

Megadados
Prova 1
30/11/2020

Prezado(a) Aluno(a),

Você terá 120 minutos a partir do início oficial desta parte da prova para concluir esta avaliação, administre bem o seu tempo. Leia atentamente as instruções a seguir e as questões da prova antes de começar a resolvê-la.

1. Esta avaliação é composta de **3** questões e um total de **3** páginas.
2. Em caso de dúvida sobre alguma questão desta avaliação, redija um texto na folha de prova explicitando-a para que o professor avalie a pertinência durante a correção.
3. Responda as perguntas diretamente no Blackboard.

Boa Prova!

Questão 1:

“Lo marqué un poco con la cabeza y un poco con la mano de Dios” – Maradona

Na Copa do Mundo de 1986 (vosso professor tinha 11 anos – eu vi!) a Argentina disputava com a Inglaterra uma partida das quartas-de-final. No início do segundo tempo Maradona sobe para cabecear e golpeia a bola com a mão, vencendo o goleiro e marcando um gol que deveria ser anulado, mas que foi validado pelo árbitro. Esse gol ficou conhecido como “A Mão de Deus”. Quatro minutos depois Maradona dribla quatro defensores ingleses e o goleiro, marcando o gol que ficou conhecido como “O Gol do Século”, considerado um dos mais bonitos de todas as Copas.

Curioso a respeito desses gols históricos, você decidiu fazer um banco de dados de gols em Copas do Mundo. Eis os requisitos:

- Cada partida envolve dois times, claro.
- O gol é marcado por um jogador, claro.
- O jogador pertence a um time, claro.
- A partida tem um local e data.
- O gol pode ter um apelido.
- O gol é marcado com uma parte do corpo: cabeça, perna direita, perna esquerda, mão direita, mão esquerda (Maradona provou que dá), etc.
- O gol é marcado em um certo minuto da partida.

- a) **(1 pt)** Construa o diagrama entidade-relacionamento para este problema.
- b) **(1 pt)** Construa o diagrama do modelo relacional para este problema.

Questão 2:

Em 1984, em um workshop científico do CERN (*Conseil Européen pour la Recherche Nucléaire* – Conselho Europeu para a Pesquisa Nuclear), surge a ideia de construir um acelerador de partículas maior que todos os existentes, para investigar a existência de partículas subatômicas apenas concebidas em teoria naquele momento. Foi o início do *Large Hadron Collider* (LHC), que após vários anos de negociação entre países, custo de vários bilhões de dólares, e mais de uma década de construção, entrou em operação em 2008. (Dizia-se na época que ele poderia produzir um buraco negro accidental que destruiria o planeta ao entrar em operação. Na dúvida, vosso professor e associados fizeram uma festa-do-fim-do-mundo na véspera, pois não se desperdiça uma oportunidade dessas.)

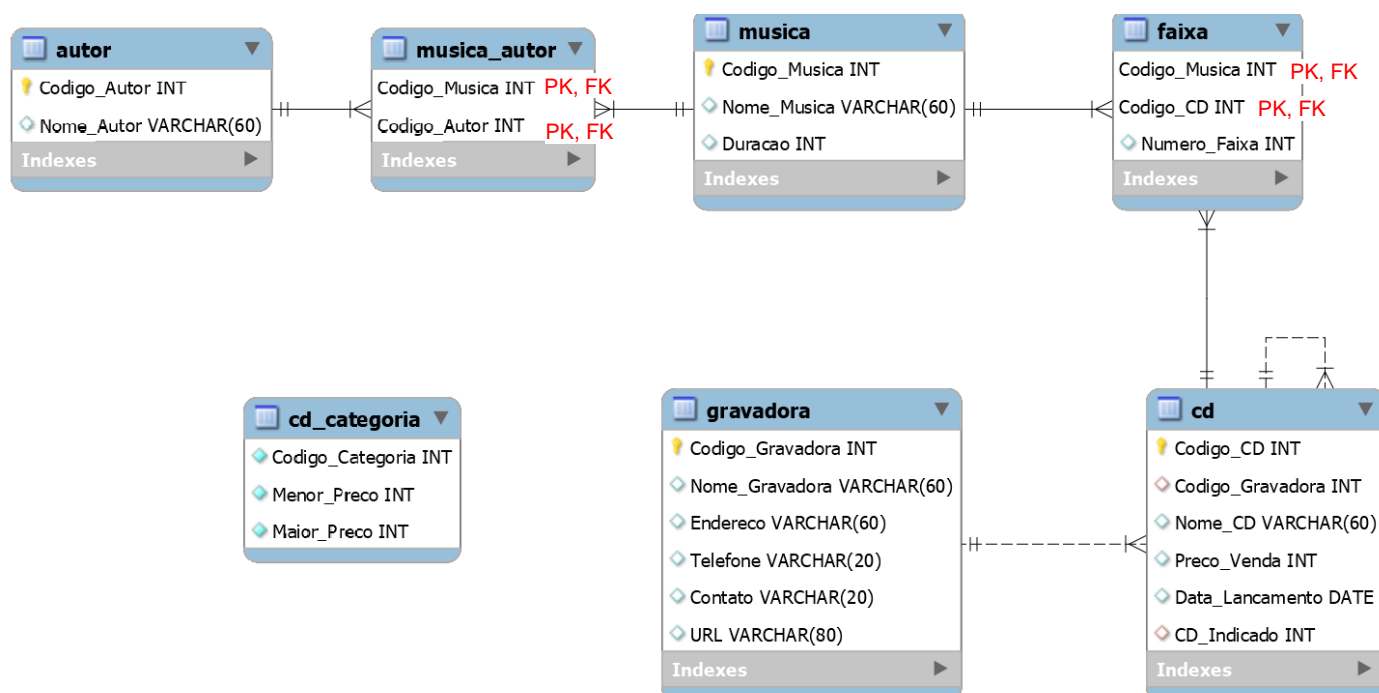
Projetos científicos dessa magnitude fazem mais do que atingir seus objetivos primários – geram ciência e tecnologia colaterais de grande importância. No LHC, um dos grandes desafios estava relacionado ao armazenamento e processamento da imensa massa de dados sendo gerada em alta velocidade pelo dispositivo – cerca de 1 petabyte por dia quando em operação, para um total de 90 petabytes por ano. Para armazenar todos esses dados o CERN desenvolveu seu próprio sistema de armazenamento de dados, o EOS (<https://eos-web.web.cern.ch/eos-web/>).

