

Normalisatie II

Boyce-Codd normaalvorm

→ Een 3NF-relatieschema $SR = \langle U, F \rangle$ is in Boyce-Codd normaalvorm a.s.a. voor geen enkel sleutelattribuut A van U geldt: A is partieel of transitief functioneel afhankelijk van een of andere kandidaatsleutel van R.

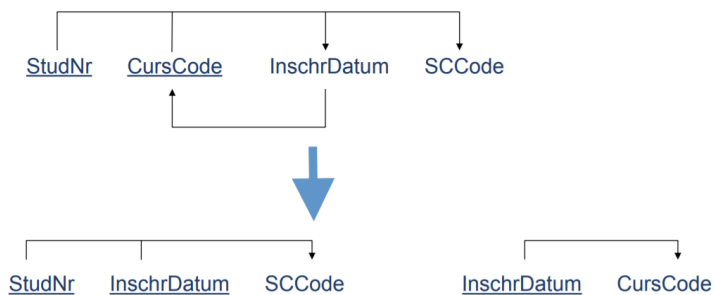
Makkelijkere definitie

→ It should be in 3NF and for any dependency $A \rightarrow B$, A should be a super key.

Oplossing

Stel $A \rightarrow B$ voldoet niet aan deze normaalvorm, maak dan een nieuwe tabel aan voor B met bijhorende functionele afhankelijkheid en behoud de originele tabel maar elimineer B.

Voorbeeld



Voordeel

- Alle ongewenste functionele afhankelijkheden (partieel en transitief, voor alle attributen) zijn weg.

Nadeel

- Niet steeds bereikbaar zonder andere problemen te creëren...
- Sommige functionele afhankelijkheden worden moeilijker te controleren

Ideale decompositie

1. Alle relaties in BCNF
2. Verliesloze decompositie
3. Afhankelijkheden bewaren

Enkel twee uit deze drie haalbaar.

Eerste poging

1. Alle relaties in 3NF
2. Afhankelijkheden bewaren

Formele algoritme (dia 19)

Informele algoritme

1. Minimale overdekking zoeken
2. Voor elke $A_n \rightarrow B_n$: Elke A die voorkomt in de linkerzijde van een functionele afhankelijkheid in een tabel stoppen waarbij A de primaire sleutel is en B de bijhorende attributen zijn.

Voorbeeld

Model#	Serial#	Price	Color	Name	Year
--------	---------	-------	-------	------	------

Dependencies

$\{M, S\} \rightarrow \{P, C\}$
 $\{S\} \rightarrow \{Y\}$
 $\{M\} \rightarrow \{N\}$
 $\{N, Y\} \rightarrow \{P\}$

Minimal cover

$\{M, S\} \rightarrow \{C\}$
 $\{S\} \rightarrow \{Y\}$
 $\{M\} \rightarrow \{N\}$
 $\{N, Y\} \rightarrow \{P\}$

Model#	Serial#	Color
--------	---------	-------

Serial#	Year
---------	------

Model#	Name
--------	------

Name	Year	Price
------	------	-------

Nadeel

- Geen volledig behoud van informatie

Verliesloze decompositie

Definition: Decomposition is lossless if it is feasible to reconstruct relation R from decomposed tables using JOINS. The information will not lose from the relation when decomposed. The JOIN would result in the same original relation.

Testen kan met behulp van de [Chase-algoritme](#):

- maakt gebruik van een matrix
 - kolommen = alle attributen A_j van originele relatie R
 - een rij per relatie R_i in de decompositie
- voor elke rij i:
 - vul a_j in op positie j als A_j in schema S_{R_i} voorkomt; anders b_{ij}
 - de a's stellen de aanwezige waarden voor
- redenering:
 - als $X \rightarrow Y$, en de waarden voor X zijn gelijk in 2 rijen, dan ook die voor Y
 - dit aanduiden in matrix
- doorgaan tot geen verandering meer optreedt (of rij a's)
- rij vol a's $\Leftrightarrow$ verliesloos

Zie dia 29 voor een voorbeeld.

