Course Syllabus

Course Information
(course number, course title, term, any specific section title)
CS 6301.001 26153 BIG DATA ANALYTICS/MANAGEMENT (3 Credits)
Tues & Thurs : 8:30am-9:45am ECSS 2.312

Professor Contact Information
(Professor's name, phone number, email, office location, office hours, other information)

Instructor: Dr. Latifur Khan

Office: ECSS (ES) 3.228

Phone: (972) 883 4137

E-mail: lkhan@utdallas.edu

Office Hours: Tuesday and Thursday: 9.45 a.m. to 10.45 a.m.

URL: http://www.utdallas.edu/~lkhan/Spring2013/CS6301.doc

Course Pre-requisites, Co-requisites, and/or Other Restrictions
(including required prior knowledge or skills)

Database Management Systems, JAVA, Machine Learning/AI (co-requisite)

Course Description

Popular relational database systems like IBM DB2, Microsoft SQLServer, Oracle, and Sybase are struggling to handle massive scale of data introduced by the Web, Social network and cyber physical systems. Now-a-days, companies have to deal with extremely large datasets. For example, on one hand, Facebook handles 15 TeraBytes of data each day into their 2.5 PetaByte Hadoop-powered data warehouse and on the other hand, eBay maintains a 6.5 PetaByte data warehouse. To handle emerging data at massive scale, "big data analytics" and "big data management" areas are emerging. Many traditional assumptions are not working, instead, new query and programming interfaces are required, and new computing models are emerging.

The course will focus on data mining and machine learning algorithms for analyzing very large amounts of data or Big data. Map Reduce and No SQL system will be used as tools/standards for creating parallel algorithms that can process very large amounts of data.

The course material will be drawn from textbooks as well as recent research literature. The following topics will be covered this year: Hadoop, Mapreduce, Association rules, Large scale supervised machine learning, Data streams, Clustering, NoSQL systems (Casadenra, Pig, Hive), and Applications including recommendation systems, Web and security.

Student Learning Objectives/Outcomes

By providing a balanced view of "theory" and "practice," the course should allow the student to understand, use, and build practical big data analytics an management systems. The course is intended to provide a basic understanding of the issues and problems involved in massive on-line repository systems, a knowledge of currently practical techniques for satisfying the needs of such a system, and an indication of the current research approaches that are likely to provide a basis for tomorrow's solutions.

Required Textbooks and Materials

The following textbook can be used this semester to augment the material presented in lectures:

- B1: Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Morgan & Claypool Publishers, 2010. http://lintool.github.com/MapReduceAlgorithms/ [Mandatory]
- B2: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Addison-Wesley April 2005. [Mandatory]
- B3: Anand Rajaraman and Jeff Ullman, Mining of Massive Datasets, Cambridge Press, http://infolab.stanford.edu/~ullman/mmds/book.pdf [Mandatory]
- B4: Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000. 550 pages. ISBN 1-55860-489-8. [Optional]
- B5: Chuck Lam, Hadoop in Action, December, 2010 | 336 pages ISBN: 9781935182191, http://netlab.ulusofona.pt/cp/HadoopinAction.pdf [Optional]

Papers Related to Big Data Analytics and Management

- P1: Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, *OSDI'04: Sixth Symposium on Operating System Design and Implementation,* San Francisco, CA, December, 2004. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/mapreduce-osdi04.pdf (shortened version: http://dl.acm.org/citation.cfm?doid=1327452.1327492)
- P2: Michael Stonebraker, Daniel Abadi, David J. DeWitt, Sam Madden, Erik Paulson, Andrew Pavlo, and Alexander Rasin. (2010) MapReduce and Parallel DBMSs: Friends or Foes? *Communications of the ACM*, 53(1):64-71.
- P3: Jeffrey Dean and Sanjay Ghemawat. (2010) MapReduce: A Flexible Data Processing Tool. *Communications of the ACM*, 53(1):72-77.
- P4: Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. (2006) Bigtable: A Distributed Storage System for Structured Data. *Proceedings*

of the 7th Symposium on Operating System Design and Implementation (OSDI 2006), pages 205-218.

