if(kakao)2021

대량의 스트림 데이터를 실시간으로 분류하기

Elasticsearch Percolator를 이용한 콘텐츠 분류

이규열 Rick.Lee
카카오

왜 스트림 데이터를 실시간으로 분류할까?

기존에 데이터를 분류하는 방식

기존 분류 방식의 문제점

Percolator는 어떻게 동작할까?

Percolator가 기존 분류 방식의 문제점을 해결하는 방법

정리

왜 스트림 데이터를 실시간으로 분류할까?

23:19

부산 25°C • 울산 24°C 현위치

홈&쿠킹 동물 스타일 자동차+ 여행맛집

세포라 입점 기념 특별 할인

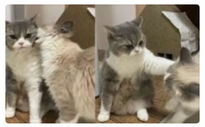
어글리시크, CH!C한 1+1 한정 특가

권은비 반려견 방송태도 논란

놀라운 아기 강아지의 점프력

우리중에 간첩이 있는것 같덕..

산책하던 중 어디선가 소시지 한 봉지를 가져온 강…

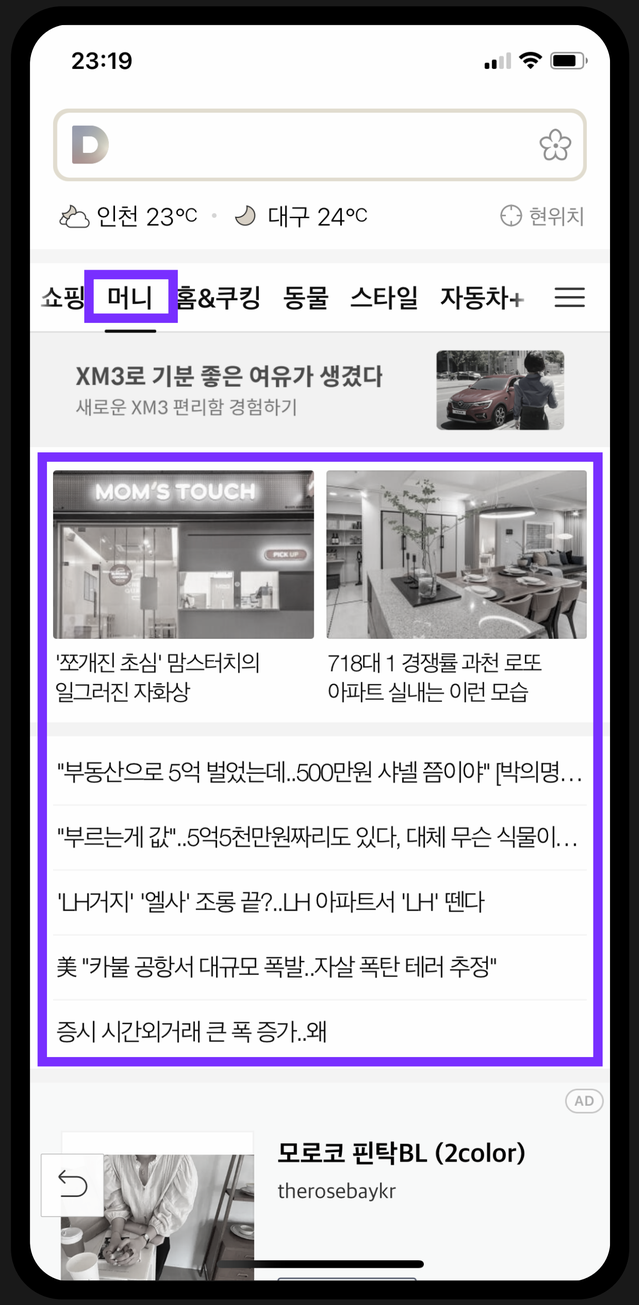


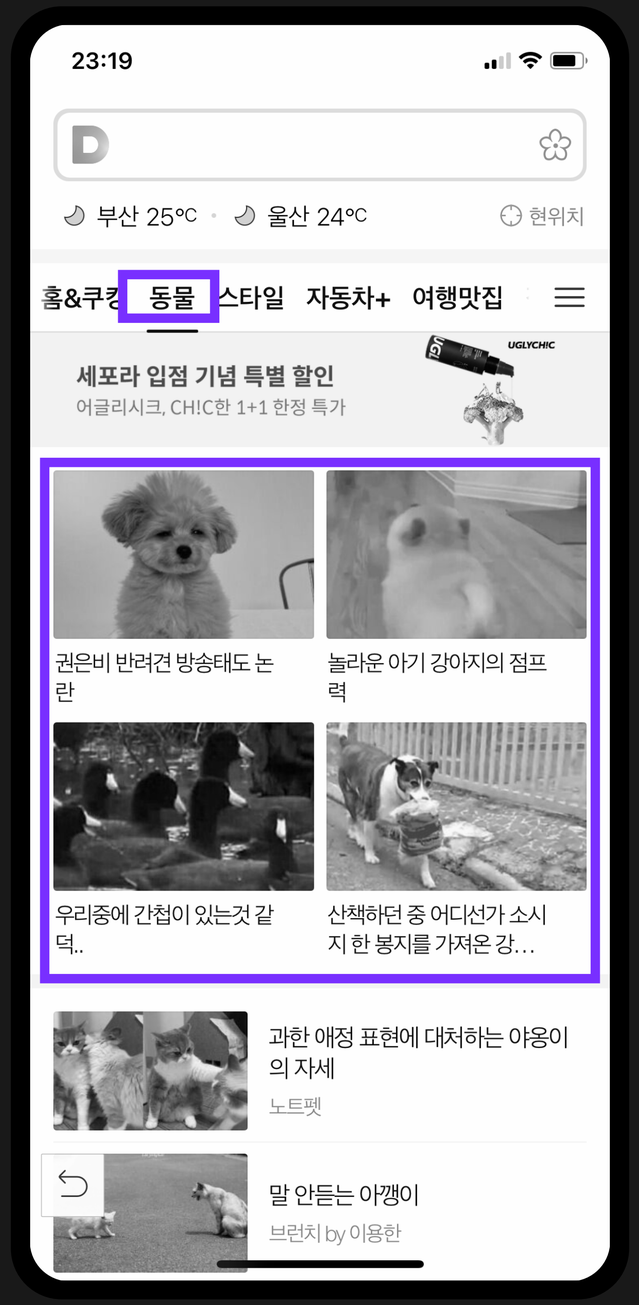
과한 애정 표현에 대처하는 야옹이의 자세

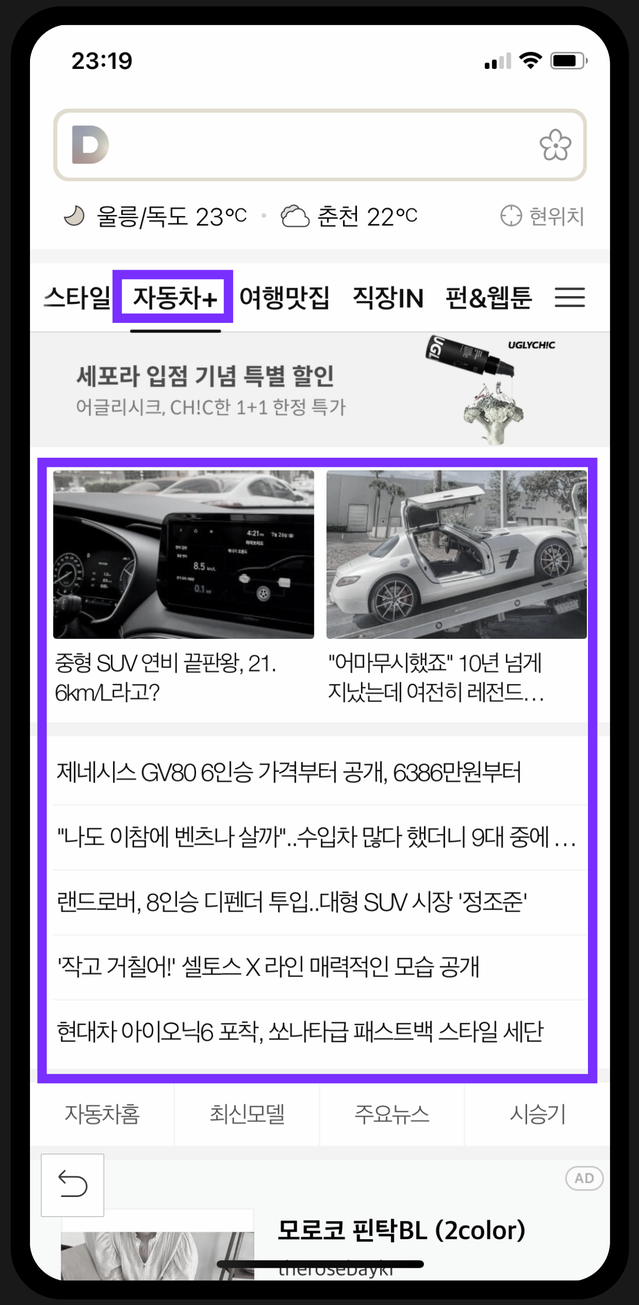
노트펫

말 안듣는 아깽이

브런치 by 이용한


23:19
인천 23℃ · 대구 24℃ 현위치
쇼핑 머니 홈&쿠킹 동물 스타일 자동차+
XM3로 기분 좋은 여유가 생겼다
새로운 XM3 편리함 경험하기
MOM'S TOUCH
'쪼개진 초심' 맘스터치의 일그러진 자화상
718대 1 경쟁률 과천 로또 아파트 실내는 이런 모습
"부동산으로 5억 벌었는데..500만원 샤넬 쯤이야" [박의명…
"부르는게 값"..5억5천만원짜리도 있다, 대체 무슨 식물이…
'LH거지' '엘사' 조롱 끝?..LH 아파트서 'LH' 뗀다
美 "카불 공항서 대규모 폭발..자살 폭탄 테러 추정"
증시 시간외거래 큰 폭 증가..왜
AD
모로코 핀탁BL (2color)
therosebaykr


23:19
부산 25℃ · 울산 24℃ 현위치
홈&쿠킹 동물 스타일 자동차+ 여행맛집
세포라 입점 기념 특별 할인
어글리시크, CH!C한 1+1 한정 특가
UGLYCHIC
권은비 반려견 방송태도 논란
놀라운 아기 강아지의 점프력
우리중에 간첩이 있는것 같덕..
산책하던 중 어디선가 소시지 한 봉지를 가져온 강…
과한 애정 표현에 대처하는 야옹이의 자세
노트펫
말 안듣는 아깽이
브런치 by 이용한


23:19
울릉/독도 23℃ · 춘천 22℃ 현위치
스타일 자동차+ 여행맛집 직장IN 펀&웹툰
세포라 입점 기념 특별 할인
어글리시크, CH!C한 1+1 한정 특가
UGLYCHIC
중형 SUV 연비 끝판왕, 21.6km/L라고?
"어마무시했죠" 10년 넘게 지났는데 여전히 레전드…
제네시스 GV80 6인승 가격부터 공개, 6386만원부터
"나도 이참에 벤츠나 살까"..수입차 많다 했더니 9대 중에 …
랜드로버, 8인승 디펜더 투입..대형 SUV 시장 '정조준'
'작고 거칠어!' 셀토스 X 라인 매력적인 모습 공개
현대차 아이오닉6 포착, 쏘나타급 패스트백 스타일 세단
자동차홈 최신모델 주요뉴스 시승기
AD
모로코 핀탁BL (2color)

