Σχεδιασμός Βάσεων Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2022-2023

Δεύτερη Σειρά Ασκήσεων

Ανάθεση: 17-05-2023
Παράδοση: 28-05-2023 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 (100 + 5 μονάδες bonus).

Άσκηση 1 [μονάδες 10]

Έστω η σχέση R(A,B,C,D,E,F) για την οποία ισχύουν τα παρακάτω:

- Η σχέση R περιέχει 1000000 εγγραφές.
- Οι τιμές του γνωρίσματος A κατανέμονται ομοιόμορφα στο διάστημα [1..10000]
- Οι τιμές του γνωρίσματος B κατανέμονται ομοιόμορφα στο διάστημα [1..1000]
- Οι τιμές του γνωρίσματος F κατανέμονται ομοιόμορφα στο διάστημα [1..10]
- Οι τιμές των γνωρισμάτων είναι μεταξύ τους ανεξάρτητες
- Υπάρχει ένα ευρετήριο συστάδων B+ δέντρο στο ζεύγος των γνωρισμάτων (A,B)
- Το ευρετήριο βρίσκεται στην μνήμη
- Σε μία σελίδα χωράνε 20 εγγραφές της σχέσης R

Ζητείται:

A. Να εκτιμήσετε τον αριθμό των εγγραφών στην έξοδο του παρακάτω επερωτήματος:

   SELECT * FROM R WHERE A=1652 AND B>500 AND F>2

B. Να υπολογίσετε το κόστος I/O εκτέλεσης του παραπάνω επερωτήματος με χρήση του ευρετηρίου.

Σε κάθε ένα από τα ερωτήματα A και B να δείξετε τον τρόπο υπολογισμού και όχι μόνο το τελικό αποτέλεσμα.

Άσκηση 2 [Μονάδες 15]

Έστω ο πίνακας T(A int, B int, C int, D int, E varchar(200), F varchar(200), G varchar(200)). Για κάθε αριθμητικό πεδίο του πίνακα (A,B,C και D) υπάρχουν τουλάχιστον 1000 μοναδικές τιμές.

Θεωρείστε τις παρακάτω τρείς ομάδες επερωτημάτων οι οποίες είναι μεταξύ τους ανεξάρτητες. Για κάθε ομάδα ξεχωριστά να προτείνετε **το πολύ δύο ευρετήρια** για τον πίνακα Τ που θεωρείτε ότι θα επιταχύνουν την εκτέλεση των επερωτημάτων και θα μειώσουν σημαντικά τον φόρτο εργασίας (query workload) της αντίστοιχης ομάδας. Σε κάθε περίπτωση να τεκμηριώσετε τους ισχυρισμούς σας.

Για κάθε ευρετήριο **α)** να παραθέσετε το γνώρισμα ή τα γνωρίσματα που σχηματίζουν το κλειδί αναζήτησης του ευρετηρίου **β)** να αναφέρετε αν πρόκειται για ευρετήριο συστάδων (clustered index) ή απλό ευρετήριο (non-clustered index) και **γ)** να προσδιορίσετε το είδος του ευρετηρίου B+tree ή Hash-Index.

Σε κάθε περίπτωση να αιτολογήσετε τις επιλογές σας.

ΟΜΑΔΑ Α	
Συχνότητα εκτέλεσης επερωτήματος	Επερώτημα
100000	E1. SELECT * FROM T WHERE B < ?
10000	E2. SELECT * FROM T WHERE C= ?

ΟΜΑΔΑ Β	
Συχνότητα εκτέλεσης επερωτήματος	Επερώτημα
100000	E1. SELECT * FROM T WHERE B < ? AND C=?
10000	E2. SELECT * FROM T WHERE D=?
1000	E3. SELECT * FROM T WHERE A=?

ΟΜΑΔΑ Γ	
Συχνότητα εκτέλεσης επερωτήματος	Επερώτημα
100000	E1. SELECT A,C FROM T WHERE B < ?
10000	E2. SELECT * FROM T WHERE D < ?

Στη θέση του χαρακτήρα ? να θεωρήσετε ότι υπάρχει μια οποιαδήποτε ακέραια τιμή.

