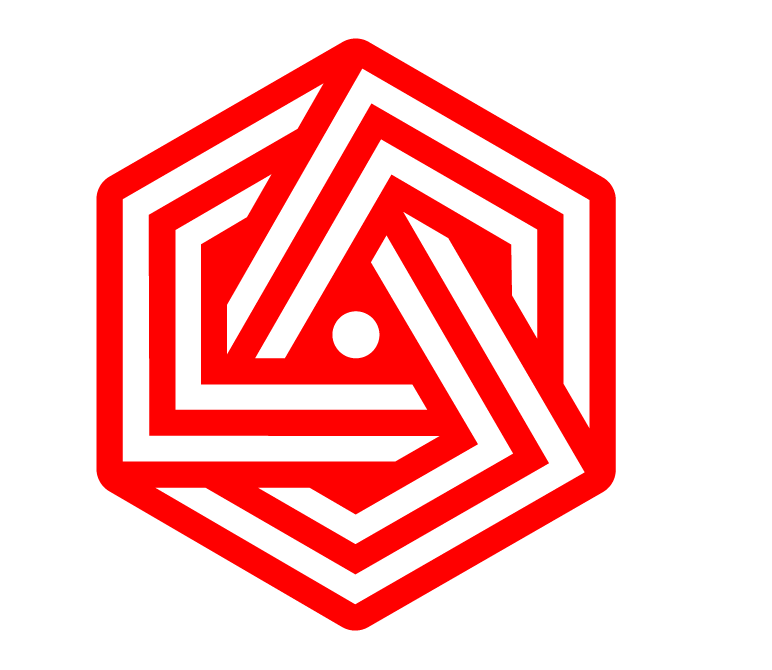
Data Engineer Technical Test

## Overview

This test is meant to help you demonstrate familiarity with common data processing workflows, especially leveraging Google Cloud Platform. It will also test your ability to recognize access patterns, design data to support those patterns, and communicate those design decisions to a diverse audience in the form of a written and an oral presentation of your results.

## Structure

1. You will be provided with disparate datasets and a number of common questions. Your first task is to **curate the provided data** for analysts to easily answer these questions.
2. You will be asked to **present your solution** during our on-site interview. Your audience will include analysts, engineers, and non-technical business users.

## Background & Dataset

Red Digital leverages data from multiple vendors to build strategies that benefit our client's marketing and advertising goals. Our Marketing team wants to develop an opinion on whether or not this will be beneficial for customers, and our analyst team has suggested that we utilize a [public dataset](https://www.kaggle.com/datasets/shivkumarganesh/tiktok-google-play-store-review) from 2022 user reviews for TikTok app available at Google Play Store, to come up with some compelling evidence either way.

TikTok, known in China as Douyin, is a video-focused social networking service owned by Chinese company ByteDance Ltd. It hosts a variety of short-form user videos, from genres like pranks, stunts, tricks, jokes, dance, and entertainment with durations from 15 seconds to ten minutes.TikTok is an international version of Douyin, which was originally released in the Chinese market in September 2016. TikTok was launched in 2017 for iOS and Android in most markets outside of mainland China; however, it became available worldwide only after merging with another Chinese social media service, Musical.ly, on 2 August 2018.

A Data Dictionary is provided for your understanding:

| **Column Name** | **Column Description** |
| --- | --- |
| userName | Name of a User |
| userImage | Profile Image that a user has |
| content | This represents the comments made by a user |
| score | Scores/Rating between 1 to 5 |
| thumbsUpCount | Number of Thumbs up received by a person |
| reviewCreatedVersion | Version number on which the review is created |
| at | Created At |
| replyContent | Reply to the comment by the Company |
| repliedAt | Date and time of the above reply |
| reviewId | unique identifier |

The analyst working on this has **already pulled the TikTok data into a Google Sheets and uploaded a number of usage CSV files into Google Cloud Storage** but is having trouble wrangling the data into a usable format. Ultimately, they would like to be able to pull this data from BigQuery into a visualization tool such as SuperSet, Tableau, or Google Data Studio with minimal SQL manipulation, intuitive column names, and appropriate data types.

## Questions for the Data

The analyst wants to use your solution to answer the following questions:

* How many unique users were studied? What was the average score for all reviews over the period represented in the dataset?
* What is the date range of the dataset?
* Do you have any recommendations for column names? Are there any patterns you were able to glean from the data?
* Does the review score change depending on the day, date, or time?
* Does a scenario exist that a user has more than one review, and if so, are the reviews approximately the same?
* What are the distinct review versions?
* How many users responded to a review? If any, which users responded?
* What is the tone of each response (angry, concerned, confused, etc.)?

## Resources

To assist the team in answering these questions, you will be provided with the following:

* Links to the **resources** mentioned above:
  + The [red-digital-interview-bucket](https://console.cloud.google.com/storage/browser/red-digital-interview-pranshu-bucket;tab=objects?project=red-digital-interview-sandbox&prefix=&forceOnObjectsSortingFiltering=false) containing usage data
  + The [Red Digital Data Engineering Interview TikTok- interview” workbook](https://docs.google.com/spreadsheets/d/1gu9NkDHezUmrXDhyHCZ-INjuGNm6qj7lSdypRFN8fpA/edit#gid=853255471), which contains a sample of the source data
  + [Red-Digital-interview-sandbox.bigquery](https://console.cloud.google.com/bigquery?referrer=search&project=red-digital-interview-sandbox) to create your tables and where the final dataset should be curated
* **Access** to the Red-Digital-interview-sandbox Google Cloud Platform project
  + Google Cloud Storage Admin privileges to the red-digital-bucket
  + Google Sheets access to the Red Digital Data Engineer Interview TikTok-interview-workbook
  + Google BigQuery User and Data Owner access to the Red-Digital-interview-sandbox dataset
  + Google BigQuery Job User access
  + Google Data Flow Developer access
  + Google Compute Viewer access
  + Additional permissions upon request
* A **service account JSON** with the same permissions, stored in the GCS bucket

## Deliverables

1. Utilize the provided data to create one or more BigQuery tables that would allow an analyst to answer the questions above.
2. Create any documentation you would want to deliver to the analyst, either as part of the technical solution or as stand-alone documents. These might include an ERD, a data dictionary, or example queries for dashboarding.
3. Create a GitHub repo, adding @jrich8573, @rv-rstack, @rv-dmalaby as collaborators to the repo.
4. Please remember any work (code, queries, diagrams) you share with us will help in evaluating the solution.

Email [rstack@redventures.com](mailto:rstack@redventures.com) and [jrich@redventures.com](mailto:jrich@redventures.com) with the subject “Red Digital Data Engineering Test” for support.

## Expectations

We expect you to spend no more than 4 to 8 hours on the test altogether. You are free to spend as much time as you would like, but the expectation is that anything exceeding that time is overkill.