Test for Data Engineer

Instructions

1. Attempting all the questions is not mandatory.
2. **Partially complete solutions are welcome**. We want to see your approach and understanding.
3. Once the task is completed, please **share the git repository link.**
4. Please share your submission at
   For any questions, you can contact us at +91 9711889358

Test

Thank you for your interest in the Data Engineer Intern position at EZ. As part of our application process, we'd like to assess your skills with a takeaway task that mirrors the type of work you would encounter in this role. Please find below the task details:

Build a Data Pipeline for Translation Memory

*

You are tasked with designing and implementing a simple data pipeline to extract, transform, and load (ETL) data from a **Translation Memory eXchange (TMX)** which can be downloaded from **https://opus.nlpl.eu/download.php?f=UN/v20090831/tmx/ar-en.tmx.gz** into a database. The TMX file contains parallel translation between English and Arabic. Your goal is to clean and structure the data before loading it into the database.

Requirements:

1. **Data Source: https://opus.nlpl.eu/download.php?f=UN/v20090831/tmx/ar-en.tmx.gz**
2. **Data Transformation**: Perform the necessary transformations to ensure reading english and Arabic characters

3. **Database**: Use any relational database of your choice (e.g., PostgreSQL, MySQL) to store the transformed data.

4. **ETL Process**:
   - Create a reader module for reading of TMX files
   - Apply the necessary transformations.
   - Load the cleaned data into the chosen database.

5. **Code Quality:** Write clean, modular, and well-documented code. Include comments where necessary to explain your thought process and any assumptions you make.

6. **Repository Structure:**
   - Create a GitHub repository or create a Zip for this task.
   - Include all relevant code, scripts, and configuration files in the repository.

7. **Documentation:**
   - Provide a README.md file that explains how to set up and run your data pipeline.
   - Include details about any prerequisites, installation steps, and execution instructions.

Submission:

Please complete the task by 13th September 2023. Once you have finished, share the GitHub repository link with us or the Zip folder. Your solution will be evaluated based on code quality, correctness, adherence to requirements, and documentation.

Feel free to reach out if you have questions or need clarification on the task. We look forward to reviewing your solution and assessing your skills.