Online Analytical Processing and Big Data Warehouses

Damianos Chatziantoniou, Associate Professor, Dept. of Management Science & Technology
Office: Hydras 28, 2nd floor, Tel: (210) 820-3953, Email: damianos@aueb.gr

Overview

Business Intelligence has evolved from a niche area for large enterprises and organizations to an essential infrastructure for all business entities, regardless size. The ability to integrate, analyze and aggregate large amounts of data in a simple and efficient manner became a necessity in the last decade or so. This course will present the motivation behind this field and discuss goals, design principles, querying approaches, processing techniques, systems, tools and applications. In addition, in the last few years, given the importance of business intelligence, specific systems have been designed to handle data warehousing, such as column-oriented and main-memory databases.

Key Outcomes

The key outcomes of this course are:

- Students will be able to understand the different phases in developing a data warehousing application: defining business goals, identifying data sources, designing star schemas, using tools/methods to extract/transform/load data, adding indexing structures to help performance, and present the results of the analysis.
- Students will be able to use a commercial or open-source system and tools to develop and deploy data warehousing applications.
- Understand and apply the additional technologies to bring business intelligence to the big data era.

Requirements and Prerequisites

Students should be familiar with relational database design and SQL.

Bibliography

Required Texts

- Multidimensional Databases and Data Warehousing (Synthesis Lectures on Data Management), by Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen (2010). Morgan and Claypool Publishers.
- Database Systems: The Complete Book, by Hector Garcia-Molina, Jeff Ullman, and Jeniffer Widom. More information at: http://infolab.stanford.edu/~ullman/dscb.html

Articles

- Daniel Abadi, Rakesh Agrawal, Anastasia Ailamaki, Magdalena Balazinska, Philip A. Bernstein, Michael J. Carey, Surajit Chaudhuri, Jeffrey Dean, AnHai Doan, Michael J. Franklin, Johannes Gehrke, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, H. V. Jagadish, Donald Kossmann, Samuel Madden, Sharad Mehrotra, Tova Milo, Jeffrey F. Naughton, Raghu Ramakrishnan, Volker Markl, Christopher Olston, Beng Chin Ooi, Christopher Ré, Dan Suciu, Michael Stonebraker, Todd Walter, Jennifer Widom: The Beckman report on database research. Commun. ACM 59(2): 92-99 (2016)
- Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States, http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf
- Surajit Chaudhuri, Umeshwar Dayal, Vivek R. Narasayya: An Overview of Business Intelligence Technology. Communications ACM 54(8): 88-98 (2011)
- Additional articles, specific to each lecture's topic, will be uploaded to moodle during the course.

Software/Computing Requirements

The two projects will require the use of a database management system and a data warehousing environment. Students have two alternatives for their projects:

- Microsoft SQL Server/Analysis Services/Tableau
- MySQL/Pentaho/Tableau

Grading

Project#1 : 20%
Project#2 : 50%
Final : 30%

Course Syllabus

Lecture 1:

An overview of database management principles; relational systems; SQL; transactions; SQL and complex aggregate queries. Motivation for data warehousing applications.

Lecture 2 & 3:

Introduction to BI; OLTP vs. OLAP; architecture, design and data modeling; the extract-transform-loading (ETL) process; data cubes; indexing for data warehouses. Systems and tools. Best practices for data warehousing.

Lecture 4:

Modern business intelligence: in-memory databases & column-oriented databases.
Case studies using SQL Server + Analysis Services and MySQL + Pentaho.

Lecture 5:

Business intelligence in the big data era: beyond relational databases and persistent data - integration of Hadoop, NoSQL engines, stream systems.

Participation

In-class contribution is a significant part of your grade and an important part of our shared learning experience. Your active participation helps us to evaluate your overall performance. You can excel in this area if you come to class on time and contribute to the course by:

- Providing strong evidence of having thought through the material.
- Advancing the discussion by contributing insightful comments and questions.
- Listening attentively in class.
- Demonstrating interest in your peers' comments, questions, and presentations.
- Giving constructive feedback to your peers when appropriate.

Please arrive to class on time and stay to the end of the class period. Chronically arriving late or leaving class early is unprofessional and disruptive to the entire class. Repeated tardiness will have an impact on your grade.

Turn off all electronic devices prior to the start of class. Cell phones tablets and other electronic devices are a distraction to everyone.

Assignments

No late assignments will be accepted.

Attendance Requirements

Attendance is required.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives express permission.

Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one’s own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Please report observed violations of this policy. Any violations will incur a fail grade at the course and reporting to the senate for further disciplinary action.