International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)

# Big Data Analytics with Cloud Databases: Efficiency and Cost Optimization

**N V Rama Sai Chalapathi Gupta Lakkimsetty**
Independent Researcher, USA

**Article Info**
Volume 6, Issue 2
Page Number : 599-607

**Publication Issue :**
March-April-2020

**Article History**
Accepted : 01 March 2020
Published : 20 March 2020

**ABSTRACT**

A revolutionary synergy, big data analytics with cloud computing allows for the analysis and processing of large datasets with previously unheard-of scalability and efficiency. The cloud overcomes the drawbacks of conventional on-premises systems by offering a flexible and affordable infrastructure for big data management, storage, and analysis. With the help of this combination, businesses can fully use big data and extract useful insights that inform choices and spur creativity. Big data analytics and cloud computing are combined to take use of cutting-edge technology like artificial intelligence, machine learning, and distributed computing. They are constantly current, which is why they use online resources instead of going to libraries in person. Users may store data on a large scale, maintain backups, and protect themselves from calamities, among many other advantages of cloud computing. Libraries may now store enormous volumes of data on the websites and electronic databases because to the development of cloud computing. All of this data will be stored securely. Cloud computing is a technique that makes a virtual platform available on library websites. It handles all of the data that is easily accessible over the internet. A popular IT strategy for meeting the requirements of several commercial and scientific Big Data applications is cloud computing. In this research, we provide a Hadoop platforms deployment technique using the Occopus cloud the orchestrator system tool for several cloud infrastructures. With the primary objective of preventing vendor locking issues, our automated solution offers a simple, portable, and scalable method of deploying the well-known Hadoop platform; that is, it does not rely on any cloud provider's prepared and provided virtual machine image or "black-box" Platform-as-a-Service mechanism. The study offers cost analysis and encouraging performance assessment outcomes.

Keywords : - Big Data Analytics, Efficiency and Scalability, Cloud Computing, Cost Analysis, Hadoop Platform Deployment, Platform-As-A-Service Mechanism, Performance Measurements, Physically Visiting Libraries.

## I. INTRODUCTION

Businesses are flooded with enormous amounts of data, both structured and unstructured, from a variety of sources, including social media, online transactions, and Internet of Things devices, in today's data-driven world. For businesses aiming to use the potential of machine learning, data science, and analytics, this phenomenon—known as big data—offers both possibilities and obstacles [1]. These enormous datasets can now be stored, processed, and analysed in a scalable and economical manner thanks to cloud computing, which has changed the game [1, 2]. Examining the core features of big data, the benefits of using cloud-based analytics, and the transition to the ELT (Extract, Load, Transform) paradigm, the essay will examine the complex link between big data and cloud computing. Additionally, it will include information on Google's Big-Query, a server less information warehouse solution that gives companies access to SQL-based analytics on petabytes of data. Furthermore, [2], we will look at actual use examples and experiments that show how well cloud-based big data analytics can drive business intelligence & innovation across a range of sectors.

*"Cloud computing can be defined as a model that facilitates widespread, convenient, on-demand network access to a shared pool of reconfigurable computing resources (such as servers, storage, apps, networks, and [2, 3]) that can be quickly provisioned and released with little involvement from service providers or management,"*

According to NIST, the National Institute of Standards and Technology. "The use of cloud computing is defined on the basis of the above model." Only lately has the term "cloud computing" been widely used and found use in a wide range of industries [2, 3]. Software as a service, hardware as a service, and platforms as a service are the three main methods via which cloud computing provides its services [3, 5]. Simply described, "cloud computing" is a method of storing and retrieving vast volumes of data from any place having an internet connection and the appropriate IT tools.

In order to digitise their data, analyse it, and get the predictions and insights that drive their operations, modern organisations worldwide are adopting a range of state-of-the-art cloud technologies [5, 6]. Numerous sectors are seeing new possibilities as a result of the convergence of IoT, cloud-based computing, & big data analytics. These businesses include, for instance, online shopping, advertising, and healthcare [5, 6]. These options could become clear as time passes. Modern businesses must become totally digital, but not all of them can afford to make significant investments in the infrastructure that they need [5, 6], IT resources (such computers, software, memory, and fast connections), and IT staff to maintain everything functioning properly.

