Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI:10.5281 /zenodo.777783

OPTIMIZED CONTINUOUS QUALITY AND STORAGE MANAGEMENT MODEL FOR BIG DATA ANALYSIS

Peerzada Hamid Ahmad[1]*, Munishwar Rai [2]

[1*]M.M. Institute of Computer Technology & Business Management, Maharishi Markandeshwar University, India
[2]M.M. Institute of Computer Technology & Business Management, Maharishi Markandeshwar University, India

[*]**Corresponding Author:** Peerzada Hamid Ahmad
[*]M.M. Institute of Computer Technology & Business Management, Maharishi Markandeshwar University, India

Abstract:
Due to the increase in devices and the availability of connecting devices, data creation increased gradually. To handle the huge amount of data that came from many different sources, the business analysis had to use effective analysis and storage optimization. So, one of the main areas of research to handle big data, keep its quality up, and get the most out of storage is how to analyses and store big data. This paper introduced a new framework for collaborative big data analysis and optimized storage using the new framework model. The proposed framework model is called D2SAE (Dynamic Domain Sample Attributes Evaluation). The proposed framework comprises four main steps: dynamic data collection from domains, sample attribute collection, evaluation metrics for quality data analytics, and an optimization process. The proposed framework was evaluated using metrics such as throughput, runtime, average latency, minimum latency, and maximum latency. Our proposed model produces better throughput and reduces runtime and latency compared to the previous framework. The results show that the throughput is increased, and the run time is reduced to 42 seconds.

Keywords: Big data analytics, Optimization, Data storage, Quality analysis**,** Evaluation metrics.

1. INTRODUCTION

Data quality and optimized data storage are major barriers in big data management and analytics. Big data applications deal with a greater volume, diversity, and velocity of data, and data quality acquires an even greater significance. Also, problems with the quality of a lot of data could lead to problems with different platforms, applications, data types, and use cases. Problems with big data quality could lead to wrong algorithm outputs and dangerous system outputs that could cause accidents and even deaths. The global data management research shows the impact of poor-quality management and its affectionate factors in Figure 1.

