By Henry Zhu, Special Thanks to my Mentor Heheng Li and my Manager David Chen

Masking Customer Data for Data Mining w/ GDPR Concern

Background Information + Constraints

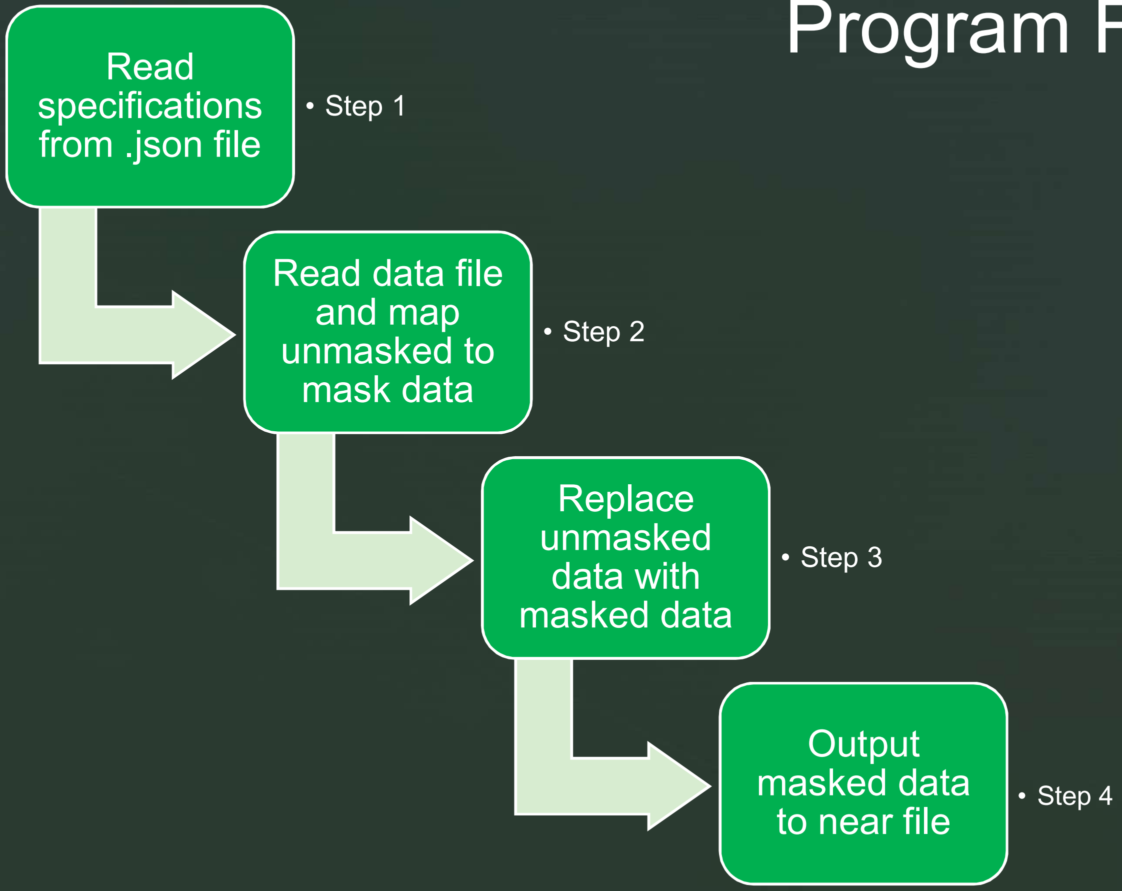
On May 25th, 2018, the European Union introduced GDPR, which protects a user’s PII (personally identifiable information from being accessed and used by companies.

My project reads in an unmasked data file containing information about Microsoft employees, and protects the employees’ PII (e.x. REDMOND\itatstsv > xostwvs\xavaeaez)

Project Constraints:

1. Keep non-PII part of the file unchanged
2. Ensure that obfuscated PII still looks similar to original
3. Ensure that PII cannot be predicted from masked data

