**Teste prático (work-at-olist-data)**

Esta é a página sobre o teste prático para os candidatos ao processo seletivo de cientistas, analistas e engenheiros de dados para o time de Business Science & Analytics do olist. Ao finalizar o teste, nos comunique pelo [**https://bit.ly/2NNVyXk**](https://bit.ly/2NNVyXk).

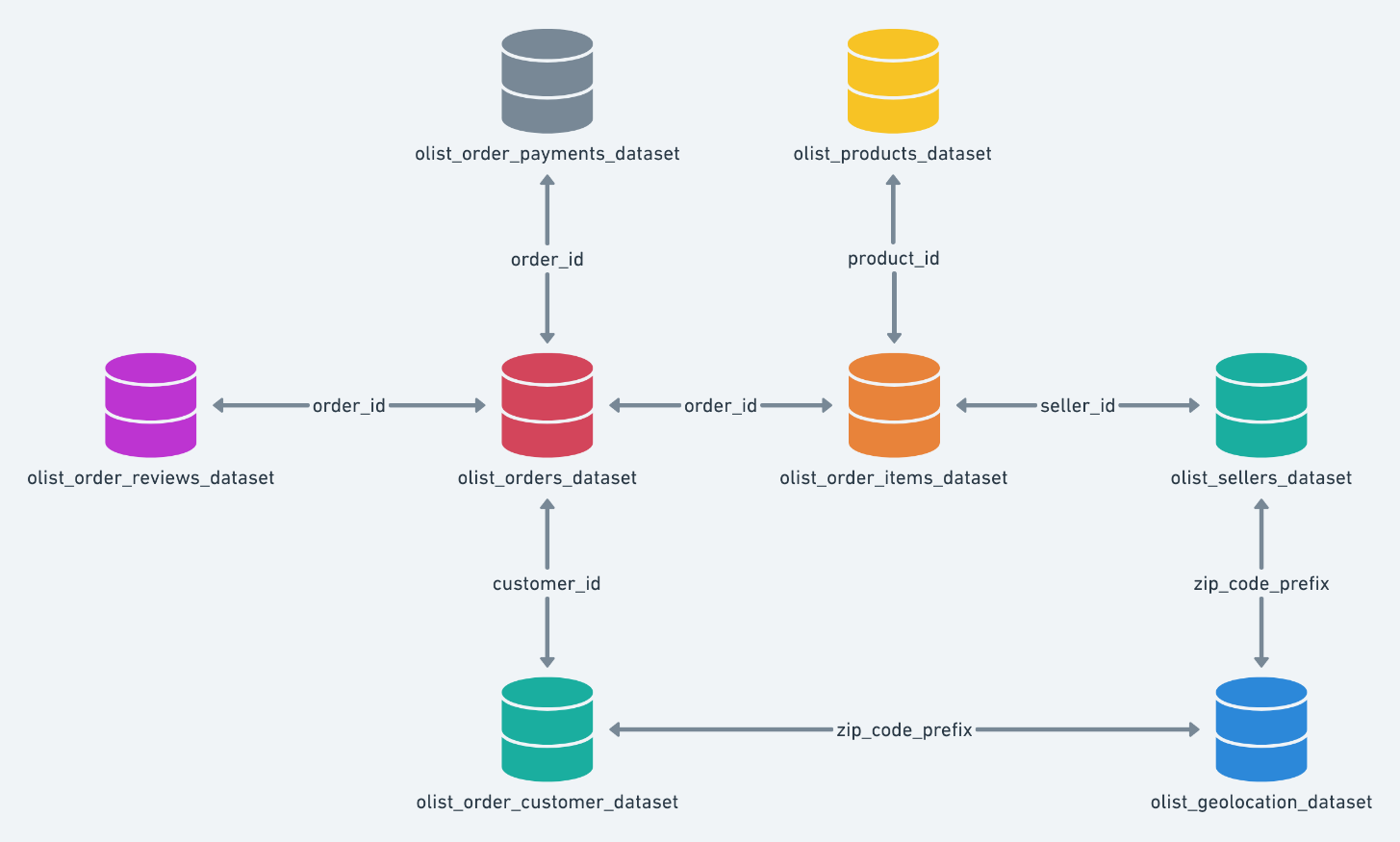
**Cenário**

O olist é a maior loja de departamentos dos marketplaces. Possui um catálogo com mais de 750 mil produtos, centenas