Normalização

Base de Dados - 2018/19

Carlos Costa

1

Introdução

- Já estudámos aspectos de desenho conceptual de base de dados e respectivo mapeamento para o modelo relacional.
- No entanto, nunca apresentámos um processo formal de analisar se determinado grupo de atributos de um esquema de relação é melhor do que outro.
- O desenho de uma base de dados relacional resulta num conjunto de relações. Existe um objectivo implícito nesse processo de desenho:
  - Preservação da informação
    - **Todos os conceitos capturados pelo desenho conceptual que são mais tarde mapeados para o desenho lógico.**
  - Minimizar a redundância dos dados
    - **Minimizar o armazenamento duplicado de dados em relações distintas, reduzindo a necessidade de múltiplos updates e consequente problema de consistência entre múltiplas cópias da mesma informação.**

2

deti

Desenho de BD - Esquemas de Relação

Análise de Qualidade:

- Critérios Informais
- Critérios Formais
  - Dependências Funcionais, Multivalor e Junção
- Processo de Normalização
  - Formas Normais
    - **Baseadas em critérios formais**

3

deti

Critérios Informais

- Clareza da semântica dos atributos da relação
- Redundância de informação no tuplo
- Redução dos NULLs nos tuplos
- Junção de relações baseada em PK e FK

4

deti

Semântica dos atributos da relação

- O desenho de um esquema de relação deve ser fácil de explicar.
- Verificar se existe uma semântica clara entre os atributos de uma relação.
  - Evitar que uma relação corresponda a uma mistura de atributos de diferentes entidades e relacionamentos.
  - Exemplos de **mau desenho**:

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

EMP_PROJ

Ssn	Pnumber	Hours	Ename	Pname	Plocation

5

deti

Redundância de Informação no Tuplo

- O objectivo é reduzir ao máximo o espaço ocupado por uma relação.
- No mau exemplo anterior verificámos que também há duplicação desnecessária de informação.

Duplicação dos dados do departamento sempre que introduzimos um novo funcionário

Update de dados departamento... ... Update todos os tuplos!!!

Redundancy

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 FireOak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555

Como introduzir um funcionário sem departamento?

Uso de NULLs...

Como Introduzir um novo departamento?

Uso de NULLs...

Ssn=NULL !!!Integridade da Entidade???

EMPLOYEE

Ename	Ssn	Bdate	Address	Dnumber
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4
Wallace, Jennifer S.	987654321	1941-06-20	291Berry, Bellaire, TX	4
Narayan, Ramesh K.	666884444	1962-09-15	975 Fire Oak, Humble, TX	5
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1

DEPARTMENT

Dname	Dnumber	Dmgr_ssn
Research	5	333445555
Administration	4	987654321
Headquarters	1	888665555

6

deti

Redução dos NULLs nos tuplos

- Há situações em que temos uma grande quantidade de atributos numa relação:
  - Muitos dos atributos não se aplicam a todos os tuplos da relação.
- Consequência: existência de muitos NULLs nesses tuplos
  - **Desperdício de espaço**
  - **Difícil interpretação do seu sentido desses atributos (Null pode ter vários significados)**
- Recomendação: Criar outra relação para esses atributos.

  Exemplo:
  - Imaginando que queremos incluir o número do gabinete na relação Employee mas só 15% dos funcionários têm esse número.
  - Solução: criar uma nova relação EMP_OFFICES(Essn, Office_number) só com tuplos de funcionários com gabinete.

7

deti

Junção de Relações baseada em PK e FK

- Devemos evitar esquemas de relação que estabeleçam relacionamentos entre duas relações baseados em atributos que não a chave primária e estrangeira.
- Mau exemplo:

EMP_LOCS

Ename	Plocation
Smith, John B.	Bellaire
Smith, John B.	Sugarland
Narayan, Ramesh K.	Houston
English, Joyce A.	Bellaire
English, Joyce A.	Sugarland
Wong, Franklin T.	Sugarland
Wong, Franklin T.	Houston
Wong, Franklin T.	Stafford

EMP_PROJ1

Ssn	Pnumber	Hours	Pname	Plocation
123456789	1	32.5	ProductX	Bellaire
123456789	2	7.5	ProductY	Sugarland
666884444	3	40.0	ProductZ	Houston
453453453	1	20.0	ProductX	Bellaire
453453453	2	20.0	ProductY	Sugarland
333445555	2	10.0	ProductY	Sugarland
333445555	3	10.0	ProductZ	Houston
333445555	10	10.0	Computerization	Stafford

