codecademy

Capstone Project: Attribution Queries

Learn SQL from Scratch

Thanh Pham

March 26, 2019

Example Table of Contents

1. Get familiar with CoolTShirts

1. Get familiar with the company, slide 1

- How many campaigns and sources does CoolTShirts use?

The table below shows that the CoolTShirt used **8 campaigns** and **6 sources**. See query 1 and query 2 (on the right of this slide). In both queries,

- I used the DISTINCT keyword on *utm_campaign* and *utm_source* columns to query distinct value on each column.
- I also used the COUNT aggregate function to count the number of rows return in the new result set for each query.

COUNT(DISTINCT utm_campaign)	COUNT(DISTINCT utm_source)
8	6

```
-- query 1: find the number of campaigns that CoolTShirt use

SELECT COUNT(DISTINCT utm_campaign)
FROM page_visits;

-- query 2: find the number of sources that CoolTShirt use

SELECT COUNT(DISTINCT utm_source)
FROM page_visits;
```

1. Get familiar with the company, slide 2

- How campaigns and sources are related? Be sure to explain the difference between `utm_campaign` and `utm_source`.

utm_campaign	utm_source
ten-crazy-cool-tshirts-facts	buzzfeed
weekly-newsletter	email
retargetting-campaign	email
retargetting-ad	facebook
paid-search	google
cool-tshirts-search	google
interview-with-cool-tshirts-founder	medium
getting-to-know-cool-tshirts	nytimes

/* the query below illustrates how to find which source is used for each campaign:

- I used DISTINCT on utm_campaign, utm_source columns because I want the query to return distinct values of the utm_campaign and the utm_source column
- The query result is show on the left of this slice
- There are 8 distinct campaigns. Each source can have multiple campaign ads */

```
SELECT DISTINCT utm_campaign, utm_source
FROM page_visits
ORDER BY utm_source;
```

1. Get familiar with the company, slide 3

- What pages are on their website?

The table below illustrates the result set of 4 distinct pages are on their website.

- The query on the right used keyword DISTINCT on the column page_name to return distinct value of the page_name column.

page_name
1 - landing_page
2 - shopping_cart
3 - checkout
4 - purchase

```
/*the query below used DISTINCT on page_name column to return distinct values of the page_name column
*/

SELECT DISTINCT page_name
FROM page_visits;
```

2. What is the user journey?

2. What is the user journey?, section 2.1

How many first touches is each campaign responsible for?

- The query is show on the right of this slice.
  - On the first part of the query is first_touch of ALL USERS in the page_visits table (the result of the GROUP BY user_id query).
  - On the second part of the query, I JOIN the first_touch table back to the page_visit table so that I can add the COUNT utm_compaign to count number of first touches for each campaign (result of another GROUP BY utm_campaign query)
- The table below shows number of first touches is responsible by each campaign

utm_campaign	COUNT(utm_campaign)
cool-tshirts-search	169
ten-crazy-cool-tshirts-facts	576
getting-to-know-cool-tshirts	612
interview-with-cool-tshirts-founder	622

```
/* the query below find the number of first touches for each campaign
*/
WITH first_touch AS (
   SELECT user_id,
      MIN(timestamp) as first_touch_at
   FROM page_visits
   GROUP BY user_id),
   ft_attr AS(
SELECT ft.user_id,
   ft.first_touch_at,
   pv.utm_source,
      pv.utm_campaign
     FROM first_touch ft
JOIN page_visits pv
   ON ft.user_id = pv.user_id
   AND ft.first_touch_at = pv.timestamp
order by utm_campaign
    )
SELECT utm_campaign, COUNT(ft_attr.user_id) AS numb_ft
```

2. What is the user journey?, section 2.2

- How many last touches is each campaign responsible for?
  - The table below shows the result set of the query on the right.

utm_campaign	numb_lt
cool-tshirts-search	60
paid-search	178
interview-with-cool-tshirts-founder	184
ten-crazy-cool-tshirts-facts	190
getting-to-know-cool-tshirts	232
retargetting-campaign	245
retargetting-ad	443

```
/* the query below find the number of last touches for
each campaign. The query is the same from previous slice
except we change MIN(timestamp) to MAX(timestamp).
*/
WITH last_touch AS (
    SELECT user_id,
        MAX(timestamp) as last_touch_at
    FROM page_visits
    GROUP BY user_id),
    lt_attr AS(
SELECT lt.user_id,
    lt.last_touch_at,
    pv.utm_source,
        pv.utm_campaign
       FROM last_touch lt
JOIN page_visits pv
    ON lt.user_id = pv.user_id
    AND lt.last_touch_at = pv.timestamp
order by utm_campaign
      )
SELECT utm_campaign, COUNT(lt_attr.user_id) AS
```

2. What is the user journey?, section 2.3

- How many visitors make a purchase?

