International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I11P106

Volume 71 Issue 11, 40-45, November 2023

Original Article

Data Governance and Quality Management in Data Engineering

Alekhya Achanta[1], Roja Boina[2]

[1]*DataOps Engineer, Continental Properties Company Inc, Wisconsin, USA.*
[2]*Independent Researcher, North Carolina, United States of America*.

Corresponding Author : alekhya.achanta@gmail.com

Received: 16 September 2023 Revised: 21 October 2023 Accepted: 08 November 2023 Published: 25 November 2023

Abstract - *Data has become one of the most valuable assets for organizations today. With the exponential growth in data, effectively governing and managing its quality is critical for gaining business insights and maintaining regulatory compliance. This paper examines the importance of data governance and quality management in data engineering. It outlines the fundamental principles, processes, and best practices for implementing robust data governance frameworks and quality management programs. The roles of key stakeholders, such as data owners, stewards, and engineers, are discussed. It also explores the challenges, such as inadequate data quality culture and lack of executive support. The focus is on new technologies, such as machine learning and automation, which can potentially improve data governance and quality. The paper concludes by emphasizing the need for a holistic strategy, strong leadership, and a collaborative culture for successful data governance and quality management outcomes.*

Keywords - *Data governance frameworks, Data profiling and monitoring, Data validation and standards, Data quality assurance, Metadata management*.

1. Introduction

In the era of digitization, the ability to harness data has revolutionized the competitive landscape for organizations. In their pursuit to be data-centric, these entities continuously amass vast troves of structured and unstructured data from diverse avenues like social media, the Internet of Things, sensors, clickstreams, and transactions (Ghavami, 2020). While the potential to extract value from this data is immense, there exists a significant research gap and practical challenge: maintaining the quality of this data. Poor-quality data can obfuscate genuine insights without robust governance and judicious management, thereby misguiding business strategies. Recent empirical studies paint a dire picture—on average, businesses suffer a staggering loss of over $15 million annually due to inferior data quality. Furthermore, this compromised data quality not only leads to flawed business decisions but also poses severe regulatory risks, which have culminated in penalties amounting to millions (Wang et al., 2022).

Given this backdrop, the criticality of instituting strong frameworks and methodologies to govern and ensure data quality cannot be overemphasized. This paper delves deep into the realm of data governance and quality management, particularly within the ambit of data engineering. Data engineering, which encapsulates the myriad processes and systems for procuring, housing, and dissecting data at scale, forms the bedrock of any data-driven decision-making mechanism. Our discourse begins by elucidating the fundamental concepts of data governance and quality. Following this, we highlight the pivotal role of data quality management in the intricate pipelines of data engineering. As we progress, the paper delineates the cardinal principles, best practices, processes, and hierarchical roles pivotal for ensuring data integrity. In addition, we probe into the challenges that organizations grapple with and spotlight emerging technological innovations poised to bolster data governance and quality. Lastly, our exploration accentuates the indispensable cultural shifts and leadership imperatives that underpin effective governance and assiduous quality management of data.

2. Defining Data Governance and Quality

Data governance refers to the overall strategy, policies, standards, and processes that ensure high-quality data assets across the organization. It establishes accountability and oversight for managing data as a critical enterprise asset. Data governance helps align regulatory, operational, and strategic objectives with data strategies. Key activities include developing data policies, standards, and procedures; governing data architecture and quality; providing stewardship and ownership; and monitoring compliance. Effective data governance requires involvement across functions - from legal, compliance, IT, and lines of business to executive leadership.

Data quality encompasses the precision, entirety, coherence, and punctuality of the data validity and uniqueness of data (Wende, 2007). High-quality data that meets these characteristics enhances business value and reduces risks. Data quality management involves practices that ensure data adheres to quality requirements through its lifecycle - from creation, acquisition, and storage to processing, distribution, and archival. It provides standards for data quality via metrics, monitoring, issue resolution, and improvement initiatives (Wende & Otto, A Contingency Approach To Data Governance, 2007).

