Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2023-2024

Δεύτερη Εργασία

Ανάθεση: 24-05-2024
Παράδοση: 02-06-2024 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 ΑΜ είναι ο αριθμός μητρώου σας. π.χ. "3210001.pdf").*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*

Αποθήκες Δεδομένων

Το αρχείο **inspections_data.txt** περιέχει ορισμένα στοιχεία των επιθεωρήσεων (inspections) που έχει διενεργήσει η υπηρεσία δημόσιας υγιεινής του υπουργείου υγείας των ΗΠΑ σε περίπου 3000 εστιατόρια κατά τα έτη 2009 έως και 2015.

Η υπηρεσία δημόσιας υγιεινής ενδιαφέρεται να αναπτύξει μια αποθήκη για την άντληση χρήσιμων πληροφοριών σχετικά με τα στοιχεία των επιθεωρήσεων. Οι απαιτήσεις της υπηρεσίας δημόσιας υγιεινής εστιάζουν μεταξύ άλλων στην ανάλυση του αριθμού των επιθεωρήσεων και των παραβάσεων ανα τύπο επιθεώρησης, κατηγορία παράβασης, περιοχή, καθώς και οποιονδήποτε συνδυασμό αυτών. Εξυπακούεται ότι στην ανάλυση των δεδομένων θα πρέπει να ληφθεί υπόψη και ο παράγοντας του χρόνου ετσι ώστε, η υπηρεσία να είναι σε θέση να παράγει στατιστικές αναφορές στοιχεία από τα ευρύματα των επιθεωρήσεων ανα έτος, μήνα, ημέρα κ.λπ.

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 inspections_data.txt και να εκτελέσετε ορισμένες επερωτήσεις 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 **inspections_data.txt** περιέχει 12571 εγγραφές. Κάθε εγγραφή αποτελείται από 15 πεδία τα οποία διαχωρίζονται με τον χαρακτήρα "|" (pipe). Ακολουθεί η περιγραφή των πεδίων.

inspections_data.txt		
rid	**integer**	Κωδικός εστιατορίου
lat	**float**	Γεωγραφικό πλάτος (latitude). Μαζί με το γεωγραφικό μήκος (longitude) δηλώνουν την τοποθεσία του εστιατορίου
lon	**flot**	Γεωγραφικό μήκος (lognitude).
insdate	**date**	Ημερομηνία επιθεώρησης σε μορφή yyyy-mm-dd.
insyear	**integer**	Έτος επιθεώρησης
insmonth	**integer**	Μήνας επιθεώρησης
insday	**integer**	Ημέρα του μήνα που έγινε η επιθεώρησης (1 έως και 31)
insweekday	**integer**	Ημέρα της εβδομάδας που έγινε η επιθεώρηση (1=Δευτέρα, 7=Κυριακή)
inscode	**integer**	Κωδικός τύπου επιθεώρησης.
instype	**nvarcahr(100)**	Τύπος επιθεώρησης (π.χ. Routine, New owner, Complaint κ.λπ.)
criticalIssue	**integer**	Αριθμός κρίσιμων ζητημάτων που εντοπίστηκαν κατά την επιθεώρηση.
nonCriticalIssue	**integer**	Αριθμός μη κρίσιμων ζητημάτων που εντοπίστηκαν κατά την επιθεώρηση.
vcode	**integer**	Κωδικός παράβασης
vdescription	**nvarcahr(255)**	Περιγραφή παράβασης
vcategory	**nvarcahr(255)**	Κατηγορία παράβασης

