A REVIEW ON CLOUD COMPUTING AND BIG DATA USED FOR SCIENTIFIC RESEARCH

ANDREI ANDRAȘ[1], MARIA-VICTORIA CODĂU[2]

Abstract: The digital era, marked by the exponential growth of data and advancements in technology, has been revolutionized by the integration of cloud computing and big data. These technologies have transformed the way data is generated, stored, and analyzed, with applications extending across various fields, including scientific research. Cloud computing provides scalable, flexible, and cost-efficient infrastructure, while big data analytics enables the extraction of valuable insights from vast, diverse datasets. Together, they address the challenges posed by the increasing volume, velocity, and variety of data, facilitating innovation and smarter decision-making. This review explores the fundamentals of cloud computing, its deployment and service models, the defining characteristics of big data, and their synergistic impact on scientific research. It highlights the transformative role of these technologies in enabling data-driven discoveries while acknowledging challenges such as data security and integration complexities. As the adoption of cloud-based big data technologies grows, their potential to drive innovation and enhance scientific collaboration continues to expand.

Keywords: cloud computing, big data, scientific research, IaaS, PaaS, SaaS.

1. INTRODUCTION

The digital era has evolved from 2000 (considered the beginning of the digital age), from the foundational computing technologies to the interconnected world powered by cloud computing and big data of today. These technologies and the new ones to come will continue to shape the future, enabling smarter decision-making, innovative services, and global connectivity. In order to understand the two terms introduced in this review—cloud computing and big data—one has to agree that the amount of digital information has increased significantly and it will continue to do so. Statistically, there is an exponential increase of the amount of data generated worldwide [1], from approximately 45 zettabytes in 2020 to 147 zettabytes in 2024 [2], and an estimation of 181 zettabytes in the year 2025 as seen in figure 1.

[1] *Prof. Ph.D. Eng., University of Petroşani, andrei.andras@gmail.com*
[2] *Managing partner, Wise Management Solutions S.R.L. Bucharest.*

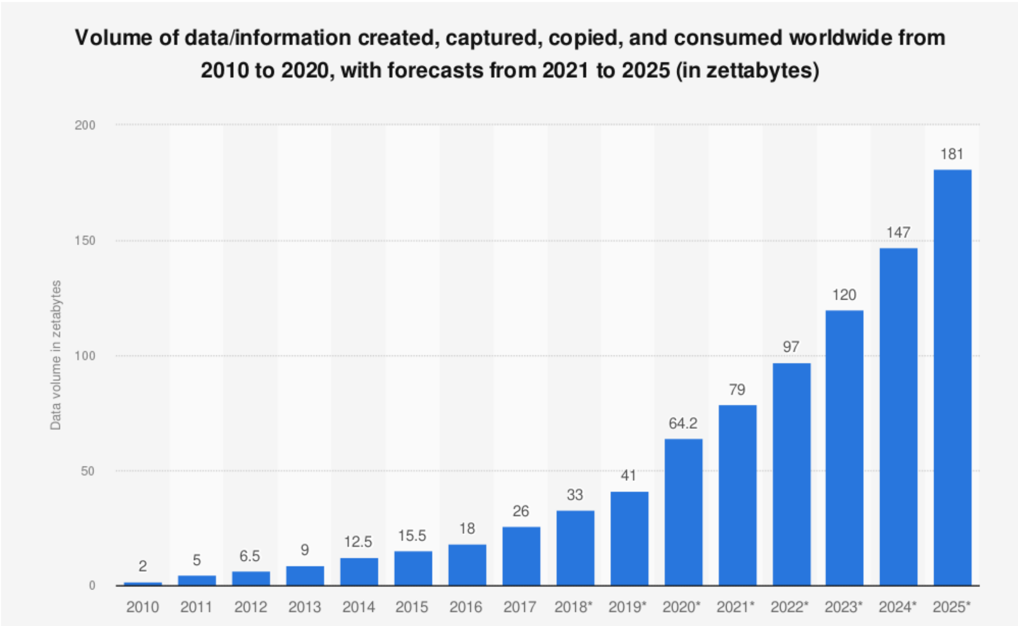

Fig. 1. Statistic on the volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024 [3]

2. DESCRIPTION OF CLOUD COMPUTING

Cloud computing refers to the provision of computing services—such as servers, storage, databases, networking, software, analytics, and intelligence—delivered over the internet ("the cloud"). It enables faster innovation, flexible resource allocation, and cost efficiency. With a pay-as-you-go model, you pay only for the services you use, helping to reduce operating costs, optimize infrastructure management, and scale resources to meet evolving business demands.

