DESIGNING CLOUD INFRASTRUCTURE FOR BIG DATA IN E-GOVERNMENT

Jelena Šuh*

Faculty of Organizational Sciences Belgrade, Jove Ilića 154, 110000 Belgrade, Serbia
jelena.suh@gmail.com

Vladimir Vujin

Faculty of Organizational Sciences Belgrade, Jove Ilića 154, 110000 Belgrade, Serbia
vujin@elab.rs

Dušan Barać

Faculty of Organizational Sciences Belgrade, Jove Ilića 154, 110000 Belgrade, Serbia
dusan@elab.rs

Zorica Bogdanović

Faculty of Organizational Sciences Belgrade, Jove Ilića 154, 110000 Belgrade, Serbia
zorica@elab.rs

Božidar Radenković

Faculty of Organizational Sciences Belgrade, Jove Ilića 154, 110000 Belgrade, Serbia
boza@elab.rs

Abstract

The development of new information services and technologies, especially in domains of mobile communications, Internet of things, and social media, has led to appearance of the large quantities of unstructured data. The pervasive computing also affects the e-government systems, where big data emerges and cannot be processed and analyzed in a traditional manner due to its complexity, heterogeneity and size. The subject of this paper is the design of the cloud infrastructure for big data storage and processing in e-government. The goal is to analyze the potential of cloud computing for big data infrastructure, and propose a model for effective storing, processing and analyzing big data in e-government. The paper provides an overview of current relevant concepts related to cloud infrastructure design that should provide support for big data. The second part of the paper gives a model of the cloud infrastructure based on the concepts of software defined networks and multi-tenancy. The final goal is to support projects in the field of big data in e-government.

Keywords: big data, cloud computing, e-government.

1 Introduction

The development of information and communication services has great impact on the occurrence of large amounts of data, but also the need to access that data and perform analysis in real time. For this reason there is a need to create a simple network infrastructure, which will also meet the requirements in terms of scalability, security and availability.

There are systems in e-government where big data emerges and cannot be processed and analyzed in a traditional manner due to its complexity, heterogeneity and size. Taking into account the fact that this data has different structure and formats, and are often time-sensitive, special attention is paid to new technologies for storing and processing big data, and tools for big data analysis. Big data is one of the new trends in IT that on the technical side gives the possibility for these data to be used for real time analysis. In order to implement big data in e-government network infrastructure needs to be redesigned.

Main subject of this paper is the design of the cloud infrastructure for big data storage and processing in e-government. The goal is to analyze the potential of cloud computing for big data infrastructure and benefits that this approach can bring to e-government services. We present a network infrastructure model for processing and analyzing big data in e-government.

This paper points out the importance of the research in the field of big data and finding solutions to handle large amounts of data since e-government currently does not make sufficient use of big data technologies. The basic concepts of big data are first described. Then Hadoop framework for distributed processing of large amounts of data, as well as Map Reduce principle for big data search is presented. Having introduced big data concepts, the remainder of the paper is organized as follows. In the next section we provide an overview of principles and technologies for cloud infrastructure design. Software-defined networks and Savanna project, which aims to facilitate the integration of Hadoop and OpenStack technology, are described. We then review and discuss application of big data technologies in e-government. Next section represents a model of the cloud infrastructure for big data in e-government based on the concepts of software-defined networks. Last section represents conclusion with a discussion of key observations and implications for future research.

2 Big data definition and concepts

Big data is a term which refers to a large amount of data which cannot be processed and analyzed in a traditional manner, due to their complexity (Liu, 2013). For a big data description 3V model (Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, 2011) is often used. This model points out three characteristics: the amount of data (volume), processing speed (velocity) and a variety of data types (variety). The amount of data in big data is measured in terabytes, and that is the reason why special