No site do EOS afirma-se que o sistema tem hoje 250 petabytes de armazenamento bruto (ou seja, disco puro, sem considerar redundâncias de armazenamento), distribuído em 1200 máquinas e em 50 mil discos.

- a) **(0,5 pts)** Qual a capacidade média de armazenamento por disco? Explique seus cálculos.
- b) **(0,5 pts)** Qual o número médio de discos por máquina? Explique seus cálculos.
- c) **(1 pt)** Os discos utilizados no EOS são HDs comuns (não SSDs). A taxa de transferência de dados em um HD é de cerca de 1 megabyte por segundo. Para que possamos transferir continuamente 1 petabytes de dados por dia para os discos, quantos discos deverão estar ativos no mínimo para receber esses dados, sem exceder suas capacidades de transferência de dados? Explique seus cálculos.

Questão 3:

O banco de dados “musica” foi utilizado em aula para vários exercícios. O diagrama do modelo relacional encontra-se abaixo:



As anotações “PK, FK” foram adicionadas por causa do bug do MySQL Workbench, que não desenha as chavinhas laranjas. No resto:

- chavinha amarela denota PK
- diamante azul vazio denota atributo que pode ser nulo
- diamante azul cheio denota atributo não-nulo
- diamante laranja vazio denota FK que pode ser nula

O script “musica.sql” (que foi passado em aula) permite criar as tabelas e populá-las com dados de teste, e está incluído nesta prova para efeito de consulta. Você pode instalar essa base no seu MySQL (se ainda não o fez – a gente usou em aula!) e testar as queries que você deverá desenvolver a seguir.

Nos itens a seguir, escreva *queries* SQL que resolvem o problema.

- (1 pt)** Quais são os autores que trabalham para apenas uma gravadora?
- (1 pt)** Crie uma view que mostre, para cada musica, a data do ultimo lançamento de CD que contenha esta música.
- (1 pt)** Quando um CD é removido, o que acontece com os CDs que o indicavam depende do *constraint* ON DELETE de chave estrangeira. Escreva o que aconteceria para cada possível constraint: CASCADE, RESTRICT, SET NULL.
- (1 pt)** Para este banco de dados, crie um exemplo de transação em nível de isolamento READ COMMITTED na qual o fenômeno de “non-repeatable reads” possa ocorrer.
- (1 pt)** Considere a tabela obtida através do *inner join* das tabelas gravadora e cd. Qual a forma normal desta tabela resultante? Explique.

Questão 4: (1 pt)

O framework Spark permite o processamento de grandes massas de dados de modo distribuido. Explique com suas palavras qual a vantagem de se usar Spark em um cluster de máquinas de capacidade média ao invés de processar uma grande quantidade de dados em um unico servidor de altissima capacidade – ou construa um cenário onde NÃO é vantajoso usar Spark, e o super-servidor é preferível!