Program Flow



Read specifications from .json file

- Step 1

Read data file and map unmasked to mask data

- Step 2

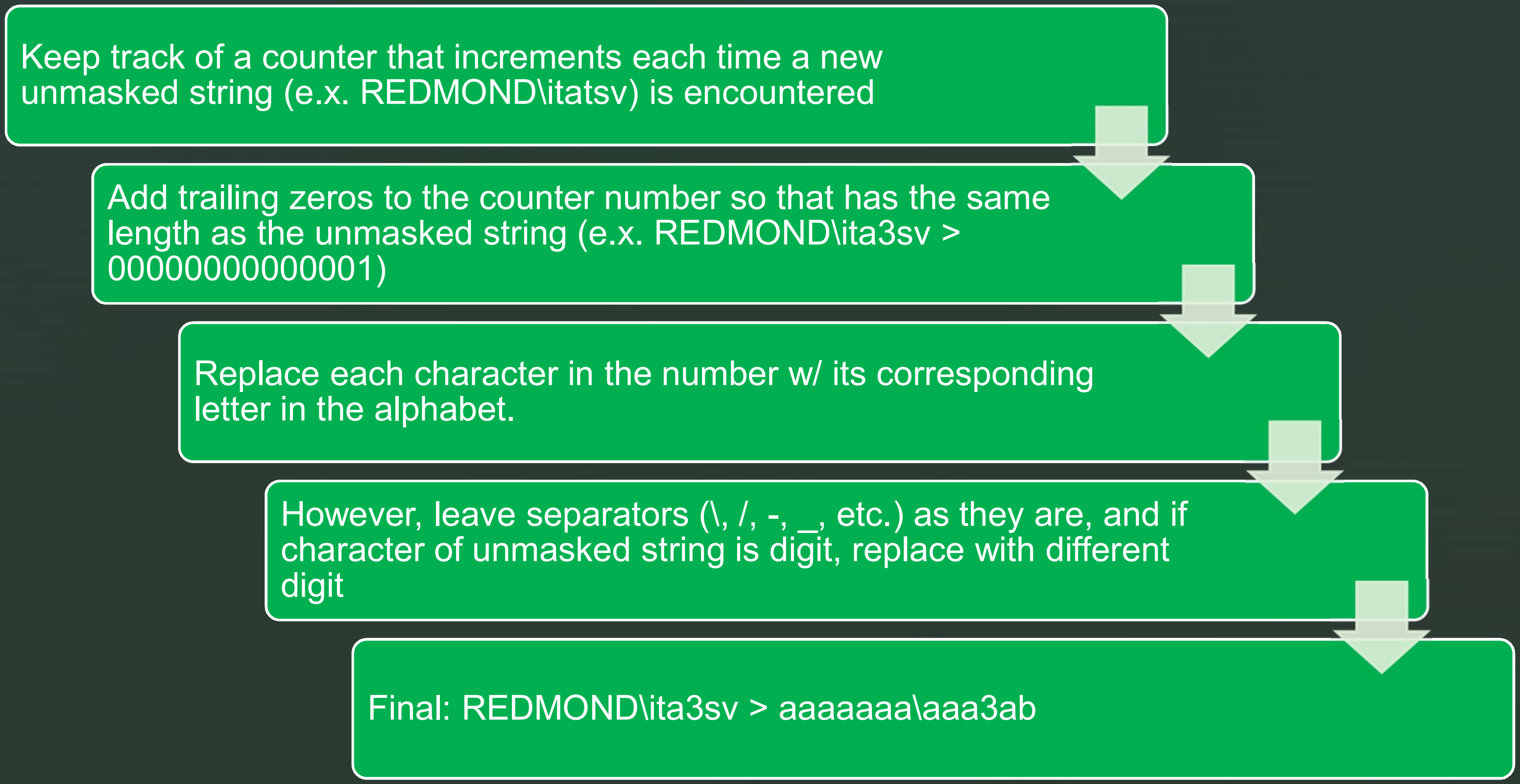
Replace unmasked data with masked data

- Step 3

Output masked data to near file

- Step 4

Masking Method: Radix



Keep track of a counter that increments each time a new unmasked string (e.x. REDMOND\itatsv) is encountered

