A Detailed Survey on the Relationship between Virtualisation, Hyperconvergence and Big Data

Harshita S[1], S Meghna[2], Saadhvi Rayasam[3], Ritesh M[4], Deepamala N[5]

[1-4] Student, Department of Computer Science and Engineering, R V College of Engineering, Bengaluru,India.

[5] Professor, Department of Computer Science and Engineering, R V College of Engineering, Bengaluru,India.

Abstract— **With the focus of world leaders in technology shifting towards Virtualization and Cloud computing, it has become a fast-growing field of technology. Earlier, corporations incurred a lot of cost in procuring hardware to deploy their applications with the absence of virtualization technology. Since the advent of virtualization it has been widely used by the industry for its cost effective deployment and flexibility. In this paper we have surveyed the impact of virtualization on the modern technological industry, how virtualization is growing and shed light on the latest developments in the field of virtualization. The types of virtualization are also discussed where we have hardware virtualization, network virtualization, server virtualization and many more. This paper also contains information about the different types of virtualization and we will also discuss the different types of hypervisors on which virtual machines are deployed and the latest developments with respect to hypervisors. Hyperconvergence is the latest advancement in virtualization and cloud computing. Hyperconvergence is combining two or more aspects of cloud computing and running both the components as one node. This has gained significance as it is more cost effective and easy to scale. This paper contains a survey on hyperconvergence and the basic technology behind it. We also enumerate the way hyperconvergence has eliminated a lot of pain points such as hardware vendor-compatibility problems, difficulty in scaling multiple machines as opposed to just one node in hyperconvergence.**

Keywords—Virtualisation,Big data, Hyperconvergence, Hypervisors

I. INTRODUCTION

With advancement in technology the IT organisations had to include multiple functionalities with their existing infrastructure but doing this resulted in excessive operating and maintenance costs, because each functionality meant an addition of another machine. This was because running multiple functionalities on a single machine hadn't been introduced until that point. Then with the introduction of virtualization the way IT organizations went about their infrastructure changed completely.

Cloud computing allows organizations to compartmentalize its software solutions such as storage, compute, operating systems and applications. This has allowed IT organizations to charge its customers according to their usage and also the customer has more flexibility on choosing a plan according to their needs[1].

Before virtualization was introduced the machines that were running were not utilizing their full resources. Which created a lot of wastage of resources but with virtualization it was found that the available resources on one particular machine can be divided logically among several virtual machines that may be running different applications. Virtualization was a game changer in terms of operating costs for IT organizations. It was also easier for the organizations to deploy new technology as there is not a lot of configuration that had to be done but just deploy an application on existing infrastructure.

Virtualization is used to create a logical layer of resources which can be used by the virtual machines deployed on the machine. Virtualization can be realised by using a hypervisor, virtual machine monitor is another term for a hypervisor [2]. A hypervisor/VMM(Virtual Machine monitor) is responsible for scheduling the resources from the physical hardware that is available on the actual machine to the virtual machines running on the host [3]. Hypervisor is the software technology for conserving and utilising resources efficiently for numerous virtual computers to work on a single physical server. A hypervisor delivers a standardised look at basic hardware, meaning that the equipment of different merchants is available on which it can work. Virtual machines can thus operate on all accessible and supported computers, given that the hypervisor rejects hardware software. This gives system administrators to see the resources as a single fabric and allot resources accordingly. Several operating systems may be installed on a single machine by using hypervisors.

Hyperconvergence is an IT architecture that integrates data centre management and scaling with storage, computing and networking into one unified system. Often on standard off-shelf servers, hyper converged solutions incorporate virtualized compute, virtualized networking and software-based storage. Multiple nodes can be combined to form pools of ease to access common computing and storage resources. [4].

Big Data plays an important role in the decision making process of organizations in the modern world. Big data can be defined as the data obtained from various sources of digital systems such as emails, transactions, blog posts, logs, posts on social media, videos, images, etc. It also involves storing, organizing, visualizing and analyzing this huge data which cannot be achieved through traditional IT practices[5]. Since it plays such an important role in decision making for various

organizations while adopting virtualization technologies the administrators should also consider accommodating Big Data practices into their infrastructure. The impact of Big Data on virtualization and hyper converged infrastructure is explained in detail in this paper.

