Avaliação prática - Engenharia de dados - Geolinkage

No Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs), estamos à procura de um Engenheiro de Dados Especialista para se juntar a uma equipe multidisciplinar em um projeto internacional e estratégico, visando impulsionar a ciência brasileira. Este profissional será encarregado de construir uma base de dados de endereços a partir de diversas fontes, a fim de dar suporte aos processos de geocodificação de eventos relacionados à saúde. O objetivo é criar o ambiente computacional necessário, implementar estratégias, processos e produtos baseados em tecnologia para atender às demandas e resolver os desafios do negócio, utilizando métodos científicos rigorosos e garantindo reprodutibilidade.

O objetivo deste teste é aferir seu domínio ou proficiência nas tecnologias e ferramentas de nossa Plataforma de Dados. Outros aspectos avaliados na entrega desta tarefa são:

- Desenho e condução de um processo de pipeline ETL;
- Manejo de bases de dados de alta dimensionalidade usando PySpark, Elasticsearch, etc;
- Capacidade de montar, automatizar e otimizar uma infraestrutura necessária para execução destas tarefas;
- Familiaridade com distribuições Linux; e
- Capacidade de comunicar os resultados.

Levaremos em conta todo o seu percurso para resolução desta tarefa, considerando o nível da vaga que está competindo e a riqueza de detalhes incluídas em seu relatório. Descreveremos a seguir o desafio.

Implemente um *pipeline* capaz de:

- Fazer download de uma base de dados de endereço para fins estatísticos (Censo de 2010), publicamente disponível em: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html. A população de interesse é a de Salvador (código IBGE 2927408), na Bahia.

- **BASE A**: Recortar cem mil registros de endereços aleatoriamente desta base baixada, selecionando as variáveis: TIPO, TITULO, LOGRADOURO, NUMERO, COMPLEMENTO e SETOR CENSITÁRIO.
- Faça uma descritiva simples da base de dados (formulação livre).
- **BASE B**: Através de uma rotina de amostragem, crie uma segunda base de dados que contenha ruídos controlados. Considere gerar os seguintes ruídos:
    - Mil registros com valores ausentes em uma das variáveis;
    - Dois mil registros com supressão de uma palavra (aleatória) nas variáveis LOGRADOURO ou COMPLEMENTO.
    - Três mil registros com supressão de duas palavras aleatórias nas variáveis LOGRADOURO ou COMPLEMENTO.
- Utilize a medida de distância Jaro-Winkler [disponível em https://pypi.org/project/jellyfish/] para criar uma matriz de distância entre todos os registros da BASE A e resultantes da rotina que gerou a BASE B. Considere apenas as variáveis “LOGRADOURO” e “COMPLEMENTO”.
- Faça uma descritiva da matriz de distância resultante (formulação livre).

Responda o email de agendamento do teste prático com o relatório em anexo para assegurar sua entrega.

Desejamos boa sorte.