Ασκήσεις σε
(i) Δομές Ευρετηρίων και Οργάνωση Αρχείων
(ii) Κανονικοποίηση

Δεκέμβριος 2016

Άσκηση 1

Θεωρείστε ότι θέλουμε να διαγράψουμε την τιμή 43 στο B+ δέντρο της Εικόνας 1. Η διαγραφή αυτή προκαλεί μείωση της πληρότητας του φύλλου που περιέχει τα 42*, 43*, κάτω του 50%. Θεωρείστε τις παρακάτω τρεις επιλογές:

(1) Συγχώνευση του φύλλου με τον δεξί αδελφό του.

(2) Συγχώνευση του φύλλου με τον αριστερό αδελφό του.

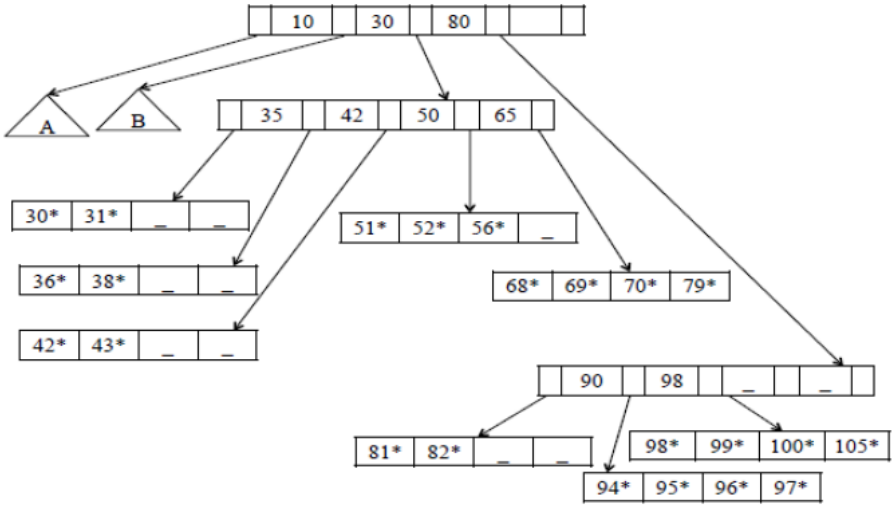
(3) Ανακατανομή των εγγραφών του φύλλου με τον δεξί αδελφό του.

και απαντήστε στις παρακάτω ερωτήσεις

(α) Δώστε τα σωστά B+-δέντρα που προκύπτουν για κάθε μια από τις επιλογές.

(β) Συγκρίνετε τις διαφορετικές επιλογές από άποψη κόστους.

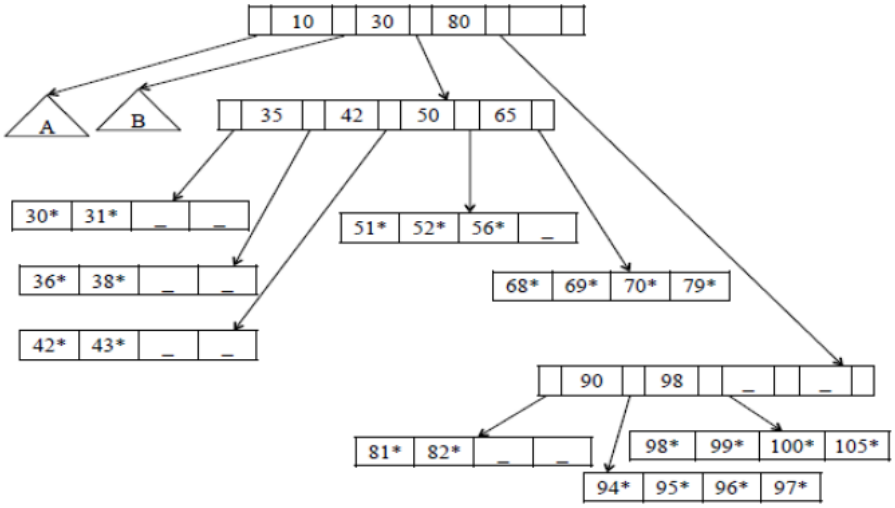
(γ) Υπάρχουν περιπτώσεις που θα προτιμούσατε τη μια από την άλλη επιλογή και γιατί. (Υπόδειξη: σκεφτείτε διαφορετικά ποσοστά εισαγωγών και διαγραφών, πχ, τι γίνετε αν έχουμε μόνο διαγραφές τιμών και τι αν έχουμε ίδιο ποσοστό διαγραφών και εισαγωγών).