attention should be given to data storage. Another important big data characteristic is fluctuations in the amount of data (Zikopoulos, Eaton, deRoos, Deutsch, & Lapis, 2012). An additional requirement is the data processing speed, since the data is often time-sensitive and require rapid transfer and analysis. A particular problem is the fact that the data are not structured and that they are in different formats: text, audio, video, log files, etc. A newer definition adds the necessity of applying new ways of data processing in order to improve decision-making and optimization process (Beyer & Laney, 2012).
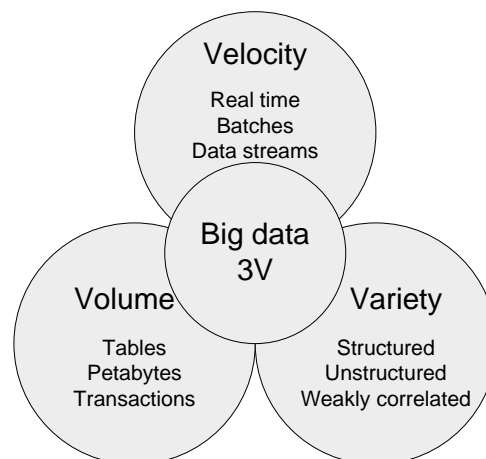
Figure 1. 3V big data definition

Bearing in mind the fact that big data are mainly unstructured data, it is necessary to apply new principles of data storing, which are different from the traditional, which use relational databases. The concept, which has become indispensable when it comes to big data, is NoSQL (Strozzi).

It should be noted that the amount of digital data in 2011th amounted to about 1.8 ZBy (1.8 trillion GBy) (Bakshi, 2012), and it is clear that there is a need for further research of big data.

3 Review of technologies

3.1   Cloud computing

Cloud computing is an area of computer science in which highly scalable IT capacity is provided in the form of services delivered via the Internet to numerous external users (Sultan, 2010). It is a model for enabling on-demand network access to shared computing resources (servers, storage, applications etc.) that can be rapidly provisioned and released with minimal management effort (Mell & Grance, 2011).

The advantages of using cloud computing are multiple (Watson, 2009): cost reduction and efficient use of resources have positive impact on development. At the same time cloud computing allows IT to focus on delivering IT services.

Main disadvantages are related to privacy and security issues and lack of legal regulative in this area.

This cloud model is composed of three service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a service (SaaS). There are four cloud deployment models (Jin, et al., 2010):

- Private cloud – internal cloud; network architecture within company or organization
- Public cloud – external cloud; on-demand resources are allocated and delivered via web services using Internet
- Hybrid cloud – combination of private and public cloud
- Community cloud – cloud infrastructure is shared among multiple organizations or companies

One of the main challenges in cloud computing environment is network infrastructure management. OpenStack is an open-source management tool for IaaS cloud computing infrastructure (OpenStack). OpenStack project is a collaboration of a large number of developers and companies whose goal is to create an open standard cloud computing platform for public and private cloud environments. This technology consists of many related projects which control different network resources via control panel (dashboard), a command line or a RESTful API (Fifield, et al., 2014).

Neutron or OpenStack Networking is one of OpenStack projects that have the task to provide the Network-as-a-Service (Neutron's Developer Documentation). The goal is to design virtual networks in a simple way without knowledge about complete underlying network infrastructure.

Simple implementation, scalability, and a range of additional features contribute to the popularity of OpenStack project that can be used in a number of networks. Multi-tenancy concept is one of the most important features because several tenants can use the same cloud infrastructure and share network resources. They can create completely isolated end-to-end network topologies based on specific requirements. Tenants can be administrators, citizens, but also e-government services and applications. In order to implement Network-as-a-Service with custom forwarding rules it is necessary to introduce the concept of software-defined networks.

3.2   Software-defined networks

Software-defined networks (SDN) have a central position in the process of design of big data network infrastructure as they are the transport medium for the transmission of large amounts of data. It is a new concept in computer networks, which allows network and service management abstracting the lower layers of the network infrastructure (Open Network Foundation) (Nadeau & Pan, 2011). Main characteristics of SDN networks are centralized management, and a complete separation from network elements, using open standard

interfaces for communication. OpenFlow protocol, a first interface for communication between the control and traffic forwarding layers, is an important factor in the development of SDN concept (McKeown, et al., 2008).

