Технічні науки

UDC 004

Haina Vladyslav
Site Reliability Engineer
(Jacksonville, Florida, USA)

CLOUD COST OPTIMIZATION STRATEGIES FOR LARGE-SCALE DATA PROCESSING

Summary. *This article presents an in-depth analysis of comprehensive strategies for cost optimization in cloud infrastructures supporting large-scale data processing. The study explores contemporary approaches to resource management, the automation of processes, and the use of hybrid and multi-cloud architectures to achieve both economic efficiency and high system resilience. Special attention is given to the integration of advanced technologies—such as artificial intelligence and machine learning—for predictive analytics and optimized allocation of computing resources, as well as to DataOps methodologies for efficient data lifecycle management. The research methodology is based on a systematic literature review, which enabled the identification of key factors contributing to the reduction of operational expenses and the enhancement of performance in cloud-based solutions. The findings demonstrate that a holistic application of the outlined approaches can significantly lower costs and improve data processing efficiency, offering high practical value for organizations undergoing digital transformation. This work is intended for scholars and practitioners in the fields of cloud computing, strategic cost management, and IT infrastructure optimization, who seek a comprehensive theoretical and practical examination of modern cost-reduction techniques in large-scale data environments. The article presents both methodological insights*

and analytical conclusions that are relevant to researchers and professionals developing effective strategies for digital transformation in the corporate sector.

***Key words**: cloud computing, cost optimization, big data, hybrid cloud solutions, multi-cloud strategies, automation, DataOps, artificial intelligence, machine learning.*

Introduction. The relevance of this topic stems from the growing shift of enterprise information systems toward cloud-based technologies. In today's environment, organizations face the increasing demand for large-scale data processing, which requires not only substantial computing power but also effective cost management of cloud infrastructure. A company's financial stability is directly tied to its ability to optimize cloud spending, especially in light of rapidly evolving business needs and the accelerating pace of technological advancement [1].

The body of literature addressing cost optimization in cloud computing for large-scale data processing demonstrates a multifaceted set of approaches, which can be broadly categorized into several thematic groups.

The first group focuses on comparative analyses of platforms and distributed architectures. For example, Shwe T. and Aritsugi M. [1] examine optimization opportunities within multi-tier computational models that include edge computing. Their study explores the trade-offs between distributed and centralized processing, highlighting the potential advantages of hybrid platforms in achieving scalability and reducing the costs of computational resources.

The second group centers on specific optimization techniques within the serverless computing paradigm. Liu X. et al. [2] propose a universal approach to mitigating cold-start latency in Function-as-a-Service (FaaS) environments, emphasizing application-level time management. This method contributes to cost reduction by improving operational efficiency and demonstrates how performance analysis of particular cloud services can inform the development of

tailored optimization algorithms that directly affect the economic performance of information systems.

A third category encompasses research focused on financial strategies and cost management in cloud infrastructures. Aydoğan M. and Batan A. [3] explore cost optimization methods for analytics tasks in public clouds, aiming to balance resource expenditures with computational performance. Bhardwaj P. [4] stresses the importance of integrating FinOps practices to structure and rationalize financial flows in cloud environments. In the same vein, Thummala V. R. and Singh P. [5] propose cloud migration strategies that address both cost reduction and regulatory compliance—reflecting the increasing relevance of compliance in modern IT ecosystems. Vadisetty R. [6] investigates how core financial factors influence the efficiency of large-scale data processing within cloud frameworks, linking technical decisions to the enterprise's broader financial landscape.

Finally, a set of studies aims to synthesize and analyze cloud resource optimization approaches more comprehensively. Nawrocki P. and Smendowski M. [7] offer an integrative view, bringing together fragmented strategies into a unified analytical framework. Meanwhile, Hassan N. A. B. [8] focuses on managing data dependencies in complex data processing pipelines, identifying challenges in optimizing performance while preserving system integrity within cloud-based solutions.

