Σχεδιασμός Βάσεων Δεδομένων
Διδάσκων: Ιωάννης Κωτίδης

Εαρινό εξάμηνο 2020-2021

Πρώτη Εργασία

Ανάθεση: **30-04-2021**
Παράδοση: **18-05-2021 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. "3170001.pdf").*
- *Πιθανή αντιγραφή θα τιμωρείται με μηδενισμό όλων των εμπλεκομένων.*

"Βάση Δεδομένων Κινηματογραφικών Ταινιών"

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 την πρακτική σας εξάσκηση θα χρησιμοποιήσετε μια βάση δεδομένων και το DBMS MICROSOFT SQL SERVER. Η βάση δεδομένων περιέχει πληροφορίες για κινηματογραφικές ταινίες.
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.

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

Βήμα 1: Από το περιβάλλον του Microsoft Sql Server Management Studio δημιουργείστε μια βάση δεδομένων με όνομα **MOVIE**.

Βήμα 2: Εκτελέστε το SQL script **"CreateMovieSchema.sql"** που δημιουργεί το λογικό σχήμα της βάσης. Πριν εκτελέσετε το script βεβαιωθείτε ότι η τρέχουσα βάση δεδομένων είναι η βάση **MOVIE** που δημιουργήσατε στο βήμα 1.

Βήμα 3: Εκτελέστε το SQL script **"LoadMovieData.sql"** το οποίο θα φορτώσει δεδομένα στους πίνακες της βάσης. Το συγκεκριμένο script περιέχει εντολές της μορφής:

BULK INSERT actors ! Πίνακας στον οποίο θα φορτωθούν τα δεδομένα

FROM 'C:\ movieData\actors.txt' ! Αρχείο που 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:\movieData '** 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 actors

FROM 'C:\DATA\actors.txt'

WITH (FIRSTROW =2, FIELDTERMINATOR= '|', ROWTERMINATOR = '\n');

Τα SQL scripts **"CreateMovieSchema.sql"** και **"LoadMovieData.sql"** καθώς επίσης και τα αρχεία με τα δεδομένα που θα φορτωθούν στους πίνακες της βάσης θα τα βρείτε στο αρχείο "**movieData.zip".**

ΣΗΜΕΙΩΣΗ: Τα περιεχόμενα των παραπάνω scripts παρατίθενται στο παράρτημα που ακολουθεί στο τέλος της εργασίας.

2. Περιγραφή των πινάκων της βάσης

Ακολουθεί η περιγραφή των πινάκων και των δεδομένων της βάσης.

ACTORS: Πίνακας με στοιχεία ηθοποιών. Αριθμός Εγγραφών=817062	
aid	Κωδικός ηθοποιού
firstName	Όνομα ηθοποιού
lastName	Επώνυμο ηθοποιού
gender	Φύλο (F=female, M=Male)

DIRECTORS: Πίνακας με στοιχεία σκηνοθετών. Αριθμός εγγραφών=86880	
did	Κωδικός σκηνοθέτη
firstName	Όνομα σκηνοθέτη
lastName	Επώνυμο σκηνοθέτη

MOVIES: πίνακας με τα στοιχεία των ταινιών. Αριθμός εγγραφών=347796.	
mid	Κωδικός ταινίας
title	Τίτλος ταινίας
pyear	Έτος κυκλοφορίας
mrank	Κατάταξη [1.0 - 9.9]

MOVIE_DIRECTORS: Πίνακας που συνδέει τις ταινίες με τους σκηνοθέτες. Αριθμός εγγραφών=319117	
mid	Κωδικός ταινίας
did	Κωδικός σκηνοθέτη

MOVIES_GENRE: Πίνακας την κατηγορία/κατηγορίες κάθε ταινίας. Αριθμός εγγραφών=387390	
mid	Κωδικός ταινίας
genre	Κατηγορία

ROLES: Πίνακας που συνδέει τις ταινίες με τους ηθοποιούς. Αριθμός εγγραφών=1093499	
mid	Κωδικός ταινίας
aid	Κωδικός ηθοποιού
a_role	Ο ρόλος του ηθοποιού στην συγκεκριμένη ταινία

USERS: Πίνακας με τα στοιχεία των χρηστών. Αριθμός εγγραφών=6039. **ΤΑ ΣΤΟΙΧΕΙΑ ΤΩΝ ΧΡΗΣΤΩΝ ΔΕΝ ΕΙΝΑΙ ΠΡΑΓΜΑΤΙΚΑ.**	
userid	Κωδικός χρήστη
uname	Ονοματεπώνυμο χρήστη
gender	Φύλο (F=female, M=Male)
age	Ηλικία [18-56]

USER_MOVIES: Πίνακας με στοιχεία αξιολόγησης των ταινιών. Αριθμός εγγραφών=996159.	
mid	Κωδικός ταινίας
userid	Κωδικός χρήστη
rating	Βαθμός αξιολόγησης [1-5]

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).

