

Βάσεις Δεδομένων II
Εργαστηριακή Άσκηση 2020/21

Όνομα	Επώνυμο	ΑΜ
Νικόλαος	Σκαμνέλος	1041878 ή 236201
Ιωάννης	Ακαρέπης	1054368

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



4 / 6 / 2021

Υπογραφή



4 / 6 / 2021

Συνημμένα αρχεία κώδικα

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

Αρχείο	Αφορά το ερώτημα	Περιγραφή/Σχόλιο
Local_PC/erotima_1.py	1.1	Περιέχει το πρώτο query
Local_PC/erotima_2.py	1.2	Περιέχει τα δεύτερο query
Local_PC/erotima_3.py	1.3	Περιέχει το τρίτο query
Local_PC/erotima_4.py	1.4	Περιέχει το τέταρτο query
Local_PC/erotima_5.py	1.5	Περιέχει το πέμπτο query

Local_PC/erotima_6.py	1.6	Περιέχει το έκτο query
Local_PC/erotima_7.py	1.7	Περιέχει το έβδομο query
Local_PC/erotima_8.py	1.8	Περιέχει το όγδοο query
Local_PC/erotima_9.py	1.9	Περιέχει το ένατο query
Local_PC/erotima_10.py	1.10	Περιέχει το δέκατο query
Livy_Server/erotima_1_LIVY.py	2.2.1	Περιέχει το πρώτο query για τον Livy server
Livy_Server/erotima_2_LIVY.py	2.2.2	Περιέχει το δεύτερο query για τον Livy server
Livy_Server/erotima_3_LIVY.py	2.2.3	Περιέχει το τρίτο query για τον Livy server
Livy_Server/erotima_4_LIVY.py	2.2.4	Περιέχει το τέταρτο query για τον Livy server
Livy_Server/erotima_5_LIVY.py	2.2.5	Περιέχει το πέμπτο query για τον Livy server
Livy_Server/erotima_6_LIVY.py	2.2.6	Περιέχει το έκτο query για τον Livy server
Livy_Server/erotima_7_LIVY.py	2.2.7	Περιέχει το έβδομο query για τον Livy server
Livy_Server/erotima_8_LIVY.py	2.2.8	Περιέχει το όγδοο query για τον Livy server
Livy_Server/erotima_9_LIVY.py	2.2.9	Περιέχει το ένατο query για τον Livy server
Livy_Server/erotima_10_LIVY.py	2.2.10	Περιέχει το δέκατο query για τον Livy server

Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

Τεχνικά χαρακτηριστικά φυσικού Η/Υ που χρησιμοποιήθηκε για την εργασία

Για τα ερωτήματα 1,3,5,7,9:

Χαρακτηριστικό	Τιμή
CPU model	Intel i3-8100
CPU clock speed	3.60GHz
Physical CPU cores	4
Logical CPU cores	4
RAM	16 GB
Secondary Storage Type	HDD/SSD/M.2 SSD

Για τα ερωτήματα 2,4,6,8,10:

Χαρακτηριστικό	Τιμή
CPU model	AMD Ryzen 3 3200G
CPU clock speed	3.60GHz
Physical CPU cores	4
Logical CPU cores	4
RAM	16 GB
Secondary Storage Type	HDD/SSD

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

Χαρακτηριστικό	Τιμή
CPU cores	4
Execution cap	100%
RAM	8Gb
VM OS	Ubuntu 20.04
VM software	VirtualBox
Host OS	Windows 10

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".	Awake (2007) Avengers, The (2012) Avatar (2009) Austin Powers: International Man of Mystery (1997) August: Osage County (2013)

