한국일보
HANKOOKILBO.COM

자동 키워드 추출 및 연관 기사 검색을 위한 AWS 기술 활용 사례

한국일보 | IT팀 김우진 | 2022년 12월 15일

발표 내용

한국일보 클라우드 환경

OpenSearch 도입 및 활용

검색 품질 개선

Comprehend 기술 활용 사례

한국일보 클라우드 환경

한국일보가 운영 중인 클라우드 환경

한국일보 클라우드 시스템의 문제점

한국일보 클라우드 통합 추진

한국일보에서 운영 중인 클라우드 환경

도입 시기가 다른 멀티 클라우드 사용

Microsoft Azure

2014 ~
한국일보 홈페이지, WCMS
PaaS 기반으로 구축

NAVER CLOUD PLATFORM

2020 ~
CMS, 쿠버네티스 기반

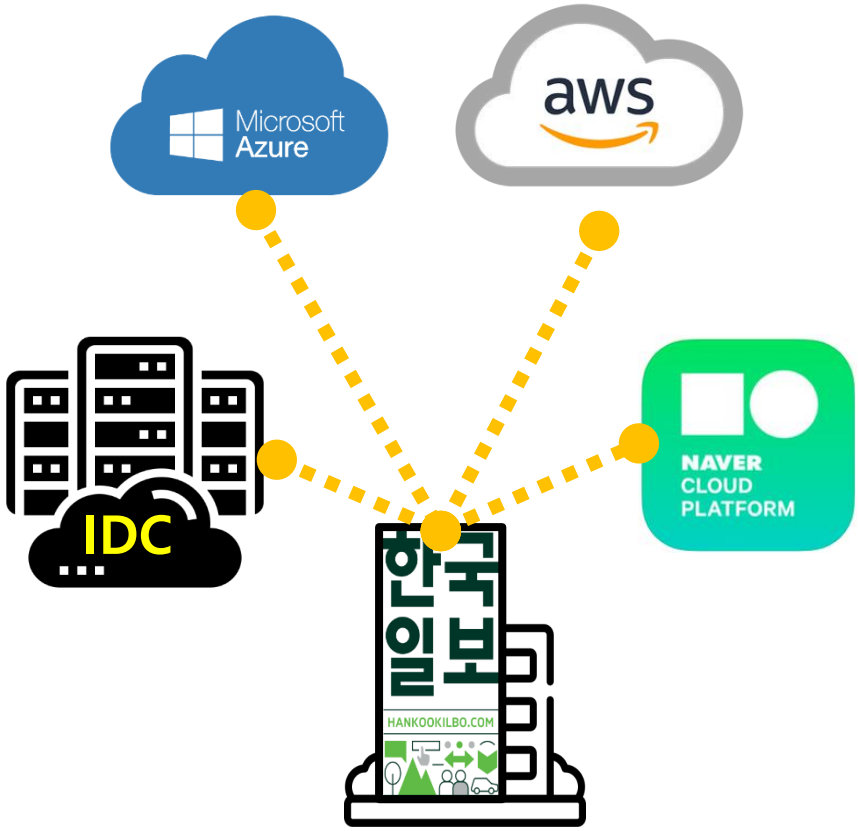
aws

2018 ~
Cloud Front(S3) 이미지, PDF서비스
OpenSearch, 아카이빙, 코리아타임스

2019 ~
CTS 시스템

멀티 클라우드 시스템의 문제점

전문 기술 인력 부족, 장애 대응 문제



잠재적 문제점

- 시스템을 관리할 수 있는 **전문 기술 인력 부족**
- 관리 포인트 사각 지대 존재
- 장애 발생 시 **대응 속도 문제**가 발생
- 물리적으로 **VM사용 비용 증가**하는 구조