SDN architecture defines three layers: infrastructure, management and application (Software-Defined Networking: The New Norm for Networks, ONF White Paper, 2013). Traffic forwarding is realized at the infrastructure layer using different network elements and devices. Network control and monitoring is a function of the control layer and user applications are placed at the application layer.

SDN networks are based on the following principles (Muglia, 2013):

- Clear separation of network software in 4 layers (plain): management, services, control and forwarding.
- Centralization of certain aspects of the management, and control plane services in order to achieve simplicity of network design and reduce costs.
- The application of cloud computing to achieve flexibility.
- Creating a platform for network applications and services and integration with management systems.
- Standardized protocols in order to achieve interoperability and support for multi-vendor environments.
- Applicability of SDN principles to all networks and network services.

SDN concept is a good solution for big data infrastructure because it provides scalability, great flexibility and support for different APIs so many different services can be implemented.

3.3 Map Reduce

Google has developed a framework, Map Reduce, for processing large amounts of data using a large number of processors (Dean & Ghemawat, 2004). The processing is divided into two phases: map and reduce, where each stage has input and output parameters in the form of key/value pairs. User defines map function, and as a result of processing a number of key/value pairs are received. Reduce function is then applied to the resulting key/value pairs and extracts all the values with the same key.
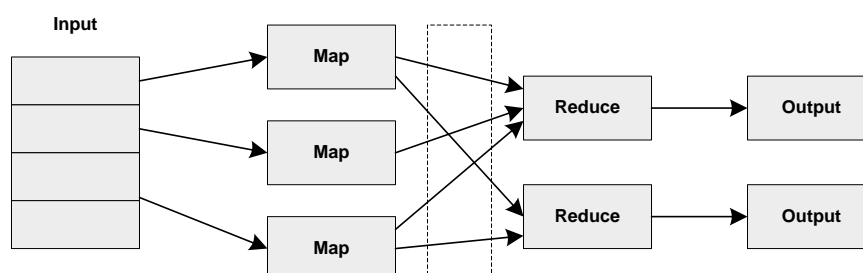

Figure 2. Map Reduce

An important feature of this method of data processing is that map and reduce functions can be executed in parallel, which enables simple implementation in the cluster.

3.4   Hadoop

Apache Hadoop is an open-source software framework for big data storage and processing in a cluster-based infrastructure (Hadoop, 2014). Hadoop uses Map Reduce approach.

Hadoop framework modules (Hadoop Project Description, 2014):
*   Hadoop Common – libraries and functions that support the other Hadoop modules
*   Hadoop Distributed File System (HDFS) – distributed file system
*   Hadoop YARN – job scheduling and cluster resource management
*   Hadoop MapReduce – system for big data parallel processing

Hadoop Map Reduce framework has master/slave architecture. There is one master server or jobtracker and more slaves or tasktrackers. Communication between users and framework is achieved through a jobtracker. Jobtracker receives map/reduce requests which are then processed on first-come/first-served basis. He is also responsible for the allocation of map/reduce tasks to tasktracker nodes.

Apache has created several Hadoop-related projects:
*   Ambari - web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters
*   Hbase - scalable, distributed database that supports structured data storage for large tables
*   Hive - data warehouse infrastructure that provides data summarization and ad hoc querying
*   Pig - high-level data-flow language and execution framework for parallel computation

Thanks to its features, such as scalability and flexibility, Hadoop project is supported by a large number of companies (Amazon Elastic MapReduce).

3.5   Big data analytics

Big data analytics is a term which refers to process of big data examination in order to discover and extract meaningful business value and to turn big data into actionable insights. Getting value out of big data is a complex process since huge volumes of data cannot be analyzed in a traditional manner. High-performance analytics must be used for faster and more accurate decision making. High-performance data mining, text mining, predictive analytics, forecasting and optimization on big data can be used. Historical data analysis is important but real-time processing as well. Final goal is to publish different reports depending on user requests and to visualize big data.

