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 2023-24, 9ο Εξάμηνο

Διδάσκων: Δημήτριος Τσουμάκος

Υπεύθυνος Εργαστηρίου: Νικόλαος Χαλβαντζής

3 Νοεμβρίου 2023

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>=3.4). Για την εγκατάσταση και διαμόρφωση του κατάλληλου περιβάλλοντος εργασίας, υπάρχει η πρόβλεψη για την χρήση εικονικών μηχανών από το public cloud *~okeanos-knossos*[1]. 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.

[1] https://okeanos-knossos.grnet.gr/home/

Δεδομένα

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[2] . 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://catalog.data.gov/dataset/crime-data-from-2010-to-2019
- https://catalog.data.gov/dataset/crime-data-from-2020-to-present

Εναλλακτικά, τα ίδια δεδομένα είναι διαθέσιμα και από αποθετήριο του δήμου του Los Angeles 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

Σε αυτούς τους συνδέσμους παρέχονται περιγραφές για κάθε ένα από τα 28 πεδία των δεδομένων, *οι οποίες θα είναι χρήσιμες στα πλαίσια της παρούσας εργασίας*, καθώς και ορισμένα σχετικά ή επεξηγηματικά σύνολα δεδομένων (στο τμήμα "**Attachments**").

Δευτερεύοντα data-sets

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

LA Police Stations: Μικρό σύνολο δεδομένων που περιέχει δεδομένα 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

Median Household Income by Zip Code (Los Angeles County): Ένα ακόμα μικρό σύνολο δεδομένων που περιέχει δεδομένα σχετικά με το μέσο εισόδημα ανά νοικοκυριό και ταχυδρομικό κώδικα (ZIP Code) στην Κομητεία του Los Angeles. Τα συγκεκριμένα δεδομένα παράχθηκαν με βάση τα αποτελέσματα απογραφών των ετών 2015, 2017, 2019 και 2021 από την ιστοσελίδα Los Angeles Almanac και είναι διαθέσιμα στους παρακάτω συνδέσμους:

- http://www.laalmanac.com/employment/em12c_2015.php
- http://www.laalmanac.com/employment/em12c_2017.php
- http://www.laalmanac.com/employment/em12c_2019.php
- http://www.laalmanac.com/employment/em12c.php

Προς διευκόλυνση, τα δεδομένα έχουν συλλεχθεί και είναι διαθέσιμα σε .csv file format στον παρακάτω σύνδεσμο:

- http://www.dblab.ece.ntua.gr/files/classes/data.tar.gz

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

Για τις ανάγκες της παρούσας εργασίας θα χρειαστεί πρόσβαση μόνο στο κομμάτι που αφορά το έτος 2015. Φυσικά, εσείς μπορείτε να πειραματιστείτε και με περισσότερα.

Reverse Geocoding: Ο όρος "geocoding"(γεωκωδικοποίηση) αναφέρεται συνήθως στη μετάφραση μιας διεύθυνσης σε μια τοποθεσία σε σύστημα συντεταγμένων. Η αντίστροφη διαδικασία, δηλαδή η αντιστοίχιση ενός ζεύγους συντεταγμένων σε μια διεύθυνση, είναι γνωστή ως "reverse geocoding" (αντίστροφη γεοκωδικοποίηση). Στα πλαίσια της εργασίας, θα χρειαστεί να γίνει αντιστοίχιση συντεταγμένων (latitude, longitude) σε ταχυδρομικούς κώδικες (ZIP Codes) εντός της πόλης του Los Angeles. Αυτό μπορεί να πραγματοποιηθεί προγραμματιστικά με τη βοήθεια web services γνωστών ως geocoders και βιβλιοθηκών όπως η geopy[3]. Επειδή η διαδικασία, λόγω του latency των web services, είναι αργή, σας παρέχεται σύνολο δεδομένων με reverse geocoding πληροφορία που καλύπτει τοποθεσίες που θα χρειαστούν στα πλαίσια της εργασίας. Φυσικά, ενθαρρύνουμε τους σπουδαστές να πειραματιστούν και εάν το επιθυμούν να δοκιμάσουν την υλοποίηση της συγκεκριμένης λειτουργίας. Το σύνολο δεδομένων είναι διαθέσιμο σε .csv file format στον παρακάτω σύνδεσμο:

- http://www.dblab.ece.ntua.gr/files/classes/data.tar.gz

Ερωτήματα

Query 1

Να βρεθούν, για **κάθε** έτος, οι 3 μήνες με τον υψηλότερο αριθμό καταγεγραμμένων εγκλημάτων. Ζητείται να τυπωθούν ανά έτος οι συγκεκριμένοι μήνες, ο συνολικός αριθμός εγκληματικών δράσεων που καταγράφηκαν τότε, καθώς και η θέση του συγκεκριμένου μήνα στην κατάταξη μέσα του αντίστοιχου έτους. Τα αποτελέσματα να δοθούν σε σειρά αύξουσα ως προς το έτος και φθίνουσα ως προς τον αριθμό καταγραφών (δείτε παράδειγμα στον Πίνακα 1).

year	month	crime_total	#
2010	2	2145	1
2010	3	1492	2
2010	5	54	3
2011	12	4632	1
2011	6	2312	2
2011	4	312	3

Πίνακας 1: Υπόδειγμα αποτελέσματος Query 1

Query 2

Να ταξινομηθούν τα τμήματα της ημέρας ανάλογα με τις καταγραφές εγκλημάτων που έλαβαν χώρα στο δρόμο (STREET), με φθίνουσα σειρά. Θεωρείστε τα εξής τμήματα μέσα στη μέρα:

- Πρωί: 5.00πμ – 11.59πμ
- Απόγευμα: 12.00μμ – 4.59μμ
- Βράδυ: 5.00μμ – 8.59μμ
- Νύχτα: 9.00μμ – 3.59πμ

[3] https://geopy.readthedocs.io/en/stable/#module-geopy.geocoders

Query 3

Να βρεθεί η καταγωγή (descent) των καταγεγραμμένων θυμάτων εγκλημάτων στο Los Angeles για το έτος 2015 στις 3 περιοχές (ZIP Codes) με το υψηλότερο και τις 3 περιοχές (ZIP Codes) με το χαμηλότερο εισόδημα ανά νοικοκυριό. Τα αποτελέσματα να τυπωθούν από το υψηλότερο στο χαμηλότερο αριθμό θυμάτων ανά φυλετικό γκρουπ (δείτε παράδειγμα αποτελέσματος στον Πίνακα 2).

Victim Descent	#
White	413
Black	274
Unknown	132
Hispanic/Latin/Mexican	12

Πίνακας 2: Υπόδειγμα αποτελέσματος Query 3

Query 3 Tips:

1. Victimless crimes exist: Φιλτράρετε εκτός του συνόλου εργασίας σας τα data points για τα οποία δεν υπάρχει καταγραφή θύματος ή της καταγωγής του.

2. Στις περιπτώσεις που στο σύνολο δεδομένων **Reverse Geocoding** αναφέρονται περισσότερα του ενός ZIP Codes για ένα ζεύγος συντεταγμένων, θα πρέπει να χρησιμοποιήσετε ένα από αυτά (π.χ., το πρώτο).

3. Οι περιοχές που καλύπτονται στο **Median Household Income by Zip Code** σύνολο δεδομένων αφορούν την ευρύτερη περιοχή της Κομητείας του Los Angeles.

4. Μπορείτε, αν θέλετε, να χρησιμοποιήσετε την αντιστοίχιση των κωδικών καταγωγής με την περιγραφή που αναφέρονται στις πληροφορίες που συνοδεύουν το σύνολο δεδομένων.

Query 4

Για το τελευταίο ερώτημα, ο στόχος είναι να εξεταστεί κατά πόσον τα εγκλήματα που καταγράφονται στην πόλη του Los Angeles αντιμετωπίζονται από το πλησιέστερο στον τόπο εγκλήματος αστυνομικό τμήμα ή όχι. Για το λόγο αυτό, θα εκτελέσουμε δύο ζεύγη παρόμοιων ερωτημάτων και θα συγκρίνουμε τα αποτελέσματα:

- α) Να υπολογιστεί ανά έτος ο αριθμός εγκλημάτων με καταγραφή χρήσης οποιασδήποτε μορφής πυροβόλων όπλων και η μέση απόσταση (σε km) των σημείων όπου αυτά έλαβαν χώρα από το **αστυνομικό τμήμα που ανέλαβε την έρευνα για το περιστατικό**. Τα αποτελέσματα να εμφανιστούν ταξινομημένα κατά έτος σε αύξουσα σειρά. β) Επίσης, να υπολογιστεί ανά αστυνομικό τμήμα ο αριθμός εγκλημάτων με καταγραφή χρήσης οποιαδήποτε μορφής όπλων που του ανατέθηκε καθώς και η μέση απόσταση του εκάστοτε τόπου εγκλήματος από το αστυνομικό τμήμα. Τα αποτελέσματα να εμφανιστούν ταξινομημένα κατά αριθμό περιστατικών, με φθίνουσα σειρά (δείτε παράδειγμα στον Πίνακα 3).

- α) Να υπολογιστεί ανά έτος ο αριθμός εγκλημάτων με καταγραφή χρήσης οποιασδήποτε μορφής πυροβόλων όπλων και η μέση απόσταση (σε km) των σημείων όπου αυτά έλαβαν χώρα από το **πλησιέστερο σε αυτά αστυνομικό τμήμα**. Τα αποτελέσματα να εμφανιστούν ταξινομημένα κατά έτος σε αύξουσα σειρά. β) Επίσης, να υπολογιστεί ανά αστυνομικό τμήμα ο αριθμός εγκλημάτων με καταγραφή χρήσης οποιαδήποτε μορφής όπλων που έλαβαν χώρα πλησιέστερα σε

αυτό καθώς και η μέση απόσταση των σημείων από το εκάστοτε αστυνομικό τμήμα. Τα αποτελέσματα να εμφανιστούν ταξινομημένα κατά αριθμό περιστατικών, με φθίνουσα σειρά (δείτε παράδειγμα στον Πίνακα 4). .

year	average_distance	#
2010	2.352	7232
2011	2.312	6763
2012	2.276	8487
2013	2.392	9745

Πίνακας 3: Υπόδειγμα αποτελέσματος Query 4a)

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

Πίνακας 4: Υπόδειγμα αποτελέσματος Query 4b)

Tips:

1. Κάποιες εγγραφές (λανθασμένα) αναφέρονται στο Null Island. Θα πρέπει να φιλτραριστούν και να μη λαμβάνονται υπόψη στον υπολογισμό.

2. Τα περιστατικά που αφορούν χρήση πυροβόλων όπλων οποιασδήποτε μορφής αντιστοιχούν σε κωδικούς της στήλης “Weapon Used Cd” της μορφής “1xx”.

3. Οι κωδικοί της στήλης “AREA ” του **Los Angeles Crime Data** αντιστοιχούν σε εκείνους της στήλης “PRECINCT” του **LA Police Stations** και αφορούν το αστυνομικό τμήμα που ανέλαβε το κάθε περιστατικό.