Cloud computing provides a Web-based architecture that enables the autonomous collection, utilisation, and management of computational resources. Thanks to its fully-equipped data centres and tightly-connected resources, cloud computing offers a wide variety of services that are accessible to both individuals and organisations [6, 7]. To meet end users' needs, it would be feasible to dynamically provide them with the resources that are accessible. By removing worries about the source, scalability, and resource constraints of that underlies service, cloud computing enables people or organisations to access and use the whole computation resource pool [6, 8]. Because of this, companies and individuals may pay for the cloud services they use rather than having to make a commitment up front. Better healthcare for people worldwide has resulted from the ease with which technology and medicine may now be combined thanks to the Internet of Things' (IoT) explosive growth and the widespread use of mobile phones [8, 9].

Data has been shown to be a highly effective and helpful tool for improving health. Using the latest developments in data science and analytics driven by

AI and ML, compliance & success are being guaranteed in the retail trade, e-commerce, healthcare, telecom, [8, 9], and been sectors. When it comes to storing, processing, and research, real-world data poses many major obstacles because of its heterogeneity. Utilising state-of-the-art big data and cloud-based components is essential for processing high velocity, variety, and volume data [8, 9]. Information streaming from several sources in a range of forms, including data that is structured as well as unstructured, is referred to as a huge volume of data [29, 30].

The components offered by cloud computing providers enable data storage, processing, and consuming on an end-to-end secured platform. However, it is still unclear what resources must be made accessible and how to construct the architecture with all of these components [8, 9]. By the end of 2020, the public cloud is expected to reach $411 million, according to Gartner, given the field's recent impressive rise in cloud computing. The widespread usage of cloud computing has made it increasingly difficult for businesses to choose among the various alternatives available a dependable provider of cloud computing that can satisfy their long-term requirements [9, 11]. There is currently no industry standard for cloud service providers, and they are all expanding at the same pace. A large number of these service providers concentrate on computing power and provide storage, database, networking, or Central Processing Unit (CPU) services to end users [10, 11].

## II. RELATED WORK

The analysis of big data has completely changed how businesses handle, store, and analyse enormous amounts of data [12,13], which has led to a rise in the use of cloud computing technologies databases. Traditional data storage solutions often find it difficult to meet the needs of scalability, flexibility, & cost-effectiveness as organisations produce ever-increasing amounts of data that is unstructured, semi-structured, and [13, 14]. A more flexible approach to large data management is offered by hybrid cloud the form of databases, which combine the advantages of both private and public cloud systems [15]. These databases provide businesses the flexibility to strike a balance between operational and production efficiency. By dividing up data processing and storage responsibilities across on-premises and cloud environments, hybrid cloud databases enable businesses to maximise their data management strategy [16]. This design improves scalability and availability while addressing a number of important issues, such as data sovereignty, latency, and compliance.

Another important element affecting the adoption of hybrid cloud databases is their cost-effectiveness [18, 19]. While public cloud services provide a pay-as-you-go approach that may drastically lower upfront costs, conventional on- premises data centres sometimes need considerable capital expenditures in hardware, maintenance, and staff [19, 20]. By using public cloud scalability for sporadic workloads and retaining control over key data bases on private infrastructure, hybrid cloud databases enable enterprises to achieve a balance between these two approaches.