Instead of owning and managing physical data centers or servers, organizations and individuals can access technology resources such as servers, storage, databases, networking, software, and more, whenever they need them, from a cloud provider. It proved to be an extremely successful endeavor of service-oriented computing, and it has revolutionized the way computing infrastructure is abstracted and used.

Cloud computing encompasses various models, but in its broadest sense, it refers to the ability to access applications, services, or platforms over the Internet, where users subscribe to a specific set of services within that framework. This approach offers continuous network access, location-independent resources, rapid scalability with quick deployment, and measurable service levels.

This brings up the question: what are cloud deployment models, and who manages the different layers? As shown in Figure 2, cloud management types highlight distinct deployment models and illustrate the varying degrees of customization available in cloud environments.

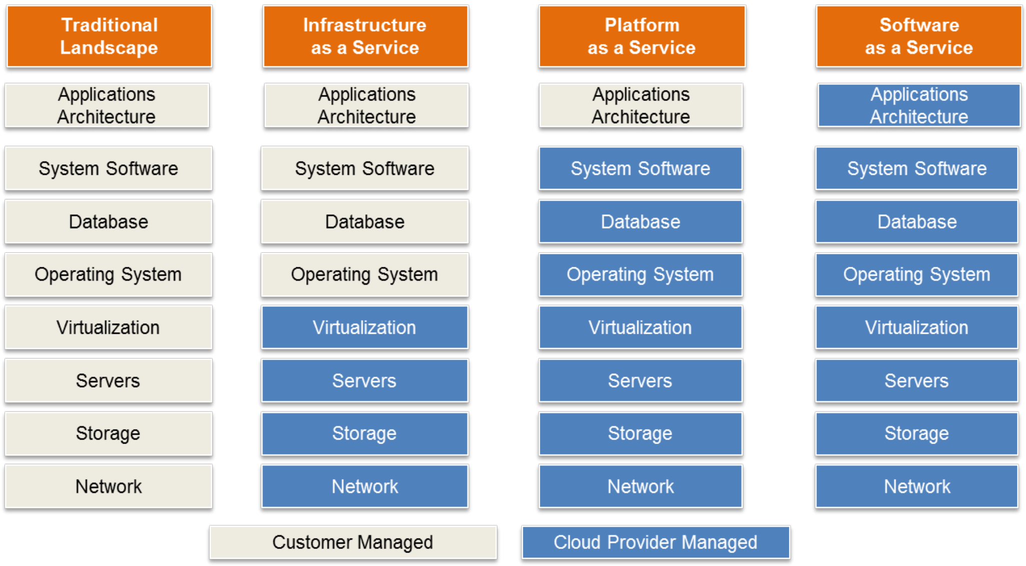

Fig. 2. Cloud Management Types (Inspired by [4])

2.1. Cloud Computing Deployment Models

Cloud computing is classified into various types based on **deployment models** and **service models**, each tailored to meet specific business needs. Deployment models describe how the cloud environment is set up and accessed. There are three main types namely:

A. Public Cloud. The public cloud is hosted by third-party providers and primarily designed for general public use over the internet and where the resources (e.g., servers, storage) are shared among multiple users (multi-tenant model). It is one of the most common deployment models. Providers like Microsoft Azure offer infrastructure and services to individuals or organizations. Users can run workloads on shared infrastructure, utilizing common computing resources with other tenants while maintaining logical data separation. The advantages of this deployment type are that it is cost-effective as no there are no upfront infrastructure costs, it is scalable and elastic, with resources based on demand and it is also accessible from anywhere. The challenges of the public cloud are its limited customization and the potential data privacy which concerns sensitive workloads.