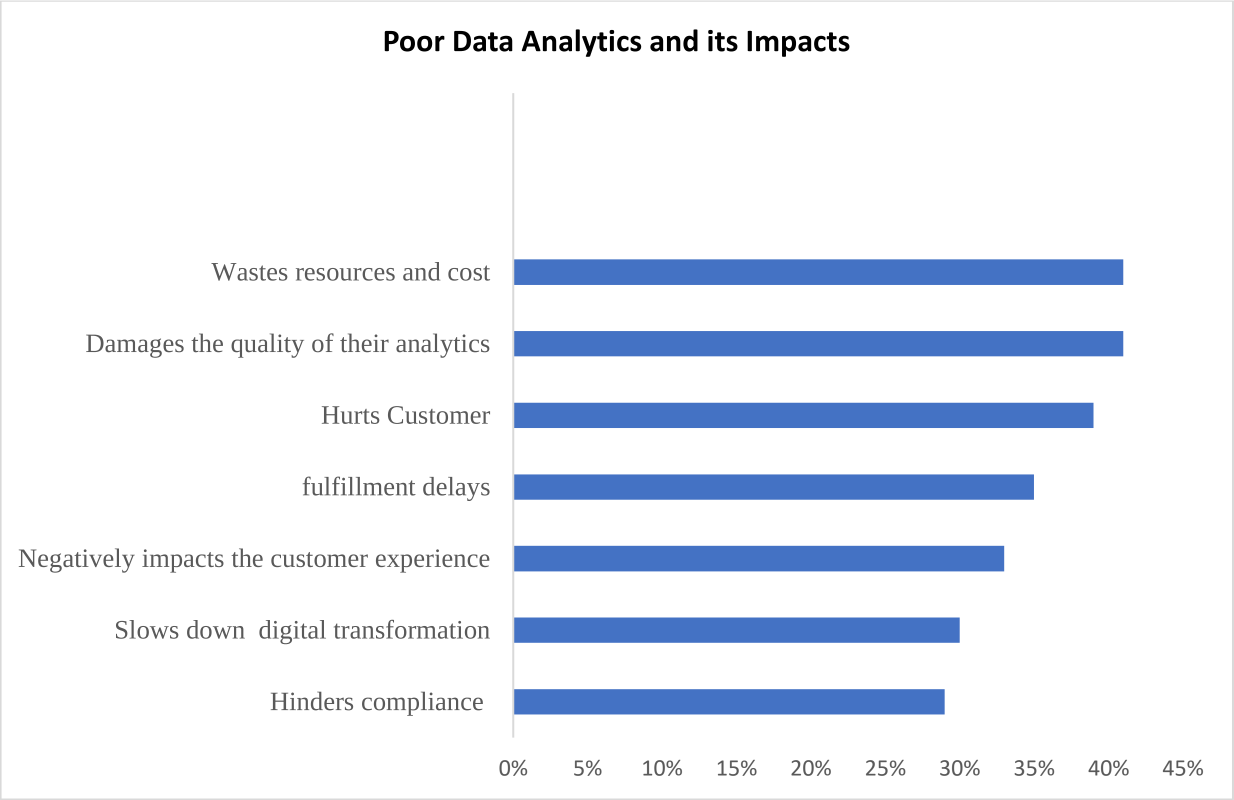

Figure 1. Poor Data Quality and its Impact's

Business users will be less inclined to trust data sets and the apps that rely on them. Additionally, if the government considers data integrity and quality to be crucial determinants in strategic business decisions, organizations may be subject to regulating the quality in big data analysis. Big data management consists of four main life cycles: data analysis, data storage, data acquisition, and data generation. Figure 1 depicts the life cycle of big data management [13].

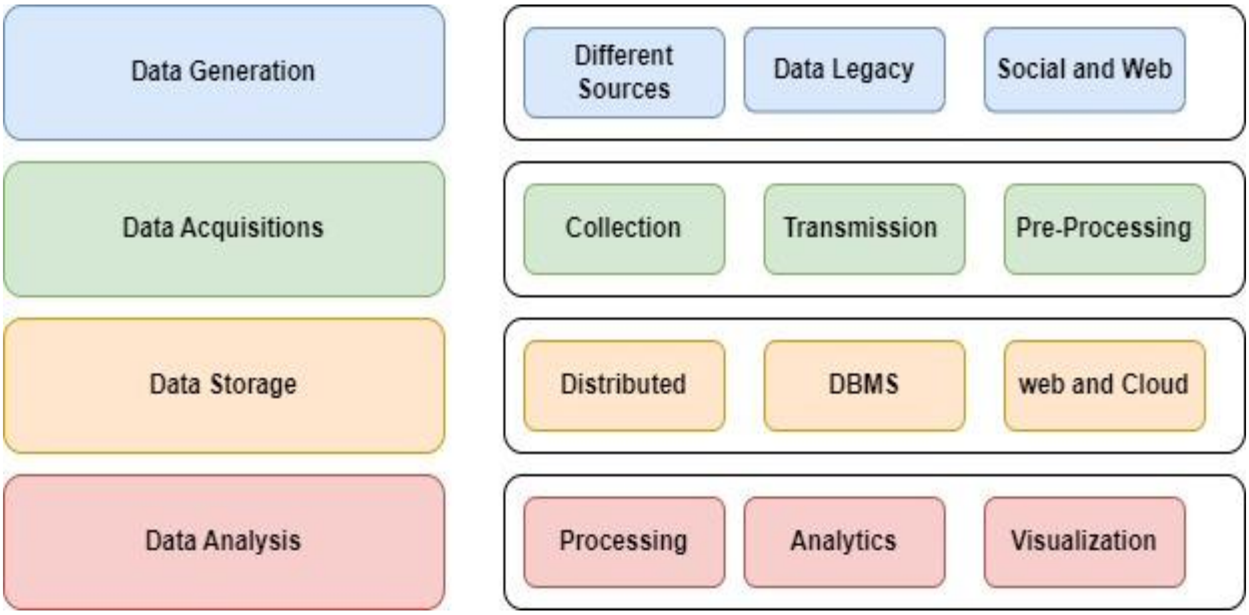

Figure 2. Life cycle for big data Management

Big data management quickly refers to structured and unstructured data at both the organizational and administrative levels. Traditional software and methods for dealing with large amounts of data are very hard to use. Ordinary tools and storage systems cannot be used to store, process, or manage data of the same magnitude [1]. A single massive data set could range from a few terabytes to many petabytes. Because of this, it is hard to store, extract, share, and analyse these huge amounts of data for several reasons. Businesses now use high-quality data to find hidden patterns that couldn't be seen before [2]. Big data analytics uses complex algorithms and methods to extract hidden data definitions. Big data technology is replacing the old tools that were used to access and change huge amounts of data through the internet [3]. The data is collected through big data platforms, which are rapidly growing. The quality of data management service is required in the different platforms and service-providing application areas [4]. With the help of different techniques, service-providing techniques graded customer service and made other suggestions that were easy to follow [5]. For better service quality, various feasible collective behaviour patterns are gathered from various sources.

