



Συστήματα Διαχείρισης Βάσεων Δεδομένων

Σειρά Ασκήσεων 1 2023-2024

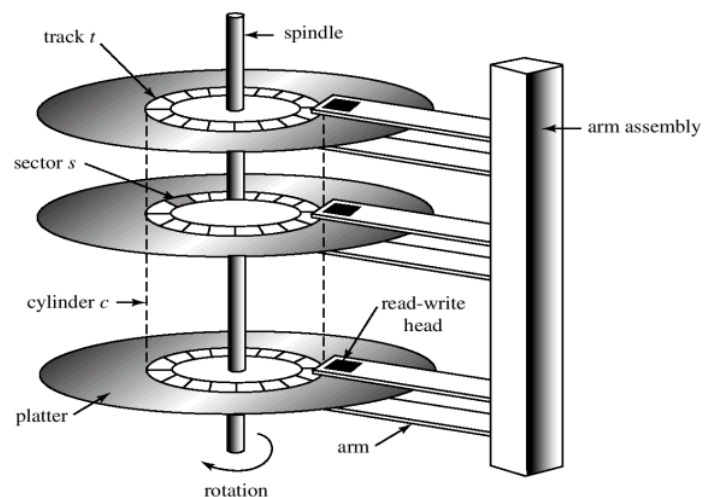
Ονοματεπώνυμο:

Κεχριώτη Ελένη

Άσκηση 1

Έχουμε τα εξής από την εκφώνηση:

- Capacity = 80GB
- 4 platters διπλής όψης
- 512 tracks/surface
- 1024 sectors/track
- rpm 7200
- χρόνος μετακίνησης κεφαλής= 8 ms (milliseconds)



1. Ο σκληρός δίσκος έχει 4 πλακέτες διπλής όψης άρα έχει 2 επιφάνειες ανά πλακέτα. Ο δίσκος έχει χωρητικότητα 80 GB, 80.000.000 KB

$KB/δίσκο = KB/επιφάνεια * επιφάνειες/πλακέτα * αριθμός πλακέτων$
 $80.000.000 KB = 85.899.345.920 B = KB/επιφάνεια * 2 * 4$
 $KB/επιφάνεια = 10.000.000 KB$

$KB/επιφάνεια = KB/ίχνος * ίχνοι/επιφάνεια$
 $10.000.000 KB = 10.737.418.240 B = KB/ίχνος * 512$
 $KB/ίχνος = 20.000 KB$

$KB/ίχνος = KB/τομέα * τομείς/ίχνος$
 $20.000 KB = 20.971.520 B = KB/τομέα * 1024$
 $KB/τομέα = 20 KB = 20.480 B$

2. Κάθε εγγραφή έχει μέγεθος 40 KB ~ $40 * 1024 = 40.960$ bytes. Άρα, εφόσον μια εγγραφή αποθηκεύεται ολόκληρη σε ένα block, κάθε block έχει μέγεθος 40KB.

Ένα ίχνος διαθέτει 1.024 τομείς των 20 KB. Κάθε μπλοκ είναι 40 KB δηλαδή αποθηκεύεται σε δύο τομείς, οπότε σε ένα ίχνος χωράνε 512 μπλοκ. Ένας κύλινδρος έχει 8 ίχνη ($4 επιφάνειες * 2 ίχνη$).

Ακόμη, ένας κύλινδρος έχει $8 ίχνη * 512 μπλοκ = 4.096$ μπλοκ.

Έχουμε 8.192 εγγραφές και μια εγγραφή αποθηκεύεται ολόκληρη σε ένα μπλοκ άρα για να αποθηκευτούν όλες οι εγγραφές θέλουμε 8.192 μπλοκ.

Οπότε για την αποθήκευση της R απαιτούνται $\frac{8.192}{4.096} = 2$ κύλινδροι.

3. Ισχύει ότι:

Συνολικός Χρόνος για το Διάβασμα της R = Μέσος Χρόνος Μετακίνησης Κεφαλής + Μέση Καθυστέρηση Περιστροφής + Χρόνος Ανάγνωσης των Μπλοκ της R.

Από την εκφώνηση έχουμε ότι ο **Μέσος Χρόνος Μετακίνησης Κεφαλής** = 8 msec = 0,008 sec

Μέση Καθυστέρηση Περιστροφής = $(60 \text{ sec} / 7200 \text{ rpm}) / 2 = 0,00415 \text{ sec}$

Για να βρούμε τον **Χρόνος Ανάγνωσης των Μπλοκ της R** υπολογίζουμε:

- Για την αποθήκευση της σχέσης R βρήκαμε προηγουμένως ότι απαιτούνται 2 κύλινδροι.
- Για το διάβασμα ενός ίχνους χρειάζονται $60 \text{ sec} / 7200 = 0,0083 \text{ sec}$.
- Για το διάβασμα ενός κυλίνδρου απαιτούνται $0,0083 * 8 = 0,0664 \text{ sec}$.
- Για το διάβασμα των 2 κυλίνδρων απαιτούνται $2 * 0,0664 = 0,1328 \text{ sec}$.

