CS 6350 Big Data Analytics and Management (Graduate Level)

Spring 2019

People:

Instructor: Dr. Latifur Khan

Office: ECSS (ES) 3.228

Phone: (972) 883 4137

E-mail: lkhan@utdallas.edu

Office Hours: Wednesday: 1.00 p.m. to 2.00 p.m. or via Email

Course Info:

CS 6350.001
24102 Big Data Management and Analytics
Monday & Wednesday
11:30am - 12:45pm
ECSW 1.315

Teaching Assistant (TA):

TBD

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/Internet of Things (IOT) devices. 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 NoSQL 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 (Spark, Cassandra, Pig, BigTable), Large scale supervised machine learning, Data streams, Clustering, advanced machine learning and Applications including recommendation systems, Web and security.

Requirements

Two pop up quiz, two Exams, a couple of assignments (preferably 4) and a group project. Your course grade will be based on 10% of the quiz, 30% of the exams (Exam I 10% & Exam II 10%), 48% of assignments, and 12% of the project. The project includes implementation and demonstration.

Class Learning Outcomes	Number of Students			Material Used
	Below Expectations	Meets Criteria	Exceeds Criteria	
Ability to understand of conceptual, logical and physical organization of big data				HW 1, 2, 3, 4 EXAMs
Ability to understand of large data processing using Map-Reduce				HW 1, 2, 3, 4 EXAMs
Ability to understand of NoSQL models, theory and practices				HW 1.2 3, 4, EXAMs
Ability to understand of data modeling, indexing, query processing for big data				HW 1, 2, 3, 4, EXAMs
Ability to understand of recommendation methods for big data				HW 4, EXAMs
Ability to understand of unsupervised learning for big data				HW 4, EXAMs
Ability to Understand of supervised learning for big data				HW 4, EXAMs
Ability to communicate and work on team software project				Project

Perquisite:

Database Management Systems, JAVA (intermediate/advanced), Linux OS, Scala (Preferable)

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: 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
- P8: 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.
- P9: M. Armbrust, R. Xin, C. Lian, Y. Huai, D. Liu, J. Bradley, X. Meng, T. Kaftan, M. Franklin, A. Ghodsi and M. Zaharia. Spark SQL: Relational Data Processing in Spark. *SIGMOD 2015*, June 2015.
- P10: M. Zaharia. An Architecture for Fast and General Data Processing on Large Clusters **(PhD Disseration)**.
- P11: M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica. Discretized Streams: Fault-Tolerant Streaming Computation at Scale, *SOSP 2013*, November 2013.

- P12: Ahsanul Haque*, Latifur Khan, Michael Baron and Charu Aggarwal. Efficient Semi-Supervised Adaptive Classification and Novel Class Detection over Data Stream, 32nd IEEE International Conference on Data Engineering, May 16-20, 2016 · Helsinki, Finland

Recommendation System:

- P13: Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." *Computer* 8 (2009): 30-37.

  http://doi.ieeecomputersociety.org/10.1109/MC.2009.263

Software

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

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

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

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

Kafka: https://cwiki.apache.org/confluence/display/KAFKA/Kafka+papers+and+presentations

Storm: http://storm-project.net/

Spark https://spark.apache.org/

Flink http://flink.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
Big Data Management: Spark, Spark SQL, Hive, NoSQL Pig Latin, Cassandra, BigTable, HBase	**Paper: P5, P7, P8, P9, P10**	
Stream Data Management: Spark Stream	Paper: P11	
Clustering Analysis	**Chapter 8 [B2]** **Chapter 9 [B2]**	Lecture Note in Clustering
Classification, Prediction, Stream Mining	**Chapter 5 [B2]** **Chapter 4 [B3]** **Paper: P6, P12**	
Recommendation System	**Chapter 9 [B3], P13**	
Advanced Analytics	**Papers TBD**	
Exam I, II	**TBD**	
Project Demonstration	**TBD (May)**	

Assignment (Tentative)

Assignments will be based on:

1. **Map Reduce Programming—Basic & Hands on HDFS setup & Advanced Map Reduce Programming**

2. **Text Data Processing and Analytics using Spark**

3. **Big Stream Text Data Management Using Spark SQL, Hive, Data Frame**

4. **Problem Solving Questions/Exercise Problems from Books; and Large Scale Recommendation Using Spark, Spark and Mlib**

Projects (Sample—Expanded further)

- **Text Mining with LDA**
- **Geo Location Mining**
- **Smart Phone Apps Fingerprinting**
- **Smart Phone Data Analytics**
- **NLP Processing for big Web Data**
- **Cyber Data Analysis**
- **Secure Data Analytics wit Trusted Execution Environment**