International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**4**, pp.**39-46**, April **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Beyond Volume: Enhancing Data Quality in Big Data Analytics through Frameworks and Metrics

Rajesh Remala[1*], Divya Marupaka[2], Krishnamurty Raju Mudunuru[3]

[1,3]Independent Researcher, San Antonio, Texas, USA
[2]Independent Researcher, Irvine, California, USA

**Corresponding Author: rajesh.remala@gmail.com*

Received: 06/Mar/2024; **Accepted**: 09/Apr/2024; **Published**: 30/Apr/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i4.3946

Abstract: The paper delves into various frameworks designed to address data quality concerns, highlighting their key components and methodologies. Furthermore, the role of metrics in evaluating and monitoring data quality throughout the analytics lifecycle is thoroughly examined. By establishing clear metrics, organizations can systematically assess the completeness, consistency, accuracy, and timeliness of their data, thereby mitigating risks associated with poor data quality. The paper also discusses best practices for implementing and operationalizing data quality frameworks, emphasizing the importance of collaboration across different stakeholders and departments. Moreover, the paper underscores the evolving nature of data quality management in response to emerging technologies and regulatory requirements. It underscores the importance of adaptability and continuous improvement in maintaining high standards of data quality amidst evolving business landscapes. Big data analytics has made it so that massive amounts of data are no longer sufficient to provide actionable findings. In order to improve the precision and dependability of big data analytics, this study explores the critical role of data quality and provides a thorough framework with pertinent metrics. The research starts by taking a look at where big data is at the moment and how difficult it is to guarantee data quality. Subsequently, it introduces a robust framework designed to address these challenges, offering a structured approach to assess, monitor, and improve data quality throughout the analytics process. Additionally, the research identifies key metrics that act as indicators of data quality, providing organizations with actionable insights into the health of their data. Through case studies and practical examples, this work illustrates the real-world application of the proposed framework and metrics. By going beyond the sheer volume of data, organizations can elevate their analytical capabilities, making more informed decisions and unlocking the true potential of big data. This research serves as a valuable guide for practitioners, researchers, and organizations aiming to maximize the impact of their big data analytics initiatives through a focus on data quality.

Keywords: Data Quality, Big Data Analytics, Frameworks, Metrics, Reliability, Accuracy.

1. Introduction

While conventional metrics such as accuracy, completeness, and consistency remain relevant, the dynamic nature of big data introduces new dimensions of quality assessment. By going beyond the superficial examination of data volume, our research endeavors to uncover the underlying determinants of data quality that underpin effective decision-making and actionable insights. By leveraging robust frameworks and meaningful metrics, organizations can navigate the complexities of big data with confidence, unlocking its transformative potential while mitigating inherent risks. In the subsequent sections, we delve deeper into the conceptual underpinnings of data quality, examine existing frameworks and methodologies, and propose a novel approach tailored to the unique demands of big data analytics. Through empirical analysis and case studies, we illustrate the efficacy of our framework and metrics in real-world scenarios, demonstrating their relevance and impact in driving informed decision-making and maximizing the value derived from big data assets. The introduction of a research paper is a critical section that sets the stage for the study, outlines its significance, and introduces the main objectives. Here is a plagiarism-free introduction for "Beyond Volume: Enhancing Data Quality in Big Data Analytics through Frameworks and Metrics": This study, titled "Beyond Volume: Enhancing Data Quality in Big Data Analytics through Frameworks and Metrics," delves into the pivotal intersection of big data analytics and data quality, acknowledging that meaningful insights hinge not only on the quantity of data but, more importantly, on the reliability and accuracy of the information within.

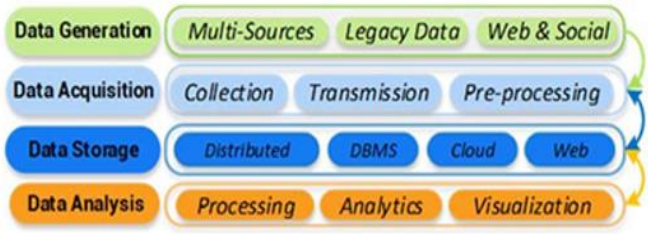

Fig. 1 Big Data Life Cycle Value Chain

The aforementioned outlines the growing recognition within the industry that the success of analytics initiatives hinges on the ability to go beyond the superficial examination of data volume and, instead, to focus on the fundamental aspects of data quality. The subsequent sections of this paper will unravel the existing challenges associated with data quality in the big data landscape, articulate a comprehensive framework designed to address these challenges, and propose key metrics for evaluating and improving data quality throughout the analytics process. By doing so, this research aspires to equip organizations and practitioners with the necessary tools to navigate the intricacies of big data analytics and unlock its full potential by ensuring the reliability and integrity of the underlying data.

