Προχωρημένα Θέματα Βάσεων Δεδομένων

Εξαμηνιαία Εργασία

Γιάννης Λεβέντης - el18168
Ζαχαρίας Παύλος Αναστασιάδης - el18161

Ιανουάριος 2025

Σύνδεσμος Αποθετηρίου

Το αποθετήριο με τον κώδικα βρίσκεται διαθέσιμο στο ακόλουθο URL:
https://github.com/zachanast18/advanced$_d b_e ce_n tua_2$024 – 2025

Query 1

Στόχος

Η κατηγοριοποίηση των θυμάτων εγκλημάτων που περιλαμβάνουν "βαριά σωματική βλάβη" σε ηλικιακές ομάδες και η κατάταξή τους κατά φθίνουσα σειρά αριθμού περιστατικών. Οι ηλικιακές ομάδες που εξετάστηκαν είναι:

- **Παιδιά**: < 18 ετών
- **Νεαροί ενήλικοι**: 18–24 ετών
- **Ενήλικοι**: 25–64 ετών
- **Ηλικιωμένοι**: > 64 ετών

Βήματα Υλοποίησης

Φόρτωση και Προεπεξεργασία Δεδομένων

- Τα δεδομένα εγκλημάτων φορτώθηκαν από το S3 bucket μέσω του API των DataFrame της Spark.
- Επιλέχθηκαν οι στήλες `Vict Age` (ηλικία θύματος) και `Crime Description` (περιγραφή εγκλήματος).
- Τα δεδομένα φιλτραρίστηκαν ώστε να περιλαμβάνουν μόνο περιπτώσεις όπου η περιγραφή του εγκλήματος περιείχε τον όρο `"aggravated assault"`.
- Εγγραφές με κενές ή μη έγκυρες τιμές ηλικίας αφαιρέθηκαν.

Καθορισμός Ηλικιακών Ομάδων

Μια νέα στήλη δημιουργήθηκε με την κατηγοριοποίηση των θυμάτων στις παραπάνω ηλικιακές ομάδες, χρησιμοποιώντας τη συνάρτηση `when` και `otherwise` της PySpark:

```
withColumn("Age Group", when(col("Vict Age") < 18, "Παιδιά").
```

Η κατηγοριοποίηση ολοκληρώθηκε για όλες τις ηλικιακές ομάδες.

Ομαδοποίηση και Ταξινόμηση

- Τα δεδομένα ομαδοποιήθηκαν με βάση τη στήλη `Age Group`.
- Υπολογίστηκε ο αριθμός περιστατικών για κάθε ηλικιακή ομάδα μέσω της συνάρτησης `count()`.
- Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά βάσει του αριθμού περιστατικών.

Σύγκριση Απόδοσης

Η ίδια διαδικασία υλοποιήθηκε και με χρήση του API των RDDs. Οι βασικές λειτουργίες που χρησιμοποιήθηκαν ήταν `filter`, `map` και `reduceByKey`.

Αποτελέσματα

Υλοποίηση με DataFrame API

Χρόνος Εκτέλεσης: 4.43s

Αποτελέσματα:

Ηλικιακή Ομάδα	Αριθμός Περιστατικών
Ενήλικοι	121,052
Νεαροί ενήλικοι	33,588
Παιδιά	15,918
Ηλικιωμένοι	5,985

Πίνακας 1: Αποτελέσματα Ανάλυσης Ηλικιακών Ομάδων Θυμάτων

Υλοποίηση με RDD API

Χρόνος Εκτέλεσης: 15.38s
Τα αποτελέσματα ήταν ίδια με αυτά του DataFrame API.

Παρατηρήσεις

Απόδοση

- Το DataFrame API είχε σημαντικά καλύτερη απόδοση σε σχέση με το RDD API.
- Αυτό οφείλεται στον Catalyst optimizer της Spark, ο οποίος βελτιστοποιεί την εκτέλεση των DataFrame queries.

Ευχρηστία

Το DataFrame API παρείχε πιο ευανάγνωστο και περιεκτικό κώδικα σε σύγκριση με το RDD API.

Συμπεράσματα

Το DataFrame API είναι προτιμότερο για εργασίες όπως του ερωτήματος λόγω της καλύτερης απόδοσης και της ευχρηστίας του.

