Syllabus

Big Data Computing

- **Instructor:**
  - Zoltan C. Toth
  - TothZ@ceu.edu
  - +36 30 291 3599

- **Credits:** 2 (4 ECTS)
- **Term:** Fall 2017-2018
- **Course level:** MSc
- **Prerequisites:**
  - Basic programming knowledge (Data Analytics 1a course)
  - SQL Knowledge (Different Shapes of Data Course is highly recommended)
  - Linux command-line knowledge (Different Shapes of Data Course is highly recommended)

Course description

This is a technology-focused course on distributed data analytics systems.

Current Data Analytics Architectures often work with an amount of data that cannot be fit on a single computer. Even companies that work with reasonably small datasets are expecting rapid growth, so they prefer to use data analytics solutions that are easy to distribute and scale. In this course you will get an overview and hands-on experience with modern distributed data-analytics (a.k.a. Big Data) systems.

Learning outcomes

At the end of this course you will have an overview of Big Data technologies applied in modern businesses. You will have a general understanding on how these technologies work and you will be able to reason about when to use or not to use them. You will be hands-on with Apache Spark, the system, which has been getting the de-facto Big Data analytics tool of choice.

Once you completed the assignments for this course, you will be hands-on with the following technologies:

- Basic Cloud Computing operations
- Spark Architecture and internal operations
- Spark SQL and DataFrames in Spark
- Advanced Optimizations in Spark
- Basics of Real-Time data processing (Streaming) Applications
- Spark's Data Pipeline based Approach for Machine learning

Reading list

- Tom White: Hadoop, the Definitive guide (sections)
- Matei Zaharia et al.: Learning Spark
- Holden Karau, Rachel Warren: High-Performance Spark

Assessment

- Start-of- the-class Quizzes (10%)
- Assignments (60%)
- Open book exam (30%)

Grading Policy

Students shall not miss more than 2 lectures and more than 1 seminar. Failing to do so will yield an administrative fail grade.

To pass, students will need to get at least 50% of the overall grade AND at least 50% of the exam. Failure to do so, will yield a Fail grade.

Course schedule and materials for each session

1. Big Data Overview
2. Basics of Cloud Computing using Amazon Web Services (AWS)
3. HDFS
4. The Hadoop Ecosystem, YARN, MapReduce
5. Apache Spark Overview
6. Spark internals
7. The Spark DataFrame and SQL API
8. Advanced Optimizations in Spark
9. Structured Streaming: Creating Continuous Applications
10. Spark ML: Machine Learning on Spark