de milhares de pedidos e uma rede de mais de 9 mil lojistas parceiros espalhados por todas as regiões do Brasil. Entendemos que a área de dados e inteligência é uma das principais alavancas de crescimento do negócio, por isso buscamos profissionais apaixonados por dados para integrar a nossa equipe de Business Science e Analytics (BSA).

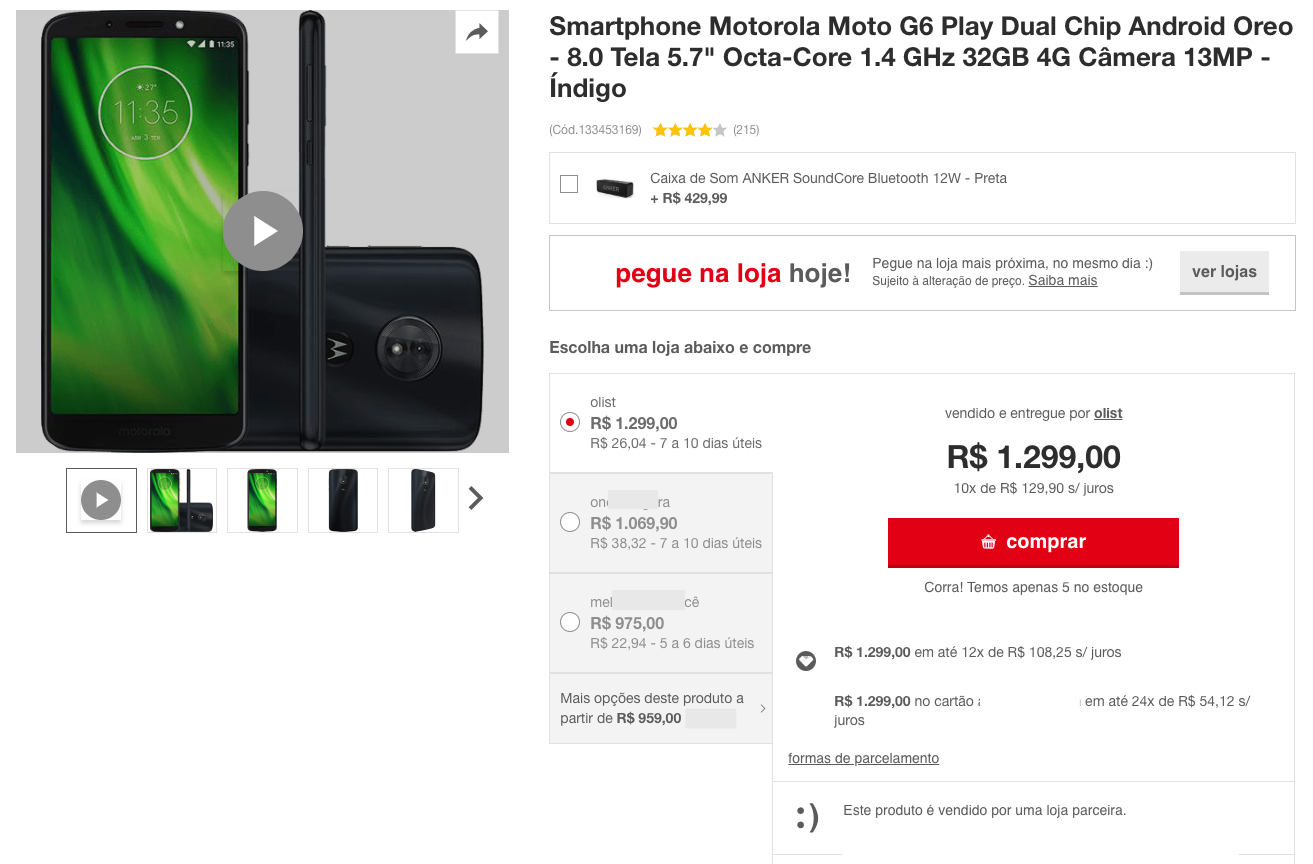
Estamos o tempo todo gerando dados, dados e muito mais dados. Nosso cenário é de big data!

