

# Desafio técnico

## Data Engineer

### Resumo

Criar um fluxo de processamento dos dados com ingestão, transformação e entrega, que possibilite analistas de BI da área de Email Marketing construirem Dashboards das principais métricas operacionais diária e mensalmente.

### O que é Email Marketing

A área de Email Marketing é responsável por gerenciar este canal de relacionamento com os contatos da empresa com o objetivo de guiá-los durante a jornada de compra, aumentando o engajamento e interesse nos produtos oferecidos para, ao final, elevar a taxa de conversão de leads em clientes.

### Conceitos importantes

Abaixo são introduzidos os principais conceitos relacionados aos requisitos apresentados na seção “O que o produto de dados deve possibilitar”.

- Funil de Marketing-Vendas : em resumo, é o conjunto de etapas da jornada de compra que inicia-se em lead e finaliza em cliente.



Abaixo, a descrição das etapas relevantes para a criação do fluxo:

Uma descrição abrangente do conceito de Inbound Marketing e do funil e Marketing-Vendas pode ser acessada neste [blog post](#).

- **Lead:** um visitante do site da empresa que preenche um formulário de contato deixando seu email (ação chamada de “conversão”). Isso ocorre, por exemplo, no momento em que o visitante faz o download de um material rico (e-book, por exemplo) criado pela RD;
  - **MQL:** Marketing Qualified Lead. São leads que demonstram interesse nas ofertas da empresa e possuem perfil adequado para serem clientes;
  - **SAL:** Sales Accepted Lead. São MQLs que foram verificados pela equipe de pré-vendas e são considerados válidos para serem abordados pelo time comercial;
  - **Etapas seguintes:** os demais estágios são Oportunidade e Cliente, as quais estão sob responsabilidade da equipe de Vendas. Como o escopo deste dashboard é a equipe de Marketing, lidaremos apenas até a etapa SAL. Mais detalhes e contextualização destas etapas são descritas neste [blog post](#);
  - **Taxas de conversão:** as taxas de conversão medem o aproveitamento entre etapas do funil. Por exemplo, se a cada 100 leads 10 se tornam MQL, a taxa de conversão Lead-MQL é de 10%. As taxas entre as etapas são métricas relevantes na avaliação da performance da empresa.
- 
- **First / Last touch:** cada conversão do lead é considerada um touch, ou ponto de contato com as iniciativas oferecidas pela empresa (materiais ricos, eventos, etc). Durante a jornada de compra os leads podem realizar múltiplas conversões. Para otimizar a geração de leads, toda a trajetória de conversões é analisada, com ênfase para a última, a qual é chamada de last touch. É desejável, também, analisar as métricas do ponto de vista da primeira conversão (first touch).
  - **Perfil dos leads:** como descrito sobre MQL, é importante que, a partir desta etapa, só continuem no funil os leads com perfil adequado. Entre diferentes atributos que constituem o perfil, o mais observado na empresa é o ICP (Ideal Customer Profile) Score: uma avaliação de A a D de acordo com a “qualidade” do lead (sendo A o melhor). Esta categorização é crucial para a operação da equipe de Email Marketing, uma vez que é desejado que sejam gerados mais leads ABC do que D.
  - **Tipo / origem do lead:** dependendo do canal usado pelo lead no last touch, classifica-se entre os tipos:
    - **MQL (mesmo nome da etapa do funil):** leads que convertem em Landing Pages (LPs) “convencionais”

- **HR (hand raise):** leads que convertem em LP sinalizando claramente o interesse de compra (ex. pedido demonstração do produto)
- **Indicação:** leads que convertem em LPs de indicação interna

Estes formulários residem, em geral, em páginas chamadas de Landing Pages, as quais são identificadas pelo atributo *identifier*.

- **Chat Online:** leads que entram em contato via aplicativo de chat do site

Essa classificação é importante para a priorização do atendimento pela equipe de pré-vendas. Além disso, leads do tipo HR possuem maior probabilidade de contratar o produto e o percentual destes no total de leads é uma métrica relevante.

