Engenheiro de dados

Teste técnico para engenheiro de dados para a plataforma de dados da B2W Digital.

O teste é simples e o tempo médio para fazer é de 2 horas.

Qual o objetivo do teste?

O objetivo é avaliar com você projetou e desenvolveu uma solução para o cenário proposto, observando a arquitetura, estilo de código e documentação.

Restrições

As restrições para o teste são:

* O código-fonte deve ser versionado com o `git`
* O projeto deve incluir um arquivo `README.md`
* O arquivo `README.md` deve incluir instruções para rodar o projeto **localmente**. Usaremos uma estação de trabalho com **macOS** ou **Ubuntu** para validar o teste
* O projeto deve ser desenvolvido usando o **Apache Beam** usando **Python** ou **Java**
* Os arquivos de entrada devem ser armazenados no diretório `input` e os de saída no diretório `output`

Onde publicar o projeto?

Você pode usar um repositório público no GitHub ou GitLab e passar a URL para clonarmos e validarmos.

Ou você pode usar um repositório privado também no GitHub ou GitLab e compartilhar conosco para clonarmos e validarmos. Nossos usuários são:

* `brunitto`
* `tiodollar`

Em último caso, também aceitamos um pacote compactado `.zip` contendo o projeto todo, incluindo o diretório `.git`.

Cenário proposto

Desenvolver um job para encontrar carrinhos abandonados pelos clientes de um e-commerce. Embora o assunto seja rico, no teste a definição de abandono de carrinho será simplificada.

Regra de abandono de carrinho

A regra de abandono de carrinho é simples.

* Definimos como sessão, uma janela de 10 minutos onde o cliente interage (visualiza páginas) no nosso site
* O tempo de sessão é renovado a cada nova interação
* O fluxo de páginas padrão de um pedido é: product -> basket -> checkout

* Um abandono pode ser identificado por um fluxo interrompido na página basket: product -> basket dentro de uma sessão

Por exemplo:

* Um cliente visualiza a página de produto (product) às 12:00
* O mesmo cliente visualiza a página de carrinho (basket) às 12:02
* O mesmo cliente visualizar a página de pagamento (checkout) às 12:04
* Nào temos uma abandono :)

Outro exemplo:

* Outro cliente visualiza a página de produto (product) às 13:00
* O mesmo cliente visualiza a página de carrinho (basket) às 13:01
* O cliente fica 15 minutos sem visualizar nenhuma página (pode ter ido tomar um café). Como sua última interação foi às 13:01, sua sessão terminou às 13:11
* Aqui temos um abandono :(

Mais um exemplo:

* Um terceiro cliente visualiza a página de produto (product) às 14:00
* O mesmo cliente visualiza a página de carrinho (basket) às 14:05
* O mesmo cliente visualizar outra página de produto (product) às 14:10
* O mesmo cliente visualiza ainda outra página de produto (product) às 14:16
* O mesmo cliente visualiza a página de carrinho (basket) novamente às 14:20
* O mesmo cliente visualiza a página de pagamento (checkout) às 14:21
* Não temos um abandono :)

Testes

Executar o job, lendo o arquivo `input/page-views.json`e escrevendo no arquivo `output/abandoned-carts.json`.

O arquivo `input/page-views.json` deve conter algumas visualizações de páginas:

```
    { "timestamp": "2019-01-01 12:00:00", "customer": "customer-1", "page": "product", "product": "product-1" }
    { "timestamp": "2019-01-01 12:02:00", "customer": "customer-1", "page": "basket", "product": "product-1" }
    { "timestamp": "2019-01-01 12:04:00", "customer": "customer-1", "page": "checkout" }
    { "timestamp": "2019-01-01 13:00:00", "customer": "customer-2", "page": "product", "product": "product-2" }
    { "timestamp": "2019-01-01 13:02:00", "customer": "customer-2", "page": "basket", "product": "product-2" }
    { "timestamp": "2019-01-01 14:00:00", "customer": "customer-3", "page": "product", "product": "product-3" }
    { "timestamp": "2019-01-01 14:05:00", "customer": "customer-3", "page": "basket", "product": "product-3" }
```

```
    { "timestamp": "2019-01-01 14:10:00", "customer": "customer-3",
"page": "product", "product": "product-4" }
    { "timestamp": "2019-01-01 14:16:00", "customer": "customer-3",
"page": "product", "product": "product-5" }
    { "timestamp": "2019-01-01 14:20:00", "customer": "customer-3",
"page": "basket", "product": "product-4" }
    { "timestamp": "2019-01-01 14:21:00", "customer": "customer-3",
"page": "checkout" }
```

O arquivo `output/abandoned-carts.json` deve conter apenas 1 abandono de carrinho:

```
    { "timestamp": "2019-01-01 13:02:00", "customer": "customer-2",
"product": "product-2" }
```

Bônus

Se sobrar um tempinho, tente:

* Ler de vários arquivos
* Escrever usando o cliente como partição
* Incluir testes unitários para o job

Referências

* https://beam.apache.org/