Άσκηση 3 [Μονάδες 30]

Έστω οι παρακάτω σχέσεις:

ΣΚΗΝΟΘΕΤΗΣ(ΚΣ, ΟΝΟΜΑ, ΕΠΩΝΥΜΟ, ΗΛΙΚΙΑ)
ΤΑΙΝΙΑ(ΚΣ, ΤΙΤΛΟΣ,ΕΤΟΣ,ΚΑΤΗΓΟΡΙΑ)

για τις οποίες ισχύει τα εξής:

- Η σχέση ΣΚΗΝΟΘΕΤΗΣ περιέχει 4000 εγγραφές και σε μία σελίδα χωράνε 40 εγγραφές της σχέσης.
- Η σχέση ΤΑΙΝΙΑ περιέχει 20000 εγγραφές και σε μια σελίδα χωράνε 20 εγγραφές της σχέσης.

Επιπλέον θεωρείστε ότι:

- Υπάρχουν 10 διαφορετικές κατηγορίες
- Στο πεδίο ηλικία της σχέσης ΣΚΗΝΟΘΕΤΗΣ το DBMS τηρεί το ακόλουθο ιστόγραμμα:

Ηλικία	Αριθμός Εγγραφών
[20..29]	500
[30..39]	1000
[40..49]	1500
[50..59]	500
[60..69]	500

- Υπάρχει ευρετήριο συστάδων (clustered index) B+ δέντρο στο πεδίο ΣΚΗΝΟΘΕΤΗΣ.ΗΛΙΚΙΑ
- Υπάρχει απλό ευρετήριο (non-clustered index) B+ δέντρο στο πεδίο ΤΑΙΝΙΑ.ΚΑΤΗΓΟΡΙΑ
- Τα ευρετήρια βρίσκονται στην μνήμη του συστήματος.
- Η διαθέσιμη μνήμη είναι M=16
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.

Ζητείται:

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

```
SELECT ΟΝΟΜΑ, ΕΠΩΝΥΜΟ,ΗΛΙΚΑ, ΤΙΤΛΟΣ,ΕΤΟΣ,ΚΑΤΗΓΟΡΑ
FROM ΣΚΗΝΟΘΕΤΗΣ,ΤΑΙΝΙΑ
WHERE ΣΚΗΝΟΘΕΤΗΣ.ΚΣ=ΤΑΙΝΙΑ.ΚΣ AND
      (ΗΛΙΚΙΑ>=38 AND ΗΛΙΚΙΑ <=55) AND ΚΑΤΗΓΟΡΙΑ='Κωμωδία'
```

B. 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 30]

Το ακόλουθο SQL επερώτημα εμφανίζει τα ονόματα των μαθητών των ιδιωτικών λυκείων της χώρας μας που διαγωνίστηκαν στις πανελλήνιες του έτους 2022 στο τέταρτο επιστημονικό πεδίο, και δήλωσαν στο μηχανογραφικό τους το τμήμα Πληροφορικής του Πανεπιστημίου Κύπρου το οποίο φέρει τον κωδικό (ΚΤ) 'ΕΠΛ'.

```
SELECT ONOMA
  FROM ΜΑΘΗΤΕΣ, ΣΧΟΛΕΙΑ, ΔΗΛΩΣΕΙΣ
WHERE ΜΑΘΗΤΕΣ.ΚΣ=ΣΧΟΛΕΙΑ.ΚΣ AND ΜΑΘΗΤΕΣ.ΑΜ=ΔΗΛΩΣΕΙΣ.ΑΜ AND
        ΠΕΔΙΟ=4  AND ΚΑΤΗΓΟΡΙΑ='Ιδιωτικό'  AND ΚΤ='ΕΠΛ'
```

Ακολουθούν ορισμένα στοιχεία για τις παραπάνω σχέσεις:

- ΜΑΘΗΤΕΣ(**ΑΜ,** ΟΝΟΜΑ, ΠΕΔΙΟ, ΚΣ) όπου ΑΜ=Αριθμός Μητρώου του μαθητή.
- ΣΧΟΛΕΙΑ(**ΚΣ**, ΟΝΟΜΑΣΙΑ, ΚΑΤΗΓΟΡΙΑ) όπου ΚΣ=Κωδικός Σχολείου. Το πεδίο ΚΑΤΗΓΟΡΙΑ παίρνει δύο τιμές 'Ιδιωτικό' ή 'Δημόσιο'
- ΔΗΛΩΣΕΙΣ(**ΑΜ,ΚΤ**) όπου ΑΜ=Αριθμός Μητρώου μαθητή και ΚΤ=Κωδικός Τμήματος.
- Η σχέση ΜΑΘΗΤΕΣ έχει 1200 εγγραφές οι οποίες χωρούν σε 100 σελίδες.
- Η σχέση ΣΧΟΛΕΙΑ έχει 600 εγγραφές οι οποίες χωρούν σε 120 σελίδες.
- Η σχέση ΔΗΛΩΣΕΙΣ έχει 12000 εγγραφές οι οποίες χωρούν σε 400 σελίδες.
- Τα πρωτεύοντα κλειδιά των σχέσεων είναι υπογραμμισμένα.

Επιπλέον δίνεται ότι:

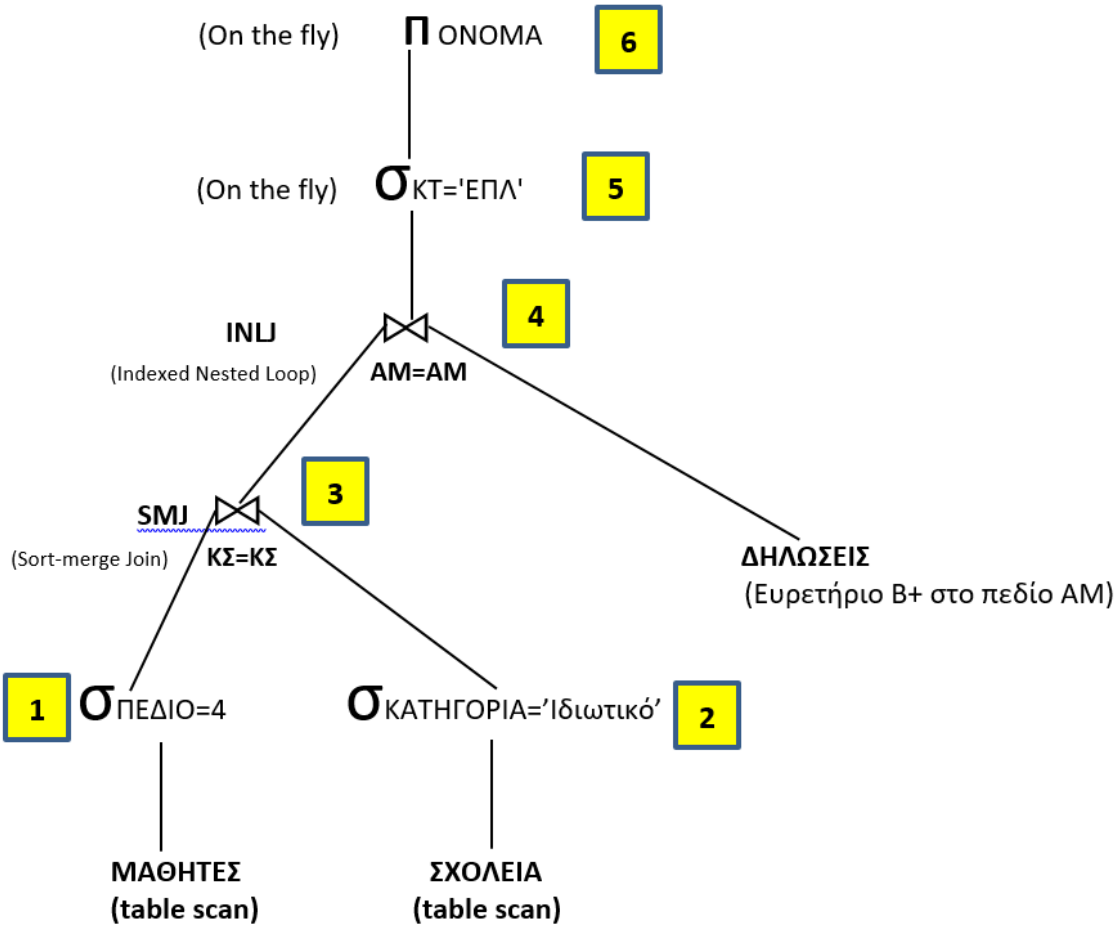
- Τα επιστημονικά πεδία είναι 4 και οι μαθητές κατανέμονται ομοιόμορφα στα 4 επιστημονικά πεδία.
- Το 5% των του συνόλου των λυκείων είναι ιδιωτικά.
- Κάθε μαθητής μπορεί να δηλώσει μέχρι και 10 τμήματα του επιστημονικού πεδίου που διαγωνίστηκε.
- Το πεδίο ΜΑΘΗΤΕΣ.ΚΣ είναι ξένο κλειδί το οποίο αναφέρεται (references) στο πεδίο ΣΧΟΛΕΙΑ.ΚΣ
- Το πεδίο ΔΗΛΩΣΕΙΣ.ΑΜ είναι ξένο κλειδί το οποίο αναφέρεται (references) στο πεδίο ΜΑΘΗΤΕΣ.ΑΜ.
- Υπάρχει ένα **απλό ευρετήριο (non-clustered index)** B+δέντρο στο πεδίο ΔΗΛΩΣΕΙΣ.ΑΜ και όλες οι σελίδες του ευρετηρίου βρίσκονται στην κύρια μνήμη. Αυτό είναι το μόνο ευρετήριο που υπάρχει. **Μην θεωρήσετε ότι στο πρωτεύον κλειδί κάθε πίνακα υπάρχει ευρετήριο συστάδων (clustered index).**
- Κανένα πεδίο των παραπάνω σχέσεων δεν δέχεται τιμές NULL.
- Το μέγεθος της διαθέσιμης μνήμης είναι M=5 σελίδες.
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.

