INTELLIGENCE · BUSINESS INTELLIGENCE

Data Quality as the Foundation of Business Intelligence

Business intelligence is only as good as the data beneath it. This article explains what data quality means, how poor data harms organisations, how to measure quality and how to improve it at its source.

ModQu research article · 1,567 words · 17 references (APA 7th edition)

Every report, dashboard and model in business intelligence rests on data. If the data are wrong, incomplete, out of date or misunderstood, the intelligence built on them will be wrong too, often in ways that are hard to see. A sales report that double-counts returns, a customer list full of duplicates or a forecast trained on mislabelled records can lead managers confidently in the wrong direction. **Data quality** is therefore not a technical detail but the foundation of everything business intelligence produces. This article explains what data quality means, why it deteriorates, how to measure it and how to improve it.

What data quality means

The simplest view equates quality with accuracy: data are good if they correctly record the facts. Research has shown that users judge data by much more. Wang and Strong (1996) asked data consumers what mattered to them and found that their judgements fell into four broad categories, summarised in Figure 1.

Four categories of data quality

Users judge data by far more than accuracy

Category	Example dimensions	Question a user asks
Intrinsic	Accuracy, objectivity, believability	Is it correct, and can I trust its source?
Contextual	Relevance, timeliness, completeness	Does it fit my task, now?
Representational	Interpretability, consistency	Can I understand it and compare it?
Accessibility	Access, security	Can I get it when I need it, safely?

Based on Wang and Strong (1996)

ModQu.com

Figure 1. Four categories of data quality. Accurate data can still fail users if they are late, irrelevant, hard to interpret or hard to obtain. Based on Wang and Strong (1996); table by ModQu.

The central insight is that quality means **fitness for use**. The same data can be good enough for one purpose and inadequate for another. Monthly sales totals accurate to within 1% may be fine for strategic planning but useless for reconciling accounts. Strong et al. (1997) showed that many data quality problems arise from the context in which data are used, not from errors in the data themselves: data produced for one purpose are reused for another without anyone checking whether they fit.

Wand and Wang (1996) offered a more formal foundation. They described an information system as a representation of the real world and defined data quality problems as failures of that representation: real-world states that are missing from the data, data that do not correspond to any real-world state, or real-world states mapped to the wrong data. Kent (1978) had earlier argued that data are never a simple mirror of reality: deciding what counts as a single customer, product or event involves choices, and different systems make them differently.

How poor data harms organisations

Redman (1998) described the impact of poor data quality on typical enterprises at three levels. **Operationally**, errors lead to wrong shipments, incorrect invoices and wasted effort spent finding and fixing mistakes. **Tactically**, poor data lead to poor decisions, because managers act on misleading reports. **Strategically**, they undermine trust: once managers find errors in reports, they stop relying on them and revert to instinct, and the investment in business intelligence is wasted.

Errors also propagate. Ballou and Pazer (1985) modelled how errors in input data combine and spread as data pass through processing steps, so that small inaccuracies in several sources can produce large errors in outputs. Sambasivan et al. (2021), studying practitioners building machine learning systems in high-stakes domains, documented what they

called **data cascades**: problems in data collection and preparation that compounded through later stages and caused failures that were discovered late and were costly to fix. They found that data work was often undervalued compared with modelling, even though it determined outcomes.

The consequences can be public. Herndon et al. (2014) found a spreadsheet error in an influential economic study that had excluded several countries from a key calculation; the corrected results weakened the study's main conclusion.

Why quality deteriorates

Data quality problems have recurring causes:

- **Entry errors.** People mistype, select the wrong option or skip fields, especially when entry is rushed or the purpose of a field is unclear.
- **Inconsistent definitions.** Different systems define and code the same thing differently.
- **Integration problems.** Combining data from several sources creates duplicates and conflicts. Fellegi and Sunter (1969) developed a statistical theory of **record linkage**, for deciding whether two records refer to the same person or entity, which remains the basis of much matching work.
- **Decay.** Data that were correct become wrong as the world changes: customers move, change names or close businesses.
- **Changes upstream.** A change to a source system, such as a new field or code, can silently break reports that depend on it.
- **No ownership.** When no one is responsible for a dataset, problems are noticed but not fixed.

Orr (1998) argued, from a systems perspective, that data quality depends on use: data that are used regularly, by people who notice and report errors, tend to improve, while data that are collected but rarely used tend to deteriorate.

Measuring data quality

Improvement requires measurement. Pipino et al. (2002) proposed combining objective measures, such as the share of records with missing values or failing validation rules, with subjective assessments by users of how well the data meet their needs. Lee et al. (2002) developed a methodology for information quality assessment that surveys users across quality dimensions and compares the results with benchmarks and between different groups, such as data collectors and data consumers, whose views often differ.

