Βάσεις Δεδομένων ΙΙ

Εργαστηριακή Άσκηση 2020/21

|  |  |  |
| --- | --- | --- |
| **Ονομα** | **Επώνυμο** | **ΑΜ** |
| Ιωάννης-Νικόλαος | Τριπολίτης | 1062662 |

Βεβαιώνω ότι είμαι συγγραφέας της παρούσας εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτήν, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για το συγκεκριμένο μάθημα/σεμινάριο/πρόγραμμα σπουδών.

Έχω ενημερωθεί ότι σύμφωνα με τον εσωτερικό κανονισμό λειτουργίας του Πανεπιστημίου Πατρών άρθρο 50§6, τυχόν προσπάθεια αντιγραφής ή εν γένει φαλκίδευσης της εξεταστικής και εκπαιδευτικής διαδικασίας από οιονδήποτε εξεταζόμενο, πέραν του μηδενισμού, συνιστά βαρύ πειθαρχικό παράπτωμα.

|  |
| --- |
| Υπογραφή  27/ 5 / 2021 |

**Συνημμένα αρχεία κώδικα**

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

|  |  |  |
| --- | --- | --- |
| Αρχείο | Αφορά το ερώτημα | Περιγραφή/Σχόλιο |
| Test.ipynb | 1,21 | Περιέχει όλα τα queries για το local και cluster mode |
| pylivy.ipynb | 22 | Περιέχει όλα τα queries για το livy mode |

# Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

## Τεχνικά χαρακτηριστικά φυσικού Η/Υ που χρησιμοποιήθηκε για την εργασία

|  |  |
| --- | --- |
| **Χαρακτηριστικό** | **Τιμή** |
| CPU model | Intel i5-9300H |
| CPU clock speed | 2.4GHz |
| Physical CPU cores | 4 |
| Logical CPU cores | 8 |
| RAM | 16 |
| Secondary Storage Type | SSD |

**Δεν χρησιμοποιήθηκε κάποιο vm**

**Διευκρινήσεις ως προς τις απαντήσεις ερωτημάτων:**

**1)Το 5ο query βρίσκει τις 10 κορυφαίες ταινίες που βγήκαν το 2015 αν θέλατε για τις 10 κορυφαίες βαθμολογίες ως προς τα ratings με year=2015 βρίσκεται σε comment στο ίδιο κελί στο Test.ipynb .Επίσης εμφανίζει μόνο το 1ο tag τους.**

**2)Τα queries 8 και 10 βρίσκουν τα genres οπουδήποτε και να βρίσκονται στην στήλη ‘genres’ δεν κοιτάνε μόνο το πρώτο**

**3)Στο query 7 μόνο 2 διαφορετικοί users έκαναν rating το έτος 1995 για αυτό και δόθηκε σαν απάντηση οι κορυφαίοι αντί για τους κορυφαίους 10**

**4)Καθώς έγινε extraction του χρόνου κυκλοφορίας της ταινίας χρησιμοποιώντας regex από την στήλη title του movie.csv ήταν 24 ταινίες που δεν ακολουθούσαν το μοτίβο «όνομα ταινίας» (έτος κυκλοφορίας) αυτές οι 24 ταινίες έγιναν dropped από το dataset**

**5)Κάποια tags στο tag.csv δεν αναφέρουν ποτέ πραγματοποιήθηκαν δηλαδή το column timestamp είναι null αυτά έγιναν dropped από το dataset**

# Ερώτημα 1: Απαντήσεις ερωτημάτων

[Μην παραθέσετε στο έντυπο όλες τις επιστρεφόμενες εγγραφές! Να καταγράψετε μόνο αυτές που αναφέρει το πρότυπο.]

