ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΛΥΣΗ ΣΤΗΝ ΔΕΥΤΕΡΗ ΑΣΚΗΣΗ

ΜΑΘΗΜΑ	**ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ**
ΑΚΑΔ. ΕΤΟΣ	2014-15
ΔΙΔΑΣΚΩΝ	Ιωάννης Βασιλείου, *Καθηγητής*, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Ερώτημα 1.

Θεωρήστε τη σχέση R (A, B, C, D) που περιέχει 1.000.000 εγγραφές, και κάθε σελίδα της σχέσης χωρά 10 εγγραφές. Η R είναι οργανωμένη σε αρχείο σωρού με πυκνά δευτερεύοντα ευρετήρια, και οι εγγραφές της είναι τυχαία ταξινομημένες. Υποθέστε πως το γνώρισμα A είναι υποψήφιο κλειδί της R, με διάστημα τιμών από 0 έως 999.999. Για κάθε ένα από τα παρακάτω αιτήματα, προσδιορίστε το πλήθος των I/O που απαιτούνται για την επεξεργασία του ερωτήματος. Ακολουθούν οι τεχνικές που πρέπει να θεωρήσετε:

- Σάρωση του αρχείου σωρού της R.
- Χρήση ενός ευρετηρίου B+ -δένδρου στο γνώρισμα R.A.
- Χρήση ενός ευρετηρίου κατακερματισμού στο γνώρισμα R.A.

Τα αιτήματα είναι:

1. Βρείτε όλες τις πλειάδες της R ώστε A < 50.
2. Βρείτε όλες τις πλειάδες της R ώστε A = 50.
3. Βρείτε όλες τις πλειάδες της R. ώστε A > 50 και A < 100.
4. Βρείτε όλες τις πλειάδες της R ώστε A ≠ 50.

ΛΥΣΗ

- Σάρωση του αρχείου σωρού R.

Οι εγγραφές είναι τυχαία ταξινομημένες, οπότε στη χειρότερη περίπτωση για να εντοπιστούν οι ζητούμενες πλειάδες πρέπει να προσπελάσουμε όλες τις σελίδες που έχουν εγγραφές της R. Αυτό ισχύει και για τα τέσσερα ερωτήματα. Οπότε, απαιτούνται $10^6/10 = 10^5$ I/O.

- Χρήση ενός ευρετηρίου B+δέντρου στο γνώρισμα R.A.

Έστω h το ύψος του B+δέντρου και M το πλήθος των εγγραφών ευρετηρίου που χωρούν σε μία σελίδα (block) του ευρετηρίου.

1. Αρχικά θα διαβαστούν τα block που αντιστοιχούν στο μονοπάτι από τη ρίζα του δέντρου έως και το πρώτο φύλλο. Άρα, απαιτούνται h I/O. Επίσης, θα πρέπει να διαβαστούν τα επόμενα στη σειρά φύλλα του δέντρου, μέχρι εκείνο που περιέχει το κλειδί με τιμή 50. Συνολικά, θα διαβαστούν $\lceil 51/\mathrm{M} \rceil$ φύλλα (μαζί με το πρώτο φύλλο), στα οποία αντιστοιχούν $\lceil 51/\mathrm{M} \rceil$ I/O. Τέλος, απαιτούνται επιπλέον 50 I/O σελίδες του αρχείου της R για να ανακτηθούν οι ζητούμενες εγγραφές, οπότε απαιτούνται επιπλέον 50 I/O. Συνολικά, απαιτούνται: $h + [\lceil 51/M \rceil - 1] + 50$ I/O.

Να σημειωθεί ότι η διόρθωση -1 απαιτείται για να μη ληφθεί υπόψη το πρώτο φύλλο δύο φορές.

2. Θα διαβαστούν τα block που αντιστοιχούν στο μονοπάτι από τη ρίζα του δέντρου έως και το φύλλο που περιέχει το κλειδί με τιμή 50 και κατόπιν η σελίδα του αρχείου με την κατάλληλη εγγραφή. Συνολικά, απαιτούνται: $h + 1$ I/O.

