Data-Intensive Analytics (Fall 2013)

Spiros Papadimitriou <spapadim@business.rutgers.edu>

Course outline:

This course provides an overview the hot topic of "big data." Although the term is overused, "big data" is commonly seen to refer to the "three Vs": volume, velocity, and variety. The aim of this course is to introduce students to each of those aspects, and provide them with general understanding of the key concepts and challenges, as well as some familiarity with the fundamental techniques and how they can be applied to derive insights from the huge and messy volumes of data collected by almost every business, or every online activity today.

The course does not assume any previous experience. However, students are expected to have knowledge of basic probability, statistics, and linear algebra, as well as a basic understanding of programming concepts.

Topics:

1. Introduction to "big data"
2. Large-scale, distributed information processing (volume)
    a. MapReduce and Hadoop
    b. Graph algorithms
    c. Text retrieval algorithms
    d. Theoretical analysis of MapReduce programs
    e. Hidden Markov Models
    f. BigTable, Hive, and Pig
3. Stream data processing and analytics (velocity)
    a. Streaming models
    b. Clustering and classification
    c. Change detection
    d. Synopsis construction
    e. Dimensionality reduction and forecasting
    f. Distributed stream mining
    g. Real-time analytics
4. Sensing and context awareness (variety)
    a. Context-aware analytics
    b. Introduction to sensing and analytics
    c. Social and mobile sensing
    d. Applications: healthcare and sciences
    e. The IOT (Internet of Things) from a data management perspective

Reference books:

1. Tom White, *Hadoop: The Definitive Guide*, O'Reilly, 2009.
2. Anand Rajaraman and Jeff Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2011.
3. Charu Aggarwal (ed.), *Data Streams: Models and Algorithms*, Springer 2006.

Reference material:

1. Spiros Papadimitriou, Jimeng Sun, and Rong Yan, *Large-Scale Data Mining: MapReduce and Beyond* , tutorial (multiple venues, including ICDM, KDD, NSA workshop on graph analytics).
2. Spiros Papadimitriou, Michail Vlachow, *Temporal Data Mining*, tutorial (broader scope, but parts on stream mining; multiple venues).
3. Jimmy Lin, *Data Intensive Information Processing Applications*, course at U. Maryland (2007, 2008, and 2010).

Additional reading (sample):

1. Charu Aggarwal (ed.), *Managing and Mining Sensor Data*, Springer 2013.
2. C.-T. Chu, S.K. Kim, Y.-A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, K. Olukotun: Map-Reduce for Machine Learning on Multicore, NIPS 2006.
3. J. Yuan, Y. Zheng, X. Xie: Discovering regions of different functions in a city using human mobility and POIs, KDD 2012.
4. C. Liu, H. Xiong, Y. Ge, W. Geng, Matt Perkins: A Stochastic Model for Context-Aware Anomaly Detection in Indoor Location Traces, ICDM 2012.
5. S. Papadimitriou, P.S. Yu: Optimal multi-scale patterns in time series streams. SIGMOD 2006.
6. S. Dhar and U. Varshney, Challenges and business models for mobile location-based services and advertising, CACM 54(5), 2011.
7. P. Anastasopoulou, S. Härtel, M. Tubic, S. Hey, Using Support Vector Regression for assessing Human Energy Expenditure using a Triaxial Accelerometer and a Barometer, MobiHealth 2012.
8. H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, J. Tian, Mining Personal Context-Aware Preferences for Mobile Users, ICDM 2012.
9. J. Huang, D. Millman, M. Quigley, D. Stavens, S. Thrun, A. Aggarwal, Efficient, generalized indoor WiFi GraphSLAM, ICRA 2011.