Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2023-2024

Πρώτη Σειρά Ασκήσεων

Ανάθεση: 22-03-2024
Παράδοση: 01-04-2024 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. "3210001.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 [20 μονάδες]

Έστω ένας σκληρός δίσκος με τα ακόλουθα χαρακτηριστικά:

- Συνολική χωρητικότητα 80 GB
- 4 πλακέτες (platters) διπλής όψης
- 512 ίχνη (tracks) ανά επιφάνεια
- 1024 τομείς (sectors) ανά ίχνος
- Ταχύτητα περιστροφής 7200 rpm
- Μέσος χρόνος μετακίνησης κεφαλής= 8 ms (milliseconds)
- Υποθέστε ότι ο χρόνος μετακίνησης της κεφαλής στο επόμενο ίχνος είναι μηδενικός.

Στον δίσκο έχουμε αποθηκεύσει μια σχέση R η οποία περιέχει 8192 εγγραφές μεγέθους 40KB (kilobytes) έκαστη. Η σχέση R είναι αποθηκευμένη σε συνεχόμενα μπλοκ του δίσκου, κάθε εγγραφή της σχέσης R αποθηκεύεται ολόκληρη σε ένα μπλοκ και ένα μπλοκ δεν μπορεί να εκτείνεται σε δύο ίχνη.

Να υπολογίσετε

1. Το μέγεθος του τομέα του δίσκου σε KB (kilobytes).
2. Τον αριθμό των κυλίνδρων που καταλαμβάνει η σχέση R.
3. Τον χρόνο που απαιτείται για την ανάγνωση ολόκληρης της σχέσης R
4. Τον χρόνο που απαιτείται για την ανάγνωση 100 τυχαίων εγγραφών της σχέσης R θεωρώντας ότι το μέγεθος ενός μπλοκ (σελίδας) του δίσκου είναι το ελάχιστο δυνατό.

Άσκηση 2 [20 μονάδες]

Θεωρήστε ένα δίσκο με μέγεθος μπλοκ 1024 bytes. Δίνεται επίσης ότι ένας δείκτης μπλοκ έχει μήκος 6 bytes. Ένα αρχείο περιέχει 60000 εγγραφές σταθερού μήκους της σχέσης ΦΟΡΟΛΟΓΟΥΜΕΝΟΣ. Κάθε εγγραφή του αρχείου αποτελείται από τα ακόλουθα πεδία

- ΑΦΜ (9 bytes, πρωτεύον κλειδί),
- ΟΝΟΜΑ (30 bytes)
- ΚΩΔΙΚΟΣ_ΕΦΟΡΙΑΣ (10 bytes)
- ΑΡΙΘΜΟΣ_ΤΗΛΕΦΩΝΟΥ (9 bytes)
- ΗΜΕΡΟΜΗΝΙΑ_ΓΕΝΝΗΣΗΣ (8 bytes)
- ΦΥΛΟ (1 byte)
- ΔΙΕΥΘΥΝΣΗ (40 bytes)
- ΚΩΔΙΚΟΣ_ΕΡΓΑΣΙΑΣ (4 bytes)
- ΕΤΗΣΙΟ_ΕΙΣΟΔΗΜΑ (4 bytes).

Υποθέστε ότι το αρχείο είναι διατεταγμένο ως προς το πεδίο ΑΦΜ (κλειδί) και θέλουμε να κατασκευάσουμε ένα απλό πολυπεπίπεδο αραιό πρωτεύον ευρετήριο στο πεδίο ΑΦΜ. Το τελευταίο επίπεδο του ευρετηρίου θα πρέπει να χωράει σε ένα μπλοκ (σελίδα).

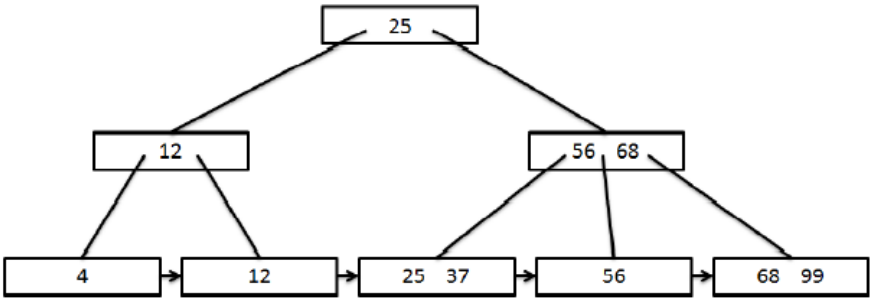
Ζητείται να υπολογίσετε:

1. Το πλήθος των καταχωρήσεων του πρώτου επιπέδου του ευρετηρίου.
2. Το πλήθος των μπλοκ του πρώτου επιπέδου του ευρετηρίου.
3. Τον αριθμό των επιπέδων του ευρετήριου.
4. Τον συνολικό αριθμό των μπλοκ που απαιτούνται για το πολυεπίπεδο ευρετήριο
5. Τον αριθμό των μπλοκ που πρέπει να προσπελαστούν για την αναζήτηση και ανάκτηση μιας εγγραφής από το αρχείο, όταν δίνεται η τιμή του ΑΦΜ, με χρήση του πολυεπίπεδου ευρετηρίου.

