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 (100 + 5 μονάδες bonus).

Λύση Άσκησης 1

Α)

Το επερώτημα Q1 έχει στην έξοδο:

T(R1)/(V(R1,a) * V(R1,b))=5.000.000/(100*50)=1000 πλειάδες. Επειδή το ευρετήριο στο ζεύγος των γνωρισμάτων (a,b) είναι απλό θα χρειαστεί να διαβάσει από το δίσκο 1000 σελίδες (κάθε εγγραφή μπορεί να βρίσκεται σε διαφορετική σελίδα). Συνεπώς το Κόστος εκτέλεσης του επερωτήματος Q1 είναι **1000 I/O.**

Το επερώτημα Q2 έχει στην έξοδο:

4*(2500/100)+100*(500/100)+20*(4000/200)=4*25+500+20*20=1000. Επειδή το ευρετήριο στο γνώρισμα s είναι απλό θα χρειαστεί να διαβάσει 1000 σελίδες (κάθε εγγραφή μπορεί να βρίσκεται σε διαφορετική σελίδα). Συνεπώς το Κόστος εκτέλεσης του επερωτήματος Q2 είναι 1000 I/O.

Συμπερασματικά και τα δύο επερωτήματα έχουν το ίδιο κόστος εκτέλεσης σε I/O.

Β)

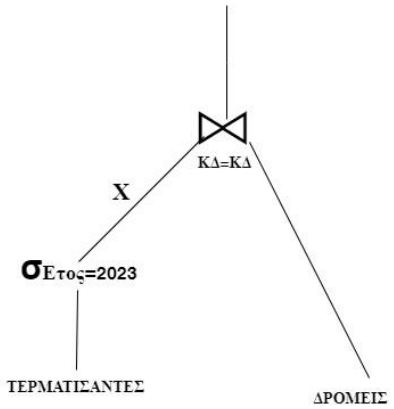
Μία σελίδα χωράει T(R1)/B(R1) = 5.000.000/50.000=100 εγγραφές της σχέσης R1. Άρα οι 1000 εγγραφές της σχέσης R1 χωράνε σε 1000/100=10 σελίδες.

Μία σελίδα χωράει T(R2)/B(R2) = 10000/2000=5 εγγραφές της σχέσης R2. Άρα οι 1000 εγγραφές της σχέσης R1 χωράνε σε 1000/5=200 σελίδες.

Συνεπώς στην περίπτωση που και τα δύο ευρετήρια είναι ευρετήρια συστάδων το επερώτημα Q1 θα εκτλεστεί πιο γρήγορα διότι έχει μικρότερο κόστος I/O.

Λύση Άσκησης 2

A)



B)

Σε μια σελίδα χωράνε T(ΔΡΟΜΕΙΣ)/B(ΔΡΟΜΕΙΣ)=40000/200=200 εγγραφές της σχέσης ΔΡΟΜΕΙΣ.

Σε μία σελίδα χωράνε T(ΤΕΡΜΑΤΙΣΑΝΤΕΣ)/B(ΤΕΡΜΑΤΙΣΑΝΤΕΣ)=60000/600=100 εγγραφές της σχέσης ΤΕΡΜΑΤΙΣΑΝΤΕΣ.

T(X)=60000/4=15000 και B(X)=15000/100=150, **cost(X)=B(X)=150** (Λόγω του ευρετηρίου συστάδων στο γνώρισμα έτος).

SMJ

Η σχέση ΔΡΟΜΕΙΣ είναι ταξινομημένη ως προς το πεδίο ΔΡΟΜΕΙΣ.ΚΔ λόγω του ευρετηρίου συστάδων στο γνώρισμα ΚΔ. Επειδή B(X)=150 < 21*21:

cost(SMJ)=cost(X) + 2*B(X) + B(ΔΡΟΜΕΙΣ) = 150 + 2*150 + 200 = **650 I/O**

Κόστος NLJ με εξωτερική την σχέση X

cost(NLJ)=cost(X) + ceil(B(X)/M-1)*B(ΔΡΟΜΕΙΣ)=150 + ceil(150/20)*200=150+8*200=**1750**