Συνολικός Χρόνος για το Διάβασμα του Αρχείου = $0,008 \text{ sec} + 0,00415 + 0,1328 \text{ sec} = 0.14495 \text{ sec}$

4. Ισχύει ότι:

Χρόνος Ανάγνωσης Ενός Μπλοκ = Μέσος Χρόνος Μετακίνησης Κεφαλής + Μέση Καθυστέρησης Περιστροφής + Χρόνος Μεταφοράς του Μπλοκ

Σε $0,0083 \text{ sec}$ διαβάζουμε ένα ίχνος το οποίο περιέχει 512 μπλοκ, δηλαδή διαβάζουμε ένα μπλοκ σε $0,0083/512=0.00001621 \text{ sec}$.

Άρα, ο χρόνος που απαιτείται για να διαβαστούν 100 τυχαία μπλοκ είναι ίσος με $(0,008+0,00415+0.00001621)*100 = 1.216621 \text{ sec}$.

Άσκηση 2

Κάθε εγγραφή αποτελείται από $9 + 30 + 10 + 9 + 8 + 1 + 40 + 4 + 4 = 115 \text{ bytes}$.
Το μέγεθος της εγγραφής του ευρετηρίου <ΑΦΜ, δείκτης> είναι $9 + 6 = 15 \text{ bytes}$.

Σε ένα μπλοκ μπορούν να αποθηκευτούν $\lfloor 1024 / 115 \rfloor = 8 \text{ εγγραφές}$.

Για να αποθηκεύσουμε 60.000 εγγραφές χρειαζόμαστε:

$\lceil 60.000 / 8 \rceil = 7.500 \text{ μπλοκ}$.

Σε ένα μπλοκ μπορούμε να αποθηκεύσουμε $\lfloor 1024 / 15 \rfloor = 68 \text{ εγγραφές της μορφής <ΑΦΜ, δείκτης>}$.

1. Στο δίσκο θα αποθηκευτούν σειριακά οι εγγραφές της σχέσης ΦΟΡΟΛΟΓΟΥΜΕΝΟΣ σε 7.500 μπλοκ συνολικά. Το ευρετήριο θα γίνει πάνω στη σχέση ΦΟΡΟΛΟΓΟΥΜΕΝΟΣ με κλειδί το ΑΦΜ, που είναι και μοναδικό. Εφόσον το ευρετήριο είναι αραιό θα αποθηκεύεται μια εγγραφή <ΑΦΜ, δείκτης> για κάθε μπλοκ με εγγραφές.

Επομένως στο πρώτο επίπεδο του ευρετηρίου θα καταχωρηθούν 6.667 εγγραφές.

2. Για 7.500 εγγραφές της μορφής <ΑΦΜ, δείκτης> θα χρειαστούμε για την αποθήκευση τους $\lceil 7.500/68 \rceil = 111$ μπλοκ.

3. Θέλουμε το τελευταίο επίπεδο του ευρετηρίου να χωράει σε 1 μπλοκ. Επομένως, προσθέτουμε 1 ακόμη επίπεδο ευρετηρίου πάνω στο προηγούμενο ευρετήριο που αναλύσαμε με εγγραφές ξανά της μορφής <ΑΦΜ, δείκτης>.

Για το δεύτερο επίπεδο του ευρετηρίου θα χρειαστούμε $\lceil 111/68 \rceil = 2$ μπλοκ.

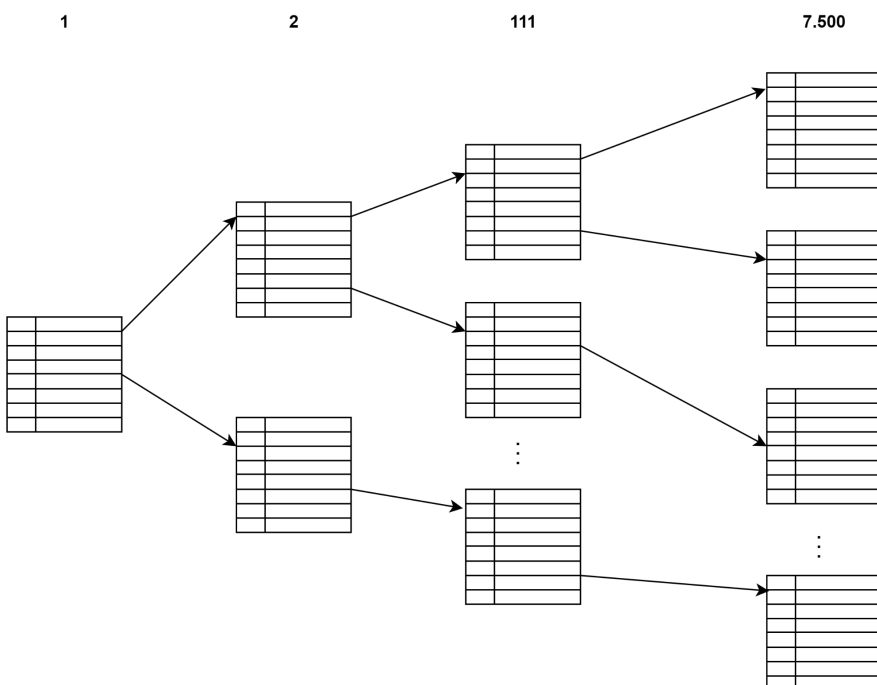
Με την προσθήκη ενός ακόμη επιπέδου ευρετηρίου θα χρειαστούμε για την αποθήκευση των 2 εγγραφών 1 μπλοκ.

