ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ ΚΑΙ ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ
ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ 2015-16

1. Διδάσκων

Επικ. Καθηγ. Δαμιανός Χατζηαντωνίου
Email: damianos@aueb.gr
Ώρες γραφείου: Τετάρτη, 1μμ-3μμ, Ύδρας 28, γραφείο 202, 2ος όροφος, τηλ. 210-8203953.
Ώρες Μαθημάτων: Τρίτη 11πμ-1μμ, Αίθουσα Α25 και Τετάρτη 9πμ-11πμ, Αίθουσα ΔΟ.
Σημειώσεις Διαλέξεων: http://edu.dmst.aueb.gr

2. Βοηθός

Θα ανακοινωθεί

3. Περιγραφή-Στόχος

Η επιχειρηματική ευφυΐα είναι ένας ιδιαίτερα καινούργιος τομέας στις βάσεις δεδομένων. Έχει ως σκοπό την εκμετάλλευση των λειτουργικών δεδομένων ενός οργανισμού καθώς και την εννοποίηση εξωτερικών στοιχείων με τελικό στόχο την παραγωγή στατιστικών αναφορών και προτύπων για την υποστήριξη επιχειρηματικών αποφάσεων. Είναι μια τεχνολογία πληροφορίας για να βοηθήσει τον εργαζόμενο γνώσης (στέλεχος, manager, αναλυτή) να παίρνει πιο σύντομες και καλύτερες αποφάσεις.

Επίσης, τα τελευταία χρόνια, τα δεδομένα που χρησιμοποιούνται για την παραγωγή γνώσης όπως περιγράφεται παραπάνω, έχουν μεταβληθεί σε όγκο και μορφή. Πλέον μιλάμε για PB δεδομένων, που πιθανόν να περιλαμβάνουν και κείμενο, εικόνα και video. Επιπροσθέτως, σε κάποιες εφαρμογές απαιτείται ειδική αναπαράσταση των δεδομένων, για παράδειγμα σε μορφή γράφων. Για το λόγο αυτό, τα τελευταία χρόνια αναπτύσσεται μία νέα γενιά συστημάτων διαχείρισης δεδομένων, όπως το Hadoop, Redis, Cassandra, Neo4j και άλλα.

Ο σκοπός του μαθήματος είναι να παρουσιάσει τις τεχνικές που χρησιμοποιούνται για τη δημιουργία μίας αποθήκης δεδομένων (data warehouse) και κατόπιν να μελετήσει τις διάφορες μεθοδολογίες (αναλυτική επεξεργασία, πολυδιάστατη ανάλυση, εξόρυξη δεδομένων – data mining) για την παραγωγή *γνώσης* από αυτήν. Για την καλύτερη κατανόηση αυτών των εννοιών χρησιμοποείται ένα δημοφιλές εμπορικό σύστημα για την ανάπτυξη μίας πραγματικής εφαρμογής. Επίσης παρουσιάζονται σύντομα τα νέα συστήματα διαχείρισης δεδομένων που έχουν αναπτυχθεί τα τελευταία χρόνια για την υποστήριξη εφαρμογών big data.

4. Θέματα

- **Εβδομάδα 1**: Εισαγωγή. Περιγραφή μαθήματος, γενικές έννοιες για αποθήκες δεδομένων και εξόρυξη γνώσης, ειδικά θέματα.
- **Εβδομάδα 2**: Επανάληψη βασικών αρχών βάσεων δεδομένων. Μοντελοποίηση (ER, σχεσιακή σχεδίαση), SQL, Ευρετήρια, Βελτιστοποίηση, Συναλλαγές.
- **Εβδομάδα 3**: Παρουσίαση των ερευνητικών προσπαθειών που οδήγησαν στις έννοιες των αποθηκών δεδομένων και της εξόρυξης γνώσης.
- **Εβδομάδες 4-6**: Αποθήκες Δεδομένων: Γενικές αρχές και παραδείγματα. Αρχιτεκτονική, μοντέλα και σχεδίαση. Εξαγωγή, μετατροπή και εισαγωγή (ETL διαδικασία). Συντήρηση και ενημέρωση. Data marts. Ανάλυτική επεξεργασία (OLAP). Θέματα υλοποίησης και απόδοσης.
- **Εβδομάδα 7-9:** Εξόρυξη Γνώσης: Αρχιτεκτονική, διαδικασία KDD, μοντέλα, παραδείγματα. Συσταδοποίηση, Κατηγοριοποίηση, Κανόνες συσχέτισης, Χρονολογικές σειρές.

- **Εβδομάδα 10:** Εισαγωγή στα συστήματα διαχείρισης εφαρμογών big data.
- **Εβδομάδα 11-12:** Παρουσιάσεις συστημάτων από ομάδες φοιτητών.

5. Βιβλία

- **Υποχρεωτικό 1η επιλογή**: «Data Mining: Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα», Margaret Dunham, Εκδόσεις Νέων Τεχνολογιών, 2004.
- **Υποχρεωτικό 2η επιλογή**: «Εξόρυξη Γνώσης από Βάσεις Δεδομένων», Μ. Βαζιργιάννης και Μ. Χαλκίδη, Εκδόσεις Τηπωθήτω, 2003.
- Σημειώσεις και διαφάνειες του μαθήματος.

6. Βαθμολογία

Η βαθμολογία σας βασίζεται στα εξής:

- Εργασία 50%
- Παρουσίαση συστήματος 30%
- Εξετάσεις 20%

Θετικά υπολογίζεται η παρουσία στο μάθημα.
Υπάρχει η δυνατότητα απαλλακτικής ερευνητικής εργασίας (όχι ως προς το 20% των εξετάσεων).
Απαραίτητη η καλή γνώση προγραμματισμού και βάσεων δεδομένων.

7. Εργασία – Υποχρεωτική (50% του βαθμού).

α) [40%] Θα βρεθεί ένα μεγάλο data set, το οποίο θα πρέπει να καθαριστεί και να εισαχθεί σε μία αποθήκη δεδομένων, την οποία θα σχεδιάσουν οι ομάδες (2 φοιτητές ανά ομάδα). Θα δημιουργηθεί ένας κύβος δεδομένων και θα χρησιμοποιηθεί κάποιο εργαλείο για πολυδιάστατη ανάλυση (π.χ. Excel)

β) [40%] Τα δεδομένα της αποθήκης θα χρησιμοποιηθούν για κάποιες λειτουργίες εξόρυξης δεδομένων, όπως για παράδειγμα κατηγοριοποίηση, κανόνες συσχέτισης, συσταδοποίηση, κ.ο.κ. χρησιμοποιώντας μεθόδους και μοντέλα εμπορικού συστήματος ή ενός open-source εργαλείου.

γ) [20%] Θα εγκατασταθεί το Hadoop – ή θα χρησιμοποιηθεί στο cloud –και για το παραπάνω σύνολο δεδομένων θα γραφούν δύο απλά MapReduce jobs.

Για την εύρεση των datasets, μπορείτε να κάνετε αναζήτηση στο internet για σχετικούς ιστοτόπους (π.χ. http://archive.ics.uci.edu/ml/, http://www.kdnuggets.com/datasets/index.html)