Workshop

Big Data Analytics and Stream Processing on Apache Spark

Associate Professor Krešimir Pripužić, Ph.D.

Faculty of Electrical Engineering and Computing

University of Zagreb

29 April 2022, 12:00 - 1:30 PM (on-line)

Summary:

Companies collect enormous amounts of data about their customers, suppliers, and operations, while billions of connected devices on the Internet of things (IoT) and we as individuals additionally produce vast amounts of data. Therefore, we have witnessed an exponential growth in the amount of newly created data for more than a decade. We define Big Data as data that either contains greater variety, arrives in increasing volumes or is produced with higher velocity. Many different open-source platforms have been developed recently for dealing with mentioned challenges using cluster computing, such as the Apache Hadoop YARN (MapReduce2), Apache Spark, Apache Flink, Apache Storm, etc. Probably the most popular among them are Apache Hadoop YARN and Apache Spark. This lecture briefly presents the Apache Spark platform and then demonstrates its analytics and stream processing   capabilities.

Biography:

Krešimir Pripužić is an Associate Professor at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia (UNIZG-FER), where he leads the Data Streams Laboratory. He has been affiliated with the Department of Telecommunications at UNIZGFER since 2003. He received his diploma degree in electrical engineering with a major in telecommunications and informatics from the UNIZG-FER in 2003. In 2005 he started his Ph.D. studies at UNIZG-FER, which he successfully finished in 2010 by defending his dissertation. As a part of his Ph.D. studies, he spent academic year 2006-2007 at the Distributed Information Systems Laboratory at EPFL (Ecole Polytechnique Fédérale de Lausanne), Switzerland, as a scholarship holder of the Swiss Government scholarship for university, fine arts and music schools for foreign students. He has co-authored over 40 scientific journal and conference papers. His main research interests are large-scale distributed systems, algorithms and data structures, big data analytics, data stream processing and internet of things.