Integrating Distributed Hadoop System into the Existing Infrastructure

Stefka PETROVA[1], Liliya MILEVA[2], Pavel PETROV[3], Plamen YANKOV[4], Julian VASILEV[5]

[1] University of Economics, Varna, Bulgaria
s.petrova@ue-varna.bg

[2] University of Economics, Varna, Bulgaria
l.mileva@ue-varna.bg

[3] University of Economics, Varna, Bulgaria
petrov@ue-varna.bg

[4] University of Economics, Varna, Bulgaria
yankov.plamen@ue-varna.bg

[5] University of Economics, Varna, Bulgaria
vasilev@ue-varna.bg

Abstract. A distributed Hadoop system can integrate clusters of different organizations. The purpose of this article is to consider the options for building an architecture of a distributed Hadoop system, so that it is, on the one hand, to integrate a logically complete Hadoop system, and on the other hand – to define the conceptual framework of individual components that technically implement this. The scope of the research covers the problem of integrating remote Hadoop cluster to other one. A several findings are made and comparison between different ways of organization in data processing – multiple clusters and multi-lease are made. Both approaches have their advantages and disadvantages which could be used in practice.

Key words: Hadoop, big data, distributed systems, information systems, cluster management.

1. Introduction

Acknowledgments. This work is supported by the "Digitization of the Economy in an Environment of Big Data" Project - BG05M2OP001-1.002-0002.

The functions of a distributed Hadoop system are to integrate clusters of different organizations. Software, hardware, network, and organizational set of events are needed to realize this task. Hadoop clusters are designed to store and process large amounts of data in a distributed heterogeneous computing environment. Characteristically, they are scalable, and if they do not have enough computing resources due to the large amount of data, new nodes can be added to the cluster in order to solve the task and increase productivity. Since the same data is stored on different nodes in one cluster, in case of failure of one of the nodes the work can be continued using data from the other nodes (fig. 1). The number of copies can be set, and the default is 3.

One Hadoop cluster can connect and use data from another Hadoop cluster through different approaches.

The distributed Hadoop system, which integrates clusters of different organizations, must be connected via high-speed networks to the other components of the big data ecosystem:

a) LAN - with exchange speeds of 10 Gbps, 50 Gbps, 100 Gbps
b) MAN - with an exchange rate of 1 Gbps, 10 Gbps
c) WAN - with an exchange rate of 1 Gbps, 5 Gbps, 10 Gbps
d) LoraWAN - with exchange rate 1Mps, 10Mps, 100Mbps (fig. 2)
e) GSM network - with exchange speed 10Mps, 100Mbps, 1Gbps
f) Firewall with IPS / IDS capabilities (fig. 3)

The following specialists categories are required for a distributed Hadoop system to integrate clusters of different organizations:

a) Server hardware maintenance technicians - with tasks to keep the servers operational. Duration - until the end of the project;

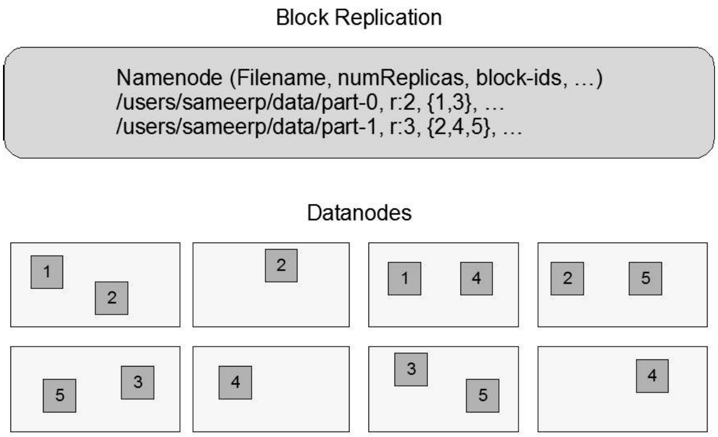
Block Replication
Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …
Datanodes
1
2
2
1
4
2
5
5
3
4
3
5
4

