International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Cloud data centers and networks: Applications and optimization techniques

Pierre Subeh [1] and Bushara AR [2,*]

[1] *Marketing Programs Advisory Committee. Full Sail University, USA.*
[2] *Department of ECE, KMEA Engineering College, APJ Abdul Kalam Technological University, India.*

International Journal of Science and Research Archive, 2024, 13(02), 218–226

Publication history: Received 23 September 2024; revised on 31 October 2024; accepted on 02 November 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2100

Abstract

As the volume and complexity of big data continue to escalate, optimizing the performance, scalability, and energy efficiency of big data applications within cloud data centers has become increasingly crucial. This journal presents a comprehensive survey of current optimization techniques, focusing on data placement, job scheduling, and network configurations tailored for cloud environments. We explore the impact of various data center topologies on the performance of big data frameworks like Hadoop, emphasizing the trade-offs between performance and energy efficiency. Advanced methodologies, including dynamic data placement strategies, locality-aware scheduling, and innovative reduce task placement techniques, are reviewed in depth. Additionally, we highlight the importance of network power effectiveness (NPE) and examine the role of optical and electronic switching technologies in enhancing data center efficiency. By synthesizing findings from recent studies, this paper provides valuable insights into the optimization of cloud data centers, offering recommendations for improving resource utilization and reducing job completion times while maintaining energy efficiency. The findings contribute to the ongoing efforts to scale and adapt cloud data infrastructures for the rapidly growing demands of big data applications.

Keywords: Cloud Data Centers; Big Data Optimization; Hadoop Clusters; Job Scheduling Algorithms; Network Topology Efficiency

1. Introduction

The big data landscape and the infrastructure for collecting, transferring and storing information is changing faster than ever [1]. As opposed to a traditional dataset, big data is unstructured and wherein produced rapidly for extended periods of time through user or device proliferation [2] & [3]. Most of these large workloads are taken care of in well-equipped cloud data centers, which enable huge storage and computational power to address for real-time as well as batch processing [4] & [5]. Big data processing aims to process big and complex volumes of (structured/un- structured) queries based on the five V-model [6 7] comprising Volume, Velocity, Variety, Veracity and Value. Volume: refers to the enormous data sizes, which are usually quantified in exabytes or zettabytes [8] is shown in FIgure 1.

Velocity indicates at which speed data is 'produced' and hence underlines that it loses its value quickly [9]. Variety deals with the many types of data, from structured [10]. Veracity whether the data is accurate and can be trusted or not Value [11] - [13]. Insights,inference that one can derive from the data. A global data volume of around 40,000 exabytes is expected to be recorded in the year 2020 and this gigantic amount will get accumulated as a progression across many applications generating large content attached with human activities [14] - [15]. It was expected that online business transactions would reach 450 billion per day in the year of 2020, and as social media platforms further enrich internet content by contributing several terabytes on a daily basis. Big Data mining is a key to understanding how users behave and what they want, enabling the next generation of innovations in psychology, economics & product development [16] - [18].

