FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Integração e Processamento Analítico de Informação

Coletânea de Exercícios

António Ferreira e André Falcão

amferreira@fc.ul.pt e aofalcao@fc.ul.pt

Março de 2022

Introdução

Esta coletânea de exercícios reúne a quase totalidade das perguntas dos testes e exames de Integração e Processamento Analítico de Informação, desde o ano letivo de 2007/08.

Os exercícios estão organizados em partes segundo a ordem de exposição da matéria ao longo do semestre para que esta coletânea possa ser utilizada como elemento de estudo pelos alunos, quer na revisão das questões apresentadas nas aulas, quer na preparação para os testes e exames.

Cada exercício está marcado com o ano e mês em que foi usado pela primeira vez num teste ou exame, como por exemplo **2016.05**. Alguns exercícios foram reformulados para serem mais facilmente compreendidos neste novo contexto de apresentação.

Nesta edição constam exercícios reunidos por António Ferreira a partir de testes e exames escritos pelo próprio a partir de 2009/2010 e por André Falcão entre 2007/08 e 2008/09.

Os autores.

Parte I — Sistemas de Apoio à Decisão e *Data Warehouses*

1. **2019.06** Desenhe um diagrama com as etapas do sistema ETL de um *data warehouse*, desde que os dados saem dos sistemas operacionais até que chegam aos decisores, e indique uma tarefa concreta de cada uma dessas etapas. Tenha o cuidado de incluir a *data presentation area* e *data staging area* no diagrama.

2. **2019.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre *data warehouses*.
   __ Na abordagem *bottom-up*, os *data marts* são feitos a pedido e derivam do *data warehouse*.
   __ A deteção de alterações nos dados é uma tarefa da etapa de transformação do sistema ETL.
   __ O cálculo e armazenamento em disco de valores agregados são feitos na etapa de carregamento.
   __ A vertente de aplicações de um projeto pode ser realizada autonomamente da vertente de tecnologias.
   __ Facilitam o acesso a dados históricos, sendo esta uma vantagem face às bases de dados federadas.

3. **2018.06** Indique duas razões que levaram ao surgimento de sistemas OLAP, em complemento dos tradicionais sistemas OLTP, tendo o cuidado de mencionar os respetivos públicos-alvo. Diga também por que razão continuam a existir sistemas OLTP nas organizações.

4. **2018.06** Compare os modelos de dados dos sistemas OLTP e OLAP em função da quantidade e tamanho das tabelas, em termos relativos. Adicionalmente, dê um exemplo de operação de leitura e outro de escrita sobre os dados para cada um destes tipos de sistema. Por fim, indique se são as leituras ou as escritas, ou ambas, que dominam na utilização habitual de cada tipo de sistema.

5. **2018.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre bases de dados federadas.
   __ Relatórios de apoio à decisão com dados históricos costumam ser viáveis.
   __ O desempenho dos sistemas operacionais é afetado pelas interrogações analíticas.
   __ O mediador permite abstrair o tipo dos vários sistemas operacionais envolvidos.
   __ As respostas às perguntas analíticas têm dados mais antigos do que com *data warehouses*.
   __ As interrogações analíticas são difíceis de construir pois não existe um modelo dimensional.

6. **2018.04** 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)*.
   __ 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.
   __ Oferecem fontes de dados externas adicionais a um *data warehouse*.

7. **2018.04** Descreva uma responsabilidade de cada etapa do sistema ETL de um *data warehouse*, e justifique a importância das mesmas para o apoio à decisão ou para o funcionamento diário da organização. Por exemplo: se essa responsabilidade não existisse, o decisor deixaria de poder fazer isto.

8. **2018.04** Preencha o texto em falta em cada afirmação sobre a construção de um *data warehouse*.
   - A atividade de escolha e instalação do produto está incluída na vertente de…
   - Cria-se primeiro o *data warehouse*, do qual derivam os *data marts*, na abordagem…
   - O desenho e desenvolvimento do sistema ETL faz parte da vertente de…
   - Segundo Kimball, o primeiro *data mart* a ser construído é para o…
   - As dimensões mais usadas na organização são mostradas na…

9. **2017.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre a gestão de dados mestre.
   __ Só é útil quando a organização dispõe de um *data warehouse*.

__ Procura atingir o objetivo de haver uma versão única da verdade sobre os dados.
__ Facilita o carregamento de dados para as dimensões de um *data warehouse*.
__ A sua introdução numa organização não afeta o funcionamento dos sistemas operacionais.
__ Requer a atribuição de um dono e de uma autoridade responsável a todos os itens de dados.

10. **2016.06** 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.

11. **2016.06** 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. Identifique essas vertentes, justifique se podem ser realizadas autonomamente umas das outras, e descreva o propósito de uma das atividades de cada vertente.

12. **2016.04** 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*.