3. Θα διαβαστούν τα block που αντιστοιχούν στο μονοπάτι από τη ρίζα του δέντρου έως και το φύλλο που περιέχει το κλειδί με τιμή 50. Άρα, απαιτούνται h I/O. Επιπλέον, θα διαβαστούν όλα τα επόμενα στη σειρά φύλλα, μέχρι εκείνο που περιέχει το κλειδί με τιμή 100. Για την ανάγνωση των φύλλων απαιτούνται $\lceil 51/\mathrm{M} \rceil$ I/O (συμπεριλαμβανομένου του πρώτου φύλλου που διαβάστηκε). Τέλος, απαιτούνται επιπλέον 48 I/O για να διαβαστούν οι σελίδες του αρχείου της R και να ανακτηθούν οι ζητούμενες εγγραφές. Συνολικά, απαιτούνται: $h + [\lceil 51/\mathrm{M} \rceil - 1] + 48$ I/O.
4. Αρχικά θα διαβαστούν τα block που αντιστοιχούν στο μονοπάτι από τη ρίζα του δέντρου έως και το πρώτο φύλλο. Απαιτούνται h I/O. Θα διαβαστούν όλα τα φύλλα του δέντρου για να γίνει ο έλεγχος της ανισότητας. Απαιτούνται $\lceil 10^6/\mathrm{M} \rceil$ I/O, συμπεριλαμβανομένου του πρώτου φύλλου. Τέλος, απαιτούνται επιπλέον $10^6 - 1$ I/O για να διαβαστούν οι εγγραφές της R. Συνολικά, απαιτούνται: $h + [\lceil 10^6/\mathrm{M} \rceil - 1] + 10^6 - 1$ I/O.

- Χρήση ενός ευρετηρίου κατακερματισμού στο γνώρισμα R.A.

Έστω M το πλήθος των εγγραφών ευρετηρίου που χωρούν σε μία σελίδα (block) του ευρετηρίου. Θεωρούμε ότι οι τιμές κλειδιού κατανέμονται ομοιόμορφα σε όλους τους κάδους ώστε να μην υπάρχουν υπερχειλίσεις και να ισχύει ότι σε κάθε κάδο αντιστοιχεί ένα block.
1. Δεδομένου ότι ότι η συνάρτηση κατακερματισμού διασκορπίζει τις τιμές κλειδιού τυχαία, οι τιμές στο ζητούμενο εύρος είναι πιθανό να είναι διάσπαρτες σε πολλούς ή και όλους τους κάδους. Συνεπώς, στη χειρότερη περίπτωση, θα πρέπει να διαβαστούν όλα τα block του ευρετηρίου και οι σελίδες του αρχείου που περιέχουν τις ζητούμενες εγγραφές. Συνεπώς, απαιτούνται: $\lceil 10^6/M \rceil + 50$ I/O.
2. Θα διαβαστεί το block του ευρετηρίου που περιέχει την τιμή κλειδιού 50 και κατόπιν η κατάλληλη σελίδα του αρχείου. Συνεπώς, απαιτούνται: 2 I/O.
3. Αντίστοιχα με το ερώτημα 1, θα διαβαστούν όλα τα block του ευρετηρίου και οι σελίδες του αρχείου με τις ζητούμενες εγγραφές. Συνεπώς, απαιτούνται: $\lceil 10^6/M \rceil + 48$ I/O.
4. Θα διαβαστούν όλα τα block του ευρετηρίου και οι σελίδες του αρχείου με τις ζητούμενες εγγραφές. Συνεπώς, απαιτούνται $\lceil 10^6/M \rceil + 10^6 - 1$ I/O.

Ερώτημα 2.
Θεωρήστε μία σχέση R(A,B) οργανωμένη ως κατακερματισμένο αρχείο στο δίσκο. Το κατακερματισμένο αρχείο οργανώνεται σε 1024 blocks (κάδους - buckets) που περιέχουν τις εγγραφές της σχέσης. Για την αποθήκευση μίας εγγραφής (a,b) εφαρμόζουμε πρώτα τη συνάρτηση κατακερματισμού h1 στο πεδίο a λαμβάνοντας X bits. Στη συνέχεια εφαρμόζουμε τη συνάρτηση h2 στο πεδίο b λαμβάνοντας 10-X bits. Τα 10 bits συνολικά ορίζουν τη διεύθυνση του block στο οποίο θα αποθηκευτεί η εγγραφή (a,b).
Θεωρήστε ότι το 20% των ερωτημάτων που αφορούν στη σχέση R είναι της μορφής Q1: SELECT * FROM R WHERE A = a, ενώ το υπόλοιπο 80% είναι της μορφής: Q2: SELECT * FROM R WHERE B = b, όπου a και b είναι σταθερές που δίνονται από τους συντάκτες των ερωτημάτων.
1. Πόσα blocks πρέπει να προσπελαστούν για να απαντηθούν τα ερωτήματα τύπου Q1 και πόσα για τα τύπου Q2; Η απάντησή σας θα είναι συνάρτηση των X bits.
2. Δώστε ένα τύπο που να εκτιμά το μέσο αριθμό blocks που χρειάζεται να προσπελαστούν για την απάντηση ερωτημάτων στη σχέση R.