B. Private Cloud. The private cloud is tailored for exclusive use by a single organization or enterprise with multiple users. This model enhances data security and privacy by employing firewalls and internal hosting. It combines the flexibility and scalability of cloud computing with the access control and customization of on-premises infrastructure.

C. Community Cloud. A community cloud serves a group of organizations with shared goals or interests, distributing resources among them. This infrastructure can be managed internally or by a third party, and it may be hosted on-site or externally. By sharing software and hardware resources, this model helps users lower operational costs.

D. Hybrid Cloud. A hybrid cloud integrates an on-premises private cloud with third-party public cloud services, enabling data and applications to be shared between the two environments. This model allows businesses to scale computing power up or down as needed, accommodating peaks in demand without over-investing in on-premises infrastructure. Additionally, it saves time and money by reducing the need to purchase, deploy, and manage additional servers that may only be occasionally required.

Each deployment model contains characteristics suitable for specific solutions which can be summarized as follows:

Private Cloud	Public Cloud	Hybrid Cloud	Community Cloud
The cloud infrastructure is operated exclusively for an organisation. It may be managed by the organization or a third party and may exist on premise or off premise. It offers greater level of privacy with limited scalability.	The cloud infrastructure is made available to the general public or a large industry group and is owned by an organisation selling cloud services. High degree of scalability with potential privacy concern.	A composition of two or more clouds (private, community, or public) that remain unique entities but are bound together to enable data and application portability, with high scalability and potential interconnectivity challenges among different clouds.	The cloud infrastructure is shared by several organisations and supports a specific community that has shared concerns (e.g., mission, security, policy, and compliance considerations). It offers unimaginable shared cost advantages with limited scalability.

Fig. 3. Cloud Computing deployment model characteristics [5]

Reviewing these characteristics, it is evident that each deployment models has its advantages and disadvantages, with the most desirable model being dependent on the organizational requirements and budget.

2.2. Service Models of Cloud Computing

Cloud computing offers a range of service models (see figure 4) designed to cater to different levels of control, flexibility, and convenience for users.

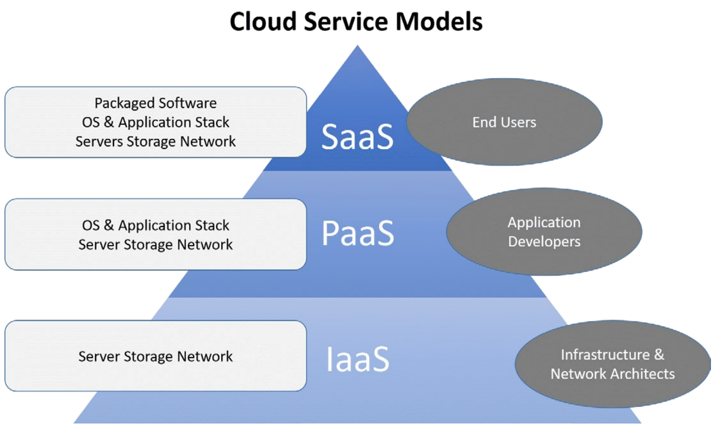

Fig. 4. Cloud Computing service models

These models—Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS)—provide varying degrees of abstraction and management responsibility, making it easier for businesses to choose the right solution for their needs.

Software as a Service (SaaS) is an application delivery model based on a standard, highly scalable software solution provided to multiple customers, with some basic configuration, with limited or no customization. SAP's SaaS deployed solutions like Business by Design, SuccesFactors, Ariba and Concur all share the common benefits of fast and effective deployment with accelerated return on investment, offering vast number of 'out of the box' capabilities with lower up-front costs. The analogy of a hand-tailored fully customized versus a generic made reflects not only the different price and the time, but also the degrees of fit to which SaaS is being deployed.