한국일보 클라우드 통합 추진

클라우드 통합, 서버리스 서비스 도입

하나의 클라우드 환경으로 통합

- Azure 아카이빙 시스템 AWS 이관
- Azure 코리아타임스 시스템 AWS 이관
- CTS 시스템 AWS 클라우드로 이관 예정

서버리스 서비스 적극 도입

- **OpenSearch :** 검색 엔진 서비스 통합 사례
- **Comprehend** : 키워드 추출 및 연관 기사 검색 사례

OpenSearch 도입 및 활용

OpenSearch 도입 배경

OpenSearch 활용 범위

OpenSearch 도입 배경

최고 전문 기술 인력 보유, 지속적인 운영 관리 가능

아카이빙시스템 - WISEnut
상용 검색 엔진 사용
단순 유지만 가능하고
지속적인 개선의 한계

CMS허브시스템 - ElasticSearch
검색 엔진 설치 버전 사용
전문인력 부재로 인한
품질개선 및 운영 관리의 어려움

AWS 서비스 안정과 신뢰
최고 전문 기술 인력 보유
지속적인 운영 관리 가능

키워드 추출 및 연관 기사 검색 등
다른 서비스 연계 시 기능 확장 가능

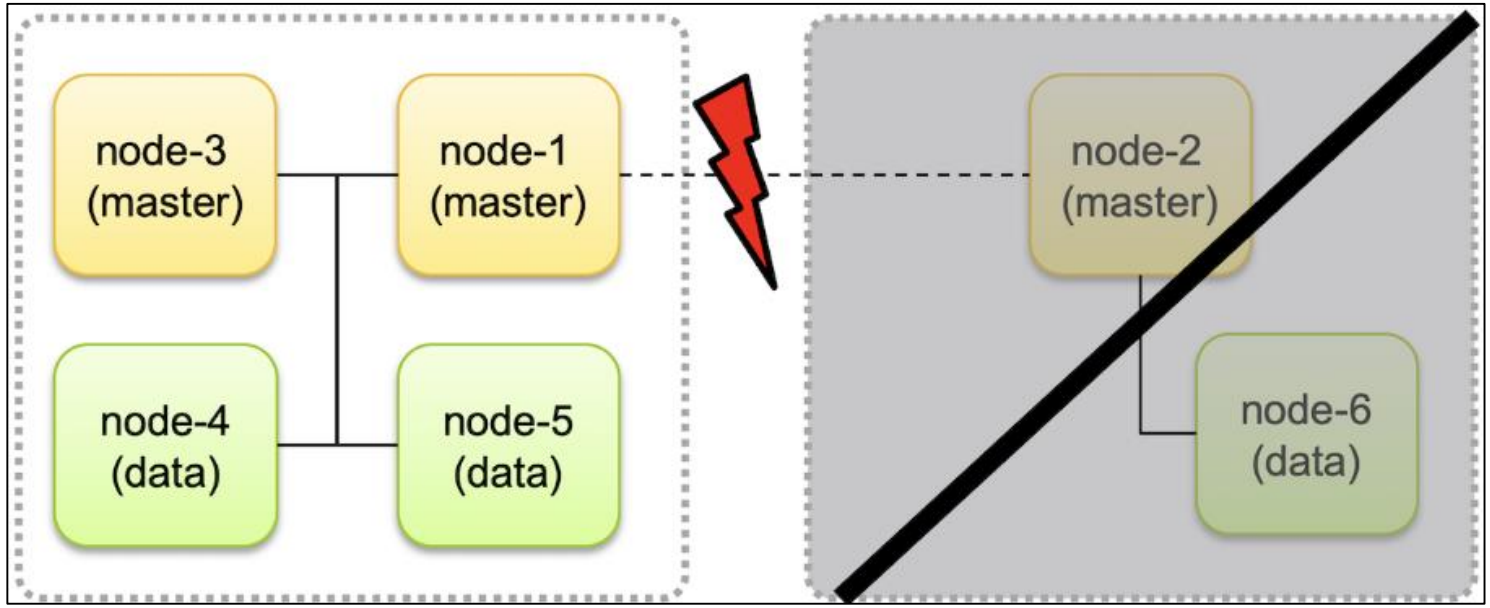
OpenSearch 도입 배경

운영 안정성 확보, 최신 기술 유지

Region 단위 DR 시스템이 구축되어 있어 네트웍 단절 시 **서비스 연속성 보장**

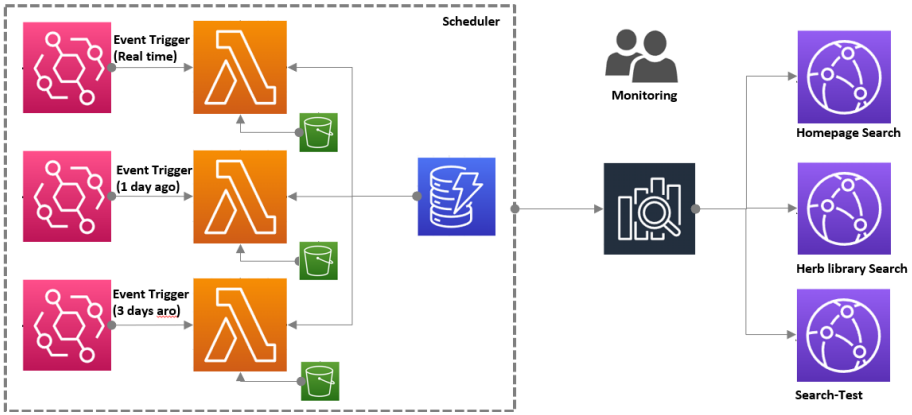
주기적인 버전 업그레이드를 통한 **최신 기술 유지 가능** (v5->v6)