ΛΥΣΗ

1. Για τα ερωτήματα τύπου Q1 πρέπει να λάβουμε υπόψη μας τα 10-X bits από την εφαρμογή της h2 hash συνάρτησης. Άρα θα προσπελαστούν: $2^{(10-X)}$ blocks.
Αντιθέτως, για τα ερωτήματα τύπου Q2 λαμβάνουμε υπόψη μας τα X bits από τη h1 hash συνάρτηση. Άρα θα προσπελαστούν: 2^X blocks.
2. Με δεδομένο ότι το 20% των ερωτημάτων είναι τύπου Q1 και το 80% τύπου Q2, ο μέσος αριθμός blocks που χρειάζεται να προσπελαστούν για την απάντηση ερωτημάτων στην R δίνεται από τον τύπο: $0.2 \times 2^{(10-X)} + 0.8 \times 2^X$.

Ερώτημα 3.

1. Χρησιμοποιείστε τα αξιώματα του Armstrong προκειμένου να αποδείξετε την εγκυρότητα του κανόνα αποσύνθεσης.

$$if\ \alpha \to \beta\gamma, then\ \alpha \to \beta\ and\ \alpha \to \gamma$$

2. Εξετάστε τον παρακάτω προτεινόμενο κανόνα για λειτουργικές εξαρτήσεις: Αν $\alpha \to \beta$ και $\gamma \to \beta$, τότε $\alpha \to \gamma$. Αποδείξτε ότι αυτός ο κανόνας δεν είναι έγκυρος.
Υπόδειξη: Δώστε μια σχέση r που ικανοποιεί το $\alpha \to \beta$ και το $\gamma \to \beta$, αλλά δεν ικανοποιεί το $\alpha \to \gamma$.

ΛΥΣΗ

1. $\alpha \to \beta\gamma$
   $\beta\gamma \to \beta$ (ανακλαστικός κανόνας)
   $\alpha \to \beta$ (μεταβατικός κανόνας)
   $\beta\gamma \to \gamma$ (ανακλαστικός κανόνας)
   $\alpha \to \gamma$ (μεταβατικός κανόνας)

2. Η παρακάτω σχέση r αποτελεί ένα αντιπαράδειγμα για τον κανόνα.

$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
$\boldsymbol{a_1}$	b_1	c_1
$\boldsymbol{a_1}$	b_1	c_2

Για την r ισχύει η $\alpha \to \beta$ και η $\gamma \to \beta$. Δεδομένου ότι οι δύο tuples δεν έχουν την ίδια τιμή στο γ, η $\gamma \to \beta$ και πάλι ισχύει. Ωστόσο, η $\alpha \to \gamma$ δεν ισχύει, καθώς αν και οι δύο tuples έχουν την ίδια τιμή στο α έχουν διαφορετικές τιμές στο γ.

Ερώτημα 4.

Δίνεται το σχήμα R(A, B, C, D, E). Δείξτε ότι η παρακάτω αποσύνθεση δεν είναι μία αποσύνθεση χωρίς απώλεια (lossless-join decomposition):

R1 (A, B, C)

R2 (C, D, E).

Υπόδειξη: Δώστε ένα παράδειγμα ενός στιγμιοτύπου του R όπου δεν ισχύει η ικανή και αναγκαία συνθήκη για lossless-join decomposition.

ΛΥΣΗ

Έστω r ένα στιγμιότυπο της R. Η αποσύνθεση της R στις R1 και R2 είναι χωρίς απώλεια αν για κάθε r ισχύει: $\Pi_{R1}(\mathrm{r}) \bowtie \Pi_{R2}(\mathrm{r}) = \mathrm{r}$.

Το ακόλουθο στιγμιότυπο r αποτελεί ένα αντιπαράδειγμα:

$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{C}$	$\boldsymbol{D}$	$\boldsymbol{E}$
$\boldsymbol{a_1}$	b_1	c_1	d_1	e_1
$\boldsymbol{a_2}$	b_2	c_1	d_2	e_2

Το $\Pi_{R1}(\mathrm{r})$ είναι:

A	**B**	**C**
$\boldsymbol{a_1}$	b_1	c_1
$\boldsymbol{a_2}$	b_2	c_1

Το $\Pi_{R2}(\mathrm{r})$ είναι:

C	**D**	**E**
$\boldsymbol{c_1}$	d_1	e_1
$\boldsymbol{c_1}$	d_2	e_2

Υπολογίζουμε το $\Pi_{R1}(\mathrm{r}) \bowtie \Pi_{R2}(\mathrm{r})$:

A	**B**	**C**	**D**	**E**
$\boldsymbol{a_1}$	b_1	c_1	d_1	e_1
$\boldsymbol{a_1}$	b_1	c_1	d_2	e_2
$\boldsymbol{a_2}$	b_2	c_1	d_1	e_1
$\boldsymbol{a_2}$	b_2	c_1	d_2	e_2

Παρατηρούμε ότι $\Pi_{R1}(\mathrm{r}) \bowtie \Pi_{R2}(\mathrm{r}) \neq \mathrm{r}$. Συνεπώς, είναι μια αποσύνθεση με απώλεια (lossy join).