librasophia : Pages:1–13

Original Research

Cost Optimization Strategies for Big Data Analytics in Public Cloud Infrastructures

Mert Aydoğan[1] and Ali Batan[2]

[1]Kırşehir Institute of Technology, Department of Computer Science, Bağbaşı Mah. 112. Sokak No:18, Kırşehir, Turkey.
[2]Bolu Technical University, Department of Computer Engineering, D100 Karayolu, Bolu, Turkey.

Abstract
This paper presents a comprehensive investigation into cost optimization strategies for big data analytics deployed within public cloud infrastructures. The analysis focuses on how advanced resource allocation, elastic provisioning, and adaptable workload scheduling can collectively address the challenges of high operational expenses when handling massive datasets. By examining both the theoretical underpinnings and the practical mechanisms of allocating computing, storage, and networking resources in dynamic cloud environments, this paper contributes a methodological roadmap for reducing costs. A major emphasis is placed on nuanced resource management, where analytical workloads are partitioned, scheduled, and executed in ways that minimize overhead and idle times. Mathematical models are developed to illuminate the multi-dimensional nature of cost components, including on-demand pricing, data transfer fees, and potential penalties associated with performance degradation. Each proposed solution is evaluated under simulated operational loads that capture realistic traffic patterns and peak demands. The findings indicate that a combination of model-driven and heuristic optimization techniques can produce substantial cost savings without sacrificing the performance requirements mandated by large-scale analytics tasks. Limitations of the proposed models in heterogeneous environments and potential trade-offs between immediate cost savings and long-term sustainability are also considered. In conclusion, this paper underscores the importance of flexible strategies that balance cost with computational performance in modern big data workflows.

1. Introduction

The prevalence of big data analytics in numerous industries has rapidly increased the demand for scalable computational power and massive storage facilities [1]. Public cloud platforms have emerged as an attractive solution to address this insatiable demand, largely due to their elasticity, flexibility, and on-demand resource provisioning. As organizations migrate analytic workflows to cloud environments, they encounter intricate challenges in resource management that can significantly impact the overall operational expenditure [2]. The tension lies in providing robust computing and networking capabilities to handle surges of analytical tasks while minimizing idle time and resource wastage when workloads diminish. This challenge becomes especially acute in large-scale data contexts, wherein computational tasks are both varied and time-sensitive. Many organizations lack the internal expertise or the sophisticated toolchains required to optimize cost, resulting in suboptimal provisioning strategies and inflated expenses. [3]

To comprehend this issue, it is crucial to explore the inherent complexities associated with big data analytics, including the evolving nature of queries, the probabilistic patterns of data arrival and processing times, and the constraints imposed by service-level agreements. Data-intensive workflows often require orchestrating numerous clusters across diverse geographical regions. Clouds operate on different pricing models, typically based on pay-as-you-go schemes for computing, storage, and data transfer [4]. Organizations often pay additional costs for special functionalities such as real-time analytics and fault tolerance. These disparate cost components are not always transparent and can lead to unexpected spikes in monthly expenditures [5]. The multi-tenant nature of public cloud environments introduces additional variables, as the resources are shared among different tenants with heterogeneous workloads, creating contention and potential performance unpredictabilities.

Cost optimization in cloud-based analytics can be approached from multiple angles. One immediate method is to adjust provisioning strategies to match resource availability with forecasted demand [6]. A second approach is to apply algorithmic and heuristic strategies in scheduling and workload distribution, with the objective of minimizing underutilized resources and reaping savings from reserved instances or spot pricing when applicable. Some organizations adopt tailored data lifecycle management policies, moving older, less accessed data to cheaper storage tiers, or even offloading data to cold storage. However, the dynamic and sometimes spiky nature of big data analytics demands more holistic solutions that extend beyond simplistic scaling rules. [7, 8]

Another dimension of complexity is the interplay between computational cost and performance deadlines. Many big data workloads include latency-sensitive tasks, such as near real-time fraud detection or trend analysis for live dashboards [9]. Sacrificing performance for cost may not be acceptable in these settings, particularly if compliance or business continuity requirements dictate stringent performance targets. Furthermore, the architecture of the analytics platform—whether it employs MapReduce, Spark, or other distributed frameworks—can significantly influence the feasibility of certain cost-reduction strategies. For instance, the overhead in transferring large volumes of data across regions may negate the financial benefits gained from cheaper compute instances in those regions. [10]

Advanced modeling of cost components is pivotal. Multiple facets like compute node usage rates, networking charges, block storage costs, and data retrieval fees must be integrated into a single objective function. This mathematical representation enables the derivation of theoretically optimal or near-optimal solutions for scheduling, resource assignment, and data placement [11]. Incorporating realistic constraints such as service-level agreements, job deadlines, and user-specified performance benchmarks ensures that these solutions are directly applicable to operational environments. Nonetheless, achieving true cost efficiency requires balancing diverse objectives—one may minimize the sum of compute-hour usage while inadvertently increasing data transfer costs [12]. The solution space is often governed by trade-offs and must be navigated with advanced optimization frameworks that can handle large-scale data volumes.

This paper investigates the manifold approaches to reducing costs for big data analytics in public cloud settings. The subsequent sections explore foundational cost optimization concepts, develop formal mathematical models for resource allocation and workload scheduling, and then propose heuristic and algorithmic techniques that account for the inherent uncertainties in such systems [13]. The results section reports the performance of these strategies under varying conditions, demonstrating both significant cost reductions and areas of potential improvement. Finally, the discussion highlights limitations, underscores directions for future research, and emphasizes how an integrated approach can yield sustainable cost savings while retaining the computational guarantees demanded by modern big data workflows.

2. Foundations of Cost Optimization in Big Data Analytics

The optimization of costs in big data analytics demands a fundamental understanding of how various cloud resources contribute to the overall expenditure [14]. In general, these resources can be classified into compute, storage, and network bandwidth. The interdependence of these resources complicates any attempts at reduction that do not consider the entire lifecycle of analytics tasks [15]. For instance, a decision to deploy more powerful compute nodes might reduce job runtime but could lead to higher storage costs due to uncompressed intermediates or more expensive instance types. Similarly, attempts to minimize data storage fees by using cheaper, lower-performance tiers may result in longer compute times, thereby increasing the overall runtime charges [16].

