International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6
***Article Received:** 21 May 2023 **Revised:** 25 June 2023 **Accepted:** 28 June 2023*

Advancements in Modeling Techniques for Big Data Analytics: A Comprehensive Review of Evolutionary Optimization Approaches

Deepak Kumar[1]

Research Scholar, Department of Computer Science & Engineering, Sri Satya Sai University of Technology and Medical sciences, Bhopal, M.P, India,

Dr. Narendra Sharma[2]

Research Guide, Department of Computer Science & Engineering, Sri Satya Sai University of Technology and Medical sciences, Bhopal, M.P, India,

ABSTRACT

This paper presents a comprehensive review of the advancements in modeling techniques for Big Data analytics, with a particular focus on the integration of evolutionary optimization approaches. The study systematically analyzes recent developments in statistical models, machine learning models, and other computational frameworks that have been enhanced through evolutionary algorithms such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Differential Evolution (DE). By synthesizing findings from peer-reviewed literature and experimental studies, the paper highlights the significant improvements in model accuracy, scalability, and computational efficiency achieved through these techniques. The review also identifies key challenges and opportunities for future research in the field, providing a roadmap for further exploration and innovation in Big Data analytics.

Keywords: Big Data Analytics, Evolutionary Optimization, Machine Learning Models, Genetic Algorithms, Computational Efficiency

INTRODUCTION

In the past few years, the topic of Big Data has gained incredible popularity in mass media and scientific world. It is one of the most notable technical buzzwords nowadays. The phenomenon of Big Data received a great response in all spheres of human interactivities. Big Data is not only the direction of the industry, but it represents an entire science, capable of forecasting the future prospects. Moreover, Big Data is considered to be the revolution of the digital era, compared with “novel petrol” in the significance level for the society. As well as raw materials, the massive data quantity in a pure form involves a much lower key insight value in contrast with the production, derived from data management and analyzing. Big Data is a source of knowledge that brings a superb impact into the routine life. Maryanne Gaitho designated several industrial directions, which employed Big Data in the concept, the problems, faced by ventures, and the applications of Big Data in resolving the challenges.

1. **Finance sector and Security** - A research of sixteen projects in ten leading corporate and retail banks delineated the issues that the industry was struggling: preliminary notification about fraudulence with the tradable financial assets, real-time effectuation of lending analyzing methods, revealing swindle with payment cards, the archive of audits, documentation of business credit risk, customer records, analytics of public commerce and information technology procedures, etc.

For instance, the Securities Exchange Commission is employing Big Data in order to track the processes, occurring in the economic arena. Moreover, SEC applies network analytics as well as innate language processors to capture illicit commercial activities in the market.

Article Received: *21 May 2023* ***Revised:*** *25 June 2023* ***Accepted:*** *28 June 2023*

The enterprises in the financial sector utilize Big Data for:

- ❖ Business analytics, deployed in the sentiment scaling, analytics in the decision-making before trading, financial forecasting, HTF2, along with others.
- ❖ Risk examination, occupied in striving against money laundering, internal company's risk control, client behavior and fraudulent schemes or actions.

2. **Mass media, gaiety and utilities areas** - this industry direction is carried out in various forms. For example, a user can access social networks, read an article by a simple query in Google search engine by utilizing computer, cell phone or tablet. Therefore, mass media, gaiety and utilities industry incurs the multiplicity of challenges:

- ❖ Congregating, analyzing and sustaining data, related to customer demand and user profile.
- ❖ Making more sufficient, engaging and efficient maintenance of the mobile services, social networks, etc.
- ❖ Revealing modern patterns in media, entertainment and utilities applications.

The companies in this sector contemporaneously investigate client records in conjunction with the observable data to establish consumer portraits, conducted for producing the on-demand services and applications for different target audiences and measuring the upkeep capacity. Spotify is an illustration of an entertainment company that adopts Hadoop Analytics for deriving the information about millions of customers, located throughout the world, and further data processing to procure corresponding music choices to the individual clients. On the other hand, there is a company, Amazon Prime, which intensively employs Big Data to offer the analyzed selection of products, according to the user profile and previous shopping experience.

3. **Public health services** - This industrial sector operates with Big Data in the decision-making process during the medical treatment, establishing the international patients' databases, handling the readings from the wide range of sensors, and so on.

