ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Συστήματα Διαχείρισης Δεδομένων Μεγάλης Κλίμακας

(Χειμερινό Εξάμηνο 2024-2025)

Διδάσκων: Ιωάννης Κωτίδης (kotidis@aueb.gr)

Βοηθός: Χρυσόστομος Καπέτης (mkap@aueb.gr)

Οδηγίες

- Η εργασία είναι **ΑΤΟΜΙΚΗ**
- Η εργασία είναι **ΥΠΟΧΡΕΩΤΙΚΗ** και προσμετρά **50%** στον τελικό βαθμό του μαθήματος.

Ανάθεση: **06-12-2024**
Παράδοση: **29-12-2024 Ώρα (23:55)**

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 ετσι 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 επερωτήσεις για την παραγωγή χρήσιμων στατιστικών στοιχείων. Την αποθήκη δεδομένων θα την υλοποιήσετε με την χρήση του συστήματος SPARK. Για την παρουσίαση των αποτελεσμάτων θα αξιοποιήσετε τις δυνατότητες του Power BI.

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(5)**	Απόσταση του αγώνα (50km, 50mi, 100km, 100mi)
raceCountry	**char(3)**	Κωδικός Χώρας διοργάνωσης του αγώνα.
runnerID	**integer**	Κωδικός δρομέα.
runnerBirthYear	**integer**	Έτος γέννησης δρομέα.
runnerGender	**char(1)**	Φύλο δρομέα (F=female, M=male).
runnerCountry	**Integer**	Κωδικός Χώρας προέλευσης του δρομέα.
ageCategoryCode	**varchar(4)**	Κωδικός ηλικιακής κατηγορίας
ageCategoryTitle	**varcahr(100)**	Τίτλος ηλικιακής κατηγορίας
performance	**varchar(10)**	Ο Χρόνος τερματισμού του δρομέας 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 [Μονάδες 50]

1. Σχεδιάστε ένα διάγραμμα που να απεικονίζει το σχήμα αστέρα της αποθήκης δεδομένων. Στο διάγραμμα θα πρέπει να φαίνονται οι πίνακες με τα πεδία τους και οι συσχετίσεις των πινάκων διαστάσεων με τον πίνακα των γεγονότων. Για την δημιουργία του διαγράμματος μπορείτε να χρησιμοποιήσετε οποιοδήποτε σχεδιαστικό πρόγραμμα της αρεσκείας σας.

2. Καλείστε να γράψετε ένα πρόγραμμα σε Apache Spark, χρησιμοποιώντας μια από τις γλώσσες προγραμματισμού Scala/Python/Java το οποίο θα υλοποιεί τα παρακάτω:

   2.1 Θα δημιουργεί το λογικό σχήμα της αποθήκης δεδομένων το οποίο θα πρέπει να έχει την μορφή αστέρα (star schema) και θα το τροφοδοτεί με τα απαραίτητα δεδομένα από το αρχείο raceData.txt.

   2.2 Χρησιμοποιώντας το σχήμα της αποθήκης δεδομένων **(όχι το αρχείο raceData.txt)** να δημιουργεί τις παρακάτω στατιστικές αναφορές:

      2.2.1 Να παράγει αναφορά με τον 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.

      2.2.2 Να παράγει αναφορά με τον μέσο χρόνο τερματισμού (finishTime) των δρομέων ανά ηλικιακή κατηγορία στους αγώνες 50km: «Ηλικιακή Κατηγορία, Μέσος Χρόνος Τερματισμού». Η αναφορά να είναι ταξινομημένη με την Ηλικιακή Κατηγορία σε αύξουσα διάταξη.

      2.2.3 Να παράγει αναφορά με τον αριθμό των Ελλήνων δρομέων (όχι των ελληνικών συμμετοχών) που έλαβαν μέρος στους αγώνες ανά έτος. Η αναφορά πρέπει να έχει την μορφή «Έτος, Αριθμός Δρομέων» και θα είναι ταξινομημένη με βάση το έτος σε αύξουσα διάταξη.

