Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2022-2023

Πρώτη Εργασία

Ανάθεση: 03-05-2023
Παράδοση: 16-05-2023 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.*

"Retail Database"

Στόχος της εργασίας είναι η πρακτική εφαρμογή των γνώσεων που αποκομίσατε από τις διαλέξεις του μαθήματος σχετικά με την δημιουργία ευρετηρίων και την βελτιστοποίηση των επερωτήσεων SQL. Για τον σκοπό της εργασίας θα χρησιμοποιήσετε την βάση δεδομένων **RETAILDB** η οποία περιέχει δεδομένα πωλήσεων μιας εικονικής (μη υπαρκτής) πολυεθνικής εταιρίας που εμπορεύεται έναν σημαντικό αριθμό προϊόντων. Οι βασικές οντότητες της βάσης αφορούν σε στοιχεία πελατών, προμηθευτών, προϊόντων και παραγγελιών. Τα δεδομένα των πινάκων δεν είναι πραγματικά αλλά έχουν δημιουργηθεί τυχαία για το σκοπό της συγκεκριμένης εργασίας.

Αρχικά θα δημιουργήσετε την βάση δεδομένων και θα φορτώσετε τα δεδομένα στους πίνακες, ακολουθώντας τις παρακάτω οδηγίες. Στη συνέχεια θα απαντήσετε στα ζητούμενα της εργασίας.

1. Οδηγίες για την δημιουργία της βάσης δεδομένων RETAILDB

Για να δημιουργήσετε την βάση δεδομένων και να φορτώσετε τις εγγραφές ακολουθείστε **ΠΡΟΣΕΚΤΙΚΑ** τα παρακάτω βήματα:

Βήμα 1: Κατεβάστε το αρχείο **retaildata.zip** (μέγεθος αρχείου 234 MB) από τον σύνδεσμο: http://pages.aueb.gr/users/mkap/retaildata.zip

Βήμα 2: Αποσυμπιέστε το αρχείο **retailldata.zip** στον φάκελο **C:\retaildata** (μέγεθος φακέλου 813MB)**.**

Βήμα 3: Από το περιβάλλον του Microsoft Sql Server Management Studio εκτελέστε το SQL script **"CreateRetailDB.sql"** που δημιουργεί το λογικό σχήμα της βάσης.

Βήμα 4: Εκτελέστε το SQL script **"LoadRetailData.sql"** το οποίο θα φορτώσει δεδομένα στους πίνακες της βάσης. Το συγκεκριμένο script περιέχει εντολές της μορφής:

```
BULK INSERT customers                    ! Πίνακας στον οποίο θα φορτωθούν τα δεδομένα
 FROM 'C:\retaildata\customers.txt'     ! Αρχείο το οποίο περιέχει τα δεδομένα.
WITH (FIRSTROW =2, FIELDTERMINATOR= '|', ROWTERMINATOR = '\n');
```

Παράμετροι:

FIRSTROW=2 : Η πρώτη γραμμή του αρχείου περιέχει τα ονόματα των πεδίων και αγνοείται.

FIELDTERMINATOR ='|' : Ο χαρακτήρας '|' δηλώνει το τέλος κάθε πεδίου της εγγραφής.

ROWTERMINATOR='\n' : Ο χαρακτήρας αλλαγής γραμμής δηλώνει το τέλος κάθε εγγραφής του αρχείου.

ΠΡΟΣΟΧΗ: Αν τοποθετήσετε τα δεδομένα σε φάκελο διαφορετικό από τον **'C:\retaildata'** θα πρέπει να τροποποιήσετε ανάλογα το path. Για παράδειγμα αν τοποθετήσετε τα δεδομένα στον φάκελο **'C:\DATA'** η παραπάνω εντολή πρέπει να αλλάξει ως εξής:

```
BULK INSERT customers
  FROM 'C:\DATA\customers.txt'
WITH (FIRSTROW =2, FIELDTERMINATOR= '|', ROWTERMINATOR = '\n');
```

Σημείωση: Για την διαδικασία της μαζικής εισαγωγής των δεδομένων απαιτούνται περίπου 60 sec σε ένα υπολογιστή με δίσκο SSD. Το μέγεθος της βάσης είναι 2.1 GB (data & log files).

