Σχεδιασμός Βάσεων Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2023-2024

Δεύτερη Σειρά Ασκήσεων

Ανάθεση: 14-05-2024
Παράδοση: 24-05-2024 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. "3200001.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.

Άσκηση 1 [Μονάδες 20]

Έστω οι σχέσεις R1 (a,b,c,d) και R2 (p,q,r,s,t,u,v) για τι οποίες ισχύουν:

1. T(R1)=5000000, B(R1)=50000, V(R1,a)=100, V(R1,b)=50
2. T(R2)=10000, B(R2)=2000
3. Στο γνώρισμα R2.s το DBMS τηρεί το ακόλουθο ιστόγραμμα:

s	Αριθμός Εγγραφών
[1..100]	2500
[101..200]	500
[201..400]	4000
[401..500]	3000

4. Υπάρχει απλό (non-clustered) ευρετήριο B+ δέντρο στο ζεύγος γνωρισμάτων (a,b) της σχέσης R1.
5. Υπάρχει απλό (non-clustered) ευρετήριο B+ δέντρο στο γνώρισμα s της σχέσης R2.
6. Οπού απαιτείται να υποθέσετε ότι οι τιμές κατανέμονται ομοιόμορφα. Επίσης θεωρείστε ότι οι κατανομές των τιμών διαφορετικών γνωρισμάτων είναι μεταξύ τους ανεξάρτητες.
7. Τα ευρετήρια βρίσκονται στην μνήμη.

Δίνονται τα παρακάτω δύο επερωτήματα σε γλώσσα SQL:

Q1: SELECT * FROM R1 WHERE a=40 and b=50

Q2: SELECT * FROM R2 WHERE s>96 AND s<=220

Α) Ποιό επερώτημα έχει το μικρότερο κόστος I/O; Να αιτιολογήσετε την απάντησή σας.

Β) Ποιά θα ήταν η απάντηση σας στο παραπάνω ερώτημα αν και τα **δύο ευρετήρια Β+δέντρου** ήταν ευρετήρια συστάδων (clustered index).

Άσκηση 2 [Μονάδες 35]

Έστω οι παρακάτω σχέσεις τα πρωτεύοντα κλειδιά των οποίων είναι υπογραμμισμένα:

ΔΡΟΜΕΙΣ(ΚΔ, Όνομα, Επώνυμο, Ηλικία, Φύλο, Τηλέφωνο, Email, Χώρα)
ΤΕΡΜΑΤΙΣΑΝΤΕΣ (ΚΔ,Ετος,bibno, cpStart, cp5, cp10, cp21, cp30, cpFinish)

Η σχέση δρομείς περιέχει τα στοιχεία των δρομέων που έχουν τερματίσει τουλάχιστον μια φορά τον αυθεντικό μαραθώνιο της Αθήνας τα τελευταία τέσσερα έτη (2020-2023).

Η σχέση ΤΕΡΜΑΤΙΣΑΝΤΕΣ περιέχει τον αριθμό συμμετοχής (bibno) κάθε δρομέα καθώς και την ημερομηνία και ώρα διέλευσής του (timestamp) από τα σημεία ελέγχου (check point) της διαδρομής του αγώνα (Εκκίνηση, πέμπτο χιλιόμετρο, δέκατο χιλιόμετρο κ.λπ.). Υπάρχουν δρομείς οι οποίοι έχουν τερματίσει στον μαραθώνιο της Αθήνας και τα τέσσερα έτη (2020-2023).

Για τις παραπάνω σχέσεις ισχύουν τα εξής:

1. Τ(ΔΡΟΜΕΙΣ)=40000 και Β(ΔΡΟΜΕΙΣ)=200
2. Τ(ΤΕΡΜΑΤΙΣΑΝΤΕΣ)=60000 και Β(ΤΕΡΜΑΤΙΣΑΝΤΕΣ)=600
3. Υπάρχει ευρετήριο συστάδων (clusterded index) Β+ δέντρο στο γνώρισμα ΤΕΡΜΑΤΙΣΑΝΤΕΣ.Ετος
4. Υπάρχει ευρετήριο συστάδων Β+ δέντρο στο γνώρισμα ΔΡΟΜΕΙΣ.ΚΔ.
5. Τα ευρετήρια βρίσκονται στην μνήμη του συστήματος.
6. Η διαθέσιμη μνήμη είναι 21 σελίδες (Μ=21).
7. Όπου απαιτείται να υποθέσετε ότι τα δεδομένα κατανέμονται ομοιόμορφα.
8. Τα ευρετήρια που δίνονται είναι τα μόνα που υπάρχουν. **Μην θεωρήσετε ότι στο πρωτεύον κλειδί κάθε σχέσης υπάρχει ευρετήριο συστάδων (clustered index).**

Ζητείται:

Α) Να σχεδιάσετε το τελικό βελτιστοποιημένο λογικό πλάνο του παρακάτω επερωτήματος. Δεν χρειάζεται να δείξετε τα ενδιάμεσα βήματα.

```
SELECT *
   FROM  ΔΡΟΜΕΙΣ, ΤΕΡΜΑΤΙΣΑΝΤΕΣ
 WHERE ΔΡΟΜΕΙΣ.ΚΔ=ΤΕΡΜΑΤΙΣΑΝΤΕΣ.ΚΔ AND Έτος='2023'
```

Β) Να υπολογίσετε το **ελάχιστο** κόστος σε I/O εκτέλεσης του επερωτήματος χρησιμοποιώντας τους αλγορίθμους α) SMJ (Sort Merge Join) και β) NLJ (Block Nested Loop Join). Να δείξετε τον τρόπο υπολογισμού, όχι μόνο το τελικό αποτέλεσμα.

Άσκηση 3 [Μονάδες 20]

Η άσκηση βασίζετε στα δεδομένα της δεύτερης άσκησης (διαβάστε την εκφώνηση εκείνης). Θεωρείστε ότι ισχύουν όλα όσα αναφέρονται στην εκφώνηση της δεύτερης άσκησης εκτός από τους ισχυρισμούς 3 και 4. Δηλαδή **δεν υπάρχει κανένα** ευρετήριο στις σχέσεις ΔΡΟΜΕΙΣ και ΤΕΡΜΑΤΙΣΑΝΤΕΣ.

Έστω το παρακάτω επερώτημα το οποίο εμφανίζει τα στοιχεία ενός δρομέα και τις πληροφορίες των τερματισμών του.

```
SELECT *
   FROM  ΔΡΟΜΕΙΣ, ΤΕΡΜΑΤΙΣΑΝΤΕΣ
 WHERE ΔΡΟΜΕΙΣ.ΚΔ=ΤΕΡΜΑΤΙΣΑΝΤΕΣ.ΚΔ AND  ΔΡΟΜΕΙΣ.ΚΔ=κδ.
```

όπου κδ ο κωδικός ενός δρομέα.

Ζητείται να προτείνετε τον **ελάχιστο** αριθμό ευρετηρίων και ένα φυσικό πλάνο εκτέλεσης του παραπάνω επερωτήματος, το οποίο θεωρείτε ότι έχει το **ελάχιστο** κόστος. Να σχεδιάσετε το φυσικό πλάνο. Στη συνέχεια να υπολογίσετε το κόστος του προτεινόμενου πλάνου και να αιτιολογήσετε γιατί θεωρείτε ότι είναι το βέλτιστο. Μπορείτε να δημιουργήσετε ένα ή περισσότερα ευρτετήρια Β+δέντρου (εξηγήστε για το κάθε ένα αν θα είναι απλό ή συστάδων), και να χρησιμοποιήσετε οποιαδήποτε αλγόριθμο θελήσετε για την υλοποίηση της ισοσύνδεσης (join). Σε περίπτωση που δημιουργήσετε κάποιο ευρετήριο να θεωρήσετε ότι αυτό βρίσκεται στη μνήμη.

Άσκηση 4 [μονάδες 30]

Στο ακόλουθο σχήμα έχουν καταγραφεί τα αποτελέσματα μιας έρευνας σχετικά με τις προτιμήσεις των ακροατών διαφορετικών τραγουδιών. Τα κλειδιά των σχέσεων είναι υπογραμμισμένα.

ΑΚΡΟΑΤΕΣ (<u>ΚΑ</u> , Ηλικία)
ΤΡΑΓΟΥΔΙΑ (<u>Τίτλος</u> , Συνθέτης)
ΑΡΕΣΕΙ (<u>ΚΑ , Τίτλος</u>)

Ακολουθούν ορισμένα στοιχεία για τις παραπάνω σχέσεις:

- **Σχέση ΑΚΡΟΑΤΕΣ**
    1. Η σχέση ΑΚΡΟΑΤΕΣ περιέχει 30.000 εγγραφές και σε μία σελίδα χωράνε 10 εγγραφές της σχέσης.
    2. Υπάρχει ένα ευρετήριο συστάδων (clustered index) Β+ δέντρο στο γνώρισμα Ηλικία.
    3. Υπάρχει ένα απλό ευρετήριο κατακερματισμού (hash index) στο γνώρισμα ΚΑ (Κωδικός Ακροατή).
    4. Η έρευνα διεξήχθη σε ακροατές ηλικίας 21 έως και 60 ετών.
- **Σχέση ΤΡΑΓΟΥΔΙΑ**
    5. Η σχέση ΤΡΑΓΟΥΔΙΑ περιέχει 1.000 εγγραφές και σε μία σελίδα χωράνε 5 εγγραφές της σχέσης.
    6. Υπάρχει ένα απλό ευρετήριο κατακερματισμού (hash index) στο γνώρισμα Συνθέτης.
    7. Υπάρχουν 100 διαφορετικοί συνθέτες.

