ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διαχείριση Δεδομένων Μεγάλης Κλίμακας

Ακαδημαϊκό έτος 2024-25, Εαρινό Εξάμηνο

Διδάσκοντες: Δημήτριος Τσουμάκος, Ιωάννης Κωνσταντίνου

Υπεύθυνος Εργαστηρίου: Νικόλαος Χαλβαντζής

11 Απριλίου 2025

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 που σας διατίθενται στο ειδικά διαμορφωμένο περιβάλλον που έχει δημιουργηθεί για εσάς και στο οποίο μπορείτε να αποκτήσετε πρόσβαση ακολουθώντας τους εργαστηριακούς οδηγούς του μαθήματος. 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 συλλεχθεί από διαφορετικές πηγές. Όλα τα δεδομένα έχουν ήδη αποθηκευτεί στο κατανεμημένο σύστημα αρχείων HDFS της υποδομής του μαθήματος. Στον Πίνακα 1 μπορείτε να δείτε τα URIs που μπορείτε να χρησιμοποιήσετε για να αποκτήσετε πρόσβαση σε αυτά.

Σύνολο Δεδομένων	HDFS URI
Los Angeles Crime Data (2010-2019)	hdfs://hdfs-namenode:9000/user/root/data/LA_Crime_Data_2010_2019.csv
Los Angeles Crime Data (2020-)	hdfs://hdfs-namenode:9000/user/root/data/LA_Crime_Data_2020_2025.csv
LA Police Stations	hdfs://hdfs-namenode:9000/user/root/data/LA_Police_Stations.csv
Median Household Income by Zip Code	hdfs://hdfs-namenode:9000/user/root/data/LA_income_2015.csv
2010 Census Populations by Zip Code	hdfs://hdfs-namenode:9000/user/root/data/2010_Census_Populations_by_Zip_Code.csv
MO Codes in Numerical Order	hdfs://hdfs-namenode:9000/user/root/data/MO_codes.txt

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

Βασικό data-set: Los Angeles Crime Data

Το βασικό σύνολο δεδομένων που θα χρησιμοποιηθεί στην εργασία προέρχεται από το δημόσιο αποθετήριο δεδομένων της κυβέρνησης των Ηνωμένων Πολιτειών της Αμερικής[1] . Συγκεκριμένα, περιλαμβάνει δεδομένα καταγραφής εγκλημάτων για το Los Angeles από το 2010 μέχρι σήμερα. Τα δεδομένα είναι διαθέσιμα σε csv file format στους παρακάτω συνδέσμους:

- https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z
- https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

Στους ίδιους συνδέσμους παρέχονται περιγραφές για κάθε ένα από τα 28 πεδία των δεδομένων.

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 :

2010 Census Populations by Zip Code: Ένα σύνολο δεδομένων που παρουσιάζει απογραφικά στοιχεία που αφορούν στην Κομητεία του Los Angeles για το έτος 2010. Είναι διαθέσιμο στον παρακάτω σύνδεσμο:

- https://data.lacity.org/Community-Economic-Development/2010-Census-Populations-by-Zip-Code/nxs9-385f/

Median Household Income by Zip Code (Los Angeles County): Ένα ακόμα μικρό σύνολο που περιέχει δεδομένα σχετικά με το μέσο εισόδημα ανά νοικοκυριό και ταχυδρομικό κώδικα (ZIP Code) στην Κομητεία του Los Angeles. Για διευκόλυνση, τα δεδομένα έχουν συλλεχθεί και αποθηκευθεί σε csv file format. Τα συγκεκριμένο σύνολο δεδομένων παράχθηκε με βάση τα αποτελέσματα της απογραφής του έτους 2015 και είναι διαθέσιμα στον παρακάτω σύνδεσμο:

- http://www.laalmanac.com/employment/em12c_2015.php

LA Police Stations: Μικρό σύνολο δεδομένων που περιέχει πληροφορία σχετικά με την τοποθεσία των 21 αστυνομικών τμημάτων που βρίσκονται στην πόλη του Los Angeles. Τα συγκεκριμένα δεδομένα προέρχονται από δημόσιο αποθετήριο δεδομένων του δήμου του Los Angeles και είναι διαθέσιμα σε csv file format στον παρακάτω σύνδεσμο:

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

MO Codes in Numerical Order: Σύνολο δεδομένων που αναφέρεται σε δραστηριότητες ή χαρακτηριστικά των δραστών ("*Modus Operandi*"). Οι συγκεκριμένοι κωδικοί αντιστοιχούν στη στήλη Mocodes του **Los Angeles Crime Data**. Για διευκόλυνση σας παρέχεται σε μορφή txt, όπου ένας κωδικός βρίσκεται στην αρχή κάθε γραμμής και διαχωρίζεται από την περιγραφή με ένα κενό. Τα συγκεκριμένα δεδομένα προέρχονται από δημόσιο αποθετήριο δεδομένων του δήμου του Los Angeles και είναι διαθέσιμα σε pdf file format στον παρακάτω σύνδεσμο:

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

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

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:

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

Query 2

Να βρεθούν, για κάθε έτος, τα 3 Αστυνομικά Τμήματα με το υψηλότερο ποσοστό κλεισμένων (περατωμένων) υποθέσεων. Να τυπωθούν το έτος, τα ονόματα (τοποθεσίες) των τμημάτων, τα ποσοστά τους καθώς και οι αριθμοί του ranking τους στην ετήσια κατάταξη. Τα αποτελέσματα να δοθούν σε σειρά αύξουσα ως προς το έτος και το ranking (δείτε παράδειγμα στον Πίνακα 2).

year	precinct	closed_case_rate	#
2010	West Valley	30.57974335472044	1
2010	N Hollywood	29.23808669119627	2
2010	Mission	27.58372669119627	3

Πίνακας 2: Υπόδειγμα αποτελέσματος Query 2

Query 3

Χρησιμοποιώντας ως αναφορά τα δεδομένα της απογραφής 2010 για τον πληθυσμό και εκείνα της απογραφής του 2015 για το εισόδημα ανα νοικοκυριό, να υπολογίσετε για κάθε ZipCode του Los Angeles το μέσο ετήσιο εισόδημα ανά άτομο.

Query 4

Να υπολογιστεί, ανά αστυνομικό τμήμα, ο αριθμός εγκλημάτων που έλαβαν χώρα πλησιέστερα σε αυτό με εμπλοκή όπλων (πυροβόλων ή όχι), 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.

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. Ως ανοικτές υποθέσεις θεωρούμε όλες εκείνες που έχουν κατάσταση "*άγνωστη*" ή "*έρευνα σε εξέλιξη*" (Status Desc == UNK||Invest Cont).

3. Ως εγκλήματα στα οποία εμπλέκονται όπλα θεωρούμε όλα εκείνα τα περιστατικά που αντιστοιχούν σε Mocodes η περιγραφή των οποίων περιλαμβάνει τους όρους "*gun*" ή "*weapon*".

4. Κάποιες εγγραφές (λανθασμένα) αναφέρονται στο Null Island. Θα πρέπει να φιλτραριστούν εκτός του συνόλου δεδομένων και να μη λαμβάνονται υπόψη στον υπολογισμό, γιατί θα επηρεάσουν αρνητικά τα αποτελέσματα των queries σας σχετικά με την απόσταση!

5. Είστε ελεύθεροι να επιλέξετε την υλοποίηση του υπολογισμού απόστασης μεταξύ δύο σημείων με οποιονδήποτε τρόπο της αρεσκείας σας.

Ζητούμενα

1. Να ολοκληρωθεί η διαδικασία σύνδεσης με την απομακρυσμένη υποδομή kubernetes που περιγράφεται στους οδηγούς του μαθήματος. Επίσης, να ολοκληρωθεί η διαδικασία παραμετροποίησης του Spark Job History Server μέσω docker και docker compose τοπικά, στο μηχάνημά σας, ώστε να αντλεί τα δεδομένα από τις εκτελέσεις σας στην απομακρυσμένη υποδομή (HDFS). **(5%)**

2. Να γραφτεί κώδικας που θα διαβάσει τα αρχεία δεδομένων και με την κατάλληλη επεξεργασία θα τα αποθηκεύσει σε μορφή parquet στο HDFS, στο παρακάτω path:

   ```
   hdfs://hdfs-namenode:9000/user/{username}/data/parquet/
   ```

   Δώστε ιδιαίτερη προσοχή στο αρχείο που αντιστοιχεί στο dataset **MO Codes in Numerical Order**, καθώς δε βρίσκεται σε εξαρχής σε μορφή που το Spark μπορεί να αναγνωρίσει. **(10%)**

