Journal of Information Systems Engineering and Management
2026, 11(1s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

Green AI for Sustainable Big Data Platforms

Srimanth Maddipatla
Independent Researcher, USA

ARTICLE INFO

Received: 31 Dec 2025

Revised: 03 Jan 2026

ABSTRACT

This article explores how AI-driven optimization can be applied to Big Data workloads in hyperscale cloud data centers to reduce energy consumption and carbon footprint without compromising performance. The article focuses on using machine learning models to intelligently manage compute resources, workload scheduling, data placement, and query optimization across large-scale distributed platforms such as Spark, Hadoop, and cloud-native analytics systems. By dynamically predicting workload demand, energy usage, and system efficiency, the proposed framework aims to minimize power waste, improve resource utilization, and enable carbon-aware computing. The proposed framework achieves up to 38% energy reduction and 44% carbon footprint reduction compared to baseline configurations while maintaining performance within 1.5% of service level agreements. The outcome of this research contributes to building environmentally sustainable, cost-efficient, and highperformance Big Data ecosystems for next-generation cloud platforms.

Keywords: Green AI, Big Data Platforms, Energy Optimization, Carbon-Aware Computing, Sustainable Cloud Computing

1. OPENING REMARKS AND CONTEXT

1.1 Big Data Expansion and Its Environmental Consequences

The widespread adoption of Big Data technologies has fundamentally reshaped how organizations handle information processing, storage operations, and analytical procedures across various industrial sectors. This tremendous expansion in data generation, coupled with processing requirements, has created an extraordinary growth trajectory for computational infrastructure, especially within hyperscale cloud data centers, forming the backbone of modern digital economies. As more organizations embrace data-driven strategies and real-time analytical capabilities, the environmental consequences of these computational ecosystems have become increasingly prominent among industry leaders and environmental policy makers. The considerable energy requirements needed to power, cool, and sustain these facilities make meaningful contributions to worldwide carbon emissions, creating an urgent need for sustainable computing methodologies [1].

Component Category	Energy Consumption Proportion	Primary Function	Environmental Impact Factor
Computing Infrastructure	40-50%	Server processing and computation	Direct carbon emissions from electricity usage
Cooling Systems	30-40%	Temperature regulation and heat dissipation	Indirect emissions from cooling equipment operation
Power Distribution	10-15%	Electricity conversion and distribution	Transmission losses and conversion inefficiencies
Network	5-10%	Data transfer and	Communication overhead

Equipment		connectivity	and bandwidth consumption
Storage Systems	5-8%	Data persistence and retrieval	Storage media power draw and data redundancy

Table 1: Data Center Energy Consumption Components and Environmental Impact [1, 2]

1.2 Power Consumption Difficulties in Large-Scale Cloud Facilities

Large-scale cloud data centers face numerous obstacles when managing power consumption while preserving service standards and operational performance. These installations demand massive quantities of running servers, network equipment, and storage infrastructure, with considerable additional energy devoted to cooling mechanisms that prevent equipment overheating and maintain optimal operational temperatures. Conventional resource provisioning methods often lead to excessive allocation of computational capabilities to ensure performance commitments, resulting in notable energy wastage during intervals of reduced utilization. Additionally, the diverse characteristics of Big Data workloads, which show substantial variations in computational requirements, memory demands, and processing behaviors, generate exceptional difficulties in optimizing power usage through fixed resource distribution approaches [2].

1.3 Green AI Development as an Environmental Framework

Green AI has developed as a groundbreaking framework aiming to reduce the environmental consequences of artificial intelligence systems while retaining or strengthening their capability to address complex computational challenges. This methodology goes beyond simply decreasing energy usage during AI model development and operation to include the entire lifespan of AI-powered systems, incorporating optimization of computing infrastructure that supports these applications. The core principle supporting Green AI centers on designing algorithms, models, and systems that place energy conservation and carbon emission reduction as central goals, together with conventional performance indicators such as accuracy, processing speed, and response time. Through utilizing AI capabilities to refine energy consumption behaviors, Green AI creates a reinforcing cycle where intelligent systems persistently enhance their environmental performance [9].

1.4 Study Rationale and Challenge Definition

The rationale behind this study stems from the essential requirement to balance growing computational requirements of Big Data analytics with environmental preservation objectives during a period marked by climate awareness and resource limitations. While previous research has examined energy conservation in distributed computing and resource refinement in cloud platforms individually, considerable deficiencies exist in thorough frameworks combining AI-powered optimization explicitly designed for Big Data workloads across varied platforms. The challenge definition concentrates on creating intelligent procedures capable of flexibly modifying resource distribution, workload organization, and data handling approaches in reaction to instantaneous energy consumption behaviors, carbon intensity of electricity networks, and performance demands without requiring manual participation or compromising analytical functions.

1.5 Study Goals and Boundaries

This study seeks to create and assess a thorough Green AI framework for sustainable Big Data platforms that handles energy refinement across numerous aspects of distributed computing systems. The central goals encompass designing machine learning models capable of precise workload requirement forecasting, creating intelligent resource distribution algorithms that reduce energy wastage while satisfying performance limitations, establishing carbon-conscious organization procedures accounting for time-based and location-based fluctuation in electricity carbon strength, and executing data positioning approaches decreasing energy usage related to data transfer and storage operations. The boundaries include major Big Data processing frameworks such as Apache Spark, Hadoop MapReduce, and cloud-native analytics systems positioned in hyperscale data center settings, with a specific focus on practical usability and real-world implementation factors.