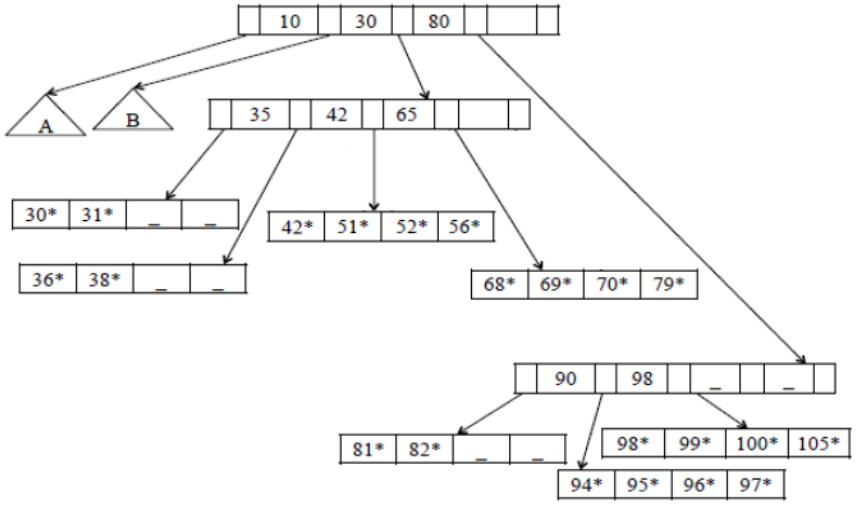
Άσκηση 1

(1) Συγχώνευση του φύλλου με τον δεξί αδελφό του.

Αρχικό B+δέντρο



Μετά τη συγχώνευση

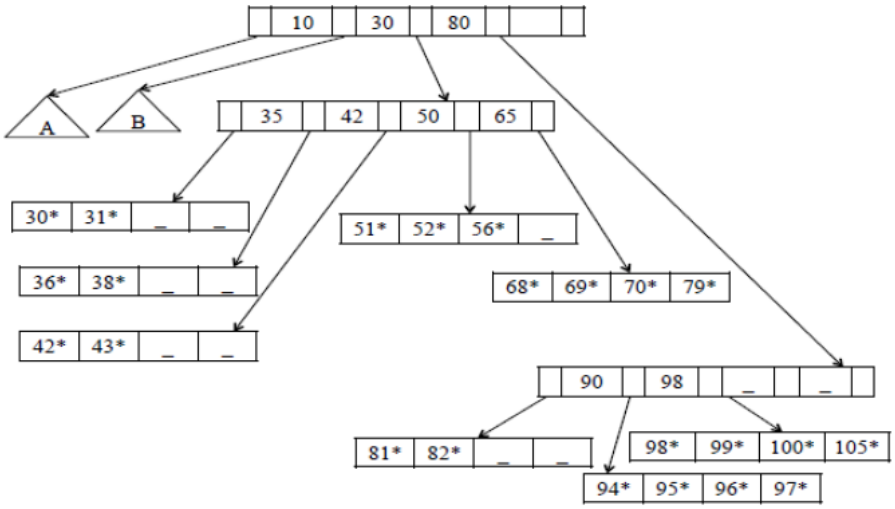


Κόστος ενημέρωσης:
θα διαβαστούν οι κόμβοι από τη ρίζα ως το φύλλο που περιέχει την τιμή 43,
θα διαβαστεί ο δεξιός αδελφός του φύλλου,
έτσι θα έχουμε 4 reads,
θα ενημερωθεί ο κόμβος-αδελφός,
καθώς και ο κοινός γονέας τους,
θα έχουμε 2 writes.
Συνολικά θα γίνουν 6 I/Os.

