Προχωρημένα Θέματα Βάσεων Δεδομένων

Ονοματεπώνυμο: Μπάρμπα Παναγιώτα-Νικολέττα
ΑΜ : 03118604
Εξάμηνο : 11ο
Ομάδα : 41
Github : Github Link

Ζητούμενο 1

Η εγκατάσταση και διαμόρφωση της πλατφόρμας εκτέλεσης Απαχε Σπαρκ ώστε να εκτελείται πάνω από το διαχειριστή πόρων του Απαχε Ηαδοοπ, ΨΑΡΝ, έγινε σε 2 εικονικά μηχανήματα σε τοπικό μηχάνημα (δεν χρησιμοποιήθηκε το cloud service okeanos). Η διαμόρφωση των εργαλείων που χρησιμοποιήθηκαν περιγράφεται στο README αρχείο του Github αποθετηρίου.

Οι web διεπαφές των Apache Spark και Apache Hadoop είναι προσβάσιμες από τους παρακάτω συνδέσμους:

- Apache Spark : http://192.168.64.9:8080/
- Apache Hadoop : http://192.168.64.9:9870/
- Apache Hadoop YARN : http://192.168.64.9:8088/

Ζητούμενο 2

Αυτό όπως και τα επόμενα ζητούμενα υλοποιήθηκαν με χρήση του PySpark και της γλώσσας προγραμματισμού Python3.

Δημιουργήθηκε ένα DataFrame από το βασικό σύνολο δεδομένων και διατηρώντας τα ονόματα των στηλών, προσαρμόστηκαν οι τύποι ορισμένων στηλών ως εξής:

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

Επίσης, στο αρχειο IncomeData2015.csv η στήλη "Estimated Median Income" έχει τύπο string της μορφής: '$number¨, οπότε αφαιρέθηκε το ′$′ και έγινε μετατροπή σε integer.

Τέλος, ενώθηκαν τα DataFrame που περιέχουν τα δεδομένα καταγραφής εγκλημάτων για το Los Angeles από το 2010 μέχρι το 2019 και από το 2020 μέχρι σήμερα, τα δεδομένα με reverse geocoding πληροφορία και τα δεδομένα σχετικά με το μέσο εισόδημα ανά νοικοκυριό και ταχυδρομικό κώδικα δημιουργώντας ένα νέο DataFrame, το οποίο αποθηκεύτηκε, με την εξής μορφή:

```
Total Rows: 3001575
root
 |-- DR_NO: integer (nullable = true)
 |-- Date Rptd: date (nullable = true)
 |-- DATE OCC: date (nullable = true)
 |-- TIME OCC: integer (nullable = true)
 |-- AREA : integer (nullable = true)
 |-- AREA NAME: string (nullable = true)
 |-- Rpt Dist No: integer (nullable = true)
 |-- Part 1-2: integer (nullable = true)
 |-- Crm Cd: integer (nullable = true)
 |-- Crm Cd Desc: string (nullable = true)
 |-- Mocodes: string (nullable = true)
 |-- Vict Age: integer (nullable = true)
 |-- Vict Sex: string (nullable = true)
 |-- Vict Descent: string (nullable = true)
 |-- Premis Cd: integer (nullable = true)
 |-- Premis Desc: string (nullable = true)
 |-- Weapon Used Cd: integer (nullable = true)
 |-- Weapon Desc: string (nullable = true)
 |-- Status: string (nullable = true)
 |-- Status Desc: string (nullable = true)
 |-- Crm Cd 1: integer (nullable = true)
 |-- Crm Cd 2: integer (nullable = true)
 |-- Crm Cd 3: integer (nullable = true)
 |-- Crm Cd 4: integer (nullable = true)
 |-- LOCATION: string (nullable = true)
 |-- Cross Street: string (nullable = true)
 |-- LAT: double (nullable = true)
 |-- LON: double (nullable = true)
 |-- AREA: integer (nullable = true)
 |-- ZIPcode: string (nullable = true)
 |-- Community: string (nullable = true)
 |-- Estimated Median Income: integer (nullable = true)
```

Ζητούμενο 3

Ζητούμενο 4