Άρα, τελικά το ευρετήριο αποτελείται από 3 επίπεδα.

4. Συνολικά απαιτούνται $111 + 2 + 1 = 114$ μπλοκ για το πολυεπίπεδο ευρετήριο.

5. Κατά την αναζήτηση στο ευρετήριο ανακτάται ένα μπλοκ δίσκου σε κάθε επίπεδο. Επομένως προσπελάζονται 3 μπλοκ για μια αναζήτηση ευρετηρίου. Για την ανάκτηση της εγγραφής θα χρειαστεί να προσπελάσουμε ακόμη 1 μπλοκ. Άρα, συνολικά θα προσπελαστούν 4 μπλοκ για την αναζήτηση και την ανάκτηση της εγγραφής.

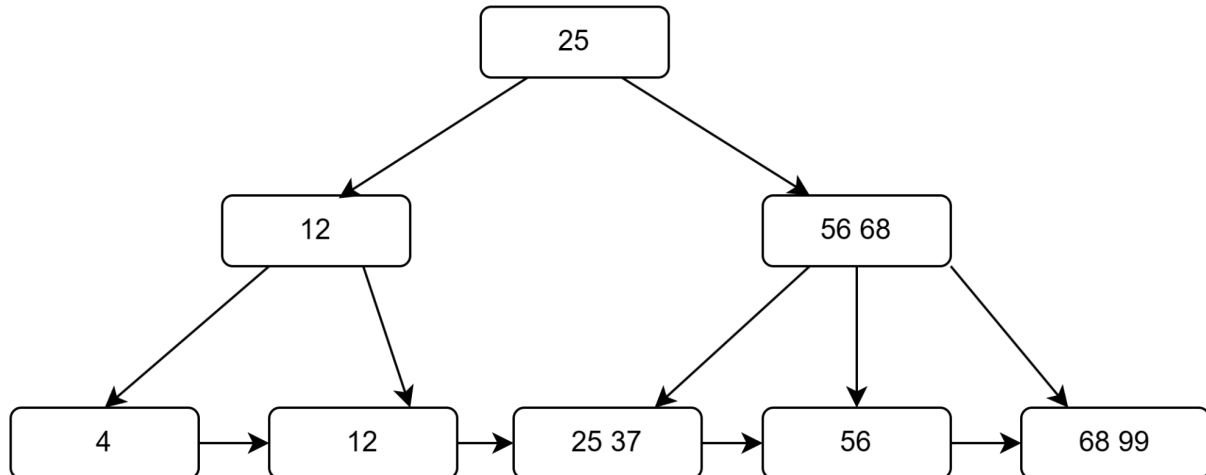
Παρακάτω φαίνεται ένα υποτυπώδες σχήμα για το πολυεπίπεδο ευρετήριο.



Άσκηση 3

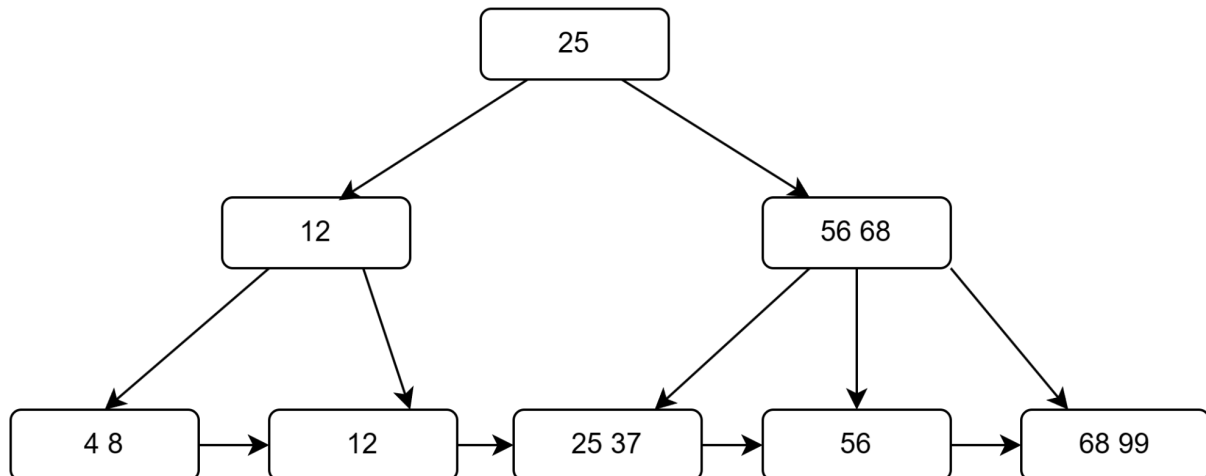
1.