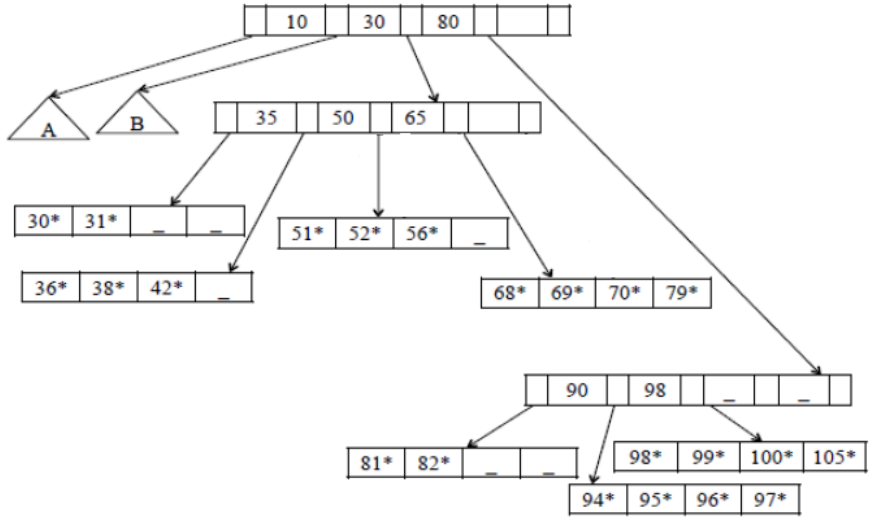
Άσκηση 1

(2) Συγχώνευση του φύλλου με τον αριστερό αδελφό του.

Αρχικό B′+δέντρο



Μετά τη συγχώνευση

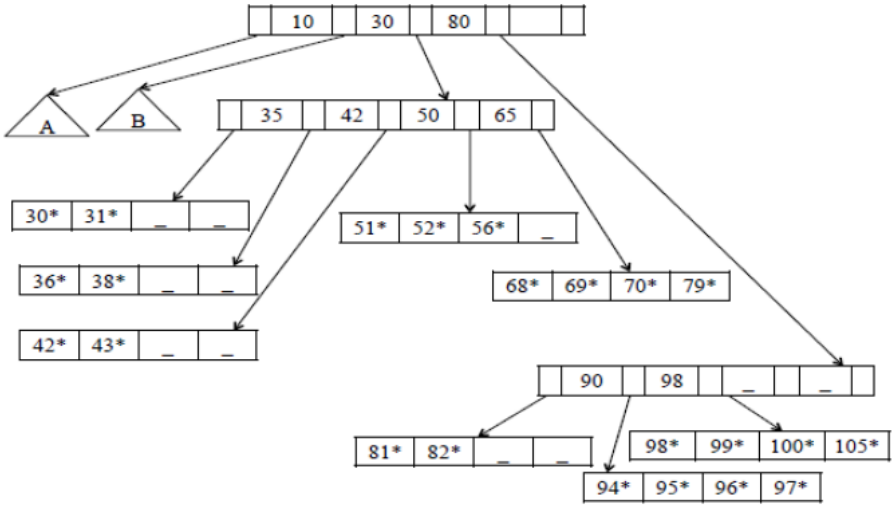


Κόστος ενημέρωσης:
όμοια με (1) 6 I/Os.

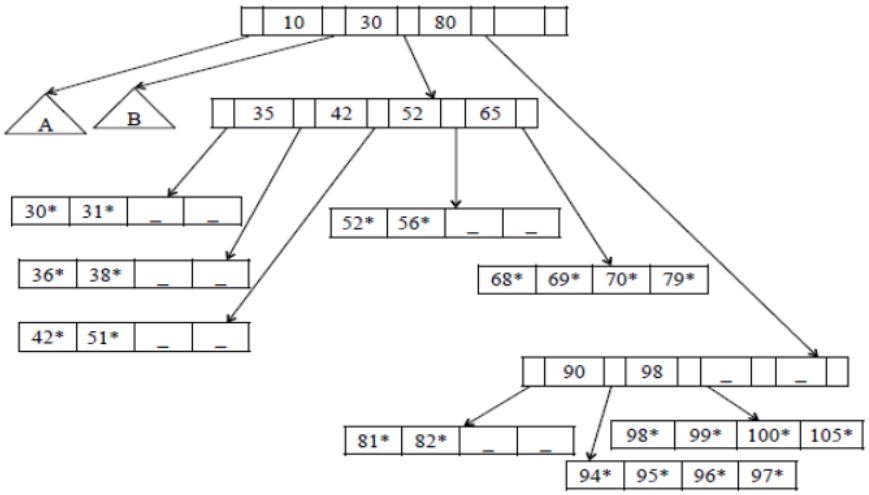
Άσκηση 1

(3) Ανακατανομή των εγγραφών του φύλλου με τον δεξί αδελφό του.

