ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΒΑΣΕΩΝ ΓΝΩΣΕΩΝ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Προχωρημένα Θέματα Βάσεων Δεδομένων

Ακαδημαϊκό έτος 2025-26, 9ο Εξάμηνο

Διδάσκων: Δημήτριος Τσουμάκος

Υπεύθυνος Εργαστηρίου: Νικόλαος Χαλβαντζής

10 Νοεμβρίου 2025

Εξαμηνιαία Εργασία

Περιγραφή

Στην παρούσα εξαμηνιαία εργασία ζητείται ανάλυση σε (μεγάλα) σύνολα δεδομένων, εφαρμόζοντας επεξεργασία με τεχνικές που εφαρμόζονται σε data science projects. Τα εργαλεία που θα χρησιμοποιηθούν στα πλαίσια του project είναι τα Apache Hadoop (version>=3.0) και Apache Spark (version>=3.5). Καλείστε να χρησιμοποιήσετε τους πόρους στο ειδικά διαμορφωμένο περιβάλλον που σας έχει παραχωρηθεί στο AWS cloud. Συνοπτικά, ο σκοπός της εργασίας είναι:

- η εξοικείωση και ανάπτυξη των δεξιοτήτων των σπουδαστών στην εγκατάσταση και διαχείριση των κατανεμημένων συστημάτων Apache Spark και Apache Hadoop.
- Η χρήση σύγχρονων τεχνικών μέσω των API του Spark για την ανάλυση δεδομένων όγκου.
- Η κατανόηση των δυνατοτήτων και περιορισμών των εργαλείων αυτών σε σχέση με τους διαθέσιμους πόρους και τις ρυθμίσεις που έχουν επιλεγεί.

Δεδομένα

Στην παράγραφο αυτή θα παρουσιαστούν τα δεδομένα που θα κληθείτε να χρησιμοποιήσετε στα πλαίσια της εξαμηνιαίας εργασίας. Πρόκειται για δημοσίως διαθέσιμα και δωρεάν σύνολα δεδομένων που έχουν περισυλλεγεί από διαφορετικές πηγές.

Προς διευκόλυνσή σας, όλα τα απαραίτητα σύνολα δεδομένων είναι προσβάσιμα στο παρακάτω S3 bucket του AWS cloud: s3://initial-notebook-data-bucket-dblab-905418150721/.

Σύνολο Δεδομένων	S3 URI
Los Angeles Crime Data (2010-2019)	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/LA_Crime_Data/LA_Crime_Data_2010_2019.csv
Los Angeles Crime Data (2020-)	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/LA_Crime_Data/LA_Crime_Data_2020_2025.csv
Census Blocks	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/LA_Census_Blocks_2020.geojson
Census Blocks Fields	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/LA_Census_Blocks_2020_fields.csv
Median Household Income by Zip Code	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/LA_income_2021.csv
LA Police Stations	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/LA_Police_Stations.csv
Race and Ethnicity Codes	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/RE_codes.csv
MO Codes	s3://initial-notebook-data-bucket-dblab-905418150721/project_data/MO_codes.txt

Πίνακας 1: Σύνολα Δεδομένων και οι τοποθεσίες όπου βρίσκονται στο S3 cloud.

Βασικό data-set: Los Angeles Crime Data[1] [2]

Το βασικό σύνολο δεδομένων που θα χρησιμοποιηθεί στην εργασία προέρχεται από το δημόσιο αποθετήριο δεδομένων της κυβέρνησης των Ηνωμένων Πολιτειών της Αμερικής[3] . Περιλαμβάνει δεδομένα καταγραφής εγκλημάτων για την πόλη Los Angeles από το 2010 μέχρι σήμερα χωρισμένα σε δύο μέρη (2010-2019, 2020-). Στους σχετικούς συνδέσμους μπορείτε να βρείτε τις περιγραφές των δεδομένων για κάθε πεδίο (column).

Δευτερεύοντα data-sets

Συμπληρωματικά με τα παραπάνω δεδομένα, θα χρησιμοποιηθεί μια σειρά δεδομένων μικρότερου όγκου τα οποία επίσης είναι διαθέσιμα σε δημόσια αποθετήρια ή πηγές :

Census Blocks (Los Angeles County)[4] : Σύνολο δεδομένων που περιέχει απογραφικά στοιχεία της Κομητεία του Los Angeles για το έτος 2020 σε geojson format. Συνοδεύεται από αρχείο με περιγραφές των πεδίων του (Census Blocks Fields).