Query 2

Στόχος

Να εντοπιστούν τα τρία αστυνομικά τμήματα με το υψηλότερο ποσοστό περατωμένων υποθέσεων για κάθε έτος. Τα αποτελέσματα περιλαμβάνουν:

- **Έτος**
- **Όνομα Αστυνομικού Τμήματος**
- **Ποσοστό Περατωμένων Υποθέσεων**
- **Κατάταξη**

Βήματα Υλοποίησης

Φόρτωση Δεδομένων

- Τα δεδομένα εγκλημάτων φορτώθηκαν από το S3 bucket μέσω του API των DataFrame της Spark.
- Επιλέχθηκαν οι στήλες `Date Reported`, `Area Name`, και `Status`.

Προετοιμασία Δεδομένων

- Από τη στήλη `Date Reported` εξήχθη το έτος (`Year`).
- Αναλύθηκε η στήλη `Status` ώστε να διαχωριστούν οι υποθέσεις σε "κλειστές" (περατωμένες) και "ανοιχτές".
- Εγγραφές με μη έγκυρα ή κενά δεδομένα αφαιρέθηκαν.

Υπολογισμός Ποσοστού Περατωμένων Υποθέσεων

Για κάθε συνδυασμό `Year` και `Area Name` υπολογίστηκαν:

- Το συνολικό πλήθος υποθέσεων.
- Το πλήθος των κλειστών υποθέσεων με χρήση της συνάρτησης συνθήκης `when`.

Το ποσοστό υπολογίστηκε ως:

$$\text{Ποσοστό Περατωμένων Υποθέσεων} = \frac{\text{Αριθμός Κλειστών Υποθέσεων}}{\text{Σύνολο Υποθέσεων}} \times 100$$

Κατάταξη Τμημάτων

- Τα αστυνομικά τμήματα ταξινομήθηκαν κατά φθίνουσα σειρά βάσει του ποσοστού περατωμένων υποθέσεων.
- Εξήχθησαν τα τρία κορυφαία τμήματα για κάθε έτος.

Μετατροπή σε Μορφή Parquet

- Τα δεδομένα αποθηκεύτηκαν σε μορφή `Parquet` για αξιολόγηση της απόδοσης.
- Για την υλοποίηση DataFrame, συγκρίθηκαν οι χρόνοι εκτέλεσης μεταξύ εισαγωγής δεδομένων σε μορφή `CSV` και `Parquet`.

Αποτελέσματα

Υλοποίηση με DataFrame API

Χρόνος Εκτέλεσης (CSV Input): 9.38s
Χρόνος Εκτέλεσης (Parquet Input): 1.77s

Παράδειγμα Αποτελεσμάτων:

Έτος	Αστυνομικό Τμήμα	Ποσοστό Περατωμένων Υποθέσεων (%)	Κατάταξη
2010	Rampart	32.85	1
2010	Olympic	31.51	2
2010	Harbor	29.36	3

Πίνακας 2: Κορυφαία Αστυνομικά Τμήματα με Βάση το Ποσοστό Περατωμένων Υποθέσεων για το 2010

Υλοποίηση με SQL API

Χρόνος Εκτέλεσης (CSV Input): 3.84s
Τα αποτελέσματα ήταν συνεπή με αυτά του DataFrame API.

Παρατηρήσεις

Απόδοση

- Το SQL API είχε ελαφρώς καλύτερη απόδοση από το DataFrame API, λόγω της βελτιστοποίησης των SQL queries μέσω του Catalyst optimizer.
- Η μορφή `Parquet` βελτίωσε σημαντικά την ταχύτητα εκτέλεσης (περίπου 4x ταχύτερα) σε σύγκριση με τη μορφή `CSV`.

Συμπεράσματα για τα Τμήματα

- Το τμήμα `Rampart` είχε το υψηλότερο ποσοστό περατωμένων υποθέσεων για το έτος 2010.
- Παρατηρήθηκαν διαφορές στις επιδόσεις των τμημάτων, οι οποίες πιθανώς οφείλονται στη διαχείριση πόρων ή άλλους παράγοντες.