4. Είστε ελεύθεροι να επιλέξετε την υλοποίηση του υπολογισμού απόστασης μεταξύ δύο σημείων με οποιονδήποτε τρόπο της αρεσκείας σας. Ενδεικτικά, σας δίνεται μια υλοποίηση σε Python, με χρήση της βιβλιοθήκης geopy[4].

```python
import geopy.distance

# calculate the distance between two points [lat1, long1], [lat2, long2] in km
def get_distance(lat1, long1, lat2, long2):
    return geopy.distance.geodesic((lat1, long1), (lat2, long2)).km
```

Ζητούμενα

1. Να εγκαταστήσετε και διαμορφώσετε κατάλληλα την πλατφόρμα εκτέλεσης Apache Spark ώστε να εκτελείται πάνω από το διαχειριστή πόρων του Apache Hadoop, YARN. Να μορφοποιήσετε κατάλληλα το Apache Hadoop Distributed File System για την είσοδο και έξοδο του λογισμικού που θα αναπτύξετε. Το περιβάλλον εργασίας σας θα πρέπει να είναι πλήρως κατανεμημένο, με 2 ή περισσότερους, εάν το επιθυμείτε, κόμβους. Θα πρέπει, τέλος, οι web εφαρμογές των HDFS, YARN και Spark History Server να είναι διαθέσιμες και προσβάσιμες. (10%)

[4] https://geopy.readthedocs.io/en/stable/

2. Να δημιουργηθεί ένα DataFrame από το βασικό σύνολο δεδομένων. Διατηρώντας τα αρχικά ονόματα στηλών, να προσαρμόσετε τους τύπους των δεδομένων όπως αναφέρεται παρακάτω:

   - Date Rptd: date
   - DATE OCC: date
   - Vict Age: integer
   - LAT: double
   - LON: double

   Να τυπωθεί ο συνολικός αριθμός γραμμών του συγκεκριμένου συνόλου δεδομένων, και ο τύπος κάθε στήλης. (5%)

3. Να υλοποιηθεί το **Query 1** χρησιμοποιώντας τα DataFrame και SQL APIs. Να εκτελέστε και τις δύο υλοποιήσεις με 4 Spark executors. Υπάρχει διαφορά στην επίδοση μεταξύ των δύο APIs; Να αιτιολογήσετε την απάντησή σας. (15%)

4. Να υλοποιηθεί το **Query 2** χρησιμοποιώντας τα DataFrame/SQL και RDD API. Να αναφέρετε και να συγκρίνετε τους χρόνους εκτέλεσης για 4 Spark executors. (20%)

5. Να υλοποιηθεί το **Query 3** χρησιμοποιώντας το DataFrame/SQL API. Να αναφέρετε και να σχολιάσετε τους χρόνους εκτέλεσης για 2, 3 και 4 Spark executors. (20%)

6. Να υλοποιηθεί το **Query 4** χρησιμοποιώντας το DataFrame/SQL API. (20%)

7. Για τα joins των **Query 3** και **Query 4**, χρησιμοποιήστε τις μεθόδους hint & explain των DataFrame/SQL APIs ώστε να εκτελεστούν με διαφορετικό τρόπο (BROADCAST, MERGE, SHUFFLE_HASH, SHUFFLE_REPLICATE_NL) και να πάρετε το πλάνο μέσω κειμένου και γραφικά απο το Spark History UI. Να σχολιάσετε ποιά (ποιές) από τις διαθέσιμες στρατηγικές join του Spark είναι καταλληλότερη(ες) και γιατί. (10%)

Παραδοτέα - Όροι Υποβολής

- Η εργασία να εκπονηθεί σε ομάδες το πολύ των 2 ατόμων.
  ΠΡΟΘΕΣΜΙΑ ΥΠΟΒΟΛΗΣ: ΤΕΤΑΡΤΗ 10 ΙΑΝΟΥΑΡΙΟΥ 2024, 23.59.

- Το παραδοτέο της εργασίας θα υποβληθεί στο helios στην σελίδα του μαθήματος σε link που θα ανοίξει αργότερα.

- Η εργασία αποτελεί το 30% 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 αναρτηθεί 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, etc.) 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. Επιπλέον, σας δίνεται η δυνατότητα να χρησιμοποιήσετε δικούς σας πόρους (π.χ. προσωπικούς Η/Υ, VM) ή πόρους από την υπηρεσία *~okeanos-knossos*. 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 τις απορίες σας στα email των διδασκόντων/βοηθών αλλά να τις υποβάλλετε όπως αναφέρεται.