Elasticsearch 시작하기(1회차)

목차

- Elasticsearch 에 대하여
- Elasticsearch 의 용어 및 개념정리
- Elasticsearch 설치하기
- Q & A

Elasticsearch 에 대하여

Elasticsearch 에 대하여

Elastisearch 란?

Elasticsearch 는 고가용성의 확장 가능한 오픈소스 이면서,

1) Full-Text 검색엔진이자,

2) 분석 엔진

이다.

Elasticsearch 에 대하여

Lucene

- Apache Software 재단의 검색엔진 상위 프로젝트
- Java 언어로 이루어진 정보 검색 라이브러리
- Free and Open-source Software
- Doug Cutting 에 의해 개발
- Apache License 하에 배포

Elasticsearch 에 대하여

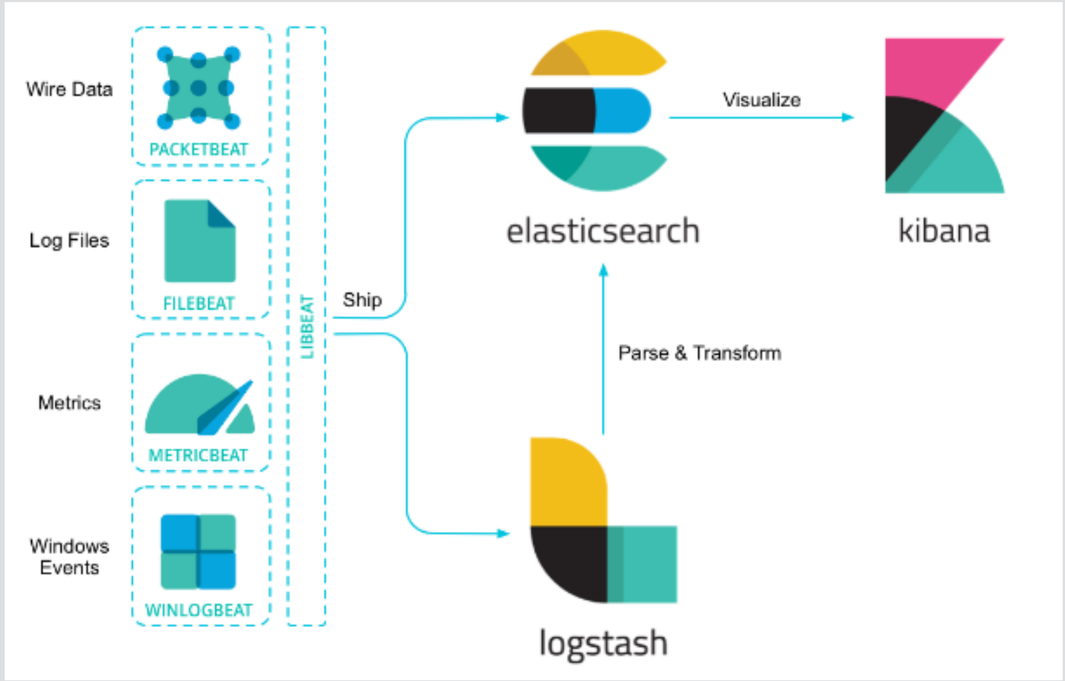
Full-Text 검색엔진 으로서의 Elastisearch

- Shay Banon 이 Lucene 을 기반으로 만든 검색엔진
- Apache 2.0 License 에 의거
- HTTP Web Interface 와 Schema 에 자유로운 Json 형태의 문서 지원
- 준 실시간 분산형 검색엔진(NRT - Near Real Time Engine)

Elasticsearch 에 대하여

분석 엔진으로서의 Elastisearch

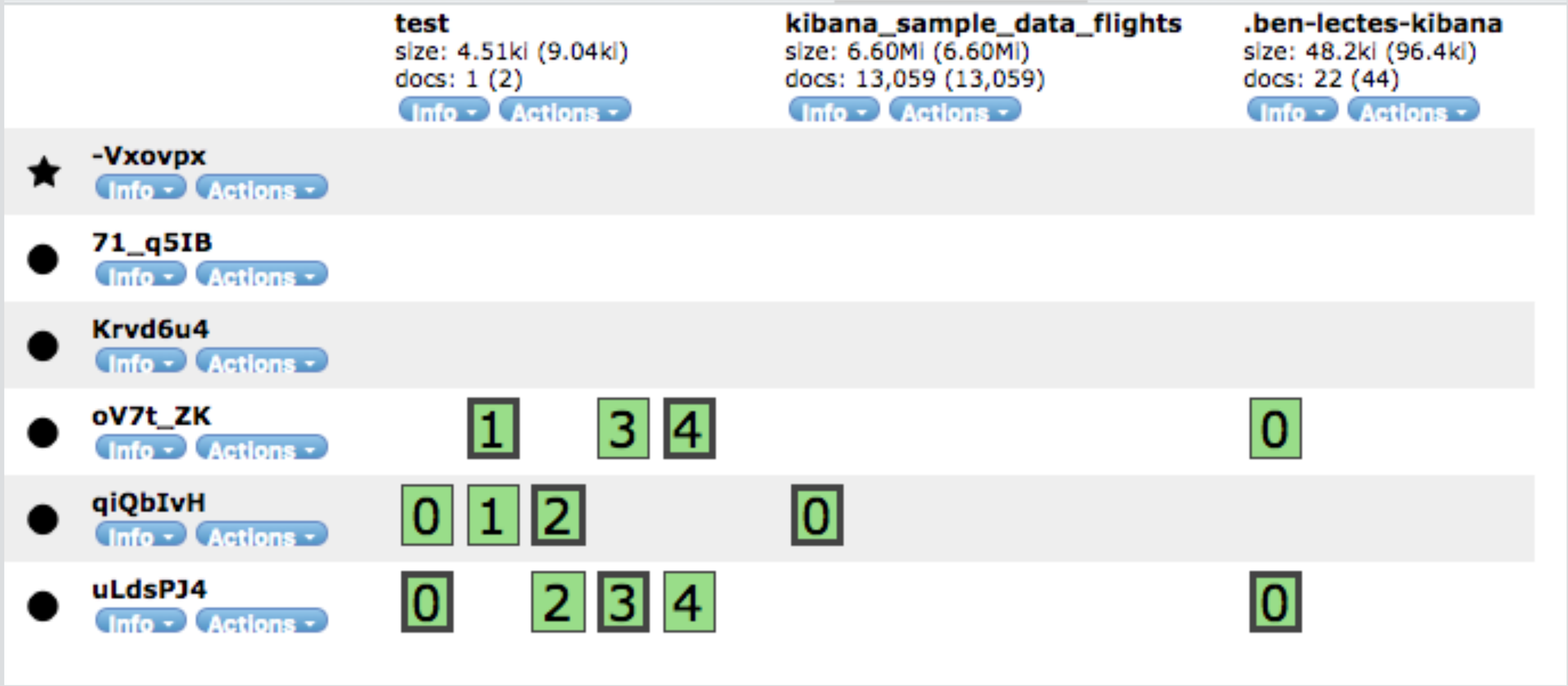
- 검색엔진은 단독으로 서비스
- Beats, Logstash, Kibana 를 통해 분석 엔진으로 사용
- 준 실시간 분석이 가능한 분석엔진
- 로그 등을 스트리밍하는 Beats
- 필터링을 통해 Elasticsearch 로 문서를 저장하는 Logstash
- 수집된 데이터를 통계/집계 내어 시각화하는 Kibana



Elasticsearch 의 용어 및 개념 정리

Elasticsearch 의 용어 및 개념 정리

ElasticSearch Cluster



Elasticsearch 의 용어 및 개념 정리

문서(Document)

- 문서는 JSON(Java Script Object Notatoin) 형태의
  실제 의미있는 데이터를 가진 Elasticsearch 저장 기본단위
- RDB 의 row 와 비슷한 개념
- 문서는 ElasticSearch 에 저장될 때 고유한 문서 ID 를 갖음
- 문서 ID 는 Random 값이나 사용자 정의된 값으로 정의
- 문서 ID 는 문서 데이터를 찾아가는 Meta key