* Corresponding author: Bushara AR https://orcid.org/0009-0003-4504-2802

Copyright © 2024 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

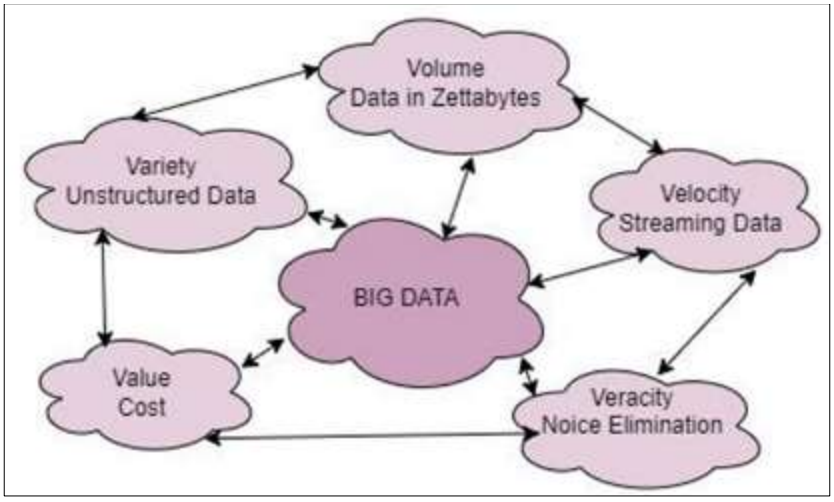

Figure 1 Representation of 5 V models in Big Data

To deal with the challenges of processing large datasets, a variety of technologies such as MapReduce has developed that delivers an efficient parallel computing model for distributed data. This has given rise to a variety of distributed programming solutions and cloud computing applications [19]. Large data applications involve almost every field, including earth sciences, astronomy where it combines with both archive and simulation/data-intensive scientific areas such as genomics or bio-informatics. Historically, such applications were supported by high-performance computing (HPC) clusters But cloud computing infrastructures are slowly but surely replacing existing systems in the face of datasets that have long surpassed what old technology can handle, sacrificing performance and cost along the way [20]. There is also Mobile Cloud Computing (MCC), where heavy computational and storage requirement applications can offload their processing tasks to the cloud. The approach allows data-rich services to be streamed in high video quality or played online games on a low-specced device. In addition, the internet of things (IoT) interconnects a large number of physical devices that help in uniting data collection and analysis within applications [21].

Realising the promise of big data needs joint efforts from many disciplines with heavy investment in improving infrastructures for data processing, management and analytics. This will require security, privacy and governance as big challenges when building a large scale- energy efficiency processing system. In this journal we present a survey of optimization strategies for cloud data centres and networks, including application-level, cloud networking-level as well as data centre level approaches to enhance the performance while improving resource utilisation and energy efficiency on big data applications.

1.1. Data Placement Optimization in Hadoop Clusters

Early research in optimizing data placement within Homogeneous Hadoop clusters focused on developing dynamic algorithms to distribute data fragments according to the processing capacities of individual nodes [22]. This strategy was designed to minimize data movement between over-utilized and under-utilized servers, which can enhance the overall efficiency of the cluster.

However, a significant limitation of this approach was the lack of consideration for data replication. Replication is a critical aspect of ensuring high data availability across the cluster, but it also introduces energy inefficiency as it requires most nodes to remain active. Later studies addressed this issue by proposing methods to dynamically switch off unused nodes while ensuring that replicas of all data sets are maintained in a subset of nodes that are always active. While this approach effectively reduced energy consumption, it also resulted in increased job running times due to the reduced number of active nodes available for processing.

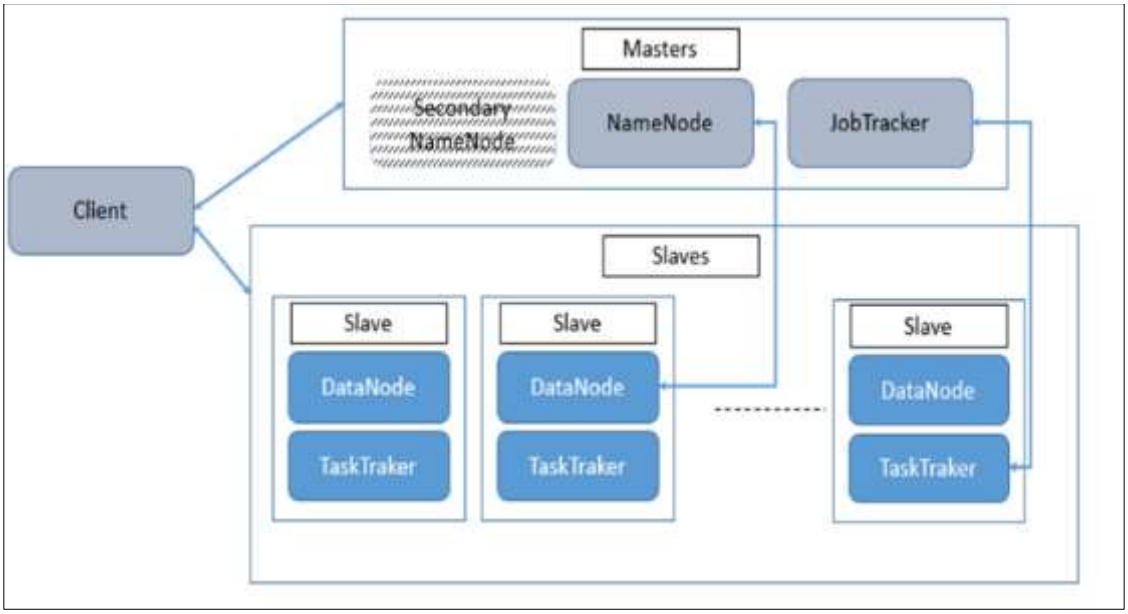

Figure 2 Hadoop Architecture [23]

1.2. Balancing Performance and Energy Efficiency

Optimizing the balance between performance and energy efficiency in dedicated MapReduce clusters has been a major focus of subsequent research. Dynamic sizing and locality-aware scheduling techniques were developed to handle diverse workloads, including batch MapReduce jobs, web applications, and interactive MapReduce jobs. By delaying batch workloads and carefully considering data locality and delay constraints for web and interactive workloads, these techniques achieved a balance between energy savings and job performance. For instance, energy savings of up to 59% were reported compared to static resource allocation strategies. Moreover, specialized algorithms like the Dependency Aware Locality for MapReduce (DALM) were introduced to process highly skewed and dependent input data. These algorithms significantly reduced cross-server traffic, with DALM achieving a reduction of up to 50%, thereby enhancing overall cluster performance [24] - [27].

1.3. Advanced Reduce Task Placement Techniques

The placement of reduced tasks within Hadoop clusters has been another area of significant research. Traditional greedy approaches focused on maximizing intermediate data locality for current reduced tasks but often at the cost of performance in subsequent map tasks. To address this, more sophisticated methods were developed, including approaches that optimized reduce task placement while considering the impact on future map tasks. One such study achieved up to a 20% improvement in performance by balancing the needs of sequential MapReduce jobs. Additionally, efforts to improve query performance in Hadoop clusters led to the development of data block allocation strategies that utilized k-means clustering. This method co-located related data blocks within the same cluster, improving query performance by up to 25% and reducing overall job execution time [28]-[31].

1.4. Innovations in Job Scheduling

Common default scheduling engines for hadoop like FIFO, capacity scheduler and HFS have not been sophisticated enough in dealing with mixed workloads, leading to suboptimal performance. A number of innovative scheduling algorithms spawned from research in this area has been developed to make the optimal use of resources, minimize job makespan and maximize energy efficiency. To alleviate the conflict between scheduling fairness and data locality, delay sched-uling was proposed. This reduced the trade-off with scheduling-to-maximize-data-locality and subsequently had high success rates in achieving perfect data locality while placing very few restrictions on overall job performance. Other scheduling innovations included. Resource-aware Adaptive Scheduling that dynamically changes device slots per machine based on job requirements and FLEX, which is a flexible allocation scheme for optimizing response time while meeting Service Level Agreements [32] - [35].

1.5. Reducing Job Completion Times in Cloud Environments

Reducing job completion time is a fundamental objective in designing cloud data centers for performance, and to meet Service Level Agreements (SLAs) in addition to minimizing power consumption. Many strategies have been derived, such as the Resource and Deadline-aware Hadoop scheduler (RDS), which uses online optimization to meet job deadlines on dynamically allocating resources in Hadoop clusters with self-learning completion time estimator. RDS has been very well received in environments where a combination of energy source and workload need to synchronously deliver deadline constrained jobs. Going a step further, the Bipartite Graph MapReduce scheduler (BGMRS), which specifically targeted clusters with heterogeneous nodes and fluctuating job execution durations, demonstrated an aggregated reduction of 79% in deadline miss ratios and 36% in total job completion times as compared to existing dynamic mapreduce schedulers.

Other approaches like BEneath the Task Level (BeTL) have been proposed to tackle fault tolerance and skew mitigation respectively, achieving more on further improving Hadoop performance under failure conditions and skewed workloads. These approaches have demonstrated impressive reductions in job execution time; reducing up to 88.79% of the total runtime in memory-constrained environments [36] - [39].

