International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

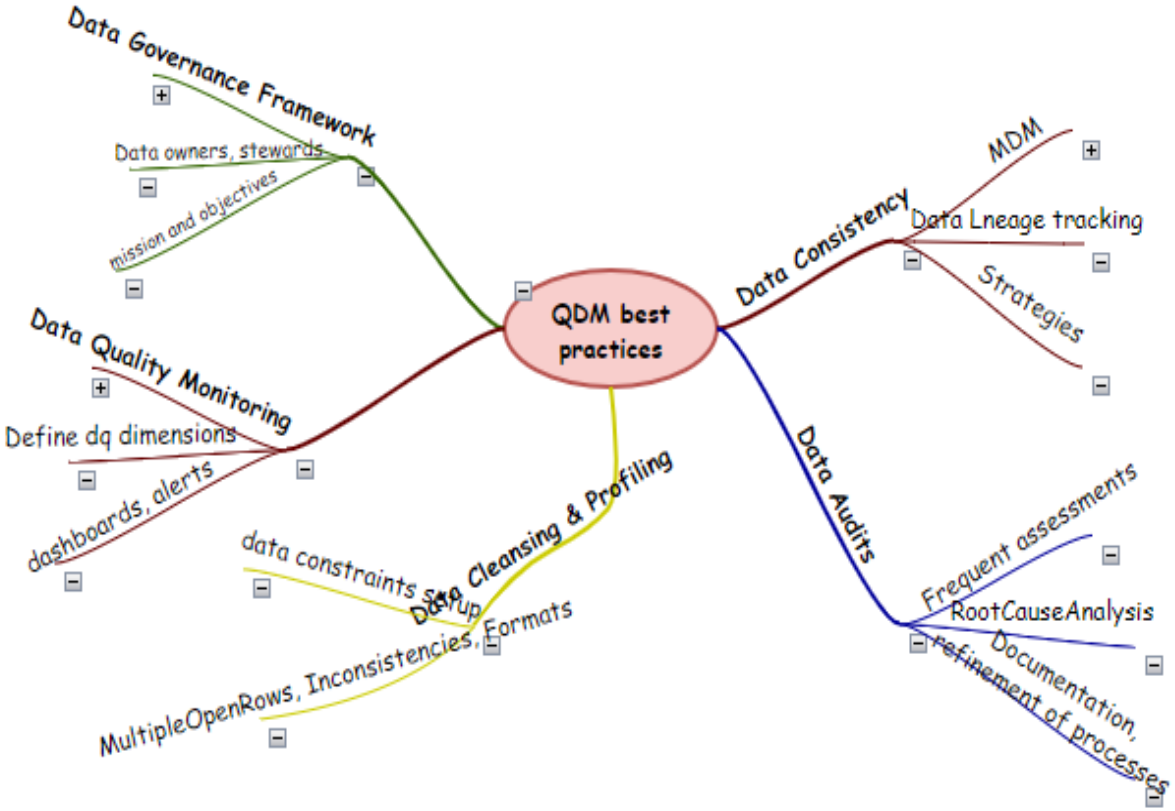
Quality Data Management (QDM) Best Practices for Preserving Data Integrity

Rameshbabu Lakshmanasamy

Senior Data Engineer, Verizon Data Services
ramesh.lakshman@gmail.com

Abstract:
Today in data-driven retail landscape, maintaining high-quality data is not optional but a mandatory practice that is the top priority. Data Integrity is crucial for making informed decisions, customer experience improvement, staying ahead of competition. This article outlines the best practices for quality data management in the IT organization, especially based on our real-time experiences from retail realm. Being given the responsibility of data quality of data marts and data warehouse, here am trying to capture the practices that we implemented and that helped us mitigate the risks and prevent any dangers of corrupt data. Let's dive into and explore what helped us in our journey of preserving quality data.

Keywords: QDM, Data Integrity, Data Governance, Compliance, Risk mitigation, Automation, Best practices.



Introduction:

Foundation of effective data quality management relies on few important steps – Establishing a Data Governance Framework, Data Consistency across touchpoints, Data Quality Monitoring process setup, Data Standardization, Data Profiling and Cleansing, Ensuring Data Accuracy at point of entry, Advanced Analytics, Regular data Audits, Compliance of Data Regulations. All these implementations in tandem greatly helped us in mitigating the risks, and helped us measure the data quality index.

Data Governance Framework:

Defining the governance principles and policies is the first step. Start with this by including subject matter experts of all key datasets and systems. First create the quality standards and metrics clearly defined.
Each metric should have tolerance level as well. Implementing data security and privacy measures is also emphasized.
Clearly defined roles and responsibilities for Data owners, Data stewards, Data policies, Data standards.

Data Quality Monitoring:

In retail or any domain, data often/usually flows between multiple systems. Main entry point would be the point-of-sale system or the shopping URL. This is a continuous process monitoring
Define key metrics relevant to data quality dimensions– like – completeness, accuracy, consistency, uniqueness, validity, integrity to name a few.

