


Βάσεις Δεδομένων II
Εργαστηριακή Άσκηση 2020/21

Όνομα	Επώνυμο	ΑΜ
Αριάδνη	Μαχιά	1059556
Αθηνά	Φουσέκη	1059623

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

10 / 06 / 2021

Υπογραφή

10 / 06 / 2021

Συνημμένα αρχεία κώδικα

Μαζί με την παρούσα αναφορά υποβάλλουμε τα παρακάτω αρχεία κώδικα

Αρχείο	Αφορά το ερώτημα	Περιγραφή/Σχόλιο
query1.ipynb	1	Περιέχει όλα τα ερωτήματα για το ερώτημα 1.
query2.ipynb	2	Περιέχει όλα τα ερωτήματα για το ερώτημα 2.
query3.ipynb	3	Περιέχει όλα τα ερωτήματα για το ερώτημα 3.
query4.ipynb	4	Περιέχει όλα τα ερωτήματα για το ερώτημα 4.

query5.ipynb	5	Περιέχει όλα τα ερωτήματα για το ερώτημα 5.
query6.ipynb	6	Περιέχει όλα τα ερωτήματα για το ερώτημα 6.
query7.ipynb	7	Περιέχει όλα τα ερωτήματα για το ερώτημα 7.
query8.ipynb	8	Περιέχει όλα τα ερωτήματα για το ερώτημα 8.
query9.ipynb	9	Περιέχει όλα τα ερωτήματα για το ερώτημα 9.
query10.ipynb	10	Περιέχει όλα τα ερωτήματα για το ερώτημα 10.
db2_graph_code.py	-	Περιέχει τον κώδικα που εκτελέσαμε για τη δημιουργία του γραφήματος

Τεχνικά χαρακτηριστικά περιβάλλοντος λειτουργίας

Τεχνικά χαρακτηριστικά φυσικού Η/Υ που χρησιμοποιήθηκε για την εργασία

Χαρακτηριστικό	Τιμή
CPU model	Intel i7-9750H
CPU clock speed	2.6GHz
Physical CPU cores	6
Logical CPU cores	12
RAM	16
Secondary Storage Type	SSD

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

Χαρακτηριστικό	Τιμή
CPU cores	6
Execution cap	100%
RAM	6Gb
VM OS	Ubuntu 20.04
VM software	VirtualBox
Host OS	macOS Big Sur 11.4

Ερώτημα 1: Απαντήσεις ερωτημάτων

Διευκρινήσεις: Στο ερώτημα 8 και 10, δεν κρατήσαμε μόνο την πρώτη κατηγορία που ανήκει η ταινία, αλλά πραγματοποιήσαμε το ερώτημα για όλες τις κατηγορίες κάθε ταινίας. Στο ερώτημα 2 και 3, έχουμε συμπεριλάβει και του σύνθετους όρους με τις λέξεις «boring» και «Bollywood» αντίστοιχα. Σε σχόλια υπάρχουν οι αντίστοιχες εντολές χωρίς τους σύνθετους όρους. (Πιστεύαμε ότι ήταν στην δικιά μας ευχέρεια εάν θα αγνοήσουμε τελικά τους σύνθετους όρους ή όχι. Σε περίπτωση που καταλάβαμε λάθος, βάλαμε την εναλλακτική εντολή σε σχόλιο).

Ερώτημα	Απάντηση
Δώστε το πλήθος των χρηστών που είδαν την ταινία "Jumanji".	22243
Δώστε τα ονόματα των ταινιών που οι χρήστες χαρακτήρισαν ως "boring".	(500) Days of Summer (2009) 101 Reykjavik (101 Reykjavík) (2000) 12 Years a Slave (2013) 1408 (2007) 1492: Conquest of Paradise (1992)
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.	Έτος 2005: 1) Before the Fall (NaPoIA - Elite für den Führer) (2004) 2) Dancemaker (1998) 3) Fear Strikes Out (1957) 4) Gate of Heavenly Peace, The (1995) 5) Life Is Rosy (a.k.a. Life Is Beautiful) (Vie est belle, La) (1987) 6) Married to It (1991) 7) My Life and Times With Antonin Artaud (En compagnie d'Antonin Artaud) (1993) 8) Not Love, Just Frenzy (Más que amor, frenesí) (1996) 9) Paris Was a Woman (1995) 10) Take Care of My Cat (Goyangileul butaghae) (2001)

