TCSS 490-590 Introduction to Big Data Management

Catalog Description

The course will discuss data management techniques for storing and analyzing very large amounts of data. The emphasis will be on columnar databases and on Map Reduce as a tool for creating parallel algorithms that can process very large amounts of data.

Topics include: Big Data applications, Columnar stores, distributed databases, Hadoop, Locality Sensitive Hashing (LSH), Dimensionality reduction, Data streams, unstructured data processing, NoSQL, and NewSQL.

Prerequisite: a minimum grade of 3.0 in TCSS 343 and TCSS445 (or equivalent course).

Preconditions

- Recognize and use mathematical formalisms (e.g., sets, logic, summations, proof).
- Translate problem descriptions into mathematical formalisms.
- Correctly employ programming language features by reading and interpreting the associated published API documentation.
- Recognize and apply different algorithm design techniques (including divide-and-conquer, decrease-and-conquer, transform-and-conquer, dynamic programming, and greedy approaches).
- Analyze the running times of algorithms.
- Design, implement, and document a medium-sized program that uses algorithms presented in class.

Student Learning Goals (to be added to syllabus handed out to students)

- Design a map reduce algorithm
- Identify implementation issues of a columnar database.
- Express data warehousing problems as Structured Query Language in database systems.
- Analyze the running times of data manipulation, storage and retrieval algorithms that constitute database systems internals and applications.
- Develop applications involving database systems such as Hive, InfiniDB, MemSQL, and Postgres.

CSS Degree Student Learning Outcomes that this course contributes to (to be added to syllabus handed out to students)

a. an ability to apply knowledge of computing and mathematics appropriate to the discipline;
b. an ability to analyze a problem, identify and define the computing requirements appropriate to its solution;
c. an ability to design, implement and evaluate a computer-based system, process, component, or program to meet desired needs;
i. an ability to use current techniques, skills, and tools necessary for computing practice.
UWT Student Learning Goals that this course contributes to (to be added to syllabus handed out to students)

Inquiry and Critical Thinking
Students will acquire skills and familiarity with modes of inquiry and examination from diverse disciplinary perspectives, enabling them to access, interpret, analyze, quantitatively reason, and synthesize information critically.

Topics covered
- Big Data Applications
- Columnar Storage and organization
- Distributed databases
- Hadoop
- Locality Sensitive Hashing (LSH)
- Dimensionality reduction
- Data streams
- Challenges in Unstructured data processing
- NoSQL
- NewSQL
- (time permitting) Introduction to Mining massive datasets

Additional Information
This is an introductory course in implementation issues of database systems being used for the big data challenges for CS majors. Columnar Databases and Hadoop are becoming increasingly important in today’s interconnected world. These database technologies drive the way we store, manage and extract information. Hence the study of big data database systems encompasses various characteristics of the study of information and computing systems. Such database systems are programs that require a fundamental understanding of data structures. We will engage in seminar style presentation and discussions, short in-class quizzes, and programming assignments.

Grading scheme will be: 50% class activity including presentations, 10% quizzes, 40% assignment(s). Learn by doing. The more you practice working with some of these open-source database systems and frameworks to solving relevant problems the more you will learn. This course is intended to be part of a 3 course sequence TCSS445, this course OR TCSS590 Data Analytics, TCSS555 Data Mining. Graduate students enrolled in 590 will undertake an additional project focusing on an application domain.

Textbook and Reading Materials:
Reading material will be provided by the instructor.