Departamento de Engenharia Informática

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

Número: _____ **Nome:** Gonçalo _____

Grupo I - Modelação Dimensional
(6,5 valores)

Uma empresa de comércio eletrónico pretende tirar partido dos dados existentes sobre os utilizadores registados no seu sítio *web*, assim como dos dados que vão sendo armazenados no *log* do seu servidor. Em particular, sobre cada utilizador registado é conhecido: o utilizador (diferente para cada utilizador); o nome completo; a morada; o código postal; o sexo; a profissão; o nome da empresa; a escolaridade (*e.g.*, licenciatura, mestrado); o e-mail; a data do último acesso ao sítio web; o n.º total de compras efetuadas; o respetivo valor dessas compras; e, a data da última compra realizada. A cada código postal encontra-se associado uma localidade.

O sítio *web* é composto por várias páginas nas quais os utilizadores podem encontrar e adquirir os produtos que a empresa comercializa, organizados em função da família e categoria do produto. Em particular, sobre cada página *web* é conhecido: um identificador único; o tipo (*i.e.*, estática ou dinâmica); a função (*e.g.*, descrição detalhada do produto; informação sobre a empresa); o nome físico; o tamanho (em *bytes*); e, a data da sua criação. Diversos eventos sobre as páginas podem ser gerados pelos utilizadores enquanto navegam/exploram o sítio *web* da empresa. Todos estes eventos vão sendo armazenados no *log* do servidor *web*. Cada evento distinto é caracterizado por um identificador e pelo respetivo descritivo (*e.g.*, abrir página; botão *refresh* do *browser* premido; botão *stop* do *browser* premido; colocação do produto no carrinho de compras). Além da família e categoria do produto, é ainda conhecido: o código; a descrição; o preço de venda atual; o preço de custo atual; o preço de custo ponderado; e, o stock mínimo que, quando atingido, despoleta a sua encomenda.

Sempre que um utilizador origina um evento numa página *web*, com base nos dados existentes sobre utilizadores registados e nos dados armazenados no *log* do servidor *web*, é possível saber-se: a data e hora local em que o evento ocorreu; a data e hora universal em que o evento ocorreu; e, o identificador único de acesso do utilizador, gerado automaticamente no momento em que este entrou no sítio *web* com as suas credenciais de acesso (utilizador e *password*). Caso o utilizador coloque algum produto no carrinho de compras numa dada página *web*, é ainda possível saber-se: o código do produto; o n.º de unidades colocadas; o preço unitário de venda do produto; e, o respetivo valor da venda. Note-se, no entanto, que muitas vezes um utilizador acede a uma página *web* mas não adiciona qualquer produto ao seu carrinho de compras.

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 eventos originados pelos utilizadores nas páginas *web*, 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 no final do dia é conhecido o valor do stock de cada produto que a empresa comercializa, resultante das vendas efetuadas e de eventuais novas entradas desse produto provenientes de fornecedores. O que acrescentaria ou alteraria a nível do modelo dimensional para suportar análises/consultas de dados diárias aos stocks dos produtos ?

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 correta.

1. Uma característica comum/usual nos armazéns de dados é:
   - ☐ Principal finalidade consiste em suportar a tomada de decisões operacionais.
   - ☐ Necessitam de consolidar dados provenientes de sistemas operacionais diferentes.
   - ☐ Volume de dados iguala o volume de dados total existente nos vários sistemas operacionais que o abastecem.
   - ☐ Regista transações curtas e isoladas que envolvem dados no estado atómico.

2. A inclusão de um atributo do tipo dimensão degenerada (*degenerate dimension*) numa tabela de factos pode justificar-se:
   - ☐ Para permitir a unicidade da sua chave primária.
   - ☐ Como forma de estabelecer uma ligação com os correspondentes registos do sistema operacional.
   - ☐ Por qualquer um ou ambos os motivos apresentados nas alíneas anteriores.
   - ☐ Por outros motivos que não os apresentados nas alíneas anteriores.

