We are excited to present you with a unique technical challenge that forms an essential part of our interview process. This exercise is designed to assess your proficiency in Python, your ability to extract meaningful insights from complex data, and your creativity in approaching open-ended questions.

Introduction:

The task at hand involves analyzing clickstream data that captures the various paths, or "journeys," of URL navigation that users take before arriving at **TripAdvisor**. These journeys are rich in information and can uncover insights into user behavior, preferences, engagement, and more. Your challenge is to delve into this data, explore the journeys, and answer specific questions we've laid out for you.

Directions:

You must complete task #1 and you can optionally do the other 2 tasks to increase your score. You have a week counting from today to do so.

The tasks outlined in this challenge don’t have a closed answer to them to allow for creativity and exploration. We encourage you to think outside the box and approach problems from unique angles. Well-documented code that explains your thought process and methodologies will be highly valued. Additionally, we expect the code to be of high quality and well-organized, adhering to the best practices in Python programming. We seek to understand your data science, interpretation and programming skills with this challenge. The submission format should align with our team's standards and goals, reflecting your ability to create accurate, intricate, and organized code. If you have any specific questions about these expectations, please don't hesitate to reach out.

You can find the data at the following s3 location and access it with these credentials

s3_uri=s3://ng-data-science-interviews/clickstream2

s3_region= us-east-2

s3_access_key_id=AKIASQRQO2RJDCVD7SAA

s3_secret_access_key=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

The dataset is divided into 47 Parquet files, each with an approximate size of 250MB. We recommend using a single Parquet file for prototyping purposes. However, if feasible, we encourage you to deliver your results using the entire dataset, leveraging the computing power available on your local machine or considering the use of cloud computing resources. Utilizing the full dataset will lead to a higher-quality evaluation of your results.

Data format

Sample files in Parquet format with the following structure:

userid	Hashed randomized user id
eventdate	Date when the event was saved on the server, UTC
eventtimestamp	Event timestamp collected on device, UTC
useragent	Browser user agent, if available
countrycode	ISO-3 country code
city	City name, if available
postalcode	Postal code, if available
platform	Desktop or Mobile
referrerurl	Referrer URL which sent traffic to targeturl
targeturl	Target URL which was clicked by the user
datasetcode	Supplier code
httpcode	HTTP Response Code, If available

Task 1: Understanding User Journeys (Mandatory)

Challenge: Analyze the clickstream data to identify the most common user journeys leading to TripAdvisor. What patterns or sequences of sites or pages do users typically navigate through before reaching TripAdvisor? Are there specific categories, themes, or domains that are common in these pre-TripAdvisor sessions? Interpret your findings (visualization and statistical analysis is optional, but it will increase your valuation as a candidate).

Tip: In this context, a "journey" or "session" refers to a series of clicks, page views, and user actions occurring within a specific time frame that ultimately leads to a predefined goal or destination, such as on TripAdvisor. These journeys offer insights into user preferences, behaviors, and decision-making processes.

The dataset provides information on how a user navigated from a "Referrer" URL to a "Target" URL. It's worth noting that the Referrer URL in one row could be the "target" URL in the previous row if the data is ordered by timestamp. To enhance data clarity and analysis, you may consider restructuring the dataset by consolidating URL information into a single column while introducing another column to indicate the sequence of user navigation. This would entail transforming the dataset format from its current structure:

Referrer URL	Target URL
a.com	b.com
b.com	c.com
c.com	d.com

To this structure:

URL	Sequence
a.com	1
b.com	2
c.com	3
d.com	4

This approach will enable you to sequentially analyze the user's navigation path leading up to their arrival at our reference URL, TripAdvisor. It's offered as a suggestion, but you're welcome to explore alternative techniques if you believe they would be more efficient for your analysis.

Task 2: Finding the Longest Way to TripAdvisor (Optional)

Challenge: Identify the longest journey (in terms of the number of clicks or unique pages visited) a user took before reaching TripAdvisor. Provide a comprehensive analysis of this journey, explaining what it may reveal about the user's behavior and intentions.

Task 3: User Engagement and Retention Analysis (Optional)

Challenge: Analyze user engagement and retention with respect to TripAdvisor links. Are there particular features or pages that lead to higher engagement? Identify and visualize drop-off points or areas where users might abandon the journey to TripAdvisor.