ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Project

Όνομα 1: **Παναγιώτης Ντυμένος** *ΑΜ 1:* **3160120**

Όνομα 2: **Σταμάτης Φακορέλλης** *ΑΜ 2:* **3160185**

Αρχικά Σχόλια:

Ο βασικός πίνακας είναι ο πίνακας Metadata ο οποίος και έχει αναλυθεί περαιτέρω καθώς παρατηρήσαμε πως μπορεί να γίνει μια κανονικοποίηση.

(Στο τέλος υπάρχει και φωτογραφία με τις σχέσεις των πινάκων εκτός του διαγράμματος ***ERD***)

Τα πεδία JSON θα παρατηρήσετε πως είναι πεδία TEXT στη βάση καθώς είχαν τεράστια προβλήματα κατά την δημιουργία τους. Ωστόσο δημιουργήθηκαν πίνακες έτσι ώστε να γίνει μια κανονικοποίηση ακόμα και αν δεν τους χρησιμοποιήσουμε στο δεύτερο μέρος. Τα μόνα πεδία που παραμένουν TEXT είναι τα πεδία cast και crew του πίνακα Credits καθώς είχε τόση πολλή φασαρία στο να τα αλλάξουμε και να θέσουμε και ως primary key το id, επομένως λόγω πίεσης χρόνου προχωρήσαμε στο δεύτερο μέρος.

Εφόσον έγινε μια μικρή περιγραφή της κατάστασης της βάσης, ακολουθεί περιγραφή του κώδικα SQL του Α' μέρους:

Βήμα 1ο

Δημιουργήθηκαν οι πίνακες με την κατάληξη ***..CSV*** μόνο και μόνο για να υπάρχουν όλα τα δεδομένα στην βάση. Οι πίνακες είναι αρχικοί και δεν θα είναι αυτοί που θα βρείτε στα παραδοτέα.

Επομένως έγινε το κατάλληλο import για κάθε πίνακα με το αντίστοιχο αρχείο CSV.

Τα πεδία με μεγάλη έκταση (πχ JSON) έχουν χαρακτηριστεί ως TEXT αν και αργότερα θα γίνει το κατάλληλο Edit. Επίσης προς το παρόν οι τιμές BOOLEAN είναι VARCHAR.

Βήμα 2ο

Άλλαξαν τα πεδία TEXT έτσι ώστε τα μονά Quotes να γίνουν διπλά για να μπορεί να γίνει η κατάλληλη μετατροπή τους σε JSON. (*σειρά 60*).

Πρώτα μετατράπηκε το πεδίο genres από την στιγμή που είδαμε πως χρειάζεται και στο δεύτερο μέρος.

Έτσι, τα δεδομένα του πεδίου δημιούργησαν τους πίνακες **Metadata_Genres** και **Genres**. Ο Genres έχει όλα τα είδη ταινιών που υπάρχουν στην βάση και ο πίνακας **Metadata_Genres** είναι ο ενδιάμεσος

πίνακας που ενώνει του πίνακες **Metadata** και **Genres**. Έτσι επιτυγχάνεται και η σχέση πολλά-πολλά μεταξύ τους.

Στις σειρές 79-84 γίνονται και οι δημιουργίες/διαγραφές των constraints. Στην σειρά 87 διαγράφεται και το πεδίο genres από τον Metadata εφόσον δεν υπάρχει λόγος να το κρατάμε.

Βήμα 3ο

Οι πίνακες ***Links***, ***Credits***, ***Ratings*** και ***Keywords*** δημιουργούνται βάση των αντίστοιχων αρχείων CSV ΑΛΛΆ μόνο με τα δεδομένα που βρίσκονται και στον **Metadata**.

Μετά από κάθε δημιουργία πίνακα θα δείτε και τις δημιουργίες των constraints που χρειάζεται.

Βήμα 4ο

Τροποποιήθηκε το πεδίο rating_timestamp από VARCHAR σε TIMESTAMP. Μετονομάστηκε σε date_time.

Τροποποιήθηκε το πεδίο imdb_id έτσι ώστε να έχει πάντα τους χαρακτήρες '**tt**' στην αρχή. (Η αλλαγή έγινε βάση του πραγματικού API της imdb).

