**Teste ao exame**

\_F\_ A introdução da gestão de dados mestre numa organização não afeta o funcionamento dos sistemas operacionais.

\_V\_ A gestão de dados mestre requer a atribuição de um dono e de uma autoridade responsável a todos os itens de dados.

\_F\_ Na construção de data warehouses a vertente de aplicações de um projeto pode ser realizada autonomamente da vertente de tecnologias.

\_F\_ Na construção de data warehouses, na abordagem bottom-up, os data marts são feitos a pedido e derivam do data warehouse.

\_V\_ Data warehouses facilitam o acesso a dados históricos, sendo esta uma vantagem face às bases de dados federadas.  
\_V\_ Em relatórios analíticos as extensões OLAP ao SQL permitem optimizar o cálculo de valores agregados.

\_V\_ Em modelação dimensional relatórios já existentes devem ser a base de partida para a modelação.

\_F\_ Chaves substitutas em dimensões impedem o registo do histórico de mudanças lentas.

\_F\_ Em modelação dimensional as tabelas de factos sem factos servem para acompanhamento simples de eventos.

\_V\_ No desenho físico de data warehouses duas das técnicas são as vistas materializadas e as bifurcações de tabelas.

\_V\_ No desenho físico de data warehouses o resultado de uma interrogação pode ser totalmente obtido por via de uma vista materializada.

\_V\_ A importância do desenho físico de data warehouses aumenta com a quantidade de dados que precisam de ser processados.

**Indique duas razões para a inclusão de extensões OLAP na linguagem SQL no contexto dos relatórios analíticos**

Sem a inclusão de extensões OLAP na linguagem SQL os relatórios analíticos ficariam muito extensos, com muitas interrogações, muita informação e muitas queries, o que o tornaria mais lento. As extensões OLAP na linguagem SQL permitem o cálculo de dados agregados para todas as dimensões e combinações.

**Indique frases verdadeiras ou falsas (V/F) sobre sistemas de informação para executivos.**

\_v\_ Fornecem muitos relatórios predefinidos para cobrir o máximo de cenários de decisão.

\_v\_ Mostram tendências nos dados ao longo do tempo, abrangendo vários anos se necessário.

\_f\_ Disponibilizam e cruzam dados de múltiplas fontes, desde que apenas internas à organização.

\_v\_ Requerem pouco ou nenhum treino para poderem ser usados pelos decisores.

\_f\_ Representam os dados apenas em forma de gráficos, em vez de texto ou tabelas.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre bases de dados federadas.**

\_f\_ Guardam e usam uma cópia dos dados, à semelhança dos *data warehouses*.

\_f\_ Interrogações analíticas afetam pouco o desempenho dos sistemas operacionais.

\_v\_ Mediador permite abstrair o tipo dos vários sistemas operacionais envolvidos.

\_v\_ Interrogações analíticas são respondidas com dados atuais dos sistemas operacionais.

\_f\_ Relatórios de apoio à decisão com dados históricos são fáceis de obter.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre bases de dados federadas.**

\_f\_ Relatórios de apoio à decisão com dados históricos costumam ser viáveis.

\_v\_ O desempenho dos sistemas operacionais é afetado pelas interrogações analíticas.

\_v\_ O mediador permite abstrair o tipo dos vários sistemas operacionais envolvidos.

\_f\_ As respostas às perguntas analíticas têm dados mais antigos do que com data warehouses.

\_v\_ As interrogações analíticas são difíceis de construir pois não existe um modelo dimensional.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre dados abertos.**

\_\_ Podem ser combinados com outros dados abertos, e redistribuídos livremente.

\_\_ São capturados de forma passiva por indivíduos, tal como os grandes dados (big data).

\_f\_ Devem ser guardados em formato PDF, pois pode ser lido em quase todos os computadores.

\_\_ Os enriquecedores colecionam e analisam dados, e cobram pelos seus resultados.

\_v\_ Oferecem fontes de dados externas adicionais a um data warehouse

**Indique afirmações verdadeiras (V) ou falsas (F) sobre a gestão de dados mestre.**

\_f\_ Só é útil quando a organização dispõe de um data warehouse.

\_v\_ Procura atingir o objetivo de haver uma versão única da verdade sobre os dados.

\_v\_ Facilita o carregamento de dados para as dimensões de um data warehouse.

\_f\_ A sua introdução numa organização não afeta o funcionamento dos sistemas operacionais.

\_v\_ Requer a atribuição de um dono e de uma autoridade responsável a todos os itens de dados.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre sistemas de apoio à decisão.**

\_v\_ As transações costumam ser curtas relativamente às dos sistemas operacionais.

\_v\_ As análises de dados tendem a ser exploratórias e a visar estudos de tendências.

\_f\_ A navegação nos dados é versátil mas requer muito treino dos executivos.

\_v\_ Os dados disponibilizados para a tomada de decisão costumam ser apenas agregados.

\_f\_ Devido ao grande volume de dados o tempo de resposta do sistema costuma ser longo.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos em SQL.**

\_f\_ Com SQL clássico, o número de SELECTs é igual ao número de dimensões no relatório.

\_v\_ A cláusula GROUP BY CUBE faz parte das extensões OLAP para SQL.

\_f\_ GROUP BY ROLLUP (A, B, C) inclui cálculos agregados em função só de C.

\_v\_ É incorreto usar SELECT A, B, SUM(C) FROM T GROUP BY ROLLUP (A, B, C).

\_v\_ Um só SELECT de SQL clássico pode gerar um relatório com totais em colunas e linhas.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.**

\_f\_ A operação de dicing é um caso particularmente simples de slicing.

\_v\_ As medidas só podem ser colocadas num relatório após os atributos das dimensões.

\_f\_ Fazer drill-down geralmente revela menos valores do que os que estavam disponíveis.

\_v\_ Cada medida M colocada num relatório acrescenta, por omissão, SUM(M) ao SELECT em SQL.

\_f\_ Um atributo A posto nas linhas de um relatório vazio gera SELECT A … GROUP BY CUBE (A)

**Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.**

\_f\_ São calculadas menos somas se for usado CUBE em vez de ROLLUP num GROUP BY.

\_v\_ As extensões OLAP ao SQL permitem otimizar o cálculo de valores agregados.

\_f\_ A operação de *slice* é um caso particularmente simples da operação *dice*.

\_f\_ *Roll-up serve para procurar explicações detalhadas para os grandes totais.*

\_v\_ Podem ser incluídos atributos de mais de duas dimensões num relatório.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.**

\_f\_ Só pode ser aplicada uma operação de *slice* em cada relatório.

\_v\_ As hierarquias de atributos de dimensões possibilitam o *roll-up* e o *drill-down*.

\_v\_ O *roll-up* serve para passar de dados mais detalhados para dados mais agregados.

\_f\_ As extensões OLAP ao SQL não simplificam a escrita de comandos que geram relatórios.

\_v\_ As medidas só podem ser colocadas num relatório após os atributos das dimensões.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.**

\_v\_ Os atributos das dimensões podem ser usados para esmiuçar as medidas numéricas.

\_v\_ O *slicing* permite aplicar restrições por intervalo de valores numa dimensão.

\_f\_ A cláusula GROUP BY ROLLUP permite gerar totais para todas as combinações de atributos.

\_v\_ A operação de *dice* pode ser usada para restringir os valores que aparecem num relatório.

\_f\_ O *roll-up* serve para passar de dados mais agregados para dados mais detalhados.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre esquemas em estrela.**

\_f\_ Costumam ser usados em *data warehouses* e também em sistemas operacionais.

\_v\_ A tabela de factos está ligada a todas as tabelas de dimensões.

\_f\_ A dimensão com mais atributos está tipicamente no centro do esquema.

\_f\_ As medidas só são permitidas na tabela de factos.

\_f\_ É comum duas ou mais dimensões estarem ligadas diretamente entre si.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre tabelas de ponte.**

\_v\_ Permitem navegar em hierarquias de profundidade variável.

\_f\_ Servem para ligar duas tabelas de factos e permitir relatórios transdepartamentais.

\_f\_ Cada linha da tabela inclui apontadores para os níveis ascendente e descendente.

\_v\_ Guardam caminhos entre todos os valores de todos os níveis de uma hierarquia.

\_v\_ Cada tabela costuma ter menos linhas do que a tabela de dimensão que lhe serve de base.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre tabelas de ponte.**  
\_v\_ Guardam caminhos entre todos os elementos de todos os níveis de uma hierarquia.  
\_v\_ Permitem agregar medidas para cada elemento abaixo de um dado elemento.  
\_v\_ São habitualmente usadas entre uma dimensão e uma tabela de factos.  
\_f\_ Não podem ser usadas para representar hierarquias de profundidade variável.  
\_f\_ A remoção de um elemento da hierarquia tem pouco impacto nas linhas da tabela

**Indique afirmações verdadeiras (V) ou falsas (F) sobre modelação dimensional.**

\_f\_ As dimensões e as medidas costumam ser substantivos e os factos costumam ser verbos.

\_v\_ Numa tabela de factos apenas as medidas de negócio podem ser atributos numéricos.