Αρχικό B′+δέντρο



Μετά την ανακατανομή



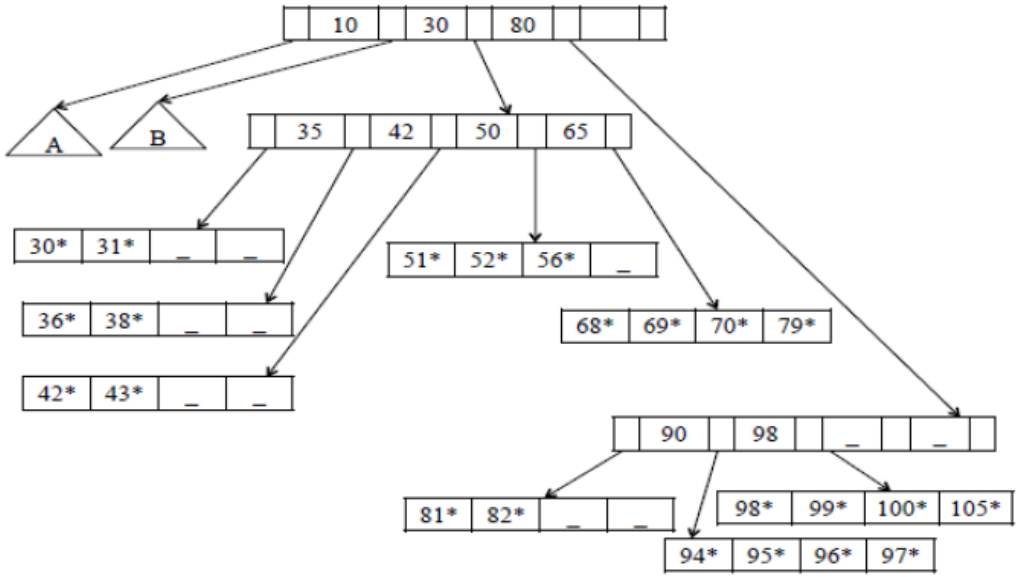
Κόστος ενημέρωσης:
θα διαβαστούν οι κόμβοι από τη ρίζα ως το φύλλο που περιέχει την τιμή 43,
θα διαβαστεί και ο δεξιός αδελφός του,
έτσι θα έχουμε 4 reads,
θα εγγραφούν αυτοί οι δύο κόμβοι-φύλλα κάνοντας ανακατανομή των τιμών τους,
έτσι θα έχουμε 2 writes,
θα ενημερωθεί και ο κόμβος γονέας τους, ώστε τα κλειδιά του να ικανοποιούν τις συνθήκες που πρέπει να ισχύουν στα B+ δένδρα, δηλαδή η μεσαία τιμή κλειδιού θα διορθωθεί στον γονέα,
θα έχουμε άλλο 1 write.
Συνολικά θα γίνουν 7 I/Os.

Άσκηση 1

- Αν περιμένουμε ότι θα γίνουν περισσότερες ***εισαγωγές***
  - *Θέλουμε τα φύλλα να έχουν όσο το δυνατόν περισσότερες κενές θέσεις, ώστε να μην χρειάζεται να γίνονται πολύ συχνά διασπάσεις κόμβων*
  - *Θα προτιμούσαμε την περίπτωση 3.*
- Αν περιμένουμε ότι θα γίνουν περισσότερες ***διαγραφές***
  - *Θέλουμε να μένουν όσο το δυνατόν πιο γεμάτα τα φύλλα.*
  - *Θα προτιμούσαμε τις περιπτώσεις 1 ή 2 επειδή έχουν λιγότερο κόστος ανά διαγραφή και επειδή αφήνουν πιο γεμάτα τα φύλλα.*
- Αν περιμένουμε παρόμοια ποσοστά ***εισαγωγών*** *και* ***διαγραφών***
  - *Οι περιπτώσεις 1 και 2 μειώνουν το πλήθος των κλειδιών του γονέα.*
  - *Αυτό θα προκαλέσει συγχώνευση του γονέα-κόμβου όταν μείνει μισοάδειος, γεγονός που θα επηρεάσει και τον κόμβο γονέα του. Οι συγχωνεύσεις μπορεί να προωθηθούν αναδρομικά προς τα επάνω ως τη ρίζα, μειώνοντας το ύψος του δένδρου.*
  - *Αντίστοιχα μετά από κάποιες εισαγωγές μπορεί να χρειαστεί να γίνει διάσπαση των φύλλων και των κόμβων και να αυξηθεί πάλι το ύψος.*
  - *Θα προτιμούσαμε την περίπτωση 3, για να μην αυξομειώνουμε πολύ συχνά το δένδρο προκαλώντας πολλές λειτουργίες I/O κάθε φορά.*

Άσκηση 2

Θεωρείστε το παρακάτω B+ δέντρο και υποθέστε ότι τα φύλλα των υπό-δέντρων A και B είναι όσο το δυνατόν πιο άδεια. Υποθέστε επίσης ότι το δέντρο ευρετηριοποιεί μια σχέση R, το πεδίο ευρετηριοποίησης είναι υποψήφιο κλειδί και κλειδί διάταξης για την R και ότι κάθε block της R χωράει 20 εγγραφές της. Δώστε μια εκτίμηση για το πόσες πλειάδες έχει η R.



