CSE 3244 (Proposed): Data Management in the Cloud

Course Description

Systematic organization of data on cloud computing architectures; basic indexing techniques, including B-tree and hash-based indexing; fundamentals of query optimization, including access path selection and cardinality estimation; full and partial replication; data partitioning and distributed task scheduling.

Transcript Abbreviation: Data Mgmt in Cloud
Grading Plan: Letter Grade
Course Deliveries: Classroom
Course Levels: Undergrad
Student Ranks: Junior, Senior
Course Offerings:
Flex Scheduled Course: Never
Course Frequency: Every Year
Course Length: 14 Week
Credits: 3.0
Repeatable: No
Time Distribution: 3.0 hr Lec
Expected out-of-class hours per week: 6.0
Graded Component: Lecture
Credit by Examination: No
Admission Condition: No
Off Campus: Never
Campus Locations: Columbus
Prerequisites and Co-requisites: 3241 or 5241; 2421 or 3430.
Exclusions:
Cross-Listings:

Course Rationale: This will become a required course for data analytics majors. Also, it fills a gap in CSE curriculum between 3241 and 5242.

The course is required for this unit's degrees, majors, and/or minors: Yes
The course is a GEC: No
The course is an elective (for this or other units) or is a service course for other units: Yes

Subject/CIP Code:
Subsidy Level: Baccalaureate Course

Programs

Abbreviation	Description
BS CSE	BS Computer Science and Engineering

Course Goals

Master using fundamental concepts in indexing and optimization techniques, including B-trees, hash-based indexing and cardinality estimation
Master using mechanisms of distributed data management, including full and partial replication strategies, data partitioning, fault tolerance models and consistency trade-offs
Be competent with data warehousing techniques, including on-line analytical processing (OLAP)

Be competent with distributed algorithms and task scheduling in cloud environments
Be exposed to current cloud-based data management technologies

Course Topics

Topic	Lec	Rec	Lab	Cli	IS	Sem	FE	Wor
Review of relevant 3241/5241 Material	3.0							
Indexing and Optimization	12.0							
Data warehousing and OLAP	2.0							
Cloud computing principles	1.0							
Replication and partitioning strategies	4.0							
Algorithms and platforms for cloud	6.0							

Representative Assignments

Analyze public financial data in a cloud format (e.g. JSON) with a SQL-on-Hadoop tool. Evaluate query response time at scale on an IaaS platform.
Use an open-source cloud data processing platform (e.g. Hadoop) to create an inverted index from natural language dataset, e.g. Wikipedia.
Deploy iterative matrix-vector multiplication kernels on failure-prone cloud platforms. Evaluate the effects of replication on throughput.

Grades

Aspect	Percent
Homework	25%
Lab assignments	25%
Midterm Exam	25%
Final Exam	25%

Representative Textbooks and Other Course Materials

Title	Author
Fundamentals of Database Systems	Ramez Elmasri and Shamkant Navathe
Data-Intensive Text Processing with MapReduce	Jimmy Lin and Chris Dyer

ABET-EAC Criterion 3 Outcomes

Course Contribution		College Outcome
***	a	An ability to apply knowledge of mathematics, science, and engineering.
*	b	An ability to design and conduct experiments, as well as to analyze and interpret data.
**	c	An ability to design a system, component, or process to meet desired needs.
*	d	An ability to function on multi-disciplinary teams.
***	e	An ability to identify, formulate, and solve engineering problems.
*	f	An understanding of professional and ethical responsibility.
*	g	An ability to communicate effectively.
*	h	The broad education necessary to understand the impact of engineering solutions in a global and societal context.

Course Contribution		College Outcome
*	i	A recognition of the need for, and an ability to engage in life-long learning.
**	j	A knowledge of contemporary issues.
***	k	An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice.

BS CSE Program Outcomes

Course Contribution		Program Outcome
***	a	an ability to apply knowledge of computing, mathematics including discrete mathematics as well as probability and statistics, science, and engineering;
*	b	an ability to design and conduct experiments, as well as to analyze and interpret data;
**	c	an ability to design, implement, and evaluate a software or a software/hardware system, component, or process to meet desired needs within realistic constraints such as memory, runtime efficiency, as well as appropriate constraints related to economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability considerations;
*	d	an ability to function on multi-disciplinary teams;
***	e	an ability to identify, formulate, and solve engineering problems;
*	f	an understanding of professional, ethical, legal, security and social issues and responsibilities;
*	g	an ability to communicate effectively with a range of audiences;
*	h	an ability to analyze the local and global impact of computing on individuals, organizations, and society;
*	i	a recognition of the need for, and an ability to engage in life-long learning and continuing professional development;
**	j	a knowledge of contemporary issues;
***	k	an ability to use the techniques, skills, and modern engineering tools necessary for practice as a CSE professional;
**	l	an ability to analyze a problem, and identify and define the computing requirements appropriate to its solution;
**	m	an ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modeling and design of computer-based systems in a way that demonstrates comprehension of the tradeoffs involved in design choices;
**	n	an ability to apply design and development principles in the construction of software systems of varying complexity.

Prepared by: Kathryn Reeves