Desafio Final de Engenharia de Dados: Acelere Sua Carreira com a triggo.ai!

Parabéns por chegarem até aqui! Dois meses de imersão intensa no mundo da Engenharia de Dados se passaram, e agora é a hora de consolidar todo o conhecimento adquirido em Databricks, Snowflake e dbt. A **triggo.ai** está incrivelmente entusiasmada para testemunhar o potencial de cada um de vocês neste desafio final – um marco que, sem dúvida, ficará gravado na memória como um passo crucial no início da sua carreira profissional.

Este não é apenas um teste, é uma oportunidade de ouro! Os melhores projetos terão a chance de fazer parte do nosso time de Engenheiros de Dados. Além disso, a triggo.ai tem uma vasta rede de parceiros de tecnologia e clientes. Isso significa que os trabalhos mais excepcionais serão compartilhados, abrindo portas e ampliando exponencialmente o reconhecimento e as oportunidades de vocês no mercado. Preparem-se para brilhar!

O Desafio: Construindo uma Solução de Dados para a "Health Insights Brasil"

O Cenário

Imagine que vocês são os novos Engenheiros de Dados da **Health Insights Brasil**, uma startup focada em otimizar a gestão e a análise de dados de saúde pública no país. O Brasil gera uma quantidade massiva de informações através do Sistema Único de Saúde (SUS), disponibilizadas publicamente pelo **DataSUS**. No entanto, a vastidão e a complexidade desses dados tornam um desafio para hospitais, pesquisadores e formuladores de políticas públicas extraírem insights relevantes e em tempo hábil.

A Health Insights Brasil busca criar uma plataforma que torne os dados do DataSUS mais acessíveis, compreensíveis e acionáveis, auxiliando na tomada de decisões estratégicas para a saúde pública. Atualmente, eles lidam com dados brutos em diversos formatos, e a equipe de análise precisa de uma fonte de dados confiável, organizada e performática para construir dashboards e relatórios que apoiem a saúde da população. É aqui que vocês entram!

O Objetivo

O desafio é projetar e implementar uma pipeline de dados completa que ingere, transforma e modela esses dados brutos do DataSUS, tornando-os prontos para análise. Vocês devem entregar um projeto que simule uma solução de engenharia de dados real, demonstrando proficiência nas ferramentas aprendidas.

Estrutura do Projeto e Requisitos Técnicos

Vocês terão **10 dias** para desenvolver e entregar este projeto. Lembrem-se: a qualidade e a robustez da solução são primordiais.

1. Coleta e Ingestão de Dados (15 pontos)

- **Dados:** Vocês deverão utilizar dados de acesso público do **DataSUS**. Recomenda-se focar em um ou dois conjuntos de dados específicos que permitam uma análise coerente (ex: Internações Hospitalares - SIH, Ambulatoriais - SIA, Nascidos Vivos - SINASC, ou Óbitos - SIM). O importante é escolher uma base de dados realista e relevante para a saúde pública.
- **Requisito:** Demonstre a ingestão desses dados para a plataforma escolhida (Databricks ou Snowflake). Pensem em como simular o download e o armazenamento inicial desses dados em um "data lake" (pode ser simulado com arquivos no armazenamento de objetos da plataforma ou localmente).

2. Transformação e Modelagem de Dados com dbt (40 pontos)

- **Modelagem Dimensional:** Crie um modelo de dados dimensional (Star Schema ou Snowflake Schema) adequado para análise em saúde pública. Pensem em tabelas de fatos e dimensões que permitam análises sobre padrões de doenças, internações, procedimentos, demografia da população atendida, etc.
- Exemplo de Dimensões (baseado no DataSUS): dim_tempo, dim_localidade (município, estado), dim_doenca (CID-10), dim_procedimento, dim_faixa_etaria_sexo.
- Exemplo de Fato: fato_atendimento_hospitalar (se usar SIH), fato_nascimento (se usar SINASC).
- **dbt:** Utilize o dbt para construir e gerenciar todas as transformações de dados.
- Defina modelos (models) para cada estágio da pipeline (staging, intermediate, mart).
- Implemente testes (tests) para garantir a qualidade dos dados (ex: unicidade de IDs, não-nulidade de campos críticos, validação de códigos do DataSUS).
- Documente seus modelos (docs) no dbt.
- Utilize o conceito de materializações (materializations) do dbt de forma apropriada (e.g., view, table, incremental).

3. Escolha da Plataforma (25 pontos)

- **Opção 1 (Databricks + dbt):** Utilize o Databricks para o processamento e armazenamento dos dados. Demonstre o uso de Delta Lake para otimização do armazenamento e performance, especialmente importante para grandes volumes de dados como os do DataSUS.
- **Opção 2 (Snowflake + dbt):** Utilize o Snowflake como seu data warehouse. Explore recursos como Time Travel e Zero-Copy Cloning se possível, para demonstrar familiaridade com a plataforma e otimizar o gerenciamento de dados complexos de saúde.
- **Bônus (10 pontos):** Se você conseguir integrar **Databricks e Snowflake (simulando uma arquitetura híbrida ou de interoperabilidade)**, usando um para ingestão/transformação e outro para consumo final, e utilizando dbt para orquestrar as transformações em ambos, você ganhará pontos extras significativos. Essa é uma excelente oportunidade para demonstrar domínio completo!

