Ατομική Εργασία για το Μάθημα Αποθήκες και Εξόρυξη Δεδομένων 2009-2010
(2 μονάδες[1])

Γενικές Οδηγίες

Υλοποιείστε τους αλγορίθμους *Multiway* και *BUC,* χρησιμοποιώντας προγραμματιστική γλώσσα της αρεσκείας σας. Οι αλγόριθμοι περιγράφονται στις παρακάτω εργασίες:

- Yihong Zhao, Prasad Deshpande, Jeffrey F. Naughton: An Array-Based Algorithm for Simultaneous Multidimensional Aggregates. SIGMOD Conference 1997: 159-170
- Kevin S. Beyer, Raghu Ramakrishnan: Bottom-Up Computation of Sparse and Iceberg CUBEs. SIGMOD Conference 1999: 359-370

και σε βιβλία όπως το «Data Mining: Concepts and Techniques» των Jiawei Han, Micheline Kamber.

Οι αλγόριθμοι θα διαβάζουν εγγραφές *n* διαστάσεων και ο κύβος θα αποθηκεύει το *πλήθος* των εγγραφών για κάθε συνδυασμό τιμών. Οι αλγόριθμοι πρέπει να υλοποιηθούν με τέτοιο τρόπο ώστε να υποστηρίζουν ερωτήματα παγόβουνου με τη συνθήκη *HAVING COUNT(*) ≥ X.* Τα δεδομένα θα προέρχονται από το αρχείο data-100K-6d-12.txt. Το αρχείο αυτό περιέχει 100000 εγγραφές με 6 διαστάσεις. Οι τιμές σε κάθε διάσταση είναι από 1 μέχρι 12.

Οι παράμετροι που θα δέχεται το πρόγραμμα υλοποίησης θα είναι: α) το όνομα του αρχείου εισόδου β) ο αριθμός των διαστάσεων *n* (αν *n<6* τότε κάποιες από τις στήλες του αρχείου εισόδου θα αγνοούνται) γ) το *X,* δ) το μέγεθος *chunk* για τον αλγόριθμο *Multiway.*

Η έξοδος του κάθε αλγορίθμου πρέπει να είναι: α) ο χρόνος εκτέλεσης του αλγορίθμου (χωρίς την αρχική ανάγνωση του αρχείου) β) η τιμή για κάθε κελί κάθε κυβοειδούς, π.χ (12,*,*,*,5,*): 2134, με τη σειρά υπολογισμού.

Φοιτητές των οποίων ο ΑΜ λήγει σε 1, 3, 5, 7 ή 9.

Να συγκρίνετε τον χρόνο εκτέλεσης των δύο αλγορίθμων για όλους τους συνδυασμούς των παρακάτω παραμέτρων:

- Για τον αλγόριθμο *Multiway* το μέγεθος κάθε *chunk* ανά διάσταση είναι α) 1, β) 4.
- Ο αριθμός των διαστάσεων είναι από 3 μέχρι 6 (οι επιπλέον διαστάσεις στο αρχείο αγνοούνται).
- Το *X* είναι σταθερό στο 10.

Φοιτητές των οποίων ο ΑΜ λήγει σε 0, 2, 4, 6 ή 8.

Να συγκρίνετε τον χρόνο εκτέλεσης των δύο αλγορίθμων για όλους τους συνδυασμούς των παρακάτω παραμέτρων:

- Για τον αλγόριθμο *Multiway* το μέγεθος κάθε *chunk* ανά διάσταση είναι σταθερά 1.
- Ο αριθμός των διαστάσεων είναι 3 (οι επιπλέον διαστάσεις στο αρχείο αγνοούνται).
- Το *X* παίρνει τις τιμές 1,10, 20, 30, …, 100.

Παραδοτέο: Ένας φάκελλος με εκτυπωμένη την εργασία και CD με τον πηγαίο και τον εκτελέσιμο κώδικα για περιβάλλον Windows. Η εργασία να αναφέρει α) το γενικό σκεπτικό υλοποίησης, β) τα πειραματικά αποτελέσματα και το περιβάλλον εκτέλεσης, γ) σχολιασμό-ερμηνεία των πειραματικών αποτελεσμάτων και δ) οδηγίες χρήσης. Ο πηγαίος κώδικας να έχει εκτενή σχόλια.

[1] Οι γραπτές εξετάσεις θα είναι με άριστα το 9.