

# e-Fólio B

**U.C. 21103**

**Sistemas de Gestão de Bases de Dados**

**2024-2025**

## **Resolução e Critérios de Correção**

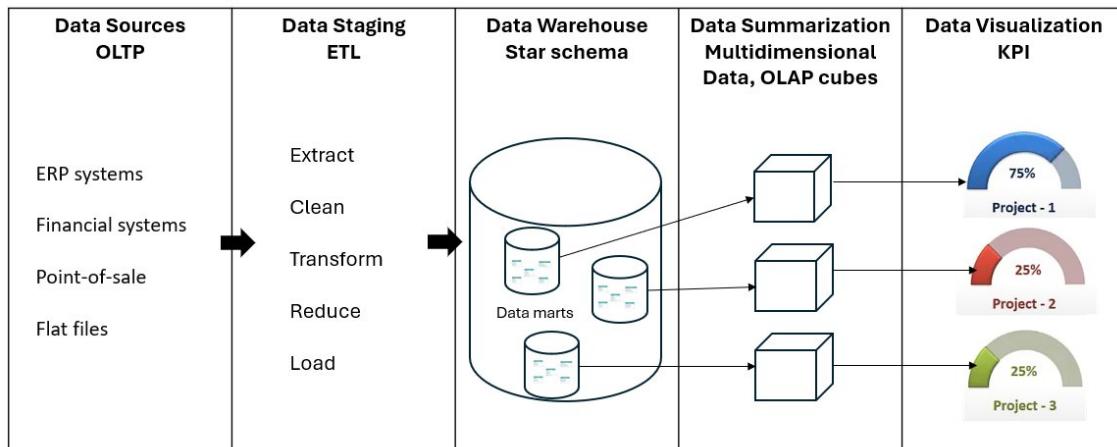
### **INSTRUÇÕES**

- 1) O e-fólio é constituído por 5 perguntas. A cotação global é de 5 valores.
- 2) O e-fólio deve ser entregue num único ficheiro PDF, não zipado, com fundo branco, com perguntas numeradas e sem necessidade de rodar o texto para o ler. Cada pergunta com uma ou mais páginas, deve ser iniciada numa nova página. Penalização de 10% a 100%.
- 3) Não são aceites e-fólios manuscritos, i.e., tem penalização de 100%.
- 4) O nome do ficheiro: <nome estudante> + “eFolioB”.
- 5) Na primeira página do e-fólio deve constar o nome completo do estudante bem como o seu número. Penalização de 10% a 100%.
- 6) Durante a realização do e-fólio, os estudantes devem concentrar-se na resolução do seu trabalho individual, não sendo permitida a colocação de perguntas ao professor ou entre colegas.
- 7) Nesta avaliação, não deve utilizar ferramentas de IA generativa, como o ChatGPT.
- 8) A interpretação das perguntas também faz parte da sua resolução, se encontrar alguma ambiguidade deve indicar claramente como foi resolvida.
- 9) A legibilidade, a objetividade e a clareza nas respostas serão valorizadas, pelo que, a falta destas qualidades será penalizada.
- 10) Critérios de correção gerais: todas as respostas devem ser justificadas, incluir imagens e exemplos com vista a clarificar os argumentos expostos. Devem ser utilizadas referências das páginas da bibliografia adotada e recomendada.

