DM_retention_cohort - 설계노트 (Version v1 scope)

1, 목적**(Why)**

가입 코호트(cohort_month)별로 가입 후 **day 0~180** 동안의 **Retention curve**를 표준화해서 만든다.

Active 정의를 “해당 day에 세션이 1개 이상”으로 고정해, 코호트 비교(월별) 및 Story의 장기 잔존/리듬 분석에 재사용한다.

2, Grain

1 cohort_month × 1 day_index = 1 row

day_index는 signup_date 기준 경과일(0..180) 이다.

3, Input tables & Join key

- **users** - 코호트 기준 테이블(signup_date → cohort_month), key - **user_id**
- **sessions** - Active 판정용(해당 날짜 세션 존재), key - **user_id**
- **Join key**: sessions.user_id = users.user_id
- 참고**(**스캔 최적화**)**: bounds(min_signup, max_signup)로 sessions를 **min_signup ≤ session_date < max_signup + 180**일로 먼저 필터한 뒤, 유저별 window 조건을 다시 적용한다.

4, Partition / Clustering

- **PARTITION BY :** cohort_month 코호트(가입 월) 단위 조회/비교가 핵심
- **CLUSTER BY :** day_index day_index(0~180) 구간 필터/집계가 잦음(리텐션 커브 조회)

5, Window 정의

- 기준일**: signup_date**
- 전처리 **(global filter,** 스캔 절감 목적**)**
    - **min_signup ≤ session_date < max_signup + 180**일 범위로 sessions를 먼저 필터

- 유저별 **window(**최종 집계 기준**)**
    - **0–180d: signup_date ≤ session_date < signup_date + 181**일 → **day 180** 포함
    - day_index = DATE_DIFF(session_date, signup_date, DAY)
- **Active** 정의: 해당 day_index에 세션이 **1**개 이상이면 **active**

6, Main Features and 계산 로직

A. 코호트 정의

- **cohort_month**: DATE_TRUNC(signup_date, MONTH)

B. 코호트 사이즈

- **cohort_size**: cohort_month별 users 수
    - Logic: COUNT(*) GROUP BY cohort_month

C. 활성 유저(일 단위)

- **user_active_days**: 유저별 active day_index 생성
    - Logic: sessions 유저별 window로 필터 후, DISTINCT(user_id, cohort_month, day_index)
- **active_users**: cohort_month × day_index별 active 유저 수
    - Logic: COUNT(DISTINCT user_id) GROUP BY cohort_month, day_index

D. day spine(0~180 채우기)

- day_index 0..180을 생성 후 cohort별로 cross join
- missing day는 active_users=0으로 채움 (retention curve가 끊기지 않게)

E. 최종 Retention

- **retention_rate** = active_users / cohort_size
    - Logic: SAFE_DIVIDE(IFNULL(active_users,0), cohort_size)

7. Sanity checks

- PK 유일성: COUNT(*) == COUNT(DISTINCT CONCAT(cohort_month, '-', day_index))

- 범위 체크: day_index는 **0~180**만 존재해야 함
- 값 범위:
    - 0 ≤ active_users ≤ cohort_size
    - 0 ≤ retention_rate ≤ 1

- 코호트 사이즈 합: SUM(cohort_size)가 전체 users 수와 일치(중복/누락 확인)
- 스파인 정상 동작: 각 cohort_month마다 day_index가 **0..180(181**행**)** 전부 존재하는지

8. 이 **DM**이 꼭 필요한지

- 필요 **-** 코호트별 **Retention curve**를 **Active=**세션 **1**회 이상으로 고정 정의해두면**,** **Story/**분석에서 장기 잔존 비교를 반복 재사용할 수 있다**.**