2η Εργασία: MovieLens Tables

Προθεσμία: 1/4/2022

Σκοπός:

Σε αυτή την εργασία θα δημιουργήσουμε 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. Θα δημιουργήσουμε τους πίνακες χρησιμοποιώντας SQL και θα εισάγουμε δεδομένα σε αυτούς με την εντολή `\copy`. Επίσης, θα δημιουργήσουμε περιορισμούς ξένου κλειδιού για *αναφορική ακεραιότητα* (*referential integrity*).

Περιγραφή Δεδομένων:

Τα δεδομένα της βάσης Movielens βρίσκονται διαθέσιμα στον εξωτερικό σύνδεσμο:
https://drive.google.com/file/d/1ykpU44D52ZB5p3sgNPMmAOwMaGi-JrfX/view?usp=sharing

Η αρχική μορφή των δεδομένων μπορεί να βρεθεί στον ακόλουθο σύνδεσμο:
https://www.kaggle.com/rounakbanik/the-movies-dataset?select=keywords.csv

movies_metadata.csv: Περιέχει πληροφορίες για 45.000 ταινίες από τη βάση MovieLens. Τα πεδία περιλαμβάνουν πληροφορίες για τα σκηνικά, τον προϋπολογισμό, τα έσοδα, τις ημερομηνίες κυκλοφορίας, τις γλώσσες, και τις χώρες παραγωγής των συγκεκριμένων ταινιών.

keywords.csv: Περιέχει τις λέξεις-κλειδιά της πλοκής των ταινιών της βάσης *MovieLens*. Η στήλη *id* περιέχει το αναγνωριστικό μιας ταινίας, ενώ η στήλη *keywords* περιέχει τις λέξεις κλειδιά κωδικοποιημένες ως JSON συμβολοσειρές.

credits.csv: Περιέχει πληροφορίες για τους συντελεστές των ταινιών της βάσης MovieLens. Η στήλη *id* περιέχει το αναγνωριστικό της ταινίας, ενώ οι στήλες *crew* και *cast* κωδικοποιούν τις πληροφορίες για τους συντελεστές της ταινίας σε μορφή *JSON* συμβολοσειρών.

ratings.csv: Ένα σύνολο 100.000 αξιολογήσεων από 700 χρήστες σε 9.000 ταινίες.

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, movie_collection, collection, κτλ.

Τι θα φτιάξουμε:

- Τη βάση δεδομένων MovieLens σε ένα *Postgres Cloud instance στο Azure*.
- 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.
    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 *Java* ή *Python parsers* για *JSon* συμβολοσειρές.
    c. Να εισαχθούν σε κάθε πίνακα τα αντίστοιχα δεδομένα.
    d. Να δημιουργηθούν οι **περιορισμοί πρωτεύοντος** και/ή **ξένου κλειδιού**.

Nested Json

Κάποιοι από τους πίνακες περιέχουν εμφωλευμένη πληροφορία όπως π.χ. λίστες και αντικείμενα που αναπαρίστανται σαν JSon συμβολοσειρές. Για παράδειγμα ο πίνακας 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.

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

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. Movie_Keywords) 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 τα keywords με τα πεδία id, name.

movie_id	keyword_id
862	931
862	4290

id	name
931	jealousy
4290	toy

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 πρωτογενές 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).

Απαραίτητα εργαλεία:

- Postgres Database on Azure
- Postgres psql client / pgAdmin

Οδηγίες:

- Το πρώτο γράμμα του ονόματος κάθε πίνακα να ξεκινάει με κεφαλαίο(π.χ. Calendar κτλ.).
- Τα ονόματα των πεδίων των πινάκων (attributes) να ξεκινάνε με μικρό (π.χ. id κτλ.).
- Τοποθετήστε κάθε εντολή ***`create table`*** σε ξεχωριστό αρχείο. Για παράδειγμα create_movies.sql.
- **Αν δουλεύετε στο terminal με psql**, να φτιάξετε ένα script create_tables.sql που να καλεί τα ξεχωριστά **`create table`** αρχεία χρησιμοποιώντας την εντολή **`\i`** που αναφέρεται στην συνέχεια.
- Τοποθετήστε όλες τις εντολές **`alter table`** σε ένα αρχείο, alter_tables.sql.

