Introdução

Lidar com grandes volumes de dados é um desafio comum em análise de dados e processamento de arquivos. Quando os dados não cabem na memória RAM, precisamos adotar estratégias eficientes para processá-los. Uma abordagem útil é usar o parâmetro *chunksize* ao ler arquivos, especialmente arquivos CSV, com a biblioteca Pandas.

O que é *chunksize*?

O parâmetro *chunksize* permite que leiamos e processemos um arquivo em partes menores (*chunks*) em vez de carregá-lo inteiramente na memória. Cada chunk é um DataFrame, e podemos aplicar operações a cada chunk individualmente.

Vantagens do uso de *chunksize*

Economia de Memória: Carregar grandes arquivos em blocos(*chunks)* economiza memória RAM, pois apenas uma parte dos dados é mantida em memória de cada vez.

Processamento Parcial: Podemos processar cada bloco(*chunk)* separadamente, aplicando operações específicas, como cálculos, filtragens ou agregações.

Iteração Eficiente: O uso de *chunksize* permite iterar sobre o arquivo **sem** sobrecarregar a memória.

Implementação

- Foi implementado a biblioteca panda para ler o arquivo em chunksize.

- Defini o tamanho de 10 mil registros por bloco(*chunk*)

- Para cada instrução foi definido um bloco(*chunk*), porém todos com o mesmo tamanho.

Conclusão

O uso de chunksize é uma estratégia eficaz para lidar com grandes volumes de dados, permitindo processamento eficiente e economia de memória. Ponto importante lembrar de ajustar o tamanho do bloco(*chunk*) conforme necessário para otimizar o desempenho.

Mirtes Miyuki Kagohara