Διαχείριση Μεγάλων Δεδομένων
2η Προγραμματιστική Εργασία

Διδάσκων:
Χ. Τρυφωνόπουλος

Παράδοση μέχρι: **Τετάρτη 30/06/2021 ώρα 23.59**
Εξέταση: στην τελευταία διάλεξη

ΣΗΜΑΝΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ:

1. Σε όλα τα αρχεία που θα παραδώσετε θα πρέπει **ΟΠΩΣΔΗΠΟΤΕ** να βάλετε τα ονόματα, τους Α.Μ., και τα username/email των μελών της ομάδας (ομάδες 2 ατόμων).
2. Αφού έχετε ολοκληρώσει την εργασία που θέλετε να παραδώσετε την υποβάλετε στο eclass στο υποσύστημα «Εργασίες φοιτητών». Προσοχή: μόνο 1 άτομο από την ομάδα χρειάζεται να παραδώσει την εργασία μέσω του e-class! Η υποβολή πρέπει να γίνει **ΠΡΙΝ** την καταληκτική ημερομηνία παράδοσης. Παραδίδετε όλα τα αρχεία που σας ζητούνται μαζί σε ένα συμπιεσμένο αρχείο (το οποίο θα φέρει τα ονόματα της ομάδας π.χ., TryfonopoulosPapadopoulos.zip).
3. Περιπτώσεις αντιγραφής θα μηδενίζονται και όλοι οι εμπλεκόμενοι δε θα έχουν δικαίωμα παράδοσης άλλων εργασιών.
4. Η ημερομηνία παράδοσης είναι αυστηρή, και η παράδοση γίνεται μόνο μέσω του eclass και όχι με email στο διδάσκοντα. Ασκήσεις που παραδίδονται μετά τη λήξη της προθεσμίας δε γίνονται δεκτές.

Ποιες ταινίες και σειρές παρακολουθεί το κοινό; Ποια είναι τα χαρακτηριστικά που τις κάνουν δημοφιλείς; Τι άλλο θα θέλαμε να μάθουμε για αυτές τις παραγωγές; Μια μελέτη των δεδομένων που παρέχονται μέσω της δημοφιλούς πλατφόρμας Netflix.

Η Netflix[1], η οποία ιδρύθηκε το 1997 από τους Reed Hastings και Marc Randolph στην Scotts Valley της Καλιφόρνια, είναι μια αμερικανική πλατφόρμα περιεχομένου και παραγωγής, η οποία προσφέρει διαδικτυακές συνδρομητικές υπηρεσίες ροής (streaming services) και υποστηρίζει μια μεγάλη βιβλιοθήκη ταινιών και τηλεοπτικών σειρών, συμπεριλαμβανομένων εκείνων που παράγει η ίδια η εταιρεία. Το αρχικό επιχειρηματικό μοντέλο της Netflix περιλάμβανε πωλήσεις και ενοικίαση DVD μέσω ταχυδρομείου, αλλά ο Reed Hastings εγκατέλειψε τις πωλήσεις περίπου ένα χρόνο μετά την ίδρυση της εταιρείας προκειμένου να επικεντρωθεί στην ενοικίαση DVD. Η εταιρεία διεύρυνε το επιχειρηματικό της μοντέλο το 2007 με την εισαγωγή υπηρεσιών ροής, και επεκτάθηκε διεθνώς το 2010 κάνοντας διαθέσιμο το περιεχόμενό της στον Καναδά, στη Λατινική Αμερική, και στην Καραϊβική. Η Netflix εισήλθε στη βιομηχανία παραγωγής περιεχομένου το 2013 με τη σειρά House of Cards.

Σήμερα, η Netflix είναι μέλος της Motion Picture Association[2] (MPA), της αμερικανικής εμπορικής ένωσης που εκπροσωπεί τα πέντε μεγαλύτερα κινηματογραφικά στούντιο των ΗΠΑ, παράγει και διανέμει περιεχόμενο από χώρες σε όλο τον κόσμο. Οι υπηρεσίες της πλατφόρμας Netflix είναι διαθέσιμες σε όλο τον κόσμο εκτός από την ηπειρωτική Κίνα (λόγω τοπικών περιορισμών), τη Συρία, τη Βόρεια Κορέα, και την Κριμαία (λόγω κυρώσεων των ΗΠΑ). Τον Οκτώβριο του 2020, η Netflix κατέγραψε πάνω από 195 εκατομμύρια συνδρομές παγκοσμίως, εκ των οποίων τα 73 εκατομμυρίων προέρχονταν από τις Ηνωμένες Πολιτείες. Για το 2020, το ως τώρα λειτουργικό εισόδημα της πλατφόρμας είναι 1,2 δισεκατομμύρια δολάρια.

