CS 6301 Big Data Analytics and Management (Graduate Level)

Fall 2013

People:

Instructor: Dr. Latifur Khan

Office: ECSS (ES) 3.228

Phone: (972) 883 4137

E-mail: lkhan@utdallas.edu

Office Hours: Monday and Wednesday: 2.55 p.m. to 3.55 p.m.

URL: http://www.utdallas.edu/~lkhan/Fall2013/CS6301.doc

Class Time

CS 6301.012 89746 BIG DATA MANAGEMENT & ANALYTICS (3 Credits) Mon & Wed: 4:00pm-5:15pm
Location: JO 3.516

Teaching Assistants (TA):

"M. Solaimani" <mxs121731@utdallas.edu>;

Office Hour: TBD

Office Location: TBD

"Mohammad Ridwanur Rahman" <mxr127030@utdallas.edu>;

Office Hour: TBD

Office Location: TBD

Course Summary

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, NoSQL systems (Cassandra, Pig, Hive, MongoDB, Hbase), Association rules, Large scale supervised machine learning, Data streams, Clustering, and Applications including recommendation systems, Web and security.

Class Learning:

Class Learning Outcomes	**Number of Students**				
	Below Expectations	**Progressing to Criteria**	**Meets Criteria**	**Exceeds Criteria**	
Understanding of conceptual, logical and physical organization of big data					H
Understanding of large data processing using Map-Reduce					H
Understanding of NoSQL models, theory and practices					H
Understanding of data modeling, indexing, query processing for big data					H
Understanding of recommendation methods for big data					
Understanding of unsupervised learning for big data					
Understanding of supervised learning for big data including					

Requirements

Two exams (in October & December 11), a couple of assignments (preferably 5) and a project. Your course grade will be based on 40% of the exam, 45% of assignments, and 15% on the project. The project includes 1/2 page proposal, implementation and demonstration on December (after exam 2).

Perquisite:

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

Course 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 [May Expand Further]

- 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 M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W. Hamlen, Nikunj C. Oza: Facing the reality of data stream classification: coping with scarcity of labeled data. Knowl. Inf. Syst. 33(1): 213-244 (2011)
- 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

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/

Cassandra: http://cassandra.apache.org/

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, P10, P11**	
Applications:	**Papers TBD**	
Exam I & II	**October (TBD) &** **December 11**	
Project Demonstration	**December (TBD)**	

Assignment (Tentative)

Assignments will be based on:

- **Map Reduce Programming—Basic & Hands on HDFS setup**
- **Map Reduce Programming – advanced**
- **Big Data Management Using Hive, Pig, Cassandra (including Hands on setup)**
- **Large Scale Machine Learning Using Mahout**
- **Problem Solving Questions/Exercise Problems from Books**

Projects (Sample—Expanded further)

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