International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 2, Issue 3, 43-52, 2021
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V2I3P105

Original Article

Data Quality in the Age of Big Data: Challenges and Best Practices

Sonika Darshan
Independent Researcher, USA.

Abstract *- In the era of big data, more and more organizations use huge, various, and fast-moving big data to perform data analysis, create products, and make critical decisions. Nonetheless, this has compounded the data quality issue, an important parameter for obtaining meaningful and reliable statistics from the data. Conventional concepts such as accuracy, completeness, and data consistency do not fully correspond to the conditions of the big data veins: volume, variety, velocity, and veracity. The present research aims to discuss the existing and emerging facets of data quality in big data, including various issues like heterogeneity, real-time validation for enormous datasets, absence of standardization, and lack of metadata. It discusses previous studies and tools that check the quality of data according to the multiple and complex qualities, such as availability, usability, reliability, relevance, and presentation of the data. In the following section, we discuss the data governance process and automated validation, the importance of data metadata management, and how emerging research applications can be used in identifying anomaly and data quality prediction. Moreover, the paper highlights the validity issues and the necessity to monitor, protect, and respect people's data with regard to relevant legislation. This day highlights the trends of growing integration between quality assurance and data architecture, and the future trends consider semantic expansion, quality control on the edge level, and the accountability. This work is intended to serve as a reference for researchers, practitioners, and organizations on how to create sustainable solutions for data quality management for big data systems.*

Keywords *- Big Data, Data Quality, Data Governance, Metadata Management, Data Architecture, Data Integration.*

1. Introduction

This study describes big data as an unprecedented approach to attaining, storing, processing, and utilizing information by organizations. With the increased use of social media, IoT devices, and various other enterprise and transactional systems, the generation of big data has positioned businesses and institutions with unlimited pools of data that can be of immense value. [1-3], but the base data should be correct, unified, and reliable to derive big data's advantages. Therefore, High-quality data has emerged as a critical issue of interest in the current data environment. The current approaches aimed at quality assurance are not sufficiently adequate for the context of big data. The four key dimensions, including volume, velocity, variety, and veracity of big data, are considered defining characteristics of this data type and are challenging in several ways. Data volume characteristics, on the one hand, put pressure on the available storage and computing capacity. Velocity leads to real-time quality of data feeding that needs to be integrated rapidly. The data could be in the form of structured databases, unstructured text, images, and video, thereby posing challenges in the process of standardization as well as validation. Last of them, veracity refers to the issue of how certain data is trustworthy and unambiguous. It also highlights the quality problem versus data noise or noise and misinformation.

This is the case since data quality is fed into a business, which affects its operations and decision-making process. Failure to provide proper input data may lead to a low-quality analysis, customer dissatisfaction, compliance issues, and high expenses. In the case of organizations that have incorporated data as the major strategic tool in decision-making, data quality is not just a technical problem but a strategic one. Based on these challenges, there is increased pressure to develop systematic and efficient methods for data quality management in big data systems.This ranges from state-of-the-art technology for Quality Control and Quality Assurance to organizational enablers such as sound data management and control, delineation of data ownership, and data auditing procedures. When specifically Understanding Big Data and following particular successful recommendations, organizations could improve data quality and make tangible use of large amounts of their data. The purpose of this paper is, therefore, not only to identify critical data quality issues arising out of big data environments but also to understand available solutions and trends for addressing them.

2. Related Work on Data Quality in the Age of Big Data

2.1 Challenges of Data Quality in Big Data Environments

Data quality issues in big data environments include four key factors, the 4Vs, which will be discussed herein. All of them are associated with issues that complicate data management in the context of big data. [4-6] The most significant concern currently stands out is the variety of sources and data types. Big data is derived from various sources and non-organizational systems such as social media platforms, IoT devices, mobile applications, website logs, and scientific instruments, unlike national enterprise data derived from organizational systems' databases. The data is not only big but also categorical into structured, semi-structured, and unstructured data in text, images, videos, and audio. This means that in the data integration, transformation, and standardization process, it becomes challenging to ensure that the quality of data being processed in all environments is of equal quality.

Another major challenge is that there are no commonly agreed specifications and definitions for assessing data quality in the context of big data. Earlier, data producers and consumers were usually the same, which gave the management more transparent directions regarding quality requirements. Nevertheless, the user often uses data in large-scale environments while remaining unrelated to its source. This disconnection necessitates a shift in quality assessment paradigms from producer-defined to user-centered evaluation. Therefore, several researchers have postulated a framework that categorizes data quality based on multiple factors: availability, usefulness, credibility, appropriateness, and presentation quality. Some of the indicators are as follows: each dimension is standard, but their measurement depends on the data's use and domain.

