

ΑΝΑΦΟΡΑ ΤΟΥ PROJECT

ΣΤΟ ΜΑΘΗΜΑ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

Θα περιγράψουμε τα βήματα της προεπεξεργασίας των δεδομένων.

Αρχικά φτιάχνουμε τους πίνακες της βάσης και τους κάνουμε import τα δεδομένα των δικών τους csv. Στη συνέχεια, φτιάχνουμε νέα αντίγραφα πινάκων, στα οποία θα βάλουμε τα στοιχεία των αντίστοιχων πρωτοτύπων χωρίς τα διπλότυπα. Διαγράφουμε τους παλιούς πίνακες και μετονομάζουμε τους νέους πίνακες στο όνομα των παλιών. Ύστερα ψάχνουμε με queries τους πίνακες εκτός του Movies_Metadata και τους ελέγχουμε για το αν περιέχουν κάποια ID τα οποία δεν περιέχονται στον πίνακα Movies_Metadata. Σε αυτά που βρίσκουμε (στους πίνακες Links, Ratings_Small), τα διαγράφουμε με την εντολή **DELETE**.

Επειδή κάποιες ιδιότητες είναι δοσμένες σε τύπο text, αλλά είναι αρκετά κοντά στο να είναι τύπου **JSON** (array και μη), θα τα επεξεργαστούμε κατάλληλα ώστε να μπορούμε να τα μετατρέψουμε σε JSON. Αυτό θα το κάνουμε στους πίνακες Keywords, Movies_Metadata και Credits, με την κατάλληλη χρήση της εντολής **REPLACE ***. Τώρα που τα δεδομένα είναι JSON compatible, τα μετατρέπουμε με τις κατάλληλες εντολές σε τύπο **jsonb**.

Αποφασίσαμε να διαγράψουμε τις ιδιότητες **genres** και **keywords**, που είναι πλέον τύπου jsonb, και να τις κάνουμε πίνακες (Genre και Keyword_IDs). Αυτοί οι 2 πίνακες συνδέονται με τον Movies_Metadata με τους πίνακες συσχετισμού **Movie_Genre** και **Keywords** αντίστοιχα. Το κάναμε για να μπορούν αυτά τα πεδία εύκολα να αλλάζουν, καθώς μπορούμε σε μια ταινία να προσθέσουμε και να αφαιρέσουμε keywords και νέα genres ανάλογα με τις ανάγκες της βάσης. Είναι προτιμότερο οι ιδιότητες τύπου json να αναφέρονται σε στατικές τιμές που είναι κυρίως read-only, όπως είναι οι υπόλοιπες ιδιότητες που μετατρέψαμε σε JSON (crew, movie_cast, production_companies, production_countries, spoken_languages, belongs_to_collection).

Τέλος, φτιάχνουμε σε κάθε πίνακα το (ή τα) ανάλογο κλειδί(Primary Key) που του αντιστοιχεί (αν υπάρχει) και ύστερα συνδέουμε τους πίνακές μας μεταξύ τους με τα κατάλληλα ξένα κλειδιά (Foreign keys).

Και η Βάση Δεδομένων ετοιμάστηκε.

*Αποφασίσαμε να μετατρέψουμε κάποια δεδομένα τύπου text σε τύπου JSON ώστε να έχουμε εύκολη πρόσβαση με queries και σε αυτά. Να μην χρειάζεται να χρησιμοποιούμε συναρτήσεις για strings για να τα ανακτούμε.