누거나 복사 가능
- **QueryDSL** : JSON으로 표현되는 elasticsearch의 검색 문법

참고자료

- https://www.elastic.co/kr/what-is/elasticsearch
- https://www.youtube.com/watch?v=B1Aq2GQ4E78&list=PLVNY1HnUlO24LCsgOxR_eK2Yi4sOgH9Pg&index=3
- https://bakyeono.net/post/2016-06-03-start-elasticsearch.html