2.2 Quality Criteria and Assessment Frameworks

The related academic and professional literature has proposed several multidimensional frameworks for measuring data quality in big data environments. These frameworks accept that developing one measure of data quality is difficult. Key dimensions commonly cited include:

- Accessibility compares how current the data is for the users, the availability of the system, and permissions granted.
- Usability looks at whether the data is as documented as possible and whether it has various metadata, user instructions, and clarified schema.
- Reliability is the ability of the collected data to self-check its accuracy, consistency, and completeness, as well as the possibility of tracing the data sources.
- Relevance, in this case, addresses the nature and range of the data in relation to the needs, requirements, and intended use of the user.
- Presentation Quality is the characteristic that deals with how data is organized and how the users will easily understand it when presented.

These dimensions differ by domain very much. For instance, for social media data, availability and credibility may be immediate since social media data is real-time data. For biomedical or genomic data, some challenges may include standardization due to diverse data formats and data storage techniques. This is stipulated by the fact that quality assessment models must adapt to the changing data use cases while being sensitive to particular domains.

2.3 Impact of Data Quality on Big Data Management

Poor quality data minimizes the value an organization can realize from its big data ventures. The volume and velocity of data growing from various sources and high data variety make the problem of quality control unsolvable with conventional approaches. This results in problems such as unequal data formats, missing entries, and incorrect data input into the analytical processes. These vices produce inaccurate information and lead to money losses, non-adherence to the law, and wrong business decisions. In particular, data management in the era of big data requires well-defined policies on data quality that have been put into place and agreed upon in an organization.

For the success of such projects, companies must adopt a shared responsibility model that spans IT, business, and analytics. They should be made as part of data entry and maintained consistently to cover data drift and decay concerns. Furthermore, automated tools for monitoring data quality and quick detection of outliers are also useful for disseminating quality control on a vast scale and keeping the process as real-time oriented as possible.

2.4 Trends Indicating Worsening Data Quality

While there is enhanced awareness and more innovations in data, recent research indicates that data quality issues are worsening. The survey shows that 57% of analytics professionals selected poor data quality as the most important factor in data preparation, which was relatively higher than the previous survey in 2022 with 41%. This is in light of the great advancement of big data environments and the fact that the pace of data generation currently outdoes the quality management capacity. When organizations that require accurate data, the presence of low-quality data is extremely damaging. The main points above suggest an intelligent and scalable data quality management solution that is adequate for today's growing state of data.

3. Challenges in Data Quality for Big Data

As the adoption of big data continues to grow across industries, maintaining high data quality becomes increasingly difficult. Volume, velocity, variety, and veracity are the four dimensions that make big data big data, and each of them brings a set of issues that are not easily addressed by conventional data quality approaches. [7-10] The following part outlines the key data quality issues in the context of big data concerns, threats emerging due to scale, complexity of the data, and the nature of data in terms of what it contains.

3.1 Volume, Velocity, and Variety Issues

One major factor that makes data quality a difficult subject to handle is the current vast amount of data being produced. Organizations receive terabytes and even petabytes of data daily in social media platforms, sensor networks, transactions, and many others. This sheer quantity undoes data quality tools and makes inspection or validation very difficult to be done manually. As such, there are high chances that a mistake or a replication of a record and errors within one system can go unnoticed and replicate in other systems. Equally problematic is the concept of velocity, the rate of information production, and the rate at which the data has to be analyzed.

For instance, logistics or telecommunications near real-time and real-time analytics are essential in finance. However, intense data streams perpetually draw the time for data quality checks, reducing the ability to avoid ineffective data in decision-making. There is even greater variety since big data comes from structured databases, different semi-structured logs, and text, voice, videos, images, and the like. It is almost impossible to have a uniform quality of data stored in such diverse formats, and often, the format will call for different validation methods. When combined, volume, velocity, and variety make it very challenging to ensure the quality of the data, and this requires very effective solutions that are automatable and scalable.

