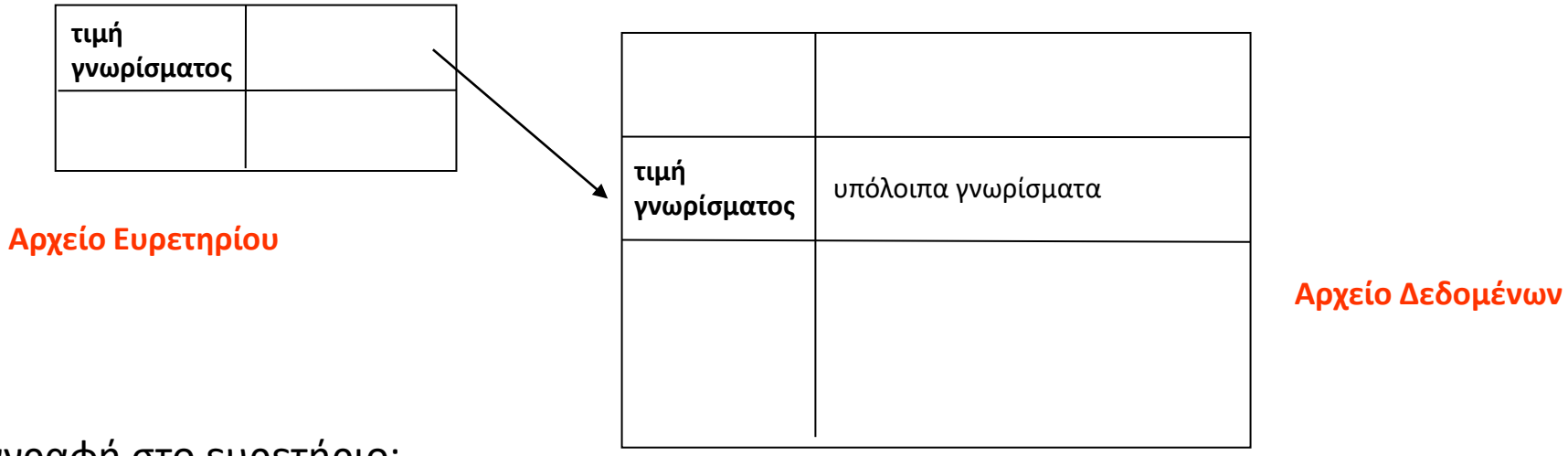
# Ευρετήρια

# Ευρετήρια

- Ένα ευρετήριο (index) είναι μια βοηθητική δομή που κάνει πιο αποδοτική την αναζήτηση μιας εγγραφής σε ένα αρχείο
- Το ευρετήριο ορίζεται (συνήθως) σε ένα γνώρισμα του αρχείου που καλείται πεδίο ευρετηριοποίησης (indexing field)
- Οι εγγραφές του ευρετηρίου είναι διατεταγμένες (διατεταγμένο αρχείο)




Εγγραφή στο ευρετήριο:

| Τιμή Πεδίου Ευρετηριοποίησης | Δείκτης στο block της εγγραφής |
|---|---|

# Παράδειγμα

## Russian_Novels

| BID | Title | Author | Published | Full_text |
|---|---|---|---|---|
| 001 | *War and Peace* | Tolstoy | 1869 | … |
| 002 | *Crime and Punishment* | Dostoyevsky | 1866 | … |
| 003 | *Anna Karenina* | Tolstoy | 1877 | … |

```
SELECT *
FROM Russian_Novels
WHERE Published > 1867
```

# Παράδειγμα

**By_Yr_Index**

| Published | BID |
|---|---|
| 1866 | |
| ***1869*** | |
| ***1877*** | |

**Russian_Novels**

| BID | Title | Author | Published | Full_text |
|---|---|---|---|---|
| 001 | *War and Peace* | Tolstoy | 1869 | … |
| 002 | *Crime and Punishment* | Dostoyevsky | 1866 | … |
| 003 | *Anna Karenina* | Tolstoy | 1877 | … |

Συνήθως, μόνο δείκτη (στη σελίδα που περιέχεται η εγγραφή (id σελίδας) ή και στη συγκεκριμένη εγγραφή στη σελίδα (id-σελίδας, id-εγγραφής)