2. Περιγραφή των πινάκων της βάσης

Ακολουθεί η περιγραφή των πινάκων και των δεδομένων της βάσης.

REGIONS: Πίνακας με τις γεωγραφικές περιοχές που δραστηριοποιείται η εταιρεία. Αριθμός Εγγραφών=5	
regionkey	Κωδικός περιοχής
region	Γεωγραφική περιοχή

NATIONS: Πίνακας με τα κράτη που δραστηριοποιείται η εταιρεία. Αριθμός Εγγραφών=25	
nationkey	Κωδικός κράτους
nation	Κράτος

CUSTOMERS: Πίνακας με τα στοιχεία των πελατών. Αριθμός εγγραφών=150.000	
custkey	Κωδικός πελάτη
cname	Όνομα πελάτη
cphone	Τηλέφωνο πελάτη
c_acctbal	Υπόλοιπο λογαριασμού πελάτη
market_segment	Εμπορικός τομέας (π.χ. BUILDING, MACHINERY, FURNITURE κ.λπ.)
nationkey	Κωδικός κράτους πελάτη.
c_comment	Σχόλια υπό την μορφή ελεύθερου κειμένου.

SUPPLIERS: Πίνακας με τα στοιχεία των προμηθευτών. Αριθμός εγγραφών=10.000	
suppkey	Κωδικός προμηθευτή
sphone	Όνομα προμηθευτή
nationkey	Τηλέφωνο προμηθευτή
s_acctbal	Υπόλοιπο λογαριασμού προμηθευτή
s_comment	Σχόλια υπό την μορφή ελεύθερου κειμένου.

PARTS: Πίνακας με τα προϊόντα που εμπορεύεται η εταιρεία. Αριθμός εγγραφών=200.000	
partkey	Κωδικός προϊόντος
ptype	Τύπος προϊόντος.
psize	Μέγεθος προϊόντος.
brand	Μάρκα προϊόντος
pname	Ονομασία προϊόντος
container	Είδος συσκευασίας.
manufacturer	Κατασκευαστής προϊόντος.
retailprice	Τιμή προϊόντος.

PARTSUPP: Πίνακας που συνδέει τα προϊόντα με τους προμηθευτές. Αριθμός εγγραφών=800.000	
partkey	Κωδικός προϊόντος
suppkey	Κωδικός προμηθευτή.
supplycost	Κόστος προμήθειας του συγκεκριμένου προϊόντος από τον συγκεκριμένο προμηθευτή.
availqty	Διαθέσιμη ποσότητα.
ps_comment	Σχόλιο υπό την μορφή ελεύθερου κειμένου.

ORDERS : Πίνακας με τις παραγγελίες των πελατών. Αριθμός εγγραφών=1.500.000	
orderkey	Κωδικός παραγγελίας
orderdate	Ημερομηνία παραγγελίας
custkey	Κωδικός πελάτη
orderpriority	Προτεραιότητα παραγγελίας
totalprice	Συνολική αξία παραγγελίας
o_comment	Σχόλια υπό την μορφή ελεύθερου κειμένου