|  |  |
| --- | --- |
| **Ερώτημα** | **Απάντηση** |
| Δώστε το πλήθος των χρηστών που είδαν την ταινία “Jumanji”. | 22243 |
| Δώστε τα ονόματα των ταινιών που οι χρήστες χαρακτήρισαν ως “boring”. | 101 Reykjavic  12 years a slave  1408  1492:Conquest of paradise  2001: A Space Odyssey |
| Δώστε τους χρήστες που έχουν χαρακτηρίσει την ταινία ως “Bollywood” και την έχουν αξιολογήσει με βαθμό >3. | 10573  19837  23333  25004  31338 |
| Βρείτε τις 10 κορυφαίες ταινίες για κάθε έτος. | Dancemaker  Fear Strikes Out  Gate Of Heavenly Peace, The  Life Is Rosy  Married To It  My Life And Times With Antonin Artaud  Not Love, Just Frenzy  Paris Was A Woman  Take Care Of My Cat  Too Much Sleep |
| Δώστε τις ετικέτες για κάθε ταινία και το όνομα της ταινίας για το έτος 2015. | A Grain Of Truth : Borys Lankosz  A Walk In The Woods : Ken Kwapis  Advantageous : Jennifer Phang  As We Were Dreaming : Based On A Book  Average Italian : Marcello Macchia |
| Δώστε το πλήθος των ratings για κάθε ταινία. | Pulp Fiction :67310  Forrest Gump :66172  Shawshank Redemption, The :63366  Silence Of The Lambs, The :63299  Jurassic Park :59715 |
| Βρείτε τους 10 πρώτους χρήστες με τα περισσότερα rating για κάθε χρονιά. | 1995:  userId:131160:3ratings  userId:28507:1rating |
| Βρείτε τις ταινίες με τα περισσότερα ratings για κάθε κατηγορία ταινίας. | Action: Jurassic Park  Adventure: Jurassic Park  Animation : Toy Story  Children : Toy Story  Comedy: Pulp Fiction |
| Δώστε το σύνολο των χρηστών που παρακολουθούν την ίδια ταινία, την ίδια μέρα και ώρα. | 4281178 |
| Δώστε το πλήθος των ταινιών, για κάθε κατηγορία, που οι χρήστες χαρακτήρισαν ως “funny” και με rating > 3.5. | Action : 363  Adventure : 401  Anime : 253  Children : 255  Comedy : 1329 |

# Ερώτημα 2: Σύγκριση επιδόσεων σε single node/virtual cluster/Livy

## Ρυθμίσεις virtual cluster

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A/A** | **Executor cores** | **Executor mem** | **Driver cores** | **Driver mem** |
| 1 | 1 | 512m | 1 | 1G |
| 2 | 2 | 1G | 1 | 1G |
| 3 | 2 | 1G | 2 | 1G |

## Χρόνοι εκτέλεσης

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ερώτημα | Local | Virtual 1 | Virtual 2 | Virtual 3 | Livy |
| 1 | 5.48s | 16.70s | 12.50s | 12.60s | 11.00s |
| 2 | 2.20s | 13.30s | 10.50s | 8.35s | 5.00s |
| 3 | 6.70s | 13.00s | 10.00s | 10.00s | 14.00s |
| 4 | 12.10s | 31.40s | 23.40s | 23.40s | 27.00s |
| 5 | 1.14s | 3.40s | 2.40s | 2.40s | 4.00s |
| 6 | 7.20s | 14.30s | 11.20s | 11.20s | 12.00s |
| 7 | 12.00s | 27.00s | 22.00s | 21.00s | 60.00s |
| 8 | 30.41s | 38.10s | 39.57s | 34.40s | 50.00s |
| 9 | 35.20s | 63.50s | 53.50s | 48.50s | 90.00s |
| 10 | 12.20s | 22.30s | 19.30s | 20.30s | 24.00s |

## Ανάλυση αποτελεσμάτων

Chart, line chart

Description automatically generated

[συνεχίζετε με το δικό σας σχολιασμό σχετικά με την αιτία των παρατηρούμενων διαφορών]

Παρατηρώ πως το local είναι καθολικά γρηγορότερο σε όλα τα queries σε σχέση με όλα τα υπόλοιπα modes. Αυτό συμβαίνει όσον αφορά τα clusters γιατί κάνω προσομοίωση σε έναν μόνον υπολογιστή διάφορους το οποίο είναι προφανές πως θα έχει κάποιο overhead,εκτός αυτού εφόσον πλέον διαθέτουμε πολλούς “υπολογιστές” πρέπει πλέον να βεβαιωθεί ο συγχρονισμός του το οποίο δημιουργεί και αυτό μια καθυστέρηση. Όσον αφορά το Livy οι εντολές πρέπει να σταλούν σε έναν άλλον υπολογιστή μέσω του δικτύου κάτι το οποίο προφανώς δημιουργεί μια καθυστέρηση και στην συνέχεια να τρέξει σε κάποιο υλικό το οποίο υπάρχει η πιθανότητα να είναι χειρότερο από το local. Εκτός αυτών γνωρίζουμε πως στο livy υπάρχουν ένας master και 2 worker αρά και εκεί πρέπει να βεβαιωθεί ο συγχρονισμός τους όπως και στα cluster. Όσον αφορά τώρα την σύγκριση των cluster μεταξύ τους παρατηρώ πως το 1ο είναι το πιο αργό και συνήθως το 2ο με το 3ο έχουν την ιδιά ταχύτητα με μερικές φορές το 3ο να είναι γρηγορότερο για πολύ λίγο. Κάτι το οποίο είναι λογικό καθώς στο 1ο διαθέτουμε λιγότερους πόρους σε σχέση με το 2ο και 3ο cluster. To livy είναι γενικά με διαφορά το αργότερο και αυτό συμβαίνει γιατί πρέπει οι εντολές να αποσταλούν μέσω δικτύου για να της εκτελέσει.

# Βιβλιογραφία

Dataset used: <https://www.kaggle.com/grouplens/movielens-20m-dataset>