II. VIRTUALISATION

Virtualization is a broad umbrella term for a variety of technologies that aim to create an abstract environment, whether virtual hardware or an operating system, in which one can run programmes. Virtualization is the process of producing an emulated version of anything. It also enables the use of various operating systems on the same computer. Virtualized architecture consists of four layers: application layer, operating system, virtualization layer and hardware layer.A hypervisor [6] (virtualisation layer) is a piece of software that allows various operating systems to coexist on the same hardware. It also monitors the memory and other resources of the system processors in order to allocate them according to the needs of each CPU. The hypervisor is far more efficient than alternative hosted designs, allowing for increased performance, resilience, and scalability. Each guest operating system will be assigned its own resources by hypervisors. As a result, there is no interaction between processes executing on different guest operating systems, nor is there communication using standard operating system primitives.

A. Types of Virtualisation

Virtualization is primarily used to replicate the execution environment, networks, and storage. To achieve virtualization, three basic methods are used: simulated or full virtualization, paravirtualization, and hardware supported virtualization[2].

1) Full virtualization

Emulated or full virtualization entails a complete software emulation of the underlying hardware platform's architecture, notably the hardware processor's instruction set architecture. An unmodified operating system binary and unaltered application binaries are run under emulated or full virtualization. The guest OS is ignorant that it is virtualized and does not require any modifications, despite the fact that the virtualization layer has totally isolated and decoupled it from the underlying hardware. All operating system instructions are translated in real time by the hypervisor to have the intended impact on the virtual hardware, while user programmes execute unaltered. On the negative side, a single guest operating system and its applications might use all physical memory, making resource consumption, management and performance isolation difficult.

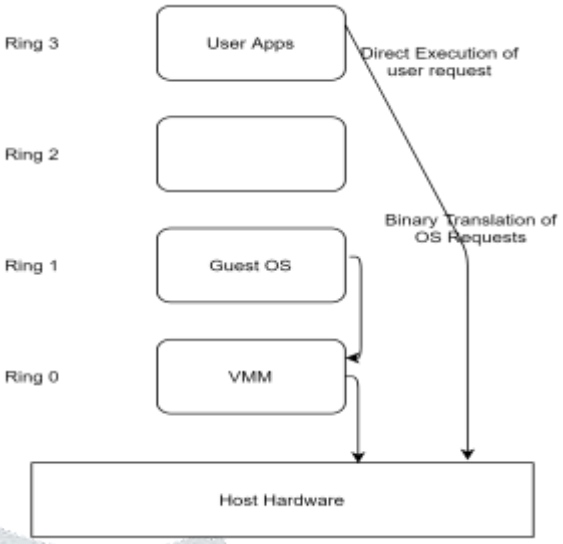

Fig. 1. Representation of Full Virtualization

2) Paravirtualization

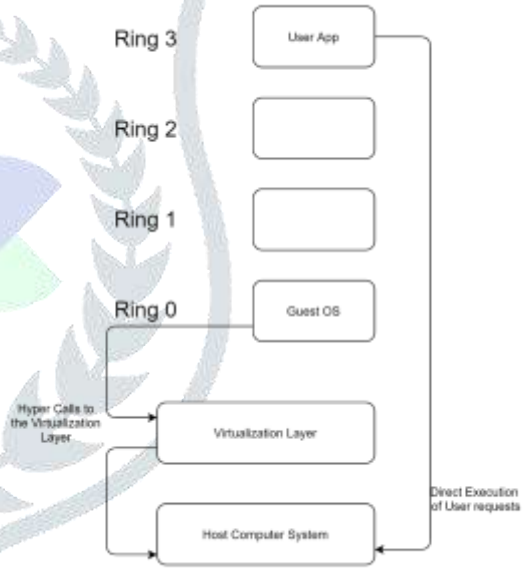

Fig. 2. Representation of Para Virtualization

In paravirtualization technology, a guest operating system and a hypervisor collaborate closely to ensure optimal performance rather than a direct software simulation of the real machine's hardware architecture as in full virtualization. However, the kernels of both the operating system and the hypervisor must be updated to support this interaction. It entails changing the operating system kernel to replace non virtualizable instructions with hypercalls that connect directly with the hypervisor. However, no changes are made to the programmes associated with the guest operating system. Hypervisors can be designed to be tightly coupled with a certain operating system or to be operating system independent. While developing sophisticated binary translation support for full virtualization is complex, modifying a guest operating system for paravirtualization is rather simple.

3) Hardware supported Virtualization

Privilege instructions are targeted with a new CPU execution mode feature in hardware-supported virtualization, allowing the hypervisor to execute in a new root mode below the privileged ring 0 for the guest OS.The hypervisor or VMM automatically traps the privileged and sensitive calls, removing the requirement for binary translation or paravirtualization. Without hardware emulation or OS modification, hardware enabled virtualization methods decrease or even remove the hypervisor's workload for trapping and simulating instructions executed within a guest operating system.