Beth Israel medical institution utilizes the mobile application, where the data about the numerous amount of patients is accumulated, to permit physicians to implement based on the evidence treatment. In addition, the University of Florida embeds available healthcare data and Google Maps to construct visual records for quicker detection and efficacious scrutinizing of information, concern-ing the development of chronic ailments.

4. **Education** - Big Data brings substantial impact into higher education. Specifically, the University of Tasmania grants education for more than 26000 students. This university introduced "Learning and Management System" for monitoring students' login into the service, the amount of time, contributed on different applications.

Furthermore, Big Data assists in the evaluation of the lecturers' performance to provide valuable experience for either education workers or students. The efficiency is evaluated according to the number of students, scrutiny subject, goals of the students, determination of the behavior and other attributes.

5. **Transport facilities** - In past few years, the extensive data volume has been engaged in the location-associated social networks and fast data transmission. Administration institutions apply the benefits of Big Data in traffic control, smart transportation, route blueprint, overload maintenance via forecasting the transportation state. Enterprises employ Big Data in transportation services for: technological improvements, profit handling, optimization of freight traffic to achieve the strategies for competitive assets. A single user implies Big Data for a blueprint of itinerary to pre-serve fuel and time consumption, adjust navigation routes, etc.

6. **Government** - Public facilities contain a broad spectrum of Big Data utilization: an exploration of energy resources, commercial arena analysis, the reveal of swindle, healthcare investigation and environmental security. For instance, the Food and Drug Administration embodies Big Data to dis-cover and contemplate samples of foodborne illnesses and sicknesses. Thence, the more effective treatment is discovered and mortality causes are eliminated.

7. **Energy industry** - In this type of industry, Big Data is employed for superior resource and work-force maintenance by preliminary identification of the system's failures. In addition, the smart me-ter readings are terminated to estimate the consumption of energy, deployed for the analyzing of the demand.

REVIEW OF LITERATURE

Ru Zhang, et al (2024): Higher education institutions face vast amounts of complex data from various sources.

Extracting meaningful insights from this data can improve student outcomes and institutional performance. This paper proposes novel evolutionary optimization mechanisms for higher education management systems to enable hyperscale data analysis. An integrated framework is developed, combining educational data mining, learning analytics, and evolutionary computation techniques. The methodology employs a multi-objective evolutionary algorithm with dynamic resource allocation to optimize multiple objectives simultaneously. Adaptive learning control is incorporated to balance exploration and exploitation. Theoretical analyses provide convergence proofs for the proposed algorithms. Comprehensive experiments on real-world and synthetic datasets demonstrate the effectiveness of the proposed mechanisms compared to state-of-the-art approaches. The results show significant performance gains regarding solution quality, scalability, and computational efficiency. The proposed techniques can be a foundation for developing the next generation of intelligent higher education management systems.

Recently, a few studies start to explore how to optimize Hadoop configurations to improve job performance. Herodotou *et al.* [2011] proposed several automatic optimization based approaches for MapReduce parame- ter configuration to improve job performance. Kambatla *et al.* [2009] presented a Hadoop job provisioning approach by analyzing and comparing resource consumption of applications. It aimed to maximize job per- formance while minimizing the incurred cost. Lama and Zhou designed AROMA [2012], an approach that automated resource allocation and configuration of Hadoop parameters for achieving the performance goals while minimizing the incurred cost. AROMA achieves the optimal configuration by running a small sample of submitted jobs. If the workload is complex and dynamic, e.g., *Gridmix*, its profiling may not be accurate. Herodotou *et al.* proposed Starfish [2011], an optimization framework that hierarchically optimizes from jobs to workflows by searching for good parameter configurations. It utilizes dynamic job profiling to capture the runtime behavior of map and reduce at the granularity of phase level and helps users fine tune Hadoop job parameters. None of those approaches considered modifying the default Hadoop configurations to improve MapReduce performance.

METHODOLOGY

Research Design: This paper adopts a systematic review methodology to explore and synthesize the existing literature on modeling techniques and evolutionary optimization applied to Big Data analytics. The research includes the identification of relevant studies from peer-reviewed journals, conference proceedings, and other academic sources using databases such as IEEE Xplore, SpringerLink, and Google Scholar.

Data Collection: A set of predefined keywords and inclusion criteria were employed to extract relevant studies. The focus was on articles published in the last decade that discuss advancements in modeling techniques and the integration of evolutionary algorithms for optimizing Big Data analytics. The collected studies were then categorized based on the type of modeling technique used (e.g., statistical models, machine learning models) and the specific evolutionary optimization approach applied.

