

## 데이터 분석을 위한 SQL 레시피 스터디

2022.10.31

발표자 : 최지원

## 스터디원 소개

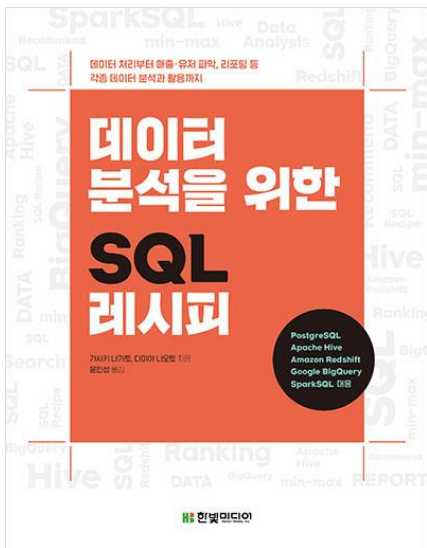


- 응용통계학과 곽수민
- 산업보안학과 남궁규민
- 산업보안학과 최지원

# 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



### 데이터 가공을 위한 SQL

#### 5장 하나의 값 조작하기 ...52

- 1 코드 값을 레이블로 변경하기 ...53
- 2 URL에서 요소 추출하기 ...55
- 3 문자열을 배열로 분해하기 ...58
- 4 날짜와 타임스탬프 다루기 ...59
- 5 결손 값을 디폴트 값으로 대체하기 ...64

#### 6장 여러 개의 값에 대한 조작 ...66

- 1 문자열 연결하기 ...67
- 2 여러 개의 값 비교하기 ...68
- 3 2개의 값 비를 계산하기 ...73
- 4 두 값의 거리 계산하기 ...77
- 5 날짜/시간 계산하기 ...79
- 6 IP 주소 다루기 ...87

#### 7장 하나의 테이블에 대한 조작 ...92

- 1 그룹의 특징 찾기 ...93
- 2 그룹 내부의 순서 ...97
- 3 새로 기반 데이터를 가로 기반으로 변환하기 ...106
- 4 가로 기반 데이터를 세로 기반으로 변환하기 ...109



### 매출을 파악하기 위한 데이터 추출

#### 9장 시계열 기반으로 데이터 집계하기 ...138

- 1 날짜별 매출 집계하기 ...139
- 2 이동평균을 사용한 날짜별 추이 보기 ...141
- 3 당월 매출 누계 구하기 ...143
- 4 월별 매출의 적대비 구하기 ...148
- 5 2 차트로 업적의 추이 확인하기 ...150
- 6 매출을 파악할 때 중요 포인트 ...157

#### 10장 다양한 축을 사용해 데이터 집계하기 ...161

- 1 카터그리브 매출과 소계 계산하기 ...162
- 2 ABC 분석으로 잘 팔리는 상품 판별하기 ...166
- 3 편 차트로 상품의 매출 증가를 확인하기 ...169
- 4 히스토그램으로 구매 가격대 집계하기 ...174



### 사용자를 파악하기 위한 데이터 추출

#### 11장 사용자 전체의 특징과 경향 찾기 ...186

- 1 사용자의 액션 수 집계하기 ...188
- 2 연령별 구분 집계하기 ...194

# 스터디 진행 방식

1. 주에 챕터 2개씩 각자 공부 후 공유 노선에 정리

2. 주 1회 대면 스터디 진행

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

데이터 분석을 위한 SQL... / 연구과정

## 연구과정

CUAI +

남궁규민 2

4~5강

6~7강

+ 새로 만들기

최지원 2

4~5강

6~7강

+ 새로 만들기

곽수민 3

4~5강

6~7강

8~9강

+ 새로 만들기

▪ 공유 노선

## 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
```

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

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

💡 문자열 분할: substring 시 정규표현식 사용

substring(<문자열 컬럼> 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...)

```
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

- CASE 식 활용
- NULL 전파 → NULLIF()활용

NULL을 포함한 데이터의 연산 결과가 모두 NULL이 되는 SQL 성질

```
SELECT
  dt
  , ad_id
```

## 새로운 지표 정의하기

- 페이지 뷰: 페이지가 출력된 횟수
- 방문자 수: 페이지를 출력한 사용자 수
- 페이지뷰방문자수: 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

```
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

```
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 노선

## 스터디 진행 (9/27)

### 4강 데이터 ...39

- 1 데이터의 종류 ...39
- 2 업무 데이터 ...41
- 3 로그 데이터 ...44
- 4 두 데이터를 사용해서 생성되는 값 ...47

### 5강 하나의 값 조작하기 ...52

- 1 코드 값을 레이블로 변경하기 ...53
- 2 URL에서 요소 추출하기 ...55
- 3 문자열을 배열로 분해하기 ...58
- 4 날짜와 타임스탬프 다루기 ...59
- 5 결손 값을 디폴트 값으로 대체하기 ...64

### <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)

### 6강 여러 개의 값에 대한 조작 ...66

- 1 문자열 연결하기 ...67
- 2 여러 개의 값 비교하기 ...68
- 3 2개의 값 비율 계산하기 ...73
- 4 두 값의 거리 계산하기 ...77
- 5 날짜/시간 계산하기 ...79
- 6 IP 주소 다루기 ...87

### 7강 하나의 테이블에 대한 조작 ...92

- 1 그룹의 특징 잡기 ...93
- 2 그룹 내부의 순서 ...97
- 3 세로 기반 데이터를 가로 기반으로 변환하기 ...106
- 4 가로 기반 데이터를 세로 기반으로 변환하기 ...109

#### <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)

### 6강 여러 개의 값에 대한 조작 ...66

- 1 문자열 연결하기 ...67
- 2 여러 개의 값 비교하기 ...68
- 3 2개의 값 비율 계산하기 ...73
- 4 두 값의 거리 계산하기 ...77
- 5 날짜/시간 계산하기 ...79
- 6 IP 주소 다루기 ...87

### 7강 하나의 테이블에 대한 조작 ...92

- 1 그룹의 특징 잡기 ...93
- 2 그룹 내부의 순서 ...97
- 3 세로 기반 데이터를 가로 기반으로 변환하기 ...106
- 4 가로 기반 데이터를 세로 기반으로 변환하기 ...109

### <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