13. **2016.04** Indique frases verdadeiras ou falsas (V/F) sobre sistemas de informação para executivos.
__ Fornecem muitos relatórios predefinidos para cobrir o máximo de cenários de decisão.
__ Mostram tendências nos dados ao longo do tempo, abrangendo vários anos se necessário.
__ Disponibilizam e cruzam dados de múltiplas fontes, desde que apenas internas à organização.
__ Requerem pouco ou nenhum treino para poderem ser usados pelos decisores.
__ Representam os dados apenas em forma de gráficos, em vez de texto ou tabelas.

14. **2015.06** 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.

15. **2014.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre sistemas de apoio à decisão.
__ As transações costumam ser curtas relativamente às dos sistemas operacionais.
__ As análises de dados tendem a ser exploratórias e a visar estudos de tendências.
__ A navegação nos dados é versátil, mas requer muito treino dos executivos.
__ Os dados disponibilizados para a tomada de decisão costumam ser apenas agregados.
__ Devido ao grande volume de dados o tempo de resposta do sistema costuma ser longo.

16. **2013.04** 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.

17. **2013.04** 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.

18. **2012.04** 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.

19. **2012.04** 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.

20. **2011.06** 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.

21. **2011.06** 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.

22. **2011.06** 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.

23. **2011.06** 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.

24. **2011.04** 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.

25. **2011.04** 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.

26. **2011.04** 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.

27. **2010.07** 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.

28. **2010.06** 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*.

29. **2009.06** No processo de ETL, justifique que tipo de técnica de mudança lenta é mais difícil de gerir, ilustrando os passos a seguir.

30. **2009.04** No contexto de um *data warehouse* defina sucintamente os seguintes conceitos: a) *data mart*; b) *data staging area*; e c) *drill across*.

Parte II — Relatórios Analíticos

31. **2019.04** Preencha o texto em falta em cada afirmação sobre relatórios analíticos.
    - Uma vantagem das extensões OLAP ao SQL é…
    - Ao ser arrastada uma medida para um relatório vazio é exibido…
    - O arrastar de um atributo de uma dimensão para um relatório serve para…
    - Para obter dados mais detalhados deve ser usada a operação de…
    - *Slicing* é um caso particularmente simples de…

32. **2017.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.
    __ A operação de *dicing* é um caso particularmente simples de *slicing*.
    __ As medidas só podem ser colocadas num relatório após os atributos das dimensões.
    __ Fazer *drill-down* geralmente revela menos valores do que os que estavam disponíveis.
    __ Cada medida M colocada num relatório acrescenta, por omissão, SUM(M) ao SELECT em SQL.
    __ Um atributo A posto nas linhas de um relatório vazio gera SELECT A … GROUP BY CUBE (A).

33. **2016.07** Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos em SQL.
    __ Com SQL clássico, o número de SELECTs é igual ao número de dimensões no relatório.
    __ A cláusula GROUP BY CUBE faz parte das extensões OLAP para SQL.
    __ GROUP BY ROLLUP (A, B, C) inclui cálculos agregados em função só de C.
    __ É incorreto usar SELECT A, B, SUM(C) FROM T GROUP BY ROLLUP (A, B, C).
    __ Um só SELECT de SQL clássico pode gerar um relatório com totais em colunas e linhas.

34. **2016.06** 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).

35. **2016.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.
    __ São calculadas menos somas se for usado CUBE em vez de ROLLUP num GROUP BY.
    __ As extensões OLAP ao SQL permitem otimizar o cálculo de valores agregados.
    __ A operação de *slice* é um caso particularmente simples da operação *dice*.
    __ *Roll-up* serve para procurar explicações detalhadas para os grandes totais.
    __ Podem ser incluídos atributos de mais de duas dimensões num relatório.

36. **2015.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.
    __ Só pode ser aplicada uma operação de *slice* em cada relatório.
    __ As hierarquias de atributos de dimensões possibilitam o *roll-up* e o *drill-down*.
    __ O *roll-up* serve para passar de dados mais detalhados para dados mais agregados.
    __ As extensões OLAP ao SQL não simplificam a escrita de comandos que geram relatórios.
    __ As medidas só podem ser colocadas num relatório após os atributos das dimensões.

37. **2014.04** 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.

38. **2014.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre relatórios analíticos.
    __ Os atributos das dimensões podem ser usados para esmiuçar as medidas numéricas.
    __ O *slicing* permite aplicar restrições por intervalo de valores numa dimensão.
    __ A cláusula GROUP BY ROLLUP permite gerar totais para todas as combinações de atributos.
    __ A operação de *dice* pode ser usada para restringir os valores que aparecem num relatório.

__ O *roll-up* serve para passar de dados mais agregados para dados mais detalhados.

39. **2013.04** 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).

40. **2012.06** 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.

41. **2012.06** 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).

