2η Εργασία: MovieLens Tables

Προθεσμία: 24/4/2023

Σκοπός:

Σε αυτή την εργασία θα δημιουργήσετε την βάση δεδομένων ταινιών *Movielens* (https://movielens.org). Η συγκεκριμένη βάση δεδομένων περιέχει πληροφορίες για ταινίες, τους συντελεστές τους, και τις αξιολογήσεις τους. Για να ορίσουμε το σχήμα της βάσης θα βασιστούμε στους τύπους των δεδομένων εισόδου που βρίσκονται στα αρχεία csv. Για την εισαγωγή των δεδομένων θα πρέπει να χρησιμοποιηθεί το SQL Server extension του Azure Data Studio

Περιγραφή Δεδομένων:

Τα δεδομένα της άσκησης βρίσκονται διαθέσιμα στον εξωτερικό σύνδεσμο:

https://drive.google.com/file/d/1BVnSB9uePCjjSoDvEPKtDdfgg3v7Z-Nz/view?usp=share_link

Η αρχική μορφή των δεδομένων περιέχει πληροφορίες για 45.000 ταινίες από τη βάση MovieLens, λέξεις-κλειδιά της πλοκής των ταινιών της βάσης *MovieLens*, πληροφορίες για τους συντελεστές των ταινιών καθώς και αξιολογήσεις από χρήστες σε ταινίες. Η αρχική μορφή των δεδομένων μπορεί να βρεθεί στον σύνδεσμο:

https://www.kaggle.com/rounakbanik/the-movies-dataset?select=keywords.csv

Στα δεδομένα που θα δείτε, έχει γίνει μία προεπεξεργασία των αρχικών csv αρχείων με σκοπό την απαλοιφή των JSon κελιών. Έτσι, π.χ., το αρχείο movies_metadata έχει χωριστεί σε περισσότερα του ενός csv αρχεία→ movie, belongsToCollection, collection, κτλ.

Τα csv αρχεία που σας δίνονται είναι τα ακόλουθα:

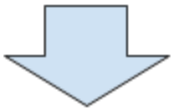
- movie.csv: Το αρχείο περιέχει πληροφορίες για διάφορες ταινίες, όπως π.χ. το αναγνωριστικό id μιας ταινίας, τον τίτλο της, το κόστος δημιουργίας της, τα κέρδη της και την περιγραφή της.
- genre.csv: Το αρχείο αυτό περιέχει τις διάφορες κατηγορίες ταινιών (π.χ. περιπέτεια, τρόμου, κωμωδία). Περιέχει ένα αναγνωριστικό id για κάθε είδος και το όνομα του είδους.
- productioncompany.csv: Το αρχείο αυτό περιέχει τις διάφορες εταιρίες παραγωγής. Περιέχει ένα αναγνωριστικό id για κάθε εταιρία παραγωγής και το όνομά της.
- collection.csv: Αναφέρεται σε μία συλλογή από ταινίες (π.χ. τριλογίες). Περιέχει ένα αναγνωριστικό id για κάθε συλλογή και το όνομά της.
- movie_cast: Περιέχει πληροφορίες για το cast της ταινίας (τους χαρακτήρες της, και τους ηθοποιούς που παίζουν τους χαρακτήρες). Το πεδίο cid προσδιορίζει μοναδικά κάθε επιλογή που έχει γίνει στο casting.

