GRADUATE PROGRAMS IN SOFTWARE
University of St. Thomas
St. Paul, MN 55105

SYLLABUS SPRING 2014
SEIS736

Course Title
Big Data Architecture

Meeting Time/Place
Tuesdays, 5:45-9:00 PM, OSS-432

Credit Value
3 semester credits

Prerequisites
SEIS601 Foundations of Software Development or Java Programming Experience
SEIS630 Database Management Systems and Design (concurrent enrollment allowed)

Instructor
Bradley S. Rubin, Ph.D. (Brad)

E-Mail Address
bsrubin@stthomas.edu

Phone Number
651-260-5676 (cell)
651-962-5506 (UST, only during office hours)

Office Address
OSS-312

Office Hours
Tuesdays and Thursdays, 3:30 – 5:30 PM
(Email and phone calls welcome too)

Course Description
This course covers emerging big data architectures, predominately Hadoop and related technologies that deal with large amounts of unstructured and semi-structured data. Topics include operating system, architecture, security, big data structure and storage. The primary applications discussed in this class focus on information retrieval, specifically text processing techniques and algorithms, such as parsing, stemming, compression, and string searching. Information retrieval is also a great case study for broader issues in building systems that scale and perform, so we discuss associated issues in data structures, algorithms, computational complexity, and measurement.

Course Objectives

- Gather factual knowledge (terminology, classifications, methods, trends) in information retrieval
- Learn to apply course material (to improve thinking, problem solving, and decisions)
- Develop specific skills, competencies, and points of view needed by professionals in the field most closely related to big data and information retrieval

Required Texts

- White, Tom, "Hadoop: The Definitive Guide," 3rd edition
- Lin, Jimmy and Dyer, Chris, "Data-Intensive Text Processing with MapReduce"
  Note: Free online at http://lintool.github.io/MapReduceAlgorithms/

Grading Policy

-Weekly homework (25%)
- Written exercises and programming assignments
- 10% of the possible points off for each day late (max 1 week late)
- Assigned Tuesday, due following Tuesday at midnight, returned on the 2nd Tuesday (turned in on paper during class (preferred) or emailed to me)

-Midterm (25%)
-Closed book/note

-Final (25%)
-Closed book/note

-Project (25%)
- Individual

Blackboard

- Lecture notes (pdf format), homework assignments/answers, and grades will be available on Blackboard each week (usually one day in advance of class)

Weekly Class Schedule

90 min Lecture
15 min Break
30 min Lecture, Homework Assignment
60 min In-class lab on most nights, no points but usually a prerequisite for the homework assignment

Course Outline

Week 1
Introduction to Big Data
Introduction to Hadoop

Week 2
Hadoop Distributed File System (HDFS)
Hadoop I/O

Week 3
Combiners
Developing for the Cluster

Week 4
- How MapReduce Works
- Hadoop Streaming
- Partitioners
- MRUnit

Week 5
- MapReduce Types and Formats
- MapReduce Features
- Course Projects

Week 6
- Demos: Maven/Eclipse, Flume, Hive, Amazon Elastic MapReduce, IPMI, Cloudera Manager
- Midterm Review Q&A

Week 7
- Midterm Exam

Week 8
- Midterm Exam Results
- Introduction to Information Retrieval
- Information Retrieval Models

Week 9
- Query Languages
- Regular Expressions
- Relevance Feedback
- Text Characteristics

Week 10
- Text Operations
- Text Compression

Week 11
- TFIDF in MapReduce
- Secondary Sorts and Joins
- Distributed Cache and MapFiles

Week 12
- Computational Complexity
- PageRank
- Data Structures for Indexing

Week 13
- Big Data Privacy and Security Issues
- Data Storage Technologies and Trends
- Case Study or Guest Speaker

Week 14
- Final Exam
- Projects Due

Attendance Policy

- Attendance sheet must be initialed each week
- Maximum of two absences expected

Academic Integrity

Academic integrity is defined as not cheating and not plagiarizing; honesty and trust among students and between students and faculty are essential for a strong, functioning academic community. Consequently, students are expected to do their own work on all academic assignments, tests, projects and research/term papers. Academic dishonesty, whether cheating, plagiarism or some other form of dishonest conduct related to academic coursework and listed in the Student Policy Book under "Discipline: Rules of Conduct" will automatically result in failure for the work involved. But academic dishonesty could also result in failure for the course and, in the event of a second incident of academic dishonesty, suspension from the University.

Here are the common ways to violate the academic integrity code:

Cheating - Intentionally using or attempting to use unauthorized materials, information, or study aids in any academic exercise. The term academic exercise includes all forms of work submitted for credit.

Fabrication **-** Intentional and unauthorized falsification or invention of any information or citation in an academic exercise.

Facilitating Academic Dishonesty - Intentionally or knowingly helping or attempting to help another to violate a provision of the institutional code of academic integrity.

Plagiarism - The deliberate adoption or reproduction of ideas or words or statements of another person as one's own without acknowledgment. You commit plagiarism whenever you use a source in any way without indicating that you have used it.

Cheating

In cases of cheating, the instructor will impose a minimum sanction of failure of work involved. The instructor will inform the student and the program director in writing of:

1. the nature of the offense,
2. the penalty imposed within the course;
3. the recommendation of the instructor as to whether further disciplinary action by the director is warranted.