A critical concept is elasticity, which allows quick scaling up or down of resources in response to shifting workloads [17]. While elasticity can be a powerful tool for cost containment, its effectiveness depends on accurate workload prediction. Static provisioning can lead to underused resources in periods of low demand, and over-reliance on automatic scaling rules may result in abrupt cost increases if rules are triggered too frequently by transient workload spikes. An in-depth understanding of usage patterns helps balance these concerns. [18, 19]

Data velocity, variety, and volume define the fundamental features of big data analytics. The velocity refers to the rapid generation rate of new data, the variety acknowledges the heterogeneity in data formats, and the volume captures the massive scale of data involved [20]. In mathematical terms, let $\lambda(t)$ represent the rate at which new data arrives at time t. The analytics tasks often involve functions such as transformations, aggregations, and machine learning model training, which we can denote by a job set $\{J_1, J_2, \ldots, J_n\}$. Each job J_i has a resource demand vector $\mathbf{r}_i(t)$ that indicates compute, memory, storage, and networking requirements over time.

Define a cost function $C(\mathbf{x}, \mathbf{y}, \mathbf{z})$ that expresses the monetary impact of resource usage. Here, $\mathbf{x}$ describes the allocation of compute resources across tasks, $\mathbf{y}$ accounts for data storage configuration, and $\mathbf{z}$ manages the networking paths and bandwidth allocations. The general objective is to minimize:

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \quad C(\mathbf{x}, \mathbf{y}, \mathbf{z}) \quad \text{subject to various system constraints.}$$

The constraints may include user-defined performance thresholds such as job completion times or throughput requirements [21]. Let d_i be the maximum allowable delay for job J_i. The constraints thus specify:

$$T_i(\mathbf{x}) \leq d_i \quad \text{for all } i,$$

where $T_i(\mathbf{x})$ denotes the completion time of job J_i given the resource allocation strategy. Additional constraints can capture dependencies between jobs, as some analytics tasks must process the output of preceding tasks. [22]

The pricing structure in public cloud environments often involves several components, such as per-hour usage of compute instances, per-gigabyte storage, and per-gigabyte data transfer rates. The complexities introduced by spot instances, reserved instances, and dedicated hosts increase the dimensionality of the optimization problem [23]. In certain frameworks, partial hours are billed as full hours, creating a discrete optimization scenario where selecting instance lifetimes must be handled carefully. A job might be scheduled to run on a highly efficient but expensive set of instances if the job is short and the performance gains reduce the total compute hours. Conversely, a longer job may derive more benefit from slower, cheaper instances to achieve overall lower costs. [24]

Another foundation lies in the statistical characteristics of workloads. The arrival of analytics tasks can often be modeled as a non-homogeneous Poisson process, where the arrival rate function $\lambda(t)$ reflects typical business cycles or bursty patterns triggered by external factors. Characterizing these patterns can improve the accuracy of short- and medium-term forecasts [25]. Advanced filtering techniques, including Kalman filters or exponential smoothing, could refine capacity planning and dynamic provisioning. By minimizing overprovisioning, cost savings can be realized in compute-intensive tasks. [26]

In many cases, the number of simultaneous tasks can be large, leading to concurrency issues and potential data bottlenecks. The concurrency level might also be intentionally restricted to limit cost. As an example, if the concurrency limit is set at k, then only k tasks may run simultaneously, while others wait in a queue [27]. Formally, let $q(t)$ be the number of tasks queued at time t. The concurrency-limited scheduling policy must satisfy:

$$q(t) \geq 0, \qquad \sum_{\text{running tasks}} r_{\text{CPU}}(t) \leq k \cdot R_{\text{CPU}}^{\text{unit}},$$

where $R_{\text{CPU}}^{\text{unit}}$ represents the CPU capacity of a single resource unit. Queueing delay may adversely affect performance measures, and hence the total cost function $C(\mathbf{x}, \mathbf{y}, \mathbf{z})$ can incorporate penalties for excessive queueing times.

Although these foundational concepts guide cost optimization strategies, actual results in production environments vary due to unpredictable network latencies, concurrency from co-located tenants, and fluctuations in spot instance availability [28]. For truly effective cost reduction, these uncertainties must be embedded in the problem formulation and solution techniques. Methods such as robust optimization, scenario-based stochastic programming, or distributionally robust approaches can account for these uncertainties [29, 30]. Such techniques, however, often produce more conservative solutions, implying that cost gains may be tempered by risk-averse provisioning choices.

3. Mathematical Modeling of Cloud Resource Allocation

To address cost optimization in a rigorous manner, the problem can be formulated as a mathematical model that captures the salient features of cloud-based big data analytics. Let there be n analytics tasks to schedule, indexed by $i = 1, 2, \ldots, n$ [31]. Each task i possesses a set of characteristics:

1. A computational requirement function $f_i(t)$, representing the CPU time or compute resources needed over time. 2. A storage demand s_i, indicating the persistent data requirements for the duration of the task. [32] 3. A deadline d_i or a service-level objective that bounds completion time. 4. A cost function component that includes any specialized charges, such as GPU usage or high-bandwidth networking. [33]

Define a binary decision variable $x_{i,m}$ that indicates whether task i is allocated to a particular instance type m. Let the set of available instance types be $\{m \in M\}$, each type m having a cost rate α_m for compute time and a capacity constraint. The capacity constraint could be represented by the maximum number of tasks or the maximum usage of CPU cycles. If tasks are migratable, we introduce additional decision variables to model the penalty for relocating tasks between instances, which may incur data transfer fees or overheads.

The total cost incurred is given by: [34]

$$\sum_{m \in M} (\alpha_m \cdot H_m) + \sum_{i=1}^{n} \beta \cdot g_i,$$

where H_m is the total number of instance hours for type m, β is the cost rate for data transfer, and g_i measures the gigabytes moved for task i. The scheduling constraints must ensure that task i meets its deadline when allocated to instance type m. Introducing the concept of start time u_i and finish time v_i: [35]

$$u_i \geq 0, \quad v_i = u_i + \delta_i(x_{i,m}), \quad v_i \leq d_i,$$

where $\delta_i(x_{i,m})$ is a function that determines the task duration based on the instance type allocated. If tasks are allowed to overlap, capacity constraints ensure that the sum of computational demands on a single instance type does not exceed its allocated resources within any given time interval.

