Proactive Data Quality Framework

Tejas Girijakant Naik
Department of MCA
RV College of Engineering
Bengaluru, India
tejasgn.mca21@rvce.edu.in

Dr..S.S.Nagamuthu.Krishnan
Department of MCA
RV College of Engineering
Bengaluru, India
ssnk@rvce.edu.in

Abstract*— Maintaining accurate and trustworthy data has become crucial for organisations in the quickly changing world of data-driven decision-making. A complete strategy to guarantee data integrity and improve the efficiency of decision-making processes is provided by the proactive data quality framework. This framework automates data quality checks on huge datasets and integrates data integration, migration, warehousing, and analytics, allowing for the quick diagnosis and correction of quality concerns. The framework increases data consistency and accuracy by utilising cutting-edge technologies like big data processing, real-time analytics, and machine learning algorithms. Additionally, it places a focus on securing private information, ensuring regulatory compliance, and protecting data security. The proactive data quality framework's use enables organisations to make smart business decisions, improving outcomes and a competitive advantage in today's data-centric environment.

***Keywords*— Data quality, Data-driven Decision-making, Proactive Data Quality, Data Integration, Data Migration, Data Warehousing, Data Analytics, Automation, Big Data Processing, Real-Time Analytics, Compliance.**

I. Introduction

The main purpose of a data quality framework is to establish a standardized and repeatable procedure for managing data quality. It begins with defining data quality requirements and metrics specific to the organization's needs and industry standards. This involves identifying critical data elements, establishing quality thresholds, and determining appropriate validation rules and data profiling techniques.

Once the data quality requirements are defined, the framework includes mechanisms for data profiling, data cleansing, and data enrichment. Data profiling involves analyzing the content, structure, and relationships within the data to identify anomalies, inconsistencies, and patterns. Data cleansing aims to rectify or remove errors, duplicates, and outliers from the dataset. Data enrichment involves augmenting the data and adding more information from outside sources to enhance its completeness and accuracy.

Another crucial aspect of a data quality framework is the ongoing monitoring and measurement of data quality. This involves establishing data quality metrics, defining key performance indicators (KPIs), and implementing regular audits and assessments to evaluate the effectiveness of data quality initiatives. Monitoring allows organizations to identify trends, detect emerging data quality issues, and take proactive measures to address them promptly.

Furthermore, it often incorporates data governance principles to ensure accountability, ownership, and compliance with data quality policies and regulations. It defines roles and responsibilities for data stewards and those in charge of upholding and enforcing data quality standards.

II. LITERATURE REVIEW

Big Data is an essential research area for governments, institutions, and private agencies to support their analytics decisions. Big Data refers to all about data, how it is collected, processed, and analyzed to generate value-added data-driven insights and decisions. Degradation in Data Quality may result in unpredictable consequences. In this case, confidence and worthiness in the data and its source are lost. [2].

In research by Maxi Kindling and Dorothea Strecker. [1] they present the outcomes of a survey conducted among research data repositories, investigating the state of data quality assurance practices. The results indicate that most repositories engage in data quality measures and significantly contribute to data quality. However, the study highlights the diverse and multifaceted nature of data quality assurance approaches, with individualized methods prevailing. Several challenges are identified, including the recognition of data reviewers and repositories, the influence of publication review processes on data review, and the absence of adequate data quality information. Notably, the study shows that a repository's level of accreditation is not always correlated with the depth of its quality assurance procedures. Overall, this research emphasizes the high expectations and resource requirements associated with data quality assurance in research data repositories.

Another study by Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui & Rachida Dssouli et al. [2] discusses the importance of data quality in the Big Data domain and proposes a Big Data Quality (BDQ) management framework. The framework aims to enhance pre-processing activities and strengthen data control by introducing the concept of Big Data Quality Profile (DQP) to capture quality requirements, attributes, and scores. The framework incorporates data profiling and sampling components to estimate data quality before and after pre-processing. It utilizes predefined quality metrics and generates quality

rules to assess important data quality dimensions. The paper emphasizes the need for quality assessment throughout the Big Data lifecycle and the challenges posed by unstructured and schema-less data from multiple sources. The proposed framework aims to support data-driven decision-making by providing quality management mechanisms and facilitating cross-process data quality enforcement. The paper concludes by discussing the implementation and dataflow management aspects of the framework.

