1η Εργασία στη Σχεδίαση ΒΔ 2021-2022

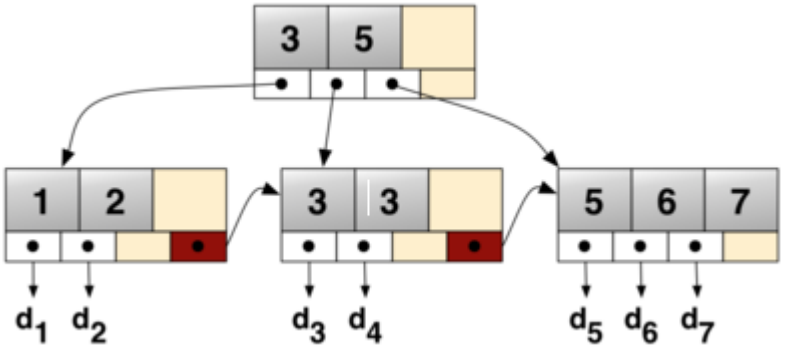
Ευρετήρια

Στόχος των εργασιών είναι η εξοικείωση με θεωρητικά και πρακτικά προβλήματα των Βάσεων Δεδομένων, μέσα από χρηστικά παραδείγματα. Στην πρώτη εργασία θα ασχοληθούμε με τις εντολές δημιουργίας ευρετηρίων και με τον υπολογισμό των χαρακτηριστικών ενός ευρετηρίου.

Α – Ερωτήματα

1ο μέρος – Ευρετήριο με Β+δέντρο

Ένα αρχείο έχει 20.000.000 εγγραφές αταξινόμητες σε σωρό. Οι εγγραφές έχουν κατά μέσο όρο μέγεθος 250 bytes και αποθηκεύονται σε αρχεία σε block μεγέθους 2048 bytes με μη εκτατή καταχώρηση.



Πάνω σε ένα αριθμητικό πεδίο (μεγέθους 8 bytes) που μπορεί και να περιέχει διπλότυπα, έχει χτιστεί ευρετήριο με χρήση Β+ δέντρου. Κάθε δείκτης προς τις εγγραφές του αρχείου (δείκτες d στην εικόνα) έχει μέγεθος 16 Bytes, κάθε δείκτης προς block του ευρετηρίου μαζί και ο δείκτης προς επόμενο φύλλο έχει μέγεθος 8 Bytes. Το ευρετήριο δημιουργείται μαζικά για όλο το αρχείο και μπορείτε να θεωρήσετε ότι οι κόμβοι ***στο τελευταίο επίπεδο*** του δέντρου είναι όσο γίνεται πλήρεις.
Υπολογίστε:
1) Το μέγεθος του αρχείου αν είναι αρχείο σωρού.
2) Το μέγεθος του αρχείου αν είναι αρχείο κατακερματισμού.
3) Πόσα επίπεδα έχει το Β+ δέντρο συμπεριλαμβανομένου και του τελευταίου επιπέδου
4) Πόσους κόμβους θα περιέχει το κάθε επίπεδό του, ποιο το μέγεθος του ευρετηρίου συνολικά.
5) Αν το δέντρο σας γίνει Β* ποια η απάντηση στο 4;
Εξηγήστε τη διαδικασία επίλυσης σε κάθε βήμα.
6) Ποιο θα είναι το κόστος αναζήτησης ισότητας για μια συγκεκριμένη τιμή που γνωρίζετε ότι εμφανίζεται 5 φορές σε όλο το αρχείο;

2ο μέρος – Εντολές δημιουργίας ευρετηρίων

Βρείτε τη σύνταξη της εντολής CREATE INDEX σε ORACLE, MySQL και PostgreSQL. Εξηγείστε τις παραμέτρους κάθε εντολής σε σχέση με τους διαθέσιμους τύπους ευρετηρίου που έχουμε δει στο μάθημα. Παρουσιάστε συγκριτικά (σε ένα πίνακα) τις δυνατότητες των τριών ΣΔΒΔ, ποια ευρετήρια υποστηρίζουν και ποια όχι η κάθε μια.
Αν χρησιμοποιήσατε πηγές από το δίκτυο να τις παραθέσετε (URL διευθύνσεις).

3ο μέρος – Δημιουργία και γέμισμα σχήματος

Στη ΒΔ σας θα πρέπει να δημιουργήσετε τρεις σχέσεις **Customers**, **Orders**, **Products** δεδομένα για τις οποίες θα πρέπει να αντιγράψετε από τη ΒΔ ***XSALES*** και συγκεκριμένα από τους πίνακες ***customers, products (and categories), orders*** και ***order_items***.

