ΣΧΕΔΙΑΣΜΟΣ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ

Α’ μέρος

Σύνοψη

- Σχεσιακός Σχεδιασμός - Στόχοι
- Κριτήρια / Οδηγίες για ένα “καλό” Σχεδιασμό
- Συναρτησιακές Εξαρτήσεις - Οι Κανονικές Μορφές ΒΔ
- Αποσυνθέσεις Σχέσεων (Decompositions)
- Συνθήκη Διατήρησης Εξαρτήσεων (dependency preservation) και Συνενώσεις άνευ απωλειών (lossless-joins)
- Άλλες Συναρτησιακές Εξαρτήσεις (Multivalued, Join)

Σχεσιακές Βάσεις – Λογικός Σχεδιασμός

- Σχεδιασμός Σχέσεων:

  Ομαδοποίηση Γνωρισμάτων έτσι ώστε να επιτευχθούν “καλά” σχήματα Σχέσεων

- Άτυπα κριτήρια:
  - Προσπάθεια για εννοιολογική καθαρότητα
  - Προσπάθεια για αποδοτικότητα χώρου (αποφυγή πλεονασμού)
  - Προσπάθεια για ακεραιότητα (αποφυγή ανωμαλιών ενημέρωσης)
  - Προσπάθεια για πληρότητα (αποφυγή NULL τιμών σε πλειάδες)
  - Προσπάθεια για γλωσσολογική αποδοτικότητα
  - Προσπάθεια για καλή επίδοση (performance)

Εννοιολογική καθαρότητα

- Ανεπίσημα (και στην ιδανική περίπτωση), **κάθε πλειάδα** πρέπει να παριστά **ακριβώς μια** οντότητα ή **ένα** στιγμιότυπο συσχέτισης
- Με άλλα λόγια, οι πλειάδες δεν πρέπει να είναι **υπερφορτωμένες** με εννοιολογική πληροφορία (δηλαδή, δεν πρέπει να παριστούν περισσότερα του ενός γεγονότα)
- Διακριτές οντότητες δεν πρέπει να αναμιγνύονται – ο μόνος τρόπος για ένδο-αναφορές μεταξύ Σχέσεων πρέπει να είναι το “εξωτερικό κλειδί”

Αποδοτικότητα χώρου

- Αποφυγή πλεονασμού

- Πλεονασμός χώρου (redundancy) σημαίνει χάσιμο χώρου

- Ο πλεονασμός φέρνει **ανωμαλίες**
  (ονομάζονται, ανωμαλίες ενημέρωσης - *update anomalies)*

Προσπάθεια για ακεραιότητα

- Αποφυγή ανωμαλιών ενημέρωσης

- Η ακεραιότητα της Βάσης φθείρεται όταν συμβαίνουν ανωμαλίες εισαγωγής, διαγραφής και τροποποίησης

- Έχουμε ανωμαλία όταν μια ενημέρωση σε ένα σημείο επιφέρει έναν μη-προσδιορίσιμο αριθμό ενημερώσεων σε άλλα σημεία της Βάσης.

Προσπάθεια για πληρότητα

- Αποφυγή NULL τιμών σε πλειάδες
- Οι τιμές Null σημαίνουν "έλλειψη γνώσης" ή "μη-εφαρμοσιμότητα" και συχνά προκαλούν λάθη σε ερωτήσεις
- Για να γίνει σωστή η εκτέλεση των ερωτήσεων απαιτείται διαφορετική λογική από αυτή που χρησιμοποιούμε σε πρακτικά/εμπορικά συστήματα – λογική δύο τιμών αληθείας (true /false)

Προσπάθεια για γλωσσολογική αποδοτικότητα

- Όσο πιο απλά μπορούν να εκφραστούν οι ερωτήσεις στην εφαρμογή – τόσο το καλύτερο για τον προγραμματιστή / χρήστη και (συνήθως) για τον Βελτιστοποιητή Ερωτήσεων του συστήματος

- Οι Ερωτήσεις γίνονται πιο εύκολα σε Σχέσεις που έχουν πολλές πληροφορίες / γνωρίσματα (π.χ., δεν χρειάζονται πολλές συνενώσεις)

Προσπάθεια για καλή επίδοση

- Απόδοση: Αριθμός ερωτημάτων ή ενημερώσεων που μπορούν να υποστούν επεξεργασία ανά μονάδα χρόνου
- Χρόνος απόκρισης: Ο χρόνος που χρειάζεται για να ολοκληρωθεί μια μόνο συναλλαγή από την αρχή ως το τέλος της (στη μέση ή στη χειρότερη περίπτωση)
- Όπως και στην προηγούμενη περίπτωση, Σχέσεις με λίγα Γνωρίσματα (π.χ., δυαδικές), επιφέρουν ένα μεγάλο αριθμό συνενώσεων για την εκτέλεση ερωτήσεων – χειρότερη απόδοση

Παράδειγμα «Κακού» Σχεδιασμού

- Έστω η Βάση με ένα Σχήμα Σχέσης:

ED(<u>SSN</u>, EName, Salary, DNumber, DName, Location, MgrSSN)

Q1: *Βρες όλους τους Υπαλλήλους που βγάζουν πιο πολλά από τον Προϊστάμενό τους*

```
select   e.EName
from     ED e, ED m
where    e.MgrSSN = m.SSN  and e.Salary > m.Salary
```

Q2: *Για κάθε Τμήμα, Βρες τον μέγιστο μισθό*

```
select      DName, max(Salary)
from        ED
group by    DNumber
```

Παράδειγμα «Κακού» Σχεδιασμού (2)

ED(<u>SSN</u>, EName, Salary, DNumber, DName, Location, MgrSSN)

- ΣΧΟΛΙΑ
  - Οι παραπάνω ερωτήσεις γίνονται με απλό τρόπο (**linguistic efficiency**) ΑΛΛΑ:

(α)***Πλεονασμοί***: Οι πληροφορίες ενός Τμήματος **επαναλαμβάνονται για κάθε υπάλληλο** του Τμήματος

(β)***Ανωμαλίες Τροποποιήσεων***: Σε μια απλή τροποποίηση (π.χ., αλλάζοντας τον Προϊστάμενο ενός Τμήματος) ένας **ανεξέλεγκτος αριθμός πλειάδων πρέπει να τροποποιηθεί**

(γ) ***Ανωμαλίες Εισαγωγής***: Πληροφορίες για ένα νέο Τμήμα **δεν είναι δυνατό να εισαχθούν** στη Βάση αν δεν υπάρχει κάποιος Υπάλληλος που (ήδη) εργάζεται στο Τμήμα.

