Processo Seletivo "Data Engineer"

Etapa Desafio Prático

Olá, tudo bem?

Você está recebendo um teste que tem como objetivo avaliar como você se comporta diante de um conjunto de dados e questões que queremos responder, dentro do escopo do engenheiro de dados. Fique a vontade para explorar soluções além dos requisitos propostos!

Os dados para a resolução do desafio estão disponíveis em anexo. Eles correspondem a uma amostra dos dados gerados pela plataforma do Passei Direto e são melhor detalhados a seguir, junto com as questões do teste.

1ª Parte

O Passei Direto é uma plataforma aberta com um modelo de negócios freemium, baseado em assinaturas. É muito importante para a saúde do negócio entendermos bem as características dos nossos usuários, principalmente os pagantes.

Nessa etapa será usada a *BASE A*, composta pelos seguintes Datasets:

- **students.json** - Amostra de usuários que acessaram o Passei Direto no mês de Novembro de 2017. Todos os outros datasets referenciam esses usuários
- **sessions.json** - Visitas que os usuários fizeram ao longo do mês
- **subscriptions.json** - Assinaturas dos usuários que aderiram ao Plano Premium do Passei Direto
- **universities.json** - Lista de Universidades cadastradas no Passei Direto
- **courses.json** - Lista de Cursos cadastrados no Passei Direto
- **subjects.json** - Lista de Disciplinas cadastradas no Passei Direto
- **student_follow_subject.json** - Disciplinas que cada usuário está seguindo

A sua tarefa é preparar uma base de BI a partir desses dados fornecidos. Essa tarefa envolve o seguinte:

- Definir a modelagem de uma base analítica para que seja otimizada a análise do perfil de nossos usuários e suas segmentações relevantes
- Incluir os dados que você julgar importantes

- Definir um pipeline dos dados. Estamos fornecendo uma base estática como exemplo, mas no mundo real esses dados são populados constantemente. Como você estruturaria a solução para que a base analítica seja mantida atualizada?
- Para essa etapa o uso de tecnologias é livre. Pode utilizar qualquer linguagem de programação e banco de dados que achar melhor. Tenha em mente apenas uma busca por simplicidade para a solução do problema

2ª Parte

Além dos dados da nossa base relacional, coletamos eventos dos usuários à medida que eles navegam na nossa plataforma. Esse tipo de informação é muito útil para entendermos melhor o comportamento do usuário, o que nos ajuda a planejar soluções melhores.

Nessa etapa será usada a *BASE B*, composta de apenas um Dataset:

- **part-[0000x].json** - Eventos de Page View que nossos usuários realizaram no dia 16 de Novembro de 2017. Esse Dataset foi dividido em alguns arquivos

O objetivo dessa parte é fazer um processamento desses eventos para entender melhor o comportamento dos usuários. Essa tarefa envolve o seguinte:

- Fazer o processamento da *BASE B* em conjunto com a *BASE A*
- Criar um pipeline para extrair informações importantes que permitam analisar o que os usuários do PD acessam na nossa plataforma e como fazem isso. Analise os dados para extrair as informações que você julgue serem relevantes para entender o comportamento do usuário
- Em relação à tecnologia, o único pré-requisito é a utilização do framework Spark para o processamento dos dados

Entregáveis

- Código-fonte utilizado para fazer o processamento e análise dos dados
    - Incluir os pré-requisitos e instruções para a reprodução da solução
    - Idealmente colocar em um repositório no Github
    - Enviar por email até o final do prazo combinado

Boa sorte, e nos vemos em breve.