- 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 [3 μονάδες]

1. Να δημιουργήσετε **ένα μόνο** ευρετήριο που να επιταχύνει την εκτέλεση και των τριών παρακάτω επερωτημάτων. Να παραθέσετε την εντολή δημιουργίας του ευρετηρίου και να αιτιολογήσετε την επιλογή σας.

```
select title from movies where pyear between 1990 and 2000

select pyear, title from movies where pyear between 1990 and 2000

select title, pyear  from movies where pyear between 1990 and 2000
order by pyear, title
```

2. Τα παρακάτω δύο επερωτήματα εμφανίζουν τον συνολικό αριθμό των αξιολογήσεων ανά κωδικό ταινίας και κωδικό χρήστη αντίστοιχα. Η συχνότητα εκτέλεσης των επερωτημάτων είναι η ίδια. Θέλετε και τα δύο επερωτήματα να εκτελούνται με τρόπο αποδοτικό. Έστω ότι μπορείτε να δημιουργήσετε **ένα μόνο** ευρετήριο, ποιο ευρετήριο θα επιλέγατε να δημιουργήσετε; Να παραθέσετε την εντολή δημιουργίας του ευρετηρίου και να αιτιολογήσετε την επιλογή σας.

```
select mid, count(rating)
  from user_movies group by mid order by mid

select userid, count(rating)
  from user_movies group by userid order by userid
```

Για κάθε μια από τις παραπάνω περιπτώσεις να παραθέσετε στοιχεία που να αποδεικνύουν ότι τα ευρετήρια που δημιουργήσατε επιταχύνουν την εκτέλεση των επερωτημάτων.

Ζήτημα Δεύτερο [3 μονάδες]

1. Το παρακάτω ερώτημα εμφανίζει τους τίτλους των ταινιών που ανήκουν στην κατηγορία 'Adventure' ή στην κατηγορία 'Action' ή και στις δύο κατηγορίες:

```
select  title
  from movies, movies_genre
 where movies.mid=movies_genre.mid and genre='Adventure'

 UNION

select  title
  from movies, movies_genre
 where movies.mid=movies_genre.mid and genre ='Action'
```

Να γράψετε ένα αποδοτικότερο επερώτημα σε SQL το οποίο να δίνει τα ίδια αποτελέσματα με το παραπάνω (η σειρά των αποτελεσμάτων δεν έχει σημασία). Να δημιουργήσετε κατάλληλο ευρετήριο ή ευρετήρια που να επιταχύνουν την εκτέλεση του επερωτήματος που γράψατε.

2. Θεωρήστε το παρακάτω ερώτημα σε φυσική γλώσσα:

"Εμφάνισε τους τίτλους των ταινιών στις οποίες συμμετέχουν μόνο άντρες ηθοποιοί".

Να γράψετε τουλάχιστον δύο διαφορετικά επερωτήματα σε SQL που να απαντούν στο παραπάνω ερώτημα. Στη συνέχεια να δημιουργήσετε κατάλληλα ευρετήρια που να επιταχύνουν την εκτέλεση των επερωτημάτων. Το ζητούμενο είναι να καταλήξετε σε ένα επερώτημα, το οποίο σε συνδυασμό με κατάλληλα ευρετήρια, θεωρείτε ότι είναι το πλέον αποδοτικό (μικρότερο κόστος εκτέλεσης).

Να παραθέσετε στοιχεία που επιβεβαιώνουν την ορθότητα των ισχυρισμών σας.

Ζήτημα Τρίτο [4 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.

ΠΑΡΑΡΤΗΜΑ Α

CreateMovieSchema.sql

```
create table actors
(aid int primary key,
 firstName varchar(100),
 lastname varchar(100),
 gender char(1)
 );

 create table directors
(did int primary key,
 firstName varchar(100),
 lastname varchar(100)
 );

create table movies
 (mid int primary key,
  title varchar(200),
  pyear int,
  mrank decimal(2,1)
  );

 create table movie_directors
 ( did int foreign key references directors(did),
   mid int foreign key references movies(mid)
   primary key(mid, did),
 );

 create table roles
 ( aid int foreign key references actors(aid),
   mid int foreign key references movies(mid),
   a_role varchar(100),
   primary key (mid,aid)
  );

  create table movies_genre
  (mid int foreign key references movies(mid),
   genre varchar(100),

   );

create table users
(userid int primary key,
 uname varchar(50),
 gender char(1),
 age int
 );

 create table user_movies
 (userid int foreign key references users(userid),
  mid int foreign key references movies(mid),
  rating int,
  primary key (mid,userid)
  );
```

loadMovieData.sql

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

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

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

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

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

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

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

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