Big data analytics enables quicker response to market trends and mass customization of services. It can provide solution to different problems in domain of video analytics, click stream analytics, social media analysis, financial analytics, fraud detection etc.

3.6   OpenStack and big data

Sahara project was launched with the idea to implement Hadoop clusters in the cloud environment based on OpenStack technology (Sahara, 2014). The aim is to enable users to easily create a Hadoop cluster by defining a few parameters, such as Hadoop version, cluster topology, hardware characteristics of nodes and so on. Management is accomplished via REST API, and there is a user interface within the OpenStack dashboard.

Sahara supports the concept of templates to simplify the configuration process. In order to make easier the process of implementation of certain versions of Hadoop, or distribution in different network topologies and with support for a variety of tools for monitoring and control, Sahara supports the concept of plug-ins.

4 Big data in e-government

E-government systems are complex and have many services for citizens and enterprise. The data volume in these systems is measured in terabytes so we can speak about big data. In order to provide better services and support for future applications for citizens and business users e-government must consider implementation of big data technologies.

There are several big data projects successfully implemented in big companies such as Amazon, Yahoo, Facebook, Adobe etc. (Hadoop Wiki PoweredBy, 2014). Many of these use cases can be implemented in e-government. Big data and Hadoop, as most significant big data framework, has already applied in the following areas (Hadoop Use Cases and Case Studies): data storage, health care, education, retail, energy, logistics, image/video processing, travel, financial services, and politics.

There are several big data use cases especially significant for e-government. Social network analysis is very important bearing in mind social network popularity. Advanced analytic tool can analyze unstructured data from social media and determine user sentiment related to particular issue. In order to determine the effectiveness of different marketing campaigns big data can be used to improve accuracy of analysis.

Health care is another area where big data can make significant improvements. Beside storing and processing medical records big data analytics can help hospitals to further personalize patient care (Groves, Kayyali, Knott, & Van Kuiken, 2013).

Smart city concept is based on intelligent management and integrated ICTs and requires active citizen participation. It is becoming more and more popular and implies usage of large

number of different sensors. Many biological and industrial sensors generate large volumes of data so sensor data analysis is domain where big data must be used.

Intelligent transportation systems are applications whose main task is to provide services related to traffic management. Another important characteristic is simplification of traffic information placement towards citizens (The Case for Smarter Transportation - IBM Whitepaper, 2010). Auto-navigation, e-payment, smart parking and smart crossroad are just some of the examples where large amount of data are present so big data technologies can make significant improvements. Public safety services can use big data analysis for image/video processing for video surveillance which is very important for transportation systems.

Big data technologies can also help in fraud detection analyzing users' behavior, historical and transactional data. Goal is threat prediction and prevention looking for patterns and anomalous activity. In health care domain this can ensure that eligible citizens receive benefits. General security can be significantly improved using big data analytics for crime prediction and prevention and in emergency situations this approach can refine disaster response and information-collecting mechanisms.

Implementation of big data technologies can contribute to overall e-government effectiveness. One significant aspect is cost reduction, but equally important is better interaction between citizens and business users and government.

5 Cloud infrastructure model for big data in e-government

E-government systems are complex and special attention should be given to network infrastructure design since it is necessary to create scalable and secure environment as a base for numbered e-government services. The cloud computing infrastructure is the solution that can adequate fulfill these requirements. In order to provide certain network programmability SDN concepts must be used too. Different e-government services require access to large amounts of data, so big data technologies must be implemented.

The main requirement that the network infrastructure must meet is the dynamic configuration of network resources based on user-defined requirements, which will vary depending on the nature of the service or application in use. In addition to basic network services for communication it is necessary to provide integration with external institutions and support for big data. Guidelines for network infrastructure design are: support for all available e-government services, on-demand resource reservation and big data support. Equally important are simple network management and new services and applications development without complex knowledge of underlying network.

