

Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων

Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2024-2025

Λύσεις Δεύτερης Σειράς Ασκήσεων

Λύση Άσκησης 1

1.

Υπολογισμός T(A)

Από τα δεδομένα έχουμε $T(\text{Cities})=5000$ και $V(\text{Cities},\text{country})=100$. Συνεπώς,

$$T(A)=T(\text{Cities})/V(\text{cities},\text{country})=5000/100 = 50$$

Υπολογισμός T(B)

Από τα δεδομένα ισχύει ότι $T(\text{Flights})=100.000$

Χρησιμοποιώντας το ιστόγραμμα στο γνώρισμα duration έχουμε:

$$T(B)=(100.000*0,3)/2 + 100.000*0,4 + (100.000 *0,15)/3 = 15000 + 40000 + 5000=60.000$$

Υπολογισμός T(Γ)

Επειδή το γνώρισμα `from_cid` είναι ξένο κλειδί το οποίο αναφέρεται στο πρωτεύον κλειδί `cid` της σχέσης `cities` **κάθε μια εγγραφή** της σχέσης `flights` κάνει join **με μία μόνο εγγραφή** της σχέσης `cities`. Στην περίπτωση μας πρέπει να υπολογίσουμε πόσες από τις 60000 εγγραφές πτήσεων που έχουμε στην έξοδο B ταιριάζουν με τις 50 πόλεις στην έξοδο A. Συνεπώς έχουμε:

$$T(\Gamma) = T(B)*T(A)/T(\text{cities})=60000*50/5000=600$$

Υπολογισμός T(Δ)

Οι εγγραφές στην έξοδο Γ συνδέονται με την σχέση `Airlines` με την σχέση πρωτεύοντος ξένου κλειδιού (`Flights.aid=Airlines.aid`). Δηλαδή κάθε μια από τις 600 εγγραφές στην έξοδο Γ κάνει join με μια ακριβώς εγγραφή της σχέσης `Airlines`. Συνεπώς,

$$T(\Delta)=T(\Gamma)=600$$

2.

Έστω ότι Q1 είναι η πρώτη επερώτηση του τελεστή UNION και Q2 η δεύτερη.

```
SELECT * FROM Flights WHERE fid > 50000 (Q1)
```

```
UNION
```

```
SELECT * FROM Flights WHERE duration >=2 and duration<6 (Q2)
```

Από τα δεδομένα της άσκησης γνωρίζουμε ότι $T(\text{Flights})=100000$ και ότι το πρωτεύον κλειδί της σχέσης λαμβάνει ακέραιες τιμές στο διάστημα $[1..100000]$. Συνεπώς **$T(Q1)=50000$** εγγραφές.

Ο αριθμός των εγγραφών στην έξοδο του T(Q2) είναι το T(B) που υπολογίσαμε στο πρώτο ζητούμενο της άσκησης. Συνεπώς **T(Q2)=60000**

Ο τελεστής UNION υπολογίζει το σύνολο των εγγραφών που προκύπτουν από τις δύο επερωτήσεις. Σε περίπτωση που υπάρχουν κοινές εγγραφές αυτές εμφανίζονται μόνο μία φορά. Συνεπώς πρέπει να αφαιρέσουμε τις κοινές εγγραφές.

Η πιθανότητα μια εγγραφή να εμφανίζεται στο αποτέλεσμα και των δύο επερωτήσεων (Q1) και (Q2) είναι: $50000/100000 * 60000/100000 = 0,5 * 0,6 = 0,3$. Οπότε από τις 100.000 εγγραφές της σχέσης Flights οι $0,3 * 100000 = 30000$ μπορεί να είναι κοινές.

Με βάση τα παραπάνω έχουμε: $T(UNION) = 50000 + 60000 - 30000 = 80.000$ εγγραφές.

Λύση Άσκησης 2

1. Ο αλγόριθμος SMJ θα δημιουργήσει τις παρακάτω ταξινομημένες λίστες για την σχέση R1:

- Στο πρώτο πέρασμα θα διαβάσει τη σχέση R1 και θα γράψει 120 ταξινομημένες λίστες μεγέθους 5 σελίδων έκαστη. **Κόστος= $2 * B(R1)$.**
- Στο δεύτερο πέρασμα θα διαβάσει τις 120 ταξινομημένες λίστες και θα τις συγχωνεύσει σε 24 λίστες μεγέθους 25 σελίδων έκαστη. **Κόστος= $2 * B(R1)$.**
- Στο τρίτο πέρασμα θα διαβάσει τις 24 λίστες και θα τις συγχωνεύσει σε 5. Τέσσερις λίστες μεγέθους 125 σελίδων έκαστη και μία λίστα μεγέθους 100 σελίδων. **Κόστος= $2 * B(R1)$.**
- Στο τέταρτο πέρασμα θα διαβάσει τις 5 λίστες και θα τις συγχωνεύσει σε μία λίστα 600 σελίδων. **Κόστος= $2 * B(R1)$.**

