Συστήματα Λογισμικού για διαχείριση και ανάλυση μεγάλου όγκου δεδομένων

Εργαστηριακή Άσκηση

Για τη συγκεκριμένη εργαστηριακή άσκηση θα χρησιμοποιηθεί ένα σύνολο δεδομένων με στοιχεία όπως η ποσότητα και το είδος των απορριμμάτων που συλλέγονται προς ανακύκλωση στη πόλη του Μπάφαλο.

Τα συγκεκριμένα δεδομένα παρέχονται από την ιστοσελίδα https://data.buffalony.gov/Quality-of-Life/Monthly-Recycling-and-Waste-Collection-Statistics/2cjd-uvx7

Η Κάθε γραμμή του αρχείου αντιστοιχεί σε διαφορετικό ανακυκλώσιμο είδος, αφορά ένα μήνα καταγραφής για τα έτη 2011-2022 και έχει την παρακάτω μορφή:

Date, Month, Type, Total (in tons)

Αρχικά σας ζητείται να υλοποιήσετε πρόγραμμα στο περιβάλλον του Apache Spark, που θα ενσωματώνει τα παραπάνω δεδομένα και στη συνέχεια να απαντήσετε στα ακόλουθα συνδυαστικά ερωτήματα:

1. Για πόσες και ποιες χρονιές ήταν η συνολική ποσότητα ενός είδους της επιλογής σας, υψηλότερη από 5 άλλα είδη της επιλογής σας
2. Ποια είναι τα 5 είδη με τις μεγαλύτερες ποσότητες ανακύκλωσης ανά έτος
3. Για ποια χρονιά το κάθε είδος είχε τη μικρότερη ποσότητα ανακύκλωσης και ποια τη μεγαλύτερη
4. Ποιοι είναι οι 3 μήνες με τη μεγαλύτερη ποσότητα ανακυκλώσιμων ειδών, ανεξαρτήτως έτους και είδους
5. Παρουσίαση συνολικών ποσοτήτων ανακυκλώσιμων ειδών ανά έτος

Παραδοτέα

1. **Γραπτή Αναφορά** (σε αρχείο pdf ή word) που θα περιλαμβάνει:

    - **Αναλυτική περιγραφή της διαδικασίας που ακολουθήσατε (και για την εγκατάσταση – ενσωμάτωση των δεδομένων στη πλατφόρμα Databricks (ή VM) στο Apache Spark**
    - **Τον κώδικα εμπλουτισμένο με αναλυτικό σχολιασμό**
    - **Screenshots παραδειγμάτων της εφαρμογής καθώς και της εγκατάστασης στο Databricks ή se vm (σε κάθε βήμα να υπάρχει αναλυτική περιγραφή)**
    - **Σχόλια - Παραδοχές που τυχόν έγιναν κατά την ανάπτυξη της εργασίας**
    - **Αρχείο powerpoint παρουσίασης της εργασίας (για 10 λεπτά παρουσίαση)**

2. Συμπιεσμένα σε ένα αρχείο zip:

    - **Την πιο πάνω γραπτή αναφορά**
    - **Τον ΤΕΛΙΚΟ κώδικα**
    - **Το αρχείο powerpoint**

Το αρχείο zip πρέπει να έχει όνομα τον **αριθμό μητρώου** του φοιτητή (π.χ. 3972.zip), και να ανεβεί (**ΥΠΟΧΡΕΩΤΙΚΑ**) **στο e-class.**
Σε ξεχωριστό αρχείο .txt μέσα στο zip να αναφέρεται το **ονοματεπώνυμο, ο αριθμός μητρώου και η e-mail διεύθυνση του φοιτητή.**

Οδηγίες

Ο κώδικας που απαντά σε κάθε ερώτημα **θα πρέπει να περιέχει αναλυτικό σχολιασμό**.

Επιπλέον, στα παραδοτέα της άσκησης πρέπει να περιλαμβάνεται αναφορά της διαδικασίας σε μορφή word ή Pdf, στην οποία θα αποσαφηνίζονται τα στάδια της εγκατάστασης, τα βασικά σημεία του κώδικά σας, καθώς και screenshots από τα αποτελέσματα, ξεχωριστά για κάθε ερώτημα.

Σημείωση: Για την επίλυση της άσκησης είναι απαραίτητη η χρήση **Spark SQL και Dataframes ή SQL μόνο** σε περιβάλλον Apache Spark, μέσω της πλατφόρμας Databricks

(www.databricks.com) ή αν επιθυμείτε μπορείτε να εγκαταστήσετε το Apache Spark σε ένα vm.

Διευκρινήσεις

1. Η άσκηση είναι **ατομική**
2. Οριστική ημερομηνία παράδοσης είναι η εξεταστική περιόδους Ιουνίου 2022 ΜΟΝΟ! Αναλόγως θα καθοριστεί και η ημερομηνία της παρουσίασης της εργασίας ή των εργασιών που θα επιλεγούν.
3. Για τυχόν απορίες ή υποδείξεις μπορείτε να απευθύνεστε με e-mail στο mvonitsanos@ceid.upatras.gr ή στις Συζητήσεις στο e-class του μαθήματος .

Καλή σας επιτυχία!