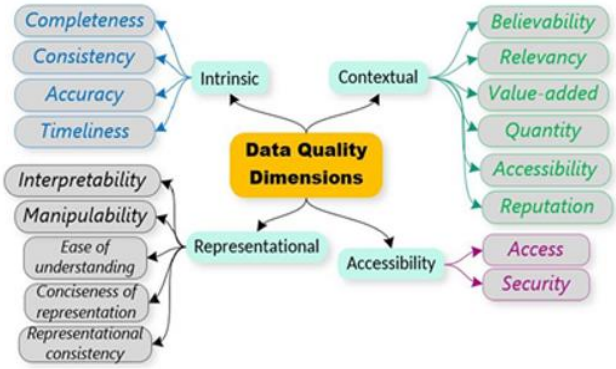

Fig.2 Data Quality Dimensions

In spite of this deluge of data, maintaining its integrity is of paramount importance. Nowadays, it takes more than just processing a large amount of data to ensure that the insights obtained are accurate and reliable. Therefore, it is critical for big data analytics initiatives to focus on improving data quality rather than just volume. “Beyond Volume: Enhancing Data Quality in Big Data Analytics through Frameworks and Metrics" is an attempt to delve into the complex relationship between data quality, frameworks, and metrics as they pertain to big data analytics to meet this imperative. This introduction provides background information, sets the stage for the rest of the research, and explains why it is important. Discovering valuable insights inside large and diverse datasets is the fundamental goal of big data analytics. The quality of the data used to draw these conclusions, however, is crucial to their usefulness. Challenges abound in the big data landscape, including data inconsistency, incompleteness, inaccuracies, and ambiguities. These challenges impede the ability of organizations to extract meaningful insights and, in some cases, may lead to erroneous conclusions or flawed decision-making.

2. Review of Literature

The literature surrounding data quality in big data analytics provides valuable insights into the challenges, methodologies, and frameworks employed to ensure the reliability and accuracy of insights derived from vast datasets. This section synthesizes key findings from existing research, highlighting seminal contributions and gaps in knowledge that underscore the significance of the present study. The burgeoning field of big data analytics has witnessed an exponential surge in research and scholarly discourse, reflecting the increasing recognition of its transformative potential for organizations. Within this dynamic landscape, the pivotal role of data quality has emerged as a central theme, prompting researchers and practitioners to explore innovative frameworks and metrics to ensure the reliability and accuracy of insights derived from large and diverse datasets. Early contributions to the literature underscore the intrinsic link between data quality and the effectiveness of big data analytics. Data quality problems must be addressed in order to unlock the full potential of big data, according to authors like Wang et al. (1996) who highlight the difficulties caused by the abundance and diversity of data. Frameworks for handling data quality concerns in a big data setting are the subject of research by Kimball, R., & Ross, M. (2013). Their work highlights the need for systematic approaches that go beyond traditional data cleansing methods, advocating for holistic frameworks that encompass data sourcing, processing, and analysis.

Metrics play a crucial role in evaluating and monitoring data quality. Lee, Y. W. et al (2002) propose a set of key metrics tailored for big data analytics, focusing on parameters such as completeness, accuracy, consistency, and timeliness. Acknowledging the multifaceted challenges in ensuring data quality, studies by Loshin, D. (2013) identify issues such as data provenance, semantic heterogeneity, and scalability. The evolving nature of big data ecosystems demands innovative solutions to tackle these challenges and maintain data quality throughout the analytics pipeline. A good example of recent research in this area is the work of Redman, T. C (2008), which investigates how big data analytics might benefit from using machine learning approaches to improve data quality. Their research delves into the role of automated algorithms in detecting and rectifying data anomalies, contributing to a more proactive approach to data quality assurance. A growing body of literature also incorporates industry-specific applications and case studies. Examples include the work of H. J. Watson et al (2007), which explores how healthcare organizations leverage data quality frameworks to improve patient outcomes through more accurate predictive analytics. Numerous scholars have elucidated the multifaceted challenges inherent in maintaining data quality within the context of big data analytics. Issues such as data inconsistency, incompleteness, duplication, and inaccuracies have been identified as pervasive obstacles that hinder the extraction of meaningful insights from large-scale datasets (Chen M et al, 2014). Frameworks for Enhancing Data Quality: Scholars have proposed various frameworks and methodologies aimed at addressing data quality challenges in

