Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2024-2025

Δεύτερη Σειρά Ασκήσεων

Ανάθεση: 02-05-2025
Παράδοση: 13-05-2025 Ώρα (23:55)

Οδηγίες

- *Η δεύτερη σειρά ασκήσεων είναι **ατομική** και **υποχρεωτική**.*
- *Η υποβολή της εργασίας πρέπει να γίνει στο eclass.*
- *Το παραδοτέο σας θα πρέπει να είναι ένα αρχείο PDF με όνομα AM.pdf (όπου AM είναι ο αριθμός μητρώου σας. π.χ. "3220001.pdf").*
- *Τα διαγράμματα πρέπει να είναι κατασκευασμένα σε κάποιο πρόγραμμα (της επιλογής σας) και όχι σκαναρισμένα χειρόγραφα.*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*
- *Για την επίλυση των ασκήσεων να μελετήσετε τις διαφάνειες των διαλέξεων του μαθήματος.*
- **Εργασίες που παραδίδονται σε μορφή σκαναρισμένων χειργράφων δεν θα αξιολογηθούν.**

Το σύνολο της βαθμολογίας αθροίζει στις 105 μονάδες

Άσκηση 1 [μονάδες 25]
Δίνονται οι ακόλουθες σχέσεις των οποίων τα πρωτεύοντα κλειδιά είναι υπογραμμισμένα:

Cities (<u>cid</u>, city, country)
Airlines(<u>aid</u>, name)
Flights(<u>fid</u>, from_cid, to_cid, aid, duration)

για τις οποίες ισχύουν τα εξής:

Cities

- Η σχέση cities περιέχει 5000 εγγραφές
- Υπάρχουν 100 διαφορετικές χώρες (country)

Airlines

- Η σχέση Airlines περιέχει 300 εγγραφές.

Flights

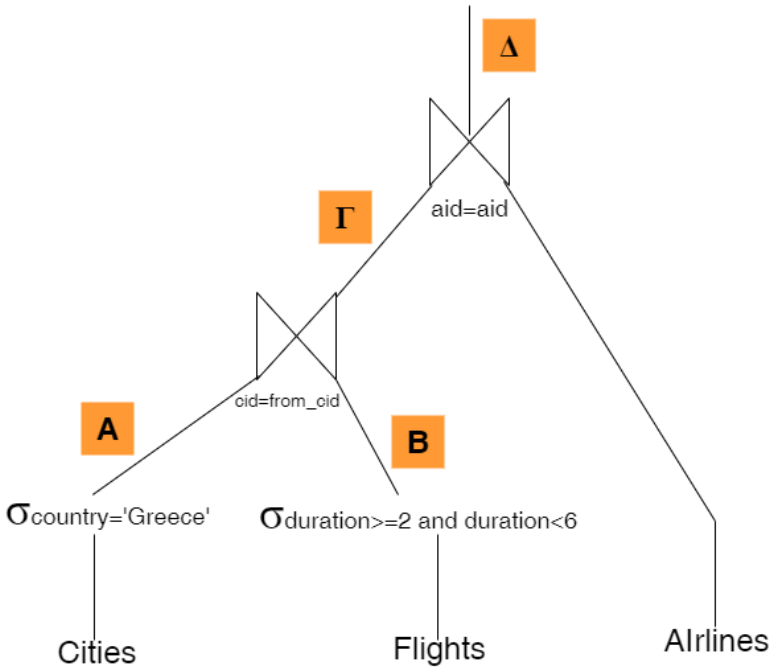
- Η σχέση flights περιέχει 100.000 εγγραφές.
- Οι τιμές του πρωτεύοντος κλειδιού (fid) ανήκουν στο διάστημα [1..100.000]
- Τα γνωρίσματα from_cid και to_cid είναι ξένα κλειδιά, εκ των οποίων το κάθε ένα αναφέρεται στο πρωτεύον κλειδί της σχέσης cities (REFERENCES cities(cid)).
- Το γνώρισμα aid είναι ξένο κλειδί το οποίο αναφέρεται στο πρωτεύον κλειδί της σχέσης Airlines.
- Το γνώρισμα duration δηλώνει την διάρκεια της πτήσης σε ώρες και λαμβάνει ακέραιες τιμές η κατανομή των οποίων ακολουθεί το παρακάτω ιστόγραμμα:

Duration (ώρες)	1-2	3-4	5-7	8-10	10-15
Ποσοστό πτήσεων	30%	40%	15%	10%	5%

Επιπλέον να θεωρήσετε ότι:

- Οι τιμές των γνωρισμάτων είναι μεταξύ τους ανεξάρτητες
- Οι επιλογές είναι μεταξύ τους ανεξάρτητες
- Όπου απαιτείται να θεωρήσετε ότι οι τιμές κατανέμονται ομοιόμορφα.
- Κανένα γνώρισμα δεν δέχεται τιμές NULL.

