- 잘못 연결된 정보로 인해 분석 결과가 왜곡되는 일 최소화
- 운영자가 모든 데이터를 눈으로 확인하지 않아도 신뢰도 높은 데이터 확보 가능

Q. 실시간 데이터 수집은 어떤 방식으로 구현되는가?

A. 이벤트가 발생하면 이를 빠르게 감지하고, 즉시 처리하는 스트리밍 기반 수집 방식을 채택하고 있습니다.이 방식은 수 초 내 반영이 가능하며, 지연이나 누락 여부는 실시간 감시를 통해 추적됩니다.

실시간 이벤트 감지 및 처리 구조,실시간 데이터 변환 및 전처리 기능,지연 및 장애 상황 자동 대응을 통해 중요한 정보가 지연 없이 시스템에 반영되며 실시간 대시보드, 알림, 자동 처리 등과 연계하기 적합합니다.

Q. 데이터 수집 단계에서 문서보안 시스템 연계가 가능한가?

A. 사내 시스템과의 연계를 위한 다양한 API 및 권한 통제 체계를 제공하고 있으며, 전자문서 권한관리 기능도 포함되어 있습니다. 또한 고객사에서 사용하고 있는 문서보안 시스템에서 파싱을 위한 복호화 API 및 key 라이브러리등을 제공할 경우 수집 전처리 파싱단계에서 연동 개발을 하여 적용 지원이 가능합니다.

Q. 문서를 Vector로 바꾸면 용량이 얼마나 증가하나요?

A. 문서를 Vector로 변환하여 저장할 경우 최소 10배에서 많게는 30배 가량 증가합니다. 여러 프로젝트에서 용량을 측정 및 시뮬레이션 한 결과이며 청크 사이즈 및 수집대상의 문서나 Text 종류별로 다를 수 있습니다. 짧은 단어나 문장은 Vector 사이즈(SSL 키값수준의 길이) 만큼 늘어나서 많이 늘어나고 신문기사나 논문과 같은 경우 적게 증가하는것을 확인 할 수 있었습니다.

하지만 검색 정확도를 위하여 청킹 시 윈도우가 겹치는 구간을 많이 하면 할수록 데이터 저장공간이 더 많이 필요 할 수 있습니다. 즉, 청크 갯수와 비례해서 증가합니다.

Q. Vector 임베딩은 어떤 과정으로 진행되는 것인지?

A. bge-m3, E5 등 다국어 모델을 사용하며, 전처리된 문서 단위로 청킹 후 벡터 임베딩을 수행합니다.