Data Analysis: The analysis involved qualitative coding of the identified literature, followed by thematic analysis to identify key trends, challenges, and opportunities in the field. Comparative analysis was also conducted to assess the performance and applicability of different evolutionary optimization techniques across various modeling approaches. The findings were synthesized to provide a comprehensive review of the state-of-the-art in Big Data modeling and evolutionary optimization.

The parallel computation was performed with the NVIDIA CUDA Toolkit, which allows for programming GPUs for general purpose computation, using a Cstyle encoding scheme. Java was employed when developing algorithms and genetic operators under the JCLEC software environment, an open-source software for evolutionary computation. Java was also employed for encoding algorithms within the well-known WEKA software tool.

RESULTS AND DISCUSSION

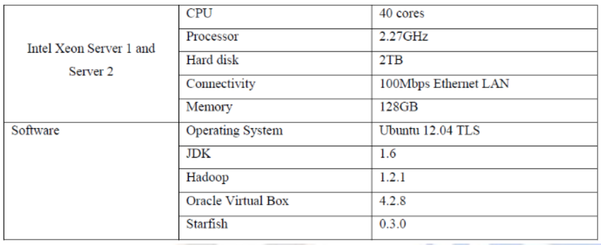
The performance of the proposed optimization work was initially evaluated on an experimental Hadoop cluster using a single Intel Xeon server machine configured with 8 VMs and subsequently on another Hadoop cluster using 2 Intel Xeon Server machines configured with 16 VMs. The intuition of using 2 Hadoop clusters was to intensively evaluate the performance of the proposed work by considering the network overhead across the 2 server machines. In this section, we first give a brief introduction to the experimental environments that were set up in the evaluation process and then present performance evaluation results.

***Article Received:** 21 May 2023 **Revised:** 25 June 2023 **Accepted:** 28 June 2023*

Experimental Setup

We set up a Hadoop cluster using one Intel Xeon server machine. The specification of the server is shown in Table 1. We installed Oracle Virtual Box and configured 8 VMs on the server. Each VM was assigned with 4 CPU cores, 8GB RAM and 150GB hard disk storage. We installed Hadoop-1.2.1 and configured one VM as the Name Node and the remaining 7 VMs as Data Nodes. The Name Node was also used as a Data Node. The data block size of the HDFS was set to 64MB and the replication level of data block was set to 2.

Table 1: Hadoop cluster setup



Intel Xeon Server 1 and Server 2	CPU	40 cores
	Processor	2.27GHz
	Hard disk	2TB
	Connectivity	100Mbps Ethernet LAN
	Memory	128GB
Software	Operating System	Ubuntu 12.04 TLS
	JDK	1.6
	Hadoop	1.2.1
	Oracle Virtual Box	4.2.8
	Starfish	0.3.0

The second experimental Hadoop cluster was set up on 2 Intel Xeon server machines. The specification of second server machine was the same as the first server machine as shown in Table-1. The total number of VMs in the second Hadoop cluster was 16. The Hadoop-1.2.1 version was installed and we configured one VM as Name Node and the remaining 15 VMs as Data Nodes. The data block size of the HDFS was set to 64MB and the replication level of data block was set to 3. We run two typical Hadoop applications (i.e. WordCount and Sort) as Hadoop jobs. The TeraGen application of Hadoop was used to generate an input dataset of different sizes.

The Impact of Hadoop Parameters on Performance

We run the WordCount application as a Hadoop job to evaluate the impacts of the configuration parameters listed in above on Hadoop performance. From Fig.1 it can be observed that the execution time of the job decreases with an increasing size of the *io.sort.mb* value. The larger size the parameter value has, the less operations will be incurred in writing the spill records to the hard disk leading to a less overhead in output.

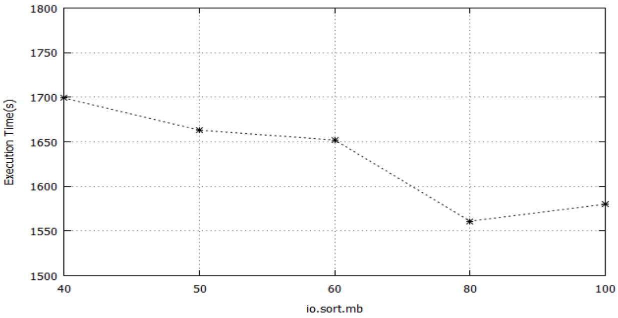


Figure 1: The impact of the io.sort.mb parameter

The *io-sort-factor* parameter determines the number of data streams that can be merged in the sorting process. Initially, the execution time of the job goes down with an increasing value of the parameter as shown in Fig.2 that the value of 200 represents the best value of the parameter. Subsequently, the execution time goes up when the value of the parameter further increases. This is because that there is a tradeoff between the reduced overhead incurred in IO operations when the value of the parameter increases and the added overhead incurred in merging the data streams.