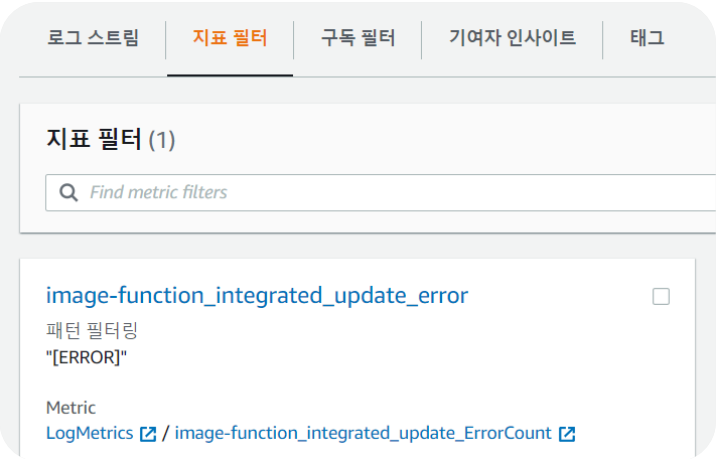
노드 구성 최적화로 데이터 안정성 확보 가능



OpenSearch 도입 배경

OpenSearch 운영 시 다른 클라우드 서비스와 연동 가능





Lambda Function

주기적인 이벤트 수신

입수 기사, 사진 증분 색인 처리

CloudWatch

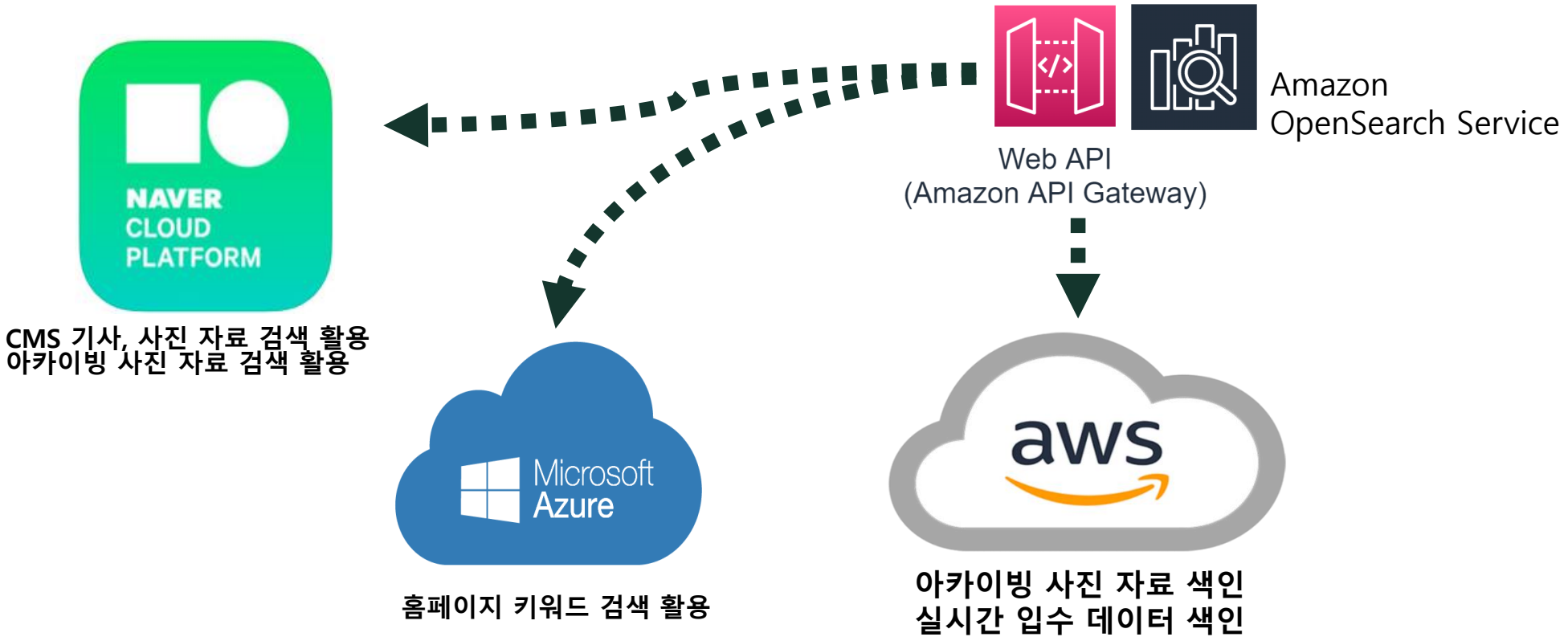
색인 로그 관리

지표 필터를 통한 에러 메시지 감지

Web Hook URL 방식으로 에러 전송 가능

OpenSearch 활용 범위

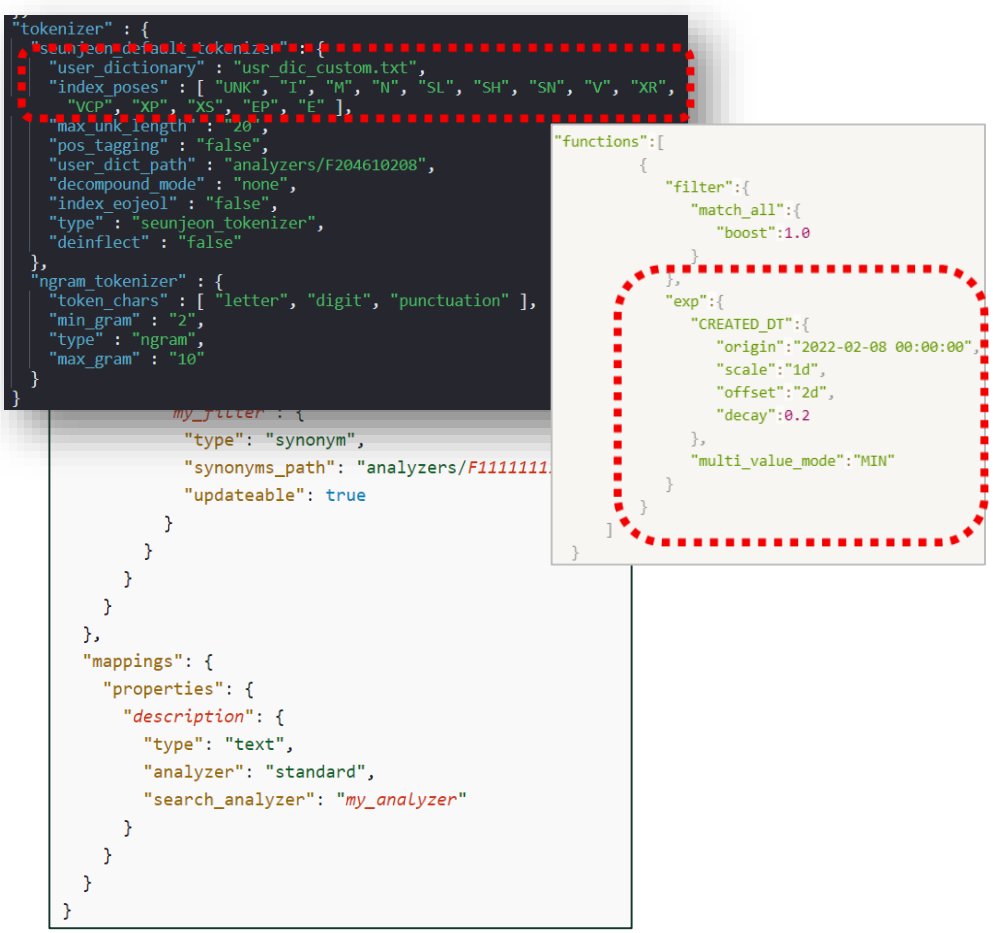
CMS기사/사진 자료 검색, 홈페이지 검색



검색 품질 개선

네이버 포탈과 한국일보 검색 결과 품질 비교

Index 튜닝 및 옵션 최적화로 검색 품질 개선



색인 및 검색 옵션 튜닝

아마존 검색 최고 전문가의 기술지원 활용

사용자 사전관리를 통한 품질 개선

tokenizer 품사 설정 및 특문 제거 처리

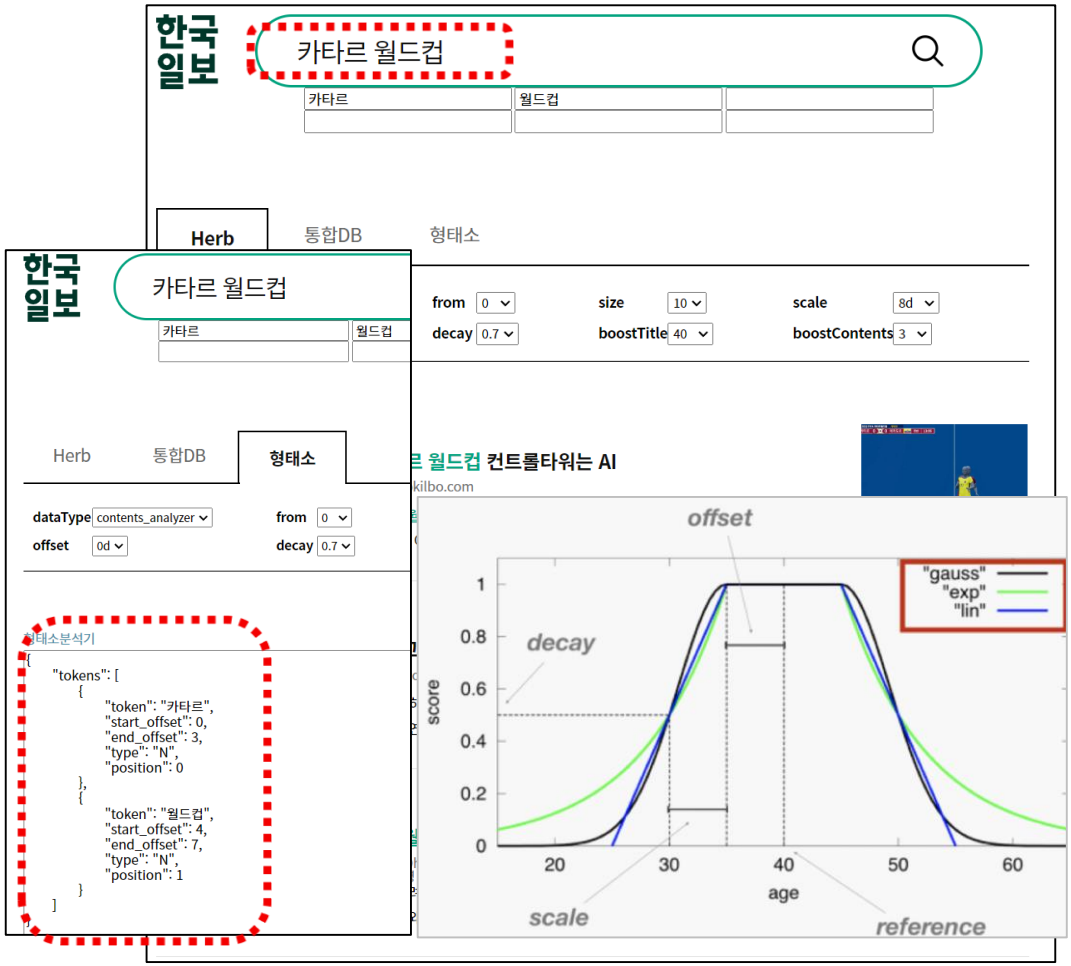
쿼리DSL Aging Function 옵션 튜닝

주기적인 재색인 처리

Kibana 데시보드를 통한 색인 관리

Scheduler를 통한 재색인 처리

검색 품질 테스트 및 지속적인 성능 업그레이드



검색 테스트 페이지

네이버 검색과 비교 분석에 활용

키워드에 대한 형태소 분석 결과 튜닝

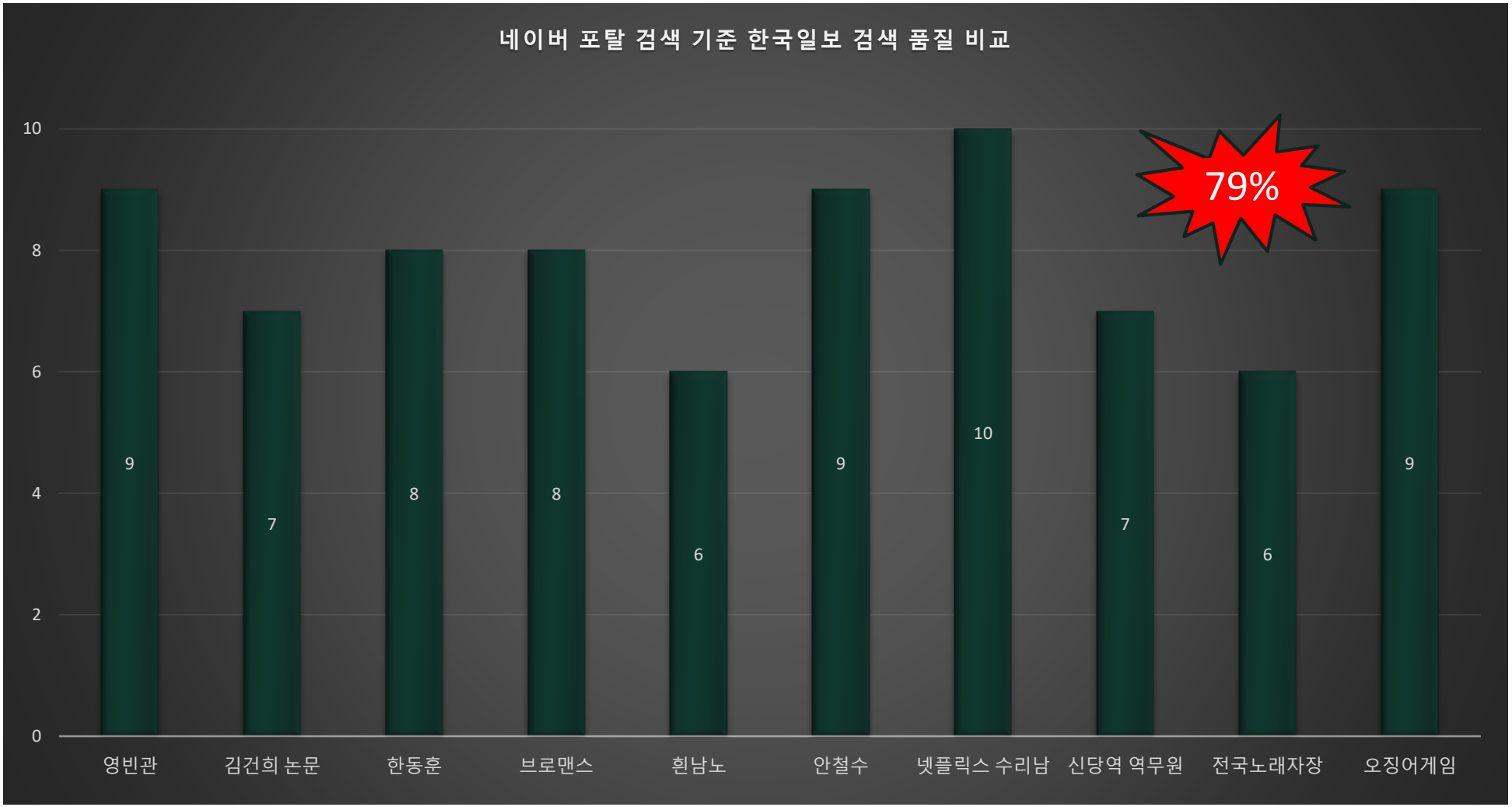
Aging Function 파라메타 옵션 조정을 통한 검색 결과 개선 진행

