ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Γλαμπεδάκης Βασίλειος
Μεταπτυχιακός Φοιτητής

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

Τετάρτη, 20/09/2017, 16:00
Αίθουσα B108 ,Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"Ενα σύστημα ανάλυσης μεγάλων δεδομένων βασισμένο σε μία γλώσσα επερωτήσεων υψηλού επιπέδου χρησιμοποιώντας το Apache Spark "

ΠΕΡΙΛΗΨΗ

Η ανάλυση μεγάλων δεδομένων είναι μία από τις πιο ενεργές ερευνητικές περιοχές σήμερα και συνοδεύεται από πολλές θεωρητικές και πρακτικές προκλήσεις. Αυτή η εργασία συμβάλλει σε αυτήν την περιοχή υλοποιώντας την γλώσσα HIFUN χρησιμοποιώντας το framework Apache Spark. Η HIFUN είναι μία γλώσσα επερωτήσεων υψηλού επιπέδου η οποία έχει προταθεί για την έκφραση επερωτήσεων ανάλυσης πάνω σε δεδομένα μεγάλου όγκου. Αυτή η γλώσσα κάνει έναν σαφή διαχωρισμό μεταξύ του εννοιολογικού και του φυσικού επιπέδου. Μία επερώτηση ανάλυσης και η απάντησή της ορίζονται στο εννοιολογικό επίπεδο ανεξάρτητα από την φύση και την τοποθεσία των δεδομένων. Κατόπιν, αυτοί οι αφηρημένοι ορισμοί αντιστοιχίζονται σε μηχανισμούς αποτίμησης χαμηλότερου επιπέδου, λαμβάνοντας υπόψιν την φύση και την τοποθεσία των δεδομένων, όπως επίσης και άλλους σχετιζόμενους παράγοντες. Σε αυτήν την εργασία, αξιοποιούμε αυτήν την γλώσσα για την σχεδίαση και την υλοποίηση ενός συστήματος το οποίο επιτρέπει σε έναν χρήστη να αναλύσει, να οπτικοποιήσει και να ανακαλύψει πληροφορία η οποία μπορεί να είναι χρήσιμη στη λήψη αποφάσεων και η οποία είναι κρυμμένη σε δεδομένα μεγάλου όγκου. Στο φυσικό επίπεδο, οι επερωτήσεις της HIFUN αντιστοιχίζονται σε μηχανισμούς αποτίμησης χαμηλότερου επιπέδου του Apache Spark, ακολουθώντας το προτεινόμενο από την HIFUN εννοιολογικό πλάνο αποτίμησης, υποστηρίζοντας μία μεγάλη γκάμα μορφών δεδομένων. Στο εννοιολογικό επίπεδο, εφαρμόζουμε τους κανόνες

επανεγγραφής επερωτήσεων και δημιουργούμε πλάνα εκτέλεσης επερωτήσεων, προτεινόμενα από την HIFUN. Η εργασία αυτή δείχνει ότι το τυπικό μοντέλο της HIFUN είναι χρήσιμο στην πράξη και η πειραματική αξιολόγησή του συστήματος αποδεικνύει ότι η προσέγγιση του μοντέλου στην επανεγγραφή την επερωτήσεων και στην παραγωγή πλάνων εκτέλεσης επερωτήσεων επιτυγχάνει τη μείωση του υπολογιστικού κόστους ανεξάρτητα από την φύση των δεδομένων..

Glabedakis Vasilis
M.Sc. Thesis

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

Wednesday, 20/09/2017, 16:00
Room B108, Computer Science Dept., University of Crete

"A Big Data Analytics System Based on a High Level Query Language Using Apache Spark "

ABSTRACT

Big data analytics is one of the most active research areas today with a lot of challenges, both theoretical and practical. This thesis makes a contribution to the area of big data analytics by implementing the HIFUN language using the Apache Spark framework. HIFUN is a high level query language, proposed for expressing analytic queries over big data sets. This language makes a clear separation between the conceptual and the physical level. An analytic query and its answer are defined at the conceptual level independently of the nature and location of data. The abstract definitions are then mapped to lower level evaluation mechanisms, taking into account the nature and location of data, as well as other related aspects. In this thesis, we leverage this language to design and implement a system which allows a user to analyse, visualize and discover information useful for decision making, which is "hidden" in large-scale data sets. In the physical level, HIFUN queries are mapped to lower level evaluation mechanisms of the Apache Spark framework following the conceptual evaluation scheme proposed by HIFUN and supporting a big range of data set formats. In the conceptual level, we apply the query rewriting rules and create query execution plans, proposed by HIFUN. Our work shows

that the HIFUN formal model is useful in practice and the experimental evaluation of the system proves that the model's approach to query rewriting and to the generation of query execution plans succeeds in reducing the computational costs regardless of the nature of the data.