

## Αποκεντρωμένα Συστήματα Διαχείρισης Μεγάλου Όγκου Δεδομένων

Τμήμα Μηχανικών Η/Υ & Πληροφορικής - Πανεπιστήμιο Πατρών

ΔΠΜΣ ΥΔΑ

Διδάσκοντας: Σπυρίδων Σιούτας

### 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. Ποιος ήταν ο μέσος όρος στις ποσότητες των απορριμμάτων για κάθε είδος
2. Για πόσες και ποιες χρονιές ήταν η συνολική ποσότητα του είδους Scrap Metal (ή ενός είδους της επιλογής σας) υψηλότερος από 5 άλλα είδη (της επιλογής σας)
3. Ποια είναι τα 5 είδη με τις μεγαλύτερες ποσότητες ανακύκλωσης ανά έτος
4. Για ποια χρονιά το κάθε είδος είχε τη μικρότερη ποσότητα ανακύκλωσης και ποια τη μεγαλύτερη
5. Ποιοι είναι οι 5 μήνες με τη μεγαλύτερη ποσότητα ανακυκλώσιμων ειδών, ανεξαρτήτως έτους και είδους
6. Παρουσίαση συνολικών ποσοτήτων ανακυκλώσιμων ειδών ανά έτος
7. Συνολική παρουσίαση ανακυκλώσιμων ειδών και των αντίστοιχων ποσοτήτων τους

Μπορείτε να επιλέξετε ως γλώσσα υλοποίησης είτε τη Scala είτε την Python, ωστόσο πρέπει να **χρησιμοποιηθεί η ίδια γλώσσα προγραμματισμού** για όλα τα ερωτήματα. Ο κώδικας που απαντά σε κάθε ερώτημα **θα πρέπει να περιέχει αναλυτικό σχολιασμό** και να βρίσκεται σε ένα μόνο αρχείο.

Η ονομασία του να ακολουθεί την σύμβαση:

`Query[αριθμός_ερωτήματος].[επέκταση_γλώσσας]` (π.χ. Query1.scala ή Query2.py).

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** σε περιβάλλον Apache Spark, μέσω της πλατφόρμας Databricks ([www.databricks.com](http://www.databricks.com)) ή αν επιθυμείτε μπορείτε να εγκαταστήσετε το Apache Spark σε ένα vm.

## 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 σε γλώσσα python ή scala εμπλουτισμένο με αναλυτικό σχολιασμό**
  - **Screenshots παραδειγμάτων της εφαρμογής καθώς και της εγκατάστασης στο Databricks ή σε 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 σε python ή scala.**
  - **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.**

## Διευκρινήσεις

1. Η άσκηση είναι **ατομική**
2. Οριστική ημερομηνία παράδοσης είναι το τέλος της εξεταστικής εξέτασης περιόδου Ιουνίου 2022 ΜΟΝΟ! Αναλόγως θα καθοριστεί και η ημερομηνία της παρουσίασης της εργασίας.
3. Για τυχόν απορίες ή υποδείξεις μπορείτε να απευθύνεστε με e-mail στο [mvonitsanos@ceid.upatras.gr](mailto:mvonitsanos@ceid.upatras.gr) ή στις Συζητήσεις στο e-class του μαθήματος <https://eclass.upatras.gr/courses/CEID1175/>

## Παράρτημα

### Apache Spark Documentation

Ο σημαντικότερος βοηθός σας κατά την εκπόνηση της εργασίας δεν είναι άλλος από τη τεκμηρίωση που θα βρείτε στην ιστοσελίδα του Apache Spark. Ένα καλό σημείο για να ξεκινήσετε την ενασχόληση σας είναι ο παρακάτω οδηγός (δείτε την ενότητα “Self-Contained Applications” για τη δημιουργία αυτόνομων προγραμμάτων):

<https://spark.apache.org/docs/latest/quick-start.html>

Τις κλάσεις και τις συναρτήσεις που θα χρειαστείτε μπορείτε να τις αναζητήσετε στα αντίστοιχα API docs:

<https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.package>

<https://spark.apache.org/docs/latest/api/python/index.html>

Για την υλοποίηση της εργασίας ιδιαίτερα χρήσιμες θα σας φανούν οι συναρτήσεις που προσφέρει η Spark SQL και μπορείτε να βρείτε στα παρακάτω links:

<https://spark.apache.org/docs/latest/api/python/pyspark.sql.html#module-pyspark.sql.functions>

Περισσότερες πληροφορίες καθώς και παραδείγματα κώδικα σχετικά με την Spark SQL και τα DataFrames υπάρχουν εδώ:

<https://spark.apache.org/docs/latest/sql-programming-guide.html>