멘티 활동 일지

작성일자: 2023년 9월 21일

도메인	빅데이터 (분산)
진행 일시	2023-09-21 15:05 ~15:40

내용	**빅데이터 파이프라인 관련 피드백**
	
	• ELK 를 사용하면 분산 처리가 해결되니 좋은 것 같음
	• 병렬적으로 가는 방법 대신 ELK 로 통합해보는 것을 고려하면 좋겠음
	  ◦ 프로젝트가 진행된지 오래돼 QueryDSL 과 병렬로 가는 것도 괜찮음
	  ◦ RDS 에서 Logstash 로 데이터를 가져오는 것이 아니라, API 호출 시 데이터가 json 형식으로 들어오면 바로 Logstash 에서 ES 로 보내는 방법도 괜찮음
	• Kafka 를 사용해 ES 에 데이터를 넣는 방법도 있음
	
	유사도 계산 알고리즘 피드백
	
	• 유사도 계산 시 자연어 처리가 아니라 룰베이스 기반 처리한 로직이 좋음
	• Topic modeling 이라는 유사도 계산 알고리즘이 있는데 참고하면 좋겠음
	  ◦ 단어-단어, 단어-문서 간 유사도 score 계산 가능
	  ◦ 실시간으로 업데이트 해야 하는 서비스가 아니어서 고려해볼 만함
	
	기획 관련 피드백
	
	• 질문: 현업에서는 도메인 전문가와의 미팅을 어떻게, 어느 주기로 진행하시나요?
	  ◦ 개발자는 요구사항을 받아 개발을 한다
	  ◦ 기획 시에는 기획 회의만 한 달 정도 진행
	  ◦ 개발 단계에서도 필요할 때마다 요구사항과 맞는지 점검하면서 개발 진행
	• 질문: 대량의 정형 데이터 분산 조회 및 병렬 처리를 어떻게 하는게 좋을까요?
	  ◦ Kafka 를 이용해 병렬 처리를 하기도 하는데, 현업에서도 그렇게 할 수 있는 사람이 많지는 않음
	  ◦ 분산과 병렬 처리가 결이 비슷한데, 현업에서는 솔루션 제품을 사용하기도 함
	• 질문: 분산처리 스택 중 특정 스택을 선택하는 작업이 어려웠는데, 현업에서는 어떻게 결정하나요?
	  ◦ 멘토님의 회사에서는 로그 데이터를 수집하기 때문에 ELK 와 Kafka 를 사용. 이때도 분산 기능은 사용하지 않는데, 요즘 솔루션 성능이 좋아서 분산을 이용하지 않아도 가능한 경우가 많음
	  ◦ ELK 사용하면 통계 자료도 쉽게 볼 수 있음
	
	프로젝트 진행 관련 피드백
	
	• 검색 로직을 QueryDSL 과 ELK 스택을 병렬적으로 수행한다는 점에서 여유로워 보임
	• 프론트엔드 로직을 들어보면, 거의 다 끝나가는 것 처럼 보임
	• 프로젝트 완성 후 서로 정보를 공유하는 시간을 가지고 포트폴리오/면접 대비를 하면 도움이 많이 될 것임