Bootcamp: Arquiteto(a) de Big Data

Desafio Final

Módulo 5º: Desafio Final do Bootcamp

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Realizar coleta de dados em arquivos.
2. Manipulação e visualização de dados.
3. Criar modelo entidade e relacionamento para armazenamento de dados.
4. Realizar carga de dados no banco de dados MySQL.
5. Tratamento de dados.
6. Realizar consultas na linguagem SQL.
7. Conhecimento teórico ministrado nas videoaulas

Enunciado

Esta pesquisa fictícia foi desenvolvida com o propósito de treinar e aprimorar todo o processo de Automação de Processos Robóticos (RPA) desde a coleta de dados, passando pelo tratamento, integração e armazenamento de informações. A simulação envolve a coleta de dados sobre preferências pessoais, hábitos e características demográficas dos participantes, fornecendo uma base prática para o desenvolvimento de habilidades em RPA.

Atividades de Coleta e Armazenamento

A principal atividade deste projeto é a coleta de dados por meio de datasets estruturados e a subsequente armazenagem dessas informações em um banco de dados. O processo é dividido em várias etapas essenciais para garantir a qualidade e a integridade dos dados, além de proporcionar uma experiência completa de automação:

1. Desenvolvimento do Questionário

Elaboramos um questionário detalhado para capturar informações sobre preferências pessoais (como bebida favorita, hobbies), características demográficas (como gênero e data de nascimento) e outros aspectos relevantes (como presença de animais de estimação e percepções sobre o clima). – Esse item já foi desenvolvido. Vamos focar na coleta dos dados.

2. Coleta de Dados

A coleta de dados é realizada por meio de entrevistas ou preenchimento de formulários on-line. Os participantes fornecem suas respostas, que são então capturadas automaticamente por boots de RPA.

3. Tratamento de Dados

Verificação de Dados Duplicados: utilizando scripts automatizados, removemos registros duplicados para evitar redundâncias e garantir a precisão dos dados.

Tratamento de Dados Ausentes: identificamos e lidamos com valores ausentes, utilizando métodos automáticos para preenchimento com valores padrão ou imputação baseada em algoritmos.

Validação e Limpeza de Dados: boots de RPA validam automaticamente que todos os dados estão no formato adequado e dentro dos intervalos válidos, corrigindo inconsistências ou erros de entrada.

