Σχεδιασμός Βάσεων Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2018-2019

Πρώτη Σειρά Ασκήσεων

Ανάθεση: 21-03-2019
Παράδοση: 31-03-2019 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. "3170001.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 καθώς επίσης και τα κεφάλαια 17 και 18 από το βιβλίο R.Elmasri, S.B.Navathe, "Θεμελιώδεις Αρχές Συστημάτων Βάσεων Δεδομένων", 6η Έκδοση Αναθεωρημένη", ISBN 978-960-531-2817-7.***

Άσκηση 1

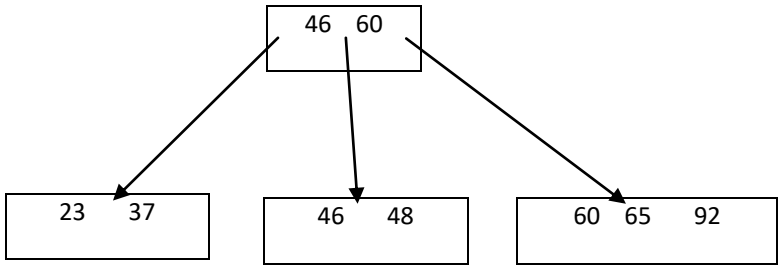
Έχουμε ένα ταξινομημένο αρχείο με 100.000 εγγραφές όπου κάθε εγγραφή έχει μήκος 240 byte. Το αρχείο είναι αποθηκευμένο σε ένα δίσκο με μέγεθος block B=2400 byte, μέσο χρόνο μετακίνησης κεφαλής ts=12 ms, μέση καθυστέρηση περιστροφής trd=8,3 ms και χρόνος μεταφοράς του block tb = 0,8ms.

Θέλουμε να διαβάσουμε Χ τυχαίες εγγραφές από το αρχείο. Μπορούμε να διαβάσουμε τυχαία Χ μπλοκ ή να διαβάσουμε ολόκληρο το αρχείο και να αναζητήσουμε τις Χ εγγραφές. Το ερώτημα είναι πότε η ανάγνωση όλου του αρχείου είναι ποιο αποδοτική από το να εκτελέσουμε Χ τυχαία διαβάσματα; Δηλαδή ποιά είναι η τιμή του Χ για την οποία η ανάγνωση ολόκληρου του αρχείου είναι πιο αποδοτική από Χ τυχαία διαβάσματα;

Σημείωση: Για λόγους απλούστευσης θεωρείστε ότι στην περίπτωση διαβάσματος όλου του αρχείου **ο χρόνος μετακίνησης στο επόμενο ίχνος** είναι μηδαμινός (δηλαδή δεν λαμβάνεται υπόψη).

Άσκηση 2

Δίνεται το παρακάτω B+ δένδρο με μέγιστο αριθμό **τριών** κλειδιών **(n=3)** ανά κόμβο/φύλλο το οποίο έχει προκύψει κατόπιν εισαγωγής των τιμών: 23,65,37,60,46,92,48.



Ζητείται να εισαγάγετε με την σειρά τις τιμές 47, 100 και 50. Σε κάθε εισαγωγή να παρουσιάσετε την νέα μορφή του δένδρου και να εξηγήσετε πως ακριβώς προέκυψε.

Άσκηση 3

Θεωρείστε ένα δίσκο με μέγεθος μπλοκ B=1024 byte. Ένας δείκτης μπλοκ (P) έχει μέγεθος 6 byte και ένας δείκτης εγγραφής (PR) έχει μήκος 7 byte. Ένα αρχείο έχει r=50000 εγγραφές βιβλίων. Κάθε εγγραφή αποτελείται από ορισμένα πεδία μεταξύ των οποίων και το πεδίο ISBN μεγέθους M=14 byte.

Υποθέστε ότι το αρχείο δεν είναι διατεταγμένο ως προς το πεδίο κλειδί ISBN και θέλουμε να δημιουργήσουμε ένα ευρετήριο B+ δένδρου πάνω στο πεδίο ISBN.

Ζητείται να υπολογίσετε:

α) τις τάξεις p και pleaf του B+ δένδρου. Όπου p είναι η τάξη των ενδιάμεσων κόμβων και pleaf η τάξη των φύλλων.