\_v\_ Numa mesma tabela de dimensão podem coexistir várias hierarquias de atributos.

\_f\_ As tabelas de factos costumam ter muitas colunas e as de dimensões tendem a ser estreitas.

\_f\_ Uma mini-dimensão está ligada diretamente à dimensão "monstra" da qual derivou.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre modelação dimensional.**  
\_\_ Relatórios já existentes devem ser a base de partida para a modelação.  
\_f\_ As tabelas de factos sem factos servem para acompanhamento simples de eventos.  
\_v\_ A técnica 3 para mudanças lentas é adequada para valores antigos e atuais de muitas colunas.  
\_v\_ Uma dimensão conformada tem nomes e valores de atributos comuns em vários processos.  
\_f\_ Chaves substitutas em dimensões impedem o registo do histórico de mudanças lentas.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre a modelação dimensional.**  
\_f\_ A escolha do processo de negócio é feita depois de comparadas tabelas de factos alternativas.  
\_f\_ Cada processo de negócio tem a sua própria matriz de exequibilidade/valor.  
\_v\_ As medidas só devem ser identificadas depois de definidos os atributos das dimensões.  
\_v\_ Os processos prioritários estão no quadrante superior direito da matriz de processos.  
\_v\_ O grão determina os conceitos participantes em cada evento guardado na tabela de factos

**Indique afirmações verdadeiras (V) ou falsas (F) sobre discos rígidos.**

\_f\_ O disco rígido é a componente que menos limita o desempenho de um *data warehouse*.

\_v\_ O tempo de posicionamento aumenta com a distância da cabeça de leitura à pista dos dados.

\_v\_ A transferência de dados em posições seguidas é mais rápida do que se estiverem dispersas.

\_v\_ O atraso de rotação é máximo quando os dados estão diametralmente opostos à cabeça.

\_f\_ Numa *solid state drive* (SSD) o tempo de acesso não depende da localização dos dados.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre desenho físico de data warehouses.**  
\_v\_ Duas partições de uma mesma tabela podem ter índices sobre atributos diferentes.  
\_f\_ Os índices de árvore B+ e de função de dispersão são apropriados para pesquisas por intervalo.  
\_f\_ A compressão de linha é mais eficaz quando há prefixos iguais nos atributos de um registo. \_v\_ O resultado de uma interrogação pode ser totalmente obtido por via de uma vista materializada.  
\_v\_ Num disco rígido, o atraso de rotação ocupa a maior fatia do tempo de pesquisa de dados.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre partições de dados e níveis RAID.**  
\_v\_ Uma tabela muito grande pode ser decomposta em partições, cada uma no seu próprio disco.  
\_v\_ Duas estratégias de partição de dados são por intervalo e por lista de valores da partition key.  
\_f\_ Na arquitetura shared memory, os nós do servidor paralelo não partilham o mesmo disco. \_f\_ O nível RAID 0 oferece menor desempenho e maior confiança face aos outros níveis.  
\_v\_ Independentemente dos níveis RAID, o utilizador vê apenas um disco lógico.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre o desenho físico de data warehouses.**  
\_f\_ Deve ser realizado antes da modelação dimensional.  
\_f\_ Procura minimizar a utilização de recursos do sistema informático.  
\_v\_ Duas das técnicas são as vistas materializadas e as bifurcações de tabelas.  
\_v\_ Alterações ao desenho físico obrigam a alterações nas aplicações dos decisores.  
\_\_ A sua importância aumenta com a quantidade de dados que precisam de ser processados.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre compressão de dados.**  
\_v\_ A compressão de página costuma poupar mais espaço em disco que a de registo/linha.  
\_f\_ As aplicações dos utilizadores não precisam de saber se os dados estão ou não comprimidos.  
\_v\_ O dicionário de página guarda prefixos e valores comuns a várias páginas da mesma tabela.  
\_v\_ A compressão é especialmente indicada quando os dados têm muitos valores repetidos.  
\_f\_ Requer a descompressão dos dados lidos do disco, tornando mais lentos os relatórios.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre vistas materializadas.**  
\_v\_ Cada vista tem um SELECT associado, cujo resultado fica guardado em disco.  
\_v\_ Evitam o cálculo repetitivo de operações muito requisitadas, tais como somas e médias.  
\_f\_ As aplicações dos utilizadores precisam de saber os nomes das vistas para as poderem usar.  
\_f\_ No comando CREATE MATERIALIZED VIEW, o SELECT não pode incluir GROUP BY.  
\_v\_ Quando há atualização de dados, as vistas também são automaticamente atualizadas.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre índices.**  
\_f\_ Num índice multi-atributo a ordem dos atributos a indexar (ex. X,Y ou Y,X) é irrelevante.  
\_v\_ Um índice bitmap tem tantas colunas quantos os valores distintos do atributo indexado.  
\_v\_ Num índice agrupado os registos de dados na tabela seguem a ordem das entradas no índice.  
\_v\_ Os índices de dispersão são mais eficientes que os de árvore B+ em pesquisas por igualdade.  
\_f\_ Num índice de dispersão cada bucket guarda entradas com valores indexados consecutivos.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre o ciclo virtuoso da prospeção de dados.**  
\_\_ A identificação do problema de negócio é feita usando os dados dos sistemas operacionais. \_\_ A decisão de alterar o negócio tem por base os modelos gerados pela prospeção de dados. \_\_ O data warehouse só participa na etapa da geração dos modelos dos dados.  
\_\_ A medição dos resultados da decisão é feita consultando os sistemas operacionais.  
\_\_ Pode ser necessário transformar os dados para poderem ser usados em métodos de prospeção.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre métodos de agrupamento.**

\_\_ É necessário, mas não suficiente, um bom cluster ter vários elementos.  
\_\_ O método k-means gera clusters a partir de pares com os subclusters mais próximos entre si.  
\_\_ São supervisionados, pois o decisor fornece as caraterísticas que deseja encontrar nos dados.  
\_\_ No método hierárquico aglomerativo existem inicialmente tantos clusters quantos os dados.  
\_\_ O dendrograma pode ser usado para mostrar os clusters gerados pelo método k-means.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre árvores de decisão.**

\_v\_ As folhas são decisões possíveis.  
\_f\_ Os atributos que melhor diferenciam os dados ficam junto das folhas.  
\_v\_ O atributo colocado no topo da árvore é o que tiver maior ganho de informação.  
\_v\_ A entropia de um conjunto de dados é zero quando todos os dados são iguais.  
\_f\_ As árvores maiores costumam ser preferíveis, pois estão menos sujeitas a sobreajustamento.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre métodos de prospeção de dados.**  
\_\_ Os métodos de agrupamento servem para descobrir conjuntos frequentes de itens.  
\_\_ No método k-means são considerados k clusters iniciais de forma aleatória.  
\_\_ Se um conjunto de itens for frequente então os seus subconjuntos também são frequentes. \_\_ As redes neuronais agrupam dados com caraterísticas idênticas desconhecidas à partida.  
\_\_ O atributo no topo de uma árvore de decisão é o que tem maior ganho de informação.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre prospeção de dados.**  
\_\_ Os conjuntos de treino e de teste são essenciais com métodos não supervisionados.  
\_\_ Para aceitar um bom cluster é suficiente verificar se inclui vários pontos de dados.  
\_\_ O atributo no topo de uma árvore de decisão é o que melhor diferencia os dados.  
\_\_ Uma regra de associação com confiança elevada tem necessariamente suporte elevado.  
\_\_ As redes neuronais aprendem lentamente e são rápidas a classificar.

**Indique afirmações verdadeiras (V) ou falsas (F) sobre prospeção de dados.**  
\_\_ O sobre-ajustamento (overfitting) causa más classificações nos dados de teste/avaliação.  
\_\_ Um conjunto de treino deve incluir dados representativos de todas as classes existentes.  
\_\_ O uso de vários métodos aumenta a confiança e dispensa a comparação com a realidade.  
\_\_ Se existir causalidade entre duas variáveis então estão correlacionadas, e vice-versa.  
\_\_ Os outliers podem ser causados por ruído ou más medições dos dados.

**Os data warehouses disponibilizam mais dados, melhores dados, e melhores ferramentas para o apoio à decisão. Indique qual o tipo de sistema informático que está a servir de base de comparação e apresente uma justificação para cada um dos três aspetos referidos.**

O sistema informático que está a servir de base de comparação são os sistemas OLTP.

Os data warehouses utilizam sistemas OLAP, disponibilizando assim mais e melhores dados e melhores ferramentas de apoio a decisão, isto porque, existe a extração periódica de dados de sistemas operacionais, a consolidação de várias fontes de dados num único repositório, dados referentes a longos períodos de tempo (com histórico), processos específicos para transformação de dados, a limpeza de dados preenche os valores ausentes e elimina duplicados, a modelagem de dados padroniza nomes e unidades de medida, visões multidimensionais e inteligíveis dos dados, análise exploratória em vários níveis de detalhe e aplicação de métodos de mineração de dados.