Αρχικό δέντρο



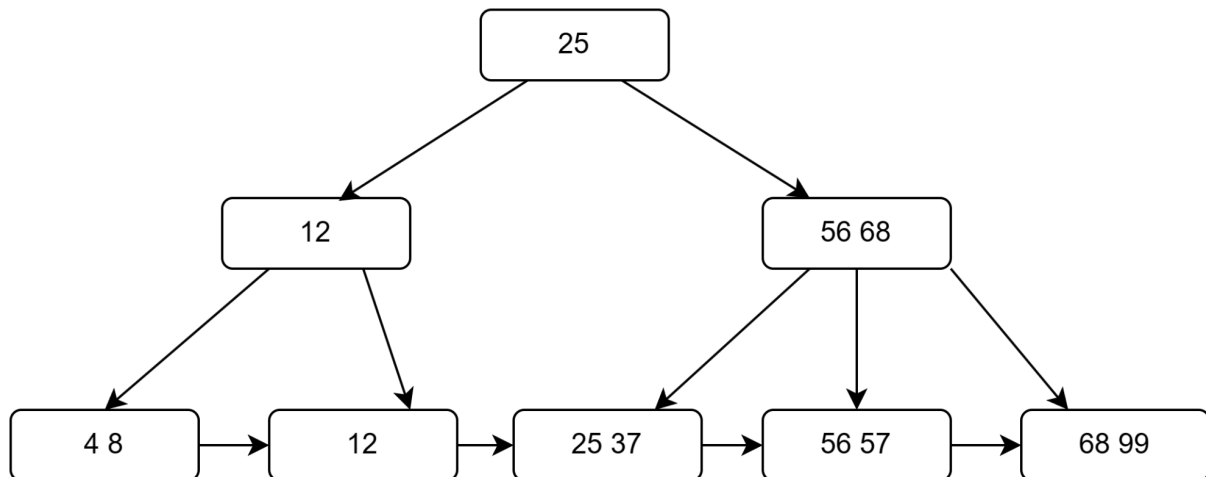
Εισαγωγή του 8

- Η αναζήτηση μας οδηγεί στον πρώτο κόμβο αριστερά καθώς η τιμή 8 είναι $12 < 25$.
- Εφόσον ο κόμβος χωράει το πολύ δύο κλειδιά και ο συγκεκριμένος έχει μόνο μια τιμή (4) χωράει και η τιμή 8. Θα τοποθετηθεί δεξιά του 4 καθώς $8 > 4$.



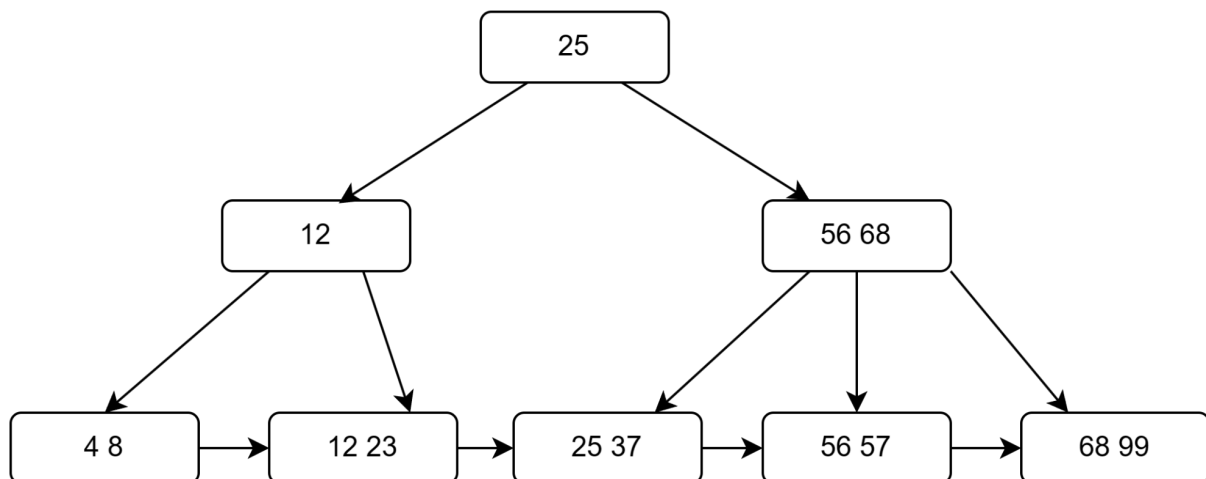
Εισαγωγή του 57

- Η αναζήτηση για την τιμή με 57 μας οδηγεί στον προτελευταίο κόμβο από δεξιά αφού $57 > 25$ και $56 < 57 < 68$.
- Όμοια με πριν, η τιμή 57 χωράει στον κόμβο οπότε τοποθετείται εκεί.