Διορθώθηκαν μικρό-λάθη όπως χαρακτήρες της μορφής '**\xa0**' οι οποίοι δεν επέτρεπαν την μετατροπή των πεδίων TEXT σε JSON.

Βήμα 5ο

Αναλύθηκαν πεδία όπως production_countries, production_companies, keywords, spoken_languages και εκτός από το ότι μετατράπηκαν σε μορφή JSON, δημιούργησαν και τους αντίστοιχους πίνακες με αποτέλεσμα την κανονικοποίηση της Βάσης.(βλ. *Βήμα 2ο*)

Όσα πεδία βρίσκονται στην μορφή VARCHAR αλλά είναι φανερά BOOL, έχουν υποστεί τις απαραίτητες μετατροπές. (πχ adult)

Τέλος, αφαιρέθηκαν τα παραπάνω πεδία από τον πίνακα Metadata.

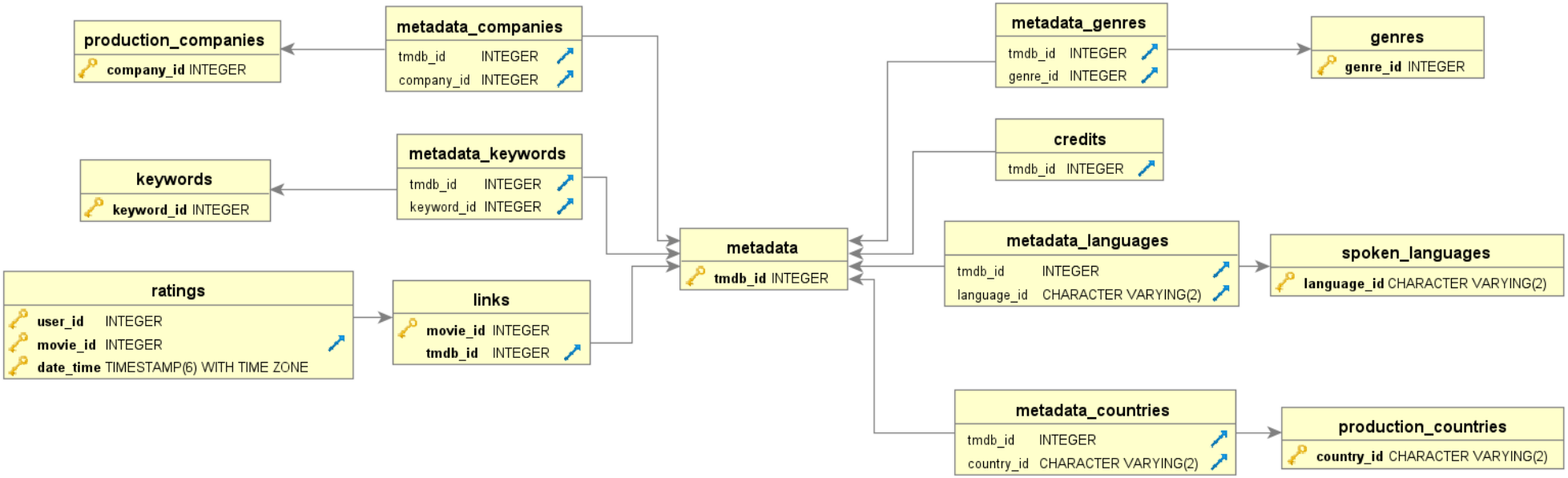
(*ΣΗΜΕΙΩΣΗ: Η βάση δεν βρίσκεται σε μορφή BCNF*)

