Data Engineer Case Study

Format

You have a week to prepare and deliver this case study that will be reviewed and discussed during the Case Assessment interview with our engineers.

Tip: It might be useful to prepare 3-4 slides or similar to structure the conversation and send it prior to the discussion. Feel free to use any tool you prefer.

Exercises

1. Data Architecture

Superside delivers creative designs continuously to multiple customers impacting people over the world.

In order to better understand our customers' brand style and voice, we aim to extract insights from their corporate sites. The data gathered will enhance our understanding of our customer traits and identify what is resonating best with them, thereby aiding in enhancing our future service deliveries.

Your task is to design a proposal explaining how you would scrape, clean, preprocess and store this data from any corporate website.

Expected deliveries

- Explain how you would approach the challenge. Tip: You can use plain English supported with code snippets.

- Support it with an Architecture diagram.
- Which data would you store? Why and how would you store it?

📖 This exercise aims to know more about your data architecture mindset, how would you approach a distributed challenge and which things would you take into account when developing such ideas. Also, the data structure to support the system would be really important.

2. Data Code

Scraping

Based on the data architecture part, implement a python script that can scrape content from Superside landing page and another corporate website landing page of your choosing.

Expected Deliverables

- Develop a PoC in Python for this scraping module. You can use any library or framework of your choice.
- Explain your solution.
- Justify your choice of each library or framework used.

📖 This exercise aims to know more about your coding skills, how you structure the different methods, classes, and which patterns and best practices you put into practise (tip: add some relevant unit tests) .

Generative AI

For this part, we want you to leverage the power of OpenAI API. The goal is to create a module that integrates this API for generative purposes using the data scraped from the corporate websites.

Expected Deliverables

- Develop a PoC in Python or Scala for this Generative AI module. You can use any library or framework of your choice.
- Explain your solution.
- Justify your choice of each library or framework used.

📖 This exercise aims to understand your experience with integrating AI APIs, how you utilize generative AI in real-world applications, and your ability to innovate using AI.

Additional Instructions

Your solution should focus on generative AI and robust web scraping methods in line with the job description. Elaborate on how the data collected and processed can be used for our AI initiatives, how to ensure its overall quality and reliability, and how to build profile models based on this data.

When explaining your architecture, consider how to best connect our internal systems with external data sources and services via APIs, and how to build near real-time services for swift data processing and delivery.

The ability to visualize and effectively communicate complex insights is essential. Hence, your explanations and presentations should be understandable to both technical and non-technical stakeholders.

Finally, your PoCs should demonstrate your skills in Python. Tip: We love clean and tested code.

Whether you use any AI-powered tools to develop and/or implement your solution (e.g. chatGPT to generate a first proposal or GitHub Copilot to generate code), please also mention which ones you have used and explain your approach to interact with them (e.g. system and user prompts, chain of thought, etc).

Remember, this is an opportunity to showcase your data engineering skills, your understanding of generative AI, and your creativity in solving complex data problems

Good luck!