- P5: Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. (2008) Pig Latin: A Not-So-Foreign Language for Data Processing. *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* pages 1099-1110.
- P6: Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham**,** Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints, *IEEE Transactions on Knowledge & Data Engineering (TKDE), 2011*, IEEE Computer Society, June 2011, Vol. 23, No. 6, Page 859-874.
- P7: Mohammad Husain, Mohammad Mehedy Masud, James McGlothlin, and Latifur Khan, Heuristics Based Query Processing for Large RDF Graphs Using Cloud Computing, *IEEE Transactions on Knowledge & Data Engineering (TKDE),* 2011, IEEE Computer Society, September 2011, Vol. 23, No. 9, Page 1312-1327.
- P8: Mohammad Masud, Tahseen Al-Khateeb, Latifur Khan , Charu Aggarwal, and Jiawei Han. Recurring and Novel Class Detection using Class-Based Ensemble, In *Proc. of IEEE International Conference on Data Mining(ICDM),* 2012, Belgium, Dec 2012.
- P9: Haoyuan Li, Yi Wang, Dong Zhang, Ming Zhang, and Edward Y. Chang. PFP: Parallel FP-Growth for Query Recommendation. In Proceedings of the 2008 ACM conference on Recommender systems.
- P10: Avinash Lakshman, Prashant Malik, Cassandra: a decentralized structured storage system, ACM SIGOPS Operating Systems Review archive, Volume 44 Issue 2, April 2010, Pages 35-40, ACM New York, NY, USA
- P11: Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, Utkarsh Srivastava, Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experience, VLDB 2009.

Clustering

- P12: Tian Zhang, Raghu Ramakrishnan, Miron Livny, BIRCH: A New Data Clustering Algorithm and Its Applications, *Data Mining and Knowledge Discovery*, Volume 1, Issue 2, 1997, 141-182.
- P13: Feng Luo, Latifur Khan et al. , A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles, *BIOINFORMATICS*, 20 (16): 2605-2617 NOV 1 2004.http://bioinformatics.oxfordjournals.org/cgi/reprint/20/16/2605

Classification

- P14: Indranil Palit and Chandan K. Reddy, "Scalable and Parallel Boosting with MapReduce", *IEEE Transactions on Knowledge and Data Engineering (TKDE),* Vol.24, No.10, pp.1904-1916, October 2012. [pdf]

Hive:

- P15: Thusoo, A.; Sarma, J.S.; Jain, N.; Zheng Shao; Chakka, P.; Ning Zhang; Antony, S.; Hao Liu; Murthy, R., "Hive - a petabyte scale data warehouse using

Hadoop," *Data Engineering (ICDE), 2010 IEEE 26th International Conference on* , vol., no., pp.996,1005, 1-6 March 2010 doi: 10.1109/ICDE.2010.5447738

Suggested Course Materials

Software

Mahout: http://mahout.apache.org/

Hive: https://cwiki.apache.org/confluence/display/Hive/Home

Piglatin: http://pig.apache.org/docs/r0.7.0/tutorial.html

Hadoop: http://hadoop.apache.org/

Casadenra: http://cassandra.apache.org/

Assignments & Academic Calendar
(Topics, Reading Assignments, Due Dates, Exam Dates)

Lectures

Topic	Chapters/Papesr	Homework/Lecture Notes
Hadoop+ Mapreduce	**Chapter 1, 2, 3 [B1], [B3]. Paper: P1, P2, P3, P4**	Hadoop with Mapreduce
Clustering Analysis	**Chapter 8 [B2]** **Chapter 9 [B2]** **Paper: P11, P12**	Lecture Note in Clustering DGSOT BIRCH
Mining Association Rules in Large Database Advanced Topics	**Chapter 6 [B2]** **Chapter 7 [B2]**	Association Rules
Recommendation System	**Chapter 9 [B3], P9**	
Classification, Prediction, Stream Mining	**Chapter 5 [B2]** **Chapter 4 [B3]** **Paper: P6, P7, P8, P13**	
Big Data Management: Pig Latin, Hive, Casadenra	**Paper: P5, P7, P10, P11, P15**	
Applications:	**Papers TBD**	
Exam	**April 23**	
Project Demonstration	**May 2**	

Assignment (Tentative)

Assignments will be based on:

- **Map Reduce Programming—Basic**
- **Map Reduce Programming – advanced**
- **Big Data Management Using Hive, Pig, Casadenra**
- **Large Scale Machine Learning Using Mahout**
- **Problem Solving Questions/Exercise Problems from Books**

Projects (Sample)

- **Tweeter Data Management**
- **Anomaly Detection**
- **Stream Mining for Tweets**
- **Text Mining with LDA**
- **Sentiment Analysis**

Grading Policy
(including percentages for assignments, grade scale, etc.)