2. Literature Survey

Authors	Methodology/Techniques	Key Points
Hedayati et al. [40]	Systematic review of Hadoop MapReduce scheduling algorithms	Categorized schedulers based on performance, proposed frameworks to enhance data locality, resource utilization, and fairness.
Ahmed et al. [41]	Comparative performance analysis of Hadoop and Spark using HiBench	Spark outperformed Hadoop in memory-intensive tasks, while Hadoop was more efficient in disk I/O operations; benchmarked different workloads.
Dokeroglu et al. [42]	Optimizing Hadoop Hive performance by sharing scan and computation tasks	Introduced query sharing to reduce redundant tasks, leading to performance improvements in Hadoop Hive.
Lee et al. [43]	Dynamic data placement strategy for Hadoop in heterogeneous environments	Focused on load balancing and reducing data movement; proposed dynamic placement to match node capacities.
Xiong et al. [44]	Optimizing data placement in heterogeneous Hadoop clusters	Suggested energy-efficient strategies for data placement to improve MapReduce performance and reduce cross-server traffic.
Wu et al. [45]	Energy-efficient Hadoop for big data analytics	Proposed methods to reduce energy consumption in Hadoop clusters while maintaining high data availability.
Ma et al. [46]	Dependency Aware Locality for MapReduce (DALM)	Reduced cross-server traffic by considering data dependencies during task scheduling in Hadoop.
Zhao et al. [47]	Improved data placement strategy in a heterogeneous Hadoop cluster	Proposed a strategy to enhance fault tolerance and optimize data block placement to improve processing efficiency.
Dokeroglu et al., [48]	SharedHive: Optimizing Hadoop Hive performance through query sharing	Improved performance by merging queries with similar operations; reduced MapReduce tasks and HDFS I/O.
Mashayekhy et al., [49]	Energy-Aware Scheduling of Hadoop MapReduce Jobs	Focused on scheduling techniques that reduce energy consumption while maintaining performance in big data applications.
Samadi et al. [50]	Comparative analysis of Hadoop and Spark using HiBench benchmarks	Benchmarked and compared Hadoop and Spark across various tasks, highlighting strengths and weaknesses in different workloads.
Lee et al. [51]	Data placement strategy for Hadoop in heterogeneous environments	Developed dynamic algorithms for optimizing data placement based on node capacities, reducing data movement costs.

The reviewed literature reflects significant advancements in the optimization of Hadoop clusters, particularly in the areas of data placement, scheduling algorithms, and performance enhancement techniques. Initial research efforts, such as those by Lee et al. focused on dynamic data placement strategies tailored for heterogeneous Hadoop environments. These methods aimed to optimize load balancing and minimize data movement by aligning data distribution with the processing capacities of individual nodes. However, the heterogeneous nature of these environments posed challenges in maintaining consistent performance levels. To mitigate these issues, Wu et al. introduced energy-efficient data placement strategies that not only reduced energy consumption but also ensured high data availability across the cluster, thereby addressing the trade-off between energy efficiency and data redundancy.

In the realm of scheduling, significant advancements have been made to enhance resource utilization and job performance. Zhao et al. provided a systematic review of Hadoop MapReduce scheduling algorithms, categorizing them based on their impact on performance metrics such as data locality, resource utilization, and fairness. Additionally, query optimization techniques, such as those explored by Dokeroglu et al. focused on enhancing Hadoop Hive performance through shared computation tasks, effectively reducing redundant MapReduce operations. These technological developments collectively contribute to the ongoing efforts to improve the scalability, efficiency, and adaptability of Hadoop clusters in managing complex and diverse data processing workloads. The optimization of big data applications within data centers has become increasingly critical as the volume and complexity of data continue to grow. This section summarizes various studies focused on enhancing the performance, scalability, flexibility, and energy efficiency of big data applications, specifically in relation to their hosting data centers. The studies are categorized based on their approach to improving data center design, routing protocols, and job scheduling to optimize the overall performance of big data infrastructures.

2.1. Performance and Energy Efficiency in Data Center Topologies

Several studies have evaluated the impact of different data center topologies on the performance and energy efficiency of big data applications. Data center Network Topology is shown in Figure 3. The authors in [52]-[53] conducted comprehensive analyses using various models and simulators to assess how network configurations influence Hadoop clusters. Modeled Hadoop clusters with a topology of up to four ToR switches and a core switch, finding that Hadoop scales effectively across nine different configurations. Similarly, [205] used the MRPerf simulator to study how data center topology affects Hadoop's performance, considering factors like CPU, RAM, disk resources, and data placement. A significant finding was that the DCell topology improved sorting performance by 99% compared to double-rack clusters. Further studies [54]-[55] extended this work by examining different data center topologies and switching technologies. The CloudSimExMapReduce simulator in [55] compared several topologies, with CamCube showing superior performance, particularly in handling intermediate data shuffling. The study in [56] highlighted that optical switching technologies could reduce power consumption by 54% compared to electronic switches, while maintaining comparable performance.