### 2.1 Big Data Analytics in Cloud Computing

Because cloud computing offers scalable and affordable infrastructure and services, it has completely changed how businesses manage big data analytics [20]. With products like Infrastructure as a Service ( IaaS) as well as Platform as a Service (PaaS), it makes it possible to store, process, and analyse large datasets efficiently. Using cloud computing for large-scale data analytics has many important benefits (Fig. 1):

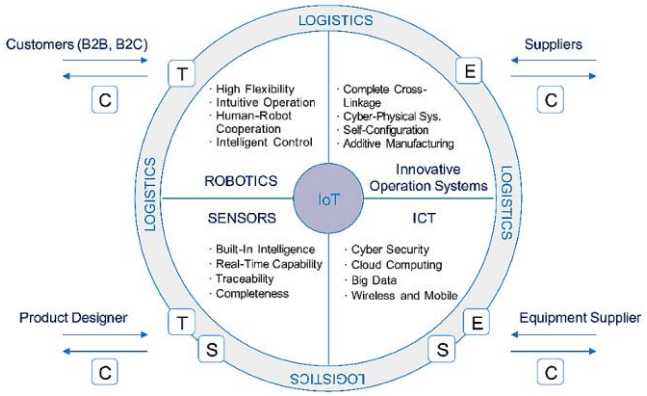


**Fig. 1** CJ Logistics' technology, engineering, system, and solution plus consultancy (TES + C). [18]

- **Scalability:** As data processing requirements increase, cloud computing makes it simple for businesses to expand their computing capacity [19, 20], guaranteeing they have the resources needed to manage growing data volumes.
- **Cost Optimization:** Businesses may drastically cut the expenses of developing and maintaining physical facilities for massive-scale data analytics by using cloud-based services [20, 21].
- **Faster Innovation:** Cloud-based analytics systems help businesses remain ahead of the curve by facilitating rapid deployment and testing with innovative analytics solutions. But it's crucial to remember that cloud data governance management may be more complicated than physical solutions, and that putting based on the cloud big data analytics into practice calls for certain knowledge and abilities [21, 22]. [22, 23].

To maximise their spending, organisations should also use cost management techniques and routinely assess how much time they spend on the cloud. Gathering data from various sources, storing it in a "the landing zone," converting it, and finally combining it for analytical activities are the usual steps in the big data analytics cycle [23]. The ELT (Extract-Load-Transform) paradigm, which performs the compute-intensive transformation on the cloud, is replacing the conventional ETL (Extract-Transform-Load) paradigm [23].

The vast amount of organised, semi-structured, and unstructured data that may potentially be mined to find out details is referred to as "big data." In this instance, however, the data is too big, complicated, and expanding quickly to be processed using conventional databases and software methods [24]. To examine this vast amount of data, conventional sequential data processing techniques are insufficient. This phenomena necessitates the development of novel data processing and analysis methods, tools, and tactics.

Over the years, cloud infrastructures have been developing steadily and are essential to the solution of applications based on big data [25]. These great challenge apps need a lot of IT resources, which cloud providers can make available to users on demand and in an easy-to-use manner by using virtualisation technologies and quick elasticity, among other things. However, once they begin to plan the usage or implementation of any Big Data platform on cloud(s), data scientists encounter a number of challenges [25].

These scientists might execute their Big Data applications more efficiently by combining Hadoop, Cloud, and an orchestration tool for dynamically building up Hadoop clusters [22]. With all of its configuration & network architecture, complex virtual infrastructures like Hadoop need specialised end-user planning, maintenance, and expertise to operate properly.

The paper's strategy ignores data localisation in favour of concentrating on Hadoop's dynamic deployment and scalability. Data transport between data storage as Hadoop is one of the most costly activities when it comes to data localisation, and it may be substantial at the petabyte level. Two simple situations may be used to decrease the cost of data transport [21]. The cost may be considerably decreased, for instance, if the Hadoop deployment and storage are situated on the same networking segment, as is the case in our lab. Utilising the transmitted data and doing several computations on it

might also lower the cost (per calculation). Other application-specific and environment-specific solutions may exist in addition to these two straightforward instances [11].

