**Instruções para o Teste de Estágio em Engenharia de Dados**

Bem-vindo(a) ao teste prático para a vaga de estágio em engenharia de dados. Este teste foi desenhado como um exercício de ETL para avaliar suas habilidades de pesquisa, aprendizado autônomo e aplicação prática de conhecimentos em um contexto real de trabalho.

**Tarefa:**

O Exercício, portanto, consistirá num case prático de ETL para dados da RAIS[[1]](#footnote-2). Você deverá realizar o seguinte passo a passo:

* Realizar o download (Extração – E) dos dados da RAIS para a Região Norte do Brasil para o Ano de 2021, os dados estão disponíveis no servidor FTP do Ministério do Trabalho e Emprego: ftp://ftp.mtps.gov.br/pdet/microdados/. Entre na pasta RAIS, lá você verá pastas com dados de 1985 a 2021, seleciona a pasta 2021 e lá baixe apenas o arquivo da Região Norte.
* Além de realizar o download dos dados, você precisará baixar o arquivo com Layout do arquivo de dados no seguinte link: <ftp://ftp.mtps.gov.br/pdet/microdados/RAIS/Layouts/v%C3%ADnculos/RAIS_vinculos_layout2020.xls>. Este arquivo contém os nomes das variáveis que existem no arquivo, os seus tipos e delimitações. Isso será importante para as etapas posteriores.
* Feita a Extração, o próximo passo será o Tratamento (T), a base inicial possui 3.977.463 linhas, serão dois tratamentos:
  + Primeiro, filtrar e eliminar da base todos os registros em que a variável “EMP EM 31/12”apresente valor igual a 0 (zero). Ou seja, essa variável indica aquelas pessoas que estavam empregadas ou não no dia 31/12 d oano em questão, ao eliminar os casos iguais a 0, estamos retirando da base quem não estava empregado em 31/12.
  + A partir da variável “CLAS CNAE 20”,criar uma nova variável chamada “DIV CNAE 20” que corresponde aos dois primeiros dígitos da variável “CLAS CNAE 20”. Isso criará uma nova variável para indicar o código de uma atividade econômica.[[2]](#footnote-3)

Os dados devem ser carregados (Load – L) em um banco SQL (Mysql, Postgre, ou outro de sua preferência).

* Por fim, realize duas análises:
  + Apresente as 5 atividades que mais empregam pessoas na Região Norte do Brasil, para isso utilize a variável que você criou “DIV CNAE 20”.
  + Apresente os 5 municípios que mais empregam pessoas na Região Norte do Brasil.

**Pontos de Atenção:**

* Armazenamento: Otimize o armazenamento dos dados no banco para ocupar o mínimo espaço possível.
* Processamento: Implemente medidas para tornar as consultas SQL no banco mais eficientes.

**Diretrizes:**

* Você deve pesquisar e decidir quais ferramentas e métodos utilizar para cumprir as tarefas acima.
* Após a conclusão, você deverá apresentar uma justificativa para cada escolha técnica que fez durante o processo.
* Encorajamos o uso de qualquer fonte de pesquisa que considere relevante (incluindo Google, ChatGPT, fóruns, documentações oficiais, etc).

**Critérios de Avaliação:**

* Correto armazenamento e processamento dos dados.
* Eficiência e otimização aplicadas ao banco de dados.
* Segurança e adequação no compartilhamento de acesso.
* Qualidade e coerência das justificativas apresentadas.

**Entrega:**

Compartilhe todo o código utilizado neste teste até o dia 14/02/2024. Pode compartilhar via Notebook Jupyter, Github, Google Colab, ou qualquer um de sua preferência.

Você deverá documentar o processo e as escolhas feitas e enviar um relatório juntamente com os detalhes de acesso ao banco de dados configurado.

Sucesso!

1. <https://ces.ibge.gov.br/base-de-dados/metadados/mte/relacao-anual-de-informacoes-sociais-rais.html>. [↑](#footnote-ref-2)
2. Para visualizar atividades econômicas acesse: <https://concla.ibge.gov.br/busca-online-cnae.html?view=estrutura>. [↑](#footnote-ref-3)