Kahoot!

Data K!rew

Data Engineer Challenge

Instructions

- Use whatever programming language/tool you are most comfortable with
- The assessment is designed to get progressively harder. Do not feel you have to answer all the questions if you get stuck
- Make it clear in your answers if you have made any assumptions about the data/problems

What to return

- All source code, relevant files or written answers to questions
- Instructions for setting up and running your solution

Part 1

The below database table schema captures the historical state for every change a user makes to their profile. The current user profile can be found in the record with the latest created timestamp for any given user_id.

Table: user_changes	
Column	**Description**
uuid	String: The unique record id
user_id	String: The unique id of the user
user_created	Timestamp: When the user was created
created	Timestamp: When this record was created
name	String: The full name of the user
email	String: The email address of the user

Q1. Write an SQL query using vendor neutral ANSI SQL to find the user_id, current name and current email address for all users. Do not worry too much about the performance of the query, favour readability.

Q2. Write an SQL query using vendor neutral ANSI SQL to find the median time between the second and third profile edit. Do not worry too much about the performance of the query, favour readability.

Part 2

The purpose of this part is to create a proof of concept application that can ingest tweets from the public Twitter stream and output some insights.

The Kahoot! marketing team has identified some key questions that they believe if they could answer, would enable them to adjust their social media strategy and make it more effective.

Q2. Create a proof-of-concept for a tool that allows the user to specify a search term and receive every five seconds an updated output of some metrics about tweets that contain the search term.

The specific insights the tool should provide in its output are:

- ➔ **What is the total count of tweets matching the search term seen so far?**
- ➔ **How many tweets containing the search term were there in the last 1, 5 and 15 minutes?**
- ➔ **What are the ten most frequent terms (excluding the search term) that appear in tweets containing the search term over the last 1, 5 and 15 minutes?**
- ➔ **Within tweets matching the search term, who were the top ten tweeps (Twitter users) who tweeted the most in the last 1, 5 and 15 minutes?**
- ➔ **What is the sentiment of tweets matching the search term over the last 1, 5 and 15 minutes?**

This is not intended to be a fully production ready application. We are not expecting any fancy gui - command line is fine, or the ability to change the search term after the application is started. Also we are not expecting you to create your own sentiment algorithm, there are plenty of open source libraries that are good enough for this proof of concept. You will need to set up a Twitter developer account if you don't have one already.