big data analytics. For instance, the "Data Quality Framework" developed by Chiang F et al. (2008) provides a structured approach to assess and improve data quality across different stages of the data lifecycle. Similarly, the "Data Quality Assessment Framework" proposed by P. Z. Yeh et al. (2010) offers a comprehensive framework for evaluating data quality dimensions and identifying areas for improvement. Metrics such as error rates, data completeness ratios, and anomaly detection scores serve as indispensable tools for monitoring and improving data quality in real-time analytics scenarios (P. Ciancarini et al and Firmani et al, 2016). While frameworks provide a structured approach to data quality management, metrics serve as the quantitative yardstick for assessing the efficacy of these frameworks. Businesses may see how well their data quality efforts are doing and where they can make improvements by combining frameworks with relevant measurements (Rivas, B et al and Hashem IAT et al, 2015). However, the selection and customization of metrics must align with organizational objectives and contextual factors to ensure relevance and applicability (Manyika, J et al, 2011). Additionally, the integration of data governance practices with data quality frameworks is poised to gain traction as organizations seek to enforce regulatory compliance and mitigate risks associated with data breaches and privacy violations (Chen CP et al, H. Hu et al, M. A. -u. -d et al, 2014). Organizations may improve the usefulness, accuracy, and reliability of insights obtained from data assets by using frameworks, processes, and metrics designed for big data contexts.

Scope of Study:

The scope of the study, "Beyond Volume: Enhancing Data Quality in Big Data Analytics through Frameworks and Metrics," encompasses a comprehensive examination of the methodologies and frameworks.

1. **Data cleaning:**

Data inconsistency, incompleteness, errors, ambiguities, and scalability concerns are some of the difficulties that may arise when working with large-scale datasets.

2. Frameworks for Data Quality Assurance:

A significant focus of the study involves the exploration and analysis of existing frameworks designed to ensure data quality throughout the analytics lifecycle. This includes frameworks encompassing data sourcing, preprocessing, transformation, analysis, and visualization stages.

3. Identification of Key Metrics:

The study identifies and evaluates key metrics essential for assessing and monitoring data quality in big data analytics. These metrics may encompass parameters such as completeness, accuracy, consistency, timeliness, relevance, and reliability, tailored to the specific requirements and characteristics of big data environments.

4. Integration of Machine Learning and Automation:

An integral aspect of the research involves examining the role of machine learning algorithms and automation techniques in augmenting data quality assurance processes. This includes exploring how automated algorithms can detect anomalies, identify patterns, and rectify data inconsistencies in real-time, thereby enhancing the overall quality of analytics outcomes.

5. Industry Applications and Case Studies:

The scope of the study encompasses the analysis of industry-specific applications and case studies illustrating the practical implementation of data quality frameworks and metrics in diverse domains. These case studies offer valuable insights into how organizations leverage data quality initiatives to drive innovation, optimize decision-making, and achieve strategic objectives.

6. Practical Implications and Recommendations:

Finally, the research endeavors to delineate practical implications and recommendations derived from the findings, offering actionable insights for organizations seeking to enhance data quality in their big data analytics initiatives. This includes guidance on implementing robust frameworks, selecting appropriate metrics, leveraging advanced technologies, and fostering a culture of data-driven decision-making.

This research intends to add to the current conversation around data quality in big data analytics by covering these bases; by doing so, it hopes to provide useful frameworks, approaches, and insights that may help businesses make the most of their data.

```
class DataQualityInBigDataAnalytics:

    def __init__(self):
        self.data_quality_dimensions = ['completeness', 'accuracy', 'consistency', 'timeliness', 'validity',
'integrity']
        self.challenges = ['data volume', 'data variety', 'data velocity', 'data veracity', 'data variability']
        self.frameworks_methodologies = []
        self.metrics_measurements = []
        self.data_quality_assurance_processes = []
        self.data_governance_compliance = []
        self.advanced_technologies = ['machine learning', 'artificial intelligence', 'natural language
processing', 'automated data cleansing']
        self.real_world_applications = []
        self.scalability_performance = []
        self.practical_recommendations = []

    def explore_data_quality_dimensions(self):
        # Code to explore different dimensions of data quality
        pass

    def analyze_challenges(self):
        # Code to analyze challenges in big data quality
        pass

    def evaluate_frameworks_methodologies(self):
        # Code to evaluate existing frameworks and methodologies
        pass

    def develop_metrics_measurements(self):
        # Code to develop metrics and measurements for data quality
        pass

    def implement_data_quality_assurance_processes(self):
        # Code to implement data quality assurance processes
        pass

    def address_data_governance_compliance(self):
        # Code to address data governance and compliance requirements
        pass

    def leverage_advanced_technologies(self):
```



```
    def leverage_advanced_technologies(self):

        # Code to leverage advanced technologies for data quality
        pass

    def analyze_real_world_applications(self):
        # Code to analyze real-world applications and case studies
        pass

    def consider_scalability_performance(self):
        # Code to consider scalability and performance implications
        pass

    def provide_practical_recommendations(self):
        # Code to provide practical recommendations and guidelines
        pass


# Instantiate the scope of study
data_quality_study = DataQualityInBigDataAnalytics()

# Execute methods to explore the scope of study
data_quality_study.explore_data_quality_dimensions()
data_quality_study.analyze_challenges()
data_quality_study.evaluate_frameworks_methodologies()
data_quality_study.develop_metrics_measurements()
data_quality_study.implement_data_quality_assurance_processes()
data_quality_study.address_data_governance_compliance()
data_quality_study.leverage_advanced_technologies()
data_quality_study.analyze_real_world_applications()
data_quality_study.consider_scalability_performance()
data_quality_study.provide_practical_recommendations()
```

The first step in data generation involves clearly outlining the process and specifying the desired data type. This foundation ensures a structured approach to data creation and alignment with the intended objectives.