Practical measures include:

- **Completeness:** the share of required fields that are filled.
- **Validity:** the share of values that conform to rules, such as valid postal codes or dates.
- **Uniqueness:** the share of records that are not duplicates.
- **Consistency:** the share of records that agree across systems.
- **Timeliness:** the age of data when used, relative to need.
- **Accuracy:** the share of records that match a trusted reference or a sample checked against reality.

Batini et al. (2009), reviewing methodologies for data quality assessment and improvement, found that most share a common sequence: reconstruct the state of data and processes, measure quality, and then choose improvement actions based on costs and benefits.

Shewhart's (1931) methods of statistical process control apply here as well. Tracking quality measures over time, with limits that show normal variation, reveals when a process has changed and needs attention.

Improving quality at the source

English (1999) argued that the most effective way to improve data quality is to fix the processes that create data, not to clean data after the fact. Cleaning is necessary, but if the source process keeps generating errors, cleaning becomes a permanent and growing cost. English drew on quality management principles from manufacturing: prevent defects rather than inspect them out.

Practical steps include:

1. **Design entry for accuracy.** Validate data at the point of entry, use lists rather than free text where possible, and explain why each field matters.
2. **Assign ownership.** Khatri and Brown (2010) identify data quality as one of the core domains of data governance. Each important dataset needs an owner with authority to set standards and fix problems.
3. **Agree definitions.** Maintain a shared glossary of business terms and measures. DAMA International (2017) treats metadata and data quality management as core functions of data management.
4. **Monitor continuously.** Automate quality checks and track them over time.
5. **Close the loop.** When users find errors, route them back to the source process and fix the cause.
6. **Invest in data work.** Recognise and reward the work of preparing and maintaining data, which Sambasivan et al. (2021) found was often treated as less important than building models.

How much quality is enough?

Perfect data are neither possible nor necessary. Improving quality has costs, in staff time, system changes and slower data entry, and those costs rise steeply as data approach perfection. The right level of quality depends on the uses the data serve. Batini et al. (2009) emphasise that improvement actions should be chosen by weighing their costs against the benefits they bring.

A practical approach is to classify data by the consequences of error. Data used for financial reporting, regulatory filings, safety or decisions about individual customers need high accuracy and strong controls. Data used for broad trend analysis can tolerate more error, provided the errors are random rather than systematic and users know their likely size. What matters most is that users understand the quality of the data they are using: a figure known to be accurate within 5% is more useful than one presented as exact but of unknown reliability. Publishing quality measures alongside reports, and labelling data that do not meet the standard for a given purpose, helps users judge how much weight to give them.

An example

A wholesaler's customer database contains about 40,000 records. An assessment finds that about 8% are duplicates, 12% have outdated addresses and many lack an industry classification. As a result, sales reports overstate the number of active customers, marketing mailings are returned and analysis of sales by industry is unreliable.

The company merges duplicates using matching rules and a review of uncertain matches, then fixes the causes: sales staff can now search for existing customers before creating new ones, the industry field becomes mandatory with a defined list, and addresses are checked against a postal database at entry. A data steward in the sales team reviews quality measures monthly. Within a year, duplicates fall below 1%, and sales and marketing reports become trusted enough to drive decisions.

Conclusion

Data quality is the foundation of business intelligence. It means fitness for use, judged across accuracy, relevance, timeliness, interpretability and accessibility. Poor data harm operations, decisions and trust, and errors compound as they flow through systems. Measuring quality, assigning ownership and fixing the processes that create data are far more effective than repeated cleaning. Organisations that treat data quality as a continuing discipline, rather than a one-time project, build business intelligence that people can trust.

References

Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management Science, 31*(2), 150–162. https://doi.org/10.1287/mnsc.31.2.150

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), Article 16. https://doi.org/10.1145/1541880.1541883

DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications.

English, L. P. (1999). *Improving data warehouse and business information quality: Methods for reducing costs and increasing profits*. Wiley.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics, 38*(2), 257–279. https://doi.org/10.1093/cje/bet075

Kent, W. (1978). *Data and reality: Basic assumptions in data processing reconsidered*. North-Holland.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM, 53*(1), 148–152. https://doi.org/10.1145/1629175.1629210

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management*, *40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM*, *41*(2), 66–71. https://doi.org/10.1145/269012.269023

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82. https://doi.org/10.1145/269012.269025

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 39). ACM. https://doi.org/10.1145/3411764.3445518

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Van Nostrand.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099