**-아래-** 는 첨부된 **project.tar.gz** 파일의 **README.md** 내용입니다. **project.tar.gz** 파일을 내려받아 과제를 수행해주세요.

**-아래-**

# [데이터Biz캠프] 데이터 엔지니어 지원자 과제

## 1. 평가 항목

1. 요구사항을 분석하고 목적에 부합하도록 데이터를 가공/처리하는 능력
2. 성능을 고려하여 효율적으로 SQL 및 Python 프로그램을 개발하는 능력
3. 컨테이너 (Docker) 기반 개발환경에 대한 이해와 활용 능력
4. 본 과제는 지원자의 과제 풀이 과정을 확인하려는 목적으로, 모든 문제를 구현하지 않아도 됩니다.

## 2. 문제 관련 공통 안내 사항

### 2-1. 문제 요약

카카오뱅크 앱을 사용하면서 발생하는 사용자 로그 데이터(메뉴 로그, 사용자 정보 변경 로그)를 SQL을 이용해 데이터 마트로 구축합니다. 또한, 해당 데이터마트를 정해진 주기에 따라 생성할 수 있는 잡 스케줄러를 Python 으로 개발합니다.

### 2-2. 로그 데이터 설명

#### **메뉴 로그 (KABANG.MENU\_LOG)**

| **컬럼명** | **속성명** | **컬럼유형** | **설명** |
| --- | --- | --- | --- |
| log\_tktm | 로그일시 | String | 메뉴에 접근한 일시 (yyyyMMddHHmmss) |
| log\_id | 로그ID | String | 로그를 식별할 수 있는 unique 값 |
| usr\_no | 사용자번호 | String | 사용자를 식별할 수 있는 값 |

#### **사용자 정보 변경 로그 (KABANG.USR\_INFO\_CHG\_LOG)**

| **컬럼명** | **속성명** | **컬럼유형** | **설명** |
| --- | --- | --- | --- |
| log\_tktm | 로그일시 | String | 메뉴에 접근한 일시 (yyyyMMddHHmmss) |
| log\_id | 로그ID | String | 로그를 식별할 수 있는 unique 값 |
| usr\_no | 사용자번호 | String | 사용자를 식별할 수 있는 값 |
| rsdt\_no | 주민등록번호 | String | 고객의 주민번호로, 7자리 이후에는 "\*"값으로 표기 |
| loc\_nm | 지역명 | String | 현재 거주중인 지역의 이름 |
| mcco\_nm | 이동통신사명 | String | 현재 가입한 통신사의 이름 |

### 2-3. 과제 제출 참고사항

* 첨부된 프로젝트 구조를 참고해 과제물을 제출해주세요.   
  (첨부된 docker-compose.yml 파일은 참고 사항입니다. 꼭 도커 환경이 아니더라도 개인 환경에 맞춰 편하게 구성하시고 실행 방법 등 별도로 설명을 남겨주시면 됩니다.)
  + mysql:8.0.29, python:3.7.13 버전 기반으로 프로젝트를 작성합니다.
  + context-files/mysql/docker-entrypoint-initdb.d 경로에 위치한 초기 데이터셋을 활용합니다.
  + 서드파티 python 패키지를 이용하는 경우 context-files/python/requirements.txt 추가해주세요.
  + 과제물에서 결과물이 있는 경우 result/ 경로에 관리하고, 과제물 소스코드는 assignments/ 경로 하위에 다음과 같은 디렉토리 구조로 작성해주시면 좋습니다.  
    

## 

## 3. 과제

### 3-1. 문제#1

카카오뱅크 앱을 사용하는 모든 고객의 정보를 확인할 수 있는 테이블을 만들고자 합니다. 다음 요구사항을 만족하는 테이블을 생성한 다음(01\_table\_creation.sql), 해당 테이블에 데이터를 적재하는 SQL을 작성하고(02\_data\_insertion.sql), 사용자번호를 기준으로 오름차순 조회한 결과를 기록하세요.(03\_data\_selection.sql, #1\_data.txt)

#### **요구사항**

* 다음과 같은 정보를 포함하는 테이블을 설계합니다. (테이블 및 컬럼명은 자유롭게 작성합니다.)
  + 사용자번호
  + 성별
  + 생년월일
  + 나이
    - 2022.07.27. 일자 기준 만 나이
  + 최초가입일
    - 사용자정보변경로그에 최초로 데이터가 확인되는 시점이 최초 가입시점
  + 통신사명
  + 거주지역명
  + 최빈메뉴
    - 적재되는 컬럼값은 {메뉴명(XX건)} 로 적재해주세요. (e.g. 내카드(10건))
* 사용자 정보는 사용자 정보 변경 로그 에서 각 항목별로 가장 최근에 수정된 정보를 기준으로 합니다.
* 성별, 나이, 거주지역명 등 사용자 정보가 없는 경우에는 해당 값을 "-" 로 적재합니다.