The main objective of this work is to provide optimized quality of service management using big data analysis in the health sector. In this work, propose an effective, optimized framework for analysing big data and providing quality services in different domains and data distributions. Different qualitative parameters can be used with the proposed framework to improve the process of managing big data. The porpose work Dynamic Domain Sample Attributes Evaluation (D2SAE) is used Cuckoo–Grey Wolf-based Correlative Naive Bayes classifier (CGCNB) is used for optimization. The Correlative Naive Bayes classifier optimization model combines the correlation among the different attributes into different hypotheses.

The rest of the article is organized as follows: Section II provides the different related work supported to this proposed work, and Section III provides the optimized framework for providing the quality of service in big data management. Section IV provides the implementation and comparison details for the proposed work, and V provides the conclusion for the proposed work.

2. RELATED WORKS

In terms of real-time structured and unstructured data, big data is different from conventional methods [10, 11]. The authors of Chen et al. [6] thoroughly analysed massive data ontologies for data analysis. The authors of [7] proposed a framework called BlockMon for flexible, high performance and use case data analytics. In this work, authors use different Boolean combiners, bloom filters, and URLs for a better telemarketer. The authors of [8] proposed a FIU-Miner model to integrate a fast and user-friendly system for different application data analysis. The proposed system architecture is used in different applications such as spatial analysis, manufacturing etc. The authors of [9] proposed a Hadoop-based platform on huge clusters of commodity machines, and this work, Twitter infrastructure was used for data analysis. The proposed work provided detail on design principles, procedures, and opportunities brought on big data analysis. The authors of M. Chen et al. [12] conducted a brief research analysis for data management, applications, data gathering, storage, and technology in terms of potential future use. The taxonomy of architectures, various information system platforms, and how to adopt new approaches are also covered in a detailed manner. Pre-processing data prior to analytics is a long-standing practice. A different number of issues have been addressed in the Big Data value chain by the author [14]. Data quality is one of the main issues carefully considered in the context of big data. Working with several data sources causes problems with data quality [15]. Because the requirement for data cleansing has significantly increased. In addition, the cleaning processes are strained by the enormous techniques that come in at an unpredictable rate. The authors of [16] proposed a data cleaning system called NADEEF for Big Data cleaning-based streaming data. The authors of [17] proposed a system for managing data quality with a data cleaning process using the functional dependencies rules. The author of [18] proposed a framework for the analysis and handling of storage, pre-processing, and processing of Big Data. The author of [19] proposed a brief state of the art and challenges of big data management. This work discussed the importance of quality assessment in large data management, the attributes for recommendations for quality, solutions available to map or solve quality issues that can occur from these Big Data Vs. A Big Data Management and Framework to end-to-end analysis of big data management is presented. Ikbal Taleb et al. [13] presented a different framework for pre-processing and controlling the quality of big data. This work describes the different qualities, requirements, dimensions, and rules. Kumar, Jitendra, et al. [20] presented data workflow and described quality issues related to a web-based portal for big data management. The web-based portal provided the quality of the corrective action for symbolic equations. Akuna's et al [21] provided quality assurance for supply chain management using Lean six Sigma to continuously improve big data analysis quality. This work presents and applies the five main qualities of big data analytics, such as velocity, volume, veracity and variety, and values, in supply chain management. The some of the most relevance framework and corresponding limitations are presented in Table 1.

S. No	Methods	Functions and Advantages	Limitations
1	FIU-Miner [8]	Without code, analysis the complex data using analysis of Programme.	Huge distributed environment difficult to configuring complex data.
2	Framework for Data Quality Profile. [13]	The data quality and rules are introduced the assignment and evaluation.	The framework not used quality evaluation metrics for evaluation.
3	NADEEF [16]	The platform introduced clearing rule for quality data.	Not used the summarization techniques and evaluation parameter for data clearing process.
4	Data Quality [18]	Used veracity for increasing the recognition of quality rules.	Not used any evaluation metrics for quality rules.
5	Six Sigma [21]	Used to fault detection and elimination of feedback from supply chain management.	Quality assurance parameters not described in the theoretical framework.