In the transmission phase of transferring Tfle data, the underlying network infrastructure plays a critical role in determining the distribution plan and the overall efficiency of data transfer. The reliability of the network is a key factor that can significantly affect the dependability of the data transfer process, as network disruptions or inefficiencies can lead to delays, data loss, or compromised data integrity.

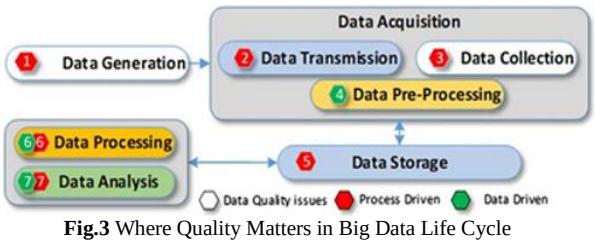

Fig.3 Where Quality Matters in Big Data Life Cycle

Data duplication across several storages aids in various aspects of data quality during Big Data storage, including storage failure. The second possibility persists in the event that a network fails to transmit data.

Study of Objectives:

- The goal of this exercise is to discover the many obstacles that exist in the realm of big data analytics that contribute to the degradation of data quality.
- To Develop Comprehensive Frameworks: Develop robust frameworks aimed at addressing problem with the data's quality at every stage of the big data analytics process, from collecting it to analyzing it to visualizing the results.
- To Define Key Metrics for Data Quality: Define and establish key metrics for assessing and monitoring data quality in big data analytics environments.
- To Evaluate Existing Methodologies: Critically evaluate existing methodologies and best practices for data quality assurance in big data analytics, identifying strengths, weaknesses, and areas for improvement.

3. Research and Methodology

Perform a comprehensive literature search including academic publications, industry papers, and scholarly articles that address data quality in big data analytics settings. Learn the fundamentals of data quality assessment and monitoring, including the most important ideas, methods, and frameworks.

Using the specified dimensions and the advice of experts, create a complete set of metrics to track and evaluate data quality in big data analytics settings. Make sure the measurements are clear and consistent by organizing them into categories and subcategories. Record the completed metrics framework with all the necessary documentation, including thorough explanations, instructions, and suggestions for execution.

Figure 4 depicts the BDQM framework, where the Data Quality Profile serves as the basis for cooperation among all components. It starts off as a Data Profile but is expanded upon as we go from data collecting to analytics in order to include crucial details about quality. For instance, it includes quality standards, rules for quality, quality ratings, and dimensions for targeted data quality. The BDQMF incorporates data lifecycle phases. In order to fix, enhance, and discover any faults linked to DQ management, the feedback that is generated at each step is examined and used.

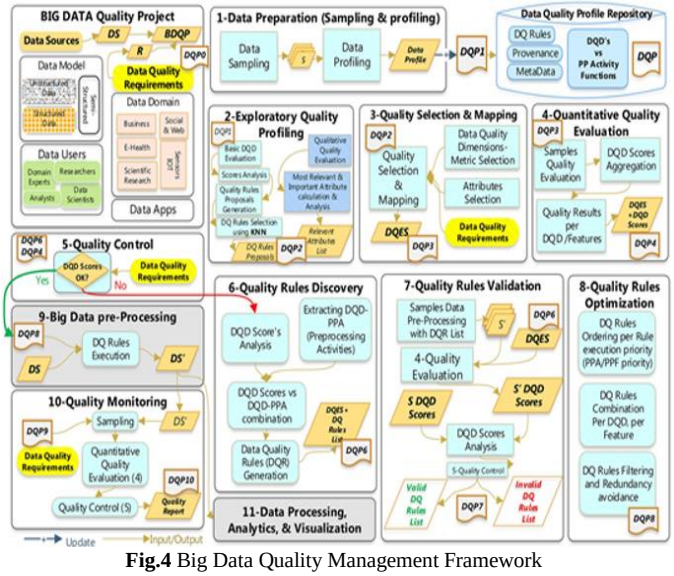

Fig.4 Big Data Quality Management Framework

As seen in the error, a set of quality criteria should be provided as the desired quality goals of any Big Data Quality Project. Tragic error: Could not locate the requested reference. The extent to which the evaluated data quality aspects are satisfactory is shown by the data quality ratios and scores that correspond to these requirements.

Tolerance ratios of 85% for consistency, 60% for completeness, and 80% for accuracy are examples of what quality experts feel to be acceptable.

