



## Course Information

- MSIS 2696: Big Data
- Class hours: Tuesday & Thursday 5:45pm - 7:00pm
- Class room: Lucas Hall 210
- Adjunct Lecturer: Mahmoud Parsian  
Ph.D in Computer Science  
LinkedIn profile: <http://www.linkedin.com/in/mahmoudparsian>
- Email: [mparsian@yahoo.com](mailto:mparsian@yahoo.com)
- GitHub: <https://github.com/mahmoudparsian/BigData-MapReduce-Course>
- Office Location: Lucas Hall 210
- Office Hours: TBDL

## Course Description:

<http://www.scu.edu/business/msis/program/msis-specialization.cfm#faq8>

## Texts and Papers:

- BOOK-1: Big Data Now  
<http://www.oreilly.com/data/free/files/bigdatanow2013.pdf>
- BOOK-2: Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer  
<http://lintool.github.io/MapReduceAlgorithms/ed1n/MapReduce-algorithms.pdf>
- BOOK-3: Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman  
<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](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](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](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

### 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 6	<ul style="list-style-type: none"><li>• <b>Introduction to Big Data</b></li><li>• Pages 1-40 of <a href="#">BOOK-1</a></li><li>• Chapter 1 of <a href="#">BOOK-2</a></li><li>• A Very Brief Introduction to MapReduce <a href="http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf">http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf</a></li><li>• Introduction to MapReduce <a href="http://lsd.ls.fi.upm.es/lsd/nuevas-tendencias-en-sistemas-distribuidos/IntroToMapReduce_2.pdf">http://lsd.ls.fi.upm.es/lsd/nuevas-tendencias-en-sistemas-distribuidos/IntroToMapReduce_2.pdf</a></li></ul>
2	January 8	<ul style="list-style-type: none"><li>• <b>Introduction to MapReduce</b></li><li>• Pages 41-80 of <a href="#">BOOK-1</a></li><li>• Chapters 2 of <a href="#">BOOK-2</a></li><li>• Introduction to MapReduce and Hadoop by Matei Zaharia <a href="http://www.cs.berkeley.edu/~matei/talks/2010/amp-mapreduce.pdf">http://www.cs.berkeley.edu/~matei/talks/2010/amp-mapreduce.pdf</a></li></ul>
3	January 13	<ul style="list-style-type: none"><li>• <b>Introduction to MapReduce</b></li><li>• Pages 81-120 of <a href="#">BOOK-1</a></li><li>• Chapters 3 of <a href="#">BOOK-2</a></li><li>• Chapters 2 of <a href="#">BOOK-3</a></li></ul>
4	January 15	<ul style="list-style-type: none"><li>• <b>Introduction to MapReduce/Hadoop</b></li><li>• Pages 121-187 of <a href="#">BOOK-1</a></li><li>• Chapters 4 of <a href="#">BOOK-2</a></li><li>• Introduction to Hadoop</li></ul>
5	January 20	<ul style="list-style-type: none"><li>• <b>Practice MapReduce/Spark/Hadoop</b></li><li>• Making Big Data Simple: by Matei Zaharia: <a href="https://www.youtube.com/watch?v=Nev1s6fHwMI">https://www.youtube.com/watch?v=Nev1s6fHwMI</a></li><li>• Introduction to Spark: <a href="http://www.slideshare.net/jeykottalam/spark-sqlamp-camp2014?qid=a9e4b467-26b7-49a9-a361-8425876e9904&amp;v=qf1&amp;b=&amp;from_search=4">http://www.slideshare.net/jeykottalam/spark-sqlamp-camp2014?qid=a9e4b467-26b7-49a9-a361-8425876e9904&amp;v=qf1&amp;b=&amp;from_search=4</a></li></ul>
6	January 22	<ul style="list-style-type: none"><li>• <b>Spark/Hadoop Nuts and Bolts</b></li><li>• Making Big Data Simple: by Matei Zaharia: <a href="https://www.youtube.com/watch?v=Nev1s6fHwMI">https://www.youtube.com/watch?v=Nev1s6fHwMI</a></li><li>• Introduction to Spark: <a href="http://www.slideshare.net/jeykottalam/spark-sqlamp-camp2014?qid=a9e4b467-26b7-49a9-a361-8425876e9904&amp;v=qf1&amp;b=&amp;from_search=4">http://www.slideshare.net/jeykottalam/spark-sqlamp-camp2014?qid=a9e4b467-26b7-49a9-a361-8425876e9904&amp;v=qf1&amp;b=&amp;from_search=4</a></li><li>• Parallel-Programming-With-Spark-Matei-Zaharia: <a href="http://ampcamp.berkeley.edu/wp-content/uploads/2013/02/Parallel-Programming-With-Spark-Matei-Zaharia-Strata-2013.pdf">http://ampcamp.berkeley.edu/wp-content/uploads/2013/02/Parallel-Programming-With-Spark-Matei-Zaharia-Strata-2013.pdf</a> pptx</li></ul>
7	January 27	<ul style="list-style-type: none"><li>• <b>MapReduce Algorithm Design</b></li><li>• Chapter 3 of <a href="#">BOOK-2</a></li></ul>