To capture concurrency limitations, define a capacity parameter κ_m for instance type m [36]. Then, for each time epoch or discrete time slot t, we have:

$$\sum_{i: u_i \leq t \leq v_i} x_{i,m} \cdot r_i(t) \leq \kappa_m,$$

where $r_i(t)$ is the resource usage profile of task i. In continuous time models, one typically integrates or sums resource consumption over intervals.

The storage cost function can be similarly modeled. Let γ denote the storage cost per gigabyte-hour, and let s_i be the storage requirement of task i from time u_i to v_i. The total storage cost is: [37]

$$\sum_{i=1}^{n} \int_{u_i}^{v_i} \gamma \cdot s_i \, dt.$$

If ephemeral storage is used, the cost may be tied to the instance usage, whereas persistent storage can be modeled as a separate resource, potentially incurring additional transfer charges if tasks migrate between regions. For tasks needing massive data input, network cost $\beta \cdot g_i$ might dominate [38]. The trade-off between storing data near the compute cluster versus transferring it from a remote data lake must also be accounted for in the cost formulation.

The model can be further extended to handle uncertain workloads by introducing stochastic variables. One might assume that each task i has an uncertain arrival time ω_i and uncertain size distribution [39]. A robust optimization approach replaces constraints such as $v_i \leq d_i$ with:

$$v_i(\omega) \leq d_i, \quad \forall \omega \in \Omega,$$

where Ω is the uncertainty set describing possible variations in task arrivals and resource needs [40, 41]. Though robust solutions tend to be more expensive, they mitigate the risk of failing to meet deadlines under worst-case scenarios. Alternatively, chance-constrained approaches enforce the probability that tasks miss deadlines to be below a chosen threshold [42]. In such a scenario, the model becomes:

$$\Pr(v_i > d_i) \leq \epsilon,$$

for each critical task i [43]. Balancing these constraints leads to interesting mathematical structures, such as mixed-integer convex programs with nonlinear constraints, often addressed via decomposition methods or heuristic approximations.

In certain big data workflows, tasks may exhibit precedence relationships. Let $J_i \longrightarrow J_k$ signify that J_k can only start when J_i completes [44]. This relationship gives rise to scheduling constraints of the form:

$$u_k \geq v_i, \text{[45]}$$

if J_k is dependent on J_i. Graph structures representing directed acyclic graphs (DAGs) typically model these dependencies. The complexity of solving DAG-based scheduling problems in large-scale environments is substantial, which has led to the proliferation of heuristics tailored to topological ordering or path-based resource planning. [46]

This rigorous mathematical framework, though abstract, forms the bedrock upon which practical solutions for big data analytics are constructed. Real-world systems implement approximations of these models, leveraging either exact solvers for smaller problem instances or heuristic approaches for large workloads. The following section delves into advanced heuristic and algorithmic methods that complement these models, aiming to find near-optimal solutions within computationally tractable time frames.

4. Advanced Heuristic and Algorithmic Approaches

Although rigorous optimization formulations capture the essence of cost minimization in big data analytics, they can quickly become intractable as problem size grows. The combinatorial explosion of potential schedules, instance selections, and data placement decisions often renders classical solvers impractical for large-scale production workflows [47]. As a result, sophisticated heuristic and meta-heuristic approaches, as well as specialized algorithms, have gained prominence as a means of producing high-quality solutions within acceptable computational overhead.

One class of heuristics focuses on dynamic scheduling, where resource allocation decisions are updated in real time or near real time based on the current state of the system. As tasks arrive or complete, heuristics recalculate resource assignments and re-evaluate ongoing usage in order to release unnecessary capacity [48, 49]. These approaches might use reactive policies, in which certain threshold triggers cause expansions or contractions of allocated instances. In more advanced versions, the policy itself is tuned using methods such as reinforcement learning, so that the system can discover better allocation patterns through repeated interactions with real-world workloads.

Another popular class of methods relies on evolutionary algorithms such as genetic algorithms or particle swarm optimization [50]. In these approaches, a population of candidate solutions (i.e., schedules or provisioning schemes) evolves through iterative transformations, such as crossover and mutation in the case of genetic algorithms. The cost function derived from the underlying optimization model serves as the fitness measure that guides the selection of candidates for subsequent generations [51]. Over time, these methods have been adapted to account for domain-specific constraints in big data analytics, such as multi-tenant concurrency or region-based storage restrictions. Hybrid variants that combine evolutionary search with local improvement heuristics or integer programming subroutines have proven to be effective at handling large-scale problems.

Swarm intelligence approaches like ant colony optimization or bee colony optimization also make appearances in big data scheduling, often focusing on partitioning tasks across geographically distributed data centers to minimize both compute and transfer costs [52]. These methods harness collective behavior analogies, where virtual “ants” or “bees” iteratively refine task allocations by following or avoiding pheromone trails that signify beneficial resource assignments from prior iterations.

Another line of research applies approximation algorithms with provable performance bounds. These algorithms do not guarantee an exact optimum but are accompanied by a theoretical ratio indicating how close their solution is to the optimal one [53]. In a common scenario, if the algorithm yields a solution with cost C_{alg} and the optimal cost is C_{opt}, it holds that $C_{\text{alg}} \leq \alpha \cdot C_{\text{opt}}$ for some constant α. For certain scheduling problems in the cloud domain, polynomial-time approximation schemes may exist, especially under restrictions such as uniform machine capacities or simple cost structures. However, the intricacies of real-world public clouds—diverse instance types, discontinuous pricing tiers, network egress charges, and ephemeral or spot market fluctuations—complicate the direct application of approximation algorithms [54, 55]. Even so, insights from these simplified models can guide the design of more robust heuristics.