Boost : 가중치
Decay : Score 감소 비율
Scale : decay가 이루어지는 단위
Offset : decay에 영향 받지 않는 범위

검색 품질 개선

네이버 포탈과 한국일보 검색 결과 품질 비교

* **비교 기준** : 한국일보 기사 기준 네이버 검색 결과 상위 10개 기사와 한국일보 검색 결과 매칭 비율 (%)



AWS Comprehend 기술 활용 사례

(Comprehend를 활용한 키워드 추출 및 연관 기사 검색)

AWS 프로토타이핑팀 + 한국일보 IT팀 협업 개발

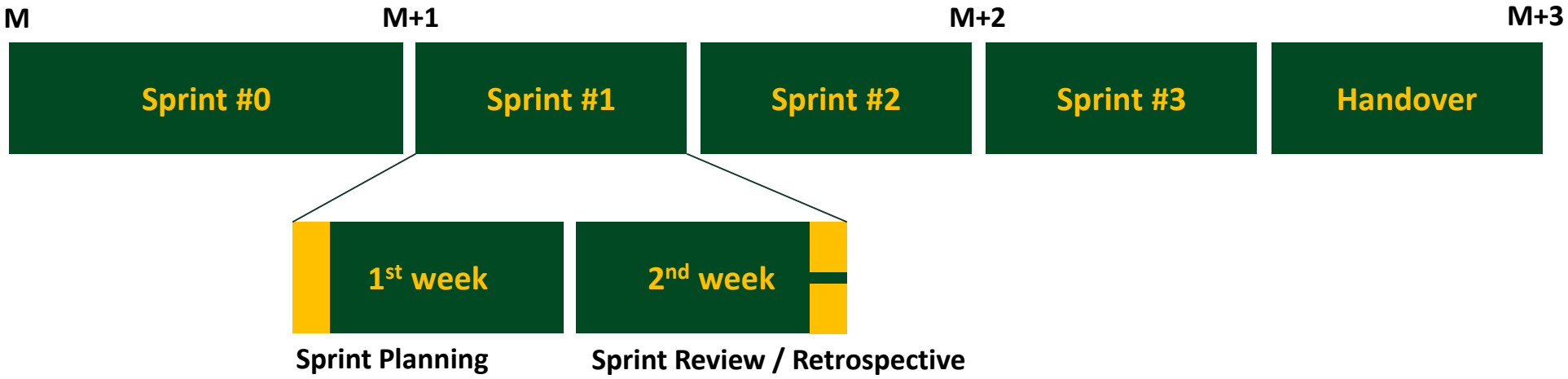
프로토타이핑 프로젝트 진행 과정 소개

프로토타이핑 프로젝트에 사용된 기술 용어

- **Prototype** : PoC 이후 기능 검증용 시제품
- **Comprehend** : 문서 콘텐츠에 대한 자연어 처리(NLP) 기술
- **Serverless** : AWS 클라우드 서비스, 이벤트 방식 처리(FaaS)
- **DevOps** : 개발(Development) + 운영(Operation) / Agile 방식
- **CDK** : Cloud Development Kit, 리소스 모델링 및 프로비저닝 도구
- **IaC** : 코드형 인프라 (Infrastructure as Code)
- **CI/CD** : 지속적 통합/배포(Continuous Integration, Continuous Deployment)

AWS 프로토타이핑팀과의 협업을 통한 개발

- 비즈니스 아이디어는 있지만 서버리스 서비스 개발 방법을 모름
- AWS 전문가의 도움을 통해서 개발이 가능
- 최대 6주간 AWS 전문가(Prototyping Engineer) 지원
- DevOps 방식의 협업 개발 (Agile방법론 기반Code Level협업)
- 의사 결정을 보조하고 상용화에 대한 최적 가이드 제공