[1] https://www.netflix.com/
[2] https://www.motionpictures.org

Η Netflix είναι γνωστή για τους μηχανισμούς συστάσεων (recommendation engines) που χρησιμοποιεί. Πρόκειται για ένα συνδυασμό μοντέλων φιλτραρίσματος βάσει περιεχομένου και αλληλεπίδρασης για να προτείνει σε χρήστες της πλατφόρμας ταινίες και τηλεοπτικές σειρές που θα τους ενδιαφέρουν.

Σκοπός της εργασίας

Σκοπός αυτής της εργασίας είναι να διαχειριστείτε τα δεδομένα της Netflix μέσω μιας NoSQL βάσης δεδομένων. Θα πρέπει να αποθηκεύσετε τα δεδομένα σας στο **MongoDB**[3] document store, κι έπειτα να είστε σε θέση να απαντάτε ερωτήσεις όπως:

- Τι είδους περιεχόμενο είναι διαθέσιμο μέσω της πλατφόρμας Netflix;
- Ποιες είναι οι 20 κορυφαίες χώρες παραγωγής περιεχομένου;
- Ποιοι είναι οι 20 πιο συχνά εμφανιζόμενοι ηθοποιοί στις ταινίες και σειρές της Netflix;
- Ποια είναι η κατανομή της διάρκειας του διαθέσιμου περιεχομένου;
- Ποια είναι τα 10 κορυφαία είδη περιεχομένου που διατίθενται μέσω της πλατφόρμας; Πώς συσχετίζονται μεταξύ τους;
- Πώς εξελίσσεται το διαθέσιμο περιεχόμενο στο πέρασμα των μηνών και των χρόνων;
- Σε τι είδους κοινό απευθύνονται κυρίως οι ταινίες και οι σειρές που παράγονται από τη Netflix;

Εκτός από τις διαφάνειες του μαθήματος, στις οποίες συζητήσαμε για τη MongoDB, πληροφορίες μπορείτε να βρείτε και στην επίσημη ιστοσελίδα της βάσης. Η εργασία θα εκπονηθεί από ομάδες των **2 ατόμων** και μπορεί να γίνει σε οποιοδήποτε λειτουργικό σύστημα.

Για τις ανάγκες της εργασίας, θα κατεβάσετε και θα εγκαταστήσετε τη **δωρεάν stable έκδοση** της **MongoDB v4.4** στο μηχάνημά σας.

1. Αποθήκευση δεδομένων στο MongoDB [10 μονάδες]

Τα δεδομένα είναι διαθέσιμα στο αρχείο **netflix_titles.csv** (2,4 MB) στο eclass του μαθήματος. Εκεί θα βρείτε 6.235 εγγραφές, οι οποίες αφορούν σε υλικό που έγινε διαθέσιμο μέσω της πλατφόρμας Netflix την περίοδο 2013 - 2019 (date_added). Για τις εγγραφές κρατούνται τα παρακάτω χαρακτηριστικά/πεδία (attributes):

- show_id - Unique ID for every entry
- type - An Identifier; it can be either Movie or TV Show
- title - Title of the Movie / TV Show
- director - Director of the Movie
- cast - Actors involved in the Movie / TV show
- country - Country where the Movie / TV show was produced
- date_added - Date the Movie / TV show was added on Netflix
- release_year - Actual Release year of the Movie / TV show
- rating - TV Rating of the Movie / TV show
- duration - Total Duration in minutes for movies or number of seasons for TV shows
- listed_in - Genre (Documentaries, Stand-Up Comedy, Dramas, Comedies, Kids' TV, International Movies, Independent Movies, etc.)
- description - The summary description

