Data Engineer Take-Home Assignment

# Background

AllTrails is a freemium consumer subscription app and wants to better understand what factors contribute to revenue growth. One assumption of the company is that encouraging users to record activities (hikes, bike-rides, etc.) soon after their first signup could positively impact Pro subscription rates.Therefore, the company would like you to help the analysts find insights around the relationship between when users sign-up and when they complete their first GPS recording. To aid the data analysts in this task, your job is to write a batch-processing job to create a single clean, structured dataset that can be more-easily manipulated to help them find the data they need.

# Guidelines

* We expect this project to take around 3 hours, but feel free to let us know if you happen to spend more time on any bonus work
* Include and share a github project repo with code and documentation on how your code works and how to run it
* Include any next steps you would take if you had more time or access to additional data/tools
* List any key assumptions you made about the data

# Tasks

To help simplify the directionality for this assignment, we’ve come up with a list of tasks that we hope will help direct you:

1. Write a batch-processing job in the scripting language of your choice (Python is preferred)
2. The job should ingest the two TSVs provided into either dataframes or equivalent (ETL) or into storage in a data platform (ELT)
3. The job should then join and transform the two datasets into a single dataset with the following characteristics:
   1. Should include all columns in the users\_2017 table (feel free to rename them to “cleaner” values)
   2. Should include the following columns of the first recording created by a user:
      1. Recording\_ID
      2. Date\_Time
      3. Pseudo\_User\_ID
      4. Activity\_Type
   3. Should also include and expand the contents of the Recording\_Summary column into multiple columns that represent each inner field of the summary JSON data
   4. Should include a column for the time difference between the user’s signup date and their first recordings in hours
   5. Should properly handle any outliers (such as unusual data) while maximizing analytical capabilities and provide the documentation necessary to describe how these are handled.
4. The final dataset should be outputted to a single table or file
5. Documentation around the materialized table/datasets’s columns should be included either in code or as a document in your Github repository

Bonus Opportunities:

* Explain or create a table representing the user’s second recording and their properties / time-differences between second recording and signup date.
* Add tests/assertions to ensure the tables were created properly and outliers were handled properly
* Demonstrating proficiency around pipeline orchestration and infrastructure

# Dataset

Can be found at this link:

<https://drive.google.com/drive/folders/1LOv38Rjsn77eqRQFHW8GrSowfj0whkhq?usp=sharing>

For this project, you are given two data sets: 1) Recordings Data in 2017 and 2) Users Data for those registered in 2017 (including their pro subscription start date). Users can record their activities as either a free user or paid subscriber. There are many other ways that users can engage with the product, but for this exercise we will only look at recordings data.

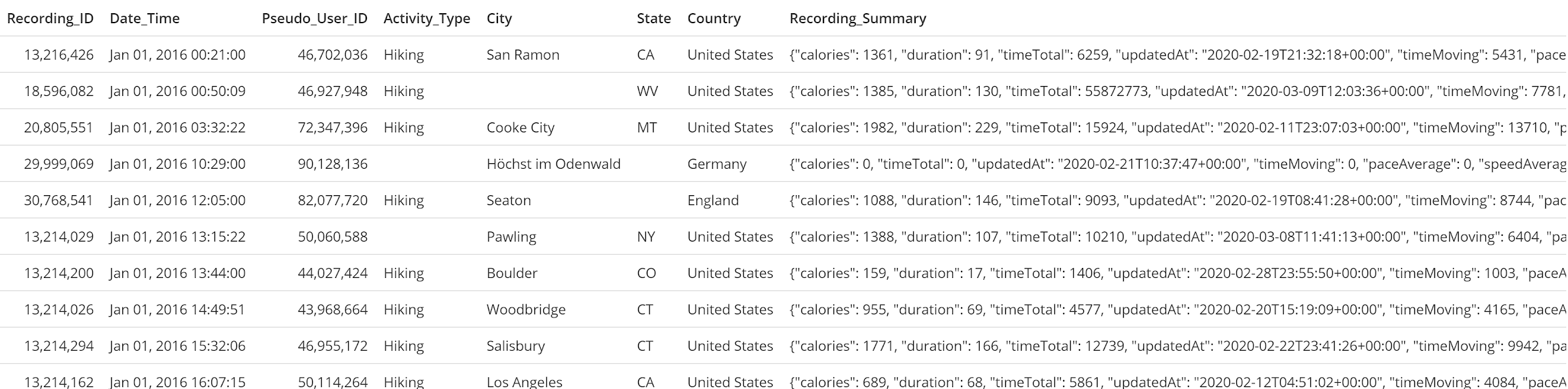
## Recordings Data in 2017

Filename: recordings\_2017.tsv

Schema:

| Column Header | Column Description |
| --- | --- |
| Recording\_ID | Unique ID given to recording on creation |
| Date\_Time | Date and time of recording |
| Pseudo\_User\_ID | ID of user who performed the activity |
| Activity\_Type | Recording activity type |
| Recording\_Summary | JSON field with recording stats |
| City | City where recording took place |
| State | State where recording took place |
| Country | Country where recording took place |

Example:



## 

## Users Registered with Recordings in 2017

Filename: users\_2017.tsv

Schema:

| Column Header | Column Description |
| --- | --- |
| Pseudo\_User\_ID | Unique User ID |
| signup\_date | The date that the user signed-up for an account with AllTrails |
| start\_date | The date that the user started their initial Pro subscription (date of conversion) |

Example:

