"Internet Movies Database"

Στόχος της άσκησης είναι η δημιουργία μιας δεδομένων με πληροφορίες κινηματογραφικών ταινιών και στη συνέχεια η συγγραφή, εκτέλεση και βελτιστοποίηση SQL επερωτήσεων. 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 **IMDB**.

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

Βήμα 3: Εκτελέστε το SQL script **"LoadImdbData.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:\ imdbData\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:\imdbData '** 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 **"CreateImdbSchema.sql"** και **"LoadImdbData.sql"** καθώς επίσης και τα αρχεία με τα δεδομένα που θα φορτωθούν στους πίνακες της βάσης θα τα βρείτε στο αρχείο "**imdbData.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]

Ασκήσεις

1. Να δημιουργήσετε κατάλληλα ευρετήρια που να επιταχύνει την εκτέλεση των παρακάτω επερωτήσεων. Για κάθε επερώτηση να παραθέσετε την εντολή δημιουργίας του ευρετηρίου και να αιτιολογήσετε την επιλογή σας.

   a) `select title from movies where pyear between 1995 and 2005`

   b) `select pyear, title from movies where pyear between 1995 and 2005`

   c) `select title, pyear from movies where pyear between 1995 and 2005 order by pyear, title`

2. Να γράψετε τις παρακάτω SQL επερωτήσεις. Στη συνέχεια για κάθε επερώτηση να δημιουργήσετε κατάλληλα ευρετήρια (ένα ή περισσότερα) που να επιταχύνουν την εκτέλεσή της. Να παραθέσετε την εντολή δημιουργίας του ευρετηρίου και να αιτιολογήσετε την επιλογή σας.

   a) Εμφανίστε ένα κατάλογο με τον τίτλο και το έτος κυκλοφορίας των ταινιών τις οποίες έχει σκηνοθετήσει ο σκηνοθέτης με επώνυμο «Zygadlo».

   b) Εμφανίστε έναν κατάλογο με τον τίτλο, το έτος κυκλοφορίας και την κατάταξη (mrank) των ταινιών που ανήκουν στην κατηγορία “Comedy” και έχουν κατάταξη μεγαλύτερη του 7.

   c) Εμφανίστε ένα κατάλογο με τον τίτλο, και την κατάταξη (mrank) των ταινιών που κυκλοφόρησαν το 2000 και έχουν κατάταξη μεγαλύτερη του 5.

3. Να γράψετε τουλάχιστον δύο διαφορετικά επερωτήματα σε SQL που να εμφανίζουν τους τίτλους των ταινιών στις οποίες συμμετέχουν μόνο άντρες ηθοποιοί. 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 αιτιολογήσετε την επιλογή σας.

4. Διατυπώστε δύο ερωτήματα σε φυσική γλώσσα και στην συνέχεια γράψτε κατάλληλες επερωτήσεις SQL που απαντούν στα ερωτήματα. Δημιουργείστε κατάλληλα ευρετήρια που να επιταχύνουν την εκτέλεση των επερωτήσεων.

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 επερωτήσεων.

- Κάθε φορά πριν την εκτέλεση μίας επερώτησης, 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, η επερώτηση που θα εκτελέσετε δεν θα χρησιμοποιήσει τυχόν σελίδες που υπάρχουν στην μνήμη από προηγούμενες εκτελέσεις της ιδίας ή/και άλλων επερωτήσεων. Σε αντίθετη περίπτωση μπορεί να οδηγηθείτε σε λάθος συμπεράσματα.