ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργ. Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΕΦΑΡΜΟΓΗ ΤΟΥ MAP-REDUCE FRAMEWORK ΓΙΑ SEQUENCE MATCHING ΣΕ ΑΛΛΗΛΟΥΧΙΕΣ ΒΙΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Τάσος Αρβανίτης, 210 772 1436, anarv@dblab.ece.ntua.gr
Θανάσης Βεργούλης, 210 772 1436, bergoulis@dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Στόχος της διπλωματικής εργασίας είναι η υλοποίηση μιας εφαρμογής για τον εντοπισμό υποψήφιων αλληλουχιών (sequence matching) μέσα σε βιολογικά δεδομένα κάνοντας χρήση του MapReduce framework.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Το MapReduce είναι ένα framework λογισμικού που χρησιμοποιείται για την ανάπτυξη εφαρμογών που τρέχουν σε παράλληλα συστήματα βασισμένα σε μεγάλα clusters. Τα συστήματα αυτά μπορούν να φτάνουν σε μέγεθος ακόμα και τους χιλιάδες κόμβους. Το MapReduce μπορεί να χρησιμοποιηθεί για εφαρμογές που απαιτούν επεξεργασία τεράστιων όγκων δεδομένων, συνήθως σε κλίμακα πολλών terabytes. Οι ευκολίες που παρέχει για την ανάπτυξη τέτοιων εφαρμογών σε παράλληλα συστήματα είναι πολλές: αναλαμβάνει τον διαμοιρασμό των πόρων, τον προγραμματισμό των εργασιών (task scheduling), παρέχει ανοχή σε σφάλματα και δικτυακή επικοινωνία και συνεργασία μεταξύ των κόμβων του cluster, κ.α. Το MapReduce χρησιμοποιείται αυτή την στιγμή από πλήθος μεγάλων εταιριών πληροφορικής όπως οι Google, Yahoo, Amazon, Facebook, IBM για την αποδοτική διαχείριση των τεράστιων όγκων δεδομένων που διαθέτουν.

Στα πλαίσια της διπλωματικής εργασίας θα γίνει χρήση του Hadoop, μίας open-source υλοποίησης του MapReduce για την ανάπτυξη ενός αλγορίθμου για sequence matching σε βιολογικά δεδομένα. Το πρόβλημα του sequence matching έχει ως εξής: Έστω μια βάση δεδομένων από μεγάλες ακολουθίες συμβόλων S1, S2, …, Sn (π.χ. ακολουθίες βάσεων A (αδενίνη), G (γουανίνη), T (θυμίνη), C (κυτοσίνη) σε γονίδια). Έστω επίσης μια μικρή ακολουθία συμβόλων s. Το πρόβλημα του sequence matching αφορά στον εντοπισμό των περιοχών στις S1, S2, …, Sn όπου ταιριάζει η ακολουθία s. Το ταίριασμα μπορεί να είναι ακριβές (exact) ή προσεγγιστικό (approximate) αν επιτρέπονται και κάποια λάθη.

Πιο αναλυτικά η εκπόνηση της διπλωματικής εργασίας περιλαμβάνει τα ακόλουθα θέματα:

- Μελέτη διαφόρων αλγορίθμων για την επίλυση του προβλήματος του exact και approximate sequence matching.
- Σχεδιασμός εναλλακτικών λύσεων βασισμένες στο προγραμματιστικό μοντέλο του MapReduce για την επίλυση του συγκεκριμένου προβλήματος.
- Υλοποίηση κάποιων από τους εξεταζόμενους αλγορίθμους κάνοντας χρήση του API που παρέχεται από το Hadoop project.
- Πειραματική αξιολόγηση του συστήματος μεταβάλλοντας διάφορες παραμέτρους όπως:
    a) run-time παραμέτρους του συστήματος όπως ο αριθμός των κόμβων του cluster (τόσο των mappers όσο και των reducers), η μέση πιθανότητα αστοχίας κάποιου κόμβου, η μέγιστη VM που διατίθεται για την εκτέλεση κάθε task ανά κόμβο κ.α.

b) παραμέτρους των υπό εξέταση αλγορίθμων όπως ο μέγιστος επιτρεπόμενος αριθμός λανθασμένων συμβόλων για να θεωρηθεί μια ακολουθία ως matched, το μέγεθος των ακολουθιών βιολογικών δεδομένων κ.α.

ΣΧΕΤΙΚΕΣ ΑΝΑΦΟΡΕΣ:

1) Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Symposium on Operating System Design and Implementation (OSDI) 2004
2) Yang, H., Dasdan, A., Hsiao, R., Parker, D.: Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters. SIGMOD Conference 2007
3) http://hadoop.apache.org/