DATA QUALITY EXPLAINED: MEASURING, ENFORCING & IMPROVING DATA QUALITY

Data drives business decisions that determine how well business organizations perform in the real world. Vast volumes of data are generated every day, but not all data is reliable in its raw form to drive a mission-critical business decision.

Today, data has a credibility problem. Business leaders and decision makers need to understand the impact of data quality. In this article, we will discuss:

- Data quality, particularly in the enterprise
- Measuring data quality
- Enforcing data quality, including getting started and key roles
- Improving data quality

Let's get started!

What is data quality?

Data Quality refers to the characteristics that determine the reliability of information to serve an intended purpose (often, in business these include planning, decision making, and operations).

Data quality refers to the utility of data as a function of attributes that determine its fitness and reliability to satisfy the intended use. These attributes—in the form of metrics, KPIs, and any other qualitative or quantitative requirements—may be subjective and justifiable for a unique set of use cases and context.

If that feels unclear, that's because data is perceived differently depending on the perspective. After all, the way you define a quality dinner, for instance, may be different from a Michelin-starred chef. Consider data quality from these perspectives:

- Consumer
- Business
- Scientific
- Standards
- Other perspectives

In order to understand the quality of a dataset, a good place to start is to understand the degree to which it compares to a desired state. For example, a dataset free of errors, consistent in its format, and complete in its features, may meet all requirements or expectations that determine data quality.

(Understand how data quality compares to data integrity.)

Data quality in the enterprise

Now let's discuss data quality from a standards perspective, as it is widely used particularly in the domains of:

- Database management
- Big data
- Enterprise IT

Let's first look at the definition of 'quality' according to the ISO 9000:2015 standard:

> *Quality is the degree to which inherent characteristics of an object meet requirements.*

We can apply this definition to data and the way it is used in the IT industry. In the domain of database management, the term 'dimensions' describes the characteristics or measurable features of a dataset.

The quality of data is also subject to external and extrinsic factors, such as availability and compliance. So, here's holistic and standards-based definition for quality data in big data applications:

> *Data quality is the degree to which dimensions of data meet requirements.*

It's important to note that the term dimensions does not refer to the categories used in datasets. Instead, it's talking about the measurable features that describe particular characteristics of the dataset. When compared to the desired state of data, you can use these characteristics to understand and quantify data quality in measurable terms.

For instance, some of the common dimensions of data quality are:

- **Accuracy.** The degree of closeness to real data.
- **Availability.** The degree to which the data can be accessed by users or systems.
- **Completeness.** The degree to which all data attributes, records, files, values and metadata is present and described.
- **Compliance.** The degree to which data complies with applicable laws.
- **Consistency.** The degree to which data across multiple datasets or range complies with defined rules.
- **Integrity.** The degree of absence of corruption, manipulation, loss, leakage, or unauthorized access to the dataset.
- **Latency.** The delay in production and availability of data.
- **Objectivity.** The degree with which data is created and can be evaluated without bias.
- **Plausibility.** The degree to which dataset is relevant for real-world scenarios.
- **Redundancy.** The presence of logically identical information in the data.
- **Traceability.** The ability to verify the lineage of data.
- **Validity.** The degree to which data complies with existing rules.
- **Volatility.** The degree to which dataset values change over time.

DAMA-NL provides a detailed list of 60 Data Quality Dimensions, available in PDF.

Why quality data is so critical

OK, so we get what data quality is – now, let's look at why you need it:

- **Cost optimization.** Poor data quality is bad for business and has a significant cost as it relates to time and effort. In fact, Gartner estimates that the financial impact of the average financial impact of poor data quality on organizations is around $15 million per year. Another study by Ovum indicates that poor data quality costs business at least 30% of revenues.
- **Effective, more innovative marketing.** Accurate, high-velocity data is critical to making choices about who to market to—and how. This leads to better targeting and more effective marketing campaigns that reach the right demographics.
- **Better decision-making.** A company is only as good as its ability to make accurate decisions in timely manner—which driven by the inputs you have. The better the data quality, the more confident enterprise business leaders will be in mitigating risk in the outcomes and driving efficient decision-making.
- **Productivity.** According to Forrester, "Nearly one-third of analysts spend more than 40 percent of their time vetting and validating their analytics data before it can be used for strategic decision-making." Thus, when a data management process produces consistent, high-quality data more automation can occur.
- **Compliance.** Collecting, storing, and using data poses compliance regulations and responsibilities, often resulting in ongoing, routine processes. Dashboard-type analytics stemming from good data have become an important way for organizations to understand, at a glance, your compliance posture.

How to measure data quality

Now that you know what you expect from your data—and why—you're ready to get started with measuring data quality.

Data profiling

Data profiling is a good starting point for measuring your data. It's a straight-forward assessment that involves looking at each data object in your system and determining if it's complete and accurate.

This is often a preliminary measure for companies who use existing data but want to have a data quality management approach.

