ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μεταπτυχιακός Φοιτητής
Ζερβουδάκης Πέτρος

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Δ. Πλεξουσάκης

Τετάρτη, 04/03/2020, 10:00
Αίθουσα Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Αυξητική αποτίμηση συνεχών αναλυτικών επερωτήσεων βασιζόμενοι σε μια γλώσσα επερωτήσεων υψηλού επιπέδου "

Η διαδικασία ανάλυσης δεδομένων έχει λάβει σημαντική προσοχή τα τελευταία χρόνια καθώς τεράστιες ποσότητες δεδομένων παράγονται καθημερινά από διάφορες πηγές. Η ανάλυση αυτών των τεράστιων δεδομένων αποτελεί ένα ενδιαφέρον αλλά και δύσκολο έργο και απαιτεί νέες μορφές επεξεργασίας ώστε να είναι εφικτή η λήψη αποφάσεων, η ανακάλυψη γνώσεων και η βελτίωση των διαδικασιών. Επιπλέον, εκτός από τον συνεχώς αυξανόμενο όγκο τους, τα σύνολα δεδομένων αλλάζουν συνεχώς, και ως εκ τούτου, τα αποτελέσματα σε συνεχόμενα ερωτήματα πρέπει να ενημερώνονται σε σύντομα χρονικά διαστήματα. Σε αυτή την εργασία, αντιμετωπίζουμε το πρόβλημα της αποτίμησης συνεχών ερωτημάτων σε μεγάλες ροές δεδομένων που αλλάζουν συχνά. Προς αυτή την κατεύθυνση, υιοθετούμε την HIFUN, μια γλώσσα ερωτημάτων υψηλού επιπέδου, που προτείνεται για την έκφραση αναλυτικών ερωτημάτων σε μεγάλα σύνολα

δεδομένων. Η HIFUN προσφέρει ένα σαφή διαχωρισμό μεταξύ του εννοιολογικού επιπέδου, όπου τα αναλυτικά ερωτήματα ορίζονται ανεξάρτητα από τη φύση και τη θέση των δεδομένων, και το φυσικό επίπεδο όπου τα ερωτήματα αυτά αποτιμώνται, εκφράζοντας τα είτε ως MapReduce διαδικασίες είτε ως SQL ερωτήματα υποστηρίζοντας έτσι διαφορετικούς τύπους δεδομένων. Χρησιμοποιώντας τη HIFUN, σχεδιάζουμε έναν αλγόριθμο για την αυξητική αποτίμηση συνεχών ερωτημάτων, επεξεργάζοντας μόνο το πιο πρόσφατο διαμέρισμα δεδομένων και εκμεταλλευόμενοι τις ήδη υπολογισμένες πληροφορίες, χωρίς να απαιτείται η αποτίμηση του ερωτήματος πάνω από το πλήρες σύνολο δεδομένων. Στη συνέχεια, μεταφράζουμε τον γενικό αλγόριθμο σε SQL και MapReduce χρησιμοποιώντας το SPARK, εκμεταλλεύοντας τις μεθόδους επανεγγραφής ερωτημάτων που παρέχονται από τη HIFUN. Χρησιμοποιώντας ένα συνθετικό σύνολο δεδομένων, επιδεικνύουμε την αποτελεσματικότητα της προσέγγισης μας στην επίτευξη της απόδοσης αποτίμησης της επερώτησης. Τέλος, αποδεικνύουμε ότι υιοθετώντας τις επίσημες μεθόδους επανεγγραφής επερωτήσεων της HIFUN, επιτυγχάνουμε την περαιτέρω μείωση του υπολογιστικού κόστους, προσθέτοντας άλλο ένα επίπεδο βελτιστοποίησης των ερωτημάτων στην υλοποίηση μας.

Zervoudakis Petros
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Professor, D. Pleksousakis

Wednesday, 04/03/2020, 10:00
Room K206, Computer Science Dept., University of Crete

“Incremental Evaluation of Continuous Analytic Queries in a High-Level Query Language”

ABSTRACT

Data analytics have received a significant attention in recent years, as huge amounts of data is generated each day from various sources. Analysis of these massive data poses an

interesting but challenging task and requires new forms of processing to enable enhanced decision making, insight discovery and process optimization. In addition, besides their ever increasing volume, data sets change frequently, and as such, results to continuous queries have to be updated at short intervals. In this thesis, we address the problem of evaluating continuous queries over big data streams that are frequently updated. To this end, we adopt HIFUN, a high-level query language, proposed for expressing analytic queries over big data sets. HIFUN offers a clear separation between the conceptual layer, where analytic queries are defined independently of the nature and location of data, and the physical layer where queries are evaluated, by encoding them as map-reduce jobs or as SQL group-by queries, thus supporting different types of data set formats. Using HIFUN, we design an algorithm for incremental evaluation of continuous queries, processing only the most recent data batch, and exploiting already computed information, without requiring the evaluation of the query over the complete data set. Subsequently, we translate the generic algorithm to both SQL and MapReduce using SPARK, exploiting the query rewriting methods provided by HIFUN. Using a synthetic data set, we demonstrate the effectiveness of our approach in achieving query answering efficiency. Finally, we show that by exploiting the formal query rewriting methods of HIFUN, we can further reduce the computational cost, adding another layer of query optimization in our implementation.