Despite the evident diversity of approaches—from architectural comparisons and serverless optimization to financial management and pipeline orchestration—the literature reveals certain tensions. On one hand, some studies prioritize technical aspects of optimization, while others emphasize financial and organizational dimensions, creating a gap between innovation and managerial strategy. Although there is significant attention to latency reduction and resource efficiency, issues such as the integration of comprehensive migration strategies with compliance requirements and the management of complex dependencies in distributed data pipelines remain underexplored.

This disconnect highlights the need for further empirical research aimed at harmonizing the technical and financial facets of cost optimization in the context of continuously evolving cloud technologies.

The aim of this paper is to explore current strategies for cost optimization in cloud environments supporting large-scale data processing.

The novelty of the study lies in its development of a model that integrates hybrid and multi-cloud strategies with advanced AI/ML technologies for predictive resource allocation and optimization. This approach not only reduces operational costs but also improves the overall performance of data processing systems—offering a critical competitive edge for enterprises undergoing digital transformation. Additionally, the paper presents original recommendations for implementing cost optimization strategies in large-scale cloud infrastructures, based on direct practical experience. These include the design and successful migration of complex cloud environments and the optimization of a multi-cloud platform's performance.

The central hypothesis proposes that the integrated application of automation tools, hybrid cloud architectures, and AI/ML-based predictive analytics within a unified management system can reduce operating expenses while ensuring high flexibility and adaptability of infrastructure for large-scale data processing.

The research methodology is based on a systematic review of contemporary literature in this domain.

1. Resource Management and Process Automation

Efficient resource management and process automation are foundational to cost optimization in cloud infrastructures designed for large-scale data processing. A central challenge in this context is achieving "right-sizing"—the precise allocation of computing resources to meet actual demand. Overprovisioning leads to unnecessary costs, whereas optimizing virtual machine configurations can reduce expenses [1]. Monitoring tools such as AWS Cost

Explorer, GCP Billing Reports and Azure Cost Management provide granular insights into resource consumption, enabling the timely identification of inefficiencies.

Dynamic auto-scaling mechanisms further enhance resource efficiency by automatically adjusting the number of instances based on real-time workload fluctuations. This capability is particularly critical when processing volumes are unpredictable. Long-term pricing models like Reserved Instances and Savings Plans have proven effective for predictable workloads, offering considerable cost reductions [2].

A summary of key resource management techniques is presented below:

Table 1

Comparative analysis of resource management methods

Method	Description	Benefits
Right-Sizing	Optimizing virtual machine configurations to match actual workload needs	Reduces overprovisioning, improves efficiency
Auto-Scaling	Automatically adjusts the number of instances based on workload changes	Fast adaptation, high scalability
Reserved Instances	Long-term contracts at fixed pricing for predictable usage	Cost stability, predictable spending
Savings Plans	Flexible pricing with fixed rates for resource usage	Flexibility in payment models

Source: adapted from [1]

The adoption of DevOps practices—particularly those built on Infrastructure as Code (IaC) and Continuous Integration/Continuous Deployment (CI/CD)—minimizes manual intervention, shortens deployment cycles, and improves responsiveness to shifting business demands [5]. Moreover, the integration of infrastructure monitoring and management solutions, along with third-party tools, ensures resource usage transparency and facilitates real-time

configuration adjustments. Automation also extends to predictive analytics systems powered by machine learning, which enable proactive identification of peak loads and dynamic adaptation of resource allocation.

In sum, combining right-sizing practices with process automation is essential for achieving optimal efficiency in cloud operations. A systems-oriented strategy that integrates hybrid and multi-cloud architectures with automated DevOps workflows not only drives down costs but also enhances flexibility and resilience across cloud ecosystems.

2. Hybrid and Multi-Cloud Strategies

The implementation of hybrid and multi-cloud strategies has emerged as a critical direction in cost optimization for cloud infrastructures supporting large-scale data processing. By combining on-premises systems with public cloud resources, organizations can maintain control over sensitive data while flexibly offloading variable workloads to the cloud [4, 6]. This approach reduces the total cost of ownership (TCO) by optimizing resource allocation and enhancing infrastructure resilience.