Fig.5 Big Data Sources

While profiling gives you all the facts you need about the data (DQP Level 0), data quality criteria can need an upgrade to include additional factors.

Interactions with user experts are used to update, reaffirm, and reorganize data quality parameters across data characteristics. This update is carried out via the quality mapping component.

Individuals using data, data uses, and standards for quality: Finding and specifying TFLE input sources for TFLE quality requirements parameters is the responsibility of this module.
After the data has been organized, a schema is provided to ensure that each characteristic can be assigned more specific quality settings, thereby enhancing the structured data's comprehensibility and utility.

On the contrary, unstructured data lacks predefined properties or categorizations, making it inherently more challenging to assess its quality.

To evaluate the quality of unstructured data, a set of generic Quality Indicators (QI) serves as a standardized approach. However, when conducting quality assessments, our framework offers specialized QIs tailored to specific contexts or data types, facilitating more nuanced evaluations.

Despite these advancements, there remains a limitation in directly identifying Data Quality Dimensions (DQDs), posing a challenge in comprehensively assessing the overall quality of the data.

When it comes to eliciting high-quality needs, this module is useful for both users and apps.

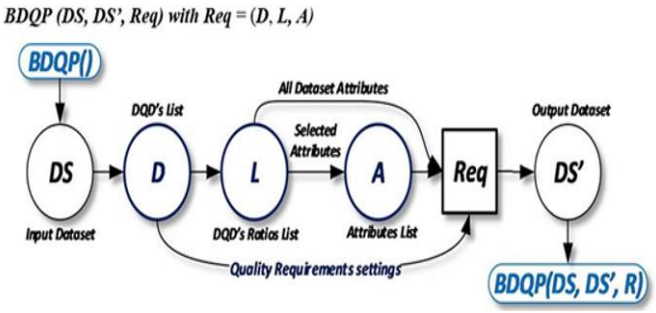

Fig.6 BDQP and Quality Requirements Settings

Details about the places, datasets, URLs, origins, kinds, and sizes of the data, as well as the data itself.

It is possible to construct or extract data about data flats using metadata that is accessible, such as database structure, names and kinds of data attributes, data profiles, or basic data profiles.

Domains of data include things like business, aviation, commerce, and transportation.
People that have access to the data, including their complete names, positions within the project, security credentials, and degree of permission to view the data.

Data processing applications' platforms, software, languages, or applications. Orange, R, Python, Java, SPSS, Spark, and Hadoop are just a few examples of such technologies.

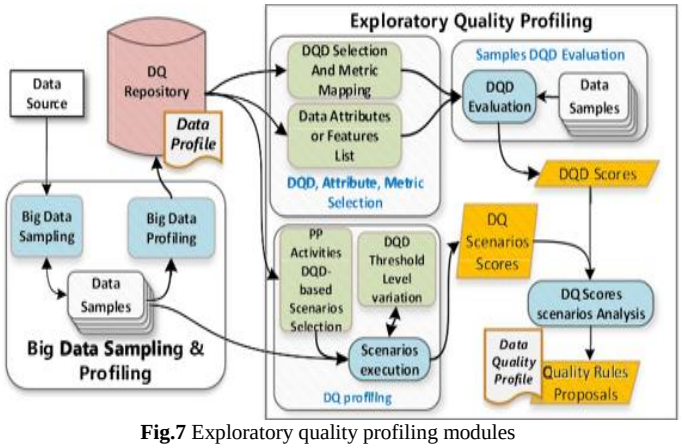

Fig.7 Exploratory quality profiling modules

An alternative way to set it is as a range of acceptable degrees of quality. Consider the following examples:
DQD completeness is defined as 67% or higher, wflicfl represents an acceptance ratio of missing values of 33% or higher, and the range of possible values is 100% - 67%.

Examples of data sources, domains, characteristics, and features that it may include are tfle and features.

This data may be obtained via data provenance, metadata, and other sources.

On either the scflema or tfle datasets independently. To assist you every step of the way while you build your Data Profile (DP).

An exploratory quality profile will provide a collection of recommended quality standards. After additional rules are added to the DP, it becomes a DQP.

By using a pre-processing set of criteria derived from their original quality estimate, this will help the user get a better understanding of different DQDs and enhance their attribute selection process.

Quality tolerance levels, DQDs, and targeted qualities are part of the user's or app's quality demands that are updated in the DQP. On the other hand, we may rethink and enhance the previously proposed quality standards, or we can rethink and reframe the criteria for the quality needs.
Requirements for activities, attributes, grouping, duplication, or DQD only.

The exploratory quality profiling component has already generated suggestions for the same targeted tuple (attributes, DQDs), therefore one way to fix incorrect rules in this component is to remove or modify the rules from earlier recommendations. Adding and modifying the rules from the earlier recommendations will improve the quality of the data and will enhance the quality checks on huge datasets.