One exam (in April), five assignments and a project. The course grade will be based on 35% of the exam, 40% of assignments, and 25% on the project. The project includes 1/2 page proposal, implementation and demonstration on May.

Course & Instructor Policies
(make-up exams, extra credit, late work, special assignments, class attendance, classroom citizenship, etc.)

No make up exam.

Field Trip Policies
Off-campus Instruction and Course Activities

Off-campus, out-of-state, and foreign instruction and activities are subject to state law and University policies and procedures regarding travel and risk-related activities. Information regarding these rules and regulations may be found at the website address http://www.utdallas.edu/BusinessAffairs/Travel_Risk_Activities.htm. Additional information is available from the office of the school dean. Below is a description of any travel and/or risk-related activity associated with this course.

Student Conduct & Discipline

The University of Texas System and The University of Texas at Dallas have rules and regulations for the orderly and efficient conduct of their business. It is the responsibility of each student and each student organization to be knowledgeable about the rules and regulations which govern student conduct and activities. General information on student conduct and discipline is contained in the UTD publication, *A to Z Guide*, which is provided to all registered students each academic year.

The University of Texas at Dallas administers student discipline within the procedures of recognized and established due process. Procedures are defined and described in the *Rules and Regulations, Board of Regents, The University of Texas System, Part 1, Chapter VI, Section 3*, and in Title V, Rules on Student Services and Activities of the university's *Handbook of Operating Procedures*. Copies of these rules and regulations are available to students in the Office of the Dean of Students, where staff members are available to assist students in interpreting the rules and regulations (SU 1.602, 972/883-6391).

A student at the university neither loses the rights nor escapes the responsibilities of citizenship. He or she is expected to obey federal, state, and local laws as well as the Regents' Rules, university regulations, and administrative rules. Students are subject to discipline for violating the standards of conduct whether such conduct takes place on or off campus, or whether civil or criminal penalties are also imposed for such conduct.

Academic Integrity

The faculty expects from its students a high level of responsibility and academic honesty. Because the value of an academic degree depends upon the absolute integrity of the work done by the student for that degree, it is imperative that a student demonstrate a high standard of individual honor in his or her scholastic work.

Scholastic dishonesty includes, but is not limited to, statements, acts or omissions related to applications for enrollment or the award of a degree, and/or the submission as one's own work or material that is not one's own. As a general rule, scholastic dishonesty involves one of the following acts: cheating, plagiarism, collusion and/or falsifying academic records. Students suspected of academic dishonesty are subject to disciplinary proceedings.

Plagiarism, especially from the web, from portions of papers for other classes, and from any other source is unacceptable and will be dealt with under the university's policy on plagiarism (see general catalog for details). This course will use the resources of turnitin.com, which searches the web for possible plagiarism and is over 90% effective.

Email Use

The University of Texas at Dallas recognizes the value and efficiency of communication between faculty/staff and students through electronic mail. At the same time, email raises some issues concerning security and the identity of each individual in an email exchange. The university encourages all official student email correspondence be sent only to a student's U.T. Dallas email address and that faculty and staff consider email from students official only if it originates from a UTD student account. This allows the university to maintain a high degree of confidence in the identity of all individual corresponding and the security of the transmitted information. UTD furnishes each student with a free email account that is to be used in all communication with university personnel. The Department of Information Resources at U.T. Dallas provides a method for students to have their U.T. Dallas mail forwarded to other accounts.

Withdrawal from Class

The administration of this institution has set deadlines for withdrawal of any college-level courses. These dates and times are published in that semester's course catalog. Administration procedures must be followed. It is the student's responsibility to handle withdrawal requirements from any class. In other words, I cannot drop or withdraw any student. You must do the proper paperwork to ensure that you will not receive a final grade of "F" in a course if you choose not to attend the class once you are enrolled.