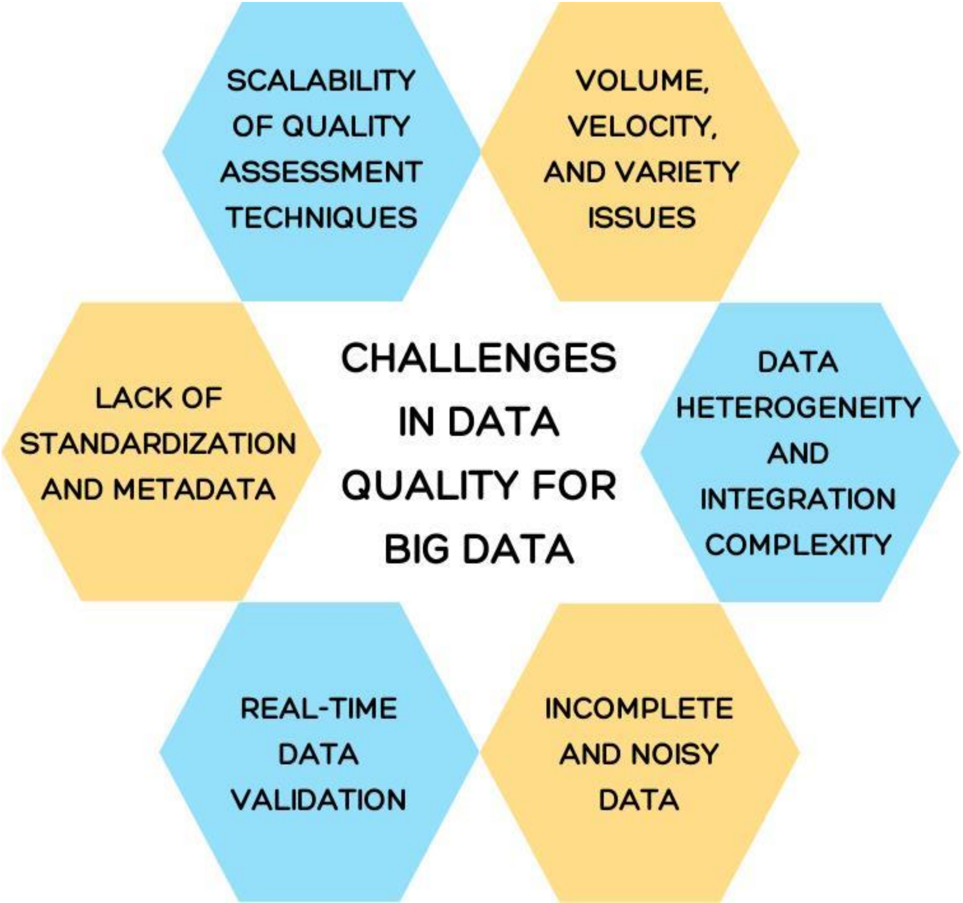
SCALABILITY OF QUALITY ASSESSMENT TECHNIQUES
VOLUME, VELOCITY, AND VARIETY ISSUES
LACK OF STANDARDIZATION AND METADATA
CHALLENGES IN DATA QUALITY FOR BIG DATA
DATA HETEROGENEITY AND INTEGRATION COMPLEXITY
REAL-TIME DATA VALIDATION
INCOMPLETE AND NOISY DATA

Fig 1: Key Challenges in Data Quality for Big Data

3.2 Data Heterogeneity and Integration Complexity

Big data environments are difficult to homogeneous because they consist of very diverse types of data and come in different formats consisting of data with different structures and meanings. This kind of data heterogeneity constitutes the biggest challenge to the integration process since it is the preliminary step toward attaining an all-encompassing data analysis. Many systems have dissimilar data standards and terminologies, resulting in semantic differences and compatibility issues whenever one try to integrate data into one system. For instance, if the data are gathered from a mobile app, it may not follow the same format as the data in a CRM system or even data derived from social networks. Another problem in dealing with multiple schemas is that there is no common model to understand the location of attributes, merge records that cross schema boundaries, and handle context information.

Furthermore, using third-party or open data sources significantly complicates the data's actuality and credibility. As a result, to address these problems, organizations will have to work towards procuring data mapping tools, metadata, and ontology frameworks. Nevertheless, these solutions are rather complex and need considerable resources; even in this case, inconsistency cannot be wholly eliminated. As a result, data heterogeneity and integration, which increase the challenges of achieving high-quality data, are still evident in big data systems.

3.3 Incomplete and Noisy Data

Among challenges associated with big data, data incompleteness, and noise are important factors that have significant impacts. Often, data is collected ad hoc rather than systematically, interfering with data quality issues, such as gaps within the data, missing values, and records. For instance, there might be issues with sensor devices, such as temporary malfunction. Social media data might also be less detailed in terms of demographics, and in most cases, user-generated content can be unformatted or confusing. Thus, noisy data is defined as irrelevant, extraneous, or incorrect data and can greatly affect the quality of analytics. This refers to typing errors, existing records, unnecessary entries, or noisy data owned by a faulty instrument.

The use of noise not only distorts insights but also has a negative impact on the training and competence of automated validation systems since they depend much on the quality of input. Problems of this nature are well managed using data cleaning techniques, but traditional approaches are inadequate given the size and rate of big data. Sophisticated techniques used in the current data cleaning include anomaly detection, data imputation, and emerging research. Finally, it is important to note that the listed methods are imperfect and should be adjusted from time to time.