**Como é o nosso banco de dados?**

Considere que a figura abaixo seja a estrutura do nosso banco de dados de e-commerce e que cada tabela desse banco de dados esteja disponível como um dataset. [Aqui estão os datasets](https://github.com/olist/work-at-olist-data/blob/master/datasets). Perceba que os dados são amostrais e estão apresentados no modelo relacional, pois foram extraídos diretamente, sem qualquer tratamento, do ambiente transacional.

[](https://github.com/olist/work-at-olist-data/blob/master/images/schema.png)

Bom, e aqui você pode observar como um produto é exibido na plataforma de e-commerce.

[](https://github.com/olist/work-at-olist-data/blob/master/images/example.png)

**Desafio**

Esta plataforma de e-commerce transaciona dados do varejo brasileiro. Assim, as áreas de negócio da empresa estão com algumas dúvidas e desafios como:

* Será que nossos diferentes lojistas associados conseguem manter o preço do mesmo produto sem grandes discrepâncias?
* Podemos dar os mesmos benefícios para todos os lojistas (sellers)? Ou existe algum que merece destaque?
* Existe diferença no valor do frete praticado em regiões/cidades diferentes? Ou podemos aplicar as mesmas regras de subsídio de frete para qualquer localidade?
* Será que nosso catálogo de produtos é abrangente? Ou tem foco em categorias específicas?
* Será que sempre vendemos os mesmos produtos? Ou existem sazonalidades?
* Será que existe um modelo preditivo para nos preparar para o futuro?
* Será que o atual banco de dados vai suportar o nosso crescimento? Ou existe uma opção mais escalável?

Ficou confuso(a). Segue abaixo, dicas mais específicas para cada função.

**Cientista de dados**

* Que tal uma análise textual dos clientes que deixaram comentários sobre suas compras?
* Alguns clientes não escreveram um comentário. Mas por que eles estão satisfeitos?
* Com as informações da data de compra, você poderá prever vendas futuras!
* Você também poderá focar na logística e encontrar maneiras de otimizar os tempos de entrega.
* Esses dados possuem coordenadas de geolocalização. Há diferenças no padrão de consumo por regiões?
* Divirta-se descobrindo as categorias de produtos mais propensas à insatisfação do cliente.
* Crie recursos deste rico conjunto de dados, feature engineering ou anexe algumas informações públicas externas a ele.
* E um modelo para precificar os produtos do nosso catálogo? Modelagem matemática para otimizar rotas? Testes de hipóteses para validar algum questionamento?
* O que acha de escrever um relatório ou slides detalhando as suas descobertas?
* Fique livre para criar sua própria abordagem, caso considere que as dicas anteriores não sejam pertinentes.

**Analista de dados e Business Intelligence**

* Pense em alguns KPIs para monitoramento. Talvez outros para direcionamento dos gestores!
* Um cruzamento dos dados poderia gerar relatórios interessantes. Afinal, quem são os Top 10 em vendas? Que tipo produtos eles vendem? Qual é o impacto deles para o negócio?
* Que tal realizar uma análise exploratória dos dados. E então? Algo lhe chama a atenção?
* Você poderia apresentar esses dados em um dashboard. Isso daria agilidade na tomada de decisão!
* Temos interesse em suas habilidades com matemática aplicada e estatística descritiva. O que você pode nos mostrar com os dados?
* O que acha de escrever um relatório ou slides detalhando as suas descobertas?
* Fique livre para criar sua própria abordagem, caso considere que as dicas anteriores não sejam pertinentes.