3. Numa dimensão Cliente de elevado volume de dados pode recorrer-se à criação de uma mini-dimensão. Uma mini-dimensão:
   - ☐ É criada mediante a combinação dos valores distintos dos diferentes atributos.
   - ☐ Incluiu os atributos que são mais suscetíveis a alterações frequentes.
   - ☐ Por questões de performance, o número de combinações dos valores distintos dos atributos não deve resultar num n.º de registos superior a 100000.
   - ☐ Possui todas as características apresentadas nas alíneas anteriores.

4. Em determinadas situações pode ser necessário proceder à correção de factos que já tenham sido carregados na tabela de factos. Por questões de auditoria/fiscais a abordagem adequada consiste em:
   - ☐ Negar o facto (medidas com valores negativos) e voltar a proceder ao seu carregamento.
   - ☐ Apagar logicamente o facto e voltar a proceder ao seu carregamento.
   - ☐ Utilizar qualquer uma das duas abordagens anteriores.
   - ☐ Apagar fisicamente o facto e voltar a proceder ao seu carregamento.

5. Ao passar-se a armazenar valores agregados numa tabela de factos existente que se encontra no nível atómico/elementar:
   - ☐ O número de registos a criar nessa tabela de factos será superior ao número de registos que seria criado numa nova tabela apenas com esses valores agregados.
   - ☐ Os atributos que armazenam os factos/medidas não necessitam de sofrer qualquer alteração.
   - ☐ Implica que o esquema/estrutura das dimensões anteriormente existentes seja alterado.
   - ☐ Tal situação não é possível, em virtude de não se poder armazenar factos com granularidades diferentes na mesma tabela de factos.

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. Uma tabela de factos pode armazenar três tipos de medidas: aditivas; semi-aditivas; e, não aditivas.

Falso - A tabela de factos não pode armazenar medidas não-aditivas

2. Sabe-se que a generalidade das análises/consultas de dados efetuadas num armazém de dados têm sempre o aspeto temporal (data e/ou tempo) presente. Assim, na definição da chave primária da tabela de factos e, consequentemente, do respetivo índice, há que ter este aspeto em consideração.

Verdadeiro - pois é importante, eis a justificação: . A inclusão do aspeto temporal na chave primária pode ajudar a garantir que os dados sejam organizados de maneira eficaz para consultas analíticas que envolvem análises temporais, como tendências, comparações entre períodos e assim por diante.

3. A implementação de um mecanismo de *Slowly Changing Dimension* (SCD) – Tipo 2 para armazenamento do histórico das alterações que ocorrem aos atributos de uma dimensão, pode ser plenamente alcançada utilizando unicamente um atributo que indica se é o registo mais atual ou não (*isCurrent*).

Falso - O uso dos atributos EffectiveDate e ExpiredDate é imperativo no caso de atributos scd tipo 2 e o iscurrent por si não suporta essa condição e nem existe o conceito de histórico.

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

Dois dos principais tipos de *On-Line Analytical Processing* (*OLAP*) são o *Multidimensional OLAP* (*MOLAP*) e o *Relational OLAP* (*ROLAP*). Apresente as características, vantagens e desvantagens de cada um destes tipos de *OLAP*.

MOLAP(guarda os dados numa matriz multidimensional, um hipercubo e precisa do pré-processamento de dados):

Vantagens: Melhor performance, computação de agregados de alto-nível automatizada e não existe a necessidade de uma ligação constante ao DW.

Desvantagens: O carregamento de dados pode ser demorado, especialmente com volumes de dados grandes (necessita de fazer atualizações periódicas);
necessita de espaço adicional, pois guarda uma cópia dos dados no servidor OLAP.

ROLAP(não precisa de pré-processamento, acede aos dados em tabelas relacionais):

Vantagens: Melhor escalabilidade, consegue lidar melhor com volumes de dados maiores, atualizações frequentes não causam problemas e necessita de espaço de armazenamento.

Desvantagens: A estrutura é relacional, logo SQL tem de ser usado por motivos de custo; consultas podem ser mais demoradas