3. FRAMEWORK FOR DATA QUALITY OF SERVICE AND OPTIMIZATION

In this paper, we propose a novel framework for optimized continuous quality analytical service management for big data management. We proposed a framework model called D2SAE (Dynamic Domain Sample Attributes Evaluation) in this work. The structure of the proposed framework is shown in Figure 3. The proposed work has four main steps, such as dynamic data collection from domains, sample attribute collection, evaluation metrics for quality data analytics, and an optimization process [29, 30]. The data collection is used to get new raw information from dynamic websites, and the sample attributes collection is used to get different attributes for the website and sources. Using the different metrics, the evaluation metrics are used to judge the data from the sources. Finally, optimization is used to speed up the process of evaluation continuously. In the next few sections of the proposed work, information about the four steps of the process is given briefly.

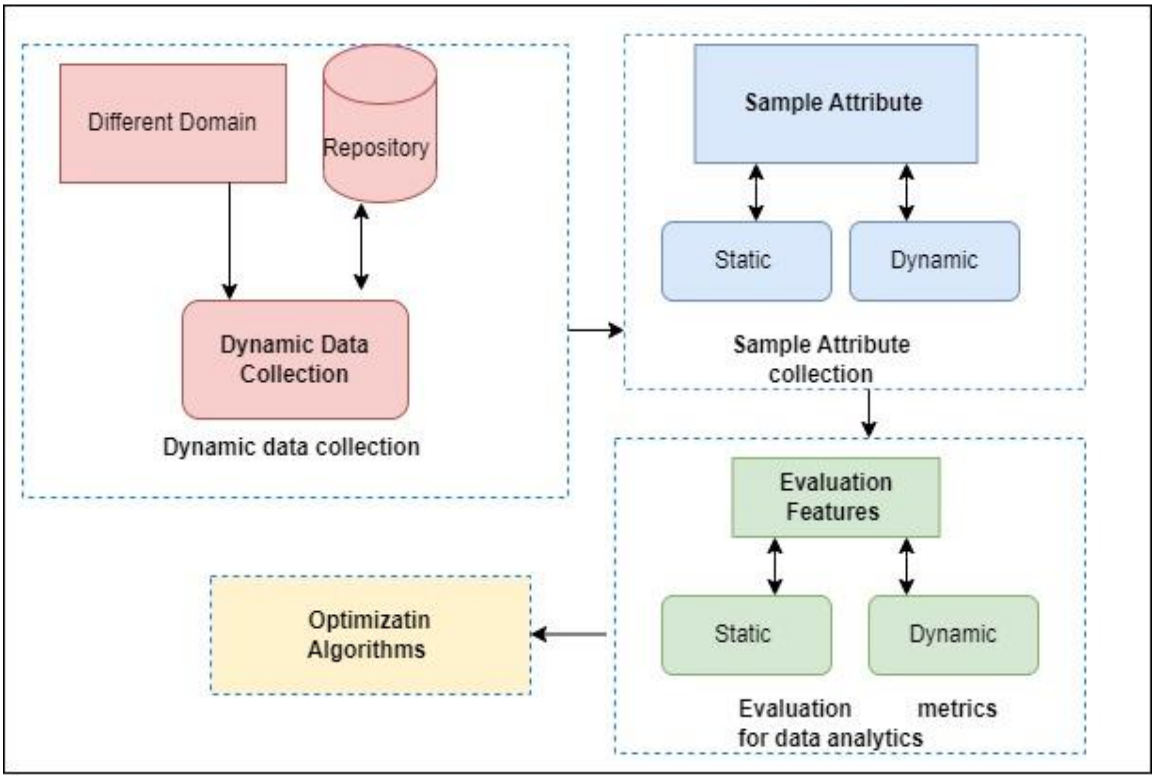

Figure 3. Framework for continuous data analytics

3.1. Dynamic data collection form domain

The data collection process is used to gather basic information for data analytics. Data was collected from two sources: various domains and manual domain inputs. In this work, we have mainly focused on automatically creating different keyword collections from different domains. The main qualitative domain keywords are words, semantics, and views. The quantitative analysis considered the table, numbers, years, and contents. To extract the information, qualitative and qualitative keywords are used. In the previous work, only static quantitative and qualitative words were used to extract the information. This work's framework changes the quantitative and qualitative words dynamically using the different intervals or time periods. So, for data analytics, the keywords extract the information dynamically based on the theme or our requirements. Another way of collecting the sources of dynamic information is pattern matching [27, 28]. The collected information matches based on the pattern, and it's stored in the repository for data analysis. The initial display for the proposed work is based on the quantitative and qualitative metrics and with the help of different keywords.