Machine learning-based methods have also been studied for cost optimization. In predictive autoscaling, a forecasting model anticipates future workload levels, enabling the system to proactively spin up or down computing resources [56]. When combined with capacity planning that selects the most cost-effective instance types, this approach helps prevent both underprovisioning (which risks missing deadlines) and overprovisioning (which inflates costs). A recurrent neural network might be used for time-series prediction of incoming workloads, while a deep reinforcement learning agent determines which instance types to allocate. An elegant solution can thus arise that continuously refines its policies by evaluating reward signals derived from cost metrics and performance satisfaction. [57]

Queueing theory also offers insights. In multi-class queueing networks, tasks of different priorities or service requirements can be mapped to distinct classes [58]. The system’s dynamics are characterized by arrival rates, service times, and capacity constraints on each server, where a server might represent a type of cloud instance. Classical results in queueing theory, such as heavy-traffic approximations or stability boundaries, can guide how many instances should be provisioned at minimum to handle peak load. Yet big data analytics extends beyond standard queueing models because tasks often require parallel processing [59]. Instead, queueing networks must be adapted to represent parallelizable jobs or DAG structures. Methods for approximate mean response times or queue lengths can then be integrated into cost optimization heuristics to yield more informed scheduling decisions.

At the infrastructure management layer, container orchestration platforms such as Kubernetes, Mesos, or Yarn can be exploited to manage computing tasks [60]. These systems offer built-in autoscaling features, but they usually rely on heuristic-based resource management policies that do not explicitly optimize for cost. An emerging approach is to overlay a cost-aware scheduling layer atop these platforms, intercepting resource requests and deciding how to map them onto the available cloud resources [61]. The scheduling layer might apply integer linear programming or advanced heuristics to batch tasks together efficiently, leveraging ephemeral container lifetimes to minimize idle time. The synergy between orchestrators and cost-aware schedulers enables organizations to adapt quickly to changing workloads and dynamically explore different resource configurations.

All of these heuristic and algorithmic methods must contend with the complexity of logging and monitoring in big data systems [62]. Detailed instrumentation is vital to capture operational metrics such as runtime durations, concurrency levels, data movement, and overall usage. These metrics inform the cost function and guide algorithmic decisions. Tools that provide real-time insights, combined with advanced anomaly detection methods, can react to unanticipated workload surges or performance degradations [63, 64]. Without robust monitoring, even the most sophisticated algorithms may fail to yield consistent and reproducible cost savings in actual environments.

Ultimately, a multi-tiered approach is often advisable, where a baseline heuristic handles real-time scheduling decisions while more powerful offline or batch-based optimization methods periodically reconfigure the system [65]. By partitioning the problem in this fashion—short-term dynamic decisions handled separately from long-term strategic planning—one can strike a balance between responsiveness and exhaustive search capabilities. The next section explores empirical performance evaluations of these techniques, demonstrating tangible cost reductions, as well as highlighting areas where these approaches may falter under extreme workload conditions or rapidly shifting usage patterns.

5. Performance Evaluation and Discussion

To evaluate the effectiveness of the discussed cost optimization strategies, a series of experiments were designed to replicate realistic big data analytics workflows in public cloud environments [66]. Synthetic workloads were generated to mimic the arrival patterns and data volumes encountered in e-commerce, social media, and streaming applications. Additionally, a set of real-world traces derived from production logs was included to ensure that results extend beyond purely idealized scenarios.

A prototypical test system was deployed on major public cloud platforms, using a combination of on-demand and spot instances [67]. The experiments encompassed multiple resource configurations, each aligned to a particular cost optimization strategy, ranging from simple static provisioning to advanced heuristic-based autoscaling. The baseline scenario involved manually scaling compute resources according to rough rules of thumb, a practice common in many organizations [68]. The more sophisticated techniques employed a combination of queueing-based capacity planning, evolutionary scheduling, and cost-aware data placement policies.

Results indicated that dynamic and heuristic methods consistently reduced operating costs by 20% to 35% compared to the baseline, with the magnitude of savings depending heavily on workload burstiness. For workloads with moderate, predictable patterns, heuristic-based autoscaling outperformed both fully static provisioning and purely reactive threshold-based scaling [69]. However, in environments with extremely volatile load fluctuations, the overhead associated with frequent provisioning and deprovisioning events sometimes eroded cost savings. This phenomenon underscores the importance of fine-tuning the elasticity mechanism's aggressiveness. Excessively rapid scaling reacts to transient spikes that might not warrant a full reconfiguration, incurring extra charges for short-lived instance usage. [70, 71]

In addition to cost metrics, performance indicators such as average job completion time and missed deadline rates were measured. Although certain strategies favored lower expenses, they sometimes yielded increased completion times for the most latency-sensitive jobs [72]. Methods that incorporate robust or chance constraints in their optimization formulations were more effective at satisfying stringent performance targets but tended to cost more overall. A trade-off thus emerges: risk-averse strategies ensure deadlines are consistently met, albeit with higher resource usage, while purely cost-driven policies risk crossing performance thresholds under heavy or unpredictable loads.

A significant factor influencing these results was data movement [73]. In tests that involved geographically distributed data, transferring large volumes to cheaper compute regions produced initial cost improvements but introduced significant network egress fees, often canceling out the gains. Strategies that dynamically co-located data and compute resources, or that partitioned datasets more efficiently, generally performed better. Yet the overhead in orchestrating data partitions, especially when combined with advanced analytics frameworks, sometimes led to non-negligible latencies [74]. This illustrates the complexity of designing a single, universal cost optimization framework. Different workload characteristics require different priorities; a strategy that excels for compute-intensive tasks may be suboptimal for data-intensive ones. [75]

Moreover, the comparison of heuristic versus machine learning-based approaches uncovered interesting insights. Machine learning-based predictive autoscaling tended to produce near-optimal resource allocation when workloads followed discernible patterns. However, these methods sometimes struggled with sudden, atypical surges [76]. Their adaptability depends on retraining intervals and the diversity of the data used for model building. In highly volatile settings, evolutionary algorithms and swarm-based techniques exhibited more robust performance, as they relied less on historical patterns and more on continuous solution refinement.

The experiments also revealed several limitations [77]. First, most optimization models assume partial usage of an hour is billed fractionally, but certain cloud providers still enforce per-hour billing for some instance categories. This discrepancy can alter the cost function significantly [78]. Second, the presence of co-located tenants can introduce interference that the models did not fully account for, leading to deviations between theoretical predictions and actual costs. Third, the performance of advanced optimization schemes can degrade in multi-cloud or hybrid-cloud deployments, where data transfer times across different vendors are even less predictable. Finally, the overhead of implementing complex heuristics and maintaining robust monitoring systems can eat into the cost savings, particularly for smaller-scale deployments [79]. Organizations must weigh the engineering complexity and maintenance burden against the potential financial benefits.