Ορισμένα είδη ευρετηρίου την ίδια την εγγραφή

# Ευρετήρια

- Στόχος: αποδοτικές *λειτουργίες αναζήτησης*
- Οι λειτουργίες ενημέρωσης γίνονται γενικά πιο αργές, γιατί απαιτούν ενημέρωση **και** του ευρετηρίου

***Ποιες*** εγγραφές μπαίνουν στο ευρετήριο;

Ανάλογα με το πεδίο ευρετηριοποίησης:

(α) πεδίο διάταξης του αρχείου ή όχι

(β) κλειδί ή όχι

- (πρωτεύον/δευτερεύον) – διαφορετικοί ορισμοί στα βιβλία

# Ευρετήρια

- Πυκνό ευρετήριο: μια καταχώρηση για κάθε εγγραφή του αρχείου
- Μη πυκνό ευρετήριο

# Πρωτεύον Ευρετήριο

Πρωτεύον ευρετήριο (primary index): ορισμένο στο *κλειδί διάταξης* του αρχείου

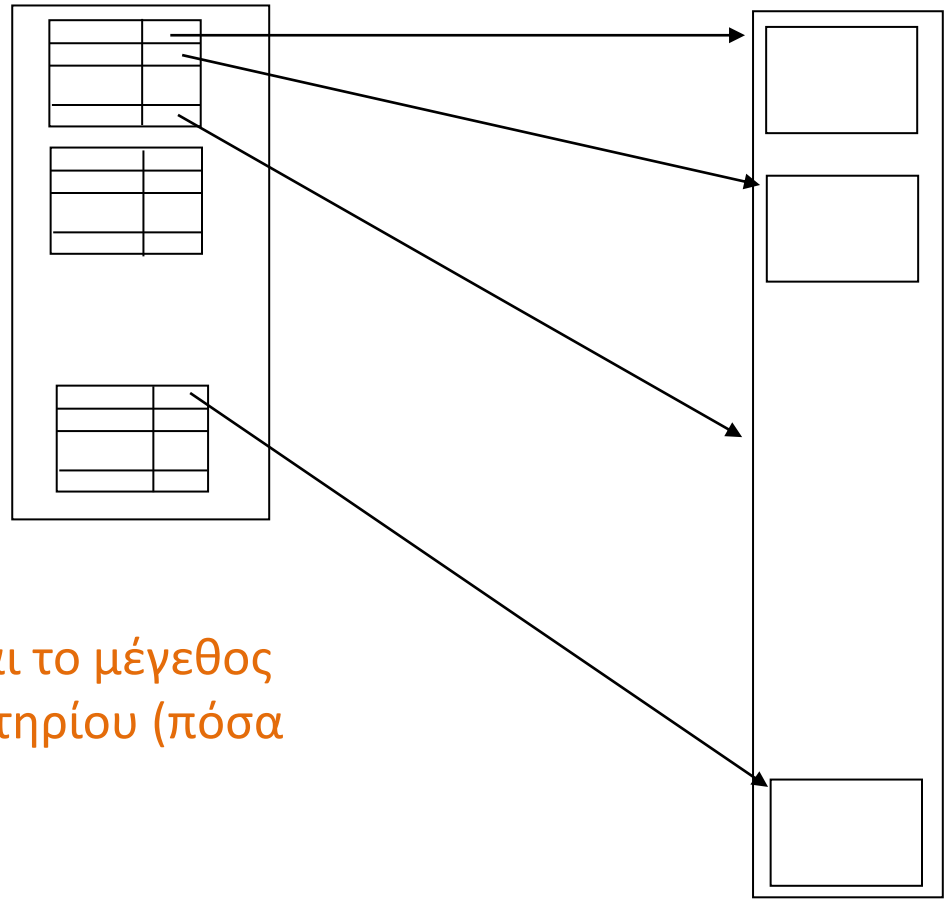
Για κάθε block του αρχείου (μη πυκνό ευρετήριο) η εγγραφή i του ευρετηρίου είναι της μορφής **(<K(i), P(i)>)** όπου:

- **K(i):** η τιμή του πρωτεύοντος κλειδιού της πρώτης εγγραφής του block (*άγκυρα* του block)
- **P(i):** δείκτης προς το block

✓ Ένα ευρετήριο στο πεδίο διάταξης (+ κλειδί) είναι ένα *μη πυκνό* ευρετήριο

# Πρωτεύον Ευρετήριο

Αρχείο Ευρετηρίου



Αρχείο Δεδομένων

Ποιο είναι το μέγεθος του ευρετηρίου (πόσα blocks);

# Πρωτεύον Ευρετήριο

## Παράδειγμα (υπολογισμός μεγέθους αρχείου ευρετηρίου)

Έστω διατεταγμένο αρχείο με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, το κλειδί διάταξης έχει μέγεθος $V_A$ = 9 bytes, μη εκτεινόμενη καταχώρηση.

Κατασκευάζουμε πρωτεύον ευρετήριο, μέγεθος δείκτη block P = 6 bytes

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου: (68 εγγραφές/block), 45 blocks

# Πρωτεύον Ευρετήριο

## Αναζήτηση

*Δυαδική αναζήτηση* στο πρωτεύον ευρετήριο

Ανάγνωση του block από το αρχείο δεδομένων

# Πρωτεύον Ευρετήριο

## Παράδειγμα (υπολογισμός κόστους αναζήτησης)

Δεδομένα όπως πριν

(Έστω διατεταγμένο αρχείο με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, κλειδί διάταξης έχει μέγεθος $V_A$ = 9 bytes, μη εκτεινόμενη καταχώρηση. Κατασκευάζουμε πρωτεύον ευρετήριο, μέγεθος δείκτη block P = 6 bytes)

$bfr_A = 10$

$bfr_E = 68$

*Μέγεθος αρχείου δεδομένων: 3.000 blocks*

*Μέγεθος αρχείου ευρετηρίου: 45 blocks*

Δυαδική γιατί το αρχείο διατεταγμένο

Αναζήτηση χωρίς ευρετήριο: $\lceil \log 3.000 \rceil = 12$ blocks

Αναζήτηση με ευρετήριο: $\lceil \log 45 \rceil + 1 = 7$ blocks

block ευρετηρίου

block αρχείου

# Πρωτεύον Ευρετήριο

## Εισαγωγή εγγραφής

αλλαγές και στο πρωτεύον ευρετήριο

μη διατεταγμένο αρχείο υπερχείλισης

συνδεδεμένη λίστα εγγραφών υπερχείλισης

## Διαγραφή εγγραφής

αλλαγές και στο πρωτεύον ευρετήριο

χρήση σημαδιών διαγραφής

# Ευρετήρια

Access paths (μονοπάτια προσπέλασης)

- Το ευρετήριο αρχείου είναι (πάντα) ένα *διατεταγμένο* αρχείο με σταθερού μήκους εγγραφές
- Το αρχείο ευρετηρίου καταλαμβάνει *μικρότερο χώρο* από το ίδιο το αρχείο δεδομένων (οι καταχωρήσεις είναι μικρότερες και (αν μη πυκνό) λιγότερες)
- Κάνοντας *δυαδική αναζήτηση* στο ευρετήριο (γιατί το ευρετήριο είναι διατεταγμένο αρχείο) βρίσκουμε το block όπου αποθηκεύεται η εγγραφή που αναζητούμε

# Ευρετήριο σε πεδίο διάταξης (όχι κλειδί)

Ευρετήριο συστάδων (clustering index): ορισμένο στο πεδίο διάταξης [το οποίο όμως δεν είναι κλειδί]

Υπάρχει *μια εγγραφή για κάθε διακεκριμένη τιμή* του πεδίου διάταξης (συστάδας) του αρχείου που περιέχει:

- την τιμή αυτή
- ένα δείκτη προς το πρώτο block του αρχείου δεδομένων που περιέχει μια εγγραφή με την τιμή αυτή στο πεδίο συστάδας

- Το ευρετήριο στο πεδίο διάταξης είναι ένα *μη πυκνό* ευρετήριο

# Ευρετήριο Συστάδων

- Ευρετήριο συστάδων ή συγκροτημένο ευρετήριο

  Όταν η διάταξη του ευρετηρίου ακολουθεί αυτή του αρχείου δεδομένων

# Ευρετήριο Συστάδων

## Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)