42. **2011.06** 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.

43. **2011.04** Considere a interrogação SELECT País, Ano, SUM(Euros) FROM Vendas GROUP BY CUBE (País, Ano) do caso HyperVending. Descreva o resultado que seria obtido.

44. **2010.07** 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.

45. **2008.04** 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.

Parte III — Modelação Dimensional

46. **2019.06** Dê um exemplo de tabela de factos de tipo instantâneo periódico, com pelo menos uma medida, mencionando se cada facto abrange períodos variáveis ou fixos de tempo, se o carregamento de novos dados envolve só inserções, só atualizações, ou ambas as operações de escrita, e justifique se essa medida (ou uma de entre várias que tenha escolhido) é aditiva ou apenas semiaditiva.

47. **2019.04** Explique de que forma as bifurcações e as minidimensões podem ajudar a controlar o crescimento de dimensões "monstras," salientando o que têm em comum bem como as principais diferenças, e dando um exemplo adequado a cada uma dessas duas técnicas.

48. **2018.06** Complete cada afirmação sobre modelação dimensional.
    - A seguir à modelação das dimensões vem o passo de…
    - Inventários semanais de armazéns, de natureza periódica, são aditivos ao longo da dimensão…
    - A percentagem de espaço ocupado pelas tabelas de factos costuma ser aproximadamente…
    - Uma tabela de factos sem factos pode ser usada para…
    - A técnica de tipo 1 para mudanças lentas não permite…

49. **2018.06** Considere a dimensão 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 incluem a evolução histórica dos atributos Rendimentos e NúmeroDeFilhos. Supondo que estes atributos são movidos para uma minidimensão, descreva as adaptações que seria necessário fazer aos dados dos atributos, desenhe um diagrama com o novo esquema em estrela, descreva um caso em que a minidimensão permitiria abrandar o crescimento da dimensão, e explique como poderia ser analisada a evolução histórica dos referidos atributos.

50. **2018.06** Desenhe um esquema em estrela adequado ao registo de factos como o seguinte: no dia D1, o aluno A foi admitido no curso C da faculdade F da universidade U, com nota de entrada NE, tendo concluído o curso no dia D2, depois de se ter inscrito em L anos letivos, nos quais fez um total de T ECTS, com média final de MF valores. Indique também, para cada medida, se é ou não aditiva, justificando sucintamente, e proponha uma medida adicional que seja aditiva.

51. **2018.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre tabelas de ponte.
    __ Guardam caminhos entre todos os elementos de todos os níveis de uma hierarquia.
    __ Permitem agregar medidas para cada elemento abaixo de um dado elemento.
    __ São habitualmente usadas entre uma dimensão e uma tabela de factos.
    __ Não podem ser usadas para representar hierarquias de profundidade variável.
    __ A remoção de um elemento da hierarquia tem pouco impacto nas linhas da tabela.

52. **2018.06** Indique três critérios para a escolha de atributos a mover para uma bifurcação de uma dimensão "monstra". Depois, para apenas um desses critérios, justifique se o mesmo também se aplica às minidimensões.

53. **2018.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre modelação dimensional.
    __ Relatórios já existentes devem ser a base de partida para a modelação.
    __ As tabelas de factos sem factos servem para acompanhamento simples de eventos.
    __ A técnica 3 para mudanças lentas é adequada para valores antigos e atuais de muitas colunas.
    __ Uma dimensão conformada tem nomes e valores de atributos comuns em vários processos.
    __ Chaves substitutas em dimensões impedem o registo do histórico de mudanças lentas.

54. **2018.04** Complete cada afirmação sobre modelação dimensional.
    - A decomposição de uma tabela de dimensão em tabelas mais pequenas designa-se…
    - Saldos mensais de contas, de natureza periódica, são aditivos ao longo da dimensão…
    - As dimensões que não têm uma tabela própria chamam-se…
    - A mesma tabela de dimensão participar várias vezes num facto designa-se…
    - A seguir à identificação dos processos de negócio a modelar, vem a etapa de…

55. **2017.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre a modelação dimensional.
    __ A escolha do processo de negócio é feita depois de comparadas tabelas de factos alternativas.
    __ Cada processo de negócio tem a sua própria matriz de exequibilidade/valor.
    __ As medidas só devem ser identificadas depois de definidos os atributos das dimensões.
    __ Os processos prioritários estão no quadrante superior direito da matriz de processos.
    __ O grão determina os conceitos participantes em cada evento guardado na tabela de factos.

