Project - Βάσεις Δεδομένων

Προθεσμία: 15/6/2023

Σκοπός

Στο project θα ασχοληθούμε με την ανάλυση και οπτικοποίηση των δεδομένων της βάσης Movielens την οποία και εξετάσαμε στα πλαίσια των προηγούμενων εργασιών. Η συγκεκριμένη βάση δεδομένων περιέχει πληροφορίες για ταινίες, τους συντελεστές τους, και τις αξιολογήσεις των ταινιών από χρήστες.

Ομάδες

Για την ολοκλήρωση του project θα πρέπει να σχηματίσετε ομάδες των 2 ή 3 ατόμων και όχι παραπάνω ή λιγότερα άτομα ανά ομάδα.

Δεδομένα

Μπορείτε χρησιμοποιήσετε την βάση που φτιάξατε στις προηγούμενες ασκήσεις. Σε περίπτωση που δεν έχετε ολοκληρώσει τις προηγούμενες ασκήσεις, θα πρέπει να τρέξετε τα βήματα που περιγράφονται στις Ασκήσεις 1, 2.

Ζητούμενα Άσκησης

Ερωτήματα συνάθροισης με ομαδοποίηση

Να απαντήσετε στα ακόλουθα ερωτήματα χρησιμοποιώντας συναρτήσεις συνάθροισης με ομαδοποίηση σε SQL γλώσσα. Να χρησιμοποιηθούν οι τελεστές `Group By` και `Having` (όπου χρειάζεται), ενώ να γίνει και η κατάλληλη ταξινόμηση του αποτελέσματος.

1. Αριθμός ταινιών ανά έτος `(year, movies_per_year)` για ταινίες με συνολικό budget μεγαλύτερο από 1,000,000.
2. Αριθμός ταινιών ανά είδος `(genre, movies_per_genre)` για ταινίες που έχουν συνολικό budget μεγαλύτερο από 1,000,000 ή διάρκεια μεγαλύτερη από 2 ώρες.
3. Αριθμός ταινιών ανά είδος και ανά έτος `(genre, year, movies_per_gy)`.
4. Για τον αγαπημένο σας ηθοποιό, το σύνολο των εσόδων (revenue) για τις ταινίες στις οποίες έχει συμμετάσχει ανά έτος `(year, revenues_per_year)`.
5. Το υψηλότερο budget ταινίας ανά έτος `(year, max_budget)`, όταν το `budget` αυτό δεν είναι μηδενικό.
6. Τις συλλογές του πίνακα `Collection` που αναφέρονται σε τριλογίες, δηλαδή η συλλογή έχει ακριβώς 3 ταινίες `(trilogy_name)`.
7. Για κάθε χρήστη του πίνακα Ratings, να επιστραφούν η μέση βαθμολογία του χρήστη και ο αριθμός των βαθμολογιών του `(avg_rating, rating_count)`.

Ερωτήματα Κατάταξης

8. Οι 10 ταινίες με το υψηλότερο budget `(movie_title, budget)`. Σε περίπτωση που έχουμε ισοβαθμία μεταξύ δύο ή περισσοτέρων ταινιών για την θέση 10 και μετά, να επιστραφεί μία από τις δύο.
   (hint: Να χρησιμοποιηθεί ο τελεστής `ORDER BY` σε συνδυασμό με τον τελεστή `TOP` της Microsoft SQL)
9. Χρησιμοποιώντας το ερώτημα 5, βρείτε με εμφώλευση την ταινία (ταινίες) με το μεγαλύτερο budget ανά χρονιά `(year, movies_with_max_revenues)`, έχοντας ταξινόμηση ως προς το έτος και το όνομα της ταινίας.

