Data Engineer Assignment

Your primary task as a data engineer will be managing and refining our data flow, focusing specifically on counterparty information. This assignment has two main parts: Duplicate Removal and Entity Resolution.

Dataset

You will be provided with a CSV file containing counterparty data entries. Each entry has three fields: 'id', 'name', and 'iban'. The dataset contains duplicates and counterparties that are essentially the same but listed under slightly different names.

Tasks

Duplicate Removal

- Input: The aforementioned CSV file with 'id', 'name', and 'iban' fields. The file contains duplicate entries.
- Task: Implement a process that identifies and removes duplicate counterparty data entries. Consider how this process might scale with increasing data volume.
- Output: A cleaned CSV file that does not contain any duplicate entries. Each entry should have a unique 'id'.

Entity Resolution

- Input: The cleaned CSV file from the Duplicate Removal task. This file contains counterparties that are essentially the same but are listed under slightly different names or ibans.
- Task: Develop a method to identify and link similar counterparties. Given that counterparties could be listed under slightly different names or ibans, consider using a suitable method to link these entities, taking into account scalability and accuracy.
- Output: A final CSV file where each unique counterparty is represented only once and has a list of all names or ibans under which they were previously listed.

Your deliverable should be a script that performs these tasks, written in any programming language you're comfortable with. Please include an explanation of your approach, a brief overview of the code, and instructions on how to run the scripts on a new dataset.

Evaluation Criteria

We will evaluate your assignment based on the following criteria:

Quality of the code (readability, structure, performance)

- Scalability of the solution
- Creativity in dealing with the described problems
- Completeness and clarity of the documentation

We appreciate the time and effort you put into this assignment and look forward to reviewing your submission.