Δώστε τους χρήστες που έχουν χαρακτηρίσει την ταινία ως "Bollywood" και την έχουν αξιολογήσει με βαθμό >3.	User IDs: 65, 910, 1741, 1741, 8513
Βρείτε τις 10 κορυφαίες ταινίες για κάθε έτος.	Alaska: Silence & Solitude (2005) Adam's Apples (Adams æbler) (2005) Ax, The (couperet, Le) (2005) After Innocence (2005) All the Invisible Children (2005)
Δώστε τις ετικέτες για κάθε ταινία και το όνομα της ταινίας για το έτος 2015.	A Grain of Truth (2015) A Walk in the Woods (2015) Advantageous (2015) As We Were Dreaming (2015) Average Italian (2015)
Δώστε το πλήθος των ratings για κάθε ταινία.	Pulp Fiction (1994) Forrest Gump (1994) Shawshank Redemption, The (1994) Silence of the Lambs, The (1991) Jurassic Park (1993)
Βρείτε τους 10 πρώτους χρήστες με τα περισσότερα rating για κάθε χρονιά.	1995: 131160, 28507
Βρείτε τις ταινίες με τα περισσότερα ratings για κάθε κατηγορία ταινίας.	Action -> Jurassic Park (1993) Adventure -> Jurassic Park (1993) Animation -> Toy Story (1995) Children -> Toy Story (1995) Comedy -> Pulp Fiction (1994)
Δώστε το σύνολο των χρηστών που παρακολουθούν την ίδια ταινία, την ίδια μέρα και ώρα.	6820
Δώστε το πλήθος των ταινιών, για κάθε κατηγορία, που οι χρήστες χαρακτήρισαν ως "funny" και με rating > 3.5.	Action -> 261 Adventure -> 335 Animation -> 253 Children -> 228 Comedy -> 1036

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

Ρυθμίσεις virtual cluster

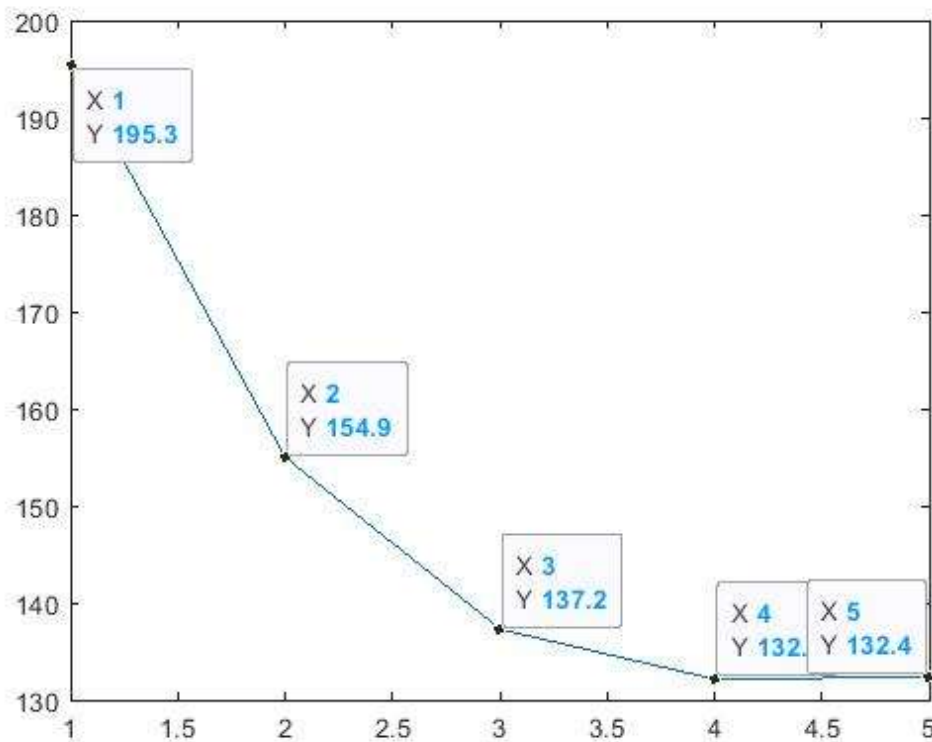
A/A	Executor cores	Executor mem	Driver cores	Driver mem
1	1	0.5G	1	1G
2	2	1G	1	1G
3	2	1G	2	1G

Χρόνοι εκτέλεσης

Ερώτημα	Local	Virtual 1	Virtual 2	Virtual 3	Livy
1	7 sec	48 sec	37 sec	38 sec	1.3 min
2	2 sec	43 sec	31 sec	30 sec	1.3 min

