

국민연금 관련 여론/감성 분석 (Sentiment Analysis)

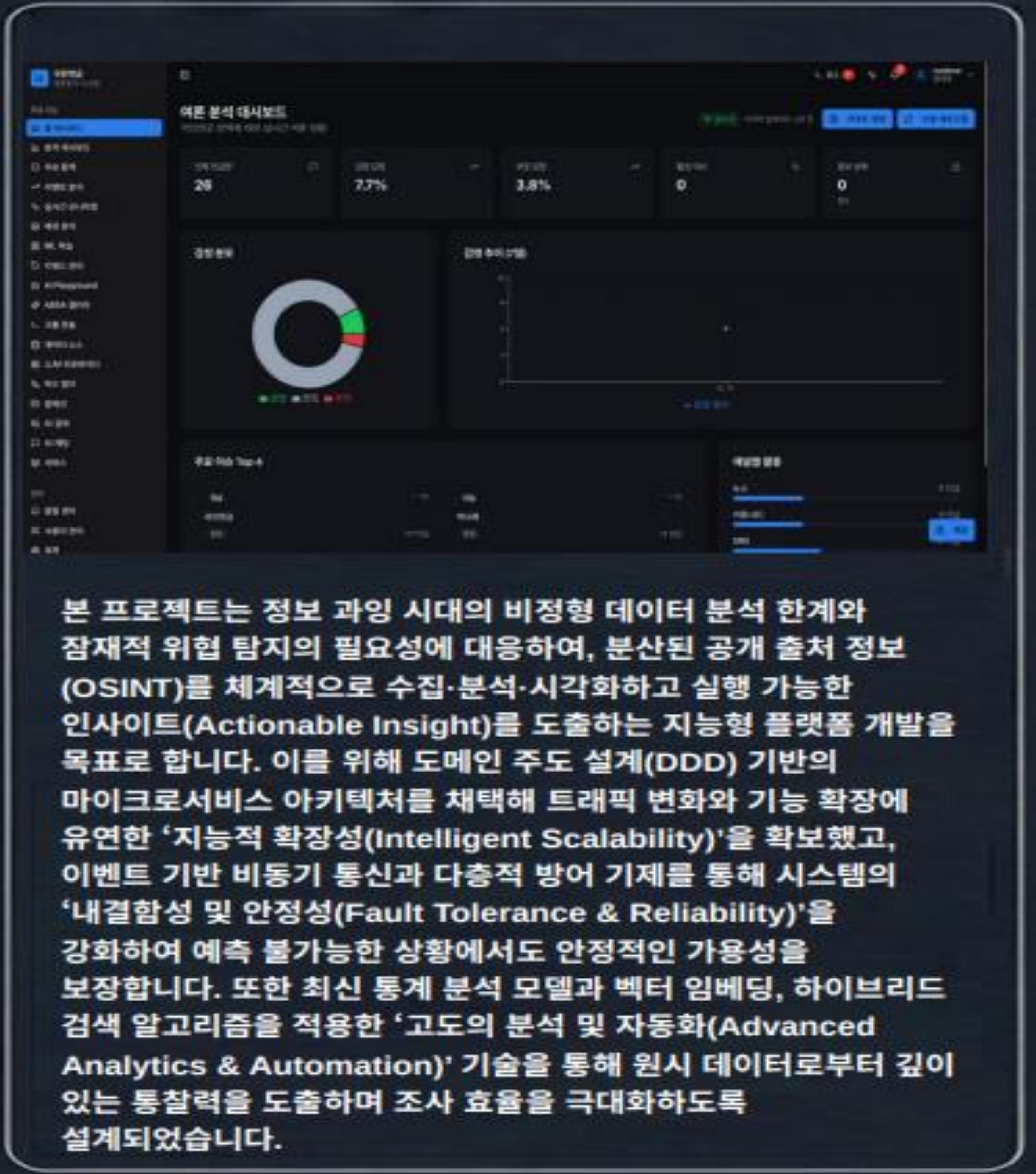
2025학년도 2학기
SW 캡스톤디자인
경진대회

팀 명 바이너리
지도교수 김윤경

팀 원 최시몬(컴퓨터공학부, 4), 정찬우(통계학과, 4)
산업체 국민연금공단

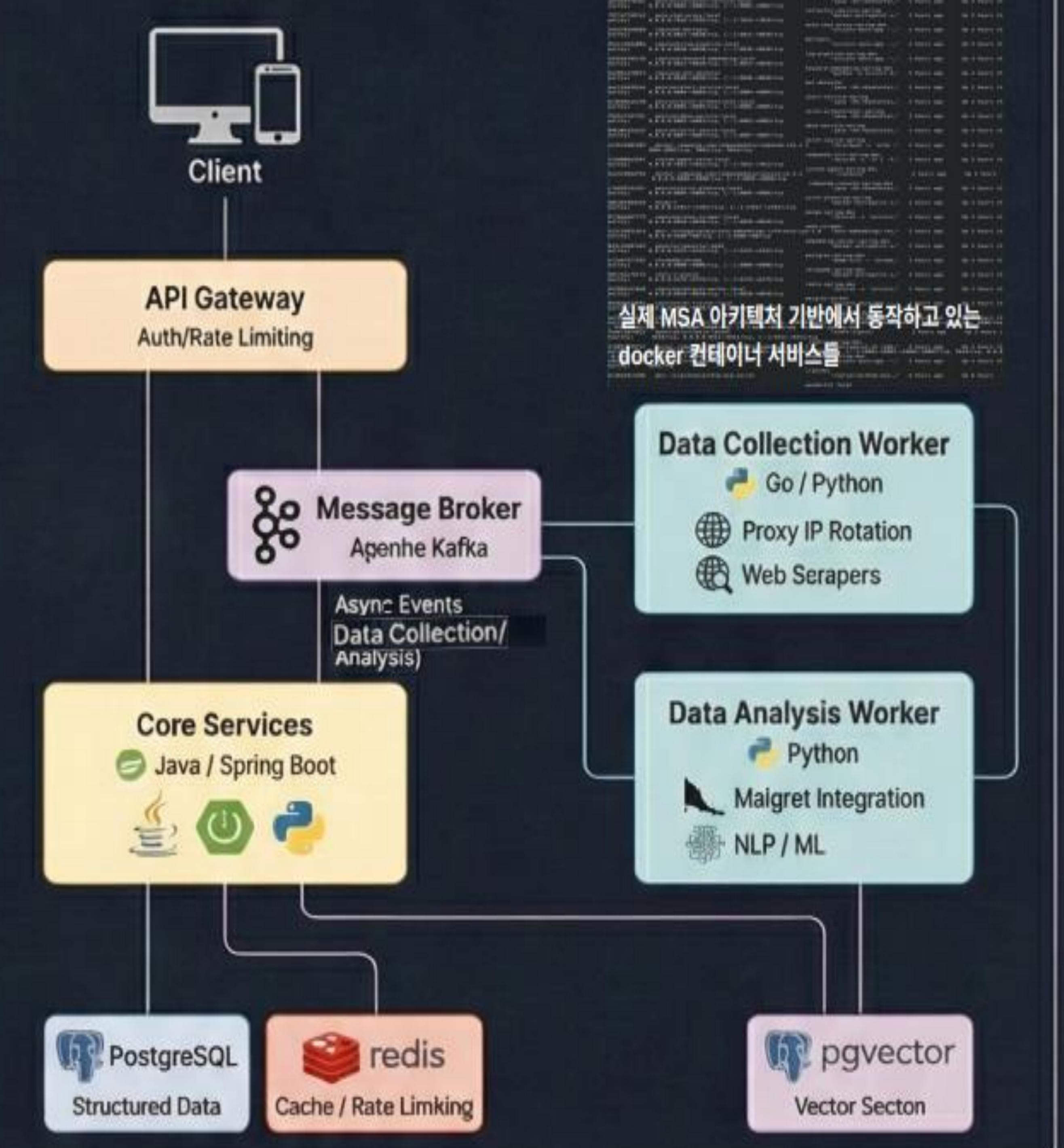
개발 동기 및 목적

1. INTRODUCTION



주요 기술

SYSTEM ARCHITECTURE



개발 내용

PILLAR I: PRINCIPLED DESIGN

검증된 SW 설계 패턴과 최적의 언어 사용으로 유지보수성과 확장성에 기반한 설계 (+ 특정 언어가 더 좋다가 아닌 트렌드에 오프가 존재한다는 점에 감안)

- 스프링부트 : 안정적인 트랜잭션이 필요한 핵심 비즈니스 로직 구현
- 파이썬 : 풍부한 라이브러리가 잘 지원되어 있어, AI/ML 워커 및 동적 크롤러 구현
- GO : 고성능 네트워크 처리가 필요한 프록시 관리 구현에 사용

Facade Pattern

PILLAR II: RESILIENT PIPELINE

장애 발생은 무조건 일어난다는 생각을 바탕으로 설계된 대용량 크롤링 데이터 스트림 처리 파이프라인 구축