Figure 3 shows conceptual model of the big data infrastructure. Citizens or enterprise users use certain service via e-government portal. Users' requests are sent to big data infrastructure in order to be processed. Big data infrastructure has databases which are not centralized. Integration with databases of different ministries, government agencies and local authorities is realized. After big data processing and analysis user gets adequate response and data visualization via e-government portal.
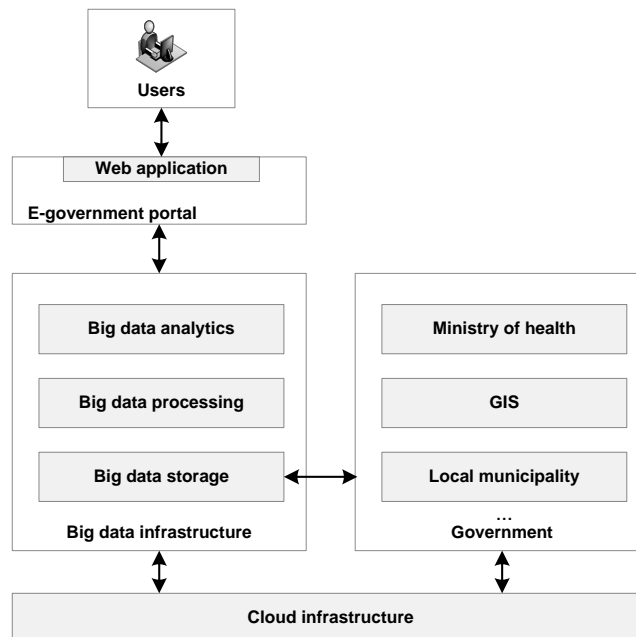

Figure 3. Conceptual model of the big data infrastructure

The big data infrastructure is based on cloud computing and SDN concepts. Hadoop framework is used for big data storage and processing. Integration with OpenStack cloud is realized using Sahara controller and HDP plug-in. The cloud-based model of the big data infrastructure is shown in Figure 4. The advantage of this model is the use of open-source software so implementation costs for this solution are significantly reduced.
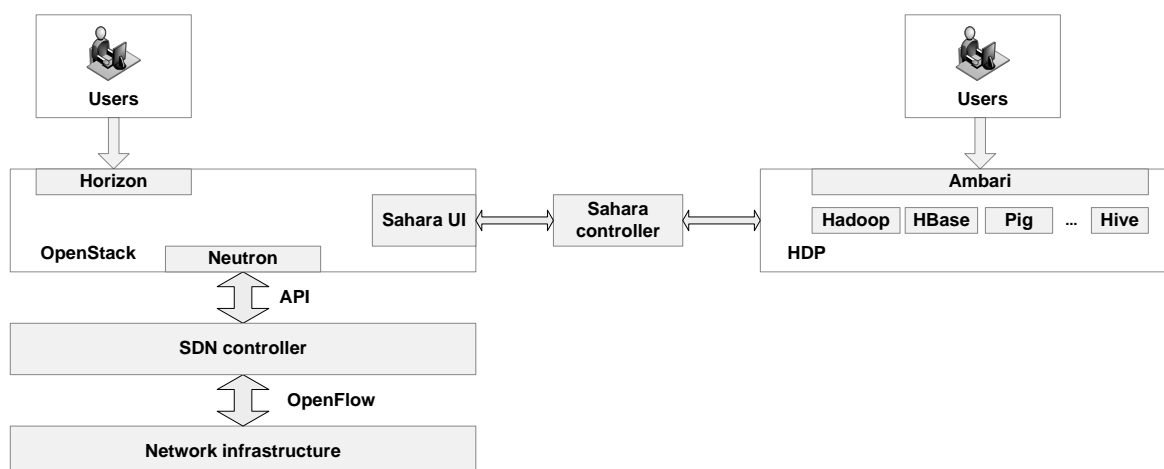