56. **2017.04** Considere os seguintes factos referentes a um só contexto: o armazém A recebeu U1 unidades do produto P no dia D1; e, do mesmo produto P saíram U2 unidades do mesmo armazém A no dia D2.
    a) Assumindo que se pretendem manter distintos os dois factos (e outros semelhantes), indique o tipo de tabela de factos e desenhe um esquema em estrela contendo as tabelas de factos e de dimensões, medidas numéricas, e vários atributos plausíveis, para além dos essenciais.
    b) Pretende-se ter uma segunda tabela de factos, que regista a quantidade de produtos de cada tipo em cada armazém no final de cada dia. Indique o tipo desta tabela de factos, desenhe o esquema em estrela, mencione que dimensões podem ser reutilizadas (não precisa mostrar os seus atributos), e justifique se as medidas são aditivas.

57. **2017.11** Considere uma dimensão onde está a ser usada a técnica de tipo 1 para registar mudanças lentas. Indique uma limitação dessa técnica, que pode ser ultrapassada com a adoção da técnica de tipo 2, e explique que alterações seria necessário fazer na estrutura e modo de utilização da tabela da dimensão.

58. **2017.04** Compare as bifurcações e as minidimensões em termos de: a) frequência de atualização dos valores dos atributos; e b) relação entre os atributos. Apresente também um exemplo adequado a uma destas técnicas para controlar o crescimento de dimensões "monstras", justificando.

59. **2016.07** Considere este facto: no jogo da final do campeonato europeu de futebol, realizado no dia 10 de julho de 2016, no estádio Stade de France com 80000 lugares e lotação esgotada, que fica na cidade de Saint-Denis, a equipa de Portugal venceu a da França por 1 a 0.

    Supondo que se pretende uma tabela de factos para guardar todos os jogos (ex. também quartos de final e meias finais) de campeonatos com equipas nacionais (ex. também os mundiais de futebol), indique: o tipo de tabela de factos, as medidas numéricas (caso existam), as dimensões, se há *role playing*, e alguns atributos representativos, incluindo chaves substitutas e estrangeiras. Nota: pode desenhar um diagrama de dados.

60. **2016.07** Indique afirmações verdadeiras (V) ou falsas (F) sobre esquemas em estrela.
    __ Costumam ser usados em *data warehouses* e também em sistemas operacionais.
    __ A tabela de factos está ligada a todas as tabelas de dimensões.
    __ A dimensão com mais atributos está tipicamente no centro do esquema.
    __ As medidas só são permitidas na tabela de factos.
    __ É comum duas ou mais dimensões estarem ligadas diretamente entre si.

61. **2016.06** Para o facto seguinte, indique o tipo de tabela de factos, as medidas numéricas (caso existam), as dimensões, se há *role playing*, e alguns atributos representativos, incluindo chaves substitutas e estrangeiras: no referendo R, do dia D, houve S eleitores a votar 'sim' e N a votar 'não', de um total de T eleitores da freguesia F, do concelho C, e distrito I, sendo também registadas as condições atmosféricas nessa freguesia e dia (ex. se choveu, nevou, ou fez sol, se houve vento forte ou fraco). Nota: pode desenhar um diagrama de dados.

62. **2016.06** 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 só será atualizado em 2020.
    a) Descreva como seria registada uma atualização do número de filhos de um cliente, incluindo o que seria guardado nos atributos demográficos, e cobrindo todos os atributos cujos valores seriam alterados.
    b) Explique como poderia ser controlado o crescimento da dimensão Cliente, particularmente no que diz respeito ao conjunto de atributos demográficos. Nota: pode fazer um diagrama, se for adequado.

63. **2016.04** Indique afirmações verdadeiras (V) ou falsas (F) sobre tabelas de ponte.
    __ Permitem navegar em hierarquias de profundidade variável.
    __ Servem para ligar duas tabelas de factos e permitir relatórios transdepartamentais.
    __ Cada linha da tabela inclui apontadores para os níveis ascendente e descendente.
    __ Guardam caminhos entre todos os valores de todos os níveis de uma hierarquia.
    __ Costumam ter menos linhas do que as tabelas de dimensões que lhes servem de base.

64. **2016.04** Para o facto seguinte, indique as medidas numéricas (se existirem), as dimensões, se há *role playing*, alguns atributos representativos, incluindo chaves substitutas e estrangeiras, e o tipo de tabela de factos: a pessoa P transferiu E euros na data D da conta C da agência A do banco B para a conta D da agência O, sinalizada como estando em local *offshore*, do banco F. Nota: pode desenhar um diagrama de dados.

65. **2015.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre modelação dimensional.
    __ As dimensões e as medidas costumam ser substantivos e os factos costumam ser verbos.
    __ Numa tabela de factos apenas as medidas de negócio podem ser atributos numéricos.
    __ Numa mesma tabela de dimensão podem coexistir várias hierarquias de atributos.
    __ As tabelas de factos costumam ter muitas colunas e as de dimensões tendem a ser estreitas.
    __ Uma minidimensão está ligada diretamente à dimensão "monstra" da qual derivou.

