Take-home Assignment - Data Engineer (Crawling Focus) 🚀

As part of our hiring process, we would like you to complete a take-home assignment that will assess your skills and experience in line with the requirements of the role. Please carefully read the instructions and complete the tasks to the best of your ability.

Note:

- **Candidate Profile**: professional and personal information of a person. For example: name, experience, skills, company…
- **Deadline**: Three days after the assignment was sent.

Assignment Instructions:

Data Crawling and Transformation

You have been assigned the task of building a data pipeline to crawl and transform candidate data from multiple sources. The sources include social media platforms, job sites, and other web-based platforms. Perform the following steps:

- Please come up with a data schema with all possible data points to create a holistic view of the candidate profile.
- Identify at least three different data sources that you can use to create the candidate profile from step 1. . These can include social media platforms, job sites, or any other web-based platforms that provide public data.
- Design and implement a sample data extraction process to retrieve data from the identified sources. You can use any web scraping or data crawling techniques you are familiar with to accomplish this task. Please make sure to consider the security and anti-scraping of these sources, and how you can get through them.
- Transform the extracted data into a unified format suitable for further processing and analysis. Perform any necessary data cleansing, normalization, or enrichment operations as part of the transformation process.
- Store the transformed data in a data store or file format of your choice, considering factors such as scalability, ease of access, and data integrity. And mapping data between different sources.

Additional Guidelines:

- Use Python for scripting and any necessary libraries for data processing, web scraping, and data pipeline orchestration.

- Document your code and provide clear instructions on how to run and test the solution.
- Ensure that your solution is well-structured, modular, and follows best practices for maintainability and scalability.
- If you encounter any challenges or limitations during the assignment, document them along with your proposed solutions or workarounds.

Submission:

Please submit the following items:

- The code and scripts for data extraction, transformation, and data pipeline orchestration.
- The documentation document provides an overview of the solution, including diagrams and code snippets.

We value clean, well-documented, and efficient code, as well as thoughtful design considerations.