CSE 891: Computational Techniques for Large-Scale Data Analysis for MSBA (Spring 2014)

GENERAL INFORMATION
Lecture Hours: Mon, Wed 10:20 – 11:40am
Lecture Room: 306 Natural Resources Building
Class Web page: http://www.cse.msu.edu/~cse891 (Section 6)
Supporting Blog: http://alproductions.us/cse891

INSTRUCTOR
Arend Hintze
Office: BPS 2228
Office Hours: Mon, Tue, 1.30pm - 3.00pm (email for additional meetings)
Phone Number: 517-355-8733 (ask for Arend)
Email: hintze at msu dot edu

COURSE OVERVIEW
The new millennium has ushered in the era of big data and data-intensive computing. As storage becomes cheaper and computers become more powerful, the need for advanced analytical solutions to harness the potential of big data has become increasingly important. This course is intended for graduate students who are interested in gaining hands-on experience applying computational techniques to solve large-scale data analysis problems. For example: How do you retrieve data using APIs from Twitter and other social media web sites? How do you segment customers in order to provide precise, personalized services? How do we predict customers who will churn to a competitor?

COURSE OBJECTIVES
• To introduce the general concepts and methods for large-scale computational data analysis
• To teach basic programming skills and tools for collecting, storing, querying, and analyzing large-scale data

COURSE OUTCOMES
• Students will have a good understanding of the capabilities provided by various data analysis methods and apply the appropriate ones to solve real-world problems
• Students will gain hands-on experience in using data analysis tools
• Students will gain experience writing programs for collecting, querying, and analyzing data

PREREQUISITE AND BACKGROUND
CSE 232 (Introduction to Programming II) or equivalent. Students should demonstrate a basic understanding of algorithms and data structures and methods to create reliable programs. Students must have working knowledge of abstract data types and classes.
Such background is needed to ensure that students can understand the sample programs provided in class and able to write their own programs for homework assignments and class project. An integral part of the course is the Hadoop distributed computing platform, which is written in Java programming language. Thus, familiarity with data abstraction, classes, and object-oriented concepts (such as inheritance and polymorphism) will be needed.

REFERENCE BOOKS
There is no main textbook for the class. Lecture notes will be posted online and are based on materials from the following books:
• Introduction to Data Mining (Pang-Ning Tan, Michael Steinbach, and Vipin Kumar), Addison Wesley, 2006
• Mining of Massive Datasets (Anand Rajaraman, Jeff Ullman), Cambridge University Press, 2011
• Hadoop In Action (Chuck Lam), Manning Publications, 2011
• Data Science for Business (Foster Provost & Tom Fawcett), O’Reilly Media, 2013
COURSE OUTLINE

- Data collection, storage, and preprocessing (2.5 weeks)
- Data analysis (4-5 weeks)
- Implementation (5-6 weeks)
- Case studies and applications (1.5 weeks)

COURSE ASSESSMENT

Homework (40%), Exams (20%), Exercise (20%), and Project (20%)

GRADING CRITERIA

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>90%+</td>
</tr>
<tr>
<td>3.5</td>
<td>85%</td>
</tr>
<tr>
<td>3.0</td>
<td>80%</td>
</tr>
<tr>
<td>2.5</td>
<td>75%</td>
</tr>
<tr>
<td>2.0</td>
<td>70%</td>
</tr>
<tr>
<td>1.5</td>
<td>65%</td>
</tr>
<tr>
<td>1.0</td>
<td>60%–</td>
</tr>
</tbody>
</table>

COURSE POLICY

- Homework assignments are due before 10am on the due date. Late submissions will be deducted 25% of the total possible assignment grade (if submitted after the deadline but on the same day) or 50% (if submitted a day after the deadline). Assignments submitted more than one day after the deadline will not be accepted (unless you receive permission from the instructor). You must use the handin system (https://secure.cse.msu.edu/handin/) to submit your homework.
- In-class exercises must be completed before noon on the day it was assigned and submitted via handin. Not all exercises will be graded. The three lowest scores you receive for the in-class exercises will not be counted towards your final grade.
- You are encouraged to form study groups to learn the materials in class. However, all submitted assignments (including computer programs) must be your own work. If plagiarism is detected, students will automatically receive a 0 for the grade and will be reported to the university.
- The instructor reserves the right to modify the course content and class schedule during the semester.
- Make-ups for examinations may be arranged if your absence is caused by documented illness or personal emergency. A written explanation (including supporting documentation) must be submitted to your lecture instructor; if the explanation is acceptable, an alternative to the examination will be arranged. When possible, make-up arrangements must be made in advance.
- All students are expected to be responsible users of the computer system provided for this course. Account usage guidelines published by the Department of Computer Science and Engineering are posted at http://www.cse.msu.edu/facility/policy.php.
- The Department of Computer Science expects all students to adhere to MSU’s policy on Integrity of Scholarship and Grades. Information about MSU policy regarding academic integrity is available at https://www.msu.edu/~ombud/academic-integrity/index.html.
- Students who require accommodation under the Americans with Disabilities Act (ADA) with MSU's Resource Centers for Disabilities (RCPD) should bring their Verified Individualized Services and Accommodations (VISA) form to the instructor at the beginning of the semester as possible.