66. **2015.06** Para o facto seguinte, indique o tipo de tabela de factos, as medidas numéricas (caso existam), as dimensões, se há *role playing*, alguns atributos representativos, incluindo chaves substitutas e estrangeiras, e uma estimativa do número de linhas em cada tabela: a peça P, com um custo de E1 euros, do automóvel A, foi encomendada no dia D1 e chegou no dia D2 à oficina O, tendo sido aí feita uma reparação concluída no dia D3 no valor de E2 euros para H horas de mão-de-obra. Notas: pode desenhar um diagrama de dados, e considere que se trata de uma empresa de aluguer com centenas de automóveis de apenas três modelos.

67. **2014.06** Compare as bifurcações e minidimensões 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.

68. **2014.06** Compare os três tipos de tabelas de factos em função do grão e dê um exemplo plausível de cada tipo.

69. **2014.06** 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.

70. **2014.06** Explique a diferença entre medidas aditivas e semiaditivas, 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.

71. **2014.06** 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, em geral, a hierarquias em condições semelhantes).

72. **2014.06** 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.

73. **2014.04** 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.

74. **2013.06** Considere a dimensão Cliente, com dados desde 2010, para já com apenas mil registos, na qual está a ser usada a técnica 2 para guardar mudanças lentas que vão ocorrendo. De entre os atributos da dimensão, existe um conjunto dedicado à demografia da região do cliente, o qual só será atualizado em 2020.
    a) Assuma que cada registo ocupa em média 2000 bytes dos quais 1000 bytes são para os valores dos atributos demográficos, que chaves primárias e estrangeiras ocupam 4 bytes, e ainda que existem 5 regiões diferentes. Calcule o espaço total ocupado pela dimensão e compare com o que seria necessário caso fosse utilizada uma bifurcação para controlar o crescimento da dimensão Cliente.
    b) Sabendo que as bifurcações, quando bem aplicadas, permitem poupanças significativas de espaço em disco, justifique se devem ou não ser usadas intensivamente no modelo de dados de um *data warehouse*.

75. **2012.04** 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.

76. **2012.04** 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.

77. **2011.06** 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.

78. **2011.06** 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.

79. **2011.06** 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.

80. **2011.06** Apresente um exemplo de *role playing* de dimensão e indique uma forma de o realizar usando conceitos de SQL, com o mínimo de recursos e de forma transparente para o utilizador.

81. **2011.06** 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."

82. **2011.06** 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.

83. **2011.04** 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.

84. **2010.07** 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.

85. **2010.07** 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.

86. **2010.06** 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.

87. **2009.07** 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.

88. **2009.07** Bifurcações e minidimensões 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.

89. **2009.06** 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.

90. **2009.04** 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. Desenhe 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.

91. **2008.09** 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.

92. **2008.09** 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.

93. **2008.05** Indique três caraterísticas das tabelas de factos do tipo instantâneo periódico *(periodic snapshot)*, ilustrando com um exemplo.

94. **2008.04** 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.

95. **2008.04** No contexto de um *data warehouse* defina sucintamente os seguintes conceitos: a) grão; b) tabelas de factos sem factos; c) dimensões degeneradas.

Parte IV — Desenho Físico de *Data Warehouses*

96. **2019.05** Descreva a técnica de compressão de página e justifique se é apropriada para uma dimensão Cliente onde, entre outros, são guardados dados demográficos da região de cada cliente, sendo que existem poucas regiões distintas. Neste contexto, mostre o estado de uma página antes e depois de comprimida.

97. **2019.05** Complete cada afirmação sobre desenho físico de *data warehouses*.
    - O não suporte de pesquisas por intervalo *(dice)* é uma desvantagem do índice de tipo…
    - Quando os registos da tabela base seguem a ordem das entradas num índice, este designa-se…
    - Num índice de mapa de *bits* sobre um atributo existem tantas colunas quanto…
    - Uma interrogação SQL com o seu resultado guardado em disco designa-se…
    - Uma desvantagem do nível RAID 0 *(striping)* é…

98. **2018.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre desenho físico de *data warehouses.*
    __ Duas partições de uma mesma tabela podem ter índices sobre atributos diferentes.
    __ Os índices de árvore B+ e de função de dispersão são apropriados para pesquisas por intervalo.
    __ A compressão de linha é mais eficaz quando há prefixos iguais nos atributos de um registo.
    __ O resultado de uma interrogação pode ser totalmente obtido por via de uma vista materializada.
    __ Num disco rígido, o atraso de rotação ocupa a maior fatia do tempo de pesquisa de dados.

