USC Viterbi School of Engineering

DSCI 551: Foundations of Data Management
Units: 4

Term—Day—Time:
Fall 2022

32413D: **RTH 217**, Tuesday 10:30-1:50pm
32443D: DEN
32447D: Online

32414D: **SLH 200**, Tuesday 3:30-6:50pm

All sections: please access D2L website for course content. Also see the website for more up-to-date syllabus.

Instructor: Wensheng Wu
Office Hours: To be announced on web site.
Contact Info: wenshenw@usc.edu

Course producers: To be announced on web site.
Office Hours: TBA.

A. Catalogue Course Description

Function and design of modern storage systems, including cloud; data management techniques; data modeling; network attached storage, clusters and data centers; relational databases; the map-reduce paradigm.

B. Expanded Course Description

This course is one of the foundation courses in the Applied Data Science program. It prepares the students with the fundamental knowledge on the data management. Such knowledge is critical for the students to succeed in more advanced data management courses in the program. It also exposes students to the cutting-edge data management concepts, systems, and techniques for managing a large scale of data, to ensure that students have adequate background to further explore big data analytics in the follow-up courses.

The course may be divided into three parts. (1) Fundamental of data management: data storage, file system, file format, relational data vs. semi-structured data such as XML and JSON, conceptual modeling, relational modeling, relational algebra, SQL, views, constraints, query processing and optimization. (2) Big data analytics: NoSQL, key-value and document stores, cloud data storage, distributed file system, and MapReduce. (3) Advanced topics in data management (if time permits): data cleaning, data transformation, data warehousing, and data integration.

The course will also provide students with hand-on experiences on RDBMS, e.g., MySQL, NoSQL & cloud databases such as Google Firebase, Amazon DynamoDB, MongoDB, and big data platform & software stacks, e.g., AWS EC2, Apache Hadoop, and Spark.

C. Recommended Preparation:

DSCI 550 taken previously or concurrently. Basic understanding of operating systems, networks, and databases. A basic understanding of engineering principles is required, including basic programming skills; familiarity with the Python is required & knowing Java programming language is desirable.

D. Course Notes

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings and students are encouraged to do the readings prior to the discussion in class. All course materials, including the readings, lecture slides, homework will be posted online on the course Web site.

E. Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in Python (preferably Java too). Students are also expected to have their own laptop or desktop computer where they can install and run software to complete the homework assignments and project.

F. Recommended Readings and Supplementary Materials

- [**AA**] Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau. *Operating Systems: Three Easy Pieces*, 2015 (selected chapters only). Available free at: http://pages.cs.wisc.edu/~remzi/OSTEP/
- [**GUW**] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. Database Systems: The Complete Book (Second Edition), Prentice Hall, 2009 (selected chapters only, see schedule below). Book web site: http://infolab.stanford.edu/~ullman/dscb.html
- [SQL] Alan Beaulieu. Learning SQL: Generate, Manipulate, and Retrieve Data. 3rd Edition. O'Reilly, 2022. Freely accessible from USC library.
- [MongoDB] Kristina Chodorow. MongoDB: The Definitive Guide, 2nd Edition. O'Reilly, 2013. Freely accessible from USC library.
- [Hadoop] Tom White. Hadoop: The Definitive Guide. O'Reilly Media; 2010. Freely accessible from USC library.
- [Spark] Chambers, Bill ; Zaharia, Matei. Spark: Big Data Processing Made Simple. O'Reilly Media, 2018. Freely accessible from USC library.

Note that the last four books are freely accessible from USC library. Links can be found above.

In addition to the textbook, students may be given additional reading materials such as research papers. Students are responsible for all reading assignments.

G. Grading Scheme

Homework Assignments: There will be 5 homework assignments. The assignments must be done individually. Each assignment is typically graded on a scale of 0-100 and the specific rubric for each assignment will be provided for the assignment.

Exams: There will be two midterms and a final exam. The final exam will be comprehensive, but focus on materials after the 2nd midterm.

Lab sessions: There will be 4 hand-on exercises on NoSQL and big data software.

Course project: Students are also expected to complete a term project on managing data for data science. Details will be announced separately. Note that in addition to in-class demo, each group is required to submit up to 20-minute video for the detailed presentation and demo of their project.