(δ) ***Ανωμαλίες Διαγραφής***: Όταν διαγράφεται και ο τελευταίος υπάλληλος που εργάζεται σε ένα Τμήμα, **χάνουμε τις πληροφορίες για το ίδιο το Τμήμα**

Παράδειγμα «Καλύτερου» Σχεδιασμού

- Μπορούμε να επιλέξουμε έναν **ισοδύναμο** (του προηγουμένου) Σχεδιασμό με δύο Σχέσεις:

EMPLOYEE(<u>SSN</u>, EName, Salary, DNumber)
DEPARTMENT(<u>DNumber</u>, DName, Location, MgrSSN)

Q1: *Βρες όλους τους Υπαλλήλους που βγάζουν πιο πολλά από τον Προϊστάμενό τους*

```
select  e.EName
from    EMPLOYEE e, EMPLOYEE m, DEPARTMENT d
where   e.DNumber = d.DNumber and e.MgrSSN = m.SSN
        and e.Salary > m.Salary
```

Παράδειγμα «Καλύτερου» Σχεδιασμού (2)

Q2: *Για κάθε Τμήμα, Βρες τον μέγιστο μισθό*

```
select      d.DName, max(e.Salary)
from        EMPLOYEE e, DEPARTMENT d
where       d.DNumber = e.DNumber
group by    d.DNumber
```

- Αυτές οι ερωτήσεις είναι σύνθετες **(linguistic inefficiency)** και απαιτούν περισσότερες συνενώσεις **(performance)**, ΑΛΛΑ ΑΠΟΦΕΥΓΟΥΝ ΟΛΕΣ ΤΙΣ ανωμαλίες τροποποίησης **(update anomalies)**
- *Ο Στόχος μας στη συνέχεια είναι να δούμε ΠΩΣ η κατάληξη σε καλούς Σχεδιασμούς μπορεί να γίνει* ***ΤΥΠΙΚΑ*** *και* ***ΣΥΣΤΗΜΑΤΙΚΑ***

Μια Απλή περίπτωση Κανονικοποίησης (α)

- Έστω μια Σχέση για EMPLOYEES
- Μερικοί Υπάλληλοι έχουν περισσότερα του ενός γραφεία
- ΠΩΣ το παρουσιάζουμε αυτό???

EID	LastName	FirstName	Salary	Office
1	Long	Mark	10k	Austin
2	Smith	Lisa	45K	Boston
3	Harris	Dave	60K	Atlanta
4	Davis	James	42K	Dallas
5	Jackson	Harold	35K	Denver

Μια Απλή περίπτωση Κανονικοποίησης (β)

- Μερικοί Υπάλληλοι έχουν περισσότερα του ενός γραφεία
- ΠΩΣ το παρουσιάζουμε αυτό???
- ΛΥΣΗ με πολλές στήλες

(δαπανά χώρο, κάνει δύσκολα μερικά ερωτήματα, κλπ)

EID	LastName	FirstName	Salary	Office1	Office2
1	Long	Mark	10k	Austin	
2	Smith	Lisa	45K	Boston	Seattle
3	Harris	Dave	60K	Atlanta	
4	Davis	James	42K	Dallas	
5	Jackson	Harold	35K	Denver	Miami

Μια Απλή περίπτωση Κανονικοποίησης (γ)

- Μερικοί Υπάλληλοι έχουν περισσότερα του ενός γραφεία
- ΛΥΣΗ με Κανονικοποίηση σε 2 Σχέσεις

Συχνά αποκαλείται: **«Σχέσεις με Κύριο / Εξωτερικό Κλειδί»**

EID	LastName	FirstName	Salary
1	Long	Mark	10k
2	Smith	Lisa	45K
3	Harris	Dave	60K
4	Davis	James	42K
5	Jackson	Harold	35K

EID	Office
1	Austin
2	Boston
2	Seattle
3	Atlanta
4	Dallas
5	Denver
5	Miami

ΣΥΝΑΡΤΗΣΙΑΚΕΣ ΕΞΑΡΤΗΣΕΙΣ

- Οι Συναρτησιακές Εξαρτήσεις (Functional Dependencies – FD) είναι το πιο κοινό **τυπικό** μέτρο του "καλού" για Σχεσιακούς Σχεδιασμούς
- Χρησιμοποιούνται για τον ορισμό των **Κανονικών Μορφών** για Σχέσεις
- Μια FD αποτελεί έναν ***περιορισμό*** για όλα τα στιγμιότυπα της Σχέσης **r(R),** αλλά είναι ***μια ιδιότητα*** των γνωρισμάτων στο Σχήμα **R**

- *ΤΥΠΙΚΟΣ ΟΡΙΣΜΟΣ*: Έστω X, Y σύνολα Γνωρισμάτων στο Σχήμα Σχέσης R. Λέμε ότι η Συναρτησιακή Εξάρτηση (FD):

$$X \rightarrow Y$$

ισχύει αν η X-τιμή *καθορίζει μοναδικά την* Y-τιμή.

FD Ορισμοί

- *ΕΝΑΛΛΑΚΤΙΚΟΣ ΟΡΙΣΜΟΣ*: Έστω X, Y σύνολα Γνωρισμάτων στο Σχήμα Σχέσης R. Λέμε ότι η Συναρτησιακή Εξάρτηση (FD):

  $$X \rightarrow Y$$

  ισχύει αν, όποτε δύο πλειάδες σε ένα στιγμιότυπο r(R) έχουν την ίδια τιμή για το X, τότε *πρέπει να έχουν* την ίδια τιμή για το Y

Δηλαδή:

για οποιεσδήποτε δύο πλειάδες t_1, t_2 σε κάθε σχέση ***r(R):***

Αν $t_1[X] = t_2[X]$ *τότε* $t_1[Y] = t_2[Y]$

- Αν το K είναι **Κλειδί στην R**, τότε το K εξαρτά συναρτησιακά ΟΛΑ τα Γνωρίσματα στην R (μια και δεν μπορούμε ποτέ να έχουμε διακριτές πλειάδες t1, t2 με t1[K] = t2[K])