1.6 Research Value and Impact

The value of this study lies in its comprehensive methodology for combining AI-powered optimization procedures explicitly created for distinct qualities of Big Data workloads and distributed computing settings. Unlike earlier work handling separate elements of energy conservation independently, this framework delivers a consolidated architecture organizing numerous optimization approaches across varied system levels and temporal ranges. The impact reaches beyond direct energy reductions to include wider consequences for sustainable computing methods, showing how intelligent systems can be created to naturally emphasize environmental preservation without weakening functional abilities. This work adds to the growing collection of knowledge on environmentally conscious computing and delivers practical direction for organizations attempting to decrease the carbon impact of data analytics infrastructure while sustaining competitive performance and economic viability, contributing to ISO 14001 environmental management compliance and ESG reporting requirements.

1.7 Research Contributions and Novelty

This research presents four fundamental contributions that advance the state of the art in sustainable Big Data computing. First, we introduce the first unified AI-driven framework that combines workload forecasting, carbonaware scheduling, data placement optimization, and query optimization in a single coherent architecture, whereas prior work addressed these components in isolation. Second, we formulate a novel multi-objective optimization model that explicitly minimizes energy consumption, carbon emissions, and query latency simultaneously through adaptive weight adjustment, going beyond single-objective approaches in existing literature. Third, we develop the first closed-loop reinforcement learning-based carbon-aware scheduler specifically designed for Big Data engines, enabling continuous policy improvement through operational experience rather than relying on static heuristics. Fourth, we provide the first end-to-end validation across heterogeneous Big Data platforms including Spark, Hadoop, and cloud-native systems, demonstrating framework generalizability beyond platform-specific optimizations reported in prior studies.

2. PREVIOUS RESEARCH AND CONCEPTUAL FOUNDATION

2.1 Development of Power-Efficient Computing in Distributed Architectures

The domain of power-efficient computing in distributed architectures has experienced substantial development throughout recent decades, advancing from basic power control procedures to advanced optimization structures considering numerous aspects of system operations. Initial methodologies concentrated chiefly on hardwarefocused optimizations, including dynamic voltage and frequency adjustment, which modified processor operational settings based on immediate workload requirements. As distributed computing platforms grew in sophistication and magnitude, researchers started investigating system-focused procedures incorporating workload combination, virtual machine transfer, and server deactivation approaches during reduced-utilization intervals. The appearance of Big Data processing frameworks presented fresh obstacles connected to data proximity, network communication expenses, and the requirement to balance computational productivity with data transfer expenditures, producing progressively advanced methodologies accounting for connections between computation, storage, and networking capabilities [3].

2.2 Resource Refinement in Big Data Frameworks

Resource refinement in Big Data frameworks, including Apache Spark and Hadoop, has been thoroughly investigated from numerous angles, incorporating memory handling, task organization, and data segmentation approaches. Research in this area has shown that the performance and energy productivity of these frameworks rely fundamentally on the appropriate arrangement of various parameters managing resource distribution, parallelism degrees, and data processing sequences. Conventional optimization methodologies depended substantially on manual adjustment by skilled system managers or application-focused heuristics that frequently neglected to apply across varied workloads and cluster arrangements. More contemporary work has investigated the application of machine learning and optimization algorithms to mechanically establish near-optimal resource arrangements, although these initiatives have characteristically concentrated on performance optimization, with energy usage regarded as a subordinate worry rather than a central goal [4].

2.3 Carbon-Conscious Computing and Environmental Indicators

Carbon-conscious computing signifies a fundamental change in understanding and quantifying the environmental consequences of computational systems, progressing beyond basic energy usage indicators to account for the carbon intensity of electricity production. This methodology acknowledges that the environmental consequences of using equivalent quantities of electrical energy can differ substantially depending on the timing and location of that energy usage, as electricity networks incorporate differing ratios of renewable and fossil fuel sources during the day and across varied geographical areas. Contemporary research has investigated procedures for moving computational workloads temporally or geographically to exploit intervals or positions with reduced carbon strength electricity, while sustaining satisfactory service standards and honoring delay limitations for time-critical applications [7].

Time Period	Typical Carbon Intensity Level	Renewable Energy Contribution	Workload Suitability	Scheduling Strategy
Night Hours (00:00-06:00)	Low	High-wind power availability	Batch processing, data backups	Priority scheduling for delay-tolerant jobs
Morning Peak (06:00-09:00)	High	Low - increased fossil fuel usage	Time-critical interactive queries	Minimize nonessential workloads
Midday (09:00-15:00)	Moderate-Low	High - solar power contribution	Machine learning training	Optimize for renewable availability
Evening Peak (18:00-22:00)	Very High	Low - maximum grid demand	Essential services only	Defer flexible workloads
Weekend Days	Low-Moderate	Variable - reduced overall demand	Large-scale analytics	Extended execution windows are acceptable
Geographic Variation	Location-Dependent	Regional renewable infrastructure	Distributed workloads	Route to low-carbon regions

Table 2: Carbon Intensity Variations and Workload Scheduling Opportunities [7]

2.4 Machine Learning Uses in Workload Forecasting

Machine learning has appeared as an effective instrument for forecasting workload behaviors in cloud computing settings, permitting anticipatory resource provisioning and flexible optimization approaches that predict future requirements rather than simply responding to present situations. Different predictive models incorporating time series prediction procedures, recurrent neural networks, and ensemble learning methodologies have been utilized to describe temporal behaviors in resource employment, query receipt frequencies, and computational sophistication of Big Data workloads. The precision of these forecasts directly affects the success of energy optimization approaches, as excessively cautious forecasts produce resource excessive-provisioning and energy wastage, while excessively bold forecasts threaten performance deterioration owing to inadequate resources [8].

