Departamento de Engenharia Informática

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

Número: _____ **Nome:** Salazar

Grupo I - Modelação Dimensional
(6,5 valores)

Um determinado aeroporto possui um sistema operacional que armazena dados sobre os voos dos passageiros. Cada passageiro é caracterizado por: número que o identifica de forma única; nome; nacionalidade; tipo de documento de identificação (bilhete de identidade; cartão de cidadão; ou, passaporte); respetivo número do documento; data de nascimento; contato telefónico; e-mail; endereço; código postal; cidade; e, país de residência. Os voos são organizados por companhias de aviação, sendo que cada uma destas é caracterizada por: código companhia (identificador único); nome; contato telefónico; endereço; código postal; cidade; país; data de início de atividade da companhia para o aeroporto em questão; e, eventualmente, data de fim dessa atividade. A cada código postal, de cada país, corresponde a respetiva localidade.

Os voos são realizados nos aviões que as companhias de aviação possuem. Cada avião é ainda caracterizado pelo: identificador único; construtor (e.g.: Boeing; AirBus); modelo (e.g.: 737; A380); nome dado pela companhia; capacidade (em n.º de passageiro); e, peso (em toneladas). Os voos são organizados pelas companhias de aviação, a partir do aeroporto em questão para aeroportos de destino. Cada voo é ainda caracterizado por: código do voo (diferente para cada companhia); frequência de realização (e.g.: diário; semanal; 2ª-feira, 4ª-feira e 6ª-feira); hora prevista de partida; e, hora prevista de chegada (hora local). Cada aeroporto de destino é caracterizado por: código internacional de aeroporto (único para cada aeroporto); nome; cidade; país; e, fuso horário.

Sempre que um passageiro realiza um voo para um outro aeroporto de destino, o sistema operacional do aeroporto regista: o passageiro envolvido; o código do voo; a data em que ocorreu; a hora de partida (pode não ser a mesma que a prevista para o voo devido a algum motivo excecional); o avião utilizado da companhia de aviação; o n.º de bagagens de mão transportadas pelo passageiro para a cabine; o respetivo peso destas; o n.º de bagagens transportadas no porão do avião; e, o respetivo peso destas.

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 aos voos realizados pelos passageiros, de acordo com a realidade que acabou de ser descrita. Apresente todos os factos, dimensões, granularidade e todos os aspectos relevantes para o projeto de *data mart*.

2. Admita que se pretendem realizar as seguintes análises de dados:

Para resolver este problema, seria criada uma nova dimensão, DimDia

- Dado um código de voo (e.g., LH701), saber-se quais os dias em que o voo não se realizou (devido a algum motivo de força maior).
- Dada uma data, saber-se quais os voos que não se realizaram nessa data (devido a algum motivo de força maior).

O que acrescentaria ou alteraria a nível do modelo dimensional (tabelas de factos e/ou dimensões) para suportar a realização deste tipo de análises? Explique como poderia realizar as referidas consultas.

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. *Ralph Kimball* (*Bus Architecture*) e *Bill Inmon* (*CIF Architecture*) defendem:
   - ☐ Que os dados armazenados nos armazéns de dados devem estar sempre no nível mais atómico (elementar).
   - ☐ Que o maior poder/flexibilidade que os dados oferecem encontra-se no nível mais atómico.
   - ☐ Que o maior poder/flexibilidade que os dados oferecem resulta destes estarem agregados.
   - ☐ Outras ideias/posições que não as referidas nas alíneas anteriores.

2. Numa situação em que seja relevante poder continuar a efetuar análises de dados como se uma dada alteração não tivesse ocorrido, a estratégia de *Slowly Changing Dimension* (SCD) adequada é de:
   - ☐ Tipo 1.
   - ☐ Tipo 2.
   - ☐ Tipo 3.
   - ☐ Tipo 2 ou Tipo 3.

3. A operação de *roll-up* suportada pelas ferramentas OLAP permite:
   - ☐ Extrair um sub-cubo a partir do cubo de dados original.
   - ☐ Visualizar os dados com igual nível de detalhe/granularidade, mas de diferentes perspetivas.
   - ☐ Efetuar análises com um maior nível de detalhe/granularidade.
   - ☐ Efetuar análises com um menor nível de detalhe/granularidade.

4. Uma mini-dimensão é utilizada em armazéns de dados para:
   - ☐ Armazenar as combinações distintas dos valores dos atributos demográficos, sejam estes contínuos (e.g., peso) ou discretos (e.g., idade).
   - ☐ Armazenar as combinações distintas dos valores dos atributos demográficos, desde que todos estes sejam discretos ou tenham sido transformados em discretos.
   - ☐ Armazenar atributos do tipo *flag* e outros atributos que contêm um conjunto reduzido de valores discretos.
   - ☐ Armazenar atributos do tipo *flag* e outros atributos que contêm um conjunto de valores contínuos e discretos.

5. O particionamento horizontal é uma estratégia de otimização que pode ser usada em armazéns de dados em que:
   - ☐ Os dados podem ser particionados por intervalos de valores ou listas de valores.
   - ☐ Os dados apenas podem ser particionados por intervalos de valores.
   - ☐ A estrutura das tabelas resultante são diferentes de particionamento para particionamento.

- ☐ Os atributos da tabela original dão origem a diferentes tabelas, repetindo-se apenas a chave primária da tabela original.

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 margem bruta ((*valor_vendas* – *custo_vendas*) / *valor_vendas*) obtida a partir de uma tabela de factos de um *data mart* de vendas constitui um facto semi-aditivo.

Falso - Representa uma medida não-aditiva. Não pode ser somado ao longo das dimensões.

2. Para além dos atributos que formam a chave primária, uma tabela de factos inclui sempre um conjunto de outros atributos numéricos cuja relevância é importante para a área de negócio em questão, sendo estes designados de factos/medidas.

Falso - A tabela de factos pode não conter medidas, sendo uma factless table

3. Para além do índice associado à chave primária de uma tabela de factos, não se justifica a criação de qualquer outro índice.

Falso - Se uma dada métrica for muito usada, pode ser criado um índice tb

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

Uma das estratégias de otimização vulgarmente utilizada em armazéns de dados envolve a criação de agregações. Explique em que consiste esta estratégia de otimização, quais as vantagens e desvantagens, assim como as técnicas que podem ser adotadas para armazenar as agregações.

O que são agregações? -> Resumir um conjunto de métricas guardadas na tabela de factos, com o propósito de acelerar queries.

Vantagens: Melhor performance do DW, transparente para com os end-users, partilhado por vários utilizadores.

Desvantagens: Constante atenção por parte do admin da DW, gasta espaço de armazenamento e apenas acelera consultas previmanente calculadas