ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διαχείριση Δεδομένων Μεγάλης Κλίμακας

Ακαδημαϊκό έτος 2023-24, Εαρινό Εξάμηνο

Διδάσκοντες: Δημήτριος Τσουμάκος, Ιωάννης Κωνσταντίνου

Υπεύθυνος Εργαστηρίου: Νικόλαος Χαλβαντζής

10 Μαΐου 2024

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.

Δεδομένα

Βασικό data-set: Los Angeles Crime Data

Το βασικό σύνολο δεδομένων που θα χρησιμοποιηθεί στην εργασία προέρχεται από το δημόσιο αποθετήριο δεδομένων της πόλης του Los Angeles[2] . Συγκεκριμένα, περιλαμβάνει δεδομένα καταγραφής

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

[2] https://data.lacity.org/

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/api/views/63jg-8b9z/rows.csv?accessType=DOWNLOAD
- https://data.lacity.org/api/views/2nrs-mtv8/rows.csv?accessType=DOWNLOAD

Εκτός από τα δεδομένα σε μορφή .csv, στους παρακάτω συνδέσμους παρέχονται περιγραφές για κάθε ένα από τα 28 πεδία του dataset οι οποίες θα είναι χρήσιμες στα πλαίσια της εργασίας, καθώς και ορισμένα σχετικά ή επεξηγηματικά σύνολα δεδομένων (στο τμήμα "**Attachments**").

- 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

Δευτερεύοντα data-sets

Συμπληρωματικά με τα παραπάνω δεδομένα, θα χρησιμοποιηθεί μια σειρά data-sets μικρότερου όγκου τα οποία επίσης είναι δημοσίως διαθέσιμα:

LA Police Stations: Σύνολο δεδομένων που περιέχει την τοποθεσία των 21 αστυνομικών τμημάτων της πόλης του Los Angeles. 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 σχετικά μικρό σύνολο δεδομένων που περιέχει πληροφορίες σχετικά με το μέσο εισόδημα ανά νοικοκυριό για τις περιοχές της Κομητείας του Los Angeles ανά ταχυδρομικό κώδικα (ZIP Code). Τα συγκεκριμένα δεδομένα βασίζονται στα αποτελέσματα των 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

Για τις ανάγκες της παρούσας εργασίας θα χρειαστεί η πρόσβαση μόνο στο κομμάτι που αφορά το έτος 2015. Προς διευκόλυνση, τα σύνολα δεδομένων έχουν συλλεχθεί και είναι διαθέσιμα σε .csv file format στον παρακάτω σύνδεσμο:

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

Reverse Geocoding: Ο όρος "geocoding"(γεωκωδικοποίηση) αναφέρεται 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 latency των web services – είναι αργή, σας παρέχεται σύνολο δεδομένων που καλύπτει τοποθεσίες που θα χρειαστούν στα πλαίσια της εργασίας. Το σύνολο δεδομένων είναι διαθέσιμο σε .csv file format στον παρακάτω σύνδεσμο:

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

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

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 περιστατικών, καθώς και η θέση του συγκεκριμένου μήνα στην κατάταξη μέσα στο αντίστοιχο έτος. 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	ranking
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 – 4.59πμ

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 (σε **δύο ξεχωριστούς** πίνακες). Τα αποτελέσματα να τυπωθούν από το υψηλότερο στο χαμηλότερο αριθμό θυμάτων ανά φυλετικό γκρουπ (δείτε παράδειγμα αποτελέσματος στον Πίνακα 2).

victim descent	total victims
White	413
Black	274
Unknown	132
Hispanic/Latin/Mexican	12

Πίνακας 2: Υπόδειγμα αποτελέσματος 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 (που είναι μεγαλύτερη από την πόλη του Los Angeles). Σε περίπτωση join μεταξύ του βασικού συνόλου δεδομένων και του συγκεκριμένου, επιλέξτε προσεκτικά το είδος του join που θα χρησιμοποιήσετε.

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, να υπολογιστεί ανά αστυνομικό τμήμα ο αριθμός εγκλημάτων με καταγραφή χρήσης οποιαδήποτε μορφής πυροβόλων όπλων που αυτό ανέλαβε, καθώς και η μέση απόσταση του εκάστοτε περιστατικού από το αστυνομικό τμήμα. Τα αποτελέσματα να εμφανιστούν ταξινομημένα κατά αριθμό περιστατικών, με φθίνουσα σειρά (δείτε παράδειγμα στον Πίνακα 3).

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

2. 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 έχει αναλάβει το κάθε περιστατικό. Για να υπολογίσετε τις αποστάσεις, θα χρειαστεί να γίνει ένα join μεταξύ των δύο data-sets πάνω στις συγκεκριμένες στήλες.

3. 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!