Elasticsearch 의 용어 및 개념 정리

문서(Document)

- JSON(Java Script Object Notation) 은
  사람이 읽고 쓰기 쉬운 형태의 경량 데이터 구조

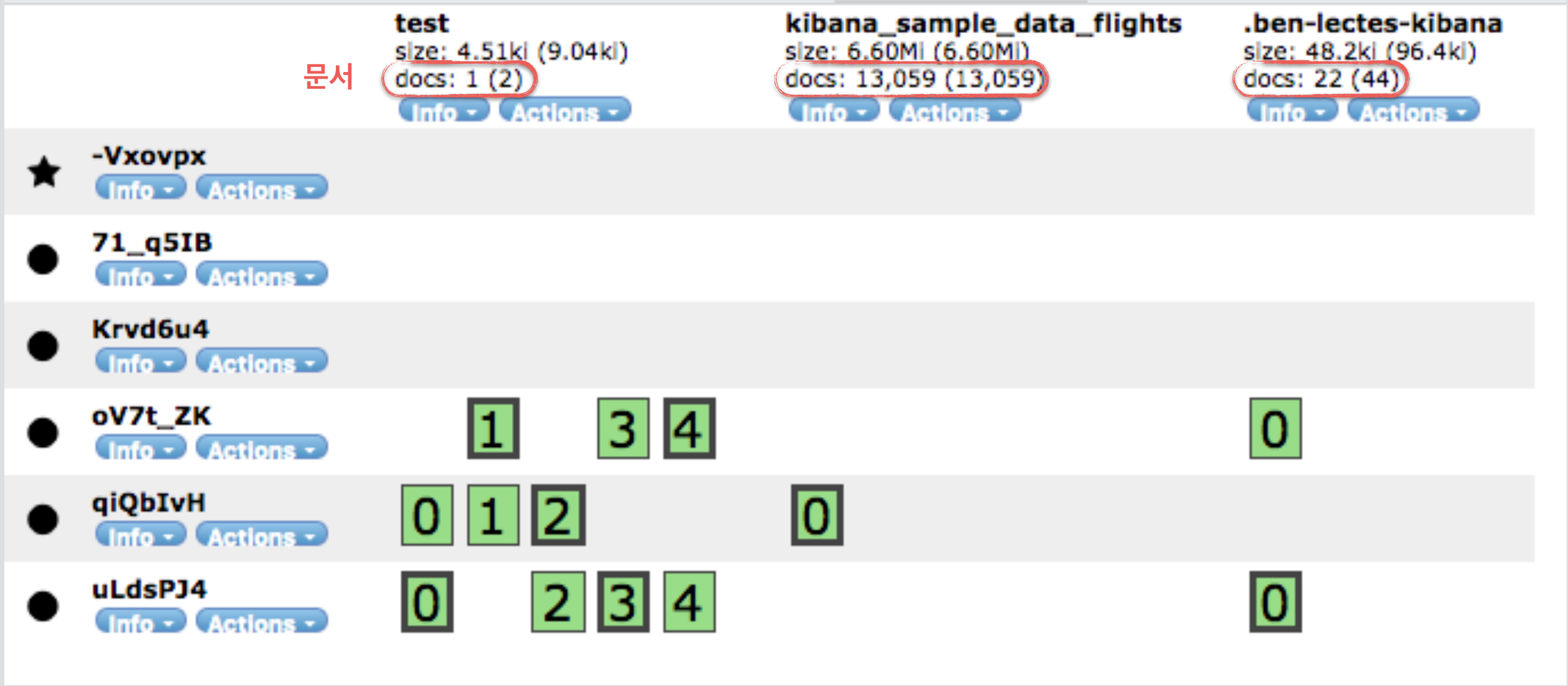
- Key:Value 의 쌍으로 사용

- Value 내에 key:value 형태의 object 나 list 도 포함

```
{
  "name": "PSH",
  "id": "benjamin",
  "phone": [
    "010-0000-0000",
    "010-1111-1111"
  ],
  "address": {
    "city": "kyung-ki",
    "town": "yong-in"
  },
  "part": "system engineer"
}
```

Elasticsearch 의 용어 및 개념 정리

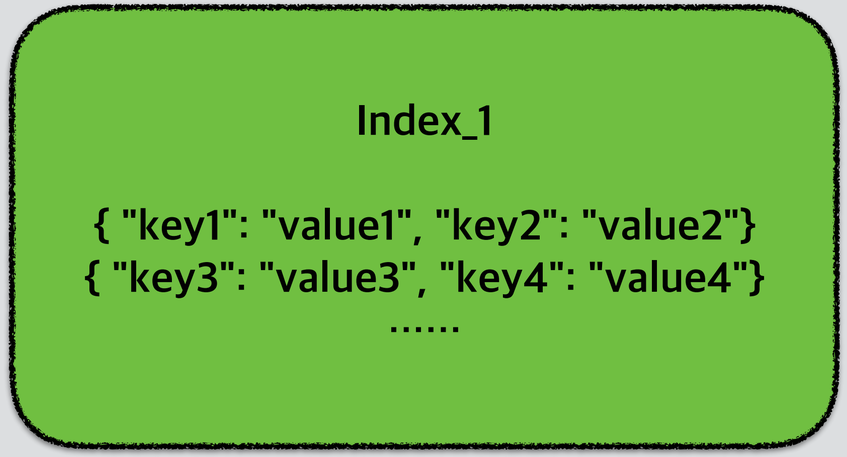
문서(Document)



Elasticsearch 의 용어 및 개념 정리

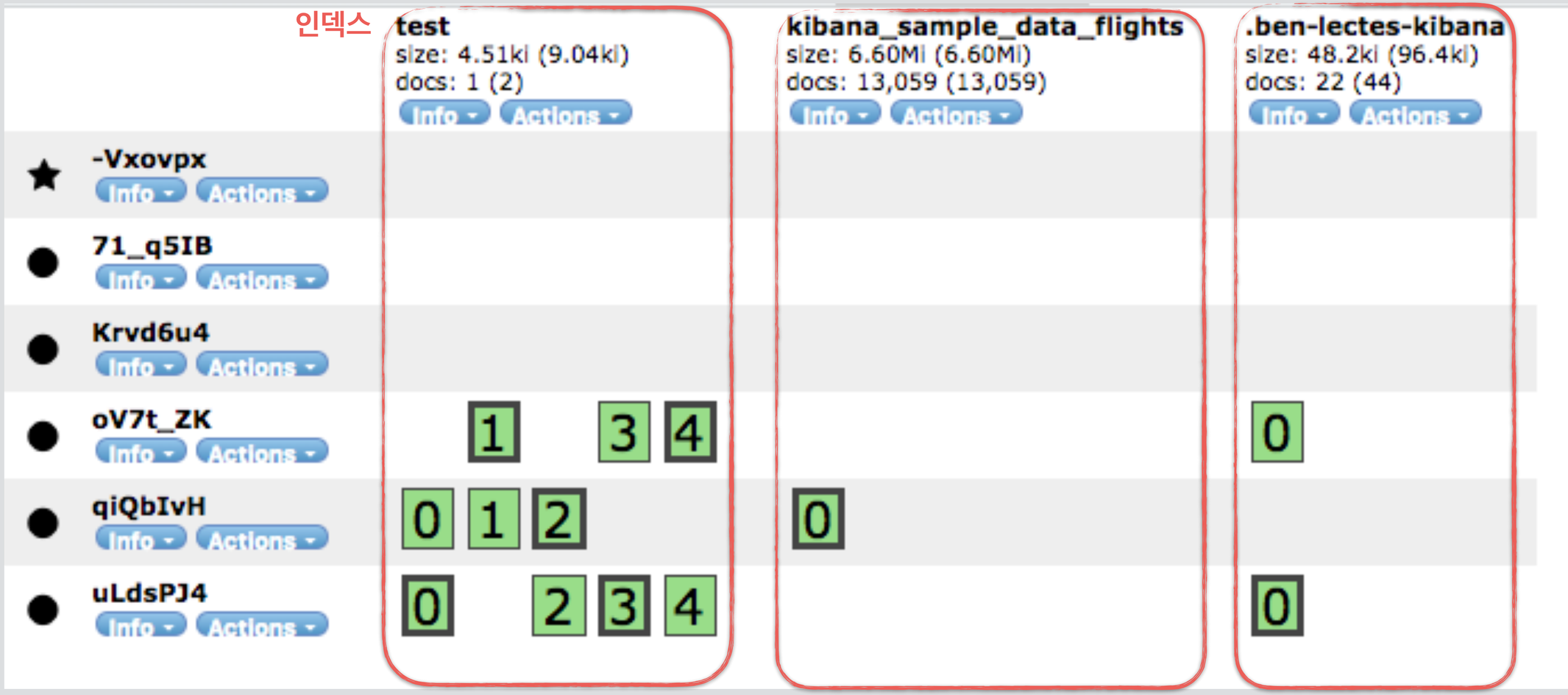
인덱스(Index)

