CS 6303 – Big Data Analytics

Catalog Description: The course introduces concepts and techniques in managing and analyzing large data sets for data discovery and modeling. Topics include big data storage systems, parallel processing platforms, and scalable machine learning algorithms.

Course Prerequisites: Credit in CS 5310 (Data Mining), or with consent of the instructor.

Learning Objectives: After taking this course, students should be able to

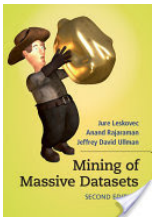
LO1. Explain how big-data has become a new norm of life and what challenges are associated with it in building models and ultimately, discovering knowledge.

LO2. Master the topics on the tools, algorithms and platforms required to store and analyze big data. In particular, the students will obtain knowledge on parallel processing platforms such as Hadoop and Spark. They will learn several data storage methods such as Hadoop Distributed File System (HDFS), HBase, document database and graph database. They will be introduced scalable machine learning algorithms via software tools such as to Apache's Mahout.

LO3. Develop visualizations for large datasets using JavaScript and D3 (Data-Driven Documents). Typically, these web-based visual representations are designed to communicate important elements of the data.

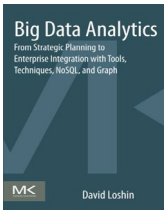
LO4. Design highly scalable systems for storing and analyzing large volumes of unstructured data.

Required Textbooks



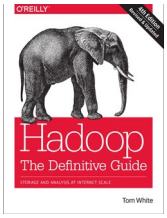
- Mining Massive Datasets, by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, ISBN: 978-1-107-07723-2.

References

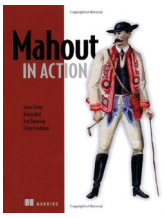


- Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph, by David Loshin, 2013, 1st Edition, Morgan Kaufmann publisher.
  Paper Book-ISBN 13: 978-0124173194

- Hadoop: The Definitive Guide by Tom White, 2015, 4th Edition, O'Reilly Media.
  Paper Book-ISBN 13: 978-1491901632



- Mahout in Action, by John Foreman by Sean Owen, Robin Anil, Ted Dunning, 2011, 1st Edition, Manning Publications.
  ISBN: 978-1935182689



Course Topics: The following topics will be covered as time permits.

1. Introduction to Big Data Analytics
2. Big Data Analytics Platforms
3. Big Data Storage and Processing
4. Big Data Analytics Algorithms: Recommendations
5. Big Data Analytics Algorithms: Clustering
6. Big Data Analytics Algorithms: Classification
7. Big Data Visualization

Workload: 5-7 hours/week

Course Grade: Course grades will be determined as follows:

Assignment	Weight
Exam-1	20 %
Exam-2	20 %
Final Exam	30 %
Labs	10 %
Research Topic Review (10-page report 10% + 15 minute Presentation 10%)	20 %

The grade on research writing and presentation will be determined as 50% writing and 50% presentation. The presentation will be peer-evaluated by anonymous survey. In order to ensure the validity of the survey, the missing portion of the evaluation from non-participating peers will be filled in by the instructor's evaluation.

Your final course grade will be determined by the standard college formula based on your course average:

90-100 → "A", 80-89 → "B", 70-79 → "C", 60-69 → "D", 0-59 → "F"

Topic Prerequisites: The course is essentially self-contained. The necessary material from statistics is integrated into the course.

Online Course Support: The Blackboard system (https://bb.uhd.edu/) will be used for online course material. As the semester progresses, various materials will be posted there including lecture notes, projects, and course announcements.

MAKE-UP POLICIES

- **Course projects/Homework assignments:** are to be completed and turned in *by the due date*. **For each late day, 15% of the total possible points will be deducted** (a day ends at the due time). No work will be accepted more than 5 days late.
- **Exams:** Make-up exams will *only* be given in cases of documented emergencies. It is your responsibility to contact your instructor with documentation of your emergency as soon as possible.
- **Quizzes:** No Make-ups for quizzes.
- All missed grades will be recorded as zeros.

CLASS POLICIES

- **Student Conduct In Class Policy:** Any acts of classroom disruption that go beyond the normal rights of students to question and discuss with instructors the educational process relative to subject content will not be tolerated, in accordance with the Academic Code of Conduct described in the Student Handbook.
- **Children In Class Policy:** Only in extreme cases are children allowed in classroom or laboratory facilities, and then only with approval of the instructor prior to class.
- **Electronic Devices In Class Policy:** Cellular phones, pagers, CD players, radios, and similar devices are prohibited in the classroom and laboratory facilities. Calculators and computers are prohibited during examinations and quizzes, unless specified. Laptops and tablets may be used in lecture for the purpose of taking notes.
- **Academic Dishonesty:** For this class, all work must be done individually -- no group work is allowed. You are encouraged to generally discuss assignments with fellow students, but may not copy their solution or code. **Doing so constitutes academic dishonesty which will be sanctioned with a grade of F in the course.** See http://www.uhd.edu/about/hr/PS03A19.pdf for more information on UHD's policy on academic dishonesty.
- **Statement on Reasonable Accommodations:** The University of Houston-Downtown complies with Section 504 of the Rehabilitation Act of 1973 and the Americans with Disabilities Act of 1990, pertaining to the provision of reasonable academic adjustments/auxiliary aids for students with a disability. In accordance with Section 504 and ADA guidelines, UHD strives to provide reasonable academic adjustments/auxiliary aids to students who request and require them. If you believe that you have a documented disability requiring academic adjustments/auxiliary aids, please contact the Office of Disability Services, One Main St., Suite 409-South, Houston, TX 77002.
  Contact info: 713-226-5227, disabilityservices@uhd.edu, www.uhd.edu/disability/

CS 6303 - Course Schedule

(This schedule is subject to update. You should check the schedule regularly for assignments and due dates)

Week	Monday (virtual)	Tuesday	Wednesday (virtual)	Thursday
1	7/7 Chapter 1	7/8 Chapter 2 Hadoop lab	7/9 Chapter 2 Chapter 3	7/10 Chapter 3
2	7/14 Chapter 3	7/15 **Midterm Exam 1** Chapter 4	7/16 Chapter 4	7/17 Chapter 4 Spark lab
3	7/21 Chapter 5	7/22 Chapter 5	7/23 Chapter 5	7/24 Chapter 5 SparkR lab
4	7/28 Chapter 5	7/29 **Midterm Exam 2** Chapter 7	7/30 Chapter 7	7/31 Chapter 9
5	8/4 Chapter 9	8/5 Writing projects presentation	8/6 Review	8/7 **Final Exam**