EMPL_LOCS ⋈ EMP_PROJ1

	Ssn	Pnumber	Hours	Pname	Plocation	Ename
	123456789	1	32.5	ProductX	Bellaire	Smith, John B.
*	123456789	1	32.5	ProductX	Bellaire	English, Joyce A.
	123456789	2	7.5	ProductY	Sugarland	Smith, John B.
*	123456789	2	7.5	ProductY	Sugarland	English, Joyce A.
*	123456789	2	7.5	ProductY	Sugarland	Wong, Franklin T.
	666884444	3	40.0	ProductZ	Houston	Narayan, Ramesh K.
*	666884444	3	40.0	ProductZ	Houston	Wong, Franklin T.
*	453453453	1	20.0	ProductX	Bellaire	Smith, John B.
	453453453	1	20.0	ProductX	Bellaire	English, Joyce A.
*	453453453	2	20.0	ProductY	Sugarland	Smith, John B.
	453453453	2	20.0	ProductY	Sugarland	English, Joyce A.

Temos situações de junção errada de tuplos:

** spurious tuples*

8

deti

Dependências Funcionais

9

deti

Dependências Funcionais (DP)

- Considerando a relação:
  - R(A1, A2, ... , An)
  - Subconjunto de atributos X,Y ⊆ R
- Dependência Funcional: X→Y
  - tuplos: t1, t2 ∈ R
  - t1[X] = t2[X] ⇒ t1[Y] = t2[Y]

Restrição

- Formalismo de análise de esquemas relacionais.
  - Permite descrever restrições dos atributos que os tuplos devem respeitar em todo o momento (invariantes).
  - Permite detectar e descrever problemas com precisão.

10

deti

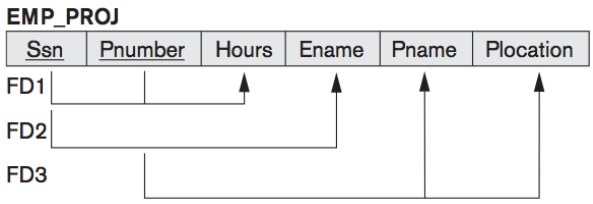
Dependências Funcionais

- X→Y ... por outras palavras:
  - Y é funcionalmente dependente de X.
  - Os valores da componente X do tuplo define de forma única a componente Y do respectivo tuplo.
- Uma DP é uma propriedade do esquema de relação R que não pode ser inferido de uma qualquer instância de R, i.e. r(R).
  - Deve ser definida por alguém que conhece a semântica dos atributos da relação.

11

deti

Dependências Funcionais - Exemplo



- Pela semântica dos atributos da relação EMP_PROJ podemos inferir as seguintes DF:
  - Ssn → Ename
  - Pnumber → {Pname, Plocation}
  - {Ssn, Pnumber} → Hours

O Ssn determina de forma única o nome do funcionários.

O número do projecto determina de forma única o seu nome e localização.

O Ssn e o número do projecto determinam de forma única o número de horas que um funcionário trabalha para o projecto.

FD: Functional Dependency

deti

Tipos de Dependências Funcionais

- Dependência Parcial
  - atributo depende de parte dos atributos que compõem a chave da relação.
- Dependência Total
  - atributo depende de toda a chave da relação.
- Dependência Transitiva
  - atributo que não faz parte da chave da relação depende de um atributo que também não faz parte da chave da relação.

14

deti

Normalização

15

deti

Introdução

- Objectivo: Reduzir a Redundância
- Utilizámos DF para especificar alguns aspectos semânticos do esquema da relação.
  - Mas a redundância está associada a DF não desejadas!
- Vamos assumir que:
  - Existe um conjunto de DF associadas a cada esquema de relação;
  - Que cada relação tem uma chave primária definida;
- Processo de Normalização:
  - Formas Normais
    - **Conjunto de testes (condições) para validação de cada forma.**
    - **Cada forma superior tem menos DF que a anterior.**

16

deti

Formas Normais

- O processo de normalização consiste em efetuar um conjunto de testes para certificar se um desenho de BD relacional satisfaz determinada Forma Normal (FN).
  - Relações que não satisfazem os testes de determinada forma normal são decompostas em relações menores.
- Codd propôs três FN baseadas em DF
  - Primeira (1FN), Segunda (2FN) e Terceira (3FN)
  - A 3FN satisfaz as condições da 2FN e esta as da 1FN
- Mais tarde Boyce e Codd propuseram uma definição mais restritiva da 3NF à qual se chamou:
  - Boyce-Codd Normal Form (BCNF)
- Foram ainda propostas a 4FN e 5FN baseadas respectivamente em dependências multivalor e de junção.

