BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **20**, No 3
Sofia • 2020 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2020-0025

Agile Elastic Desktop Corporate Architecture for Big Data

Valentin Kisimov, Dorina Kabakchieva, Aleksandar Naydenov, Kamelia Stefanova

University of National and World Economy, 1700 Sofia, Bulgaria
E-mails: vkisimov@unwe.bg dkabakchieva@unwe.bg anaydenov@unwe.bg kstefanova@unwe.bg

Abstract**:** *New challenges in the dynamically changing business environment require companies to experience digital transformation and more effective use of Big Data generated in their expanding online business activities. A possible solution for solving real business problems concerning Big Data resources is proposed in this paper. The defined Agile Elastic Desktop Corporate Architecture for Big Data is based on virtualizing the unused desktop resources and organizing them in order to serve the needs of Big Data processing, thus saving resources needed for additional infrastructure in an organization. The specific corporate business needs are analyzed within the developed R&D environment and, based on that, the unused desktop resources are customized and configured into required Big Data tools. The R&D environment of the proposed Agile Elastic Desktop Corporate Architecture for Big Data could be implemented on the available unused resources of hundreds desktops.*

Keywords**:** *Agile Elastic Desktop Corporate Architecture, Desktop Virtualization, Big Data, Digital Transformation.*

1. Introduction

Science and research are facing dynamic and challenging changes that are influencing terminologies, methodologies, practical investigations in all fields of knowledge within the digital era today. The terms “Digitalization” and “Digital transformation” are related to the scalable processes encompassing all aspects of society and economy. Specifically, economy needs a very concrete focus on the changes related to Digital transformation.

The implementation of advanced technologies is not the only critical aspect for the economy today. All traditional business models and processes need to be reconsidered and innovated. Prominent experts in the Information and Communication Technology (ICT) field define “Digitalization of economy” as “… the use of digital technologies to change a business model and provide new revenue and value-producing opportunities…” [1]. Another step forward in the process of moving to a digital business is the “Digital Transformation of economy” clearly explained as "…the cultural, organizational and operational change of an organization, industry or ecosystem through a smart integration of digital technologies, processes and competencies across all levels and functions in a staged way…" [2]. Digital transformation is predominantly used in a business context to encompass all aspects of digitalization, including intelligent use of technologies for creating value for various stakeholders, innovating and adapting to changing circumstances.

Digital Transformation of economy leads to another phenomena – Big Data that becomes a force influencing all types and size of companies. They should adequately change their use of technology, processes, personnel, and data. According to the IDC (International Data Corporation) forecasts published in April 2019 [3], "… revenues for Big Data and Business Analytics solutions will reach $189.1 Billion this year with double-digit annual growth through 2022".

The use of Big Data allows companies to make meaningful, strategic adjustments leading to minimizing costs and maximizing results. Big Data analytics is the main focus in all industries today [4]. However, Big Data challenges are numerous, including volumes, variety of data formats, increasing velocity of generation, to mention just a few [5]. The exponential growth of data [6], and especially concerning the growth of unstructured and semi-structured data, coming from emails, photos, videos, monitoring devices, PDFs, audio, etc., poses important issues for storage, mining and analysing data.

The conducted research reveals that companies, introducing business processes digitalization, do not propose standard approaches for organisation and integration of Big Data. Classically, the digitalized business processes work with structured data but most of the newly generated data are predominantly unstructured or semi-structured (about 80%). Many companies that are even world leaders in the digitalization of business processes, nowadays are also confronted with the Big Data challenges because of semi-structured and unstructured data which should be transformed into structured format in order to be further used in business processes digital transformation. Business practice is facing the need for searching innovative models and new approaches to develop and implement solutions for working with Big Data volumes and new data formats.

In order to remain competitive, Bulgarian companies need to experience digital transformation and receive added value by adequately using Big Data generated in everyday business transactions. In response to that need, a new project titled "Digital Transformation of Economy in Big Data Environment" is initiated, funded by the Operational Program "Science and Education for Smart Growth" 2014-2020. The project is aimed at creating a unique ICT infrastructure to be used for developing innovative applicable business solutions for digital transformation and Big Data processing. The infrastructure will be based on the latest scientific and technological world achievements, international and industrial standards, and best business practices, and will provide multi-access hosting and cloud features.

When approaching a Big Data system development, the problem of storage is always addressed. For the purposes of this research, a "Big Data system" is considered as an ICT system focused on Big Data processing and application. The business practice usually underestimates the role of desktop Personal Computer (PC) resources. Desktop devices' components (e.g., CPU, RAM) are nowadays produced with a powerful capacity and it is even not possible to produce a desktop (including laptop) with low computational capacity and memory size. Therefore, the desktop virtualization problem is on the research agenda, leading to providing solutions with different levels of desktop virtualization at the current stage of technology.

A possible solution for solving real business problems concerning Big Data is proposed in this paper. The idea is to more effectively utilize the available unused desktop resources in an organization by virtualizing and organizing them in order to serve the needs of Big Data processing.

The paper is organized in 5 sections. The motivation for the research work and the current state-of-the-art in the field is described in the Introduction section. The second section is devoted to the proposed Agile Elastic Desktop Corporate Architecture for Big Data, including a definition of the architecture, implementation of a desktop virtualization, and presentation of the Big Data virtualized components. In the third and fourth sections, the parameters for establishment of the proposed Architecture and the development of an R&D environment are considered, including the proposed solution for an effective use of desktop resources for Big Data processing. The paper ends with a conclusion section in which the main research findings are discussed.

2. Core concepts of the Agile Elastic Desktop Corporate Architecture for Big Data

The purpose of this section is to describe the essence of the proposed Agile Elastic Desktop Corporate Architecture for Big Data that is the authors' vision for solving an existing problem in many companies – not using fully and adequately the existing desktop resources while additional computer resources are needed for processing and analysing Big Data. Authors introduce a definition of the architecture, propose a model for the desktop virtualization, and respectively describe the Big Data virtualized components.