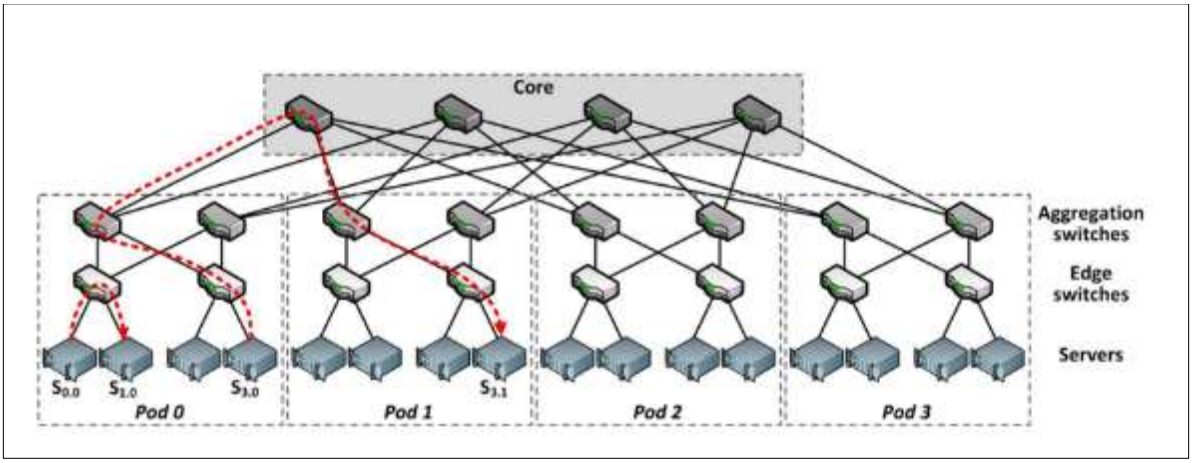

Figure 3 Datacenter Network Topology [58]

Additionally, research in [57] introduced the concept of Network Power Effectiveness (NPE), evaluating different electronic switching topologies under various routing protocols. The study found that topologies like FBFLY achieved the highest NPE, particularly when considering server-centric designs.

2.2. Enhancing Big Data Applications in Data Centers

As big data grows even bigger, and more complex in its use of the infrastructure beneath it, maintaining performance optimization to applications running at data center scale is critical. Introduction This section gives an introductory documentary on a few important original research and surveys related to optimising the data centre design, routing protocols as well as job scheduling for better performance of big data infrastructures.

2.3. Evaluating Data Center Topologies for Optimal Performance and Efficiency

Several studies investigated how different topologies of data centers have an influence on performance and energy efficiency in dealing with big data applications. Several studies [59]–[60], employed various models and simulators to evaluate the performance implications of network configurations in Hadoop clusters. A Hadoop clusters of up to four ToR switches and a core switch came across the fact that recurrence networks scaled well in case they were different through a sum of nine configurations. [61] also used the MRPerf simulator but they consider a slightly larger model and focus on how data center topology affects performance of Hadoop, including CPU, RAM disk resources as well is concerning with which nodes will be storing chunks or files. This brings us to the DCell topology, which improved sort performance 99.4% over double-rack clusters. Authors in [62] use Cloud SimEx MapReduce simulator to show comparison of various topologies and CamCube providing good results specifically for intermediate data shuffling jobs. Also the work in [63] pointed out that optical switching technologies could cut power consumption by 54%, compared with conventional electronic switches but provide equivalent levels of efficiency.

2.4. Network Power Effectiveness and Topology Optimization

Building on these analyses, research in [64] introduced the concept of Network Power Effectiveness (NPE), which evaluates different electronic switching topologies under varying routing protocols. The study identified that topologies like FBFLY achieved the highest NPE, especially when focusing on server-centric designs. It also examined the impact of design choices such as link speeds and oversubscription ratios on performance, revealing that specific configurations could significantly influence the energy efficiency and scalability of data center operations. These studies collectively contribute to a deeper understanding of how data center topology and network configurations can be optimized to enhance the performance and energy efficiency of big data applications

3. Conclusion

As big data continues to expand in both volume and complexity, optimizing cloud data centers has become increasingly critical. This paper has examined key optimization strategies, focusing on data placement, job scheduling, and network topology enhancements within cloud environments, particularly in Hadoop clusters. Dynamic data placement strategies have proven essential for balancing load distribution and minimizing data movement within heterogeneous clusters. However, these strategies face challenges in maintaining energy efficiency without compromising performance. Advanced job scheduling algorithms, such as locality-aware scheduling and dynamic resource allocation, have shown significant potential in reducing job completion times and improving resource utilization, particularly in environments with diverse workloads.