3	4 sec	41 sec	31 sec	30 sec	35 sec
4	11 min	6.5 min	5.5 min	5.9 min	4.6 min
5	0.2 sec	34 sec	26 sec	25 sec	1.1 min
6	21 sec	51 sec	41 sec	41 sec	47 sec
7	2.7 min	4 min	3.8 min	3.3 min	■
8	6.9 min	3.9 min	3.7 min	3.5 min	3.1 min
9	23 sec	1.2 min	1.1 min	1 min	2.4 min
10	11 min	6.6 min	6.0 min	5.6 min	4.9 min

Ανάλυση αποτελεσμάτων



Γραφική παράσταση μέσου όρου χρόνου εκτέλεσης ερωτημάτων (μέσω Matlab)

Στην ανωτέρω γραφική παράσταση ο άξονας X ορίζεται ως εξής

- X = 1 → Local
- X = 2 → Virtual 1
- X = 3 → Virtual 2
- X = 4 → Virtual 3
- X = 5 → Livy Server

Αρχικά μπορούμε εύκολα να παρατηρήσουμε ότι υπάρχει βελτίωση χρόνου εκτέλεσης όταν το πρόγραμμα εκτελείται στα συστήματα **Virtual 3** και **Livy**. Αυτό ήταν αναμενόμενο καθώς και στα δύο συστήματα χρησιμοποιούνται περισσότεροι πόροι και συνεπώς έχουμε μεγαλύτερη υπολογιστική ισχύ.

Όπως φαίνεται και στον πίνακα με τους χρόνους εκτέλεσης, υπάρχουν ερωτήματα για τα οποία το Local σύστημα χρειάστηκε πολύ λιγότερο χρόνο εκτέλεσης έναντι των άλλων servers

(για παράδειγμα στο πέμπτο ερώτημα). Ο λόγος που ενδεχομένως συμβαίνει αυτό είναι επειδή τα `virtuals` καθώς και ο `Livy` είναι `servers` διαφορετικοί του προσωπικού μας υπολογιστή και επικοινωνούν μέσω ενός δικτύου, συνεπώς υπάρχει κάποιος απαιτούμενος χρόνος αποστολής του προγράμματος προς εκτέλεση και ύστερα επιστροφής των αποτελεσμάτων πίσω στην συσκευή μας. Αυτό σημαίνει ότι για `queries` που απαιτούν μικρό χρόνο εκτέλεσης (δηλαδή απλά **queries** με λίγα **φίλτρα** και ελάχιστα **join operations**) ίσως είναι προτιμότερος ο προσωπικός μας υπολογιστής. Αντιθέτως, για πιο πολύπλοκα **queries** (π.χ ερωτήματα 4, 10 με χρήση βρόχου **for**) παρατηρούμε ότι χρειαζόμαστε περισσότερους πόρους και η χρήση των `servers` αυτών είναι προτιμότερη για την εκτέλεσή τους.

Το ερώτημα 7 δεν μπορούσε να εκτελεστεί στον `livy server` επειδή πάντα ξεπερνούσε τον μέγιστο χρόνο που το `session` ήταν ενεργό. Επομένως, αποφασίσαμε για την γραφική παράσταση να χρησιμοποιήσουμε στην θέση του τον μέσο όρο της στήλης (όλων των ερωτημάτων του `livy server`).

Βιβλιογραφία

- [1] "Spark Apache Documentation," Apache, [Online]. Available: <https://spark.apache.org/docs/3.1.1/api/python/reference/>.
- [2] E. Heitman, "A Neanderthal's Guide to Apache Spark in Python," 14 June 2019. [Online]. Available: <https://towardsdatascience.com/a-neanderthals-guide-to-apache-spark-in-python-9ef1f156d427>.
- [3] A. Kumar, "Guide to install Spark and use PySpark from Jupyter in Windows," 19 March 2019. [Online]. Available: <https://bigdata-madesimple.com/guide-to-install-spark-and-use-pyspark-from-jupyter-in-windows/>.