4. Orquestração e Automação (10 pontos)

- Descreva como a pipeline de dados poderia ser orquestrada (ex: Airflow, Databricks Workflows, Snowflake Tasks). Não é necessário implementar a orquestração completa, mas um plano bem detalhado e justificado será avaliado. Se possível, demonstre um fluxo simples.

5. Inovação e Diferenciação (10 pontos)

Este é o seu espaço para brilhar e mostrar sua criatividade!

- Pense em como você pode ir além dos requisitos básicos.
- Ideias para inovação (não obrigatórias, apenas sugestões):
- Implementar um mecanismo básico para identificar tendências ou alertas epidemiológicos a partir dos dados modelados.
- Criar um pequeno dashboard de visualização (ex: Power BI, Tableau, Streamlit) consumindo os dados do seu modelo dbt, mostrando um insight relevante sobre a saúde pública.
- Propor uma solução para a integração de novas fontes de dados do SUS ou de outras bases de saúde (ex: dados de atendimentos de emergência em tempo quase real).
- Abordar desafios de privacidade e anonimização de dados (sem a necessidade de implementá-los, mas com uma discussão sobre o tema).
- Adicionar um **GitOps** para o seu projeto dbt, demonstrando um fluxo de trabalho de CI/CD simples.

Critérios de Avaliação e Pontuação

- **Coleta e Ingestão de Dados (15 pontos):**
- Clareza e eficiência do processo de ingestão dos dados do DataSUS.
- Organização dos dados brutos na plataforma escolhida.
- **Transformação e Modelagem com dbt (40 pontos):**
- Qualidade do modelo dimensional (20 pontos).
- Uso adequado do dbt (models, tests, docs, materializations) (15 pontos).
- Qualidade e cobertura dos testes de dados (5 pontos).
- **Uso da Plataforma (25 pontos):**
- Proficiência na plataforma escolhida (Databricks ou Snowflake) (20 pontos).
- Exploração de recursos avançados da plataforma (5 pontos).
- **Uso das Três Tecnologias (Databricks + Snowflake + dbt) (Bônus de 10 pontos):**
- Demonstração clara da integração e interoperabilidade entre as plataformas, orquestrada pelo dbt.
- **Orquestração e Automação (10 pontos):**
- Proposta clara e justificada de orquestração.
- (Opcional) Implementação de um fluxo de trabalho simples.
- **Inovação e Diferenciação (10 pontos):**
- Originalidade e impacto das funcionalidades adicionais para o contexto da saúde.
- Qualidade da implementação das ideias inovadoras.

O Pitch Final: Seu Momento de Brilhar (5 minutos)

Após a entrega do projeto, vocês terão a oportunidade de apresentar sua solução em um pitch de 5 minutos. Esta é a chance de vender sua ideia, explicar suas escolhas técnicas e demonstrar o valor do seu trabalho.

No pitch, abordem:

- **Problema e Solução:** Qual o problema que vocês resolveram para a Health Insights Brasil e como a sua solução o endereça, utilizando dados do DataSUS.
- **Design da Arquitetura:** Uma visão geral da arquitetura de dados que vocês implementaram.
- **Tecnologias Utilizadas:** Como Databricks/Snowflake e dbt foram cruciais para a solução.
- **Desafios e Aprendizados:** Quais foram os principais obstáculos encontrados ao lidar com dados do DataSUS e como vocês os superaram.
- **Inovação:** Destaque suas ideias inovadoras e como elas agregam valor para a saúde pública.
- **Próximos Passos:** Sugestões para o futuro da solução.

Entrega do Projeto

Vocês devem entregar o projeto em um repositório Git (GitHub, GitLab ou Bitbucket) contendo:

- **Código dbt:** Todos os modelos, testes e documentação.
- **Scripts de Ingestão:** Código ou scripts utilizados para ingestão dos dados brutos do DataSUS.
- **Documentação:** Um arquivo README.md detalhado que explique:
- Como rodar o projeto.
- A arquitetura da solução.
- Decisões de design e justificativas.
- Resultados e insights obtidos (ex: um exemplo de consulta que gera um insight relevante sobre saúde pública).
- Qualquer inovação implementada.
- Um link para o dbt docs gerado (se for o caso).

O prazo final para a entrega do repositório é de **10 dias** a partir da divulgação deste desafio. A triggo.ai está ansiosa para ver as soluções incríveis que vocês irão criar! Este é um passo gigantesco em suas jornadas profissionais. Boa sorte e mãos à obra!