2η Σειρά Ασκήσεων Σχεδιασμός Βάσεων Δεδομένων:

Γεωργίου Αλέξιος-Λάζαρος 3180027

Άσκηση 1:

$$T(R) = 1000000, B(R) = 20000, V(R,a) = n$$

1.1.1) Clustered index για select “a = 2”:

$$Cost = \frac{B(R)}{V(R,a)} = \frac{20000}{n}$$

1.1.2) Clustered index για select “k<=a<=l”:

$$Cost = \frac{B(R)}{\frac{n}{10}} = \frac{200000}{n}, έχουμε \frac{n}{10}\ διακρίτες\ τιμές$$

1.2.1) Non-clustered index για select “a = 2”:

$$Cost = \frac{T(R)}{V(R,a)} = \frac{1000000}{n}$$

1.2.2) Non-clustered index για select “k<=a<=l”:

$$Cost = \frac{T(R)}{\frac{n}{10}} = \frac{10000000}{n}$$

2) Αν δεν χρησιμοποιήσουμε κάποιο ευρετήριο τότε και για τα δύο queries το κόστος είναι το κόστος όλων των μπλοκ (table scan).

$$Cost = B(R) = 20000$$

Άρα για να μας συμφέρει το ευρετήριο (non clustered) για το πρώτο query:

$$\frac{1000000}{n} \leq 20000 => n \geq 50$$

Άρα για να μας συμφέρει το ευρετήριο (non clustered) για το δεύτερο query:

$$\frac{10000000}{n} \leq 20000 => n \geq 500$$

Άσκηση 2:

$$T(R) = 0 + 80 + 100 + 20 + 30 = 230, T(S) = 10 + 100 + 60 + 60 + 0 = 230$$

a) Με το ιστόγραμμα έχω:

$$V_i(R, b) = V_i(S, b) = 20 \; για \; κάθε \; i \; διάστημα$$

Για κάθε διάστημα έχω περίπου τόσες πλειάδες όσο το γινόμενο των εγγραφών δια τον αριθμό των πιθανών τιμών του διαστήματος.

$$\sum_{i=1}^{5} \frac{T_i(R) * S_i(R)}{20} = 0 + \frac{8000}{20} + \frac{6000}{20} + \frac{1200}{20} + 0 = 400 + 300 + 60 = 760 \; εγγραφές$$

b) Αν δεν έχω ιστόγραμμα:

$$V(R, b) = V(S, b) = 100 \; διακριτές \; τιμές, ακέραιοι \; στο \; [1{,}100]$$

Στατιστικά έχω περίπου

$$T_i(R) = S_i(R) = \frac{230}{100} = 2.3 \; εγγραφές \; σε \; κάθε \; τιμή \; σύμφωνα \; με \; την \; ομοιόμορφη \; κατανομή$$

Άρα θα έχουμε περίπου

$$\sum_{i=1}^{100} \frac{T_i(R) * S_i(R)}{1} = 100 * 2.3^2 = 529 \; εγγραφές$$

Άσκηση 3:

$$R(a,b,\boldsymbol{c}), S(\boldsymbol{c},d,e)\ , T(R)=20000, T(S)=45000$$

$$B(R)=\frac{T(R)}{25}=\frac{20000}{25}=800\ \sigma\varepsilon\lambda\acute{\iota}\delta\varepsilon\varsigma$$

$$B(S)=\frac{T(S)}{30}=\frac{45000}{30}=1500\ \sigma\varepsilon\lambda\acute{\iota}\delta\varepsilon\varsigma$$

$$M=41$$

1) I/O της σύζευξης R⋈S

a)NLJ

Θα διαβάσουμε το R μια φορά άρα κόστος $B(R)=800$

$$Αριθμός\ block\ =\frac{B(S)}{M-1}=\frac{1500}{40}=37.5$$

Για κάθε block πρέπει να διαβάσουμε όλο το S άρα:

$$Κόστος=800+37.5*B(R)=800+37.5*800=30800$$

b)SMJ

$$Cost=5*B(S)+B(R)=11500$$

Ισχύουν οι απαιτήσεις μνήμης $41\geq\sqrt{\max\big(B(R),B(S)\big)}\cong 38.72$

c)Hash join

$$Cost=3*B(S)+B(R)=6900$$

2) Χρησιμοποιώντας τον βελτιστοποιημένο SMJ, έχουμε

$$Κόστος=3*B(S)+B(R)=6900$$

Για να το πετύχουμε αυτό, πρέπει να αυξήσουμε την μνήμη

Σε $M\geq\sqrt{B(R)+B(S)}\cong 47.95$, δηλαδή σε τουλάχιστον $M=48$

Άσκηση 4:

1)

Έχουμε 50000 εκδότες στην βάση και 500 διαφορετικούς εκδότες άρα με βάση την ομοιόμορφη κατανομή θα έχουμε περίπου 100 εγγραφές βιβλίων με εκδότη τον Σαββάλα $T(\sigma_{Σαββάλας}) = 100$.

