

Case Técnico Data Architect - iFood

1

Objetivo

2

Esse desafio permitirá que você demonstre suas habilidades em Engenharia de Dados/Software, Análise e Modelagem de Dados.

3

Neste case técnico, você deverá fazer a ingestão de alguns dados em nosso Data Lake e pensar em uma forma de disponibilizá-los para os consumidores. Para finalizar, você deverá realizar análises sobre os dados disponibilizados.

4

Você deverá:

5

- Desenvolver uma solução para fazer a ingestão de dados referentes às corridas de táxis de NY em nosso Data Lake;
- Disponibilizar os dados para os usuários consumirem (através de SQL, por exemplo);
- Realizar algumas análises dos dados e mostrar os resultados;

6

7

8

Dados Disponíveis

9

Os dados estão disponíveis no site da agência responsável por licenciar e regular os táxis na cidade de NY: <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

10

Num primeiro momento, precisamos que sejam armazenados e disponibilizados os dados de Janeiro a Maio de 2023.

11



12

Considerações

13

- Você pode considerar inicialmente armazenar todos os arquivos originais em uma landing zone (que pode ser um bucket do S3, por exemplo, ou qualquer outra tecnologia de sua escolha);

14

- Você pode considerar armazenar os dados estruturados/ transformados em uma camada de consumo (que pode ser um bucket do S3, por exemplo, ou qualquer outra tecnologia de sua escolha);	15
- Você pode considerar manipular/limpar os dados que julgar necessário;	16
- Você precisa garantir que as colunas **VendorID** , **passenger_count** , **total_amount** , **tpep_pickup_datetime** e **tpep_dropoff_datetime** estejam presentes na camada de consumo. As outras colunas podem ser ignoradas;	17
- Você pode considerar que no Data Lake não existe nenhuma tabela criada, portanto, precisam ser modeladas e criadas.	18
## O Desafio	19
Você deverá entregar:	20
1. **Solução para ler os dados originais, fazer a ingestão no Data Lake e disponibilizar para os usuários finais:**	21
2. Deve utilizar PySpark em alguma etapa;	22
3. Recomendamos usar Databricks Community Edition (https://community.cloud.databricks.com/);	23
4. A escolha da tecnologia de metadados fica a seu critério;	24
5. A escolha da linguagem de consulta (SQL, PySpark e etc) para os usuários finais fica a seu critério;	25
2. **Código SQL ou PySpark estruturado da forma que preferir com as respostas para as seguintes perguntas:**	26
2. Qual a média de valor total (total_amount) recebido em um mês considerando todos os yellow táxis da frota?	27
3. Qual a média de passageiros (passenger_count) por cada hora do dia que pegaram táxi no mês de maio considerando todos os táxis da frota?	28

Estrutura do Repositório 29

ifood-case/	</>	30
└─ src/	# Código fonte da solução	31
└─ analysis/	# Scripts/Notebooks com as repostas das perguntas	32
└─ README.md		33
└─ requirements.txt		34

Critérios de Avaliação 35

Serão avaliados:	36
- Qualidade e organização do código	37
- Processo de análise exploratória	38
- Justificativa das escolhas técnicas	39
- Criatividade na solução proposta	40
- Clareza na comunicação dos resultados	41

Instruções de Entrega 42

1. Crie um repositório público ou privado no GitHub	43
2. Desenvolva sua solução	44
3. Atualize o README com instruções de execução	45
4. Envie o link do seu repositório	46

Dúvidas 47

Em caso de dúvidas sobre o desafio, abra uma issue neste repositório.	48
-----------------------------------------------------------------------	----

Boa sorte!	49
------------	----