Βάσεις Δεδομένων ΙΙ

Εργαστηριακή Άσκηση 2020/21

Όνομα	**Επώνυμο**	**ΑΜ**
Νικηφόρος – Γεώργιος	Παπαγεωργίου	1059633
Νικόλαος	Σταμόπουλος	1057764

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

22 / 09 / 2021

Υπογραφή

22 / 09 / 2021

Συνημμένα αρχεία κώδικα

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

Αρχείο	**Αφορά το ερώτημα**	**Περιγραφή/Σχόλιο**
queryX.ipynb	1	Περιέχει τον κώδικα για το ερώτημα X, όπου $X \in [1, 10]$

Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

Τεχνικά χαρακτηριστικά φυσικού Η/Υ που χρησιμοποιήθηκε για την εργασία

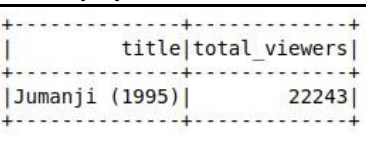
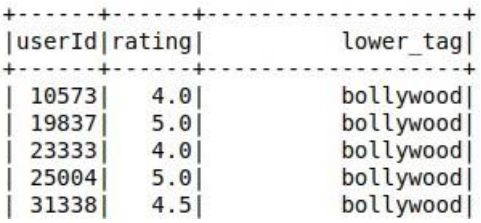
Χαρακτηριστικό	**Τιμή**
CPU model	AMD Ryzen 5 2600 Six-Core Processor
CPU clock speed	3.4GHz
Physical CPU cores	6
Logical CPU cores	12
RAM	16Gb
Secondary Storage Type	SSD

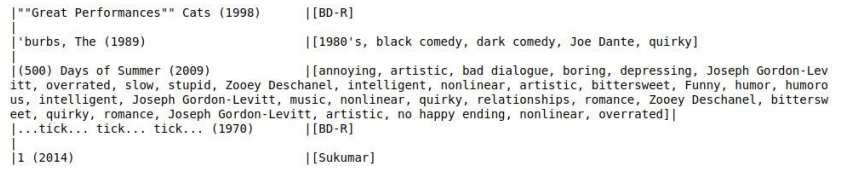
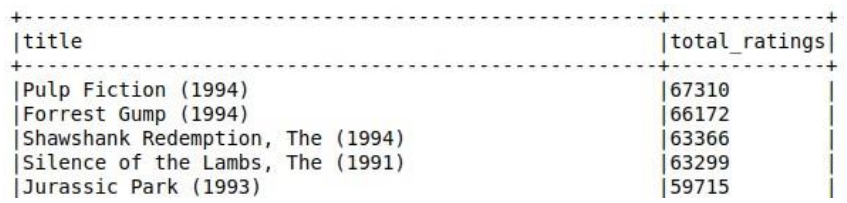
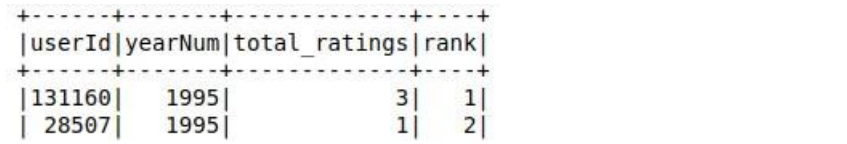
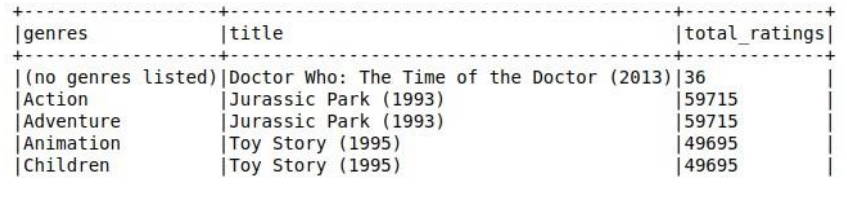
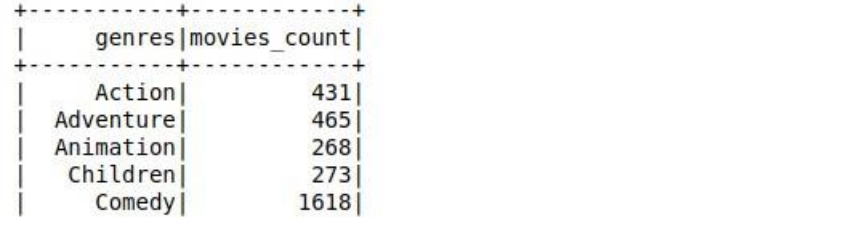
Τεχνικά χαρακτηριστικά εικονικής μηχανής (VM) που χρησιμοποιήθηκε για την εργασία