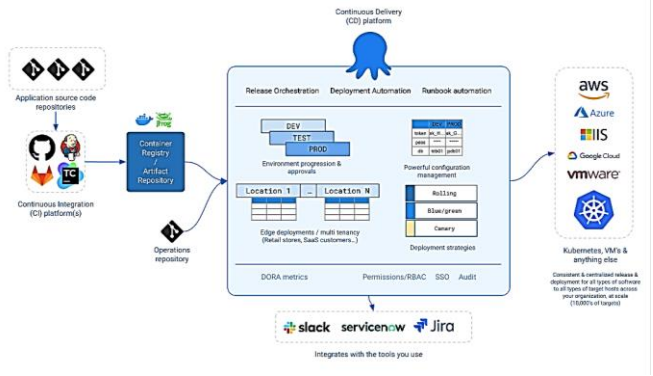


**Fig. 2** Occopus cloud orchestrator characteristics. [11, 13]

## III. DESIGN CONSIDERATIONS AND IMPLEMENTATION DETAILS

A higher degree of abstraction of the suggested and realised design is shown in Figure 3. To guarantee the safe and dependable functioning of configuration management systems, advanced virtual infrastructure maintainers (experts) may write descriptors for these systems and keep track of subcomponent descriptions [28]. As a result, end users of virtual infrastructure may use sophisticated applications without worrying about their performance, dependability, or integrity [14], regardless of the load. We will talk about the successfully developed descriptions for a cluster of Hadoop databases setup. On the target cloud, end users may build and manage virtual infrastructures as needed. As a kind of "throw-away" virtual infrastructure, it may be constructed for a little period of time [14, 15] to support a single Hadoop task and for an extended period of time, provided that it is decided to be destroyed [15, 27].

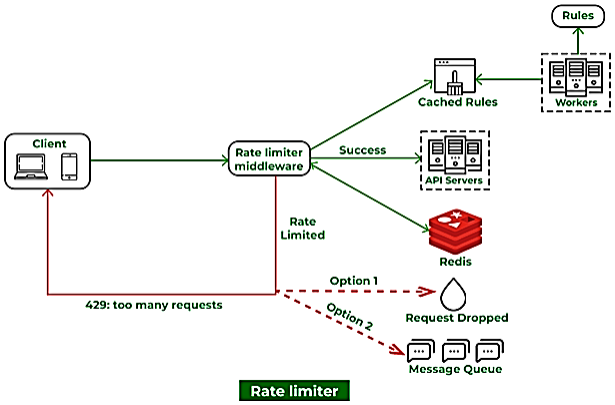


**Fig. 3** High-level architecture view with selected major components. [19]

### 3.1 Virtual infrastructure descriptors for Hadoop in Occopus

A Hadoop Master and a Hadoop Slave node are defined in Figure 3 of the infrastructure description. In our definition of dependence between them, the slave node relies on the master node [18, 19]. This guarantees that Occopus will launch the Master nodes first, followed by the concurrent operation of all slaves.

### 3.2 Extended Hadoop cluster architecture overview

One Hadoop Master Node and several Hadoop Slave nodes may be included in a Hadoop cluster, according to the Hadoop cluster architecture [19]. Occopus deploys every node automatically depending on the descriptors.

### 3.3 Hadoop cluster customisation and fine-tuning

Hadoop configuration XML files are published to the core-site, hd-fs-site, yarn-site, and map-red-site nodes when it is cloud-based [19, 20]. Before the infrastructure is launched, end users have the ability to modify the configurational files.

### 3.4 Automatic scaling of Hadoop cluster

As previously mentioned, our Hadoop infrastructure's master and slave component are already ready for manual scaling, i.e., [19], and new slaves may join or leave the cluster. The implementation of an extra control loop is a step towards autonomous scaling.

## IV. PERFORMANCE EVALUATION AND COST ANALYSIS

Numerous variables greatly influence how quickly the deployment procedures proceeds, including;

(i) The target cloud's and its interface' performance, [19, 20],
(ii) The virtual machines' performance and data connectivity speed after launch,
(iii) How quickly the Hadoop software was installed, and
(iv) The Occopus tools own effectiveness.

Thus, the effectiveness of the suggested solution Table 1 only slightly affects the amount of time needed to establish a Hadoop cluster on the cloud [20, 23]. We have conducted two kinds of measurements such as indicators of profitability and technical characteristics (deployment and scale-up time).