17

deti

Primeira Forma Normal (1NF)

- Definição formal de uma relação básica do modelo relacional:
  - Atributos são atómicos (simples e indivisíveis)
    - **Não permite atributos composto ou multivalor**
  - Não suporta relações dentro de relações (Nested Relation)
    - **Não é possível utilizar uma relação como valor de um atributo de um tuplo.**

EMP_PROJ

		Projs	
Ssn	Ename	Pnumber	Hours

Nested Relation

EMP_PROJ

Ssn	Ename	Pnumber	Hours
123456789	Smith, John B.	1	32.5
		2	7.5
666884444	Narayan, Ramesh K.	3	40.0
453453453	English, Joyce A.	1	20.0
		2	20.0

EMP_PROJ(Ssn, Ename, {PROJS(Pnumber, Hours)})

{ } significa que o atributo PROJS é multivalor

18

deti

1FN - Exemplo 1

(a)
DEPARTMENT

Dname	Dnumber	Dmgr_ssn	Dlocations

Esquema Relação

Não está na 1FN

(b)
DEPARTMENT

Dname	Dnumber	Dmgr_ssn	Dlocations
Research	5	333445555	{Bellaire, Sugarland, Houston}
Administration	4	987654321	{Stafford}
Headquarters	1	888665555	{Houston}

Instância

- Dlocation - atributo multivalor!
- 3 aproximações para converter a relação na 1FN...

19

deti

1FN – Exemplo 1 (soluções)

DEPARTMENT

Dname	Dnumber	Dmgr_ssn
Research	5	333445555
Administration	4	987654321
Headquarters	1	888665555

DEPT_LOCATIONS

Dnumber	Dlocation
1	Houston
4	Stafford
5	Bellaire
5	Sugarland
5	Houston

Aproximação 1

Melhor solução

Decomposição da Relação

DEPARTMENT

Dname	Dnumber	Dmgr_ssn	Dlocation
Research	5	333445555	Bellaire
Research	5	333445555	Sugarland
Research	5	333445555	Houston
Administration	4	987654321	Stafford
Headquarters	1	888665555	Houston

Aproximação 2

Está na1FN mas...

Problema: Redundância

Aproximação 3

Se K é o nº máximo de Dlocation,

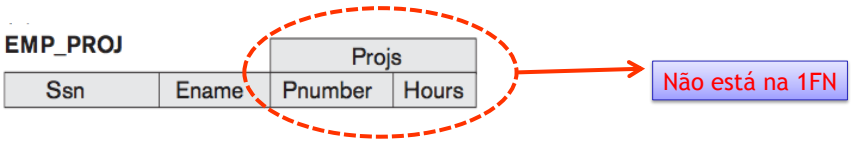
Criar K atributos distintos (Dlocation1, Dlocation2,.., DlocationK)

Problemas: NULL values, consultas Dlocation difíceis, …

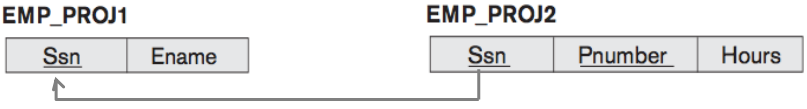
20

deti

1FN – Exemplo 2 (Nested Relation)



- Ssn é chave primária da relação EMP_PROJ
- Pnumber é chave parcial da Nested Relation (Projs)
- Solução:
  - Decompor a relação em duas relações na 1FN:

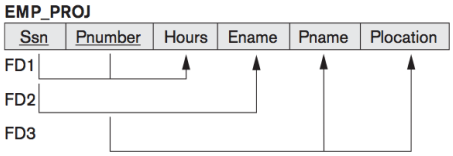


21

deti

Segunda Forma Normal (2FN)

- A relação está na 1FN e…
- …todos os atributos não pertencentes a qualquer chave candidata devem depender totalmente da chave e não de parte dela.
  - i.e. não existem dependências parciais