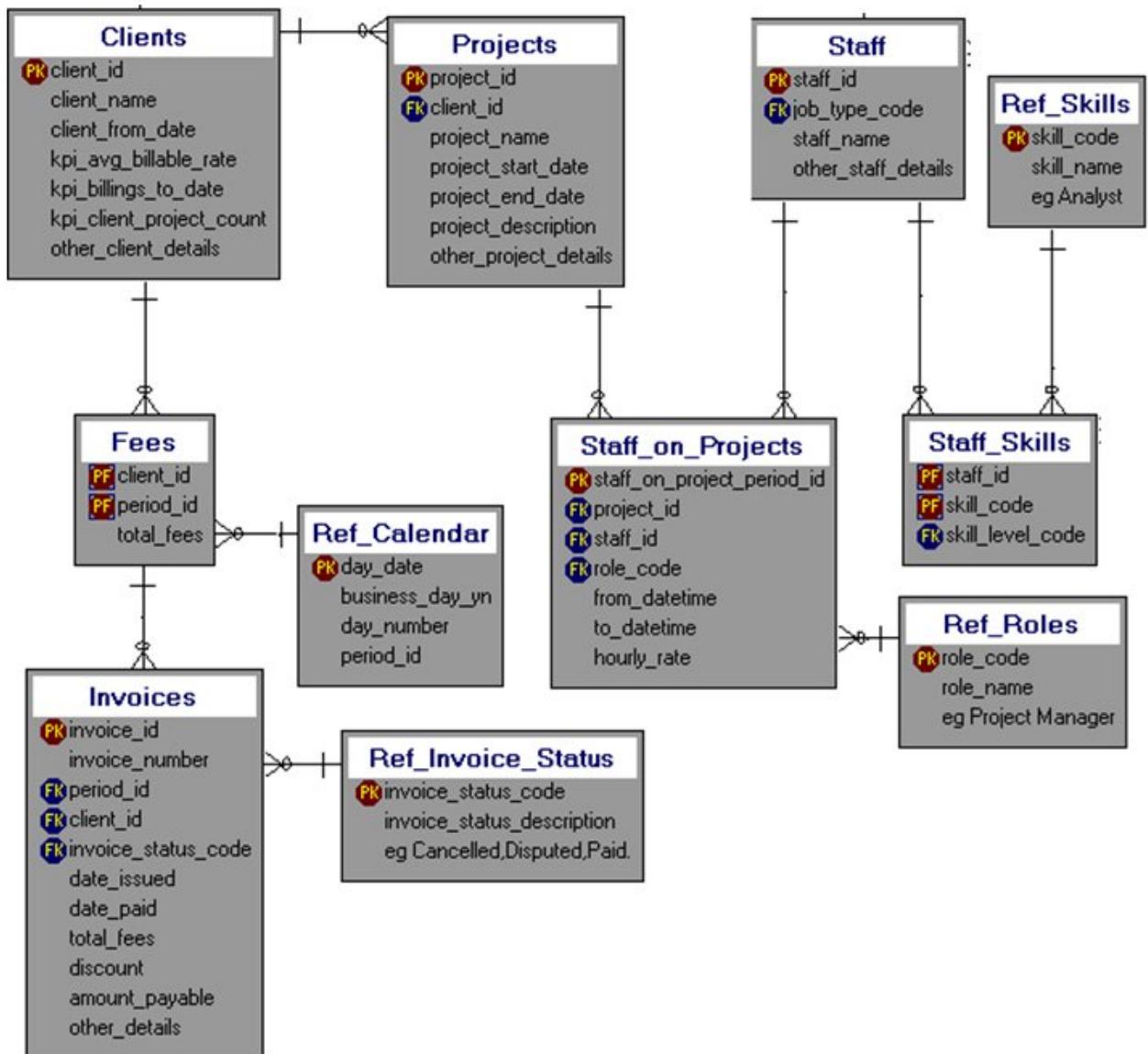
#### **Vetor Cotações**

1 2 3 4 5 pergunta  
10 10 10 10 10 décimas

Um Data Warehouse (DW) é um sistema projetado para armazenar grandes volumes de dados originados de diversas fontes, com o objetivo de apoiar a tomada de decisões numa organização.

