Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2022-2023

Δεύτερη Εργασία

Ανάθεση: 26-05-2023
Παράδοση: 06-06-2023 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. "3200001.pdf").*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*

Δημιουργία Αποθήκης Δεδομένων

Το αρχείο **"CardsTransactions.txt"** περιέχει στοιχεία συναλλαγών χρεωστικών και πιστωτικών καρτών των πελατών μιας τράπεζας για τα έτη 2015 έως και 2020.

Η διοίκηση της τράπεζας ενδιαφέρεται να αναπτύξει μια αποθήκη δεδομένων με σκοπό την παραγωγή στατιστικών αναφορών και την άντληση χρήσιμων πληροφοριών σχετικά με τις ηλεκτρονικές συναλλαγές που πραγματοποιούνται μέσω καρτών.

Οι απαιτήσεις της διοίκησης εστιάζουν στην ανάλυση του αριθμού και της αξίας των συναλλαγών, βάσει της επωνυμίας και του είδους των καρτών, το φύλο και την ηλικία των κατόχων τους, τον τύπο της συναλλαγών, την πόλη στην οποία έλαβαν χώρα καθώς και οποιονδήποτε μεταξύ τους συνδυασμό. Εξυπακούεται ότι στην ανάλυση των δεδομένων θα πρέπει να ληφθεί υπόψη και ο παράγοντας του χρόνου έτσι ώστε, η διοίκηση του οργανισμού να είναι σε θέση να παράγει στατιστικές αναφορές με τα στοιχεία των συναλλαγών 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 " **CardsTransactions.txt** " και να εκτελέσετε ορισμένες επερωτήσεις για την παραγωγή χρήσιμων στατιστικών αναφορών .

Περιγραφή Αρχείου CardsTransactions.txt

Το αρχείο **CardsTransactions.txt** περιέχει 4.386.859 εγγραφές. Κάθε εγγραφή αποτελείται από 13 πεδία τα οποία διαχωρίζονται με τον χαρακτήρα "|" (pipe). Ακολουθεί η περιγραφή των πεδίων.

CardsTransactions.txt		
pid	**int**	Κωδικός κατόχου κάρτας
pname	**varchar(50)**	Όνομα κατόχου
age	**int**	Ηλικία κατόχου
gender	**char(1)**	Φύλο κατόχου
cardno	**char(16)**	Αριθμός κάρτας
card_brand	**varchar(30)**	Επωνυμία κάρτας (π.χ. Visa, Mastercard)
card_type	**varchar(20)**	Είδος κάρτας (π.χ. Debit, Credit)
tdate	**datetime**	Ημερομηνία και ώρα συναλλαγής
amount	**decimal(6,2)**	Ποσό συναλλαγής
ttc	**int**	Κωδικός είδους συναλλαγής
trans_type	**varchar(30)**	Τύπος συναλλαγής
mcc	**int**	Κωδικός πόλης στην οποία πραγματοποιήθηκε η συναλλαγή
merchant_city	**varchar(50)**	Πόλη στην οποία έλαβε χώρα η συναλλαγή

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 **CTDW (Cards Transactions Data Warehouse)**. Στη συνέχεια να δημιουργήσετε τον πίνακα **CardsTransactions** στον οποίο να φορτώσετε τα δεδομένα του αρχείου **CardsTransactions.txt** χρησιμοποιώντας την παρακάτω εντολή:

```
BULK INSERT CardsTransactions
FROM 'C:\data\CardsTransactions.txt'   !!! Προσαρμόστε το path
WITH (FIRSTROW =2, FIELDTERMINATOR='|', ROWTERMINATOR = '\n');
```

   Ακολουθήστε τον παρακάτω σύνδεσμο για να κάνετε λήψη του αρχείου **CardsTransactions.txt:**

   http://pages.aueb.gr/users/mkap/CardsTransactions.zip

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 **CardsTransactions.**
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 της τράπεζας:

1. Εμφανίστε έναν κατάλογο με την αξία των συναλλαγών ανά πόλη. Ο κατάλογος πρέπει να είναι ταξινομημένος με βάση την πόλη σε αύξουσα διάταξη.

2. Εμφανίστε έναν κατάλογο με την αξία των συναλλαγών ανά έτος και φύλο. Ο κατάλογος πρέπει να είναι ταξινομημένος με βάση το έτος σε φθίνουσα διάταξη.

3. Εμφανίστε έναν κατάλογο με τον αριθμό και την αξία των συναλλαγών ανά επωνυμία (card_brand) είδος (card_type) κάρτας.

4. Εμφανίστε έναν κατάλογο με ανάλυση της αξίας των συναλλαγών ανά τύπο συναλλαγής (trans_type) σε τριμηνιαία βάση για το έτος 2019.

5. Η διοίκηση της τράπεζας θέλει μία αναφορά που θα περιέχει τις ακόλουθες πληροφορίες για τις online συναλλαγές (trans_type='Online Transaction').
    a. Την συνολική αξία των online συναλλαγών.
    b. Την αξία των online συναλλαγών ανά έτος.
    c. Την αξία των online συναλλαγών ανά έτος και φύλο.
    d. Την αξία των online συναλλαγών ανά έτος, φύλο και ηλικία.

    Γράψτε **μια μόνο επερώτηση** (δίχως την χρήση του τελεστή UNION) 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 συναλλαγών για έναν συγκεκριμένο συνδυασμό τιμών: έτος, επωνυμία κάρτας (card_brand) και φύλο.

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.

Σημείωση: Στις απαντήσεις του δεύτερου και τρίτου ζητήματος **ΔΕΝ ΠΡΕΠΕΙ** να χρησιμοποιηθεί ο πίνακας **CardsTransactions.**

Ζήτημα τέταρτο [10 μονάδες bonus]

Σημείωση: Το συγκεκριμένο ζήτημα θα αξιολογηθεί μόνο στην περίπτωση που έχουν απαντηθεί **και τα τρία** παραπάνω ζητήματα.

Δημιουργήστε μια αναφορά (report) με το power BI (ή κάποιο αντίστοιχο εργαλείο π.χ. tableau) με τα παρακάτω:

1. Κατάλληλο γράφημα στο οποίο θα απεικονίζονται τα αποτελέσματα της επερώτησης 2 του δευτέρου ζητήματος. Δηλαδή η αξία των συναλλαγών ανά έτος και φύλο.
2. Κατάλληλο γράφημα ή/και πίνακα για την αναπαράσταση των δεδομένων του κύβου που ζητείται στο τρίτο ζήτημα.

ΠΑΡΑΔΟΤΕΑ

Τα παραδοτέα της εργασίας σας θα είναι 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

- Τον κώδικα με τις επερωτήσεις SQL του τρίτου ζητήματος.

Παραδοτέα τέταρτου ζητήματος

- Τα γραφήματα που δημιουργήσατε σε μορφή εικόνας.