

Número: _____ Nome: Apicultor

Ex3 resolvido do moodle
com umas ligeiras
alterações, mas a lógica
permanece a mesma

Grupo I - Modelação Dimensional **(6,5 valores)**

A Federação Portuguesa de Futebol possui um sistema operacional que armazena dados variados sobre as diversas competições profissionais e amadoras de futebol em Portugal. Em particular, sobre cada competição é armazenado: um identificador único; uma descrição (*e.g.*, primeira liga; taça de Portugal); o n.º de jornadas/eliminatórias; o n.º de clubes participantes; mês e n.º do fim de semana da 1ª jornada/eliminatória; mês e n.º do fim de semana da última jornada/eliminatória; e, indicação se é uma competição amadora ou profissional. As diversas competições decorrem no âmbito de épocas futebolísticas. Cada época é caracterizada por um identificador único; por um descritivo (*e.g.*, época 2014/15); e, pelas datas oficiais de início e fim de época.

As diversas competições, em cada época futebolística, encontram-se organizadas em jornadas/eliminatórias que decorrem ao fim de semana. Cada jornada/eliminatória é caracterizada por um identificador único e por um descritivo (*e.g.*, 16ª jornada; meia-final). Independentemente de serem profissionais ou amadores, os atores principais são os jogadores de futebol. Sobre cada jogador é armazenado: o número de inscrição do jogador na Federação; o nome completo; a alcunha pelo qual é conhecido; a posição em que joga; a morada; o código postal; a data de celebração do último contrato; a duração (em meses) do contrato; 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. Ao longo do tempo, além de poder jogar por clubes diferentes, o mesmo jogador normalmente joga em competições diferentes (*e.g.*, primeira liga e taça da liga). Sobre o clube é armazenado: o seu número de registo na Federação; a sua designação; 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.

O sistema operacional da Federação Portuguesa de Futebol armazena dados estatísticos referentes a cada jogador, no âmbito da cada jornada/eliminatória em que efetivamente participa/joga, referente às diversas épocas futebolísticas, e às diferentes competições futebolísticas. Em particular, são armazenadas os seguintes dados estatísticos: n.º de minutos jogados; n.º de golos marcados ou sofridos; n.º de faltas cometidas; n.º de faltas sofridas; n.º de cartões amarelos (0, 1 ou 2); n.º de cartões vermelhos (0 ou 1); n.º de passes bem sucedidos; n.º de passes falhados; e, n.º de assistências para golo. Além destes dados, o sistema armazena também o ano, mês e n.º do fim de semana em que decorreu a jornada/eliminatória em questão.

1. Seguindo a metodologia *Kimball*, desenvolva o processo de análise dimensional, a fim de definir e criar o modelo dimensional para um *data mart* que permita realizar análises multidimensionais de dados variadas às estatísticas existentes, de acordo com a realidade que acabou de ser descrita. Apresente todos os factos, dimensões, granularidade e todos os aspectos relevantes para o projecto de *data mart*.

2. Admita que se pretendem efetuar as seguintes análises de dados:
- Adicionaria uma nova tabela de factos referentes aos jogadores não presentes em campo**
- Dada uma época, uma competição e uma jornada/eliminatória, saber-se quais os jogadores que nela não jogaram.
 - Dada uma época, uma competição e um jogador, saber-se quais as jornadas/eliminatórias em que este não jogou.

O que acrescenta ou altera a nível do modelo dimensional para suportar análises deste tipo? Explique de que forma essas análises podem ser realizadas.