Ανεξάρτητα από το αρχικό σχήμα, οι σχέσεις θα πρέπει να έχουν την ακόλουθη δομή:

Η Customers:

CUSTOMER_ID	GENDER	AGEGROUP	MARITAL_STATUS	INCOME_LEVEL
101542	Female	above 70	single	low
47829	Female	40-50	single	medium
4940	Female	above 70	unknown	medium
12050	Female	50-60	married	medium
19162	Female	40-50	single	medium
104407	Female	above 70	single	high

α) Το πεδίο age_group θα προκύψει από το birth_date και την τρέχουσα ημερομηνία και με χρήση κατάλληλων updates που θα ετοιμάσετε και που θα μετατρέπουν την ηλικία στις ακόλουθες ομάδες: i) under 30, ii) 30-40, iii) 40-50, iv) 50-60, v) 60-70, vi) above 70, με ισότητα στο πάνω όριο.

β) Το πεδίο income level θα προκύψει με ομαδοποίηση των τιμών που έχει το αρχικό πεδίο ως εξής: i) εισόδημα ως 109.999 low, ii) εισόδημα ως 189.999 medium, iii) εισόδημα πάνω από 11000 high, iv) σε κάθε άλλη περίπτωση. Η ομαδοποίηση θα γίνει με κατάλληλα updates.

γ) Το πεδίο marital_status θα ομαδοποιηθεί με updates που θα αντιστοιχούν τα 'Married','Mabsent','married','Mar-AF' σε married, τα υπόλοιπα σε 'single' και τα null σε unknown.

Η products:

PRODUCT_ID	PRODUCTNAME	CATEGORYNAME	LIST_PRICE
15	Envoy 256MB - 40GB	Desktop PCs	999.99
28	Unix/Windows 1-user pack	Operating Systems	199.99
113	CD-R Mini Discs	Recordable CDs	22.99
114	Music CD-R	Recordable CDs	18.99
115	CD-RW, High Speed, Pack of 10	Recordable CDs	8.99
116	CD-RW, High Speed Pack of 5	Recordable CDs	11.99
117	CD-R, Professional Grade, Pack of 10	Recordable CDs	8.99
118	OraMusic CD-R, Pack of 10	Recordable CDs	7.99
119	CD-R with Jewel Cases, pACK OF 12	Recordable CDs	6.99

Η orders:

ORDER_ID	PRODUCT_ID	CUSTOMER_ID	DAYS_TO_PROCESS	PRICE	COST	CHANNEL
4631	27	2144	150	48.09	41.54	Direct Sales
12184	27	6228	51	48.09	41.54	Direct Sales
7324	27	3245	70	48.09	41.54	Direct Sales
14363	27	8312	91	48.09	41.54	Direct Sales
8594	27	3877	93	46.57	37.02	Direct Sales
6620	27	2935	31	46.74	41.24	Internet
12845	27	6794	31	46.74	41.24	Internet
1721	28	819	0	216.38	177.31	Direct Sales

Η στήλη price είναι η αρχική amount (του πίνακα order_items), η στήλη days_to_process είναι η διαφορά μεταξύ Order_date και Order_finished σε ημέρες.

4ο μέρος – Στατιστικά στοιχεία
Υπολογίστε:
α) τον αριθμό των πλειάδων κάθε πίνακα,
β) τον καταμερισμό των πλειάδων στις διαφορετικές τιμές των πεδίων gender, agegroup, marital_status, income_level του CUSTOMERS και channel του ORDERS.
Δώστε τις εντολές που τρέξατε για κάθε υποερώτημα και τα αποτελέσματα.

Β – Οδηγίες Παράδοσης

Η εργασία θα υλοποιηθεί από ομάδες των 3 ατόμων (το πολύ), αν και επιτρέπονται μικρότερες ομάδες. Θα πρέπει τελικά να ανεβάσετε ένα zip αρχείο με ονομασία τους ΑΜ των μελών της ομάδας: π.χ. ***AM1-AM2-AM3.zip***

- Το zip θα περιλαμβάνει:
  - ένα αρχείο readme.txt
    - με τα ονοματεπώνυμα και τους ΑΜ των φοιτητών της ομάδας
  - το αρχείο pdf με την τελική εργασία

Γ – Άλλες Οδηγίες

Όσες εργασίες δεν τηρούν τις οδηγίες παράδοσης, θα έχουν επίπτωση στο βαθμό.

Όσες εργασίες κριθούν ότι είναι **αντιγραφές θα μηδενίζονται**.

Ημερομηνία παράδοσης: **Στο e-class με οριστική τελική ημερομηνία 11-11-2021**

Όσες εργασίες παραδοθούν μετά το πέρας της ημερομηνίας και μέχρι τις 14-11-2021 θα έχουν μείωση 2 μονάδων στο βαθμό.