ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ &
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Μάθημα: Προχωρημένα Θέματα Βάσεων Δεδομένων

Εξαμηνιαία Εργασία 19/01/2025

Όνομα Φοιτητή: Στυλιανός Χαραλάμπους

Αριθμός Ομάδας: 41 Α.Μ: 03118716

Github Repo Link:

https://github.com/ntua-el18716/advancedDatabaseSystems.git

Ζητούμενο 1

Στο **Query 1** ζητείται η ταξινόμηση ηλικιακών ομάδων των θυμάτων σε περιστατικά που περιλαμβάνουν οποιαδήποτε μορφή “βαριάς σωματικής βλάβης”(aggravated assault).

Το **Query 1** έγινε με δύο διαφορετικές υλοποιήσεις, RDD APIs και Dataframe. Και οι δύο υλοποιήσεις εκτελέστηκαν με 4 Spark Executors.

	RDD APIs	**Dataframe**
1	102.97 sec	32.38 sec
2	84.13 sec	28.69 sec
3	58.85 sec	38.04 sec
4	96.41 sec	32.21
5	88.23 sec	31.55 sec
Average	**86.12 sec**	**32.57 sec**

Παρατηρούμε ότι η υλοποίηση με RDD APIs είναι αρκετά πιο αργή από αυτή με Dataframe. Αυτό ήταν αναμενόμενο αφού τα dataframes αξιοποιούν query optimizations και αποδοτική διαχείριση μνήμης.

Αποτέλεσμα Query 1:

```
+-----------------+------+
|Vict Age Category| count|
+-----------------+------+
|            Adult|121093|
|      Young Adult| 33605|
|         Children| 15928|
|          Elderly|  5985|
+-----------------+------+
```

Ζητούμενο 2

(α)

Στο **Query 2** ζητείται η εύρεση για κάθε έτος τα 3 Αστυνομικά Τμήματα με το υψηλότερο ποσοστό περατωμένων υποθέσεων.

Το **Query 2** έγινε με δύο διαφορετικές υλοποιήσεις, Dataframe και SQL APIs. Παρατηρούμε πως και οι δύο υλοποιήσεις έχουν παρόμοιους χρόνους εκτέλεσης.

	Dataframe	**SQL APIs**
1	17.10 sec	18.41 sec
2	17.71 sec	16.59 sec
3	19.21 sec	16.02 sec
4	16.76 sec	15.23 sec
5	24.89 sec	26.39 sec
Average	**19.13 sec**	**18.52 sec**

(β)

