**Expectations:**

Code needs to be uploaded in github and share the URL with proper comments and instructions to run and validate your code.

**Python Questions:**

**TASK 1**

One of our products is in charge of downloading and ingesting millions of records from our clients.

Recently during ingesting a large dataset we had our entire DB(postgres) go down and the entire ingestion process from a pandas dataframe to sql took around 2-3 hours because of the RAM unavailability. Now this has two simple fixes

- Increase ram/ scale the db on demand

- change our code to accommodate these restrictions and make the entire ingestion process much faster on the way.

How would you approach this? We are not looking for a full blown ingestion logic. Just a small script to take a given csv file and upload it to DB in an efficient manner.

Write code to take a large csv file( > 1GB ) and ingest it to table - public.test\_od

=======================================================

**TASK 2**

=======================================================

Our customer records are sometimes a little messy and contain duplicate entires.

For instance, we might have two records like "Pfizer" and "Pfizer Incorporated (old DO NOT USE)".

Please write a small basic function to merge these records together. We've provided a few example inputs below.

We are aware this is a very difficult problem: we're looking for what you can do quickly as a basic case, not an ideal solution.

Equipment ONLY - Saama Technologies

Saama Technologies

SaamaTech, Inc.

Takeda Pharmaceutical SA - Central Office

\*\*\* DO NOT USE \*\*\* Takeda Pharmaceutical

Takeda Pharmaceutical, SA

Ship to AstraZeneca

AstraZeneca, gmbh Munich

AstraZeneca (use AstraZeneca, gmbh Munich acct 84719482-A)

Use your own interpretation of the question and feel free to provide a written explanation for your choices as well.

=======================================================

**SQL Questions**

=======================================================

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Visits** |  |  |  | **Customer** |  |  |  |
| Customer\_id | City\_id\_visited | Date\_visited |  | Customer\_id | Customer\_name | Gender | Age |
| 1001 | 2003 | 1-Jan-03 |  | 1001 | John | M | 25 |
| 1001 | 2004 | 1-Jan-04 |  | 1002 | Mark | M | 40 |
| 1002 | 2001 | 1-Jan-01 |  | 1003 | Martha | F | 55 |
| 1004 | 2003 | 1-Jan-03 |  | 1004 | Selena | F | 34 |

|  |  |  |
| --- | --- | --- |
| **City** |  |  |
| City\_id | City\_name | Expense |
| 2001 | Chicago | 500 |
| 2002 | Newyork | 1000 |
| 2003 | SFO | 2000 |
| 2004 | Florida | 800 |

======================================================

**TASK 1**

=======================================================

1) Cities frequently visited?

2) Customers visited more than 1 city?

3) Cities visited breakdown by gender?

4) List the city names that are not visited by every customer and order them by the expense budget in ascending order?

5) Visit/travel Percentage for every customer?

6) Total expense incurred by customers on their visits?

7) list the Customer details along with the city they first visited and the date of visit?