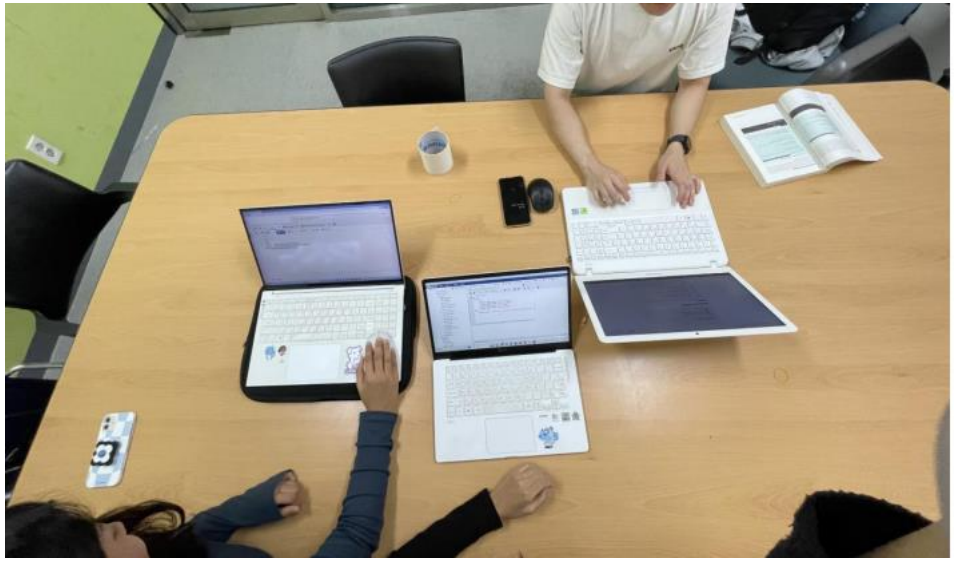
# 데이터 분석을 위한 SQL 레시피 스터디

2022.10.31

발표자 : 최지원

# 스터디원 소개

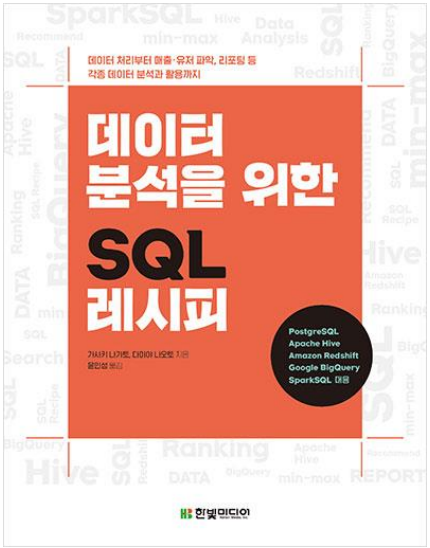


- 응용통계학과 곽수민
- 산업보안학과 남궁규민
- 산업보안학과 최지원

# 스터디 진행 방식

**선정 도서: <데이터 분석을 위한 SQL 레시피>**

단순 분석서가 아닌, 실무에서 사용한 코드를 기반으로 한 지침서
데이터 가공과 매출 파악, 사용자 파악, 웹사이트 내 사용자 행동 파악, 이상수치 검출 등



# 스터디 진행 방식

1. **주에 챕터 2개씩 각자 공부 후 공유 노션에 정리**

2. **주 1회 대면 스터디 진행**

   랜덤으로 정리한 내용 발표 + 각자 어려웠던 부분 공유 + 다른 방법으로 작성한 코드 공유


- **공유 노션**

## URL에서 요소 추출하기

### 레퍼러 분석

- 어떤 웹페이지를 거쳐 넘어왔는지 판별하기 위함 ⇒ 호스트 단위 집계가 일반적

```
-[ RECORD 1 ]----------------------------------------------------
stamp    | 2016-08-26 12:02:00
referrer | http://www.other.com/path1/index.php?k1=v1&k2=v2#Ref1
url      | http://www.example.com/video/detail?id=001
-[ RECORD 2 ]----------------------------------------------------
stamp    | 2016-08-26 12:02:01
referrer | http://www.other.net/path1/index.php?k1=v1&k2=v2#Ref1
url      | http://www.example.com/video#ref
-[ RECORD 3 ]----------------------------------------------------
stamp    | 2016-08-26 12:02:01
referrer | https://www.other.com/
url      | http://www.example.com/book/detail?id=002
```

```sql
SELECT
  stamp,
  , substring(referrer from 'https?://([^/]*)') AS referrer_host
FROM
  access_log
;
```