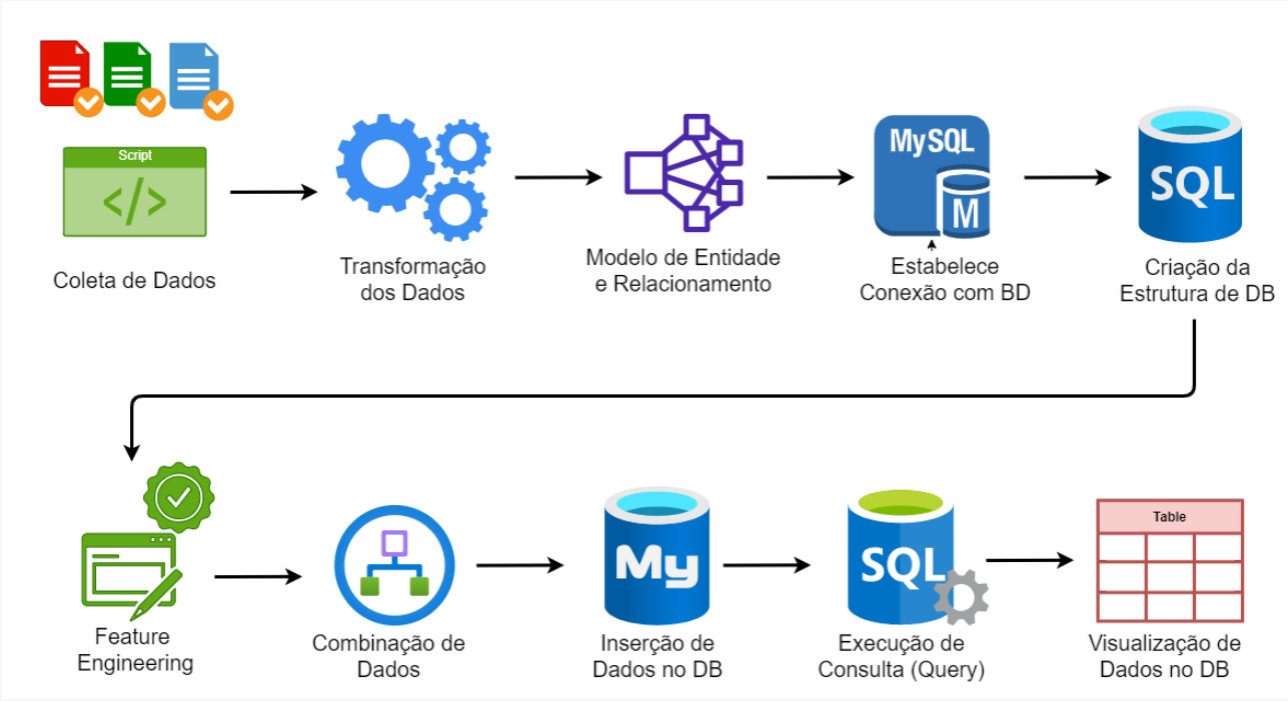
4. Integração de Dados

Os dados coletados e tratados são integrados de maneira automática com sistemas existentes, garantindo que todas as informações estejam disponíveis para análise e uso imediato.

5. Armazenamento de Dados

Após o tratamento e a integração, os dados são armazenados em um banco de dados estruturado. Este processo inclui a criação automatizada de tabelas e esquemas que organizam os dados de forma eficiente e acessível, além de implementar medidas de integridade referencial para assegurar a consistência dos dados entre diferentes tabelas.

Fluxograma das etapas a serem desenvolvidas neste trabalho.



Atenção! Para garantir a obtenção dos mesmos resultados do projeto, é recomendável o uso das mesmas versões das bibliotecas

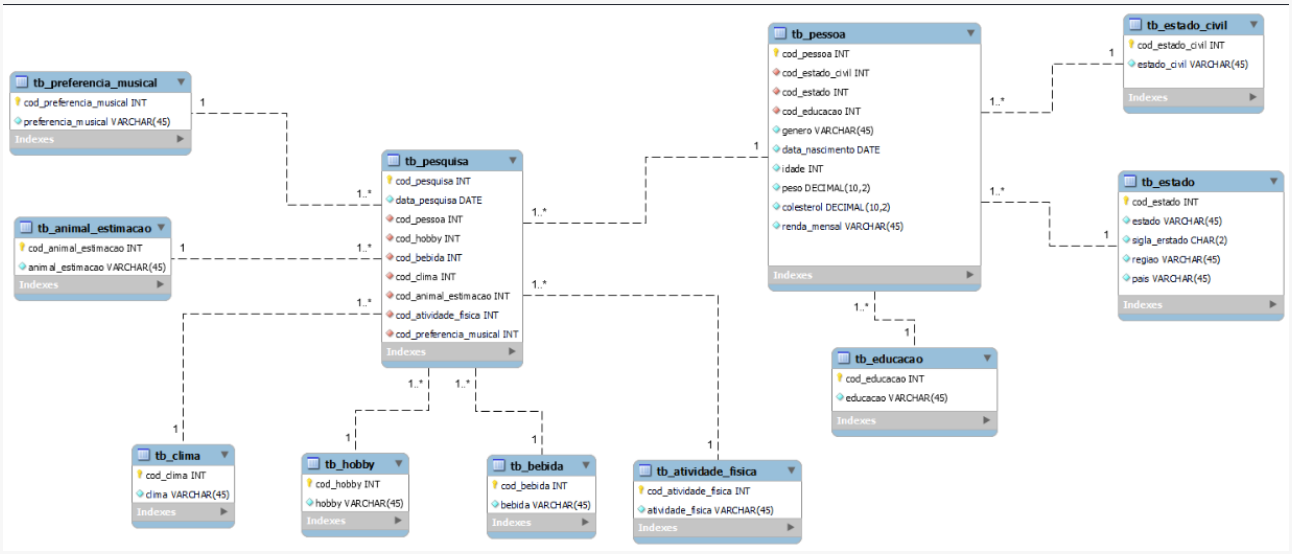
VERSÕES BIBLIOTECAS UTILZIADAS
Pandas: 2.2.2