3. Significance in Data Engineering Context

In data engineering pipelines, large volumes of data flow through various phases like acquisition, storage, processing, and consumption. Governance and quality must be ingrained across these pipelines to drive confidence in the data products. Data engineers are deeply involved in developing and operating these pipelines. Hence, they must adopt governance practices and build quality into data infrastructure and processes (Dai, et al., 2016). Some key reasons are:

3.1. Compliance

Various regulations like GDPR mandate data governance through accountability, transparency, and quality. This requires implementing policies, access controls, lineage tracking, and quality checks in data platforms.

3.2. Trustworthy Analytics

Quality issues like errors, duplication, inconsistencies, and incompleteness can propagate and get magnified in downstream analytics, leading to incorrect insights. Governance and quality help prevent "garbage-in, garbage-out."

3.3. Metadata

Governance requires rich metadata with definitions, standards, and rules enabled by data catalogues, dictionaries, and linage tools. This aids discoverability and interoperability (Lis & Otto, 2020).

3.4. Monitoring

Continuous, automated data quality monitoring via statistical profiling, rules, and machine learning models helps identify issues early. Data engineers need to build these capabilities.

3.5. Automation

Technologies like data quality rules engines and machine learning can automate quality checks and corrections, freeing up engineers.

With robust governance and quality, data teams save significant time in non-value cleaning and reconciliation. Governance provides standards, while quality management helps systematically improve data assets (Koltay, 2016).

4. Fundamental Principles and Best Practices

Here is an elaboration on the fundamental principles and best practices for data governance and quality:

4.1. Business Alignment

Business outcomes and requirements should drive data quality initiatives rather than just IT preferences. The priorities and use cases should come from business teams, while technology teams enable implementation. This ensures governance and quality efforts deliver maximum business value.

4.2. Shared Accountability

Data quality cannot be the sole responsibility of IT teams. Business teams who enter or depend on data for decisions are equally accountable for reporting issues, resolving them at source, and implementing quality practices. A collaborative culture between data producers, consumers, and enablers is essential.

4.3. Data Lifecycle Approach

Data quality must be assessed and managed across the entire lifecycle - from creation, acquisition, storage, and processing to consumption and archival. For example, assess quality at intake, build checks into the ETL process, profile before analysis, retain integrity during archival, etc. This end-to-end view is critical.

4.4. Continuous Monitoring

Quality levels must be systematically measured via metrics and monitored through dashboards. Issues must be rapidly identified and resolved. Increase monitoring coverage through automation using data quality tools.

Issue resolution - Superficial bug fixes result in quality issues reappearing downstream. Root cause analysis to identify systemic gaps and address those to prevent recurrence.

Prevention over correction - Defining quality standards upfront and building validations into systems versus retrofitting quality is more efficient. Controls during entry and processing prevent issues downstream.

4.5. Incremental Improvement

Start with a few critical metrics and data sets. Quick wins build momentum for expanding systematically across other data assets based on value, risk, etc.

4.6. Reusable Frameworks

Leverage consistent governance models across data initiatives. Do not reinvent the wheel. Promote reuse of metrics, policies, standards, and patterns.

4.7. Risk-Based Approach

Priorities for data quality must be driven by business risk and impact analysis. Focus on high-value, sensitive, or compliance-related data first.

4.8. Master Data Foundation

High-quality customer, product, and financial master data are necessary for reliable downstream analytics. Bad master data amplifies downstream issues.

4.9. Security and Privacy

The governance framework must consistently incorporate data security controls, access policies, and privacy protections. These principles require change management, executive mandate, shared accountability, and robust processes. With persistent execution, data quality becomes an organizational capability.

5. Key Roles and Responsibilities

The critical roles involved are Chief Data Officer, Data/Domain Owners, Data Stewards, Data Engineers, Data Architects, Data Analysts, and Legal/Compliance.

