Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής και Τηλεματικής

Συστήματα Διαχείρισης Δεδομένων Μεγάλης Κλίμακας
Άσκηση

Διδάσκων **Δημήτρης Μιχαήλ**
Ακ. Έτος **2021-2022**

Οδηγίες Παράδοσης

1. Η άσκηση μπορεί να γίνει σε ομάδες 2 φοιτητών.
2. Η παράδοση της άσκησης πρέπει να γίνει ηλεκτρονικά μέσω της πλατφόρμας `http://eclass.hua.gr`. Μπορείτε να ανεβάσετε την άσκηση σας μέχρι και την ημέρα της παράδοσης.
3. Το παραπάνω zip αρχείο πρέπει να περιέχει
   (a) ένα φάκελο **src** με τον πηγαίο κώδικα της άσκησης
   (b) ένα .pdf αρχείο με την αναφορά.

   Το αρχείο πρέπει να περιέχει μόνο τον πηγαίο κώδικα και όχι και τα εκτελέσιμα αρχεία.
4. Η αναφορά πρέπει να περιέχει εισαγωγή στο θέμα, λεπτομερή ανάλυση της λύσης που υλοποιήσατε μαζί με τον κώδικα που γράψατε καθώς και παραδείγματα εκτέλεσης του.
5. Σε περίπτωση αντιγραφής θα μηδενίζονται **όλες** οι εμπλεκόμενες ασκήσεις.

Άσκηση

Στην άσκηση αυτή καλείστε να χρησιμοποιήστε το Spark για να πραγματοποιήσετε βασικές λειτουργίες με σχεσιακή άλγεβρα. Στο πρώτο μέρος της άσκησης καλείστε σε επίπεδο RDD να υλοποιήσετε κάποια βασικά query με joins ενώ στο δεύτερο μέρος της άσκησης καλείστε να υλοποιήσετε τα ίδια και μερικά επιπλέον ερωτήματα με την χρήση Spark DataFrames.

Dataset

Για τους σκοπούς της εργασίας αυτής θα χρησιμοποιήσουμε το movielens dataset (`https://grouplens.org/datasets/movielens/`). Το συγκεκριμένο dataset υπάρχει σε διάφορα μεγέθη. Μπορείτε να χρησιμοποιήσετε το 10m για την υλοποίηση σας.

Μέρος 1ο

Στο πρώτο μέρος της εργασίας θα πρέπει να υλοποιήσετε τα παρακάτω ερωτήματα απευθείας με την χρήση RDD.

- Βρείτε τις 25 ταινίες που έχουν γίνει rate περισσότερο από τους χρήστες.
- Μετρήστε όλες τις κωμωδίες που κάποιος χρήστης βαθμολογεί τουλάχιστον 3.0.
- Βρείτε τις top 10 ρομαντικές ταινίες όσο αφορά το rating τον Δεκέμβριο.

Χρησιμοποιείστε το Spark για να διαβάσετε τα csv αρχεία σε RDDs και για να κάνετε τα απαραίτητα joins.

Μέρος 2ο

Στο δεύτερο μέρος της εργασίας θα πρέπει να χρησιμοποιήσετε την λειτουργικότητα του Spark DataFrames API για να υλοποιήσετε τα παρακάτω ερωτήματα.

- Βρείτε τις 25 ταινίες που έχουν γίνει rate περισσότερο από τους χρήστες.
- Μετρήστε όλες τις κωμωδίες που κάποιος χρήστης βαθμολογεί τουλάχιστον 3.0.
- Βρείτε τις top 10 ρομαντικές ταινίες όσο αφορά το rating τον Δεκέμβριο.
- Βρείτε τις ταινίες που οι περισσότεροι χρήστες έκαναν rate τον Δεκέμβριο.

Προσοχή πως πρέπει να υλοποιήσετε τα ερωτήματα με το DataFrames API. Για δοκιμή ορθότητας μπορείτε να τα γράψετε και σε SQL και να ζητήσετε απευθείας από το Spark να τα εκτελέσει χρησιμοποιώντας π.χ την `spark.sql`.

Τεχνολογιές

Για την υλοποίηση σας είναι υποχρεωτικό να χρησιμοποιήσετε το σύστημα Spark.

- Προσοχή, η βαθμολογία σας θα εξαρτηθεί σε σημαντικό βαθμό στο ποσοστό του κώδικα σας που θα τρέχει στο κατανεμημένο σύστημα σε αντίθεση με το ποσοστό που θα τρέχει τοπικά στον driver.
- Επιτρέπεται να χρησιμοποιήσετε είτε Java είτε Python αλλά όχι άλλες γλώσσες προγραμματισμού.

Βαθμολογία

Κριτήρια

- Καλή μοντελοποίηση, χρήση λίγων RDD και σωστό caching των ενδιάμεσων αποτελεσμάτων.
- Σωστή ονοματολογία μεταβλητών και συναρτήσεων.
- Σωστή λειτουργικότητα.
- Αποδοτική υλοποίηση.
- Εύκολη μεταγλώττιση και εκτέλεση.
- Ολοκληρωμένη και σωστή τεκμηρίωση και περιγραφή στην αναφορά.

Μπορειτε να χρησιμοποιήστε (a) Java και Maven όπως στις εργαστηριακές ασκήσεις ή (b) Python. Μπορείτε να χρησιμοποιήσετε ως σκελετό τον κώδικα του εργαστηρίου. Σε περίπτωση Python θα πρέπει να είναι τουλάχιστον έκδοση 3 και να παρέχεται και ένα requirements.txt με όλα τα dependencies.

Εκτός από τα αρχεία με τον κώδικα πρέπει να γράψετε και μια αναφορά. Η αναφορά πρέπει να εξηγεί τις διάφορες επιλογές που κάνατε, γιατί μοντελοποιήσατε έτσι το πρόβλημα καθώς και να σχολιάζει τον κώδικα σας. Η αναφορά πρέπει να είναι υποχρεωτικά σε μορφή *pdf*. Επίσης φροντίστε ο κώδικας σας να περιέχει και ένα README αρχείο που να εξηγεί με ακρίβεια πως κάνει κάποιος compile και πως το εκτελεί.