Data Quality Assessment Framework

A more intricate way to assess data is to do it with a Data Quality Assessment Framework (DQAF). The DQAF process flow starts out like data profiling, but the data is measured against certain specific qualities of good data. These are:

- **Integrity.** How does the data stack up against pre-established data quality standards?
- **Completeness.** How much of the data has been acquired?
- **Validity.** Foes the data conform to the values of a given data set?
- **Uniqueness.** How often does a piece of data appear in a set?
- **Accuracy.** How accurate is the data?
- **Consistency.** In different datasets, does the same data hold the same value?

Using these core principles about good data as a baseline, data engineers and data scientists can analyze data against their own real standards for each. For instance, a unit of data being evaluated for timeliness can be looked at in terms of the range of best to average delivery times within the organization.

Data quality metrics

There are a few standardized ways to analyze data, as described above. But it's also important for organizations to come up with their own metrics with which to judge data quality. Here are some examples of data quality metrics:

- **Data-to-errors ratio** analyzes the number of errors in a data set taking into account its size.
- **Empty values** assess how much of the data set contains empty values.
- **Percentage of "dark data",** or unusable data, shows how much data in a given set is usable.
- **The time-to-value ratio** represents how long it takes you to use and access important data after input into the system. It can tell you if data being entered is useful.

(Learn more about dark data.)

How to enforce data quality

Data quality management (DQM) is a principle in which all of a business' critical resources—people, processes, and technology—work harmoniously to create good data. More specifically, data quality management is a set of processes designed to improve data quality with the goal of actionably achieving pre-defined business outcomes.

Data quality requires a foundation to be in place for optimal success. These core pillars include the following:

- The right organizational structure
- A defined standard for data quality
- Routine data profiling audits to ensure quality
- Data reporting and monitoring
- Processes for correcting errors in bad and incomplete data

Getting started

If you are like many organizations, it's likely that you are just getting settled in with big data. Here are our recommendations for implementing a strategy that focuses on data quality;

- **Assess current data efforts**. An honest look at your current state of data management capabilities is necessary before moving forward.
- **Set benchmarks for data.** This will be the foundation of your new DQM practices. To set the right benchmarks, organizations must assess what's important to them. Is data being used to super-serve customers or to create a better user experience on the company website? First, determine business purposes for data and work backward from there.
- **Ensure organizational infrastructure.** Having the proper data management system means having the right minds in place who are up for the challenge of ensuring data quality. For many organizations, that means promoting employees or even adding new employees.

DQM roles & responsibilities

An organization committed to ensuring their data is high quality should consider the following roles are a part of their data team:

- **The DQM Program Manager** sets the tone with regard to data quality and helps to establish data quality requirements. This person is also responsible for keeping a handle on day-to-day data quality management tasks, ensuring the team is on schedule, within budget, and meeting predetermined data quality standards.
- **The Organization Change Manager** is instrumental in the change management shift that occurs when data is used effectively, and this person makes decisions about data infrastructure and processes.
- **Data Analyst/Business Analyst** interprets and reports on data.
- **The Data Steward** is charged with managing data as a corporate asset.

Leverage technology

Data quality solutions can make the process easier. Leveraging the right technology for an

enterprise organization will increase efficiency and data quality for employees and end users.

Improving data quality: best practices

Data quality can be improved in many ways. Data quality depends on how you've selected, defined, and measured the quality attributes and dimensions.

In a business setting, there are many ways to measure and enforce data quality. IT organizations can take the following steps to ensure that data quality is objectively high and is used to train models that produce the profitable business impact:

- **Find the most appropriate data quality dimensions from a business, operational, and user perspective.** Not all 60 data quality dimensions are necessary for every use case. Likely, even the 12 included above are too many for one use case.
- **Relate each data quality dimension to a greater objective and goal.** This goal can be intangible, like user satisfaction and brand loyalty. The dimensions can be highly correlated to several objectives—IT should determine how to optimize each dimension in order to maximize the larger set of objectives.
- **Establish the right KPIs, metrics, and indicators** to accurately measure against each data quality dimension. Choose the right metrics, and understand how to benchmark them properly.
- **Improve data quality at the source.** Enforce data cleanup practices at the edge of the network where data is generated (if possible).
- **Eliminate the root causes that introduce errors and lapses in data quality.** You might take a shortcut when you find a bad data point, correcting it manually, but that means you haven't prevented what caused the issue in the first place. Root cause analysis is a necessary and worthwhile practice for data.
- **Communicate with the stakeholders and partners involved in supplying data.** Data cleanup may require a shift in responsibility at the source that may be external to the organization. By getting the right messages across to data creators, organizations can find ways to source high quality data that favors everyone in the data supply pipeline.

Finally, identify and understand the patterns, insights, and abstraction hidden within the data instead of deploying models that churn raw data into predefined features with limited relevance to the real-world business objectives.

Related reading

- BMC Machine Learning & Big Data Blog
- Data Analytics vs Data Analysis: What's The Difference?
- 3 Keys to Building Resilient Data Pipelines
- Data Management vs Data Governance: Main differences
- Big Data vs Analytics vs Data Science: What's The Difference?
- Data Visualization Guide, a series of tutorials on graphs, charts, Tableau Online & more