Noha Mostafa, Haitham Saad Mohamed Ramadan, Omar Elfarouk et al. [3] .This paper suggests that implementing big data analytics in the energy sector, particularly for renewable energy power stations and smart grids, offers significant benefits and challenges. For the possible use of big data analytics in smart grids and renewable energy power utilities, a framework is created. The study uses a five-step approach, employing different machine learning methods, to predict smart grid stability. The results indicate that the penalized linear regression model achieved 96% accuracy in predicting system stability, while the Convolutional Neural Network (CNN) model achieved 87% accuracy in classifying smart grid stability. The study highlights the importance of data analysis in optimizing the stability and efficiency of smart grid systems, with the CNN model providing the best accuracy and faster processing compared to other machine learning algorithms. The paper acknowledges the limitation of the relatively small dataset and suggests future research incorporating larger datasets from diverse renewable energy sources and countries to enhance the effectiveness of big data analytics in the energy sector.

In another research by By Dan Zhang, L.G. Pee, Shan L. Pan, Lili Cui et al. [4] assert that although managing and integrating various data sources is improving, big data still plays a limited role in the creation of sustainable smart cities. Through a case study of Wuhu, China, the paper identifies three phases of development in a sustainable smart city and proposes a framework for orchestrating different data sources. The study highlights the importance of addressing data-related issues, such as data quality, privacy, and political power, and provides insights into the interactions between data orchestration and other resources. By analyzing various data applications, the study contributes to both theoretical understanding and practical implementation of smart cities and big data. The results highlight the importance of efficient data management and integration to fully utilise big data's potential for enhancing urban management and accomplishing sustainable development objectives.

Another research by Vinaya Keskar, Jyoti Yadav, Ajay Kumar et al. [5]. This paper suggests that the examination of big data has begun in many businesses, and with the increasing advancement of internet technology, organizations are dealing with larger and more complex data. The study focuses on the banking sector and analyzes inconsistencies in credit card data due to big data technology, proposing methods to address these inconsistencies. The paper discusses the significance of data volume, variety, velocity, value, and veracity in the context of big data. It also explores different application areas such as business, healthcare, government, education, and scientific research. The study highlights the importance of data cleaning, storage, and analysis techniques like data mining and machine learning. It also discusses real-time data analytics, text analytics, multimedia analytics, and spatio-textual analytics. The article gives a broad overview of the opportunities and problems related to big data management across many industries.

by Lina Zhang, Dongwon Jeong, and Sukhoon Lee, et al. [6] This paper suggests that IoT data quality is crucial in various application areas, and choosing the appropriate technique for managing data quality in IoT systems can be challenging due to diverse requirements. To address this challenge, the paper surveys data quality frameworks, methodologies, and international standards for IoT data, comparing them in terms of data types, definitions, dimensions, metrics, and assessment dimensions. The survey aims to assist in narrowing down the choices for IoT data quality management techniques. The paper highlights the exponential growth of IoT platforms and the increasing amount of data generated, emphasizing the need for reliable data quality to avoid system failures and incorrect decision outcomes. By providing an overview of existing methodologies and frameworks, the paper contributes to the understanding of IoT data quality management and assists practitioners and researchers in selecting appropriate techniques for assessing IoT data quality.

Yassine Ramdane, Omar Boussaid, Doulkifli Boukraà, Nadia Kabachi and Fadila Bentayeb, et al. [7] . This study proposes a novel method to enhance the efficiency of OLAP queries in a distributed system built on top of Hadoop. The strategy emphasises improving star join and group-by aggregation operations, which are known to be costly in Hadoop database systems. The authors propose a data placement strategy that enhances these operations, allowing the query optimizer to perform a star join process locally in one Spark stage without a shuffle phase. They also introduce a dynamic approach to improve group-by aggregation. Experimental results on a 15-node cluster demonstrate that the proposed method outperforms existing approaches in terms of OLAP query evaluation time. The paper contributes to the field of data warehouse optimization in distributed systems and provides technical details and experiments to support the proposed approach.

Giuseppe Cattaneo, Raffaele Giancarlo, Umberto Ferraro Petrillo and Gianluca Roscigno, et al. [8] This paper suggests that the MapReduce computing paradigm, along with its implementations Hadoop and Spark, is gaining popularity in the field of bioinformatics due to its scalability and ease of use. The authors conduct a qualitative evaluation of several important MapReduce bioinformatics applications and highlight the importance of properly engineering applications to fully utilize the potential of distributed systems. The paper contributes to understanding the benefits of using MapReduce in bioinformatics and emphasizes the need for optimizing application design in distributed systems.