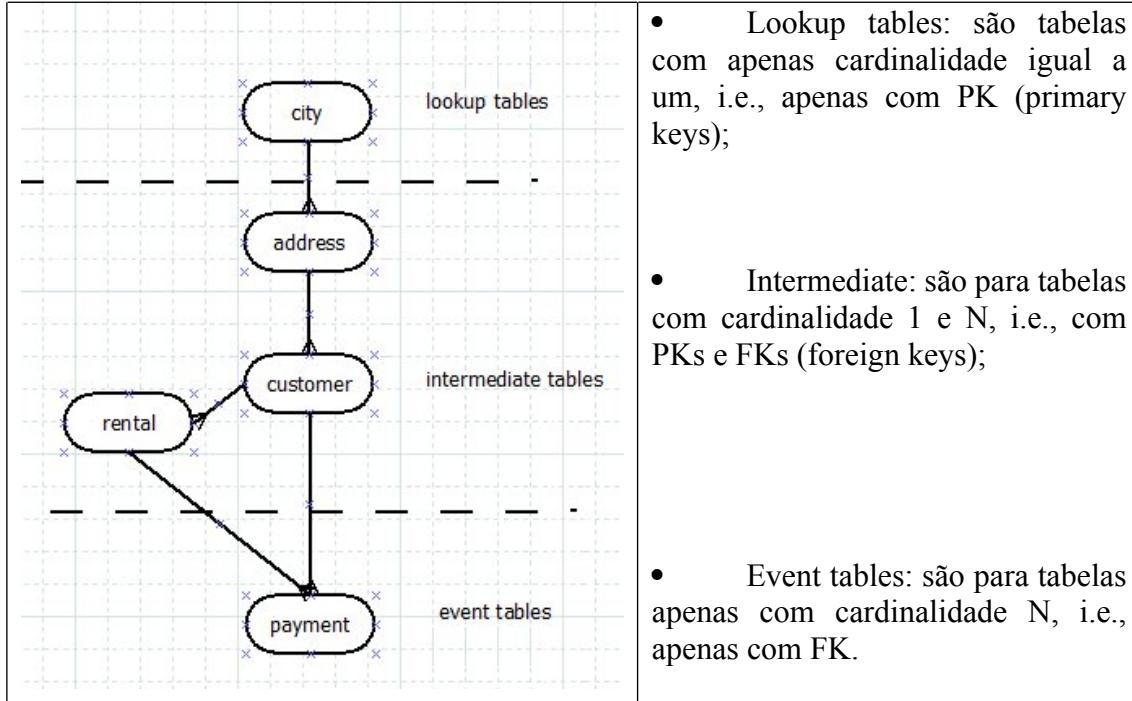


Considere a seguinte base de dados OLTP de uma empresa com projetos, para as perguntas seguintes:



## 1) (1 valor) Extração dos dados

Considere a base de dados de uma empresa com projetos. Considere os seguintes tipos de tabelas:



- **Lookup tables:** são tabelas com apenas cardinalidade igual a um, i.e., apenas com PK (primary keys);
- **Intermediate:** são para tabelas com cardinalidade 1 e N, i.e., com PKs e FKs (foreign keys);
- **Event tables:** são para tabelas apenas com cardinalidade N, i.e., apenas com FK.

Considere ainda, as seguintes formas desnormalizadas (FD):

- 1FD – constituída por uma poli-árvore, com a replicação das tabelas intermédias e de lookup que forem necessárias para evitar caminhos múltiplos;
- 2FD – constituída por várias árvores separadas, com a replicação das tabelas intermédias e de lookup que forem necessárias para evitar caminho múltiplos; esta FD é equivalente ao esquema em estrela ou ao esquema floco-de-neve.
- 3FD – o processo de desnormalização termina com a junção de todas as tabelas da árvore com vista a uma rápida leitura dos dados.

1.1) Quais as “event tables” que encontra no esquema OLTP da empresa de projetos? Represente graficamente os dados na 2FD da seguinte forma: as ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo. Depois de representar as tabelas classifique-as segundo a tipologia indicada (lookup, intermédia, eventos) como na figura.