Figure 4. The model of the big data infrastructure for cloud-based e-government

Users send requests to the network using an appropriate interface. The request is then forwarded to SDN controller, which is a central place in the architecture. SDN controller has the task to configure network infrastructure and define the rules for traffic forwarding based on the specific service. Mentioned phases are shown in the Figure 5.
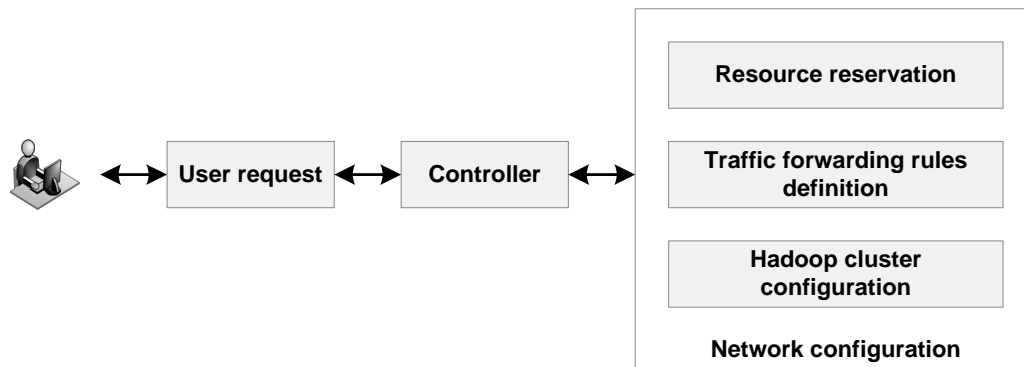

Figure 5. User's interaction phases

For the evaluation of the model the big data infrastructure prototype in e-learning is developed in Laboratory for E-business at the Faculty of Organizational Sciences in Belgrade. This network infrastructure can provide on-demand big data infrastructure for different users (students, teachers, laboratories, etc.). The analysis of data from Moodle LMS (students' data, their activities, logs etc.) is implemented within the Hadoop infrastructure in order to increase efficiency and improve the quality of teaching. Figure 6 shows conceptual diagram for developed infrastructure prototype in e-learning.
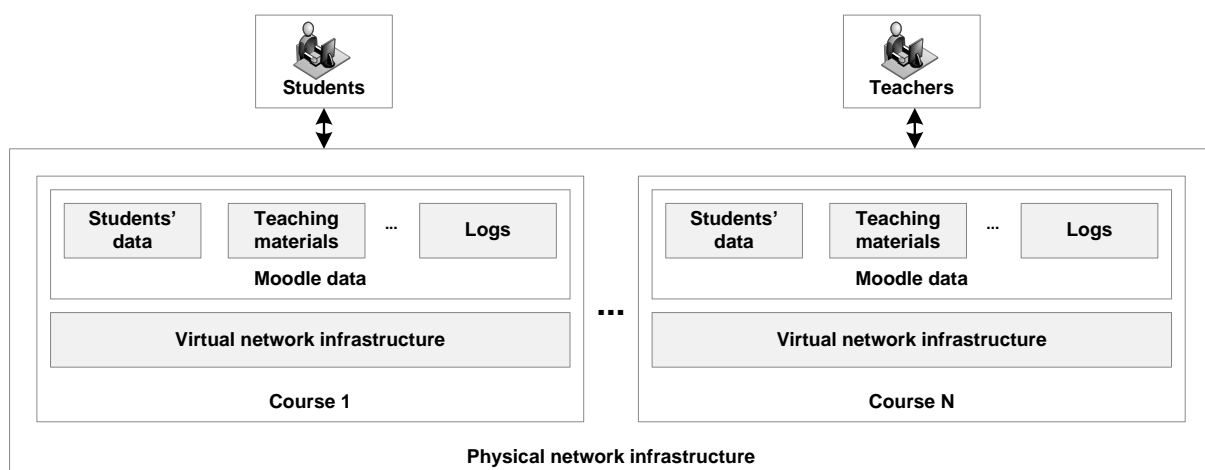

Figure 6. Conceptual diagram for infrastructure prototype in Laboratory for E-business

Following the example of the prototype the same model can be applied to any e-government segment where big data resources can be provided as a service to users and institutions depending on organizational policies and all that using common network infrastructure. This model is based on open-source technologies so main advantages of the proposed model are efficiency and cost-effectiveness. This relatively simple network infrastructure can fulfill different requirements and final goal is to improve the usage of existing data. General issues

such as privacy and data protection are present here too, so future research should be devoted to these problems.