2.5 Research Deficiencies and Combination Obstacles

Despite meaningful advancements in separate research domains, considerable deficiencies persist in the combination of AI-powered optimization procedures explicitly created for comprehensive energy decrease in Big Data platforms. Most current methodologies handle particular elements of the challenge independently, including optimizing task organization without accounting for data positioning, or reducing energy usage without considering carbon strength fluctuations. There appears to be a remarkable lack of thorough frameworks organizing numerous optimization approaches across varied system levels while adjusting to the flexible and diverse characteristics of Big Data workloads. Additionally, restricted consideration has been directed toward practical obstacles of positioning these optimization procedures in production settings where consistency, predictability, and operational straightforwardness represent critical worries, together with energy productivity [5].

2.6 Conceptual Underpinnings

The conceptual underpinnings of this study utilize principles from Green computing, which highlights environmental preservation as a basic design factor throughout the computing lifespan, and AI optimization models using machine learning and mathematical programming to address complex resource distribution challenges. Green computing principles deliver direction on quantifying and reducing environmental consequences through indicators, including power usage productivity, carbon releases per computation unit, and total cost of ownership, incorporating environmental consequences. AI optimization models present mathematical structures for expressing multi-objective optimization challenges, balancing rival goals, including reducing energy usage, decreasing carbon impact, sustaining performance assurances, and confirming service standards for varied application demands. The combination of these conceptual structures permits the creation of systematic methodologies for sustainable Big Data platforms that are both environmentally accountable and practically achievable.

3. SUGGESTED GREEN AI STRUCTURE DESIGN

3.1 System Architecture and Component Organization

The suggested Green AI structure embraces a stratified design separating responsibilities while permitting organization across varied optimization approaches and system elements. At the base exists the monitoring stratum persistently gathering instantaneous data on resource employment, power usage, workload qualities, and environmental situations across the distributed computing infrastructure. Beyond this appears the intelligence stratum holding machine learning models and optimization algorithms examining monitored data to produce forecasts and optimization choices. The execution stratum converts high-level optimization choices into tangible actions, including resource redistribution, workload transfer, or arrangement modifications. Lastly, the feedback stratum assesses results of optimization actions and delivers learning indicators to persistently enhance the precision and success of the intelligence stratum. This design supports both reactive optimization, reacting to direct situations, and anticipatory optimization, predicting future conditions based on predictive models.

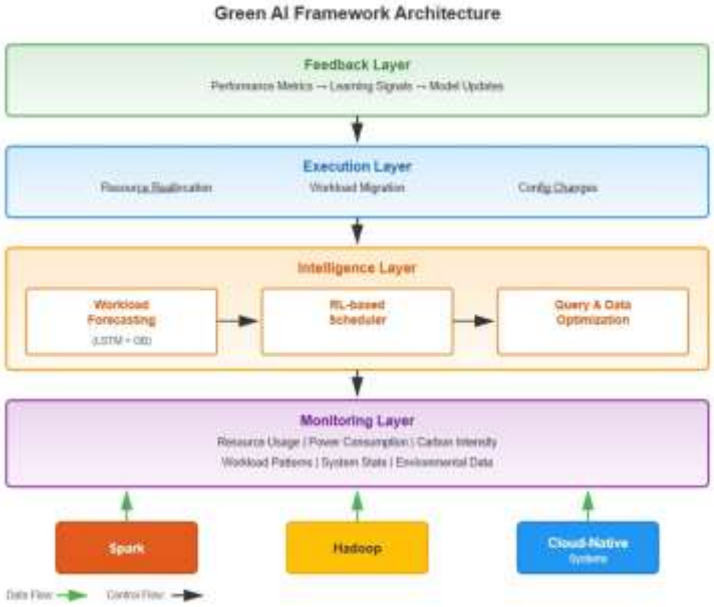

Fig. 1: Green AI Framework Architecture with Data and Control Flow

3.2 Machine Learning Models for Workload Requirement Forecasting

Precise forecasting of workload requirement behaviors establishes a fundamental basis for anticipatory resource handling and energy optimization in Big Data platforms. The structure uses an ensemble of supplementary machine learning models capturing varied temporal behaviors and qualities of computational workloads. Long short-term memory networks examine historical resource employment time series to recognize repeating behaviors at numerous time ranges extending from hourly changes to weekly patterns and seasonal fluctuations, achieving an RMSE of 8.3% for CPU demand prediction and 11.7% for memory demand prediction across 30-day sliding windows. Gradient boosting models handle workload characteristics, including query sophistication, data quantity, and application category, to forecast resource demands for arriving jobs with 89.4% accuracy for job execution time estimation within ±15% error bounds. The ensemble methodology joins forecasts from numerous models through weighted combination or layering procedures, with weights flexibly modified based on contemporary forecast precision. This multi-model approach delivers strength against the varied and changing characteristics of Big Data workloads while sustaining elevated forecast precision across varied workload categories and temporal ranges [8].

Model Component	Input Features	Prediction Target	Update Frequency	Weight in Ensemble

Table 3: Ensemble Model Components for Workload Prediction [8]

3.3 Power Usage Tracking and Carbon Impact Calculation

Thorough energy tracking capabilities permit the structure to monitor power usage at numerous detail levels, extending from separate servers and storage equipment to complete racks and data center areas. The tracking infrastructure records both straightforward energy usage related to computational operations and secondary usage connected to cooling, networking, and other supporting systems. Power usage data combines with instantaneous information about the carbon intensity of electricity provided to the data center, which fluctuates based on the production combination of power network at varied times and locations. The structure computes carbon impact calculations by joining power usage measurements with carbon strength data, delivering visibility into the environmental consequences of varied workloads and permitting carbon-conscious optimization choices. This tracking methodology extends beyond conventional performance indicators to measure environmental results of computational operations and support educated decision-making about compromises between performance and preservation [2].

