Data Migration Analyst Take Home Assignment

September 2022

Executive Summary

The final deliverable for this project will be a zipped project directory containing the following items:

1. Python Files:
    a. Python function that removes duplicate transactions (Step 1)
    b. Function for creating the TransactionByDay.csv report from (Step 2)
    c. Function for creating the CustomerTotalsByDay.csv report from (Step 2)
    d. Function for loading CSV data into a SQLite database from (Step 3)
2. CSV File Directory:
    a. Transactions.csv (Provided)
    b. Customers.csv (Provided)
    c. CleanedTransactions.csv (Step 1)
    d. TransactionsByDay.csv (Step 2)
    e. CustomerTotalsByDay.csv (Step 2)
3. SQLite .db file with the following tables:
    a. Transactions (Step 3)
    b. Customers (Step 3)
    c. CleanedTransactions (Step 3)
    d. TransactionByDay (Step 4)
    e. CustomerTotalsByDay (Step 4)
4. SQL File Directory:
    a. SQL file with the query necessary to create the TransactionByDay Table (Step 4)
    b. SQL file with the query necessary to create the CustomerTotalsByDay Table (Step 4)

Instructions

There are two csv files included with this assessment:

    a. Transactions.csv: Contains transaction data for a 3 day span.
    b. Customers.csv: Contains a lookup for CustomerID to CustomerName

Create a Python project to complete the following tasks using Transactions.csv and Customers.csv:

1. During the data extract for the Transactions.csv file, some duplicate rows were created. Please created a new csv file called CleanedTransactions.csv and store it within the project directory. (**A duplicate row is a row that contains a duplicated TransactionID in the transaction file.**) The TransactionID column should be unique in the CleanedTransactions.csv file.
2. Using Python, create 2 new csv reports and store them within the project directory.
    a. Report 1: TransactionsByDay.csv – A count of the total number of transactions that occurred each day. This report will have three rows in it, each one corresponding to a single day within CleanedTransactions.csv. Please have the following columns in the report (Date, CountOfTransactions).
    b. Report 2: CustomerTotalsByDay.csv– The sum of transactions for each customer by day. Please have the following columns in this report: (CustomerName, Date, Total).

3. We would like to have this data stored in a database for future use. Inside of the project folder create a SQLite database with the following tables: Transactions, CleanedTransactions, Customers.
4. Once the data is loaded into the SQLite database recreate the two csv reports created with Python by using SQL to create the reports. Use SQL queries to create the two reports as tables within the SQLite database. The two table names should correspond to the CSV report names from step 2. Save the queries used as .sql files within the project directory.
5. Zip the entire project directory and return to the interview team. Be prepared to step through the code you have written and explain any logic or design choices during the first part of the in-person digital interview.

Hints:

1. If you choose to view the .csv files during your work, please do not open and then save them with Microsoft Excel or similar software. There are certain string formatting properties that will be stripped away if you do so. If you need to look at the data, use a text editor such as NotePad or NotePad++.
2. SQLite was chosen due to its portability and light weight nature. However, it is missing a SQL idiom that you may find necessary for step 4. If you are needing to use SELECT INTO to create the report table, please use CREATE TABLE table_name as SELECT format instead.