99. **2018.05** Complete cada afirmação sobre desenho físico de *data warehouses*.
    - Os índices de função de dispersão são mais eficientes que as árvores B+ na operação de…
    - O primeiro atributo de um índice multi-atributo de uma tabela de factos deve ser…
    - Quando os registos da tabela base seguem a ordem das entradas no índice, este designa-se…
    - Os índices de mapas de *bits* são apropriados para atributos com…
    - Um índice oferece tempos de pesquisa de dados mais curtos à custa de…

100. **2017.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre partições de dados e níveis RAID.
    __ Uma tabela muito grande pode ser decomposta em partições, cada uma no seu próprio disco.
    __ Duas estratégias de partição de dados são por intervalo e por lista de valores da *partition key*.
    __ Na arquitetura *shared memory*, os nós do servidor paralelo não partilham o mesmo disco.
    __ O nível RAID 0 oferece menor desempenho e maior confiança face aos outros níveis.
    __ Independentemente dos níveis RAID, o utilizador vê apenas um disco lógico.

101. **2017.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre o desenho físico de *data warehouses*.
    __ Deve ser realizado antes da modelação dimensional.
    __ Procura minimizar a utilização de recursos do sistema informático.
    __ Duas das técnicas são as vistas materializadas e as bifurcações de tabelas.
    __ 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.

102. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre compressão de dados.
    __ A compressão de página costuma poupar mais espaço em disco que a de registo/linha.
    __ As aplicações dos utilizadores não precisam de saber se os dados estão ou não comprimidos.
    __ O dicionário de página guarda prefixos e valores comuns a várias páginas da mesma tabela.
    __ A compressão é especialmente indicada quando os dados têm muitos valores repetidos.
    __ Requer a descompressão dos dados lidos do disco, tornando mais lentos os relatórios.

103. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre vistas materializadas.

__ Cada vista tem um SELECT associado, cujo resultado fica guardado em disco.
__ Evitam o cálculo repetitivo de operações muito requisitadas, tais como somas e médias.
__ As aplicações dos utilizadores precisam de saber os nomes das vistas para as poderem usar.
__ No comando CREATE MATERIALIZED VIEW, o SELECT não pode incluir GROUP BY.
__ Quando há atualização de dados, as vistas também são automaticamente atualizadas.

104. **2016.05** Indique uma razão para a taxa de compressão de dados ser habitualmente elevada nos *data warehouses*, e 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.

105. **2016.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre índices.
__ Num índice multi-atributo a ordem dos atributos a indexar (ex. X,Y ou Y,X) é irrelevante.
__ Um índice *bitmap* tem tantas colunas quantos os valores distintos do atributo indexado.
__ Num índice agrupado os registos de dados na tabela seguem a ordem das entradas no índice.
__ Os índices de dispersão são mais eficientes que os de árvore B+ em pesquisas por igualdade.
__ Num índice de dispersão cada *bucket* guarda entradas com valores indexados consecutivos.

106. **2014.06** 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.

107. **2014.05** 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.

108. **2014.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre discos rígidos.
__ O disco rígido é a componente que menos limita o desempenho de um *data warehouse*.
__ O tempo de posicionamento aumenta com a distância da cabeça de leitura à pista dos dados.
__ A transferência de dados em posições seguidas é mais rápida do que se estiverem dispersas.
__ O atraso de rotação é máximo quando os dados estão diametralmente opostos à cabeça.
__ Numa *solid state drive* (SSD) o tempo de acesso não depende da localização dos dados.

109. **2013.06** Descreva a estrutura e conteúdo de um índice *bitmap* quando aplicado a um atributo com elevada cardinalidade, como os países do mundo, em termos do número de colunas, número de linhas, e valores típicos, e justifique se este cenário é o mais apropriado a este tipo de índice.

110. **2013.05** 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.

111. **2013.05** 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.

112. **2013.05** 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 de índices seria viável num sistema operacional.

113. **2012.05** 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.

114. **2012.05** 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.

115. **2011.06** 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.

116. **2011.06** 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.

117. **2011.05** 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.

118. **2011.05** 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.

119. **2010.07** 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.

120. **2010.06** Considere 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.

121. **2010.06** 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 = 'azul'). Justifique se o uso de índices de tipo *bitmap* em sexo e olhos permitiria dar uma resposta eficiente à interrogação.

122. **2009.06** Indique uma vantagem e uma desvantagem dos índices *bitmap* e descreva uma situação em que os usaria.

123. **2009.06** Na estruturação de um *cluster* de discos para *data warehousing*, justifique que tipo de distribuição de dados pelos discos, ou nível RAID, escolheria.

124. **2009.06** Comente a seguinte afirmação: "o maior fator condicionante do desempenho de sistemas de *data warehousing* é o custo de *input/output*."

125. **2008.05** Indique as vantagens e desvantagens de guardar resultados agregados em disco no contexto da sua utilização em *data warehousing*.

