Bootcamp: Arquiteto(a) de Big Data

Desafio Prático

Módulo 2: Coleta e obtenção de dados

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Realizar coleta de dados em arquivos utilizando linguagem Python;
2. Criar tabelas de acordo com o modelo de entidade e relacionamento fornecido;
3. Realizar carga dos dados coletados no servidor de banco de dados relacional MySQL;
4. Manipular arquivo JSON para carga de dados em banco de dados não relacional MongoDB;
5. Criar e manipular dados no banco não relacional MongoDB;
6. Realizar prática de comandos SQL e NoSQL.

Enunciado

Uma das atividades de um arquiteto de Big Data é saber coletar, tratar e manipular bem os dados. Além disso, é fundamental que o profissional saiba lidar com diferentes tipos de base de dados. É muito comum no dia a dia de um arquiteto de Big Data se deparar com fonte de dados distintas. E esses dados podem vir em diversas fontes:

- Dados de arquivos de texto;
- Dados coletados na Web;

- Dados oriundos de planilhas eletrônicas;
- Dados de questionários;
- Dados não estruturados;
- Dados persistidos em banco de dados.

O objetivo desta atividade é treinar e capacitar o aluno a realizar a coleta, armazenamento e manipulação de dados. Desta forma, os alunos deverão realizar a coleta de dados em dois arquivos:

1. dados_jogadores.csv
2. lista_jogadores_chess.json

O *dataset* de jogadores é um dado estatístico fictício utilizado para o desenvolvimento deste trabalho. Nesse arquivo temos dados de jogadores que realizaram partidas de jogos. Os jogos podem ser: Damas, Xadrez ou jogo da velha. O arquivo possui os seguintes atributos:

1. jogador
    - Informação do jogador
2. Gênero
    - Gênero do jogador
3. data_nascimento
    - Data de nascimento do jogador
4. Jogo
    - Tipo do jogo
5. País

- O país do jogador

6 num_vitorias

- O número de vitórias que o jogador possui

7 num_derrotas

- O número de derrotas que o jogador possui

8 total_partidas

- Total de partidas que o jogador possui

Já os dados que contêm a lista de jogadores do Chess foram extraídos através de uma coleta de dados utilizando API do jogo de xadrez (Chess) ilustrado em uma videoaula deste curso. Desta forma, essa lista contém dados reais de jogadores de Xadrez reconhecidos em todo o mundo. Os dados se encontram em um arquivo JSON e que precisam ser tratados e armazenados. Os atributos dessa lista são: 'avatar', 'player_id', '@id', 'url', 'name', 'username', 'title', 'followers', 'country', 'last_online', 'joined', 'status', 'is_streamer'

Atividade I

Os alunos deverão desempenhar as seguintes atividades:

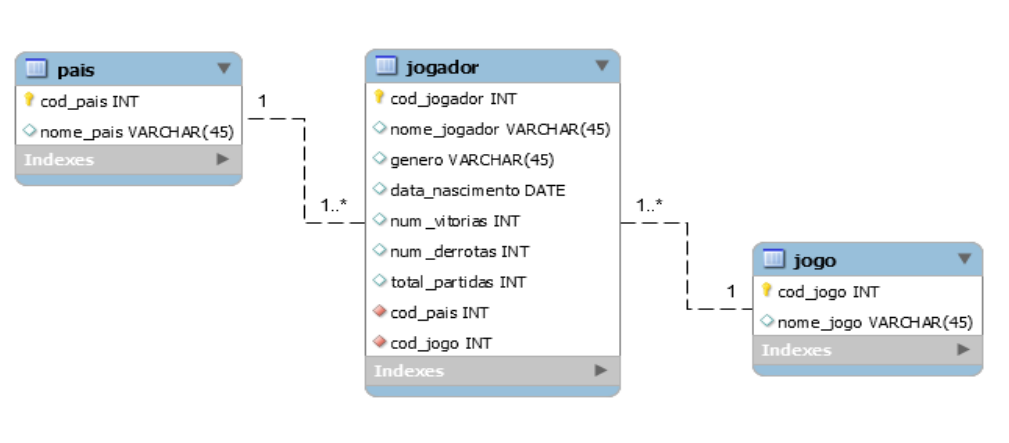
Para responder as questões de 1 a 5, os alunos deverão:

1. Coletar os dados fornecidos (dados_jogadores.csv);
2. Criar um banco de dados relacional MySQL;
3. Criar estrutura de tabelas no banco de dados MySQL;
4. Avaliar dados ausentes das colunas e corrigi-los;
5. Inserir dados coletados na estrutura criada;

6. Realizar comandos SQL para extrair informações da base de dados.

Observações Importantes:

1. Os alunos deverão ter atenção com valores nulos na base de dados e corrigi-los antes de inserir no banco de dados e realizar as operações SQL.
    a. Avalie cada caso e veja qual a melhor escolha para o tratamento.
    b. Avalie cada feature e suas relações umas com as outras.
2. O banco MySQL tem o formato de inserção de datas o ano, mês e dia (Y/m/d). Dessa forma, o aluno deverá mudar a estrutura da data do arquivo de dia, mês e ano (d/m/Y) para fazer a inserção da data no banco.
3. Abaixo segue uma sugestão para modelagem dos dados no banco MySQL:



Atividade II

Os alunos deverão desempenhar as seguintes atividades:

1. Coletar os dados fornecidos (lista_jogadores_chess.json);
2. Criar um banco de dados não relacional MongoDB;
3. Criar uma coleção no banco de dados;
4. Inserir os dados coletados da lista de jogadores na coleção;

5. Realizar comandos NoSQL para extrair informações da base de dados.

Observações Importantes:

1. Os alunos deverão consultar a documentação da API para buscar informações sobre os dados contidos no arquivo JSON.

   https://www.chess.com/news/view/published-data-api

   As descrições de cada atributo dos dados estão na seção (Description: Get additional details about a player in a game.)

2. Após compreender a especificação de cada atributo, resolver as questões objetivas.

Dicas do professor:

1. Sigam rigorosamente as instruções do trabalho.
   a. Não existem pegadinhas nas questões.
2. Tenham cuidado para trabalhar com os valores nulos.
3. Antes de enviar as respostas, verifiquem se o gabarito esteja correto.
4. Tenham atenção no que pede cada questão.
5. Os *datasets* utilizados no desafio podem ser obtidos no link:

   https://github.com/ProfLeandroLessa/desafio-M2-ABD

Bom trabalho prático a todos!