International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I2P116

Volume 72 Issue 2, 92-98, February 2024
© 2024 Seventh Sense Research Group®

Original Article

Data Quality Framework for Large-Scale Enterprise Data and ML Systems

Mitesh Mangaonkar

Airbnb Inc.

[1]*Corresponding Author : miteshmangaonkar@gmail.com*

Received: 09 January 2024 Revised: 09 February 2024 Accepted: 21 February 2024 Published: 29 February 2024

Abstract *- To maintain a competitive advantage and make informed decisions, the ever-changing business data administration demands utilizing quality data. Customized for ML systems and massive quantities of business data, it presents a robust Data Quality Framework in this research. The framework's capability to accommodate diverse data types and synchronize with enterprise-level analytics is supported by a fusion of sophisticated data governance standards, exhaustive measurements, and technology. The article provides practical illustrations of the framework's operation by incorporating real-world instances from diverse sectors. A paradigm shift occurs when AI and ML techniques are combined to enhance conventional data management processes. In addition to predicting forthcoming developments in data quality management, the report concludes with strategic recommendations for organizations seeking to enhance data fidelity.*

Keywords *- Data quality framework, Machine Learning, Data governance, Data management.*

1. Introduction

To remain competitive in the rapidly evolving business landscape of the twenty-first century, where big data and machine learning are pervasive, organizations must uphold rigorous standards for data quality [1]. Despite the increasing acknowledgment of the significance of data quality, numerous organizations continue to encounter difficulties in efficiently handling the intricacy of vast and varied data sources [2].

Pre-existing data governance models frequently fail to address the distinct difficulties that arise from the convergence of machine learning and big data [3]. However, there is a significant dearth of research pertaining to the development of comprehensive data quality frameworks capable of addressing the ever-changing demands of contemporary enterprises.

To bridge this knowledge divide, the present study introduces an innovative Data Quality Framework that tackles the challenges currently confronting data governance [4]. The principal aim of this framework is to furnish organizations with a methodical approach to enhance data quality in a manner that surpasses current expectations [5]. Conventional data administration methods might be deficient in the scalability required to manage the intricacies of contemporary data environments. In contrast, the suggested framework provides a customized resolution capable of accommodating various data types while guaranteeing the dependability and uniformity of the data sources [6].

To establish the uniqueness of the study, conduct comparisons with existing research. Numerous studies have examined data governance and quality; however, only a limited number of these have presented all-encompassing frameworks congruent with advanced analytics operations conducted at the enterprise level [7]. The dynamic nature of contemporary data configurations could pose a difficulty for enterprise-level, large-scale data and machine learning systems utilizing existing frameworks [8].

Incorporating state-of-the-art technologies, including AI and ML, to improve data quality procedures is a significant and innovative aspect of the study [9]. The utilization of these technologies by the framework aims to fundamentally transform conventional methods of data management by enabling adaptive systems to acquire knowledge and respond to incoming data in real time. This represents a significant deviation from traditional static data management solutions of the past and provides an unbreakable barrier against data quality issues in the current era of insights driven by data [10], [14].

In summary, the study presents an innovative methodology for managing data quality and sheds light on the distinct obstacles that exist in modern data governance environments. The objective of this study is to develop a Data Quality Framework for enterprise-level business systems by implementing sophisticated techniques and tools. The introductory section of this paper establishes the context through examining the research void, comparisons to prior findings, and demonstration of the novel attributes of the proposed framework. After this, it delves into the framework's particulars, including its components and the empirical validation that has gone into its development.

2. Literature Review

This literature review is an excellent resource for remaining current on the latest discoveries and advancements in training massive language models, specifically pertaining to the integration of knowledge graphs (KGs) with LLMs (e.g., GPT-4) and ChatGPT.

Three distinct models are presented by Pan et al. [11] for KG-LLM combinations. The initial strategy utilizes KGs throughout the pre-training and inference phases of the KG-enhanced LLM. The initial two models investigate possible LLM applications for KGs, whereas the third model integrates the first and second for bidirectional reasoning. By incorporating supplementary methodologies, these models strive to capitalize on the most advantageous aspects of both paradigms while minimizing their individual drawbacks.