Despite these limitations, the experimental outcomes reinforce that a carefully tuned combination of mathematical modeling, heuristic allocation, and adaptive autoscaling can significantly reduce the costs of big data analytics in public cloud environments. However, the search for a universal, one-size-fits-all solution is likely to remain elusive, given the diversity of workloads and the ever-evolving nature of cloud service offerings [80]. The ultimate success of any optimization strategy depends on continuous refinement, real-time monitoring, and a deep understanding of the specific workloads at hand.

6. Conclusion

This paper has explored the multifaceted landscape of cost optimization for big data analytics deployed in public cloud infrastructures, encompassing both theoretical models and practical heuristics [81]. By dissecting the complexity of pricing structures, resource allocation, and workload characteristics, it has been shown that effective cost reduction requires more than simple overprovisioning avoidance. Instead, an orchestrated approach blending rigorous mathematical modeling, dynamic scheduling, and data-driven predictive techniques emerges as the most promising pathway for controlling operational expenditures.

A core insight is that big data analytics often spans interdependent tasks bound by performance, storage, and networking constraints [82]. Formal optimization formulations grounded in linear, integer, or robust modeling help articulate these relationships, yielding frameworks for analyzing trade-offs among various cost drivers. Heuristic and metaheuristic methods, together with machine learning-based scheduling and autoscaling approaches, can solve large-scale instances that defy tractable exact solutions. The results from empirical evaluations underscore the potential for substantial cost savings, sometimes exceeding 30% relative to baseline provisioning strategies. [83]

It is also evident that cloud cost optimization cannot be undertaken in isolation from performance considerations. Many analytics pipelines carry latency requirements that limit how aggressively costs can be minimized before performance degrades beyond acceptable thresholds [84]. Furthermore, the movement of massive datasets across regions to exploit cheaper compute resources frequently carries a network charge that mitigates or reverses any anticipated benefits. Complexity arises from uncertain workloads, multi-tenant interference, and evolving service-level agreements, making it essential for optimization frameworks to incorporate robust or stochastic elements that account for these variations.

Among the main limitations highlighted are the difficulty in precisely modeling co-located tenants, the overhead of frequent scaling operations in rapidly changing environments, and the occasional mismatch between theoretical assumptions about billing increments and real-world provider policies [85]. These issues reflect the ever-evolving nature of cloud services, necessitating adaptive strategies that can assimilate new service offerings, billing methods, and workload patterns. In practice, the maintenance cost of implementing advanced optimization must be balanced against the projected savings, particularly in organizations with limited engineering bandwidth.

In conclusion, cost optimization in big data analytics is best achieved through a holistic strategy that acknowledges both the technical and economic realities of modern cloud environments [86]. Techniques such as dynamic provisioning, advanced scheduling, robust modeling, and machine learning-based prediction can produce significant efficiencies while preserving critical performance guarantees. Nevertheless, continuous refinements and vigilant monitoring remain indispensable, as the dynamic nature of big data and the public cloud ecosystem demands iterative improvement rather than static solutions. This research has demonstrated that the marriage of mathematical rigor and practical heuristics can yield marked cost reductions, and it opens the door for further exploration into hybrid-cloud scenarios, federated analytics frameworks, and emerging service models that will shape the next generation of cost-aware big data processing. [87]

References

[1] A. P. Carrieri, W. P. M. Rowe, M. Winn, and E. O. Pyzer-Knapp, "Aaai - a fast machine learning workflow for rapid phenotype prediction from whole shotgun metagenomes," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9434–9439, 7 2019.

[2] J. Woo and M. Mishra, "Predicting the ratings of amazon products using big data," *WIREs Data Mining and Knowledge Discovery*, vol. 11, 12 2020.

[3] F. Musazade, N. Jamalova, and J. Hasanov, "Review of techniques and models used in optical chemical structure recognition in images and scanned documents.," *Journal of cheminformatics*, vol. 14, pp. 61–, 9 2022.

[4] C. Clott and B. Hartman, "Do maritime innovation centers produce results?," *WMU Journal of Maritime Affairs*, vol. 21, pp. 283–326, 8 2022.

[5] J. Grandinetti, "Welcome to a new generation of entertainment: Amazon web services and the normalization of big data analytics and rfid tracking," *Surveillance & Society*, vol. 17, pp. 169–175, 3 2019.

[6] B. Yum, A. Adelsheimer, R. Tafreshi, R. Hagopian, and J. Kim, "Clinical implications of machine learning, artificial intelligence, and radiomics in cardiac imaging," *Current Treatment Options in Cardiovascular Medicine*, vol. 22, pp. 1–13, 10 2020.

[7] B. Guo, S. Hampton, and S. D. Mathur, "Can we observe fuzzballs or firewalls," *Journal of High Energy Physics*, vol. 2018, pp. 162–, 7 2018.

[8] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Approximate query processing for big data in heterogeneous databases," in *2020 IEEE international conference on big data (big data)*, pp. 5765–5767, IEEE, 2020.

[9] S. Monteith, T. Glenn, J. R. Geddes, E. Severus, P. C. Whybrow, and M. Bauer, "Internet of things issues related to psychiatry.," *International journal of bipolar disorders*, vol. 9, pp. 11–11, 4 2021.

[10] H. A. Zurqani, C. J. Post, E. A. Mikhailova, M. P. Cope, J. S. Allen, and B. Lytle, "Evaluating the integrity of forested riparian buffers over a large area using lidar data and google earth engine," *Scientific reports*, vol. 10, pp. 14096–14096, 8 2020.

[11] H. Gebre-Amlak, A. M. A. Jabbari, Y. Chen, B.-Y. Choi, C.-T. Huang, and S. Song, "Agile polymorphic software-defined fog computing platform for mobile wireless controllers and sensors," *International Journal of Internet Technology and Secured Transactions*, vol. 9, no. 4, pp. 426–, 2019.

[12] J. Akhavan, J. Lyu, and S. Manoochehri, "A deep learning solution for real-time quality assessment and control in additive manufacturing using point cloud data," *Journal of Intelligent Manufacturing*, vol. 35, pp. 1389–1406, 4 2023.

[13] M. Obschonka and D. B. Audretsch, "Artificial intelligence and big data in entrepreneurship: A new era has begun," *Small Business Economics*, vol. 55, pp. 529–539, 6 2019.

