Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2021-2022

Δεύτερη Σειρά Ασκήσεων

Ανάθεση: 18-05-2022
Παράδοση: 30-05-2022 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 ΑΜ είναι ο αριθμός μητρώου σας. π.χ. "3190001.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 [25 μονάδες]

Έστω οι σχέσεις R και S για τι οποίες ισχύουν:

- Η σχέση R έχει 24.000 εγγραφές και σε μια σελίδα χωρούν 20 εγγραφές της σχέσης R.
- Η σχέση S έχει 40.000 εγγραφές και σε μία σελίδα χωρούν 50 εγγραφές της σχέσης S.

Το μέγεθος της διαθέσιμης μνήμης είναι **M=31** σελίδες.

Ζητείται:

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

2. Δεδομένου ότι η διαθέσιμη μνήμη είναι **M=10** να υπολογίσετε το **βέλτιστο κόστος** του αλγορίθμου SMJ για την ισοσύνδεση των σχέσεων **R⋈S** και να δείξετε αναλυτικά πως αυτό προκύπτει. Με άλλα λόγια, να δείξετε τον αριθμό και το μέγεθος (σε σελίδες) όλων των ταξινομημένων λιστών που θα δημιουργήσει ο αλγόριθμος για κάθε μία από τις σχέσεις R και S, καθώς επίσης και το συνολικό κόστος (σε I/O) για την επίτευξη της ισοσύνδεσης.

Άσκηση 2 [25 μονάδες]

Έστω οι σχέσεις R (a,b,c,d) και S (d,e,f) για τι οποίες ισχύουν τα ακόλουθα:

1. T(R)=30000, B(R)=1500 και V(R,c)=50
2. T(S)=50000, B(S)=2000 και V(S,f)=100
3. Υπάρχει ένα απλό ευρετήριο (non-clustered index) στο γνώρισμα R.c
4. Υπάρχει ένα ευρετήριο συστάδων (clustered index) στο γνώρισμα S.f

Το μέγεθος της διαθέσιμης μνήμης είναι **M=10** σελίδες, τα ευρετήρια βρίσκονται στην μνήμη και τα δεδομένα κατανέμονται ομοιόμορφα.

Έστω το παρακάτω επερώτημα διατυπωμένο σε σχεσιακή άλγεβρα:

$$\pi_{R.a}\,(\sigma_{c=100 \text{ and } f=50}\,(R \bowtie S))$$

Ζητείται:

1. Να σχεδιάσετε το αρχικό λογικό πλάνο του επερωτήματος. Αφού εφαρμόσετε τους απαραίτητους μετασχηματισμούς να σχεδιάσετε το τελικό λογικό πλάνο (δεν χρειάζεται να δείξετε τα ενδιάμεσα πλάνα).

2. Να υπολογίσετε το **ελάχιστο δυνατό** κόστος εκτέλεσης του πλάνου όταν ο αλγόριθμος που χρησιμοποιείται για την σύζευξη είναι:

   a. NLJ (Block Nested Loop Join)
   b. SMJ (Sort Merge Join)

Άσκηση 3 [25 μονάδες]

Ο πίνακας players περιέχει πληροφορίες για τους 60000 ποδοσφαιριστές που αγωνίζονται στα πρωταθλήματα της Ευρώπης:

Players

Γνώρισμα	**Τύπος**	**Μέγεθος σε bytes**
pid (primaryKey)	int	4
lastname	char	40
firstname	char	34
age	tinyint	1
rating	tinyint	1

Κανένα από τα πεδία του πίνακα δεν δέχεται τιμές NULL. Για το πεδίο **age** το DBMS τηρεί το ακόλουθο ιστόγραμμα:

Διάστημα Τιμών Πεδίου Age	**Αριθμός εγγραφών του Πίνακα Players**
[17..21]	12000
[22..26]	24000
[27..31]	18000
[32..36]	6000

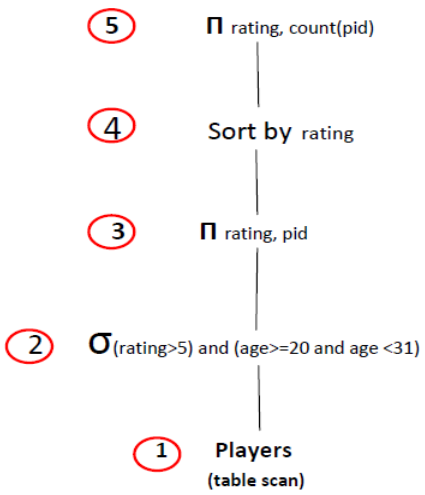
Για το πεδίο **rating** υπάρχουν **10** διακριτές τιμές **[1..10]** οι οποίες κατανέμονται ομοιόμορφα.