정규표현식을 통해 추출한 문자열을 기존 문자열에서 나눔
http:// 또는 https:// 다음으로 /가 오기 전까지 추출 ⇒ 레퍼러 부분 잡힘

> 문자열 분할; substring 시 정규표현식 사용
>
> substring( <문자열 칼럼> from <정규표현식>) AS <값>

### 정수형 자료형의 데이터 나누기

- CTR(Click Through Rate): 클릭 / 노출 수

> CAST 함수 : 형 변환
> CAST(컬럼명, 값 AS 변경하려는 TYPE명)
> double precision: 8바이트까지 표현

무엇이든 임베드하세요(PDF, Google Docs, Google Maps, Spotify...)

```sql
SELECT
dt
,ad_id
, clicks / impressions AS ctr
, 100.0 * clicks/ impressions AS ctr_percentage
FROM `ch3.advertising_stats`
WHERE dt = '2017-04-01'
ORDER BY dt, ad_id;
```

### 0으로 나누는 것 피하기

- 2017-04-02 데이터의 경우 impressions가 0임

⇒ 0으로 나누게 되어 오류가 발생할 수 있음

1. CASE 식 활용
2. NULL 전파 → NULLIF()활용
   NULL을 포함한 데이터의 연산 결과가 모두 NULL이 되는 SQL 성질

```sql
SELECT
dt
,ad_id
```

## 새로운 지표 정의하기

- 페이지 뷰 : 페이지가 출력된 횟수
- 방문자 수 : 페이지를 출력한 사용자 수
- 페이지뷰 / 방문자수 : 사용자 한 명 당 방문하는 페이지 수
- 이 외에도 CTR(Click Through Rate), CVR(Conversion Rate) 등등 새로이 정의 가능

### 1. 문자열 연결하기

- 서울시, 강서구라는 컬럼이 있으면 이 두 컬럼을 연결

```sql
select user_id,
  concat(pref_name, city_name) as pref_city
from mst_user_location
```

- CONCAT 함수 대신에 || 연산자로 붙일 수도 있음(파이썬의 문자열 +와 같은 기능)

### 2. 여러 개의 값 비교하기

- 분기별 매출 증감 판정

```sql
select year, q1, q2,
  case when q1 < q2 then '+'
  when q1 = q2 then ''
  else '-' end as judge_q1_q2,
  q2 - q1 as diff_q2_q1,
  sign(q2-q1) as sign_q2_q1
from quarterly_sales

-- case when으로 판정해도 되고, SIGN 함수로 판정해도 됨
```

## ▪ 공유 노션

# 스터디 진행 (9/27)



**<주요 문법>**

CASE WHEN ; 조건식

SUBSTRING ; 문자열 분리

CAST ; 자료의 형 변환

EXTRACT ; 날짜/시각에서 특정 필드 추출

COALESCE ; 결측 값 대체

# 스터디 진행 (10/03)



**<주요 문법>**

CONCAT; 문자열 연결

SIGN ; 값의 범위에 따라 -1,0,1 반환

NULLIF ; 특정 값인 경우 결측치로 전환

CURRENT_DATE ; 현재 날짜 반환

FLOOR ; 소수점 버림

# 스터디 진행 (10/03)



**<주요 문법>**

**윈도우 함수**

```
SELECT WINDOW_FUNCTION (ARGUMENTS) OVER
( [PARTITION BY 컬럼] [ORDER BY 컬럼] [WINDOWING 절] )
FROM 테이블명 ;
```

**순위 관련 함수**

RANK, DENSE_RANK, ROW_NUMBER

STRING_AGG ; 데이터 구분하여 문자열로 합침

UNNEST ; 배열을 레코드 분할하여 반환