

Άσκηση

Η παρακάτω SQL επερώτηση εμφανίζει τα ονόματα των φοιτητών που κατοικούν στην Αθήνα, σπουδάζουν σε ένα από τα 9 καλύτερα πανεπιστήμια της χώρας (ΚΑΤΑΤΑΞΗ < 10), και έχουν προεγγραφεί σε ένα Μεταπτυχιακό Πρόγραμμα Σπουδών με γνωστικό αντικείμενο την επιστήμη των δεδομένων “Data Science”.

```
SELECT ΟΝΟΜΑ
FROM ΦΟΙΤΗΤΕΣ, ΠΑΝΕΠΙΣΤΗΜΙΑ, ΠΡΟΕΓΓΡΑΦΕΣ
WHERE ΦΟΙΤΗΤΕΣ.ΚΠ=ΠΑΝΕΠΙΣΤΗΜΙΑ.ΚΠ AND
      ΦΟΙΤΗΤΕΣ.ΑΜ=ΠΡΟΕΓΓΡΑΦΕΣ.ΑΜ AND
      ΠΟΛΗ='ΑΘΗΝΑ' AND ΚΑΤΑΤΑΞΗ < 10 AND ΜΠΣ='DS'
```

Ο πίνακας που ακολουθεί περιέχει στοιχεία για τις παραπάνω σχέσεις:

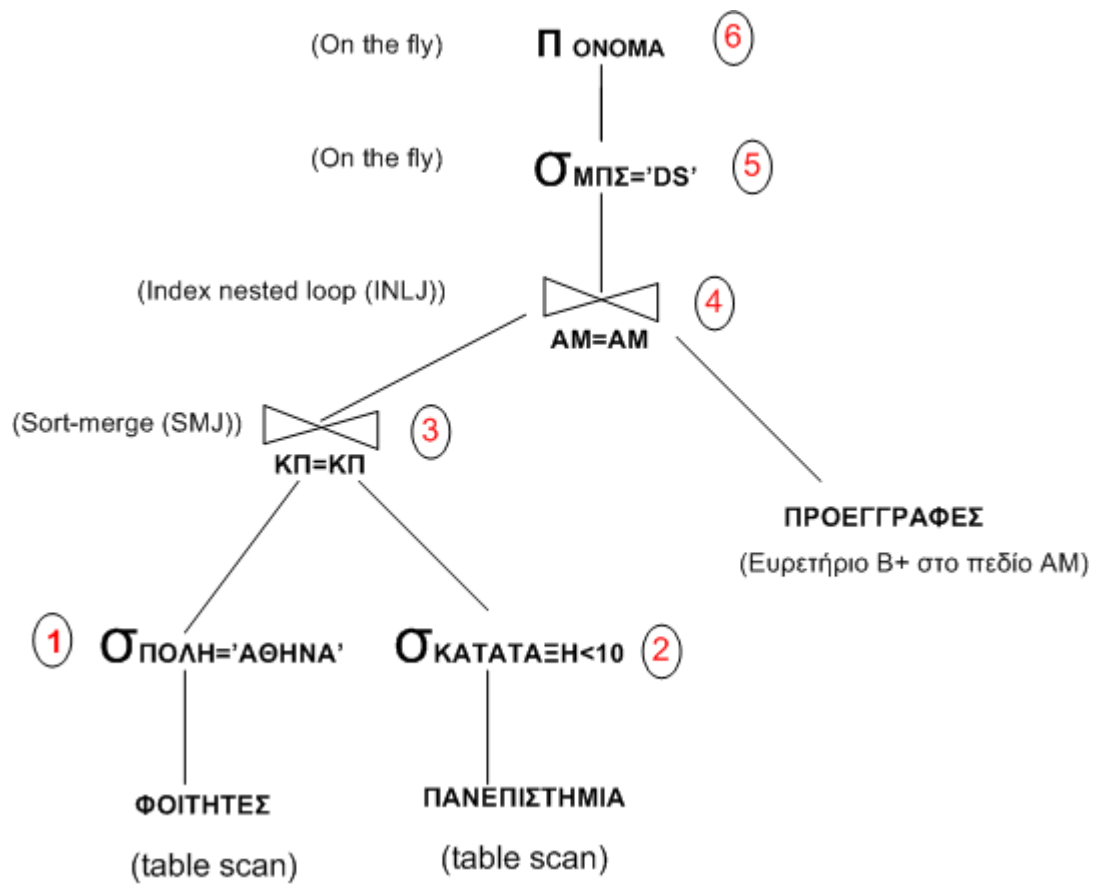
ΣΧΕΣΗ	ΑΡΙΘΜΟΣ ΕΓΓΡΑΦΩΝ	ΑΡΙΘΜΟΣ ΣΕΛΙΔΩΝ	ΠΡΩΤΕΥΟΝ ΚΛΕΙΔΙ
ΦΟΙΤΗΤΕΣ(<u>ΑΜ</u> , ΟΝΟΜΑ, ΠΟΛΗ, ΚΠ)	2000	100	ΑΜ (αριθμός μητρώου)
ΠΑΝΕΠΙΣΤΗΜΙΑ (<u>ΚΠ</u> , ΟΝΟΜΑΣΙΑ, ΚΑΤΑΤΑΞΗ)	100	10	ΚΠ (κωδικός πανεπιστημίου)
ΠΡΟΕΓΓΡΑΦΕΣ(<u>ΑΜ,ΜΠΣ</u>)	3000	200	(ΑΜ, ΜΠΣ) αριθμός μητρώου και κωδικός μεταπτυχιακού προγράμματος.