Για την αποθήκευση του πίνακα players απαιτούνται **600 σελίδες (blocks)** στον δίσκο και η διαθέσιμη μνήμης είναι **M=8** σελίδες.

Έστω ότι θέλουμε να εκτελέσουμε το παρακάτω ερώτημα:

```
Select  rating, count(pid)
  from players
where (rating > 5) and (age >=20 and age <31)
group by rating
```

Δεδομένου ότι δεν υπάρχει κάποιο ευρετήριο στον πίνακα players το DBMS παράγει το ακόλουθο πλάνο εκτέλεσης:

	1. Διαβάζει τον πίνακα players (table scan). 2. Επιλέγει τις πλειάδες που πληρούν το κριτήριο αναζήτησης. 3. Προβάλει τα πεδία rating και pid του αποτελέσματος της επιλογής. 4. Ταξινομεί το αποτέλεσμα της προβολής για να σχηματίσει τις ομάδες. 5. Προβάλει το τελικό αποτέλεσμα.

Να θεωρήσετε ότι οι επιλογές είναι μεταξύ τους ανεξάρτητες.

Ζητείται:

1. Για κάθε μια από τις **5** αριθμημένες λειτουργίες να προσδιορίσετε τ**ον αριθμό των εγγραφών που προκύπτουν στην έξοδο** καθώς **και το κόστος I/O** (εφόσον υφίσταται) και να δείξετε πως αυτά υπολογίζονται.

2. Να υπολογίσετε το **συνολικό κόστος σε I/O** του παραπάνω φυσικού πλάνου εκτέλεσης.

3. Πως διαφοροποιείται το συνολικό κόστος αν η διαθέσιμη μνήμη χωράει 15 σελίδες (M=15);

Άσκηση 4 [25 μονάδες]

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

ΠΡΟΙΟΝΤΑ (Barcode, Όνομα, Κατηγορία)
ΚΑΤΑΣΤΗΜΑΤΑ (Κωδικός, Κατάστημα, Πόλη, Διεύθυνση)
ΠΩΛΗΣΕΙΣ (Barcode, Κωδικός, Ημερομηνία)

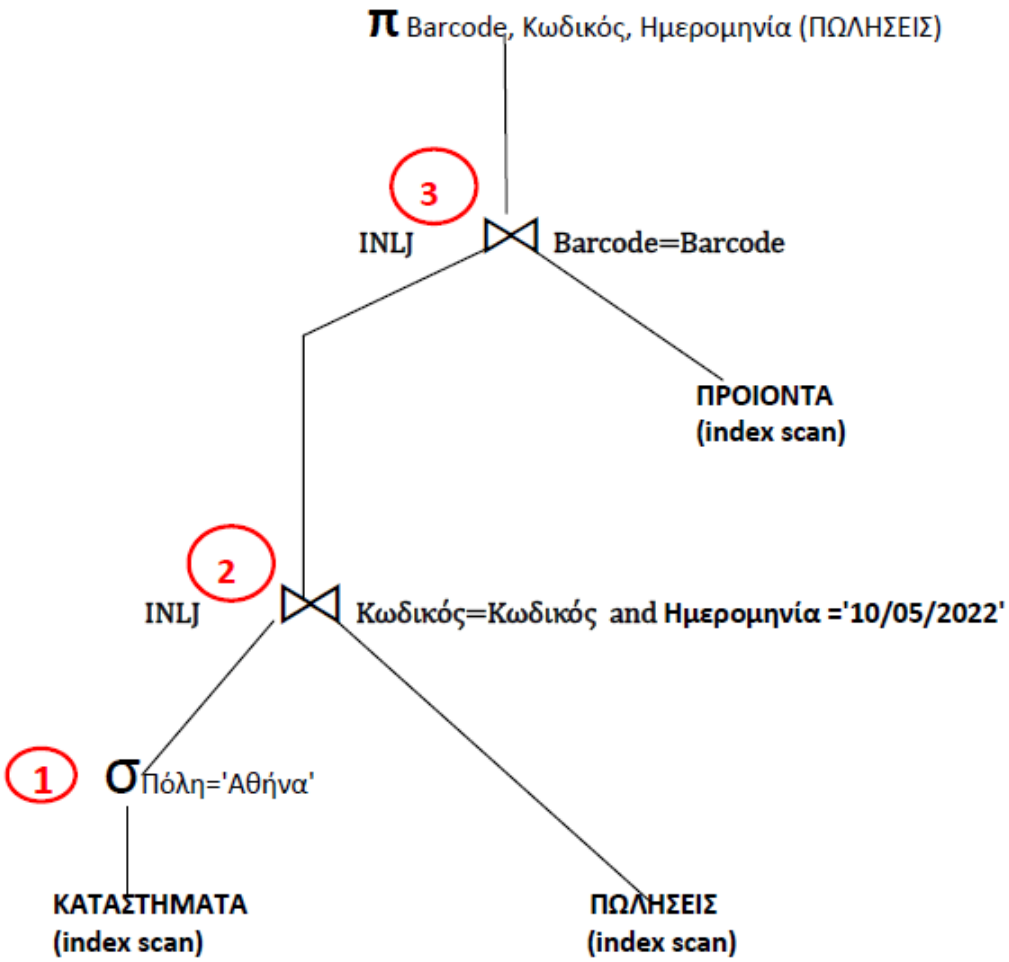
Θεωρείστε ότι:

1. Η σχέση ΠΡΟΙΟΝΤΑ περιέχει 50.000 πλειάδες και σε μία σελίδα χωρούν 10 εγγραφές.
2. Υπάρχει ένα απλό (non clustered) B+ ευρετήριο στο πεδίο ΠΡΟΙΟΝΤΑ.Barcode.
3. Η σχέση ΚΑΤΑΣΤΗΜΑΤΑ περιέχει 1000 εγγραφές και σε μια σελίδα χωρούν 5 εγγραφές.
4. Υπάρχει ένα B+ ευρετήριο συστάδων (clustered index) στο γνώρισμα Πόλη.
5. Το δίκτυο των καταστημάτων καλύπτει 100 διαφορετικές πόλεις.
6. Η σχέση ΠΩΛΗΣΕΙΣ περιέχει 500.000 εγγραφές και σε μια σελίδα χωρούν 25 εγγραφές.
7. Καθημερινά κάθε υποκατάστημα πωλεί 10 προϊόντα.
8. Υπάρχει ένα B+ ευρετήριο συστάδων (clustered index) στο ζεύγος των πεδίων (Κωδικός, ημερομηνία) της σχέσης ΠΩΛΗΣΕΙΣ.
9. Το μέγεθος της διαθέσιμης μνήμης είναι 50 σελίδες (M=50).
10. Όλα τα ευρετήρια βρίσκονται στην μνήμη.
11. Όπου απαιτείται θεωρείστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.
12. Οι επιλογές είναι μεταξύ τους ανεξάρτητες.

Δίνεται το ακόλουθο επερώτημα:

```
SELECT  ΠΩΛΗΣΕΙΣ.*
   FROM ΠΡΟΙΟΝΤΑ, ΚΑΤΑΣΤΗΜΑΤΑ, ΠΩΛΗΣΕΙΣ
WHERE ΠΡΟΙΟΝΤΑ.Barcode=ΠΩΛΗΣΕΙΣ.Barcode AND
          ΠΩΛΗΣΕΙΣ.Κωδικός=ΚΑΤΑΣΤΗΜΑΤΑ.Κωδικός AND
          Πόλη='Αθήνα' AND   Ημερομηνία='10/05/2022'
```

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 3 αριθμημένες επιμέρους λειτουργίες του πλάνου και να δείξετε πως αυτό προκύπτει. Επιπλέον για κάθε μια από τις 3 λειτουργίες να προσδιορίσετε τον αριθμό των εγγραφών που προκύπτουν στην έξοδο και να δείξετε πως αυτός υπολογίζεται.


π Barcode, Κωδικός, Ημερομηνία (ΠΩΛΗΣΕΙΣ)
3
INLJ
Barcode=Barcode
ΠΡΟΙΟΝΤΑ
(index scan)
2
INLJ
Κωδικός=Κωδικός and Ημερομηνία ='10/05/2022'
1
σ Πόλη='Αθήνα'
ΚΑΤΑΣΤΗΜΑΤΑ
(index scan)
ΠΩΛΗΣΕΙΣ
(index scan)