한국일보 HANKOOKILBO.COM

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

프로토타이핑 목표

- 키워드 추출 및 연관 기사 추천 테스트 페이지 생성
- 하나의 기사에서 10개의 핵심 키워드를 자동으로 추출

  → **한국언론진흥재단 빅카인즈 키워드 추출 결과와 품질 비교**
- 특정 기간 색인된 기사 중 유사도가 높은 기사 5개 추천

현행 업무

- 기사 작성 후 기사의 핵심 키워드를 기자가 직접 수동으로 작성 중

한국일보
HANKOOKILBO.COM

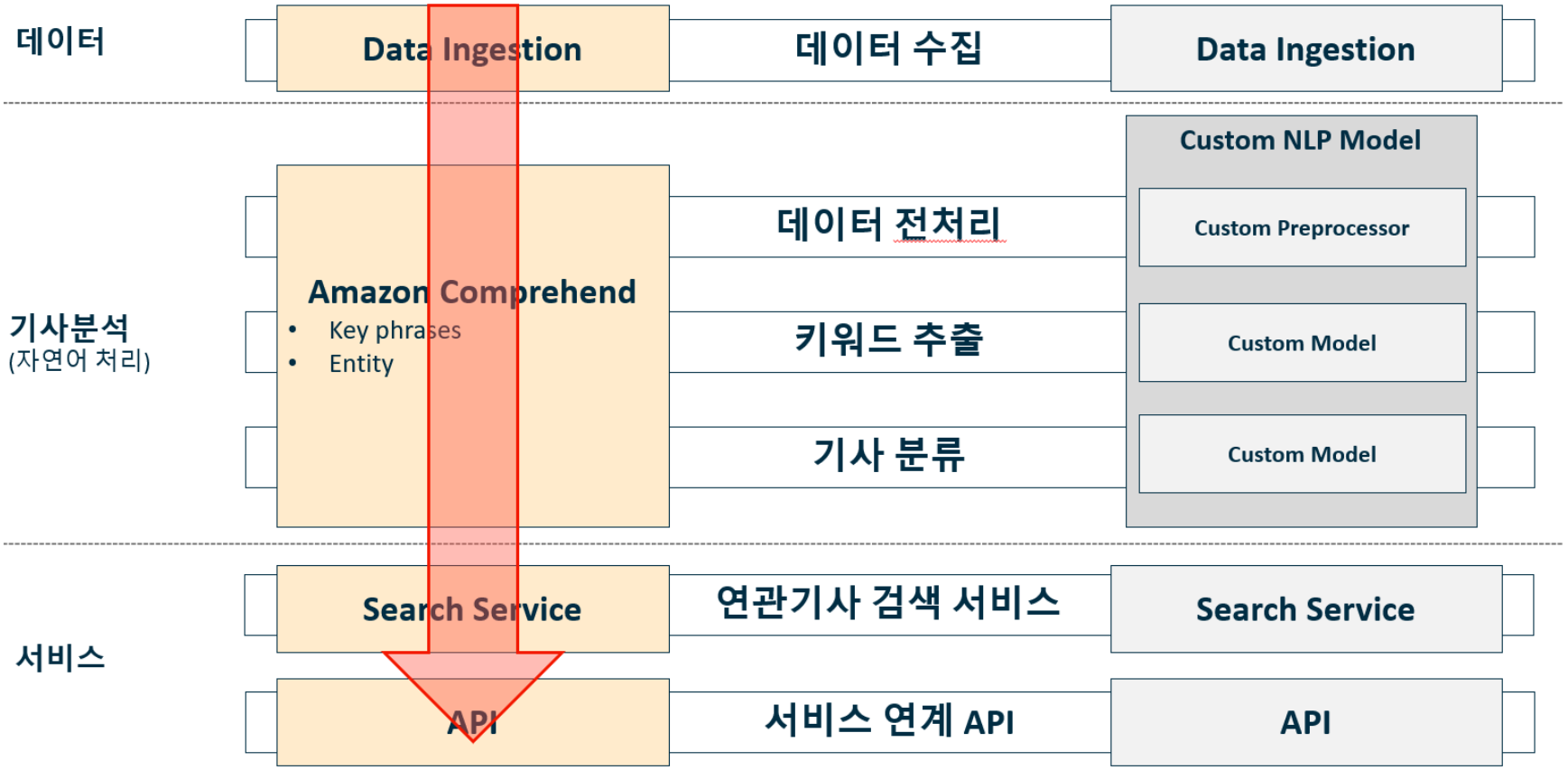
Comprehend를 활용한 키워드 추출 및 연관 기사 검색

프로토타이핑 기대 효과

- 기사작성 시 핵심 키워드 추출 기능 제공을 통한 업무 시간 단축
- 텍스트 처리 솔루션 개발을 위한 구현 역량 확보
- AWS Cloud 기반 IaC, DevOps 기술 내재화
- IaC, CICD, Serverless, DevOps 구축으로 운영 부담 최소화

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

프로토타이핑 개발 범위



Comprehend를 활용한 키워드 추출 및 연관 기사 검색

프로토타이핑 참여 인원

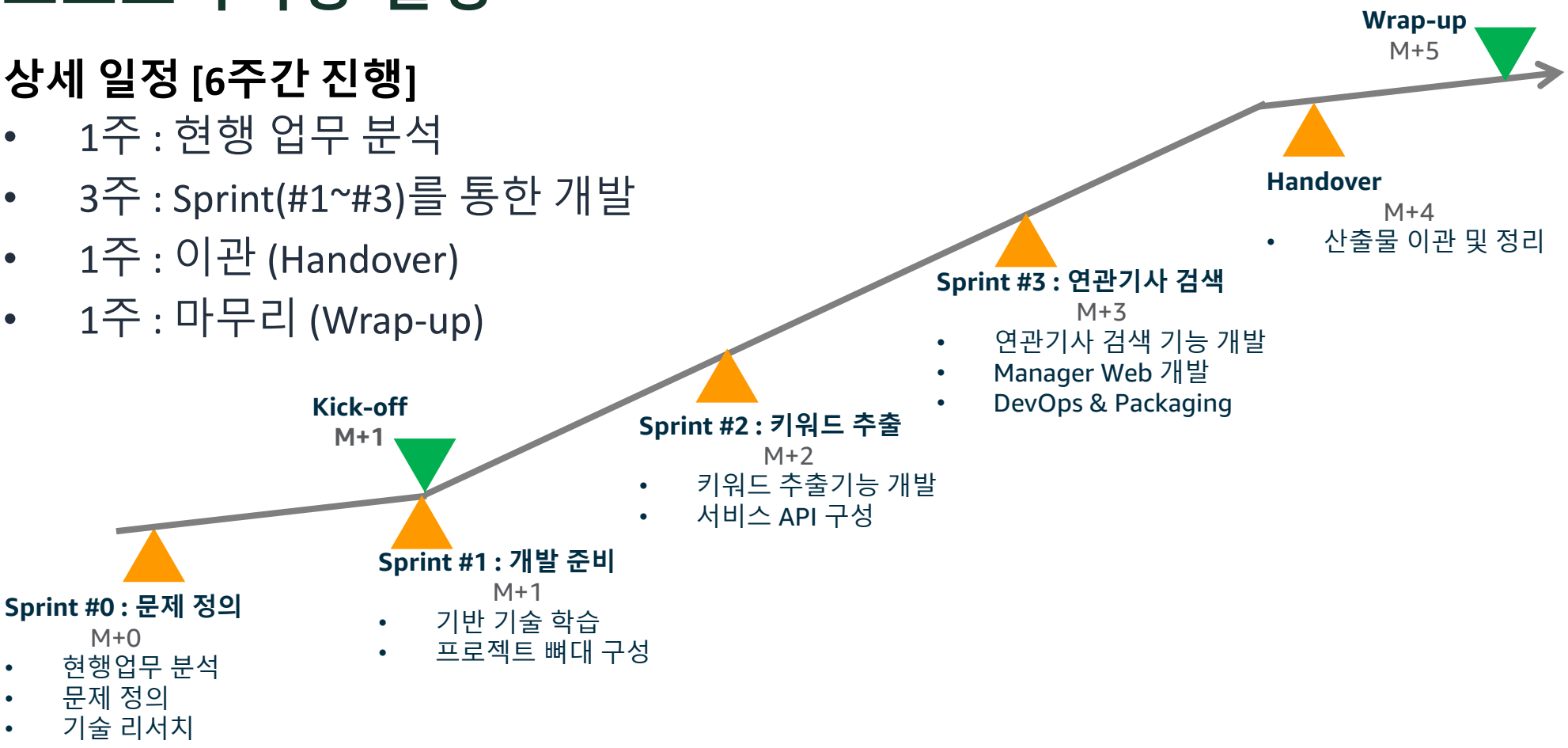
소속	담당 파트	비고
한국일보	Project Manager	
한국일보	**Developer**	**Backend (키워드 추출, 연관기사검색)**
한국일보	**Developer**	**DevOps (AWS Cloud, IaC)**
AWS	**Project Lead / Developer**	**Architect, FrontEnd**
AWS	Account Manager	
AWS	Solutions Architect	

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

프로토타이핑 일정

상세 일정 [6주간 진행]