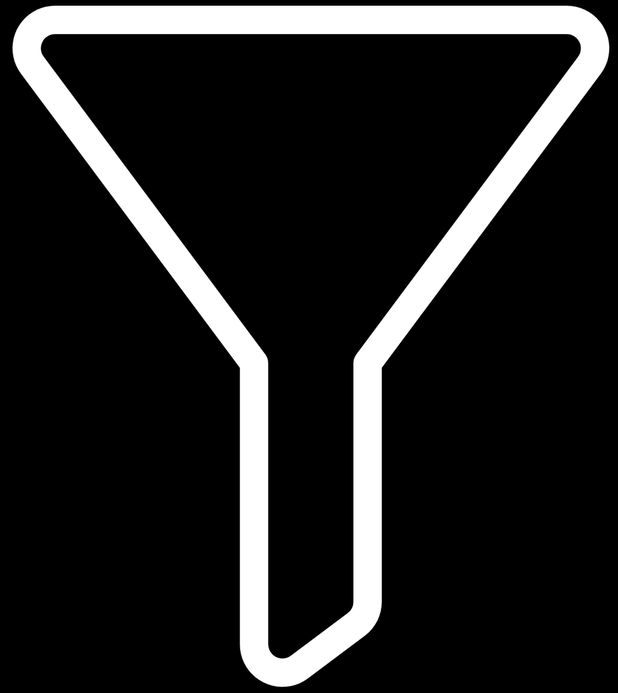
필터를 이용해 콘텐츠를 분류

- 콘텐츠 정보를 조건으로 설정
- 예) 동물 이미지 필터

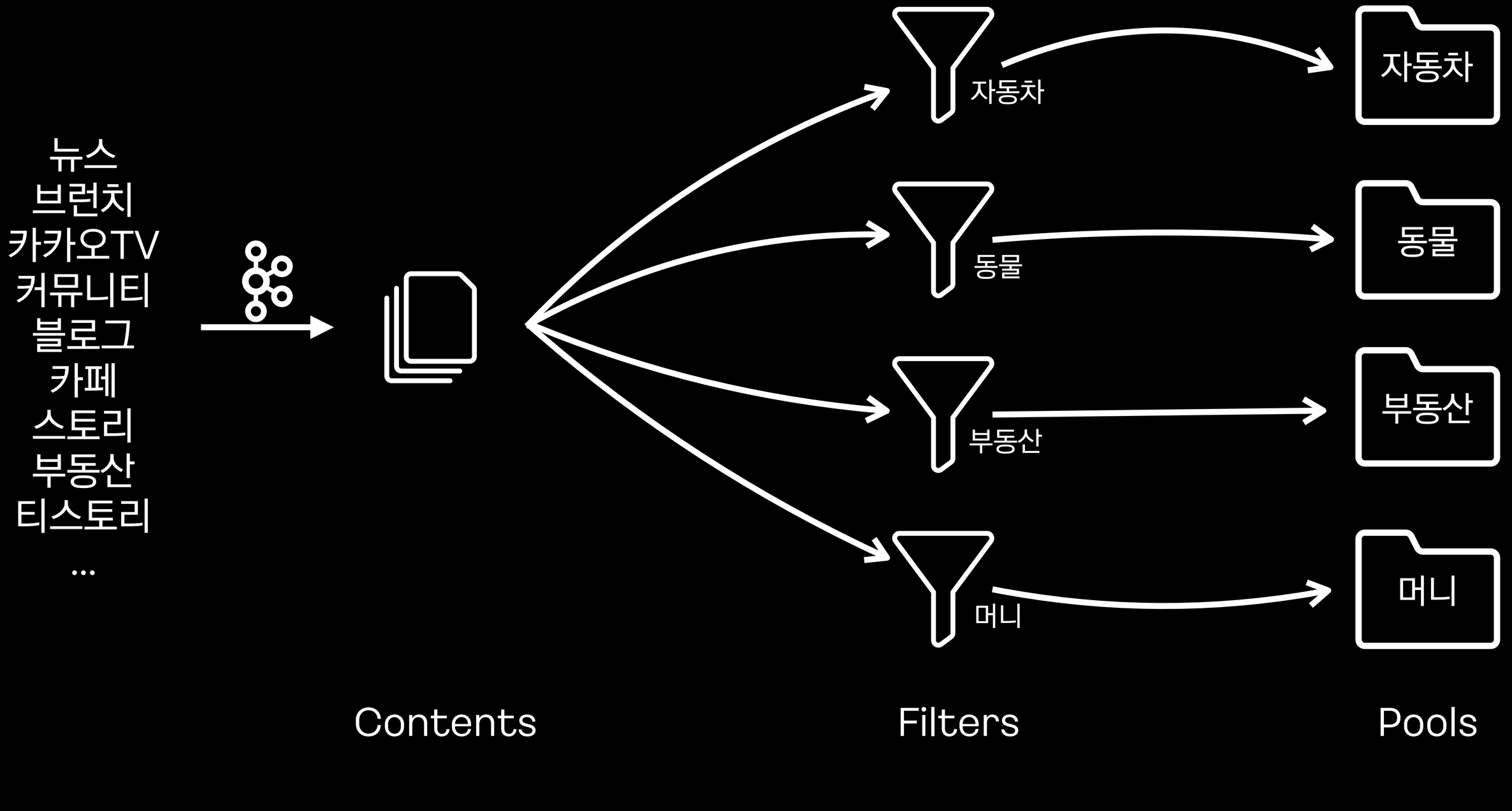
  제목: 강아지, 멍멍이, 고양이, 냥이 키워드 포함

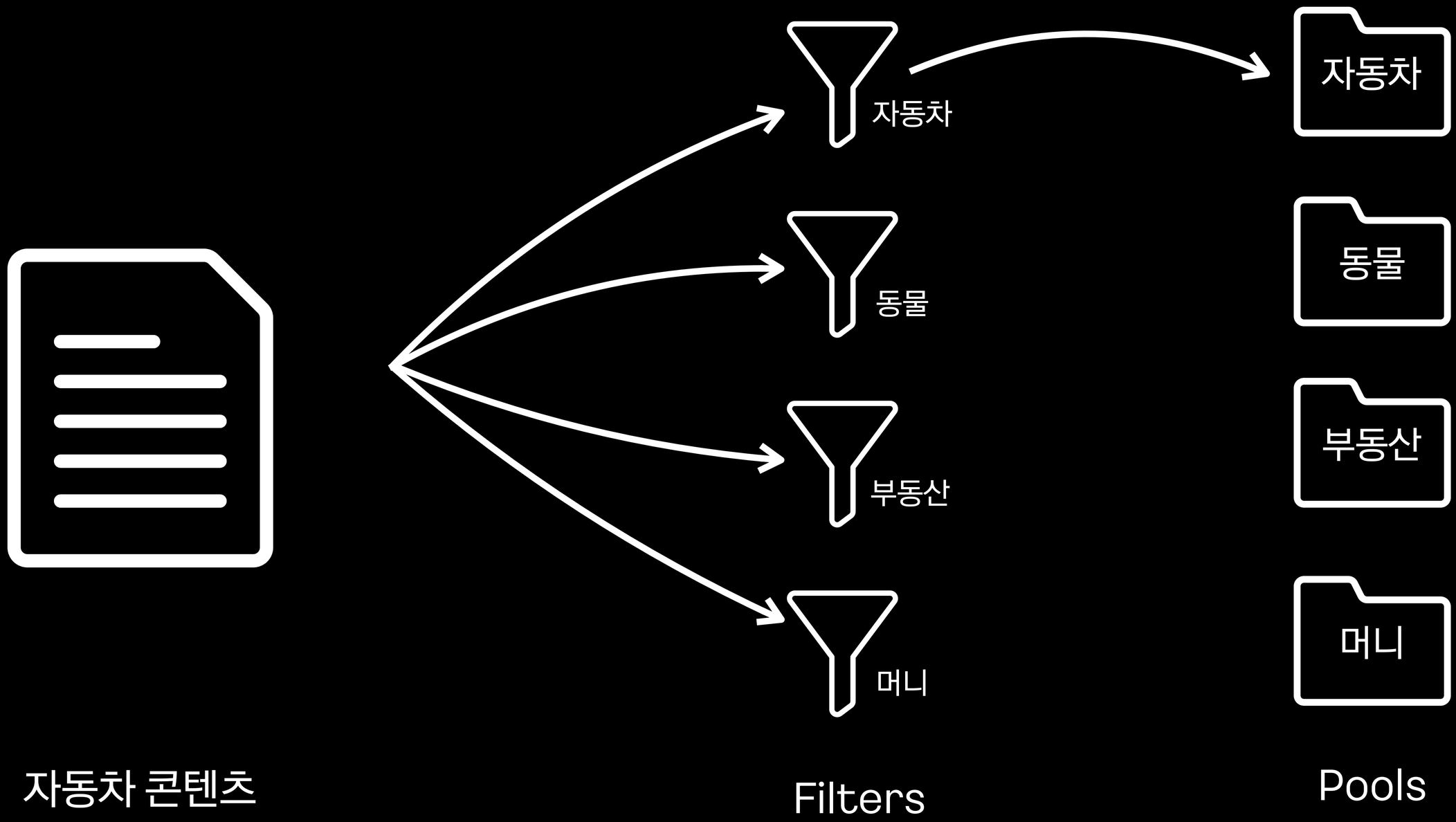