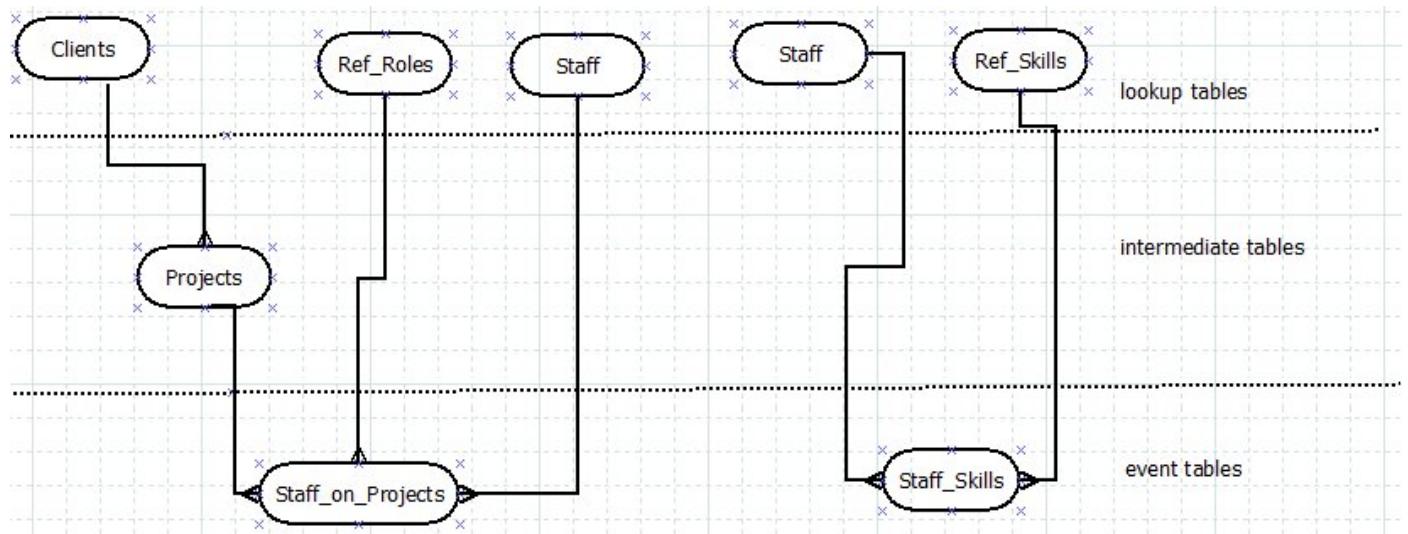
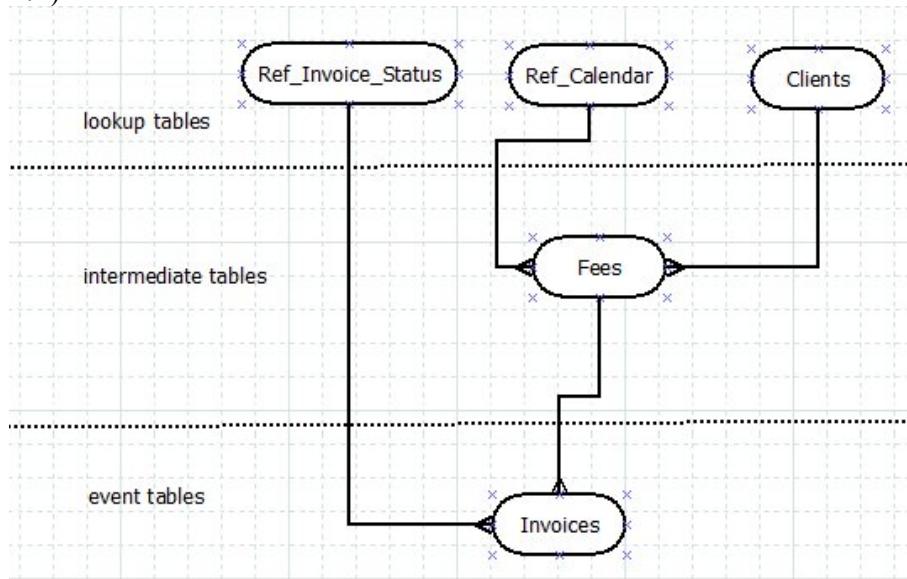
1.2) Considere os seguintes tipos de atributos da tabela de eventos:

- **Aditivos:** são atributos que podem ser agregados (somados) por todas as dimensões, ex: valor da venda (usar Sum() sempre)
- **Semi-aditivos:** são atributos que podem ser agregados (somados) por algumas as dimensões, ex: quantidade (usar Sum() em condições particulares)
- **Não-aditivos:** são atributos que não podem ser agregados (somados), ex: preço unitário (usar Average() por exemplo)
- **Sem factos:** só existem identificadores (usar a função Count() dos identificadores).

Para cada tabela de evento encontrada defina os atributos aditivos, semi-aditivos, não-aditivos e sem factos. Justifique a resposta.

## Resposta parcial:

1.1)



A base de dados original não tem caminhos múltiplos, pelo que está na 1FD. As ‘event tables’ são três: `Invoices`, `Staff_on_Projects` e `Staff_Skills`.

As figuras em cima representam a base de dados na 2FD constituída por árvores separadas, com replicação de tabelas.

## 1.2) Tipos de atributos da tabela de eventos

Invoices: Total\_fees e Amount\_payable são aditivos

Staff\_on\_Projects: hourly\_rate é não-aditivo

Staff\_Skills: sem factos

Critérios de correção:

- 1.1) 7 décimas, base dados na 2FD e tipologia das tabelas
- 1.2) 3 décimas, tipos de atributos da tabela de eventos
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

## 2) (1 valor) Qualidade dos dados

Considere as seguintes dimensões na Qualidade de Dados, ou Limpeza de Dados:

dimensão (PT)	dimension	description	example error	automatic assessment
pontualidade	timeliness	is the information up-to-date?	check last update	easy to check
unicidade	uniqueness	# duplicated data	duplicated ID or name	easy to check
completude	completeness	# missing values	'NA', ''	easy to check
precisão	accuracy	# data errors, # typos	'Brawn' instead of 'Brown'	check with lookup tables
validade	validity	# validity format violations	format violations in dates or emails	check specific syntaxes, e.g. date, email, #phone
consistência	consistency	# inconsistencies	name gender John Snow female	more difficult to check; (compare pairs of attributes)

