Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2021-2022

Πρώτη Σειρά Ασκήσεων

Ανάθεση: 01-04-2022
Παράδοση: 10-04-2022 Ώρα (23:55)

Οδηγίες

- *Η πρώτη σειρά ασκήσεων είναι **ατομική** και **υποχρεωτική**.*
- *Η υποβολή της εργασίας πρέπει να γίνει στο eclass.*
- *Το παραδοτέο σας θα πρέπει να είναι ένα αρχείο PDF με όνομα AM.pdf (όπου ΑΜ είναι ο αριθμός μητρώου σας. π.χ. "3190001.pdf").*
- *Τα διαγράμματα πρέπει να είναι κατασκευασμένα σε κάποιο πρόγραμμα (της επιλογής σας) και όχι σκαναρισμένα χειρόγραφα.*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*
- ***Για την επίλυση των ασκήσεων να μελετήσετε τις διαφάνειες των διαλέξεων του μαθήματος.***

  Η συνολική βαθμολογία των ασκήσεων ανέρχεται σε 105 μονάδες (100+5 μονάδες bonus).

Άσκηση 1 [15 μονάδες]

Ένας σκληρός δίσκος έχει τα παρακάτω χαρακτηριστικά:

- Μέγεθος τομέα (sector) 1024 bytes
- 1000 ίχνη (tracks) ανά επιφάνεια (surface)
- 10 πλακέτες (platters) διπλής όψης
- 100 τομείς ανά ίχνος
- Μέσος Χρόνος Μετακίνησης Κεφαλής 8 ms
- Ταχύτητα περιστροφής 7200 rpm
- Μέγεθος μπλοκ (σελίδας) 2048

Θέλουμε να αποθηκεύσουμε στον παραπάνω δίσκο ένα αρχείο με 120000 εγγραφές μεγέθους 100 bytes έκαστη.

Δεδομένου ότι κάθε εγγραφή αποθηκεύεται ολόκληρη σε ένα μπλοκ και ένα μπλοκ δεν μπορεί να εκτείνεται σε δύο ίχνη να υπολογίσετε τα ακόλουθα:

1. Τον αριθμό των εγγραφών που χωρούν σε ένα μπλοκ.
2. Τον συνολικό αριθμό των απαιτούμενων μπλοκ για την αποθήκευση του αρχείου.
3. Τον αριθμό των κυλίνδρων που καταλαμβάνει το αρχείο αν τα μπλοκ του αποθηκεύονται συνεχόμενα στον δίσκο.
4. Τον συνολικό αριθμό των εγγραφών που χωράνε στον δίσκο.
5. Τον απαιτούμενο χρόνο για την ανάγνωση ολόκληρου του αρχείου δεδομένου ότι τα μπλοκ του αποθηκεύονται συνεχόμενα στο δίσκο. Για λόγους απλούστευσης θεωρείστε ότι ο χρόνος μετακίνησης στο επόμενο ίχνος είναι μηδαμινός (δηλαδή δεν λαμβάνεται υπόψη).
6. Τον απαιτούμενο χρόνο για την ανάγνωση 10000 τυχαίων εγγραφών του αρχείου.

Άσκηση 2 [Μονάδες 20]

Θεωρείστε την σχέση R(a,b,c,d) στο γνώρισμα **b** της οποία έχουμε δημιουργήσει ένα δευτερεύων ευρετήριο (Secondary Index), έστω I1. Σε ένα μπλοκ (σελίδα) του δίσκου, στον οποίο αποθηκεύεται η σχέση R και το ευρετήριο I1, χωράνε 10 εγγραφές της σχέσης R, ή 50 ζεύγη τιμών κλειδιού-δείκτη (Key-pointer) ή 200 δείκτες (pointers). Δεδομένου ότι το ευρετήριο I1 έχει την δομή που παρουσιάζεται στην **σελίδα 51 (αριθμός διαφάνειας 79)** του αρχείου **"03-Indexing.pdf"** (βλέπε Έγγραφα\03-Indexing.pdf) ζητείται:

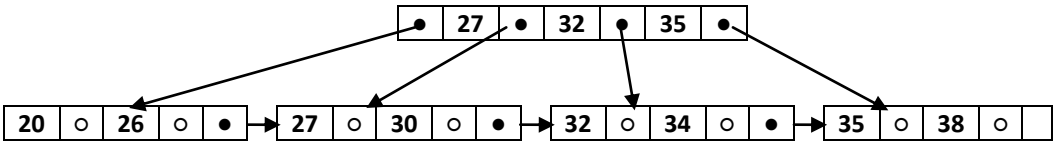
α) Να υπολογίσετε τον αριθμό των μπλοκ που απαιτούνται για την αποθήκευση 6000 εγγραφών της σχέσης R και του αντίστοιχου ευρετηρίου I1. Να υποθέσετε ότι μία συγκεκριμένη τιμή του γνωρίσματος b εμφανίζεται κατά μέσο όρο σε 15 εγγραφές της σχέσης R.

β) Να υπολογίσετε τον **ελάχιστο** και τον **μέγιστο** αριθμό των μπλοκ που απαιτούνται για την αποθήκευση 6000 εγγραφών της σχέσης R και του ευρετηρίου I1, θεωρώντας ότι δεν υπάρχει κανένας περιορισμός στον αριθμό των εγγραφών της σχέσης R που μπορεί να έχουν μια συγκεκριμένη τιμή του γνωρίσματος b.

γ) Βάσει της παραδοχής του ερωτήματος (α) να υπολογίσετε τον **μέσο αριθμό των μπλοκ (σελίδων)** που πρέπει να διαβαστούν για τον εντοπισμό και την ανάκτηση των 15 εγγραφών που αντιστοιχούν σε μια συγκεκριμένη τιμή του γνωρίσματος b. Υποθέστε ότι η μνήμη δεν περιέχει αρχικά τίποτα και ότι η αναζήτηση στο ευρετήριο γίνεται **σειριακά** μέχρι να βρεθεί η τιμή που αναζητούμε η να διαπιστωθεί ότι αυτή δεν υπάρχει.

Άσκηση 3 [20 Μονάδες]

Δίνεται το παρακάτω B+ δέντρο με μέγιστο αριθμό **τριών** κλειδιών **(n=3)** ανά κόμβο/φύλλο.



● **Δείκτης προς κόμβο του δένδρου**

○ **Δείκτης δεδομένων**

Δείκτης δένδρου με τιμή NULL

Ζητείται:

α) Να εισαγάγετε με την σειρά που δίνονται τις τιμές **25,19,23,22**. Σε κάθε εισαγωγή να παρουσιάσετε τη νέα μορφή του δέντρου και να εξηγήσετε πως ακριβώς προέκυψε.

β) Στο δέντρο που θα προκύψει μετά την εισαγωγή των τιμών του παραπάνω ερωτήματος **πόσοι** και **ποιοί** κόμβοι πρέπει να προσπελαστούν για να ανακτηθούν όλες οι εγγραφές με κλειδί αναζήτησης **key>=22 AND key<=34;** i) όταν το κλειδί αναζήτησης είναι μοναδικό και ii) όταν το κλειδί αναζήτησης δεν είναι μοναδικό.

Άσκηση 4 [Μονάδες 25]

Θεωρείστε ένα δίσκο με μέγεθος μπλοκ **B=2048** byte στο οποίο είναι αποθηκευμένο ένα αρχείο το οποίο περιέχει **200000** εγγραφές προϊόντων. Κάθε εγγραφή του αρχείου προσδιορίζεται μοναδικά από το πεδίο BARCODE. Υποθέστε ότι το αρχείο δεν είναι διατεταγμένο ως προς το κλειδί και θέλουμε να δημιουργήσουμε ένα ευρετήριο **B+** δέντρου πάνω στο πεδίο κλειδί (BARCODE) το οποίο έχει μέγεθος **N=20** bytes.

Ένας δείκτης δέντρου **(p)** έχει μέγεθος 7 bytes και το μέγεθος ενός δείκτη δεδομένων **(q)** είναι 9 bytes.

Ζητείται να υπολογίσετε:

α) τις τάξεις **P** και **PL** του B+ δέντρου. Όπου **P** είναι η τάξη των ενδιάμεσων κόμβων και **PL** η τάξη των φύλλων.

β) το πλήθος των μπλοκ που απαιτούνται για τους κόμβους-φύλλα του B+ δέντρου αν τα μπλοκ είναι κατά 70% περίπου πλήρη (με στρογγυλοποίηση προς τα πάνω για ευκολία).