  콘텐츠 타입: 이미지

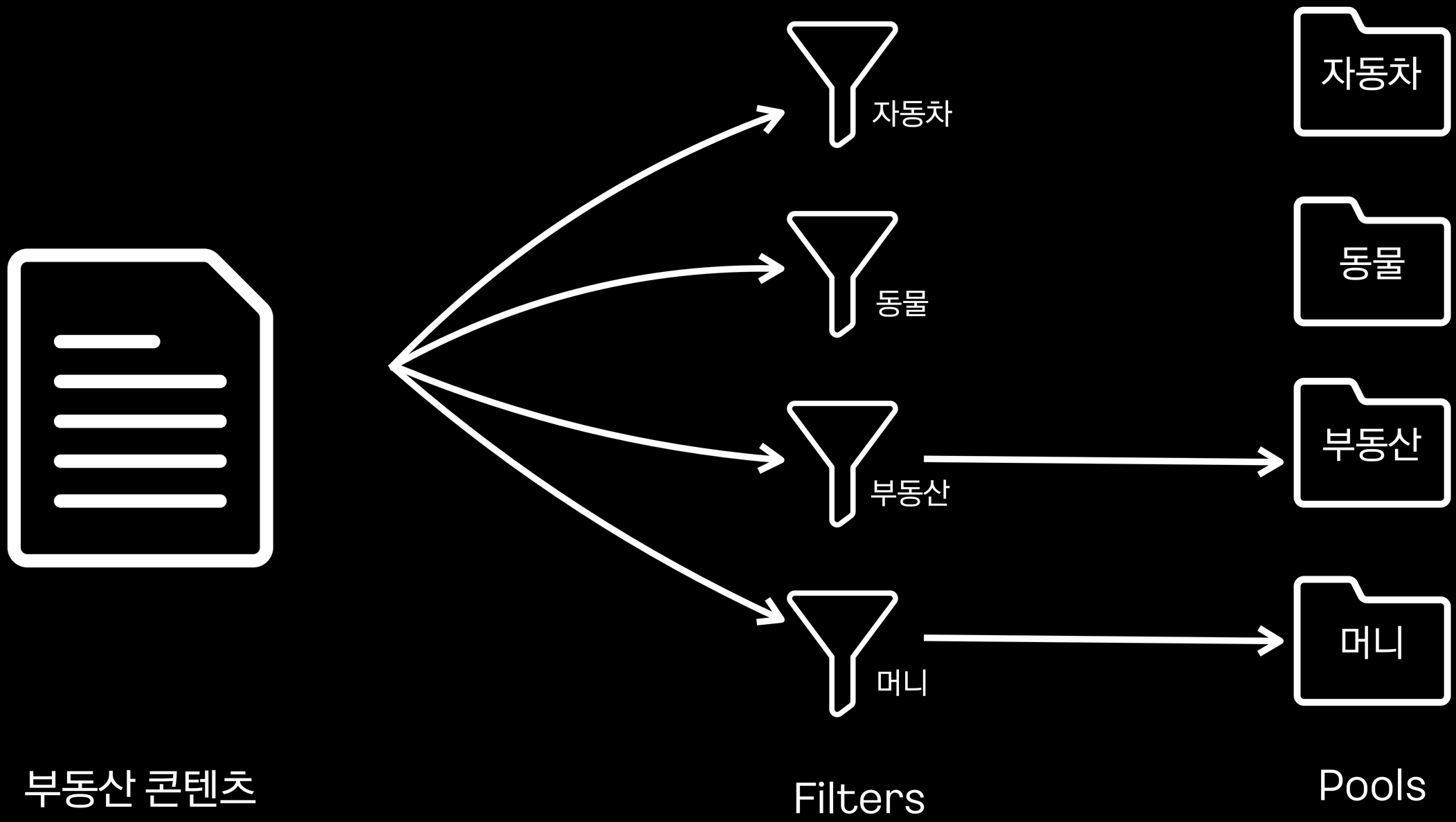
  이미지 수: 2개 이상



동물 이미지 필터


뉴스
브런치
카카오TV
커뮤니티
블로그
카페
스토리
부동산
티스토리
...
자동차
동물
부동산
머니
자동차
동물
부동산
머니
Contents
Filters
Pools