Sonia Cisneros-Cabrera a, Anna-Valentini Michailidou b, Sandra Sampaio, Pedro Sampaio and Anastasios Gounaris, et al. [9] This study makes the argument that by implementing quality requirements, query processing techniques combined with data quality management techniques can improve the quality of query returns. Studies evaluating the effectiveness and scalability of data quality

evaluation activities during query processing are, nevertheless, scarce. The authors undertake an empirical study utilising the Apache Spark big data computing framework to assess the effectiveness and scalability of data quality querying jobs across various computational platforms in order to close this gap. On massive traffic data sets, they apply a series of DQ-aware query processing tasks, taking timeliness, completeness, and correctness into account. The results show that employing optimised data science libraries along with the parallel and distributed capabilities of big data computing results in significant speed and scalability gains. The findings emphasise the advantages of employing big data frameworks and offer recommendations for choosing the best computing infrastructure for running data quality-aware queries.

Corinna Cichy, Stefan Rass, et al. [10] The significance of data quality in organisations and the difficulties in reaching and upholding high standards of data quality are the key topics of this presentation. The authors review and compare a number of data quality frameworks, together with their definitions, evaluation techniques, and process for improvement. The purpose is to give a thorough review of various frameworks and provide advice on how to choose the best strategy for data quality management based on particular requirements. The study focuses on the effects of bad data quality on firms, including monetary losses and hampered decision-making. The authors shed light on the multidisciplinary nature of the subject and the difficulties of measuring data quality. A decision-making guide is provided at the end of the paper to help organisations select the best data quality framework.

III. METHODOLOGY

1. Data Acquisition

The data acquisition module is responsible for gathering data from various sources, such as databases, APIs, or data streams. It ensures the retrieval of relevant and accurate data, considering factors like data completeness, data integrity, and data consistency. This module may involve data extraction, transformation, and loading (ETL) processes to collect and consolidate data from multiple sources into a centralized repository for further processing and analysis.

2. Metric Acquisition

To determine the degree of data quality, the metric acquisition module gathers and measures data quality metrics. It captures relevant data quality indicators and calculates metrics such as error rates, completeness percentages, and consistency scores. This module ensures the systematic tracking and measurement of data quality, enabling organizations to monitor their data quality performance over time and take corrective actions when necessary.

3. Data Quality Checks

The data quality checks module focuses on evaluating the quality of the acquired data. It employs predefined data quality metrics and rules to assess factors such as accuracy, completeness, consistency, and timeliness. This module performs automated data quality checks, flagging any data anomalies or inconsistencies. It helps identify and rectify data issues at an early stage, ensuring that the data meets the required quality standards.

4. Report Generation.

The report generation module is responsible for creating comprehensive and actionable reports based on the results of the data quality checks. It generates visualizations, summaries, and detailed insights regarding data quality issues, trends, and improvements. These reports are often generated in the form of tables, presenting data quality metrics, error statistics, and other relevant information. The reports are then sent to clients or stakeholders, enabling them to assess the quality of the data and make informed decisions. The data presented in the reports helps stakeholders understand the overall data quality status and facilitates discussions around data governance, data management, and process improvements.

5. Data Delivery

The data delivery module focuses on securely and efficiently transmitting the processed and validated data to its intended destination. It ensures data is appropriately packaged, formatted, and sent to the designated systems, databases, or applications. This module may involve data transformation, encryption, and the use of network protocols to enable seamless data transfer while maintaining data integrity and confidentiality. The data delivery module ensures that high-quality data reaches the right recipients for further analysis, decision-making, or downstream processes.