2.1. Defining the Agile Elastic Desktop Corporate Architecture for Big Data

ICT Architecture [7] comprises all static and dynamic information and telecommunication structures, including hardware, software, network devices and communication devices – ICT components. At the same time, there is no single definition for "ICT Corporate Architecture" and it is usually defined as an ICT Architecture that supports and fulfils the corporate business requirements with corresponding ICT guidelines and standards.

In order to approach the solution of the problem stated above, authors need to introduce a working definition of "Agile Elastic ICT Corporate Architecture" that to be a guiding pillar of the research purposes. In general, the term "Agile" means flexible, quickly and easily changing. The term "Elastic" considers spontaneous extending/reducing/resuming resources that in this research case are available computer capacities. The traditional methodology for ICT Architecture development requires first elaboration of a detailed plan that is a basis for further implementation [8] resulting in building the architecture. When an Agile development methodology is followed, the high-level plan for a system is broken down into small pieces which are independently completed in separate work sessions [9].

The "Agile ICT Corporate Architecture" in this research is defined as an ICT Architecture which is not designed by using the traditional ICT Corporate Architecture design methodology, but is based on a design philosophy for implementing an integrated set of separate dynamic architectural parts where some of the components have the ability and flexibility to be transformed (re-designed) into other architectural components during the process of ICT Architecture deployment or while running it. In other words, the Agile ICT architecture consists of integrated architectural parts which can be transformed (re-designed) or re-tuned with deferent configuration parameters during the deployment and running phase. In that way, the Agility incorporates three stages: Agility#1 – Creation of the ICT Architecture from different architectural parts; Agility#2 – Transformation or re-design of the architectural parts during the architecture deployment and/or using them; and Agility#3 – Re-tuning the parameters of the architectural parts during the architecture use by running a software script – a program dynamically changing the parameters of the architectural parts.

In this respect, the term "Agile Elastic ICT Architecture" is defined as a set of architectural components where each of them can spontaneously extend/reduce/resume its functional parameters as a result of running application programs (end user programs) on that ICT Architecture, while activating one or more architectural parts. In other words, the elasticity is related to the ICT architecture dynamics, depending on the business needs during the execution of different application programs.

Following the stated above definition, a "Desktop Architecture" could be considered as a set of desktop computers with devices for their networking and security services [10]. Approaching the objective of this research for creating a flexible dynamic solution and encompassing the above two definitions, then the term "Agile Desktop Corporate Architecture for Big Data" could be defined as a combination of a Desktop Architecture with the principles of an Agile ICT Corporate Architecture, offering ability and flexibility for transforming the existing Corporate Desktop architecture with additional scalability for shaping new desktop based architectural components powered with Big Data functionality via software driven processes, and ensuring improved desktop resources utilization.

2.2. From Desktop Corporate Virtualization to Agile Elastic Desktop Corporate Architecture

More than 15 years ago, the biggest corporations had hundreds of physical separate servers, where generally a single application was executed on a server (server application). The utilization of those servers was about 20% of their resources or less. During the years, the computational power of servers has increased, and the server resources have been even less utilized. The virtualization principles are then introduced, where many Virtual Machines (VMs) are implemented on a single physical server and in each VM a single server application is run. One of the most important benefits achieved from server virtualization is the indicator for maximum server resources utilization.

In analogy to the server resources utilization problem leading to a solution of server virtualization, today the desktop virtualization problem is on the research agenda. The CPU, RAM memory and all other components of desktop devices are nowadays produced with a powerful capacity and it is already not possible to produce a desktop (including laptop) with small computational capacity and low memory size. Similarly to the server virtualization solution, there is also a solution to the desktop resource unitization – desktop virtualization [11].

Different levels of desktop virtualization are defined at the current stage of technology development:

- **Remote Desktop Virtualization (RDV).** Operates in client-server computing model, where server-based application and operating system work with desktop (display, keyboard, and mouse) via remote display protocol. This desktop can be any "thin client", including tablet and smartphone.
- **Desktop with Presentation virtualization.** Operates under the common name "Terminal Services", where the desktop can access remotely another desktop, and using RDP protocol or the approach used by Citrix XenApp, to operate with the remote application.
- **Desktop with Application virtualization.** Operates with separation of the application from its operating system, where the application runs on the desktop, and the operating system of that application runs on a remote system. A hypervisor [12], installed on the desktop, acting as kernel of an operating system, ensures the application at runtime to act as if it is interfacing with the original operating system (running on a remote system). The hypervisor, as software working with the desktop hardware and firmware, has another role, which is most popular – to create and run Virtual Machines, supporting the virtualization process.
- **Desktop with User virtualization.** Operates with separation of user personality (profile, group policy) from the application and its operating system. This approach is used by Citrix, Microsoft, VMware, etc.
- **Virtual Desktop Infrastructure (VDI)** runs every desktop application in a separate VM installed in a remote server and users are logged in on the desktop (independently if it is thin client device or fully equipped desktop device). In this way, VDI offers the possibility for corporate users to access variety of desktop applications running on a variety of Operating Systems (OSs) from a single device – desktop. The remote server can be in a cloud or in-house premise.
- **Desktop as a Service (DaaS).** DaaS [13] offers the possibility of VDI to be hosted in a cloud service provider environment, applying a multi-tenancy architecture, which means that a single instance of a desktop application is served to multiple desktop users, referred to as "tenants". In most cases, the desktop runs a special hypervisor, a VM on top of the hypervisor and inside that VM a DaaS software to operate with the cloud based multi-tenant servers, which desktop generally is not thin client device in corporate systems.

