SPECIALISTS TRAINING IN BIG DATA USING DISTRIBUTED ARCHITECTURAL SOLUTIONS SERVICES

M. Batura[1]
Ректор Белорусского государственного университета информатики и радиоэлектроники (БГУИР), доктор технических наук, профессор, академик Международной академии наук высшей школы, заслуженный работник образования Республики Беларусь

S. Dzik[2]
Проректор по учебной и воспитательной работе БГУИР, кандидат физико-математических наук, доцент, академик Белорусской инженерной академии, Республика Беларусь

I. Tsyrelchuk[3]
Заведующий кафедрой проектирования информационно-компьютерных систем БГУИР, кандидат технических наук, доцент, Республика Беларусь

V. Komlitchenko[4]
Заведующий кафедрой экономической информатики БГУИР, кандидат технических наук, доцент, Республика Беларусь

E. Unuchek[5]
Старший преподаватель кафедры экономической информатики БГУИР, Республика Беларусь

[1] *Belarusian State University of Informatics and Radioelectronics, rector@bsuir.by,*
[2] *Belarusian State University of Informatics and Radioelectronics, sdick@bsuir.by,*
[3] *Belarusian State University of Informatics and Radioelectronics, tsyrelchuk@bsuir.by,*
[4] *Belarusian State University of Informatics and Radioelectronics, v.komlitchenko@gmail.com,*
[5] *Belarusian State University of Informatics and Radioelectronics, e.unuchek@gmail.com*

Application integration and business process improvement connected with Distributed Computation Services and Predictive Analytics can bring new concurrent capabilities. This article considers organization of Big Data learning and Big Data experience exchange in Belarus. The experience of the organizing and the use of virtual distributed computing infrastructure for Big Data trainings in BSUIR is proposed.

Application integration and optimization of distributed business processes require analysis of transferred data streams and Big Data opportunities to obtain new competitive advantages, revealing hidden knowledge using special analytics. The problem is reduced

to the integration of different interfaces for connecting external systems based on different protocols, as well as predefined routes messages. Such integrated transactions will allow to extract and provide key information in the field of complex computations, interpreting large amounts of information and decision management solutions in business problems.

At the same time, Big Data analytics is not self-sufficient. Using Big Data methods and technologies could lead to the minimization of human intervention in the business processes in the future within the areas where natural limit of human memory capacity and speed interferes with the actions performance, and the actions themselves are based on large amounts of data and computations, where decision making is based on special analytics and new knowledge elements derived from Big Data. Business Process Management (BPM) and Big Data combination will not only increase the efficiency of business processes of organizations, but also will improve their quality, reduce expenses, increase flexibility, build new business models.

The appeal of this subject area takes one of the first most popular places among the areas of information technology development. It is noteworthy that recently there is a gradual increase of interest to Big Data technologies and Advanced analytics in Belarus. These technologies are often seen as an alternative to the traditional. The main generator of this evolution is due to the highlight of this particular subject area within the framework of specialized conferences (e.g., XXII International Forum “TIBO-2015” with the theme “Databases and data, call-centers, Big Data”) [1], to highly specialized training and to special public communities, supported by the leading companies-residents of the High Technologies Park.

It is notable that there are several communities connecting people with the interests in the field of Big Data in Belarus now. The most famous are “Belarus Big Data User Group” and “Data Talks Belarus”. Both communities include several hundred stakeholders and offer opportunities for professional dialogue with experts, information and knowledge about the latest experience in applying methods of analytics and tools to solve practical problems. Among the regular participants of the events organized by these communities one can find analysts, researchers, project managers, and all those who use or is going to use in their work analysis or complex mathematical calculations on large amounts of data either for reporting and decision-making or for information systems creation.

Two most numerous groups among the community members are clearly defined.

The first group consists of people with an interest to analytics, methods and tools of data mining and machine learning algorithms.

The second group of participants shows a particular interest to the engineering and technical component of technologies providing Big Data functioning. Their research interests are the methods, technologies and tools of distributed batch, real-time and stream-

ing data processing, high-performance, scalable data storage and access technologies, etc.