**Um projeto típico de construção de um *data warehouse* contempla três vertentes principais logo a seguir ao planeamento e definição de requisitos de negócio, e antes da implantação e exploração na organização. Descreva essas vertentes, justifique se podem ser realizadas autonomamente umas das outras, e indique o propósito de uma das atividades de cada vertente.**

As três vertentes principais de um data warehouse a seguir ao planeamento e definição de requisitos de negócio são as tecnologias, dados e aplicação.

A vertente das tecnologias engloba o projeto de arquitetura (normalmente baseado em plataformas existentes) e a seleção e instalação de software.

Na vertente dos dados define-se a modelagem dimensional (processos, fatos, dimensões, medidas, etc), o projeto físico (acesso ao disco rígido, distribuição e partição de dados, índices, compactação de dados, etc) e o projeto e desenvolvimento de processos ETL (extração, transformação e carregamento de dados para o data warehouse). Por último, na aplicação/utilização, surge o design das aplicações analíticas (painéis, relatórios dinâmicos, etc) e o desenvolvimento de aplicações analíticas.

As três vertentes principais de um data warehouse, para uma definição clara e planeada do projeto, podem ser realizadas autonomamente entre si, contribuindo para um resultado final comum. Com o desenrolar do processo de negócio (crescimento e manutenção) podem existir alterações posteriores nas três vertentes.

**Descreva o propósito da gestão de dados mestre numa organização, indique de que forma este serviço deve ser usado pelos sistemas operacionais, e justifique se facilita o desenvolvimento de *data warehouses***

Pode-se considerar que os dados mestre estão até acima do Data Warehouse, é o “repositório da verdade” e crítico para os sistemas OLAP.

Cada instância de um tipo de dados ou mesmo de cada dado, de cada linha da tabela, quando esses dados entram no Master Data estão completamente standardizados e validados (acentos, abreviaturas, etc). Os dados mestre geram dados críticos comuns aos processos de negócios, focado na qualidade dos dados da organização.

A relação dos dados mestre com sistemas operacionais ocorre na medida em que estes respondem a solicitações de dados com valores corretos e coerentes (standardizados) e aceita novos pedidos de registo de dados para validação

Em relação ao desenvolvimento de data warehouses, a gestão de dados mestre, simplifica a limpeza e integração de dados de várias fontes e facilita o carregamento de dados em dimensões conformadas mestre. Desta forma é esperada uma redução de custos devido a dados inconsistentes e redundantes, maior agilidade para iniciar projetos que dependem de dados confiáveis e melhor conformidade com diretrizes e normas.

**Justifique qual a ordem de desenvolvimento e operacionalização de sistemas numa cadeia de lojas de roupa: se primeiro um sistema OLAP e depois um OLTP, ou se o contrário. Descreva também duas operações típicas em cada tipo de sistema e o respetivo público-alvo.**

Um sistema OLTP deve surgir primeiro que um sistema OLAP, caso essa cadeia consiga fazer toda a análise com sistemas OLTP, ou seja, esteja mais orientada para transações, com poucos dados, sem preocupação com o histórico e atualizações continuas tendo em consideração apenas fontes de dados internas. Caso essa cadeia pretenda iniciar, por exemplo, uma expansão para outras cidades ou países ou criar um site na web, deve optar pelo desenvolvimento de um sistema OLAP, para fazer análises de padrões, dados agregados e relatórios analíticos, com histórico de fontes internas e externas. Sistemas OLTP dizem respeito a processos orientados para o mercado e sistemas OLAP a processos orientados para o cliente. Operações OLTP: INSERT e UPDATE; Operações OLAP: SELECT e GROUP BY

**Justifique a existência de um sistema OLTP numa cadeia de lojas de roupa e descreva duas operações que nele possam ser executadas recorrentemente. Indique também uma razão para o desenvolvimento de um sistema OLAP na mesma organização, tendo o cuidado de mencionar os públicos-alvo dos dois tipos de sistema.**

Um sistema OLTP pode justificar-se numa cadeia de loja de roupa, caso essa cadeia consiga fazer toda a análise com sistemas OLTP, ou seja, esteja mais orientada para transações, com poucos dados, sem preocupação com o histórico e atualizações continuas tendo em consideração apenas fontes de dados internas. Caso essa cadeia pretenda iniciar, por exemplo, uma expansão para outras cidades ou países ou criar um site na web, deve optar pelo desenvolvimento de um sistemas OLAP, para fazer análises de padrões, dados agregados e relatórios analíticos, com histórico de fontes internas e externas.

**Os dados mais recentes dos sistemas OLAP costumam ter um atraso temporal face aos dos sistemas OLTP. Apresente uma razão para esse desfasamento, e justifique se este é ou não aceitável para os decisores, relacionando com a frequência de atualização dos dados OLAP.**

Os dados mais recentes dos sistemas OLAP costumam ter um atraso temporal face aos dos sistemas OLTP porque nos sistemas OLAP a atualização é planeada e em sistemas OLTP é atualização contínua. Apesar deste delay de informação, o desfasamento é aceitável para os decision makers porque apesar do desfasamento, os sistemas OLAP permitem analisar padrões, impactos de descontos e saldos, o histórico, dados agregados, entre outras vantagens.

**Descreva as abordagens *top-down* e *bottom-up* de construção de *data warehouses*, indicando o foco inicial dos trabalhos e a forma de expansão para satisfazer necessidades futuras.**

**A abordagem Top-Down (de Inmon)** possui uma visão geral que vai do alto nível ao mais detalhado. Em primeiro lugar, detalha o plano completo do data warehouse, cria/transforma dimensões simplificadas em dimensões muito especializadas e os data marts são construídos por solicitação, provenientes do principal data warehouse. Esta abordagem demonstra-se complexa, já que a visão geral possui grande abrangência, onde requer a construção completa do data warehouse para a disponibilização posterior dos data marts. Top-Down é assim apenas indicado quando todo o processo é facilmente compreendido e simples de ser implementado.

Já a **abordagem Bottom-Up (de Kimball)** vai do menor ao maior nível de detalhe e visa “dividir para conquistar”. Torna o primeiro data mart para o processo o mais importante, sendo que o conjunto de todos os datar marts em construção formam o data warehouse. A matriz de negócio funciona como uma linha de orientação para integração de dados e partilha de dimensões.

A abordagem **Bottom-Up** é, tipicamente, a abordagem com melhores resultados, já que o projeto tem o desenvolvimento evolutivo e possui menores riscos.

**Antes do surgimento, nos anos 70, dos *executive information systems* (EIS), as organizações tinham ao seu dispor *management information systems* (MIS). Indique o tipo de relatórios que eram gerados pelos MIS e apresente duas razões pelas quais os executivos sentiram necessidade dos EIS.**

Os executive information systems (MIS) são sistemas de informações usados para a tomadas de decisão e para a coordenação, controle, análise e visualização de informações numa organização. Esse estudo envolve, obviamente, pessoas, processos e tecnologia num contexto organizacional. Num ambiente corporativo, como tipicamente acontece, o objetivo final do uso de um sistema de informações é aumentar o valor e os lucros dos negócios. Os MIS são geradores periódicos de relatórios pré-definidos. A partir dos anos 70, passou a existir uma análise de tarefas específicas de decisão, usados por gestores, mas não por executivos. Os Executivos descontentes com a pouca flexibilidade dos relatórios e pouca relevância dos dados fornecidos pelos sistemas de informação e com o aparecimento dos executive information system (EIS) foram fazendo gradualmente a transição. O EIS é um tipo de sistema de suporte à gestão que facilita e dá suporte às informações de executivos superiores e às necessidades das tomadas de decisão, para além do fácil acesso a informações internas e externas relevantes para os objetivos organizacionais. O EIS enfatiza telas gráficas e interfaces fáceis de usar, perceber e analisar (mesmo pelos executivos), uma forte capacidade de gerar relatórios detalhados importantes nos processos de negócio. Através de EIS é possível comparar e destacar tendências, para delinear o desempenho e identificar oportunidades e problemas.

**Indique duas razões que levaram ao surgimento de *executive information systems* (EIS) dirigidos aos executivos de organizações, em complemento dos tradicionais sistemas OLTP. Diga também por que razão continuam a existir sistemas OLTP nas organizações.**

Os executive information systems (EIS) ganharam terreno face aos tradicionais sistemas OLTP pois facilitam e dão suporte às informações de executivos superiores e às necessidades das tomadas de decisão, para além do fácil acesso a informações internas e externas relevantes para os objetivos organizacionais. O EIS enfatiza telas gráficas e interfaces fáceis de usar, perceber e analisar (mesmo pelos executivos), uma forte capacidade de gerar relatórios e detalhes importantes nos processos de negócio. Através de EIS é possível comparar e destacar tendências, para delinear o desempenho e identificar oportunidades e problemas. Pese embora tudo isto, continuam a existir sistemas OLTP nas organizações pois o custo computacional é muito menor e caso essa organização esteja mais orientada para transações, com poucos dados, sem preocupação com o histórico e atualizações continuas tendo em consideração apenas fontes de dados internas.