Χαρακτηριστικό	**Τιμή**
CPU cores	6
Execution cap	100%
RAM	9Gb
VM OS	Ubuntu 20.04.2 LTS
VM software	VirtualBox
Host OS	Windows 10

Ερώτημα 1: Απαντήσεις ερωτημάτων

Σημείωση: Τα ζητούμενα αποτελέσματα παρατίθενται σε μορφή screenshot, για την καλύτερη απεικόνισή τους.

Ερώτημα	**Απάντηση**
Δώστε το πλήθος των χρηστών που είδαν την ταινία “Jumanji”.	
Δώστε τα ονόματα των ταινιών που οι χρήστες χαρακτήρισαν ως “boring”.	
Δώστε τους χρήστες που έχουν χαρακτηρίσει την ταινία ως “Bollywood” και την έχουν αξιολογήσει με βαθμό >3.	

Βρείτε τις 10 κορυφαίες ταινίες για κάθε έτος.	
Δώστε τις ετικέτες για κάθε ταινία και το όνομα της ταινίας για το έτος 2015.	
Δώστε το πλήθος των ratings για κάθε ταινία.	
Βρείτε τους 10 πρώτους χρήστες με τα περισσότερα rating για κάθε χρονιά.	
Βρείτε τις ταινίες με τα περισσότερα ratings για κάθε κατηγορία ταινίας.	
Δώστε το σύνολο των χρηστών που παρακολουθούν την ίδια ταινία, την ίδια μέρα και ώρα.	
Δώστε το πλήθος των ταινιών, για κάθε κατηγορία, που οι χρήστες χαρακτήρισαν ως “funny” και με rating > 3.5.	

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

Ρυθμίσεις virtual cluster

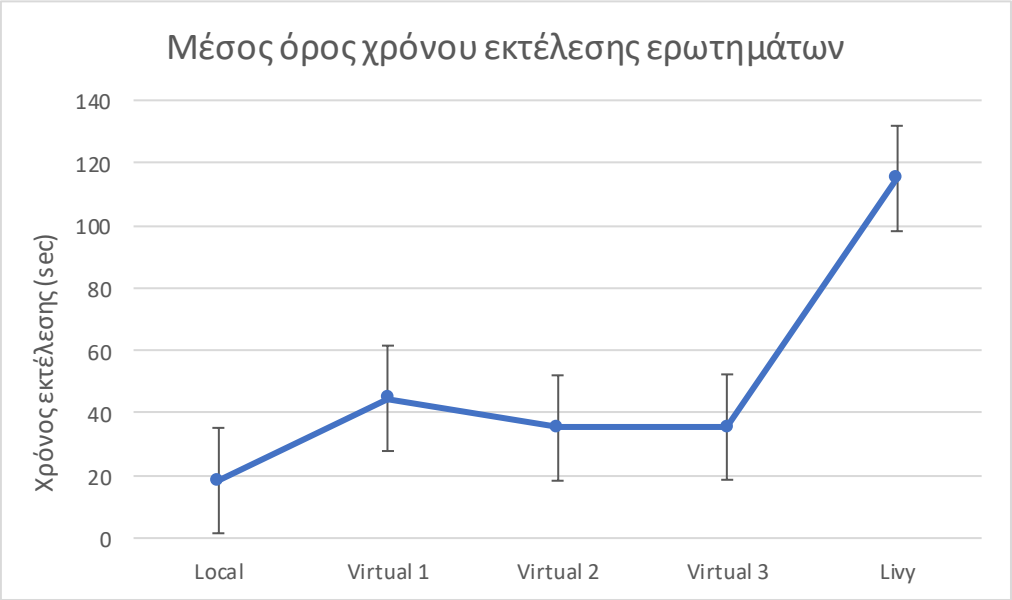
A/A	Executor cores	Executor mem	Driver cores	Driver mem
1	1	1G	1	1G
2	2	2G	1	1G
3	2	2G	2	2G