3.4 Real-Time Data Validation

Since real-time data is used in various aspects of an organization, such as analytics, decision-making, and automation, the quality of the incoming data is a considerable concern. [11-13] While, for example, batch data can be preprocessed and validated to a greater extent, the streaming data must be analyzed in real time, and there is little time for rigorous data quality checks. This is especially important in financial trading, healthcare surveillance, or industrial IoT, where even a small percentage of deviated measurement yields dangerous consequences. Real-time data validation checks the validity of the data's accuracy, completeness, consistency, and reasonability as it is fed to the system.

However, the data speed and amount require fast, lightweight, and automated check- and balance mechanisms that do not add delay. Furthermore, trends and outliers have to be pinpointed ‘in real-time.’ This may occur when control is unavailable due to the lack of a sizable historical database combined with inadequate training. Such systems’ creation calls for implementing a technique known as event-driven architectures and Complex Event Processing (CEP) suited to issue real-time alerts after detecting anomalous data patterns. However, real-time validation continues to be less of an achievement because introducing such systems is technically demanding and computationally expensive.

3.5 Lack of Standardization and Metadata

Lack of data quality is defined in the big data environments by the inconsistency of data formats, structures, and semantics. Since big data emanates from a number of sources where data elements from one source exhibit dissimilar formats, naming standards, units of measure, and encoding conventions to those of another source, integration, and interpretation of the total data set in a meaningful manner is not possible without the format. Some problems that may be observed when working with datasets include absence or insufficient documentation and ambiguous or inconsistent data schemas. The lack of metadata is semantically related, which is crucial for data lineage, background information, and applicability.

As such, metadata assists a user in knowing when, how, and by whom the data was created and under which circumstances. Due to the unavailability of proper metadata, the user cannot be assured of the data's relevance, its credibility, and the use they may put it. Also, poor quality impacts metadata because it affects such activities as discovery, transformation, and data governance. The strategies for enhancing standardization include leveraging data standards focusing on industry-specific data dictionaries and ontologies. Still, the adoption has been deemed to lack uniformity across different organizations and sectors. Creating sustainable metadata is equally difficult, especially in complex environments where data must be ingested and processed quickly without technical supervision.

3.6 Scalability of Quality Assessment Techniques

A potential problem of big data quality management is the scalability issue of the assessment of big data. With the volume, variations, and velocities of Big Data, conventional data quality methods like auditing, Business rules checking, and Sample checking are not effective. As datasets and pipelines expand into terabytes and petabytes, QA across the several phases becomes time-consuming and uneconomical to do manually without automation. To complement these claims, the data quality solution implementations must be deployable on Hadoop or any other distributed system and be able to take advantage of parallelism to operate on large datasets.

This includes creating a range of auto-populating profiling instruments, a data quality assessment scale, and a real-time working activity dashboard that can function on the cloud and hybrid systems. Emerging research also benefits the identification of a large number of quality problems through data abnormality detection, value prediction, and value classification. At the same time, these innovations do not allow for achieving the necessary scalability and adaptability of the data quality management system. It is important to note that all provided solutions are generally disparate and not reusable. Moreover, most companies cannot freely implement these tools in their organizations' existing architectures and investment plans. Therefore, building up reliable and easily expandable methods for data quality evaluation remains a critical and emergent issue in the big data domain.

4. Architectural Approaches to Data Quality Management

The context of big data is not just the theory, the general framework, and the best practices that are needed but the actual architecture that can effectively support the flow and the management of data at a large scale and a high velocity. [14-16] the architectural diagram depicted in Figure 2 identifies some of the components, tools, and processes required in handling data quality, from the ingestion process to its usage in analytics. Such a view is critical to review how the various technical layers in data technologies enable data consistency and accessibility in vast and complex data environments.

These are entries from the Internet of Things devices, social media, enterprise databases, APIs, or weblogs. In each of them, more structural and semantic specificities influence the variety and truth issues typical of big data systems. These raw inputs feed through the ingestion and integration layer, which includes batch and streaming data pipelines, EL and ETL processes, and integration tools that cleanse data that passes through quality control. The so-called Quality Assessment & Validation block is located at the core of the presented architecture, where various methods are used to regulate and enhance data characteristics.

Schema evaluation promotes methods that check for the adherence of the data with immense structures, while duplicate detection and missing value handling promote the handling of data completeness. There is profiling and checks of internal consistencies. Then, there are more complex anomaly detection methods, which may be based on emerging research and aimed at catching subtle or emerging issues. Standardization and normalization are the module's responsibility, which will help ensure that data is compatible and can readily be integrated and analyzed.

Supplying these operations is a strong governance and metadata management structure. Some of these are metadata repositories, lineage trackers, and access control systems, which track transformation and ensure policy compliance. These systems are particularly responsible for visibility and accountability since the context in which data is used is captured, and the data stewards get a chance to make quality-related decisions. Metadata is also used for automatic quality control and helps the user comprehend the meaning and source of data.