Ο αλγόριθμος SMJ θα δημιουργήσει τις παρακάτω ταξινομημένες λίστες για την σχέση R2:

- Στο πρώτο πέρασμα θα διαβάσει την σχέση R2 και θα γράψει 80 ταξινομημένες λίστες μεγέθους 5 σελίδων έκαστη. **Κόστος= $2 * B(R2)$.**
- Στο δεύτερο πέρασμα θα διαβάσει τις 80 ταξινομημένες λίστες και θα τις συγχωνεύσει σε 16 λίστες μεγέθους 25 σελίδες έκαστη. **Κόστος= $2 * B(R2)$.**
- Στο τρίτο πέρασμα θα διαβάσει τις 16 λίστες και θα τις συγχωνεύσει σε 4. Τρεις λίστες μεγέθους 125 σελίδων και μία λίστα μεγέθους 75 σελίδων. **Κόστος= $2 * B(R2)$.**

Στο σημείο αυτό ο αλγόριθμος SMJ έχει δημιουργήσει στον δίσκο 1 ταξινομημένη λίστα για τη σχέση R1 και 4 ταξινομημένες λίστες για την σχέση R2. Η μνήμη χωράει 5 σελίδες οπότε ο αλγόριθμος SMJ θα προχωρήσει στην συγχώνευση των λιστών και θα δώσει το αποτέλεσμα στην έξοδο. Το κόστος του output δεν προσμετράται. **Κόστος= $B(R1) + B(R2)$**

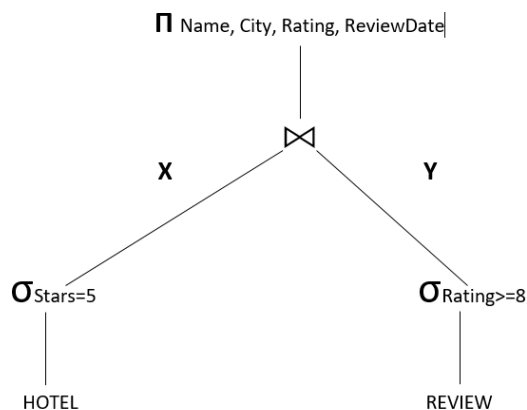
Συνολικό Κόστος = $9 * B(R1) + 7 * B(R2) = 9 * 600 + 7 * 400 = 5400 + 2800 = 8200$ I/O.

2. Δημιουργία δύο ευρετηρίων συστάδων B+tree στα πεδία R1.b και R2.b.

$COST(SMJ) = B(R1) + B(R2) = 600 + 400 = 1000$.

Λύση Άσκησης 3

A)



B)

Από τα δεδομένα της άσκησης έχουμε ότι:

- $T(\text{HOTEL})=20.000$ και $B(\text{HOTEL})=20.000/50=400$
- $T(\text{REVIEW})=400.000$ και $B(\text{REVIEW})=400.000/500=800$

Έστω ότι $X=\sigma_{\text{Stars}=5}\text{HOTEL}$ και $Y=\sigma_{\text{Rating}\geq 8}\text{REVIEW}$

Επειδή τα ξενοδοχεία 5 αστέρων ανέρχονται στο 10% του συνόλου έχουμε:

$T(X)=20000 * 10/100=2000$ και $B(X)=2000/50=40$

Το κόστος υπολογισμού του X με table scan είναι $B(\text{HOTEL})=400$ ενώ με την χρήση του ευρετηρίου στο πεδίο HOTEL.Stars είναι $T(X)=2000$. Επιλέγουμε το μικρότερο οπότε κόστος $\text{cost}(X)=400$.

Με βάση το ιστόγραμμα στο πεδίο REVIEW.Rating έχουμε:

$T(Y)=(400.000*3/10)=120000$, $B(Y)=120000/500=240$

Το κόστος υπολογισμού του Y με table scan είναι $B(\text{REVIEW})=800$ ενώ με την χρήση του ευρετηρίου συστάδων στο πεδίο REVIEW.Rating είναι $B(Y)=240$. Επιλέγουμε το μικρότερο οπότε κόστος $\text{cost}(Y)=240$.