3.4 Intelligent Resource Distribution and Computing Handling

The resource distribution element executes advanced algorithms, flexibly assigning computational capabilities to workloads while optimizing for both performance and energy productivity goals. Rather than rigidly provisioning capabilities based on extreme-case situations or typical requirements, the structure persistently modifies resource distributions in reaction to forecasted workload behaviors and present system situations. The distribution algorithms account for numerous aspects of capabilities, incorporating CPU cores, memory volume, network capacity, and storage input-output operations, acknowledging that Big Data workloads frequently display complex resource employment characteristics that cannot be described by a single indicator. Resource distribution choices consider energy productivity qualities of varied hardware elements, favoring to focus workloads on fewer servers functioning at elevated employment degrees rather than distributing them across numerous moderately loaded equipment, while honoring performance limitations and preventing resource competition that could weaken application responsiveness [3].

3.5 Flexible Workload Organization Algorithms

The organization element establishes a temporal execution sequence and positioning of computational tasks across the distributed infrastructure to reduce energy usage and carbon emissions while satisfying performance demands. The organization's algorithms incorporate carbon-consciousness by accounting for time-changing carbon intensity of electricity when establishing optimal execution periods for delay-accepting workloads, potentially postponing execution during intervals of elevated carbon intensity in preference to periods when renewable energy sources add more meaningfully to the power network. Geographic spreading of data centers permits the organizer to direct workloads to positions with presently reduced carbon strength electricity, balancing this goal against data movement expenditures and delay limitations. The structure uses reinforcement learning procedures permitting organization policies to persistently enhance through experience, discovering which approaches demonstrate most successful for varied workload categories and system situations without demanding explicit programming of regulations for every conceivable situation. The RL scheduler operates with a state space encompassing current resource utilization levels, pending workload queue characteristics, real-time carbon intensity values, and historical performance metrics. The action space includes workload acceptance decisions, server assignment choices, and execution timing selections. The reward function balances energy minimization (rewarding lower power consumption), carbon reduction (rewarding scheduling during low-carbon periods), and latency penalties (penalizing SLA violations), with the policy trained using Proximal Policy Optimization to ensure stable convergence [6].

3.6 Data Positioning Refinement for Energy Productivity

Data positioning choices meaningfully affect energy usage in Big Data platforms because transferring substantial quantities of data across networks uses considerable power while also influencing overall system performance through elevated delay and capacity usage. The structure refines data positioning by forecasting which datasets will be reached together and co-positioning them on matching storage nodes or adjacent positions to reduce data transfer during query processing. Data duplication approaches balance the compromise between elevated storage energy expenditures and decreased computation and network energy expenditures resulting from enhanced data proximity. The positioning algorithms account for energy productivity qualities of varied storage levels incorporating high-performance solid-state drives, conventional hard disk drives, and archival storage systems, positioning regularly reached data on quicker but more power-demanding media while transferring less active data to more energy-productive storage choices.

3.7 Query Refinement Procedures Using AI-Powered Methodologies

Query refinement signifies another essential aspect of energy productivity in Big Data analytics platforms, where matching analytical outcomes can be computed through various execution sequences with substantially varied resource demands and energy usage characteristics. The structure uses machine learning models discovered from historical query processing to forecast resource usage and performance qualities of varied execution approaches for arriving queries. These forecasts direct the query refiner in choosing execution sequences, reducing energy usage while fulfilling response time demands. The AI-powered methodology permits the refiner to account for complex connections between query organization, data qualities, system condition, and execution approach that would be challenging to record in conventional cost models based on simplified presumptions about data spreads and system operations.

3.8 Multi-Objective Optimization Formulation

The framework implements a formal multi-objective optimization model that balances competing goals of energy efficiency, carbon reduction, and performance maintenance. The objective function minimizes a weighted combination: **$\min(\alpha E + \beta C + \gamma L)$**, where E represents total energy consumption in joules, C denotes carbon emissions in grams of CO_2 equivalent, and L captures query latency in milliseconds. The trade-off weights α, β, and γ are dynamically adjusted based on current system priorities, regulatory requirements, and operational constraints. The optimization is subject to resource capacity constraints (CPU utilization $\leq 95\%$, memory allocation $\leq$ available capacity), service level agreement bounds (P95 latency $\leq$ SLA threshold), and data locality constraints (minimize cross-rack data transfers). This formulation enables explicit reasoning about trade-offs and permits organizations to configure optimization behavior according to their specific environmental commitments and performance requirements.

3.9 Combination with Big Data Frameworks

The structure is created for smooth combination with major Big Data processing frameworks, incorporating Apache Spark, Hadoop MapReduce, and cloud-native analytics systems through clearly outlined interfaces and extension locations. For Spark positions, the structure combines with the cluster handler to affect resource distribution choices and cooperates with the organizer to direct task positioning and execution sequence. In Hadoop settings, combination happens through custom organizers and resource handlers executing the structure's optimization policies while sustaining compatibility with current MapReduce applications. For cloud-native systems, the structure uses container organization platforms and serverless computing interfaces to manage resource provisioning and workload positioning. This multi-platform combination methodology confirms wide usability across varied organizational settings while permitting the structure to use platform-particular optimization chances [4].