Άσκηση 2

- Θεωρούμε ότι τα υποδένδρα Α και Β έχουν ίδιο ύψος με τα άλλα υποδένδρα που φαίνονται στο σχήμα.

  Έχουν από 3 έως 5 φύλλα το καθένα.
- Κάθε τέτοιο φύλλο είναι όσο το δυνατόν πιο άδειο.

  Άρα, έχει ακριβώς 2 δείκτες προς εγγραφές της σχέσης.
- Η R είναι ταξινομημένη ως προς το υποψήφιο κλειδί.

  Επομένως, αρκεί κάθε δείκτης του ευρετηρίου να δείχνει στην 1η εγγραφή κάθε block.

- Στο υποδένδρο Α θα υπάρχουν από 2*3=6 έως 2*5=10 δείκτες προς τη σχέση.

  Όμοια και για το Β, άλλοι 6 έως 10 δείκτες.

  Στα 8 φύλλα όπως φαίνονται στο σχήμα περιέχονται 23 δείκτες προς blocks της σχέσης.

  Συνολικά από τα φύλλα του ευρετηρίου υπάρχουν από 23+6+6=35 έως 23+10+10=43 δείκτες προς τη σχέση.
- Επομένως η R έχει από 35 έως 43 blocks, δηλαδή από 700 έως 860 εγγραφές.

Άσκηση 3

Θεωρείστε στατικό πίνακα κατακερματισμού (οργάνωση αρχείου) όπου κάθε κάδος χωρά μέρι 2 εγγραφές. Η συνάρτηση κατακερματισμού είναι h(k)=k mod 3.

Ως συνήθως ένας κάδος έχει την χωρητικότητα ενός μπλοκ στον δίσκο.

(i) Δείξτε τα περιεχόμενα του πίνακα κατακερματισμού σε κάθε βήμα κατά την εισαγωγή των ακόλουθων κλειδιών: 27, 5, 18, 30, 10, 32, 38. Για τη διαχείριση υπερχειλίσεων χρησιμοποιείστε αλυσιδωτή σύνδεση (chaining).

(ii) Πόσο κοστίζει (σε I/O) η προσπέλαση της εγγραφής με κλειδί 30 και πόσο η προσπέλαση της εγγραφής με κλειδί 32;

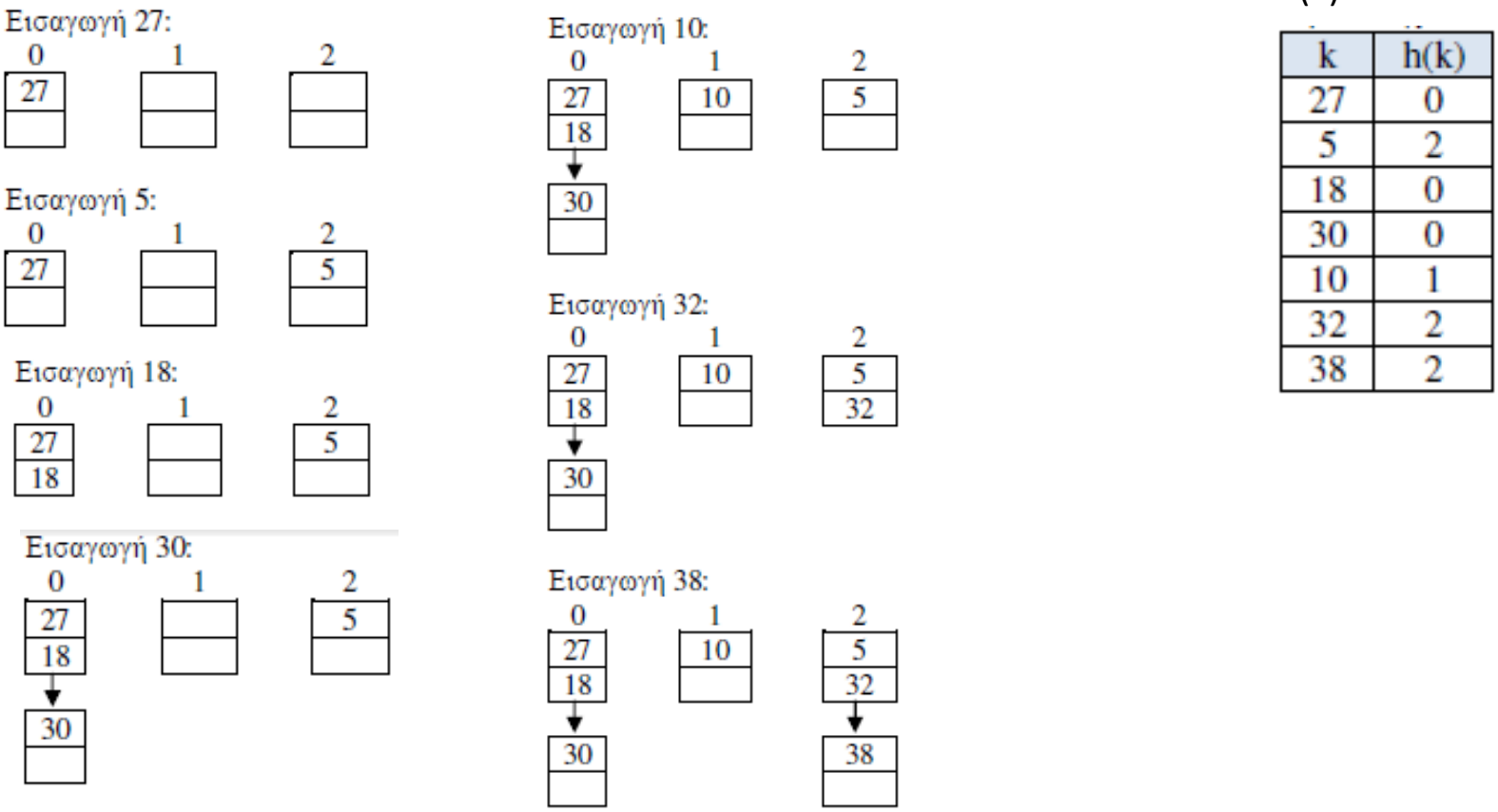
(iii) Ποιος ο μέσος αριθμός I/O μιας αποτυχημένης αναζήτησης;