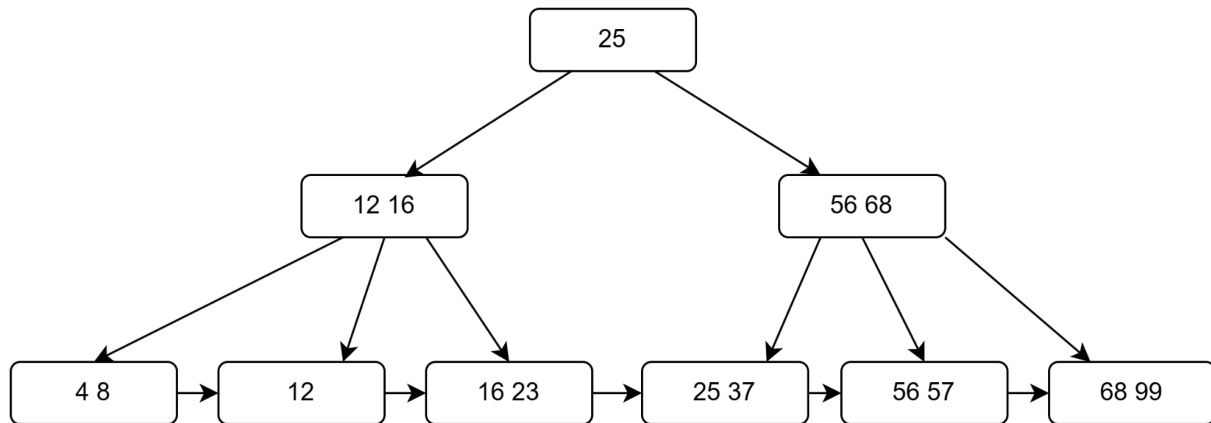
Εισαγωγή του 23

- Η αναζήτηση για την τιμή 23 μας οδηγεί στον δευτερο κόμβο από τα αριστερά εφόσον $23 < 25$ και $23 > 12$.
- Με την ίδια λογική η τιμή 23 αποθηκεύεται στον δεύτερο κόμβο φύλλο.



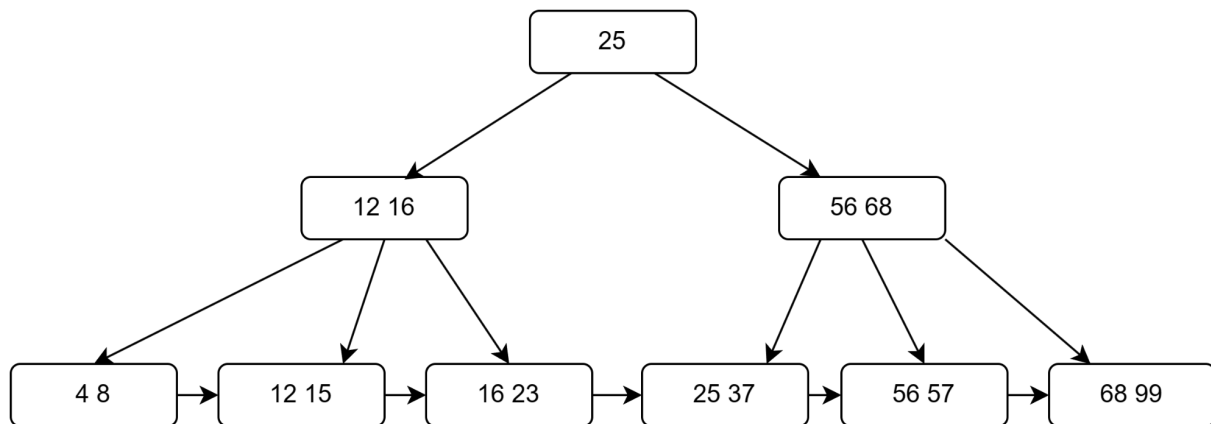
Εισαγωγή του 16

- Η τιμή 16, με την ίδια λογική τοποθετείται και αυτή μεταξύ του 12 και 23. Ωστόσο προκύπτει υπερχείλιση αφού σε έναν κόμβο χωράνε το πολύ δύο τιμές.
- Θα δημιουργηθεί ένας νέος κόμβος με τις τιμές 16 και 23 και η τιμή 16 θα αναγραφεί στο πάνω κόμβο μαζί με ένα δείκτη προς τον νέο κόμβο.
- Ο πάνω κόμβος περιέχει μόνο την τιμή 12, επομένως μπορεί να αποθηκευτεί και η τιμή 16, χωρίς να προκύψει υπερχείλιση.