- 문서가 저장되는 가장 큰 단위
- RDB 의 데이터베이스와 비슷한 개념
- 하나의 인덱스에 여러 문서 저장
- 최초 데이터를 클러스터에 저장할 때 인덱스가 자동으로 생성



Elasticsearch 의 용어 및 개념 정리

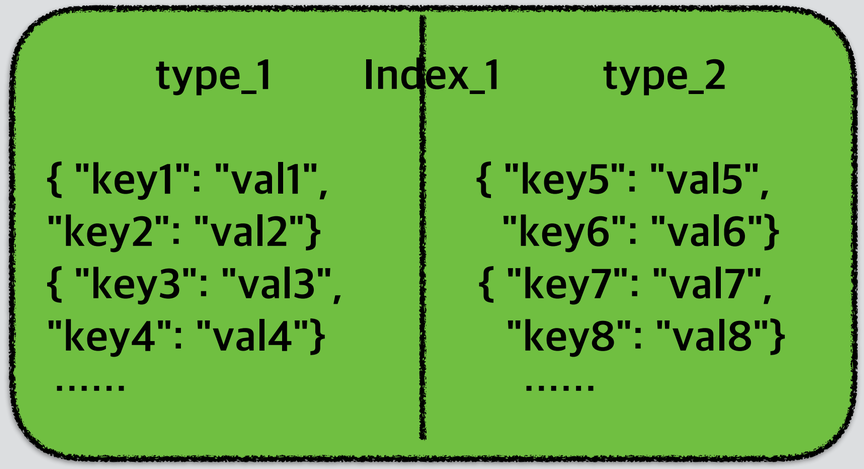
인덱스(Index)



Elasticsearch 의 용어 및 개념 정리

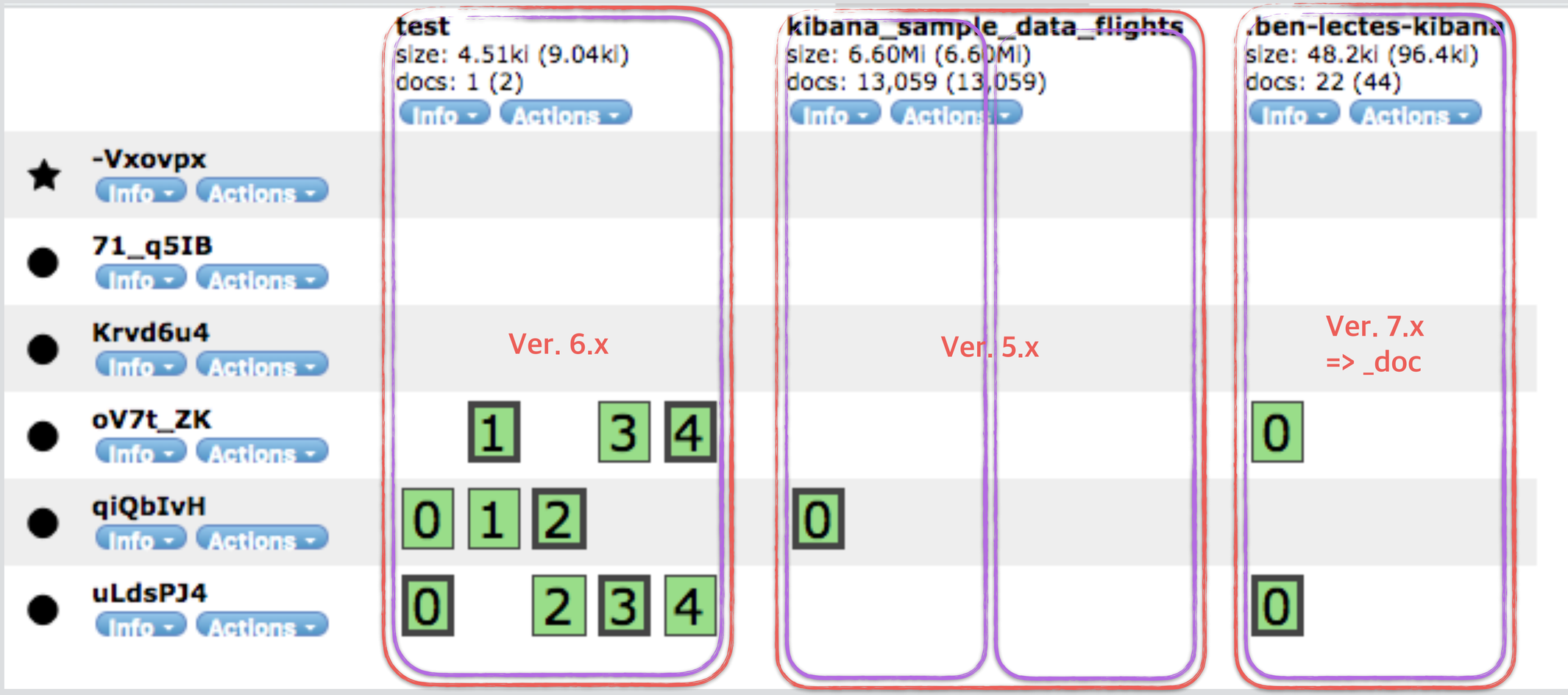
타입(Type)

- 인덱스의 파티션
- RDB 의 테이블과 비슷한 개념
- ElasticSearch 6.x 버전부터는 Multi Type Deprecated
- 하나의 인덱스에 Single Type 권고(_doc)
- ElasticSearch 7.x 버전부터는 _doc 로 고정



Elasticsearch 의 용어 및 개념 정리

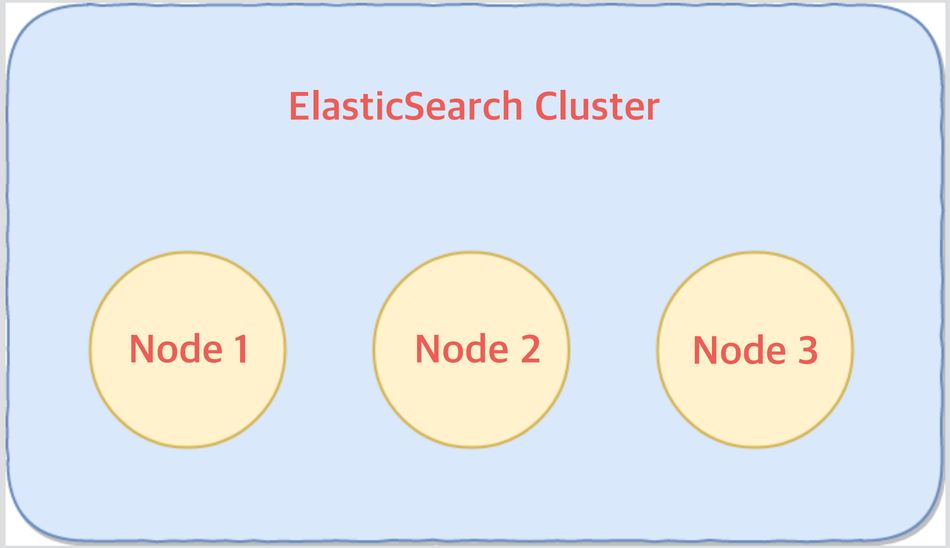
타입(Type)



Elasticsearch 의 용어 및 개념 정리

클러스터(Cluster)

