CSCI E-63 Big Data Analytics (24038) 2017 Spring term (4 credits)
Zoran B. Djordjević, PhD, Senior Enterprise Architect, NTT Data, Inc.
Lectures: Fridays starting on September 1st, 2017, from 5:30 to 7:30 PM (EST), 1 Story Street, Room 306, Cambridge, MA

Optional Online Sections: Saturdays, starting September 2nd, 2017 at 12:00 noon (EST) with an introduction to AWS Cloud and Linux OS.

The emphasis of this course is on mastering two most important big data technologies: Spark 2 and Deep Learning with TensorFlow. Spark is an evolution of Hadoop and Map/Reduce but with massive speedup and scalability improvements. TensorFlow is Google's open-source framework for distributed neural networks-based machine learning. The explosion of social media and the computerization of every aspect of social and economic activity results in the creation of large volumes of semi-structured data: web logs, videos, speech recordings, photographs, e-mails, Tweets, and similar data. In a parallel development, computers keep getting ever more powerful and storage ever cheaper. Today, we can reliably and cheaply store huge volumes of data, efficiently analyze them, and extract business and socially relevant information. This course familiarizes the students with the most important information technologies used in manipulating, storing, and analyzing big data. We examine the basic tools for statistical analysis, R and Python, and several machine learning algorithms. We examine Spark Core, Spark ML (machine learning) API, and Spark Streaming which allows analysis of data in flight, that is, in near real time. We learn to use TensorFlow for several standard data analysis practices, including regression, clustering, classification, and others. We learn about so-called NoSQL storage solutions exemplified by Cassandra for their critical features: speed of reads and writes, and the ability to scale to extreme volumes. We learn about memory-resident databases and graph databases (Spark GraphX and Ne4J). We acquire practical skills in scalable messaging systems like Kafka and Amazon Kinesis. We will conduct most of our exercises in Amazon Cloud, so the students will master the most important AWS services. At the end of the course, students will be able to initiate and design highly scalable systems that can accept, store, and analyze large volumes of unstructured data in batch mode and/or real time. Most lectures are presented using Python examples. Some lectures use Java and R.

Prerequisites: Familiarity with intermediate Python, Java, Scala, or R. Most assignments could easily be done in one of those four languages, though we consider Python and Java the most convenient. We assume no familiarity with Linux and introduce all essential Linux features and commands. We assume no familiarity with AWS Cloud and will introduce all essential services. Students need access to a computer with a 64-bit operating system and at least 4 GB of RAM. Note: 8 GB or more of RAM is strongly advised.

Lectures: Lectures will be delivered live and made available after lectures for online viewing through Zoom Web Conferencing tool. Recordings of video streams will also be available. Zoom Links to recorded lectures and lab sessions will be accessible on the course Web site one or two hours after the end of the lecture or lab session. Streaming videos of recorded lectures will become available with a delay, hopefully by Saturday morning.

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 a few weeks before the end of the class. For the final project, 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 to be presented to the whole class on the day of the final presentations. Several students will be invited to present their final projects live to the entire class.

Grades: 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, Canvas class site and Piazza, once class starts.

Tentative List of Class Topics:

	Date	**Topic**
1	09/01/2017	**Basic Statistics and R.** we will cover basic statistical concepts with a brief review of R, a language very much used by statisticians. This course will use Python much more that R but we want to acknowledge the importance of R and its libraries.
2	09/08/2017	Relationships and Representations, Graph Databases. We will use Neo4J graph database to represent relationships among objects in IT space, as well as concepts and words in spoken languages. Various types of relationship discovery and representations are essential in solving many of our data analysis problems. Neo4J and similar technologies make our understanding of complex problems much easier.
3	09/15/2017	Introduction to Spark 2.0. Spark 2.0 replaced Hadoop as the dominant mainstream framework for processing of large data volumes on large computational clusters. Initially, we will learn how to formulate our calculations so that they could process big data in batch mode. We will discuss setup of Spark clusters in “on premise” environments and in the Cloud
4	09/22/2017	Language processing with Spark 2.0. Processing large volumes of textual data is very important step in many business analysis applications. We will learn how to combine tools for natural language processing (NLP) with computational efficiency of Spark 2.0. In this lecture we will introduce Cassandra, one of NoSQL databases for fast storage and retrieval of big volumes of (textual) data.
5	09/29/2017	Analysis of Streaming Data with Spark 2.0. While many applications could profit handsomely from batch processing of large volumes of data, some application must process a lot of data practically in real time. Spark provides its Streaming API as a powerful tool for such scenarios. In this lecture we will introduce two special messaging system (Kafka and Kinesis) each of which acts as a buffer between actual data sources and Spark processing engine.
6	10/06/2017	Applications of Spark ML Library. Spark comes with a Machine Learning (ML) API, which allows us to perform many routing ML task at Spark speed. We will learn how to select use cases or scenarios in which Spark ML library is the most appropriate tool. We will also learn how to use Spark’s computational engine for Neural Network processing.

7	10/13/2017	Basic Neural Network and Tensor Flow. Neural Networks and Deep Learning are emerging as the highest precision tools for many large scale classification and pattern recognition problems. We will learn how to use Tensor Flow both on GPU and CPU machines.
	10/20/2017	Spring Break
8	10/27/2017	Advance Tensor Flow. We will analyze some more complex configurations of Neural Networks and also learn how to integrate NN engines into practical systems for large scale analysis. In particular we will learn how to integrate NNs with fast NoSQL storage systems like Mongo DB and Cassandra.
9	11/03/2017	Assessing Quality of Big Data Analysis. We will learn “standard” procedures for accessing quality of ML algorithms. We will learn also learn how to access precision of other large scale calculations.
10	11/10/2017	Analysis of Images, OCR Applications. Analysis of images and pattern recognition are part and parcel of many applications. We will learn how to use some standard API-s to perform such analysis at big data speed.
11	11/17/2017	Analysis of Speech Signal. Many intelligent devices can now speak back to us. We will learn how to build large scale systems that can process speech in real time.
	11/24/2017	Thanksgiving Holiday
12	11/24/2017	Question Answer Systems are the true test of our ability to build intelligent machines. We will learn how to build QA systems using TensorFlow and two intelligent AWS API-s Polly, and Amazon Lex. We will also examine benefits of AWS Rekognition API
13	12/01/2017	Page Rank like Search systems. Searching through large volumes of textual data at very high speed is what made Google.com possible. We will learn how such systems are build and analyze possibilities to search through large volumes of sound and video data.
14	12/08/2017	Analysis of Streaming Data and Time Series with Tensor Flow and VoltDB, Data Flow Engines and other memory databases. We will understand comparative advantages of different technologies for processing of fast moving data.
15	12/15/2017	Final Project Presentations