

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Συστήματα Διαχείρισης Δεδομένων Μεγάλης Κλίμακας
(Χειμερινό Εξάμηνο 2025-2026)

Διδάσκων: Ιωάννης Κωτίδης (kotidis@aueb.gr)

Βοηθός: Χρυσόστομος Καπέτης (mkar@aueb.gr)

Οδηγίες

- Η εργασία είναι **ΟΜΑΔΙΚΗ (ομάδες δύο ατόμων)**
- Η εργασία είναι **ΥΠΟΧΡΕΩΤΙΚΗ** και προσμετρά **25%** στον τελικό βαθμό του μαθήματος.

Ανάθεση: **21-11-2025**

Παράδοση: **08-12-2025 Ώρα (23:55)**

Αποθήκες Δεδομένων

Το αρχείο **crimeData.csv** περιέχει δεδομένα εγκληματικότητας για το Los Angeles για τα έτη 2020 έως και 2023.

Η αστυνομική διεύθυνση του Los Angeles (LAPD) επιθυμεί να αναπτύξει μια αποθήκη δεδομένων με σκοπό την άντληση στατιστικών στοιχείων σχετικά με τα περιστατικά εγκληματικότητας, τις τοποθεσίες που λαμβάνουν χώρα, τα είδη των εγκλημάτων και τα δημογραφικά χαρακτηριστικά των θυμάτων. Ειδικότερα η αστυνομική διεύθυνση ενδιαφέρεται για την κατηγοριοποίηση των εγκληματικών περιστατικών ανά είδος εγκλήματος, περιοχή, τύπο εγκαταστάσεων, οχημάτων ή τοποθεσιών που διαπράχθηκαν τα εγκλήματα, είδος όπλων που χρησιμοποιήθηκαν καθώς επίσης και με τα δημογραφικά χαρακτηριστικά των θυμάτων (καταγωγή, φύλο, ηλικία). Επιπλέον η αστυνομική διεύθυνση ενδιαφέρεται και για την κατηγοριοποίηση των εγκληματικών περιστατικών με βάση την κατάστασή τους (π.χ. IC=Investigation Continues, AA=Adult Arrest).

Εξυπακούεται ότι στην ανάλυση των εγκληματικών περιστατικών θα πρέπει να ληφθεί υπόψη και ο παράγοντας του χρόνου έτσι ώστε, η αστυνομική διεύθυνση να μπορεί να παρακολουθεί τάσεις εγκληματικότητας και να παράγει αναφορές με στατιστικά στοιχεία ανα έτος, μήνα και ημέρα.

Καλείστε να σχεδιάσετε και να υλοποιήσετε μία αποθήκη δεδομένων που θα επιτρέπει την εύκολη και αποδοτική παραγωγή στατιστικών στοιχείων. Στην συνέχεια να τροφοδοτήσετε την αποθήκη με τα δεδομένα του αρχείου **crimeData.csv** και να εκτελέσετε ορισμένες επερωτήσεις για την παραγωγή χρήσιμων στατιστικών αναφορών. Την αποθήκη δεδομένων θα την υλοποιήσετε με την χρήση του συστήματος **SPARK**. Για την παρουσίαση των αποτελεσμάτων θα αξιοποιήσετε τις δυνατότητες του **Power BI**.

Ακολουθεί αναλυτική περιγραφή των δεδομένων και των ζητούμενων της εργασίας.

Το αρχείο **crimeData.csv** περιέχει 246416 εγγραφές. Κάθε εγγραφή αποτελείται από 18 πεδία τα οποία διαχωρίζονται με τον χαρακτήρα ";". Ακολουθεί η περιγραφή των πεδίων.

crimeData.txt		
CaseID	integer	Κωδικός περιστατικού εγκλήματος.
DateOccured	date	Ημερομηνία που συνέβη το περιστατικό (yyyy-mm-dd).
AreaCode	integer	Κωδικός γεωγραφικής περιοχής.
Area	string	Όνομα περιοχής.
CrimeCode	integer	Κωδικός που προσδιορίζει το είδος του εγκλήματος.
CrimeDescription	integer	Περιγραφή του είδους εγκλήματος.
VictimAge	integer	Ηλικία θύματος.
VictimSex	string	Φύλο θύματος.
VictimDescentCode	string	Κωδικός καταγωγής θύματος (ταξινόμηση LAPD).
VictimDescent	string	Περιγραφή καταγωγής θύματος.
PremisCode	integer	Κωδικός τύπου εγκατάστασης, οχήματος ή τοποθεσίας στην οποία έλαβε χώρα το περιστατικό.
PremisDescription	string	Περιγραφή τύπου εγκατάστασης, οχήματος ή τοποθεσίας στην οποία έλαβε χώρα το περιστατικό.
WeaponCode	integer	Κωδικός που προσδιορίζει το είδος του όπλου (0=Unknown/not used).
Weapon	string	Περιγραφή όπλου.
CaseStatusCode	String	Κωδικός κατάστασης περιστατικού (π.χ. IC=Investigation Continues, AA=Adult Arrest).
CaseStatusDescription	String	Περιγραφή κατάστασης περιστατικού.
Latitude	decimal(9,6)	Γεωγραφικό πλάτος τοποθεσίας που συνέβη το περιστατικό.
Longitude	decimal(9,6)	Γεωγραφικό μήκος τοποθεσίας που συνέβη το περιστατικό.