Κόστος SMJ

Δεδομένου ότι η διαθέσιμη μνήμη είναι $M=21$ ισχύει ότι $M^2 \geq B(X) + B(Y)$ ($21*21 > 40+240$) μπορούμε να εφαρμόσουμε την αποδοτική έκδοση του αλγορίθμου:

$\text{cost}(\text{SMJ})=\text{cost}(X)+\text{cost}(Y) + 2*B(X)+2*B(Y)= 400+240+ 2*40 + 2*240 = 1200$ I/O.

Κόστος NLJ με εξωτερική την σχέση X

$\text{cost}(\text{NLJ})=\text{cost}(X) + \text{ceil}(B(X)/M-1)*\text{cost}(Y)=400 + \text{ceil}(40/20)*240=400+2*240= 880$ I/O.

Κόστος NLJ με εξωτερική την σχέση Y

$\text{cost}(\text{NLJ})=\text{cost}(Y) + \text{ceil}(B(Y)/M-1)*\text{cost}(X)=240 + \text{ceil}(240/20)*400=240 + 12*400=5040$ I/O.

Λύση άσκησης 4

1. Ευρετήριο συστάδων (clustered index) στο γνώρισμα Ειδικότητα

Τα φύλλα του ευρετηρίου στο γνώρισμα ειδικότητα καταλαμβάνουν $10000/4 = 2500$ σελίδες στον δίσκο. Επειδή η συνθήκη Ειδικότητα='Ηλεκτρολόγος' θα ανακτήσει το 10% των εγγραφών θα χρειαστεί να διαβάσουμε μόνο το 10% των φύλλων του ευρετηρίου. Στην συνέχεια θα πρέπει να διαβάσουμε το 10% των σελίδων της σχέσης ώστε να ανακτήσουμε τις τιμές για τα γνωρίσματα «Επώνυμο» και «Τμήμα».

$$\text{COST}(1) = 2500 * 10/100 + 10000 * 10/100 = 250 + 1000 = 1250 \text{ σελίδες.}$$

2. Ευρετήριο συστάδων (clustered index) στο ζεύγος των γνωρισμάτων (Ειδικότητα, Τμήμα) με την σειρά που δίνεται.

Τα φύλλα του ευρετηρίου στο ζεύγος των γνωρισμάτων (Ειδικότητα, Τμήμα) καταλαμβάνουν $10000/2 = 5000$ σελίδες στον δίσκο. Επειδή η συνθήκη 'Ηλεκτρολόγος' AND Τμήμα=Παραγωγή' θα ανακτήσει το 5% των εγγραφών θα χρειαστεί να διαβάσουμε μόνο το 5% των φύλλων του ευρετηρίου. Στην συνέχεια θα πρέπει να διαβάσουμε το 5% των σελίδων της σχέσης ώστε να ανακτήσουμε τις τιμές του γνωρίσματος «Επώνυμο».

$$\text{COST}(2) = 5000 * 5/100 + 10000 * 5/100 = 250 + 500 = 750 \text{ σελίδες.}$$

3. Ευρετήριο συστάδων (clustered index) στα ακόλουθα τρία γνωρίσματα (Τμήμα, Ειδικότητα, Επώνυμο) με την σειρά που αυτά δίνονται .

Τα φύλλα του ευρετηρίου στην τριάδα των γνωρισμάτων (Τμήμα, Ειδικότητα, Επώνυμο) καταλαμβάνουν $10000 * 3/4 = 7500$ σελίδες στον δίσκο. Επειδή η συνθήκη 'Ηλεκτρολόγος' AND Τμήμα='Παραγωγή' θα ανακτήσει το 5% των εγγραφών θα χρειαστεί να διαβάσουμε μόνο το 5% των φύλλων του ευρετηρίου. Το ευρετήριο περιέχει και τα τρία γνωρίσματα οπότε δεν χρειάζεται να διαβάσουμε την σχέση.

$$\text{COST}(3) = 7500 * 5/100 = 375 \text{ σελίδες.}$$

4. Ευρετήριο συστάδων (clustered index) στα ακόλουθα τρία γνωρίσματα (Επώνυμο, Ειδικότητα, Τμήμα) με τη σειρά που αυτά δίνονται.

Τα φύλλα του ευρετηρίου στην τριάδα των γνωρισμάτων (Επώνυμο, Ειδικότητα, Τμήμα) καταλαμβάνουν $10000 * 3/4 = 7500$ σελίδες στον δίσκο. Επειδή το πρώτο γνώρισμα του ευρετηρίου δεν είναι κάποιο από τα πεδία της συνθήκης θα χρειαστεί να διαβάσουμε ολόκληρο το ευρετήριο.

$$\text{COST}(4) = 7500 \text{ σελίδες.}$$