ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2016-17 (ενημέρωση 13/9/2016)

1) Ανάλυση Ιατρικών Δεδομένων

Η παρούσα πτυχιακή σκοπεύει να επικεντρωθεί στην ανάλυση πραγματικών (ανώνυμων) ιατρικών δεδομένων. Πιο συγκεκριμένα, στοχεύει στην απάντηση των εξής δύο ερωτημάτων:
Α) «ποιοί ασθενείς είναι ο πιο όμοιοι με κάποια συγκεκριμένη περίπτωση»;
Β) «μπορούν να εξαχθούν με αυτοματοποιημένο τρόπο σχέσεις αιτιότητας αναλύοντας τα δεδομένα, π.χ., το τάδε χαρακτηριστικό οδηγεί στο τάδε σύμπτωμα»;
Για το ερώτημα (Α), θα πρέπει να αξιολογηθούν οι εναλλακτικές μετρικές ομοιότητας-απόστασης, ώστε η τεχνική εύρεσης κοντινότερων γειτόνων να είναι αποτελεσματική.
Για το (Β), χρησιμοποιώντας ως σημείο εκκίνησης τις τεχνικές δημιουργίας μοντέλων αιτιότητας στο [1] και στο εργαλείο Tetrad (http://www.phil.cmu.edu/tetrad/), ο σκοπός της εργασίας είναι να δημιουργηθούν γραφικά μοντέλα αιτιότητας.
Η εργασία απαιτεί καλή γνώση στατιστικής.
[1] *Cosma Rohilla Shalizi: Advanced Data Analysis from an Elementary Point of View (http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/)*

2) Ανάπτυξη εργαλείου εξατομικευμένων προσφορών προϊόντων σε online αγορές

Η εργασία αυτή στοχεύει στην ανάπτυξη μίας εφαρμογής για online αγορές προϊόντων σούπερ-μάρκετ, η οποία δυναμικά θα παρουσιάζει προσφορές στους πελάτες με εξατομικευμένο τρόπο. Στόχος είναι η υλοποίηση και αξιολόγηση ενός συνόλου διαφορετικών προσεγγίσεων που χρησιμοποιούνται σήμερα από μεγάλες εταιρείες όπως η Google και η Amazon (π.χ., εύρεση κανόνων συσχέτισης, διαφημίσεις στον Παγκόσμιο Ιστό, εξατομικευμένα προτασιακά συστήματα για μεγάλα δεδομένα [1]). Επιπλέον, να περιλαμβάνει προηγμένα τεχνικά χαρακτηριστικά, όπως η ανάλυση του χρόνου που ένας χρήστης βλέπει ένα συγκεκριμένο προϊόν πριν το αγοράσει. Η εργασία προϋποθέτει προγραμματιστικές δεξιότητες (κατά προτίμηση και σε περιβάλλον android) και καλή γνώση προηγμένων τεχνικών εξόρυξης δεδομένων.
[2] *Jure Leskovec, Anand Rajaraman, Jeff Ullmam: Mining of Massive Datasets (http://mmds.org/)*

3) Μαζικά παράλληλες τεχνικές εντοπισμού ανωμαλιών σε ροές δεδομένων (1 ή 2 άτομα)

Στα πλαίσια αυτής της εργασίας, ο φοιτητής καλείται να επεκτείνει ένα υπάρχον σύστημα μαζικά παράλληλης προηγμένης επεξεργασίας δεδομένων με τεχνικές εντοπισμού ανωμαλιών σε ροές δεδομένων. Οι τελευταίες είναι ήδη υλοποιημένες (στη μη παράλληλη εκδοχή τους) στο σύστημα ανοιχτού κώδικα MOA (http://moa.cms.waikato.ac.nz) [3] και ζητείται η παραλληλοποίησή τους. Οι εναλλακτικές για το σύστημα που θα επεκταθεί περιλαμβάνουν α) το υποσύστημα MLlib του Spark (http://spark.apache.org/), β) το SAMOA (http://samoa.incubator.apache.org/) και γ) το Flink (http://flink.apache.org/). Απαιτείται καλή γνώση προγραμματισμού και βασική γνώση εξόρυξης δεδομένων. Η εργασία προϋποθέτει ευχέρεια στη μελέτη και χρήση συστημάτων, στον προγραμματισμό σε υπάρχον σύστημα και στην εκπόνηση ερευνητικής εργασίας για την αποδοτική παραλληλοποίηση των τεχνικών.
[3] *Dimitrios Georgiadis, Maria Kontaki, Anastasios Gounaris, Apostolos N. Papadopoulos, Kostas Tsichlas, Yannis Manolopoulos: Continuous Outlier Detection in Data Streams: an Extensible Framework and State-of-the-Art Algorithms. SIGMOD 2013 - demo (http://delab.csd.auth.gr/papers/SIGMOD2013gkgptm.pdf)*

