USC Viterbi School of Engineering

INF 551: Foundations of Data Management
Units: 3

Term—Day—Time:
Spring 2016 – MW – 3:30-4:50 pm (section 32405D)
Location: SLH100

Instructor: Seon Ho Kim
Office: PHE 304
Office Hours: 2:00 pm - 3:00 pm MW; or by appointment
Contact Info: seonkim@usc.edu, 213-740-2483

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

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.

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, SimpleDB, and Dynamo, and big data solution stacks, e.g., Apache Hadoop and Spark.

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.

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

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.

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/

In addition to the textbook, students may be given additional reading materials such as research papers. Students are responsible for all assigned reading assignments.

Description and Assessment of Assignments

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.

Grading Breakdown
Homework: There will be 5 homework based on the topics of the class each week.
Midterm Exam: There will be a midterm exam, usually in the 7-th or 8-th week of the semester.
Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class.
Class Participation: Students are expected to come to class and participate in the class discussions and discussion board.

Grading Schema:

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

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. NO late homework will be accepted. Lowest homework grade will be dropped.

Course Schedule: A Weekly Breakdown (Tentative, may revise when the course progresses)

Week	Topic	Readings	Homework
1 (8/24)	Data Management Overview		
2 (8/31)	Storage System	[AA] Chapter 37	
3 (9/7)	RAID	[AA] Chapter 38	Homework 1 assigned
4 (9/14)	File System & Network File System	[AA] Chapters 39, 40, 48	Homework 1 due
5 (9/21)	File Format, Cloud data storage, XML, JSON	Amazon S3	Homework 2 assigned
6 (9/28)	Data Modeling (ER & relational)	[GUW] Sec. 4.1-4.6, 2.1-2.1	Homework 2 due
7 (10/5)	Relational Algebra SQL	[GUW] Sec. 2.4, Sec. 5.1-5.2 [GUW] Sec. 2.3, 6.1-6.5	Midterm Exam
8 (10/12)	Data Warehousing	[HKP] Chapter 1	
9 (10/19)	OLAP	[HKP] Chapter 3	Homework 3 assigned
10 (10/26)	NoSQL	R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011. 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.	Homework 3 due
11 (11/2)	NoSQL	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,"	Homework 4 assigned

		in SOSP, 2007, pp. 205-220.	
12 (11/9)	Hadoop & MapReduce	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.	Homework 4 due
13 (11/16)	Hadoop & MapReduce (Spark)	D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-based highperformance Data Mining of large Data on MapReduce Clusters," in Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on, 2009, pp. 296-301. 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 assigned
14 (11/23)	Metadata Management	Dublin Core, RDF	Homework 5 due
15 (11/30)	Advanced topics	Data integration	
Final (12/14)	Final Exam		

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.