Έστω διατεταγμένο αρχείο με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο διάταξης έχει μέγεθος $V_A$ = 9 bytes και υπάρχουν *1000 διαφορετικές* τιμές και οι εγγραφές είναι ομοιόμορφα κατανεμημένες ως προς τις τιμές αυτές. Υποθέτουμε ότι χρησιμοποιούνται άγκυρες block, κάθε νέα τιμή του πεδίου διάταξης αρχίζει στην αρχή ενός νέου block. Κατασκευάζουμε ευρετήριο συστάδων, μέγεθος δείκτη block P = 6 bytes

$bfr_A = 10$

$bfr_E = 68$

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος ευρετηρίου συστάδων: 15 blocks

# Ευρετήριο Συστάδων

## Αναζήτηση

Δυαδική αναζήτηση στο ευρετήριο

Ανάγνωση blocks (τώρα μπορεί να είναι παραπάνω από ένα) από το αρχείο δεδομένων που περιέχουν την τιμή

# Ευρετήριο Συστάδων

## Παράδειγμα (υπολογισμός κόστους αναζήτησης)

(στοιχεία όπως πριν) Έστω διατεταγμένο αρχείο με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο διάταξης έχει μέγεθος $V_A$ = 9 bytes και υπάρχουν 1000 διαφορετικές τιμές και οι εγγραφές είναι ομοιόμορφα κατανεμημένες ως προς τις τιμές αυτές. Υποθέτουμε ότι χρησιμοποιούνται άγκυρες block, κάθε νέα τιμή του πεδίου διάταξης αρχίζει στην αρχή ενός νέου block. Κατασκευάζουμε ευρετήριο συστάδων, μέγεθος δείκτη block P = 6 bytes

*Μέγεθος αρχείου δεδομένων: 3.000 blocks*

*Μέγεθος αρχείου ευρετηρίου: 15 blocks*

Αναζήτηση χωρίς ευρετήριο: $\lceil \log 3.000 \rceil$ + ταιριάσματα (= 3) ≈ 15 blocks

Αναζήτηση με ευρετήριο: $\lceil \log 15 \rceil + 3$ = 7 blocks

# Δευτερεύον Ευρετήριο

Δευτερεύον ευρετήριο (secondary index): ορισμένο σε πεδίο διαφορετικό του πεδίου διάταξης

Θα εξετάσουμε την περίπτωση που το πεδίο ευρετηριοποίησης είναι κλειδί και την περίπτωση που δεν είναι

# Δευτερεύον Ευρετήριο

Περίπτωση 1: Το πεδίο ευρετηριοποίησης είναι *κλειδί* (καλείται και *δευτερεύον κλειδί*)

Υπάρχει *μια εγγραφή για κάθε εγγραφή του αρχείου* που περιέχει:

- την τιμή του κλειδιού για αυτήν την εγγραφή
- ένα δείκτη προς το block (ή την εγγραφή) του αρχείου δεδομένων που περιέχει την εγγραφή με την τιμή αυτή

✓ Το ευρετήριο σε πεδίο ΟΧΙ διάταξης (+ κλειδί) είναι ένα *πυκνό* ευρετήριο

# Δευτερεύον Ευρετήριο

## Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)

Έστω αρχείο με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, μη εκτεινόμενη καταχώρηση, όπου το *πεδίο κλειδιού* έχει μέγεθος $V_A$ = 9 bytes αλλά *δεν είναι πεδίο διάταξης*. Κατασκευάζουμε δευτερεύον ευρετήριο, μέγεθος δείκτη block P = 6 bytes

$bfr_A = 10$

$bfr_E = 68$

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου: 442 blocks

45 για πρωτεύον

# Δευτερεύον Ευρετήριο

## Παράδειγμα (υπολογισμός κόστους αναζήτησης)

Στοιχεία όπως πριν

(Έστω αρχείο με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο κλειδιού έχει μέγεθος $V_A$ = 9 bytes αλλά δεν είναι πεδίο διάταξης. Κατασκευάζουμε δευτερεύον ευρετήριο, μέγεθος δείκτη block P = 6 bytes)

