Documentação do Projeto: Processamento de Dados de Grande Volume

Objetivo do Projeto

O objetivo deste projeto é avaliar a capacidade de manipular e analisar grandes volumes de dados em Python, de forma eficiente e com um consumo otimizado de recursos. Para isso, utilizamos um arquivo CSV denominado `vendas.csv`, que contém dados de vendas de uma cadeia de varejo, com aproximadamente 5GB de tamanho.

Requisitos

O desafio consiste nas seguintes tarefas:

1. Ler o arquivo `vendas.csv` de forma eficiente, considerando o grande volume de dados.
2. Identificar o produto mais vendido em termos de quantidade e canal.
3. Determinar qual país e região teve o maior volume de vendas (em valor).
4. Calcular a média de vendas mensais por produto.

Processo de Implementação

O projeto começou com a criação de uma sessão do Spark, que é a base para a execução de qualquer operação de dados no ambiente do Apache Spark. Utilizando o Spark, foi possível lidar eficientemente com o arquivo `vendas.csv`, que possui um tamanho considerável de 5GB.

Para ler os dados, utilizamos o método `spark.read.csv`, o que permitiu a leitura direta do arquivo armazenado no Databricks. Isso foi feito com a configuração adequada para garantir que a primeira linha fosse reconhecida como cabeçalho, facilitando a identificação das colunas.

Após a leitura do arquivo, a próxima etapa foi a limpeza e transformação dos dados. Um DataFrame foi criado e foram removidas as linhas com valores nulos nas colunas essenciais para as análises. Além disso, foram realizados ajustes nos tipos de dados,

assegurando que as colunas de quantidade e preço fossem convertidas para os formatos apropriados, como inteiro e float, respectivamente.

O uso do Spark facilitou a agregação dos dados. Para identificar o produto mais vendido, agrupamos os dados por tipo de item e canal de vendas, somando as unidades vendidas. Isso foi feito em uma única operação de grupo, permitindo que o Spark realizasse o processamento em paralelo, otimizando o tempo de execução.

Em seguida, para determinar qual país e região tiveram o maior volume de vendas, calculamos o valor total de vendas multiplicando as unidades vendidas pelo preço unitário. Essa operação também foi realizada de maneira eficiente, agrupando os dados por país e região.

Para calcular a média de vendas mensais por produto, extraímos o mês e o ano da data do pedido e agrupamos os dados de acordo. Novamente, o uso do Spark permitiu que essas operações complexas fossem executadas rapidamente, mesmo em um conjunto de dados grande.

Por fim, os resultados foram exibidos de forma clara, incluindo o produto mais vendido, o país e a região com o maior volume de vendas e a média de vendas mensais por produto. O Spark proporcionou um desempenho excelente ao lidar com essas operações em um ambiente que poderia, de outra forma, ser desafiador devido ao tamanho do conjunto de dados.