

**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**  
**ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

**Συστήματα Διαχείρισης Δεδομένων Μεγάλης Κλίμακας**  
**(Χειμερινό Εξάμηνο 2022-2023)**

Διδάσκων: Ιωάννης Κωτίδης (kotidis@aueb.gr)

Βοηθός: Χρυσόστομος Καπέτης (mkar@aueb.gr)

Οδηγίες

- Η εργασία είναι ΑΤΟΜΙΚΗ
- Η εργασία είναι ΥΠΟΧΡΕΩΤΙΚΗ και προσμετρά 25% στον τελικό βαθμό του μαθήματος.

Ανάθεση: **11-11-2022**

Παράδοση: **21-11-2022 Ώρα (23:55)**

**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 έτσι ώστε, η υπηρεσία να είναι σε θέση να παράγει στατιστικές αναφορές με τα στοιχεία τα ευρήματα 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. Για την παρουσίαση των αποτελεσμάτων θα αξιοποιήσετε τις δυνατότητες του Power BI.

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	float	Γεωγραφικό μήκος (longitude).
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	nvarchar(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	nvarchar(255)	Περιγραφή παράβασης
vcategory	nvarchar(255)	Κατηγορία παράβασης

## Ζήτημα Πρώτο

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

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 κατηγορία παράβασης (vcategory) και έτος επιθεώρησης (insyear).

## Ζήτημα 3

1. Δημιουργήστε μια αναφορά (report) με το power BI με τα παρακάτω:
  - 1.1 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.
  - 1.2 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.
  - 1.3 Κατάλληλο γράφημα ή/και πίνακα για την αναπαράσταση των δεδομένων του κύβου που ζητείται στο προηγούμενο ζήτημα.
2. Εξερευνήστε τα δεδομένα της αποθήκης ώστε να εντοπίσετε επιπλέον πληροφορίες οι οποίες κατά την γνώμη σας είναι χρήσιμες για την υπηρεσία δημόσιας υγιεινής. Δημιουργήστε ένα καλαίσθητο dashboard με κατάλληλα γραφήματα και άλλα στοιχεία παρουσίασης (π.χ. χάρτες, πίνακες κ.λπ.) στο οποίο να απεικονίζονται οι πλέον χρήσιμες κατά την γνώμη σας πληροφορίες.

## **ΠΑΡΑΔΟΤΕΑ**

Τα παραδοτέα της εργασίας σας θα είναι τα εξής:

1. Ένα αρχείο pdf με όνομα AM.pdf το οποίο θα περιέχει:

### **Παραδοτέα πρώτου 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.

### **Παραδοτέα δεύτερου ζητήματος**

- Τον κώδικα με τις επερωτήσεις SQL του δεύτερου ζητήματος.

### **Παραδοτέα τρίτου ζητήματος**

- Την αναφορά (report) που δημιουργήσατε στο PowerBI
- Το dashboard που δημιουργήσατε

2. Το αρχείο .pbix που δημιουργήσατε με το power bi. Το όνομα του αρχείου να είναι της μορφής: AM.pbix

Τα δύο παραπάνω αρχεία (PDF, και ZIP) μπορείτε να τα τοποθετήσετε σε έναν φάκελο τον οποίο στην συνέχεια θα τον συμπιέσετε σε ένα αρχείο της μορφής AM.zip (όπου AM είναι ο αριθμός μητρώου σας) και θα το ανεβάσετε στο eclass.