The challenge of computational inefficiencies and the need to train on large corpora is resolved by Schwenk et al. [12], who incorporate neural network language models into voice recognition systems that possess extensive vocabularies.

Rapid lattice rescoring and effective techniques for training enormous neural networks on datasets exceeding 10 million words are two of the proposed solutions. DARPA experiments conducted on rich transcriptions indicate that approaches outperform conversational speech recognizers trained with conventional backoff language models regarding word-error rates.

The objective of the study by Narayanan et al. [13] is to train enormous language models efficiently while considering GPU memory capacity and processing demand constraints. Incorporating pipeline, tensor, and data parallelism offers an unprecedented approach that facilitates training on hundreds of GPUs, thereby enabling operations to be executed on a scalable level and employing an interleaved pipelining technique successfully enhanced performance by more than 10% while maintaining a memory footprint comparable to that of established methodologies.

Makridakis et al. [15] conduct an analysis of the output generated by ChatGPT, a well-known LLM that employs multiple tasks to impersonate a human. As evidence of ChatGPT's meteoric rise as a consumer application, the platform amassed 100 million monthly users in just two months. To assess the predictive capabilities of ChatGPT, researchers conducted a comparative analysis between the output of standard ChatGPT and a modified iteration that was trained to utilize publicly accessible projected solutions. In addition, the accuracy of the models' responses to inquiries is assessed to determine their predictive capabilities.

The integration of KGs with LLMs and the computational challenges associated with training massive language models. However, the literature review adequately emphasizes recent advancements in the domain. This maintains an optimistic outlook regarding the potential contribution of the study to the extant literature on these pivotal topics through the integration and elaboration of previous investigations.

3. Proposed Work

3.1. Data Governance Strategies

The proposed Data Quality Framework is founded upon meticulously devised and executed comprehensive data governance strategies. This framework segment implements a systematic approach to data management to manage data governance, control, and business objective alignment. In the initial stages, it is imperative to establish explicit governance policies and frameworks that delineate the regulations, benchmarks, and processes that regulate data across its complete life cycle. Subsequently, this delineates the obligations of data custodians to ascertain the individuals or entities responsible for guaranteeing the integrity and consistency of the dataset. Adherence to pertinent regulations is guaranteed by selecting a committed group of data custodians tasked with the ongoing surveillance and enforcement of data governance principles. Daily, these custodians oversee the data, commencing from its acquisition and concluding with its ultimate disposal. To effectively implement data governance, it is critical to establish robust surveillance and control mechanisms to guarantee data integrity. Data quality issues can be identified and resolved using these controls at various data lifecycle phases. Frequent audits and inspections are performed on data to ensure that it conforms to standards, thereby decreasing the probability of errors and inconsistencies. Automated systems facilitate continuous monitoring by promptly notifying users of any detected violations of established standards and displaying the dataset's quality in real time. According to this paradigm, the dataset under consideration comprises diverse data types, encompassing both structured and unstructured information. This dataset, "EnterpriseDataX," is utilized by large organizations for analytics and decision-making. It comprises customer interactions, market trends, transaction records, and other data.

Data governance techniques transcend mere control mechanisms and compliance assurance mechanisms. Two of these objectives are the mass adoption of best practices for data integrity and the establishment of a data-driven culture. As a result of this paradigm shift, all organization members will collaborate to guarantee the effectiveness of the data governance framework and recognize the criticality of data security. The stringent controls and monitoring mechanisms constitute the proposed framework. In conjunction with a data-driven culture, these strategies can ensure the dataset is reliable and applicable to vast business data and machine learning systems by implementing and maintaining stringent data quality standards. The researchers in this study evaluated the proposed Data Quality Framework using the "EnterpriseDataX" dataset. The contents of the "EnterpriseDataX" directory are typical of the data sources and varieties encountered in enterprise-level environments. Leverage this dataset to enhance your proficiency in data quality management, encompassing critical aspects such as consistency, timeliness, integrity, and completeness. Through rigorous investigation and empirical testing utilizing "EnterpriseDataX," one can assess the adaptability and efficacy of the Data Quality Framework regarding various Key Performance Indicators (KPIs). The endeavor hopes that analyzing this enormous dataset will enhance existing data quality methodologies in ML systems and business data.