Άσκηση 3

(i) Δείξτε τα περιεχόμενα του πίνακα κατακερματισμού σε κάθε βήμα κατά την εισαγωγή των ακόλουθων κλειδιών: 27, 5, 18, 30, 10, 32, 38. Για τη διαχείριση υπερχειλίσεων χρησιμοποιείστε αλυσιδωτή σύνδεση (chaining).

h(k)=k mod 3

k	h(k)
27	0
5	2
18	0
30	0
10	1
32	2
38	2



Άσκηση 3

(ii) Πόσο κοστίζει (σε I/O) η προσπέλαση της εγγραφής με κλειδί 30 και πόσο η προσπέλαση της εγγραφής με κλειδί 32;

- Η εγγραφή με τιμή κλειδιού 30 βρίσκεται στην περιοχή υπερχείλισης του bucket 0.
  Συνεπώς απαιτούνται 2 I/O.

- Η εγγραφή με τιμή κλειδιού 32 βρίσκεται στο bucket 2 του πίνακα κατακερματισμού,
  όχι όμως σε περιοχή υπερχείλισης.
  Συνεπώς απαιτείται 1 I/O.

(iii) Ποιος ο μέσος αριθμός I/O μιας αποτυχημένης αναζήτησης;

- Στην περίπτωση αποτυχημένης αναζήτησης:
  - Για κάθε ένα από τα δύο bucket με την περιοχή υπερχείλισης απαιτούνται 2 I/O.
  - Για το bucket χωρίς περιοχή υπερχείλισης απαιτείται 1 I/O.
- Συνολικά, για όλα τα bucket του πίνακα απαιτούνται 5 I/O.
- Υπολογίζοντας τον μέσο όρο έχουμε:

$$\textit{Μέσος αριθμός προσβάσεων σε αποτυχημένη αναζήτηση} = 5/3 = 1.67$$

Άσκηση 4

Έστω ότι έχετε μια σχέση με 100.000 εγγραφές. Επιθυμείτε να κατακερματίσετε τη σχέση σε ένα πίνακα κατακερματισμού με 1.000 κάδους. Ένα μπλοκ δίσκου μπορεί να αποθηκεύσει 100 εγγραφές (μαζί με ένα επιπρόσθετο δείκτη προς ένα μπλοκ υπερχείλισης). Θεωρείστε ότι ένα μπλοκ δίσκου δεν μπορεί να αποθηκεύσει εγγραφές από δύο διαφορετικούς κάδους.

