Συστήματα Διαχείρισης και Ανάλυσης Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2023-2024

Πρώτη Εργασία

Ανάθεση: 26-04-2024
Παράδοση: 13-05-2024 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 AM είναι ο αριθμός μητρώου σας. π.χ. "3210001.pdf").*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*

"StackOverflow 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 του ιστότοπου StackOverFlow, η οποία περιέχει ερωτήσεις και απαντήσεις για ένα ευρύ φάσμα θεμάτων σχετικά με τον προγραμματισμό υπολογιστών. Πρόκειται ουσιαστικά για μια απλοποιημένη έκδοση η οποία υλοποιεί μέρος του λογικού σχήματος της πραγματικής βάσης του ιστότοπου και περιέχει δοκιμαστικά δεδομένα για τα έτη 2008 έως και 2010. Για περισσότερες πληροφορίες σχετικά με τη βάση δεδομένων StackOverflow ακολουθείστε τους παρακάτω συνδέσμους:

https://en.wikipedia.org/wiki/Stack_Overflow
https://stackoverflow.com/

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 StackOverFlow.

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

Βήμα 1: Κατεβάστε το αρχείο **sodata.zip** (μέγεθος αρχείου 1.12 GB) από τον παρακάτω σύνδεσμο:

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

Βήμα 2: Αποσυμπιέστε το αρχείο **sodata.zip** στον φάκελο **C:\sodata** (μέγεθος φακέλου 5GB)**.**

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

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

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

Παράμετροι:

DATAFILETYPE = 'widechar' : Το αρχείο περιέχει δεδομένα σε μορφή unicode.