This processed data is then pushed to several repositories for raw or low-level structured data in data lakes, for the structured data ready for analysis in data warehouses, or for NoSQL/object storage for unstructured formats. Then, it gets consumed by dashboards, API, and Notebooks for business intelligence, data visualization, and highly sophisticated data science. This type of coordination ensures that each output layer enjoys the quality controls implemented earlier in the process, guaranteeing that analytical results are derived from sound, high-quality inputs.

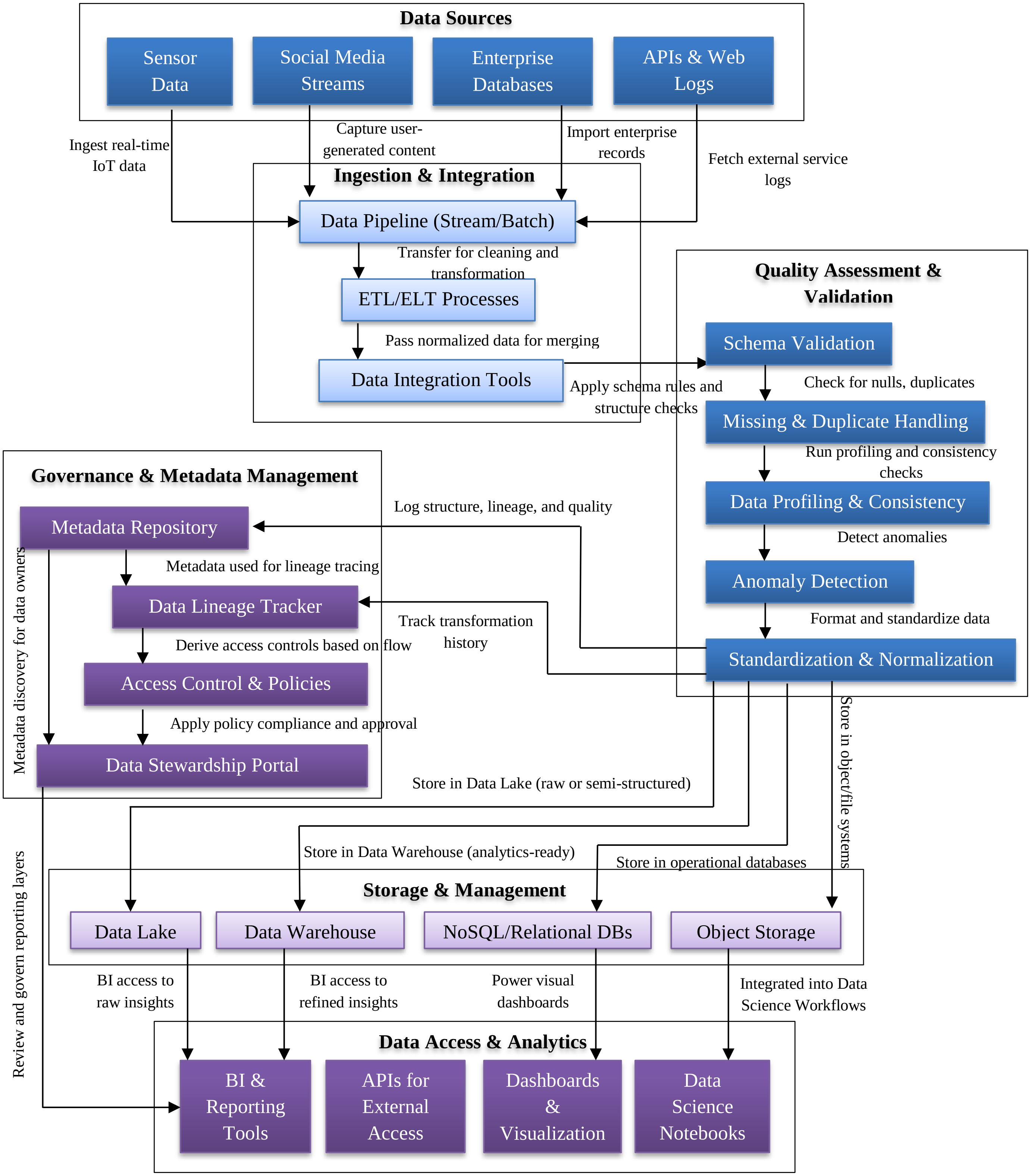
Data Sources
Sensor Data
Social Media Streams
Enterprise Databases
APIs & Web Logs
Ingest real-time IoT data
Capture user-generated content
Import enterprise records
Fetch external service logs
Ingestion & Integration
Data Pipeline (Stream/Batch)
Transfer for cleaning and transformation
ETL/ELT Processes
Pass normalized data for merging
Data Integration Tools
Apply schema rules and structure checks
Quality Assessment & Validation
Schema Validation
Check for nulls, duplicates
Missing & Duplicate Handling
Run profiling and consistency checks
Data Profiling & Consistency
Detect anomalies
Anomaly Detection
Format and standardize data
Standardization & Normalization
Governance & Metadata Management
Log structure, lineage, and quality
Metadata Repository
Metadata used for lineage tracing
Data Lineage Tracker
Track transformation history
Derive access controls based on flow
Access Control & Policies
Apply policy compliance and approval
Data Stewardship Portal
Metadata discovery for data owners
Store in Data Lake (raw or semi-structured)
Store in Data Warehouse (analytics-ready)
Store in operational databases
Store in object/file systems
Storage & Management
Data Lake
Data Warehouse
NoSQL/Relational DBs
Object Storage
BI access to raw insights
BI access to refined insights
Power visual dashboards
Integrated into Data Science Workflows
Review and govern reporting layers
Data Access & Analytics
BI & Reporting Tools
APIs for External Access
Dashboards & Visualization
Data Science Notebooks