Οπτικοποίηση Στατιστικών

Σε αυτό το μέρος θα πρέπει να οπτικοποιήσετε τα παρακάτω στατιστικά με χρήση SQL και Python (μέσω σύνδεσης στην βάση σας). Συγκεκριμένα οπτικοποιήστε όλα τα προηγούμενα ερωτήματα, **εκτός των ερωτημάτων 6, 8, 9**. Όπου χρειάζεται χρησιμοποιείστε είτε δισδιάστατο, είτε τρισδιάστατο γράφημα, είτε γράφημα διασποράς (scatter plot).

Συστάσεις Ταινιών

Θεωρείστε ότι ένα ζεύγος ταινιών είναι δημοφιλές όταν υπάρχουν πάνω από 10 χρήστες που έχουν βαθμολογήσει και τις 2 ταινίες με άνω του 4 βαθμολογία. Η δημοφιλία ενός τέτοιου ζεύγους προσδιορίζεται από τον αριθμό των συγκεκριμένων χρηστών.

10. Προσδιορίστε το ερώτημα το οποίο θα χρησιμοποιηθεί για να φτιάξουμε ένα materialized view (υλοποιημένη όψη) με το όνομα `Popular_Movie_Pairs` που περιέχει όλα τα ζεύγη ταινιών και την δημοφιλία τους.

Παραδοτέο

Θα πρέπει να παραδώσετε τα ακόλουθα:

A. Ένα αρχείο sql που να περιέχει τις εντολές που χρησιμοποιήσατε για να υπολογίσετε όλα τα παραπάνω (Ερωτήματα 1-10).
B. Το αρχείο Python που πραγματοποιεί την σύνδεση με την βάση (με κωδικούς examiner), εκτελεί το ερώτημα, και οπτικοποιεί τα δεδομένα.
C. Ένα pdf το οποίο περιέχει τις οπτικοποιήσεις μαζί με μία σύντομη επεξήγηση της πληροφορίας που απεικονίζεται.

Χρήσιμα links:

Python

- Azure Connection:
https://docs.microsoft.com/en-us/azure/postgresql/connect-python
- Matplotlib:
https://matplotlib.org/
- Transpose Matrix:
https://numpy.org/doc/stable/reference/generated/numpy.transpose.html

SQL Server

- Εξαγωγή ημερομηνίας Date:
https://learn.microsoft.com/en-us/sql/t-sql/functions/getdate-transact-sql?view=sql-server-ver16
- Order by:
https://learn.microsoft.com/en-us/sql/t-sql/queries/select-order-by-clause-transact-sql?view=sql-server-ver16
- Εντολή Select:
https://learn.microsoft.com/en-us/sql/t-sql/queries/select-transact-sql?view=sql-server-ver16

Τελικά Παραδοτέα

- Δημιουργήστε ένα .txt αρχείο στο οποίο θα αναγράφονται το endpoint του Azure instance σας (Server name στο Overview tab του Azure), το όνομα της βάσης σας και το username και το password ενός χρήστη με read-only δικαιώματα, ώστε να μπορούμε να δούμε τους πίνακες της βάσης σας. Το .txt αρχείο θα πρέπει να έχει την παρακάτω μορφή:

  Endpoint: <name_of_the_endpoint>
  Username: <username>
  Password: <password>
  Database: <name_of_the_database>

- Βάλτε σε ένα φάκελο
  - το txt αρχείο,
  - τα παραδοτέα που σχετίζονται με sql ερωτήματα,
  - τον κώδικα σε python, καθώς και
  - τις απεικονίσεις μαζί με τις επεξηγήσεις τους.
- Το όνομα του φακέλου πρέπει να αποτελείται από τους αριθμούς μητρώου σας χωρισμένους με παύλα, δηλαδή αριθμός_μητρώου_1- αριθμός_μητρώου_2-αριθμός_μητρώου_3. Δημιουργήστε ένα .zip αρχείο αυτού του φακέλου, το οποίο θα έχει το ίδιο όνομα με τον φάκελο.
- Κάντε υποβολή το .zip αρχείο στο eclass στην ενότητα Εργασίες / Project.