Figure 1. Replication of data blocks on DataNode servers

Source: HDFS Architecture, https://hadoop.apache.org/docs/r3.3.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

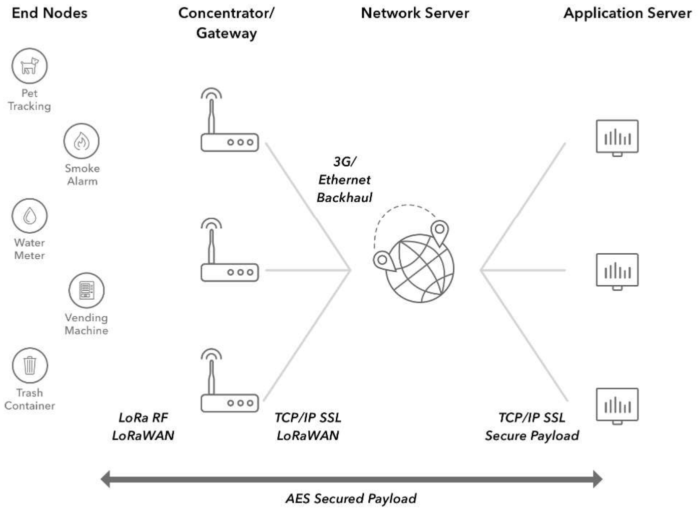

Figure 2. Using the LoRaWAN architecture as a means of communication

Source: LoRaWAN Architecture, https://www.thethingsnetwork.org/docs/lorawan/architecture.html

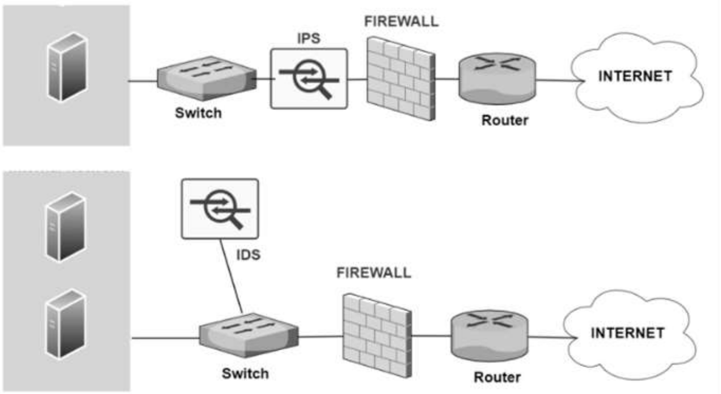

Figure 3. Using the Intrusion Prevention System (IPS) and the Intrusion Detection System (IDS) in combination with a firewall

Source: Comparison and Differences Between IPS vs IDS vs Firewall vs WAF,
https://www.networkstraining.com/firewall-vs-ips-vs-ids-vs-waf/

b) Technical specialists in the hardware maintenance of the network components - with tasks to maintain in working order the network devices and those for the information security. Duration - until the end of the project;

c) Software specialists for administration of Hadoop system - with the task to be able to re-install Hadoop system, to maintain in working condition Hadoop system; to configure the respective modules - services of Hadoop system. Duration - until the end of the project;

d) Hadoop system application software specialists - with the task to be able to develop application programs using the main software components - Hadoop system services. Duration - after the launch of the Hadoop system until the end of the period of development and testing of application programs;

e) Software specialists for data analysis from Hadoop system - with the task to choose a suitable software tool for data analysis (Sulova 2019) located in Hadoop system, as well as to be able to operate with these tools (Stoyanova 2020). Duration - after the launch of the Hadoop system until the end of the period of development and testing of application programs.

2. Literature review

When integrating an external distributed Hadoop system (cluster) into another existing Hadoop infrastructure, it is possible to use different approaches (Radev 2019; Aleksandrova 2018), depending on the objectives pursued.

