USC Viterbi School of Engineering

DSCI 351: Foundations of Data Management
Units: 4

Term—Day—Time:
Fall 2023 (section 32402D**) – MW – 10-11:50am**
Location: OHE230

Please visit https://blackboard.usc.edu/webapps/login/ for course contents

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

TA: To be announced on course website
Office Hours: TBA
Contact Info: TBA

A. Course Description

Catalog:

Data management course focused on data modeling, data storage, indexing, relational databases, key-value/document store, NoSQL, distributed file system, parallel computation, and big-data analytics.

Extended:

This course provides students with the fundamental knowledge and key skills for managing large-scale diverse data. After taking DSCI 351, students will have solid knowledge of data modeling, data formats, and query languages; basic understanding of relational and NoSQL databases; and exposure to systems and techniques for managing and analyzing large-scale data.

Major topics in DSCI 351 are as follows: (a) Fundamentals of data management: conceptual data modeling, relational data model, and JSON; data storage, data organization, indexing, and relational databases; structured query languages such as SQL. (b) Management of non-relational data: document stores such as MongoDB, and row stores such as Amazon DynamoDB. (c) MapReduce parallel computation framework, and big data platform & software such as Amazon EC2, Apache Hadoop, and Spark.

B. Prerequisites: DSCI 250: Introduction to Data Science; ITP 115: Programming in Python

C. 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 finish the readings prior to the discussion in class. All of the course materials, including the readings, lecture slides, homework, and programming assignments will be posted on the course website.

D. 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 do the weekly homework assignments.

E. Recommended Readings and Supplementary Materials

- [**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. Students are responsible for all reading assignments.

F. Course Structure

Homework Assignments
There will be 6 homework/programing assignments on major topics of the course. Assignments must be completed independently. Each assignment is typically graded on a scale of 0-100 and grading rubric for each assignment will be provided.

Exams: There will be two midterm exams and a final exam. Closed-notes and book. The final exam will be accumulative, but focus on materials after the 2nd midterm.

Class Participation: Students are expected to come to class and participate in the class discussions. There will also be online forums (Piazza) created to facilitate out-of-class discussions of class materials.

Project: Students are also expected to complete a course project on managing data for data science. 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.

Grading Scheme:

Homework	30%
Midterm 1	15%
Midterm 2	15%
Comprehensive exam	25%
Project	15%

_____

Total 100%

Grades will range from A through F. The following is the breakdown for grading:

[**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). **Note the cut-off for A is 94**. 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).

Assignment Submission 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. **No late submissions will be accepted**. You are responsible for making sure you have stable network connection for the submission.

Makeup for exams and extension of coursework deadlines may be considered only when there are documented medical emergencies. Doctor notes are needed as proof. Any requests after the exam time and homework deadline will not be considered. Two-week in advance notices are required for scheduling a makeup of exam. 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.

Regrading requests must be made (by emailing TAs or following the instructions posted by TAs) within one week after the solutions or grades have been posted. Grades are final after the regrading period.

G. Course Schedule: A Weekly Breakdown (tentative, may be revised as the course progresses)

Week	Topic	Readings	Hands-on	Homework & project
1 (8/21)	• Introduction • Amazon EC2			
2 (8/28)	• Data modeling I: ER & design principles	• [GUW] Sec. 4.1-4.6	• Set up EC2 instance	
3 (9/4)	• No class on Monday, 9/4, Labor day • Data modeling II: Relational	• [GUW] Sec. 2		HW1 out
4	• SQL I: single-relation	• [GUW] Sec. 2.3, 6.1-6.5	• Install & run	HW1 in

(9/11)	query	• [SQL] Chapters 3-6	MySQL on EC2	
5 (9/18)	• SQL II: join, natural join, theta join, outer join	• [SQL] Chapters 7-10		HW2 out
6 (9/25)	• SQL III: subquery, aggregation			HW2 in
7 (10/2)	• Constraints & Views • Midterm 1 on 10/4, Wednesday, in class	[GUW] Sec. 7.1-7.2, 8,1		HW3 out Project proposal due
8 (10/9)	• XML/XPath			HW3 in
9 (10/16)	• NoSQL1: MongoDB & JSON	• R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011. • [MongoDB] Parts 1 and 2	• Install & run MongoDB on EC2	HW4 out
10 (10/23)	• NoSQL2: Amazon DynamoDB & row store	• 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-	DynamoDB: setup and querying	HW4 in
11 (10/30)	• BigData 1: Hadoop MapReduce • Midterm 2 on 11/1, Wednesday, in class	J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008.	• Install & run Hadoop on EC2	HW5 out Project progress report due
12 (11/6)	• BigData 1: Hadoop MapReduce			HW5 in
13 (11/13)	• Big data2: Apache Spark	• Zaharia, Matei and Chowdhury, Mosharaf and Franklin, Michael J. and Shenker, Scott and Stoica, Ion. Spark: cluster computing with working sets. HotCloud, 2010. • Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI,	• Install & run Spark on EC2	HW6 out

		2012.		
14 (11/20)	• Big data2: Apache Spark			HW6 in
15 (11/27)	• Project in-class demo • Comprehensive exam, in-class			
Final week	• Project final report due			

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
Non-emergency assistance or information.