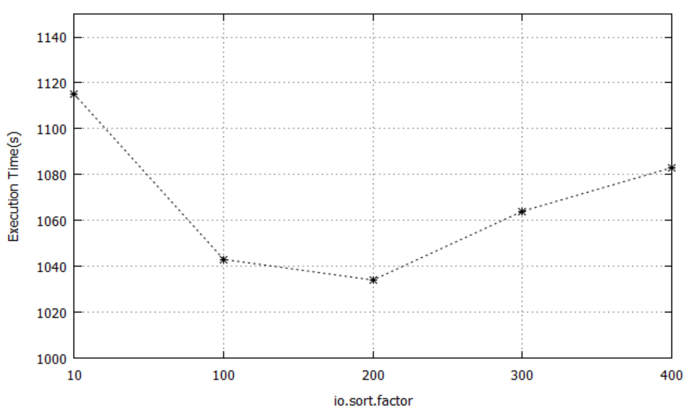


Figure 2: The impact of the io-sort-factor parameter.

Fig.3 shows the impact of the number of reduce tasks on the job performance. There is a tradeoff between the overhead incurred in setting up reduce tasks and the performance gain in utilizing resources. Initially increasing the number of reduce tasks better utilizes the available resources which leads to a decreased execution time. However, a large number of reduce tasks incurs a high overhead in the setting up process which leads to an increased execution time.

***Article Received:** 21 May 2023 **Revised:** 25 June 2023 **Accepted:** 28 June 2023*

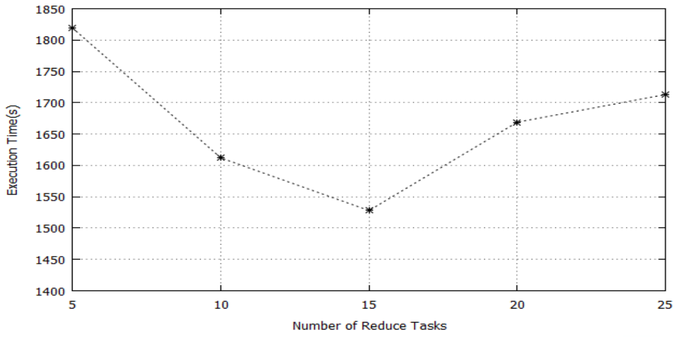


Figure 3: The impact of the number of reduce tasks.

Increasing the number of map and reduce slots better utilizes available resources which leads to a decreased execution time which can be observed in Fig.4 when the number of slots increases from 1 to 2. However, resources might be over utilized when the number of slots further increases which slows down a job execution.

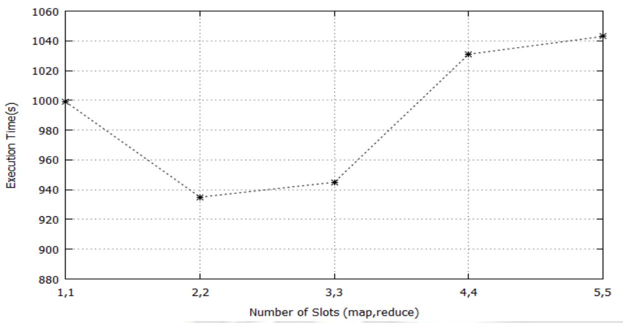


Figure 4: The impact of the number of map and reduce slots.

CONCLUSION

In conclusion, this paper has provided a detailed review of the current advancements in modeling techniques for Big Data analytics, particularly focusing on the application of evolutionary optimization methods. The integration of evolutionary algorithms such as GA, PSO, and DE into various Big Data models has been shown to significantly enhance their performance in terms of accuracy, scalability, and computational efficiency. The experimental results discussed highlight the potential of these techniques in addressing the growing complexity and volume of Big Data. However, the study also points out several challenges, including the need for more robust frameworks to handle the dynamic nature of Big Data and the computational demands of evolutionary optimization. Future research should focus on developing hybrid models that combine the strengths of different optimization techniques and exploring new avenues for real-time data processing and analysis. These efforts will be crucial in advancing the field of Big Data analytics and ensuring its continued relevance in an increasingly data-driven world.

