

SQL Open Challenges

The following section includes open-ended challenges that do not have predefined answers. The goal is to encourage you to explore the DWH data independently, extract relevant information, and use Excel to create visualizations that support your conclusions.

After working through the data, we'll come together to discuss your insights, the queries you used, and the results you found.

These challenges are more advanced than the earlier exercises, so we recommend completing the previous ones first to build confidence before diving into these.

Challenges

- 1- This is based on a real Data Captain ticket we completed for Jamie D. around August 23, 2024. The objective is for you to replicate the resolution process yourself.

Description of the ticket was:

Export the distribution of night-length (days diff between check-in and checkout) of bookings.

An Excel delivery with different buckets (1 night, 2 nights, 3-5, etc) and the amounts would be perfect.

Goal: retrieve

- a) Calculate the **number of bookings per night-duration bucket** (e.g., 1 night, 2 nights, 3-5, 6-10, etc.).
- b) Provide insights of any clear outlier (ex: do we have checkouts happening before check-in? do we have extremely long bookings?)

Bonus:

- c) Do the results change if **duplicate bookings** are excluded?
- d) What if we only consider approved bookings?
- e) Do you notice a shift in the distribution depending on the booking creation year (e.g., 2022 vs 2023 vs 2024)?

Tip: Use Excel visualizations (e.g., histograms or bar charts) to analyse the distribution and support your conclusions.

- 2- This is a real Data Captain ticket that we did for Alex A. around the 25th of July 2024. The goal would be that you can replicate the resolution of the ticket by yourselves:

Description of the ticket was:

Discussing with Alex Anderson yesterday about the "Customer segmentation" on the KPIs, it seems they are using a combination of the business type and the listing volume to classify customers.

Business type is coming from Account type field in the dashboard (backend, Core), but for some live cases, data might not be up-to-date, especially those that have had been around for a very old time.

Goal: retrieve

- a) Identify **accounts with a missing Account Type** that show signs of **recent activity** (e.g., bookings or revenue).

Include useful host/deal attributes such as:

- a. Host name
- b. Go-live date
- c. Deal name

Bonus: also retrieve

- b) Those accounts that have this field missing that are not active anymore but had a certain huge amount of activity in the past (bookings, revenue, etc), in case it can be back-filled for history purposes

Notes: Terms like "activeness" and "huge activity" are vague. Use your common sense and domain knowledge to define relevant thresholds (e.g., last booking within 6 months, total revenue above £X).

Tip: the account type might not be directly appearing in any table, but rather the id of the account type. You might need to use this case-when query to retrieve it:

```
case
when id_account_type = 1 then 'PMC - Property Management Company'
when id_account_type = 2 then 'PMS - Property Management System'
when id_account_type = 3 then 'Guest Portal'
when id_account_type = 4 then 'OTA - Online Travel Agency'
when id_account_type = 5 then 'Test'
```

- 3- One of the main entities that we use for reporting and analytical purposes is the identifier of the Client, namely the Deal ID. In DWH, Deal ID is referred to id_deal and can appear (and not appear) in different places: Hubspot (CRM), Xero (Invoicing) and Core (the name used for the Backend).

Goal: provide an extract that gathers any deal appearing in any of the 3 different sources (Hubspot, Xero, Core). For each deal, identify

- a) whether this Deal appears in Hubspot or not
- b) whether this Deal appears in Xero or not
- c) whether this Deal appears in Core or not

Bonus: For each deal, provide:

- d) the time that went live according to Hubspot (if the Deal exists in Hubspot)
- e) the first time the deal had a booking created (if the Deal appears in Core)
- f) the last time the deal had a booking created (if the Deal appears in Core)
- g) the first time the deal had an invoice (if the Deal appears in Xero)
- h) the last time the deal had an invoice (if the Deal appears in Xero)
- i) the time the deal was created in Core (if the Deal appears in Core). Note that we do not have necessarily this information since the backend uses the user (platform user). A deal can have one or more users, so you might need to provide an approximated creation date.

Notes: there's different ways to retrieve the data. Using Staging tables will provide the most similar output as what we're retrieving from the abovementioned data sources. However, it might be easier to retrieve part of this information from the Intermediate KPIs tables and views - but we are aware that at the moment, we only consider Core Deals as the source of truth.