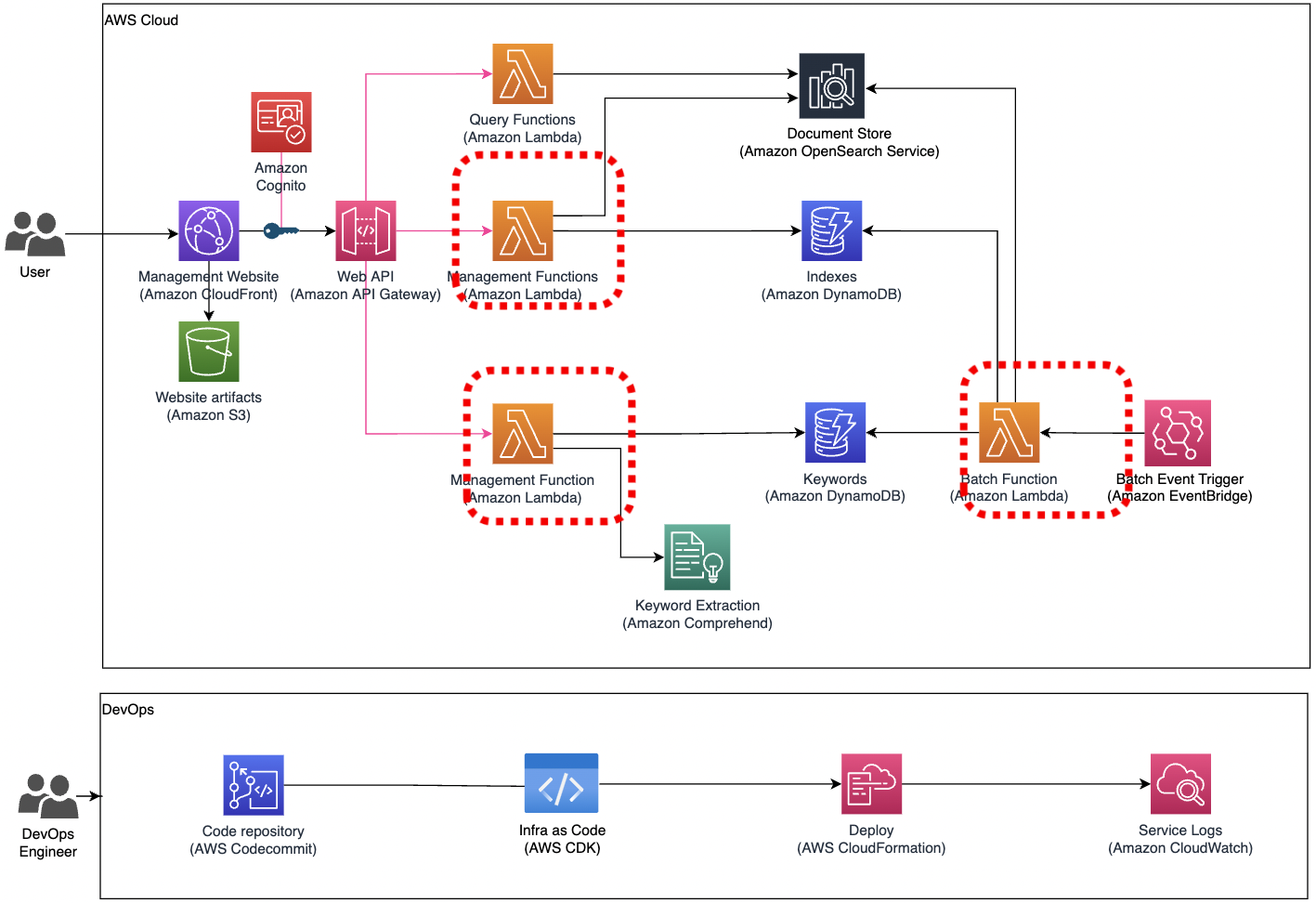
- 1주 : 현행 업무 분석
- 3주 : Sprint(#1~#3)를 통한 개발
- 1주 : 이관 (Handover)
- 1주 : 마무리 (Wrap-up)



AWS Comprehend 기술 활용 사례

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

AWS Serverless 서비스를 활용한 키워드 추출 및 연관기사 검색 초기 설계



Kickoff ver.

