Contribution ID: **120** Type: **Sectional**

Methodical aspects of training data scientists using the Data GRID in a virtual computer lab environment

Friday, 4 October 2019 09:45 (15 minutes)

Today it is crucial to train Data Scientists that serve as the bridge between cutting-edge technology and digital economy needs. It is essential to teach them to improve access to Big Data, analytics tools, innovative research methods. They should be able to design and deploy Data GRID clusters use and advise on such tools as machine learning, natural language processing, web scraping, big data platforms, and data visualization techniques. Virtual Computer Lab (VCL) provides a set of software and hardware-based virtualization and containerization tools that enable the flexible and on-demand provision and use of computing resources. The central methodical aspect of the VCL is the principle of self-organization, that makes the transition from a complex system of granular group security policies with a large number of restrictions to the formation of personal responsibility and respect for colleagues, which should be a solid foundation for strengthening and developing classical cultural values in the educational environment. Education in the VCL with integrated Knowledge Management System is the process of facilitating learning, or the acquisition of knowledge, skills, values, beliefs, habits. Educational methods include storytelling, discussion, teaching, training, directed research. Technology enhances relationships between teachers and students. When teachers effectively integrate technology into subject areas, teachers grow into the roles of adviser, content expert, coach. Technology helps make teaching and learning more meaningful and fun. Using VCL, students learn to design and deploy a Data GRID cluster based on Apache Hadoop, perform basic cluster administration tasks, upload real-world data. Based on the uploaded data, they study the main components of the cluster and the essential analytics tools. VCL allows us to train Data Scientists who can productively solve actual business and scientific problems in the field of Big Data. Data Scientists are integral to supporting both leaders and developers in creating better products and paradigms. Also, as their role in big business becomes more and more important, they are in increasingly short supply.

Summary

This paper discusses methodical aspects of training data scientists using the data grid in a virtual computer lab environment. Data scientists serve as the bridge between cutting-edge technology and digital economy needs. It is important to teach them to improve access to big data, analytics tools, and innovative research methods. They also should be able to design and deploy Data GRID clusters use and advise on such tools as machine learning, natural language processing, web scraping, big data platforms, and data visualization techniques and their application to relevant business needs and public policy issues. Virtual computer lab is a powerful innovative tool for training IT-professionals, created and successfully operated by the experts of the System Analysis and Control Department at the Dubna State University.

Keywords: virtual computer lab, virtualization, containerization, Grid, Data Grid, Hadoop, Map Reduce, Spark, HCatalog, Hive, Impala, Solr, Sqoop, Hue, cluster, data mining, distributed systems, mathematical modeling, education, data analytics, IT training, IT education, innovative education.

Primary author: TOKAREVA, Nadezhda (Dubna Univeristy)

Co-authors: Mrs CHEREMISINA, Evgenia (Dubna International University of Nature, Society and Man. Institute of system analysis and management); Mr BELOV, Mikhail (Dubna State Univeristy); POTEMKINA, Snezhana (Dubna State University); Dr KORENKOV, Vladimir (JINR)

Presenter: TOKAREVA, Nadezhda (Dubna Univeristy)

Session Classification: Innovative IT Education

Track Classification: Innovative IT Education