Big data is not just a buzzword

by 
Martin Sarnovský,
Assistant Professor,
Department of Cybernetics and AI


František Babič,
Head of Centre of Business Informatics,
Department of Cybernetics and AI


Peter Bednár,
Assistant Professor,
Department of Cybernetics and AI

In a fast-changing digital economy, the innovation drivers like automation, cloud computing, big data and artificial intelligence are poised to change the future of work. In response to the growing popularity of the big data topic, we designed a new subject entitled Technologies for big data processing for our IT-oriented study program Business Informatics at the Technical University of Košice.

Instead of developing new solutions in a vacuum, successful digital technologies require more sharing, customer-centricity and co-innovation. This transformation also requires strong partnerships between companies, and universities to create an innovative ecosystem capable of adapting rapidly and reacting to changes.

Business Informatics, an IT-oriented study program at the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering, Technical University of Košice, offers graduates a combination of informatics and economic knowledge and skills at all three levels (bachelor, master, PhD.). Our education is in line with current IT trends such as data analytics, big data, processing data from various Smart devices, cloud computing, web or mobile application development, User Experience, IT management, etc. We cooperate with several leading Slovak IT companies, members of BSC Forum or IT Valley Košice.

Big data is one of the key current trends in information technologies with enormous potential in various areas of real-world applications as well as in research. As this particular area is becoming more and more important, the issues with integration of these topics within the education of the data analysis and data mining areas are gradually increasing. A newly designed course entitled Technologies for big data processing focuses on the development, implementation and validation of new methods and forms of education in the area of processing and analysis of big data. We incorporated this subject into our study program three year ago.

The course objective is to provide students with the theoretical information and practical skills in big data processing, methods, approaches and technologies. Students will gain knowledge of basic concepts of distributed computing paradigms, the parallel and distributed programming methods, as well as distributed, NoSQL, in-memory and graph databases. The theoretical part mostly consists of lectures and seminars and is accompanied by practical lab sessions. During the semester, students attend several hands-on lab sessions. During these sessions they explore the platform for big data analysis and the technologies used in real-world applications, see the demonstration via various examples and learn to write their own code using the presented technologies.

The lab sessions' design represents the challenge from two different aspects. It relies on the need of resources and infrastructure, and, on the other hand, requires expertise in formulating the practical tasks. Choosing the right technologies is crucial, as the development of new technologies is taking a stride in this area. Setting up the proper infrastructure to handle the teaching process is also important.

The lab's infrastructure represents our private cloud based on a cluster virtualisation using Hyper-V technology. The cumulative computational resources consist of 156 CPU, 724 GB RAM, 6TB storage and 30 TB of network attached storage. On top of this infrastructure, a platform for big data analysis is deployed.

We decided to use the Apache Spark as it currently represents the framework which is suitable for different processing tasks ranging from batch processing of large volumes of data to real-time stream processing. The platform also includes a machine learning library (MLlib), tools to work with structured data (Spark SQL) as well as a graph computing library (GraphX), which makes it ideal for a wide range of data analytical tasks.

An important part of the lab package is the configuration of the lab environment itself. The first part of the lab sessions teaches the students how to configure and deploy such environment. Students should be able to properly configure the environment and explore the running platform – on their own workstations. During the course, we use a Spark standalone cluster created during one of the lab sessions by the students themselves. Then, during the multiple sessions, Apache Spark is introduced through examples, demonstrations and source code examples in Python language.

During the semester, students work on project-based assignments. In a group of 3-4 students, they need to solve a real-world data analytics task tasks using Apache Spark technology stack. Students are provided with the datasets, data description and task formulations needed to implement the project and to execute the implemented scripts in cluster environment.

Students can use this experience in their master theses in related areas. Some successful examples include: social network analysis (figure), traffic or industrial data analysis, or prediction of delays in air traffic.