Sqlalchemy: 2.0.30

É crucial reconhecer que a linguagem de programação Python e suas bibliotecas associadas estão em constante evolução. Como resultado, pode ocorrer que funções ou métodos específicos, que costumavam estar disponíveis em versões anteriores, deixem de existir ou passem a ser implementados de maneira diferente em versões mais recentes.

Essas atualizações são realizadas para melhorar a eficiência, corrigir erros e fornecer novos recursos aos desenvolvedores. No entanto, essa dinâmica de mudança também pode criar desafios, especialmente quando se trabalha com código legado ou ao compartilhar código com outros membros da equipe. Portanto, é de extrema importância que os alunos estejam cientes dessas mudanças e estejam dispostos a se adaptar a elas.

Modelo de entidade e relacionamento que deverá ser criado.



É importante observar que ao inserir dados em tabelas que dependem de informações de outras tabelas para concluir com sucesso a operação de inserção, como o caso da tabela 'pesquisa' que requer que a todas as outras tabelas já estejam populadas, é necessário seguir uma ordem estratégica de inserção.

Além disso, utilize a tabela de 'stage' para fazer um processo parecido com o PROCV do Excel para inserir os dados.

ATENÇÃO PARA TRATAMENTO DE DADOS

Avalie se será necessário realizar tratamento de dados ausentes nos datasets disponibilizados.

Instruções para correção de dados ausentes

1. Média arredondada para 2 casas decimais para as variáveis do tipo numéricas;
2. Moda para as variáveis categóricas.

Atividades

Para esta atividade, os alunos deverão realizar as seguintes tarefas:

1. Coletar os dados fornecidos através da lista de arquivos;
2. Criar estrutura de tabelas no banco de dados MySQL;
3. Inserir dados coletados na estrutura criada;
4. Realizar comandos SQL para extrair informações da base de dados.

Dicas do professor:

1. Antes de enviar as respostas, verifique se o gabarito está correto.
2. Analise se existem dados duplicados e elimine-os se necessário.
3. Siga fielmente todos os passos contidos no enunciado das questões.
4. É fundamental observar a configuração de autoincremento ao criar tabelas que requerem a geração automática de códigos para representar os dados.
5. Os dados disponibilizados no dataset são fictícios, ou seja, não têm relação com o mundo real.
6. Utilize o artigo abaixo para criar a idade.
    a. https://leandrolessa.com.br/tutoriais/pandas-3-passos-para-converter-data-de-nascimento-em-idade/
    b. Para o cálculo da idade, utilize a data de referência:
        i. data_ref = pd.to_datetime('2024-06-20')
7. Siga os procedimentos realizados nas videoaulas. O sucesso do experimento depende de seguir a mesma estratégia.
8. Os datasets utilizado no trabalho pode ser obtido no link:
    a. Análise de Preferências e Comportamentos Demográficos
    b. Dados de Estados
        i. https://leandrolessa.com.br/datasets/

9. Caso queiram automatizar a coleta e extração acesse o link:

    a. https://leandrolessa.com.br/tutoriais/automatizando-coleta-e-extracao-de-arquivos-zip-na-web-com-python/
    b. https://leandrolessa.com.br/tutoriais/coleta-de-dados-em-3-passos-com-python-simples-e-direto-para-listas-de-arquivos/