[3] https://www.mongodb.com

Δείτε ένα μικρό υποσύνολο των δεδομένων (για λόγους παρουσίασης παραλείπονται χαρακτηριστικά):

show_id	type	title	director	cast	duration
80057700	Movie	The Runner	Austin Stark	Nicolas Cage, Sarah Paulson, Connie Nielsen, Wendell Pierce, Bryan Batt, Peter Fonda, Dana Gourrier	90 min
81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole Howard, Jennifer Cameron, Jonathan Holmes, Lee Tockar, Lisa Durupt, Maya Kay, Michael Dobson	90 min
70153404	TV Show	Friends		Jennifer Aniston, Courteney Cox, Lisa Kudrow, Matt LeBlanc, Matthew Perry, David Schwimmer	10 Seasons

Μελετήστε τα δεδομένα σας, δείτε προσεκτικά τις τιμές για κάθε χαρακτηριστικό, παρατηρήστε ότι κάποιες εγγραφές έχουν πολλαπλές τιμές σε κάποιο πεδίο ή έχουν κενά κάποια από τα πεδία τους, προσέξτε επίσης ότι για το ίδιο χαρακτηριστικό μπορεί να έχουμε διαφορετικού τύπου τιμές ανάλογα με το αν αφορούν σε ταινία ή σε σειρά, κοκ. Σε μια σχεσιακή βάση δεδομένων, θα έπρεπε να αποφασίσουμε ποιο είναι το schema των δεδομένων μας και πώς θα αντιμετωπίσουμε τέτοιου είδους διαφοροποιήσεις στις τιμές των χαρακτηριστικών. Στα πλαίσια αυτής της εργασίας, τίποτα από αυτά δε θα μας απασχολήσει καθώς θα δουλέψουμε με μια NoSQL βάση δεδομένων, και συγκεκριμένα με τη MongoDB όπου τα δεδομένα αποθηκεύονται ως JSON έγγραφα.

Το JSON (JavaScript Object Notation) είναι ένα format περιγραφής κι ανταλλαγής δεδομένων, το οποίο είναι εύκολα κατανοητό από τους ανθρώπους, αλλά κυρίως προσφέρει έναν εύκολο τρόπο για αποθήκευση, ανάκτηση και ανάλυση δεδομένων στα μηχανήματα. Με λίγα λόγια, το JSON είναι μια μορφή κειμένου που είναι εντελώς ανεξάρτητη από τη γλώσσα, υποστηρίζει ενσωματωμένα πεδία (επομένως σχετιζόμενα δεδομένα ή λίστες δεδομένων) και χρησιμοποιεί συμβάσεις που είναι γνωστές στους προγραμματιστές· αυτές οι ιδιότητες καθιστούν το JSON ιδανικό για αποθήκευση και ανταλλαγή δεδομένων. Εκτός από τις διαφάνειες του μαθήματος, περισσότερα για το JSON και για την μοντελοποίηση των δεδομένων στη MongoBD μπορείτε να βρείτε και στους παρακάτω συνδέσμους:

[7] https://docs.mongodb.com/manual/core/data-modeling-introduction/
[8] https://www.json.org/

Στα πλαίσια αυτής της εργασίας, θα πρέπει να μετατρέψετε τα δεδομένα που σας δίνονται στο αρχείο netflix_titles.csv **σε JSON format**. Δείτε για παράδειγμα την παρακάτω εγγραφή:

```
{"show_id": 80057700,
 "type": "Movie",
 "title": "The Runner",
 "director": "Austin Stark",
 "cast": [ "Nicolas Cage", "Sarah Paulson", "Connie Nielsen", "Wendell Pierce", "Bryan Batt", "Peter Fonda", "Dana Gourrier" ],
 …
 "duration": "90 min"}
```

Σημειώστε ότι η παραπάνω εγγραφή είναι μόνο ένα παράδειγμα, το οποίο δε δείχνει όλη της δυναμική του JSON format. Σε αυτό το παράδειγμα, οι ηθοποιοί είναι αποθηκευμένοι σε πίνακα, αλλά υπάρχουν και άλλες επιλογές/δομές που μπορούν να υποστηρίξουν πολλαπλές τιμές σε κάποιο πεδίο.

Για την μετατροπή των δεδομένων σε JSON μπορείτε να χρησιμοποιήσετε κάποιον από τους csv σε json μετατροπείς που είναι διαθέσιμοι online (αρκεί να αναφέρετε την πηγή σας στην αναφορά) ή να φτιάξετε κάποιον δικό σας (μια καλή ιδέα θα ήταν με χρήση shell/python script). Σε κάθε περίπτωση σκεφτείτε προσεκτικά πώς θα διαχειριστείτε τις πολλαπλές τιμές σε ένα πεδίο, όπως για παράδειγμα στα πεδία cast ή listed_in, καθώς επίσης και τις μη αλφαριθμητικές τιμές, όπως για παράδειγμα τις ημερομηνίες.