Considere ainda o seguinte conjunto de dados desnормalizado:

customer_id	firstname	surname	gender	date_of_birth	acum_purchases	e-mail
100	Anne	Kirk	F	22/08/1971	120,00 €	anne.kirk@gmail.com
101	Elise	Cloney	F	25/01/1975	795,00 €	e.cloney@gmail.com
102	Zoe	Robinson	F	01/02/1962	904,00 €	ZoeR62@gmail.com
103	Mary	Grey	F	12/07/1985	10,00 €	MaryG1985@gmail.com
103	kate	Braawn		05-13-1980	832,00	KateBrown@gmail.com
106	Eva	Smith	F	18/07/1974	73,00 €	EvaSmith@gmail#.com
108	Molly	Smith	F	05/11/1963	264,00 €	MollyS1980@gmail.com
110	Marc	Smith	M	30/05/1967	742,00 €	Marc@gmail.com
114	Alister	Mongomery	M	19/04/1999	324,50 €	AlisterM1999@gmail.com
116	Hugh	Whiteley	M	28/08/1973	277,50 €	HughW1973@gmail.com

2.1) Proceda à verificação dos erros no conjunto de dados, identificando para cada elemento (linha, coluna) o tipo de erro/dimensão. Mostre a nova tabela com erros a cor amarela. Recupere manualmente os dados com erros e mostre o resultado noutra tabela.

2.2) Para medir a qualidade de cada dimensão use a métrica  $= 1 - \frac{\# \text{ elementos errados}}{\# \text{ total de elementos}} \times 100$  e complete a tabela seguinte:

dimensão	métrica qualidade (%)
Unicidade	
Completude	
Precisão	
Validade	
Consistência	
Geral	

## Resposta parcial:

2.1)

customer_id	firstname	surname	gender	date_of_birth	acum_purchases	e-mail
100	Anne	Kirk	F	22/08/1971	120,00 €	anne.kirk@gmail.com
101	Elise	Cloney	F	25/01/1975	795,00 €	e.cloney@gmail.com
102	Zoe	Robinson	F	01/02/1962	904,00 €	ZoeR62@gmail.com
103	Mary	Grey	F	12/07/1985	10,00 €	MaryG1985@gmail.com
103	kate	Braawn		05-13-1980	832,00	KateBrown@gmail.com
106	Eva	Smith	F	18/07/1974	73,00 €	EvaSmith@gmail#.com
108	Molly	Smith	F	05/11/1963	264,00 €	MollyS1980@gmail.com
110	Marc	Smith	M	30/05/1967	742,00 €	Marc@gmail.com
114	Alister	Mongomery	M	19/04/1999	324,50 €	AlisterM1999@gmail.com
116	Hugh	Whiteley	M	28/08/1973	277,50 €	HughW1973@gmail.com
unicidade	precisão	precisão	completude	validade	validade	validade
duplicados	typo	typo	missing	formato	formato	formato

customer_id	firstname	surname	gender	date_of_birth	acum_purchases	e-mail
100	Anne	Kirk	F	22/08/1971	120,00 €	anne.kirk@gmail.com
101	Elise	Cloney	F	25/01/1975	795,00 €	e.cloney@gmail.com
102	Zoe	Robinson	F	01/02/1962	904,00 €	ZoeR62@gmail.com
103	Mary	Grey	F	12/07/1985	10,00 €	MaryG1985@gmail.com
104	Kate	Brown	F	13/05/1980	832,00 €	KateBrown@gmail.com
106	Eva	Smith	F	18/07/1974	73,00 €	EvaSmith@gmail.com
108	Molly	Smith	F	05/11/1963	264,00 €	MollyS1980@gmail.com
110	Marc	Smith	M	30/05/1967	742,00 €	Marc@gmail.com
114	Alister	Mongomery	M	19/04/1999	324,50 €	AlisterM1999@gmail.com
116	Hugh	Whiteley	M	28/08/1973	277,50 €	HughW1973@gmail.com

2.2)

dimensão	elementos errados	total	erros	qualidade
Unicidade	1	70	1%	99%
Completude	1	70	1%	99%
Precisão	2	70	3%	97%
Validade	3	70	4%	96%
Consistência	0	70	0%	100%
Geral	7	70	10%	90%