In summary, all kinds of virtualization increase computer resource usage by allowing many applications to operate on the same computer and be performed by separate users. From an environmental standpoint, it results in reduced power usage, fewer machines being employed, and, presumably,

fewer equipment being manufactured, which reduces carbon emissions upstream.

B. Application of Virtualisation

Virtualization is being used in six key applications: server virtualization, storage virtualization, network virtualization, desktop virtualization and application virtualization[7].

1) Server Virtualization

Server virtualization allows multiple operating systems and their associated applications to run on top of the hardware system without relying on the host operating system.server virtualisation allows multiple applications to run on a single server hardware or comparatively lesser number of hardware platforms. The physical features of the server are completely abstracted from the software applications that run on it. The programmes running on a VM perceive a dedicated operating system and server.The hypervisor, also known as VMM, distributes the server's memory, CPU, and input/output resources to each VM. Even though server virtualization allows software applications to run concurrently on the same physical server, they remain completely abstracted from one another and from the underlying hardware.

2) Storage virtualization

Storage virtualization is appropriate for meeting the requirement of the increasing expansion of stored data, there is a desire for storage systems that can be expanded without shutting down systems or limiting the number of physical discs supporting the storage system. Storage virtualization aggregates physical storage from multiple network devices into one virtual storage unit. Storage discs may be added or replaced using virtualization software, and replications, backups, snapshots, and mirrors may be produced without causing downtime. It boosts IT workers' efficiency by allowing them to consolidate various storage-related maintenance tasks. Storage virtualization can also be used to disguise or hide storage volumes from servers that are not permitted to access them, adding an extra layer of protection.

3) Network virtualization

The number of network servers that may be efficiently integrated on a single physical computer is directly proportional to the efficiency of network usage. Network virtualization allows multiple groups to access the same physical network while remaining logically separate to the point of being invisible to each other. It is achieved by partitioning a single physical network into multiple virtual networks, thereby not only minimising capital and operational expenses and carbon emissions, but also providing an organisation with rapid scaling up capability to meet new business needs by incrementally adding new partitions. Network virtualization improves application performance by dynamically increasing network asset use while lowering operating costs.

4) Desktop virtualization

Desktop virtualization is a relatively new use of virtualization when compared to other types of virtualization. It entails providing the end-user with a desktop environment that provides access to any approved programme regardless of its location. The technique of generating a virtual version of a user's operating system and desktop environment that is independent of the end user's computing hardware or client is known as desktop virtualization. The user may now access his or her desktop from any computer device. There are two major desktop virtualization architectures: client hosted and virtual desktop infrastructure.

5) Application virtualization

In the application layer, virtualization isolates software programmes from the hardware and the operating system, essentially enclosing them as separate, mobile objects that may be transferred without disrupting other systems. It delivers an application to the end user without requiring the programme to be installed on the end user's client or local system. Local resources are used by applications that operate locally. Application virtualization provides more flexible IT application deployment, streamlines administration, and helps resolve application conflicts. Application virtualization minimises the complexity and IT labour associated in installing, updating, and administering applications by transforming them into virtual services that are managed and hosted centrally but operate on demand locally. As applications are no longer competing for shared resources in their environment, testing for incompatibilities with existing programmes is considerably reduced. Application virtualization also enables applications to be updated or rolled back while still in use.

III. BIG DATA

Big data is the terminology that is used to describe the large volume of data that can be both structured and unstructured that a business or an organisation faces on a daily basis. This data is too immense to be handled with an approach of data handling done traditionally or the existing software techniques. In order to address this issue, virtualization has introduced a broader range of measurable benefits in recent decades. In recent times there is an increased and widespread recognition of the importance and value of data and the results obtained by analysing it, in fact, Big data is the new promising buzzword in the industry and the talk of the town. It has become the centre of truth for businesses over the last decade as they use collective data insights to align strategy, assist teams in collaboration, uncover growth avenues, and compete in the global marketplace. Originally, big data had three defining properties, viz, the three Vs; volume, velocity, and variety. However, two more Vs, namely, veracity and value were also added. These characteristics indicate that 'Big data' is so much more than just a huge amount of data[8].

.