Ιδιότητες των FD

- ***Τετριμμένη Συναρτησιακή Εξάρτηση:***

  Όποτε $Y \subseteq X$, τότε $X \rightarrow Y$

  Παράδειγμα: SSN, Salary → Salary

- ***Πλήρης Συναρτησιακή Εξάρτηση:***

  Λέμε ότι ένα σύνολο γνωρισμάτων Y είναι **πλήρως εξαρτώμενο συναρτησιακά** από ένα σύνολο X, αν είναι συναρτησιακά εξαρτώμενο από το X **και δεν είναι** συναρτησιακά εξαρτώμενο από κάποιο υποσύνολο του X,

  Δηλαδή, το Y είναι **πλήρως εξαρτώμενο συναρτησιακά** από το X, αν $X \rightarrow Y$ και **δεν υπάρχει** W έτσι ώστε $W \subset X$ and $W \rightarrow Y$

  Παράδειγμα: SSN, PNumber → HoursPW

Ιδιότητες των FD (2)

- ***Μερική Εξάρτηση:*** Σαν συνέπεια του παραπάνω ορισμού, λέμε ότι το Υ είναι **μερικό εξαρτώμενο** από το Χ, αν $X \rightarrow Y$ και δεν υπάρχει κάποιο W έτσι ώστε $W \subset X$ και $W \rightarrow Y$

  Παράδειγμα: SSN,Salary $\rightarrow$ Address (**αλλά και**, SSN $\rightarrow$ Address)

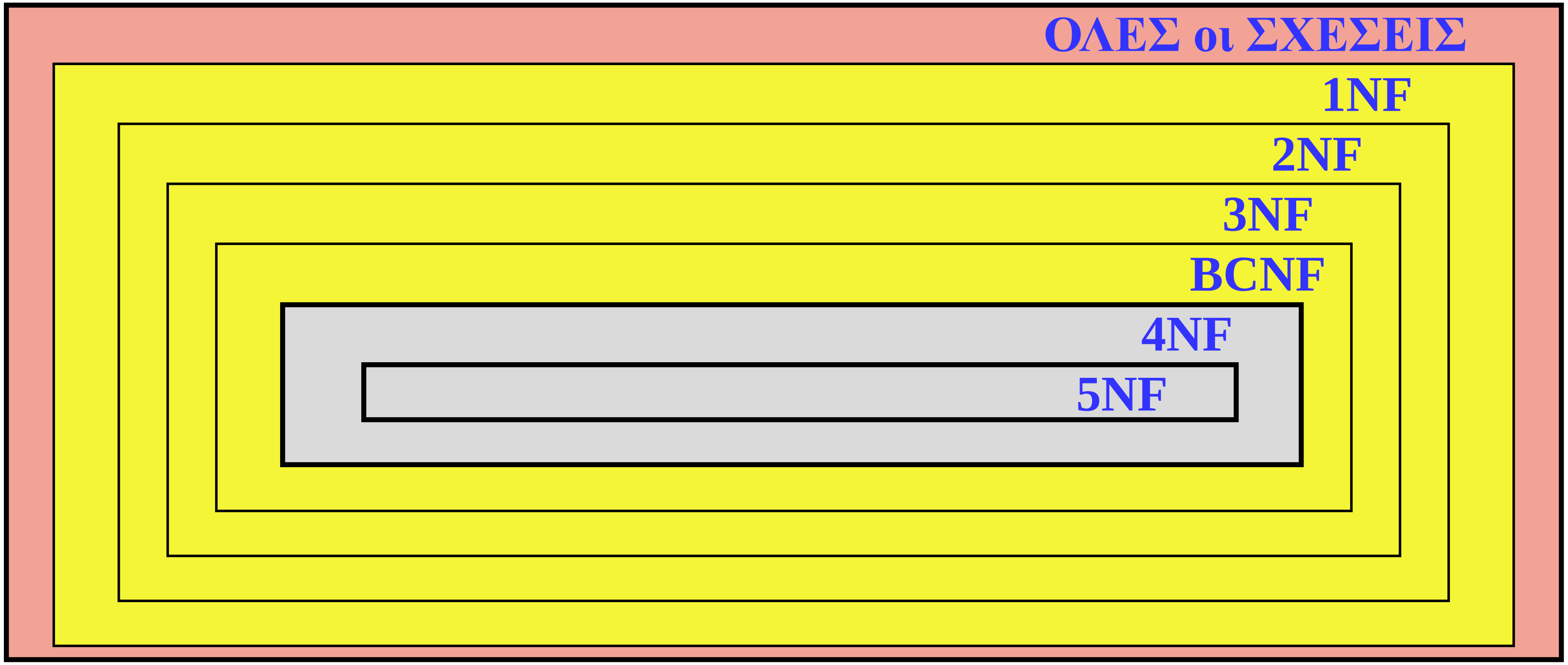
- ***Μεταβατική Εξάρτηση:*** Μια συναρτησιακή εξάρτηση $X \rightarrow Z$ είναι **μεταβατική** αν μπορεί να προέλθει / απορρέει από δύο άλλες FD με μεταβατικότητα ($X \rightarrow Y$ και $Y \rightarrow Z$)

  Παράδειγμα: SSN $\rightarrow$ DNumber **και** DNumber $\rightarrow$ MgrSSN,
  συνεπάγονται: SSN $\rightarrow$ MgrSSN

- ***Πρωτεύον Γνώρισμα:*** Ένα Γνώρισμα που είναι μέλος κάποιου υποψήφιου κλειδιού

ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ

- Η διαδικασία της Κανονικοποίησης ***ενσωματώνει*** την εννοιολογική έννοια της FD στα **Σχήματα Σχέσεων.**
- Υπάρχουν **Κανονικές Μορφές Σχημάτων,** όπου τα παρακάτω αποδεικνύονται:



Κανονικές Μορφές (1)

- **Πρώτη Κανονική Μορφή (1NF):** Η R είναι σε 1NF **αν κάθε Γνώρισμα παίρνει ατομικές / αδιαίρετες τιμές.** Θεωρούμε ότι οι Σχέσεις που δουλεύουμε είναι τουλάχιστον σε 1NF (αυτό σε σύγχρονα DBMS δεν ισχύει – π.χ., πολυμέσα, γεωγραφικές πληροφορίες, κλπ).