Εισαγωγή του 15

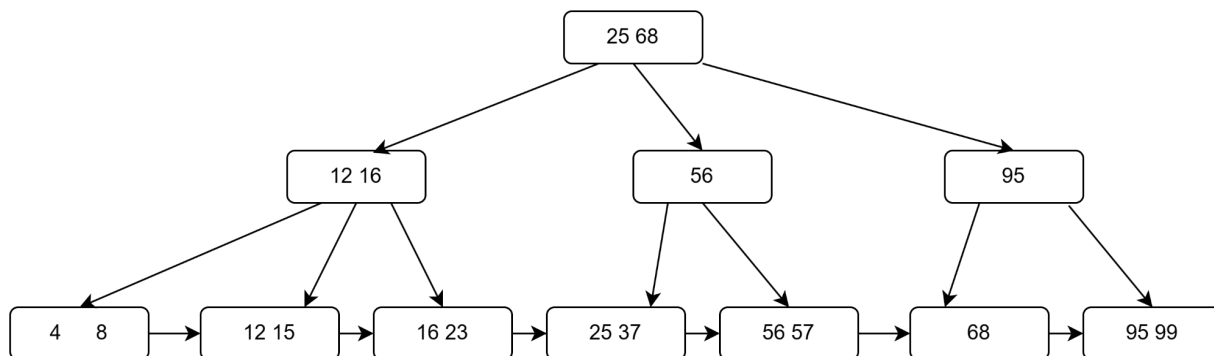
- Η αναζήτηση για την τιμή 15 μας οδηγεί στον δευτερο κόμβο από τα αριστερά εφόσον $15 < 25$ και $12 < 15 < 16$.
- Η τιμή 15 αποθηκεύεται στον δεύτερο κόμβο φύλλο, αφού χωράει.



Εισαγωγή του 95

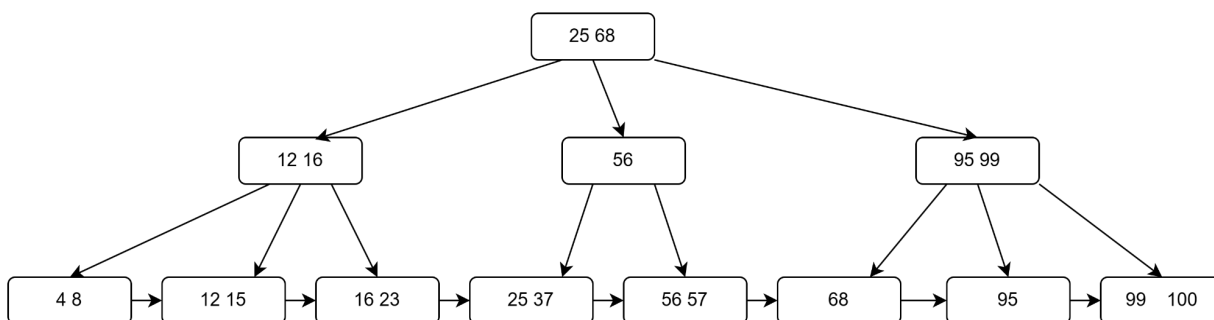
- Η αναζήτηση για την τιμή 95 μας οδηγεί στο τελευταίο κόμβο, καθώς $95 > 25$ και $95 > 68$.
- Η τιμή 95 θα τοποθετηθεί στο τελευταίο φύλλο στη μέση, αφού $68 < 95 < 99$. Αυτή η εισαγωγή όμως προκαλεί υπερχείλιση λόγω του ότι ήταν ήδη δύο τιμές αποθηκευμένες στο κόμβο αυτό.
- Θα δημιουργηθεί νέος κόμβος με τις τιμές 95 και 99 και η τιμή 95 θα αναγραφεί στον πάνω κόμβο μαζί με ένα δείκτη προς τον νέο κόμβο.
- Στον πάνω κόμβο όμως θα προκύψει υπερχείλιση, καθώς είναι ήδη δύο τιμές τοποθετημένες.

- Ανεβάζουμε τη τιμή 68(median) στη ρίζα και προσθέτουμε έναν δείκτη προς τον νέο κόμβο.