자동차
동물
부동산
머니
자동차
동물
부동산
머니
자동차 콘텐츠
Filters
Pools

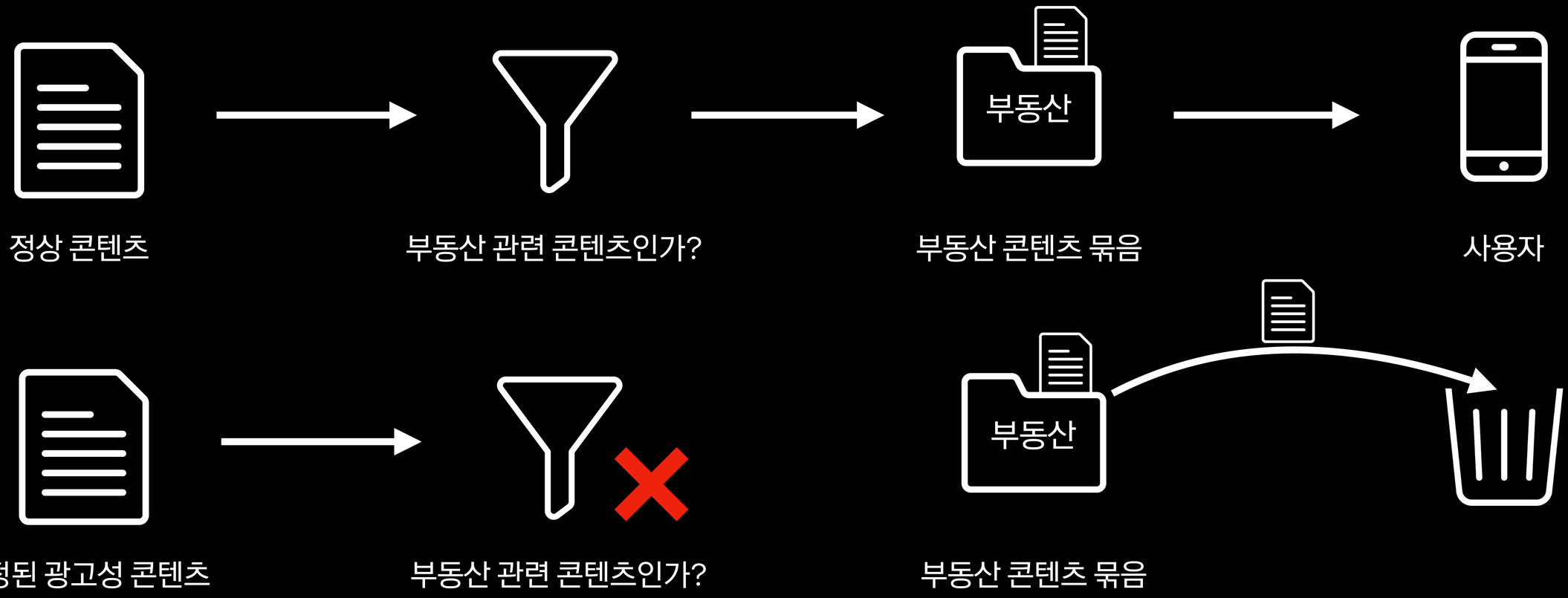

자동차
동물
부동산
머니
자동차
동물
부동산
머니
부동산 콘텐츠
Filters
Pools

왜 스트림 데이터를 실시간으로 분류할까?

- 뉴스의 경우 빠른 노출이 중요

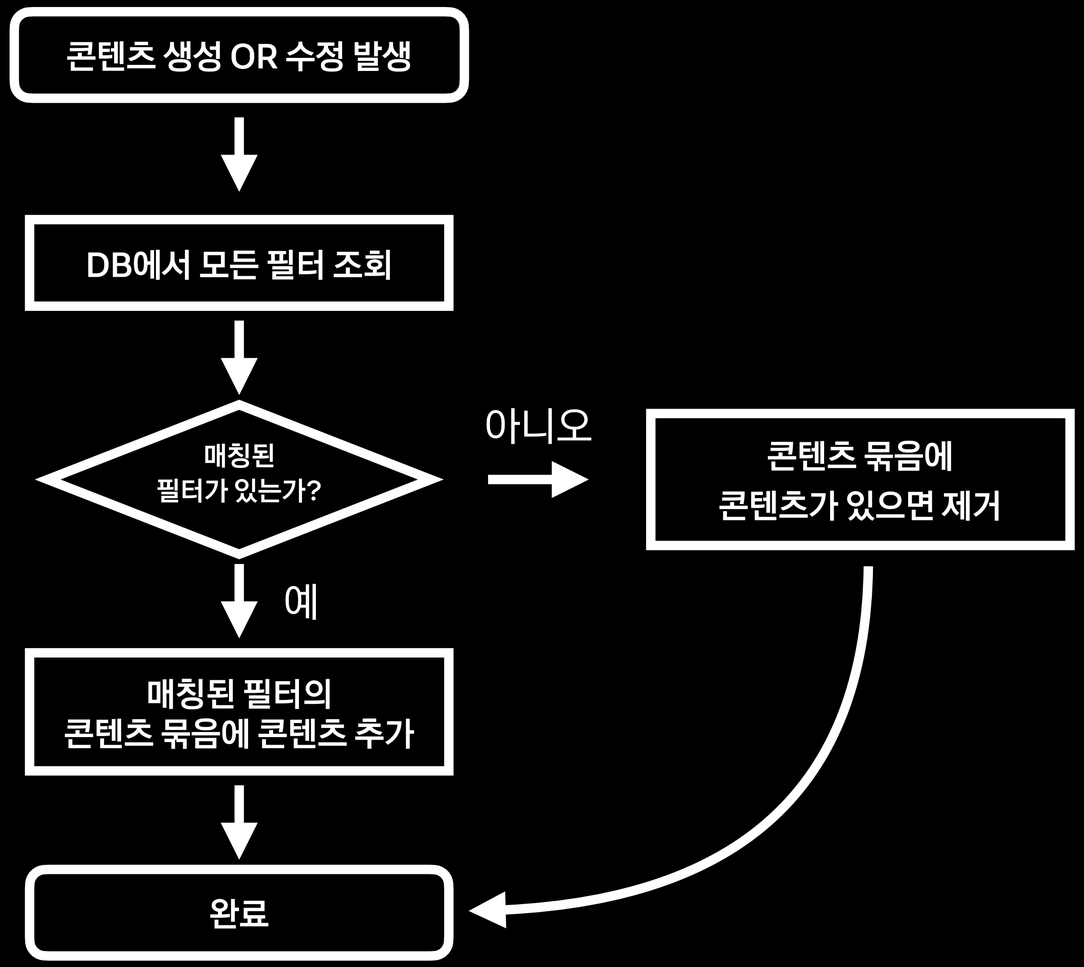
왜 스트림 데이터를 실시간으로 분류할까?

