Data Engineer Take Home Challenge

Here I would explain why I would design and write the code in the submitted fashion and clarify my logic behind the code.

Overall Logic

In this challenge, we have chosen to process the information using the Python library Pandas whereas it is powerful with many formats of data and illustrates data well using its dataframe. The reason we did not choose to write everything using Apache Hadoop framework or Apache Spark framework is due to the small size of the dataset. With only 1000 lines, pandas is more than enough for this task.

Widget List

After careful observation, the widget list contains information of a list of widgets for each user, however, some users may have more than one widgets and there also exist users who possess none. We have decomposed widgets into their own rows. It is worth to mention here that since some users may have multiple widgets, there exist lines where one user's information appear more than once.

We could also convert the dataframe in the format where all other user information appear only once, however, making them appear more than once make querying much easier. But alternatives can be done easily through other commands.

Anonymisation

For this specific task, we have opted for the base64 encryption and decryption. There exist even safer ways to encode them, such as SHA, however, SHA is irreversible, hence making it impossible for use to recover an email once an anonymised value is given.

The anonymised column returns a list of byte string, they could be converted to *utf-8* string, we have provided that code inside the file. We have also written the function where an anonymised value is given, we could recover the original email address. One should notice that once the recovered email address is found, we have chosen to delete the column of decoded emails since they shall remain anonymous.

The base64 encryption gives us a **unique** byte string for a unique email address, this is useful for machine learning training and inference later on, with the idea of considering the column of anonymised email address as a vector, we could convert them using one-hot encoding whereas the an embedding matrix could be constructed eventually.

Conclusion

As we all know, there exists many alternative solutions for various tasks involved in this challenge, they could be easily done. Here, we have only demonstrated one option to complete this task. Other option could be easily done per the request of the receiving teams.