METHODICAL ASPECTS OF TRAINING DATA SCIENTISTS USING THE DATA GRID IN A VIRTUAL COMPUTER LAB ENVIRONMENT

M.A. Belov[1,a], V.V. Korenkov[1,2,3,b], S.V. Potemkina[1,c], M.V. Lishilin[1,d], E.N. Cheremisina[1,e], N.A. Tokareva[1,f], Y.A. Krukov[1,g]

[1] *System Analysis and Control Department, Dubna State University, Universitetskaya 19, 141980, Dubna, Russia*

[2] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, Joliot-Curie 6, 141980, Dubna, Russia*

[3] *Plekhanov Russian University of Economics, Stremyanny lane 36, 117997 Moscow, Russia*

E-mail: [a] belov@uni-dubna.ru, [b] korenkov@jinr.ru, [c] snezhanna@uni-dubna.ru, [d] lishilin@uni-dubna.ru, [e] chere@uni-dubna.ru, [f] tokareva@uni-dubna.ru, [g] kua@uni-dubna.ru

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, containerization, cluster, data mining, distributed systems, mathematical modeling, IT training, IT education, innovative education

Mikhail Belov, Vladimir Korenkov, Snezhana Potemkina, Mikhail Lishilin, Evgeniya Cheremisina, Nadezhda Tokareva, Yury Krukov

1. Introduction

The tasks of distributed data storage and processing, data mining and mathematical modeling based on these data are priorities within the agenda of the digital economy development program in the Russian Federation.

Today it is very important to train data scientists that 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.

The aim of an innovative IT education is the possibility of training specialists who can effectively solve such problems as conducting researches that explores methods to harness technology and innovation to advance equity, mobility and inclusion in cities, building innovative data products and pipelines to produce novel data that can help tackle pressing issues from a new angle, building systems and processes to collect, analyze, and combine multiple sources of data in novel ways.

2. Hardware and software tools for educational process

In order to provide students with the opportunity to design the Data GRID cluster for personal researches in the field of data analysis and mathematical modeling, we decided to replace physical computers with virtual machines in the virtual computer lab, which was established at the Institute of System Analysis and Control since 2007 by M. Belov.

The virtual computer lab (VCL) provides a set of software and hardware-based virtualization and containerizations tools that enable the flexible and on-demand provision and use of computing resources in the form of cloud Internet services with integrated knowledge management system based on the principles of self-organization, functioning as a homogeneous environment with elements of cognitive representation of internal operational resources based on visual models and partial automation of basic technological operations with expert system for carrying out research projects, resource-intensive computational calculations and tasks related to the development of complex corporate and other distributed information systems. The service also provides dedicated virtual servers for innovative projects that are carried out by students and staff at the Institute of System Analysis and Control.

The main features of a virtual computer lab are the principles of self-organization, which make 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 [1-9].

Data GRID is the general term which utilize the multiple sites or clusters for distribute the processing and storage among them, so the Hadoop HDFS is a method or a way for data grid implementation since many other Hadoop frameworks like MapReduce or Spark used for distributed data processing.

Traditional GRID computing is a processor architecture that combines computer resources from various domains to reach a main objective. In grid computing, the computers on the network can work on a task together, thus functioning as a supercomputer. Another way to look at is that GRID computing is now the traditional high-performance system with a flavor of MPI [10].

We look at GRID as a distributed system concept – a way to use computers distributed over a network to solve a problem. GRID is a group of physical machines connected to make a GRID Computer and Hadoop is the software running on these machines, therefore Hadoop is a subset of Grid computing.

In order to provide the ability to quickly deploy Data GRID clusters, we added new blade servers (to optimize the space they occupy in a server room) with SSD disks of increased wear resistance and increased RAM.

In order to minimize costs, we use the VMware vCenter technology platform with an integrated set of proprietary software tools, for the productive implementation of educational tasks.

The organization of an effective process for the goal-directed training of IT experts has demanded a speedy solution to the following problems: an often insufficient number of classroom hours for students to cover a necessary and sufficient set of practical exercises that help students learn complex information systems; on a typical personal computer with average capabilities it is impossible to get real practical experience working with multi-component information systems because the hardware requirements for such systems often go beyond what is offered on typical home, office and laptop computers; sometimes there are difficulties installing and supporting some information systems, and these problems cannot be solved without gaining experience about how to use such systems; the single-user license cost is too high, and in most cases, such a license is required only for the duration of the learning process.

