Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2021-2022

Δεύτερη Εργασία

Ανάθεση: 30-05-2022
Παράδοση: 13-06-2022 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. "3190001.pdf").*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*

Δημιουργία Αποθήκης Δεδομένων

Το αρχείο **ACCDATA.TXT** (Accidents Data) περιέχει στοιχεία τροχαίων ατυχημάτων που κατέγραψαν οι αστυνομικές αρχές του Ηνωμένου Βασιλείου κατά τα έτη 2005 έως και 2015.

Το Υπουργείο Μεταφορών της Αγγλίας ενδιαφέρεται να αναπτύξει μια αποθήκη δεδομένων με σκοπό την άντληση χρήσιμων πληροφοριών για την χάραξη πολιτικής σχετικά με την οδική ασφάλεια και την μείωση των θανατηφόρων τροχαίων ατυχημάτων.

Οι απαιτήσεις του υπουργείου εστιάζουν στην ανάλυση του αριθμού των τροχαίων ατυχημάτων, του αριθμού των θυμάτων και του αριθμού των εμπλεκόμενων οχημάτων, βάσει του φύλου και της ηλικίας των του υπαιτίου οδηγού, το είδος του οχήματός που προκάλεσε το ατύχημα, την σοβαρότητα του ατυχήματος (severity), την κατάσταση του οδοστρώματος (surface conditions) στο σημείο του ατυχήματος, καθώς και οποιονδήποτε συνδυασμό αυτών. Εξυπακούεται ότι στην ανάλυση των δεδομένων θα πρέπει να ληφθεί υπόψη και ο παράγοντας του χρόνου έτσι ώστε, οι αρμόδιες αρχές να είναι σε θέση να παράγουν στατιστικές αναφορές με τα στοιχεία των ατυχημάτων ανά μήνα, τρίμηνο και έτος.

Καλείστε να σχεδιάσετε και να υλοποιήσετε την παραπάνω αποθήκη δεδομένων προκειμένου να αυξήσετε την αποτελεσματικότητα της διεξαγωγής χρήσιμων στατιστικών στοιχείων, μειώνοντας ταυτόχρονα τον χρόνο εκτέλεσης των επερωτήσεων. Στην συνέχεια να τροφοδοτήσετε την αποθήκη με τα δεδομένα του αρχείου “ACCDATA.TXT” και να εκτελέσετε ορισμένες επερωτήσεις για την παραγωγή χρήσιμων στατιστικών αναφορών .

Περιγραφή Αρχείου ACCDATA.TXT

Το αρχείο **ACCDATA.TXT** περιέχει 2.119.396 εγγραφές. Κάθε εγγραφή αποτελείται από 14 πεδία τα οποία διαχωρίζονται με τον χαρακτήρα "|" (pipe). Ακολουθεί η περιγραφή των πεδίων.

		ACCDATA.TXT
accident_id	**varchar(15)**	Κωδικός ατυχήματος
severity_id	**integer**	Κωδικός χαρακτηρισμού σοβαρότητας
severity	**varchar(10)**	Χαρακτηρισμός της σοβαρότητας του ατυχήματος (fatal, serious κ.λπ.)
road_surface_conditions_id	**integer**	Κωδικός Κατάστασης οδοστρώματος
road_surface_conditions	**varchar(50)**	Κατάσταση του οδοστρώματος στο σημείου του ατυχήματος
accident_date	**date**	Ημερομηνία ατυχήματος
number_of_vehicles	**integer**	Ο αριθμός των οχημάτων που ενεπλάκησαν στο ατύχημα.
vehicle_type_id	**integer**	Κωδικός κατηγορίας του οχήματος που προκάλεσε το ατύχημα.
vehicle_type	**varchar(50)**	Κατηγορία του οχήματος που προκάλεσε το ατύχημα.
driver_class_id	**integer**	Κωδικός κατηγορίας οδηγού. Εκφράζει την κατηγορία στην οποία κατατάσσεται ο οδηγός που φέρει την ευθύνη του ατυχήματος ανάλογα με φύλο και την ηλικία. Π.χ. Οι γυναίκες οδηγοί 20 ετών κατατάσσονται στην κατηγορία 5 (5, Female, 20).
sex_of_driver	**varchar(6)**	Το φύλο του οδηγού που ευθύνεται για το ατύχημα
age_of_driver	**integer**	Η ηλικία του οδηγού που ευθύνεται για το ατύχημα
sex_of_casualty	**varchar(6)**	Φύλο θύματος. Ως θύμα νοείται οποιοδήποτε άτομο έχασε την ζωή του, τραυματίστηκε βαριά, ή ελαφριά εξαιτίας του ατυχήματος.
age_of_casualty	**integer**	Ηλικία θύματος