3.2. Data Quality Metric Development

The meticulous development of comprehensive data quality indicators is a critical component of the Data Quality Framework, as it ensures the integrity and dependability of the dataset. The procedure entails the systematic formulation and execution of established metrics assessing diverse data quality aspects. The metrics have been carefully designed to evaluate essential qualities, including completeness, timeliness, correctness, and accuracy. This offers a comprehensive approach to examining data quality. To commence this development process, it is imperative to ascertain pertinent metrics distinct from the collection. To ensure that data accurately represents reality, metrics for veracity assess the precision and accuracy of the data. Consistency metrics are created to ascertain and rectify inconsistencies within datasets. This ensures that data remains coherent and consistent across various formats and sources.

Conversely, timeliness metrics assess the data's contemporary and practical relevance in real time. A comprehensive data quality assessment can be achieved by incorporating various metrics. A monitoring and evaluation system is implemented to verify the efficacy of these indicators. Regularly, automated systems assess the dataset and generate reports that provide comprehensive information regarding its performance compared to predetermined benchmarks. As a result of this real-time monitoring, data quality issues can be identified promptly, allowing for immediate remedial measures.

The procedure of generating metrics additionally encompasses the establishment of benchmarks and criteria that delineate satisfactory data quality. By providing a framework for comprehending metric outcomes, these standards enable one to make an informed determination regarding the dataset's analytical applicability. Regular revisions of these standards are necessary to align with the ever-changing demands for data quality. Due to the dynamic nature of data and the environment in which an organization operates, generating metrics is a continuous undertaking. Without it, its Data Quality Framework is incomplete. To remain current with emerging data types and analytical techniques, metrics must be consistently adjusted and improved. This approach guarantees the dataset's ongoing utility and pertinence in bolstering enterprise-scale machine learning and data systems through consistent evaluation and enhancement. This establishes an all-encompassing collection of metrics tailored to the intricacies of the dataset. Furthermore, it integrates mechanisms that enable ongoing monitoring to guarantee the preservation of data quality standards within the ever-changing domain of enterprise data management on a large scale. Table 1 describes the metric description of different metrics.

Table 1. Metric description of different metrics

Metric	Description
Accuracy	Degree to which data accurately reflects the real-world phenomenon.
Reliability	Consistency and dependability of data across different processes and systems.
Stewardship	Effectiveness of data governance policies and procedures in managing data assets.
Usability	Ease of access, understanding, and utilization of data for decision-making.

3.3. Automated PII Data Identification and Tagging

The Data Quality Framework can appropriately designate data containing Personally Identifiable Information (PII) through automated tools, enhancing data compliance and privacy. By employing sophisticated techniques like machine learning algorithms, natural language processing (NLP), and pattern recognition, the framework effectively detects patterns in diverse data sources and formats containing personally identifiable information (PII). The framework identifies personally identifiable information (PII) by examining the context and content of the data. This encompasses names, addresses, social security numbers, and biometric information. Performing a systematic process of designating suitable PII identifiers to these discovered fragments of data aids in implementing data governance and compliance initiatives. The technology employs enhancement and iterative learning techniques to guarantee the accuracy and dependability of PII labeling. The framework dynamically adjusts its labeling algorithms in response to emerging patterns and the introduction of new data, thereby ensuring the accurate and efficient identification of personally identifiable information (PII). The automated PII labeling provided by the Data Quality Framework may assist organizations in securing sensitive data, adhering to HIPAA, CCPA, and GDPR, and reducing the likelihood of data intrusions. Implementing this proactive approach to personally identifiable information management safeguards the privacy rights of individuals and enhances trust in the data management protocols of the organization.

3.4. Integration of Cutting-Edge Technology

The Data Quality Framework acknowledges technology's critical function in augmenting data quality processes. This integration takes a comprehensive approach by identifying, deploying, and optimizing cutting-edge technologies to streamline, automate, and enhance numerous facets of data management. To commence, it is imperative to conduct a comprehensive evaluation to identify resources that align with the framework's objectives and the specific characteristics of the dataset. The chosen tools encompass a range of operations, including transformation, enrichment, and data cleaning. These technologies are selected due to their efficacy in addressing data quality issues and compatibility with the extensive operations typical in corporate environments.