Innovative network topologies, including the integration of optical networking technologies, have been identified as crucial for achieving substantial energy savings and reducing latency in data center operations. Despite these advancements, several challenges persist. Balancing energy efficiency with high performance remains a key area of focus, and there is a need for more sophisticated algorithms capable of real-time adaptation to fluctuating workloads and data center conditions. Furthermore, the integration of next-generation technologies aimed at further reducing energy consumption without sacrificing computational performance is crucial. Security, privacy, and governance also need to be addressed in the ongoing evolution of cloud data center optimization strategies

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., ... & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. *The scientific world journal*, *2014*(1), 712826.

[2] Gupta, D., & Rani, R. (2019). A study of big data evolution and research challenges. *Journal of information science*, *45*(3), 322-340.

[3] Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., & Alor-Hernández, G. (2016). A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing*, *72*, 3073-3113.

[4] Ruiu, P., Scionti, A., Nider, J., & Rapoport, M. (2016, July). Workload management for power efficiency in heterogeneous data centers. In *2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)* (pp. 23-30). IEEE.

[5] Li, J., Bao, Z., & Li, Z. (2014). Modeling demand response capability by internet data centers processing batch computing jobs. *IEEE Transactions on Smart Grid*, *6*(2), 737-747.

[6] Keskar, V., Yadav, J., & Kumar, A. H. (2020). 5V's of Big Data Attributes and their Relevance and Importance across Domains. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN*, 2278-3075.

[7] Deng, D., Leung, C. K., Wodi, B. H., Yu, J., Zhang, H., & Cuzzocrea, A. (2018, July). An innovative framework for supporting cognitive-based big data analytics for frequent pattern mining. In *2018 IEEE International Conference on Cognitive Computing (ICCC)* (pp. 49-56). IEEE.

[8] Hammad, K. A. I., Fakharaldien, M. A. I., Zain, J., & Majid, M. (2015, September). Big data analysis and storage. In *International conference on operations excellence and service engineering* (pp. 10-11).

[9] Arrowsmith, S. J., Trugman, D. T., MacCarthy, J., Bergen, K. J., Lumley, D., & Magnani, M. B. (2022). Big data seismology. *Reviews of Geophysics*, *60*(2), e2021RG000769.

[10] Fang, B., & Zhang, P. (2016). Big data in finance. *Big data concepts, theories, and applications*, 391-412.

[11] Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible?. *Advances in Classification Research Online*, *24*(1), 4.

[12] Bushara A. R., Vinod Kumar R. S, T. Gopalakrishnan, S. Hari Kumar (2024. International Research Journal of Education and Technology. 6(5), pp. 369--375.

[13] Pendyala, V. (2018). Veracity of big data. *Machine Learning and Other Approaches to Verifying Truthfulness*.

[14] Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., & Alor-Hernández, G. (2016). A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing*, *72*, 3073-3113.

[15] babu Nuthalapati, S., & Nuthalapati, A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning. *International Journal of Science and Research Archive*, *12*(2), 408-422.

[16] Antons, D., & Breidbach, C. F. (2018). Big data, big insights? Advancing service innovation and design with machine learning. *Journal of Service Research*, *21*(1), 17-39.

[17] Nuthalapati, S. B. (2022). Transforming Agriculture with Deep Learning Approaches to Plant Health Monitoring. *Remittances Review*, *7*(1), 227-238.

[18] Mariani, M. M., & Wamba, S. F. (2020). Exploring how consumer goods companies innovate in the digital age: The role of big data analytics companies. *Journal of Business Research*, *121*, 338-352.

[19] Grolinger, K., Hayes, M., Higashino, W. A., L'Heureux, A., Allison, D. S., & Capretz, M. A. (2014, June). Challenges for mapreduce in big data. In *2014 IEEE world congress on services* (pp. 182-189). IEEE.

[20] Cafaro, M., & Aloisio, G. (2011). *Grids, clouds, and virtualization* (pp. 1-21). Springer London.

[21] Ali, Z. H., Ali, H. A., & Badawy, M. M. (2015). Internet of Things (IoT): definitions, challenges and recent research directions. *International Journal of Computer Applications*, *128*(1), 37-47.

[22] Jain, R., Saxena, A., & Manoriya, M. (2015). Analysis of Dynamic Data Placement Strategy for Heterogeneous Hadoop Cluster. *International Journal of Emerging Trends and Technology in Computer Science, Expected*, *4*.

[23] ELomari, A., Hassouni, L., & MAIZATE, A. (2021). New data placement strategy in the HADOOP framework. *International Journal of Advanced Computer Science and Applications*, *12*(7).

