Σχεδιασμός Βάσεων Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2020-2021

Δεύτερη Σειρά Ασκήσεων

Ανάθεση: 10-05-2021
Παράδοση: 23-05-2021 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. "3180001.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) για την οποία ισχύουν τα εξής:

- Η σχέση R περιέχει 1.000.000 εγγραφές για την αποθήκευση των οποίων απαιτούνται 20.000 σελίδες.
- Δίνεται ότι V(R,a)=n, όπου n είναι θετικός ακέραιος αριθμός.
- Υπάρχει το ευρετήριο P στο γνώρισμα a της σχέσης R.
- Το ευρετήριο βρίσκεται στην μνήμη.

Θεωρείστε τα ακόλουθα επερωτήματα:

a. $\sigma_{a=2}(R)$
b. $\sigma_{K<=a<=L}(R)$, όπου K και L είναι σταθερές, έτσι ώστε n/10 τιμές να ανήκουν στο διάστημα [K...L].

Ζητείται:

1. Να υπολογίσετε το κόστος σε I/O των επερωτημάτων **a** και **b,** ως συνάρτηση του **n**, για κάθε μια από τις παρακάτω περιπτώσεις:
   i. Το ευρετήριο P είναι ευρετήριο συστάδων (clustered index)
   ii. Το ευρετήριο P είναι απλό ευρετήριο (nonclustered index)

2. Στην περίπτωση που το ευρετήριο P είναι απλό μπορούμε να εκτελέσουμε τα επερωτήματα a και b σαρώνοντας κάθε φορά την σχέση R (table scan), ή χρησιμοποιώντας το ευρετήριο αντίστοιχα. Για ποιές τιμές του **n** συμφέρει να χρησιμοποιήσουμε το ευρετήριο για να την εκτέλεση των επερωτημάτων a και b;

   Να απαντήσετε ξεχωριστά για κάθε επερώτημα.

Άσκηση 2 [Μονάδες 15]

Έστω οι σχέσεις R(a,b) και S(b,c) για τις οποίες υπάρχουν δύο ιστογράμματα. Ένα ιστόγραμμα για το γνώρισμα R.b και ένα ιστόγραμμα για το γνώρισμα S.b. Η μορφή των ιστογραμμάτων παρουσιάζεται στον παρακάτω πίνακα.

Διάστημα Τιμών	**R**	**S**
[1..20]	0	10
[21..40]	80	100
[41..60]	100	60
[61..80]	20	60
[81..100]	30	0

Κάθε γραμμή του πίνακα μας δίνει την πληροφορία ότι υπάρχουν T1 πλειάδες της σχέσης R και T2 πλειάδες της σχέσεις S, για τις οποίες οι τιμές των γνωρισμάτων R.b και S.b ανήκουν στο αντίστοιχο διάστημα. Για παράδειγμα η τρίτη γραμμή του πίνακα δηλώνει ότι υπάρχουν 100 πλειάδες της σχέσης R με τιμές στο γνώρισμα R.b που ανήκουν στο διάστημα [41..60], καθώς επίσης και 60 πλειάδες της σχέσης S με τιμές στο γνώρισμα S.b που ανήκουν στο ίδιο διάστημα.

Ζητείται να εκτιμήσετε το πλήθος των πλειάδων που θα εμφανίσει το παρακάτω επερώτημα:

```
SELECT a,c FROM R,S WHERE R.b=S.b
```

a. Βάσει των στατιστικών των ιστογραμμάτων R και S

b. Βάσει της υπόθεσης ότι όλες οι τιμές κατανέμονται ομοιόμορφα.

Άσκηση 3 [Μονάδες 30]

Έστω οι σχέσεις R (a,b,c) και S(c,d,e) για τις οποίες ισχύουν τα εξής:

1. Η σχέση R έχει 20.000 εγγραφές και σε μια σελίδα χωράνε 25 εγγραφές της σχέσης R.
2. Η σχέση S έχει 45.000 εγγραφές και σε μια σελίδα χωράνε 30 εγγραφές της σχέσης S.

Το μέγεθος της διαθέσιμης μνήμης είναι M=41 σελίδες.

Ζητείται:

1. Να εκτιμήσετε το κόστος σε I/O της σύζευξης **R⋈S** για κάθε έναν από τους παρακάτω αλγορίθμους:
    a. NLJ (Block Nested Loop Join)
    b. SMJ (Sort Merge Join)
    c. Hash Join

2. Συγκεκριμένα για τον αλγόριθμο SMJ ποιο είναι το ελάχιστο δυνατό κόστος (σε I/O) που μπορούμε να πετύχουμε για την υλοποίηση της σύζευξης **R⋈S**; Αναφέρετε δύο τρόπους για να πετύχουμε το ελάχιστο δυνατό κόστος.