[14] A. Farhan, R. Camacho Barranco, M. Akbar, and M. S. Hossain, “Temporal word embedding with predictive capability,” *Knowledge and Information Systems*, vol. 65, pp. 5159–5194, 7 2023.

[15] J. A. Kaduk, S. J. L. Billinge, R. E. Dinnebier, N. Henderson, I. Madsen, R. Černý, M. Leoni, L. Lutterotti, S. Thakral, and D. Chateigner, “Powder diffraction,” *Nature Reviews Methods Primers*, vol. 1, 11 2021.

[16] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “Automatic visual recommendation for data science and analytics,” in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, pp. 125–132, Springer, 2020.

[17] S. Grossberg, “The resonant brain: How attentive conscious seeing regulates action sequences that interact with attentive cognitive learning, recognition, and prediction,” *Attention, perception & psychophysics*, vol. 81, pp. 2237–2264, 6 2019.

[18] N. A. Cressie and C. K. Wikle, “Measuring, mapping, and uncertainty quantification in the space-time cube,” *Revista Matemática Complutense*, vol. 33, pp. 643–660, 6 2020.

[19] R. Avula, “Optimizing data quality in electronic medical records: Addressing fragmentation, inconsistencies, and data integrity issues in healthcare,” *Journal of Big-Data Analytics and Cloud Computing*, vol. 4, no. 5, pp. 1–25, 2019.

[20] A. Miglio, V. Brousseau-Couture, E. Godbout, G. Antonius, Y.-H. Chan, S. G. Louie, M. Côté, M. Giantomassi, and X. Gonze, “Predominance of non-adiabatic effects in zero-point renormalization of the electronic band gap,” *npj Computational Materials*, vol. 6, pp. 167–, 11 2020.

[21] J. Bloch, J. Glesaaen, J. J. M. Verbaarschot, and S. Zafeiropoulos, “Complex langevin simulation of a random matrix model at nonzero chemical potential,” *Journal of High Energy Physics*, vol. 2018, pp. 1–33, 3 2018.

[22] M. Demertzis, S. Merler, and G. B. Wolff, “Capital markets union and the fintech opportunity,” *Journal of Financial Regulation*, vol. 4, pp. 157–165, 1 2018.

[23] J. H. Wise, J. A. Regan, B. W. O’Shea, M. L. Norman, T. P. Downes, and H. Xu, “Formation of massive black holes in rapidly growing pre-galactic gas clouds,” *Nature*, vol. 566, pp. 85–88, 1 2019.

[24] B. Qolomany, I. Mohammed, A. Al-Fuqaha, M. Guizani, and J. Qadir, “Trust-based cloud machine learning model selection for industrial iot and smart city services,” *IEEE Internet of Things Journal*, vol. 8, pp. 2943–2958, 2 2021.

[25] J. Bunder, I. G. Kevrekidis, and A. J. Roberts, “Equation-free patch scheme for efficient computational homogenisation via self-adjoint coupling,” *Numerische Mathematik*, vol. 149, pp. 229–272, 10 2021.

[26] R. Lu, “Merger and acquisitions of internet companies: A comprehensive analysis,” *Advances in Economics, Management and Political Sciences*, vol. 44, pp. 221–226, 11 2023.

[27] J. A. Warrick, K. Vos, A. E. East, and S. Vitousek, “Fire (plus) flood (equals) beach: coastal response to an exceptional river sediment discharge event.,” *Scientific reports*, vol. 12, pp. 3848–, 3 2022.

[28] M. Shokrof and M. Abouelhoda, “Ioncram: a reference-based compression tool for ion torrent sequence files,” *BMC bioinformatics*, vol. 21, pp. 1–16, 9 2020.

[29] N. Sahu, D. Peng, and H. Sharif, “Diagnosis-steganography-transmission: An innovative integrated paradigm for ecg healthcare,” *SN Computer Science*, vol. 2, pp. 1–22, 6 2021.

[30] A. K. Saxena, “Evaluating the regulatory and policy recommendations for promoting information diversity in the digital age,” *International Journal of Responsible Artificial Intelligence*, vol. 11, no. 8, pp. 33–42, 2021.

[31] D. Ojika, A. Gordon-Ross, H. Lam, and B. Patel, “Faam: Fpga-as-a-microservice - a case study for data compression,” *EPJ Web of Conferences*, vol. 214, pp. 07029–, 9 2019.

[32] K. Neshatpour, M. Malik, A. Sasan, S. Rafatirad, and H. Homayoun, “Hardware accelerated mappers for hadoop mapreduce streaming,” *IEEE Transactions on Multi-Scale Computing Systems*, vol. 4, pp. 734–748, 10 2018.

[33] A. Megahed, J. A. Pazour, A. Nazeem, S. Tata, and M. Mohamed, “Monitoring services in the internet of things: an optimization approach,” *Computing*, vol. 101, pp. 1119–1145, 9 2018.

[34] A. Kocheturov, P. M. Pardalos, and A. Karakitsiou, “Massive datasets and machine learning for computational biomedicine: trends and challenges,” *Annals of Operations Research*, vol. 276, pp. 5–34, 5 2018.

[35] S. Dhal, B. M. Wyatt, S. Mahanta, N. Bhattarai, S. Sharma, T. Rout, P. Saud, and B. S. Acharya, "Internet of things (iot) in digital agriculture: An overview," *Agronomy Journal*, vol. 116, pp. 1144–1163, 6 2023.

[36] L. M. M. Peralta, "Resisting techno-orientalism and mimicry stereotypes in and through data science education," *TechTrends*, vol. 67, pp. 426–434, 3 2023.

[37] C. C. Wang, N. C. Su, J. P. Hou, and D. I. Lee, "Evaluation of the 2.5-km cloud-resolving storm simulator in predicting local afternoon convection during the summer in taiwan," *Asia-Pacific Journal of Atmospheric Sciences*, vol. 54, pp. 489–498, 9 2018.

[38] M. Karouzos, "Black holes," *Nature Astronomy*, vol. 3, pp. 2–5, 1 2019.

[39] I. Tanaka, K. Rajan, and C. Wolverton, "Data-centric science for materials innovation," *MRS Bulletin*, vol. 43, pp. 659–663, 9 2018.

[40] C. Sur, "Deepseq: learning browsing log data based personalized security vulnerabilities and counter intelligent measures," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 3573–3602, 10 2018.