- 콘텐츠 수정 OR 삭제 시 재분류 필요



기존에 데이터를 분류하는 방식

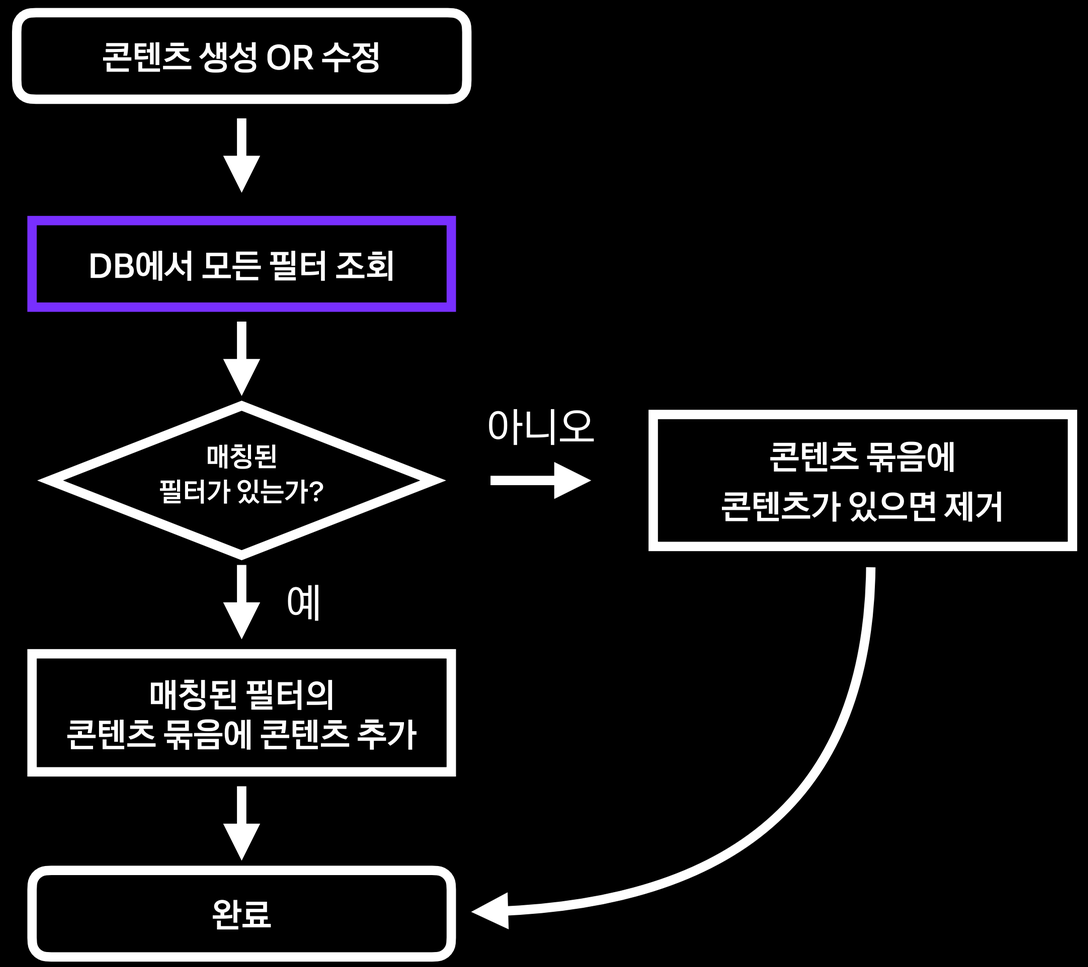
순서도


콘텐츠 생성 OR 수정 발생
DB에서 모든 필터 조회
매칭된
필터가 있는가?
아니오
콘텐츠 묶음에
콘텐츠가 있으면 제거
예
매칭된 필터의
콘텐츠 묶음에 콘텐츠 추가
완료

기존 분류 방식의 문제점

🤦 분류 필터를 DB에서 매번 조회

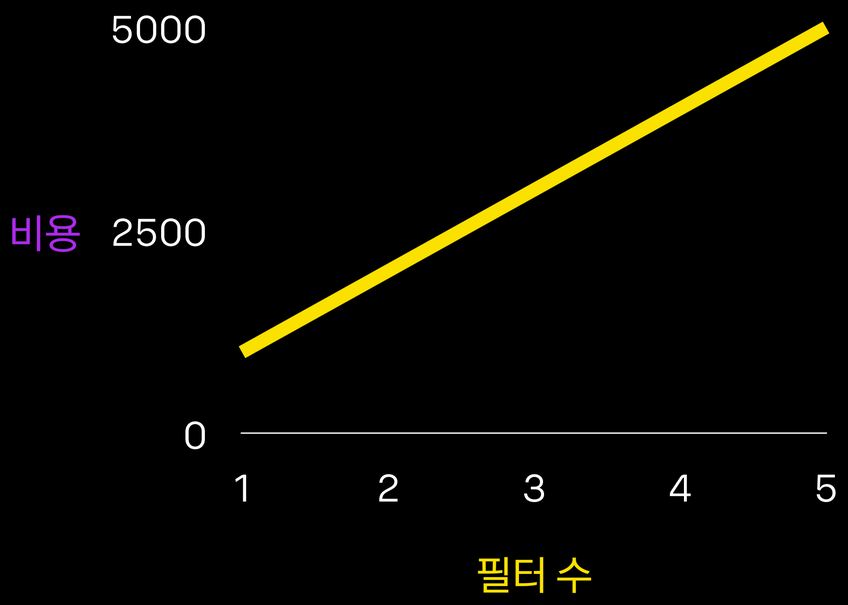
- 최신 필터 정보를 위해 매번 조회 필요
- 필터가 많아지면 조회 비용 증가
- 캐싱과 동기화 로직 필요