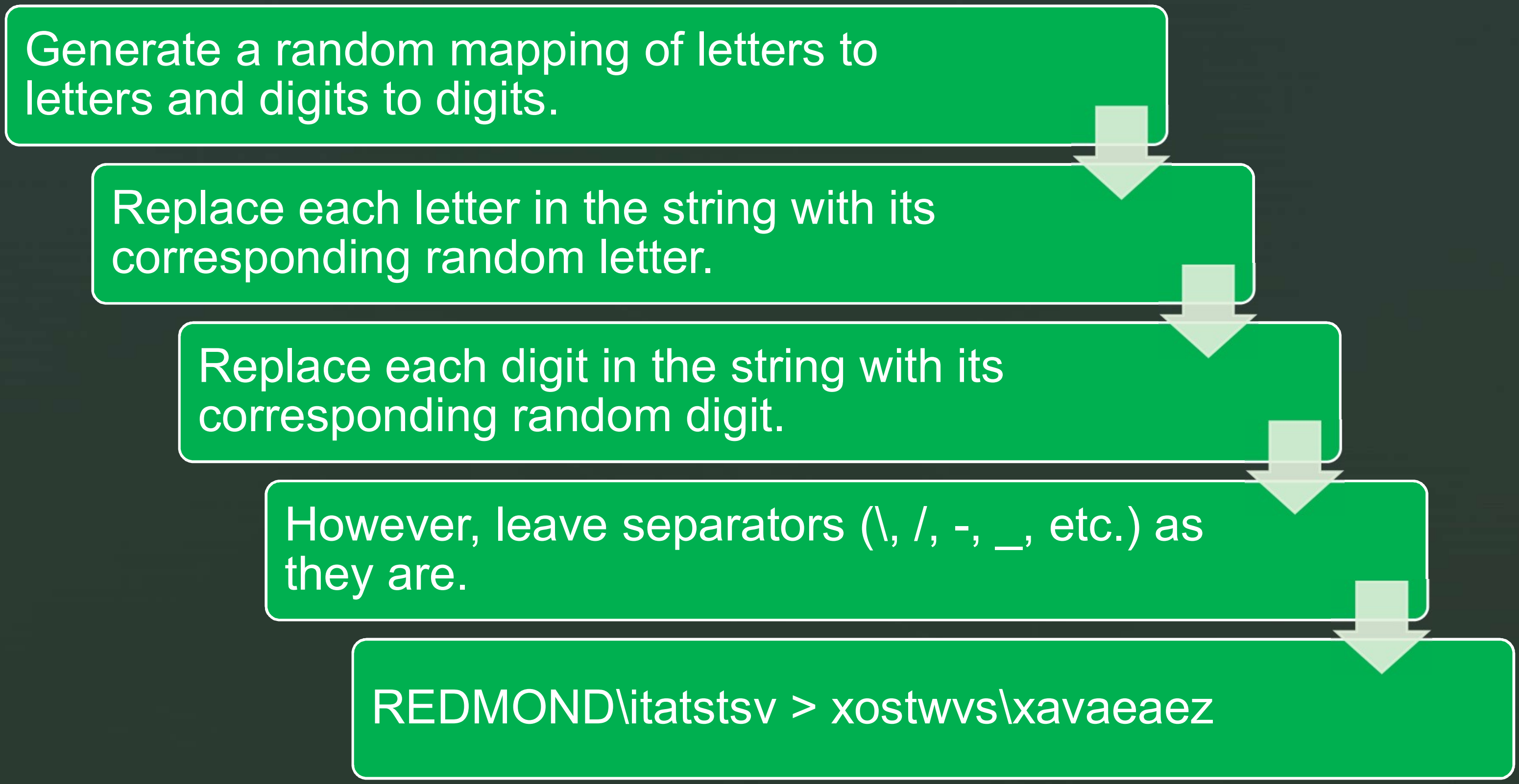
Add trailing zeros to the counter number so that has the same length as the unmasked string (e.x. REDMOND\ita3sv > 00000000000001)

Replace each character in the number w/ its corresponding letter in the alphabet.

However, leave separators (\, /, -, _, etc.) as they are, and if character of unmasked string is digit, replace with different digit

Final: REDMOND\ita3sv > aaaaaaa\aaa3ab

Masking Method: Random



Generate a random mapping of letters to letters and digits to digits.

Replace each letter in the string with its corresponding random letter.

Replace each digit in the string with its corresponding random digit.

However, leave separators (\, /, -, _, etc.) as they are.

REDMOND\itatstsv > xostwvs\xavaeaez

Improvements / The Future

1. Partial masking (REDMOND is not PII)

   Example: REDMOND\itasv3 > REDMOND\aaaaa5

2. Maintain capitalization

3. Foreign key, change all references to REDMOND\abc

4. Creating a GUI

Appendix

JSON Specification: categoryIndexes indicate which columns are to be masked, and method indicates what strategy of masking is to be used.

```
{
    "categoryIndexes": [ 2, 3, 4, 5, 6, 7, 8, 9, 11, 16 ],
    "method": "random"
}
```