A. Characteristics of Big Data

1) Volume

As the name suggests, 'Big Data' refers to the sheer enormity of the size of data. These volumes of data can reach unprecedented heights in fact. The global datasphere is expected to reach 175 zettabytes by 2025. By 2023, the big data industry is estimated to be worth approximately the big figure of $77 billion. Big data can be visualised as a pyramid, in which the volume would be the base.

2) Velocity

Furthermore, data is being produced at an alarming rate. Velocity essentially measures the increasing rate at which data is generated in order to gauge the increasing rate at which data can be analysed, processed, and stored. As the importance of data is increasing and the business environment today is data-driven,,this rate is best described as "unprecedented" and "torrential". This data should now be apprehended as soon as possible.

3) Variety

Refers to the processing of various data types gathered from numerous data sources. These data sources could include both external and internal business units. Big data can be classified as structured, semi-structured, or unstructured data. The immense volume of data that most organizations capture and generate may appear disorganized and unstructured. In fact, unstructured data like photos, mobile data, videos and social media content comprise close to 80% of global data.

4) Veracity

Refers to the assurance of quality or credibility of the collected data. Considering the vastness of big data, and the fact that it is sourced from numerous sources, there is a good chance that the entirety of the collected data will not be of good quality or accurate in nature. Thus, it is of utmost importance to ensure that the validity and precision of the data is checked before proceeding for processing.

5) Value

Refers to the worthiness of the data with respect to positively impacting analytics that in turn enable informed decision. Aggregation of data doesn't equal value addition, i.e. aggregated data can only be valuable if meaningful insights can be gained from it. It must be ensured that the value of big data is significant and profitable to be investing time and effort into.

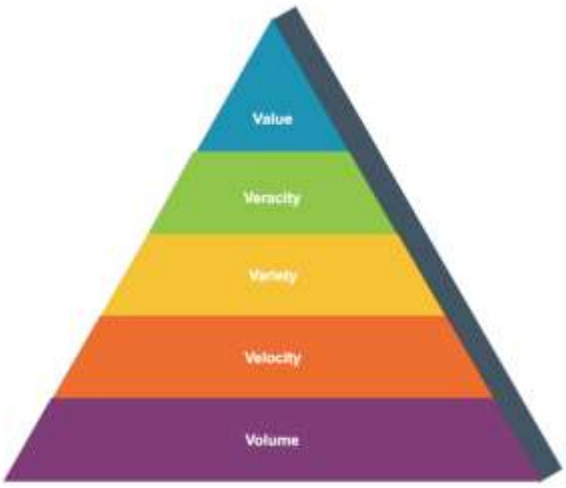

Fig. 3. Representation of the 5 Vs of Big Data

IV. BIG DATA AND VIRTUALISATION

To handle Big Data challenges, it is necessary to have effective management of massively distributed data and data-integrated applications.Virtualisation has made big data management a reality by adding an extra degree of efficiency to the process of constructing big data platforms. This is not the only available option to optimise the management process of big data, but the employment of virtualisation helps in making the data organisation more generic.In comparison to our traditional storage solutions, which are incapable of satisfying the criteria for dealing with big data, virtualization enables greater flexibility in consuming and controlling large data from any location. Big Data and virtualization have diverse uses, yet they are related in several ways. Some of which is discussed below [5].

1) Big Data and Server virtualization

Virtualization enables all the physical resources (RAM, network, CPU, etc) to be virtualized and converted into a set of virtual machines where each VM has its own OS and applications running on them.It is an authorized software approach that allows many applications and operating systems to operate on the same hardware platform concurrently. Hypervisors or Virtual Machine Monitors (VMM) are utilised to offer efficacy in the usage of physical resources.

Server virtualization aids in Big Data analysis management by increasing IT agility, suppleness, and scalability. Workload deployment is quicker, output and accessibility are improved, and core procedures are performed automatically. It is useful to validate that any platform can expand as needed to handle massive amounts of data of various sorts, including Big Data analysis. It keeps costs down. Server virtualisation serves as the foundation for many of the cloud services utilised as data sources in big data analytics. It also improves cloud efficiency, making many complicated systems easier to optimise.

2) Big Data And Storage virtualization

Storage virtualization combines data from multiple physical storage devices into a single unit that appears to be a single storage device.Storage virtualisation plays a critical role in big data analytics, when data has to be stored and accessed from different locations. To enhance the process and network-independent storage management, the approach involves abstracting and masking the basic characteristics of a storage device from the host devices or the network. By doing so, it has been considerably able to cut storage costs and therefore make it simpler to manage and distribute all data.