Επιπλέον δίνεται ότι:

- Κάθε πανεπιστήμιο έχει έναν μοναδικό αριθμό κατάταξης (ΚΑΤΑΤΑΞΗ) από το 1 μέχρι το 100.
- Υπάρχουν 20 διαφορετικές πόλεις.
- Κάθε υποψήφιος φοιτητής μπορεί να προεγγραφεί το πολύ σε 5 Μεταπτυχιακά Προγράμματα Σπουδών.
- Το πεδίο ΦΟΙΤΗΤΕΣ.ΚΠ είναι ξένο κλειδί το οποίο αναφέρεται (references) στο πεδίο ΠΑΝΕΠΙΣΤΗΜΙΑ.ΚΠ.
- Το πεδίο ΠΡΟΕΓΓΡΑΦΕΣ.ΑΜ είναι ξένο κλειδί το οποίο αναφέρεται (references) στο πεδίο ΦΟΙΤΗΤΕΣ.ΑΜ.
- Υπάρχει ένα απλό ευρετήριο (non clustered) στο πεδίο ΠΡΟΕΓΓΡΑΦΕΣ.ΑΜ και όλες οι σελίδες του ευρετηρίου βρίσκονται στην κύρια μνήμη. Αυτό είναι το μόνο ευρετήριο που υπάρχει. Μην θεωρήσετε ότι για κάθε πρωτεύον κλειδί του πίνακα υπάρχει ευρετήριο συστάδων (clustered index).
- Το μέγεθος της διαθέσιμης μνήμης είναι Μ=8 σελίδες.
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.

Ζητείται να υπολογίσετε το κόστος σε I/O του φυσικού πλάνου εκτέλεσης που ακολουθεί. Να υπολογίσετε το κόστος σε I/O (εφόσον υφίσταται) για κάθε μία από τις 6 επιμέρους λειτουργίες του πλάνου και να δείξετε πως αυτό προκύπτει.

ΦΥΣΙΚΟ ΠΛΑΝΟ ΕΚΤΕΛΕΣΗΣ



Λύση

Κόστος Βήματος 1: 100 I/O

Από τα δεδομένα της άσκησης έχουμε $T(\text{ΦΟΙΤΗΤΕΣ})=2000$, $B(\text{ΦΟΙΤΗΤΕΣ})=100$ και $V(\text{ΦΟΙΤΗΤΕΣ}, \text{ΠΟΛΗ})=20$. Το κόστος για το διάβασμα της σχέσης ΦΟΙΤΗΤΕΣ (table scan) είναι **100 I/O** και το αποτέλεσμα της επιλογής είναι $2000/20=100$ εγγραφές δηλαδή $100/20=5$ σελίδες (σε μία σελίδα χωράνε $2000/100=20$ εγγραφές της σχέσης ΦΟΙΤΗΤΕΣ).

