DSCI 351: Foundations of Data Management
Units: 4

Term—Day—Time:
Spring 2024 (section 32454R) – MW – 8-9:50am
Location: WPH B27
Spring 2024 (section 32400R) – MW – 10-11:50am
Location: DMC 151

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

   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/programming 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:

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework</td>
<td>30%</td>
</tr>
<tr>
<td>Midterm 1</td>
<td>15%</td>
</tr>
<tr>
<td>Midterm 2</td>
<td>15%</td>
</tr>
<tr>
<td>Comprehensive exam</td>
<td>25%</td>
</tr>
</tbody>
</table>
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)**

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Readings</th>
<th>Hands-on</th>
<th>Homework &amp; project</th>
</tr>
</thead>
</table>
| 1 (1/8) | • Introduction  
          • Amazon EC2 | | | |
| 2 (1/15) | • Data modeling I: ER  
            & design principles  
            • No class on Monday,  
            1/15, MLK Birthday | [GUW] Sec. 4.1-4.6 | • Set up EC2 instance |
| 3 (1/22) | • Data modeling II:  
            Relational | [GUW] Sec. 2 | | HW1 out |
<table>
<thead>
<tr>
<th>Week</th>
<th>Details</th>
</tr>
</thead>
</table>
| 4 1/29 | • SQL I: single-relation query  
• [GUW] Sec. 2.3, 6.1-6.5  
• [SQL] Chapters 3-6  
• Install & run MySQL on EC2  
| HW1 in  |
| 5 2/5 | • SQL II: join, natural join, theta join, outer join  
• [SQL] Chapters 7-10  
| HW2 out  |
| 6 2/12 | • SQL III: subquery, aggregation  
• Constraints & Views  
| HW2 in  |
| 7 2/19 | • No class on President’s Day, 2/19  
• Midterm 1 on 2/21, Wednesday, in class  
• [GUW] Sec. 7.1-7.2, 8,1  
| HW3 out  
Project proposal due |
| 8 2/26 | • XML/XPath  
| HW3 in  |
| 9 3/4 | • NoSQL1: MongoDB & JSON  
• [MongoDB] Parts 1 and 2  
• Install & run MongoDB on EC2  
| HW4 out  |
| 10 3/11 | Spring Recess  
|  |  |
| 11 3/18 | • 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  |
| 12 3/25 | • BigData 1: Hadoop MapReduce  
• Midterm 2 on 3/27, Wednesday, in class  
• Install & run Hadoop on EC2  
| HW5 out  
Project progress report due |
| 13 4/1 | • BigData 1: Hadoop MapReduce  
| HW5 in  |
| 14 4/8 | • Big data2: Apache Spark  
• Resilient Distributed  
• Install & run Spark on EC2  
| HW6 out  |
**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](http://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](http://policy.usc.edu/scientific-misconduct).

**Support Systems:**

*Counseling and Mental Health* - (213) 740-9355 – 24/7 on call
[studenthealth.usc.edu/counseling](http://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](http://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](http://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](http://equity.usc.edu), [titleix.usc.edu](http://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
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.