  Παράδειγμα: **R(ENumber, ChildrenNames)** **ΔΕΝ ΕΙΝΑΙ** σε 1NF

- **Δεύτερη Κανονική Μορφή (2NF):** Η R είναι σε 2NF **αν είναι σε 1NF και κανένα μη-πρωτεύον Γνώρισμα *δεν είναι μερικώς εξαρτώμενο* από ένα υποψήφιο κλειδί.**

  Παράδειγμα : **SUPPLIER(<u>SNumber</u>, SName, <u>ItemNumber</u>, Price)**
  ΔΕΝ ΕΙΝΑΙ σε 2NF, γιατί ο συνδυασμός <u>SNumber</u>, <u>ItemNumber</u> **είναι υποψήφιο κλειδί, αλλά επίσης** SNumber $\rightarrow$ SName **ισχύει**

Κανονικές Μορφές (2)

- **Τρίτη Κανονική Μορφή (3NF):** Η R είναι σε 3NF **αν είναι σε 2NF και κανένα μη-πρωτεύον Γνώρισμα *δεν είναι μεταβατικά εξαρτώμενο* από ένα υποψήφιο κλειδί**

Παράδειγμα: Η Σχέση:

ED(SSN, EName, Salary, DNumber, DName, Location, MgrSSN)

είναι σε 2NF **ΑΛΛΑ ΔΕΝ ΕΙΝΑΙ** σε 3NF, διότι

Το SSN είναι (το μόνο) υποψήφιο κλειδί και έχουμε:

SSN $\rightarrow$ Dnumber και DNumber $\rightarrow$ MgrSSN

(δηλαδή, το MgrSSN είναι μεταβατικά εξαρτώμενο από το SSN)

Κανονικές Μορφές (3)

- **Ισοδύναμος Ορισμός της ΤΡΙΤΗΣ ΚΑΝΟΝΙΚΗΣ ΜΟΡΦΗΣ (3NF):** Η R είναι σε 3NF **αν και μόνο εάν** για κάθε FD $\mathbf{X} \rightarrow \mathbf{A}$, όπου X είναι ένα σύνολο γνωρισμάτων της R και A είναι απλό Γνώρισμα, **τουλάχιστο ένα από τα παρακάτω τρία ισχύει**

1.- $A \subseteq X$ (η FD είναι τετριμμένη)

2.- $K \subseteq X$ (με το K υποψήφιο κλειδί του R)

3.- $A \subseteq K$ (με το K υποψήφιο κλειδί του R)

Κανονικές Μορφές (4)

- **Κανονική Μορφή BOYCE-CODD (BCNF):** Η R είναι σε BCNF **αν και μόνο εάν όποτε** μια **FD** X → Y **ισχύει τότε το X είναι υποψήφιο κλειδί** (όπου X, Y είναι σύνολα Γνωρισμάτων του R)

- **Ισοδύναμος Ορισμός του BOYCE-CODD NORMAL FORM (BCNF):** Η R είναι σε BCNF **αν και μόνο εάν για κάθε FD** X → A, όπου το X είναι είναι σύνολο γνωρισμάτων του R και A είναι απλό γνώρισμα, **τουλάχιστον ένα από τα παρακάτω ισχύει**:

  1.- $A \subseteq X$ (η FD είναι τετριμμένη)

  2.- $K \subseteq X$ (με το K υποψήφιο κλειδί του R)

  Σημείωση: Ακριβώς όπως η 3NF χωρίς την επιλογή 3., που άμεσα δείχνει ότι η **BCNF συνεπάγεται την 3NF**

Κανονικές Μορφές (5)

Παράδειγμα: Η Σχέση:

RESTAURANT(<u>Client</u>, <u>Food</u>, ReceiptNumber)

είναι σε 3NF **ΑΛΛΑ ΔΕΝ ΕΙΝΑΙ σε** BCNF, διότι ισχύουν:

Client, Food $\rightarrow$ ReceiptNumber, *και*

ReceiptNumber $\rightarrow$ Food (αυτή δεν είναι εξάρτηση από κλειδί)

- Κάθε Σχέση **είναι δυνατόν** να μετασχηματιστεί σε **ισοδύναμες** σχέσεις σε 3NF (χρησιμοποιώντας καλά ορισμένους αλγορίθμους). Η διαδικασία αυτή ονομάζεται ***Αποσύνθεση*** ή ***3NF Κανονικοποίηση.***
- Δυστυχώς, υπάρχουν κάποιες 3NF σχέσεις που **δεν είναι δυνατόν να** μετασχηματιστούν σε BCNF (όπως το χαρακτηριστικό παράδειγμα παραπάνω)

Κανονικες Μορφές Περαν της Τρίτης

- **4η Κανονική Μορφή (4NF)**

Απομόνωση των ανεξάρτητων πολλαπλών σχέσεων μεταξύ δεδομένων

- **5η Κανονική Μορφή (5NF)**

Απομόνωση των εννοιολογικά σχετιζόμενων πολλαπλών σχέσεων μεταξύ δεδομένων

Παράδειγμα 4ης Κανονικής Μορφής (α)

Έστω η Σχέση (για Πωλήσεις Προϊόντων) σε 3 NF

ΠΡΟΙΟΝ	ΧΡΩΜΑ	ΜΕΓΕΘΟΣ
Footer	Κόκκινο	Small
Footer	Κόκκινο	Medium
Footer	Κόκκινο	Large
Footer	Πράσινο	Small
Footer	Πράσινο	Medium
Footer	Πράσινο	Large

Παράδειγμα 4ης Κανονικής Μορφής (β)

σε 4η NF με ΔΥΟ ΠΙΝΑΚΕΣ

ΠΡΟΙΟΝ	ΧΡΩΜΑ
Footer	Κόκκινο
Footer	Πράσινο

ΠΡΟΙΟΝ	ΜΕΓΕΘΟΣ
Footer	Small
Footer	Medium
Footer	Large

Παράδειγμα 5ης Κανονικής Μορφής (α)

Έστω η Σχέση (για Πωλήσεις Προϊόντων) σε 3 NF

ΠΡΟΙΟΝ	ΧΡΩΜΑ	ΜΕΓΕΘΟΣ
Footer	Κόκκινο	Small
Footer	Κόκκινο	Medium
Footer	Κόκκινο	Large
Footer	Πράσινο	Small
Footer	Πράσινο	Medium

Παράδειγμα 5ης Κανονικής Μορφής (β)

Σε 5η NF με ΤΡΕΙΣ ΠΙΝΑΚΕΣ

ΠΡΟΙΟΝ	ΧΡΩΜΑ
Footer	Κόκκινο
Footer	Πράσινο

ΠΡΟΙΟΝ	ΜΕΓΕΘΟΣ
Footer	Small
Footer	Medium
Footer	Large

ΧΡΩΜΑ	ΜΕΓΕΘΟΣ
Κόκκινο	Small
Κόκκινο	Medium
Κόκκινο	Large
Πράσινο	Small
Πράσινο	Medium