Hypervisors are used in most of the contemporary desktop virtualizations. The VMs created on top of the hypervisors can play different roles in the desktop virtualization. At the same time, the hypervisor runs and creates VMs under the control of a software script, which is either integrated with the GUI to look like a management tool or activated via Application Programming Interface (API). There are two types of hypervisors – "Type-1", where the hypervisors interact with the underlying physical resources (hardware and firmware), interacting directly with its CPU, memory, and physical storage (called also bare-metal hypervisor), and "Type-2", where the hypervisor runs as an application on top of an existing OS.

For an Agile ICT infrastructure, the hypervisor Type-1 is most appropriate, and the contemporary ICT technology has evolved hypervisor Type-1 as a mature product. When a VM is created on top of a hypervisor, the role, functions and existence of this VM is also established and modified via a software script. In this way, a desktop with installed hypervisor is the core component of an Agile ICT Architecture, including the Agile Desktop Corporate Architecture.

One of the fundamental technologies used in desktop virtualization is "CPU virtualization", allowing the physical CPU cores (even one) to be logically split into more virtual cores (vCores/vCPU), on which vCPU to build hypervisor, VMs, Operating systems, applications. Many of today's CPUs (AMD & Intel) have extended instruction sets that support CPU virtualization, called "hardware CPU virtualization" (different providers give different names of the hardware CPU virtualization – Intel VT, AMD-V, etc.). The enabling of vCPU is managed by desktop BIOS, and often the CPU virtualization is disabled by default in the BIOS and needs to be enabled [14], for example to make "Intel Virtualization Technology Enabled" for Intel CPU based desktop. In this way, the CPU virtualization involves a single CPU to act as if it were multiple separate CPUs/cores in the desktop. CPU virtualization features enable faithful abstraction of the real number of cores/CPUs that is used. Most of the today's hypervisors and VMs successfully use CPU virtualization technology.

Another technology, used for virtualizing the desktop environment into a bigger one, for the purpose of implementing different system and application software, is the "Virtual memory/RAM". This is one of the parameters that is used to define the VM, which is built on top of a hypervisor. In Virtual memory, Virtual address space is used as an alternate to the physical memory address space (virtual address space offers using bigger memory capacity than the physical address space). For example, virtual memory might contain twice as many addresses as physical main memory. In this virtualization, the software uses these virtual addresses rather than real addresses to store instructions and data. When the program is executed, the virtual addresses are converted into real physical memory addresses. The full content of the virtual memory is not able to fit in the main memory all at once. Nevertheless, the desktop could execute such a software by copying into the main memory those portions of the virtual memory needed at any given point during the execution. The amount of memory that is possible to be allocated for a VM is the amount of memory that the guest operating system detects. Minimum memory size is 4 MB for VMs that use BIOS firmware. The memory size must be a multiple of 4 MB. If the VM memory is greater than the host memory size, swapping occurs, which can have a severe effect on the VM performance. The maximum size of the virtual memory depends on the hypervisor and form of the VM performance. The current hypervisors can manage terabytes of maximum virtual memory size, which for desktop systems is practical without serious limitation. The hypervisor performs the swapping between physical memory and disk, manages the position of a page from a VM memory into a physical memory and provides addressing from a CPU to the right physical address.

I/O virtualization provides abstraction of the upper layer protocols from the physical connections and enables one physical adapter card to appear as multiple virtual Network Interface Cards (vNICs) and virtual Host Bus Adapters (vHBAs). For the Desktop ICT Architecture this type of virtualization cannot play a crucial role and will not be part of the Agile Desktop Corporate Architecture.

Many of the comparisons presented in the literature, between Thin clients/Zero clients and Desktops (known as "Thick client"), are based on the assumption that the desktop is a full computer system with full OS and application software stack, without having in mind the specified above six types of desktop virtualization technologies. The Thin client systems TCO (Total Cost of Ownership) is not essentially lower than the Thick client, if the cost of the supporting servers and the price of their reliability are added [15]. Most of the desktop virtualization technologies, applied on standard desktop equipment, eliminate the essential advantages of Thin client machines to the Desktop machines. On top of this, the standard desktop systems offer additional computational and network features, which the current corporate staff requires, based on their increased computational literacy. Today the big corporations, including big banks, are using tens of thousands desktops in their corporate environments, applying or in a way to apply different virtualization technologies. Gartner informed [16] that worldwide PC (Desk-Based and Notebook) market grew 1.5% in the second quarter of 2019, with shipment to be 183 million units for 2019. At the same time, the global market of Thin clients is decreasing [17] and as a total number of units, they are about 2-3% of the total numbers of PCs. This is the reason for the authors to consider the high importance of approaching the available computational resources on a corporate base and use their scalability for meeting the needs of Big Data system development. Referring to the above presented approaches, the proposed Agile Desktop Corporate Architecture is defined based on desktops and thin/zero clients.

2.3. Big Data virtualized components for the Agile Elastic Desktop Corporate Architecture

The Agile Elastic Desktop Corporate Architecture is based on desktop PCs with their networking devices. The minimum level of the contemporary PC computational power – CPU power, RAM capacity, I/O bandwidth and Disk capacity are so high, that a PC with low computational power cannot exist in the market. That means each new PC, dedicated for desktop, has as minimum configuration so high level of computational power that the desktop applications cannot use it. The resource utilization of a desktop PC at this moment is similar to the servers' utilization situation of 10-15 years ago which led to the virtualization philosophy of servers. Today, there is low utilization of the desktop PCs serving only the desktop needs, so there is a need to move to virtualization philosophy of desktop PCs. The proposed Agile Desktop Corporate Architecture is based exactly on that philosophy – to virtualize the desktop PC resources unused for desktop applications and organize them to serve the needs of Big Data processing. Using Agile Desktop Corporate Architecture will help the corporations to utilize the existing unused desktop capacity. Having in mind that the corporations have increasing needs for Big Data systems, these available desktop PC unused resources will be oriented to creating components for Big Data systems. It such a way, the corporations will not buy dedicated Big Data systems and their components but will create such components by using the available unused desktop PC resources.

