Problem Description: Extracting and Analysing Tax Invoice statement

Please refer to the attached PDF file (title *Test PDF*).

You are given a tax invoice statement in the form of a PDF file that contains transaction details. Your task is to design a system that can extract transactions from the PDF, eliminate any duplicate entries, and perform operations on the cleaned dataset.

Note: Candidates are encouraged to use appropriate programming languages and tools for PDF parsing, data manipulation, and SQL operations in their solutions.

Part 1: Data Extraction

Develop a script or program to take a pdf as input and extract transaction details from the provided tax invoice PDF file. The information includes App ID, Xref, Settlement Date, Broker, Sub Broker, Borrower Name, Description, Total Loan Amount, Commission Rate, Upfront, Upfront Incl GST.

Part 2: Data Storage

Store the extracted data in a database(file can also be used as datastore), so that the information is available when required.

Part 3: Deduplication

Implement a deduplication mechanism to identify and remove any duplicate transactions from the extracted dataset. Transactions should be considered duplicates if they have the same Xref and Total Loan Amount. If the same file is uploaded multiple times, the datastore should not store multiple instances of the same transaction.

Part 4: SQL Operations

Design a set of SQL operations to analyse the dataset. Perform the following tasks:

1. Calculate the total loan amount during a specific time period.
2. Calculate the highest loan amount given by a broker.

Part 5: Reporting

1. Generate a report for the broker, sorting loan amounts in descending order from maximum to minimum, covering daily, weekly, and monthly periods.
2. Generate a report of the total loan amount grouped by date.
3. Define tier level of each transaction, based on the following criteria
    a. Tier 1 : Total Loan Amount > 1,00,000
    b. Tier 2 : Total Loan Amount > 50,000
    c. Tier 3 : Total Loan Amount > 10,000
4. Generate a report of the number of loans under each tier group by date.

Part 6: Results

Kindly submit the code via GitHub.

Additional Considerations

- The solution should be scalable to handle large bank statement PDFs with a substantially higher number of transactions.
- Consider error handling mechanisms to manage variations in the PDF structure or unexpected data formats.
- Document the process thoroughly to ensure clarity and replicability.

Evaluation Criteria

- Accuracy of data extraction and deduplication.
- Correct implementation of SQL operations.
- Efficiency and performance of the solution, especially with large datasets.
- Clarity and organisation of the code.
- Documentation quality.