# Documentação - Solução para Desafio de Engenharia de Dados

Este documento descreve a implementação e explicação das soluções propostas para as questões práticas do desafio de engenharia de dados. As soluções foram implementadas em PySpark e abordam cálculos de médias de faturamento e consolidação de respostas em arquivos JSON.

## Questão 2: Média de Faturamento por Conta

Esta questão consiste em calcular a média de faturamento de cada conta nos últimos 3 e 6 meses retroativos a janeiro de 2020, utilizando os dados do arquivo invoices.csv. Quando não houver dados suficientes para a janela temporal, o valor será registrado como null.

Passos implementados:

1. 1. Leitura do arquivo invoices.csv utilizando PySpark.
2. 2. Conversão da coluna de datas para o formato apropriado.
3. 3. Filtragem dos dados até a data de referência (31 de janeiro de 2020).
4. 4. Criação de janelas de tempo de 3 e 6 meses usando funções de janela do PySpark.
5. 5. Cálculo das médias de faturamento para cada janela temporal.

6. Geração do resultado consolidado.

## Questão 3: Relatório Consolidado de Respostas

Esta questão exige a consolidação das respostas fornecidas pelos usuários em dois arquivos JSON (hour=13.json e hour=14.json). O objetivo é gerar um relatório tabular contendo informações como as primeiras e últimas interações válidas e os campos de resposta.

Passos implementados:

1. 1. Leitura dos arquivos JSON utilizando PySpark.
2. 2. União dos dados provenientes dos dois arquivos.
3. 3. Explosão do campo content para capturar as respostas fornecidas pelos usuários.
4. 4. Filtragem para remover respostas inválidas (vazias).
5. 5. Agregação dos dados para determinar os horários das primeiras e últimas interações, bem como organizar as respostas.

6. Geração do relatório consolidado em formato tabular.

## Instruções de Execução

Pré-requisitos:

1. 1. PySpark instalado.
2. 2. Arquivos de entrada (invoices.csv, hour=13.json, hour=14.json) disponíveis nos caminhos especificados no código.

Passos para execução:

1. 1. Salve os scripts fornecidos em arquivos com extensão .py.
2. 2. Execute os scripts utilizando o comando spark-submit, especificando o nome do arquivo. Por exemplo:

spark-submit questao\_2.py

spark-submit questao\_3.py