Bootcamp: Arquiteto de Big Data

Módulo 5º Desafio Final

Projeto de Coleta e Armazenamento de Dados - Arquitetura de Big Data

Como arquiteto de big data recém-contratado, você recebeu a tarefa de criar um banco de dados eficiente utilizando o MySQL Server para armazenar e gerenciar os dados provenientes de uma pesquisa realizada por um instituto renomado. Essa pesquisa coletou informações sobre as preferências das pessoas em relação a diferentes assuntos, incluindo escolhas de animais de estimação, preferências climáticas, bebidas favoritas e hobbies.

Descrição da Atividade

Sua missão é realizar o design e a implementação do banco de dados, bem como desenvolver os procedimentos necessários para coletar e armazenar esses dados de Licenças Médicas. Siga as etapas abaixo para completar a atividade.

Análise de Licenças Médicas

Examine detalhadamente os conjuntos de dados das licenças Médicas para compreender as informações coletadas e identificar os tipos de dados associados as entidades envolvidas no conjunto de dados.

Projeto do Banco de Dados Relacional

Desenvolva um esquema de banco de dados relacional que represente de maneira eficiente as relações entre as categorias e os atributos da pesquisa. Considere a normalização do banco de dados para evitar redundâncias.

Configuração do MySQL Server

Instale e configure o MySQL Server, ajustando as configurações conforme necessário para atender aos requisitos específicos do projeto.

Criação de Tabelas e Relacionamentos

Implemente as tabelas necessárias no MySQL Server de acordo com o esquema desenvolvido, estabelecendo os relacionamentos apropriados entre elas.

Procedimentos de Coleta de Dados

Elabore procedimentos ou scripts para coletar dados das Licenças Médicas e inseri-los no banco de dados MySQL Server. Certifique-se de considerar a integridade e a consistência dos dados durante o processo de inserção.

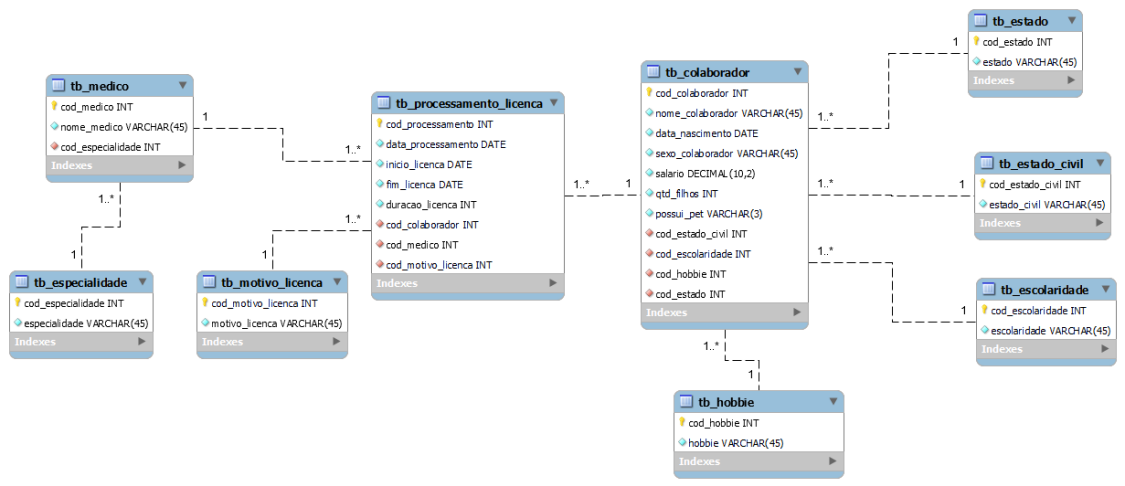
Objetivos geral do desafio

Exercitar os seguintes conceitos trabalhados no curso:

1. Coleta de dados estruturados.
2. Coleta de dados na Web.
3. Criação de estrutura de armazenamento em banco de dados.
4. Tratamento, limpeza e processamento de dados.
5. Análise de dados.
6. Visualização de dados.
7. Práticas de manipulação de dados.
8. Exercitar comandos Python e SQL.

Dicas e Instruções do professor

1. Utilizem o diagrama de entidade e relacionamento a seguir para criar a estrutura de dados no MySQL.



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 'tb_medico' que requer que a tabela 'tb_especialidade' já esteja populada, 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. Abaixo um exemplo de código.

```sql
insert into tb_medico (nome_medico, cod_especialidade)
(
    SELECT distinct nome_medico, esp.cod_especialidade
    FROM stg_licenca stg
    INNER JOIN tb_especialidade esp on esp.especialidade = stg.especialidade
);
```

Certifiquem-se de que estamos buscando o nome da especialidade tanto na tabela 'stage' quanto na tabela 'tb_especialidade', retornando apenas o

código correspondente. Repitam esse processo para todas as tabelas que se encontrem nessa mesma situação.

2. Cuidado para não esquecerem de selecionar a opção de autoincremento na criação das tabelas do banco relacional.

3. Os dados de Licenças médicas são fictícios, porém baseados em um estudo de caso no mundo real.

4. Os datasets utilizados no trabalho podem ser obtidos no link:

    a. https://leandrolessa.com.br/datasets/

    - Processamento de licenças médicas.

5. Encontre neste link dicas sobre como realizar a coleta e extração automatizada de um arquivo ZIP:

    a. https://leandrolessa.com.br/tutoriais/automatizando-coleta-e-extracao-de-arquivos-zip-na-web-com-python/

6. Antes de enviar as respostas verifique se o gabarito está correto.

7. Analise se existem dados duplicados e elimine-os.

8. Siga fielmente todos os passos contidos no enunciado das questões.

9. Siga os procedimentos realizados nas videoaulas. O sucesso do experimento depende de seguir a mesma estratégia.

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

```
VERSÕES UTILIZADAS DAS BIBLIOTECAS:

Pandas: 1.5.2

Sqlalchemy: 1.4.44
```

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 fornece 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.

ATENÇÃO PARA TRATAMENTO DE DADOS

Avaliem 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:

Exceção para qtd_filhos utilize round().

2. Moda para as variáveis categóricas.

Para as questões das idades utilize o código abaixo apresentando nas videoaulas.

```
query = '''
create temporary table tb_idade
select cod_colaborador , data_nascimento
    ,TIMESTAMPDIFF(YEAR, data_nascimento, NOW()) AS idade
from tb_colaborador;

'''
conn.execute(query)
```

Atenção para as questões que solicitam a média de idades. Os resultados podem ser diferentes dependendo do dia que for realizado o cálculo do indicador. Mas não se preocupem! Esse comportamento já é esperado. Essa diferença pode acontecer devido à idade ser baseada entre a diferença do dia atual e a data de nascimento. Desta forma, ao realizar o cálculo em dias diferentes, as idades dos entrevistados podem ter variações. De qualquer modo, esse detalhe não invalida a questão. Obs.: geralmente as diferenças ocorrem nas casas decimais.