According to the Hadoop documentation, the operating modes of the system that are officially supported are three:

1. Local (stand-alone) mode. This is the default mode when installing on a computer, which does not use the HDFS file system, but the local file system of the computer when performing input / output operations. In this variant of using Hadoop, for the purposes of integration with existing infrastructure, it is necessary to make appropriate changes in the configuration files core-site.xml, hdfs-site.xml and mapred-site.xml in order to move to full distributed mode of operation (Apache 2020a).

2. Pseudo-distributed mode. It is a single-node cluster that uses HDFS. In this mode, both the NameNode, which contains the list of all directories and files, and the DataNode, where the data is stored, are located on the same computer. For the purposes of integration with existing infrastructure, it is necessary to make appropriate changes to the configuration files core-site.xml, hdfs-site.xml and mapred-site.xml in order to switch to fully distributed mode (Apache 2020b).

3. Fully distributed mode. It is a cluster with many nodes in which the data is distributed and processed in each of them. In this situation, the following configuration parameters are set in the main configuration files for connection with the other nodes in the cluster:

- File /etc/hadoop/conf/core-site.xml - HDFS settings;
- File /etc/hadoop/conf/hdfs-site.xml - settings for NameNode (and possibly Secondary NameNode) and multiple DataNode;
- File /etc/hadoop/conf/yarn-site.xml - settings for ResourceManager, NodeManager and History Server;
- /etc/hadoop/conf/mapred-site.xml file - MapReduce settings.

Combined big data processing in a fully distributed mode of operation of one cluster with another cluster can be performed in three main operating scenarios.

In the first scenario, the two clusters operate relatively independently of each other, whereby the data from one cluster is transferred (copied) for processing to the other, after which the result of the processing is eventually transferred back. The DistCp tool (Cloudera 2019) can be used for this purpose when transferring large volumes of data. Another option is to use the real-time HFTP tool, which allows data to be transferred between remote Hadoop clusters that use HDFS. HFTP works on HTTP and data access is read-only, i.e., post-processing write operations must be performed in the local cluster.

In the second scenario, the two clusters merge statically to work as one large cluster and the resources of one distributed Hadoop system become part of the resources of the other Hadoop system (Ryu 2018; IBM Knowledge Center 2019). Special care is required when setting network settings in terms of increasing security.

In the third scenario, the resources of the two clusters are dynamically pooled by specially designed systems that build on Hadoop (Wang et al. 2013; Jeon et al. 2014).

Each of the listed scenarios has its advantages and disadvantages, which makes it suitable for one situation or another. Each of the main scenarios may have different variations (Kuyumdzhiev 2019) related to the details of a particular implementation.

3. Main functional components of the distributed architecture

The main functional components (blocks) of the architecture of a distributed Hadoop system, which integrates clusters of different organizations, are the following:

a) Main central cluster, part of the infrastructure of the Center of Competence. It can include clusters of external organizations.

b) Additional Hadoop cluster of a research company / organization, consisting of several components of type DataNodes and type NameNode / Secondary NameNode (fig. 4). The main system software components are: Hadoop Distributed File System (HDFS) - for distributed data storage in a file system, and Hadoop YARN - for managing computing resources and their distribution among many different users.