- **Σχέση ΑΡΕΣΕΙ**
    8. Η σχέση ΑΡΕΣΕΙ περιέχει 500.000 εγγραφές και σε μία σελίδα χωρούν 50 εγγραφές της σχέσης.
    9. Υπάρχει ένα απλό ευρετήριο κατακερματισμού (hash index) στο γνώρισμα Τίτλος.
    10. Υπάρχει ένα ευρετήριο συστάδων (clustered index) B+ δέντρο στο γνώρισμα ΚΑ (Κωδικός Ακροατή).
    11. Θεωρείστε ότι κατά μέσο όρο ένας ακροατής δήλωσε ότι του αρέσουν 17 διαφορετικά τραγούδια.

Επιπλέον θεωρείστε ότι:

- Το μέγεθος της διαθέσιμης μνήμης (buffer) είναι M=62 σελίδες.
- Όλες οι σελίδες όλων των ευρετηρίων βρίσκονται στην μνήμη.
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.
- Οι επιλογές είναι μεταξύ τους ανεξάρτητες.
- Τα ευρετήρια που δίνονται είναι τα μόνα που υπάρχουν. **Μην θεωρήσετε ότι στο πρωτεύον κλειδί κάθε σχέσης υπάρχει ευρετήριο συστάδων (clustered index).**

Έστω το παρακάτω επερώτησα σε γλώσσα SQL:

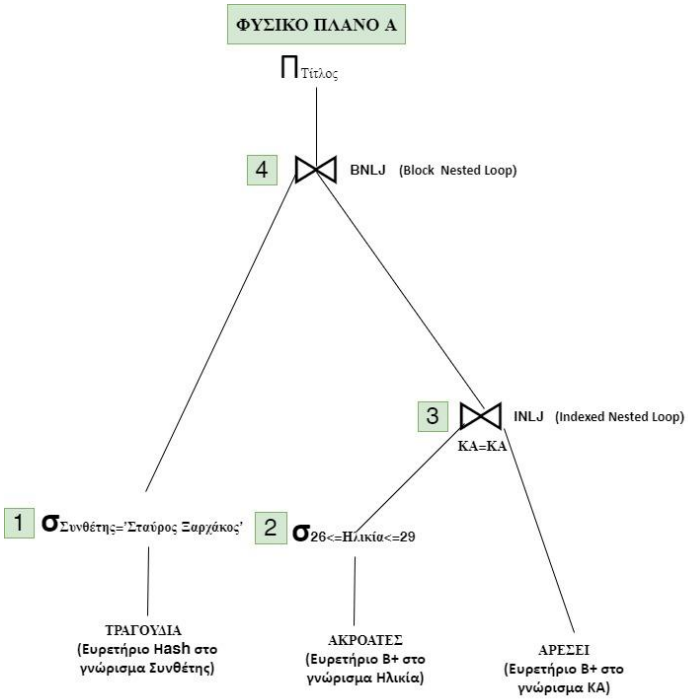
```
SELECT ΤΡΑΓΟΥΔΙΑ.Τίτλος
  FROM ΑΚΡΟΑΤΕΣ, ΑΡΕΣΕΙ, ΤΡΑΓΟΥΔΙΑ
WHERE ΑΚΡΟΑΤΕΣ.ΚΑ=ΑΡΕΣΕΙ.ΚΑ AND ΑΡΕΣΕΙ.Τίτλος=ΤΡΑΓΟΥΔΙΑ.Τίτλος AND
      (Συνθέτης='Σταύρος Ξαρχάκος') AND (Ηλικία>= 26 AND Ηλικία <=29)
```