3.2. Sample Attribute collection

One of the main parts of data analytics is collecting sample attributes. Previously, the researchers used different methods for data sampling and collecting the attributes. The sample collection methods are classified into two types: probability sampling and non-probability sampling methods. The probability sampling method is classified into random sampling, systematic sampling, stratified sampling, the cluster sampling method, etc. Similarly, the non-probability method is classified into convenience sampling, response sampling, purposeful sampling, snowball sampling, etc. In this work, we have used cluster sampling for static and dynamic data grouping. Initially, the given manual samples are divided into subgroups. The subgroup is formed based on characteristics and is given manual samples. The automatic data analytics subgroup is randomly grouped using the different possible inputs or automatically takes the dynamic keywords or possible main keywords and features. The advantage of cluster sampling is that it requires fewer resources for grouping the entire population and sampling, making it more feasible for homogeneous and heterogeneous data analysis. Compared to other clustering methods such as random sampling, it is simple and provides effective sampling.

Accuracy
Validity
Completeness
Data Quality Metrics
Uniqueness
Consistency
Timeliness

Figure 4. Evaluation metrics for data quality

3.3. Evaluation metrics for Quality data analytics

In our framework, we considered the six-evaluation metrics to ensure data quality. These six metrics check the quality and track the metrics, accuracy, completeness, consistency, timeliness, uniqueness, and validity [22]. The relationship between the evaluation metrics is shown in Figure 4.

3.3.1. Accuracy

Measures the types of errors in the dataset or from the data sources. Some of the inaccurate metrics considered in the dataset include anomalous values, microvolume, anomalous strings, intercolumn relationships, and distribution.

3.3.2. Completeness

Completeness is used to consider the full population of data. Consider both complete and incomplete data when determining completeness. It is tracked using different fields of record.

3.3.3. Consistency

Consistency measures the data points from two or more data sources. The data point is difficult, it combines different sources or domains.

3.3.4. Timeliness

The timeliness is considering the aging of the data from the various sources. The current data sources are considering different activities such as accuracy, relevance, customer requirements, errors, and pipeline processing according to timing.

3.3.5. Uniqueness

The uniqueness is considering the tracking of duplicate data. Initially, the data is collected from different sources, and the duplicate data is deleted from the repository.

3.3.6. Validity

The validity measures the standard of data with stable and unstable parameters. Each data set should consider the format and the required information.

3.4. Optimization process

The optimization process is an important part of our proposed framework for reducing storage and process. In this framework, the Bayesian principle is used to classify and reduce the storage requirements of data. The Cuckoo–Grey Wolf-based Correlative Naive Bayes classifier (CGCNB) is used for optimization [23, 24]. The CNB optimization model combines the correlation among the different attributes into different hypotheses. The grey wolf optimization method is integrated with the CNB classifier for better performance improvement.

The working process of the proposed work consists of two parts such as analysis of static data and analysis of dynamic data from the various domains. The static and dynamic analysis data compared with the sample attributes and first evaluated with the fixed parameters form the evaluation parameters. In the framework any new relevance metrics are added in to the evaluation features. Parallelly fixed evaluation metrics mentioned in the section 3.3 are evaluation and relationship between the parameters are analysis. After that the Correlative Naive Bayes classifier optimization method is applied for optimization of the different parameters and classification of analysis.

4. RESULT AND DISCUSSIONS

This is section presented implementation and comparison details with existing method. For comparison DSBIGDA Framework is used with different parameters such as throughput, runtime, average latency, min latency and maximum latency. The dataset details for implementation of training, testing and domain information are described in the Table 1 [25]. The dataset consists of treatment, side effect, associate details, dugs, medical entities, semantic types, and unique identifiers. The D2SAE is implemented using health dataset, python programming language and semantic script for extraction of information form the dataset. Table 2 shown the hardware and software details for implementation.

Table 1. Dataset Information for Implementation

Datasets	**Type**	**Domain**	**Pairs**
SNLI (2015)	Inference Pair	Open – Domain	550,152 (train)
MultiNLI (2017)	Inference Pair	Open – Domain	392,702 (train)
Quora (2017)	Similarity pairs	Open – Domain	404,279
New Test Data (CHQs)	Entailment pairs	Consumer Health Questions	850

Table 2. Implementation Environment

CPU	Intel (R) Xeon (R) 4110
Memory	128 GB RAM
Operating System	Ubuntu 16.04 LTS
Programming Language	Python v.3.5

The general functions are used to get the inputs from the sources to read and write inputs and outputs to the system. The functionalities are experimented with using different scenarios, and basic metric measures are calculated using the above-mentioned parameters. The different parameter comparisons between proposed work and existing work are shown in Table 3.

