Data Engineer 과제

과제 주제

- 법제처에서 제공하는 API를 통해 법령 및 법령 조항 데이터를 관리하는 데이터 파이프라인 구조를 설계합니다.
- 데이터 파이프라인은 다음과 같은 역할을 합니다.
    - 법제처 제공 데이터를 정합성을 유지하여 데이터베이스에 적재
        - 법령 - 법령조항 간의 릴레이션 유지
        - 법령 최신성 유지 (최신 날짜 기반 업데이트)
    - 주기적인 배치작업 수행을 통해 최신 데이터 반영
    - 작업 실패 등 이슈사항에 대한 관리자 알람
    - 기타 지원자가 필요하다고 생각하는 데이터 파이프라인의 기능

과제 주요 평가 사항

- 기본적인 파이프라인 역할을 위한 구현
- 데이터 저장 및 배치작업의 효율적인 파이프라인 및 로직 설계
- 실시간 무중단 서비스를 고려한 데이터 관리 파이프라인 및 로직 설계

대상 데이터

실제 API 연동 없이, 아래와 같이 정의된 2가지 Mock API가 있다고 가정하고 작업을 수행합니다.

- **현행 법령 목록 조회 API**
- **법령 본문 조회 API**

1. **현행 법령 목록 조회 API**

- **엔드포인트 :** `GET /laws`
- **역할:** 현재 시행되고 있는 법령의 목록을 반환합니다.
- **output 예시:** `mock_law_list.json`

2. **법령 본문 조회 API**

- **엔드포인트 :** `GET /law/{법령ID}?promulgationNo={공포번호}`)
- **역할:** `법령ID` 와 `공포번호` 를 파라미터로 받아, 해당 버전의 상세 본문(기본정보, 조문, 부칙, 별표 등) 전체를 반환합니다.
- **output 예시 :** `mock_law_detail_001322_02889.json`

중요 필드는 다음과 같습니다.

- `법령ID` : 각 법령의 고유 ID (e.g., "001322")
- `공포번호` : 법령 개정 시마다 변경되는 번호 (e.g., "02889")
- `시행일자` : 법령 개정시 마다 최신날짜로 변경

데이터 구조 요건

- 수집된 데이터는 `법령` 단위 / `조항` 단위로 선택적 조회하여 활용 가능하도록 설계
- 조회되는 법령 및 조항은 최신 날짜 기준 데이터 활용되도록 설계

사용 기술 스택

- **언어:** Python
- **데이터베이스:** MySQL
- **기타 프레임워크:** Airflow / Kafka 등 기능에 필요 프레임워크 자유 선택

과제 제출물

- 전체 데이터 파이프라인 설계 문서 필수 제출
- e2e 실행 가능한 프로젝트 셋팅보다는 각 컴포넌트 별 간단한 구현 스크립트 제출
    - Python 스크립트 / Docker 파일 등
    - 데이터베이스 생성 및 데이터 적재 스크립트
    - 배치작업 실행 스크립트
    - 기타 지원자가 수행한 결과물