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 **INSDW (Inspections Data Warehouse)**. Στη συνέχεια να δημιουργήσετε τον πίνακα **inspections_data** στον οποίο να φορτώσετε τα δεδομένα του αρχείου **inspections_data.txt** χρησιμοποιώντας την παρακάτω εντολή:

```
BULK INSERT inspections_data
FROM 'C:\MY_DATA\inspections_data.txt'   !!!ΠΡΟΣΟΧΗ: Προσαρμόστε το PATH
WITH (DATAFILETYPE = 'widechar', 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 **inspections_data.**
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.

ΠΡΟΣΟΧΗ: Για αλφαριθμητικά πεδία να χρησιμοποιήσετε τον τύπο **nvarchar** (όχι varchar) διότι τα δεδομένα του αρχείου inspections_data.txt είναι σε μορφή Unicode.

Ζήτημα Δεύτερο [40 μονάδες]

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 επιθεωρήσεων ανά έτος και τύπο επιθεώρησης (instype). Ο κατάλογος πρέπει να είναι ταξινομημένος με βάση το έτος σε φθίνουσα διάταξη.

2. Εμφανίστε έναν κατάλογο με τα παρακάτω στοιχεία για κάθε εστιατόριο:
    - Κωδικός εστιατορίου
    - Συντεταγμένες
    - Αριθμός μη κρίσιμων ζητημάτων που εντοπίστηκαν συνολικά από όλες τις επιθεωρήσεις που έγιναν στο συγκεκριμένο εστιατόριο.
    - Αριθμός κρίσιμων ζητημάτων που εντοπίστηκαν συνολικά από όλες τις επιθεωρήσεις που έγιναν στο συγκεκριμένο εστιατόριο.
    - Αριθμός ζητημάτων (κρίσιμων και μη κρίσιμων) που εντοπίστηκαν συνολικά από όλες τις επιθεωρήσεις που έγιναν στο συγκεκριμένο εστιατόριο.

    Ο κατάλογος να είναι ταξινομημένος σε φθίνουσα διάταξη βάσει του αριθμού ζητημάτων (κρίσιμων και μη κρίσιμων).

3. Γράψτε μια επερώτηση σε γλώσσα SQL το αποτέλεσμα της οποίας είναι η δημιουργία ενός κύβου (data cube), κάθε κελί του οποίου περιέχει τον συνολικό αριθμό των κρίσιμων ζητημάτων που εντοπίστηκαν από όλες τις επιθεωρήσεις ανα τύπο επιθεώρησης (instype), κατηγορία παράβασης (vcategory) και έτος επιθεώρησης (insyear).

Ζήτημα Τρίτο [10 μονάδες Bonus]

Δημιουργήστε μια αναφορά (report) με το power BI με τα παρακάτω:

1. Κατάλληλο γράφημα στο οποίο θα απεικονίζονται τα αποτελέσματα της πρώτης επερώτησης που γράψατε στο προηγούμενο ζήτημα. Δηλαδή ο αριθμός των επιθεωρήσεων ανα έτος και τύπο επιθεώρησης.
2. Απεικονίστε σε έναν χάρτη την τοποθεσία των είκοσι εστιατορίων στα οποία εντοπίστηκαν τα περισσότερα ζητήματα (κρίσιμα και μη) από όλες τις επιθεωρήσεις. Πρόκειται για τα εστιατόρια του καταλόγου που δημιουργήσατε στο δεύτερο ερώτημα του προηγούμενου ζητήματος. Ο χάρτης να συνοδεύεται από ένα τρίστηλο πίνακα κάθε γραμμή του οποίου θα περιέχει τον κωδικό και τις συντεταγμένες ενός εστιατορίου.

ΠΑΡΑΔΟΤΕΑ

Τα παραδοτέα της εργασίας σας θα είναι τα εξής:

1. Ένα αρχείο pdf με όνομα AM.pdf το οποίο θα περιέχει τα παραδοτέα του πρώτου και του δεύτερου ζητήματος:

    - 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.

2. Το αρχείο .pbix που δημιουργήσατε με το Power BI. Το όνομα του αρχείου να είναι της μορφής: AM.pbix (Παραδοτέο τρίτου ζητήμαος).

ΠΡΟΣΟΧΗ

- **Οσοι δεν απαντήσετε στο τρίτο ζητήμα θα ανεβάσετε στο eclass μόνο το αρχείο PDF με τα παραδοτέα του πρώτου και του δεύτερου ζητήματος (AM.pdf) (όπου ΑΜ είναι ο αριθμός μητρώου σας).**

- **Οσοι απαντήσετε στο τρίτο ζήτημα θα τοποθετήσετε σε ένα φάκελο τα δύο αρχεία (PDF και PBIX) τον οποίο φάκελο στη συνέχεια θα τον συμπιέσετε σε ένα αρχείο της μορφής AM.zip το οποίο και θα ανεβάσετε στο eclass.**