USC Viterbi School of Engineering

**INF 559: Introduction to Data Management**
**Units: 3**

**Term—Day—Time:**
**Spring 2017 – MW – 5-6:20pm**
**Location: KDC 235**

**Instructor: Wensheng Wu**
**Office:** GER 204
**Office Hours:** TBD
**Contact Info:** wenshenw@usc.edu

**TA:** TBD
**Office:** SAL computing lab
**Office Hours: TBD**
**Contact Info: 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 introductory courses in the Informatics program. It prepares the students with the fundamental knowledge on the data management. Such 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, and query processing. (2) Big data analytics: NoSQL, key-value and document stores, cloud data storage, distributed file system, and MapReduce; and (3) Advanced topics in data management: data warehousing, data cleaning, and data integration.

The course will also provide students with hand-on experiences on RDBMS, e.g., MySQL, cloud data storage, e.g., Amazon S3, NoSQL databases suchas Amazon DynamoDB, Apache Cassandra, and big data solution stacks, e.g., Apache Hadoop, Hive, and Spark.

**C. Recommended Preparation**:

Basic understanding of engineering principles, including basic programming skills, knowledge of operating systems, networks, and databases; familiarity with the Python (preferably Java too) programming language is desired.

**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**

Familiarity with Python and Java programming languages is highly desired.
Students are also expected to have their own laptop or desktop computer where they can install and run software for the homework assignments and labs.

**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). 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 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.
**Lab sessions:** Hand-on exercises on database and big data software.
**Student presentation:** Students may also be expected to present research or application papers on subjects related to class materials.

| Grade breakdown: | |
|---|---|
| Homework | 30% |
| Quizzes | 30% |
| Final exam | 30% |
| Labs and 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. Course Schedule: A Weekly Breakdown (may be revised when the course progresses)**

| Week | Topic | Readings | Homework | Lab |
|---|---|---|---|---|
| 1 (1/9) | • Data Management Overview<br>• Computer system review | | | Lab 1: Amazon EC2 |
| 2 (1/16) | • Storage System<br>(no class on 1/16, University Holiday) | [AA] Chapter 37 | | |
| 3 (1/23) | • Storage System<br>• File System | [AA] Chapters 39, 40, 48 | Homework 1 assigned | |
| 4 (1/30) | • Network File System<br>• HDFS | • 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. | | Lab 2: File system & HDFS |
| 5 (2/6) | • File Format<br>• XML, JSON | | Homework 1 due<br>Homework 2 assigned | |
| 6 (2/13) | • Data Modeling (ER & relational) | [GUW] Sec. 4.1-4.6, 2.1-2.1 | | |
| 7 (2/20) | • Relational Algebra<br>(No class on 2/20, | [GUW] Sec. 2.4, Sec. 5.1-5.2 | Homework 2 due<br>Homework 3 | Lab 3: XML & XPath |

| | | | | |
|---|---|---|---|---|
| | University Holiday) | | assigned | |
| 8 (2/27) | • SQL | [GUW] Sec. 2.3, 6.1-6.5 | | |
| 9 (3/6) | • Data organization & external sorting<br>• Indexing (B+-tree) | [GUW] Sec. 14.1-14.6 | Homework 3 due<br>Homework 4 assigned | |
| 3/13-3/17 | Spring recess | | | |
| 10 (3/20) | • Query execution | [GUW] Chapter 15 | | |
| 11 (3/27) | • Cloud data storage<br>• NoSQL databases<br>• Amazon DynamoDB<br>• Eventual consistency | • R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011.<br>• 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. | | Lab 4: Amazon S3 & DynamoDB |
| 12 (4/3) | • Apache Hadoop<br>• MapReduce framework | • J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008. [copy and paste this link to your browser, note the double slashes //archive: http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf] | Homework 4 due<br>Homework 5 assigned | |
| 13 (4/10) | • Apache Spark | • Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012.<br>• Zaharia, Matei and Chowdhury, Mosharaf and Franklin, Michael J. and | | |

| | | | | |
|---|---|---|---|---|
| | | Shenker, Scott and Stoica, Ion. Spark: cluster computing with working sets. HotCloud, 2010. | | |
| 14 (4/17) | • Apache Cassandra<br>• MongoDB | • Lakshman and P. Malik, Cassandra: a decentralized structured storage system," ACM SIGOPS Operating Systems Review, vol. 44, pp. 35-40, 2010.<br>• 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. | | Lab 5: MongoDB |
| 15 (4/24) | • Data warehousing<br>• Apache Hive<br>• Final review | • [HKP] Chapter 1<br>• Hive – A Petabyte Scale Data Warehouse Using Hadoop. Thusoo et. al., ICDE 2010. | Homework 5 due | |
| Final exam | • TBD | | | |

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.