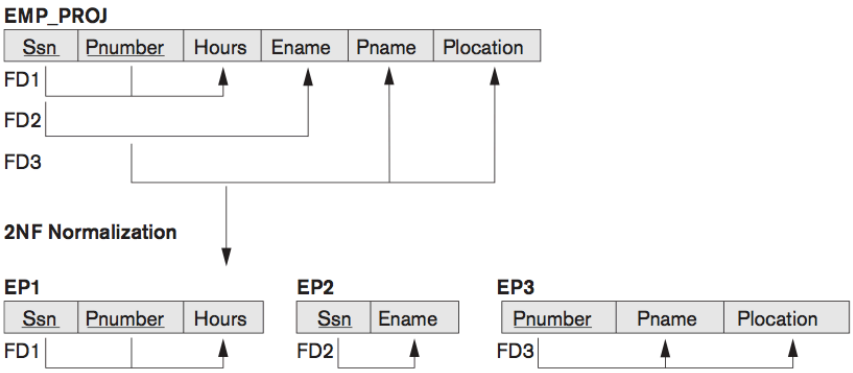
- Exemplo:
  - está na 1FN
  - dependência total:
    - **FD1 ({Ssn,Pnumber} → Hours)**
  - Problema de dependências parciais:
    - **FD2 (Ssn → Ename)**
    - **FD3 (Pnumber → {Pname, Plocation})**

Solução: Decompor a relação...

22

deti

2FN - Exemplo



- Todas as dependência parciais deram resultado a uma nova relação.
- Verificar se as novas relações só têm dependências totais.

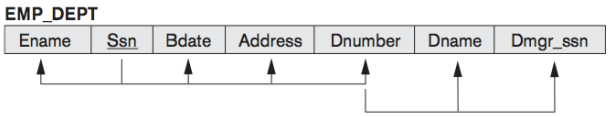
23

deti

Terceira Forma Normal (3FN)

- A relação está na 2FN e…
- …não existem dependências funcionais entre atributos não chave.
  - i.e. não existem dependências transitivas

- Exemplo:
  - está na 2FN
  - Problema de dependências transitiva:
    Ssn → Dnumber e
    Dnumber → Dname
    Dnumber → Dmgr_ssn

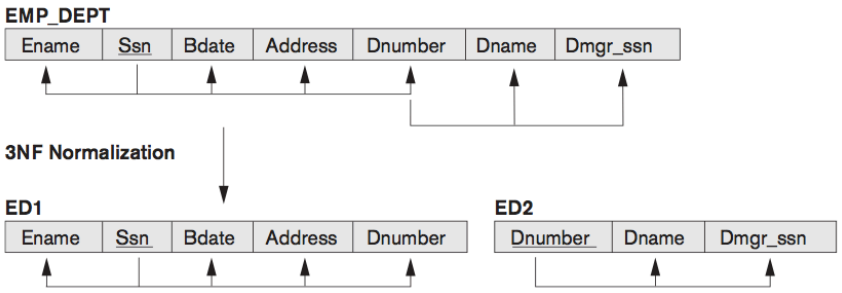


Solução: Decompor a relação...

24

deti

3FN – Exemplo



- As dependências transitivas relativamente a Dnumber deu origem a uma nova relação (ED2) em que Dnumber é a sua chave primária.
- Dnumber mantém-se na relação inicial como chave estrangeira.

25

deti

Quadro Resumo: 1FN, 2FN e 3FN

Summary of Normal Forms Based on Primary Keys and Corresponding Normalization

Normal Form	Test	Remedy (Normalization)
First (1NF)	Relation should have no multivalued attributes or nested relations.	Form new relations for each multivalued attribute or nested relation.
Second (2NF)	For relations where primary key contains multiple attributes, no nonkey attribute should be functionally dependent on a part of the primary key.	Decompose and set up a new relation for each partial key with its dependent attribute(s). Make sure to keep a relation with the original primary key and any attributes that are fully functionally dependent on it.
Third (3NF)	Relation should not have a nonkey attribute functionally determined by another nonkey attribute (or by a set of nonkey attributes). That is, there should be no transitive dependency of a nonkey attribute on the primary key.	Decompose and set up a relation that includes the nonkey attribute(s) that functionally determine(s) other nonkey attribute(s).