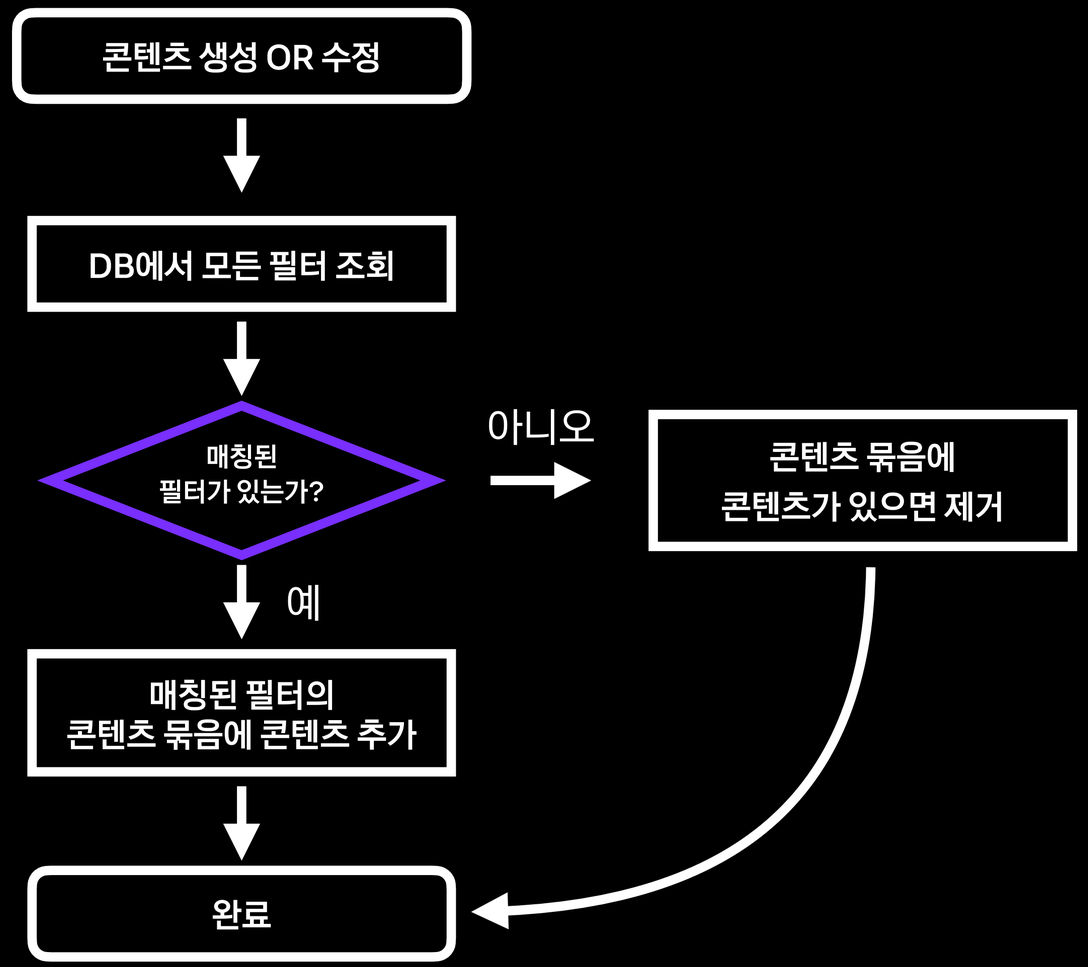


🤦 콘텐츠가 매칭된 필터 찾기

- 반복문으로 모든 조건을 비교
- 필터 추가시 계산 비용이 크게 증가

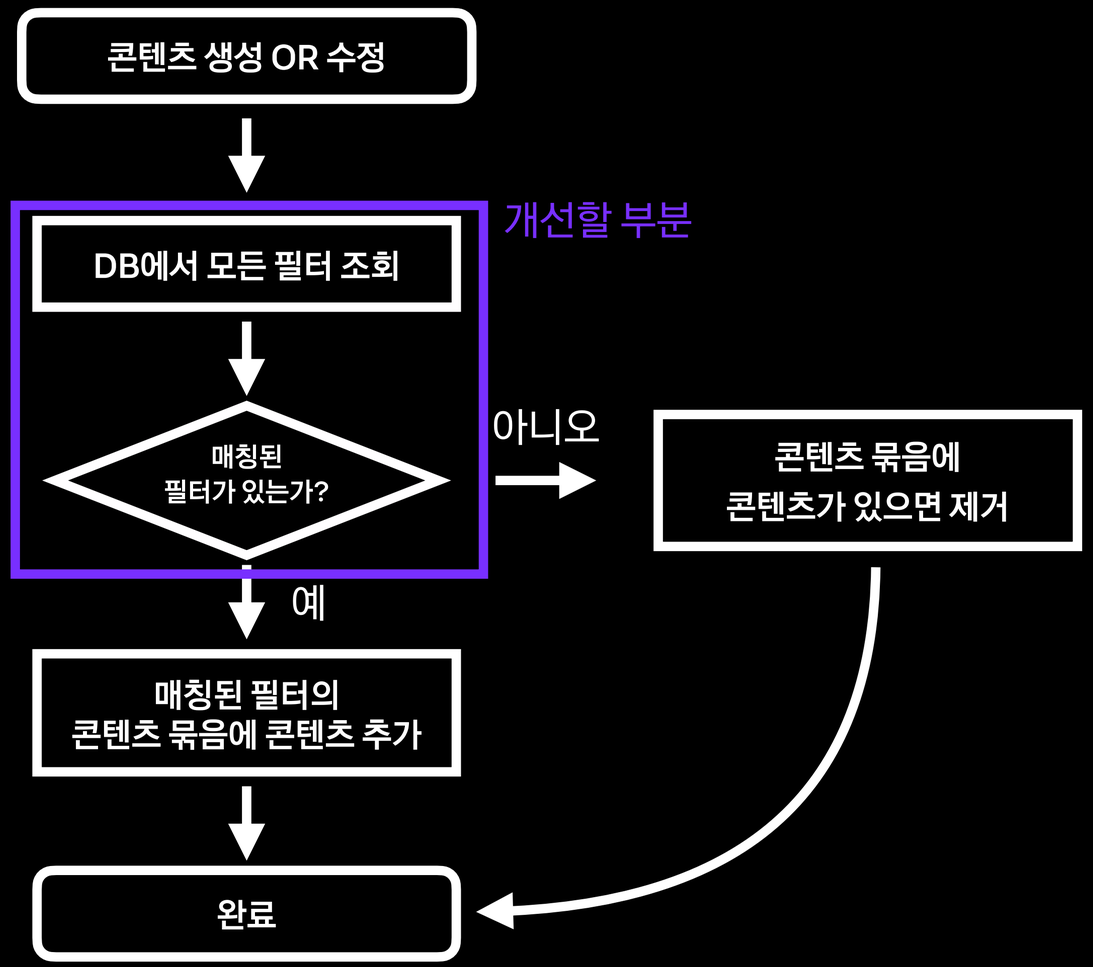
  비용 = 필터 수 X 콘텐츠 수





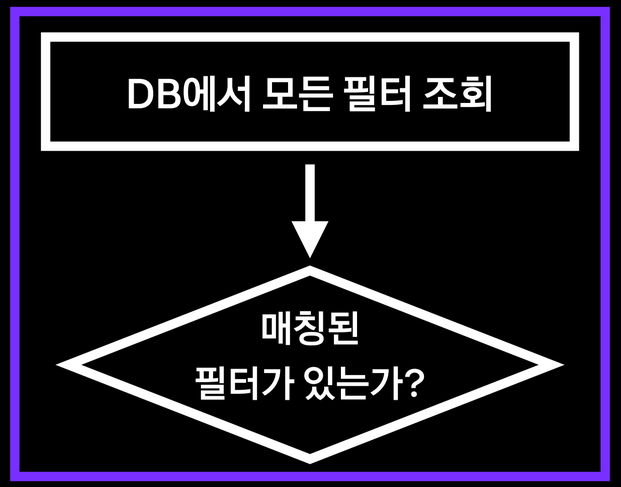
🤦 콘텐츠와 필터는 계속 증가한다

- 필터 정보 캐싱과 동기화 기능 필요
- Scale-out이 용이한 구조로 변경
- 효율적인 매칭 알고리즘으로 검색 속도 향상



Elasticsearch Percolator

콘텐츠 필터 매칭 로직을 Percolator로 전환



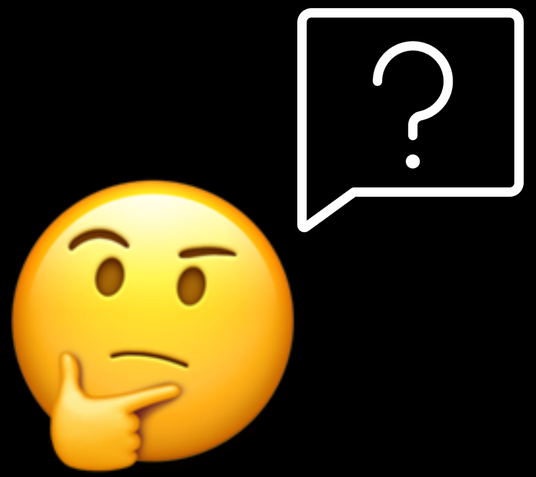
전환

Elasticsearch
Percolator

Percolator는 어떻게 동작할까?

Percolator란?