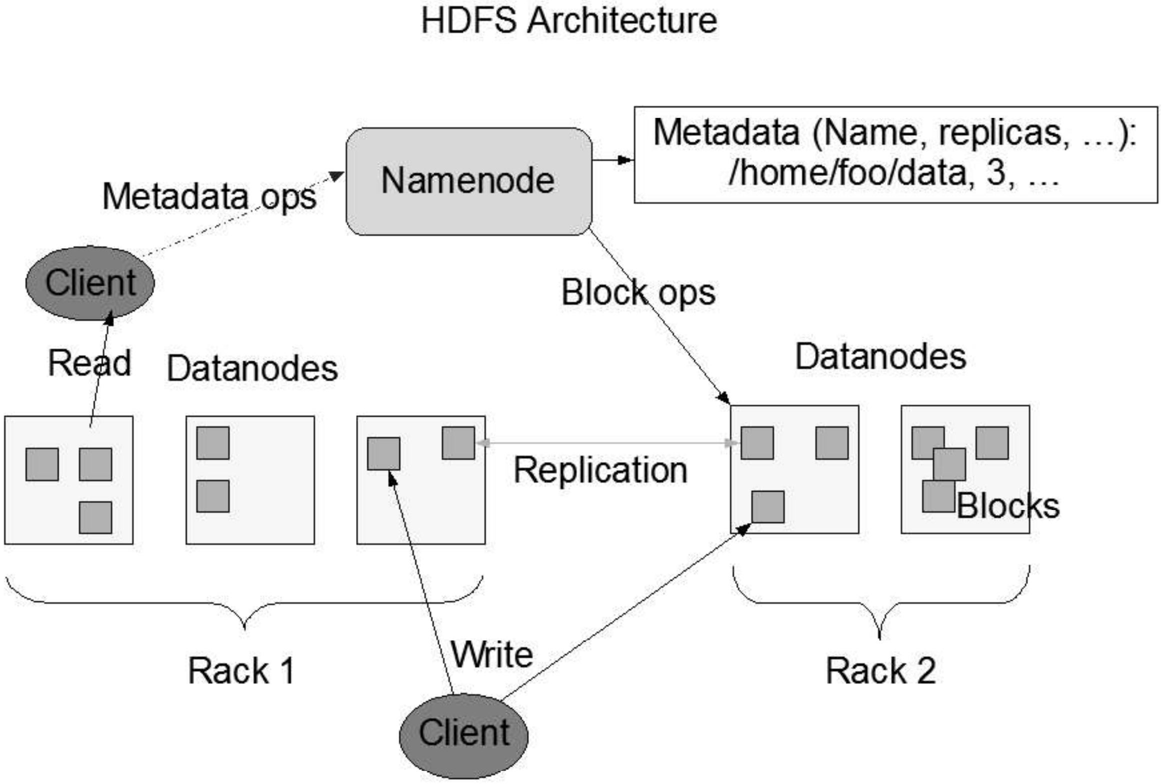

Figure 4. Replication of data blocks on different DataNode servers

Source: HDFS Architecture, https://hadoop.apache.org/docs/r3.3.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

The scenario in which two clusters merge statically to work as a large "single" cluster and the resources of one distributed Hadoop system become part of the resources of the other Hadoop system sets certain requirements. In practice, the nodes of the new "single" Hadoop cluster are located not in one, but in several data centers, geographically located in different places. It is necessary for all nodes in the cluster to be accessible through the network, i.e. their IP addresses to be available to other nodes in the network.

There are several options for merging the networks of different geographically remote data centers:

1. To use a separate telecommunication line between the two networks for connection and thus to build a common LAN. This is a relatively expensive option in case a rent had to be paid for a leased line to a telecommunications company.

2. To use the public Internet for connection. It is possible to apply two main tactics:

a) All nodes have public IP addresses - providing a large number of IP addresses version 4 can be a problem, as their number is limited and scarce nowadays. The use of IP addresses version 6, where there are no problems with a shortage of IP addresses, is not possible at this stage, as it is not officially supported (HadoopIPv6 2019). A significant problem is the protection of each node from various attacks specific to the Internet space. The guidelines provided by Cloudera (Ahluwalia 2017) can be used for this purpose.

b) Use of virtual private networks - VPN (Virtual Private Networks). Virtual private network technology makes it possible to provide an encrypted connection between private networks by using a public network, such as the Internet, as the transmission medium. Possible options are the use of GRE (Generic Routing Encapsulation), PPTP (Point-to-Point Tunnelling Protocol), L2F (Layer 2 Forwarding), L2TP (Layer 2 Tunnelling Protocol), IPSec (Internet Protocol Security) and MPLS (MultiProtocol Label Switching). Each of the options has its advantages and disadvantages in the implementation of the so-called. "Site-to-site VPN" and is recommended for use by various companies, such as Microsoft and Cisco.