LINEITEM: Πίνακας με τα προϊόντα των παραγγελιών. Αριθμός εγγραφών=4.423.659	
orderkey	Κωδικός παραγγελίας.
linenumber	Γραμμή παραγγελίας. Μία γραμμή παραγγελίας περιέχει μια συγκεκριμένη ποσότητα ενός συγκεκριμένου προϊόντος (partkey) το οποίο έχει προμηθευτεί από συγκεκριμένο προμηθευτή (suppkey).
discount	Συντελεστής έκπτωσης γραμμής παραγγελίας.
price	Τιμή γραμμής παραγγελίας (price = quantity*supplycost).
suppkey	Κωδικός προμηθευτή.
quantity	Ποσότητα γραμμής παραγγελίας.
returnflag	Ένδειξη επιστροφής (R=το προϊόν της συγκεκριμένης γραμμής παραγγελίας επεστράφη).
partkey	Κωδικός προϊόντος.
tax	Συντελεστής φορολογίας γραμμής παραγγελίας.
shipdate	Προγραμματισμένη ημερομηνία αποστολής.
receiptdate	Ημερομηνία παραλαβής.
commitdate	Καταληκτική ημερομηνία παράδοσης: ημερομηνία μέχρι την οποία το προϊόν της γραμμής παραγγελίας πρέπει να έχει παραδοθεί στον πελάτη. Συνήθως η καταληκτική ημερομηνία παράδοσης ορίζεται σε κάποια σύμβαση ή συμφωνία με τον πελάτη.
shipmode	Τρόπος αποστολής.
shipinstruct	Οδηγίες αποστολής
l_comment	Σχόλια υπό την μορφή ελεύθερου κειμένου.

3. Ζητούμενα εργασίας

Ακολουθούν τα ζητούμενα της εργασίας. Για την απάντηση των ζητημάτων **δεν επιτρέπεται καμία απολύτως τροποποίηση του σχήματος** εκτός φυσικά από την δημιουργία των ζητούμενων ευρετηρίων. Επίσης **απαγορεύεται** η δημιουργία και η χρήση όψεων (views).

Σε κάθε ζήτημα δεν αρκεί μόνο να παραθέσετε τα επερωτήματα σε γλώσσα SQL ή/και τις εντολές δημιουργίας των ευρετηρίων που ζητούνται. Σε κάθε περίπτωση **πρέπει να τεκμηριώσετε τις απαντήσεις σας και να παραθέσετε στοιχεία που επιβεβαιώνουν τους ισχυρισμούς σας.** Για παράδειγμα:

- Σε περιπτώσεις που ζητείται να αποδείξετε ότι ένα ευρετήριο επιταχύνει ένα ερώτημα, εκτελέστε το επερώτημα δίχως το ευρετήριο και εξετάστε το πλάνο εκτέλεσης. Αφού δημιουργήσετε το ευρετήριο εκτελέστε εκ νέου το επερώτημα και επανεξετάστε το πλάνο εκτέλεσης. Συγκρίνοντας τα δύο πλάνα μπορείτε να καταλήξετε σε συμπεράσματα σχετικά με την καταλληλότητα του ευρετηρίου.

- Σε περιπτώσεις που πρέπει να συγκρίνετε ένα η περισσότερα επερωτήματα, εκτελέστε τα όλα μαζί σε δέσμη και εξετάστε τα πλάνα εκτέλεσης. Ο SQL server δείχνει το κόστος κάθε επερωτήματος ως ποσοστό επί του συνολικού κόστους εκτέλεσης της δέσμης.

- Ενεργοποιείστε τα στατιστικά στοιχεία I/O με την εντολή: **set statistics io on.** Με τον τρόπο αυτό μπορείτε να βλέπετε κάθε φορά που εκτελείτε ένα επερώτημα πόσες σελίδες διαβάζονται από τον δίσκο ή/και από την μνήμη (buffer).

- Μπορείτε να ενεργοποιήσετε τα στατιστικά στοιχεία σχετικά με τον χρόνο εκτέλεσης του επερωτήματος με την εντολή **set statistics time on.**

- Κάθε φορά πριν την εκτέλεση ενός επερωτήματος, εκτελέστε τις παρακάτω εντολές που "καθαρίζουν" τους buffers που χρησιμοποιεί ο SQL server για την αποθήκευση των δεδομένων και των πλάνων εκτέλεσης:

  checkpoint
  dbcc dropcleanbuffers

  Με τον τρόπο αυτό διασφαλίζετε ότι, το επερώτημα που θα εκτελέσετε δεν θα χρησιμοποιήσει τυχόν σελίδες που υπάρχουν στην μνήμη από προηγούμενες εκτελέσεις του ιδίου ή/και άλλων επερωτημάτων. Σε αντίθετη περίπτωση μπορεί να οδηγηθείτε σε λάθος συμπεράσματα.

