Προχωρημένα Θέματα Βάσεων Δεδομένων

Εξαμηνιαία Εργασία

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, ΕΜΠ

Ομάδα 7

Οικονόμου Νικόλαος (03120014)

Ραφτόπουλος Μιχαήλ (03120114)

Αποθετήριο GitHub

Ιανουάριος 2025

Περιεχόμενα

Query 1

Εκτελέστηκε το query 1, τόσο με το DataFrame API, όσο και με το RDD API του Spark. Συγκρίνοντας τον χρόνο των δύο υλοποιήσεων, η υλοποίηση με DataFrame API αποδείχθηκε ταχύτερη[1]. Το αποτέλεσμα αυτό είναι αναμενόμενο, καθώς το DataFrame API αποτελεί μια υψηλωτερη αφαιρετικά διεπαφή, με πολλές βελτιστοποιήσεις να πραγματοποιούνται στο υπόβαθρο (από τον Catalyst Optimizer). Με το RDD API μπορούμε θεωρητικά να πετύχουμε την ίδια αποδοτικότητα (αφού και το DataFrame API χρησιμοποιεί στο υπόβαθρο RDDs), αλλά απαιτείται πολλή εμειρία και προσεκτικός χειρισμός.

Age Group	Count
Adults	121093
Young Adults	33605
Children	15928
Elderly	5985

Πίνακας 1: Αποτελέσματα query 1.

Query 2

(α) Υπολογίστηκαν για κάθε έτος τα 3 Αστυνομικά Τμήματα με το υψηλότερο ποσοστό κλεισμένων υποθέσεων και ταξινομήθηκαν ανά έτος και ανά ποσοστό. Χρησιμοποιήθηκαν δύο διαφορετικά APIs του Spark: το DataFrame API και το SQL API. Μετά από αρκετές επαναλήψεις, η υλοποίσηση με SQL API αποδείχθηκε ταχύτερη από το αυτήν με DataFrame API. Θεωρητικά δε θα αναμέναμε ουσιαστική διαφορά μεταξύ των δύο, καθώς αποτελούν απλά διαφορετικές διεπαφές του ίδιου optimizer. Η απόκλιση λοιπόν των δύο υλοποιήσεων μάλλον οφείλεται στον τρόπο που αυτές είναι γραμμένες, με τον κώδικα σε DataFrames να οδηγεί σε περισσότερα operations.

(β) Σε αυτό το σημείο, έγινε η μετατροπή των δεδομένων εισόδου από `.csv` σε `.parquet`. Εκτελέστηκε η ίδια υλοποίηση SQL, χρησιμοπιώντας το δεύτερο format. Η υλοποίηση με τα δεδομένα σε `.parquet` ήταν ταχύτερη. Αυτό είναι αναμενόμενο, καθώς το `.parquet` αποτελεί έναν τύπο αρχείου σχεδιασμένο για κατανεμημένα filesystems.

year	AREA NAME	closed_rate	#
2010	Rampart	32.84713448949121	1
2010	Olympic	31.515289821999087	2
2010	Harbor	29.36028339237341	3
2011	Olympic	35.0400600901352	1
2011	Rampart	32.496447181430604	2
...			

Πίνακας 2: Αποτελέσματα query 2 (φαίνονται μόνο οι πρώτες γραμμές).

[1]Σε αυτό το σημείο να αναφερθεί ότι στις διάφορες δοκιμές του χρόνου εκτέλεσης των queries τα αποτελέσματα παρουσιάζαν πολύ μεγάλη διακύμανση, που πιθανώς οφείλεται στο cloud περιβάλλον εκτέλεσης. Οι σχετικοί χρόνοι όμως ανάμεσα στις υλοποιήσεις παρέμεναν σταθεροί. Για το λόγο αυτό, στο μεγαλύτερο μέρος της εργασίας αναφερόμαστε σε σχετικούς χρόνους και όχι απόλυτα νούμερα.

Query 3

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

COMM	...	Median Income Per Person	...	Crimes Per Person Ration
Elysian Park		13871.32276		1.08487
Longwood		13420.05226		0.73017
Cadillac-Corning		19572.7847		0.66692
...				

Πίνακας 3: Αποτελέσματα query 3 (φαίνονται μόνο οι πρώτες γραμμές και επιλεγμένες στήλες).

Query 4

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

vict_descent	total_victims	...
White	8429	
Other	1125	
Hispanic/Latin/Mexican	868	
Unknown	651	
...		

Πίνακας 4: Αποτελέσματα query 4 για την περίπτωση των περιοχών με το υψηλότερο κατά κεφαλήν εισόδημα (φαίνονται μόνο οι πρώτες γραμμές και επιλεγμένες στήλες).

Query 5

Πραγματοποιήθηκε η υλοποίηση του Query 5 και εκτελέστηκε με τα 3 ζητούμενα configuration. Για τον υπολογισμό του χρόνου εκτέλεσης, το πρόγραμμα επαναλήφθηκε 10 φορές για κάθε configuration και υπολογίστηκε ο μέσος χρόνος για κάθε περίπτωση. Παρακάτω φαίνονται τα αποτελέσματα:

- 2 executors $\times$ 4 cores/8GB memory: **9.82s**

- 4 executors × 2 cores/4GB memory: **7.60s**
- 8 executors × 1 core/2GB memory: **7.25s**

Βλέπουμε ότι πολλοί «αδύναμοι» executors αποδίδουν καλύτερα απ' ο, τι λίγοι «ισχυροί». Αυτή η παρατήρηση δεν αποτελεί έκπληξη. Από τη «φύση» της, η κατανεμημένη βάση δεδομένων χρησιμοποιεί κατά κόρον την παραλληλοποίηση. Οι υψηλά παραλληλοποιήσιμες αυτές ενέργειες εποφελούνται από περισσότερες, έστω και πιο «αδύναμες», διεργσίες.

DIVISION	crime_count	mean_distance
HOLLYWOOD	213080	2.269
VAN NUYS	211457	3.181
WILSHIRE	198150	2.921
SOUTHWEST	186742	2.395
...		

Πίνακας 5: Αποτελέσματα query 5 (φαίνονται μόνο οι πρώτες γραμμές).