- ElasticSearch 는 클러스터로 구성됨
- 클러스터는 하나 이상의 노드로 구성됨
- 사용자는 클러스터를 대상으로 데이터를 저장하거나 검색 요청
- 클러스터 별 cluster_name 과 cluster_uuid 사용



Elasticsearch 의 용어 및 개념 정리

노드(Node)

- 클러스터를 구성하는 ES 프로세스
- 노드 간 헬스 체크, 문서를 색인하여 저장, 검색 요청에 의해 데이터를 리턴
- 클러스터와 마찬가지로 노드 별 name 과 UUID 사용
- 노드의 역할 정의에 따라 master, data, ingest, client 노드로 사용

Elasticsearch 의 용어 및 개념 정리

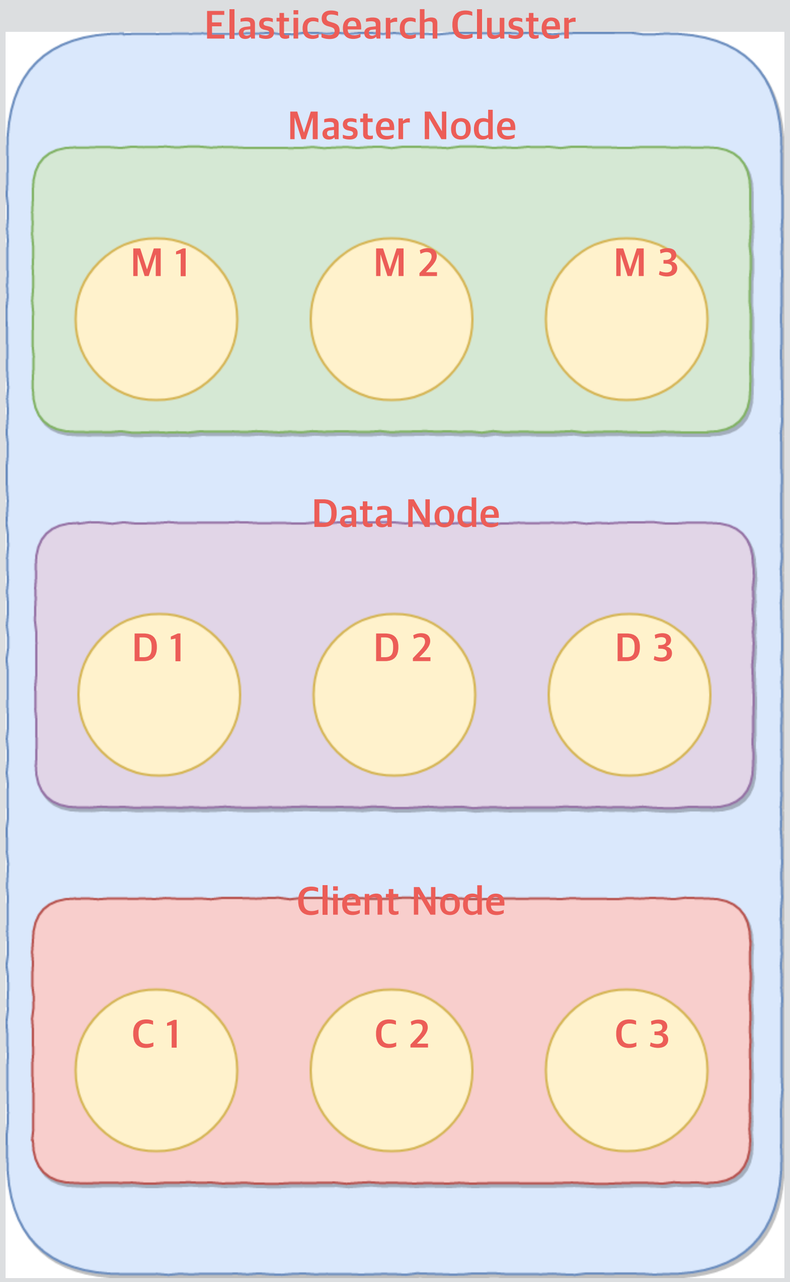
노드(Node)

master 노드
- 클러스터 구성의 기준이 되는 노드
- 클러스터 내 노드들의 헬스 체크를 담당
- 클러스터 내에서 벌어지는 모든 업데이트 사항을 보고받는 노드

data 노드
- 사용자의 문서가 저장되는 노드
- 사용자의 문서 요청에 데이터를 리턴해주는 노드

all 노드
- 마스터와 데이터 노드의 구분이 필요 없을 때 사용
- 데이터 노드 확장이 거의 필요 없는 상태일 때 비용 절감을 위해 사용

client 노드
- 사용자의 쿼리를 받기 위한 노드
- 사용자에게 받은 쿼리를 데이터 노드에게 전달하는 노드
- 데이터 노드가 리턴해 준 문서를 취합하여 사용자에게 리턴해주는 노드



Elasticsearch 의 용어 및 개념 정리

클러스터(Cluster) & 노드(Node)

Elasticsearch 의 용어 및 개념 정리

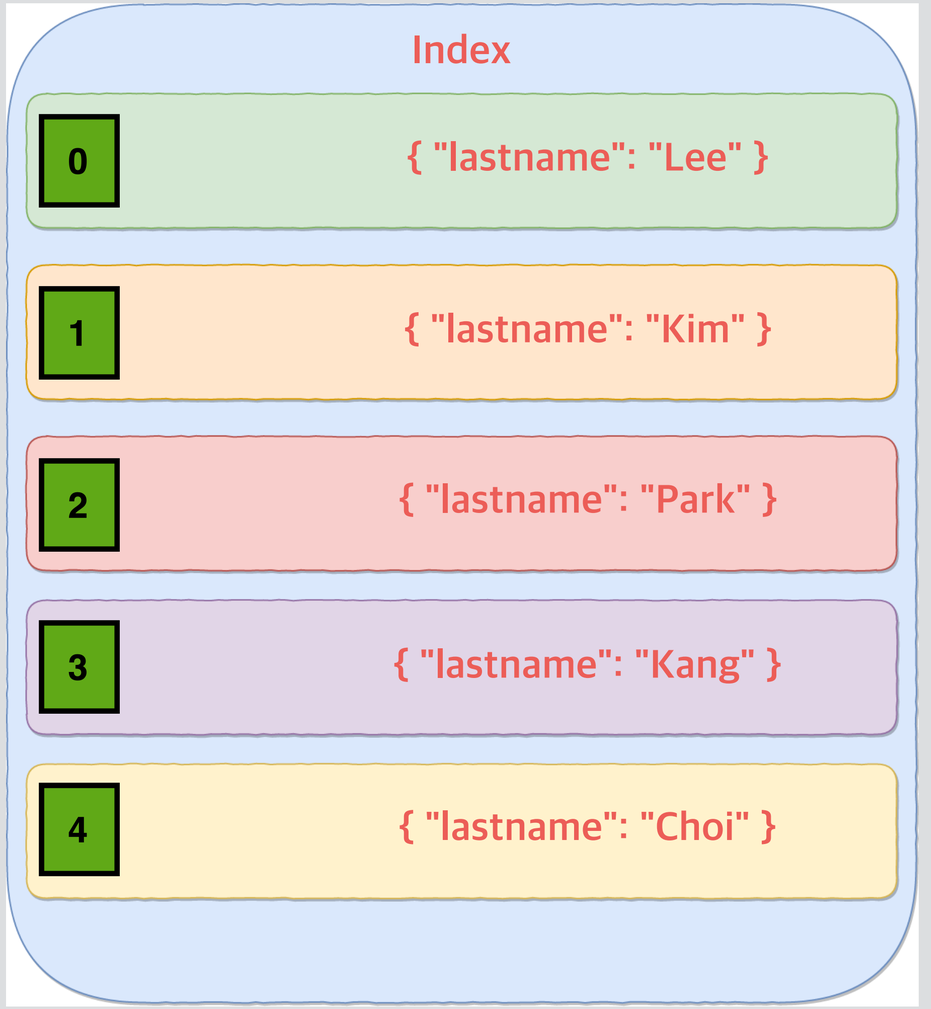
샤드(Shard)