3. The advantages of using the virtual computer lab in the educational process

The organization of an effective educational process for the goal-directed training of IT experts has demanded a speedy solution to the following problems: an often insufficient number of classroom hours for students to cover a necessary and sufficient set of practical exercises that help students learn complex information systems; on a typical personal computer with average capabilities it is impossible to get real practical experience working with multi-component Data GRID cluster because the hardware requirements for such systems often go beyond what is offered on typical home, office and laptop computers; the single-user license cost for some software components or professional technical support is too high, and in most cases, such a license is required only for the duration of the learning process.

The training of the «consumers» should be cut off in the process of the IT specialists' education, and we should spare no effort to training of the «creative doers». For this purpose, it is important to study the ways of creating the information systems from the scratch, paying attention to the configuring and adjustment of the equipment, connection and integration of all the necessary parts of the system without any help, and only after that to accomplish issue-oriented tasks.

The data scientist of the future is an expert, which has not only the fundamental scientific knowledge, but he is a promising engineer with an outstanding potential and is able to compose and make the capable data analysis solutions suitable for the project. Only the skilled professionals of this level can create the right conditions for the science development and its practical applications at an increasing rate.

All above-mentioned problems can be solved in the virtual computer lab, which has become not only the innovative tool for the training of the high skilled IT specialists, but also a demanded space for the technical cooperation between a final-year student and a potential employer. It gives an opportunity to show qualification in real time, and to present the employer's problem in the virtual format and try to solve it together, attracting the young minds and sometimes people with different ways of thinking, for example, the history of the neural network expansion and the idea of calculation of the back propagation errors, using the gradient descent method and so on.

A centralized management portal as well as a knowledge management system were created in order to manage the virtual computer laboratory. The need to create such a system was conditioned by the fact that students are able to learn about Data GRID clusters, so it is important to create a social network between all participants as well as to create an environment that allows pupils the opportunity to independently engage in such processes as the identification, acquisition, presentation, and use (distribution) of knowledge without the direct involvement of the instructor.

Methods of use (propagation) are directly related to storage methods and, consequently, the technological tools that may be used for the transmission of formal knowledge include knowledge bases with various search functionality; blogs, wikis, and social networks; "Wiki Textbooks" that allow all participants to collaboratively create and update educational content and exchange practical problems (including from real companies); as well as user blogs, forums, and group chat systems.

That is why the priority of the university is to create the most favorable conditions for the forming of the professional competence in IT, which will help the graduates to solve a wide range of the tasks, happening during all the stages of the Data GRID development, including the design itself. It is evident that to form the professional competence the students should do the following in order master a lot of literature, do many practical tasks and make research works on the modern data analysis systems, their deployment, maintenance and effective appliance for solving the problem-oriented tasks.

The main way to solve these problems has been to create a virtual computer lab that is able to solve the problem of insufficient computing and software resources and to provide an adequate level of technological and methodological support; to teach how to use cutting-edge technologies to work with distributed information systems on the example of Hadoop Data GRID Cluster; to organize group work with educational materials by involving users in the process of improving these materials and allowing them to communicate freely with each other on the basis of self-organizational principles.

4. The results of the educational process

Education is the process of facilitating learning, or the acquisition of knowledge, skills, values, beliefs, and habits. Educational methods include storytelling, discussion, teaching, training, and directed research. Technology can enhance relationships between teachers and students. When teachers effectively integrate technology into subject areas, teachers grow into roles of adviser, content expert, and coach. Technology helps make teaching and learning more meaningful and fun. Using Virtual Computer Lab students learn to design and deploy a Data GRID cluster based on Apache Hadoop software using most common topologies (Basic Horizontal topology, Federation topology, Monadic topology, Hierarchical topology, Hybrid Topology), perform basic cluster administration tasks, such as adding or removing hosts and service instances, changing the replication factor, adjusting the amount of allocated memory for execution containers, etc. Learners upload real-world data from various data sources into distributed HDFS file system, perform data rebalancing. Based on the uploaded data, they study the main components of the cluster and most important analytics tools (MapReduce, Spark and utility tools HUE, HCatalog, Hive, Impala, Pig Latin, Sqoop, Solr, Oozie, CDSW). For example, based on several tens of millions posts from technical forum, evaluate the popularity of programming languages, the effectiveness of moderation, the tonality of a statement on a given product, etc.