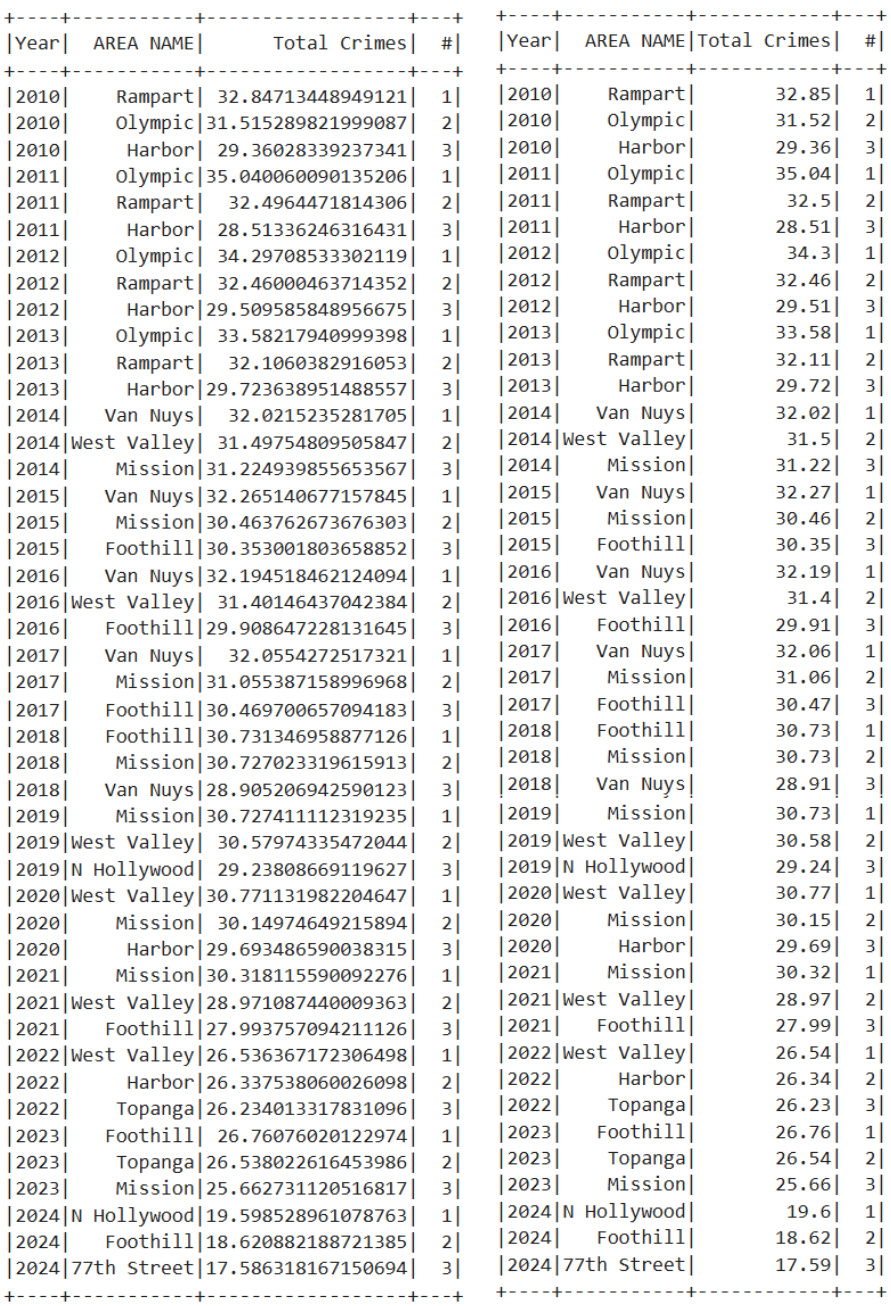
Για το δεύτερο σκέλος του ζητήματος δύο κάνουμε join τα δύο datasets(crimes 2010-2019, 2020-today) και στην συνέχεια τα μετατρέπουμε σε ένα parquet αρχείο το οποίο και αποθηκεύουμε στο S3 bucket της ομάδας.

Συγκρίνοντας τους χρόνους με το dataframe παρατηρούμε ότι με το parquet είναι λίγο πιο γρήγορο.

	Dataframe	**Parquet**
1	17.10 sec	8.59 sec
2	17.71 sec	6.59 sec
3	19.21 sec	6.68 sec

4	16.76 sec	6.31 sec
5	24.89 sec	6.87 sec
Average	**19.13 sec**	**7.01 sec**

Αποτέλεσμα Query 2:



Year	AREA NAME	Total Crimes	#
2010	Rampart	32.84713448949121	1
2010	Olympic	31.515289821999087	2
2010	Harbor	29.36028339237341	3
2011	Olympic	35.040060090135206	1
2011	Rampart	32.4964471814306	2
2011	Harbor	28.51336246316431	3
2012	Olympic	34.29708533302119	1
2012	Rampart	32.46000463714352	2
2012	Harbor	29.509585848956675	3
2013	Olympic	33.58217940999398	1
2013	Rampart	32.1060382916053	2
2013	Harbor	29.723638951488557	3
2014	Van Nuys	32.0215235281705	1
2014	West Valley	31.49754809505847	2
2014	Mission	31.224939855653567	3
2015	Van Nuys	32.265140677157845	1
2015	Mission	30.463762673676303	2
2015	Foothill	30.353001803658852	3
2016	Van Nuys	32.194518462124094	1
2016	West Valley	31.40146437042384	2
2016	Foothill	29.908647228131645	3
2017	Van Nuys	32.0554272517321	1
2017	Mission	31.055387158996968	2
2017	Foothill	30.469700657094183	3
2018	Foothill	30.731346958877126	1
2018	Mission	30.727023319615913	2
2018	Van Nuys	28.905206942590123	3
2019	Mission	30.727411112319235	1
2019	West Valley	30.57974335472044	2
2019	N Hollywood	29.23808669119627	3
2020	West Valley	30.771131982204647	1
2020	Mission	30.14974649215894	2
2020	Harbor	29.693486590038315	3
2021	Mission	30.318115590092276	1
2021	West Valley	28.971087440009363	2
2021	Foothill	27.993757094211126	3
2022	West Valley	26.536367172306498	1
2022	Harbor	26.337538060026098	2
2022	Topanga	26.234013317831096	3
2023	Foothill	26.76076020122974	1
2023	Topanga	26.538022616453986	2
2023	Mission	25.662731120516817	3
2024	N Hollywood	19.598528961078763	1
2024	Foothill	18.620882188721385	2
2024	77th Street	17.586318167150694	3

Year	AREA NAME	Total Crimes	#
2010	Rampart	32.85	1
2010	Olympic	31.52	2
2010	Harbor	29.36	3
2011	Olympic	35.04	1
2011	Rampart	32.5	2
2011	Harbor	28.51	3
2012	Olympic	34.3	1
2012	Rampart	32.46	2
2012	Harbor	29.51	3
2013	Olympic	33.58	1
2013	Rampart	32.11	2
2013	Harbor	29.72	3
2014	Van Nuys	32.02	1
2014	West Valley	31.5	2
2014	Mission	31.22	3
2015	Van Nuys	32.27	1
2015	Mission	30.46	2
2015	Foothill	30.35	3
2016	Van Nuys	32.19	1
2016	West Valley	31.4	2
2016	Foothill	29.91	3
2017	Van Nuys	32.06	1
2017	Mission	31.06	2
2017	Foothill	30.47	3
2018	Foothill	30.73	1
2018	Mission	30.73	2
2018	Van Nuys	28.91	3
2019	Mission	30.73	1
2019	West Valley	30.58	2
2019	N Hollywood	29.24	3
2020	West Valley	30.77	1
2020	Mission	30.15	2
2020	Harbor	29.69	3
2021	Mission	30.32	1
2021	West Valley	28.97	2
2021	Foothill	27.99	3
2022	West Valley	26.54	1
2022	Harbor	26.34	2
2022	Topanga	26.23	3
2023	Foothill	26.76	1
2023	Topanga	26.54	2
2023	Mission	25.66	3
2024	N Hollywood	19.6	1
2024	Foothill	18.62	2
2024	77th Street	17.59	3

Ζητούμενο 3

Στο **Query 3** ζητείται η εύρεση του κατά κεφαλήν μέσου ετήσιου εισοδήματος και η αναλογία συνολικού αριθμού εγκλημάτων ανά άτομο για κάθε περιοχή του Los Angeles.

Για τον σκοπό αυτό χρησιμοποιούμε τα εξής datasets:

- 2010 Census Blocks
- Los Angeles Crime Data (2010-2019 & 2020-)
- Medium House Per Income by Zip Code
- Για την συνένωση των διαφόρων datasets κάναμε το εξής:

Για την εκτέλεση του ερωτήματος χρειάστηκε η συνένωση των πιο πάνω data sets. Η διαδικασία που ακολουθήθηκε είναι η εξής:

1. Κάνουμε join του [2010 Census Blocks] με το [Medium House Per Income by Zip Code] στο Zip Code
2. Κάνουμε groupBy το αποτέλεσμα με βάση την στήλη COMM και κάνοντας aggregate ως εξής: άθροιση του πληθυσμού, άθροιση του αριθμού των νοικοκυριών, εύρεση μέσου όρου των μέσων εισοδημάτων
3. Στην συνέχεια για να κάνουμε join τα [Los Angeles Crime Data] με το αποτέλεσμα του προηγούμενου βήματος θα χρειαστεί αρχικά να χρησιμοποιήσουμε το Sedona ώστε να βρούμε από το [Census Blocks 2010] σε ποια περιοχή αντιστοιχεί το κάθε COMM. Στην συνέχεια και πάλι με χρήση του Sedona κάνουμε join τα [Los Angeles Crime Data] (με βάση τα LAT, LON) ώστε να βρούμε τον αριθμό εγκλημάτων σε κάθε COMM.
4. Τέλος κάνουμε join με βάση την στήλη COMM το αποτέλεσμα μαζί με το αποτέλεσμα του βήματος (2).