- movie_crew: Περιέχει πληροφορίες το crew μιας ταινίας (σκηνοθέτης, σεναριογράφος, κτλ.). Το πεδίο cid προσδιορίζει μοναδικά κάθε επιλογή που έχει γίνει για το crew μιας ταινίας.
- belongsTocollection.csv: Συσχετίζει μία ταινία με την συλλογή στην οποία ανήκει. Περιέχει τα πεδία movie_id, collection_id.
- hasGenre.csv: Συσχετίζει μία ταινία με τα είδη στα οποία ανήκει. Περιέχει τα πεδία movie_id, genre_id.
- hasproductioncompany.csv: Συσχετίζει μία ταινία με τις εταιρείες παραγωγής της. Περιέχει τα πεδία movie_id, pc_id
- ratings.csv: Συσχετίζει έναν χρήστη με μία ταινία δίνοντας την αντίστοιχη βαθμολογία του χρήστη για την ταινία. Περιέχει τα πεδία user_id, movie_id, rating.
- **keywords.csv: Πρέπει να γίνει προεπεξεργασία για το συγκεκριμένο αρχείο.**

Nested Json

Το αρχείο keywords.csv περιέχει εμφωλευμένη πληροφορία όπως π.χ. λίστες και αντικείμενα που αναπαρίστανται σαν JSon συμβολοσειρές. Πιο συγκεκριμένα, το αρχείο keywords περιέχει 2 πεδία, το ένα είναι το id μιας ταινίας και το άλλο είναι μία JSon συμβολοσειρά που αναφέρεται σε keywords που περιγράφουν την ταινία.

Keywords	
id	**keywords**
862	[{'id': 931, 'name': 'jealousy'}, {'id': 4290, 'name': 'toy'}, {'id': 5202, 'name': 'boy'}, {'id': 6054, 'name': 'friendship'}]
...	



Keyword

id	name
931	jealousy
4290	toy
5202	boy
6054	friendship
...	

haskeyword

movie_id	keyword_id
862	931
862	4290
862	5202
862	6054
...	

Προκειμένου να είναι επίπεδοι οι πίνακές μας, θα πρέπει από την λίστα που υπάρχει εμφωλευμένη στο πεδίο keywords να προκύψουν μία ή περισσότερες εγγραφές που να συσχετίζουν την ταινία με τα αναγνωριστικά των keywords που την περιγράφουν. Για αυτό τον

σκοπό θα δημιουργηθεί ένας καινούριος πίνακας (π.χ. hasKeyword) ο οποίος θα περιέχει το αναγνωριστικό κάθε ταινίας μαζί με το αναγνωριστικό της λέξης-κλειδί. Επίσης θα πρέπει να δημιουργηθεί ένας καινούργιος πίνακας (π.χ. keyword) ο οποίος θα περιέχει τις λέξεις κλειδιά με τα πεδία id, name.

Σε αυτήν την περίπτωση θα πρέπει να φτιαχτούν τα αντίστοιχα *πρωτεύοντα* και *ξένα κλειδιά* για τους δύο καινούργιους πίνακες. Να σημειωθεί ότι, για να δηλωθεί πρωτεύον κλειδί στον πίνακα Keyword, χρειάζεται μία προεπεξεργασία των δεδομένων κατά την οποία θα αφαιρεθούν τα διπλότυπα. Η διαδικασία αυτή μπορεί να γίνει χρησιμοποιώντας την κλάση `set` της Python (μπορεί να γίνει και σε SQL χρησιμοποιώντας κάποιον ενδιάμεσο πίνακα).

Τι θα φτιάξετε:

- Τους πίνακες του MovieLens στο *Azure database instance που φτιάξατε.*
- Η βάση αυτή θα πρέπει να περιέχει πίνακες για τους οποίους θα ισχύουν τα εξής:
    a. Από κάθε csv αρχείο να προκύψουν ένας ή περισσότεροι πίνακες της βάσης δεδομένων. Συγκεκριμένα, 2 πίνακες θα προκύψουν μόνο από το αρχείο keywords.
    b. Σε περίπτωση που το csv αρχείο περιέχει μία στήλη με εμφωλευμένη πληροφορία στην μορφή *JSon* συμβολοσειράς, η συγκεκριμένη πληροφορία να *εξαχθεί* και να αναπαρασταθεί με τον κατάλληλο τρόπο στον υπάρχοντα ή σε νέο πίνακα. Η εξαγωγή της πληροφορίας να γίνει χρησιμοποιώντας *Python parsers* για *JSon* συμβολοσειρές.
    c. Να εισαχθούν σε κάθε πίνακα τα αντίστοιχα δεδομένα χρησιμοποιώντας το SQL Server Import plugin του Azure data studio.
    d. Χρησιμοποιώντας την εντολή **alter table**, να δημιουργηθούν οι **περιορισμοί πρωτεύοντος κλειδιού** για τους πίνακες:
        ■ movie, genre, productioncompany, collection, movie_cast, movie_crew, keyword.
    e. Χρησιμοποιώντας την εντολή **alter table**, να δημιουργηθούν οι **περιορισμοί ξένου κλειδιού** (ενδεχομένως έχω περισσότερα από ένα ξένα κλειδιά για κάθε πίνακα) για τους πίνακες:
        ■ belongsTocollection, hasGenre, hasProductionCompany, Ratings, movie_cast, movie_crew, hasKeyword.
    f. Τοποθετήστε όλες τις εντολές **`alter table`** σε ένα αρχείο, alter_tables.sql.

Απαραίτητα εργαλεία:

- Azure data studio

- SQL Server Import plugin του Azure data studio
- ή εναλλακτικά (δεν το συνιστούμε) το πρόγραμμα bcp για αντιγραφή δεδομένων μεταξύ του SQL Server και αρχείων δεδομένων με συγκεκριμένο format.

Συμβουλές για την υλοποίηση:

- Σας παραθέτουμε στον φάκελο της άσκησης αναλυτική περιγραφή των βημάτων που θα πρέπει να ακολουθήσετε.

Χρήσιμα links:

Εντολές alter table για δημιουργία και διαγραφή πρωτεύοντος και ξένου κλειδιού:

https://learn.microsoft.com/en-us/sql/relational-databases/tables/create-primary-keys

https://learn.microsoft.com/en-us/sql/relational-databases/tables/create-foreign-key-relationships

https://learn.microsoft.com/en-us/sql/relational-databases/tables/delete-primary-keys

Εντολές create table (δεν θα χρειαστούν αν χρησιμοποιήσετε το SQL Server Import :

https://learn.microsoft.com/en-us/sql/t-sql/statements/create-table-transact-sql

Azure Data Studio:

https://learn.microsoft.com/en-us/sql/azure-data-studio/

SQL Server Import extension:

https://learn.microsoft.com/en-us/sql/azure-data-studio/extensions/sql-server-import-extension?view=sql-server-ver16

Παραδοτέα:

- Δημιουργήστε ένα .txt αρχείο στο οποίο θα αναγράφονται τα εξής στοιχεία: ονοματεπώνυμα και αριθμοί μητρώου των μελών της ομάδας, το endpoint του Azure instance σας, το όνομα της βάσης σας και το username και το password του χρήστη examiner ή ενός άλλου 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:

<Ονοματεπώνυμο 1> - <A.M. 1>

<Ονοματεπώνυμο 2> - <A.M. 2>

Endpoint: <name_of_the_endpoint>

Username: <username>

Password: <password>

Database: <name_of_the_database>

- Βάλτε σε ένα φάκελο
    a. τον **python** κώδικα για την επεξεργασία του αρχείου keywords.csv,
    b. το alter_tables.sql αρχείο,
    c. το **αρχείο .txt**,
    d. καθώς και μία **συνοπτική αναφορά** (~ 1 σελίδα) για το τι κάνατε σε κάθε βήμα της άσκησης.

  Το όνομα του φακέλου πρέπει να αποτελείται από τους αριθμούς μητρώου σας χωρισμένους με παύλα, δηλαδή *αριθμός_μητρώου_1-αριθμός_μητρώου_2*. Δημιουργήστε ένα .zip αρχείο αυτού του φακέλου, το οποίο θα έχει το ίδιο όνομα με τον φάκελο.
- Ανεβάστε το .zip αρχείο στο eclass στην ενότητα *Εργασίες / 2η Εργασία*.