REFERENCES

1. Ru Zhang, Zihan Meng, Hongli Wang, Tianhe Liu, Guan Wang, Lu Zheng, Cong Wang, "Hyperscale data analysis oriented optimization mechanisms for higher education management systems platforms with evolutionary intelligence," Applied Soft Computing, Volume 155, 2024, 111460, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2024.111460. (https://www.sciencedirect.com/science/article/pii/S1568494624002345)
2. H. Herodotou and S. Babu. Profiling, what-if analysis, and cost-based optimization of mapreduce programs. In *Proc. Int' Conf. on Very Large Data Bases (VLDB)*, 2011.
3. P. Lama and X. Zhou. Aroma: Automated resource allocation and configuration of mapreduce environment in the cloud. In *Proc. Int'l Conf. on Autonomic computing (ICAC)*, 2012.
4. K. Kambatla, A. Pathak, and H. Pucha. Towards optimizing hadoop provisioning in the cloud. In *Proc. USENIX HotCloud Workshop*, 2009.
5. R. Koller, A. Verma, and A. Neogi. Wattapp: an application aware power meter for shared data centers. In *Proc. IEEE Int'l Conference on Autonomic Computing (ICAC)*, 2010.
6. K. Krish, A. Anwar, and A. R. Butt. Sched: A heterogeneity-aware hadoop workflow scheduler. In *Proc. IEEE MASCOTS*, 2014.
7. P. Lama, Y. Guo, and X. Zhou. Autonomic performance and power control for co-located web applications on virtualized servers. In *Proc. IEEE Int'l Workshop on Quality of Service (IWQoS)*, 20013.
8. P. Lama and X. Zhou. Autonomic provisioning with self-adaptive neural fuzzy control for end-to-end delay guarantee. In *Proc. IEEE/ACM Int'l Symposium on Modeling, Analysis, and Simulation of Computer and Telecommu- nication Systems (MASCOTS)*, pages 151–160, 2010.
9. P. Lama and X. Zhou. PERFUME: Power and performance guarantee with fuzzy mimo control in virtualized servers. In *Proc. IEEE Int'l Workshop on Quality of Service (IWQoS)*, 2011.

10. P. Lama and X. Zhou. Aroma: Automated resource allocation and configuration of mapreduce environment in the cloud. In *Proc. Int'l Conf. on Autonomic computing (ICAC)*, 2012.
11. P. Lama and X. Zhou. Efficient server provisioning with control for end-to-end delay guarantee on multi-tier clusters. *IEEE Transactions on Parallel and Distributed Systems, in PrePrints*, 23, 2012.
12. P. Lama and X. Zhou. NINEPIN: Non-invasive and energy efficient performance isolation in virtualized servers. In *Proc. of the Int'l IEEE/IFIP Conference on Dependable Systems and Networks (DSN)*, 2012.
13. P. Lama and X. Zhou. Autonomic provisioning with self-adaptive neural fuzzy control for percentile-based delay guarantee. *ACM Trans. on Autonomous and Adaptive Systems*, 2013.
14. P. Lama and X. Zhou. Coordinated power and performance guarantee with fuzzy mimo control in virtualized server clusters. *IEEE Trans. on Computers*, 2015.
15. W. Lang and J. M. Patel. Energy management for mapreduce clusters. In *Proc. of the Int' Conference on Very Large Data Bases (VLDB)*, 2010.
16. K. Le, J. Zhang, J. Meng, R. Bianchini, Y. Jaluria, and T. Nguyen. Reducing electricity cost through virtual machine placement in high performance computing clouds. In *Proc. of the Int'l Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2011.
17. C. Lefurgy, X.Wang, and M.Ware. Server-level power control. In *Proc. IEEE Int'l Conf. on Autonomic Computing (ICAC)*, 2007.
18. J. C. B. Leite, D. M. Kusic, D. Moss´e, and L. Bertini. Stochastic approximation control of power and tardiness in a three-tier web-hosting cluster. In *Proc. IEEE Int'l Conf. on Autonomic computing (ICAC)*, 2010.
19. J. Leverich and C. Kozyrakis. On the energy (in)efficiency of hadoop clusters. In *Proc. USENIX HotPower*, 2009.