쿼리를 등록해두고 도큐먼트를 담은 퍼컬레이트 요청을
보내 매칭된 쿼리를 반환해주는 Elasticsearch의 기능



percolator

1. 퍼컬레이터 2. 여과기 3. 여과하는 사람

---

coffee percolator

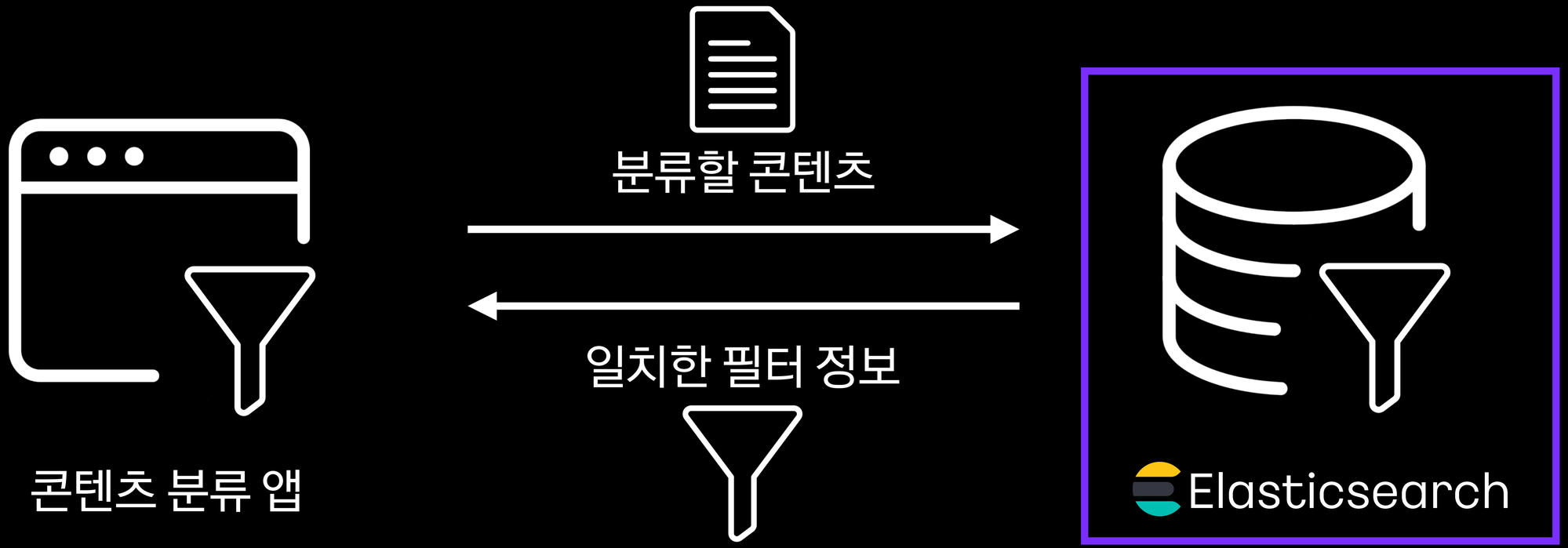
(여과식) 커피 끓이개

Percolator의 동작 구조 - 쿼리 등록

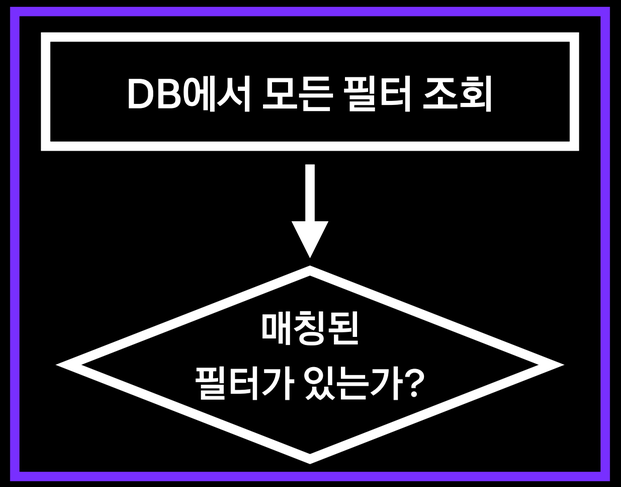
필터 관리 앱

Elasticsearch

Percolator의 동작 구조 - 매칭 쿼리 요청



콘텐츠 필터 매칭 로직을 Percolator로 전환



전환

Elasticsearch
Percolator

Percolator Index 생성

- 필터 조건으로 사용될 필드 정의
- 쿼리가 저장될 필드 지정

```
// PUT /pool-filters
{
  "mappings": {
    "properties": {
      "title": {
        "type": "text"
      },
      "category": {
        "type": "keyword"
      },
      "bodyImageCount": {
        "type": "integer"
      },
      "query": {
        "type": "percolator"
      }
    }
  }
}
```

Percolator Query 등록

- 라이언과 어피치 필터 등록

  타이틀에 "라이언" 또는 "어피치"가 포함

  카테고리가 "연예" 또는 "경제"

  본문 이미지 수가 2개 이상

```
// PUT /pool-filters/_doc/1?refresh
{
  "query": {
    "bool": {
      "must": [
        {
          "bool": {
            "should": [
              {
                "match": {
                  "title": "라이언"
                }
              },
              {
                "match": {
                  "title": "어피치"
                }
              }
            ]
          }
        },
        {
          "terms": {
            "category": ["연예", "경제"]
          }
        },
        {
          "range": {
            "bodyImageCount": {
              "from": 2
            }
          }
        }
      }
```

Percolator Query 검색

- 조건에 일치하는 쿼리를 찾기 위한 콘텐츠 전송

  타이틀에 "라이언" 또는 "어피치"가 포함

  카테고리가 "연예" 또는 "경제"

  본문 이미지 수가 2개 이상

```
// GET /pool-filters/_search
{
  "query": {
    "percolate": {
      "field": "query",
      "document": {
        "title": "라이언 신규 상품 출시!!",
        "cateogry": "경제",
        "bodyImageCount": 3
      }
    }
  }
}
```

Percolator Query 검색

- 조건에 일치하는 쿼리를 찾기 위한 콘텐츠 전송

  타이틀에 "라이언" 또는 "어피치"가 포함

  카테고리가 "연예" 또는 "경제"

  본문 이미지 수가 2개 이상

```
// GET /pool-filters/_search
{
  "query": {
    "percolate": {
      "field": "query",
      "document": {
        "title": "어피치 광고 촬영 현장!",
        "cateogry": "경제",
        "bodyImageCount": 1
      }
    }
  }
}
```

Percolator Query 결과 분석

took : 걸린 시간 (ms)

hits.total.value: 매칭된 쿼리 수

hits.hits: 매칭된 쿼리 목록

Percolator가 기존 분류 방식의 문제점을 해결하는 방법

필터 정보 캐싱과 동기화

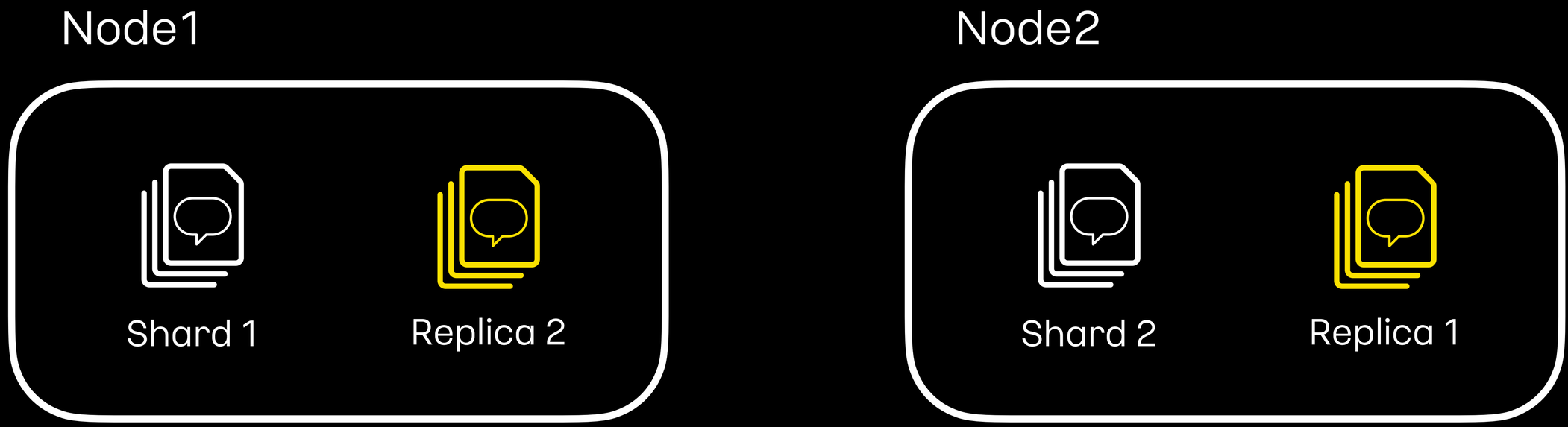
- Elasticsearch 가 쿼리(필터)를 관리하므로 캐싱 및 동기화를 신경쓰지 않아도 된다
- 각 노드 자동으로 동기화 되고 힙 메모리에 캐싱됨

Scale-out이 쉽게 가능한 구조로 변경 1/2

- Document를 여러 샤드(shard)에 분산 저장
- 샤드 당 스레드(thread) 하나가 할당
- 각 스레드에서 처리하는 Document 수가 감소하여 검색 속도 향상

