castor

Overview

The assignment is expected to take your 4 hours maximum. Please let us know if you have any doubts related to finishing the test or if it takes longer than 4 hours for you to complete. We will be ready to help.

Good luck and we can not wait to see what you will bring to the table.

Test Instructions

- Use Python to complete the task
- All code should be authored by the applicant. You may use any libraries that you consider production quality.
- Your code should demonstrate a production ready standard and be:
    - Well structured
    - Extensible
    - Testable
    - Readable
    - Commented
    - Etc.
- You should include some tests for your code
- Any 3rd party code (at all) should be clearly attributed
- We prefer that you do not use AI coding tools to author code, but if you do feel that these tools are essential, please clearly describe which parts were AI generated.

castor

How to submit the solution

If you choose to make the repo private, please make sure that you invite the following github users (@castorian, @sandldan and @alexgr-castor) so that we can review your code.

If all else fails … submit it to us as a .zip!

Your Challenge

Create a small Python command line application that allows a user to transform an existing dataset into a dataset with a different format.

For this assignment:

- The input files arrive in CSV format, where the first row represents the column headers and the remaining rows are the data.
- The resulting output file should be a CSV
- The output CSV should be able to have different column order to the input.
- Allow a transformation to be applied to any existing column before it is output: e.g Format a date, change a timezone, recode enumeration values, strip sensitive information, perform a calculation.
- Your implementation should have the following 3 transformations available to configure on columns in the sample data set:
    - Convert UUIDs into a simple integer sequence, whilst maintaining their uniqueness
    - Redact fields to replace data that is sensitive with similar looking random data
    - Convert timestamps to the appropriate date in the year-month-day format.
- The `user_sample.csv` input data file is provided, this has 100 lines.

Please include instructions on how to run your tool to generate a transformed output and also include the output file generated by following the instructions in your submission.

In the README of your submission. Please also include a bullet level discussion of how your implementation would scale to 1,000,000 line input files and what changes would be required to minimise the processing time and to scale to effectively utilise the available resources of the machine being used.