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:\sodata'** 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 Users
  FROM 'C:\DATA\Users.txt'
WITH (DATAFILETYPE = 'widechar', FIRSTROW =2, FIELDTERMINATOR= '|', ROWTERMINATOR = '\n');
```

Σημείωση: Για την διαδικασία της μαζικής εισαγωγής των δεδομένων απαιτούνται περίπου 7 λεπτά σε ένα υπολογιστή με δίσκο SSD και περίπου 11 λεπτά σε έναν υπολογιστή με συμβατικό σκληρό δίσκο. Το μέγεθος της βάσης είναι περίπου 10 GB (data & log files).

2. Περιγραφή των πινάκων της βάσης

Ακολουθεί η περιγραφή των πινάκων και των δεδομένων της βάσης.

PostTypes: Πίνακας με τους τύπους των αναρτήσεων. Αριθμός Εγγραφών=8	
postTypeId	Κωδικός
PostTypeName	Τύπος Ανάρτησης (π.χ. Question, Answer, Wiki κ.λπ.).

VoteTypes: Πίνακας με τα είδη των ψήφων. Αριθμός Εγγραφών=8	
VoteTypeId	Κωδικός
VoteTypeName	Είδος (π.χ. UpVote, DownVote κ.λπ.).

Users: Πίνακας με τα στοιχεία των χρηστών. Αριθμός εγγραφών=299397	
UserId	Κωδικός Χρήστη
AboutMe	Πληροφορίες που καταχωρεί ο χρήστης στην ενότητα "AboutMe" του προφίλ του.
CreationDate	Ημερομηνία δημιουργίας του χρήστη.
displayName	Όνομα Χρήστη.
DownVotes	Αριθμός αρνητικών ψήφων που έχει δώσει ο χρήστης σε αναρτήσεις άλλων χρηστών.
LastAccessDate	Ημερομηνία της πιο πρόσφατης επίσκεψης του χρήστη στον ιστότοπο.
UserLocation	Τοποθεσία Χρήστη.
Reputation	Φήμη. Μέτρηση η οποία εκφράζει τον βαθμό εμπιστοσύνης των μελών της κοινότητας στον συγκεκριμένο χρήστη. Αντικατοπτρίζει, ως ένα βαθμό, την εξοικείωση του χρήστη με τον ιστότοπο, την εξειδίκευσή του πάνω σε συγκεκριμένα ζητήματα και γενικότερα την συνεισφορά του στην κοινότητα.
UpVotes	Αριθμός θετικών ψήφων που έχει δώσει ο χρήστης σε αναρτήσεις άλλων χρηστών.
ProfileViews	Αριθμός προβολών του προφίλ του χρήστη
WebsiteUrl	Διεύθυνση Ιστοσελίδας του Χρήστη

Badges: Πίνακας με τα σήματα (διακριτικά) των χρηστών (π.χ. Guru, Teacher, Popular Question, Nice Answer κ.λπ.). Τα σήματα είναι "ψήγματα" ψηφιακής αίσθησης που κερδίζουν οι χρήστες για σχεδόν κάθε είδος δραστηριότητας στον ιστότοπο stackOverflow. Αριθμός Εγγραφρών=1102017	
Bid	Μοναδικός Κωδικός Εγγραφής
Bname	Ονομασία σήματος
UserId	Κωδικός χρήστη
Bdate	Ημερομηνία απόδοσης του διακριτικού

Posts: Πίνακας με τις αναρτήσεις των χρηστών. Αριθμός εγγραφών=3729155	
PostrId	Κωδικός ανάρτησης
AcceptedAnswerID	Κωδικός αποδεκτής απάντησης. Το πεδίο έχει τιμή μόνο για αναρτήσεις οι οποίες είναι ερωτήσεις (Question). Διαφορετικά έχει την τιμή NULL.
AnsweCount	Συνολικός αριθμός απαντήσεων για την συγκεκριμένη ανάρτηση εφόσον είναι ερώτηση. Για αναρτήσεις άλλου τύπου το πεδίο έχει την τιμή 0 (μηδέν).
Body	Κείμενο ανάρτησης
CommentCount	Δηλώνει τον αριθμό των σχολίων (comments) των χρηστών για την συγκεκριμένη ανάρτηση.
CreationDate	Ημερομηνία δημιουργίας της ανάρτησης.
LastActivityDate	Δηλώνει την πιο πρόσφατη ημερομηνία κατά την οποία έλαβε χώρα μια δραστηριότητα σχετική με την ανάρτηση (π.χ. δημιουργία, τροποποίηση κ.λπ.).
FavoriteCount	Αριθμός που δείχνει πόσο δημοφιλής είναι η ανάρτηση.
LastEditDate	Ημερομηνία πιο πρόσφατης τροποποίησης της ανάρτησης.
LastEditorDisplayName	Όνομα χρήστη που έκανε την πιο πρόσφατη τροποποίηση.
OwnerUserId	Ο κωδικός του Χρήστη που έκανε την ανάρτηση.
ParentId	Το πεδίο έχει τιμή μόνο για αναρτήσεις οι οποίες αποτελούν απάντηση κάποιας ερώτησης (Answer). Περιέχει τον κωδικό της ερώτησης στην οποία αναφέρεται η απάντηση και ουσιαστικά συνδέει κάθε ερώτηση με τις απαντήσεις της.
PostTypeId	Κωδικός που δηλώνει τον τύπο της ανάρτησης.
Score	Το Score της ανάρτησης το οποίο υπολογίζεται ως η διαφορά ανάμεσα στις θετικές και τις αρνητικές ψήφους που έχει λάβει η ανάρτηση.
Title	Ο τίτλος της ανάρτησης.
ViewCount	Ο αριθμός επισκέψεων/εμφανίσεων της ανάρτησης.

Comments: Πίνακας με τα σχόλια των χρηστών. Αριθμός Εγγραφρών=3874841	
Cid	Μοναδικός κωδικός εγγραφής.
CreationDate	Ημερομηνία δημιουργίας σχολίου.
PostId	Ο κωδικός της ανάρτησης στην οποία αναφέρεται το σχόλιο.
Score	Το σύνολο των θετικών ψήφων που έχει λάβει το σχόλιο.
Text	Το κείμενο του σχολίου.
UserId	Ο κωδικός του χρήστη που δημιούργησε το σχόλιο.

postLinks: Πίνακας για την διασύνδεση σχετικών αναρτήσεων. Αριθμός Εγγραφρών=135391	
id	Μοναδικός κωδικός εγγραφής.
CreationDate	Ημερομηνία δημιουργίας της σύνδεσης.
PostId	Κωδικός ανάρτησης.
RelatedPostId	Κωδικός σχετικής ανάρτησης.

Tags: Πίνακας με ετικέτες (λέξεις κλειδιά) που αποδίδουν οι χρήστες στις αναρτήσεις τους. Αριθμός Εγγραφρών=24218.	
Tagid	Κωδικός ετικέτας.
tagName	ετικέτα.

PostTags: Πίνακας που συσχετίζει τις ετικέτες με τις αναρτήσεις. Αριθμός Εγγραφρών=3136519.	
PostId	Κωδικός ανάρτησης.
Tagid	Κωδικός ετικέτας.

Votes: Πίνακας με τις ψήφους των χρηστών. Αριθμός Εγγραφρών=8079643	
VoteId	Μοναδικός κωδικός εγγραφής.
PostId	Κωδικός ανάρτησης στην οποία δίνεται η ψήφος.
UserId	Ο κωδικός του χρήστη που έδωσε την ψήφο.
BountyAmount	Μονάδες πριμοδότησης.
VoteTypeId	Κωδικός που δηλώνει το είδος της ψήφου.
CreationDate	Ημερομηνία απόδοσης της ψήφου.

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). Αν για κάποιο ζήτημα **δεν ισχύουν** οι συγκεκριμένοι περιορισμοί τότε αυτό θα δηλώνεται ρητά στην εκφώνηση του ζητήματος.

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 τα ονόματα και τον αριθμό προβολών του προφίλ των χρηστών που εγγράφηκαν στην υπηρεσία το έτος 2010. Τα αποτελέσματα εμφανίζονται ταξινομημένα με βάση την ημερομηνία εγγραφής και τον αριθμό των προβολών του προφίλ των χρηστών.

