INLS 690-163
Intro to Big Data and NoSQL
Fall 2013

Thursday, 5:30 – 8:15pm, Smith 107

Instructor:

Arcot Rajasekar
Office: Manning 021
Office Hours: 3:30 – 5:00pm Thu, and by appointment
Email: rajasekar at unc dot edu

Course Description: This class will prepare students on current and emerging practices for dealing with big data (data in peta- and exa-scale) and large-scale database systems used by many social networking services. Information is being generated at an exponential scale and handling and analyzing these data need new types of tools and processes. Many areas are experiencing this growth from astronomy to finance to medicine to zebra fish genomics. Similarly, social networking sites, such as Facebook and Twitter, also deal with data that need different database requirements compared to traditional business applications. These applications are highly data intensive and require heavy read/write workloads. They also do not need some of the stringent ACID properties that are central to relational databases. These databases belong to an emerging genre called NoSQL databases that are mainly open source, non-schema oriented, having weak consistency properties and heavily distributed over large and evolving clusters of off-the-shelf server systems. The use and management these systems differ very much from traditional relational database systems. We will look at several such systems in this course,

Topics covered will include: fundamentals of big data and big data analytics, and NoSQL systems; examples of big data analytics such as Map Reduce an Hadoop; examples of NoSQL systems such as Google's BigTable, Amazon's Dynamo, Apache Cassandra (used by Facebook), Apache HBase (used by Yahoo and twitter), and other NoSQL systems such as MongoDB, Voldemort, CouchDB, SimpleDB, etc. Other topics include supporting systems such as the Google's File System, Chubby file system, etc.

Prerequisite(s): INLS 523 or permission of instructor

Textbook: None. The course will use papers published in the literature. The course will be somewhat reading intensive as new ideas are explored in each of these topics.

Grading Scheme:

1.	Class participation	5%
2.	Blog participation	10%
3.	Journal	15%
4.	Projects	40%
5.	Current Technology Paper and Presentation	10%
6.	Final exam	20%

1. Course Objectives:

- Study the requirements of non-traditional large-scale data applications
- Compare the properties of social networking sites against ACID properties of traditional databases
- Explore the fundamentals of NoSQL systems and big data analytics
- Examine application of very large clusters of COTS computing systems for solving large data problems
- Gain experience with NoSQL systems and Hadoop through hands-on projects.

2. Hardware and Software Requirements

We will be using open source software which will require installation and administration. RENCI/NCDS will be providing compute clusters for running programs such as Hadoop, Cassandra, MongoDB and other emerging data analytics. You may also be required to install and administer some of these on your laptop for testing purposes.

3. Graded Work

Your grade will be based on class and blog participation, keeping a journal, a technology paper and presentation, and through projects and a final take-home exam, weighted as shown under “Grade Weighting” on the first page.

Participation

I require all students to participate actively in class discussions throughout the class. At the beginning of each class, we will have a common discussion period, where we will discuss current events related to topics in the course. I expect that every student reads the ‘required reading’ list, posted at least a week before the class. As the class proceeds, I will be looking for questions, comments and a lively dialogue on the presented material as well as on the required reading materials. Apart from class participation, I also expect students to actively participate in blog posts on topics related to the course. Sometimes I will start a thread of conversation, but I also expect students to take initiative in starting new threads of discussions. The sakai site has facilities for blogs. I have also turned on the chat feature for our course in sakai to enable interactive discussions. There will be no homework – apart from the assigned reading list.

Journal

Each student is expected to maintain a journal. This is something of a personal digital library where one will keep all materials related to this course, gathered in the course or elsewhere. I expect material beyond the reading list to be part of your journal. Current events and class discussion topics can also be part of your journal. I also expect tags, metadata and your own commentary added for each material as an outcome of your reading the material. I would strongly recommend the use of the SILS Lifetime Library (http://lifetime-library.ils.unc.edu/) for maintaining the journal as it allows controlled sharing. Please make the material readable by me so that I can evaluate the progress. This journal will be a persistent digital library that may help you later after the course and which you can grow as you gather more relevant material.