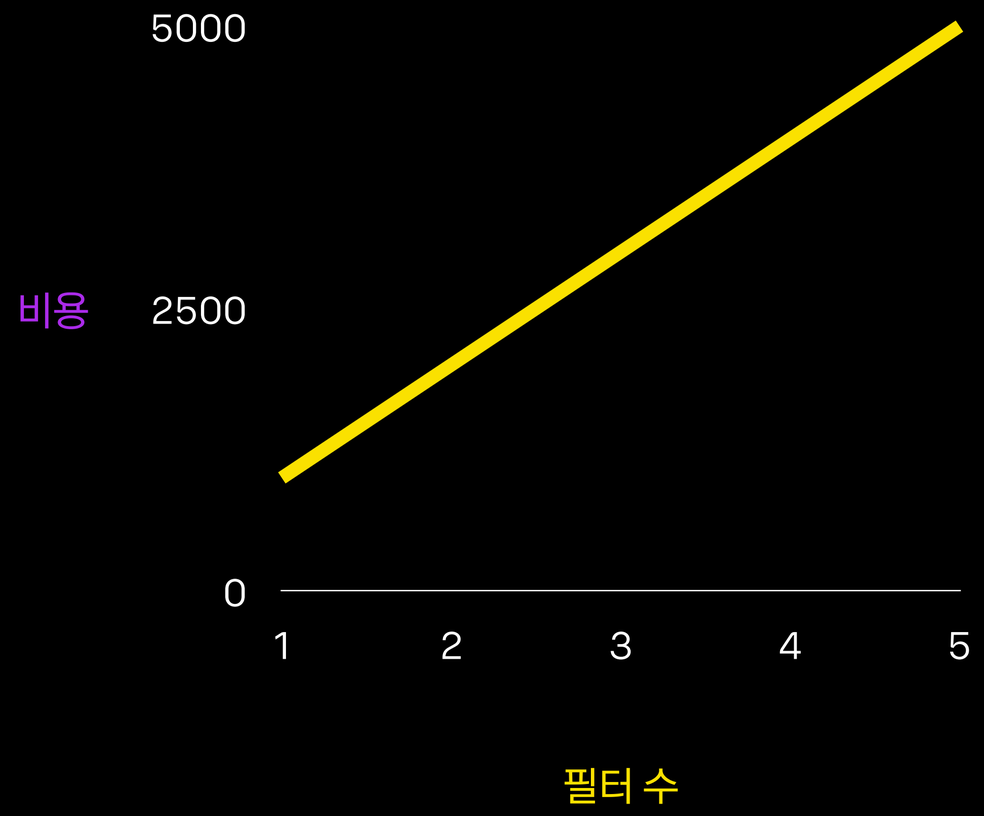
Scale-out이 쉽게 가능한 구조로 변경 2/2

- Replica 를 여러개 생성해서 전체 처리량(throughput)을 증가시킬 수 있다
- 스레드 4개 활용



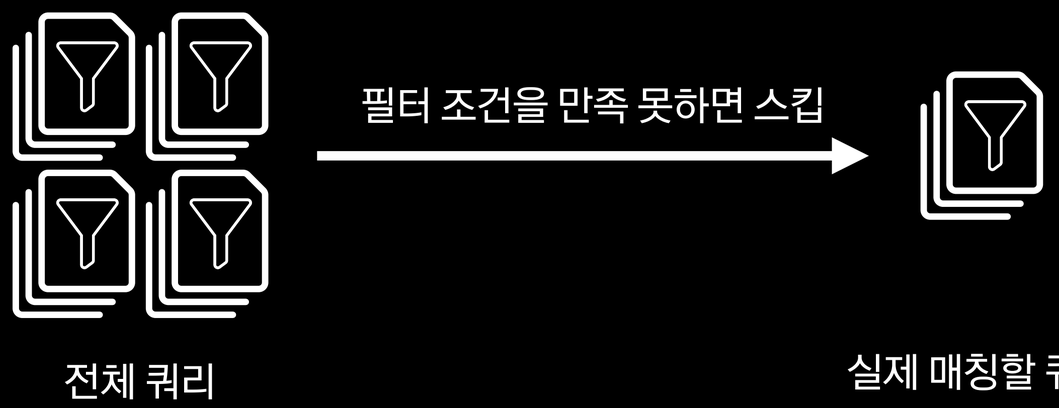
효율적인 매칭 알고리즘으로 검색 속도 향상 1/2

- 수평적 확장과 병행해서 효율적인 매칭 알고리즘 적용 필요



효율적인 매칭 알고리즘으로 검색 속도 향상 2/2

- filter 조건을 이용해 전체 쿼리 중 검사 대상 쿼리 선별
- filter 는 캐싱되어 검사 대상 쿼리를 빠르게 선별

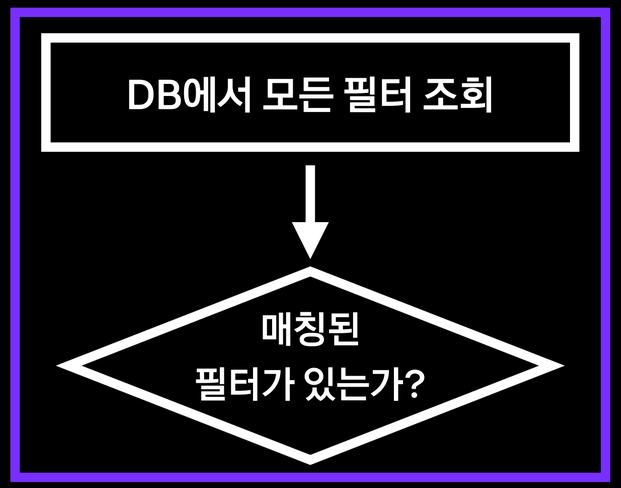



```
// PUT /pool-filters/_doc/1?refresh
{
  "query": {
    "bool": {
      "filter": [
        {
          "terms": {
            "category": ["경제"]
          }
        }
      ]
      ... // 나머지 쿼리
    }
  }
}
```

기존 구현과 Percolator 비교

	기존 구현	Percolator
쿼리(필터) 캐싱 및 동기화	구현 필요	ES 에서 기본 지원
콘텐츠& 필터 증가에 따른 Scale-out 기능	인스턴스 추가	ES의 샤딩과 레플리카를 활용
빠른 매칭을 위한 알고리즘	직접 구현 필요	ES의 매칭 알고리즘 활용
Full-text 검색	직접 구현 필요	ES의 가장 강력한 기능
속도	느림	빠름

기존 로직 대비 성능 비교



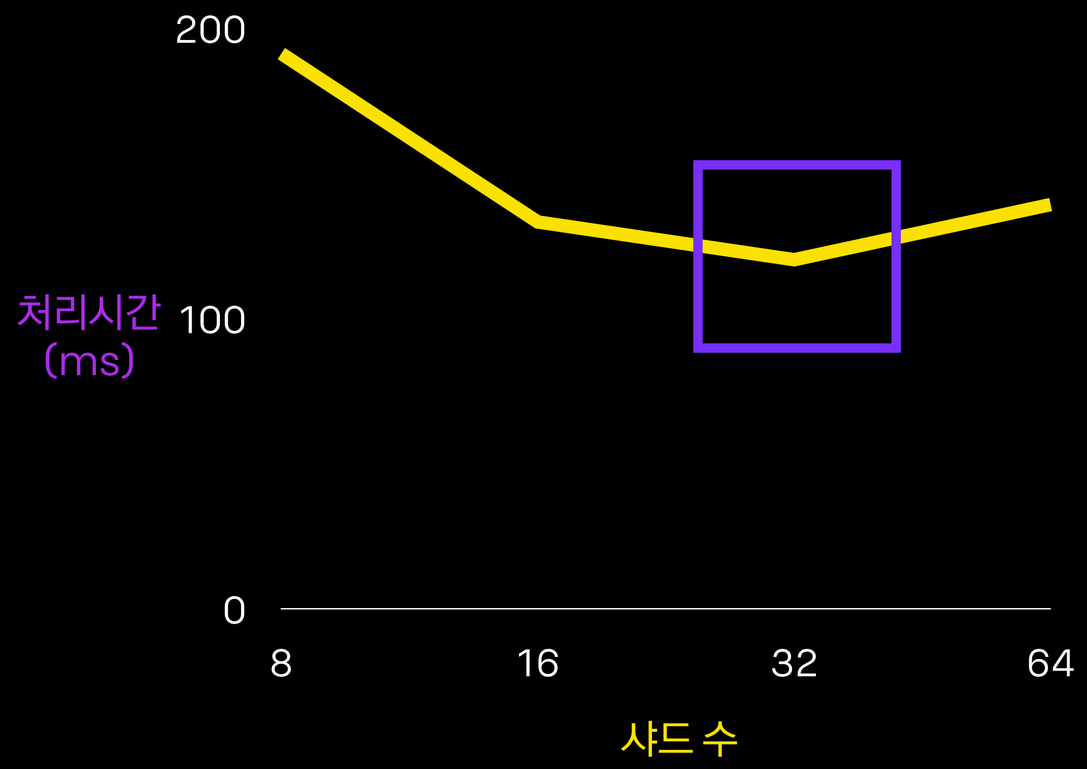
202ms

Elasticsearch
Percolator

86ms

샤드 수에 따른 성능 비교

샤드 수	8	16	32	64
필터 수	111,143	111,143	111,143	111,143
처리 시간 (ms)	191	133	120	139



정리

Recap

- 분류할 데이터나 필터가 많은데 빠르게 분류하고 싶을 때
- 샤딩과 레플리카를 통한 Scale-out이 가능
- Elasticsearch의 효율적인 알고리즘 활용
- 최적의 샤드 수를 위해 실제 데이터와 쿼리로 테스트 필요

또 다른 활용 1/2

- 앱 로그를 이용한 실시간 알람 시스템 개발

  데이터: 로그 메시지

  필터: 로그 메시지에 특정 문자열이 포함됨

또 다른 활용 2/2

- 특정 조건을 만족하는 중고 제품이 올라오면 알람

  데이터: 상품 정보

  필터: 제품명, 가격, 지역 등

E.O.D

Resources

1) https://www.findinpath.com/elasticsearch-percolator-primer/
2) https://www.elastic.co/kr/blog/when-and-how-to-percolate-1
3) https://www.elastic.co/kr/blog/when-and-how-to-percolate-2