### 3-2. 문제#2

카카오뱅크 앱 사용 현황을 시간대별로 분석하기 위해 스냅샷 테이블을 만들고자 합니다. 다음 요구사항을 만족하는 테이블을 생성해(01\_table\_creation.sql) 해당 테이블에 데이터를 적재하는 SQL을 작성하고(02\_data\_insertion.sql), 기준일자, 사용번호, TOP5메뉴구분을 기준으로 오름차순 조회한 결과를 기록하세요.(03\_data\_selection.sql, #2\_data.txt)

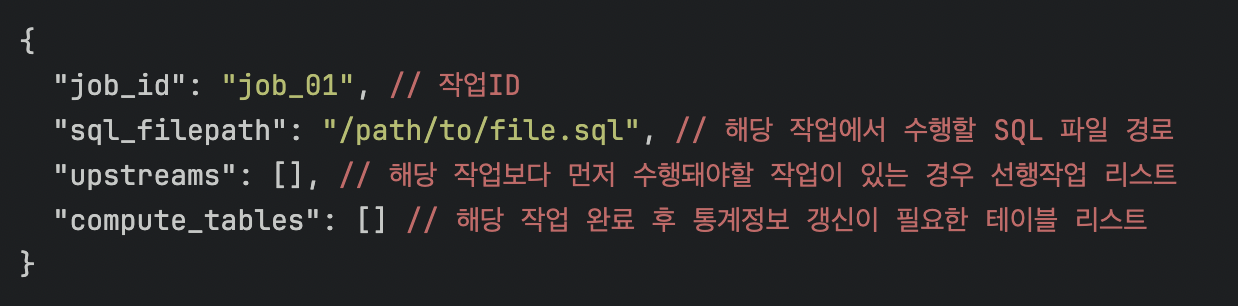
#### **요구사항**

* 다음과 같은 정보를 포함하는 테이블을 설계합니다. (테이블 및 컬럼명은 자유롭게 작성합니다.)
  + 기준일자
  + 사용자번호
  + TOP5메뉴구분
    - TOP5메뉴는 적재된 메뉴로그데이터 중 일별 클릭건수가 높은 5종류 메뉴입니다.
    - 적재되는 값은 다음과 같이 적재되도록 구현해주세요.
      * TOP1\_{메뉴명}
      * TOP2\_{메뉴명}
      * TOP3\_{메뉴명}
      * TOP4\_{메뉴명}
      * TOP5\_{메뉴명}
  + 00\_06시간\_클릭건수
  + 06\_09시간\_클릭건수
  + 09\_13시간\_클릭건수
  + 13\_16시간\_클릭건수
  + 16\_19시간\_클릭건수
  + 19\_21시간\_클릭건수
  + 21\_24시간\_클릭건수
* 메뉴명이 로그인, 로그아웃인경우 집계 대상에서 제외합니다.
* 문제#1에서 생성된 테이블에 존재하지 않는 고객은 적재 대상에서 제외합니다.

### 3-3. 문제#3

문제#1, 문제#2에서 작성한 SQL파일을 바탕으로, 특정 작업 명세에 따라 배치작업을 수행하는 프로그램을 개발하려고 합니다. 아래 요구사항 및 고려사항에 맞게 Python 으로 개발해주세요.

#### **요구사항**

* 파이프라인을 실행하기 위한 공통 설정들은 \*.cfg 파일로 관리합니다.
* 작업에 필요한 정보는 작업별 \*.json 파일로 관리합니다. (e.g. job\_01.json)  
  
* 작업 간의 선/후행 관계가 있는 경우 선행작업들(upstreams)이 정상처리 되었는 지 확인한 뒤 실행될 수 있도록 합니다.
* 선행작업 미완료 시 선행작업이 완료될 때까지 대기해야 합니다.
* 선행작업 체크 주기는 기본값으로 5분, 체크횟수는 기본값으로 60회로 설정합니다.
* 문제#1, 문제#2 데이터 생성 의존관계를 유의하여 선후행 작업을 등록합니다.
* 작업종료시 작업 명세서의 compute\_tables 에 정의된 테이블 통계정보를 생성합니다.
  + e.g. ANALYZE TABLE KABANG.PRODUCT\_SUMMARY;
* 작업시작, 종료 시 배치파이프라인 작업수행에 대한 상태 및 로그정보를 다음과 같은 테이블에 적재합니다.
  + 대상테이블: KABANG.BATCH\_LOG
  + 로그정보
    - base\_day (기준일자)
    - job\_id (작업ID)
    - start\_time (작업시작일시)
    - end\_time (작업종료일시)
    - job\_status (작업상태)
      * RUNNING: 작업시작
      * ERROR: 작업중 오류발생
      * FINISHED: 작업완료
* 파이프라인 작업ID별 수행되는 쿼리 순서 및 소요시간에 대한 정보를 테이블에 적재합니다.
  + 대상테이블: KABANG.BATCH\_LOG\_DETAIL
  + 로그정보
    - base\_day (기준일자)
    - job\_id (작업ID)
    - order\_number (실행순서)
    - elapsed\_time (작업소요시간)

#### 

#### 

#### **고려 사항**

* 절차적 프로그래밍(PP)을 지양하고 객체지향 프로그래밍으로 작성해주세요. (OOP)
* 프로그램 코드에 대한 신뢰성을 높이기 위해 테스트코드를 작성하고, 그 결과를 확인할 수 있어야 합니다. (Testing)
* docstring, typing, comment 를 최대한 활용해주세요. (Documentation)
* RuntimeError 등 프로그램에서 발생할 수 있는 예외 상황을 고려해 작성해주세요. (Exception Handling)