Συμπεράσματα

Το SQL API είναι ιδιαίτερα αποδοτικό για σύνθετες εργασίες ομαδοποίησης και κατάταξης, ενώ η χρήση της μορφής `Parquet` βελτιστοποιεί σημαντικά την απόδοση σε μεγάλα σύνολα δεδομένων.

Query 3

Στόχος

Ο στόχος του ερωτήματος είναι να υπολογιστούν τα παρακάτω για κάθε περιοχή του Los Angeles:

- Το μέσο ετήσιο εισόδημα ανά άτομο.
- Η αναλογία του συνολικού αριθμού εγκλημάτων προς τον πληθυσμό.

Τα αποτελέσματα συγκεντρώθηκαν σε έναν πίνακα για ανάλυση και σύγκριση.

Βήματα Υλοποίησης

Προετοιμασία Δεδομένων

- Τα δεδομένα εισοδήματος (απογραφή 2015) συνδέθηκαν με τα δεδομένα πληθυσμού (απογραφή 2010) χρησιμοποιώντας τον `ZIP Code`.
- Η στήλη `Estimated Median Income` καθαρίστηκε από μη αριθμητικούς χαρακτήρες και μετατράπηκε σε ακέραιους αριθμούς.
- Υπολογίστηκε το συνολικό εισόδημα:

$$\text{Συνολικό Εισόδημα} = \text{Αριθμός Νοικοκυριών} \times \text{Μέσο Εισόδημα Νοικοκυριού}$$

Υπολογισμός Μεγεθών

- Το κατά κεφαλήν εισόδημα υπολογίστηκε ως:

$$\text{Κατά Κεφαλήν Εισόδημα} = \frac{\text{Συνολικό Εισόδημα}}{\text{Πληθυσμός}}$$

- Η αναλογία εγκλημάτων υπολογίστηκε ως:

$$\text{Αναλογία Εγκλημάτων} = \frac{\text{Συνολικός Αριθμός Εγκλημάτων}}{\text{Πληθυσμός}}$$

Ανάλυση Δημογραφικών και Εγκληματικότητας

- Τα δεδομένα εγκλημάτων συσχετίστηκαν με τις περιοχές χρησιμοποιώντας γεωχωρικές λειτουργίες της Apache Sedona.
- Οι περιοχές ταξινομήθηκαν βάσει του κατά κεφαλήν εισοδήματος.

Αποτελέσματα

Δείγμα Αποτελεσμάτων

Περιοχή	Αναλογία Εγκλημάτων	Κατά Κεφαλήν Εισόδημα ($)
Pacific Palisades	0.3797	70,673.56
Palisades Highlands	0.1878	66,867.44
Marina Peninsula	0.6000	65,235.69
Bel Air	0.3992	63,041.34
Beverly Crest	0.3690	60,947.49
Brentwood	0.4059	60,846.85

Πίνακας 3: Αποτελέσματα Ανάλυσης Κατά Κεφαλήν Εισοδήματος και Αναλογίας Εγκλημάτων

Χρόνοι Εκτέλεσης για Στρατηγικές Join

Στρατηγική Join	Χρόνος Εκτέλεσης (δευτ.)
Shuffle Replicate NL + Shuffle Replicate NL	46.51
Shuffle Replicate NL + Merge	26.26
Merge + Merge	27.02
Broadcast + Broadcast	29.63
Shuffle Hash + Merge	23.19
Shuffle Hash + Broadcast	20.93
Shuffle Hash + Shuffle Hash	17.76

Πίνακας 4: Χρόνοι Εκτέλεσης για Διαφορετικές Στρατηγικές Join

Παρατηρήσεις

- Οι στρατηγικές `Broadcast` και `Shuffle Hash` ήταν πιο αποδοτικές σε σενάρια με μικρά και μεσαία σύνολα δεδομένων.
- Η χρήση της `Shuffle Replicate NL` οδήγησε σε υψηλούς χρόνους εκτέλεσης λόγω των μεγάλων απαιτήσεων μνήμης.
- Οι περιοχές με υψηλό εισόδημα παρουσίασαν χαμηλότερη αναλογία εγκλημάτων, ενώ οι περιοχές με χαμηλό εισόδημα εμφάνισαν υψηλότερη αναλογία.

