Προχωρημένα θέματα Βάσεων Δεδομένων

http://eclass.uth.gr/eclass/courses/INFS129

Διδάσκων: Ευάγγελος Θεοδωρίδης (evangelos.theodoridis@gmail.com)
www.evangelostheodoridis.org

Περιγραφή Μαθήματος:

Η εξέλιξη της τεχνολογίας σε συνδυασμό με την καθολική χρήση του διαδικτύου έχει συμβάλλει σημαντικά στη συσσώρευση τεράστιων όγκων δεδομένων. Τα πεδία των εφαρμογών που απαιτούν εντατική επεξεργασία/ανάλυση δεδομένων είναι ποικίλα όπως εφαρμογές ιατρικής, βιοπληροφορικής, ανάλυση κοινωνικών δικτύων, ανάλυση πωλήσεων, επεξεργασία και ανάλυση επιχειρησιακών δεδομένων (δίκτυο πωλήσεων και πελατών, δεδομένων ERP), επεξεργασία και ανάλυση ροών δεδομένων από δίκτυα αισθητήρων κλπ. Στόχος του μαθήματος είναι η κατανόηση σύγχρονων «συστημάτων διαχείρισης βάσεων δεδομένων» (ΣΔΒΔ) και τεχνολογιών που επιτρέπουν την διαχείριση και ανάλυση μεγάλου όγκου δεδομένων.

Τα θέματα που θα καλύψει το μάθημα πιο αναλυτικά είναι:

- Βασικές έννοιες Συστημάτων Διαχείρισης Βάσεων Δεδομένων , Σχεσιακών ΣΔΒΔ, Ευρετηρίων.
- Σύγχρονες Τεχνολογίες:
    1. NoSQL/NewSQL (Column, Document/XML, Object/Relational, Graph/RDF, Key-Value, In-Memory) ΣΔΒΔ
    2. Χωρικών και Χρονικών ΣΔΒΔ
    3. Πολυμεσικών ΣΔΒΔ
    4. Data Warehousing και OLAP Συστημάτων
    5. Κατανεμημένων Συστημάτων διαχείρισης βάσεων δεδομένων (ΣΔΒΔ),
    6. Συστημάτων Εντατικής Επεξεργασίας Δεδομένων (Hadoop/Map Reduce, Streaming)
- Παρουσίαση θεμάτων ανάλυσης και εξόρυξης μεγάλου όγκου δεδομένων με στόχο την κατανόηση προτύπων, την εύρεση ομοιοτήτων, τον προσδιορισμό συσχετίσεων, τον εντοπισμό ομαλοτήτων ή ανωμαλιών (Data Analytics και Data Mining μέθοδοι).
- Θέματα Βελτιστοποίησης σε ΣΔΒΔ και Δομές Ευρετηρίων
    - Βελτιστοποίηση Ερωτημάτων
    - Χωρικά Ευρετήρια
    - Ευρετήρια Κειμένων
    - Χρονικά Ευρετήρια
    - Δεικτοδότηση στην δευτερεύουσα μνήμη
    - Κατανεμημένα Ευρετήρια

Προαπαιτούμενα Μαθήματα:

- Βάσεις Δεδομένων
- Δομές Δεδομένων και Αλγόριθμοι
- Ανάκτηση Πληροφορίας
- Εξόρυξη Γνώσης

Εργασίες Φοιτητών και Βαθμολόγηση

- Εκπόνηση Προγραμματιστικής Εργασίας [60%]
- Παρουσίαση Θεωρητικού Θέματος [40%]

Βιβλιογραφία

1. Database Management Systems (3 ed.). Raghu Ramakrishnan and Johannes Gehrke. 2002. McGraw-Hill, Inc., New York, NY, USA.
2. Database Systems Concepts (5 ed.). Abraham Silberschatz, Henry Korth, and S. Sudarshan. 2005. McGraw-Hill, Inc., New York, NY, USA.
3. Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
4. Data Mining: Concepts and Techniques. J. Han and M. Kamber, Morgan Kaufmann Publishers, Second Edition, 2006.
5. Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Cambridge University Press, New York, NY, USA.
6. Handbook of Massive Data Sets. James Abello, Panos M. Pardalos, and Mauricio G. C. Resende (Eds.). 2002. Kluwer Academic Publishers, Norwell, MA, USA.
7. Foundations of Multidimensional and Metric Data Structures. Hanan Samet. 2005. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Πλάνο Μαθημάτων:

1. Βασικές έννοιες Συστημάτων Διαχείρισης Βάσεων Δεδομένων , Σχεσιακών ΣΔΒΔ,
2. Δομές Δεικτοδότησης – Ευρετήρια Α
    a. One-dimensional (B-Trees, Hashing)
    b. Multidimensional – Spatial
3. Δομές Δεικτοδότησης – Ευρετήρια Β
    a. Textual
    b. Temporal
4. Ανάκτηση Πληροφορίας
5. Εξόρυξη Γνώσης
6. NoSQL/NewSQL (Document/XML, Object/Relational) ΣΔΒΔ
7. NoSQL/NewSQL (Column, Graph/RDF, Key-Value, In-Memory) ΣΔΒΔ
8. Συστήματα Εντατικής Επεξεργασίας Δεδομένων
    a. Batch
    b. Streaming
9. Παρουσιάσεις Εργασιών – Α
10. Παρουσιάσεις Εργασιών – Β

Σενάρια Προγραμματιστικής Εργασίας

- Retail –relational data
- Social Graph (Twitter) – graph data
- Arxiv.org – textual data

Θέματα Εργασιών:

1. RDBMS: Θέματα Βελτιστοποίησης Ερωτημάτων

I. **Query Processing** – [2] Chapter 21
II. Cagri Balkesen, Gustavo Alonso, Jens Teubner, M. Tamer Özsu: **Multi-Core, Main-Memory Joins: Sort vs. Hash Revisited**. 85-96 VLDB Endowment Vol7. No1, 2013
III. Mohammed Elseidy, Abdallah Elguindy, Aleksandar Vitorovic, Christoph Koch: **Scalable and Adaptive Online Joins**. 441-452 VLDB Endowment Vol7. No6, 2014
IV. **CONCURRENCY CONTROL** [1] Chapter 17
V. **PHYSICAL DATABASE DESIGN AND TUNING** [1] Chapter 20

2. NoSQL/NewSQL

I. Robert Escriva, Bernard Wong, and Emin Gün Sirer. 2012. **HyperDex: a distributed, searchable key-value store**. *SIGCOMM Comput. Commun. Rev.* 42, 4 (August 2012), 25-36. DOI=10.1145/2377677.2377681 http://doi.acm.org/10.1145/2377677.2377681
II. Biplob Debnath, Sudipta Sengupta, and Jin Li. 2010. **FlashStore: high throughput persistent key-value store**. *Proc. VLDB Endow.* 3, 1-2 (September 2010), 1414-1425. DOI=10.14778/1920841.1921015 http://dx.doi.org/10.14778/1920841.1921015
III. Ugur Cetintemel, Jiang Du, Tim Kraska, Samuel Madden, David Maier, John Meehan, Andrew Pavlo, Michael Stonebraker, Erik Sutherland, Nesime Tatbul, Kristin Tufte, Hao Wang, and Stanley Zdonik. 2014. **S-Store: a streaming NewSQL system for big velocity applications**. *Proc. VLDB Endow.* 7, 13 (August 2014), 1633-1636.
IV. Vishal Sikka, Franz Färber, Wolfgang Lehner, Sang Kyun Cha, Thomas Peh, and Christof Bornhövd. 2012**. Efficient transaction processing in SAP HANA database: the end of a column store myth**. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (SIGMOD '12). ACM, New York, NY, USA, 731-742. DOI=10.1145/2213836.2213946 **http://doi.acm.org/10.1145/2213836.2213946**
V. Daniel J. Abadi, Peter A. Boncz, and Stavros Harizopoulos. 2009. **Column-oriented database systems**. *Proc. VLDB Endow.* 2, 2 (August 2009), 1664-1665. DOI=10.14778/1687553.1687625 http://dx.doi.org/10.14778/1687553.1687625
VI. Thomas Neumann and Gerhard Weikum. 2010. **x-RDF-3X: fast querying, high update rates, and consistency for RDF databases**. *Proc. VLDB Endow.* 3, 1-2 (September 2010), 256-263. DOI=10.14778/1920841.1920877 http://dx.doi.org/10.14778/1920841.1920877
VII. Fabian Prasser, Alfons Kemper, and Klaus A. Kuhn. 2012. **Efficient distributed query processing for autonomous RDF databases**. In *Proceedings of the 15th International Conference on Extending Database Technology* (EDBT '12), Elke Rundensteiner, Volker Markl, Ioana Manolescu, Sihem Amer-Yahia, Felix Naumann, and Ismail Ari (Eds.). ACM, New York, NY, USA, 372-383. DOI=10.1145/2247596.2247640 http://doi.acm.org/10.1145/2247596.2247640
VIII. Stephan Müller, Hasso Plattner: **Aggregates Caching in Columnar In-Memory Databases**. 69-81 http://db.disi.unitn.eu/pages/VLDBProgram/pdf/IMDM/paper8.pdf
IX. Martin Faust, David Schwalb, Jens Krüger: **Fast Column Scans: Paged Indices for In-Memory Column Stores.** 15-27 http://db.disi.unitn.eu/pages/VLDBProgram/pdf/IMDM/paper2.pdf