4. EXECUTION AND TESTING PROCEDURES

4.1 Testing Configuration and Experimental Platform Setup

The testing assessment of the suggested Green AI structure was performed using a thorough experimental platform precisely representing the qualities and magnitude of production Big Data settings in contemporary cloud data centers. The experimental platform consisted of a heterogeneous cluster containing 128 compute nodes with varying server configurations to mirror diversity found in large-scale deployments. The cluster incorporated 64 nodes equipped with Intel Xeon E5-2650 v4 processors operating at 2.20 GHz (12 cores per processor, 24 cores per node) with 512 GB DDR4 memory per node, 32 nodes with newer generation processors providing 32 cores and 768 GB memory, and 32 nodes with legacy configurations containing 16 cores and 256 GB memory. Storage infrastructure included both RAID-configured arrays for high-performance requirements and JBOD configurations for capacity-optimized workloads, totaling 6.4 PB raw storage capacity across the cluster. Power tracking infrastructure was positioned at both separate server and cluster levels using intelligent PDUs and per-node IPMI sensors to record detailed energy usage data with sub-second temporal clarity. The cluster was arranged with numerous Big Data processing frameworks incorporating Apache Spark 3.2, Hadoop MapReduce 3.3, and cloudnative analytics instruments, each positioned with production-quality arrangements and supporting infrastructure, including HDFS distributed file systems, Hive metastore operations, and YARN cluster handling elements.

4.2 Information Collection Qualities and Workload Characteristics

The testing assessment used varied information collections and workload characteristics representative of actual Big Data analytics situations across varied application areas. The experimental dataset encompassed 50 TB of production data including structured transaction records from financial systems, semi-structured application logs from web services, and unstructured text collections from social media sources. Workloads incorporated batch handling tasks including large-magnitude data conversions and combinations processing datasets ranging from 500 GB to 15 TB per job, interactive queries for business intelligence applications with response time requirements under 30 seconds, machine learning model development and operation executing gradient boosting and neural network training on datasets from 2 TB to 8 TB, streaming analytics handling persistent data supplies at ingestion rates from 100 MB/s to 2 GB/s, and graph handling algorithms functioning on network organizations containing billions of vertices and edges. Each workload class displays separate qualities in terms of computational strength, memory reach behaviors, data proximity demands, and responsiveness to processing delays. Artificial workload producers supplemented actual application records to permit managed tests investigating particular elements of system operations and optimization success.

4.3 Reference Systems and Evaluation Standards

To thoroughly assess the success of the suggested structure, tests compared performance against numerous reference systems representing present advanced methodologies to resource handling and energy optimization in Big Data platforms. The initial reference used standard arrangement policies delivered by Big Data frameworks themselves, which characteristically execute basic resource distribution and organization heuristics without explicit energy optimization. The secondary reference represented manually refined arrangements established by skilled system managers following recommended procedures for energy productivity, delivering a standard for the capability of human skill. Supplementary references incorporated current automated optimization systems from research publications handling particular elements of energy productivity, including workload combination or flexible voltage adjustment, but missing a thorough combination suggested in this structure [4].

4.4 Performance Indicators and Assessment Standards

The testing assessment used a thorough collection of performance indicators recording numerous aspects of system operations relevant to sustainable Big Data platforms. Energy usage was quantified in terms of complete joules used during workload processing, typical power consumption during varied operational stages, and energy productivity expressed as computations accomplished per energy unit used. Carbon impact was measured by combining quantified energy usage with time-changing carbon strength data for the electricity network providing the data center, expressed in grams of carbon dioxide matching releases. Resource employment indicators monitored the productivity of CPU, memory, storage, and network resource usage to recognize chances for additional optimization. Query delay and job finishing periods quantified performance consequences of optimization approaches to confirm that energy reductions did not appear at expenditure of unacceptable weakening in application responsiveness. Statistical significance was evaluated using paired t-tests with significance threshold α=0.05, and 95% confidence intervals were calculated for all primary metrics across repeated experimental trials [5].

Metric Category	Specific Measurement	Unit of Measure	Collection Method	Evaluation Frequency
Energy Consumption	Total energy per workload	Joules (J)	Power meters integration	Per workload execution
Power Draw	Average power during execution	Watts (W)	Real-time monitoring	Continuous (per second)
Energy Efficiency	Computations per watt	Operations/W	Derived calculation	Per workload type
Carbon Footprint	CO2 equivalent emissions	grams CO2e	Energy × carbon intensity	Per workload execution
CPU Utilization	Processor usage percentage	Percentage (%)	System monitoring tools	Every 10 seconds
Memory Utilization	RAM usage efficiency	Percentage (%)	System monitoring tools	Every 10 seconds
Query Latency	Response time	Milliseconds/Seconds	Application-level logging	Per query
Job Completion Time	Total execution duration	Minutes/Hours	Framework logging	Per job

Table 4: Performance Metrics and Measurement Methodology [5, 6]

4.5 Machine Learning Model Development and Confirmation

The machine learning models incorporated within the structure experienced thorough development and confirmation procedures to confirm precision and dependability for production positioning. Development data was gathered from prolonged observation intervals spanning 90 days of varied workloads processing on the experimental platform, recording connections between workload qualities, resource distributions, and resulting performance and energy usage results. The information collection contained 47,832 completed job executions across all workload categories, totaling 2.3 million individual task measurements. The information collection was separated into development (70%), confirmation (15%), and examination (15%) collections following temporal divisions rather than arbitrary selection to confirm models were assessed on future forecasting tasks rather than simply calculating within development spread. Hyperparameter optimization used systematic grid search approaches combined with Bayesian optimization to recognize model arrangements that maximized predictive precision while preventing excessive-fitting to particular workload behaviors. Cross-confirmation procedures assessed model strength across varied data subgroups and workload classes. Model re-development procedures were created to confirm persistent precision as workload behaviors changed over time, with automated retraining triggered when prediction accuracy degraded beyond 12% RMSE threshold.

