**Desafio​ ​Técnico**

**Time de Inteligência**



**Orientações**

Você está recebendo o desafio técnico para a vaga do time de inteligência de dados. Esse desafio está dividido em 4 etapas.

1. Ingestão dos dados (**PostgreSQL**);
2. Responder aos questionamentos da área de negócio com consultas **SQL**;
3. Responder aos questionamentos da área de negócio com **Python/Pandas**;
4. Proposta de Arquitetura de Dados;

**Regras​ gerais:**

* Você​ deve utilizar​ todos​ os​ arquivos​ enviados.
* Comente o seu​ código,​ pois​ isso ajuda​ muito no​ processo​ de​ avaliação, lembre-se​ que​ outra​ pessoa​ irá avaliar​ seu​ código.
* Você​ deve​ enviar​ todos​ os​ arquivos​ utilizados​ durante​ a resolução dos​ cases.

**Observações:**

* Airflow, Scikit-Learn, Kinesis, Kafka, Spark são conhecimentos **desejáveis**. Ou seja, conhecimento sobre essas tecnologias serão **vistos como um diferencial e não como algo​ obrigatório!**

**Contexto:**

Você recebeu a tarefa de ajudar os analistas de BI a responder alguns questionamentos sobre o faturamento da Petlove. Depois de algumas pesquisas e conversas com desenvolvedores você conseguiu acesso aos dados em vários formatos de arquivo, abaixo temos a relação dos arquivos obtidos e suas respectivas descrições:

**faturamento.parquet.gzip -** Dados das notas faturadas.

|  |  |
| --- | --- |
| **Campo** | **Descrição** |
| id\_nota | Id da nota. |
| sku | Código do produto. |
| dia\_emissao\_nota | Data de emissão da nota. |
| uf\_entrega | UF de entrega da mercadoria. |
| quantidade | Quantidade faturada na nota. |
| receita | Valor faturado. |

**familiasetor.json** - Familia e setor por SKU.

|  |  |
| --- | --- |
| **Campo** | **Descrição** |
| sku | Código do produto. |
| setor | Setor do produto (Ex: Alimentos, Brinquedos, e etc.) |
| familia | Agrupamento por tipo de pet (Ex.Gatos,Cachorros e peixes.) |

**peso\_unitario.txt** - Peso unitário do SKU. Usado para o custo de frete.

|  |  |
| --- | --- |
| **Campo** | **Descrição** |
| sku | Código do produto. |
| peso\_unitario | Peso unitário do produto. |

**frete.csv** - Valor gasto com o frete.

|  |  |
| --- | --- |
| **Campo** | **Descrição** |
| uf\_entrega | UF de estado que foi entregue o produto. |
| dia | Data da entrega do produto. |
| custo\_frete | Custo total do frete. |

**Etapa 1: Ingestão dos dados.**

Nessa etapa você deve inserir os arquivos recebidos em um banco de dados (SGBD), esse banco deve ser um PostgreSQL, você pode usar python ou as ferramentas disponibilizadas pelo próprio PostgreSQL para completar a tarefa.

**Etapa 2: Consultas SQL.**

Nessa etapa você deve escrever as consultas em SQL que foram solicitadas pelos analistas de BI. Você deve salvar o resultado de cada consulta em um arquivo .csv e enviar junto com os *scripts* .sql.

**Q1** -> Escreva uma única consulta que apresente a quantidade total de SKUs distintos comprados em 2018 na família "Cachorros" e também o total de SKUs distintos comprados em cada mês em 2018 na família "Cachorros".

**Q2** -> Escreva uma única consulta que aplique a regra de rateio do frete definida com a àrea de negócio e apresente o custo total de frete gasto no Setor "Alimentos".

\*\*\*Regra rateio = (Custo de frete / Peso total) \* Peso unitário

**Q3** -> Escreva uma única consulta que apresente o crescimento percentual de vendas em relação ao último mês para todos os meses a partir de 2018. Apresente também a variação em relação ao mesmo mês do ano anterior para todos os meses.

**Q4** -> Escreva uma única consulta que apresente o percentual de vendas das famílias Cachorros, Gato e Peixes em relação ao total para todos os meses em 2018.

**Etapa 3: Consultas Python/Pandas.**

Nessa etapa você deve reescrever as consultas que foram feitas com SQL em Python/Pandas. Você deve salvar o resultado de cada consulta em um arquivo .parquet e enviar junto com os *scripts* .py (ou .ipynb caso use jupyter notebook)

**Etapa 4: Visualização dos dados.**

Usando os dados que foram compartilhados e a ferramenta de visualização que você tiver familiaridade, monte um dashboard simples para mostrar os indicadores que que você julgar mais interessantes (não precisa ser necessariamente os resultados das queries solicitadas nas perguntas anteriores).

**Etapa 5: Proposta de Arquitetura de Dados.**

Nessa etapa, você deve propor uma arquitetura de dados que ajude as áreas de negócio da petlove responder mais perguntas sem depender (muito) da área de engenharia de dados/BI.

Na conversa que foi feita com os desenvolvedores, você apurou as seguintes características do ambiente de tecnologia/dados da petlove:

* Cada arquivo que você recebeu foi extraído de um sistema diferente.
  + O arquivo de “**faturamento.parquet.gzip**” foi extraído do banco de dados do ERP.
  + O arquivo ”**peso\_unitario.txt**” foi extraído do sistema de remessa de mercadorias, através da interface do sistema por um operador no centro de distribuição. Você tem acesso ao banco de dados por uma VPN.
  + O arquivo “**familiasetor.json**” foi extraído da API do CRM.
  + O arquivo “**frete.csv**” foi extraído da API do sistema contábil.

* Algumas restrições para extração dos dados também foram identificadas:
  + O ERP e o sistema de remessa só permitem conexão às bases de dados no horário fixo das 00:00 até 04:00
  + A API do CRM tem momentos curtos (1 à 5 minutos) de indisponibilidade devido ao grande volume de acessos.
  + API do sistema contábil tem um “*endpoint*” com histórico disponível.

Com essas informações em mãos, descreva e justifique uma arquitetura de dados que seja capaz de lidar com a extração e armazenamento dessas fontes de dados, disponibilizando-os para uso por qualquer área de empresa. Os dados serão atualizados **diariamente (D-1)**. Você pode utilizar a linguagem de programação e as tecnologias que mais lhe agradam ou as que você julgar mais adequadas.

Boa sorte !