**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

**Συστήματα Διαχείρισης Δεδομένων Μεγάλης Κλίμακας**

Διδάσκων: Ιωάννης Κωτίδης (kotidis@aueb.gr)

Βοηθός: Χρυσόστομος Καπέτης (mkap@aueb.gr)

**Ατομική εργασία - Η εργασία είναι ΥΠΟΧΡΕΩΤΙΚΗ και προσμετρά 25% στον τελικό βαθμό του μαθήματος.**

Ανάθεση: **8-12-2023**
Παράδοση: **03-01-2024 Ώρα (23:55)**

**Εργασία σε Spark**

**"Στοιχεία Εγκληματικότητας"**

Τα παρακάτω γραμμογραφημένα αρχεία κειμένου περιέχουν δεδομένα για περιστατικά εγκληματικότητας που έχουν διαπραχθεί στο Λος Άντελες από το 2020 μέχρι και σήμερα. Τα πεδία των εγγραφών κάθε αρχείου διαχωρίζονται με τον χαρακτήρα "|".

| **areas.csv: Αρχείο με τις περιοχές στις οποίες έλαβαν χώρα τα περιστατικά εγκληματικότητας** | |
|---|---|
| area_id | Κωδικός περιοχής (ακέραιος αριθμός) |
| area | Όνομα περιοχής |

| **crimes.csv: Αρχείο με τους τύπους των διαπραχθέντων εγκλημάτων.** | |
|---|---|
| crime_id | Κωδικός που προσδιορίζει τον τύπο του εγκλήματος (ακέραιος αριθμός) |
| crime_desc | Τύπος (είδος) εγκλήματος |

| **premises.csv: Αρχείο με τους τύπους των εγκαταστάσεων, οχημάτων ή τοποθεσιών που διαπράχθηκαν τα εγκλήματα** | |
|---|---|
| premis_id | Κωδικός (ακέραιος αριθμός) |
| Premis_desc | Περιγραφή είδους εγκατάστασης, οχήματος, τοποθεσίας |

| **weapons.csv: Αρχείο με τα είδη των όπλων που χρησιμοποιήθηκαν στα εγκλήματα.** | |
|---|---|
| weapon_id | Κωδικός που προσδιορίζει το είδος του όπλου (ακέραιος αριθμός) |
| weapon | Περιγραφή όπλου |

| **victim_descent.csv: Αρχείο με τις χώρες καταγωγής των θυμάτων.** | |
|---|---|
| descent_id | Κωδικός χώρας καταγωγής (ένας χαρακτήρας π.χ. A,B,X) |
| descent | Καταγωγή |

| **case_status.csv: Αρχείο με τις καταστάσεις των περιστατικών** | |
|---|---|
| status_id | Κωδικός που προσδιορίζει την κατάσταση ενός περιστατικού. Ο κωδικός σχηματίζεται από δύο χαρακτήρες (IC, AA κ.λπ.) |
| Status_desc | Κατάσταση περιστατικού (Invest Cont, Adult Arrest, Juv Arrest κ.λπ.) |

| **criminal_cases.csv: Αρχείο με τα περιστατικά εγκληματικότητας** | |
|---|---|
| case_id | Κωδικός περιστατικού (Ακέραιος αριθμός) |
| date_occured | Ημερομηνία στην οποία το περιστατικό έλαβε χώρα (YYYY-MM-DD) |
| area_id | Κωδικός περιοχής στην οποία συνέβη το περιστατικό |
| crime_id | Ο τύπος του εγκλήματος που διαπράχθηκε |
| victim_age | Ηλικία θύματος. |
| victim_sex | Φύλο θύματος. |
| victim_descent_id | Κωδικός χώρας καταγωγής του θύματος |
| Premis_id | Κωδικός τύπου εγκατάστασης, οχήματος ή τοποθεσίας στην οποία έλαβε χώρα το περιστατικό. |
| Weapon_used_id | Τύπος όπλου που χρησιμοποιήθηκε (εφόσον χρησιμοποιήθηκε κάποιο όπλο) |
| case_status_id | Κωδικός κατάστασης περιστατικού |

Καλείστε να γράψετε ένα πρόγραμμα σε Apache Spark, χρησιμοποιώντας μια από τις γλώσσες προγραμματισμού Scala/Python/Java για την παραγωγή στατιστικών αναφορών. Συγκεκριμένα το πρόγραμμα πρέπει:

1. Να παράγει αναφορά με τον συνολικό αριθμό των περιστατικών εγκληματικότητας ανά περιοχή και είδος εγκατάστασης, οχήματος ή τοποθεσίας της μορφής: «Περιοχή, Είδος (premis_desc), Αριθμός_Περιστατικών». Η αναφορά να είναι ταξινομημένη με την περιοχή σε αύξουσα διάταξη και τον αριθμό των περισταστικών σε φθίνουσα διάταξη.

2. Να παράγει αναφορά με τα 10 κορυφαία είδη εγλημάτων της μορφής: «Τύπος_Εγκλήματος, Αριθμός_Περιστατικών». Η αναφορά να είναι ταξινομημένη με βάση τον συνολικό αριθμό των περιστατικών σε φθίνουσα διάταξη.

3. Να παραγει αναφορά με τον μηνιαίο αριθμό εγκληματικών περιστατικών κάθε έτους. Η αναφορά να είναι ταξινομημένη με βάση το έτος και τον μήνα σε αύξουσα διάταξη.

4. Να παράγει αναφορά με την κατάσταση των περιστατικών ανά είδος εγκλήματος της μορφής: «Είδος_Εγκλήματος, Κατάσταση_Περιστατικού, Αριθμός_Περιστατικών». Η αναφορά να

είναι ταξινομημένη αλφαβητικά με το είδος του εγλήματος και την κατάσταση του περιστατικού.

5. Να δημιουργεί έναν κύβο (data cube), κάθε κελί του οποίου περιέχει τον συνολικό αριθμό των εγληματικών περιστατικών ανά χώρα καταγωγής, φύλο και ηλικία θύματος.

6. Να δημιουργεί κατάλληλα γραφήματα (π.χ. Histogram, Pie chart) για την παρουσίαση των περιεχομένων της δεύτερηςς και της τρίτης αναφοράς (βλέπε 2 και 3).

**Σημείωση:** Το πρόγραμμα πρέπει να αποθηκεύει τα αποτελέσματα κάθε αναφοράς καθώς και τα γραφήματα σε ξεχωριστό αρχείο (ένα αρχείο για κάθε αναφορά/γράφημα).

**Παραδοτέα:**
Θα πρέπει να αναρτήσετε στο eclass εναν συμπιεσμένο φάκελο το όνομα του οποίου θα είναι ο αριθμός μητρώου. Ο φάκελος θα περιέχει:

1. Το αρχείο (ή αρχεία) με τον πηγαίο κώδικα του προγράμματος. Στην αρχή του αρχείου να γράψετε υπό την μορφή σχολίου το ονοματεπώνυμό σας και τον αριθμό μητρώου.
2. Τα αρχεία με τα αποτελέσματα των αναφορών και τα γραφήματα.

**Εξέταση**
Θα κληθείτε ατομικά να προυσιάσετε την εφαρμογή σε συγκεκριμένη ημέρα και ώρα. Κατά την διάρκεια της παρουσίασης θα ζητηθεί να μεταγλωττίσετε την εφαρμογή, να επιδείξετε την λειτουργικότητά της και να απαντήσετε σε σχετικές ερωτήσεις.

.