One of the main problems in connecting several geographically remote clusters is that Hadoop generates a very large amount of traffic between nodes in the common cluster. In addition, time latency reduces the speed of data processing, as there is more waiting between the individual nodes that work together. It is very likely that time latency will be the "bottleneck" in the cluster and limit overall performance.

In the scenario where two Hadoop clusters operate relatively independently of each other, it is possible that part of the resources of one cluster will be made available for use if these resources are larger than the resources of the other cluster or need to be run in parallel. big data processing. In both cases, data from one cluster must be transferred for processing to the other and possibly the result must be transferred back (for example via the distcp tool).

The provision of some of the resources to the Hadoop cluster is known in the literature as "multitenancy" or "multi-lease". Generally speaking, this is: "a software architecture in which a single instance of an application running on a server serves multiple client organizations (so-called tenants)." (CIO Media 2014). In case of multirental, the Hadoop cluster software works simultaneously with several different configurations and data from several organizations. Each of the client organizations has access to a separate instance of the virtual application and sees only its data.

The YARN tool can be used to organize multi-tenancy (Nguyen&Won 2017). It separates resource management from big data processing workflows. This is done through separate program modules that work together - common to all ResourceManager module and a separate module for each application ApplicationMaster. The ResourceManager module allocates resources between applications by communicating with the NodeManager module, which is responsible for managing the resources of a physical machine - processor time, memory, disk space and network. The ApplicationMaster module, which is responsible for one application, declares the use of resources by ResourceManager and works with one or more NodeManager to perform work tasks (fig. 5).

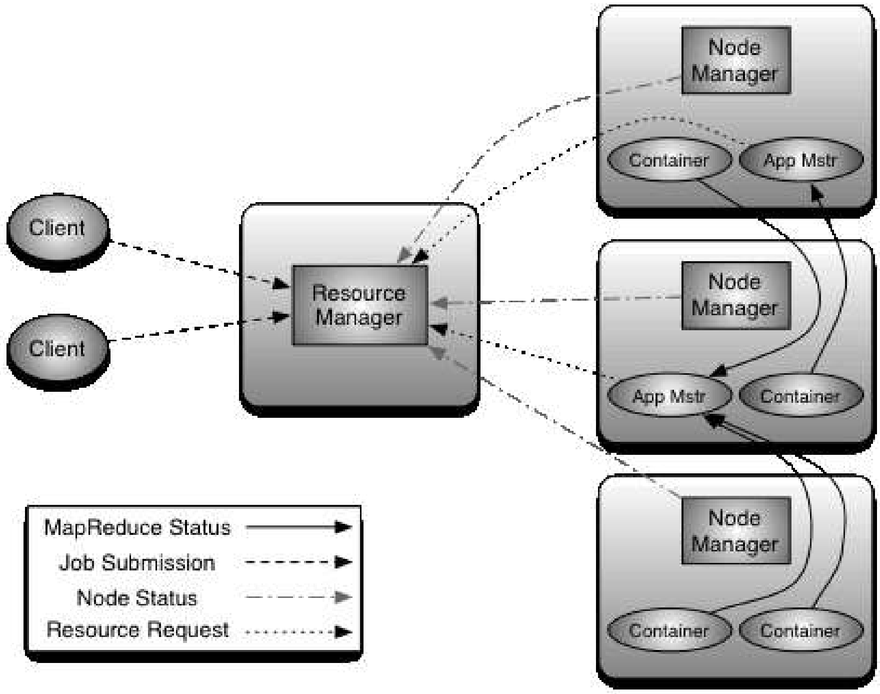

Figure 5. Interaction between ResourceManager and NodeManager
Source: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

ResourceManager has two main components:

1. Scheduler - is responsible for allocating resources between different running applications, subject to set restrictions. The rules under which resources are allocated can be changed or built-in allocation schemes such as

CapacityScheduler and FairScheduler can be used. The differences between the two schemes are not large and it should be borne in mind that Clodera in its product Cloudera Data Platform (CDP) only supports CapacityScheduler and offers a tool fs2cs to convert the rules from FairScheduler to CapacityScheduler (Clodera Docs 2020). For this reason, the use of CapacityScheduler is probably more promising.