Platform as a Service (PaaS) is an abstraction of software frameworks used with the underlying operating system that allows application development without the knowledge of underlying layers. In simple terms, the cloud customers are managing their applications architecture layer without any need to manage their operating system admin tasks. Microsoft Azure offers a reliable PaaS model which has Web, Storage, Content Delivery, SQL Azure, access controls and the mass utilization of resources. This offers some phenomenal benefits with redundancy built in through the platform using vendor's architecture scalability. As a result, applications that are scheduled to run in the PaaS model need to be adapted to this environment that may require a certain degree of technical skills required to design feature-rich applications. Important things to consider are the innovation concepts needed to maintain this model building add-ons when developing new solutions. This model contains some natural risks such as cross-platform integration points and ownership of updates. For example, PaaS integration may require further tests to sustain business process related transformations, things like routing, application programming interface (API) calls and ongoing integration with other solutions. In simple terms, Microsoft's approach might not be entirely synchronized with SAP's and vice-versa.

Infrastructure as a Service (IaaS) can be defined as a model for enabling convenient, on-demand network access to a shared pool of configurable resources that can be rapidly provisioned and released with minimal management or service provider interaction. Amazon Web Services (AWS) is a typical example of IaaS, where AWS provide server, network, storage resources and gives the customer the choice of operating system. The virtual environments available through AWS behave exactly like local infrastructure enabling the customers to configure the infrastructure according to their demand.

The service models of cloud computing cater to different levels of technical expertise and business needs. IaaS provides full control over infrastructure, making it suitable for enterprises with specialized IT requirements. PaaS accelerates development by abstracting infrastructure management. SaaS simplifies access to software applications for end-users. Together, these models offer the flexibility and efficiency needed to drive innovation and optimize operations in the cloud.

3. WHAT IS BIG DATA?

This term refers to the vast volumes of data—structured, semi-structured, and unstructured—that are generated at high velocity and in a variety of formats. This data is so large and complex that traditional data processing tools and techniques struggle to store, manage, and analyze it efficiently. The term not only encompasses the data itself but also the processes, technologies, and methodologies used to derive meaningful insights from it. Big Data, unlike traditional data, refers to rapidly expanding datasets that encompass a variety of formats, including structured, unstructured, and semi-structured data. Due to its complexity, Big Data demands robust technologies and advanced algorithms, making traditional static Business Intelligence tools insufficient for handling its applications.

Most data scientists and experts define Big Data through the initial key characteristics, commonly known as the 3Vs [6]: The first and most fundamental characteristic of big data is its **volume**. The sheer amount of data being generated is staggering, with over 3.8 billion people out of 7.6 billion population of the world connected to the internet, and 8.06 billion out of 13.4 billion devices having a mobile connection. [7] This vast volume of data, ranging from social media interactions to machine-generated data, requires new and innovative technologies to collect, store, and process it effectively. The second key aspect of big data is its **variety**. Big data encompasses a wide range of data types, including structured, semi-structured, and unstructured data, such as text, images, video, and audio. This diversity in data formats presents a unique challenge, as traditional data processing methods are often not equipped to handle the heterogeneous nature of big data. The third characteristic of big data is its **velocity**, which refers to the speed at which data is being generated and the need for real-time or near-real-time processing and analysis. The New York Stock Exchange, for example, captures 1TB of trade information during each trading session, while experiments at the Large Hadron Collider at CERN generate 40 terabytes every second [8].

