



INTERNATIONAL RESEARCH JOURNAL OF HUMANITIES AND INTERDISCIPLINARY STUDIES

(Peer-reviewed, Refereed, Indexed & Open Access Journal)

DOI : 03.2021-11278686

ISSN : 2582-8568

IMPACT FACTOR : 8.031 (SJIF 2025)

Data Quality Awareness: A Shift from Traditional Data Management to Data Science Systems

Dr. Vyankat Vishnupant Munde

Assistant Professor,
Department of Computer Science,
Vaidyanath College,
Parli-V, Dist. Beed (Maharashtra, India)
E-mail: mvv.parli@gmail.com

DOI No. **03.2021-11278686** DOI Link :: <https://doi-ds.org/doi/10.2025-68237197/IRJHIS2501006>

Abstract:

Artificial intelligence (AI) has revolutionized a number of industries and profoundly affected our day-to-day existence. The success of AI is largely dependent on high-quality data. The progression of data quality (DQ) awareness from conventional data management systems to contemporary data-driven AI systems—which are essential to data science—is thoroughly reviewed in this work. We summarize the body of research, emphasizing the quality issues and methods that have developed from traditional data management to data science, which includes big data and machine learning. Although data science solutions facilitate a variety of tasks, in this study we specifically address the analytics component powered by machine learning. To provide a more comprehensive knowledge of growing DQ difficulties and the associated quality awareness strategies in data science systems, we leverage the cause-effect relationship between the quality challenges of big data and machine learning. We believe this work is the first to review DQ awareness across both classic and emerging data science platforms. We hope that this exploration of the development of data quality awareness will be useful and enlightening to readers.

Keywords: Data Quality (DQ), Big Data, Data Science, Machine Learning (ML)

Introduction:

Data quality (DQ) is crucial in data-centric technologies like big data and machine learning, as it reflects the growing importance of data-centric technologies. Traditional data management methods still have limitations in improving DQ awareness in data-driven AI systems. These systems focus on scalable and adaptive quality standards, adapting to changing data landscapes and user needs. The shift from traditional data management to big data and machine learning reflects a changing environment of data quality challenges.

Old data management systems often oversimplified data quality issues, leading to a proliferation in big data contexts due to various data types, large datasets, and real-time, time-evolving data. These factors, including the need for new data quality dimensions and evaluation techniques, can impact the accuracy and insights gained from these systems. Two important areas where the big data world poses data quality management issues are delineated by Saha and Srivastava [23].

This paper discusses data quality (DQ) issues in machine learning (ML) systems, focusing on conventional systems and large data systems. It examines DQ techniques, new features, and approaches, and discusses quality awareness in ML pipelines. The study concludes with key findings and suggestions for further research to improve data quality awareness in modern data science platforms.

Essential Concept of quality in data:

The paper explores data quality issues in machine learning systems, focusing on both conventional and large data systems. It explores techniques, new features, and approaches, and suggests further research for improving data quality awareness [22]. DQ Dimensions and DQ Metrics are crucial tools for assessing data quality, focusing on factors like timeliness, accuracy, completeness, and consistency, thereby ensuring data serves its intended purpose effectively [4, 30]. The process of systematically comparing data to relevant quality characteristics to determine whether or not it is adequate for a given purpose is known as data quality assessment. By implementing corrective measures based on assessment results, data quality enhancement aims to raise the standard for data quality in order to meet or exceed new targets. Typically, there are two main types of improvement initiatives: process-driven and data-driven. One data-driven method that immediately modifies data values is data cleaning. Process-driven techniques focus on redesigning data creation or modification processes, such as adding steps for data format validation.

Traditional data quality management:

1. Quality Awareness from:

DQ Dimensions and DQ Metrics are essential for assessing data quality, focusing on timeliness, accuracy, completeness, and consistency. Data quality enhancement involves corrective measures, either process-driven or data-driven, with data cleaning being a data-driven method.

The Data Point of View (responsiveness, accessibility).

The quality awareness data viewpoint focusses mostly on data characterisation. Dimensions are used to define various data properties [26]. Literature has separated the most common DQ aspects into four groups: accessible quality, contextual quality, representational quality, and intrinsic quality [31]. The quality dimensions of services are contextual, representational, and intrinsic, depending on context and internal characteristics of the data. Quality dimensions are divided into

two groups: those that provide declarative explanations of services (accuracy, completeness, timeliness) and those helpful for perceptual assessment (responsiveness, consistency, and accessibility).

1.1 Quality Awareness from the Viewpoint of the User:

Quality-aware query processing approaches, which involve query language extensions and adaption, ensure data quality meets user requests and preferences. They allow for explicit definition of quality measurements and limitations across various quality factors, ensuring data quality.

A SQL extension was proposed in [34] to employ hierarchical prioritization to indicate user preferences for data quality. Quality views (QV), which are domain-specific quality restrictions, are included into query processing as part of an XQuery upgrade in [In]. By connecting quality contracts to data sources, the data quality profiling idea in [5] may be implemented. Establishing a quality profile. According to [34], conditional DQ profiling connects the quality of specific data attributes to user query circumstances. To ensure that the quality of data meets user demands and preferences, quality-aware query processing techniques—which mainly involve query language extensions and query processing adaptation—are frequently used.

2. Users Perspective Dimensions:

Big data applications require quality criteria based on user viewpoints, considering context-dependent and effective approaches for data assessment. Effective evaluation involves analyzing quality characteristics relevant to user interactions, including presentation quality, usefulness, relevance, availability, and reliability.

2.1 Big data qualities affect big data characteristics:

Big data applications require quality criteria based on user viewpoints. Two main approaches for assessing data are analyzing the relationship between big data characteristics and data quality dimensions, identifying specific dimensions, and evaluating them. However, there is still a gap between big data characteristics and data quality aspects that needs to be addressed to understand applications. The quality and value of data are conceptually different due to the impact it has on conclusions.[2]. Previous studies looked at the connection between data quality and big data, especially for companies in the financial sector, in an effort to close the gap between quality dimensions and Big Data Characteristics (BDC) [29] suggested that data variety in the financial sector, where several data sources are employed, is the most important BDC impacting the majority of DQ characteristics, including correctness, consistency, security timeliness, and completeness. Additionally, it was suggested that velocity and timeliness are often linked due to the relationship between timely use of data and the pace at which it is created and processed.

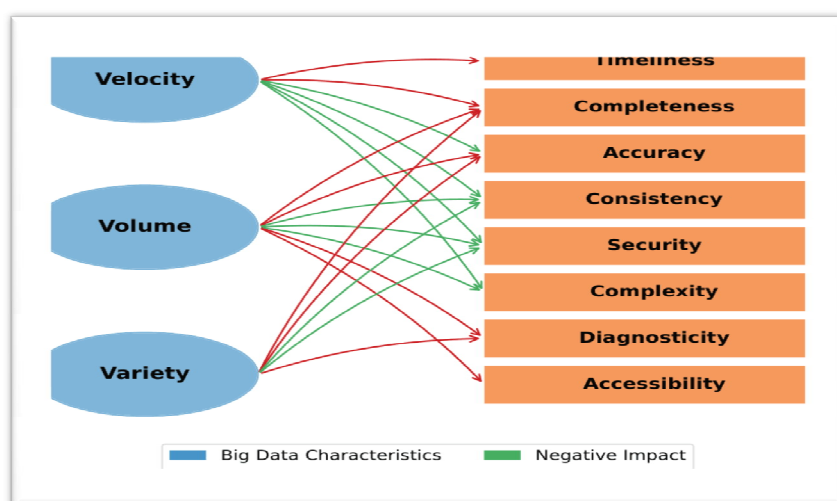


Figure 1: The impact of the big data characteristics on data quality dimensions

2.1.1 Parallel Computing.

The study reveals that contextual data quality (DQ) dimensions, such as timeliness and accessibility, have the strongest link with Big Data Characteristics (BDCs) for user applications. However, existing studies have issues, as DQ factors beyond data security and accessibility may affect big data analytics use. The narrow focus on user validation makes generalization challenging.

2.2.2. Sketching and sampling methods.

Sampling techniques can reduce the time required to compute data quality since they enable findings to be estimated. Frequently used strategies include stratified sampling, cluster sampling, systematic sampling, simple random sampling, and reservoir sampling [6,11,17,25,28]. These techniques help determine sample sizes and sample selection while considering the properties of the data source and relevant quality dimensions in order to accurately evaluate quality metadata.

2.2.3 Techniques for data fusion and integration.

These techniques integrate data from several sources to create a coherent dataset. Social media, internal corporate databases, and Internet of Things devices are just a few of the sources of data that may be found in a big data environment. It may also vary in quality, structure, and organisation [15]. Inconsistencies across various databases, such as date differences, incompatibilities in data formats, and potential data conflicts or duplication, are addressed via data fusion and integration [9].

2.2.3 Gradual Evaluation of Big Data Quality.

Data profiling is challenging due to the rapid expansion of data, making it difficult to assess quality. Effective data profiling methods should manage growing data without requiring a complete dataset reprofile. Continuous and incremental profiling are suggested to improve data calculation, updating data as it is input.

3 Quality consciences in the new data pipeline:

Data analytics in big data relies on data quality for precise insights. Machine learning uses these insights to create predictive models and automate operations. Emerging systems require an iterative approach to address issues like label quality, data drift, imbalance, and inaccurate data entry.

3.1 Quality dimensions based on machine learning:

The full lifespan is covered by DQ dimensions designed for ML pipelines, from data preparation to model training and monitoring. Based on a number of criteria, we examine the literature in this part on the dimensions of data quality in machine learning (ML) pipelines [8,18, 19,20]. Although these studies provide useful classifications of data quality dimensions, their frameworks do not highlight the causal relationship between the big data and machine learning quality challenges. This would enable a more comprehensive understanding of the new data quality issues in data science systems, especially in machine learning pipelines.

3.1.1 Dimensions of data.

Each component of machine learning uses the data under consideration as input. [27] separates serving data from training data. Training data is the dataset used to train an ML component during the development phase, while serving data is the dataset used to produce real-time predictions during the deployment phase.

3.1.2. Dimensions based on models.

The quality of an ML model is influenced by various factors including the type of model, data used for training, evaluation of developed artifacts, specific task being addressed, and data separation for training and validation.

3.2 Data Cleaning and Validation.

Data validation techniques verify that the data satisfies certain criteria [10] and spot anomalies before they have an impact on the model's performance. One fundamental approach is to use descriptive statistics, which calculate measures such as mean, median, variance, and standard deviation to summarise the shape, dispersion, and central tendency of the data distribution [12]. In this area, robust statistical methods have recently been developed that improve the detection of outliers and abnormalities in large datasets [24]. Schema validation ensures that the data adheres to a predefined schema by checking for consistency in data types, ranges, and formats [3].

ML Stages	Techniques	Quality Dimensions in the ML Pipeline				
		Data-based Dimensions	Model-based Dimensions	Process-based Dimensions	Use Case/Context-based Dimensions	Stakeholder-based Dimensions
Data Preparation	Sampling	Representativeness, Balancedness			Contextual Relevance	
	Data Validation	Correctness, Completeness				
	Data Cleaning	Correctness, Completeness				
	Data Imputation	Correctness, Completeness, Intra-Consistency				
	Data Labeling	Correctness, Absence of Bias				
Model Training	Feature Engineering	Representativeness, Train/Test Independence, Balancedness	Performance, Model Complexity		Use Case Specificity	
	Fairness and Bias Checking	Absence of Bias	Fairness			Ethical Alignment
	Data Augmentation	Balancedness	Fairness		Use Case Specificity	
Model Validation & Monitoring	Data Drift Detection	Currentness	Robustness	Real-time Performance		
	Monitoring Model		Performance, Scalability, Trust	Reliability, Documentation Quality, Auditability, Reproducibility	Trust	Transparency

Table 1: Classification of Quality Dimensions and Techniques Across ML Pipeline Stages

In order to guarantee data quality throughout the training and deployment stages, Google's TensorFlow Extended (TFX) [21], which is used in Google Play's recommendation algorithms, provides schema inspection and anomaly detection capabilities. Supervised learning relies on good labelled data, with advanced tools automating data validation and labelling. Monitoring machine learning is crucial, with serving data serving data being the main focus for quality improvement, ensuring contextual similarity to training data.

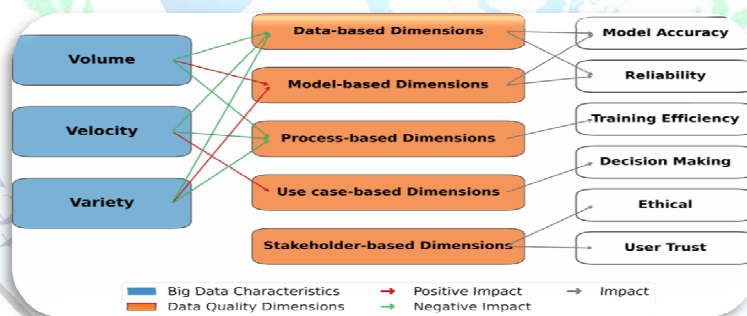


Figure 2: DQ impact in ML Pipelines

3.3 Impact of Data Quality in ML Pipelines:

The quality of machine learning (ML)-based data can impact various stages of ML pipelines, intermediate processes, and decision-making results. It affects data preparation and modelling, leading to improved conclusions and forecasts, impacting technical performance and decision-making.

4 The DATA Science system offers new opportunities for data quality awareness.

There are several opportunities to enhance quality awareness techniques since big data and machine learning in data science systems provide unique data quality challenges. This section looks

at potential research directions that might result from these challenges, focusing on the impact of large language models (LLMs) .

4.1 Improving Quality Awareness in DS Systems:

Multimodal Data Difficulties Current data quality algorithms struggle to manage the different properties of multimodal data, such as the distinctions between discrete text data and continuous picture data [16]. Synchronization issues with time-dependent data formats complicate model performance. Future research should provide consistent quality indicators and cross-modal consistency tests for multimodal settings [31]. Investigating the intersection of data cleaning and fairness in machine learning pipelines is crucial. Large language models (LLMs) are essential for data science systems to extract insights and make data-driven choices.

5. Conclusion:

This study examines how data quality awareness evolved from traditional data management to contemporary data science, with an emphasis on big data and machine learning contexts. Managing massive volumes, rapid data intake, and a wide range of data types are some of the main problems we have seen with big data, necessitating flexible and scalable quality assessment. Additionally, the significance of data quality in ML pipelines was explored, demonstrating its impact on model reliability and accuracy. We discussed ways to maintain the integrity of the model, such as data augmentation, real-time validation, and drift detection. We also looked into new opportunities in data quality for data science.

References:

- [1] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2017. Data profiling: A tutorial. In Proceedings of the 2017 ACM International Conference on Management of Data. 1747–1751.
- [2] Serge Abiteboul, Xin Luna Dong, Oren Etzioni, Divesh Srivastava, Gerhard Weikum, Julia Stoyanovich, and Fabian M. Suchanek. 2015. The elephant in the room: getting value from Big Data. In WebDB. ACM, 1–5.
- [3] Serge Abiteboul, Richard Hull, and Victor Vianu. 1995. Foundations of databases. Vol. 8. Addison-Wesley Reading.
- [4] Carlo Batini and Monica Scannapieco. 2006. Data Quality: Concepts, Methodologies and Techniques. <https://doi.org/10.1007/3-540-33173-5>
- [5] Laure Berti-Equille. [n. d.]. Quality-Adaptive Query Processing Over Distributed Sources. In 9th International Conference on Information Quality (IQ).
- [6] P. J. Bickel and D. A. Freedman. 1984. Asymptotic Normality and the Bootstrap in Stratified Sampling. The Annals of Statistics 12, 2 (1984), 470 – 482.

- [7] Li Cai and Yangyong Zhu. 2015. The challenges of data quality and data quality assessment in the big data era. *Data science journal* 14 (2015).
- [8] Giordano d'Aloisio, Antiniscia Di Marco, and Giovanni Stilo. 2022. Modeling Quality and Machine Learning Pipelines through Extended Feature Models. In
- [9] Giordano d'Aloisio, Antiniscia Di Marco, and Giovanni Stilo. 2022. Modeling Quality and Machine Learning Pipelines through Extended Feature Models. arXiv:2207.07528
- [10] Xin Luna Dong and Divesh Srivastava. 2013. Big data integration. In 2013 IEEE 29th international conference on data engineering (ICDE). IEEE, 1245–1248.
- [11] Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, and et al. 2021. Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. arXiv:2108.05935 [cs.LG]
- [12] Alon Halevy, Anand Rajaraman, and Joann Ordille. 2006. Data integration: The teenage years. In *Proceedings of the 32nd international conference on Very large data bases*. 9–16.
- [13] R. H. Henderson and T. Sundaresan. 1982. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization* 60, 2 (1982), 253 – 260.
- [14] Peter J Huber and Elvezio M Ronchetti. 2011. *Robust statistics*. John Wiley & Sons.
- [15] Vimukthi Jayawardene, Shazia Sadiq, and Marta Indulska. 2013. The curse of dimensionality in data quality. (2013).
- [16] Vimuthki Jayawardene, Shazia Sadiq, and Marta Indulska. 2015. An analysis of data quality dimensions. (2015).
- [17] Nawsher Khan, Ibrar Yaqoob, Ibrahim AbakerTargio Hashem, Zakira Inayat, and et al. Mahmoud Ali. 2014. Big data: survey, technologies, opportunities, and challenges. *The scientific world journal* 2014, 1 (2014), 712826.
- [18] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1092–1102.
- [19] Hong Liu, Zhenhua Sang, and Sameer Karali. 2019. Approximate Quality Assessment with Sampling Approaches. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI).
- [20] Sedir Mohammed, Hazar Harmouch, Felix Naumann, and Divesh Srivastava. 2024. Data Quality Assessment: Challenges and Opportunities. arXiv preprint

arXiv:2403.00526 (2024).

[21] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. 2021. From Cleaning before ML to Cleaning for ML. *IEEE Data Eng. Bull.* 44, 1 (2021), 24–41.

[22] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[23] Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. 2019. Data Validation for Machine Learning. In *Proceedings of Machine Learning and Systems*, Vol. 1. 334–347.

[24] Thomas C Redman. 1997. *Data quality for the information age*. Artech House, Inc.

[25] Barna Saha and Divesh Srivastava. 2014. Data quality: The other face of Big Data. 2014 IEEE 30th International Conference on Data Engineering (2014), 1294–1297.

[26] Yiyuan She and Art B Owen. 2011. Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc.* 106, 494 (2011), 626–639.

[27] Han Shomorony and A Salman Avestimehr. 2014. Sampling large data on graphs. In 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 933–936.

[28] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. Data quality: A survey of data quality dimensions. In 2012 International Conference on Information Retrieval & Knowledge Management. IEEE, 300–304.

[29] Stefan Studer, Thanh Binh Bui, Christian Drescher, and et al. 2021. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. arXiv:2003.05155 [cs.LG]

[30] Jeffrey Scott Vitter. 1985. Random sampling with a reservoir. *ACM Trans. Math. Software* 11 (1985), 37–57.

[31] Agung Wahyudi, Adiska Farhani, and Marijn Janssen. 2018. Relating Big Data and Data Quality in Financial Service Organizations. In *IFIP International Conference on e-Business, e-Services, and e-Society*.

[32] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

[33] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* 12, 4 (1996), 5–33.

[34] Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022. Noiserobust

cross-modal interactive learning with text2image mask for multi-modal neural machine translation. In Proceedings of the 29th International Conference on Computational Linguistics. 5098–5108.

[35] Naiem K. Yeganeh, Shazia Sadiq, and Mohamed A. Sharaf. 2014. A framework for data quality aware query.