Στη συνέχεια, αφού εγκαταστήσετε τη MongoDB στον υπολογιστή σας, θα πρέπει να εισάγετε τα δεδομένα σας στη βάση. Η εισαγωγή των δεδομένων μπορεί να γίνει μέσω του **mongo Shell** (διανέμεται μαζί με την MongoDB). Αντί αυτού, μια καλή ιδέα θα ήταν να χρησιμοποιήσετε το εργαλείο **mongoimport** για μαζική εισαγωγή δεδομένων στο document store. Δυστυχώς αυτό το εργαλείο δε διανέμεται μαζί με την τελευταία έκδοση (v4.4) της MongoDB, οπότε, αν θέλετε να το χρησιμοποιήσετε, θα πρέπει να το κατεβάσετε και να το εγκαταστήσετε ανεξάρτητα.

Στη συνέχεια, ανακτήστε όλα τα έγγραφα από τη συλλογή σας, δείτε κι ελέγξτε τα αποτελέσματά σας, σιγουρευτείτε ότι όλα πήγαν καλά ως εδώ.

2. Ανάλυση των δεδομένων

Τι μπορείτε να κάνετε με αυτά τα δεδομένα; Θα αρχίσετε απαντώντας μερικές απλές ερωτήσεις κι έπειτα θα κάνετε μια μικρή μελέτη του περιεχομένου της πλατφόρμας Netflix. Η διατύπωση των ερωτημάτων σας πάνω από τα αποθηκευμένα δεδομένα (json documents) μπορεί να γίνει μέσω του **mongo Shell** και κάνοντας χρήση της mongoDB γλώσσας ερωτήσεων[4] (mongoQL). Αντί αυτού, μια καλή ιδέα θα ήταν να χρησιμοποιήσετε **MongoDB Compass**[5] που ουσιαστικά είναι το GUI για το MongoDB, οπότε και δεν χρειάζεται να αποκτήσετε κάποια ιδιαίτερη εξοικείωση με τη σύνταξη ερωτημάτων σε mongoQL. Αν θέλετε να χρησιμοποιήσετε το compass θα πρέπει να το κατεβάσετε και να το εγκαταστήσετε ανεξάρτητα.

Για κάθε ένα από τα παρακάτω ερωτήματα θα εξάγετε όλες τις απαντήσεις σε αρχείο αποτελεσμάτων, το οποίο και θα ανεβάσετε στο eclass μαζί με τα υπόλοιπα παραδοτέα της άσκησης. Επίσης, για κάθε ένα από τα ερωτήματα θα συμπεριλάβετε τις 20 πρώτες εγγραφές των αποτελεσμάτων στη γραπτή αναφορά σας μαζί με τα σχόλιά σας. Αν οι απαντήσεις για κάποιο ερώτημα είναι λιγότερες από 20, τότε θα τις συμπεριλάβετε όλες στην αναφορά σας.

i. Διαθέσιμο περιεχόμενο το 2019 [10 μονάδες]
Βρείτε το περιεχόμενο που έγινε διαθέσιμο μέσω της πλατφόρμας Netflix το 2019 (date_added). Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε τα show_id, type, και title.

ii. Χώρες παραγωγής σειρών [15 μονάδες]
Βρείτε το πλήθος των σειρών (TV Show) ανά διαφορετική χώρα παραγωγής (country). Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε τη χώρα παραγωγής και το πλήθος σειρών που έχει παραγάγει. Τα αποτελέσματά σας θα πρέπει να είναι ταξινομημένα σε φθίνουσα σειρά ως προς το πλήθος των σειρών.

iii. Είδη διαθέσιμου περιεχομένου [15 μονάδες]
Βρείτε όλα τα διαφορετικά είδη (genre) περιεχομένου που γίνονται διαθέσιμα μέσω της Netflix και το σύνολο των ταινιών και σειρών (συνολικά) που ανήκουν σε κάθε είδος. Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε το είδος περιεχομένου και το πλήθος παραγωγών που ανήκει σε αυτό. Τα αποτελέσματά σας θα πρέπει να είναι ταξινομημένα σε φθίνουσα σειρά ως προς το πλήθος των παραγωγών.