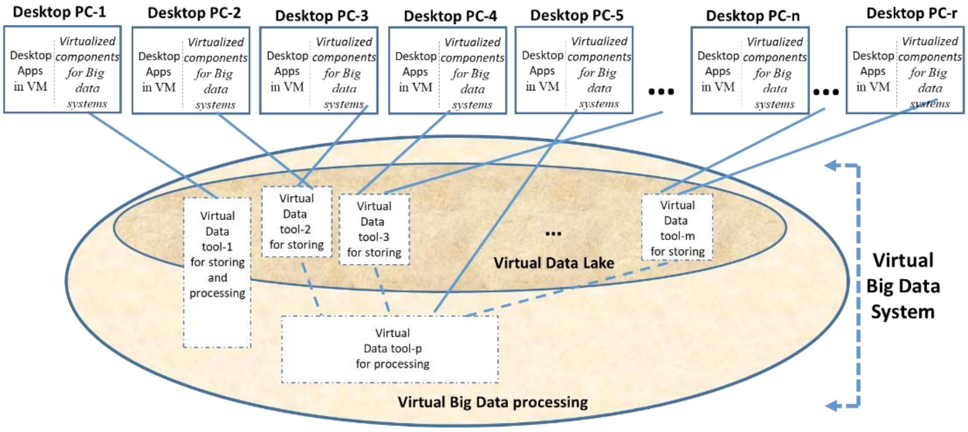

Desktop PC-1
Desktop PC-2
Desktop PC-3
Desktop PC-4
Desktop PC-5
Desktop PC-n
Desktop PC-r
Desktop Apps in VM
Virtualized components for Big data systems
•••
Virtual Data tool-1 for storing and processing
Virtual Data tool-2 for storing
Virtual Data tool-3 for storing
...
Virtual Data tool-m for storing
Virtual Data Lake
Virtual Data tool-p for processing
Virtual Big Data processing
Virtual Big Data System

Fig. 1. Unused desktop PC resources can be organized into Virtual Big Data System

The proposed Agile Desktop Corporate Architecture is presented in Fig. 1 and is based on Desktop PCs. The Big Data system will be using the unused desktop PC resources, which will be configured as Virtual Data tools – for storing Big Data, forming a Data Lake and for processing Big Data. Each Desktop PC will be configured in 2 parts – two Virtual Machines (VM) with their CPU cores, RAM, disk space and networking. The first VM is for the desktop functionality and the second VM is to provide Virtualized components for Big Data systems – supporting the needs of Big Data operations. Each Virtualized component will provide appropriate functioning to create a full Virtual Data tool or a part of it. A single Virtual Data tool can be built using the functioning of one or a few Virtualized components. For example, a virtual datastore, being a Virtual Data tool, can be created by five Virtual components from five desktop PCs. Various Virtual Data tools exist – the first group is only for making a Data Lake (centralized repository allowing to store structured, semi-structured and unstructured data at any scale [18]), and the second group is to process data (read/write/modify/delete/conclude/calculate/relate/analyse/present/report) from the Data Lake. These two groups of tools are used to build “Big Data systems”. For example, Virtual tools from the first group are used to store and access data in the form of spreadsheets, word files, presentations, log files, audio and video data, emails; Virtual tools from the second group are used for operating databases (relational and non-relational), Data Nodes and Name/Management Nodes for Hadoop system, elements from Hadoop ecosystem like Spark, etc.

Specific feature of the Virtual Data tools from the second group, processing data from the Virtual Data Lake, is that they can use already created Virtual Data tools, supporting the Virtual Data Lake. For example, if a few Virtual Data tools for Virtual SAN (tools supporting the Virtual Data Lake) are created, they can be used by another Virtual Data tool for DataNode of Hadoop system (tools that process data from Virtual Data Lake). Also, a Virtual Data tool can be from the first group, i.e., for supporting the Virtual Data Lake, and at the same time functioning as a Virtual Data tool from the second group, i.e., for processing data.

After analysing the possibilities for virtualization of the Desktop infrastructure and the needed functions of Big Data systems, the following set of 9 types of Virtual Data tools have been defined for the Agile Elastic Desktop Corporate Architecture for Big Data (AECDABD), where the VM can store data or/and access data for appropriate management, including:

i. *Physical hard drive*, named "Raw Datastore";

ii. *File data store*, named "File Datastore". It is a storage container for files, which could be located on local hard drives or across the network on a SAN [19]. This type of data store can have capacity up to 64 TB, but it depends on different vendors [20];

iii. *Device data store,* named "Virtual Volumes Datastore" or "vVols Datastore". It is a storage that can be provisioned dynamically according to the needs of the VM [21]. It is high-level virtualization of data store, which includes a data store, a set of data services, and metadata. This type of data store can bound up to 4096 desktop PCs at one time with capacity up to 64 TB [20], but it depends on different vendors;

iv. *Virtualized SAN*, named "vSAN". Virtual SAN aggregates local or direct-attached storage disks in a cluster of desktop PCs and creates a single shared storage pool [22]. One vSAN datastore can use up to 64 desktop PCs, and when High-availability feature (HA) is applied, the number of "embraced" desktop PCs to make a single vSAN is doubled [23,24];

v. *Relational DBMS* – where structured data will be stored, named "SQL database";

vi. *Non-relational DBMS*, named "Non-SQL database";

vii. *Nodes for Hadoop system* such as DataNode and NameNode;

viii. *Management of Hadoop system* such as ManagementNode, for example Cloudera Manager;

ix. *Analytical and Machine Learning engine* as part of Hadoop ecosystem, for example Spark software.