If the instructor or the director of the program determines that further disciplinary action is warranted, a disciplinary hearing shall be commenced at the request of either the instructor or the director. (If there is a previous offense of this nature on the student's record, a hearing is mandatory.)

Here are examples of various kinds of plagiarism. In each instance, the source is a passage from p. 102 of E.R. Dodd's The Greek and the Irrational (Berkeley, 1971; reprinted: Boston: Beacon, 1957). First here is the original note, copied accurately from the book Functions, Dodds 12, p. 102: "If the waking world has certain advantages of solidity and continuity its social opportunities are terribly restricted. In it we need as a rule, only the neighbors whereas the dream world offers the chance of intercourse, however fugitive, with our distant friends, our dead and gods. For normal men it is the sole experience in which they escape the offensive and incomprehensible bondage of time and space."

Here are five ways of plagiarizing this source: (If you have any questions about plagiarism ask the instructor)

1. Word-for-word continuous copying without quotation marks or mention of the author's name.

Dreams help us satisfy another important psychic need - our need to vary our social life. This need is regularly thwarted in our waking moments. If the waking world has certain advantages of solidity and continuity, its social opportunities are terribly restricted. In it we need, as a rule, only the neighbors,

whereas the dream world offers the change of intercourse, however fugitive, with our distant friends, our dead, and our gods. We awaken from such encounters feeling refreshed, the dream having liberated us from the here and now...

2. Copying many words and phrases without quotation marks or mention of the author's name.
Dreams help us satisfy another important psychic need - our need to vary our social life. In the waking world our social opportunities, for example, are terribly restricted. As a rule, we usually encounter only the neighbors. In the dream world, on the other hand, we have the chance of meeting our distant friends. For most of us it is the sole experience in which we escape the bondage of time and space....
3. Copying an occasional key word or phrase without quotation marks or mention of the author's name.
Dreams help us satisfy another important psychic need - our need to vary our social life. During our waking hours our social opportunities are terribly restricted. We see only the people next door and our business associates. In contrast, whenever we dream, we can see our distant friends. Even though the encounter is brief, we awaken refreshed, having freed ourselves from the bondage of the here and now...
4. Paraphrasing without mention of the author's name.
Dreams help us satisfy another important psychic need - our need to vary our social life. When awake, we are creatures of this time and this place. Those we meet are usually those we live near and work with. When dreaming, on the other hand, we can meet far-off friends. We awaken refreshed by our flight from the here and now.
5. Taking the author's idea without acknowledging the source.
Dreams help us to satisfy another important psychic need - the need for a change. They liberate us from the here and now, taking us out of the world we normally live in....

If you quote anything at all, even a phrase, you must put quotation marks around it, or set it off from your text; if you summarize or paraphrase an author's words, you must clearly indicate where the summary or paraphrase begins and ends; if you use an author's idea, you must say that you are doing so. In every instance, you also must formally acknowledge the written source from which you took the material. **(This includes material taken from the World Wide Web and other Internet sources.)** Reprinted from "Writing: A College Handbook" by James A.W. Herrerman and John E. Lincoln. By Permission W.W. Norton & Co. Inc., Copyright 1982 by W.W. Norton & Co. Inc. Students are encouraged to report incidents of academic dishonesty to course instructors.

When academic dishonesty occurs, the following procedures will be followed.
A. The instructor will impose a minimum sanction of failure for the work involved. The instructor will notify the student and the appropriate academic dean/director in writing of the nature of the offense and that the minimum sanction has been imposed. The instructor may recommend to the dean that further penalties should be imposed. If further penalties are imposed, the dean/director will notify the student immediately and the student will have five working days to respond to the intention to impose additional penalties. The student has the right to respond to the charge of academic dishonesty and may request in writing that the dean review the charge of academic dishonesty as fully as possible. If the dean/director determines that no further sanctions will be applied, the instructor's sanction will stand and the instructor's letter to the dean/director and student will be placed in the student's file. If no further charges of academic dishonesty involving the student occur during the student's tenure at St. Thomas, the materials will be be removed from the file upon graduation.
B. If the student has been involved in a previous incident of academic dishonesty, the dean will convene a hearing, following guidelines listed under "Hearings and Procedures" in the Student Policy Book. During the hearing, all violations of academic integrity will be reviewed. The student and the faculty member charging the most recent incident will be present at the hearing.

C. In either situation, A or B, if the dean/director determines that further sanctions are warranted, the student will be informed in writing. Among the sanctions considered by the dean/director will be the following: failure for the course in which the incident occurred; suspension from the university for the following semester; expulsion from the university; community service; a written assignment in which the student explores the principles of honesty and trust; other appropriate action or sanctions listed under "Sanctions" in the Student Policy Book. The materials relating to the incident including the instructor's original letter to the student and dean and the dean's decision following the hearing, will become part of the student's file.

Students with Disabilities

I want to ensure that the classroom environment is conducive to your learning and ask that you discuss with me any concerns that are interfering with your learning as they arise. Classroom accommodations will be provided for students with documented disabilities. Students must contact the Disability Resources Office about accommodations for this course as early in the semester as possible. Appointments can be made by calling 651-962-6315 or 800-328-6819, extension 6315, or in person in Rm 110 Murray Herrick Center on the St. Paul campus. Further information is available at: www.stthomas.edu/enhancementprog/.

Recording of Classroom Activities

All recordings of class sessions using any device is expressly prohibited without the written permission of the instructor. (See **Class Session Recording Permission Form.**)