Median Household Income by Zip Code (Los Angeles County)[5] : Σύνολο δεδομένων που περιέχει δεδομένα σχετικά με το μέσο εισόδημα ανά νοικοκυριό για κάθε ταχυδρομικό κώδικα (ZIP Code) στην Κομητεία του Los Angeles. Για διευκόλυνση, τα δεδομένα έχουν συλλεχθεί και αποθηκευθεί σε csv format – ως delimitter, ο χαρακτήρας ";". Αναφέρεται σε στατιστικά στοιχεία του έτους 2021.

LA Police Stations[6] : Μικρό σύνολο δεδομένων που περιέχει δεδομένα σχετικά με την τοποθεσία των 21 αστυνομικών τμημάτων που βρίσκονται στην πόλη του Los Angeles.

Race and Ethnicity codes: Μικρό σύνολο δεδομένων που περιέχει τις πλήρες περιγραφές που αντιστοιχούν στην κωδικοποίηση του φυλετικού προφίλ που χρησιμοποιείται στο βασικό σύνολο δεδομένων.

MO Codes[7] : Σύνολο δεδομένων με τις περιγραφές που αντιστοιχούν στους κωδικούς της στήλης "Mocodes" του **Los Angeles Crime Data**. Παρέχεται σε μορφή txt, όπου ένας κωδικός βρίσκεται στην αρχή κάθε γραμμής και διαχωρίζεται από την περιγραφή με ένα κενό.

Ερωτήματα

Query 1

Να ταξινομηθούν, σε φθίνουσα σειρά, οι ηλικιακές ομάδες των θυμάτων σε περιστατικά που περιλαμβάνουν οποιαδήποτε μορφή "*βαριάς σωματικής βλάβης*". Θεωρείστε τις εξής ηλικιακές ομάδες:

[1] https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z

[2] https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

[3] https://catalog.data.gov/dataset

[4] https://data.lacounty.gov/maps/lacounty::2020-census-blocks

[5] http://www.laalmanac.com/employment/em12c_2021.php

[6] https://geohub.lacity.org/datasets/lahub::lapd-police-stations/explore

[7] https://data.lacity.org/api/views/63jg-8b9z/files/e14442b9-a6b8-4531-83f3-f7ba980b1377

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

Query 2

Ανά έτος, να βρεθούν τα 3 φυλετικά γκρουπ με τα περισσότερα θύματα καταγεγραμμένων εγκλημάτων (Vict Descent) στο Los Angeles. Τα αποτελέσματα να εμφανιστούν με φθίνουσα σειρά αριθμού θυμάτων ανά φυλετικό γκρουπ – να υπολογιστεί και να εμφανιστεί επίσης το ποσοστό επί του συνολικού αριθμού θυμάτων ανα περίπτωση (δείτε παράδειγμα αποτελέσματος στον Πίνακα 2).

year	Victim Descent	#	%
2024	White	413	32.5
2024	Black	274	25.4
2024	Unknown	132	22.3
2023	Hispanic/Latin/Mexican	512	30.2
⋮			

Πίνακας 2: Υπόδειγμα αποτελέσματος Query 1

Query 3

Να ταξινομηθούν και να εμφανιστούν με φθίνουσα σειρά συχνότητας εμφάνισης οι μέθοδοι διάπραξης εγκλημάτων και οι αντίστοιχοι κωδικοί τους (Mocodes). Χρησιμοποιήστε το σύνολο **MO Codes** για να αντιστοιχίσετε τους κωδικούς με τις περιγραφές τους.

Query 4

Να υπολογιστεί, ανά αστυνομικό τμήμα, ο αριθμός εγκλημάτων που έλαβαν χώρα πλησιέστερα σε αυτό από οποιοδήποτε άλλο, καθώς και η μέση απόστασή του από τις τοποθεσίες όπου σημειώθηκαν τα συγκεκριμένα περιστατικά. Τα αποτελέσματα να εμφανιστούν ταξινομημένα κατά αριθμό περιστατικών, με φθίνουσα σειρά (δείτε παράδειγμα στον Πίνακα 3).

division	average_distance	#
77TH STREET	2.208	7045
RAMPART	2.009	4595
FOOTHILL	3.597	3047
PACIFIC	2.739	2132

Πίνακας 3: Υπόδειγμα αποτελέσματος Query 4.