**Engenheiro(a) de dados**

* Gostaríamos de analisar suas habilidades com SQL, modelagem dimensional e integração de dados. Mostre seus conhecimento em processos de ETL e conceitos de Data Warehouse. Que tal replicar nossos datasets, remodelar em um banco de dados e apresentar as melhorias realizadas em sua criação.
* É possível utilizar o modelo proposto em um ambiente cloud? Quais plataformas ou serviços você utilizaria? Quais as vantagens do modelo escolhido em questões de performance?
  + Sim. O projeto proposto foi elaborado na plataforma Microsoft Azure. A vantagem é a escalabilidade e redundância.
* Alguns membros do time dizem que a atual modelagem do banco de dados é adequada para o uso dos cientistas de dados e analistas de BI, porém, outros dizem que existem formas de modelar bancos de dados que trarão mais eficiência. Qual é a sua opinião sobre isso?
  + Acredito em 2 formatos de disponibilização dos dados. O primeiro basedo no data warehouse existente e outro com a normalização das informações para que os cientistas possam utilizar em suas aplicações.
* Estamos preocupados com o vertiginoso aumento do volume em nosso banco de dados atual? Você consideraria uma opção mais escalável ou devemos manter a estrutura existente?
  + A vantagem da utilização da cloud para o banco de dados é a escalabilidade rápida sem a necessidade de investimento em infra-estrutura.
* Nossa ferramenta de visualização de dashboards está lenta e o nosso time detectou que o problema está na infraestrutura de dados. Como você abordaria esta situação do ponto vista de arquitetura de dados?
  + Criar um modelo intermediário que já faça os cálculos e facilite a consulta. Uma possibilidade é a utilização do Microsoft Analysis Services que consegue realizar deixar os cálculos salvos em um modelo que poderá ser plugado no power bi, aumentando a velocidade de consulta.
* Nosso banco de dados está hospedado na nuvem e nossas ferramentas de análise de dados são "on premisses". Você manteria este arranjo ou faria mudanças visando mais performance?
  + Manteria o formato existente. Para os casos que já estejam estáveis, publicaria na nuvem para que os dashboards/ferramentas consigam acessar.
* Nossa área operacional necessita de informações em tempo real, porém os diretores da empresa, que acompanham somente informações de KPIs mensais, alegam que isso é desnecessário e acarretaria custos. Qual é o seu posicionamento sobre isso?
  + *Não creio que acarretaria custos elevados e desnecessários. A informação em tempo real de vendas e de toda a parte comercial seria importantíssimo o tempo real para que o direcionamento seja rápido e preciso.*
* Nosso time que está focado em Governança de Dados alega que documentar os processos é mais importante do que refatorar os mais de 500 scripts que estão funcionando com lentidão. Como você atuaria neste impasse, se tivesse que priorizar o trabalho?
  + *Com a equipe, faria um trabalho em conjunto que no mesmo momento que alguém da equipe esteja refatorando os scripts já teria alguém documentando todos os processos em tempo real e ainda ajudando com possíveis melhoras.*
* Aqui no olist, somos muito mão na massa! Como Engenheiro(a) de dados, mostre pra gente o que você consegue fazer na prática com esse nosso banco de dados. (Sabemos que é uma amostra, mas imagine que o todo pode ser petabytes de dados)
  + *Mão na massa + dados = sucesso!*
* O que acha de escrever um relatório ou slides sobre a sua abordagem na solução de alguns desses problemas?
* Fique livre para criar sua própria abordagem, caso considere que as dicas anteriores não sejam pertinentes.**Engenheiro(a) de dados**