Since Doug Laney first presented the "three Vs of big data" in 2001, these have since been expanded to include additional characteristics that define the true nature of big data [9], with additional characteristics, up to the currently accepted 10 Vs (figure 5). **Veracity**, the fourth V, refers to the reliability and trustworthiness of the data, as inaccurate or unreliable data can lead to flawed insights and decision-making. The fifth V, **Value**, highlights the importance of extracting meaningful and actionable insights from the vast amounts of data, as the true value of big data lies in its ability to drive business decisions and strategic advantage. The sixth V, **Variability**, acknowledges the fluctuating and dynamic nature of big data, which can be influenced by a variety of factors. These include seasonal trends, such as increased online shopping during the holidays, user behavior patterns that may shift over time, and external events like natural disasters or socio-political changes that can impact data collection and analysis. The variability of big data requires organizations to be adaptable and responsive in their data management and analytical strategies. The seventh V, **Visualization**, emphasizes the importance of effective data visualization techniques in making sense of the vast amounts of data. Visualization tools can help transform complex data into intuitive and easily understandable formats, enabling stakeholders to quickly identify patterns, trends, and insights that can inform decision-making. The eighth V**, Verifiability**, underscores the need for data provenance and the ability to verify the origin, authenticity, and integrity of the data. Ensuring the verifiability of big data is crucial, as it allows organizations to establish trust in the data and make informed decisions based on reliable information. This involves maintaining a clear record of data sources, tracking any transformations or processing applied to the data, and implementing robust verification mechanisms to validate the accuracy and completeness of the data. Verifiability also enables organizations to comply with regulatory requirements and industry standards related to data governance and data quality. The ninth V, **Volatility**, acknowledges the ephemeral nature of some big data, where the value or relevance of certain data may be time-sensitive or short-lived. This characteristic emphasizes the importance of understanding the lifecycle and shelf-life of data, as well as the need for effective data retention and archiving strategies to ensure that valuable insights are not lost. Finally, the tenth V, **Virality**, refers to the exponential growth and rapid spread of certain types of data, particularly in the context of social media and user-generated content.

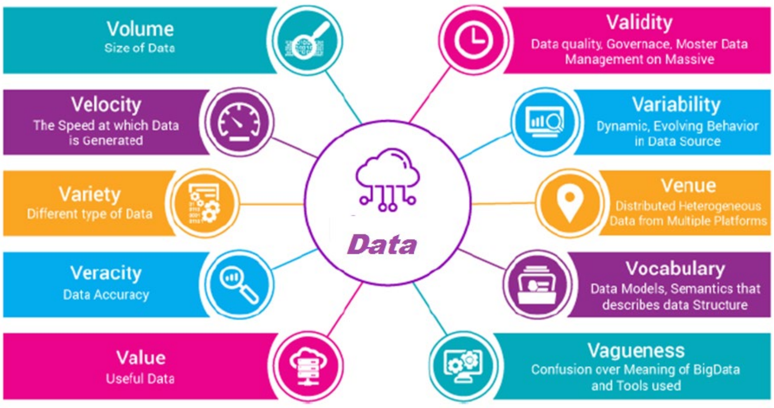

Fig. 5. The characteristics of Big Data also known as the 10 Vs.

4. CLOUD COMPUTING AND BIG DATA IN SCIENTIFIC RESEARCH

Both Big Data and Cloud Computing have numerous applications in all domains of everyday life, such as transportation, health, public utilities, energy, IoT, education, etc. Cloud computing and big data have also become increasingly integral to scientific research, revolutionizing the way researchers gather, store, process, and analyze large and complex datasets. The rapid growth in the volume, velocity, and variety of data generated by advanced scientific instruments and technologies, such as large-scale surveys, high-performance computing, and high-throughput experimentation, has driven the rise of "big data" in scientific research. This shift towards big data has presented significant challenges for traditional computing infrastructure, as the storage and processing requirements of these massive datasets often overwhelm the capabilities of on-premises systems. Cloud computing platforms provide scalable and on-demand access to computing resources, enabling researchers to rapidly deploy and execute data-intensive applications. At the same time, the rise of big data analytics has transformed the way researchers capture, process, and extract insights from vast amounts of data. One of the key aspects of this convergence is the effective management and analysis of large-scale data. Cloud computing platforms provide the necessary scalability and flexibility to handle the ever-increasing volume, velocity, and variety of scientific data, enabling researchers to store, process, and analyze their data more efficiently. The integration of cloud-based big data processing techniques, such as distributed computing and in-memory analytics, has revolutionized the way researchers approach data-driven discoveries [10-13].