Query 5

Χρησιμοποιώντας ως αναφορά τα δεδομένα της απογραφής του 2020 για τον πληθυσμό και τα οικονομικά στοιχεία του 2021 για το εισόδημα ανα νοικοκυριό, να υπολογίσετε μέσα στη διετία 2020-2021 τη συσχέτιση μέσου ετήσιου κατακεφαλήν εισοδήματος με την ετήσια μέση αναλογία εγκλημάτων ανά άτομο για κάθε περιοχή του Λος Άντζελες. Επαναλάβετε τον υπολογισμό εξετάζοντας μόνο τις 10 περιοχές με το υψηλότερο και τις 10 με το χαμηλότερο ετήσιο κατακεφαλήν εισόδημα.

Tips:

1. Ως εγκλήματα που περιλαμβάνουν οποιαδήποτε μορφή "*βαριάς σωματικής βλάβης*" θεωρούμε όλα εκείνα τα περιστατικά που περιέχουν τον όρο "*aggravated assault*" στη σχετική περιγραφή.
2. Για την υλοποίηση queries που περιλαμβάνουν geospatial analytics θα πρέπει να χρησιμοποιήσετε τη βιβλιοθήκη Apache Sedona (version 1.6.1), που έχει εγκατασταθεί στο περιβάλλον εργασίας σας. Ενδεικτικά, σας δίνεται ένας οδηγός χρήσης σε σχετικό notebook που μπορείτε να βρείτε στο αντίστοιχο section του λογαριασμού σας. Περισσότερες πληροφορίες μπορείτε να βρείτε στο documentation και την ιστοσελίδα: https://sedona.apache.org/1.6.1/.
3. Θεωρήστε ότι οι διάφορες περιοχές του Los Angeles ορίζονται από τη στήλη "COMM" του **Census Blocks**.
4. Κάποιες εγγραφές του βασικού συνόλου δεδομένων λανθασμένα αναφέρονται στο Null Island (0,0). Θα πρέπει να φιλτραριστούν και να μη λαμβάνονται υπόψη στον υπολογισμό των αποστάσεων.

Ζητούμενα

1. Να υλοποιηθεί το **Query 1** χρησιμοποιώντας τα DataFrame (με και χωρίς UDF) και RDD APIs. Να εκτελέσετε και τις δύο υλοποιήσεις με 4 executors (1 core, 2GB memory). Υπάρχει διαφορά στην επίδοση μεταξύ των τριών υλοποιήσεων; Να σχολιάσετε τα ευρήματά σας. (20%)
2. Να υλοποιηθεί το **Query 2** χρησιμοποιώντας τα DataFrame και SQL APIs. Να εκτελέσετε και τις δύο υλοποιήσεις με 4 executors (1 core, 2GB memory). Να συγκρίνετε και να σχολιάσετε τους χρόνους εκτέλεσης μεταξύ των δύο υλοποιήσεων. (20%)
3. Να υλοποιηθεί το **Query 3** χρησιμοποιώντας τα DataFrame και RDD APIs. Να εκτελέσετε και τις δύο υλοποιήσεις με 4 executors (1 cores, 2GB memory). Να συγκριθούν και να σχολιαστούν οι χρόνοι εκτέλεσεις για την κάθε υλοποίηση. Στη περίπτωση του DataFrame API, χρησιμοποιήστε τις μεθόδους hint και explain για να βρείτε ποιες στρατηγικές join χρησιμοποιεί ο catalyst optimizer. Πειραματιστείτε προτρέποντας το Spark να χρησιμοποιήσει διαφορετικές στρατηγικές (μεταξύ των BROADCAST, MERGE, SHUFFLE_HASH, SHUFFLE_REPLICATE_NL) και σχολιάστε την επίδραση στην επίδοση. Ποιά (ή ποιές) από τις διαθέσιμες στρατηγικές join του Spark είναι καταλληλότερη(ες) και γιατί; (20%)
4. Να υλοποιηθεί το **Query 4** χρησιμοποιώντας το DataFrame ή SQL API. Για τα joins που εκτελούνται, να αναφέρετε ποιες στρατηγικές επιλέγει ο optimizer του Spark και να σχολιάσετε τις επιλογές του. Να εκτελέσετε την υλοποίησή σας εφαρμόζοντας κλιμάκωση στο σύνολο των υπολογιστικών πόρων που θα χρησιμοποιήσετε: Συγκεκριμένα, καλείστε να εκτελέστε την υλοποίησή σας σε 2 executors με τα ακόλουθα configurations:
   - 1 core, 2 GB memory
   - 2 cores, 4GB memory
   - 4 cores, 8GB memory