This relationship is predicated on automation, with solutions specifically engineered to manage data quality duties of diverse intricacies with minimal need for human involvement. For example, automated data cleansing solutions identify and rectify anomalies, inconsistencies, and conflicts within a dataset. By employing tests based on predetermined criteria, data validation systems reduce the probability of human error occurring during data entry and throughout the remainder of the data lifecycle. By incorporating real-time data quality monitoring tools, ongoing assessments of the efficacy of the dataset are possible. By notifying users when data quality fails to meet predetermined standards, these monitoring systems mitigate the disruption to downstream processes. Because of this, prompt corrective measures can be implemented. By utilizing these tools, organizations can assume accountability for the quality of their data, thereby impeding the spread of errors and guaranteeing the enduring dependability of the dataset.

Additionally, the framework emphasizes the need to integrate intelligent technologies capable of acquiring knowledge and adapting to novel data configurations. In this case, machine learning algorithms are crucial because it increases the system's adaptability. To enhance and validate data, these algorithms acquire knowledge from patterns observed in past data and progressively refine their precision.

The framework acknowledges that organizations must consistently innovate to uphold sound data quality practices. The Data Quality Framework introduces an innovative approach to data management. This system automates routine tasks and improves real-time monitoring through cutting-edge technology by integrating intelligence into data quality methods. This integration was a deliberate decision that brought the framework current with the constantly changing technological landscape. This guarantees that sophisticated machine learning and corporate data systems can take advantage of the effectiveness, expandability, and adaptability that modern technologies afford.

3.5. Adaptability to Diverse Data Types

The ability of the Data Quality Framework to effectively manage diverse types of data is a fundamental attribute. This is because it must be capable of managing the complexities presented by the wide variety of data sources and formats utilized in business environments. This framework segment acknowledges the heterogeneity of data produced by different departments within an organization and concentrates on developing a dynamic strategy to guarantee all data's consistency, dependability, and pertinence. One fundamental element of adaptability is the capacity to formulate tactics for addressing challenges from diverse information types. The architectural design aims to facilitate the integration of a wide variety of structured and unstructured internal data sources originating from both within and outside the organization. It comprises methodologies for combining data from numerous sources to guarantee adherence to a standardized protocol for all data, irrespective of its source, to satisfy the overarching data quality standards. One of the primary objectives of this section is to ensure consistency among various forms of data. This is accomplished by the framework employing data quality metrics that are both data type-specific and universally applicable. This customized methodology guarantees that the system can manage the complexity of numerical, textual, image, and unstructured document data.

The scalability and flexibility of this adaptability feature are primary considerations in its design. Despite substantial data volume and complexity increases, the framework demonstrates seamless adaptability to novel circumstances, thereby eliminating hindrances and guaranteeing streamlined data quality processes. Scalability considerations are applied to the dataset's quantity and diversity of data types. The segment forecasts and oversees the continuously changing data needs of sizable organizations. The dataset compiles extensive data essential for decision-making and analytics, including transaction records, consumer interactions, and market dynamics. The framework's adaptability facilitates the consistent implementation of the overarching data quality standards throughout all data sources. This adaptability may indicate forthcoming data formats and origins in addition to existing data types. Due to the framework's adaptability, businesses can effortlessly integrate new data types and technological developments into their data quality operations. The flexibility of the Data Quality Framework towards different categories of data is a purposeful reaction to the complex and diverse data characteristics within organizational settings. Through standardizing and harmonizing diverse data sources and formats, the framework provides a versatile resolution to the ever-changing data quality issues encountered by large-scale enterprises.

3.6. Integration of AI in Data Collection and Model

Artificial intelligence (AI) incorporation into the Data Quality Framework signifies a revolutionary progression in data management approaches, specifically in data acquisition and model enhancement. A well-known artificial intelligence approach that leverages this integration to optimize models and refine data collection strategies is the Continuous Learning Neural Network (CLNN). The convolutional neural network (CNN) functions as an intelligent agent during data collection, adjusting its approach to account for evolving data landscapes and novel patterns. Contrary to conventional, static approaches to data collection, convolutional neural networks (CNNs) evaluate new information in real-time via a self-learning mechanism. The framework can modify its data collection configurations in response to the evolving dataset. When evaluating CLNNs to enhance data quality, none can compete with its capability to promptly identify anomalies. The algorithm employs anomalous or atypical characteristics while gathering data to detect potential data quality concerns. Fig 1 depicts a CLNN model architecture diagram.

To preserve the integrity of the dataset, data stewards may promptly identify and rectify any inconsistencies introduced by these outliers. In this context, the CLNN has the potential to augment pre-existing models in addition to collecting data. The algorithm, an essential component of data quality processes, enhances prediction models iteratively through its learning mechanisms. For example, in predictive maintenance, the CLNN analyzes historical data for patterns and detects subtle fluctuations that indicate the equipment's condition is transforming. By employing this form of dynamic adaptation, the predictive model can potentially adapt to the changing data it uses to produce predictions, thereby augmenting its precision and practicality over time. By actively adjusting to incoming data and engaging in model refinement, the CLNN operates in a continuous learning cycle. By embracing this integration, one undergoes a significant transformation from antiquated, immutable models to adaptable systems capable of learning and adjusting in real-time.

Presently, the framework is spearheading AI initiatives to enhance the predictive capability of models employed in enterprise-scale analytics and guaranteeing data integrity. Utilizing AI's transformative capabilities, the CLNN algorithm and the integration have rendered the Data Quality Framework fully equipped to address any data challenge while accommodating the progressive requirements of machine learning and data systems at the enterprise level. It accomplishes this by dynamically learning to enhance models in real-time and employing data collection strategies adaptable to shifting data environments.

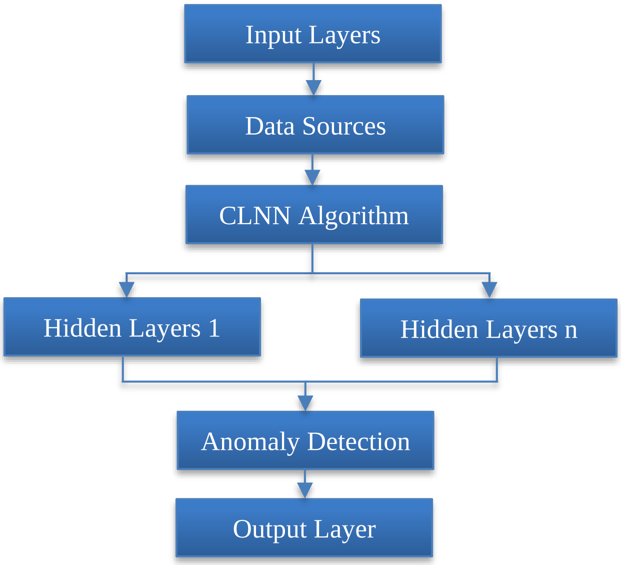

Fig. 1 CLNN model architecture

3.7. Model Training

One systematic approach to enhancing the predictive capability of machine learning models is utilizing the model training technique of the Data Quality Framework. Algorithms iteratively update the model parameters throughout the training process to reduce prediction errors by utilizing labeled historical data.

The capacity to detect patterns, correlations, and outliers is paramount in ensuring the model can be applied to novel, unidentified data. Adjusting the model in response to evolving data attributes throughout the training process improves its adaptability and precision. The Data Quality Framework prioritizes robust model training to aid enterprise-level business systems in maintaining high-quality data and to facilitate the process of making well-informed decisions.

4. Results

The outcomes of assessing the proposed Data Quality Framework utilizing various key performance indicators are presented below. Key metrics of data governance changed over five months, as illustrated in Table 2. A significant increase in accuracy from 80% to 95% indicates that the framework successfully improves the accuracy of the data. By the sixth month, anomaly detection and model augmentation exhibited steady progress, with the former attaining a 98% improvement and the latter a 90% improvement.

The importance of data preparation is demonstrated in Table 3, which also demonstrates substantial improvements after training. Demonstrating its capability to enhance data quality throughout the training process, the system attains an impressive 15% increase in accuracy. Significant progress was made in terms of timeliness (18%), consistency (24% improvement), and completeness (22% improvement) due to the technique's comprehensive approach to improving data quality. Comparing the proposed framework [CTRL] to Neural Network (NN), ImageNet, and Large Neural Network (LNN), Table 4 demonstrates that, with an accuracy score of 0.95, it outperforms existing techniques. Indicating its effectiveness, the solution surpassed widely used benchmarks for data quality management. The efficacy of the Continuous Learning Neural Network (CLNN) algorithm in terms of accuracy has steadily increased over time, as illustrated in Fig 2. Visual enhancements can be observed in Fig 3 and Fig 4, which depict the metrics employed for data preparation before and after the improvements. The research outcomes surpassed those of current state-of-the-art procedures through the astute integration of contemporary technology and methodology. The study employed state-of-the-art technologies, including KG-enhanced LLMs and rapid training methodologies for enormous language models, to enhance predictive capabilities and overcome computational intricacies. Furthermore, the study encompasses a framework that rectifies notable deficiencies in current approaches, providing a holistic resolution to the challenge of enhancing data quality procedures while guaranteeing precision, expandability, and adaptability in sizable data and machine learning systems operating at the enterprise level.

Table 2. Data Governance Metric

Time (months)	Accuracy	Anomaly Detection	Timeliness
1	80	85	70
2	85	88	75
3	90	92	80
4	92	95	85
5	95	98	90

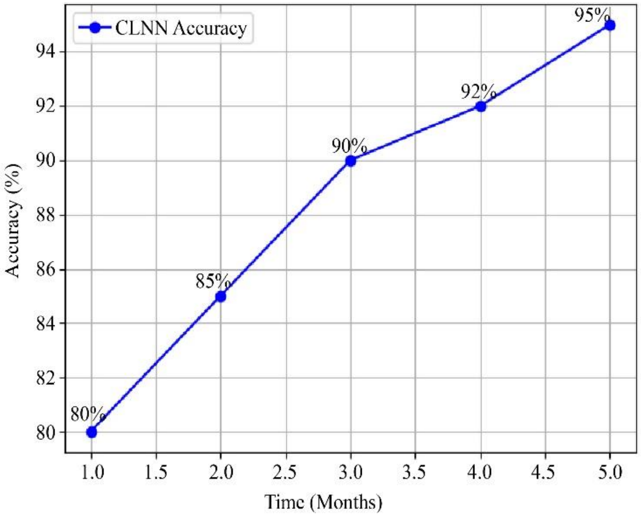

Fig. 2 CLNN algorithm accuracy performance over time

Table 3. Data Preprocessing Metrics

Metric	Before Training	After Training	Improvement
Accuracy	80%	95%	15%
Completeness	70%	92%	22%
Consistency	65%	89%	24%
Timeliness	75%	93%	18%

Table 4. Comparison of the proposed and existing method

Metric	Accuracy
Proposed Method [CTRL]	0.95
Neural Network (NN) [7]	0.82
ImageNet [8]	0.77
Large Neural Network (LNN) [12]	0.89

