Data Scientist - Dito

Esse desafio está divido em duas partes, sendo a primeira delas um problema que envolve SQL e a segunda Análise Exploratória de Dados. Em ambas é esperado que você envie as soluções através de um repositório público no Github.

SQL

Contexto

Na Dito utilizamos o Google BigQuery como fonte para diversas análises. Os dados estão estruturados em uma única tabela. As linhas representam eventos e possuem a coluna `type` que pode conter os valores `"track"` e `"identify"`.

Eventos de "identify"

As linhas com `type = "identify"` são eventos que marcam a alteração de características das pessoas como nome, email e telefone. Para uma mesma pessoa podem existir vários eventos de identify ao longo do tempo, que correspondem às diversas vezes que essa pessoas foi identificada ou teve seus dados atualizados na Dito.

Um exemplo: uma pessoa pode mudar seu telefone ao longo do tempo; portanto o telefone atual de uma pessoa corresponde ao registro do tipo identify mais recente onde a coluna telefone esteja preenchida.

As características das pessoas estão guardadas no objeto `traits`. Por exemplo, acessar nome e email seria, `traits.name` e `traits.email`, respectivamente.

Eventos de "track"

As linhas com `type = "track"` representam ações das pessoas. Um evento do tipo "track" pode ser uma compra, um login no site ou a navegação online em uma página de produto. As ações desses eventos são diferenciados pela coluna `properties.action`.

> No contexto desse exercício, apenas o evento `"buy"` existe nos dados.

Os atributos dos eventos estão guardados no nó `properties`. Por exemplo: a receita gerada a partir de uma compra pode ser acessada pelo campo `properties.revenue`.

Timestamp dos eventos

Tanto os eventos `"identify"` quanto os eventos `"track"` possuem uma coluna chamada `timestamp`, que corresponde à data/hora que de fato aconteceu a identificação do usuário

ou a ação. Ambos os tipos possuem a coluna `id` que identifica uma pessoa.

Preparação

Antes de tudo você deve criar uma conta Google Cloud. O processo é bem rápido e gratuito, sendo necessário apenas estar logado com algum email Google.

O próximo passo é criar um projeto Google Cloud. Isso pode ser feito nesse link.

Nesse momento você já consegue acessar o Dataset público criado pela Dito para esse desafio. Basta abrir o Console do BigQuery e fazer a seguinte consulta de teste, que lista os 5 produtos que mais geraram receita:

```
SELECT
  properties.product,
  SUM(properties.revenue) revenue
FROM `dito-data-scientist-challenge.tracking.dito`
WHERE type = 'track'
GROUP BY properties.product
ORDER BY revenue DESC
LIMIT 5
```

Desafio

O desafio será responder duas perguntas a partir do Dataset fornecido e usando **apenas SQL**.

1. Qual o nome, email e telefone das 5 pessoas que mais geraram receita?
2. De quantos em quantos dias, em média, as pessoas compram? Use a mediana como média.

Você deve **entregar as consultas SQL** das duas perguntas (não apenas as respostas) e uma breve explicação de como você pensou.

Send Time Optimization

Contexto

Os clientes da Dito podem entrar em contato com seus usuários por diversos canais através do envio de notificações.

Atualmente, o canal mais utilizado é o E-mail. Queremos ajudar nossos clientes a obterem a maior taxa de abertura de e-mails possível.

> Taxa de abertura é o número de pessoas que abriu a mensagem dividido pelo número de pessoas que recebeu a mensagem.

Para isto, vamos implementar a funcionalidade de Send Time Optimization.

Desafio

O objetivo do Send Time Optimization é encontrar, para cada destinatário, qual a **hora do dia** para se fazer o envio de um e-mail que maximize a taxa de abertura.

Você deve utilizar os dados desse CSV para criar uma solução de Data Science. O arquivo contém eventos relacionados ao disparo de emails, com os seguintes campos:

- id: identificador do usuário
- timestamp: data/hora que aconteceu o evento, no formato `2018-05-25 14:59:02 UTC`
- email_id: identificador que diferencia disparos de e-mail
- action: qual ação representa o evento

Você vai encontrar as seguintes ações:

- received: usuário recebeu o e-mail
- open: usuário abriu o e-mail
- click: usuário acessou algum link presente no e-mail
- spamreport: usuário reportou spam
- unsubscribe: usuário se descadastrou

Você deve fazer uma **Análise Exploratória dos Dados** e entregar um **storytelling** que embase a sua solução de Send Time Optimization.

Para nós é importante entender todas a análises feitas para chegar na solução. Por isso, envie os códigos e qualquer outro material utilizado no processo.

#dito/careers