Table 3. Performance Metrics Comparison

Metrics	**DSBIGDA Framework [26]**	**D^{2}SAE (Proposed Method)**
Throughput	195321	198765
Runtime	521	480
Average Latency	3.021	2.8
Min Latency	0.513	0.4
Max Latency	304	208

Table 3 shows the performance metrics for the proposed work, and the comparison shows that throughput and other metrics are reduced. Figures 4 and 5 depict a graphical comparison of throughput and other metrics. Compared to the throughput, due to the conditions of the different metrics mentioned in Figure 3, the throughput is increased gradually in the different situations. Due to the relationship between the metrics of each metric, such as accuracy, completeness, consistency, timeliness, uniqueness, and validity, we received better results in terms of throughput and other metrics such as minimum latency, maximum latency, and average latency.

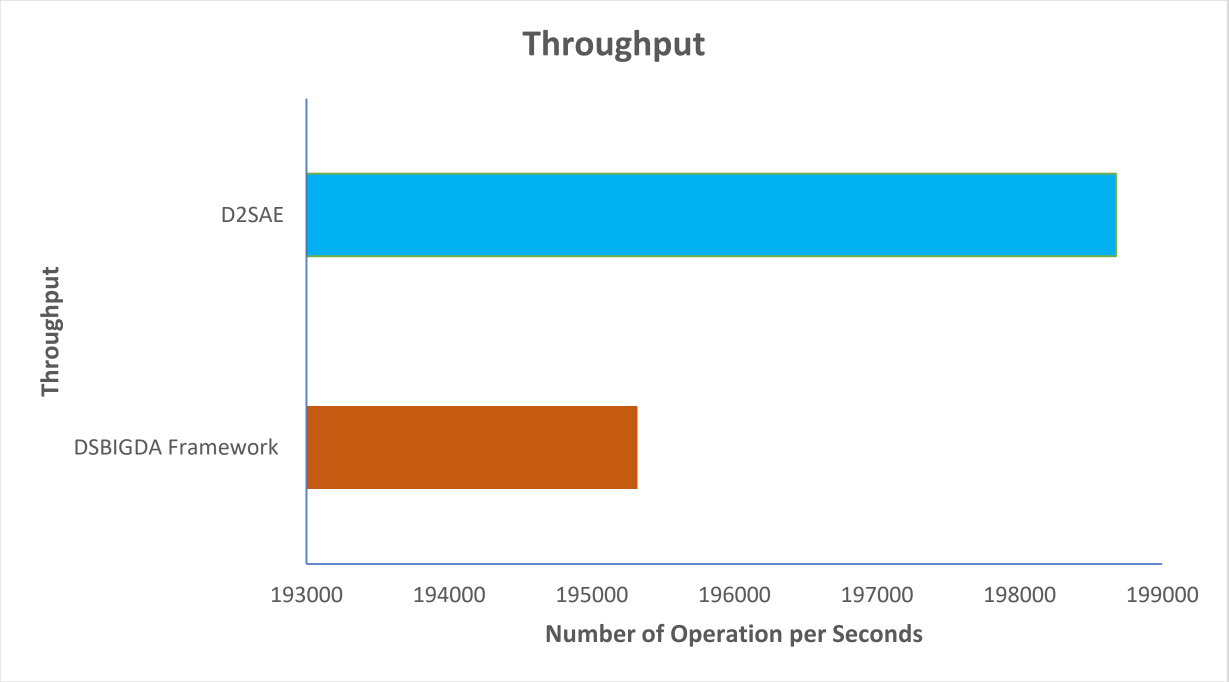

Figure 4. Comparison of Throughput

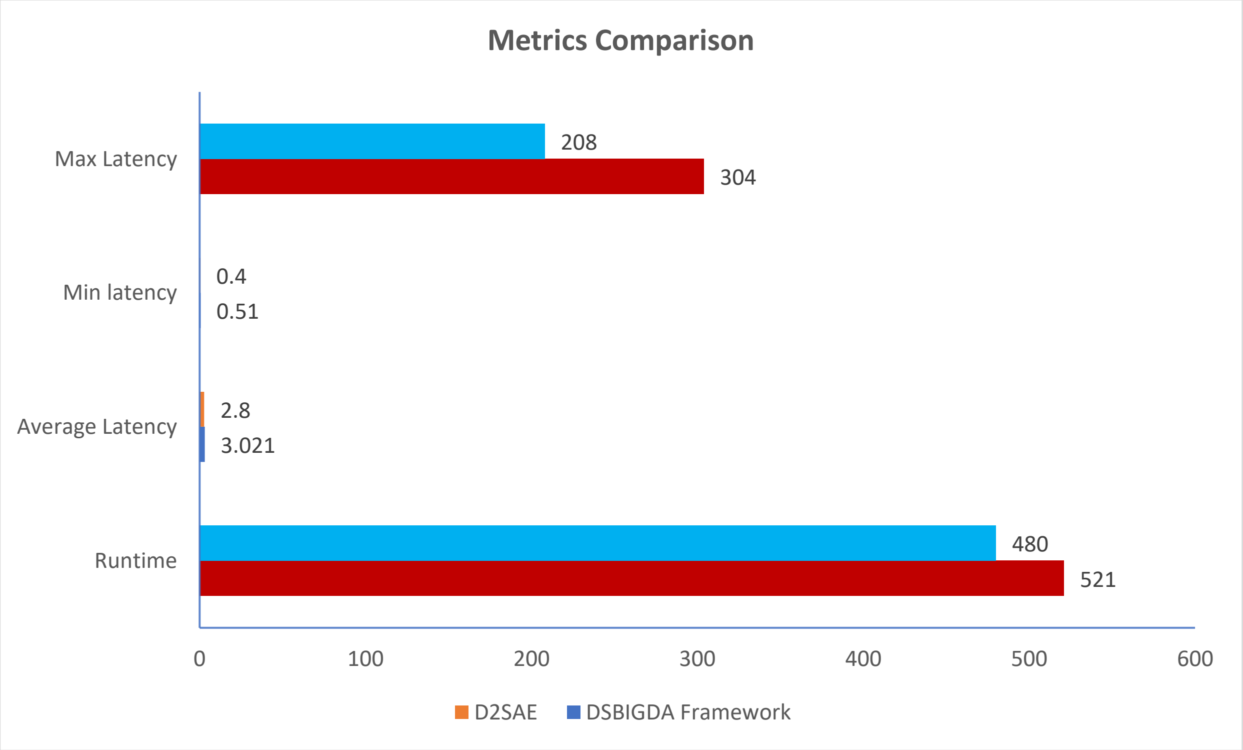

Figure 5. Comparison of different metrics

The runtime between the previous work and the proposed work is reduced by 42 seconds due to the effective metrics and optimization method. The CGCNB optimization combined different metrics and correlations with various features. Similarly, the latency between the previous work and the proposed work decreased as time passed. In our work, the time intervals are calculated using different scenarios, such as minimum latency, maximum latency, and average latency. Compared to the other method, the overall latency is reduced in different situations, such as read and write operations. Figure 5 shows the different latency comparison with the previous work, and overall latency is reduced in three situations such as minimum, maximum, and average.

5. CONCLUSION