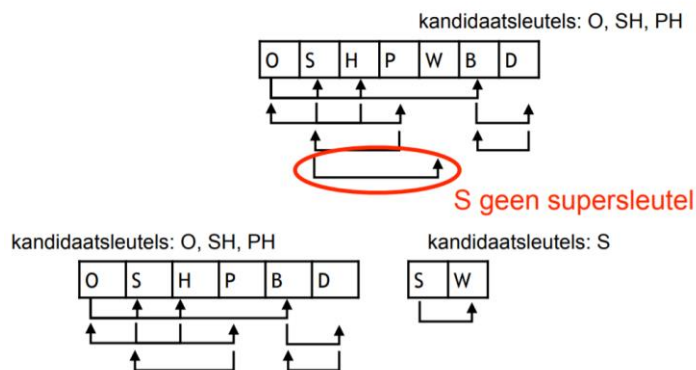
Tweede poging

1. Alle relaties in BCNF
2. Verliesloze decompositie

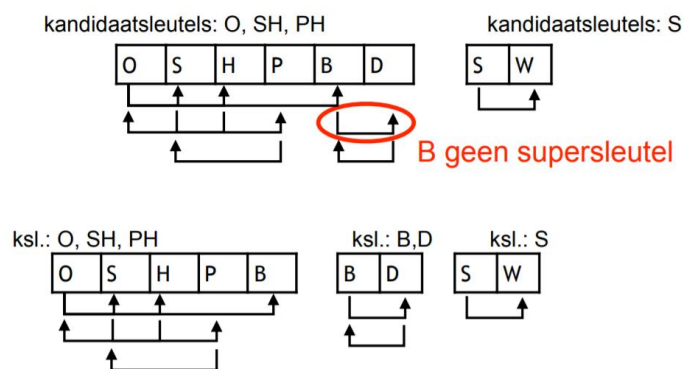
Naar BCNF brengen.

Voorbeeld

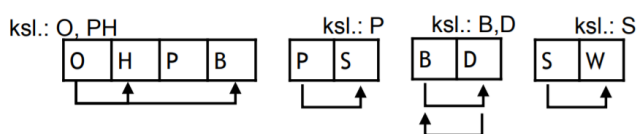
$$F_{R1} = \{ O \rightarrow BSH, B \rightarrow D, D \rightarrow B, SH \rightarrow OP, P \rightarrow S, S \rightarrow W \}$$



$$F_{R1} = \{ O \rightarrow BSH, B \rightarrow D, D \rightarrow B, SH \rightarrow OP, P \rightarrow S, S \rightarrow W \}$$



...



in BCNF!

Derde poging

1. Alle relaties in 3NF
2. Verliesloze decompositie
3. Afhankelijkheden bewaren

Verliesloze en afhankelijkhedenbewarende decompositie naar 3NF

Formele algoritme (dia 47)

Informele algoritme

1. Minimale overdekking zoeken
2. Voor elke $A_n \rightarrow B_n$: Elke A die voorkomt in de linkerzijde van een functionele afhankelijkheid in een tabel stoppen waarbij A de primaire sleutel is en B de bijhorende attributen zijn.
3. Als geen van de relatieschema's een sleutel van R bevat dan maak een extra relatieschema dat attributen bevat die een sleutel voor R vormen.

Voorbeeld

Emp_ssn	Pno	Esal	Ephone	Dno	Pname	Plocation
---------	-----	------	--------	-----	-------	-----------

Afhankelijkheden

$\{Emp_ssn\} \rightarrow \{Esal, Ephone, Dno\}$

$\{Pno\} \rightarrow \{Pname, Plocation\}$

$\{Emp_ssn, Pno\} \rightarrow \{Esal, Ephone, Dno, Pname, Plocation\}$

1. Minimale overdekking

$Emp_ssn \rightarrow Esal$

$Emp_ssn \rightarrow Ephone$

$Emp_ssn \rightarrow Dno$

$Pno \rightarrow Pname$

$Pno \rightarrow Plocation$

2.

Emp_ssn	Esal	Ephone	Dno
---------	------	--------	-----

Pno	Pname	Plocation
-----	-------	-----------

3.

Emp_ssn	Pno
---------	-----

Algoritme voor het vinden van een sleutel

1. Find the attributes that are neither on the left and right side
2. Find attributes that are only on the right side
3. Find attributes that are only on the left side
4. Combine the attributes on step 1 and 3
5. Test if the closures of attributes on step 4 are all the attributes
 - a. If the closure is itself $\rightarrow$ Find [relation exterior](#)¹
 - b. Test each attribute one by one together with step 4.

Voorbeeld

- $R(A, B, C, D, E)$
- $F = \{AB \rightarrow C, CD \rightarrow E, DE \rightarrow B\}$

- Kandidaatsleutels?

(1) $K := R$
(2) Voor elk attribuut A in K
 bereken $(K - A)^+$ t.o.v. F
 als $(K - A)^+$ alle attributen van R bevat, dan $K := K - \{A\}$

- $\{A, B, D\}, \{A, C, D\}, \{A, D, E\}$

¹ Attributes not included in step 2 and 4 together.

