Teste Técnico - Analytics Engineer & Data Engineer

Introdução

O PicPay é uma plataforma de pagamento criada para quebrar barreiras e eliminar burocracias. Nós existimos para melhorar a vida das pessoas e só conseguimos fazer isso porque temos pessoas fantásticas aqui dentro. Aqui você irá compor um time de talentos focado em construir um ambiente que proporcione a melhor experiência ao nosso usuário, independente do produto que ele busca.

Os profissionais de engenharia de dados devem entender o contexto do negócio e ajudar o time a construir uma base de dados sólida através das melhores práticas de mercado e identificando melhorias para os processos que sustentamos. Essa pessoa trabalhará de forma sincronizada, em um time que pensará em como disponibilizar dados de uma forma simples para destravar necessidades da empresa e empoderar os usuários de uma tomada de decisão baseada em dados.

Este teste tem como objetivo avaliar suas habilidades e conhecimento em coleta, ingestão e modelagem de dados, que são atividades frequentes na atuação como engenheiro de dados no PicPay. Será apresentado a você um conjunto de dados provenientes de uma base pública, que é amplamente utilizada no mercado (e por nós) e será seu desafio trazer e entender os dados, modelá-los em um formato que possibilite sua análise e responder às perguntas propostas. Fique à vontade para desenvolver a solução que, na sua visão, melhor atenda à necessidade proposta e de expandir quaisquer outras análises e visualizações que você julgue necessárias para o entendimento completo dos dados.

Recomendações

O teste pode ser realizado em linguagem Python, Pyspark e/ou SQL, na plataforma de sua escolha. Nós usamos Databricks, então se desejar, você pode utilizar o Databricks Community, basta criar uma conta gratuita e iniciar sua resolução. Ela é uma ferramenta que permite realizar todas as etapas necessárias para o teste.

Em se tratando da solução, mais uma vez reforçamos que você tem liberdade de utilizar as técnicas e soluções que melhor entender, mas recomendamos que os dados brutos sejam trazidos via request a partir da url pública (existe a possibilidade de download em máquina local também) e que sua manipulação seja feita como parte do código a ser criado para execução do teste (chamada, extração e materialização).

O resultado do teste deve ser encaminhado por email para o endereço marlon.pacheco@picpay.com em até **5 dias** da data do envio do teste, com o seguinte assunto: *[CASE TÉCNICO] Nome do Candidato - Credit Data Engineering*.

Abaixo você encontrará a descrição do caso e mais detalhes do que esperamos como sua solução. Agradecemos desde já pelo seu interesse e tempo dedicado para realização do teste proposto. Vamos lá!

Proposta

Você trabalha no time de engenharia de dados do Crédito, que atende de maneira cross diversas áreas consumidoras de dados. Chegou para o PM (Product Manager) um pedido de uma nova estrutura de dados que ainda não trabalhamos, que atenderá principalmente o time de modelagem de crédito, composto por cientistas de dados, e o time de políticas de crédito, composto por analistas de negócio. As informações foram passadas para você pelo PM por mensagem e são as seguintes:

Contexto

Recebemos um pedido de consumo dos dados do SIAPE do time de Modelagem e do time de Políticas. O time de modelagem gostaria de incluir os dados de remuneração nos modelos de renda e de risco de crédito; já o time de políticas, gostaria de realizar alguns estudos para adaptação da política de concessão de crédito pessoal e consignado para alguns grupos de ocupação. Apesar de serem necessidades distintas, ambos nos pediram para que a base fosse consumida em sua posição mais atual, que é sempre M-2, além de 3 meses de histórico. Será necessário adaptar a tabela final para receber novas posições, pois atualizaremos de maneira recorrente a partir da produtização da tabela.

Nos foi passado também que esses dados em partes já foram estudados pelos times, mas a complexidade é muito grande, pois são muitas bases e muitas colunas para manipular, e o processo está sendo muito moroso e manual, então precisamos entregar um produto de dados robusto, automatizado, mas simples para consumo. Sabemos de algumas premissas que podemos antecipar:

- Os estudos e modelagens serão feitos a nível de CPF e por posição do arquivo (é importante que seja possível voltar no tempo nessa base conforme formos alimentando ela); e
- É necessário estar identificada a fonte de cada informação, caso seja criada apenas uma tabela final.

Informações técnicas

Os dados que devem ser consumidos vem do Portal da Transparência. Os links para acesso aos dados são os seguintes:

- Download manual
- Dicionário de dados

Cada uma das requisições trará uma estrutura com 3 ou 4 arquivos: remuneração, observações, cadastro e afastamento.

Como não temos um *schema* definido para esse tipo de informação, vamos criar um *schema* chamado *public_informations*. Esse *schema* será posteriormente alimentado com outras tabelas de mesma natureza.

Por se tratarem de muitos dados, e que faremos um append das informações mensalmente, precisamos particionar os dados. Isso deve ser feito pela posição da base.

Definição de finalizado

Para concluirmos essa tarefa, precisamos garantir que as seguintes entregas sejam feitas:

1. Nome da(s) tabela(s) a ser(em) produtizada(s);
2. Código fonte de criação das tabelas, com todas as manipulações que serão feitas, pronto para subida em produção; e
3. Código para levantamento das estatísticas descritivas do(s) produto(s) final(is).