β) το πλήθος των μπλοκ που απαιτούνται για τους κόμβους-φύλλα του B+ δένδρου αν τα μπλοκ είναι κατά 69% περίπου πλήρη (με στρογγυλοποίηση προς τα πάνω για ευκολία).

γ) τον αριθμό των επιπέδων του δένδρου αν οι εσωτερικοί κόμβοι είναι επίσης κατά 69% πλήρεις (με στρογγυλοποίηση προς τα πάνω για ευκολία).

δ) το συνολικό πλήθος των μπλοκ που απαιτούνται για το B+ δένδρο 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 ISBN, με την χρήση του B+ ευρετηρίου του ερωτήματος δ).

Προσοχή:

Η τάξη ενός κόμβου ορίζεται ως ο μέγιστος αριθμός δεικτών του κόμβου. Επειδή στα B+ δένδρα οι δομές των εσωτερικών κόμβων και των κόμβων-φύλλων είναι διαφορετικές, η τάξη τους μπορεί να διαφέρει. Στην συγκεκριμένη άσκηση ορίζουμε ως **p** την τάξη ενός ενδιάμεσου κόμβου και ως **pleaf** την τάξη ενός κόμβο φύλλου. Ειδικά για τα φύλλα του δένδρου, στην τάξη προσμετρούνται μόνο οι δείκτες προς τα δεδομένα. Στον υπολογισμό της χωρητικότητας του φύλλου θα πρέπει όμως να προσμετρήσετε και το χώρο που απαιτείται για τον δείκτη προς το επόμενο φύλλο.

Σχετικά με την τάξη και την δομή των κόμβων δείτε τις διαφάνειες 85-93 της παρουσίασης "04a-Indexing.pdf" των διαλέξεων του μαθήματος. Προσέξτε όμως ότι στη συγκεκριμένη άσκηση η τάξη των ενδιάμεσων κόμβων και των κόμβων φύλλων μπορεί να είναι διαφορετικές.

Άσκηση 4

Ένας πίνακας περιέχει εγγραφές φοιτητών. Το πρωτεύον κλειδί του πίνακα είναι ο Αριθμός μητρώου (K#). Ο πίνακας περιέχει 15 εγγραφές με τις παρακάτω τιμές για το K#.

Record	K#
R1	2369
R2	3760
R3	4692
R4	4871
R5	5659
R6	1821
R7	1074
R8	7115
R9	1620
R10	2428
R11	3943
R12	4750
R13	6975
R14	4981
R15	9208

Έστω ένα αρχείο ευρετηρίου που χρησιμοποιεί την μέθοδο του γραμμικού κατακερματισμού με αρχικό μέγεθος **2** buckets (**m=1)** χωρητικότητας δύο εγγραφών έκαστο. Για την κατανομή των τιμών χρησιμοποιούνται τα **i=1** λιγότερο σημαντικά bits, ενώ η συνάρτηση κατακερματισμού είναι η **h(K)= K mod 8.** Το m πρέπει να αυξάνεται όταν το utilization του ευρετηρίου γίνει μεγαλύτερο ή ίσο του **80%**. Το i αυξάνεται μόνο όταν κρίνεται απαραίτητο. Επίσης, δεν υπάρχει όριο στον αριθμό σελίδων υπερχείλισης.

Εισαγάγετε τα κλειδιά των παραπάνω εγγραφών με την σειρά που σας δίνονται στον παραπάνω πίνακα. Εμφανίστε την μορφή του ευρετηρίου μετά από κάθε εισαγωγή κλειδιού (κάθε πράξη εισαγωγής πρέπει να εκτελείται στο αποτέλεσμα της προηγούμενης και όχι στο αρχικό ευρετήριο).

Τέλος να υπολογίσετε τον μέσο αριθμό προσπελάσεων για την ανάκτηση μιας εγγραφής όταν δίνεται το K#.

Προς διευκόλυνσή σας ακολουθεί η εισαγωγή των πρώτων δύο τιμών:

Αρχική Κατάσταση

0 1

Εισαγωγή Κλειδιού 2369: h(2369)= 2369 mod 8 = 1 (0001)

	2369

0 1

Utilization=(1/4)=25%

Εισαγωγή Κλειδιού 3760: h(3760)= 3760 mod 8 = 0 (0000)

3760	2369

0 1

Utilization=(2/4)=50%