DS 220 Data Management for Data Sciences (3)

Description

An introductory course in advanced relational databases and issues related to managing non-relational data sets.

This three-credit introductory course teaches students techniques and processes for managing large data sets. It builds upon knowledge gained in IST 210 Organization of Data. This course has two major components: (1) advance students' knowledge in relational database and their skills in using SQL and database indexing; and (2) introduce NoSQL databases such as document-oriented database, key-value database, column-oriented database, graph database, and Hadoop system.

In the first component, the course will review the techniques learned in IST210, strengthen students' skills in using SQL queries and introduce students about indexing and scalability issue in relational database.

While relational database is still frequently used, the emergence of storage for big data and various types of data has driven a new of class of non-relational databases commonly referred to NoSQL database. This course will introduce the real-world needs for NoSQL databases and the characteristics that distinguish them from relational database. We will introduce both the concepts of NoSQL databases and how the concepts are implemented in the database systems. We will focus on three main NoSQL data models: key-value, column family, and document. You will learn the concepts of these data models and know how to use them in the database systems. We will also introduce graph databases, hadoop systems, and data warehousing. Finally, we will present criteria that decision makers should consider when choosing between relational and non-relational databases and techniques for selecting the NoSQL database that best addresses specific use cases.

Personnel

Prasenjit Mitra, Professor, College of Information Sciences and Technology
pmitra@psu.edu 814-865-4454
Office Hours: MWF 11-12

Yafei Wang, Teaching Assistant, Ph.D. Student, College of IST
yxw184@ist.psu.edu
Office Hours: TBD

Prasenjit Mitra is a Professor in the College of Information Sciences and Technology and the chair of the Faculty Council at the College. He serves on the graduate faculty of the Department of Computer Sciences. His current research interests are in the areas of big data analytics, applied machine learning, and visual analytics. In the past, he has contributed to research in the areas of data interoperation, data cleaning, and digital libraries especially in tabular data extraction, and citation recommendation.

Mitra received his Ph.D. from Stanford University in 2004 in Electrical Engineering. In 1994, he obtained an M.S. in Computer Science from The University of Texas at Austin. Prior to that, Mitra obtained his Bachelor of Technology, with honors in Computer Science & Engineering from the Indian Institute of Technology, Kharagpur in 1993. From 1995 to 2000, he worked at the Server Technologies Division at Oracle Corporation as a Senior Member on the Oracle Parallel Server in the Languages and Relational Technologies group. From 2014-2016, he served as a Principal Scientist at the Qatar Computing Research Institute. He has served as a consultant for several startups including the Board of Advisors of Global IDs, Inc.

At Penn State, he has pursued research on a broad range of topics ranging from data mining on the web and social media, scalable data cleaning, political text mining, chemical formula and name extraction from documents, and the extraction of data and metadata from figures and tables in digital documents.

He was the principal investigator of the DOES project funded by the NSF CAREER Award. Mitra serves as the director of the Cancer Informatics Initiative at Penn State. His research has been supported by the NSF, Microsoft Corporation, DoD, DHS, DoE, NGA, and DTRA.

Mitra has co-authored approximately 150 articles at top conferences and journals. His work along with his co-authors has resulted in a visual analytics system that was awarded the IEEE VAST '08 Grand Challenge award in the Data Integration area. He has supervised over 15 Ph.D. students; and several M.S. students.

Schedule

- Week 1 – Introduction: data types (text, images, web, transactions), data size, database applications
- Week 2 – Advanced database query practices
- Week 3 – Database indexing
- Week 4 – NoSQL foundations
- Week 5 – NoSQL database overview
- Week 6 – Document-oriented database
- Week 7 – Document-oriented database + Key-value stores
- Week 8 – Key-value stores
- Week 9 – Key-value stores
- Week 10 – Midterm + Guest Lecture
- Week 11 – Column-Oriented Store
- Week 12 – Column-Oriented Store
- Week 13 – Concepts of graph database
- Week 14 – Concepts of Hadoop system
- Week 15 – Concepts of data warehousing

Grading Policy

Class Attendance & Participation: 9%
Quizzes: 8%
Exams: Midterm 20% Final 30%
Homeworks: 16%
Project: 17%

Mid-terminal examination will be held on October 11th, in class.

All work must be submitted on the due date to get full credit. Late submissions will get reduced points based on how many days they are late. In case of emergencies or illnesses, etc., in most cases, the instructor will excuse the student's absence from class, and allow late submission. For excused absences, the student will notify the instructor about the reason for the absence as much in advance as possible. For known events, please let me know one week in advance. For unforeseen events, please notify us as soon as possible. Once the answers to an assignment have been posted, for obvious reasons, no further late submission for that assignment will be possible.

All work must be done individually unless otherwise mentioned. Consultation with fellow students is encouraged when the student's learning is enhanced by such consultations and discussions. However, the student is encouraged to write her or his own answers. Similar answers with substantial overlap especially when the student cannot explain the answer is a violation of the academic integrity policy of the college and will be reported and dealt with according to college policy.

94-100: A
88-94: A-
81-88: B+
75-81: B
70-75: B-
65-70: C+
60-65: C
50-60: D
<50: F

If the class puts in good effort but the exams are exceptionally hard, then the grading will be done on the curve at the discretion of the instructor. Grading on the curve can only improve the grade of a student. In the past, approximately, the top 20-25% of the students were be awarded A's, about 50% awarded B's, 15% C's, 10% D's, and 5% Fs. These ratios are meant for approximate guidance only.

Academic Integrity

Short Version: Do not copy from other students or sources without naming them. Do your own work. Be prepared to explain any part of any work you submit.

Please see the college's policy at: https://ist.psu.edu/students/academic_integrity

Statement of Nondiscrimination

The Pennsylvania State University is committed to the policy that all persons shall have equal access to programs, facilities, admission, and employment without regard to personal characteristics not related to ability, performance, or qualifications as determined by University policy or by state of federal authorities. The Pennsylvania State University does not discriminate against any person because of age, ancestry, color, disability or handicap, national origin, race, religious creed, sex, sexual orientation, or veteran status.

For additional information and for all inquiries regarding the nondiscrimination policy please contact:

Affirmative Action Director
The Pennsylvania State University
201 Willard Building
University Park, PA 16802-2801
Telephone: (814) 863-0471 U.Ed.OVP98-4

Disability Access Statement

Penn State welcomes students with disabilities into the University's educational programs. Every Penn State campus has an office for students with disabilities. Student Disability Resources (SDR) Web site provides contact information for every Penn State campus: http://equity.psu.edu/sdr/disability-coordinator. For further information, please visit Student Disability Resources Web site: http://equity.psu.edu/sdr.

In order to receive consideration for reasonable accommodations please contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: http://equity.psu.edu/sdr/guidelines. If the documentation supports your request for reasonable accommodations, your campus's disability services office will provide you with an accommodation letter. Please share this letter with your instructors and discuss the accommodations with them as early in your courses as possible. You must follow this process for every semester that you request accommodations.