Table 1 Hadoop cluster deployment duration across several cloud computing infrastructures. [20, 21]

| Number of Hadoop Slave nodes | Deployment time (min) | | |
|---|---|---|---|
| MTA Cloud | 5.08 | 4.96 | 5.09 |
| LPDS Cloud | 4.09 | 2.01 | 2.96 |
| Cloud Sigma | 3.96 | 3.69 | 4.96 |

### 4.1 Scale-up time of Hadoop clusters

The scenario we demonstrate involves scaling up a cluster from two Hadoop Slave nodes to ten Hadoop Slave nodes. We analysed the time interval between the fired controls are used and the expanded cluster's fully functioning structure [20, 21]. The outcomes may vary greatly for the reasons listed above. The average scale-up time to the specified number of nodes is the end result of eight iterations of the tests conducted on the MTA Cloud [22]. The findings' diagram, or the Hadoop cluster's elasticity, is shown in Fig. 4.

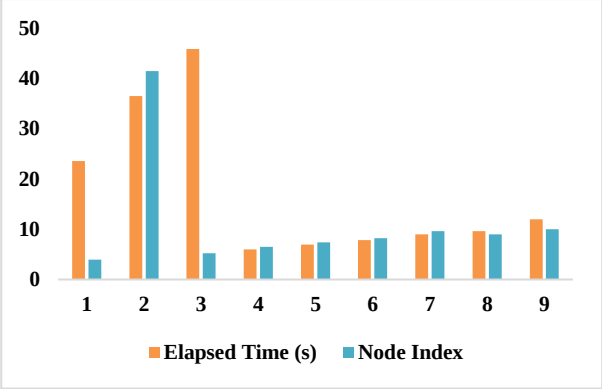


**Fig. 4** Hadoop cluster's scale-up duration. [22]

### 4.2 Cost Analysis

In our last test, [23,24], we looked at how much money might be saved by using the Occopus option rather than utilising HDInsight [25], a service offered by the Azure Hadoop platform. Hadoop clusters have been established in the Microsoft Azure clouds using virtual machines driven by Occopus throughout the investigation (Fig. 5). Comparing the same number of powerful virtual computers inside the Hadoop cluster was also crucial for this investigation [25, 27]. We utilised HDInsight's basic A3 virtual machines, which are in the weakest category, running Hadoop version 2.6.0 on Linux. Each worker and head node in this cluster contains four cores, seven gigabytes of random-access memory, eight hard drives, and load distribution. The anticipated cost of the HDInsight cluster is around e1.08 per hour [26].

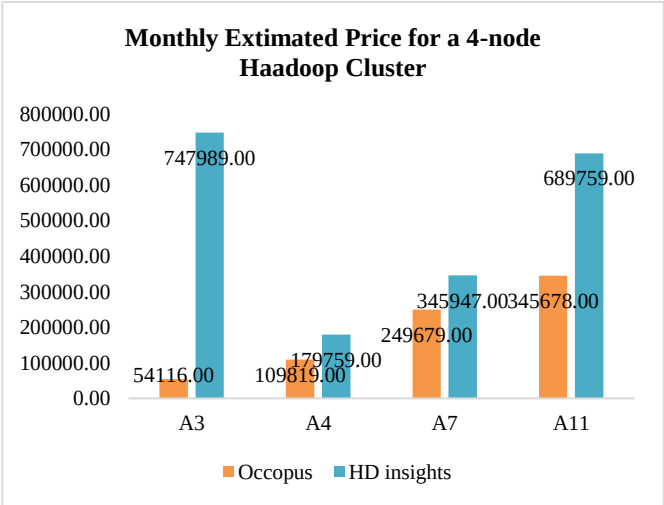


**Fig. 5** Cost comparison between Occopus' Hadoop cluster and off-the-shelf HDInsight on the Microsoft Azure cloud.

## V. CONCLUSION

We have discussed a number of problems with the current commercial and scientific virtual Hadoop infrastructure options, including Amazon's Elastic MapReduce and other scholarly tools, in this study. We have given an overview of Apache Hadoop and its operation, with a focus on operators. We looked at the Occopus orchestration tool's functionality and benefits. This study described Occopus's cloud-based Hadoop cluster implementation, which is completely automated and scalable. With the help of this technology, MTA Cloud data scientists may build intricate virtual Hadoop infrastructures for scientific investigations with short or lengthy life cycles.