2. ApplicationsManager - is responsible for accepting job requests, negotiating a container in which to run ApplicationMaster and restarting the ApplicationMaster container if necessary. In turn, ApplicationMaster is responsible for negotiating with the Scheduler of containers with appropriate resources (CPU time, memory, disk space and network) and monitors the operation of the containers.

Several factors need to be considered when choosing a multi-tenancy approach. A Hadoop cluster can generally be easily expanded with new hardware resources, but opportunities (such as power supply, air conditioning, etc.) and space (space in the server room) must be provided in advance. Therefore, it is necessary to correctly determine the initial size of the cluster, so as not to subsequently require frequent changes in terms of hardware resources. Based on the data processing needs, it is possible to choose an approach for several small clusters managed individually or one large cluster serving many users. Both approaches have their advantages and disadvantages (table 1).

Table 1.

Comparison between different ways of organization in data processing - multiple clusters and multi-lease

Way of organization / **Parameter**	**Several smaller Hadoop clusters**	**Large single Hadoop cluster with multi-lease**
Data duplication	Specific data is stored in each cluster. Duplicate data between clusters is possible.	The data of several research companies / organizations are stored. It is possible to use a "data lake" type structure, avoiding duplication when storing shared data. Company / organization-specific data can be stored in separate directories with access control provided.
Load of hardware resources	Each cluster has its own hardware resources. Their use can be both weak and overloaded. If the cluster is designed to handle peak loads, it is very likely that hardware resources are underloaded most of the time.	Hardware resources can be utilized efficiently through YARN. By creating a schedule for using the system within a day or a longer period, a fuller use (load) of hardware resources can be achieved.
Cluster management	Different system procedures (such as upgrades and debugging) must be performed on each of the clusters separately. These increases maintenance costs compared to a centralized management option. The advantage is that problems in one of the clusters do not directly affect the work of the others.	Maintenance activities are performed centrally for the entire cluster, which reduces costs. The disadvantage is that in these periods the work on all projects can be blocked.

Source: Own observations

In case when all servers in the clusters must have public IP addresses nftables network tool and nftables.conf configuration files with the same content could be used. The settings in the configuration files block access to any server from a computer on the Internet, except those with explicitly specified IP addresses. In this case, the goal of using nftables in a Hadoop cluster is to allow access to each server in one cluster only when the network traffic is from a server in the other Hadoop cluster. Eventually only the traffic to ports 22 (SSH protocol), 80 (HTTP protocol) and 443 (HTTPS protocol) could be publicly allowed for the purpose of remote administration of the specific server and output of publicly available system information via a web server.

4. Conclusion

Several conclusions could be made as a result of this research. Integrating an external distributed Hadoop system (cluster) into another existing Hadoop infrastructure is difficult to implement because, by default, the system is most often used in a local area network that is separate from the Internet. However, it is possible to use different approaches, depending on the objectives pursued. The operating modes of the system, according to the official documentation of Hadoop, are three: Local (standalone) mode, Pseudo-distributed mode, and fully distributed mode. The latter mode can be used to integrate distributed Hadoop system to the existing infrastructure.

When connecting several geographically remote Hadoop clusters a very large amount of traffic between nodes are generated. Moreover, the speed of data processing is reduced due to the time latency, and it is very likely that the time latency will be the "bottleneck" and will limit overall performance.

Literature

Ahluwalia, M., 2017. How to secure "Internet exposed" Apache Hadoop, https://blog.cloudera.com/how-to-secure-internet-exposed-apache-hadoop/

Aleksandrova, Y., 2018. Application of Machine Learning for Churn Prediction Based on Transactional Data (RFM Analysis). *18 International Multidisciplinary Scientific Geoconference SGEM 2018: Conference Proceedings*, 2-8 July 2018, Albena, Bulgaria: Vol. 18. Informatics, Geoinformatics and Remote Sensing. Iss. 2.1. Informatics, Sofia: STEF92 Technology Ltd., pp.125-132.