```
SELECT displayName, profileviews
   FROM users
 WHERE YEAR(CreationDate)=2010
 ORDER BY creationDate, profileViews
```

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 Δεύτερο [20 μονάδες]

Το παρακάτω επερώτημα εμφανίζει τα στοιχεία των χρηστών που έχουν αναρτήσει τουλάχιστον μία απάντηση αλλά δεν έχουν αναρτήσει κάποια ερώτηση.

```
SELECT users.*
  FROM users, posts, postTypes
WHERE users.userid=posts.ownerUserid and
      posts.postTypeid=PostTypes.postTypeid and postTypeName='Answer'

EXCEPT

SELECT users.*
  FROM users, posts, postTypes
WHERE users.userid=posts.ownerUserid and
      posts.postTypeid=PostTypes.postTypeid and postTypeName='Question'
```

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. Η σειρά εμφάνισης των εγγραφών στην έξοδο δεν έχει σημασία. **Να παραθέσετε στοιχεία που τεκμηριώνουν την απάντησή σας.**

Ζήτημα Τρίτο [20 μονάδες]

Το παρακάτω επερώτημα εμφανίζει τον αριθμό των θετικών ψήφων που έχει πάρει ένας χρήστης (στην συγκεκριμένη περίπτωση ο χρήστης `'Dominik Weber'`) για κάθε λέξη κλειδί (ετικέτα) που έχει αποδώσει στις αναρτήσεις του.

```
SELECT TagName, COUNT(*) AS UpVotes
FROM Tags
    INNER JOIN PostTags ON PostTags.TagId = Tags.tagid
    INNER JOIN Posts ON Posts.ParentId = PostTags.PostId
       INNER JOIN Users ON Posts.OwnerUserId=Users.Userid
    INNER JOIN Votes ON Votes.PostId = Posts.postId
       INNER JOIN VoteTypes ON Votes.voteTypeId=VoteTypes.VoteTypeID
WHERE
    VoteTypeName='UpVote' AND  Users.displayName = 'Dominik Weber'
GROUP BY TagName
ORDER BY UpVotes DESC
```

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

Ζήτημα Τέταρτο [20 μονάδες]

Το επερώτημα που ακολουθεί εμφανίζει τους κωδικούς 100 (top 100) χρηστών που σπάνια ψηφίζουν θετικά σε σύγκριση με τον εκτιμώμενο αριθμό θετικών ψήφων που έλαβαν. Οι θετικές ψήφοι που λαμβάνει ένας χρήστης αντικατοπτρίζονται στο πεδίο Reputation. Για κάθε θετική ψήφο που λαμβάνει μια ανάρτηση του χρήστη το πεδίο reputation αυξάνεται κατά 10 μονάδες.

```
select top 100
  userid,
  round((100.0 * (Reputation/10)) / (Upvotes+1), 2) as [Ratio %],
  Reputation,
  UpVotes,
  DownVotes
from Users
where Reputation > 1000
  and Upvotes > 100
order by [Ratio %] desc
```

Να προτείνετε δύο εναλλακτικά ευρετήρια τα οποία θεωρείτε ότι επιταχύνουν την εκτέλεση του επερωτήματος και να επιλέξετε το πλέον κατάλληλο. Να αιτιολογήσετε την επιλογή σας και να παραθέσετε στοιχεία που να αποδεικνύουν ότι το προτεινόμενο ευρετήριο επιταχύνει την εκτέλεση του επερωτήματος.

Ζήτημα 5 [20 μονάδες]

1. Να διατυπώσετε ένα ερώτημα σε φυσική γλώσσα και στη συνέχεια να γράψετε την αντίστοιχη εντολή σε γλώσσα SQL ώστε να απαντηθεί το ερώτημα που διατυπώσατε.

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

Φροντίστε το επερώτημα που θα γράψετε να δίνει χρήσιμες πληροφορίες, να μην είναι εντελώς απλοϊκό και να μην χρησιμοποιεί μόνο ευρετήρια που δημιουργήσατε για να απαντήσετε τα προηγούμενα ζητήματα.