The table below is the result set of the query shows on the right of this slice. There are 361 visitors made their purchase.

COUNT(DISTINCT user_id)	page_name
361	4 - purchase

```
-- the query below finds the number of visitors make a purchase

SELECT COUNT(DISTINCT user_id), page_name
FROM page_visits
WHERE page_name = '4 - purchase';
```

2. What is the user journey?, section 2.4

- How many last touches *on the purchase page* is each campaign responsible for?
- The table below show the source, the campaign ad and the number of users made their purchases.

utm_source	utm_campaign	numb_lt
google	cool-tshirts-search	2
medium	interview-with-cool-tshirts-founder	7
nytimes	getting-to-know-cool-tshirts	9
buzzfeed	ten-crazy-cool-tshirts-facts	9
google	paid-search	52
email	retargetting-campaign	54
facebook	retargetting-ad	113

```
/*query below  find  the number of last touches on the purchase page for each campaign.  I added the WHERE to last_touch table query.  The result set is show on the left table.
*/
WITH last_touch AS (
    SELECT user_id,
        MAX(timestamp) as last_touch_at
    FROM page_visits
    WHERE page_name = '4 - purchase'
    GROUP BY user_id),
    lt_attr AS(
SELECT lt.user_id,
    lt.last_touch_at,
    pv.utm_source,
        pv.utm_campaign
       FROM last_touch lt
JOIN page_visits pv
    ON lt.user_id = pv.user_id
    AND lt.last_touch_at = pv.timestamp
order by utm_campaign
      )
SELECT utm_campaign, COUNT(lt_attr.user_id) AS numb_lt
FROM lt_attr
GROUP BY utm_campaign
```

2. What is the user journey?, section 2.5

- What is the typical journey?

As I looked back the journey, I have several opinions:

- users came to the CoolTShirts website through multiple campaigns ads as indicated in the result set tables for first_touch attribution and last_touch attribution queries in the slice 8 and 9, specifically, part 2, section 2.1 and section 2.2. The table in slice 8 show number of first touches is responsible by each campaign and the table in slice 9 shows the number of last touches for each campaign.
- The total number of users came to the website was very high if we add all the number from first touches and last touches together but only 361 users did make a final purchases. See the result set table in the slice 10, section 2.3 for details.
- The result set table in the slice 11, section 2.4 indicated the number of purchases from lowest to highest depend on the source that ran the campaign. After analyzed the result table data, I conclude that email and facebook are the two sources where visitors are drawn back to a website, especially for making a final purchase.

3. Optimize the campaign budget?

3. Optimize the campaign budget, section 3.1

CoolTShirts can re-invest in 5 campaigns. Which should they pick and why?

CoolTShirts can re-invest in 5 campaigns listed below:

1. Weekly-newsletter
2. Retargetting-ad
3. Retargetting-campaign
4. Paid-search
5. Getting-to-know-cool-tshirts

Why?

Please see the explanation in the next slice

utm_source	utm_campaign	numb_lt
google	cool-tshirts-search	2
medium	interview-with-cool-tshirts-founder	7
nytimes	getting-to-know-cool-tshirts	9
buzzfeed	ten-crazy-cool-tshirts-facts	9
google	paid-search	52
email	retargetting-campaign	54
facebook	retargetting-ad	113
email	weekly-newsletter	115

3. Optimize the campaign budget, section 3.2

CoolTShirts can re-invest in 5 campaigns. Which should they pick and why?

Why?

- The right table shows the last touch attribution that the last source for each customer. So those campaign listed above indicated high number of customers making a final purchase.
- Although, the “getting-to-know-cool-tshirts” and “ten_crazy-cool-tshirts-facts” campaigns have the same number of purchase is 9. But I chose “getting-to-know-cool-tshirts” campaign because the first touch attribution table in slice 8 and the last touch attribution table in slice 9 show that it has higher number of initial visit and last visit from customers.

utm_source	utm_campaign	numb_lt
google	cool-tshirts-search	2
medium	interview-with-cool-tshirts-founder	7
nytimes	getting-to-know-cool-tshirts	9
buzzfeed	ten-crazy-cool-tshirts-facts	9
google	paid-search	52
email	retargetting-campaign	54
facebook	retargetting-ad	113
email	weekly-newsletter	115