Specter

Welcome to the Specter Data Engineering Challenge! 👻

The aim of this take home assignment is to assess your ability to perform common data engineering tasks that we have here at Specter. Namely, parsing the output from a web scrape, picking out important details and inserting that data in a structured format into a database.

Please use Python for this challenge.

Please note, that in addition to the coding challenge there are a series of questions that we'd like you to answer. You may consult any online documentation that you like, but you will be asked to explain and justify your answers in your final interview. As this is a Senior level challenge we'll likely use these questions to go into more depth so you'll want to read around the subject.

Coding challenge

Parse the HTML scrapes

1. 5 scrapes of 5 websites have been made (LINK HERE). These correspond to the websites here:
    a. https://www.similarweb.com/website/tryspecter.com/#overview
    b. https://www.similarweb.com/website/byte-trading.com/#overview
    c. https://www.similarweb.com/website/crunchbase.com/#overview
    d. https://www.similarweb.com/website/pitchbook.com/#overview
    e. https://www.similarweb.com/website/stripe.com/#overview
    f. https://www.similarweb.com/website/google.com/#overview
2. Extract the following data points:
    a. Global Rank
    b. Total Visits
    c. Bounce Rate
    d. Pages per Visit
    e. Avg Visit Duration
    f. The change in rank over October, November and December
    g. The total number of visits in October, November and December
    h. Last Month Change in traffic
    i. Top Countries
    j. Age Distribution
3. Extract the data and store them in a single CSV file.

Transform the data you've extracted, and load it into a SQLite file

1. Clean and normalise the data for example
    a. the string "87.0B" is better represented as 87,000,000,000
    b. the duration "00:10:35" is better represented as seconds 635
2. Import the CSV into SQLite

3. Attempt to cast the data to data types, please don't cast all the data to Text

Analyse the data

1. Calculate statistics on the data
    a. Month-on-month growth on the web visits
    b. Month-on-month growth on the rank changes
    c. Rank the results based on growth in visits and rank using a relative scale
2. Plot the results on appropriate graphs and save the graphs as image files i.e. a JPEG or PNG

Please submit your code, your CSV files, your SQLite file and your plots either as a github repository or as a zipped file.

Questions

Please provide answers to the following questions, a couple paragraphs for each question will suffice.

1. We run scrapes continuously, both on the same websites as data changes over time and on new websites that we find interesting. How would you monitor the activity of the scrapers to make sure they were functioning and functioning correctly?
2. We join data from lots of sources and this can lead to sparsity in the data, often it's a case of identifying when we are missing data and differentiating that from when data simply isn't available. How could you determine missing data in a scalable way?
3. We release data on a weekly cadence, as time goes on we query more data and it can take longer to scrape and process the data we need. How would you scale the system to do more work within a shorter period of time?
4. A recent change to the codebase has caused a feature to begin failing, the failure has made it's way to production and needs to be resolved. What would you do to get the system back on track and reduce these sorts of incidents happening in future?

Good Luck! 🍀