   Σχολιάστε τα αποτελέσματα. (20%)
5. Να υλοποιηθεί το **Query 5** χρησιμοποιώντας το DataFrame ή SQL API. Για τα joins που εκτελούνται, να αναφέρετε ποιες στρατηγικές επιλέγει ο optimizer του Spark και να σχολιάσετε τις επιλογές του. Να εκτελέσετε την υλοποίησή σας χρησιμοποιώντας συνολικούς πόρους 8 cores και 16GB μνήμης με τα παρακάτω configurations:
   - 2 executors $\times$ 4 cores, 8GB memory
   - 4 executors $\times$ 2 cores, 4GB memory
   - 8 executors $\times$ 1 core, 2GB memory

   Σχολιάστε τα αποτελέσματα. (20%)

Παραδοτέα - Όροι Υποβολής

- Η εργασία να εκπονηθεί σε ομάδες το πολύ των 2 ατόμων.
- Η προθεσμία παράδοσης θα καθοριστεί στο helios σε link που θα ανοίξει σύντομα. Εκπρόθεσμες υποβολές με καθυστέρηση μέχρι μία (1) ημέρα θα έχουν ποινή 50% του βαθμού. Πέραν αυτής της καθυστέρησης, καμία υποβολή δεν θα βαθμολογείται. Υποβολές με μέσο άλλο από το helios δεν γίνονται δεκτές.
- Η εργασία αποτελεί το 30% του συνολικού βαθμού του μαθήματος. Για να καταχωριθεί βαθμός, η κάθε ομάδα θα πρέπει να υποβάλει αναφορά **και** να περάσει επιτυχώς την υποχρεωτική προφορική εξέταση στο αντικείμενο της εργασίας. Η εξέταση θα γίνει μετά την παράδοση της εργασίας (θα αναρτηθεί σχετικό πρόγραμμα στο helios).
- Ως παραδοτέο θα υποβληθεί ένα .pdf αρχείο με όνομα τους ΑΜ των μελών της ομάδας χωρισμένα με κάτω παύλα (ή το ΑΜ του φοιτητή σε περίπτωση μονομελούς ομάδας), π.χ. 03100000.zip, ή 03100000_03100001.zip (ανάλογα με το πλήθος των ατόμων της ομάδας). Το αρχείο θα περιέχει μία αναφορά (αυστηρά με όσα ζητούνται στην εκφώνηση) η οποία θα περιέχει τις απαντήσεις στα ζητούμενα καθώς και απαραιτήτως ένα link σε αποθετήριο (github, gitlab, bitbucket, etc.) με τους κώδικες που έχετε υλοποιήσει, όπως και πιθανά scripts/howtos για την εκτέλεση του κώδικά σας. Όλες οι υποβολές υπόκεινται αυστηρά στον κώδικα ακαδημαϊκής ηθικής του ΕΜΠ και της ΣΗΜΜΥ.
  Ο κώδικάς σας δεν πρέπει να αλλάξει από την ημέρα παράδοσης της αναφοράς μέχρι και τη βαθμολόγηση του μαθήματος. Αν συμβεί αυτό η βαθμολογία σας θα είναι ΜΗΔΕΝ (0).
- Η κάθε ομάδα μπορεί να υλοποιήσει τον κώδικά της σε Scala, Java ή Python. Επιπλέον, σας δίνεται η δυνατότητα να χρησιμοποιήσετε δικούς σας πόρους (π.χ., προσωπικούς Η/Υ, VM σε άλλο cloud provider), αρκεί να απαντώνται τα ζητούμενα της εργασίας. Σε κάθε περίπτωση, η εξέταση θα απαιτήσει τη ζωντανή επίδειξη του κώδικά σας.
- Απορίες/επεξηγήσεις για την εργασία θα γίνονται μέσω forum στη σελίδα του μαθήματος στο helios. Μη στέλνετε απορίες στα email των διδασκόντων/βοηθών.