Ζήτημα Πρώτο [μονάδες 35]

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 **ACCIDENTSDW (Accidents Data Warehouse)**. Στη συνέχεια να δημιουργήσετε τον πίνακα **accdata** στον οποίο να φορτώσετε τα δεδομένα του αρχείου **ACCDATA.TXT** χρησιμοποιώντας την παρακάτω εντολή:

```
BULK INSERT accdata
FROM 'C:\data\ACCDATA.TXT'   !!! Προσαρμόστε 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 **accdata.**
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 [μονάδες 35]

Χρησιμοποιώντας την αποθήκη δεδομένων που δημιουργήσατε στο προηγούμενο ζήτημα, να γράψετε και να εκτελέσετε επερωτήσεις σε γλώσσα SQL, οι οποίες να απαντούν στα ακόλουθα ερωτήματα (απαιτήσεις) της διοίκησης του υπουργείου:

1. Εμφανίστε έναν κατάλογο με τον αριθμό των ατυχημάτων ανά έτος και σοβαρότητα (severity) ατυχήματος. Ο κατάλογος πρέπει να είναι ταξινομημένος με βάση το έτος σε φθίνουσας διάταξη.

2. Εμφανίστε έναν κατάλογο με τον αριθμό των **θανατηφόρων** ατυχημάτων (Fatal) και τον συνολικό αριθμό των θυμάτων τους ανά φύλο και ηλικία υπαίτιου οδηγού (οδηγού που φέρει την ευθύνη του ατυχήματος).

3. Εμφανίστε έναν κατάλογο με ανάλυση του αριθμού των ατυχημάτων βάσει της κατάστασης του οδοστρώματος (road surface conditions) και την σοβαρότητα του ατυχήματος.

4. Εμφανίστε έναν κατάλογο με τον αριθμό των ατυχημάτων και τον αριθμό των θυμάτων τους ανά έτος και τύπο του οχήματος (vehicle type) που προκάλεσε το ατύχημα. Ο κατάλογος να υπολογίζεται μόνο για τα ατυχήματα στα οποία εμπλέκονται περισσότερα από 2 οχήματα.

5. Η διοίκηση του υπουργείου θέλει μία αναφορά που θα περιέχει τις ακόλουθες πληροφορίες.
    a. Τον συνολικό αριθμό των ατυχημάτων, τον συνολικό αριθμό των οχημάτων που ενεπλάκησαν σε κάποιο ατύχημα και τον συνολικό αριθμό των θυμάτων στην διάρκεια της δεκαετίας (2005 έως 2015).
    b. Τον αριθμό των ατυχημάτων, τον αριθμό των οχημάτων που ενεπλάκησαν σε ατύχημα και τον αριθμό των θυμάτων σε ετήσια βάση.
    c. Τον αριθμό των ατυχημάτων, τον αριθμό των οχημάτων που ενεπλάκησαν σε ατύχημα και τον αριθμό των θυμάτων ανά τρίμηνο και μήνα κάθε έτους.

   Γράψτε **μια επερώτηση** σε γλώσσα SQL η οποία να παράγει την παραπάνω αναφορά.

Ζήτημα τρίτο [μονάδες 30]

1. Γράψτε μια επερώτηση σε γλώσσα SQL το αποτέλεσμα της οποίας είναι η δημιουργία ενός κύβου (data cube), κάθε κελί του οποίου περιέχει τον αριθμό των ατυχημάτων για έναν συγκεκριμένο συνδυασμό τιμών: σοβαρότητα ατυχήματος (severity), κατάσταση οδοστρώματος (road_surface_conditions) και τύπο οχήματος που προκάλεσε το ατύχημα (vehicle type).

2. Θεωρείστε ότι το DBMS δεν υποστηρίζει τον τελεστή **CUBE** για την δημιουργία του παραπάνω κύβου ούτε την εντολή **GROUP BY GROUPING SETS** παρά μόνο την εντολή **GROUP BY**. Δημιουργήστε μια **MATERIALIZED όψη (INDEXED VIEW στον SQL SERVER)** η οποία θα περιέχει το αποτέλεσμα **ενός μόνο** GROUP BY του κύβου του ερωτήματος 1. Γράψτε κατάλληλες εντολές SQL ώστε να παράγετε τα υπόλοιπα GROUP BY του κύβου **από την όψη** που δημιουργήσατε.