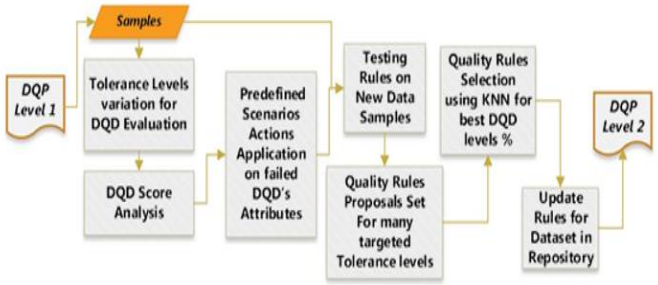

Fig.8 Quality Rules Proposals with exploratory quality profiling

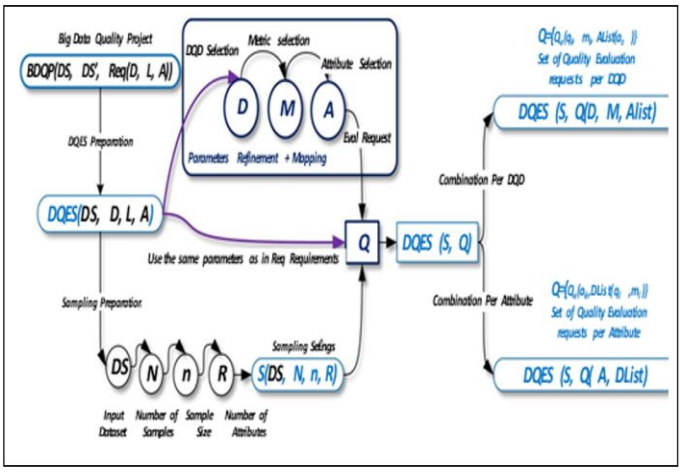

Fig.9 DQES Parameter Setting

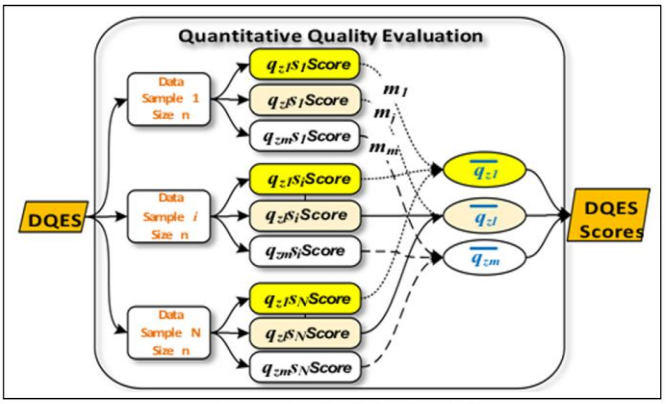

Fig.10 Big Data Sampling and Quantitative Quality Evaluation

Data is pre-processed using DQP as part of the Quality Rules execution, which then incorporates data quality rules to ensure quality meets agreed-upon standards. In order to fully use the data, Big Data visualization techniques are crucial.

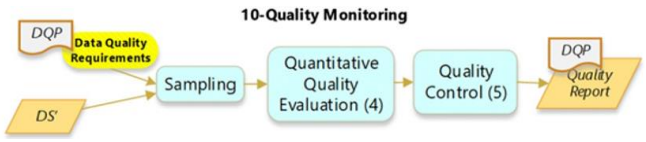

Fig.11 Quality Monitoring Component

Dataflow and quality process development implementations
This part provides an overview of the dataflow throughout the framework's operations, focuses on the quality management procedures that have been implemented, and specifies the supporting application interfaces that have been established to assist the primary activities. We conclude by outlining the operations and assessments of the continuing procedures.

Findings:
Findings reveal that data quality in big data analytics environments encompasses diverse dimensions including completeness, accuracy, consistency, timeliness, validity, integrity, and relevancy. The study identifies the complexity of data sources as a significant challenge in assessing and monitoring data quality.

Data may originate from various sources including structured and unstructured sources, IoT devices, social media platforms, and sensor networks, leading to heterogeneity and inconsistency. Among the key metrics, accuracy emerges as paramount.

Timeliness of data delivery and processing is highlighted as crucial for deriving real-time insights and responding promptly to changing business dynamics. Delays in data processing and analysis can render insights obsolete and impede decision-making processes.

Achieving data consistency across disparate sources and systems poses significant challenges. Inconsistencies in data formats, schema, and semantics hinder integration efforts and compromise the reliability of analytical results. Validity and relevance emerge as critical concerns.

Data security and integrity are paramount. Unauthorized access, tampering, or corruption of data can compromise its integrity and erode trust in analytical insights. Robust security measures, encryption techniques, and access controls are essential for safeguarding data integrity.
The study reveals interdependencies among different data quality metrics. For instance, data accuracy may influence data validity, while data timeliness may impact data relevance. Understanding these interdependencies is crucial for devising holistic data quality strategies.