Θεωρία Κανονικοποίησης - Κάνονες

- Δοθέντος ενός συνόλου από FD F μπορούμε να προσδιορίσουμε άλλες FD που ισχύουν όποτε οι FD στην F ισχύουν - για να προσδιορίσουμε τέτοιες FD χρειαζόμαστε ***κανόνες συμπερασμού.***

A1. Αν $Y \subseteq X$ τότε $X \rightarrow Y$ (Ανακλαστικός -*Reflexivity*)

A2. Αν $X \rightarrow Y$ τότε $XZ \rightarrow YZ$ (Επαυξητικός -*Augmentation*)

A3. Αν $X \rightarrow Y$ και $Y \rightarrow Z$ τότε $X \rightarrow Z$ (Μεταβατικός - *Transitivity*)

Ο *Armstrong απέδειξε ότι*: {A1, A2, A3) είναι ένα **βάσιμο και πλήρες σύνολο** κανόνων συμπερασμού

δηλαδή, οι κανόνες δημιουργούν μόνο *βάσιμες* FD και επίσης, όλες οι FD που μπορούμε να συμπεράνουμε, *δημιουργούνται με αυτούς τους κανόνες.*

Κανόνες Συμπερασμού

- Επιπλέον Κανόνες Συμπερασμού ισχύουν :

A4. Αν $X \rightarrow YZ$ τότε $X \rightarrow Y$ και $X \rightarrow Z$

(Διασπαστικός - *Decomposition*)

A5. Αν $X \rightarrow Y$ και $X \rightarrow Z$ τότε $X \rightarrow YZ$ (Ενωτικός - *Union*)

A6. Αν $X \rightarrow Y$ και $WY \rightarrow Z$ τότε $WX \rightarrow Z$

(Ψευδομεταβατικός -*Pseudotransitivity*)

- Το ***κλείσιμο - closure*** F+ ενός συνόλου FD F είναι το σύνολο **όλων** των FD που συμπεραίνονται από το F (με την εφαρμογή των κανόνων Armstrong)

- Το ***κλείσιμο - closure*** X+ ενός συνόλου γνωρισμάτων X σε σχέση με ένα σύνολο FD F είναι το σύνολο όλων των γνωρισμάτων που εξαρτώνται συναρτησιακά από το X (με την εφαρμογή των κανόνων Armstrong)

Παραδείγματα

- $R = (A, B, C, G, H, I)$
  $F = \{ A \to B$
  $A \to C$
  $CG \to H$
  $CG \to I$
  $B \to H\}$
- Συναρτησιακές εξαρτήσεις στην F^+
  - $A \to H$
    - » Με μεταβατικότητα $A \to B$ *και* $B \to H$
  - $AG \to I$
    - » Επαυξητικά από το $A \to C$ με το G, παίρνουμε το $AG \to CG$ και μετά με μεταβατικότητα με το $CG \to I$
  - $CG \to HI$
    - » Από το $CG \to H$ *και το* $CG \to I$: "κανόνας ένωσης"

Υπολογισμός του F^+

- Αλγόριθμος:

$F^+ = F$
repeat
    for each functional dependency f in F^+
        apply reflexivity and augmentation rules on f
        add the resulting functional dependencies to F^+
    for each pair of functional dependencies f_1 and f_2 in F^+
        if f_1 and f_2 can be combined using transitivity
            then add the resulting functional dependency to F^+
until F^+ does not change any further

«ΑΡΓΟΣ» Αλγόριθμος

Υπολογισμός Κλεισίματος Συνόλου Γνωρισμάτων (α^+)

- Αλγόριθμος υπολογισμού α^+, με την ύπαρξη της F

result := α;
while (changes to *result*) **do**
for each $\beta \rightarrow \gamma$ **in** F **do**
begin
if $\beta \subseteq$ *result* **then** *result* := *result* $\cup\ \gamma$
end

«ΓΡΗΓΟΡΟΣ» Αλγόριθμος

Παραδείγματα του Κλεισίματος Συνόλου Γνωρισμάτων (α^+)

- $R = (A, B, C, G, H, I)$
- $F = \{A \rightarrow B$
  $A \rightarrow C$
  $CG \rightarrow H$
  $CG \rightarrow I$
  $B \rightarrow H\}$
- $(AG)^+$
  1. *result* = *AG*
  2. *result* = *ABCG* ($A \rightarrow C$ and $A \rightarrow B$)
  3. *result* = *ABCGH* ($CG \rightarrow H$ and $CG \subseteq AGBC$)
  4. *result* = *ABCGHI* ($CG \rightarrow I$ and $CG \subseteq AGBCH$)
- Είναι το *AG* υποψήφιο κλειδί?
  1. Είναι το AG κλειδί?
     1. *Η Ερώτηση AG* $\rightarrow$ *R? Είναι ισοδύναμη με την « είναι το* $(AG)^+ \supseteq R$»?
  2. Είναι κάποιο υποσύνολο του AG κλειδί?
     1. Είναι το $A \rightarrow R$? == Είναι το $(A)^+ \supseteq R$
     2. Είναι το $G \rightarrow R$? == Είναι το $(G)^+ \supseteq R$

Χρήσεις του Κλεισίματος Συνόλου Γνωρισμάτων (α^+)

- ΕΛΕΓΧΟΣ ΚΛΕΙΔΙΟΥ:
  - Για τον έλεγχο αν το α είναι κλειδί, υπολογίζουμε το α^+, και ελέγχουμε αν το α^+ περιέχει όλα τα γνωρίσματα του *R*.
- ΕΛΕΓΧΟΣ ΣΥΝΑΡΤΗΣΙΑΚΩΝ ΕΞΑΡΤΗΣΕΩΝ
  - Για τον έλεγχο αν η συναρτησιακή εξάρτηση $\alpha \rightarrow \beta$ ισχύει (με άλλα λόγια, αν είναι στο F^+), έλεγξε αν $\beta \subseteq \alpha^+$.
  - Δηλαδή, υπολογίζουμε το α^+ με το γνωστό αλγόριθμο, και ελέγχουμε αν περιέχει το β.
  - Είναι ένα εύκολος και γρήγορος έλεγχος
- ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΚΛΕΙΣΙΜΑΤΟΣ ΤΟΥ F
  - Για κάθε $\gamma \subseteq R$, υπολογίζουμε το γ^+, και για κάθε $S \subseteq \gamma^+$, παίρνουμε μια συναρτησιακή εξάρτηση $\gamma \rightarrow S$.

