topminds challenge stone.co

stone

# Data Challenge

uma iniciativa Stone.

## Case Engenharia

## Data Wrangling - Ingestion Engine

Responsável: Leonardo de Almeida Barbosa

Lead Data Engineer (BU Produtos Financeiros)

# Descrição do case

“We’re entering a new world in which data may be more important than software.”
– Tim O’Reilly, founder, O’Reilly Media.

O trabalho da engenharia de dados é coletar, catalogar, limpar, enriquecer e disponibilizar a maior quantidade de dados relevantes para nosso negócio. Esse tipo de tarefa envolve diferentes plataformas de dados, API’s, websites, tópicos e centenas de formatos diferentes de arquivos.

Além dos terabytes de dados gerados internamente na Stone, precisamos acompanhar indicadores que são públicos e estão disponíveis em diversos portais do governo, dentre eles a Procuradoria Geral da Fazenda Nacional. Essa base contém um conjunto de informações sobre débitos com a Fazenda Nacional e FGTS inscritos em Dívida Ativa em todas as situações, incluindo seus devedores, na condição de devedor principal, corresponsável ou solidário, e é atualizada trimestralmente.

Outro portal muito rico em informações é o do Banco Central, mais especificamente o SGS - Sistema Gerenciador de Séries Temporais, onde são disponibilizados indicadores de crédito, indicadores do mercado financeiro, indicadores de atividade econômica, entre outras séries.

O desafio é criar pipelines de dados que se conectem nesses portais e disponibilizem os dados em formato de tabelas para análise posterior.

# Dados

Para este desafio, vamos utilizar os dados abertos do site da Procuradoria-Geral da Fazenda Nacional referentes à dívida ativa geral e os dados do Bacen referentes aos indicadores de crédito, contidos nas séries 21388 a 21395.

Sua tarefa implica inicialmente em:

- Coletar todo o histórico disponível e armazenar ambas as bases no s3, respeitando as boas práticas de tipos de arquivos, particionamento, zonas de armazenamento comuns em um Datalake e anonimização dos dados exigidos pela LGPD.
- Criar tabelas no Athena de forma que os cientistas de dados possam analisar o histórico e as correlações entre os dados de dívidas e os indicadores de crédito.
- Criar uma chave única para consultar a base de dívidas e outra chave temporal para cruzamento com a base de indicadores.

# Requisitos

- Uma conta gratuita na AWS (necessário cartão de crédito para registrar)
- Uma conta gratuita no Github

# Entrega 1

Desenho da solução + código fonte no Github (enviar link do repositório privado no formulário). Esse código será utilizado para testar o funcionamento da carga dos dados em outra conta que não a do desenvolvedor.

Observação Importante: **Nunca publique as credenciais de acesso no seu repositório de códigos.**

# Entrega 2

Apresentação do projeto em 20 min, detalhando o problema e a solução implementada.

# Avaliação do case

Serão avaliados da seguinte forma:

A. Deployability (Roteiro de implantação)
B. Performance
C. Estrutura de códigos (Manutenibilidade)
D. Estrutura/organização de dados
E. Recuperação em caso de falhas (reprocessamento)
F. Arquitetura da solução (caixograma, custo da solução, etc.)
G. Documentação (caixograma)

Pesos:

| | | | | | | |
|---|---|---|---|---|---|---|
| A=1.5 | B=1.5 | C=2.0 | D=2.0 | E=1 | F=1 | G=1 |

A nota final será a soma das notas da seguinte maneira:

**nota final** = (A+B+C+D+E+F+G)

## 3. Avaliação da apresentação

- Conhecimento da solução implementada
- Storytelling
- Maiores dificuldades
- Benefícios da solução

# Banca avaliadora

**Composta por engenheiros e cientistas de dados da Stone.**