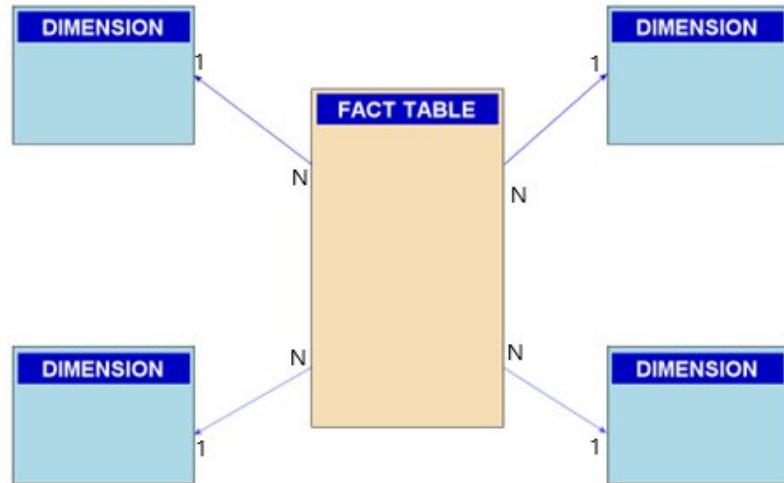
Critérios de correção:

- 2.1) 6 décimas, dimensão da qualidade e recuperação erro
- 2.2) 4 décimas, taxa de erros e qualidade do dataset
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

### 3) (1 valor) Data warehouse: star schema

Considere novamente a base de dados de uma empresa com projetos, em particular a tabela de factos com a informação de Invoices. Considere ainda que  $\text{Amount\_payable} = \text{Total\_fees} - \text{Discount}$ .

3.1) Crie um Data Mart, com um esquema em estrela, com os dados de Invoices com pelo menos três dimensões como o da figura.

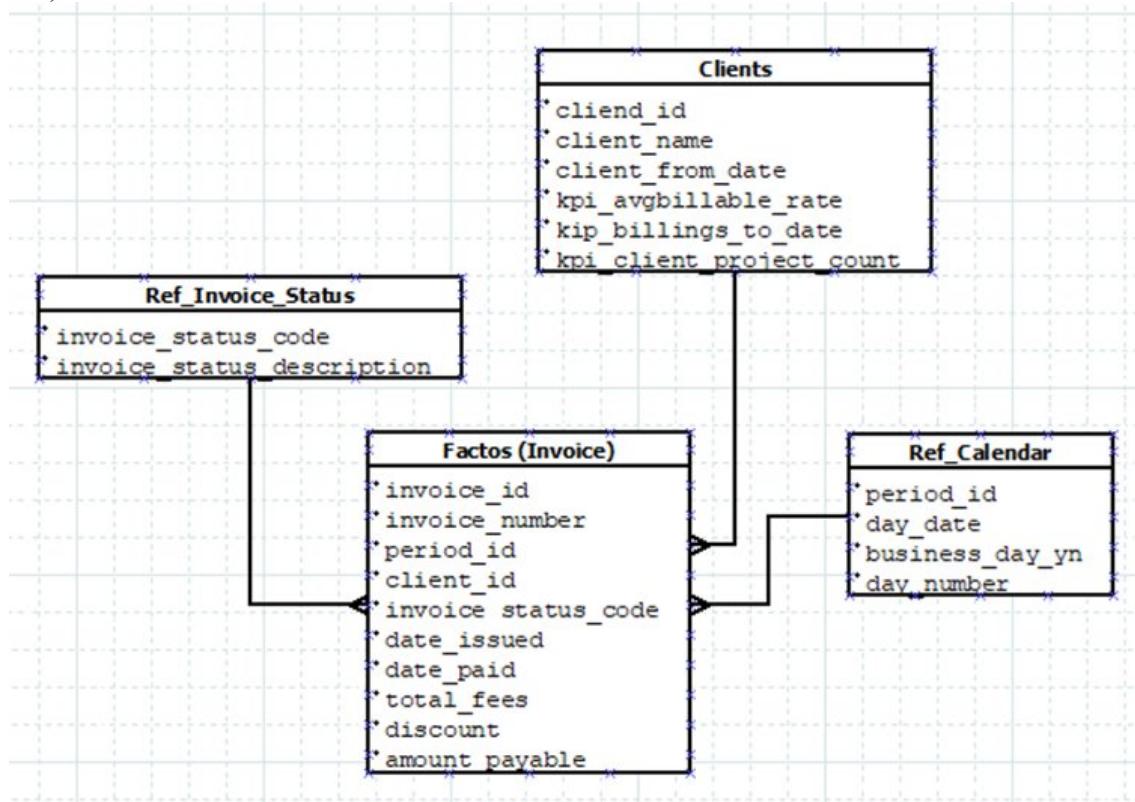


3.2) Diga quais das seguintes perguntas são passíveis de serem realizadas (responda verdadeiro ou falso). Justifique as respostas.