To the applications and guest operating systems inside each VM, the storage subsystem – i), ii), and iii) appears as a virtual SCSI controller connected to disks. All of them can be used in VMs and for this reason they can be addressed via a common name VMDK (disk for VM).

3. Parameters for creation of Agile Elastic Desktop Corporate Architecture for Big Data Systems

In the process of creation and tuning of the Agile Elastic Desktop Corporate Architecture the following parameters for virtualization of the desktop environment (desktop PCs and their networking) can be identified.

3.1. Parameters for Desktop Virtualization

i. vCPU – virtual core of a CPU;
ii. vRAM – RAM used for VM;
iii. vNICs – virtual network interface cards;
iv. vHBAs – virtual host bus adapters;
v. IOps – I/O operations per second;
vi. IO buffer size;
vii. size of queue for IO operations;
viii. I/O size;
ix. VMDK stripe width – number of disks used;
x. Ratio SDD/HDD in a storage subsystem;
xi. Number of VMs per host cluster.

3.2. Parameters for Virtualized Data tools

A. Parameters for storage subsystem – Raw Datastore, File Datastore, vVols Datastore:
i. Capacity of storage subsystem;
ii. Number of disks in a VM deployed for the Storage subsystem cluster;
iii. Number of VMs deployed for the Storage subsystem cluster;
B. Parameters for vSAN:
iv. Number of VMs deployed on the vSAN cluster;
v. Number of vCPUs per VM for vSAN;
vi. Size of memory per VM for vSAN;
vii. Size of VMDK per VM for vSAN;
viii. Number of HDD in a vSAN group;
ix. Size of SSD in a vSAN group;
x. Ratio of SSD vs HDD capacity;
xi. Stripe width, defining the number of disks across which a single data element (object) is striped;
xii. Tolerance to fail, defining the number of host in the vSAN cluster, that are tolerated to fail without data loss;
xiii. number of nodes in the vSAN cluster;
xiv. number of disk groups in each node for vSAN cluster;
C. Parameters for SQL database:
xv. Capacity of SQL database;
xvi. Number of disks in a VM deployed for the SQL database cluster;
xvii. Number of VMs deployed for the SQL database cluster;

D. Parameters for Non-SQL database:

xviii. Capacity of Non-SQL database;

xix. Number of disks in a VM deployed for the Non-SQL database cluster;

xx. Number of VMs deployed for the Non-SQL database cluster;

E. Parameters for DataNode and NameNode:

xxi. io.file.buffer.size – buffer size for I/O(read/write) operation on sequence files;

xxii. fs.inmemory.size.mb – the size of the in-memory file system instance in MB;

xxiii. io.sort.factor – the number of streams to merge at once while sorting files;

xxiv. io.map.index.skip – number of index entries to skip between each entry, helping to facilitate opening large map files using less memory;

xxv. io.map.index.interval – for every interval an entry (record-key, data-file-position) is written in the index file, for the purpose of optimizing the ration time for random access against the memory usage;

xxvi. fs.df.interval – the interval in which a statistics is written for disk usage;

F. Parameters for ManagementNode:

xxvii. Capacity of the used internal database;

xxviii. Number of disks in a VM deployed for the ManagementNode cluster;

xxix. Number of VMs deployed for the ManagementNode cluster.

3.3. Tools for Configuration of Elasticity and Agility

i. vMgmCenter – for configuration of a Center for Management of the virtualized desktop environment;

ii. LDAP server managing the access rights of the configurator-administrator;

iii. Hadoop Benchmarking evaluation tools.

4. Research and Development (R&D) Environment for Agile Elastic Desktop Corporate Architecture for Big Data systems (AEDCABD)

The Desktop environment for a big corporate architecture consists of tens of thousands desktop PCs and their networking. Creating a formal design solution for implementing these tens of thousands desktop PCs when realizing a variety of Big Data processing is not a practical approach, firstly because such a formal approach is not existent and secondly, due to the possible low efficiency resulting in serious mis-usage of those tens of thousands desktops.

There is no suitable architecture design methodology for Big Data systems on desktop environment that could be used. Therefore, a R&D environment has to be built where core tools and their parameters, together with essential Big Data application algorithms, must be tested to extrapolate possible components and structure of a solution for an Agile Elastic Desktop Corporate Architecture for Big Data (AEDCABD).

The general purpose of the R&D environment is to approbate and evaluate specific corporate business needs for Big Data System components and their possible virtual realization with the available unused desktop PC resources in the corporate ICT system. During the approbation and evaluation, three types of Agilities and 1 type of Elasticity will be applied. That will result in identifying the required customization and configuration of the unused desktop PC resources, leading to the creation of the document "Detail design" of the AEDCABD. In that way, lacking a methodology for the design of a Big Data System on desktop PCs, the R&D environment is proposed, which will guide the design of the AEDCABD.