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. Na base da criação de um modelo dimensional encontram-se objetivos relacionados com:
- ☐ Simplicidade do modelo, o que o torna facilmente compreensível por utilizadores que não sejam informáticos.
 - ☐ Performance dos inquéritos a realizar, atendendo a que o motor da base de dados pode tirar partido dos *joins* existentes entre tabelas com volumes de dados reduzidos (dimensões) com uma tabela com um volume de dados elevado (tabela de factos).
 - ☐ Flexibilidade na inclusão de novos atributos nas dimensões, novas medidas/factos na tabela de factos, e novas dimensões a um modelo já existente.
 - ☒ Todos os motivos apresentados nas alíneas anteriores.
2. Na arquitetura *BUS* (*Kimball*):
- ☒ Os *data mart* agregados representam o mesmo que os *data mart* existentes na arquitetura *CIF* (*Inmom*).
 - ☐ O armazém de dados resulta unicamente do conjunto dos vários *data mart* atómicos.
 - ☐ A existência de esquemas de dados no formato relacional não é aceitável em nenhum componente.
 - ☐ Há um repositório de dados físico equivalente ao armazém de dados da arquitetura *CIF* (*Inmom*).
3. Em armazéns de dados é frequente o requisito de se poderem realizar análises de dados conjuntas que envolvem moedas diferentes. Uma forma de suportar essas análises de dados consiste em:
- ☐ Incluir colunas nas dimensões para cada uma das moedas pretendidas.
 - ☐ Adotar uma moeda de uniformização (*e.g.*, Euros) e representar todos os valores monetários unicamente nesta moeda, mediante a realização de operações de conversão (câmbio).
 - ☐ Criar uma tabela de factos que contém os fatores de conversão (câmbios) com a granularidade pretendida (*e.g.*, diária) entre cada par de moedas.
 - ☒ Utilizar qualquer uma das abordagens apresentadas nas alíneas anteriores.
4. Uma estratégia de otimização frequentemente utilizada em armazéns de dados envolve a criação de índices. Em particular, numa tabela de factos:

- ☐ Não se justifica proceder à criação de índices individuais para cada uma das chaves estrangeiras para as dimensões.
- ☒ A ordem dos atributos na formação da chave primária e, consequentemente, do respetivo índice é determinante na performance das consultas.
- ☐ Os factos/medidas existentes numa tabela de factos nunca são indexados.
- ☐ Todas as afirmações anteriores estão corretas.

5. Existem diversos tipos de *On-Line Analytical Processing (OLAP)*, tendo o *Multidimensional OLAP (MOLAP)* sido o primeiro a surgir. Qual das seguintes afirmações relativas ao *MOLAP* não está correta ?
- ☐ Como os dados são armazenados a nível de memória primária, isso limita o volume de dados que pode ser manipulado.
 - ☒ A realização das consultas/análises de dados é mais lenta do que no *Relational OLAP*.
 - ☐ É necessária a realização de uma operação de processamento e armazenamento dos dados num hipercubo que pode ser lenta.
 - ☐ O hipercubo necessita de atualizações periódicas de modo a manter-se corrente e atualizado.

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. *Kimball (Bus architecture)* e *Inmon (CIF architecture)* concordam que o maior poder dos dados se encontra no nível mais atómico.

Verdadeiro - Tendo a informação no nível mais elementar qualquer tipo e consulta (detalhada ou genérica) é possível ser feita.

Isto está nos slides: "Power of information resides in the atomic data, which embed all available information dimensionality"

2. Uma tabela de factos é sempre composta pela chave primária, formada pelo conjunto ou subconjunto de chaves estrangeiras, por um conjunto de medidas (factos) cuja análise é relevante para o negócio e, eventualmente, uma ou mais dimensões degeneradas.

Falso - A tabela de factos pode não conter medidas, ou seja, ser um factless table.

3. Numa situação em que a probabilidade de ocorrer uma alteração ao valor de um atributo de uma dimensão seja muito baixa, mas em que seja relevante manter um histórico da alteração, uma estratégia de *Slow Changing Dimension (SCD)* – Tipo 3 é apropriada.

Verdadeiro - Sendo que a probabilidade é baixa e apenas queremos manter um histórico da alteração, o tipo 3 suporta essas condições, mantendo o valor alterado e o já existente.

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

Na extração dos dados a partir dos sistemas fonte (e.g., sistemas operacionais), há duas estratégias distintas que podem ser adotadas. A estratégia a adotar depende das características que os sistemas fonte possuam. Apresente e caracterize cada uma dessas estratégias de extração de dados.

Existem 2 técnicas de extração, estática e incremental. A extração estática, normalmente, realizada primeiro captura todos os dados fonte numa dada altura. A extração incremental, normalmente, realizada posteriormente captura os dados fonte em comparação com a última extração realizada.