AWS Comprehend 기술 활용 사례

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

독립개발 및 배포 가능한 단위로 Stack을 구성하여 유연한 확장과 효율적인 협업이 가능

■ **인증 및 보안**

- IdentityStack

■ **기사분석**

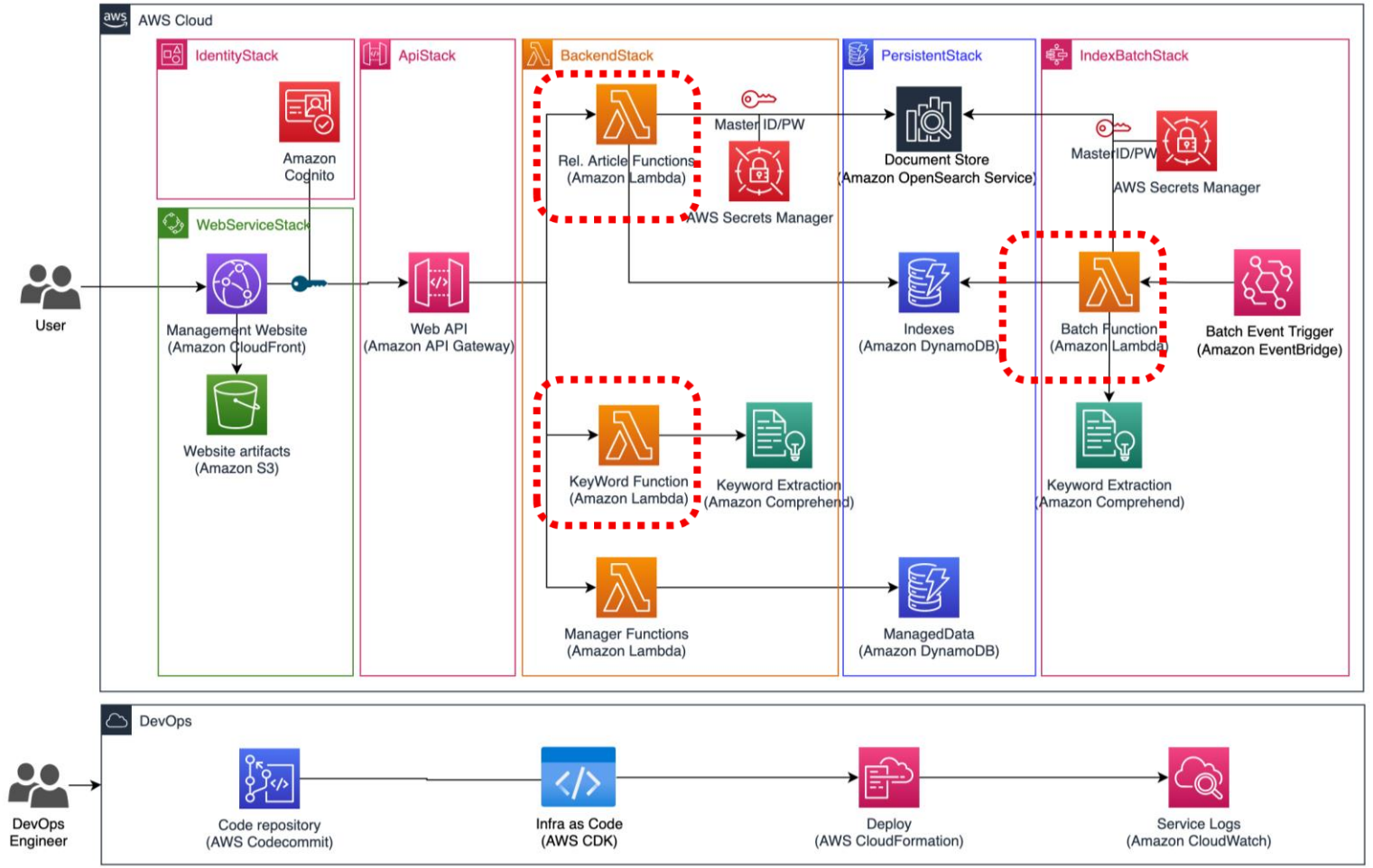
- BackendStack
- IndexBatchStack

■ **UI 및 Web API**

- WebServiceStack
- ApiStack

■ **저장소**

- PersistentStack

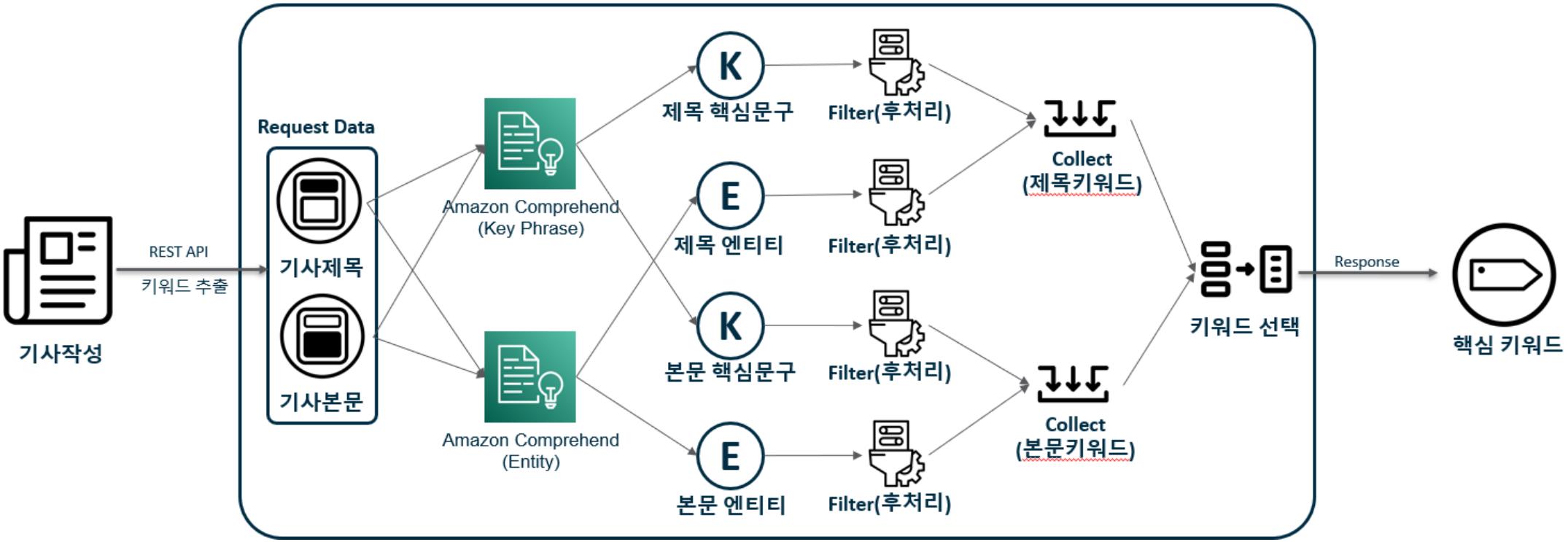


Sprint #3 ver.

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

키워드 추출 및 후처리 과정

- **Filter(후처리)** : Score, 단어 개수, 최소 글자수, 숫자 허용여부, 엔티티 종류
- **키워드 선택** : 제목 키워드/본문 키워드 선택 비율



