ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Τμήμα Εφαρμοσμένης Πληροφορικής
Ανάλυση Δεδομένων Μεγάλου Όγκου
Ακαδημαϊκό έτος 2022-2023

Δεύτερη ομαδική εργασία
Ημερομηνία παράδοσης: Παρασκευή, 03 Φεβρουαρίου 2023

Σας δίνεται ένα σύνολο δεδομένων, το οποίο προέρχεται από το Kaggle και αφορά συναλλαγές καταναλωτών. Κάθε γραμμή αφορά μία συναλλαγή και περιέχει κωδικούς προϊόντων που αγοράστηκαν. Το σύνολο δεδομένων μπορείτε να το κατεβάσετε ώς εξής:

wget snf-803830.vm.okeanos.grnet.gr/data.zip

1. Χρησιμοποιώντας τον υλοποίηση του αλγορίθμου FP-Growth της βιβλιοθήκης MLLib, που αποτελεί μέρος του Apache Spark, κατασκευάστε πρόγραμμα το οποίο θα υπολογίζει συχνά στοιχειοσύνολα και θα λαμβάνει ως παράμετρο κατά την εκτέλεση παράμετρο N, όπου N η ελάχιστη συχνότητα εμφάνισης του κάθε προϊόντος και θα παράγει ως έξοδο σχετικό αρχείο.
2. Βάζοντας ως κατώφλια τις παρακάτω τιμές συναλλαγών: 5000, 10000, 50000, Πραγματοποιήστε 3 εκτελέσεις και καταγράψτε,:
    a. τον ελάχιστο, μέγιστο και μέσο χρόνο που απαιτήθηκε για τον υπολογισμό.
    b. Τα παραπάνω θα τα πραγματοποιήσετε:
        i. Για έναν worker
        ii. Για δύο workers
3. Σχολιάστε τα αποτελέσματα των παραπάνω μετρήσεων.

Τον χρόνο κάθε εκτέλεσης μπορείτε να τον δείτε στο πρόγραμμα περιήγησής σας στη διεύθυνση <ip-του-master>:8080

Οδηγίες

1. **Η εργασία είναι ομαδική αποκλειστικά σε ομάδες των 2 ατόμων.**
2. Θα παραδώσετε:
    2.1. Ένα αρχείο κώδικα και την εντολή που χρησιμοποιείτε για να εκτελεστεί (δείτε και στο παράδειγμα του μαθήματος).
    2.2. Μία αναφορά έως 3 σελίδες σχετικά με τα βήματα που ακολουθήσατε, τα προβλήματα που αντιμετωπίσετε και τους τρόπους επίλυσής τους.
    2.3. Τις εντολές και παραμέτρους που χρησιμοποιήσατε σε ένα .txt αρχείο.
    2.4. Ένα αρχείο με τα συχνά στοιχειοσύνολα που προέκυψαν.
    2.5. Την έξοδο της εκτέλεσης του Spark.
3. Η εργασία θα πρέπει να παραδοθεί **ηλεκτρονικά** μέσω openeclass σε μορφή **ενός αρχείου** zip.

Ενδεικτικές Αναφορές:

1. Κεφάλαιο 6 από το βιβλίο «Εξόρυξη από Μεγάλα Σύνολα Δεδομένων», διαθέσιμο και online στο http://www.mmds.org/ (αγγλικά)
2. Κεφάλαιο 5.6 από το βιβλίο «Εισαγωγή στην Εξόρυξη Δεδομένων» των Ning Tan, Steinbach, Kumar.
3. «Data-Intensive Text Processing with MapReduce». Διαθέσιμο online στο https://lintool.github.io/MapReduceAlgorithms/
4. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.fpm.FPGrowth.html#pyspark.ml.fpm.FPGrowth
5. https://spark.apache.org/docs/latest/ml-frequent-pattern-mining.html