3) Big data and network virtualization

Maintaining all of the storage and data from one location would be a difficult effort without a reliable network. The physical network's reliance and management will be reduced if virtual networks can be established and used efficiently.Instead, a virtual network can eliminate unacceptably frequent interruptions while also improving the capacity to manage vast amounts of dispersed data necessary for big data research.

V. HYPERCONVERGENCE

Hyperconvergence is a framework that combines at least any two of storage, compute and network in order to improve scalability while decreasing complexity of the data center. Hyper-Converged Infrastructure (HCI) is an infrastructure that has evolved from silos of hardware-based solutions in which the compute is separate from a storage network to offer highly virtualized solutions in which the compute, storage, and network are all virtualized. A hypervisor in a hyperconverged platform is used for virtualized computing, software-defined storage, and virtualized networking.

For simplified management, converged infrastructure was introduced by IT companies. This is a preconfigured software and hardware into a single system. But, the storage, compute and networking components are discrete and can be separated. With a hyper converged environment, components are more tightly infused on a software level and cannot be separated. Organizations can easily expand capacity by deploying additional modules. Hyperconvergence thus improves on abstraction and automation[4].

A. Advantages of Hyper Convergence

Organizations are increasingly looking to adopt hyper convergence to simplify management of resources and lower costs by combining storage, networking and computation into a single system[4]. The major benefits of hyper convergence are as follows.

1) Software-defined storage

Storage nodes in a hyperconverged environment are software defined, making them highly reliable. It is essentially a redundant pool of storage. Thus, resilience is key to ensuring excellent uptime.

2) Agility and scalability

There is single administration in hyper converged infrastructures. So migration of workloads from one location to another is easier.

As mentioned earlier, the node-based architecture makes it simple to scale up the data centre.

3) *Disaster recovery and backup*

Hyper convergence offers data protection by seamlessly allowing restoration of data. This is achieved by certain components within the hyper converged infrastructure which are dedicated to provide backup and disaster recovery.

4) *Economical solution*

Lesser equipment, lower maintenance and support costs mean data centres are now less taxing financially. Thus, any organization with a data centre would be eager to switch to the hyper converged model.

VI. BIG DATA AND HYPERCONVERGENCE

The biggest challenge with Big Data is that the volume keeps rocketing. Since all data could be insightful, there is no longer historical, obsolete data that can be discarded, which means storage space grows incessantly. Cloud storage could be a potential solution to the constantly growing footprint, and the challenge of scaling storage, however, businesses are apprehensive about this due to security concerns as they would rather have complete control over their prized data. Other challenges are accommodating the necessary servers required to process huge volumes of data in the data centre. The most popular tool for big data analyses, Hadoop, relies on a scalable cluster of servers, with calculation tasks happening close to the data. That means businesses need to build a large suite of servers, which will mostly be idle most of the time. This can be a costly affair, even with Hadoop's ability to use relatively inexpensive, commodity servers.

Hyperconverged infrastructure can solve the challenge of fitting big data into data centres of businesses, with a smaller footprint. This is because, by design, hyperconverged infrastructure treats storage and compute as a single unit. It can be made to scale out by simply adding nodes. This perfectly satisfies the demand that comes with Hadoop and big data. Furthermore, hyperconverged infrastructure offers virtualization which makes big data fit in the data storage even more efficiently. It serves for the increase in the server utilization, and also makes the estimations cost-effective. It simplifies the data complexities and also increases the flexibility, scalability, and efficiency of data storage infrastructure. This makes it possible to manage more data in less time and with the available data storage capacity which can also be increased by easy and seamless scalability. Thus, hyperconvergence does not just solve the handling of large volumes of data with maximum efficiency and optimization of the resources, but also offers a cost-effective solution to the big data problem.

VII. CONCLUSION

The paper began by introducing the terms big data, virtualization and hyper convergence. In the next section, types and applications of virtualization are explained. Enterprises use virtualization to reduce IT costs and to effectively utilize resources. The next section displayed the need for virtualization, and how it can be leveraged to make the most of big data. Virtualization also improves scalability and performance of Map reduce engines.

The next section introduced hyper convergence and how it is better than converged infrastructure. Organizations can expand data centre(s) without any hassle. The other major advantages of hyper converged infrastructure for an enterprise was also highlighted. This compelling argument sparked the discussion for using hyperconvergence as a solution to the big data problem, i.e storing and processing the ever increasing volumes of data. Hyperconvergence does not just solve the handling of large volumes of data with maximum efficiency and optimization of the resources, but also offers a cost-effective solution to the big data problem.