[24] Mohamed, S. H., El-Gorashi, T. E., & Elmirghani, J. M. (2019). A survey of big data machine learning applications optimization in cloud data centers and networks. *arXiv preprint arXiv:1910.00731*.

[25] babu Nuthalapati, S. (2023). AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking. *Educational Administration: Theory and Practice*, *29*(1), 357-368.

[26] AR, B., RS, V. K., & SS, K. (2023). LCD-capsule network for the detection and classification of lung cancer on computed tomography images. *Multimedia Tools and Applications*, *82*(24), 37573-37592.

[27] Jain, R., Saxena, A., & Manoriya, M. (2015). Analysis of Dynamic Data Placement Strategy for Heterogeneous Hadoop Cluster. *International Journal of Emerging Trends and Technology in Computer Science, Expected*, *4*.

[28] Durut, M. (2012). *Distributed clustering algorithms over a cloud computing platform* (Doctoral dissertation, Télécom ParisTech).

[29] Pedone, I., Canavese, D., & Lioy, A. (2020). Trusted computing technology and proposals for resolving cloud computing security problems. In *Cloud Computing Security* (pp. 373-386). CRC Press.

[30] Muhammed Kunju, A. K., Baskar, S., Zafar, S., & AR, B. (2024). A transformer based real-time photo captioning framework for visually impaired people with visual attention. *Multimedia Tools and Applications*, 1-20.

[31] Yang, H. C., Dasdan, A., Hsiao, R. L., & Parker, D. S. (2007, June). Map-reduce-merge: simplified relational data processing on large clusters. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 1029-1040).

[32] Assudani, P. J., & Abimannan, S. (2018). Power efficient and workload aware scheduling in cloud. *Int J Eng Technol*, *7*(17), 44-52.

[33] Bushara, A. R., RS, V. K., & Kumar, S. S. (2024). The Implications of Varying Batch-Size in the Classification of Patch-Based Lung Nodules Using Convolutional Neural Network Architecture on Computed Tomography Images. *Journal of Biomedical Photonics & Engineering*, *10*(1), 39-47.

[34] Nuthalapati, A. (2023). Smart Fraud Detection Leveraging Machine Learning For Credit Card Security. *Educational Administration: Theory and Practice*, *29*(2), 433-443.

[35] Bawankule, K. L., Dewang, R. K., & Singh, A. K. (2021). Historical data based approach for straggler avoidance in a heterogeneous Hadoop cluster. *Journal of Ambient Intelligence and Humanized Computing*, *12*(10), 9573-9589.

[36] Podlesny, N. J. (2023). *Quasi-identifier discovery to prevent privacy violating inferences in large high dimensional datasets* (Doctoral dissertation, Universität Potsdam).

[37] Jishamol, T. R., & Bushara, A. R (2016). Enhancement of Uplink Achievable Rate and Power Allocation in LTE-Advanced Network System. International Journal of Science Technology and Engineering (IJSTE). 211-217.

[38] Nuthalapati, A. (2022). Optimizing Lending Risk Analysis & Management with Machine Learning, Big Data, and Cloud Computing. *Remittances Review*, *7*(2), 172-184.

[39] Zhao, Y., MacKinnon, D. J., Gallup, S. P., Dakwala, N., Chick, T. A., Cagle, L., ... & Lipke, W. (2015). Data Mining and Measurements. *CrossTalk*.

[40] Hedayati, S., Maleki, N., Olsson, T., Ahlgren, F., Seyednezhad, M., & Berahmand, K. (2023). MapReduce scheduling algorithms in Hadoop: a systematic study. *Journal of Cloud Computing*, *12*(1), 143.

[41] Ahmed, N., Barczak, A. L., Susnjak, T., & Rashid, M. A. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench. *Journal of Big Data*, *7*(1), 110.

[42] Dokeroglu, T., Ozal, S., Bayir, M. A., Cinar, M. S., & Cosar, A. (2014). Improving the performance of Hadoop Hive by sharing scan and computation tasks. *Journal of Cloud Computing*, *3*, 1-11.

[43] Lee, C. W., Hsieh, K. Y., Hsieh, S. Y., & Hsiao, H. C. (2014). A dynamic data placement strategy for hadoop in heterogeneous environments. *Big Data Research*, *1*, 14-22.

[44] Xiong, R., Luo, J., & Dong, F. (2015). Optimizing data placement in heterogeneous Hadoop clusters. *Cluster Computing*, *18*, 1465-1480.