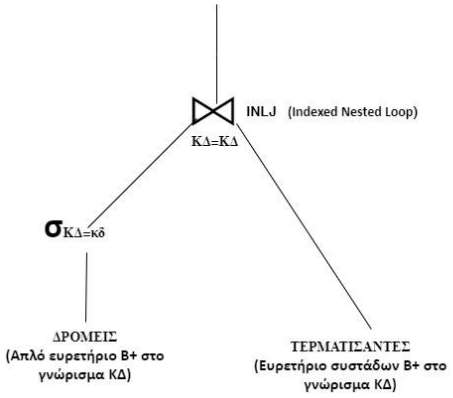
Κόστος NLJ με εξωτερική την σχέση ΔΡΟΜΕΙΣ

cost(NLJ)=B(ΔΡΟΜΕΙΣ)+ ceil(B(ΔΡΟΜΕΙΣ)/M-1)*cost(X) =200+(200/20)*150=200+1500=**1700 I/O.**

Λύση Άσκησης 3

1. Δημιουργώ απλό ευρετήριο B+ δέντρο στο πρωτεύον κλειδί (ΚΔ) της σχέσης ΔΡΟΜΕΙΣ.
2. Δημιουργώ ευρετήριο συστάδων B+ στο πεδίο ΚΔ της σχέσης ΤΕΡΜΑΤΙΣΑΝΤΕΣ
3. Επιλέγω τον αλγόριθμο INLJ για την ισοσύνδεση.

Φυσικό Πλάνο



Υπολογισμός Κόστους

Δεδομένου ότι το γνώρισμα ΚΔ είναι πρωτεύον κλειδί στον πίνακα ΔΡΟΜΕΙΣ χρησιμοποιώντας το απλό ευρετήριο στο γνώρισμα ΔΡΟΜΕΙΣ.ΚΔ θα διαβάσουμε μόνο μία σελίδα, η οποία περιέχει την εγγραφή του δρομέα με ΚΔ=κδ.

Για την υλοποίηση της ισοσύνδεσης χρησιμοποιούμε τον αλγόριθμο INLJ με εξωτερική την σχέση ΔΡΟΜΕΙΣ και εσωτερική την σχέση ΤΕΡΜΑΡΤΙΣΑΝΤΕΣ. Ο αλγόριθμος INLJ θα δεχθεί στην είσοδο τον κωδικό του δρομέα κδ και χρησιμοποιώντας το ευρετήριο συστάδων στο γνώρισμα ΤΕΡΜΑΤΙΣΑΝΤΕΣ.ΚΔ θα χρειαστεί να διαβάσει μόνο την σελίδα της σχέσης ΤΕΡΜΑΤΙΣΑΝΤΕΣ η οποία θα περιέχει τις συμμετοχές του δρομέα (οι οποίες είναι το πολύ 4 και χωράνε σε μια σελίδα).

ΚΟΣΤΟΣ= 1 +1 =2 I/O

Το πλάνο είναι το βέλτιστο διότι διαβάζουμε μόνο τις σελίδες που περιέχουν τα δεδομένα που θέλουμε.

Λύση Άσκησης 4

Από τα δεδομένα της άσκησης έχουμε:

Τ(ΑΚΡΟΑΤΕΣ)=30000, Β(ΑΚΡΟΑΤΕΣ)=30000/10=3000
Τ(ΤΡΑΓΟΥΔΙΑ)=1000, Β(ΤΡΑΓΟΥΔΙΑ)=1000/5=200
Τ(ΑΡΕΣΕΙ)=500000,Β(ΑΡΕΣΕΙ)=500000/10000=50

Υπολογισμός Κόστους Ι/Ο του Πλάνου Α

COST(1)
Δεδομένου ότι Τ(ΤΡΑΓΟΥΔΙΑ)=1000 και υπάρχουν 100 διαφορετικοί συνθέτες έχουμε ότι: Τ(1)=1000/100=10 και Β(1)=2. Συνεπώς **COST(1)=Τ(1)=10 Ι/Ο.**