Fig 2: Data Quality Architecture

5. Best Practices for Ensuring Data Quality

5.1 Data Governance and Stewardship

Data quality management can be only effective when organizations have the necessary data governance and stewardship framework. [17-20] It manages data as a valuable resource by formulating rules and guidelines. As part of big data governance, this involves the question of who owns the big data, which can use it, and how it can be managed amongst the business units. Others are regulatory compliances like GDPR HIPAA or any industry-specific standards, many of which are strict on data accuracy and adequacy. Data stewardship can be considered a good fit with governance since it handles the specific aspects of its implementation. The stewards' responsibilities include observing data quality to determine the best way of correcting errors and standardizing data. In the current data management structures, specialized personnel can utilize stewardship portals to alter rules regarding the form of data to be processed, oversee quality reports of data, and then approve or reject datasets before integration into analytics processes. When combined, governance and stewardship promote the responsible use and production of data by ensuring accountability for the data in the organization, which is very important in large organizations.

5.2 Automated Data Cleaning and Validation

Automated Data Cleaning and Validation are impossible due to the volume and rate at which big data is generated and processed. These include problems related to duplicate data, missing data, improper formats, and outliers that can be quickly detected and handled by means of automation. Data cleaning is usually performed within the data pipeline, specifically as part of ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform) activities where the incoming data is standardized, augmented, and prepared for analysis.

Automated validation mechanisms are mostly rule and/or algorithm-based procedures that check on the compliance of the data with the chosen schema, ensure compliance with logical business standards of data processing, and call out data that violates laid-out expected patterns. Higher-level forms also use to automate the analysis and learning of error profiles and deviations from them. For instance, a time series analysis model can detect slight changes in patterns within the sensor profiles, which may signify a failure or corruption in equipment. Not only does it decrease the work for data engineers, but the effectiveness and efficiency of data checking are also improved.

5.3 Metadata Management

Metadata is typically defined as data about data, but big data is crucial in guaranteeing integrity in big data systems. They give information about the data’s source, format, ownership, history of previous conversions, and purpose. This can pose a problem since users may find it difficult to know whether or not a particular dataset is suitable or reliable for their study. Metadata management systems, therefore, extend to identifying and capturing this contextual data throughout the whole data chain. The contemporary world's metadata management frameworks comprise metadata repositories, lineage tracking systems, and data catalogs. These enable the identification of data, monitor data movement across systems, and allow for viewing transformation processes.

For instance, the data lineage tracker illustrates how raw data obtained from IoT sensors was processed and transformed for analysis. This makes auditors’ work easier, enables the reproduction of results, and allows the tracing of potential errors. Further, metadata is well applied to automation as most workflows and processes can use data attributes or quality scores to be reformed automatically. The fact that metadata are incorporated in quality management systems essentially offers increased visibility to the data while enabling users to make adequate decisions. In this way, metadata management is key, as it provides the basic support for big data processing, which is needed to bring efficiency and quality to big data analytics.

5.4 Continuous Monitoring and Quality Metrics

In big data architecture, one of the significant factors emphasized is the ongoing assurance that the data being used in analysis and operating systems is of high quality. Data quality cannot be seen as a one-off project but rather a continual process that needs to be monitored and maintained regularly. Continuous monitoring systems are based on the use of complex routines according to which results of the data quality analysis are reviewed repeatedly with a focus on a number of aspects, including the accuracy, completeness, consistency, time relevance, and uniqueness of the data in question. All these systems provide quality metrics and dashboards, which enable the data teams to get a pattern or pattern they see in case they need help identifying an issue to solve.

