USC Viterbi School of Engineering

INF 551: Foundations of Data Management
Units: 4

Term—Day—Time:
Fall 2016 – MW 10-11:50am (section 32418D**)**
Location: GFS 101

Fall 2016 – MW 4-5:50pm (section 32431D**)**
Location: THH 208

Instructor: Wensheng Wu
Office: GER 204
Office Hours: MW 3-4pm
Contact Info: wenshenw@usc.edu

TA: xxx
Office: xxx
Office Hours: xxx
Contact Info: xxx

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) Advanced topics in data management: data warehousing, data cleaning, ETL, data integration, and metadata management; (3) Big data analytics: NoSQL, key-value and document stores, cloud data storage, distributed file system, and MapReduce.

The course will also provide students with hand-on experiences on RDBMS, e.g., MySQL, cloud data storage, e.g., Amazon S3/Dynamo, CouchDB, Cassandra, and big data solution stacks, e.g., Apache Hadoop, Pig, 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

- [**GUW**] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. Database Systems: The Complete Book (Second Edition), Prentice Hall, 2009 (selected chapters only). 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).
- [**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/
- **[RLU]** Rajaraman, J. Leskovec and J. D. Ullman, Mining of Massive Datasets. Cambridge University Press, 2012. Available free at: http://infolab.stanford.edu/~ullman/mmds.html

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 Structure

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.
Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class.
Student presentation: Students are expected to present research or application papers on subjects related to class materials.

Grade breakdown:

Homework	30%
Weekly quizzes	30%
Final	30%
Student presentation	10%
Total	100%

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

94 - 100 = A	74 - 76 = C
90 – 93 = A-	70 - 73 = C-
87 – 89 = B+	67 - 69 = D+
84 – 86 = B	64 - 66 = D
80 – 83 = B-	60 - 63 = D-
77 – 79 =C+	Below 60 is an F

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, quizzes, and midterm exam regrading requests must be made within a week after the solutions have been posted. Grades are final after the regrading period.

I. Important Administrative Dates

Classes Begin	Mon	August 22
Labor Day	Mon	September 5
Thanksgiving	Wed-Sun	November 23-27
Classes End	Fri	December 2
Study Days	Sat-Tue	December 3-6
Exams	Wed-Wed	December 7-14
Winter Recess	Thu-Sun	December 15-January 8

J. Course Schedule: A Weekly Breakdown

Week	Topic	Readings	Homework
1 (8/22)	• Data Management Overview		
2 (8/29)	• Storage System • Disk scheduling	[AA] Chapter 37	
3 (9/5)	• RAID (no class on 9/5)	[AA] Chapter 38	Homework 1 assigned
4 (9/12)	• File System • Network File System	[AA] Chapters 39, 40, 48	Homework 1 due
5	• File Format	Amazon S3	Homework 2

(9/19)	• Cloud data storage • XML, JSON		assigned
6 (9/26)	• Data Modeling (ER & relational)	[GUW] Sec. 4.1-4.6, 2.1-2.1	
7 (10/3)	• Relational Algebra • SQL	[GUW] Sec. 2.4, Sec. 5.1-5.2 [GUW] Sec. 2.3, 6.1-6.5	Homework 2 due
8 (10/10)	• Data organization • Indexing	[GUW] Sec. 14.1-14.6	
9 (10/17)	• Query execution • Query optimization	[GUW] Chapter 15	
10 (10/24)	• Data Warehousing	[HKP] Chapter 1	Homework 3 assigned
11 (10/31)	• OLAP • Cube computation	[HKP] Chapter 3	Homework 3 due
12 (11/7)	• NoSQL • Apache CouchDB • Apache Cassandra	• R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011. • Lakshman and P. Malik, Cassandra: a decentralized structured storage system," ACM SIGOPS Operating Systems Review, vol. 44, pp. 35-40, 2010. • 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. 205-220. • F. Chang, J. Dean, S. Ghemwat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A	Homework 4 assigned

		distributed storage system for structured data," ACM Transactions on Computer Systems (TOCS), vol. 26, p. 4, 2008.	
13 (11/14)	• In-memory cluster computing • Apache Spark	• Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012. • Zaharia, Matei and Chowdhury, Mosharaf and Franklin, Michael J. and Shenker, Scott and Stoica, Ion. Spark: cluster computing with working sets. HotCloud, 2010.	Homework 4 due
14 (11/21)	• Hadoop & MapReduce • Large-scale ETL and data warehousing • Apache Pig (no class on 11/23)	• J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008. • 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. • Pig Latin: A Not-So-Foreign Language for Data Processing, Christopher Olston, et. al., SIGMOD 2008. • Matrix multiplication ([RLU] Section 2.3.10) • HITS algorithm ([RLU] Section 5.5)	Homework 5 assigned
15 (11/28)	• Large-scale stream data processing • Apache Spark (stream processing) • NoSQL 2: Apache HBase, MongoDB	• Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters. Zaharia, et.al. USENIX HotCloud, 2012. • Hive – A Petabyte Scale Data	Homework 5 due

	• Apache Hive: Warehousing on Hadoop • Wrap-up & review	Warehouse Using Hadoop. Thusoo et. al., ICDE 2010.	
Final exam	To be announced		

K. **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.