

Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων

Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2024-2025

Πρώτη Εργασία

Ανάθεση: **16-05-2025**

Παράδοση: **27-05-2025 Ώρα (23:55)**

Οδηγίες

- Η εργασία 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").
- Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.

Αποθήκες Δεδομένων

Το αρχείο **raceData.txt** περιέχει δεδομένα αγώνων τρεξίματος υπεραποστάσεων (50km, 50mi, 100km και 100mi) που έλαβαν χώρα κατά τα έτη 2015-2022.

Η Παγκόσμια Ομοσπονδία Υπερμαραθωνοδρόμων (Internatianla Association of Ultrarunners) επιθυμεί να αναπτύξει μια αποθήκη δεδομένων με σκοπό την άντληση στατιστικών στοιχείων για τους αγώνες υπεραποστάσεων και τις επιδόσεις των δρομέων. Οι απαιτήσεις της ομοσπονδίας εστιάζουν μεταξύ άλλων στην κατηγοριοποίηση των αγώνων ανα απόσταση και χώρα διεξαγωγής, καθώς επίσης και στην ανάλυση των επιδόσεων των δρομέων ανα ηλικιακή κατηγορία, φύλο και χώρα προέλευσης. Εξυπακούεται ότι στην ανάλυση των δεδομένων θα πρέπει να ληφθεί υπόψη και ο παράγοντας του χρόνου έτσι ώστε, η ομοσπονδία να μπορεί να παρακολουθεί τάσεις και επιδόσεις σε διάφορες χρονικές περιόδους και να παράγει αναφορές με στατιστικά στοιχεία ανα έτος, μήνα και ημέρα.

Καλείστε να σχεδιάσετε και να υλοποιήσετε την παραπάνω αποθήκη δεδομένων προκειμένου να αυξήσετε την αποτελεσματικότητα της διεξαγωγής χρήσιμων στατιστικών στοιχείων. Στην συνέχεια να τροφοδοτήσετε την αποθήκη με τα δεδομένα του αρχείου **raceData.txt** και να εκτελέσετε ορισμένες ερωτήσεις για την παραγωγή χρήσιμων στατιστικών στοιχείων. Την αποθήκη δεδομένων θα την υλοποιήσετε με την χρήση του DBMS SQL SERVER.

Ακολουθεί αναλυτική περιγραφή των δεδομένων και των ζητούμενων της εργασίας.

Το αρχείο **raceData.txt** περιέχει 973554 εγγραφές. Κάθε εγγραφή αποτελείται από 14 πεδία τα οποία διαχωρίζονται με τον χαρακτήρα "|" (pipe). Ακολουθεί η περιγραφή των πεδίων.

raceData.txt		
raceID	integer	Κωδικός αγώνα.
raceDate	date	Ημερομηνία διεξαγωγής του αγώνα (yyyy-mm-dd)
raceName	varchar(200)	Ονομασία αγώνα.
raceDistance	varchar(10)	Απόσταση του αγώνα (50km, 50mi, 100km, 100mi)
raceCountry	varchar(3)	Κωδικός Χώρας διοργάνωσης του αγώνα.
runnerID	integer	Κωδικός δρομέα.
runnerBirthYear	integer	Έτος γέννησης δρομέα.
runnerGender	varchar(1)	Φύλο δρομέα (F=female, M=male).
runnerCountry	varchar(3)	Κωδικός Χώρας προέλευσης του δρομέα.
ageCategoryCode	varchar(4)	Κωδικός ηλικιακής κατηγορίας
ageCategoryTitle	varchar(50)	Τίτλος ηλικιακής κατηγορίας
performance	varchar(20)	Ο Χρόνος τερματισμού του δρομέα σε μορφή hh:mm:ss (ώρες, λεπτά, δευτερόλεπτα)
finishTime	decimal(4,2)	Ο χρόνος τερματισμού του δρομέα σε ώρες.
averageSpeed	decimal(5,3)	Η μέση ταχύτητα του δρομέα ανά χιλιόμετρο.

