Σχεδιασμός Βάσεων Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2019-2020

Δεύτερη Σειρά Ασκήσεων

Ανάθεση: 01-05-2020
Παράδοση: 12-05-2020 Ώρα (23:55)

Οδηγίες

- *Η δεύτερη 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 AM 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.*

Άσκηση 1 [μονάδες 25]

Έστω η σχέση R(a,b,c,d) η οποία περιέχει 45000 εγγραφές. Η R είναι αποθηκευμένη σε ένα αρχείο σωρού κάθε block (σελίδα) του οποίου χωράει 100 εγγραφές. Υποθέστε ότι το DBMS χρησιμοποιεί τον αλγόριθμο ταξινόμησης "Two Phase Sort" όπως περιγράφεται στο αρχείο "Εγγραφα\09-sorting.pdf", και ότι μπορεί να χρησιμοποιηθεί ενδιάμεση μνήμη μεγέθους 4 blocks. Θέλουμε να ταξινομήσουμε τις εγγραφές της σχέσης R ως προς το γνώρισμα **a**.

Ζητείται να απαντήσετε τα παρακάτω ερωτήματα.

1. Πόσες ταξινομημένες λίστες θα υπάρχουν μετά το πρώτο πέρασμα ταξινόμησης και ποιό θα είναι το μέγεθος (σε blocks) της κάθε μιας;
2. Πόσα περάσματα θα χρειαστούν συνολικά για την ταξινόμηση της σχέσης R;
3. Ποιό θα είναι το συνολικό κόστος σε I/0 για την ταξινόμηση της σχέσης R; Μην προσμετρήσετε το κόστος για το output του τελικού αποτελέσματος.
4. Ποιό πρέπει να είναι το ελάχιστο μέγεθος (σε blocks) της ενδιάμεσης μνήμης ώστε η σχέση R να μπορεί να ταξινομηθεί σε δύο μόνο περάσματα;

Άσκηση 2 [μονάδες 25]

Έστω οι σχέσεις R1(x,y,z) και R2(x,c,d) οι οποίες περιέχουν 15000 και 9000 εγγραφές αντίστοιχα. Σε μία σελίδα χωράνε 100 εγγραφές της σχέσης R1 ή 100 εγγραφές της σχέσης R2. Το μέγεθος του διαθέσιμου ενταμιευτή μνήμης είναι M=13 σελίδες και δεν υπάρχουν διαθέσιμα ευρετήρια. Υποθέστε ότι το DBMS υποστηρίζει τους παρακάτω αλγορίθμους για την υλοποίηση της ισοσύνδεσης:

a. Nested Loop Join
b. Sort-Merge Join
c. Hash Join

Ζητείται:

1. Να προσδιορίσετε τον βέλτιστο αλγόριθμο ισοσύνδεσης για την εκτέλεση του παρακάτω ερωτήματος.

   Select * from R1,R2 where R1.x = R2.x

2. Αν αυξήσουμε το μέγεθος του ενταμιευτή σε 100 σελίδες πως επηρεάζεται το κόστος των παραπάνω αλγορίθμων;

Άσκηση 3 [μονάδες 25]

Έστω οι παρακάτω σχέσεις:

ΣΧΕΣΗ	**ΑΡΙΘΜΌΣ ΕΓΓΡΑΦΩΝ**	**ΑΡΙΘΜΟΣ ΣΕΛΙΔΩΝ**
ΕΡΓΑΖΟΜΕΝΟΙ (ΚΕ, ΟΝΟΜΑ, ΤΜΗΜΑ)	5000	250
ΕΡΓΑ(ΚΕ, ΕΡΓΟ, ΕΙΔΙΚΟΤΗΤΑ)	2000	50

```
SELECT ONOMA, TMHMA, ΕΡΓΟ, ΕΙΔΙΚΟΤΗΤΑ
   FROM ΕΡΓΑΖΟΜΕΝΟΙ, ΕΡΓΑ
WHERE ΕΡΓΑΖΟΜΕΝΟΙ.ΚΕ=ΕΡΓΑ.ΚΕ AND
        ΤΜΗΜΑ='ΜΗΧΑΝΟΓΡΑΦΗΣΗ' AND ΕΙΔΙΚΟΤΗΤΑ='ΠΡΟΓΡΑΜΜΑΤΙΣΤΗΣ'
```

Επιπλέον θεωρείστε ότι:

- Υπάρχουν 10 διαφορετικά τμήματα και 5 διαφορετικές ειδικότητες.
- Υπάρχει ένα απλό ευρετήριο (non clustered) στο πεδίο ΕΡΓΑ.ΕΙΔΙΚΟΤΗΤΑ
- Υπάρχει ευρετήριο συστάδων (clustered index) στο πεδίο ΕΡΓΑΖΟΜΕΝΟΙ.ΤΜΗΜΑ
- Τα ευρετήρια βρίσκονται στην μνήμη.
- Το μέγεθος της διαθέσιμης μνήμης είναι M=8 σελίδες.
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.