ΠΡΟΣΟΧΗ: Κάθε ζήτημα πρέπει να το αντιμετωπίσετε ανεξάρτητα από τα υπόλοιπα και να το υλοποιήσετε στο αρχικό στιγμιότυπο της βάσης. Για παράδειγμα αν θέλετε να εξετάσετε κατά πόσο ένα ευρετήριο κάνει πιο αποδοτικό ένα ερώτημα, βεβαιωθείτε ότι έχετε διαγράψει (drop index) τα ευρετήρια που έχετε δημιουργήσει για την βελτιστοποίηση άλλων επερωτημάτων.

Ζήτημα Πρώτο[20 μονάδες]

Το παρακάτω επερώτημα εμφανίζει την αξία των παραγγελιών που έχουν γίνει μέχρι και το τέλος του έτους 1993 και πρέπει να αποσταλούν στους πελάτες το πρώτο δίμηνο του έτους 1994. Οι παραγγελίες εμφανίζονται με φθίνουσα σειρά της αξίας τους.

```
Select cname, orders.orderkey,sum(price) as total
  from customers, orders, lineitem
where
  customers.custkey = orders.custkey and
  lineitem.orderkey = orders.orderkey and
  orderdate <= '1993-12-31' and
  shipdate between '1994-01-01' and '1994-02-28'
group by cname,orders.orderkey
order by total desc
```

Ζητείται να δημιουργήσετε ένα ή περισσότερα ευρετήρια που θα επιταχύνουν την εκτέλεση του επερωτήματος. Να παραθέσετε στοιχεία που να αποδεικνύουν ότι το ευρετήριο (ή τα ευρετήρια) που δημιουργήσατε επιταχύνει την εκτέλεση του επερωτήματος

Ζήτημα Δεύτερο [20 μονάδες]

Μία σημαντική απαίτηση της εταιρείας αφορά στον εντοπισμό των κατάλληλων προμηθευτών για την εξυπηρέτηση της παραγγελίας προϊόντων, συγκεκριμένου τύπου και μεγέθους.
Για παράδειγμα το ακόλουθο επερώτημα εντοπίζει τους προμηθευτές που μπορούν να προμηθεύσουν την εταιρεία με προϊόντα συγκεκριμένου τύπου “LARGE PLATED IN” και μεγέθους “31”, σε μια συγκεκριμένη περιοχή “EUROPE”, με το χαμηλότερο κόστος.

```
select  s_acctbal,sname, nation, parts.partkey, manufacturer, sphone, s_comment
   from parts, suppliers,partsupp, nations,regions
where parts.partkey = partsupp.partkey and
      suppliers.suppkey = partsupp.suppkey and
      suppliers.nationkey=nations.nationkey and
      nations.regionkey=regions.regionkey and
      psize = 31 and
      ptype = 'LARGE PLATED TIN' and
      region = 'EUROPE' and
      supplycost = ( select min(supplycost)
                        from partsupp, suppliers, nations, regions
                     where parts.partkey = partsupp.partkey and
                           suppliers.suppkey = partsupp.suppkey and
                           suppliers.nationkey=nations.nationkey and
                           nations.regionkey=regions.regionkey and
                           region='EUROPE')
order by s_acctbal desc, nation, sname, parts.partkey;
```

Ζητείται να δημιουργήσετε ένα ή περισσότερα ευρετήρια που θα επιταχύνουν την εκτέλεση του παραπάνω επερωτήματος. Να παραθέσετε στοιχεία που να τεκμηριώνουν την απάντησή σας.

