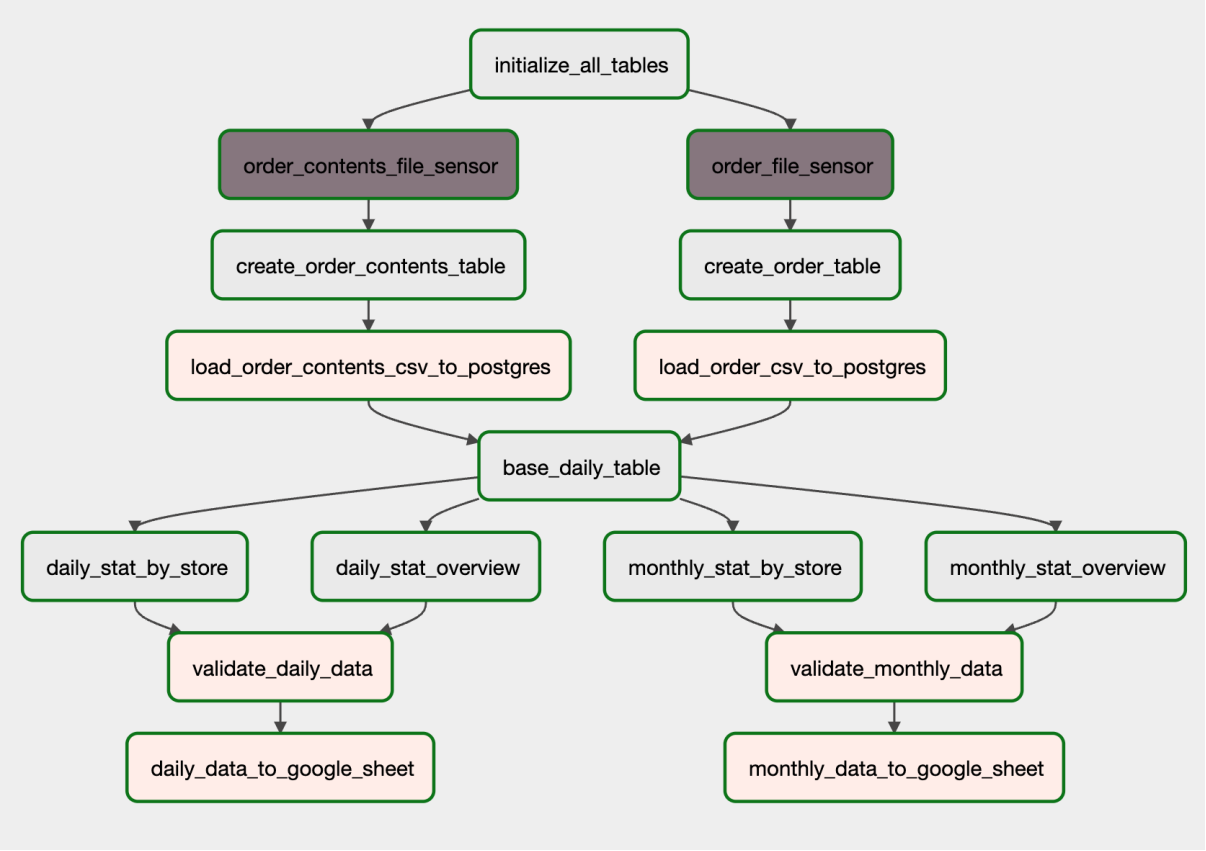
데이터 파이프라인 요약 - 신상훈

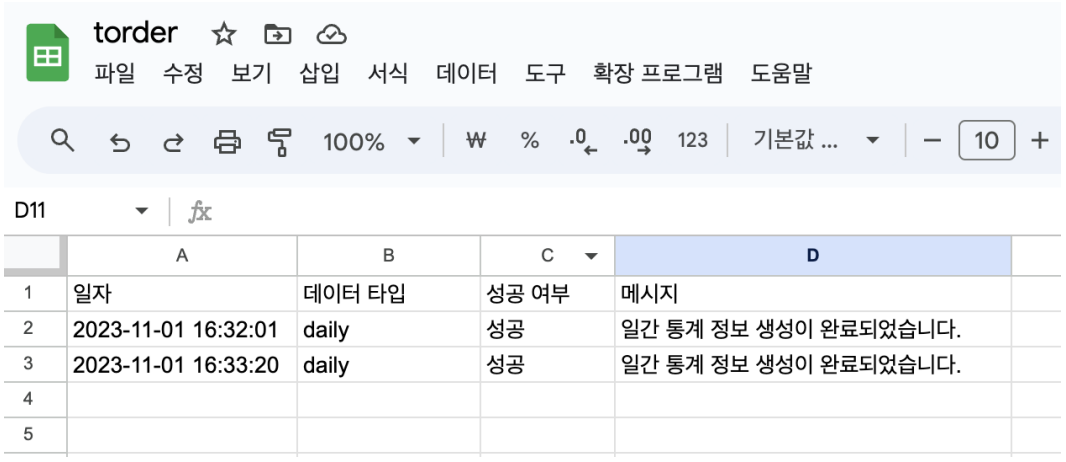
데이터 파이프라인

Airflow DAG



- `FileSensor`를 활용하여 csv 파일이 있는지 확인하여 파일이 있으면 DB로 적재를 하도록 하였습니다. csv 데이터를 적재하기 위한 테이블은 만들고 적재하는 과정을 daily, monthly로 나누어서 파이프라인을 설계하였습니다.
- 데이터가 전부 적재되고 난 후에 일별, 월별 집계를 위한 데이터 처리를 진행합니다. `PostgresOperator`를 사용하여 진행하였으며 일별로 집계된 테이블을 기준으로 파생 테이블을 생성하도록 파이프라인을 설계하였습니다.
- 에러률은 아래와 같은 공식으로 계산하였습니다.
  - 에러률: 에러 주문 수 / 전체 주문 수
- 만들어진 테이블을 대상으로 validation 작업을 파이프라인에 포함시켰습니다. 실제로 validation 작업을 하는 코드를 구현하지는 않았습니다.

- validation 작업이 완료된 이후에는 gspread 라이브러리를 활용하여 구글 시트에 작업 날짜와 메시지를 입력할 수 있게 하였습니다.



	A	B	C	D
1	일자	데이터 타입	성공 여부	메시지
2	2023-11-01 16:32:01	daily	성공	일간 통계 정보 생성이 완료되었습니다.
3	2023-11-01 16:33:20	daily	성공	일간 통계 정보 생성이 완료되었습니다.
4				
5				

- (과제를 위한) 재실행 편의를 위해 테이블을 초기화(테이블 삭제) 기능을 파이프라인 제일 앞에 넣었습니다.
- 스케쥴링에 관한 부분은 따로 요구사항에 없어 임의로 입력하였습니다.

> 💡 실제 비슷한 업무를 하게 되면 고려해야할 부분이 더욱 많겠지만, 최대한 과제에서 설명한 내용과 유사하게 작업을 하려고 하였습니다.

Table Info

daily_stat

	store_code	business_day	order_count	error_count
1	store_1202	2022-11-18	87	0
2	store_1202	2022-11-04	85	0
3	store_1345	2022-11-19	83	0
4	store_1202	2022-11-12	81	0
5	store_1202	2022-11-25	81	0

- store_code: 상점 코드
- business_day: 영업일(7:00AM ~ 7:00AM 기준으로 변환)
- order_count: 주문 수
- error_count: 에러 수

daily_stat_overview

