USC Viterbi School
of Engineering

INF 351: Foundations of Data Management
Units: 4

Term—Day—Time:
Spring 2020 (section 32452R) – MW – 10-11:50am
Location: SGM 601

Instructor: Wensheng Wu
Office: GER 204
Office Hours: right after class or 9-9:45 am MW by appointment
Contact Info: wenshenw@usc.edu

TA: XX
Office: TBD
Office Hours: TBD
Contact Info:

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 provide students with the fundamental knowledge and key skills for managing large-scale diverse data. After taking INF 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 INF 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 Google Firebase, MongoDB, and row stores such as Amazon Dynamo. (c) Systems and techniques for managing and analyzing large-scale data: distributed file system such as HDFS, MapReduce parallel computation framework, and big data software such as Apache Hadoop and Spark.

B. **Prerequisites**: INF 250: Introduction to Data Informatics; 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 online.

D. Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in a language such as Python or Java. 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. Required 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
- [**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/

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 12 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 a midterm exam and a final exam. Closed-notes and book. The final exam will cover the materials after the midterm.

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

Grading Scheme:

Homework	40%
Midterm	30%
Final	30%
Total	100%

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

[93, 100] = A
[90, 93) = A-
[87, 90) = B+
[83, 87) = B
[80, 83) = B-
[77, 80) = C+
[73, 76) = C
[70, 73) = C-
[67, 70) = D+
[63, 67) = D
[60, 63) = D-
Below 60 is an F

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. Late homework will be deducted 10% of its points for every 24 hours that it is late. No credit will be given after 72 hours of its due time.

Makeups for exams are not permitted unless there are medical emergencies. Doctor notes are needed as proof. Typically 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 due to these situations. Students are required to contact the Student Advocacy Services office (contact information will be provided in class) to submit proper documents for the verification of emergency.

Homework regrading requests must be made within a week after the solutions or grades have been posted. Grades are final after the regrading period. Exam grades are finalized after exam grading review hours (which are typically announced shortly after the exams).

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

Week	Topic	Readings	Homework
1 (1/13)	• Introduction • Amazon EC2		
2 (1/20)	• Data Storage • No class on 1/20 (Martin Luther King's Birthday)	• [AA] Chapter 37	HW1 assigned
3 (1/27)	• Data Storage • File System	• [AA] Chapter 39 • [AA] Chapter 40	HW1 due HW2 assigned
4 (2/3)	• File System • Network File System	• [AA] Chapter 48	HW2 due HW3 assigned
5 (2/10)	• HDFS • NoSQL1: Firebase & JSON	• 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.	HW3 due HW4 assigned
6 (2/17)	• Conceptual data modeling • No class on 2/17 (President's Day)	• [GUW] Sec. 4.1-4.6	HW4 due HW5 assigned

7 (2/24)	• Relational data modeling	• [GUW] Sec. 2	HW5 due HW6 assigned
8 (3/2)	• SQL • Midterm (3/4, Wednesday, in class)	• [GUW] Sec. 2.3, 6.1-6.5	
9 (3/9)	• SQL • Constraints & views	• [GUW] Sec. 7.1-7.2, 8,1	HW6 due HW7 assigned
(3/16)	• Spring recess		
10 (3/23)	• Data organization & external sorting • NoSQL2: MongoDB		HW7 due HW8 assigned
11 (3/30)	• Indexing (B+-tree) • Query execution	• [GUW] Sec. 14.1-14.2 • [GUW] Chapter 15	HW8 due HW9 assigned
12 (4/6)	• Query execution		HW9 due HW10 assigned
13 (4/13)	• NoSQL3: Amazon DynamoDB • Big data1: Hadoop MapReduce	• R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011. • 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- • J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008.	HW10 due HW11 assigned
14 (4/20)	• Big data1: Hadoop MapReduce • 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	HW11 due HW12 assigned

		Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012.	
15 (4/27)	• Big data2: Apache Spark • Final review		HW12 due
Final exam	• Monday, May 11, 8-10am • Same classroom, closed-notes and book		

H. 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 Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* *http://emergency.usc.edu* will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.