Data Engineer - Technical Test

Welcome to the Data Engineer Technical Test. Being hands on is one of the key requirements for this role. We ask you to complete the following four tasks. To submit the results, or if you have any questions, you can send an email to the person you have been in contact during the selection process.

1) You are given the following SQL tables:

a) **streamers**: it contains time series data, at a 1-min granularity, of all the channels that broadcast on Twitch. The columns of the table are:
   - **username**: Channel username
   - **timestamp:** Epoch timestamp, in seconds, corresponding to the moment the data was captured
   - **game**: Name of the game that the user was playing at that time
   - **viewers**: Number of concurrent viewers that the user had at that time
   - **followers**: Number of total followers that the channel had at that time

b) **games_metadata**: it contains information of all the games that have ever been broadcasted on Twitch. The columns of the table are:
   - **game**: Name of the game
   - **release_date:** Timestamp, in seconds, corresponding to the date when the game was released
   - **publisher:** Publisher of the game
   - **genre:** Genre of the game

Write an SQL query to:

- Obtain, for each month of 2018, how many streamers broadcasted on Twitch and how many hours of content were broadcasted. The output should contain **month**, **unique_streamers** and **hours_broadcast**.
- Obtain the Top 10 streamers that have percentually gained more followers during January 2019, and that primarily stream FPS games. The output should contain the **username** and **follower_growth**.
- Obtain the Top 10 publishers that have been watched the most during the first quarter of 2019. The output should contain **publisher** and **hours_watched**.
  Note: Hours watched can be defined as the total amount of hours watched by all the viewers combined. Ie: 10 viewers watching for 2 hours will generate 20 Hours Watched.

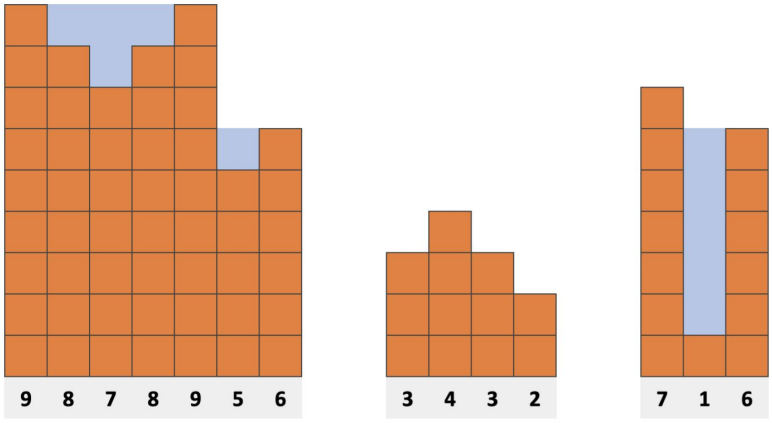
2) Imagine a new streaming platform has recently launched. They provide an API endpoint that allows third-parties to obtain, at any given time, the list of all the channels broadcasting in the platform, how many concurrent viewers each channel has, what game is each channel playing, etc.
   At Stream Hatchet we want to capture that information and offer it to our clients through our web app, providing rankings of top-performing streamers and games for each day, week, month, etc.
   Explain, in detail, how would you design and implement a system that is able to achieve that. From the data gathering to serving the information to the web app so that the end user can consume it, detail how you would implement each step, focusing on scalability and reliability.
   Describe what specific technologies, frameworks, and tools you would use and how you would deploy the system on a cloud-native infrastructure.

3) A 4-year-old is trying to build a tub for his goldfish out of Lego. Every Lego piece is stuck to the piece to its left and its right (except for the first and last one). All the pieces have a width of 1 unit.
   Write a program, using the programming language of your choice, that given the heights (in units) of the lego pieces from left to right, outputs the total amount of water held over the pieces that the kid built.

   Example input 1: [9 8 7 8 9 5 6]
   Example output 1: 5

   Example input 2: [3 4 3 2]
   Example output 2: 0

   Example input 3: [7 1 6]
   Example output 3: 5



4) Take a look at Stream Hatchet's BI
   a) Focusing on one or two sections of your choice, explain what insights you can extract from the data that is being represented.
   b) Propose a new section for the BI that offers a different perspective.
      Assume that all the metrics present in the BI (game genres, publishers, channels, tournaments, chat, etc.) are available for all the platforms and date ranges.
      What new insights could a business extract from this new section?
   c) Looking at the metrics that are available in the BI, think of a dataset(s) that you would use to apply Machine Learning to extract new information. Explain what techniques you would use and how the new information would be valuable.