	business_day	order_count_all	error_count_all	error_rate_all
1	2022-11-19	20,633	13	0.001
2	2022-11-26	19,991	21	0.001
3	2022-11-05	19,352	14	0.001
4	2022-11-12	19,048	17	0.001
5	2022-11-18	17,242	20	0.001

- business_day: 영업일(7:00AM ~ 7:00AM 기준으로 변환)
- order_count_all: 일별 주문 수
- error_count_all: 일별 에러 수
- error_rate_all: 에러률 (에러 수/주문 수)

daily_stat_by_store

	store_code	business_day	order_count	error_count	error_rate
1	store_1202	2022-11-18	87	0	0
2	store_1202	2022-11-04	85	0	0
3	store_1345	2022-11-19	83	0	0
4	store_1202	2022-11-12	81	0	0
5	store_1202	2022-11-25	81	0	0

- store_code: 상점 코드
- business_day: 영업일(7:00AM ~ 7:00AM 기준으로 변환)
- order_count: 일별 주문 수
- error_count: 일별 에러 수
- error_rate: 에러률 (에러 수/주문 수)

monthly_stat_overview

	business_month	order_count_all	error_count_all	error_rate_all
1	2022-11-01	380,751	327	0.001

- business_month: 영업월
- order_count_all: 월별 주문 수
- error_count_all: 월별 에러 수
- error_rate_all: 에러률 (에러 수/주문 수)

monthly_stat_by_store

	store_code	business_month	order_count	error_count	error_rate
1	store_914	2022-11-01	45	0	0
2	store_1029	2022-11-01	119	0	0
3	store_1325	2022-11-01	230	0	0
4	store_2426	2022-11-01	9	0	0
5	store_846	2022-11-01	461	0	0

- store_code: 상점 코드
- business_month: 영업월
- order_count: 일별 주문 수
- error_count: 일별 에러 수
- error_rate: 에러률 (에러 수/주문 수)