5.1. Responsibilities

5.1.1. Chief Data Officer (CDO)

- Responsible for data strategy and governance at the executive level. Establishes policies and standards and focuses on data quality.
- Evangelizes the importance of data quality across the organization.
- Sponsors data governance programs and drives adoption top-down.
- Chairs data stewardship committees and councils that define governance practices.
- Secures funding and investments for improvement initiatives.
- Measures effectiveness of governance through quality metrics and benefits tracking.

5.1.2. Data/Domain Owners

- Business teams who generate, consume and are accountable for domain data.
- Define business data requirements and quality needs for their domain or functions.
- Participate in stewardship committees to evaluate proposals and issues.
- Implement data quality practices mandated by governance within their team.
- Fix data quality issues at the source systems under their control.

5.1.3. Data Stewards

- Cross-functional team that defines and oversees data standards and quality.
- Document data definitions, standards, rules, metrics, and SOPs for governance.
- Support data certification for critical data assets to assure quality compliance.
- Help troubleshoot data quality issues, analyze root causes, and guide remediation.
- Provide tools and training to data producers on quality practices.
- Monitor quality metrics and track issue resolution.

5.1.4. Data Engineers

- Implement data quality checks, automation, and monitoring per governance rules.
- Embed data quality capabilities within data infrastructure like ETL pipelines and models.
- Support integration of new data quality tools like profiling, cleansing, and matching.
- Monitor data quality metrics across the data lifecycle and report issues.
- Remediate quality issues found in upstream systems and pipelines.

5.1.5. Data Architects

- Develop overall data models, architecture principles, and standards for consistency.
- Define the master data model for critical domains like customer, product, finance, etc.
- Create data dictionaries, taxonomy, and metadata standards for governance.
- Ensure architectural components follow recommended data quality patterns and capabilities.

5.1.6. Data Analysts

- Analyze reports and dashboards to detect data quality issues that affect insights.
- Report upstream data issues found in consumption systems like BI tools.
- Validate reports and metrics for accuracy and completeness.
- Provide inputs to improve quality based on analytical needs and pain points.

5.1.7. Legal/Compliance

- Recommend policies based on regulatory and compliance needs like GDPR and CCPA.
- Conduct audits to ensure governance practices meet compliance requirements.
- Determine retention rules and access policies per regulatory guidelines.
- Enforce security standards for sensitive data like PII.
- Validate quality practices to meet compliance expectations around reporting accuracy.

5.2. Key Governance Processes Enabled by the Above Stakeholders

- Data quality standards - Dimensions, metrics, acceptance criteria, frequency
- Data policy and principles - Usage, integrity, retention, security, access
- Data models - Master, transactional, analytics, integration models
- Metadata standards - Define and maintain data taxonomy, dictionaries, lineage
- Issue tracking - Document issues found, severity, status, root cause, remediation
- Data quality tools - Select and implement automated profiling, monitoring, and metadata tools
- Training and communication - Increase quality awareness and skills across teams

The interplay between the roles, supported by standardized governance processes and executive sponsorship, can help engrain data quality accountability across the data lifecycle.

Stewards and IT teams enable the above processes while analysts and engineers embed quality in upstream pipelines. Collaboration between roles is vital for end-to-end governance.