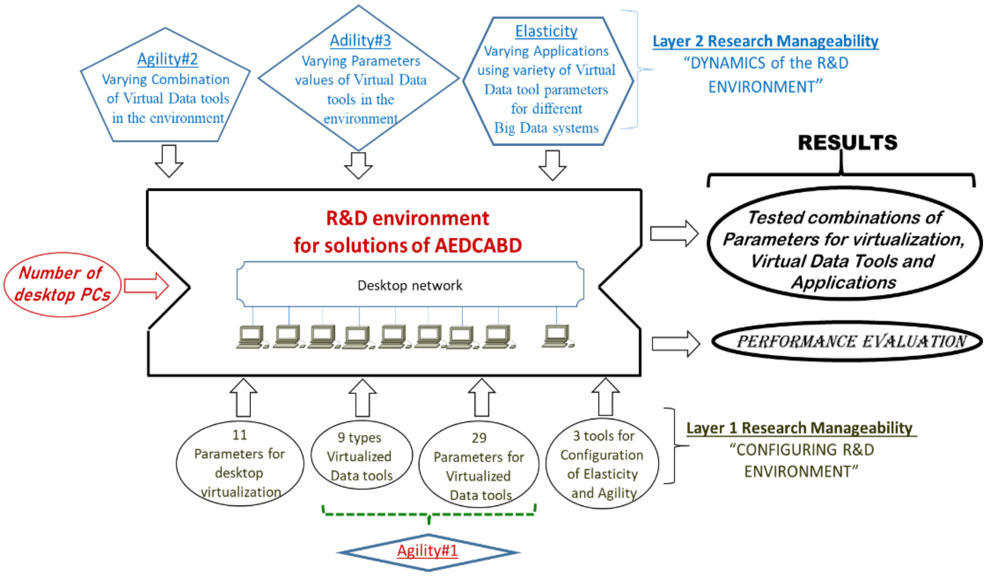

Agility#2
Varying Combination of Virtual Data tools in the environment
Agility#3
Varying Parameters values of Virtual Data tools in the environment
Elasticity
Varying Applications using variety of Virtual Data tool parameters for different Big Data systems
Layer 2 Research Manageability
"DYNAMICS of the R&D ENVIRONMENT"
RESULTS
R&D environment for solutions of AEDCABD
Desktop network
Number of desktop PCs
Tested combinations of Parameters for virtualization, Virtual Data Tools and Applications
PERFORMANCE EVALUATION
11 Parameters for desktop virtualization
9 types Virtualized Data tools
29 Parameters for Virtualized Data tools
3 tools for Configuration of Elasticity and Agility
Layer 1 Research Manageability
"CONFIGURING R&D ENVIRONMENT"
Agility#1

Fig. 2. R&D Environment helps to define the AEDCABD detail design

The concept of the R&D environment is presented in Fig. 2. The R&D environment is based on a number of desktop PCs with their networking, on which two layers of Research manageability are implemented to receive the final results: Tested combinations of Parameters for virtualization, Virtual Data Tools and focused Applications for using Big Data systems; and Performance evaluation of those parameters, tools and applications. Based on those results, the real Desktop Corporate Architecture will be designed. The first layer of the Research manageability of the R&D environment, "Configuring R&D Environment", includes the 11 Parameters for Desktop virtualization (explained in 3.1), the Agility#1 components (covering the selection of up to 9 Virtualized Data tools and up to 29 types of Parameters for those Virtualized Data tools) and up to three tools for configuring the Elasticity and Agility of the architecture. The second layer of the Research manageability of the R&D environment, "Dynamics of the R&D environment", includes Agility#2 components (varying Combination of Virtual Data tools), Agility#3 components (varying Parameter values of Virtual Data tools) and the Elasticity features (following the varying Applications using variety of Virtual Data tools for different Big Data systems).

In order to identify how many desktop PCs are required for the R&D environment, two different approaches are used. The first approach considers Layer 1 for Research manageability as a basis, i.e., Layer 1 is with slow dynamics, while the Layer 2 for Research manageability is with higher dynamics, i.e. the parameters of Layer 2 components and features are changing dynamically for each status of Layer 1, depending on the Agility#2, Agility#3 and Elasticity. In this case, an enormous R&D environment of 8613 desktop PCs should be developed (the full combination of all components of Layer 1 Research manageability is 8613 – 11 Parameters for desktop virtualization × 9 types Virtualized Data tools × 29 types of Parameters for Virtualized Data tools × 3 tools for Configuration of Elasticity and Agility).

The second approach for identifying the necessary desktop PCs in the R&D environment is based on the statistical method for estimation of the general number of computers collecting experimental results, with a 95% confidence level, with a maximum error in the processes of research and testing of ±2 (number of PCs), and applying the standard deviation between the values of the test results, for the purpose of Big Data system development. The calculations are based on the next equation:

(1) $$n = \frac{z^2 \cdot \sigma^2}{\Delta^2},$$

where: z is defined by the level of confidence of 95%, e.g., z is equal to 1.96 (according to the Normal distribution); the maximal error of ±2% of number of needed PCs in the research and testing defines $\mathbf{\Delta}$=2; σ is the standard deviation in the results achieved from varying parameters in desktop virtualization for the purpose of Big Data systems creation.

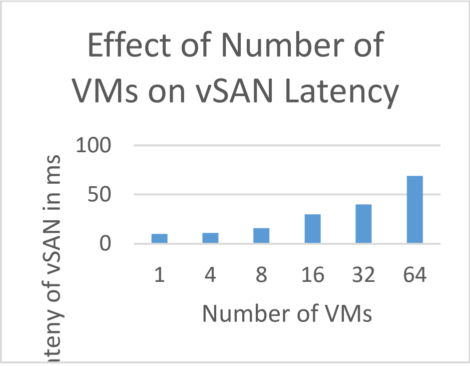
Equation (1) defines the minimum sample size that is needed to estimate the true population parameter with the required maximum error and certain confidence level. The formula is based on the Central Limit Theorem from which the maximum error originates defined as follows [25]:

(2) $$\Delta = z.\frac{\sigma}{\sqrt{n}},$$

Therefore,

$$(\Delta)^2 = \left(z.\frac{\sigma}{\sqrt{n}}\right)^2 \rightarrow \ \Delta^2 = z^2.\frac{\sigma^2}{n} \quad \rightarrow \ n = \frac{z^2.\sigma^2}{\Delta^2}.$$

There are many results in the literature, where the standard deviation σ is varying around the value of 22-23. For example: the effect of the number of VMs on vSAN Latency in Mixed R/W workload in vSAN is with σ=22.69 [26] – Fig. 3; the effect of the number of VMs on vSAN I/O operations in Mixed R/W workload in vSAN is with σ=22.41 [26] – Fig. 4; the effect of the number of desktop PCs for vSAN I/O operations in Mixed R/W workload is with σ=22.77 [26] – Fig. 5; the time for running different tests on Hadoop DataNodes made up of two VMs (PI, TestDFSIO-write, TeraGen 1TB, TeraValidate 1TB) is with σ=22.87 [27] – Fig. 6.



