USC Viterbi School of Engineering

INF 351: Foundations of Data Management
Units: 4

Term—Day—Time:
Fall 2020 (section 32401R**) – MW – 10-11:50am**
Location: online

Instructor: Wensheng Wu
Office Hours: Tuesday 10am-12pm
Contact Info: wenshenw@usc.edu

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

Note: The course will be online. We will be using Blackboard and Zoom for class meetings and content delivery. Students enrolled in the class will be emailed with instructions before the first class meeting.

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 MongoDB, and row stores such as Amazon DynamoDB. (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. **We will be using Blackboard this semester.**

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. 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

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 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.

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

Grading Scheme:

Homework	30%
Midterm	30%
Final	30%
Project	10%

Total	100%

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

[93, 100] = A	[73, 76) = C
[90, 93) = 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

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	Handon	Homework
1 (8/17)	• Introduction • Amazon EC2			
2 (8/24)	• Data modeling I: ER & design principles	• [GUW] Sec. 4.1-4.6	• Set up EC2 instance	HW1 out
3 (8/31)	• Data modeling II: Relational	• [GUW] Sec. 2		HW1 in
4 (9/7)	• SQL I: single-relation query • No class on 9/7: labor day	• [GUW] Sec. 2.3, 6.1-6.5	• Install & run MySQL on EC2	HW2 out
5 (9/14)	• SQL II: join, subquery, aggregation			HW2 in
6 (9/21)	• Constraints & Views	• [GUW] Sec. 7.1-7.2, 8,1		HW3 out
7 (9/28)	• External sorting • Midterm			HW3 in
8 (10/5)	• Indexing: B+-tree • Query execution: overview • Project proposal due	• [GUW] Sec. 14.1-14.2		HW4 out
9 (10/12)	• NoSQL1: MongoDB & JSON	• R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol.	• Install & run MongoDB on EC2	HW4 in

		39, pp. 12-27, 2011.		
10 (10/19)	• NoSQL2: Amazon DynamoDB & row store • Project progress report due	• 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	HW5 out
11 (10/26)	• BigData 1: Hadoop MapReduce	• 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 in
12 (11/2)	• BigData 1: 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 Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012.	• Install & run Spark on EC2	HW6 out
13 (11/9)	• Big data2: Apache Spark • Final review • Project demo & final report due			HW6 in
Final exam	• Monday, November 23, 8-10am			

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.