1. Στο παρακάτω λογικό πλάνο τα σημεία εξόδου κάθε τελεστή (επιλογή ή ισοσύνδεση) ονοματίζονται με τα γράμματα Α,Β,Γ και Δ. Ζητείται να υπολογίσετε τον αριθμό των εγγραφών σε κάθε σημείο εξόδου, δηλαδή να υπολογίσετε τα Τ(Α), Τ(Β), Τ(Γ) και Τ(Δ). Να δείξετε τον τρόπο υπολογισμού και όχι μόνο το τελικό αποτέλεσμα.



2. Να υπολογίσετε τον αριθμό των εγγραφών στο αποτέλεσμα της παρακάτω επερώτησης:

```
SELECT * FROM Flights WHERE fid > 50000
  UNION
SELECT * FROM Flights WHERE duration >=2 and duration<6
```

Να δείξετε τον τρόπο υπολογισμού και όχι μόνο το τελικό αποτέλεσμα.

Άσκηση 2 [μονάδες 25]

Έστω οι σχέσεις R1(a,b,c) και R2(b,e,f) οι οποίες καταλαμβάνουν αντίστοιχα 600 και 400 μπλοκ (σελίδες) στον δίσκο. Δεδομένου ότι η διαθέσιμη μνήμη είναι M=5 σελίδες ζητείται:

1. Να υπολογίσετε το βέλτιστο κόστος του αλγορίθμου SMJ για την σύζευξη των σχέσεων **R1⋈ R2 (**R1.b=R2.b) και να δείξετε αναλυτικά πως αυτό προκύπτει. Με άλλα λόγια, να δείξετε τον αριθμό και το μέγεθος (σε σελίδες) όλων των ταξινομημένων λιστών που θα δημιουργήσει ο αλγόριθμος για κάθε μία από τις σχέσεις R1 και R2, καθώς επίσης και το συνολικό κόστος (σε I/O) για την επίτευξη της σύζευξης.

2. Να προτείνετε κατάλληλα ευρετήρια ώστε να μειωθεί στο **ελάχιστο δυνατό** το κόστος (σε I/O) της σύζευξης των σχέσεων R1 και R2 με χρήση του αλγορίθμου SMJ. Να αναφέρετε την μορφή (b+tree, hash) καθώς και τον τύπο κάθε ευρετηρίου (clustered, non-clustered). Να υπολογίσετε το κόστος της σύζευξης των δύο σχέσεων μετά την χρήση των ευρετηρίων.

Άσκηση 3 [Μονάδες 30]

Έστω οι παρακάτω σχέσεις:

HOTEL (HotelID, Name, City, Stars)
REVIEW (HotelID, CustomeID, ReviewDate, Rating)

για τις οποίες ισχύουν τα εξής:

- Η σχέση HOTEL περιέχει 20.000 εγγραφές και σε μία σελίδα χωράνε 50 εγγραφές της σχέσης.
- Η σχέση REVIEW περιέχει 400.000 εγγραφές και σε μια σελίδα χωράνε 500 εγγραφές της σχέσης.

Επιπλέον θεωρείστε ότι:

- Το γνώρισμα Stars είναι ακέραιος αριθμός, με τιμές από 1 έως και 5, και αντιπροσωπεύει την κατηγορία ξενοδοχείου σύμφωνα με το επίπεδο των παρεχόμενων υπηρεσιών και των υποδομών. Τα ξενοδοχεία πέντε αστέρων (Stars=5) ανέρχονται στο 10% του συνόλου των ξενοδοχείων.
- Το πεδίο Rating της σχέσης REVIEW είναι ακέραιος αριθμός που καταγράφει την αξιολόγηση ενός ξενοδοχείου από έναν πελάτη. Οι τιμές του πεδίου κατανέμονται ομοιόμορφα στο διάστημα [1..10].
- Υπάρχει ευρετήριο συστάδων (clustered index) B+ δέντρο στο πεδίο REVIEW.Rating
- Υπάρχει απλό ευρετήριο (non-clustered index) B+ δέντρο στο πεδίο HOTEL.Stars
- Τα ευρετήρια βρίσκονται στην μνήμη του συστήματος.
- Η διαθέσιμη μνήμη είναι M=21 σελίδες.
- Όπου απαιτείται υποθέστε ότι τα δεδομένα κατανέμονται ομοιόμορφα.

