Processo Seletivo
Engenheiro de Dados

Desafio Técnico

Case #01 - Modelagem de dados

A equipe de produtos da Geofusion tem recebido feedbacks da equipe comercial e eles dizem que alguns clientes do setor de alimentação (restaurantes, pizzarias, bares, etc.) estão procurando soluções que os ajudem a entender melhor seus concorrentes. Eles gostariam de saber qual é a faixa de preço praticada pelos concorrentes, como é o fluxo de pessoas nesses locais, qual é a população e a densidade demográfica dos bairros onde os concorrentes estão, e etc.

A Geofusion decidiu investir no desenvolvimento de uma solução para resolver os problemas desses clientes. A intenção da empresa é inserir esta solução nos produtos já existentes.

Atualmente um dos nossos desafios é criar serviços de dados escaláveis, otimizados para o armazenamento e para o acesso aos dados. Diante disso, seu primeiro desafio será definir um modelo de dados que otimize o armazenamento e o acesso a estes dados.

Como resultado, você deve enviar uma descrição da arquitetura de dados proposta, incluindo o(s) modelo(s) de dados e as tecnologias sugeridas. Você precisará enviar um documento no formato pdf.

Os parágrafos a seguir irão te contextualizar quanto as regras de negócio e os arquivos que devem ser utilizados durante o desenvolvimento.

Regras de negócio:

Nossos Analistas analisaram os dados de fluxo de pessoas e concluíram que a melhor forma de apresentar essa informação seria segmentando o fluxo por dias da semana e períodos do dia (manhã, tarde e noite), ou seja, os clientes precisam saber quantas pessoas em média frequentam seus concorrentes em cada dia da semana e em cada período do dia. Para

encontrar fluxo médio de pessoas é preciso considerar os eventos dos mesmos dias da semana e dos mesmos períodos do dia.

A densidade demográfica de um bairro é uma informação muito importante para nossos clientes e é uma informação que precisa ser calculada. A densidade demográfica de um bairro é o resultado da divisão da população do bairro pela área do bairro.

Arquivos:

Os arquivos que você deve utilizar durante o desenvolvimento são de diferentes fontes, da Geofusion, do IBGE e de um parceiro e foram enviados em anexo. Segue a descrição de cada um:

- **eventos_de_fluxo.csv**: contém os dados do fluxo de pessoas. São eventos registrados a partir dos celulares de pessoas que permanecem mais de 5 minutos em um estabelecimento comercial. Esses dados são enviados diariamente para a Geofusion. Este arquivo possui uma pequena amostra dos dados (248.590 de eventos), entretanto, o volume real passa dos 15 milhões de eventos. O arquivo possui 3 atributos:

CODIGO	Código do evento
DATETIME	Data e hora do evento
CODIGO_CONCORRENTE	Código do concorrente

- **populacao.json**: contém a quantidade de habitantes por bairro. Esse arquivo contém 2 atributos:

CODIGO	Código do bairro
POPULACAO	Quantidade de habitantes

- **bairros.csv**: contém as informações dos bairros. Esse arquivo contém 5 atributos:

CODIGO	Código do bairro
NOME	Nome do bairro
MUNICIPIO	Cidade
UF	Estado
AREA	Área do bairro em km²

- **concorrentes.csv**: contém os dados de concorrentes. Esse arquivo contém 8 atributos:

CODIGO	Código do concorrente
NOME	Nome do concorrente
CATEGORIA	Atividade econômica
FAIXA_PRECO	Faixa de preço praticada
ENDERECO	Endereço
MUNICIPIO	Cidade
UF	Estado
CODIGO_BAIRRO	Bairro

Case #02 - ETL

Agora que definimos o(s) modelo(s) de dados, podemos seguir para a etapa de implementação.
Neste case seu desafio será implementar o que foi definido no case anterior para o armazenamento dos dados. O foco maior desse case será na etapa de transformação dos dados.
Como resultado da implementação, você deve enviar todos os scripts desenvolvidos e os arquivos csv contendo os dados finais de cada uma das tabelas criadas. É importante que você envie tudo que foi utilizado durante esta etapa, pois estes scripts serão inseridos e testados em uma base de dados. Portanto, caso esteja faltando algum script não será possível efetuar a validação. Se atente para não esquecer de enviar os scripts responsáveis pela criação das tabelas.
Use e abuse da sua criatividade. Lembre-se que aqui você participará de todo o desenvolvimento do processo ETL, desde a definição do modelo até a disponibilização dos dados.

Case #03 - Serviço para disponibilizar os dados

Agora que temos todos os dados armazenados, podemos seguir com a etapa de disponibilização.
Neste case seu desafio será implementar um serviço REST que retorne informações dos concorrentes como código, nome, endereço, preço praticado, fluxo médio de pessoas por dia da semana e por período do dia, bairro e a população e a densidade demográfica do bairro.

Case #04 - Análise dos dados

Uma coisa muito importante quando trabalhamos com uma grande quantidade de dados, é quanto valor conseguimos extrair desses dados. Neste case, seu desafio é fazer uma análise dos dados já trabalhados juntamente com alguns dados que foram levantados por um cliente. A ideia é analisar os dados e extrair insights dessa análise para entender tendências ou padrões nesses dados. A estrutura dos dados entregue pelo cliente é a seguinte:

- **potencial.csv**: contém dados levantados por um cliente com informações relevantes para o negócio dele para cada bairro. Esse arquivo contém 5 atributos:

CODIGO	Código do bairro
QTD_AGENCIAS	Quantidade de agências bancárias no bairro
EMPRESAS	Quantidade de empresas no bairro
EMPREGADOS	Quantidade estimada de trabalhadores no bairro
RENDA	Renda predominante do bairro
FATURAMENTO	Faturamento das lojas do cliente no bairro

Faça a análise desses dados da forma que achar melhor, utilizando ferramentas e técnicas que achar mais pertinente. Não existe certo ou errado, mas utilize bons argumentos para justificar sua análise e explicar as suas conclusões.

Este case é opcional, isso significa que você não será penalizado por não entregá-lo, mas a entrega dele com uma boa análise contará muitos pontos positivos no seu desafio.

Regras gerais

1. Você precisa utilizar todos os arquivos listados no case #01.
2. O desenvolvimento de todo o desafio utilizando Docker facilita bastante para você e para quem vai avaliar seu trabalho. Portanto, o desafio **deve** ser entregue em um ambiente Docker.
3. Não esqueça de comentar seu código, pois isto ajuda muito no processo de manutenção. Lembre-se que outra pessoa irá avaliar seu código. Quanto mais bem documentado, mais fácil e menos passível de ambiguidades será a análise.
4. Você **deve** disponibilizar o projeto em um repositório de códigos como GitHub ou BitBucket, por exemplo.

Observações

Para realizar esse desafio, você pode utilizar a linguagem de programação e as tecnologias que mais lhe agradam ou as que você julgar mais adequadas.
Embora não seja obrigatória a utilização de ferramentas de Big Data (Spark, MapReduce, etc), isto será visto como um diferencial. Portanto, fique a vontade para utilizá-las caso se sinta confortável.