COST(2)
Δεδομένου ότι ή έρευνα διεξήχθη σε ακροατές 21 έως και 60 ετών (40 διακριτές τιμές) και τα δεδομένα κατανέμονται ομοιόμορφα έχουμε ότι: Τ(2)=(30000/40)*4=3000 και Β(2)=3000/10=300. **Συνεπώς COST(2)=Β(2)=300.**

COST(3)
Για κάθε μία από τις 3000 εγγραφές που δέχεται στην είσοδο ο αλγόριθμος INLJ (βήμα 3) χρησιμοποιεί απλό ευρετήριο Β+ που υπάρχει στο γνώρισμα ΚΑ της σχέσης ΑΡΕΣΕΙ (look-up/probe). Δεδομένου ότι κάθε ακρατής έχει δηλώσει ότι του αρέσουν 17 τραγούδια, για κάθε κωδικό ακροατή (ΚΑ) θα ανακτηθούν το πολύ 17 εγγραφές της σχέσης ΑΡΕΣΕΙ, οι οποίες χωράνε σε μία σελίδα. **Συνεπώς COST(3)=3000.**

COST(4)
Όπως υπολογίσαμε στο βήμα 1 το κόστος της επιλογής με βάση τον συνθέτη είναι 10 Ι/Ο. Δεδομένου ότι η διαθέσιμη μνήμη είναι 62 σελίδες, μπορούμε να κρατήσουμε στην μνήμη τις 10 σελίδες που προκύπτουν στην έξοδο της επιλογής του βήματος (1) και στην συνέχεια να εκτελέσουμε τα βήματα 2 και 3. Με τον τρόπο αυτό οι εγγραφές που προκύπτουν στην έξοδο του τελεστή INLJ (βήμα 3) διοχετεύονται ως είσοδο στην ισοσύνδεση BLNJ (βήμα 4) η οποία μπορεί να υπολογιστεί δίχως επιπλέον κόστος Ι/Ο. **Συνεπώς COST(4)=0.**

Συνολικό κόστος πλάνου Α: 10+300+3000 = 3310 Ι/Ο.

Υπολογισμός Κόστους Ι/Ο του Πλάνου Β

COST(1)
Δεδομένου ότι Τ(ΤΡΑΓΟΥΔΙΑ)=1000 και υπάρχουν 100 διαφορετικοί συνθέτες έχουμε ότι: Τ(1)=1000/100=10 και Β(1)=2. Συνεπώς **COST(1)=Τ(1)=10 Ι/Ο.**

COST(2)
Για κάθε μία από τις 10 εγγραφές τραγουδιών που δέχεται στην είσοδο ο αλγόριθμος INLJ (ισοσύνδεση βήματος 2) χρησιμοποιεί το απλό ευρετήριο κατακερματισμού που υπάρχει στο πεδίο τίτλος της σχέσης αρέσει. Δεδομένου ότι: α) υπάρχουν 1000 διαφορετικά τραγούδια, β) η σχέση αρέσει περιέχει 500000 εγγραφές και γ) υποθέτουμε ομοιόμορφη κατανομή, για κάθε τραγούδι που δέχεται στην είσοδο ο αλγόριθμος INLJ θα ανακτήσει 500000/1000=500 εγγραφές της σχέσης αρέσει. Συνεπώς **COST(2)=10*500=5000 Ι/Ο.**

COST(3)
Για κάθε μία από τις 5000 εγγραφές που δέχεται στην είσοδο ο αλγόριθμος INLJ (βήμα 3) χρησιμοποιεί το απλό ευρετήριο κατακερματισμού που υπάρχει στο γνώρισμα ΚΑ της σχέσης ακροατές για να ανακτήσει τα στοιχεία του ακροατή. **Συνεπώς COST(3)=5000.**

COST(4)
Η επιλογή μπορεί να γίνει στην μνήμη δίχως κόστος Ι/Ο. **Συνεπώς COST(4)=0**

Συνολικό κόστος πλάνου Β: 10+5000+5000=10.010 Ι/Ο.