인덱스의 데이터를 나누는 단위
- 인덱스는 많은 개수의 문서들을 저장
- 무한정 데이터를 인덱스에 넣다보면

1) 단일 노드의 디스크 볼륨 크기의 유한성으로
   더 이상 문서 저장을 할 수 없는 순간이 오게 됨

2) 단일 노드의 유한한 CPU, 혹은 Memory 자원으로
   색인이나 검색 성능 저하

- 이러한 문제를 해결하기 위해 도입된 개념이 샤딩(Sharding)



Elasticsearch 의 용어 및 개념 정리

샤드(Shard)

- 각 Elasticsearch 샤드는 Lucene 인덱스

- 단일 Lucene 인덱스가 포함할 수 있는 문서 수의 최대 한도가 2,147,483,519 건

- Replica Shard 가 있기 때문에 샤드/노드 오류가 발생하더라도 ElasticSearch 클러스터의 고가용성이 유지

- 모든 Replica Shard 에서 병렬 방식으로 검색을 실행할 수 있으므로 검색 처리량 확장 가능

- 색인할 때 Primary Shard 의 복제를 하는 과정이 추가되기 때문에
  1) I/O 가 두배로 발생하기 때문에 색인 성능 저하
  2) 디스크 볼륨도 실제 문서의 두배 필요

Elasticsearch 의 용어 및 개념 정리

원본 샤드(Primary Shard)

- Primary Shard 는 색인되어 저장되는 문서의 원본 Shard 를 의미
- 인덱스에 문서가 색인될 때 가장 처음에 생성되는 Shard
- Primary Shard 에는 Shard 번호가 할당되어 어느 노드에 어떤 샤드가 할당되어 있는지 식별 가능
- 한 번 지정한 샤드 갯수는 불변
- 6.x 버전에서는 5개의 Primary Shard를 기본으로 설정
- 7.x 버전에서는 1개의 Primary Shard를 기본으로 설정

Elasticsearch 의 용어 및 개념 정리

복제본 샤드(Replica Shard)

- Replica Shard 는 색인되어 들어온 문서의 복제본 Shard 를 의미

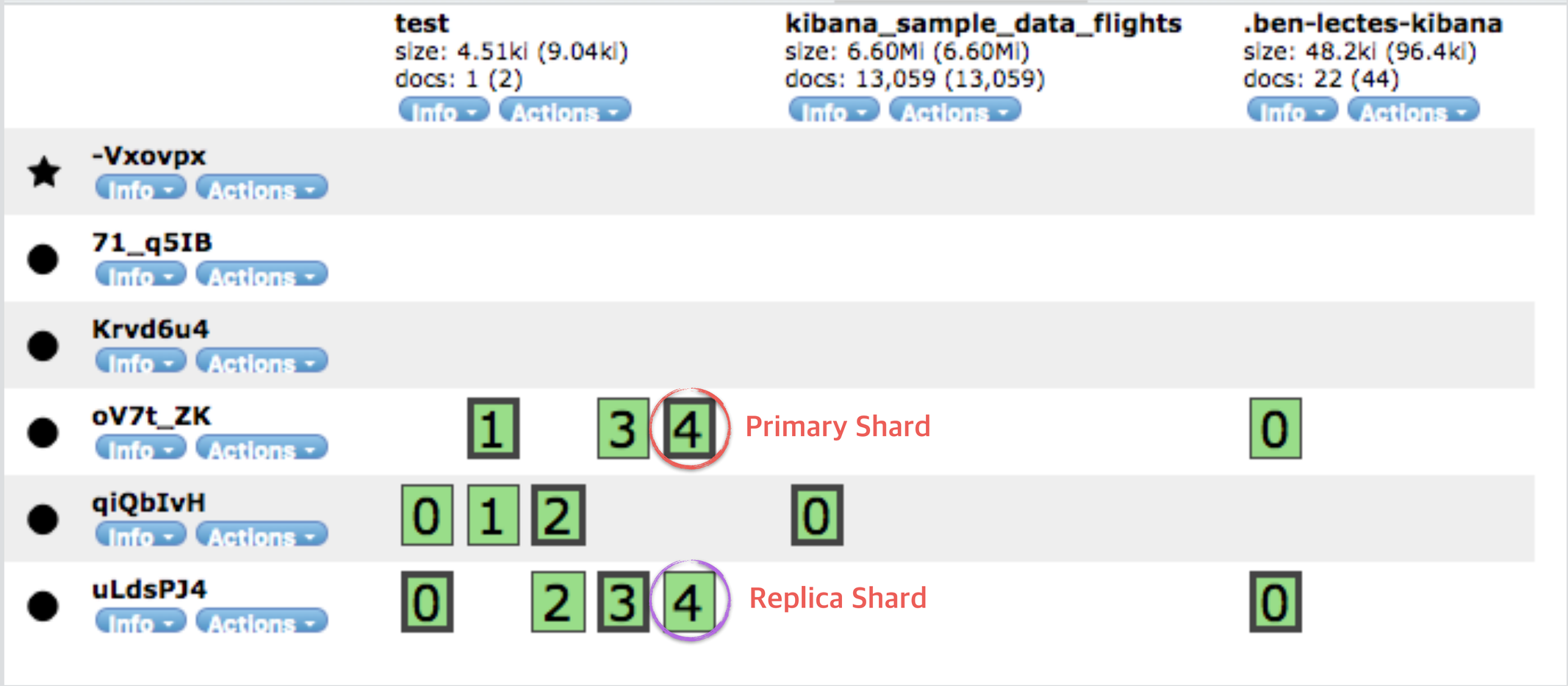
- ElasticSearch 에 Primary Shard 가 색인된 후,
  Primary Shard 가 저장된 데이터 노드와 다른 데이터 노드에 복제되는 Shard

- Replica Shard 도 마찬가지로 번호가 할당되어 어떤 Primary Shard 의 복제본인지 식별 가능

- 1개의 Replica Shard를 기본으로 설정

Elasticsearch 의 용어 및 개념 정리

샤드(Shard)



Elasticsearch 의 용어 및 개념 정리

샤드(Shard)

- 각 Elasticsearch Shard 는 Lucene 인덱스
- 단일 Lucene 인덱스가 포함할 수 있는 문서 수의 최대 한도가 2,147,483,519 건
- Replica Shard 가 있기 때문에 샤드/노드 오류가 발생하더라도 ElasticSearch 클러스터의 고가용성이 유지
- 모든 Replica Shard 에서 병렬 방식으로 검색을 실행할 수 있으므로 검색 처리량 확장 가능
- 색인할 때 Primary Shard 의 복제를 하는 과정이 추가로 발생
  1) I/O 가 두 배로 발생하기 때문에 색인 성능 저하
  2) 디스크 볼륨도 실제 문서 용량의 두배 필요