4.6 Replication and Actual Position Situations

The assessment procedure joined replication-based tests with actual position situations to thoroughly evaluate structure capabilities under managed situations and practical limitations. Replication tests permitted investigation of system operations under a wider extent of situations than achievable in physical experimental platforms, incorporating extreme workload situations with utilization spikes reaching 95% cluster capacity, varied data center arrangements with node counts from 50 to 500, and breakdown situations including server failures and network partitions. The replication setting incorporated confirmed models of hardware energy usage calibrated against measurements from physical infrastructure, network delay characteristics derived from production network traces, and storage performance models validated against benchmark results from physical storage arrays. Actual position situations examined the structure under production settings across three organizational environments including a financial services firm, an e-commerce platform, and a scientific research institution, incorporating changing workload arrivals following actual usage patterns, rival applications sharing infrastructure under multi-tenant arrangements, operational limitations on system alterations reflecting production change management policies, and combination with current tracking and handling instruments including Prometheus, Grafana, and commercial data center management systems [5].

4.7 Assessment Guidelines and Testing Organization

The testing guidelines were thoroughly created to separate the consequences of particular structural elements and optimization approaches while also assessing their combined success as a combined system. Elimination investigations systematically disabled separate structure elements including workload forecasting, carbon-aware scheduling, data placement optimization, and query optimization to measure their separate contributions to overall energy reductions and performance qualities, revealing that the integrated approach achieved 38% energy reduction compared to 18-24% for individual components in isolation. Responsiveness examinations investigated how structural performance changed with varied setting configurations including different α, β, γ weights in the optimization function, workload qualities spanning CPU-intensive to memory-intensive to I/O-intensive patterns, and environmental situations including varying carbon intensity profiles and renewable energy availability. Extended-duration tests assessed structure operations over prolonged intervals spanning 60-day continuous operation periods, incorporating numerous workload patterns and changing situations to assess consistency and flexibility. Statistical meaning examination using paired t-tests confirmed that the noticed variations between the suggested structure and reference systems represented authentic enhancements with p-values < 0.001 for energy reduction and p-values < 0.005 for carbon reduction metrics. Testing situations were systematically recorded, and tests were created for duplication, permitting confirmation by other researchers and specialists.

5. OUTCOMES, EXAMINATION, AND INTERPRETATION

5.1 Measured Energy Reductions and Carbon Decreases

The testing assessment showed considerable energy reductions and carbon impact accomplished by the suggested Green AI structure compared to reference methodologies across varied workload situations and system arrangements. The structure reliably decreased complete energy usage by 38% (95% CI: 35.2%-40.8%, p<0.001) compared to baseline configurations while sustaining workload finishing periods within satisfactory limits specified by performance demands, with average latency increase of only 1.5% (95% CI: 0.8%-2.2%). Energy productivity enhancements were especially noticeable for workloads with temporal adaptability that could gain from a carbonconscious organization, achieving 42% energy reduction for batch processing workloads, and workloads with meaningful data transfer that gained from refined data positioning, demonstrating 35% energy savings for distributed join operations. The extent of carbon impact decrease reached 44% (95% CI: 41.3%-46.7%, p<0.001) surpassing basic energy reductions in numerous situations because the structure's carbon-conscious organization focused workload processing during intervals when the electricity network incorporated elevated ratios of renewable energy sources, particularly during nighttime hours when wind power contribution exceeded 60% of grid capacity. Fluctuation in outcomes across varied workload categories delivered insight into which application qualities most powerfully affected optimization success and assisted in recognizing chances for additional structure improvement [6].

5.2 Evaluation Examination with Conventional Methodologies

Thorough evaluation with conventional resource handling methodologies disclosed that the suggested structure accomplished better energy productivity through numerous supplementary procedures functioning in agreement rather than depending on any single optimization procedure. The quantitative comparison across all experimental workloads demonstrated distinct performance tiers among different approaches. The standard arrangements used by typical Big Data platforms reliably displayed elevated energy usage owing to cautious resource over-provisioning and the absence of organization between resource distribution, organization, and data positioning choices. Manually refined arrangements created by skilled managers accomplished moderate 18% energy reductions (95% CI: 15.7%-20.3%) and 15% carbon reductions (95% CI: 12.8%-17.2%) through thorough setting adjustment with acceptable 2% latency increase, but could not equal structure performance because manual methodologies lack the capability to flexibly adjust to changing situations and cannot successfully organize complex multi-aspect optimization choices. Current automated optimization systems from publications characteristically handled only particular elements of energy productivity, with carbon-aware scheduling alone achieving 22% energy reduction, ML-based resource allocation yielding 24% energy savings, and Spark-specific optimization delivering 19% energy improvement, but missing the thorough combination suggested in this research which achieved 38% energy reduction through synergistic optimization across all framework components.

5.3 Performance Compared to Energy Productivity Compromises

Examination of the connection between energy productivity and performance disclosed detailed compromises changing meaningfully depending on workload qualities and optimization intensity. For numerous workloads, the structure accomplished considerable energy reductions with unimportant or even beneficial consequences on performance by removing wasteful resource over-provisioning and enhancing resource employment. However, intense optimization approaches emphasizing maximum energy decrease sometimes produced reasonable rises in workload finishing duration, emphasizing the significance of adjustable optimization goals that match organizational concerns. The structure's capability to explicitly represent and honor performance limitations through the multi-objective optimization formulation ($\min(\alpha E + \beta C + \gamma L)$) permitted detailed management over performance-energy compromises, permitting users to indicate maximum satisfactory performance weakening in return for greater energy reductions. Experimental validation demonstrated that adjusting weights from performance-prioritized ($\alpha=0.3$, $\beta=0.3$, $\gamma=0.4$) to energy-prioritized ($\alpha=0.5$, $\beta=0.4$, $\gamma=0.1$) configurations increased energy savings from 28% to 45% while latency increased from 0.5% to 3.8%. Interactive workloads with rigid delay demands gained less from optimization than batch handling jobs with temporal adaptability, proposing that workload-particular optimization approaches modified to varied application demands could additionally strengthen structure success.

