# **Τμήμα Μηχανικών Η/Υ & Πληροφορικής ΔΠΜΣ ΥΔΑ**

# **Αποκεντρωμένα Συστήματα Διαχείρισης Μεγάλου Όγκου Δεδομένων**

# **Άσκηση Ακαδημαϊκού Έτους 2018-2019**

Στο αρχείο tempm.txt περιέχονται μετρήσεις για τη θερμοκρασία (σε °C) στην πόλη του Aarhus την περίοδο από 13/2/2014 έως και 8/6/2014. Όμοια, στο αρχείο hum.txt υπάρχουν οι καταγραφές των τιμών της υγρασίας (σε ποσοστό %) για την ίδια χρονική περίοδο. Κάθε γραμμή των αρχείων αυτών αντιστοιχεί και σε μία ημέρα μετρήσεων ενώ οι γραμμές έχουν την παρακάτω μορφή:

{“timestamp1”: “value1”, “timestamp2”: “value2”, … , “timestampN”: “valueN”}

Σας ζητείται να υλοποιήσετε τέσσερα διαφορετικά προγράμματα για το περιβάλλον του Apache Spark που να χρησιμοποιούν ως είσοδο τα παραπάνω αρχεία και το καθένα να απαντά σε ένα από τα ακόλουθα ερωτήματα:

1. Σε πόσες μέρες η θερμοκρασία κυμάνθηκε από 18 °C έως 22 °C;
2. Ποιες ήταν οι 10 πιο κρύες μέρες και οι 10 πιο ζεστές μέρες;
3. Ποιος μήνας είχε την υψηλότερη τυπική απόκλιση στις τιμές της υγρασίας;
4. Ποια ήταν η μέγιστη και ποια η ελάχιστη τιμή του δείκτη δυσφορίας (Σημείωση 1);

Μπορείτε να επιλέξετε ως γλώσσα υλοποίησης είτε τη Scala είτε την Python, ωστόσο πρέπει να χρησιμοποιηθεί η ίδια γλώσσα προγραμματισμού σε όλη την άσκηση. Ο κώδικας που απαντά σε κάθε ερώτημα θα πρέπει να βρίσκεται σε ένα μόνο αρχείο και η ονομασία του να ακολουθεί την σύμβαση: *Query[αριθμός\_ερωτήματος].[επέκταση\_γλώσσας]*(π.χ. Query1.scala ή Query2.py). Τέλος, στα παραδοτέα της άσκησης πρέπει να περιλαμβάνεται μια σύντομη αναφορά σε μορφή pdf στην οποία θα αποσαφηνίζονται τα βασικά σημεία του κώδικά σας.

Σημείωση 1: Για τον υπολογισμό του δείκτη δυσφορίας χρησιμοποιείστε το τύπο:

DI = T - 0,55 (1 - 0,01\*RH) (T - 14,5)

T: θερμοκρασία, RH: υγρασία

Σημείωση 2: Για την επίλυση της άσκησης συνίσταται η χρήση Spark SQL και Dataframes.

# **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 2.7 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:
   1. Τρέξτε την εντολή: gedit ~/.profile
   2. Προσθέστε στο τέλος του αρχείου τη γραμμή:   
      export PATH=$PATH:*path\_to\_spark\_installation*/bin
   3. 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 στον ιστότοπο 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 εγασίας 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/scala/index.html#org.apache.spark.sql.functions>

<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>