Κεφ. 13 - Εξωτερική ταξινόμηση – Διαφάνεια 5

Έστω ότι έχω:
Β=1000 //πλήθος πλαισίων στην ενδιάμεση μνήμη
Ν=500.000 σελίδες //πλήθος σελίδων του πίνακα

Τότε:

ΠΕΡΑΣΜΑ 0: Διαβάζω το αρχείο σε δόσεις των 1000 σελίδων και δημιουργώ ταξινομημένους συρμούς των 1000 σελίδων. Πόσους συρμούς; 500.000/1000=500

ΠΕΡΑΣΜΑ 1: Χρησιμοποιώ τα 999 πλαίσια ως πλαίσια εισόδου και το 1 πλαίσιο ως πλαίσιο εξόδου. Έχω (μόνο) 500 συρμούς, άρα τους συγχωνεύω όλους με ένα πέρασμα. Ως έξοδο παίρνω 1 συρμό των 500.000 σελίδων.

Παραλλαγή: Έστω ότι είχα 5.500.000 σελίδες στον πίνακα.

Τότε

ΠΕΡΑΣΜΑ 0: Θα πάρω 5.500.000/1000=5500 συρμούς των 1000 σελίδων

ΠΕΡΑΣΜΑ 1: Με 999 πλαίσια εισόδου, θα διαβάσω τους 5500 συρμούς σε 6 δόσεις (999+999+999+999+999+505). Σε κάθε δόση (εκτός της τελευταίας) θα παράξω ένα συρμό των 999*1000=999.000 σελίδων.

ΠΕΡΑΣΜΑ 2: Συγχωνεύω τους 6 συρμούς στον ένα τελικό συρμό.

Για να δούμε τον τύπο:

πλήθος_περασμάτων = $1+\lceil \log_{B-1}(\lceil N/B \rceil) \rceil = 1+\lceil \log_{999}(5500) \rceil = 1 + 2 = 3$

Κόστος: 2 * 5.500.000 * 3 = 33.000.000 I/O

Κεφ. 14Α - Σύξευξη

X >< Y

Πίνακας Χ:
RPP(X)=200 εγγραφές ανά σελίδα,
CARD(X)=500 σελίδες,
σύνολο 100000 εγγραφές

Πίνακας Β:
RPP(Y)=400 εγγραφές ανά σελίδα,
CARD(Y)=300 σελίδες,
σύνολο 120000 εγγραφές

[1] Σύζευξη με Απλή Εμφώλευση Βρόχων (Simple Nested Loops Join) – Διαφάνεια 5

Με Χ εξωτερικό:
Κόστος = CARD(X) + RPP(X)*CARD(X)*CARD(Y) =
500 + 200*500*300 = 500 + 30.000.000 = 30.000.500 I/O

Με Υ εξωτερικό:
Κόστος = CARD(Y) + RPP(Y)*CARD(Y)*CARD(X) =
300 + 400*300*500 = 300 + 60.000.000 = 60.000.300 I/O

[2] Σύζευξη με Εμφώλευση Βρόχων κατά Σελίδα (Page Nested Loops Join) Διαφάνεια 5

Με X εξωτερικό:
Κόστος = CARD(X) + CARD(X)*CARD(Y) = 500+500*300=500+150.000=150.500 I/O

Με Y εξωτερικό:
Κόστος = CARD(Y) + CARD(Y)*CARD(X) = 300+300*500=300+150.000=150.300 I/O

[3] Σύζευξη με Εμφ. Βρόχων και Ευρετήριο (Index Nested Loops Join) – Διαφάνεια 6

Υποθέτουμε μη-συγκροτημένα ευρετήρια Εναλλακτικής (2) ή (3)

Έστω ότι υπάρχει ευρετήριο στη στήλη σύζευξης στον Y.
Κόστος =
CARD(X) + RPP(X)*CARD(X)*κόστος_εύρεσης_ταιριαστών_εγγραφών_στον_Y=
500 + 200*500*(έστω 1+2) =
500 + 100.000*3 =
300.500 I/O

Έστω ότι υπάρχει ευρετήριο στη στήλη σύζευξης στον X.
Κόστος =
CARD(Y) + RPP(Y)*CARD(Y)*κόστος_εύρεσης_ταιριαστών_εγγραφών_στον_X=
300 + 400*300*(έστω 1+1) =
300 + 120.000*2 =
240.300 I/O

[4] Σύζευξη με Εμφ. Βρόχων κατά Μπλοκ (Block Nested Loops Join) – Διαφάνεια 8

Στα προηγούμενα αρκούσαν 3 πλαίσια μόνο (2 εισόδου και ένα εξόδου). Τώρα θα εκμεταλλευτώ όλα τα διαθέσιμα πλαίσια, έστω B=20.