Apache, 2020a. Hadoop 3.3.0: Setting up a Single Node Cluster. Available at: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html, accessed 20.08.2020.

Apache, 2020b. Hadoop 3.3.0: Hadoop Cluster Setup. Available at: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html, accessed 20.08.2020.

Apache, 2020c. Comparison and Differences Between IPS vs IDS vs Firewall vs WAF. Available at: https://www.networkstraining.com/firewall-vs-ips-vs-ids-vs-waf/

Apache, 2020d. HDFS Architecture. Available at: https://hadoop.apache.org/docs/r3.3.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

Apache, 2020e. Apache Hadoop YARN. Available at: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

CIO Media, 2014. Predskazaniya za SaaS prez sledvashtata godina. Available at: https://cio.bg/management/2014/10/04/3440091_predskazaniia_za_saas_prez_sledvashtata_godina/

Cloudera Docs, 2020. Comparison of Fair Scheduler with Capacity Scheduler. Available at: https://docs.cloudera.com/runtime/7.2.1/yarn-reference/topics/yarn-FS-vs-CS.html

Cloudera, 2019. Administering HDFS, Using DistCp to Copy Files. Available at: https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.5/administration/content/using-distcp-to-copy-files.html, accessed 20.08.2020.

HadoopIPv6, 2019. Hadoop and IPv6. Available at: https://cwiki.apache.org/confluence/display/HADOOP2/HadoopIPv6

IBM Knowledge Center, 2019. Overview of InfoSphere Information Server on Hadoop. Available at: https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ishadoop.doc/topics/overview.html

Jeon, K., Chandrashekhara, S., Shen, F., Mehra, S., Kennedy, O., & Ko, S. Y., 2014. Pigout: Making multiple hadoop clusters work together. In *2014 IEEE International Conference on Big Data (Big Data)*, pp.100-109.

Kuyumdzhiev, I., 2019. *Metodologicheski i tehnologichni aspekti pri arhiviraneto na bazi ot danni*. Varna: Nauka i ikonomika, Bibl. Prof. TSani Kalyandzhiev; Kn.55.

Nguyen, M. C., & Won, H. S., 2017. Advanced Multitenant Hadoop in Smart Open Data Platform. In *Proceedings of the International Conference on Big Data and Internet of Thing*, pp.48-51.

Radev, M., 2019. Organizational Variants of IT Department at the University. *Information and Communication Technologies in Business and Education*: Proceedings of the International Conference Dedicated to the 50th Anniversary of the Department of Informatics, Varna: Science and Economic Publ. House, pp.284-290.

Ryu, W., 2018. Implementation of dynamic node management in Hadoop cluster. *International Conference on Electronics, Information, and Communication (ICEIC)*, Honolulu, HI, 2018, pp. 1-2, doi: 10.23919/ELINFOCOM.2018.8330612.

Stoyanova, M., 2020. Good Practices and Recommendations for Success in Construction Digitalization. *TEM Journal - Technology, Education, Management, Informatics* / Association for Information Communication Technology Education and Science, Novi Pazar, Serbia: UIKTEN - Association for Information Communication Technology Education and Science, 9(1), pp.42-47.

Sulova, S., 2019. An Approach to Storing Data Based on the Data Lake Concept to Facilitate Intelligence Data Analysis. *Serdica Journal of Computing*, Sofia: Institute of Mathematics and Informatics. BAS, 13(3-4), pp.171-182.

The Things Network, 2019. LoRaWAN Architecture. Available at: https://www.thethingsnetwork.org/docs/lorawan/architecture.html

Wang, L., Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J., & Chen, D., 2013. G-Hadoop: MapReduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems, 29*(3), pp.739-750.