Testes de engenheiro de dados (Jhonatan Aguiar)

Para abordar a solução de processamento de grandes volumes de dados, utilizei ferramentas especializadas em Python, como a IDE PyCharm. Iniciei os testes no Jupyter Notebook, isolando o projeto em um ambiente novo dedicado exclusivamente para testes. Validei os resultados dos dados utilizando o Power BI, garantindo a precisão e confiabilidade das informações.

Durante o desenvolvimento, implementei a técnica de chunking, que se mostrou mais eficiente para processar grandes volumes de dados neste projeto comparado ao streaming. Reconheço a importância do streaming em cenários onde a atualização contínua dos dados é crucial.

Para obter o melhor desempenho testei diferentes tamanhos de chunking com base na memória disponível, criei o arquivo py "teste chunksize" para validar o desempenho. Chunks pequenos sobrecarregam o disco, enquanto grandes consomem mais RAM. Equilibrar isso garante eficiência no processamento de dados, otimizando recursos e tempo de execução. O chunking de 450000 me proporcionou a leitura completa em 10.07 segundos, servindo como referência para testes subsequentes.

Essa metodologia não apenas garantiu a eficiência e precisão no processamento de dados, mas também demonstrou a flexibilidade e adaptabilidade das ferramentas e técnicas utilizadas para atender às necessidades específicas do projeto.

Resultados

0) 0 teste chunksize.py (Otimização de Performance com Chunking)

```
Memória usada durante a leitura do arquivo: 14.49 MiB
Tempo de execução: 10.07 segundos
Total de linhas processadas: 5000000
```

1) 1 teste.py (produto mais vendido em termos de quantidade e canal.)

```
   Sales Channel Item Type  Units Sold
2        Offline    Cereal  1044443977
22        Online    Snacks  1044143121

Process finished with exit code 0
```

2) 2 teste.py (país e região teve o maior volume de vendas em valor)

```
      País  País Units              Região  Região Units
0  Liberia    136188169  Sub-Saharan Africa    6486855992

Process finished with exit code 0
```

3) 3 teste.py (média de vendas mensais por produto)

```
         Month        Item Type  Total Units Sold  Average Units Sold
0      2020-12          Clothes            434934        14030.129032
1      2020-12           Cereal            402180        12973.548387
```