Query 4

Στόχος

Το ερώτημα στοχεύει:

- Στον υπολογισμό του φυλετικού προφίλ των θυμάτων εγκλημάτων στις 3 περιοχές με το υψηλότερο κατά κεφαλήν εισόδημα για το έτος 2015.
- Στην εκτέλεση της ίδιας ανάλυσης για τις 3 περιοχές με το χαμηλότερο κατά κεφαλήν εισόδημα.

Η αντιστοίχιση των φυλετικών κωδικών πραγματοποιήθηκε με χρήση του συνόλου δεδομένων `Race and Ethnicity Codes`. Τα αποτελέσματα εμφανίζονται σε δύο ξεχωριστούς πίνακες.

Βήματα Υλοποίησης

Προετοιμασία Δεδομένων

- Τα δεδομένα εγκλημάτων για το έτος 2015 φιλτραρίστηκαν ώστε να περιλαμβάνουν μόνο έγκυρους φυλετικούς κωδικούς (`Vict Descent`).
- Τα δεδομένα εισοδήματος και πληθυσμού συνδυάστηκαν μέσω του `ZIP Code`.
- Υπολογίστηκε το κατά κεφαλήν εισόδημα ως:

$$\text{Κατά Κεφαλήν Εισόδημα} = \frac{\text{Συνολικό Εισόδημα}}{\text{Πληθυσμός}}$$

- Τα γεωχωρικά δεδομένα επεξεργάστηκαν με χρήση της Apache Sedona.

Ανάλυση Περιοχών

- Οι 3 περιοχές με το υψηλότερο κατά κεφαλήν εισόδημα και οι 3 με το χαμηλότερο επιλέχθηκαν μέσω ταξινόμησης.
- Τα δεδομένα εγκλημάτων συσχετίστηκαν γεωχωρικά με τις περιοχές χρησιμοποιώντας τη συνάρτηση `ST_Within`.
- Ομαδοποιήθηκαν τα δεδομένα βάσει φυλετικών κωδικών, και οι περιγραφές προστέθηκαν από το σύνολο δεδομένων `Race and Ethnicity Codes`.

Κλιμάκωση Υπολογιστικών Πόρων

Η υλοποίηση εκτελέστηκε με τις εξής διαμορφώσεις:

- 2 εκτελεστές με 1 πυρήνα και 2 GB μνήμης.
- 2 εκτελεστές με 2 πυρήνες και 4 GB μνήμης.
- 2 εκτελεστές με 4 πυρήνες και 8 GB μνήμης.

Αποτελέσματα

Περιοχές με Υψηλό Εισόδημα

Φυλετική Ομάδα Θυμάτων	Αριθμός Θυμάτων
White	649
Other	72
Hispanic/Latin/Mexican	66
Unknown	38
Black	37
Other Asian	21
Chinese	1
American Indian/Alaska Native	1

Πίνακας 5: Ανάλυση Θυμάτων στις Περιοχές με Υψηλό Εισόδημα

Περιοχές με Χαμηλό Εισόδημα

Φυλετική Ομάδα Θυμάτων	Αριθμός Θυμάτων
Hispanic/Latin/Mexican	2,815
Black	761
White	330
Other	187
Other Asian	113
Unknown	22
American Indian/Alaska Native	21
Korean	5
Chinese	3
Asian Indian	1
Filipino	1

Πίνακας 6: Ανάλυση Θυμάτων στις Περιοχές με Χαμηλό Εισόδημα

Χρόνοι Εκτέλεσης

Διαμόρφωση Εκτελεστών	Χρόνος Εκτέλεσης (δευτ.)
1 πυρήνας / 2 GB μνήμης	46.36
2 πυρήνες / 4 GB μνήμης	47.82
4 πυρήνες / 8 GB μνήμης	88.80

Πίνακας 7: Χρόνοι Εκτέλεσης με Διαφορετικές Διαμορφώσεις

Παρατηρήσεις

- Στις περιοχές με υψηλό εισόδημα, τα περισσότερα θύματα ανήκουν στη φυλετική ομάδα `White`.
- Στις περιοχές με χαμηλό εισόδημα, η πλειοψηφία των θυμάτων ανήκει στη φυλετική ομάδα `Hispanic/Latin/Mexican`.
- Η κλιμάκωση των πόρων δεν βελτίωσε τη συνολική απόδοση στις μεγαλύτερες διαμορφώσεις, πιθανώς λόγω κορεσμού της υποδομής.

