Entry level data engineer – coding assignment

Author: Vasilis Vasileiou

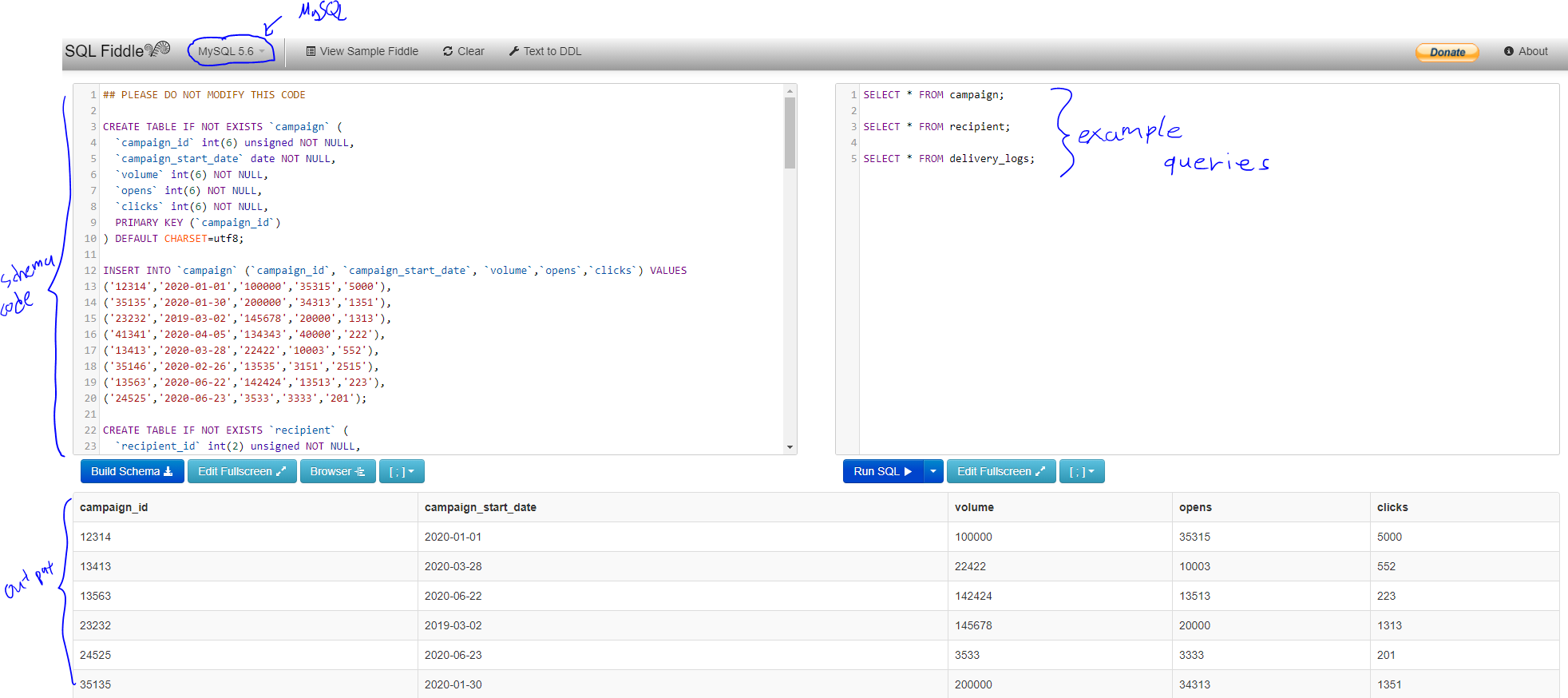
# Introduction

The day-to-day tasks of the role involve a fair amount of coding in Python and SQL. Therefore, we created this assignment, in order to better understand the level of your skills in these two programming languages. You will need to send us back two files: one .txt/.sql and one .py. If you can’t solve some of the questions, please add your incomplete code or thought process as comments in your code as we would still like to see the way you think. If you are unsure about what the question asks you to do then you can state your assumptions as comments and proceed with what you believe is needed. You are allowed to use Google but please work on your own and make sure you 100% understand your code. You could be asked questions about anything in the code you deliver.

# SQL

For the SQL part of the assignment we will make use of the website <http://sqlfiddle.com/>

The window of the left is the Schema builder panel and all the code that goes there will build the tables that we’ll be querying. The window on the right is the playground to test your queries and the window in the bottom are the query results. Please make sure that MySQL 5.6 is selected and then copy and paste the code called inside the file “schema\_code.sql” in the schema window (you can open the .sql file using any notepad). Click the button “Build Schema”. Now you are good to go! You can write queries in the window next to it and see the results. Make sure you see something like this:



The campaign table contains top level information for each campaign such as the volume, start date etc. The recipient table contains info for recipients (this is a sample of all recipients). The delivery\_logs table contains sample info for each time a recipient received an e-mail (did they open? Click? etc). An open is when someone opens an e-mail. A click is when someone clicks on a link inside an e-mail. A click always requires an open first. The open\_flag and click\_flag columns are 0 if the recipient didn’t open/click and 1 otherwise. Open rate is the number of opens per e-mails sent and Click rate is the number of clicks per e-mails sent. Write one SQL query for each of the below questions and put them all in the .txt/.sql file.

Questions:

1. Return the open rate for each campaign id.
2. Return the average open rate and average click rate per month (regardless of the year).
3. Return a count of how many recipients are called Dave.
4. For each recipient return the date that they ever received an e-mail.
5. Return the three weekdays (Tuesday, Thursday, etc) with the highest total amount of clicks.
6. Return the full names of all recipients that received at least one e-mail on the last delivery date.
7. Return the average days of difference between the date that the campaign started and the date that an e-mail of that campaign reached a recipient.

# Python

For the Python part of the assignment, we’ll be looking at NBA data! Please return a .py file answering the questions below. The data is on the URL:

<https://raw.githubusercontent.com/fivethirtyeight/data/master/nba-elo/nbaallelo.csv>

Please spend some time exploring the structure of the data. You will notice that in the questions below we recommend using the Pandas package but if you feel more confident using something else please feel free to do so.

Questions:

1. Read the data into a pandas dataframe.
2. Create a new column called ‘date\_game\_ts’ that has the timestamp of the date of the game.
3. Create a new column called ‘great\_game\_flag’ that is 1 if at least one of the teams scored more than 100 points and 0 otherwise (pts and opp\_pts are the points for each team).
4. Split the data intro two different pandas dataframes: aba, nba based on the column ‘lg\_id’.
5. Create a pandas dataframe that is named “stats\_aba” and contains the number of great games (the definition is above) per year (year\_id) in the aba dataframe.
6. Create a pandas dataframe that is called nba\_stats where every row is a basketball team (fran\_id) and every column is a year from 1947 to 2015 and inside the table is the number of wins (game\_result) that this team has had in that respective year, according to the nba dataframe.