This work introduced an effective framework for analysing data quality and optimizing storage using different frameworks. In this paper, a new framework for analysing big data continuously and storing it in the best way possible was described. The proposed framework model is called D2SAE (Dynamic Domain Sample Attributes Evaluation). The proposed framework is made up of four main steps: dynamic data collection from domains, sample attribute collection, evaluation metrics for quality data analytics, and an optimization process. Metrics such as throughput, runtime, average latency, minimum latency, and maximum latency were used to judge the proposed framework. In the future, better metrics and optimization methods will need to be considered for an effective model.

REFERENCES

[1]. Angelov, P. Grefen, and D. Greefhorst, “A framework for analysis and design of software reference architectures,” Inf. Softw. Technol. 54 (2012) 417–431.

[2]. M. Galster, and P. Avgeriou, “Empirically-grounded reference architectures: a proposal,” in: Joint ACM SIGSOFT Conference on Quality of Software Architectures and ACM SIGSOFT Conference on Quality of Software Architectures and ACM SIGSOFT Symposium on Architecting Critical Systems, Boulder, Colorado, USA, June 20–24, 2011.

[3]. Y. Demchenko, and C. Ngo, P. Membrey, “Architecture framework and components for the Big Data Ecosystem,” SNE Technical Report, University of Amsterdam, September 12, 2013.

[4]. Mishne, “Fast data in the era of big data: Twitter’s real-time related query suggestion architecture,” in: The 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 22–27 June 2013.

[5]. R. Sumbaly, J. Kreps, and S. Shah, “The “Big Data” Ecosystem at LinkedIn, in: 2013 ACM SIGMOD International Conference on Management of Data,” New York, New York, USA, 22–27 June, 2013.

[6]. D. Simoncelli, M. Dusi, F. Gringoli, and S. Niccolini, “Stream-monitoring with BlockMon: convergence of network measurements and data analytics platforms,” ACM SIGCOMM Commun. Rev. 43 (2013) 29–35.

[7]. M. Dusi, et al., “BlockMon: flexible and high- performance big data stream analytics platform and its use cases,” NEC Tech. J. 7 (2012) 102–106.