Grade breakdown:

Homework	20%
Midterm1	15%
Midterm2	15%
Final	25%
Lab task	5%
Course project	20%
Total	100%

Letter grades will range from A through F. The following are the cut-offs:

[**94**, 100] = A	[73, 76) = C
[90, **94**) = A-	[70, 73) = C-
[87, 90) = B+	[67, 70) = D+
[83, 87) = B	[63, 67) = D
[80, 83) = B-	[60, 63) = D-
[77, 80) = C+	Below 60 is an F

Note that this is an absolute grading (no curving will be applied). **<u>Note also that the cut-off for A is 94</u>**. Grades are not negotiable. No rounds up will be performed. Requests for rounding up, asking for special treatments, etc. will be ignored and may be subject to penalty (e.g., 10% deduction of the grade).

H. Grading Policy

Your coursework (including homework assignments, labs, and project deliverables) is due at 11:59pm on the due date and should be submitted on the course Web site as announced. **<u>No late submissions will be accepted.</u>**

Makeup for exams will be not permitted unless there are documented medical emergencies. Doctor notes are needed as proof. Two-week in advance notices are required for scheduling a makeup. No makeups will be given for situations such as interview, job fairs, etc. Students are responsible for scheduling to avoid conflicts with class meeting times and for any missing coursework under these situations. Students may be required to contact the Student Advocacy Services office (contact information will be provided in class) to submit proper documents for the verification of emergency.

Regrading requests must be made (by emailing TAs) within one week after the solutions or grades have been posted. Grades are final after the regrading period. Final exam grades and all class grades are final after final exam grading review hours (which are typically announced shortly after the final exam).

I. Course Schedule: A Weekly Breakdown (may be revised when the course progresses)

Week	Topic	Readings	Homework/Project	Lab
1 (8/22)	• Data Management Overview	• [AA] Chapter 2 (optional) • [AA] Chapter 4 (optional)		
2 (8/29)	• NoSQL 1: Firebase & JSON	• [AA] Chapter 37 (storage system)		
3 (9/5)	• Storage System • File System	• [AA] Chapter 39 • [AA] Chapter 40	HW1 out	Lab 1: Amazon EC2
4 (9/12)	• Hadoop HDFS	• K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, 2010, pp. 1-10. • [Hadoop] Chapter 3	HW1 due	
5 (9/19)	• File Format • XML & XPath	• [GVW] Sec. 11.1-3, 12.1	Project proposal due HW2 out	Lab 2: HDFS
6 (9/26)	• Data Modeling (ER & relational) • Midterm 1 (in-class)	• [GUW] Sec. 4.1-4.6, 2.1-2.1	HW2 due	
7 (10/3)	• SQL	• [GUW] Sec. 2.3, 6.1-6.5 • [SQL] Chapters 3-6		
8 (10/10)	• SQL	• [GUW] Sec. 2.3, 6.1-6.5 • [SQL] Chapters 7-10	HW3 out	
9 (10/17)	• Constraints & views • Data organization & external sorting	• [GUW] Sec. 7.1-7.2, 8,1, 8.3 • [GUW] Sec. 13.5, 13.7	HW3 due	
10 (10/24)	• NoSQL 2: MongoDB Midterm 2 (in class)	[MongoDB] Parts 1 and 2		
11 (10/31)	• Indexing (B+-tree) • Query execution	• [GUW] Sec. 14.1-14.2 • [GUW] Chapter 15	HW4 out Project midterm report due	Lab 3: MongoDB
12 (11/7)	• Query execution NoSQL 3: DynamoDB	G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available key-value store," in SOSP, 2007, pp. 205-220.	HW4 due	
13 (11/14)	• Hadoop MapReduce	• J. Dean and S. Ghemawat, MapReduce: simplified	HW5 out	Lab 4: DynamoDB

		data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008. • F. Chang, J. Dean, S. Ghemwat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems (TOCS), vol. 26, p. 4, 2008. • R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011. • [Hadoop] Chapter 2		
14 (11/21)	• Apache Spark	• Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012. • Zaharia, Matei and Chowdhury, Mosharaf and Franklin, Michael J. and Shenker, Scott and Stoica, Ion. Spark: cluster computing with working sets. HotCloud, 2010.	HW5 due	
15 (11/28)	• Project demo		Project final report & demo video due	
Final exam	• Morning section: Thursday, December 8, 11am-1pm • Afternoon section: Tuesday, December 13, 2-4pm			

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, “Behavior Violating University Standards” policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL), press “0” after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) - (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Campus Support and Intervention - (213) 821-4710
campussupport.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost’s Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu

Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
[dps.usc.edu](dps.usc.edu)
Non-emergency assistance or information.