Συμπεράσματα

Οι περιοχές με υψηλό εισόδημα είχαν χαμηλότερη εγκληματικότητα, ενώ οι περιοχές με χαμηλό εισόδημα παρουσίασαν υψηλότερη εγκληματική δραστηριότητα.

Query 5

Στόχος

Ο στόχος του ερωτήματος είναι:

- Ο υπολογισμός του αριθμού εγκλημάτων που έλαβαν χώρα πλησιέστερα σε κάθε αστυνομικό τμήμα.
- Ο υπολογισμός της μέσης απόστασης μεταξύ των εγκλημάτων και του πλησιέστερου αστυνομικού τμήματος.

Τα αποτελέσματα ταξινομήθηκαν κατά φθίνουσα σειρά αριθμού εγκλημάτων.

Βήματα Υλοποίησης

Προετοιμασία Δεδομένων

- Τα γεωχωρικά δεδομένα των αστυνομικών τμημάτων μετατράπηκαν σε γεωμετρικά σημεία (`ST_Point`).
- Τα δεδομένα εγκλημάτων και τμημάτων συνδυάστηκαν μέσω διασταυρούμενης ένωσης (`cross join`) για τον υπολογισμό των αποστάσεων.

Υπολογισμός Πλησιέστερου Τμήματος

Για κάθε έγκλημα:

- Υπολογίστηκε η απόσταση από όλα τα τμήματα.
- Επιλέχθηκε το πλησιέστερο τμήμα χρησιμοποιώντας την ελάχιστη απόσταση (`MIN(distance)`).

Ανάλυση Αποτελεσμάτων

Τα εγκλήματα ομαδοποιήθηκαν ανά τμήμα για την εξαγωγή:

- Του συνολικού αριθμού εγκλημάτων.
- Της μέσης απόστασης εγκλημάτων από το τμήμα.

Κλιμάκωση Υπολογιστικών Πόρων

Η υλοποίηση εκτελέστηκε με τρεις διαμορφώσεις πόρων:

- 2 εκτελεστές $\times$ 4 πυρήνες / 8 GB μνήμης.
- 4 εκτελεστές $\times$ 2 πυρήνες / 4 GB μνήμης.
- 8 εκτελεστές $\times$ 1 πυρήνας / 2 GB μνήμης.

Αποτελέσματα

Ανάλυση Εγκλημάτων

Αστυνομικό Τμήμα	Μέση Απόσταση (km)	Αριθμός Εγκλημάτων
Hollywood	2.08	224,340
Van Nuys	2.95	210,134
Southwest	2.19	188,901
Wilshire	2.59	185,996
77th Street	1.72	171,827
Olympic	1.72	170,897
North Hollywood	2.64	167,854
Pacific	3.85	161,359
Central	0.99	153,871
Rampart	1.53	152,736

Πίνακας 8: Ανάλυση Εγκλημάτων ανά Αστυνομικό Τμήμα

Χρόνοι Εκτέλεσης

Διαμόρφωση Εκτελεστών	Χρόνος Εκτέλεσης (δευτ.)
2 εκτελεστές × 4 πυρήνες / 8 GB	23.82
4 εκτελεστές × 2 πυρήνες / 4 GB	31.21
8 εκτελεστές × 1 πυρήνας / 2 GB	18.53

Πίνακας 9: Χρόνοι Εκτέλεσης για Διαφορετικές Διαμορφώσεις

Παρατηρήσεις

- Το τμήμα `Hollywood` είχε τον μεγαλύτερο αριθμό εγκλημάτων λόγω της υψηλής πυκνότητας πληθυσμού και δραστηριότητας στην περιοχή.
- Η μέση απόσταση για το τμήμα `Central` ήταν η μικρότερη (0.99 km), υποδεικνύοντας εγκληματική δραστηριότητα συγκεντρωμένη κοντά στο τμήμα.
- Η αύξηση του αριθμού εκτελεστών βελτίωσε την απόδοση, με τη διαμόρφωση `8 εκτελεστές × 1 πυρήνας / 2 GB` να έχει τον καλύτερο χρόνο εκτέλεσης.