Ισοδυναμία Συνόλων από FD

- Δύο σύνολα από FD F και G είναι ***ισοδύναμα*** αν κάθε FD στο F είναι δυνατόν να παραχθεί από το G και κάθε FD στο G είναι δυνατόν να παραχθεί από το F (δηλαδή, F και G είναι ισοδύναμα αν ισχύει: **F+ = G+**)
- Το F ***καλύπτει το*** G αν κάθε FD στο G είναι δυνατόν να παραχθεί από το F (δηλαδή ισχύει: **G+** $\subseteq$ **F+**)
- Ένα σύνολο F από FD είναι ***ελάχιστη κάλυψη*** αν ικανοποιεί τα παρακάτω:
  (1) Κάθε εξάρτηση στο F είναι της μορφής: $X \rightarrow A$, όπου A είναι ένα απλό γνώρισμα
  (2) Δεν μπορούμε να αποσύρουμε μια FD από το F και να εξακολουθούμε να έχουμε ένα σύνολο ισοδύναμο του F
  (3) Δεν μπορούμε να αντικαταστήσουμε μια FD $X \rightarrow A$ από το F με μια FD $Y \rightarrow A$, όπου $Y \subset X$ και να εξακολουθούμε να έχουμε ένα σύνολο ισοδύναμο του F

Υπολογισμός Ελάχιστης Κάλυψης του F

- Αλγόριθμος:

repeat

Use the union rule to replace any dependencies in *F*

$\alpha_1 \rightarrow \beta_1$ and $\alpha_1 \rightarrow \beta_2$ with $\alpha_1 \rightarrow \beta_1 \beta_2$

Find a functional dependency $\alpha \rightarrow \beta$ with an extraneous attribute either in α or in β

If an extraneous attribute is found, delete it from $\alpha \rightarrow \beta$

until *F* does not change

Παράδειγμα

- $R = (A, B, C)$
  $F = \{A \rightarrow BC$
  $B \rightarrow C$
  $A \rightarrow B$
  $AB \rightarrow C\}$
- Combine $A \rightarrow BC$ and $A \rightarrow B$ into $A \rightarrow BC$
  - Set is now $\{A \rightarrow BC, B \rightarrow C, AB \rightarrow C\}$
- A is extraneous in $AB \rightarrow C$
  - Check if the result of deleting A from $AB \rightarrow C$ is implied by the other dependencies
    - » Yes: in fact, $B \rightarrow C$ is already present!
  - Set is now $\{A \rightarrow BC, A \rightarrow B, B \rightarrow C\}$
- C is extraneous in $A \rightarrow BC$
  - Check if $A \rightarrow C$ is logically implied by $A \rightarrow B$ and the other dependencies
    - » Yes: using transitivity on $A \rightarrow B$ *and* $B \rightarrow$ C.
      - Can use attribute closure of A in more complex cases
- Η ελάχιστη κάλυψη είναι: $A \rightarrow B$
  $B \rightarrow C$

Αποτελέσματα Θεωρίας Συναρτησιακών Εξαρτήσεων

- Υπάρχει **απλός αλγόριθμος** για να ελέγχει την ισοδυναμία μεταξύ συνόλων από FD
- Κάθε σύνολο από FD έχει ένα **ελάχιστο ισοδύναμο** σύνολο
- **Δεν υπάρχει απλός (efficient) αλγόριθμος** για να υπολογίζει το ελάχιστο ισοδύναμο σύνολο από FD που είναι ισοδύναμο με ένα σύνολο F από FD
- Η ύπαρξη (και υπολογισμός) ενός ελάχιστου ισοδύναμου συνόλου είναι **σημαντική** για κάθε αλγόριθμο σχεδιασμού
- Επιπλέον κριτήρια απαιτούνται για να επιτύχουμε έναν **«καλό»** λογικό σχεδιασμό σχέσεων (***ιδιότητα συνενώσεων άνευ απωλειών - lossless join property, ιδιότητα διατήρησης εξαρτήσεων - dependency preserving property***).
- Οι αλγόριθμοι σχεδιασμού πληρούν τέτοιες ιδιότητες

Σχεσιακή Αποσύνθεση

- ΤΟ ΑΡΧΙΚΟ ΣΗΜΕΙΟ όλων των αλγορίθμων είναι ένα ***σχήμα καθολικής σχέσης*** ***R*** που περιέχει ***όλα*** τα γνωρίσματα της Βάσης
- ΣΤΟΧΟΣ του σχεδιασμού είναι **μια αποσύνθεση (decomposition) D του R** σε **m** Σχήματα Σχέσεων $\mathbf{R_1, R_2, R_3, \ldots R_m}$ όπου κάθε $\mathbf{R_i}$ περιέχει ένα υποσύνολο γνωρισμάτων του **R** και κάθε γνώρισμα του **R** πρέπει να παρουσιάζεται σε **τουλάχιστον ένα** Σχήμα Σχέσης $\mathbf{R_i}$ **(ιδιότητα διατήρησης Γνωρισμάτων)**

- Στην ουσία, όταν κάνουμε αποσύνθεση του R, σκοπεύουμε να αποθηκεύσουμε στιγμιότυπα των Σχέσεων που προκύπτουν από την αποσύνθεση, αντί για στιγμιότυπα του R.

Προβλήματα με Αποσυνθέσεις

- **Αν η αποσύνθεση γίνει με τυχαίο τρόπο, παρουσιάζονται προβλήματα:**
  - ☆ Μερικές ερωτήσεις καθίστανται πιο ακριβές (μη-αποδοτικές) .
    - » Π.χ., Οι ερωτήσεις πιθανόν να απαιτούν πολλές Συνενώσεις
  - Δοθέντων στιγμιότυπων των αποσυνθεμένων Σχέσεων, είναι πιθανόν να μη μπορούμε να ανακατασκευάσουμε το στιγμιότυπο της αρχικής Σχέσης!
    - » Η περίπτωση των ¨πλασματικών¨ πλειάδων
  - Για τον έλεγχο κάποιων εξαρτήσεων, πιθανόν να χρειάζεται η Συνένωση των αποσυνθεμένων στιγμιότυπων.
    - » Τα γνωρίσματα στις FD είναι τώρα σε περισσότερες της μιας Σχέσεις

Διατήρηση Εξαρτήσεων (1)

