USC Viterbi School of Engineering

INF 551: Foundations of Data Management
Units: 4

Term—Day—Time:
Fall 2019
Morning section: MW 10-11:50am (THH 118)
Afternoon section: MW 4-5:50pm (KAP 144)

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

Course producers:
- **TBD**

Office: SAL computing lab
Office Hours: TBD

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 Informatics program. It prepares the students with the fundamental knowledge on the data management. Such a 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 large scale of data, to ensure that students have adequate background for further exploring big data analytics in 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 warehousing, data cleaning, 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 solution stacks, e.g., Apache Hadoop and Spark.

C. Recommended Preparation:

INF 550 taken previously or concurrently. Basic understanding of operating systems, networks, and databases. A basic understanding engineering principles is required, including basic programming skills; familiarity with the Python & 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 of the course materials, including the readings, lecture slides, home works will be posted online

E. 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.

F. Required 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
- [**HKP**] Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2011, 3rd Edition (selected chapters only).

In addition to the textbook, students may be given additional reading materials such as research papers. Students are responsible for all assigned 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.
Weekly quizzes: There will be weekly quizzes, typically based on the lectures in the past week.
Exams: There will be a midterm and a final exam.
Lab sessions: Hand-on exercises on database and big data software.
Course project: Develop a cloud-based Web/mobile database application.

Grade breakdown:

Homework	20%
Weekly quizzes	25%
Midterm	15%
Final	25%
Lab session	5%
Course project	10%
Total	100%

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

[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

Note that [90, 93) means that your score is greater than or equal to 90 but less than 93. Note that every point in your coursework counts. We will strictly follow the above cut-off and NO roundup will be performed. Note that grades are NOT negotiable!

H. Grading 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.

Makeup for quizzes and 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.

Homework and quizzes regrading requests must be made within a week after the solutions or grades have been posted. Grades are final after the regrading period. Final exam grades are finalized 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 (8/26)	• Data Management Overview	• [AA] Chapter 2 (optional) • [AA] Chapter 4 (optional)		
2 (9/2)	• Firebase & JSON No class on 9/2 (Labor day)	• [AA] Chapter 37		
3 (9/9)	• Storage System • File System	• [AA] Chapter 39 • [AA] Chapter 40	Homework 1 assigned	Lab 1: Amazon EC2
4 (9/16)	• Network File System • HDFS • Project proposal presentation	• [AA] Chapter 48 • 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.	Homework 1 due Project proposal due	
5 (9/23)	• File Format • XML, JSON	• [GVW] Sec. 11.1-3, 12.1	Homework 2 out	Lab 2: HDFS

6 (9/30)	• Data Modeling (ER & relational)	• [GUW] Sec. 4.1-4.6, 2.1-2.1	Homework 2 due	
7 (10/7)	• SQL • Midterm (10/9), in-class			
8 (10/14)	• SQL • Fall recess (10/17-10/18)	• [GUW] Sec. 2.3, 6.1-6.5	Homework 3 assigned	
9 (10/21)	• Constraints & views • Data organization & external sorting • MongoDB	• [GUW] Sec. 7.1-7.2, 8,1, 8.3 • [GUW] Sec. 13.5, 13.7	Project midterm report due	Lab 3: MongoDB
10 (10/28)	• Indexing (B+-tree)	[GUW] Sec. 14.1-14.2	Homework 3 due Homework 4 assigned	
11 (11/4)	• Query execution	• [GUW] Chapter 15		
12 (11/11)	• NoSQL & Amazon DynamoDB • Hadoop MapReduce	• J. Dean and S. Ghemawat, MapReduce: simplified 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. • G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly	Homework 4 due Homework 5 assigned	Lab 4: DynamoDB

		available key-value store," in SOSP, 2007, pp. 205-220.		
13 (11/18)	• Hadoop MapReduce • Apache Spark	• Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012.		
14 (11/25)	• Apache Spark • Thanksgiving break (11/27-12/1)	• Zaharia, Matei and Chowdhury, Mosharaf and Franklin, Michael J. and Shenker, Scott and Stoica, Ion. Spark: cluster computing with working sets. HotCloud, 2010.	• Homework 5 due	
15 (12/2)	• Project demo • Final review		Project final report due	
Final exam	• Morning section: 12/16, Monday, 8-10am • Afternoon section: 12/11, Wednesday, 4:30-6:30pm • Same classroom, closed-notes and book			

J. **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.