Ζήτημα Πρώτο [Μονάδες 50]

1. Σχεδιάστε ένα διάγραμμα που να απεικονίζει το σχήμα αστέρα της αποθήκης δεδομένων. Στο διάγραμμα θα πρέπει να φαίνονται οι πίνακες με τα πεδία τους και οι συσχετίσεις των πινάκων διαστάσεων με τον πίνακα των γεγονότων. Για την δημιουργία του διαγράμματος μπορείτε να χρησιμοποιήσετε οποιοδήποτε σχεδιαστικό πρόγραμμα της αρεσκείας σας.
2. Καλείστε να γράψετε ένα πρόγραμμα σε Apache Spark, χρησιμοποιώντας μια από τις γλώσσες προγραμματισμού Scala/Python/Java το οποίο θα υλοποιεί τα παρακάτω:
 - 2.1 Θα δημιουργεί το λογικό σχήμα της αποθήκης δεδομένων το οποίο θα πρέπει να έχει την μορφή αστέρα (star schema) και θα το τροφοδοτεί με τα απαραίτητα δεδομένα από το αρχείο crimeData.csv.
 - 2.2 Χρησιμοποιώντας το σχήμα της αποθήκης δεδομένων (**όχι το αρχείο crimeData.csv**) να δημιουργεί τις παρακάτω στατιστικές αναφορές:
 - 2.2.1 Να παράγει αναφορά με τον συνολικό αριθμό των περιστατικών εγκληματικότητας ανά περιοχή και τύπο εγκατάστασης, οχήματος ή τοποθεσίας της μορφής: «Περιοχή, Τύπος (premisDescription), Αριθμός_Περιστατικών». Η αναφορά να είναι ταξινομημένη με την περιοχή σε αύξουσα διάταξη και τον αριθμό των περιστατικών σε φθίνουσα διάταξη.

2.2.2 Να παράγει αναφορά με τα 10 κορυφαία είδη εγκλημάτων (CrimeDescription) της μορφής: «Είδος_Εγκλήματος, Αριθμός_Περιστατικών». Η αναφορά να είναι ταξινομημένη με βάση τον συνολικό αριθμό των περιστατικών σε φθίνουσα διάταξη.

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

2.2.4 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 εγκλήματος και την κατάσταση του περιστατικού.

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

Σημείωση: Τα αποτελέσματα κάθε αναφοράς θα πρέπει να αποθηκεύονται σε ξεχωριστό αρχείο (ένα αρχείο για κάθε αναφορά).

2.3 Τέλος το πρόγραμμα πρέπει να αποθηκεύει τα δεδομένα που περιέχονται στους πίνακες της αποθήκης (πίνακες διαστάσεων και πίνακα γεγονότων) σε ξεχωριστά αρχεία CSV, ένα αρχείο CSV για κάθε πίνακα. Ο στόχος είναι τα αρχεία αυτά να χρησιμοποιηθούν για την μεταφόρτωση του σχήματος αστερά στο Power BI.

Ζήτημα Δεύτερο [Μονάδες 50]

Στο σημείο αυτό καλείστε να χρησιμοποιήσετε το εργαλείο Power BI για να την δημιουργία γραφημάτων. Συγκεκριμένα:

1. Μεταφορτώστε στο PowerBI το σχήμα και τα δεδομένα της αποθήκης που δημιουργήσατε στο προηγούμενο ζήτημα. Με άλλα λόγια θα πρέπει να φορτώσετε τα αρχεία CSV που δημιουργήσατε στο παραπάνω υποερώτημα 2.3.
2. Δημιουργήστε κατάλληλα γραφήματα για την παρουσίαση των αποτελεσμάτων των πέντε αναφορών που δημιουργήσατε στο υποερώτημα 2.2 του πρώτου ζητήματος. Να δημιουργήσετε ένα γράφημα για κάθε αναφορά.
3. Εξερευνήστε τα δεδομένα της αποθήκης και δημιουργήστε ένα καλαίσθητο dashboard με κατάλληλα γραφήματα και άλλα στοιχεία παρουσίασης στο οποίο να απεικονίζονται οι πλέον χρήσιμες κατά την γνώμη σας πληροφορίες σχετικά με τα περιστατικά εγκληματικότητας.

ΠΑΡΑΔΟΤΕΑ

Θα πρέπει να αναρτήσετε στο eclass έναν 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 (Ζήτημα Πρώτο:2). Στην αρχή του αρχείου να γράψετε υπό την μορφή σχολίου το ονοματεπώνυμό σας και τον αριθμό μητρώου.

2. Ένα αρχείο pdf το οποίο θα περιέχει:

- Το διάγραμμα που θα απεικονίζει το σχήμα αστέρα της αποθήκης (Ζήτημα – Πρώτο:1)
- Τα γραφήματα και το dashboard που θα δημιουργήσετε στο δεύτερο ζήτημα

ΕΞΕΤΑΣΗ

Θα κληθείτε να παρουσιάσετε την εργασία σε συγκεκριμένη ημέρα και ώρα. Κατά την διάρκεια της παρουσίασης θα ζητηθεί να μεταγλωττίσετε και να εκτελέσετε την εφαρμογή, να επιδείξετε την λειτουργικότητά της και να απαντήσετε σε σχετικές ερωτήσεις.