Τμήμα Μηχανικών Η/Υ & Πληροφορικής
ΔΠΜΣ ΥΔΑ

Αποκεντρωμένα Συστήματα Διαχείρισης Μεγάλου Όγκου Δεδομένων

Άσκηση Ακαδημαϊκού Έτους 2020-2021

Α) Το αρχείο tour_occ_ninat.xls περιέχει δεδομένα της Eurostat σχετικά με τις διανυκτερεύσεις τουριστών σε 36 ευρωπαικές χώρες για το διάστημα 2006-2015.

Αρχικά σας ζητείται να υλοποιήσετε πρόγράμμα στο περιβάλλον του Apache Spark, που θα ενσωματώνει τα δεδομένα, και να απαντήσετε στη συνέχεια, στα 4 ακόλουθα συνδυαστικά ερωτήματα:

1. Ποιος ήταν ο μέσος όρος διανυκτερεύσεων για κάθε χώρα για το χρονικό διάστημα 2007-2014
2. Για πόσες και ποιες χρονιές ήταν ο αριθμός διανυκτερεύσεων της χώρας Ελλάδα υψηλότερος από 5 άλλες ευρωπαικές χώρες (της επιλογής σας)
3. Ποιες ήταν οι χώρες με το μεγαλύτερο αριθμό διανυκτερεύσεων ανά έτος
4. Ποια ήταν η χρονιά που η κάθε χώρα είχε το μικρότερο αριθμό διανυκτερεύσεων σε σχέση με όλες τις υπόλοιπες ευρωπαϊκές χώρες.

Β) Το αρχείο International_tourist_arrivals περιέχει δεδομένα του UNWTO σχετικά με τις αφίξεις τουριστών σε όλο το κόσμο για τα έτη 1990, 1995,2000, 2005, 2014, 2015.

Αρχικά σας ζητείται να υλοποιήσετε πρόγράμμα στο περιβάλλον του Apache Spark, που θα ενσωματώνει τα δεδομένα, και να απαντήσετε στη συνέχεια, στα 2 ακόλουθα συνδυαστικά ερωτήματα:

1. Υπολογίστε τις τιμές που λείπουν σχετικά με τα αθροίσματα για τις επιλογές: Europe, Asia and the Pacific, Americas, Africa
2. Υπολογίστε τη ποσοστιαία μεταβολή για κάθε γεωγραφική περιοχή μεταξύ των διαφορετικών χρονικών σημείων

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 σχόλια 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 μια σύντομη αναφορά σε μορφή word, στην οποία θα αποσαφηνίζονται 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, που θα εγκαταστήσετε σε ένα vm (βάσει των οδηγιών του παραρτήματος), είτε σε περιβάλλον Databricks (www.databricks.com)

Η εργασία μπορεί να είναι ατομική ή ομαδική. Για τυχόν απορίες ή υποδείξεις μπορείτε να απευθύνεστε με e-mail στο mvonitsanos@ceid.upatras.gr ή στις Συζητήσεις στο e-class του μαθήματος https://eclass.upatras.gr/courses/CEID1175/

Παράρτημα

Εγκατάσταση Apache Spark

Για την εγκατάσταση του Apache Spark απαιτείται να έχετε εγκατεστημένη στον υπολογιστή σας κάποια διανομή Linux. Στον παρακάτω σύνδεσμο θα βρείτε λεπτομερείς οδηγίες για την εγκατάσταση Ubuntu σε Virtual Machine με χρήση VirtualBox:

https://www.wikihow.com/Install-Ubuntu-on-VirtualBox

Στη συνέχεια, θα πρέπει να εγκαταστήσετε στο σύστημά σας το JDK. Σε περιβάλλον Ubuntu αυτό γίνεται με τις ακόλουθες εντολές:

```
sudo apt-get update
sudo apt-get install default-jdk
```

Το επόμενο βήμα είναι η εγκατάσταση της γλώσσας Scala και του Sbt (Scala Built Tool). Αναλυτικά τα βήματα μπορείτε να τα βρείτε στον ακόλουθο σύνδεσμο:

https://www.scala-sbt.org/1.x/docs/Installing-sbt-on-Linux.html

Τέλος, θα πρέπει για να ολοκληρώσετε την εγκατάσταση του Apache Spark, ακολουθήστε τα παρακάτω βήματα:

1. Κατεβάστε το από εδώ (επιλογή “Pre-built for Apache Hadoop 3.2 and later”): https://spark.apache.org/downloads.html
2. Αποσυμπιέστε το αρχείο και μεταφέρετε τα περιεχόμενά του στον επιθυμητό φάκελο (προτείνεται ο /usr/local/spark). Σημείωση: Για την μεταφορά των αρχείων σε κάποιο φάκελο που για λόγους ασφάλειας δεν έχετε δικαιώματα εγγραφής, μπορείτε να τρέξετε την εντολή: `sudo mv move_from move_to`
3. (Προαιρετικό) Προσθέστε το φάκελο της εγκατάστασης στο PATH σας ώστε να μπορείτε να τρέξετε τα εκτελέσιμα αρχεία του Spark από οποιοδήποτε σημείο:
    a. Τρέξτε την εντολή: `gedit ~/.profile`
    b. Προσθέστε στο τέλος του αρχείου τη γραμμή:
    `export PATH=$PATH:path_to_spark_installation/bin`
    c. Αποσυνδεθείτε από το λογαριασμό σας και συνδεθείτε εκ νέου

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