4) Αποδοτική εκτέλεση ερωτημάτων υποστήριξης αποφάσεων στο Spark (1 ή 2 άτομα)

Τα ερωτήματα υποστήριξης αποφάσεων είναι κατά κανόνα χρονοβόρα. Στα πλαίσια αυτής της εργασίας, θα γίνει μελέτη του κατά πόσο οι προτεινόμενες υλοποιήσεις των ερωτημάτων υποστήριξης αποφάσεων του TPC-DS benchmark στο Spark (https://github.com/databricks/spark-sql-perf) μπορούν να βελτιωθούν. Οι επιδιωκόμενες βελτιώσεις θα κινηθούν σε δύο βασικούς άξονες: α) στην αποδοτικότερη υλοποίηση συνδέσεων, π.χ., με χρήση τεχνικών στην εργασία [4] και β) στην εύρεση αποδοτικότερης σειράς εκτέλεσης των επιμέρους βημάτων βάσει των τεχνικών στην εργασία [5]. Η εργασία προϋποθέτει καλή κατανόηση

συμπεριφοράς συστημάτων και ερευνητικών εργασιών που περιγράφουν προηγμένες τεχνικές διαχείρισης δεδομένων. Επίσης απαιτεί καλές γνώσεις προγραμματισμού.

[4] *Foto N. Afrati, Jeffrey D. Ullman: Optimizing Multiway Joins in a Map-Reduce Environment. IEEE Trans. Knowl. Data Eng. 23(9): 1282-1298 (2011)*

[5] *Astrid Rheinländer, Arvid Heise, Fabian Hueske, Ulf Leser, Felix Naumann: SOFA: An extensible logical optimizer for UDF-heavy data flows. Inf. Syst. 52: 96-125 (2015)*

5) Ανάπτυξη και μελέτη συμπεριφοράς ελαστικής υποδομής NoSQL βάσων δεδομένων πάνω σε LXD

Τα LXD (https://linuxcontainers.org/lxd/) έχουν χαρακτηριστικά που επιτρέπουν την πολύ γρήγορη εγκατάσταση εικονικών μηχανημάτων πάνω σε ένα φυσικό μηχάνημα. Ο σκοπός αυτής της εργασίας είναι α) η δημιουργία μίας υποδομής νέφους για υποστήριξη πολύ μεγάλων NoSQL βάσεων δεδομένων με χρήση LXDs και β) η μελέτη της συμπεριοφοράς της βάσης δεδομένων κάτω από μεταβαλλόμενες συνθήκες αναφορικά με το πλήθος, τη διαθέσιμη μνήμη, τους πυρήνες και την φυσική τοποθεσία των εικονικών μηχανών. Ο τελικός πειραματισμός θα λάβει χώρα στην υποδομή ιδιωτικού νέφους του εργαστηρίου DELAB. Απαιτούνται ικανότητες διαχείρισης παράλληλων συστημάτων και εντατικός πειραματισμός.

6) Ανάλυση απόδοσης και βελτιστοποίηση επιχειρησιακών διαδικασιών

Οι επιχειρησιακές διαδικασίες (business processes) αυτοματοποιούνται με τη χρήση ειδικών μηχανών ροών εργασίας (workflow engines), οι οποίες όμως έχουν λιγότερα στοιχεία βελτιστοποίησης από τις αντίστοιχες μηχανές για ερωτήματα βάσεων δεδομένων. Η εργασία αυτή είναι διαθεματική. Στοχεύει αρχικά στην κατανόηση της μελέτης απόδοσης των επιχειρησιακών διαδικασιών (http://design.inf.usi.ch/research/projects/benchflow, https://github.com/benchflow/benchflow) στο εργαλείο Camunda (https://camunda.org/) και σε δεύτερη φάση, στην βελτιστοποίηση επιχειρησιακών διαδικασιών προσαρμόζοντας τεχνικές από την διαχείριση δεδομένων [6]. Ο τελικός στόχος είναι να προκύψει προστιθέμενη αξία από τη μεταφορά τεχνογνωσίας μεταξύ των δύο περιοχών. Συνολικά, η εργάσια περιλαμβάνει μελέτη βιβλιογραφίας, πειραματισμό με εργαλεία, βασική έρευνα και προγραμματισμό.

[6] *Georgia Kougka, Anastasios Gounaris: Optimization of Data-intensive Flows: Is it Needed? Is it Solved? DOLAP 2014: 95-98 (http://delab.csd.auth.gr/~gounaris/2014dolap.pdf)*