**Descreva duas razões para o surgimento de sistemas OLAP em organizações que já dispunham de sistemas OLTP, justificando, em particular, que público-alvo mais sentiu a necessidade de mudança.**

Caso as organizações consigam fazer toda a análise com sistemas OLTP, ou seja, estejam mais orientadas para transações, com poucos dados, sem preocupação com o histórico e atualizações continuas tendo em consideração apenas fontes de dados internas não existe a necessidades de implementar sistemas OLAP. Caso essas organizações pretendam iniciar, por exemplo, uma expansão para outras cidades ou países ou criar um site na web, devem optar pelo desenvolvimento de um sistema OLAP, para fazer análises de padrões, dados agregados e relatórios analíticos, com histórico de fontes internas e externas. Sistemas OLAP facilitam a tomada de decisão dos executivos porque os processos são orientados para o cliente.

**Compare os sistemas OLTP e OLAP segundo os critérios seguintes: a) número e origem de fontes de dados usadas nos relatórios; e b) frequência e momento de atualização dos dados.**

1. OLTP: Número de fontes reduzido, sendo essas fontes de dados internas e sem histórico;  
   OLAP: Número de fontes de dados alto, sendo essas fontes de dados internas e externas, com histórico
2. OLTP: A atualização dos dados é contínua, com uma alta disponibilidade e processamento repetitivo;  
   OLAP: A atualização dos dados é, tipicamente, planeada, com uma disponibilidade relaxada e processamento exploratório.

**Enquadre as etapas do processo de *extraction*, *transformation*, e *loading* (ETL) relativamente à *data staging area*, à *data presentation area*, e aos sistemas operacionais, e indique uma responsabilidade específica de cada etapa. Notas: pode fazer um esboço anotado se achar mais prático e não se limite a traduzir os termos para português.**

O processo ETL costuma ser um processo complexo, moroso e transversal a toda a construção do data warehouse, pois envolve a extração, transformação e carregamento dos dados. Quando se extraem os dados, estes passam para uma data staging area, que se situa entre os sistemas operacionais e a data presentation data e onde são guardados os dados extraídos dos sistemas operacionais. Na data staging area os dados são limpos, combinados e standardizados. Após a passagem por essa “área de estágio” os dados são transformados e carregados para uma data presentation area, onde os dados prontos a usar são carregados, incluindo já índices e agregados pré-calculados, para o data warehouse e posteriormente para os data marts necessários ao processo de negócio.

**Justifique se interrogações analíticas realizadas sobre uma base de dados federada afetam ou não o desempenho das bases de dados operacionais da organização, bem como se é ou não simplificada a conjugação de dados históricos nessas análises.**

As interrogações analíticas realizadas sobre uma base de dados federada colocam problemas aos sistemas operacionais, afetando assim o desempenho das bases de dados operacionais da organização. É dificil fazer otimização de interrogações numa base de dados federada bem como pode ser difícil obter informação histórica nessas análises.

**Descreva o papel dos sistemas OLTP e OLAP no contexto de uma organização, identifique os públicos-alvo respetivos, e dê um exemplo de operação típica em cada tipo de sistema.**

Caso as organizações consigam fazer toda a análise com sistemas OLTP, ou seja, estejam mais orientadas para transações, com poucos dados, sem preocupação com o histórico e atualizações continuas tendo em consideração apenas fontes de dados internas não existe a necessidades de implementar sistemas OLAP. Caso essas organizações pretendam iniciar, por exemplo, uma expansão para outras cidades ou países ou criar um site na web, devem optar pelo desenvolvimento de um sistema OLAP, para fazer análises de padrões, dados agregados e relatórios analíticos, com histórico de fontes internas e externas. Sistemas OLTP dizem respeito a processos orientados para o mercado e sistemas OLAP a processos orientados para o cliente. Operações OLTP: INSERTS; Operações OLAP: SELECTS.

**Compare os sistemas OLTP e OLAP segundo os três critérios seguintes: a) estrutura dos relatórios; b) frequência de atualização dos dados; e c) abrangência temporal dos dados.**

1. OLTP: Relatórios pré definidos;  
   OLAP: Relatórios personalizados.
2. OLTP: A atualização dos dados é contínua;  
   OLAP: A atualização dos dados é, tipicamente, planeada.
3. OLTP: Reflete apenas o estado atual dos dados;  
   OLAP: Para além da atualidade dos dados, vem acompanhado do histórico

**Compare os *data warehouses* e as bases de dados federadas segundo os três critérios seguintes: a) necessidade de cópia dos dados dos sistemas operacionais; b) impacto das interrogações analíticas nos sistemas operacionais; e c) possibilidade de obtenção de dados históricos.**

1. Data Warehouses: Cria uma cópia dos dados operacionais e as interrogações analíticas são feitas sobre a cópia;  
   Bases de Dados Federadas: As Interrogações analíticas são feitas sobre dados dos sistemas operacionais.
2. Data Warehouses: Para além de serem feitas sobre a cópia, as interrogações analíticas nos sistemas operacionais não têm grande impacto, apesar de continuarem a existir dados redundantes  
   Bases de Dados Federadas: Apesar de não existirem dados redundantes, as interrogações analíticas colocam problemas aos sistemas operacionais.
3. Data Warehouses: Obtenção de dados históricos, tipicamente, fáceis de obter;  
   Bases de Dados Federadas: Pode ser difícil obter informação histórica.

**Os *data warehouses* consolidam dados de múltiplas fontes num só repositório. Em alternativa também podem ser usadas bases de dados federadas para responder a interrogações analíticas. Descreva uma desvantagem desta última opção face à primeira.**

Uma das desvantagens das bases de dados federadas face aos data warehouses reside no facto de puder ser difícil obter informações históricas ficando as consultas restritas a dados atuais.

**Explique qual o âmbito típico de um *data mart* dentro do contexto organizacional e apresente duas razões que levam estes últimos a serem usados em vez de os utilizadores acederem diretamente ao *data warehouse*.**

Um data mart é definido como um conjunto flexível de dados, idealmente com base nos dados mais atómicos (granulares) possíveis de extrair de uma fonte operacional e apresentado num modelo simétrico (dimensional) que é mais resiliente quando confrontado com consultas inesperadas do user. Um data mart vai permitir imensas querys. É uma organização dos dados para permitir análise de determinados processos de negócio e visualizações no data warehouse sobre processos de negócios muito específicos. Um data mart ocupa um menor espaço, é mais focado e mais eficaz do que aceder diretamente ao data warehouse.

**No processo de ETL, justifique que tipo de técnica de mudança lenta é mais difícil de gerir, ilustrando os passos a seguir.**

**No contexto de um *data warehouse* defina sucintamente os seguintes conceitos: a) *data mart*; b) *data staging area*; e c) *drill across*.**

Um data mart é definido como um conjunto flexível de dados, idealmente com base nos dados mais atómicos (granulares) possíveis de extrair de uma fonte operacional e apresentado num modelo simétrico (dimensional) que é mais resiliente quando confrontado com consultas inesperadas do user. Um data mart vai permitir imensas querys. É uma organização dos dados para permitir análise de determinados processos de negócio e visualizações no data warehouse sobre processos de negócios muito específicos. Um data mart ocupa um menor espaço, é mais focado e mais eficaz.

A data staging area é uma área de armazenamento intermediária usada para processamento de dados durante o processo de extração, transformação e carregamento (ETL). A área de armazenamento temporário de dados está entre o sistema de origem (fontes) e o servidor de apresentação de dados, que geralmente são data warehouses, data marts ou outros repositórios de dados.

Drill across significa fazer consultas separadas em duas ou mais tabelas de fatos num único relatório, onde os cabeçalhos de linha de cada consulta consistem em atributos conformes idênticos. Os conjuntos de respostas das duas consultas, tipicamente feitas de forma separada, são alinhadas executando uma operação de classificação e junção nos cabeçalhos de linha do atributo de dimensão comum.

**Relatórios Analíticos**

**No contexto da HyperVending, indique que totais e subtotais seriam calculados para a interrogação SELECT ano, sucursal, cliente, SUM(euros) FROM vendas WHERE (ano BETWEEN 2010 AND 2016) GROUP BY ROLLUP (ano, sucursal, cliente).**

Esta interrogação calcularia:  
- Soma (Euros) de vendas por sucursal por cada ano entre 2010 e 2016;  
- Soma (Euros) de vendas por clientes por cada ano entre 2010 e 2016;  
- Soma (Euros) total de vendas por sucursal entre 2010 e 2016;  
- Soma (Euros) total de vendas por clientes entra 2010 e 2016;  
- Soma (Euros) total de vendas de todas as sucursais entre 2010 e 2016

**Suponha um relatório com as vendas em euros por sucursal ao longo dos anos, inspirado no caso HyperVending. Identifique uma hierarquia de atributos numa das dimensões e descreva um cenário de *drill-down*, apresentando uma razão plausível para tal operação.**

Uma hierarquia possível seria: ano – semestre – trimestre – mês. Aplicando um drill-down, ou seja, aumentando o nível de detalhe, poderia ser feita uma consulta a partir do 2º trimestre a qual o mês com mais vendas em cada sucursal. Razões plausíveis para essa consulta seria, por exemplo, ver o efeito de uma promoção em determinada sucursal num mês do 2º trimestre.