8	January 29	<ul style="list-style-type: none"> <li>• <b>Language Models and Inverted Indexing</b></li> <li>• Chapter 4 of BOOK-2</li> <li>• <a href="#">PAPER-3</a></li> </ul>
9	February 3	<ul style="list-style-type: none"> <li>• <b>Midterm Exam</b></li> <li>• Exams are closed book, closed notes, closed computers</li> <li>• No makeup quizzes or exams will be given</li> </ul>
10	February 5	<ul style="list-style-type: none"> <li>• <b>Similar Item Detection</b></li> <li>•</li> </ul>
11	February 10	<ul style="list-style-type: none"> <li>• <b>Clustering and Classification</b></li> <li>•</li> </ul>
12	February 12	<ul style="list-style-type: none"> <li>• <b>Clustering and Classification</b></li> <li>•</li> </ul>
13	February 17	<ul style="list-style-type: none"> <li>• <b>NoSQL, MongoDB, Redis</b></li> <li>• Chapter 4 of BOOK-2</li> <li>• PAPER-4</li> </ul>
14	February 19	<ul style="list-style-type: none"> <li>• <b>NoSQL, MongoDB, Redis</b></li> <li>•</li> <li>•</li> </ul>
15	February 24	<ul style="list-style-type: none"> <li>• <b>Frequent Itemsets</b></li> <li>• Chapter 6 of BOOK-3</li> <li>•</li> </ul>
16	February 26	<ul style="list-style-type: none"> <li>• <b>MapReduce Design Patterns</b></li> <li>• Chapter 6 of BOOK-3</li> <li>•</li> </ul>
17	March 3	<ul style="list-style-type: none"> <li>• <b>Relational Algebra and MapReduce</b></li> <li>• Chapter 6 of BOOK-2</li> <li>• Relational Algebra and MapReduce: <a href="http://www.eurecom.fr/~michiard/teaching/slides/clouds/tutorial-high_level.pdf">http://www.eurecom.fr/~michiard/teaching/slides/clouds/tutorial-high_level.pdf</a></li> <li>• MapReduce examples: <a href="http://courses.cs.washington.edu/courses/cse344/11sp/sections/section8/section8-mapreduce-solution.pdf">http://courses.cs.washington.edu/courses/cse344/11sp/sections/section8/section8-mapreduce-solution.pdf</a></li> <li>• MapReduce and relational algebra: <a href="http://www.cs.kent.edu/~jin/Cloud12Spring/DatabaseMapReduce.pptx">http://www.cs.kent.edu/~jin/Cloud12Spring/DatabaseMapReduce.pptx</a></li> </ul>
18	March 5	<ul style="list-style-type: none"> <li>• <b>Regression Algorithms</b></li> <li>• Linear Regression</li> <li>• Ttest</li> </ul>
19	March 10	<ul style="list-style-type: none"> <li>• <b>Overview of Big Data and MapReduce</b></li> </ul>
	March 11	<ul style="list-style-type: none"> <li>• <b>Classes End</b></li> </ul>
	March 16-19	<ul style="list-style-type: none"> <li>• <b>Winter Final Examinations</b></li> </ul>