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 2024-25, 9ο Εξάμηνο

Διδάσκων: Δημήτριος Τσουμάκος

Υπεύθυνος Εργαστηρίου: Νικόλαος Χαλβαντζής

9 Δεκεμβρίου 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>=3.5). 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/CrimeData/Crime_Data_from_2010_to_2019_20241101.csv
Los Angeles Crime Data (2020-)	s3://initial-notebook-data-bucket-dblab-905418150721/CrimeData/Crime_Data_from_2020_to_Present_20241101.csv
LA Police Stations	s3://initial-notebook-data-bucket-dblab-905418150721/LA_Police_Stations.csv
Median Household Income by Zip Code	s3://initial-notebook-data-bucket-dblab-905418150721/LA_income_2015.csv
2010 Census Blocks	s3://initial-notebook-data-bucket-dblab-905418150721/2010_Census_Blocks.geojson
Race and Ethnicity Codes	s3://initial-notebook-data-bucket-dblab-905418150721/RE_codes.csv

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

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 Blocks (Los Angeles County): Ένα σύνολο δεδομένων που παρουσιάζει απογραφικά στοιχεία που αφορούν στην Κομητεία του Los Angeles για το έτος 2010 σε `geojson` format. Συνοδεύεται από αρχείο με περιγραφές των πεδίων του (`2010_Census_Blocks_fields.csv`). Είναι διαθέσιμο στον παρακάτω σύνδεσμο:

- `https://data.lacounty.gov/maps/lacounty::2010-census-blocks`

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. 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 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`

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

[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 περιοχή του Los Angeles τα παρακάτω: Το μέσο ετήσιο εισόδημα ανά άτομο και την αναλογία συνολικού αριθμού εγκλημάτων ανά άτομο. Τα αποτελέσματα να συγκεντρωθούν σε ένα πίνακα.

Query 4

Να βρεθεί το φυλετικό προφίλ των καταγεγραμμένων θυμάτων εγκλημάτων (`Vict Descent`) στο Los Angeles για το έτος 2015 στις 3 περιοχές με το υψηλότερο κατά κεφαλήν εισόδημα. Να γίνει το ίδιο για τις 3 περιοχές με το χαμηλότερο εισόδημα. Να χρησιμοποιήσετε την αντιστοίχιση των κωδικών καταγωγής με την πλήρη περιγραφή από το σύνολο δεδομενων **Race and Ethnicity codes**. Τα αποτελέσματα να τυπωθούν σε δύο ξεχωριστούς πίνακες από το υψηλότερο στο χαμηλότερο αριθμό θυμάτων ανά φυλετικό γκρουπ (δείτε παράδειγμα αποτελέσματος στον Πίνακα 3).

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

Πίνακας 3: Υπόδειγμα αποτελέσματος Query 4

Query 5

Να υπολογιστεί, ανά αστυνομικό τμήμα, ο αριθμός εγκλημάτων που έλαβαν χώρα πλησιέστερα σε αυτό, 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 4).

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 5.

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 περιοχές του Los Angeles ορίζονται από τη στήλη `COMM` του **2010 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 και RDD APIs. Να εκτελέσετε και τις δύο υλοποιήσεις με 4 Spark executors. Υπάρχει διαφορά στην επίδοση μεταξύ των δύο APIs; Αιτιολογήσετε την απάντησή σας. (20%)

2. α) Να υλοποιηθεί το **Query 2** χρησιμοποιώντας τα DataFrame και SQL APIs. Να αναφέρετε και να συγκρίνετε τους χρόνους εκτέλεσης μεταξύ των δύο υλοποιήσεων. (10%)

   β) Να γράψετε κώδικα Spark που μετατρέπει το κυρίως data set σε `parquet`[2] file format και αποθηκεύει ένα μοναδικό `.parquet` αρχείο στο S3 bucket της ομάδας σας. Επιλέξτε μία από τις δύο υλοποιήσεις του υποερωτήματος α) (DataFrame ή SQL) και συγκρίνετε τους χρόνους εκτέλεσης της εφαρμογής σας όταν τα δεδομένα εισάγονται σαν `.csv` και σαν `.parquet`. (10%)

3. Να υλοποιηθεί το **Query 3** χρησιμοποιώντας DataFrame ή SQL API. Χρησιμοποιήστε τις μεθόδους `hint` & `explain` για να βρείτε ποιες στρατηγικές join χρησιμοποιεί ο catalyst optimizer. Πειραματιστείτε αναγκάζοντας το Spark να χρησιμοποιήσει διαφορετικές στρατηγικές (μεταξύ

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

των `BROADCAST, MERGE, SHUFFLE_HASH, SHUFFLE_REPLICATE_NL`) και σχολιάστε τα αποτελέσματα που παρατηρείτε. 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. 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. 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/2 GB 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.