Parte V — Prospeção de Dados

126. **2019.06** Complete cada afirmação sobre prospeção de dados.
    - O atributo no topo de uma árvore de decisão é…
    - A geração de regras de associação tem em conta um valor mínimo de…
    - No ciclo virtuoso da prospeção de dados, depois da geração de modelos vem a etapa de…
    - Um subconjunto dos dados em que a classe (ou decisão) é sempre a mesma tem entropia…
    - A proporção de verdadeiros positivos nos positivos detetados pelo modelo é dada pela métrica de…

127. **2019.05** Complete cada afirmação sobre prospeção de dados.
    - Duas desvantagens das redes neuronais são serem caixas opacas e…
    - Os bons agrupamentos *(clusters)* devem ter simultaneamente vários dados e…
    - A proporção de dados relevantes obtidos face a todos os dados relevantes é dada pela métrica de…
    - O conjunto de dados usado para estimar a taxa de erro de um modelo designa-se…
    - No método *k-means*, existem tantos agrupamentos iniciais quanto…

128. **2019.05** O algoritmo Apriori permite reduzir o espaço de descoberta de conjuntos frequentes de itens em cestos de compras. Indique a propriedade em que se baseia este algoritmo e descreva como são obtidos os conjuntos e qual a condição de paragem. Por último, explique como são geradas as regras de associação a partir dos conjuntos frequentes de itens.

129. **2018.06** Descreva um cenário exemplificativo dos quatro passos do ciclo virtuoso da prospeção de dados e indique em quais dos passos participam o *data warehouse* e os sistemas operacionais.

130. **2018.06** No âmbito das regras de associação, justifique se haveria ou não vantagem para a tomada de decisão em passar a aplicar o método a cestos contendo os produtos adquiridos em cada mês, em vez dos cestos de compras de cada visita ao supermercado. Indique também se essa mudança faria aumentar ou diminuir o suporte de cada produto, mencionando explicitamente os casos de produtos comprados frequentemente, como o pão, comprados rotineiramente, mas menos vezes, como o detergente de lavar a loiça, e comprados raramente, como um forno micro-ondas.

131. **2017.06** 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.

132. **2017.05** 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*.

133. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre árvores de decisão.
    __ As folhas são decisões possíveis.
    __ Os atributos que melhor diferenciam os dados ficam junto das folhas.
    __ O atributo colocado no topo da árvore é que tiver maior ganho de informação.
    __ A entropia de um conjunto de dados é zero quando todos os dados são iguais.

__ As árvores maiores costumam ser preferíveis, pois estão menos sujeitas a sobre-ajustamento.

134. **2017.05** Indique para que tipo de método de prospeção de dados fazem sentido os conjuntos de teste, treino, e validação, descrevendo o propósito de cada conjunto e algumas das suas caraterísticas desejáveis, como o tamanho relativo e os seus elementos. Nota: mostre qual a sequência lógica de uso destes conjuntos.

135. **2016.07** 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.

136. **2016.05** Descreva os passos do método de agrupamento hierárquico aglomerativo, indicando quantos são e onde estão os agrupamentos iniciais, como são criados novos agrupamentos, e quais as condições de paragem.

137. **2015.06** 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.

138. **2014.06** Na avaliação de modelos de classificação podem ser usadas as métricas de precisão e rechamada. Defina estas métricas e justifique se devem ambas ser consideradas em simultâneo ou se, regra geral, é suficiente calcular apenas uma delas.

139. **2014.06** Descreva o significado de entropia dos dados e explique, genericamente, de que forma pode ser usada na construção de árvores de decisão. Nota: tenha o cuidado de indicar se a entropia é aplicada a um atributo ou separadamente a cada valor de um atributo.

140. **2014.05** 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.

141. **2014.05** Identifique os papéis dos sistemas operacionais, *data warehouse*, e prospeção de dados num contexto organizacional e faça o seu enquadramento nas etapas do ciclo virtuoso da prospeção de dados.

142. **2013.05** Explique o algoritmo *k-means*, mencionando os pontos iniciais (quantos são e onde estão), como vão sendo criados os agrupamentos, qual a condição de paragem, e se é ou não um método supervisionado.

143. **2013.05** Descreva o significado de uma regra de associação ter confiança muito elevada e explique se do ponto de vista da tomada de decisão isso se traduz sempre em algo de útil. Por fim, justifique se uma regra nessas circunstâncias pode ser ignorada pelo algoritmo Apriori.

144. **2013.05** Explique de que forma as redes neuronais podem auxiliar a tomada de decisão e indique se estas são adequadas a casos em que se pretende que o modelo de prospeção de dados vá evoluindo com a chegada de dados novos em vez de ter de ser criado um modelo de raiz.

145. **2012.05** Suponha o início de construção de uma árvore de decisão, estando em análise dois atributos, A e B, os quais admitem os valores "sim" ou "não". Explique o que significa Entropia(A_{sim}) = 0, e, assumindo que Entropia(B_{sim}) = 0, indique em que circunstâncias o atributo A seria o principal diferenciador, e ainda, justifique abaixo de que ramo da árvore, A_{sim} ou $A_{não}$, seria colocado o atributo B.