      2.2.4 Να παράγει αναφορά με τον πιο γρήγορο αγώνα για κάθε απόσταση. Ως πιο γρήγορος αγώνας για κάθε απόσταση θεωρείται ο αγώνας στον οποίο οι δρομείς πέτυχαν κατά μέσο όρο την μεγαλύτερη μέση ταχύτητα (averageSpeed) στο σύνολο των ετών που διοργανώθηκε. Η αναφορά θα έχει την μορφή «Απόσταση Αγώνα, Ονομασία Αγώνα, Μέση Επίδοση Δρομέων»

2.2.5 Να δημιουργεί έναν κύβο δεδομένων κάθε κελί του οποίου θα περιέχει την αριθμό των συμμετοχών ανά χώρα διοργάνωσης, απόσταση αγώνα και φύλο δρομέα.

Σημείωση: Τα αποτελέσματα κάθε αναφοράς θα πρέπει να αποθηκεύονται σε ξεχωριστό αρχείο (ένα αρχείο για κάθε αναφορά).

2.3 Τέλος το πρόγραμμα πρέπει να αποθηκεύει τα δεδομένα που περιέχονται στους πίνακες της αποθήκης (πίνακες διαστάσεων και πίνακα γεγονότων) σε ξεχωριστά αρχεία CSV, ένα αρχείο CSV για κάθε πίνακα. Ο στόχος είναι τα αρχεία αυτά να χρησιμοποιηθούν για την μεταφόρτωση του σχήματος αστέρα στο Power BI.

Ζήτημα Δεύτερο [Μονάδες 50]

Στο σημείο αυτό καλείστε να χρησιμοποιήσετε το εργαλείο Power BI για να την δημιουργία γραφημάτων. Συγκεκριμένα:

1. Μεταφορτώστε στο PowerBI το σχήμα και τα δεδομένα της αποθήκης που δημιουργήσατε στο προηγούμενο ζήτημα. Με άλλα λόγια θα πρέπει να φορτώσετε τα αρχεία CSV που δημιουργήσατε στο παραπάνω υποερώτημα 2.3.

2. Δημιουργήστε κατάλληλα γραφήματα για την παρουσίαση των αποτελεσμάτων των πέντε αναφορών που δημιουργήσατε στο υποερώτημα 2.2 του πρώτου ζητήματος. Να δημιουργήσετε ένα γράφημα για κάθε αναφορά.

3. Εξερευνήστε τα δεδομένα της αποθήκης και δημιουργήστε ένα καλαίσθητο dashboard με κατάλληλα γραφήματα και άλλα στοιχεία παρουσίασης στο οποίο να απεικονίζονται οι πλέον χρήσιμες κατά την γνώμη σας πληροφορίες για τους αγώνες υπεραποστάσεων και τις επιδόσεις των δρομέων.

ΠΑΡΑΔΟΤΕΑ

Θα πρέπει να αναρτήσετε στο eclass έναν συμπιεσμένο φάκελο το όνομα του οποίου θα είναι ο αριθμός μητρώου. Ο φάκελος θα περιέχει:

1. Το αρχείο (ή αρχεία) με τον πηγαίο κώδικα του προγράμματος (Ζήτημα Πρώτο:2). Στην αρχή του αρχείου να γράψετε υπό την μορφή σχολίου το ονοματεπώνυμό σας και τον αριθμό μητρώου.

2. Ένα αρχείο pdf το οποίο θα περιέχει:

    - Το διάγραμμα που θα απεικονίζει το σχήμα αστέρα της αποθήκης (Ζήτημα – Πρώτο:1)
    - Τα γραφήματα και το dashboard που θα δημιουργήσετε στο δεύτερο ζήτημα

ΕΞΕΤΑΣΗ

Θα κληθείτε ατομικά να παρουσιάσετε την εργασία σε συγκεκριμένη ημέρα και ώρα. Κατά την διάρκεια της παρουσίασης θα ζητηθεί να μεταγλωττίσετε και να εκτελέσετε την εφαρμογή, να επιδείξετε την λειτουργικότητά της και να απαντήσετε σε σχετικές ερωτήσεις.