(α) Ποιος είναι ο μέγιστος αριθμός από μπλοκ δίσκου που απαιτούνται για την αποθήκευση της σχέσης;

(β) Ποιος είναι ο ελάχιστος αριθμός από μπλοκ δίσκου που απαιτούνται για τη σχέση;

(γ) Ποια είναι η απάντηση στο ερώτημα (α) αν η σχέση είχε 100.099 εγγραφές;

Άσκηση 4

(α) Ποιος είναι ο μέγιστος αριθμός από μπλοκ δίσκου που απαιτούνται για την αποθήκευση της σχέσης;

- Στην χειρότερη περίπτωση αποθηκεύεται 1 εγγραφή/κάδο στους 999 κάδους, και όλες οι υπόλοιπες 100.000-999 = 99.001 εγγραφές αποθηκεύονται όλες σε ένα κάδο.
- Για τους 999 κάδους δεσμεύεται από 1 block στον καθένα, που θα περιέχει 1 εγγραφή. Συνολικά 999 blocks.
- Στον τελευταίο κάδο θα έχουμε υπερχείλιση.

  Επειδή δεν επιτρέπεται κάποιο block δίσκου να περιέχει εγγραφές από διαφορετικούς κάδους, δεν μπορούμε να έχουμε ανοιχτή διευθυνσιοδότηση, ούτε πολλαπλό κατακερματισμό.

  Έτσι θα έχουμε αλυσιδωτή σύνδεση με blocks υπερχείλισης στον κάδο αυτό.

  Για να αποθηκευθούν οι υπόλοιπες εγγραφές θα χρειαστούν (99.001 / 100) = 991 blocks. (990 γεμάτα και 1 με μια εγγραφή).
- Συνολικά για την αποθήκευση της σχέσης θα χρειαστούν το πολύ 999 + 991 = 1.990 blocks.

(β) Ποιος είναι ο ελάχιστος αριθμός από μπλοκ δίσκου που απαιτούνται για τη σχέση;

- Στην καλύτερη περίπτωση όλες οι εγγραφές μπορούν να κατανεμηθούν ομοιόμορφα στους 1.000 κάδους, θα έχουμε 100.000 / 1.000 = 100 εγγραφές/κάδο.

  Αρκεί 1 block ανά κάδο, που θα είναι γεμάτο με 100 εγγραφές.
- Συνολικά αρκούν 1.000 blocks (όσα και οι κάδοι).

Άσκηση 4

(γ) Ποια είναι η απάντηση στο ερώτημα (α) αν η σχέση είχε 100.099 εγγραφές;

- Θα αποθηκευθεί 1 εγγραφή/κάδο σε 999 κάδους, και όλες οι υπόλοιπες 100.099-999 = 99.100 εγγραφές αντιστοιχούν όλες σε ένα κάδο.
- Για τους 999 κάδους δεσμεύεται από 1 block στον καθένα, που θα περιέχει 1 εγγραφή. Συνολικά 999 blocks.
- Στον τελευταίο κάδο θα έχουμε υπερχείλιση και θα κάνουμε αλυσιδωτή σύνδεση με blocks υπερχείλισης.

  Για να αποθηκευθούν οι υπόλοιπες εγγραφές θα χρειαστούν (99.100 / 100) = 991 blocks. (όλα γεμάτα με εγγραφές).
- Συνολικά για την αποθήκευση της σχέσης θα χρειαστούν και σε αυτή την περίπτωση το πολύ 1.990 blocks.

Άσκηση 5

Χρησιμοποιείστε τα αξιώματα του Armstrong για να αποδείξετε την ορθότητα των ακόλουθων κανόνων:

- Ένωση: $if\ \alpha \to \beta\ and\ \alpha \to \gamma, then\ \alpha \to \beta\gamma$
- Αποσύνθεση: $if\ \alpha \to \beta\gamma, then\ \alpha \to \beta\ and\ \alpha \to \gamma$
- Ψευδο-μεταβατικότητα: $if\ \alpha \to \beta\ and\ \gamma\beta \to \delta, then\ \alpha\gamma \to \delta$

Αξιώματα Armstrong

- Αντανακλαστικότητα: Αν το α είναι ένα σύνολο από γνωρίσματα, και $\beta \subseteq \alpha$, τότε $\alpha \to \beta$.

- Προσαύξηση: Αν $\alpha \to \beta$, και γ είναι ένα σύνολο γνωρισμάτων, τότε $\gamma\alpha \to \gamma\beta$.

- Μεταβατικότητα: Αν $\alpha \to \beta$ και $\beta \to \gamma$, τότε $\alpha \to \gamma$.