- **ΙΔΙΟΤΗΤΑ ΔΙΑΤΗΡΗΣΗΣ ΕΞΑΡΤΗΣΕΩΝ:** Η αποσύνθεση D πρέπει **να διατηρεί τις συναρτησιακές εξαρτήσεις**; δηλαδή, η συλλογή όλων των εξαρτήσεων που ισχύουν για τις σχέσεις R_i πρέπει να είναι **ισοδύναμη με την** F (οι FD που ισχύουν στην R)
- **ΤΥΠΙΚΑ:**

Ορισμός: Η **ΠΡΟΒΟΛΗ** του F στο R_i, η $\boldsymbol{\Pi_F(R_i)}$, είναι το σύνολο των FD $X \rightarrow Y$ στο F^+ έτσι ώστε: $(X \cup Y) \subseteq R_i$

Μια αποσύνθεση D ικανοποιεί την ιδιότητα διατήρησης εξαρτήσεων αν:

$$(\Pi_F(R_1) \cup \Pi_F(R_2) \ldots \cup \Pi_F(R_m))^+ = F^+$$

Διατήρηση Εξαρτήσεων (2)

- Υπάρχει αλγόριθμος, ο οποίος ονομάζεται ***αλγόριθμος σχεσιακής σύνθεσης,*** που αποσυνθέτει την καθολική σχέση R σε ένα σύνολο σχέσεων $D = \{R_1, R_2, R_3, \ldots R_m\}$ διατηρώντας τις εξαρτήσεις σε σχέση με το F, **έτσι ώστε κάθε** R_i **είναι σε 3NF**

- Ο Αλγόριθμος βασίζεται σε **Ελάχιστες Καλύψεις** και, όπως έχει ειπωθεί και νωρίτερα, ΔΕΝ υπάρχουν αποδοτικοί τρόποι για να βρεθούν ελάχιστες καλύψεις.

Συνενώσεις Άνευ Απωλειών

- **ΙΔΙΟΤΗΤΑ ΣΥΝΕΝΩΣΕΩΝ ΑΝΕΥ ΑΠΩΛΕΙΩΝ:** Η ιδιότητα αυτή φροντίζει ώστε να μη παρουσιάζονται μη-υπαρκτές (πλασματικές) πλειάδες σε Σχέσεις, όταν συνενώνονται αποσυνθεμένες Σχέσεις

Τυπικά, Μια αποσύνθεση $D = \{R_1, R_2, R_3, \ldots R_m\}$ του R έχει την ***ιδιότητα συνενώσεων άνευ απωλειών*** σε σχέση με ένα σύνολο από FD F, αν ***για κάθε*** στιγμιότυπο ***r(R)*** του οποίου οι πλειάδες ικανοποιούν όλες τις FD στο F, έχουμε:

$$(\Pi_{R1}(r(R)) \bowtie \Pi_{R2}(r(R)) \bowtie \ldots \bowtie \Pi_{Rm}(r(R))\ ) = r(R)$$

- Αυτή είναι πολύ σημαντική ιδιότητα για Αποσυνθέσεις, διότι επηρεάζει τον τρόπο που Ερωτήσεις με ΝΟΗΜΑ γίνονται στα Σχεσιακά Σχήματα

Παράδειγμα Πλασματικών Πλειάδων

- Έστω η Βάση Δεδομένων COMPANY και μια Σχέση E_P(SSN, PNumber, HoursPW, EName, PName, Location) που δημιουργείται με την Συνένωση των EMPLOYEE, PROJECT και WORKS_ON (σε στιγμιότυπα) και στην συνέχεια με την Προβολή στα παραπάνω Γνωρίσματα

  Πεισθείτε ότι η αποσύνθεση της E_P στις:

  R1 (EName, Location)

  R2 (SSN, PNumber, HoursPW, PName, Location)

  ΔΕΝ ΙΚΑΝΟΠΟΙΕΙ την ιδιότητα συνένωσης άνευ απωλειών

  ΠΩΣ?: Κάντε προβολές σε στιγμιότυπα του E_P (για R_1, R_2) και συνενώστε τις Προβολές. ΔΕΝ θα πάρετε πίσω το E_P !

Περισσότερα για Lossless Joins

- Η αποσύνθεση του R στα X και Y είναι lossless-join σε σχέση με το F αν και μόνο εάν το κλείσιμο του F περιέχει την:
  - $X \cap Y \rightarrow X$, ή την
  - $X \cap Y \rightarrow Y$
- Συγκεκριμένα, η αποσύνθεση της R σε UV και R - V είναι lossless-join αν η $U \rightarrow V$ ισχύει στην R.

A	B	C
1	2	3
4	5	6
7	2	8

A	B
1	2
4	5
7	2

B	C
2	3
5	6
2	8

A	B	C
1	2	3
4	5	6
7	2	8
1	2	8
7	2	3

Θεωρητικά Αποτελέσματα

- Υπάρχει **απλός αλγόριθμος** για να ελέγχει αν μια αποσύνθεση D ικανοποιεί την ιδιότητα ***lossless join*** σε σχέση με ένα σύνολο F από FD.
- Υπάρχει **απλός αλγόριθμος** για Αποσύνθεση του R σε BCNF Σχέσεις έτσι ώστε η αποσύνθεση D να ικανοποιεί την ιδιότητα ***lossless join*** σε σχέση με ένα σύνολο F από FD στο R
- **ΔΕΝ ΥΠΑΡΧΕΙ αλγόριθμος** για Αποσύνθεση σε BCNF Σχέσεις που ικανοποιούν την ιδιότητα ***διατήρησης εξαρτήσεων***
- **Υπάρχει απλός αλγόριθμος (τροποποίηση του αλγορίθμου σχεσιακής σύνθεσης) για αποσυνθέσεις σε 3NF (όχι BCNF) Σχέσεις που ικανοποιούν και τις δύο ιδιότητες (*διατήρηση εξαρτήσεων και lossless join*)**

Αποσύνθεση σε 3NF

ΑΛΓΟΡΙΘΜΟΣ:

1.- Βρες ένα ***ελάχιστο σύνολο*** *G* από FD ισοδύναμο του F

2.- Για κάθε X μιας FD $X \rightarrow Y$ στο G

Δημιούργησε ένα Σχήμα Σχέσης R_i στο D με τα γνωρίσματα $\{ X \cup A_1 \cup A_2 \cup ..., \cup A_k\}$, όπου κάθε A_j είναι ένα γνώρισμα που υπάρχει σε μια FD του G με το X στα αριστερά

3.- Αν γνωρίσματα του R δεν έχουν μπει σε κάποιο R_i τότε ***δημιούργησε*** μια νέα Σχέση στο D για αυτά τα γνωρίσματα

4.- Αν καμία σχέση στο D δεν περιέχει ένα κλειδί του R, ***δημιούργησε*** μια σχέση που περιέχει ένα κλειδί του R και ***πρόσθεσέ*** την στο D

- Ο Αλγόριθμος δημιουργεί Σχέσεις σε 3NF που ικανοποιούν και τις δύο ιδιότητες.