**Indique duas razões para a inclusão de extensões OLAP na linguagem SQL no contexto dos relatórios analíticos e descreva quais os totais e subtotais calculados na resposta à interrogação SELECT Cliente, Ano, SUM(Unidades) FROM Vendas GROUP BY ROLLUP (Cliente, Ano).**

Esta interrogação calcularia:  
- Soma (unidades) de vendas por cada cliente em cada ano;  
- Soma (unidades) total de vendas de todos os clientes em todos os anos  
- Soma (unidades) total de vendas de unidades de todos os clientes

**Suponha um relatório dinâmico sobre audiências televisivas, envolvendo programas e espetadores. Identifique uma hierarquia de atributos numa das dimensões e descreva um cenário de realização da operação *roll-up* com uma medida numérica plausível.**

**No contexto da HyperVending, mostre exemplos de linhas resultantes da interrogação SELECT País, Ano, SUM(Euros) FROM Vendas GROUP BY ROLLUP (País, Ano) e indique que linhas adicionais seriam mostradas caso tivesse sido usado GROUP BY CUBE (País, Ano).**

Esta interrogação calcularia:  
- Soma (euros) de vendas por pais por cada ano;  
- Soma (euros) de vendas total por países;  
- Soma (euros) de vendas total de todos os países.

Com a GROUP BY CUBE adiciona:  
- Soma (euros) de vendas total por ano em todos os países

**Suponha um relatório dinâmico da empresa HyperVending abrangendo sucursais e produtos. Identifique uma possível hierarquia de atributos e descreva um cenário de realização da operação *roll-up*, para o qual deverá também identificar uma medida numérica plausível.**

Uma possível hierarquia tendo em conta o cenário descrito poderia ser: Europa – União Europeia - Península Ibérica – Portugal – Cidades

Uma operação roll-up, ou seja diminuir o nível de detalhe, poderia ser por exemplo, passar do total de vendas por cidade (Lisboa, Porto, Coimbra) para vendas por país (Portugal).

**Considere a interrogação SELECT País, Ano, SUM(Euros) FROM Vendas GROUP BY CUBE (País, Ano) do caso HyperVending. Descreva o resultado que seria obtido.**

Esta interrogação calcularia:  
- Soma (euros) de vendas por pais por cada ano;  
- Soma (euros) de vendas total por países;  
- Soma (euros) de vendas total de todos os países;  
- Soma (euros) de vendas total por ano em todos os países.

**Suponha um relatório dinâmico que atualmente mostra as vendas em euros por sucursal ao longo dos anos, inspirado no caso HyperVending. Descreva o conteúdo do relatório se fizesse *drill-down* em 2010, bem como a sua assunção sobre a hierarquia de atributos subjacente.**

Através do caso descrito, fazendo um drill-down nos dados de 2010, ou seja aumentando o detalhe, seria dado o acesso às vendas em euros por sucursal em cada semestre, trimestre e até mesmo mês ou dia se fosse necessário. Esta hierarquia permitiria ver a variação mais detalhada das vendas em cada sucursal, bem como a influência dos saldos ou datas importantes como o Natal, Dia do Pai ou Dia da Mãe.

**Considere a seguinte tabela de folha de cálculo, criada para armazenar dados de audiências televisivas: Audiências, NomeDoEspetador, FaixaEtária, ClasseSocEconómica, DimAgregadoFamiliar, Data, HoraDeInício, HoraDeFim, MinutosVistos, NomeDoPrograma, TipoDePrograma, Canal. Nota: HoraDeInício e HoraDeFim referem-se respetivamente ao momento no tempo em que um determinado espetador sintonizou um canal e o deixou.**

**Usando as extensões OLAP do SQL, refira como faria para construir uma *pivot table* que descrevesse o número de espetadores que viram o programa semanal Batatoon distribuídos por faixa etária, nas semanas de Junho de 2007.**

SELECT Data, FaixaEtária, DimAgregadoFamiliar, NomeDoPrograma  
FROM Audiências  
WHERE (MONTH = 6 AND YEAR = 2007) AND (NomeDoPrograma = “Batatoon”)  
GROUP BY (FaixaEtária, DimAgregadoFamiliar, NomeDoPrograma).

**Modelação Dimensional**

**Compare as bifurcações e mini-dimensões em termos de: a) frequência de atualização dos atributos; b) relação entre os atributos; e c) modo de juntar os dados com os da dimensão “monstra” correspondente.**

1. Bifurcações: Os atributos devem mudar lentamente

Mini-dimensões: Os atributos devem mudar com frequência

1. Bifurcações: Os atributos estão relacionados entre si

Mini-dimensões: Os atributos não necessitam de estar relacionados entre si

1. Bifurcações: Snowflaking de uma dimensão “monstra” que transfere atributos que mudam pouco para outra tabela;

Mini-dimensões: Guarda as combinações possíveis de valores dos atributos através de tabela de factos com chave estrangeira adicional para mini-dimensão.

**Compare os três tipos de tabelas de factos em função do grão e dê um exemplo plausível de cada tipo.**

**Tabela de factos de transações**: Cada evento é armazenado na tabela de fatos apenas uma vez, possui uma coluna de data indicando quando o evento ocorreu e uma coluna que identifica cada evento. O número de linhas é igual ao da tabela de origem. Uma linha numa tabela de fatos da transação corresponde a um evento no espaço e no tempo

Um exemplo poderá ser quando existe uma tabela de vendas com pedidos de clientes. Grão: Uma linha por transação.

**Tabela de factos instantâneos periódicos:** Uma linha numa tabela de factos de snapshot periódico resume muitos eventos de medição que ocorrem durante um período padrão, como um dia, uma semana ou um mês. Todo o sistema de origem é copiado na tabela de fatos regularmente, sendo que o mesmo evento pode ser armazenado várias vezes. Possui uma coluna de data do snapshot, indicando quando uma cópia da tabela de origem foi criada. O grão é o período, não a transação individual.

Um exemplo deste tipo de tabela de factos poderá ser o saldo da conta bancária. No final de cada dia, os saldos de cada conta de cliente no banco são armazenados nesta tabela de saldo da conta. Grão: uma linha por saldo bancário por mês (período).

**Tabela de factos instantâneos acumulativo**: Resume os eventos que ocorrem em etapas previsíveis entre o início e o final de um processo. Há uma chave estrangeira da dimensão data na tabela de fatos para cada evento crítico no processo. Além das datas de chaves estrangeiras associadas a cada etapa crítica do processo, as tabelas de fatos de snapshot acumulativo acumulativas contêm chaves estrangeiras para outras dimensões e, opcionalmente, dimensões degeneradas. A atualização das linhas de fatos é exclusiva entre os três tipos de tabelas de fatos.

Exemplo: Numa compra numa loja online, o cliente é de Lisboa e o artigo estava num armazém no Porto. Terá de existir a data do pedido, a data de saída do armazém do Porto, a data de chegada ao armazém de Lisboa, a data de saída do armazém de Lisboa e a data de entrega. Grão: uma nova linha na tabela de fatos sempre que algo nesse processo de entrega tenha mudado (ciclo de vida).

**Justifique se, numa tabela de factos de tipo transacional sobre encomendas, usaria ou não como medidas o preço unitário de cada produto e a percentagem de desconto aplicada ao preço base do produto para cada cliente que fez uma encomenda. Em caso negativo, indique que medidas usaria em alternativa.**

A aplicação de percentagens de desconto face a preços base ou preços unitários é mais típica de tabelas de factos sem factos. A percentagem não pode ser considerada como medida na transação, logo não a usaria numa tabela de factos desse tipo. As medidas que utilizaria seriam o preço unitário normal e o preço unitário com desconto.

**Explique a diferença entre medidas aditivas e semi-aditivas, dando um exemplo de cada, e justifique qual dos dois tipos é preferível em *data warehouses*. Nota: tenha especialmente em conta o ponto de vista do decisor que está a compor relatórios.**

As medidas aditivas permitem a soma de valores ao longo de todas as dimensões, sendo esta uma das operações mais úteis num data warehouse, enquanto que as medidas semi aditivas dizem respeito a factos de natureza periódica e onde as medidas são aditivas mas não em todas as dimensões. Normalmente as medidas que surgem na tabela de factos, do ponto de vista do decisor, são medidas somáveis, logo as medidas aditivas são mais utilizadas, embora não seja de todo de descartar a utilização de medidas semi-aditivas até porque conseguem também fazer somas, mas não em todas as dimensões.

**Apresente um exemplo de duas hierarquias para atributos muito relacionados entre si e pertencentes a uma mesma dimensão, e justifique se pode ou não ser útil um decisor que está a compor relatórios ter acesso a qualquer uma dessas hierarquias (ou, no caso geral, a hierarquias em condições semelhantes).**

Tipicamente a dimensão data apresenta atributos muito relacionados entre si, por exemplo: Ano-Semestre-Trimestre-Mês-Dia e Ano fiscal-Semestre fiscal-Trimestre fiscal-Dia.