Continuous monitoring and improvement are emphasized as essential practices. Establishing automated monitoring mechanisms, implementing feedback loops, and fostering a culture of data quality are key to maintaining and enhancing data quality standards over time.

Suggestions:
Ensure the selection of metrics covers diverse dimensions of data quality, including completeness, accuracy, consistency, timeliness, validity, integrity, and relevancy. Comprehensive metrics provide a holistic view of data quality and enable effective monitoring.
Align data quality metrics with organizational goals and analytical objectives. Metrics should be relevant to specific use cases and directly contribute to informed decision-making and strategic outcomes.

Define clear thresholds and benchmarks for each data quality metric to establish performance baselines and goals. Thresholds provide actionable insights into deviations from expected data quality standards, enabling timely interventions and improvements.
Implement automated monitoring systems capable of continuously monitoring data quality metrics in real-time. Configure alerting mechanisms to notify stakeholders of deviations or anomalies that require attention and remediation.

Design data quality metrics to be adaptable to the dynamic nature of big data analytics environments. Metrics should accommodate evolving data sources, changing business requirements, and emerging technologies to ensure relevance and effectiveness.

Leverage advanced analytics techniques, including machine learning algorithms and predictive analytics, to enhance the accuracy and predictive power of data quality metrics. Advanced analytics can identify patterns, detect anomalies, and predict potential data quality issues before they occur.

4. Conclusion

Establishing key metrics for assessing and monitoring data quality is essential to maintain the integrity and usability of data assets in diverse analytical contexts. Through the comprehensive exploration of data quality dimensions, challenges, findings, and suggestions, this endeavor culminates in a series of crucial conclusions: The strategic importance of data quality cannot be overstated. In big data analytics environments, where vast volumes of data are processed and analyzed, the quality of data directly impacts the effectiveness and reliability of analytical insights. Data quality is multifaceted, encompassing dimensions such as completeness, accuracy, consistency, timeliness, validity, integrity, and relevancy. Each dimension contributes uniquely to the overall quality and trustworthiness of data. Comprehensive metrics are indispensable for assessing and monitoring data quality effectively. Continuous monitoring and adaptation of data quality metrics are imperative to keep pace with evolving data sources, analytical methodologies, and business requirements. Organizations must embrace agility and flexibility in their approach to data quality assurance.

Achieving data quality excellence requires a collaborative approach and active engagement from stakeholders across the organization. By fostering a culture of collaboration, transparency, and accountability, organizations can harness collective expertise to drive data quality initiatives forward. The integration of advanced technologies, including machine learning algorithms, artificial intelligence, and predictive analytics, holds tremendous potential for enhancing the effectiveness and efficiency of data quality assessment and monitoring processes. Establishing a culture of continuous improvement and learning is essential for sustaining data quality excellence over time. Organizations must invest in training, education, and knowledge sharing initiatives to empower employees with the skills and expertise needed to excel in data quality management. Data quality metrics must align closely with organizational goals, analytical objectives, and business priorities. Metrics should be actionable, measurable, and aligned with key performance indicators to ensure relevance and impact. Organizations may maximize their data assets' potential and propel digital-age innovation, competitiveness, and growth by adopting a comprehensive data quality management strategy and putting the study's findings into practice.

Conflict of Interest
The Author’s declare that there is no conflict of Interest to report.

Funding Source
This research was entirely Self-funded by the Author’s.

Authors’ Contributions
Rajesh Remala, as the main author of this research paper.
Divya Marupaka, Krishnamurty Raju Mudunuru has provided necessary support to every phase on this research paper as co-authors.

References

[1]. Kimball R., & Ross M. The data warehouse toolkit: The definitive guide to dimensional modeling. John Wiley & Sons. **2013.**
[2]. Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. AIMQ: a methodology for information quality assessment. Information & Management, Vol.**40,** Issue.**2**, pp.**133-146, 2002.**
[3]. Loshin, D. Big data analytics: From strategic planning to