[8]. C. Zeng, et al., “FIU-miner: a fast, integrated, and user-friendly system for data mining in distributed environment,” in 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 11–14 August 2013.

[9]. G.L. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy, “The unified logging infrastructure for data analytics at Twitter,” in: The 38th International Conference on Very Large Databases, Istanbul, Turkey, 27–31 August 2012.

[10]. C. Wang, K. Schwan, V. Talwar, G. Eisenhauer, L. Hu, and M. Wolf, "A Flexible Architecture Integrating Monitoring and Analytics for Managing Large-Scale Data Centers,” in the proceedings of ACM ICAC’11, 2011, Germany, pp. 141 -150.

[11]. C.L.P. Chen, and C. Zhang, “Data-intensive applications, challenges, techniques and technologies: a survey on Big Data,” Inf. Sci. 275 (2014) 314–347.

[12]. M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” Mob. Netw. Appl. 18 (2014).

[13]. Taleb, Ikbal, et al. "Big data quality framework: a holistic approach to continuous quality management." *Journal of Big Data* 8.1 (2021): 1-41.

[14]. Hu H, Wen Y, Chua T-S, Li X. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access.2014; 2:652–87. https:// doi.org/ 10. 1109/ ACCESS. 2014. 23324 53.

[15]. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000; 23:3–13.

[16]. Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I.F., Ouzzani, M., Tang, N., 2013. NADEEF: A Commodity Data Cleaning System, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD ’13. ACM, New York, NY, USA, pp. 541–552. https:// doi. org/ 10. 1145/ 24636 76. 24653 27.

[17]. Tang N. Big Data Cleaning. In: Chen L, Jia Y, Sellis T, Liu G, editors. Web Technologies and Applications. Lecture Notes in Computer Science: Springer International Publishing; 2014. p. 13–24.

[18]. Saha, B., Srivastava, D., 2014. Data quality: The other face of Big Data, in: 2014 IEEE 30th International Conference on Data Engineering (ICDE). Presented at the 2014 IEEE 30th International Conference on Data Engineering (ICDE), pp.1294–1297. https:// doi. org/ 10. 1109/ ICDE. 2014. 68167 6.

[19]. Ge M, Dohnal V. Quality management in big data informatics. 2018; 5:19. https:// doi. org/ 10. 3390/ infor matic s5020019.

[20]. Kumar, Jitendra, et al. "Provenance–aware workflow for data quality management and improvement for large continuous scientific data streams." *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.

[21]. Akuns, Uloh, and Sebastian Okafor. "Big data analytics: virtuosity in Lean Six Sigma for quality assurance in supply chain management." Interdisciplinary Journal of Economics and Business Law 11.2 (2022): 44-72.

[22]. https://firsteigen.com/blog/6-key-data-quality-metrics-you-should-be-tracking/#:~:text=The%20six%20key%20data%20quality,timeliness%2C%20uniqueness%2C%20and%20validity.

[23]. Banchhor, Chitrakant, and N. Srinivasu. "Analysis of Bayesian optimization algorithms for big data classification based on Map Reduce framework." *Journal of big data* 8.1 (2021): 1-19.

[24]. Banchhor, Chitrakant, and N. Srinivasu. "Integrating Cuckoo Search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification." *Data & Knowledge Engineering* 127 (2020): 101788.

[25]. Ben Abacha, A., Demner-Fushman, D. A question-entailment approach to question answering. BMC Bioinformatics 20, 511 (2019). https://doi.org/10.1186/s12859-019-3119-4.

[26]. Dar Masroof, and Munishwar Rai. "A Novel Framework for Enhancing QoS of Big Data." *International Journal of Advanced Computer Science and Applications* 11.4 (2020).

[27]. Salim, Faizan, et al., "Consensus Algorithm for Healthcare Using Blockchain." *Digitization of Healthcare Data Using Blockchain*, pp. 93-116, 2022.

[28]. Naqushbandi, Faizan Salim, and A. John, "Sequence of Actions Recognition Using Continual Learning." *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. IEEE, 2022.

[29]. Himeur, Yassine, et al. "AI-big data analytics for building automation and management

systems: a survey, actual challenges and future perspectives." Artificial Intelligence Review 56.6 (2023): 4929-5021.

[30]. Jagadeesan, J., and D. Nancy Kirupanithi. "An Optimized Ensemble Support Vector Machine-Based Extreme Learning Model for Real-Time Big Data Analytics and Disaster Prediction." Cognitive Computation (2023): 1-23.