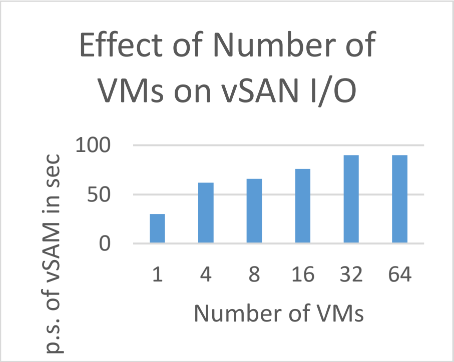


Fig. 3. Effect of Number of VMs on vSAN Latency Fig. 4. Effect of Number of VMs on vSAN I/O

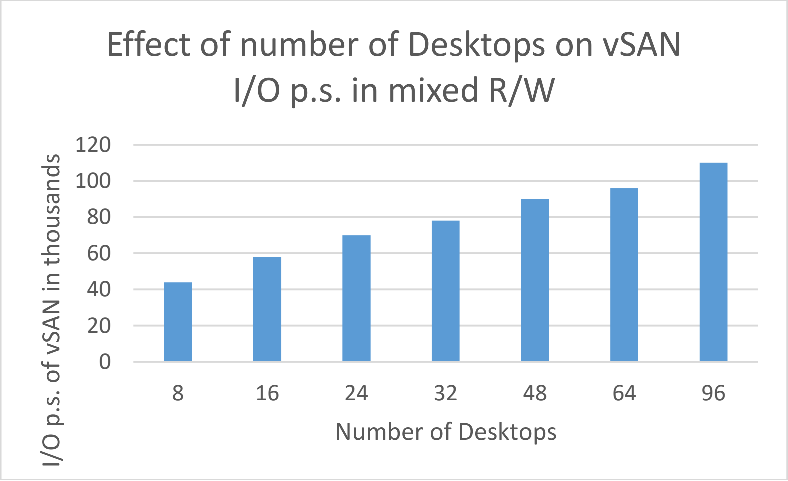


Fig. 5. Effect of number of Desktops on vSAN I/O p.s. in mixed R/W

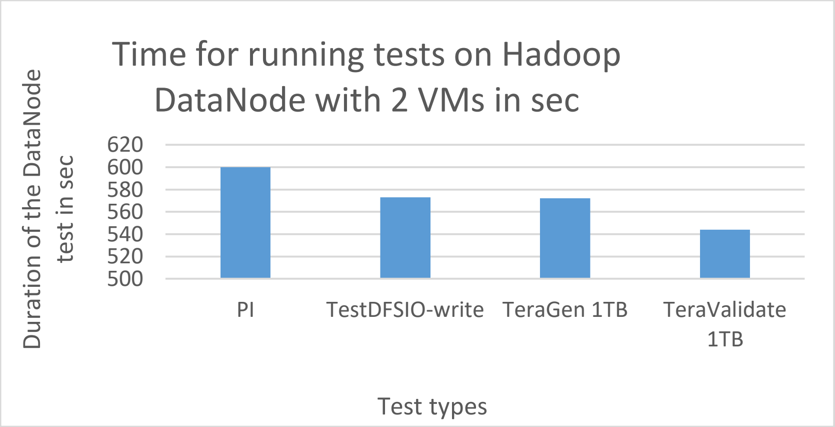


Fig. 6. Time for running tests on Hadoop DataNode with two VMs (in s)

The values of the standard deviation σ for Figs 3-6 are calculated based on the next equation:

(3) $$\sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{(n-1)}},$$

where x takes on each value in the sample (for each Figs 3-6, these are the corresponding values presented on the vertical axis); $\overline{x}$ is the sample mean; n is the sample size.

Based on the mentioned results and in order to calculate the required number (rounded number) of desktop PCs for the R&D environment of the proposed Agile Elastic Desktop Corporate Architecture for Big Data, the standard deviation value of 22.82 is considered. The calculation is based on equation (4) and the calculated number of required desktop PCs is about 500:

(4) $$n = \frac{z^2 \cdot \sigma^2}{\Delta^2} = \frac{1.96^2 \times 22.82^2}{2^2} \approx 500.$$

If the R&D environment has less than 500 desktop PCs, the problem will be not the duration in days for the tuning and testing, but for some Big Data system, the results cannot be achieved using lower number of PCs. For example, for the implementation of a small-size Hadoop system with three NameNodes and 20 DataNodes, each node uses five VMs, applying for each VM a vSAN with 4 VMs for each DataNode and a vSAN with two VMs for NameNode, where in each desktop PC two VMs are established – the first VM for the desktop functioning and the second VM dedicated to the Big Data systems, means that 430 PCs are required in the R&D environment. The conclusion is that just for testing and tuning the described small-sized Hadoop, without any additional Big Data components, at least 430 desktop PCs are required in the R&D environment.

5. Conclusion

An Agile Elastic Desktop Corporate Architecture for Big Data is proposed in this paper. That is a possible solution for solving real business problems concerning Big Data processing in companies with large ICT systems (with at least 500 desktops).

The research results reveal that it is possible to more effectively utilize the available unused desktop resources for the needs of Big Data processing. Instead of having a stand-alone R&D environment, the corporations can use their legacy desktops for Research and development purposes, as long as the number of the available desktop PCs is at least 500. In the case of big corporations with tens of thousands desktops, only part of the available PCs – about 500, could be used for the design of an Agile Elastic Desktop Corporate Architecture for Big Data systems.