6 Conclusion

The development of new information services and technologies, especially in domains of mobile communications, Internet of things, and social media, has led to appearance of the large quantities of unstructured data. Analysis of these data is of great importance not only in the internal business processes, but also in science, government, education, and all of society (Bryant, Katz, & Lazovska, 2008). E-government systems have large amounts of data that cannot be processed and analyzed in a traditional manner. The basis for the development of software solutions and tools for the big data analysis is a network infrastructure, which should provide a scalable, reliable and secure environment for storing and processing data.

In this paper, we presented main concepts of big data and principles and technologies for cloud infrastructure design. We developed a model for big data infrastructure in e-government based on cloud computing and SDN concepts. This model can significantly improve efficiency in e-government for citizens and enterprise users. Future research should be focused on improving current model, exploring possibilities for mobile services support and QoS implementation.

References

1. *Amazon Elastic MapReduce*. Retrieved from http://aws.amazon.com/elasticmapreduce/
2. Bakshi, K. (2012). Considerations for big data: Architecture and approach. *IEEE Aerospace Conference.* Big Sky, MT.
3. Beyer, M. A., & Laney, D. (2012). *The Importance of 'Big Data': A Definition.* Gartner.
4. Bryant, R. E., Katz, R. H., & Lazovska, E. D. (2008). *Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society.* Retrieved from http://www.cra.org/ccc/files/docs/init/Big_Data.pdf
5. Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *OSDI.*
6. Fifield, T., Fleming, D., Gentle, A., Hochstein, L., Proulx, J., Toews, E., & Topjian, J. (2014). *OpenStack Operations Guide.* O'Reilly Media.
7. *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data.* (2011). Retrieved from http://www.gartner.com/newsroom/id/1731916/
8. Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). *The big data revolution in healthcare.* Retrieved from

http://www.google.rs/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8
&sqi=2&ved=0CDAQFjAB&url=http%3A%2F%2Fwww.mckinsey.com%2F~%2Fmedia
%2Fmckinsey%2Fdotcom%2Fclient_service%2Fhealthcare%2520systems%2520and%25
20services%2Fpdfs%2Fthe_big_data_revolution_in_healthcare.ashx&ei=pbxzU7CLCuby
ygPgh4CABw&usg=AFQjCNFIpUfpqQQcDlpSXHwco5atclHrag&bvm=bv.66699033,d.
bGQ

9. *Hadoop*. Retrieved from http://hadoop.apache.org/

10. *Hadoop Wiki PoweredBy*. Retrieved from http://wiki.apache.org/hadoop/PoweredBy/

11. *Hadoop Project Description*. Retrieved from
    http://wiki.apache.org/hadoop/ProjectDescription/

12. *Hadoop Use Cases*. Retrieved from
    http://hadoopilluminated.com/hadoop_book/Hadoop_Use_Cases.html

13. Jin, H., Ibrahim, S., Bell, T., Gao, W., Huang, D., & Wu, S. (2010). Cloud Types and
    Services. In *Handbook of Cloud Computing* (pp. 335-355). Springer.

14. Liu, L. (2013). Computing infrastructure for big data processing. *Front. Comput. Sci.,
    7*(2), pp. 165-170.

15. McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., . . .
    Turner, J. (2008). OpenFlow: Enabling Innovation in Campus Networks. *ACM
    SIGCOMM Computer Communication Review, 38*(2), pp. 69-74.

16. Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing*. National
    Institute of Standards and Technology, Special Publication 800-145.