6. Challenges in Implementation

- Lack of executive support - For data governance to be effective, it must be championed and funded by executive leadership. They must set the vision, strategy, and urgency for enterprise-wide governance. Without their active sponsorship, governance councils and working groups will lack authority and struggle to drive adoption. Executive focus on data quality and backing for improvement programs is also essential.
- Poor data culture - Data quality needs to be everyone's responsibility, not just IT's. A culture that values high-quality data avoids "data dumping" and fixes issues at source requires behavioral change across teams. It is difficult to promote shared accountability and break siloed attitudes.
- Data producers need to take more ownership, while consumers should provide feedback.
- Distributed systems - With data and workloads increasingly distributed across multi-cloud architectures in siloed groups, maintaining consistent governance standards becomes challenging. Data in legacy systems also creates fragmentation. Governance processes break down due to a lack of visibility and coordination across systems.
- Inconsistent metrics - Measuring data quality through standardized dimensions and metrics provides necessary visibility. But, consistent definitions and calculations across teams lead to clarity. The lack of automated monitoring using quality metrics also hampers systematic improvement.
- Technical complexity - With hundreds of upstream data sources, ETL pipelines, databases, and systems, implementing end-to-end governance with checks, controls, and monitoring is complex. Diverse technologies and integration points make it worse.
- Manual processes - Governance processes like validating data against standards, visual inspection, reconciliation, issue reporting, etc., remain predominantly manual. This leads to limited coverage, errors, delays, and rework.
- Legacy practices - Ingrained processes not designed for data quality continue unchecked.
- Siloed teams unwilling to adopt new governance practices and standards thwart progress.
- Skill gaps - Data engineers, stewards, and architects need diverse skills in data modeling, metadata management, and statistical quality techniques. Lack of these skills impedes design and implementation.

Addressing these requires a strategic focus on culture, skills, systems integration, metrics standardization, issue resolution processes, and increased automation.

7. Emerging Trends and Technologies

Several emerging technologies enable organizations to enhance data governance and quality in a scalable manner:

1. Machine learning and rules engines automate quality checks and issue remediation versus manual processes. They can also find complex data relationships and root cause issues.
2. Metadata management, data catalogs, and data lineage tools provide visibility into data assets and their usage across systems. This aids governance, reporting, and issue resolution.
3. Data virtualization creates abstraction layers that insulate consumers from underlying changes while providing transformation capabilities. This facilitates quality enhancement.
4. Data quality monitoring tools perform statistical profiling, custom checks, and anomaly detection over data. They generate metrics and dashboards to track quality over time.
5. Natural language interfaces allow users to state quality requirements and definitions in business terms rather than technical queries. This simplifies governance processes.
6. Cloud-scale data platforms provide built-in governance capabilities for access control, lineage, provenance, and compliance monitoring. Leveraging them accelerates implementation.

7. Augmented data management via platforms that automatically detect issues, recommend fixes, and enable remediation helps continuously improve quality.

Adopting the above technologies helps enhance quality, reduce manual efforts, and improve adoption across the data lifecycle. However, more than technology is needed to address the cultural and organizational challenges outlined earlier. Holistic governance frameworks and change management are vital for success.

8. Conclusion

Data has emerged as a strategic asset critical to the digital transformation of organizations. As data volumes grow across disparate systems, consistently governing and managing its quality provides tremendous value but poses challenges. This requires a systematic approach driven by executive leadership, shared business-IT accountability, and collaborative culture. Core principles include continuous monitoring, prevention versus correction, and incremental data quality improvement supported by standard processes.

Technologies like machine learning and cloud-scale data platforms are valuable enablers. However, cultural transformation and persistence are necessary to embed data governance and quality across the data value chain - from acquisition to consumption. With robust implementation, organizations can accelerate their data-driven ambitions and unlock more business value responsibly.

References

[1] Peter Ghavami, *Big Data Management: Data Governance Principles for Big Data Analytics*, Walter De Gruyter GmbH and Co KG, pp. 1-174, 2020. [Google Scholar] [Publisher Link]

[2] Miye Wang et al., "Big Data Health Care Platform with Multisource Heterogeneous Data Integration and Massive High-Dimensional Data Governance for Large Hospitals: Design, Development, and Application," *JMIR Medical Informatics*, vol. 10, no. 4, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Kristin Wende, "A Model for Data Governance-Organising Accountabilities for Data Quality Management," *Association for Information Systems Electronic Library*, vol. 80, pp. 1-10, 2007. [Google Scholar] [Publisher Link]

[4] Kristin Wende, and Boris Otto, "A Contingency Approach to Data Governance," *International Consultation on Incontinence Questionnaire*, pp. 163-176, 2007. [Google Scholar] [Publisher Link]