Δώστε τις ετικέτες για κάθε ταινία και το όνομα της ταινίας για το έτος 2015.	<pre> +-----+-----+ title tag_lc +-----+-----+ A Grain of Truth ... social commentary A Grain of Truth ... robert wieckiewicz A Grain of Truth ... dark A Grain of Truth ... sandomierz A Grain of Truth ... do zassania A Grain of Truth ... investigation A Grain of Truth ... polish-jewish his... A Grain of Truth ... abel korzeniowski A Grain of Truth ... jerzy trela A Grain of Truth ... borys lankosz A Grain of Truth ... crime A Grain of Truth ... poland A Grain of Truth ... thriller A Walk in the Woo... ken kwapis Advantageous (2015) jennifer phang As We Were Dreami... based on a book Average Italian (... drugs Average Italian (... marcello macchia +-----+-----+ </pre>
Δώστε το πλήθος των ratings για κάθε ταινία.	<pre> +-----+-----+-----+ movieId title numOfRatings +-----+-----+-----+ 296 Pulp Fiction (1994) 67310 356 Forrest Gump (1994) 66172 318 Shawshank Redempt... 63366 593 Silence of the La... 63299 480 Jurassic Park (1993) 59715 +-----+-----+-----+ </pre>
Βρείτε τους 10 πρώτους χρήστες με τα περισσότερα rating για κάθε χρονιά.	<pre> +-----+-----+-----+ userId year count +-----+-----+-----+ 28507 1995 1 131160 1995 3 +-----+-----+-----+ </pre>
Βρείτε τις ταινίες με τα περισσότερα ratings για κάθε κατηγορία ταινίας.	<pre> +-----+-----+-----+ title lgenres max(max(numOfRatings)) +-----+-----+-----+ Jurassic Park (1993) Action 59715 Jurassic Park (1993) Adventure 59715 Toy Story (1995) Animation 49695 Toy Story (1995) Children 49695 Pulp Fiction (1994) Comedy 67310 +-----+-----+-----+ </pre>
Δώστε το σύνολο των χρηστών που παρακολουθούν την ίδια ταινία, την ίδια μέρα και ώρα.	4280240

Δώστε το πλήθος των ταινιών, για κάθε κατηγορία, που οι χρήστες χαρακτήρισαν ως “funny” και με rating > 3.5.	<pre> +-----+-----+ lgenres count +-----+-----+ Action 125 Adventure 127 Animation 73 Children 88 Comedy 540 +-----+-----+ </pre>
--	--