Βήμα 6ο

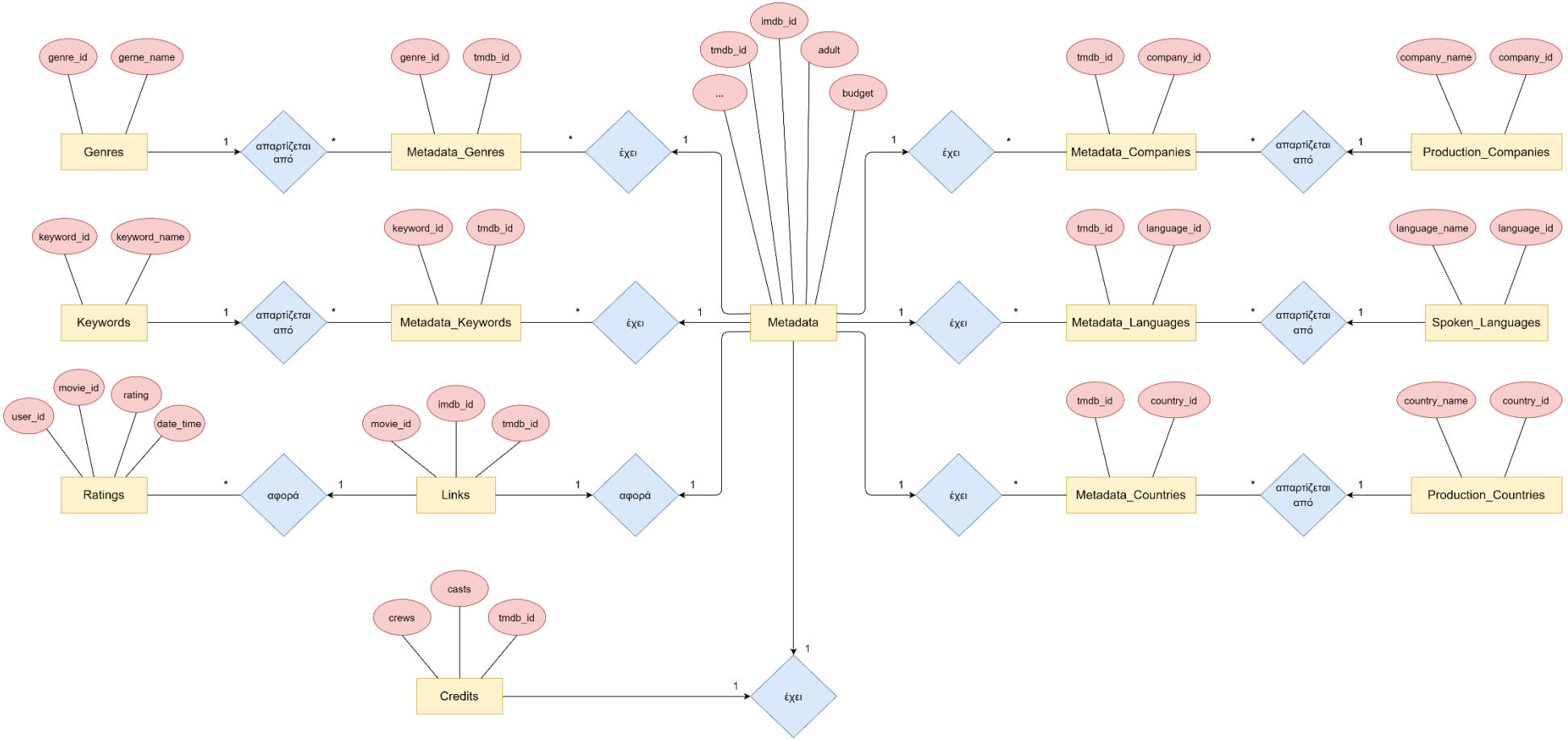
Διαγραφή όλων των πινάκων που χρειάστηκαν μόνο για την υλοποίηση της Βάσης(πχ οι πίνακες με κατάληξη ..CSV)

Τέλος, δόθηκαν τα απαραίτητα δικαιώματα στον χρήστη examiner να εκτελεί Queries πάνω στην Βάση MoviesDB και δημιουργήθηκε το διάγραμμα **ER**.

TABLE RELATIONS



ER Diagram


genre_id
gerne_name
Genres
απαρτίζεται από
genre_id
tmdb_id
Metadata_Genres
έχει
imdb_id
tmdb_id
adult
...
budget
Metadata
έχει
tmdb_id
company_id
Metadata_Companies
απαρτίζεται από
company_name
company_id
Production_Companies
keyword_id
keyword_name
Keywords
απαρτίζεται από
keyword_id
tmdb_id
Metadata_Keywords
έχει
έχει
tmdb_id
language_id
Metadata_Languages
απαρτίζεται από
language_name
language_id
Spoken_Languages
movie_id
rating
user_id
date_time
Ratings
αφορά
movie_id
imdb_id
tmdb_id
Links
αφορά
έχει
tmdb_id
country_id
Metadata_Countries
απαρτίζεται από
country_name
country_id
Production_Countries
casts
crews
tmdb_id
Credits
έχει
1
*