Multi-cloud strategies—those that distribute workloads across multiple cloud providers—further reduce reliance on any single vendor. This diversification not only offers access to more competitive pricing models but also enables organizations to select the most suitable services for specific tasks [7]. When combined with hybrid architectures, multi-cloud approaches provide additional fault tolerance and allow dynamic redistribution of workloads during demand spikes—essential for ensuring system stability in high-volume data environments.

The integration of hybrid and multi-cloud models offers the following key advantages:

- Flexibility and Adaptability: Hybrid architectures enable organizations to retain mission-critical systems on-premises while leveraging cloud resources

to handle fluctuating workloads. This facilitates rapid responses to changing business requirements.

- Risk Management: Storing sensitive data locally while processing less critical workloads in the cloud helps mitigate security risks and reduces dependence on a single provider.
- Cost Efficiency: By taking advantage of competitive pricing among cloud vendors, multi-cloud strategies enable significant cost savings, as supported by empirical evidence [2].

The table below summarizes the comparative features of hybrid and multi-cloud strategies:

Table 2

Comparative analysis of hybrid and multi-cloud strategies

Parameter	Hybrid Strategy	Multi-Cloud Strategy
Infrastructure Integration	Combines on-premises and cloud resources for critical and variable workloads	Distributes workloads across multiple cloud providers
Flexibility	Retains control over core systems while adapting to shifting business needs	Offers high flexibility and optimal cost-to-performance ratios
Risk Management	Enhances security by keeping sensitive data on-premises	Reduces vendor lock-in and mitigates the risk of inflated pricing
Cost Savings	Decreases TCO through workload distribution	Allows you to save resources due to supplier competition
Scalability & Resilience	Ensures fault tolerance for critical systems via local infrastructure	Increases resilience by distributing loads across multiple platforms

Source: adapted from [2; 6; 7]

In summary, hybrid and multi-cloud strategies represent a holistic approach that enables enterprises to tailor infrastructure to evolving business demands, optimize spending, and bolster system resilience.

3. Specifics of Integrating Modern Technologies for Cost Forecasting and Optimization

One of the key directions in this area is the use of AI/ML technologies for predictive analytics in cloud systems. Machine learning algorithms analyze historical data on resource consumption, identify patterns, and forecast future workload peaks, allowing infrastructure configurations to be adjusted in advance[2]. Moreover, AI/ML algorithms enable the detection of anomalies in resource usage, ensuring timely interventions and realignment of computing capacity.

An essential element of the modern approach is DataOps—a methodology focused on automating the data lifecycle. DataOps encompasses data flow management, archiving, and tiered storage strategies, which help optimize costs related to data storage and processing [8]. Moreover, integrating monitoring systems with artificial intelligence algorithms enables the creation of a unified data stream for analyzing resource usage. These integrated solutions ensure continuous monitoring, automatic detection of inefficiencies, and offer recommendations for resource redistribution, which further reduces costs and enhances overall infrastructure resilience [1; 2].

Below is Table 3, which summarizes the technologies used for forecasting and optimizing costs.

Table 3

Comparative analysis of technologies for forecasting and optimizing costs [1; 2; 8]

Technology	Description	Advantages	Example of Use
AI/ML for Forecasting	Application of machine learning algorithms to analyze historical data and predict peak loads	Improved forecasting accuracy	Forecasting cloud resource utilization

Technology	Description	Advantages	Example of Use
DataOps	Automation of the data lifecycle, including archiving, tiering, and data flow management	Lower storage costs, enhanced data management	Data lifecycle optimization in Cloud Infrastructure
AI-Integrated Monitoring Systems	Integration of monitoring tools with AI to detect resource usage anomalies	Automatic inefficiency detection, real-time cost correction	Continuous monitoring and real-time optimization of resource allocation

Thus, the integration of advanced technologies such as AI/ML and DataOps with modern monitoring systems forms the foundation for a proactive cost management model in cloud infrastructures. This approach not only enables significant savings through optimized resource distribution but also increases system flexibility and resilience, which is essential for maintaining competitiveness in large-scale data processing.