*Μέγεθος αρχείου δεδομένων: 3.000 blocks*

*Μέγεθος αρχείου ευρετηρίου: 442 blocks*

$bfr_A = 10$

$bfr_E = 68$

Αναζήτηση χωρίς ευρετήριο (σειριακή αναζήτηση, γιατί το αρχείο δεδομένων δεν είναι ταξινομημένο): κατά μέσο όρο 3.000/2 = **1500** blocks

Αναζήτηση με ευρετήριο: $\lceil \log 442 \rceil + 1 =$ **10** blocks

Για πρωτεύον ήταν 45 και 7 blocks αντίστοιχα

# Δευτερεύον Ευρετήριο

Περίπτωση 2: Το πεδίο ευρετηριοποίησης *δεν είναι κλειδί*

1. Πυκνό ευρετήριο όπως πριν με μία καταχώρηση για κάθε εγγραφή

2. *Μεταβλητού μήκους εγγραφές με ένα επαναλαμβανόμενο πεδίο για το δείκτη*

3. Πυκνό ευρετήριο με δύο επίπεδα:

- Μία εγγραφή ευρετηρίου για κάθε τιμή του πεδίου ευρετηριοποίησης +
- Ένα ενδιάμεσο επίπεδο για την διαχείριση των πολλαπλών δεικτών

# Δευτερεύον Ευρετήριο

## Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)

Έστω μη διατεταγμένο αρχείο (αρχείο σωρού) με $r_A$ = 30.000 εγγραφές, μέγεθος block B = 1024 bytes, σταθερού μεγέθους εγγραφές μεγέθους $R_A$ = 100 bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο ευρετηριοποίησης (δηλαδή, το πεδίο στο οποίο θα κατασκευάσουμε το ευρετήριο) έχει μέγεθος $V_A$ = 9 bytes. Υπάρχουν 1000 διαφορετικές τιμές και οι εγγραφές είναι ομοιόμορφα κατανεμημένες ως προς τις τιμές αυτές. Κατασκευάζουμε ευρετήριο συστάδων χρησιμοποιώντας την επιλογή (3), μέγεθος δείκτη block P = 6 bytes

Ευρετήριο $bfr_E$ = 68 $b_E$ = 15 κόστος αναζήτησης;

Ενδιάμεσο επίπεδο (ΕΕ) -- *Ποια είναι η οργάνωση του;*

$bfr_{EE}$ = 170 $b_{EE}$ = 177 blocks

# Δευτερεύον Ευρετήριο

## Αναζήτηση

Δυαδική αναζήτηση στο δευτερεύον ευρετήριο

Ανάγνωση του block (ή των blocks) από το ενδιάμεσο επίπεδο

Ανάγνωση των blocks με τα ταιριάσματα (στη χειρότερη περίπτωση όσες οι εγγραφές που ταιριάζουν, γιατί δεν υπάρχει διάταξη) από το αρχείο δεδομένων

## Εισαγωγή

Απλή αν δεν αφορά εισαγωγή νέας τιμής στο ευρετήριο

# Ευρετήρια

- Επιπρόσθετες δομές για την πιο αποδοτική εκτέλεση ερωτήσεων/αναζητήσεων – προκαλούν όμως επιβάρυνση στις ενημερώσεις
- Εύκολη η λογική διάταξη των εγγραφών με βάση το πεδίο ευρετηριοποίησης
- Ανακτήσεις με *σύνθετες συνθήκες*, μπορεί να γίνουν χρησιμοποιώντας τα blocks του ευρετηρίου

# Ευρετήρια (σύνοψη)

Οι εγγραφές εξαρτώνται από το πεδίο ευρετηριοποίησης:

- Κλειδί
- Πεδίο διάταξης

Είδη ευρετηρίων

- **Πυκνό**: μια εγγραφή στο ευρετήριο για κάθε εγγραφή στο αρχείο δεδομένων (πλειάδα του πίνακα)
  Μη πυκνό

- **Πρωτεύον**: ευρετήριο σε πεδίο που είναι κλειδί και πεδίο διάταξης

- Συστάδων: σε πεδίο που είναι πεδίο διάταξης

# Ερωτήσεις;