- Redis 기반 토큰 배치 알고리즘으로 순간적인 고속 요청 트래픽 차단, Kafka consumer group 수평적 확장으로 병목 감소, 처리 불가능한 비상황에서 DLQ를 사용하여 보호

Rate Limiting (Token Bucket)

기반 토큰 벅트 알고리즘을 통해 고속 요청 시 Rate Limit 트리거 작동하여 시스템 부하를 최소화

Backpressure Handling

데이터 급증 시 컨슈머 그룹 확장 및 대규모 디중 스레드 동시 요청 시에는 차단

Pipeline Protection (DLO)

Kafka를 통한 로그 전송 과정에서의 파이프라인 protection 잘못된 형식 차단 실체 동작

Hybrid Search (BM25 + Vector)

BM25 알고리즘과 임베딩 벡터 검색을 결합한 후 RRF 알고리즘을 통해 두 검색 결과 순위를 지능적으로 재조정할 수 있도록 구현, 사용자의 의도를 파악한 최적의 검색 결과를 제공하기 위함

Chain of Responsibility Pattern

Reciprocal Rank Fusion (RRF)

Query: '보안'을 사용했을 때 RRF 알고리즘 실행 동작과정 로그

$$\text{score} = \sum (\text{weight}_i / (k + \text{rank}_i))$$

요청을 처리할 수 있는 여러 객체를 체인으로 연결해서, 앞에서부터 차례대로 요청을 처리하도록 하는 설계 방법. 이 필터의 순서에 따른 트래픽 과부하의 편차가 심하므로 이를 감안하여 순서를 배치

결과 및 분석

PILLAR IV: ROBUST SECURITY (건고한 보안)

Role-Based Access Control (RBAC)
역할에 따른 시스템 자원에 대한 접근 권한을 세분화하여 통제 (Admin, User, Researcher 등)

Go Proxy IP Rotation

실제 IP rotation을 통한 크롤링 시 복 차단에 대응한 동적 프록시 관리 로그 (국가 우회, IP 우회 등 여러 방안 동적으로 적용)

4. RESULT

전문가 수준의 디지털 포렌식 기능 구현
Malgret와 같은 전세계 수천 개의 사이트에 동시 디지털 포렌식 기능 작동하도록 프로젝트에 통합하여 심층적인 정보 분석 환경을 제공

대용량 데이터 처리 성능 최적화

풀리그랫 MSA 아키텍처, Kafka를 통한 이벤트 기반 처리를 적용한 대규모 데이터 처리와 자동 확장 기반 크롤러를 적용하여 쉽게 사용 가능

5. TECH STACK