Ζητείται:

1. Να σχεδιάσετε το τελικό, βελτιστοποιημένο λογικό πλάνο της παρακάτω επερώτησης. Δεν χρειάζεται να δείξετε τα ενδιάμεσα βήματα.

```
SELECT ΟΝΟΜΑ, ΤΜΗΜΑ, ΕΡΓΟ, ΕΙΔΙΚΟΤΗΤΑ
    FROM ΕΡΓΑΖΟΜΕΝΟΙ, ΕΡΓΑ
WHERE ΕΡΓΑΖΟΜΕΝΟΙ.ΚΕ=ΕΡΓΑ.ΚΕ AND
        ΤΜΗΜΑ='ΜΗΧΑΝΟΓΡΑΦΗΣΗΣ' AND ΕΙΔΙΚΟΤΗΤΑ='ΠΡΟΓΡΑΜΜΑΤΙΣΤΗΣ'
```

2. Να υπολογίσετε το ελάχιστο κόστος (σε I/O) εκτέλεσης της επερώτησης χρησιμοποιώντας του αλγορίθμους α) SMJ (Sort Merge Join) και b) NLJ (Block Nested Loop Join).

Άσκηση 4 [μονάδες 25]

Η παρακάτω SQL επερώτηση εμφανίζει τα ονόματα των φοιτητών που κατοικούν στην Αθήνα, σπουδάζουν σε ένα από τα 9 καλύτερα πανεπιστήμια της χώρας (ΚΑΤΑΤΑΞΗ < 10), και έχουν προεγγραφεί σε ένα Μεταπτυχιακό Πρόγραμμα Σπουδών με γνωστικό αντικείμενο την επιστήμη των δεδομένων "Data Science".

```
SELECT ΟΝΟΜΑ
  FROM ΦΟΙΤΗΤΕΣ, ΠΑΝΕΠΙΣΤΗΜΙΑ, ΠΡΟΕΓΓΡΑΦΕΣ
WHERE ΦΟΙΤΗΤΕΣ.ΚΠ=ΠΑΝΕΠΙΣΤΗΜΙΑ.ΚΠ AND
        ΦΟΙΤΗΤΕΣ.ΑΜ=ΠΡΟΕΓΓΡΑΦΕΣ.ΑΜ AND
        ΠΟΛΗ='ΑΘΗΝΑ'  AND ΚΑΤΑΤΑΞΗ < 10 AND ΜΠΣ='DS'
```

Ο πίνακας που ακολουθεί περιέχει στοιχεία για τις παραπάνω σχέσεις:

ΣΧΕΣΗ	**ΑΡΙΘΜΌΣ ΕΓΓΡΑΦΩΝ**	**ΑΡΙΘΜΟΣ ΣΕΛΙΔΩΝ**	**ΠΡΩΤΕΥΟΝ ΚΛΕΙΔΙ**
ΦΟΙΤΗΤΕΣ(**ΑΜ,** ΟΝΟΜΑ, ΠΟΛΗ, ΚΠ)	2000	100	ΑΜ (αριθμός μητρώου)
ΠΑΝΕΠΙΣΤΗΜΙΑ (**ΚΠ**, ΟΝΟΜΑΣΙΑ, ΚΑΤΑΤΑΞΗ)	100	10	ΚΠ (κωδικός πανεπιστημίου)
ΠΡΟΕΓΓΡΑΦΕΣ(**ΑΜ,ΜΠΣ**)	3000	200	(ΑΜ, ΜΠΣ) αριθμός μητρώου και κωδικός μεταπτυχιακού προγράμματος.

Επιπλέον δίνεται ότι:

- Κάθε πανεπιστήμιο έχει έναν μοναδικό αριθμό κατάταξης (ΚΑΤΑΤΑΞΗ) από το 1 μέχρι το 100.
- Υπάρχουν 20 διαφορετικές πόλεις.
- Κάθε υποψήφιος φοιτητής μπορεί να προεγγραφεί το πολύ σε 5 Μεταπτυχιακά Προγράμματα Σπουδών.

- Το πεδίο ΦΟΙΤΗΤΕΣ.ΚΠ είναι ξένο κλειδί το οποίο αναφέρεται (references) στο πεδίο ΠΑΝΕΠΙΣΤΗΜΙΑ.ΚΠ.
- Το πεδίο ΠΡΟΕΓΓΡΑΦΕΣ.ΑΜ είναι ξένο κλειδί το οποίο αναφέρεται (references) στο πεδίο ΦΟΙΤΗΤΕΣ.ΑΜ.
- Υπάρχει ένα απλό ευρετήριο (non clustered) στο πεδίο ΠΡΟΕΓΓΡΑΦΕΣ.ΑΜ και όλες οι σελίδες του ευρετηρίου βρίσκονται στην κύρια μνήμη. Αυτό είναι το μόνο ευρετήριο που υπάρχει. Μην θεωρήσετε ότι για κάθε πρωτεύον κλειδί του πίνακα υπάρχει ευρετήριο συστάδων (clustered index).
- Το μέγεθος της διαθέσιμης μνήμης είναι M=8 σελίδες.
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.

Ζητείται να υπολογίσετε το κόστος σε I/O του φυσικού πλάνου εκτέλεσης που ακλουθεί. Να υπολογίσετε το κόστος σε I/O (εφόσον υφίσταται) για κάθε μία από τις 6 επιμέρους λειτουργίες του πλάνου και να δείξετε πως αυτό προκύπτει.

ΦΥΣΙΚΟ ΠΛΑΝΟ ΕΚΤΕΛΕΣΗΣ