

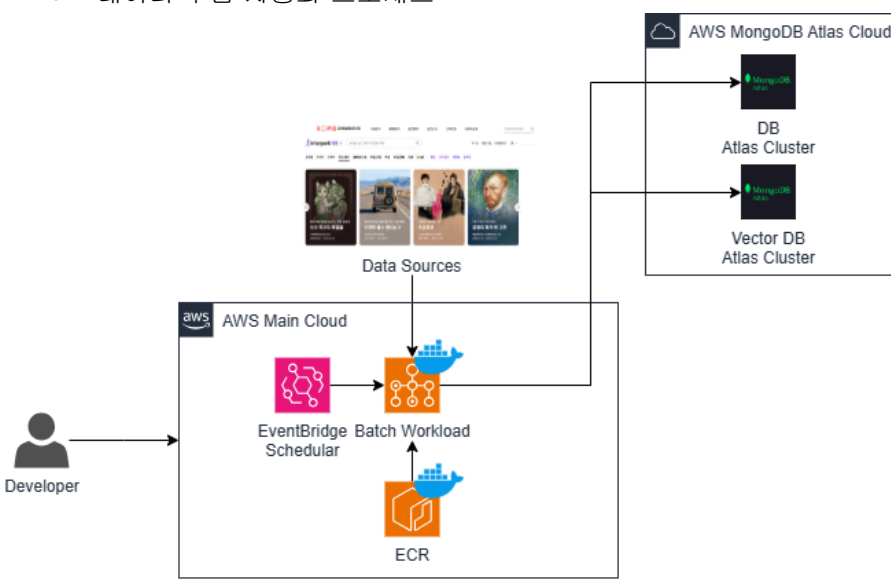
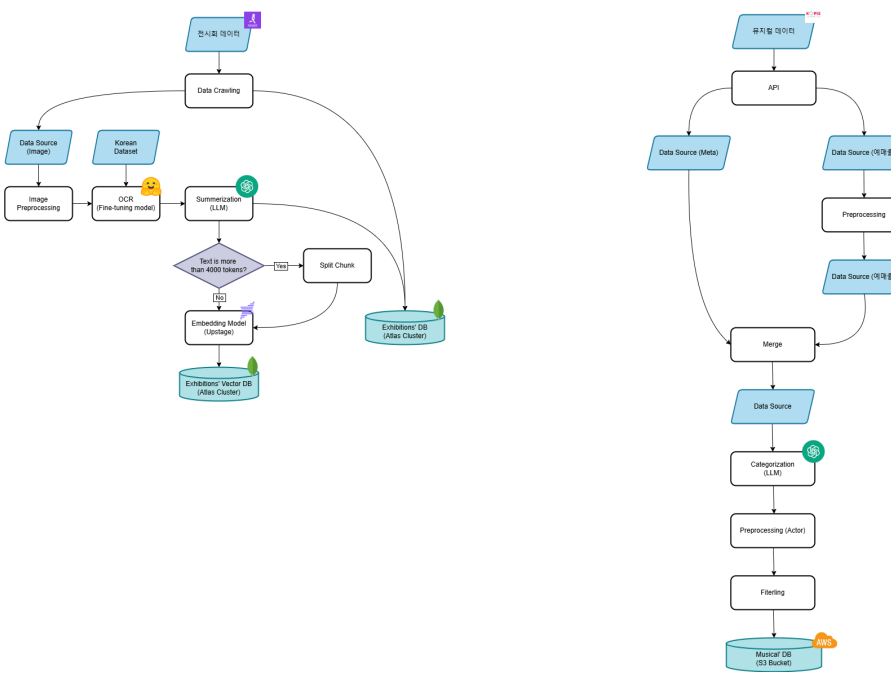
SK네트웍스 Family AI과정 3기

모델링 및 평가 수집된 데이터 및 전처리 문서

□ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 수집된 데이터 및 전처리 문서
- 제출 일자 : 2024.12.26
- 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-2Team>
- 작성 팀원 : 이준석

개요	<ul style="list-style-type: none">• 데이터 설명<ol style="list-style-type: none">1. 전시회 데이터<ol style="list-style-type: none">a. 한국어 데이터셋 (AI Hub)b. 전시회 데이터 (인터파크 티켓 크롤링)c. 사용자 데이터 (Kakao Talk)2. 뮤지컬 데이터<ol style="list-style-type: none">a. 뮤지컬 데이터 (Kopis API)• 데이터 수집목적<ol style="list-style-type: none">1. 전시회 데이터<ol style="list-style-type: none">a. 한국어 데이터셋 (AI Hub) - EasyOCR 모델 파인 튜닝b. 전시회 데이터 (인터파크 티켓 크롤링) - 메타 데이터c. 전시회 데이터 (인터파크 티켓 크롤링) - OCR 후 임베딩d. 사용자 데이터 (Kakao Talk) - 사용자2. 뮤지컬 데이터<ol style="list-style-type: none">a. 뮤지컬 데이터 (Kopis API) - DeepFM 모델 예매율 예측 (배우기반)
----	--

<p>데이터 자동화 및 검증</p>	<ul style="list-style-type: none"> 데이터 수집 자동화 프로세스 
<p>데이터 저장 및 관리</p>	<ul style="list-style-type: none"> 데이터 저장 방식 <ol style="list-style-type: none"> 이벤트 스케줄러로 일주일 마다 자동으로 API 호출 및 크롤링 MongoDB에 메타 데이터 및 임베딩 데이터 적재 일주일이 지나면 새 데이터 추출 및 기간 지난 데이터 삭제
<p>데이터 전처리 과정</p>	<ul style="list-style-type: none"> 전처리 단계 및 방법 설명 

데이터 전처리
결과

- 결과
 1. 사용자에게 맞춤형 데이터 제공
 - a. 전시회 데이터: 이미지 데이터를 분석하여 유사도 기반으로 개인화된 전시회 정보를 추천
 - b. 뮤지컬 데이터: 배우 및 장르를 기반으로 사용자의 선호도를 반영한 뮤지컬 데이터를 추천
 - c. 사용자 경험 개선을 위해 직관적이고 간편한 데이터 탐색 기능 제공
 2. 데이터 활용 확장성
 - a. 전시회와 뮤지컬 외 다른 문화 콘텐츠 데이터로 확장 가능.
 - b. 다양한 언어와 지역에 맞춘 데이터 제공으로 글로벌 사용자 대상 서비스 확장
- 향후 데이터 사용계획
 1. 주기적 업데이트
 - a. 뮤지컬 및 전시회 데이터는 시간과 영향을 받으므로 정기적으로 최신 정보를 반영하여 업데이트
 - b. 데이터의 정확성을 유지하기 위해 신뢰성 있는 데이터 소스를 지속적으로 확보
 2. 서비스 고도화
 - a. AI 및 추천 시스템 성능 향상을 위한 추가 학습 데이터 수집
 - b. 사용자 피드백을 반영하여 추천 알고리즘의 정밀도 개선