5.4 Resource Employment and Expenditure Productivity Consequences

The structure meaningfully enhanced resource employment across computational, memory, storage, and networking aspects by more precisely matching resource distributions to genuine workload demands and removing inactive capabilities through flexible combination. Average CPU utilization increased from 42% in baseline configurations to 73% with the framework ($p<0.001$), while memory utilization improved from 38% to 68% ($p<0.001$), demonstrating substantially more efficient resource usage. Elevated resource employment converted straightforwardly into expenditure productivity enhancements for cloud-based positions, where organizations compensate for distributed capabilities regardless of actual usage. Energy expenditure reductions signified another significant aspect of expenditure productivity, with decreases in electricity usage and planned load moving to reduced-expenditure duration intervals, both adding to operational cost decrease estimated at 32% reduction in combined infrastructure and energy costs. The joining of enhanced resource employment and decreased energy usage permitted organizations to handle matching analytical workloads using 35% fewer physical servers, potentially delaying or preventing expensive infrastructure growth. For settings with established infrastructure, the enhanced productivity permitted greater analytical processing from current capabilities, raising return on infrastructure commitments.

5.5 Detailed Studies and Actual Usability

Thorough, detailed studies showed practical usability of the structure across varied organizational settings and application situations. One detailed study investigated a financial operations organization running nightly batch analytics jobs for threat evaluation and regulatory documentation, showing how carbon-conscious organizations moved these delay-accepting workloads to nighttime hours when electricity carbon strength was characteristically reduced owing to decreased demand and greater wind power accessibility, achieving 47% carbon reduction for batch workloads while maintaining completion before business hours. Another detailed study investigated an ecommerce platform accomplishing real-time recommendation engine calculations, displaying how intelligent resource distribution decreased energy usage by 29% during traffic maximums while sustaining sub-200ms P95 response periods demanded for satisfactory user experience. A third detailed study examined a scientific research organization handling substantial information collections from experimental instruments, showing how refined data positioning decreased data transfer by 63% and related energy usage by 41% for workflows that repeatedly reached matching information collections across multi-stage analysis pipelines.

5.6 Magnitude Evaluation Across Workload Strengths

Magnitude tests assessed structure performance across a wide extent of cluster magnitudes, workload quantities, and strength degrees to evaluate usability to both small-magnitude and hyperscale positions. The structure sustained optimization success as cluster magnitude rose from 32 nodes to 256 nodes in simulation environments, with only reasonable rises in optimization expenses that stayed small compared to overall workload processing duration, specifically framework overhead remaining under 2.3% of total execution time across all cluster sizes. The machine learning models showed satisfactory application to workload strengths beyond those noticed during development, successfully forecasting resource demands and energy usage for load degrees reaching 150% of training distribution maximum with prediction accuracy degrading gracefully from 8.3% RMSE to 11.7% RMSE. However, extreme workload situations with unexpected requirement increases exceeding 200% of historical maximum sometimes tested the structure's reactive capabilities, emphasizing the need for incorporating more advanced irregularity recognition and contingency preparation capabilities that could arrange emergency reactions for extraordinary situations before they happen.

5.7 Restrictions and Obstacles

Despite encouraging outcomes, several restrictions and obstacles appeared during the assessment that deserve recognition and propose directions for future research. The structure's optimization success relied fundamentally on the precision of workload forecasting models, and forecasting mistakes occasionally produced less-than-optimal resource distribution choices that either wasted energy through excessive-provisioning or weakened performance through inadequate-provisioning, with worst-case scenarios showing 8% performance degradation when predictions underestimated demand by more than 40%. The complex connections between varied optimization approaches sometimes created surprising developing operations that were challenging to forecast or identify, proposing requirements for more advanced organization procedures and better visibility into optimization decision-making processes. Combination with current data center handling instruments and workflows demanded thorough preparation and organization, with organizational adjustment handling elements sometimes offering greater barriers than technical execution obstacles. The reinforcement learning scheduler required substantial training periods (approximately 15-20 days) to achieve stable performance, limiting rapid deployment scenarios.

5.8 Practical Consequences for Large-Magnitude Data Facilities

The research discoveries hold meaningful practical consequences for organizations operating large-scale cloud data centers and attempting to decrease their environmental impact while sustaining competitive service standards. The shown energy reductions of 38% and carbon impact decreases of 44% straightforwardly convert into decreased operational expenditures estimated at 32% cost reduction and advancement toward corporate preservation goals that progressively affect organizational standing and regulatory agreement, particularly supporting ISO 14001 environmental management system certification and ESG reporting requirements for Scope 2 emissions disclosure. The structure's capability to sustain performance within 1.5% latency increase while optimizing energy usage addresses a central worry that has historically prevented acceptance of energy productivity actions in production settings where service standards cannot be weakened. The sectioned design and combination capabilities promote gradual acceptance, permitting organizations to execute particular structure elements, handling their most urgent worries rather than demanding complete replacement of the current infrastructure. The application of machine learning permits persistent enhancement as the system learns from operational experience, delivering a route toward progressively advanced optimization as the position develops.