Project work

I am planning on a series of projects – at least two: one with Hadoop, another with Cassandra or MongoDB. More information will be available as the course proceeds. We will be using platforms provided by RENCI/NCDS. Mr. Erik Scott from RENCI has agreed to help us in setting and administering these systems.

Current technology paper and presentation

Every student will sign up for a “current technology” paper and presentation during the semester. These will be completed in the form of a written document and an in-class presentation on a current issue or topic important to NoSQL database implementation, administration, or design. Part of the goal of this assignment is for you to gain confidence in reading, understanding, and presenting current topics about large-scale database research and technology.

Exam

There will be one take-home final exam.

4. Grading Policies

The following grade scale will be used AS A GUIDELINE (subject to any curve):

Graduate Percentage	Undergraduate Percentage
H 100-95%	A 100-90%
P+ 94-90%	B 89-80%
P 85-89%	C 79-70%
P- 80-84%	D 69-60%
L 70-79%	F Below 60%
F Below 70%	

This scale will be used as a GUIDELINE ONLY. The final grade scale may differ.

Due Dates and Late work
Project and paper assignment will have a due date and time and will include instructions for submission. Late submissions will be given a late penalty. Typically, a late penalty of 10% per day will be applied unless prior arrangements have been made with the instructor.

Requests for extensions and Absences
Any request for an extension must be made, preferably by email, at least 24 hours prior to the due date. Written documentation is required for illness. If a serious illness prevents you from taking part, send your instructor an e-mail message, or a friend with a note, describing your condition before schedule. Also, to establish a valid excuse for an illness you must get a note from a physician or the University infirmary.

Statute of limitations
Any questions or complaints regarding the grading of an assignment or test must be raised within one week after the score or graded assignment is made available (not when you pick it up).

5. Course Communication (Sakai)

Sakai-based course website has been set up and it is the responsibility of every student to **check the Sakai website regularly** for announcements and materials. The Announcements section of the website will be the source for all **official announcements** related to the class. Your instructor may announce tests, assignments, or changes to assignments in class, but there is no guarantee or promise that such announcements will be made in class. The Announcements section of the website is the **only** official, reliable source for announcements, changes, etc. from the instructor. If something the instructor says in class conflicts with information posted by the instructor on the website, then the information posted on by the instructor **on the Sakai website takes precedence**. Verbal instructions are easily misinterpreted, and they do not leave a documentation trail. All students should be able to access the system.

6. Honor Code

The UNC Honor Code is in effect for all work in this course. When work or ideas are not your own, you must attribute them. Unless otherwise stated, all assignments in this class are individual assignments, meaning that the substance of the work you turn in must be your own. If you have any doubts or questions about a course of action or a specific situation, please ask for clarification. Students should NOT receive (or give) major creative assistance or ongoing minor support on individual assignments. If you have any questions about this, please ask me.

7. Special Accommodations

If any student needs special accommodations, please contact the instructor during the first week of classes.

8. Timeline

No.	Date	Topic
1	08/22	Introduction, RDB Review, Motivation
2	08/29	Motivation, CAP Theorem, Big Table
3	09/05	Big Table, GFS, Chubby,
4	09/12	Bloom Filters, Map-Reduce
5	09/19	Hadoop, Pig and Hive
6	09/26	MongoDB, Dynamo,
7	10/03	Dynamo, Cassandra
8	10/10	Voldemort, MemCached, Memcachedb, TokyoCabinet
	10/17	Fall Break – No Class
9	10/24	HBase, HDFS, Zookeeper, Accumulo
10	10/31	SimpleDB, CouchDB, Neo4J, JENA
11	11/01	NewSQL, NuoSQL
12	11/07	More Big Data Analytics – Other Architectures
13	11/14	Presentations
14	11/21	Presentations, Discussions, Wrap Up – Last Class
	11/21	Project Reports Due
Exam	TBD	Take Home - Comprehensive