Course Information

- MSIS 2627: Big Data
- Class hours:
  - Tuesday, 5:45pm - 7:00pm PST
  - Thursday, 5:45pm - 7:00pm PST
- Class room: Lucas Hall 310
- Adjunct Lecturer: Mahmoud Parsian
  Ph.D in Computer Science
  LinkedIn profile: http://www.linkedin.com/in/mahmoudparsian
- Email: `mparsian@yahoo.com`
- GitHub: https://github.com/mahmoudparsian/big-data-mapreduce-course
- Office Location: Lucas Hall 221X
- Office Hours: TBDL

Course Description:
http://www.scu.edu/business/msis/program/msis-specialization.cfm#faq8

Texts and Papers:

- Big Data Now [book]
  http://www.oreilly.com/data/free/files/bigdatanow2013.pdf
- Data-Intensive Text Processing with MapReduce by Jimmy Lin and Chris Dyer [book]
  http://lintool.github.io/MapReduceAlgorithms/ed1n/MapReduce-algorithms.pdf

- Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman [book]
  http://infolab.stanford.edu/~ullman/mmds/book.pdf
- PAPER-1: MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat
  http://lintool.github.io/MapReduce-course-2013s/material/Dean_Ghemawat_OSDI2004.pdf
- PAPER-2: Ullman. (2012) Designing Good Mapreduce Algorithms
  http://lintool.github.io/MapReduce-course-2013s/material/Ullman_2012.pdf
- PAPER-3: Large Language Models in Machine Translation, Thorsten Brants et. al.
  https://www.aclweb.org/anthology/D/D07/D07-1090.pdf
- PAPER-4: Bigtable: A Distributed Storage System for Structured Data
  http://lintool.github.io/MapReduce-course-2013s/material/ChangFay_etal_OSDI2006.pdf

Course Objectives:
At the completion of this course, students will be able to understand:

1. Elements of Big Data
2. Introduction to MapReduce, DAG
3. MapReduce algorithms and some design patterns
4. NoSQL Databases
5. Fundamentals of Hadoop framework
6. Fundamentals of Spark framework
7. Scale out vs. Scale up
8. Basics of Indexing (Solr/Lucene/ElasticSearch)

Grade Distribution:

Assignment-1	10%
Assignment-2	10%
Assignment-3	10%
Assignment-4	10%
Assignment-5	10%
Midterm Exam	20%
Final Exam	30%

Letter Grade Distribution:

A	4.0
A-	3.7
B+	3.3
B	3.0
B-	2.7
C+	2.3
C	2.0
C-	1.7
F	0.0

Course Policies:

- **General**
  - Quizzes and exams are closed book, closed notes.
  - **No makeup quizzes or exams will be given.**
- **Grades**
  - Grades in the **C** range represent performance that **meets expectations**; Grades in the **B** range represent performance that is **substantially better** than the expectations; Grades in the **A** range represent work that is **excellent**.
- **Labs and Assignments**
  - Students are expected to work independently. **Offering** and **accepting** solutions from others is an act of **plagiarism**, which is a serious offense and **all involved parties will be penalized according to the Academic Honesty Policy**. Discussion amongst students is encouraged, but when in doubt, direct your questions to the professor.
  - **No late assignments will be accepted under any circumstances**.
- **Attendance and Absences**
  - Attendance is expected.
  - Students are responsible for all missed work, regardless of the reason for absence. It is also the absentee's responsibility to get all missing notes or materials.
- **SANTA CLARA UNIVERSITY Academic Integrity Protocol:**
  http://www.scu.edu/provost/policy/academicpolicy/upload/revised-Academic-Integrity-protocol-6-17-12.pdf

- **Instructor's Intended Purpose**

  The student's work must match the instructor's intended purpose for an assignment. While the instructor will establish the intent of an assignment, each student must clarify outstanding questions of that intent for a given assignment.

- **Unauthorized/Excessive Assistance**

  The student may not give or get any unauthorized or excessive assistance in the preparation of any work.

- **Authorship**

  The student must clearly establish authorship of a work. Referenced work must be clearly documented, cited, and attributed, regardless of media or distribution.

