Departamento de Engenharia Informática

Exame **sem consulta** com duração de **1h30**

Número: _____ **Nome:** Anabela _____

Exercício resolvido nº 3 do moodle

Grupo I - Modelação Dimensional
(6,5 valores)

A Liga Portuguesa de Futebol possui um sistema informático que armazena variados dados sobre a principal Liga de Futebol desta época. Em particular, sobre os jogadores é armazenado: o número de inscrição do jogador na Liga, o nome completo, a alcunha pelo qual é conhecido no mundo do futebol, a posição em que joga, a morada, o código postal, a data em que foi celebrado o último contracto, a duração (em meses) do contracto, o valor de aquisição do passe, e o valor da cláusula de rescisão. Para além destes dados, é também armazenado o clube atual do jogador. Sobre o clube é armazenado: o seu número de registo na Liga, o nome pelo qual é conhecido, a morada, o código postal, o número de identificação de pessoa colectiva, e a data de fundação. A cada código postal encontra-se sempre associado uma localidade.

A Liga de Futebol encontra-se organizada em 30 jornadas. Cada jornada é realizada em torno do fim-de-semana, sendo caracterizada por uma descrição (e.g., 8ª jornada), pelo número do fim-de-semana do mês (1 a 5), mês, e ano em que decorre. Em cada jornada é também armazenado o n.º de jogos que decorrem na sexta-feira, no sábado, no domingo, e na segunda-feira, num total de 8 jogos.

O sistema operacional da Liga armazena dados sobre os diversos eventos relacionados com cada jogador, no âmbito da cada jornada da Liga em que participa/joga e o respectivo tempo de jogo (e.g.: 67 minutos e 23 segundos) em que ocorreram. Cada evento é caracterizado por um identificador único, uma descrição (e.g.: falta cometida; falta sofrida; golo marcado; penalty falhado; assistência para golo; passe falhado; amostragem de cartão amarelo/vermelho), e pela indicação se o evento corresponde (e.g., uma falta) ou não (e.g., uma passe falhado) a uma infracção às regras do futebol.

1. Seguindo a metodologia *Kimball*, desenvolva o processo de análise dimensional, a fim de definir e criar o esquema conceptual para um *data mart* que permita realizar análises multidimensionais de dados variadas aos eventos registados, de acordo com a realidade descrita. Apresente todos os factos, dimensões, granularidade e todos os aspectos relevantes para o projecto de *data mart*.

   A título de exemplo considere o género de análises que se pretendem realizar:

   - Total de golos marcados, ao longo das várias jornadas, já no tempo de descontos (após os 45 ou 90 minutos);
   - Total de faltas sofridas no 1º, 2º ou 3º quarto de hora, da 1ª ou 2ª parte, por um determinado jogador, numa dada jornada.

2. Admita que se pretende armazenar a(s) habilidade(s)/capacidade(s) específicas que cada jogador tem (*e.g.*: jogador A: marcador de livres, marcador de penalties; jogador B: marcador de cantos diretos; jogador C: marcador de penalties; marcador de cantos diretos; passes longos; dribles). O que acrescentava/alterava a nível de dimensões para suportar o armazenamento destes dados?

Grupo II - Múltipla Escolha

(1 valor cada questão correcta/-0,5 cada questão errada)

Nas questões seguintes assinale apenas uma só alternativa correspondendo à que considera correcta.

1. Qual das seguintes operações não é válida na área de manipulação de dados (*data staging area*), existente nas arquiteturas *BUS* (*Ralph Kimball*) e *CIF* (*Bill Inmon*):
   - ☐ Limpeza de dados.
   - ☐ Atribuição de chaves de substituição.
   - ☐ Análise de dados pelos utilizadores.
   - ☐ Integração de dados provenientes de múltiplos sistemas operacionais.

2. A tabela de factos de um *data mart*:
   - ☐ Armazena as medidas aditivas, semi-aditivas e não aditivas cuja análise é relevante para o negócio.
   - ☐ Possui uma chave primária que é sempre composta pela totalidade das chaves primárias das dimensões existentes.
   - ☐ Pode conter um atributo que representa uma dimensão degenerada (*degenerate dimension*).
   - ☐ Possui todas as características apresentadas nas alíneas anteriores.

3. Existem diversos tipos de índices que podem ser usados em bases de dados, em função das características dos atributos. Em particular:
   - ☐ Um índice do tipo *B-Tree* não pode envolver múltiplos atributos.
   - ☐ Um índice do tipo *Bitmap* é apropriado para atributos com elevada cardinalidade.
   - ☐ A grande maioria das bases de dados suporta índices do tipo *Hash*.
   - ☐ Todas as afirmações anteriores são falsas.

4. Uma estratégia de optimização em armazéns de dados envolve a criação de agregações. A criação dessas agregações pode ser feita:
   - ☐ Através de uma operação de *Group By* sobre uma cópia/réplica dos dados transacionais que se encontram na *Data Staging Area*.
   - ☐ Através de uma operação de *Group By* sobre os dados que se encontram numa tabela de factos a um nível de granularidade mais elementar.
   - ☐ Através da adição dos dados transacionais a acumuladores/totalizadores especialmente criados para o efeito na *Data Staging Area*.
   - ☐ Por qualquer um das formas descritas nas alíneas anteriores.

5. No On-Line Analytical Processing (OLAP) há diversas operações de análise de dados que os utilizadores podem realizar, nomeadamente:
   - ☐ Operação de *Drill-down* que consiste em efetuar análises a um menor nível de detalhe.
   - ☐ Operação de *Dice* que pode ser combinada com a operação de *Slice*.
   - ☐ Operação de *Dice* que consiste em seleccionar um sub-cubo de dados composto por uma só dimensão.
   - ☐ Nenhuma das afirmações que constam dos pontos anteriores é válida.

Grupo III – Verdadeiros ou Falsos com Justificação
(2 valores cada questão)

Indique se as seguintes afirmações são verdadeiras ou falsas, apresentando a respectiva justificação.

1. A utilização de chaves de substituição nas dimensões, em detrimento das chaves dos sistemas operacionais, justifica-se unicamente por questões de performance.

2. Na extração de dados estática (a partir dos sistemas operacionais), a única estratégia disponível consiste em utilizar força bruta, ou seja, comparar os atributos dos registos do sistema operacional com os atributos dos registos no armazém de dados.

3. Sempre que se optimiza um armazém de dados através da introdução de agregações, isso implica que sejam criadas novas tabelas (de factos) para as armazenar.

Grupo IV – Questão de Desenvolvimento
(2,5 valores)

Durante o processo de desenvolvimento de um armazém de dados surge, frequentemente, o problema dos valores monetários estarem em moedas diferentes. Dependo da situação, há três estratégias diferentes que podem ser adoptadas para solucionar o problema. Explique e exemplifique a nível do modelo dimensional cada uma dessas três estratégias.

1. Falso. Existem outros motivos por além de performance também queremos garantir integridade referencial.
2. Falso - Existem outras estratégias como extração através de uma abordagem, Cyclic-Redudancy Checksum.
3. Falso. São criadas tabelas adicionais, mas não de factos.

Resposta Aberta:

As 3 alternativas são:

- Pôr todos os dados na tabela de factos, no entanto o número de colunas será bastante elevado.

- Criar uma dimensão de suporte, DimCurrency. Apesar de ser uma melhor solução, o número de pares (Standard e local) é elevado na mesma.

- Possuir a DimCurrency e um FactCurrencyRates, sendo a solução mais flexível e poderosa