This approach does not rely on pre-compiled images or proprietary management solutions that cloud providers provide as black box services, and it is compatible with a variety of major clouds (including EC2, Nova, and others) thanks to the Occopus tool. In comparison to the hours or days required to execute big Hadoop applications, the overhead associated with setting up and taking down a Hadoop cluster is essentially nothing. To test and assess the deployed solution on various cloud infrastructures from both a technical and budgetary standpoint, a number of tests were conducted. As of Q1 2017, MTA Cloud customers may access the described solution openly.

The description of a sustainable Hadoop cluster and the Occopus tool's source code are accessible online.

With the primary goal of creating a new knowledge centre and precision agriculture decision support system based on sensor data gathered from Hungarian farms and semi-structured data from global aggregated repositories, the developed solutions are being further developed towards an Internet-of-things back-end different platforms in the AgroDat.hu project.

## REFERENCES


[1]. Pradhananga, Y., Karande, S., & Karande, C. High performance analytics of big data with dynamic and optimized Hadoop cluster. IEEE.

[2]. Dawelbeit, O., & McCrindle, R. A novel cloud based elastic framework for big data preprocessing. In IEEE Conference Publications.

[3]. Gonzales, J. U., & Krishnan, S. P. T. Building your next big thing with Google Cloud Platform.

[4]. Singh, M. P., Hoque, M. A., & Tarkoma, S. A survey of systems for massive stream analytics.

[5]. Ambeth Kumar, V. D., Ashok Kumar, V. D., Divakar, H., & Gokul, R. Cloud enabled media streaming using Amazon Web Services.

[6]. Subia, S. (2018). Data Storage SpringerDOI: 978-3-319-21569-3_7 10, Procedia Computer Science.

[7]. Nakhimovsky, A., & Myers, T. Google, Amazon, and beyond: Creating and consuming Web services.

[8]. Mohanty, H., Bhuyan, P., & Chenthati, D. Chapter 2: Big data architecture. In Big data: A primer.

[9]. Begam, S. S., Selvachandran, G., Ngan, T. T., & Sharma, R. (2020). Similarity measure of lattice ordered multi-fuzzy soft sets based on set theoretic approach and its application in decision making. Mathematics, 8, 1255.

[10]. Thanh, V., Rohit, S., Raghvendra, K., Le Hoang, S., Thai, P. B., Dieu, T. B., Ishaani, P., Manash, S., & Tuong, L. (2020). Crime rate detection using social media of different crime locations and Twitter part-ofspeech tagger with Brown clustering. Journal of Intelligent & Fuzzy Systems, 38, 4287–4299.

[11]. The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP with 180,000 Page Images DocEng '20: In Proceedings of the ACM

Symposium on Document Engineering, September 2020. Article No.: 19, pp. 1–4.

[12]. Ta, V.-D., Liu, C.-M., & Nkabinde, G. W. (2016). Big data stream computing in healthcare real-time analytics. In 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 37–42.

[13]. Sui, X., Liu, D., Li, L., Wang, H., & Yang, H. (2019). Virtual machine scheduling strategy based on machine learning algorithms for load balancing. EURASIP Journal on Wireless Communications and Networking, 2019(1), 1-16.

[14]. Tao, D., Lin, Z., & Wang, B. (2017). Load feedback-based resource scheduling and dynamic migration-based data locality for virtual hadoop clusters in openstack-based clouds. Tsinghua Science and Technology, 22(2), 149-159.