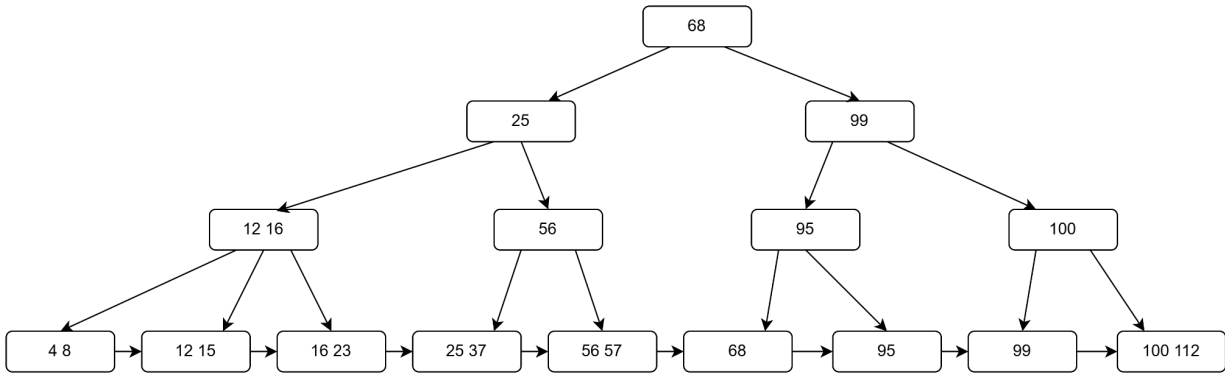
Εισαγωγή του 100

- Η αναζήτηση μας οδηγεί στο τελευταίο φύλλο αφού το $100 > 68$ και $100 > 95$.
- Η τοποθέτηση της τιμης 100 προκαλεί υπερχείλιση, οπότε δημιουργείται νέος κόμβος με τις τιμές 99 και 100.
- Ανεβαίνει η τιμή 99 στον πάνω κόμβο και αποθηκεύεται ένας δείκτης για τον νέο κόμβο.



Εισαγωγή του 112

- Η αναζήτηση μας οδηγεί στο τελευταίο φύλλο αφού το $112 > 68$ και $112 > 99$.
- Η τοποθέτηση της τιμης 112 προκαλεί υπερχείλιση, οπότε δημιουργείται νέος κόμβος με τις τιμές 100 και 112.
- Η τιμή 100 ανεβαίνει στον πάνω κόμβο, όπου προκαλεί υπερχείλιση και ο κόμβος διασπάται.
- Δημιουργείται νέος κόμβος με τη τιμή 100, και αποθηκεύονται και οι δείκτες στα φύλλα.
- Η τιμή 99 ανεβαίνει στον πάνω κόμβο ρίζα, προκαλείται υπερχείλιση και διασπάται ο κόμβος.
- Δημιουργείται νέα ρίζα με τιμή το 68(median).

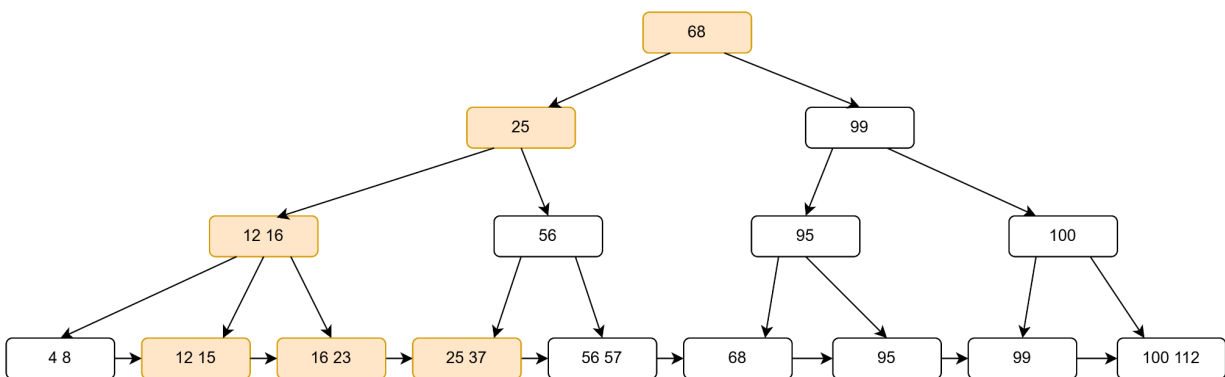


2. Ψάχνουμε τις τιμές του A με εύρος [12,37], όπου A γνωρίζουμε ότι είναι το πρωτεύον κλειδί της σχέσης R και είναι μοναδικό.

Αναζητούμε την πρώτη τιμή του εύρους 12 και προσπελαύνουμε τους κόμβους από την ρίζα προς τα κάτω (4 κόμβοι). Γνωρίζουμε ότι αριστερότερα από τον κόμβο (12 15) θα υπάρχουν τιμές μικρότερες του 12 και συνεχίζουμε να μετακινούμαστε μέσω των δεικτών προς τα δεξιά όπου οι τιμές είναι >12 . Μετακινούμαστε δεξιά στους κόμβους όσο δεν βρίσκουμε τιμή ≥ 37 . Μόλις εντοπίσουμε την τιμή 37 που είναι το τέλος του εύρους σταματάμε τη διάσχιση, εφόσον γνωρίζουμε ότι δεν θα υπάρχει άλλη εγγραφή με τιμή 37, γιατί το κλειδί αναζήτησης είναι μοναδικό. Τελικά προσπελάστηκαν 6 κόμβοι που φαίνονται με πορτοκαλί παρακάτω.

Για να ανακτήσουμε αυτές τις εγγραφές από τον δίσκο, αφού δεν είναι αποθηκευμένες στα φύλλα, θα χρειαστεί να προσπελάσουμε άλλα 6 μπλοκ (όσα και οι εγγραφές στο εύρος), αυτά που δείχνουν οι δείκτες στα φύλλα.

Συνολικά, θα προσπελαστούν 12 μπλοκ.