Με τον X εξωτερικό
Χρησιμοποιώ B-2=18 πλαίσια για να διαβάζω τον X κατά μπλοκ. Για κάθε μπλοκ του X διαβάζω όλον τον Y στο ένα πλαίσιο και γράφω στο πλαίσιο εξόδου τις ταιριαστές εγγραφές.

Κόστος = CARD(X) + πλήθος_μπλοκ_X * CARD(Y) = 500 + 500/18 * 300 =
500 + 28 * 300 = 8900 I/O

Με τον Y εξωτερικό
Χρησιμοποιώ B-2=18 πλαίσια για να διαβάζω τον Y κατά μπλοκ. Για κάθε μπλοκ του Y διαβάζω όλον τον X στο ένα πλαίσιο και γράφω στο πλαίσιο εξόδου τις ταιριαστές εγγραφές.

Κόστος = CARD(Y) + πλήθος_μπλοκ_Y * CARD(X) = 300 + 300/18 * 500 =
300 + 17 * 500 = 8800 I/O

[5] Σύζευξη με Ταξινόμηση και Συγχώνευση (Sort-Merge Join) – Διαφάνεια 12

Κόστος = Κόστος_Ταξ(X) + Κόστος_Ταξ(Y) + CARD(X) + CARD(Y)

Για B=20
Περάσματα X = $1+\lceil \log_{B-1}(\lceil N/B \rceil) \rceil = 1+\lceil \log_{19}(\lceil 500/20 \rceil) \rceil = 1+\lceil \log_{19}(25) \rceil = 3$
Κόστος_Ταξ(X) = 2*500*3=3000 I/O

Περάσματα Y = $1+\lceil \log_{B-1}(\lceil N/B \rceil) \rceil = 1+\lceil \log_{19}(\lceil 300/20 \rceil) \rceil = 1+\lceil \log_{19}(15) \rceil = 2$
Κόστος_Ταξ(Y) = 2*300*2=1200 I/O

Κόστος = 3000 + 1200 + 500 + 300 = 5000 I/O

[6] Σύζευξη με Κατακερματισμό (Hash Join) – Διαφάνεια 13

Διαμερισμός X
Διαβάζω X και γράφω B-1=19 διαμερίσματα των 500/19 = 27 σελίδων

Διαμερισμός Y
Διαβάζω Y και γράφω B-1=19 διαμερίσματα των 300/19 = 16 σελίδων

Δυστυχώς έχω sqrt(CARD(X)) = 23 > B, άρα θα χρησιμοποιήσω ως εξωτερικό τον Y (sqrt(CARD(Y) = 18 < B).

Κόστος = Κόστος_διαμερισμού_X + Κόστος_διαμερισμού_Y + CARD(Y) + SUM(CARD(Xi))
= 2*500 + 2*300 + 300 + 500 = 2400 I/O

Αν είχα B=23, τότε με το ίδιο κόστος θα μπορούσα να έχω τον X εξωτερικό.

[5] Σύζευξη με Ταξινόμηση και Συγχώνευση (Sort-Merge Join) – Διαφάνεια 12

Αν είχα B > sqrt(CARD(μεγαλύτερου_πίνακα)), δηλαδή B = 23, και χρησιμοποιούσα τη βελτιστοποίηση που παράγει συρμούς 2B στο Πέρασμα 0, τότε

X: Το Πέρασμα 0 παράγει 11 συρμούς των 46 σελίδων
Y: Το Πέρασμα 0 παράγει 7 συρμούς των 46 σελίδων

Πέρασμα 1 ταξινόμησης γίνεται μαζί με τη φάση Σύζευξης με Συγχώνευση:
Χρησιμοποιώ 11 πλαίσια για X, 7 πλαίσια για Y και 1 πλαίσιο για έξοδο. Διαβάζω X και Y μια φορά.

Κόστος = 2*CARD(X) + 2*CARD(Y) + CARD(X) + CARD(Y) = 2400 I/O