Υπάρχει απλό ευρετήριο (non-clustered) για την ιδιότητα του εκδότη στην σχέση ΒΙΒΛΙΑ.

Άρα το κόστος σε I/O είναι:

$$Cost = \frac{T(\text{ΒΙΒΛΙΑ})}{V(\text{ΒΙΒΛΙΑ}, Εκδότης)} = \frac{50000}{500} = 100$$

2) $\sigma_{Σαββάλας}\bowtie$ ΔΑΝΕΙΣΜΟΙ

Συνδέουμε τον πίνακα ΔΑΝΕΙΣΜΟΙ με το αποτέλεσμα του SELECT από το προηγούμενο ερώτημα με βάση την ιδιότητα ΚΒ.

Έχουμε 300000 εγγραφές στην σχέση ΔΑΝΕΙΣΜΟΙ όποτε υποθέτουμε (ομοιόμορφη κατανομή) ότι οι

$\frac{300000}{50000} = 6\ εγγραφές\ δανεισμού\ για\ κάθε\ βιβλίο$

Οπότε θα έχουμε $6 * 100\ βιβλία = 600\ εγγραφές\ δανεισμού$

Υπάρχει ένα ευρετήριο συστάδων (clustered index) στο γνώρισμα ΚΒ της σχέσης ΔΑΝΕΙΣΜΟΙ

$$B(\sigma_{Σαββάλας}) = \frac{5000 * 100}{50000} = 10$$

$$X = 6$$

$$Κόστος = B(\sigma_{Σαββάλας}) + T(\sigma_{Σαββάλας}) * \frac{X}{\frac{300000}{15000}} = 10 + 100 * \frac{6}{\frac{300}{15}} = 40$$

3)$ΔΑΝΕΙΣΜΟΙΣΑΒΒΑΛΑ\bowtie$ΔΑΝΕΙΖΟΜΕΝΟΙ

Έχουμε 600 εγγραφές από δανεισμούς σε βιβλία του Σαββάλα και με την σύνδεση του πίνακα στο γνώρισμα ΚΔ δεν θα αλλάξει ο αριθμός των εγγραφών, αφού απλά συνδέουμε την πληροφορία του δανειζόμενου στον δανεισμό και γνωρίζουμε ότι ένας δανεισμός έχει μόνο έναν δανειζόμενο.

NLJ cost:

$$B(ΔΑΝΕΙΣΜΟΙΣΑΒΒΑΛΑ) = B(\sigma_{Σαββάλας}) + B(\text{ΔΑΝΕΙΣΜΟΙ}) = 10 + \frac{300000}{15000} = 30$$

$$Αριθμός\ block\ = \frac{B(\text{ΔΑΝΕΙΖΟΜΕΝΟΙ})}{M-1} = \frac{1000}{19} \cong 52.63$$

$$Κόστος = B(ΔΑΝΕΙΣΜΟΙΣΑΒΒΑΛΑ) + 52.63 * B(ΔΑΝΕΙΣΜΟΙΣΑΒΒΑΛΑ) = 30 + 52.63 * 30 \cong 1608$$

4)

Έχουμε 600 εγγραφές και πρέπει να διαλέξουμε τις εγγραφές εκείνες στις οποίες ο δανειζόμενος έχει ηλικία [13,19] (μεγαλύτερες του 12 και μικρότερες του 20).

Γνωρίζουμε ότι οι ηλικίες είναι οι φυσικοί αριθμοί στο [7,24], δηλαδή έχουμε 24-7 + 1 = 18 διακριτές τιμές σύνολο και μας ενδιαφέρουν οι 19-13 + 1 = 7 διακριτές τιμές. Άρα σύμφωνα με την ομοιόμορφη κατανομή περιμένουμε να έχουμε περίπου $600 * \frac{7}{18} \cong 233.33\ εγγραφές$

Δεν υπάρχει κάποιο ευρετήριο για την ηλικία, άρα το κόστος σε I/O είναι:

$$Cost = B(ΔΑΝΕΙΖΟΜΕΝΟΙ ΔΑΝΕΙΣΜΟΙΣΑΒΒΑΛΑ)$$

$$= B(ΔΑΝΕΙΣΜΟΙΣΑΒΒΑΛΑ) + B(ΔΑΝΕΙΖΟΜΕΝΟΙ) = 30 + \frac{10000}{1000} = 40$$

Άσκηση 5:

1)