iv. Εμφανιζόμενοι ηθοποιοί [15 μονάδες]
Βρείτε τους 20 πιο συχνά εμφανιζόμενους ηθοποιούς στις ταινίες και τις σειρές που γίνονται διαθέσιμες μέσω της Netflix, και για κάθε ηθοποιό το πλήθος των ταινιών και σειρών που έχει συμμετάσχει συνολικά. Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε το ονοματεπώνυμο του ηθοποιού και το πλήθος των παραγωγών που έχει αυτός συμμετάσχει. Τα αποτελέσματά σας θα πρέπει να είναι ταξινομημένα σε φθίνουσα σειρά ως προς το πλήθος των παραγωγών.

v. Κορυφαίες προτιμήσεις ηθοποιών [20 μονάδες]
Βρείτε σε ποιο είδος (genre) περιεχομένου εμφανίζεται κυρίως κάθε ηθοποιός και σε πόσες παραγωγές αυτού του είδους έχει συμμετάσχει. Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε τον ηθοποιό, το πιο συχνό είδος περιεχομένου που συμμετέχει, και το πλήθος δουλειών που έχει συμμετάσχει και ανήκουν σε αυτό το είδος. Τα αποτελέσματά σας θα πρέπει να είναι ταξινομημένα σε αύξουσα σειρά ως προς τα ονόματα των ηθοποιών.

[4] https://docs.mongodb.com/manual/tutorial/query-documents/
[5] https://docs.mongodb.com/compass/

Οπτικοποίηση/γραφική απεικόνιση αποτελεσμάτων (15 μονάδες)

Χρησιμοποιείστε κάποιο εργαλείο/εφαρμογή για να κάνετε γραφική απεικόνιση των αποτελεσμάτων σε κάθε ερώτημα. Για τις γραφικές απεικονίσεις θα μπορούσατε να χρησιμοποιήσετε Gnuplot[6], robomongo[7], Grafana[8], Visual Studio Code[9], ή κάποιο άλλο (εκτός από Excel) που εσείς προτιμάτε ή έχετε χρησιμοποιήσει σε συνεννόηση με το διδάσκοντα.

Η επιλογή του εργαλείου απεικόνισης (μαζί με λεπτομέρειες που αξίζει να σημειωθούν), οι γραφικές απεικονίσεις των αποτελεσμάτων, και τα σχόλια σας θα πρέπει να συμπεριληφθούν στην αναφορά.

Παραδοτέα και βαθμολόγηση

Πλήρης θεωρείται η εργασία η οποία **υλοποιεί σωστά τις απαιτήσεις/ερωτήσεις** που περιγράφονται παραπάνω. Ασκήσεις που υλοποιούν μόνο ένα μέρος των βασικών απαιτήσεων λαμβάνουν και αντίστοιχο μέρος του βαθμού.

Τα παραδοτέα της εργασίας είναι:

- τα scripts που χρησιμοποιήσατε
- τα αρχεία με τα αποτελέσματα για κάθε ένα από τα παραπάνω ερωτήματα
- μία γραπτή αναφορά που θα περιέχει:
    - τις ενέργειες σας (μαζί με κατάλληλη αιτιολόγηση) σε σχέση με την μετατροπή των δεδομένων σας από csv σε json,
    - τα εργαλεία που εγκαταστήσατε και χρησιμοποιήσατε (μαζί με λεπτομέρειες που αξίζει να σημειωθούν),
    - τα ερωτήματα που υποβάλατε στη βάση σας,
    - οδηγίες για την εκτέλεση των ερωτημάτων,
    - τις 20 πρώτες εγγραφές των αποτελεσμάτων για κάθε ένα από τα ερωτήματα,
    - τις γραφικές απεικονίσεις των αποτελεσμάτων, και
    - τα σχόλιά σας σε σχέση με τα αποτελέσματα (π.χ., τι αναμένατε και τι πήρατε ως αποτέλεσμα, κα.) για κάθε ένα από τα ερωτήματα.

Καλή επιτυχία!

[6] http://www.gnuplot.info
[7] https://robomongo.org
[8] https://grafana.com
[9] https://code.visualstudio.com