

AWS - Data Clean Room

데이터를 보지 않고 분석할 수 있게 해주는 도구

배경

카드사

누가 얼마 썼는지 데이터



데이터 교환 불가

유통사

누가 광고를 봤는지 데이터

광고가 진짜 효과 있었는지 보려면 두 데이터를 사람 단위로 붙여야 함.
그런데 privacy 때문에 그냥 데이터를 공유할 수가 없음.

AWS란?

아마존이 빌려주는 서버와 저장소. 서버실을 따로 두지 않고 여기에 데이터를 올려둬.

여러 명이 접속해서 SQL 돌리는 공유 DataBase와 거의 같음.

S3	파일 보관소. CSV 데이터를 여기 올린다.
-----------	--------------------------

Table	올린 파일을 테이블로 등록한다. 이때부터 SQL로 조회된다.
--------------	-----------------------------------

Clean Rooms	여러 회사의 테이블을 서로 안 보여주고 SQL로 접근만 가능해진다.
--------------------	---------------------------------------

AWS에서 데이터를 공유하지 않고 분석할 수 있는 도구를 제공함.

데이터는 각자 계정에 그대로 있고 공유하지 않음. 출력이 되는건 코딩의 결과뿐임.

Clean Room – 데이터 이동 금지

분석 환경이 양쪽 데이터를 읽기만 함. 원본은 각자 계정 안에 그대로 남음.



Clean Room – 데이터 출력 금지 규정

보통 DB는 SELECT * 한 줄이면 전체가 출력됨. 여기서는 세 겹으로 막혀 있음.

1

SELECT * 를 아예 차단

행 자체를 출력하는 쿼리는 실행이 거부됨.

COUNT · SUM · AVG 같은
통계량만 출력할 수 있음.

2

최소 인원 임계값

"결과를 보려면 최소 100명
이상의 결과만 출력 가능"
조건을 검.

소수의 결과는 시스템이
출력을 금지함.

3

노이즈 프라이버시

결과 숫자에 노이즈를 섞음.

결과로 10,000이 나와도
진짜 결과는 10,001이거나
9,999일 수 있음.

Clean Room – 각 변수별 지정한 역할만 수행 가능

데이터 주인이 컬럼마다 쓸 수 있는 역할을 미리 지정함. 그 역할 밖으로는 아예 못 씀.

컬럼	지정된 역할 (SQL)	Python pandas	결과 출력
이메일	JOIN 조건에만 사용	merge	절대 출력 불가
금액	SUM · AVG · COUNT	agg	집계값으로만
카테고리	GROUP BY · WHERE	groupby, if	그대로 출력

역할 밖으로 나가면 실행 거부

- ✗ `SELECT * FROM 카드사`
df 원본 통째로
- ✗ `SELECT 이메일 FROM 카드사`
df['이메일']
- ✗ `... LIMIT 10`
df.head(10)
- ✗ 서브쿼리 · `WITH` · `UNION`
tmp = df[df.email > 2] , pd.concat([...])

역할을 벗어난 쿼리는 실행 전 문법 단계에서 거부됨. 사람이 쿼리를 하나씩 검토할 필요가 없음.

Clean Room – 출력 결과

카테고리	구매자 수	매출 합계
식품	4,210	1억 8,205만
가전	1,038	9,640만
명품	—	—

← 실제로는 1억 8,207만

← 실제로는 87명. 100명 미만이라
시스템이 줄짜로 지웠음.

개인 단위 정보는 이 표에 들어올 자리가 없음. 출력 가능한 형태가 집계표뿐임.
어떤 쿼리를 돌렸는지 데이터 주인이 로그로 전부 확인 가능함.

Clean Room – Python · ML 실행 방식

SQL 말고 Python 코드나 ML 모델을 돌리는 방법은 두 가지임.

1. AWS가 제공하는 모델 사용하기

- 현재는 유사고객(lookalike) 찾기 모델만 제공함

2. 내가 만든 모델을 상대방에게 승인받기

- 코드를 py파일로 S3에 업로드함
- 데이터 주인이 코드를 읽고 승인해야 실행됨

SQL은 문법에 자동 차단조건을 걸고, Python·ML은 상대가 승인한 코드만 사용이 가능함. 막는 방식이 다름.

Clean Room – 실행 절차

1. 협업(Collaboration) 생성

상대 AWS 계정 ID 입력. 최대 5곳

2. 데이터 업로드

각자 자기 S3에. 외부로 나가지 않음

3. 변수별 역할 지정

집계 허용 · 최소 인원 임계값 등

상대가 실행 가능한 코드 범위가 여기서 전부 결정됨

4. Table 연결

생성한 협업에 각자 table을 연결함

5. 쿼리 또는 코드 실행

결과는 사전 지정된 수신자에게만 전달됨