**USC VITERBI SCHOOL OF ENGINEERING INFORMATICS PROGRAM**

# INF 551: Overview of Data Informatics in Large Data Environments

**Section: 32405D**

**Spring 2017 (4 units), 3:30 – 5:20 PM, MW, SOS B44**

*Syllabus*


Instructor: Dr. Seon Ho Kim
Email: seonkim@usc.edu
Office: PHE304
Phone: 213.740.2483

Assistant: TBD
Email:


**Instructor's Office Hours:**
Thursday 3:00 p.m. to 5:00 p.m. in PHE304. Other hours by appointment only. Students are advised to make appointments with the professor ahead of time in any event and be specific with the subject matter to be discussed. Students should also be prepared for their appointment by bringing all applicable materials and information.

**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.

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.

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.

**Books and Readings:**

***Recommended Books (some selected chapters):***

[**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.

***Lecture notes* will be available in the Class Blackboard.**

**Grading:**

Exams: 50%, Homework 40%, Quiz: 10%

** The instructor reserves the right to make changes to the grading criteria, assignments and course outline to best serve the needs of the class.*

**Course Schedule**

| Week | Topic | Readings | Homework |
|---|---|---|---|
| 1 (1/9, 1/11) | • Data Management Overview | | |
| 2 (1/18) | Storage System<br>Disk scheduling<br>(no class on 1/16, University Holiday) | [AA] Chapter 37 | |
| 3 (1/23, 1/25) | • RAID | [AA] Chapter 38 | |
| 4 (1/30, 2/1) | • File System<br>• Network File System | [AA] Chapters 39, 40, 48 | |
| 5 (2/6, 2/8) | • File Format<br>• Cloud data storage<br>• XML, JSON | Amazon S3 | |
| 6 (2/13, 2/15) | • Data Modeling (ER & relational) Review<br>• SQL | [GUW] Sec. 4.1-4.6, 2.1-2.1<br>[GUW] Sec. 2.4, Sec. 5.1-5.2 [GUW] Sec. 2.3, 6.1-6.5 | |

| | | | |
|---|---|---|---|
| | | | |
| 7 (2/22) | • Data organization<br>• Indexing<br>(No class on 2/20, University Holiday) | [GUW] Sec. 14.1-14.6 | |
| 8 (2/27, 3/1) | | | **Midterm** |
| 9 (3/6, 3/8) | • Query execution & Optimization<br>• Data Warehousing | [GUW] Chapter 15<br>[HKP] Chapter 1 | |
| 3/13-3/17 | Spring Break | | |
| 10 (3/20, 3/22) | • OLAP<br>• Cube computation | [HKP] Chapter 3 | |
| 11 (3/27, 3/29) | • NoSQL | • 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. | |
| 12 (4/3, 4/5) | • Apache CouchDB<br>• Apache Cassandra<br>• Amazon DynamoDB | • R. Cattell, "Scalable SQL and NoSQL data stores," ACM SIGMOD Record, vol. 39, pp. 12-27, 2011.<br>• Lakshman and P. Malik, Cassandra: a decentralized structured storage system," ACM SIGOPS Operating Systems Review, vol. 44, pp. 35-40, 2010.<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. | |
| 13 (4/10, 4/12) | • In-memory cluster computing<br>• 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, 4/19) | • Hadoop & MapReduce<br>• Large-scale ETL and data warehousing<br>• Apache Pig | • 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]<br>• 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.<br>• Pig Latin: A Not-So-Foreign Language for Data Processing, Christopher Olston, et. al., SIGMOD 2008.<br>• Matrix multiplication ([RLU] Section 2.3.10)<br>• HITS algorithm ([RLU] Section 5.5) | |
| 15 (4/24, 4/26) | • Large-scale stream data processing<br>• Apache Spark (stream processing)<br>• NoSQL 2: Apache HBase, MongoDB<br>• Apache Hive<br>• Wrap-up & review | • Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters. Zaharia, et.al. USENIX HotCloud, 2012.<br>• Hive – A Petabyte Scale Data Warehouse Using Hadoop. Thusoo et. al., ICDE 2010. | |
| **16** | • TBD | | **Final exam** |

**Students with Disabilities:**

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me as early in the semester as possible. Your letter must be specific as to the nature of any accommodations granted. DSP is located in STU 301 and is open 8:30 am to 5:30 pm, Monday through Friday. The telephone number for DSP is (213) 740-0776.

**Academic Integrity:**

USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. SCampus, the Student Guidebook, (www.usc.edu/scampus or http://scampus.usc.edu) contains the University Student Conduct Code (see University Governance, Section 11.00), while the recommended sanctions are located in Appendix A.

The University, as an instrument of learning, is predicated on the existence of an environment of integrity. As members of the academic community, faculty, students, and administrative officials share the responsibility for maintaining this environment. Faculties have the primary responsibility for establishing and maintaining an atmosphere and attitude of academic integrity such that the enterprise may flourish in an open and honest way. Students share this responsibility for maintaining standards of academic performance and classroom behavior conducive to the learning process. Administrative officials are responsible for the establishment and maintenance of procedures to support and enforce those academic standards. Thus, the entire University community bears the responsibility for maintaining an environment of integrity and for taking appropriate action to sanction individuals involved in any violation. When there is a clear indication that such individuals are unwilling or unable to support these standards, they should not be allowed to remain in the University."
(http://policies.usc.edu/p4acad_stud/facultyhandbook.pdf)

Academic dishonesty includes: (http://policies.usc.edu/p4acad_stud/facultyhandbook.pdf)

- Examination behavior – any use of external assistance during an examination shall be considered academically dishonest unless expressly permitted by the teacher.

- Fabrication – any intentional falsification or invention of data or citation in an academic exercise will be considered a violation of academic integrity.

- Plagiarism – the appropriation and subsequent passing off of another's ideas or words as one's own. If the words or ideas of another are used, acknowledgment of the original source must be made through recognized referencing practices.

- Other Types of Academic Dishonesty – submitting a paper written by or obtained from another, using a paper or essay in more than one class without the teacher's express permission, obtaining a copy of an examination in advance without the knowledge and consent of the teacher, changing academic records outside of normal procedures and/or petitions, using another person to complete homework assignments or take-home exams without the knowledge or consent of the teacher.

The use of unauthorized material, communication with fellow students for course assignments, or during a mid-term examination, attempting to benefit from work of another student, past or present and similar behavior that defeats the intent of an assignment or mid-term examination, is unacceptable to the University. It is often difficult to distinguish between a culpable act and inadvertent behavior resulting from the nervous tensions accompanying examinations. Where a clear violation has occurred, however, the instructor may disqualify the student's work as unacceptable and assign a failing mark on the paper.

**Return of Course Assignments:**

Returned paperwork, unclaimed by a student, will be discarded after a year and hence, will not be available should a grade appeal be pursued following receipt of his/her grade.

**Emergency Preparedness/Course Continuity in a Crisis:**

In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies.