



Desafio Data Engineer

O objetivo deste desafio é analisar sua capacidade em pesquisar e desenvolver um pipeline de dados através da ferramenta [Apache Beam](#). Para isto, esperamos receber de você um script desenvolvido em **Python** e que seja capaz de executar todos os passos necessários para entregar os resultados solicitados.

Os arquivos necessários estão disponíveis através do arquivo compactado .zip [disponível para download neste link](#), no qual existem dois arquivos. Um extraído do IBGE com dados dos estados e outro um arquivo simulado de vendas:

1- EstadosIBGE.csv – Informações gerais dos estados

2- Vendas_por_dia.csv.csv – Dados fakes sobre vendas por dia e UF

Tarefa

Criar um pipeline de dados, utilizando do Apache Beam, que seja capaz de ler os dois arquivos em anexo e que o resultado desse pipeline sejam os dois seguintes arquivos:

1º arquivo

Agregado de informações por estado.

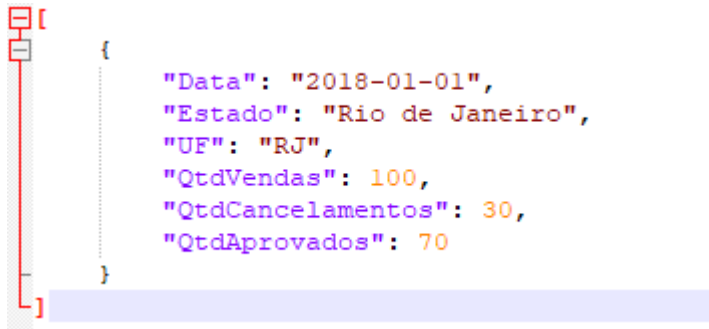
- Formato: CSV
- Informação: Data, Estado, UF, QtdVendas, QtdCancelamentos, QtdAprovados

2º arquivo

Com base nos mesmos resultados gerados no 1º arquivo, gerar um arquivo .json (válido) onde cada coluna do arquivo anterior, seja uma chave dentro desse Json.



Com base nos mesmos resultados gerados no 1º arquivo, gerar um arquivo .json (válido) onde cada coluna do arquivo anterior, seja uma key dentro desse json.

A screenshot of a code editor showing a JSON array with one object. The object has keys for date, state, city, and three counts. The text is color-coded: strings in purple, numbers in orange, and punctuation in red. A vertical line on the left indicates the array structure.

```
[  
  {  
    "Data": "2018-01-01",  
    "Estado": "Rio de Janeiro",  
    "UF": "RJ",  
    "QtdVendas": 100,  
    "QtdCancelamentos": 30,  
    "QtdAprovados": 70  
  }  
]
```

Entrega esperada

Código em um repositório público do GitHub, com o script desenvolvido junto aos dois arquivos gerados e um README.md explicando como foi desenvolvido o script e como o executar.

Serão considerados diferenciais:

- Utilização de boas práticas de desenvolvimento
- Versionamento contínuo do código no GitHub
- Documentação do código e do projeto
- Explicações sobre o porquê da utilização do Apache Beam, demonstrando domínio sobre o framework