Both in Belarus and in the world there is an acute shortage of professionals in this particular area. This is primarily due to the high qualification requirements and an extremely wide range of competences, which such a specialist should have. This is the knowledge in databases, statistics, data visualization techniques, machine learning algorithms and artificial intelligence methods, algorithmization, design and programming using unix-like systems and computer networks, English language and other skills and knowledge.

The beginning of the implementation of analytical and Big Data solutions into universities educational process were activities related to establishment of the relevant specialties for MA courses to train specialists within the second stage of higher education at BSUIR and BSU and conduct special training courses for specialists at the same universities. One of the first courses on data analysis in Belarus were the courses within the project "Yandex School of Data Analysis" [2] at the Department of Applied Mathematics and Informatics, BSU (now the courses are also held at the Faculty of Computer Systems and Networks, BSUIR), Data Vita – "Introduction to Data Science" [3] and "Big Data and Predictive Analytics" were organized by BSUIR and BEZNext company, USA. Particular interest is in organizing training in areas related to processing large amount of data, held at BSUIR.

So in October 2013 in collaboration with the American company BEZNext BSUIR organized recruitment and training as part of the training on "Big Data and Predictive Analytics". The training was in English in evenings, it was conducted by the leading US experts in the field of Big Data using remote networking communications. As a result of the six-month training 6 people defended their final projects with honors.

In the autumn semester a special training on Java-technology and work with large amount of data ("Introduction into Big Data") was held as part of the activity of the joint laboratory BSUIR-IBA. During the courses, students had the opportunity to deal with IBM advanced technologies in the field of Big Data – IBM Info Sphere Streams and IBM Info Sphere Big Insight [4].

In February 2015, BSUIR in collaboration with BEZNext organized iterative course on "Big Data and Predictive Analytics" in Russian language. The first open lecture was held as a videoconference for BSUIR professors, graduate students and undergraduates gathered more than 150 people. But the structure and content of the course have been significant changed due to the appearance and development of new solutions on the Big Data market affected by demands and technological requirements of BEZNext customers.

This course includes both lectures and laboratory classes on the following issues:

• statistical methods in the data analysis;

• correlation and regression analysis;
• time series analysis;
• data preparation and cleaning;
• machine learning algorithms;
• Map / Reduce;
• data processing with Spark
• NoSQL databases;
• ETL tools;
• data visualization tools and techniques.

One of the challenges that the participants of the training faces in the practical tasks performance was the infrastructure for applications development, executing and debugging. Partially this problem was solved by running virtual machines on the client's workstations. However, this method just simulates a real environment and contains a number of restrictions referred to the scaling and capability of large amounts of data processing.

To solve that problem there was a solution to create a virtual lab and provide access to the trainees of the courses on "Big Data and Predictive Analytics". Virtual Lab is based on Amazon Web Services (AWS). It is a set of computer web services that make computing cloud platform provided by Amazon. Virtual Laboratory (Big Data Lab) simulates most of the systems of Apache Big Data projects stack [5] starting from messages processing (Kafka) and ending with a batch processing (Map / Reduce) and real time processing (Storm, Spark, Tez) deployed and configured to be run in a distributed environment to achieve parallelism. Additionally, a distributed NoSQL database (Cassandra) is configured and works.

Big Data Lab architecture includes five infrastructure nodes and one control node (Fig. 1).