Para o decisor pode ser importante ter acesso a todas as hierarquias e até efetuar algumas operações ao cubo de dados em condições semelhantes, pois estas hierarquias dizem respeito, no fundo, ao mesmo (datas), o que se altera são os valores que compõem os atributos. Por exemplo, em 2020 as receitas fiscais baixaram porque no 2º trimestre o país entrou em estado de emergência devido à COVID-19.

**Para o facto seguinte, indique as medidas numéricas, caso existam, o tipo de tabela de factos, as dimensões, e alguns atributos representativos, incluindo chaves substitutas e estrangeiras: a equipa do país P1 treinada por T1 jogou com a equipa do país P2 treinada por T2 no dia D, no estádio E com capacidade para 55000 espetadores que fica na cidade C do país P3, tendo o resultado final sido 4 a 0.**

**Para o facto seguinte, indique as medidas numéricas, caso existam, o tipo de tabela de factos, as dimensões, e alguns atributos representativos: o avião V com P pessoas a bordo, das quais T pertencentes à tripulação, partiu do aeroporto A1 e tinha destino previsto para o aeroporto A2, mas desapareceu dos radares no dia D1, tendo sido descoberto no local L, no dia D2. Nota: pode desenhar um diagrama de dados.**

**Justifique se a existência da gestão de dados mestre numa organização facilita ou não a manutenção de dimensões conformadas.**

Entende-se por dimensões conformadas as dimensões que partilham atributos. Já a gestão de dados mestre é um serviço que gere dados críticos comuns aos processos de negócio. As características de ambos estão interligadas, sendo que a gestão de dados mestre simplifica a limpeza e integração de dados de várias fontes e facilita o carregamento de dados nas dimensões conformadas.

**Considerando os factos seguintes, indique: as medidas numéricas, caso existam; o tipo de cada tabela de factos; as dimensões e alguns atributos representativos; e se as dimensões podem participar em ambas as tabelas de factos. Nota: pode desenhar diagramas de dados.**

**a) 48 alunos realizaram o teste de IPAI no dia 12 de abril de 2011, tendo as notas sido divulgadas em 3 de maio de 2011, e havendo obtido aprovação 37 alunos.   
b) O aluno Pedro, número 43210, fez o teste de IPAI de 17 de abril de 2012, ocorrido no anfiteatro 1.3.20, com 74 lugares, com vigilância do professor António, funcionário 1234.**

**Apresente uma vantagem e uma desvantagem das técnicas de tipo 1 e 2 para registo de mudanças lentas numa dimensão, e descreva um cenário adequado a cada técnica.**

Vantagens  
Tipo 1: Fácil e simples de perceber  
Tipo 2: Histórico permite perceber o impacto das decisões

Desvantagens  
Tipo 1: Não considera o histórico   
Tipo 2: Acelera o crescimento da tabela de dimensões

Cenários  
Tipo 1: Mudanças de produto no departamento, apenas salva o departamento para onde foi mudado.  
Tipo 2: Se o produto mudar de departamento, o tipo 2 adiciona colunas novas para guardar as alterações bem como as datas das alterações.

**Justifique se a matriz de processos *(bus matrix)* deve ser definida antes ou depois da modelação detalhada das dimensões, e indique o papel da matriz na conformação de dimensões.**

Aquando da implementação da bus matrix, existe já a ideia das dimensões comuns mas não estão definidas as modelações detalhadas das mesmas, porque a bus matrix cruza dimensões dos dados com os processos de negócio. Já as dimensões são conformadas quando partilham atributos, desta forma a bus matrix facilita a visualização das dimensões conformadas.

**Considerando os factos seguintes, indique os nomes e atributos das dimensões envolvidas bem como das medidas numéricas (caso existam), e mencione o tipo de tabela de factos.**

**a) 113 pessoas candidataram-se em 2011 a cursos da FCUL através do programa Maiores de 23, tendo 12 candidatos obtido aprovação na prova realizada a 7 de Maio cujos resultados foram publicados a 31 de Maio.   
b) A fatura mensal de eletricidade do cliente C no mês M foi de E euros, dos quais T resultam do pagamento de uma taxa de audiovisual.**

**Considere a dimensão Cliente, com milhões de linhas, na qual está a ser usada a técnica 2 para guardar o histórico de mudanças que vão ocorrendo. De entre os atributos da dimensão, existe um conjunto dedicado à demografia da região do cliente, o qual é atualizado no início de cada década.**

**a) Indique como seria registada uma atualização do número de filhos de um cliente, incluindo o que seria guardado nos atributos demográficos relativamente ao estado anterior.   
b) Explique como poderia ser controlado o crescimento da quantidade de dados na dimensão Cliente, particularmente no que diz respeito ao conjunto de atributos demográficos. Nota: pode fazer um diagrama, se achar adequado.**

**Considerando os factos seguintes, indique os nomes e eventuais atributos das dimensões envolvidas bem como das medidas numéricas (caso existam), e mencione também o tipo de tabela de factos.**

**a) 17 partidos concorreram às eleições legislativas antecipadas de 5 de junho de 2011 em Portugal, dos quais 5 elegeram deputados.   
b) Nas legislativas de 5 de junho de 2011 o partido P elegeu o deputado N pelo distrito D.**

NomePartido, NumeroVotos, Distrito, Concelho, Freguesia, MesaDeVoto,

**Apresente um exemplo de *role playing* de dimensão e indique uma forma prática de o realizar, usando conceitos de SQL, com o mínimo de recursos e de forma transparente para o utilizador.**

O role-playing dá-se quando por exemplo numa dimensão data existem vários eventos com a mesma data, embora cada data tenha um significado diferente em cada dimensão/evento. Neste exemplo, as dimensões data são simuladas com vistas ou sinónimos de SQL.

**Quanto mais fino o grão dos factos mais registos tendem a existir na tabela de factos. Por exemplo, basta um registo para guardar o total de uma fatura, mas são precisos vários para guardar os subtotais por produto. Nesta linha, justifique se a afirmação seguinte é ou não correta: “em geral, quanto mais fino o grão dos factos mais dimensões tendem a ser modeladas.”**

A afirmação presente no enunciado é verdadeira, porque um grão fino possui maior nível de detalhe e logicamente mais dimensões a serem modeladas e mais atributos na chave primária.

**Suponha uma tabela de factos de tipo transacional sobre vendas de produtos, com medidas PreçoUnitário e QuantidadeVendida. Justifique se ambas as medidas são aditivas e em caso negativo proponha uma alternativa em que tal aconteça sem haver perda de informação.**

**Descreva um exemplo de tabela de factos de tipo instantâneo periódico, mencionando se cada facto abrange períodos variáveis ou fixos de tempo e se o carregamento de novos dados envolve só inserções, só atualizações, ou ambas as operações de escrita.**

As tabelas de factos de tipo instantâneo periódico guardam o desempenho acumulado em períodos fixos e os dados novos são inseridos. Algumas medidas podem valer zero.

**Considerando os dois factos seguintes, indique os atributos e nomes das dimensões envolvidas bem como das medidas numéricas (caso existam), e mencione também o tipo de tabela de factos em causa.**

**a) 27 alunos inscreveram-se no teste de IPAI do dia 20 de Abril de 2010, tendo as notas sido publicadas em 22 de Abril de 2010, e tendo obtido aprovação 23 alunos.   
b) O aluno Pedro, número 43210, fez o teste de IPAI de 12 de Abril de 2011, ocorrido no anfiteatro 8.2.47, com 220 lugares, com vigilância do professor António, funcionário 1234.**

**Considere a dimensão “monstra” Cliente, com milhões de linhas e em rápido crescimento. Os decisores têm-se queixado da lentidão com que os relatórios são gerados, em particular os que envolvem a evolução histórica dos atributos Rendimentos e NúmeroDeFilhos.**

**a) Supondo que estes atributos são transferidos para uma mini-dimensão, descreva as adaptações que seria necessário fazer aos dados destes atributos, e faça uma figura com o novo modelo em estrela, incluindo a tabela de factos.   
b) Justifique em que circunstâncias a mini-dimensão permitiria abrandar o crescimento da dimensão “monstra” e em que medida continuaria a ser possível analisar a evolução histórica dos referidos atributos.**

**Descreva um exemplo de tabela de factos de tipo instantâneo cumulativo, tendo o cuidado de mencionar se cada facto abrange períodos variáveis ou fixos de tempo e se o carregamento de novos dados envolve só inserções, só atualizações, ou se permite ambas as operações de escrita.**

As tabelas de factos de tipo instantâneo cumulativo acompanham um processo recorrente, mas de duração variável, com etapas previsíveis, cumpridas em datas imprevisíveis. Dados novos através de inserções e atualizações de linhas.