3. Να υλοποιηθεί το **Query 1** χρησιμοποιώντας τα RDD και DataFrame APIs (με udf και χωρίς). Σχολιάστε τις διαφορές στην επίδοση μεταξύ των διαφορετικών υλοποιήσεών σας. **(15%)**

4. Να υλοποιηθεί το **Query 2** χρησιμοποιώντας τα RDD, DataFrame και SQL APIs. Σχολιάστε τις επιδόσεις των υλοποιήσεών σας. **(15%)**

5. Να υλοποιηθεί το **Query 3** χρησιμοποιώντας τα RDD και DataFrame APIs. Στην υλοποίηση με DataFrames πειραματιστείτε κάνοντας την εισαγωγή των δεδομένων με χρήση αρχείων csv και parquet και σχολιάστε πώς επηρεάζεται η εκτέλεση σε κάθε περίπτvim . **(15%)**

6. Να υλοποιηθεί το **Query 4** χρησιμοποιώντας το DataFrame API. Εφαρμόστε οριζόντια και κάθετη κλιμάκωση (horizontal and vertical scaling) των πόρων που δεσμεύετε για την εκτέλεση χρησιμοποιώντας τα κατάλληλα spark configurations (spark.executor.instances, spark.executor.cores, spark.executor.memory).

   - Καλείστε να εκτελέσετε την υλοποίησή σας χρησιμοποιώντας συνολικούς πόρους 8 cores και 16GB μνήμης με τα παρακάτω configurations:
     - 2 executors × 4 cores/8GB memory
     - 4 executors × 2 cores/4GB memory
     - 8 executors × 1 core/2 GB memory
   - 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. **(25%)**

7. Για κάθε ένα από τα joins των υλοποιήσεων των **Query 3** και **Query 4** να αναφερθεί η επιλογή στρατηγικής (BROADCAST, MERGE, SHUFFLE_HASH, SHUFFLE_REPLICATE_NL κλπ.) που κάνει ο Catalyst Optimizer του Spark με χρήση της εντολής explain ή του Job History Server (να συμπεριληφθεί το σχετικό output ή screenshot). Να σχολιάσετε βάσει θεωρίας αν η επιλογή δικαιολογείται ή όχι βάσει των χαρακτηριστικών του join. **(15%)**

Παραδοτέα - Όροι Υποβολής

- Η εργασία να εκπονηθεί σε ομάδες το πολύ των 2 ατόμων.
- **ΠΡΟΘΕΣΜΙΑ ΥΠΟΒΟΛΗΣ: 22 ΙΟΥΝΙΟΥ 2025, 23:59**.
- Το παραδοτέο της εργασίας θα υποβληθεί στη σελίδα του μαθήματος στο helios, σε link που θα ανοίξει αργότερα.
- Η εργασία αποτελεί το 40% του συνολικού βαθμού του μαθήματος. Για να υπολογιστεί ο βαθμός της εργασίας, η κάθε ομάδα θα πρέπει να υποβάλει σχετική αναφορά και να περάσει επιτυχώς την υποχρεωτική προφορική εξέταση στο αντικείμενο της εργασίας. Η εξέταση θα γίνει μετά την παράδοση της εργασίας (θα ανακοινωθεί σχετικό πρόγραμμα).
- 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 αποκλειστικά τις απαντήσεις στα ζητούμενα, καθώς και ένα link σε αποθετήριο (github, gitlab, bitbucket, κλπ.) που θα περιέχει όλους 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. Είναι υποχρεωτικό να χρησιμοποιήσετε πόρους από την υποδομή του εργαστηρίου που σας έχει παραχωρηθεί για τις ανάγκες της εργασίας. Επίσης, είναι υποχρεωτικό να υπάρχει στην υποδομή του kubernetes ίχνος της τελευταίας εκτέλεσης κάθε υλοποίησής σας. 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, προκειμένου όλοι να έχουν πρόσβαση στις απαντήσεις/επεξηγήσεις. Μη στέλνετε τις απορίες σας στα email των διδασκόντων/βοηθών αλλά να τις υποβάλλετε όπως αναφέρεται.