Θα δοκιμάσουμε για το κάθε Join τις 4 στρατηγικές.

	BROADCAST	MERGE	SHUFFLE_ HASH	SHUFFLE_ REPLICATE_NL
JOIN 1	18.33 sec	16.61 sec	18.71 sec	18.43
JOIN 2	Range Join	Range Join	Range Join	Range Join
JOIN 3	30.85 sec	18.88 sec	21.56 sec	21.95

Στο Join 2 παρατηρούμε ότι δεν είναι δυνατόν να αλλάξουμε το Join Strategy αφού πάντα χρησιμοποιεί το Range Join.

Παρατηρούμε ότι όλες οι στρατηγικές έχουν περίπου τον ίδιο χρόνο εκτέλεσης(θα ανέμενα το Καρτεσιανό Γινόμενο να είναι πιο αργό)

Αποτέλεσμα Query 3 (139 γραμμές):

```
139
+------------------+----------------+-----------------------+---------------+--------------------+
|              COMM|Total Population|Median Income Per Capita|Number of Cases|   Crimes Per Capita|
+------------------+----------------+-----------------------+---------------+--------------------+
|  Pacific Palisades|           20643|        70526.2203104497|            613|0.029695296226323692|
|      Beverly Crest|           12191|       66513.90150799365|            274|0.022475596751702076|
|   Marina Peninsula|            4337|       65235.69402813004|            159|0.036661286603643074|
|Palisades Highlands|            3833|       65048.95354904471|             42|0.010957474563005479|
|            Bel Air|            8261|       63259.97685510228|            174|0.021062825323810676|
|  Mandeville Canyon|            3233|       61443.86522911051|             43|0.013300340241261985|
|          Brentwood|           29301|      60696.777650004915|            922| 0.03146650284973209|
|            Carthay|           13165|      50282.692104378286|            624| 0.04739840486137486|
|             Venice|           32625|       46575.69192582585|           2652| 0.08128735632183907|
|       Century City|           11890|       45707.53601562712|            607| 0.05105130361648444|
|      Playa Del Rey|            3158|         45522.596580114|            152|0.048131728942368585|
|        Playa Vista|            8926|      44472.100292884345|            341| 0.03820300246470984|
|    Hollywood Hills|           27895|      43713.597155829746|           1519|0.054454203262233374|
|        Studio City|           20703|       42206.35394275496|           1226| 0.05921847075303096|
|   West Los Angeles|           34165|       40983.06782689424|           1496| 0.04378750182935753|
|      South Carthay|           10093|      39642.419795898146|            465|0.046071534727038545|
|             Encino|           42349|       39546.65508835928|           1936|0.045715365179815344|
|       Miracle Mile|           16060|       38993.44837997859|            817| 0.05087173100871731|
|        Rancho Park|            6295|      38740.063860206516|            511| 0.08117553613979349|
|     Woodland Hills|           63322|      38153.839762249285|           3270| 0.05164081993619911|
+------------------+----------------+-----------------------+---------------+--------------------+
only showing top 20 rows
```

Ζητούμενο 4

Στο **Query 4** ζητείται η εύρεση του φυλετικό προφίλ των καταγεγραμμένων θυμάτων εγκλημάτων στο Los Angeles για το έτος 2015 στις 3 περιοχές με το υψηλότερο κατά κεφαλήν εισόδημα και στις 3 περιοχές με το χαμηλότερο κατά κεφαλήν εισόδημα.