Άσκηση 4 [μονάδες 30]

Έστω οι παρακάτω σχέσεις της βιβλιογραφικής βάσης δεδομένων της Βιβλιοθήκης ενός Κέντρου Νεότητας.

ΔΑΝΕΙΖΟΜΕΝΟΙ(**ΚΔ,** Επώνυμο, Όνομα, Ηλικία, Φύλο, Ιδιότητα)
ΒΙΒΛΙΑ(**ΚΒ,** Τίτλος, Εκδότης, Έκδοση, Σελίδες, Σημειώσεις)
ΔΑΝΕΙΣΜΟΙ(**ΚΔ,ΚΒ,** Ημερομηνία_Δανεισμού)

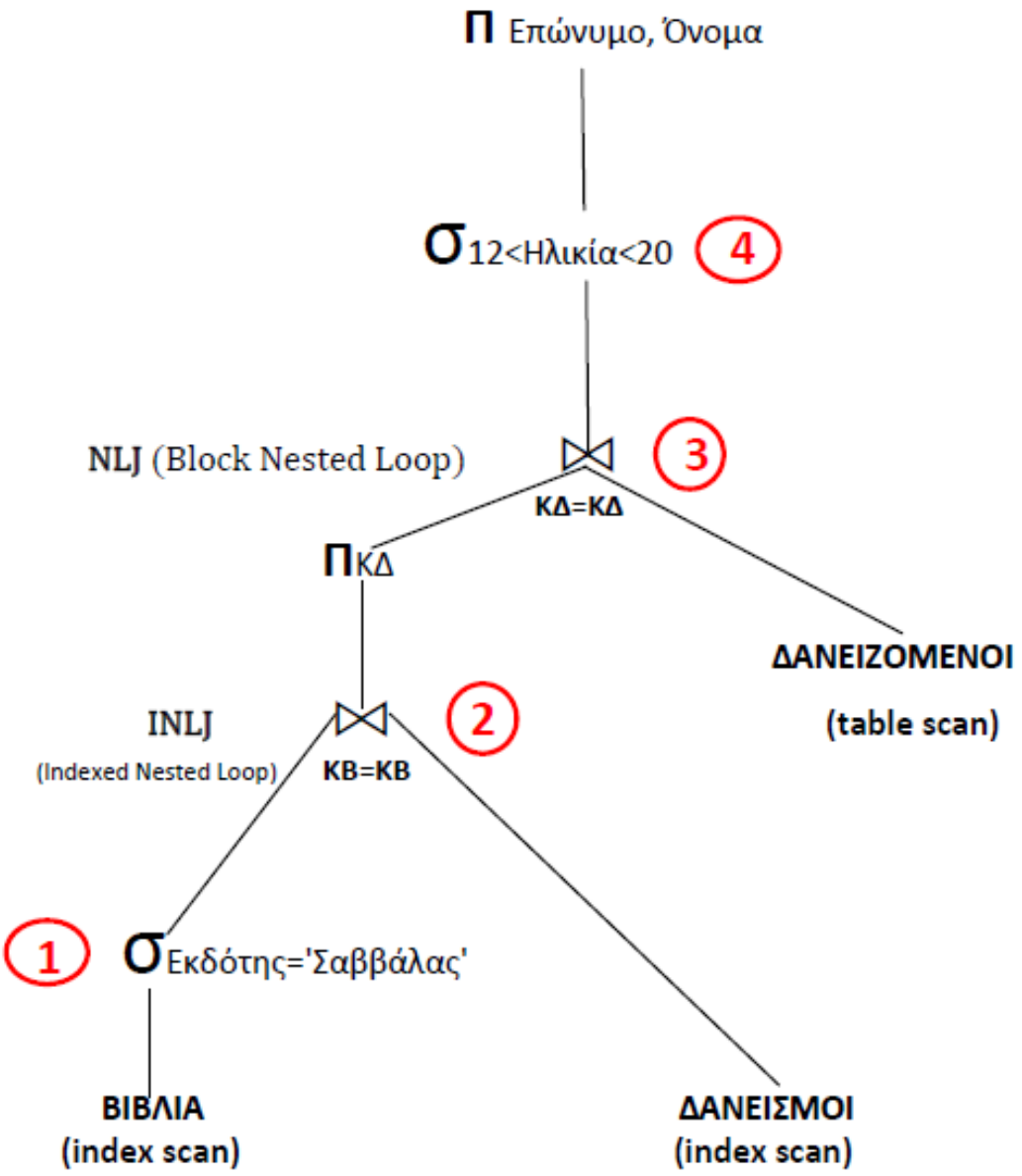
Θεωρείστε ότι:

1. Υπάρχουν 10.000 εγγραφές δανειζόμενων αποθηκευμένες σε 1.000 σελίδες.
2. Υπάρχουν 50.000 εγγραφές βιβλίων αποθηκευμένες σε 5.000 σελίδες
3. Υπάρχουν 300.000 εγγραφές δανεισμών αποθηκευμένες σε 15.000 σελίδες
4. Υπάρχουν 500 διαφορετικοί εκδότες
5. Οι ηλικίες των δανειζομένων κυμαίνονται από 7 έως και 24 έτη.
6. Το μέγεθος της διαθέσιμης μνήμης είναι 20 σελίδες (**M=20**).
7. Υπάρχει ένα απλό ευρετήριο (non clustered) στο γνώρισμα **Εκδότης** της σχέσης ΒΙΒΛΙΑ.
8. Υπάρχει ένα ευρετήριο συστάδων (clustered index) στο γνώρισμα **ΚΒ** της σχέσης ΔΑΝΕΙΣΜΟΙ.
9. Τα παραπάνω δύο ευρετήρια είναι τα **μόνα** που υπάρχουν. Μην θεωρήσετε ότι για κάθε πρωτεύον κλειδί των σχέσεων υπάρχει ευρετήριο συστάδων.
10. Τα ευρετήρια βρίσκονται στην μνήμη.
11. Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.
12. Οι επιλογές είναι μεταξύ τους ανεξάρτητες.

Δίνεται το ακόλουθο SQL επερώτημα:

```
SELECT Επώνυμο, Όνομα
   FROM ΔΑΝΕΙΖΟΜΕΝΟΙ, ΒΙΒΛΙΑ, ΔΑΝΕΙΣΜΟΊ
WHERE ΔΑΝΕΙΖΌΜΕΝΟΙ.ΚΔ=ΔΑΝΕΙΣΜΟΙ.ΚΔ AND
        ΔΑΝΕΙΣΜΟΙ.ΚΒ=ΒΙΒΛΙΑ.ΚΒ AND
        Εκδότης='Σαββάλας' AND  ( Ηλικία > 12 AND Ηλικία < 20)
```

Ζητείται να υπολογίσετε το κόστος σε I/O του φυσικού πλάνου εκτέλεσης που ακολουθεί. Συγκεκριμένα να υπολογίσετε το κόστος σε I/O (εφόσον υφίσταται) για κάθε μία από τις 4 αριθμημένες επιμέρους λειτουργίες του πλάνου και να δείξετε πως αυτό προκύπτει. Επιπλέον για κάθε μια από τις 4 λειτουργίες να προσδιορίσετε τον αριθμό των εγγραφών που προκύπτουν στην έξοδο και να δείξετε πως αυτός υπολογίζεται.


Π Επώνυμο, Όνομα
σ12<Ηλικία<20
4
NLJ (Block Nested Loop)
3
ΚΔ=ΚΔ
ΠΚΔ
ΔΑΝΕΙΖΟΜΕΝΟΙ
(table scan)
INLJ
(Indexed Nested Loop)
ΚΒ=ΚΒ
2
1
σΕκδότης='Σαββάλας'
ΒΙΒΛΙΑ
(index scan)
ΔΑΝΕΙΣΜΟΙ
(index scan)

Άσκηση 5 [BONUS Μονάδες 15]

Δίνεται το παρακάτω λογικό σχήμα:

R1(a,b), R2(b,c), R3(b,e), R4(b,f)

Ζητείται:

1. Να γράψετε ένα λογικό πλάνο αριστερού βάθους (left-deep) για το ακόλουθο SQL ερώτημα. Μπορείτε να σχεδιάσετε το πλάνο ή να το γράψετε σε σχεσιακή άλγεβρα.

   ```
   Select *
     from R1, R2,R3,R4
   where R1.b=R2.b and R2.b=R3.b and R3.b=R4.b
   ```

2. Δεδομένου ότι υπάρχουν μόνο τέσσερα ευρετήρια, ένα ευρετήριο συστάδων (clustered index) στο γνώρισμα b κάθε σχέσης, και ότι καμία σχέση ή ενδιάμεσο αποτέλεσμα δεν χωράει στην μνήμη, να προτείνετε ένα αποδοτικό φυσικό πλάνο για το λογικό πλάνο που σχεδιάσατε στο ερώτημα ένα (1). Με άλλα λόγια να προσδιορίσετε τον αλγόριθμο ισοσύνδεσης (hash join, nested loop, sortmerge κ.λπ.) καθώς και την μέθοδο πρόσβασης για το διάβασμα των σχέσεων (table scan, index scan κ.λπ.). Να εξηγήσετε γιατί το πλάνο που προτείνετε είναι αποδοτικό.

3. Να εκτιμήσετε το κόστος σε I/O του φυσικού πλάνου που προτείνατε στο ερώτημα δύο (2). Η εκτίμηση του κόστους να γίνει σε όρους B(X). Οπου B(X)=αριθμός σελίδων της σχέσης X. Να αιτιολογήσετε την απάντησή σας.