γ) τον αριθμό των επιπέδων του δέντρου αν οι εσωτερικοί κόμβοι είναι επίσης κατά 70% πλήρεις (με στρογγυλοποίηση προς τα πάνω για ευκολία).

δ) το συνολικό πλήθος των μπλοκ που απαιτούνται για το B+ δέντρο εφόσον ισχύουν οι υπολογισμοί των ερωτημάτων β) και γ)

ε) το πλήθος των προσπελάσεων για την αναζήτηση και την ανάκτηση μιας εγγραφής από το αρχείο, όταν δίνεται η τιμή του κλειδιού BARCODE, με την χρήση του B+ ευρετηρίου του ερωτήματος δ).

ΟΡΙΣΜΟΙ:

1. **Δείκτης δέντρου (p):** Δείκτης από έναν κόμβο του δέντρου (ενδιάμεσο ή φύλλο) προς έναν άλλον κόμβο του δέντρου (ενδιάμεσο ή φύλλο).
2. **Δείκτης δεδομένων (q):** Δείκτης από έναν κόμβο φύλλο προς τα μπλοκ με τις εγγραφές του αρχείου.
3. **Τάξη ενδιάμεσου κόμβου (P):** ο μέγιστος αριθμός **δεικτών** ενός ενδιάμεσου κόμβου.
4. **Τάξη κόμβου φύλλου (PL):** ο μέγιστος αριθμός **δεικτών δεδομένων ε**νός κόμβου φύλλου.

Άσκηση 5 [Μονάδες 25]

Ο πίνακας ΠΕΛΑΤΕΣ περιέχει εγγραφές πελατών ενός ηλεκτρονικού καταστήματος. Το πρωτεύον κλειδί του πίνακα είναι ο κωδικός πελάτη (ΚΠ#). Ακολουθούν τα κλειδιά (ΚΠ#) 11 εγγραφών πελατών.

Εγγραφή	ΚΠ#
R1	6800
R2	7819
R3	8820
R4	6428
R5	7325
R6	8000
R7	6830
R8	8231
R9	6803
R10	5800
R11	5401

Έστω ένα αρχείο ευρετηρίου που χρησιμοποιεί την μέθοδο του γραμμικού κατακερματισμού με αρχικό μέγεθος **2** κάδους (**m=1)** χωρητικότητας δύο εγγραφών έκαστος. Για την κατανομή των τιμών χρησιμοποιούνται τα **i=1** λιγότερο σημαντικά bits, ενώ η συνάρτηση κατακερματισμού είναι **h(key)=key mod 8.** Ο αριθμός των κάδων πρέπει να αυξάνεται όταν το utilization του ευρετηρίου γίνει **μεγαλύτερο ή ίσο του 75%.** Το **i** αυξάνεται μόνο όταν κρίνεται απαραίτητο. Επίσης, δεν υπάρχει όριο στον αριθμό σελίδων υπερχείλισης. Κάθε σελίδα υπερχείλισης χωράει και αυτή δύο εγγραφές.

Ζητείται:

α) Να εισαγάγετε τα κλειδιά των εγγραφών με την σειρά που σας δίνονται στον παραπάνω πίνακα

Να εμφανίσετε την μορφή του ευρετηρίου μετά από κάθε εισαγωγή κλειδιού δείχνοντας και όσα ενδιάμεσα βήματα απαιτούνται. Κάθε πράξη εισαγωγής πρέπει να εκτελείται στο αποτέλεσμα της προηγούμενης και όχι στο αρχικό ευρετήριο.

Προς διευκόλυνσή σας ακολουθεί η εισαγωγή του πρώτου κλειδιού:

Εισαγωγή Κλειδιού **6800**: h(6800)=6800 mod 8 = 0 (0000)

6800	
0	**1**

utilization=1/4 (25%)

β) Να υπολογίσετε τον **μέσο αριθμό προσπελάσεων** για την ανάκτηση μιας εγγραφής πελάτη βάσει του κλειδιού ΚΠ# όταν:

I. Το κλειδί υπάρχει στο ευρετήριο
II. Το κλειδί δεν υπάρχει στο ευρετήριο.