Για την εύρεση των έξι περιοχών ακολουθούμε παρόμοια διαδικασία με το **Query 3**. Στην συνέχεια ενώνουμε με χρήση του Sedona τις γεωγραφικές επικράτειες των τριών περιοχών στις δύο περιπτώσεις και το κάνουμε join με τα [Los Angeles Crime Data]. Συνεχίζουμε με join με τον [Race and Ethnicity Codes].

	2 executors 1 cores 2 GB memory	**2 executors 2 cores 4 GB memory**	**2 executors 4 core 8 GB memory**
1	123.31 sec	115.29	113.29 sec
2	113.43 sec	117.29	111.21 sec
3	116.88 sec	117.45	107.27 sec
4	119.28 sec	122.77 sec	104.35 sec
5	123.30 sec	112.83 sec	107.25 sec
Average	**119.24 sec**	**117.13 sec**	**108.67 sec**

Γενικά δεν παρατηρούμε θεαματικές διαφορές στους χρόνους εκτέλεσης. Παρόλα στην Τρίτη περίπτωση διακρίνουμε ότι η εκτέλεση του query γίνεται πιο γρήγορα.

Αποτελέσματα Query 4:

```
+--------------------+----------------------------+
|   Vict Descent Full|Victims Per Race/Ethnic Group|
+--------------------+----------------------------+
|Hispanic/Latin/Me...|                         174|
|               Black|                         118|
|               White|                          85|
|               Other|                          36|
|         Other Asian|                          12|
|             Unknown|                           4|
|             Chinese|                           2|
|            Filipino|                           1|
+--------------------+----------------------------+
```

Vict Descent Full	Victims Per Race/Ethnic Group
Hispanic/Latin/Me...	583
White	379
Other	137
Black	83
Other Asian	62
Unknown	23
Filipino	8
Chinese	2
Pacific Islander	2
Korean	1
American Indian/A...	1

Ζητούμενο 5

Στο **Query 5** ζητείται ο υπολογισμός του αριθμού εγκλημάτων που έλαβαν χώρα πλησιέστερα σε κάθε αστυνομικό τμήμα καθώς και η μέση απόσταση του από τις τοποθεσίες όπου σημειώθηκαν τα συγκεκριμένα περιστατικά.

	2 executors 4 cores 8 GB memory	**4 executors 2 cores 4 GB memory**	**8 executors 2 core 4 GB memory**
1	17.79 sec	23.26 sec	23.08 sec
2	1 1.99 sec	20. 15 sec	24.96 sec
3	14.56 sec	18.54 sec	17.29 sec
4	18.61 sec	19.00 sec	30.14 sec
5	17.42 sec	18.92	24.62 sec
Average	**16.07 sec**	**19.97 sec**	**24.02 sec**

Αποτέλεσμα Query 5:

```
root
 |-- AREA NAME: string (nullable = true)
 |-- Avg Distance: double (nullable = true)
 |-- #: long (nullable = false)

+-----------+------------------+------+
|  AREA NAME|      Avg Distance|     #|
+-----------+------------------+------+
|77th Street|2.7138402396920918|145774|
|  Southwest|2.7003222580129886|135814|
|    Pacific| 3.869783858004145|112726|
|  Southeast| 2.128449442304811|111789|
|    Mission| 4.716329029051904|103978|
|  Northeast| 3.849091085345116|100526|
|     Newton| 2.062374065896697|100304|
|   Van Nuys|  2.38293185168801|100040|
|  Hollywood| 1.549327607147048| 99277|
|    Central|1.1011963711733572| 98493|
|    Topanga| 3.791807391170099| 97953|
| Devonshire|3.8517457042194514| 96842|
|    Olympic|1.8513797092283986| 95695|
|     Harbor|  4.07414222316739| 92233|
|West Valley|3.5849667125065503| 89960|
|    Rampart| 1.652221431842629| 89880|
|   Wilshire|  2.58772993924733| 88866|
|   Foothill| 4.610083901601238| 80245|
| Hollenbeck| 2.751581540675943| 78158|
+-----------+------------------+------+
```