Course Unit Descriptor

Study Programme: Computer Science – Master

Course Unit Title: Distributed Programming

Course Unit Code: CS712

Name of Lecturer(s): Miloš Savić

Type and Level of Studies: Master Academic Degree

Course Status (compulsory/elective): Elective

Semester (winter/summer): Summer

Language of instruction: Serbian (primary), English (secondary)

Mode of course unit delivery (face-to-face/distance learning): Face-to-face

Number of ECTS Allocated: 6

Prerequisites: Object-oriented programming 1, Operating systems 1

Course Aims:

The primary objective of the course is to familiarize students with distributed programming techniques and models focusing on scalable big data platforms and modern frameworks for data-intensive distributed computing.

Learning Outcomes:

Minimum: Successful students should be capable to develop simple distributed Java applications running on Hadoop clusters.

Desirable: At the end of the course it is expected that successful students deeply understand distributed computing programming models and are able to develop distributed Java applications for large-scale data processing in various domains.

Syllabus:

Theory

Introduction to distributed computing programming models. Overview of programming languages for distributed computing. The concept of big data and abstractions for scalable, fault-tolerant, data-intensive computing. Basics of MapReduce programming model (functional programming roots, mappers and reducers, MapReduce execution framework, combiners and partitioners). Distributed file systems. Introduction to Apache Hadoop, the Hadoop software ecosystem and the Hadoop cluster architecture. MapReduce design patterns and examples of MapReduce algorithms (counting, sorting, relational algebra operations, matrix multiplication, etc). MapReduce graph algorithms. MapReduce algorithms for large-scale information retrieval and data analysis. Processing rapid, high-speed data streams. Limitations and extensions of MapReduce and alternative programming models. Introduction to Apache Pig, PigLatin and other dataflow languages.

Practice

Practical programming tasks related to the development of distributed Java applications based on the MapReduce programming model and the Hadoop framework.

Required Reading:

Jimmy Lin and Chris Dyer. *Data-Intensive Text Processing with MapReduce*. Morgan & Claypool Publishers, 2010.
Tom White. *Hadoop: The Definitive Guide*, 4th Edition. O'Reilly Media, 2015.
Donald Miner and Adam Shook. *MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems*, 1st Edition. O'Reilly Media, 2012.

Weekly Contact Hours: 4	**Lectures:** 2	**Practical work:** 2

Teaching Methods:

Theoretical classes are based on the classical teaching model involving a projector. Presented algorithms, techniques and models are augmented with illustrative case studies implemented in Java. At practical exercises organized in computer labs, students have to individually solve practical programming problems related to the development of distributed Java applications running on Hadoop clusters. To approach the oral exam, students have to pass pre-exam obligations consisting of one theoretical test and two practical programming tasks.

Knowledge Assessment (maximum of 100 points):

Pre-exam obligations	points	**Final exam**	points
Test	20	Oral exam	40
Practical problems	40		

The methods of knowledge assessment may differ; the table presents only some of the options: written exam, oral exam, project presentation, seminars, etc.