Comprehend를 활용한 키워드 추출 및 연관 기사 검색

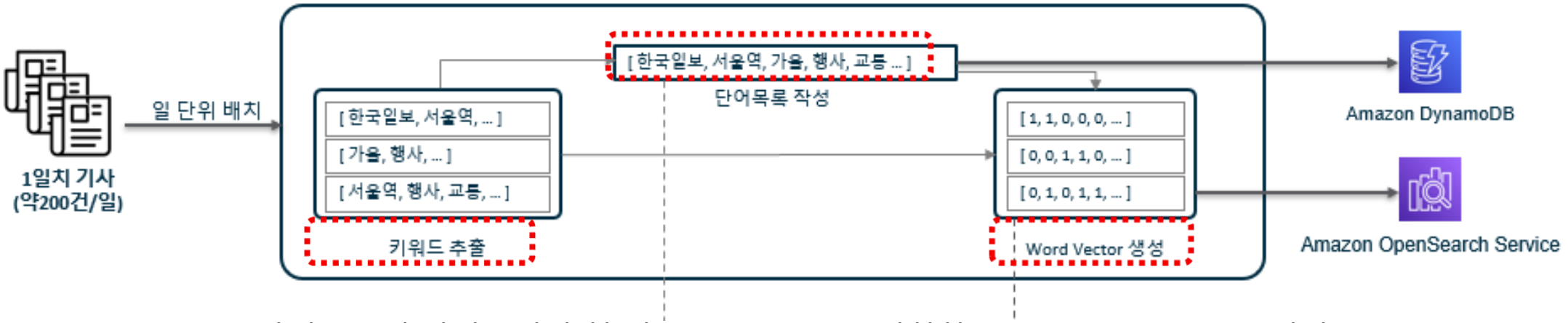
IndexBatch 작업

일 단위 배치 작업으로 하루에 생산된 모든 기사의 키워드 추출

하루 단위 단어 목록을 작성하고 DynamoDB에 저장

단어 목록을 기준으로 각 기사의 Word Vector를 생성해서 OpenSearch 필드에 저장

■ **하루 단위 배치작업 / 단어 목록 작성 / Word Vector 작성**



* **Word Vector** : 단어목록에 키워드의 출현 여부를 1/0으로 수치화한(one-hot-encodding) 배열

한국일보

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

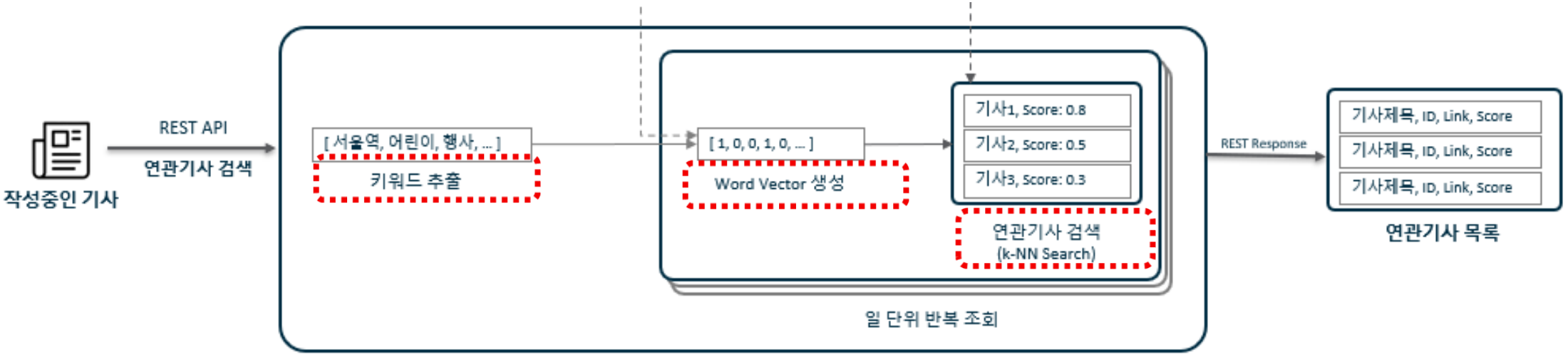
연관기사 검색

연관 기사 검색 시 REST API를 통해서 키워드를 추출하고 단어목록 기준 Word Vector 생성

Word Vector 기준 k-NN알고리즘으로 OpenSearch에 저장된 유사도가 높은 기사를 검색함

■ **REST API / 기사의 키워드 추출 / 배치작업으로 생성된 단어목록으로 Word Vector 생성**

- **K-NN 알고리즘을 이용하여 Word Vector 의 유사도가 높은 기사 순으로 조회**



* **k-NN 알고리즘** : k-Nearest Neighbor(k-최근접 이웃) 알고리즘. 가장 거리가 가까운 k 개의 이웃을 선택

한국일보
HANKOOKILBO.COM

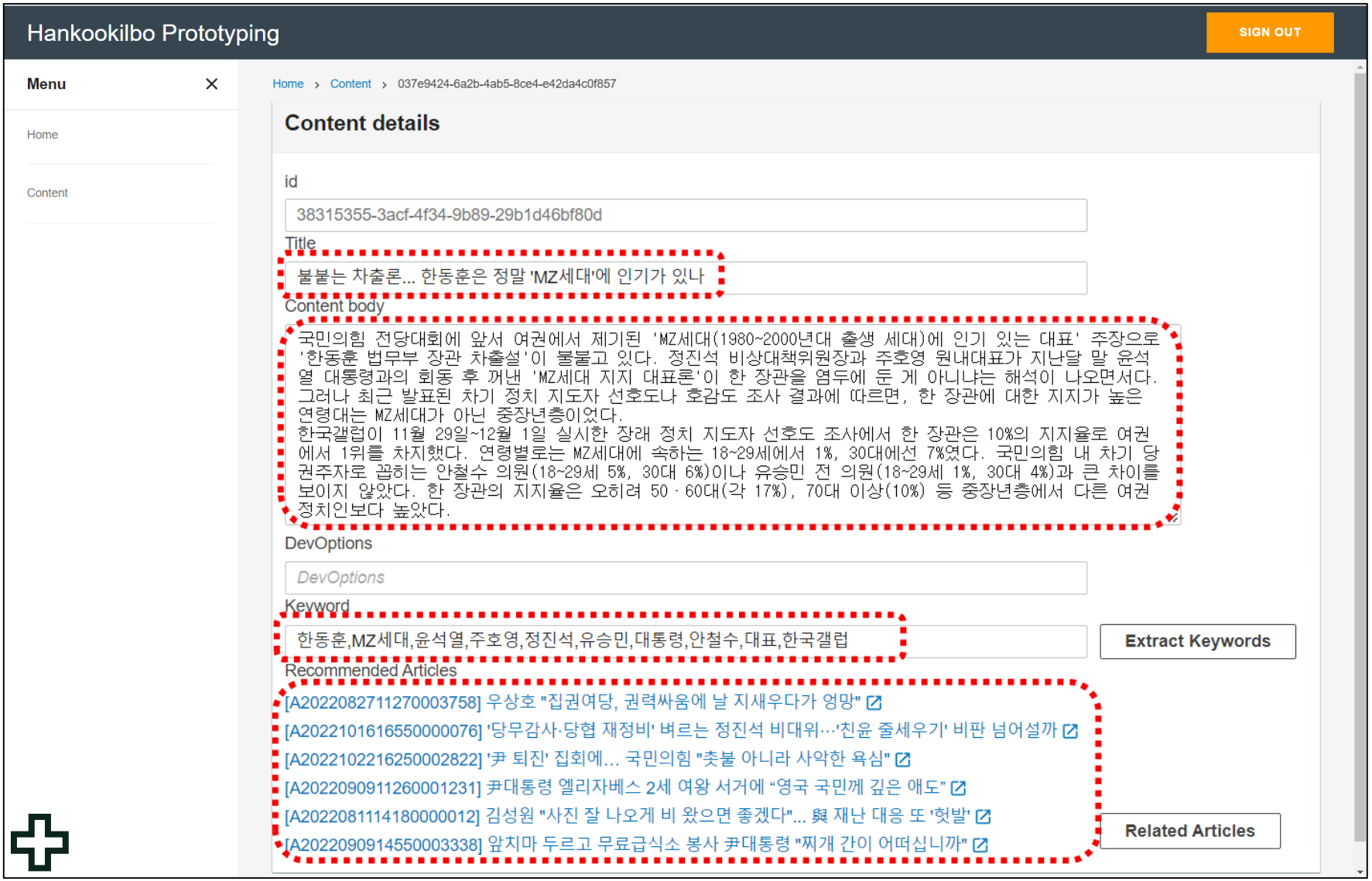
Comprehend를 활용한 키워드 추출 및 연관 기사 검색

프로토타이핑 결과

- 키워드 추출 및 연관 기사 검색 **테스트 페이지 확보**
- 한국언론재단 빅카이즈 키워드 추출 결과 **품질 비교 가능**
- Amazon Comprehend를 이용한 키워드 추출 서비스 구성 가능
- 과거 기사 색인을 통한 유사한 키워드를 가진 기사 검색에 활용 가능
- AWS Serverless 서비스 활용 경험
- 코드 구조화(IaC) DevOps 역량 확보

AWS Comprehend 기술 활용 사례

Comprehend를 활용한 키워드 추출 및 연관 기사 검색



한국일보
HANKOOKILBO.COM

Comprehend를 활용한 키워드 추출 및 연관 기사 검색

"빅카인즈" 키워드 추출 결과 품질 비교

제목 : 불붙는 차출론... 한동훈은 정말 'MZ세대'에 인기가 있나

본문 : 국민의힘 전당대회에 앞서 여권에서 제기된 'MZ세대(1980~2000년대 출생 세대)에 인기 있는 대표' 주장으로 '한동훈 법무부 장관 차출설'이 불붙고 있다. 정진석 비상대책위원장과 주호영 원내대표가 지난달 말 윤석열 대통령과의 회동 후 꺼낸 'MZ세대 지지 대표론'이 한 장관을 염두에 둔 게 아니냐는 해석이 나오면서다. 그러나 최근 발표된 차기 정치 지도자 선호도나 호감도 조사 결과에 따르면, 한 장관에 대한 지지가 높은 연령대는 MZ세대가 아닌 중장년층이었다.

한국갤럽이 11월 29일~12월 1일 실시한 장래 정치 지도자 선호도 조사에서 한 장관은 10%의 지지율로 여권에서 1위를 차지했다. 연령별로는 MZ세대에 속하는 18~29세에서 1%, 30대에선 7%였다. 국민의힘 내 차기 당권주자로 꼽히는 안철수 의원(18~29세 5%, 30대 6%)이나 유승민 전 의원(18~29세 1%, 30대 4%)과 큰 차이를 보이지 않았다. 한 장관의 지지율은 오히려 50·60대(각 17%), 70대 이상(10%) 등 중장년층에서 다른 여권 정치인보다 높았다.

호감도 조사에서도 비슷한 결과가 나왔다. 지난 10월 14, 15일 … 리서치·주간조선이 실시한 주요 정치인 호감도 조사를 보면, 한 장관은 37.5%로 안 의원(35.2%) … 비교해 가장 낮았다. 70세 이상에서 … '매우 호감이 간다'는 응답도 50.1%에 달했다.

여론조사기관 에스티아이 … 대에서 다른 여권 주자들을 한 번도 앞선 적이 없다"며 "오히려 국회 인사청문회에서 … 윤석열 대통령이 내건 '공정'이란 가치를 퇴색시킨 측면이 있다"고 분석했다. MZ세대가 민감하게 반응 … 상대적으로 약점이 있는 것으로 볼 수 있다는 설명이다.

"빅카인즈" 키워드 추출 결과

중복

- MZ세대, 호감도, 호감도 조사, 원내대표, MZ세대 지지, 정치인 호감도 조사, 선호도, 30대 호감도, 정치인 호감도, 여권 주자들

"AWS Comprehend" 키워드 추출 결과

- 한동훈, MZ세대, 윤석열, 주호영, 정진석, 유승민, 대통령, 안철수, 대표, 한국갤럽

AWS 클라우드 서비스 기술 활용 계획

다음 프로토타이핑 목표

- 한국일보에 보관 중인 방대한 사진 활용
- Amazon Rekognition서비스를 활용한 이미지 AI검색 모듈 개발

  예> 유명인사, 구글 이미지 검색 및 사진 관리, 애플 사진 관리 등

* **Amazon Rekognition** : 딥 러닝 기반 시각 분석 서비스


한국
일보
HANKOOKILBO.COM