5. Conclusion

The results that we get specialists who can create Data GRID clusters and productively solve problems in corresponding application domains. Their jobs can focus on data management, analytics modeling, and business analysis. Data scientists can be real change-makers within an organization, offering insight that can illuminate the company's trajectory toward its ultimate business goals. Data scientists are integral to supporting both leaders and developers in creating better products and paradigms. And as their role in big business becomes more and more important, they are in increasingly short supply.

The Institute of System Analysis and Control has achieved in improving the educational process represent strategic foundations for overcoming perhaps one of the most acute problems in modern education: the fact that it tends to respond to changes in the external environment weakly and slowly.

It should also be emphasized that the virtual computer lab has helped us provide an optimal and sustainable technological, educational-organizational, scientific-methodological, and regulatory-administrative environment for supporting innovative approaches to computer education. It promotes the integration of the scientific and educational potential of Dubna State University and the formation of industry and academic research partnerships with leading companies that are potential employers of graduates of the Institute of System Analysis and Control.

References

[1] Belov M.A., Kryukov Y.A., Miheev M.A., Lupanov P.E., Tokareva N.A., Cheremisina E.N., Improving the efficiency of mastering distributed information systems in a virtual computer lab based on the use of containerization and container orchestration technologies, Sovremennye informatsionnye tekhnologii i IT-obrazovanie. 2018, T.14. №4.

[2] Belov, M.A., Krukov, Y.A., Mikheev, M.A., Tokareva, N.A., Cheremisina, E.N. Essential aspects of it training technology for processing, storage and data mining using the virtual computer lab, CEUR Workshop Proceedings 2267, pp. 207-212, 2018.

[3] Belov M.A., Kryukov Y.A., Lupanov P.E., Miheev M.A., Cheremisina E.N., Koncepciya kognitivnogo vzaimodeystviya s virtual'noy komp'yuternoy laboratoriey na osnove vizual'nyh modeley i ehkspertnoy sistemy, Estestvennye i tekhnicheskie nauki, 2018, №10, S. 27-36.

[4] Belov M.A., Lupanov P.E., Tokareva N.A., Cheremisina E.N. Kontseptsiya usovershenstvovannoy arhitektury virtual'noy komp'yuternoy laboratorii dlya effektivnogo obucheniya spetsialistov po raspredelennym informatsionnym sistemam razlichnogo naznacheniya i instrumental'nym sredstvam proektirovaniya, Sovremennye informatsionnye tekhnologii i IT-obrazovanie. 2017. T. 13. № 1. S. 182-189.

[5] Cheremisina, E.N., Belov, M.A., Tokareva, N.A., Grishko, S.I., Sorokin, A.V. Embedding of containerization technology in the core of the Virtual Computing Lab, CEUR Workshop Proceedings 2023, pp. 299-302, 2018.

[6] Belov M.A., Cheremisina E.N., Potemkina S.V., Distance learning through distributed information systems using a virtual computer lab and knowledge management system, Journal of Emerging research and solutions in ICT, 2016.

[7] Lishilin M.V., Belov M.A., Tokareva N.A., Sorokin A.V., Kontseptual'naya model' sistemy upravleniya znaniyami dlya formirovaniya professional'nyh kompetentsiy v oblasti IT v srede virtual'noy komp'yuternoy laboratorii, Fundamental'nye issledovaniya. 2015. № 11-5. S. 886-890.

[8] Belov M.A., Lishilin M.V., Tokareva N.A., Antipov O.E., Ot virtual'noy komp'yuternoy laboratorii k upravleniyu znaniyami. Itogi i perspektivy, Kachestvo. Innovatsii. Obrazovanie. 2014. № 9 (112). S. 3-14.

[9] Cheremisina E.N., Belov M.A., Lishilin M.V., Integratsiya virtual'noy komp'yuternoy laboratorii i znanievogo prostranstva - novyy vzglyad na podgotovku vysokokvalifitsirovannyh it-spetsialistov, Sistemnyy analiz v nauke i obrazovanii. 2014. № 1 (23). S. 97-104.

[10] Foster, Ian., Kesselman, Carl The Grid2: Blueprint for a New Computing Infrastructure. — Morgan Kaufmann Publishers. — ISBN ISBN 1-55860-475-8, 2003.