**Compare tabelas de factos de tipo instantâneo periódico e cumulativo segundo os dois critérios seguintes: a) períodos variáveis ou fixos representados em cada facto; e b) uso ou não de inserções/atualizações aquando do carregamento de dados.**

1. Instantâneo periódico: Períodos fixos  
   Instantâneo cumulativo: Períodos variáveis
2. Instantâneo periódico: Inserções  
   Instantâneo cumulativo: Inserções e atualizações

**Descreva os quatro passos recomendados para a modelação dimensional, exemplificando com o caso do projeto da disciplina. Nota: os passos referem dimensões, grão dos factos, processos de negócio, e medidas, não necessariamente por esta ordem.**

1. Identificar processos a modelar: Matriz exequibilidade/valor salienta processos valiosos e com dados  
2. Determinar grão da tabela de factos: Nível de detalhe máximo desejado para as análises   
3. Modelar dimensões em tabelas: Dados que descrevem os processos da organização  
4. Identificar medidas numéricas na tabela de factos: Medidas que servem para avaliar o negócio

**Considere que foi feito um estudo sobre a aquisição de medicamentos em farmácias portuguesas. Para cada aquisição é registado o medicamento adquirido, o valor pago, o dia e a hora bem como informação sobre o cliente (nome, número de beneficiário, e residência). Para certos medicamentos passados com receita conhece-se ainda o médico envolvido e o valor da comparticipação. Os medicamentos têm determinados compostos ativos em quantidades diferentes dependendo da marca. Nota: a informação do cliente poderá não estar sempre disponível. Elabore o esquema para um *data warehouse* que permita recolher e analisar esta informação.**

**Bifurcações e mini-dimensões são estratégias diferentes de decomposição usadas para modelar tabelas de dimensão “monstras” (com muitas linhas e muitas colunas). Descreva em que consistem e ilustre as suas diferenças.**

Nas bifurcações, os atributos devem mudar lentamente, os atributos estão relacionados entre si e é feito um snowflaking de uma dimensão “monstra” que transfere atributos que mudam pouco para outra tabela;

Nas mini-dimensões, os atributos devem mudar com frequência, não necessitando de estar relacionados entre si. Guardam as combinações possíveis de valores dos atributos através de tabela de factos com chave estrangeira adicional para mini-dimensão.

**Elabore um esquema para um *data warehouse* que permita recolher e analisar a informação na seguinte situação, justificando as decisões tomadas.**

**Foi feito um estudo sobre a utilização de transportes públicos em várias cidades europeias. Para tal fez-se o seguimento de um conjunto de passageiros, através dos registos de entrada e saída dos seus títulos de transporte para cada viagem encetada. Conhece-se a identidade dos passageiros seguidos bem como as suas caraterísticas socioeconómicas. Está ainda disponível informação sobre os veículos, as rotas seguidas, o número de paragens em cada viagem e os condutores.**

**No contexto de um *data warehouse* referente a infrações de trânsito, a GNR necessitou da criação de uma dimensão Condutor.**

**a) Indique que atributos consideraria para a modelação dessa dimensão e justifique que tamanho teria ao fim de três anos sabendo que em média são autuados cerca de 10.000 novos condutores por mês (valores fictícios).   
b) Para o mesmo trabalho a GNR partiu de um conjunto de folhas Excel existentes em cada esquadra, que tinham a seguinte estrutura: Audiências, NomeDoCondutor, FaixaEtária, Veículo, EstadoDoVeículo, CartaDeCondução, ÁlcoolNoSangue, TipoDeInfracção, Hora, Data, Local, AgentesEmOperação, ValorDaMulta. Faça o diagrama de um *data warehouse* para este problema, expandindo, se necessário, os atributos das dimensões consideradas. Nota: pode referenciar a resposta da pergunta anterior.**

**Indique como faria para modelar uma hierarquia de profundidade fixa numa tabela de dimensão. Descreva também as diferenças caso a hierarquia fosse de profundidade variável. Nota: pode fazer uma ilustração se achar necessário.**

Numa hierarquia de profundidade fixa, todos os níveis da hierarquia têm sempre valores. Estas hierarquias são de uso generalizado e fáceis de entender. As principais diferenças residem no facto de as hierarquias de profundidade variável possuirem alguns níveis que podem não estar preenchidos e servem para relações hierárquicas mais complexas.

**Elabore um esquema para um *data warehouse* que permita recolher e analisar a informação na seguinte situação, justificando as decisões tomadas.**

**Uma empresa de sondagens faz entrevistas telefónicas recolhendo preferências de consumidores sobre refrigerantes e cervejas. Dentro das perguntas efetuadas, pergunta-se qual o grau de preferência por uma série de marcas, e até que ponto se recorda de ter visto propaganda a elas referida. A propaganda pode ser de rua, televisão, ou imprensa escrita. É recolhida também informação socioeconómica e geográfica sobre o cliente, sabendo-se ainda o momento em que a entrevista ocorreu e qual o entrevistador.**

**Indique três caraterísticas das tabelas de factos do tipo instantâneo periódico *(periodic snapshot)*, ilustrando com um exemplo.**

Guardam desempenho acumulado em períodos fixos e regulares  
Dados novos através de inserções  
Algumas medidas podem valer zero  
Ex: Saldo, comissões e juros em janeiro, fevereiro, março, …

**Indique que estratégia utilizaria para modelar uma dimensão de clientes empresariais, que contém um conjunto de atributos qualitativos que muda com muita frequência. Para o problema concreto suponha que a dimensão tem 100.000 empresas distintas, com cerca de 120 atributos, dos quais 20, sofrem alterações todas as semanas.**

**No contexto de um *data warehouse* defina sucintamente os seguintes conceitos: a) grão; b) tabelas de factos sem factos; c) dimensões degeneradas.**

a) O grão é o significado de uma linha na tabela de fatos e determina o nível máximo de detalhe;  
b) As tabelas de factos sem factos guardam factos sem medidas de negócio e servem para simples acompanhamento de eventos;  
c) As dimensões degeneradas dizem respeito a dimensões que servem apenas para agrupar factos. O agrupamento tem de ter significado forte no contexto. Não possui outros atributos próprios e não tem tabela de dimensão associada. As dimensões degeneradas podem ser descartadas, se não fizerem falta à análise ou substituídas por chaves estrangeiras para dimensões

**Desenho Físico de Data Warehouses**

**Considere que qualquer dos atributos, A, B, e C, de uma dimensão é frequentemente usado individualmente para aplicar filtros com intervalos de valores em relatórios (por exemplo, A BETWEEN 10 AND 20). Explique se faz mais sentido criar um índice composto que abranja os três atributos, ou três índices separados, um para cada atributo, e justifique qual o tipo de índice que escolheria.**

Neste caso, o que faz mais sentido é criar um índice composto que abranja os três atributos, isto porque os índices estão sujeitos a dados redundantes porque permitem o acesso eficiente aos dados relevantes de cada índice que abrange um ou mais atributos de uma tabela base. Como os atributos da dimensão são usados frequentemente com filtros de intervalo, o índice mais apropriado seria o de Árvore B+.

**Justifique se a técnica de compressão de página é apropriada para uma dimensão Cliente onde são guardados dados demográficos da região de cada cliente, sendo que existe um número reduzido de regiões diferentes. No mesmo contexto, mostre o estado de uma página antes e depois de comprimida.**

A técnica de compressão de página pode fazer sentido para uma dimensão cliente onde são guardados dados demográficos da região de cada cliente e ainda por cima com um número reduzido de regiões diferentes.   
Ex:

**Apresente dois motivos para os relatórios analíticos abrangerem grandes volumes de dados. Nestas circunstâncias, identifique o problema que o desenho físico de *data warehouses* procura resolver e descreva como, em traços gerais**

Os relatórios analíticos abrangem grandes volumes de dados porque existe a agregação de muitos valores numéricos e a inclusão de dados históricos. Assim, o desenho físico de data warehouses procura minimizar a utilização de recursos, com menos dados pesquisados por interrogações SQL, menos espaço ocupado em disco aquando de atualizações periódicas e menos tempo para gerar relatórios dinâmicos.

**Descreva o aspeto geral de um índice *bitmap* quando aplicado a um atributo com elevada cardinalidade, em termos do número de colunas, de linhas, e de valores típicos, e justifique se este cenário é o mais apropriado a este tipo de índice.**

No caso descrito, o índice de bitmap ficaria com um elevado número de colunas, de linhas e de valores com dados iguais e redundantes. A solução de um índice bitmap não foi a ideal, uma vez que que os índices bitmap são mais apropriados para atributos com baixa cardinalidade.

**Descreva o propósito do desenho físico de um *data warehouse*, justificando se este é efetuado antes ou depois da modelação dimensional, e mencionando o papel desempenhado pelo disco rígido.**

O desenho físico de um data warehouse é feito depois da modelação dimensional e pretende resolver a problemática dos relatórios analíticos que envolvem grandes volumes de dados devido à agregação de muitos valores numéricos, inclusão de dados históricos e à construção interativa e incremental de relatórios. Desta forma a geração de relatórios pode ser morosa. Assim, o desenho físico procura minimizar a utilização de recursos através de menos dados pesquisados por interrogações SQL, menos espaço ocupado em disco aquando de atualizações periódicas e menos tempo para gerar relatórios dinâmicos. Neste aspeto o disco rígido desempenha um papel importante na organização e no tempo/custo de acesso aos dados, sendo que numa organização sequencial como num disco rígido, a transferência de dados sequenciais 10 a 20 vezes mais rápida do que em posições aleatórias.