[5] Wei Dai et al., "Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking," *Information Technology: New Generations: 13th International Conference on Information Technology*, pp. 439-450, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[6] Dominik Lis, and Boris Otto, "Data Governance in Data Ecosystems-Insights from Organizations," *Association for Information Systems Electronic Library*, pp. 1-11, 2020. [Google Scholar] [Publisher Link]

[7] Tibor Koltay, "Data Governance, Data Literacy and the Management of Data Quality," *International Federation of Library and Institutions Journal*, vol. 42, no. 4, pp. 303-312, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[8] Soňa Karkošková, "Data Governance Model to Enhance Data Quality in Financial Institutions," *Information Systems Management*, vol. 40, no. 1, pp. 90-110, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sung Une Lee, Liming Zhu, and Ross Jeffery, "A Contingency-Based Approach to Data Governance Design for Platform Ecosystems," *Association for Information Systems Electronic Library*, pp. 1-15, 2018. [Google Scholar] [Publisher Link]

[10] Rene Abraham, Johannes Schneider, and Jan vom Brocke, "Data Governance: A Conceptual Framework, Structured Review, and Research Agenda," *International Journal of Information Management*, vol. 49, pp. 424-438, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ibrahim Alhassan, David Sammon, and Mary Daly, "Data Governance Activities: An Analysis of the Literature," *Journal of Decision Systems*, vol. 25, no. 1, pp. 64-75, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[12] John A. Pearce, and Shaker A. Zahra, "Board Composition from a Strategic Contingency Perspective," *Journal of Management Studies*, vol. 29, no. 4, pp. 411-438, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[13] Chunxia Wang, and Jian Xie, "Constructing a Computer Model for Discipline Data Governance using the Contingency Theory and Data Mining," *2021 4th International Conference on Information Systems and Computer Aided Education*, pp. 1967-1970, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Majid Al-Ruithe, Elhadj Benkhelifa, and Khawar Hameed, "A Systematic Literature Review of Data Governance and Cloud Data Governance," *Personal and Ubiquitous Computing*, vol. 23, pp. 839-859, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] Marijn Janssen et al., "Data Governance: Organizing Data for Trustworthy Artificial Intelligence," *Government Information Quarterly*, vol. 37, no. 3, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Zeljko Panian, "Some Practical Experiences in Data Governance," *World Academy of Science, Engineering and Technology*, vol. 62, no. 1, pp. 939-946, 2010. [Google Scholar] [Publisher Link]

[17] Boris Otto, "A Morphology of the Organisation of Data Governance," *Association for Information Systems Electronic Library*, pp. 1-13, 2011. [Google Scholar] [Publisher Link]

[18] Marina Micheli et al., "Emerging Models of Data Governance in the Age of Datafication," *Big Data and Society*, vol. 7, no. 2, pp. 1-15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Stephanie Russo Carroll, Desi Rodriguez-Lonebear, and Andrew Martinez, "Indigenous Data Governance: Strategies from United States Native Nations," *Data Science Journal*, vol. 18, no. 31, pp. 1-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] Steve Sarsfield, "*The Data Governance Imperative: A Business Strategy for Corporate Data*, IT Governance Publishing, pp. 1-161, 2009. [Google Scholar] [Publisher Link]

[21] Ibrahim Alhassan, David Sammon, and Mary Daly, "Data Governance Activities: A Comparison between Scientific and Practice-Oriented Literature," *Journal of Enterprise Information Management*, vol. 31, no. 2, pp. 300-316, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Krassimira Paskaleva et al., "Data Governance in the Sustainable Smart City," *Informatics*, vol. 4, no. 4, pp. 1-19, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[23] Huberman A. Michael, and Miles B. Matthew, "Data Management and Analysis Methods," *Handbook of Qualitative Research*, pp. 428-444, 1994. [Google Scholar] [Publisher Link]