Συμβουλές για την υλοποίηση:

- Συνδεθείτε στη βάση σας στο Azure για να δημιουργήσετε τους πίνακες που ζητά η άσκηση.
- Επειδή μερικές εντολές create table έχουν πολλά πεδία, μπορείτε να τρέξετε το python πρόγραμμα gen_ddl_python3.py (θα το βρείτε στα έγγραφα του μαθήματος στο eclass) αν έχετε python 3 στον υπολογιστή σας ή το gen_ddl_python2.py (θα το βρείτε στα έγγραφα του μαθήματος στο eclass) αν έχετε python 2, το οποίο παίρνει ως παράμετρο το .csv αρχείο των δεδομένων, π.χ. ratings.csv για τον πίνακα Ratings, και παράγει ένα αρχείο .sql με την εντολή create table για τον αντίστοιχο πίνακα. **Ελέγξτε την παραγόμενη εντολή και προσθέστε τους περιορισμούς για πρωτεύοντα κλειδιά**.
- Χρησιμοποιήστε την εντολή **`\i <filename>`** στην psql για να εκτελέσετε τον κώδικα SQL που έχετε αποθηκεύσει σε ένα αρχείο. Για παράδειγμα **`\i create_tables.sql.`** Εναλλακτικά, στο pgAdmin επιλέξτε τη βάση, πατήστε το query tool (κεραυνός πάνω αριστερά) και τρέξτε ένα sql script ως εξής: πατήστε το “Open file” (εικονίδιο φακέλου πάνω αριστερά στο query tool), επιλέξτε το sql script από τον υπολογιστή σας και πατήστε τον “κεραυνό” στη μπάρα του query tool.

- Χρησιμοποιήστε την εντολή **`\copy`** στην psql ή τη λειτουργία **`Import/Export`** στο pgAdmin για να εισάγετε τα δεδομένα.
    - Λάβετε υπόψη σας ότι η πρώτη γραμμή στα .csv αρχεία είναι ο header. Δε θέλουμε να εισάγουμε τον header στον πίνακα. Θέστε την κατάλληλη παράμετρο είτε στο **`\copy`** είτε στην λειτουργία **`Import/Export`** ώστε να προσπεράσετε αυτή τη γραμμή. Παράδειγμα εντολής \copy:
      `\copy movie FROM 'movie.csv' DELIMITER ',' QUOTE '"' ESCAPE '"' CSV HEADER;`
    - Επίσης κατά την εισαγωγή των δεδομένων να χρησιμοποιηθούν οι ακόλουθοι παράμετροι

delimiter	,
quote	"
escape	"

- Πριν εκτελέσετε την εντολή **`\copy`**, τρέξτε την εντολή **`set client_encoding to 'utf8';`** στην psql για να αποφύγετε προβλήματα με την κωδικοποίηση των χαρακτήρων.
- Προσθέστε τους περιορισμούς **ξένου κλειδιού μετά την εισαγωγή των δεδομένων στους πίνακες**.

Χρήσιμα links:

Εντολή create table:
https://www.postgresql.org/docs/9.6/sql-createtable.html

Εντολή copy:
https://www.postgresql.org/docs/9.6/sql-copy.html

Εντολή alter table:
https://www.postgresql.org/docs/9.6/sql-altertable.html

Postgres meta commands, όπως η \copy:
https://www.postgresql.org/docs/9.2/app-psql.html

pgAdmin query tool:
https://www.pgadmin.org/docs/pgadmin4/dev/query_tool.html

pgAdmin import:
https://www.pgadmin.org/docs/pgadmin4/dev/import_export_data.html

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 όλα τα .sql αρχεία, το αρχείο .txt σε ένα φάκελο. 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*.