146. **2011.06** Os métodos de prospeção de dados podem ser supervisionados ou não supervisionados. Indique em que circunstâncias um decisor pode ser levado a recorrer a cada tipo de método, ilustrando cada caso com um cenário plausível.

147. **2011.06** Comente a validade da seguinte afirmação, justificando: "as redes neuronais tendem a ser lentas durante a fase de treino/aprendizagem, mas rápidas na classificação de dados."

148. **2011.06** Explique o propósito dos conjuntos de teste, treino, e validação na classificação de dados, indicando explicitamente a sequência lógica de uso destes conjuntos, bem como a percentagem da totalidade dos dados pré-classificados que é recomendada em cada.

149. **2011.06** Descreva uma vantagem e uma desvantagem do algoritmo de *k-means* face ao de agrupamento hierárquico aglomerativo, indicando também se ambos os algoritmos seguem ou não uma abordagem de tipo *bottom-up* (dos pequenos para os grandes *clusters*).

150. **2011.05** Explique o propósito das tarefas de agrupamento e classificação, indicando um método apropriado a cada uma das tarefas, e justificando se esses métodos são ou não supervisionados.

151. **2011.05** Considere uma loja onde diariamente são feitas milhares de compras pelos clientes, e onde se detetou que "quem leva pão também costuma levar leite." Mais precisamente, em 1000 cestos, 600 continham pão e leite, 800 continham pão, e 700 continham leite.
    a) Calcule o suporte dos cestos com pão, leite, e pão e leite simultaneamente.
    b) Mostre qual das duas regras tem maior confiança: pão → leite ou leite → pão.

152. **2011.05** Justifique se uma regra de associação com confiança muito elevada (no caso limite, próxima ou até mesmo igual a 100%) pode ser ignorada pelo algoritmo Apriori. Nota: imagine um hipermercado onde também se vende ouro, prata, e diamantes.

153. **2011.05** Explique sucintamente o algoritmo de aprendizagem assente em *back-propagation*, no contexto das redes neuronais. Nota: pode considerar o caso em que o resultado obtido no neurónio de saída é superior ao resultado esperado.

154. **2010.07** Considere uma loja onde se detetou, com confiança de 100%, que "quem leva joias também leva caviar." Sabendo que em 1000 cestos, 2 continham caviar, estime o número de pessoas que levaram joias. Nota: explique o seu raciocínio.

155. **2010.06** Os métodos de prospeção de dados podem ser supervisionados ou não supervisionados. Descreva uma diferença fundamental entre estas duas abordagens e dê um exemplo de método de cada tipo.

156. **2009.07** "Correlação não é causalidade" é um dito habitual em prospeção de dados. Discuta a validade deste dito no contexto da avaliação de modelos, e indique que cuidados se devem ter para discriminar os dois conceitos.

157. **2009.06** Comente a seguinte afirmação "o *overfitting* é um problema típico em modelos de prospeção de dados mal ajustados."

158. **2009.06** Descreva os passos habituais na preparação de um conjunto de dados numéricos e alfanuméricos para a sua utilização num problema de classificação usando redes neuronais, salientando os principais cuidados a ter.

159. **2009.06** Considere que numa base de dados com cerca de 5 milhões de infrações e acidentes registados pela brigada de trânsito, encontra-se informação sobre condutores, viaturas envolvidas, e informação sobre o acidente/infração (por exemplo, local e tipo de infração, álcool e/ou drogas no sangue dos condutores envolvidos, comportamento, agentes envolvidos, entre outros). Indique como faria para procurar relações de correlação/causalidade entre os vários fatores e o tipo de acidente, referindo ainda as principais dificuldades que encontraria e que técnicas utilizaria para as ultrapassar.

160. **2008.09** Depois do ajustamento de um modelo de *data mining* para classificação, a sua utilização está sujeita a condicionantes. Refira algumas das principais, justificando a sua resposta.

161. **2008.05** Indique como faria para analisar os dados seguintes, referindo ainda quais as principais dificuldades que encontraria e que estratégias utilizaria para as ultrapassar.

    A DAS-28 é uma medida empírica com valores contínuos da severidade de sintomas de artrite reumatoide, atribuída por análise visual de pacientes. Para um estudo desta patologia, fez-se o mapeamento genético de 8 caraterísticas, que controlam a expressão de uma proteína identificada em todos os casos desta doença (mas também presente em pessoas sem a doença) para um conjunto de 300 pacientes. Para cada paciente conhece-se ainda a medicação que toma, bem como um conjunto de cerca de 200 atributos qualitativos e quantitativos referentes às suas caraterísticas, ao seu modo de vida, e aspetos relacionados com a doença. Nem todos os atributos têm valores conhecidos.