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
```

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

Ζητούμενα

1. Να εγκαταστήσετε και διαμορφώσετε κατάλληλα το σύστημα κατανεμημένης επεξεργασίας δεδομένων Apache Spark, καθώς και το κατανεμημένο σύστημα αρχείων Hadoop Distributed File System (HDFS) ακολουθώντας τις οδηγίες και εκτελώντας τα scripts που σας έχουν δοθεί. 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, και Spark Job History Server να είναι διαθέσιμες και προσβάσιμες. (5%)

2. Να δημιουργηθεί ένα directory στο HDSF όπου θα αποθηκεύσετε τα σύνολα δεδομένων σε .csv μορφή. Να αναφερθούν οι εντολές που θα χρησιμοποιήσετε για αυτό το σκοπό και να συμπεριληφθεί στην αναφορά screenshot όπου θα φαίνεται η κατάσταση του συστήματος αρχείων με τα δεδομένα διαθέσιμα. Να γράψετε κώδικα spark που μετατρέπει το κυρίως data set parquet[5] file format και αποθηκεύει τα παραγόμενα parquet αρχεία στο HDFS. (5%)

3. Να υλοποιηθεί το **Query 1** χρησιμοποιώντας τα DataFrame και SQL APIs. Να εκτελέστε και τις δύο υλοποιήσεις διαβάζοντας τα δεδομένα από τα .csv αρχεία και καταγράφοντας τους χρόνους εκτέλεσης κάθε φορά. Στη συνέχεια, να επαναλάβετε το ίδιο πείραμα χρησιμοποιώντας το parquet φορμάτ για τα αρχεία δεδομένων. Αποτυπώστε σε ένα πίνακα τις μετρήσεις σας για τους 4 δυνατούς συνδυασμούς μεταξύ προγραμματιστικών APIs και φορμάτ αρχείων εισόδου. Παρατηρείται διαφορά στην επίδοση μεταξύ των δύο τύπων αρχείων και μεταξύ των δύο APIs; Να σχολιάσετε τα αποτελέσματα σας. (20%)

4. Να υλοποιηθεί το **Query 2** χρησιμοποιώντας τα DataFrame (ή SQL) και RDD APIs. Για το διάβασμα των δεδομένων να χρησιμοποιήσετε το .csv φορμάτ. Να αναφέρετε και να σχολιάσετε τους χρόνους εκτέλεσης. (15%)

5. Να υλοποιηθεί το **Query 3** χρησιμοποιώντας το DataFrame (ή SQL) API. Για τα joins του κώδικα της υλοποίησής σας να αναφέρετε τη στρατηγική που επιλέγει ο Catalyst Optimizer. Στη συνέχεια, χρησιμοποιήστε τη μέθοδο hint των DataFrame/SQL APIs ώστε τα joins να εκτελεστούν με διαφορετικό τρόπο, παρακάμπτωντας τον Optimizer. Πειραματιστείτε με τις διαφορετικές υλοποιήσεις που προσφέρει το API του Spark (BROADCAST, MERGE, SHUFFLE_HASH, SHUFFLE_REPLICATE_NL). Μπορείτε να παρατηρήσετε τις διαφορές στο physical plan της εκτέλεσης γραφικά απο το web interface του Spark Job History Server. Να σχολιάσετε ποιά (ποιές) από τις διαθέσιμες στρατηγικές join του Spark είναι καταλληλότερη(ες), και γιατί. (20%)

6. Για το **Query 4**, θα χρειαστεί να εκτελέσετε ένα join μεταξύ του βασικού συνόλου δεδομένων (**Los Angeles Crime Data**) και του **LA Police Stations**. Καλείστε να πειραματιστείτε, υλοποιώντας δύο απλούς join αλγορίθμους και κάνοντας χρήση του RDD API. Συγκεκριμένα, σας ζητείται να υλοποιήσετε τους αλγορίθμους broadcast join και repartition join. Πληροφορίες σχετικά με τους ζητούμενους αλγορίθμους καθώς και ψευδοκώδικα, μπορείτε να βρείτε στο άρθρο που παρατίθεται στη σχετική βιβλιογραφία παρακάτω (broadcast join: sub-section 3.2, pseudocode A.4 – repartition join: sub-section 3.1, pseudocode A.4). (20%)

7. Να υλοποιηθεί το **Query 4** χρησιμοποιώντας το DataFrame (ή SQL) API. (15%)

[5] https://parquet.apache.org/docs/file-format/

Προσοχή στα εξής:

- Όπου δεν αναφέρεται σχετική οδηγία, μπορείτε να πραγματοποιήσετε την είσοδο των δεδομένων χρησιμοποιώντας είτε csv είτε parquet αρχεία.
- Για την καταγραφή των χρόνων εκτέλεσης των applications σας, μπορείτε να χρησιμοποιήσετε το Spark Job History Server ή να κάνετε την καταγραφή εντός του κώδικα.
- Φροντίστε όταν κάνετε καταγραφή χρόνων εκτέλεσης να επανεκκινείτε το cluster σας ή, σε περίπτωση που χρησιμοποιείτε το ίδιο spark session, να καθαρίζετε την cache με την εντολή:

```
spark.catalog.clearCache()
```

Σχετική Βιβλιογραφία:

1. Blanas, Spyros, et al. "A comparison of join algorithms for log processing in mapreduce." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. 2010.

Παραδοτέα - Όροι Υποβολής

- Η εργασία να εκπονηθεί σε ομάδες το πολύ των 2 ατόμων.
- **ΠΡΟΘΕΣΜΙΑ ΥΠΟΒΟΛΗΣ: 12 Ιουνίου 2024, 23:55**.
- Το παραδοτέο της εργασίας θα υποβληθεί στο helios στην σελίδα του μαθήματος 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, 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, προκειμένου όλοι να έχουν πρόσβαση στις απαντήσεις/επεξηγήσεις. Μη στέλνετε τις απορίες σας στα email των διδασκόντων/βοηθών αλλά να τις υποβάλετε όπως αναφέρεται.