

Exercício de Contagem de Palavras usando Apache Beam

Objetivo:

Desenvolver um pipeline Apache Beam em Python para contar a ocorrência de palavras em um arquivo de texto. O exercício envolve a leitura de um arquivo de entrada, a contagem de palavras e a gravação dos resultados em um arquivo de saída.

Instruções:

1. Criação do Pipeline Apache Beam:

- Desenvolva um script Python que contém o código do pipeline Apache Beam.
- O pipeline deve ler o arquivo de entrada, contar a ocorrência de cada palavra e gravar os resultados no arquivo de saída.
- Utilize a transformação “ReadFromText” para ler o arquivo de entrada e “WriteToText” para gravar os resultados.
- Implemente uma transformação personalizada (“ParDo”) para contar as ocorrências de palavras.

2. Resultados Esperados:

- O arquivo de saída deve conter as palavras encontradas no arquivo de entrada e o número de vezes que cada palavra aparece.

3. Observações:

- Teste o pipeline com diferentes entradas para garantir a correta contagem de palavras.
- Explore os recursos do Apache Beam para aprendizado sobre transformações e pipeline de dados.

Este exercício visa proporcionar uma introdução prática ao desenvolvimento de pipelines de dados usando o Apache Beam.