These metrics should be set up and monitored in the long term so that the organizations can compare them to the data quality standards and SLAs. For instance, a data product that supports customer analytics may include an SLA that is 98% accurate and ought to have no more than 0.1% of records as duplicates. Within the architecture, it is possible to incorporate monitoring into the process, often in the data ingestion process and the layer in which the data is stored. Then, teams can receive alerts once such quality control has been violated. This enhances the credibility of the data collected and reduces vulnerability when making critical operations decisions in data-intensive settings. Moreover, the gathered data can generate governance reports, enable audits, and be the basis for applying improvements through the data life cycle.

5.5 Privacy, Ethics, and Compliance

Data's ethical and legal use increases in conjunction with the volume and architecture of the systems in use. Privacy, ethics, and compliance are not the things one must add to data quality management; they must be the spine of the structure. It must also be legal, and the data acquisition process should be moral, meaning that the information used must be well protected, particularly in the form of Personally Identifiable Information (PII), financial or even health data. These problems can lead to legal consequences, loss of reputation, and customer loyalty's erosion. In order to address these concerns, big data architectures support privacy preservation methods such as data anonymization, masking, and differential privacy. These techniques ensure that one cannot reverse engineer the data and determine the identity of the bare individuals involved while at the same time achieving some good results.

Regulatory compliance tools help track data protection laws like GDPR, CCPA, and HIPAA by detecting access rights consent and keeping a record. Moreover, data management has integrated ethical principles to reciprocate organizational and social responsibilities. Integrating the principles of privacy and compliance into the information processing framework enables one to promote the use of big data while protecting the rights of the persons and ensuring public trust in digital technology. This view is becoming increasingly widespread as an acknowledgment that data quality means much more than data accuracy or completeness; it is also about legal compliance and, in effect, decency with respect to the information held.

6. Discussion

The need for high-quality data as a competitive asset for organizations has been an essential objective as more organizations maintain, use, and innovate using big data. The basic concept that used to be generally applied to data quality has to do with elements like accuracy, completeness, and consistency and has been adjusted to include factors such as timing, relevancy, and ethical considerations. Therefore, issues elicited by high volume velocity and variety must be catered for with intelligent architectures that are also flexible in this context. It is evident from the discussion of quality criteria and frameworks that data quality management must be integrated based on consumption patterns and organizational objectives. In evaluating the architectural approaches that were proposed and implemented, data quality appears to be a system's responsibility that is integrated throughout the system, including data ingestion, data processing, data storage, data access, and data governance.

In quality check process makes it easier to detect errors in datasets and help in the predictive maintenance of them; in metadata management, it makes it easy for users to be aware of data pipelines and make them more transparent. However, using such systems also poses some challenges, model drift, there is also an issue with regard to metadata and governance issues. These depicted forms must be periodically assessed and adjusted to ensure they are in lockstep with enterprise goals. Moreover, it is crucial to mention ethics, privacy, and compliance as significant aspects of any company. As more and more business organizations depend on personal and sensitive information, the possibility of leaking it, getting biased results, or non-compliance rises. Therefore, in formulating the current definition of data quality, ethical dimensions that dictate the kind of data usage shall be an important aspect since data must not be applied in a wrong, unlawful, or non-democratic manner.

Shifting data quality with privacy and other compliance laws highlights the importance of good structures and technology solutions that can detect the law change to the organization's needs. In summary, data quality in the era of big data is a complex and not a simplistic issue and cannot be resolved merely with the help of some fixed model. It requires organizational architecture and architectural thinking that materializes automation, governance, and intelligent systems at the data confluence. This discussion also implies that data quality is a continuous process complemented by organizational culture, ample infrastructure, and research collaboration. Reliable and well-founded data quality initiatives can extend to future and more innovative technologies like rules-based anomaly detection, real-time analytics.

7. Future Directions

The future, therefore, holds great changes in how data quality management in big data environments will evolve, hence changes in automation, and regulations. The forecast for the coming year is most inspiring by embedding autonomous data quality systems into the support framework. This system will not only detect and correct mistakes that may occur in a process but will also be able to predict error cases based on previous experiences. The proactive capabilities of such claims will significantly minimize the role of human intervention in the context of quality management as it exists today, which will prove highly beneficial in highly transactional areas of operation such as finance, healthcare, and the IoT, amongst others.

The use of semantic technologies and knowledge graphs at a global scale for enhancing the compatibility of means for data analysis. This is especially so because a new kind of dataset is increasingly emerging and needs contextualization and integration as big data gets even bigger and more complex. Semantic structures add to basic metadata further richer, more easily interpretable meaning, making it easier to link, assess data quality, and perform analysis across domains. This is the evolution towards Smart Metadata whereby the systems can decide when and how data is used or even if it is suitable for use. Real-time and edge-based quality management will also gain prominence more than ever as much of the data is collected from decentralized sources such as mobile devices or other smart devices and edge computing.