- Qual é o valor total de receitas geradas por período?
- Quantas faturas estão marcadas como pendentes ou canceladas e qual é o valor total associado a essas faturas?
- Qual é o desconto médio concedido por cliente ou por período?
- Quais produtos ou serviços estão associados a cada fatura?
- Qual é a satisfação do cliente com relação às faturas ou serviços?
- Quais são os métodos de pagamento utilizados para cada fatura?

Resposta parcial:

3.1)



3.2) Diga quais das seguintes perguntas são passíveis de serem realizadas (responda verdadeiro ou falso). Justifique as respostas.

- Qual é o valor total de receitas geradas por período? (Verdadeiro)
- Quantas faturas estão marcadas como pendentes ou canceladas e qual é o valor total associado a essas faturas? (Verdadeiro, considerando que 'disputed' ou contestada equivale a pendente)
- Qual é o desconto médio concedido por cliente ou por período? (Verdadeiro)
- Quais produtos ou serviços estão associados a cada fatura? (Falso)
- Qual é a satisfação do cliente com relação às faturas ou serviços? (Falso)
- Quais são os métodos de pagamento utilizados para cada fatura? (Falso)

Critérios de correção:

A tabela **Ref\_Calendar** tem como chave principal **Period\_id**.

- 3.1) 4 décimas, estrela com 3 dimensões (também correto se utilizar a tabela **Fees**)
- 3.2) 6 décimas, as respostas devem ser justificadas
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

**4) (1 valor) Sumarização de dados: operações OLAP**

Considere os seguintes dados da tabela de factos com a informação de Invoices relativos e a pergunta (b) qual é o total de receita gerada por período, cliente e Invoice status code.

invoice_number	period_id	client_id	invoice_status_code		date_issued		date_paid		total_fees		discount		amount_payable
INV-0001	2	110	paid	03/11/2024	20/11/2024	736,96 €	95,44 €	641,52 €					
INV-0002	1	104	pending	23/10/2024		828,33 €	89,73 €	738,60 €					
INV-0003	2	106	paid	11/10/2024	24/11/2024	597,99 €	64,36 €	533,63 €					
INV-0004	2	100	paid	27/09/2024	22/11/2024	359,42 €	8,15 €	351,27 €					
INV-0005	1	110	paid	13/02/2024	22/11/2024	278,57 €	68,74 €	209,83 €					
INV-0006	2	108	paid	20/10/2024	21/11/2024	264,43 €	29,79 €	234,64 €					
INV-0007	1	109	pending	28/05/2024		469,80 €	99,11 €	370,69 €					
INV-0008	1	105	pending	18/05/2024		361,51 €	1,33 €	360,18 €					
INV-0009	2	103	pending	13/09/2024		261,89 €	70,53 €	191,36 €					
INV-0010	2	103	cancelled	04/09/2024		822,10 €	43,73 €	778,37 €					

4.1) Explique e exemplifique em Excel o que entende por Roll-up e Drill-down, usando o mês da emissão da fatura e o período.

4.2) Explique e exemplifique em Excel o que entende por Slide e Dice usando o mês da emissão da fatura, o período e o status da fatura.

**Resposta parcial:**

4.1) O Roll-up e o Drill-down consideram a hierarquia dos dados. Para o efeito criou-se um novo atributo Month derivado de Date\_issued.

Sum of amount\_payable Column Labels

Row Labels

100	351,27 €	351,27 €	
103	969,73 €	969,73 €	
104	738,60 €	738,60 €	
105	360,18 €	360,18 €	
106	533,63 €	533,63 €	
108	234,64 €	234,64 €	
109	370,69 €	370,69 €	
110	209,83 €	641,52 €	851,35 €
<b>Grand Total</b>	<b>1 679,30 €</b>	<b>2 730,79 €</b>	<b>4 410,09 €</b>

Sum of amount\_payable Column Labels

Row Labels

100	351,27 €	351,27 €	351,27 €	351,27 €		
103	969,73 €	969,73 €	969,73 €	969,73 €		
104	738,60 €	738,60 €	738,60 €	738,60 €		
105	360,18 €	360,18 €	360,18 €	360,18 €		
106	533,63 €	533,63 €	533,63 €	533,63 €		
108	234,64 €	234,64 €	234,64 €	234,64 €		
109	370,69 €	370,69 €	370,69 €	370,69 €		
110	209,83 €	641,52 €	641,52 €	851,35 €		
<b>Grand Total</b>	<b>1 679,30 €</b>	<b>1 321,00 €</b>	<b>768,27 €</b>	<b>641,52 €</b>	<b>2 730,79 €</b>	<b>4 410,09 €</b>

