ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργ. Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΥΛΟΠΟΙΗΣΗ ΕΦΑΡΜΟΓΗΣ ΓΙΑ ΜΙΑ COLUMN-ORIENTED DATABASE

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

ΠΕΡΙΛΗΨΗ: Στόχος της διπλωματικής εργασίας είναι η υλοποίηση μιας εφαρμογής για μια column-oriented database κάνοντας χρήση της open-source βάσης δεδομένων HBase.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Οι column-oriented databases, σε αντίθεση με τις παραδοσιακές raw-oriented, αποθηκεύουν τα data σε στήλες (columns) αντί γραμμές (raws). Είναι αρκετά πιο αποδοτικές όταν χρησιμοποιούνται σε εφαρμογές που απαιτούν κυρίως reads και όχι τόσο writes στη βάση. Ως εκ τούτου, ενώ οι raw-oriented databases έχουν βέλτιστη απόδοση για OLTP-like workloads, οι column-oriented databases θα μπορούσαν να βρουν ευρεία χρήση για εφαρμογές data warehousing (OLAP κλπ), data analysis κ.α.. Τα τελευταία χρόνια έχουν παρουσιαστεί αρκετά column-oriented DBMS, όπως π.χ. τα BigTable της Google [1], C-Store [2] και HBase [3]. Συγκεκριμένα το BigTable χρησιμοποιείται από την Google για την επεξεργασία τεραστίων όγκων δεδομένων (της τάξης των πολλών terabytes) για διάφορες εφαρμογές όπως το web page indexing, το Google Maps και το Google Earth, το Google Analytics κ.α.. Το C-Store, επιπλέον, υποστηρίζει το σχεσιακό μοντέλο και παρέχει τη δυνατότητα ερωτήσεων γραμμένων σε γλώσσα SQL. Τέλος, το HBase είναι μια open-source, κατανεμημένη, column-oriented βάση δεδομένων, που σχεδιαστικά δανείζεται στοιχεία από το BigTable της Google. Είναι κομμάτι της αρχιτεκτονικής του συστήματος Hadoop της Apache, το οποίο παρέχει τη δυνατότητα παράλληλης επεξεργασίας τεράστιων όγκων δεδομένων πάνω από συστήματα cluster.

Στα πλαίσια της διπλωματικής εργασίας θα γίνει χρήση του HBase για την υλοποίηση μιας εφαρμογής που θα εξάγει στατιστικά στοιχεία για τα δεδομένα ταινιών που παρέχονται από το http://www.imdb.com/ .

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

- Μελέτη της αρχιτεκτονικής των column-oriented DBMS, με έμφαση στη βάση δεδομένων HBase.
- Φόρτωμα των δεδομένων του http://www.imdb.com/ στο σύστημα HBase (ενδεικτικά για κάποιους βασικούς πίνακες όπως movies, genres, directors, ratings).
- Εφαρμογή aggregate queries στα δεδομένα της βάσης. Πιθανά queries που θα μπορούσαν να υλοποιηθούν είναι π.χ.:
    - top-k movies group by genre/director/year
    - select director, avg(rating) from movies group by director
    - select director, count(*) from movies group by director
    - …
- Πειραματική αξιολόγηση του συστήματος μεταβάλλοντας διάφορες παραμέτρους όπως:

a) Το μέγεθος του dataset
b) Τον αριθμό των κόμβων hosts της κατανεμημένης βάσης δεδομένων.

- Συμπληρωματικά μπορεί να γίνει φόρτωμα των δεδομένων σε μια σχεσιακή raw-oriented βάση δεδομένων και να υλοποιηθούν τα ίδια aggregate queries ως stored procedures πάνω στα δεδομένα. Στη συνέχεια, μπορεί να γίνει σύγκριση της απόδοσης μεταξύ των δύο διαφορετικών συστημάτων για διάφορους τύπους aggregate queries, καθώς και για insertions, updates κλπ.

Σχετικές αναφορές:

1) Chang, F., Dean, J., Ghemawat, S., Hsieh, W., Wallach, D., Burrows, M., Chandra, T., Fikes, A., Gruber, R.: BigTable: A Distributed Storage System for Structured Data. Symposium on Operating System Design and Implementation (OSDI) 2006
2) Stonebraker, M., Abadi, D., Batkin, A., Chen, X., Cherniack, M., Ferreira, M., Lau, E., Lin, A., Madden, S., O' Neil, E., O' Neil, P., Rasin, A., Tran, N. Zdonik, S.: C-Store: A Column-oriented DBMS. VLDB Conference 2007
3) Abadi, D., Madden, S., Hachem, M.: Column-Stores vs. Row-Stores: How Different are they really? SIGMOD Conference 2008
4) Holloway, A., DeWitt, D.: Read-Optimized Databases in Depth. PVLDB vol 1. 2008
5) http://hadoop.apache.org/hbase/