CSCI E-63 Big Data Analytics (24038) Spring term (4 credits)

Zoran B. Djordjević, PhD, Senior Enterprise Architect, NTT Data, Inc.

Lectures: Fridays from January 29th, 2016, from 5:30 to 7:30PM, Rm306, 1 Story St, Cambridge

Optional Online Sections: Saturdays, 10-11:30AM, Starting January 30th, 2016

The recent explosion of social media and the computerization of every aspect of economic activity resulted in creation of large volumes of mostly unstructured data: web logs, videos, speech recordings, photographs, e-mails, Tweets, and similar. In a parallel development, computers keep getting ever more powerful and storage ever cheaper. Today, we have the ability to reliably and cheaply store huge volumes of data, efficiently analyze them, and extract business and socially relevant information. This course brings together several key IT technologies used in manipulating, storing, and analyzing big data. We will look at the basic tools for statistical analysis, R and Python, and key methods used in Machine Learning. We will review MapReduce techniques for parallel processing and Hadoop, an open source framework that allow us to cheaply and efficiently implement MapReduce on internet scale problems. We will spend considerable time mastering Spark and few other memory based evolutions of Hadoop. We will touch on related tools that provide SQL-like access to unstructured data like Hive. We will analyze so-called NoSQL storage solutions exemplified by Casandra and Hbase for their critical features: speed of reads and writes, data consistency, and ability to scale to extreme volumes. We will examine memory resident databases (VoltDB) and streaming technologies which allow analysis of data in flight, i.e. real time. Students will gain the ability to design highly scalable systems that can accept, store, and analyze large volumes of unstructured data in batch mode and/or real time.

Prerequisites: Familiarity with Intermediate Java is advised. Most assignments could easily be done in Ruby, Python, C#, or Perl. We will assume no familiarity with Linux and will introduce you to all essential Linux commands. Students need to have access to a computer with 64 bit operating system and at least 4 GB of RAM. 8 GB or more of RAM is preferred.

Lectures: Lectures will be delivered live and simultaneously made available for online viewing through BlackBoard Collaborate Web Conferencing tool. Streaming recording might also be available. Links to BlackBoard Collaborate recorded lectures will be accessible on the course Web site within one hour after the end of the lecture. If streaming video is provided, recorded lectured will become available with a delay of up to two days.

References: Detailed handouts with references to material on the Web will be handed out every week. There is no required text book.

Grading: Practically every class will be followed by a homework assignment. Grades on the solutions for class assignments constitute approximately 85% of the final grade. 15% of the grade will be earned through the final project. Final projects will be assigned four weeks before the end of the class. You will produce a paper (10+ pages of MS Word text, 10+ PowerPoint Slides, a working demo, 15 minute YouTube Video of your presentation and a brief 2 minute YouTube video that might be presented to the class on the day of final presentations. Several students will be invited to present their final projects live to the entire class. 95% or higher cumulative grade on all assignments and the final project gives you an A as the final grade in the course, 90-94.9% gives you an A-, 85-89.9% a B+, 80-84.9% a B, etc.

Communications: zdjordj@fas.harvard.edu, Piazza class site.

Tentative List of Class Topics:

	Date	Topic
1	01/29/16	**R** programming language for statistical analysis.
2	02/05/16	**MapReduce Framework and Hadoop.** Embarrassingly parallel processes and other design patterns for big data processing. Cloudera virtual machine. **HDFS -** Hadoop Distributed Filesystem, **YARN -** Yet Another Resource Negotiator..
3	02/12/16	**Spark.** A memory based evolution of MapReduce framework with considerable improvement in execution speed. **Spark APIs**
4	02/19/16	**Near Real Time processing with Spark.** Micro Batch Computing and Near Real Time Analytics, with Kafka and Flume. Stream Computing is a technology oriented towards data that is continuously flowing or streaming and requires inflight processing.
5	02/26/16	**Apache Flink.** Flink processes data streams as true streams, i.e., data elements are immediately "pipelined" though a streaming program as soon as they arrive. This allows to perform flexible window operations on streams.
6	03/04/16	**Basic Ideas of Machine Learning.** Algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.
7	03/11/16	**Visualizing Large Data Sets with D3**. We will introduce a Java Script API and techniques that enable more insightful use of graphs and charts presenting the content and features of large data set in your Big Data application.
	03/18/16	No Class Spring Break
8	03/25/16	**Neo4J, a Graph Database.** A storage and retrieval system based on hierarchical structures which have proven themselves very efficient for fast queries among highly correlated data.
9	04/01/16	**Natural Language Processing.** Basic mechanisms for processing and analysis of written text.
10	04/08/16	**Advanced Text Processing.** Advanced tools for analysis of speech and written text and their use in commercial applications.
11	04/15/16	**CUDA. Compute Unified Device Architecture** is a parallel computing platform and API model created by NVIDIA. It allows software developers to use graphics processing units (GPU) for general purpose processing. Basic Programming techniques.
12	04/22/16	**CUDA based Frameworks** for advanced Statistical Analysis.
13	04/29/16	**VoltDB** is a representative of a new breed of in-memory databases equipped with sophisticated tools for efficient high performance computations.
14	05/06/16	**Data Flow Computing** is a new, revolutionary way of performing computations, completely different to computing with conventional CPUs or GPUs. Dataflow computers focus on optimizing the movement of data in an application and utilize massive parallelism between thousands of tiny 'dataflow cores' to provide order of magnitude benefits in performance, space and power consumption. We will learn to perform Data Flow computations using Maxeler's technology.
14	05/13/16	**Presentations of selected student projects.**