ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ (www.cslab.ece.ntua.gr)

Διπλωματικές Εργασίες
Ακ. Έτος 2012-2013

Μελέτη Απόδοσης Κατανεμημένων Βάσεων Δεδομένων Για Γράφους (Graph Databases) για Διαχείριση Διασυνδεδεμένων Δεδομένων (Linked Data)

Εισαγωγή

Ο όγκος των δεδομένων που παράγεται και δημοσιεύεται στο διαδίκτυο καθημερινά είναι τεράστιος. Η ψηφιακή εποχή μας χαρακτηρίζεται από μία έκρηξη δεδομένων ("data explosion")! Επομένως, ένα σημαντικό πρόβλημα είναι ο σχεδιασμός μηχανισμών και αρχιτεκτονικών, ώστε να επιτυγχάνεται η αποδοτική διαχείριση των δεδομένων. Η χρήση προηγμένων συστημάτων για την οργάνωση και αποθήκευση των δεδομένων παίζει σημαντικό ρόλο στην αποδοτική και γρήγορη αναζήτηση τους.

Ένα βασικό χαρακτηριστικό που παρατηρείται στα δεδομένα αυτά είναι ότι μπορούν να σχηματίσουν τεράστιους γράφους. Ο αριθμός των χρηστών των διαδικτυακών τόπων για e-mail και κοινωνική δικτύωση ανέρχεται σε εκατοντάδες εκατομμύρια και αυξάνεται συνεχώς με ραγδαίο ρυθμό. Τα δεδομένα που δημοσιεύονται σε διαφορετικούς διαδικτυακούς τόπους μπορούν να διασυνδέονται μεταξύ τους και είναι γνωστά ως Linked Data [1].

Ο μεγάλος όγκος της διαθέσιμης πληροφορίας καθιστά την επεξεργασία της όλο και πιο δύσκολη, όταν αυτή πραγματοποιείται σε μεμονωμένους υπολογιστικούς πόρους. Για αυτόν το λόγο, οι κατανεμημένες πλατφόρμες γίνονται όλο και πιο δημοφιλείς. Ένα ευρέως διαδομένο υπολογιστικό μοντέλο που προσφέρει τη δυνατότητα για κατανεμημένη υπολογιστική επεξεργασία αποτελεί το Cloud Computing [2].

Η επεξεργασία των αναφερόμενων γράφων δεδομένων σε ισχυρά υπολογιστικά κέντρα δεδομένων (data centers) μπορεί να παίξει σημαντικό ρόλο σε πολλούς τομείς, όπως σε μηχανές αναζήτησης που προσαρμόζονται στις προτιμήσεις των χρηστών (personalized information & recommendation systems), σε συστήματα αναλύσεων και λήψης αποφάσεων, κ.α. Για το λόγο αυτό, έχουν αναπτυχθεί διάφορα συστήματα για την αποθήκευση και

επεξεργασία γράφων σε cloud υποδομές, όπως το Google Pregel [3], Google BigQuery [3], Apache Giraph [5], Titan [6], κλπ.

Σκοπός

Ο στόχος της διπλωματικής εργασίας είναι η δημιουργία μίας κατανεμημένης πλατφόρμας για την επεξεργασία linked data σε cloud περιβάλλοντα. Κατά τη διάρκεια της συγκεκριμένης διπλωματικής θα μελετηθούν θέματα που αφορούν την απόδοση κατανεμημένων συστημάτων και βάσεων δεδομένων για την επεξεργασία δεδομένων με μορφή γράφων. Αναλυτικότερα, η υλοποίηση της συγκεκριμένης εργασίας θα κινηθεί στις ακόλουθες κατευθύνσεις:

- Μελέτη των κατανεμημένων συστημάτων για επεξεργασία δεδομένων γράφων που υπάρχουν στη βιβλιογραφία.
- Εγκατάσταση των αναφερόμενων συστημάτων και σύγκριση της απόδοσης τους ως προς το χρόνο εισαγωγής δεδομένων, επεξεργασίας διαφορετικών κατηγοριών ερωτημάτων και της επεκτασιμότητας τους όσο αυξάνεται ο όγκος των δεδομένων, ο αριθμός των χρηστών και ο αριθμός των διαθέσιμων πόρων.
- Ανάπτυξη API (κατά προτίμηση σε Java) για την εισαγωγή διασυνδεμένων δεδομένων τύπου RDF στα συστήματα που μελετήθηκαν.
- Ανάπτυξη API για τη εκτέλεση διάφορων κατηγοριών ερωτημάτων, όπως SPARQL ερωτήματα, ερωτήματα για την εύρεση υπογράφων, κλπ.
- Μετρήσεις στο τελικό σύστημα που θα προκύψει και εξαγωγή τελικών συμπερασμάτων.

Επικοινωνία

Νεκτάριος Κοζύρης, Αναπ. Καθηγητής nkoziris@cslab.ece.ntua.gr
Νάσια Ασίκη, Μεταδ. Ερευνήτρια nasia@cslab.ece.ntua.gr
Δημήτριος Τσουμάκος, Επικ. Καθηγητής dtsouma@cslab.ece.ntua.gr

Βιβλιογραφία

[1] Linked Data, http://linkeddata.org/
[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "*A view of cloud computing*"
[3] G. Malewicz, M. Austern, A. Bik, Aart, J. Dehnert, I. Horn, N. Leiser and G. Czajkowski, "Pregel: a system for large-scale graph processing"
[4] Google BigQuery, https://developers.google.com/bigquery/
[5] Apache Giraph, http://incubator.apache.org/giraph/
[6] Titan, https://github.com/thinkaurelius/titan/wiki