Σημείωση: Να υποθέσετε ότι κάθε εγγραφή του αρχείου αποθηκεύεται ολόκληρη σε ένα μπλοκ.

Άσκηση 3 [20 μονάδες]

Έστω η σχέση R(A,B,C,D) με πρωτεύον κλειδί το γνώρισμα Α. Έχουμε δημιουργήσει ένα ευρετήριο B+ δέντρου στο πρωτεύον κλειδί της R. Οι εσωτερικοί κόμβοι και τα φύλλα του ευρετηρίου χωράνε το πολύ δύο κλειδιά. Στο παρακάτω σχήμα παρουσιάζεται ένα στιγμιότυπο του δέντρου (ευρετηρίου) στο οποίο έχουν εισαχθεί ορισμένα κλειδιά.



1. Εμφανίστε την μορφή του δέντρου μετά την εισαγωγή κάθε μίας από τις ακόλουθες τιμές με την σειρά την οποία δίνονται (κάθε πράξη εισαγωγής εκτελείται στο αποτέλεσμα της προηγούμενης και όχι στο αρχικό δέντρο): 8, 57, 23, 16, 15, 95, 100, 112. Όταν ένας κόμβος διασπάται οι εγγραφές θα πρέπει να μοιράζονται ως εξής: μία στο αριστερό κόμβο και δύο στο δεξί.
2. Μετά την εισαγωγή των τιμών του παραπάνω ερωτήματος πόσα μπλοκ πρέπει να προσπελαστούν στο δίσκο για να ανακτηθούν όλες οι εγγραφές της R με κλειδί αναζήτησης A>=12 AND A<=37. Να θεωρήσετε ότι το ευρετήριο είναι αποθηκευμένο στο δίσκο, και ότι στα φύλλα δεν αποθηκεύονται οι εγγραφές της σχέσης R. Να θεωρήσετε επίσης ότι κάθε εγγραφή της σχέσης R αποθηκεύεται ολόκληρη σε ένα μπλοκ του δίσκου.

Άσκηση 4 [25 μονάδες]

Έστω ένα αρχείο ευρετηρίου που χρησιμοποιεί την μέθοδο του γραμμικού κατακερματισμού με αρχικό μέγεθος **2** κάδους (**m=1)** χωρητικότητας τριών εγγραφών έκαστο. Για την κατανομή των τιμών χρησιμοποιούνται τα **i=1** λιγότερο σημαντικά bits. Ο αριθμός των κάδων πρέπει να αυξάνεται όταν το utilization του ευρετηρίου γίνει μεγαλύτερο ή ίσο του 7**0%**. Το **i** αυξάνεται μόνο όταν κρίνεται απαραίτητο. Επίσης, δεν υπάρχει όριο στον αριθμό σελίδων υπερχείλισης. Κάθε σελίδα υπερχείλισης χωράει και αυτή τρείς εγγραφές.

Ζητείται να εισαγάγετε τα παρακάτω κλειδιά (αναγράφεται η τιμή h(x) αντί για το κλειδί x) με την σειρά που σας δίνονται ξεκινώντας από αριστερά προς τα δεξιά.

[1000, 0000, 1101, 0010, 0010, 1100, 0011, 1111, 0110, 1110]

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:

1. Εισαγωγή 0000

1000	

utilization=1/6, m=1, i=1

0 **1**

Άσκηση 5 [20 μονάδες]

Έστω μια σχέση R(A,B) οργανωμένη ως αρχείο κατακερματισμού στο δίσκο. Το κατακερματισμένο αρχείο καταλαμβάνει 1024 blocks (κάδους – buckets) που περιέχουν τις εγγραφές της σχέσης. Για την αποθήκευση μιας εγγραφής (a,b) εφαρμόζουμε πρώτα τη συνάρτηση κατακερματισμού h1 στο πεδίο a λαμβάνοντας N bits. Στη συνέχεια εφαρμόζουμε τη συνάρτηση h2 στο πεδίο b λαμβάνοντας 10-N bits. Τα 10 bits συνολικά ορίζουν τη διεύθυνση του block στο οποίο θα αποθηκευτεί η εγγραφή (a,b).

Θεωρήστε ότι το 30% των επερωτήσεων που αφορούν στη σχέση R είναι της μορφής:

Q1: SELECT * FROM R WHERE A=a,

ενώ το υπόλοιπο 70% είναι της μορφής :

Q2: SELECT * FROM R WHERE B=b

όπου a και b είναι τιμές που δίνονται από τους χρήστες που υποβάλλουν τα ερωτήματα.

1. Ζητείται να προσδιορίσετε τον αριθμό των μπλοκ που πρέπει να προσπελαστούν για να απαντηθούν α) οι επερωτήσεις τύπου Q1 και β) οι επερωτήσεις τύπου Q2. Η απάντηση να δοθεί ως συνάρτηση των N bits.
2. Δώστε έναν τύπο που να εκτιμά το μέσο αριθμό των μπλοκ που πρέπει να προσπελαστούν για την απάντηση των επερωτήσεων (Q1 και Q2) στη σχέση R.