17. Muglia, B. (2013). *Decoding SDN*. Retrieved from http://forums.juniper.net/t5/The-New-
    Network/Decoding-SDN/ba-p/174651/

18. Nadeau, T., & Pan, P. (2011). Software Driven Networks Problem Statement. *IETF
    Internet-Draft*.

19. *Neutron's Developer Documentation*. Retrieved from
    http://docs.openstack.org/developer/neutron/

20. *Open Network Foundation*. Retrieved from https://www.opennetworking.org/sdn-
    resources/sdn-definition

21. *OpenStack*. Retrieved from http://www.openstack.org/

22. *Sahara*. Retrieved from http://savanna.readthedocs.org/en/latest/

23. *Software-Defined Networking: The New Norm for Networks, ONF White Paper.* (2013).
    Retrieved from https://www.opennetworking.org/images/stories/downloads/sdn-
    resources/white-papers/wp-sdn-newnorm.pdf

24. Strozzi, C. *NoSQL*. Retrieved from http://www.strozzi.it/cgi-
    bin/CSA/tw7/I/en_US/NoSQL/Home%20Page/

25. Sultan, N. (2010). Cloud Computing for education: A new dawn? *International Journal of
    Information Management, 30*(2), pp. 101-182.

26. *The Case for Smarter Transportation - IBM Whitepaper.* (2010). Retrieved from
    http://www-01.ibm.com/common/ssi/cgi-
    bin/ssialias?infotype=PM&subtype=XB&appname=SNDE_TT_TT_USEN&htmlfid=TTE
    03001USEN&attachment=TTE03001USEN.PDF

27. Watson, D. L. (2009). Hey - Get Off My Cloud! *5th International Conference on Global Security, Safety and Sustainability.* London, UK.

28. Zikopoulos, P. C., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding big data.* New York: McGraw Hill.

<div align="center">***</div>

Jelena Šuh is a PhD student at the Faculty of Organizational Sciences, University of Belgrade. From 2008 she is employed as IP/MPLS engineer at Telekom Serbia. As a Cisco instructor at Cisco Networking Academy Telekom Educational Initiative, she is involved in teaching CCNA course, covering the area of computer networks. Her current professional interests include computer networks, network management, e-business, Internet technologies and e-government. She can be reached at: jelena.suh@gmail.com.

Vujin Vladimir is research associate at Department for e-business and senior system administrator at Department for distance learning at the Faculty of Organizational Sciences, University of Belgrade. His current professional interests include e-business, computer networks, network management, Internet technologies, cloud computing and e-education. He can be reached at: vujin@elab.rs.

Barać Dušan is an assistant professor at Faculty of Organizational Sciences, University of Belgrade. He is involved in teaching courses covering the area of e-business, mobile business, Internet technologies, Internet marketing, Simulation and simulation languages, and Risk management in information systems. His current professional interests include internet technologies, mobile technologies, e-business, cloud computing and e-education. He can be reached at: dusan@elab.rs.

Bogdanović Zorica is an assistant professor at Faculty of Organizational Sciences, University of Belgrade. She is involved in teaching courses covering the area of E-business, Mobile business, Simulation and simulation languages, Internet technologies, Internet marketing, E-education, E-government, E-health, and Risk management in information systems. Her current professional interests include e-business, m-business, Internet technologies, e-government and e-education. She is a secretary of IEEE Computer Society Chapter CO-16. She can be reached at: zorica@elab.rs.

Radenković Božidar a full professor at the Faculty of Organizational Sciences, University of Belgrade. He teaches several courses in the fields of E-business, Internet technologies, E-education, Simulation and simulation languages, Internet marketing, Risk management in information systems, M-business, Internet of things. His current professional and scientific interests include e-business, internet technologies, computer networks, cloud computing, mobile business and Internet of things. He a president of IEEE Computer Society Chapter CO-16, and a full member of the following professional organizations: IEEE, ACM, SCS, YU-INFO, DOPIS, and UNESCO. He can be reached at boza@lelab.rs.