Επιπρόσθετες Εξαρτήσεις

- Με τη χρήση των FD μπορούμε να φτάσουμε μέχρι 3NF και BCNF

- Άλλες εξαρτήσεις: ***πλειότιμες εξαρτήσεις (Multivalued) (MVD), εξαρτήσεις συνένωσης (Join Dependencies), εξαρτήσεις εγκλεισμού (Inclusion Dependencies), ...***
Αυτές οι εξαρτήσεις μας πάνε σε κανονικές μορφές πέραν των 3NF και BCNF (δηλαδή, 4NF και 5NF)

- ***Πλειότιμες Συναρτήσεις:*** (άτυπος ορισμός) Ένα σύνολο γνωρισμάτων X *καθορίζει πολλαπλά* ένα σύνολο Y **αν η τιμή του** X **καθορίζει ένα σύνολο τιμών για το** Y
(*ανεξάρτητα από τα άλλα γνωρίσματα στη Σχέση)*

Πλειότιμες Εξαρτήσεις

- Μια MVD παρουσιάζεται ως $X \rightarrow \rightarrow Y$
- Υπάρχουν ***βάσιμα* και *πλήρη*** σύνολα κανόνων συμπερασμού για MVDs
- Μια FD είναι ειδική περίπτωση των MVD
- Ένα Σχεσιακό Σχήμα R είναι στην ***τέταρτη κανονική μορφή (4NF)*** σε σχέση με ένα σύνολο συναρτησιακών και πλειότιμων εξαρτήσεων F αν για κάθε μη τετριμμένη πλειότιμη εξάρτηση $X \rightarrow \rightarrow Y$ στην F^+, το X είναι υποψήφιο κλειδί του R
- Δεν υπάρχει **αποδοτικός αλγόριθμος** για την αποσύνθεση του R σε 4NF Σχέσεις, έτσι ώστε η αποσύνθεση να έχει την ιδιότητα lossless join σε σχέση με ένα σύνολο F από FD και MVD στο R

Εξαρτήσεις Συνένωσης

- Μια ***εξάρτηση συνένωσης*** $JD(R_1, R_2, R_3, \ldots R_m)$ είναι ένας δομικός περιορισμός στην R που προσδιορίζει ότι *κάθε νόμιμο στιγμιότυπο* r(R) **πρέπει να έχει μια lossless join αποσύνθεση** στα $R_1, R_2, R_3, \ldots R_m$
- Μια MVD είναι ειδική περίπτωση των JD (όπου m=2)
- Ένα Σχεσιακό Σχήμα R είναι στην ***πέμπτη κανονική μορφή (5NF)*** σε σχέση με ένα σύνολο συναρτησιακών, πλειότιμων και εξαρτήσεων συνένωσης F αν για κάθε μη τετριμμένη $JD(R_1, R_2, \ldots R_m)$ στην F^+, κάθε R_i είναι υπέρ-κλειδί του R
- Η 5NF ονομάζεται επίσης PJNF (Project-Join Normal Form)

Εξαρτήσεις Εγκλεισμού

- FD, MVD και JD ορίζονται **εντός του ιδίου Σχεσιακού Σχήματος** R (δεν συσχετίζουν γνωρίσματα που βρίσκονται σε διαφορετικές σχέσεις)
- Υπάρχουν και άλλες εξαρτήσεις, όπως οι *εξαρτήσεις εγκλεισμού* που χρησιμοποιούνται για την παράσταση της **αναφορικής ακεραιότητας και των ιεραρχιών class / subclass μεταξύ δύο σχέσεων** R καιS
- Μια ***εξάρτηση εγκλεισμού*** **R.X < S.Y** προσδιορίζει ότι σε κάθε χρονικό σημείο, αν τα r(R) και s(S) είναι στιγμιότυπα σχέσεων των R και S, τότε $\Pi_X(\mathbf{r(R)}) \subseteq \Pi_Y(\mathbf{s(S)})$

Πρακτικά Θέματα για την Κανονικοποίηση

- Ένας μεγάλος αριθμός από εμπορικά εργαλεία, δοθέντων ενός συνόλου Σχημάτων Σχέσεων / Γνωρισμάτων και ενός συνόλου συναρτησιακών εξαρτήσεων **δημιουργούν αυτόματα** Σχήματα Σχέσεων σε μορφή 3NF (σπάνια πάνε σε BCNF, 4NF και 5NF)
- Μια άλλη χρήση τέτοιων εργαλείων είναι να **ελέγχουν το επίπεδο Κανονικοποίησης** μιας Σχέσης - γενικά, η χρήση ως **ευρεστικό εργαλείο επιλογής** ενός Σχεδιασμού έναντι κάποιου άλλου
- Υπάρχουν πρακτικά αποτελέσματα της Θεωρίας που επιτρέπουν σε έναν Σχεδιαστή να κάνει ανάλυση της μορφής:
  - *Αν μια σχέση είναι σε 3NF και κάθε υποψήφιο κλειδί αποτελείται ακριβώς από ένα γνώρισμα, τότε είναι και σε 5NF* (Fagin, 1991)

ΣΧΟΛΙΑ

- Η διαδικασία Κανονικοποίησης έχει και **μειονεκτήματα**:
  - ***Δεν είναι δημιουργική*** -- με στόχο τα κριτήρια που αναφέρθηκαν προηγουμένως, δεν υπάρχει τρόπος να δημιουργηθεί μια «καλή» βάση δεδομένων
  - Συνήθως η Κανονικοποίηση γίνεται ***αφού έχουμε κάποιο Σχήμα*** (μας λέει αν είναι «καλό» ή «κακό»)
  - ***Δεν προσφέρει ένα εννοιολογικό σχήμα*** (ασχολείται μόνο με Σχέσεις και Γνωρίσματα)
- Όμως, είναι μια αξιέπαινη και πρακτικά χρήσιμη προσπάθεια να γίνουν με τυπικό και συστηματικό τρόπο πράγματα που τα κάνουμε συνήθως διαισθητικά.