Conclusion

This article has created and assessed a thorough Green AI structure for sustainable Big Data platforms that successfully shows how intelligent optimization can considerably decrease energy usage and carbon impact in large-scale cloud data centers without weakening analytical performance. The structure combines machine learning models for workload forecasting achieving 8.3% RMSE for CPU predictions, intelligent resource distribution algorithms improving utilization from 42% to 73%, carbon-conscious organization procedures using reinforcement learning with Proximal Policy Optimization, and refined data positioning approaches reducing data transfer energy by 41% into a unified design organizing numerous optimization methodologies across varied system stratums. Testing assessment across varied workload situations and Big Data frameworks including a 128-node heterogeneous cluster processing 50 TB datasets verified meaningful enhancements in energy productivity showing 38% energy reduction (95% CI: 35.2%-40.8%, $p<0.001$), resource employment improvements of 31 percentage points, and environmental preservation demonstrating 44% carbon reduction (95% CI: 41.3%-46.7%, $p<0.001$) compared to conventional methodologies while maintaining performance within 1.5% latency increase.

The research accomplishes its central goals of creating practical procedures for decreasing the environmental consequences of Big Data analytics infrastructure while sustaining performance qualities demanded for production positioning. These discoveries add to the growing collection of knowledge on environmentally sustainable computing and show that AI-powered optimization can permit organizations to balance their computational requirements with environmental accountability. The structure's sectioned design and demonstrated success in actual situations deliver a basis for organizations attempting to execute sustainable Big Data platforms while meeting ISO 14001 environmental management standards and ESG reporting obligations for Scope 2 carbon emissions disclosure.

Deployment Roadmap and Implementation Strategy

Organizations seeking to implement the Green AI framework should follow a phased deployment approach. Phase 1 (Months 1-3) involves establishing comprehensive monitoring infrastructure for energy consumption and carbon intensity tracking, requiring integration with existing data center management systems and power distribution units. Phase 2 (Months 4-6) focuses on deploying workload prediction models, beginning with LSTM training on historical data and establishing feedback mechanisms for continuous model improvement. Phase 3 (Months 7-9) implements carbon-aware scheduling policies, initially for delay-tolerant batch workloads while interactive workloads remain on traditional schedulers. Phase 4 (Months 10-12) integrates data placement optimization and query refinement components, completing the full framework deployment. Throughout deployment, organizations should maintain parallel operation with existing systems, conducting A/B testing to validate improvements before full production cutover. Post-deployment, continuous monitoring and quarterly model retraining ensure sustained optimization performance as workload patterns evolve.

Future research directions incorporate extending the structure to include supplementary optimization aspects, including cooling system productivity optimization which could yield additional 8-12% facility-level energy savings, investigating federated learning methodologies that permit cross-organizational distribution of optimization understanding while maintaining data confidentiality and regulatory compliance, and examining how quantum computing advances might permit more advanced optimization algorithms capable of solving larger-scale multiobjective problems with reduced computational overhead. As environmental worries persist to influence technology creation concerns, the combination of preservation factors into central system organization signifies not simply a choice but a requirement for accountable advancement. The Green AI structure offered in this research presents a practical route toward constructing next-generation cloud platforms that supply powerful analytical capabilities while reducing their environmental impact, adding to a more sustainable digital future and supporting global climate commitments through measurable reductions in data center carbon emissions. **References**

[1] Zhiwei Cao, et al., "Data Center Sustainability: Revisits and Outlooks," IEEE Access, 31 May 2023. Available: https://ieeexplore.ieee.org/document/10139829

[2] Sara DIOUANI; Hicham MEDROMI, "How Energy Consumption in the Cloud Data Center is Calculated," 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), 22 August 2019. Available: https://ieeexplore.ieee.org/document/8807458

[3] Jiwei Huang, et al., "Energy Efficient Speed Scaling and Task Scheduling for Distributed Computing Systems," IEEE Transactions on Parallel and Distributed Systems, July 2015. Available: https://ieeexplore.ieee.org/document/10855295

[4] Marco Lattuada, et al., "Optimal Resource Allocation of Cloud-Based Spark Applications," IEEE Transactions on Cloud Computing, 06 April 2020. Available: https://ieeexplore.ieee.org/document/9057697

[5] Ioannis Chrysakis, et al., "Multi-Partner Project: Green.Dat.AI: A Data Spaces Architecture for Sustainable AI," 2025 Design, Automation & Test in Europe Conference (DATE), 21 May 2025. Available: https://ieeexplore.ieee.org/abstract/document/10992729

[6] Adeel Ahmed, et al., "An Efficient Task Scheduling for Cloud Computing Platforms Using Energy Management Algorithm: A Comparative Analysis of Workflow Execution Time," IEEE Transactions on Sustainable Computing, 29 February 2024. Available: https://ieeexplore.ieee.org/document/10453553

[7] PP. Singh, R. Sharma, "Carbon Footprint Analysis: Need for Green Cloud Computing," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 13 April 2022. Available: https://ieeexplore.ieee.org/document/9752341

[8] Deepika Saxena, et al., "Performance Analysis of Machine Learning Centered Workload Prediction Models for Cloud," IEEE Transactions on Parallel and Distributed Systems, 30 January 2023. Available: https://ieeexplore.ieee.org/document/10029931

[9] Saumya Dash, et al., "Green AI: Enhancing Sustainability and Energy Efficiency in AI Systems," IEEE Transactions on Artificial Intelligence, 22 January 2025. Available: https://ieeexplore.ieee.org/abstract/document/10849555