[45] Wu, W., Lin, W., Hsu, C. H., & He, L. (2018). Energy-efficient hadoop for big data analytics and computing: A systematic review and research insights. *Future Generation Computer Systems*, *86*, 1351-1367.

[46] Ma, X., Fan, X., Liu, J., & Li, D. (2015). Dependency-aware data locality for MapReduce. *IEEE Transactions on Cloud Computing*, *6*(3), 667-679.

[47] Zhao, W., Meng, L., Sun, J., Ding, Y., Zhao, H., & Wang, L. (2014). An improved data placement strategy in a heterogeneous hadoop cluster. *The Open Cybernetics & Systemics Journal*, *8*(1).

[48] Dokeroglu, T., Ozal, S., Bayir, M. A., Cinar, M. S., & Cosar, A. (2014). Improving the performance of Hadoop Hive by sharing scan and computation tasks. *Journal of Cloud Computing*, *3*, 1-11.

[49] Mashayekhy, L., Nejad, M. M., Grosu, D., Zhang, Q., & Shi, W. (2014). Energy-aware scheduling of mapreduce jobs for big data applications. *IEEE transactions on Parallel and distributed systems*, *26*(10), 2720-2733.

[50] Samadi, Y., Zbakh, M., & Tadonki, C. (2016, May). Comparative study between Hadoop and Spark based on Hibench benchmarks. In *2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech)* (pp. 267-275). IEEE.

[51] Lee, C. W., Hsieh, K. Y., Hsieh, S. Y., & Hsiao, H. C. (2014). A dynamic data placement strategy for hadoop in heterogeneous environments. *Big Data Research*, *1*, 14-22.

[52] Ahmed, N., Barczak, A. L., Susnjak, T., & Rashid, M. A. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench. *Journal of Big Data*, *7*(1), 110.

[53] Polato, I., Ré, R., Goldman, A., & Kon, F. (2014). A comprehensive view of Hadoop research—A systematic literature review. *Journal of Network and Computer Applications*, *46*, 1-25.

[54] Liu, Y., Muppala, J. K., Veeraraghavan, M., Lin, D., & Hamdi, M. (2013). *Data center networks: Topologies, architectures and fault-tolerance characteristics*. Springer Science & Business Media.

[55] Liu, Y., Muppala, J. K., Veeraraghavan, M., Lin, D., & Hamdi, M. (2013). *Data center networks: Topologies, architectures and fault-tolerance characteristics*. Springer Science & Business Media.

[56] Papadimitriou, G. I., Papazoglou, C., Pomportsis, A. S., & Tutorial), I. (2003). Optical switching: switch fabrics, techniques, and architectures. *Journal of lightwave technology*, *21*(2), 384.

[57] Shang, Y., Li, D., Zhu, J., & Xu, M. (2015). On the network power effectiveness of data center architectures. *IEEE Transactions on Computers*, *64*(11), 3237-3248.

[58] Couto, R. S., Campista, M. E. M., & Costa, L. H. M. (2012, December). A reliability analysis of datacenter topologies. In *2012 IEEE Global Communications Conference (GLOBECOM)* (pp. 1890-1895). IEEE.

[59] Mohamed, S. H., El-Gorashi, T. E., & Elmirghani, J. M. (2019). A survey of big data machine learning applications optimization in cloud data centers and networks. *arXiv preprint arXiv:1910.00731*.

[60] Li, H., Ghodsi, A., Zaharia, M., Shenker, S., & Stoica, I. (2014, November). Tachyon: Reliable, memory speed storage for cluster computing frameworks. In *Proceedings of the ACM Symposium on Cloud Computing* (pp. 1-15).

[61] Liu, N., Haider, A., Sun, X. H., & Jin, D. (2015, June). Fattreesim: Modeling large-scale fat-tree networks for hpc systems and data centers using parallel and discrete event simulation. In *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation* (pp. 199-210).

[62] Kathiravelu, P. (2016). An elastic middleware platform for concurrent and distributed cloud and mapreduce simulations. *arXiv preprint arXiv:1601.03980*.

[63] Papadimitriou, G. I., Papazoglou, C., Pomportsis, A. S., & Tutorial), I. (2003). Optical switching: switch fabrics, techniques, and architectures. *Journal of lightwave technology*, *21*(2), 384.

[64] Shang, Y., Li, D., Zhu, J., & Xu, M. (2015). On the network power effectiveness of data center architectures. *IEEE Transactions on Computers*, *64*(11), 3237-3248.