## O desafio

O resultado deste trabalho deve ser um produto de dados que facilite o acompanhamento do desempenho das diferentes campanhas, gerando insights e subsídios para a equipe de Email Marketing otimizar a geração de Leads, MQL e SAL ao longo do mês.

O que o produto de dados deve implementar

- Ingestão dos arquivos CSV (RAW)
- Implementar uma camada de limpeza
  - Com remoção dos dados repetidos. Quando dois registros possuírem o mesmo ID, selecionar apenas um como válido
  - Remoção de registros inválidos, com Id = ""
  - Converter os dados conforme o tipo observado (tipos de dados: data, timestamp, inteiro e etc)
  - Converter o timezone das colunas de tipo timestamp. Os valores foram entregues em UTC na camada RAW e precisa ser convertido para UTC - 3 na camada Clean.
- Camada de entrega: Criar uma tabela com
  - Comparação da geração de leads vs. meta ("Actual vs. budget") no horizonte de tempo diário e mensal
    - Geração de leads
      - Qtde. de leads, MQL e SAL vs. respectivas metas
      - Taxas de conversão entre etapas
      - Month-to-date (MTD) das métricas acima em relação à meta diária. A meta diária pode ser a meta mensal dividida por 30.
      - Qtde. de e-mails únicos por dia.
      - Último horário de conversão para cada dia
    - Colunas de segmentação
      - Account (Brasil, Partners, Latam, Portugal)
      - ICP Score ("Perfil")

- Tipo de lead (MQL, HR, etc)
- País

## O que também deve ser entregue?

Responder questões de negócio:

Além do acompanhamento das métricas principais, deve ser entregue:

- Liste as top 5 campanhas de email que geram mais Leads, MQL e SAL? Incluir quebras por ICP Score.
- Liste as top 5 campanhas de email que geram mais leads do tipo HR?

## Dataset

Você receberá dois conjuntos de dados:

- **bi\_challenge\_rd\_bi\_funnel\_email.csv**: tabela na qual cada linha é um lead e as colunas são atributos de suas etapas no funil, com dados de 1 mês e meio (de 01/11/2018 a 15/12/2018).

Colunas de interesse:

- \*\_created\_month: mês em que o lead, MQL e SAL são contabilizados
- campaign\_last\_touch: nome da campanha em que o lead converteu em MQL
- identifier\_last\_touch: nome da landing page em que o lead converteu em MQL
- converted\_to\_\*: flags booleanas que indicam se o lead atingiu os próximos estágios do funil
- is\_\*: flags booleanas que indicam atributos que o lead possui
- account\_last\_touch: nome da conta de origem do lead
- icp\_score: perfil do lead (A a D)
- lead\_type: tipo/origem do lead (MQL, HR, etc)
- **metas\_email.csv**: arquivo com as metas mensais de geração de leads, MQL e SAL para a área de Email Marketing.

## Formato de entrega

O fluxo de processamento poderá ser desenvolvido no banco a sua escolha, desde que:

- Os dados devem ser consumidos de um banco de dados relacional open source (ex. PostgreSQL, MySQL, etc);
- As transformações de dados precisam ser majoritariamente realizadas no próprio data warehouse (ex. join, group by, last\_value).

- Para avaliar a entrega, é necessário que possamos avaliar os resultados produzidos. Entregue em CSV os dados que foram gerados nas tabelas finais.
- Elaborar instruções básicas para acessar o projeto desenvolvido (“README”, máx. ½ página)
- Compartilhar o código que foi produzido no Git ou repositório de preferência para e-mails [thiago.rudzewicz@rdstation.com](mailto:thiago.rudzewicz@rdstation.com), [fernando.brangaca@rdstation.com](mailto:fernando.brangaca@rdstation.com) e [estevao.radaelli@rdstation.com](mailto:estevao.radaelli@rdstation.com).

## Opcional

As etapas a seguir são opcionais, mas sua execução poderá destacar você no processo seletivo de Engenharia de Dados, demonstrando domínio técnico e boas práticas de desenvolvimento:

- Criar orquestração das etapas de processamento de dados;
- Documentar tabelas e colunas, incluindo suas descrições;
- Definir contratos de dados (data contracts) para validar e tratar situações em que os dados não estejam em conformidade com o tipo esperado.