Voorbeeld

ABCD

ABC → D

D → A

1. /
2. /
3. BC
4. BC
5. The closure of BC is only BC, we should find the relation exterior.
 - a. Find the relation exteriors, that is the attributes not included in step 4 and step 2. in this example it is AD
 - b. Now test the closures of attributes on step 4 + one attribute in step 6 one at a time.
 ABC closures are ABCD, so it is a candidate key
 BCD closures are ABCD, so it is also a candidate key
 so in this case we have two candidate keys, they are ABC and BCD

Dangling tuples

(a)

EMPLOYEE				
Ename	Ssn	Bdate	Address	Dnum
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	291 Berry, Bellare, TX	4
Narsyan, Ramesh K.	866884444	1982-09-15	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
Berger, Anders C.	999775555	1965-04-26	6530 Braes, Bellare, TX	NULL
Benitez, Carlos M.	888664444	1963-01-09	7654 Beech, Houston, TX	NULL

DEPARTMENT		
Dname	Dnum	Dmgr_ssn
Research	5	333445555
Administration	4	987654321
Headquarters	1	888665555

Figure 11.2

Issues with NULL-value joins. (a) Some EMPLOYEE tuples have NULL for the join attribute Dnum. (b) Result of applying NATURAL JOIN to the EMPLOYEE and DEPARTMENT relations. (c) Result of applying LEFT OUTER JOIN to EMPLOYEE and DEPARTMENT.

(b)

natural join

Ename	Ssn	Bdate	Address	Dnum	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, Bellare, TX	4	Administration	987654321
Narsyan, Ramesh K.	866884444	1982-09-15	975 Fire Oak, 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

(c)

left outer join

Ename	Ssn	Bdate	Address	Dnum	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, Bellare, TX	4	Administration	987654321
Narsyan, Ramesh K.	866884444	1982-09-15	975 Fire Oak, 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
Berger, Anders C.	999775555	1965-04-26	6530 Braes, Bellare, TX	NULL	NULL	NULL
Benitez, Carlos M.	888664444	1963-01-09	7654 Beech, Houston, TX	NULL	NULL	NULL

Nadelen decompositiemethode

- functionele afhankelijkheden moeten vooraf bekend zijn
 - kan moeilijk zijn bij grote gegevensbanken met veel attributen
- de algoritmen zijn niet deterministisch
 - minimale overdekking van F: meerdere oplossingen
 - decompositie is afhankelijk van volgorde waarin functionele afhankelijkheden bekeken worden
- na decompositie (dure) joins nodig om originele informatie terug te vinden
 - kan performantie nadelig beïnvloeden

4^{de} normaalvorm

→ It should be in the Boyce-Codd normal form and the table should not have any multi-valued dependency.

Informal definition: A table is said to have **multi-valued dependency**, if the following conditions are true

1. For a dependency $A \twoheadrightarrow B$, if for a single value of A, multiple value of B exists, then the table may have multi-valued dependency.
2. A table should have at-least 3 columns for it to have a multi-valued dependency.
3. For a relation $R(A,B,C)$, if there is a multi-valued dependency between, A and B, then B and C should be independent of each other.

Voorbeeld

Below we have a college enrolment table with columns **s_id**, **course** and **hobby**.

s_id	course	hobby
1	Science	Cricket
1	Maths	Hockey
2	C#	Cricket
2	Php	Hockey

As you can see in the table above, student with **s_id 1** has opted for two courses, **Science** and **Maths**, and has two hobbies, **Cricket** and **Hockey**.

You must be thinking what problem this can lead to, right?

Well the two records for student with **s_id 1**, will give rise to two more records, as shown below, because for one student, two hobbies exists, hence along with both the courses, these hobbies should be specified.

s_id	course	hobby
1	Science	Cricket
1	Maths	Hockey
1	Science	Hockey
1	Maths	Cricket

And, in the table above, there is no relationship between the columns **course** and **hobby**. They are independent of each other.

So there is multi-value dependency, which leads to un-necessary repetition of data and other anomalies as well.

Oplossing

To make the above relation satisfy the 4th normal form, we can decompose the table into 2 tables.

CourseOpted Table

s_id	course
1	Science
1	Maths
2	C#
2	Php

And, Hobbies Table,

s_id	hobby
1	Cricket
1	Hockey
2	Cricket
2	Hockey

Opmerking: A table can also have functional dependency along with multi-valued dependency. In that case, the functionally dependent columns are moved in a separate table and the multi-valued dependent columns are moved to separate tables.

5^{de} normaalvorm

→ A relation R is in 5NF if and only if it satisfies following conditions:

1. R should be already in 4NF.
2. If breaking down the table does not lead to loss of information or additional information

Definitie join afhankelijkheid: Zie [video](#).

Oplossing

If breaking down the table does not lead to loss of information or additional information, then **decompose** the table.