Data Engineering Test Assignment

#0 Introduction

Welcome to the technical assignment for the data engineer position at Scalable Capital!

- Within the following tasks, you can skip a step if it's too hard (and use mock data for the next step).
- If something is not clear or can be done in multiple ways, describe why you chose your approach.

Scope

For this assignment, assume you are working as a consultant for the CEO of Spotify, and you are tasked with using your data skills to help them improve the business.

Dataset

You will be working with a real-world dataset provided freely by the open source project ListenBrainz:

> The ListenBrainz project serves as an archive where users can store their music listening history. This dataset can be used to create new music recommendation engines. The provided data dumps contains over a 100 million listens in the ListenBrainz database.

As the original dataset is quite large, we will provide you with a subset of this data.

Task #1 Data Ingestion

You are provided with an export of all listens that happened on the Spotify platform. Each line of the file contains a json document with data about one listen (the song that was listened to, the user who listened to the song, the time of the listen, etc). Please download the dataset in the following Google Drive folder: Test Assignment dataset

Your job is to load this file into a database for easier analysis:

- Set up the database. Create one or more tables, think about how you structure and optimize the database to simplify later analysis.
- Write a function that reads the export-file and writes the data into your database. The function is intended to run continuously, write it in a way that it can deal with duplicate or corrupted data.

For simplicity, use a sqlite database for this assignment. Sqlite comes with python, there is no need to install additional database drivers. The following code demonstrates how to work with sqlite in python, but feel free to use other database drivers if you are more familiar with them.

In [2]:
```
import sqlite3
import pandas as pd
# Connects to an in-file database in the current working directory, or creates one, if it doesn't exist:
conn = sqlite3.connect('spotify.db')

with conn:
    # Set up your database here
    conn.execute('CREATE TABLE IF NOT EXISTS greeting(greeting TEXT);')
    conn.execute('INSERT INTO greeting VALUES ("Hello World!");')

pd.read_sql_query('SELECT * from greeting', conn)
```

Out[2]:

	greeting
0	Hello World!

Task #2 Data Analysis

In the following, we ask you to run some python code to get more information out of the provided data.

a) First, familiarize yourself with the dataset. What can you tell us about the data and the included Spotify users? You can use visualizations to show your ideas as you think it makes sense.

b) Assume that the company decides to run more personalized marketing campaigns. For this reason, the data science team starts a project on user profiling. Within this project, you as a data engineer are asked to deliver a table with features for every user. Please prepare such a table with users and their features as a pandas dataframe. Choose the features that you think make sense as input for a data science model. The model that the data scientists are going to implement can be, e.g., a clustering model. The columns of the result table should be "user_id" and the respective feature names.

c) Assume, the CEO wants you to answer the following questions:

- Who are the 10 most active users?
- How many users were active on the 1st of March 2019?
- For every user, what was the first song they listened to?

Answer the above questions by running python code against the database that you setup in task #1. You can, e.g., work with pandas dataframes.

Task #3 Management Report

The CEO asks you to provide them with a management report containing the most important metrics that you can distill from the dataset. The desired output is a table showing how these metrics develop over time.

- Consider what the most important metrics are and generate the report using sql queries or a python script.
- Visualize the chosen metrics as graphs that you would put in a management report. Please use a python plotting library for visualization.
- What other metrics (that aren't available in the given dataset) would you like to add to the report? Prioritize your top 3.
- Shortly describe your reasoning for what metrics you picked and why you decided to include them in the report.

How much time did you approximately spend on the whole assignment?