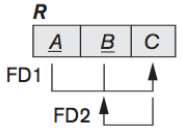
26

deti

Boyce-Codd Normal Form (BCNF)

- Usualmente, a 3FN é aquela que termina o processo de normalização.
  - No entanto, em algumas situações a 3FN ainda apresenta algumas anomalias.
- BCNF é mais restritiva que a 3FN
  - BCNF => 3FN
- Definição:

  Todos os atributos são funcionalmente dependentes da chave da relação, de toda a chave e de nada mais.
- Exemplo:
  - está na 3FN
  - FD2 viola a BCNF

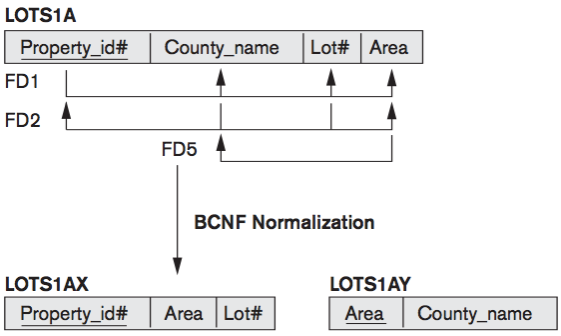


27

deti

BCNF - Exemplo

Base de dados de uma imobiliária:



Chaves candidatas: Property_id# e {County_name, Lot#}

- Relações LOTS1A, LOTS1B e LOTS2 estão na 3FN
- FD5 viola BCNF

Solução: Decomposição de LOTS1A em LOTS1AX e LOTS1AY

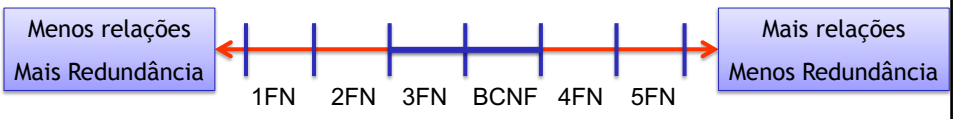
Reverso: Perdemos a FD2

28

deti

Normalização - Ponto de Equilíbrio

- Como verificámos no exemplo de BCNF, perdeu-se uma dependência funcional importante (deduzida da semântica dos atributos).
  - Que deverá ser tratada ao nível aplicacional.
- Assim, existe um ponto de equilíbrio no processo de Normalização que tipicamente fica entre a 3FN e a BCNF.



29

deti

4FN e 5FN

- Usualmente uma relação na BCNF também se encontra na 4FN e 5FN.
  - 4FN são raros e 5FN ainda mais raros
- Definição 4FN:
  - Está na BCNF
  - Não existem dependências multivalor
- Definição 5FN:
  - Está na 4FN
  - A relação não pode ser mais decomposta sem haver perda de informação
  - Não existem dependências de junção

30

deti

Dependências Multivalor

- Dependência multivalor X –» Y em R(X,Y,Z)
- Garantir a seguinte restrição em qualquer instância r(R):
  - Se dois tuplos t1 e t2 existem em r(R) tal que t1[X]=t2[X]
  - Então também devem existir dois tuplos t3 e t4 em r(R) com as seguintes características:
    - **t4[X] = t3[X] = t1[X] = t2[X]**
    - **t3[Y] = t1[Y] e t4[Y] = t2[Y]**
    - **t3[Z] = t2[Z] e t4[Z] = t1[Z]**

X	Y	Z
x1	y1	z1
x1	y2	z2
x1	y1	z2
x1	y2	z1
..	..	..

mesma r(R)

X	Y	Z
x1	y1	z1
x1	y1	z2
x1	y2	z1
x1	y2	z2
..	..	..

- Exemplo:
  - X –» Y
  - X –» Z
- Outras palavras…

  X multidetermina Y se, para cada par de tuplos de R contendo os mesmo valores de X, existe em R um par de tuplos correspondentes à troca dos valores de Y no par original.

31

deti

4FN: Dependências Multivalor - Exemplo

EMP

Ename	Pname	Dname
Smith	X	John
Smith	Y	Anna
Smith	X	Anna
Smith	Y	John

Dependências Multivalor:
Ename –» Pname
Ename –» Dname

- Solução: decomposição da relação EMP

EMP_PROJECTS

Ename	Pname
Smith	X
Smith	Y

EMP_DEPENDENTS

Ename	Dname
Smith	John
Smith	Anna

32

deti

Dependências de Junção

- Existe uma dependência de junção em R se, dadas algumas projeções de R, apenas se reconstrói R através de algumas junções bem definidas, mas não de todas.
- Muito rara na prática
  - difícil de detectar
- Exemplo:
  - Projetando R em (X,Y), (X,Z) e (Y,Z)
  - Verificamos que não é possível reconstruir R por junção de qualquer umas das projeções.
  - Só com a junção das 3 projeções é que conseguimos reconstruir R.