3. Ευρετήρια Κειμένων

I. **Searching Large Text Collections** Ricardo Baeza-Yates, Alistair Moffat, Gonzalo Navarro [6]
II. **Indexing and Searching** [5] chapter 4
III. **Web Retrieval** [5] chapter 19, 20
IV. P. Ferragina and R. Grossi. **The string B-tree: a new data structure for string search in external memory and its applications**. J. ACM, 46(2):236-280, 1999.
V. Rahul Shah, Cheng Sheng, Sharma V. Thankachan, Jeffrey Scott Vitter, **Top-k Document Retrieval in External Memory**. ESA 2013: 803-81
VI. Frederik Transier , Peter Sanders, **Engineering basic algorithms of an in-memory text search engine**, ACM Transactions on Information Systems (TOIS), v.29 n.1, p.1-37, December 2010

VII. J. Shane Culpepper, Matthias Petri , Falk Scholer**, Efficient in-memory top-k document retrieval**, Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, August 12-16, 2012, Portland, Oregon, USA

VIII. Francisco Claude, J. Ian Munro: **Document Listing on Versioned Documents**. SPIRE 2013: 72-83

IX. Antonio Fariña, Nieves R. Brisaboa, Gonzalo Navarro, Francisco Claude, Ángeles S. Places, Eduardo Rodríguez: **Word-based self-indexes for natural language text**. ACM Trans. Inf. Syst. 30(1): 1 (2012)

4. Χωρικά Ευρετήρια/ΣΔΒΔ

I. **Multidimensional Point Data** (Quadtrees, k-d trees, Grid Directory, PK-trees) [7]

II. **Object based and image based image representations** [7]

III. **Intervals and small rectangles** [7]

IV. High Dimensional Data [7]

V. Tilmann Zäschke , Christoph Zimmerli , Moira C. Norrie, **The PH-tree: a space-efficient storage structure and multi-dimensional index**, Proceedings of the 2014 ACM SIGMOD international conference on Management of data, June 22-27, 2014, Snowbird, Utah, USA

5. Δομές στην 2ουσα μνήμη

I. **External Memory Data Structures** Lars Arge [6]

II. M. A. Bender, E. Demaine, and M. Farach-Colton."**Cache Oblivious B-Trees**" Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS), pages 399-409, 2000.

III. Pankaj K. Agarwal, Lars Arge, Sathish Govindarajan, Jun Yang, Ke Yi, **Efficient external memory structures for range-aggregate queries**. Comput. Geom. 46(3): 358-370 (2013)

IV. Lars Arge, Johannes Fischer, Peter Sanders, Nodari Sitchinava, **On (Dynamic) Range Minimum Queries in External Memory.** WADS 2013: 37-48R

V. Peyman Afshani, Lars Arge, Kasper Dalgaard Larsen: **Orthogonal Range Reporting in Three and Higher Dimensions**. FOCS 2009: 149-158

6. Κατανεμημένα Ευρετήρια

I. Wang, J., S. Wu, H. Gao, J. Li, and B. C. Ooi (2010). **Indexing multi-dimensional data n a cloud system**. In Proceedings of the 2010 international conference on Management of data, SIGMOD '10, New York, NY, USA, pp. 591{602. ACM.

II. Grossman, R. and Y. Gu (2008). **Data mining using high performance data clouds: experimental studies using sector and sphere**. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, NewYork, NY, USA, pp. 920-927. ACM.

III. Sai Wu, Dawei Jiang, Beng Chin Ooi, KunLung Wu, **Efficient Btree Based Indexing for Cloud Data Processing**, Proceedings of VLDB Endowment,2010.

