Data Engineering Challenge

1 Introduction

This is a challenge assignment using an open source project of real world data. It showcases a complete end to end data engineering project.

1.1 Dataset

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

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.

1.2 Setup

Please make sure that your code can easily be executed on any operating system. Include instructions on how to execute your code on MacOS.

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).

We suggest to use duckdb for this assignment, but feel free to use any tooling you deem fit.

2 Data Analysis

In the following, we ask you to run some analysis on the dataset. The goal is to get more information out of the provided data.

a) To get started, answer the following questions:

- Who are the top 10 users with respect to number of songs listened to?
- How many users did listen to some song on the 1st of March 2019?
- For every user, what was the first song the user listened to?

b) Next, let's do a deep dive into user behaviour next. For each user, we want to know the top 3 days on which they had the most listens, and how many listens they had on each of these days. The result should include the following:

- 3 rows per user

- 3 columns: (user, number_of_listens, date)
- Please sort the result by the user and the number_of_listens column

c) [optional] Finally, we want to understand the development of active users within our userbase. For this, please write a query that calculates, on a daily basis, the absolute number of active users, and the percentage of active users among all users. We define a user to be active one some day X, if the user listened to at least one song in the time interval [X-6 days, X]. The result should adhere to the following schema:

- 1 row per day
- 3 columns: (date, number_active_users, percentage_active_users)
- Please sort the result by date.

Answer the above questions by running SQL queries against the data.

Please also outline the production setup you would choose for regularly updating the queries based on new data.

3 System Design

Suppose the data is produced by different upstream services, which publish messages to an event broker.

Design a system to enable all employees within the company to use the data for insights and dashboards (Business Intelligence). Please make reasonable assumptions if necessary to answer the following questions:

1. a) How would you approach this challenge with your team?
2. b) Which data stack architecture would you choose? Please design an architecture diagram that you present us in detail in the technical interview. (10 min)
3. c) What technical and organizational challenges might you need to overcome during implemen- tation of the project?