When the data quality is built into the edge or checked at the edge before it is transmitted to more centralized repositories, the latency is decreased, the response time is improved, and the generation of low-quality data that affects all other systems is avoided. Local governance policies and lightweight, embedded quality checks will be important for these decentralized models. The concepts important today for ethical and sustainable management of data practices will define the significance of the qualitative level in the future. Organizations will also have to incorporate ethical standards into the measurement and evaluation of data quality in a fair and sustainable way regarding handling features like the transparent data lineage, consent-based data pipeline, which will be part of future-ready data quality architecture. These are new paths to better, responsible, and fairly superior ways of managing data quality with globalization and a highly demanding world of data.

8. Conclusion

Data quality is viewed today as the basic precondition for efficient decision-making given the current highest amounts of data. There are issues such as a lack of standardized formats, completeness, and even data definitions, and a new issue rapidly growing: ethical considerations. As discussed in this paper, mitigating the mentioned challenges requires using strong data governance practices, advanced and automated validation processes, practical approaches for metadata management, and applying intelligence and learning concepts. The trends show that with the development of architectural solutions, data quality becomes not an add-on but an important and indispensable component of the modern data platform.

It can only be done tactfully; it needs consistent surveillance and adaptable structures, and the organization must demonstrate accountability to technological advancements and ethical principles. What is yet to come is a combination of trends like real-time validation at the edge, the introduction of knowledge graphs for enhancing the semantics of data, and the core concept of making more responsible in its data quality monitoring efforts. In conclusion, the distinctive ‘best practices’ of quality data management introduced by this paper will enable organizations to collect good quality data that will last long, allow for higher returns on investment, and, most importantly, supply legal and trusted data.

Reference

[1] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data science journal, 14, 2-2.

[2] Abdullah, N., Ismail, S. A., Sophiayati, S., & Sam, S. M. (2015). Data quality in big data: a review. Int. J. Advance Soft Compu. Appl, 7(3), 17-27.

[3] Ramasamy, A., & Chowdhury, S. (2020). Big data quality dimensions: a systematic literature review. JISTEM-Journal of Information Systems and Technology Management, 17, e202017003.

[4] Saha, B., & Srivastava, D. (2014, March). Data quality: The other face of big data. In 2014 IEEE 30th International Conference on data engineering (pp. 1294-1297). IEEE.

[5] Cai, L., & Zhu, Y. (2015). Data quality and data quality assessment challenges in the big data era. Data Science Journal, 14, 2-2.

[6] Taleb, I., Serhani, M. A., & Dssouli, R. (2018, July). Big data quality: A survey. In 2018 IEEE International Congress on Big Data (BigData Congress) (pp. 166-173). IEEE.

[7] Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International journal of information management, 34(3), 387-394.

[8] Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. Journal of Database Management (JDM), 26(1), 60-82.

[9] Cappiello, C., Pernici, B., & Villani, L. (2014, October). Strategies for data quality monitoring in business processes. In International Conference on Web Information Systems Engineering (pp. 226-238). Cham: Springer International Publishing.

[10] Bhaskaran, S. V. (2020). Integrating data quality services (dqs) in big data ecosystems: Challenges, best practices, and opportunities for decision-making. Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems, 4(11), 1-12.

[11] Furner, J. (2020). Definitions of “metadata”: A brief survey of international standards. Journal of the Association for Information Science and Technology, 71(6), E33-E42.

[12] Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience, and acquisition intention of big data analytics. International journal of information management, 34(3), 387-394.

[13] Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. Journal of Database Management (JDM), 26(1), 60-82.

[14] Bickmore, T. W. (1994). Real-time sensor data validation (No. E-8672).

[15] Becker, D., King, T. D., & McMullen, B. (2015, October). Big data, big data quality problem. In 2015 IEEE international conference on big data (big data) (pp. 2644-2653). IEEE.

[16] Abdallah, M. (2019, February). Big data quality challenges. In 2019 International Conference on Big Data and Computational Intelligence (ICBDCI) (pp. 1-3). IEEE.

[17] Dautov, R., & Distefano, S. (2017, December). Quantifying volume, velocity, and variety to support (Big) data-intensive application development. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2843-2852). IEEE.

[18] Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016, May). A model-driven architecture-based data quality management framework for the Internet of Things. In 2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech) (pp. 252-259). IEEE.

[19] Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. (1998). Architecture and quality in data warehouses. In Advanced Information Systems Engineering: 10th International Conference, CAiSE'98 Pisa, Italy, June 8–12, 1998 Proceedings 10 (pp. 93-113). Springer Berlin Heidelberg.

[20] Nelson, C., Lindell, M., Hopkins, E., Abramowitz, A., Hinkley, P., de Kerchove, G. & Flynn-Heapes, E. (2007). Managing quality in architecture. Routledge.