Άσκηση 4

Οι εισαγωγές των τιμών γίνονται με βάση τα i λιγότερο σημαντικά bits. Επομένως, μια τιμή εισάγεται στον κάδο που έχει ίσα τα τελευταία i bits. Π.χ. Το 1000 έχει ως τελευταίο bit το 0 ($i = 1$) και μπαίνει στο πρώτο κάδο.

1. Εισαγωγή του 1000

1000	

$$U = 1/6 \quad 16,67\%$$

0 1

2. Εισαγωγή του 0000

0000	
1000	

$$U = 2/6 \quad 33,33\%$$

0 1

3. Εισαγωγή του 1101

0000	
1000	1101

$$U = 3/6 \quad 50,00\%$$

0 1

4. Εισαγωγή του 0010

0010	
0000	
1000	1101

$$U = 4/6 \quad 66,67\%$$

0 1

5. Εισαγωγή του 0010

0010	
0010	
0000	
1000	1101
0	1

$U = 5/9 \quad 55,56\%$

6. Εισαγωγή του 1100

1100	
0010	
0010	
0000	
1000	1101
0	1

$U = 6/9 \quad 66,67\%$

7. Εισαγωγή του 0011

1100	
0010	
0010	
0000	0011
1000	1101

$$U = 7/9 \quad 77,77\%$$

0 1

To utilization > 70% -> το i = 2 και m->10

1100		
0000	0011	0010
1000	1101	0010

$$U = 7/9 \quad 77,77\%$$

00 *1 10

To utilization > 70% -> το i = 2 και m->11

1100			
0000		0010	
1000	1101	0010	0011

$$U = 7/12 \quad 58,00\%$$

00 01 10 11

8. Εισαγωγή του 1111

1100			
0000		0010	1111
1000	1101	0010	0011

$$U = 8/12 \quad 67,00\%$$

00 01 10 11

9. Εισαγωγή του 0110

1100		0110	
0000		0010	1111
1000	1101	0010	0011

$$U = 9/12 \quad 75,00\%$$

00 01 10 11

To utilization > 70% -> το i = 3 και m->100

		0110		
0000		0010	1111	
1000	1101	0010	0011	1100

$$U = 9/15 \quad 60,00\%$$

000 *01 *10 *11 100

10. Εισαγωγή του 1110

		1110		
		0110		
0000		0010	1111	
1000	1101	0010	0011	1100

$$U = 10/18 \quad 55,55\%$$

000 *01 *10 *11 100

Άσκηση 5

1. Τα 10 bits συνολικά ορίζουν τη διεύθυνση του block στο οποίο θα αποθηκευτεί η εγγραφή (a,b).

$$2^{10} = 1024$$

Κάθε συνδυασμός των 10 bits αντιστοιχεί σε ένα bucket.

Από τα 10 bits τα πρώτα N αφορούν την τιμή a και τα υπόλοιπα τη τιμή b.

Επομένως, από την συνάρτηση h_1 υπάρχουν μόνο 2^N μοναδικές τιμές κλειδί, οι οποίες επαναλαμβάνονται x φορές σε όλους τους 1024 buckets με διαφορετικά τα υπόλοιπα 10-N bits που προκύπτουν από την συνάρτηση h_2 , όπου x ο $\frac{\text{αριθμός των buckets}}{\text{μοναδικές τιμές}}$.

Στους 1024 buckets θα υπάρχουν $\frac{1024}{2^N}$ buckets που θα έχουν τα ίδια πρώτα bits και μέσα θα αποθηκεύονται εγγραφές με ίδιο a , αλλά διαφορετικό b .

Άρα αν θέλουμε να βρούμε τις εγγραφές που έχουν συγκεκριμένη τιμή a θα χρειαστεί να ψάξουμε στους κουβάδες που έχουν τα πρώτα N bits τους ίσα με την h_1 για το συγκεκριμένο a .

Όποτε για να βρώ όλες τις εγγραφές με συγκεκριμένη τιμή a ανεξαρτήτως του b θα χρειαστεί να ψάξω σε $\frac{1024}{2^N}$ buckets.

Για επερωτήσεις τύπου Q2 θα χρειαστεί να κοιτάξουμε τα τελευταία 10-N bits.

Ακολουθώντας την ίδια λογική που περιγράφηκε παραπάνω θα χρειαστεί να ψάξουμε σε $\frac{1024}{2^{10-N}}$ buckets.

2. Από την εκφώνηση γνωρίζουμε ότι το 30% των επερωτήσεων για την σχέση R είναι της μορφής Q1 και το 70% είναι της μορφής Q2.

Στο ερώτημα (1) βρήκαμε τον αριθμό των μπλοκ που προσπελάζονται ανάλογα το N για το κάθε τύπο επερώτησης.

Αφού γνωρίζουμε το ποσοστό χρησιμοποίησης των επερωτήσεων, μπορούμε να βρούμε το μέσο αριθμό των μπλοκ που θα προσπελάζονται για τα Q1 και Q2.

Κατά μέσο όρο, λοιπόν, θα προσπελάζονται:

$$0,3 \cdot \frac{1024}{2^N} + 0,7 \cdot \frac{1024}{2^{10-N}}$$