Ζήτημα Τρίτο [20 μονάδες]

Ζητήθηκε από ένα προγραμματιστή να γράψει ένα επερώτημα που εμφανίζει την συνολική αξία των παραγγελιών ανά εμπορικό τομέα (market_segment) για ένα συγκεκριμένο έτος. Ο προγραμματιστής έγραψε το ακόλουθο επερώτημα:

```
select market_segment, sum(totalprice)
   from customers, orders
  where customers.custkey=orders.custkey and YEAR(orderdate) = 1996
  group by market_segment
```

Ο υπεύθυνος του τμήματος πωλήσεων δεν είναι ικανοποιημένος από την απόδοση του επερωτήματος. Καλείστε να προτείνετε τρόπους βελτιστοποίησης του επερωτήματος. Μπορείτε να πειραματιστείτε με την δημιουργία ευρετηρίων σε συνδυασμό με την συγγραφή εναλλακτικών επερωτήσεων που παράγουν το επιθυμητό αποτέλεσμα. Να παραθέσετε στοιχεία που τεκμηριώνουν την απάντησή σας.

Ζήτημα 4 [20 μονάδες]

Ακολουθούν δύο αιτήματα του τμήματος πωλήσεων σε φυσική γλώσσα:

a. Εμφάνισε έναν κατάλογο με τον κωδικό και το όνομα των προϊόντων μιας συγκεκριμένης μάρκας (π.χ. Origin) ή ενός συγκεκριμένου Κατασκευαστή (π.χ. Domeka).
b. "Εμφάνισε έναν κατάλογο με τον κωδικό και το όνομα των προϊόντων μιας συγκεκριμένης μάρκας και ενός συγκεκριμένου κατασκευαστή".

Ένας προγραμματιστής έγραψε τα παρακάτω αντίστοιχα SQL επερωτήματα:

a.
```
select distinct partkey, pname
from parts
WHERE brand='Origin' OR manufacturer='Domkapa'
```

b.
```
select partkey, pname
   from parts
where brand='Origin'

intersect

select partkey, pname
    from parts
where manufacturer='Domkapa'
```

Για να επιταχύνει την εκτέλεση και των δύο επερωτημάτων δημιούργησε τα ακόλουθα ευρετήρια:

```
create index Q4_idx1 on parts(brand) include (pname)
create index Q4_idx2 on parts(manufacturer) include (pname)
```

Ζητείται:

1. Να εξετάσετε αν τα παραπάνω ευρετήρια αξιοποιούνται στην εκτέλεση των παραπάνω επερωτημάτων. Να απαντήσετε για κάθε ξεχωριστά για κάθε επερώτημα.
2. Να προτείνετε μια καλύτερη λύση για κάθε ένα από τα παραπάνω δύο αιτήματα του τμήματος πωλήσεων. Μπορείτε να πειραματιστείτε με την δημιουργία ευρετηρίων σε συνδυασμό με την συγγραφή εναλλακτικών επερωτημάτων που παράγουν το επιθυμητό αποτέλεσμα. Η σειρά εμφάνισης των αποτελεσμάτων δεν έχει σημασία.

   Να παραθέσετε στοιχεία που τεκμηριώνουν τις απαντήσεις σας.

Ζήτημα Πεμπτο [24 μονάδες]

1. Να διατυπώσετε δύο ερωτήματα σε φυσική γλώσσα και στη συνέχεια να γράψετε εντολές σε γλώσσα SQL ώστε να απαντηθούν τα ερωτήματα που διατυπώσατε.

2. Να δημιουργήσετε κατάλληλα ευρετήρια που επιταχύνουν την εκτέλεση των επερωτημάτων σας. Να παραθέσετε τις εντολές δημιουργίας των ευρετηρίων, καθώς επίσης και στοιχεία που να αποδεικνύουν ότι τα ευρετήρια που δημιουργήσατε επιταχύνουν την εκτέλεση των επερωτημάτων.

   Φροντίστε τα επερωτήματα που θα γράψετε να δίνουν χρήσιμες πληροφορίες, να μην είναι εντελώς απλοϊκά και να μην χρησιμοποιούν μόνο ευρετήρια που δημιουργήσατε για να απαντήσετε τα προηγούμενα ζητήματα.

ΔΙΑΓΡΑΜΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΛΟΓΙΚΟΥ ΣΧΗΜΑΤΟΣ ΤΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ RETAILDB