7. Κατανεμημένα ΣΔΒΔ

I. **Distributed Transaction Management** [2] Chapter 23

II. **Distributed Concurrency Control** [2] Chapter 23

III. Distributed Query Optimization [2] Chapter 23

8. Συστήματα Εντατικής Επεξεργασίας Δεδομένων/Επεξεργασία Ροών Δεδομένων

I. Guoping Wang, Chee-Yong Chan. **Multi-Query Optimization in MapReduce Framework**. 145-156 VLDB Endowment Vol7. No3, 2013

II. Stefan Richter, Jorge-Arnulfo Quiané-Ruiz, Stefan Schuh, and Jens Dittrich. 2014. **Towards zero-overhead static and adaptive indexing in Hadoop**. *The VLDB Journal* 23, 3 (June 2014), 469-494. DOI=10.1007/s00778-013-0332-z http://dx.doi.org/10.1007/s00778-013-0332-z

III. Randall T. Whitman, Michael B. Park, Sarah M. Ambrose, and Erik G. Hoel. 2014. **Spatial indexing and analytics on Hadoop**. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (SIGSPATIAL '14). ACM, New York, NY, USA, 73-82. DOI=10.1145/2666310.2666387 **http://doi.acm.org/10.1145/2666310.2666387**

IV. Umut A. Acar and Yan Chen. 2013. **Streaming big data with self-adjusting computation**. In*Proceedings of the 2013 workshop on Data driven functional programming* (DDFP '13).

ACM, New York, NY, USA, 15-18. DOI=10.1145/2429376.2429382 **http://doi.acm.org/10.1145/2429376.2429382**

V. Zhengping Qian, Yong He, Chunzhi Su, Zhuojie Wu, Hongyu Zhu, Taizhi Zhang, Lidong Zhou, Yuan Yu, and Zheng Zhang. 2013**. TimeStream: reliable stream computation in the cloud**. In*Proceedings of the 8th ACM European Conference on Computer Systems* (EuroSys '13). ACM, New York, NY, USA, 1-14. DOI=10.1145/2465351.2465353 http://doi.acm.org/10.1145/2465351.2465353

9. Πολυμεσικές Βάσεις

I. **Multimedia Information Retrieval http://nordbotten.com/ADM/ADM_book/MIRS-frame.htm**

II. Gopal Pingali, Agata Opalach, Ingrid Carlbom, **Multimedia retrieval through spatio-temporal activity maps**, ACM Multimedia 01, pp. 129-136.

III. **Music Indexing and Retrieval for Multimedia Digital Libraries** http://www.springerlink.com/content/p772np0k371867uq/

IV. Ching-Hua Chuan, **Audio Classification and Retrieval Using Wavelets and Gaussian Mixture Models**, International Journal of Multimedia Data Engineering & Management, v.4 n.1, p.1-20, January 2013

V. Akshat Verma , Rohit Jain , Sugata Ghosal, **A utility-based unified disk scheduling framework for shared mixed-media services**, ACM Transactions on Storage (TOS), v.3 n.4, p.1-30, February 2008

10. Εξόρυξη Δεδομένων

I. **Clustering in Massive Data Sets** Fionn Murtagh [6]

II. **Data Squashing: Constructing Summary** Data Sets William DuMouchel [6]

III. **Anomaly Detection** [4] chapter 12

IV. **Mining Data Streams** [4]

V. **Mining Time-Series Data** [4]

VI. **Graph Mining and Social Network Analysis** [4]

VII. **Mining the World Wide Web** [4]

VIII. **Text Classification** [5] chapter 13

Αναφορές Συστημάτων:

1. RDBMS:
    a. MySQL
    b. SQL Server
    c. Oracle DB
    d. PostgreSQL
2. NoSQL/NewSQL DBMS
    a. Column Databases
        i. Cassandra
        ii. HBase
        iii. BigTable
    b. Document/XML database systems
        i. Apache CouchDB
        ii. MongoDB,
    c. Object-oriented-Relational database systems
        i. ObjectDB
        ii. PostgreSQL
        iii. Oracle DB
    d. Graph/RDF Databases
        i. Neo4J,
        ii. Oracle Spatial + Graph
        iii. JENA TDB
    e. (Key,Value) Stores
        i. MemcacheDB,

        ii. Redis,
3. In-memory DBMS
    a. Memcached
    b. Hazelcast
4. Χωρικές/Χρονικές Βάσεις Δεδομένων
    a. PostgreSQL
    b. Oracle Spatial + Graph
    c. GeoMesa
5. Συστήματα Εντατικής Επεξεργασίας Δεδομένων
    a. Apache Hadoop
    b. Apache Storm
    c. Apache Spark
6. Data warehousing και OLAP Συστήματα
    a. Modrian
    b. SQL Analysis Services
    c. Oracle DB OLAP
7. Εξόρυξη Δεδομένων
    a. Weka
    b. R-Project