Ζητείται:

A. Να σχεδιάσετε το τελικό, βελτιστοποιημένο λογικό πλάνο της παρακάτω επερώτησης. Δεν χρειάζεται να δείξετε τα ενδιάμεσα βήματα.

```
SELECT Name, City, Rating, ReviewDate
FROM HOTEL JOIN REVIEW ON  HOTEL.HotelID = REVIEW.HotelID
WHERE Stars = 5 AND Rating >= 8;
```

B. Να υπολογίσετε το ελάχιστο κόστος (σε I/O) εκτέλεσης της επερώτησης χρησιμοποιώντας του αλγορίθμους α) SMJ (Sort Merge Join) και b) BNLJ (Block Nested Loop Join).

Άσκηση 4 [Μονάδες 25]

Θεωρείστε την σχέση ΥΠΑΛΛΗΛΟΣ (Επώνυμο, Ειδικότητα, Τμήμα, Διεύθυνση) η οποία καταλαμβάνει 10.000 blocks (σελίδες) στον δίσκο. Όλα τα γνωρίσματα της σχέσης είναι αλφαριθμητικά, έχουν το ίδιο μήκος και καταλαμβάνουν τον ίδιο χώρο σε bytes για την αποθήκευσή τους στο δίσκο. Για την αποθήκευση των εγγραφών της σχέσης στο δίσκο να θεωρήσετε μόνο το χώρο αποθήκευσης των γνωρισμάτων.

Δίνεται η παρακάτω επερώτηση:

```
SELECT Επώνυμο
    FROM ΥΠΑΛΛΗΛΟΣ
  WHERE Ειδικότητα='Ηλεκτρολόγος' AND Τμήμα='Παραγωγή'
```

Έστω ότι:

- Μόνο το 10% των πλειάδων της σχέσης ικανοποιούν τη συνθήκη Ειδικότητα='Ηλεκτρολόγος'
- Μόνο το 10% των πλειάδων της σχέσης ικανοποιούν τη συνθήκη Τμήμα='Παραγωγή'
- Μόνο το 5% των πλειάδων της σχέσης ικανοποιούν και τις δύο συνθήκες ταυτόχρονα.

Για **κάθε μια** από τις παρακάτω περιπτώσεις να υπολογίσετε το κόστος (σε I/O) του βέλτιστου πλάνου εκτέλεσης υποθέτοντας ότι υπάρχει **μόνο το ευρετήριο που αναφέρεται** σε κάθε περίπτωση.

1. Ευρετήριο συστάδων (clustered index) στο γνώρισμα Ειδικότητα
2. Ευρετήριο συστάδων (clustered index) στο ζεύγος των γνωρισμάτων (Ειδικότητα, Τμήμα) με την σειρά που δίνονται.
3. Ευρετήριο συστάδων (clustered index) στα ακόλουθα τρία γνωρίσματα (Τμήμα, Ειδικότητα, Επώνυμο) με την σειρά που αυτά δίνονται.
4. Ευρετήριο συστάδων (clustered index) στα ακόλουθα τρία γνωρίσματα (Επώνυμο, Ειδικότητα, Τμήμα) με τη σειρά που αυτά δίνονται.

Σε **κάθε** μία από τις παραπάνω περιπτώσεις να θεωρήσετε ότι το αντίστοιχο ευρετήριο έχει την μορφή B+ δέντρου με 3 επίπεδα: ρίζα, ενδιάμεσο επίπεδο και φύλλα, και ότι είναι αποθηκευμένο σε αρχείο στον δίσκο. Στην μνήμη του συστήματος βρίσκονται μόνο οι σελίδες που περιέχουν τη ρίζα και τους ενδιάμεσους κόμβους (έχουν γίνει cached). Τέλος για λόγους απλούστευσης των υπολογισμών να θεωρήσετε στους υπολογισμούς σας ότι οι κόμβοι του β+ δέντρου (ενδιάμεσοι και φύλλων) περιέχουν **μόνο** τα κλειδιά αναζήτησης (δεν λαμβάνεται υπόψη ο χώρος που καταλαμβάνουν οι δείκτες μεταξύ κόμβων και οι δείκτες προς τα δεδομένα παρότι αυτοί θα υπάρχουν).