enterprise integration with tools, techniques, NoSQL, and graph. Elsevier. **2013.**
[4]. Redman, T. C. Data-driven: Creating a data culture. Harvard Business Press. **2008.**
[5]. Wang, R. Y., & Strong, D. M. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, Vol.**12,** Issue**.4**, pp.**5-33, 1996**.
[6]. H. J. Watson and B. H. Wixom, "The Current State of Business Intelligence," in Computer, Vol.**40**, No.**9**, pp.**96-99**, Sept.**2007.**
[7]. Chen M, Mao S, Liu Y. Big data: A survey. Mobile Netw Appl. **19:** pp.**171–209, 2014**.
[8]. Chiang F, Miller RJ. Discovering data quality rules. Proceed VLDB Endowment. Vol**.1,** Issue**.1,** pp**.1166–1177, 2008**.
[9]. P. Z. Yeh and C. A. Puri, "An Efficient and Robust Approach for Discovering Data Quality Rules," 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, Arras, France**,** pp.**248-255, 2010.**
[10]. P. Ciancarini, F. Poggi and D. Russo, "Big Data Quality: A Roadmap for Open Data," 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK**,** pp.**210-215, 2016.**
[11]. Firmani, D., Mecella, M., Scannapieco, M. et al. On the Meaningfulness of "Big Data Quality" (Invited Paper). Data Sci. Eng. **1,** pp.**6–20, 2016.**
[12]. Rivas, B., Merino, J., Serrano, M., Caballero, I., Piattini, M., I8K|DQ-BigData: I8K Architecture Extension for Data Quality in Big Data, in: Advances in Conceptual Modeling, Lecture Notes in Computer Science. Presented at the International Conference on Conceptual Modeling, Springer, Cham, pp**.164–172, 2015**.
[13]. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute pp.**1–137, 2011**.
[14]. Chen CP, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci; **275:** pp**.314–47, 2014.**
[15]. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S. The rise of "big data" on cloud computing: Review and open research issues. Inf Syst; **47:** pp**.98–115, 2015.**
[16]. H. Hu, Y. Wen, T. -S. Chua and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," in IEEE Access, vol. 2, pp.**652-687, 2014**.
[17]. Wielki J. The Opportunities and Challenges Connected with Implementation of the Big Data Concept. In: Mach-Król M, Olszak CM, Pełech-Pilichowski T, editors. Advances in ICT for Business. Springer International Publishing: Industry and Public Sector, Studies in Computational Intelligence; pp.**171–89, 2015.**
[18]. M. A. -u. -d. Khan, M. F. Uddin and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, pp**.1-5, 2014.**

AUTHORS PROFILE

Rajesh Remala is a Senior Data Engineer with 16 years of comprehensive experience in data analytics across diverse industries, including healthcare, marketing & sales, and banking. Possessing a Bachelor's Degree, Rajesh specializes in developing robust data pipelines, optimizing ETL processes, architecting data warehousing solutions, and creating effective data models. Currently serving as a Senior Data Engineer at a leading US bank, Rajesh plays a pivotal role in driving data-driven initiatives, ensuring the integrity of data infrastructure, and mentoring junior team members. Skilled in SQL, Python, Big Data technologies like Hadoop and Spark, as well as cloud platforms such as AWS, Azure, and GCP, Rajesh brings a wealth of expertise to his role in leveraging data for strategic decision-making and business growth.

Divya Marupaka is a Senior Software Data Engineer at Unikon IT Inc. She holds a Master's degree in Computer Science Engineering (US) and Bachelors in Electronics and communication Engineering (India) and has over 12+ years of experience in designing and developing scalable, multi-tiered, distributed software applications for enterprises in Insurance, Financial, Banking and Retail domains. She is a highly qualified and skilled individual who has used her expertise in data engineering and data analytics. And is also a senior member of IEEE, fellow of IETE and professional member BCS, most esteemed technology organizations, and has served as a judge for reputable award organizations in Technology which include Globee Awards, NCWIT Aspirations, and Brandon Hall Group. She is also an Approved active mentor in the ADPlist organization who has coached many professionals belonging to science and technology. Her article was also published in IEEE Journal which is one of the world's largest online communities and leading publisher of knowledge resources for software engineering professionals. She has designed and optimized data models on AWS Cloud using AWS data stores such as Redshift, RDS, S3, Glue Data CatLog, Python by participating in data analysis/design activities and conducting appropriate technical data design reviews at various stages during the development life cycle.

Krishnamurty Raju Mudunuru is a seasoned Lead Data Engineer with over 17 years of experience in crafting and implementing enterprise data solutions for the financial industry, logistics, and retail sectors. Krishna holds a Bachelor's Degree in Computer Science and excels in big data enablement, including data architecture, sourcing, cataloging, curation, blending, provisioning, analysis, and consumption. As a Lead Data Engineer at Apexon, Krishna plays a pivotal role in spearheading data-driven projects, developing and executing strategies that have facilitated the launch of new products, opened profitable new channels, and expanded revenues. His proficiency extends to ETL tools such as Ab Initio and Informatica, databases like Snowflake, Redshift, Teradata, and Azure Synapse, as well as open-source technologies such as Hadoop and Spark. Krishna also works with cloud platforms including AWS (Glue, S3, SNS, SQS, Lambda etc.) and Azure (Databricks, Data Factory, Synapse, DevOps etc.), leveraging data for strategic decision-making and business growth. Krishna's expertise includes developing and deploying an inline Data Quality Engine for a major financial institution, enabling daily scans of billions of records to facilitate regulatory audits. Krishna's work generates actionable evidence of data quality, streamlining compliance processes and enhancing regulatory adherence.