1. **Completeness:** Measure of all required data availability. For e.g. In our case, we had a quality control to measure 'Customer profile' if any does not have a valid email address or a phone number. If not, raise an alert.
2. **Accuracy:** Whether the data rightly reflecting the real-world event being talked about. E.g. measuring the number of mobile phones inventory in various systems/apps are in sync. If not, raise an alert.
3. **Consistency:** Same source of truth. Two or more representations of same metrics must be in sync. For e.g. the special offers for a particular segment, showing same price in website, and also in mobile-app backend database.
4. **Uniqueness:** Multiple open rows for same key in the main subject areas in data marts. E.g. For slowly changing dimensional data, usually the latest row is the updated/correct snapshot of truth. Ensure multiple rows for same key are not active in these cases. Else, raise a red flag.
5. **Validity:** measure the data if it conforms to the standard syntax (format, type, length etc.). E.g. Phone numbers not having same standards across systems, resulting in risk of showing it as incorrect numbers when try to do a match.
6. **Integrity:** The consistency and accuracy of the data over its entire course of data pipeline and life-cycle. E.g. Sales data and inventory level must go hand in hand. If they are not in sync, leads to potential issue in decision making of new orders.
7. **Relevance:** How relevant is the data if it meets potential user needs. E.g. While tracking preferred brands of customer, ensure tracking preferred product categories too for effective business decisions.

How did we implement this in our data warehouse and DataMart?

Well, once we defined clearly of above metric types, identify all the subject areas and its associated databases. Define the above metric categories in a lookup/reference table. Define the threshold levels until which it can be considered as allowable deviation for each data-control-id. Map each data-control-id to

quality-check-queries which will fire periodically in set frequency (some may be daily, some are weekly, some every few hours depending the criticality and necessity of the respective data elements). The more emphasis should be on the key quality-check-queries, which will measure the data across various defined touchpoints, and various systems, right from operational and reporting applications, spanned across marketing, sales, and inventory systems.

Based on above outcome, the results were published into a curated datasets, which feeds into real-time dashboards, available for the IT leaders to keep monitoring. Email / system alerts is also on top of this for production support teams for action in case of deviation from threshold levels for each quality control.

Continuous improvement is very important. Insights from all the metrics should help refine processes and keep raising the quality standards. We cannot afford to overlook this.

Data Profiling and Cleansing:

Robust ETL handling that regularly profiles data to identify anomalies, duplicates, inaccuracies, and inconsistencies will help with this. Though there is a huge list, the key things to consider are

1. Removal of duplicates
2. Fix the deviation of formats/syntaxes (e.g. phone , email, SSN to name a few)
3. Locate the missing information
4. Capture any other inconsistencies
5. Implement real-time address check for standardization
6. Setup constraints
7. Automation of data quality controls that we talked about in previous section.

Consistency across systems/applications:

From point of sale entry point or website entry, single source of truth must be implemented. Setup Quality controls that capture the key indices across the different systems and data marts to ensure they are in sync. Regular audit across data systems for inconsistencies is very crucial.

Data Lineage tracking and metadata management will come handy for implementing and ensuring data consistency across data pipelines from entry to reporting systems.

Regular Data Audits:

This is also part of continuous improvement aspect to keep doing regular periodical audits for data quality measure.

1. Perform regular assessments
2. Documenting and performing root-cause-analysis for frequent issues
3. Keep refining the QDM processes using the audit outcomes.
4. Brainstorming with all stakeholders on the Data Quality Control Checks that is setup which is backbone of all the DQ measures.

Data Quality mindset:

All sections/stakeholders should have a culture of data quality on whatever is being developed and setup. Quality data is tied to performance evaluations, and continuous reminders on the importance of this on business outcomes. Fostering a data quality culture will lessen the risks of incorrect decision making, and help organizations to stay relevant and growing in current competitive era.

Conclusion:

Everyone is aware of the value of data and its reliability and integrity, without which it is of no use to the business users and data consumers. Implementing above QDM best practices which we had setup and from our real time experience, would help us significantly gain a competitive edge. Without enhancing the decision-making capabilities, and customer experiences, it is difficult to stay relevant. Keep in mind, this is an ongoing process that will need continuous attention and refinement.

References:

1. Batini, C., & Scannapieco, M. (2016). "Data and Information Quality: Dimensions, Principles and Techniques." Springer.
2. DAMA International. (2017). "DAMA-DMBOK: Data Management Body of Knowledge." Technics Publications. The Data Management Association's guide is an industry-standard reference for data management practices, including a substantial section on data quality.
3. Loshin, D. (2010). "The Practitioner's Guide to Data Quality Improvement." Morgan Kaufmann. This guide offers practical advice on implementing data quality programs in organizations.
4. Sadiq, S. (Ed.). (2013). "Handbook of Data Quality: Research and Practice." Springer.
   This collection of research papers and practical case studies covers various aspects of data quality management.
5. TDWI (The Data Warehousing Institute): https://tdwi.org/ : Offers research, training, and resources on various data management topics, including data quality.
6. ISO 8000: https://committee.iso.org/home/tc184sc4 : The international standard for data quality, which provides guidelines and frameworks for data quality management.
7. MITIQ (MIT Information Quality): http://mitiq.mit.edu/ : Provides research and insights from MIT's Total Data Quality Management program.
8. Preyaa Atri, "Optimizing Financial Services Through Advanced Data Engineering: A Framework for Enhanced Efficiency and Customer Satisfaction", International Journal of Science and Research (IJSR), Volume 7 Issue 12, December 2018, pp. 1593-1596, https://www.ijsr.net/getabstract.php?paperid=SR24422184930