Ένωση: $if\ \alpha \to \beta\ and\ \alpha \to \gamma, then\ \alpha \to \beta\gamma$
$\alpha \to \beta$ (δίνεται)
$\alpha\alpha \to \alpha\beta$ (προσαύξηση)
$\alpha \to \alpha\beta$ (ένωση των κοινών γνωρισμάτων)
$\alpha \to \gamma$ (δίνεται)
$\alpha\beta \to \gamma\beta$ (προσαύξηση)
$\alpha \to \beta\gamma$ (μεταβατικότητα και αντιμετάθεση γνωρισμάτων)

Άσκηση 5

Αξιώματα Armstrong

- Αντανακλαστικότητα: Αν το α είναι ένα σύνολο από γνωρίσματα, και $\beta \subseteq \alpha$, τότε $\alpha \rightarrow \beta$.

- Προσαύξηση: Αν $\alpha \rightarrow \beta$, και γ είναι ένα σύνολο γνωρισμάτων, τότε $\gamma\alpha \rightarrow \gamma\beta$.

- Μεταβατικότητα: Αν $\alpha \rightarrow \beta$ και $\beta \rightarrow \gamma$, τότε $\alpha \rightarrow \gamma$.

Αποσύνθεση: $if\ \alpha \rightarrow \beta\gamma, then\ \alpha \rightarrow \beta\ and\ \alpha \rightarrow \gamma$
$\alpha \rightarrow \beta\gamma$ (δίνεται)
$\beta\gamma \rightarrow \beta$ (αντανακλαστικότητα)
$\alpha \rightarrow \beta$ (μεταβατικότητα)
$\beta\gamma \rightarrow \gamma$ (αντανακλαστικότητα)
$\alpha \rightarrow \gamma$ (μεταβατικότητα)

Ψευδο-μεταβατικότητα: $if\ \alpha \rightarrow \beta\ and\ \gamma\beta \rightarrow \delta, then\ \alpha\gamma \rightarrow \delta$
$\alpha \rightarrow \beta$ (δίνεται)
$\alpha\gamma \rightarrow \gamma\beta$ (προσαύξηση και αντιμετάθεση γνωρισμάτων)
$\gamma\beta \rightarrow \delta$ (δίνεται)
$\alpha\gamma \rightarrow \delta$ (μεταβατικότητα)

Άσκηση 6

Δίνεται το σχήμα R(A, B, C, D, E). Δείξτε ότι η παρακάτω αποσύνθεση δεν είναι lossless-join:

R1 (A, B, C)

R2 (C, D, E).

Υπόδειξη: Δώστε ένα παράδειγμα ενός στιγμιοτύπου του R όπου δεν ισχύει η ικανή και αναγκαία συνθήκη για lossless-join decomposition.

Άσκηση 6

Έστω r ένα στιγμιότυπο της R(A, B, C, D, E).
Η αποσύνθεση της R στις R1 (A, B, C) και R2 (C, D, E) είναι lossless-join αν για κάθε r ισχύει: $\Pi_{R1}(\mathrm{r}) \bowtie \Pi_{R2}(\mathrm{r}) = \mathrm{r}$.

Το ακόλουθο στιγμιότυπο r αποτελεί ένα αντιπαράδειγμα:

A	***B***	***C***	***D***	***E***
$\boldsymbol{a_1}$	b_1	c_1	d_1	e_1
$\boldsymbol{a_2}$	b_2	c_1	d_2	e_2

Το $\Pi_{R1}(\mathrm{r})$ είναι:

A	***B***	***C***
$\boldsymbol{a_1}$	b_1	c_1
$\boldsymbol{a_2}$	b_2	c_1

Το $\Pi_{R2}(\mathrm{r})$ είναι:

C	***D***	***E***
$\boldsymbol{c_1}$	d_1	e_1
$\boldsymbol{c_1}$	d_2	e_2

Υπολογίζουμε το $\Pi_{R1}(\mathrm{r}) \bowtie \Pi_{R2}(\mathrm{r})$:

A	***B***	***C***	***D***	***E***
$\boldsymbol{a_1}$	b_1	c_1	d_1	e_1
$\boldsymbol{a_1}$	b_1	c_1	d_2	e_2
$\boldsymbol{a_2}$	b_2	c_1	d_1	e_1
$\boldsymbol{a_2}$	b_2	c_1	d_2	e_2

Παρατηρούμε ότι $\Pi_{R1}(\mathrm{r}) \bowtie \Pi_{R2}(\mathrm{r}) \neq \mathrm{r}$.
Συνεπώς, δεν είναι lossless-join decomposition.