Recent studies have highlighted the growing potential of cloud-based big data analytics for a range of scholarly applications, including improved research data management, enhanced collaborator discovery, and more effective expert finding capabilities. These cloud-based big data technologies have enabled researchers to more efficiently store, process, and analyze large-scale scientific data, driving new data-driven discoveries. The application of cloud computing to biomedical research is a prime example of the potential of this technology. Cloud-based platforms have enabled biomedical researchers to access and analyze large, complex datasets, such as genomic sequencing data and medical imaging, at scale, without the need for costly on-premises infrastructure [14-16]. However, the adoption of cloud-based big data technologies in the scientific research domain is not without its challenges. Concerns around data security, privacy, and regulatory compliance, as well as the complexity of integrating heterogeneous data sources, have emerged as significant hurdles. Despite these challenges, the convergence of cloud infrastructure and big data technologies holds immense promise for the future of scientific research. As researchers continue to explore and leverage these advancements, the potential for accelerating scientific discoveries, enhancing collaborations, and driving innovation is expected to grow.

4. CONCLUSIONS

The integration of cloud computing and big data has reshaped the landscape of scientific research, providing unparalleled opportunities for managing and analyzing vast and complex datasets. Cloud computing's scalability and flexibility, combined with big data's capacity to process high-velocity, high-volume information, have enabled researchers to overcome traditional computational and storage limitations. These technologies have paved the way for groundbreaking advancements, particularly in data-intensive fields such as genomics, healthcare, and physics. Despite challenges such as data privacy concerns, integration complexities, and regulatory compliance, the synergy of cloud computing and big data remains a cornerstone for future innovations. By addressing these challenges and leveraging the full potential of these technologies, researchers can accelerate discoveries, enhance collaboration, and drive transformative outcomes across disciplines.

Acknowledgements

This work is supported through the project Increasing the research capacity of the University of Petroşani through investments in Cloud and Big Data infrastructures, contract no. 376/390052/30.09.2021, Id POC/398/1/1/124999, co-funded by ERDF through the Competitiveness Operational Programme 2014-2020.

REFERENCES

[1]. **Berisha, B., Mëziu, E., Shabani, I.** *Big data analytics in Cloud computing: an overview.* J Cloud Comp. 2022, 11 (24). DOI: 10.1186/s13677-022-00301-w.

[2]. **Reinsel, D., Gantz, J., Rydning, J.** *Data age 2025: the evolution of data to-life critical.* International Data Corporation, Framingham, 2022.

[3]. https://www.weforum.org/stories/2024/05/data-growth-drives-ict-energy-innovation/ Accessed 16 december 2024.

[4]. https://www.slideshare.net/slideshow/windows-azure-platform/2253540 Accessed 16 december 2024.

[5]. **Mell P, Grance T.** *National Institute of Standards and Technology (NIST)*, Gaithersburg, MD, September 2011.

[6]. **Furht, B, Villanustre, F, Furht, B, Villanustre, F.** *Introduction to big data.* Big data technologies and applications. 2016, 3-11

[7]. **Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A., Salehian, S.** *The 10 Vs, Issues and Challenges of Big Data*. 2018. DOI: 10.1145/3206157.3206166.

[8]. **Ahmed, K. B., Bouhorma, M., Ahmed, M. B.** *Age of Big Data and Smart Cities: Privacy Trade-Off.* International Journal of Engineering Trends and Technology 2018. 16(6), p. 298). DOI: 10.14445/22315381/ijett-v16p258.

[9]. **Memon, M. A., Soomro, S., Jumani, A. K., Kartio, M. A.** *Big Data Analytics and Its Applications.* Annals of Emerging Technologies in Computing. 2017, 1 (1), p. 45. DOI: 10.33166/aetic.2017.01.006.

[10]. **Khan, S., Shakil, K. A., Alam, M.** *Big Data Computing Using Cloud-Based Technologies, Challenges and Future Perspectives*. Cornell University, 2017. DOI: 10.48550/arxiv.1712.05233.

[11]. **Buyya, R., Ramamohanarao, K., Leckie, C., Calheiros, R. N., Dastjerdi, A. V., Versteeg, S.** *Big Data Analytics-Enhanced Cloud Computing: Challenges, Architectural Elements, and Future Directions.* IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), Melbourne, VIC, Australia, 2015, pp. 75-84, DOI: 10.1109/ICPADS.2015.18.

[12]. **Ji, C., Li, Y., Qiu, W., Awada, U., Li, K.** *Big Data Processing in Cloud Computing Environments.* 12th International Symposium on Pervasive Systems, Algorithms and Networks, San Marcos, TX, USA, 2012, pp. 17-23, DOI: 10.1109/I-SPAN.2012.9.