[41] M. Kansara, "A structured lifecycle approach to large-scale cloud database migration: Challenges and strategies for an optimal transition," *Applied Research in Artificial Intelligence and Cloud Computing*, vol. 5, no. 1, pp. 237–261, 2022.

[42] M. H. de Menendez, R. Morales-Menendez, C. A. Escobar, and J. Arinez, "Biometric applications in education," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 15, pp. 365–380, 7 2021.

[43] M. Herrero-Huerta, P. Rodríguez-Gonzálvez, and K. M. Rainey, "Yield prediction by machine learning from uas-based mulit-sensor data fusion in soybean," *Plant methods*, vol. 16, pp. 1–16, 6 2020.

[44] E. E. Saul, R. B. Guerra, M. E. Saul, L. L. da Silva, G. F. P. Aleixo, R. M. K. Matuda, and G. Lopes, "The challenges of implementing low-dose computed tomography for lung cancer screening in low- and middle-income countries," *Nature cancer*, vol. 1, pp. 1140–1152, 11 2020.

[45] D. Lee, Y. Huh, C. Y. Lin, and C. M. Reigeluth, "Technology functions for personalized learning in learner-centered schools," *Educational Technology Research and Development*, vol. 66, pp. 1269–1302, 7 2018.

[46] H. Fukuda, A. V. Manohar, H. Murayama, and O. Telem, "Axion strings are superconducting," *Journal of High Energy Physics*, vol. 2021, pp. 1–42, 6 2021.

[47] Abdullah-Al-Faisal, A. A. Kafy, M. A. Fattah, D. M. A. Jahir, A. A. Rakib, Z. A. Rahaman, J. Ferdousi, and X. Huang, "Assessment of temporal shifting of pm2.5, lockdown effect, and influences of seasonal meteorological factors over the fastest-growing megacity, dhaka," *Spatial Information Research*, vol. 30, pp. 441–453, 3 2022.

[48] J. Fernández-Villaverde, "Has machine learning rendered simple rules obsolete," *European journal of law and economics*, vol. 52, pp. 1–15, 2 2021.

[49] R. Avula, "Strategies for minimizing delays and enhancing workflow efficiency by managing data dependencies in healthcare pipelines," *Eigenpub Review of Science and Technology*, vol. 4, no. 1, pp. 38–57, 2020.

[50] D. T. Hai, "Network coding for improving throughput in wdm optical networks with dedicated protection," *Optical and Quantum Electronics*, vol. 51, pp. 1–15, 11 2019.

[51] C. Sridhar, O. S. Lih, V. Jahmunah, J. E. W. Koh, E. J. Ciaccio, T. R. San, N. Arunkumar, S. Kadry, and U. R. Acharya, "Accurate detection of myocardial infarction using non linear features with ecg signals," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 3227–3244, 10 2020.

[52] V. de Paul Obade and C. Gaya, "Digital technology dilemma: on unlocking the soil quality index conundrum," *Bioresources and bioprocessing*, vol. 8, pp. 6–6, 1 2021.

[53] A. Rahman, Z. Tari, D. Zhu, F. Piccialli, and X. Wang, "Ieee access special section editorial: Cloud-fog-edge computing in cyber–physical–social systems (cpss)," *IEEE Access*, vol. 8, pp. 222859–222864, 2020.

[54] S. Bhadra, V. Sagan, J. Skobalski, F. Grignola, S. Sarkar, and J. Vilbig, "End-to-end 3d cnn for plot-scale soybean yield prediction using multitemporal uav-based rgb images," *Precision Agriculture*, vol. 25, pp. 834–864, 12 2023.

[55] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Integrating polystore rdbms with common in-memory data," in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 5762–5764, IEEE, 2020.

[56] M. Al-Maini, M. Maindarkar, G. D. Kitas, N. N. Khanna, D. P. Misra, A. M. Johri, L. Mantella, V. Agarwal, A. Sharma, I. M. Singh, G. Tsoulfas, J. R. Laird, G. Faa, J. Teji, M. Turk, K. Viskovic, Z. Ruzsa, S. Mavrogeni, V. Rathore, M. Miner, M. K. Kalra, E. R. Isenovic, L. Saba, M. M. Fouda, and J. S. Suri, "Artificial intelligence-based preventive, personalized and precision medicine for cardiovascular disease/stroke risk assessment in rheumatoid arthritis patients: a narrative review.," *Rheumatology international*, vol. 43, pp. 1965–1982, 8 2023.

[57] J. Woo, S.-J. Shin, W. Seo, and P. Meilanitasari, "Developing a big data analytics platform for manufacturing systems: architecture, method, and implementation," *The International Journal of Advanced Manufacturing Technology*, vol. 99, pp. 2193–2217, 7 2018.

[58] S. J. Ghoushchi, S. Dorosti, M. Khazaeili, and A. Mardani, "Extended approach by using best–worst method on the basis of importance–necessity concept and its application," *Applied Intelligence*, vol. 51, pp. 8030–8044, 3 2021.

[59] O. Rostami, M. Tavakoli, A. Tajally, and M. GhanavatiNejad, "A goal programming-based fuzzy best-worst method for the viable supplier selection problem: a case study.," *Soft computing*, vol. 27, pp. 2827–2852, 11 2022.

[60] M. Wörsdörfer, "Ordoliberalism 2.0: Towards a new regulatory policy for the digital age," *Philosophy of Management*, vol. 19, pp. 191–215, 4 2020.

[61] A. S. Villar and N. Khan, "Robotic process automation in banking industry: a case study on deutsche bank," *Journal of Banking and Financial Technology*, vol. 5, pp. 71–86, 5 2021.

[62] C. Ordonez, Y. Zhang, and S. L. Johnsson, "Scalable machine learning computing a data summarization matrix with a parallel array dbms," *Distributed and Parallel Databases*, vol. 37, pp. 329–350, 6 2018.

[63] P. O. A. Navaux, A. F. Lorenzon, and M. da Silva Serpa, "Challenges in high-performance computing," *Journal of the Brazilian Computer Society*, vol. 29, pp. 51–62, 8 2023.

[64] M. Abouelyazid and C. Xiang, "Architectures for ai integration in next-generation cloud infrastructure, development, security, and management," *International Journal of Information and Cybersecurity*, vol. 3, no. 1, pp. 1–19, 2019.

