UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

Προχωρημένη Κατανεμημένη Υπολογιστική
Χειμερινό Εξάμηνο 2019-2020
Δημήτριος Κατσαρός

Bonus Coding project: Ατομικό ή ομάδα των 2

Ημέρα ανακοίνωσης: Monday, December 02, 2019
Προθεσμία παράδοσης: Κυριακή, Φεβρουάριος 02, 2020

Περιγραφή προβλήματος

Η υλοποίηση τελεστών ομοιότητας π.χ., set similarity σε περιβάλλοντα Hadoop είναι πολύ σημαντικά στην επεξεργασία μεγάλου όγκου δεδομένων.

Σκοπός της συγκεκριμένης εργασίας είναι η υλοποίηση του αλγορίθμου FullFilteringJoin για set similarity join σε περιβάλλον Hadoop.

Παρόλο που υπάρχει πλούσια πλέον βιβλιογραφία πάνω στο αντικείμενο, θα αρκεστούμε στην υλοποίηση μιας απλής εκδοχής του set similarity join, όπως αυτή περιγράφεται στο άρθρο: https://dl.acm.org/citation.cfm?id=3242932.

Τα δεδομένα όπου θα εφαρμοστεί ο αλγόριθμος είναι πραγματικά.

- Ανακτήστε κάποια από τα δεδομένα που περιγράφονται στο προαναφερθέν άρθρο.
- Μελετήστε το πρόβλημα του set similarity join, όπως περιγράφεται στο προαναφερθέν άρθρο.
- Ζητείται να υλοποιήσετε τον προαναφερθέντα αλγόριθμο.
- Να γράψετε μια αναφορά η οποία θα περιέχει: α) τον ψευδοκώδικα (όσων) ζευγών map-reduce, β) τις λεπτομέρειες υλοποίησης του αλγορίθμου, π.χ., εξειδικευμένες δομές δεδομένων, επιμέρους αλγορίθμους εντός των φάσεων map-reduce κ.τ.λ., γ) την πειραματική αποτίμηση της επίδοσης του αλγορίθμου που αναπτύξατε. Συγκεκριμένα:
  - Πειραματιστείτε με το μέγεθος του dataset (δημιουργήστε υπο-datasets αυξανόμενου μεγέθους, π.χ., μικρότεροι πίνακες).
  - Πειραματιστείτε με το similarity threshold.
  - Πειραματιστείτε με το configuration των παραμέτρων εκτέλεσης του Hadoop.
  - Καταγράψτε
    - τον χρόνο εκτέλεσης
      - της map φάσης
      - της shuffling φάσης
      - της reduce φάσης
      - τον συνολικό χρόνο
    - το memory footprint
      - της map φάσης
      - της shuffling φάσης
      - της reduce φάσης
      - συνολικά

Χρηστικές πληροφορίες:
Η προθεσμία παράδοσης είναι αυστηρή. Είναι όμως δυνατή η παροχή παράτασης (μέχρι 7 ημέρες), αλλά μόνο αφού δώσει ο διδάσκων την έγκρισή του, και αυτή η παράταση στοιχίζει 10% ποινή στον τελικό βαθμό της. Η παράδοση γίνεται με email στο dkatsar@e-ce.uth.gr του πηγαίου κώδικα, καθώς της αναφοράς που περιέχει την (σύντομη) περιγραφή του κώδικα, και των αποτελεσμάτων της πειραματικής αξιολόγησης. Το subject του μηνύματος πρέπει να είναι: CE623-Project: AEMx-AEMy

Ερμηνεία συμβόλου:

Απαιτεί την ανάπτυξη κώδικα (σε όποια γλώσσα υποστηριζόμενη από το Hadoop επιθυμείτε, π.χ., Python, Java). Εάν χρησιμοποιήσετε έτοιμο κώδικα από κάποια πηγή απαιτείται να δηλώσετε την πηγή, καθώς και σε ποιο σημείο του project τον χρησιμοποιήσατε.