

## Bootcamp: Arquiteto(a) de Big Data

### Trabalho Prático

#### Módulo 2: Coleta e obtenção de dados

#### Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Realizar coleta de dados em arquivos.
- ✓ Manipulação e visualização de dados.
- ✓ Criar modelo entidade e relacionamento para armazenamento de dados.
- ✓ Realizar carga de dados no banco de dados MySQL.
- ✓ Tratamento de dados.
- ✓ Realizar consultas na linguagem SQL.
- ✓ Conhecimento teórico ministrado nas videoaulas.

#### Enunciado

Um instituto de pesquisa realizou no ano de 2020, uma pesquisa que tinha como objetivo coletar dados referente a preferências pessoais de seus entrevistados. Essa pesquisa coletou dados dos seguintes assuntos:

1. Animal de estimação;
2. Bebida;
3. Clima;
4. Hobbies.

A pesquisa foi realizada em dias diferentes durante todo o ano de 2020. Desta forma, cada dia da pesquisa contém informações pessoais de um ou vários entrevistados.

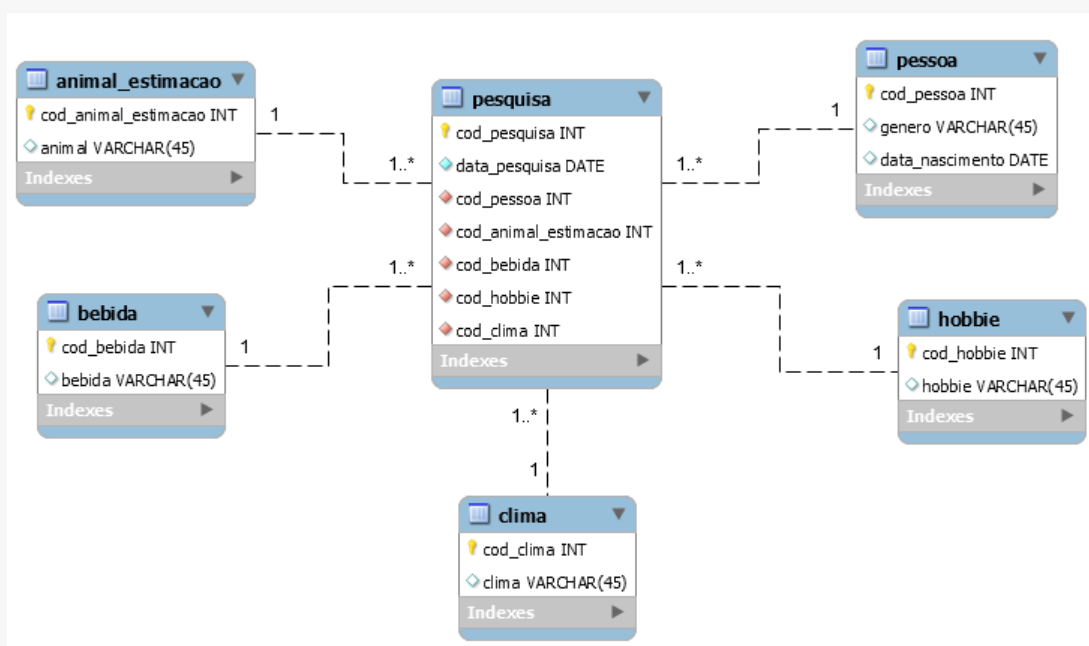
## Atividades

Para essa atividade, os alunos deverão realizar a coleta estruturada nos *datasets* disponibilizados e aplicar conceitos práticos e teóricos ministrados no curso.

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



2. Cuidado para não esquecerem de selecionar a opção de auto incremento na criação das tabelas do banco relacional.
3. 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.

4. Os dados da pesquisa são fictícios. Ou seja, não tem relação com o mundo real.
5. Os datasets utilizados no trabalho podem ser obtidos no link:  
<https://github.com/ProfLeandroLessa/TP-M2-ABD>

Bom trabalho prático à todos!