[65] S. Paul, M. Riffat, A. Yasir, M. N. Mahim, B. Y. Sharnali, I. T. Naheen, A. Rahman, and A. Kulkarni, "Industry 4.0 applications for medical/healthcare services," *Journal of Sensor and Actuator Networks*, vol. 10, pp. 43–, 6 2021.

[66] S. Spooner, Z. Li, and S. Sridhar, "Hidden phenomena during transient reaction trajectories in liquid metals processing," *Metallurgical and Materials Transactions B*, vol. 51, pp. 1301–1314, 7 2020.

[67] E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Computational Materials*, vol. 8, 4 2022.

[68] A. Magdy, L. Abdelhafeez, Y. Kang, E. Ong, and M. F. Mokbel, "Microblogs data management: a survey," *The VLDB Journal*, vol. 29, pp. 177–216, 9 2019.

[69] E. Naderi, L. Mirzaei, M. Pourakbari-Kasmaei, F. V. Cerna, and M. Lehtonen, "Optimization of active power dispatch considering unified power flow controller: application of evolutionary algorithms in a fuzzy framework," *Evolutionary Intelligence*, vol. 17, pp. 1357–1387, 1 2023.

[70] C. Turner, "Neuromedia, cognitive offloading, and intellectual perseverance," *Synthese*, vol. 200, 3 2022.

[71] R. Avula, "Applications of bayesian statistics in healthcare for improving predictive modeling, decision-making, and adaptive personalized medicine," *International Journal of Applied Health Care Analytics*, vol. 7, no. 11, pp. 29–43, 2022.

[72] H. W. Grießhammer, J. A. McGovern, and D. R. Phillips, "Comprehensive study of observables in compton scattering on the nucleon," *The European Physical Journal A*, vol. 54, pp. 37–, 3 2018.

[73] S. Zaheer, A. W. Malik, A. Rahman, and S. A. Khan, "Locality-aware process placement for parallel and distributed simulation in cloud data centers," *The Journal of Supercomputing*, vol. 75, pp. 7723–7745, 8 2019.

[74] G. Fiscon, F. Conte, and P. Paci, "Swim tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines.," *BMC bioinformatics*, vol. 19, pp. 436–436, 11 2018.

[75] A. Adamo, P. Zeidler, J. M. D. Kruijssen, M. Chevance, M. Gieles, D. Calzetti, C. Charbonnel, H. Zinnecker, and M. Krause, "Star clusters near and far tracing star formation across cosmic time," *Space Science Reviews*, vol. 216, pp. 1–55, 6 2020.

[76] C. Seiferth, L. Vogel, B. Aas, I. Brandhorst, P. Carlbring, A. Conzelmann, N. Esfandiari, M. Finkbeiner, K. Hollmann, H. Lautenbacher, E. Meinzinger, A. Newbold, A. Opitz, T. J. Renner, L. B. Sander, P. S. Santangelo, R. Schoedel, B. Schuller, C. Stachl, Y. Terhorst, J. Torous, K. Wac, A. Werner-Seidler, S. Wolf, and J. Löchner, "How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health," *Nature Mental Health*, vol. 1, pp. 542–554, 8 2023.

[77] E. M. Shakshuki, A.-U.-H. Yasar, and H. Malik, "Special issue on advances in ambient intelligence and pervasive computing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 3893–3894, 12 2019.

[78] N. Zincir-Heywood, R. Birke, E. Bou-Harb, G. Casale, K. El-Khatib, T. Inoue, N. Kumar, H. Lutfiyya, D. Puthal, A. Shami, N. Stakhanova, and F. Zulkernine, "Guest editorial: Special section on machine learning and artificial intelligence for managing networks, systems, and services—part ii," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 882–889, 2023.

[79] O. Marques, L. Zaniolo, A. Ishay, A. Elshaikh, null Whitney, A. Andrews, A. Bulatovic, H. M., L. J. Cornelissen, B. Pereira, V. Stirler, E. Fennema, P. M. A. van Ooijen, F. García-Castro, J. C. March, A. Jiménez-Pastor, D. Cuixat, M. Llambrich, E. Osorio, Ángel Alberich-Bayarri, I. Blanquer, E. Camacho-Ramos, A. Brito, C. Fetzer, A. L. Carvalho, and F. García, "Eusomii annual meeting 2019 book of abstracts.," *Insights into imaging*, vol. 11, pp. 25–, 2 2020.

[80] L. Liu, "The rise of data politics: Digital china and the world.," *Studies in comparative international development*, vol. 56, pp. 1–23, 3 2021.

[81] L. Bitomsky, O. Bürger, B. Häckel, and J. Töppel, "Value of data meets it security – assessing it security risks in data-driven value chains," *Electronic Markets*, vol. 30, pp. 589–605, 2 2020.

[82] O. Castro-Lopez, D. E. Lopez-Barron, and I. F. Vega-Lopez, "Next-generation heartbeat classification with a column-store dbms and udfs," *Journal of Intelligent Information Systems*, vol. 54, pp. 363–390, 4 2019.

[83] T. G. Myers, P. N. Ramkumar, B. F. Ricciardi, K. L. Urish, J. Kipper, and C. Ketonis, "Artificial intelligence and orthopaedics: An introduction for clinicians.," *The Journal of bone and joint surgery. American volume*, vol. 102, pp. 830–840, 2 2020.

[84] Z. Chen, S. Wei, W. Yu, J. Nguyen, and W. G. Hatcher, "A cloud/edge computing streaming system for network traffic monitoring and threat detection," *International Journal of Security and Networks*, vol. 13, no. 3, pp. 169–186, 2018.

[85] A. Mohapatra and S. Panda, "Big data analytics and cloud computing: Implementation issues and effects," *SSRN Electronic Journal*, 2023.

[86] E. Eslami, Y. Choi, Y. Lops, and A. Sayeed, "A real-time hourly ozone prediction system using deep convolutional neural network," *Neural Computing and Applications*, vol. 32, pp. 8783–8797, 6 2019.

[87] H. Sharma, M. Shrivastava, and B. Singh, "Physics informed deep neural network embedded in a chemical transport model for the amazon rainforest," *npj Climate and Atmospheric Science*, vol. 6, 3 2023.