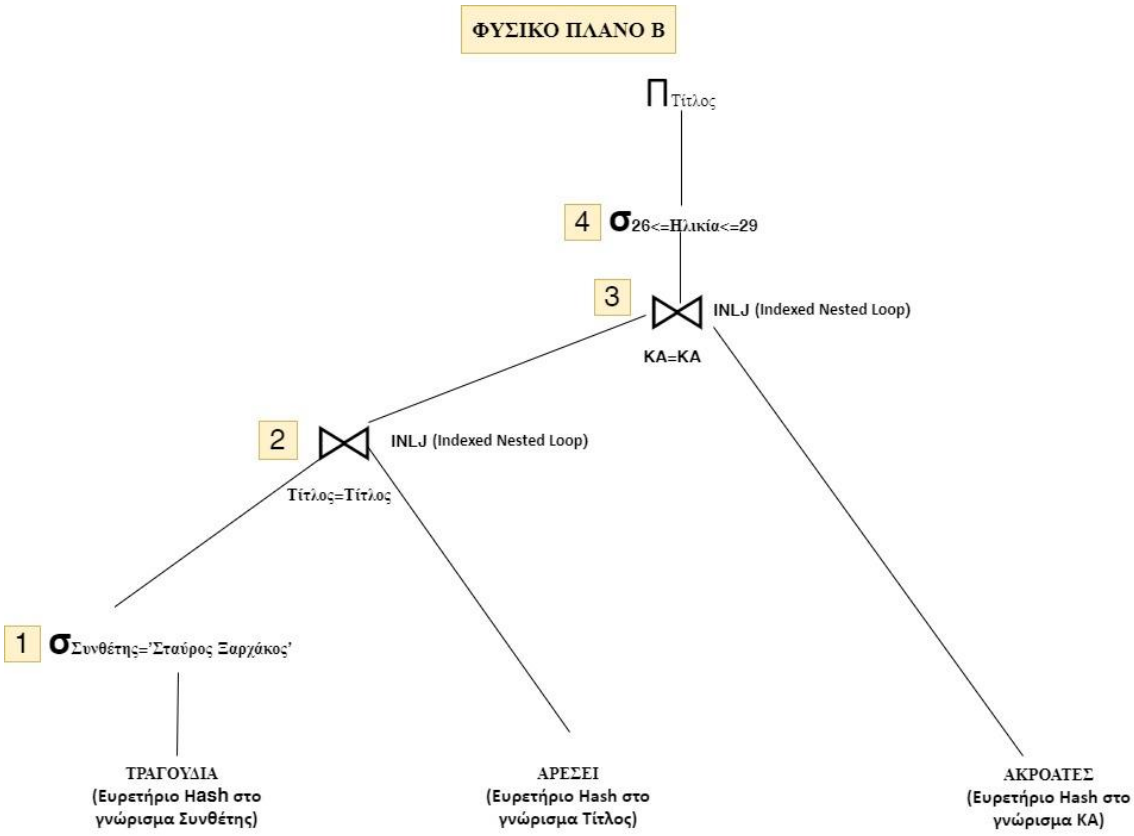
Ζητείται:

Α) Να υπολογίσετε το κόστος σε I/O (εφόσον υφίσταται) για κάθε μία από τις τέσσερις επιμέρους λειτουργίες (αριθμημένες λειτουργίες) του πλάνου Α και να δείξετε πως αυτό προκύπτει. Στη συνέχεια να υπολογίσετε το συνολικό κόστος σε I/O του πλάνου το οποίο προκύπτει από το άθροισμα του κόστους I/O των αριθμημένων λειτουργιών.

Β) Να υπολογίσετε το κόστος σε I/O (εφόσον υφίσταται) για κάθε μία από τις τέσσερις επιμέρους λειτουργίες (αριθμημένες λειτουργίες) του πλάνου Β και να δείξετε πως αυτό προκύπτει. Στη συνέχεια να υπολογίσετε το συνολικό κόστος σε I/O του πλάνου το οποίο προκύπτει από το άθροισμα του κόστους I/O των αριθμημένων λειτουργιών.

Γ) Ποιο πλάνο είναι καλύτερο και γιατί;


ΦΥΣΙΚΟ ΠΛΑΝΟ Α
Π Τίτλος
4
BNLJ (Block Nested Loop)
3
INLJ (Indexed Nested Loop)
ΚΑ=ΚΑ
1 σ Συνθέτης='Σταύρος Ξαρχάκος'
2 σ 26<=Ηλικία<=29
ΤΡΑΓΟΥΔΙΑ
(Ευρετήριο Hash στο γνώρισμα Συνθέτης)
ΑΚΡΟΑΤΕΣ
(Ευρετήριο B+ στο γνώρισμα Ηλικία)
ΑΡΕΣΕΙ
(Ευρετήριο B+ στο γνώρισμα ΚΑ)


ΦΥΣΙΚΟ ΠΛΑΝΟ Β
Π Τίτλος
4 σ 26<=Ηλικία<=29
3
INLJ (Indexed Nested Loop)
ΚΑ=ΚΑ
2
INLJ (Indexed Nested Loop)
Τίτλος=Τίτλος
1 σ Συνθέτης='Σταύρος Ξαρχάκος'
ΤΡΑΓΟΥΔΙΑ
(Ευρετήριο Hash στο γνώρισμα Συνθέτης)
ΑΡΕΣΕΙ
(Ευρετήριο Hash στο γνώρισμα Τίτλος)
ΑΚΡΟΑΤΕΣ
(Ευρετήριο Hash στο γνώρισμα ΚΑ)