The following section presents authorial recommendations for implementing cloud cost optimization strategies for large-scale data processing, derived from direct experience. It is necessary to use adaptive demand forecasting algorithms which, through machine learning and analytical models, allow for accurate estimation of future compute and network requirements. This proactive approach helps eliminate overprovisioning and optimize the use of cloud services, which is particularly critical in environments with rapidly fluctuating workloads.

The next step is the integration of automated scalability mechanisms that allow resources to be dynamically adjusted to match the evolving data processing infrastructure. Automation, implemented through auto-scaling policies, enables not only rapid response to workload peaks but also minimizes baseline expenses during periods of low activity. This, in turn, requires the deployment of real-time monitoring and cost analytics systems that utilize advanced visualization and analytical tools to support timely and well-informed decisions.

Equally important is the adoption of a multi-layered data management strategy, where distributed architecture and a microservices approach optimize the processes of data processing, storage, and transfer. This involves the use of specialized cloud services tailored to specific types of compute tasks, which helps significantly reduce operational expenses and increase overall system performance. At the same time, the comprehensive integration of services across hybrid and multi-cloud infrastructures provides the flexibility to select optimal operational models based on project specifics and business needs.

This systems-based approach, grounded in experience and research, enables not only the optimization of current costs but also the development of a long-term strategy for sustainable growth and technological modernization in the context of a rapidly evolving digital landscape.

Conclusion. This article presents a comprehensive approach to cost optimization in cloud systems for large-scale data processing, based on the integration of resource management techniques, process automation, hybrid and multi-cloud strategies, and advanced forecasting technologies powered by AI/ML. The study confirms that efficient allocation of computing resources through right-sizing, automation via IaC and CI/CD, and the adoption of hybrid and multi-cloud architectures can lead to substantial reductions in operational costs. In addition, the integration of predictive analytics and DataOps methodologies enables proactive infrastructure management, enhancing resilience and adaptability under dynamic business demands.

The findings hold practical value for organizations seeking to improve the efficiency of cloud resource utilization in the context of large-scale data processing and digital transformation. Future research may focus on developing unified frameworks tailored to the specifics of different industries, as well as expanding experimental models to incorporate emerging technological trends such as edge computing and quantum computing.

References

1. Shwe T., Aritsugi M. Optimizing data processing: a comparative study of big data platforms in edge, fog, and cloud layers. *Applied Sciences.* 2024. Vol. 14 (1). 452 p.

2. Liu X. et al. Faaslight: General application-level cold-start latency optimization for function-as-a-service in serverless computing. *ACM Transactions on Software Engineering and Methodology.* 2023. Vol. 32 (5). P. 1-29.

3. Aydoğan M., Batan A. Cost Optimization Strategies for Big Data Analytics in Public Cloud Infrastructures. *Applied Science, Engineering, and Technology Review: Innovations, Applications, and Directions.* 2024. Vol. 14 (10). P. 1-13.

4. Bhardwaj P. The Role of FinOps in Large-Scale Cloud Cost Optimization. 2024. Vol. 8 (1). P. 1-10.

5. Thummala V. R., Singh P. Developing Cloud Migration Strategies for Cost-Efficiency and Compliance. *International Journal of Multidisciplinary Innovation and Research Methodology.* 2024. P. 2960-2068.

6. Vadisetty R. Efficient large-scale data based on cloud framework using critical influences on financial landscape. *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC).* IEEE, 2024. P. 1-6.

7. Nawrocki P., Smendowski M. A Survey of Cloud Resource Consumption Optimization Methods. *Journal of Grid Computing*. 2025. Vol. 23 (1). 5 p.

8. Hassan N. A. B. Managing Data Dependencies in Cloud-Based Big Data Pipelines: Challenges, Solutions, and Performance Optimization Strategies. *Orient Journal of Emerging Paradigms in Artificial Intelligence and Autonomous Systems*. 2025. Vol. 15 (2). P. 20-28.