Κόστος Βήματος 2: 10 I/O

Από τα δεδομένα της άσκησης έχουμε $T(\text{ΠΑΝΕΠΙΣΤΗΜΙΑ})=100$, $B(\text{ΠΑΝΕΠΙΣΤΗΜΙΑ})=10$ και $V(\text{ΠΑΝΕΠΙΣΤΗΜΙΑ}, \text{ΚΑΤΑΞΗ})=100$. Το κόστος για το διάβασμα της σχέσης ΠΑΝΕΠΙΣΤΗΜΙΑ είναι **10 I/O**. Το αποτέλεσμα της επιλογής είναι **9 εγγραφές** οι οποίες χωράνε σε **1 σελίδα** (σε μία σελίδα χωράνε $100/10=10$ εγγραφές της σχέσης ΠΑΝΕΠΙΣΤΗΜΙΑ).

Κόστος Βήματος 3: 0 (μηδέν) I/O.

Η είσοδος στον τελεστή ισοσύνδεσης είναι 6 σελίδες ($5 + 1$) η λειτουργία sort-merge μπορεί να εκτελεστεί στην μνήμη ($M=8$). Με άλλα λόγια δεν υπάρχει κόστος για I/O. Στην συγκεκριμένη περίπτωση πρόκειται για ισοσύνδεση μεταξύ πρωτεύοντος και ξένου κλειδιού δηλαδή κάθε εγγραφή της σχέσης ΦΟΙΤΗΤΕΣ αντιστοιχεί **το πολύ σε μία** εγγραφή της σχέσης ΠΑΝΕΠΙΣΤΗΜΙΑ. Έχουν προηγηθεί οι επιλογές με βάση την πόλη και την κατάταξη οι οποίες όμως είναι μεταξύ τους ανεξάρτητες και μόνο 9 ($9 \cdot (1/100) \cdot 100$) φοιτητές αντιστοιχούν σε κάποιο πανεπιστήμιο με κατάταξη < 10 .

Κόστος Βήματος 4: 45 I/O.

Για κάθε μια από τις 9 εγγραφές που δέχεται ως είσοδο, ο αλγόριθμος INLJ χρησιμοποιεί το ευρετήριο που υπάρχει στο πεδίο AM της σχέσης ΠΡΟΕΓΓΡΑΦΕΣ για να εντοπίσει και να ανακτήσει τις αντίστοιχες εγγραφές της σχέσης ΠΡΟΕΓΓΡΑΦΕΣ (look-up/probe). Δεδομένου ότι κάθε υποψήφιος φοιτητής μπορεί να προεγγραφεί το πολύ σε 5 μεταπτυχιακά προγράμματα, για κάθε AM θα ανακτηθούν το πολύ 5 εγγραφές της σχέσης ΠΡΟΕΓΓΡΑΦΕΣ. Αν υποθέσουμε ότι αυτές οι εγγραφές βρίσκονται σε διαφορετικές σελίδες τότε για κάθε μία από τις 9 εγγραφές φοιτητών θα διαβάσει το πολύ 5 σελίδες. Άρα το συνολικό κόστος είναι $9 \cdot 5 = 45$ I/O.

Σημείωση: Αν το ευρετήριο στο πεδίο AM ήταν ευρετήριο συστάδων (clustered index) τότε για κάθε έναν από τους 9 φοιτητές θα διάβαζε μια σελίδα (5 εγγραφές της σχέσης προεγγραφές χωράνε σε μία σελίδα) οπότε το κόστος θα ήταν 9 I/O.

Κόστος Βήματος 5: 0 I/O.

On the fly

Κόστος Βήματος 6: 0 I/O.

On the fly

Συνολικό Κόστος: $100 + 10 + 45 = 155$