CS 4371.501 Introduction to Big Data Management and Analytics (Spring 2020)

People:

Instructor: Erick Skorupa Parolin

Office: ECSS 3.614 or ECSS 3.101

Phone: (214) 708 7760

E-mail: exs172930@utdallas.edu

Office Hours: Wednesdays / 6:00 pm to 7:00 pm

Course Info:

CS 4371.501 Introduction to Big Data Management and Analytics (3 Semester Credit Hours)

Class Meeting Time: Mondays & Wednesdays 7:00pm - 8:15pm

Class Location: ECSS 2.412

Teaching Assistant (TA):

Name: Vasundhara Komaragiri

Email: vxk180011@utdallas.edu

Office Hours: Mondays 5:30-6:30 pm and Tuesdays 4:00-5:00 pm

TA Office Room: ECSS 2.103B1

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.

Prerequisites:

CS 2336: Computer Science II

CS 4347: Database Systems

Java Programming (intermediate/advanced), Linux OS, Python/Scala programming

Grading Criteria and Requirements:

Midterm: 25%

Final Exam: 25%

Homeworks: 30%

Project (Group): 15%

Pop Quizzes and Attendance: 5%

Class Learning Outcomes	Material Used
Ability to understand conceptual, logical and physical organization of big data.	HW 1, 2, 3, 4 EXAMs
Ability to understand 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	HW 1, 2, 3, 4 EXAMs

for big data	
Ability to Understand of unsupervised learning for big data	HW 4, EXAMs
Ability to communicate and work on team software project	Project

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

- 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 Dissertation).

- 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

- 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

Softwares

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

Hadoop MapReduce - MapReduce Basics (Chapters 1 and 2)
Hadoop MapReduce - Algorithm Design (Chapter 3)
Hadoop MapReduce - Inverted Indexing (Chapter 4)

Big Data Management - Spark
Big Data Management - NoSQL (tentative)
Big Data Management - Cassandra (tentative)
Big Data Management - BigTable (tentative)

Stream Data Management - Kafka
Stream Data Management - Spark Streaming

Data Analytics and Machine Learning - Clustering Analysis and Classification (tentative)

Assignment (Tentative)

Assignments will be based on:

1. Environment Setup (Hadoop / OS)

2. Map Reduce Programming - Basic/Intermediate Hadoop Map Reduce Programming;

3. Data Processing and Analytics using Spark;

4. Big Data Management Using Spark SQL, Spark Streaming, Kafka;

Projects:

The project will be done in groups and includes implementation and demonstration. Projects will be based on trendy technologies/techniques, like:

Window-based Stream Data Analytics with SPARK and Kafka
Text Mining with LDA
Geo Location Mining
Smart Phone Apps Fingerprinting
Smart Phone Data Analytics
NLP for Big Web Data
Cyber Data Analysis
Secure Data Analytics with Trusted Execution Environment