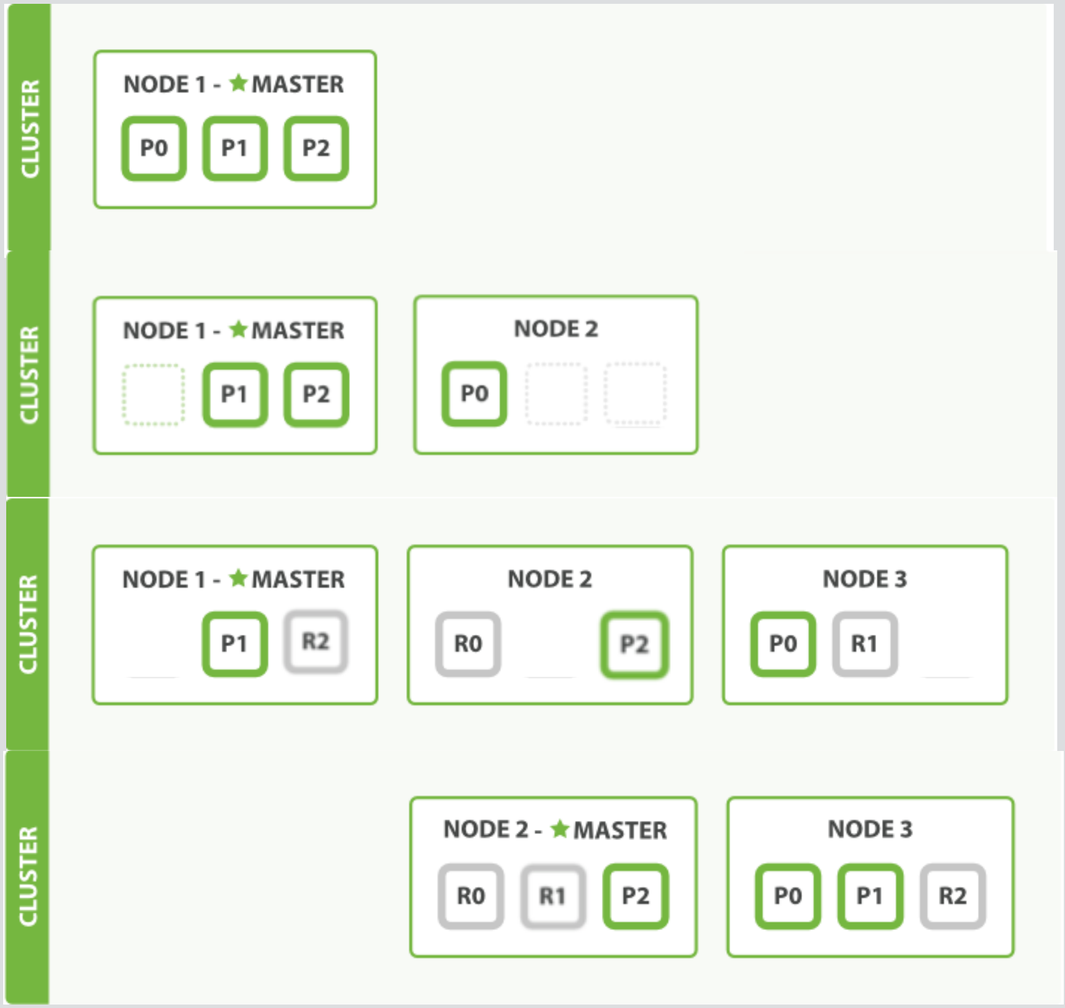
Elasticsearch 의 용어 및 개념 정리

샤드(Shard)

- 싱글 노드에 3개의 샤드로 클러스터 구성
- 시간이 지날수록, 문서는 점점 늘어나고,
  결국에는 싱글노드가 허용하는 볼륨을 모두 소진
- ElasticSearch 클러스터에 더 이상 데이터를 적재불가

- 클러스터에 동일한 설정의 노드 증설
- 클러스터가 일정 샤드들을 새로 투입된 노드로 분배
- 클러스터 용량 및 요청에 응답을 줄 수 있는 노드가 늘어남
- 노드가 한 대 Fail 이 날 경우 데이터의 안정성을 보장 불가

- 복제본인 Replica Shard 추가
- 노드 한 대 Fail 시 남아있는 나머지 노드들의
  Replica Shard 가 Primary Shard 로 승격됨
- Fail 된 노드의 Replica Shard 는 남아있는 나머지 노드들의
  Primary Shard 를 복제하여 Replica Shard 를 재구성

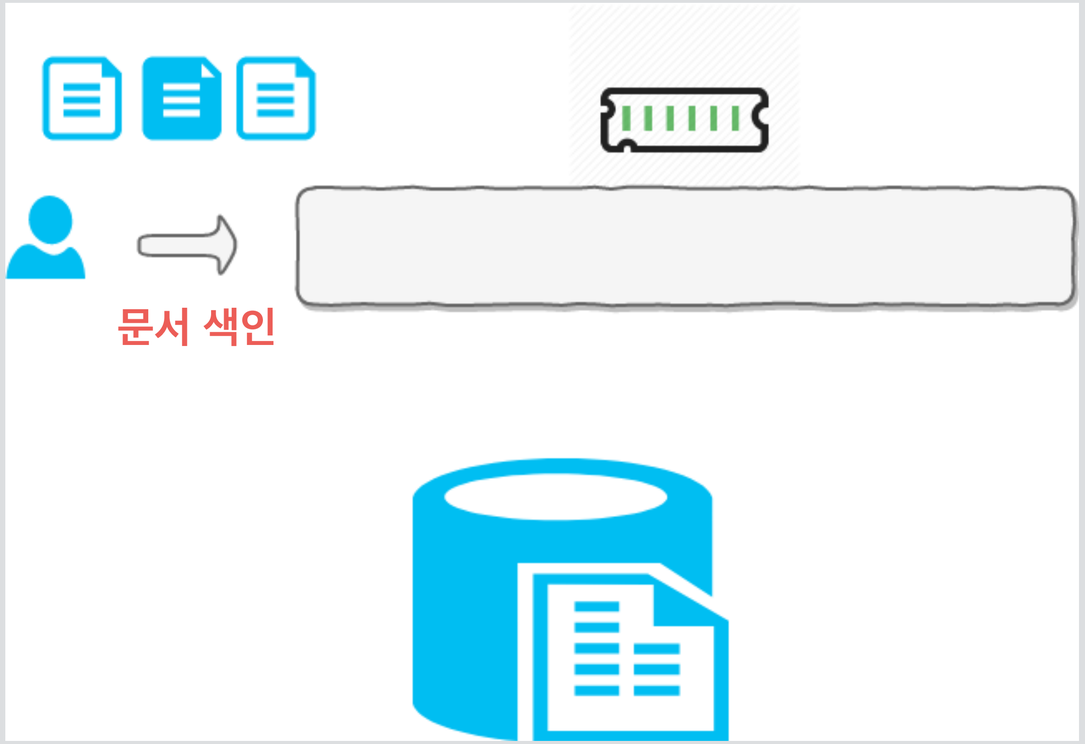


Fast campus

Elasticsearch 의 용어 및 개념 정리

세그먼트(Segment)

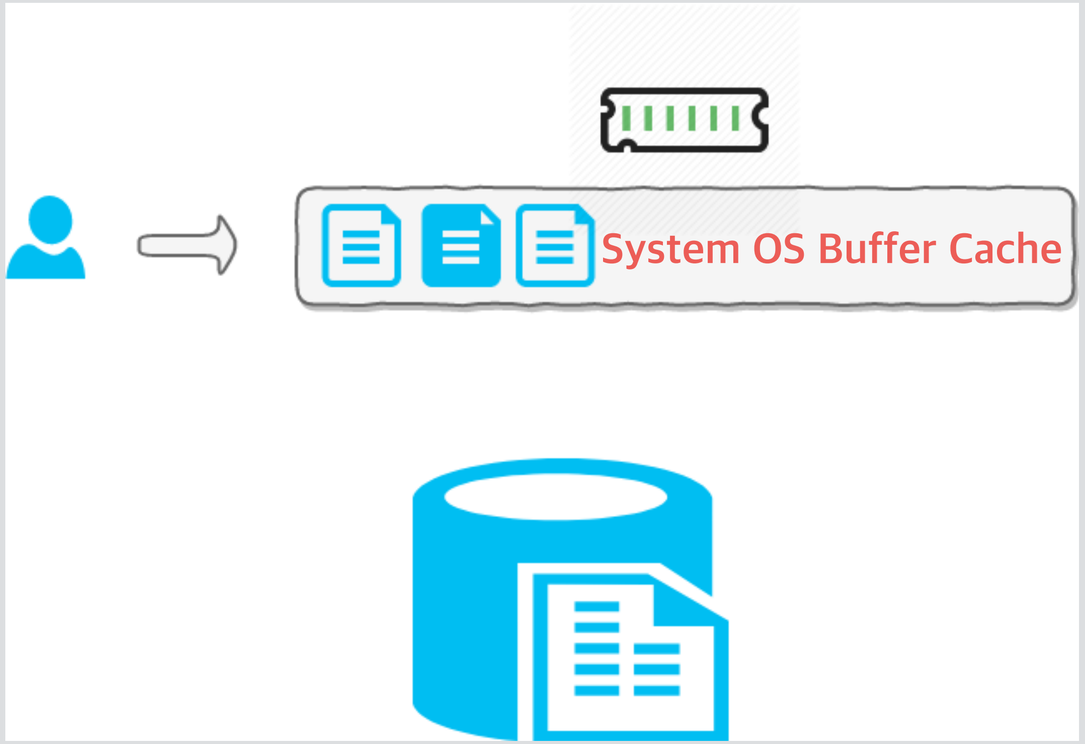
- Shard 는 다시 Segment 들로 구성
- 문서가 저장되는 최소 단위