**Considere os atributos ID, ÚltimoNome, Sexo, EstadoCivil, e RegiãoDoPaís de uma dimensão Cliente. Apresente um cenário de acesso a clientes em que seja vantajosa a utilização de índices de mapas de *bits*, indicando quais os atributos, de entre os mencionados, mais adequados a este tipo de índice, e descrevendo o conteúdo de um índice *bitmap* sobre um desses atributos.**

Os índices de bitmap são frequentemente utilizados porque são compactos (sequências de bits ocupam pouco espaço) e eficientes (disjunções e conjunções realizadas por comparação direta de bits e o processador do computador faz comparação de forma muito eficiente). Para quem se depara com o caso do enunciado, poderia querer saber as mulheres que são casadas, consultando os atributos ID, Sexo e Estado Civil. O atributo Sexo poderia ser transformado em 0 para Masculino e 1 para Feminino e o atributo EstadoCivil transformado em 0 para solteiro e 1 para casado. Assim, o resultado seria o ID que apresentasse 01 01 nos atributos selecionados.

**Os índices nos *data warehouses* de tipo ROLAP costumam ocupar três vezes o espaço dos dados. Indique o propósito de tamanha quantidade de dados indexados, mencionando se os índices incidem sobre atributos só da tabela de factos, só das dimensões, ou de todas as tabelas, e, por fim, justifique se esta abundância seria viável num sistema operacional.**

ROLAP, tabelas relacionais em disco • Sincronização rápida com fontes de dados (também relacionais) • Elevado tempo de resposta a interrogações

Em sistemas ROLAP – Existe uma relação principal, designada tabela de factos (fact table) – Relaciona cada medida de negócio com as dimensões – Dimensões em tabelas próprias, designadas dimension tables

**Os cubos de dados, incluindo os respetivos dados agregados pré-calculados, podem ser armazenados em partições. Descreva dois cenários de uso de partições que potenciam a geração mais rápida de relatórios, mencionando, se apropriada, a tecnologia de servidores paralelos.**

As partições são indicadas para tabelas de factos e dimensões “monstras” e pretendem acelerar a geração mais rápida de relatórios. Dois cenários para gerar relatórios mais rapidamente são as partições por intervalos de valores que é uma estratégia adequada para intervalos de datas e a partição de servidores paralelos através do shared nothing onde duas partições podem ser processadas em simultâneo e originam dois relatórios, focado cada um na sua partição e um outro relatório abrangendo todas as partições, em metade do tempo.

**Uma das tendências da *business intelligence* é os relatórios serem gerados a partir de cubos de dados guardados em memória RAM (volátil) do computador. Indique uma vantagem e uma desvantagem do designado *in-memory* BI face à produção de relatórios a partir do dispositivo tradicional de armazenamento de dados, e justifique se faz sentido construir e usar índices em memória RAM.**

**Considere os atributos RegiãoDoPaís, DonaDeCasa, e DiaDaSemana de um *data warehouse* de tipo ROLAP sobre audiências televisivas. Descreva o conteúdo de um índice de mapas de *bits (bitmap)* aplicado ao atributo que considerar mais adequado (dos apresentados) a este tipo de índice.**

Para o cenário retratado de um data warehouse de tipo ROLAP, o atributo que se adequa melhor a um bitmap é o DiaDaSemana. Isto porque tem um valor fixo de dias (7) e permite perceber por exemplo qual o dia da semana com maiores audiências televisivas.

**As técnicas de compressão permitem reduzir o espaço ocupado pelos dados armazenados em disco. Sabendo que algum tempo extra é gasto pelo processador para descomprimir os dados a fim de poderem ser usados, explique como é que o tempo de geração de relatórios pode, ainda assim, ser inferior relativamente ao não uso de compressão.**

As técnicas de compressão, apesar de apresentarem cerca de 10% de acréscimo no tempo de processador, este aspeto é compensado por acederam a menos dados no disco (30% a 50% inferior), o que em tabelas de factos e dimensões “monstras” são percentagens significativas de dados e naturalmente de tempo e espaço.

**Ilustre o uso da técnica de compressão de página, mostrando o estado de uma página antes e depois de comprimida, no contexto de uma dimensão onde a poupança de espaço seja relevante.**

**Considere que pretende particionar os dados de uma tabela de dimensão “monstra,” na qual as mudanças lentas têm vindo a ser registadas através da técnica de tipo 2. Assumindo que adquiriu um novo disco mais rápido que os existentes e que a maioria dos relatórios precisa apenas dos registos que estão em vigor, justifique que estratégia de partição escolheria para esses relatórios serem gerados em menos tempo.**

**Descreva um cenário em que seja vantajosa a utilização de uma vista materializada num *data warehouse* e justifique se seria necessário alterar as interrogações SQL dos relatórios existentes para tirar partido deste tipo de vista.**

Num cenário onde existe a geração de um relatório frequente (executado repetidas vezes) como, por exemplo, as somas de quantidades e euros das vendas por clientes e sucursais, com uma vista materializada as somas de quantidades e euros ficam guardadas em disco, devidamente agrupadas por cliente e sucursal. Tipicamente as vistas são usadas automaticamente, porque o otimizador de interrogações recorre a vistas se forem opção vantajosa, sendo assim adaptadas automaticamente sem necessidade de alterar as interrogações.

**Descreva duas caraterísticas das vistas materializadas que permitem reduzir o tempo de geração de relatórios que envolvam o cruzamento de dados provenientes de várias tabelas.**

Duas das características das vistas materializadas residem no facto de permitirem o pré-cálculo e armazenamento de valores agregados e evitam cálculo repetitivo de operações muito requisitadas.

**Considere a uma tabela de dimensão cliente que, entre outros, guarda um conjunto de atributos demográficos relacionados entre si e atualizados de dois em dois anos. Supondo que é aplicada a técnica de compressão de página a esta tabela, justifique de onde viriam as principais poupanças de espaço em disco.**

Perante o cenário descrito, as principais poupanças de espaço em disco residiriam no facto de a compressão de página identificar prefixos e valores comuns em registos da mesma página, guardando esses prefixos e valores comuns num dicionário de página, ou seja, quando houvesse uma atualização aos dados demográficos os prefixos e valores comuns (como por exemplo um ID da freguesia/concelho, o nome da freguesia/concelho, etc) já estariam comprimidos, havendo apenas a necessidade de atualização dos dados variáveis no tempo (por exemplo, número de habitantes)

**Considere que, no contexto de um *data warehouse* de tipo ROLAP usado numa maternidade, pretende indexar os atributos Idade (em meses), Sexo, e CorDosOlhos, pois estes são frequentemente usados na produção de relatórios dinâmicos.**

**a) Assumindo que é usual fazer a distinção entre bebés com menos de 4 meses dos restantes, justifique se seria apropriado utilizar um índice de função de dispersão para o atributo Idade.  
b) Admita que os valores possíveis para a cor dos olhos são azul, verde, e castanho. Descreva sucintamente o conteúdo de um índice de tipo *bitmap* aplicado a este atributo.  
c) Considere a interrogação SELECT AVG(f.Peso) FROM tblFactos f, dimBebé b WHERE (f.fkBebé = b.ID) AND (b.Sexo = ’F’ AND b.CorDosOlhos = ’azuis’). Justifique se o uso de índices de tipo *bitmap* em sexo e olhos permitiria dar uma resposta eficiente à interrogação.**

**Indique uma vantagem e uma desvantagem dos índices *bitmap* e descreva uma situação em que os usaria.**

Vantagem: Compactos. As sequências de bits ocupam relativamente pouco espaço.  
Desvantagem: Pouco apropriados para atributos com alta cardinalidade

Poderiam ser usados quando se pretende saber, por exemplo, as clientes do sexo feminino que são solteiras.

**Na estruturação de um *cluster* de discos para *data warehousing*, justifique que tipo de distribuição de dados pelos discos, ou nível RAID, escolheria.**

**Comente a seguinte afirmação: “o maior fator condicionante do desempenho de sistemas de *data warehousing* é o custo de *input/output*.”**

O custo input/output de sistemas data warehousing pode ser crítico. A eficiência de um sistema de data warehouse depende muito da velocidade de resposta a interrogações, processamento de dados e da forma como apresenta os resultados (input/output). Data warehousing envolve grandes volumes de dados por isso torna-se necessário minimizar a utilização de recursos sem perder a informação mais relevante. Desta forma, a forma como se vai encaixar todos estes fatores é fundamental para o desempenho de sistemas de data warehousing.

**Indique as vantagens e desvantagens de guardar resultados agregados em disco no contexto da sua utilização em *data warehousing*.**

A vantagem é que os resultados já se encontram processados e pré-agregados e a desvantagem é que se está a ocupar espaço em disco para guardar esses mesmos dados processados e pré-agregados.