Tentative Course Outline:
The weekly coverage might change as it depends on the progress of the class. However, you must keep up with the reading assignments.

Session	Date	Content
1	January 5	• **Introduction to Big Data** • Pages 1-40 of Big Data Now • Chapter 1 of Data-Intensive Text Processing with MapReduce • A Very Brief Introduction to MapReduce http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf • Introduction to MapReduce http://lsd.ls.fi.upm.es/lsd/nuevas-tendencias-en-sistemas-distribuidos/IntroToMapReduce_2.pdf
2	January 7	• **Introduction to MapReduce** • Pages 41-80 of Big Data Now • Chapters 2 of Data-Intensive Text Processing with MapReduce • Introduction to MapReduce and Hadoop by Matei Zaharia http://www.cs.berkeley.edu/~matei/talks/2010/amp_mapreduce.pdf
3	January 12	• **Introduction to MapReduce** • Pages 81-120 of Big Data Now • Chapters 3 of Data-Intensive Text Processing with MapReduce • Chapters 2 of Mining of Massive Datasets
4	January 14	• **Introduction to MapReduce/Hadoop** • Pages 121-187 of Big Data Now • Chapters 4 of Data-Intensive Text Processing with MapReduce • Introduction to Hadoop
5	January 19	• **Practice MapReduce/Spark/Hadoop** • Making Big Data Simple: by Matei Zaharia: https://www.youtube.com/watch?v=Nev1s6fHwMI • Introduction to Spark: http://www.slideshare.net/jeykottalam/spark-sqlamp-camp2014?qid=a9e4b467-26b7-49a9-a361-8425876e9904&v=qf1&b=&from_search=4
6	January 21	• **Spark/Hadoop Nuts and Bolts** • Making Big Data Simple: by Matei Zaharia: https://www.youtube.com/watch?v=Nev1s6fHwMI • Introduction to Spark: http://www.slideshare.net/jeykottalam/spark-sqlamp-camp2014?qid=a9e4b467-26b7-49a9-a361-8425876e9904&v=qf1&b=&from_search=4 • Parallel-Programming-With-Spark-Matei-Zaharia: http://ampcamp.berkeley.edu/wp-content/uploads/2013/02/Parallel-Programming-With-Spark-Matei-Zaharia-Strata-20 pptx
7	January 26	• **MapReduce Algorithm Design** • Chapter 3 of Data-Intensive Text Processing with MapReduce

8	January 28	• **Language Models and Inverted Indexing** • Chapter 4 of http://lintool.github.io/MapReduceAlgorithms/ed1n/MapReduce-algorithms.pdf • https://www.aclweb.org/anthology/D/D07/D07-1090.pdf
9	February 2	• **Midterm Exam** • Exams are closed book, closed notes, closed computers • No makeup quizzes or exams will be given
10	February 4	• **Similar Item Detection** • **Clustering and Classification**
11	February 9	• **Introduction to Search-1: Solr/Lucene**
12	February 11	• **Introduction to Search-2: Solr/Lucene**
13	February 16	• **NoSQL, MongoDB, Redis** • Chapter 4 of BOOK-2 • PAPER-4
14	February 18	• **NoSQL, MongoDB, Redis**
15	February 23	• **Frequent Itemsets** • **Implementing Recommendation Systems using Big Data** • Chapter 6 of BOOK-3
16	February 25	• **Redis Server role for Big Data** • Chapter 6 of BOOK-3
17	March 1	• **Relational Algebra and MapReduce** • Chapter 6 of BOOK-2 • Relational Algebra and MapReduce: http://www.eurecom.fr/~michiard/teaching/slides/clouds/tutorial-high_level.pdf • MapReduce examples: http://courses.cs.washington.edu/courses/cse344/11sp/sections/section8/section8-mapreduce-solution.pdf • MapReduce and relational algebra: http://www.cs.kent.edu/~jin/Cloud12Spring/DatabaseMapReduce.pptx
18	March 3	• **Regression Algorithms** • Linear Regression • Ttest
19	March 8	• **Overview of Big Data and MapReduce**
	March 10	• **Classes End**
	March 14-18	• **Final Examinations**