Elasticsearch 의 용어 및 개념 정리

세그먼트(Segment)

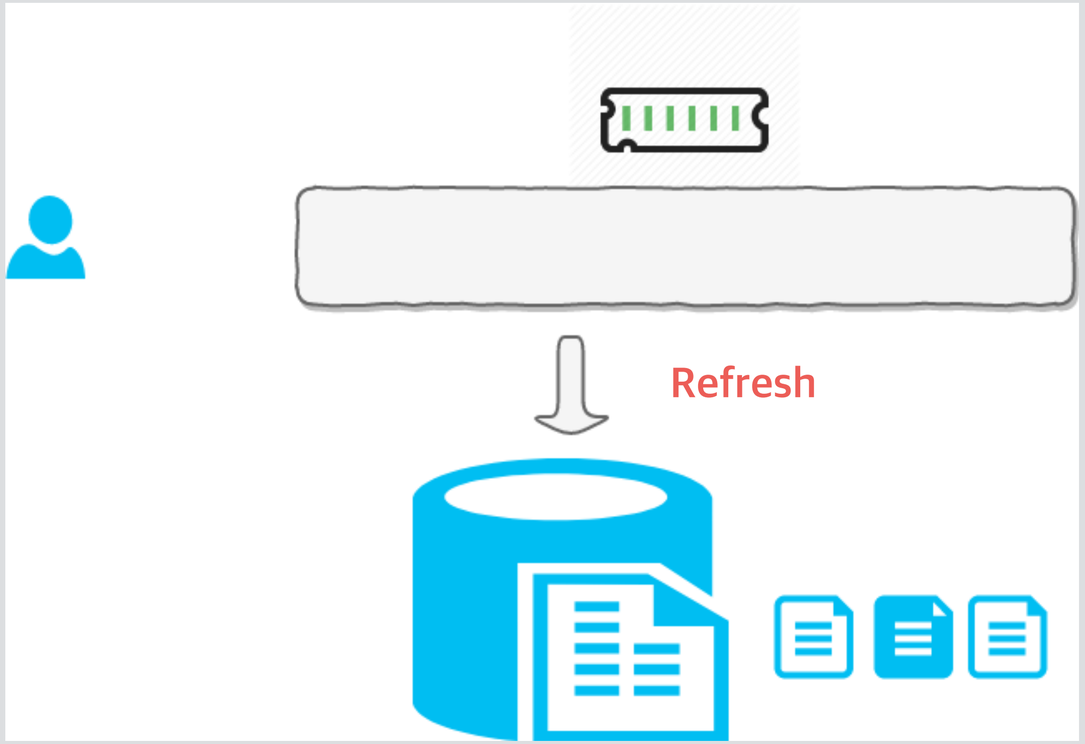
- 문서는 먼저 System OS Buffer Cache 영역에 저장



Elasticsearch 의 용어 및 개념 정리

세그먼트(Segment)

- Refresh 과정을 통해 문서를 디스크에 저장
- 저장된 데이터는 곧바로 검색 가능한 상태가 됨
- 불변의(immutable) 특성 - update 불가
- data lock 을 걸 필요 없이 일관성을 유지



- 6.x 버전까지는 기본 1초마다 refresh, 7.x 버전부터는 30초동안 검색 요청이 없으면 refresh disable

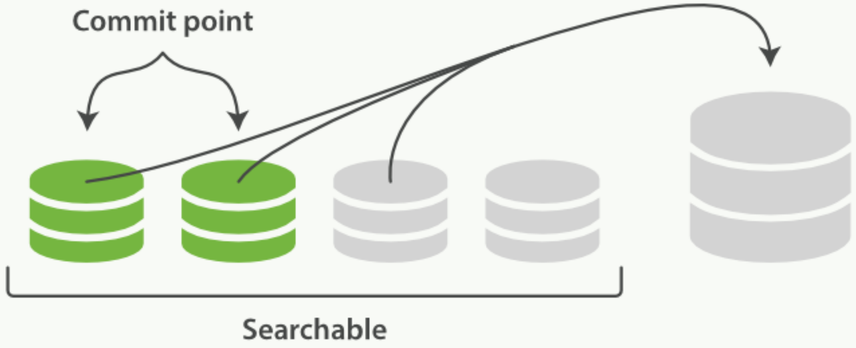
Elasticsearch 의 용어 및 개념 정리

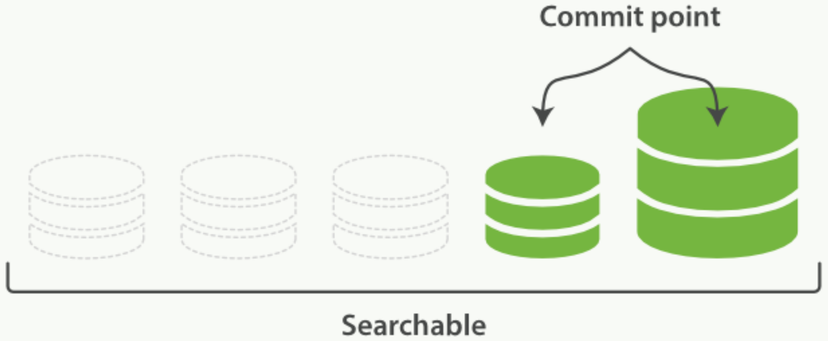
세그먼트(Segment)

- 잘게 쪼개진 데이터는 단일 요청에 의해 많은 Segment 들이 응답해야 한다는 단점

- 백그라운드에서 Segment 병합(Merging)을 진행

- 하나 혹은 그 이상의 적은 수의 Segment 로 병합된 데이터가 저비용으로 빠른 응답을 줄 수 있음





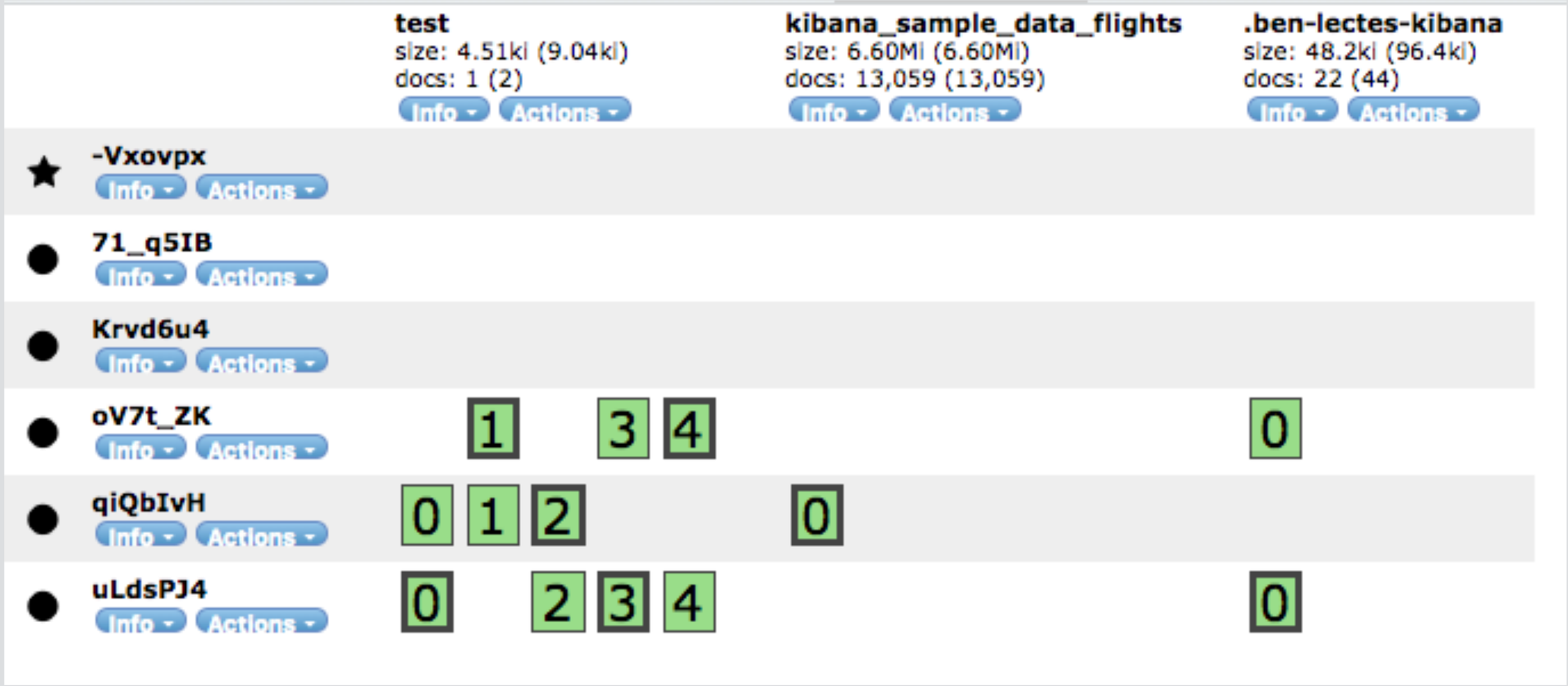
Elasticsearch 의 용어 및 개념 정리

세그먼트(Segment)

Fast campus

Elasticsearch 의 용어 및 개념 정리

ElasticSearch Cluster



Elasticsearch 설치

Elasticsearch 설치

Elasticsearch 는 Java 언어로 이루어진 정보 검색 라이브러리인 Lucene 기반으로 구성

6.x 버전까지는 Java 를 설치 필수, 7.x 버전부터는 java 내재화

Elasticsearch JVM Support Metrics
https://www.elastic.co/support/matrix#matrix_jvm

OS Support Metrics
https://www.elastic.co/support/matrix

Elasticsearch 설치

YUM Repository 등록으로 설치하기

- /etc/yum.repos.d/elasticsearch.repo 등록

```
[elasticsearch-7.x]
name=Elasticsearch repository for 7.x packages
baseurl=https://artifacts.elastic.co/packages/7.x/yum
gpgcheck=1
gpgkey=https://artifacts.elastic.co/GPG-KEY-elasticsearch
enabled=1
autorefresh=1
type=rpm-md
```

- sudo yum install elasticsearch

Elasticsearch 설치

RPM Download 하여 설치하기

- Elasticsearch RPM Download 후 설치
- elasticsearch user, group 자동생성

```
sudo yum -y install wget
wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.3.0-x86_64.rpm
sudo rpm -ivh elasticsearch-7.3.0-x86_64.rpm
```

Elasticsearch 설치

RPM Download 하여 설치하기

- elasticsearch user, group 자동생성
- /etc/elasticsearch 에 config 폴더

elasticsearch.keystore elasticsearch.yml jvm.options log4j2.properties role_mapping.yml roles.yml
users users_roles

- /etc/sysconfig 에 elastic search 환경변수 파일

- /usr/share/elasticsearch 밑에 그 외 파일들

LICENSE.txt NOTICE.txt README.textile bin lib modules plugins

Elasticsearch 설치

zip, tar Download 하여 설치하기

- Elasticsearch zip, tar Download 후 설치(non root 계정)

```
wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.3.0.zip
unzip elasticsearch-7.3.0.zip

wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.3.0.tar.gz
tar -xzf elasticsearch-7.3.0.tar.gz
```

Elasticsearch 설치

zip, tar Download 하여 설치하기

- root 가 아닌 일반 계정으로만 설치 가능
- elasticsearch 를 설치할 폴더 아래 압축 해제

bin config data lib LICENSE.txt logs modules NOTICE.txt plugins README.textile

- rpm 설치와 비교했을 때 config, data 추가로 생성됨
- config 폴더 아래 파일은 rpm 설치 시에 /etc/elasticsearch 밑의 파일

Elasticsearch 설치

Elasticsearch 실행하기

- RPM / YUM 설치

OS 버전에 따라

init : sudo service elasticsearch start
systemd : sudo systemctl start elasticsearch.service

- 소스 설치(non root)

bin/elasticsearch -d

Elasticsearch 설치

Elasticsearch 실행 확인하기

- 프로세스 확인

```
ps -ef | grep elasticsearch
```

- 어플리케이션 반응 확인

```
curl localhost:9200
```

- Elasticsearch 어플리케이션이 시작되지 않았을 때

rpm 설치 - /var/log/elasticsearch/elasticsearch.log 확인
소스 설치 - {install path}/logs/elasticsearch.log 확인

Elasticsearch 설치

Kibana Dev Tools 활용하기

- Kibana 의 기능 중 쿼리를 날릴 수 있는 Devtools 가 존재

YUM 으로 설치하기

```
[kibana-7.x]
name=Kibana repository for 7.x packages
baseurl=https://artifacts.elastic.co/packages/7.x/yum
gpgcheck=1
gpgkey=https://artifacts.elastic.co/GPG-KEY-elasticsearch
enabled=1
autorefresh=1
type=rpm-md
```

```
sudo yum install kibana
```

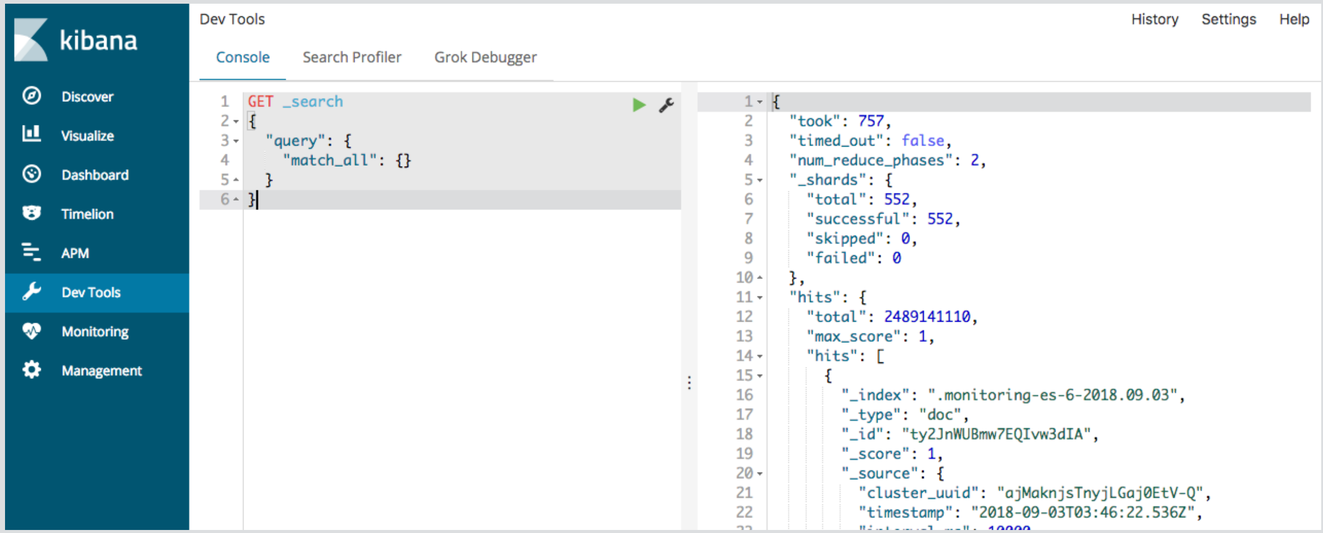
Elasticsearch 설치

Kibana Dev Tools 활용하기

- /etc/kibana/kibana.yml

```
server.host: "0.0.0.0"
elasticsearch.hosts: "http://{ES_URL}:9200"
kibana.index: ".kibana"
```

- 6.x 버전에서는 elasticsearch.url 로 설정



실습

Elasticsearch & Kibana Install
https://github.com/benjamin-btn/ES7-Tutorial/tree/master/ES-Tutorial-1

Q & A