[15]. Toosi, A. N., Calheiros, R. N., & Buyya, R. (2014). Interconnected Cloud Computing Environments: Challenges, Taxonomy, and Survey. ACM Computing Surveys, 47(1), 7-47.

[16]. S. Jiao, C. He, Y. Dou, H. Tang, "Molecular dynamics simulation: Implementation and optimization based on Hadoop", 2012 Eighth International Conference on Natural Computation (ICNC), 12031207, 2012.

[17]. K. Shvachko, H. Kuang, S. Radia, "The hadoop distributed file system", Proceedings of the 26th Symposium on Mass Storage Systems and Technologies, 1-10, 2010.

[18]. G. Kecskemeti, M. Gergely, ´ A. Visegr ´ adi, Zs. N ´ emeth, J. Kov ´ acs, P. Kacsuk, ´ "One Click Cloud Orchestrator: Bringing Complex Applications Effortlessly to the Clouds", In: Euro-Par 2014, Lecture Notes in Computer Science (8806), Springer, 38-49, 2014.

[19]. J. Kovacs, P. Kacsuk, Z. Farkas, "Orchestrating Federated Clouds by Occopus", in P. Ivnyi, B.H.V. Topping, G. Vrady, (Editors), "Proceedings of the Fifth International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering", Civil-Comp Press, Stirlingshire, UK, Paper 14, 2017.

[20]. R. Lovas, E. Nagy, J. Kovacs, "Cloud Agnostic Orchestration for Big Data Research Platforms", in P. Ivnyi, B.H.V. Topping, G. Vrady, (Editors), "Proceedings of the Fifth International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering", Civil-Comp Press, Stirlingshire, UK, Paper 15, 2017.

[21]. Vishal Reddy Vadiyala, Parikshith Reddy Baddam, and Swathi Kaluvakuri. "Demystifying Google Cloud: A Comprehensive Review of Cloud Computing Services". In: Asian Journal of Applied Science and Engineering 5.1 (2016), pp. 207–218.

[22]. Fletcher Trueblood, David Rodriguez, Jese Hernandez, Michelle Salomon, Sanjay Soundarajan, and Matin Pirouz. "Demystifying Transportation Using Big Data Analytics". In: 2019 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE. 2019, pp. 1281–1286.

[23]. G Kousalya, P Balakrishnan, C Pethuru Raj, G Kousalya, P Balakrishnan, and C Pethuru Raj. "Demystifying the Traits of Software-Defined Cloud Environments (SDCEs)". In: Automated Workflow Scheduling in Self-Adaptive Clouds: Concepts, Algorithms and Methods (2017), pp. 23– 53.

[24]. Venu Vedam and Jayanti Vemulapati. "Demystifying cloud benchmarking paradigm-an in-depth view". In: 2012 ieee 36th annual computer software and applications conference. IEEE. 2012, pp. 416–421.

[25]. Kai Hwang and Min Chen. Big-data analytics for cloud, IoT and cognitive computing. John Wiley & Sons, 2017.

[26]. Prateeksha Varshney and Yogesh Simmhan. "Demystifying fog computing: Characterizing architectures, applications and abstractions".

In: 2017 IEEE 1st international conference on fog and edge computing (ICFEC). IEEE. 2017, pp. 115–124.

[27]. Susan M Keaveney. "Customer switching behavior in service industries: An exploratory study". In: Journal of marketing 59.2 (1995), pp. 71–82.

[28]. Joseph Vignos, Philip Kim, and Richard L Metzer. "Demystifying the fog: Cloud computing from a risk management perspective". In: Special Issue: Cloud Computing (2013).

[29]. Basappa B Kodada and Demian Antony D'Mello. "Secure Data Deduplication (SD 2 e D up) in Cloud Computing: Threats, Techniques and Challenges". In: International Conference on Advanced Communication and Computational Technology. Springer. 2019, pp. 1239–1251.

[30]. Dutta, P., & Dutta, P. (2019). Comparative study of cloud services offered by Amazon, Microsoft, and Google. International Journal of Trends in scientific Research and Development (IJTSRD), 3(3), 981–985.