IV PROPOSED SYSTEM

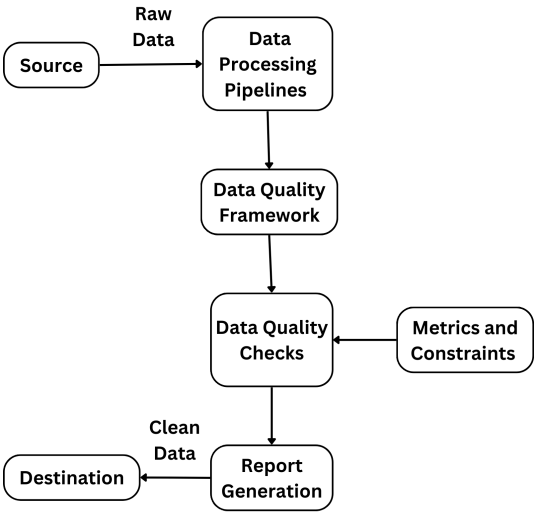

Figure 1: Architecture Diagram

The following are the main characteristics and elements of the suggested system:

The data quality checks element plays a crucial role in ensuring the accuracy and reliability of the acquired data. It involves implementing predefined rules and metrics to assess data quality factors such as completeness, accuracy, consistency, and timeliness. By conducting systematic data quality checks, the system can identify any anomalies or inconsistencies in the data and take appropriate corrective actions.

The metric acquisition is another important element of the system, where relevant data quality metrics are collected and measured. These metrics provide quantitative indicators of data quality and can include measures like error rates, completeness percentages, or consistency scores. By capturing and analyzing these metrics, the system can track the performance of data quality over time and identify areas that require improvement.

The report generation element focuses on creating comprehensive reports based on the results of the data quality checks and metric analysis. These reports present insights, trends, and key findings related to data quality in a clear and understandable manner. Visualizations, summaries, and detailed information are often included to facilitate decision-making and provide actionable recommendations for improving data quality.

By integrating these three elements, the proposed system aims to establish a proactive approach to data quality management. It ensures that data is thoroughly checked for quality, important metrics are captured and analyzed, and comprehensive reports are generated for stakeholders. This enables organizations to make informed decisions, address data quality issues, and continuously improve the overall quality of their data assets.

IV. CONCLUSION

In conclusion, this project highlights the importance of ensuring accurate and reliable data for decision-making processes. The proposed framework, covering data integration, migration, warehousing, and analytics, automates data quality checks and enables prompt identification and resolution of issues. By combining the framework with existing data governance frameworks, a comprehensive approach to data management is achieved. Exploring cloud-native deployment options enhances scalability and flexibility while addressing infrastructure management complexities. The future integration of machine learning techniques can streamline data quality improvement, reducing manual effort. These efforts contribute to improved data quality, ultimately leading to better business outcomes.

REFERENCES

[1] Maxi Kindling, Dorothea Strecker, et al., "A Survey on Data Quality Assurance at Research Data Repositories", Data Science Journal (2022)

[2] Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui & Rachida Dssouli, "Big data quality framework: a holistic approach to continuous quality management", Journal of Big Data (2021)

[3] Noha Mostafa, Haitham Saad Mohamed Ramadan, Omar Elfarouk, "Renewable energy management in smart grids by using big data analytics and machine learning", Machine Learning with Applications (2022)

[4] Dan Zhang, L.G. Pee, Shan L. Pan, Lili Cui, "Big data analytics, resource orchestration, and digital sustainability: A case study of smart city development", Government Information Quarterly (2022)

[5] Vinaya Keskar, Jyoti Yadav, Ajay Kumar, "Perspective of anomaly detection in big data for data quality improvement", Materials Today Proceedings (2022)

[6] Lina Zhang, Dongwon Jeong, and Sukhoon Lee, "Data Quality Management in the Internet of Things", Sensors (2021)

[7] Yassine Ramdane, Omar Boussaid, Doulkifli Boukraà, Nadia Kabachi, Fadila Bentayeb, "Building a novel physical design of a distributed big data warehouse over a Hadoop cluster to enhance OLAP cube query performance", Parallel Computing (2022)

[8] Giuseppe Cattaneo, Raffaele Giancarlo, Umberto Ferraro Petrillo, Gianluca Roscigno, "MapReduce in Computational Biology Via Hadoop and Spark", Encyclopedia of Bioinformatics and Computational Biology (2019)

[9] Sonia Cisneros-Cabrera, Anna-Valentini Michailidou, Sandra Sampaio, Pedro Sampaio, Anastasios Gounaris, "Experimenting with big data computing for scaling data quality-aware query processing", Expert Systems with Applications (2021)

[10] Corinna Cichy, Stefan Rass, "An Overview of Data Quality Frameworks", IEEE Access (2019)