Χρόνοι εκτέλεσης

Σημείωση: Οι παρακάτω χρόνοι εκτέλεσης μετρήθηκαν σε δευτερόλεπτα (seconds). Η χρονομέτρηση έγινε με χρήση της βιβλιοθήκης sparkMeasure.

Ερώτημα	Local	Virtual 1	Virtual 2	Virtual 3	Livy
1	14	40	28	29	120
2	5	19	15	15	15
3	20	46	40	38	123
4	26	55	44	52	180
5	4	18	15	13	16
6	15	40	36	34	108
7	27	53	41	41	168
8	25	48	41	38	119
9	32	90	62	63	181
10	16	38	30	32	120

Ανάλυση αποτελεσμάτων



Μετά την χρονομέτρηση και ανάλυση των αποτελεσμάτων εξάγουμε τις παρακάτω παρατηρήσεις:

- Σε single node μηχάνημα (local) πετυχαίνουμε τάχιστη εκτέλεση των ερωτημάτων, καθώς χρησιμοποιείται όλη η υπολογιστική δύναμη που έχουμε αναθέσει στο VM.
- Ανάμεσα στα Virtual 1, Virtual 2 και Virtual 3, παρατηρούμε πως το πρώτο καταναλώνει περισσότερο χρόνο για την εκτέλεση των ερωτημάτων σε σχέση με τα άλλα δύο και αυτό λογικά οφείλεται στην ανάθεση μόνο ενός πυρήνα (core) για κάθε worker. Επίσης παρατηρούμε πως ανάμεσα στα Virtual 2 και 3 οι διαφορές είναι μηδαμινές, συνεπώς η αύξηση των πυρήνων και της μνήμης που δεσμεύει ο driver στο Virtual 3 δεν έφερε καλύτερα αποτελέσματα από το 2.
- Η εκτέλεση των ερωτημάτων στον Livy server καταναλώνει τον μέγιστο χρόνο.
- Όσον αφορά τα ερωτήματα, παρατηρούμε πως συγκεκριμένα τα ερωτήματα 2 και 5 τρέχουν πάντα σε πολύ λιγότερο χρόνο σε σύγκριση με τα υπόλοιπα. Πιθανολογούμε πως αυτό οφείλεται στο γεγονός ότι αυτά τα δύο ερωτήματα δεν κάνουν χρήση του αρχείου rating.csv, το οποίο περιέχει το μεγαλύτερο πλήθος εγγραφών απ' όλα και συνεπώς είναι πιο «ακριβό» σε πλήθος πράξεων.

Βιβλιογραφία

1. *PySpark 3.1.2 Documentation*. http://spark.apache.org/docs/latest/api/python/
2. Α. Κομνηνός. *Φροντιστήριο 6 – Εισαγωγή στο Apache Spark*. https://eclass.upatras.gr/modules/document/index.php?course=CEID1176&openDir=/5e6f65ear83d/60756472Ab51
3. Nishant Bahri. *Movie Lens Data Analysis Using PySpark [for beginners]*. https://medium.com/analytics-vidhya/movie-lens-data-analysis-using-pyspark-for-beginners-9c0f5f21eaf5
4. Mauro Krikorian. *Movie Data Statistics with Apache Spark*. https://medium.com/southworks/movie-data-statistics-with-apache-spark-58c2ef8fe452