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

Ρυθμίσεις virtual cluster

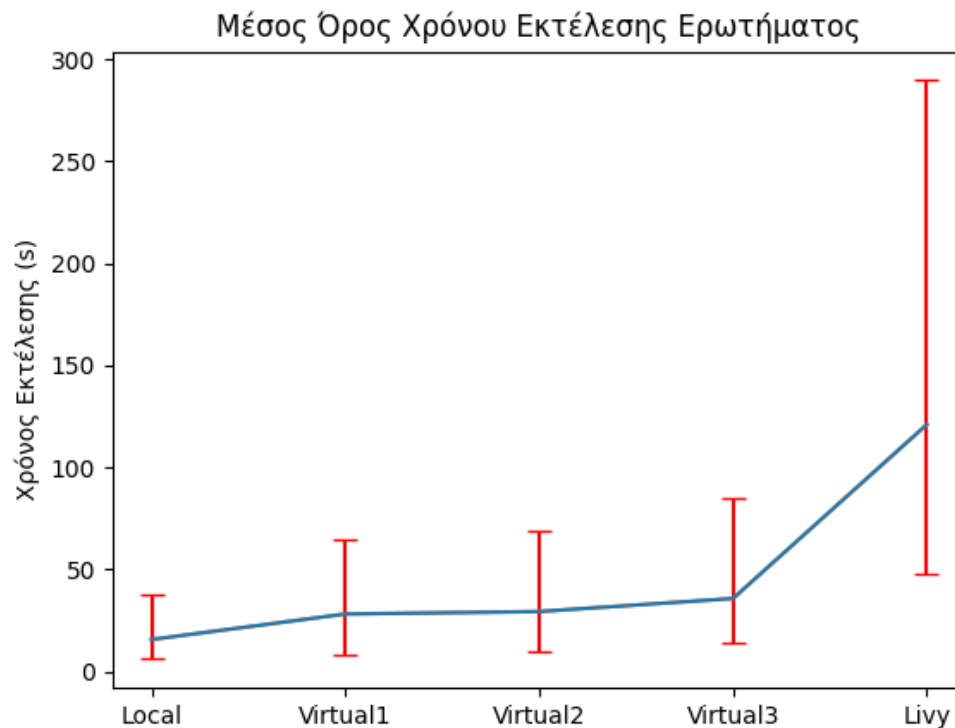
A/A	Executor cores	Executor mem	Driver cores	Driver mem
1	2	1G	2	1G
2	2	1G	1	1G
3	1	512M	1	1G

Χρόνοι εκτέλεσης

(s= δευτερόλεπτα και min=λεπτά)

Ερώτημα	Local	Virtual 1	Virtual 2	Virtual 3	Livy
1	10s	22s	18s	24s	1.9min
2	4s	10s	11s	10s	14s
3	12s	25s	29s	31s	1.9 min
4	18s	43s	39s	60s	3.3min
5	4s	13s	6s	6s	14s
6	19s	30s	43s	56s	2.1min
7	15s	28s	30s	35s	3.1min
8	25s	33s	40s	38s	2.0min
9	30s	47s	48s	60s	3.4min
10	19s	30s	29s	37s	2.0min

Ανάλυση αποτελεσμάτων



(Μπάρες σφάλματος στο διάστημα εμπιστοσύνης 95%)

[Μ.Ο.: Local= 15.6 | Virtual1= 28.1 | Virtual2= 29.3 | Virtual3= 35.7 | Livy= 121.0]

Φαίνεται από το παραπάνω διάγραμμα πως οι πιο γρήγοροι χρόνοι είναι όταν χρησιμοποιούνται όλοι οι διαθέσιμοι πόροι που διαθέτει το σύστημα και ο driver node ενεργεί τόσο ως master όσο και ως worker (Local). Ενώ όταν περιορίζεται η μνήμη ή οι πυρήνες που χρησιμοποιούνται, ο χρόνος εκτέλεσης είναι μεγαλύτερος. Στα εικονικά cluster μηχανήματα (virtual 1-3), δεν παρατηρούνται ιδιαίτερες διαφορές μεταξύ των χρόνων εκτέλεσης. Η πιο «σημαντική» διαφορά που παρατηρούμε μοιάζει να οφείλεται στην executor memory. Παρόλο που έχουμε διαθέσει λιγότερους πόρους στα εικονικά cluster μηχανήματα, οι χρόνοι είναι αρκετά κοντά με αυτούς του Local, διότι οι διεργασίες εκτελούνται παράλληλα (όσο γίνεται). Τέλος, παρατηρούμε ότι οι χρόνοι εκτέλεσης των ερωτημάτων στο Livy Server είναι σημαντικά μεγαλύτεροι, πράγμα που οφείλεται στην απομακρυσμένη ανταλλαγή δεδομένων μεταξύ του server και του μηχανήματός μας. Γι' αυτό ο Livy έχει τους μεγαλύτερους χρόνους.

Βιβλιογραφία

- <http://spark.apache.org/docs/3.1.1/api/python/reference/pyspark.sql.html>
- https://mallikarjuna_g.gitbooks.io/spark/content/spark-standalone-example-2-workers-on-1-node-cluster.html
- <https://spark.apache.org/docs/latest/spark-standalone.html#cluster-launch-scripts>
- <https://spark.apache.org/docs/latest/configuration.html#application-properties>