Beek Data Challenge - Data Engineer
User Status problem

Context

Beek is an audio content app that has a period of free trials and then the monthly or annual subscription is paid, our onboarding team is working so that more and more users start their free trials and then pay their subscription (CR FT to PS) and to measure its success we need to have a table with the necessary and updated data every day.

EDA Challenge

Database exploration and analysis

We have the current dummy database (not actual values) that has a very similar schema to our production database.

Database credentials:

- host: 35.227.110.100

- database_name: postgres

- user: postgres

- password: n$EYUJRrmZ9jz2>7o

- port: 5432

Connect to the database using your favorite SQL IDE and answer the following questions (please add each code snippet for the questions answered):
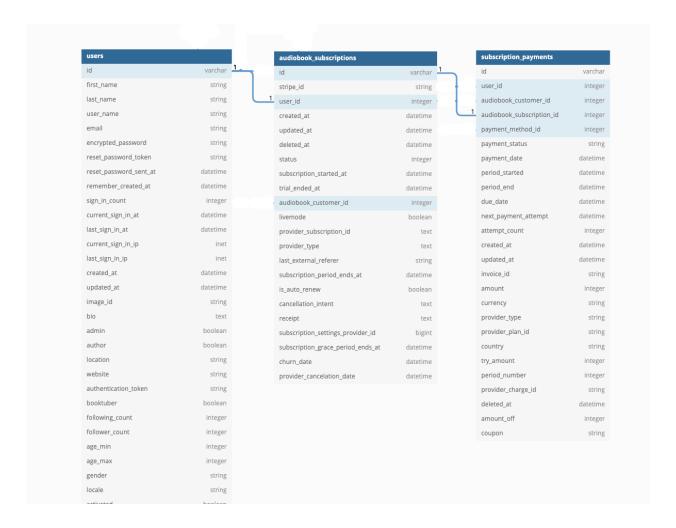
1. Calculate the WAU - weekly unique active users (where an active = a user that plays an audiobook) for the dates provided in the data.

2. What is the average time listened (in hours) per user for the last 30 days for the data provided?

3. What is the book category that has been showing more popularity? Why and how did you picked it?

4. Find two quality issues that affect the integrity of the database

ETL Challenge

Tables information

There are 3 main tables to obtain the user's information from when they sign up until they convert to pay and if not, in which phase of the free trial they were withdrawn.

- users

- audiobook_subscriptions

- subscriptions_payments

users

id	varchar
first_name	string
last_name	string
user_name	string
email	string
encrypted_password	string
reset_password_token	string
reset_password_sent_at	datetime
remember_created_at	datetime
sign_in_count	integer
current_sign_in_at	datetime
last_sign_in_at	datetime
current_sign_in_ip	inet
last_sign_in_ip	inet
created_at	datetime
updated_at	datetime
image_id	string
bio	text
admin	boolean
author	boolean
location	string
website	string
authentication_token	string
booktuber	boolean
following_count	integer
follower_count	integer
age_min	integer
age_max	integer
gender	string
locale	string
activated	boolean

audiobook_subscriptions

id	varchar
stripe_id	string
user_id	integer
created_at	datetime
updated_at	datetime
deleted_at	datetime
status	integer
subscription_started_at	datetime
trial_ended_at	datetime
audiobook_customer_id	integer
livemode	boolean
provider_subscription_id	text
provider_type	text
last_external_referer	string
subscription_period_ends_at	datetime
is_auto_renew	boolean
cancellation_intent	text
receipt	text
subscription_settings_provider_id	bigint
subscription_grace_period_ends_at	datetime
churn_date	datetime
provider_cancelation_date	datetime

subscription_payments

id	varchar
user_id	integer
audiobook_customer_id	integer
audiobook_subscription_id	integer
payment_method_id	integer
payment_status	string
payment_date	datetime
period_started	datetime
period_end	datetime
due_date	datetime
next_payment_attempt	datetime
attempt_count	integer
created_at	datetime
updated_at	datetime
invoice_id	string
amount	integer
currency	string
provider_type	string
provider_plan_id	string
country	string
try_amount	integer
period_number	integer
provider_charge_id	string
deleted_at	datetime
amount_off	integer
coupon	string

Requirements

1. Model a table that shows us the data from when a user signs up to when a user pays for a subscription. Think about the business cases and model the data to be consistent with the business.

   You can ask me all about it.

2. Build an ETL that takes the information from those tables and makes a CSV file with the resulting table defined above.

   - The ELT must run at 06:00 GMT -3

   - In case of failure of the ETL, have to send an email

   - Propose tests that validate the consistency of the data and how they will be implemented

3. The ETL code will be reviewed together in an onsite interview scheduled with our data team.

Data Warehousing Challenge

Design an MVP of a scalable Data Warehouse that responds to the questions ¿What content are users listening to?, ¿How long does a user listen to content? and explain:

- How do Data Analysts consume this data?

- How can the decision-makers use this data?

- What are the next steps to implement this DW?

- How do you monitor this DW?

Next steps

1. Send the challenge by email to andres@beek.io

2. After sending the email, a meeting will be coordinated to present the solution.

Any questions or suggestions do not hesitate to write to *andres@beek.io*
All the hits. Have a Can-Do Attitude.