[13]. **Li, X., Qiu, J.** *Cloud computing for data-intensive applications. Vol. 1.* Berlin: Springer, 2014. DOI: 10.1007/978-1-4939-1905-5.

[14]. **Khan, S., Shakil, K.A., Alam, M.** *Cloud-Based Big Data Management and Analytics for Scholarly Resources: Current Trends, Challenges and Scope for Future Research.* ArXiv, 2016. DOI: 10.48550/arXiv.1606.01808.

[15]. Holko, M., Weber, N., Lunt, C., & Brenner, S. E. *Biomedical research in the Cloud: considerations for researchers and organizations moving to (or adding) cloud computing resources.* Biocomputing 2022, pp. 536-540. DOI: 10.1142/9789811270611_0049.

[16]. Navale, V., Bourne, P. E. *Cloud computing applications for biomedical science: A perspective.* PLoS Computational Biology, 2018, 14(6). DOI: 10.1371/journal.pcbi.1006144.

[17]. Popescu, F.D., Radu, S.M., Andraş, A., Brînaş, I. *Numerical Modeling of Mine Hoist Disc Brake Temperature for Safer Operation.* Sustainability, 13, 2874, 2021. DOI: 10.3390/su13052874.

[18]. Popescu, F. D., Radu, S. M., Andras, A., Kertesz (Brînaş), I. *Simulation of the frequency response of the ERC 1400 Bucket Wheel Excavator boom, during the excavation process*, New Trends in Production Engineering, Vol. 2, Issue 1, pp. 153-167, (2019).

[19]. Kovacs, I., Andraş, I., Nan, M.S., Popescu, F.D. *Theoretical and experimental research regarding the determination of non-homogenons materials mechanical cutting characteristics*. 8th WSEAS International Conference On Simulation, Modelling And Optimization (SMO '08) Santander, Cantabria, Spain, September 23-25, 2008.

[20]. Popescu, F. *Controls ways of the transportation capacity variation for the canvas conveyer.* WSEAS Trans. Syst. Control. *3*, 393–402, 2008.

[21]. Andraş, A., Brînaş, I., Radu, S.M., Popescu, F.D., Popescu, V., Budilică, D.I. *Investigation of the Thermal Behaviour for the Disc-Pad Assembly of a Mine Hoist Brake Using COMSOL Multiphysics.* Acta Tech. Napoc.-Ser. Appl. Math. Mech. Eng. 64, 227–234, 2021.

[22]. Popescu, F.D., Radu, S.M., Andras, A., Brinas, I.K. *A grafo-numeric method of determination of the operation power of the rotor of EsRc-1400 bucket wheel excavator using computer simulation in SolidWorks.* MATEC Web Conf. 290, 04007, 2019.

[23]. Dumitrescu Iosif, Cozma Bogdan-Zeno, Itu Răzvan-Bogdan, *Safety mechanisms for mining extraction vessels*, 16th International Multidisciplinary Scientific GeoConference: SGEM, pages: 759-766, Sofia, 2016.

[24]. Popescu, F. D., Radu, S. M., Kotwica, K., Andraş, A., Kertesz, I. *Simulation of the time response of the ERc 1400-30/7 bucket wheel excavator's boom during the excavation process.* Sustainability, 11(16), 4357, 2019.

[25]. Popescu, F.D. Radu, S.M. V*ertical Hoist Systems: New Trends Optimizations*. LAP Lambert Academic Publishing, 2013.

[26]. Bogdan-Zeno Cozma, Vilhelm Itu, *Trolley for Skip Maneuvring in Lupeni Mining Plant,* 14th International Multidisciplinary Scientific GeoConference: SGEM, pages: 729-736, Sofia, 2014.

[27]. Popescu, F.D. *Calculatorul numeric în industria extractivă*, Editura Universitas, Petroşani, 2004.

[28]. Popescu, F.D. *Instalaţii de Transport pe Vertical*; Editura Focus: Petroșani, Romania, 2010.