#### 4.2) Slide e Dice

Sum of amount\_payable Column Labels

Row Labels

100	351,27 €	351,27 €	
103	969,73 €	969,73 €	
104	738,60 €	738,60 €	
105	360,18 €	360,18 €	
106	533,63 €	533,63 €	
108	234,64 €	234,64 €	
109	370,69 €	370,69 €	
110	209,83 €	641,52 €	851,35 €
<b>Grand Total</b>	<b>1 679,30 €</b>	<b>2 730,79 €</b>	<b>4 410,09 €</b>

Sum of amount\_payable Column Labels

Row Labels

100	351,27 €	351,27 €
103	969,73 €	969,73 €
106	533,63 €	533,63 €
108	234,64 €	234,64 €
110	641,52 €	641,52 €
<b>Grand Total</b>	<b>2 730,79 €</b>	<b>2 730,79 €</b>

Sum of amount\_payable Column Labels

Row Labels

103	778,37 €	778,37 €
<b>Grand Total</b>	<b>778,37 €</b>	<b>778,37 €</b>

Critérios de correção:

- 4.1) 5 décimas, Roll-up e o Drill-down
- 4.2) 5 décimas, Slide e Dice
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

## 5) (1 valor) Data Mining

Com base nas “Lecture Notes: Ciências dos Dados”, considere a Tabela 1.

Tabela 1. Tabela do modelo relacional

doente (chave)	idade	#medicamentos	pagamentos	complicação
1	52	7	121 €	sim
2	57	9	7,113 €	sim
3	43	6	75 €	sim
4	33	6	3,720 €	não
5	35	8	4,489 €	não
6	49	8	77 €	sim
7	58	4	39 €	não
8	62	3	79 €	não
9	48	0	2,797 €	não
10	37	6	90 €	sim

5.1) Altere a base de dados com vista a criar regras associativas que encontrem o medicamento X que é mais utilizado com o medicamento Y. Exemplifique com dados artificiais e encontre os dois medicamentos mais utilizados em conjunto. Os medicamentos têm nomes no intervalo de (A..Z).

5.2) Considere a seguinte matriz de medicamentos (A..Z) versus doentes (1..5).

	1	2	3	4	5
A		1			1
B	1		1	1	
C		1	1		1
D	1				
E					
F	1	1		1	1
G		1	1	1	1
H	1	1	1	1	1
I	1	1			1
J					
K	1			1	
L	1	1	1	1	
M		1	1		1
N		1			1
soma	7	9	6	6	8

Quais os 2 medicamentos mais comprados em conjunto? Quais os 3 medicamentos mais comprados em conjunto? Justifique a resposta.

## Resposta parcial:

5.1) Para além da tabela existente deve ser criada uma tabela, da seguinte forma:

Doente-> idade, #medicamentos, pagamentos complicações

Doente, linha -> medicamento

doente	linha	medicamento
1	1	A
1	2	C
1	3	F
1	4	M
1	5	N
1	6	O
1	7	P
2	1	A
2	2	B
2	3	F
2	4	M
...	...	...

5.2) Para encontrar os produtos mais comprados em conjunto, podemos criar a matriz (medicamentos versus medicamentos).

soma	A	B	C	D	F	G	H	I	K	L	M	N
A	0	0	2	0	2	2	2	2	0	1	2	2
B	0	0	1	1	2	2	3	1	2	3	1	0
C	2	1	0	0	2	3	3	2	0	2	3	2
D	0	1	0	0	1	0	1	1	1	1	0	0
F	2	2	2	1	0	3	4	3	2	3	2	2
G	2	2	3	0	3	0	4	2	1	3	3	2
H	2	3	3	1	4	4	0	3	2	4	3	2
I	2	1	2	1	3	2	3	0	1	2	2	2
K	0	2	0	1	2	1	2	1	0	2	0	0
L	1	3	2	1	3	3	4	2	2	0	2	1
M	2	1	3	0	2	3	3	2	0	2	0	2
N	2	0	2	0	2	2	2	2	0	1	2	0

Como pares de medicamentos mais comprados temos (F, H), (G, H) e (H, L).

Como trios de medicamentos mais comprados temos (G, H, F) ou (G, H, L).

Existem algoritmos específicos como o Apriori para encontrar estes subconjuntos.

Critérios de correção:

- 5.1) 3 décimas, alteração da base de dados
- 5.2) 7 décimas, medicamentos comprados em conjunto
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%