The proposed architecture will be considered during the actual implementation of the project “Digital Transformation of Economy in Big Data Environment”, especially while developing the ICT infrastructure providing multi-access hosting and cloud features.

***Acknowledgements*:** This work has been supported by the project NoBG05M2OP001-1.002-0002 “Digital Transformation of Economy in Big Data Environment”, funded by the Operational Program “Science and Education for Smart Growth” 2014–2020, ProcedureBG05M2OP001-1.002 “Establishment and development of competence centers”.

References

1. Gartner Glossary. Last retrieved on 4 March 2020. **https://www.gartner.com/en/information-technology/glossary/digitalization**
2. i-SCOOP. Digital Transformation: Online Guide to Digital Business Transformation. Last retrieved on 4 March 2020. **https://www.i-scoop.eu/digital-transformation/**
3. IDC (2019). IDC Forecasts Revenues for Big Data and Business Analytics Solutions Will Reach $189.1 Billion This Year with Double-Digit Annual Growth Through 2022. Last retrieved on 2 July 2020. **https://www.idc.com/getdoc.jsp?containerId=prUS44998419**
4. Venkatram, K., M. A. Geetha. Review on Big Data & Analytics – Concepts, Philosophy, Process and Applications. – Cybernetics and Information Technologies, Vol. **17**, 2017, No 2, pp. 3-27.
5. Popchev, I., D. Orozova. Towards Big Data Analytics in the e-Learning Space. – Cybernetics and Information Technologies, Vol. **19**, 2019, No 3, pp. 16-24.
6. insideBIGDATA (2017). The Exponential Growth of Data. White Paper. Last retrieved on 2 July 2020. **https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/**
7. ict-expert.com – ICT Architecture. Last retrieved on 4 March 2020. **http://www.ict-expert.com/services/ictarchitectureen.php**
8. Agile Development Methodology. Last retrieved on 4 March 2020. **https://pm-training.net/agile-development-methodology-wiki/**
9. Agile Project Management. Last retrieved on 4 March 2020. **https://searchcio.techtarget.com/definition/Agile-project-management**
10. Desktops Standards Architecture. Last retrieved on 4 March 2020. **https://www.ulster.ac.uk/isd/services/hardware-and-software/desktops-standards-architecture**
11. Kleyman, B. (2016). Desktop Virtualization: A Pros and Cons List. Last retrieved on 2 July 2020. **https://www.mtm.com/desktop-virtualization-pros-cons-list/**
12. Taneja Group (2020). Hypervisor Shootout: Maximizing Workload Density in the Virtualization Platform. Last retrieved on 2 July 2020. **http://tanejagroup.com/files/Hypervisor-Shootout-Maximizing-Workload1.pdf**
13. DaaS Provider Comparisons. Last retrieved on 4 March 2020. **https://www.business.com/categories/best-desktop-as-a-service**
14. How to Enable CPU Virtualization in Your Computer’s BIOS. Last retrieved on 4 March 2020. **https://www.bleepingcomputer.com/tutorials/how-to-enable-cpu-virtualization-in-your-computer-bios**
15. Thin Clients or PCs for Better TCO? Last retrieved on 4 March 2020. **http://www.devonit.com/blog/vdi-at-the-desktop**
16. Gartner Inc. Gartner Says Worldwide PC Shipments Grew 1.5% in Second Quarter of 2019. Last retrieved on 4 March 2020. **https://www.gartner.com/en/newsroom/press-releases/2019-07-11-gartner-says-worldwide-pc-shipments-grew-1point5percent-in-second-quarter-of-2019**
17. Thin Client Market Gets Even Thinner, Down Seven Per Cent in a Year. Last retrieved on 4 March 2020. **https://www.theregister.co.uk/2016/03/29/thin_client_market_gets_even_thinner_down_seven_per_cent_in_a_year/**
18. What is a Data Lake? Last retrieved on 4 March 2020. **https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/**
19. What is a Datastore? Last retrieved on 4 March 2020. **https://geek-university.com/vmware-esxi/what-is-a-datastore/**
20. Comparing Virtual Volume (VVol) Limits to VMFS/NFS Limits. Last retrieved on 4 March 2020. **http://vsphere-land.com/news/comparing-virtual-volume-vvol-limits-to-vmfsnfs-limits.html**

21. What are VMware VVOLs and How Do They Work. Last retrieved on 4 March 2020. **https://searchstorage.techtarget.com/answer/What-are-VMware-VVOLs-and-how-do-they-work**
22. VMware Virtual SAN™ 6.0 Performance. Last retrieved on 4 March 2020. **http://www.vmware.com/resources/techresources/10459**
23. VMware® vSAN™ Design and Sizing Guide. Last retrieved on 4 March 2020. **https://storagehub.vmware.com/t/vmware-vsan/vmware-r-vsan-tm-design-and-sizing-guide-2/**
24. Administering VMware vSAN. Last retrieved on 4 March 2020. **https://docs.vmware.com/en/VMware-vSphere/6.7/vsan-671-administration-guide.pdf**
25. Mendenhall, W., R. Beaver, B. Beaver. Introduction to Probability and Statistics. 15th Edition. Cengage Learning, 2019.
26. VMware Virtual SAN™ 6.0 Performance. Last retrieved on 6 March 2020. **https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/whitepaper/products/vsan/vmware-virtual-san6-scalability-performance-white-paper.pdf**
27. Performance Best Practices for VMware vSphere® 6.0. Last retrieved on 6 March 2020. **https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-perfbest-practices-vsphere6-0-white-paper.pdf**

Received: 10.03.2020; Second Version: 06.07.2020; Third Version: 17.07.2020; Accepted: 29.07.2020