Ζητείται:

A. Να υπολογίσετε το κόστος σε I/O του φυσικού πλάνου εκτέλεσης που ακολουθεί. 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.

B. Πως μεταβάλλεται το κόστος δεδομένου ότι το μέγεθος της διαθέσιμης μνήμης είναι M=32 (αντί για M=5) και το ευρετήριο που υπάρχει στο πεδίο ΔΗΛΩΣΕΙΣ.ΑΜ είναι ευρετήριο συστάδων (clustered index) και όχι απλό ευρετήριο.

ΦΥΣΙΚΟ ΠΛΑΝΟ ΕΚΤΕΛΕΣΗΣ

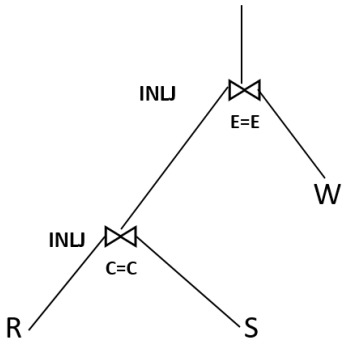


Άσκηση 5 [20 μονάδες]

Δίνονται οι σχέσεις R(**A**,B,C), S(**C**,D,E) και W(**E**,F) των οποίων τα πρωτεύοντα κλειδιά είναι υπογραμμισμένα για τις οποίες ισχύουν τα ακόλουθα:

- T(R)=1000, B(R)=100, V(R.C)=125
- T(S)=1500
- T(W)=750
- Υπάρχει ένα απλό ευρετήριο (non-clustered index) στο πεδίο S.C
- Υπάρχει ένα απλό ευρετήριο (non-clustered index) στο πεδίο W.E
- Κανένα πεδίο δεν περιέχει τιμές NULL.
- Το πεδίο R.C είναι ξένο κλειδί το οποίο αναφέρεται στο πεδίο S.C
- Το πεδίο S.E είναι ξένο κλειδί το οποίο αναφέρεται στο πεδίο W.E
- Τα ευρετήρια βρίσκονται στην μνήμη.

Θεωρείστε ότι η σύζευξη των τριών σχέσεων γίνεται σύμφωνα με το ακόλουθο φυσικό πλάνο:



Ζητείται

A. Να υπολογίσετε τον αριθμό των εγγραφών στην έξοδο ως αποτέλεσμα της σύζευξης των σχέσεων R,S και W. Να δείξετε τον τρόπο υπολογισμού όχι μόνο το τελικό αποτέλεσμα.
B. Να εκτιμήσετε το κόστος I/O του παραπάνω φυσικού πλάνου.
C. Θεωρείτε ότι η δημιουργία ενός ευρετηρίου συστάδων στο πεδίο R.C θα επιταχύνει τον υπολογισμό της παραπάνω σύζευξης; Να αιτιολογήσετε την απάντησή σας.