ACKNOWLEDGMENT

The authors received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

[1] Budhprakash, Dr.Anupam Bhatia, Dr. GurjeetsinghBhatta, A comparative study of Various Hypervisors Performance, *International Journal of Scientific & Engineering Research*, Volume 7, Issue 12, December-2016

[2] Durairaj. M, Kannan.P, "A Study Of Virtualization Techniques and Challenges in Cloud Computing", *International Journal of Scientific & Technology Research*, Volume 3, Issue11, November 2014

[3] Sonam Srivastava, S.P Singh, A Survey on Virtualization and Hypervisor-based Technology in Cloud Computing Environment, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* Volume 5 Issue 2, February 2016

[4] Shaikh Abdul Azeem, Dr. Satyendra Kumar Sharma, "Study of Converged Infrastructure & Hyper Converged Infrastructure As Future of Data Centre", *International Journal of Advanced Research in Computer Science*, Volume 8, No. 5, May-June 2017

[5] Selina Sharmin, Asoke Datta, Md. Nurain Haider," Big Data & Virtualization: Concept familiarization and relation between them", *IJEDR 2018*, Volume 6, Issue 3

[6] Bohar Singh , Gursewak Singh, "A Study of Virtualization and Hypervisor In Cloud Computing", *International Journal of Computer Science and Mobile Applications*, Vol.6 Issue. 1, January- 2018

[7] Sharma, Srinarayan & Park, Young,"Virtualization: A Review And Future Directions Executive Overview", *American Journal of Information Technology*, May 2011

[8] Mandeep Kaur Saggi, SushmaJain, "A survey towards an integration of big data analytics to big insights for value-creation", *Elsevier*, -, Volume 54, Issue 5, September 2018

[9] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[10] Shaikh Abdul Azeem and Dr. Satyendra Kumar Sharma," Study of Converged Infrastructure & Hyper Converge Infrastructre As Future of Data Centre", *International Journal of Advanced Research in Computer Science*, Volume 8, No. 5, May-June 2017

[11] Shyam Patidar, Dheeraj Rane and Pritesh Jain "A survey paper on cloud computing" *2012 Second International Conference on Advanced Computing and Communication Technologies* on 7-8 January 2012 pp. 394-398 ISBN: 978-1-4673-0471-9 DOI: 10.1109/ACCT.2012.15 2012 IEEE.

[12] Farzad Sabahi "Secure virtualization for cloud environment using hypervisor-based technology" *International Journal of Machine Learning and Cloud Computing*, Vol. 2, No. 1, February 2012.

[13] Zongzian He and Guangqing Liang "Research and evaluation of network virtualization in cloud computing environment" *2012 Third International Conference on Networking and Distributed Computing* on 21-24 October 2012 pp. 40-44 ISBN: 978-1-4673-2858-6 DOI: 10.1109/ICNDC.2012.18 2012 IEEE

[14] Krishna Tej Koganti, Eshwar Patnala, Sai Sagar Narasingu and J.N Chaitanya "Virtualization technology in cloud computing environment" *International journal of Emerging Technology and Advanced Engineering* Vol. 3, Issue 3, March 2013

[15] J. Pereira, E. da Silva, T. Batista, F. Delicato, P. Pires and S. Khan, "Cloud Adoption in Brazil", *IT Professional*, vol. 19, no. 2, pp. 50-56, 2017

[16] Zaharia M, Chowdhury M, Franklin MJ, Shenker S. "Spark SI: Cluster Computing with Working Sets". *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing* 2010, 10–10.

[17] N. Totmelina, R. Anashkin, A. Kiryanov and V. Sirotkin, "Algorithm for distributed data storage software solutions for automation of small and medium enterprises in cloud", *Modern problems of science and education*, no. 3, 2013.

[18] V. K. Manik and D. Arora, "Performance Comparison of Commercial VMM: ESXI XEN Hyper-V & KVM", *2016 3rd Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 1771-1775, 2016.

[19] A. Krasov and A. Shvidkiy, "Using the possibilities of scaling the cloud infrastructure to optimize the process of creating laboratory stands", *conference proceedings of APINO*, vol. 2, pp. 1580-1584, 2015.

[20] G. Rastogi and R. Sushil, "Cloud Computing Implementation: Key Issues and Challenges", *Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 3-7, 2015

[21] D. Marinescu, "Cloud Computing - Theory and Practice", *Morgan Kaufmann*, 2017.