Student Grievance Procedures

Procedures for student grievances are found in Title V, Rules on Student Services and Activities, of the university's *Handbook of Operating Procedures*.

In attempting to resolve any student grievance regarding grades, evaluations, or other fulfillments of academic responsibility, it is the obligation of the student first to make a serious effort to resolve the matter with the instructor, supervisor, administrator, or committee with whom the grievance originates (hereafter called "the respondent"). Individual faculty members retain primary responsibility for assigning grades and evaluations. If the matter cannot be resolved at that level, the grievance must be submitted in writing to the respondent with a copy of the respondent's School Dean. If the matter is not resolved by the written response provided by the respondent, the student may submit a written appeal to the School Dean. If the grievance is not resolved by the School Dean's decision, the student may make a written appeal to the Dean of Graduate or Undergraduate Education, and the deal will appoint and convene an Academic Appeals Panel. The decision of the Academic Appeals Panel is final. The results of the academic appeals process will be distributed to all involved parties.

Copies of these rules and regulations are available to students in the Office of the Dean of Students, where staff members are available to assist students in interpreting the rules and regulations.

Incomplete Grade Policy

As per university policy, incomplete grades will be granted only for work unavoidably missed at the semester's end and only if 70% of the course work has been completed. An incomplete grade must be resolved within eight (8) weeks from the first day of the subsequent long semester. If the required work to complete the course and to remove the incomplete grade is not submitted by the specified deadline, the incomplete grade is changed automatically to a grade of **F**.

Disability Services

The goal of Disability Services is to provide students with disabilities educational opportunities equal to those of their non-disabled peers. Disability Services is located in room 1.610 in the

Student Union. Office hours are Monday and Thursday, 8:30 a.m. to 6:30 p.m.; Tuesday and Wednesday, 8:30 a.m. to 7:30 p.m.; and Friday, 8:30 a.m. to 5:30 p.m.

The contact information for the Office of Disability Services is:
The University of Texas at Dallas, SU 22
PO Box 830688
Richardson, Texas 75083-0688
(972) 883-2098 (voice or TTY)

Essentially, the law requires that colleges and universities make those reasonable adjustments necessary to eliminate discrimination on the basis of disability. For example, it may be necessary to remove classroom prohibitions against tape recorders or animals (in the case of dog guides) for students who are blind. Occasionally an assignment requirement may be substituted (for example, a research paper versus an oral presentation for a student who is hearing impaired). Classes enrolled students with mobility impairments may have to be rescheduled in accessible facilities. The college or university may need to provide special services such as registration, note-taking, or mobility assistance.

It is the student's responsibility to notify his or her professors of the need for such an accommodation. Disability Services provides students with letters to present to faculty members to verify that the student has a disability and needs accommodations. Individuals requiring special accommodation should contact the professor after class or during office hours.

Religious Holy Days

The University of Texas at Dallas will excuse a student from class or other required activities for the travel to and observance of a religious holy day for a religion whose places of worship are exempt from property tax under Section 11.20, Tax Code, Texas Code Annotated.

The student is encouraged to notify the instructor or activity sponsor as soon as possible regarding the absence, preferably in advance of the assignment. The student, so excused, will be allowed to take the exam or complete the assignment within a reasonable time after the absence: a period equal to the length of the absence, up to a maximum of one week. A student who notifies the instructor and completes any missed exam or assignment may not be penalized for the absence. A student who fails to complete the exam or assignment within the prescribed period may receive a failing grade for that exam or assignment.

If a student or an instructor disagrees about the nature of the absence [i.e., for the purpose of observing a religious holy day] or if there is similar disagreement about whether the student has been given a reasonable time to complete any missed assignments or examinations, either the student or the instructor may request a ruling from the chief executive officer of the institution, or his or her designee. The chief executive officer or designee must take into account the legislative intent of TEC 51.911(b), and the student and instructor will abide by the decision of the chief executive officer or designee.

These descriptions and timelines are subject to change at the discretion of the Professor.