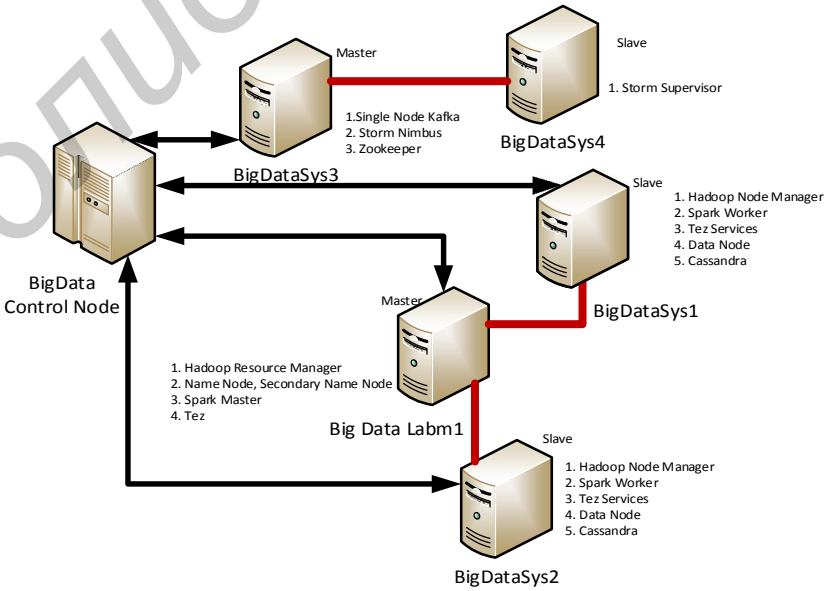

Fig. 1. Architecture of a virtual laboratory Big Data Lab

Three nodes contain Hadoop, Spark (as a separate installation), Cassandra and Tez, the remaining two nodes – contain Storm and Kafka. Kafka can run a set of brokers on a single machine. Storm is configured as master/slave to the Zookeeper service which is responsible for the interaction between them.

BigDataSys3 and BigDataSys4. Both these nodes are the nodes for messages processing. Apache Kafka is installed on BigDataSys3 with a set of brokers working simultaneously. Test topics were created. Storm daemons "Nimbus" and "UI" are also launched on the BigDataSys3 node. Topology of examples "Word Count" and "Exclamation" are launched. BigDataSys4 node is a supervisor for distributed processes environment Storm with 10 processes (workers) configured for the topologies.

BigDataLabm1. As shown on the Pic. 1, this node is the master of the cluster. YARN is deployed on it. BigDataSys1 and BigDataSys2 nodes are slaves. We can run either standard Hadoop benchmarks like "Word Count" and "Terasort" or at the same time standard Spark-tests such as "Spark PI" and "Spark word count from shell" on this machine. To run standard tests for Map Reduce and Spark one need to be connected to this node. Basic Hadoop and Spark commands can be executed from this node, for example, copy data to HDFS, retrieve a list of Hadoop jobs etc.

BigDataSys1 and BigDataSys2. BigDataSys1 and BigDataSys2 act as slave-nodes of BigDatLabm1. Also, Cassandra is deployed on both these nodes. Since there is no master/slave in Cassandra clusters, both of the nodes are the slave-nodes of Cassandra cluster named "BigDataLab". All Cassandra-cluster nodes are located in one data-center. Only one server executes seed-node functions.

BigDataControlNode. This server is the only way to get into a virtual lab and to run test applications and applications which are being developed.

Applying this approach to organizing of the environment for debugging and execution of either individual laboratory works and final project allowed to focus on practical problems solution (adjustment and installation of infrastructure requires deep knowledge and understanding of unix-like systems and computer networks administration), and also provided a high-performance, scalable computation infrastructure. The necessity of a separate rate per hour for the cluster work and high qualification requirements for the specialist who can set up and maintain performance of such a training computing environment are the difficulties and restrictions for the its usage.

A list of additional positive aspects should be mentioned.

The experience in creating a distributed virtual training laboratory shown on the example of training courses on Big Data at BSUIR reveals that there is no need in high initial investments in the network and server infrastructure, exclusion of incompatibility problems of equipment from different manufacturers, and system availability. Moreover, it can be mentioned that such a virtual laboratory is beyond the boundaries of the university, making a continuing learning process for students. Anywhere in the world where

there is an access to the Internet, trainees can distantly and continuously interact with a virtual lab environment that fully meets the concept of continuing distance learning.

Literature

1. Subject of the XXII forum Tibo [digital resource]: http://www.park.by/post-888/?lng=ru – access date: 18.05.2015.

2. Yandex School of Data Analysis [electronic resource]: https://yandexdataschool.ru/about/branches/minsk – access date: 18.05.2015.

3. Data Vita: the science of data in theory and practice [digital resource]: http://www.datavita.net/ – access date: 18.05.2015.

4. Presentation of IBM certificates was held at BSUIR [digital resource] – режим доступа: http://www.bsuir.by/online/tnj2/one_article.jsp?PageID=94597&resID=100229&lang=ru&tnj_type=2&rid=102243&tnj_id=11875&pid=100229 – access date: 18.05.2015.

5. Apache Big Data Projects [digital resource]: http://projects.apache.org/indexes/category.html#big-data – access date: 18.05.2015.