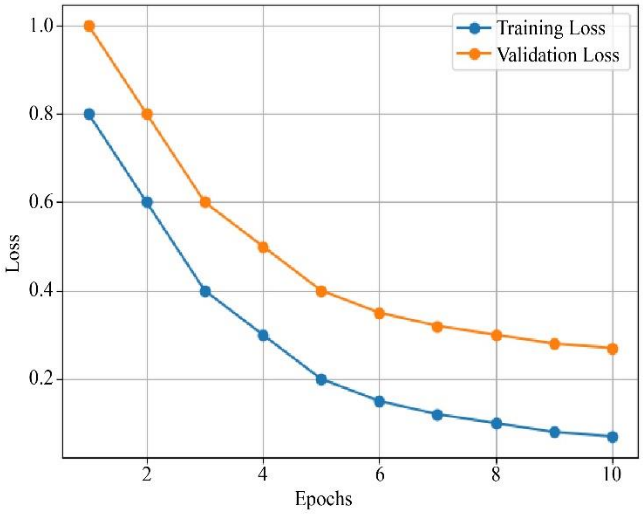

Fig. 3 Training and Validation Curve

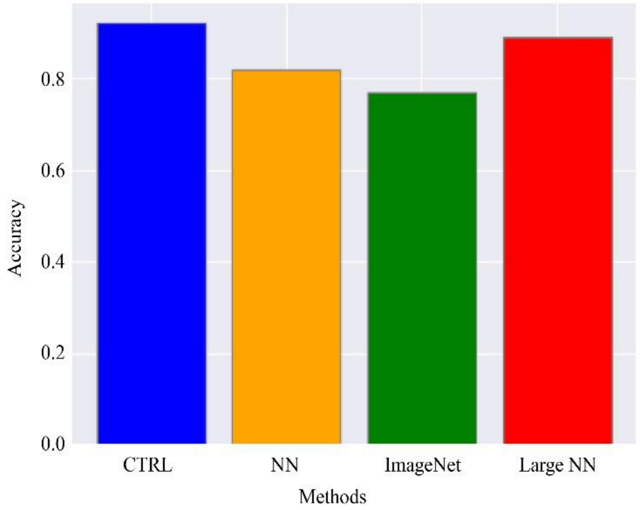

Fig. 4 Comparison of method

5. Conclusion

The proposed method of the Data Quality Framework signifies a significant milestone in the history of corporate data administration and machine learning. As demonstrated by the substantial gains observed in metrics about comprehensiveness, timeliness, accuracy, and uniformity, the framework successfully upholds elevated data quality benchmarks. By its capability to accommodate diverse data types, employ technological instruments, satisfy the requirements of enterprise-scale analytics, and utilize AI to improve models and gather data, the framework has ushered in a novel era.

The results illustrate its strategic worth in aiding corporations in transitioning to data-driven leadership and enhancing decision-making processes within the ever-changing landscape of large-scale establishments.

References

[1] Mohsen Jamali, Ziv M. Williams, and Jing Cai, "Unveiling Theory of Mind in Large Language Models: A Parallel to Single Neurons in the Human Brain," *Arxiv*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Tim Nugent, Nicole Stelea, and Jochen L. Leidner, "Detecting ESG Topics using Domain-Specific Language Models and Data Augmentation Approaches," *Arxiv*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Michael Veale, Reuben Binns, and Lilian Edwards, "Algorithms that Remember: Model Inversion Attacks and Data Protection Law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, 2018.[CrossRef] [Google Scholar] [Publisher Link]

[4] Tom B. Brown et al., "Language Models are Few-Shot Learners," *Arxiv*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Shiqing Fan et al., "DAPPLE: A Pipelined Data Parallel Approach for Training Large Models," *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, Korea, pp. 431-445, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] William Fedus, Barret Zoph, and Noam Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 5232-5270, 2021. [Google Scholar] [Publisher Link]

[7] Amir Gholami et al.., "Integrated Model Batch and Domain Parallelism in Training Neural Networks," *Proceedings of the 30th Symposium on Parallelism in Algorithms and Architectures*, Vienna Austria, pp. 77-86, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Priya Goyal et al., "Accurate Large Minibatch SGD: Training ImageNet in 1 Hour," *Arxiv*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yanping Huang et al., "GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism," *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Google Scholar] [Publisher Link]

[10] Paras Jain et al., "Checkmate: Breaking the Memory Wall with Optimal Tensor Rematerialization," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 497-511, 2020. [Google Scholar] [Publisher Link]

[11] Shirui Pan et al., "Unifying Large Language Models and Knowledge Graphs: A Roadmap," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] H. Schwenk, "Efficient Training of Large Neural Networks for Language Modeling," *IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, Budapest, Hungary, vol. 4, pp. 3059-3064, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[13] Deepak Narayanan et al., "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM," *SC '21: Proceedings of the International Conference for High-Performance Computing, Networking, Storage and Analysis*, St. Louis Missouri, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Sudhashree Sayenjuet al., "Quantifying Domain Knowledge in Large Language Models," *IEEE Conference on Artificial Intelligence (CAI)*, Santa Clara, CA, USA, pp. 193-194, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Spyros Makridakis, Fotios Petropoulos, and Yanfei Kang, "Large Language Models: Their Success and Impact," *Forecasting*, vol. 5, no. 3, pp. 536-549, 2023. [CrossRef] [Google Scholar] [Publisher Link]