Ζήτημα Πρώτο [60 Μονάδες]

Να δημιουργήσετε το λογικό σχήμα της αποθήκης δεδομένων και να το τροφοδοτήσετε με τα απαραίτητα δεδομένα. Συγκεκριμένα:

1. Να δημιουργήσετε μία βάση δεδομένων με όνομα **UMDW (Ultra Marathons Data Warehouse)**. Στη συνέχεια να εκτελέσετε το SQL script **createRaceData.sql** που δημιουργεί τον πίνακα **raceData** και φορτώνει τα δεδομένα από το αρχείο **raceData.txt**. Το SQL script **createRaceData.sql** περιέχει τις παρακάτω εντολές:

```
CREATE TABLE raceData
(
    raceID INT,
    raceDate DATE,
    raceName VARCHAR(200),
    raceDistance VARCHAR(10),
    raceCountry VARCHAR(3),
    runnerID INT,
    runnerBirthYear INT,
    runnerGender VARCHAR(1),
    runnerCountry VARCHAR(3),
    ageCategoryCode VARCHAR(4),
    ageCategoryTitle VARCHAR(50),
    performance VARCHAR(20),
    finishTime DECIMAL(4,2),
    averageSpeed DECIMAL(5,3)
);

BULK INSERT raceData
FROM 'C:\MY_DATA\raceData.txt' !!!ΠΡΟΣΟΧΗ ΠΡΟΣΑΡΜΟΣΤΕ ΤΟ PATH
WITH (FIRSTROW =2, FIELDTERMINATOR= '|', ROWTERMINATOR = '\n');
```

2. Να υλοποιήσετε το λογικό σχήμα της αποθήκης δεδομένων το οποίο θα πρέπει να έχει την μορφή αστέρα (Star Schema).
3. Να γράψετε κατάλληλες εντολές σε γλώσσα SQL, οι οποίες θα τροφοδοτούν το σχήμα της αποθήκης με τα απαραίτητα στοιχεία από τον πίνακα **raceData**.
4. Να αναπαραστήσετε διαγραμματικά το σχήμα της αποθήκης χρησιμοποιώντας την επιλογή "Database diagrams" του SQL Server Management Studio.

Η δημιουργία του λογικού σχήματος και η τροφοδότηση της αποθήκης με τα δεδομένα θα γίνουν με την εκτέλεση ενός **SQL script** το οποίο θα πρέπει να γράψετε.

Ζήτημα Δεύτερο [40 μονάδες]

Χρησιμοποιώντας την αποθήκη δεδομένων που δημιουργήσατε στο προηγούμενο ζήτημα (**όχι τον πίνακα racedata**), να γράψετε και να εκτελέσετε επερωτήσεις σε γλώσσα SQL, οι οποίες να απαντούν στα ακόλουθα ερωτήματα:

1. Εμφανίστε μια αναφορά με τον συνολικό αριθμό των αγώνων που έχουν διεξαχθεί σε κάθε χώρα, ανά έτος της μορφής: «Χώρα, Έτος, Αριθμός αγώνων». Η αναφορά να είναι ταξινομημένη ανά χώρα και έτος σε αύξουσα διάταξη.
2. Εμφανίστε μια αναφορά με τον μέσο χρόνο τερματισμού των δρομέων ανά ηλικιακή κατηγορία και φύλο στους αγώνες 50km: «Ηλικιακή Κατηγορία, φύλο, Μέσος Χρόνος Τερματισμού». Ο χρόνος τερματισμού είναι το πεδίο (finishTime). Η αναφορά να είναι ταξινομημένη με την ηλικιακή κατηγορία σε αύξουσα διάταξη.
3. Εμφανίστε μια αναφορά με μορφή «Έτος, Μήνας, Αριθμός Δρομέων» όπου για κάθε μήνα κάθε έτους να υπολογίζεται ο αριθμός των Ελλήνων δρομέων (όχι ο αριθμός των συμμετοχών) που έλαβαν μέρος σε αγώνες. Η αναφορά να είναι ταξινομημένη ανά έτος και μήνα σε φθίνουσα διάταξη.
4. Δημιουργείστε έναν κύβο δεδομένων κάθε κελί του οποίου θα περιέχει την αριθμό των συμμετοχών ανά χώρα διοργάνωσης, απόσταση αγώνα και φύλο δρομέα.

ΠΑΡΑΔΟΤΕΟ

Το παραδοτέο της εργασίας θα είναι ένα αρχείο pdf με όνομα AM.pdf (όπου AM είναι ο αριθμός μητρώου) το οποίο θα περιέχει τις απαντήσεις σας στο πρώτο και το δεύτερο ζήτημα:

- Τον κώδικα (εντολές SQL) για την δημιουργία του λογικού σχήματος της αποθήκης δεδομένων και την εισαγωγή των εγγραφών στους αντίστοιχους πίνακες.
- Το διάγραμμα του σχήματος αστέρα της αποθήκης δεδομένων.
- Τον κώδικα με τις επερωτήσεις SQL του δεύτερου ζητήματος.