r(R)

X	Y	Z
x1	y1	z1
x1	y1	z2
x1	y2	z2
x2	y3	z2
x2	y4	z2
x2	y4	z4
x2	y5	z4
x3	y2	z5

33

deti

5FN: Dependência Junção - Exemplo

SUPPLY

Sname	Part_name	Proj_name
Smith	Bolt	ProjX
Smith	Nut	ProjY
Adamsky	Bolt	ProjY
Walton	Nut	ProjZ
Adamsky	Nail	ProjX
Adamsky	Bolt	ProjX
Smith	Bolt	ProjY

Vamos Criar 3 Projecções de Supply:

R1(Sname, Part_name)

R2(Sname, Proj_name)

R3(Part_name, Proj_name)

R_1

Sname	Part_name
Smith	Bolt
Smith	Nut
Adamsky	Bolt
Walton	Nut
Adamsky	Nail

R_2

Sname	Proj_name
Smith	ProjX
Smith	ProjY
Adamsky	ProjY
Walton	ProjZ
Adamsky	ProjX

R_3

Part_name	Proj_name
Bolt	ProjX
Nut	ProjY
Bolt	ProjY
Nut	ProjZ
Nail	ProjX

- A relação SUPPLY, com dependência de junção, pode ser decomposta em 3 relações R1, R2 e R3 cada uma na 5FN.
  - Só reconstruímos Supply com a junção das 3 relações R1, R2 e R3.

34

deti

Normalização - Caso de Estudo

Gestão de Encomendas

35

deti

Esquema de Base de Dados - Início

Encomendas				
num_encomenda	num_cliente	cliente	endereco_cliente	...

...			
data_encomenda	cod_produto	produto	quantidade_prod

- É notório que o designer não tem conhecimentos de desenho de base de dados…
- Problemas:
  - Mistura de grupos de atributos de entidades (claramente) distintas.
  - Redundância de informação nos tuplos
    - **Temos de repetir num_encomenda, num_cliente, cliente, endereco_cliente e data_encomenda para registar várias linhas de uma encomenda!**

36

deti

1FN

Encomendas				
num_encomenda	num_cliente	cliente	endereco_cliente	...

...			
data_encomenda	cod_produto	produto	quantidade_prod

- Podemos dizer que existe uma segunda relação Linhas_Encomenda(cod_produto, produto, quantidade_prod) na relação Encomendas.
- Por decomposição:

Encomendas				
num_encomenda	num_cliente	cliente	endereco_cliente	data_encomenda

Linhas_Encomenda			
num_encomenda	cod_produto	produto	quantidade_prod

37

deti

2FN

Encomendas				
num_encomenda	num_cliente	cliente	endereco_cliente	data_encomenda

Linhas_Encomenda			
num_encomenda	cod_produto	produto	quantidade_prod

- Verificámos que na segunda relação há uma violação à 2FN:
  - Dep. Parcial: produto só depende de um atributo (cod_produto) da chave da relação!
- Por decomposição da relação Linhas_Encomenda:

Linhas_Encomenda		
num_encomenda	cod_produto	quantidade_prod

Produtos	
cod_produto	produto

38

deti

3FN

Encomendas				
num_encomenda	num_cliente	cliente	endereco_cliente	data_encomenda

Linhas_Encomenda		
num_encomenda	cod_produto	quantidade_prod

Produtos	
cod_produto	produto

- Verificamos que a relação Encomendas viola a 3FN:
  - Dependência transitiva dos atributos cliente e endereco_cliente!
  - Problemas: redundância, actualização de dados cliente obriga a actualizar N tuplos, só é possível registar um novo cliente quando existir uma primeira encomenda,...
- Por decomposição da relação Encomendas:

Encomendas		
num_encomenda	num_cliente	data_encomenda

Clientes		
num_cliente	cliente	endereco_cliente

39

deti

BCFN

Encomendas		
num_encomenda	num_cliente	data_encomenda

Linhas_Encomenda		
num_encomenda	cod_produto	quantidade_prod

Clientes		
num_cliente	cliente	endereco_cliente

Produtos	
cod_produto	produto

- Já está na BCFN
- Verificamos que todos os atributos só dependem de toda a chave e de nada mais.

40

deti

Resumo

- Qualidade do Desenho de Base de Dados Relacionais
- Critérios Informais
- Dependências Funcionais
- Normalização (Formas Normais)

41